

1. Ejemplos históricos:

a. La visualización de John Snow:

1854

Epidemia de cólera en Londres.

En esos años no se conocía el origen de la enfermedad ni los mecanismos de transmisión.

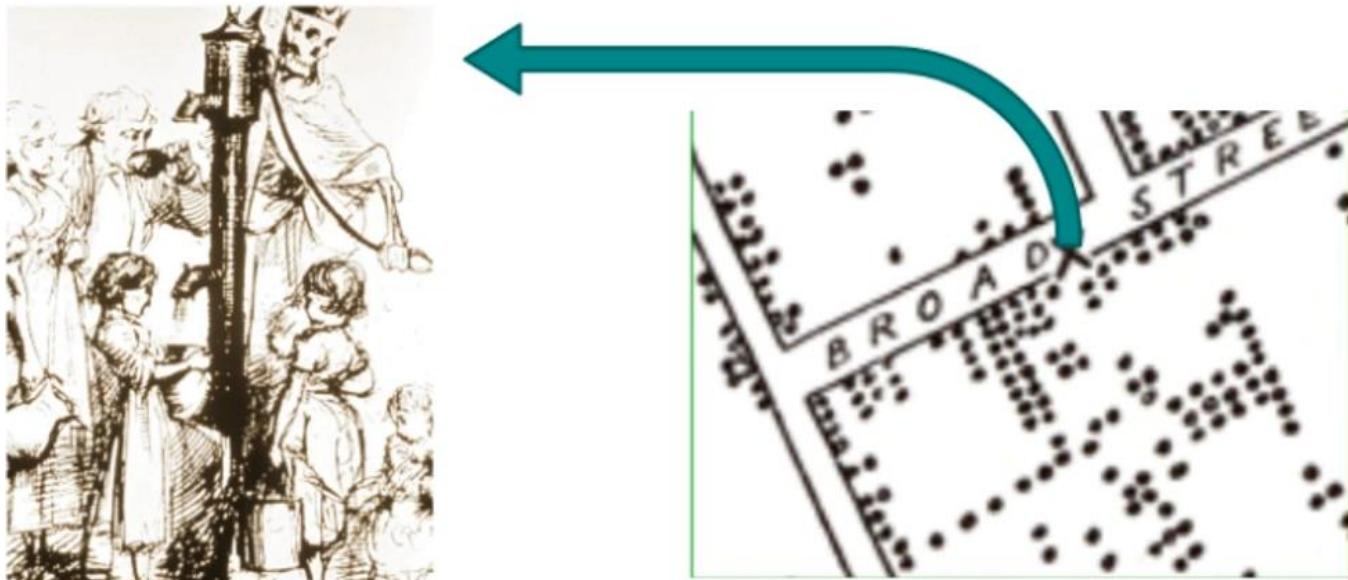
Doctor John Snow

Analiza los datos a través de una visualización:
presenta los casos como un histograma espacial.



Concluye que el problema puede estar en los bebedores de agua públicos que contagia la enfermedad.

Ordena cerrar esos pozos de agua y tiene un efecto rápido en disminuir el número de infectados.



b. **La visualización de Florence Nightingale:**

Inglaterra entra en la guerra de Crimea.

Envío de enfermeras

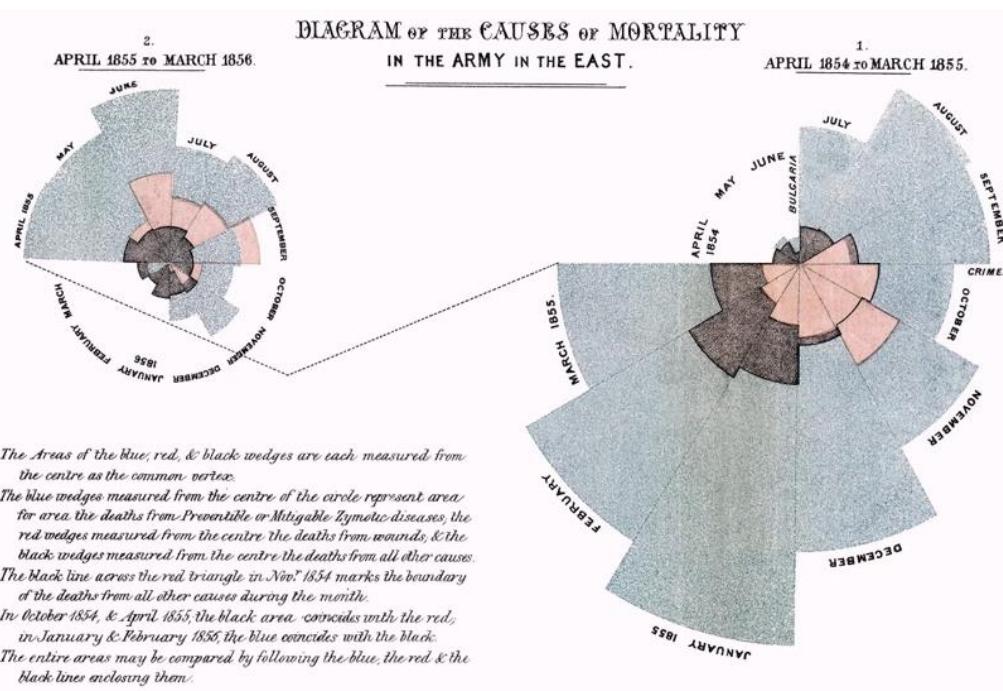
Para apoyar con las malas condiciones de los heridos, entre ellas a Florence Nightingale.

Florence Nightingale

Analiza visualmente causas de muerte usando un gráfico circular, al que algunos llaman ahora diagrama de la rosa de Nightingale.

Fallecimiento de soldados

Fallecieron diez veces más soldados de enfermedades como tifus, fiebre tifoidea, cólera y disentería que de heridas en el campo de batalla.



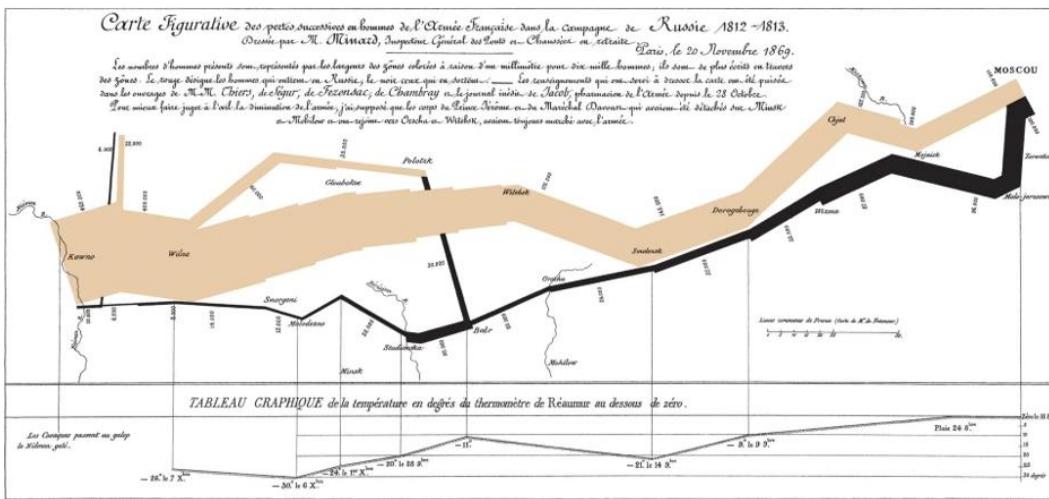
c. La visualización de Jacques Minard:

Charles Jacques Minard Ingeniero - (1781-1870)

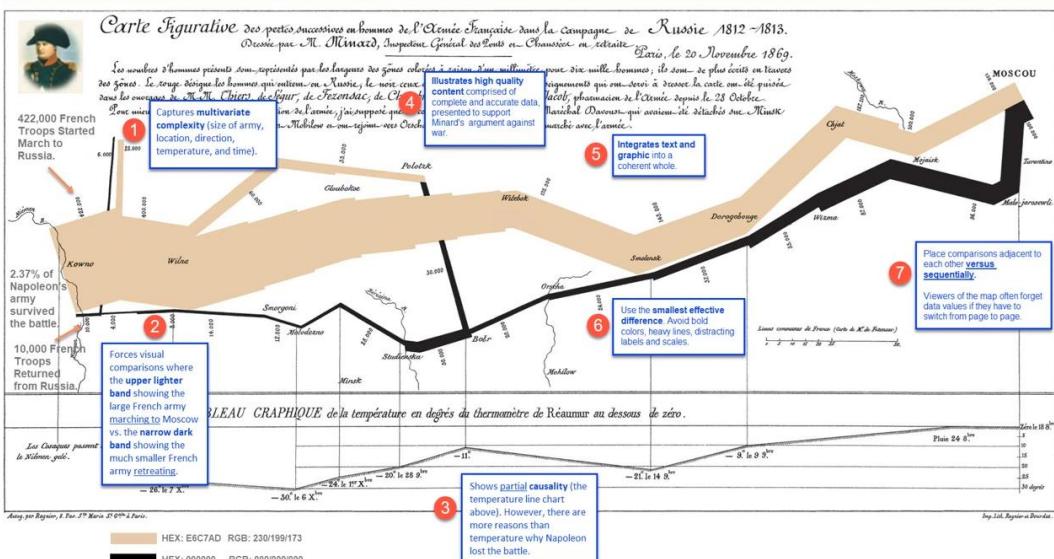
Presenta en 1869 una visualización que ilustra la campaña de Napoleón en Rusia (1812-1813).

Opinión de autores

Edward Tufte en su libro *The Visual Display of Quantitative Information*, "este podría ser el mejor gráfico estadístico dibujado".



Aspectos que captura esta visualización:



2. Conceptos y definiciones:

a. Definiciones de visualizar:

Definición de "visualizar" (RAE)

- **Visibilizar**
- **Representar mediante imágenes** óptica fenómenos de **otro carácter**.
- Formar en la **mente** una **imagen visual** de un concepto **abstracto**.

Card, Mackinlay & Shneiderman (1999)

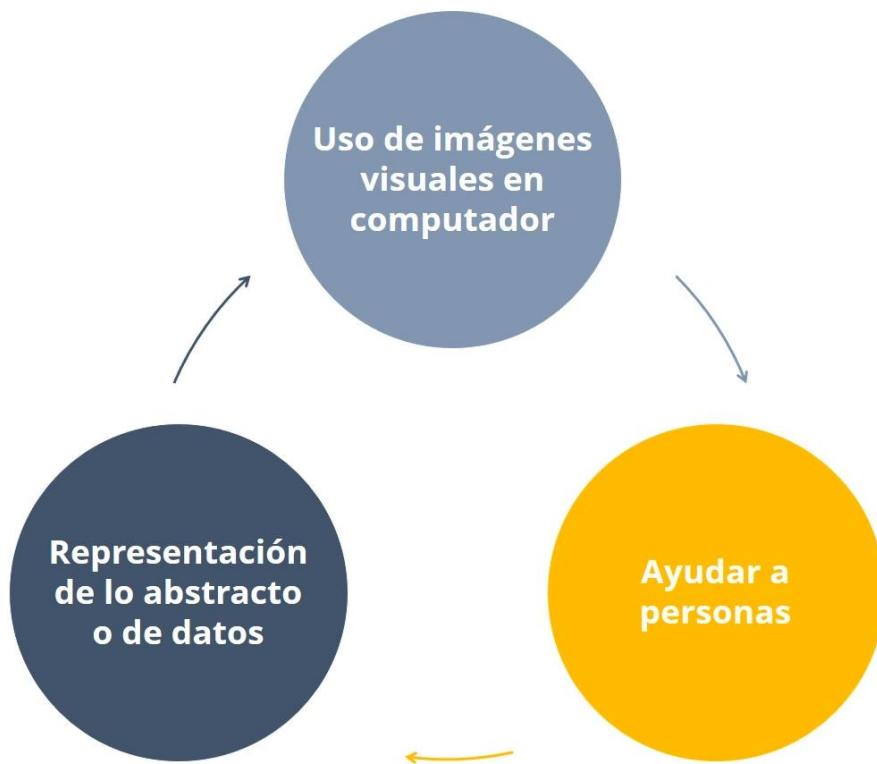
*"El uso de **representaciones visuales** de **datos**, generados por **computador**, interactivos, para amplificar nuestra **cognición**."*

Munzner (2014)

*"Sistemas de visualización por **computador** proveen **representaciones** de **datos** diseñados para ayudar a **personas** realizar **tareas** de forma más eficiente."*

b. Tres ideas de los objetivos del área:

- i. Con las 3 definiciones anteriores nos damos cuenta de que hay pilares que apunta a la definición de visualización.



c. Marcas y canales:

- i. **Marca:** elemento geométrico básico, que puede ser clasificado según el número de dimensiones especiales que requiera. Ejemplos de marca son:

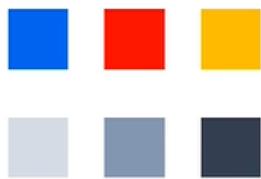
1. Puntos.
2. Líneas.
3. Áreas.

Las marcas son bloques básicos para crear visualizaciones y, por lo general, representan una identidad de datos.

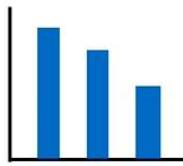
- ii. **Canales:** es un aspecto visual que permite alterar la apariencia de una marca. Ejemplos comunes son:

1. Posición vertical u horizontal.
2. Largo y tamaño.
3. Forma.

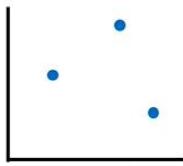
La variación de un canal se asocia a la representación visual de un valor de un dato específico.



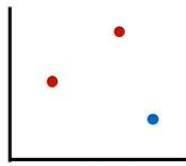
¿Canales?



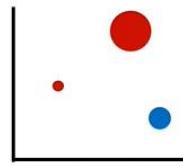
(a)



(b)



(c)



(d)

Posición horizontal y largo vertical

Posición horizontal y posición vertical

Posición horizontal, posición vertical y color

Posición horizontal, posición vertical, color y tamaño

Tipos de canales:

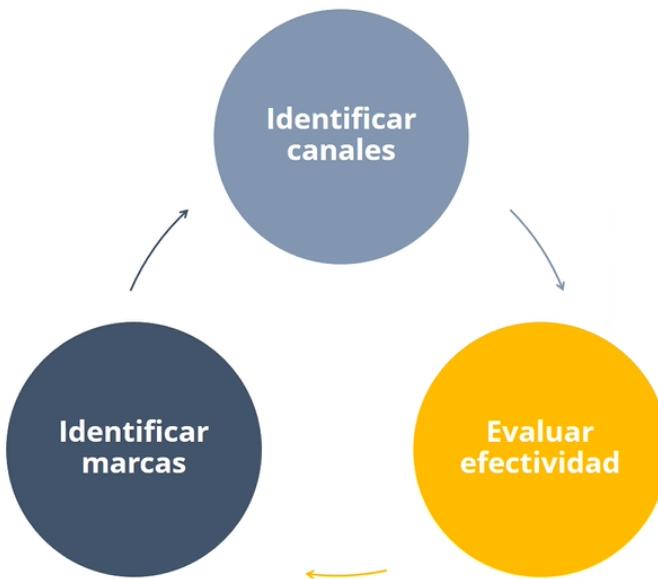
Canales de identidad: permite discernir información sobre **qué** es algo o **dónde** se encuentra.

- Posición espacial.
- Tonalidad de color.
- Forma.

Canales de magnitud: permite saber cuánto de algo existe.

- Posición.
- Largo o tamaño.
- Saturación de color.

Analizar los bloques elementales de una visualización mediante marcas y canales



iii. **Efectividad de canales:** La efectividad de un canal se puede analizar en base a los siguientes criterios:

1. Precisión.
2. Discriminación.
3. Separabilidad.
4. Detectabilidad (pop-out).



iv. Ranking de canales por efectividad:

1. Canales de identidad:

- a.** Posición espacial.
- b.** Tonalidad de color.
- c.** Forma.

2. Canales de magnitud:

- a.** Posición en escala común.
- b.** Posición en escalas distintas.
- c.** Largo (una dimensión).
- d.** Inclinación (ángulo).
- e.** Tamaño (dos dimensiones).
- f.** Profundidad (posición 3D).
- g.** Iluminación de color.
- h.** Saturación de color.
- i.** Curvatura.
- j.** Volumen (tres dimensiones).

d. Herramientas de computación:

i. Herramientas de visualización generales: permite rápidas construcciones de visualizaciones simples, sin conocimiento técnico de programación.

- 1.** Tableau.
- 2.** Power BI.

ii. Librerías de visualización de alto nivel: permite mayor variedad de opciones, a cambio de mayor manejo técnico de programación.

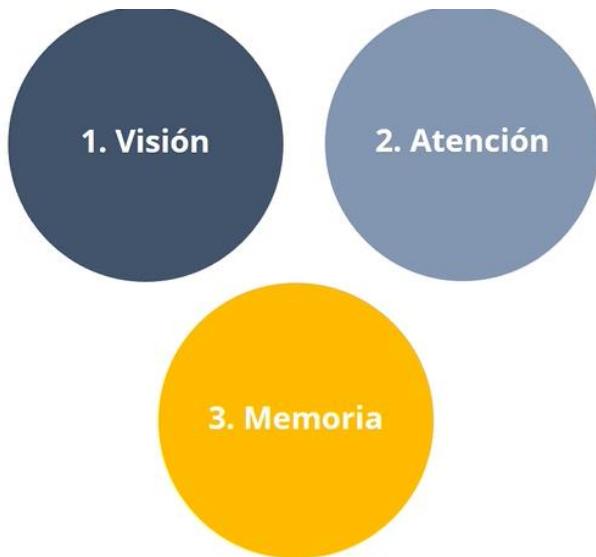
- 1.** Matplotlib (Python).
- 2.** Seaborn (Python).

iii. Librerías de visualización de bajo nivel: permite total libertad creativa de construcción, pero requiere conocimientos profundos de programación.

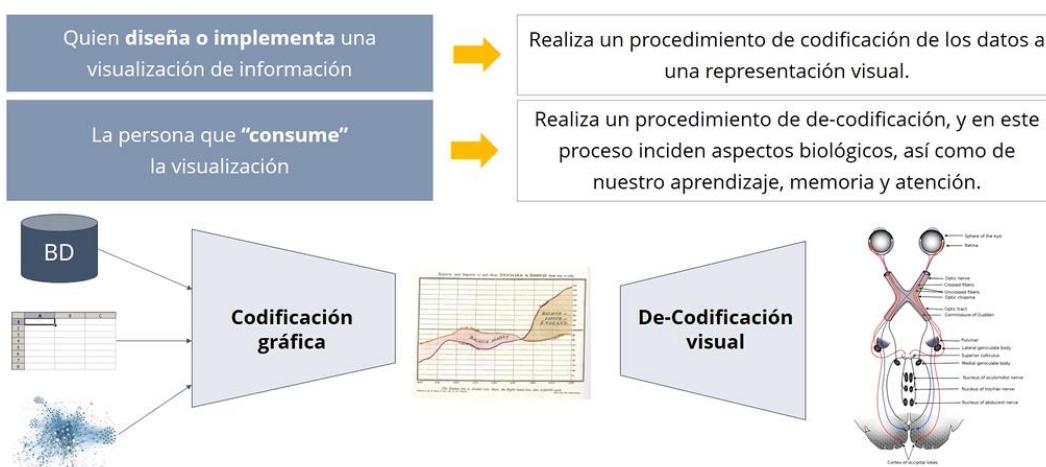
- 1.** D3 (JavaScript).
- 2.** Processing (Java).

3. Percepción visual

- a. Aspectos biológicos, de percepción son fundamentales para diseñar visualizaciones efectivas. Se revisarán:
 - i. Percepción visual.
 - ii. Atención.
 - iii. Memoria.



b. Codificación y decodificación en visualización de información:



c. ¿Qué entendemos por percepción?

Percepción

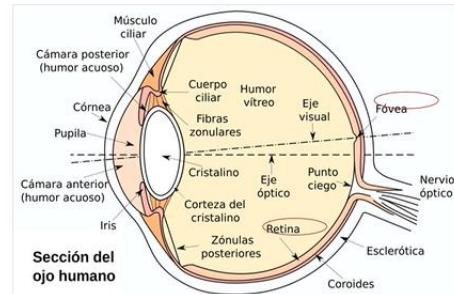


Organización, identificación e interpretación de la información sensorial.

Este complejo sistema nos permite representar y entender nuestro entorno.

Toda clase de percepción involucra señales que actúan en el sistema nervioso, a través de reacciones físico-químicas.

La **visión** ocurre por la luz que recibe la retina del ojo.



d. ¿Somos receptores pasivos de estímulos visuales?

Percepción visual

No es un proceso pasivo

No sólo recibimos estas señales desde el exterior para percibir, sino que también están fuertemente afectadas por el aprendizaje, la memoria y la atención.

Estudio de ilusiones e imágenes ambiguas

Ha demostrado que nuestros cerebros intentan (de forma subconsciente, incluso) darle sentido al input que recibimos.

Entender estos conceptos nos puede ayudar a diseñar visualizaciones más efectivas.

e. Procesos de precepción

Percepción

Dos procesos

Primer proceso

Transforma información de bajo nivel hacia alto nivel

Ejemplo

Extraer bordes y formas para luego reconocer objetos.

Segundo proceso

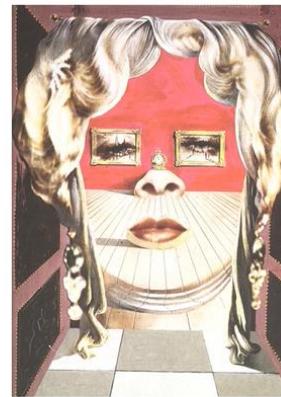
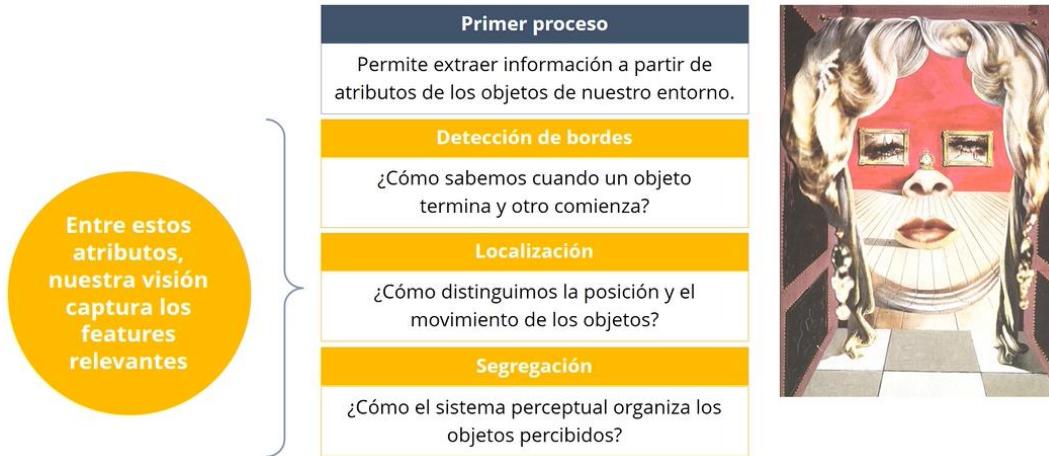
Conecta el conocimiento previo de cada persona junto con la atención.

Ejemplo

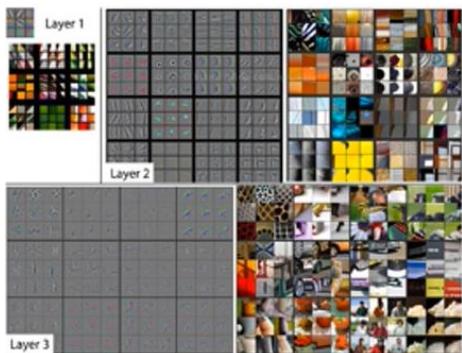
Identificar atributos de objetos en relación a su contexto.

