UNS UNIVERSIDAD NACIONAL DEL SUR

2025
Bahía Blanca, Argentina

Tesis de Doctorado en Ciencias de la Computación

Análisis Visual de Datos Multidimensionales

Ing. Antonella Soledad Antonini

Directora/or: Dra. Silvia Mabel Castro / Dra. María Luján Ganuza

Esta página ha sido intencionalmente dejada en blanco.

Prefacio

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctora en Ciencias de la Computación, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Ciencias e Ingeniería de la Computación durante el período comprendido entre el 10 de Septiembre del 2019 y el 24 de Febrero del 2025, bajo la dirección de la Dra. Silvia Mabel Castro y de la Dra. María Luján Ganuza.

Ing. Antonella Soledad Antonini

UNIVERSIDAD NACIONAL DEL SUR

Subsecretaría de Posgrado

La presente tesis ha sido aprobada el 26/06/2025, mereciendo la calificación de **10** (**Sobresaliente**)

Esta página ha sido intencionalmente dejada en blanco.

Agradecimientos

Deseo expresar mi más sincero y profundo agradecimiento a todas las personas e instituciones que han sido parte esencial en la realización de esta tesis.

A los jurados, por el tiempo dedicado a la lectura de la tesis, por sus comentarios y devoluciones que enriquecen mi trabajo.

A mis directoras, Silvia Castro y María Luján Ganuza, fuentes constantes de conocimiento, inspiración y apoyo a lo largo de este proceso. Agradezco profundamente a ambas por transmitirme su amor y dedicación por la docencia y la investigación, por su paciencia y por la libertad creativa con la que me permitieron trabajar. Ha sido un verdadero placer compartir este tiempo con ustedes, quienes me hicieron sentir siempre cómoda y me brindaron un ambiente cálido y acogedor. Su orientación, aliento y cercanía han sido invaluables para mí, y siempre recordaré con gratitud los momentos que compartimos.

A los profesionales del Departamento de Geología de la Universidad Nacional del Sur, el Dr. Ernesto Bjerg, la Dra. Gabriela Ferracutti, la Dra. Florencia Gargiulo, la Dra. Lucía Asiain y el Lic. Juan Tanzola, por su valiosa colaboración en los temas relacionados con la Visualización de Datos aplicada a las Geociencias. Valoro profundamente el conocimiento que me compartieron, así como su paciencia, dedicación y compromiso.

A la Universidad Nacional del Sur y al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), por brindarme las herramientas y recursos necesarios para mi desarrollo profesional. En especial, agradezco al CONICET por otorgarme la Beca Interna Doctoral que posibilitó la realización de esta tesis. También agradezco al Instituto de Ciencias e Ingeniería de la Computación (CONICET-UNS) por facilitarme un lugar de trabajo durante todo este proceso. Ha sido un privilegio poder acceder a una educación gratuita y de tan alta calidad.

A todos los profesores que, con su dedicación y esfuerzo, han dejado un valioso granito de arena en mi formación durante la carrera de grado y posgrado, les agradezco profun-

damente.

A mis compañeros y miembros del Laboratorio de Investigación y Desarrollo en Visualización y Computación gráfica (VyGLab), cuya cercanía, apoyo y calidez humana han sido esenciales para mi desarrollo académico y personal, contribuyendo de manera invaluable a lo largo de este proceso.

A mi familia y amigos, quienes me ofrecieron su apoyo incondicional, su amor y paciencia a lo largo de todo este proceso. Gracias por creer en mí, por estar siempre a mi lado y por celebrarme en cada paso, en cada pequeño logro. Su presencia y aliento han sido fundamentales para que pudiera llegar hasta acá.

A Fran, mi compañero de vida, por su apoyo incondicional a lo largo de todos estos años. Gracias por acompañarme en cada etapa, por brindarme calma en los momentos difíciles y por recordarme siempre lo verdaderamente importante. Tu amor, paciencia, comprensión y presencia constante fueron pilares fundamentales para que este camino fuera posible. Y aunque ya no me creas, prometo que no voy a seguir estudiando... *ponele*.

A todos ustedes, mi más profundo agradecimiento.

Antonella Antonini

Bahía Blanca, 26 de Junio de 2025

Resumen

El crecimiento exponencial en la generación y almacenamiento de datos ha planteado importantes desafíos en su procesamiento y análisis. En este contexto, la visualización de datos desempeña un papel fundamental al permitir la transformación de grandes volúmenes de datos en representaciones comprensibles para el ser humano. No obstante, a medida que los conjuntos de datos aumentan en tamaño y dimensionalidad, el diseño de visualizaciones efectivas se vuelve cada vez más complejo.

Con el objetivo de abordar estos desafíos y optimizar el proceso de visualización, esta tesis propone soluciones innovadoras que abarcan diversos aspectos clave. En primer lugar, se desarrolla un marco metodológico basado en tres pilares fundamentales de la visualización: los datos, su representación visual y los objetivos del análisis. Se realiza un análisis detallado de las técnicas de visualización existentes, organizándolas en una nueva taxonomía que facilita su clasificación y aplicación. Asimismo, con el propósito de unificar la interpretación y el uso de las tareas analíticas en visualización, se propone una taxonomía de tareas unificada que supera las diferencias terminológicas presentes en la literatura.

De manera complementaria, se diseña y desarrolla un sistema de recomendación que optimiza la selección de técnicas de visualización considerando de manera integral las características de los datos, los objetivos analíticos del usuario y sus preferencias en la representación visual. Este sistema ofrece una guía estructurada para la selección de representaciones gráficas adecuadas, permitiendo a usuarios expertos y no expertos crear visualizaciones expresivas y efectivas.

Finalmente, se presenta el diseño e implementación de herramientas especializadas para la visualización de datos multidimensionales, con un enfoque particular en el análisis de datos geológicos. Estas herramientas combinan técnicas tradicionales con modelos avanzados de aprendizaje automático, ofreciendo soluciones innovadoras para el análisis

de grandes conjuntos de datos.

Abstract

The exponential data generation and storage has posed significant challenges in data processing and analysis. In this context, data visualization plays a fundamental role in enabling the transformation of large volumes of data into human-comprehensible representations. However, as data sets increase in size and dimensionality, the design of effective visualizations becomes increasingly complex.

To address these challenges and optimize the visualization process, this thesis proposes innovative solutions encompassing several key aspects. First, a methodological framework is developed based on three fundamental pillars of visualization: data, their visual representation, and the objectives of the analysis. A detailed analysis of existing visualization techniques is conducted, organizing them into a new taxonomy that facilitates their classification and application. Additionally, to unify the interpretation and use of analytical tasks in visualization, a unified taxonomy of tasks is proposed, overcoming the terminological differences present in the literature.

In addition, a recommendation system is designed and developed that optimizes the selection of visualization techniques by comprehensively considering the characteristics of the data, the user's analytical objectives, and their preferences in visual representation. This system offers a structured guide for the selection of appropriate graphical representations, allowing expert and non-expert users to create expressive and effective visualizations.

Finally, the design and implementation of specialized tools for multidimensional data visualization are presented, with a particular approach to geological data analysis. These tools combine traditional techniques with advanced machine learning models, offering innovative solutions for the analysis of large datasets.

Esta página ha sido intencionalmente dejada en blanco.

Certificamos que fueron incluidos los cambios y correcciones sugeridos por los jurados.

Dra. Silvia Mabel Castro y Dra. María Luján Ganuza

Esta página ha sido intencionalmente dejada en blanco.

Abreviaturas

CVRS Comprehensive Visualization Recommendation System, Sistema de Recomendación Integral de Visualización

DA Datos Abstractos

DaV Datos a Visualizar

DC Datos Crudos

DMV Datos Mapeados Visualmente

DR Dimensionality Reduction, Reducción de Dimensionalidad

DV Datos Visualizados

GLC General Lines Coordinates, Coordenadas Generales de Líneas

MDS Multidimensional Scaling, Escalado Multidimensional

NP-GLC Non-Paired General Lines Coordinates, Coordenadas Generales de Líneas No Emparejadas

PCA Principal Component Analysis, Análisis de Componentes Principales

PCP Parallel Coordinate Plot, Gráfico de Coordenadas Paralelas

P-GLC Paired General Lines Coordinates, Coordenadas Generales de Líneas Emparejadas

SC Start Coordinates, Coordenadas Estrella

SOM Self-Organizing Maps, Mapas Auto-organizados

SVD Singular Value Decomposition, Descomposición de Valores Singulares

UMAP Uniform Manifold Approximation and Projection, Aproximación y Proyección Uniforme de Variedades

UVAM Unified Visual Analytics Model, Modelo Unificado de Análisis Visual

UVM Unified Visual Model, Modelo Unificado de Visualización

Esta página ha sido intencionalmente dejada en blanco.

Índice general

Índice de figuras

Esta página ha sido intencionalmente dejada en blanco.

Índice de tablas

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 1

Introducción

1.1. Contexto y Motivación

En la actualidad, nos encontramos inmersos en la era de los datos, donde la cantidad generada y almacenada crece a un ritmo sin precedentes. Los avances tecnológicos han impulsado un incremento exponencial en la producción de datos, llevando a volúmenes que superan cualquier estimación previa. En 2003, se calculaba que cada persona almacenaba aproximadamente 800 megabytes de datos digitales, duplicándose esta cifra cada tres años [LV03]. Hoy en día, esta tendencia no solo persiste, sino que se acelera: en 2023, la cantidad total de datos generados alcanzó los 120 zettabytes, y se proyecta que para 2025 esta cifra ascienda a 181 zettabytes.

Este crecimiento en la generación de datos no solo ha transformado la manera en que se almacenan y gestionan estos datos, sino que también ha introducido desafíos significativos en su procesamiento, análisis y aprovechamiento. Para que estos conjuntos de datos sean comprensibles y útiles para los seres humanos, es fundamental convertirlos en representaciones visuales adecuadas que permitan extraer información de forma clara y efectiva. Sin este paso, los datos crudos pueden resultar ininteligibles y difíciles de interpretar, incluso cuando su volumen es relativamente pequeño. En este sentido, la visualización de datos juega un papel crucial, ya que facilita la identificación de patrones, la formulación de hipótesis y la toma de decisiones fundamentadas, aprovechando así esta invaluable fuente de conocimiento.

En particular, en diversas áreas es habitual visualizar conjuntos de datos caracterizados por múltiples dimensiones, lo que introduce desafíos significativos en su repre-

sentación gráfica. A medida que la dimensionalidad de los datos aumenta, lograr una visualización efectiva se vuelve más complejo. Las técnicas tradicionales, diseñadas para datos de pocas dimensiones, muchas veces resultan insuficientes para capturar la estructura de estos conjuntos de datos. Uno de los principales obstáculos en la visualización de datos multidimensionales es la sobrecarga visual, donde la visualización se vuelve tan densa y compleja que los usuarios no pueden procesar toda la información de manera efectiva.

Otro desafío considerable es la pérdida de información que puede ocurrir al reducir las dimensiones, lo que afecta la calidad de la interpretación de los datos. Para facilitar el análisis de grandes volúmenes de datos, en muchos casos es necesario representarlos en un espacio de baja dimensionalidad, generalmente de dos o tres dimensiones [DKZ13]. Las representaciones bidimensionales son más fáciles de explorar, y la percepción humana está naturalmente adaptada al mundo tridimensional que nos rodea. En algunos casos, puede ser aceptable reducir la dimensionalidad eliminando variables, lo que Inselberg denominó "mutilar dimensiones" [GLI98]. Sin embargo, limitarse a un subconjunto de dimensiones podría resultar en una representación simplificada que no capture adecuadamente la complejidad inherente de los datos. Además, la selección de las dimensiones más relevantes es un proceso subjetivo, que depende del conocimiento profundo del dominio en cuestión y de los objetivos del análisis. En otros escenarios, es crucial analizar todas las dimensiones de manera conjunta, ya que solo al considerarlas en su totalidad se pueden revelar relaciones, patrones complejos y estructuras subyacentes que no serían evidentes si se redujeran o eliminaran dimensiones. En este contexto, la reducción de la dimensionalidad se presenta como una herramienta útil, al proyectar los elementos de datos desde un espacio de alta dimensión a uno de menor dimensión preservando la mayor cantidad posible de la información original.

La representación visual de las relaciones entre las variables también puede ser un desafío en el análisis de conjuntos de datos multidimensionales. Como señaló Jacques Bertin [Ber83], la información no solo consiste en elementos de datos individuales, sino también en las relaciones entre ellos. Sin embargo, a medida que el número de variables aumenta, muchas de las técnicas de visualización existentes no logran representar adecuadamente estas relaciones.

A pesar de estos desafíos, los avances en el poder computacional y la creciente disponi-

bilidad de herramientas y algoritmos especializados han facilitado considerablemente el análisis de datos multidimensionales. Comprender en profundidad los métodos disponibles, identificar sus capacidades y limitaciones, seleccionar cuidadosamente las técnicas más adecuadas, y desarrollar soluciones alternativas es crucial para extraer información relevante de conjuntos de datos complejos y aprovechar todo su potencial.

1.2. Objetivo General y Aportes

La era digital ha traído consigo una explosión sin precedentes en la generación y acumulación de datos, presentando retos significativos para su procesamiento y análisis. La visualización de datos desempeña un papel crucial, ya que permite convertir grandes y complejos conjuntos de datos en representaciones visuales comprensibles. Sin embargo, a medida que la dimensionalidad de los datos aumenta, el diseño de visualizaciones efectivas se convierte en un desafío cada vez mayor.

En este contexto, esta tesis propone soluciones innovadoras que optimizan el proceso de visualización mediante el desarrollo de un marco metodológico robusto, la realización de un análisis exhaustivo de las técnicas de visualización más utilizadas y sistemas de recomendación de técnicas existentes, y la implementación de herramientas innovadoras.

A lo largo de este trabajo se abordan los aspectos fundamentales de la visualización, estructurados bajo un marco metodológico basado en las tres preguntas clave *"¿Qué?-¿Por qué?-¿Cómo?"* definidas por Munzner [Mun14], que estructuran el proceso de visualización en función de los datos, su representación visual y los objetivos analíticos. Este enfoque proporciona una base sólida para diseñar visualizaciones efectivas, adaptadas a las necesidades y requerimientos específicos de los usuarios.

Además, se lleva a cabo un análisis exhaustivo de las técnicas de visualización más utilizadas, las cuales se agrupan en una nueva taxonomía que incluye enfoques basados en geometría, íconos, píxeles, jerarquías y grafos, y se propone una taxonomía unificada de tareas que facilita la interpretación y aplicación de las tareas analíticas en el diseño de visualizaciones. Asimismo, se diseña y desarrolla un sistema de recomendación integral que optimiza la selección de técnicas de visualización, considerando las propiedades de los datos, las preferencias de representación visual y los objetivos analíticos del usuario. Por otra parte, se desarrollan e implementan enfoques y herramientas especializadas para

la visualización de datos multidimensionales, con un enfoque particular en las Ciencias Geológicas. Estas soluciones combinan técnicas tradicionales de visualización con métodos avanzados de aprendizaje automático para mejorar el análisis de datos complejos.

A continuación se detallan los principales aportes de este trabajo de investigación:

1. **Relevamiento y análisis de los modelos de visualización y definición del marco metodológico.** Se define un marco metodológico basado en las tres preguntas clave *"¿Qué?-¿Por qué?-¿Cómo?"* definidas por Munzner [Mun14], con el fin de estructurar el proceso de visualización. A su vez, se lleva a cabo una exploración de los modelos de visualización presentes en la literatura, dentro del contexto de este marco propuesto. Se describe en detalle el Modelo Unificado de Análisis Visual (UVAM), que integra las características del Modelo Unificado de Visualización (UVM) con aspectos específicos del análisis visual [GULC23, MCFE03]. Este modelo presenta un marco conceptual esencial para la comprensión de los temas que se desarrollarán a lo largo de la tesis.

2. **Análisis de los tres pilares del marco *"¿Qué?-¿Por qué?-¿Cómo?"*:** Se lleva a cabo un análisis exhaustivo de los tres pilares fundamentales del marco propuesto: los datos, su representación visual y los objetivos dentro del proceso de visualización. Este análisis incluye una revisión detallada de la taxonomía de los datos, el estudio de los componentes que conforman la estructura visual, y un relevamiento exhaustivo de las diferentes taxonomías de tareas presentes en la literatura. Este relevamiento sirvió como base para desarrollar una taxonomía de tareas, que unifica las terminologías empleadas en la literatura para describir las diversas tareas analíticas. Además, se exploran las interacciones aplicables a los distintos estados y transformaciones del proceso de visualización.

3. **Relevamiento exhaustivo de técnicas de visualización para datos multidimensionales.** Se describen las técnicas de visualización más empleadas en el análisis de datos multidimensionales y se agrupan en una nueva taxonomía, que abarca enfoques basados en geometría, íconos, píxeles, jerarquías y grafos. El objetivo en este punto consiste en ofrecer una comprensión profunda de cada técnica, sus características y aplicaciones específicas.

4. **Relevamiento de sistemas de recomendación de técnicas de visualización.**

Se realiza un relevamiento de los sistemas existentes de recomendación de técnicas de visualización, evaluando sus metodologías y criterios de selección, tales como la naturaleza de los datos, los objetivos del análisis, el conocimiento del dominio y las preferencias de visualización del usuario. Este análisis establece el marco conceptual fundamental para la creación del sistema de recomendación propuesto.

5. **Diseño e implementación de un sistema de recomendación integral de visualización.** Se diseña y desarrolla un sistema de recomendación de técnicas de visualización que integra factores clave para la selección de las representaciones visuales más apropiadas. Este sistema se destaca por considerar los tres pilares fundamentales de la visualización en su recomendación: las características del conjunto de datos, los objetivos analíticos del usuario y sus preferencias de representación visual.

6. **Diseño e implementación de sistemas especializados para la visualización de datos multidimensionales.** Se diseñan e implementan diversos enfoques y herramientas innovadoras para la visualización de datos multidimensionales, con un enfoque particular en las Ciencias Geológicas. Estos sistemas combinan técnicas tradicionales de visualización con métodos avanzados de aprendizaje automático, ofreciendo soluciones integradas para la representación y análisis de conjuntos de datos complejos. Herramientas como *Spinel Web* [AGF+21], *VISUEL* [AGC22], *GLC-Frame* [LAGC24] y *SpinelVA* [ALF+24] se destacan como contribuciones clave, al facilitar la exploración y el análisis visual interactivo de grandes volúmenes de datos. Además, se ha formalizado una taxonomía que abarca todas las características capaces de describir cualquier técnica de visualización NP-GLC en 2D, así como las interacciones asociadas [ALGC23]. Esta taxonomía ha sido fundamental para estructurar el análisis de estas técnicas, y guiar el diseño y desarrollo de nuevas soluciones NP-GLC. Además, a través de casos de estudio específicos, se ha resaltado el potencial y el impacto de estos sistemas en el análisis de datos multidimensionales.

1.3. Estructura de la Tesis

En esta tesis incluimos los conceptos relevantes, de modo tal que la lectura de la misma sea autocontenida. A continuación describimos su estructura:

- **Capítulo 1 Introducción.** Este capítulo presenta los objetivos, el marco en el que se desarrolla el estudio y detalla las principales contribuciones de la investigación.

- **Capítulo 2 Los Tres Pilares de la Visualización.** En este capítulo, se propone un marco metodológico basado en las tres preguntas clave *"¿Qué?-¿Por qué?-¿Cómo?"* definidas por Munzner [Mun14], con el objetivo de estructurar el proceso de visualización. Se realiza un análisis exhaustivo de los diferentes modelos de visualización para el análisis visual presentados en la literatura, en el contexto de este marco metodológico. Asimismo, se describe en detalle el Modelo Unificado de Análisis Visual (UVAM), que integra el Modelo Unificado de Visualización (UVM) con aspectos específicos del análisis visual [GULC23, MCFE03], proporcionando un marco conceptual clave para la comprensión de los temas que se abordarán a lo largo de la tesis.

- **Capítulo 3 Datos y Tareas - Dos Pilares de la Visualización.** Este capítulo se enfoca en los dos primeros pilares fundamentales del marco metodológico propuesto: los datos y el objetivo del análisis. En particular, se aborda la representación de datos dentro del proceso de visualización, utilizando como referencia el Modelo Unificado de Análisis Visual (UVAM). Además, se realiza un análisis exhaustivo de la taxonomía propuesta por Munzner [Mun14] para clasificar los conjuntos de datos y atributos en el ámbito de la visualización, lo cual resulta esencial para identificar las propiedades de los datos que determinan la selección de las técnicas visuales adecuadas.

  A continuación, se presenta un relevamiento detallado de las taxonomías de tareas existentes, que sirvió como base para el desarrollo de una taxonomía unificada, la cual se expone en el capítulo 6. Finalmente, se realiza un análisis de las interacciones posibles en los distintos estados y transformaciones del proceso de visualización.

- **Capítulo 4 Fundamentos de la Representación Visual.** El capítulo se centra en el tercer pilar del marco metodológico propuesto: la representación visual de los datos. Se exploran en detalle los componentes que conforman la estructura visual, tales como las marcas, los canales visuales y el sustrato espacial. Además, se analizan las técnicas más comúnmente utilizadas para la visualización de datos multidimensionales, las cuales se agrupan en una nueva taxonomía que abarca enfoques

basados en geometría, íconos, píxeles, jerarquías y grafos.

- **Capítulo 5 Sistemas de Recomendación de Técnicas de Visualización.** En este capítulo, se lleva a cabo un análisis exhaustivo de los sistemas de recomendación de técnicas de visualización presentes en la literatura, examinando las metodologías y enfoques actuales que orientan la selección de las técnicas más adecuadas. Este análisis establece el marco conceptual fundamental para el desarrollo del sistema de recomendación que se presenta en el capítulo siguiente.

- **Capítulo 6 Sistema de Recomendación Integral de Visualización (CVRS).** Se diseña e implementa un sistema integral de recomendación de técnicas de visualización que considera de manera integral los tres pilares fundamentales del marco metodológico propuesto: el objetivo de análisis, las características de los datos y las preferencias individuales de representación visual. Este sistema representa un avance significativo en el campo de la visualización de datos, al combinar flexibilidad con un sólido fundamento metodológico y establece una base robusta para futuras mejoras y desarrollos en el área.

  Además, como se ha señalado previamente, la literatura presenta diversas taxonomías de tareas que emplean terminologías distintas para describir tareas que, en esencia, son conceptualmente similares. En este contexto, se propone una nueva taxonomía unificada que facilita tanto la interpretación como la aplicación de las tareas analíticas en el diseño de visualizaciones efectivas.

- **Capítulo 7 Casos de Estudio.** Este capítulo aborda el diseño e implementación de enfoques y herramientas desarrolladas para enfrentar los desafíos del análisis y la visualización de datos multidimensionales, con un enfoque particular en las Ciencias Geológicas.

  Entre las herramientas destacadas se incluyen *Spinel Web* [AGF+21], *VISUEL* [AGC22], *GLC-Frame* [LAGC24] y *SpinelVA* [ALF+24], las cuales combinan técnicas tradicionales de visualización con métodos avanzados de aprendizaje automático, y facilitan la exploración interactiva de grandes volúmenes de datos, proporcionando nuevas perspectivas para el análisis y la visualización de datos multidimensionales. Asimismo, se exploran en profundidad las técnicas GLC (*General Lines*

Coordinates) y se formaliza una taxonomía que abarca las características fundamentales para describir cualquier técnica de visualización NP-GLC (*Non-Paired General Lines Coordinates*) en 2D, incluidas las interacciones asociadas, lo que ha sido esencial para estructurar y orientar el desarrollo de nuevas soluciones en este campo [ALGC23].

Además, se presentan distintos casos de estudio que ilustran la aplicación práctica de estas herramientas en el ámbito de las Ciencias Geológicas, destacando su capacidad para optimizar la visualización de datos geoquímicos.

- **Capítulo 8 Conclusiones y Trabajo Futuro.** En este capítulo se presentan las conclusiones de nuestro trabajo y el trabajo futuro a realizar.

Capítulo 2

Los Tres Pilares de la Visualización

2.1. Introducción

En la actualidad, la creciente disponibilidad de grandes volúmenes de datos ha transformado significativamente la forma en que recopilamos, almacenamos y analizamos la información. Diariamente, se generan enormes cantidades de datos provenientes de diversas fuentes, que incluyen sensores, redes sociales y experimentos científicos, entre otras. Sin dudas, esta avalancha de información presenta nuevas oportunidades y desafíos considerables.

En este contexto, la visualización de datos se ha convertido en una herramienta fundamental para facilitar el análisis y comprensión de grandes volúmenes de datos. Aprovechando la capacidad humana para procesar información visual de manera rápida, la visualización transforma datos abstractos en representaciones expresivas y efectivas. Una representación visual es expresiva cuando transmite exclusivamente la información relevante contenida en los datos. Es decir, no crea ni retiene información adicional, sino que refleja de manera objetiva lo que necesitamos para realizar una tarea específica. Por otro lado, una representación visual es efectiva cuando se adapta a nuestros sistemas motores y sensoriales, es decir, a nuestra capacidad de observar e interactuar con el entorno. Dado que tratamos con representaciones visuales, es crucial tener en cuenta las características del sistema visual humano. En este sentido, la efectividad se refiere a la facilidad con la que podemos extraer la información necesaria de una representación visual para llevar a cabo nuestra tarea.

El proceso de visualización se refiere a la serie de pasos mediante los cuales los datos

se transforman en representaciones visuales, es decir, ilustra los estados que los datos atraviesan desde su formato bruto hasta su representación visual final. En la literatura, se han propuesto varios modelos para la visualización de datos, que han permitido avances interesantes en la formalización y semi-automatización del proceso de visualización.

En este capítulo, se presenta una síntesis de los principales modelos de visualización propuestos en la literatura. Este análisis revela que todos los modelos o procesos de visualización pueden estructurarse utilizando el marco *"¿Qué?-¿Por qué?-¿Cómo?"* presentado por Munzner [Mun14]. La importancia de este marco reside en su capacidad para describir los modelos mentales que empleamos como diseñadores de visualizaciones: comenzamos definiendo qué datos vamos a analizar *(¿Qué?)*, luego identificamos la información valiosa a extraer y el objetivo del análisis *(¿Por qué?)*, y culminamos con las decisiones sobre la representación visual de los datos *(¿Cómo?)*, incluyendo la selección de técnicas de visualización, codificaciones y opciones de diseño específicas. Este marco no solo guía el proceso de visualización, sino que también garantiza una alineación coherente entre los datos, el objetivo y su representación visual, lo que resulta en visualizaciones efectivas.

Luego, se presenta en detalle el Modelo Unificado de Análisis Visual (UVAM, por sus siglas en inglés, Unified Visual Analytics Model) propuesto por Ganuza *et al.* [GULC23], que incorpora características específicas del análisis visual al Modelo Unificado de Visualización (UVM, por sus siglas en inglés) de Martig *et al.* [MCFE03]. Se detallan los diferentes estados del modelo, así como las transformaciones y operaciones realizadas en cada uno de ellos para facilitar el análisis visual y la interacción del usuario. Este modelo proporcionará el marco conceptual para los temas tratados en capítulos posteriores.

2.2. Los Tres Pilares de la Visualización

Tamara Munzner propone un marco de tres pilares fundamentales para la visualización de datos, que orienta el proceso de trasformación de los datos en representaciones visuales efectivas [Mun14]. La autora menciona que este marco, denominado *"¿Qué?-¿Por qué?-¿Cómo?"*, no debe ser visto como una imposición restrictiva, sino como una guía útil para el desarrollo de visualizaciones. El marco proporciona un enfoque que ayuda a los diseñadores a crear visualizaciones que comuniquen la información de manera efectiva y expresiva, considerando todos los aspectos importantes del proceso.

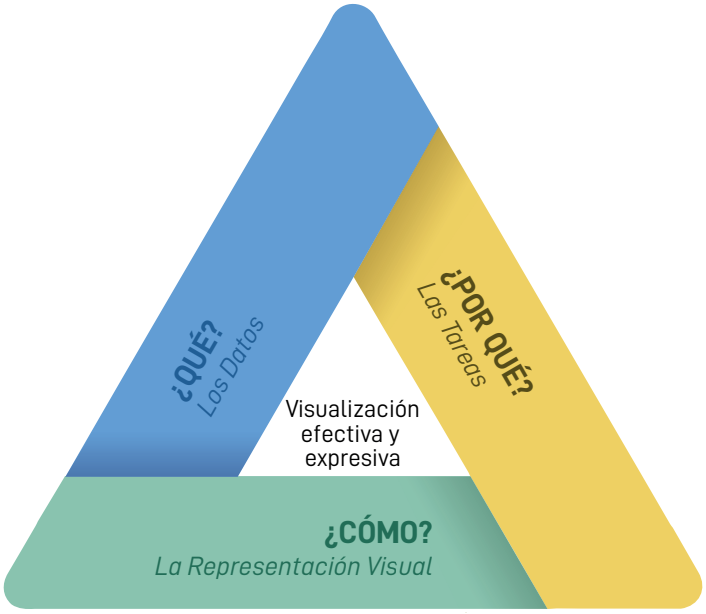


Figura 2.1: Los tres pilares de la visualización. Los modelos de visualización pueden estructurarse utilizando el marco *"¿Qué?-¿Por qué?-¿Cómo?"*: los datos, las tareas y la representación visual. Considerar en conjunto estos tres pilares permite diseñar visualizaciones tanto efectivas como expresivas.

El marco se estructura en torno a tres ejes fundamentales, como se ilustra en la figura 2.1. El primer pilar, el *"¿Qué?"*, se refiere a los datos que se van a visualizar. Esta etapa implica un análisis detallado de la calidad, naturaleza y características de los datos, incluyendo su estructura, volumen y complejidad. Comprender a fondo los datos es crucial porque la naturaleza de estos influirá directamente en las decisiones de diseño posteriores. Aquí, es crucial preguntarse: *¿Qué datos visualizará el usuario final en la representación producida?*. Esta reflexión guía el diseño hacia los datos más relevantes y significativos para el usuario.

El segundo pilar, el *"¿Por qué?"*, se centra en la razón detrás de la visualización. Para diseñar una visualización efectiva, es necesario comprender en profundidad el propósito de la visualización y qué se espera lograr con ella. Entre las tareas se pueden incluir la comparación de valores específicos, la identificación de tendencias, así como la detección de anomalías o patrones, entre otras. Comprender el *"¿Por qué?"* ayuda a guiar el diseño y asegura que la visualización cumpla con las necesidades del usuario. En esta etapa, se debe considerar: *¿Por qué es necesario desarrollar esta visualización para el usuario final? ¿Es para descubrir lo que los datos revelan, presentar hallazgos a otras personas,*

o simplemente por interés? ¿Estamos buscando tendencias, valores atípicos, similitudes o correlaciones?. Estas preguntas son fundamentales para definir el propósito y la funcionalidad de la visualización.

Finalmente, el pilar del *"¿Cómo?"* aborda las técnicas, métodos de visualización y la representación visual que se empleará para representar los datos y alcanzar los objetivos identificados. Una vez que se ha determinado qué datos se van a visualizar, la siguiente pregunta es cómo se van a representar esos datos. Esto abarca las técnicas y los métodos de visualización que se utilizarán, como gráficos de líneas, diagramas de flujo y tablas, entre otros. La elección de la representación visual adecuada es crucial para facilitar la comprensión y el análisis de los datos. Además, este pilar permite alinear las visualizaciones con las preferencias del usuario, asegurando una presentación personalizada y de calidad. Los diseñadores deben tener en cuenta aspectos específicos como la elección de las marcas, los canales visuales y el sustrato espacial. En esta fase, es importante plantearse: *¿Cómo debemos diseñar la visualización y las interacciones? ¿Cómo definiremos las opciones de diseño que conformarán la vista final?*.

Integrar los tres pilares es esencial para lograr una visualización efectiva y expresiva. Considerar apropiadamente cada pilar garantiza que la visualización no solo sea técnicamente correcta, sino también efectiva en la comunicación de los datos y en el cumplimiento de las tareas del usuario.

2.3. Modelos de Visualización

En la literatura se han propuesto diversos modelos para la visualización de datos, que han impulsado avances significativos en la formalización y automatización del proceso de visualización basado en datos de entrada.

En 1999, Card *et al.* [CMS99] introdujeron un *pipeline* fundamental para la visualización de datos (ver figura 2.2). Este modelo se considera una base esencial para la visualización de datos describiendo, paso a paso, el proceso para generar representaciones visuales. El *pipeline* se alinea perfectamente con los tres pilares del marco metodológico propuesto. En primer lugar, se enfoca en definir qué datos se van a visualizar. En esta etapa, los *Datos Crudos* se transforman en tablas detalladas que incluyen metadatos relevantes. A continuación, el modelo aborda la *Forma Visual*, donde los *Mapeos Visuales*

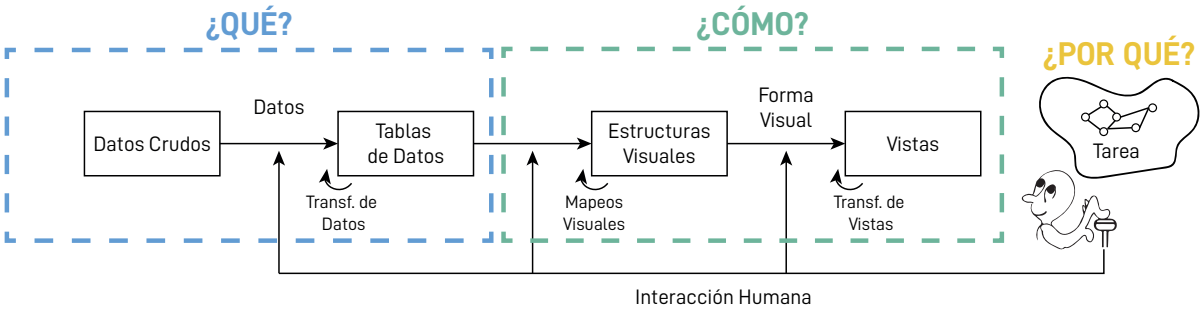


Figura 2.2: Modelo de visualización definido por Card *et al.* [CMS99] en el contexto de los tres pilares de la visualización. Figura adaptada de [CMS99].

convierten las tablas en estructuras visuales, integrando sustratos espaciales, marcas y otros métodos de codificación gráfica. Finalmente, las *Estructuras Visuales* se transforman en *Vistas* interactivas presentadas al usuario a través del sistema visual humano. A diferencia de otros modelos, el *pipeline* de Card enfatiza las interacciones humanas y considera explícitamente el objetivo o tarea final. Las interacciones se llevan a cabo en función de la tarea, permitiendo que los usuarios ajusten la visualización para cumplir con sus objetivos específicos.

Ed Chi y John Riedl [CR98, Chi00] proponen un modelo de estados de datos para el diseño de técnicas de visualización que permite al usuario interactuar en todas las etapas del proceso. Este modelo distingue entre etapas de datos (representadas por paralelogramos oscuros en la figura 2.3), transformaciones de datos y operadores en cada etapa. Las cuatro etapas de datos—*Valor, Abstracción analítica, Abstracción de visualización y Vista*—corresponden a los diferentes estados que los datos atraviesan durante el proceso de visualización. Estos estados se transforman sucesivamente a través de tres tipos de transformaciones: *Transformación de datos, Transformación de visualización y Transformación de mapeo visual*. Los operadores en cada etapa (*Operador de etapa de valor, analítica, de visualización y de vista*) actúan sin alterar la estructura de los datos procesados. La *Transformación de visualización* une dos componentes del modelo: el espacio de datos (indicado por la línea punteada celeste) y el espacio de vista (representado por la línea punteada verde). Esta transformación convierte los datos en una forma que se puede visualizar.

Años más tarde, Martig *et al.* [MCFE03] presentan un Modelo Unificado de Visualización (UVM) que abarca tanto los procesos como los estados de los datos, aplicable a cualquier tipo de visualización, sin importar el campo específico de origen. Este modelo,

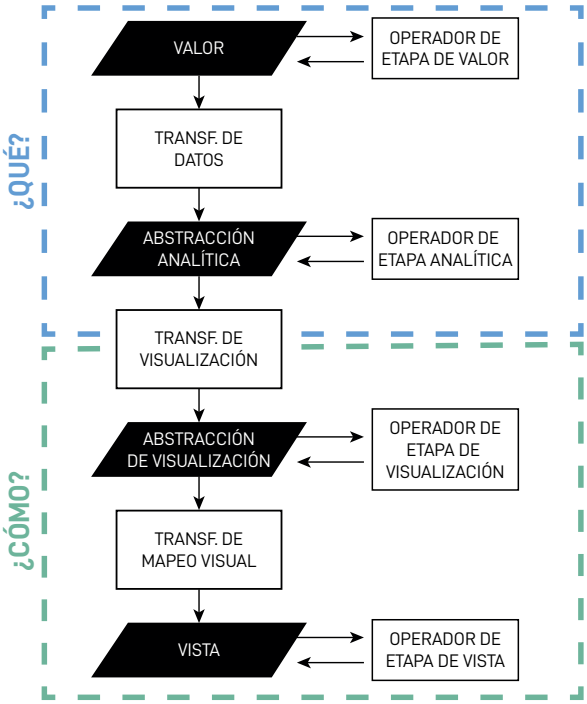


Figura 2.3: Modelo de estados de datos para la visualización de información presentado por Chi *et al.* [Chi00] en el contexto de los tres pilares de la visualización. Figura adaptada de [Chi00].

ilustrado en la figura 2.4, se representa como un flujo que describe los diferentes estados que los datos atraviesan a lo largo del proceso de visualización. El modelo incluye cinco estados: *Datos Crudos, Datos Abstractos, Datos a Visualizar, Datos Mapeados Visualmente y Datos Visualizados*. Además, define cuatro transformaciones que facilitan la transición entre estos estados: *Transformación de Datos Crudos a Datos Abstractos (Transf. DC-DA), Transformación de Datos Abstractos a Datos a Visualizar (Transf. DA-DV), Transformación de Mapeo de Visualización (Transf. Mapeo Visual) y Transformación de Visualización (Transf. de Visualiz.)*. Posteriormente, Ganuza *et al.* [GULC23] introducen el Modelo Unificado de Analítica Visual (UVAM), el cual combina el UVM [MCFE03] con las características específicas de la analítica visual. Este modelo modifica el UVM añadiendo operaciones especiales dentro de los estados para respaldar el análisis visual. Este modelo se ilustra en la figura 2.6 y se explora en detalle en la sección 2.4.

Por su parte, Tamara Munzner presenta un modelo anidado de cuatro niveles [Mun09], que inicia el proceso de visualización desde la caracterización del problema en lugar de enfocarse directamente en los datos. En el siguiente nivel, se realiza el mapeo de estos datos a operaciones abstractas y tipos de datos. El tercer nivel se centra en el diseño de la

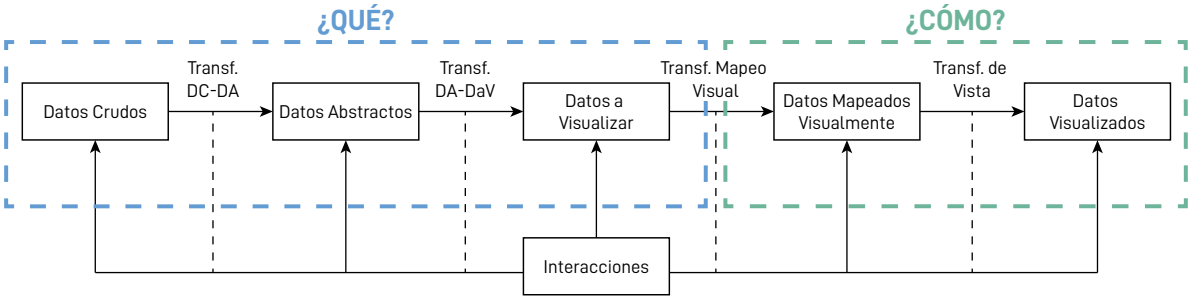


Figura 2.4: Modelo Unificado de Visualización (UVM) definido por Martig *et al.* [MCFE03] en el contexto de los tres pilares de la visualización. Figura adaptada de [MCFE03].

codificación visual y la interacción necesarias para soportar estas operaciones. Finalmente, en el nivel más interno, se desarrolla el algoritmo para ejecutar las técnicas de manera automática y eficiente. La figura 2.5 ilustra este modelo anidado.

Tras un análisis de los modelos de visualización presentados, se identifican los tres pilares fundamentales que sustentan el proceso de visualización. A pesar de las diferencias en la terminología, la complejidad y los enfoques metodológicos, los datos, la representación visual y las tareas emergen como elementos constantes en cada uno de los modelos estudiados. Estos tres pilares no solo son esenciales para la estructuración de los modelos de visualización, sino que también proporcionan un marco que permite el desarrollo de visualizaciones efectivas y expresivas.

En primer lugar, los datos actúan como el punto de partida común en la mayoría de los modelos. Sin importar la denominación empleada, ya sea “Datos Crudos”, “Valor” o simplemente “Datos”, se reconoce la existencia de información que requiere procesamiento y transformación para ser analizada. El segundo pilar es la tarea o el propósito que motiva la visualización. Aunque en algunos modelos este aspecto puede no ser del todo explícito,

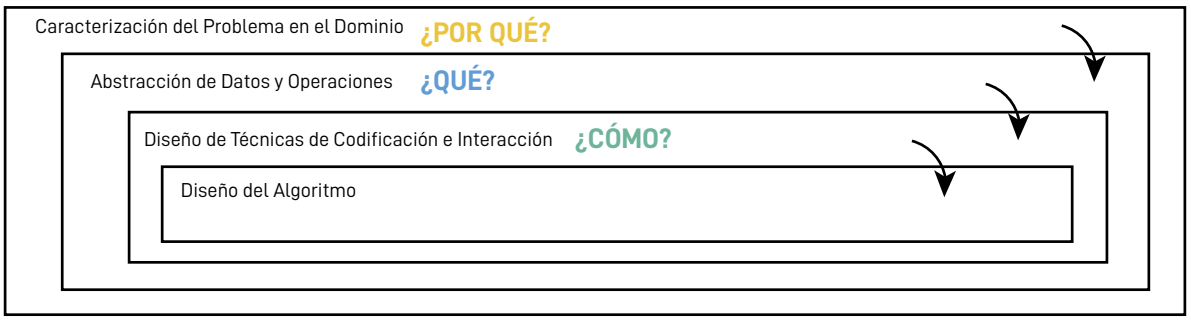


Figura 2.5: Modelo anidado de visualización definido por Munzner [Mun09]. Figura adaptada de [Mun09].

siempre está presente. Este último aspecto se manifiesta en la interacción del usuario, la generación de conocimiento o la caracterización del problema. Representa el objetivo final del proceso de visualización y análisis, orientando la decisión sobre qué datos procesar y cómo representarlos visualmente para cumplir con un propósito analítico específico. Finalmente, el tercer pilar se refiere a la representación de los datos para su visualización. Este aspecto se manifiesta de diversas formas en los modelos, desde los "Mapeos Visuales" propuestos por Card *et al.* [CMS99], hasta la "Transformación de Mapeo Visual" presente tanto en el Modelo Unificado de Visualización (UVM) como en el Modelo Unificado de Analítica Visual (UVAM). Este pilar encapsula la transformación de datos abstractos en representaciones visuales interpretables, constituyendo el núcleo de la visualización.

La presencia de estos tres pilares a lo largo de la evolución de los modelos de visualización enfatiza su trascendencia esencial.

2.4. Modelo Unificado de Análisis Visual (UVAM)

En tanto los modelos de visualización vistos anteriormente ilustran el proceso de transformación de datos a representaciones visuales, un flujo de trabajo de análisis visual pone énfasis en el razonamiento analítico, así como en la toma de decisiones con interfaces visuales interactivas.

Como se mencionó en la sección 2.3, el Modelo Unificado de Análisis Visual (UVAM) combina el Modelo Unificado de Visualización (UVM) con las particularidades propias del análisis visual. Este modelo se representa como un flujo entre los diversos estados por los que transitan los datos durante el proceso. Además, define las interacciones del usuario con los datos y sus representaciones intermedias, el control ejercido por el usuario sobre las transformaciones y las visualizaciones, y añade operaciones dentro de los estados para respaldar el análisis visual. La figura 2.6 ilustra el Modelo Unificado de Análisis Visual.

A continuación, se detallarán los distintos componentes del modelo, el cual constituirá la base para explicar y comprender mejor los contenidos abordados en los capítulos posteriores.

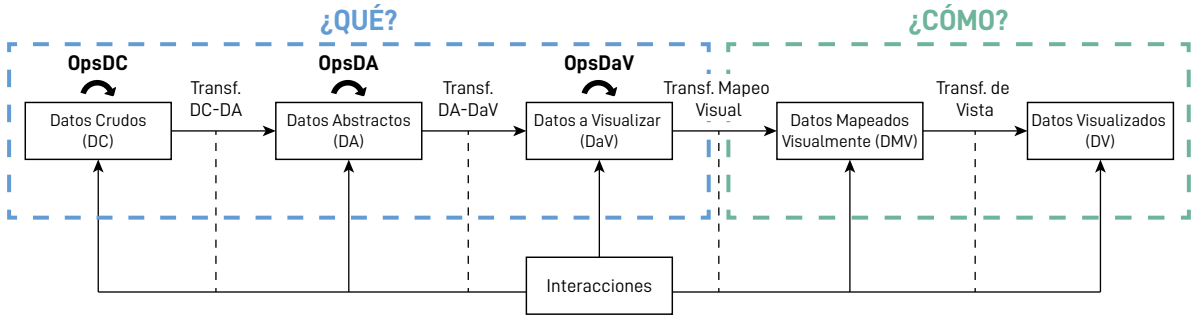


Figura 2.6: Modelo Unificado de Análisis Visual (UVAM) presentado por Ganuza *et al.* [GULC23] en el contexto de los tres pilares de la visualización. Figura adaptada de [GULC23].

2.4.1. Los Estados

Los datos atraviesan diferentes estados a lo largo del proceso. Un cambio en el estado no siempre se debe a diferencias estructurales, sino a los diferentes roles que cumplen los datos dentro del proceso completo.

- *Datos Crudos (DC).* Los DC corresponden al estado inicial de los datos en el proceso de visualización. Estos datos provienen de distintos dominios de aplicación o de visualizaciones previas, lo que resulta en una amplia variedad de formatos y fuentes diversas.

- *Datos Abstractos (DA).* Este es un estado intermedio de los datos, en el cual se encuentran en un formato manejable pero aún no es posible su visualización. Los DC, o un subconjunto de ellos, se almacenan en una representación adecuada para facilitar su gestión en las etapas posteriores del proceso de visualización. En este estado, además de los datos de interés, también se encuentran disponibles los metadatos generados durante la transformación de los datos.

- *Datos a Visualizar (DaV).* Son todos los datos que estarán presentes en la visualización. Este conjunto puede estar constituido por todos los DA o solo por una parte de ellos. Es importante considerar que se pueden generar múltiples conjuntos de DaV a partir de un mismo conjunto de DC.

- *Datos Mapeados Visualmente (DMV).* Son los datos que se obtienen al mapear visualmente los Datos a Visualizar. En esta etapa toma relevancia el tercer pilar fundamental del marco metodológico presentado en la sección 2.2.

Los DaV se enriquecen con la *estructura visual*, información necesaria para su representación visual. La *estructura visual* comprende el *sustrato espacial* y el *sustrato gráfico*. El *sustrato espacial* se refiere a la superficie subyacente en la que se muestran los elementos visuales y sus atributos (organización del espacio, diseño y tipo de ejes, y dimensiones del espacio). El *sustrato gráfico* incluye los elementos gráficos utilizados para representar los datos en la vista (*marcas visuales*) y sus propiedades gráficas asociadas (*canales visuales*, que controlan la apariencia de las marcas). Para un mismo conjunto de Datos a Visualizar pueden existir varios conjuntos de Datos Mapeados Visualmente. Esta característica es relevante en el marco de un proceso de exploración, donde el usuario puede o necesita comparar y analizar distintas maneras de visualizar un mismo conjunto de datos. Características fundamentales de la representación visual se describen en el capítulo 4.

- *Datos Visualizados (DV).* Si bien ésta es la última etapa del modelo, no es la final del proceso de visualización, ya que constituye el espacio de exploración para el usuario. Es importante enfatizar que para el mismo conjunto de DMV, se puede generar más de un DV mediante la aplicación de diferentes técnicas que satisfagan las características o las restricciones especificadas en el estado anterior.

2.4.2. Las Transformaciones

Las transformaciones son los procesos que permiten que los datos pasen de un estado a otro.

- *Transformación DC-DA.* Esta transformación se encarga de convertir los datos crudos (DC) del formato del dominio de la aplicación a un formato interno manejable a lo largo del resto del proceso. El resultado es un conjunto de datos abstractos (DA) provenientes de fuentes externas de datos u otras visualizaciones. Esta transformación puede crear un nuevo conjunto DA o incorporar nuevos datos a uno existente.

- *Transformación de DA-DaV.* Esta transformación permite la selección de los datos a visualizar y realizar tareas como proyección y filtrado. Se generan nuevos conjuntos de DaV manteniendo el mismo conjunto de DA.

- *Transformación de Mapeo Visual.* Esta transformación es fundamental en el proceso de visualización y permite al usuario definir cómo desea visualizar los datos. Se encarga de la representación visual, es decir, establecer qué estructuras visuales son apropiadas, qué atributos se mapearán espacialmente y cómo, y qué marcas y canales se utilizarán.

- *Transformación de Vista.* Esta transformación se encarga de generar la representación visual en pantalla, según lo expresado en el estado de DMV. Una vez seleccionada la técnica de visualización que soporta las restricciones presentes en el estado anterior, se aplica para generar la vista. Para un conjunto determinado de DMV, puede haber varias técnicas.

2.4.3. Las Operaciones

Para ampliar el proceso del Modelo Unificado de Visualización e incluir soporte de operaciones especiales sobre los datos para llevar a cabo el análisis visual, el Modelo Unificado de Análisis Visual incorpora operaciones dentro de los estados relacionados con los datos. Estas operaciones están representadas por flechas oscuras en la figura 2.6, que parten de un estado y regresan al mismo.

- *Operaciones en los Datos Crudos (OpsDC).* Los DC constituyen la entrada al proceso de visualización, proviniendo de diferentes dominios de aplicación en un formato dado. Estos DC pueden ser inexactos, poco fiables o ruidosos. Se deben llevar a cabo ciertas actividades de preprocesamiento de datos para resolver estos problemas y satisfacer los requisitos para las etapas siguientes. Las OpsDC a menudo se relacionan con operaciones como conversiones de datos de un formato a otro, limpieza de datos (para identificar y abordar problemas de calidad de datos en esta etapa), integración de datos (para fusionar datos de muchas fuentes), etc.

- *Operaciones en los Datos Abstractos (OpsDA).* Las OpsDA corresponden a operaciones realizadas en los DA. El resultado de estas operaciones podría ser, por ejemplo, datos derivados o mejorados. Estas operaciones también corresponden a métodos centrados en el análisis como la reducción de dimensiones, regresión, agrupamiento o clustering en un conjunto de datos, extracción de características,

muestreo de datos, etc. En este estado, además de tener los datos de interés, también se dispone de los metadatos generados como resultado de las operaciones.

- *Operaciones en los Datos a Visualizar (OpsDV).* Las OpsDV corresponden a operaciones aplicadas a los datos que estarán presentes en la visualización. Son las mismas operaciones que en el estado anterior. Sin embargo, dado que el DaV puede constituir un subconjunto de DA y las operaciones pueden ser costosas en términos de tiempo y/o espacio, realizarlas en esta etapa podría mejorar significativamente la eficiencia del proceso.

2.5. Conclusiones

En este capítulo se propone un marco metodológico basado en las tres preguntas clave definidas por Munzner [Mun14]: "*¿Qué?-¿Por qué?-¿Cómo?*", con el objetivo de estructurar el proceso de visualización. En este contexto, se realiza un análisis exhaustivo de los modelos de visualización más relevantes presentados en la literatura, revelando que, a pesar de sus diferencias en terminología y complejidad, todos pueden ser estructurados en torno a los tres pilares fundamentales de la visualización: el *"¿Qué?"* (los datos), el *"¿Por qué?"* (las tareas u objetivos), y el *"¿Cómo?"* (la representación visual). Estos pilares son esenciales para la estructuración de los modelos de visualización, y proporcionan un marco conceptual que facilita el desarrollo de visualizaciones efectivas y expresivas.

Además, se presenta en detalle el Modelo Unificado de Análisis Visual [GULC23], el cual amplía el UVM [MCFE03] para integrar las particularidades del análisis visual. Este modelo establece una base sólida para abordar los temas que se desarrollarán a lo largo de la tesis.

Capítulo 3

Datos y Tareas: Dos Pilares de la Visualización

3.1. Introducción

Como se detalló en el capítulo anterior, el proceso de visualización de datos se basa en tres pilares fundamentales: los datos, las tareas y la representación visual. En este capítulo, nos centraremos en los dos primeros pilares, abordando las preguntas *¿Qué datos vamos a visualizar?* y *¿Por qué es necesaria esta visualización?*, dejando el análisis del tercer pilar para el capítulo siguiente. Este estudio se enmarca en el Modelo Unificado de Análisis Visual (UVAM) [GULC23] y se contextualiza en el enfoque *"¿Qué?-¿Por qué?-¿Cómo?"* descrito en el capítulo 2.

En lo que respecta al primer pilar de la visualización, Munzner ofrece una clasificación exhaustiva de conjuntos de datos y atributos en el contexto de la visualización [Mun14]. En su análisis, destaca la importancia de entender tanto el tipo de datos que se va a visualizar como su semántica. La semántica de los datos se refiere a su significado en el mundo real. Por ejemplo, una palabra puede identificar un producto, una ubicación o una categoría, mientras que un número podría representar una fecha, una cantidad, o una coordenada espacial. Por otro lado, Munzner define el tipo de dato como su interpretación estructural y establece una clasificación en tres niveles: el nivel de los datos, el nivel de los conjuntos de datos y el nivel de los atributos. Entre las diversas clasificaciones de datos presentes en la literatura, consideramos que ésta es, sin duda, la más completa.

Para abordar el segundo pilar, *¿Por qué es necesaria esta visualización?*, realizaremos

un análisis detallado de las diversas taxonomías de tareas propuestas en la literatura. La tarea de visualización se refiere al objetivo o propósito detrás de la visualización de datos: ¿Qué preguntas necesitamos responder? ¿Qué decisiones deben tomarse a partir de esta visualización?. Estas taxonomías de tareas han sido desarrolladas para ayudar a estructurar y comprender las necesidades del usuario en el proceso de visualización.

Al considerar tanto los datos como las tareas, aseguramos que la visualización sea relevante, funcional y alineada con los objetivos específicos del usuario.

3.2. Los Datos

En la actualidad, nos enfrentamos a un aumento exponencial en la cantidad de datos generados y almacenados. A medida que las tecnologías de almacenamiento y recolección de datos mejoran, la cantidad de información disponible supera nuestra habilidad para procesarla y utilizarla de manera efectiva. Un reto sumamente desafiante es la identificación de métodos y modelos que permitan convertir esos datos en información útil y confiable. Gran parte del diseño de visualizaciones se encuentra determinado por la naturaleza de los datos disponibles.

En el contexto del UVAM [GULC23], el tratamiento de los datos se concentra en los tres primeros estados del proceso (ver figura 3.1). En el estado de *Datos Crudos (DC)*, se dispone de información en su forma original, sin procesar, a menudo en un formato que carece de una estructura adecuada para una visualización efectiva. Al transformarse en *Datos Abstractos (DA)*, se define una estructura apropiada para estos datos, aplicando operaciones que pueden generar datos derivados o enriquecidos. Finalmente, en el estado de *Datos a Visualizar (DaV)*, se seleccionan aquellos datos que estarán presentes en la visualización. Estos tres estados están específicamente orientados a los datos, a *"¿Qué?"* se va a visualizar.

A nivel de los datos, Munzner [Mun14] distingue cinco tipos de datos, a saber: *elementos* o *ítems*, *atributos*, *enlaces* o *conexiones*, *posiciones* y *grillas*. Un *atributo* es alguna propiedad o característica específica que se puede observar, medir o registrar. El salario de un empleado, el precio de un vehículo, el número de ventas, la duración de una canción o la temperatura, son ejemplos de atributos. Un *elemento* es una entidad individual que es discreta, como una fila en una tabla o un nodo en una red. Por ejemplo, los elementos

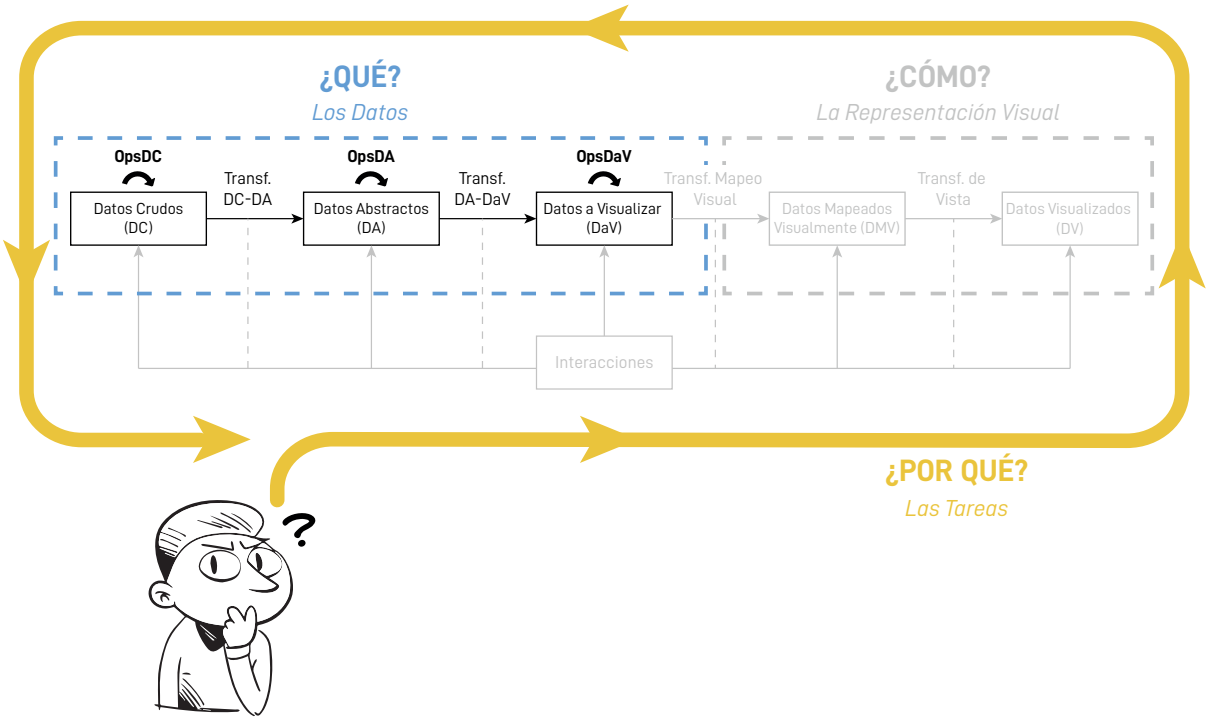


Figura 3.1: Los datos y tareas en el Modelo Unificado de Análisis Visual (UVAM) [GULC23].

pueden ser personas, sucursales, dispositivos electrónicos o ciudades. Un *enlace* es una relación entre elementos, típicamente dentro de una red. Una *grilla* es una estructura que representa el muestreo de un conjunto de datos sobre un espacio continuo permitiendo dividir un dominio continuo en celdas o regiones discretas. Y, por último, una *posición* es un dato espacial, que proporciona una ubicación en el espacio bidimensional (2D) o tridimensional (3D). Por ejemplo, una posición podría ser el par latitud-longitud que describe una ubicación geográfica o tres números *(x, y, z)* que especifican una ubicación en el espacio tridimensional.

Posteriormente, estos datos se combinan para dar lugar a distintos conjuntos de datos, los cuales se describen en detalle a continuación.

3.3. Conjuntos de los Datos Abstractos

En el contexto del UVAM [GULC23], el procesamiento y la transformación de los DC generan los *Datos Abstractos (DA)*. Esta etapa intermedia, denominada *Transformación de Datos Crudos a Datos Abstractos (Transf. DC-DA)* según [GULC23], es crucial, ya que organiza los datos en una estructura interna manejable en el proceso de visualización. Los

DA adoptan representaciones más manejables, como tablas, redes o formas geométricas, lo que facilita su manipulación en las etapas posteriores.

Un conjunto de datos, o *dataset*, se define como una colección de observaciones relacionadas, organizadas y formateadas para un propósito particular [CSK+20]. De acuerdo con [Mun14], los conjuntos de datos pueden clasificarse en cuatro tipos principales: *tablas*, *redes*, *campos* y *geometría*. La figura 3.2 ilustra estos tipos básicos de conjuntos de datos.

Como se mencionó anteriormente, los tipos básicos de conjuntos de datos se derivan de combinaciones de los tipos de datos fundamentales: elementos, atributos, enlaces, posiciones y grillas. En una *tabla*, las celdas están organizadas mediante elementos y atributos. Mientras que, en una *red*, los elementos, a menudo denominados nodos, están conectados por enlaces. Los *campos* continuos presentan una grilla basada en posiciones espaciales, con atributos en sus celdas. Por su parte, la *geometría* espacial solo proporciona información sobre la posición.

Además, Munzner señala otras maneras de agrupar elementos, como clústeres, conjuntos y listas. No obstante, no se profundiza en estos enfoques, ya que su análisis se centra en los tipos básicos, los cuales constituyen un excelente punto de partida para describir y comprender la parte del *qué* en el proceso de visualización.

Una vez que los datos han sido estructurados, el proceso avanza hacia la siguiente etapa: la selección del conjunto de datos que se visualizará, conocido como *Datos a Visualizar (DaV)*. Este conjunto puede estar compuesto por todos los DA o por un subconjunto de ellos. De un mismo conjunto de DA, es posible generar múltiples conjuntos de DaV, lo que facilita la exploración de diferentes regiones del espacio de información y su comparación. La identificación de los datos más relevantes para la tarea de visualización es un paso crucial, ya que influye de manera directa en la efectividad de la representación final.

