

Data Visualization for Big Data

Introduction & Concepts

Juan Morales del Olmo

Índice

- 1. Introducción**
2. Fundamentos
3. Casos
4. Visualización en Big Data
5. Datos Tabulares
6. Datos Temporales
7. Datos Espaciales
8. Redes y Jerarquías

Estructura del módulo

- 1. Introduction & Concepts**
2. GIS
3. Dashboards
4. Networks
5. Javascript Tools

Definición

- Los sistemas de visualización proporcionan **representaciones visuales de los datos**.
- Están diseñadas para **ayudar a los usuarios a ejecutar mejor sus tareas**.

La visualización resulta adecuada cuando hay una necesidad de mejorar la capacidad de análisis de los seres humanos en lugar de reemplazarlos con métodos de decisión exclusivamente computacionales.

- **Human in the Loop:** si la pregunta no está bien definida o los algoritmos nos son suficientemente precisos necesitamos visualizar no sólo machine learning.

Usos principales

- Comunicar
 - Reporting, infografías, ...
- Comprobar
 - Monitorización, depurar, calidad del dato ...
- Descubrir
 - Análisis exploratorio, modelado, ...

¿Qué buscamos al visualizar?

→ Trends



→ Outliers



→ Features



→ Distribution



→ Extremes



→ Dependency



→ Correlation



→ Similarity



→ Topology



→ Paths



→ Shape



Tamara Munzner

¿Por qué el canal visual?

Debido a **Factores Humanos**:

- Procesamiento en paralelo (preatentivo)
- Gran ancho de banda
- Con el sonido procesamos secuencialmente
- Con el resto de canales perceptivos no tenemos suficiente precisión

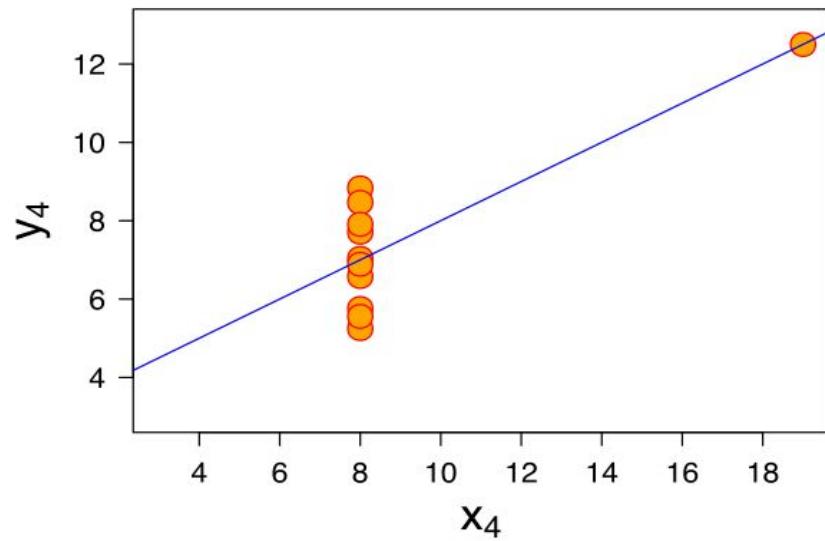
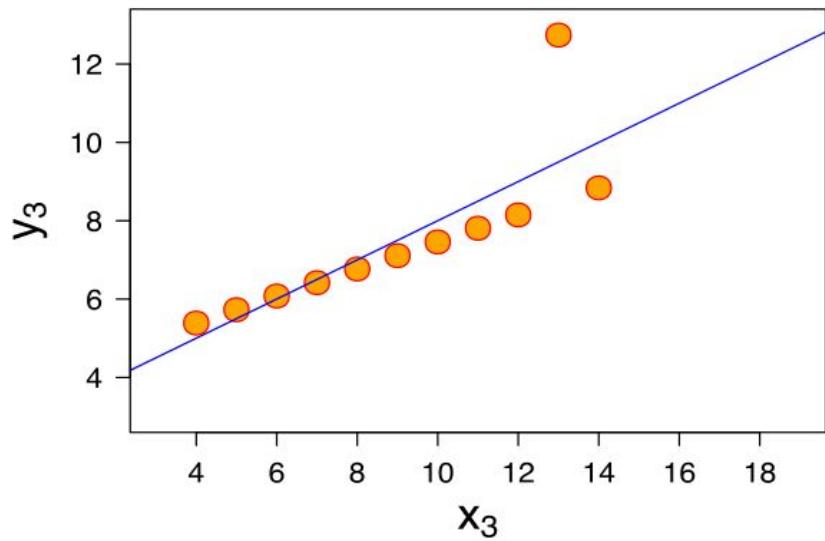
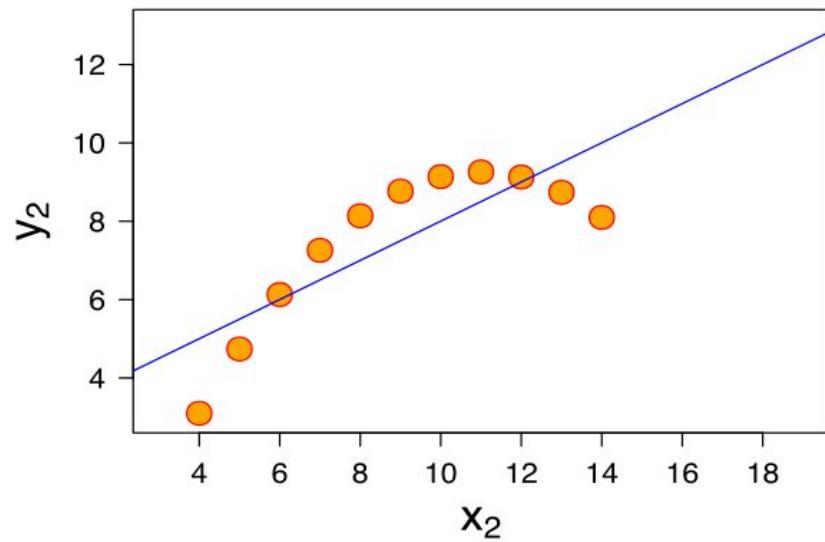
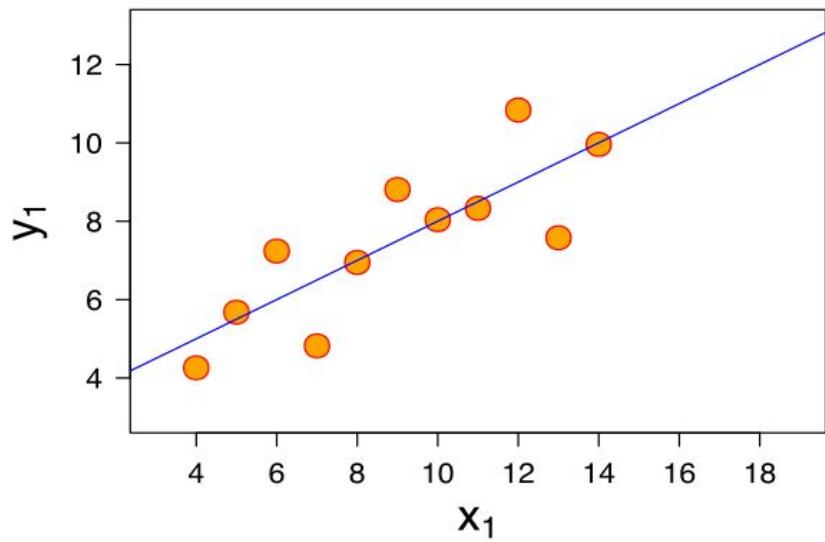
Tarea: Contar

8632145684689642244689886867687979
9129812665870980983097618798409834
2834769898958648509406926448400980
2349588900986047798704980480980498

8632145684689642244689886867687979
9129812665870980983097618798409834
2834769898958648509406926448400980
2349588900986047798704980480980498

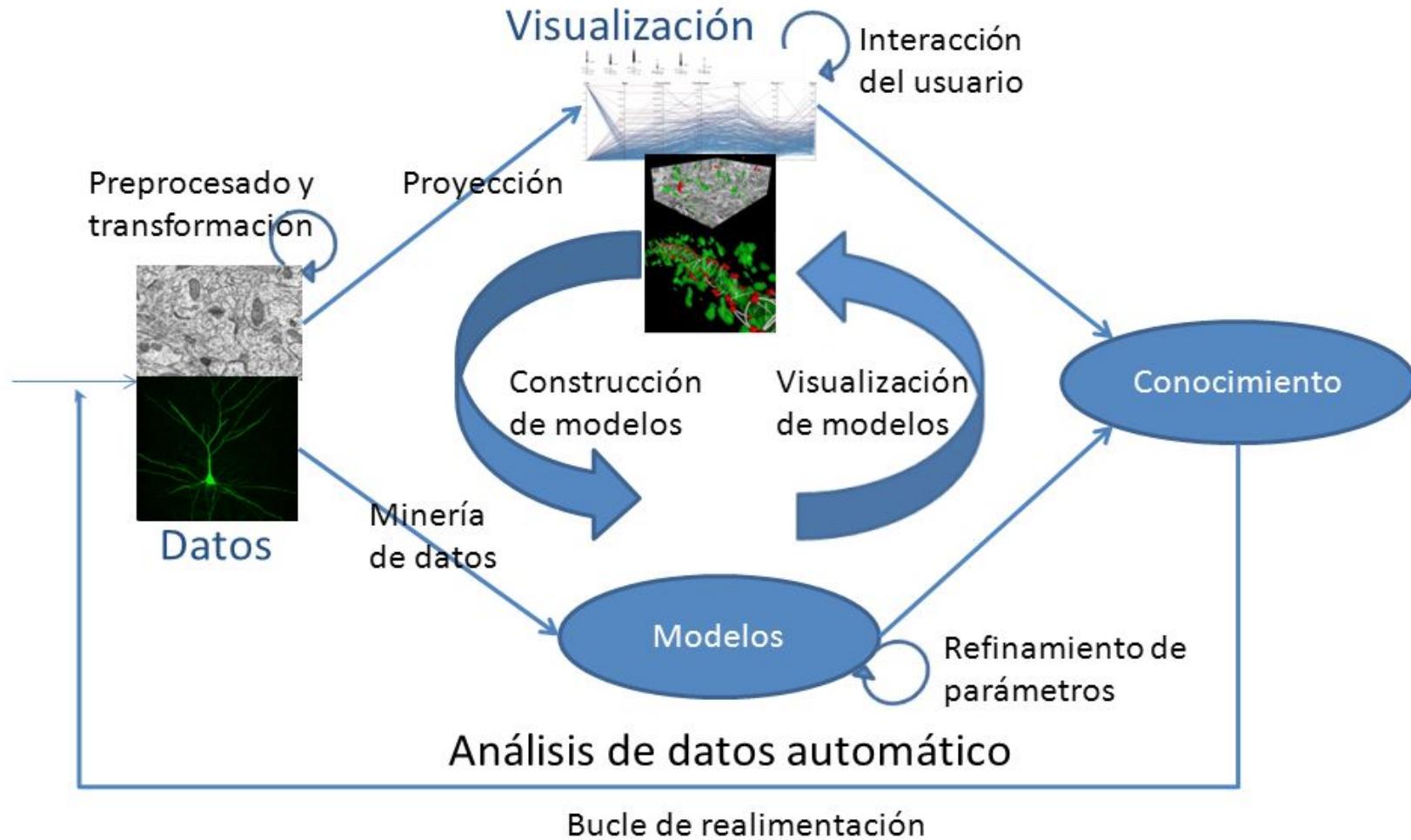
Cuarteto de Anscombe

	I	II	III	IV				
x	y	x	y	x	y	x	y	
10	8,04	10	9,14	10	7,46	8	6,58	
8	6,95	8	8,14	8	6,77	8	5,76	
13	7,58	13	8,74	13	12,74	8	7,71	
9	8,81	9	8,77	9	7,11	8	8,84	
11	8,33	11	9,26	11	7,81	8	8,47	
14	9,96	14	8,1	14	8,84	8	7,04	
6	7,24	6	6,13	6	6,08	8	5,25	
4	4,26	4	3,1	4	5,39	19	12,5	
12	10,84	12	9,13	12	8,15	8	5,56	
7	4,82	7	7,26	7	6,42	8	7,91	
5	5,68	5	4,74	5	5,73	8	6,89	
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



- La visualización nos permite ver muchos datos **sin necesidad de simplificaciones**
- Aún así, cuando trabajamos en un contexto **Big Data** habrá que utilizar técnicas avanzadas porque **no será posible representar todo**

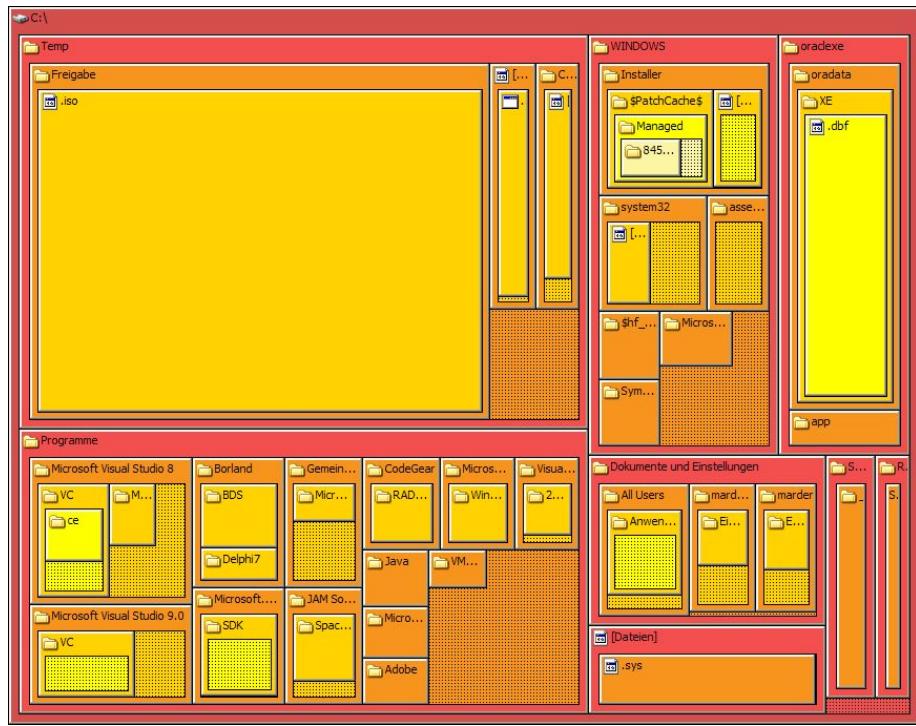
Exploración visual de los datos



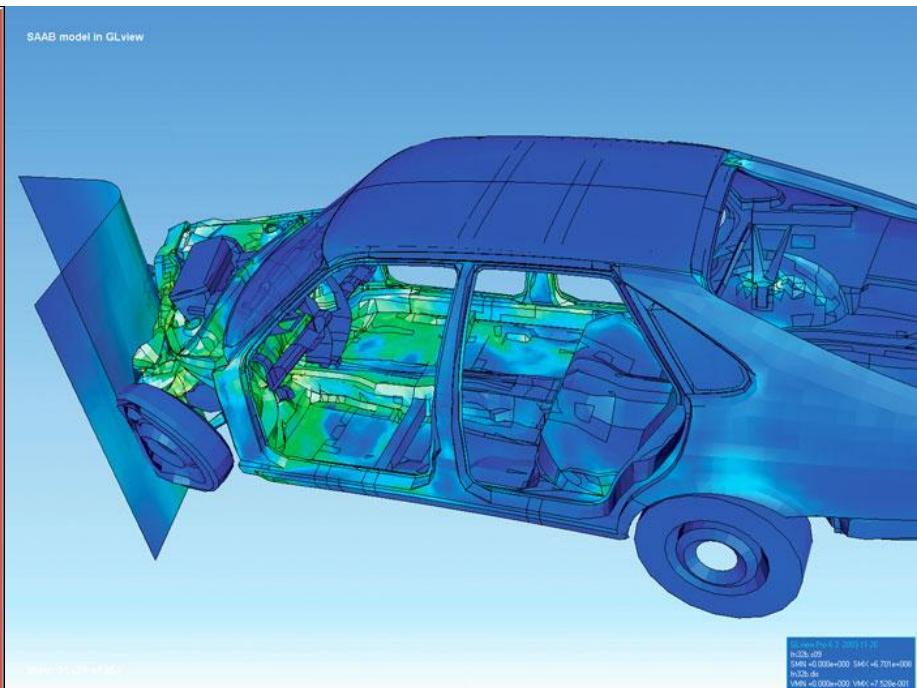
Visual Analytics

Visualización de datos: Corrientes

InfoVis



SciVis



Índice

1. Introducción
- 2. Fundamentos**
3. Casos
4. Visualización en Big Data
5. Datos Tabulares
6. Datos Temporales
7. Datos Espaciales
8. Redes y Jerarquías

Fundamentos

Ontología de datos

Tipo de datos

- Items
- Atributos
- Enlaces
- Posiciones
- Grids

Tipos de atributos

- Categóricos
- Ordinales *
- Cuantitativos *

(*) Según la dirección:

- Secuenciales
- Divergentes
- Cíclicos

La semántica es importante: clave o valor, **espacial, temporal**

Tipos de datasets

- Tabulares
 - Tablas
 - Tablas multidimensionales
- Redes
 - Grafos
 - Árboles
- Campos
- Espaciales
 - Mapas
 - Geometrías

Fundamentos

Visual mappings

Marcas y Canales

- Puntos (0D)



- Líneas (1D)



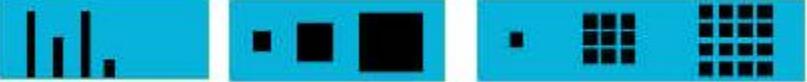
- Áreas (2D)



Marcas y Canales

- Posición:
 - En el plano (2D)
 - Profundidad (3D)
- Color:
 - Tono
 - Saturación
 - Luminancia
- Tamaño:
 - Longitud
 - Área
 - Volumen
- Ángulo
- Curvatura
- Forma
- Patrones:
 - Punteo
 - Textura
- Movimiento

Canales o Variables Visuales (Bertin)

Bertin's Original Visual Variables						
Position changes in the x, y location						
Size change in length, area or repetition						
Shape infinite number of shapes						
Value changes from light to dark						
Colour changes in hue at a given value						
Orientation changes in alignment						
Texture variation in 'grain'						

Los canales también son llamados variables visuales

Principio de expresividad

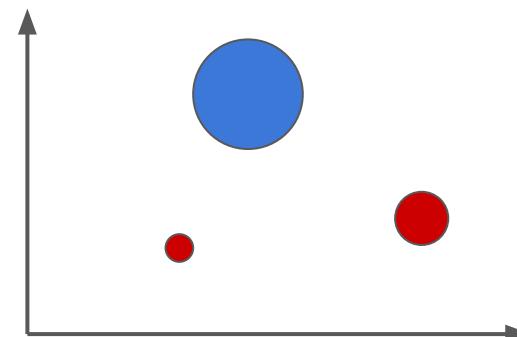
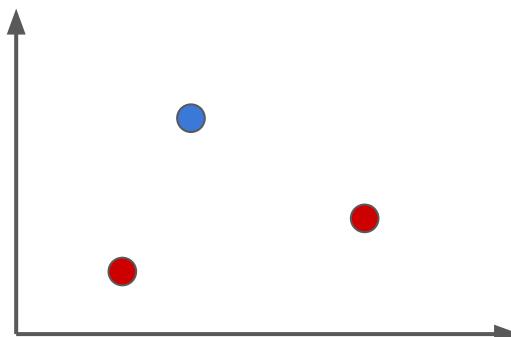
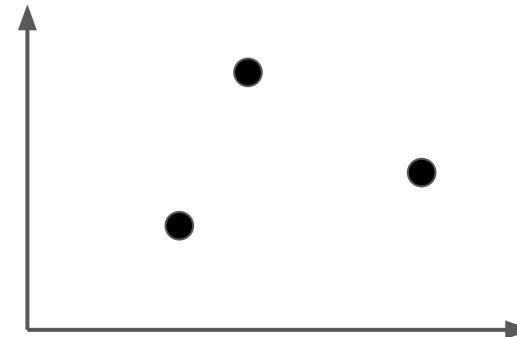
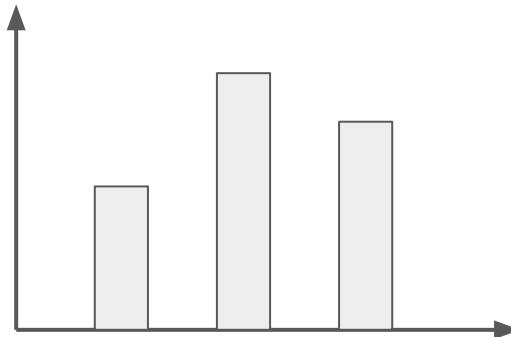
- El mapping debería expresar **toda la información** de los atributos del dataset **y sólo esa información**
- Los atributos **cuantitativos y ordinales** deberían estar mapeados a canales que expresaran su **orden y magnitud**
- Los atributos **categóricos** deberían estar mapeados a canales fácilmente **separables** y que no se interpreten como ordenados

Principio de Eficacia

- Los **atributos más importantes** deben ir mapeados con los **canales más eficaces**
- Como la **posición** es un canal tan eficaz hay que **pensar bien** qué atributo mapear