El proceso cognitivo de selectivamente concentrarse en un algún aspecto de la información recibida, mientras se ignora otro tipo de información perceptible.

f. Percepción visual de bajo nivel



g. Relación con redes neuronales artificiales



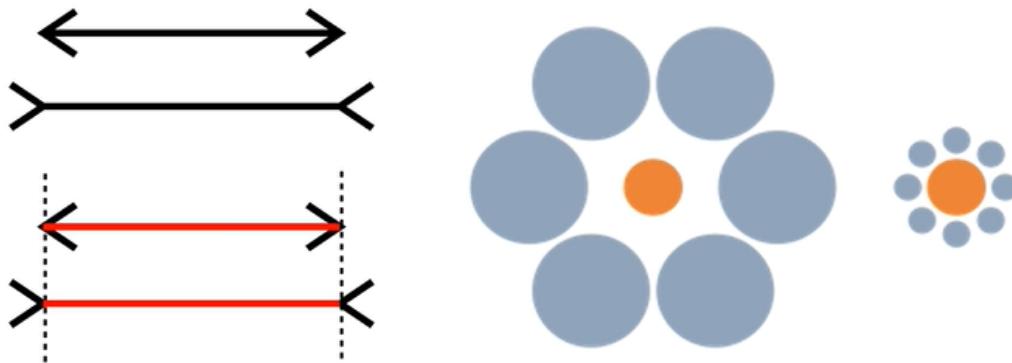
Se especula que las redes neuronales profundas aprenden de forma similar

- Aprenden características de bajo nivel (líneas, curvas, etc.).
- Se van componiendo de forma jerárquica para entender una imagen.

h. Percepción visual de alto nivel

Constancia del tamaño

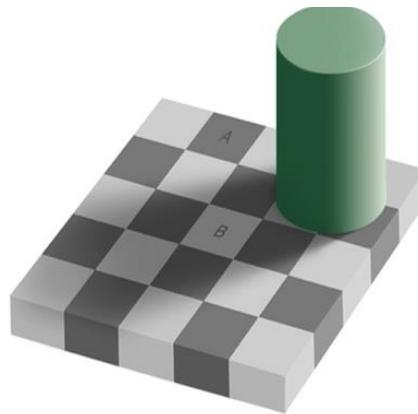
- Cuando un objeto se aleja o cambia su contexto, lo percibimos del mismo tamaño.
- Pero esta propiedad falla, según algunos contextos.



Constancia del color

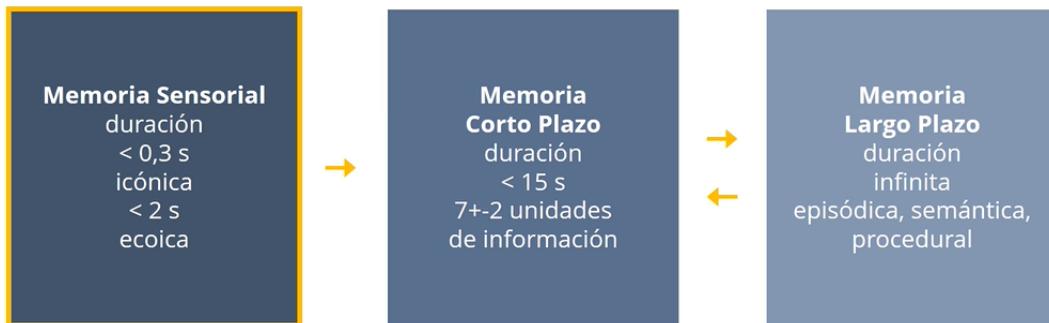
- El color percibido de los objetos permanece relativamente constante bajo condiciones de iluminación variables.

Notar que A y B son exactamente el mismo color.



i. Atención y procesamiento pre-atentivo

- Pre-atentiva** quiere decir, sin poner atención de forma consciente, sin usar memoria de corto o largo plazo, es decir, inconsciente.
- Estructura y procesamiento de memoria:**



Procesamiento inconsciente, previo a poner atención (**pre-atentiva**)



Memoria de corto alcance

- Habilidad para recordar y procesar información.
- Uso consciente.

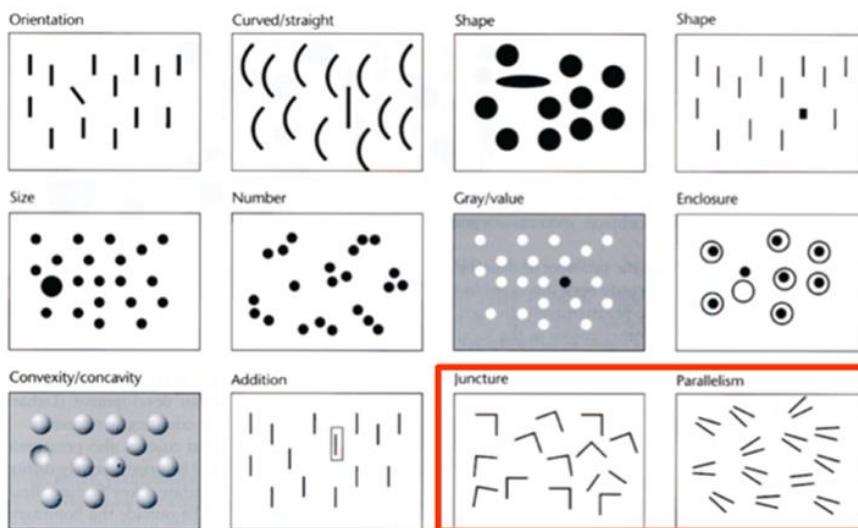


iii. Atención y características pre-atentivas en visualización

Características que podemos procesar de forma inconsciente		
Pre-atentivas		
Colores	Contraste	Curvatura, etc.
Extremos de línea	Inclinación	

"An understanding of what is processed pre-attentively is probably the most important contribution that vision science can make to data visualization."
(Colin Ware)

Ejemplos de características pre-atentivas



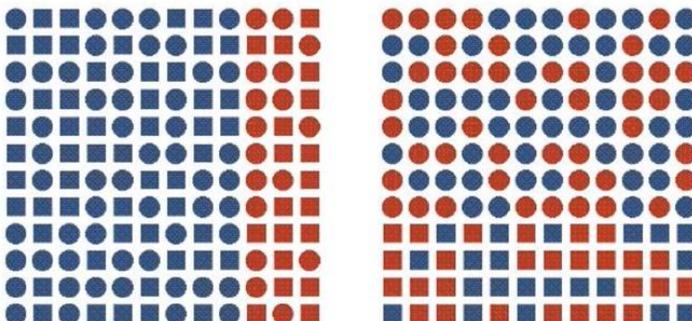
iv. Si en una visualización queremos se muestre un patrón que nos interesa que se vea se va a usar el canal de color.

v. Interferencias en características pre-atentivas

Detectar objeto considerando distractores



Trazar una línea (horizontal o vertical) que separa dos claros grupos.

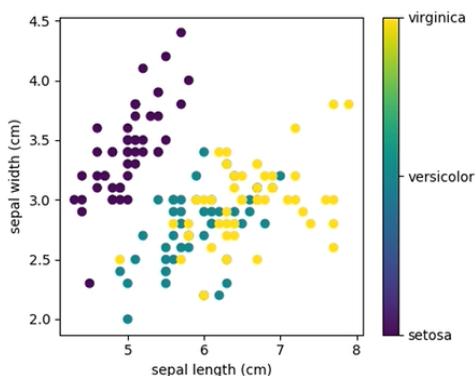


vi. Procesamiento pre-atentivo para visualización: ayuda a atraer el foco de atención sin complicación, por eso es tan importante.

Procesamiento pre-atentiva



- Puede ayudar a atraer rápidamente el foco de atención a un objetivo con una característica visual única (Healey, 2005).
- Puede servir en tareas de analítica visual, por ejemplo, para identificar clusters o outliers.



vii. Memoria de corto alcance:

Memoria de corto alcance

- Funciona procesando unidades de información.
- Indica que es bueno agrupar elementos para almacenarlos y procesarlos.

Ejemplo

- ¿En cuál es más fácil de recordar todas las letras?



L E B P M O W A S T A I A F B

F I A T O P E L B M W S A A B

FIAT OPEL BMW SAAB

4. Color

- a. **Modelo de color:** Método para expresar el color de un objeto usando algún tipo de anotación numérica.
- b. Algunos modelos que veremos son RGB, CMYK, HSL.
- c. **Modelo de color aditivo:**

Aditivo

- Un ejemplo de este modelo es el RGB.
- Funciona mediante la adición de color, es decir, la suma de los tres colores básicos superpuestos termina dando el color blanco.
- Se utiliza en medios que transmiten luz como la televisión.

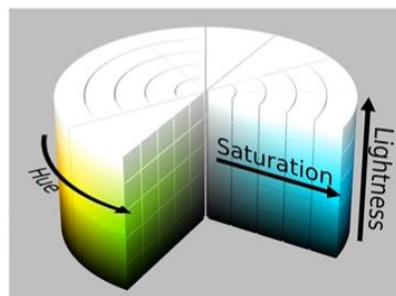
RGB



- d. **Modelo de color basado en propiedades:**

Basado en propiedades (HSL)

- Usado fuertemente por artistas y diseñadores.
- **Hue:** captura lo que normalmente conocemos como colores puros, dejando de lado la mezcla del blanco y del negro. Por ejemplo, rojo, verde, azul, amarillo, púrpura, etc.
- **Saturation:** especifica la intensidad del color (que tan "vivo" está).
- **Lightness:** especifica la cantidad de luz que recibe el color.



- i. **Identidad y magnitud:**

Interpretación término **color** → Confusa en el análisis de visualizaciones

Dos usos:

- Canal de **identidad**
- Canal de **magnitud**

Cuando se usan los términos **hue, saturation y luminance:**

- Tres canales separados
- **Hue:** canal de identidad
- **Saturation y luminance:** magnitud

Cuando usamos el término genérico **color**

- Se refiere a **percepción integral** de estos tres canales en sólo uno, analizado como un **canal de identidad**.

ii. Canal de magnitud:

Luminance	}	Apropriados para los tipos de datos ordenados .
Saturation		

Cuidado

Con usar valores cercanos en regiones **no contiguas**.

¿Por qué?

Nuestro cerebro no percibe las diferencias con exactitud, por causa de los efectos del contraste.

Síntesis

- Un modelo de color es un método para expresar el color de un objeto usando algún tipo de anotación numérica.
- El modelo más recurrido en la visualización es el basado en propiedades del color.
- *Hue* es un canal de identidad.
- Saturación y luminosidad son canales de magnitud.

5. Se usará Matplotlib, Seaborn y Panda:

- a. Hacer el ejercicio de la clase 5.

6. Modelo de visualización Tamara Muzner

a. Espacio de diseño para visualizaciones

Definición

Espacio de diseño de visualización se refiere a todas las posibles combinaciones de variables y valores de atributos que podrían usarse para diseñar una visualización.

¡La cantidad de opciones posibles a explorar para hacer una visualización es muy grande!

Es muy fácil entonces tomar una decisión de diseño que termine con una visualización poco efectiva.

Ejemplo

Dada una base de datos sobre la cual nos piden hacer una visualización:



¿Qué datos y variables elegir para visualizar?

¿Qué canal visual elegir para representar cada variable?

¿Qué tipo de gráfico es más adecuado (de barras, de líneas, de torta, múltiples ejes, etc.)?

¿Se debe agregar interacción o no?

¿Qué paleta de colores usar? Etc.

b. Solución para limitar el espacio de diseño



Buscar patrones y buenas prácticas para tomar decisiones de diseño.



Identificar decisiones de diseño independientes del dominio de aplicación (medicina, finanzas, educación, etc...).



Sintetizar aspectos anteriores en un marco de diseño.

c. Un modelo anidado para el análisis y diseño de visualizaciones:

Tamara Munzner
Investigadora



2009: creación modelo marco

Modelo: permite simplificar la toma de decisiones para diseñar visualizaciones.

Modelo anidado para el diseño y validación de visualizaciones: tres aspectos principales

- ¿Qué?
- ¿Por qué?
- ¿Cómo?

d. Las tres partes del modelo anidado de visualización:

¿Qué?

Referido a la acción de identificar y derivar los **datos** y atributos a visualizar.

¿Por qué?

Identificación de las **tareas** que la visualización debe cumplir.

¿Cómo?

Identificación de la(s) **codificaciones visuales** o gráficos a utilizar.



e. **¿Qué? Datos:**

Identificar tipos de datos, de datasets y de atributos

- 1 ¿Los datos corresponden a una tabla tradicional de base de datos (u hoja de cálculo) o a datos de red?
- 2 ¿Los atributos son de tipo categórico, ordenados, continuos?
- 3 ¿Existen datos de tipo temporal o geográfico?

f. ¿Por qué? Tareas:

Identificar tareas genéricas de visualización

- 1 Identificar *outliers* en un gráfico de dispersión
- 2 Comparar tendencias de subconjuntos de datos de un mismo dataset
- 3 Comparar distribuciones
- 4 Navegar elementos en un mapa

g. ¿Cómo? Codificación:

Identificar la codificación visual y de interacción

Consejo: hacerla después de qué y por qué.

Permite:

1 Mapear un tipo de atributo con la marca y canal más adecuado

2 Elegir el tipo de gráfico más efectivo para ciertas tareas, ejemplo: comparar distribuciones

3 Elegir si es que se requiere y qué tipo de interacción: Navegar elementos en un mapa, filtrar elementos, seleccionar elementos, etc.

h. El modelo anidado: Un marco para validar visualizaciones

i. Distintas formas de validación de una visualización:

Modelo Anidado de Munzner



Creación de visualizaciones nuevas

Análisis jerárquico de visualizaciones existentes

4 niveles



1 Dominio de Aplicación

3 Codificación Visual y de Interacción

2 Abstracción de Datos y Tareas

4 Algoritmo para visualizar

Dominio de Aplicación

Abstracción de Datos/Tareas

Codificación visual y de interacción

Algoritmo

- ii. Abstracción de datos y tareas, se refieren al ¿Qué? Y el ¿por qué?
- iii. Codificación visual y de interacción, se refieren a ¿cómo?
- iv. ¿Qué puede estar fallando en la visualización?

Si una visualización tiene problemas y los usuarios no la usan o la perciben como inefectiva,...

Se puede usar el modelo anidado para identificar el problema:

A nivel de dominio	No se comprendieron las necesidades del dominio de aplicación.
A nivel de datos/tareas	Los datos que se muestran no son los más apropiados.
A nivel visual	El gráfico o la selección de marcas y canales no es la adecuada.
A nivel de algoritmo	La implementación a nivel de software no es eficiente.

Dominio de Aplicación:

- No se comprendieron las necesidades

Abstracción de Datos/Tareas

- No se está mostrando la información adecuada

Codificación visual y de interacción:

- El gráfico es incorrecto

Algoritmo

- Código corre muy lento

v. ¿Qué hacer cuando se identifica el problema?

Pre-Implementación de visualización

S1.1 Observar y entrevistar a usuarios

S3.1 Justificar codificación visual y de interacción

Durante implementación

S4.1 Analizar complejidad computacional del algoritmo

S4.2 Medir tiempo y memoria de sistema

(L1) Problema Identificado está Equivocado
[X] S1.1

(L2) Abstracción de Datos/Tareas está equivocada

(L3) Codificación visual/interacción inefectiva

[X] S3.1

(L4) Algoritmo lento

[X] S4.1

Re-implementar

[X] S4.2

[] S3.2

[] S3.3

[] S2.1

[] S2.2

[] S1.2

vi. ¿Qué hacer cuando se identifica el problema?

Post-implementación

- S3.2 Estudio de usabilidad informal (cualquier usuario)
- S3.3 Estudio de laboratorio, medir tiempo y errores por tarea
- S2.1 Testear en usuarios finales, evidencia anecdótica
- S2.2 Estudio de campo, con Visualización en producción
- S1.2 Observar tasas de adopción

(L1) Problema Identificado está Equivocado
[] S1.1

(L2) Abstracción de Datos/Tareas está equivocada

(L3) Codificación visual/interacción inefectiva

[] S3.1

(L4) Algoritmo lento

[] S4.1

Re-implementar

[] S4.2

[X] S3.2

[X] S3.3

[X] S2.1

[X] S2.2

[X] S1.2

7. Qué: tipo de datos y datasets

a. ¿Por qué reconocer Qué?



Reconocer los datos es el primer paso en la elaboración de visualizaciones de información.



Muchas decisiones de diseño vienen impulsadas por el tipo de información con el que se cuenta.



Reconocer erróneamente trae como consecuencia diseños de visualización poco efectivos.

Semántica de dato

- Es el significado del mundo real del dato.
- Representa algún concepto externo.
- Está determinado por un contexto.

Tipo de dato

- Es su interpretación estructural o matemática.
- Va más allá de ser un valor numérico o de texto.

14, 2.8, 30, 30, 15, 1001

Albahaca, 7, S, Pera

b. Tipos de datos y Datasets

Tipos de Datos

1. Atributos
2. Ítems
3. Conexiones
4. Posiciones
5. Otros

Tipos de *Datasets*

1. Tabulares
2. Redes y árboles
3. Geométricos
4. Colecciones
5. Otros

c. Tipos de atributos y semántica

i. Tipos de atributos (2):

Categóricos

- Género
- Tipos de frutas
- Nacionalidad

Se emparejan con canales de **identidad**.

Ordenados

Ordinal

- Talla de camisetas
- Rankings

Se emparejan con canales de **magnitud** de forma discreta.

Cuantitativo

- Altura
- Precio
- Porcentaje de cambio

Se emparejan con canales de **magnitud** de forma continua.

- ii. Canales de **Identidad**, por ejemplo, el color.
- iii. Canales de **Magnitud** de forma discreta, por ejemplo, opacidad en colores, superficies.
- iv. Canales de **Magnitud** de forma continua de alta precisión, por ejemplo, posición, largo, ángulo.

v. **Tipos de semánticas:**

Atributo de llave o de valor

- Un atributo de llave actúa como un índice para poder identificar un ítem o valor.
- También, se conoce como atributo independiente.
- Un atributo de valor en cambio no permite identificar un ítem, solo lo describe.
- También, se conoce como atributo dependiente.

Atributo temporal

- Son muy ricos en jerarquía.
- Se puede tratar en múltiples escalas.
- Se puede asociar a muchos tipos de canales distintos.
- Se deben manejar con cuidado dependiendo del contexto.

8. Por qué: tareas de visualización

- En esta clase vamos a entender en qué consiste la pregunta “**por qué**” y la importancia de responderla antes de comenzar a programar una visualización.
- Reconoceremos las diferentes acciones y objetivos que se describen en esta tarea.

a. **Ejemplo: No sólo datos justifican la visualización:**

Primera etapa del modelo anidado de diseño de visualizaciones identificamos	
	Los tipos de datos y datasets: Qué
1	Responder Qué: nos da una idea de cómo codificar visualmente una visualización (etapa 3: ¿Cómo?).
2	La tarea “a resolver” por la visualización (¿Por qué?) puede cambiar el tipo de gráfico o la codificación visual.



Ejemplo

Si nos interesa visualizar la distribución del tamaño de pétalos de una flor,

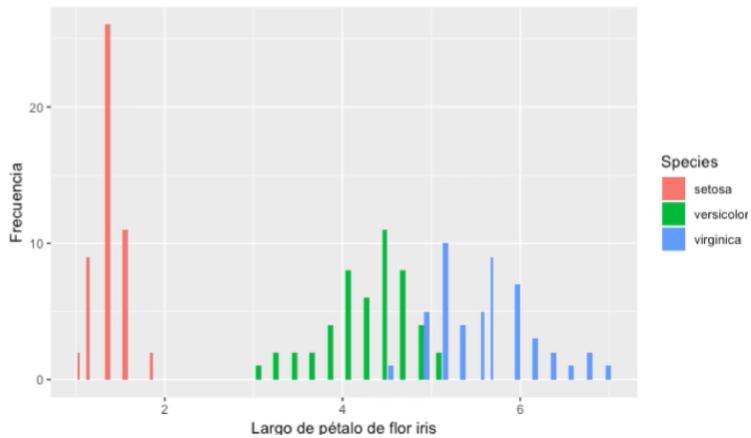
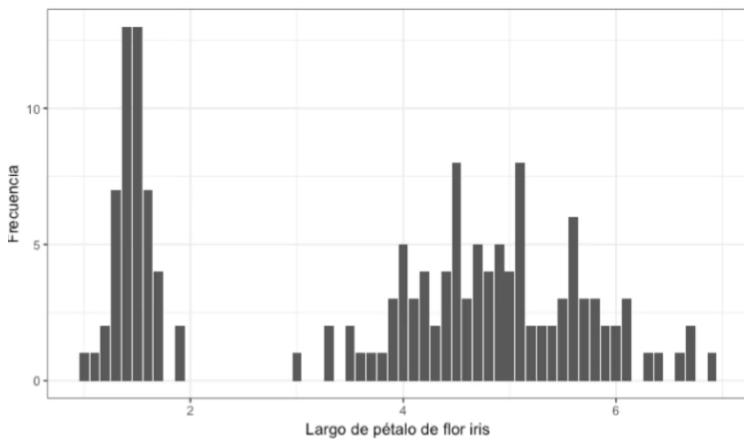
Un histograma podría ser una buena opción.

Si queremos comparar la distribución de tamaños de pétalos de 3 tipos de flores,

Una visualización que agregue el canal de color y sobreponga 3 histogramas es una mejor opción.

El por qué del gráfico,

Modifica la decisión del tipo de gráfico para los mismos datos.



Luego de identificar los datos, el marco anidado de visualización indica **identificar tareas: el por qué de la visualización.**



Las tareas se compone de **acciones** y **objetivos**.

b. Por qué: Identificar tareas genéricas de visualización



c. Acciones:



Nivel alto:
Analizar



Nivel medio:
Buscar



Nivel bajo:
Consultar

d. Acciones – Analizar:

→ Consume

→ *Discover*



→ *Present*



→ *Enjoy*

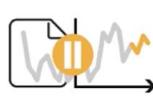


→ Produce

→ *Annotate*



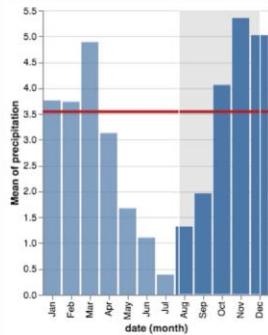
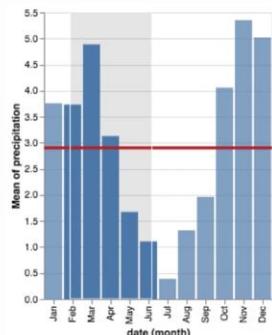
→ *Record*



→ *Derive*



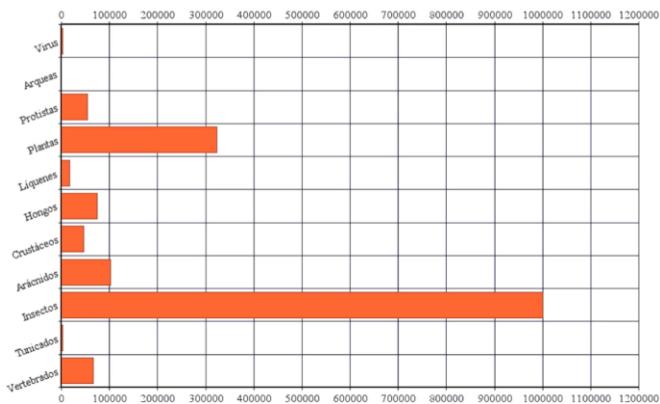
Ejemplo: Derivar



e. Acciones – Buscar:

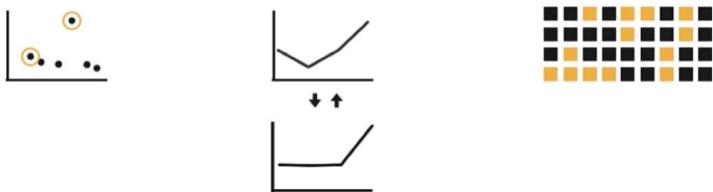
	Target known	Target unknown
Location known	• • • <i>Lookup</i>	• ?? <i>Browse</i>
Location unknown	⟨ ⟩ <i>Locate</i>	⟨ ⟩ <i>Explore</i>

Ejemplo de buscar: Locate y Lookup

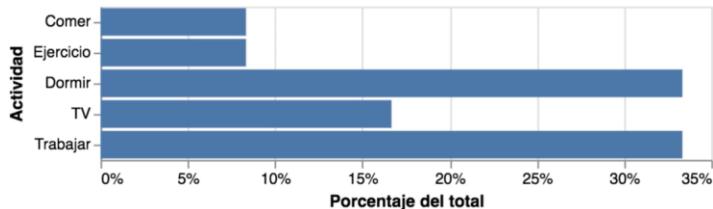


f. Acciones – Consultar:

→ Identify → Compare → Summarize



Ejemplo: Comparar



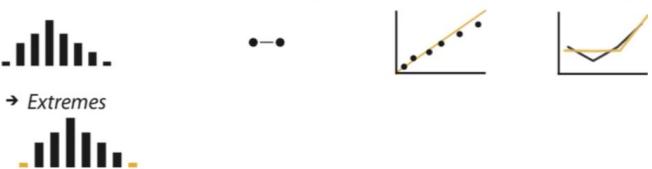
g. Objetivos

i. Objetivo I: Para datos tabulares

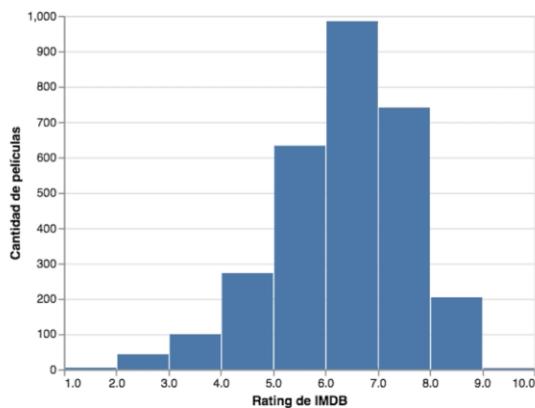
→ All Data

- Trends
 - Outliers
 - Features
- 

→ Attributes

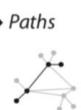
- One
 - Many
- Distribution
 - Dependency
 - Correlation
 - Similarity
- 
- Extremes
- 

Ejemplo: gráfico de un atributo como objetivo (target)



ii. Objetivos II: Datos de red o espaciales

→ Network Data

- Topology
- 
- Paths
- 

→ Spatial Data

- Shape
- 

Ejemplo: Paths como objetivo (target)

si fuera posible. Judíos y gentiles, blancos o negros. Tenemos que ayudarnos unos a otros. Los seres humanos somos así. los seres. El camino de la vida puede ser libre y hermoso, pero lo hemos perdido. La codicia ha ensuculado las alturas. nosotros. Ahora mismo mi voz llega a millones de seres en todo el mundo, a millones de hombres desesperados, mujeres y niños. a la felicidad. ¡Soldados, en nombre de la democracia, debemos unirnos todos!

que encarcelar a gente inocente. A los que puedan oírme, les digo: no desesperéis. La desdicha que padecemos no es más que la pasajera codicia y la amargura de hombres que temen seguir el camino del progreso humanaidad un futuro. Y a la vejez, seguridad. Con la promesa de esas cosas, las fieras alcanzaron el poder.

que puedas oírme, les digo: no desesperéis. La desdicha que padecemos no es más que la pasajera codicia y la amargura de hombres que temen seguir el camino del progreso humanaidad un futuro. Y a la vejez, seguridad. Con la promesa de esas cosas, las fieras alcanzaron el poder.

a nadie. En este mundo hay sitio para todos. La Tierra es rica y puede alimentar a todos los seres. El camino de la vida puede ser libre y hermoso, pero lo hemos perdido.

millones de seres en todo el mundo, a millones de hombres desesperados, mujeres y niños. Víctimas de un sistema que hace torturar a los hombres y encarcelar a gente inocente.

esos hombres desesperados, mujeres y niños. Víctimas de un sistema que hace torturar a los hombres y encarcelar a gente inocente.