3.3.1. Tablas

Los datos sin procesar pueden encontrarse en diversas formas, como en hojas de cálculo, archivos de texto, o registros de bases de datos. La primera etapa del Modelo Unificado de Análisis Visual implica organizar estos datos en estructuras internas que facilitan su posterior representación visual. Una de las formas más comunes de estructurar estos datos es en *tablas*. En una tabla, cada fila corresponde a un elemento de datos, mientras que cada columna representa un atributo del conjunto de datos. Cada celda de

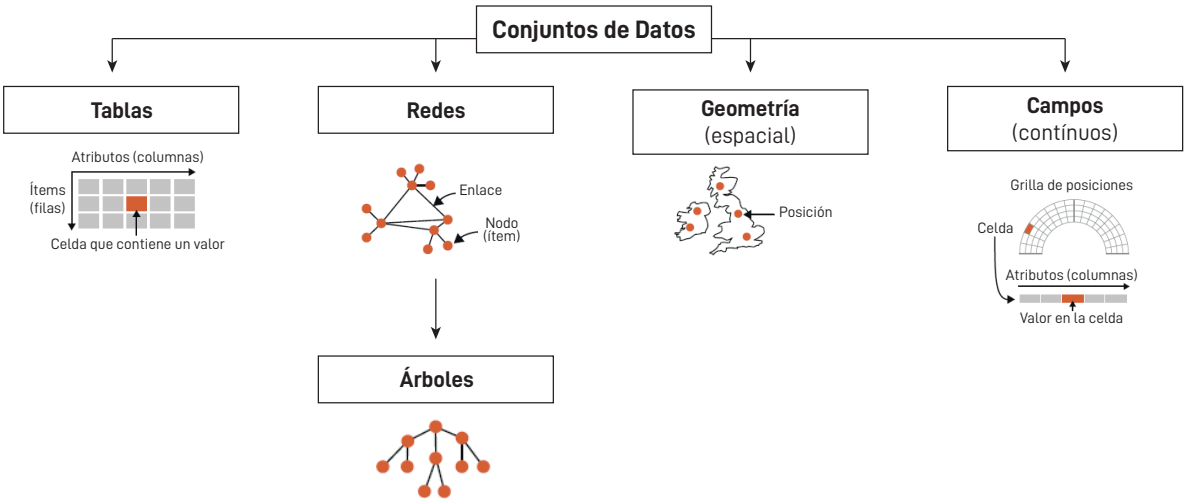


Figura 3.2: Clasificación de conjuntos de datos de Munzner [Mun14].

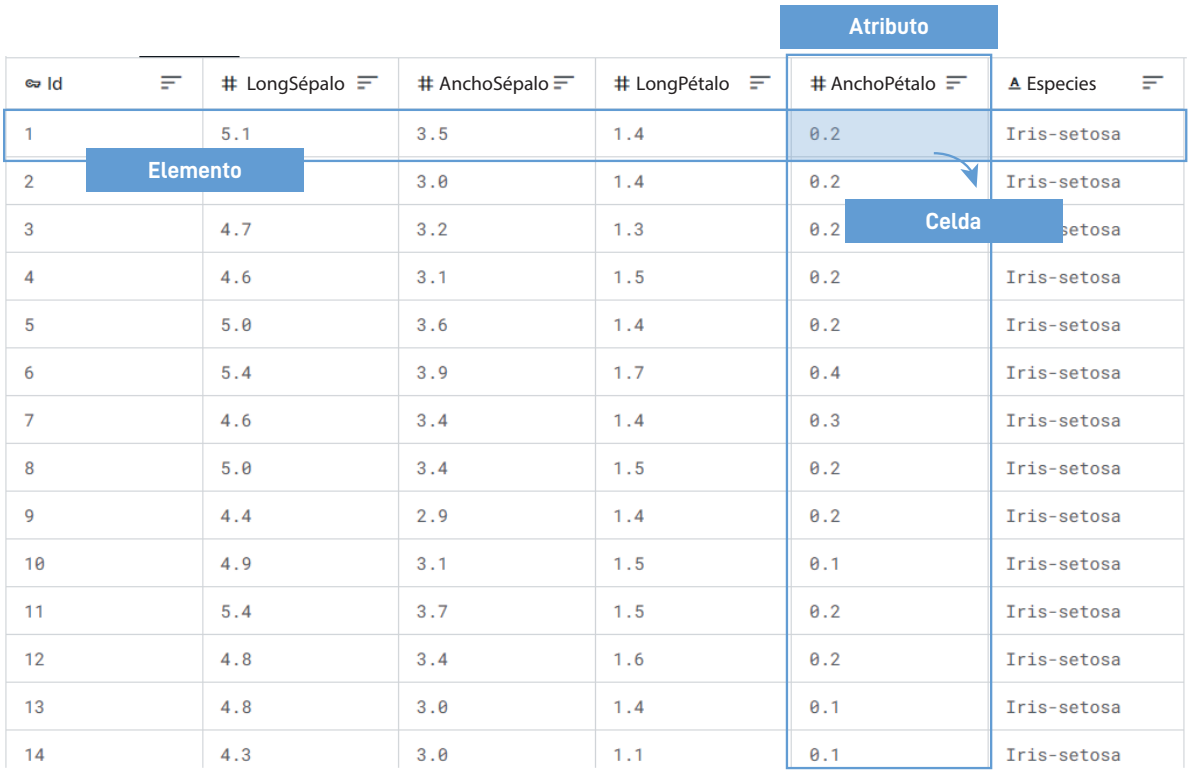
la tabla se define por la intersección de una fila y una columna —es decir, un elemento y un atributo— y contiene el valor correspondiente a ese par.

La figura 3.3 muestra una sección de la tabla del conjunto de datos Iris [Fis88], donde las filas corresponden a las observaciones individuales y las columnas representan las variables asociadas a estas observaciones, como el largo y el ancho de los sépalos y los pétalos.

3.3.2. Redes

Una *red* se utiliza para especificar relaciones entre dos o más elementos de datos. Un elemento en una red a menudo se denomina *nodo* o *vértice*, mientras que un *enlace* o *arista* representa una relación entre dos nodos. Las redes no solo representan la estructura, sino también aspectos contextuales como flujos o distancias entre los nodos. Por ejemplo, en una red social, cada persona podría ser un nodo y los enlaces entre ellos representarían sus amistades. En una red informática, los nodos podrían ser las computadoras y los enlaces, las conexiones entre éstas mediante cables físicos o conexiones inalámbricas. Tanto los nodos de la red como los enlaces pueden tener atributos asociados.

Un *árbol* es un tipo particular de red con una estructura jerárquica. A diferencia de una red general, los árboles carecen de ciclos: cada nodo hijo está vinculado a un único nodo padre. Un ejemplo de un árbol es el organigrama de una empresa, que ilustra en forma esquemática las áreas que la componen, las jerarquías y las relaciones entre ellas.



Id	LongSépalo	AnchoSépalo	LongPétalo	AnchoPétalo	Especies
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa

Figura 3.3: Ejemplo de tabla para el conjunto de datos *Iris* [Fis88]. Cada fila representa un elemento (muestra de flor), cada columna un atributo (propiedad de la muestra) y su intersección es la celda que contiene el valor para ese par.

En la figura 3.4(a) se presenta un ejemplo de árbol con 82 nodos, extraído del estudio [GTN14], donde los nodos varían en tamaño y color según su distancia a la raíz. La figura 3.4(b) utiliza un diseño basado en fuerzas para visualizar la red de co-ocurrencia de personajes en los capítulos de *Les Misérables*, la famosa novela de Víctor Hugo [HBO10]. En esta visualización, los colores de los nodos representan los clústeres identificados por un algoritmo de detección de comunidades.

3.3.3. Campos

El tipo de conjunto de datos conocido como *campo* se caracteriza por contener valores de atributos asociados con celdas en un dominio continuo. Cada celda en un campo representa medidas o cálculos, como temperatura, presión, velocidad, fuerza y densidad, que son ejemplos de fenómenos físicos que pueden ser descritos de manera continua. A diferencia de los datos organizados en tablas o redes, donde el número de elementos individuales es finito y bien definido, en los campos existen infinitos valores posibles entre

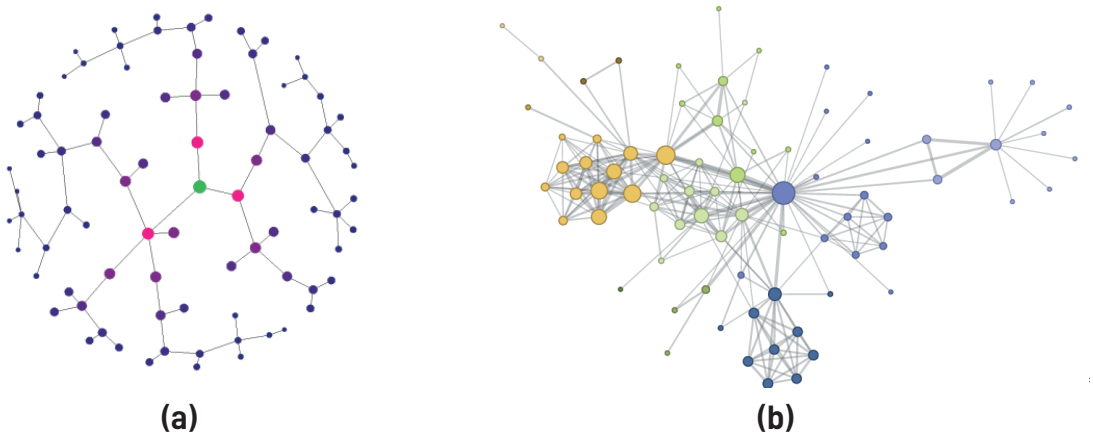


Figura 3.4: (a) Un árbol con 82 nodos. Figura extraída de [GTN14]. (b) Red de co-ocurrencia de personajes en los capítulos de la novela *Les Misérables* de Víctor Hugo. Figura extraída de [HBO10].

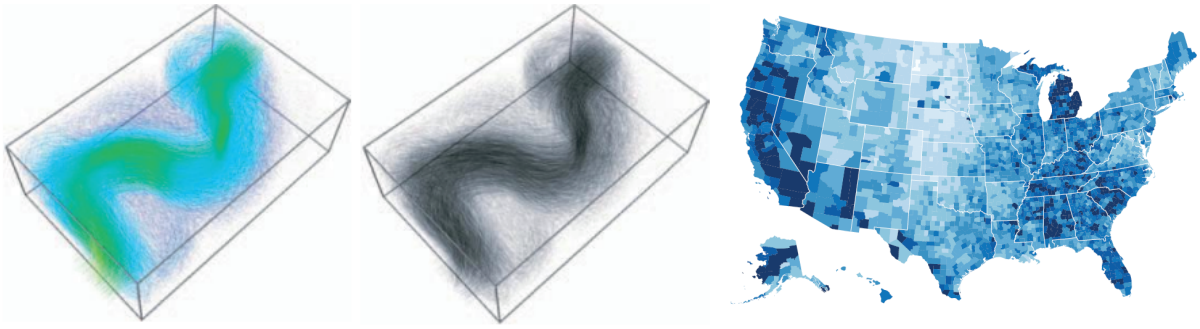
cualquier par de puntos medidos. Esto significa que siempre es posible obtener nuevas medidas entre dos valores existentes.

Trabajar con datos continuos implica abordar varias consideraciones técnicas, como la frecuencia de muestreo y la interpolación. La frecuencia de muestreo se refiere a la densidad con la que se toman las muestras, y una frecuencia insuficiente puede resultar en una representación inexacta del fenómeno continuo. Por otro lado, la interpolación es una técnica que se utiliza para estimar valores en puntos intermedios entre mediciones directas, permitiendo así reconstruir una representación más completa y precisa del campo.

En la figura 3.5(a) se muestra una visualización basada en glifos de un campo vectorial tridimensional que ilustra cómo se pueden representar de manera gráfica los valores continuos en un campo tridimensional.

3.3.4. Geometría

Los conjuntos de datos de tipo *geométrico* proporcionan información sobre la forma de los elementos con posiciones espaciales explícitas. Estos elementos pueden ser puntos (0D), líneas o curvas (1D), superficies (2D) o volúmenes (3D). Estos conjuntos de datos son intrínsecamente espaciales y a menudo incluyen una estructura jerárquica en múltiples escalas. A diferencia de los otros tipos de conjuntos de datos vistos, no tienen necesariamente atributos: transmiten información directamente a través de la posición espacial de sus elementos.



(a) (b)

Figura 3.5: (a) Visualización basada en glifos de un campo vectorial tridimensional. Figura extraída de [Tel14]. (b) Mapa de colores que representa regiones como marcas de área utilizando la geometría proporcionada. El color representa un atributo cuantitativo. Figura extraída de [Mun14].

La figura 3.5(b) muestra un ejemplo de las tasas de desempleo en los Estados Unidos de 2008 con un mapa de colores.

3.4. Atributos

Un *atributo* es una característica específica de un elemento o ítem. La figura 3.6 resume los tipos de atributos identificados por Munzner [Mun14]. Estos atributos se clasifican principalmente en *categóricos*, *ordenados* y *jerárquicos*. Dentro de los atributos ordenados, se distingue entre *ordinales* y *cuantitativos*. Además, los datos ordenados pueden presentar variaciones secuenciales desde un valor mínimo hasta un valor máximo o viceversa (*secuenciales*), pueden divergir en ambas direcciones desde un punto de origen (*divergentes*) o repetir sus valores siguiendo un patrón regular (*cíclicos*).

La tabla 3.1 permite examinar un ejemplo concreto de estos diversos tipos de datos. En ella se presenta un subconjunto de registros con información sobre los pasajeros del Titanic [Tit21]. La tabla incluye siete atributos: el número único de ticket (**Ticket**), el nombre del pasajero (**Name**), la edad (**Age**), el género (**Gender**), la clase de viaje (**Class**), el precio del billete (**Ticket**) y su estatus de supervivencia (**Survived**). De éstos, el número de ticket, el estatus de supervivencia, el género y el nombre del pasajero son atributos categóricos; la clase es un atributo ordinal, la edad es un valor cuantitativo discreto y el precio del billete es un valor cuantitativo continuo.

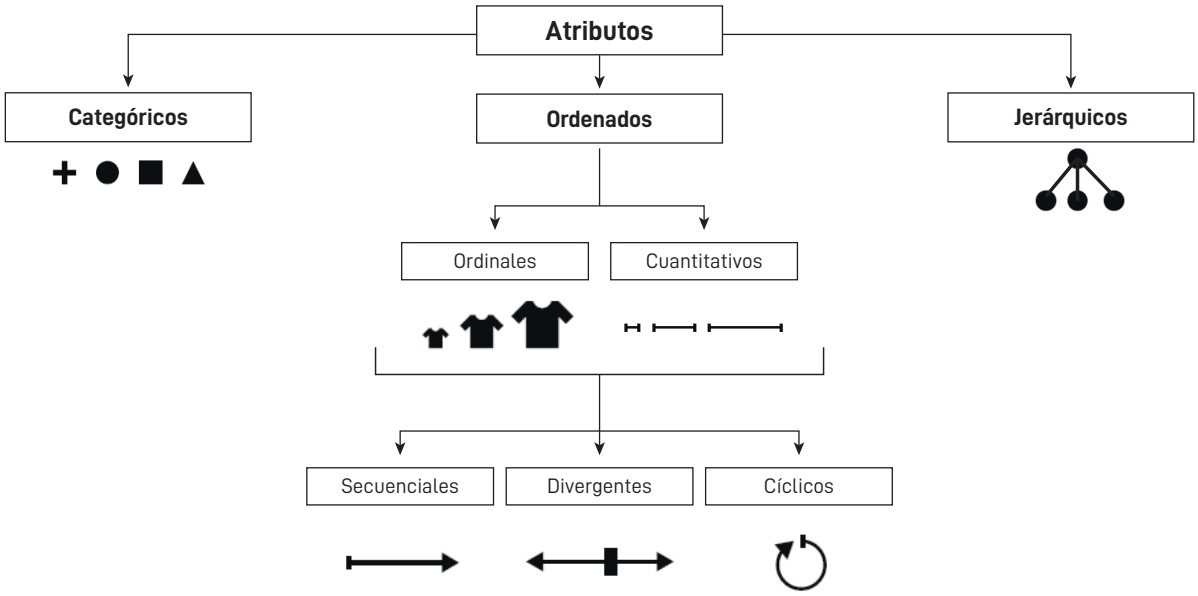


Figura 3.6: Clasificación de atributos según Munzner [Mun14].

3.4.1. Atributos Categóricos

Según Card *et al.* [CMS99], los datos *categóricos* o *nominales*, como los nombres de ciudades, géneros de películas o identificadores de productos, se definen como un conjunto no ordenado. Estos datos permiten realizar comparaciones de igualdad o diferencia entre sus elementos y pueden ser organizados según criterios externos arbitrarios, como ordenar alfabéticamente una lista de frutas por sus nombres. Sin embargo, es importante destacar que tales ordenamientos no reflejan una propiedad intrínseca del atributo, a diferencia de los datos ordenados. Es decir, para establecer cualquier tipo de orden en datos categóricos, es necesario recurrir a información adicional externa.

3.4.2. Atributos Ordenados: Ordinales y Cuantitativos

Los atributos *ordenados* se distinguen de los categóricos por poseer un orden implícito. Dentro de esta categoría, los datos *ordinales* disponen sus valores en una secuencia ordenada, aunque no permiten la realización de operaciones aritméticas sobre ellos. Por ejemplo, las tallas de una prenda (pequeña, mediana, grande) o los niveles de satisfacción de un cliente (insatisfecho, regular, satisfecho) son datos ordinales.

Otro tipo importante de datos ordenados son los datos *cuantitativos*, que permiten realizar operaciones aritméticas sobre sus valores numéricos. Ejemplos de datos cuantitativos incluyen la altura y el peso de una persona, la temperatura, el tiempo de viaje en

Ticket	Name	Age	Gender	Class	Fare	Survived
24160	Allen, Miss Elisabeth Walton	29	female	1st	211.69	Saved
113760	Carter, Mrs Lucile	36	female	1st	120.00	Saved
113760	Goldschmidt, Mr George B.	48	male	1st	76.14	Saved
113760	Harper, Mr Henry Sleeper	71	male	1st	34.13	Lost
113503	Keeping, Mr Edwin Herbert	33	male	1st	211.10	Lost
110469	Maguire, Mr John Edward	30	male	1st	26.00	Lost

Tabla 3.1: Seis registros con información sobre los pasajeros del Titanic. Fuente de datos: [Tit21]

minutos y la cantidad de agua consumida en litros. Tanto los números enteros (discretos) como los números reales (continuos) se consideran datos cuantitativos.

3.4.2.1. Secuenciales, Divergentes y Cíclicos

Los datos ordenados pueden organizarse en dos formas principales: *secuenciales* y *divergentes*. Una *secuencia* es un conjunto ordenado de elementos en el que cada término, excepto el primero y el último en el caso de ser finita, tiene un único predecesor y un único sucesor. Esta puede ser creciente, decreciente o seguir otro patrón específico. Por ejemplo, en una escala de puntuación en un examen, donde los valores van desde 0 hasta 100, los números siguen un orden secuencial en el que el puntaje aumenta monótonamente.

En contraste, en un conjunto *divergente*, los datos se organizan en dos secuencias de tal manera que sus valores se alejan en direcciones opuestas desde un punto de origen común. Un ejemplo sería la temperatura en grados Celsius (°C), donde los valores positivos indican temperaturas cálidas y los valores negativos indican temperaturas frías, con el punto de congelación (0°C) como punto común de unión entre ambas secuencias.

Los datos ordenados pueden ser *cíclicos*, lo que significa que los valores se repiten en un patrón regular y predecible a lo largo del tiempo. Estos datos suelen tener una naturaleza periódica, donde los valores se repiten en intervalos regulares. Por ejemplo, muchos tipos de medidas de tiempo son cíclicas, como la hora del día, los días de la semana o las estaciones del año.

3.4.3. Atributos Jerárquicos

Dentro de un conjunto de datos, es posible encontrar una estructura *jerárquica* tanto en un atributo específico como entre varios atributos diferentes. Los datos jerárquicos son aquellos organizados en una estructura de niveles o jerarquías, en la que los elementos se vinculan de tal manera que algunos son considerados superiores (padres) y otros subordinados (hijos). Esta estructura facilita la representación de relaciones de dependencia. Un ejemplo ilustrativo de esto es la categorización de productos en un sistema de inventario. Consideremos datos sobre productos vendidos en una cadena de tiendas. El atributo de categoría del producto puede estar organizado jerárquicamente, desde subcategorías específicas (como "remeras estampadas") hasta categorías más amplias (como "indumentaria"). Esta jerarquía permite identificar patrones de venta en distintos niveles, como variaciones entre tipos específicos de productos o tendencias generales en la categoría de ropa.

Además del tiempo, que se puede estructurar, por ejemplo, en semanas a partir de los días, y la geografía, que organiza los códigos postales de ciudades y países, hay otros atributos que también pueden presentar estructuras jerárquicas. Por ejemplo, el atributo relacionado con los empleados en una empresa puede organizarse jerárquicamente desde puestos específicos (como "desarrollador junior") hasta niveles superiores (como "gerente de departamento" o "director general"). Esta jerarquía puede revelar patrones en la asignación de recursos y responsabilidades a diferentes niveles de la organización.

3.5. Datos Multidimensionales

Un dato multidimensional (o n-dimensional) se caracteriza por tener múltiples atributos o variables [DKZ13]. Conceptualmente, podemos visualizar un conjunto de m datos multidimensionales (o n-dimensionales) como puntos en un espacio n-dimensional. En este contexto, cada muestra i-ésima se denota como:

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \quad \text{donde } i \in \{1, \ldots, m\}, \quad n > 3 \tag{3.1}$$

donde n representa el número de características o atributos, mientras que m indica el total de muestras o instancias en el conjunto de datos X. Este conjunto se puede expresar

como:

$$X = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = 1, ..., m, \quad j = 1, ..., n\} \tag{3.2}$$

Los elementos $X_1, X_2, ..., X_m$ representan puntos de datos, y debido a sus n características, se consideran datos n-dimensionales. Cada punto $X_i, i \in \{1, \ldots, m\}$ encapsula los valores característicos de un objeto específico. Cada valor x_{ij} corresponde al j-ésimo atributo en la i-ésima muestra.

Las tablas de datos se destacan entre otros formatos de conjuntos de datos básicos debido a su capacidad para mostrar de manera clara la cantidad de variables asociadas a cada conjunto de datos. Esta característica es crucial para seleccionar las visualizaciones más adecuadas para el análisis de datos.

3.6. Las Tareas

La identificación y comprensión de las tareas en el proceso de visualización son fundamentales para extraer información significativa de diversos conjuntos de datos.

En el ámbito de la visualización de datos, el término "tarea" a menudo se utiliza de manera ambigua, abarcando tanto interacciones como objetivos. Aquí, las tareas (o también llamadas "tareas de alto nivel") se refieren a los objetivos generales de los usuarios al diseñar una visualización, mientras que las interacciones o "tareas de bajo nivel" hacen alusión a las acciones y operaciones específicas que los usuarios llevan a cabo para lograr esos objetivos al interactuar con una representación visual de datos.

Las tareas de alto nivel se enfocan en metas generales y abstractas, como identificar patrones, comparar conjuntos de datos o inferir relaciones entre variables. En contraste, las tareas de bajo nivel se centran en acciones concretas en el proceso de visualización, tales como filtrar datos, ordenar elementos o seleccionar elementos específicos.

Elegir la visualización adecuada es esencial para la ejecución efectiva de estas tareas, ya que diferentes representaciones visuales son más apropiadas que otras para llevar a cabo ciertas tareas. Por ejemplo, un gráfico de barras puede ser efectivo para comparar cantidades, mientras que un gráfico de dispersión puede revelar relaciones entre variables. La comprensión de las tareas específicas que los usuarios desean abordar permite diseñar herramientas y sistemas de visualización más efectivos, facilitando la interpretación y el descubrimiento de información valiosa.

En el UVAM [GULC23], la tarea aparece de forma implícita y actúa como un hilo conductor que guía al usuario a lo largo de todo el proceso de visualización, desde la selección inicial de los DC hasta la interpretación de la representación final (ver figura 3.1). Ésta impulsa el proceso de visualización y afecta a las decisiones tomadas en cada etapa, incluidas las operaciones dentro de los estados y las transformaciones entre ellos. Además, la tarea orienta al usuario durante la interacción a lo largo del proceso de visualización, facilitando la exploración y permitiendo realizar ajustes iterativos en cualquier etapa del *pipeline*.

A principios de los años 90, Wehrend y Lewis [WL90] fueron pioneros en abordar explícitamente las tareas de alto nivel del proceso de análisis de datos, con el objetivo de facilitar la selección de representaciones visuales apropiadas. Caracterizaron tareas independientes del dominio, lo que permitió generalizar su clasificación. Propusieron una taxonomía que incluye diez tareas analíticas: *localización*, *identificación*, *distinción*, *categorización*, *agrupamiento*, *distribución*, *clasificación*, *comparación dentro de entidades*, *asociación* y *correlación*.

Shneiderman [Shn96] propuso una clasificación de los datos a visualizar en siete categorías principales: unidimensionales, bidimensionales, tridimensionales, temporales, multidimensionales, jerárquicos y en red. Asociada a esta clasificación, desarrolló una taxonomía de tareas de bajo nivel que incluye siete acciones fundamentales: *vista general*, *zoom*, *filtro*, *detalles a pedido*, *relacionar*, *historial* y *extraer*. Años más tarde, junto a Jeffrey Heer [HS12], presentan una taxonomía que consta de doce tareas de bajo nivel agrupadas en tres categorías de alto nivel, como se muestra en la tabla 3.2. Estas categorías incorporan las tareas críticas que permiten el análisis visual iterativo.

Posteriormente, Zhou y Feiner [ZF98] introdujeron otra categorización de tareas. Separaron los objetivos que tiene un usuario al utilizar una representación visual de las técnicas visuales de bajo nivel (la operación exacta realizada en un elemento presente en pantalla) mediante un nivel intermedio, las tareas visuales. Clasificaron las tareas visuales en dos categorías principales: aquellas destinadas a *informar* y las diseñadas para *permitir* al usuario realizar cálculos o exploraciones visuales. Las tareas visuales orientadas a informar pueden *resumir* o *elaborar* información. De manera similar, las tareas que permiten al usuario *explorar* o *realizar* cálculos pueden subdividirse según si permiten *buscar* un objeto específico, *verificar* un hecho o *diferenciar* o *sumar* diferentes valores.

Tarea	**Descripción**
Especificación de datos y vistas	- *Visualizar* datos eligiendo codificaciones visuales. - *Filtrar* los datos para centrarse en elementos relevantes. - *Ordenar* elementos para exponer patrones. - *Derivar* valores o modelos a partir de los datos de entrada.
Manipulación de vistas	- *Seleccionar* elementos para resaltar, filtrar o manipular. - *Navegar* para examinar patrones generales y detalles específicos. - *Coordinar* vistas para una exploración multidimensional vinculada. - *Organizar* múltiples ventanas y espacios de trabajo.
Proceso de análisis y procedencia	- *Registrar* historiales de análisis para revisarlos y compartirlos. - *Anotar* patrones para documentar los hallazgos. - *Compartir* vistas y anotaciones para permitir la colaboración. - *Guiar* a los usuarios a través de tareas de análisis o historias.

Tabla 3.2: Taxonomía de tareas propuesta por Heer y Shneiderman [HS12].

Además, Zhou y Feiner definieron un conjunto de tareas visuales de bajo nivel, agrupadas en las tareas de medio y alto nivel mencionadas anteriormente. La tabla 3.3 resume la taxonomía de tareas visuales propuesta por estos autores.

En un trabajo más reciente, Amar *et al.* [AES05] presentan un conjunto de diez tareas de bajo nivel resumidas en la tabla 3.4, que describen las actividades de los usuarios mientras utilizan herramientas de visualización para comprender sus datos.

Valiati *et al.* [VPF06] propusieron una taxonomía de tareas específicas de usuario, basada en taxonomías existentes. Esta taxonomía integra, en diferentes niveles, tareas *analíticas*, *cognitivas* y *operativas* que un usuario podría necesitar realizar al utilizar una técnica de visualización. Cinco de estas tareas pueden considerarse como objetivos que un usuario podría tener al utilizar una técnica de visualización para explorar visualmente o analizar el conjunto de datos a través de estadísticas; éstas incluyen identificar, determinar, comparar, inferir y ubicar. Las tareas restantes, visualizar y configurar, son de nivel intermedio y respaldan las tareas analíticas. Dado que estas últimas son de alto nivel, su realización podría implicar la ejecución de otras tareas.

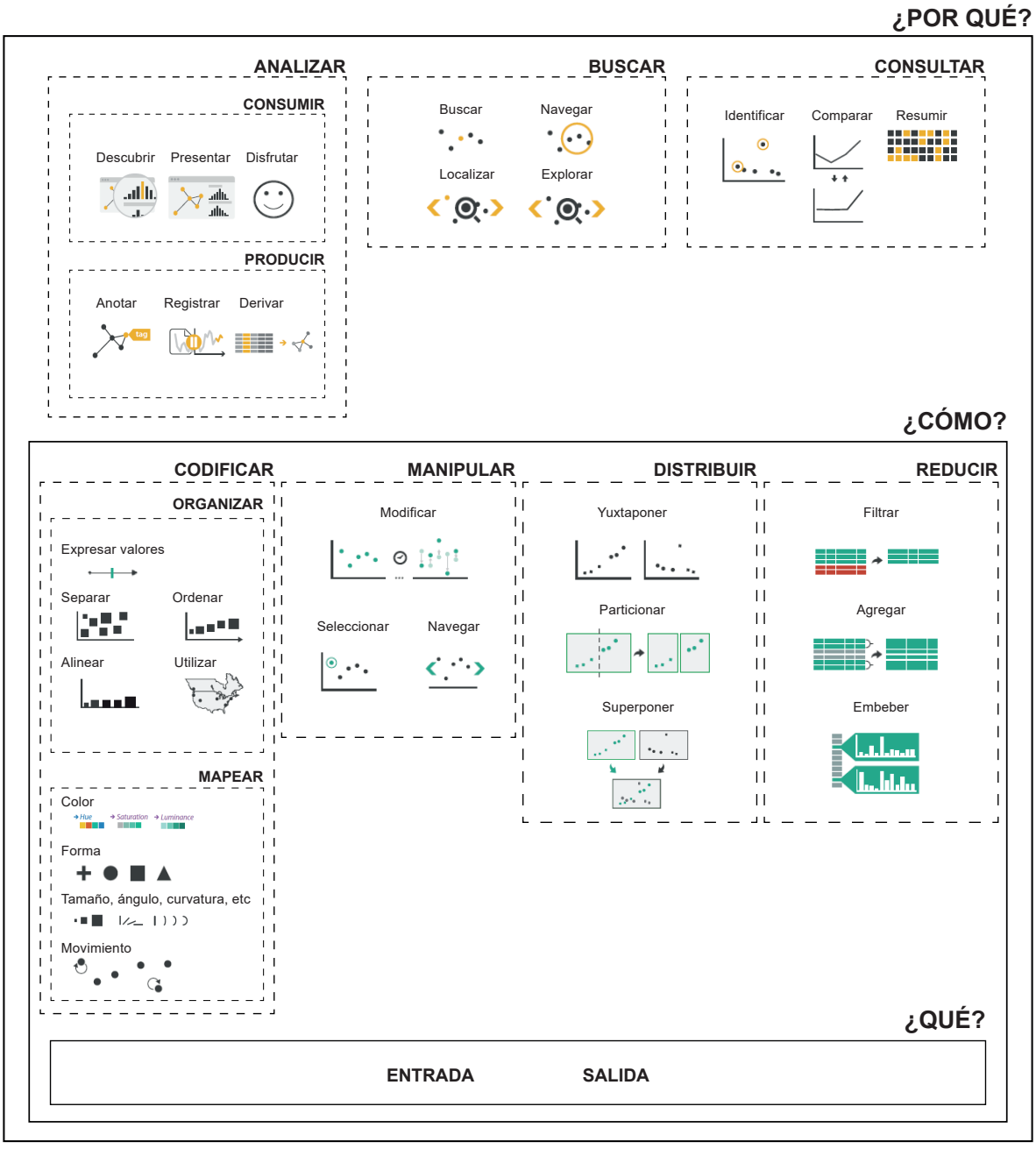


Figura 3.7: Clasificación multinivel de tareas propuesta por Brehmer y Munzner [BM13] que abarca el *por qué* se realiza una tarea, *cómo* se lleva a cabo y *qué* parámetros de entrada y salida tiene.

Alto nivel	**Nivel intermedio**	**Bajo nivel**
Informar	Elaborar	Enfatizar y revelar.
	Resumir	Asociar, background, categorizar, agrupar, comparar, correlacionar, distinguir, generalizar, identificar, ubicar y clasificar.
Permitir	Explorar	*Buscar:* categorizar, agrupar, comparar, correlacionar, distinguir, enfatizar, identificar ubicar, clasificar y revelar. *Verificar:* categorizar, comparar, correlacionar, distinguir, identificar, ubicar, clasificar y revelar.
	Computar	*Sumar:* correlacionar, ubicar y clasificar. *Diferenciar:* correlacionar, ubicar y clasificar.

Tabla 3.3: Taxonomía de tareas visuales propuesta por Zhou y Feiner [ZF98].

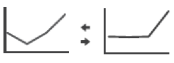
Más recientemente, Brehmer y Munzner [BM13] abordan la brecha entre los sistemas de clasificación de tareas de visualización de bajo y alto nivel. Proponen una taxonomía innovadora que descompone tareas complejas en secuencias de tareas simples e independientes, utilizando el marco *"¿Qué?-¿Por qué?-¿Cómo?"*.

En el aspecto *¿Por qué?* se describen los propósitos de la visualización en diversos niveles, tanto el análisis de información como la producción y consumo de nueva información. En el *¿Cómo?* se presentan una variedad de funciones de visualización y técnicas relacionadas con la codificación, manipulación, distribución y reducción de la información. Las técnicas de codificación determinan cómo se organizan espacialmente los elementos visuales o qué tipo de elementos se utilizan. Las técnicas de manipulación abarcan la modificación de posiciones o formas de los elementos visuales, la selección de elementos de interés y la navegación en la visualización. Las técnicas de distribución permiten comparar diferencias o similitudes entre vistas, mientras que las técnicas de reducción incluyen filtrar información innecesaria, agregar elementos visuales para ajustar la granularidad de la información o añadir explicaciones adicionales de los datos. Por su parte, en el *¿Qué?* de la clasificación se modelan los parámetros de entrada y salida de las tareas.

Tarea	Descripción
Recuperar valor	Encontrar atributos de casos específicos en un conjunto de datos.
Filtrar	Reducir un conjunto de datos a un subconjunto basado en ciertos criterios.
Calcular valor derivado	Crear un nuevo atributo basado en los valores de otros atributos en un conjunto de datos.
Encontrar extremo	Encontrar el valor mínimo o máximo de un atributo en un conjunto de datos.
Ordenar	Organizar un conjunto de datos en un orden específico basado en los valores de un atributo.
Determinar rango	Entender el rango de valores en un conjunto de datos.
Caracterizar distribución	Entender la distribución de los datos en un conjunto de datos.
Encontrar anomalías	Encontrar elementos inusuales o excepcionales en un conjunto de datos.
Agrupar	Organizar elementos en grupos basados en sus similitudes y diferencias.
Correlacionar	Encontrar relaciones entre dos o más variables en un conjunto de datos.

Tabla 3.4: Taxonomía de tareas propuesta por Amar *et al.* [AES05].

La figura 3.7 resume brevemente la clasificación multinivel propuesta por Brehmer y Munzner [BM13], y la tabla 3.5 ofrece una descripción detallada de los propósitos de la visualización en los distintos niveles.

¿Por qué?	Análisis	Consumir	Descubrir 	Encontrar nuevos conocimientos o verificar/refutar una hipótesis existente.
			Presentar 	Comunicar algo específico y ya comprendido a una audiencia.
			Disfrutar 	El usuario no está impulsado por una necesidad de verificar o generar una hipótesis, sino por una curiosidad que podría ser estimulada y satisfecha por la visualización.
		Producir	Anotar 	Adicionar anotaciones gráficas o textuales asociadas con uno o más elementos de visualización preexistentes, generalmente como una acción manual del usuario.
			Registrar 	Guardar o capturar elementos de visualización mediante capturas de pantalla, marcadores de elementos o ubicaciones, configuraciones de parámetros y registros de interacción.
			Derivar 	Producir nuevos elementos de datos basados en elementos de datos existentes.
	Búsqueda		Buscar 	Los usuarios saben lo que están buscando y dónde se encuentra.
			Localizar 	Los usuarios saben lo que están buscando pero desconocen su ubicación.
			Navegar 	Los usuarios no saben exactamente lo que están buscando, pero tienen en mente una ubicación donde buscarlo.
			Explorar 	Los usuarios no saben lo que están buscando ni dónde encontrarlo.
	Consultar		Identificar 	Devuelve un único objeto (sus características o referencias específicas).
			Comparar 	Devuelve múltiples objetivos.

Sigue en la página siguiente.

...continuación de la página anterior.

		Resumir 	Devuelve todos los objetos posibles. Proporciona una visión integral de todo.

Tabla 3.5: Clasificación de tareas de visualización *("¿Por qué?")* en tres niveles de Brehmer y Munzner [BM13].

Además, varios autores se han dedicado a estudiar y categorizar las tareas específicas según el tipo de datos a visualizar (por ejemplo, datos temporales, datos espaciales, datos relacionales, etc.), la estructura subyacente del conjunto de datos (por ejemplo, redes, árboles, tablas, etc.) y las técnicas de visualización utilizadas para representar los datos. Por ejemplo, Chan [SMS+21] clasifica los datos en temporales y no temporales, y agrega la tarea de bajo nivel *timestep* a las siete tareas clásicas propuestas por Ben Shneiderman [Shn96]. *Timestep* refiere a la acción de controlar las secuencias de tiempo y permitir a los usuarios inspeccionar los datos en un momento específico. Asimismo, Lee *et al.* [LPP+06] se han enfocado en las tareas visualización de datos en grafos, mientras que Sarikaya y Gleicher [SG18] lo hacen en diagramas de dispersión.

3.7. Interacciones en Visualización

Mediante representaciones visuales, los usuarios pueden explorar iterativa e interactivamente el espacio de sus datos para analizarlos y comprenderlos. En este contexto, las técnicas de interacción juegan un papel crucial ya que brindan al usuario la posibilidad de manipular las representaciones visuales para poder interpretarlas y extraer información.

Hasta el momento, existen múltiples y diversos trabajos que proponen taxonomías relacionadas con interacciones en visualización. Algunos categorizan las técnicas de interacción [CR96, Kei02, Wil12], mientras que otros describen las técnicas de interacción en función de sus dimensiones [Twe97, Spe07] o presentan las taxonomías relevantes ampliadas al área de *Visual Analytics*[1] [TC05, KS12]. Dado que las interacciones en visualización no constituyen el objeto de estudio específico de esta tesis, pero son sumamente relevantes en el contexto de la visualización, he decidido llevar a cabo una breve

[1] *Visual Analytics*, o Analítica Visual, se refiere a la combinación de técnicas de análisis automatizado y visualizaciones interactivas para una comprensión, razonamiento y toma de decisiones efectiva sobre conjuntos de datos grandes y complejos [KAF+08].

introducción sobre las técnicas de interacción presentes en la literatura.

Ganuza [Gan18] ofrece un relevamiento detallado y exhuastivo de los trabajos científicos presentes en la literatura relacionados con taxonomías y clasificaciones de las interacciones en el contexto de visualización y *Visual Analytics*. Además, propone una nueva clasificación multi-nivel de interacciones en el contexto del Modelo Unificado de Visualización [MCFE03], con el objetivo de proporcionar un marco conceptual sólido para el diseño y análisis visual de datos.

De acuerdo a la clasificación propuesta por Ganuza [Gan18], podemos dividir las interacciones en dos grandes grupos:

- **Interacciones a Nivel del Programador:** interacciones diseñadas para el programador del sistema de visualización, quién conoce el proceso de visualización. Estas pueden ser *Interacciones de Bajo Nivel del Programador* las cuales manipulan directamente el conjunto de datos, o *Interacciones de Alto Nivel del Programador* que manipulan los estados de los datos y las transformaciones del proceso de visualización.

- **Interacciones a Nivel del Usuario:** interacciones diseñadas para el usuario del sistema de visualización, quien no necesariamente conoce el proceso de visualización. El usuario interactuará con la vista final, sin conocer necesariamente la representación de los conjuntos de datos, los estados que atraviesan y/o las transformaciones a las que se someten.

  En función del conjunto de datos que afectan, se pueden distinguir tres clases: *Interacciones a Nivel del Usuario Sobre el Conjunto de Datos*, *Interacciones a Nivel del Usuario Sobre el Mapeo Visual* e *Interacciones a Nivel del Usuario Sobre la Vista*. Una cuarta clase denominada *Interacciones Compuestas a Nivel del Usuario* incluye las interacciones que se componen de otras interacciones a nivel del usuario.

3.7.1. Interacciones a Nivel del Programador

Estas interacciones están especialmente diseñadas para el usuario desarrollador, quien posee un conocimiento completo del *pipeline* de visualización y será responsable de programar el sistema de visualización que utilizará el usuario final.

Las interacciones a nivel del programador se clasifican en dos niveles:

- **Interacciones de Bajo Nivel del Programador:** estas interacciones se definen en base a un conjunto de operaciones definidas para las representaciones de los conjuntos de datos y para el mapeo visual. Las interacciones aplicadas a los conjuntos de datos incluyen operaciones básicas como crear, eliminar, insertar y modificar datos, atributos y valores; así como operaciones para agrupar y desagrupar datos, y consultas para obtener información sobre los *datasets*, incluyendo datos, atributos y clústeres. Por otro lado, las interacciones aplicadas al mapeo visual comprenden operaciones para crear, eliminar y configurar el mapeo visual, incluyendo la orientación y organización de los ejes; funciones para mapear datos a marcas gráficas y asociar atributos de datos a canales visuales y operaciones para crear, configurar y manipular las características de las marcas gráficas, como color, tamaño, orientación y posición.

- **Interacciones de Alto Nivel del Programador:** estas interacciones se definen en base a las interacciones de bajo nivel del programador y manipulan directamente los estados y transformaciones de los datos en el *pipeline* de visualización. Abarcan cargar fuentes de datos, así como generar, eliminar y modificar las diferentes ramas de datos abstractos, datos a visualizar, datos mapeados visualmente y datos visualizados a lo largo del *pipeline*. Además, incluyen operaciones para derivar nuevos datos y atributos, filtrar, agrupar y desagrupar conjuntos de datos. También permiten configurar el sustrato espacial y sus ejes, asociar datos y atributos a marcas gráficas y canales visuales, así como seleccionar técnicas de visualización. En la vista final, posibilitan recuperar datos y objetos gráficos, transformarlos mediante rotaciones, traslaciones y escalados, además de aplicar distorsiones para facilitar la exploración visual.

3.7.2. Interacciones a Nivel del Usuario

Estas interacciones están diseñadas para el usuario del sistema de visualización, que puede no estar familiarizado con los detalles del proceso de visualización. Se abstrae al usuario final de la complejidad inherente al procesamiento de datos y generación de representaciones visuales para que pueda enfocarse en explorar y analizar los datos a través de un conjunto de interacciones de alto nivel, sin tener conocimiento de la representación in-

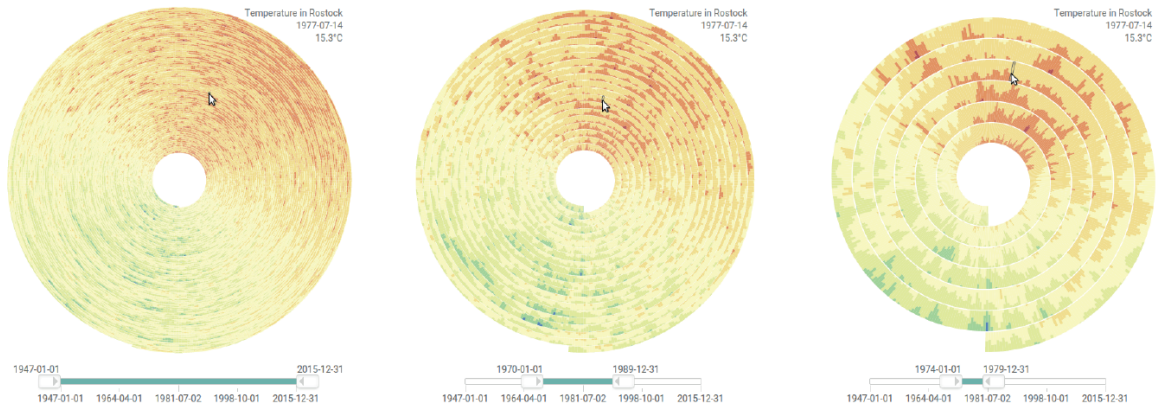


Figura 3.8: Ejemplo de interacción de filtrado a nivel del usuario sobre el conjunto de datos. Se utiliza una barra de desplazamiento (*slider*) para ajustar el período de tiempo asignado a una representación espiral. Figura extraída de [TS20].

terna de los diferentes conjuntos de datos ni de los distintos estados y/o transformaciones que atraviesan.

A diferencia del programador, el usuario final interactuará con el sistema a través de la vista generada. Sin embargo, estas interacciones pueden resolverse en etapas anteriores del *pipeline* de visualización. En función de los conjuntos de datos que afectan, se diferencian cuatro clases de interacciones a nivel del usuario:

- **Interacciones a Nivel del Usuario Sobre el Conjunto de Datos:** estas interacciones se resuelven en las primeras etapas del *pipeline* y permiten al usuario manipular directamente los conjuntos de datos, sin afectar el mapeo visual. Incluyen operaciones como cargar o eliminar conjuntos de datos, agrupar o desagrupar datos, ocultar o mostrar datos y atributos, así como generar nuevas configuraciones de visualización. La figura 3.8 ilustra un ejemplo de estas interacciones.

- **Interacciones a Nivel del Usuario Sobre el Mapeo Visual:** estas interacciones permiten al usuario manipular y configurar el mapeo visual de los datos. Incluyen operaciones como configurar los ejes de la visualización, asociar datos a marcas visuales, asociar atributos a canales visuales de las marcas, configurar los canales visuales, resaltar datos y generar nuevas vistas con diferente mapeo visual.

- **Interacciones a Nivel del Usuario Sobre la Vista:** estas interacciones manipulan directamente la vista generada, sin afectar los conjuntos de datos ni el mapeo

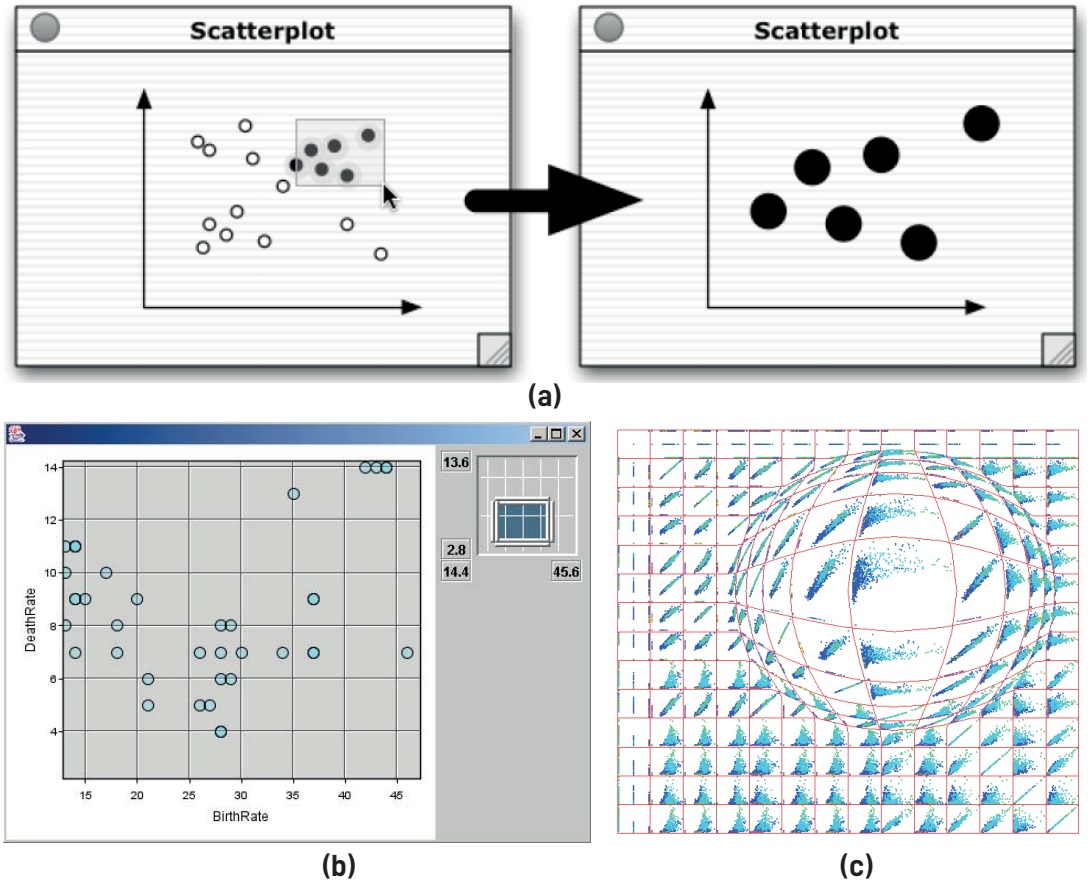


Figura 3.9: Ejemplos de interacciones a nivel del usuario sobre la vista. (a) *Zooming*: se selecciona una región rectangular para ampliar. Figura extraída de [Sii07]. (b) *Panning*: se mueve la cámara de forma horizontal y/o vertical para obtener una vista panorámica de la visualización. Esta técnica puede combinarse con *Zooming* para cambiar el enfoque y navegar al mismo tiempo por la visualización. Figura extraída de [Wil12]. (c) *Distorsión*: en esta matriz de diagramas de dispersión (SPLOM), se selecciona un centro de enfoque y se amplía utilizando una técnica de lente de distorsión. Figura extraída de [WY04].

visual subyacente. Incluyen operaciones como seleccionar la técnica de visualización aplicada, seleccionar datos directamente en la vista, obtener nuevas vistas aplicando diferentes técnicas, así como escalar (*Zooming*), trasladar (*Panning*), rotar o distorsionar la escena visualizada. La figura 3.9 ilustra ejemplos de estas interacciones.

- **Interacciones Compuestas a Nivel del Usuario:** son interacciones más complejas compuestas por combinaciones de las interacciones básicas. Incluyen técnicas como el *Zoom Semántico*, que permite revelar más información bajo demanda del usuario mediante la selección y filtrado de datos; *Foco+Contexto*, que resalta áreas de interés aplicando una distorsión en la vista; *Overview+Detalle*, que proporciona

una vista general y una detallada de un subconjunto seleccionado; *Brushing*, que permite seleccionar y resaltar datos interactivamente y *Brushing & Linking*, que coordina selecciones entre múltiples vistas. Además, incluyen la *Agregación* y *Desagregación*, que permiten agrupar y representar datos con una única marca visual o deshacer esa agrupación para representarlos individualmente. Estas interacciones suelen involucrar operaciones coordinadas de selección, filtrado, codificación visual y agregación/desagregación en los conjuntos de datos, mapeos visuales y vistas. En las figuras 3.10 y 3.11 se ilustran ejemplos de estas interacciones.

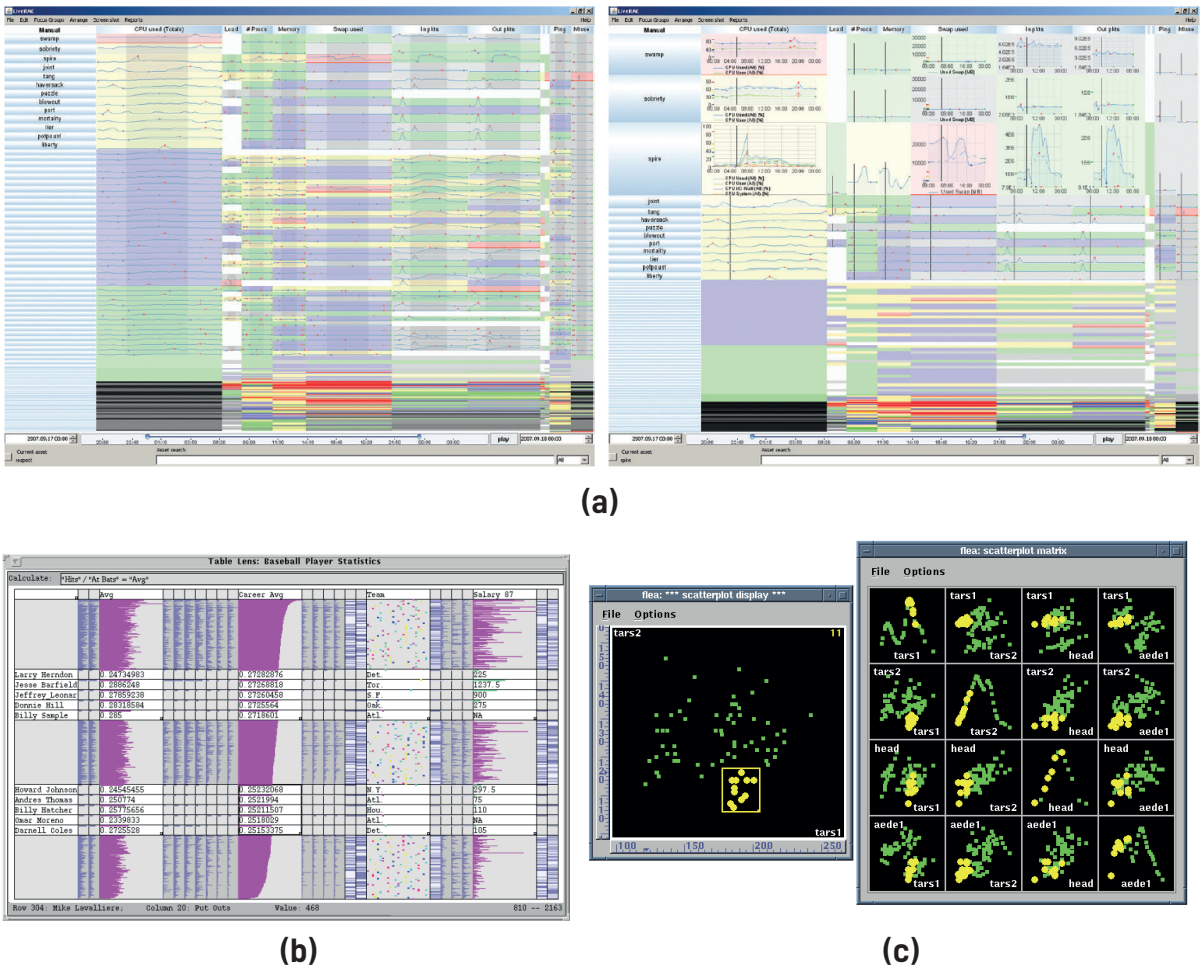


Figura 3.10: Ejemplos de interacciones compuestas a nivel del usuario. (a) *Zoom Semántico*: *LiveRAC* es un sistema que facilita la exploración de datos de series temporales de gestión de sistemas. En el gráfico de la izquierda, se ampliaron las primeras decenas de filas para mostrar gráficos de líneas simplificadas de los dispositivos. En el gráfico de la derecha, las tres filas superiores se han ampliado aún más, mostrando los gráficos con todos sus detalles. Figura extraída de [Mun14]. (b) *Foco+Contexto*: *Table Lens* [RC94] utiliza un diseño de foco+contexto para facilitar el análisis detallado de subconjuntos de datos mientras se mantiene una vista de contexto del conjunto de datos completo. Las filas de interés se pueden enfocar en una vista detallada, mientras que las filas restantes se muestran como contexto miniaturizado. Figura extraída de [RC94]. (c) *Brushing & Linking*: al seleccionar un conjunto de datos en el diagrama de dispersión, éstos se resaltan en todas las demás vistas que componen la matriz del diagrama de dispersión. Figura extraída de [Wil12].

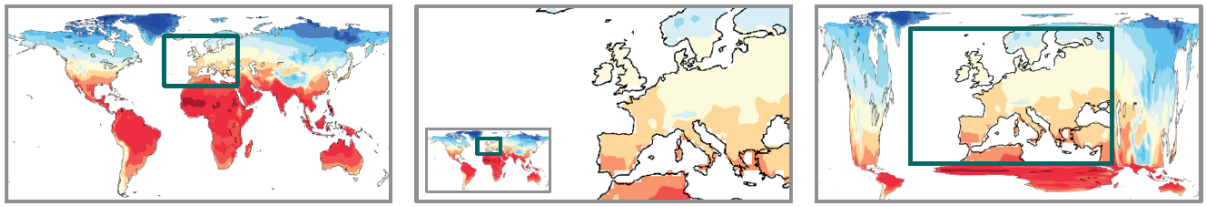


Figura 3.11: Ejemplos de interacciones compuestas a nivel del usuario. (a) El usuario marca una región de interés en la vista general del mapa. (b) *Overview+Detalle*: se proporciona una vista general de todos los datos junto con un detalle de la región de interés. Implica dos visualizaciones de los datos: una general que muestra la vista general y otra vista que proporciona el detalle. (c) *Foco+Contexto*: la zona de interés se destaca en una única vista mediante una distorsión aplicada sobre la misma. Se utiliza un enfoque de lente de distorsión sobre la zona de interés, lo que permite visualizarla con mayor detalle sin perder de vista el contexto general. Figura extraída de [TS20].

3.8. Conclusiones

En este capítulo se han explorado dos pilares fundamentales de la visualización: los datos y las tareas. Se presentó un análisis detallado de la clasificación propuesta por Munzner [Mun14], que cubre desde elementos básicos como ítems, atributos, enlaces, posiciones y grillas, hasta conjuntos de datos más complejos, tales como tablas, redes, campos y geometría. Además, se realizó un análisis exhaustivo de las taxonomías de tareas en la literatura, investigando sus distintos niveles, desde las tareas de alto nivel que definen objetivos generales hasta las interacciones específicas de bajo nivel, clasificándolas en dos grandes grupos: interacciones a nivel del programador y a nivel del usuario. Este estudio sirvió como base para el desarrollo de una taxonomía unificada de tareas, que se presenta en el capítulo 6.

La comprensión de estos pilares establece una base sólida para el diseño de visualizaciones efectivas, facilitando la selección de representaciones adecuadas según la naturaleza de los datos y los objetivos específicos del usuario.

Capítulo 4

Fundamentos de la Representación Visual

4.1. Introducción

En el proceso de visualización de datos, una vez que hemos identificado qué datos queremos visualizar y por qué lo hacemos, es crucial definir cómo vamos a representarlos visualmente. Este tercer pilar, el "*¿Cómo?*", abarca tanto el mapeo visual que implica la transformación de los datos en elementos gráficos como la selección de técnicas de visualización adecuadas para comunicar la información de manera efectiva.

La representación visual requiere una cuidadosa selección de marcas gráficas, que son los elementos básicos utilizados para representar datos, como puntos, líneas o superficies. Además, es fundamental elegir los canales visuales más adecuados—como la posición, el color, el tamaño o la orientación—que son las propiedades de las marcas gráficas que permiten codificar diferentes dimensiones o atributos de los datos. Este proceso debe considerar tanto las características intrínsecas de los datos como las preferencias, necesidades y objetivos del usuario.

Otro aspecto esencial de este pilar es la selección de las técnicas de visualización, que determina cómo se mostrarán los datos. En esta etapa, se seleccionan las técnicas más adecuadas para destacar la información relevante, permitiendo al usuario visibilizar, si existieran, patrones, tendencias y relaciones en los datos. La técnica elegida debe no solo resaltar los aspectos clave de los datos, sino también apoyar la tarea del usuario.

Es fundamental que este mapeo mantenga la integridad de los datos, asegurando que

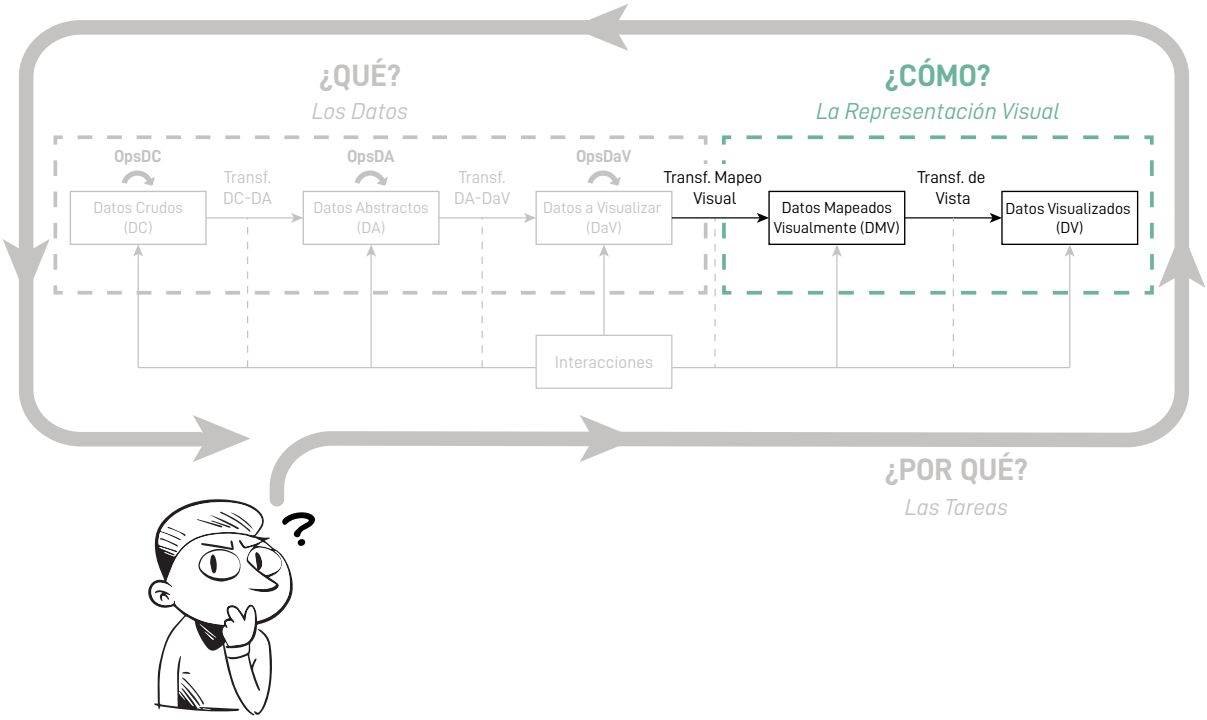


Figura 4.1: La representación visual en el Modelo Unificado de Análisis Visual (UVAM) [GULC23].

solo la información relevante esté representada visualmente, evitando la introducción de información no deseada que podría llevar a interpretaciones incorrectas.