Mappings

Los canales modifican las marcas según los atributos

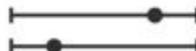


Expresividad y Eficacia

Channels: Expressiveness Types and Effectiveness Ranks

④ Magnitude Channels: Ordered Attributes

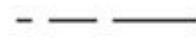
Position on common scale



Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



④ Identity Channels: Categorical Attributes

Spatial region



Color hue



Motion



Shape



↑ Most
Effectiveness
↓ Least
Same

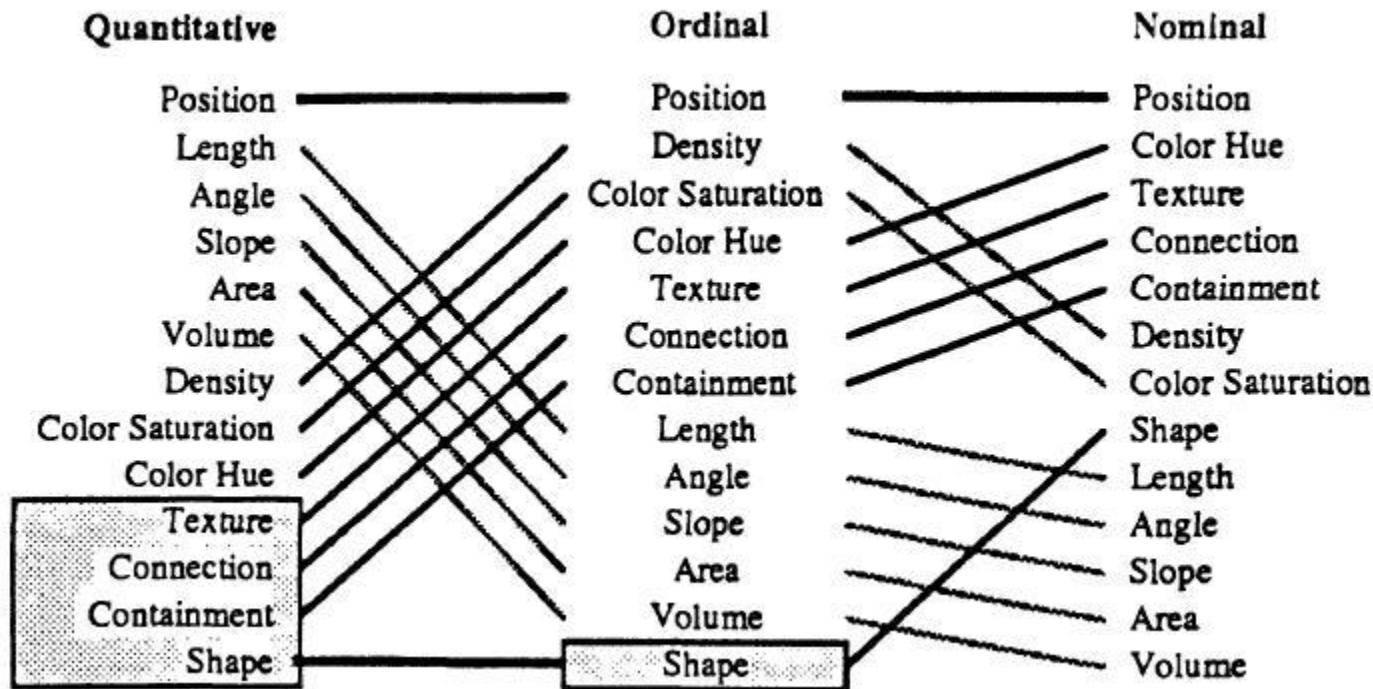
Tamara Munzner

Criterios de Eficacia

- Precisión
- Discriminación
- Separación
- Preatención
- Agrupación

Precisión

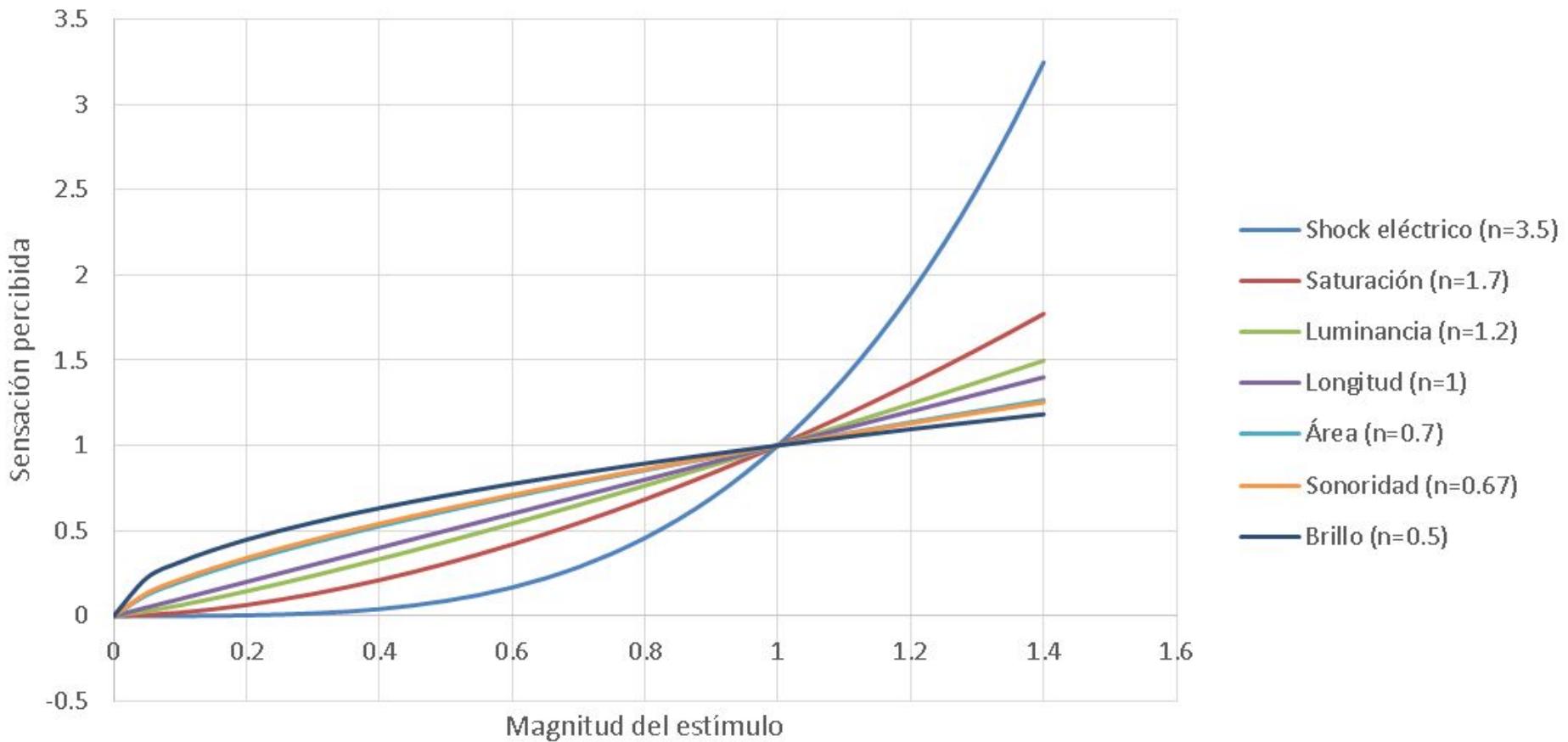
Se comparan atributos mapeados a diferentes canales y se calcula el error



Mackinlay's Perceptual Tasks

Precisión

Ley de Stevens

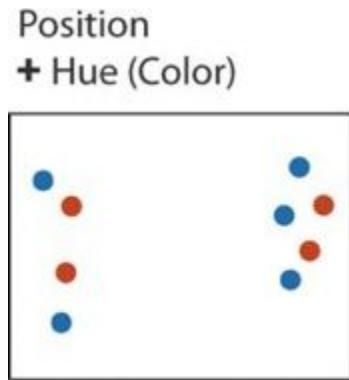


Discriminación

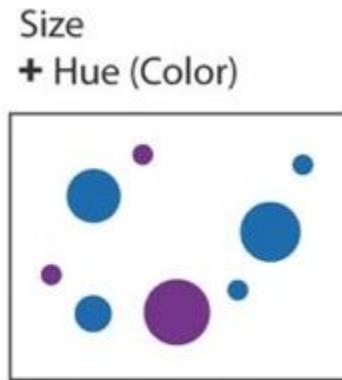
- Un canal tiene mucha discriminación si permite al usuario percibir muchos "**escalones**" de valores
- Hay que pensar en el **número de clases** que tiene el **atributo** a representar y el número de clases que el usuario puede **diferenciar** sin problemas dependiendo del **canal**. Ej: Color < 20
- Además **hay interferencias** entre canales: línea muy gruesa -> rectángulo

Separación

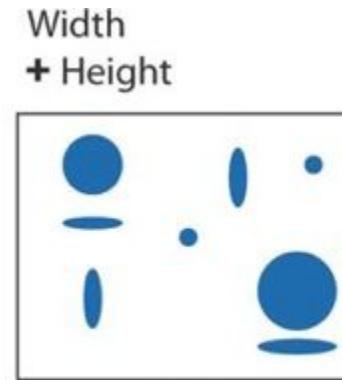
- No todos los canales son completamente independientes entre sí
- La codificación visual será fácil de realizar si se utilizan canales separables



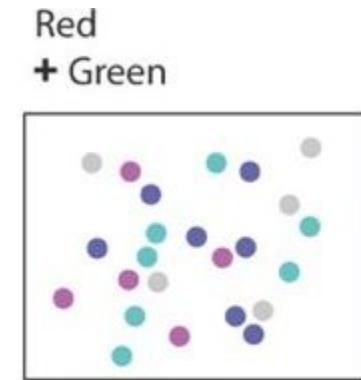
Fully separable



Some interference



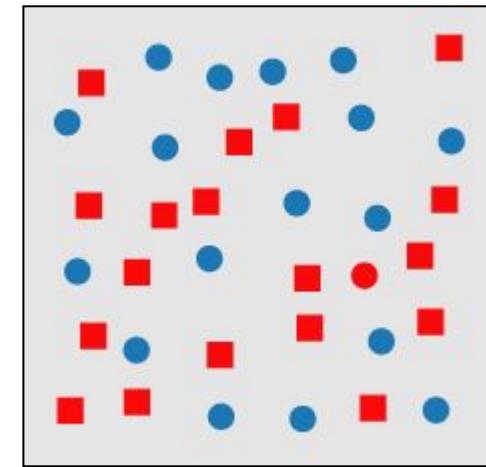
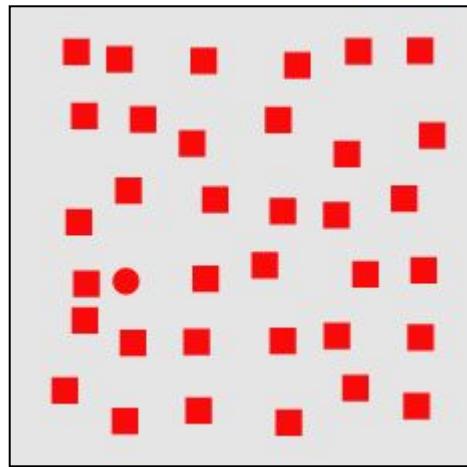
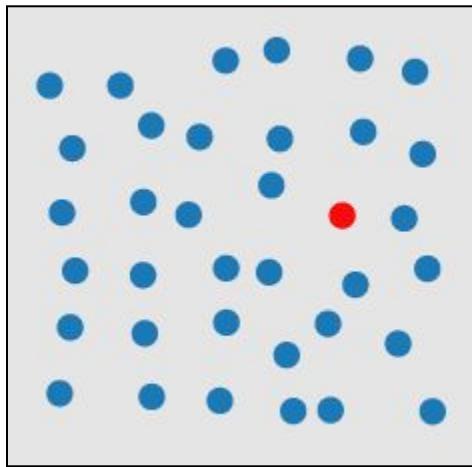
Some/significant
interference



Major interference

Tamara Munzner

Preatención

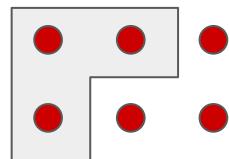


La potencia varía entre canales y además sufre con la conjunción

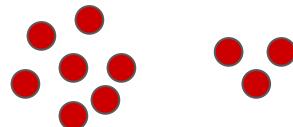
Más sobre percepción en: <http://www.csc.ncsu.edu/faculty/healey/PP/>

Agrupación

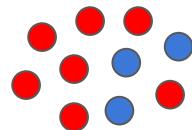
- Que varios ítems pertenezcan a un grupo puede codificarse **con marcas o con semejanza en canales**



marcas



proximidad



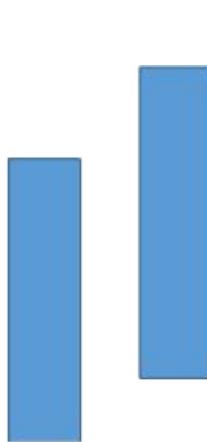
similaridad

Fundamentos

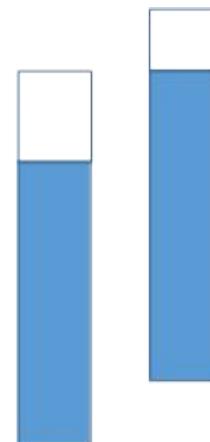
Ley de Weber

Todo es relativo

- La percepción humana sigue la **Ley de Weber**: la cantidad mínima detectable en la intensidad de un estímulo I es un valor fijo proporcional a su magnitud $I_{\min} = I * R$
- Es decir, **percibimos los estímulos de manera relativa y no absoluta**



Rectángulos
no alineados

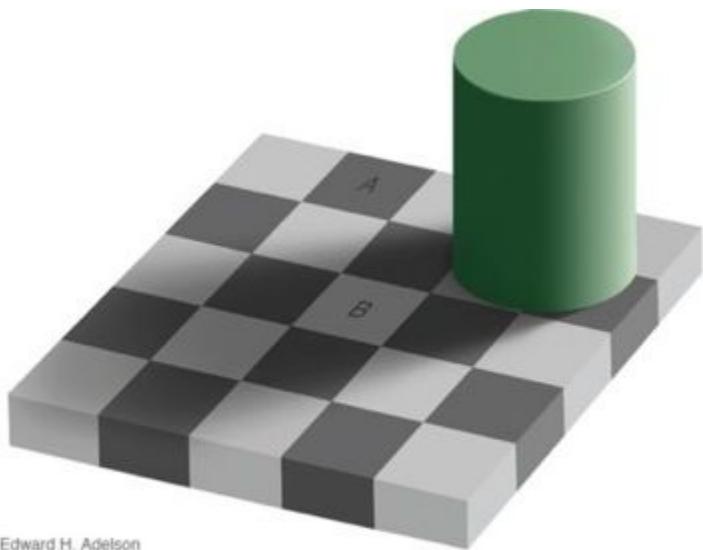


Rectángulos
no alineados

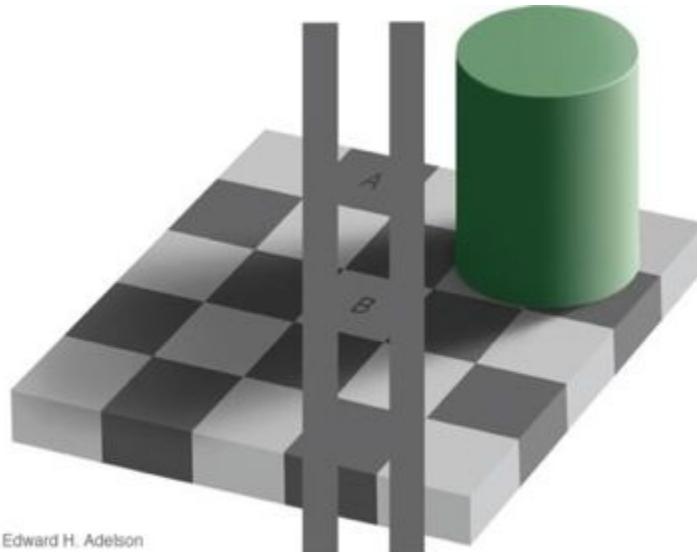


Rectángulos
alineados

Todo es relativo

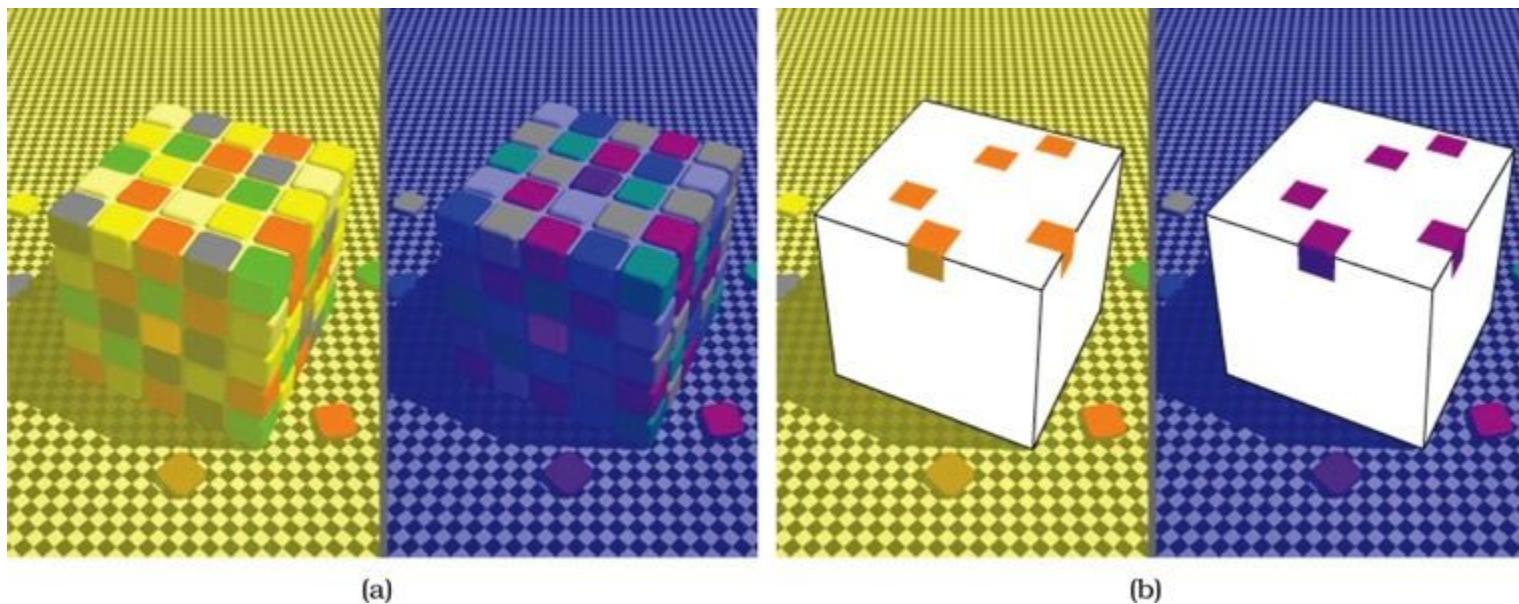


(a)



(b)

Todo es relativo

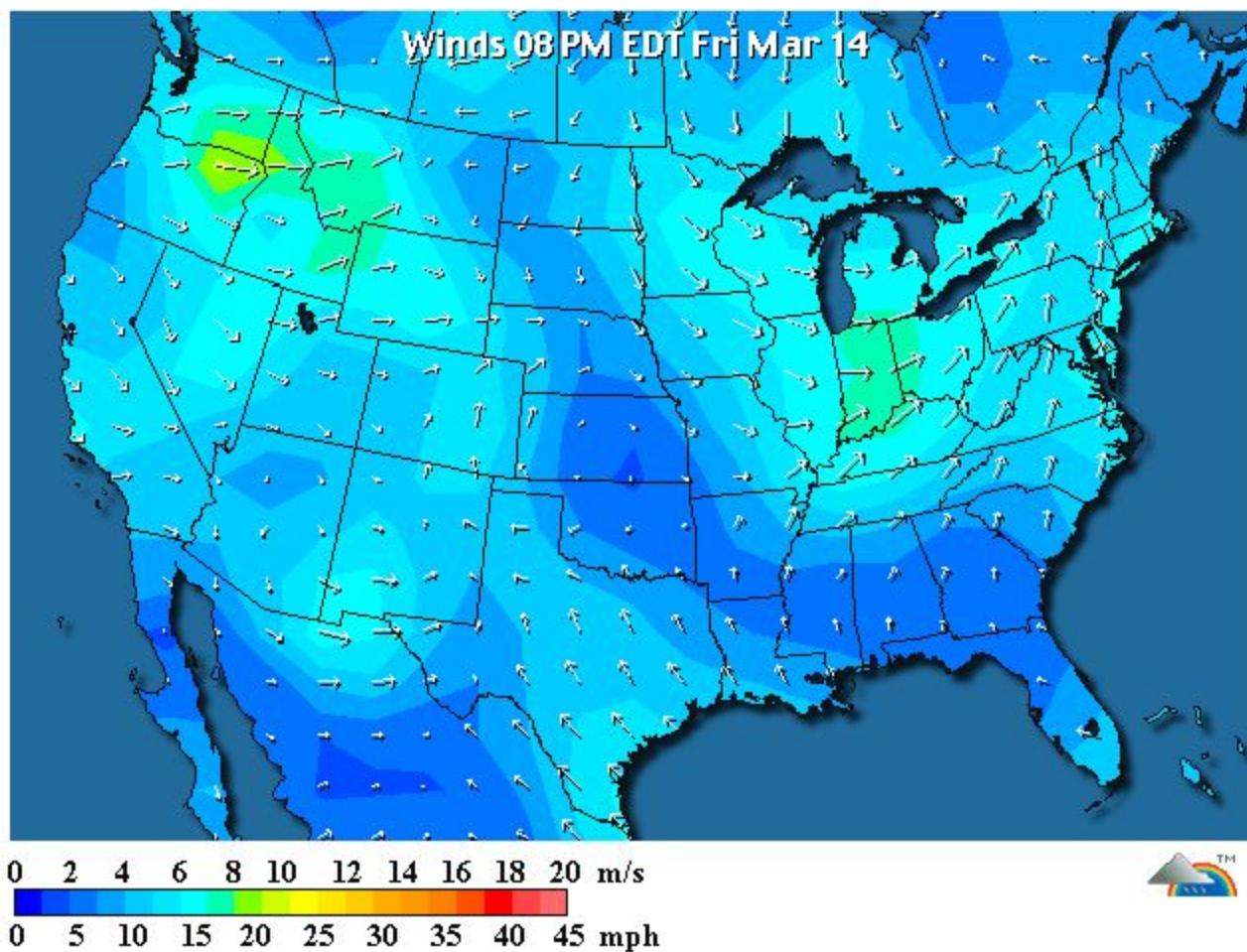


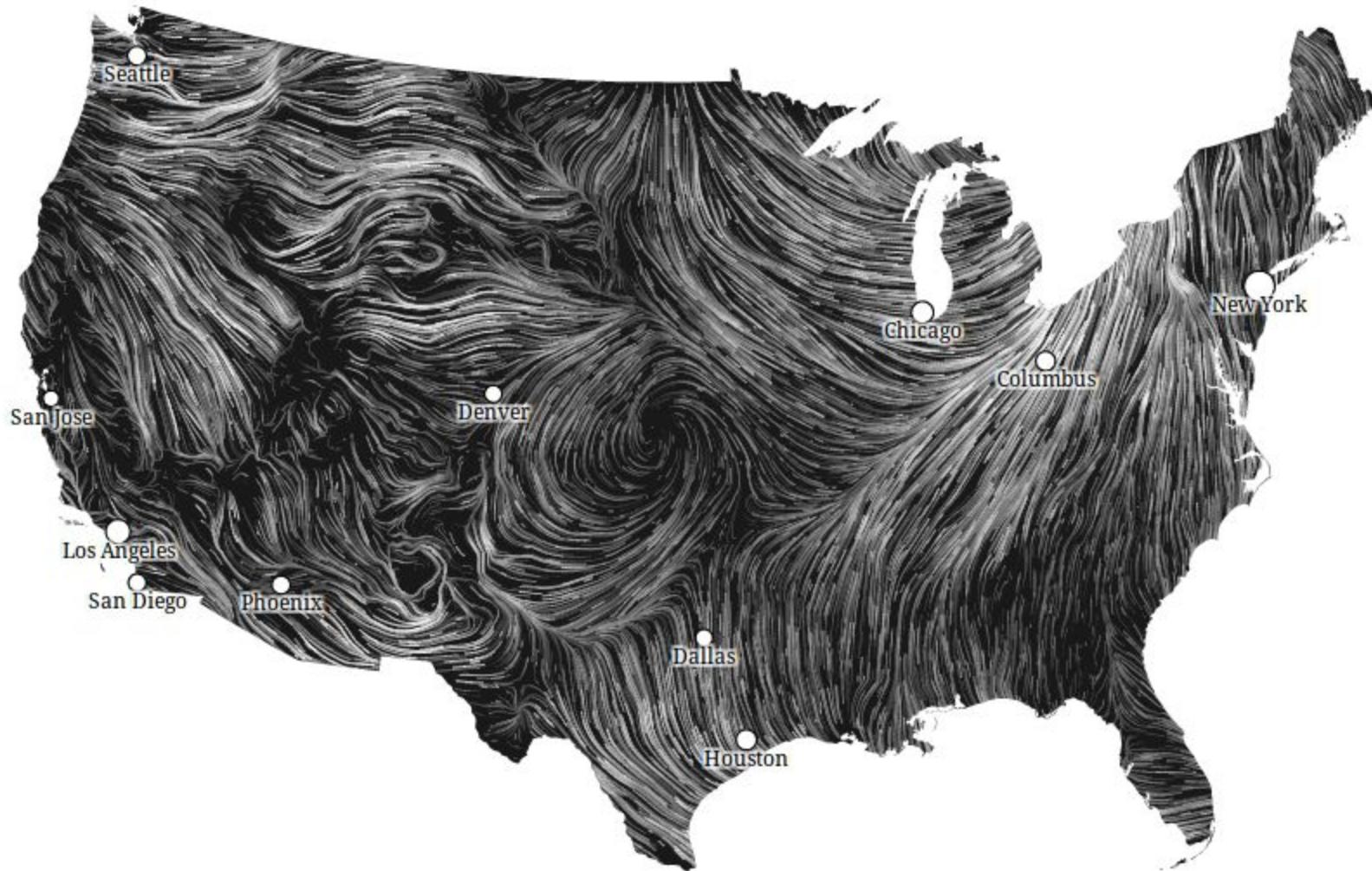
Todo es relativo



Índice

1. Introducción
2. Fundamentos
- 3. Casos**
4. Visualización en Big Data
5. Datos Tabulares
6. Datos Temporales
7. Datos Espaciales
8. Redes y Jerarquías