, sino ayudar a todos si fuera posible. Judíos y gentiles, blancos o negros. Tenemos que ayudarnos unos a otros.

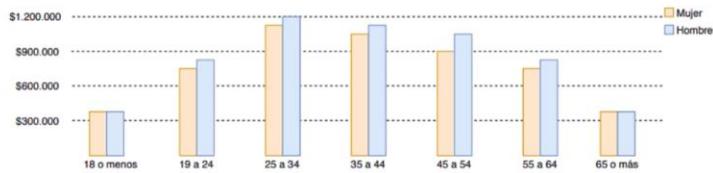
honestos que en realidad os desprecian, os rechazan, regalámonos vuestras vidas y os dicen lo que tendréis que hacer, que pensar y que sentir.

carne de caballo. No os entregáis a esos individuos inhumanos, hombres máquinas, con cerebros y corazones de máquinas. Vosotros no sois máquinas; no sois ganado.

luchar para liberar al mundo. Para derribar barreras nacionales. Para eliminar la ambición, el odio y la intolerancia. Luchemos por el mundo de la razón.

h. Ejemplo: Integrando Qué y Por qué

Sueldo líquido promedio por grupo etaria y sexo



Qué (What)	Dataset: Tabular con un atributo cuantitativo (sueldo) y 2 cualitativos (rango etaria) y sexo.
Por Qué (Why)	Presentar → Lookup → Compare

- El **Por qué** permite describir de forma abstracta por qué las personas usan una visualización a través de tareas.
- Las tareas se describen con tuplas de acciones y objetivos.

$$Tarea = \{acción, objetivo\}$$

9. Cómo: Modo de representación



Cómo una visualización puede ser construida a partir de un conjunto de elecciones de diseño.



Se compone de 4 familias: *Encode*, *Manipulate*, *Facet* y *Redution*.

a. Familias de cómo implementar

i. Acciones – **Encode**: Como se va a codificar la información

Arrange

→ Express → Separate



→ Order → Align



→ Use



Map

→ Color

→ Hue



→ Saturation



→ Luminance



→ Size, Angle, Curvature, ...



→ Shape



→ Motion

Direction, Rate, Frequency, ...



ii. Acciones – **Manipular**: Como manipular la información

④ Change



④ Select

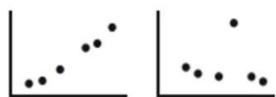


④ Navigate

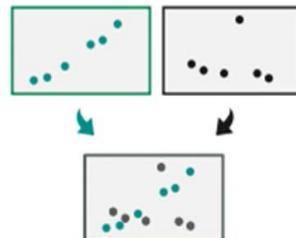


iii. **Acciones – Facet:** Como dividir la pantalla de visualización

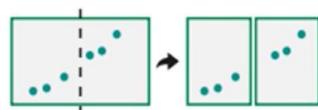
④ Juxtapose



④ Superimpose



④ Partition



iv. **Acciones – Reduce:** Se encarga de ver como reducir la complejidad de los datos

④ Filter



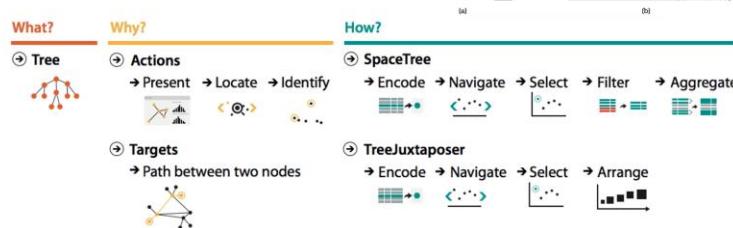
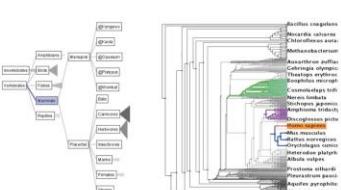
④ Embed



④ Aggregate



Ejemplo



10. Recomendaciones generales

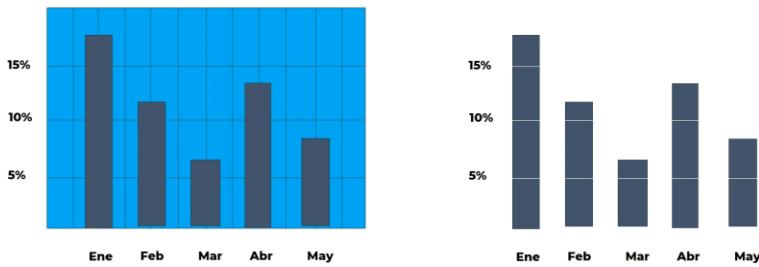
a. Data-ink ratio (pixeles con fundamento)

Principio que establece que el fundamento del uso de recursos visuales es para representar datos solamente.

$$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink}}$$

Mientras más cercana a 1 la proporción, mejor. Cada pixel está representado.

Ejemplo:



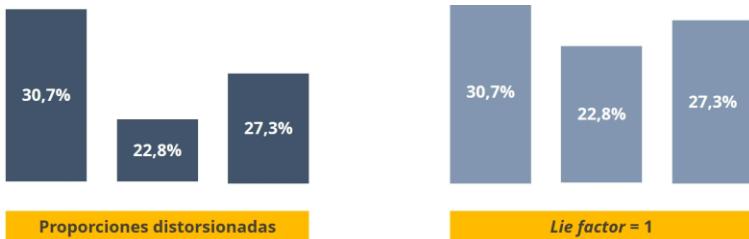
b. Lie factor (veracidad visual)

Principio que establece las proporciones de tamaños en el gráfico deben ser lo más parecidas a las proporciones entre los datos.

$$\text{Lie factor} = \frac{\text{size effect in graph}}{\text{size effect in data}}$$

Mientras más cercana a 1 la proporción, mejor. El gráfico refleja de la misma manera como se comparan los datos, pero de forma visual.

Ejemplo:



c. No 3D injustificado

- La representación 3D presenta muchos desafíos.
- Es fácil que se produzca oclusión de elementos.
- La posibilidad de navegación es costosa en tiempo.
- Distorsión por perspectiva.
- Hay situaciones donde el 3D si es justificado.
- Incluso existe 2D injustificado.

d. Primero lo general, detalle en demanda

Principio enunciado por Ben Shneiderman (1996) que pone énfasis en la interacción de dos requisitos en una visualización:



Tener una perspectiva general primero

Dar la opción de conocer detalle

e. Responsividad

Desde el área de Interacción Humano-Computador (IHC), es un mínimo que la interacción con visualizaciones tengan buen grado de responsividad:



Retroalimentación visual

Latencia temporal baja

Encontrar los costos y definir límites

f. Consistencia

En herramientas o informes de múltiples visualizaciones, es importante mantener consistencia de decisiones:



Consistencia interna

Consistencia externa

g. Otros

- Los ojos vencen a la memoria

- Auto explicación y contexto

- Tipografía

- Daltonismo

What?

Datasets

→ Data Types

→ Items → Attributes → Links → Positions → Grids

→ Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	

Attributes

→ Attribute Types

→ Categorical



→ Ordered

→ Ordinal

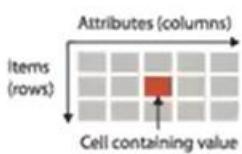


→ Quantitative

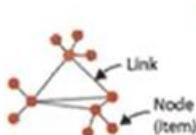


→ Dataset Types

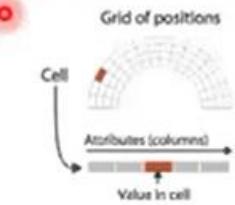
→ Tables



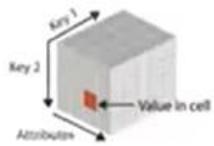
→ Networks



→ Fields (Continuous)



→ Multidimensional Table



→ Trees



→ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



→ Geometry (Spatial)



→ Dataset Availability

→ Static



→ Dynamic



→ Attribute Types

→ Categorical

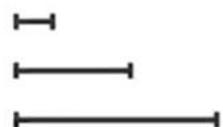


→ Ordered

→ *Ordinal*



→ Quantitative



→ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



Why?

Actions		Targets													
Analyze <ul style="list-style-type: none"> → Consume <ul style="list-style-type: none"> → Discover → Present → Enjoy → Produce <ul style="list-style-type: none"> → Annotate → Record → Derive 		All Data <ul style="list-style-type: none"> → Trends → Outliers → Features Attributes <ul style="list-style-type: none"> → One → Many → Distribution → Dependency → Correlation → Similarity → Extremes 													
Search <table border="1"> <thead> <tr> <th></th> <th>Target known</th> <th>Target unknown</th> </tr> </thead> <tbody> <tr> <td>Location known</td> <td></td> <td>Lookup </td> </tr> <tr> <td>Location unknown</td> <td></td> <td>Locate </td> </tr> <tr> <td></td> <td></td> <td>Explore </td> </tr> </tbody> </table>			Target known	Target unknown	Location known		Lookup	Location unknown		Locate			Explore	Network Data <ul style="list-style-type: none"> → Topology → Paths Spatial Data <ul style="list-style-type: none"> → Shape 	
	Target known	Target unknown													
Location known		Lookup													
Location unknown		Locate													
		Explore													
Query <ul style="list-style-type: none"> → Identify → Compare → Summarize 															

How?

Encode	Manipulate	Facet	Reduce
Arrange <ul style="list-style-type: none"> → Express → Separate Order <ul style="list-style-type: none"> → Align Use <ul style="list-style-type: none"> 	Map <ul style="list-style-type: none"> from categorical and ordered attributes → Color <ul style="list-style-type: none"> → Hue → Saturation → Luminance → Size, Angle, Curvature, ... → Shape <ul style="list-style-type: none"> → Motion <ul style="list-style-type: none"> Direction, Rate, Frequency, ... 	Change <ul style="list-style-type: none"> Select <ul style="list-style-type: none"> Navigate <ul style="list-style-type: none"> 	Juxtapose <ul style="list-style-type: none"> Partition <ul style="list-style-type: none"> Superimpose <ul style="list-style-type: none"> Embed <ul style="list-style-type: none">
			Filter <ul style="list-style-type: none"> Aggregate <ul style="list-style-type: none">

RANKING CANALES

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)	· · ■ ■ ■ ■

⑤ Identity Channels: Categorical Attributes

Spatial region	■ ■ ■ ■
Color hue	■ ■ ■ ■
Motion	• • • •
Shape	+ ● ■ ▲

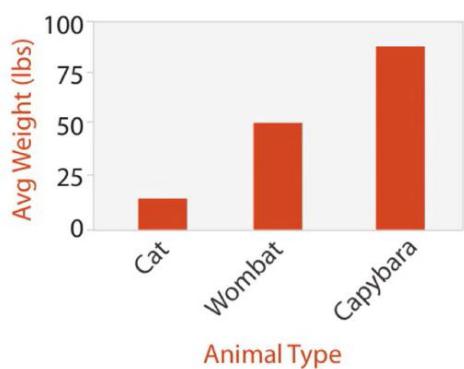
- **expressiveness principle**
 - match channel and data characteristics

11. Gráficos de barras, barras apiladas:

a. Gráfico de barras:

- Se usa principalmente para representar un **atributo categórico** y otro **numérico**.
- En el eje X se muestran los datos **categóricos** y en el Eje Y los datos **numéricos**.
- Ejemplo,

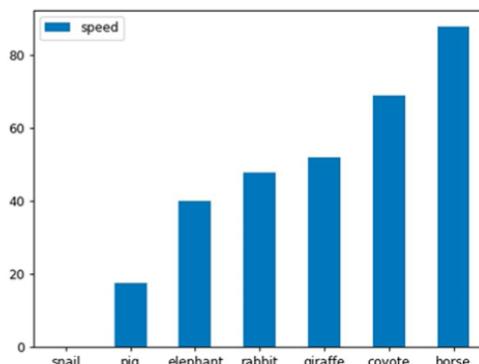
Animal	Peso Promedio (lbs)
Cat	20
Wombat	50
Capybara	80



- iv. Se utiliza el canal de **longitud** de cada barra para codificar el valor.
- v. Gráfico muy efectivo para hacer **comparaciones**.
- vi. El orden de las barras queda a decisión del diseñador de la gráfica (alfabético, de menor a mayor en función del valor Y, etc).
- vii. Ejemplo:

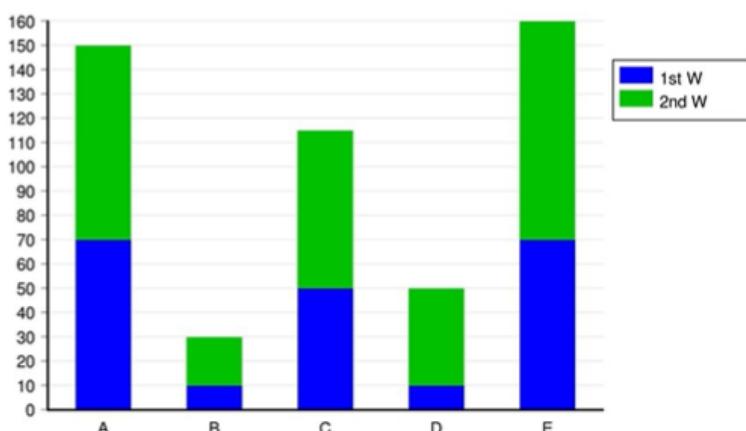
Ejemplo de gráfico de barras

- 1 El canal de color no codifica información.
- 2 El eje X presenta la variable categoría (especie).
- 3 El eje Y presenta la variable numérica (velocidad).
- 4 Gráfico efectivo hasta con decenas a cientos de valores en el eje X (valores del atributo categórico).



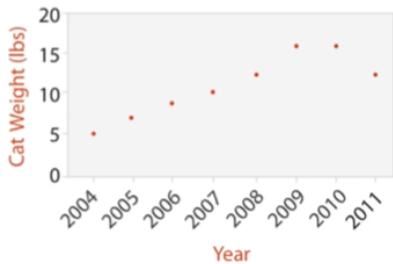
b. Gráficos de barras apiladas:

- i. Tipo especial de gráficos de barra, permite codificar visualmente 2 variables categóricas, y una numérica.
- ii. Cuando se tiene que elegir que variables categóricas codificar en el eje X y cual dentro de los valores de colores. Normalmente, el atributo categórico con más categorías se pone en el eje X. El atributo categórico con menos categorías en las barras, como colores diferentes.
- iii. La categoría en el eje X puede tener decenas de valores.
- iv. La categoría codificada dentro de las barras (con colores) puede tener como máximo 10 a 12 valores.

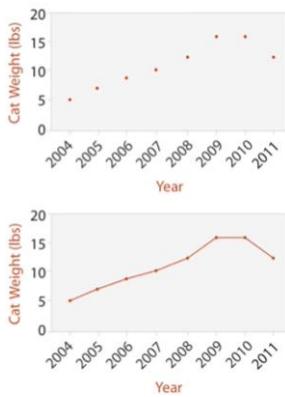


12. Gráficos de puntos y de líneas

- Utilizados principalmente para presentar tendencias.
- En el eje X se pone una variable ordinal (típicamente datos temporales).
- El caso del gráfico de puntos sólo usa puntos para codificar el valor de una posición en el eje X.



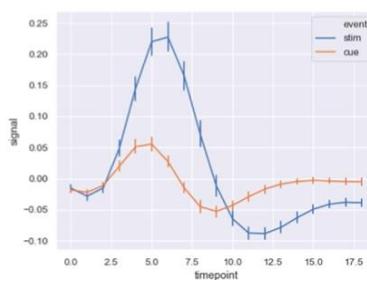
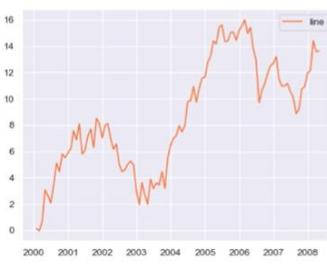
- El gráfico de líneas es similar al de puntos, sin embargo, las líneas que unen los puntos permiten dar una sensación de continuidad.
- En el eje X pueden perfectamente ponerse cientos de niveles (valores) y el gráfico escala bien visualmente.
- No confundir el gráfico de puntos con el gráfico de dispersión, donde tanto el eje X como el eje Y presentan valores de atributos numéricos.



- Ejemplo de gráficos de puntos y de líneas:

1 En algunos casos, se asocian barras de error a los puntos (Ej.: intervalos de confianza) cuando el valor indica un promedio.

2 Es común dibujar más de una serie en el gráfico.

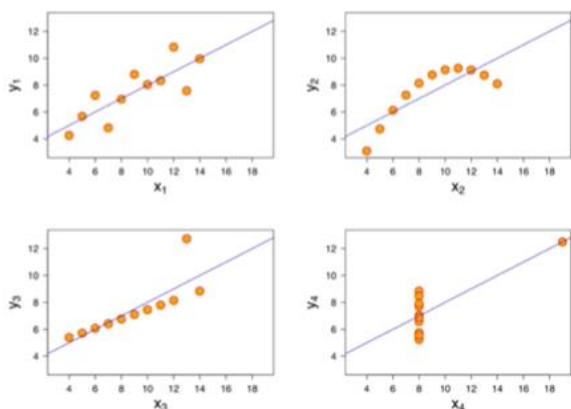


- h. Los gráficos de puntos y de líneas presentan una variable ordenada en el eje X y en el eje Y un valor numérico.
- i. Se usan principalmente para presentar tendencias: cambios de una variable en el tiempo.
- j. Matplotlib, Seaborn, y Pandas permite realizar esta visualización.

13. Gráfico de dispersión, gráfico de burbujas

a. Gráfico de dispersión:

- i. Utilizado para visualizar relaciones entre dos atributos numéricos.
- ii. Se usa la marca de punto y el canal de posición horizontal y vertical (X, Y) para la representación visual de cada marca.
- iii. Permite visualizar fácilmente si es que existe correlación, tanto positivo como negativo.
- iv. En inglés se conoce como scatterplot.



v. Ejemplo de gráfico de dispersión:

Gráfico 1
Muestra una **correlación positiva**.

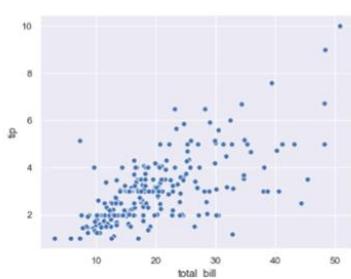
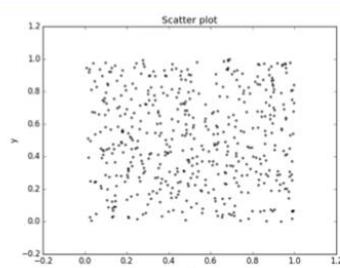
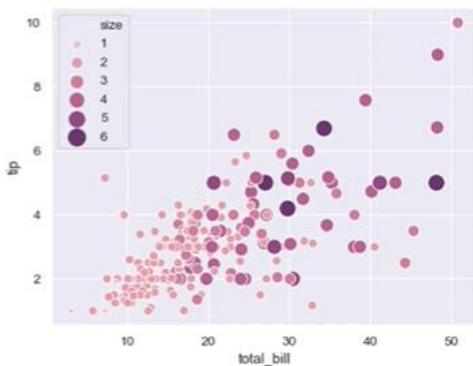


Gráfico 2
Ilustra que no hay relación entre las variables **X e Y**.



b. **Gráfico de burbuja:**

- i. Extensión del gráfico de dispersión.
- ii. Permite codificar variables categóricas adicionales usando los canales de color y tamaño de los puntos.
- iii. El caso del gráfico de puntos sólo usa puntos para codificar el valor de una posición en el eje X.

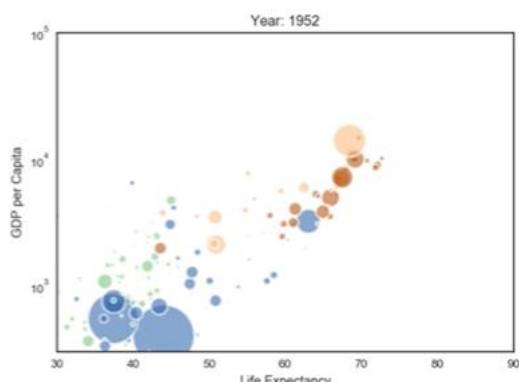


iv. Ejemplo de gráfico de burbujas:

El famoso gráfico de Hans Rosling

Muestra información de tasas de desnutrición.

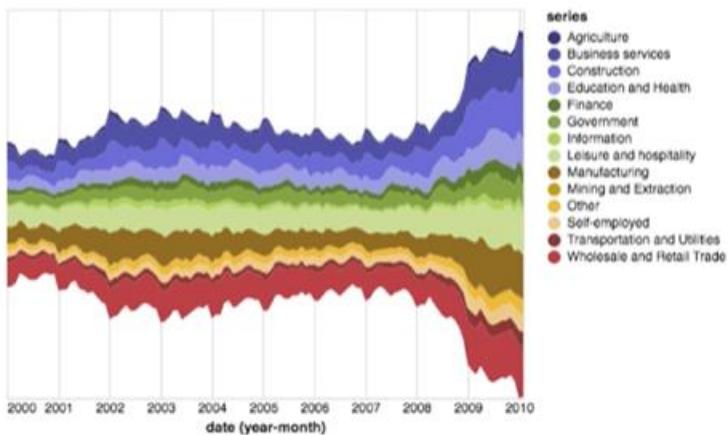
https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen



- c. Los gráficos de dispersión y de burbujas permiten presentar visualmente relaciones entre variables numéricas.
- d. El gráfico de burbujas es una extensión del gráfico de dispersión. Permite representar variables categóricas adicionales usando canales de color y tamaño.
- e. Matplotlib, Seaborn y Pandas permite realizar esta visualización.

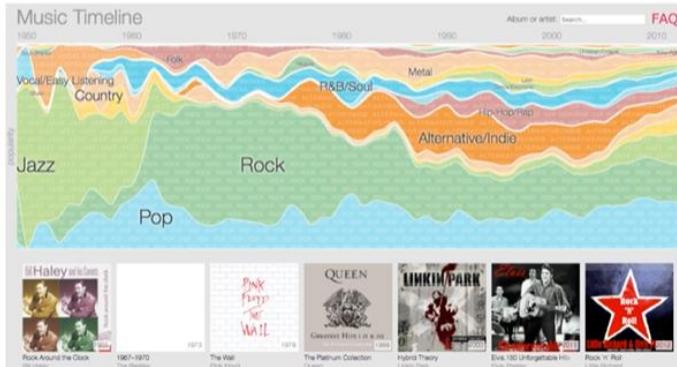
14. Gráficos de flujos

- Utilizado para visualizar tendencias en el tiempo.
- A diferencia del gráfico de líneas, usa el canal de área y de color para presentar la información de tendencias.
- Se codifican visualmente tres variables: un atributo temporal (eje X), un atributo numérico (para cuantificar el tamaño de las áreas), y un atributo categórico (para mostrar las distintas áreas, normalmente codificadas con color).
- En inglés se conoce como Streamgraph.



