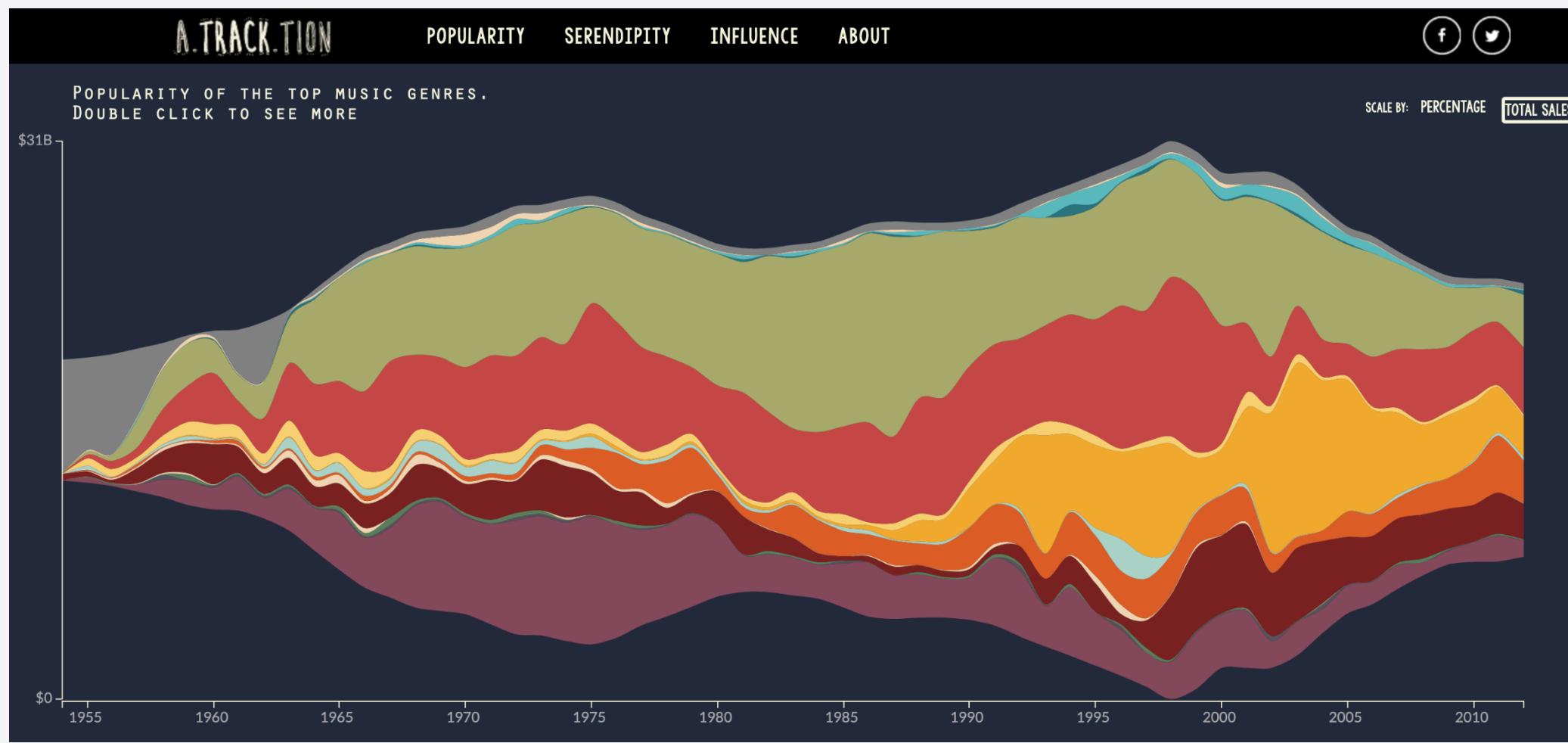




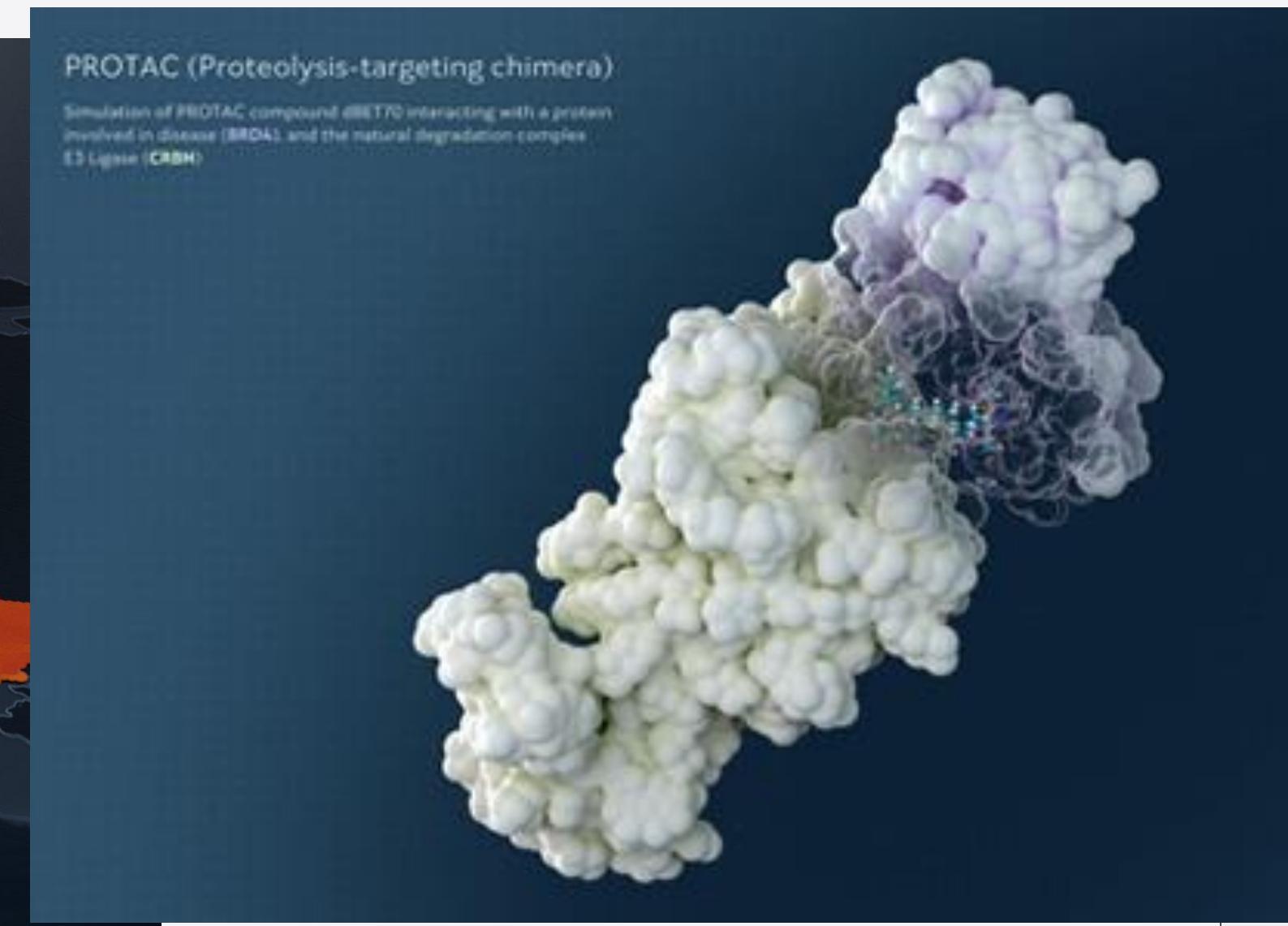
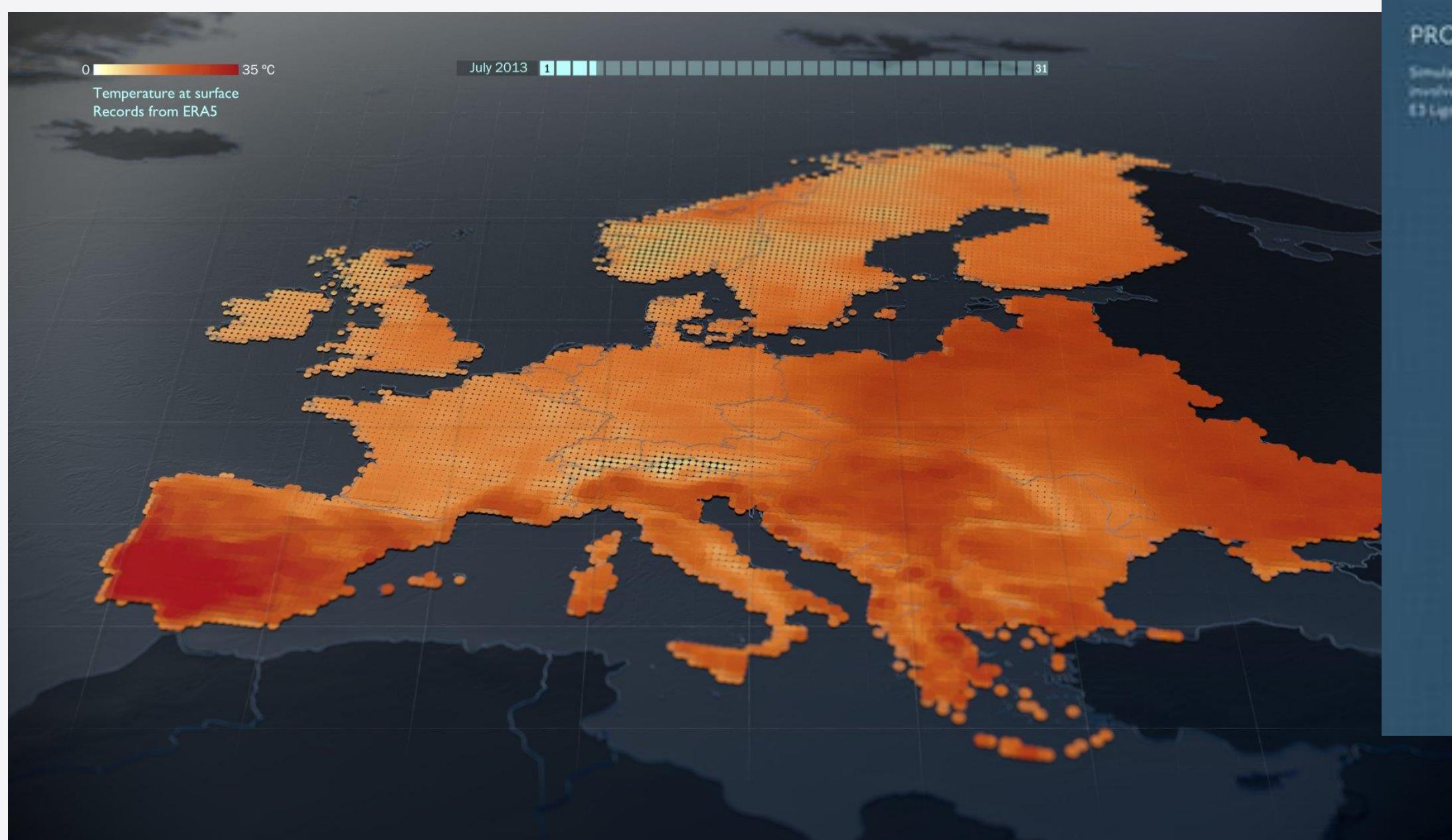
2. Tipos de datos

Guillermo Marin
guillermo.marin@uab.cat



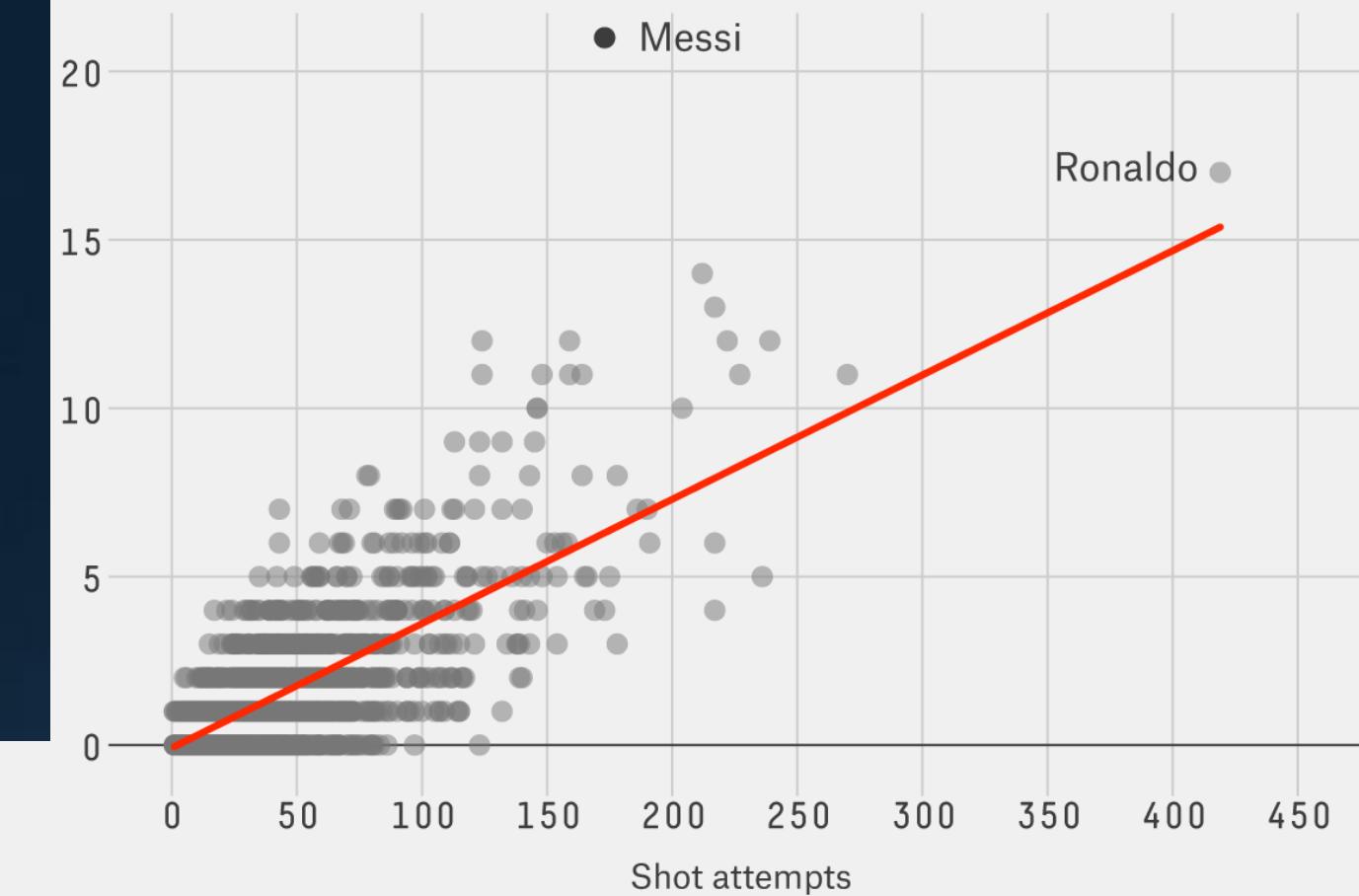
- Ejemplos muy diferentes pero que se basan en los mismos principios y técnicas
- Pueden analizarse y producirse bajo un marco común
- Características comunes a todas las gráficas que podemos usar para hacer la mejor visualización posible en cada caso.

MENSAJE PRINCIPAL: MISMAS TÉCNICAS PARA PRODUCIR CUALQUIER VISUALIZACIÓN - INDEPENDIENTE DEL ÁREA

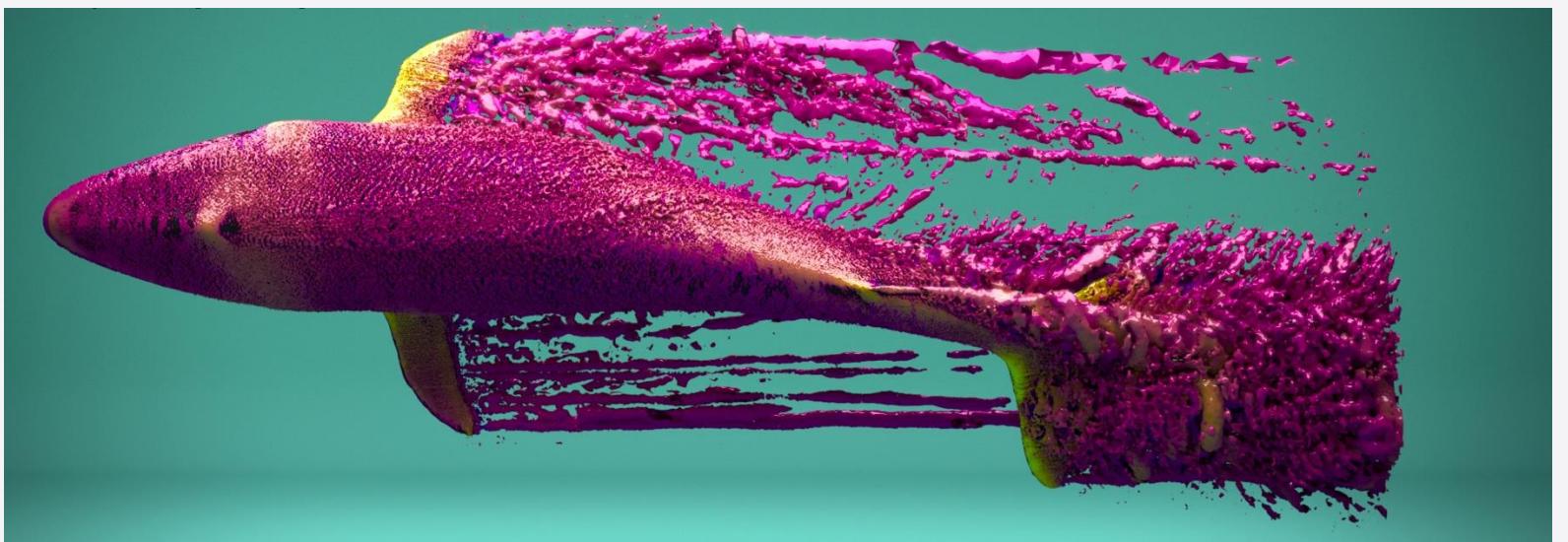


deadly From Outside the Penalty Area

Goals scored vs. shot attempts



La representación visual de información compleja de forma que ayude al razonamiento y a la comprensión

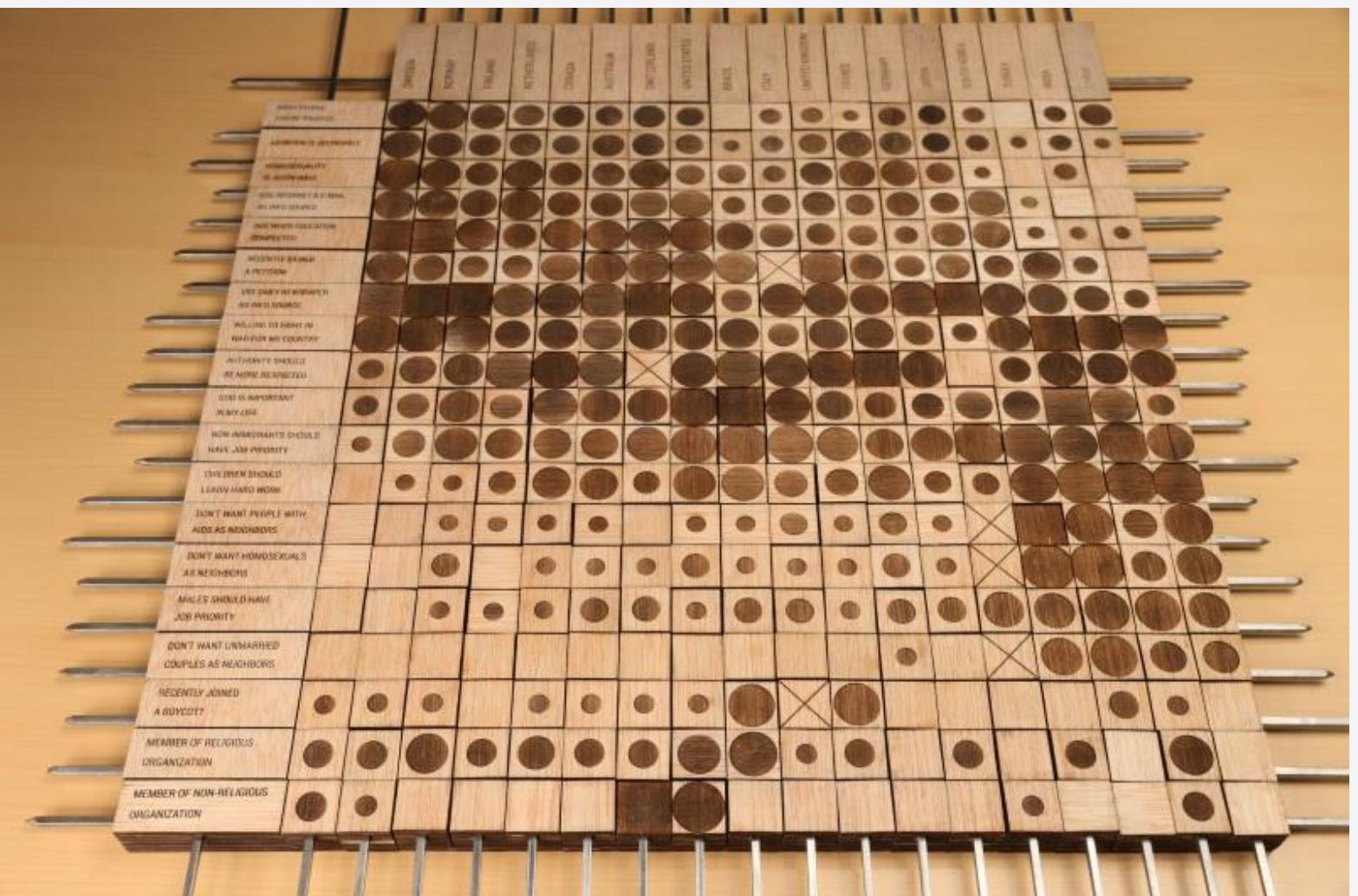


Convertir los datos en información visual, en **representaciones externas**

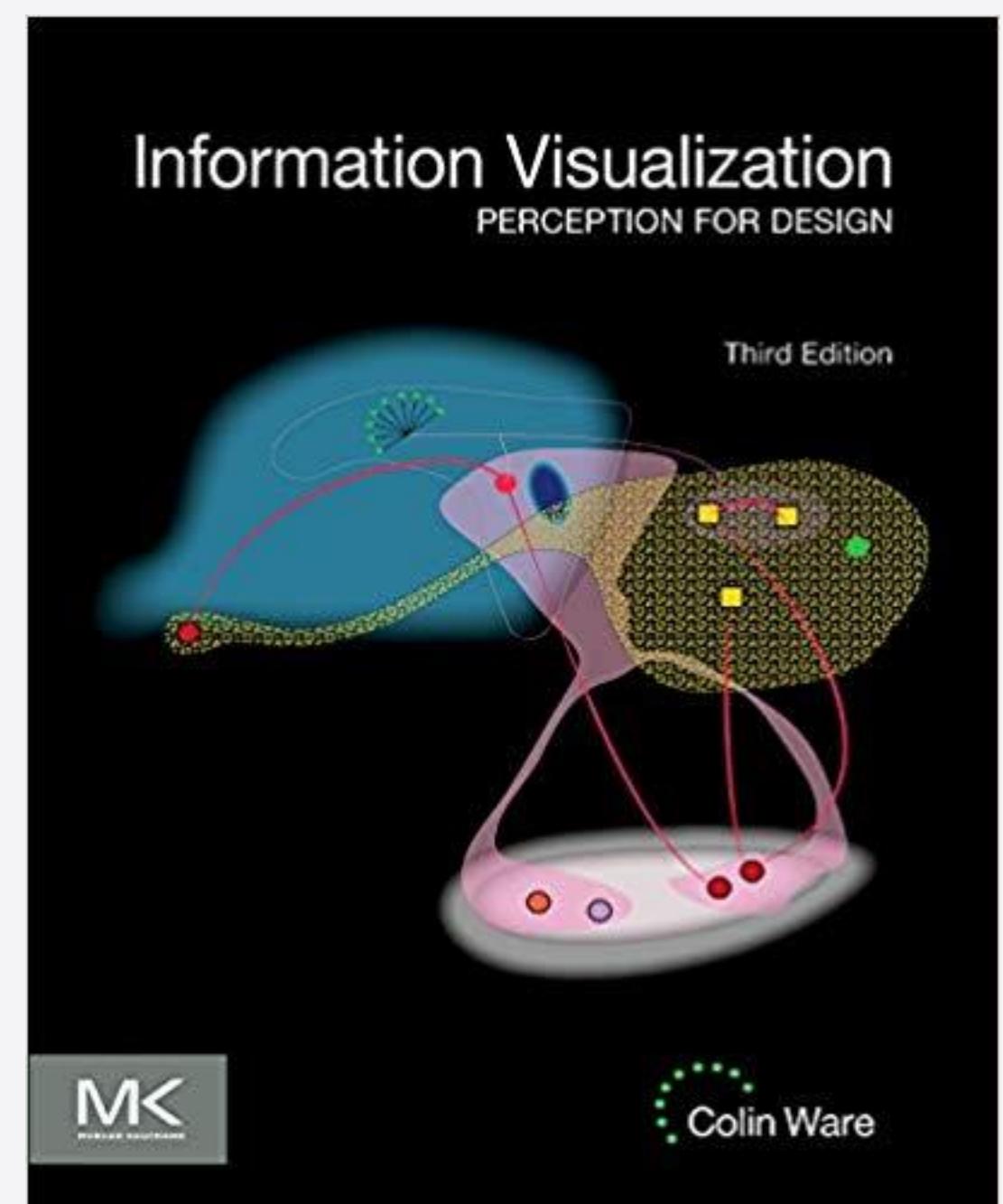
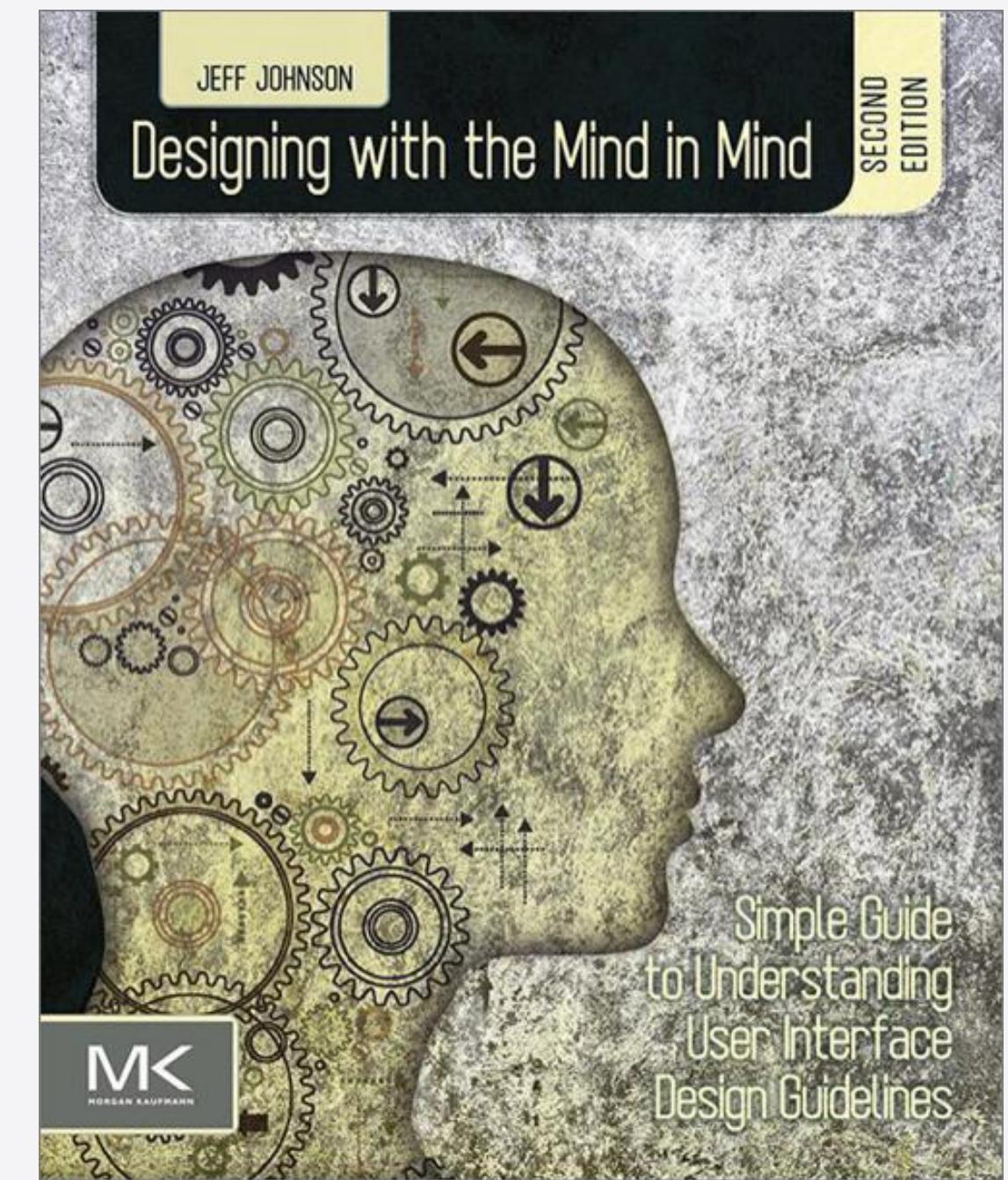
Tanto artefactos físicos (ábaco, matrices físicas de Bertin), como gráficas 2D

Porqué usar representaciones externas?

- Aumentan las capacidades humanas permitiendo superar limitaciones cognitivas y de memoria
- Herramientas cognitivas que ayudan a pensar (Ware, C.)