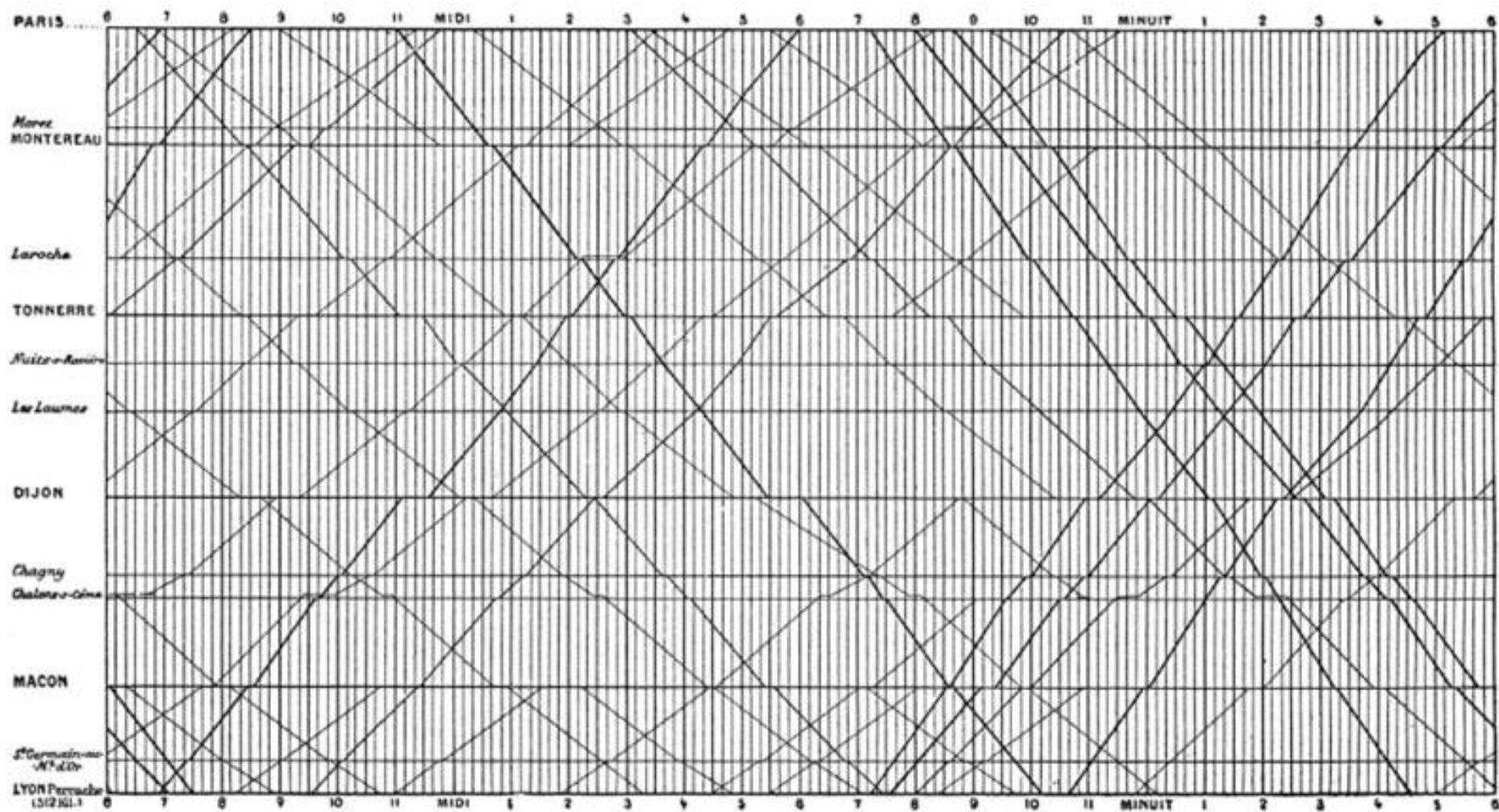




Variación del precio de la luz

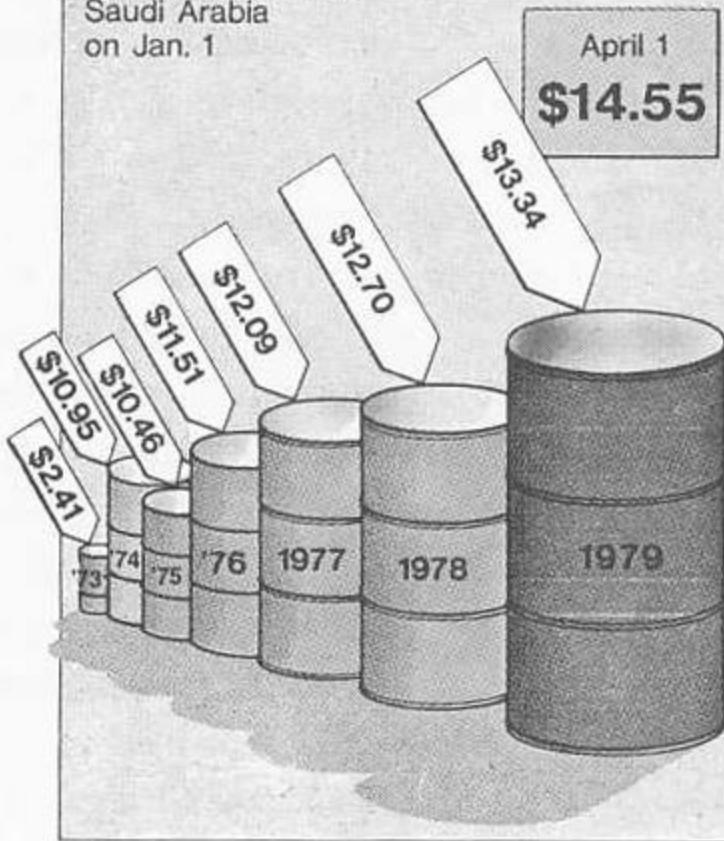






IN THE BARREL...

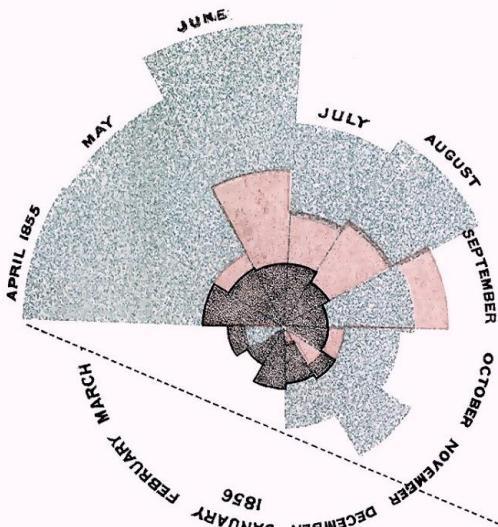
Price per bbl. of
light crude, leaving
Saudi Arabia
on Jan. 1



**DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.**

2.

APRIL 1855 TO MARCH 1856.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

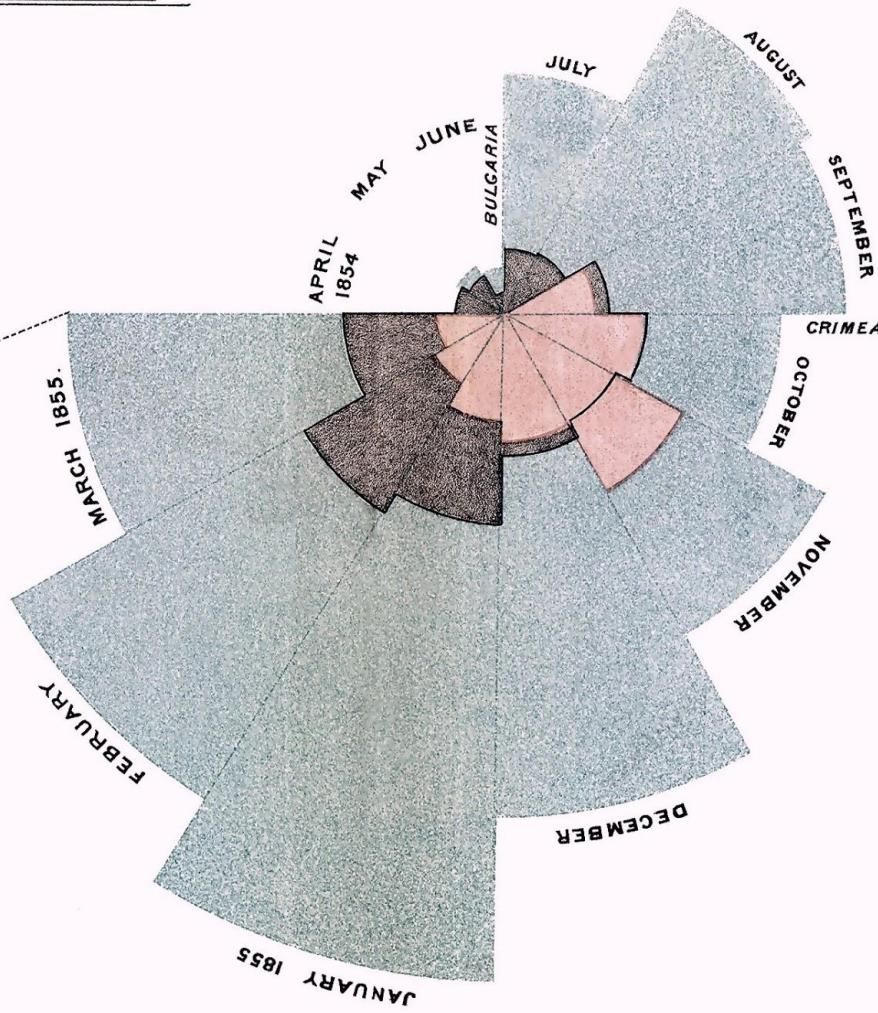
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.

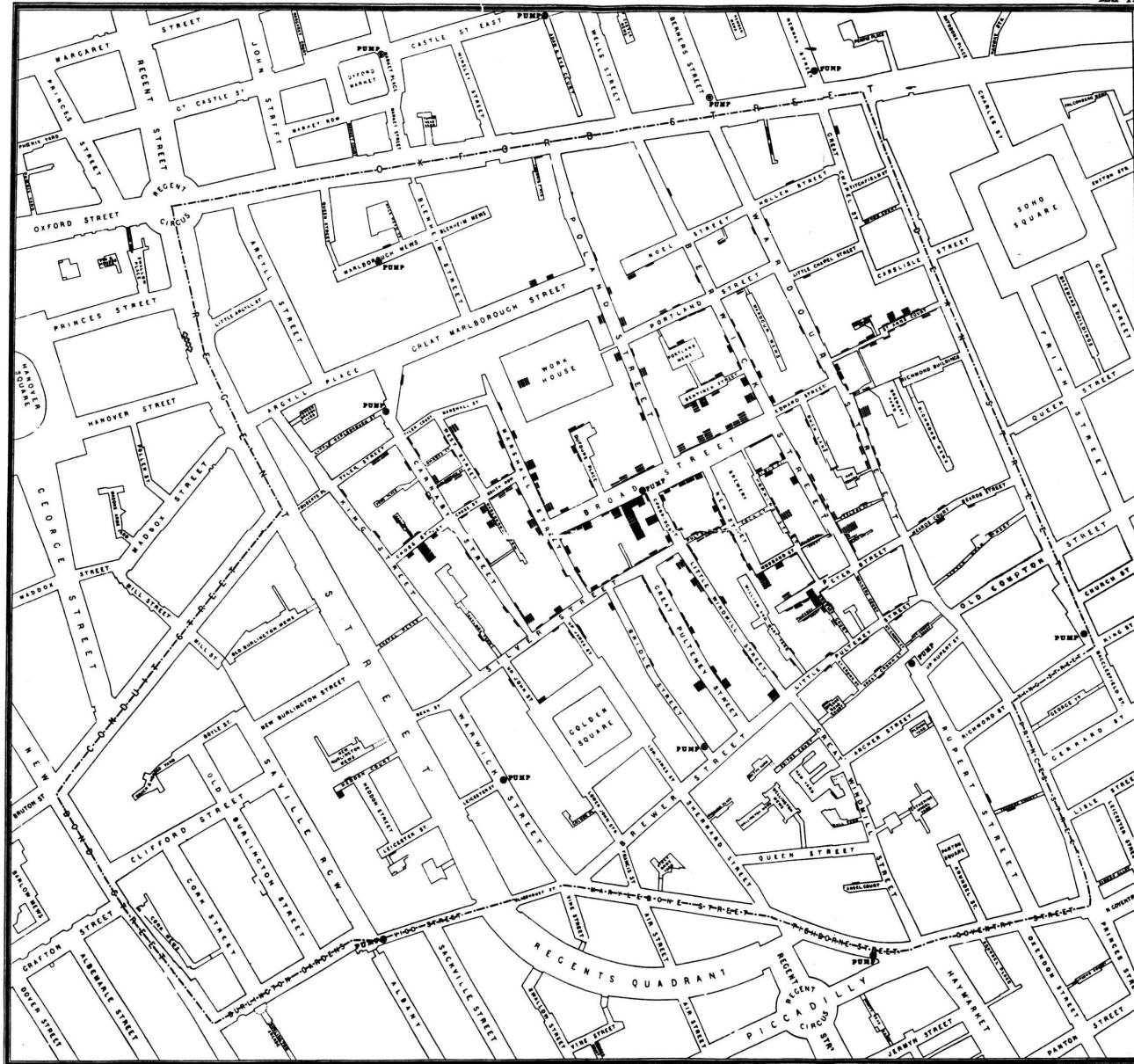
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

1

APRIL 1854 TO MARCH 1855.





Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en trèfles des zones. Le rouge désigne les hommes qui ont été en Russie; le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Séguir, de Fezensac, de Chambray et le journal médical de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk et Mohilow et qui rejoignirent Orsha et Witebsk, avaient toujours marché avec l'armée.

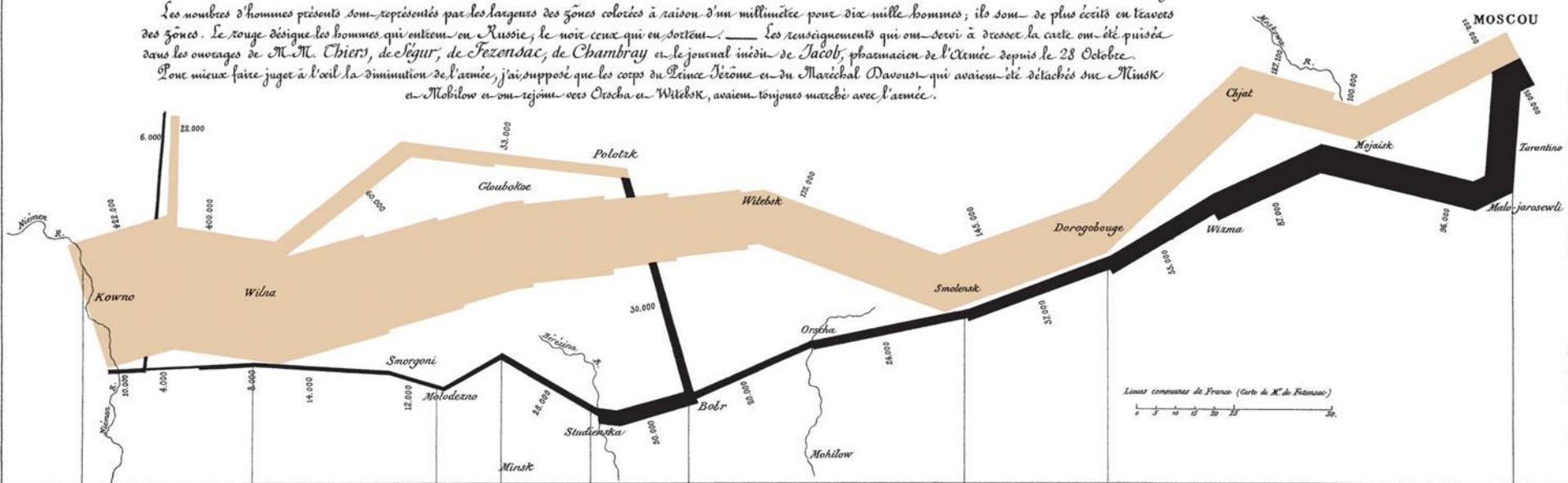
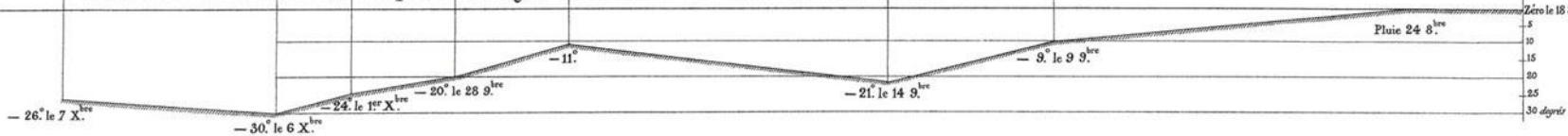
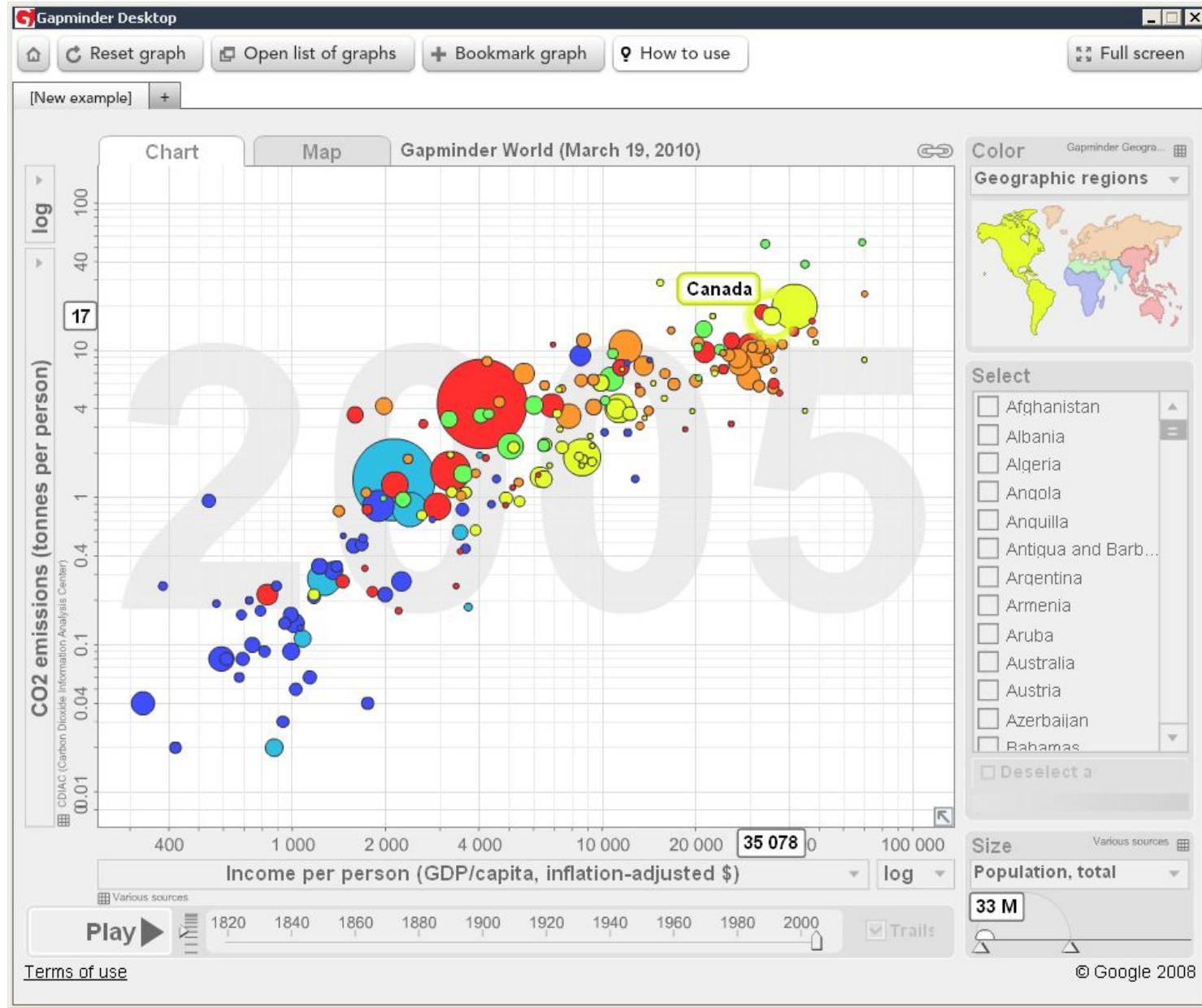


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



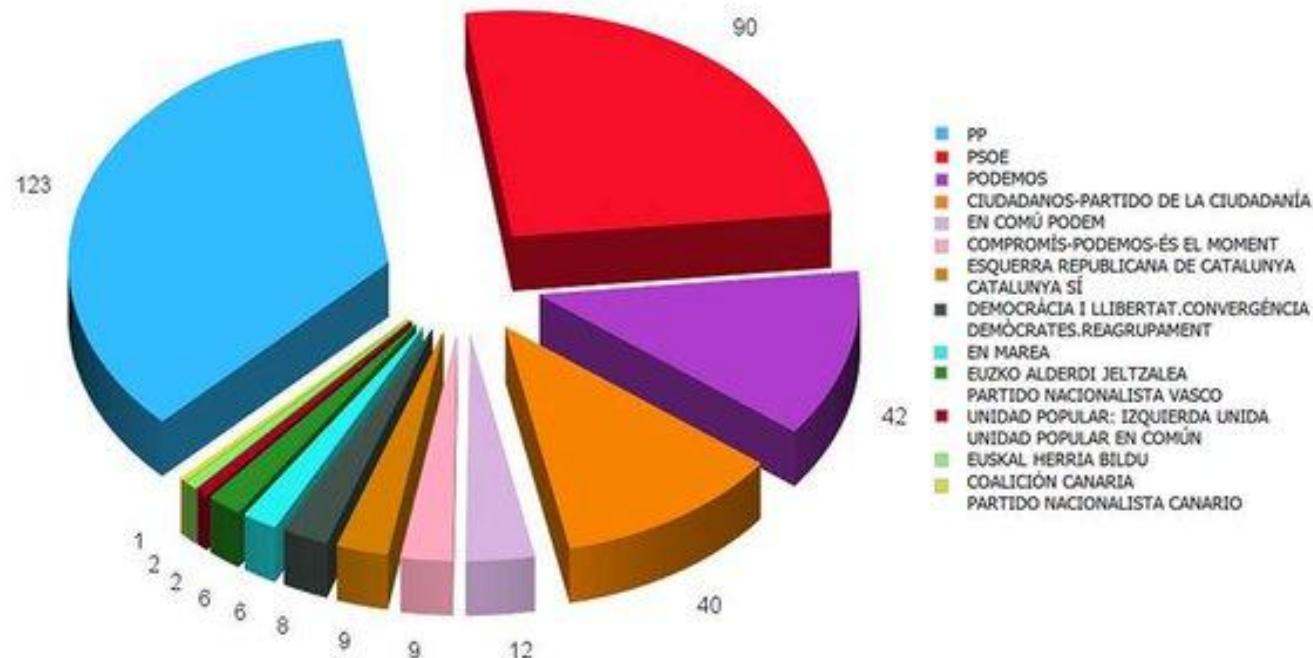
Les Cosaques passent au galop
le Niemen gelé.