- Ejemplo de gráfico de flujo:

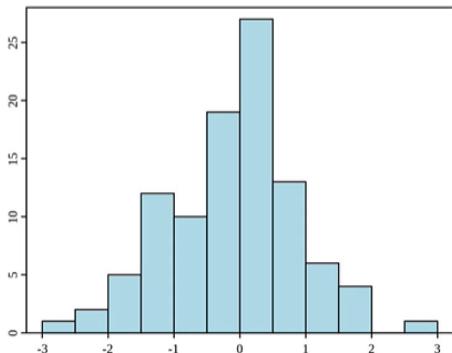
El siguiente gráfico de flujo muestra la tendencia de popularidad de distintos tipos de música.



- Los gráficos de flujo permiten codificar visualmente tendencias de distintas categorías usando canales de área y de color.
- Este gráfico es fácil de entender, ya que es algo familiar y reduce el esfuerzo cognitivo para interpretar la visualización.
- La biblioteca de software Altair permite realizar esta visualización.

15. Histograma, gráfico de caja, gráfico de violín

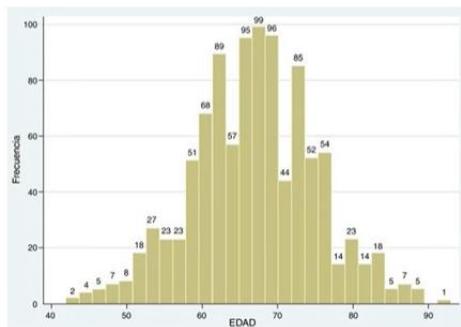
- a. Se utilizan los análisis estadísticos.
- b. **Histograma:**
 - i. Este término fue acuñado en 1891 por el matemático y estadístico Karl Pearson.
 - ii. Gráfico muy usado en estadística, especialmente análisis de datos exploratorio.
 - iii. Permite conocer la distribución de los valores (eje X) de un atributo numérico (cuantitativa y continuo).
 - iv. Ejemplo: la distribución de edades de un grupo de personas.



- v. Permite identificar rápidamente los valores más comunes (Ej.: edad más común).
- vi. También, permite identificar valores menos comunes y datos extraños (barras más pequeñas).
- vii. Similar a un gráfico de barras, pero diferente en cuanto a la obtención de datos.
- viii. La cantidad de barras, en lugar de corresponder a categorías fijas, corresponden a rangos de valores calculados automáticamente.
- ix. Ejemplo de histograma:

Gráfico de histograma muestra

La distribución de edades de un conjunto de pacientes a quienes se les realizó una biopsia prostática. La gran mayoría tenía entre 60 y 78 años de edad, dados que las barras son más altas para esos valores de edad en el eje X.



c. Gráfico de caja:

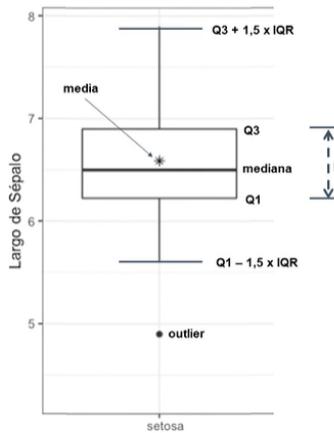
- Permite visualizar la distribución de datos de una variable numérica (cuantitativa continua), enfatizando medidas estadísticas como la mediana, cuartiles y outliers. En inglés se denomina box plot.
- La caja central está delimitada por el cuartil Q1 (percentil 25) y el cuartil Q3 (percentil 75). La línea dentro de la caja indica la mediana (percentil 50).
- La media aritmética no siempre se incluye en el gráfico de caja. Cuando se incluye, se usa una marca adicional (un asterisco, por ejemplo).
- Se agregan dos líneas adicionales fuera de la caja que indican unos límites mínimos y máximos esperados, equivalentes a:

$$\text{Min} = \text{Q1} - 1.5 \times \text{IQR}$$

$$\text{Max} = \text{Q3} + 1.5 \times \text{IQR}$$

Donde IQR significa “**rango intercuartil** = $\text{Q3} - \text{Q1}$ ”

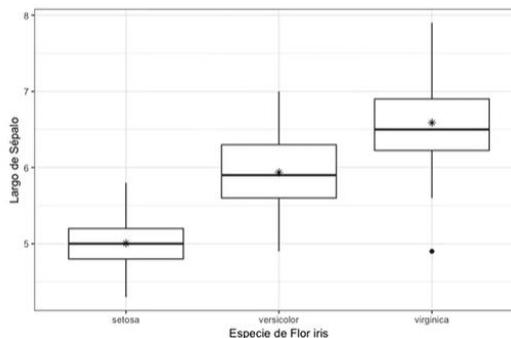
- Fuera de los límites mínimos y máximos, se encuentran los “outliers”.



- Ejemplo de gráfico de caja:

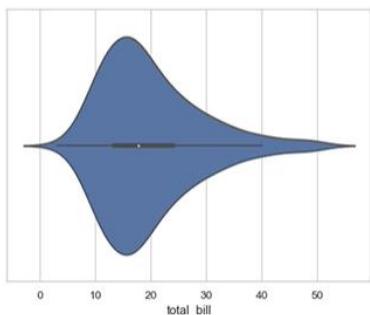
Gráfico de caja por cada variable categórica (especie de flor)

Permite comparar fácilmente la distribución de valores de la variable mostrada en el eje Y (largo de sépalo) entre las distintas categorías (especies de flor iris).



d. **Gráfico de violín:**

- i. Similar al gráfico de caja en cuanto a su propósito: Presentar información resumida, visualmente, de una distribución de datos.
- ii. Diferencia a los gráficos de caja: Porque al usar un pre-cálculo de la distribución de probabilidad estimada, dibuja una forma suavizada de la distribución directamente desde los datos.
- iii. Es más efectivo que el gráfico de caja en términos visuales: Pero como la forma suavizada es aproximada, puede contener sesgos si hay muy pocos datos.

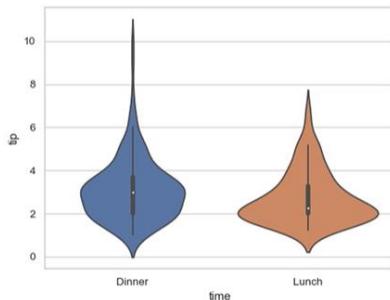


iv. Ejemplo de gráfico de violín:

Uso de gráfico de violín

Comparar la distribución de propinas que se dan en un restaurante en el almuerzo (*lunch*) versus en la cena (*dinner*).

Se observa una distribución de datos con mayores propinas en la cena.



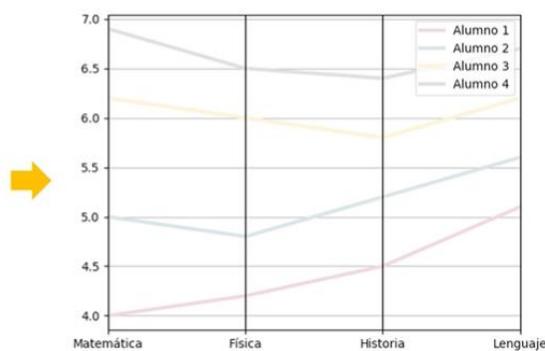
- e. Los histogramas, los gráficos de caja y los gráficos de violín permiten mostrar de forma resumida.
- f. El gráfico de caja presenta visualmente información como la mediana, cuartiles y outliers.
- g. El gráfico de violín intenta presentar información de la distribución que es ocultada visualmente por el gráfico de caja.
- h. Matplotlib, Seaborn y Pandas permiten realizar estas visualizaciones.

16. Múltiples ejes, gráficos radiales y de torta

a. Gráficos con múltiples ejes:

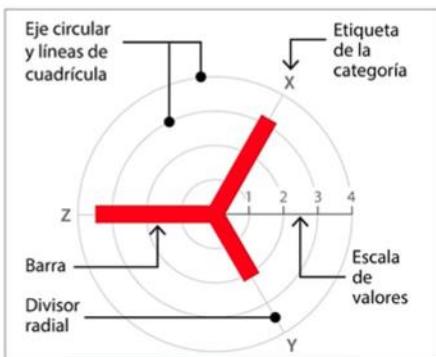
- i. El más conocido es el **Parallel Coordinates**.
- ii. Utiliza el canal especial para visualizar toda la información.
- iii. Se utiliza en datasets que presentan múltiples atributos.
- iv. Los atributos se codifican como ejes paralelos mientras que cada ítem del dataset es una línea que cruza 1 vez cada eje.
- v. Se utiliza generalmente, para **mostrar la relación/patrón entre 2 atributos**, pero visualizando varios pares de atributos al mismo tiempo.
- vi. Ejemplo:

Matemática	Física	Historia	Lenguaje
4	4.2	4.5	5.1
5	4.8	5.2	5.6
6.2	6	5.8	6.2
6.9	6.5	6.4	6.7



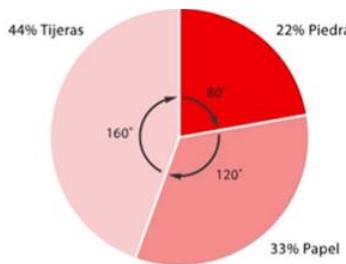
b. Gráficos radiales:

- i. Visualización que utiliza una disposición circular y el largo de barra para codificar la información.
- ii. El largo de la barra representa un valor en una escala y los divisores radiales se utilizan para indicar cada categoría.
- iii. Representado también como un gráfico de barra.
- iv. Se recurre a esta forma cuando **se quiere representar una periodicidad**. Por ejemplo, datos referidos a los meses del año.



c. **Gráficos de torta:**

- i. Se utiliza una disposición circular para visualizar los datos, pero se reemplaza el canal del largo de la barra por el canal de ángulo.
- ii. Cada elemento es codificado por un segmento del círculo y su valor es representado por el ángulo de dicho segmento.
- iii. Se utiliza principalmente para **mostrar proporciones dentro de un total.**



Datos			
Piedra	Papel	Tijeras	TOTAL
2	3	4	9
Para calcular porcentajes			
$2/9=22\%$	$3/9=33\%$	$4/9=44\%$	100%
Grados para cada «porción de tarta»			
$(2/9) \times 360 = 80^\circ$	$(3/9) \times 360 = 120^\circ$	$(4/9) \times 360 = 160^\circ$	360°

- d. Parallel Coordinates permite mostrar la relación/patrón entre 2 atributos, pero visualizando varios pares de atributos al mismo tiempo.
- e. Los gráficos radiales permiten codificar la misma información que un gráfico de barra, pero agregar la atribución de periodicidad a los datos.
- f. El gráfico de torta permite codificar proporciones dentro de un total.
- g. El módulo plotting de pandas permite realizar el Parallel Coordinates.
- h. Matplotlib permite realizar los gráficos radiales y de torta.

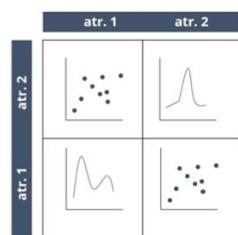
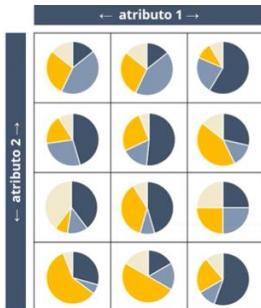
17. Matriz de gráficos

- a. Permite codificar mucha información en poco espacio.
- b. Alineamiento matricial:
 - i. Una matriz es una forma de usar espacio bidimensional.
 - ii. Datasets con al menos dos atributos de tipo llave.
 - iii. Visualizar las combinaciones entre valores de atributos.
 - iv. Visualizar las combinaciones de pares de atributos.

		← atributo 1 →
← atributo 2 →		

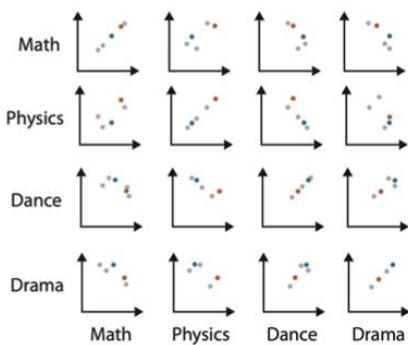
		atr. 1 atr. 2
atr. 2	atr. 1	

- c. Cada celda de la matriz es otra codificación por sí sola.
- d. Cada codificación no necesariamente debe ser la misma.
- e. Datasets con múltiples atributos suelen ser compatibles.
- f. Es una forma eficiente de mostrar la relación entre variadas combinaciones de atributos.
- g. Ejemplo:



h. Matriz de scatterplots o gráficos de dispersión:

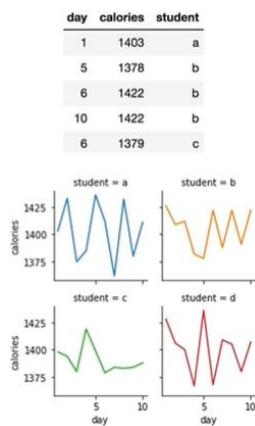
- i. Un caso común de matriz de gráficos.
- ii. Un scatterplot permite apreciar una posible relación entre dos atributos cuantitativos.
- iii. Un dataset con más de dos atributos cuantitativos es muy compatible a esta codificación.



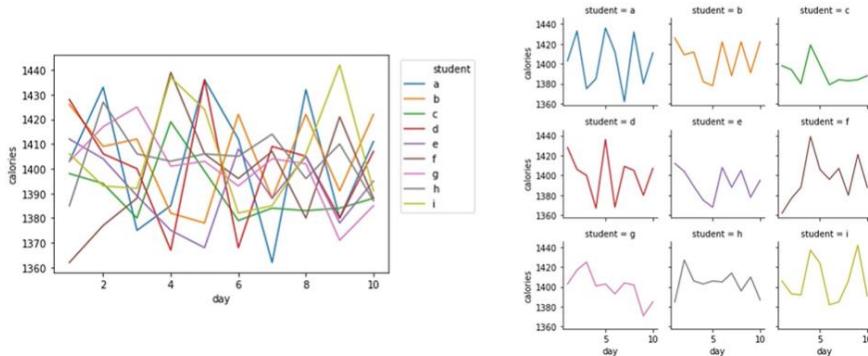
- i. Matriz de gráficos que permite codificar más información.
- j. Matriz de scatterplots permite revisar la relación entre varios pares de atributos al mismo tiempo.
- k. Seaborn permite realizar matrices de gráficos.

18. Pequeños múltiples o Small Múltiples

- a. Un caso particular de matriz de gráficos.
- b. Cada celda presenta la misma codificación.
- c. Todos los gráficos comparten los mismos atributos.
- d. Cada gráfico posee sólo 1 categoría posible respecto al atributo definido.
- e. Utilizado comúnmente **cuando hay un atributo con demasiadas categorías**.
- f. Otra aplicación: como alternativa a las animaciones, donde cada paso de la animación se ve en un gráfico distinto en vez de mostrar un paso a la vez.



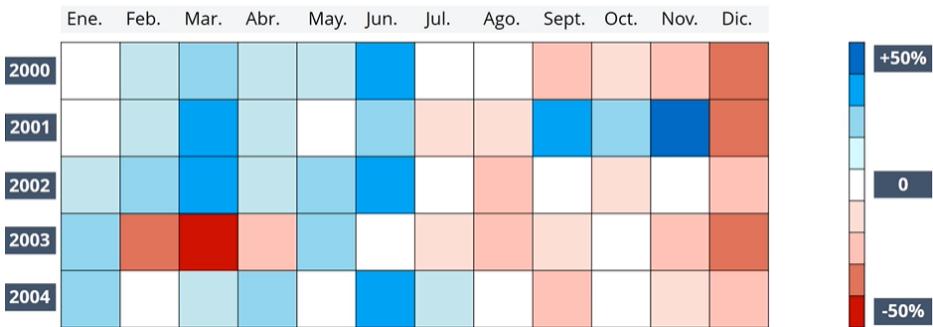
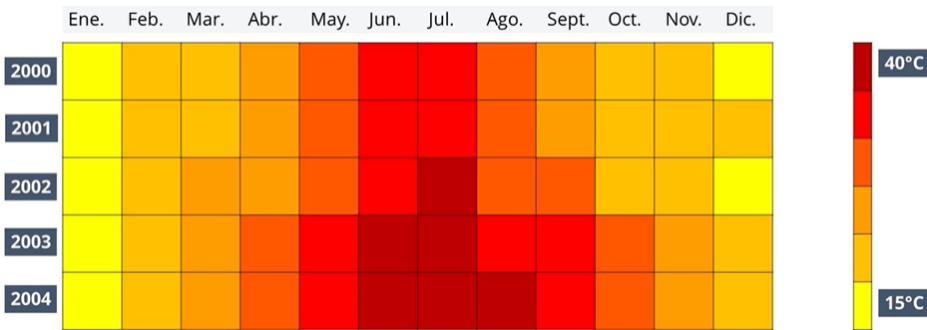
- g. Ejemplo:



- h. Esta visualización es un caso particular de matriz de gráficos.
- i. Utilizados comúnmente para un atributo que representa demasiadas categorías o para visualizar animaciones.
- j. Seaborn permite realizar esta visualización.

19. Mapa de calor

- a. Caso específico de alineamiento matricial de dos atributos.
- b. Cada celda de la matriz codifica un tercer atributo.
- c. Atributo cuantitativo, secuencial o divergente.
- d. Permite ver la distribución del tercer atributo a lo largo de las combinaciones de las filas y columnas.
- e. Eficiente en espacio.
- f. Ejemplo:

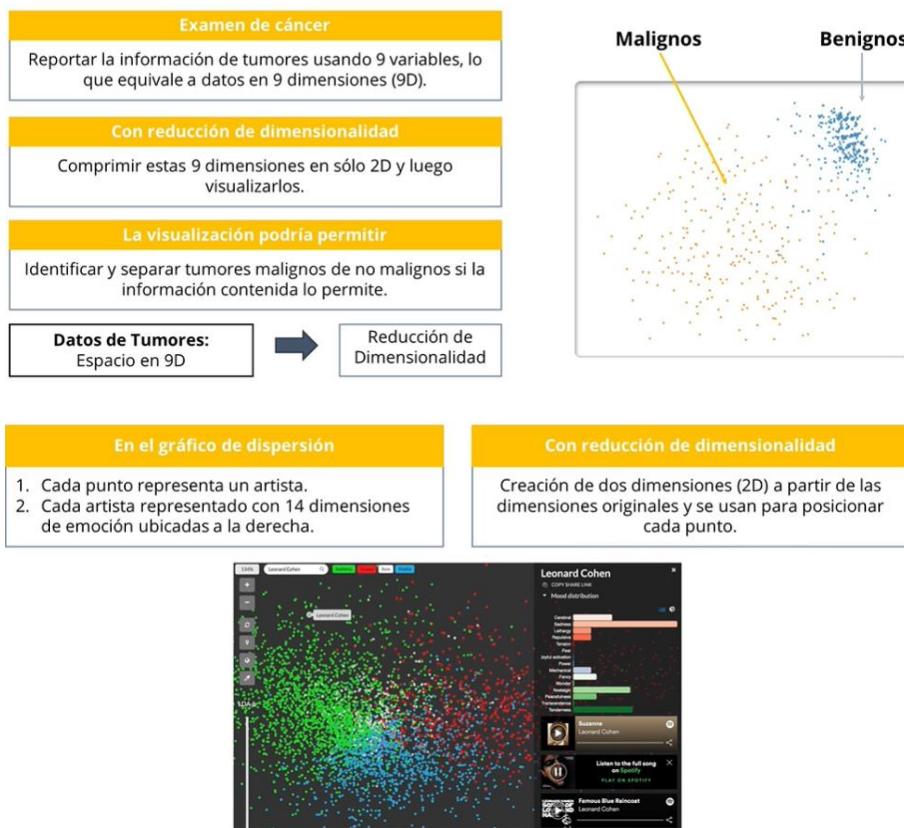


- g. Los mapas de calor son un caso especial de matrices de gráficos.
- h. Los mapas de calor son especialmente útiles para codificar distribución y comportamiento.
- i. Seaborn permite realizar mapas de calor.

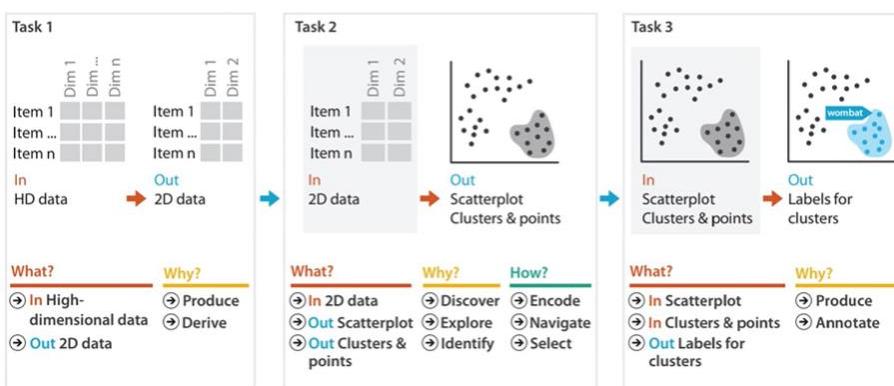
20. Reducción de dimensionalidad lineal

- a. Los datos tabulares tienen una gran cantidad de atributos (visualización y análisis).
- b. Un dataset que describe información de personas con decenas o cientos de atributos.
- c. Datos como imágenes o texto: se pueden describir con miles de dimensiones.
- d. ¿Podríamos representar estos datos en un gráfico de dos dimensiones?
- e. ¿Cómo podemos hacer uso de reducción de dimensionalidad con un método lineal y su correspondiente visualización en Python?

- f. La reducción de dimensionalidad (**RD**) se puede entender como agregación atributos para obtener una representación comprimida de los datos originales.
- g. La RD es diferente a simplemente seleccionar o filtrar un subconjunto de atributos. La RD implica un proceso de construcción de nuevos atributos a partir de los originales.
- h. Datos que representan con vectores de muchas dimensiones. Son textos, imágenes o videos. La RD permite verlos en 2D o 3D.
- i. Se asocia la reducción de dimensionalidad a términos como factores latentes o variables escondidas.
- j. Ejemplos:



k. Relación con modelo anidado y evaluación:



I. Evaluación de una reducción de dimensionalidad:

- i. Al hacer RD, los datos se transforman de alta a baja dimensión, y en el proceso se pierde información.
- ii. Esta pérdida de información puede producir errores o distorsiones al visualizar los datos en menor dimensión.
- iii. Existencia de diferentes métricas para evaluar cuánta información se ha perdido de la original:
 1. Varianza acumulada: cuando se presenta en los nuevos atributos.
 2. Stress: diferencia entre distancias de los puntos tanto en alta como en baja dimensión, como se ve en la fórmula.

$$\text{stress}(D, \Delta) = \sqrt{\frac{\sum_{ij} (d_{ij} - \delta_{ij})^2}{\sum_{ij} \delta_{ij}^2}}$$

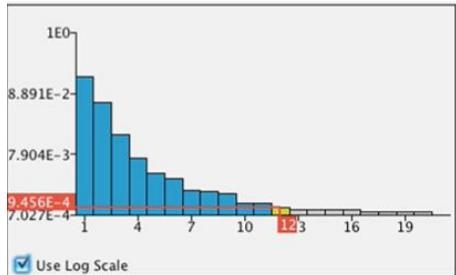
- D : matrix of lowD distances
- Δ : matrix of hiD distances δ_{ij}

m. Evaluación con scree plots:

- i. Permite de forma visual medir cuantas nuevas dimensiones crear al hacer RD.
- ii. Ejemplo:

Ejemplo: Dataset original

Tiene 294 dimensiones y se puede ver que con 20 dimensiones se preserva la mayor cantidad de información (varianza) de los datos originales.



iii. PCA:

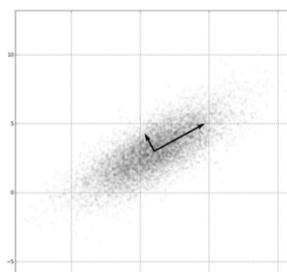
1. La reducción de dimensionalidad lineal:

Diversos métodos de reducción de dimensionalidad

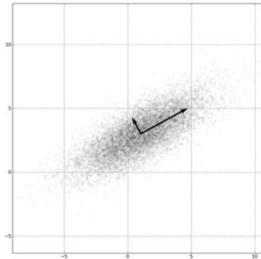
Más popular el llamado PCA (*Principal Component Analysis*).

Método PCA

Crea nuevos atributos (dimensiones, ejes o componentes principales) sobre los cuales se pueden proyectar los datos originales.



Datos originales
Pueden ser representados como una combinación lineal de los nuevos ejes (componentes principales).



Los nuevos ejes se ordenan del que captura más varianza al que captura menos

Basta con elegir los primeros ejes (componentes principales) para reducir la dimensionalidad de los datos.

2. PCA usando técnica de matriz de covarianza

Una formulación típica para calcular PCA es:



- 1 **Normalizar**
Los datos (columnas) y calcular la matriz de covarianza.
- 2 **Calcular**
Valores y vectores propios de la matriz de covarianza.
- 3 **Proyectar**
Los datos originales en la nueva base (vectores propios).

$$C = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^T$$

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}$$

3. PCA usando técnica SVD

Otra forma de realizar reducción de dimensionalidad
Usando PCA es a través de Singular Value Decomposition (SVD).

$$\mathbf{M} = \begin{bmatrix} 1 & v & v & v & s \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$M = U\Sigma V^T$$

Factorización de la matriz de datos original
(donde las columnas son los atributos y las filas los datos)
Usando SVD.