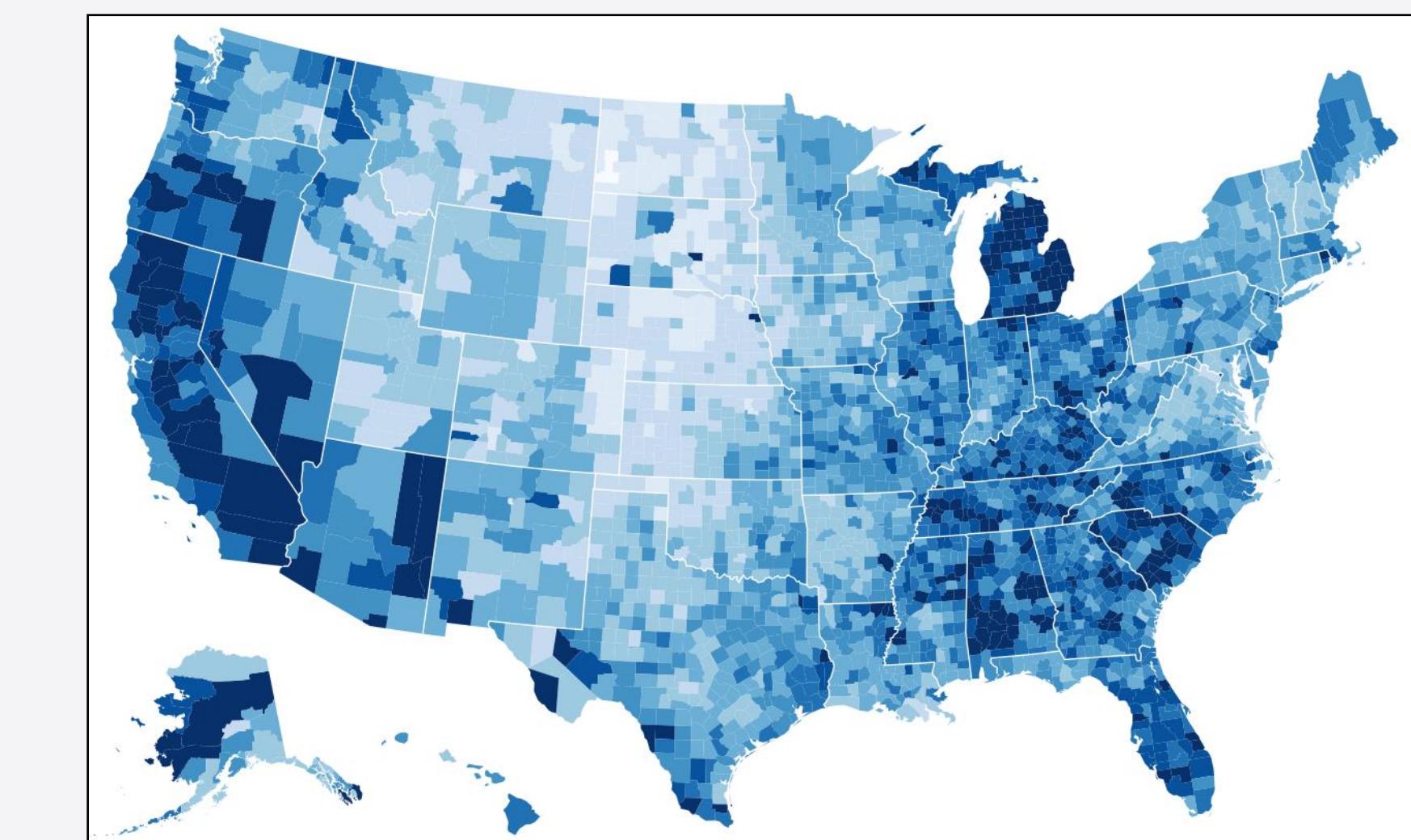
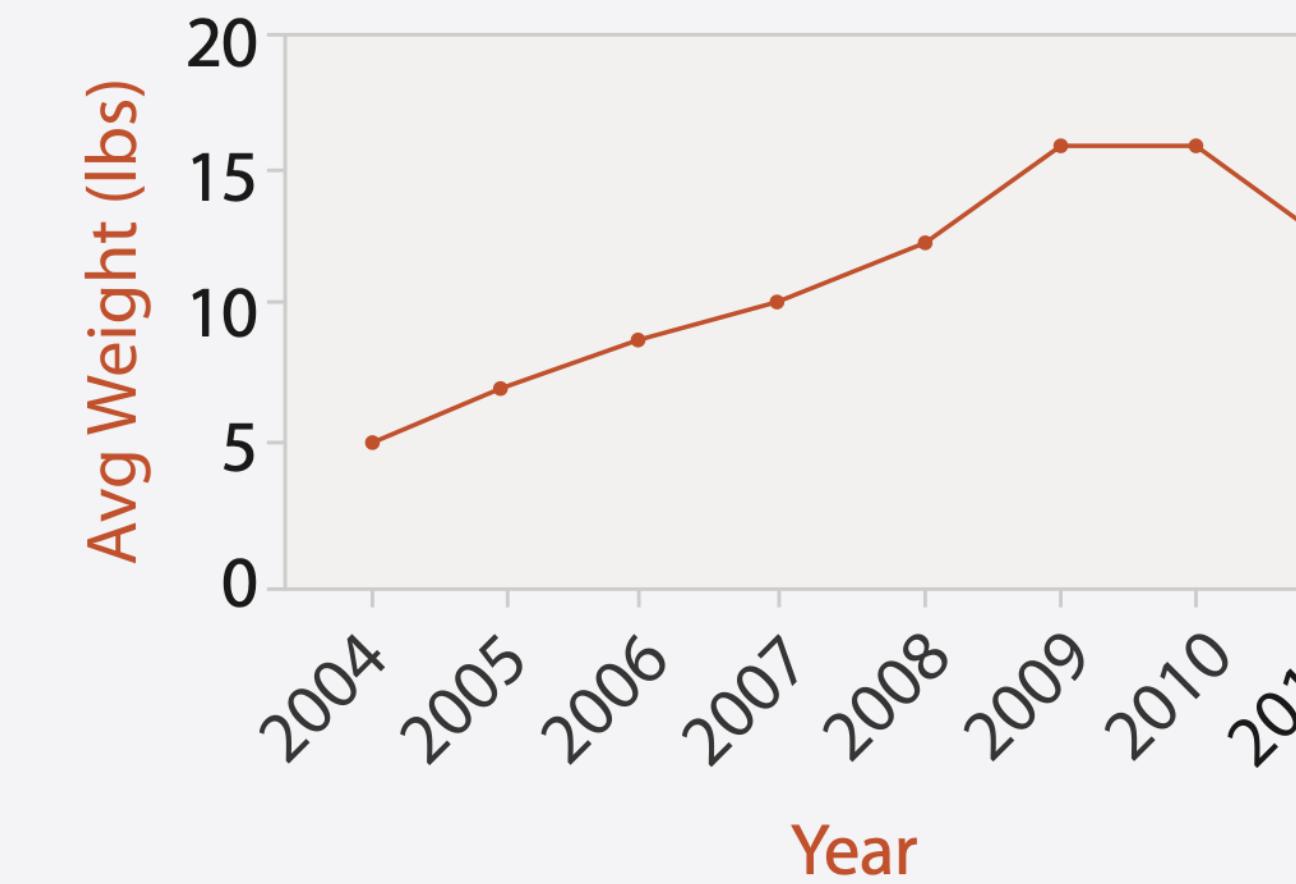
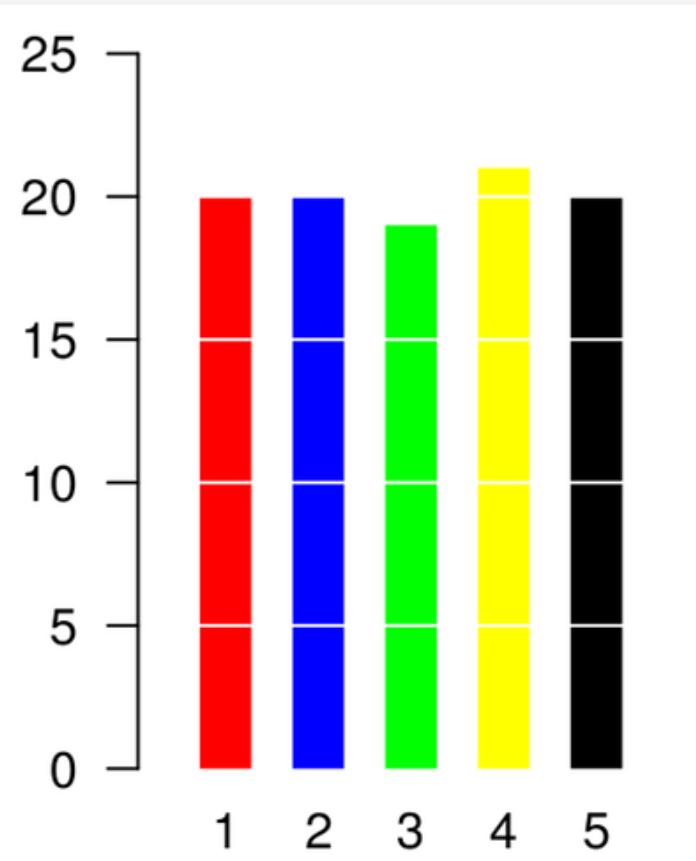
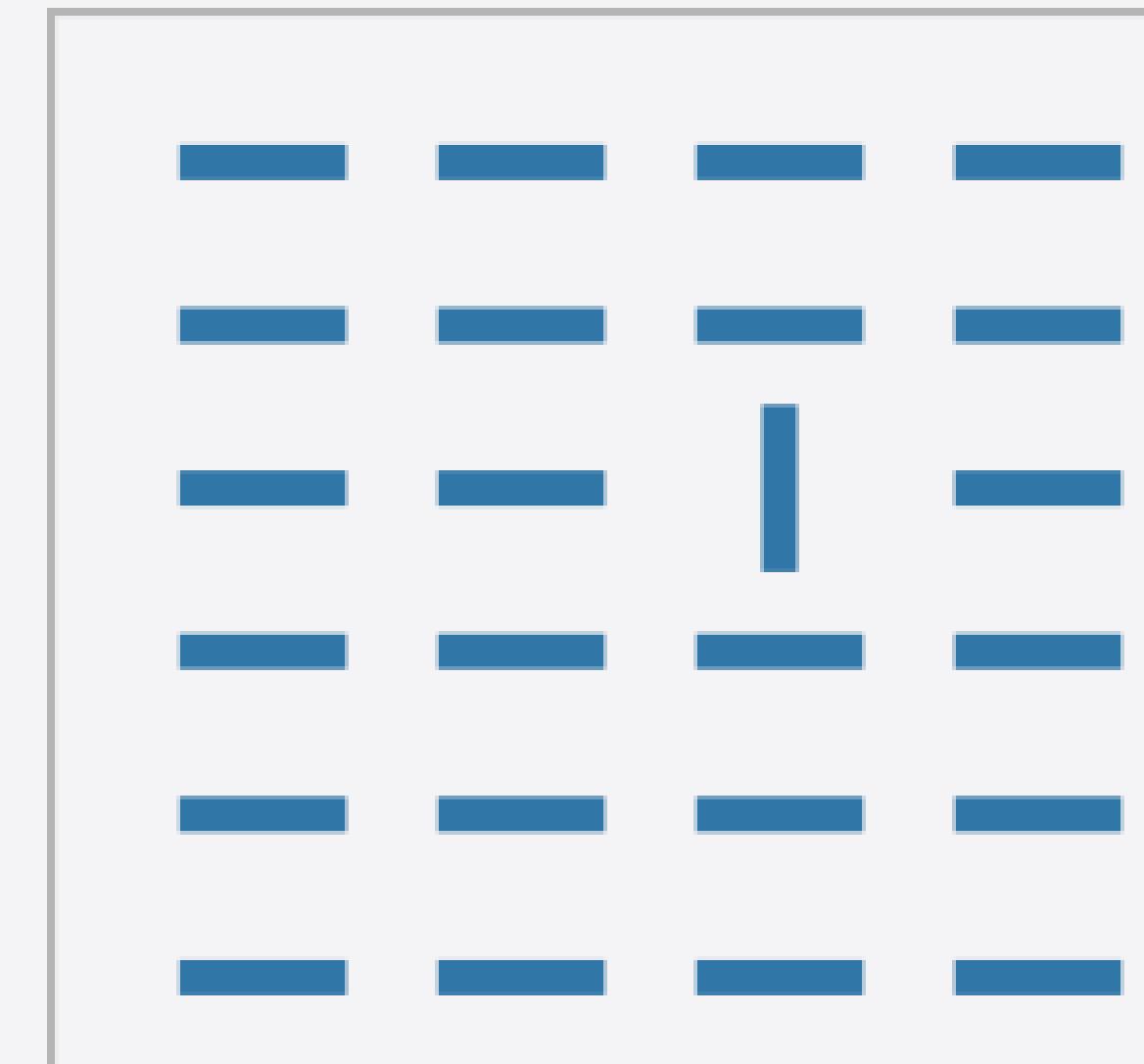
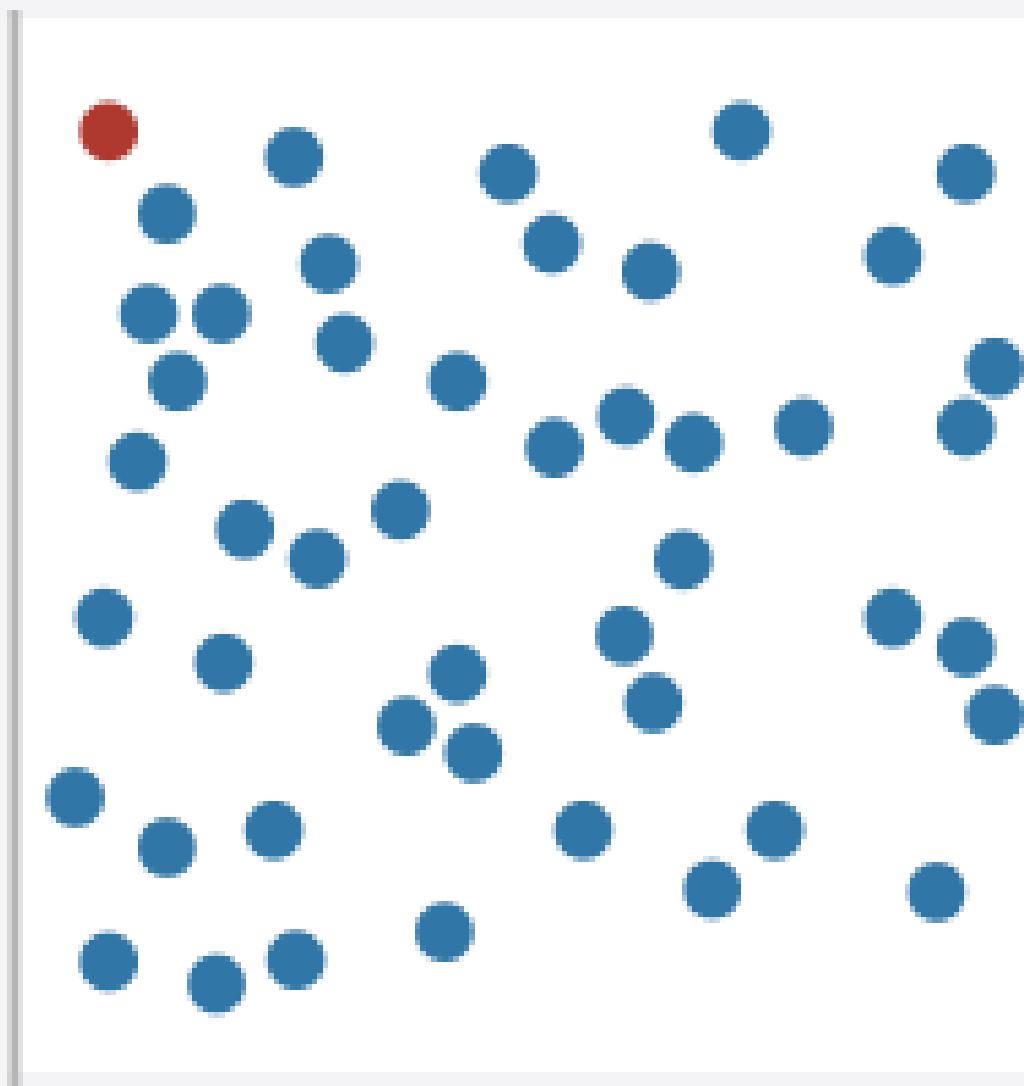


Physical Matrix. Bertin, J.- <https://aviz.fr/diyMatrix/>



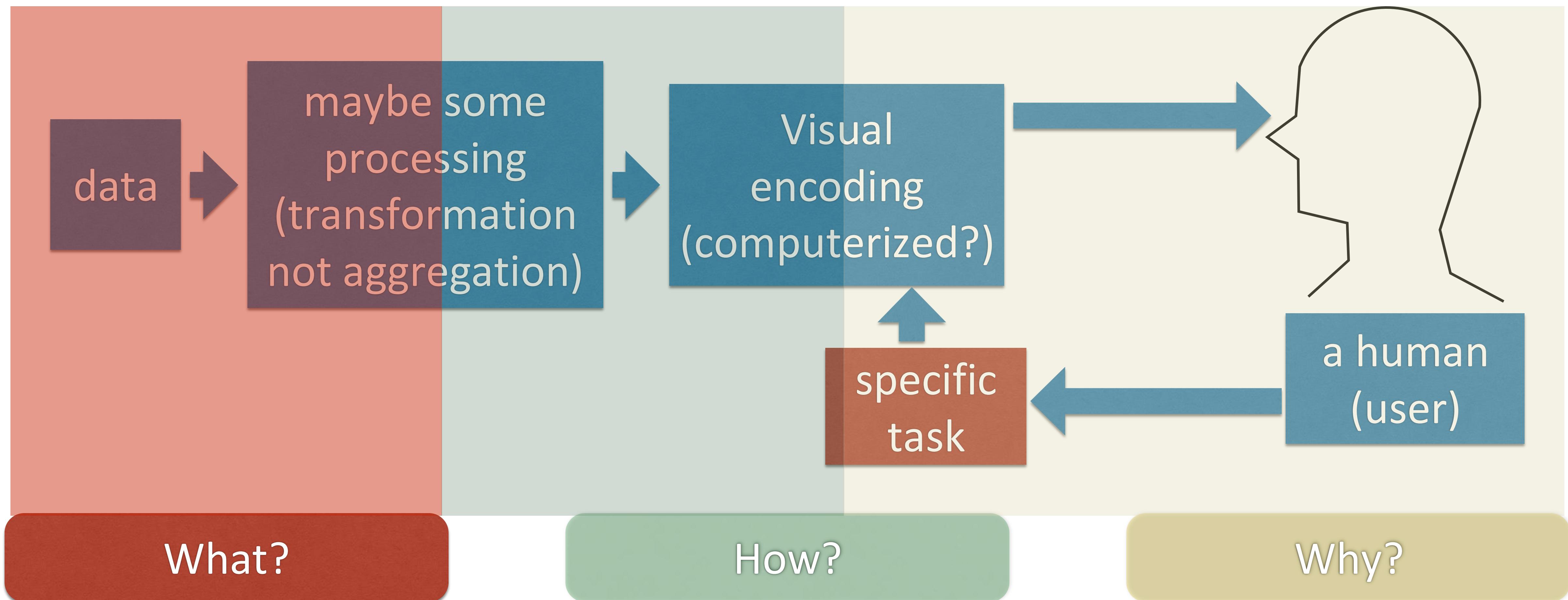
Visualizations exploit perceptual and cognitive mechanisms in our minds, tuned for pattern detection

Find the red dot



Data Visualisation

*Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively
(T.Munzner)*



Marcas y Canales

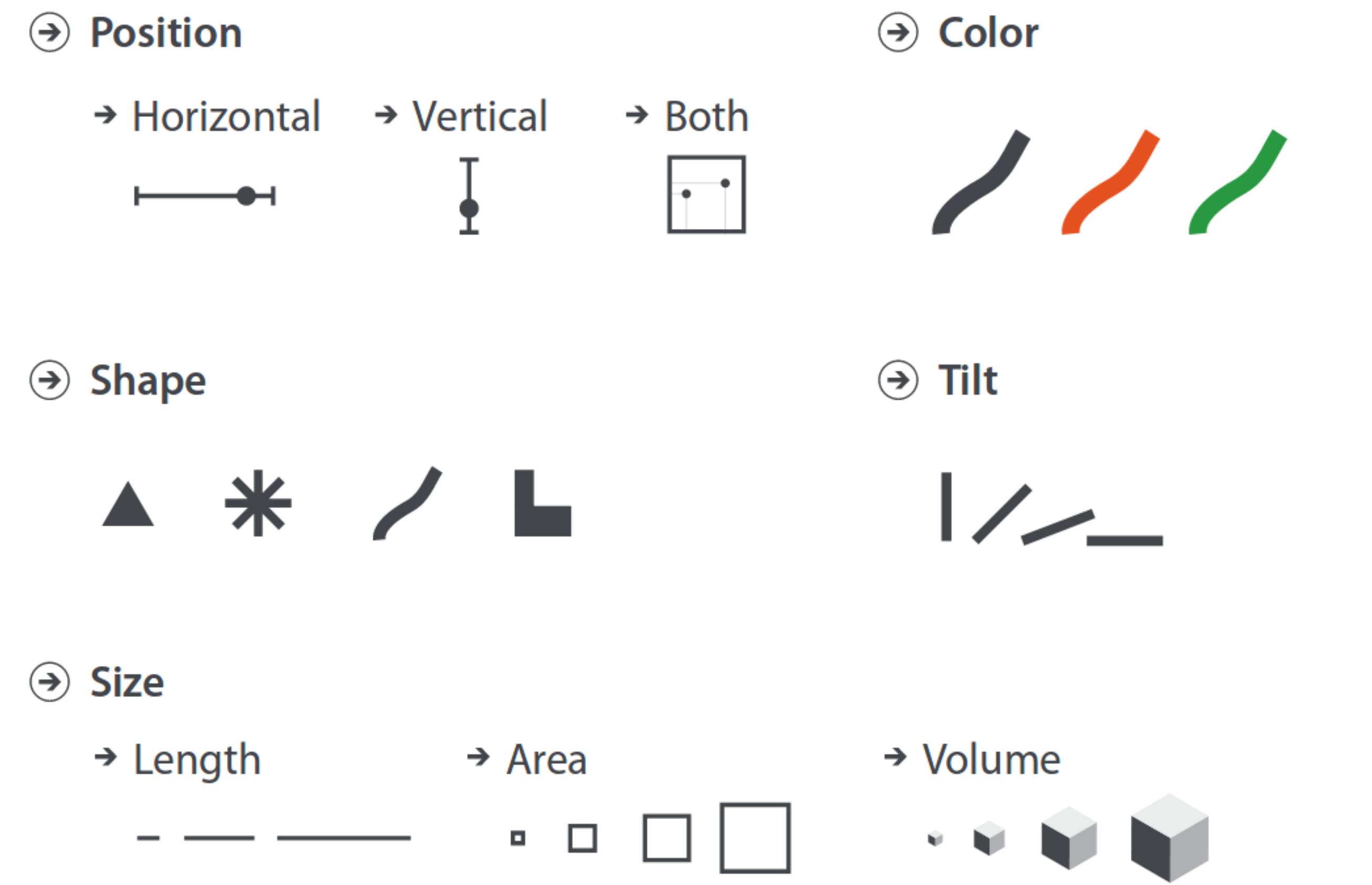
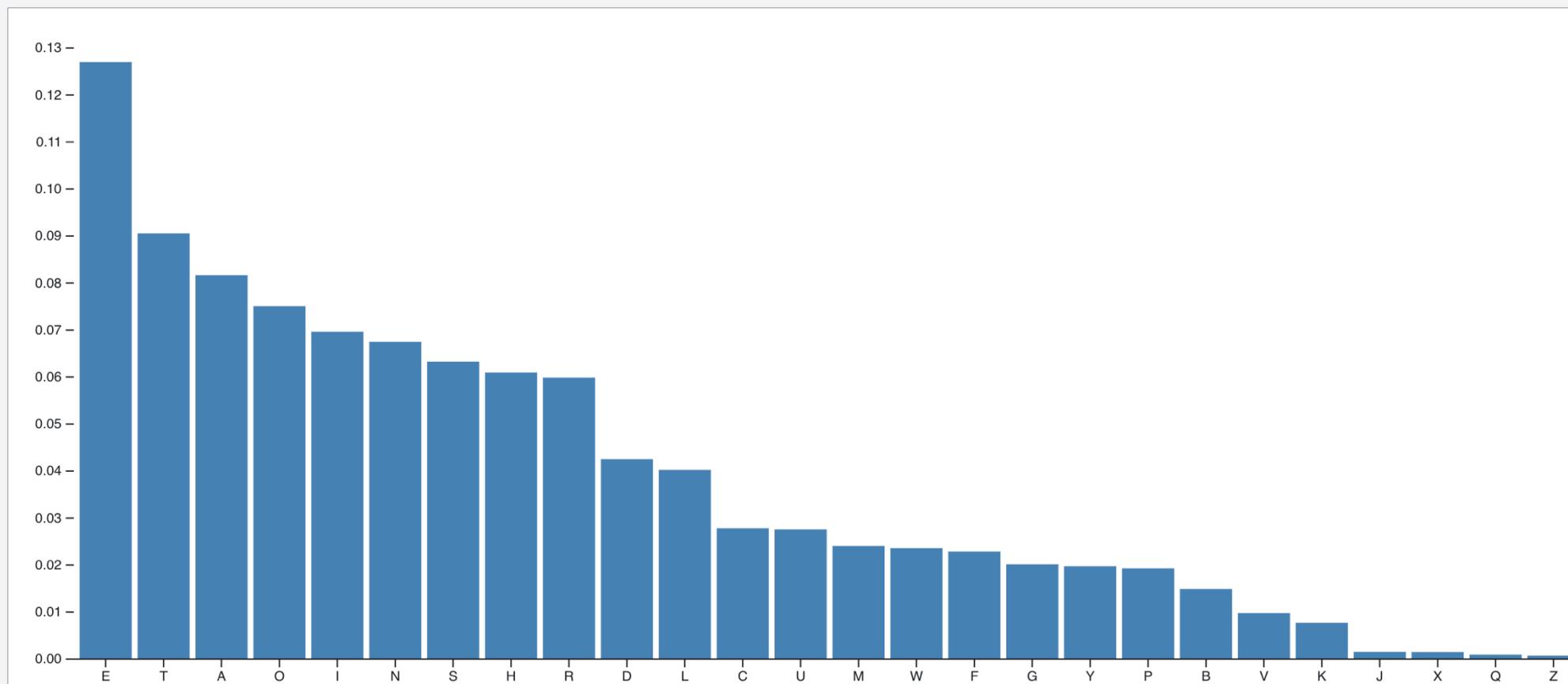
Asignación de propiedades gráficas a los datos

Marcas

Elementos geométricos básicos clasificados por dimensiones

Canales

Los canales visuales que modulan la apariencia de las marcas



Tipos de Canales

- No todos los canales son iguales
- Percibimos dos tipos generales de modalidades sensoriales:
- Canales de Identidad – Dan información de **Qué** es algo.
- Canales de Magnitud – Nos dice **Cuanto** hay de algo.

Identidad

→ Shape



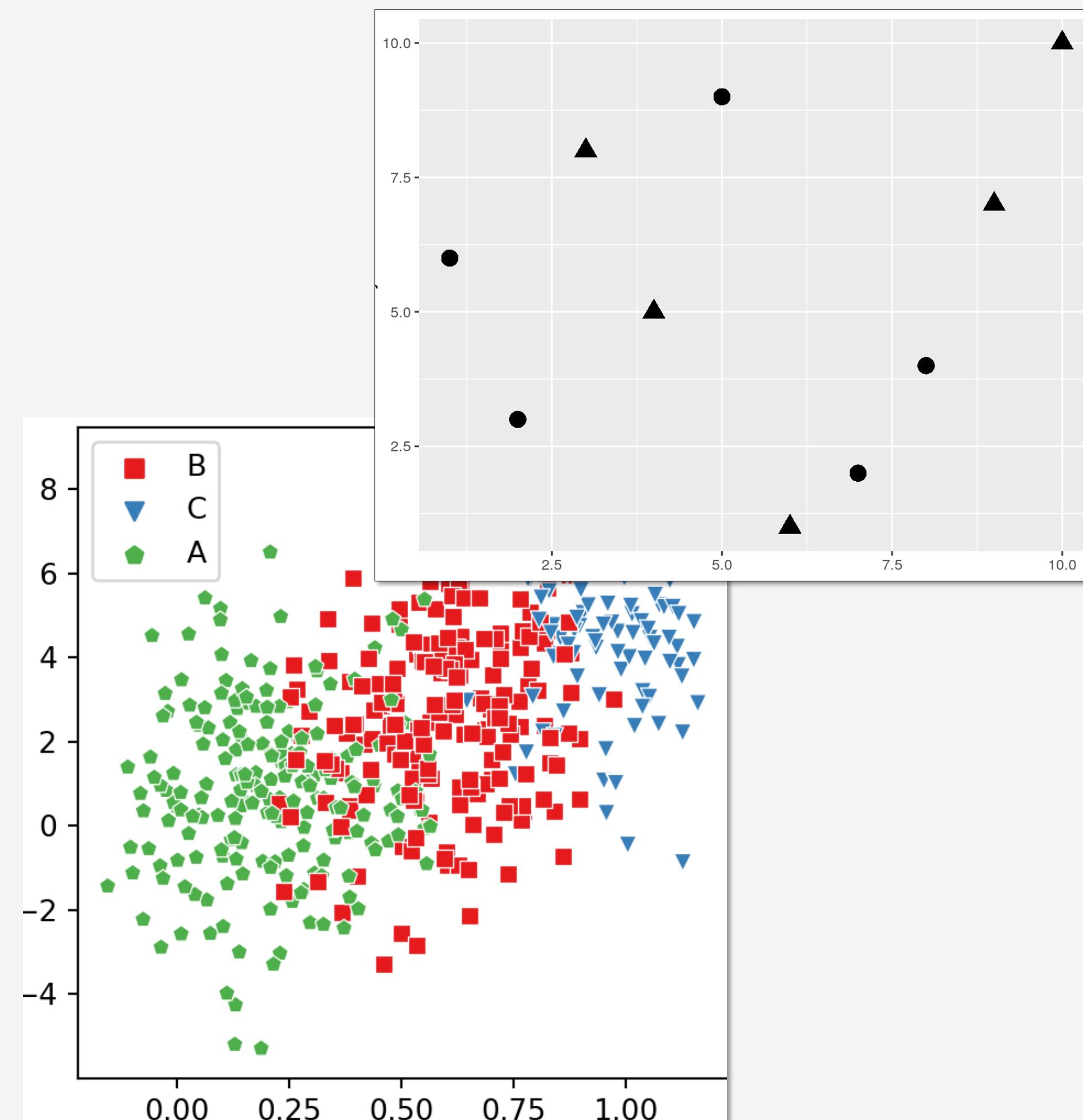
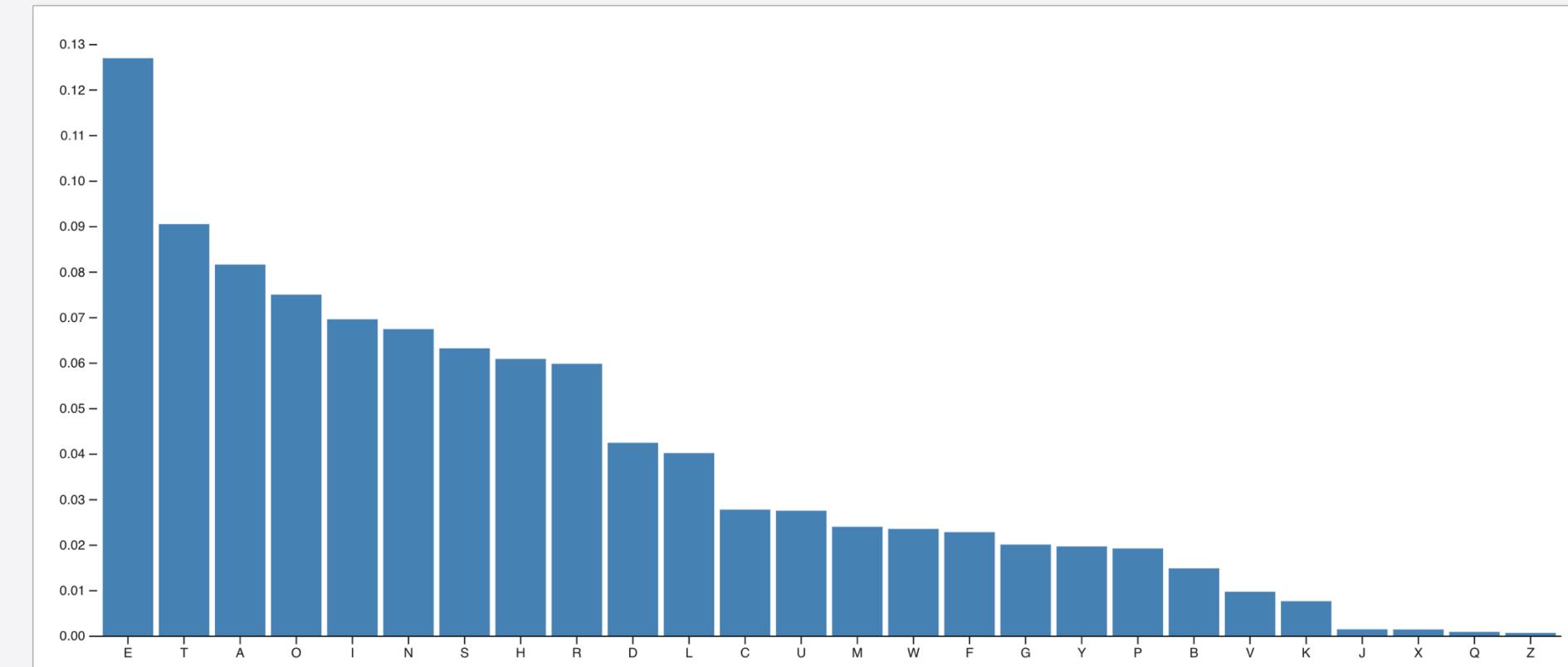
e.g., podemos decir **Qué forma vemos**
¿Tiene sentido hacer preguntas de magnitud?

Magnitud

→ Length



e.g., podemos hacer preguntas de magnitud para la longitud. ¿Preguntar identidad?



Tipos

- No tiene
- Percepción
- Canción
- Canción

Identidad

→ Shape



e.g., poe

¿Tiene se

Car



Tipos

- No to
- Perci
- Canc
- Canc

Identid

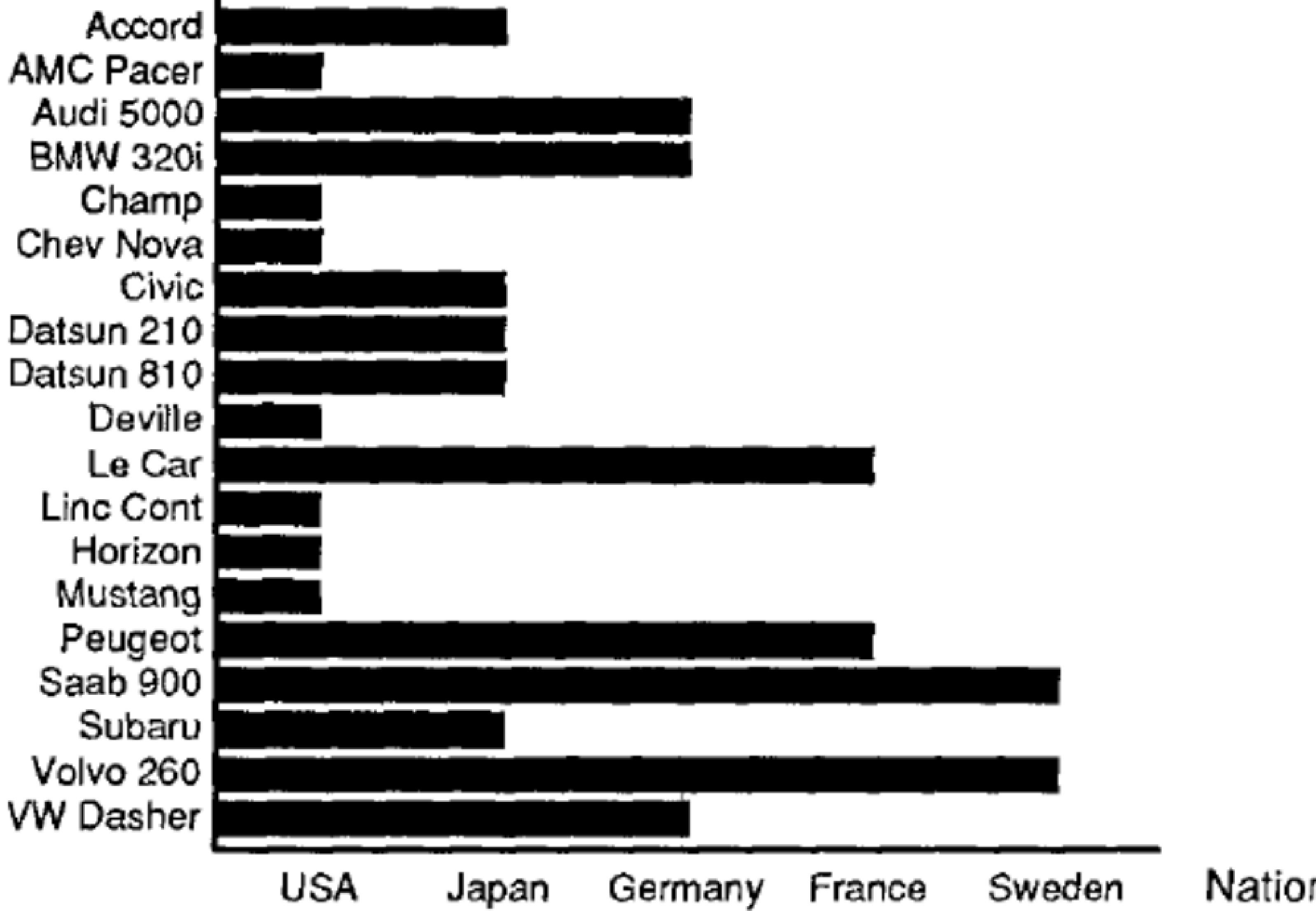
→ Shap



e.g., po

¿Tiene se

Car



Car nationality for 1979

Tipos de Canales

- **No todos los canales son iguales**
- Percibimos dos tipos generales de modalidades sensoriales:
- Canales de Identidad – Dan información de **Qué** es algo.
- Canales de Magnitud – Nos dice **Cuanto** hay de algo.

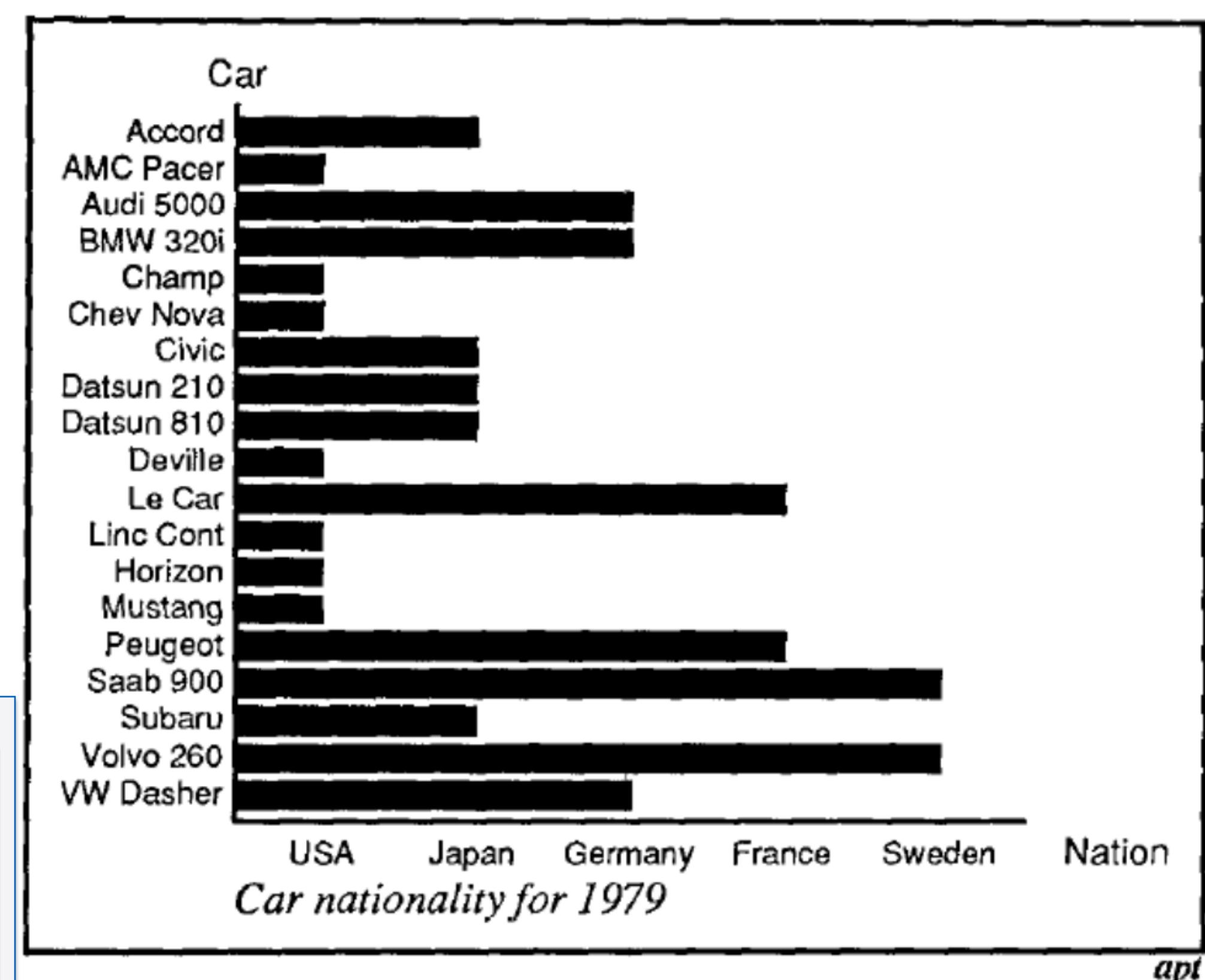
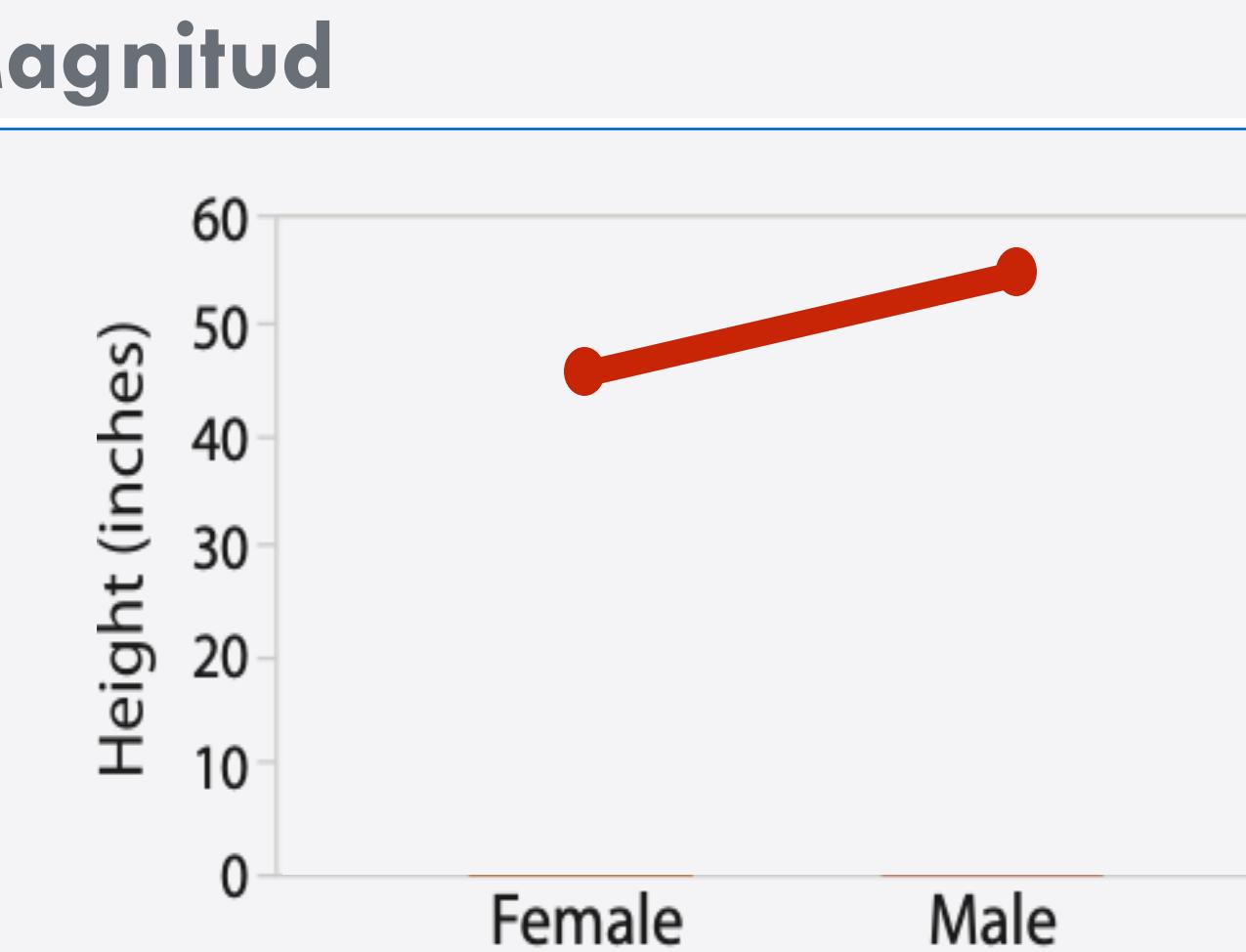
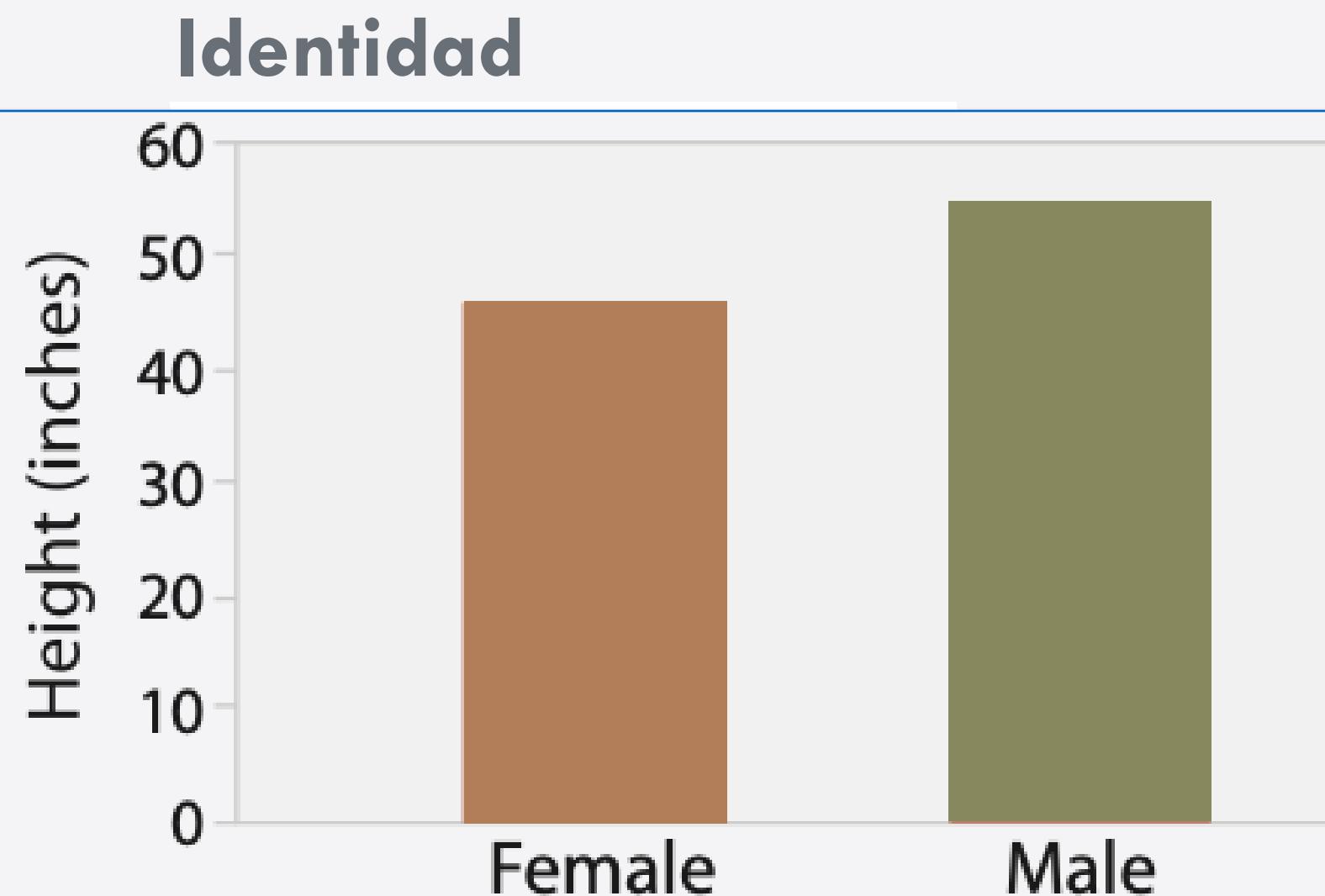