En el contexto del Modelo Unificado de Análisis Visual (UVAM) [GULC23], la representación visual se concentra en los dos estados finales del proceso (ver figura 4.1). Los *Datos Mapeados Visualmente (DMV)*, son los *Datos a Visualizar (DaV)* que han sido enriquecidos con la estructura visual. La estructura visual constituye la información de soporte necesaria para la representación visual de los datos. Los datos en este estado pueden visualizarse aplicando una técnica que los soporte. Finalmente, en el estado de *Datos Visualizados (DV)*, una vez seleccionada la técnica de visualización adecuada, se crea la vista correspondiente. Para un conjunto específico de *DMV*, pueden existir varias técnicas disponibles, por lo que el usuario deberá seleccionar una de ellas para obtener los *DV*. Esta segunda parte del proceso está específicamente orientada a *"¿Cómo?"* vamos a representar visualmente los datos.

4.2. Estructura Visual

Todo proceso de visualización puede pensarse como una transformación de los datos en una representación visual. Como se mencionó previamente, la estructura visual proporciona la información de soporte necesaria para representar visualmente los datos. Especifica cómo el usuario desea visualizar los datos y proporciona todos los elementos necesarios para mostrarlos de acuerdo con las preferencias seleccionadas.

Las estructuras visuales están formadas por un sustrato espacial, marcas y las propiedades gráficas de esas marcas. La ecuación 4.1 y figura 4.2 describen la composición de la estructura visual. Los componentes individuales se detallan más adelante.

$$\begin{aligned} \text{Estructura Visual} &= \text{Sustrato Espacial} + \text{Sustrato Gráfico} \\ \text{Sustrato Gráfico} &= \text{Marcas} + \text{Canales Visuales} \end{aligned} \tag{4.1}$$

4.2.1. Sustrato Espacial

Enge *et al.* [ERI+21] definen el sustrato espacial como el contenedor donde se posicionan las marcas. El sustrato espacial refleja la organización del espacio que utilizará la técnica de visualización y contiene información sobre las características geométricas y topológicas de la representación a generar.

Card *et al.* [CMS99] describen el sustrato espacial en términos de ejes y sus propiedades, y definen cuatro tipos elementales de ejes:

- Ejes no estructurados: en esta representación no hay ejes. En la figura 4.3(a) se muestra una red con codificación de tamaño para los atributos de los nodos [Mun14].
- Ejes nominales o categóricos: se divide el espacio en regiones que representan diferentes categorías. La figura 4.3(b) muestra un diagrama de conjuntos paralelos [BKH05], técnica con ejes categóricos, donde se visualizan los puentes en Pittsburgh de acuerdo a diferentes atributos [Wil19].
- Ejes ordinales: se divide el espacio en regiones que representan diferentes categorías las cuales tienen algún orden dentro del espacio. En la figura 4.3(c) se presenta una visualización de una historia clínica basada en un caso real utilizando *LifeLines* [PMS+03], donde todos los ejes comparten un ordenamiento temporal.

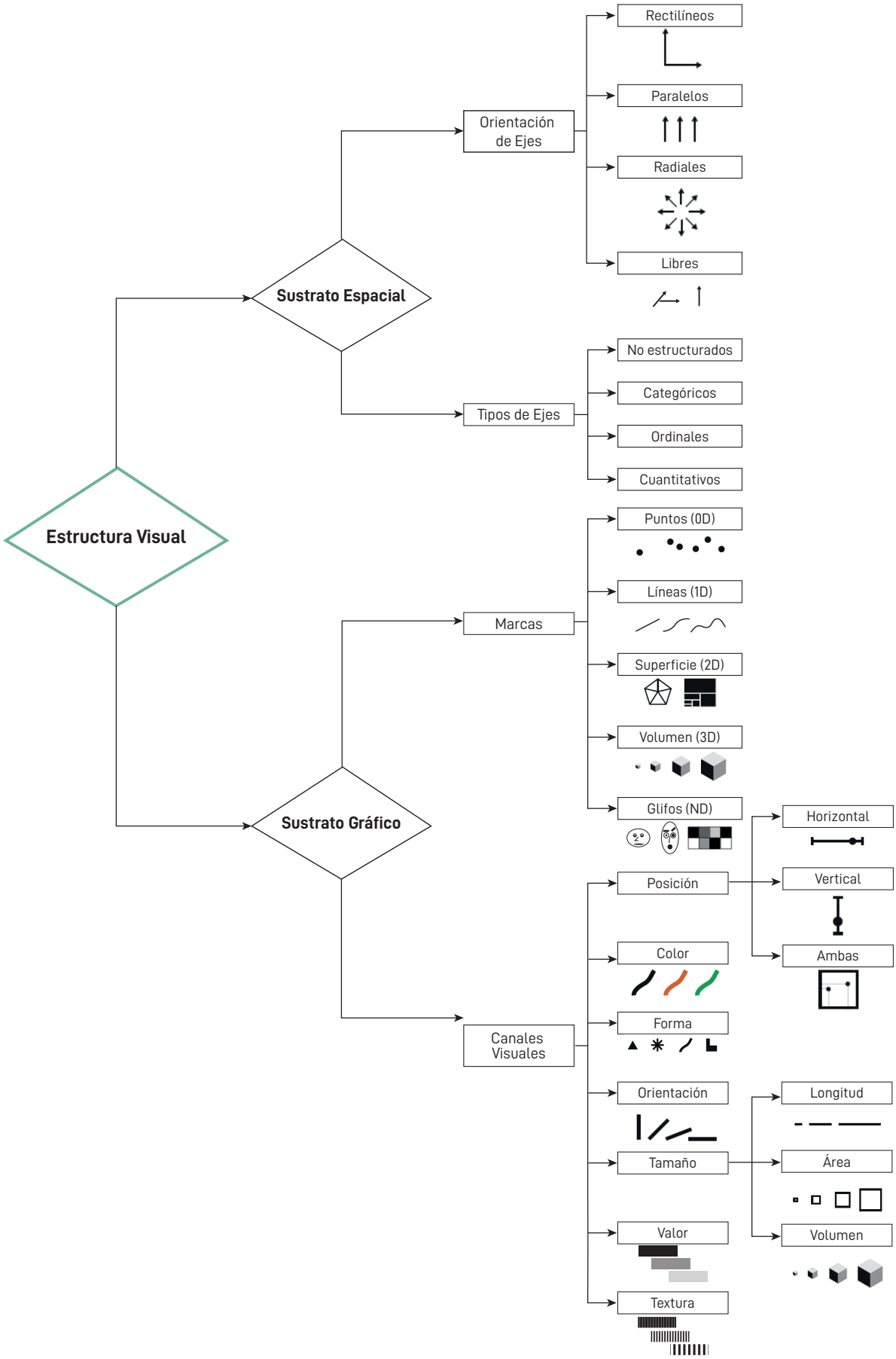


Figura 4.2: Estructura visual basada en el trabajo de Munzner [Mun14]. Se han incorporado los canales visuales de valor y textura propuestos por Bertin [Ber83], y la orientación de ejes libre introducido por Ganuza [Gan18].

- Ejes cuantitativos: representan regiones que poseen una métrica y suelen estar asociadas a atributos de este tipo. La figura 4.3(d) ilustra un ejemplo de visualización con ejes cuantitativos, en el cual se distribuyen los pasajeros del Titanic según su edad y sexo utilizando dos histogramas [Wil19].

Otra característica clave que se debe definir es la orientación de los ejes. Munzner propone tres orientaciones posibles: *rectilínea*, *paralela* y *radial* [Mun14]. Ganuza añade una cuarta orientación, que denomina "*libre*", para incluir todas las demás configuraciones no consideradas en esas tres alternativas [Gan18].

- Orientación de ejes rectilínea: las regiones o elementos se distribuyen a lo largo de ejes perpendiculares. En la figura 4.4(a) se utiliza un diseño rectilíneo para mostrar datos meteorológicos utilizando una matriz de diagramas de dispersión [TS20].

- Orientación de ejes radial: en una organización de ejes radial los ejes se extienden desde un punto central hacia el exterior en distintas direcciones, como radios de un círculo. El sistema de coordenadas natural para este tipo de organización es el sistema de coordenadas polar, donde una dimensión se mide como un ángulo desde una línea de inicio y la otra se mide como una distancia desde el punto central. Los glifos de estrella, presentados en la figura 4.4(b), utilizan este tipo de orientación para sus ejes [LRB03].

- Orientación de ejes paralela: en este tipo de organización, los ejes se ubican paralelos entre sí. La figura 4.4(c) presenta una visualización de coordenadas paralelas [TS20].

- Orientación de ejes libre: la orientación de los ejes no se restringe a ninguna de las tres alternativas descritas previamente. En este caso, cada uno de los ejes pueden tener cualquier orientación y estar distribuidos libremente en el espacio. En la figura 4.4(d) se ilustra un ejemplo. La técnica *TimeWheel* considera un eje de tiempo central, posicionado horizontalmente, alrededor del cual se sitúan los demás ejes en una disposición circular. Los datos se representan con líneas que unen un valor temporal dado con cada uno de los valores de los otros atributos en los respectivos ejes [TS20].

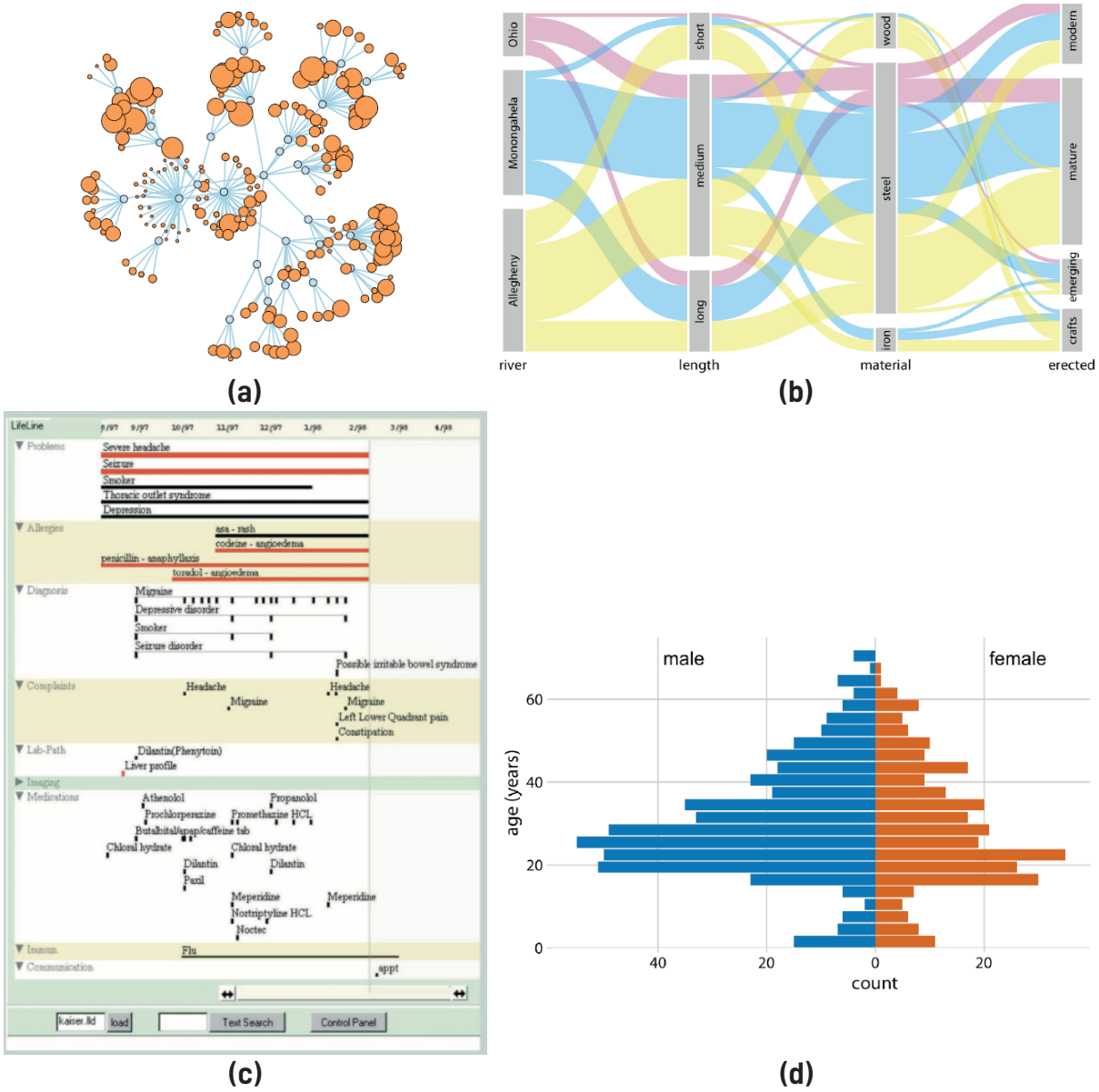


Figura 4.3: Ejemplos de los distintos tipos de ejes. (a) *No estructurados*: red con codificación de tamaño para atributos de nodo. Figura extraída de [Mun14]. (b) *Nominales*: visualización de puentes en Pittsburgh separados por río, época de construcción, longitud y material de construcción utilizando la técnica de conjuntos paralelos. Figura extraída de [Wil19]. (c) *Ordinales*: historia clínica basada en un caso real que incluye diferentes aspectos como problemas, alergias, diagnósticos, etc. utilizando *LifeLines*. Figura extraída de [BWP+19]. (d) *Cuantitativos*: distribución de los pasajeros del Titanic, según su edad y sexo utilizando dos histogramas. Figura extraída de [Wil19].

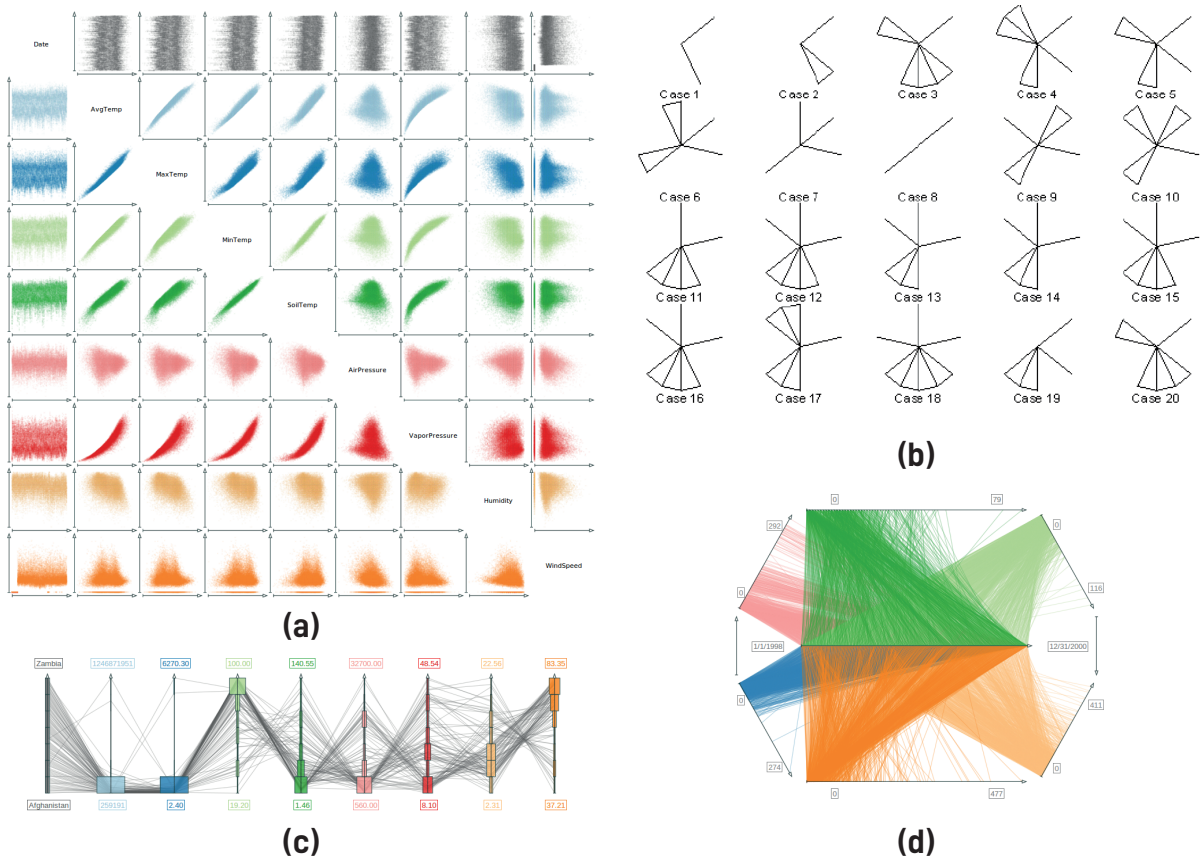


Figura 4.4: Ejemplos de las distintas orientaciones de los ejes. (a) *Rectilínea*: matriz de diagramas de dispersión de datos meteorológicos. Figura extraída de [TS20]. (b) *Radial*: glifos de estrellas del conjunto de datos de animales. Figura extraída de [LRB03]. (c) *Paralela*: coordenadas paralelas con histogramas que muestran datos demográficos. Figura extraída de [TS20]. (d) *Libre*: visualización de *TimeWheel* de datos de salud humana. Figura extraída de [TS20].

4.2.2. Sustrato Gráfico

El sustrato gráfico está compuesto por los elementos que se utilizarán para representar los datos en la vista a generar y sus propiedades gráficas. La información presente en el sustrato gráfico se puede descomponer en *marcas* (elementos visuales) y *canales visuales* (atributos de esos elementos).

4.2.2.1. Marcas

Una marca es un objeto geométrico primitivo que se clasifica según el número de dimensiones espaciales que ocupa. Card *et al.* [CMS99] identifica cuatro tipos de marcas: puntos (0D), líneas (1D), áreas o superficies (2D) y volúmenes (3D). Cada marca siempre representa un elemento.

Por ejemplo, en un gráfico de dispersión 2D, los puntos son marcas que se posicionan a lo largo de dos ejes cuantitativos ortogonales. En un mapa coroplético (o mapa de colores), los países son marcas de área situadas en un sustrato espacial geográfico. Las líneas en un gráfico de coordenadas paralelas [Ins85] son un ejemplo clásico de marcas unidimensionales (1D).

Además de estas marcas básicas, existen glifos o íconos que pueden representar datos con múltiples dimensiones, y son considerados marcas n-dimensionales. Los glifos permiten representar datos con múltiples variables mediante un solo objeto gráfico. Un ejemplo destacado de glifos n-dimensionales son las caras de Chernoff [Che73]. Estos glifos utilizan la forma y las características faciales para representar datos multidimensionales. Cada atributo facial (como la forma de la cabeza, el tamaño de los ojos, la curvatura de la boca, el grosor de las cejas, etc.) corresponde a una variable en el conjunto de datos. Por ejemplo, una cara de Chernoff podría mostrar diferentes tamaños de ojos para representar una variable cuantitativa, la forma de la boca para otra y la posición de la nariz para una tercera, permitiendo así visualizar simultáneamente múltiples dimensiones en un solo glifo de forma compacta y visualmente intuitiva. Otro ejemplo de glifos n-dimensionales son los *star glyphs*, o glifos de estrella [JFK16, LRB03], que permiten representar múltiples variables utilizando un sistema de coordenadas radiales. Cada eje radial corresponde a una variable o dimensión del dato, y la longitud del segmento en cada eje indica el valor de esa variable. Los valores en los ejes se conectan para formar un polígono que proporciona una representación visual de un determinado dato. Estos polígonos permiten comparar múltiples datos, ya que las diferencias en la forma o tamaño de los mismos reflejan variaciones en los valores de las variables.

La sección 4.3 ofrece una explicación detallada sobre éstas y otras técnicas de visualización basadas en íconos.

4.2.2.2. Canales Visuales

Un canal visual, también conocido como variable retiniana [Ber83], es una característica gráfica utilizada para controlar y variar la apariencia de las marcas en una visualización. Estos canales visuales permiten codificar información de manera efectiva mediante propiedades visuales que los humanos perciben de manera inmediata.

En su libro [Ber83], Jacques Bertin identifica siete variables retinianas principales:

la *posición*, que se refiere a la ubicación de un elemento gráfico en el plano; el *tamaño*, que representa magnitudes; la *forma*, que diferencia elementos por su geometría; el *valor*, que expresa diferencias de cantidad o intensidad a través de claros y oscuros; la *textura*, que utiliza patrones repetidos para distinguir categorías; el *color* (*hue* o matiz), empleado para clasificar o agrupar elementos y la *orientación*, que utiliza la inclinación para marcar diferencias direccionales. Cada uno de estos canales visuales tiene un papel específico en la representación de datos y en cómo se perciben las marcas en una visualización. La elección adecuada de estos canales es crucial para garantizar que la información se comunique de manera clara y efectiva.

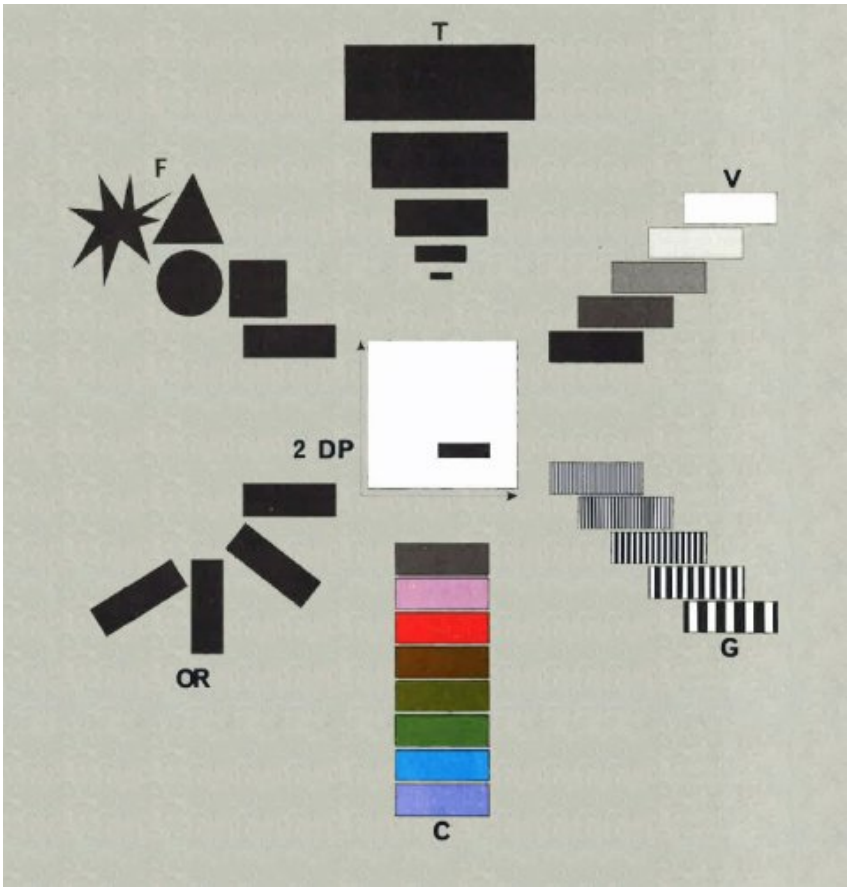


Figura 4.5: Variables visuales de Bertin. Figura extraída de [Ber83].

La figura 4.5 ilustra un ejemplo de variación de canales visuales en un rectángulo. Cada variación representa una forma distinta de codificar información en el mismo objeto gráfico.

4.3. Técnicas de Visualización para Datos Multidimensionales

En el amplio ámbito de la visualización de datos, existen múltiples técnicas y métodos diseñados para representar información de manera expresiva y efectiva. Las técnicas de visualización de datos van desde gráficos sencillos, como *scatterplots*[1] o diagramas de dispersión, gráficos de barras[2] y *pie chart*[3] o gráfico de tortas, hasta representaciones más complejas que incluyen visualizaciones multidimensionales. Cada técnica tiene sus propias fortalezas y limitaciones, y la elección de una u otra depende de varios factores. Sin embargo, cuando se trata de datos con múltiples dimensiones o características, las técnicas tradicionales pueden resultar insuficientes para captar toda la complejidad y riqueza de la información.

La visualización de datos multidimensionales ha sido objeto de estudio por expertos de diversas áreas mucho antes de que la informática adquiriera su relevancia actual. Desde la llegada de la era computacional, se han desarrollado numerosas técnicas de visualización y se han ampliado las técnicas existentes para manejar volúmenes de datos cada vez más grandes. La búsqueda de técnicas de visualización expresivas y efectivas para el análisis y exploración de estos conjuntos de datos ha crecido desde entonces.

Existen múltiples técnicas bien conocidas para visualizar datos multidimensionales como las coordenadas paralelas [ID09], las caras de Chernoff [Che73, FR81], la matriz de diagramas de dispersión [CLN87, EDF08], entre otras. Clasificar estas técnicas de visualización es una tarea compleja. Esto se debe a la diversidad de criterios que podrían considerarse, como el objetivo de la visualización, la naturaleza y la dimensionalidad de los datos, además de las propiedades específicas de las técnicas utilizadas.

Wong y Bergeron [WB94] clasificaron las técnicas de visualización multidimensio-

[1] *Scatterplot.* Representación gráfica que utiliza puntos para visualizar la relación entre dos variables en un sistema de ejes ortogonal. Cada punto en el gráfico representa un ítem del conjunto de datos y está posicionado de acuerdo con los valores de las dos variables.

[2] *Gráfico de barras.* Representación gráfica en la que las categorías se muestran en el eje horizontal y la magnitud o frecuencia de cada categoría en el eje vertical. Cada categoría se representa mediante una barra cuya altura o longitud es proporcional al valor que representa.

[3] *Pie chart.* Representación gráfica en la que un círculo se divide en sectores para mostrar la proporción de cada categoría respecto al total. Cada sector representa un porcentaje del total, facilitando la comparación de las partes con el todo.

nales basadas en visualizaciones bivariadas (como la matriz de diagramas de dispersión [CLN87]), visualizaciones multivariadas (que incluyen la matriz de diagramas de dispersión con *brushing*[4], matriz de paneles, visualizaciones basadas en íconos, visualizaciones jerárquicas y no cartesianas) y animaciones. Keim y Kriegel [KK96], por su parte, clasificaron las técnicas de visualización en seis clases según su modo de visualización: (1) técnicas de proyección geométrica, (2) basadas en íconos, (3) orientadas a píxeles, (4) jerárquicas, (5) basadas en grafos e (6) híbridas. Años más tarde, Tominski y Schumann [TS20] también presentan seis clases para categorizar técnicas multidimensionales, algunas de las cuales se solapan con las de Keim y Kriegel. Estas corresponden a técnicas (1) basadas en tablas, (2) bivariadas combinadas, (3) basadas en polilíneas, (4) basadas en íconos, (5) basadas en píxeles y (6) anidadas.

Una consideración relevante es si los datos se visualizarán con o sin pérdidas de información. Esto tiene que ver con la forma en que se representa la información original en la visualización. En una visualización sin pérdidas, se mantiene toda la información original de los datos. Los datos se representan de forma completa y exacta sin sacrificar ningún detalle. Algunas de las técnicas de visualización multidimensional más comúnmente utilizadas, como las matrices de diagramas de dispersión [CLN87] y las coordenadas paralelas [Ins85] son ejemplos de técnicas de visualización sin pérdida. En cambio, en la visualización con pérdidas, se sacrifica parte de la información original para hacer la visualización más comprensible para el usuario. Los métodos de reducción de dimensionalidad, como el análisis de componentes principales (PCA) [Jol02] y el escalado multidimensional (MDS) [Tor52], son ejemplos típicos. Estos métodos proyectan los datos originales en un espacio de menor dimensión, a menudo 2D o 3D, para facilitar su interpretación. Existen numerosos métodos que pueden emplearse para reducir la dimensionalidad. El propósito de estos métodos es representar los datos multidimensionales en un espacio de baja dimensionalidad de manera que se conserven ciertas propiedades del conjunto de datos (como distancias, topología u otras proximidades) de la manera más fiel posible.

Otro aspecto clave a considerar es la reversibilidad de la técnica. La reversibilidad implica la capacidad de recuperar el conjunto de datos original de manera exacta a partir de su visualización. En técnicas como PCA [Jol02] o t-SNE [vdMH08], aunque no haya pérdida de información importante, a menudo no es posible recuperar de manera exacta los

[4] *Brushing.* Interacción que permite seleccionar los objetos contenidos en un área establecida [Cle93].

datos originales debido a la reducción de dimensiones o a las transformaciones aplicadas, que afectan la estructura o la representación de los datos. Por lo tanto, aunque algunas visualizaciones intentan minimizar la pérdida de información al preservar las relaciones importantes entre las variables, no todas son estrictamente reversibles. La reversibilidad dependerá de la naturaleza de la transformación empleada y de si se conserva suficiente información para reconstruir los datos originales.

Con base en todo lo anterior, hemos elaborado una clasificación combinada que comprende cinco categorías principales, incluyendo así las técnicas más utilizadas en la visualización de conjuntos de datos multidimensionales en la actualidad. La tabla 4.1 ofrece una breve descripción de cada una de las clases y proporciona ejemplos de cada una de ellas. A continuación, se explican en detalle y se ejemplifican las técnicas de visualización más representativas para datos multidimensionales, encuadrándose en la clasificación propuesta. Es importante señalar que, en cada caso, se consideraron las técnicas en su forma original, excluyéndose las variantes derivadas. Esta decisión tiene como finalidad preservar un enfoque claro y coherente, evitar redundancias y asegurar que el análisis se concentre en los principios conceptuales fundamentales de cada técnica, sin desviar la atención hacia adaptaciones específicas que no aportan innovaciones significativas.

4.3.1. Técnicas Basadas en Geometría

Las técnicas de visualización multidimensional basadas en geometría son un conjunto de métodos que fundamentan su funcionamiento en principios y transformaciones geométricas. Estas técnicas utilizan formas geométricas elementales como puntos o líneas para mostrar conjuntos de datos. Mediante transformaciones geométricas como rotaciones, escalados y proyecciones, se busca preservar las relaciones espaciales y las estructuras subyacentes de los datos, facilitando su interpretación y análisis visual. Estas técnicas son fundamentales en el análisis de datos complejos, ya que permiten explorar patrones y relaciones de manera accesible y comprensible.

4.3.1.1. Matriz de Diagramas de Dispersión (SPLOM)

Un diagrama de dispersión o *scatterplot* suele ser la primera opción para explorar un conjunto de datos, ya que permite fácilmente encontra relaciones entre dos o tres variables (*scatterplot 3D*) e identificar patrones en los datos. No obstante, conforme aumenta el

Categoría	**Descripción y Ejemplos**
Basadas en geometría	Utilizan principios y transformaciones fundamentales, como rotaciones, escalados y proyecciones, así como formas geométricas como puntos o líneas, para representar conjuntos de datos. - *Scatterplot matrix* [CLN87] - *General Line Coordinates* [Kov14] - *RadViz* [HGM$^+$97] - *Star coordinates* [Kan00, Kan01] - *Table lens* [RC94] - *Principal Component Analysis (PCA)* [Jol02] - *Uniform Manifold Approximation and Projection (UMAP)* [LJJ18]
Basadas en íconos	Mapean cada elemento de dato multidimensional a un ícono o glifo. Las características visuales varían según los valores de los atributos de los datos. - *Chernoff faces* [Che73] - *Star Glyph* [JFK16] - *Stick figures* [PG88] - *Shape Coding* [Bed90] - *Color icons* [Lev91]
Basadas en píxeles	Representan el valor de un atributo mediante un píxel basado en alguna escala de colores. - *Independientes de la consulta* [KKA95] - *Dependientes de la consulta* [Kei95, Kei96, AKK96]
Jerárquicas	Subdividen el espacio n-dimensional y presentan los subespacios de manera jerárquica. Las técnicas en esta categoría se centran principalmente en datos jerárquicos. - *Dimensional stacking* [LWW90] - *Treemap* [JS91] - *Dendrogram* [HKP12]
Basadas en grafos	Permiten visualizar de manera efectiva grafos utilizando algoritmos de diseño específicos, lenguajes de consulta y técnicas de abstracción. Se dividen en tres categorías principales: representaciones de nodos y enlaces, de matrices e implícitas. - *Force-directed Layouts* [FR91] - *Matrix View* [SM07] - *Arc Diagram* [LWW90]

Tabla 4.1: Taxonomía propuesta a partir de las taxonomías de Keim y Kriegel [KK96] y de Tominski y Schumann [TS20].

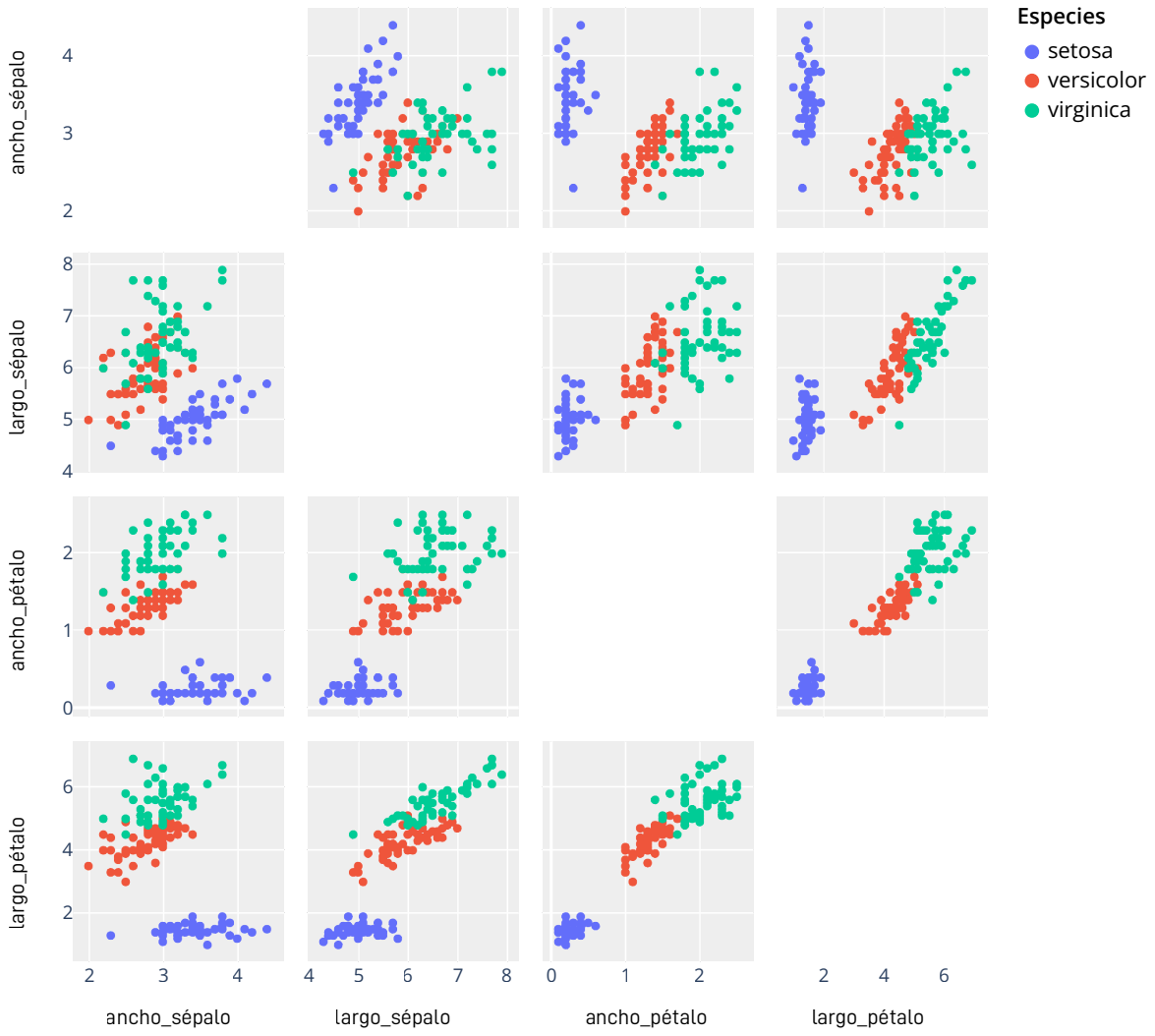


Figura 4.6: Matriz de diagramas de dispersión para el conjunto de datos *Iris* [Fis88].

número de dimensiones, el enfoque más común es utilizar una matriz de diagramas de dispersión [CLN87, EDF08] (también conocido como SPLOM o *scatterplot matrix*). Ésta consiste en una cuadrícula de diagramas de dispersión de n^2 celdas, donde n es el número de dimensiones del conjunto de datos. Cada celda de la matriz representa la relación entre pares de variables en un conjunto de datos, lo que permite visualizar simultáneamente las relaciones entre todas las combinaciones posibles de dos variables. El orden de las dimensiones suele ser el mismo para las orientaciones horizontal y vertical, lo que resulta en una matriz simétrica. La figura 4.6 presenta una matriz de gráficos de dispersión para el conjunto de datos *Iris* [Fis88].

4.3.1.2. Coordenadas Paralelas y Radiales

Las coordenadas paralelas, denominadas en inglés PCP (*Parallel Coordinates Plot*), fueron mencionadas por primera vez en la obra del matemático francés Maurice d'Ocagne en 1885 [d'O85]. Durante el siglo XX, la técnica fue explorada de manera esporádica en diferentes contextos para visualizar datos multidimensionales. Sin embargo, se popularizaron en la década de 1980 gracias al trabajo del matemático y científico de la computación Alfred Inselberg [Ins85]. Inselberg formalizó y redefinió la técnica, presentándola como una herramienta eficaz para la visualización de datos de alta dimensionalidad. Desde entonces, diversos investigadores han explorado y perfeccionado esta técnica para su aplicación en el análisis de datos multidimensionales [Sii00, GK03, CvW11, DSM17, XZM17].

En un gráfico de coordenadas paralelas, se generan tantos ejes, como atributos de un dato. Para cada valor de atributo de un determinado dato, se calcula una posición en el eje del atributo correspondiente. Las posiciones obtenidas se conectan para formar la polilínea que representa el dato completo. Este proceso se repite para cada muestra del conjunto de datos.

Existen variantes de las coordenadas paralelas, como las radiales y las jerárquicas. En las coordenadas radiales, los ejes se extienden desde un punto central hacia el exterior en distintas direcciones, separados por el mismo ángulo, como radios de un círculo [DLR09]. A menudo, los extremos de las líneas se conectan para formar una figura cerrada con forma de estrella. Si el valor del último atributo no se une con el del primero, las coordenadas radiales y paralelas son matemáticamente equivalentes, y la diferencia radica en la orientación (radial o paralelo) de los ejes coordenados. La figura 4.7 ilustra el mismo conjunto de datos en ambas representaciones.

Estas técnicas resultan útiles para la identificación de valores atípicos, relaciones, tendencias y clústeres. Si bien son técnicas reversibles y sin pérdida de información, no son suficientes para abordar algunos desafíos de visualización, como la oclusión, que es común en conjuntos de datos de mayor tamaño. Diversas alternativas [GXWY10, CvW11, GPS$^+$11, Kov14] han propuesto diferentes enfoques para abordar estos problemas, logrando, en cierta medida, superarlos en el contexto de grandes volúmenes de datos.

Con el fin de explorar soluciones y alternativas para este desafío, a partir de 2014, Kovalerchuk *et al.* [Kov14, Kov18, KG19] introdujo las técnicas GLC (*General Line Coordinates*) para visualizar, sin pérdida de información, datos multidimensionales en 2D y 3D

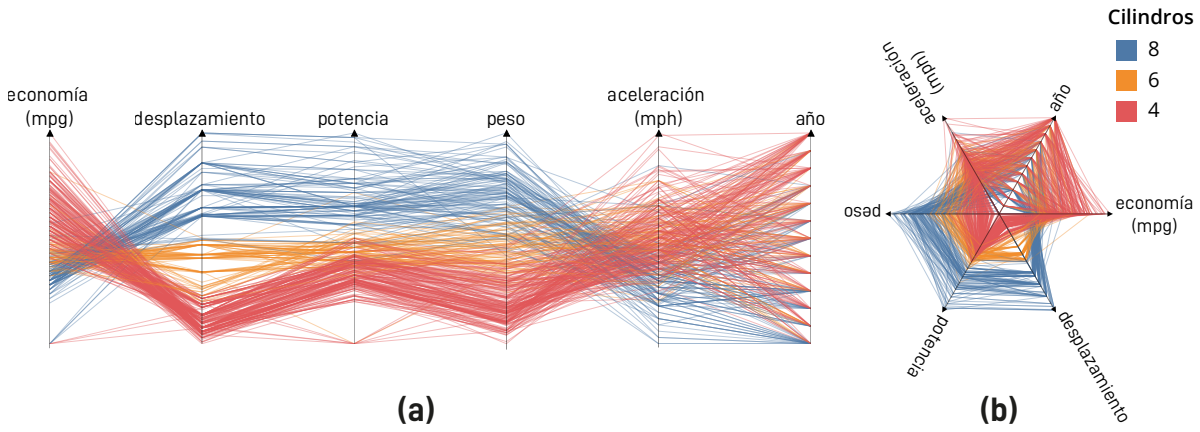


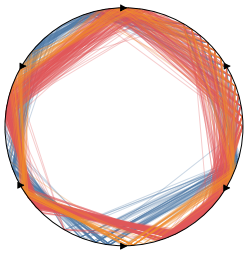
Figura 4.7: Coordenadas (a) paralelas y (b) radiales para el conjunto de datos de automóviles (Auto MPG) [AN07].

e identificó dos clases de GLC: las *Non-Paired General Line Coordinates* (NP-GLC) que generalizan las coordenadas paralelas y radiales e incluyen también las coordenadas *Circular*, *N-Gon*, *Bush*, *Generic Sequential*, *In-Line*, y *Non-Sequential*, y las *Paired General Line Coordinates* (P-GLC) que generalizan las coordenadas cartesianas, incluyendo las coordenadas *Collocated*, *Shifted*, *Anchored*, *Radial*, *Elliptic* y *Crown Paired Coordinates*.

En las P-GLC, un dato de n dimensiones $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, donde $i \in \{1, \ldots, m\}$, se divide en pares consecutivos disjuntos $(x_{i1}, x_{i2}), \ldots, (x_{in-1}, x_{in})$, cada uno representando un punto 2D asociado a un par de ejes. Así, un dato de n dimensiones se representa como un grafo dirigido basado en la idea de una colección de pares de dimensiones, lo que produce mucho menos desorden que los métodos NP-GLC. Además, requieren la mitad de las líneas que los NP-GLC. En contraste, en los NP-GLC, cada dimensión se representa en un eje, y cada dato X_i se visualiza mediante una poligonal que intersecta cada eje en el punto correspondiente a x_{ij}, donde $\{i = 1, \ldots, m, \quad j = 1, \ldots, n\}$. Considerando diferentes formas de disponer los ejes en la visualización podemos distinguir distintas representaciones de NP-GLC.

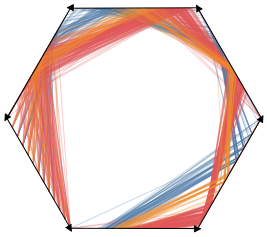
Non-Paired General Line Coordinates (*NP-GLC*)

Circular Coordinates



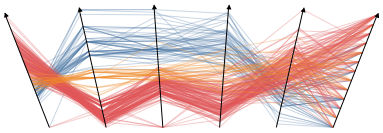
Los ejes son segmentos curvos dispuestos en un círculo.

N-Gon Coordinates



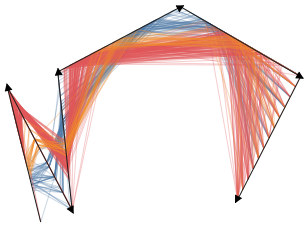
Los ejes están dispuestos formando un polígono (triángulo, cuadrado, pentágono, etc.). Cada lado del polígono corresponde a un eje y todos ellos están orientados en sentido horario o antihorario.

Bush Coordinates



Este diseño es una variante de las coordenadas paralelas, en la cual únicamente el eje de coordenadas central se orienta verticalmente. Los ejes restantes se disponen de forma inclinada, aumentando gradualmente su ángulo respecto al eje central, lo que crea una estructura similar a un arbusto.

Generic Sequential Coordinates

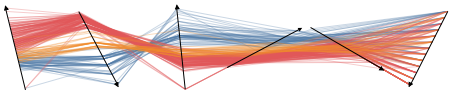


Los ejes están dispuestos consecutivamente en el plano uno tras otro, en sentido horario o antihorario, y cada eje está orientado en cualquier dirección.

In-Line Coordinates

Los ejes son horizontales, colineales y están orientados en la misma dirección; pueden superponerse o no.

Generic Non-Sequential Coordinates



Los ejes están orientados en cualquier dirección, separados, y dispuestos consecutivamente en el plano de izquierda a derecha.

Paired General Line Coordinates (*P-GLC*)

Collocated Paired Coordinates

Shifted Paired Coordinates

Sigue en la página siguiente.

...continuación de la página anterior.

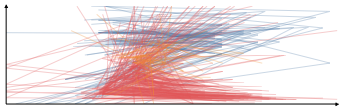
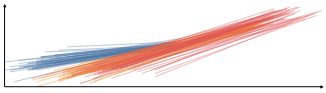
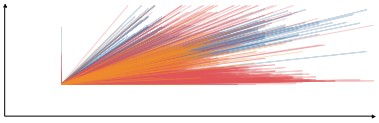
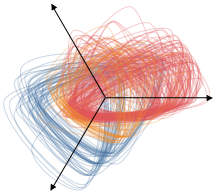
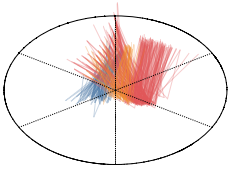
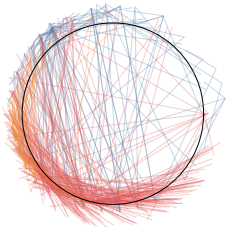
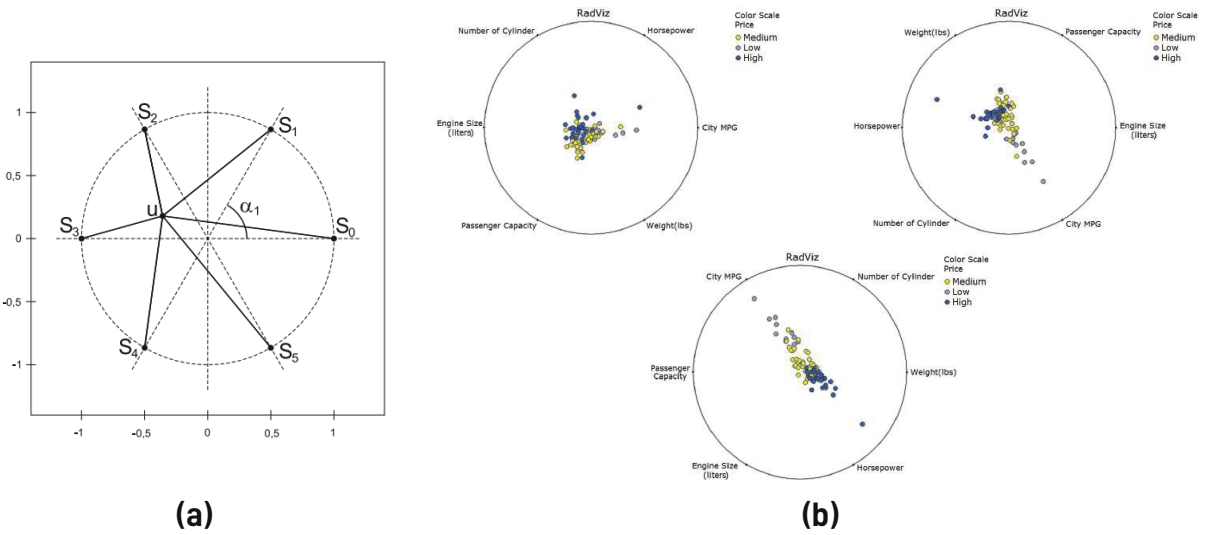
 Cada par se dibuja como un punto 2D en los mismos dos ejes en el plano, y estos puntos 2D están conectados para formar un grafo dirigido.	 Después del primer par, cada par siguiente se dibuja en un sistema de coordenadas desplazado.
Anchored Paired Coordinates  Después del primer par, cada par siguiente se dibuja desplazado, es decir, las coordenadas se desplazan a la ubicación de un par dado (ancla).	Paired Radial Coordinates  El círculo está dividido en $n/2$ sectores iguales. Cada par se representa en su propio sector como un punto, ubicado a una distancia r del centro.
Ellipse Paired Coordinates  Cada par se representa como la intersección de sus elipses correspondientes. Estos puntos están conectados para formar una polilínea.	Paired Crown Coordinates  Dado un casco convexo, las coordenadas impares se trazan sobre ella y las coordenadas pares se trazan ortogonalmente a ellas como una función de la coordenada impar.

Tabla 4.2: Resumen de las diferentes representaciones de GLC para el conjunto de datos de automóviles (Auto MPG) [AN07].

Las GLC proporcionan una gran variedad de nuevas representaciones visuales para datos multidimensionales, sin reducción de dimensiones. Esto aumenta significativamente las posibilidades de identificar patrones, tendencias y relaciones relevantes en diversos conjuntos de datos multidimensionales. En la tabla 4.2 se definen e ilustran estas técnicas.



(a) (b)

Figura 4.8: Visualización radial (RadViz). (a) Mapeo para $N = 6$. Figura extraída de [NŠ09]. (b) Diferentes vistas del conjunto de datos de automóviles (Auto MPG) [AN07]. Los automóviles están codificados por color en función de su costo (bajo, medio y alto) y se ha realizado una reorganización manual de las dimensiones. Figura extraída de [WGK10].

4.3.1.3. Visualización Radial (RadViz)

La visualización de coordenadas radiales, conocidas como RadViz (del inglés *Radial Visualization*), es una técnica de visualización utilizada para representar datos de alta dimensión en un plano [HGM+97, HG01, DGRG12]. Esta técnica se basa en el principio físico de la ley de Hooke[5], y modela un sistema físico de resortes donde las dimensiones constituyen puntos de anclaje.

En RadViz, los datos se representan como puntos dentro de un círculo. Alrededor de la circunferencia de este círculo se disponen puntos de anclaje, también llamados "anclas", que representan los diferentes atributos o dimensiones del conjunto de datos. Para un conjunto de datos con 6 dimensiones por ejemplo, habría 6 anclas distribuidas uniformemente alrededor del círculo (ver figura 4.8(a)). Cada punto de datos está conectado a cada una de las anclas mediante un resorte virtual. La posición final de cada punto dentro del círculo se determina por un proceso de equilibrio de fuerzas, donde la suma de las fuerzas ejercidas por todos los resortes debe ser igual a cero.

La característica clave de la técnica es que la rigidez de cada resorte está determinada

[5]La ley de Hooke establece que la fuerza F ejercida por un resorte es proporcional a la elongación δ del resorte: $F = -k\delta$, donde k es la constante elástica del resorte.

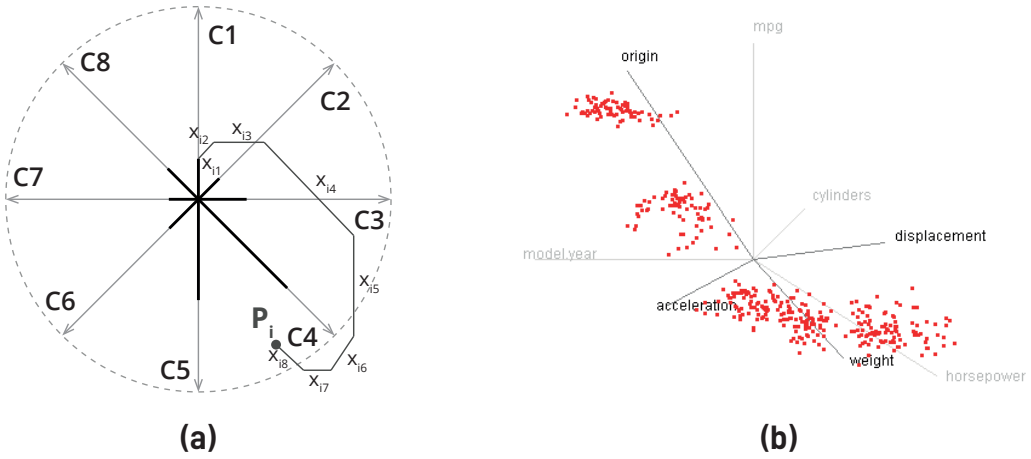


Figura 4.9: Coordenadas estrella. (a) Cálculo de la posición final del punto P_i en el espacio bidimensional, representación de la muestra X_i de un conjunto de datos de ocho dimensiones. Figura adaptada de [Kan00]. (b) Ejemplo de coordenadas estrella utilizando el conjunto de datos de automóviles (Auto MPG) [AN07]. Figura extraída de [Kan00].

por los atributos del elemento de datos correspondiente. Estos atributos deben ser no negativos y representan la importancia relativa de cada dimensión para ese punto de datos específico. Esta técnica permite visualizar relaciones complejas entre múltiples variables simultáneamente. Los puntos que se agrupan cerca de ciertas anclas indican una mayor relevancia de esas características particulares, mientras que los puntos ubicados en el centro del círculo sugieren una distribución equilibrada entre todas las características.

El RadViz es particularmente útil para identificar patrones, clústeres [NŠ09] y valores atípicos en conjuntos de datos multidimensionales, ofreciendo una representación intuitiva y visualmente accesible de datos complejos.

Es importante destacar que diferentes ubicaciones y ordenamientos de los anclajes darán diferentes resultados, y que puntos que son distintos en n dimensiones pueden mapearse a la misma ubicación en 2D. En la figura 4.8(b) se observan diferentes vistas del conjunto de datos de automóviles (Auto MPG) [AN07]. Los automóviles están codificados por colores de acuerdo a su costo (bajo, medio y alto), y se emplea el reordenamiento manual de dimensiones.

4.3.1.4. Coordenadas Estrella (SC)

Las coordenadas estrella (SC, por sus siglas en inglés de *Star Coordinates*), introducidas por Eser Kandogan [Kan00, Kan01], es una técnica de visualización particularmente

útil para la exploración y el análisis visual de conjuntos de datos complejos que genera mapeos lineales para representar datos de alta dimensionalidad en un espacio de menor dimensión, típicamente bidimensional o tridimensional.

El gráfico se construye utilizando un conjunto de vectores C_j de n dimensiones, para $j = 1, ..., n$, con un punto de origen común que representa los ejes radiales. Cada vector C_j está asociado a la i-ésima variable de los datos. En la representación visual inicial, todos los ejes tienen la misma longitud y están uniformemente distribuidos alrededor de un círculo unitario en el plano. Para cada punto de datos, se calcula la longitud de cada vector para representar el valor en la dimensión correspondiente. En la figura 4.9(a), se ilustra un ejemplo de coordenadas estrella con ocho ejes $C_1, \ldots, C_8$ que representan las ocho dimensiones. La representación de un dato en particular comienza en el centro del círculo, moviéndose a lo largo del eje C_1 con longitud x_{i1} (valor del primer atributo para la i-ésima muestra), continuando paralelamente al eje C_2 con longitud x_{i2}, y así sucesivamente hasta dibujar todas las componentes. La posición final de un punto P_i en el espacio bidimensional es una representación de la muestra $X_i = (x_{i1}, \ldots, x_{in})$ con n características o atributos. En el sistema de coordenadas estrella, los vectores son linealmente dependientes y la representación de un punto no es única.

Los usuarios pueden aplicar transformaciones de escala para ajustar la longitud de un eje, aumentando o disminuyendo así la contribución de un atributo, o aplicar transformaciones de rotación para cambiar la orientación de un eje, lo que altera los ángulos y, en consecuencia, modifica el grado de correlación de un atributo con respecto a los demás. En la figura 4.9(b) se presenta un ejemplo de coordenadas estrella aplicado al conjunto de datos de automóviles (Auto MPG) [AN07]. Al ajustar la escala de la coordenada "origen", se pueden identificar dos grupos principales.

4.3.1.5. Visualización Basada en Tablas

Los datos multidimensionales suelen representarse en forma de tablas, donde las columnas indican atributos y las filas corresponden a muestras u observaciones. La visualización basada en tablas mantiene el formato tabular de las hojas de cálculo, pero sustituye la representación textual de los valores por una representación visual. Por ejemplo, en lugar de mostrar valores numéricos, se pueden incorporar barras dentro de las celdas de la tabla, ajustando la longitud de las barras de acuerdo con los valores de los

Figura 4.10: Ejemplo de visualización utilizando la técnica de *table lens*, la cual permite explorar datos demográficos de 111 países. En la visualización, se han ampliado las alturas de las filas de interés para resaltar información específica. Figura extraída de [TS20].

datos. Además, estas visualizaciones suelen conservar funciones típicas de las hojas de cálculo, como la reorganización de columnas y el ordenamiento de filas según los valores de los atributos.

La técnica denominada "lente de tabla" o *table lens* en inglés [RC94] es un enfoque de visualización que combina la estructura de tabla conocida con representaciones gráficas y un conjunto restringido de funcionalidades para mejorar el análisis de datos. Por ejemplo, *table lens* utiliza la técnica de Foco + Contexto [6] que permite visualizar la tabla en su totalidad y hacer *zoom* en elementos específicos mediante lentes de tabla locales. Además de controlar qué filas y columnas están "enfocadas", esta técnica ofrece la posibilidad de reorganizar columnas y crear nuevas columnas calculadas utilizando fórmulas basadas en otras. Para representar variables cuantitativas, emplea pequeñas barras horizontales cuya longitud codifica valores específicos de los datos. Esta representación es especialmente útil para obtener una visión general de los datos. Asimismo, los usuarios pueden seleccionar filas para enfocar su atención, facilitando la lectura de los valores de datos impresos y su comprensión. La figura 4.10 muestra un ejemplo de visualización de datos demográficos de 111 países.

A diferencia de las hojas de cálculo tradicionales, que, aunque intuitivas, pueden no ser adecuadas para presentar grandes volúmenes de datos de manera concisa, la lente de tabla resuelve este problema mediante el uso de barras delgadas para codificar los valores, reduciendo así el espacio de visualización necesario.

[6] La técnica Foco + Contexto distorsiona una vista, mediante un lente de distorsión, para destacar áreas de interés sin perder el contexto general ni alterar los datos subyacentes.

Las técnicas basadas en tablas pueden interpretarse desde una perspectiva geométrica, considerando la estructura tabular como un sistema de coordenadas en el que el eje vertical corresponde a las observaciones y el eje horizontal a las dimensiones o atributos. Las celdas funcionan como contenedores que utilizan formas geométricas básicas, como barras, para representar los datos y transformaciones de escalado para representar cuantitativamente los valores dentro del espacio limitado de cada celda.

4.3.1.6. Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA, por sus siglas en inglés, *Principal Component Analysis*) es una técnica utilizada para la reducción de dimensionalidad en conjuntos de datos multidimensionales. Su objetivo es transformar las variables originales en un nuevo conjunto, llamadas "componentes principales", que capturen la mayor parte de la variabilidad presente en los datos, reduciendo así el número de dimensiones necesarias para su representación sin perder información significativa.

La primera propuesta del método fue presentada por Karl Pearson [Pea01], pero el término "componentes principales" y su desarrollo teórico formal se atribuyen a Harold Hotelling [Hot33], quien estableció un método para la extracción de factores. Esta reducción dimensional se basa en la obtención de combinaciones lineales de las variables originales, y la posterior selección de un subconjunto de éstas [Jol02, Jac03].

El algoritmo se inicia calculando las nuevas variables como combinaciones lineales de las variables originales. La primera componente principal se define de tal forma que tenga la mayor varianza posible, lo que implica que capturará la mayor parte de la variabilidad presente en los datos. La segunda componente se calcula bajo la restricción de ser ortogonal a la primera y maximizar la varianza restante. Este proceso se repite con las componentes sucesivas, que se calculan de manera similar y son siempre ortogonales a las anteriores. Por ejemplo, si consideramos un conjunto de datos en dos dimensiones, como el que se ilustra en la figura 4.11, el análisis de componentes principales puede visualizarse como una rotación de los ejes originales de las variables hacia nuevos ejes ortogonales que coinciden con la dirección de máxima varianza de los datos. Esta primera dirección define la primera componente principal (PC1). Las componentes siguientes maximizan la varianza restante en direcciones ortogonales a las anteriores.