Resultados electorales 2015



Congreso de
los Diputados



Fuente: Ministerio del Interior

Índice

1. Introducción
2. Fundamentos
3. Casos
- 4. Visualización en Big Data**
5. Datos Tabulares
6. Datos Temporales
7. Datos Espaciales
8. Redes y Jerarquías

Problemas añadidos en Big Data

Escalabilidad

Computacional

- La interacción necesaria para explorar es demasiado lenta
- Imposible renderizar por falta de recursos computacionales

Escalabilidad Visual

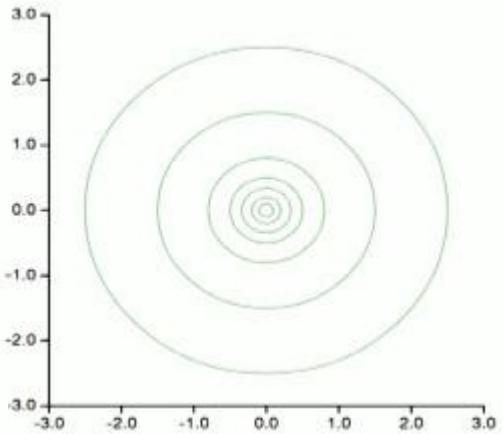
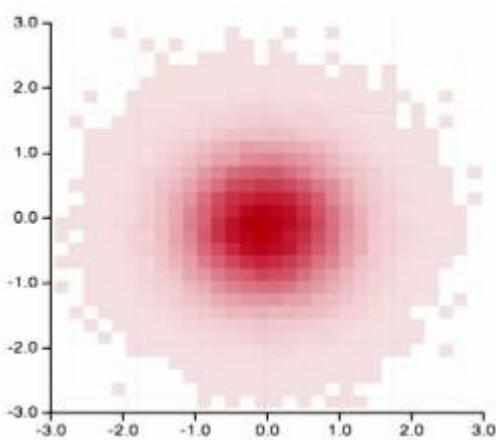
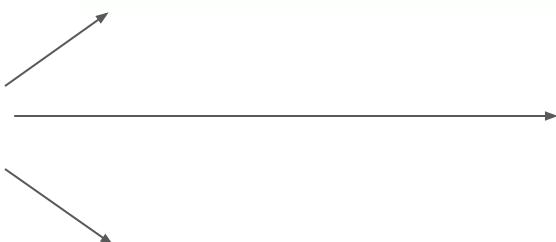
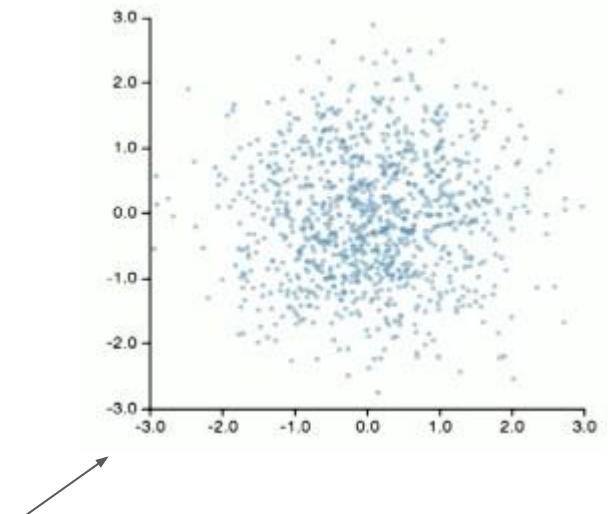
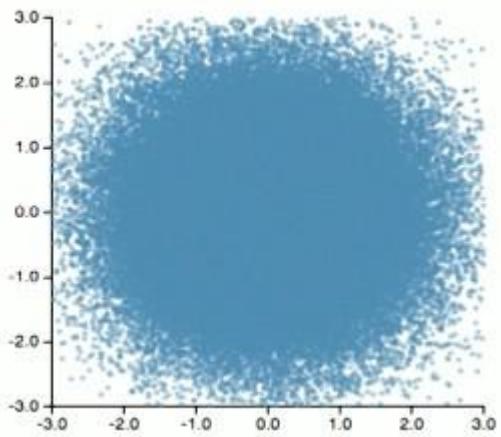
- Cada visualización tiene sus propios problemas de escalado
- "Bajar opacidad y mostrar todo" también se queda pequeño

Tratamiento diferenciado en Big Data

Nos podemos encontrar con datos complejos en:

- Número de Items
 - Problemas al hacer el mapping visual
 - Problemas para tener interacción fluída
- Número de dimensiones
 - Reducción de dimensionalidad (PCA, t-SNE, StarPlots ...)
 - Submuestreo de dimensiones (interactivo, selección de variables ML)
- Variedad de tipos de datos
 - Dashboards: Si tenemos que monitorizar
 - Múltiples vistas enlazadas: Si es para análisis
 - Power Walls: En centros de control y tareas críticas

Estrategias



Estrategias

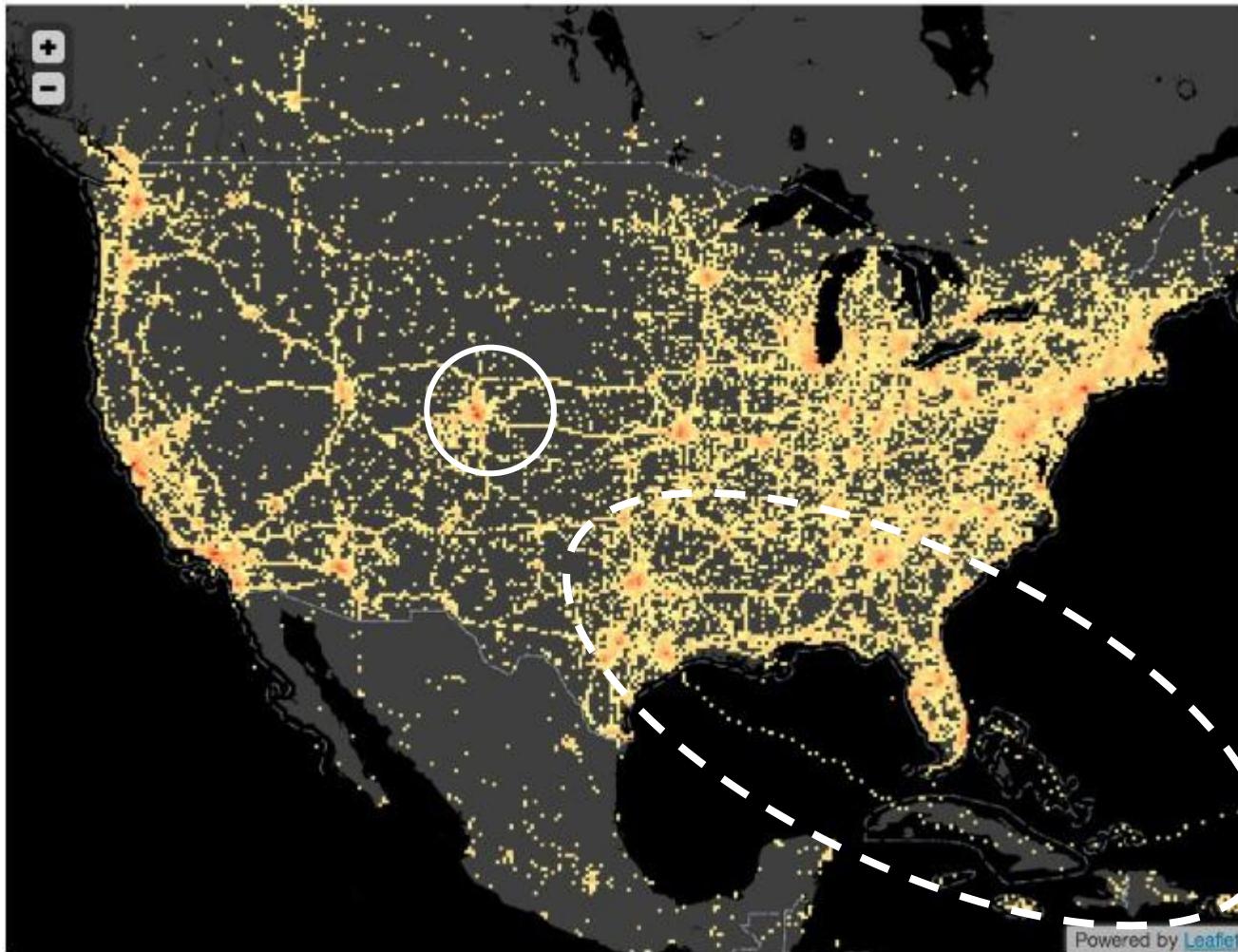
- Muestreo aleatorio → Reducimos el número de puntos
- Modelado → Restringe la exploración de datos
- Agregaciones
 - De datos → Perdemos resolución
 - Métricas resúmen → Modelado Básico
 - Densidad teniendo en cuenta percepción → **Todos los datos se perciben**
- Visualización Bottom - Up
 - Empieza con una búsqueda
 - Expande el resultado

3000 Millones de Checkins: Sampling



imMens - Jeff Heer

3000 Millones de Checkins: Binning



imMens - Jeff Heer

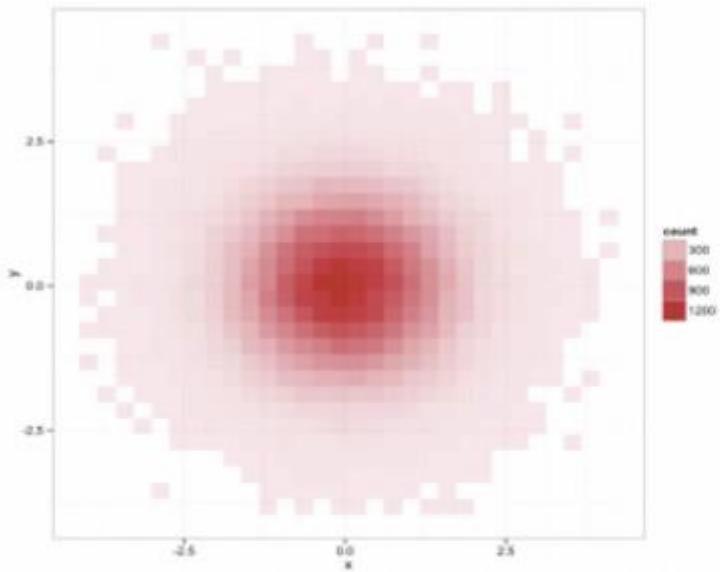
¿Cómo Agregamos?

- Los Categóricos ya están agregados pero:
 - Si hay muchas categorías: Crear "others" o mostrar "las primeras"
- Los Cuantitativos:
 - Seleccionar los intervalos (uniformes, cuantiles, ...)
 - Límite superior de bins: Píxeles de la pantalla
- Temporales:
 - Día, Semana, Mes...
- La operación de agregación también se diseña: Count, Sum, Avg, ...

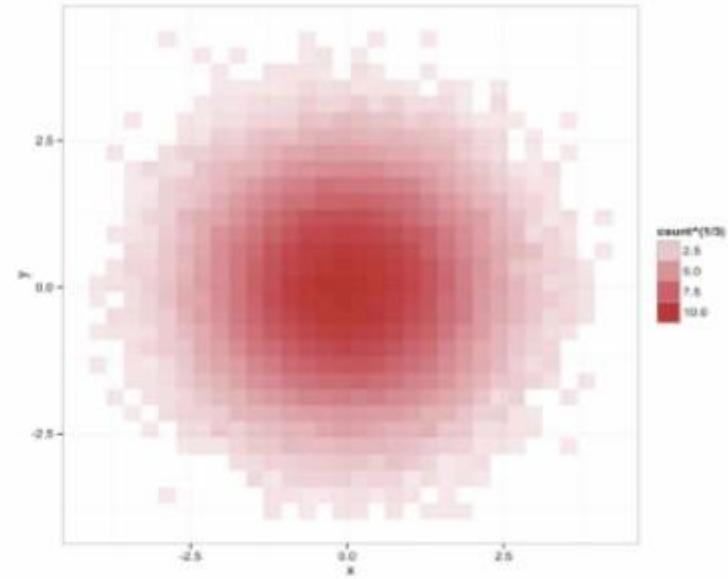
¿Cómo representamos visualmente?

- Agregación a 1D: Utilizamos lo más preciso
 - Longitudes ó posiciones
- Agregación a 2D: Áreas o Colores
 - Aunque las áreas son más precisas
 - El **color es preferible**, escala mejor por uso de pocos píxeles

Escala de color



Linear Alpha Interpolation
is not perceptually linear.



Cube-Root Alpha Interpolation
approximates perceptual linearity.

imMens - Jeff Heer

Percepción lineal

$$Y = \alpha + \left(\frac{\hat{x} - x_{min}}{x_{max} - x_{min}} \right)^\gamma (1 - \alpha)$$

↓ ↓ ↓

Discontinuidad para ver outliers: e.g. 0,15

El valor del bin después de la agregación

Ajusta la rampa de valores a la curva luminosa que percibimos: 1 / 3

↑

El valor de luminosidad final: [0 , 1]

Interacción Fluída

- Los bins son esenciales:
 - **El tiempo de renderizado va a ser constante**
- Por resolver: **tiempo de Query**
 - Difícil con Spark o Storm (Hadoop impossible)
 - Bases de datos OLAP:
 - Hiper cubos
 - Localidad de memoria (Caches)
 - Queries en GPUs

Índice

1. Introducción
2. Fundamentos
3. Casos
4. Visualización en Big Data
- 5. Datos Tabulares**
6. Datos Temporales
7. Datos Espaciales
8. Redes y Jerarquías

Características

- También conocidos como datos Multivariados o Multidimensionales
- Formatos comunes:
 - Excel(xls, xlsx)
 - texto (csv, tsv, ...)
 - Bases de datos
- Formado por **items**, muestras o individuos (filas) y **atributos**, dimensiones o variables (columnas)

El problema

- **Muchas dimensiones que representar gráficamente de forma efectiva**
- Informalmente: “No podemos ver más de 3 dimensiones”
 - Sólo cierto si cada dimensión se mapea, en un mismo espacio, utilizando la variable visual “**Posición**”.

Objetos de estudio típicos

- Correlaciones
- Regresiones
- Clasificaciones
- Agrupaciones (clustering)
- Análisis de varianza

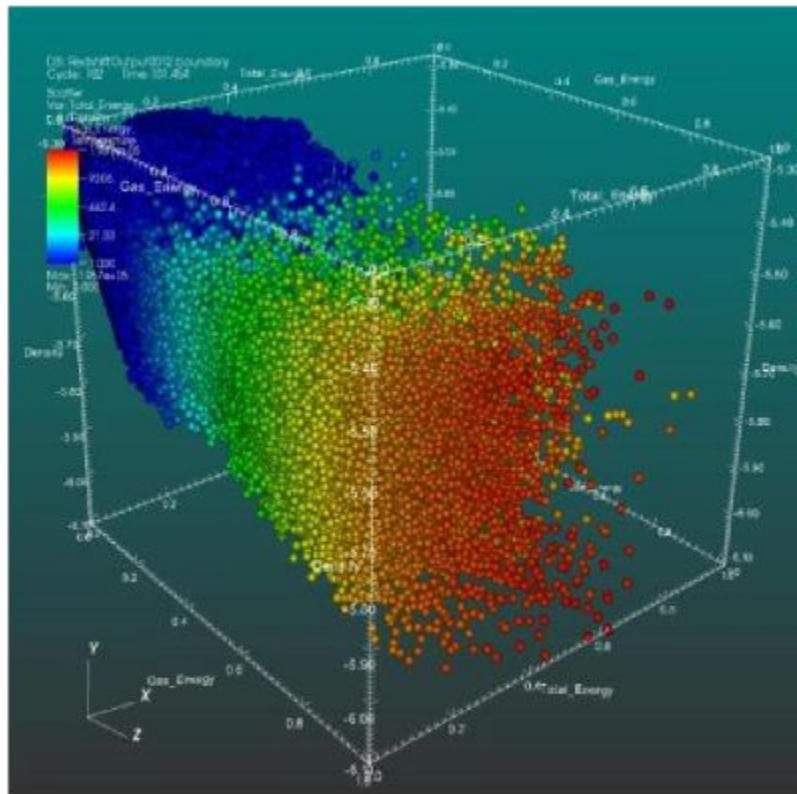
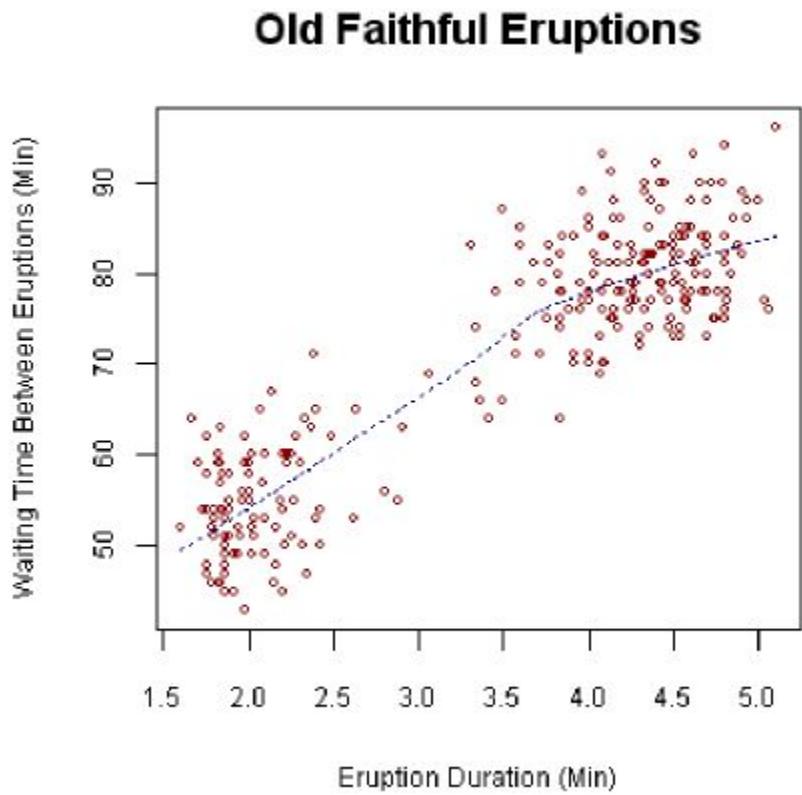
Datos Tabulares

Representaciones basadas en puntos

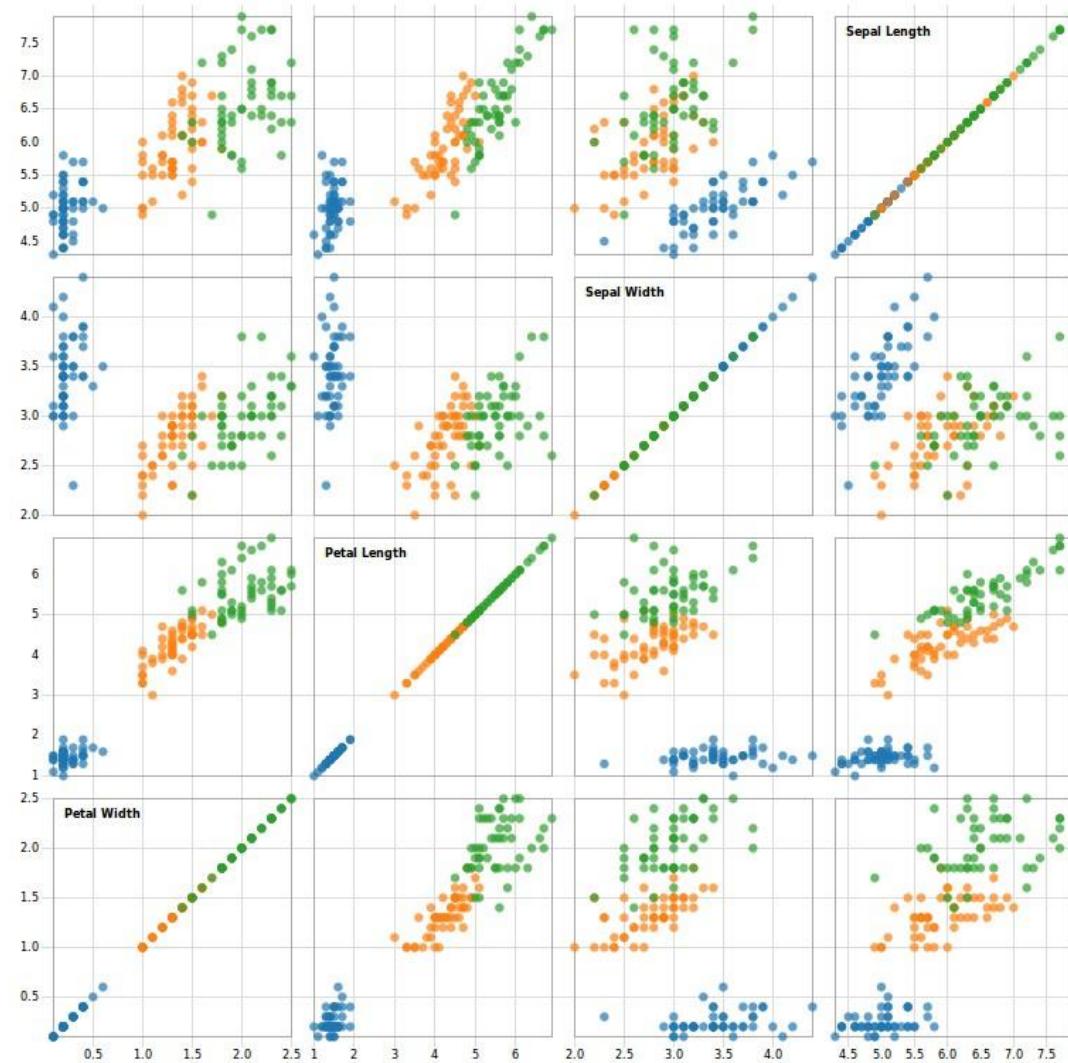
Estrategias

- Submuestreo de dimensiones
- Multiples vistas
- Reducción de la dimensionalidad