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

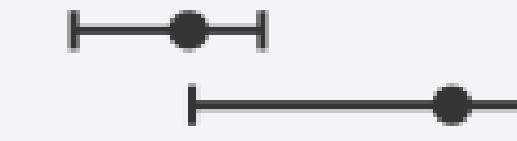
Channels: Expressiveness Types And Effectiveness Ranks

→ **Magnitude** Channels: **Ordered Attributes**

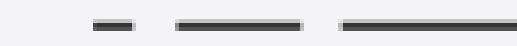
Position on common scale



Position on unaligned scale



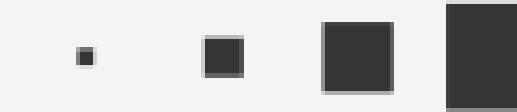
Length (1D size)



Tilt/angle



Area (2D size)



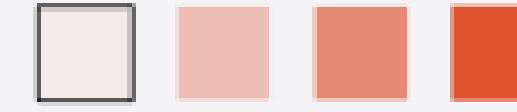
Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



→ **Identity** Channels: **Categorical Attributes**

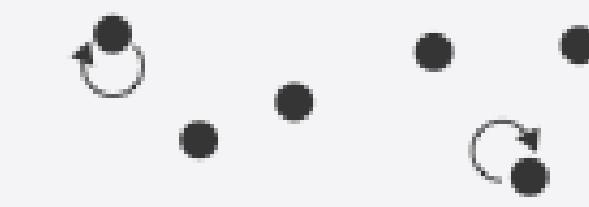
Spatial region



Color hue



Motion



Shape



Best ↑

Effectiveness ↓

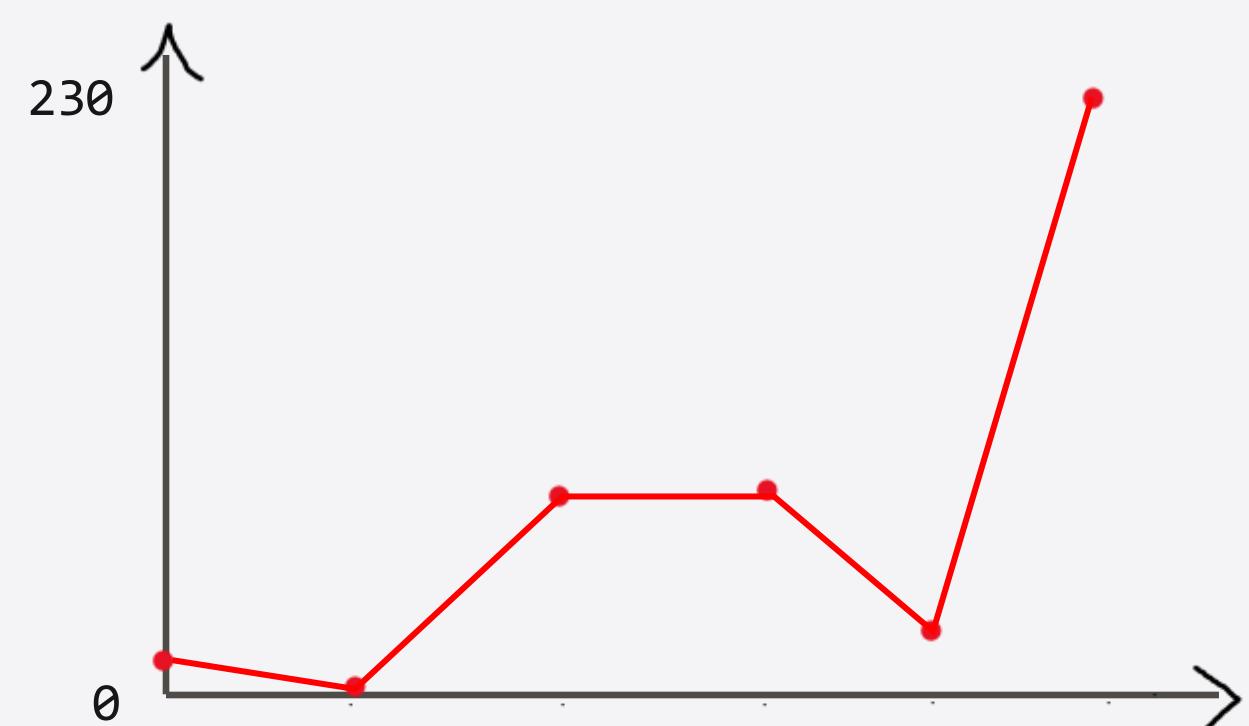
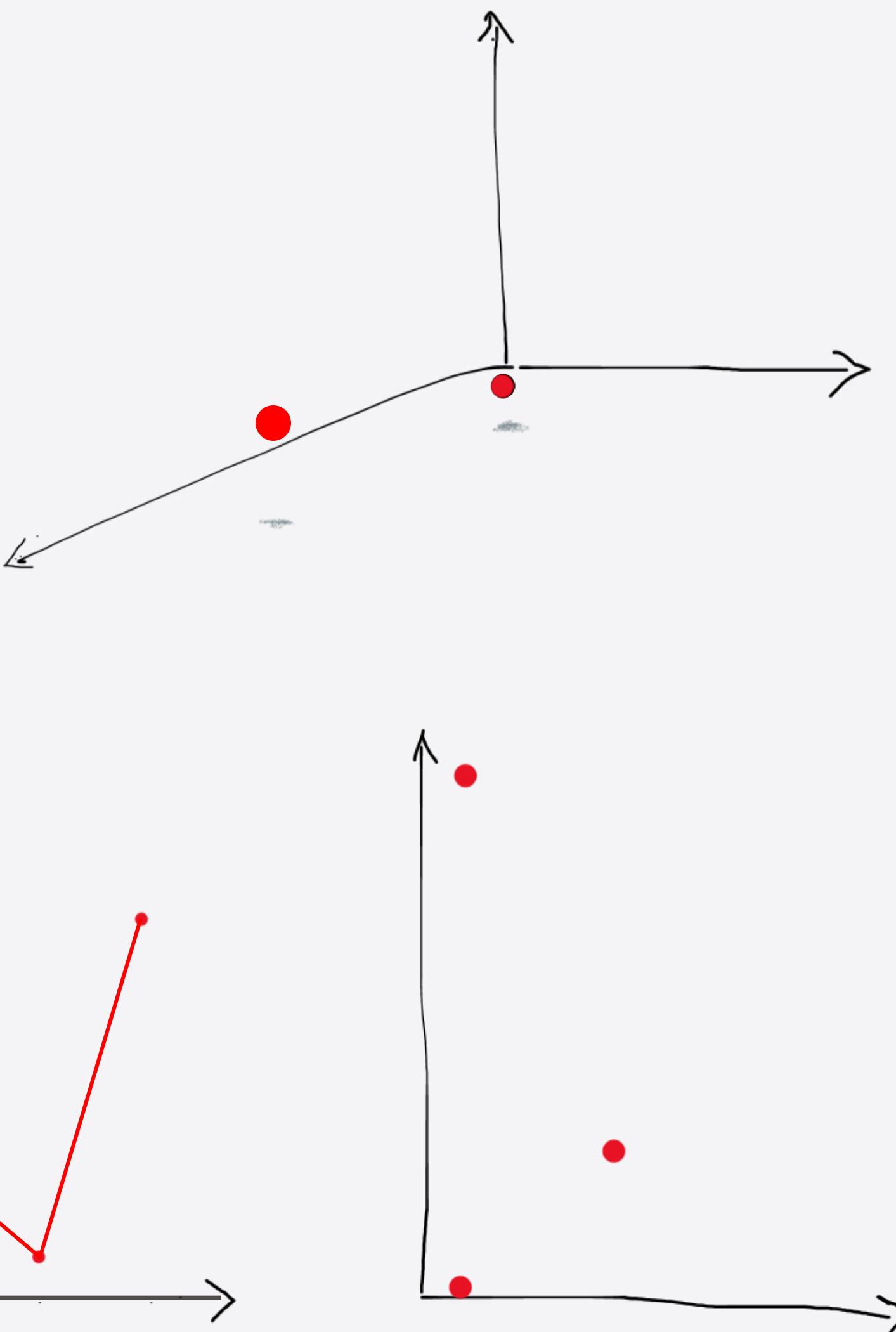
Least ▼

Canales visuales separados en Magnitud e Identidad
Orden vertical según efectividad

Hay características de los datos que determinan cómo los representamos.

Dos características esenciales:

- Semántica - Qué representa en el mundo real
- Tipo – Interpretación matemática Y estructural
- Muchas taxonomías y formas de clasificarlos



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583	
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125	
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13	
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18	
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225	
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292	
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6

What?

Datos

- Hay muchas formas de clasificar tipos de datos. Usamos una con implicaciones directas en la forma de representarlos visualmente y que es independiente del área.
- Dataset=Colección de información que es objeto de análisis
- Tipos básicos de dataset:
 - Tablas, redes, campos, geometría, etc
- Tipos básicos de datos:
 - Items, atributos, links, posiciones y grids
- Los atributos pueden ser:
 - categóricos, ordinales, cuantitativos

Datasets

→ Data Types

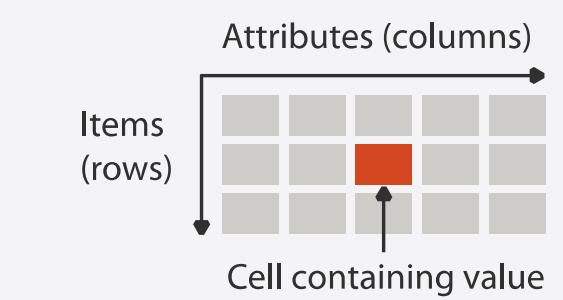
→ Items → Attributes → Links → Positions → Grids

→ Data and Dataset Types

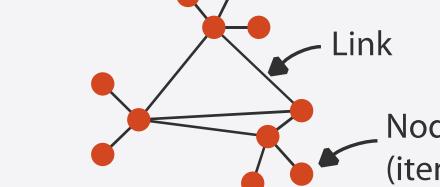
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items Attributes	Items (nodes) Links Attributes	Grids Positions Attributes	Items Positions	Items

→ Dataset Types

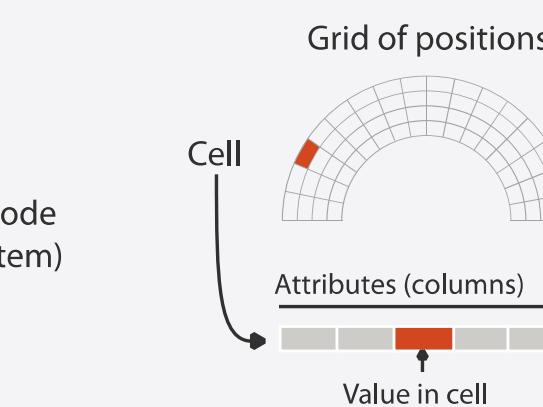
→ Tables



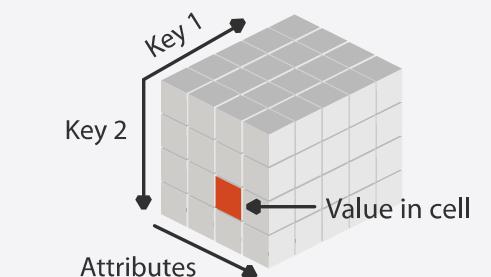
→ Networks



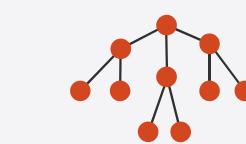
→ Fields (Continuous)



→ Multidimensional Table



→ Trees



→ Geometry (Spatial)



Attributes

→ Attribute Types

→ Categorical



→ Ordered

→ Ordinal



→ Quantitative



→ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



Tipos de Datasets

→ Tables

Attributes (columns)

Items (rows)

Cell containing value

→ Multidimensional Table

Key 1

Key 2

Attributes

Value in cell

→ Networks

Link

Node (item)

→ Trees

Grid of positions

Cell

Attributes (columns)

Value in cell

→ Spatial

→ Fields (Continuous)

Value in cell

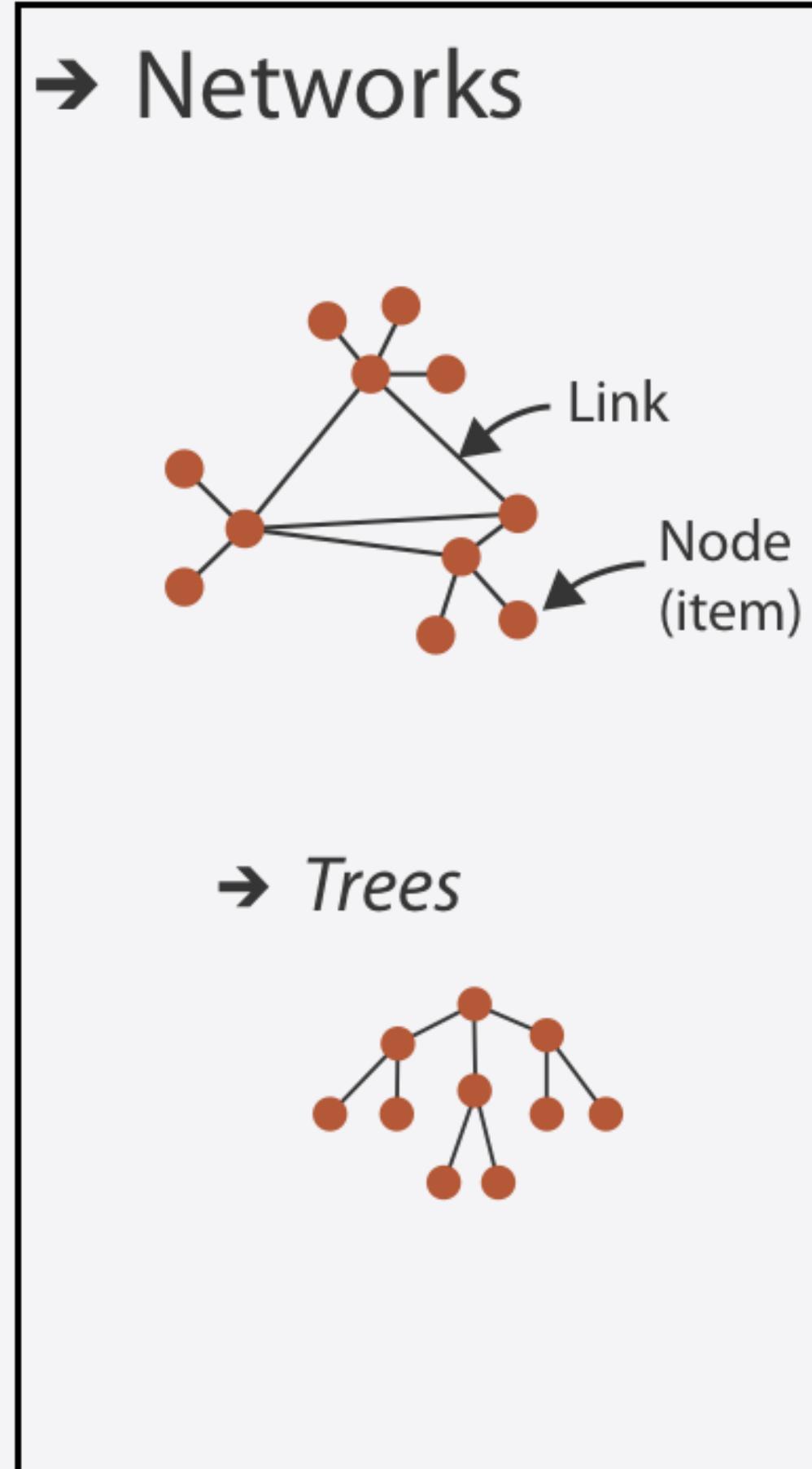
→ Geometry (Spatial)

Position

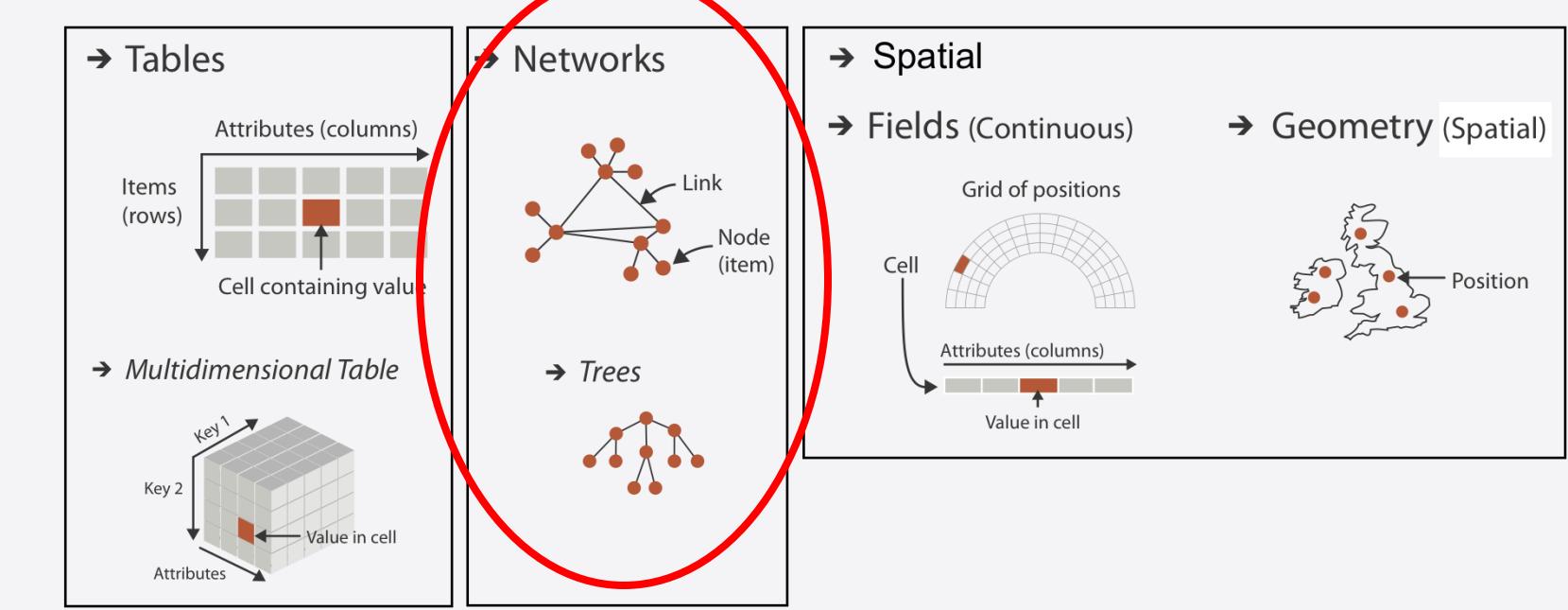
⇒ Data and Dataset Types

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

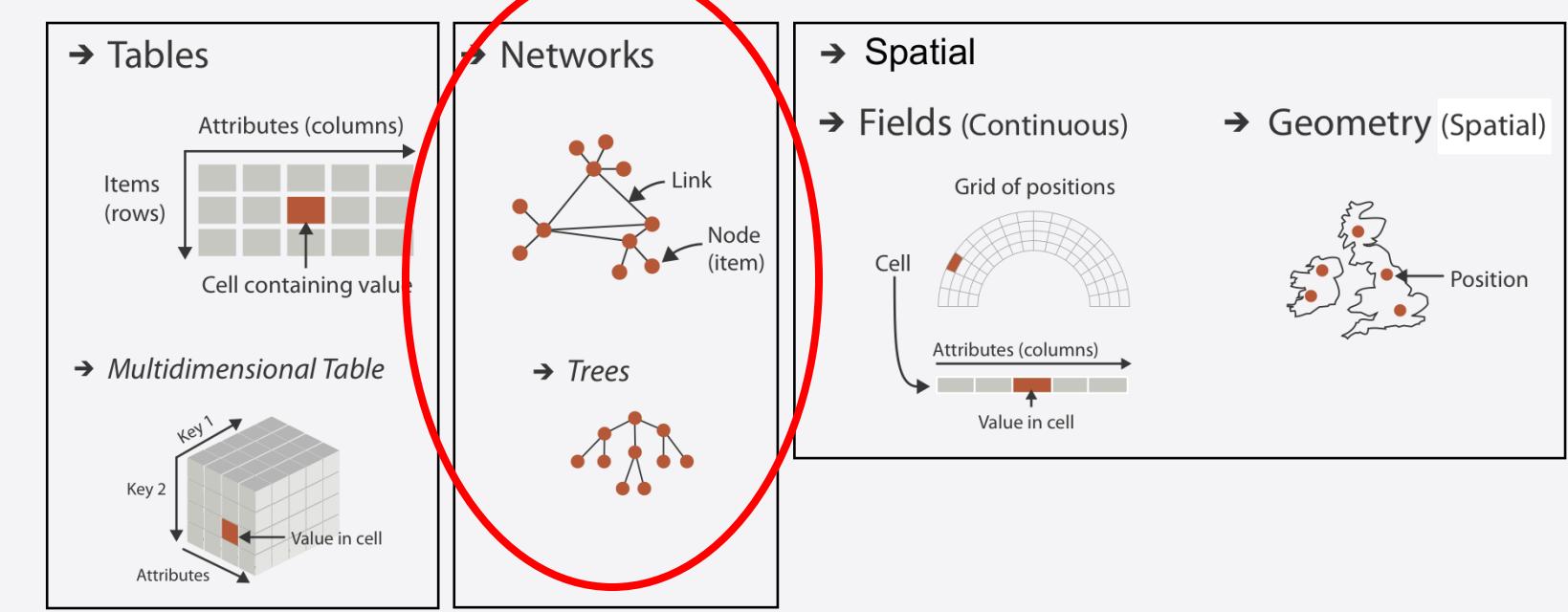
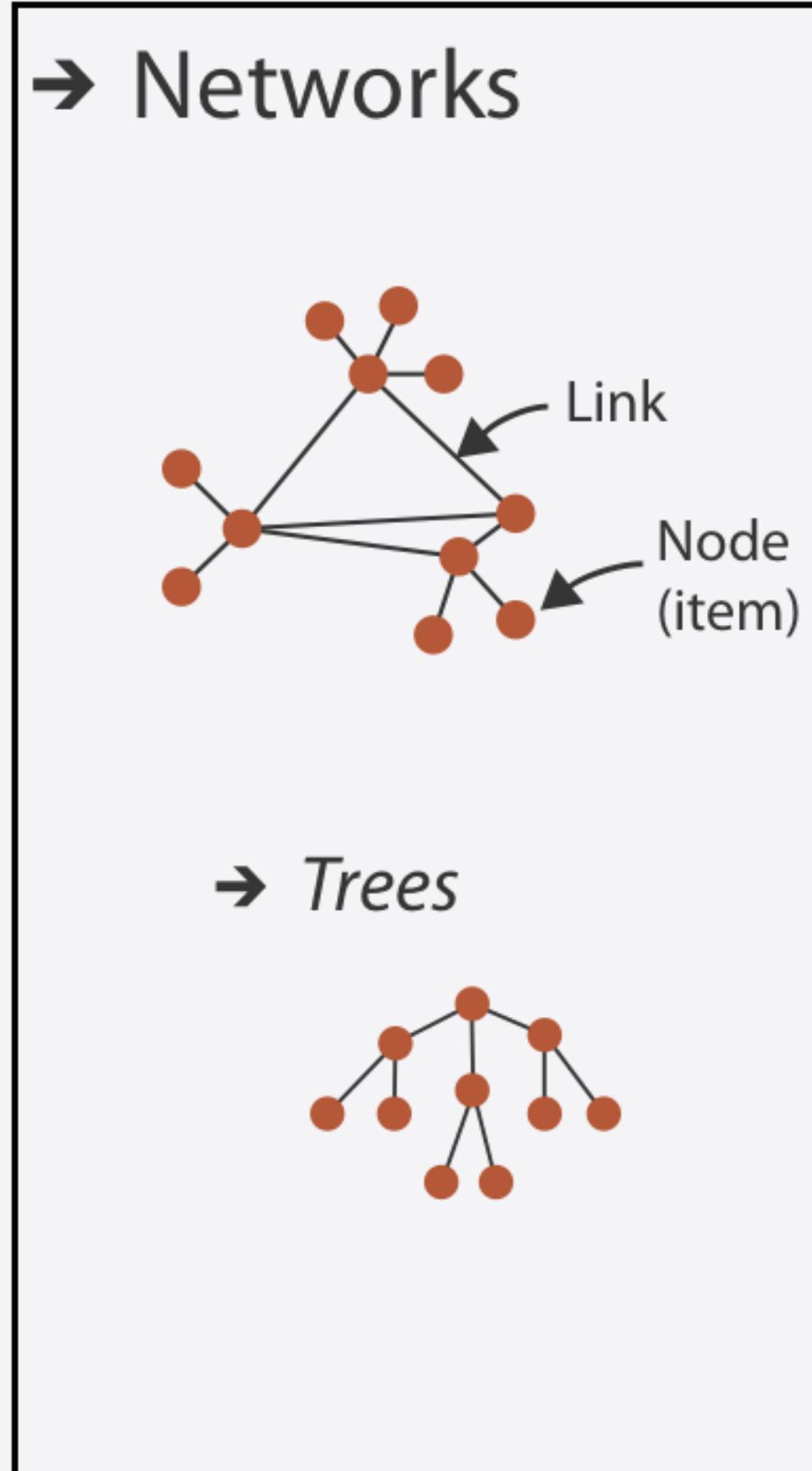
Redes y Árboles



- En una **red** las observaciones son *nodos*. Cada nodo puede estar conectado con otros en una cantidad arbitraria.
- Las conexiones se llaman *links*, y pueden tener atributos como peso y label.
- Los nodos además pueden tener otros atributos derivados de las características de la red: degree, centrality, transitivity, etc.
- Redes pueden ser dirigidas o no dirigidas. Los árboles son redes con una estructura jerárquica (root, branches, leaves)



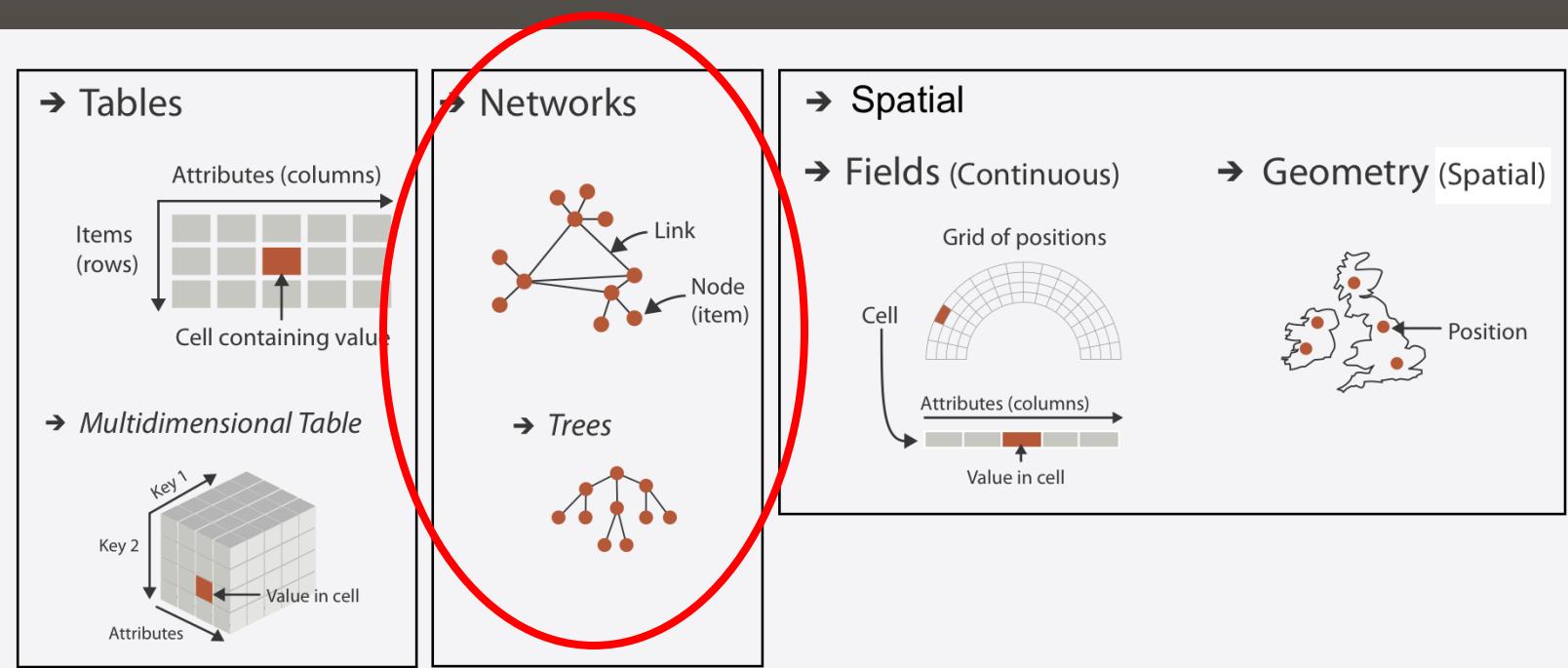
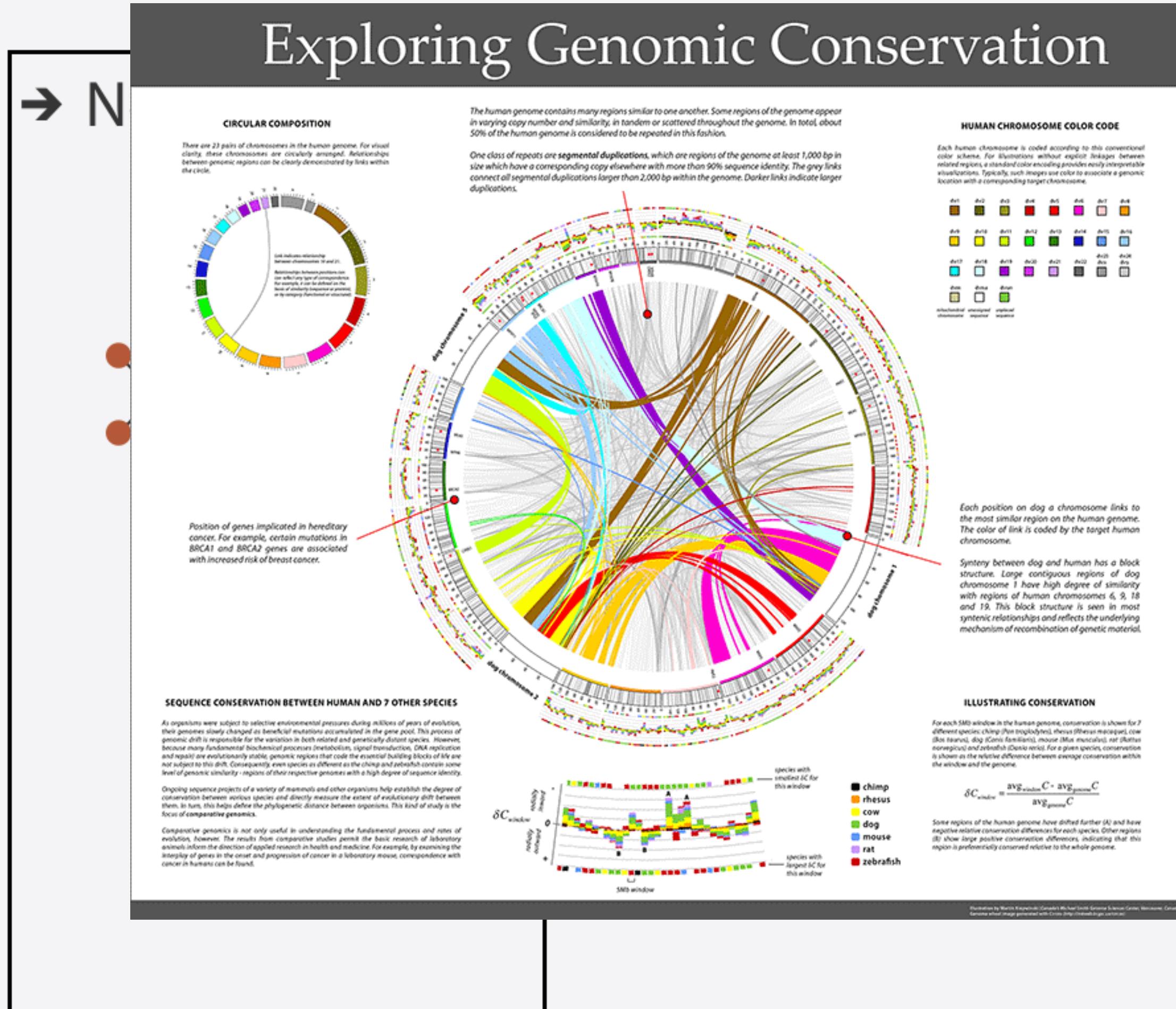
Redes y Árboles



EJEMPLOS
Autodesk organization scheme

<https://youtu.be/mkJ-Uy5dt5g>

Redes y Árboles

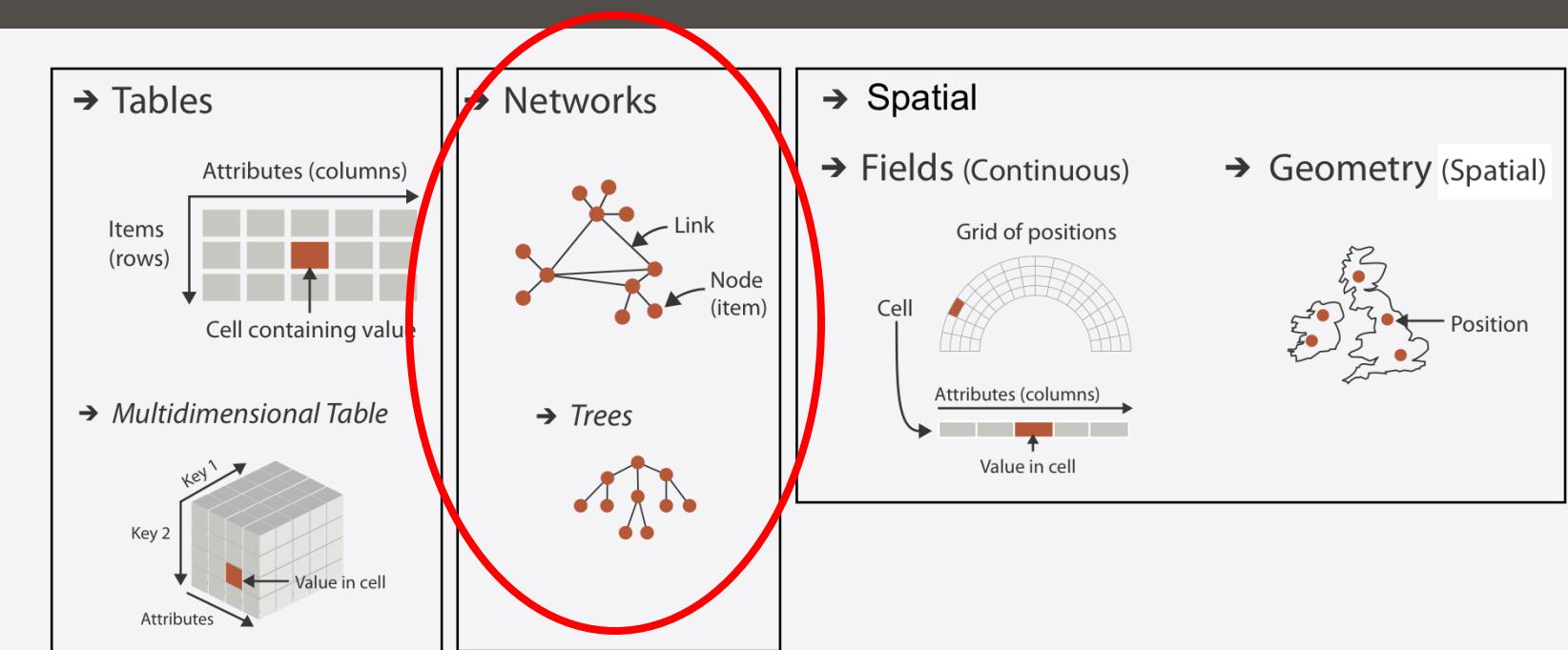


- Otra forma de representar grafos son los diagramas circulares.
- Populares en genética por el software Circos, ideado para representar conexiones.
- También se usa fuera de ese ámbito.