La representación gráfica utilizada habitualmente corresponde a los diagramas de

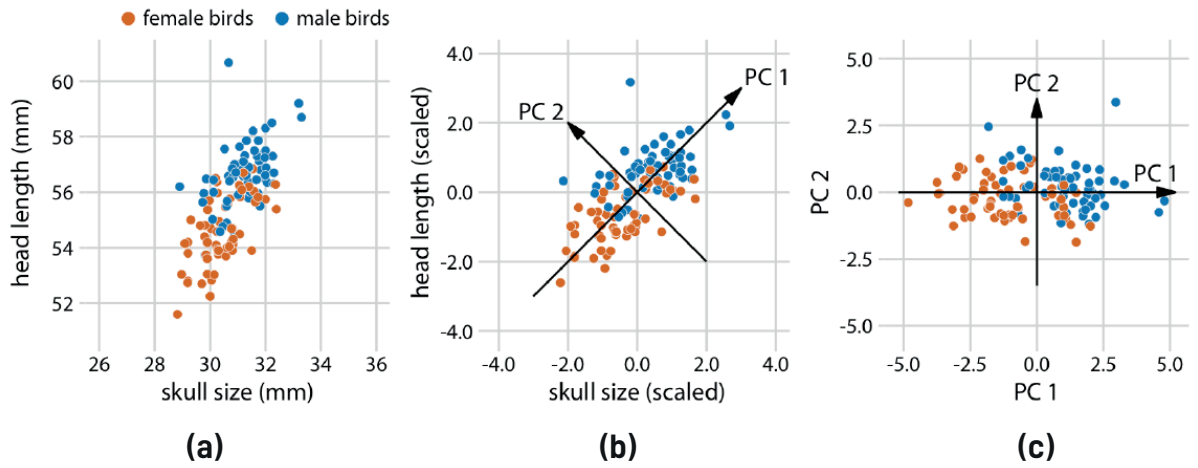


Figura 4.11: Análisis de componentes principales (PCA). (a) Datos originales: mediciones de longitud de cabeza y tamaño de cráneo del conjunto de datos del arrendajo azul, con machos y hembras diferenciados por color. (b) Primer paso del PCA: se escalan los datos originales a media cero y varianza unitaria, y luego se definen nuevas variables (componentes principales) en las direcciones de máxima variación de los datos. (c) Proyección de los datos en las nuevas coordenadas, lo que equivale matemáticamente a una rotación de los puntos de datos alrededor del origen. Figura extraída de [Wil19].

dispersión en 2D o 3D, donde los ejes corresponden a las dos o tres primeras componentes principales, que explican la mayor variabilidad de los datos en el espacio transformado.

Una de las razones de la popularidad de esta técnica es su relativa simplicidad de implementación y comprensión. Además, ha sido objeto de amplios estudios y cuenta con una sólida base teórica. Por otro lado, el análisis de componentes principales también se destaca por su eficiencia computacional, lo que le permite manejar grandes conjuntos de datos de manera efectiva.

4.3.1.7. UMAP

UMAP (por sus siglas en inglés, *Uniform Manifold Approximation and Projection*) es un algoritmo no lineal de reducción de dimensiones que tiene como objetivo preservar tanto la estructura local como una parte significativa de la estructura global de los datos. Desarrollado por McInnes *et al.* [LJJ18], este método se fundamenta en principios de la topología algebraica y la geometría Riemanniana. El algoritmo parte de la idea de que los datos en alta dimensión pueden ser proyectados a un espacio de menor dimensión, conocido como *manifold* [Row] (o *variedad* en español), sin perder la información esencial.

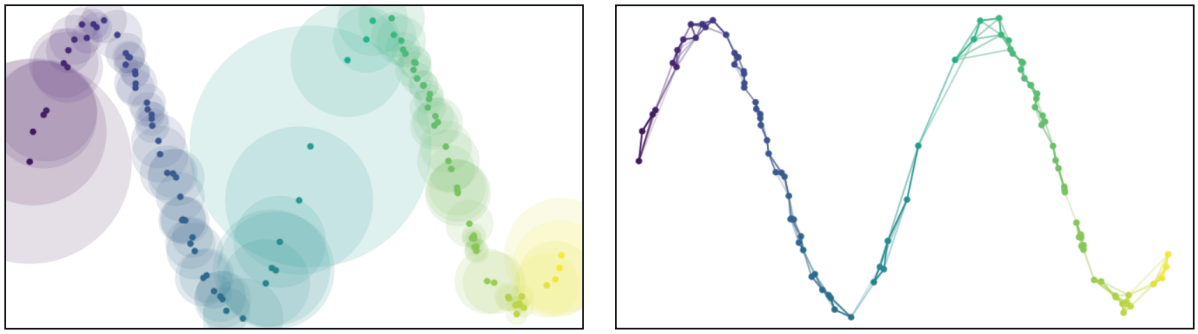


Figura 4.12: *Uniform Manifold Approximation and Projection* (UMAP). (a) Ejemplo en dos dimensiones de esferas unitarias construidas considerando los 5 vecinos más cercanos. (b) *Complejo de símplices* obtenido considerando la distancia entre vecinos. Figuras extraídas de [UMA].

Se inicia con la construcción de un grafo que representa las relaciones entre los puntos en el espacio de alta dimensión. Para ello, el algoritmo identifica los vecinos más cercanos de cada punto, es decir, aquellos que están a menor distancia, utilizando esferas que abarcan los puntos más próximos alrededor de cada uno. Una característica distintiva de UMAP es su enfoque adaptativo para medir distancias: ajusta la medida de distancia en torno a cada punto en función de su vecino más cercano, lo que permite obtener una distribución uniforme en el *manifold*. Posteriormente, se evalúa la similitud entre estos vecinos para comprender las relaciones locales en el conjunto de datos y se le asigna una importancia relativa a la posible conexión de puntos, en función de la distancia. La figura 4.12(a) muestra las esferas unitarias construidas cuando se consideran los 5 vecinos más cercanos para cada punto.

Luego, se forma el *complejo de símplices* a partir de los vecinos encontrados y los pesos relativos como se muestra en la figura 4.12(b). Un *complejo de símplices* es una estructura compleja formada por *símplices*, que capturan las relaciones topológicas entre los datos. Los *símplices* son formas geométricas básicas empleadas para construir objetos n-dimensionales. En términos geométricos, un *n-simplex* es un *simplex*[7] de n dimensiones, construido a partir de la cobertura convexa de $n + 1$ puntos. La figura 4.13 ilustra los primeros 4 *símplices* de menor dimensión.

Posteriormente, el algoritmo proyecta los datos a un espacio de menor dimensión y optimiza la disposición de los puntos en la dimensión reducida para asegurar la preservación

[7] *Simplex*: singular de *símplices*.

de las relaciones topológicas cruciales, manteniendo tanto la estructura global como local de los datos en la representación final. El objetivo es que los puntos cercanos en el espacio original continúen siendo próximos en el espacio reducido, y que las relaciones generales entre grupos de puntos también se mantengan.

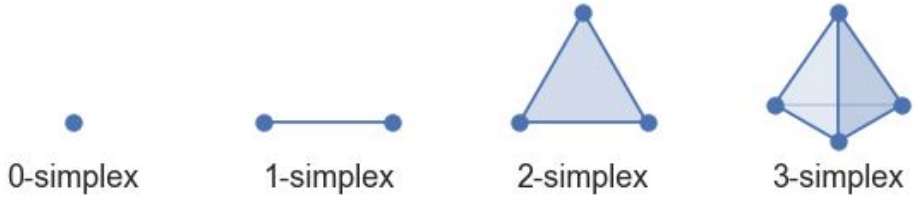


Figura 4.13: Primeros 4 *símplices* de menor dimensión. Figura extraída de [UMA].

La eficiencia del algoritmo depende en gran medida de la elección de los hiperparámetros, especialmente de la cantidad de vecinos a considerar ($n\_neighbors$) y la distancia mínima (min_dist) (ver figura 4.14). El parámetro $n\_neighbors$ nos permite controlar la forma en la que UMAP construye la estructura de cada uno de los puntos con sus puntos vecinos. Un valor más bajo de este parámetro implica que el algoritmo se centra en una estructura más local de los datos, mientras que un valor mayor tiende a capturar más la estructura global de los datos. Por otro lado, min_dist controla la distancia mínima que hay entre los puntos representados en el espacio de dimensión reducida, evitando así que se aglomeren demasiado entre ellos. La configuración de estos hiperparámetros influye significativamente en cómo UMAP balancea la preservación de la estructura local versus global de los datos y en la distribución final de los puntos en el espacio reducido.

En resumen, UMAP asegura la conservación de las propiedades topológicas de los datos en alta dimensión en la representación en baja dimensión, destacándose además por su rapidez y eficacia, lo que lo convierte en una herramienta valiosa para la interpretación de relaciones en conjuntos de datos complejos.

4.3.2. Técnicas Basadas en Íconos

Una categoría destacada en el análisis visual de datos multidimensionales es la que se basa en íconos o glifos [War08]. Estas técnicas representan cada muestra del conjunto de datos mediante un glifo, cuyas características visuales se ajustan en función de los

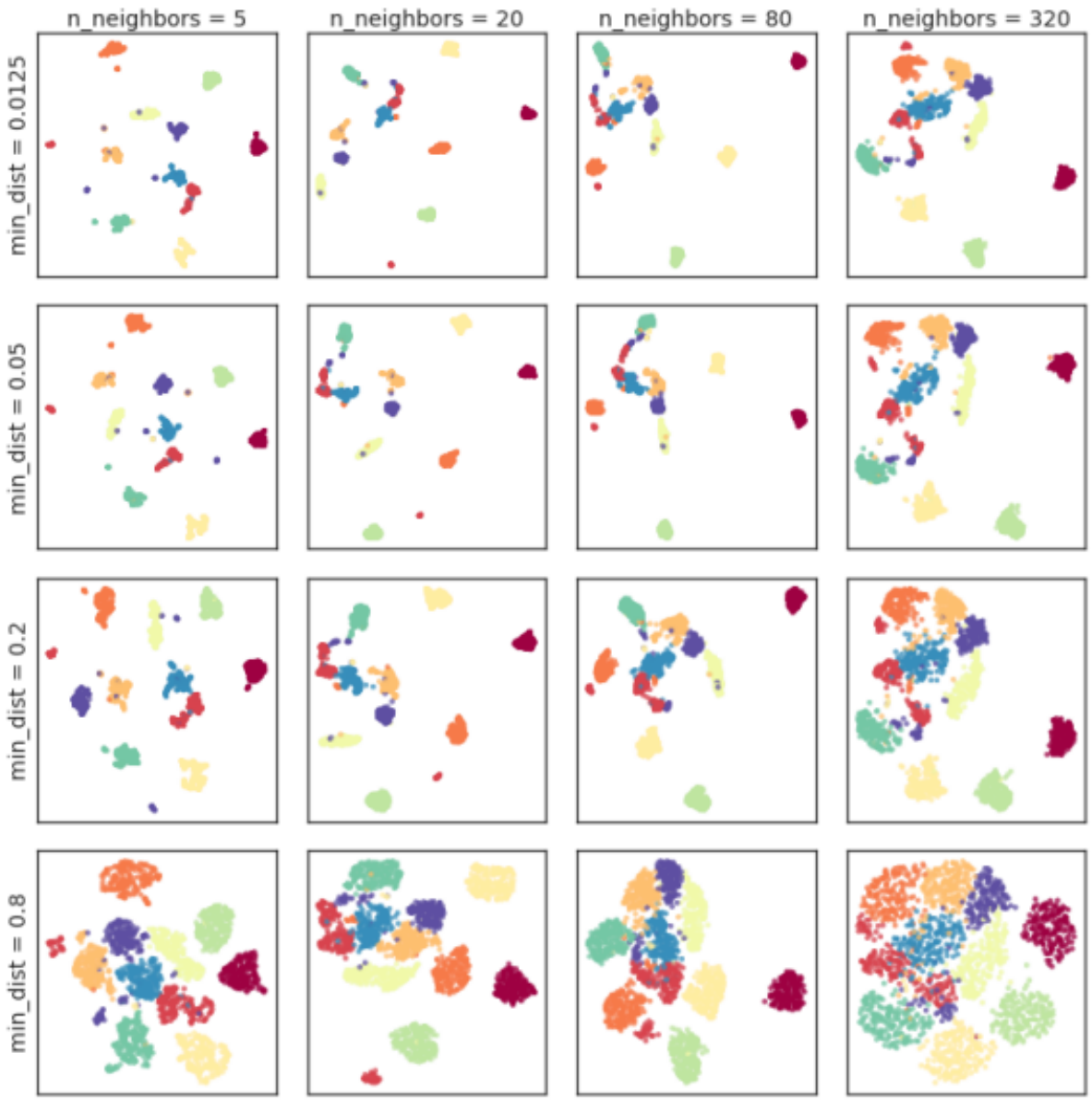


Figura 4.14: Variación de los hiperparámetros *n_neighbors* (cantidad de vecinos) y *min_dist* (distancia mínima entre puntos) en UMAP. Se utilizó el conjunto de datos *PenDigits*, donde cada punto corresponde a una imagen en escala de grises de 8x8 píxeles de un dígito escrito a mano. Figura extraída de [LJJ18].

valores de los atributos correspondientes. Los glifos poseen diversas propiedades gráficas, como forma, color y tamaño, que pueden asignarse a distintos atributos de los datos. La figura 4.15 ilustra ejemplos de glifos descritos en la literatura.

Las técnicas basadas en íconos resultan útiles para visualizar conjuntos de datos multidimensionales, permitiendo una representación compacta y fácilmente interpretable. Sin embargo, el diseño de íconos no es trivial [BKC+13], y puede generar dificultades interpretativas cuando el número de dimensiones es alto o cuando las diferencias entre los valores de los atributos son sutiles.

Una vez diseñados los glifos, es fundamental determinar su ubicación dentro del espacio de visualización. Estos glifos son marcas n-dimensionales que tienen la capacidad de adaptarse a diversas configuraciones del sustrato espacial (ver sección 4.2.1). Una correcta disposición de los glifos en pantalla no solo facilita la representación visual efectiva de los datos, sino que también puede reflejar las relaciones entre los elementos de datos.

Figura 4.15: Ejemplos de glifos. Figura adaptada de [War08].

A lo largo de los años, se han desarrollado múltiples estrategias para su ubicación [War02]. La estrategia más sencilla consiste en disponer los glifos de manera secuencial, ubicándolos uno junto a otro sin seguir un criterio específico. La figura 4.18(b) ilustra un subconjunto de las caras de Chernoff [Che73], correspondientes a 24 especímenes fósiles hallados en Jamaica, organizados en un orden secuencial. Otra alternativa es utilizar directamente algunas dimensiones de los datos como componentes de posición. En este enfoque, una,

dos o tres dimensiones de los datos se emplean para determinar la posición, las cuales también pueden ser representadas de manera redundante en el propio glifo. Entre las principales ventajas de esta estrategia destacan su facilidad de interpretación, su bajo costo computacional y su efectividad para revelar correlaciones entre las dimensiones que definen la ubicación espacial. Adicionalmente, los glifos pueden ser posicionados a partir de datos derivados que utiliza un proceso analítico para generar posiciones utilizando los valores de los datos como entrada. De esta manera, en lugar de que una ubicación refleje solo algunas dimensiones de los datos, se utiliza una combinación de todas las dimensiones por ejemplo mediante el uso de técnicas de reducción de dimensionalidad como PCA [Jol02] y MDS [Tor52]. Por otra parte, es posible emplear algoritmos basados en la estructura de los datos, los cuales presuponen la existencia de una conectividad o relación, ya sea implícita o explícita, entre los puntos de datos. Esta relación puede no estar explícitamente representada en los valores originales de los datos, pero puede derivarse de información contextual. Por ejemplo, podría haber un orden temporal o una relación jerárquica que no esté incluida como uno de los campos de datos.

4.3.2.1. Diagramas de Caja

El diagrama de caja univariante, conocido también como *boxplot* o diagrama de caja y bigotes, fue introducido por Tukey *et al.* [Tuk77, MTL78] como una herramienta para resumir las cantidades estadísticas descriptivas de un conjunto. En la figura 4.16(a) se representa un diagrama de caja típico. Estos resúmenes se basan en la mediana e incluyen el valor mínimo, la mediana de la primera mitad de los datos (primer cuartil, $Q1$), la mediana (segundo cuartil, $Q2$), la mediana de la segunda mitad de los datos (tercer cuartil, $Q3$) y el valor máximo. El área entre el primer y el tercer cuartil se conoce como el rango intercuartílico y da una indicación de la dispersión en los datos ($IQR = Q3–Q1$). El IQR corresponde visualmente a la caja y cubre aproximadamente el 50 % de las muestras más cercanas a la mediana. La línea en el centro de la caja indica la mediana, o segundo cuartil, que marca el punto medio de los datos. Cuando los datos individuales se encuentran fuera de los límites (bigotes) establecidos para representar el rango intercuartílico extendido (por ejemplo, $1{,}5IQR$), se consideran valores atípicos y se representan como puntos individuales en el diagrama.

El diagrama de caja permite una rápida comprensión de la distribución de los datos.

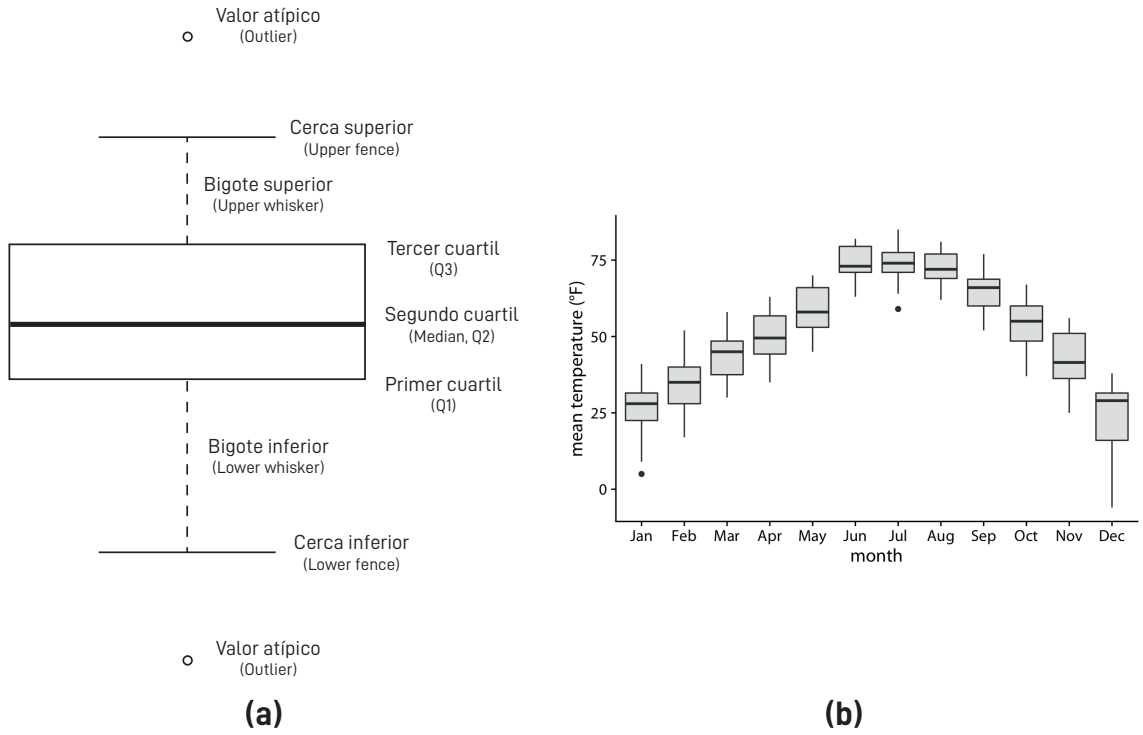


Figura 4.16: (a) Diagrama de caja típico. Figura adaptada de [IBAAR+18]. (b) Temperaturas medias diarias en Lincoln (Nebraska, Estados Unidos) visualizadas con diagramas de caja. Los glifos se han ubicado de acuerdo al atributo temporal correspondiente al mes en que se tomaron las mediciones. Figura extraída de [Wil19].

Su capacidad para resumir y visualizar estos aspectos lo convierte en una herramienta esencial para el análisis de datos, facilitando la comparación entre diferentes conjuntos de datos y proporcionando información sobre la variabilidad y la tendencia central. La figura 4.16(b) presenta las temperaturas medias diarias en Lincoln, ciudad de Nebraska, Estados Unidos, donde los glifos se han dispuesto de acuerdo al atributo temporal correspondiente al mes en que se tomaron las mediciones. Se observa un alto grado de sesgo en diciembre, en contraste con otros meses como julio, donde el sesgo es considerablemente menor.

En diversas ocasiones, los diagramas de caja han sido sustituidos por gráficos de violín, los cuales representan una alternativa más sofisticada y reveladora en la visualización de datos. El gráfico de violín o *violin chart* fue introducido formalmente por Hintze y Nelson [HN98] como una adaptación del diagrama de caja combinado con trazas de densidad. A diferencia de los diagramas de caja, que muestran percentiles seleccionados de la distribución, los gráficos de violín están diseñados para proporcionar una visión

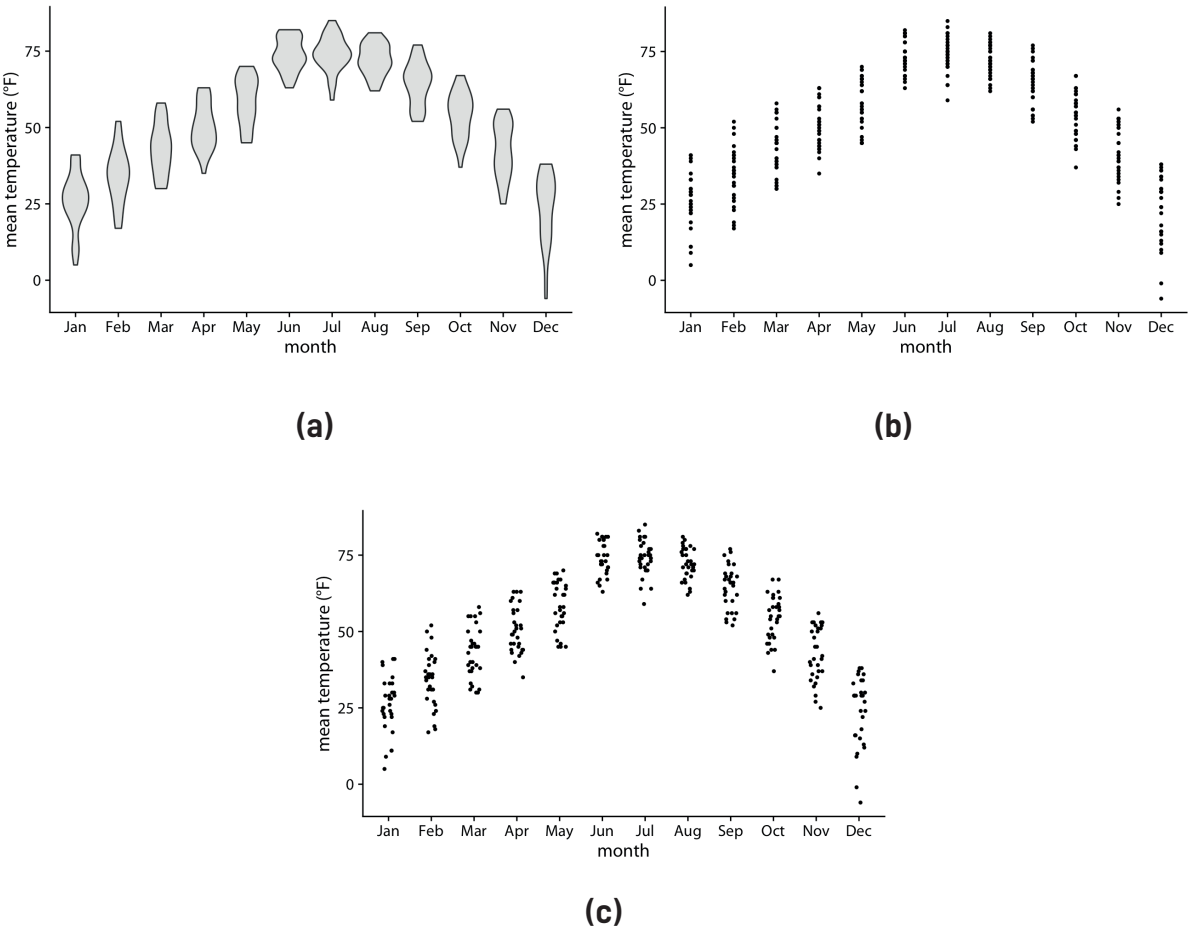


Figura 4.17: Temperaturas medias diarias en Lincoln visualizadas con (a) gráficos de violín y (b) gráficos de tiras. (c) Los puntos en los gráficos de tiras han sido ligeramente desplazados a lo largo del eje x mediante la técnica de *jittering*, lo que permite visualizar mejor la densidad de puntos en cada valor de temperatura. Figuras extraídas de [Wil19].

integral de la distribución de un conjunto de datos.

En un gráfico de violín, los valores se representan en el eje y. El ancho del violín en un valor y dado, refleja la densidad de puntos en esa posición, lo que permite una comprensión visual de la distribución de los datos a lo largo de la variable analizada. Los extremos del gráfico se extienden desde el valor mínimo hasta el máximo de los datos, mientras que la sección más ancha del violín indica la mayor concentración de puntos, proporcionando así una clara representación de las áreas donde los datos están más agrupados. En la figura 4.17(a) se visualizan los datos de temperatura de Lincoln con gráficos de violín. Se puede observar que algunos meses, como noviembre, presentan datos moderadamente bimodales.

A pesar de sus ventajas, los gráficos de violín también presentan algunas desventajas. Una de las críticas más frecuentes es que pueden dar la impresión de que la densidad de un conjunto de datos es muy alta cuando, en realidad, está bastante dispersa. Para abordar estas limitaciones, una solución efectiva es trazar todos los puntos de datos individuales de forma directa (ver figura 4.17(b)), mediante un enfoque conocido como gráfico de tiras o *strip chart*. Este método permite visualizar cada observación de manera explícita, proporcionando así una representación más precisa de la distribución de los datos.

Los gráficos de tiras son particularmente útiles, siempre que se evite la superposición excesiva de puntos en el mismo lugar, lo que podría complicar la interpretación de los resultados. Para ello se puede aplicar una técnica llamada *jittering*, que consiste en introducir un ligero desplazamiento aleatorio en la posición de los puntos a lo largo del eje x (ver figura 4.17(c)).

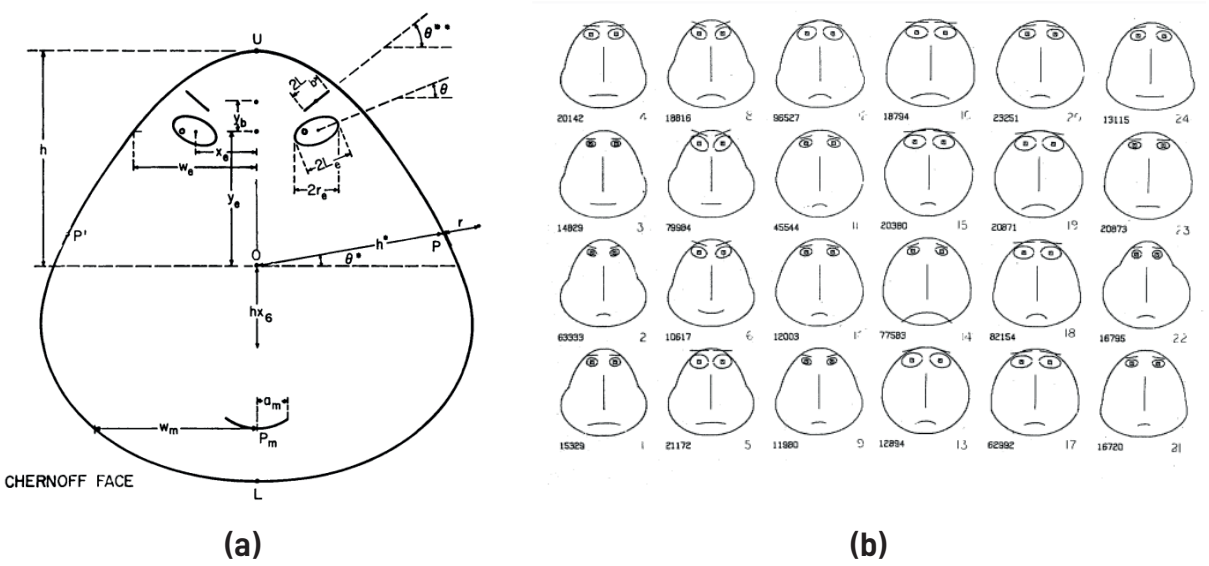


Figura 4.18: Caras de Chernoff. (a) Mapeo de características. Figura extraída de [HMB81]. (b) Subconjunto de 24 caras generadas a partir de ocho mediciones realizadas en 87 especímenes fósiles encontrados en Jamaica. Figura adaptada de [Che73].

4.3.2.2. Caras de Chernoff

Los primeros enfoques de visualización basados en íconos incluyen las reconocidas caras de Chernoff [Che73]. Esta técnica representa cada muestra de un conjunto de datos multidimensional mediante un rostro cuyas características visuales dependen de los valores de los atributos de la muestra. Así, los distintos parámetros visuales determinan la

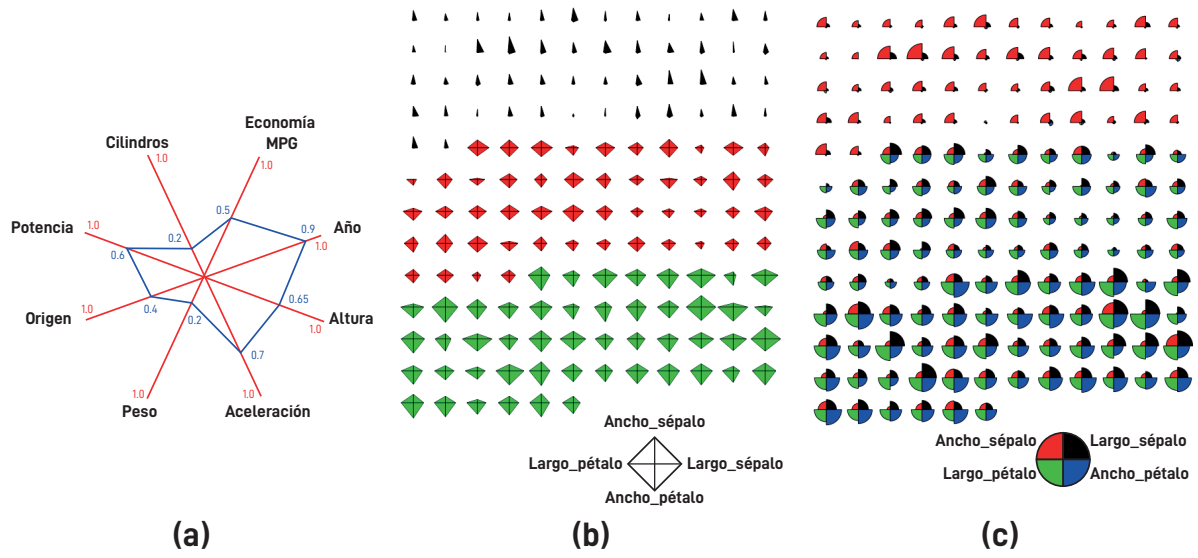


Figura 4.19: (a) Ejemplo de glifo de estrella. La línea azul conecta los diferentes puntos de valores de datos en cada eje para definir el glifo. Figura adaptada de [Ber14]. Ejemplos de (b) glifos de estrella y (c) diagramas de segmentos para el conjunto de datos *Iris* [Fis88]. Figuras adaptadas de [HBRD22].

apariencia de la cara, como se ilustra en la figura 4.18(a). Las caras de Chernoff permiten diversas configuraciones del sustrato espacial. Por ejemplo, es posible utilizar dos dimensiones de los datos para definir la posición 2D de las caras en un sistema de coordenadas rectilíneo, mientras que los atributos restantes se asignan a las propiedades visuales del rostro, tales como la forma de la boca, la nariz y los ojos. Otra configuración consiste en disponer las caras de manera secuencial, sin seguir un criterio específico, como se muestra en la figura 4.18(b).

Las caras son elementos visuales especiales porque los humanos están naturalmente familiarizados para reconocer las caras. Por lo tanto, sigue siendo un desafío asignar las características de los datos a las características faciales apropiadas para maximizar la efectividad de este método de visualización.

4.3.2.3. Glifo de Estrella

El glifo de estrella, o *star glyph*, es una técnica de visualización que permite representar múltiples dimensiones de datos en un único gráfico [JFK16, LRB03]. Cada observación se representa mediante un ícono en forma de estrella. Los rayos de la estrella, distribuidos uniformemente alrededor de un círculo, corresponden a las diferentes variables de un

elemento de datos. Los valores de cada atributo se normalizan en el intervalo [0,1], de manera que los atributos cercanos a cero se sitúan cerca del centro del círculo, mientras que aquellos cercanos a uno se localizan cerca del perímetro. A continuación, se traza una línea que conecta los puntos de los valores en cada eje (ver figura 4.19(a)).

Esta técnica es particularmente útil para conjuntos de datos multidimensionales de tamaño moderado. Sin embargo, su principal limitación radica en que la visualización puede volverse abrumadora a medida que aumenta el número de elementos de datos.

Una alternativa a los glifos de estrella son los diagramas de segmentos, en los cuales cada variable se utiliza para crear segmentos de colores. Los tamaños de estos segmentos son proporcionales al valor de la muestra representada en la dimensión correspondiente. En las figuras 4.19(b) y 4.19(c), se presenta el conjunto de datos *Iris* [Fis88] utilizando glifos de estrella y diagramas de segmentos, respectivamente. En ambos casos, los glifos están dispuestos en orden secuencial, sin utilizar un criterio específico.

4.3.2.4. Figuras de Palos

Pickett y Grinstein [PG88] introdujeron la técnica de visualización basada en íconos utilizando figuras de palo, conocidas como *stick figures*. En la figura 4.20(a) se ilustra un ejemplo típico de este tipo de ícono. Los segmentos de línea que lo componen se denominan "extremidades" y cada ícono posee una extremidad especial, denominada "cuerpo", que actúa como punto de referencia para las diversas transformaciones geométricas que se apliquen en el ícono. Cada extremidad tiene al menos tres parámetros que se pueden mapear a atributos de datos: el ángulo, la intensidad y la longitud.

La figura 4.20(b) presenta una familia de doce miembros, cada uno definido por la forma en que sus extremidades se conectan al cuerpo o entre sí. En uno de los miembros, las cuatro extremidades están directamente unidas al cuerpo, mientras que en otro, todas las extremidades se conectan a un único extremo del cuerpo, formando un brazo largo con tres articulaciones. Los diez miembros restantes exhiben diversas combinaciones de conexiones secuenciales y paralelas entre las extremidades.

Esta técnica resulta especialmente útil cuando los conjuntos de datos son relativamente densos en relación con las dimensiones de la pantalla, ya que los íconos compactos facilitan la visualización de patrones de textura que varían según las características de los datos (ver figura 4.20(c)).

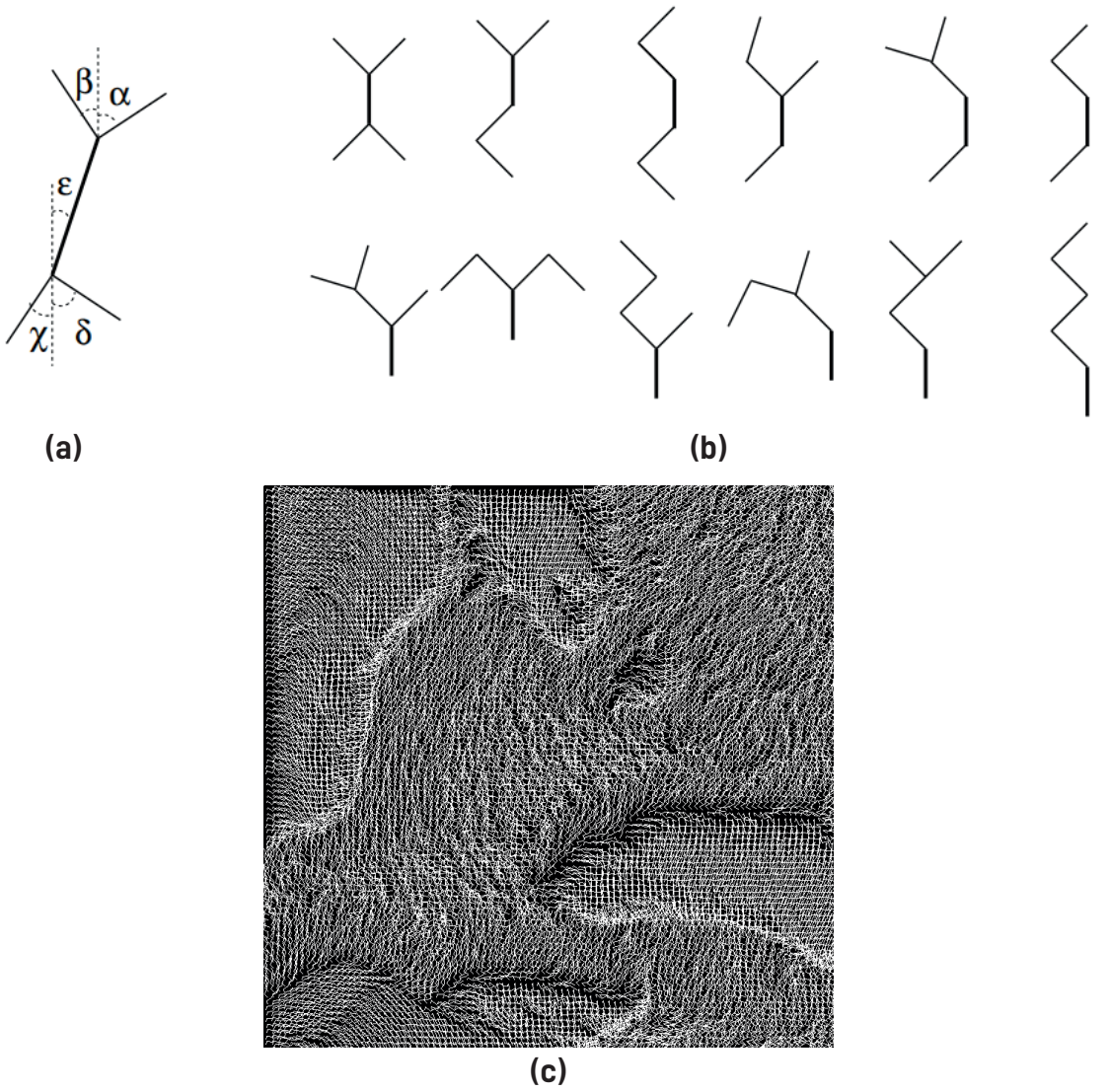


Figura 4.20: Figuras de palos. (a) Ícono de figura de cinco palos. Figura extraída de [WB94]. (b) Familia de íconos de figura de palos. Cada uno tiene un cuerpo y cuatro extremidades. Figura extraída de [WB94]. (c) Representación de datos de 5 dimensiones de la región Great Lake. Figura extraída de [Cha06].

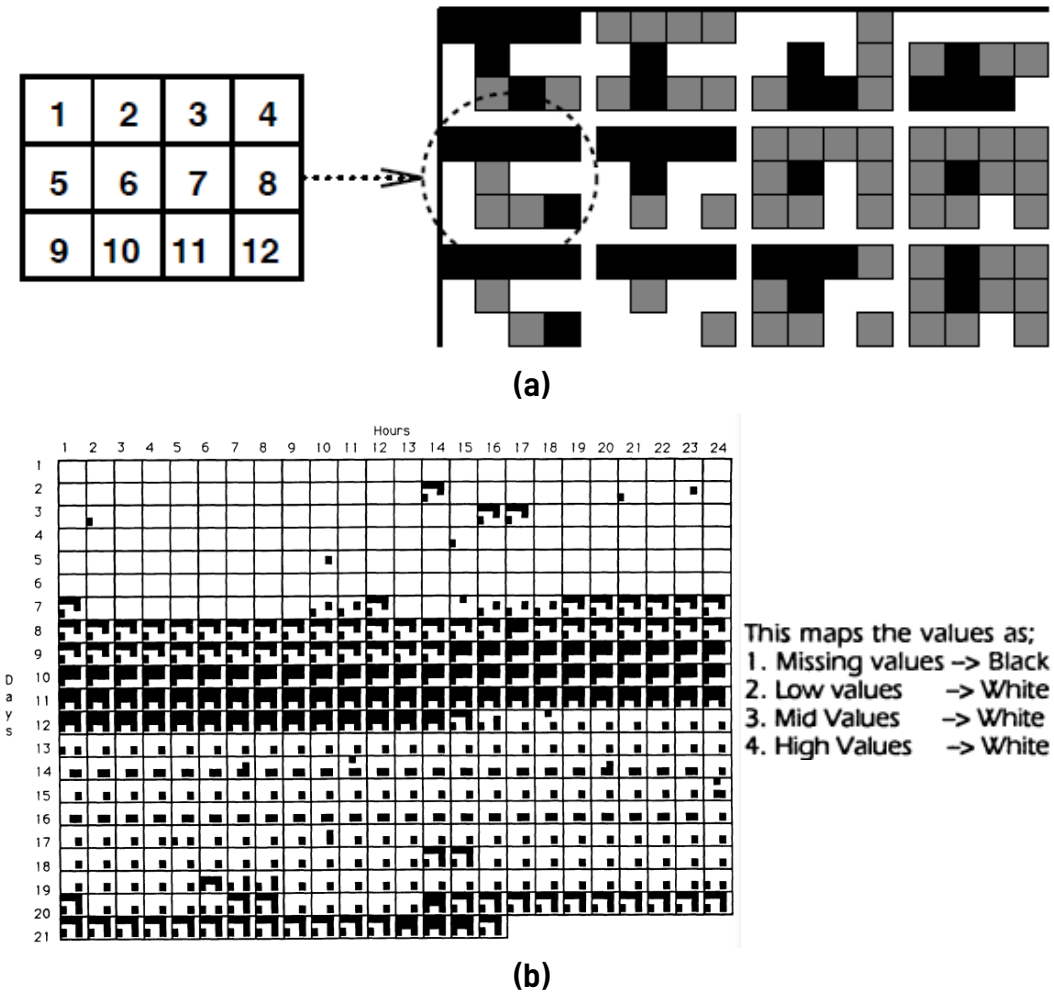


Figura 4.21: Codificación de Formas. (a) Matriz con doce atributos. Figura extraída de [WB94]. (b) Visualización de un conjunto de datos de la magnetosfera y del viento solar, con trece parámetros, recolectados cada hora durante 21 días en 1976. Figura extraída de [Bed90].

4.3.2.5. Codificación por Formas

Jeff Beddow [Bed90] presenta la técnica de codificación por formas, denominada *shape coding*, que utiliza pequeñas matrices de píxeles para representar información. Cada elemento de datos se visualiza a través de una de estas matrices, en la que los píxeles se mapean a una escala de colores según los valores de sus atributos.

Consideremos un conjunto de datos que consta de doce variables. En primer lugar, cada variable se normaliza y se clasifica en tres grupos según la función de la desviación estándar, además de un grupo que representa los valores faltantes. Los grupos de valores altos y bajos se representan con los colores negro y gris, respectivamente, mientras que los valores restantes se muestran en blanco. A continuación, cada variable se asigna a un cuadro dentro del glifo rectangular, como se muestra en la figura 4.21(a). Las matrices se disponen de manera secuencial, fila por fila, tal como se ilustra en la figura 4.21(b), donde los valores faltantes se representan en color negro y los demás valores en blanco. En [Bed90], el autor también ofrece ejemplos que utilizan colores primarios en lugar de negro, gris y blanco; sin embargo, no hemos logrado obtener la imagen en su versión a color.

4.3.2.6. Íconos de Color

Haim Levkowitz [Lev91] introduce un ícono de color (*color icon*), como un área en la pantalla que permite asignar atributos como color, forma, tamaño, orientación, límites y subdivisiones de área. La figura 4.22(a) muestra un ícono de color en forma de cuadrado que mapea hasta seis variables. Cada variable se asigna a una de las seis líneas gruesas, mientras que las líneas delgadas sirven como límites para separar los íconos vecinos.

Existen dos métodos distintos para colorear un ícono. En el primer enfoque, se asigna un color específico a cada línea gruesa según el valor del atributo de mapeo. Luego, la imagen del ícono en color se genera interpolando los colores asignados a todas las líneas gruesas. El segundo enfoque implica asignar un color diferente a cada subárea según los valores de los atributos de mapeo. Mientras que la primera opción ofrece una mejor combinación de parámetros, ya que al interpolar los colores de las líneas gruesas se logra una transición suave entre los valores de los atributos, lo que facilita la percepción de cómo estos varían a lo largo de la figura; la segunda opción, por su parte, proporciona una mejor separación de los atributos.

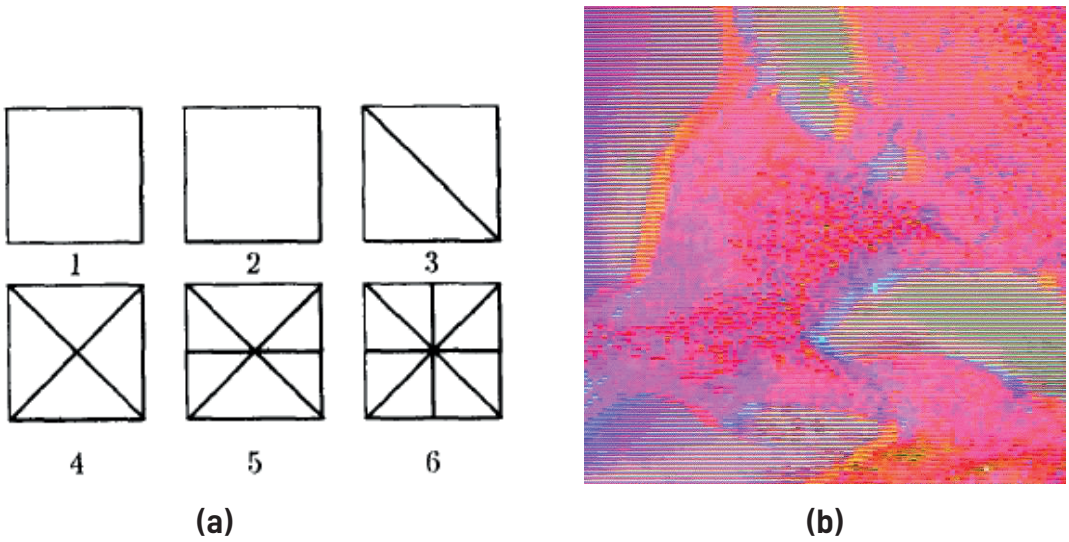


Figura 4.22: Íconos de color. (a) Mapeo de datos de una a seis variables (líneas más gruesas). Figura extraída de [Lev91]. (b) Ejemplo de visualización. Figura extraída de [WB94].

El número de variables mapeadas puede aumentarse utilizando íconos de diferentes formas, como hexagonales, en lugar de íconos cuadrados, o haciendo que cada variable controle uno de los valores HSV[8] (*Hue-Saturation-Value*). Esto significa que cada variable influirá en una componente distinta del color del ícono: una en el matiz, otra en la saturación, y la tercera en el valor. Un ejemplo de visualización de íconos de color se muestra en la figura 4.22(b).

4.3.3. Técnicas Basadas en Píxeles

Las técnicas orientadas a píxeles [Kei96] codifican cada valor de los datos mediante un píxel, lo que permite aprovechar al máximo el espacio de la pantalla, permitiendo mostrar conjuntos de datos con millones de valores en una sola pantalla. La idea central de estas técnicas radica en asignar cada valor de datos a un píxel coloreado, distribuyendo los valores correspondientes a distintas dimensiones o atributos en subventanas independientes.

La mayoría de estas técnicas segmentan la pantalla en áreas adyacentes, de manera que, para conjuntos de datos n-dimensionales, se generan n subventanas, una para cada dimensión. Los píxeles correspondientes a una observación, aparecen en la misma posición

[8]HSV (Hue, Saturation, Value): es un modelo de color que representa colores en términos de tres componentes: Matiz (*Hue*), que indica el tipo de color; Saturación (*Saturation*), que refleja la intensidad o pureza del color; y Valor (*Value*), que representa el brillo del color.

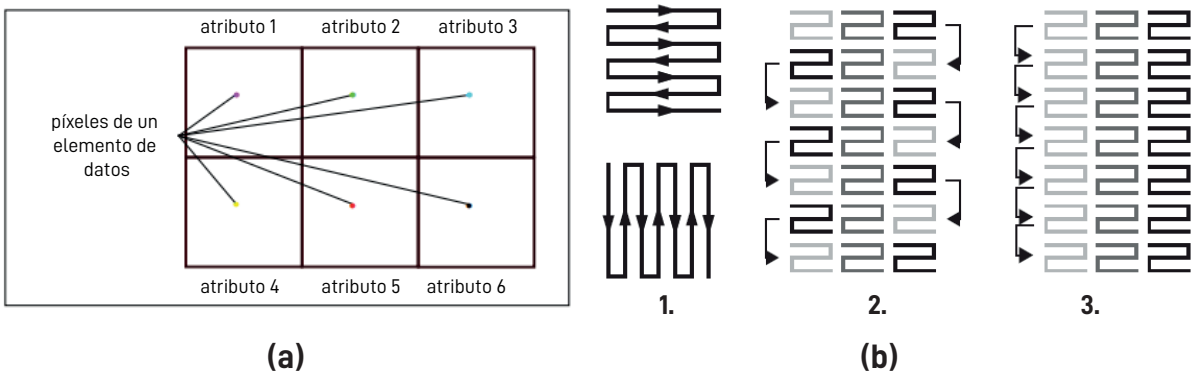


Figura 4.23: (a) Ubicación de un elemento de dato con seis atributos en técnicas basadas en píxeles. Figura adaptada de [Kei96]. (b) Organización de píxeles en técnica de patrón recursivo. Figura extraída de [KKA95].

en cada ventana (ver figura 4.23(a)). Con una disposición adecuada, el usuario puede identificar relaciones entre los atributos, así como detectar tendencias y patrones en los datos subyacentes.

Al emplear técnicas orientadas a píxeles, es fundamental considerar varios aspectos clave de diseño, como el mapeo de los valores de datos a colores, la disposición de los píxeles dentro de las subventanas y el ordenamiento de las subventanas en la pantalla. Las técnicas basadas en píxeles se dividen en técnicas independientes de la consulta como las matrices de correlación [RtBS84] o la técnica de patrón recursivo [KKA95], las cuales visualizan directamente los datos (o un subconjunto de ellos), y técnicas dependientes de la consulta como las técnicas de espiral generalizada [Kei95] y de sectores circulares [AKK96], que visualizan la relevancia de los elementos de datos con respecto a una consulta.

4.3.3.1. Técnicas Independientes de la Consulta

Como se mencionó anteriormente, las técnicas de visualización independientes de la consulta representan directamente los datos o un subconjunto de ellos. Uno de los principales desafíos al aplicar estas técnicas es encontrar disposiciones significativas para los píxeles en la pantalla. Una opción es organizar los elementos de datos de manera lineal, ya sea de izquierda a derecha o de arriba hacia abajo en columnas. Otra alternativa es organizar los píxeles en pequeños grupos y disponer estos grupos para formar algún patrón global, como sugieren las visualizaciones de patrones recursivos [KKA95].

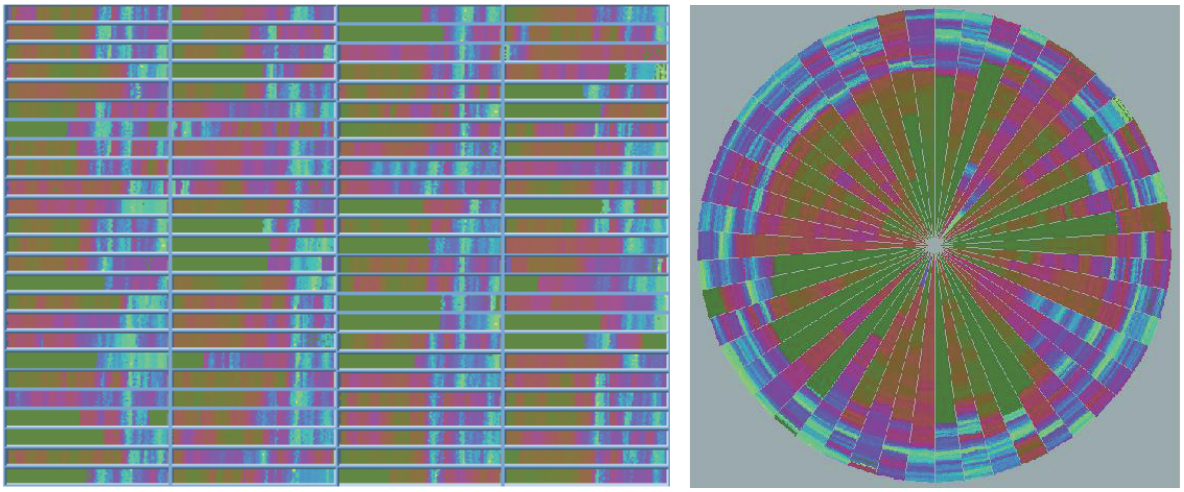


Figura 4.24: (a) Uso de la técnica de patrón recursivo para visualizar 532.900 valores de datos del índice FAZ desde enero de 1974 hasta abril de 1995. La configuración de parámetros elegida es $(w1, h1) = (1, 22)$ y $(w2, h2) = (243, 1)$, y el mapeo de colores asigna colores claros a los valores altos y colores oscuros a los valores bajos. El espacio vacío debido a la falta de datos se llena con el color de fondo. Figura extraída de [KKA95]. (b) Visualización de sectores circulares que representa 50 precios de acciones de la bolsa de valores de Frankfurt (índice FAZ) durante un período de 10 años, lo que da como resultado alrededor de 265.000 valores de datos. Figura extraída de [AKK96].

La técnica de patrón recursivo o *recursive pattern*, se basa en un esquema genérico de disposición recursiva que permite al usuario controlar la disposición de los píxeles [KKA95]. Esta técnica visualiza cada atributo en una subventana separada. Dentro de una subventana, cada valor del atributo está representado por un píxel coloreado, donde el color refleja el valor del atributo. El elemento base recursivo es un patrón de altura $h1$ y ancho $w1$, según lo especificado por el usuario. Primero, los elementos del patrón corresponden a píxeles individuales que se organizan dentro de un rectángulo de altura $h1$ y ancho $w1$ de izquierda a derecha, luego hacia abajo, de derecha a izquierda, luego nuevamente de izquierda a derecha, y así sucesivamente (ver figura 4.23(b1)). El mismo arreglo básico se utiliza en todos los niveles de recursión, con la única diferencia de que los elementos básicos que se organizan en el nivel i son los patrones resultantes de los arreglos del nivel (*i-1*) (ver figura 4.23(b2) para $(w1, h1) = (3, 3)$ y $(w2, h2) = (3, 7)$, y figura 4.23(b3) para $(w1, h1) = (3, 3)$, $(w2, h2) = (3, 1)$ y $(w3, h3) = (1, 7)$).

Como se mencionó anteriormente, esta técnica emplea un algoritmo genérico para

la organización de los píxeles, lo que otorga al usuario control sobre su disposición. Al ajustar la altura y el ancho en cada nivel de recursión, los usuarios pueden personalizar las visualizaciones según sus necesidades específicas. En la figura 4.24(a) se presenta una visualización de patrón recursivo de los precios diarios de acciones de 100 empresas a lo largo de 20 años.

La matriz de correlación [RtBS84] es otra técnica ampliamente reconocida basada en píxeles. Esta matriz puede representarse de diversas formas, como números, píxeles sombreados o puntos. Todos estos enfoques reflejan tanto el signo como la magnitud del valor de correlación. Por lo general, se utiliza un esquema de color de dos tonalidades para codificar el signo de la correlación, y la intensidad del color se ajusta en proporción a la magnitud de la correlación.

Otro ejemplo de este grupo de técnicas son los mapas auto-organizados, conocidos como *Self-Organizing Maps* (SOM). Los mapas auto-organizados son una clase de redes neuronales artificiales que se entrenan utilizando técnicas de aprendizaje no supervisado y permiten visualizar datos multidimensionales en un espacio bidimensional. Similar a otros métodos de reducción dimensional, el objetivo de los SOM es preservar las propiedades topológicas de los datos originales. Este modelo, propuesto en 1982 por el profesor finlandés Teuvo Kohonen [Koh90], se distingue por emplear un enfoque de aprendizaje competitivo. En el aprendizaje competitivo las neuronas compiten unas con otras con el fin de llevar a cabo una tarea dada. Se pretende que cuando se presente a la red una muestra de entrada, sólo una de las neuronas de salida (o un grupo de vecinas) se active.

Los mapas auto-organizados pueden clasificarse como una técnica orientada a píxeles porque cada neurona en la capa de salida se visualiza como un píxel que codifica información mediante variaciones de color o intensidad, siguiendo el principio fundamental de las técnicas basadas en píxeles. Aunque la generación del mapa incluye procesos complejos, la representación visual resultante utiliza una codificación pixel-valor para transmitir la información, alineándose con las características distintivas de este grupo de técnicas.

Un modelo SOM consta de dos capas principales: una capa de entrada, que recibe la información del entorno y la transmite a la capa de salida, y una capa de salida, compuesta por p neuronas, responsables de procesar los datos y generar el mapa. La cantidad de neuronas (p) se fija previamente, y éstas pueden estar organizadas en una disposición cuadrada o hexagonal en el espacio de salida. Las figuras 4.25(a) y (b) ilustran ambas

configuraciones topológicas.

El algoritmo SOM se divide en dos fases: la fase competitiva y la fase cooperativa. En la figura 4.25(c) se ilustra el algoritmo. En la fase competitiva, cada neurona de la capa de salida j se asocia con un vector de pesos $W_j = (W_{j1}, W_{j2}, ..., W_{jn})$, cuya dimensionalidad coincide con la del espacio de entrada. Posteriormente, se calculan las distancias entre cada neurona de la capa de salida y las observaciones del conjunto de entrada. La neurona de salida con la menor distancia euclidiana será designada como el "nodo ganador". En la segunda fase, se ajustan los pesos tanto de la neurona ganadora como de sus neuronas vecinas en la red, para que las neuronas se acerquen a la observación de entrada. Antes de actualizar los pesos, hay que determinar qué neuronas pertenecen a la vecindad de la neurona ganadora. Para ello, primero hay que calcular el tamaño del radio de vecindad centrado en la neurona ganadora y determinar qué neuronas están dentro de dicho radio. Entre las neuronas de la capa de salida, puede decirse que existen conexiones laterales de excitación e inhibición implícitas, pues aunque no estén conectadas, cada una de estas neuronas va a tener cierta influencia sobre sus vecinas.

En la figura 4.26 se muestra el resultado de la agrupación del conjunto de datos *Iris* [Fis88] utilizando 2D-SOM y 3D-SOM. En el 3D-SOM, el eje z representa cinco capas diferentes, donde cada capa tiene 25 neuronas de salida.

4.3.3.2. Técnicas Dependientes de la Consulta

Las técnicas de visualización dependientes de la consulta se centran en representar los datos en el contexto de una consulta específica del usuario, con el fin de proporcionar realimentación y orientar su búsqueda. Como se mencionó anteriormente, estas técnicas asignan colores a los píxeles basándose en las distancias entre los valores de los atributos y la consulta. El objetivo principal de estas técnicas es resaltar la importancia de los elementos de datos en relación con la consulta realizada, por lo que se utilizan diversos métodos de disposición de píxeles. Estas técnicas sitúan los elementos de datos más relevantes para una consulta en el centro de la visualización.

En particular, abordaremos las técnicas de espiral generalizada [Kei95] y de sectores circulares [AKK96], que son generalizaciones de la técnica propuesta por Keim [KK94a, Kei94]. Este enfoque sitúa los elementos de datos más relevantes (aquellos que cumplen con la consulta) en el centro de la ventana, mientras que los elementos de datos menos

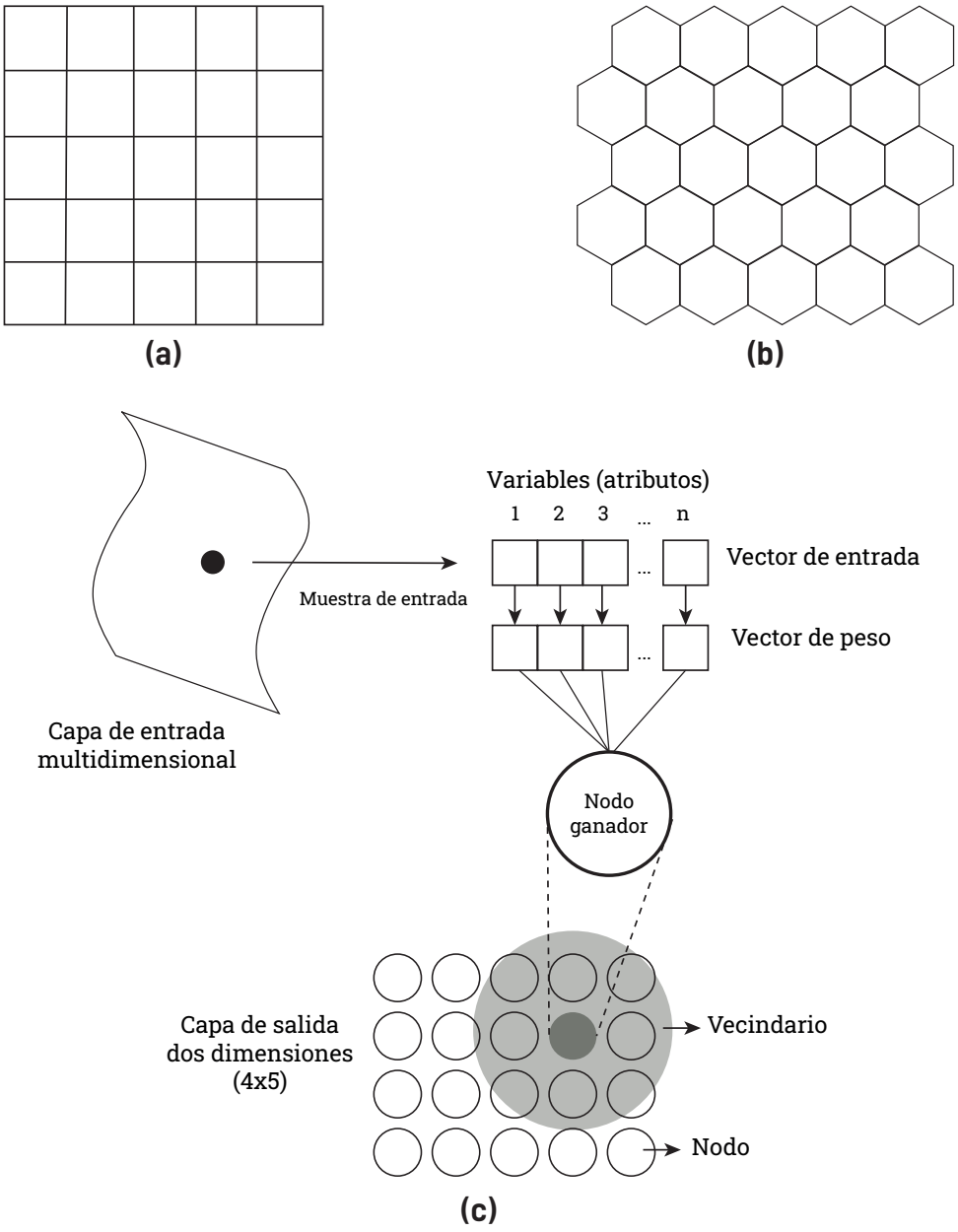


Figura 4.25: Mapas auto-organizados (SOM). Configuración topológica (a) rectangular y (b) hexagonal. (c) Ilustración gráfica del algoritmo. Figuras adaptadas de [AE12].