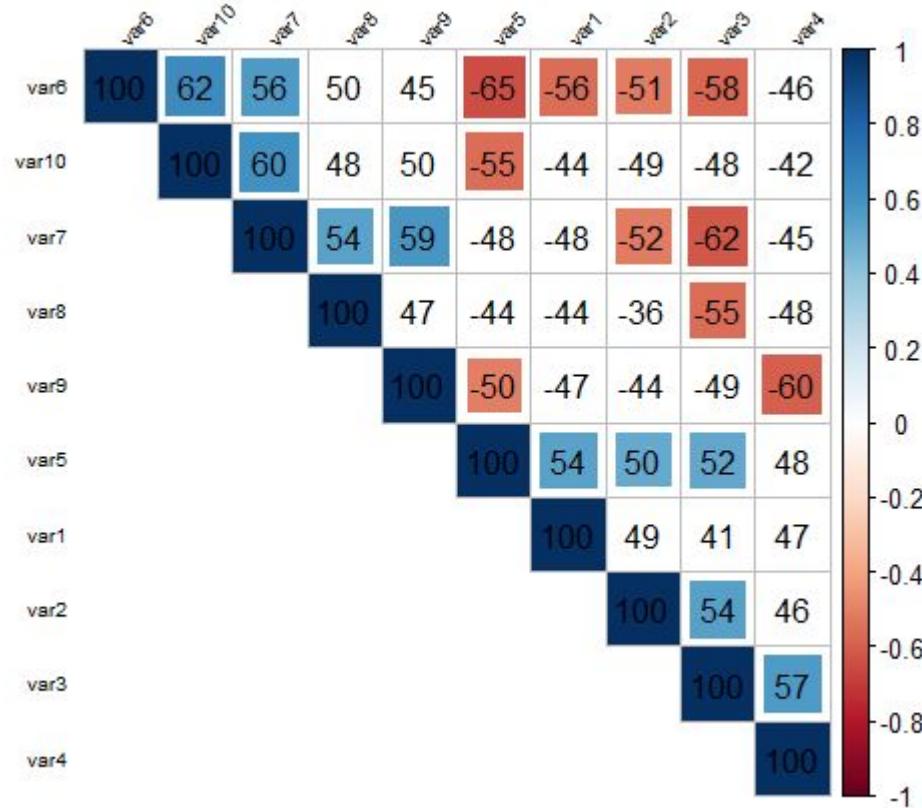
Scatter Plot



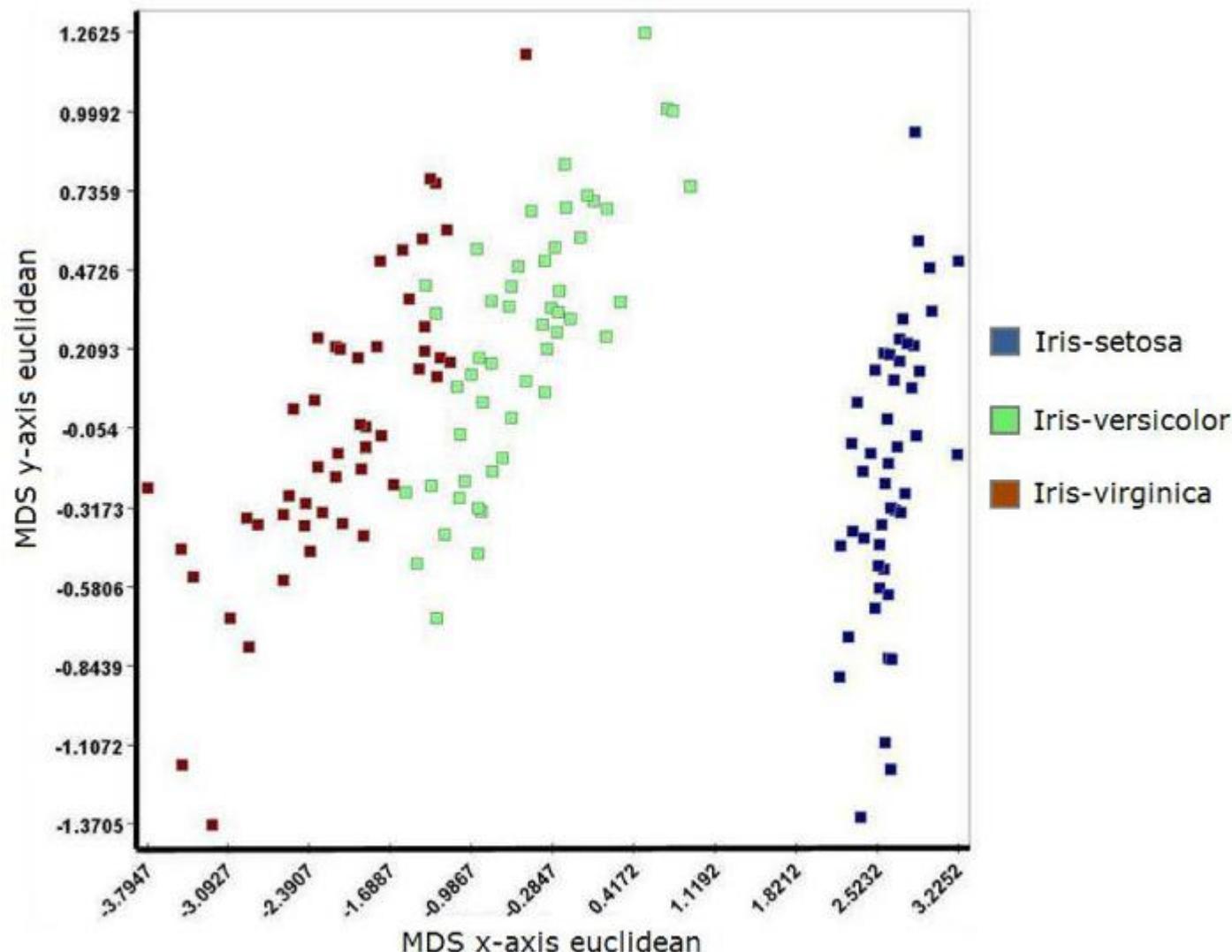
Scatter Plots Matrix



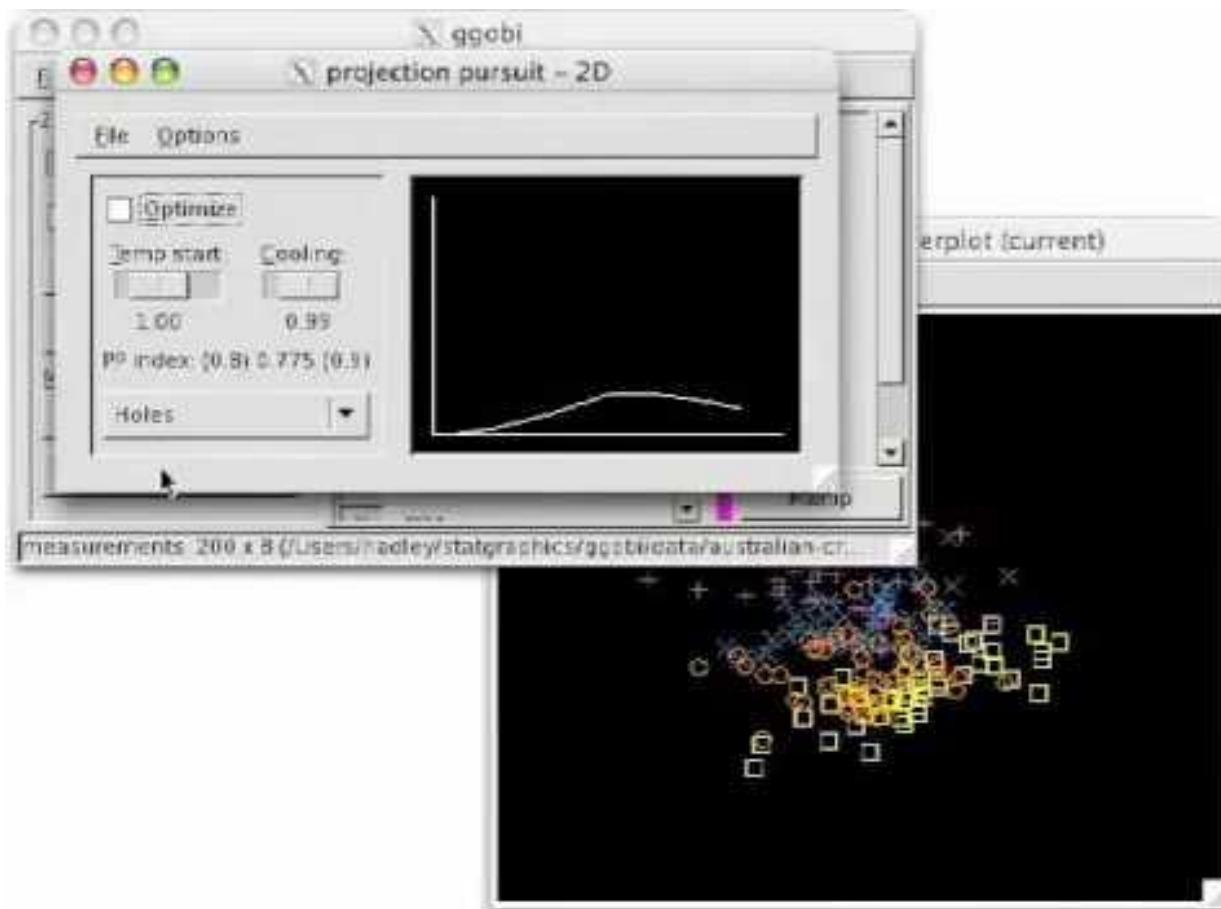
Correlation Matrix



Reducción de dimensionalidad



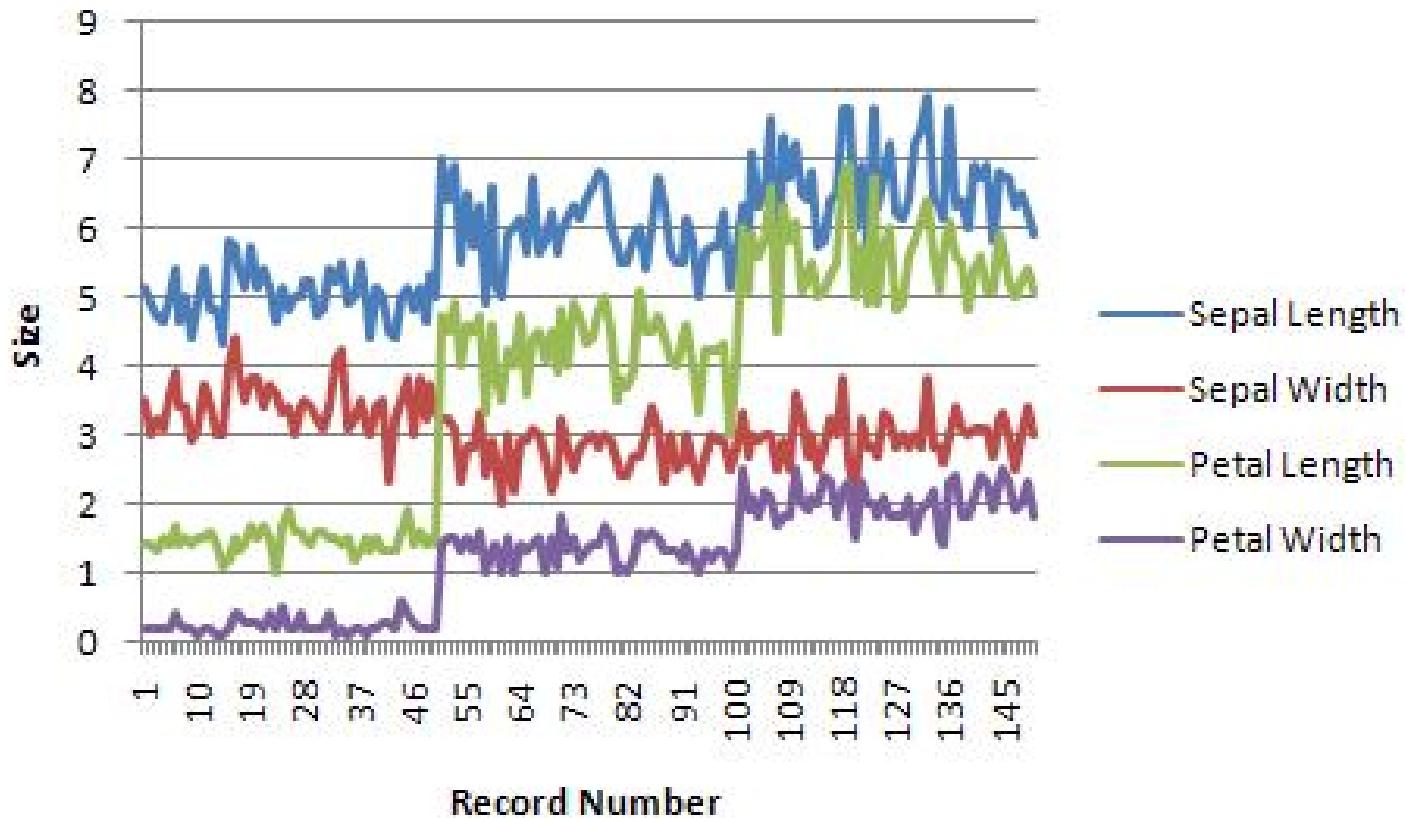
Grand Tour



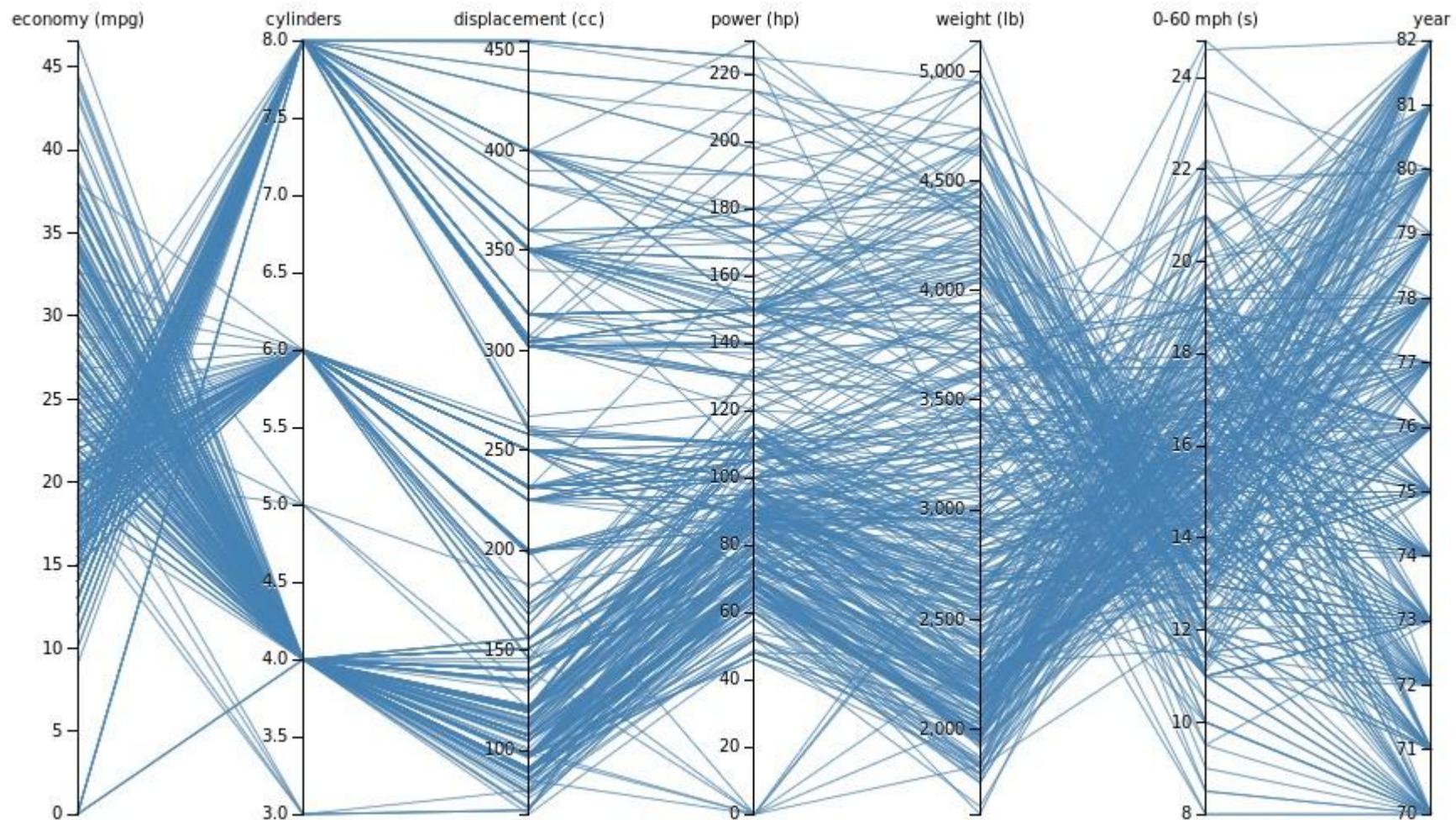
Datos Tabulares

Representaciones basadas en líneas

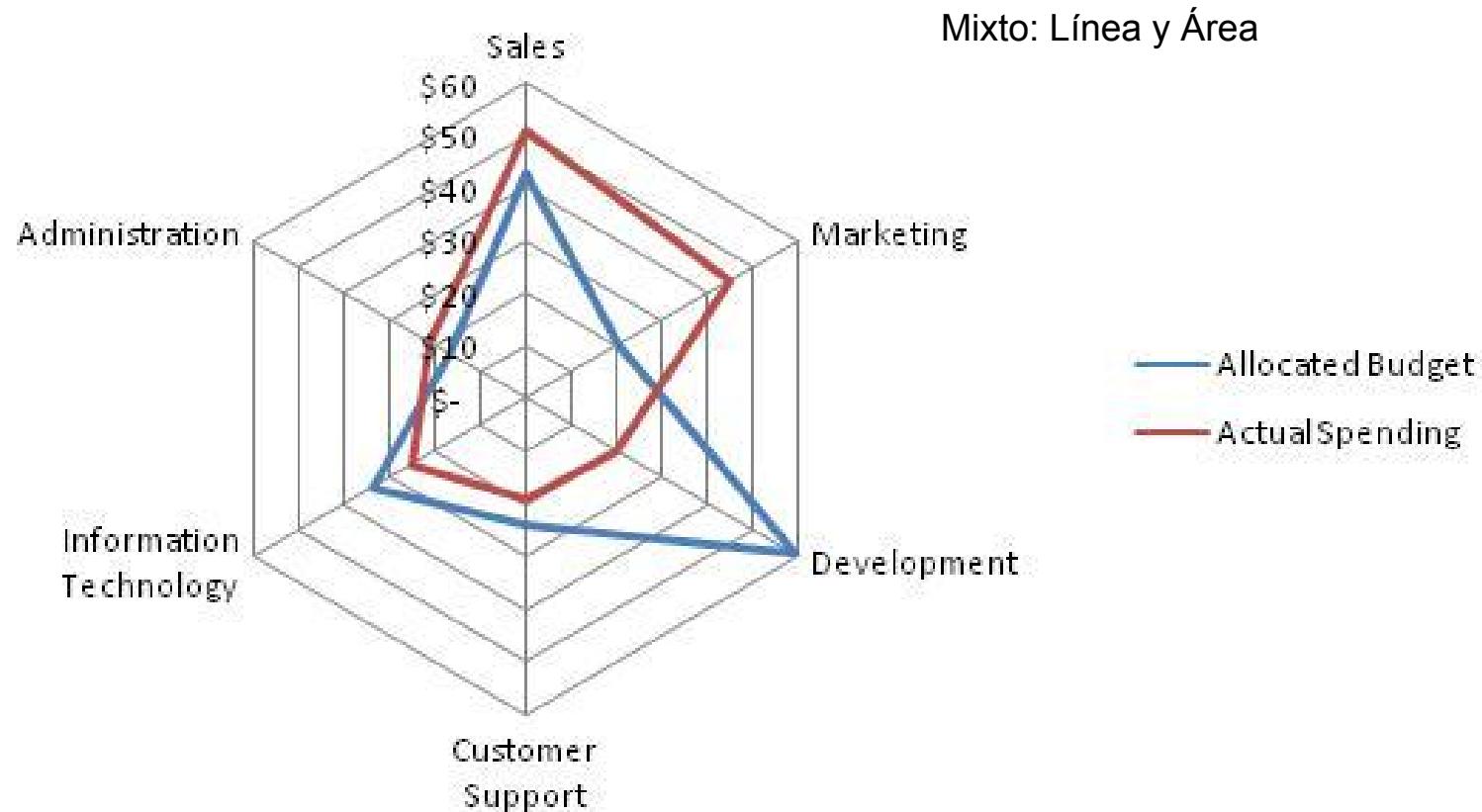
Gráfico de líneas multivariado



Coordenadas Paralelas



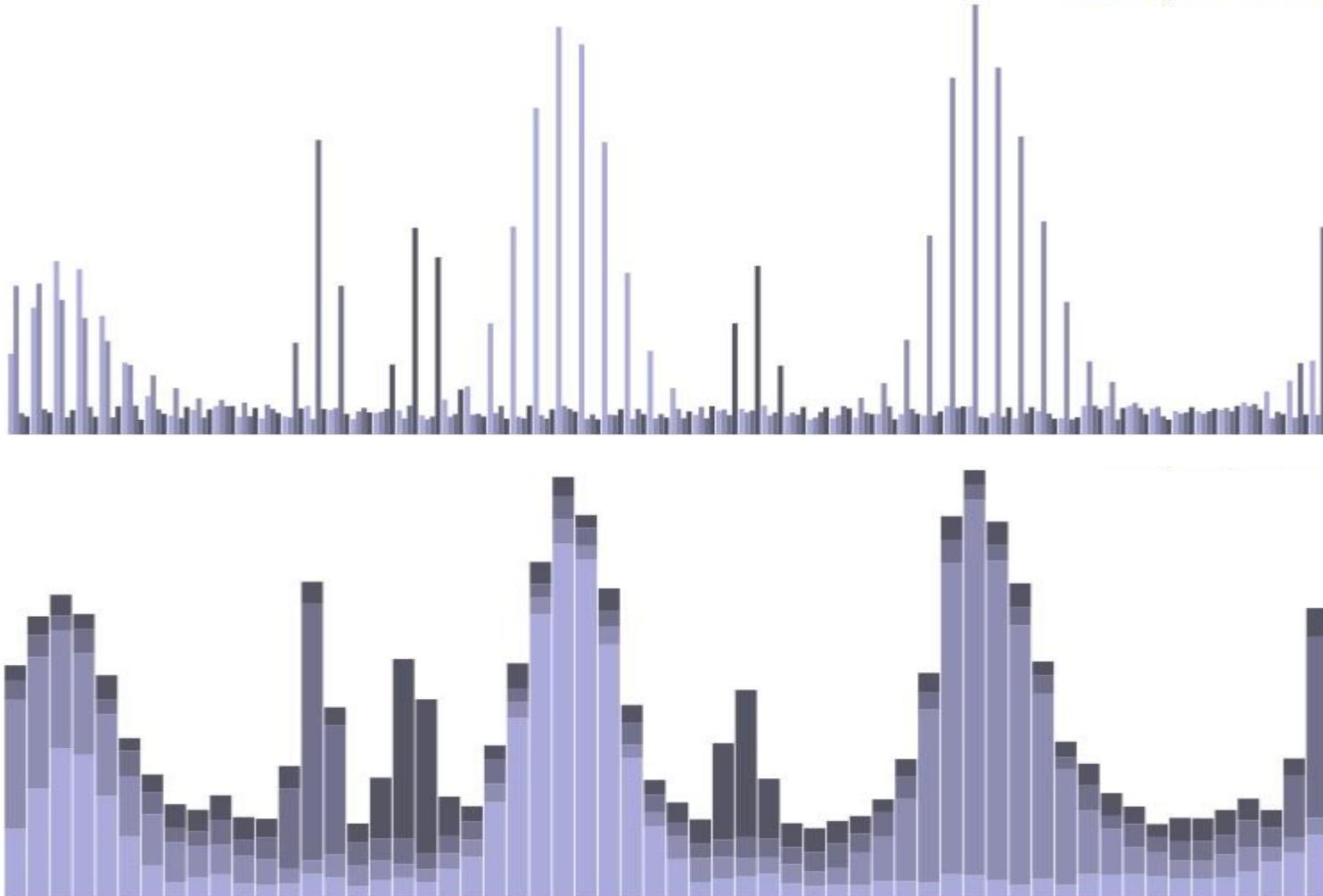
Gráfica de Radar



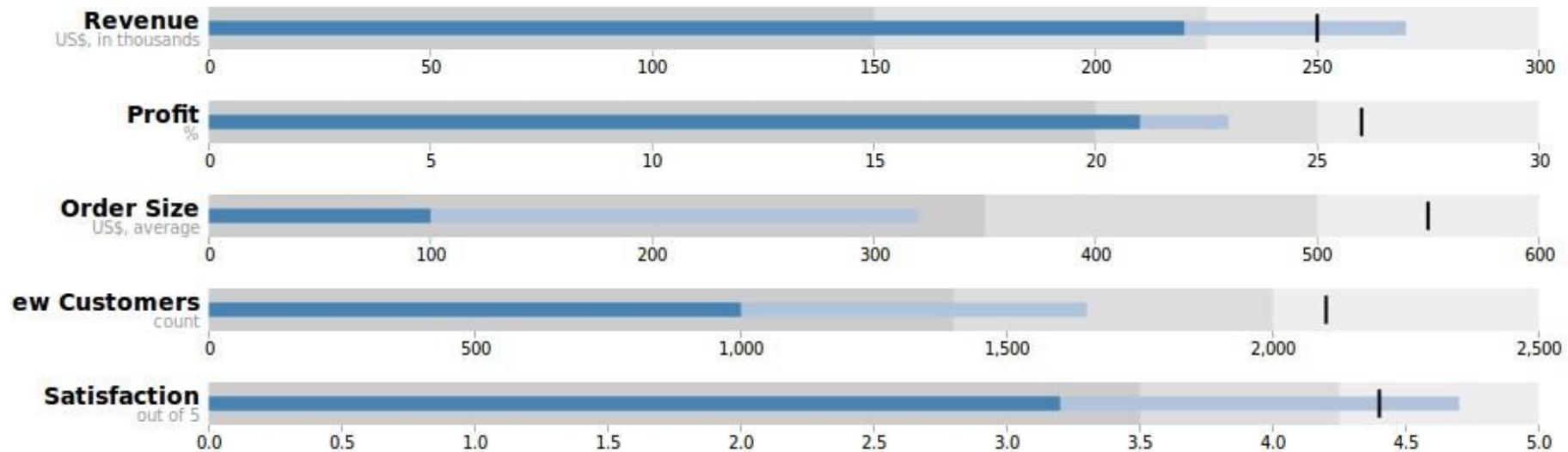
Datos Tabulares

Representaciones basadas en áreas

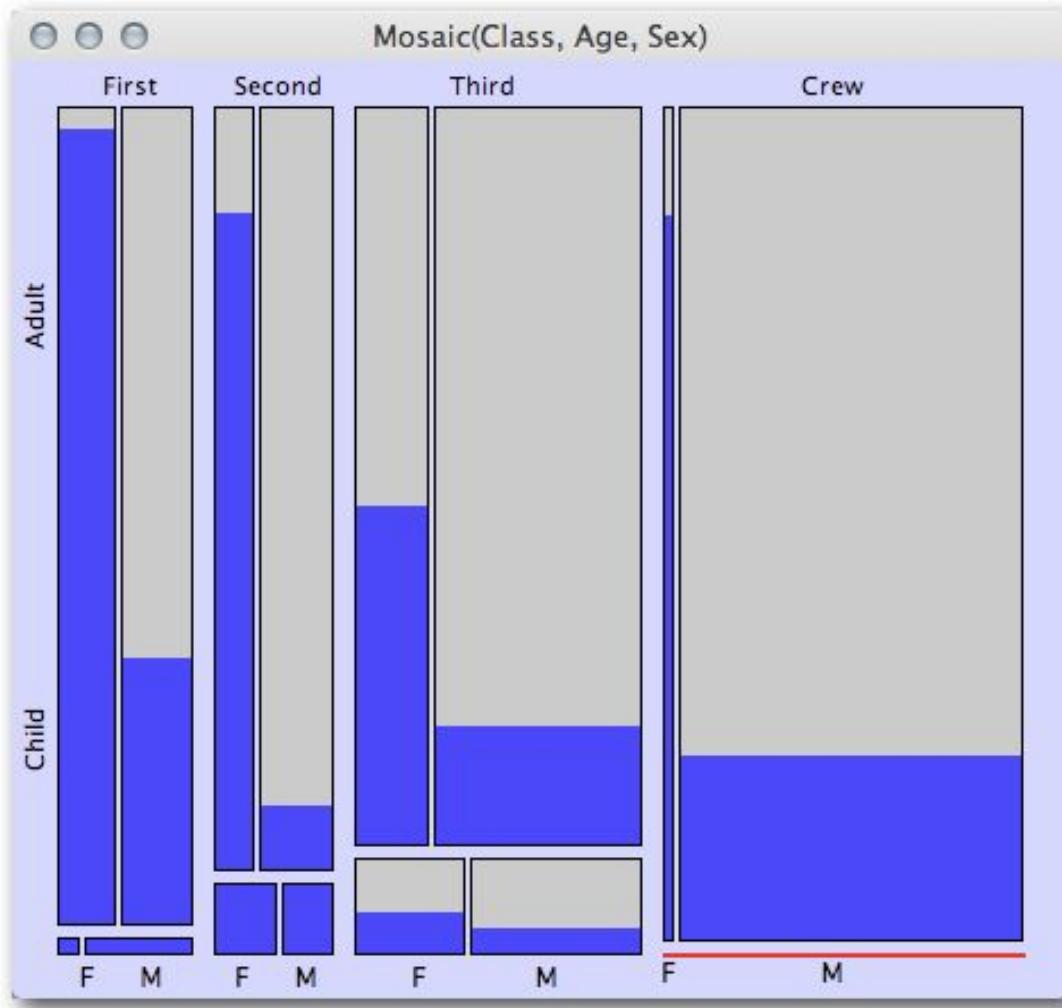
Diagrama de barras agrupado y apilado



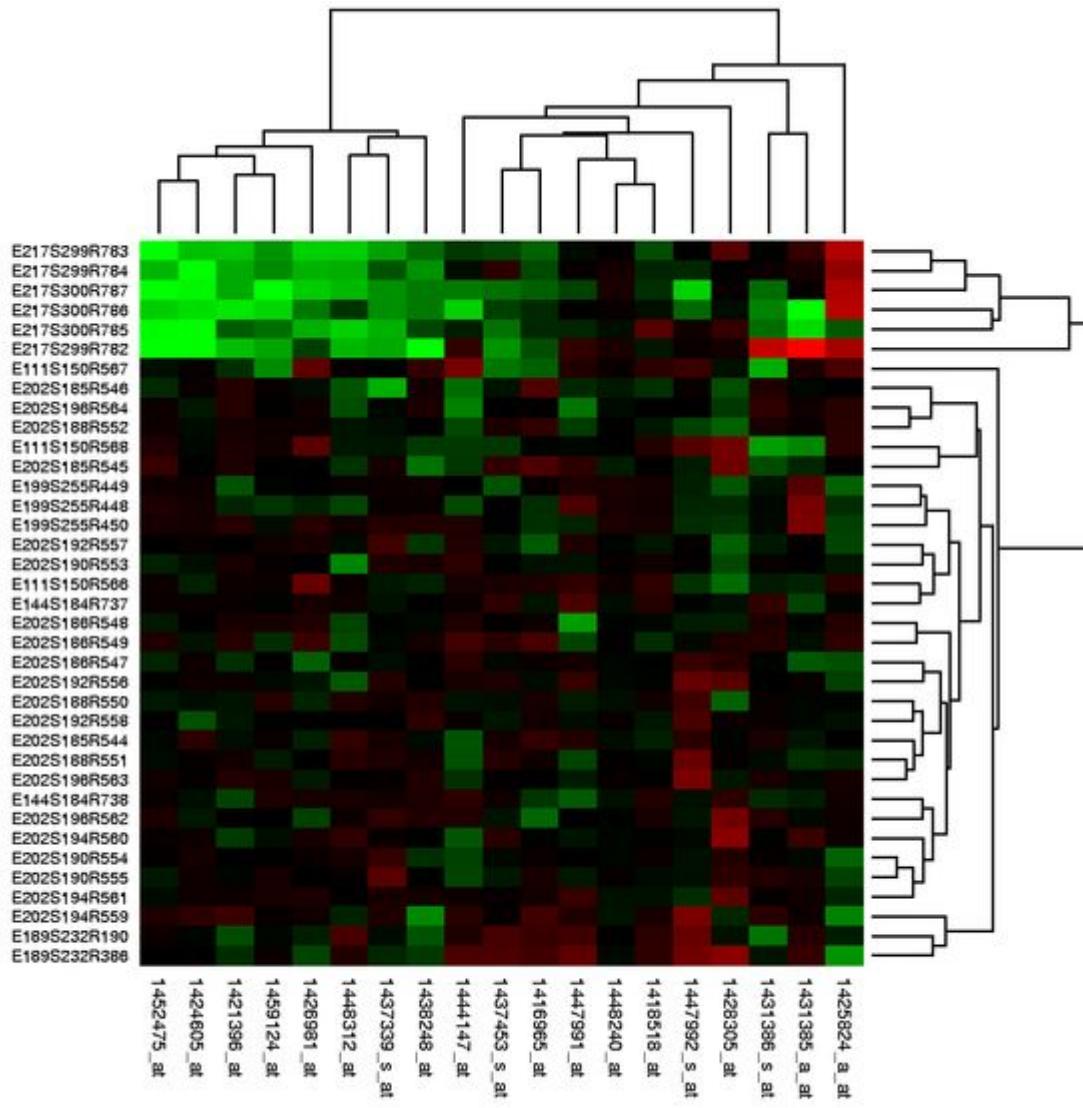
Bullet Chart



Gráficos en mosaico



Heatmap



Datos Tabulares

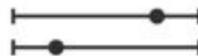