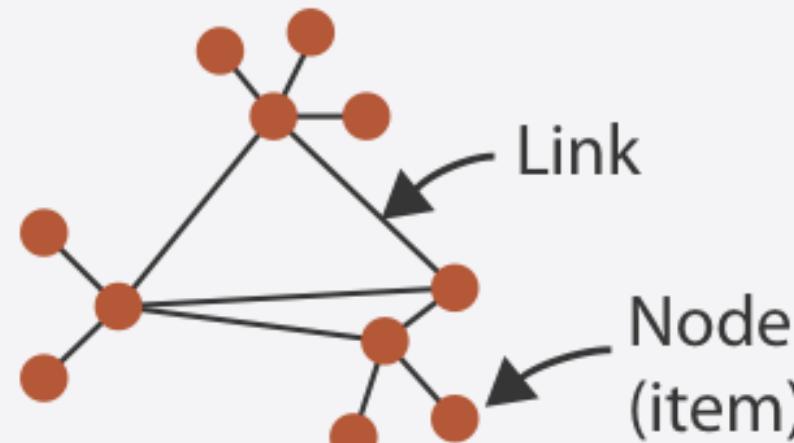
EJEMPLOS

- **Circos**
- **The Global Flow of People**
http://download.gsb.bund.de/BIB/global_flow/
- **The Human Connectome**
<https://www.bsc.es/viz/connectome/>

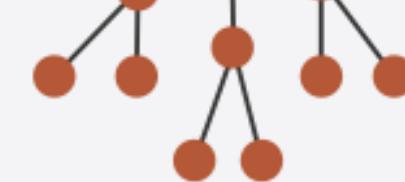
Redes y Árboles



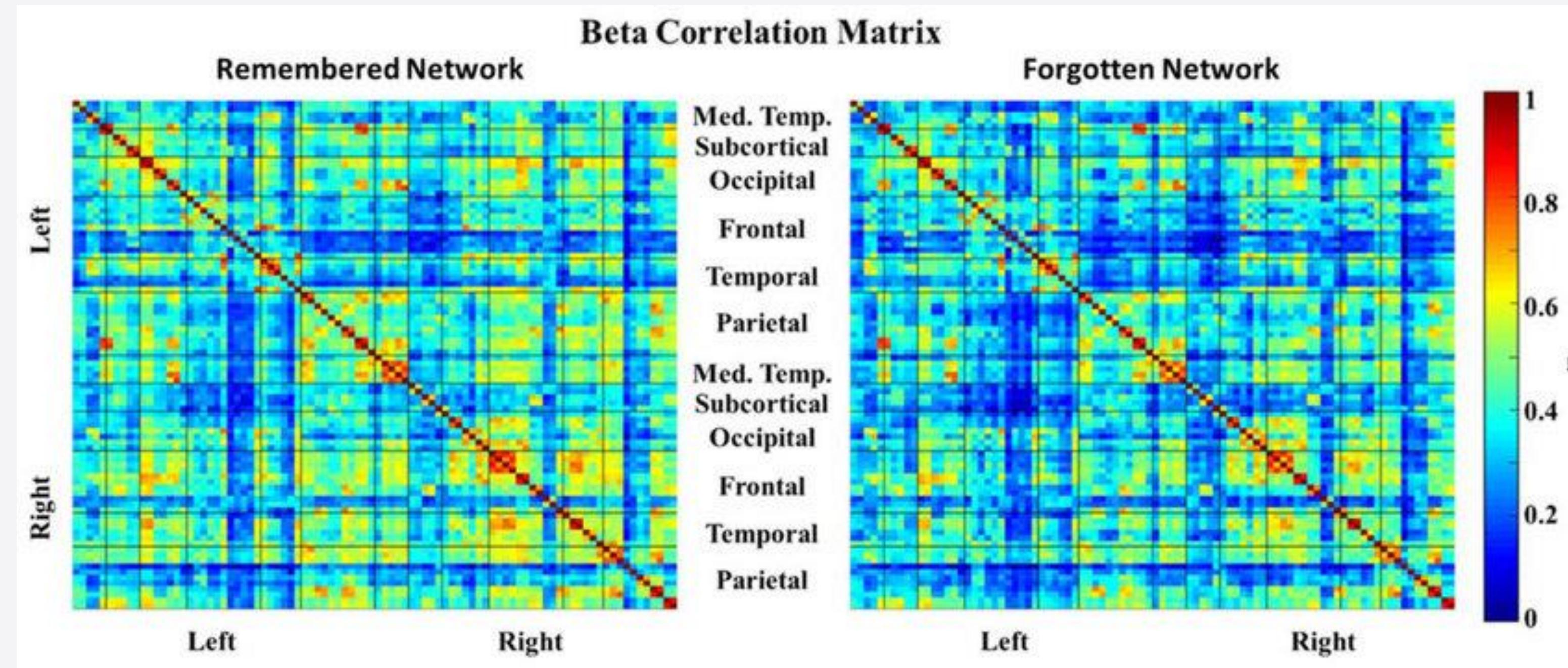
→ Networks



→ Trees

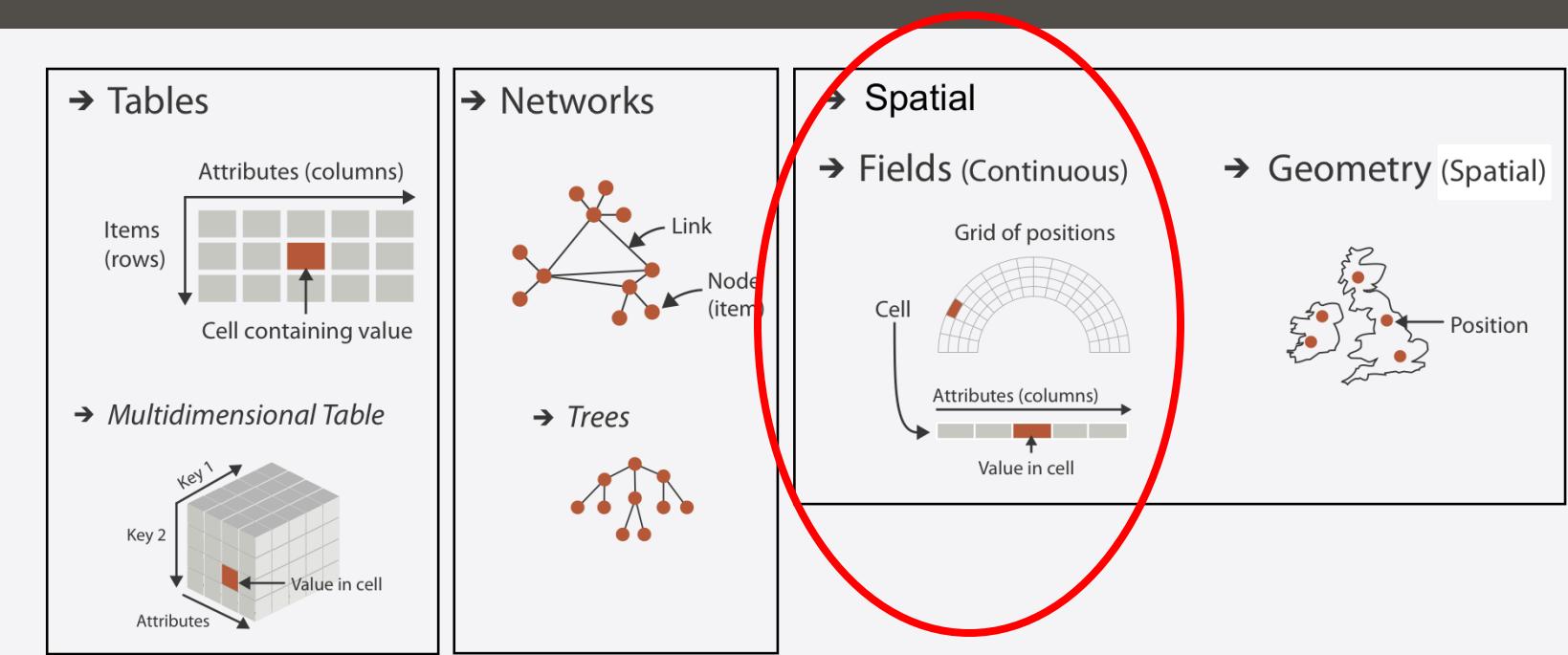


- Otra forma de representar redes son heatmaps de matrices de adyacencia.



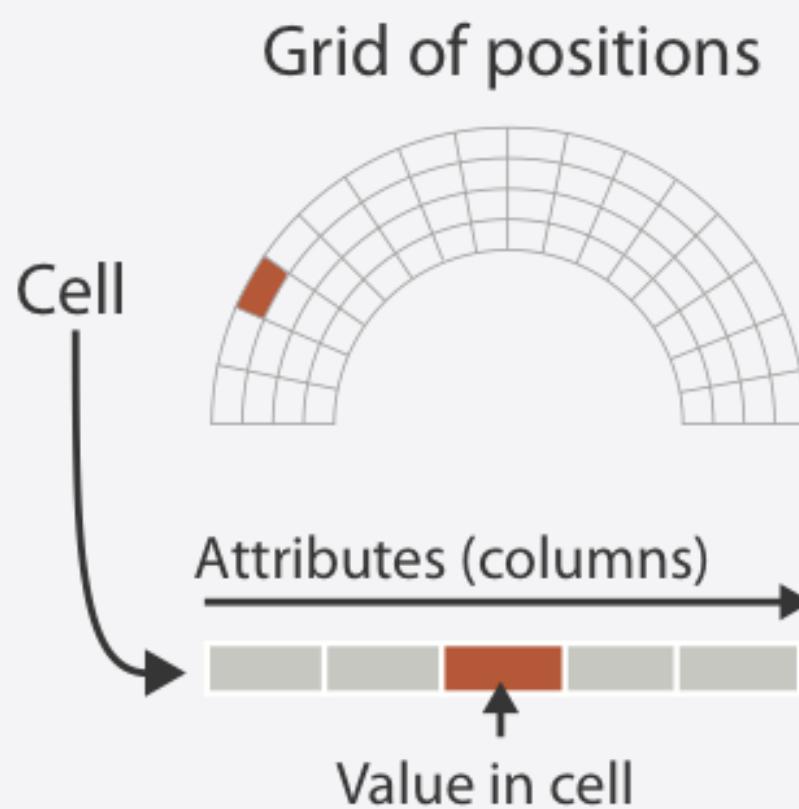
From hippocampus to whole-brain: The role of integrative processing in episodic memory retrieval

Datasets espaciales



→ Spatial

→ Fields (Continuous)



- **Espaciales:**
 - Datos con una posición inherente en el espacio
- **Campos**
 - Modelan fenómenos continuos en un espacio (infinitos puntos). e.g. luz.
 - Puede ser discretizado en un grid regular o irregular.
 - Usualmente provienen de simulaciones, modelos de fenómenos físicos, sensores, etc.
 - Muchas veces, explorar y entender aspectos de su estructura -sobre todo la forma-, es tan importante como visualizar los atributos

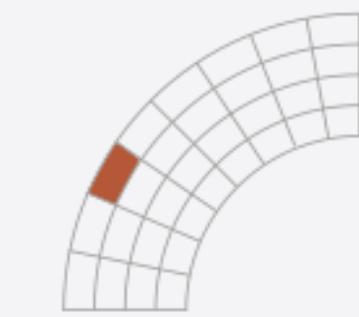
Campos

→ Spatial

→ Fields (Continuas)

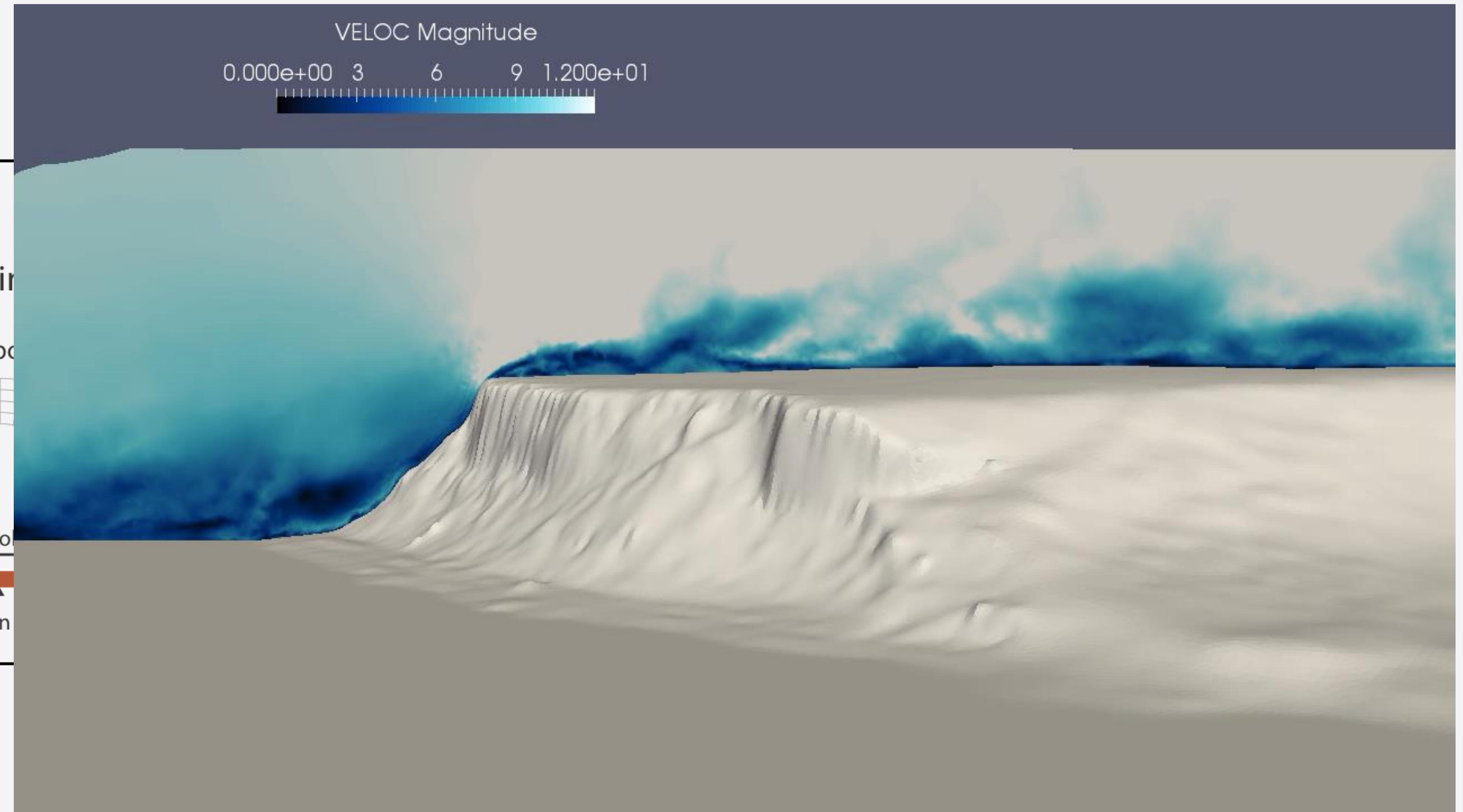
Grid of po

Cell



Attributes (color)

Value in



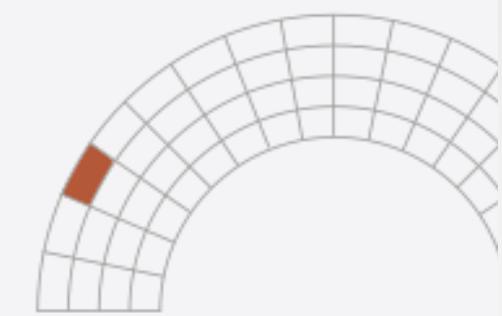
Campos

→ Spatial

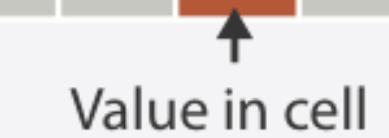
→ Fields (Continuous)

Grid of positions

Cell



Attributes (column)



Value in cell

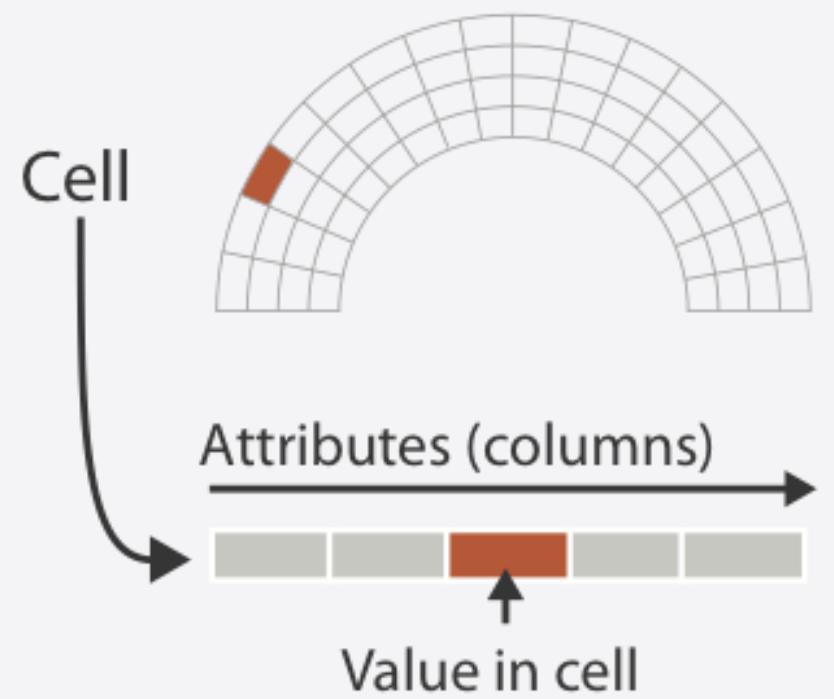


Campos

→ Spatial

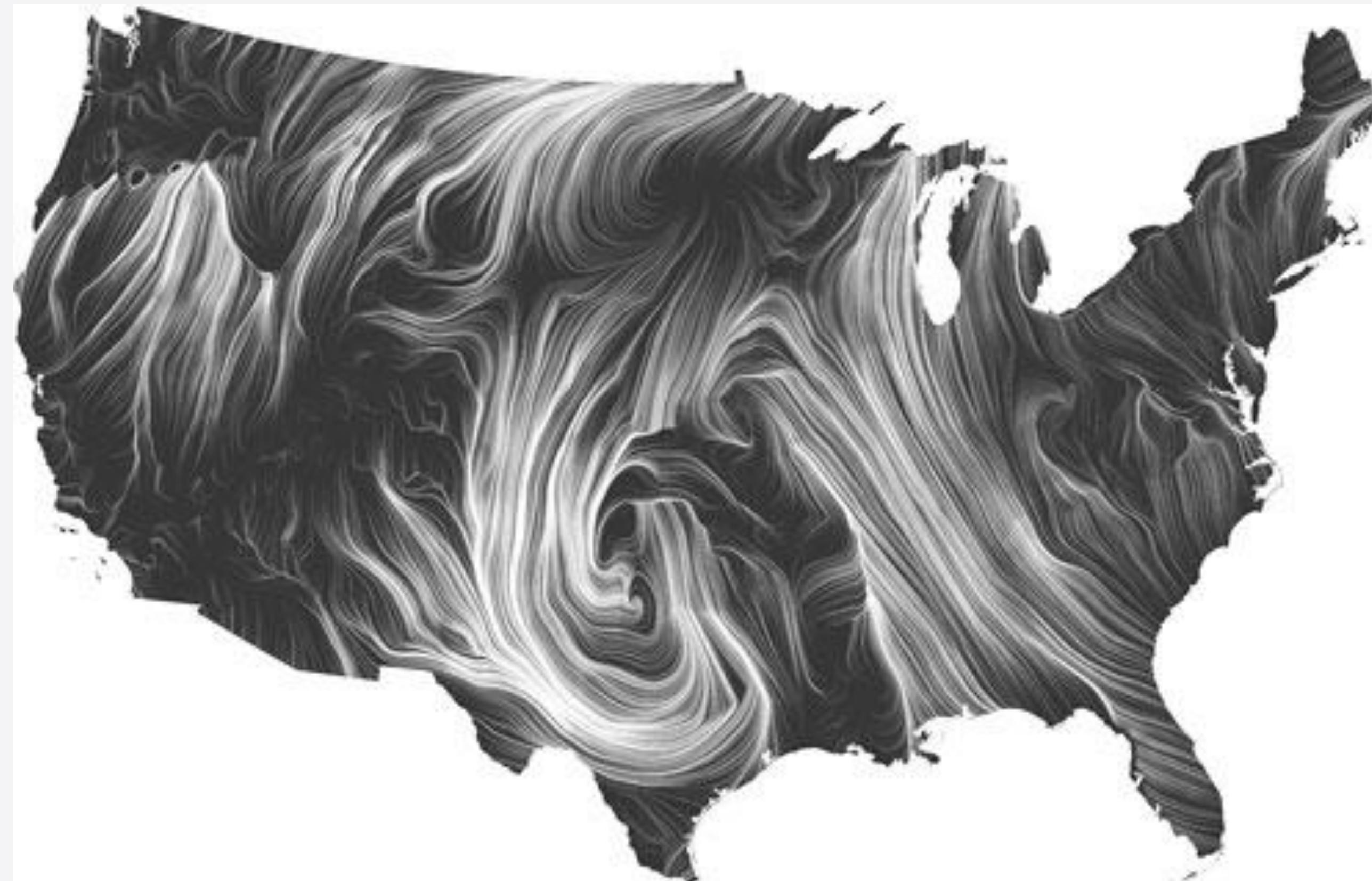
→ Fields (Continuous)

Grid of positions



- Wind map – Viegas & Wattenberg

<http://hint.fm/wind/>



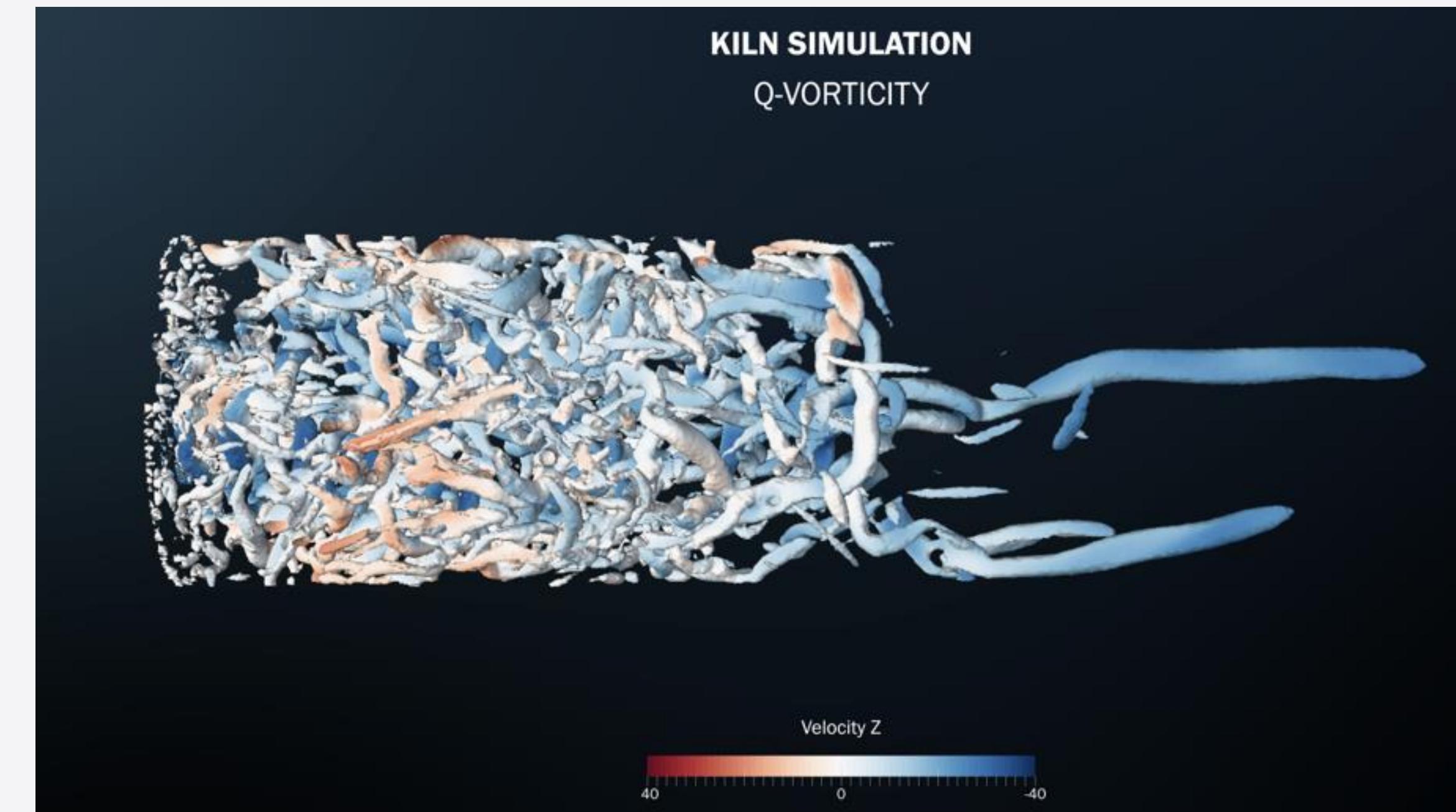
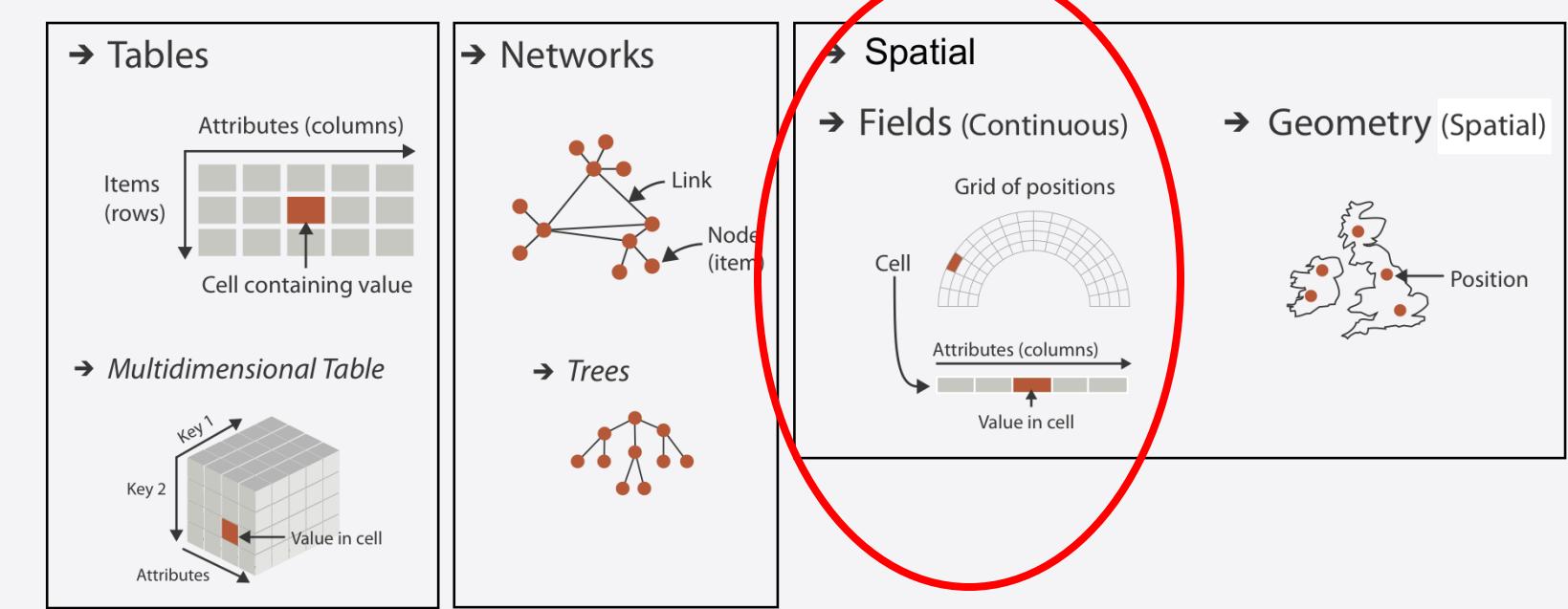
Geometría

→ Spatial

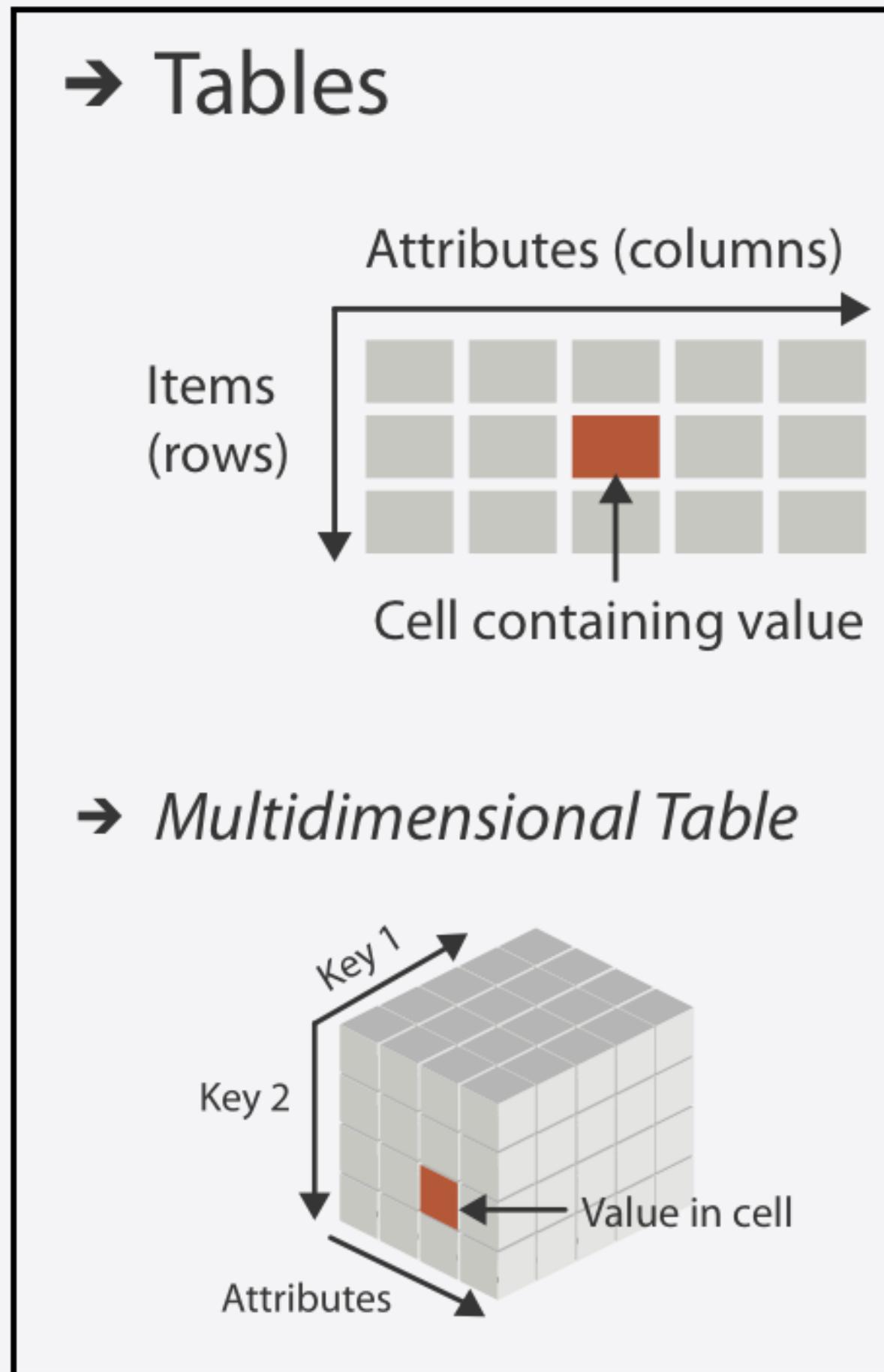
→ Geometry (Spatial)



- La forma se especifica a través de la posición espacial.
- No tienen necesariamente atributos.
- Pueden ser solamente el marco sobre el que se sobrepone información adicional de otras fuentes (mapa)
- También puede ser geometría derivada de un campo espacial (e.g., isosuperficies o contornos)