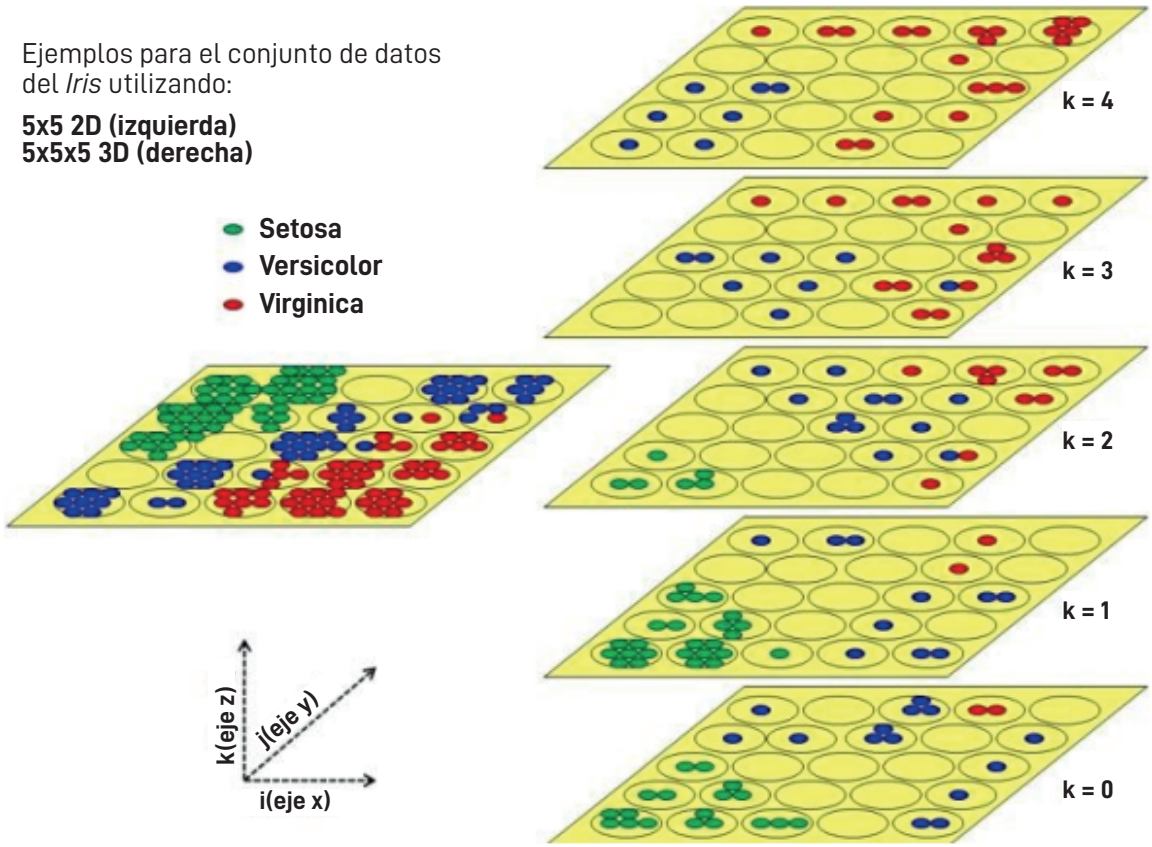


Figura 4.26: Clustering del conjunto de datos *Iris* [Fis88] utilizando SOM en dos dimensiones (izquierda) y tres dimensiones (derecha). Figura adaptada de [MZKMY12].

relevantes se disponen en forma de espiral rectangular hacia el exterior de la ventana (ver figura 4.27(a)). En el caso de la técnica de espiral generalizada, la disposición espiral original [KK94a] se extiende a una forma genérica parecida a una serpiente (ver figura 4.27(b)) o a una curva de Peano-Hilbert o Morton, de la cual el usuario puede elegir la altura (ver figuras 4.27(c) y 4.27(d)). Al igual que en el caso de las técnicas de visualización independientes de la consulta, se genera una visualización separada para cada uno de los atributos, y los píxeles para cada elemento de datos se colocan en la misma posición. Una ventana adicional muestra las distancias generales.

La técnica de sectores circulares [AKK96], como su nombre lo indica, utiliza sectores de un círculo para representar las dimensiones de los datos. La idea fundamental de esta técnica es mostrar las dimensiones de los datos como sectores de un círculo. Para datos de n dimensiones, el círculo se divide en n sectores, cada uno representando una dimensión de los datos. La figura 4.27(e) ilustra esta división del círculo para un conjunto de datos de ocho dimensiones. Dentro de cada sector, los píxeles se disponen comenzando en el

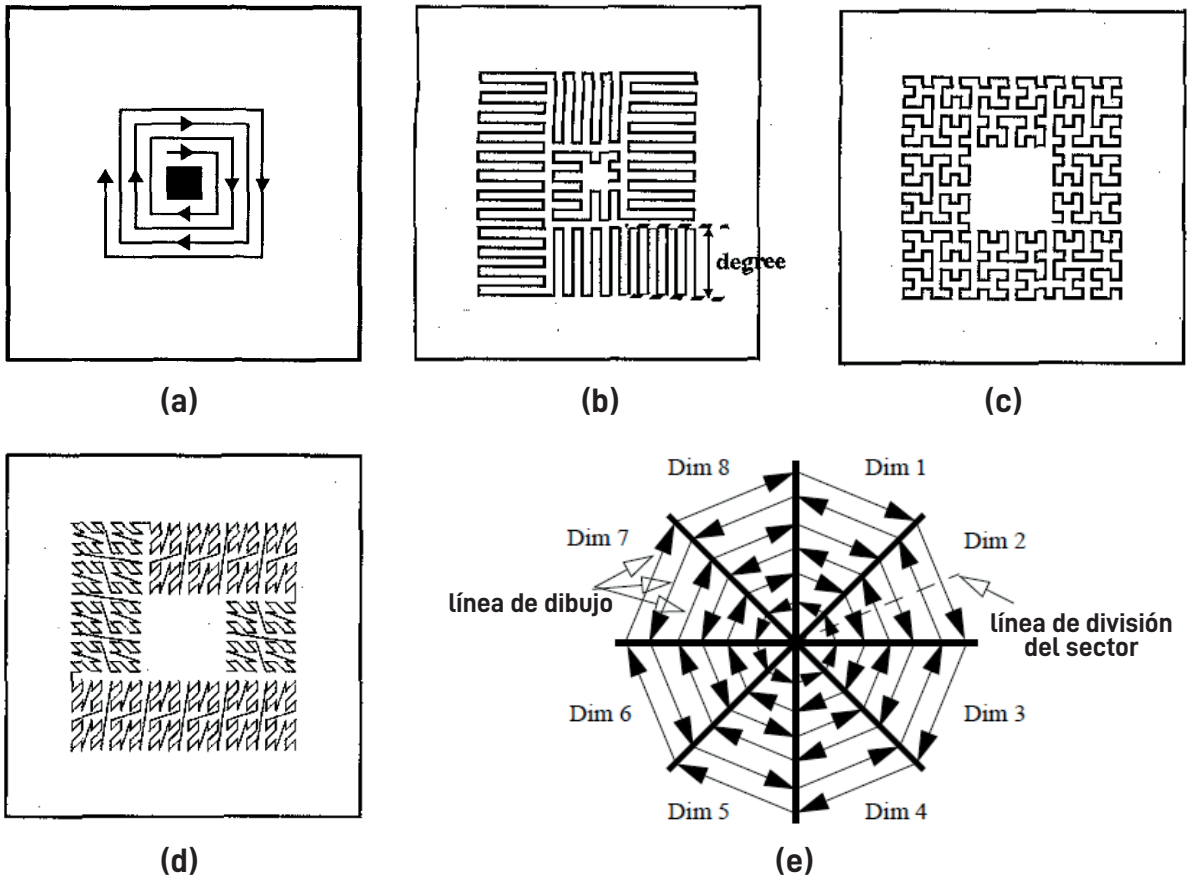


Figura 4.27: Técnicas basadas en píxeles dependientes de la consulta. (a) Técnica de espiral. Disposición en espiral generalizada de un atributo utilizando: (b) Técnica de espiral de serpiente, (c) Técnica de espiral de Peano-Hilbert y (d) Técnica de espiral de Morton. (e) Técnica de sectores circulares para datos de 8 dimensiones. Figuras extraídas de [KK96, Kei96].

centro del círculo y extendiéndose de manera alternada hacia el borde exterior del sector, en forma de línea por línea. Estas líneas de dibujo, sobre las que se disponen los píxeles, son ortogonales a las líneas de división del sector. Un ejemplo de esta técnica puede observarse en la figura 4.24(b).

4.3.4. Técnicas Jerárquicas

Las técnicas jerárquicas de visualización son métodos que representan datos multidimensionales a través de estructuras que reflejan relaciones de dependencia o subordinación entre los elementos. Estas técnicas dividen el espacio de datos n-dimensional en subespacios, presentándolos en diferentes niveles jerárquicos.

Un árbol es la representación más básica de una estructura jerárquica [Lim14]. Como

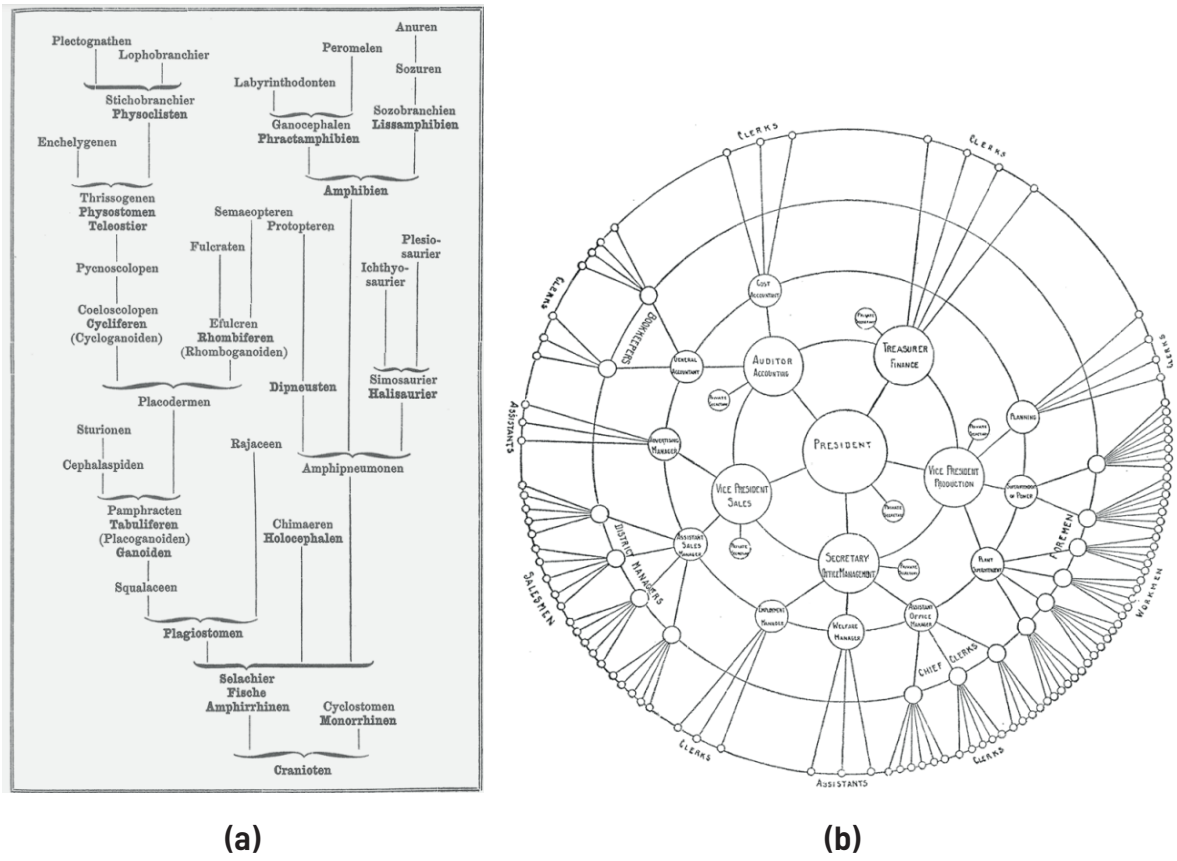


Figura 4.28: (a) Ejemplo de árbol vertical: árbol genealógico de vertebrados amniotas, animales con cuatro extremidades y columna vertebral. Este árbol vertical muestra el estilo gráfico entre llaves. (b) Ejemplo de árbol circular o radial: el árbol destaca el proceso de toma de decisiones centralizado en la mayoría de las empresas, con el presidente en el núcleo del gráfico, seguido por niveles sucesivos de gerentes y trabajadores representados por anillos concéntricos secuenciales. Figuras extraídas de [Lim14].

su nombre lo indica, esta estructura se asemeja a un árbol, donde hay un nodo raíz que puede conectarse a múltiples nodos mediante ramas y estos nodos pueden a su vez ramificarse hacia otros nodos. Los nodos finales (aquellos sin ramificaciones) se llaman hojas. En la figura 4.28(a) se presenta un ejemplo de árbol genealógico de vertebrados amniotas. A medida que un árbol aumenta su número de relaciones, puede expandirse hacia afuera de forma radial, dando lugar a un árbol circular (ver figura 4.28(b)).

4.3.4.1. Apilamiento Dimensional

LeBlanc *et al.* [LWW90] presentaron una técnica llamada apilamiento dimensional o *dimensional stacking*, que subdivide el espacio n-dimensional en subespacios 2D que se

apilan entre sí. La figura 4.29(b) ilustra un ejemplo de una visualización de apilamiento dimensional. La idea básica consiste en incrustar un sistema de coordenadas dentro de otro sistema de coordenadas; es decir, dos atributos forman el sistema de coordenadas exterior, dos atributos adicionales se incrustan en el sistema exterior, y así sucesivamente.

El algoritmo consiste en primer lugar seleccionar el par de atributos más importantes g_1 y g_1, y se define una cuadrícula 2D de g_1 versus g_2 (ver figura 4.29(a)). Los atributos más importantes deben ser elegidos para los niveles externos del apilamiento. Luego, se lleva a cabo una subdivisión recursiva de cada celda de la cuadrícula usando el siguiente par de parámetros más importantes. Finalmente, se colorean las celdas de la cuadrícula final.

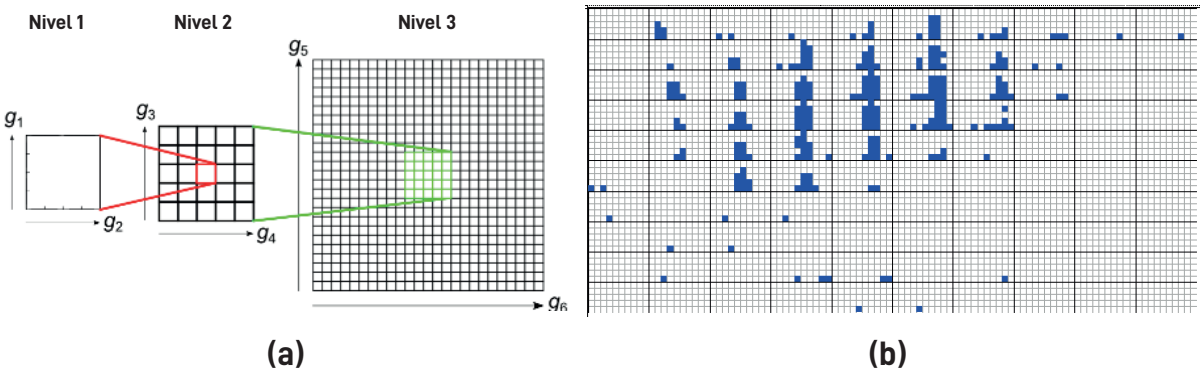


Figura 4.29: (a) Particionamiento de apilamiento dimensional. Figura adaptada de [GBRQ14]. (b) Ejemplo de visualización mediante apilamiento dimensional de un conjunto de datos de perforación, compuesto por tres dimensiones espaciales y una cuarta dimensión que representa la ley del mineral. Figura extraída de [WGK10].

4.3.4.2. Treemap

Johnson y Shneiderman introdujeron el *treemap*, una técnica de visualización que mapea la estructura jerárquica a rectángulos anidados [JS91, Shn92].

Un *treemap* se construye mediante subdivisiones recursivas, es decir, un nodo se divide en varios rectángulos según el tamaño de los hijos de este nodo (ver figura 4.30). La dirección de la subdivisión alterna, un rectángulo se divide en una dirección (por ejemplo, horizontalmente) y para el siguiente nivel esta dirección se alterna. Los *treemaps* proporcionan una representación visual compacta de datos jerárquicos complejos.

En [STLD20] se detallan otras alternativas al *treemap* tradicional.

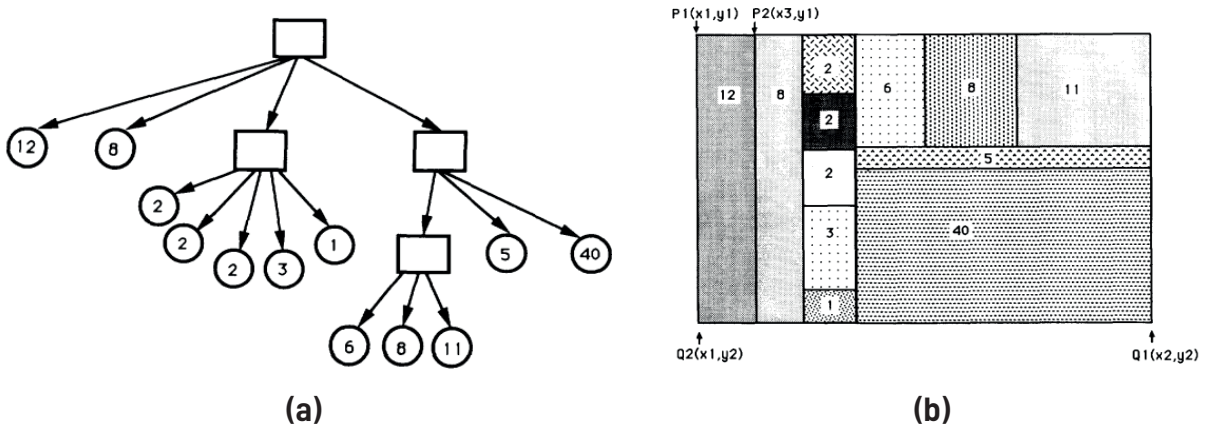


Figura 4.30: (a) Estructura de árbol tradicional de 3 niveles. Los números indican el tamaño de cada nodo hoja. (b) *Treemap* de la estructura definida en (a). Imágenes extraídas de [Shn92].

4.3.4.3. Dendrograma

El clustering jerárquico es una técnica empleada para dividir un conjunto de datos en grupos, organizando los elementos en una estructura jerárquica o "árbol" de clústeres. Se han desarrollado varios algoritmos de clustering jerárquico [HKP12] para abordar distintos tipos de datos y requisitos de aplicación. Según la forma en que generan los clústeres, estos algoritmos se dividen en dos categorías: el clustering jerárquico aglomerativo, que sigue un enfoque ascendente fusionando progresivamente clústeres más pequeños en otros más grandes, y el clustering jerárquico divisivo, que adopta un enfoque descendente dividiendo iterativamente un clúster inicial en grupos más pequeños.

El clustering jerárquico suele representarse gráficamente mediante un diagrama en forma de árbol llamado dendrograma o *dendrogram* (ver figura 4.31). El dendrograma es una técnica de visualización jerárquica utilizada para representar la estructura de agrupamiento de datos en forma de un árbol. Cada nodo del dendrograma representa un grupo de elementos, y las uniones entre nodos indican la similitud o proximidad entre los grupos. La altura de las uniones en el dendrograma refleja el nivel de similitud en el que se fusionan los grupos; cuanto más alta es la unión, menor es la similitud entre los grupos. Al cortar el dendrograma en un nivel deseado, se obtiene una agrupación de los elementos de datos en grupos disjuntos.

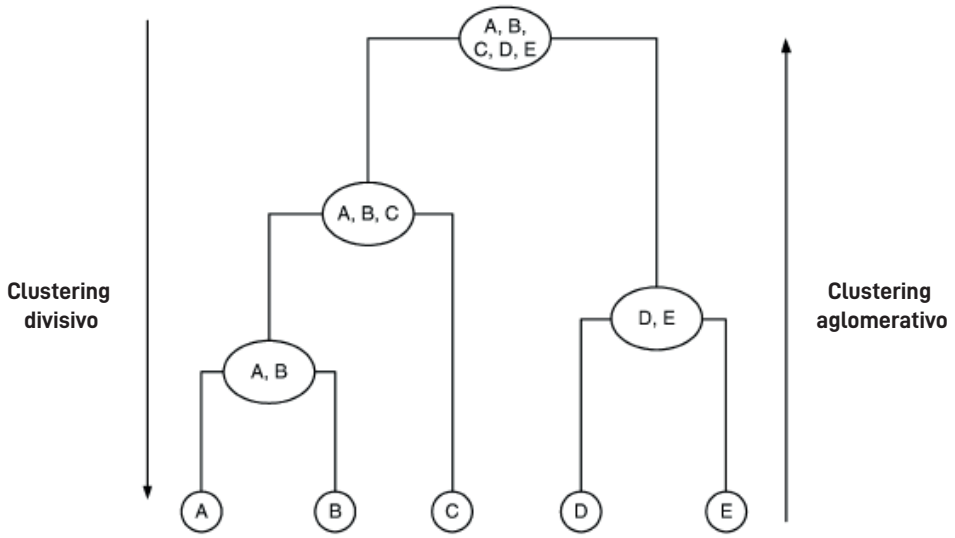


Figura 4.31: Representación de clustering jerárquico mediante un dendrograma. Figura adaptada de [Hal18].

4.3.5. Técnicas Basadas en Grafos

Las técnicas basadas en grafos se dividen en tres categorías principales: representaciones de nodos y enlaces, representaciones de matrices y representaciones implícitas [TS20].

Los diagramas de nodos y enlaces visualizan los nodos como puntos y los enlaces entre ellos como líneas o arcos. Un aspecto crucial en este tipo de representaciones, es determinar la ubicación de los nodos y los enlaces. Según el grado de libertad que tenga un algoritmo para posicionar los nodos, se pueden identificar tres tipos de diseño: diseño libre, fijo y estilizado [SS06]. En un diseño libre, no se imponen restricciones específicas sobre la ubicación de los nodos, y las representaciones suelen crearse mediante algoritmos de diseño dirigidos por fuerzas [FR91]. En cambio, en los diseños fijos, los nodos se ubican en posiciones previamente definidas. Entre los diseños libres y los fijos, se encuentran los llamados diseños estilizados, donde las posiciones de los nodos están limitadas por un esquema predefinido. Un ejemplo común es cuando los nodos deben ubicarse en un círculo o alinearse a lo largo de un eje. Los diagramas de arco [Wat02] son una representación típica de este tipo de disposición.

Por su parte, las representaciones de matrices ofrecen una interpretación visual de la matriz de adyacencia de un grafo. Cada nodo está representado por una fila y una columna. Si existe un enlace entre el nodo i y el nodo j, se marca la celda correspondiente en esa posición de la matriz. La ventaja principal de este tipo de representación es que

Figura 4.32: (a) Diseño dirigido por fuerzas de las palabras más importantes en la novela *Moby Dick*. El tamaño de los nodos indica la frecuencia de las palabras. Los enlaces conectan palabras que comúnmente se encuentran cerca unas de otras. Se ha ampliado el clúster alrededor de la palabra "*whale*" mediante un lente focal. Figura extraída de [Wil09]. (b) Red de co-ocurrencia de personajes en los capítulos de la novela *Les Misérables* de Víctor Hugo, representado con una vista de matriz. Figura extraída de [HBO10].

ofrece un mapeo visual directo de los datos del grafo, centrándose en los enlaces, mientras que los nodos se representan únicamente como etiquetas. Esta técnica se detalla en la sección 4.3.5.3.

Las representaciones de nodo-enlace y de matrices, representan los enlaces mediante primitivas gráficas, como celdas o líneas. Por este motivo, se les puede denominar representaciones explícitas. En contraste, las representaciones implícitas no muestran los enlaces de manera directa; en su lugar, las relaciones entre nodos se representan implícitamente a través de su posición relativa. Estas representaciones solo son viables cuando la estructura del grafo posee ciertas características. Por ejemplo, son particularmente efectivas en el caso de árboles, dada la relación jerárquica entre padres e hijos. En [SHS10] se muestran ejemplos de representaciones implícitas.

4.3.5.1. Diseños Dirigidos por Fuerzas

La técnica de visualización de grafos conocida como *force-directed layouts* o diseños dirigidos por fuerzas [FR91], corresponde a las representaciones de nodos y enlaces con diseño libre. Esta técnica se basa en la simulación de un sistema físico en el que los nodos del grafo se comportan como partículas cargadas que se repelen mutuamente, mientras que los enlaces actúan como resortes que los atraen y conectan. El objetivo del algoritmo es alcanzar un estado de equilibrio, minimizando la energía del sistema, lo que genera una disposición de los nodos que resulta visualmente clara y comprensible.

Este enfoque es especialmente eficaz para revelar patrones y estructuras latentes en los datos, ya que tiende a agrupar de forma natural los nodos más interconectados y a distanciar los menos relacionados. Esto facilita la identificación de clústeres y permite una mejor comprensión de las relaciones entre los elementos del grafo. En la figura 4.32(a), se utiliza un diseño dirigido por fuerzas para visualizar las palabras más importantes en la novela *Moby Dick*. El tamaño de los nodos indica la frecuencia de las palabras. Los enlaces conectan palabras que comúnmente se encuentran cerca unas de otras.

4.3.5.2. Diagramas de Arco

Los diagramas de arco [Wat02], conocidos en inglés como *arc diagrams*, son una técnica de visualización de grafos de la categoría nodo-enlace con diseño estilizado, donde todos los nodos se disponen en una línea recta, generalmente horizontal. Las conexiones entre los nodos se muestran como arcos semicirculares por encima o por debajo de esta línea, tal como se ilustra en la figura 4.33. La altura de cada arco suele ser proporcional a la distancia entre los nodos que conecta, creando una representación visualmente atractiva y fácil de interpretar. Esta disposición es particularmente efectiva para visualizar patrones en datos secuenciales o para resaltar conexiones en estructuras lineales como textos o líneas temporales. Su principal desventaja es que pueden volverse complejos con grafos densamente conectados.

4.3.5.3. Vistas de Matrices

Las vistas de matrices, o *matrix views*, son una técnica de visualización de grafos que convierte las relaciones entre nodos en una representación tabular [SM07]. En esta visualización, los nodos del grafo se disponen tanto en las filas como en las columnas

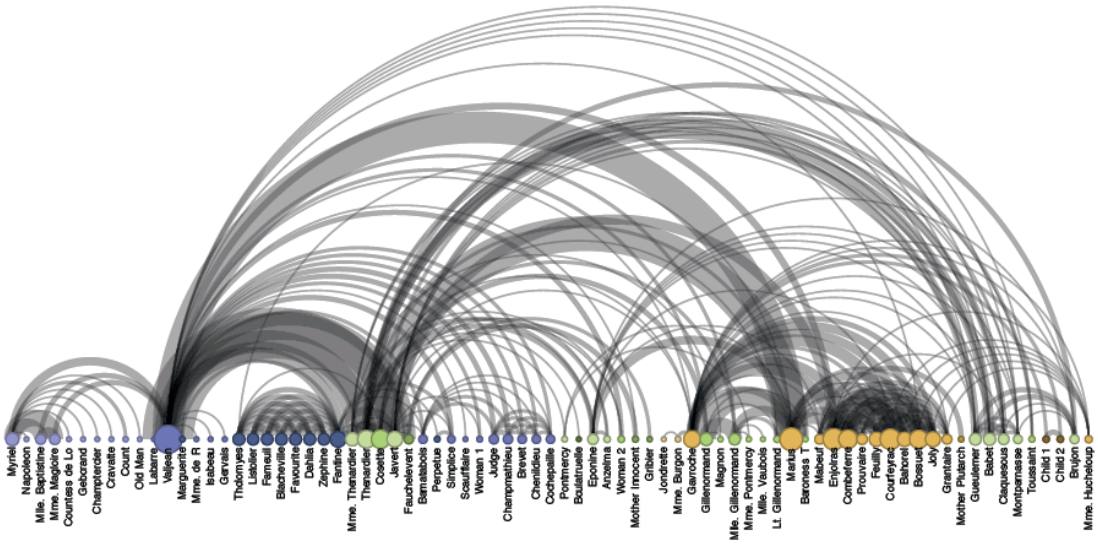


Figura 4.33: Red de co-ocurrencia de personajes en los capítulos de la novela *Les Misérables* de Víctor Hugo, representado con un diagrama de arcos. Figura extraída de [HBO10].

de una matriz bidimensional. Las celdas de la matriz indican la presencia o ausencia de conexiones entre los nodos correspondientes, generalmente utilizando colores o valores numéricos para representar la fuerza o tipo de relación. Esta técnica resulta especialmente útil en grafos densos o redes complejas, ya que facilita una visualización compacta de todas las conexiones posibles. Las vistas de matrices son particularmente efectivas para identificar clústeres dentro de la red y para comparar la conectividad entre distintas regiones del grafo. En la figura 4.32(b), un ordenamiento eficiente de las filas y columnas permite identificar clústeres, en la red de co-ocurrencia de personajes en los capítulos de la novela *Les Misérables* de Víctor Hugo.

4.4. Conclusiones

En este capítulo se han explorado en profundidad los fundamentos de la representación visual de datos, centrándose en el tercer pilar del proceso de visualización: el "*¿Cómo?*".

Se han examinado en detalle los componentes esenciales que constituyen la estructura visual, abarcando desde las marcas y canales visuales hasta el sustrato espacial. Además, se ha propuesto una taxonomía de las técnicas de visualización multidimensional más utilizadas, clasificándolas en enfoques basados en geometría, íconos, píxeles, jerarquías y grafos.

Este capítulo completa el análisis de los tres pilares fundamentales —"*¿Qué visualizar?*", "*¿Por qué visualizar?*" y "*¿Cómo visualizar?*"— que orientan el proceso de visualización de datos, proporcionando una perspectiva integral para el diseño y la implementación de representaciones visuales efectivas y expresivas.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 5

Sistemas de Recomendación de Técnicas de Visualización

5.1. Introducción

En la actualidad, la generación masiva y continua de datos ha fomentado un notable crecimiento en el uso de técnicas y herramientas de visualización. Sin lugar a dudas, la visualización de datos se ha consolidado como una herramienta indispensable, permitiéndonos explorar, interpretar y comunicar información de manera efectiva y expresiva.

Lograr una representación adecuada de los datos implica comprender su naturaleza y sus características, y el propósito de la visualización. ¿Qué queremos lograr con nuestro gráfico? ¿Queremos comparar datos, identificar patrones o relaciones, o ambas cosas? ¿Qué aspectos son más relevantes para nuestro análisis? ¿Qué dimensiones y características tienen los datos que estamos tratando de visualizar? ¿Cómo se distribuyen los datos?. Además, es fundamental tomar decisiones sobre la representación visual, lo que implica transformar los datos en elementos gráficos, comprender y mapear sus atributos gráficos, y seleccionar las técnicas de visualización más adecuadas.

Con el incremento en la cantidad y variedad de técnicas de visualización, la tarea de seleccionar la visualización más adecuada se ha vuelto cada vez más desafiante. Por ejemplo, en la librería *D3.js* (*Data-Driven Documents*) [BOH11] se ofrecen más de 300 visualizaciones distintas. Evidentemente elegir la visualización adecuada es un desafío considerable.

En este contexto, han surgido los sistemas de recomendación de técnicas de visua-

lización, herramientas que asisten a usuarios no expertos en la toma de decisiones al ofrecer sugerencias y orientación sobre qué técnicas de visualización serían más efectivas para sus datos y objetivos. Definimos a un "usuario no experto" como aquél que carece de conocimientos profesionales o especializados en visualización de datos.

5.2. Trabajo Relacionado

Los sistemas de recomendación de técnicas de visualización han surgido para contrarrestar la creciente complejidad y el aumento exponencial de los datos, en un contexto en el que la capacidad humana de análisis es limitada y cada vez más profesionales de diversas áreas deben interpretar datos complejos. En este contexto, los sistemas de recomendación de técnicas de visualización juegan un papel fundamental al actuar como un puente entre los datos complejos y los usuarios que necesitan interpretarlos, sin importar su nivel de experiencia, con el objetivo de optimizar el proceso de análisis y visualización.

Con el transcurso de los años, los investigadores han expandido el conjunto de requisitos que los sistemas deben abordar, con el objetivo de crear visualizaciones más efectivas y mejor alineadas con las necesidades específicas de los usuarios. Según Vartak *et al.* [VHS+17] los sistemas de recomendación deben considerar una combinación de cinco ejes o factores de recomendación:

1. *Características de los Datos.* El objetivo de un sistema de recomendación de técnicas de visualización es facilitar la exploración de valores, tendencias y patrones relevantes. Para ello, hay ciertas características de los datos que un sistema debe considerar [Was13], por ejemplo:

    - Resúmenes, como histogramas, que ofrecen una visión general del conjunto de datos y sus distribuciones;
    - Correlaciones, como el coeficiente de correlación de Pearson [Pea95], que revelan qué atributos están relacionados y en qué grado;
    - Patrones y tendencias, como la regresión, las reglas de asociación y agrupamiento;
    - Estadísticas avanzadas, como la suma de rangos de Wilcoxon [Wil92], que facilitan un análisis más detallado.

2. *Tarea o Insight*[1] *previsto.* Otro aspecto importante es el propósito u objetivo del usuario al realizar el análisis. Esto incluye los siguientes aspectos:

   - Estilo de análisis: por ejemplo, exploratorio, comparativo, predictivo o dirigido;
   - Objetos de análisis: subconjunto de datos y atributos de interés;
   - Objetivo del análisis: por ejemplo, explicar un comportamiento, comparar entre subconjuntos de datos o encontrar patrones atípicos.

3. *Semántica y Conocimiento del Dominio.* Como se mencionó en capítulos anteriores, los datos están enriquecidos con información semántica. Esta semántica abarca el tipo de datos almacenados, la información que proporciona cada atributo y las relaciones entre ellos. Además, un aspecto crucial, aunque más difícil de capturar, es el conocimiento especializado del usuario sobre el dominio, que orienta el análisis de los datos.

4. *Facilidad de Entendimiento Visual.* Un factor fundamental es garantizar que los datos se presenten de manera clara y comprensible para el usuario. Esto se refiere a la facilidad con la que un usuario puede interpretar la información representada en una visualización.

5. *Preferencia del Usuario.* Un aspecto clave es considerar una variedad de factores relacionados con las preferencias del usuario. El sistema puede examinar el historial de interacción, tanto de usuarios individuales como de grupos con datos similares, para identificar patrones de uso y comportamientos recurrentes. Los aportes explícitos del usuario, como configuraciones elegidas y el conocimiento específico del dominio, son igualmente esenciales para mejorar la precisión de las recomendaciones. Asimismo, es posible evaluar los atributos que el usuario suele visualizar, sugiriendo atributos relacionados o similares para enriquecer el análisis, así como identificar los tipos de visualización que el usuario prefiere.

De acuerdo a estos factores, Kaur y Owonibi [KO17] distinguen cuatro tipos de sistemas de recomendación de técnicas de visualización:

[1] *Insight* es un término utilizado en Psicología proveniente del inglés que se puede traducir al español como "visión interna" o más genéricamente "percepción" o "entendimiento". Mediante un *insight* el sujeto capta, internaliza o comprende, una verdad revelada.

- *Orientados a las Características de los Datos*: sistemas que sugieren visualizaciones en función de las propiedades de los datos, considerando factores como el tipo de atributos, la distribución y la cardinalidad del conjunto de datos, entre otros.

- *Orientados a la Tarea*: estos sistemas recomiendan visualizaciones según el objetivo o propósito específico del análisis.

- *Orientados al Conocimiento del Dominio*: optimizan el proceso de recomendación de visualizaciones al incorporar directamente en el sistema de visualización conocimiento especializado del usuario sobre el dominio.

- *Orientados a las Preferencias del Usuario*: estos sistemas recogen información sobre las preferencias y objetivos del usuario mediante interacciones explícitas con el sistema de visualización, para luego generar recomendaciones basadas en los datos obtenidos de dichas interacciones y selecciones.

5.2.1. Recomendadores Orientados a las Características de los Datos

Los sistemas de recomendación orientados a las características de los datos, recomiendan las técnicas de visualización más adecuadas según la naturaleza de los datos, considerando sus propiedades estadísticas, estructurales y semánticas. Estos sistemas evalúan factores como el tipo de atributos (ver sección 3.4), la distribución de los datos, la presencia de valores atípicos y las relaciones entre los atributos.

Uno de los primeros sistemas, denominado **BHARAT**, se basaba en un algoritmo simple que sugería gráficos de líneas, torta o barras dependiendo de las características de los datos [Gna81]. El sistema proponía el uso de gráficos de líneas para datos continuos, y gráficos de torta para representar datos que se pueden dividir en partes significativas de un todo. En todos los demás casos, se recomendaban gráficos de barras.

En 1986, Jock Mackinlay presentó el sistema **APT** (*A Presentation Tool*), siendo uno de los primeros en implementar el mapeo automático de las características de los datos a gráficos bidimensionales [Mac86]. Este sistema asigna visualmente los atributos de los datos a variables visuales como color, forma, tamaño, textura y orientación, y genera de manera sistemática una amplia gama de diseños mediante un álgebra de composición que

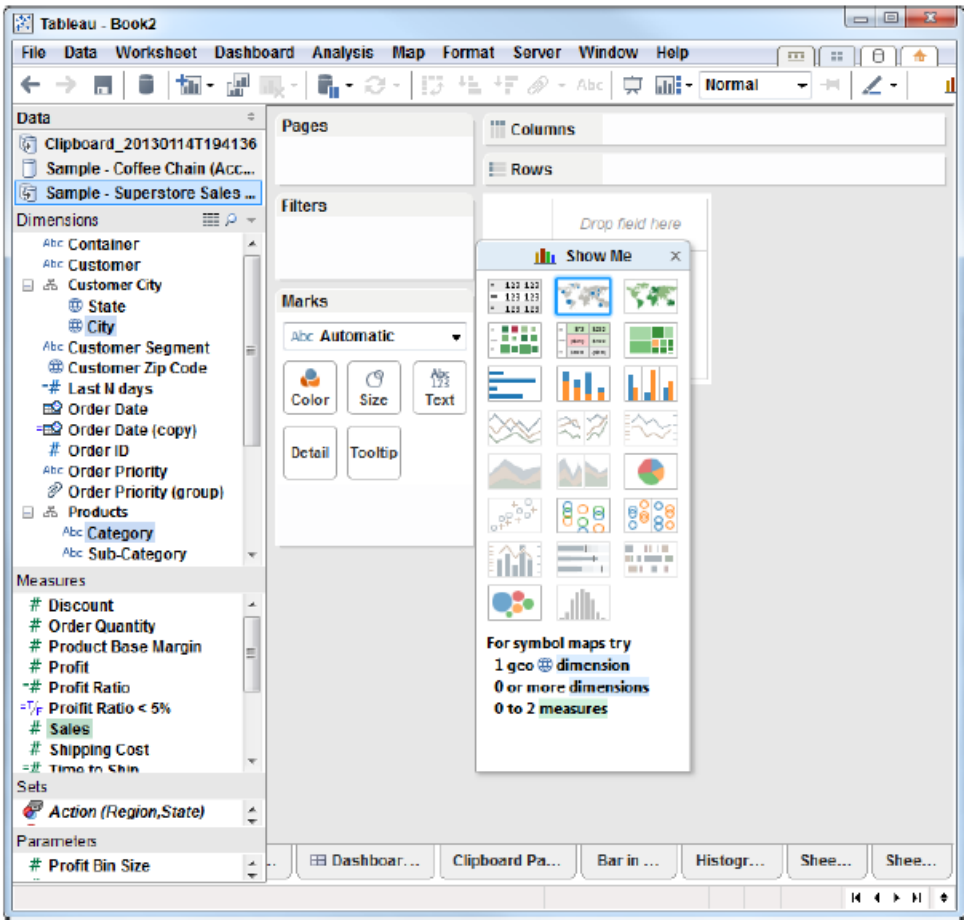


Figura 5.1: Interfaz de usuario de **Tableau Software**. Figura extraída de [Mur13].

combina un conjunto de lenguajes gráficos primitivos y operadores de composición. Las especificaciones de Mackinlay fueron posteriormente empleadas para desarrollar el sistema denominado **Polaris** [STH02], que más tarde se comercializó como **Tableau Software**[2] (ver figura 5.1). Estas especificaciones fueron refinadas más adelante en un lenguaje visual declarativo formal llamado **VizQL** [Han06]. **Tableau Software** introdujo una función llamada *Show Me* [MHS07]. *Show Me* utiliza especificaciones VizQL para recomendar automáticamente visualizaciones. Cuando el usuario selecciona los atributos de datos de su interés, el sistema determina la técnica de visualización más adecuada observando los tipos de atributos presentes en los datos, ya que cada visualización requiere atributos específicos.

Otro sistema destacado fue **Many Eyes**, que permitía a los usuarios crear visualizaciones de forma colaborativa al asociar un conjunto de datos con un componente de visualización, dependiendo de los atributos del mismo [VWvH+07]. **Many Eyes** ofrecía

[2]https://public.tableau.com/s/

más de una docena de componentes básicos de visualización, incluyendo gráficos de barras, *treemaps*, gráficos de burbujas y nubes de palabras. No todas las técnicas son adecuadas para visualizar el mismo tipo de datos. Por ejemplo, un *treemap* necesita varios datos textuales para definir la jerarquía, además de dos columnas cuantitativas que se asignan al tamaño y al color. En cambio, un diagrama de dispersión tradicional requiere dos columnas cuantitativas (una para cada eje), una columna cuantitativa opcional para especificar el tamaño de cada punto y una columna textual para las etiquetas. Para facilitar esta especificación, **Many Eyes** utilizaba un esquema de cinco tipos de "*slots*": no estructurados (U), numéricos (N), textuales (T), numéricos múltiples (N^+) y textuales múltiples (T^+). Estos tipos de *slots* expresan las necesidades de datos de cada componente de visualización. Por ejemplo, el esquema de un diagrama de dispersión se puede representar así: $\{Xaxis : N, Yaxis : N, label : T, [Dotsize : N]\}$. El conjunto de datos y la visualización producida pueden compartirse con otros usuarios para comentarios, realimentación y futuras mejoras, proporcionando así un banco de trabajo colaborativo para la creación de visualizaciones.

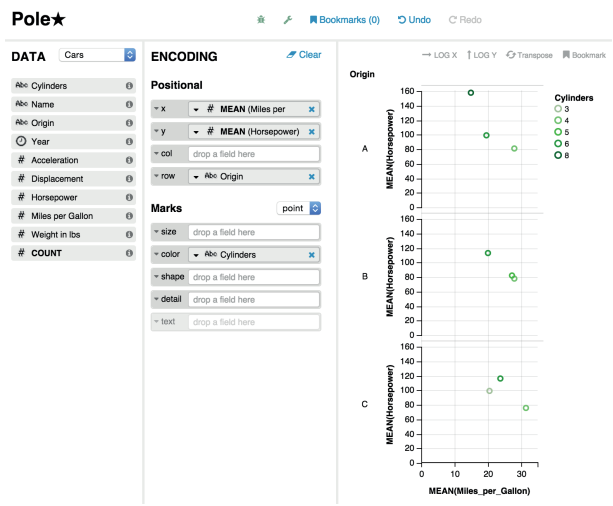
VizDeck es una herramienta web diseñada para la creación de dashboards, que sugiere visualizaciones basadas en las propiedades estadísticas del conjunto de datos [KHPA12]. Adopta una metáfora de juego de cartas: cuando un usuario carga un conjunto de datos, el sistema analiza automáticamente sus propiedades estadísticas, identificando patrones, relaciones y características relevantes para visualizar. A partir de este análisis, genera una "mano" de visualizaciones recomendadas, presentadas al usuario como "cartas" que pueden añadirse al tablero de control. Estas visualizaciones se clasifican y ordenan según su relevancia para el conjunto de datos específico. Los usuarios pueden seleccionar las visualizaciones que les resulten más útiles y descartar las que no les resulten interesantes. A través de estas interacciones, el sistema aprende y ajusta sus recomendaciones, proponiendo visualizaciones similares en el futuro. De manera similar, **Microsoft Excel**[3] incorporó una función de *Gráficos recomendados* que sugiere visualizaciones adecuadas según los datos seleccionados, mientras que **Google Sheets**[4] se sumó con su función *Explorar*, que utiliza inteligencia artificial y procesamiento de lenguaje natural para recomendar preguntas y visualizaciones relevantes para los datos del usuario. Sin embargo,

[3]https://www.microsoft.com/es-ar/microsoft-365/excel

[4]https://support.google.com/

(a)

```
{
  "data": {"url": "data/cars.json"},
  "marktype": "point",
  "encoding": {
    "x": {
      "name": "Miles_per_Gallon",
      "type": "Q",
      "summarize": "mean"
    },
    "y": {
      "name": "Horsepower",
      "type": "Q",
      "summarize": "mean"
    },
    "row": {
      "name": "Origin",
      "type": "N",
      "sort": [{"name": "Horsepower",
          "summarize": "mean", "reverse": true}]
    },
    "color": {"name": "Cylinders","type": "N"}
  }
}
```



(b)

Figura 5.2: (a) Interfaz de usuario de **Voyager**. (b) Una especificación *Vega-lite* [SMWH17] (izquierda) de la visualización (derecha). Figuras extraídas de [WMA+15].

el proceso exacto de selección de visualizaciones no se ha revelado completamente en ninguno de estos casos.

En 2015, Vartak *et al.* emplearon métodos estadísticos para recomendar diversas representaciones de gráficos de barras y gráficos de líneas [VRM+15]. Su prototipo, llamado **SEEDB**, utiliza una métrica sencilla para evaluar la utilidad de una visualización: se considera que una visualización es interesante si muestra una gran desviación respecto a un conjunto de datos de referencia (que puede ser generado a partir del conjunto de datos completo o a partir de un conjunto de comparación definido por el usuario). Es decir, **SEEDB** calcula una desviación del subconjunto de los datos en comparación con el conjunto de datos de referencia. Luego, recomienda aquellas visualizaciones para las cuales los datos subyacentes presentan una alta desviación. Los autores sostienen que los usuarios consideran más interesantes y expresivas, aquellas visualizaciones con altas desviaciones.

Por su parte, **Voyager**, una aplicación web desarrollada por Wongsuphasawat *et al.*, sugiere una lista de visualizaciones basadas en las propiedades estadísticas de los datos [WMA+15, WQM+17]. Utiliza el motor de recomendación *CompassQL* [WMA+16] y un lenguaje de especificación de alto nivel llamado *Vega-lite* [SMWH17]. Una especificación en *Vega-lite* es un objeto JSON (*JavaScript Object Notation*) que describe una fuente de datos, el tipo de marca, las codificaciones visuales de las variables de datos y las transformaciones de datos, incluyendo filtros y funciones de agregación (ver figura 5.2). Asimismo, **Foresight** también se basa en propiedades estadísticas de los datos, como la alta correlación entre atributos o agrupaciones significativas de valores, para recomendar visualizaciones [DHPP17]. Esta herramienta ayuda a los usuarios a descubrir rápidamente *insights* visuales en conjuntos de datos grandes y de alta dimensionalidad.

En 2018, Kubernátová *et al.* introdujeron un modelo basado en preguntas denominado **NEViM** (*Non-Expert Visualization Model*) que utiliza un árbol de decisiones y una jerarquía de clasificación de visualización de datos para recomendar una técnica [KFvD18, KFVD19]. Además, incorpora perspectivas tanto impulsadas por las tareas como por las características de los datos.

La mayoría de los sistemas de recomendación descritos codifican pautas de visualización como una colección de declaraciones o reglas para generar visualizaciones automáticamente. En contraste, los sistemas basados en aprendizaje automático aprenden

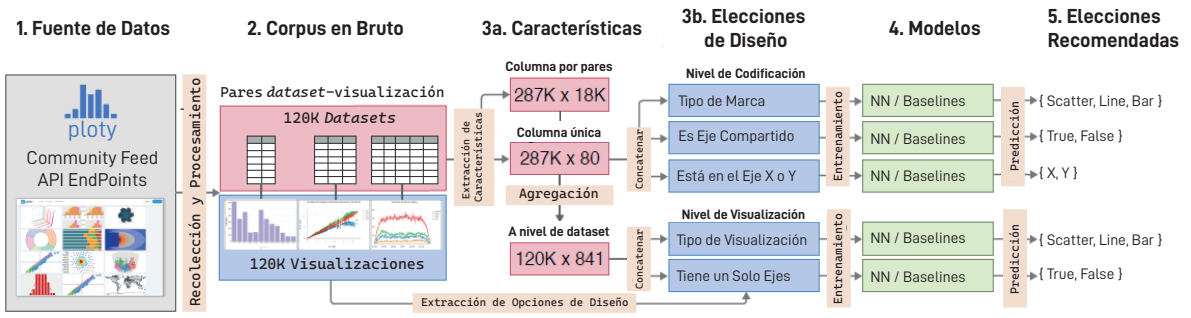


Figura 5.3: Descripción general de la estructura de **VizML**. (1) El proceso comienza con la recopilación de datos y visualizaciones del repositorio de *Plotly* [Plo18]. (2) Luego, se eliminan los duplicados, asegurando que cada par *dataset*-visualización sea único. (3) Se extraen características tanto del conjunto de datos como de la codificación de las visualizaciones. (4) Se entrenan modelos específicos para predecir decisiones de diseño visual, como el tipo de gráfico. (5) Finalmente, los modelos generan sugerencias de diseño para ayudar a los usuarios a seleccionar las mejores opciones visuales. Figura adaptada de [HBL+19].

directamente la relación entre los datos y las visualizaciones a través del entrenamiento de modelos. Por ejemplo, **VizML** utiliza aprendizaje automático para predecir decisiones de diseño de visualización [HBL+19]. El modelo se entrena con un millón de pares únicos de conjuntos de datos y visualizaciones extraídos del Feed de la Comunidad de *Plotly* [Plo18]. En la figura 5.3 se detalla el flujo de procesamiento y análisis de datos. Por su parte, **DeepEye** combina la generación de visualizaciones basada en reglas con modelos entrenados para clasificar una visualización como "buena" o "mala" [LQT+18, LQTL18, QLTL18]. El sistema aprende reglas de decisión sobre qué visualizaciones se consideran buenas o malas; por ejemplo, se pueden establecer reglas que indiquen que los gráficos de torta son ideales para comparar partes de un todo, mientras que los gráficos de barras son más apropiados para comparar diferentes categorías de datos.

Es importante mencionar que algunas de las herramientas descritas no son sistemas de recomendación en sí. Sin embargo, he decidido incluirlas porque han desempeñado un papel fundamental en la evolución de este campo y han sentado las bases para otros sistemas de recomendación.

5.2.2. Recomendadores Orientados a las Tareas

Los sistemas de recomendación de visualizaciones orientados a las tareas asisten a los usuarios en la selección de las representaciones visuales más adecuadas para sus datos, y en particular, para sus objetivos o propósitos de análisis.

En 1990, Roth y Mattis fueron pioneros en considerar los objetivos del usuario al visualizar datos [RM90]. Los autores argumentan que las variaciones en estos objetivos pueden influir de manera considerable en la efectividad de las técnicas de visualización utilizadas. Identifican distintos objetivos para la visualización de datos independientes del dominio, como comparación, distribución y correlación, entre otros. Posteriormente, desarrollaron **SageTools**, una plataforma innovadora que integra dos herramientas de diseño interactivas llamadas *SageBrush* y *SAGE* [RKMC95], y un repositorio de gráficos denominado *SageBook* [CRK+95]. *SageBrush* actúa como una interfaz de manipulación directa, permitiendo a los usuarios crear elementos gráficos de manera sencilla. Los usuarios pueden elaborar bocetos que se convierten en directivas de diseño para *SAGE*, un sistema automático de presentación, que interpreta estas especificaciones para generar gráficos. Por su parte, *SageBook* funciona como un repositorio, y permite a los usuarios navegar y recuperar los gráficos diseñados previamente.

Wehrend y Lewis desarrollaron un esquema de clasificación de técnicas de visualización basado en conjuntos de objetivos [WL90]. Su enfoque se estructuró en una matriz bidimensional, donde las columnas representaban los atributos de los datos, las filas correspondían a los objetivos del usuario y las celdas contenían diferentes tipos de visualizaciones. Para utilizar este esquema, el usuario debía descomponer un problema complejo en subproblemas más sencillos y luego identificar la entrada adecuada en la matriz para cada subproblema. Al combinar las técnicas de representación seleccionadas para los distintos subproblemas, se podía obtener una representación integral del problema original. Sin embargo, la matriz completa no fue publicada, lo que dificulta el acceso a los tipos específicos de visualizaciones que incluía. Por su parte, Stephen Casner propone **BOZ**, una herramienta automatizada que analiza una descripción lógica de la tarea que debe realizar un usuario y diseña una tarea perceptual equivalente mediante la sustitución de inferencias lógicas por inferencias perceptuales en la descripción de la tarea [Cas91].

En 2015, Saket *et al.* evaluaron la efectividad de cinco tipos de visualización bidimensional a pequeña escala (tablas, gráficos de líneas, gráficos de barras, gráficos de

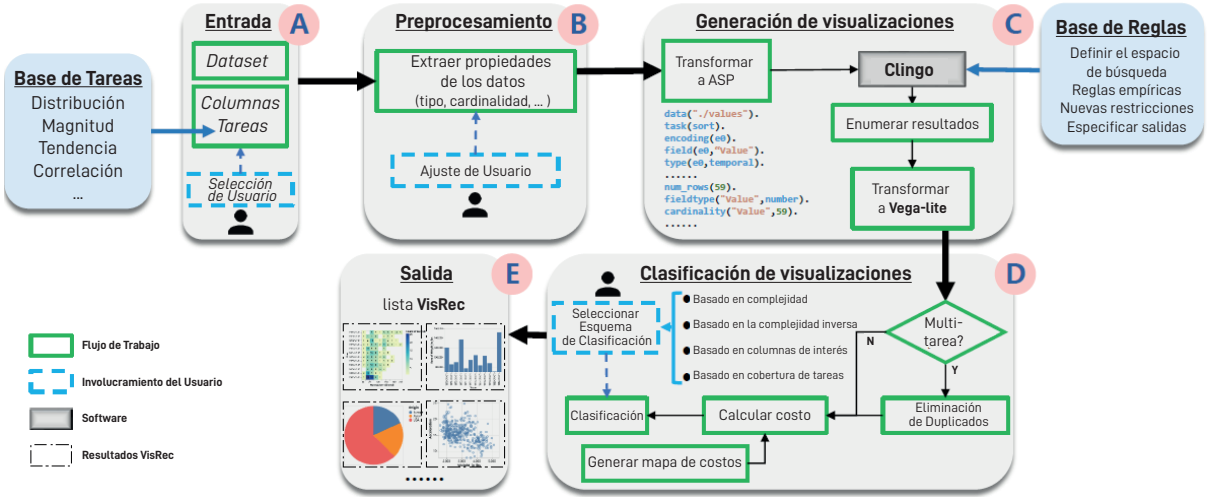


Figura 5.4: Descripción general de la estructura de **Taskvis**. (A) Entrada: el sistema recibe los datos suministrados por los usuarios. (B) Preprocesamiento: se extraen las características de los datos. (C) Generación de visualizaciones: se listan todos los candidatos válidos. (D) Clasificación de visualizaciones: se ordenan todas las visualizaciones de acuerdo con el esquema elegido. (E) Salida: se muestran los resultados de las recomendaciones a los usuarios. Figura adaptada de [SST+21].

dispersión y gráfico circular o de torta) en diez tareas comunes de análisis de datos, utilizando dos conjuntos de datos [SED19]. A partir de los resultados, desarrollaron un árbol de decisión y crearon **Kopol**, una herramienta que ofrece recomendaciones basadas en tareas específicas de los usuarios y conjuntos de datos. Más recientemente, se presentó **TaskVis**, un enfoque de recomendación de visualizaciones que combina 18 tareas clásicas de análisis de bajo nivel con los tipos de gráficos más adecuados, basándose en encuestas realizadas en entornos académicos e industriales [SST+21]. La figura 5.4 proporciona una descripción general de la estructura de la herramienta.

Estos sistemas ofrecen recomendaciones de visualizaciones individuales basadas en las tareas que los usuarios desean realizar. Otros estudios, sin embargo, se centran en recomendaciones orientadas a la creación de paneles de control o *dashboards*, que consisten en colecciones de múltiples vistas. Un ejemplo destacado es el trabajo de Pandey *et al.*, en el que se presenta **MEDLEY**, una interfaz diseñada para facilitar la composición de paneles de control al recomendar un conjunto de vistas lógicamente agrupadas y *widgets* de filtrado [PSS22]. Los usuarios pueden definir las tareas explícitamente a través de un panel de entrada o implícitamente mediante la selección de atributos de datos y vistas

Figura 5.5: Interfaz de usuario de **MEDLEY**. (a) Panel de entrada de atributos de datos y tareas. (b) Panel con recomendaciones. (c) Panel de control con las vistas agregadas. Figura extraída de [PSS22].

de interés (ver figura 5.5). Se identifican cuatro tareas clave: análisis de medidas (resumir uno o más atributos cuantitativos), análisis de cambios (visualizar variaciones en los datos a lo largo del tiempo), análisis de categorías (comparar diferentes categorías) y análisis de distribución (mostrar la distribución de registros a través de los campos de datos disponibles). Por otro lado, **MultiVision** emplea técnicas de aprendizaje automático para recomendar paneles de control analíticos, infiriendo columnas de datos potencialmente relevantes y seleccionando las vistas más apropiadas [WWZ+22].

5.2.3. Recomendadores Orientados al Conocimiento del Dominio

Una visualización efectiva depende en gran medida de la habilidad del usuario para interpretar correctamente una representación visual y extraer inferencias válidas. En este proceso, el conocimiento del dominio desempeña un papel crucial, ya que permite al usuario transformar los datos en información significativa. Los sistemas de recomendación que se fundamentan en este conocimiento buscan captar y utilizar esta información crítica dentro del entorno de visualización. Al integrar el conocimiento del dominio, estos sistemas pueden proponer visualizaciones más efectivas y relevantes, alineándose con los in-

tereses del usuario. Tal como señalan Meddes y McKenzie en su investigación, es esencial que el usuario aporte su conocimiento del dominio al contexto de los datos visualizados para interpretar adecuadamente una representación visual [MM00].

RAVE es el sistema más antiguo de recomendación de visualización orientado al conocimiento del dominio y fue desarrollado para la visualización de datos de medición espacial de la NASA [KAS94]. **RAVE** permite al usuario seleccionar un tipo de visualización u objetivo de una lista proporcionada, lo que activa una técnica asociada y genera los gráficos resultantes. Este sistema cuenta con una base de conocimiento que incluye objetos de visualización y reglas que vinculan los objetivos de visualización con dichos objetos. Las reglas de la base de conocimiento establecen qué constituye una visualización adecuada para un conjunto de datos y un objetivo específico. Cada objeto representa una técnica de visualización e incluye información como su nombre, los objetivos que puede satisfacer, los refinamientos que permite, los dominios en los que es aplicable y el programa que lo implementa en un marco de visualización. Por ejemplo, el objeto correspondiente al diagrama de dispersión puede satisfacer el objetivo de determinar si un atributo x está relacionado con un atributo y. Este objeto admite refinamientos como *zoom* y aplicación de color, y se puede aplicar en cualquier dominio donde se comparen atributos de valores numéricos. **RAVE** permite al analista explorar y comparar la efectividad de múltiples visualizaciones para alcanzar un objetivo, refinando o descartando las visualizaciones según sea necesario.