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & \textcolor{red}{0} & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

3 Proyección de la matriz original M
Haciendo algunos valores de la matriz diagonal Σ igual a cero y luego multiplicando $\mathbf{U} * \Sigma$.

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$M = U\Sigma V^T$$

4 Proyección de los datos originales de M multiplicando M^*V
Luego, hacer = 0 las últimas columnas.

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & \textcolor{red}{0} & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

- n. La reducción de dimensionalidad reduce el número de dimensione de los datos originales, permitiendo visualizaciones y análisis de datos difíciles de hacer con los datos originales.
- o. Es posible usar métricas para medir qué tan bien representados están los datos originales luego de usar reducción de dimensionalidad, como el stress y los scree plots.
- p. La técnica PCA es la más utilizada para hacer reducción de dimensionalidad y podemos usarla para visualizar en 2D y 3D datos que originalmente está en muchas más dimensiones.

21. Reducción de dimensionalidad no lineal

a. Ver PDF.

22. ayudantía

Gráfico de barras



Ayudantes del Curso

- El gráfico de barras se usa para presentar valores numéricos asociados a categorías.
- Permite comparación efectiva debido al uso del canal de largo.
- El gráfico de barras apiladas permite agregar un atributo categórico adicional.

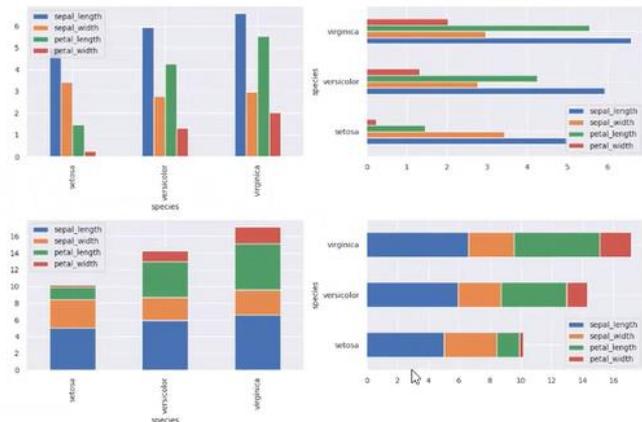


Gráfico de puntos y líneas



Ayudantes del Curso

- Utilizados principalmente para presentar tendencias.
- En el eje X se pone una variable ordinal (típicamente datos temporales).
- El caso del gráfico de puntos sólo usa puntos para codificar el valor de una posición en el eje X.

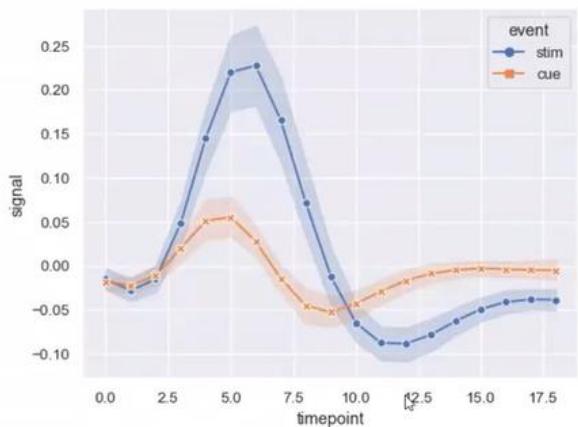


Gráfico de dispersión



- Utilizado para visualizar relaciones entre dos atributos numéricos.
- Se usa la marca de punto y el canal de posición horizontal y vertical (X,Y) para la representación visual de cada marca.
- Permite visualizar fácilmente si es que existe correlación, tanto positiva como negativa.

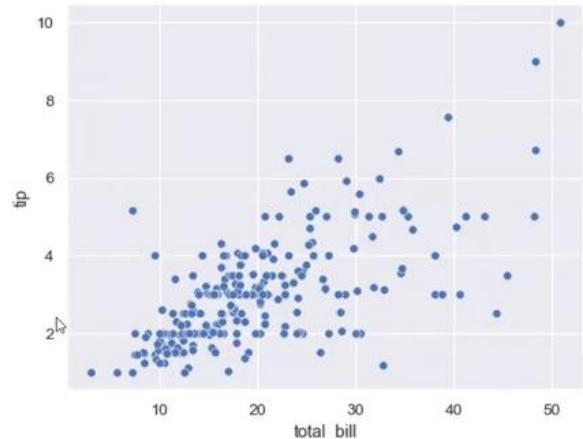


Gráfico de burbuja



- Permite codificar variables categóricas adicionales usando los canales de color y tamaño de los puntos.

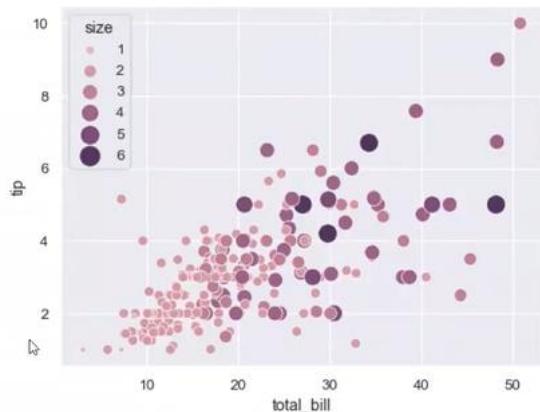
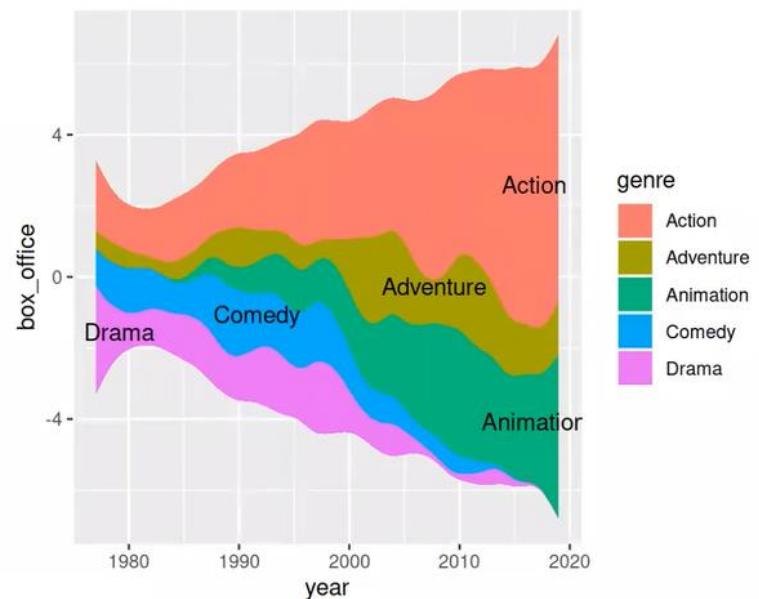


Gráfico de flujos



- Utilizado para visualizar tendencias en el tiempo
- A diferencia del gráfico de líneas, usa el canal de área y de color para presentar la información de tendencias
- Se codifican visualmente tres variables: un atributo temporal (eje X), un atributo numérico (para cuantificar el tamaño de las áreas), y un atributo categórico (para mostrar las distintas áreas, normalmente codificadas con color)



Histograma



- Permite conocer la distribución de los valores de un atributo numérico (cuantitativo y continuo).
- Permite identificar rápidamente los valores más comunes, menos comunes y datos extraños
- Las cantidad de barras, en lugar de corresponder a categorías fijas, corresponden a rangos de valores calculados automáticamente.

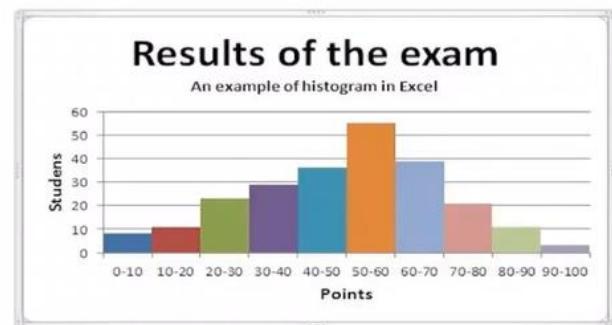


Gráfico de caja



- Permite visualizar la distribución de datos de una variable numérica (cuantitativa continua), enfatizando medidas estadísticas como la mediana, cuartiles y outliers.

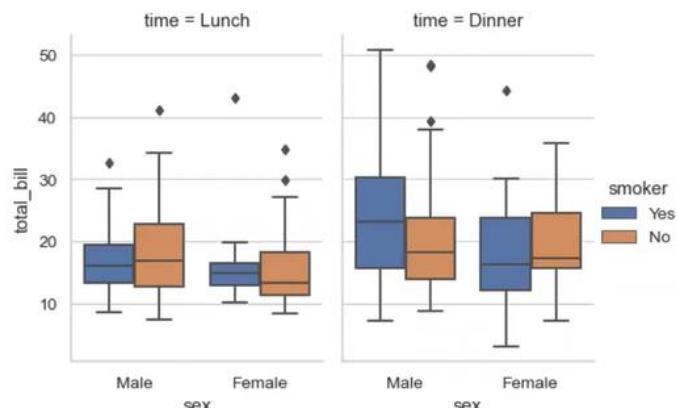
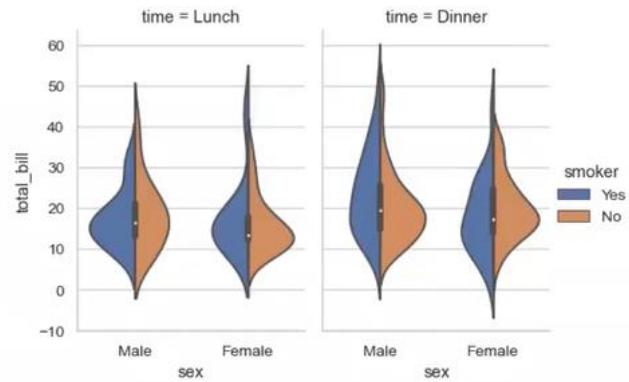


Gráfico de violín



- Presentar información resumida, visualmente, de una distribución de datos.
- Usando un pre-cálculo de la distribución de probabilidad estimada, dibuja una forma suavizada de la distribución directamente desde los datos.
- Pero como la forma suavizada es aproximada, puede contener sesgos si hay muy pocos datos.

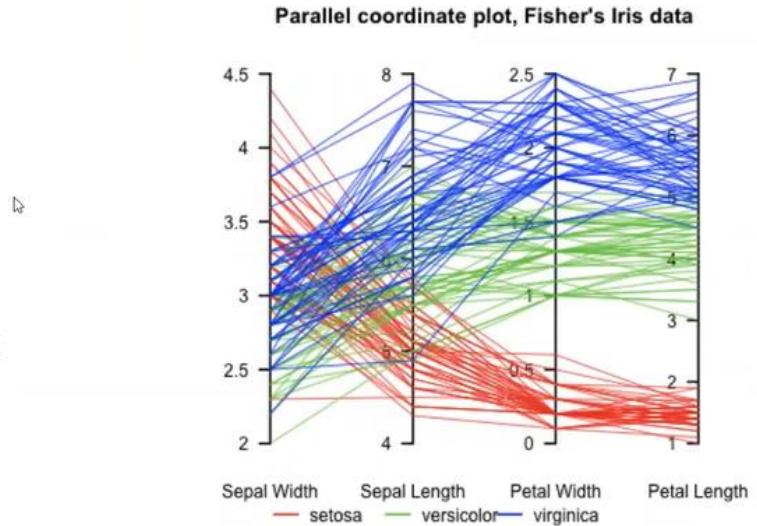


Se utiliza para visualizar la distribución de los datos y su densidad de probabilidad. Este gráfico es una combinación de un diagrama de cajas y bigotes y un diagrama de densidad girado y colocado a cada lado, para mostrar la forma de distribución de los datos.

Gráficos con múltiples ejes



- Utiliza el canal espacial para visualizar toda la información.
- Se utiliza en datasets que presentan múltiples atributos.
- Los atributos se codifican como ejes paralelos mientras que cada ítem del dataset es una línea que cruza 1 vez cada eje.



Gráficos radiales



- Visualización que utiliza una disposición circular y el largo de barra para codificar la información.
- El largo de la barra representa un valor en una escala y los divisores radiales se utilizan para indicar cada categoría.
- Se recurre a esta forma cuando se quiere representar una periodicidad. Por ejemplo, datos referidos a los meses del año.

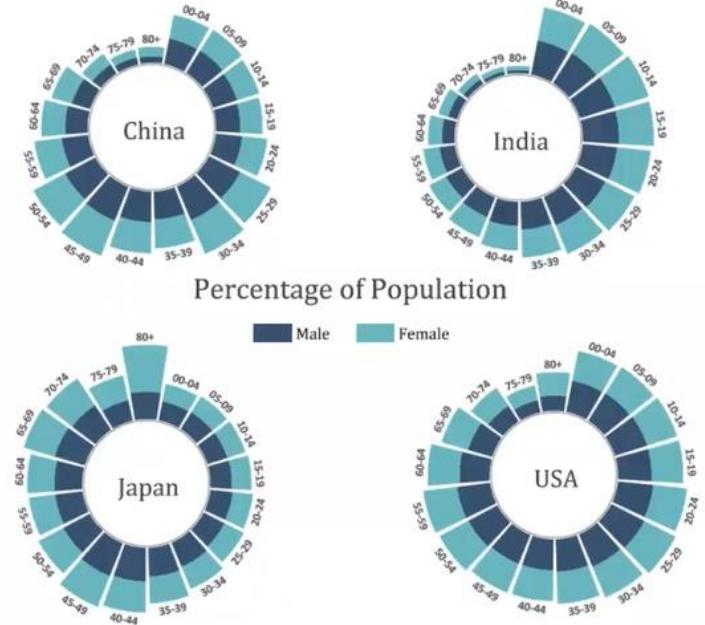
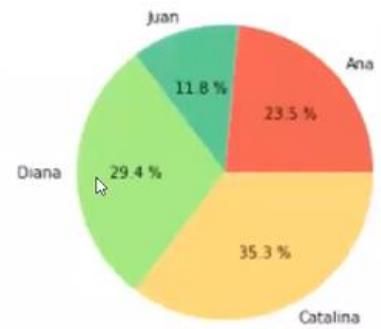


Gráfico de torta

- Se utiliza una disposición circular para visualizar los datos, pero se reemplaza el canal del largo de la barra por el canal de ángulo.
- Cada elemento es codificado por un segmento del círculo y su valor es representado por el ángulo de dicho segmento.
- Se utiliza principalmente para mostrar proporciones dentro de un total.



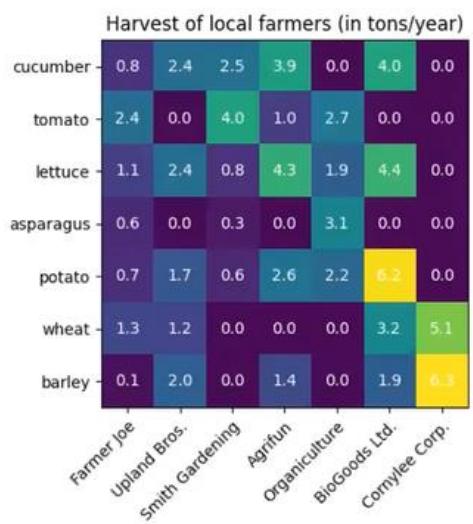
Mapa de calor

Caso específico de alineamiento matricial de dos atributos.

Cada celda de la matriz codifica un tercer atributo.

Atributo cuantitativo secuencial o divergente.

Permite ver la distribución del tercer atributo a lo largo de las combinaciones de las filas y columnas.



23. Conceptos: red, grafos y árboles

a. Introducción:

- i. Los datasets de tipo tabular son los más tradicionales para visualizar, pero los dataset de red también son comunes.
- ii. Diversos dominios, como redes sociales en línea (Facebook, Twitter, Instagram), redes de genes en aplicaciones de bioinformática, redes de transporte o redes de computadores implican visualización de redes.

b. Historia:

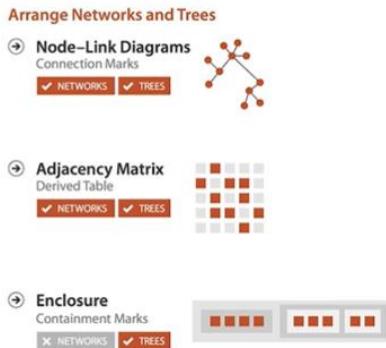
- i. **Easley y Kleinberg (2011)**, indican que una red es un concepto abstracto de: “Cualquier colección de objetos en los cuales algunos pares de esos objetos están conectados”.
- ii. **Ven Euler (1735)**, crea las bases de la teoría de grafos, creando un modelo para representar el problema de “Los 7 puentes de Konigsberg”.
- iii. Un grafo, es una forma particular no abstracta de representar las redes.

c. Relación entre red, grafos y árbol:

- i. Un **grafo**, es una forma de especificar relaciones entre una colección de elementos.
- ii. Un **grafo**, consiste en un conjunto de objetos, llamados nodos, con ciertos pares de estos objetos conectados por líneas llamados enlaces.
- iii. Un **grafo**, es una forma particular de representar visualmente una red.
- iv. Los **gráficos**, son útiles porque sirven como modelos matemáticos de la estructura de red.
- v. Un **árbol**, es un tipo particular de red donde no hay ciclos: hay un nodo raíz, luego ramas y finalmente hojas.

d. Visualización de red según el modelo anidado de Munzner:

- i. Hay tres formas de visualizar redes:
 1. Los diagramas nodo-enlace (grafos).
 2. Las matrices de adyacencia.
 3. Los gráficos tipo encierro.



ii. Tareas típicas al visualizar redes son:

1. Identificar la estructura de la red.
2. Identificar senderos entre pares de nodos.
3. Identificar grupos similares o cercanos de nodos (clusters o comunidades).
4. Presentar nodos importantes según su centralidad (número de enlaces de los nodos, page-rank o centralidad del tipo betweennes).

iii. Formatos de datos para datasets de red:

1. Existen diversos formatos de datasets para redes:
 - a. Distintos formatos tienen niveles diferentes de información para codificar.
 - b. Cuanta información se codifique, tendrá luego relación con los canales utilizados para visualizar.

	Edge List/Matrix Structure	XML Structure	Edge Weight	Attributes	Visualization Attributes	Attribute Default Value	Hierarchical Graphs	Dynamics
CSV								
DL Ucinet								
DOT Graphviz								
GDF								
GEXF								
GML								
GraphML								
NET Pajek								
TLP Tulip								
VNA Netdraw								
Spreadsheet*								

iv. Ejemplo, Formato GDF:

1. Permite incluir más información que el formato CSV.
2. Definición lista de nodos con **nodedef>** y a cada nodo se le puede indicar mucha información.
3. Se indica la lista de enlace con **edgedef>** y para cada enlace se indica mucha información.

```

nodedef>name VARCHAR,label VARCHAR
s1,Site number 1
s2,Site number 2
s3,Site number 3
edgedef>node1 VARCHAR,node2 VARCHAR, weight DOUBLE
s1,s2,1.2341
s2,s3,0.453
s3,s2, 2.34
s3,s1, 0.871

```

v. Ejemplo, Formato GraphML:

1. Permite incluir tanto información como el formato GDF, pero en base al formato XML.

```

<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
    http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <graph id="G" edgedefault="undirected">
    <node id="n0"/>
    <node id="n1"/>
    <edge id="e1" source="n0" target="n1"/>
  </graph>
</graphml>

```

e. Síntesis:

- i. Las redes son un tipo de dataset que requiere tareas y codificaciones visuales a las de datasets tabulares.
- ii. Mientras que red es un concepto general y abstracto, grafo es un modelo para representar y visualizar redes.
- iii. Existen diferentes formas de visualización de redes, así como tareas de visualización.
- iv. Hay diferentes formatos para representar redes que permiten codificar diversos niveles de información.

24. Diagramas nodo-enlace

a. Introducción:

- i. La forma más tradicional de visualizar redes es a través de grafos Munzner (2014) los llama diagramas nodo-enlace.
- ii. Los diagramas de nodo-enlace son fáciles de entender y permiten explorar varios canales para poder visualizar datos de redes.
- iii. A medida que crece el tamaño de la red (en número de nodos y de enlaces), se hace complejo dibujar estos diagramas.
- iv. Hay diferentes algoritmos para abordar este problema, pero es complejo y es un tema que sigue en investigación.

b. Marcas y canales en diagramas nodo-enlace:

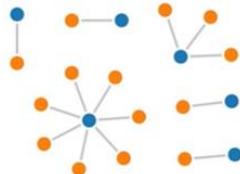
- i. Los diagramas nodo-enlace representan visualmente redes.
- ii. Usa dos tipos de marcas:
 1. **Marcas de puntos**: para los nodos.
 2. **Marcas de líneas**: para mostrar relaciones entre pares de nodos.



3. Codificar visualmente información de diferentes tipos de nodos (tamaño, color, posición).
4. Los enlaces, codifica información a través de canales (ancho, color, tipo de línea).
5. También puede usarse una flecha en lugar de línea para identificar tipos de relaciones.

c. Tareas en diagramas nodo-enlace:

- i. Efectivo en tareas relacionadas con entender la topología de una red (conexiones directas o indirectas entre nodos).

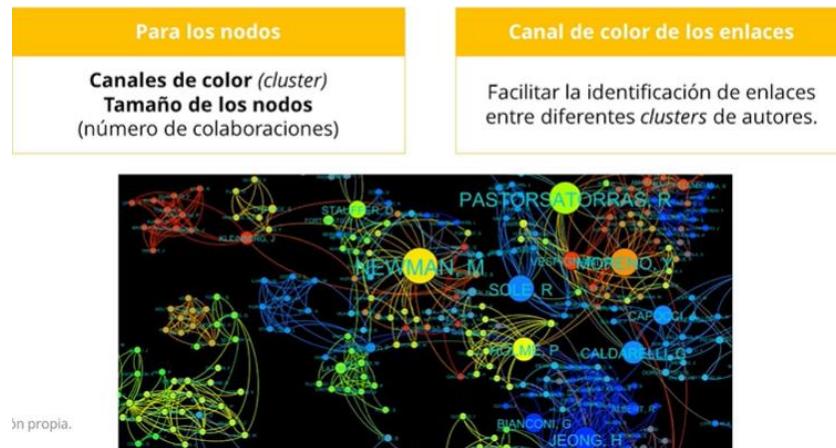


ii. Ejemplos:

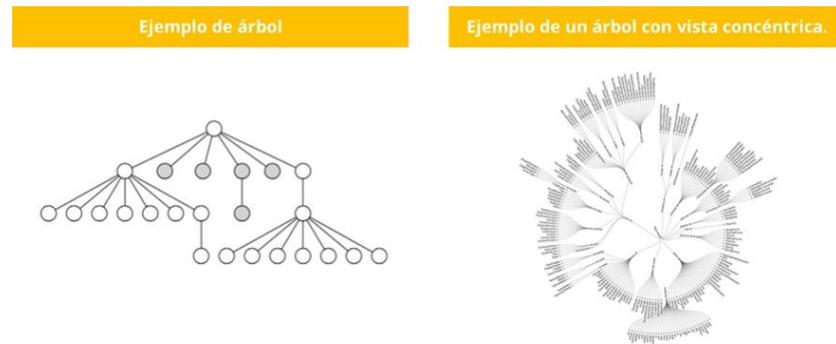
Ejemplos de tareas
Encontrar
<ul style="list-style-type: none">▪ Todos los caminos entre dos nodos.▪ Todos los nodos adyacentes en 1 salto.▪ Los nodos puente de una red.

Número de saltos (<i>hops</i>) en una ruta entre dos nodos
<p>Es un métrica de distancia discreta (no continua, como en un plano 2D)</p>

iii. Ejemplo 1:



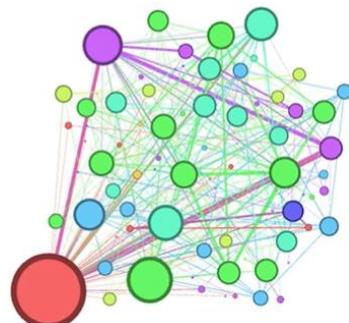
iv. Ejemplo 2:



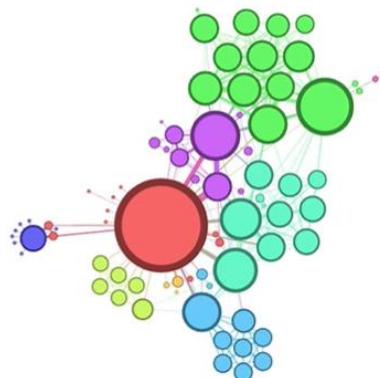
d. Algoritmos de posicionamiento basado en fuerza:

i. Algoritmos para dibujar diagramas nodo-enlace:

1. Se presentan dibujados en 2 dimensiones.
2. El algoritmo “Posicionamiento basados en fuerza” permite dibujar automáticamente los diagramas nodo enlace.
3. El posicionamiento basado en fuerza se conoce también como Algoritmos de minimización de energía, embeddings de resorte u optimización no líneal.