Tablas

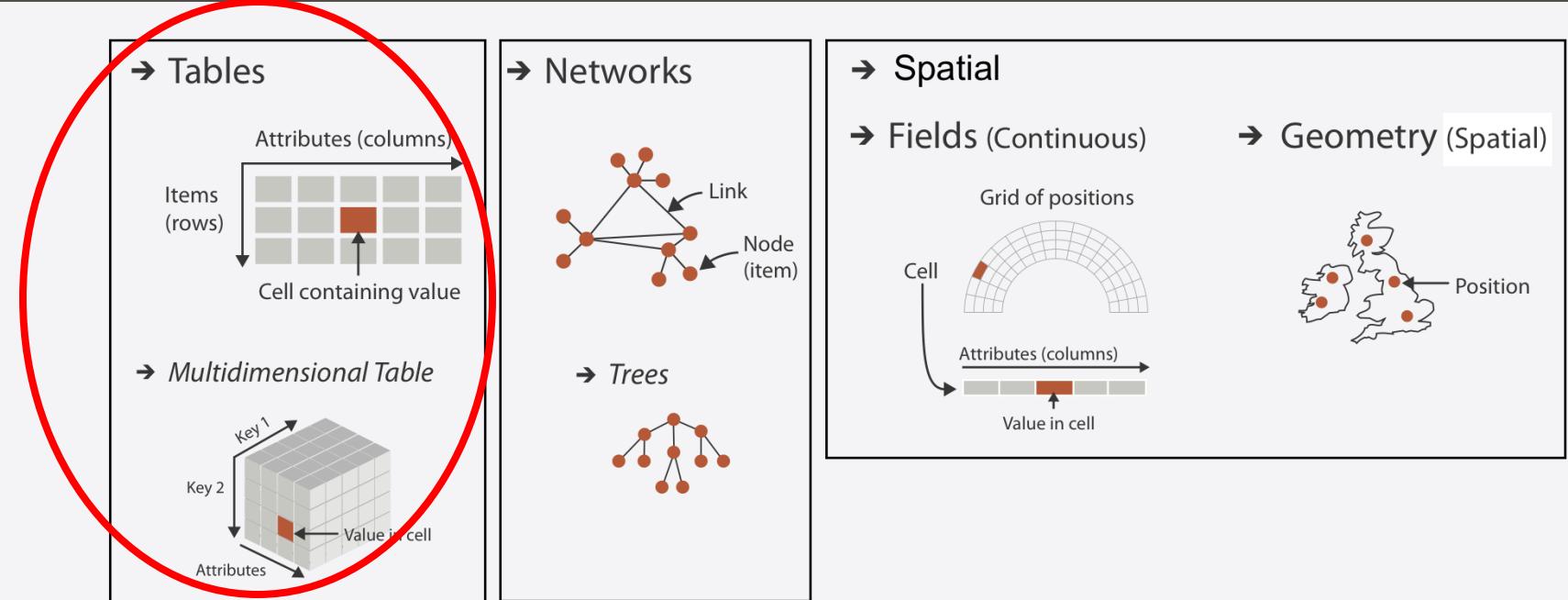


Filas

-> Items / Observaciones

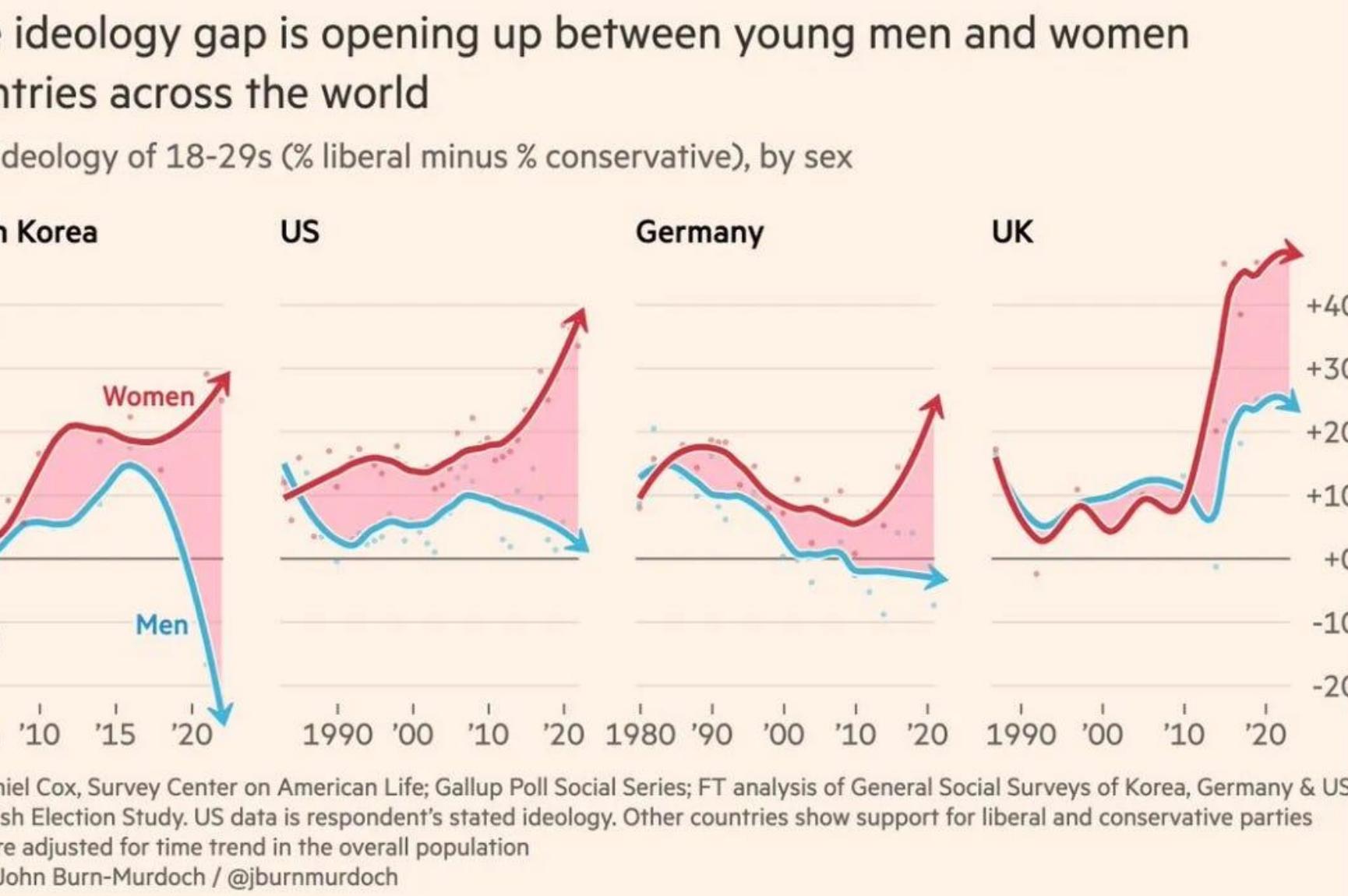
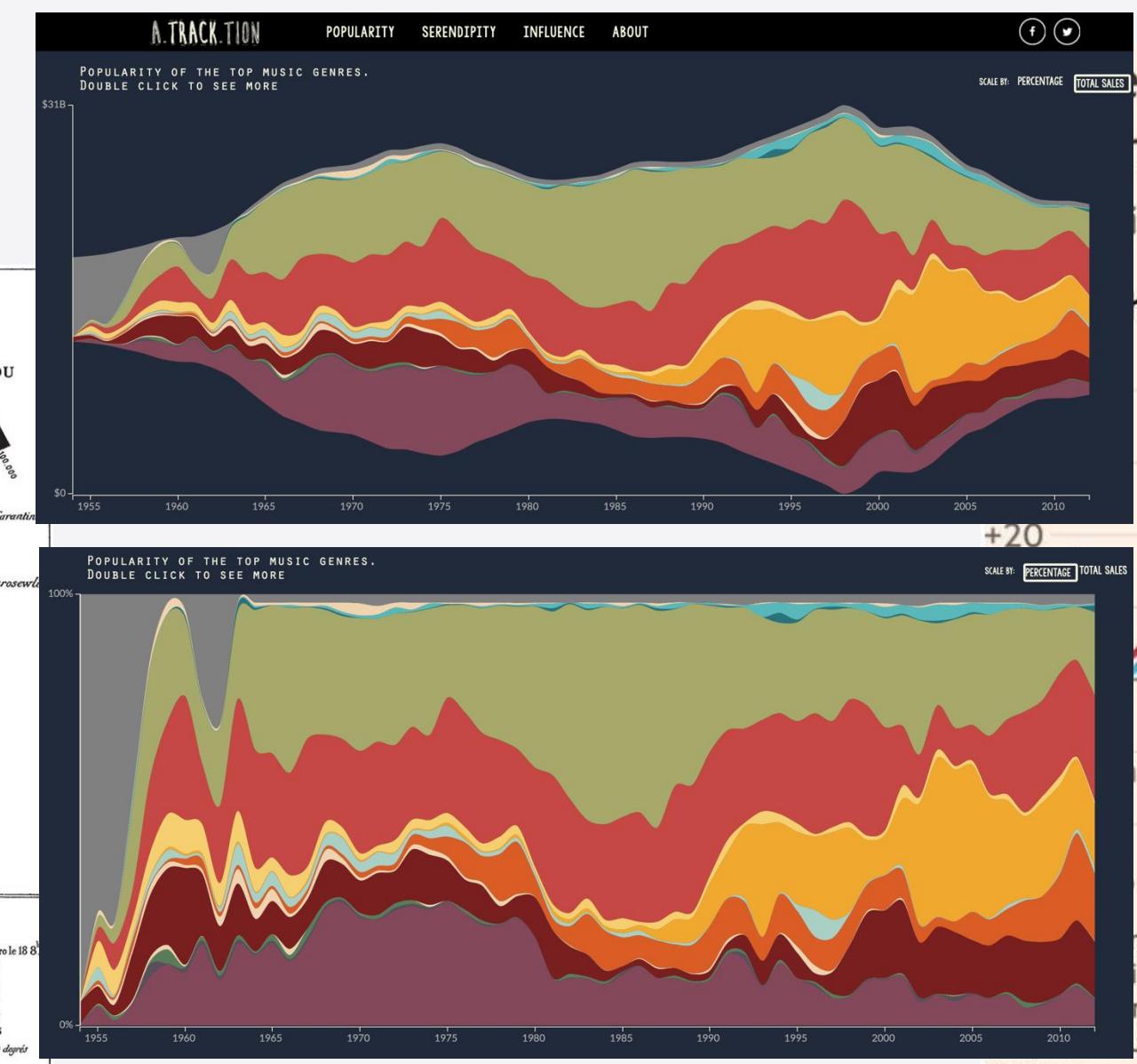
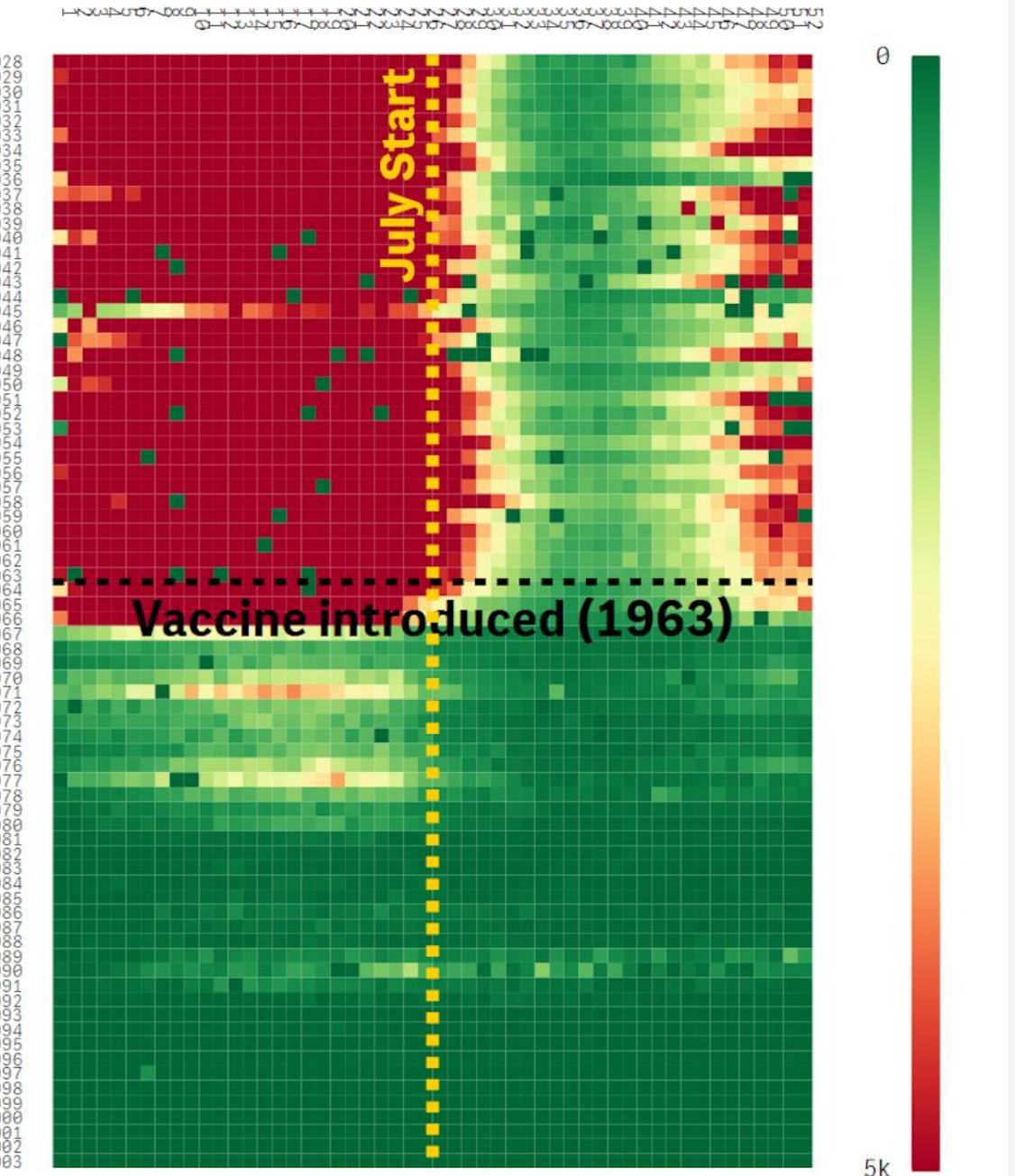
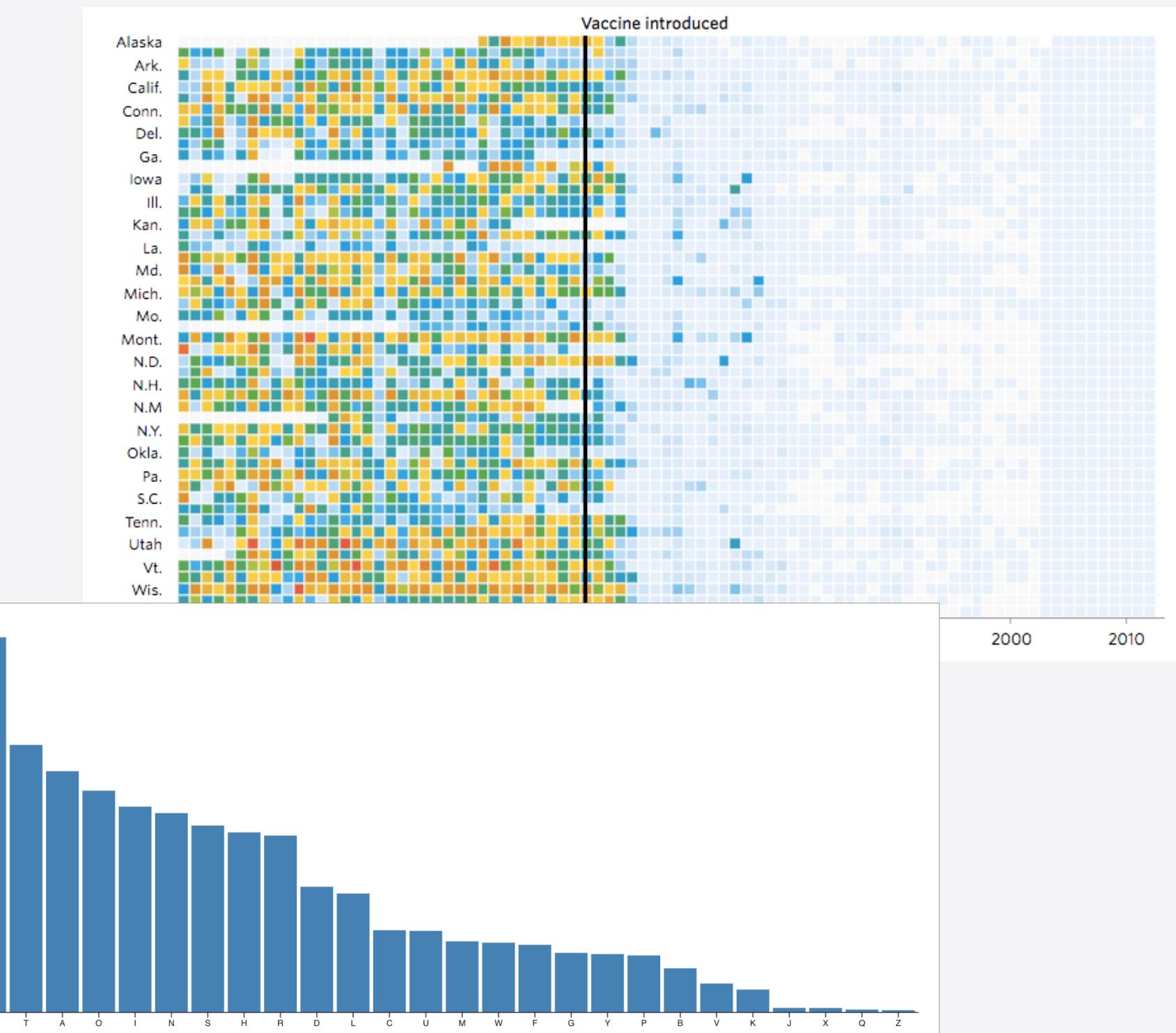
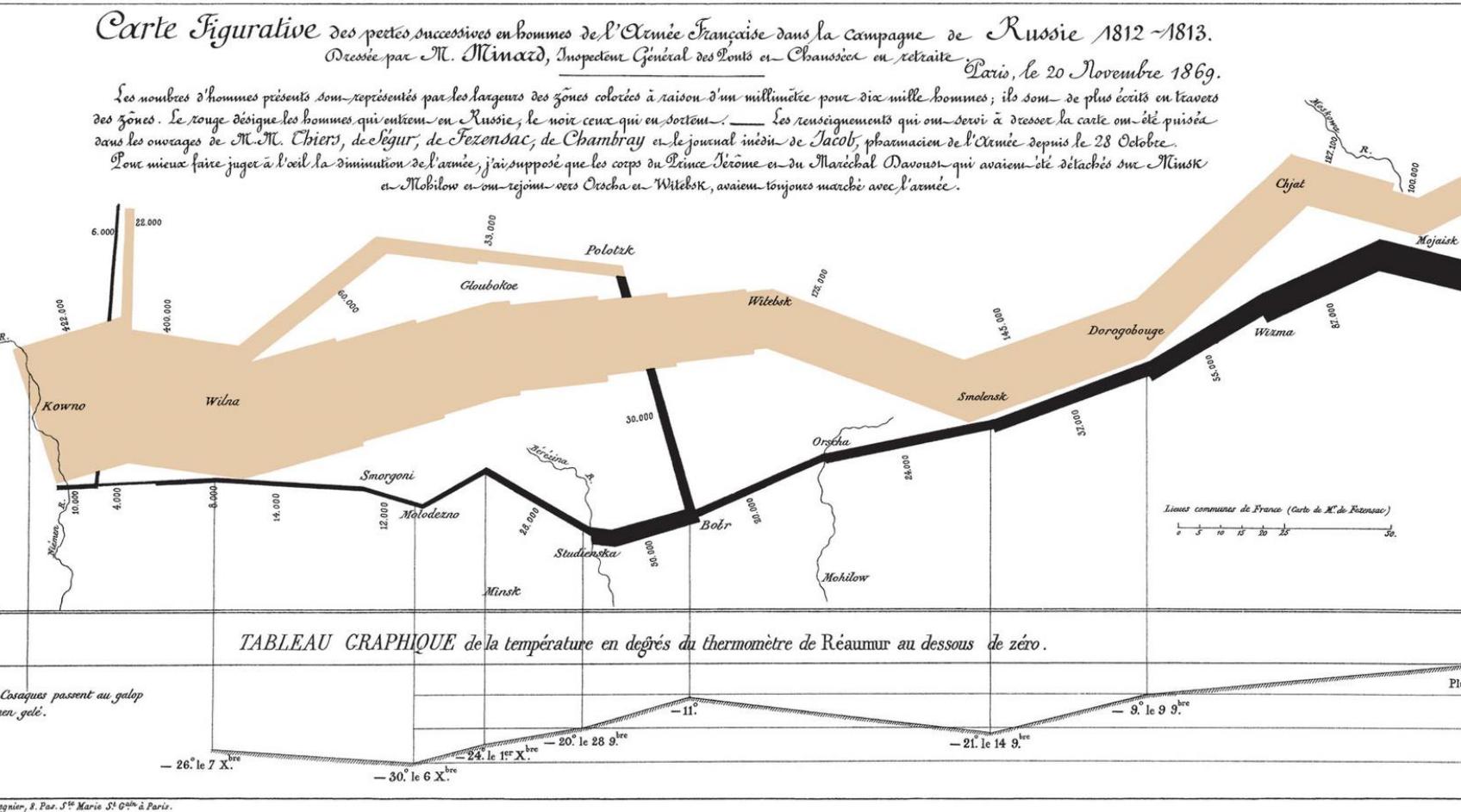
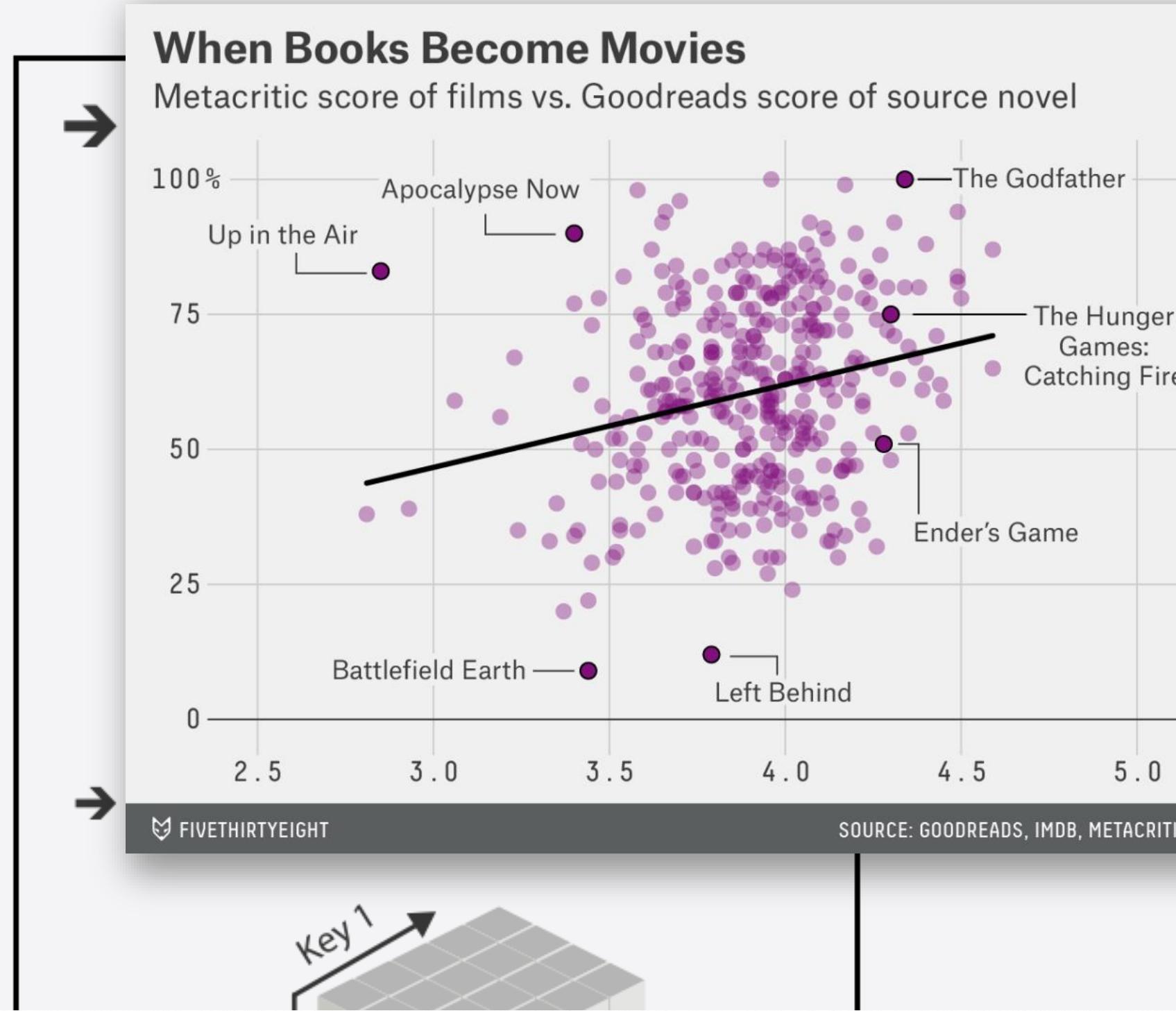
Columnas -> Atributos (AKA variables)

Celdas -> Valores

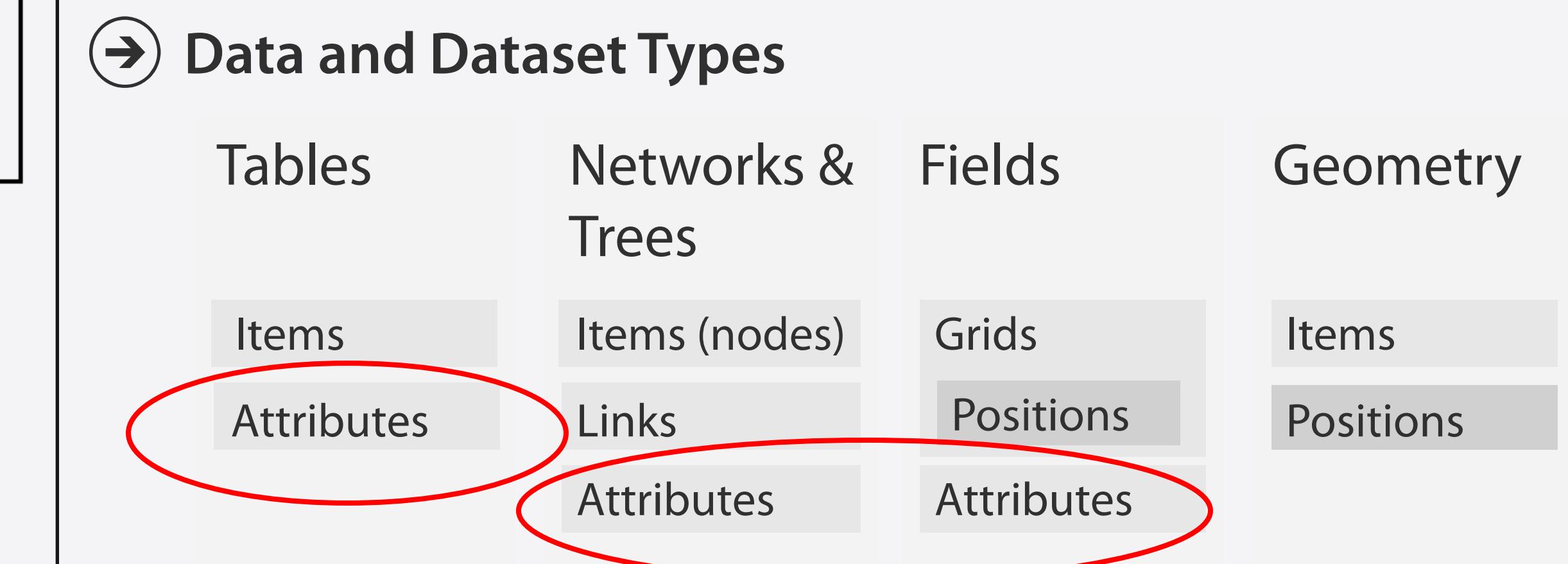
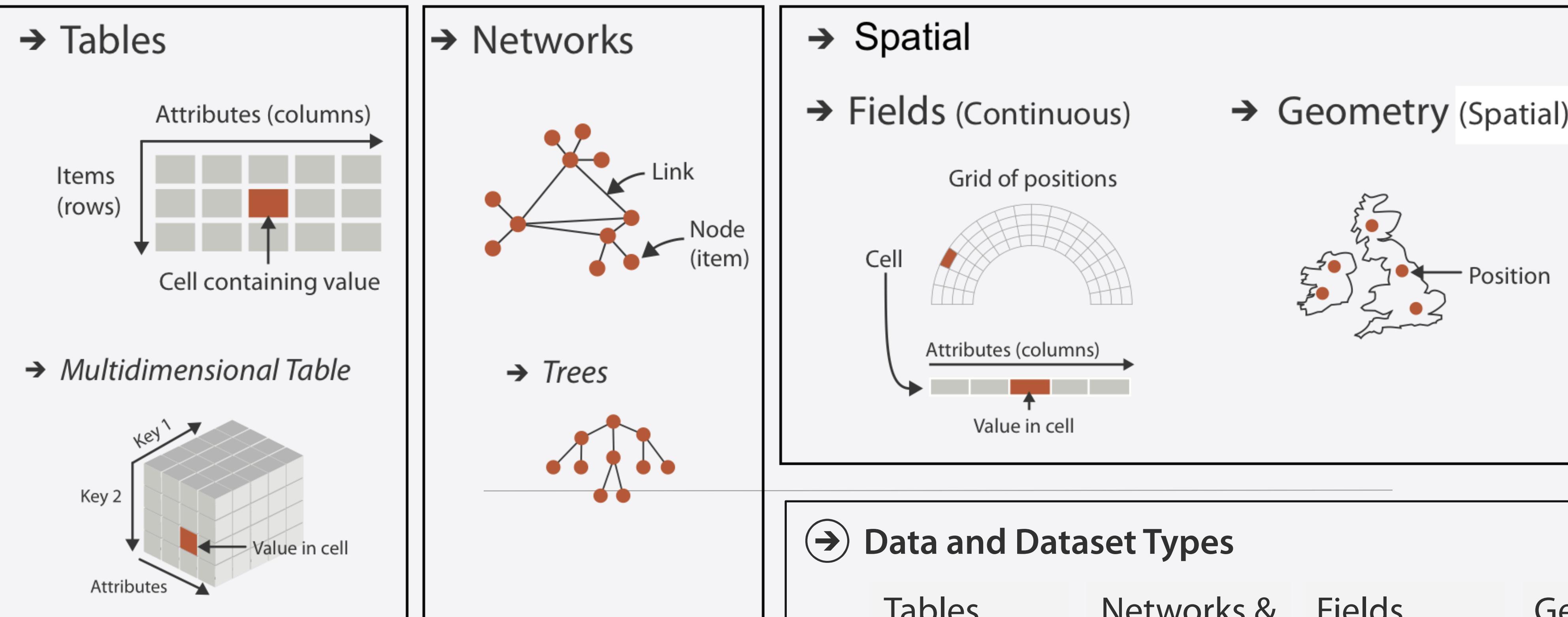


Year	Afghanistan	Haiti	Japan	Spain	Switzerland	World Average		
2000	54,8	58,6	81,1	79,1	79,7	66,3		
2001	55,3	58,9	81,5	79,4	80,2	66,7		
2002	56,2	59,3	81,8	79,5	80,4	66,9		
2003	56,7	59,7	81,9	79,4	80,5	67,1		
2004	57	58,7	82,1	80,1	81	67,4		
2005	57,3	60,5	82	80,1	81,1	67,8		
2006	57,3	61,1	82,4	80,8	81,5	68,6		
2007	57,5	61,8	82,6	80,9	81,7	68,7		
2008	58,1	62,1	82,7	81,3	82	69		
2009	58,6	62,5	83	81,6	82,1	69,4		
2010	58,8	36,3	83	81,9	82,3	69,6		
2011	59,2	62,3	82,5	82,1	82,6	70,2		
2012	59,5	62,3	83,3	82	82,7	70,4		
2013	59,9	62,7	83,5	82,4	83	70,7		
2014	59,9	63,1	83,5	82,6	83,2	71		
2015	60,5	63,5	83,7	82,8	83,4	71,3		

Tablas



Tipos de Datasets



Tipos de Atributos

➔ Attribute Types

→ Categorical



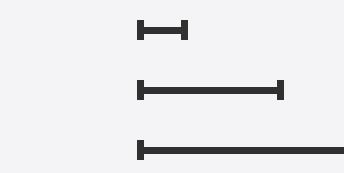
→ Ordered



→ Ordinal



→ Quantitative



➔ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



➔ Data and Dataset Types

Tables

Items

Attributes

Networks & Trees

Items (nodes)

Links

Attributes

Fields

Grids

Positions

Attributes

Geometry

Items

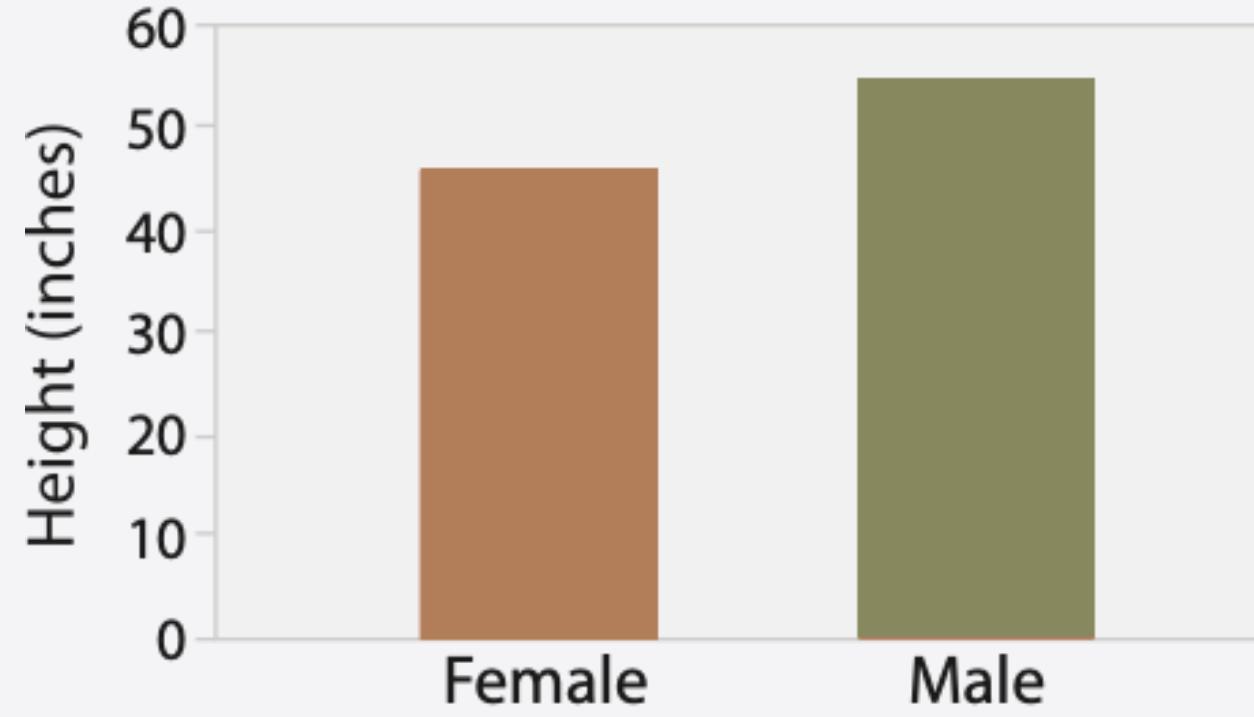
Positions

Clusters, Sets, Lists

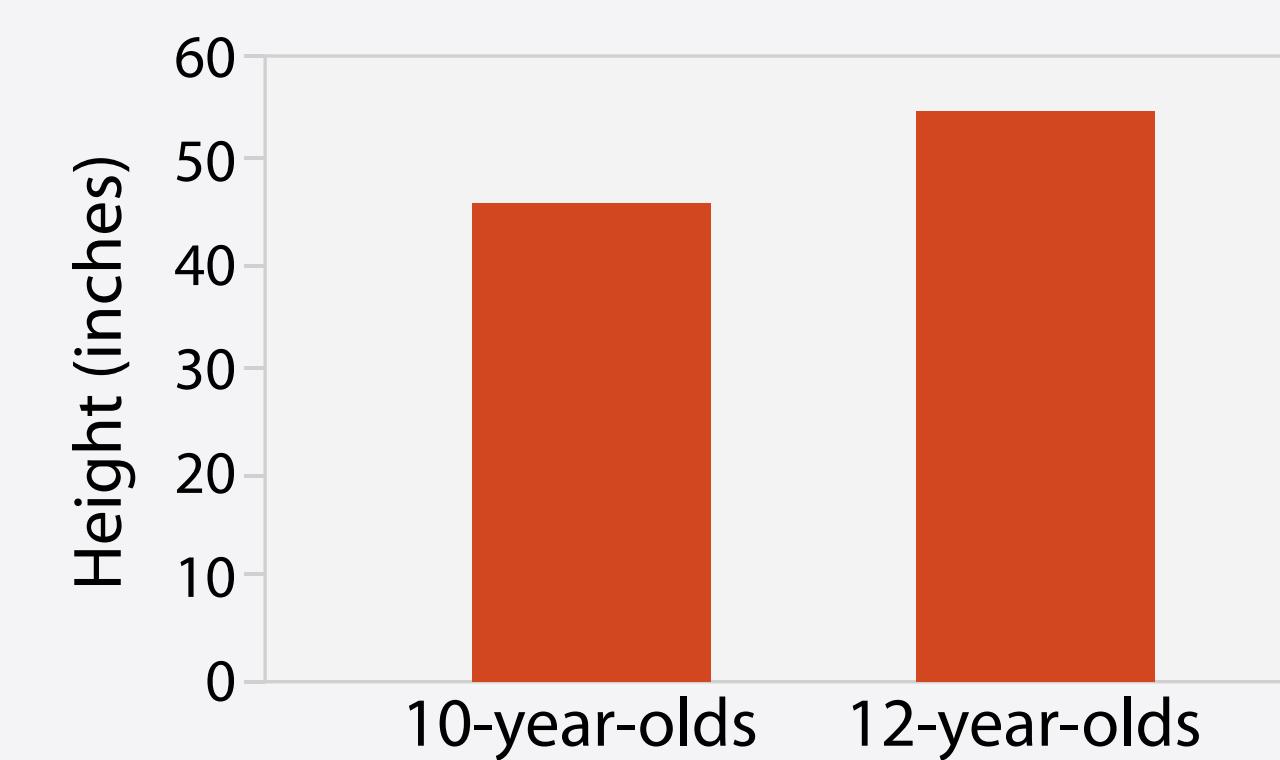
Items

- Distinción crucial para diseñar una visualización porque determina varias de las decisiones que podemos tomar