Años más tarde, Gilson *et al.* proponen un enfoque para la generación automática de visualizaciones a partir de datos específicos de dominio disponibles en la web [GSGC08]. Describen un *pipeline* que combina mapeo de ontologías y técnicas de razonamiento probabilístico. En primer lugar, la página web se mapeaba a una ontología de dominio, que almacena la semántica de un dominio temático específico (por ejemplo, listas de música). Luego, esta ontología se mapea a una o más ontologías de representación visual, que capturan la semántica de una técnica de visualización particular. Finalmente, cada ontología de representación visual se mapea a una visualización utilizando un kit de herramientas de visualización externo. Para implementar este enfoque, desarrollaron un prototipo llamado **SemViz** que, al combinarse con software de dominio público, puede generar visualizaciones adecuadas automáticamente a partir de una gran colección de páginas web populares.

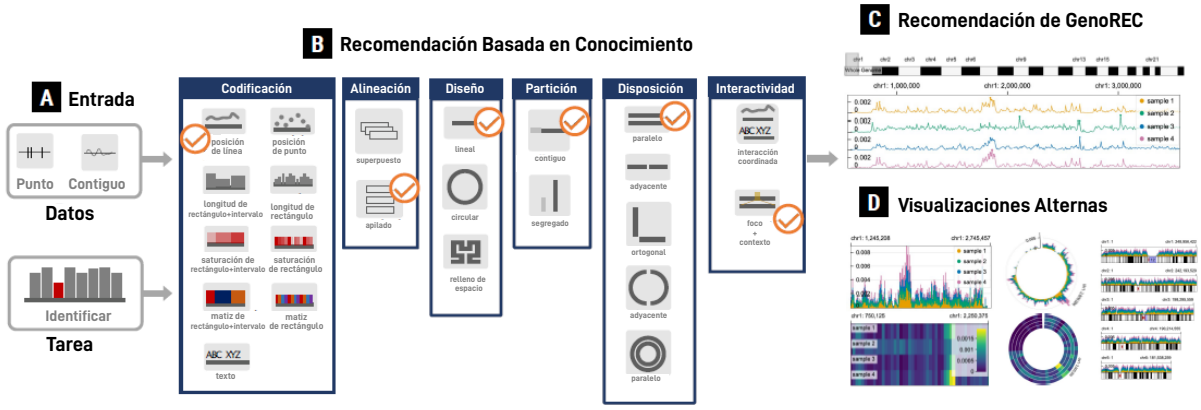


Figura 5.6: Descripción general de la estructura de **GenoREC**. (a) El sistema permite a los analistas especificar sus datos y tarea. (b) La recomendación basada en conocimiento muestra el modelo de **GenoREC** dividido en componentes y las decisiones tomadas en cada paso. (c) Con base en el modelo de recomendación, se recomienda una visualización adecuada para el usuario y (d) una variedad de opciones de visualización similares. Figura adaptada de [PLW+22].

Otro ejemplo de sistemas de recomendación en esta categoría es **GenoREC** [PLW+22]. Pandey *et al.* presentan un sistema innovador para la recomendación de visualizaciones interactivas de datos genómicos, que incluye información sobre los genes, su estructura, función y variaciones genéticas. El modelo utiliza reglas basadas en principios de visualización y conocimientos específicos del dominio de la genómica para sugerir visualizaciones efectivas en función de un conjunto de datos y las tareas de análisis correspondientes. Este sistema cuenta con una interfaz web que permite a los analistas especificar sus requisitos de datos y tareas mediante formatos comunes de archivos genómicos y descripciones de tareas. A partir de esta información, genera y recomienda visualizaciones apropiadas. Las recomendaciones se fundamentan en las mejores prácticas de visualización, que incluyen el uso de canales visuales efectivos para codificar datos cuantitativos y cualitativos, así como en el conocimiento del dominio, que sugiere diseños lineales para tareas de identificación y comparación. Además, **GenoREC** propone variaciones de diseño para fomentar la exploración y el descubrimiento.

5.2.4. Recomendadores Orientados a las Preferencias del Usuario

Los sistemas de recomendación de técnicas de visualización mencionados anteriormente presentan una limitación significativa: sugieren las mismas visualizaciones a todos los usuarios, sin considerar que los intereses, intenciones y preferencias de visualización

pueden variar considerablemente entre ellos. En esta categoría se agrupan aquellos sistemas de recomendación que recopilan explícitamente las intenciones y preferencias de los usuarios a partir de su comportamiento al interactuar con el sistema de visualización. Algunos estudios en este ámbito también integran técnicas probabilísticas y de aprendizaje automático para predecir las tendencias de elección de los usuarios, basándose en los datos de interacción recopilados.

Gotz y Wen propusieron un enfoque innovador denominado **BDVR** (*Behaviour Driven Visualization Recommendation*), el cual recomienda visualizaciones basadas en las acciones y comportamientos observados durante el uso del sistema [GW09]. El sistema está diseñado para escenarios en los que el usuario interactúa con un conjunto de datos de manera exploratoria, sin un objetivo de análisis completamente definido. En estas situaciones, **BDVR** identifica patrones en la forma en que el usuario manipula los datos y utiliza esa información para inferir sus intenciones analíticas.

BDVR se compone de dos fases diferenciadas: la fase de detección de patrones y la fase de recomendación de visualización. En la primera fase, se identifican cuatro patrones predefinidos basados en la actividad del usuario: el patrón de exploración (*scan pattern*), en el que el usuario compara visualmente los atributos de los objetos explorados; el patrón de alternancia (*flip pattern*), que indica la comparación de múltiples conjuntos de datos al cambiar iterativamente las restricciones de filtro; el patrón de intercambio (*swap pattern*), donde el usuario reordena las dimensiones de los datos para comparar correlaciones entre ellas; y el patrón de profundización (*drill-down pattern*), que ocurre cuando el usuario aplica filtros de manera repetida, lo que refleja su intención de centrar el análisis en un subconjunto específico de datos. En la segunda fase, un motor de recomendación infiere la intención del usuario a partir de estos patrones detectados, y sugiere las visualizaciones más adecuadas para satisfacer sus necesidades.

En 2015, Mutlu *et al.* presentaron **VizRec**, un sistema de dos etapas diseñado para generar visualizaciones personalizadas. Este sistema responde a consultas en texto libre, como las que se utilizan en motores de búsqueda, proporcionando una lista de visualizaciones ordenadas en formato $top - n$ [MVTS15b, MVTS15a, MVT17]. En la primera etapa, el sistema utiliza un algoritmo de mapeo basado en reglas que aplica directrices de codificación visual para crear una colección de visualizaciones apropiadas para los datos. Las directrices de codificación visual son principios generales que definen las relaciones

Figura 5.7: Descripción general de la estructura de **VizRec**. El sistema consta de dos etapas: una etapa basada en reglas que aplica directivas de codificación visual para generar una colección de visualizaciones apropiadas para los datos y una etapa de personalización que aplica las preferencias de los usuarios y filtra las visualizaciones según las necesidades e intereses de los usuarios. Figura adaptada de [MVT17].

entre los componentes visuales de una visualización y las características de los datos. La segunda etapa se centra en la personalización de las recomendaciones mediante un filtrado colaborativo basado en usuarios. En esta fase, se recopilan las calificaciones de los usuarios sobre la utilidad de diferentes visualizaciones, lo que permite predecir de manera más efectiva las preferencias del usuario activo y filtrar las visualizaciones según sus necesidades e intereses. La figura 5.7 muestra una representación esquemática del sistema.

Algunos sistemas han adoptado técnicas de aprendizaje automático que se basan en las interacciones de los usuarios para generar recomendaciones de visualizaciones [DPD16, HPP+18, LQTL18, PDDP16]. Qian *et al.* desarrollan un marco de aprendizaje automático para la personalización de gráficos, considerando tanto interacciones anteriores de los usuarios con diferentes visualizaciones como sus preferencias individuales [QRD+22]. Este marco también tiene la capacidad de extraer conocimiento de visualizaciones que son relevantes para otros usuarios, incluso si provienen de conjuntos de datos diferentes. Sin embargo, al igual que otros modelos [GW09, ZGCS21], se fundamenta en información previamente etiquetada, lo que restringe su flexibilidad para adaptarse a las variaciones en las preferencias durante la exploración. Por el contrario, **VisGuide** emplea un modelo de aprendizaje en línea que le permite adaptarse de manera dinámica a los cambios en las preferencias de los usuarios a lo largo del proceso de exploración [CLPL22].

5.3. Conclusiones

En este capítulo, se ha analizado el papel fundamental de los sistemas de recomendación de técnicas de visualización en un contexto caracterizado por la creciente complejidad de los datos.

Mediante un estudio detallado, se han identificado cuatro enfoques principales: recomendadores basados en las características de los datos, en las tareas, en el conocimiento del dominio y en las preferencias del usuario. Cada uno de estos enfoques aborda desafíos específicos, desde la estructura y distribución de los datos hasta las necesidades personalizadas de los usuarios y el conocimiento especializado en distintos dominios. Además, se han revisado herramientas y modelos clave que han sido fundamentales en la evolución de este campo, permitiendo que usuarios con distintos niveles de experiencia generen visualizaciones efectivas para optimizar el análisis de datos complejos.

A partir del estudio detallado de los sistemas recopilados, se observa que ninguno aborda de manera integral los enfoques definidos en este capítulo. En particular, no existe un sistema que considere simultáneamente las características de los datos, las tareas analíticas y las preferencias del usuario. Ante esta ausencia, se ha desarrollado un nuevo sistema de recomendación que integra estos tres factores. Este capítulo establece el marco conceptual necesario para su creación, mientras que en el siguiente se presenta su diseño y funcionamiento.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 6

Sistema de Recomendación Integral de Visualización (CVRS)

6.1. Introducción

El diseño de visualizaciones efectivas requiere, en muchas ocasiones, que el usuario cuente con un conocimiento profundo del dominio del conjunto de datos, además de habilidades específicas en diseño visual y análisis de información. Este proceso exige tiempo y dedicación, especialmente cuando implica definir manualmente parámetros como las marcas, los canales visuales y las técnicas de visualización adecuadas para representar la información. Estas tareas pueden resultar tediosas y, al llevarse a cabo en un espacio de búsqueda tan amplio, probablemente no lleve a un resultado satisfactorio, especialmente para aquellos usuarios no expertos en análisis visual. Para abordar estas dificultades y optimizar el flujo de trabajo, muchas herramientas de visualización incorporan mecanismos de recomendación. Estas recomendaciones no solo aceleran el proceso, sino que también contribuyen a mejorar la calidad de las visualizaciones.

En esta tesis, se ha adoptado un enfoque basado en tres pilares fundamentales para abordar el diseño de representaciones visuales de datos n-dimensionales: el nivel de datos, que responde a la pregunta "*¿Qué visualizar?*"; el nivel de tarea, que aborda "*¿Por qué visualizar?*" y el nivel de representación visual, centrado en "*¿Cómo visualizar?*". Estos pilares proporcionan un marco conceptual adecuado para guiar tanto el diseño de visualizaciones como el desarrollo de sistemas de recomendación.

En este capítulo, se introduce un Sistema de Recomendación Integral de Visualización

llamado **CVRS** (por sus siglas en inglés, *Comprehensive Visualization Recommendation System*). Este sistema guía el proceso de visualización, permitiendo a los usuarios gestionar, configurar y analizar sus conjuntos de datos, y asegura una alineación coherente entre los datos, los objetivos y la representación visual, lo que da como resultado visualizaciones efectivas y expresivas.

CVRS adopta un enfoque fundamentado en los tres pilares y abarca todo el proceso, desde la importación de datos hasta la recomendación de técnicas visuales específicas. En la etapa inicial, los usuarios cargan sus conjuntos de datos en la aplicación y seleccionan columnas que identifiquen las muestras, definiendo su clase o categoría. Esta etapa incorpora un sistema de validación diseñado para verificar la estructura del conjunto de datos y el tipo de cada atributo, asegurando así su consistencia y preparación para el análisis. Este mecanismo es fundamental para minimizar posibles errores en fases posteriores, garantizando que el análisis y la visualización se construyan sobre datos confiables. A continuación, el sistema facilita la configuración detallada de la representación visual y de los objetivos de análisis. En esta etapa, los usuarios pueden personalizar diversos aspectos de la visualización, como el tipo de marcas a utilizar (puntos, líneas, barras, etc.), la orientación y tipo de ejes, así como las tareas analíticas que desean llevar a cabo. La arquitectura descrita en esta sección está diseñada para ser flexible y dinámica, adaptándose en tiempo real a las elecciones del usuario, lo que proporciona una experiencia interactiva y ajustada a sus necesidades. Finalmente, el sistema recomendador ofrece un conjunto de técnicas de visualización que se ajustan mejor a las preferencias del usuario.

Este sistema, a diferencia de los analizados en el capítulo 5, se destaca por abordar de manera integral tres de los cuatro enfoques o factores de recomendación definidos en dicho capítulo: las características de los datos, las tareas analíticas y las preferencias del usuario. Esto permite ofrecer recomendaciones más precisas y personalizadas, alineadas tanto con los objetivos analíticos del usuario como con la naturaleza de los datos, lo que maximiza la relevancia y efectividad de las visualizaciones sugeridas, mejorando la experiencia del usuario y promoviendo un análisis más ágil, intuitivo y ajustado a sus necesidades. Además, a diferencia de la mayoría de los sistemas actuales, que restringen al usuario a configuraciones predefinidas o recomendaciones automáticas, **CVRS** amplía las posibilidades de personalización. El sistema permite al usuario ajustar aspectos clave

del mapeo visual, como las marcas y canales visuales, la orientación y los tipos de ejes, facilitando la adaptación de la representación gráfica a las necesidades específicas del análisis, a través de una interfaz intuitiva que no requiere conocimientos previos.

6.2. Los Datos

Permitir que los usuarios importen sus propios conjuntos de datos representa una ventaja significativa, ya que les brinda flexibilidad y autonomía en el análisis. Sin embargo, esta capacidad también plantea desafíos, ya que los datos ingresados deben adaptarse a la estructura interna del sistema para garantizar su correcto procesamiento y representación. Por esta razón, es fundamental establecer ciertas restricciones que aseguren la compatibilidad y faciliten la integración de los datos proporcionados por los usuarios.

En este sentido, **CVRS** adopta el formato tabular como estructura de datos principal, debido a su amplia aplicabilidad en diversos contextos, su facilidad de comprensión y su capacidad para representar información de manera clara y estructurada. La representación tabular de los datos ha sido esencial en la organización de la información a lo largo del tiempo. Aunque existen otros formatos para representar datos (ver sección 3.3), el formato tabular se distingue por su versatilidad y su amplia aceptación en diversas disciplinas. En este formato, cada fila representa una entidad única, como una persona, un evento o mediciones tomadas mediante un experimento, mientras que cada columna refleja una característica o atributo específico de esa entidad, como la edad, la especie u otras propiedades relevantes. En este sistema en particular, cada columna tiene un tipo determinado, que puede ser categórico o cuantitativo.

En esta implementación inicial del sistema, se define una estructura específica para el procesamiento de datos tabulares. Esta estructura exige que el conjunto de datos incluya al menos tres atributos cuantitativos y permite la incorporación de hasta dos atributos categóricos opcionales: un identificador único para cada ítem de datos y un indicador que defina su clase o categoría. Para ilustrar mejor esta estructura de datos, podemos considerar el conjunto de datos Iris [Fis88], un referente en el campo de la visualización y análisis de datos. Este conjunto incluye mediciones de 4 características de las flores de iris de tres especies diferentes (Setosa, Versicolor y Virginica). Cada muestra o ítem se caracteriza por cuatro atributos cuantitativos (longitud y ancho tanto del sépalo como del pétalo),

un atributo categórico que especifica la especie, y opcionalmente, un identificador único. La matriz de diagramas de dispersión, por ejemplo, es una técnica que aprovecha esta organización de los datos. Ésta requiere al menos dos atributos cuantitativos y permite generar gráficos donde las muestras se distinguen visualmente mediante colores según su especie y puede aprovechar el identificador único para permitir la identificación individual de muestras en un contexto interactivo.

6.3. Las Tareas

Hemos identificado el objetivo o propósito del análisis como un segundo aspecto clave en el diseño de visualizaciones. Reconocemos que, en la literatura, existen diversas taxonomías que utilizan terminologías distintas para describir tareas que, en esencia, son similares, o incluso iguales [WL90, AES05, VPF06, Mun14]. Para proporcionar un marco unificado que simplifique la interpretación y aplicación de las tareas, y tras realizar un análisis exhaustivo de las clasificaciones propuestas en las últimas décadas (ver sección 3.6), hemos desarrollado una nueva taxonomía que unifica las terminologías y definiciones existentes. Esta taxonomía integra las 10 tareas más relevantes en el contexto de la visualización de datos, de las cuales 6 son identificadas como tareas principales. Además, una de estas tareas, *Explorar*, se desglosa en 5 sub-tareas específicas.

Es relevante señalar que las diferentes tareas incluidas en nuestra taxonomía son flexibles y pueden aplicarse a una amplia variedad de objetos de análisis, los cuales corresponden a todas las posibles combinaciones de subconjuntos de datos que pueden presentarse en una tabla. En la tabla 6.1 se presenta una definición de cada objeto de análisis, incluyendo su nombre y las diversas tareas que se pueden llevar a cabo sobre cada uno de ellos. Por ejemplo, identificar relaciones entre los valores de los atributos resulta especialmente útil al trabajar con conjuntos de datos que contienen múltiples atributos, es decir, cuando se analizan objetos como *datasets* y tuplas de atributos. Esta tarea permite a los analistas identificar relaciones significativas que pueden proporcionar información valiosa sobre el comportamiento de los datos. No obstante, es fundamental que el análisis de relaciones se realice considerando el conjunto completo de muestras disponibles en el conjunto de datos. Limitar el análisis a un subconjunto de los datos puede generar conclusiones erróneas que no sean representativas del conjunto total.

A continuación, se presenta una descripción detallada de cada una de las tareas incluidas en la taxonomía, acompañada de ejemplos ilustrativos.

1. **Comparar**

   Descripción general: *Dado un conjunto de datos, examinar similitudes y diferencias entre objetos en un conjunto de datos [KK94b]. Una vez que los datos han sido localizados e identificados, compararlos analizando dimensiones, elementos de datos, valores, clústeres, propiedades, proporciones, ubicaciones y distancias, así como características visuales [VPF06].*

   Ejemplo: "¿Qué automóviles son más eficientes en combustible, los coches japoneses o los coches americanos?"

   La tarea de *Comparar* emerge como un objetivo central y recurrente en la literatura, manifestándose de diversas formas según los diferentes autores. Esta tarea fue definida por Wehrend y Lewis [WL90], quienes distinguieron dos tipos de comparación: dentro de relaciones y entre relaciones. La comparación dentro de relaciones consiste en evaluar los atributos de un mismo conjunto de datos, donde el usuario identifica similitudes y diferencias entre elementos similares. Por otro lado, la comparación entre relaciones implica comparar diferentes conjuntos de datos, lo que permite entender mejor las interacciones y relaciones entre ellos. Tamara Munzner [Mun14] y Andrew Abela [Abe08] también mencionan la comparación como una tarea esencial, reconociendo su papel fundamental en el proceso de análisis visual. Un avance significativo en la conceptualización de las tareas de comparación fue realizado por Valiati *et al.* [VPF06]. Su contribución expandió el alcance de la comparación al introducir un marco más granular que abarca la comparación a través de dimensiones, elementos de datos, valores y propiedades. Más recientemente, Shen *et al.* [SST+21], identifican tres tareas que se pueden considerar derivadas de la tarea de comparar. La primera, *Desviación*, implica comparar datos con respecto a un valor de referencia, como cero o la media. La segunda tarea, *Magnitud*, se refiere a mostrar comparaciones de tamaño relativo o absoluto entre diferentes elementos. Por último, *Tendencia*, utiliza técnicas de regresión para mostrar la variación a lo largo del tiempo. Esta última tarea puede considerarse como una composición de otras tareas, incluida la comparación, así como la distribución, agregación y recuperación

de valores.

2. **Agrupar**

   Descripción general: *Dado un conjunto de datos, identificar y determinar conjuntos de elementos con atributos similares [KK94b, AES05].*

   Ejemplo: "¿Existen grupos de automóviles con características similares de cilindros/peso/aceleración? ¿Podemos agrupar los automóviles de origen europeo por su peso y capacidad de aceleración?"

   El objetivo principal de esta tarea es encontrar, dentro de un conjunto de datos, clústeres o grupos con valores de atributos similares. En este enfoque, se integran las tareas de categorizar y clusterizar dentro de la tarea más amplia de agrupar [WL90]. La categorización se refiere a la clasificación de objetos en diferentes categorías definidas por el usuario, mientras que la clusterización implica que el sistema identifica categorías y muestra los objetos relacionados agrupados juntos. Ambas tareas contribuyen a la comprensión y organización de los datos en función de sus similitudes y diferencias.

3. **Determinar distribución**

   Descripción general: *Dado un conjunto de datos y un atributo de interés, determinar la distribución de los valores de ese atributo en el conjunto [AES05].*

   Ejemplo: "¿Cuál es la distribución del peso de los automóviles del conjunto de datos? ¿Y la de los automóviles de origen japonés en particular?"

4. **Determinar relaciones**

   Descripción general: *Dado un conjunto de datos y dos atributos, determinar relaciones útiles entre los valores de esos atributos [AES05].*

   Ejemplo: "¿Existe relación entre el año de origen de los automóviles y el peso de los mismos?"

5. **Explorar**

   Esta tarea integra múltiples subtareas para examinar y comprender en profundidad un conjunto de datos. Permite al usuario investigar las características, comportamientos y peculiaridades de los datos mediante cinco operaciones fundamentales:

a) **Recuperar valor**

Descripción general: *Dado un conjunto de datos específico, encontrar atributos de esos datos [AES05]. A menudo actúa como una subtarea para otras tareas.*

Ejemplos: "¿Cuál es el rendimiento de millas por galón del Audi TT? ¿Cuál es el peso del Renault 12?"

b) **Buscar anomalías**

Descripción general: *Detectar cualquier anomalía dentro de un conjunto de datos específico en relación con una expectativa o relación dada, como pueden ser los valores estadísticos atípicos [AES05].*

Ejemplos:"¿Existen excepciones a la relación entre la potencia (caballos de fuerza) y la aceleración?"

c) **Determinar rango**

Descripción general: *Dado un conjunto de elementos de datos y un atributo de interés, encontrar el rango de valores dentro del conjunto [AES05].*

Ejemplos: "¿Cuál es el rango de potencia (caballos de fuerza) de los automóviles?"

d) **Buscar valores extremos**

Descripción general: *Identificar elementos de datos que presentan un valor extremo de un atributo dentro de su rango en el conjunto de datos [AES05].*

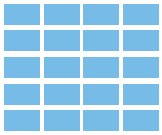
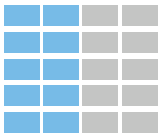
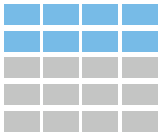
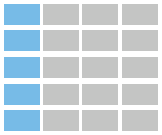
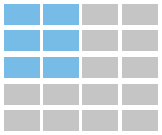
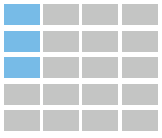
Ejemplos: "¿Cuál es el coche con el mayor MPG?"

Es importante destacar que esta tarea se distingue de *Ordenar*, ya que no siempre se requiere realizar un ordenamiento completo para identificar un valor extremo. Además, también se diferencia de *Buscar anomalías*, ya que las anomalías no necesariamente corresponden a valores extremos.

e) **Recuperar muestra**

Descripción general: *Dadas algunas condiciones concretas sobre los valores de los atributos, encontrar elementos de datos que satisfacen esas condiciones [SST+21].*

Ejemplos: "¿Cuáles son los automóviles de origen japonés?"

Objeto de Análisis	Definición	Tareas
 Dataset	m filas, n columnas	Comparar Agrupar Determinar distribución Determinar relaciones Explorar: Recuperar valor Explorar: Buscar anomalías Explorar: Determinar rango Explorar: Buscar valores extremos Explorar: Recuperar muestra Ordenar
 Tupla de atributos	m filas, k columnas $0 < k < n$	
 Tupla de ítems	p filas, n columnas $0 < p < m$	Comparar Agrupar Determinar distribución Explorar: Recuperar valor Explorar: Buscar anomalías Explorar: Determinar rango Explorar: Buscar valores extremos Explorar: Recuperar muestra Ordenar
 Atributo	m filas, 1 columna	
 Subconjunto de atributos de ítems	p filas, k columnas $0 < p < m$ $0 < k < n$	
 Un atributo de un subconjunto de ítems	p filas, 1 columna $0 < p < m$	

Sigue en la página siguiente.

...continuación de la página anterior.

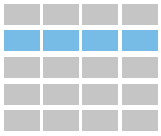
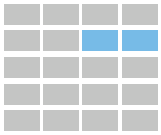
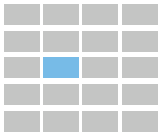
 Ítem  Subconjunto de atributos de un ítem	1 fila, n columnas 1 fila, p columnas $0 < p < n$	Comparar Explorar: Recuperar valor
 Celda	1 fila, 1 columna	Explorar: Recuperar valor

Tabla 6.1: Descripción de los objetos de análisis, incluyendo su denominación y las tareas asociadas que pueden realizarse con cada uno.

La tarea de *Explorar* abarca múltiples subtareas y se presenta en diferentes trabajos bajo distintos términos y objetivos. Munzner introduce tareas de alto nivel como *Descubrir*, definida como el uso de la visualización para revelar conocimientos previamente desconocidos, y *Buscar*, que se refiere a analizar visualizaciones para localizar elementos de interés [Mun14]. La búsqueda se divide en cuatro tipos, según el conocimiento previo de la identidad y ubicación del objetivo, y pueden pensarse como casos especiales de las subtareas **Recuperar valor** y **Recuperar muestra** que hemos definido: *Encontrar*, donde los usuarios saben qué están buscando y dónde encontrarlo; *Localizar*, en la que se conoce el objetivo pero no su ubicación; *Navegar*, donde la identidad del objetivo no es conocida pero se especifica mediante características; y *Explorar*, cuando los usuarios desconocen tanto la identidad como la ubicación del objetivo.

Munzner también detalla la tarea de *Consultar*, que se ejecuta una vez que un objetivo ha sido localizado, y permite al usuario *Identificar* un objeto, es decir,

describir un objeto previamente desconocido o *Resumir*, que proporciona una visión general de los datos.

Autores como Wehrend y Lewis [WL90], y Valiati *et al.* [VPF06] también mencionan tareas como *Localizar*, que implica la búsqueda de un objeto conocido por el usuario, y *Visualizar*, definida como la representación gráfica del espacio de datos para facilitar la navegación a través del conjunto. Por su parte, Shen *et al.* [SST+21] sintetizan 18 tareas de análisis de bajo nivel, entre las que se encuentran *Determinar rango*, *Buscar anomalías*, *Buscar valores extremos*, *Recuperar valor* y *Filtrar*, que se refiere a la tarea de buscar casos de datos que satisfacen las restricciones dadas.

Todas estas tareas pueden integrarse en una tarea más amplia, denominada *Explorar*.

6. **Ordenar**

   Descripción general: *Dado un conjunto de datos, ordenarlos de acuerdo con alguna métrica ordinal [AES05].*

   Ejemplos: Ordenar los automóviles por peso. Ordenar los automóviles de origen japonés por la potencia (caballos de fuerza) y el peso.

   La tarea de ordenar se presenta en diversas publicaciones bajo distintos términos [WL90, AES05]. En la literatura, esta tarea raramente se presenta como una tarea independiente, y suele ser considerada un paso preliminar para la identificación de valores extremos.

6.4. La Representación Visual

Como se mencionó anteriormente, el proceso de visualización de datos involucra no solo la elección de *qué* datos visualizar y *por qué*, sino también la forma en que se van a representar visualmente. En este sentido, el sistema **CVRS** se distingue por su enfoque centrado en el usuario, ofreciendo un nivel de personalización superior al de otros sistemas de visualización descritos en la literatura. Mientras que la mayoría de los sistemas actuales limitan al usuario a configuraciones predeterminadas o recomendaciones automáticas basadas en los datos y objetivos, **CVRS** permite una personalización detallada de la visualización, incluyendo la selección de marcas, canales visuales, orientación

y tipos de ejes. Los usuarios pueden ajustar el mapeo visual según sus preferencias y necesidades analíticas, mediante una interfaz intuitiva, sin tener que lidiar con lenguajes de bajo nivel. Sin embargo, este ajuste se realiza dentro de los límites establecidos por el sistema, que asegura que las modificaciones se realicen de manera coherente y adecuada, evitando configuraciones erróneas o inapropiadas. Este enfoque representa una mejora significativa frente a los sistemas tradicionales, ya que ofrece una capacidad de personalización que se ajusta a la diversidad de requerimientos de los usuarios.

Esta capacidad de personalización es fundamental por varias razones. En primera instancia, cada usuario tiene una experiencia y una familiaridad diferente con ciertos tipos de representaciones gráficas. Por ejemplo, algunos usuarios pueden trabajar más eficientemente con visualizaciones sin ejes, mientras que otros prefieren configuraciones con ejes cuantitativos tradicionales. De manera similar, las preferencias en la orientación de los ejes pueden variar significativamente: algunos usuarios se sienten más cómodos con representaciones ortogonales convencionales, mientras que otros pueden preferir diseños radiales o configuraciones más libres. Esta capacidad de personalización no es meramente una cuestión estética, sino que impacta directamente en la eficiencia del análisis de datos. Al permitir que los usuarios personalicen las visualizaciones según sus preferencias y experiencia, se facilita el descubrimiento y se optimiza el proceso analítico.

Además, esta personalización mejora la comprensión y la interpretación de los datos. Un sistema de visualización que impone configuraciones predeterminadas podría, en algunos casos, dificultar la detección de patrones o la identificación de relaciones clave en los datos. **CVRS**, al permitir al usuario elegir entre diferentes tipos de representaciones, facilita la exploración de los datos desde distintas perspectivas.

Otro aspecto crucial de esta personalización es su capacidad para adaptarse al contexto o dominio específico del usuario, ya que cada campo de estudio o área profesional ha desarrollado, a lo largo del tiempo y la práctica, convenciones y preferencias específicas para la representación visual de sus datos. La capacidad de ajustar el mapeo visual según estas convenciones específicas de dominio puede impactar directamente en la eficacia del análisis. Además, la posibilidad de modificar múltiples aspectos de la visualización fomenta la exploración, ya que los usuarios suelen explorar diversas formas de representar los datos hasta encontrar la más adecuada.

En este contexto, **CVRS** ofrece más de una docena de técnicas de visualización,

brindando a los usuarios la posibilidad de seleccionar entre diversas opciones, cada una vinculada a configuraciones particulares de mapeos visuales. La tabla 6.2 resume las técnicas de visualización actualmente soportadas por **CVRS**. Para cada una de éstas se especifican las configuraciones de mapeo visual disponibles y las tareas analíticas para las que son más efectivas.

6.5. Arquitectura del Sistema

CVRS se estructura en tres etapas principales: 1) Preprocesamiento de Datos, 2) Configuración de Tareas y Preferencias de Visualización y 3) Clasificación.

En la primera etapa, el usuario carga el conjunto de datos que desea analizar. Posteriormente, el sistema propone un conjunto de parámetros relacionados con los objetivos de análisis y la codificación visual, que el usuario seleccionará según sus necesidades y preferencias específicas. En la última etapa, el sistema clasifica las técnicas de visualización recomendadas según su adecuación a los parámetros seleccionados, priorizando aquellas que mejor se ajusten a las preferencias y requisitos definidos por el usuario.

1. **Preprocesamiento de los Datos**

   En la fase de preprocesamiento de datos, se analiza la estructura sintáctica de los mismos y se elabora una descripción inicial, lo cual constituye un paso fundamental dentro del primer pilar del marco metodológico, el "*¿Qué?*".

   En esta etapa, el preprocesador identifica los tipos de datos y determina su sintaxis, clasificando cada columna en datos categóricos y numéricos. Esta reflexión preliminar sobre "¿Qué datos visualizará el usuario final?' es clave, ya que orienta las decisiones de diseño subsecuentes.

   Además, el sistema ofrece una interfaz interactiva que permite al usuario seleccionar manualmente, entre los datos categóricos detectados automáticamente, aquellas columnas que contienen identificadores únicos de muestras y variables categóricas que describen grupos o categorías. Esta distinción es esencial, dado que el sistema está optimizado para procesar datos numéricos.

2. **Configuración de Tareas y Preferencias de Visualización**

			Correlation Matrix	Box graph	Table (LineUp)	SPLOM	Parallel coordinates	Radial coordinates	Caras de Chernoff	RadViz	Star Coordinates	Strip plot	Violin chart	UMAP	SOM	Principal Component Analysis (PCA)
Técnicas																
Visualización Directa			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
Sustrato espacial	Ejes	No estructurados	■	■	■				■				■		■	
		Cuantitativos		■		■	■	■	■	■	■	■	■	■	■	■
	Orientación	Ortogonal		■		■			■			■	■	■		■
		Paralela		■			■			■	■		■			
		Radial		■				■					■			
		Libre		■		■				■	■	■	■	■		■
Sustrato gráfico	Marcas	Punto(0D)	■			■				■	■	■		■		■
		Línea(1D)			■		■	■								
		Área(2D)													■	
		Glifo(ND)		■					■				■			
	Canales Visuales	Posición		■		■	■	■	■	■	■	■	■	■		■
		Color	■	■	■	■	■	■	■	■	■	■	■		■	
		Forma	■	■					■				■			
		Tamaño	■		■				■	■						
		Orientación	■				■	■	■							
Tareas		Comparar		■	■	■	■	■	■	■	■	■	■			
		Agrupar					■	■						■	■	■
		Determinar distribución		■	■							■	■			
		Determinar relaciones	■			■	■	■		■	■					
		Ordenar			■		■	■								
		Explorar		■	■	■	■	■		■	■	■	■			

Tabla 6.2: Resumen de las técnicas de visualización soportadas por **CVRS**, indicando las configuraciones de mapeo visual disponibles y las tareas analíticas para las que son más adecuadas.

En la segunda fase de la arquitectura de **CVRS**, el sistema permite a los usuarios definir los parámetros que orientarán la visualización de los datos de acuerdo con sus preferencias de representación visual y objetivos de análisis. Una vez que el usuario ha especificado estas preferencias, el sistema las procesa y las transforma en parámetros que se utilizan como entrada para la siguiente etapa de clasificación.

Este paso es clave para el segundo pilar del marco metodológico, el "*¿Por qué?*", ya que implica la especificación de las tareas que el usuario desea realizar. Además, esta fase está vinculada al mapeo visual, que corresponde al tercer pilar, el "*¿Cómo?*", ya que implica la transformación de los datos en elementos gráficos mediante configuraciones visuales seleccionadas.

En esta etapa, **CVRS** permite a los usuarios personalizar la visualización según sus necesidades y preferencias analíticas. Los usuarios eligen tanto las tareas que desean realizar —tales como comparar valores, identificar patrones de agrupamiento o explorar relaciones en los datos— como sus preferencias de codificación visual. Para ello, el sistema presenta opciones de mapeo visual, permitiendo que el usuario seleccione configuraciones de preferencia, como el tipo de marcas, canales visuales, tipos de ejes y otros atributos visuales que desee incluir en la visualización.

Un aspecto fundamental en el proceso de selección del mapeo visual de acuerdo a sus prioridades, es la manera en que se presenta al usuario. El sistema incorpora una "lista de prioridad", mediante la cual el usuario elige las configuraciones que desea incluir en su visualización y establece el orden de preferencia de cada elemento del mapeo visual. Por ejemplo, el usuario puede asignar mayor prioridad a las marcas y a las tareas, mientras que considera de menor importancia el tipo o la orientación de los ejes. La lista de prioridad está estructurada en una serie de casillas ordenadas de alta a baja prioridad, donde el usuario puede "arrastrar" los distintos elementos de configuración entre las casillas para modificar su importancia dentro del mapeo deseado. Como se ilustra en la figura 6.1(a), el sistema presenta inicialmente tres casillas: una correspondiente a prioridad alta (con un peso de 6), otra a prioridad baja (con un peso de 1) y una casilla con peso cero para aquellos elementos en los que el usuario no considera necesario asignar una prioridad. Además, el sistema permite la inclusión de casillas con prioridad intermedia (con pesos que varían entre 2 y 5), según lo requiera el usuario (ver figura 6.1(b)).

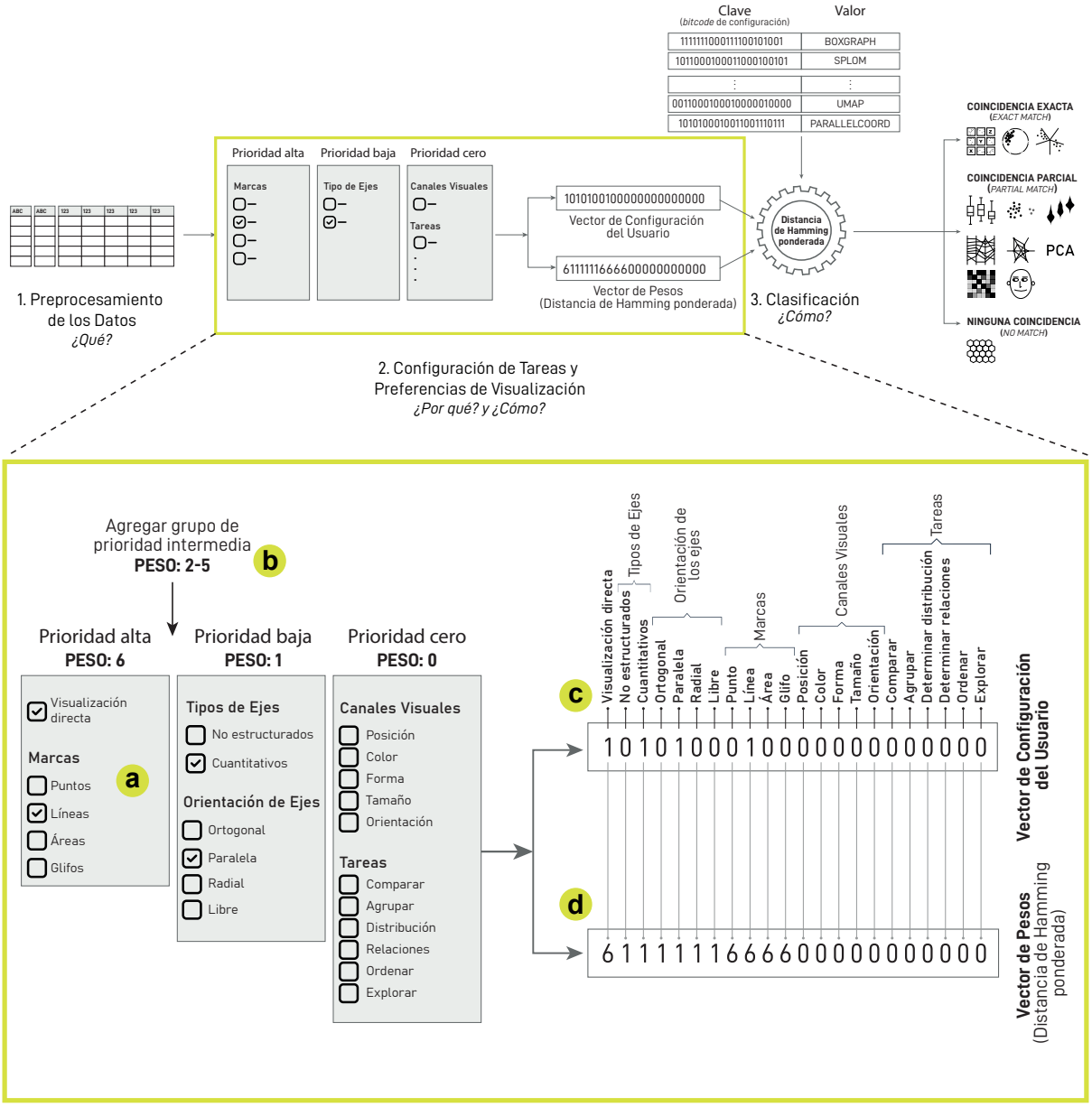


Figura 6.1: Arquitectura del sistema **CVRS**. (1) *Preprocesamiento de los datos*: se clasifican los datos en categóricos y numéricos, permitiendo al usuario seleccionar las columnas clave para su visualización. (2) *Configuración de tareas y preferencias de visualización*: el usuario define los parámetros que orientarán la visualización de los datos de acuerdo con sus preferencias de representación visual y objetivos de análisis. (3) *Clasificación*: el sistema selecciona las técnicas de visualización más adecuadas para comunicar de manera efectiva la información.

El sistema permite al usuario elegir y priorizar entre varios elementos de configuración visual que se pueden ajustar según sus preferencias analíticas. Estos elementos incluyen opciones de **visualización directa**[1], **tipos de ejes** (como *no estructurado* o *cuantitativo*) y su **orientación** (ya sea *ortogonal*, *paralelo*, *radial* o *libre*). También se pueden seleccionar las **marcas**, como *puntos*, *líneas*, *áreas* o *glifos*, y definir los **canales visuales** como *posición*, color, *forma*, *tamaño* u *orientación*. Además, el usuario puede establecer el enfoque de las **tareas** analíticas que desea realizar, incluyendo opciones como *comparar*, *agrupar*, *determinar distribución*, *determinar relaciones*, *explorar* y *ordenar*.

Una vez que el usuario ha especificado sus preferencias de codificación visual y objetivos analíticos, el sistema procesa estas especificaciones y las traduce en dos vectores que constituyen los parámetros de entrada para la etapa de clasificación.

Vector de configuración

El primer vector, denominado vector de configuración, se implementa mediante un código binario (*bitcode*), que constituye una representación digital de las preferencias del usuario en formato binario. La estructura del vector se organiza de manera que cada posición (bit) corresponde a una característica específica de configuración visual (ver figura 6.1(c)). El sistema emplea una lógica binaria donde se asigna un valor de 1 cuando el usuario selecciona una configuración como deseable (por ejemplo, la utilización de líneas como marcas o la implementación de ejes paralelos), mientras que se mantiene un valor de 0 para aquellas opciones no seleccionadas. Esta codificación binaria permite una identificación precisa y sistemática de las preferencias del usuario, facilitando el posterior procesamiento y selección de técnicas de visualización adecuadas.

Vector de pesos

El segundo vector generado es un vector de pesos (ver figura 6.1(d)), donde cada

[1]La visualización directa de datos consiste en una representación gráfica del conjunto de datos que permite una comprensión cualitativa de la información de manera natural e intuitiva [DKZ13]. Entre los métodos más utilizados se encuentran las matrices de diagramas de dispersión [CLN87], las coordenadas paralelas [Ins85], las caras de Chernoff [Che73] y el apilamiento dimensional [LWW90], entre otros. En estos enfoques, cada variable que caracteriza a los ítems se representa en una forma visual comprensible para el ser humano.

componente corresponde a una configuración visual y refleja la prioridad que el usuario le asignó a dicha configuración.

La asignación de valores se lleva a cabo mediante un sistema de pesos escalonados, lo que posibilita que el usuario asigne diferentes niveles de importancia a cada elemento de la configuración visual. Por ejemplo, si el usuario considera que las marcas son de máxima relevancia y las clasifica en la categoría de mayor peso (peso 6), el vector asignará automáticamente este valor numérico a los componentes correspondientes a esas marcas. Por el contrario, aquellas configuraciones que el usuario considera de menor relevancia, como los canales visuales, por ejemplo, se asignan al peso 0, lo que indica que, según el sistema, esos componentes no tienen prioridad en la visualización, reflejándose en el vector con un valor de 0. En el caso presentado en la figura 6.1 el usuario no expresa preferencia particular sobre el tipo de canal visual a utilizar (color, tamaño, posición, etc.), pero si manifiesta de forma explícita su preferencia de que se utilicen líneas como marcas.

3. **Clasificación**

   La etapa de clasificación dentro del sistema **CVRS** se centra en la selección de las técnicas de visualización, un aspecto clave del tercer pilar del proceso de visualización, el "*¿Cómo?*". En esta fase, el sistema identifica las técnicas más adecuadas para comunicar la información de manera efectiva, considerando las preferencias y prioridades definidas previamente por el usuario.

   Los vectores generados (el **vector de configuración** y el **vector de pesos**) son utilizados para comparar las preferencias del usuario con las configuraciones admitidas por cada técnica de visualización disponible en el sistema. Cada técnica de visualización tiene su propio *bitcode* de configuración, que es una representación binaria que describe las configuraciones que dicha técnica admite. En la tabla 6.2 se resumen las técnicas disponibles en esta primera implementación de **CVRS**, junto con las configuraciones que cada una de ellas es capaz de soportar.

   El primer paso es verificar si alguna técnica de visualización admite en su totalidad las preferencias definidas por el usuario. Para ello, se compara el vector de configuración del usuario con los vectores de configuración de las técnicas. Si alguna técnica responde a la configuración del usuario (es decir, su *bitcode* de configuración

coincide con al menos todos los valores 1 presentes en el vector de configuración del usuario), entonces esa técnica se clasifica en el grupo de "*Exact Match*" (ver figura 6.2(a)). Esto significa que la técnica responde en su totalidad a las preferencias del usuario.

El siguiente paso es identificar las técnicas que no coinciden en absoluto con las preferencias del usuario. Para ello, se revisan las técnicas restantes que no fueron clasificadas en el grupo de "*Exact Match*". Si una técnica no cumple con ninguna de las configuraciones definidas por el usuario (es decir, su *bitcode* de configuración no tiene ningún valor de 1 que coincida con los valores de 1 del vector de configuración del usuario), entonces esta técnica se clasifica en el grupo de "*No Match*" (ver figura 6.2(b)). Es decir, no hay ninguna coincidencia entre lo que la técnica soporta y lo que el usuario prefiere.

Finalmente, para las técnicas que no tienen una coincidencia exacta, pero que sí presentan coincidencias parciales con las preferencias del usuario, se realiza un cálculo de "*Partial Match*". Esto significa que se evalúa qué tan bien se ajusta la configuración de cada técnica a las preferencias del usuario, calculando un índice de coincidencia.

Para calcular el índice de coincidencia, se emplea la distancia de Hamming ponderada (*Weighted Hamming distance*) [ZZT+13]. El cálculo compara el vector de configuración proporcionado por el usuario con el *bitcode* de configuración asociado a cada técnica. La distancia de Hamming tradicional evalúa cuántos bits difieren entre ambos vectores, pero aquí se restringe únicamente a los casos donde el vector de configuración tiene un valor de 1 y el *bitcode* tiene un valor de 0. Este enfoque permite identificar aquellas configuraciones que el usuario considera indispensables, pero que la técnica no es capaz de soportar. En contraste, no se toman en cuenta los casos en los que el *bitcode* tiene un valor de 1 y el vector de configuración tiene un valor de 0, ya que representan configuraciones adicionales soportadas por la técnica que no son requeridas por el usuario. Luego, para reflejar la importancia de cada configuración, se pondera esa distancia utilizando el vector de pesos.

Para cada bit en el vector de configuración y en el *bitcode* de configuración de la técnica, se compara si los valores son iguales o diferentes. Si los valores son

Algoritmo 1 Cálculo de la Distancia de Hamming Ponderada

Entrada:
  $w = [w_1, w_2, \ldots, w_n]$ ▷ Vector de pesos
  $U = [u_1, u_2, \ldots, u_n]$ ▷ Vector de configuración del usuario
  $T_i = [t_{i1}, t_{i2}, \ldots, t_{in}]$ ▷ Vector de configuración de la técnica i, para $i = 0$ a k
Salida:
  $D(U, T_i)$ ▷ Distancia ponderada entre U y T_i
para cada i desde 0 hasta k **realizar**
    $D(U, T_i) \leftarrow 0$ ▷ Inicializar la distancia ponderada para la técnica i
    para cada bit j desde 1 hasta n **realizar**
      si $u_j = 1 \land t_{ij} \neq 1$ **entonces**
        $D(U, T_i) \leftarrow D(U, T_i) + w_j$ ▷ Sumar el peso w_j si los bits son diferentes
      fin si
    fin para
fin para
Retornar $D(U, T_i)$ ▷ Distancia ponderada calculada para cada técnica i

diferentes (es decir, el bit del vector de configuración es 1 y el bit de la técnica es 0), se multiplica la diferencia por el peso correspondiente del vector de pesos (ver figura 6.2(c)). Luego, se suman todas las diferencias ponderadas. Cuanto mayor sea el valor de la distancia de Hamming ponderada, menor será la coincidencia entre la configuración definida por el usuario y la configuración soportada por la técnica de visualización. Este valor permite evaluar qué tan bien se ajusta la configuración de cada técnica a las preferencias del usuario.

6.6. Espacio de Trabajo de CVRS

La interfaz de usuario de **CVRS** se estructuró en dos pantallas principales: la pantalla de configuración y la pantalla del *dashboard* principal. Los controles de navegación, situados en la esquina superior derecha de la aplicación (ver figura 6.3(a)), permiten alternar entre ambas pantallas.

La pantalla de configuración constituye la etapa inicial en la interacción del usuario con la aplicación, proporcionando un entorno para la preparación de los datos, la definición de

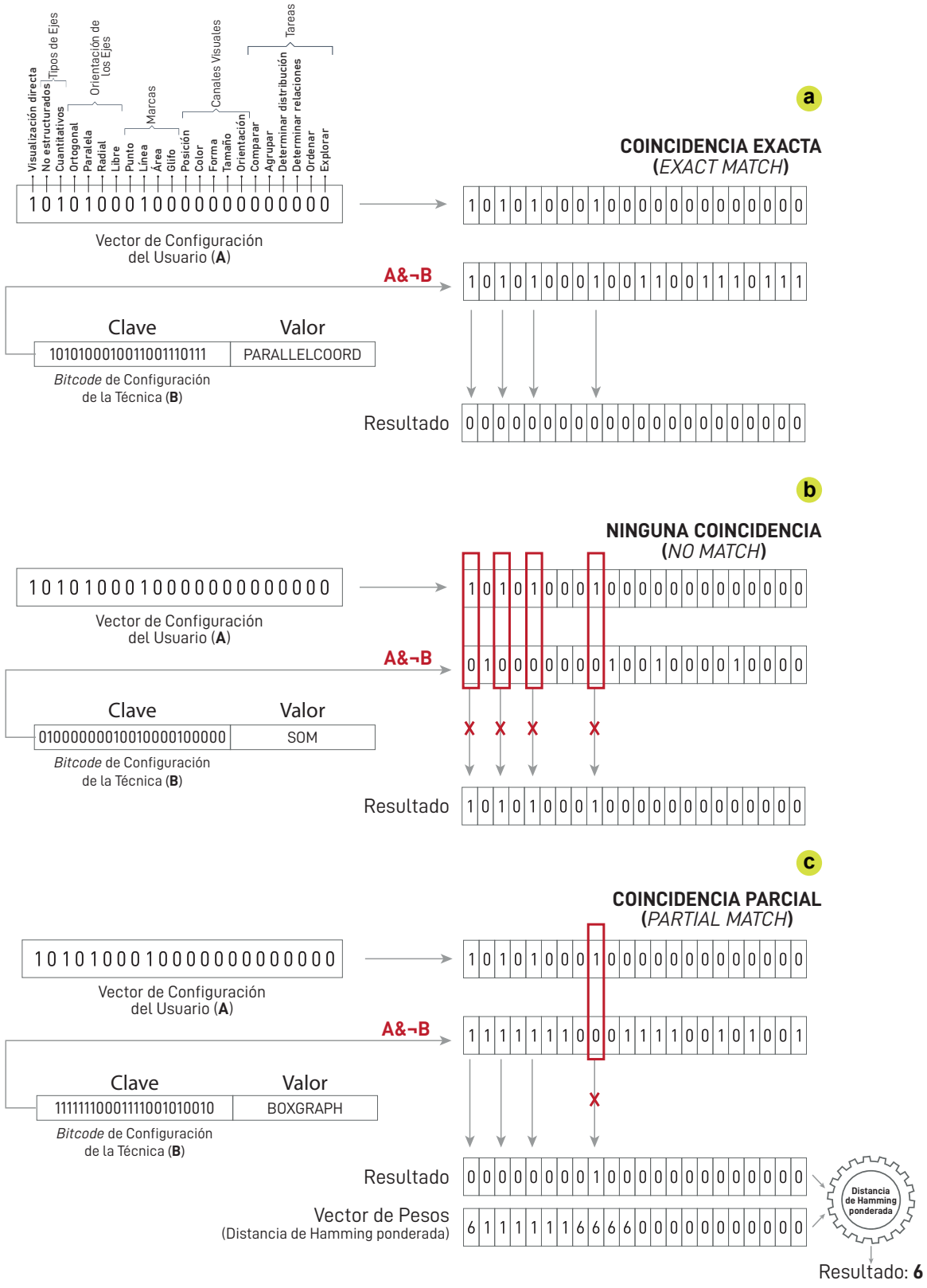


Figura 6.2: Etapa de clasificación en **CVRS**. (a) Exact match: la técnica de visualización soporta en su totalidad las preferencias definidas por el usuario. (b) No match: la técnica no cumple con ninguna de las configuraciones definidas por el usuario. (c) Partial match: Para las técnicas con coincidencias parciales con las preferencias del usuario, se calcula un índice de coincidencia utilizando la distancia de Hamming ponderada (Weighted Hamming distance).

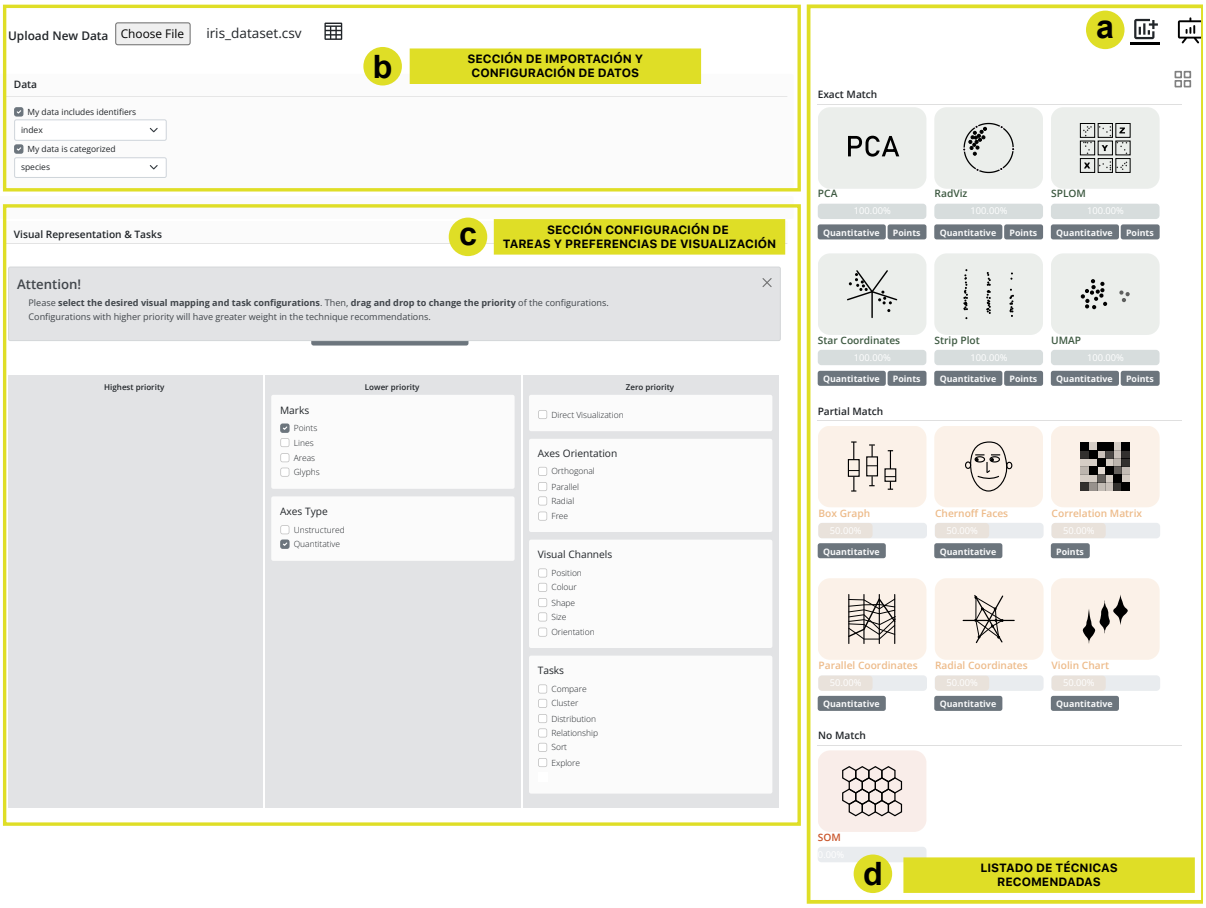


Figura 6.3: Interfaz de usuario de **CVRS**. (a) Controles de navegación para alternar entre la pantalla de configuración y la pantalla del *dashboard* principal. (b) Sección de importación y configuración de datos. (c) Sección de configuración de tareas y preferencias de visualización. (d) Listado de técnicas recomendadas agrupadas en tres categorías según el nivel de coincidencia de su configuración con las preferencias del usuario.

tareas analíticas y preferencias de visualización, así como para la selección de las técnicas de visualización. Esta pantalla se estructura en tres secciones:

1. *Sección de Importación y Configuración de Datos*

   Esta sección permite al usuario cargar su conjunto de datos en formato CSV (*.csv). El sistema clasifica automáticamente las columnas del conjunto de datos importado en atributos categóricos y numéricos. Posteriormente, en la sección de configuración de datos (ver figura 6.3(b)), el usuario puede especificar manualmente si el conjunto de datos incluye identificadores únicos para las muestras y variables categóricas que describan grupos o categorías.

Figura 6.4: Pantalla del *dashboard* principal de **CVRS** con múltiples vistas coordinadas. Se ha cargado el conjunto de datos *Iris* [Fis88] y se ha representado mediante múltiples vistas, incluyendo: (a) una matriz de diagramas de dispersión, (b) coordenadas paralelas, (c) UMAP y (d) coordenadas radiales. (e) Utilizando la interacción *Brushing & Linking*, se ha aplicado una selección sobre uno de los ejes de la visualización de coordenadas paralelas para filtrar las muestras con un largo de sépalo (*sepal length*) inferior a 6. Esta acción resalta de manera automática los elementos correspondientes en todas las demás vistas.

2. *Sección de Configuración de Tareas y Preferencias de Visualización*

   En esta sección, el usuario define los objetivos principales de su análisis y selecciona sus preferencias de representación visual. El sistema incluye la "lista de prioridad", a través de la cual el usuario elige las configuraciones a incluir en su visualización y establece el orden de preferencia de cada elemento del mapeo visual.

   Por ejemplo, el usuario puede asignar mayor prioridad a las marcas y al tipo de ejes, mientras que considera de menor importancia los canales visuales o la orientación de los ejes (ver figura 6.3(c)). La lista de prioridad está organizada en una serie de casillas ordenadas de mayor a menor prioridad, permitiendo al usuario "arrastrar" los distintos elementos de configuración entre las casillas para ajustar su importancia en el mapeo visual.

3. *Listado de Técnicas Recomendadas*

   Finalmente, en esta última sección, se presentan al usuario las técnicas de visualización soportadas con la aplicación (ver figura 6.3(d)). Las técnicas se agrupan en tres categorías según el nivel de coincidencia de su configuración con las preferencias proporcionadas por el usuario. Se muestra el porcentaje de coincidencia junto con las etiquetas de la configuración del usuario que son compatibles con la técnica. Al seleccionar una técnica, se presenta una descripción detallada acompañada de ejemplos ilustrativos.

Por otro lado, la pantalla del *dashboard* principal incorpora todas las vistas que el usuario ha añadido desde la pantalla de configuración. Todas las vistas son interactivas, están vinculadas entre sí e integradas en un sistema de múltiples vistas coordinadas (ver figura 6.4). La técnica de múltiples vistas coordinadas es un enfoque consolidado en el análisis visual [Rob07]. Su idea central consiste en mostrar un conjunto de datos simultáneamente mediante distintas técnicas de visualización, ofreciendo diversas perspectivas de los datos y permitiendo la implementación de los conceptos de *Brushing & Linking*. Como vimos en la sección 4.3, el *Brushing* permite al usuario seleccionar interactivamente (resaltar) subconjuntos de datos en una vista, mientras que el *Linking* garantiza que se resalten de manera coherente los elementos correspondientes en todas las vistas coordinadas. En la figura 6.4, se ha aplicado una selección sobre uno de los ejes de la visualización de coordenadas paralelas para filtrar las muestras con un largo

de sépalo (*sepal length*) inferior a 6. Como resultado, los elementos correspondientes se resaltan automáticamente en todas las demás vistas.

6.7. Limitaciones

Si bien el sistema **CVRS** ofrece un marco conceptual sólido y una arquitectura flexible para la recomendación de técnicas de visualización, presenta ciertas limitaciones que deben ser reconocidas tanto en términos de diseño como de alcance funcional.

En primer lugar, **CVRS** se basa en decisiones explícitas del usuario para configurar los parámetros de visualización, tales como las tareas analíticas, las marcas y canales visuales, y los tipos de ejes. Esta elección deliberada de diseño, ofrece transparencia, control e interpretabilidad sobre las recomendaciones, y se implementa mediante un sistema de clasificación basado en coincidencia de configuraciones. Sin embargo, este enfoque puede resultar demandante para usuarios con escasa experiencia en análisis visual. Por esta razón, como línea de trabajo futuro se plantea la incorporación de una funcionalidad de visualización automática que, sin requerir decisiones manuales, sea capaz de generar propuestas iniciales considerando tanto la naturaleza del conjunto de datos como, eventualmente, información derivada de historiales de uso.

Además, el sistema se basa en una estructura de datos tabular que, si bien es ampliamente utilizada y versátil, impone restricciones sobre los tipos de datos que pueden ser procesados. **CVRS** requiere al menos tres atributos cuantitativos y, opcionalmente, hasta dos atributos categóricos. Esto impide su aplicación en conjuntos de datos que no cumplen con esta estructura mínima o que presentan otras estructuras como redes, jerarquías o datos temporales. En este sentido, se prevé como línea de trabajo futura la extensión del sistema para soportar otro tipo de estructuras, ampliando así su aplicabilidad a un conjunto más diverso de dominios y problemas analíticos.