4. Simulando fuerzas físicas que empujan los nodos entre sí, mientras los enlaces actúan como resortes para acercar sus nodos de comienzo a fin.
5. Estos algoritmos parten ubicando los nodos de forma aleatoria dentro de una región espacial. Refina sus posiciones de acuerdo con simulación de fuerzas.
6. El algoritmo intenta disminuir cruce de enlaces y sobreposición de nodos.



7. Primera versión del algoritmo por Peter Eades (1984), "A Heuristic for Graph Drawing. Congressus Numerantium, 42, 149-160".
8. La idea principal del punto anterior, "Dado un grafo, reemplazamos los vértices por aros de acero y los enlaces por resortes para formar un sistema mecánico".
- 9. Ventaja**, relativamente fácil de implementar.
- 10. Ventaja**, fácil de explicar la intuición detrás del algoritmo.
- 11. Limitaciones**, posicionamiento no es determinístico (no siempre permite explotar memoria espacial).
- 12. Limitaciones**, escalabilidad (computacional y visualmente (bolas de pelo o hairballs)).

e. Límites del diagrama nodo-enlace:

- i. Cuando el número de nodos y enlaces crece demasiado, el diagrama de nodo-enlace se vuelve ilegible (llamado, bola de pelos o hairball).
- ii. Según Munzner, idealmente usar como máximo de decenas a cientos de nodos y del orden de cientos enlaces.
- iii. Una densidad nodo/enlace indica que debe haber un número de enlaces menor de cuatro veces el número de nodos.

f. Alternativa: SFDP (Scalable Force-Directed Placement – Alternativa de escalamiento):

- i. Permite mostrar miles de nodos y decenas de miles de enlaces dejando de mostrar algunos enlaces y nodos pocos importantes.

- ii. Presenta problemas de escalamiento, diagrama nodo-enlace con 26 mil nodos a la derecha. Un poco más de 100 mil enlaces, donde es difícil ver estructura de red.

g. Alternativa II: Hive Plots

- i. Fija las posiciones de los nodos en ejes y así lo hace determinístico (posiciones de los nodos no cambian al correr de nuevo el algoritmo).

h. Síntesis:

- i. El diagrama nodo-enlace es el más popular para presentar redes visualmente.
- ii. Para dibujar redes con este tipo de visualización, se utilizan algoritmos, la gran mayoría pertenecen a la familia “posicionamiento basado en fuerza”.
- iii. El diagrama nodo-enlace es muy útil y fácil de entender, sin embargo, cuando la cantidad de nodos y enlaces crecen mucho, se vuelve ilegible.

25. Matrices de adyacencia:

- a. Un conjunto de datos tipo red o grafos puede ser representado por una matriz de adyacencia.
- b. Es un caso específico de **alineamiento** matricial de datos (dos atributos). Tanto en las filas como en las columnas se suelen aplicar nodos en la red (red cuadrada), mientras que en las celdas que son las coincidencias entre filas y columnas representan la posible coincidencia entre el par de nodos.
- c. Al colorear la celda se representa la existencia de conexión en la red.
- d. El ser el dataset datos tabular será necesario realizar procesamientos de los datos para alcanzar un estándar tabular de ellos, donde cada ítem represente una conexión en la red.
- e. El uso de color en las celdas permite codificar datos cuantitativos en las conexiones, pero es limitado al número de pixeles utilizados en cada celda.
- f. Este método es eficiente en espacio para mostrar la completitud de toda la red en un espacio bidimensional.

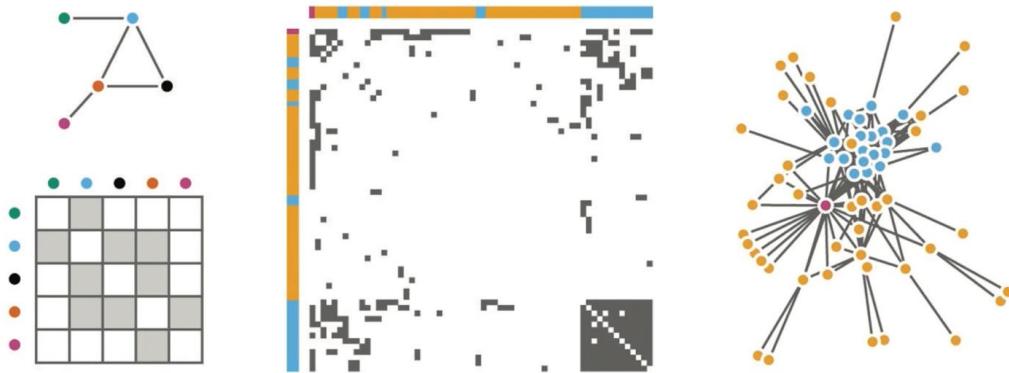
Caso específico de alineamiento matricial de dos atributos.

Matriz cuadrada, cada celda codifica una conexión.

Es posible sea necesario transformar dataset.

Permite ver la distribución de conexiones e identificar clusters.

Eficiente en espacio.

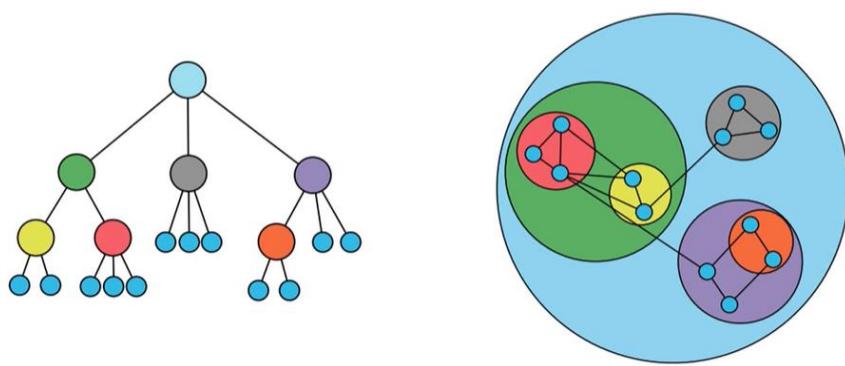


26. Gráficos tipo encierro

- a. Gráfico alternativo a gráfico node-enlace cuando el dataset es un árbol.
- b. Utilizado para mostrar una información más completa de la estructura jerárquica.
- c. Todos los nodos de un grupo quedan encerrados dentro de un área mayor generando una disposición anidada.
- d. Si los nodos poseen un atributo numérico, generalmente, se utiliza el área como canal para visualizar dicho atributo.

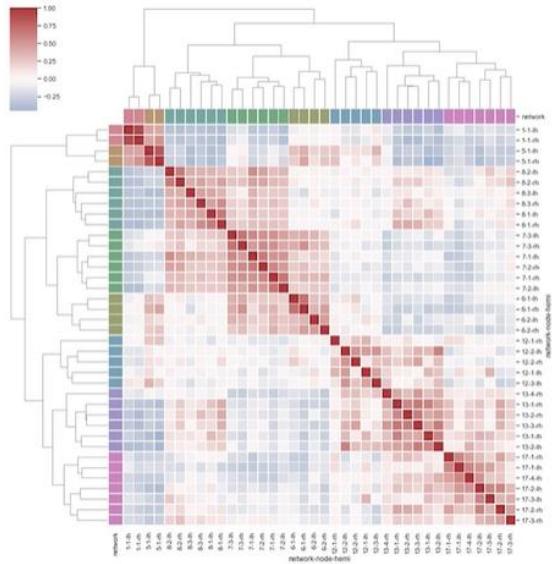


- e. No solo pueden ser gráficos cuadrados (Treemaps), pueden ser de otra forma (GrouseFlocks).



27. Clustermap

- De las distintas maneras de codificar visualmente redes, algunas tienen ciertas ventajas sobre escalamiento (matriz de adyacencia) y otras sobre familiaridad para el usuario (diagrama nodo-enlace).
- Una opción interesante es una visualización híbrida que combina las fortalezas de diferentes visualizaciones: este es un clustermap.
- Motivación:
 - Si tenemos una red donde los enlaces tienen asociados atributos de peso, es posible extender la matriz de adyacencia usando un mapa de calor (colores en las celdas para representar fortaleza de enlaces).
 - Si al mapa de calor le agregamos la estructura jerárquica de la red, considerando que algunos grupos de nodos forman comunidades (clusters), podremos visualizar un clustermap.
- Ejemplo:
 - Una matriz: filas y columnas representan los nodos.
 - Colores dentro de una matriz: representan los enlaces según su fortaleza.
 - Dendograma externo de la matriz: permite visualizar estructura jerárquica en los enlaces de la red.

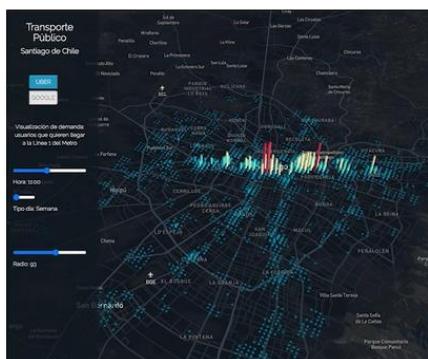


e. Marcas y canales:

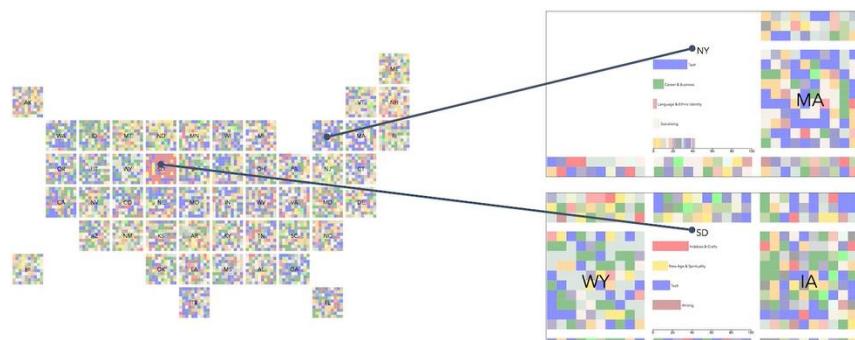
- i. Marcas principales: marcas de área en un alineamiento matricial 2D (matriz de calor).
- ii. Canales principales: color y posición.
- iii. Marcas del dendograma: indica la estructura de clusters o jerarquía de nodos evidenciada gracias a la visualización.

28. Visualizaciones en dominios de aplicación

- a. Distintos dominios de aplicación y ejemplos de visualización para ciertos casos: visualización espacio-temporal, visualización de texto y visualización para aplicaciones de inteligencia artificial.
- b. Un estilo de aplicación donde la visualización de información se utiliza para contar historias: visual story telling.
- c. Visualización en dominio espacio-temporal:
 - i. Ejemplo, el Transporte público en Santiago de Chile, combina visualización espacial de uso de estaciones de metro, comparando tráfico entre 11 am vs 6 pm.

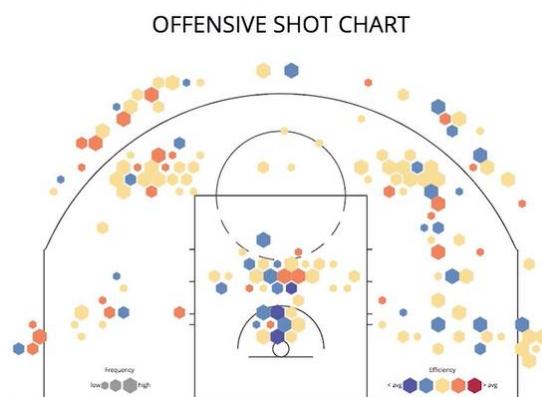


- ii. Ejemplo, Actividades en diferentes estados de EEUU, tipo de visualización temporal usando un GridMap.

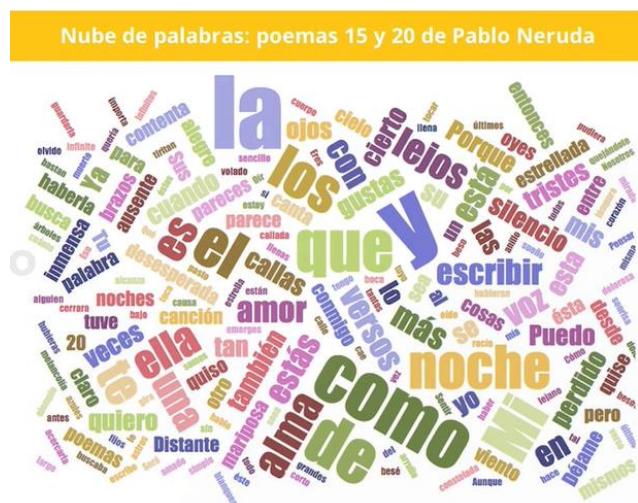


- ### iii. Ejemplo, Deporte: Mapa hexagonal en basketball.

1. Kirk Goldsberry (PhD en Geografía), popularizó el mapa de calor hexagonal para analizar especialmente la efectividad del ataque y defensas en la NBA.

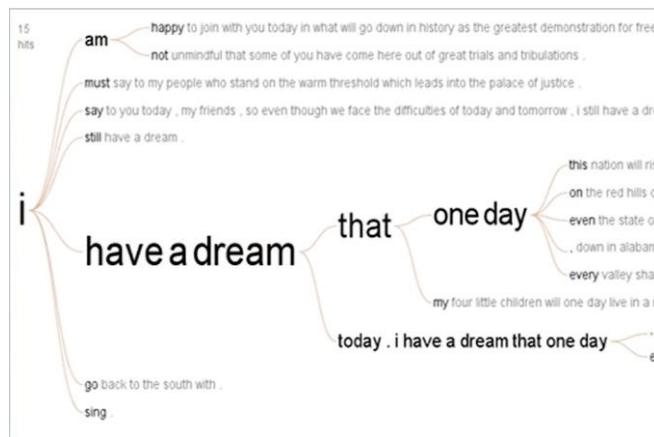


- iv. Ejemplo, Nube de palabras: La forma más tradicional de presentar en forma resumida y visual un documento o colección de texto.

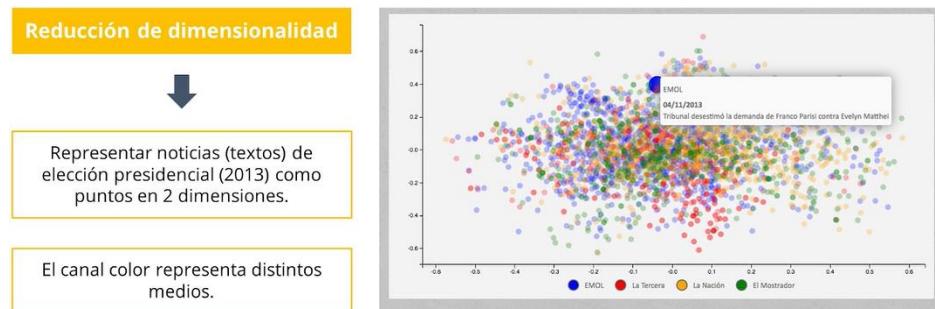


v. Ejemplo, Word Tree, creado por Martin Wattenberg y Fernanda Viegas. Permite:

1. Tomar un texto.
2. Generar un árbol de patrones.

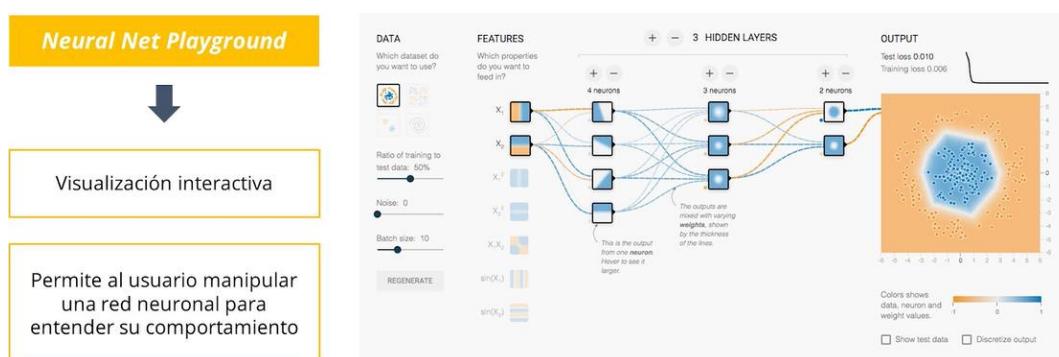


vi. Ejemplo, Texto y reducción de dimensionalidad,



d. Visualizaciones para aplicaciones de inteligencia artificial:

i. Entendiendo las redes neuronales:



ii. Conectando visualización espacio-temporal y predicciones

Seebacher y co-autores (2018)

Desarrollo aplicación interactiva
Muestra tanto el aspecto espacio-temporal como las predicciones.

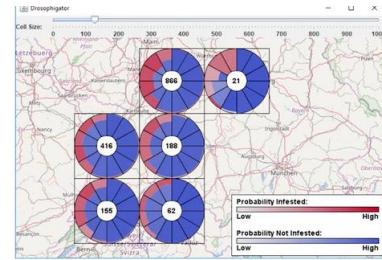
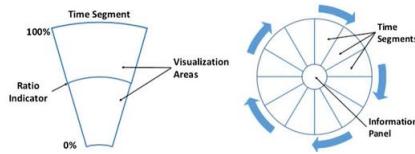
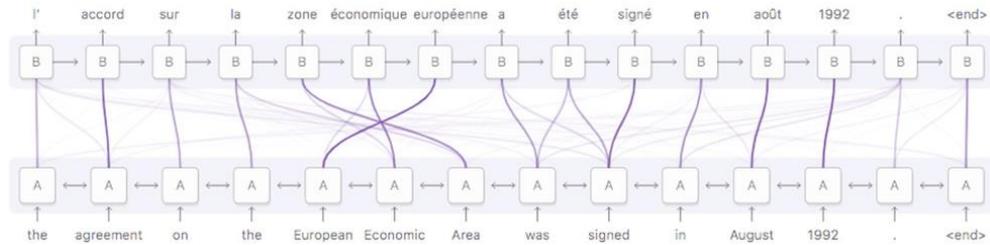


Figure 2: An overview of the *Drosophigator* application for the visual analysis of spatio-temporal event predictions.

III. Como funciona un sistema de traducción con IA:



e. Contando historias con visualización (Visual Story Telling)

i. New York Times: Opinión de Republicaciones sobre Clima

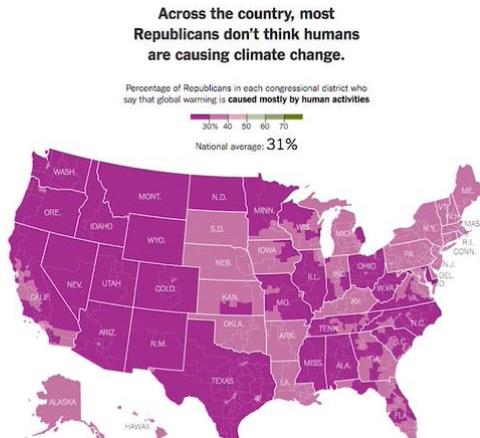
**Reportaje del New York Times
(Diciembre de 2017)**

↓

Reportaje enriquecido visualmente

Mapa
con el porcentaje de republicanos en cada "county".

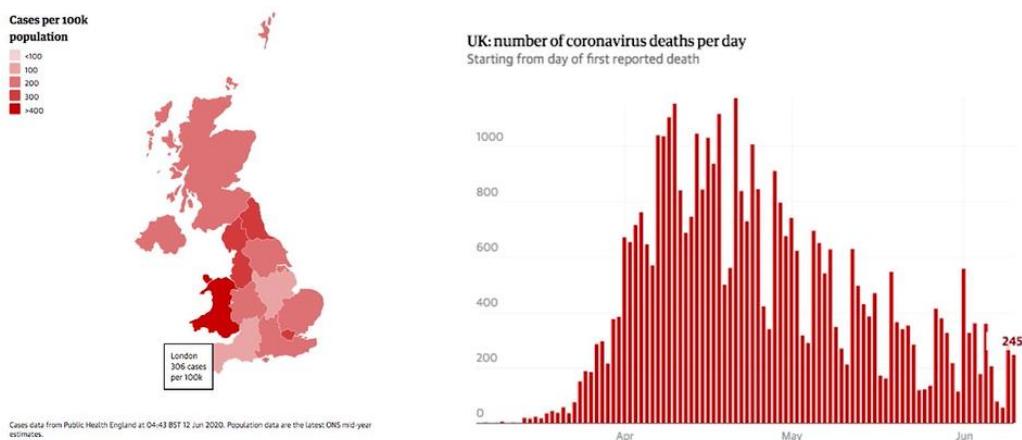
Creencia
Calentamiento global causado por humanos.



ii. La tercera: reportaje sobre maltrato infantil



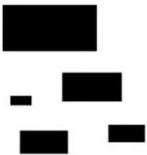
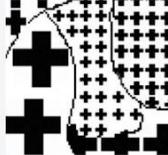
iii. Evolución del coronavirus en Inglaterra



- f. La visualización de información se usa intensivamente en diferentes dominios de aplicación.
- g. Ejemplos de dichas aplicaciones con énfasis en espacio-temporalidad, visualización de texto y de aplicaciones de inteligencia artificial.
- h. El uso de visualización para contar historias: visual story telling.

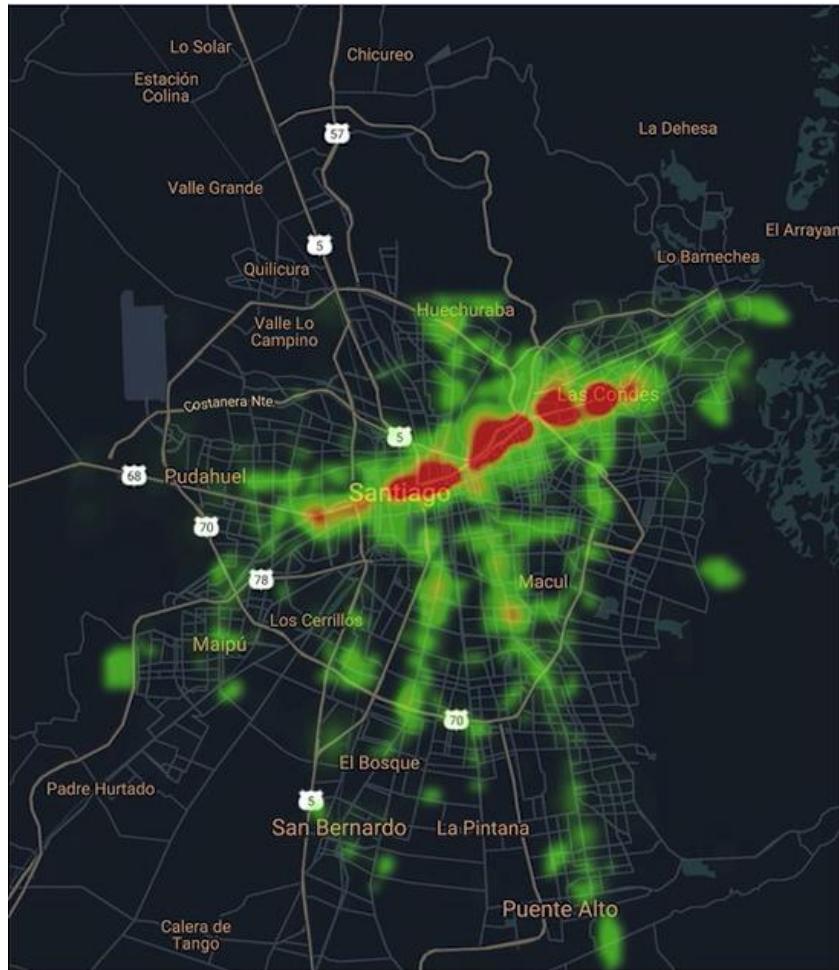
29. Visualización espacial y espacio-temporal

- a. La visualización de datos espaciales tiene como característica inherente el describir objetos y fenómenos con una ubicación específica en el mundo por medio de:
 - i. Mapas y sus proyecciones.
 - ii. Tipos de datos a representar.
 - iii. Gráficos específicos de acuerdo con la información a presentar.
- b. Visualización geográfica:

Punto	Línea	Superficie
<p>0 dimensiones Latitud / longitud</p> 	<p>1 dimensión Pares de latitud/longitud</p> 	<p>2 dimensiones Región por puntos con atributos para cada fenómeno</p> 

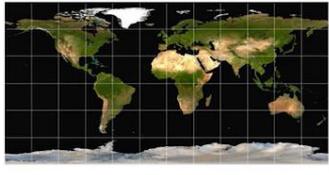
c. ¿Qué son los mapas?

- i. Es un mundo reducido a un conjunto de puntos, líneas y áreas, definidos por su posición en un sistema de coordenadas.
- ii. Es un conjunto de fenómenos espaciales y sus relaciones en el contexto del sistema de coordenadas definido.