**Representaciones basadas en
explotar los canales gráficos**

Un canal por dimensión y varias marcas

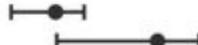
Channels: Expressiveness Types and Effectiveness Ranks

④ **Magnitude Channels: Ordered Attributes**

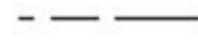
Position on common scale



Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



④ **Identity Channels: Categorical Attributes**

Spatial region



Color hue



Motion



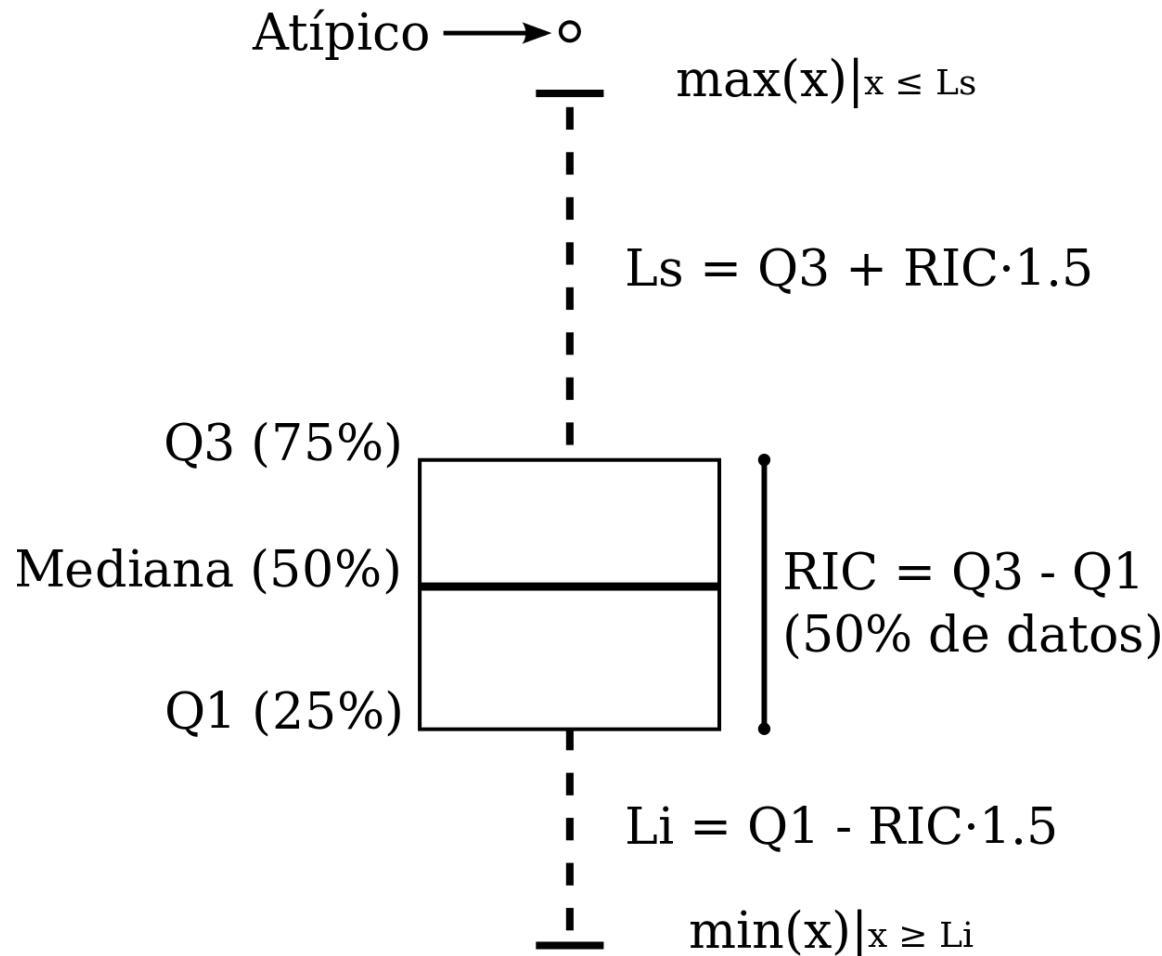
Shape



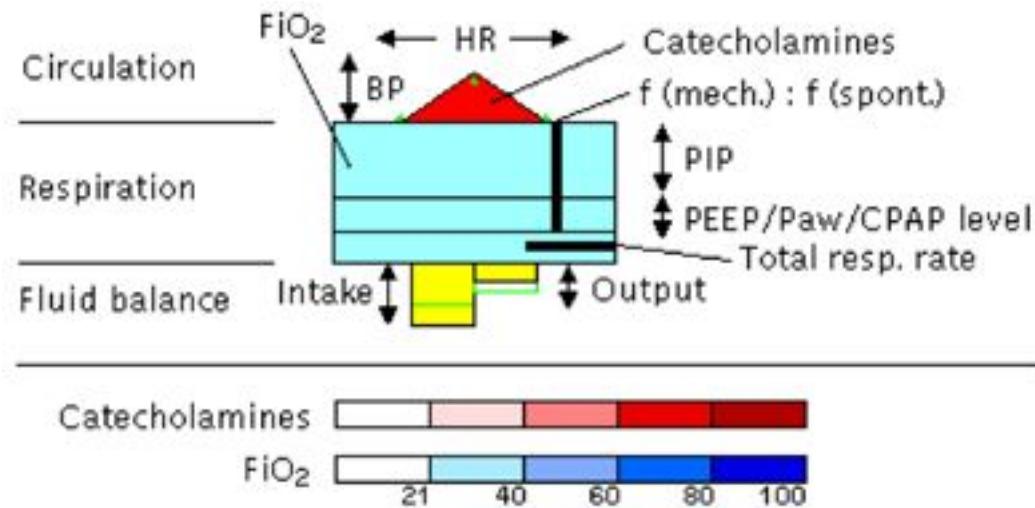
▲ Most
Effectiveness
Same
Least ▾

Tamara Munzner

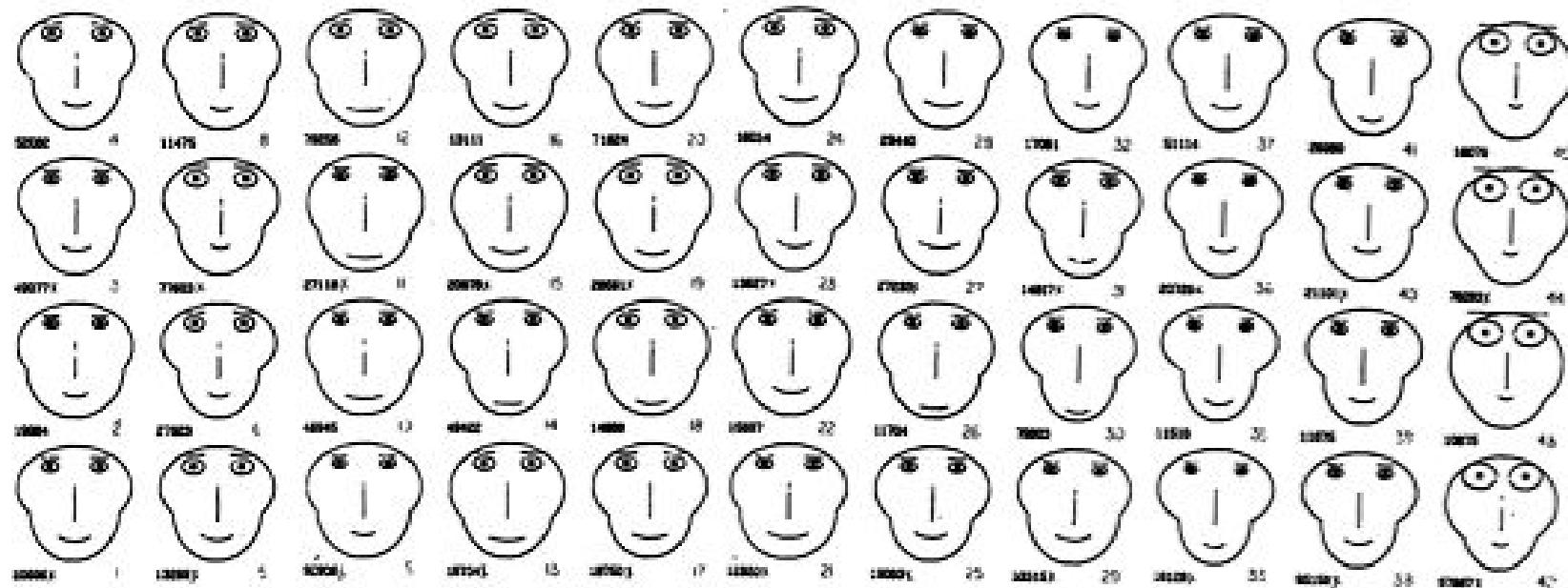
Box Plot



Glifos



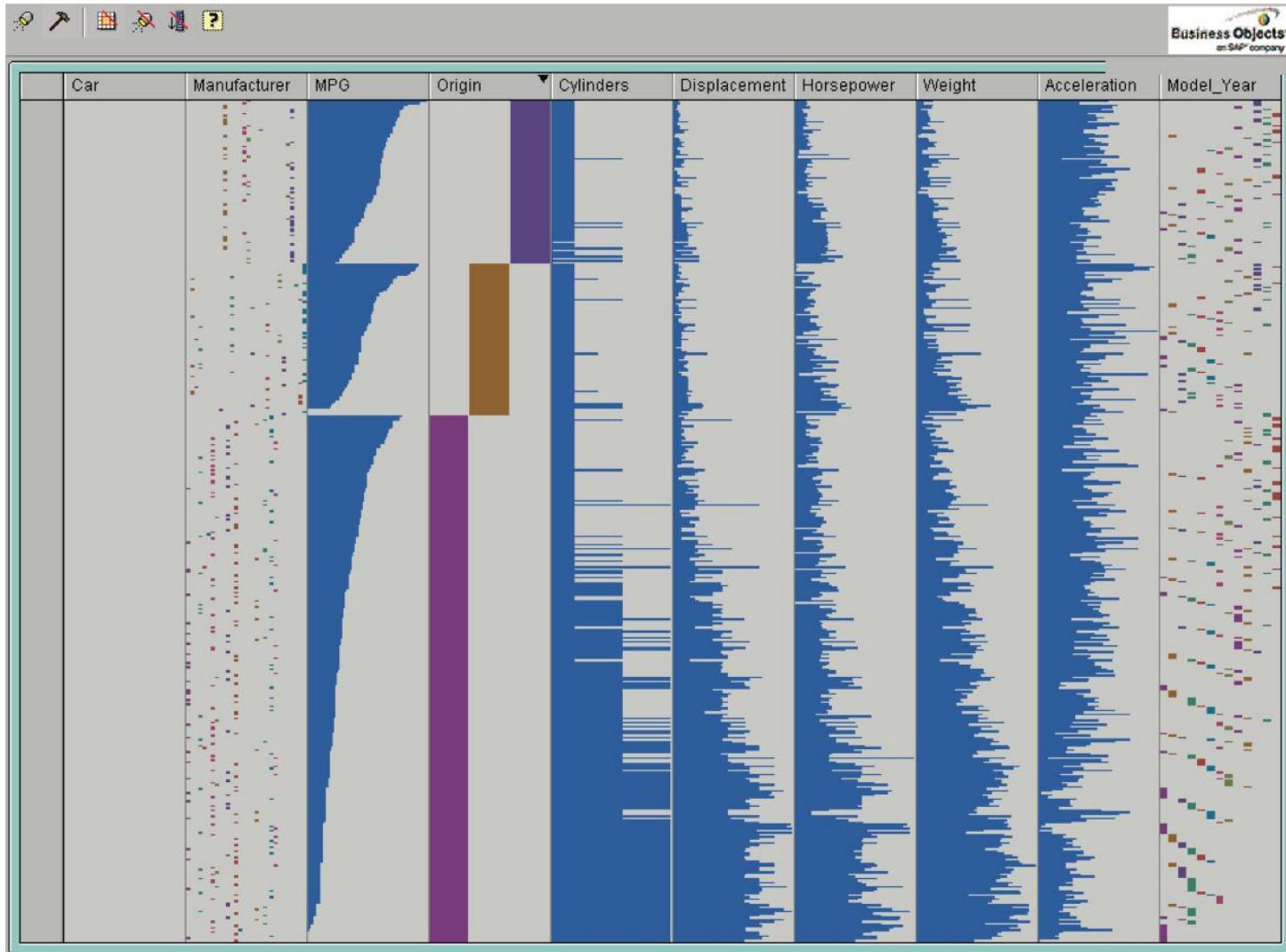
Caras de Chernoff



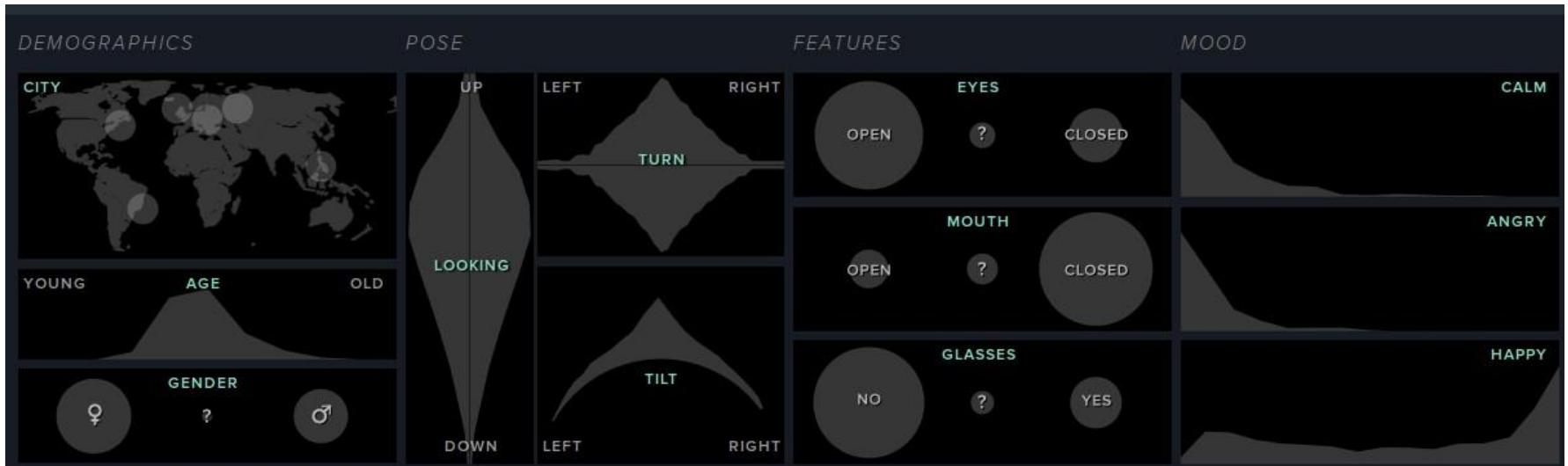
Datos Tabulares

Estrategias Interactivas

Ordenado según campos



Vistas enlazadas



3840 of 3840 selfies.