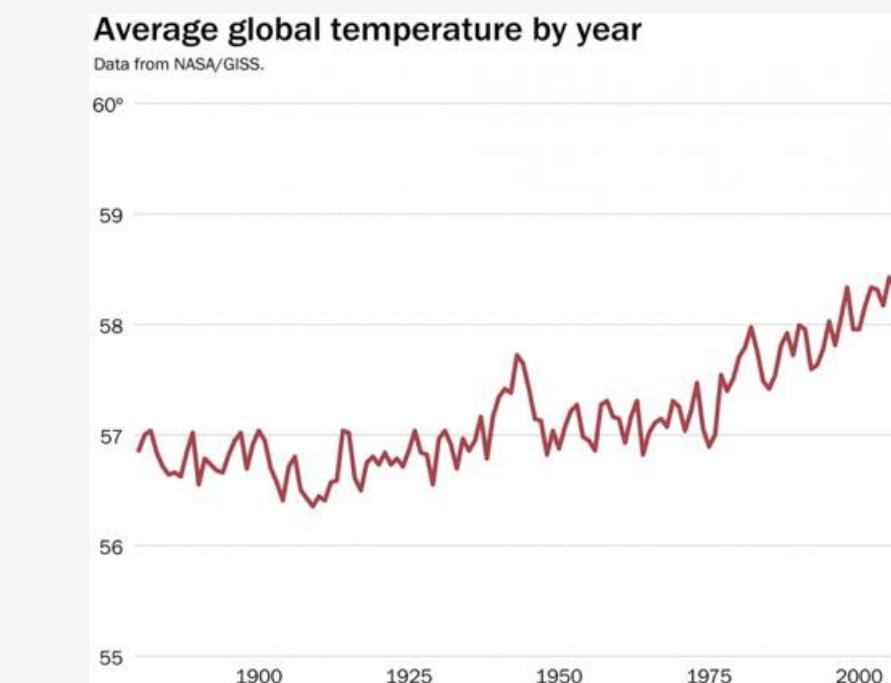
Categóricos



Ordinales

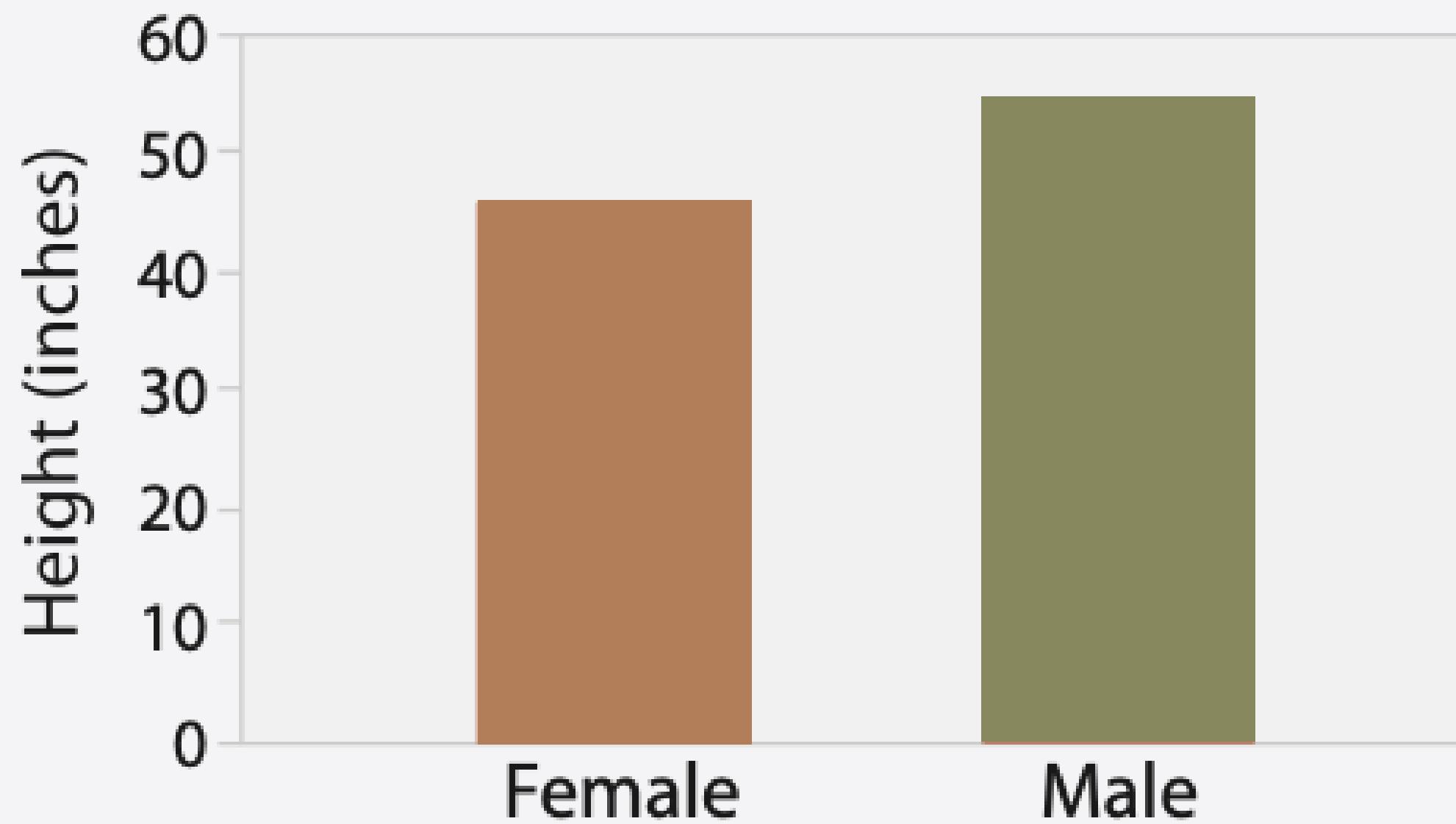


Cuantitativos

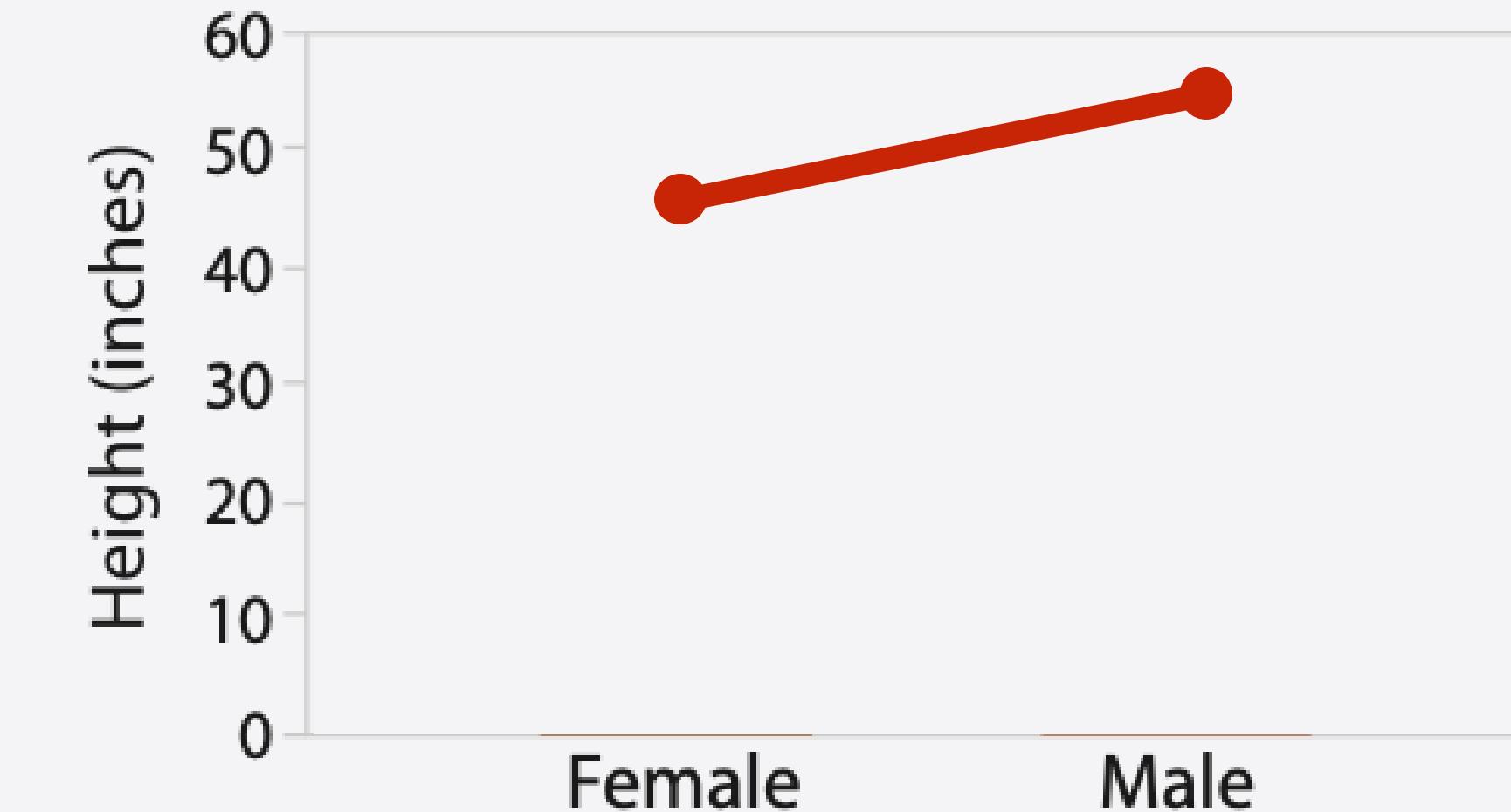


Tipos de Atributos

Categóricos

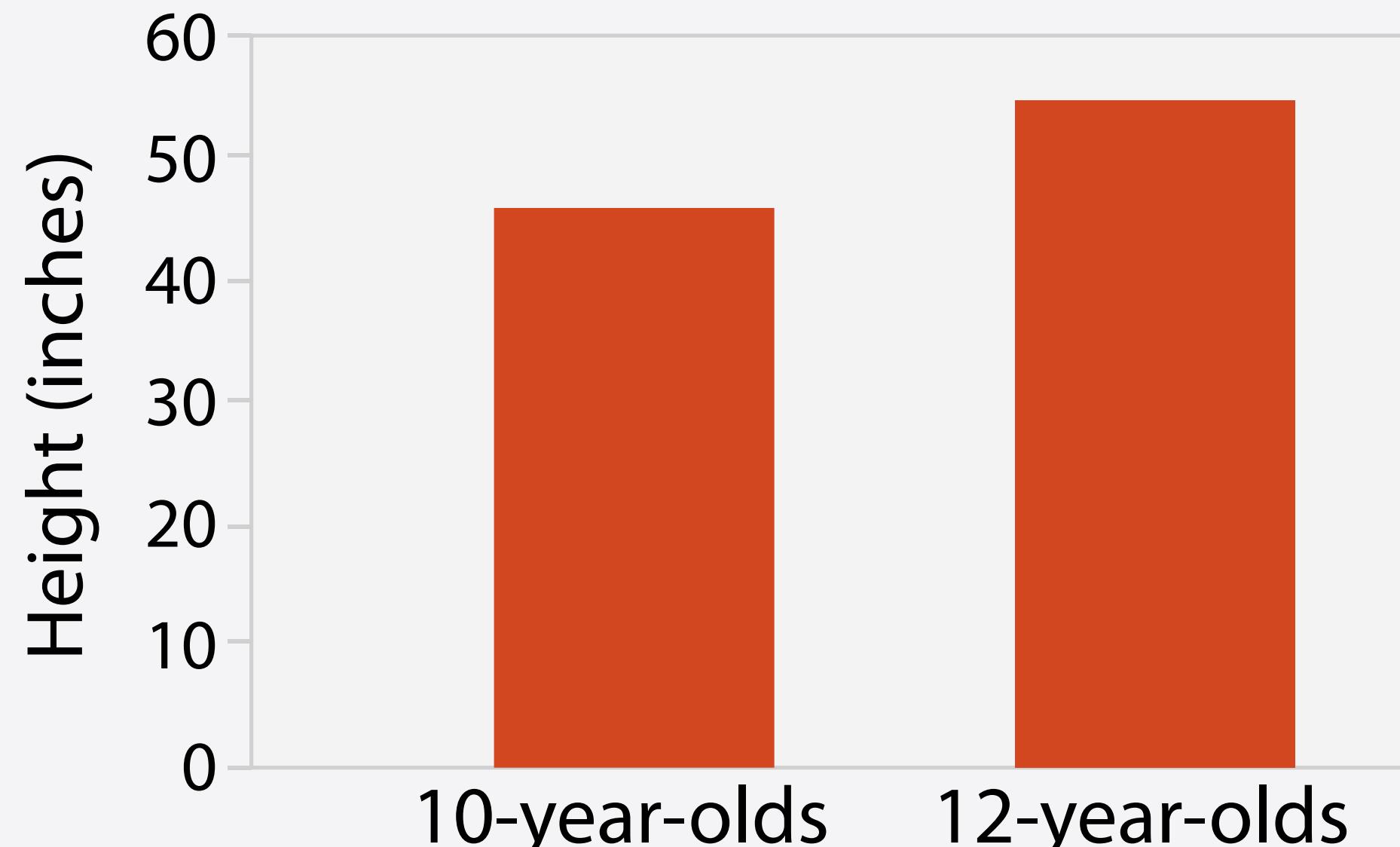


- AKA Nominales
- Sin relación cuantitativa
- Sin orden inherente
- Usar gráficas que expresen pertenencia a distintas categorías; o evitar gráficas que impliquen una relación de continuidad cuando no la hay

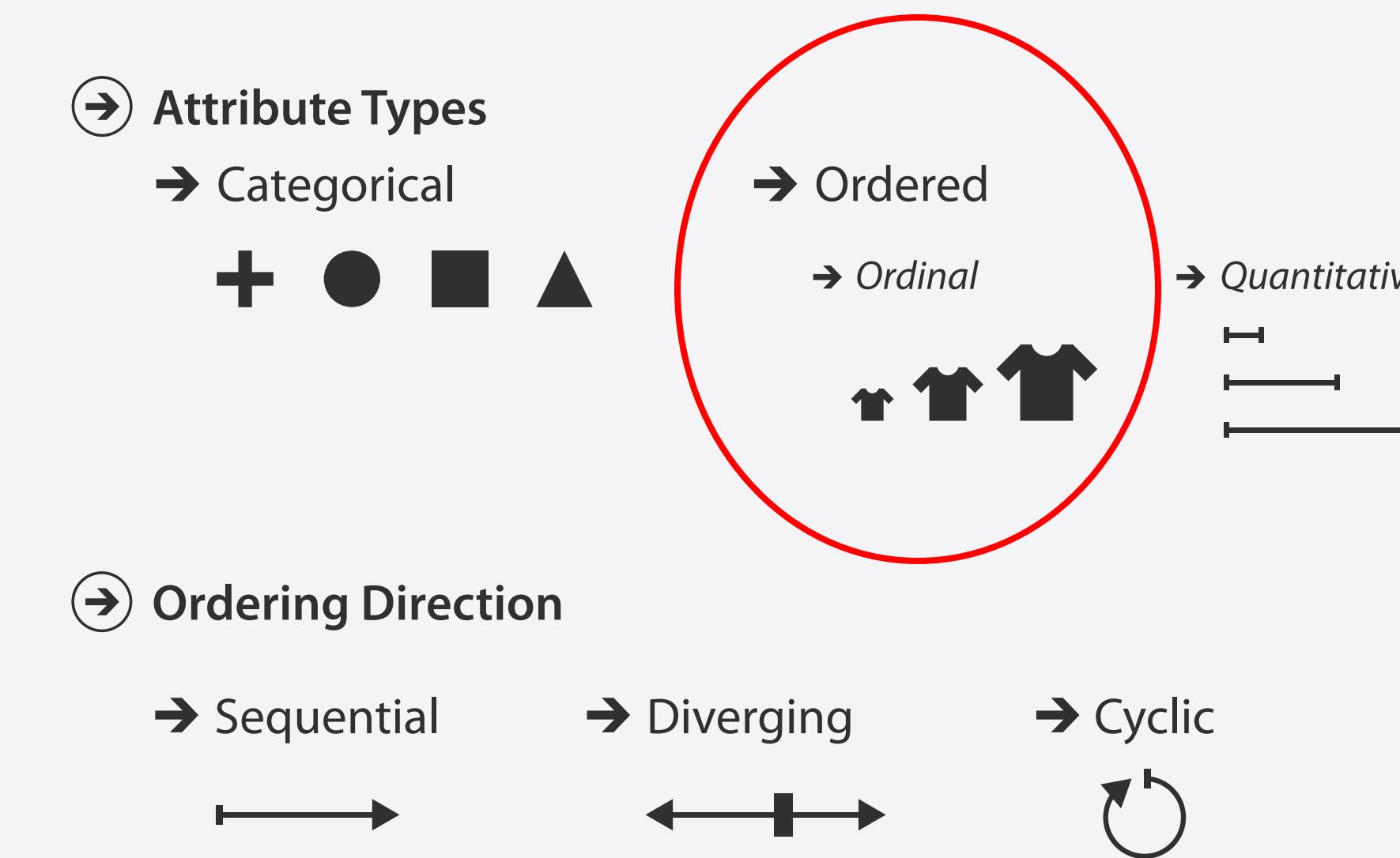


Tipos de Atributos

Ordinales

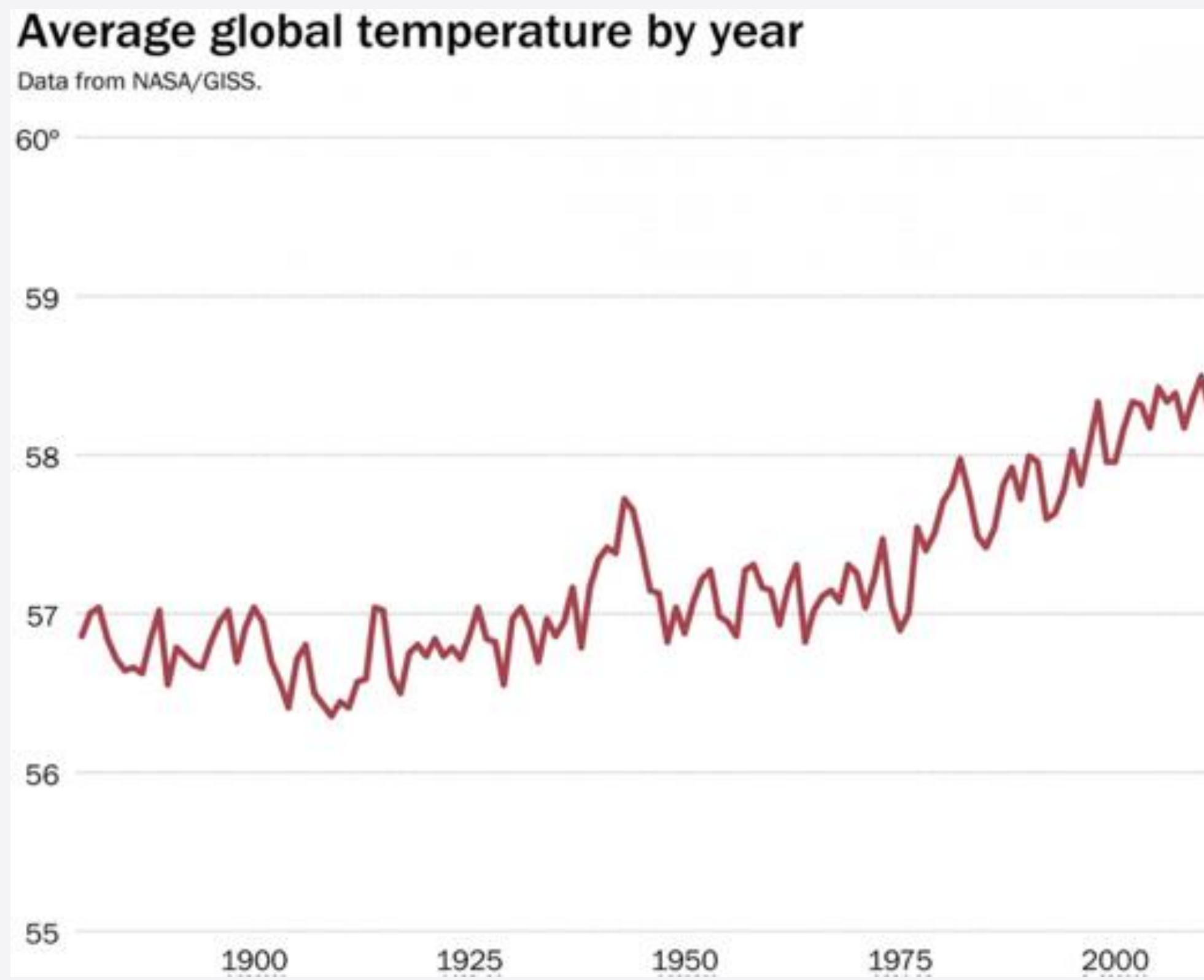


- Se pueden ordenar
- Grado de diferencia no medible

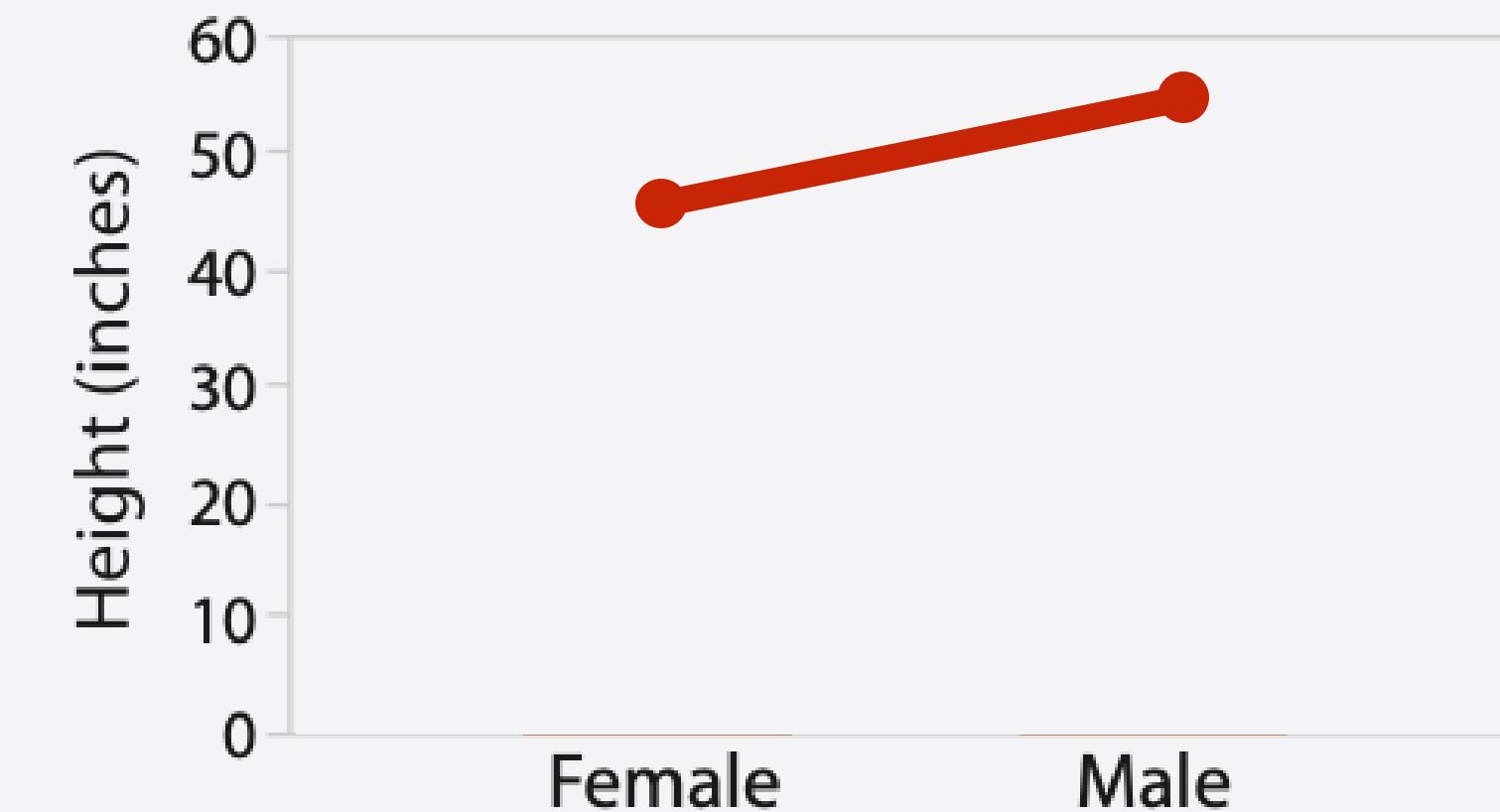


Tipos de Atributos

Cuantitativos



- Se pueden medir y manipular numéricamente
- Línea continua para serie temporal de registros continuos
- Línea continua NO para datos sin continuidad (e.g. categóricos)



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925	
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583	
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05	
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275	
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16	
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125	
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13	
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18	
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225	
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26	
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292	
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6

Tipos de Atributos

Textuales



- Análisis de lenguaje natural, redes sociales, etc.
- Normalmente se grafican métricas derivadas del análisis

Tipos de Atributos

➔ Attribute Types

➔ Categorical

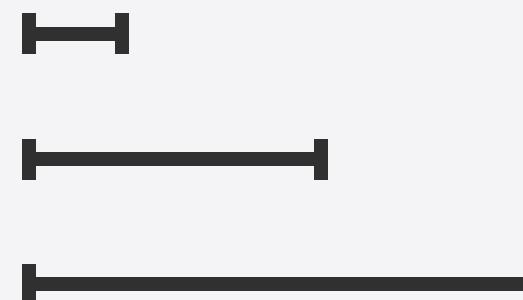


➔ Ordered

➔ *Ordinal*

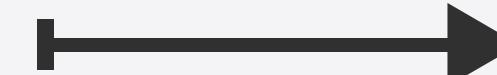


➔ Quantitative



➔ Ordering Direction

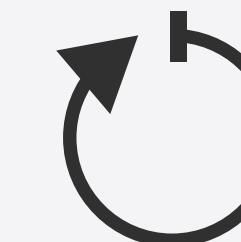
➔ Sequential



➔ Diverging



➔ Cyclic



Dirección de los Atributos

Secuencial

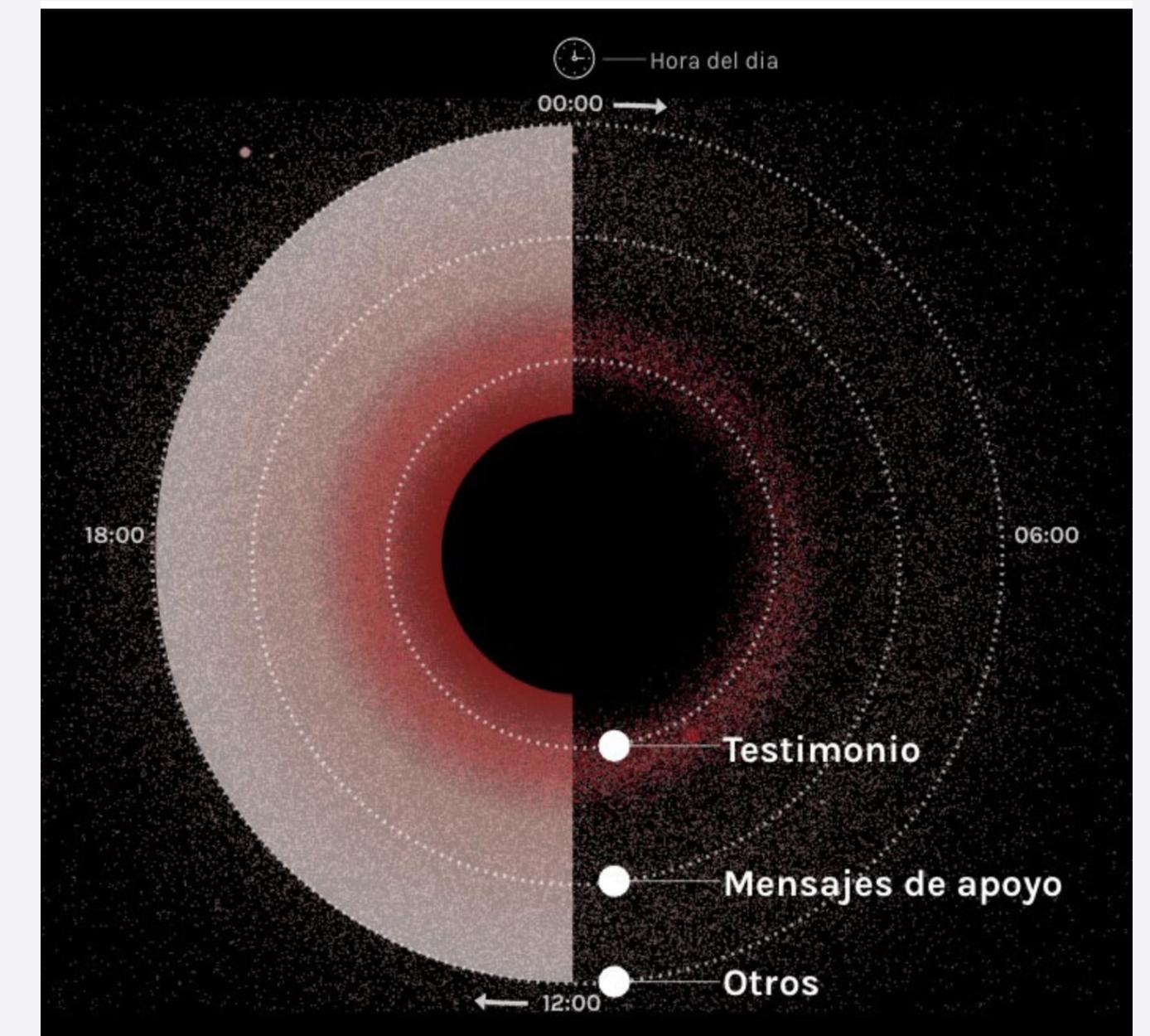
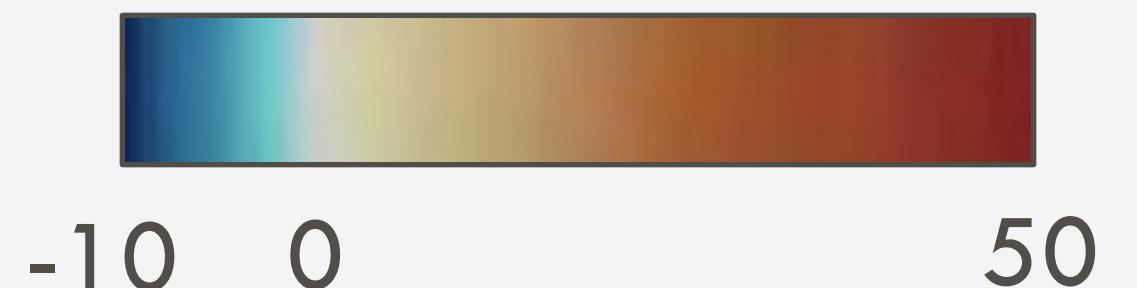
Varía en **una sola dirección** respecto a un punto de origen

Divergente

Varía en **direcciones opuestas** respecto a un punto de origen

Cíclico

Varía en una dirección respecto a un punto de origen y regresa al punto de origen



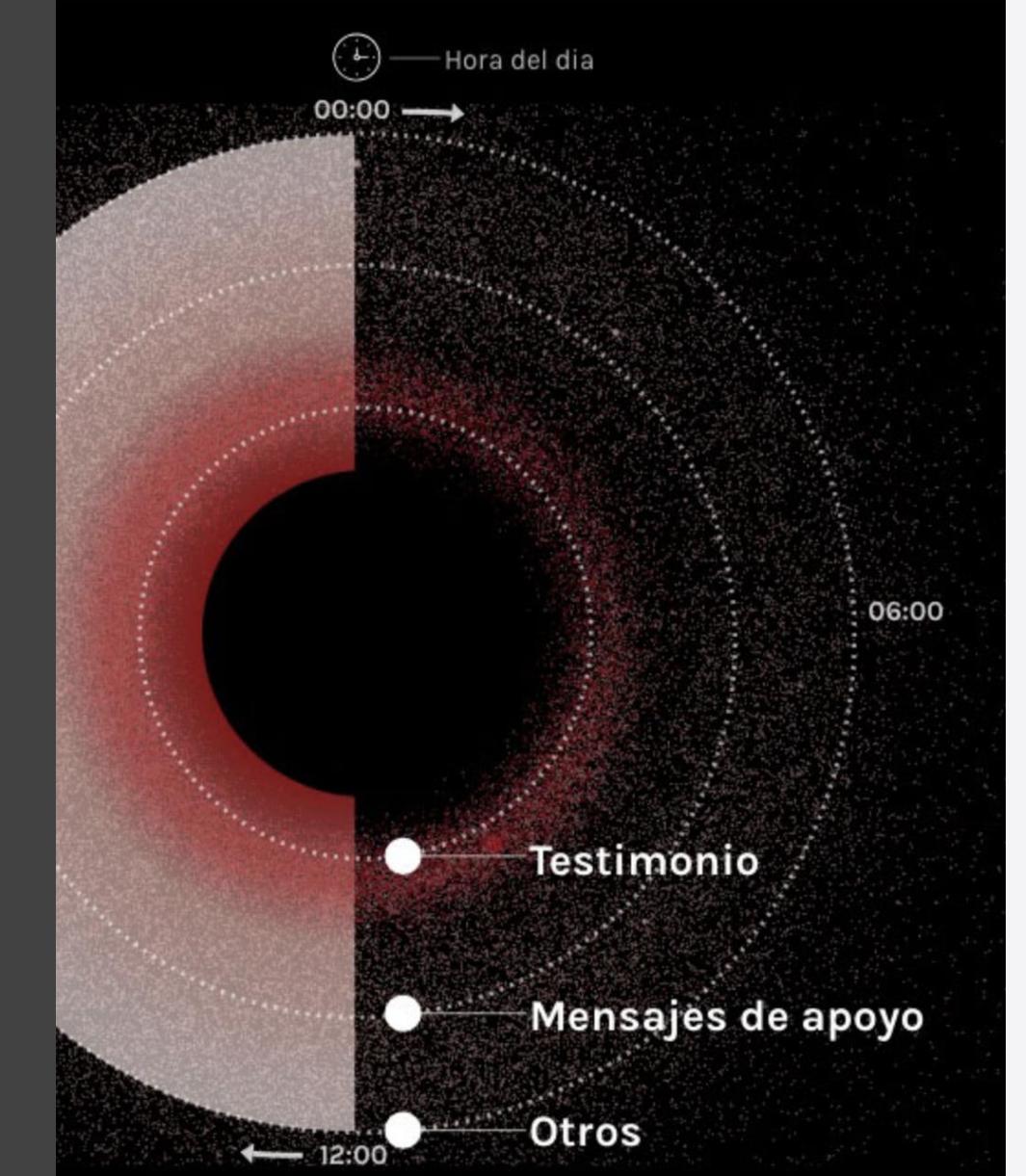
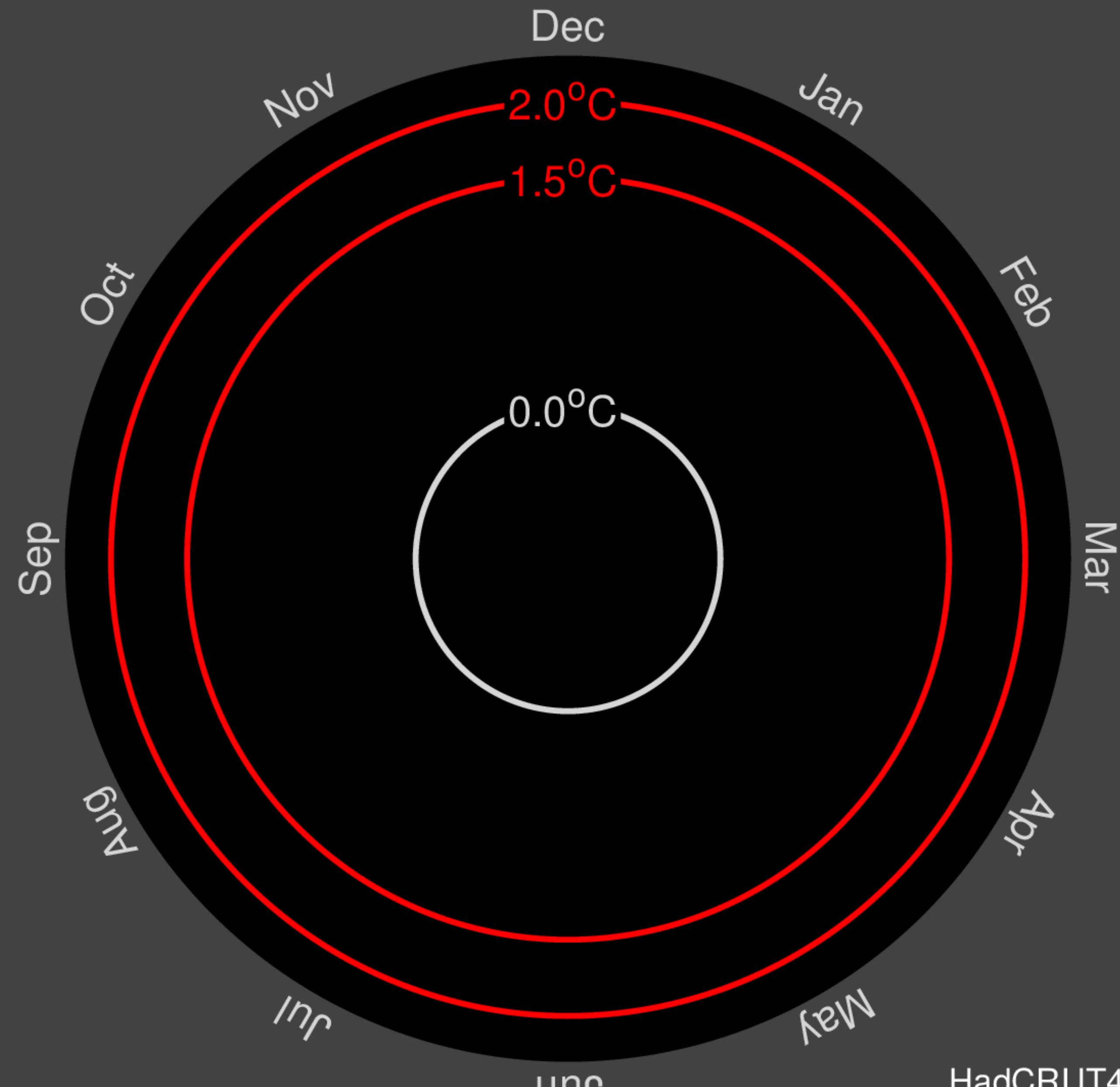
Dirección de los