Por último, **CVRS**, en su estado actual, solo considera un conjunto fijo de técnicas de visualización. Si bien estas técnicas han sido cuidadosamente seleccionadas para cubrir un amplio rango de tareas analíticas, la escalabilidad del sistema frente a la incorporación de nuevas técnicas aún requiere ser evaluada, especialmente en términos de rendimiento y rediseño de la interfaz para mantener la claridad y usabilidad. Cabe destacar que **CVRS** fue concebido bajo un marco modular, lo que facilita la integración incremental de nuevas

técnicas y permite abordar de forma flexible los desafíos asociados a la extensibilidad del sistema.

Estas limitaciones delinean un conjunto claro de oportunidades para futuras extensiones del sistema, orientadas a optimizar la experiencia del usuario y ampliar el alcance funcional y técnico de **CVRS** hacia nuevos tipos de datos y contextos de aplicación.

6.8. Conclusiones

En este capítulo se presenta **CVRS**, un Sistema de Recomendación Integral de Visualización creado para abordar los desafíos de diseñar visualizaciones efectivas. Además, con el objetivo de facilitar la interpretación y aplicación de las tareas, se propone una taxonomía unificada que supera las diferencias terminológicas presentes en la literatura, en la que diversas taxonomías describen tareas conceptualmente similares utilizando terminologías diferentes.

CVRS se distingue por abordar de manera integral tres pilares clave: las características de los datos, las tareas analíticas y las preferencias de representación visual. Su arquitectura flexible se organiza en tres etapas fundamentales: preprocesamiento de los datos, configuración de tareas y de preferencias de visualización, y clasificación de técnicas. El sistema emplea un algoritmo de ponderación basado en la distancia de Hamming para seleccionar y recomendar las técnicas de visualización más adecuadas, según las necesidades y preferencias particulares de cada usuario. Además, incorpora un sistema de validación de datos y soporte para múltiples vistas coordinadas, lo que facilita un análisis más completo y desde diferentes perspectivas. Su enfoque centrado en el usuario posibilita una personalización detallada de las visualizaciones a través de una interfaz intuitiva, permitiendo asignar prioridades a los elementos visuales según el contexto y las necesidades específicas del usuario.

Este sistema representa un avance significativo en el campo de la visualización de datos al combinar flexibilidad con un sólido fundamento metodológico. Su arquitectura establece una base sólida para futuras mejoras y desarrollos en los sistemas de recomendación. La implementación actual del sistema muestra que es posible combinar la automatización con el control del usuario, permitiendo a usuarios expertos como no expertos crear visualizaciones expresivas y efectivas.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 7

Casos de Estudio

7.1. Introducción

La visualización constituye una herramienta fundamental para detectar, analizar e interpretar posibles relaciones entre los datos [Cle93], especialmente en contextos multidimensionales como los presentes en las Ciencias Geológicas. En el marco de esta tesis, se desarrollaron diversas herramientas de visualización orientadas a la exploración y análisis de datos complejos, con aplicaciones específicas en este campo y en otros dominios.

Este capítulo presenta los sistemas desarrollados, destacando *Spinel Web* [AGF+21] y *SpinelVA* [ALF+24], herramientas diseñadas para el análisis visual de datos geoquímicos y la clasificación de minerales del grupo de los espinelos mediante la integración de técnicas avanzadas de aprendizaje automático con enfoques tradicionales de visualización. Ambos sistemas fueron desarrollados en colaboración interdisciplinaria entre el VyGLab (Laboratorio de Investigación en Visualización y Computación Gráfica) del Departamento de Ciencias e Ingeniería de la Computación de la Universidad Nacional del Sur y el INGEOSUR (Instituto Geológico del Sur, UNS-CONICET). También, se presentan herramientas de carácter más general, como VISUEL [AGC22], *npGLC-Vis* [LGAC21], *GLC-Frame* y *GLC-Vis* [LAGC24], que son adecuadas para la exploración y el análisis visual de datos multidimensionales en diversos dominios.

Este capítulo incluye casos de estudio que ilustran la aplicación de estas herramientas en contextos reales, subrayando su capacidad para manejar múltiples dimensiones y relaciones complejas. Este enfoque no solo contribuye al avance en el ámbito de las Ciencias Geológicas, sino que también presenta un considerable potencial para su aplicación en

otros dominios que requieran el análisis de datos multidimensionales.

7.2. Visualización de Datos Multidimensionales

En diversas disciplinas, se presentan problemas que requieren el manejo de grandes volúmenes de datos. A medida que los conjuntos de datos aumentan en tamaño y dimensionalidad, las técnicas tradicionales, diseñadas para conjuntos de datos de menor dimensión, a menudo resultan insuficientes para capturar adecuadamente la estructura de la información [CMS99, War19, Spe07, WGK10].

Por esta razón, resulta de suma importancia disponer de un conjunto de metáforas visuales y de técnicas de visualización asociadas que faciliten el análisis de datos multidimensionales. Con este propósito, hemos desarrollado diversas herramientas orientadas al análisis visual de dichos tipos de datos. En 2022, se presentó *VISUEL* [AGC22], una aplicación de acceso libre que integra un conjunto de técnicas de visualización coordinadas en un único *dashboard*. Esta herramienta puede ser empleada en diversas disciplinas y ofrece a los usuarios la posibilidad de explorar un mismo conjunto de datos a través de distintas representaciones visuales. *VISUEL* es totalmente interactivo y permite al usuario configurar la representación visual de sus datos.

Además, se han desarrollado diversas herramientas que integran métodos de visualización reversibles y sin pérdida de información para datos multidimensionales, como *npGLC-Vis* [LGAC21], una librería que soporta distintas representaciones de las *Non-Paired General Line Coordinates* (NP-GLC) y facilita interacciones tradicionales como selección, escalado y desplazamiento. *GLC-Vis* [LAGC24], por su parte, amplía las capacidades de *npGLC-Vis*, incorporando soporte para las *Paired General Line Coordinates* (P-GLC). Además, presentamos *GLC-Frame* [LAGC24], una herramienta web de exploración que permite cargar conjuntos de datos multidimensionales y explorar interactivamente distintas configuraciones GLC.

7.2.1. *VISUEL*

VISUEL es una herramienta web gratuita, desarrollada para facilitar el análisis de conjuntos de datos multidimensionales [AGC22]. Esta aplicación integra diversas técnicas de visualización dentro de un entorno dinámico e interactivo. Permite a los usuarios

cargar sus propios conjuntos de datos y personalizar las visualizaciones de acuerdo con sus necesidades y preferencias específicas.

Las técnicas de visualización soportadas están diseñadas para abordar una amplia gama de necesidades analíticas. Los gráficos de dispersión y los diagramas de caja son útiles para explorar distribuciones y detectar valores atípicos [SG18, Tuk77]. Las coordenadas paralelas y radiales permiten representar simultáneamente varias dimensiones en un espacio bidimensional [ID09]. Los mapas coropléticos y de burbujas permiten visualizar datos enriquecidos con información geoespacial.

VISUEL es una herramienta completamente interactiva. Entre las interacciones que soporta, destaca el *brushing & linking*, que permite seleccionar subconjuntos específicos de datos en una vista para un análisis más detallado, asegurando que los elementos correspondientes se resalten de manera coherente en las demás vistas coordinadas. Además, las interacciones de *zoom* y navegación permiten una exploración detallada de las visualizaciones, mientras que la capacidad de reordenar los ejes en los diagramas de coordenadas paralelas optimiza el análisis de las relaciones entre dimensiones. Además, ofrece a los usuarios la posibilidad de personalizar la representación visual de sus datos. Esto incluye la selección de canales visuales como el color y la forma de las marcas visuales, permitiendo a los usuarios adaptar las representaciones a sus necesidades específicas y preferencias visuales.

Un caso práctico que ilustra el potencial de *VISUEL* es su aplicación en el análisis de datos de la industria vitivinícola argentina. A través de visualizaciones como mapas coropléticos, mapas de burbujas y gráficos de barras, se han identificado tendencias clave en la producción, consumo y exportación de vino en diferentes regiones del país. Por ejemplo, las distribuciones espaciales de los viñedos y bodegas, representadas mediante mapas interactivos, revelan patrones significativos en la concentración de actividades vitivinícolas en provincias como Mendoza y San Juan. Asimismo, la integración de gráficos de series temporales ha permitido analizar la evolución de la superficie cultivada y la producción de vino a lo largo de los años, destacando el impacto de factores climáticos y económicos en la actividad vitivinícola.

En conclusión, *VISUEL* es una herramienta de fácil acceso e innovadora que integra múltiples técnicas avanzadas de visualización en una interfaz altamente interactiva. Su capacidad para integrar múltiples vistas coordinadas y su flexibilidad en la personalización

de visualizaciones la convierten en una solución adecuada para la exploración y el análisis de datos en diversos campos.

7.2.2. *Coordenadas Generales de Líneas (GLC)*

Aunque se han desarrollado numerosos métodos de visualización para conjuntos de datos n-dimensionales, muchos de ellos no son reversibles y/o presentan pérdidas de información, lo que significa que no es posible restaurar los datos completos a partir de su representación reducida, o bien, no representan el conjunto de datos n-dimensional en su totalidad.

En el campo de la visualización de datos, estos desafíos se han abordado mediante el desarrollo de representaciones visuales reversibles y sin pérdida de información, entre las cuales destacan las coordenadas paralelas y las radiales [ID87, ID09, DLR09]. Ambas técnicas han sido utilizadas durante muchos años y han demostrado su efectividad; no obstante, presentan el problema de la oclusión. En este contexto, han surgido las coordenadas generales de líneas, conocidas como GLC, por sus siglas en inglés (*General Lines Coordinates*) [Kov14, Kov18, KG19]. Las GLC hacen referencia a diversas alternativas para la visualización de conjuntos de datos n-dimensionales en 2 o 3 dimensiones, de forma reversible y sin pérdida de información. Se distinguen dos tipos de GLC: las coordenadas generales de líneas no emparejadas (NP-GLC, por sus siglas en inglés: *Non-Paired General Line Coordinates*) y las coordenadas generales de líneas emparejadas (P-GLC, por sus siglas en inglés: *Paired General Line Coordinates*). Las NP-GLC generalizan las coordenadas paralelas y radiales, incluyendo las coordenadas *N-Gon*, *Circular*, *In-Line*, *Dynamic*, y *Bush Coordinates*. Las P-GLC generalizan las coordenadas cartesianas, incluyendo las *Paired Orthogonal*, *Non-orthogonal*, *Collocated*, *Partially Collocated*, *Shifted*, *Radial*, *Elliptic*, y *Crown Coordinates*. Kovalerchuk, en diversos trabajos, describe las GLC y sus ventajas, además de incluir el análisis de varios conjuntos de datos [Kov14, Kov18].

En este contexto, llevamos a cabo una revisión sistemática de los artículos existentes en la literatura relacionados con técnicas de visualización NP-GLC en 2D [ALGC23]. En particular, nos centramos en las NP-GLC, que generalizan las coordenadas paralelas y radiales, y que han demostrado ser altamente adecuadas para visualizar datos multidimensionales. Optamos por no incluir las técnicas P-GLC debido no solo al escaso número de contribuciones en esta área específica, sino también a que aún no se han consolidado

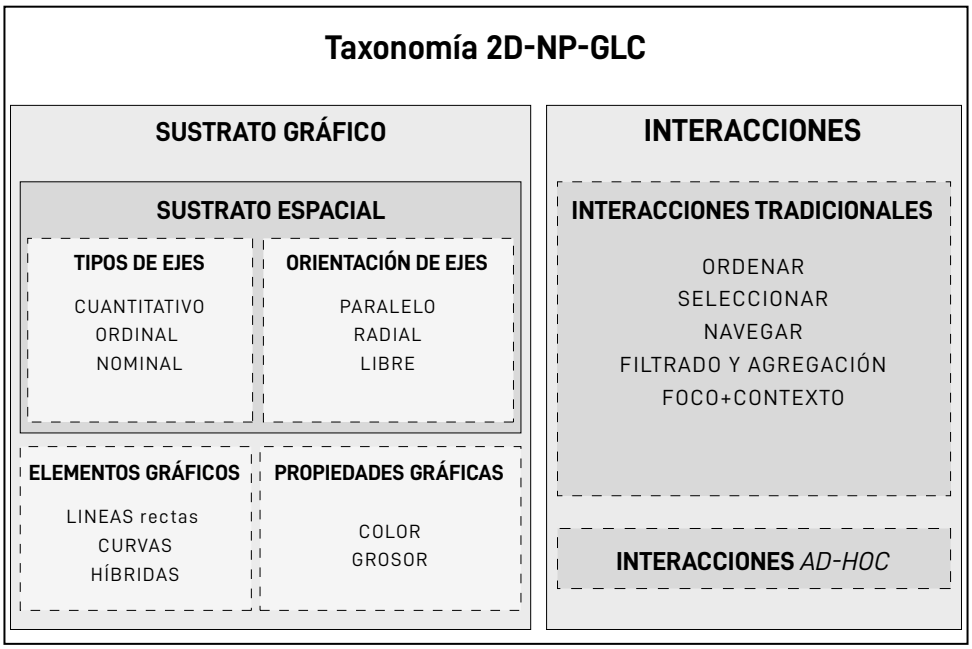


Figura 7.1: Taxonomía 2D-NP-GLC que considera tanto la estructura visual de las técnicas como las interacciones que estas soportan. Figura adaptada de [ALGC23].

como enfoques establecidos para la visualización de datos multidimensionales.

Para estructurar y analizar las contribuciones, las organizamos dentro de un marco de referencia unificado que considera tanto la estructura visual de las técnicas (ver la sección 4.2) como las interacciones que éstas soportan. Para definir la estructura visual, ampliamos la organización de Card *et al.* [CMS99] para incluir en el sustrato espacial los diseños de ejes asociados con las coordenadas NP-GLC. En cuanto a las interacciones, clasificamos la literatura recopilada en dos grupos principales: aquellos que aplican interacciones tradicionales y aquellos que definen interacciones *ad-hoc* diseñadas específicamente para su técnica NP-GLC. En base a todo lo mencionado, hemos definido una taxonomía denominada 2D-NP-GLC, que abarca todas las características capaces de describir cualquier técnica de visualización NP-GLC en 2D, así como las interacciones asociadas (ver figura 7.1). Este enfoque integrado permite evaluar las técnicas considerando estos dos criterios y, al mismo tiempo, establecer un marco de referencia que no solo guía el diseño y desarrollo de nuevas soluciones NP-GLC en 2D, sino que también facilita la identificación de áreas poco exploradas que podrían ser candidatas para futuras investigaciones.

Durante la revisión, identificamos que a lo largo del tiempo, las contribuciones han mostrado características comunes que permiten distinguir diferentes períodos en su de-

sarrollo. Las primeras contribuciones se enfocan en métodos basados en líneas como una forma eficiente de visualizar datos multidimensionales [ID90], junto con diversas variantes diseñadas para abordar, desde una perspectiva de implementación, problemas asociados al desorden visual [FWR99, AdOL04]. Ante las limitaciones impuestas por el espacio en pantalla, la siguiente etapa se caracterizó por el uso de métodos para reducir el número de polilíneas cuando los usuarios manejan grandes cantidades de datos. Estos métodos incluyen técnicas de agrupamiento, como SVD, k-means y jerárquico, así como enfoques visuales [ZYQ+08]. Por su parte, la tercera etapa se centra en el manejo de grandes conjuntos de datos para la exploración interactiva, donde las interacciones, como la navegación y el filtrado, resultan clave para manipular datos y facilitar el proceso iterativo de obtención de conocimiento. En las coordenadas paralelas, estas interacciones suelen aplicarse sobre los ejes (atributos de datos) o las polilíneas (elementos de datos). Durante esta etapa, Kovalerchuk junto a otros autores [Kov14, Kov18, KG19], publicaron una serie de trabajos relacionados con la mejora de la representación de datos multidimensionales mediante polilíneas y nuevos diseños de coordenadas (GLC). La última etapa, corresponde a la introducción de técnicas de *Machine Learning* para mejorar o complementar la calidad de los resultados que el usuario puede obtener de los métodos visuales [PKC19]. La mayoría de las contribuciones en esta etapa se enfocan en identificar qué dimensiones son clave para describir el comportamiento de los datos y determinar el orden óptimo de estas dimensiones, con el fin de evitar el desorden [LHH12].

Hasta entonces, algunos de los tipos de NP-GLC definidos por Kovalerchuk [Kov18], no tenían aplicaciones ni usos publicados en la literatura académica, aunque era posible encontrar gráficos y menciones en fuentes no académicas. En este contexto, consideramos relevante el desarrollo de una librería, llamada *npGLC-Vis*, que permita aplicar estos métodos en el análisis de grandes conjuntos de datos [LGAC21]. *npGLC-Vis* es una herramienta integral que soporta la representación de diferentes conjuntos de datos utilizando todos los métodos de coordenadas generales de líneas no emparejadas definidos por Kovalerchuk [Kov18]. Además, la librería permite interacciones tradicionales como selección (*Brushing*), escalado (*Zooming*) y desplazamiento (*Panning*), lo que facilita una exploración iterativa e interactiva del espacio de datos.

Para el diseño de la librería, se consideraron todas las representaciones de NP-GLC, y se elaboró una clasificación propia, en donde cada representación de GLC tiene una

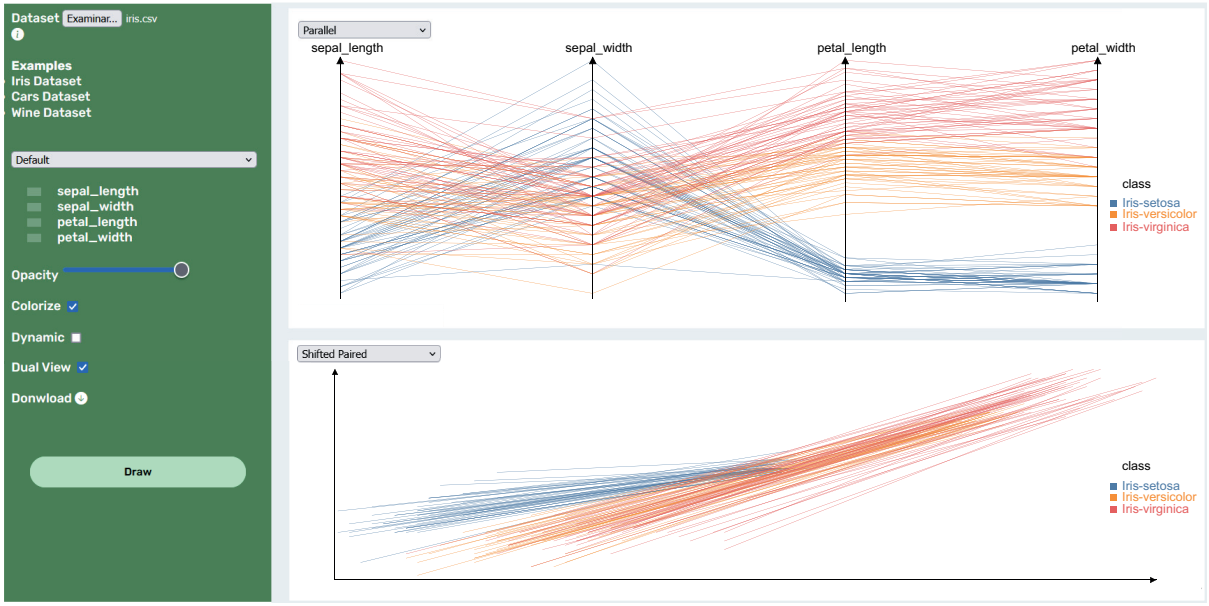


Figura 7.2: La interfaz de usuario de *GLC-Frame* [LAGC24] se divide en dos secciones: la sección de configuración (a la izquierda) y la de visualización (a la derecha). La sección de configuración permite cargar un conjunto de datos personalizado o utilizar los predeterminados. También ofrece opciones para reordenar los ejes y visualizar los elementos de los datos. Por su parte, la sección de visualización contiene paneles con las técnicas GLC seleccionadas, junto con sus opciones de manipulación.

colección asociada de ejes junto con su disposición, una colección de elementos de datos y un atributo que describe si los elementos de datos se dibujarán de manera estática o dinámica. Cada eje en la colección de ejes tiene asociados sus atributos de dirección, rango, tipo, escala, longitud y posición. Cada elemento de datos en la colección de elementos puede ser representado con curvas o líneas rectas.

Sin lugar a dudas, *npGLC-Vis* constituyó un excelente punto de partida para la exploración de las técnicas GLC; no obstante, la necesidad de una librería unificada que también ofreciera soporte para las técnicas GLC emparejadas seguía siendo una necesidad pendiente. Por ello, presentamos *GLC-Vis*, una librería de visualización de datos de código abierto basada en la web, que admite todas las técnicas GLC en 2D e interacciones tradicionales [LAGC24]. Esta librería permite a los usuarios crear técnicas de visualización personalizadas y es altamente compatible con diversos navegadores web.

También presentamos *GLC-Frame*, una herramienta web de exploración que permite cargar conjuntos de datos multidimensionales y explorar interactivamente diferentes configuraciones GLC sin necesidad de escribir código. Esta herramienta ofrece un mecanismo

de vista dual que muestra un conjunto de datos bajo dos representaciones diferentes: una técnica NP-GLC y una técnica P-GLC (ver figura 7.2). Esta funcionalidad posibilita a los usuarios comparar visualmente el mismo conjunto de datos desde múltiples perspectivas, lo que resulta extremadamente útil para obtener una comprensión más profunda e identificar patrones que podrían no ser evidentes con una sola representación. *GLC-Frame* soporta un amplio conjunto de interacciones, entre las cuales se incluyen la selección de la técnica de visualización, el reordenamiento de ejes, la modificación de canales visuales como opacidad y color, el filtrado por clases, la activación de la vista dual, así como escalado (*Zooming*) y desplazamiento (*Panning*). Además, se incorpora la interacción de *Brushing & Linking* entre estas vistas.

7.3. Visualización de Datos Geoquímicos

En diversas áreas, incluidas las Ciencias Geológicas, se generan datos multidimensionales. En particular, los geólogos enfrentan desafíos significativos en el procesamiento y análisis visual de datos multidimensionales, impulsados por el notable incremento, en las últimas décadas, tanto en la cantidad de análisis realizados como de elementos determinados. Los geólogos suelen trabajar con rocas que tienen hasta varios miles de millones de años de antigüedad. Un desafío importante para ellos es caracterizar una región geológica en función de su ambiente geotectónico. Definir el ambiente tectónico ayuda a los geólogos a comprender la evolución geológica de la Tierra y a localizar, por ejemplo, depósitos minerales. Para lograr estos objetivos, los geólogos utilizan, entre otros criterios, la composición química de los minerales del grupo de los espinelos. Debido a que son muy sensibles a las condiciones que prevalecen durante la cristalización de las rocas y muy resistentes a las modificaciones químicas posteriores a la cristalización, los minerales del grupo de los espinelos proporcionan información valiosa sobre el entorno geológico en el que se formaron las rocas que contienen estos minerales [Lin91, Roe94, BR01]. Por ello, constituyen excelentes indicadores de entornos geológicos y son de ayuda invaluable en la búsqueda de depósitos minerales de interés económico.

Los datos correspondientes a los minerales del grupo de los espinelos plantean importantes desafíos para su visualización, especialmente debido a su naturaleza multidimensional. Para cada muestra de espinelo se analizan 11 elementos químicos mayoritarios

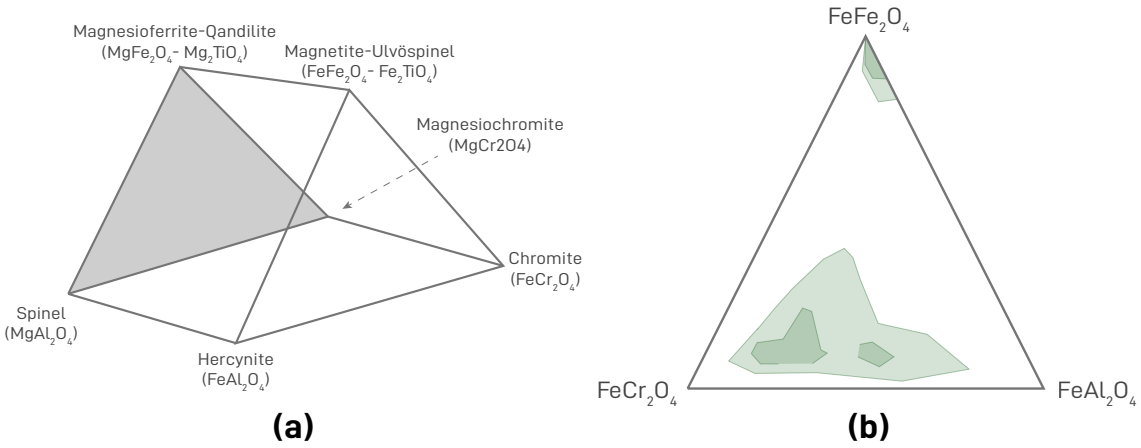


Figura 7.3: (a) Prismas de Espinelos (b) Proyección triangular del Prisma de Magnetita. Muestra los contornos de densidad del campo composicional de basaltos definido por Barnes y Roeder [BR01]. Figura adaptada de [GGF+15].

(aquellos cuyo contenido es superior al 0,01 % en peso), a partir de los cuales se obtienen 22 metadatos, denominados miembros finales, que son combinaciones de estos elementos. Aunque existen técnicas de visualización capaces de manejar grandes conjuntos de datos multidimensionales en distintos contextos, el análisis geoquímico tradicional se lleva a cabo utilizando gráficos especiales que analizan solo un subconjunto limitado de estos atributos a la vez. Sólo 8 de estos miembros finales son usados habitualmente por los geólogos para su representación en gráficos específicos [BR01].

Uno de los gráficos específicos empleados por los geólogos es el Prisma de Espinelos (ver figura 7.3(a)). En este prisma se representa cada análisis químico obtenido. El Prisma de Espinelos admite dos representaciones: el Prisma de Magnetita y el Prisma de Ulvöespinelo. El Prisma de Magnetita se utiliza para representar las composiciones químicas de la solución sólida integrada por los miembros finales *Hercynite-Spinel-Magnesioferrite-Magnetite-Magnesiochromite-Chromite*. Por su parte, el Prisma de Ulvöespinelo se emplea para graficar las composiciones químicas de la solución sólida representada por los miembros finales *Hercynite-Spinel-Ulvöspinel-Qandilite-Magnesiochromite-Chromite*. Adicionalmente, el prisma puede dividirse en distintos campos que representan los distintos ambientes tectónicos. Los análisis que corresponden a espinelos provenientes de un determinado ambiente tectónico se agrupan en un patrón de referencia determinado y único, en una determinada región del prisma [CDLS97, Lin91]. Para evaluar las relaciones entre los elementos geoquímicos se analizan tanto los diagramas prismáticos 3D como la

información proyectada sobre las caras del prisma.

En 2001, Barnes y Roeder [BR01] compilaron una base de datos correspondiente a más de 26000 análisis de espinelos de rocas ígneas y metamórficas de todo el mundo, que les permitió caracterizar cerca de 40 ambientes tectónicos posibles. Para cada uno de estos ambientes tectónicos se extrajeron campos composicionales característicos y se construyeron gráficos con contornos, que suelen ser utilizados por los geólogos para estimar el ambiente tectónico donde una muestra de espinelo podría haberse formado. Para determinarlo, se digitalizan los gráficos de los contornos de Barnes y Roeder para luego poder compararlos manualmente con los diagramas generados previamente. Esta comparación es exhaustiva y debe realizarse para cada uno de los ambientes tectónicos y proyecciones sobre las caras del prisma. Es claro que son tareas propensas a errores y sumamente tediosas y por lo tanto, excelentes candidatas a ser automatizadas. En la figura 7.3(b) se muestra la proyección triangular del Prisma de Magnetita con el contorno de densidad del campo composicional de basaltos definido por Barnes y Roeder [BR01]. Para este campo de composición en particular, se definieron dos campos para los contornos del percentil 90 (verde claro) y tres campos para los contornos del percentil 50 (verde oscuro).

A lo largo de los años se han desarrollado diversas técnicas de visualización de datos multidimensionales que han probado ser de gran utilidad en diversos contextos [PEF02, EH11, ELP+16], y en particular, en el contexto de las Ciencias Geológicas, se han desarrollado varios proyectos [AMB+03, MRY+04] como IGPet [CG17], MinPet [Ric95], Tri Plot [GM00], Tern-Plot [Mar96], etc. Sin embargo, en lo que respecta a la visualización de espinelos, ninguna de las aplicaciones mencionadas integra las herramientas utilizadas en el flujo tradicional de análisis de espinelos sin la necesidad de recurrir a herramientas alternativas de análisis o diagramado. Profesionales del Departamento de Ciencias e Ingeniería de la Computación y del Departamento de Geología de la Universidad Nacional del Sur han trabajado en conjunto en temas relacionados con la Visualización de Datos aplicada a las Geociencias en general, y a los minerales del grupo de los espinelos en particular [GCF+12, GFG+14, GGF+15, FGG+15, GFG+17]. Estos esfuerzos han sido fundamentales para el avance y desarrollo de herramientas que responden a las necesidades específicas de este campo.

Por esta razón, es crucial avanzar en el diseño y desarrollo de métodos de visualización apropiados para datos multidimensionales, que faciliten al geólogo la integración y evalua-

ción de toda la información obtenida durante los análisis. La consideración conjunta de todos estos datos podría mejorar la comprensión de la interacción entre los procesos geológicos responsables de la formación de ambientes geológicos, las composiciones de minerales y rocas, y el entorno geotectónico de una región específica.

En este contexto, se desarrollaron dos herramientas para la visualización de minerales del grupo de los espinelos. En 2021 se presentó *Spinel Web* [AGF+21], una herramienta interactiva para explorar la composición química de estos minerales. Incluye gráficos binarios, diagramas ternarios, proyecciones del Prisma de Espinelos y coordenadas paralelas, permitiendo un análisis visual multidimensional. Su funcionalidad principal es la categorización semi-automática de ambientes tectónicos mediante los volúmenes generados a partir de la base de datos definida por Barnes y Roeder [BR01]. Posteriormente, en 2024, se desarrolló *SpinelVA* [ALF+24], una aplicación web que integra métodos de visualización interactiva, reducción de dimensionalidad y *Machine Learning* para clasificar minerales del grupo de los espinelos. La herramienta utiliza vistas coordinadas y un modelo de clasificación que considera tanto los cationes considerados por Barnes y Roeder como otros atributos, con el objetivo de apoyar el análisis visual de datos geoquímicos y la identificación de ambientes tectónicos.

7.3.1. *Spinel Web*

Spinel Web [AGF+21] es una herramienta interactiva web de análisis visual, diseñada para explorar la composición química de los minerales del grupo de los espinelos, aportando soluciones a los desafíos en la exploración de datos geoquímicos multidimensionales. Este sistema integra gráficos binarios en 2D, diagramas ternarios, representaciones tridimensionales del Prisma de Espinelos y coordenadas paralelas, permitiendo un análisis visual completo.

Una de las características principales de *Spinel Web* es su capacidad para categorizar de manera semi-automática los ambientes tectónicos de formación de un conjunto de datos de minerales del grupo de los espinelos. Esta categorización se realiza mediante la comparación de los volúmenes generados a partir de la base de datos definida por Barnes y Roeder [BR01] con los volúmenes correspondientes a los datos del usuario. Una contribución significativa de este trabajo fue la generación de estos volúmenes a partir de los contornos bidimensionales originales de Barnes y Roeder, proporcionando por primera

vez a la comunidad geológica una representación espacial completa de estos dominios composicionales permitiendo la comparación con los datos del usuario. Estos volúmenes se superponen a los Prismas de Espinelos en 3D y sus respectivas proyecciones, permitiendo identificar patrones composicionales distintivos de diferentes entornos geológicos. Además, el usuario tiene la posibilidad de generar contornos basados en la densidad para su propio conjunto de datos y compararlos con los contornos definidos por Barnes y Roeder.

La aplicación emplea el principio de múltiples vistas coordinadas, un enfoque consolidado en el análisis visual que permite representar un conjunto de datos desde diversas perspectivas mediante diferentes métodos de visualización. En este enfoque, las interacciones realizadas en una vista específica se reflejan de manera inmediata en las vistas vinculadas, lo que garantiza una visualización coherente y completa de los datos.

Desde el marco de los tres pilares fundamentales de la visualización (ver sección 2), *Spinel Web* se destaca como una herramienta que abarca de manera integral el *"¿Qué?-¿Por qué?-¿Cómo?"* de la visualización de datos geoquímicos.

- ¿Qué?: *Spinel Web* permite a los usuarios cargar sus propios conjuntos de datos, obtenidos a partir de análisis de minerales del grupo de los espinelos. Estos datos incluyen elementos químicos principales expresados como óxidos, y valores calculados como los miembros finales, que representan combinaciones químicas específicas. La flexibilidad de la herramienta para aceptar datos personalizados en formato CSV, preprocesados con herramientas como EMG [FGG+15], facilita su adaptación a las necesidades específicas de cada investigador.

- ¿Por qué?: *Spinel Web* está diseñado para respaldar tareas específicas relacionadas con el análisis visual de datos geoquímicos. Entre sus objetivos principales se encuentra la identificación de ambientes tectónicos definidos a partir de la composición química de los minerales. Esto se logra mediante la superposición de contornos de Barnes y Roeder [BR01], la generación de nuevos contornos basados en densidades de datos y la categorización semi-automática de las muestras analizadas.

- ¿Cómo?: *Spinel Web* ofrece múltiples técnicas de visualización, como diagramas binarios en 2D, ternarios y representaciones tridimensionales de los Prismas de Espinelos (Magnetita y Ulvöespinelo), junto con coordenadas paralelas. Estas técnicas permiten a los usuarios elegir cómo visualizar sus datos, según las relaciones

químicas que deseen destacar. Además, la plataforma permite al usuario configurar canales visuales según sus preferencias, como el color y la forma de las marcas. *Spinel Web* admite múltiples vistas coordinas e interacciones como *Brushing & Linking*, donde los usuarios pueden seleccionar subconjuntos de datos en una vista y observar cómo estas selecciones se reflejan en todas las vistas coordinadas. Esta capacidad de interacción proporciona un gran poder analítico porque permite examinar y explorar los datos desde distintas perspectivas.

7.3.1.1. Ejemplo de Aplicación

En esta sección se muestran los resultados del análisis visual de los datos químicos de oxiespinelos provenientes de xenolitos del manto recolectados cerca de Comallo, en la provincia de Río Negro, Patagonia Argentina. Este caso de estudio se presentó como parte de la contribución al 14° Congreso de Minerología, Petrología Ígnea y Metamórfica, y Metalogénesis (MinMet), celebrado en la Universidad Nacional del Sur en 2023 [AFG+23]. Los detalles completos de este caso de estudio pueden consultarse en el artículo correspondiente.

El objetivo fue emplear diagramas tradicionales 2D, gráficos de coordenadas paralelas y Primas de Espinelos, junto con sus proyecciones, para llevar a cabo un análisis visual de las composiciones químicas de las muestras analizadas. Para este análisis, inicialmente se utilizaron los datos químicos obtenidos mediante microsonda electrónica y el programa EMG [FGG+15] para el cálculo de los miembros finales de los minerales del grupo de los espinelos. Posteriormente, se elaboró la tabla de datos de entrada para *Spinel Web*, siguiendo las especificaciones detalladas en [AGF+21].

Para iniciar el análisis visual, se cargó el archivo recopilado en el sistema y se configuró la vista de coordenadas paralelas, seleccionando tantos ejes como miembros finales se representan en el Prisma de Magnetita (ver figura 7.4(f)). Al seleccionar las muestras donde la proporción de Spinel ($MgAl_2O_4$) está entre el 35 % y el 45 %, y el Magnesiochromite ($MgCr_2O_4$) oscila entre el 25 % y el 35 %, se observó que podría existir un intercambio entre Al y Cr en estas muestras. Para verificarlo, se construyó un diagrama binario Al_2O_3 vs. Cr_2O_3, evidenciando una tendencia lineal negativa entre estos elementos (ver figura 7.4(c)). Además, debido a la coordinación de las vistas, se dedujo que los datos con mayor proporción de Magnesiochromite ($MgCr_2O_4$) presentaban mayores niveles de

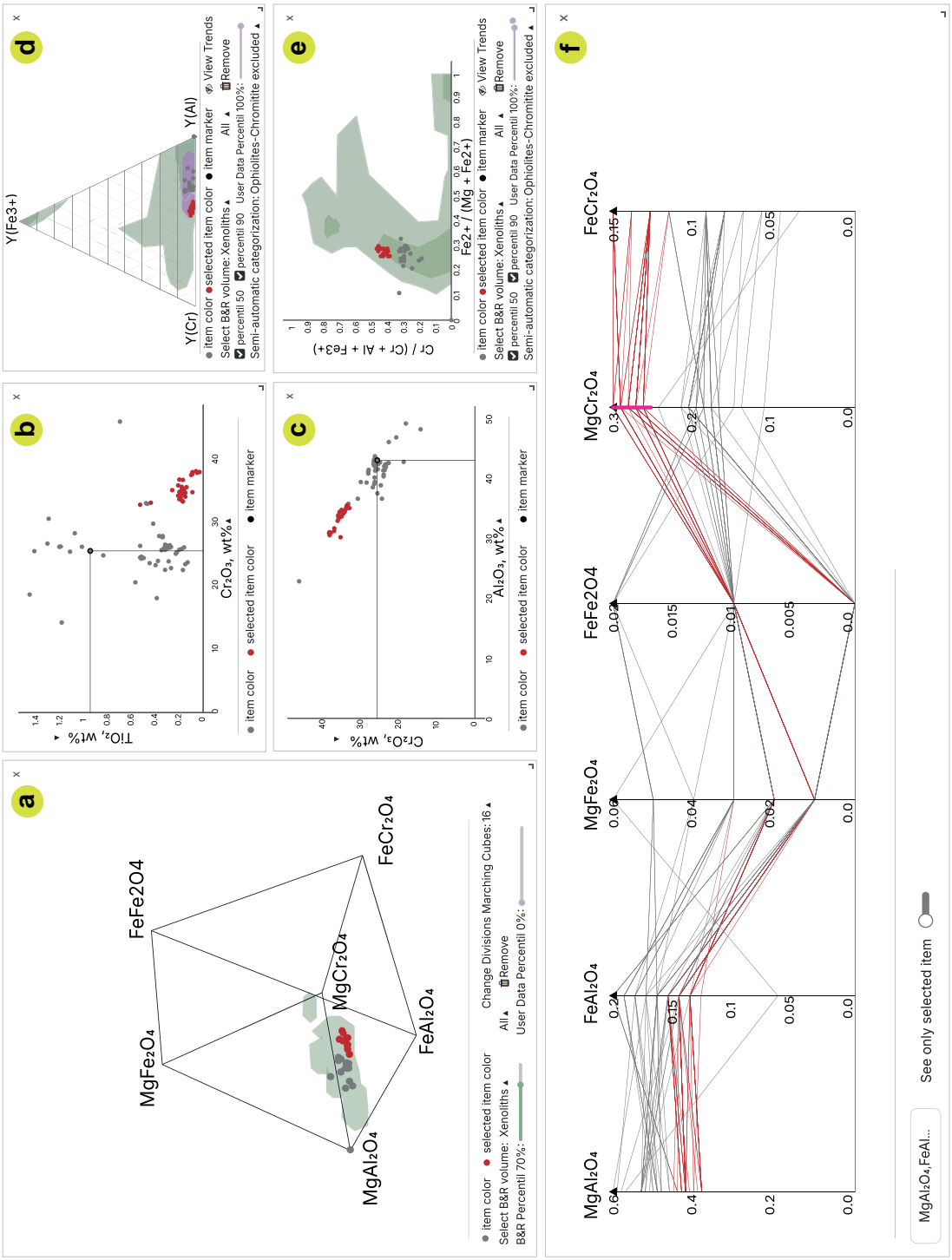


Figura 7.4: Sesión de análisis en *Spinel Web* [AGF+21]. Una selección en fucsia en el eje del miembro final Magnesiochromite ($MgCr_2O_4$) sobre (f) el diagrama de coordenadas paralelas resalta en rojo las muestras correspondientes a ese intervalo, no solo en este diagrama, sino también en los demás diagramas visualizados: (a) el Prisma de Magnetita, los diagramas binarios (b) Cr_2O_3 vs. TiO_2 y (c) Al_2O_3 vs. Cr_2O_3, y (d) y (e) las proyecciones del prisma. En lila se muestra el contorno de densidad del conjunto de datos analizados, y en verde se superponen los contornos de densidad del campo composicional de xenolitos definido por Barnes y Roeder [BR01].

Cr_2O_3 y menores de Al_2O_3.

Con el fin de detectar patrones adicionales en otros atributos, se observó que las muestras con mayores contenidos de Magnesiochromite ($MgCr_2O_4$) corresponden predominantemente a lherzolitas con textura equigranular tabular, cuyas concentraciones de TiO_2 son menores al 0,30 % en peso. Para comprobar esta tendencia, se realizó un gráfico binario Cr_2O_3 vs. TiO_2 (ver figura 7.4(b)), con el que se confirmó que el contenido de TiO_2 es reducido en estas muestras.

En cuanto al análisis en el diagrama de coordenadas paralelas, se observó que los miembros finales de mayor proporción se localizan en la base del prisma. Esto se verificó en la distribución del conjunto de datos en el Prisma de Magnetita (ver figura 7.4(a)), donde se agrupan en el extremo magnesiano del prisma, entre Spinel ($MgAl_2O_4$) y Magnesiochromite ($MgCr_2O_4$), y en el diagrama binario correspondiente a la proyección en la base del prisma (ver figura 7.4(e)). La figura 7.4(d) muestra la proyección triangular del prisma, destacando con puntos rojos las muestras que tienen una alta proporción de Magnesiochromite ($MgCr_2O_4$).

Spinel Web también permite trazar contornos de densidad para los datos analizados, los cuales se muestran en color lila. Estos contornos pueden superponerse con los contornos de densidad en color verde, que corresponden a los campos de categorización propuestos por Barnes y Roeder [BR01]. Esto muestra que el conjunto de datos se ajusta al campo de los xenolitos (ver figura 7.4(d)). Además, es posible generar el volumen correspondiente al campo de los xenolitos en la vista en 3D del Prisma de Magnetita para contrastar con el conjunto de datos analizado (ver figura 7.4(a)).

En resumen, *Spinel Web* ofrece una herramienta integral para visualizar los diferentes diagramas realizados, con selecciones coordinadas entre ellos, facilitando así un análisis visual eficiente de las características químicas de las muestras de oxiespinelos estudiadas.

7.3.2. *SpinelVA*

SpinelVA [ALF+24] es una aplicación web que integra técnicas tradicionales de análisis visual con métodos de aprendizaje automático, optimizando la exploración, clasificación y análisis de los minerales del grupo de los espinelos.

La aplicación permite visualizar los datos mediante múltiples vistas coordinadas, que incluyen proyecciones del Prisma de Magnetita, diagramas de reducción dimensional basa-

dos en UMAP [LJJ18], gráficos de distribución de elementos químicos y tablas exploratorias. Estas vistas están sincronizadas, de modo que cualquier interacción en una de ellas se refleja automáticamente en las demás, facilitando la exploración simultánea de diversas dimensiones y relaciones entre los datos.

Una de las ventajas de *SpinelVA* es la mejora en la precisión al categorizar los ambientes tectónicos de formación de minerales del grupo de los espinelos, al integrar la experiencia humana con técnicas avanzadas de visualización y aprendizaje automático. Esto facilita la exploración simultánea de múltiples dimensiones y relaciones. Se realizó una evaluación de varios modelos (por ejemplo, máquina de soporte vectorial, redes neuronales, árboles de decisión, etc.) para determinar cuál era el más eficiente en términos de precisión y rendimiento. Se obtuvo como resultado que Random Forest [Bre01], al considerar tanto los cationes empleados por Barnes y Roeder [BR01] como otros atributos derivados de análisis químicos, proporcionaba resultados precisos sin añadir complejidad al análisis. Este modelo nos permite generar un *ranking* de los ambientes tectónicos más probables para cada muestra, asistiendo a los geólogos en la identificación del ambiente tectónico de formación de las muestras analizadas.

Además, *SpinelVA* integra técnicas de Reducción de Dimensionalidad (DR) para mostrar de manera sencilla las estructuras entre los datos mediante representaciones 2D. Utiliza UMAP [LJJ18] como técnica principal para representar la agrupación de categorías y los atributos compartidos. El sistema ofrece dos configuraciones de UMAP: una que incluye dimensiones relacionadas con ratios ($Ti\#$, $Cr\#$, $Fe^{3+}\#$, y $Fe^{2+}\#$), cationes (Ti, Cr, Fe^{2+}, Fe^{3+}, MgO, y Al) y sus óxidos, y otra que agrega información de los ocho miembros finales ($MgAl_2O_4$, $FeAl_2O_4$, $MgFe_2O_4$, $FeFe_2O_4$, $MgCr_2O_4$, $FeCr_2O_4$, Mg_2TiO_4, Fe_2TiO_4). Los usuarios pueden colorear las muestras según el ambiente tectónico o un atributo seleccionado, facilitando la visualización de patrones químicos. También es posible seleccionar puntos usando laso o clic individual, y las vistas se complementan con una vista enfocada para un análisis más detallado.

Por último, podemos analizar *SpinelVA* en el marco de los tres pilares fundamentales de la visualización de datos (ver sección 2):

- ¿Qué?: *SpinelVA* ofrece la posibilidad de cargar conjuntos de datos personalizados en formato CSV que incluyan cationes, óxidos y miembros finales. Este enfoque asegura que los datos analizados sean directamente relevantes para los objetivos del

usuario.

- ¿Por qué?: La aplicación permite a los geólogos categorizar los ambientes tectónicos de formación de minerales del grupo de los espinelos, utilizando un modelo de clasificación que incorpora métodos de aprendizaje automático y que considera tanto los cationes utilizados por Barnes y Roeder [BR01] como otros atributos. Además, las tareas exploratorias, como la identificación de agrupaciones o el análisis de distribuciones químicas, se ven mejoradas por las herramientas de reducción de dimensionalidad y las opciones de selección avanzada disponibles en la aplicación.

- ¿Cómo?: *SpinelVA* integra múltiples representaciones gráficas interactivas, como proyecciones del Prisma de Magnetita, diagramas de reducción de dimensionalidad, gráficos de distribución de elementos químicos y tablas exploratorias. Todas las vistas están coordinadas y admiten interacciones, como *Brushing & Linking*, para explorar relaciones y patrones dentro de los datos.

7.3.2.1. Ejemplo de Aplicación

En este ejemplo, se presenta brevemente la categorización de muestras de minerales del grupo de los espinelos utilizando el sistema *SpinelVA*, un análisis que fue presentado como caso de estudio en el trabajo de Antonini *et al.* [ALF+24]. Los detalles completos de este análisis pueden consultarse en el artículo correspondiente.

El conjunto de datos empleado corresponde a espinelos provenientes de xenolitos recolectados en diversas localidades de la Patagonia Argentina, con un total de 693 muestras. Cada muestra contiene 60 atributos, que incluyen valores de óxidos, cationes y miembros finales. En este contexto, el objetivo del análisis es verificar la hipótesis de los geólogos de que las muestras provienen de xenolitos del manto terrestre.

La figura 7.5 muestra una captura de pantalla de una sesión de análisis en la que los geólogos interactúan con diferentes visualizaciones del conjunto de datos. En primer lugar, el sistema se configura para representar la base de datos mediante (a) diagramas de reducción de dimensionalidad y (b) gráficos triangulares y de dispersión correspondientes a las proyecciones del Prisma de Espinelos. En estos gráficos, se observa que una parte significativa de las muestras es clasificada como xenolitos. Además, los geólogos superponen los contornos de los xenolitos definidos por Barnes y Roeder sobre las proyecciones

del Prisma de Espinelos (ver figura 7.5(b)). Esta superposición confirma que gran parte de las muestras que se seleccionaron en el diagrama de reducción de dimensionalidad se sitúan dentro del percentil 50 del campo correspondiente a los xenolitos.

No obstante, algunas muestras son categorizadas como ofiolitas o basaltos. Esta situación podría explicarse por la similitud en los rangos de composición de los espinelos en ciertas categorías, lo que puede ocasionar clasificaciones no adecuadas. Para profundizar en este aspecto, se investigaron en mayor detalle aquellas muestras clasificadas como basaltos (ver figura 7.6). Para ello, se superpusieron los contornos correspondientes a los basaltos (línea delgada) y a los xenolitos (línea gruesa) sobre (b) las proyecciones del Prisma de Espinelos. Este análisis revela que la mayoría de los puntos destacados del conjunto de datos se localizan en la intersección de ambos contornos.

Los expertos continuaron la exploración revisando la información cuantitativa disponible en la tabla de exploración (ver figura 7.6(c)). En esta tabla, se observa que las muestras seleccionadas por el geólogo (destacadas en gris en la columna #), específicamente las comprendidas entre los números #157 y #162, fueron clasificadas inicialmente como basaltos por el modelo, aunque la categoría de xenolitos aparece entre las tres principales sugerencias de clasificación.

En resumen, *SpinelVA* integra técnicas avanzadas de visualización y aprendizaje automático para clasificar muestras de minerales del grupo de los espinelos según su composición química. Su interfaz interactiva, que incluye vistas coordinadas, permite explorar los datos desde diversas perspectivas, facilitando un análisis geológico más eficiente. Gracias a su capacidad para agilizar este proceso, *SpinelVA* se presenta como una herramienta valiosa para geólogos e investigadores en el estudio de los minerales.

7.3.3. *Machine Learning* Interpretable para la Clasificación de Rocas Ígneas

La adopción de técnicas de aprendizaje automático en diversos campos puede ser recibida con reticencia, debido a que muchos de estos métodos actúan como "cajas negras", lo que limita la transparencia y la interpretabilidad de sus procesos. La falta de claridad sobre el proceso de toma de decisiones puede generar desconfianza en los usuarios, quienes a menudo encuentran dificultades para comprender los fundamentos detrás de los resultados obtenidos. En este contexto, se desarrolló un enfoque innovador [ATA+24]

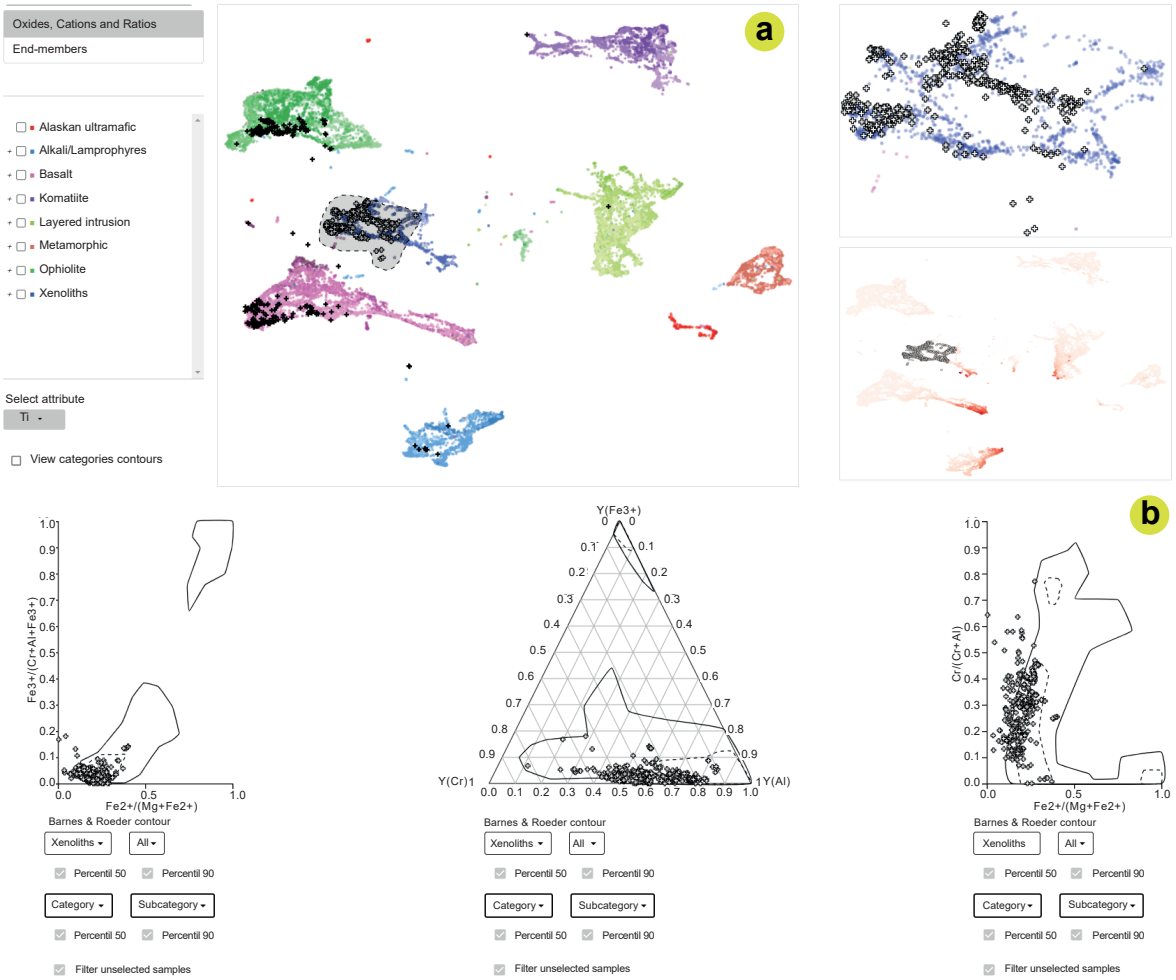
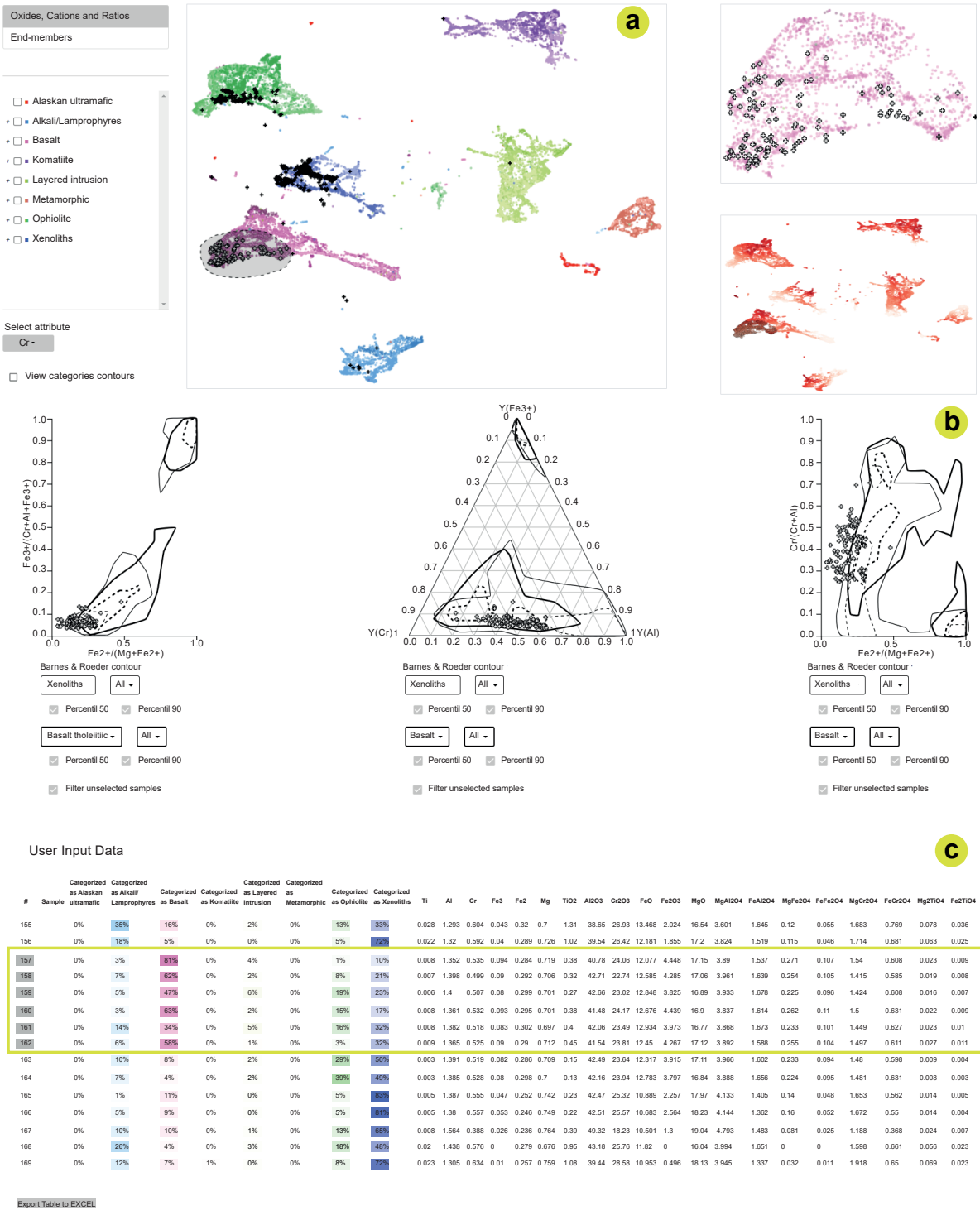


Figura 7.5: Sesión de análisis en *SpinelVA*[ALF^{+}24] con selección de muestras clasificadas como xenolitos. Se muestran 693 muestras de espinelos provenientes de xenolitos recolectados en la Patagonia Argentina, representadas mediante (a) diagramas de reducción de dimensionalidad y (b) gráficos triangulares y de dispersión correspondientes a las proyecciones del Prisma de Espinelos. Además, en (b) se superponen los contornos de los xenolitos definidos por Barnes y Roeder[BR01].



#	Sample	Categorized as Alaskan ultramafic	Categorized as Alkali/ Lamprophyres	Categorized as Basalt	Categorized as Komatiite	Categorized as Layered intrusion	Categorized as Metamorphic	Categorized as Ophiolite	Categorized as Xenoliths	Ti	Al	Cr	Fe3	Fe2	Mg	TiO2	Al2O3	Cr2O3	FeO	Fe2O3	MgO	MgAl2O4	FeAl2O4	MgFe2O4	FeFe2O4	MgCr2O4	FeCr2O4	Mg2TiO4	Fe2TiO4
155		0%	35%	16%	0%	2%	0%	13%	33%	0.028	1.293	0.604	0.043	0.32	0.7	1.31	38.65	26.93	13.468	2.024	16.54	3.601	1.645	0.12	0.055	1.683	0.769	0.078	0.036
156		0%	18%	5%	0%	0%	0%	5%	72%	0.022	1.32	0.592	0.04	0.289	0.726	1.02	39.54	26.42	12.181	1.855	17.2	3.824	1.519	0.115	0.046	1.714	0.681	0.063	0.025
157		0%	3%	81%	0%	4%	0%	1%	10%	0.008	1.352	0.535	0.094	0.284	0.719	0.38	40.78	24.06	12.077	4.448	17.15	3.89	1.537	0.271	0.107	1.54	0.608	0.023	0.009
158		0%	7%	62%	0%	2%	0%	8%	21%	0.007	1.398	0.499	0.09	0.292	0.706	0.32	42.71	22.74	12.585	4.285	17.06	3.961	1.639	0.254	0.105	1.415	0.585	0.019	0.008
159		0%	5%	47%	0%	6%	0%	19%	23%	0.006	1.4	0.507	0.08	0.299	0.701	0.27	42.66	23.02	12.848	3.825	16.89	3.933	1.678	0.225	0.096	1.424	0.608	0.016	0.007
160		0%	3%	63%	0%	2%	0%	15%	17%	0.008	1.361	0.532	0.093	0.295	0.701	0.38	41.48	24.17	12.676	4.439	16.9	3.837	1.614	0.262	0.11	1.5	0.631	0.022	0.009
161		0%	14%	34%	0%	5%	0%	16%	32%	0.008	1.382	0.518	0.083	0.302	0.697	0.4	42.06	23.49	12.934	3.973	16.77	3.868	1.673	0.233	0.101	1.449	0.627	0.023	0.01
162		0%	6%	58%	0%	1%	0%	3%	32%	0.009	1.365	0.525	0.09	0.29	0.712	0.45	41.54	23.81	12.45	4.267	17.12	3.892	1.588	0.255	0.104	1.497	0.611	0.027	0.011
163		0%	10%	8%	0%	2%	0%	29%	50%	0.003	1.391	0.519	0.082	0.286	0.709	0.15	42.49	23.64	12.317	3.915	17.11	3.966	1.602	0.233	0.094	1.48	0.598	0.009	0.004
164		0%	7%	4%	0%	2%	0%	39%	49%	0.003	1.385	0.528	0.08	0.298	0.7	0.13	42.16	23.94	12.783	3.797	16.84	3.888	1.656	0.224	0.095	1.481	0.631	0.008	0.003
165		0%	1%	11%	0%	0%	0%	5%	83%	0.005	1.387	0.555	0.047	0.252	0.742	0.23	42.47	25.32	10.889	2.257	17.97	4.133	1.405	0.14	0.048	1.653	0.562	0.014	0.005
166		0%	5%	9%	0%	0%	0%	5%	81%	0.005	1.38	0.557	0.053	0.246	0.749	0.22	42.51	25.57	10.683	2.564	18.23	4.144	1.362	0.16	0.052	1.672	0.55	0.014	0.004
167		0%	10%	10%	0%	1%	0%	13%	65%	0.008	1.564	0.388	0.026	0.236	0.764	0.39	49.32	18.23	10.501	1.3	19.04	4.793	1.483	0.081	0.025	1.188	0.368	0.024	0.007
168		0%	26%	4%	0%	3%	0%	18%	48%	0.02	1.438	0.576	0	0.279	0.676	0.95	43.18	25.76	11.82	0	16.04	3.994	1.651	0	0	1.598	0.661	0.056	0.023
169		0%	12%	7%	1%	0%	0%	8%	72%	0.023	1.305	0.634	0.01	0.257	0.759	1.08	39.44	28.58	10.953	0.496	18.13	3.945	1.337	0.032	0.011	1.918	0.65	0.069	0.023

Figura 7.6: Sesión de análisis en *SpinelVA* [ALF$^+$24] con (a) selección de muestras clasificadas como basaltos. (b) Se superponen los contornos de los basaltos (línea delgada) y los xenolitos (línea gruesa) sobre las proyecciones del Prisma de Espinelos de Barnes y Roeder. (c) La tabla de exploración muestra que las muestras seleccionadas (resaltadas en gris en la columna #) fueron clasificadas como basaltos, aunque los xenolitos figuran entre las tres categorías principales. Figura adaptada de [ALF$^+$24].

para conocer la naturaleza máfica o ultramáfica de las rocas observadas en los testigos de perforación en el cuerpo intrusivo El Fierro, uno de los cuerpos que componen la faja máfica-ultramáfica La Jovita–Las Águilas, ubicada en la Sierra Grande de San Luis, Argentina. Este enfoque se centra en la aplicación de técnicas de aprendizaje automático, en particular del algoritmo *Random Forest* [Bre01], combinado con el método SHAP (*SHapley Additive exPlanations*) [LL17], con el objetivo de mejorar tanto la precisión como la interpretabilidad de los modelos de clasificación geológica.