Índice

1. Introducción
2. Fundamentos
3. Casos
4. Visualización en Big Data
5. Datos Tabulares
- 6. Datos Temporales**
7. Datos Espaciales
8. Redes y Jerarquías

Dataset temporales

- Un conjunto de datos es temporal si **al menos una de sus variables es temporal**
- El tiempo no es más que una variable cuantitativa pero:
 - Nos da la dirección de la **causalidad** si esta existe
 - Es importante en la mayoría de escenarios
 - **Múltiples escalas** que se usan indistintamente (segundos, días, semanas, quincenas, meses, ...)
- Se suele agregar: (Variación entre semanas, trimestres, años, ...)
- Suele derivarse: (Medias móviles, Filtrado paso alto/bajo, T. Fourier, ...)

Según el muestreo

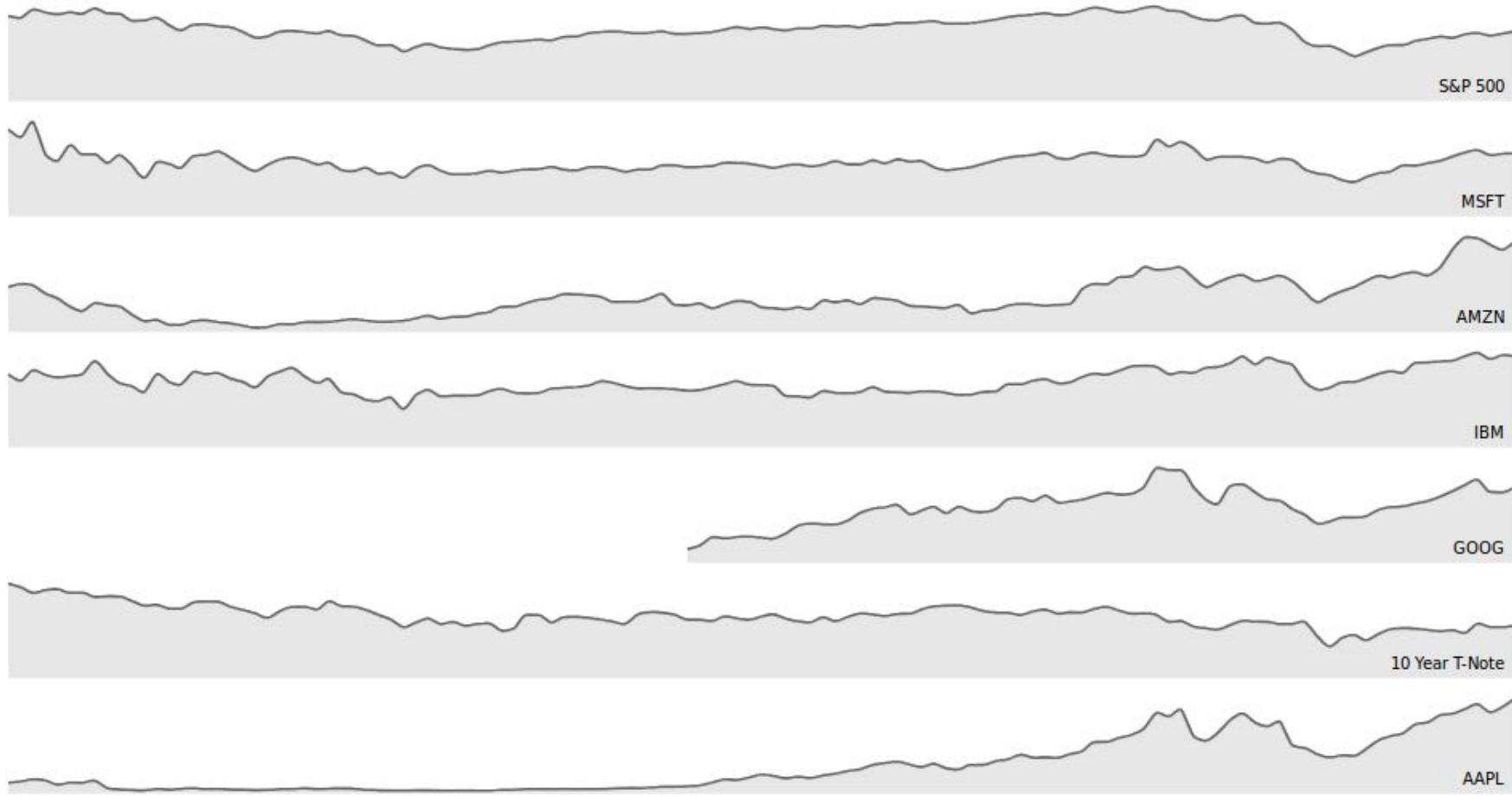
- **Series temporales:**
 - Datos registrados con una cadencia fija (stocks, sensores, ...)
- **Eventos temporales:**
 - Hay cierta relación temporal entre los items (intervalos, eventos puntuales)

Datos Temporales

Series temporales

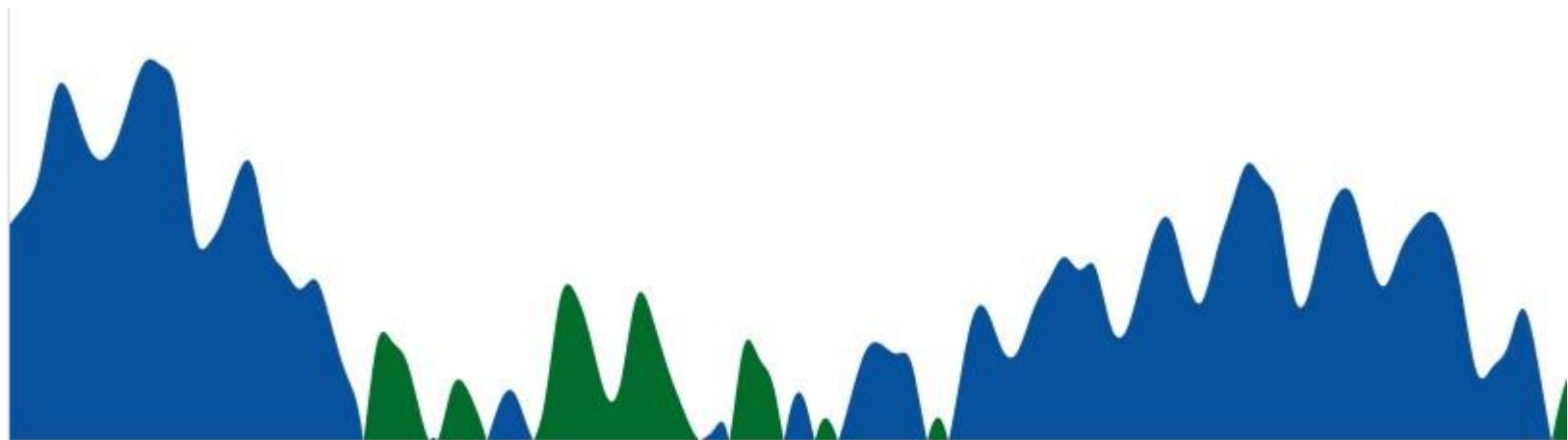
Line Chart

- La línea da sensación de continuidad, mejor que barras y puntos.
- El eje X suele ser el tiempo



Horizon Chart

- Consigue mayor resolución en el eje Y

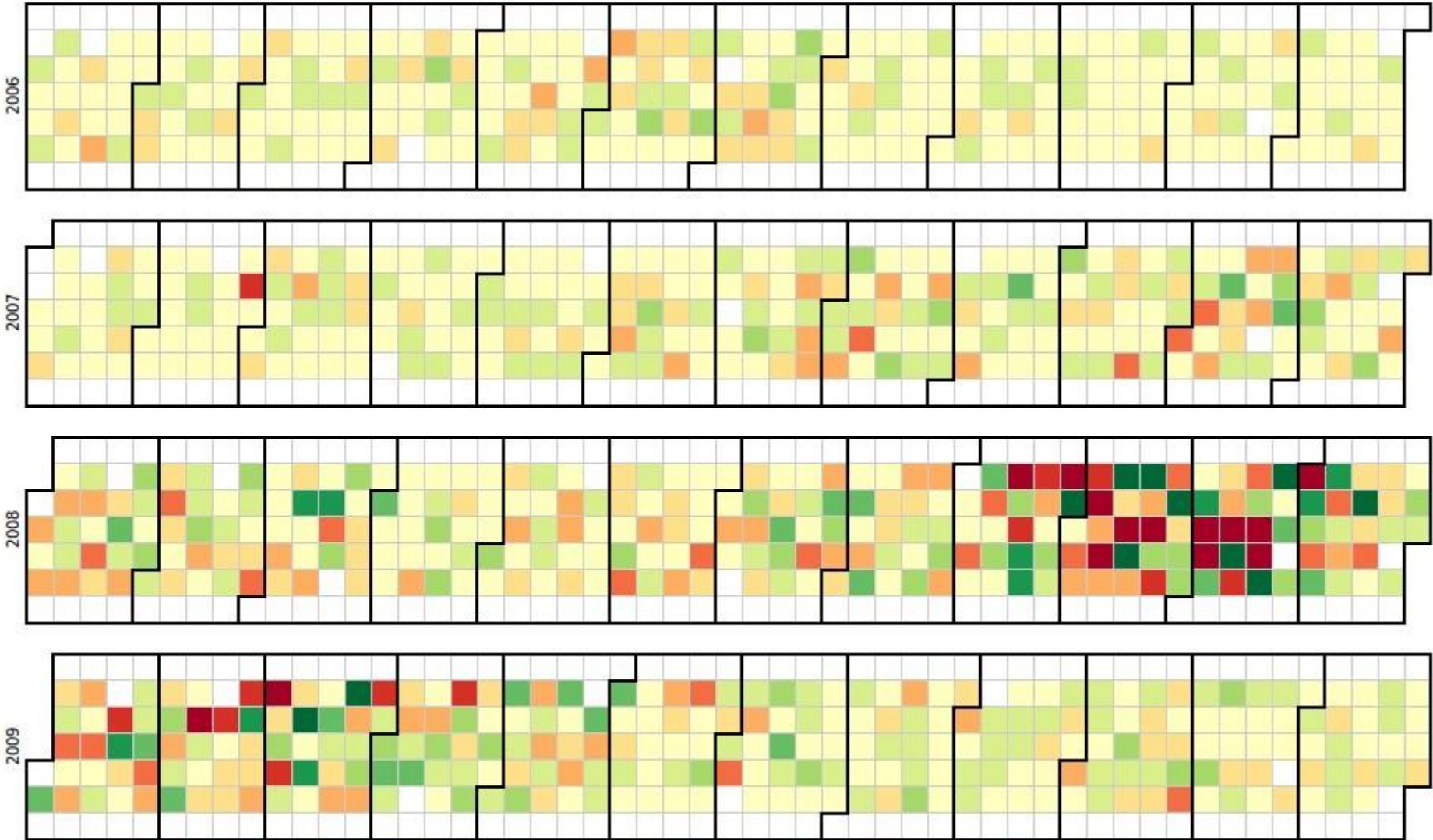


Candlesticks

- Agregaciones temporales como forma de ver patrones

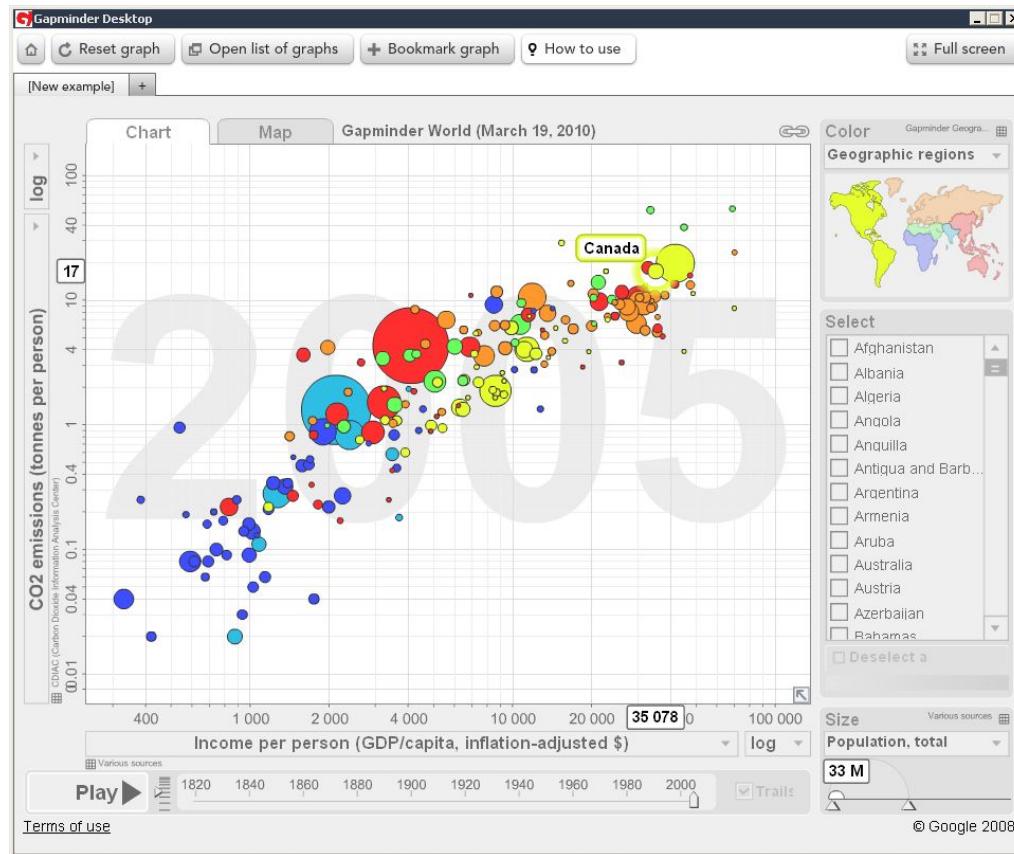


Calendar View



Gapminder

- Tiempo mapeado a animación

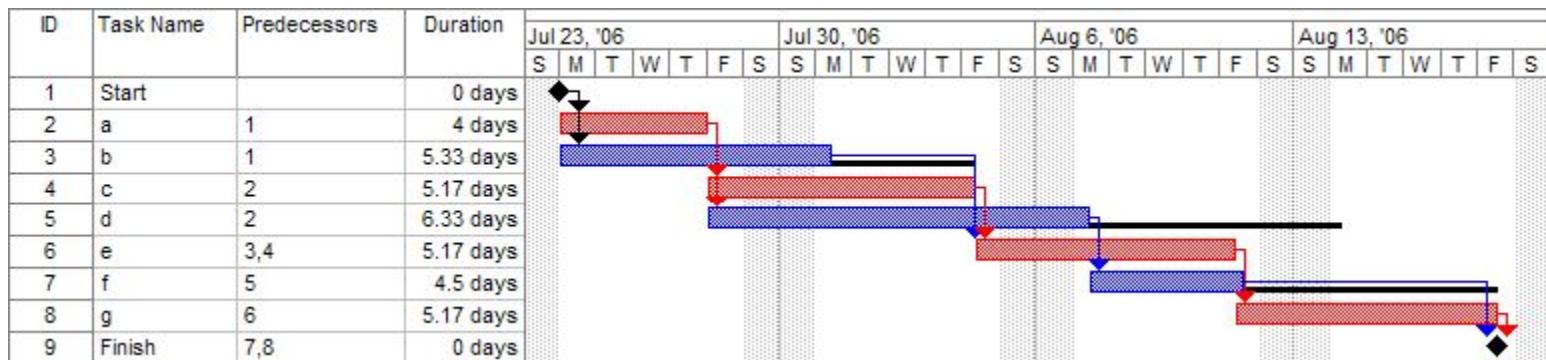


Datos Temporales

Eventos temporales

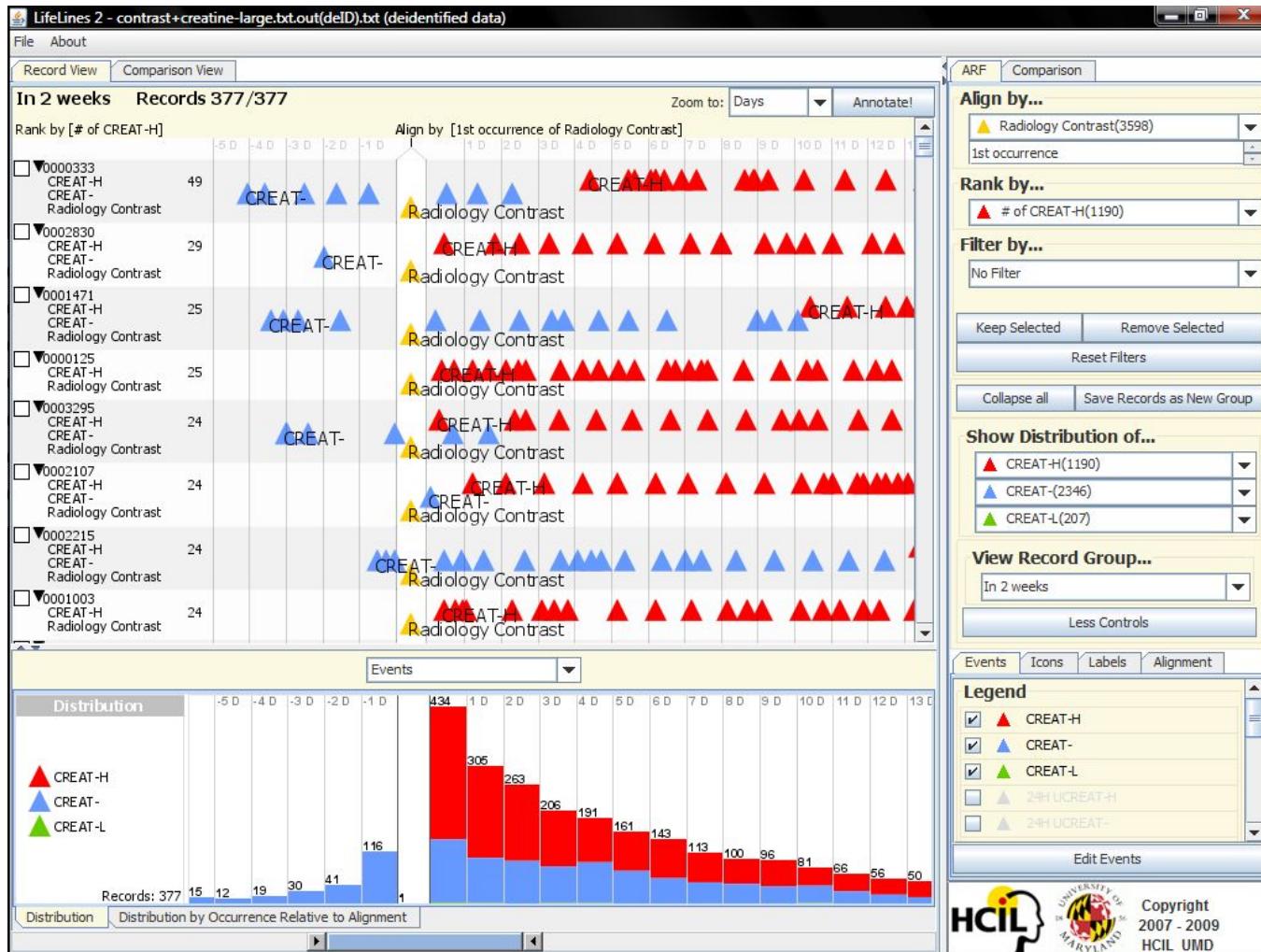
Gantt Chart

- Para representar intervalos y eventos puntuales



LifeLines

- Alineado relativo para "causa → consecuencia"



Índice

1. Introducción
2. Fundamentos
3. Casos
4. Visualización en Big Data
5. Datos Tabulares
6. Datos Temporales
- 7. Datos Espaciales**
8. Redes y Jerarquías

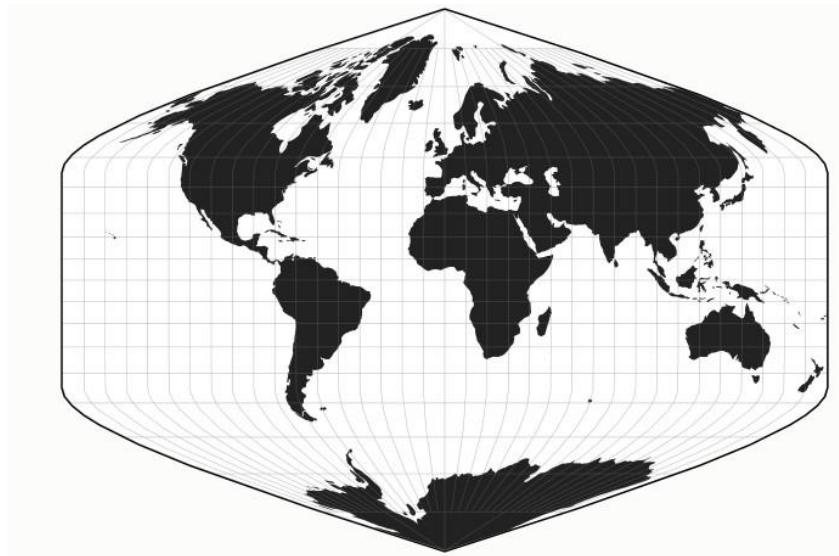
Datasets espaciales

Un conjunto de datos es espacial si la **posición viene dada por los datos**

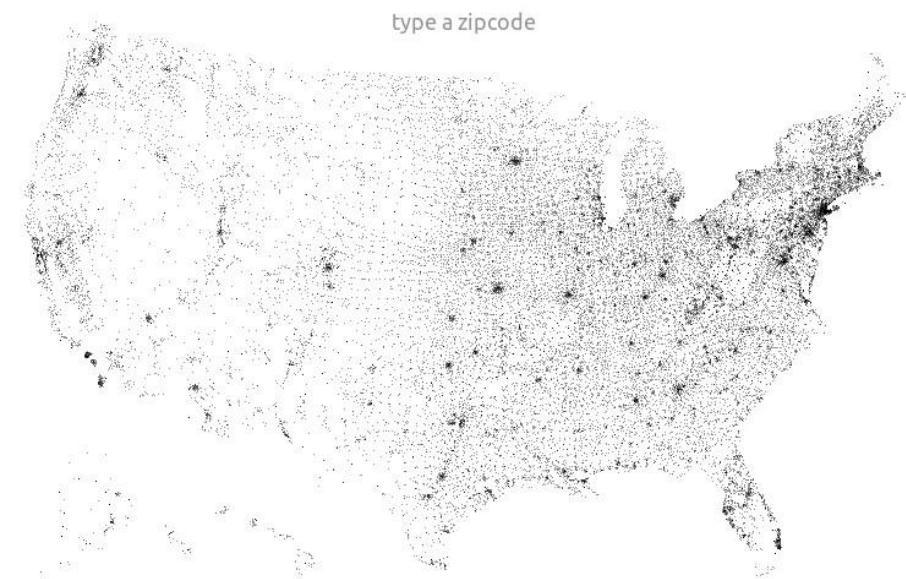
- **1D, 2D ó 3D:**
 - Ligado a Visualización científica
- **GeoEspeciales**
 - Mapas.
 - Cartografía: Otra corriente separada de InfoVis