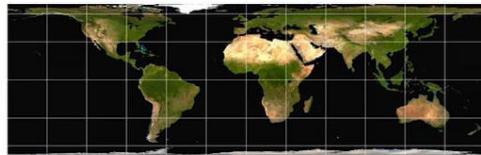


d. Proyección cartográfica:

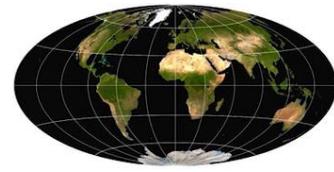
- i. Mapeo de posición en el globo (esfera) a posiciones de pantalla (superficie plana).
- ii. Software gratuito G Projector de la NASA, explora una gran variedad de estas proyecciones.



Equirectangular



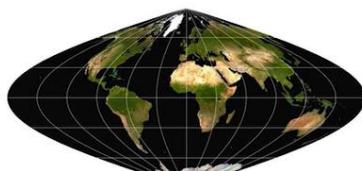
Lambert cylindrical



Hammer-Aitoff



Mollweide



Cosinusoidal

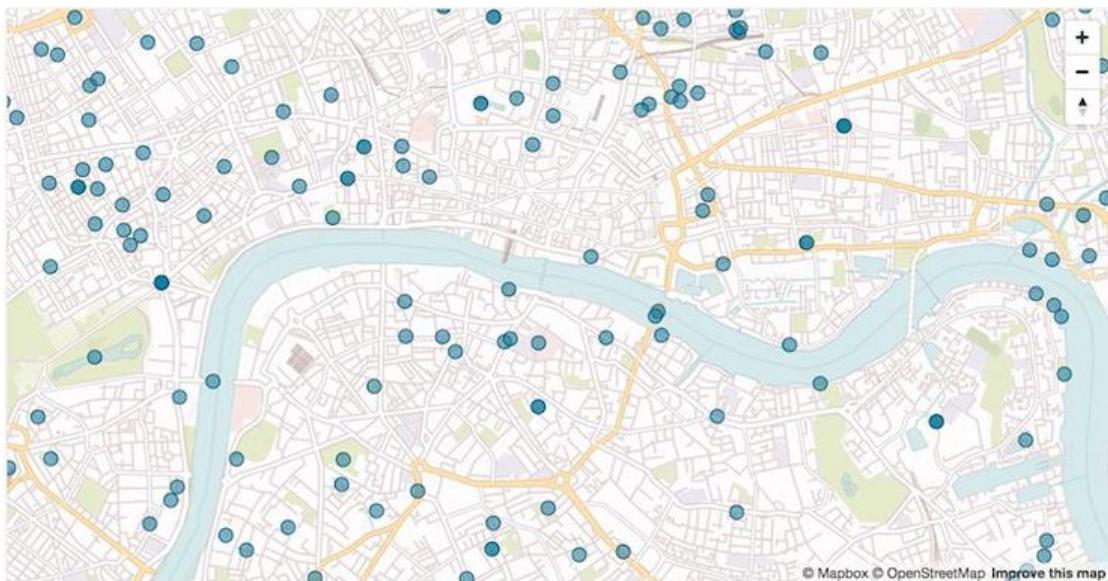


Albers equal-area conic

e. Tipos de gráficos más comunes:

i. Mapa de puntos:

1. Permite visualizar fenómenos puntuales colocando un símbolo o píxel donde se produce el fenómeno.
2. Símbolos o marca (Círculo, barras, cuadrados, etc.), el valor está codificado por tamaño o color.

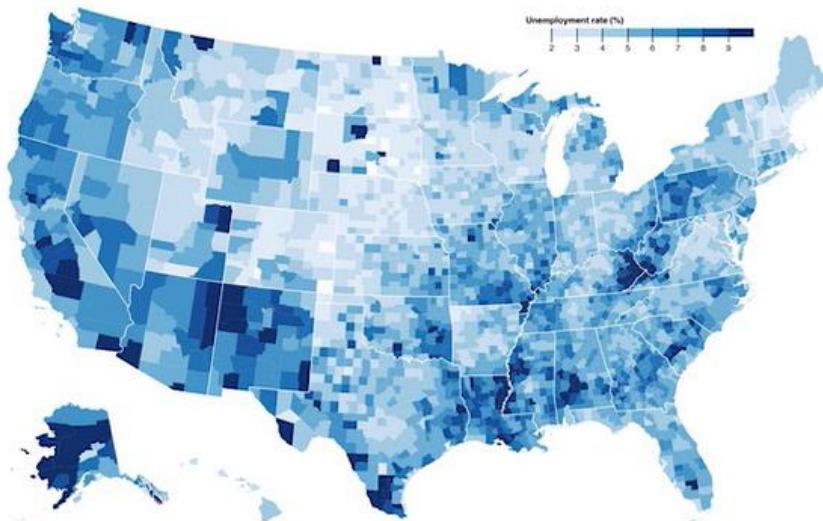


3. Ventajas (Gran facilidad de compresión e ilustran la variación de una cierta densidad espacial).
4. Desventajas (Su relación lleva tiempo, es necesario estudiar los factores que controlen la distribución de la variable en el mapa, adquirir información, etc.)

ii. Mapa de coropleta:

1. Muestra áreas geográficas divididas o regiones coloreadas en relación con una variable numérica.

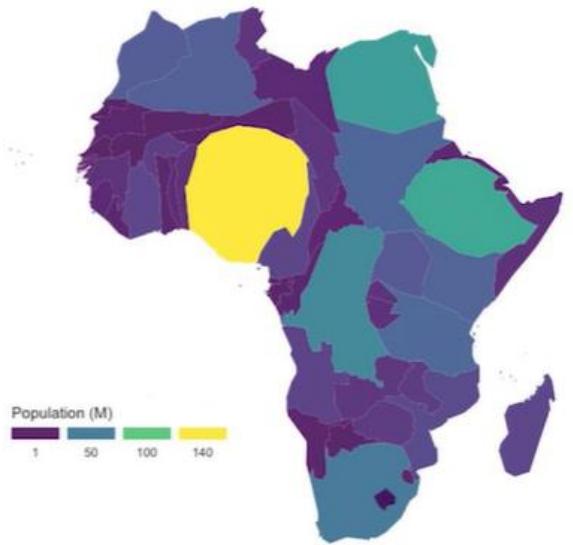
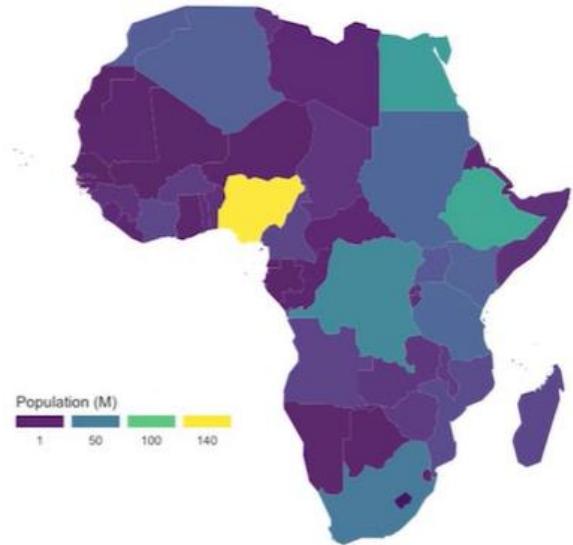
2. Ventajas (permite estudiar cómo evoluciona una variable a lo largo de un territorio).
3. Desventaja (las regiones con tamaños más grandes tienden a tener un mayor peso en la interpretación del mapa, que incluye un sesgo).



Tasa de desempleo el 2006 en cada "county" de EEUU.

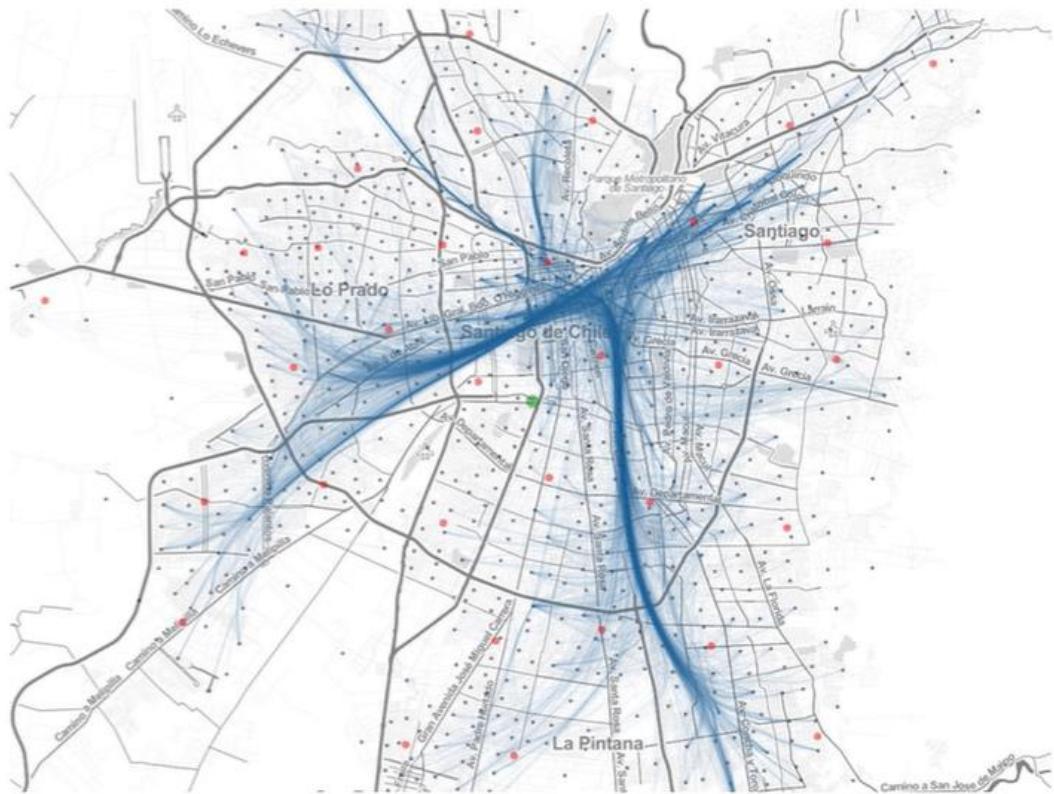
iii. Cartograma

1. Tipo específico de transformación de mapa, donde las regiones se redimensionan de acuerdo con una variable relacionada geográficamente.
2. Símbolos o marca (las regiones se colorean de acuerdo con una variable numérica).
3. Ventaja (evitan el problema de los mapas coropléticos a través de la distorsión).
4. Desventajas (distorsiona los límites reales, y por lo tanto, hace que el mapa sea más difícil de identificar).

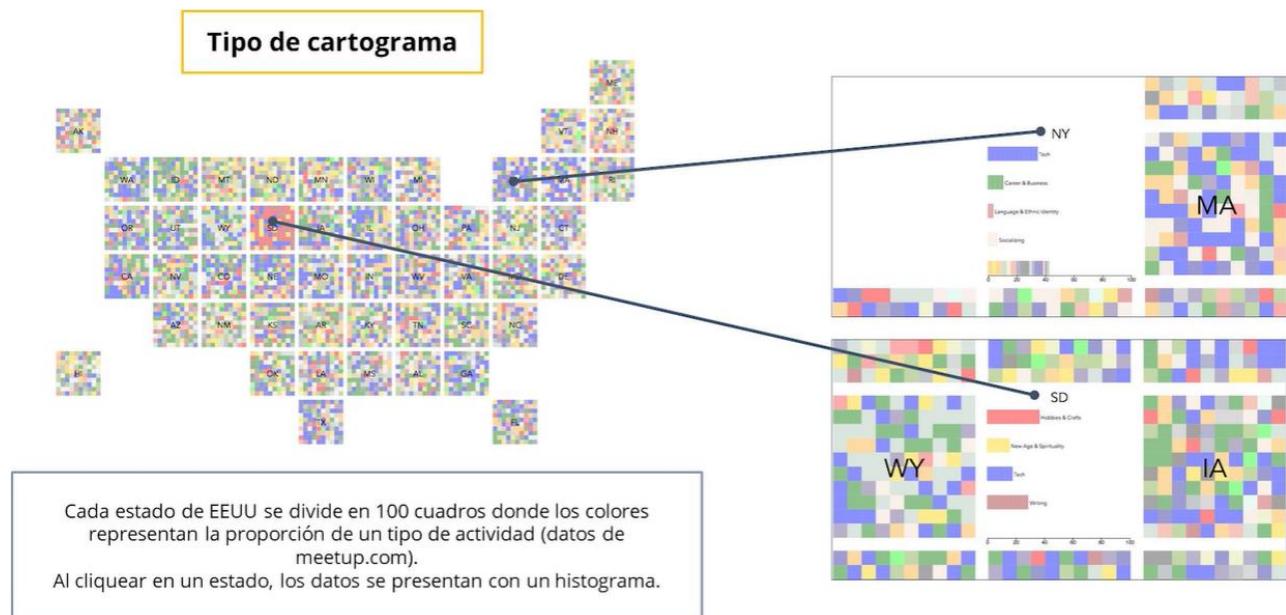


iv. Mapas de flujos

1. Muestra movimientos lineales. Se usan líneas en forma de flecha indicando dirección y sentido del flujo.
2. Símbolos o marcas (se usan líneas para representar, A) Qué tipo de movimiento es el que da., B) Qué cantidad de movimiento se está dando.
3. El ancho de las líneas es proporcional a las cantidades que representan.



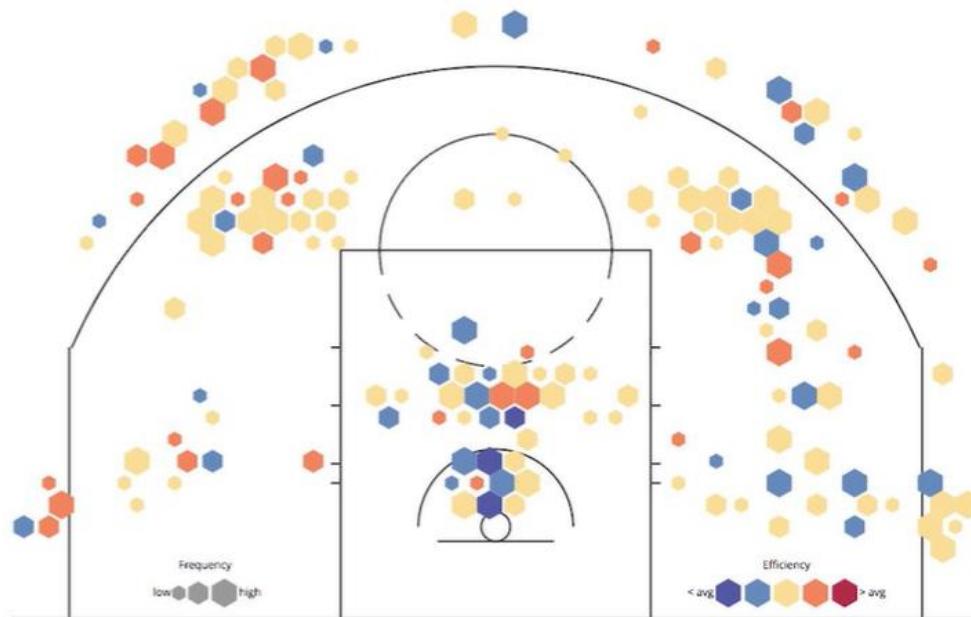
v. GridMap



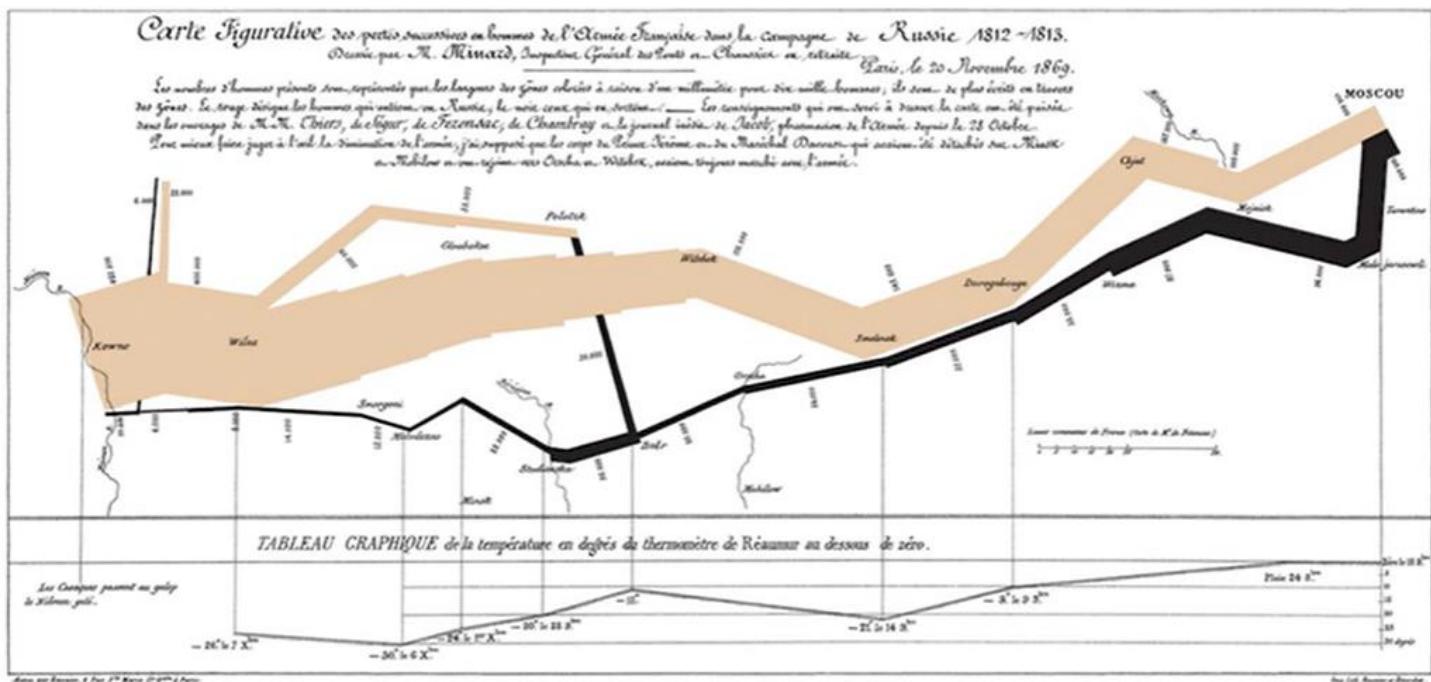
vi. Mapa hexagonal

1. Permite crear divisiones sobre un mapa y mostrar activaciones sin distorsionar el mapa como lo haría un cartograma.

OFFENSIVE SHOT CHART



vii. Mapa de flujo (Invasión de Napoleón)



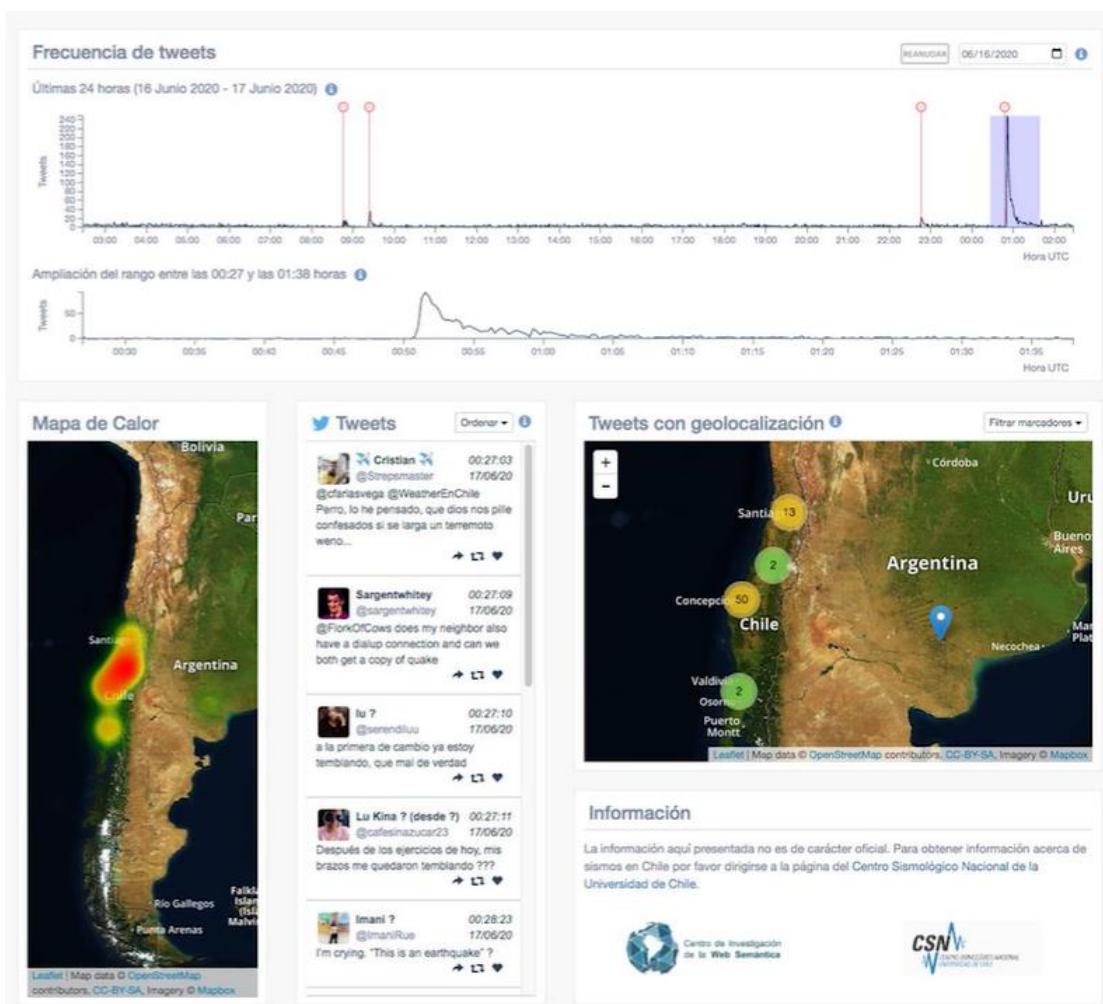
viii. Transporte público en Santiago de Chile

1. Visualización espacio-temporal de uso de estaciones de metro, comparando tráfico 11am vs 6pm.



ix. Twicalli

1. Sistema para monitoreo de la percepción de sismos a partir de mensajes en las redes sociales.
2. Dashboard con una serie de gráficos, se complementan para mostrar información espacio temporal.



30. Visualización de colecciones de documentos

- a. El texto se define como un tipo de dato no estructurado: no tiene la estructura tabular de filas y columnas que facilita la elección de visualizaciones tradicionales.
- b. El primer paso en visualizar texto y documentos es la elección de cómo representarlo.
- c. La estructura sintáctica y semántica del texto hace también difícil que puedan utilizarse gráficos tradicionales.
- d. Ejemplo de creación de nube de palabras usando Python.
- e. Representación “**Bolsa de Palabras**”:
 - i. **Un supuesto:** Las palabras son importantes, pero su orden podemos obviarlo.
 - ii. **“Bolsa de palabras”:** Permite reducir la complejidad asociada a representar texto, y por ende, a visualizarlo.
 - iii. Modelo de espacio vectorial para documentos.
- f. De la bolsa de palabras a matrices y vectores:

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
...						

g. La matriz término-documento

Documento 1: Un auto rojo										ID	palabra
Documento 2: Un tomate rojo y un globo rojo.										1	amarillo
Documento 3: Un plátano amarillo y un tomate verde.										2	auto
1	2	3	4	5	6	7	8	9		3	globo
Doc. 1	0	1	0	0	1	0	1	0	0	4	plátano
Doc. 2	0	0	1	0	2	1	2	0	1	5	rojo
Doc. 3	0	0	0	1	0	1	2	1	1	6	tomate
										7	un
										8	verde
										9	y

h. Pesos dentro de la matriz:

TF: Term Frequency

Es la frecuencia de un término en el documento, pero normalmente se le aplica logaritmo para ajustar su importancia:

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $tf_{t,d} \rightarrow w_{t,d}$: $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4$, etc.

IDF: Inverse Document Frequency

Se refiere a darle más importancia a términos que aparecen en pocos documentos, como "paralelepípedo", porque son más informativos del contenido.

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Donde N es la cantidad de documentos en el **corpus** y $|\{d \in D : t \in d\}|$ es la cantidad de documentos en los que aparece la palabra t .

i. TF-IDF aplicado:

- Para representar los documentos multiplicamos la frecuencia de cada palabra TF por el peso IDF.

Documento 1: Un auto rojo

Documento 2: Un tomate rojo y un globo rojo.

Documento 3: Un plátano amarillo y un tomate verde.