De este modo, los geólogos lograron una alta precisión en la clasificación, al mismo tiempo que ofrecieron interpretaciones claras y comprensibles de los factores que influían en las decisiones del modelo. El método SHAP permitió identificar las variables más relevantes en las predicciones, como el óxido de aluminio (Al_2O_3), el óxido de magnesio (MgO) y el estroncio (Sr). Estos resultados coinciden con el conocimiento geológico previo, ya que dichos elementos están fuertemente asociados con las diferencias mineralógicas entre las rocas máficas y ultramáficas.

El uso de técnicas de aprendizaje automático en las Ciencias Geológicas supone un avance significativo en relación con los métodos tradicionales de clasificación, que a menudo se basan en descripciones macroscópicas imprecisas. La implementación del algoritmo *Random Forest* mejora la precisión y optimiza la clasificación de las muestras. Además, el método SHAP permite comprender las contribuciones de cada variable geoquímica en la clasificación de las rocas, lo que puede orientar futuras investigaciones y exploraciones. Este enfoque tiene el potencial de convertirse en una herramienta valiosa en las Ciencias Geológicas, complementando los métodos tradicionales y facilitando nuevos descubrimientos en la clasificación y caracterización de rocas.

7.4. Conclusiones

En este capítulo, se han presentado el diseño e implementación de diversas herramientas y enfoques desarrollados para abordar los desafíos inherentes al análisis y visualización de datos multidimensionales, particularmente en el contexto de las Ciencias Geológicas. A través de casos de estudio específicos, se ha resaltado el potencial y el impacto de estos sistemas en el análisis de datos multidimensionales.

En el ámbito de la visualización de datos multidimensionales en general, se ha destaca-

do el desarrollo de *VISUEL* [AGC22], un sistema que integra diversas técnicas de visualización en un entorno interactivo, permitiendo a los usuarios explorar grandes volúmenes de datos desde múltiples perspectivas. Su diseño flexible y su capacidad para coordinar vistas han favorecido la identificación de patrones y relaciones en conjuntos de datos multidimensionales, optimizando el análisis exploratorio en distintos dominios. El estudio de las coordenadas generales de líneas ha resultado en la creación de *npGLC-Vis* [LGAC21], *GLC-Vis* [LAGC24] y *GLC-Frame* [LAGC24], herramientas que han permitido avanzar en la representación y exploración de datos mediante técnicas de visualización reversibles y sin pérdida de información. La formalización de una taxonomía, que abarca todas las características capaces de describir cualquier técnica de visualización NP-GLC en 2D y las interacciones asociadas, ha sido clave para estructurar el análisis de estas técnicas [ALGC23]. Estas herramientas han permitido trabajar con distintos métodos y configuraciones de las GLC, evaluando su aplicabilidad en diversos contextos.

En el análisis visual de datos geoquímicos, *Spinel Web* [AGF+21] y *SpinelVA* [ALF+24] fueron fundamentales para simplificar y enriquecer la exploración de los minerales del grupo de los espinelos. Mientras que *Spinel Web* ofrece una plataforma interactiva para la exploración de datos geoquímicos y la categorización semi-automática de ambientes tectónicos, permitiendo la creación de contornos basados en densidad y la superposición de los contornos de Barnes y Roeder en 2D, junto con los volúmenes generados a partir de su base de datos para el análisis y comparación con los volúmenes correspondientes a los datos del usuario, *SpinelVA* integra técnicas de reducción de dimensionalidad y aprendizaje automático para mejorar la precisión y eficiencia en la clasificación de muestras. Finalmente, la implementación de enfoques interpretables de *Machine Learning*, supuso una mejora sustancial en la clasificación de rocas ígneas, combinando precisión y transparencia en el proceso de toma de decisiones. Este enfoque no solo mejora la confianza de los usuarios en los resultados, sino que también proporciona *insights* valiosos que pueden guiar futuras investigaciones geológicas.

En conjunto, estos desarrollos han fortalecido la visualización y el análisis de datos multidimensionales, proporcionando herramientas más eficientes, interactivas y accesibles para la comunidad científica, y estableciendo una base sólida para futuras mejoras en diversas áreas, como las Ciencias Geológicas.

Capítulo 8

Conclusiones y Trabajo a Futuro

En un contexto marcado por la constante generación de grandes volúmenes de datos, la capacidad para analizar y comprender dichos conjuntos complejos se constituye como una competencia crucial en diversos ámbitos.

La visualización, como un puente entre los datos y la interpretación humana, emerge como una herramienta clave para enfrentar estos desafíos. No obstante, el diseño y la selección de representaciones visuales efectivas y expresivas, requieren no solo una comprensión profunda de las características intrínsecas de los datos, sino también una comprensión integral de los objetivos analíticos y las preferencias del usuario.

En este contexto, la presente tesis integra avances conceptuales y prácticos destinados a abordar los desafíos de la visualización de datos multidimensionales. Este trabajo propone soluciones que optimizan el proceso de visualización mediante el desarrollo de un marco metodológico robusto, la realización de un análisis exhaustivo de técnicas de visualización y sistemas de recomendación existentes, y la implementación de herramientas innovadoras. Este enfoque integral no solo mejora la representación y comprensión de los datos multidimensionales, sino que también sienta las bases para futuros avances en el análisis de datos en un entorno cada vez más dinámico y desafiante.

En síntesis, en esta tesis se ha propuesto:

- **Un Marco Metodológico para Estructurar el Proceso de Visualización.**

  Se propuso un marco metodológico fundamentado en las tres preguntas clave propuestas por Munzner [Mun14] *"¿Qué?-¿Por qué?-¿Cómo?"*. Su propósito es proporcionar una estructura clara y sistemática que guíe el proceso de visualización

desde la identificación de las características de los datos hasta la implementación de representaciones visuales efectivas y expresivas.

El marco se complementó con un análisis exhaustivo de los tres pilares fundamentales: los datos, mediante una revisión detallada de su taxonomía; la representación visual, a través del estudio de los componentes que conforman la estructura visual; y las tareas, mediante un relevamiento riguroso de las taxonomías existentes en la literatura y la formulación de una nueva propuesta. Dado que la literatura presenta diversas taxonomías con terminologías distintas para describir tareas que, en esencia, son similares [WL90, AES05, VPF06, Mun14], esta nueva propuesta proporciona un marco unificado que facilita la interpretación y aplicación de las tareas analíticas.

- **Un Relevamiento Exhaustivo de las Técnicas de Visualización para Datos Multidimensionales y Sistemas de Recomendación de Técnicas.**

  Se realizó un análisis exhaustivo de las principales técnicas de visualización utilizadas en el análisis de datos multidimensionales, evaluando sus fortalezas, limitaciones y aplicaciones específicas, y se integraron en una nueva taxonomía que abarca enfoques basados en geometría, íconos, píxeles, jerarquías y grafos. Simultáneamente, se realizó un estudio sobre los sistemas de recomendación de técnicas de visualización presentes en la literatura, en el que se examinaron los métodos empleados y los factores que se tienen en cuenta al seleccionar representaciones visuales. El análisis reveló una amplia gama de enfoques, que van desde sistemas basados en reglas simples hasta modelos más complejos que incorporan múltiples aspectos, como la naturaleza de los datos, los objetivos específicos del análisis, las preferencias del usuario y el conocimiento del dominio.

- **El Diseño y Desarrollo de un Sistema de Recomendación Integral de Visualización.**

  Se diseñó e implementó un Sistema de Recomendación Integral de Visualización (**CVRS**) que considera los tres pilares fundamentales del proceso de visualización. Este sistema se fundamenta en el marco metodológico y en la taxonomía de tareas propuesta, y considera de manera integral la sintaxis y estructura del conjunto de datos, los objetivos analíticos del usuario y sus preferencias de representación visual.

La arquitectura del sistema fue diseñada para ser modular y extensible, facilitando la incorporación de nuevas técnicas de visualización y criterios de recomendación según evolucionen las necesidades de los usuarios.

Este sistema no solo optimiza el proceso de selección, sino que también promueve un análisis más eficiente y personalizado, constituyendo una contribución significativa al campo de la visualización de datos multidimensionales.

- **El Desarrollo de Herramientas y Enfoques para la Visualización y Análisis de Datos Multidimensionales.**

  Se desarrollaron diversas herramientas y enfoques innovadores para abordar los desafíos del análisis y la visualización de datos multidimensionales, con aplicaciones en distintos dominios, especialmente en las Ciencias Geológicas.

  En el contexto de la visualización de datos n-dimensionales en general, se diseñó e implementó *VISUEL* [AGC22], una herramienta que permitió explorar grandes volúmenes de información mediante un entorno interactivo que combina múltiples técnicas de visualización. Su diseño flexible y la integración de múltiples vistas coordinadas facilitó la identificación de patrones y relaciones complejas, optimizando el análisis exploratorio en diversos dominios. Además, dadas sus características, *VISUEL* se presenta como una plataforma ideal de exploración y evaluación de diferentes técnicas de visualización, adaptándose a las necesidades de cada tipo de análisis. Por otro lado, el estudio de las coordenadas generales de líneas impulsó el desarrollo de herramientas como *npGLC-Vis* [LGAC21], *GLC-Vis* [LAGC24] y *GLC-Frame* [LAGC24], enfocadas en la representación de datos a través de técnicas de visualización reversibles y sin pérdida de información. La formalización de una taxonomía específica ha sido fundamental para estructurar y analizar estas técnicas, permitiendo la experimentación con distintas configuraciones y estableciendo un marco metodológico sólido para su aplicación en diferentes contextos.

  En el campo de la Geoquímica, *Spinel Web* [AGF$^+$21] y *SpinelVA* [ALF$^+$24] optimizaron la exploración de minerales del grupo de los espinelos. Mientras que *Spinel Web* permitió la exploración y categorización semi-automática de ambientes tectónicos, la creación de contornos basados en densidad y la superposición de los contornos de Barnes y Roeder en 2D, junto con los volúmenes generados a partir de

su base de datos para el análisis y comparación con los volúmenes correspondientes a los datos del usuario, *SpinelVA* incorporó técnicas de reducción de dimensionalidad y aprendizaje automático para mejorar la precisión y eficiencia en la clasificación de muestras. Finalmente, la implementación de enfoques interpretables de *Machine Learning* supuso un avance significativo en la clasificación de rocas ígneas, combinando precisión y transparencia en la toma de decisiones [ATA+24]. Este enfoque no solo refuerza la confianza en los resultados, sino que también aporta información clave para futuras investigaciones geológicas.

En conjunto, estos desarrollos contribuyeron al avance de la visualización y el análisis de datos multidimensionales, proporcionando herramientas más accesibles, interactivas y efectivas.

8.1. Publicaciones

A continuación se detallan los trabajos científicos más relevantes publicados durante el proceso de desarrollo de este trabajo de investigación:

1. **Publicaciones en Revistas Indexadas**

2024 **Antonini, A.**, Tanzola, J., Asiain, L., Ferracutti, G., Castro, S., Bjerg, E., Ganuza, M. L., *Machine Learning Model Interpretability Using SHAP Values: Application to Igneous Rock Classification Task.* Applied Computing & Geosciences (ELSEVIER). 9 pp., Volume 23, September 2024. DOI: 10.1016/j.acags.2024.100178

Antonini, A., Luque, L., Ferracutti, G., Bjerg, E., Castro, S., Ganuza, M. L., *SpinelVA. A New Perspective for the Visual Analysis and Classification of Spinel Group Minerals.* Earth Science Informatics (Springer), pp. 3851–3861, Volume 17 (4), July 2024. DOI: 10.1007/s12145-024-01393-5

Luque, L., **Antonini, A.**, Ganuza, M. L., Castro, S., *GLC-Frame: A Framework and Library for Exploration of Multidimensional Data with General Line Coordinates.* Journal of Computer Science and Technology, pp. 14-28, Volume 24 (1), April 2024. Online ISSN: 1666-6038 Print ISSN: 1666-6046. DOI: 10.24215/16666038.24.e02.

Tanzola, J., Ferracutti, G., Asiain, L., **Antonini, A.**, Ganuza, M. L., *Applied Geochemistry for the Discrimination Between Mafic and Ultramafic Rocks in Cu-Ni-PGE-bearing Layered Complexes: A Case Study at the la Jovita-Las Águilas Belt, Sierra Grande de San Luis, Argentina.* Journal of South American Earth Sciences, Volume 134, February 2024, 104755. DOI: 10.1016/j.jsames.2023.104755.

Ferracutti, G., Asiain, L., **Antonini, A.**, Tanzola, J., Ganuza, M. L., *OxyEMG: An Application for Determination of the Oxyspinel Group End-members Based on Electron Microprobe Analyses.* European Journal of Mineralogy, pp. 87-98, Volume 36(1), January 2024. DOI: 10.5194/ejm-36-87-2024

2023 Luque, L., **Antonini, A.**, Ganuza, M. L., Castro, S., *Towards a Taxonomy for 2D Non-Paired General Line Coordinates: A Comprehensive Survey.* International Journal of Data Science and Analytics (JDSA), pp. 133–158, Volume 15, March 2023. ISSN: 2364-415X. Springer Nature. DOI: 10.1007/s41060-022-00361-w.

2022 **Antonini, A.**, Ganuza, M. L., Castro, S., *VISUEL - A Web Dynamic Dashboard for Data Visualization.* Journal of Computer Science and Technology. pp. 42-57. Vol. 22(1). April 2022. https://doi.org/10.24215/16666038.22.e03. ISSN: 1666-6038.

2021 **Antonini, A.**, Ganuza, M. L., Ferracuti, G., Gargiulo, M. F., Matkovic, K., Gröller, E., Bjerg, E., Castro, S., *Spinel Web: An Interactive Web Application for Visualizing the Chemical Composition of Spinel Group Minerals*, Earth Sciences Informatics, pp. 521–528, Vol. 14(1), March 2021. DOI: 10.1007/s12145-020-00542-w.

2. **Capítulos de Libros**

2021 Luque, L., Ganuza, M. L., **Antonini, A.**, Castro, S., *npGLC-Vis Library for Multidimensional Data Visualization*, pp. 188-202. In: Cloud Computing, Big Data & Emerging Topics, Communications in Computer and Information Science 1444, IX Jornadas de Cloud Computing, Big Data & Emerging Topics. Selected Papers, Vol. 1444, Chapter 14., Springer. International, 2021. DOI:10.1007/978-3-030-84825-5_14. Print ISBN: 978-3-030-84824-8. Electronic ISBN: 978-3-030-84825-5

3. **Publicaciones en Congresos Nacionales**

2024 Ferracutti, G., Asiain, L., **Antonini, A.**, Ganuza, M. L., Bouhier, V. (2024). *Los Minerales Ígneos del Grupo de los Oxiespinelos ¿son Indicadores Petrogenéticos?* En Actas del XXII Congreso Geológico Argentino.

Antonini, A., Luque, L., Tanzola, J., Asiain, L., Ferracutti, G., Bjerg, E., Castro, S., Ganuza, M. L., *Machine Learning en el Análisis Visual de Minerales del Grupo de los Espinelos*, pp. 230-233, Actas del XXVI Workshop de Investigadores en Ciencias de la Computación – WICC 2024, 18 y 19 de abril de 2024, Puerto Madryn, Chubut, Argentina.

Ganuza, M. L., **Antonini, A.**, Luque, L., Selzer, M., Larrea, M., Tanzola, J., Asiain, L., Ferracutti, G., Gargiulo, M. F., Bjerg, E., Castro, S., *Análisis Visual de Datos Multidimensionales*, pp. 270-274, Actas del XXVI Workshop de Investigadores en Ciencias de la Computación – WICC 2024, 18 y 19 de abril de 2024, Puerto Madryn, Chubut, Argentina.

Larrea, M., Urribarri, D., Ganuza, M. L., Selzer, M., Cobo M. L.,**Antonini, A.**, Vecslir L., *Verificación y Validación de Software en la Industria* 5.0, pp. 370-374, Actas del XXVI Workshop de Investigadores en Ciencias de la Computación – WICC 2024, 18 y 19 de abril de 2024, Puerto Madryn, Chubut, Argentina.

Selzer, M., Ganuza, M. L., **Antonini, A.**, Luque, L., Urribarri, D., Larrea, M., Ferracutti, G., Asiain, L., Bjerg, E., Castro, S., *Innovación Tecnológica para la Exploración Geocientífica: Enfoque en Inmersión y Visualización Situada*, pp. 245-249, Actas del XXVI Workshop de Investigadores en Ciencias de la Computación – WICC 2024, 18 y 19 de abril de 2024, Puerto Madryn, Chubut, Argentina.

2023 **Antonini, A.**, Ferracutti, G., Gargiulo, M. F., Bjerg, E., Castro, S., Ganuza, M. L., *Análisis Visual de Datos Multidimensionales de Química Mineral Correspondientes a Oxiespinelos de Xenolitos del Manto con Spinel Web.* Caso de estudio. pp. 36-42. Actas del Congreso de Mineralogía, Petrología Ígnea y Metamórfica, y Metalogénesis. ISSN 0328-2767. Agosto 2023. Bahía Blanca, Argentina.

Antonini, A., Ganuza, M. L., Gargiulo, M. F., Ferracutti, G., Bjerg, E., Castro, S., Matković, K., Gröller, E., *Visualización de Datos Multidimensionales Procedentes de las Geociencias*, pp. 242-246, Actas del XXV Workshop de Investigadores en Ciencias de la Computación – WICC 2023, 13 y 14 de abril de 2023, Junín, Buenos Aires, Argentina. ISBN 978-987-3724-64-0. CEDi Centro de Edición y Diseño. UNNOBA (editor), Ciudad Autónoma de Buenos Aires).

Antonini, A., Luque, L., Tanzola, J., Asiain, L., Ferracutti, G., Bjerg, E., Castro, S., Ganuza, M. L., *Análisis Visual Guiado para el Estudio de Datos Geoquímicos*, pp. 247-251, Actas del XXV Workshop de Investigadores en Ciencias de la Computación – WICC 2023, 13 y 14 de abril de 2023, Junín, Buenos Aires, Argentina. ISBN 978-987-3724-64-0. CEDi Centro de Edición y Diseño. UNNOBA (editor), Ciudad Autónoma de Buenos Aires).

Ganuza, M. L., Selzer, M., **Antonini, A.**, Luque, L., Urribarri, D., Larrea, M., Ferracutti, G., Asiain, L., Bjerg, E., Castro, S., *Tecnologías Inmersivas y Visualización Situada Aplicadas a Geociencias*, pp. 237-241, Actas del XXV Workshop de Investigadores en Ciencias de la Computación – WICC 2023, 13 y 14 de abril de 2023, Junín, Buenos Aires, Argentina. ISBN 978-987-3724-64-0. CEDi Centro de Edición y Diseño. UNNOBA (editor), Ciudad Autónoma de Buenos Aires).

2021 **Antonini, A.**, Ganuza, M. L., Gargiulo, M. F., Ferracutti, G., Bjerg, E., Castro, S., Matković, K., Gröller, E., *Análisis Visual de Datos Multidimensionales*, Libro de Actas del XXIII Workshop de Investigadores en Ciencias de la Computación – WICC 2021, https://wicc2021.undec.edu.ar/libros-de-actas-y-posters/, pp. 256-259, 15 y 16 de abril de 2021, Chilecito, La Rioja, Argentina. ISBN 978-987-24611-3-3. Inst. Organizadora: UN Chilecito y Red de Universidades con Carreras de Informática (RedUNCI)

Luque, L., Ganuza, M. L., **Antonini, A.**, Castro, S., *npGLC-Vis Library for Multidimensional Data Visualization*, pp. 188-202. In: Cloud Computing, Big Data & Emerging Topics, Communications in Computer and Information Science 1444, IX Jornadas de Cloud Computing, Big Data & Emerging Topics. Selected Papers, Vol. 1444, Chapter 14., Springer. International, 2021. DOI:10.1007/978-3-030-84825-5_14. Print ISBN: 978-3-030-84824-8. Electronic ISBN: 978-3-030-84825-5

2019 Ganuza, M. L., **Antonini, A.**, Gargiulo, M. F., Ferracutti, G., Bjerg, E., Castro, S., Matković, K., Gröller, E. *Análisis Visual de Datos Multidimensionales Provenientes de las Ciencias Geológicas.*, XXI Workshop de Investigadores en Ciencias de la Computación – WICC 2019, pp. 277-281, 25 y 26 de abril de 2019, Fac. de Cs. Exactas, Físicas y Naturales, Universidad Nacional de San Juan (UNSJ), San Juan. ISBN: 978-987-3984-85-3.

8.2. Direcciones Futuras de Investigación

En base al trabajo realizado, se identificaron diversas direcciones futuras de investigación que complementan y amplían el enfoque desarrollado. A continuación, se detallan algunas de estas posibles líneas de investigación:

- **Avance en el Desarrollo e Implementación de Nuevas Técnicas de Visualización para el Análisis de Datos Multidimensionales.**

  El propósito es ampliar y perfeccionar las técnicas de visualización actuales, creando nuevos modelos, herramientas y enfoques que faciliten la exploración de datos por geólogos y otros expertos.

  Aunque el enfoque principal está dirigido a las Ciencias Geológicas, estas metodologías tienen una aplicación más amplia y pueden ser empleadas en diversas disciplinas y áreas de aplicación que también demandan el análisis de datos multidimensionales.

- **Extensión del CVRS para el Análisis y Representación de Redes.**

  En el prototipo inicial de **CVRS** (ver capítulo 6), el sistema de recomendación de técnicas de visualización se ha diseñado principalmente para trabajar con conjuntos de datos tabulares, un formato accesible, fácil de comprender y ampliamente utilizado en múltiples áreas de aplicación. Sin embargo, el análisis de datos no siempre se ajusta a una estructura tabular. En algunos contextos, se requiere el análisis de datos organizados como redes (ver sección 3.3.2), en los cuales los elementos (nodos) están interconectados mediante enlaces, formando estructuras complejas.

  Como parte del trabajo futuro, se plantea ampliar las capacidades del sistema

CVRS para incluir el análisis y la recomendación de técnicas de visualización dirigidas a conjuntos de datos estructurados como redes. Esta extensión permitirá que el sistema se pueda aplicar en una mayor variedad de contextos y disciplinas.

- **Diseño e Implementación de un Sistema de Recomendación Integral de Visualización para el Análisis de Datos Geoquímicos.**

  Se propone el diseño y desarrollo de un sistema integral para la recomendación de técnicas de visualización dirigidas al análisis de datos geoquímicos. Este sistema incorporará gráficos especializados utilizados por geólogos, así como métodos de visualización apropiados para representar datos multidimensionales.

  Además, para ofrecer recomendaciones más precisas, el sistema incorporará un cuarto factor clave en el proceso de recomendación, sumado a los ya considerados en el sistema propuesto anteriormente (ver capítulo 6): el conocimiento del dominio, tal como lo destacan Kaur y Owonibi [KO17]. De este modo, el sistema no solo sugerirá las técnicas de visualización más apropiadas en función de los datos, los objetivos del análisis y las preferencias de representación visual, sino que también asegurará una alineación precisa con las prácticas y requisitos específicos del análisis geológico.

Bibliografía

[Abe08] Abela, A. *Advanced Presentations by Design: Creating Communication That Drives Action.* John Wiley & Sons, 2008.

[AdOL04] Artero, A. O., de Oliveira, M. C. F., and Levkowitz, H. Uncovering clusters in crowded parallel coordinates visualizations. In *IEEE Symposium on Information Visualization* (2004), pp. 81–88.

[AE12] Asan, U., and Ercan, S. *An Introduction to Self-Organizing Maps.* Atlantis Press, Paris, 2012, pp. 295–315.

[AES05] Amar, R., Eagan, J., and Stasko, J. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (New York, NY, USA, 2005), IEEE, IEEE, pp. 111–117.

[AFG+23] Antonini, A., Ferracutti, G., Gargiulo, F., Bjerg, E., Castro, S., and Ganuza, M. L. Análisis visual de datos multidimensionales de química mineral correspondientes a oxiespinelos de xenolitos del manto con spinel web. caso de estudio. In *Actas del 14° Congreso de Mineralogía, Petrología Ígnea y Metamórfica, y Metalogénesis - 14° MinMet y 5° PIMMA - Serie D, Publicación Especial 16* (2023), pp. 36–42.

[AGC22] Antonini, A. S., Ganuza, M. L., and Castro, S. M. Visuel - a web dynamic dashboard for data visualization. *Journal of Computer Science and Technology 22*, 1 (2022).

[AGF+21] Antonini, A. S., Ganuza, M. L., Ferracutti, G., Gargiulo, M. F., Matković, K., Gröller, E., Bjerg, E. A., and Castro, S. M.

Spinel web: An interactive web application for visualizing the chemical composition of spinel group minerals. *Earth Science Informatics 14*, 1 (2021), 521–528.

[AKK96] Ankerst, M., Keim, D. A., and Kriegel, H.-P. Circle segments: A technique for visually exploring large multidimensional data sets. In *Proceedings of the IEEE Visualization '96, Hot Topic Session* (1996), IEEE Computer Society.

[ALF+24] Antonini, A. S., Luque, L., Ferracutti, G. R., Bjerg, E. A., Castro, S. M., and Ganuza, M. L. Spinelva. a new perspective for the visual analysis and classification of spinel group minerals. *Earth Science Informatics 17*, 4 (2024), 3851–3861.

[ALGC23] Antonini, A. S., Luque, L., Ganuza, M. L., and Castro, S. M. Toward a taxonomy for 2d non-paired general line coordinates: A comprehensive survey. *International Journal of Data Science and Analytics 15*, 2 (2023), 133–158.

[AMB+03] Artimo, A., Mäkinen, J., Berg, R. C., Abert, C. C., and Salonen, V.-P. Three-dimensional geologic modeling and visualization of the virttaankangas aquifer, southwestern finland. *Hydrogeology Journal 11* (2003), 378–386.

[AN07] Asuncion, A., and Newman, D. UCI machine learning repository, 2007. Accessed on 2025-02-12.

[ATA+24] Antonini, A. S., Tanzola, J., Asiain, L., Ferracutti, G. R., Castro, S. M., Bjerg, E. A., and Ganuza, M. L. Machine learning model interpretability using shap values: Application to igneous rock classification task. *Applied Computing and Geosciences 23* (2024), 100178.

[Bed90] Beddow, J. Shape coding of multidimensional data on a microcomputer display. In *Proceedings of the 1st Conference on Visualization '90 (VIS '90)* (Washington, DC, USA, 1990), IEEE Computer Society Press, pp. 238–246.

[Ber83] Bertin, J. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, WI, 1983.

[Ber14] Berná, A. G. *A Visual Framework to Accelerate Knowledge Discovery Based on Dimensionality Reduction Minimizing Degradation of Quality*. PhD thesis, Technical University of Madrid, Spain, 2014.

[BKC+13] Borgo, R., Kehrer, J., Chung, D., Maguire, E., Laramee, R., Hauser, H., Ward, M., and Chen, M. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics 2013 - State of the Art Reports* (2013), The Eurographics Association, pp. 39–63.

[BKH05] Bendix, F., Kosara, R., and Hauser, H. Parallel sets: Visual analysis of categorical data. In *IEEE Symposium on Information Visualization (INFOVIS 2005)* (2005), vol. 12, pp. 133–140.

[BM13] Brehmer, M., and Munzner, T. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2376–2385.

[BOH11] Bostock, M., Ogievetsky, V., and Heer, J. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2301–2309.

[BR01] Barnes, S. J., and Roeder, P. L. The range of spinel compositions in terrestrial mafic and ultramafic rocks. *Journal of Petrology 42*, 12 (2001), 2279–2302.

[Bre01] Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.

[BWP+19] Belden, J., Wegier, P., Patel, J., Plaisant, C., Moore, J., Lowrance, N., Boren, S., and Koopman, R. Designing a Medication Timeline for Patients and Physicians. *Journal of the American Medical Informatics Association 26*, 2 (2019), 95–105.

[Cas91] Casner, S. M. Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics (TOG) 10*, 2 (1991), 111–151.

[CDLS97] Castro, S. M., Delrieux, C., Larrea, M., and Silvetti, A. Low-cost volume visualization. In *Proceedings International Congress on Imaging Science, Systems and Technology (CISST'97)* (Nevada, USA, 1997), pp. 489–493.

[CG17] Carr, M. J., and Gazel, E. Igpet software for modeling igneous processes: Examples of application using the open educational version. *Mineralogy and Petrology 111* (2017), 283–289.

[Cha06] Chan, W. W.-Y. A survey on multivariate data visualization. *Department of Computer Science and Engineering. Hong Kong University of Science and Technology 8*, 6 (2006), 1–29.

[Che73] Chernoff, H. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association 68*, 342 (1973), 361–368.

[Chi00] Chi, E. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (2000), pp. 69–75.

[Cle93] Cleveland, W. S. *Visualizing Data*. Hobart Press, 1993.

[CLN87] Carr, D. B., Littlefield, R. J., and Nicholoson, W. L. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association 82*, 398 (1987), 424–436.

[CLPL22] Cao, Y.-R., Li, X.-H., Pan, J.-Y., and Lin, W.-C. Visguide: User-oriented recommendations for data event extraction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)* (New York, NY, USA, 2022), Association for Computing Machinery, pp. 1–13.

[CMS99] Card, S., Mackinlay, J., and Shneiderman, B. *Readings in Information Visualization: Using Vision To Think*. Morgan Kaufmann, 1999.

[CR96] Chuah, M. C., and Roth, S. F. On the semantics of interactive visualizations. In *Proceedings of the 1996 IEEE Symposium on Information Visualization* (USA, 1996), INFOVIS '96, IEEE, IEEE Computer Society, pp. 29–36.

[CR98] Chi, E. H.-H., and Riedl, J. An operator interaction framework for visualization systems. In *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)* (1998), pp. 63–70.

[CRK+95] Chuah, M. C., Roth, S. F., Kolojejchick, J., Mattis, J., and Juarez, O. Sagebook: Searching data-graphics by content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)* (USA, 1995), ACM Press/Addison-Wesley Publishing Co., pp. 338–345.

[CSK+20] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., and Groth, P. Dataset search: A survey. *The VLDB Journal 29*, 1 (2020), 251–272.

[CvW11] Claessen, J. H. T., and van Wijk, J. J. Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2310–2316.

[DGRG12] Daniels, K., Grinstein, G., Russell, A., and Glidden, M. Properties of normalized radial visualizations. *Information Visualization 11*, 4 (2012), 273–300.

[DHPP17] Demiralp, c., Haas, P. J., Parthasarathy, S., and Pedapati, T. Foresight: Recommending visual insights. *Proceedings of the VLDB Endowment 10*, 12 (2017), 1937–1940.

[DKZ13] Dzemyda, G., Kurasova, O., and Zilinskas, J. *Multidimensional Data Visualization: Methods and Applications*, vol. 75 of *Springer Optimization and Its Applications*. Springer, 2013.

[DLR09] Draper, G. M., Livnat, Y., and Riesenfeld, R. F. A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics 15*, 5 (2009), 759–776.

[d'O85] d'Ocagne, M. *Coordonnées Parallèles et Axiales: Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèlles.* Gauthier-Villars, Paris, France, 1885.

[DPD16] Dimitriadou, K., Papaemmanouil, O., and Diao, Y. Aide: An active learning-based approach for interactive data exploration. *IEEE Transactions on Knowledge and Data Engineering 28*, 11 (2016), 2842–2856.

[DSM17] Divino, R., Santos, C., and Meiguins, B. A visual representation of clusters characteristics using edge bundling for parallel coordinates. In *2017 21st International Conference Information Visualisation (IV)* (2017), IEEE, pp. 90–95.

[EDF08] Elmqvist, N., Dragicevic, P., and Fekete, J.-D. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1539–1148.

[EH11] Everitt, B., and Hothorn, T. *Looking at Multivariate Data: Visualisation.* Springer, New York, NY, USA, 2011, pp. 25–60.

[ELP+16] Etemadpour, R., Linsen, L., Paiva, J. G., Crick, C., and Forbes, A. G. Choosing visualization techniques for multidimensional data projection tasks: A guideline with examples. In *Computer Vision, Imaging and Computer Graphics Theory and Applications* (2016), Springer International Publishing, pp. 166–186.

[ERI+21] Enge, K., Rind, A., Iber, M., Höldrich, R., and Aigner, W. It's about time: Adopting theoretical constructs from visualization for sonification. In *Proceedings of the 16th International Audio Mostly Conference* (New York, NY, USA, 2021), Association for Computing Machinery, pp. 64–71.

[FGG+15] Ferracutti, G. R., Gargiulo, M. F., Ganuza, M. L., Bjerg, E. A., and Castro, S. M. Determination of the spinel group end-members based on electron microprobe analyses. *Mineralogy and Petrology 109* (2015), 153–160.

[Fis88] Fisher, R. A. Iris. UCI Machine Learning Repository, 1988.

[FR81] Flury, B., and Riedwyl, H. Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association 76*, 376 (1981), 757–765.

[FR91] Fruchterman, T. M. J., and Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience 21*, 11 (1991), 1129–1164.

[FWR99] Fua, Y.-H., Ward, M. O., and Rundensteiner, E. A. *Hierarchical Parallel Coordinates for Exploration of Large Datasets.* IEEE Computer Society Press, Washington, DC, USA, 1999.

[Gan18] Ganuza, M. L. *Interacciones en visualización.* PhD thesis, Universidad Nacional del Sur, Bahía Blanca, Argentina, 2018.

[GBRQ14] Gemmell, P., Burrage, K., Rodriguez, B., and Quinn, T. A. Population of computational rabbit-specific ventricular action potential models for investigating sources of variability in cellular repolarisation. *PLOS ONE 9*, 2 (2014), 1–13.

[GCF+12] Ganuza, M. L., Castro, S. M., Ferracutti, G., Bjerg, E. A., and Martig, S. R. Spinelviz: An interactive 3d application for visualizing spinel group minerals. *Computers & Geosciences 48* (2012), 50–56.

[GFG+14] Ganuza, M. L., Ferracutti, G., Gargiulo, M. F., Castro, S. M., Bjerg, E., Gröller, E., and Matković, K. The spinel explorer—interactive visual analysis of spinel group minerals. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1913–1922.

[GFG+17] Ganuza, M. L., Ferracutti, G., Gargiulo, F., Castro, S. M., Bjerg, E. A., Gröller, E., and Matković, K. Interactive visual categorization of spinel-group minerals. In *Proceedings of the 33rd Spring Conference on Computer Graphics (SCCG '17)* (New York, NY, USA, 2017), Association for Computing Machinery, pp. 1–11.

[GGF+15] Ganuza, M. L., Gargiulo, F., Ferracutti, G., Castro, S., Bjerg, E., Gröller, E., and Matković, K. Interactive semi-automatic categorization for spinel group minerals. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), IEEE, pp. 197–198.

[GK03] Graham, M., and Kennedy, J. Using curves to enhance parallel coordinate visualisations. In *Proceedings of the Seventh International Conference on Information Visualization* (USA, 2003), IV '03, IEEE Computer Society, pp. 10–16.

[GLI98] Grinstein, G., Laskowski, S., and Inselberg, A. Key problems and thorny issues in multidimensional visualization. In *Proceedings of the conference on Visualization'98* (1998), IEEE Computer Society Press, pp. 505–506.

[GM00] Graham, D. J., and Midgley, N. G. Graphical representation of particle shape using triangular diagrams: An excel spreadsheet method. *Earth Surface Processes and Landforms 25*, 13 (2000), 1473–1477.

[Gna81] Gnanamgari, S. *Information Presentation Through Default Displays.* PhD thesis, University of Pennsylvania, USA, 1981.

[GPS+11] Geng, Z., Peng, Z., S.Laramee, R., C. Roberts, J., and Walker, R. Angular histograms: Frequency-based visualizations for large, high dimensional data. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2572–2580.

[GSGC08] Gilson, O., Silva, N., Grant, P. W., and Chen, M. From web data to visualization via ontology mapping. *Computer Graphics Forum 27* (2008), 959–966.

[GTN14] Ghassemi Toosi, F., and Nikolov, N. Circular tree drawing by simulating network synchronisation dynamics and scaling. *Lecture Notes in Computer Science 8871* (2014), 511–512.

[GULC23] Ganuza, M., Urribarri, D., Larrea, M., and Castro, S. Uvam: The unified visual analytics model. the unified visualization model revisited. In *Cloud Computing, Big Data & Emerging Topics* (2023), Springer Nature Switzerland, pp. 189–203.

[GW09] Gotz, D., and Wen, Z. Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)* (New York, NY, USA, 2009), Association for Computing Machinery, pp. 315–324.

[GXWY10] Guo, P., Xiao, H., Wang, Z., and Yuan, X. Interactive local clustering operations for high dimensional data in parallel coordinates. In *Proceedings of the 2010 IEEE Pacific Visualization Symposium (PacificVis)* (2010), pp. 97–104.

[Hal18] Halkidi, M. Hierarchical clustering. In *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Springer, New York, NY, USA, 2018, pp. 1684–1689.

[Han06] Hanrahan, P. Vizql: A language for query, analysis and visualization. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)* (New York, NY, USA, 2006), Association for Computing Machinery, pp. 721–721.

[HBL+19] Hu, K., Bakker, M. A., Li, S., Kraska, T., and Hidalgo, C. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (New York, NY, USA, 2019), Association for Computing Machinery, pp. 1–12.

[HBO10] Heer, J., Bostock, M., and Ogievetsky, V. A tour through the visualization zoo. *Commun. ACM 53*, 6 (2010), 59–67.

[HBRD22] Harrison, K. R., Bidgoli, A. A., Rahnamayan, S., and Deb, K. Image-based benchmarking and visualization for large-scale global optimization. *Applied Intelligence 52*, 4 (2022), 4161–4191.

[HG01] Hoffman, P. E., and Grinstein, G. G. *A Survey of Visualizations for High-Dimensional data Mining.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 47–82.

[HGM+97] Hoffman, P., Grinstein, G., Marx, K., Grosse, I., and Stanley, E. Dna visual and analytic data mining. In *Proceedings of the 8th Conference on Visualization '97* (Washington, DC, USA, 1997), IEEE Computer Society Press, pp. 437–442.

[HKP12] Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*, third ed. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Waltham, MA, USA, 2012.

[HMB81] Huff, D. L., Mahajan, V., and Black, W. C. Facial representation of multivariate data. *Journal of Marketing 45*, 4 (1981), 53–59.

[HN98] Hintze, J. L., and Nelson, R. D. Violin plots: A box plot-density trace synergism. *The American Statistician 52*, 2 (1998), pp. 181–184.

[Hot33] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 6 (1933), 498–520.

[HPP+18] Huang, E., Peng, L., Palma, L. D., Abdelkafi, A., Liu, A., and Diao, Y. Optimization for active learning-based interactive database exploration. *Proceedings of the VLDB Endowment 12*, 1 (2018), 71–84.

[HS12] Heer, J., and Shneiderman, B. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue 10*, 2 (2012), 30–55.

[IBAAR+18] Ibrahim Babura, B., Adam, M., Abdul Rahim, A., Samad, A., Fitrianto, A., and Yusif, B. Analysis and assessment of boxplot characters for extreme data. *Journal of Physics Conference Series 1132*, 1 (2018).

[ID87] Inselberg, A., and Dimsdale, B. Parallel coordinates for visualizing multi-dimensional geometry. In *Computer Graphics 1987: Proceedings of CG International'87* (Tokyo, 1987), T. L. Kunii, Ed., Springer, pp. 25–44.

[ID90] Inselberg, A., and Dimsdale, B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization '90 (VIS '90)* (Washington, DC, USA, 1990), IEEE Computer Society Press, pp. 361–378.

[ID09] Inselberg, A., and Dimsdale, B. Parallel coordinates. *Human-Machine Interactive Systems* (2009), 199–233.

[Ins85] Inselberg, A. The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91.

[Jac03] Jackson, J. E. *A User's Guide to Principal Components*, 1st ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, USA, 2003.

[JFK16] Jäckle, D., Fuchs, J., and Keim, D. Star glyph insets for overview preservation of multivariate data. *Electronic Imaging 28*, 1 (2016).

[Jol02] Jolliffe, I. T. *Principal Component Analysis*. Springer, New York, NY, USA, 2002.

[JS91] Johnson, B., and Shneiderman, B. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd Conference on Visualization (VIS '91)* (Washington, DC, USA, 1991), IEEE Computer Society Press, pp. 284–291.

[KAF+08] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. Visual analytics: Definition, process, and challenges. In *Information Visualization: Human-Centered Issues and Perspectives* (Berlin, Heidelberg, 2008), Springer, pp. 154–175.

[Kan00] Kandogan, E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the 2000*

IEEE Symposium on Information Visualization (InfoVis 2000) (2000), IEEE, pp. 11–16.

[Kan01] Kandogan, E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2001), KDD '01, Association for Computing Machinery, pp. 107–116.

[KAS94] Klumpar, D. M., Anderson, K., and Simoudis, A. Rave: Rapid visualization environment. In *The 1994 Goddard Conference on Space Applications of Artificial Intelligence* (1994), vol. 3268, pp. 29–38.

[Kei94] Keim, D. A. *Visual Support for Query Specification and Data Mining System.* PhD thesis, Ludwig Maximilians Universität, Munich, Germany, 1994.

[Kei95] Keim, D. A. Enhancing the visual clustering of query-dependent database visualization techniques using screen-filling curves. In *Proceedings of the IEEE Visualization '95 Workshop on Database Issues for Data Visualization* (Berlin, Heidelberg, 1995), Springer, pp. 101–110.

[Kei96] Keim, D. A. Pixel-oriented database visualizations. *SIGMOD Rec. 25*, 4 (1996), 35–39.

[Kei02] Keim, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 1–8.

[KFvD18] Kubernátová, P., Friedjungová, M., and van Duijn, M. Knowledge at first glance: A model for a data visualization recommender system suited for non-expert users. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)* (Setubal, Portugal, 2018), SCITEPRESS - Science and Technology Publications, Lda., pp. 208–219.

[KFVD19] Kubernátová, P., Friedjungová, M., and Van Duijn, M. Constructing a data visualization recommender system. In *Data Management*

Technologies and Applications: 7th International Conference (DATA 2018) (Porto, Portugal, 2019), vol. 862, pp. 1–25.

[KG19] Kovalerchuk, B., and Grishin, V. Adjustable general line coordinates for visual knowledge discovery in n-d data. *Information Visualization 18*, 1 (2019), 3–32.

[KHPA12] Key, A., Howe, B., Perry, D., and Aragon, C. VizDeck: Self-Organizing Dashboards for Visual Analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2012), Association for Computing Machinery, pp. 681–684.

[KK94a] Keim, D. A., and Krigel, H.-P. Visdb: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications 14*, 5 (1994), 40–49.

[KK94b] Keller, P. R., and Keller, M. M. *Visual Cues: Practical Data Visualization*, vol. 2. IEEE Computer Society Press, Washington, DC, USA, 1994.

[KK96] Keim, D. A., and Kriegel, H.-P. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering 8*, 6 (1996), 923–938.

[KKA95] Keim, D., Kriegel, H.-P., and Ankerst, M. Recursive pattern: A technique for visualizing very large amounts of data. In *Proceedings Visualization '95* (1995), pp. 279–286.

[KO17] Kaur, P., and Owonibi, M. A review on visualization recommendation strategies. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017) - IVAPP* (2017), vol. 3, SciTePress, pp. 266–273.

[Koh90] Kohonen, T. The self-organizing map. *Proceedings of the IEEE 78*, 9 (1990), 1464–1480.

[Kov14] Kovalerchuk, B. Visualization of multidimensional data with collocated paired coordinates and general line coordinates. In *Visualization and Data*

Analysis (Bellingham, WA, USA, 2014), P. C. Wong, D. L. Kao, M. C. Hao, and C. Chen, Eds., vol. 9017, International Society for Optics and Photonics, SPIE, pp. 191–205.

[Kov18] Kovalerchuk, B. *Visual Knowledge Discovery and Machine Learning*, vol. 144. Springer Publishing Company, Incorporated, 2018.

[KS12] Kerren, A., and Schreiber, F. Toward the role of interaction in visual analytics. In *Proceedings of the 2012 Winter Simulation Conference (WSC)* (2012), WSC '12, IEEE, Winter Simulation Conference, pp. 1–13.

[LAGC24] Luque, L., Antonini, A. S., Ganuza, M. L., and Castro, S. M. Glc-frame: A framework and library for exploration of multidimensional data with general line coordinates. *Journal of Computer Science and Technology 24*, 1 (2024).

[Lev91] Levkowitz, H. Color icons: Merging color and texture perception for integrated visualization of multiple parameters. In *Proceedings of the 2nd Conference on Visualization (VIS '91)* (Washington, DC, USA, 1991), IEEE Computer Society Press, pp. 164–170.

[LGAC21] Luque, L. E., Ganuza, M. L., Antonini, A. S., and Castro, S. M. npglc-vis library for multidimensional data visualization. In *Cloud Computing, Big Data & Emerging Topics* (2021), Springer International Publishing, pp. 188–202.

[LHH12] Lu, L. F., Huang, M. L., and Huang, T.-H. A new axes re-ordering method in parallel coordinates visualization. In *2012 11th International Conference on Machine Learning and Applications* (2012), vol. 2, pp. 252–257.

[Lim14] Lima, M. *The Book of Trees: Visualizing Branches of Knowledge*. Princeton Architectural Press, New York, NY, USA, 2014.

[Lin91] Lindsley, D. H. Oxide minerals: petrologic and magnetic significance. *Review in Mineralogy 25* (1991).

[LJJ18] Leland, M., John, H., and James, M. Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints arxiv:1802.03426 3* (2018).

[LL17] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (Red Hook, NY, USA, 2017), vol. 25, Curran Associates Inc., pp. 4768–4777.

[LPP+06] Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (New York, NY, USA, 2006), BELIV '06, Association for Computing Machinery, pp. 1–5.

[LQT+18] Luo, Y., Qin, X., Tang, N., Li, G., and Wang, X. Deepeye: Creating good data visualizations by keyword search. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)* (New York, NY, USA, 2018), Association for Computing Machinery, p. 1733–1736.

[LQTL18] Luo, Y., Qin, X., Tang, N., and Li, G. Deepeye: Towards automatic data visualization. In *2018 IEEE 34th international conference on data engineering (ICDE)* (2018), pp. 101–112.

[LRB03] Lee, M. D., Reilly, R. E., and Butavicius, M. A. An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation* (Adelaide, Australia, 2003), vol. 24, Australian Computer Society, Inc., pp. 1–10.

[LV03] Lyman, P., and Varian, H. R. How much information? *The Journal of Electronic Publishing 9*, 1 (2003).

[LWW90] LeBlanc, J., Ward, M. O., and Wittels, N. Exploring n-dimensional databases. In *Proceedings of the 1st Conference on Visualization (VIS '90)* (Washington, DC, USA, 1990), IEEE Computer Society Press, pp. 230–237.

[Mac86] Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG) 5*, 2 (1986), 110–141.

[Mar96] Marshall, D. Ternplot: An excel spreadsheet for ternary diagrams. *Computers & Geosciences 22*, 6 (1996), 697–699.

[MCFE03] Martig, S. R., Castro, S. M., Fillottrani, P. R., and Estévez, E. C. Un modelo unificado de visualización. In *IX Congreso Argentino de Ciencias de la Computación* (2003), Universidad Nacional de La Plata, pp. 881–892.

[MHS07] Mackinlay, J., Hanrahan, P., and Stolte, C. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1137–1144.

[MM00] Meddes, J., and McKenzie, E. Improving visualization by capturing domain knowledge. In *Visual Data Exploration and Analysis VII* (2000), vol. 3960, pp. 186–195.

[MRY+04] Masumoto, S., Raghavan, V., Yonezawa, G., Nemoto, T., and Shiono, K. Construction and visualization of a three dimensional geologic model using grass gis. *Transactions in GIS 8*, 2 (2004), 211–223.

[MTL78] McGill, R., Tukey, J. W., and Larsen, W. A. Variations of box plots. *The American Statistician 32*, 1 (1978), pp. 12–16.

[Mun09] Munzner, T. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 921–928.

[Mun14] Munzner, T. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC press, 2014.

[Mur13] Murray, D. *Tableau Your Data! Fast and Easy Visual Analysis with Tableau Software*, 1st ed. Wiley Publishing, 2013.

[MVT17] Mutlu, B., Veas, E., and Trattner, C. Tags, titles or qas? choosing content descriptors for visual recommender systems. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '1)* (New York, NY, USA, 2017), Association for Computing Machinery, pp. 265–274.

[MVTS15a] Mutlu, B., Veas, E., Trattner, C., and Sabol, V. Towards a recommender engine for personalized visualizations. In *User Modeling, Adaptation and Personalization* (2015), Springer International Publishing, pp. 169–182.

[MVTS15b] Mutlu, B., Veas, E., Trattner, C., and Sabol, V. Vizrec: A two-stage recommender system for personalized visualizations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion (IUI '15 Companion)* (New York, NY, USA, 2015), Association for Computing Machinery, pp. 49–52.

[MZKMY12] Mohd Zin, Z., Khalid, M., Mesbahi, E., and Yusof, R. Data clustering and topology preservation using 3d visualization of self organizing maps. In *Lecture Notes in Engineering and Computer Science* (2012), vol. 2198 of *Lecture Notes in Engineering and Computer Science*, Newswood Limited, pp. 696–701. World Congress on Engineering 2012 (WCE 2012).

[NŠ09] Nováková, L., and Štepanková, O. Radviz and identification of clusters in multidimensional data. In *2009 13th International Conference Information Visualisation* (Los Alamitos, CA, USA, 2009), IEEE Computer Society, pp. 104–109.

[PDDP16] Papaemmanouil, O., Diao, Y., Dimitriadou, K., and Peng, L. Interactive data exploration via machine learning models. *IEEE Data Engineering Bulletin 39*, 4 (2016), 38–49.

[Pea95] Pearson, K. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58* (1895), 240–242.

[Pea01] Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*, 11 (1901), 559–572.

[PEF02] Pastizzo, M., Erbacher, R., and Feldman, L. Multidimensional data visualization. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc 34*, 2 (2002), 158–162.

[PG88] Pickett, R., and Grinstein, G. Iconographic displays for visualizing multidimensional data. In *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics* (1988), vol. 1, pp. 514–519.

[PKC19] Piotrowski, W., Kipouros, T., and Clarkson, P. J. Enhanced Interactive Parallel Coordinates using Machine Learning and Uncertainty Propagation for Engineering Design . In *2019 15th International Conference on eScience (eScience)* (Los Alamitos, CA, USA, 2019), IEEE Computer Society, pp. 339–348.

[Plo18] Plotly. Plotly community feed, 2018. Accessed on February, 2025.

[PLW+22] Pandey, A., L'Yi, S., Wang, Q., Borkin, M. A., and Gehlenborg, N. Genorec: A recommendation system for interactive genomics data visualization. *IEEE Transactions on Visualization and Computer Graphics 29*, 1 (2022), 570–580.

[PMS+03] Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., and Shneiderman, B. LifeLines: Using visualization to enhance navigation and analysis of patient records. In *The Craft of Information Visualization.* Morgan Kaufmann, San Francisco, CA, USA, 2003, pp. 308–312.

[PSS22] Pandey, A., Srinivasan, A., and Setlur, V. Medley: Intent-based recommendations to support dashboard composition. *IEEE Transactions on Visualization and Computer Graphics PP* (2022), 1–11.

[QLTL18] Qin, X., Luo, Y., Tang, N., and Li, G. Deepeye: An automatic big data visualization framework. *Big Data Mining and Analytics 1*, 1 (2018), 75–82.

[QRD+22] Qian, X., Rossi, R. A., Du, F., Kim, S., Koh, E., Malik, S., Lee, T. Y., and Ahmed, N. K. Personalized visualization recommendation. *ACM Transactions on the Web (TWEB) 16*, 3 (2022), 1–47.

[RC94] Rao, R., and Card, S. K. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1994), CHI '94, Association for Computing Machinery, pp. 318–322.

[Ric95] Richard, L. Minpet: Mineralogical and petrological data processing system, version 2.02. *MinPet Geological Software, Québec, Canada* (1995).

[RKMC95] Roth, S. F., Kolojejchick, J., Mattis, J., and Chuah, M. C. Sagetools: An intelligent environment for sketching, browsing, and customizing data-graphics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)* (New York, NY, USA, 1995), Association for Computing Machinery, pp. 409–410.

[RM90] Roth, S. F., and Mattis, J. Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)* (New York, NY, USA, 1990), Association for Computing Machinery, pp. 193–200.

[Rob07] Roberts, J. C. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '07)* (USA, 2007), IEEE Computer Society, pp. 61–71.

[Roe94] Roeder, P. L. Chromite; from the fiery rain of chondrules to the kilauea iki lava lake. *The Canadian Mineralogist 32*, 4 (1994), 729–746.

[Row] Rowland, T. Manifold. `https://mathworld.wolfram.com/Manifold.html`. Accessed: 2025-02-13.

[RtBS84] Ramsay, J. O., ten Berge, J., and Styan, G. P. Matrix correlation. *Psychometrika 49*, 3 (1984), 403–423.

[SED19] Saket, B., Endert, A., and Demiralp, Ç. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics 25*, 7 (2019), 2505–2512.

[SG18] Sarikaya, A., and Gleicher, M. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 402–412.

[Shn92] Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG) 11*, 1 (1992), 92–99.

[Shn96] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)*. IEEE Computer Society, Los Alamitos, CA, USA, 1996, pp. 336–343.

[SHS10] Schulz, H.-J., Hadlak, S., and Schumann, H. The design space of implicit hierarchy visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics 17*, 4 (2010), 393–411.

[Sii00] Siirtola, H. Direct manipulation of parallel coordinates. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2000), Association for Computing Machinery, pp. 119–120.

[Sii07] Siirtola, H. *Interactive Visualization of Multidimensional Data*. Dissertations in Interactive Technology. Tampere University Press, Tampere, Finland, 2007.

[SM07] Shen, Z., and Ma, K.-L. Path visualization for adjacency matrices. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC Conference on Visualization (EUROVIS '07)* (Goslar, Germany, 2007), Eurographics Association, pp. 83–90.

[SMS+21] Siang, C. V., Mohamed, F. B., Salleh, F. M., Isham, M. I. B. M., Basori, A. H., and Selamat, A. B. An overview of immersive data visualisation methods using type by task taxonomy. In *2021 IEEE International Conference on Computing (ICOCO)* (2021), IEEE, pp. 347–352.

[SMWH17] Satyanarayan, A., Moritz, D., Wongsuphasawat, K., and Heer, J. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 341–350.

[Spe07] Spence, R. *Information Visualization: Design for Interaction (2nd Edition)*. Prentice-Hall, Inc., USA, 2007.

[SS06] Schulz, H.-J., and Schumann, H. Visualizing graphs - a generalized view. In *Proceedings of the Conference on Information Visualization (IV '06)* (USA, 2006), IEEE Computer Society, pp. 166–173.

[SST+21] Shen, L., Shen, E., Tai, Z., Song, Y., and Wang, J. TaskVis: Task-oriented Visualization Recommendation. In *EuroVis 2021 - Short Papers* (2021), The Eurographics Association, pp. 91–95.

[STH02] Stolte, C., Tang, D., and Hanrahan, P. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 52–65.

[STLD20] Scheibel, W., Trapp, M., Limberger, D., and Döllner, J. A taxonomy of treemap visualization techniques. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)* (2020), vol. 3, SciTePress, pp. 273–280.

[TC05] Thomas, J. J., and Cook, K. A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE, Piscataway, NJ, USA, 2005.

[Tel14] Telea, A. C. *Data Visualization: Principles and Practice*, 2nd ed. CRC Press, Boca Raton, FL, USA, 2014.

[Tit21] Titanica, E. Titanic people explorer. https://www.encyclopedia-titanica.org/explorer/, 2021. Accessed on 2025-02-10.

[Tor52] Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika 17* (1952), 401–419.

[TS20] Tominski, C., and Schumann, H. *Interactive Visual Data Analysis*. AK Peters Visualization Series. CRC Press, 2020.

[Tuk77] Tukey, J. W. *Exploratory Data Analysis*, vol. 2 of *Addison-Wesley series in behavioral science*. Addison-Wesley Publishing Company, 1977.

[Twe97] Tweedie, L. Characterizing interactive externalizations. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1997), CHI '97, Association for Computing Machinery, pp. 375–382.

[UMA] UMAP Development Team. Umap documentation. `https://umap-learn.readthedocs.io/en/latest/index.html`. Accessed: 2025-02-13.

[vdMH08] van der Maaten, L., and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[VHS+17] Vartak, M., Huang, S., Siddiqui, T., Madden, S., and Parameswaran, A. Towards visualization recommendation systems. *SIGMOD Rec. 45*, 4 (2017), 34–39.

[VPF06] Valiati, E. R. A., Pimenta, M. S., and Freitas, C. M. D. S. A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (New York, NY, USA, 2006), BELIV '06, Association for Computing Machinery, pp. 1–6.

[VRM+15] Vartak, M., Rahman, S., Madden, S., Parameswaran, A., and Polyzotis, N. Seedb: Efficient data-driven visualization recommendations to support visual analytics. *Proceedings of the VLDB Endowment 8*, 13 (2015), 2182–2193.

[VWvH+07] Viegas, F. B., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1121–1128.

[War02] Ward, M. O. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization 1*, 3-4 (2002), 194–210.

[War08] Ward, M. O. *Multivariate Data Glyphs: Principles and Practice*. Springer, Berlin, Heidelberg, 2008, pp. 179–198.

[War19] Ware, C. *Information Visualization: Perception for Design*, d ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2019.

[Was13] Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, New York, NY, USA, 2013.

[Wat02] Wattenberg, M. Arc diagrams: Visualizing structure in strings. *IEEE Symposium on Information Visualization (INFOVIS 2002)* (2002), 110–116.

[WB94] Wong, P. C., and Bergeron, R. D. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques* (USA, 1994), IEEE Computer Society, pp. 3–33.

[WGK10] Ward, M., Grinstein, G., and Keim, D. *Interactive Data Visualization: Foundations, Techniques, and Applications*. AK Peters/CRC Press, 2010.

[Wil92] Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N. L. Johnson, Eds., vol. 2. Springer, New York, NY, USA, 1992, pp. 196–202.

[Wil09] Wills, G. Visualizing network data. In *Encyclopedia of Database Systems* (Boston, MA, 2009), Springer, pp. 3432–3437.

[Wil12] Wilkinson, L. *The Grammar of Graphics*. Statistics and Computing. Springer, New York, NY, 2012.

[Wil19] Wilke, C. O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 2019.

[WL90] Wehrend, S., and Lewis, C. A problem-oriented classification of visualization techniques. In *Proceedings of the 1st Conference on Visualization (VIS '90)* (Washington, DC, USA, 1990), IEEE Computer Society Press, pp. 139–143.

[WMA+15] Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 649–658.

[WMA+16] Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. Towards a general-purpose query language for visualization recommendation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '16)* (New York, NY, USA, 2016), Association for Computing Machinery, pp. 1–6.

[WQM+17] Wongsuphasawat, K., Qu, Z., Moritz, D., Chang, R., Ouk, F., Anand, A., Mackinlay, J., Howe, B., and Heer, J. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)* (New York, NY, USA, 2017), Association for Computing Machinery, p. 2648–2659.

[WWZ+22] Wu, A., Wang, Y., Zhou, M., He, X., Zhang, H., Qu, H., and Zhang, D. Multivision: Designing analytical dashboards with deep learning based recommendation. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2022), 162–172.

[WY04] Ward, M. O., and Yang, J. Interaction spaces in data and information visualization. In *Proceedings of the Sixth Joint Eurographics - IEEE TCVG Conference on Visualization* (Goslar, DEU, 2004), VISSYM'04, Eurographics Association, pp. 137–146.

[XZM17] Xie, C., Zhong, W., and Mueller, K. A visual analytics approach for categorical joint distribution reconstruction from marginal projections. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 51–60.

[ZF98] Zhou, M. X., and Feiner, S. K. Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems (New York, NY, USA, 1998), ACM Press/Addison-Wesley Publishing Co., pp. 392–399.

[ZGCS21] Zhang, X., Ge, X., Chrysanthis, P. K., and Sharaf, M. A. Viewseeker: An interactive view recommendation framework. *Big Data Research 25*, C (2021), 100238.

[ZYQ+08] Zhou, H., Yuan, X., Qu, H., Cui, W., and Chen, B. Visual clustering in parallel coordinates. *Computer Graphics Forum 27*, 3 (2008), 1047–1054.

[ZZT+13] Zhang, L., Zhang, Y., Tang, J., Lu, K., and Tian, Q. Binary code ranking with weighted hamming distance. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition* (2013), pp. 1586–1593.