Secuencial

Divergente

Cíclico

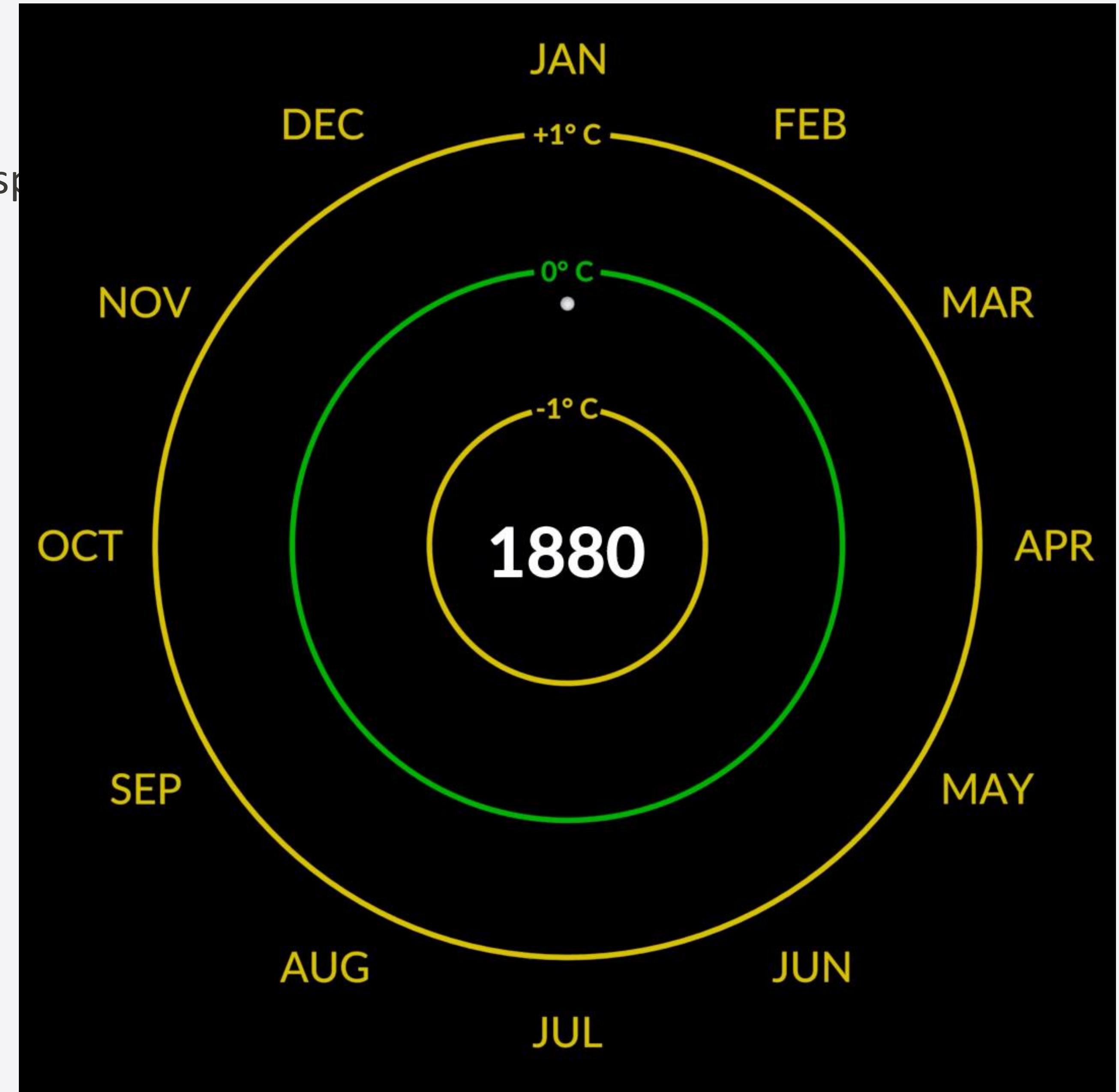
Global temperature change (1850-2020)



Dirección de los Atributos

Cíclico

Varía en una dirección respecto al punto de origen



#Cuéntalo

En abril de 2018 se lanzó el hashtag #Cuéntalo, invitando a las mujeres a relatar las agresiones sufridas. Tras analizar los datos de los primeros 14 días, éste es el resultado, abrumador e inédito. Se trata de un documento histórico que compone una nueva memoria colectiva de la violencia machista narrada en palabras de las propias mujeres.

El **objetivo** de #Cuéntalo es evidenciar la veracidad de las denuncias y la dimensión del conflicto.

#Cuéntalo en cifras

2.75
MILLONES

INTERVENCIONES

790
MIL

USUARIAS ÚNICAS

160
MIL

TUITS ORIGINALES

40
MIL

11
MIL

50
MIL

4
MIL

21
MIL

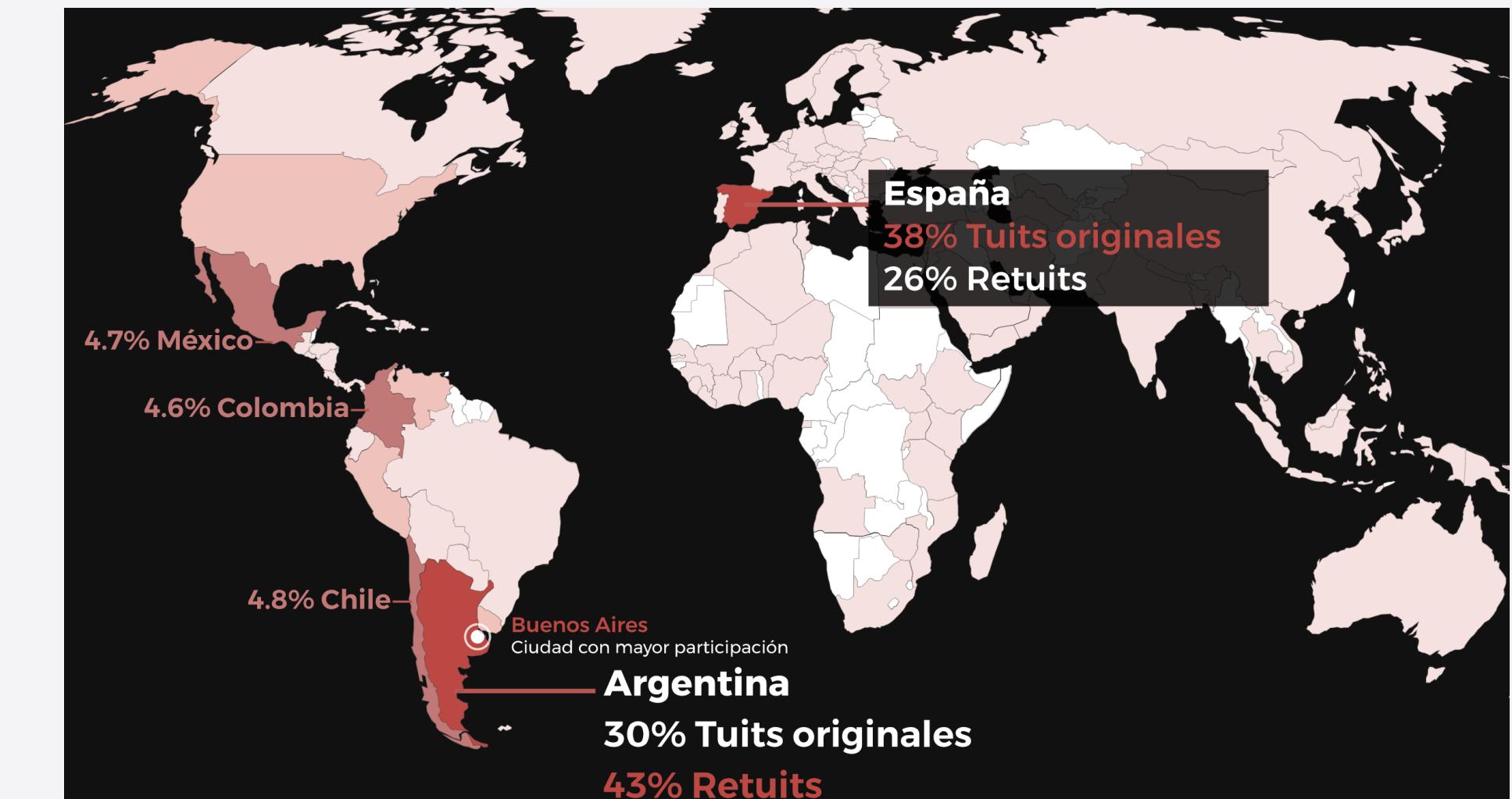
EN 1^a PERSONA

LO CUENTO YO PORQUE...

APOYO

#ANTI-CUÉNTALO

INCLASIFICADOS





<http://www.bsc.es/viz/cuentalo/>

CÓMO LEER ESTA GRÁFICA

18:00

06:00

12:00

Hora del dia

00:00 →



Testimonio



Mensajes de apoyo



Otros

La visualización muestra cerca de 130.000 Tweets, clasificados con redes neuronales.

La **POSICIÓN** en el círculo representa el tipo de mensaje. De dentro hacia afuera respectivamente, es la certeza con la que el algoritmo los categoriza en Testimonios, Mensajes de Apoyo, y Otros.

El **COLOR** representa el contenido del mensaje, rojo para los que hablan de agresiones físicas (asesinato, violación, agresión sexual y maltrato), y rosa pálido para el resto.

El círculo se lee como un reloj de 24 horas, los Tweets se organizan según la hora en que fueron escritos.

Visualización hecha por BSC Viz Team del Barcelona Supercomputing Center.

Para mas detalle sobre la visualización, visita nuestro blog.

Otros proyectos

Visualizar algoritmos

 Mike Bostock's Block `bd012e7bbe5f66c41d39`
Updated February 8, 2016

Popular / About

Quicksort III



This block displays a visualization of the Quicksort III algorithm. It shows a series of vertical bars of varying heights, representing an array of elements. The bars are initially unsorted, forming a dense cluster at the top. As the algorithm progresses, the bars are rearranged into a sorted order from left to right. The visualization uses a minimalist black and white color scheme, with the bars being black on a white background.

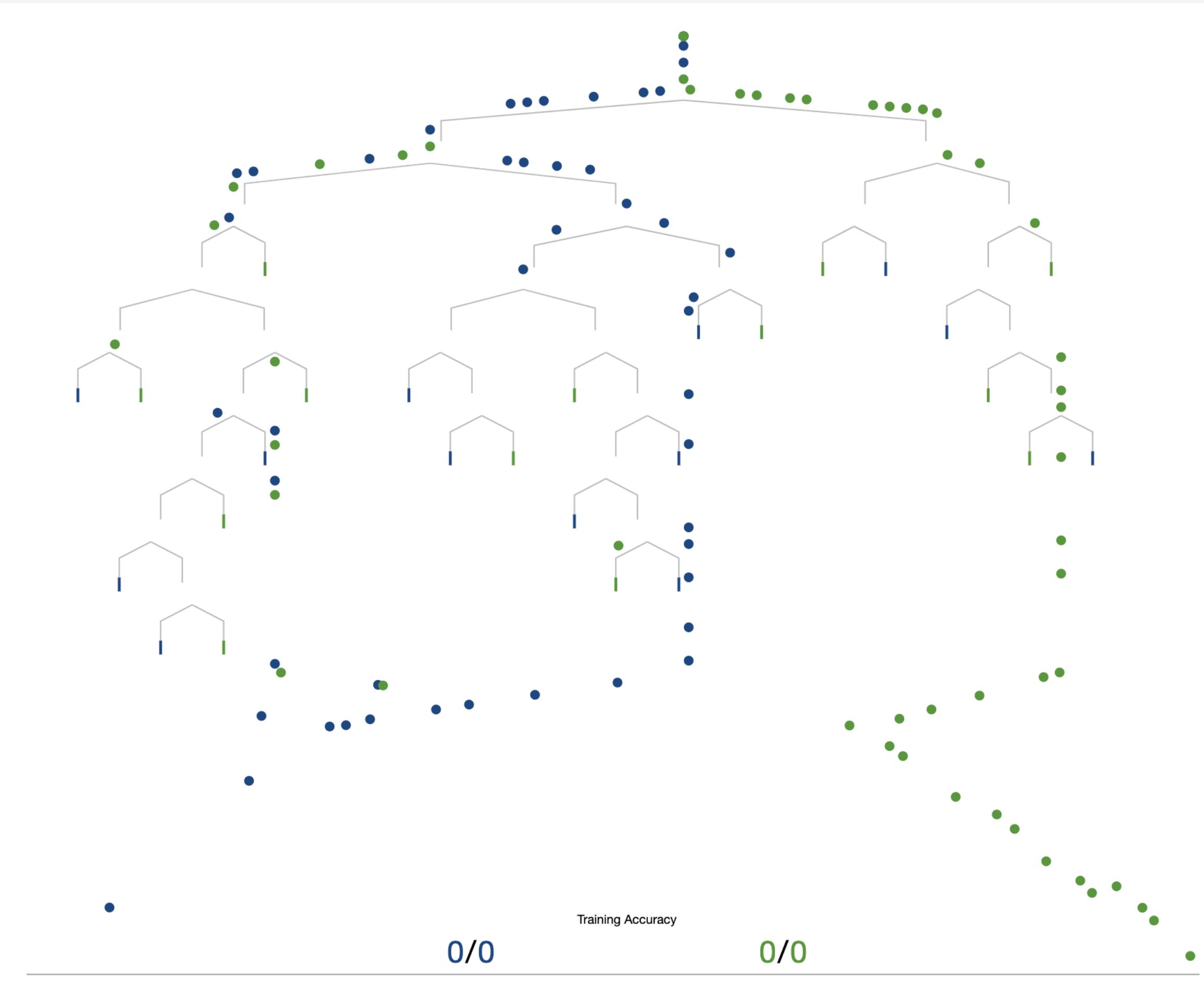
<https://bostocks.org/mike/algorithms/>

Otros proyectos

Visualizar algoritmos

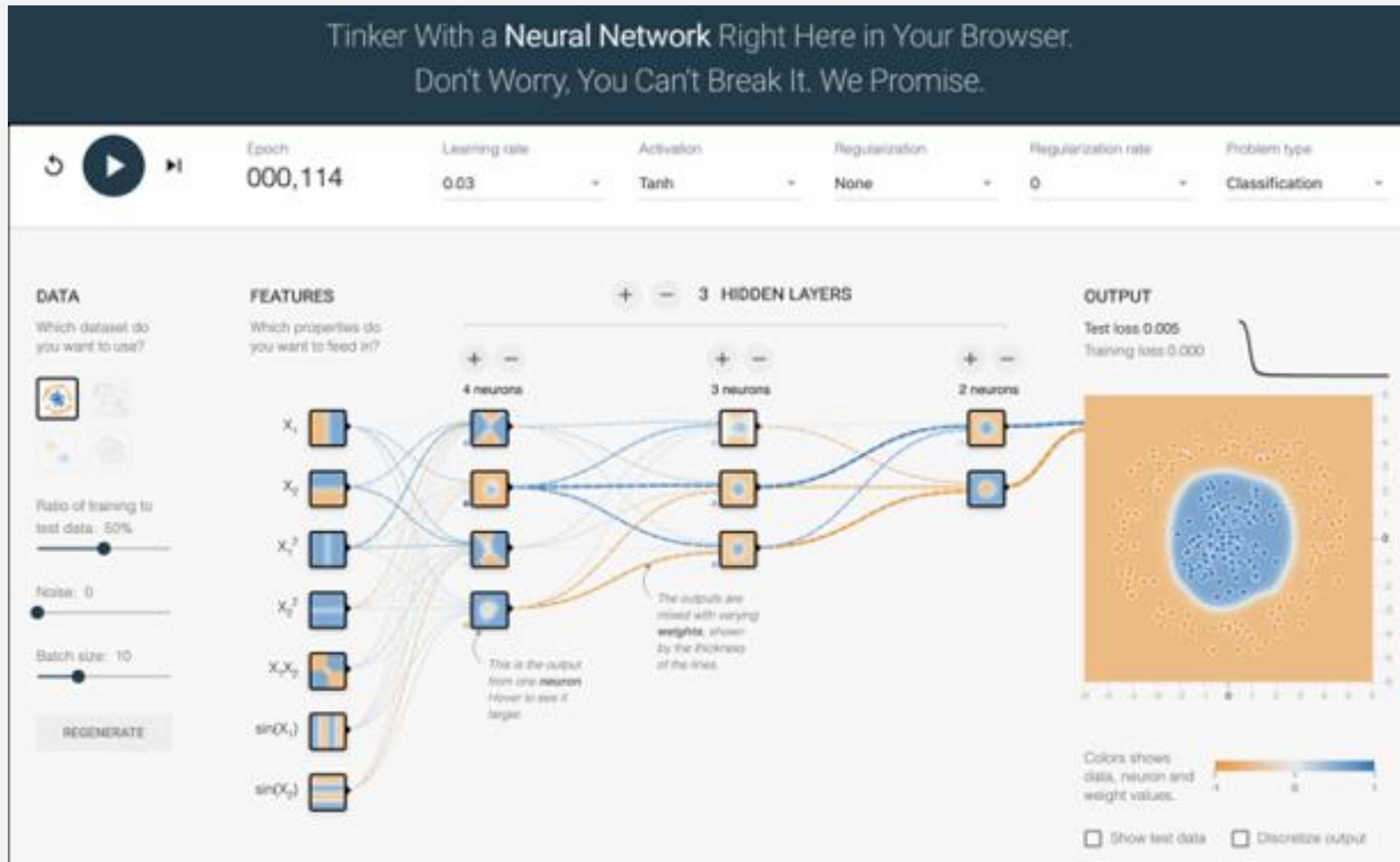
Making predictions

The newly-trained decision tree model determines whether a home is in San Francisco or New York by running each data point through the branches.



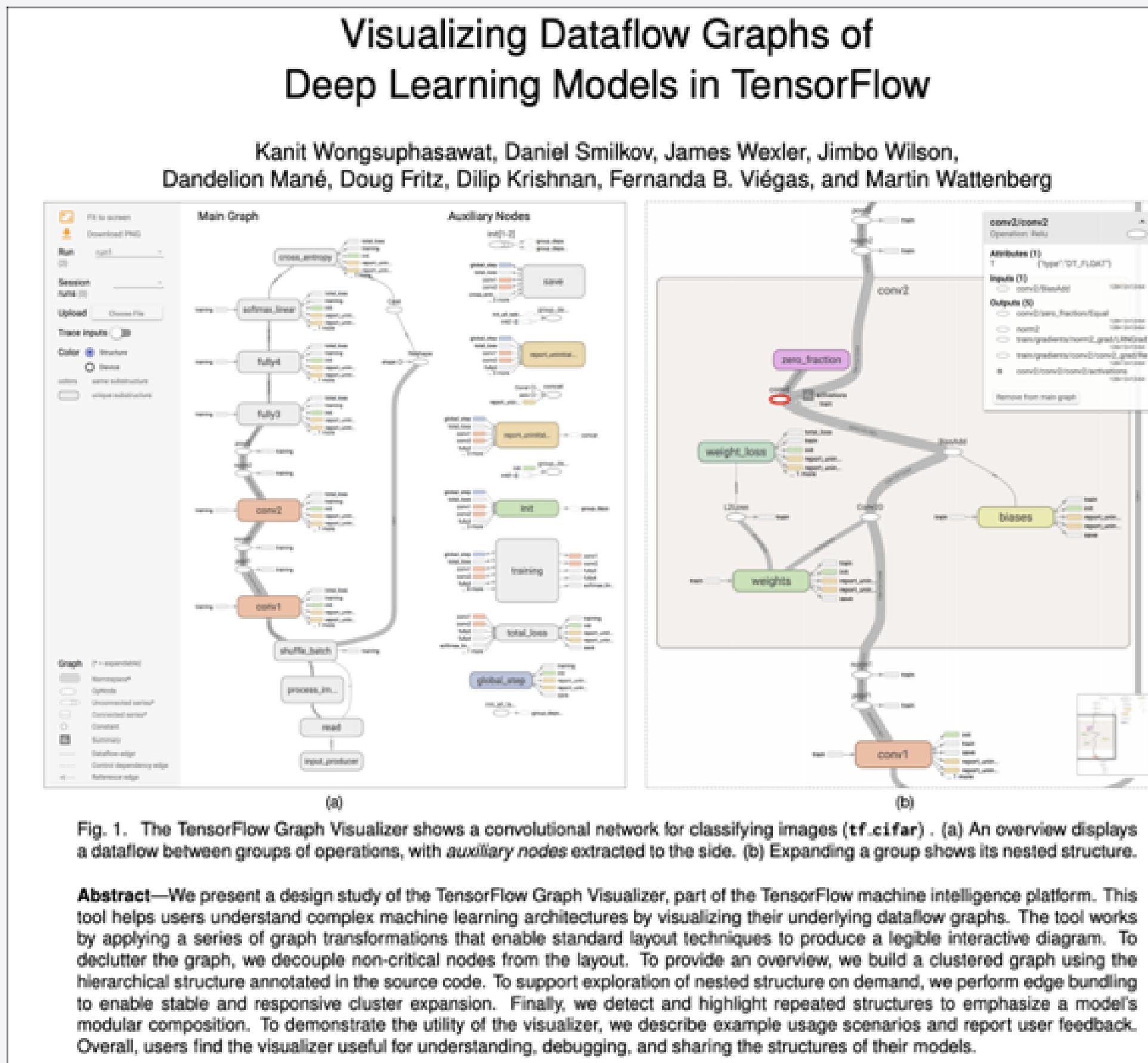
Otros proyectos

Redes neuronales



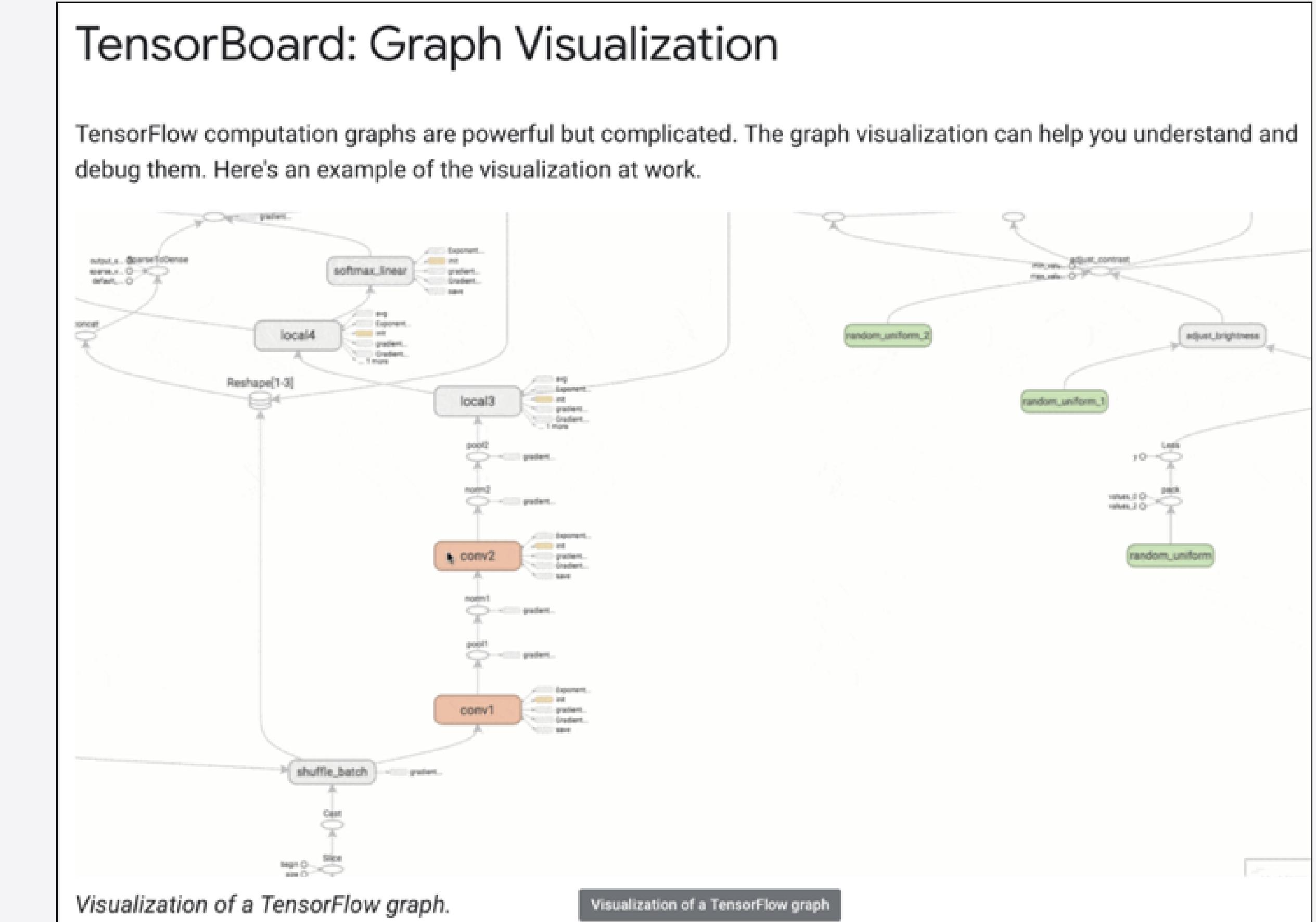
Otros proyectos

Redes neuronales



Wongsuphasawat, et al. 2017

Visualizing dataflow graphs of deep learning models in tensorflow



TensorBoard: Graph Visualization

What?

Datos

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

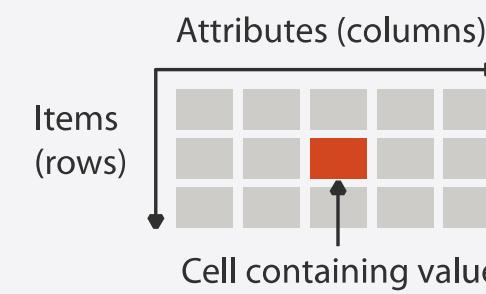
Positions

Attributes

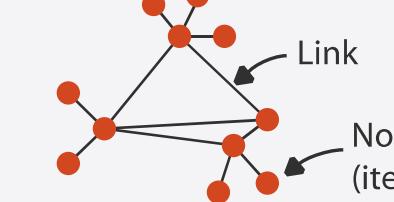
Attributes

➔ Dataset Types

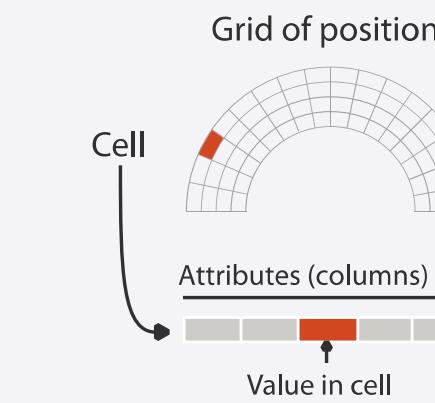
→ Tables



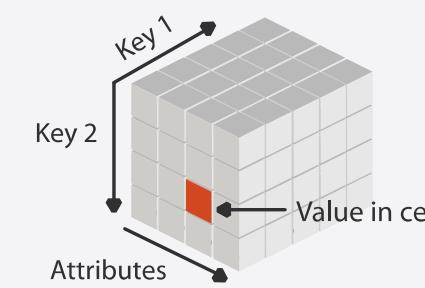
→ Networks



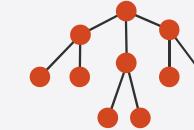
→ Fields (Continuous)



→ Multidimensional Table



→ Trees



→ Geometry (Spatial)



Attributes

➔ Attribute Types

→ Categorical



→ Ordered

→ Ordinal

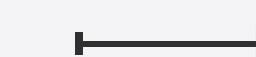


→ Quantitative



➔ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



EJERCICIO 1 - 10 min

- Descargar el archivo **filmdeathcounts.csv** (alternativa **.xlsx**)
- Abrir el archivo (Excel, Preview de mac, GoogleSheets, etc) y analizar el dataset.
 - Anota qué tipo de dataset es, qué tipos de datos contiene, y de qué tipo son los atributos.
- Abrir la web www.datawrapper.de /Click en “Start Creating”. Cargar el archivo
- [Opcional: Crear una cuenta para guardar las gráficas]

Round 1

- Apunta el tipo de variables que hay en el dataset
- ¿Qué gráficas puedes hacer y con qué variables para explorar el dataset?



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

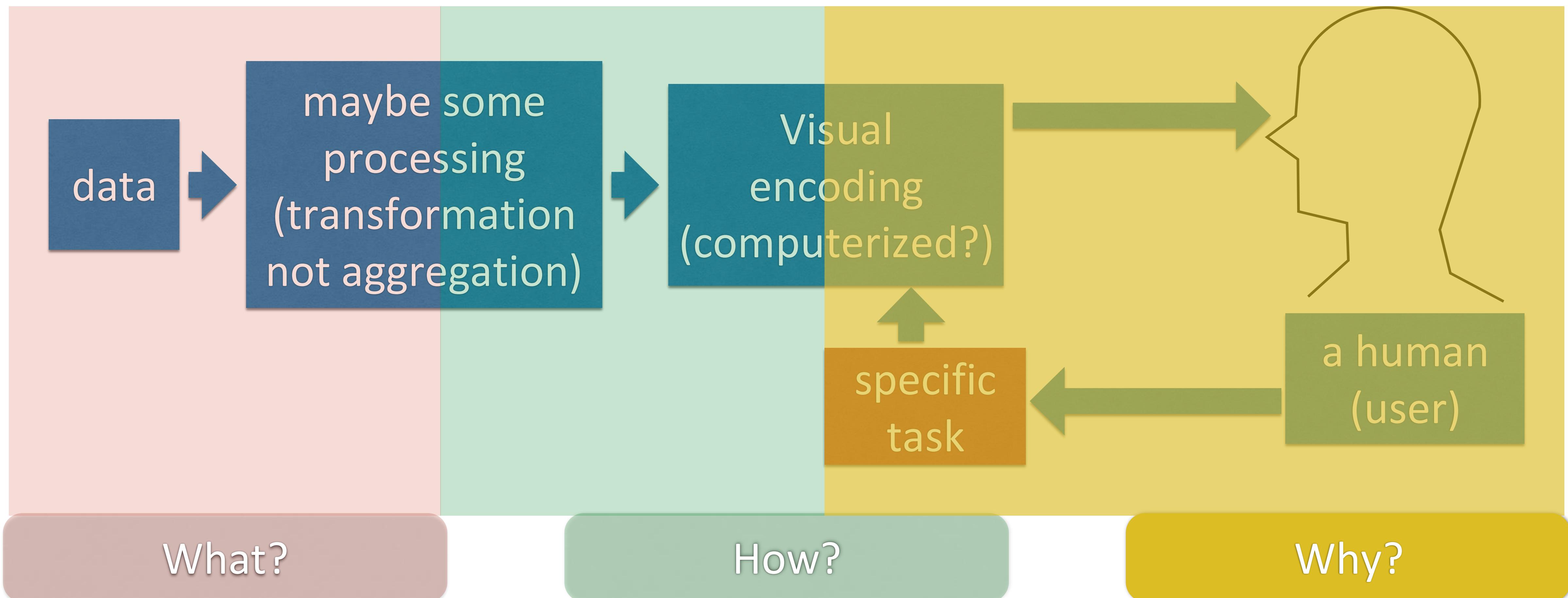


2. Abstracción de Tareas

Guillermo Marin
guillermo.marin@uab.cat

Data Visualisation

*Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively
(T.Munzner)*



Para qué se utiliza una visualización

Tareas

una acción + un objetivo

Tareas

acción + objetivo

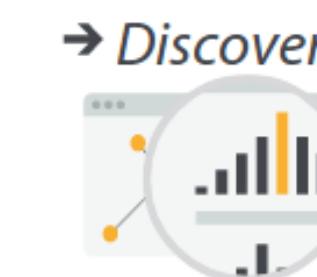
Why?

Actions

Targets

→ Analyze

→ Consume



→ Present



→ Enjoy



→ Produce

→ Annotate



→ Record



→ Derive



→ All Data

→ Trends



→ Outliers



→ Features



→ Search

	Target known	Target unknown
Location known <i>Lookup</i> <i>Browse</i>
Location unknown	<i>Locate</i>	<i>Explore</i>

→ Query

→ Identify



→ Compare



→ Summarize



→ Network Data

→ Topology



→ Paths



→ Spatial Data

→ Shape



What?

Why?

How?

Tareas

acción + objetivo

Why?

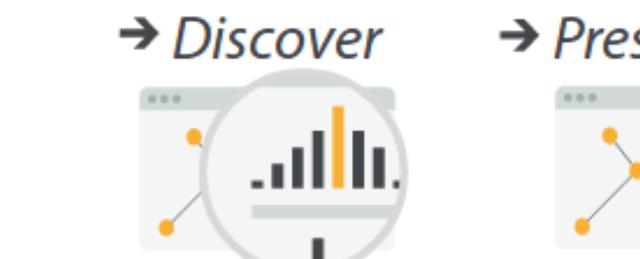
Actions

Targets

→ Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



→ All Data

→ Trends



→ Outliers



→ Features



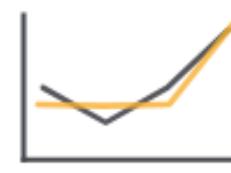
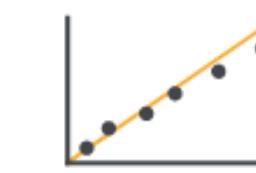
→ Attributes

Many

→ Dependency

→ Correlation

→ Similarity



→ Shape



What?