Mapas

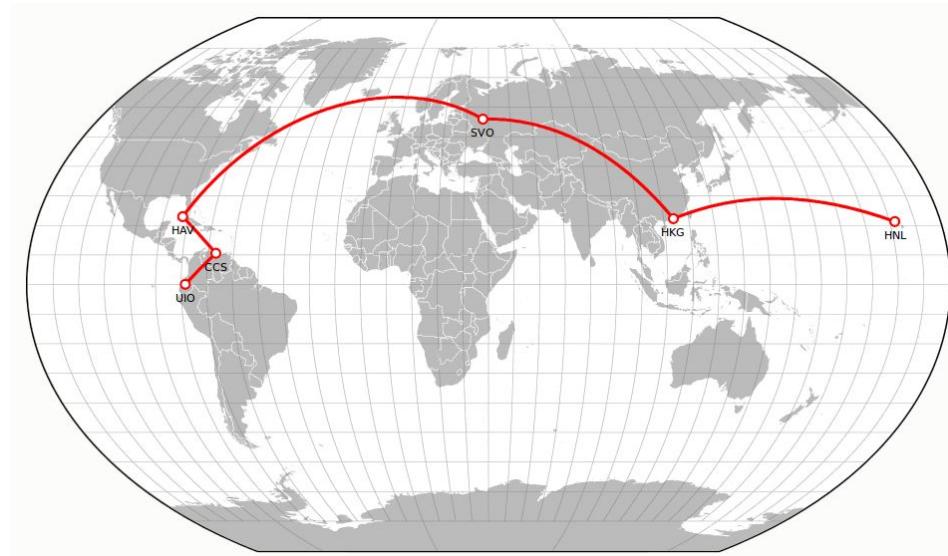
- La visualización fundamental para datos Geoespaciales es el mapa
- Representar en mapas da mucho contexto al resto de datos
- Warning: Se corre el riesgo de repetir una y otra vez el mapa de **densidad de población**.
- Hay mucho trabajo hecho sobre proyecciones:



Representar items sobre mapas

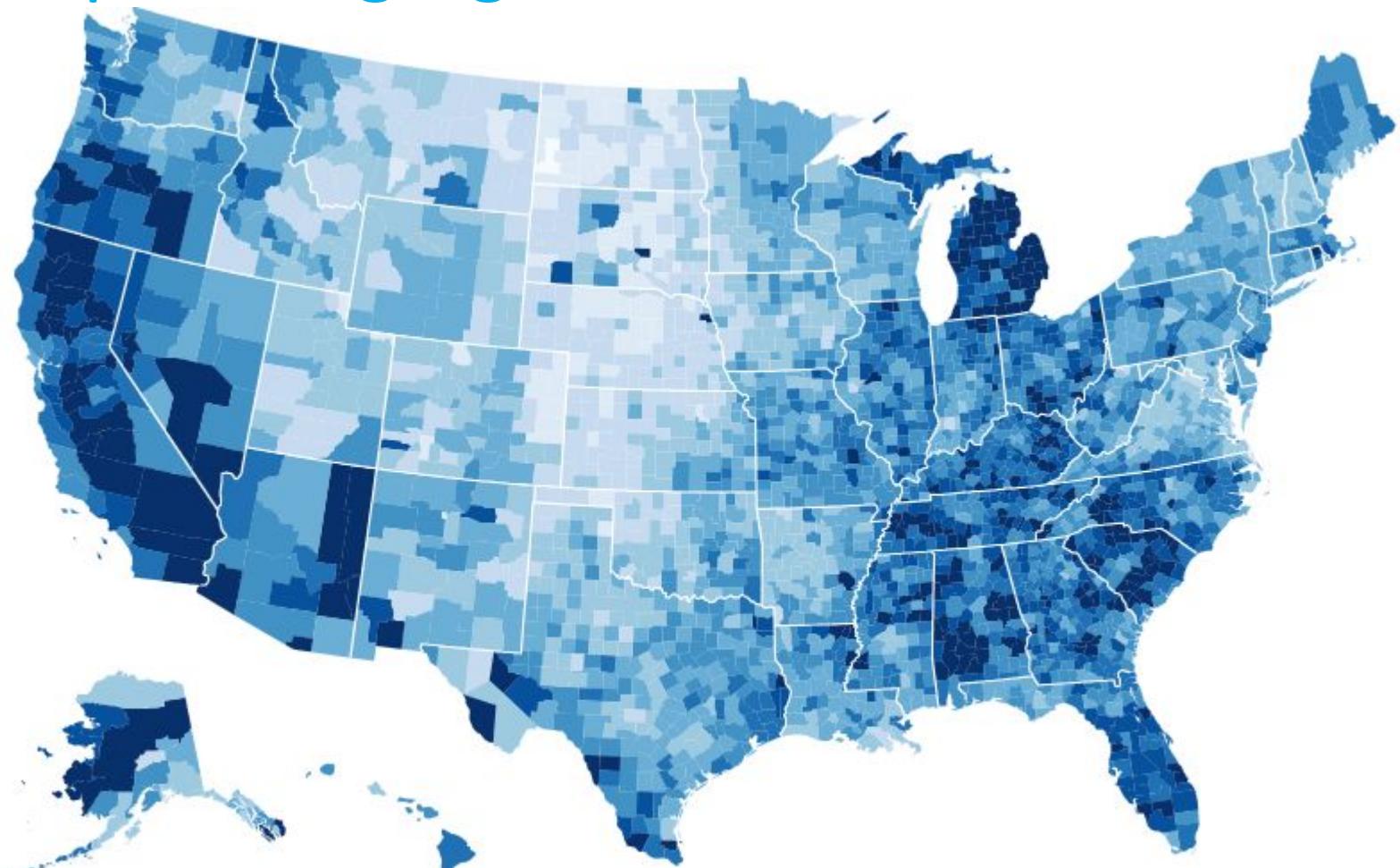


Puntos

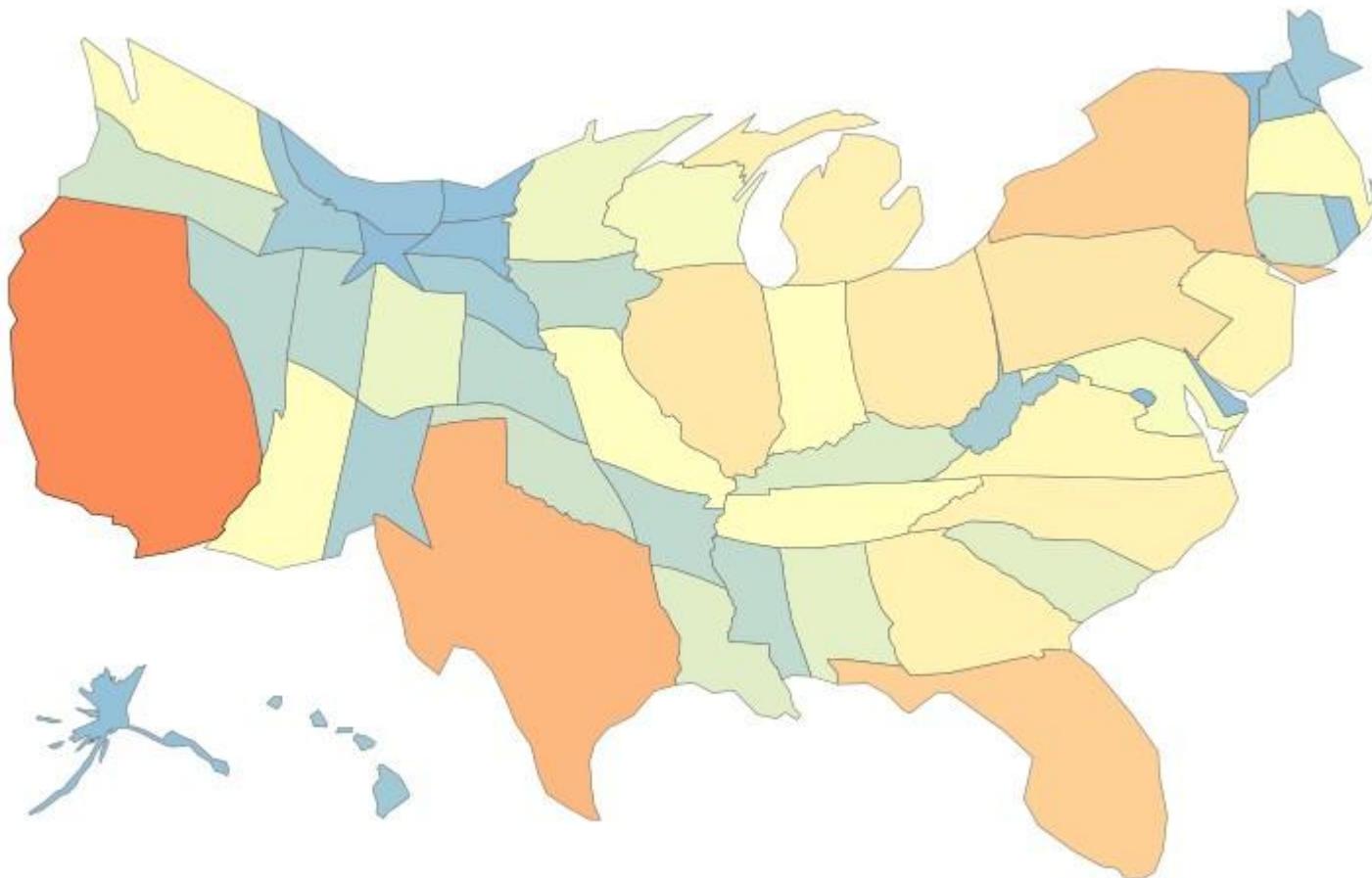


Paths

Choropleth: Agregaciones naturales



Cartograms



GridMaps



- El área de la región no afecta a la percepción del color del atributo que se quiere mapear
- Es más difícil de reconocer. Limitado a mapas con los que estemos "muy familiarizados"

Índice

1. Introducción
2. Fundamentos
3. Casos
4. Visualización en Big Data
5. Datos Tabulares
6. Datos Temporales
7. Datos Espaciales
- 8. Redes y Jerarquías**

Redes y Jerarquías

- Una **red** se modela como un **grafo**
- Una **jerarquía** se modela como un **árbol**, que a su vez es un tipo de grafo (conexo, sin pesos y acíclico).
- Este tipo de dataset no **se centra** en la representación de los items y sus atributos sino **en las relaciones** que se establecen entre los items.

Tipos de relaciones

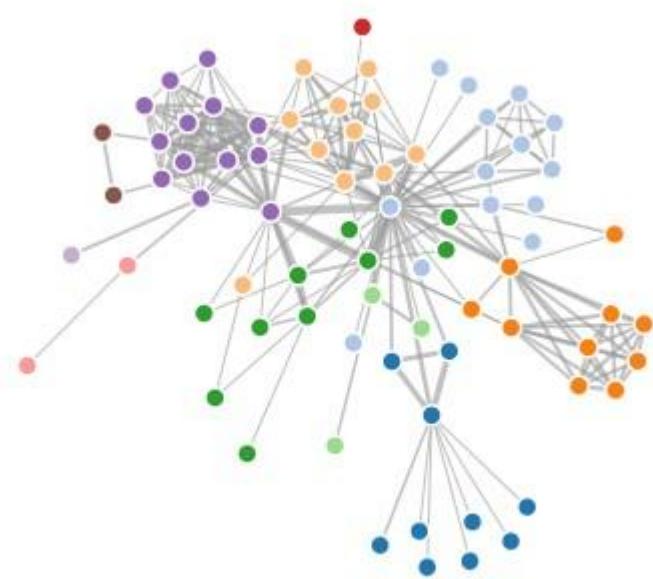
- Según su semántica:
 - parte/sub-parte, padre/hijo, es-un
 - conectividad (como entre ciudades unidas por carreteras)
 - derivado-de (como en flujo de datos)
 - pertenencia a grupo
 - similitud entre items o entre atributos
- Según su complejidad:
 - unidireccional / bidireccional
 - con pesos / sin pesos
 - incierto / cierto

Representación: Arco - Nodo

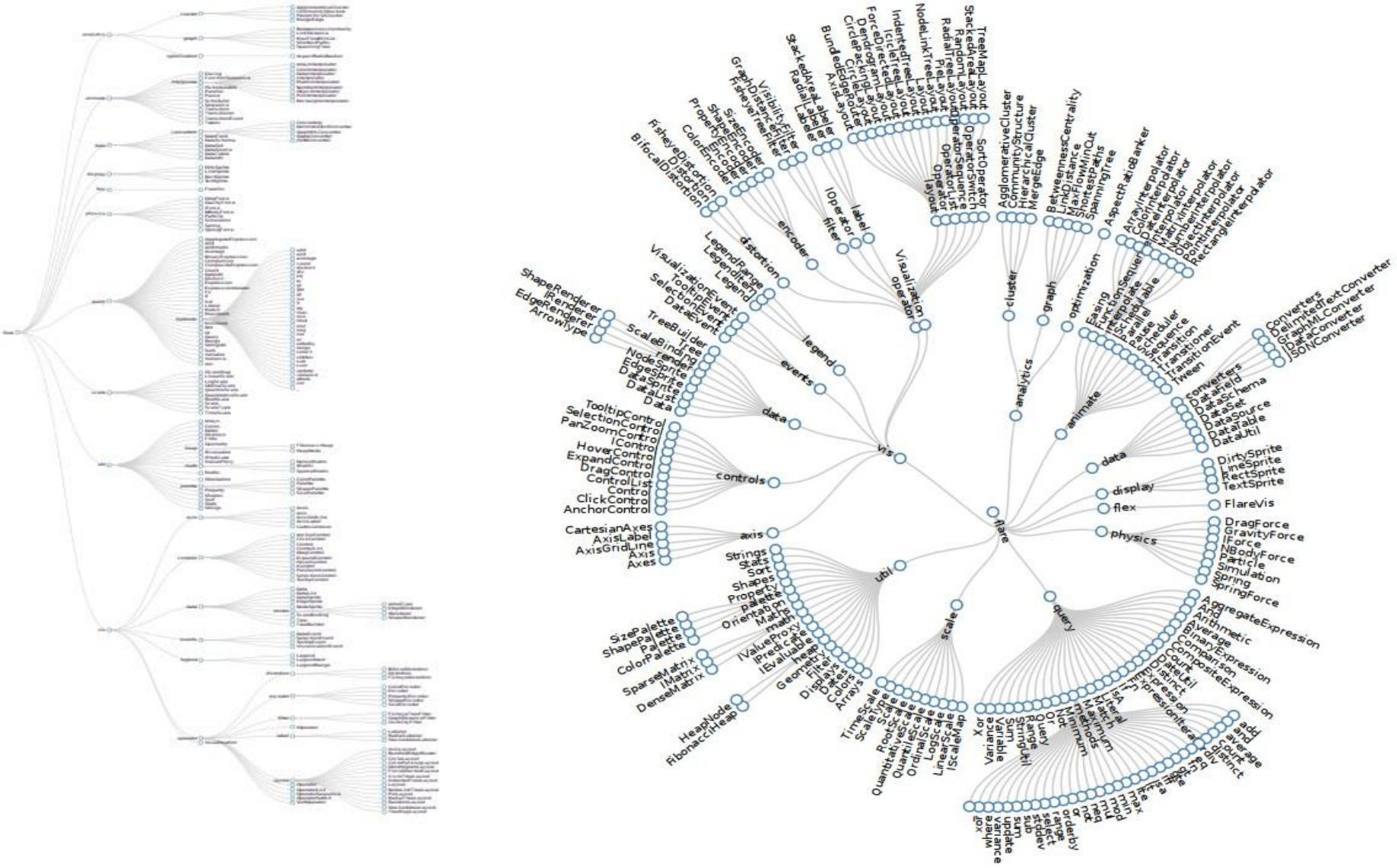
Muchos tipos de Layouts

- Se busca:
 - Minimizar arcos que se cruzan
 - Minimizar el área total de dibujo
 - Mantener un buen aspect ratio
- Se puede buscar:
 - Centrar algún nodo concreto
 - Posición fija de los nodos (ej: sobre un mapa)
 - Posicionar hermanos en la misma línea

Grafos: Arco - Nodo



Árboles: Arco - Nodo

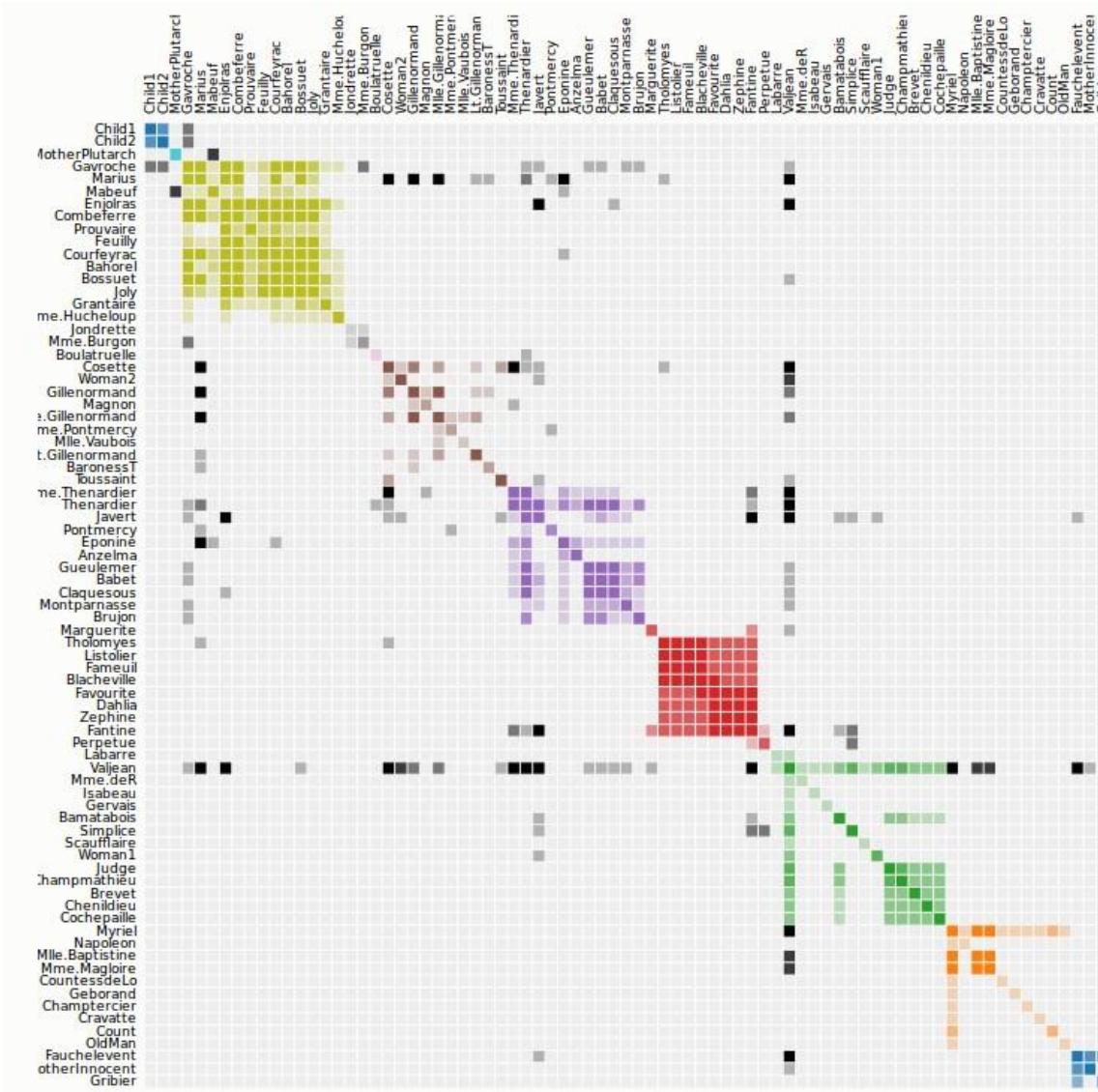


Representación: Space Filling

- Mayor escalabilidad: Puede representar muchos más arcos y nodos
- Con respecto a Arco-Nodo
 - Mejora la densidad de información
 - Suele ser menos intuitivo

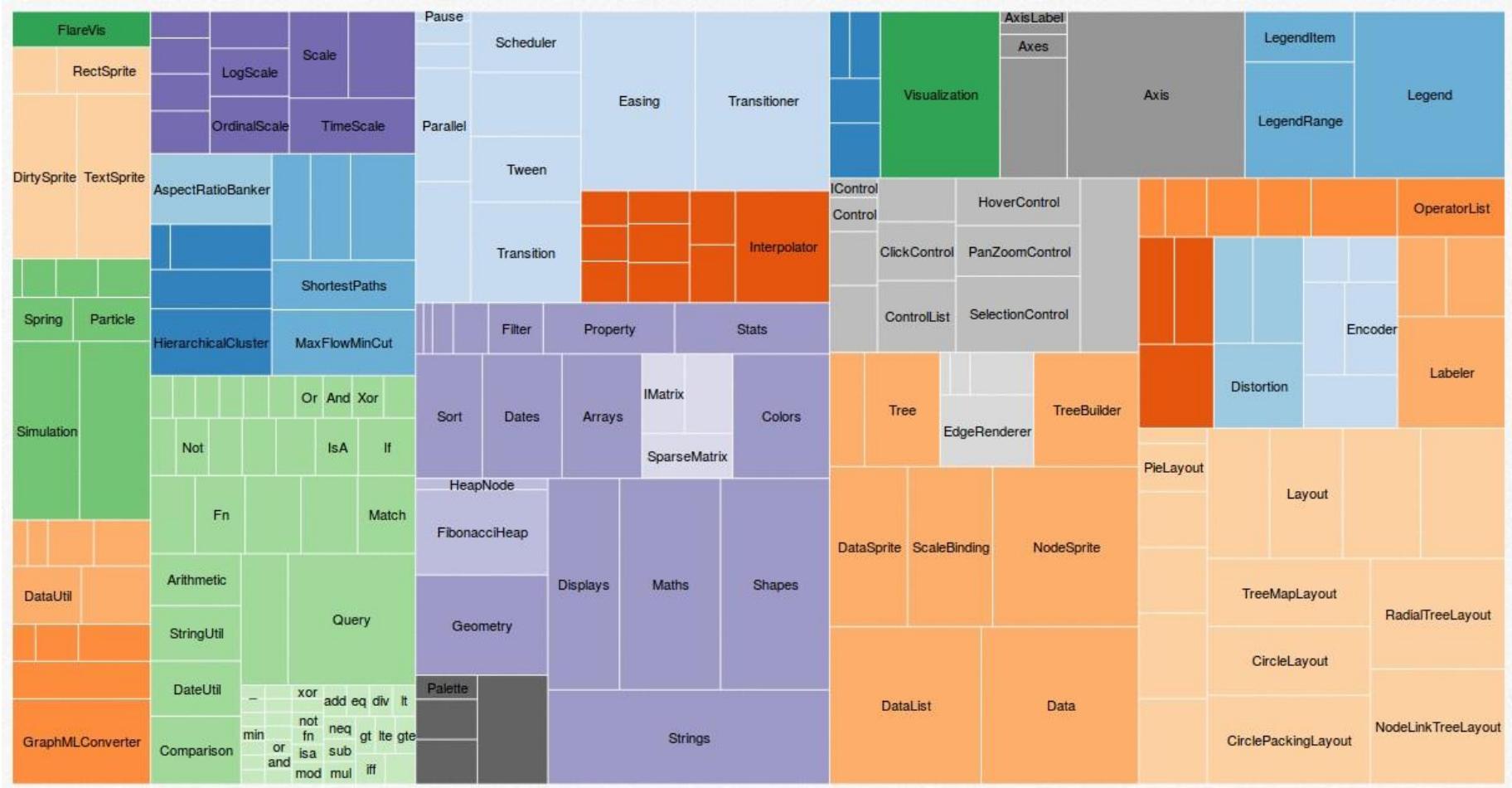
Grafos: Space Filling

Matriz de convectividad



Árboles: Space Filling

Treemap



Árboles: Space Filling

Sunburst