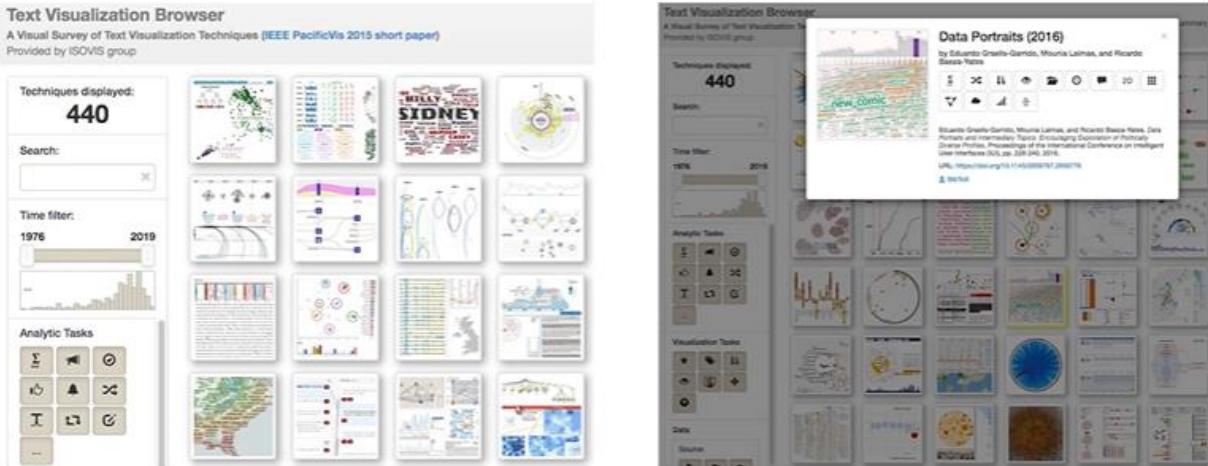
	1	2	3	4	5	6	7	8	9
Doc. 1	0	0,48	0	0	0.17	0	0	0	0
Doc. 2	0	0	0	0	0.34	0	0	0	0
Doc. 3	0,48	0	0	0,48	0	0.17	0	0,48	0,17

ID	palabra	idf
1	amarillo	0,48
2	auto	0,48
3	globo	0,48
4	plátano	0,48
5	rojo	0,17
6	tomate	0,17
7	un	0
8	verde	0,48
9	y	0,17

j. ¿Qué podemos hacer con estos datos y representación?

Text visualization browser del ISOVIS Group

(Linnaeus University, Suecia)



k. Nube de palabras:

- i. La forma visual más popular de representar texto.
 - ii. Herramientas comunes para su generación:
 - 1. Wordle (wordle.net)
 - 2. Jonathan Feinberg y su implementación (2010), capítulo 3 en Beautiful Visualization. Eds. Steele & Uliinsky.



iii. Debilidades de las nubes de palabras:

1. El tamaño de las palabras: No considera el largo de la palabra.
 2. Los colores de palabras: No tienen significado.
 3. La fuente: Es meramente estético.

iv. ¿Cómo mejorar la nube de palabras?

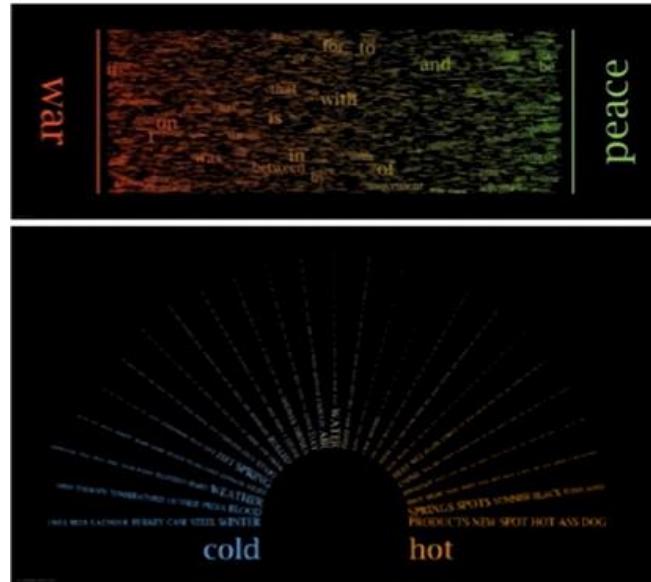


Figure 4 Topical bias. Word clouds for biographies of women (top) and men (bottom), with birth date before 1900 (left) and since 1900 (right). Spaces in bi-grams are replaced with an underscore. Font size is proportional to PMI with each gender. Colors depict the four categories: gender in orange, family in green, relationship in violet, and other in blue. Beside professional and topical areas, words in the gender, relationship, and family categories are more dominant in articles about women born before 1900. Gender-specific differences are much less pronounced in articles about people born since 1900.

1. Agregar interacción (ejemplo, al hacer clic en una palabra se ve su detalle).



2. Usar semánticas de colores y posiciones (Ejemplo, visualización de Chris Harrison).



v. ¿Qué funciona bien en la nube de palabras?:

Resultados

Usar barras alineadas mejora la lectura de valores con precisión

Tamaño de la fuente e intensidad de color funcionan mejor para búsqueda de palabras

1 Barras bajo palabras

**Listas simples de palabras
funcionan bien**

2 Probar otras disposiciones espaciales

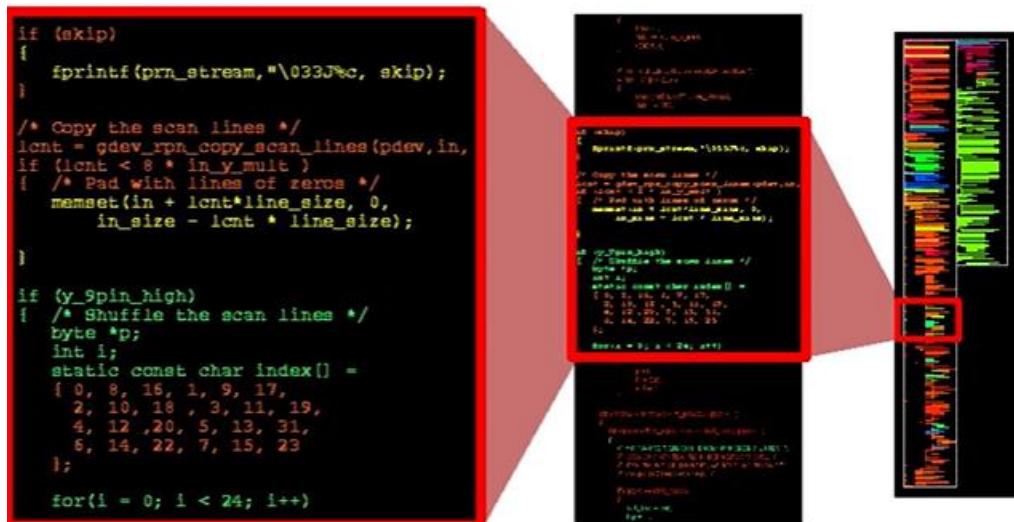
Pensar en ordenar

I. Otra forma de visualizar texto:

i. Distinciones en visualizar de texto:

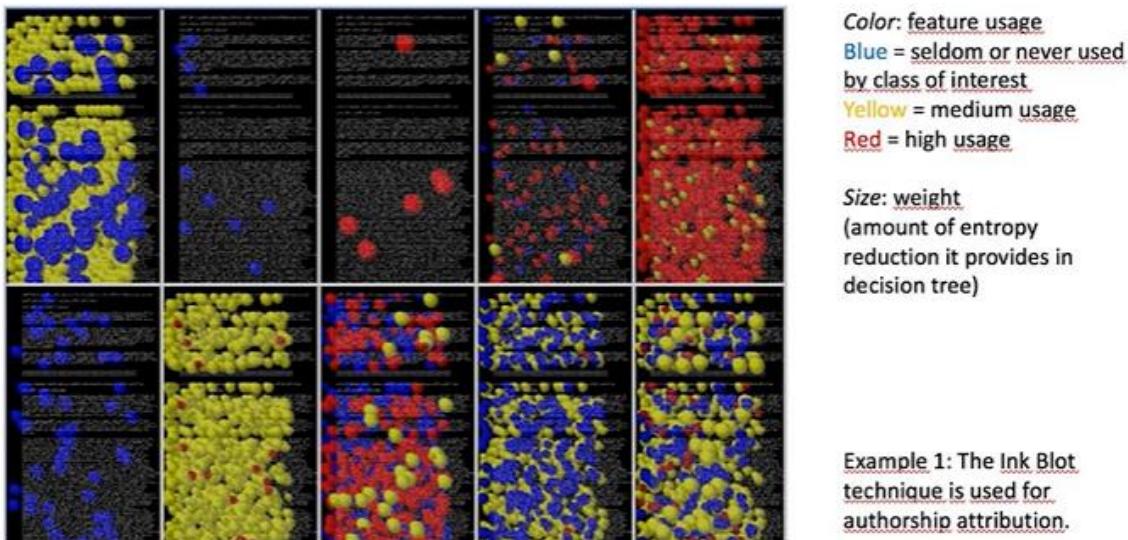
1. Visualización de características (features) del texto.
2. Visualización de la estructura del documento.
3. Visualización del corpus y metadata del documento.

ii. Seesoft:



Color = estadística de interés, en este caso: antiguedad del código

iii. InkBlots:



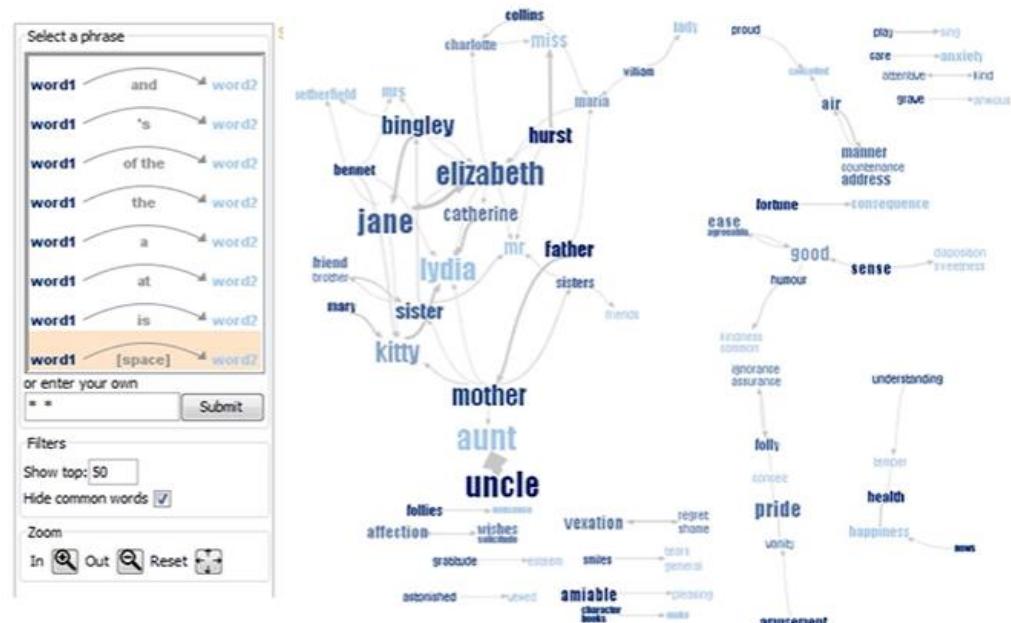
iv. Visualización de la estructura del documento:

1. Visualización de características (features) del texto.
2. **Visualización de la estructura del documento.**
3. Visualización del corpus y metadato del documento.

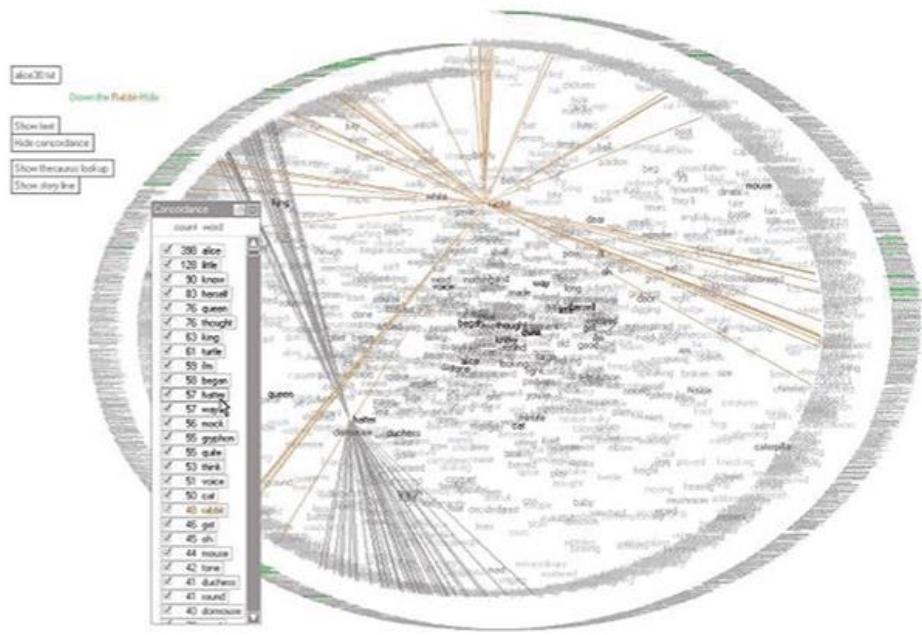
v. WordTree:



vi. Phrase Net:



vii. TextArc:



viii. Visualización del corpus y metadata del documento:

1. Visualización de características (features) del texto.
 2. Visualización de la estructura del documento.
 3. **Visualización del corpus y metadata del documento.**

ix. LDAVis:

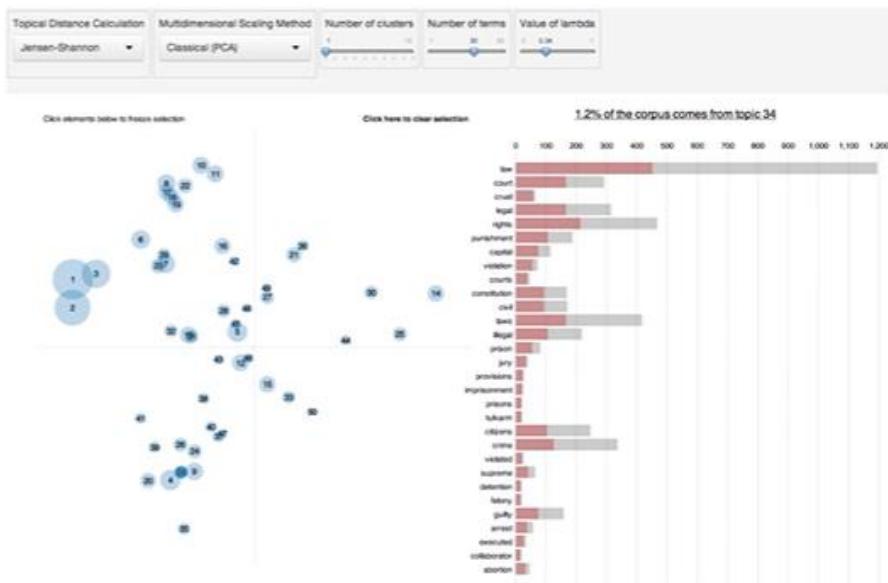
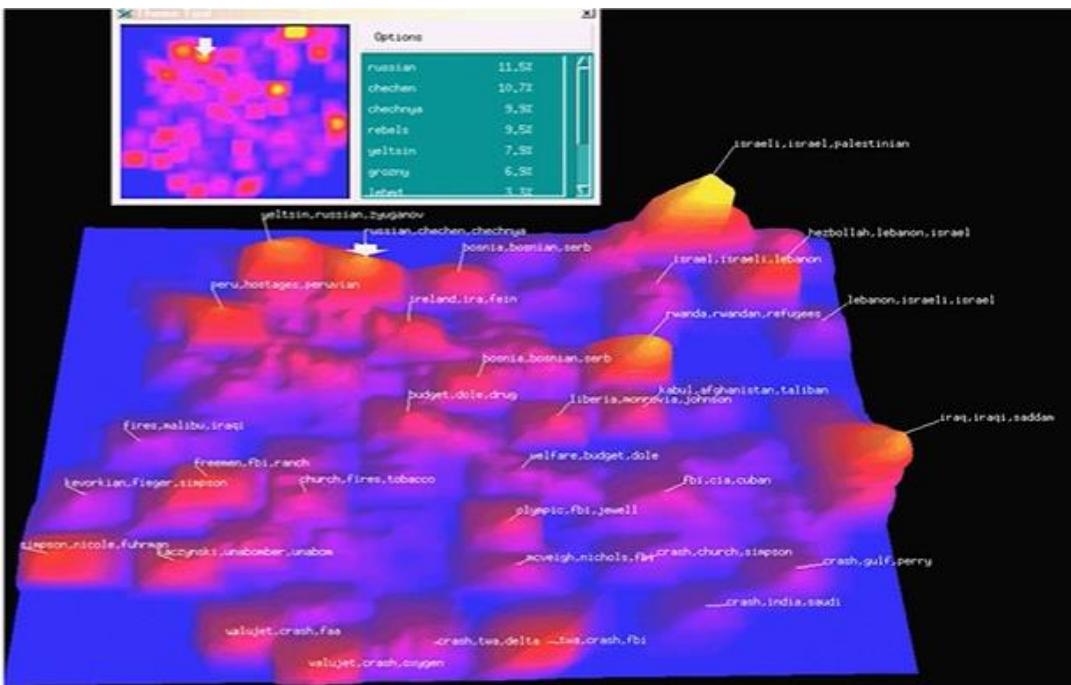


Figure 1: The layout of LDAvis, with the global topic view on the left, and the term barcharts (with Topic 34 selected) on the right. Linked selections allow users to reveal aspects of the topic-term relationships compactly.

x. ThemeScape:



xi. History Flow:

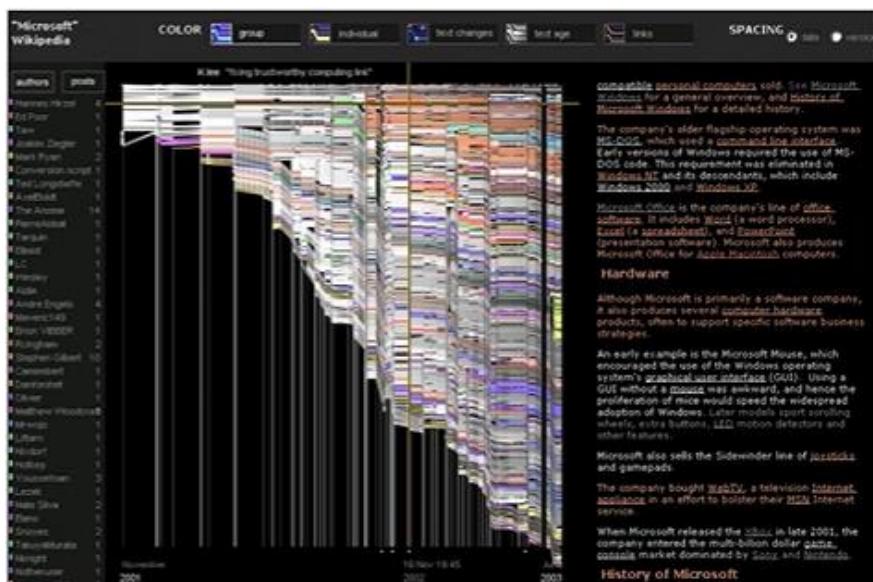


Fig 3: history flow user interface showing the Microsoft page on Wikipedia; on the right we see the contents of the page, on the left we see all the authors who have contributed to this page; the center panel shows the visualization

- m. El modelo de bolsa de palabras para representar texto, así como el modelo de espacio vectorial, que permite representar datos no estructurados en datos estructurados.
- n. Ejemplo de distintas visualizaciones de texto en función de sus características, de la estructura del texto, así como en función del corpus y metadata.

31. Visualización de vectores de palabras

- a. La representación de texto y su posterior visualización se ha visto impactada en los últimos años con el desarrollo de varios métodos de redes neuronales.
- b. A partir de la hipótesis distribucional en lingüística, un modelo de red neuronal con word2vec o GloVe produce vectores por palabra y podemos explorar sus propiedades a través de visualizaciones.
- c. Veremos un ejemplo de visualización de estos modelos de vectores de palabras usando lenguaje Python.
- d. **Conceptos de modelos de vectores:**
 - i. **Representación de palabras:**
 1. La hipótesis distribucional indica: Palabras similares ocurren en contextos similares.
 2. Contexto de una palabra: Corresponde a las palabras alrededor de una palabra.
 3. Una palabra se conoce por su compañía, "J. R. Firth (1957)"
 4. Ejemplo:

1

Cocinar **tomates** / hacer ensalada de **tomates** / dejar **tomates** limpios

2

Cocinar **lechugas** / hacer ensalada de **lechuga** / dejar **lechuga** limpia

Las palabras "**tomates**" y "**lechugas**" son similares porque ocurren en contextos similares:

Las palabras que las rodean son las mismas o muy similares.

- 5. Otro ejemplo de regularidad semántica:

¿Cómo capturar esto con vectores?

1

Mujer es a reina como hombre es a rey

2

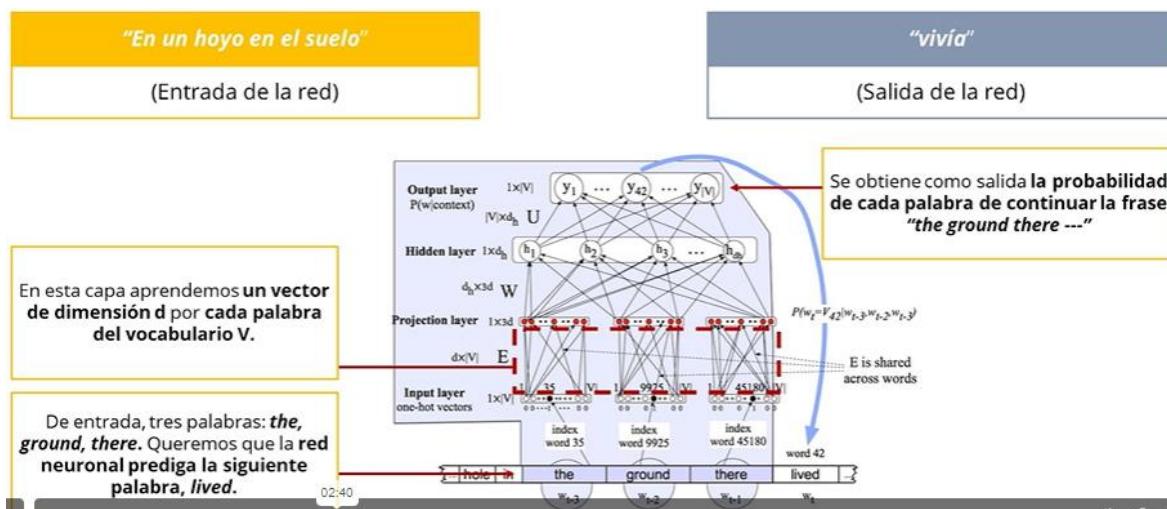
Mujer es a hermana como hombre es a hermano

3

Correr es a corrió como sentir es a sintió

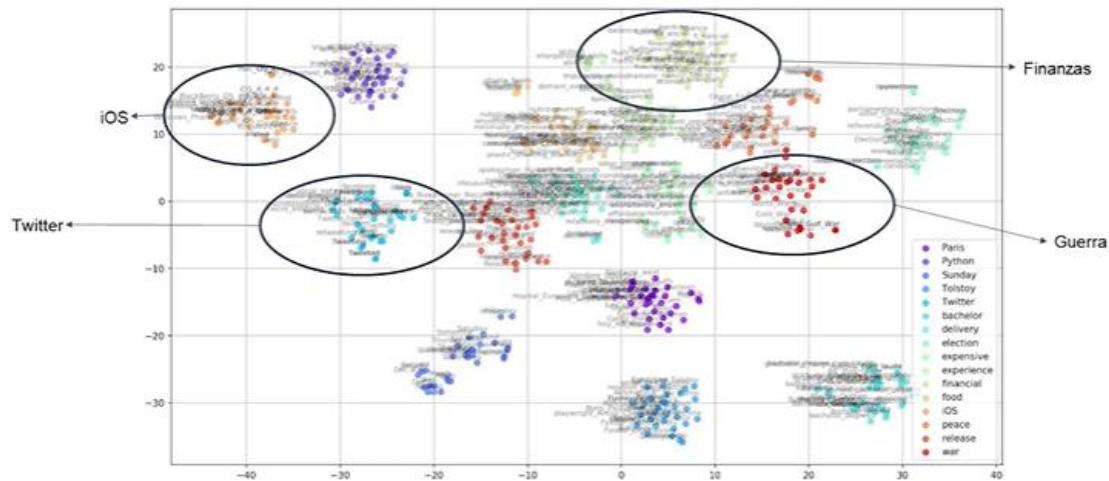
e. Representando palabras con vectores

i. Modelo de lenguaje usando redes neuronales:



ii. ¿Qué visualización con estos vectores?

Reducción de dimensionalidad y un bubble chart



- f. Los vectores de palabras y la hipótesis distribucional.
- g. Términos generales sobre cómo se calculan los vectores de palabras.
- h. Un ejemplo de Python usando la biblioteca gensim.