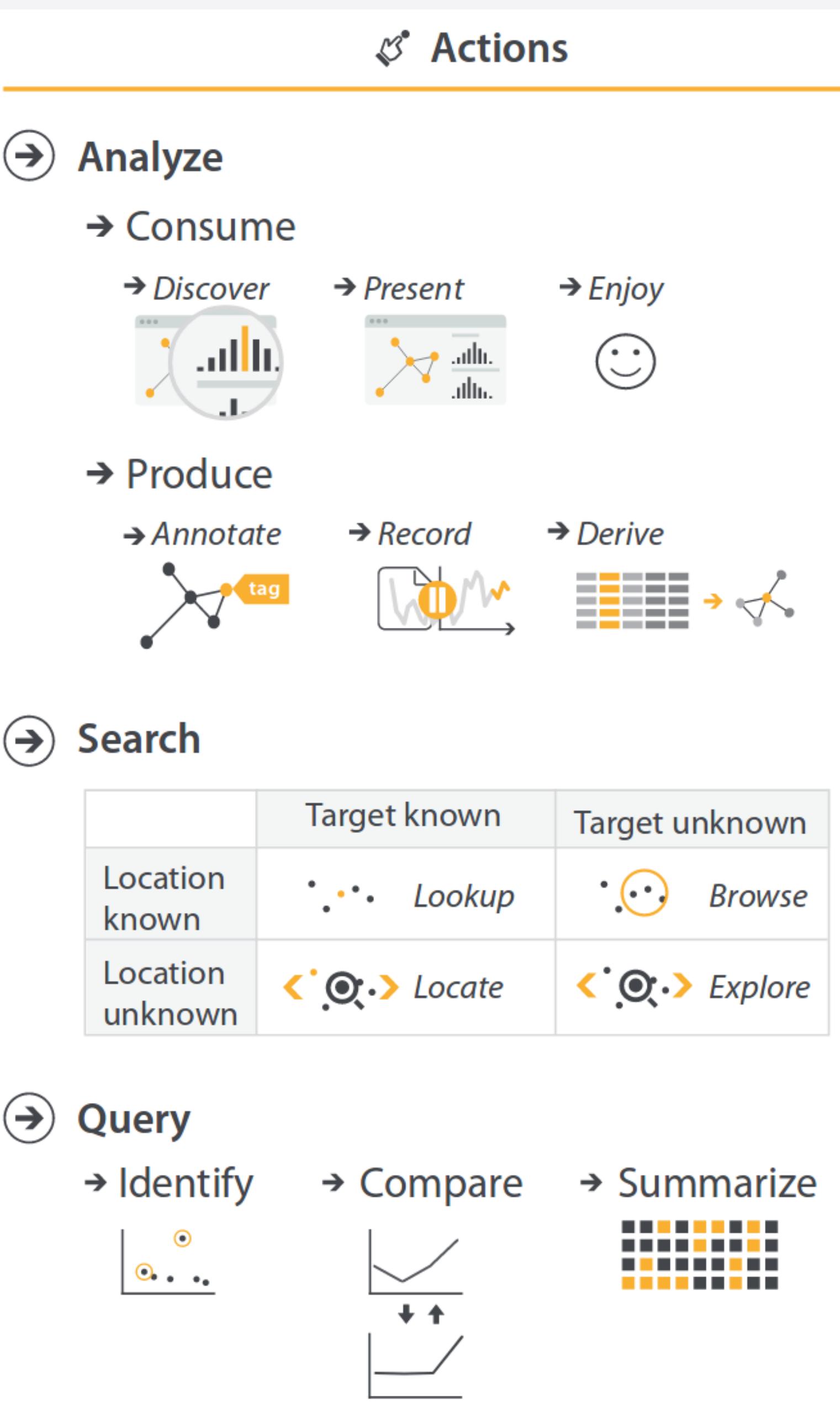
Why?

How?

Acción

Tres fases:

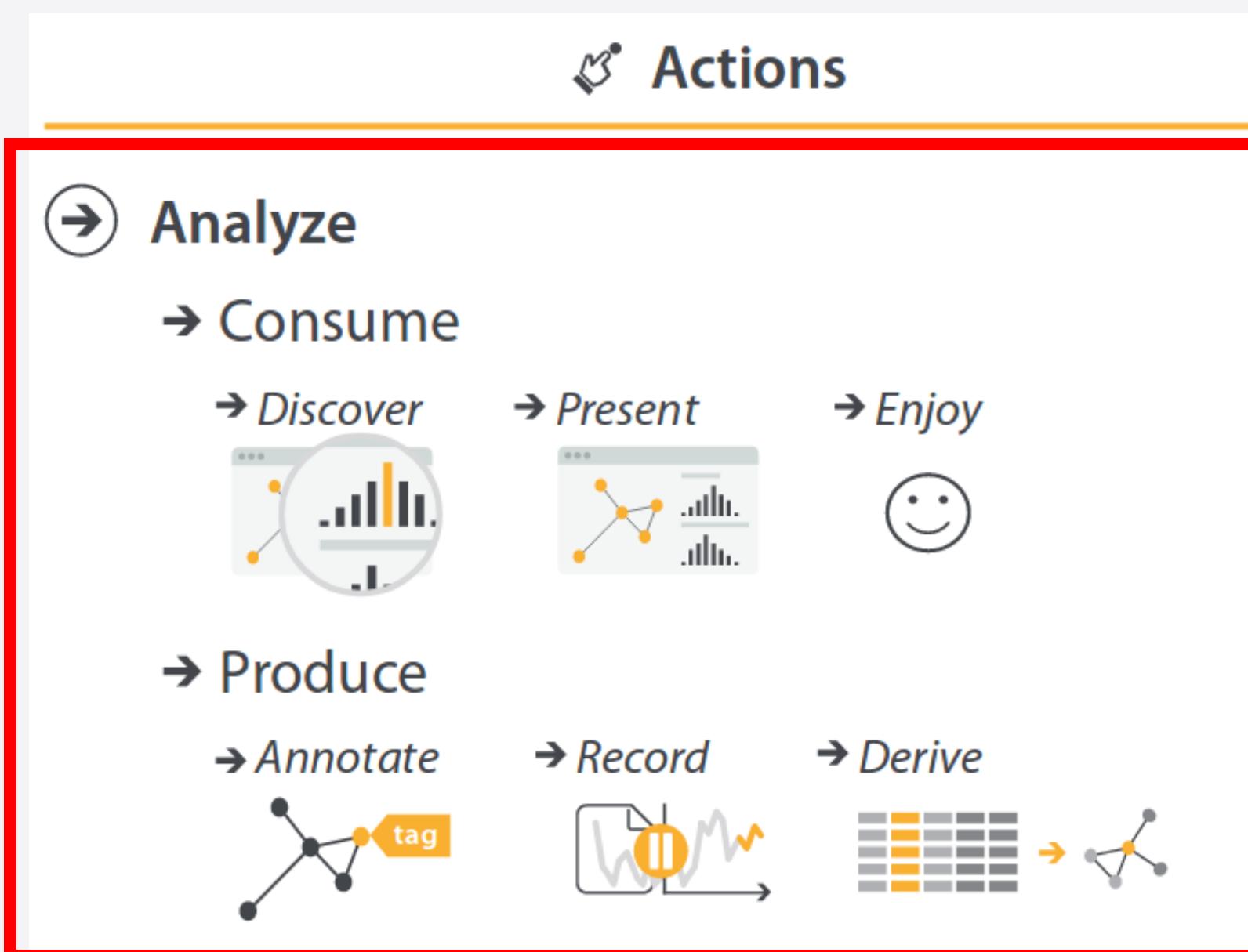
- Analizar -> Consumir / Producir información
 - Buscar -> Todos los casos anteriores de Analizar requieren que el usuario **busque** elementos de interés. Esta búsqueda está modulada por el conocimiento previo del target y su posición.
 - Consultar -> Una vez que se han encontrado los targets, un objetivo del usuario es hacer una **consulta** sobre éstos.
 - Identificar. Consulta sobre un objetivo
 - Comparar. Dos o más
 - Resumir. Consulta sobre el set completo de targets posibles.



Acción - Analizar

- Consumir Información:

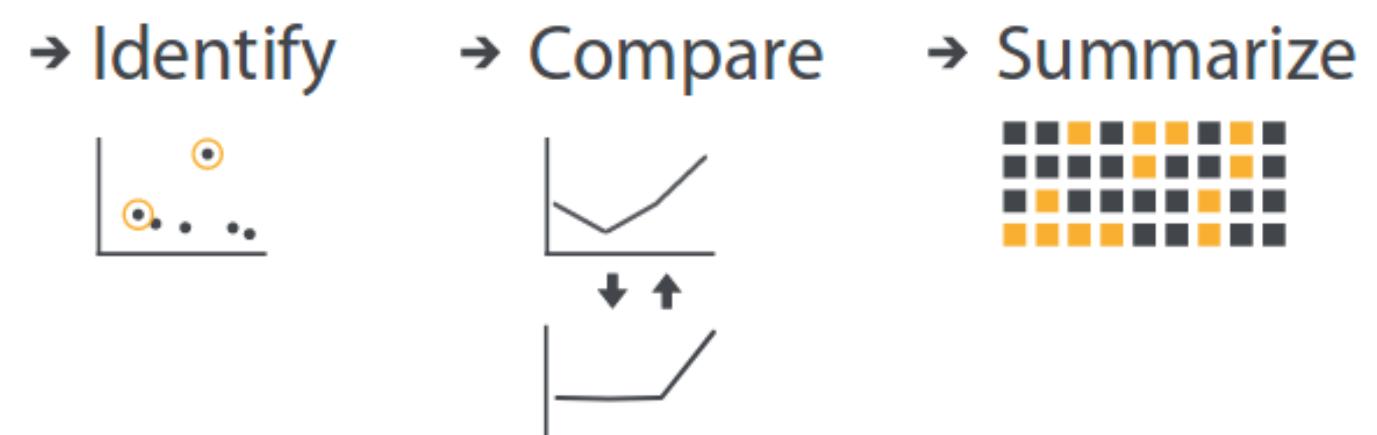
- Descubrir. Esto es lo que usualmente asociamos a una visualización, usarla para descubrir algo que no sabíamos.
- Presentar. Queremos entregar un mensaje específico a la audiencia.
- Disfrutar. A veces queremos explorar los datos solo por el placer de hacerlo. Eso también es una tarea.



→ Search

	Target known	Target unknown
Location known		
Location unknown		

→ Query



Acción - Analizar

NameVoyager: Explore baby names and name trends letter by letter

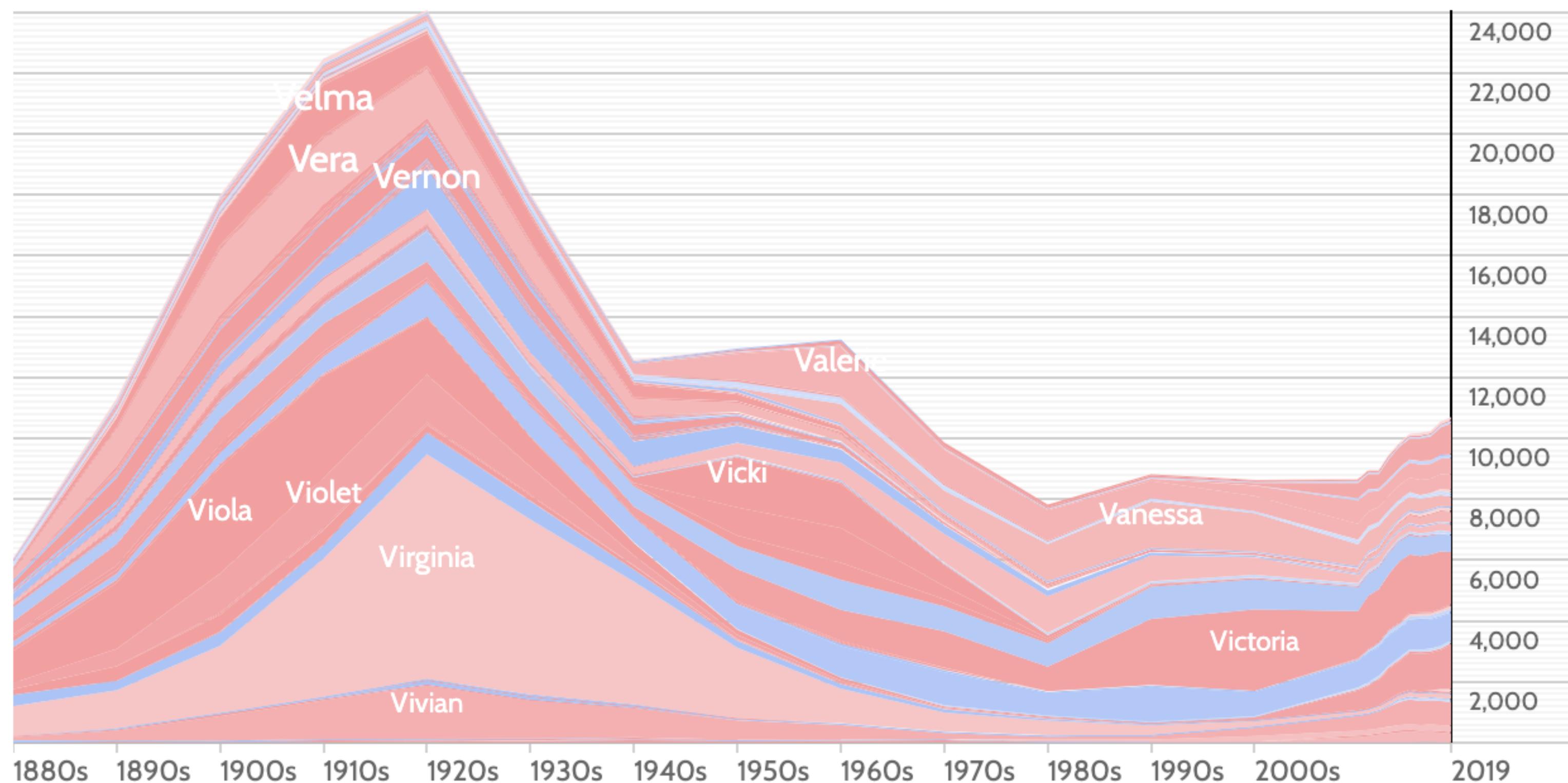
Baby Name > Both Boys Girls

boys	1,000	500	100	25	1
girls	1,000	500	100	25	1

Current rank:

per million births

Names starting with 'V' per million babies



Click a name graph to view that name. Double-click to read more about it.

[enlarge](#)

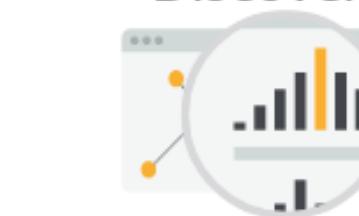
See top names by decade, trends by letters, and more with [Expert Name Voyager!](#)

Actions

→ Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



→ Produce

→ Annotate



→ Record



→ Derive



→ Search

	Target known	Target unknown
Location known	•••	Lookup
Location unknown	🔍	Locate

••• *Lookup*

🔍 *Locate*

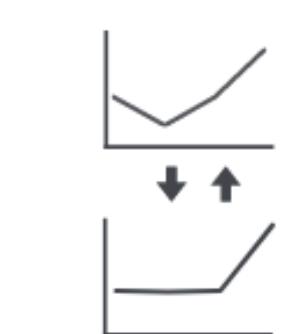
🔍 *Explore*

→ Query

→ Identify



→ Compare

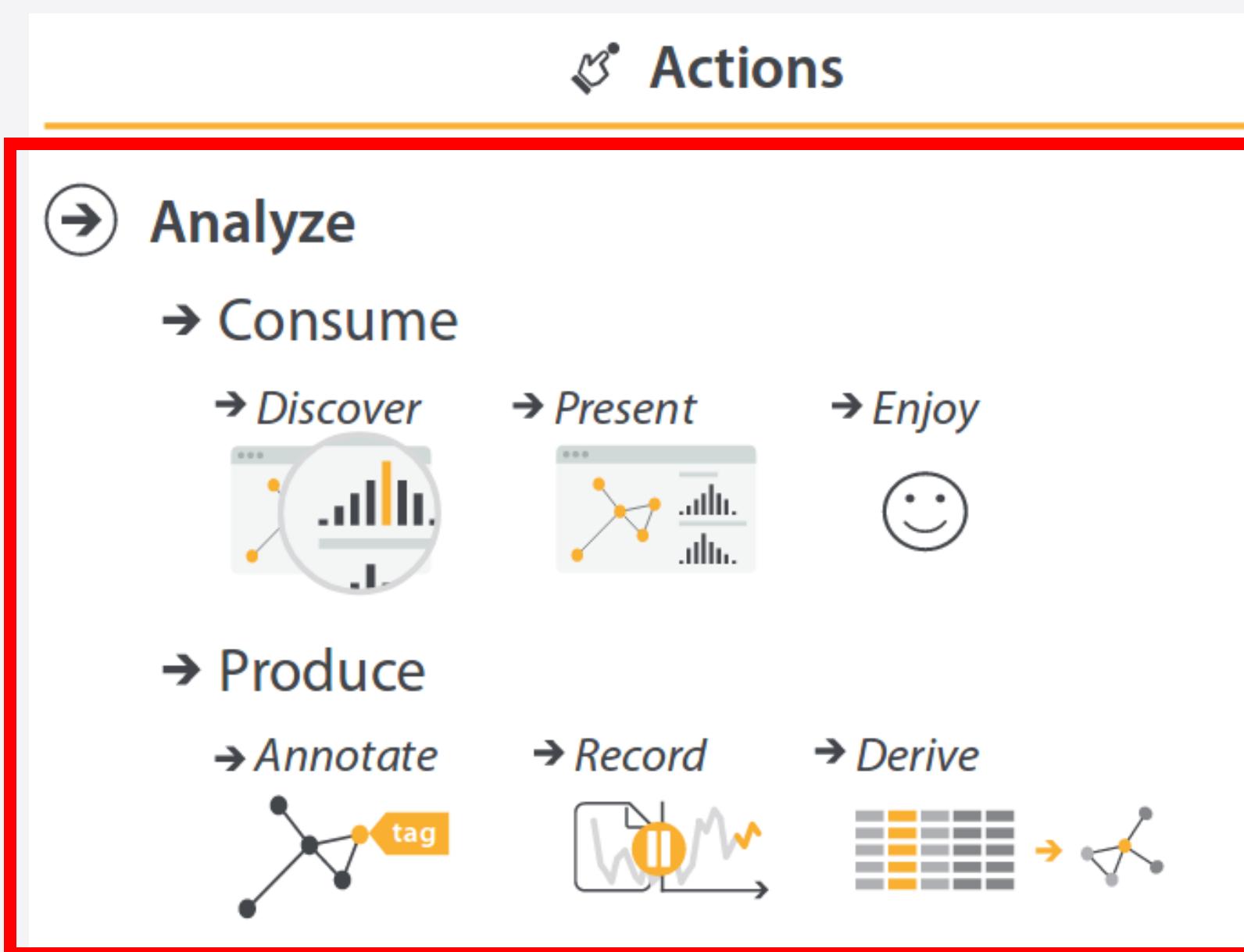


→ Summarize



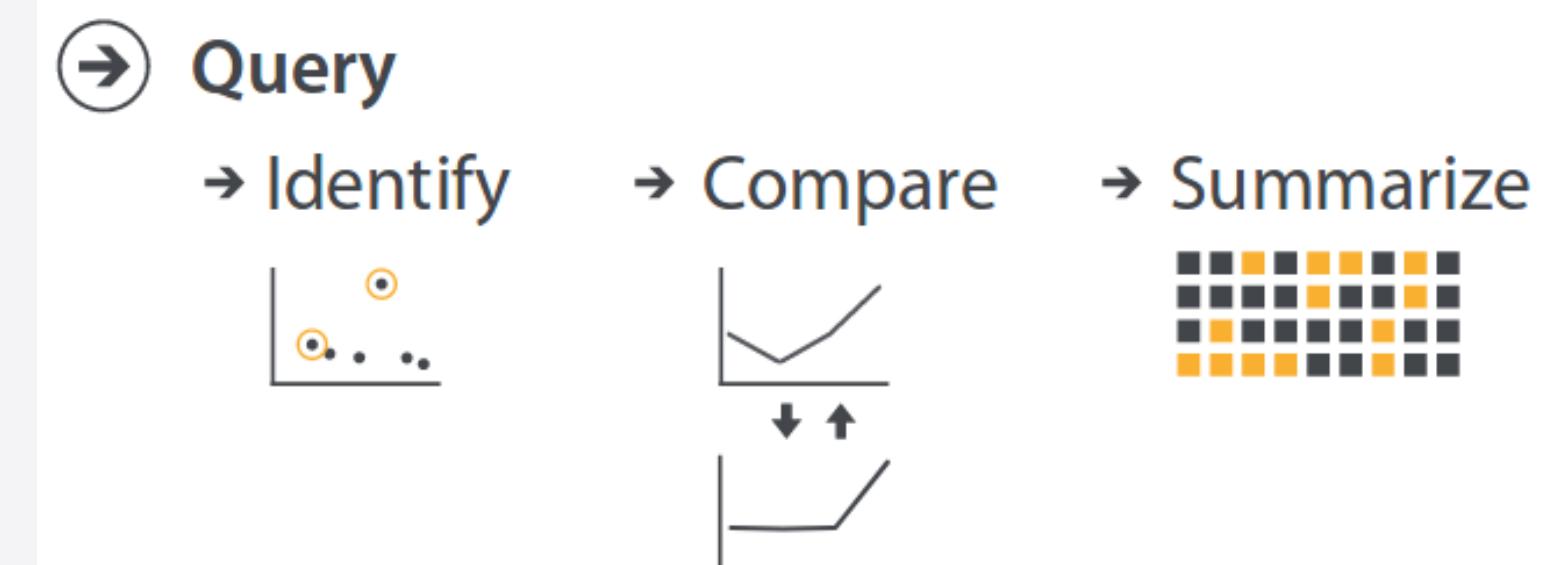
Acción - Analizar

- **Producir información:**
 - Anotar. Agregamos información que aporta contexto, explicaciones, etiquetas, etc.
 - Grabar. Cuando interactuamos con una visualización en un dispositivo realizamos un recorrido que es susceptible de ser repetido, tanto con los mismos datos como con otros.
 - Derivar. Transformar estos datos en otros, posiblemente con otra estructura.



→ Search

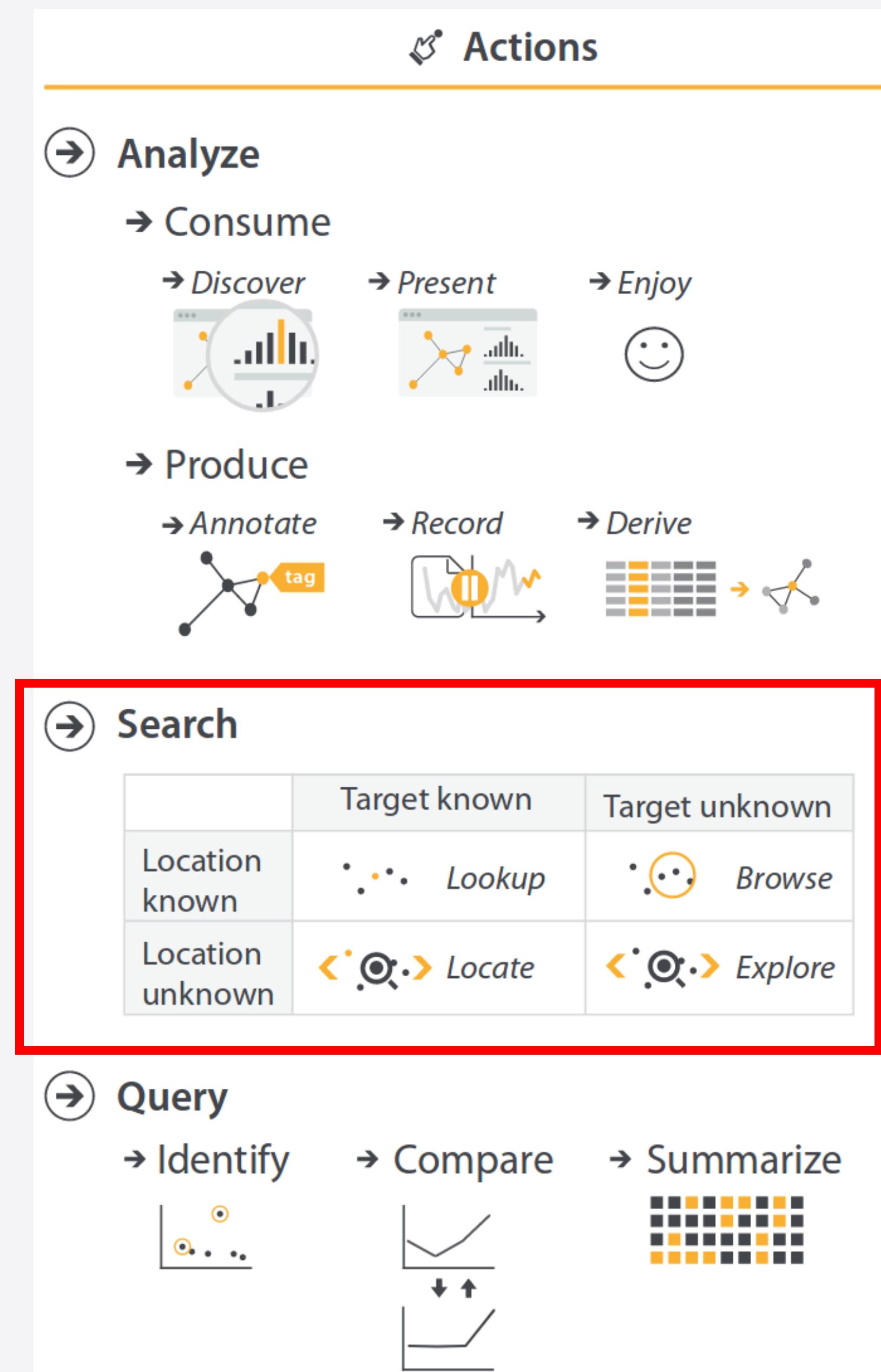
	Target known	Target unknown
Location known	Lookup	Browse
Location unknown	Locate	Explore



Acción - Buscar

Buscar -> Todos los casos anteriores de Analizar requieren que el usuario **busque** elementos de interés. Esta búsqueda está modulada por el conocimiento previo del target y su posición.

The screenshot shows a map of the area around the University Autónoma de Barcelona (UAB) in Cerdanyola del Vallès. A red marker indicates the location of the 'Escola d'Enginyeria - UAB'. The map displays various roads, including the AP-7, C-58, N-150, and B-30. Other landmarks shown include Aeropuerto de Sabadell (LELL), Barberà del Vallès, Badia del Vallès, Parque dels Pinetos, and the Parque Tecnológico del Vallès. On the left side of the screen, there is a sidebar for the 'Escola d'Enginyeria - UAB' listing, which includes a thumbnail image of the building, the name, address (Carrer de les Sitges, 08193 Cerdanyola del Vallès, Barcelona), opening hours (8:00–22:00), and a link to their website (uab.cat). There are also buttons for 'Cómo llegar' (Get directions), 'Guardar' (Save), 'Cercano' (Nearby), 'Enviar a tu teléfono' (Send to phone), and 'Compartir' (Share).



Acción - Buscar

Buscar -> Todo
usuario busca
el conocimiento



Escola d'Enginyeria - UA
UAB Escola d'Enginyeria
4,2 ★★★★☆ 79 reseñas
Escuela universitaria

Cómo llegar Guardado Cercano Entete
Guardada en Favoritos

Añadir nota Ver lista

Carrer de les Sitges, 08193 Cerdanyola del Vallès, Barcelona

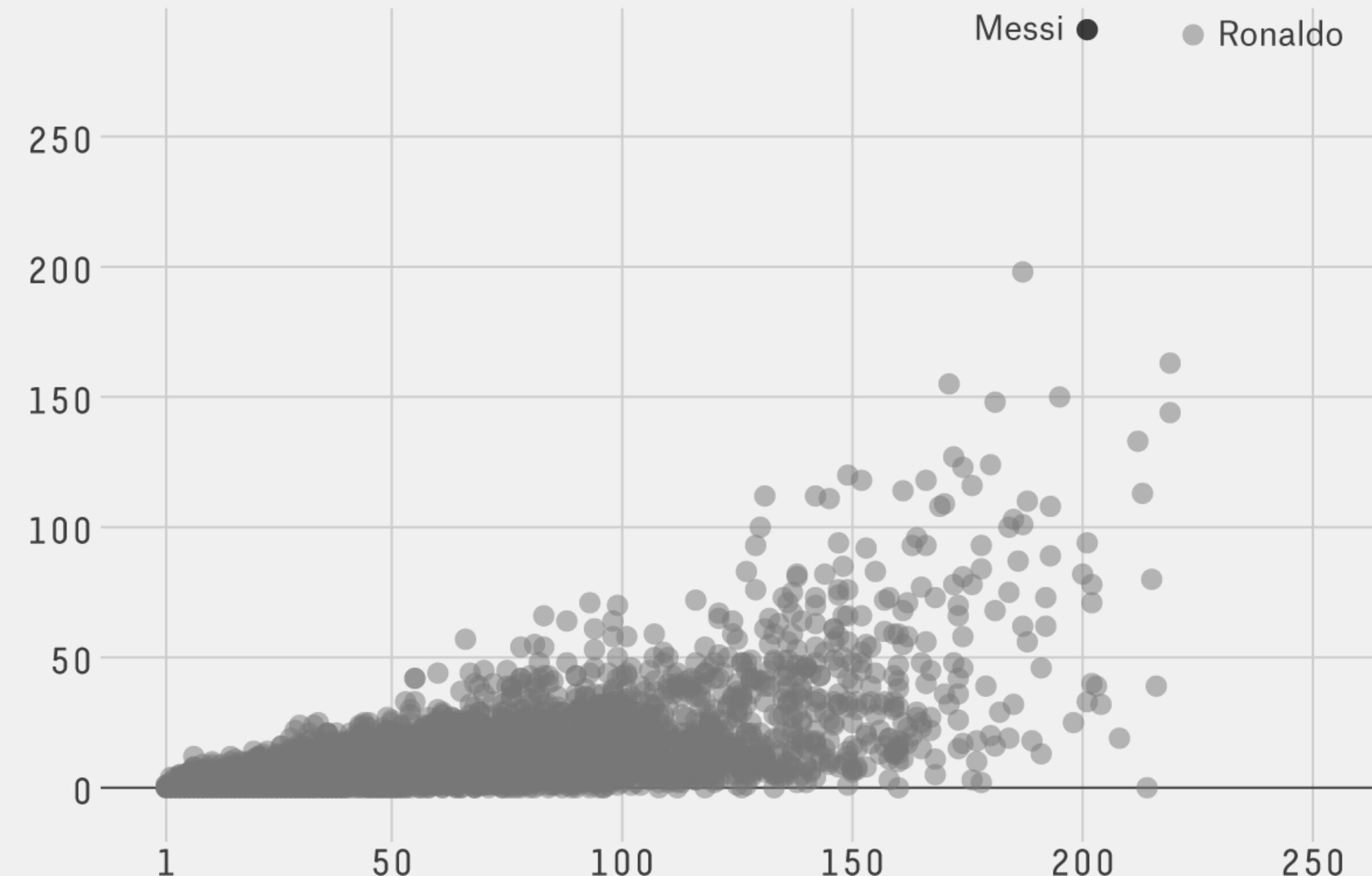
Se encuentra en: Universidad Autónoma de Barcelona

Abierto ahora: 8:00–22:00

uab.cat

Overall Scoring Production

Total goals and assists vs. games played since 2010 World Cup

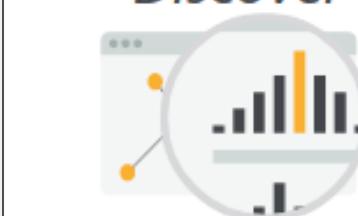


Actions

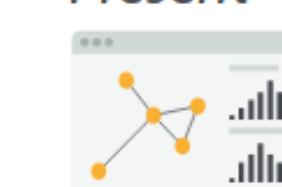
Analyze

→ Consume

→ Discover



→ Present



→ Enjoy



Produce

→ Annotate



→ Record



→ Derive

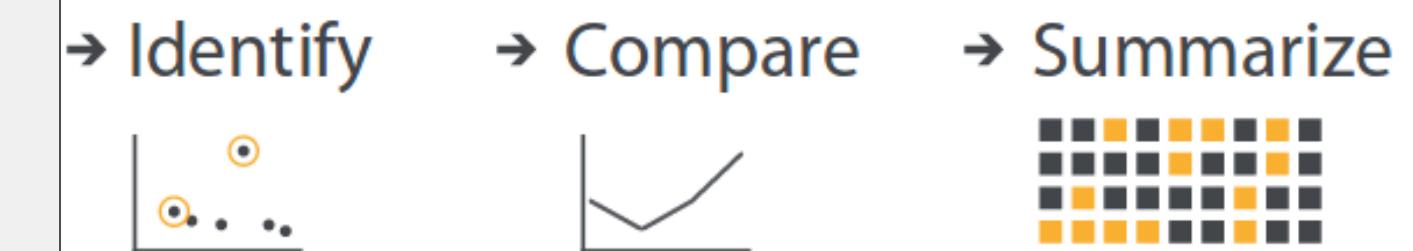


Search

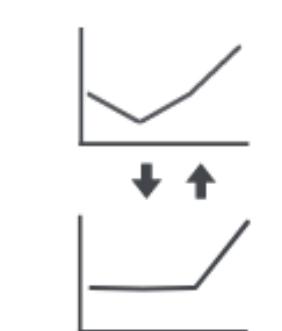
	Target known	Target unknown
Location known	••• ••• Lookup	•••○○ Browse
Location unknown	○○○○○ Locate	○○○○○ Explore

Query

→ Identify



→ Compare

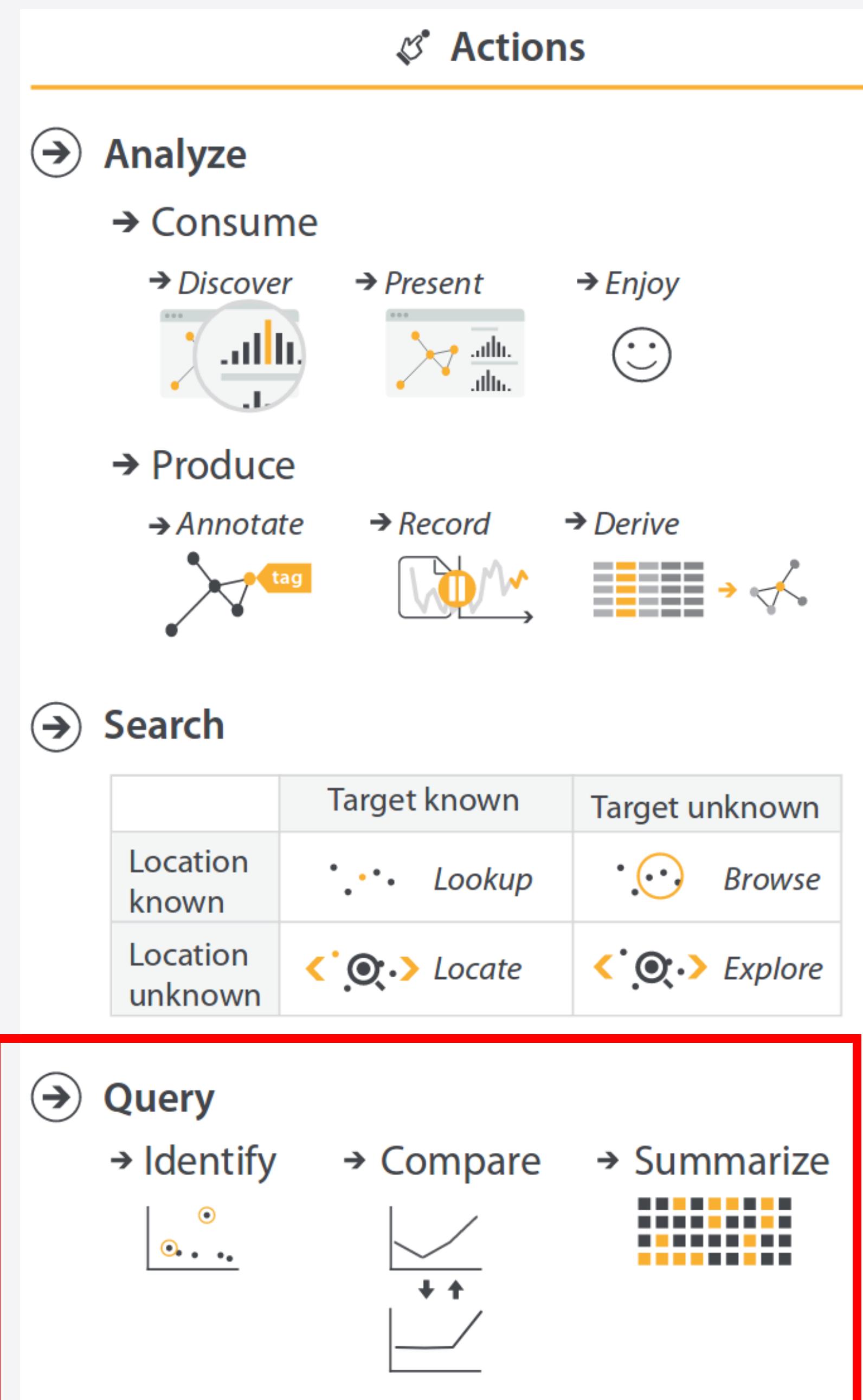


→ Summarize



Acción - Consultar (query)

- Consultar -> Una vez que se han encontrado los targets, un objetivo del usuario es hacer una **consulta** sobre éstos.
 - Identificar. Necesitamos la observación, grupo de observaciones, columnas, datasets, etc., que cumplan con un criterio específico.
 - Comparar. Conocer similitudes y diferencias entre targets.
 - Resumir. Dada una observación, grupo de observaciones, columnas, datasets, etc., necesitamos conocer una agregación de éstos que los describa.



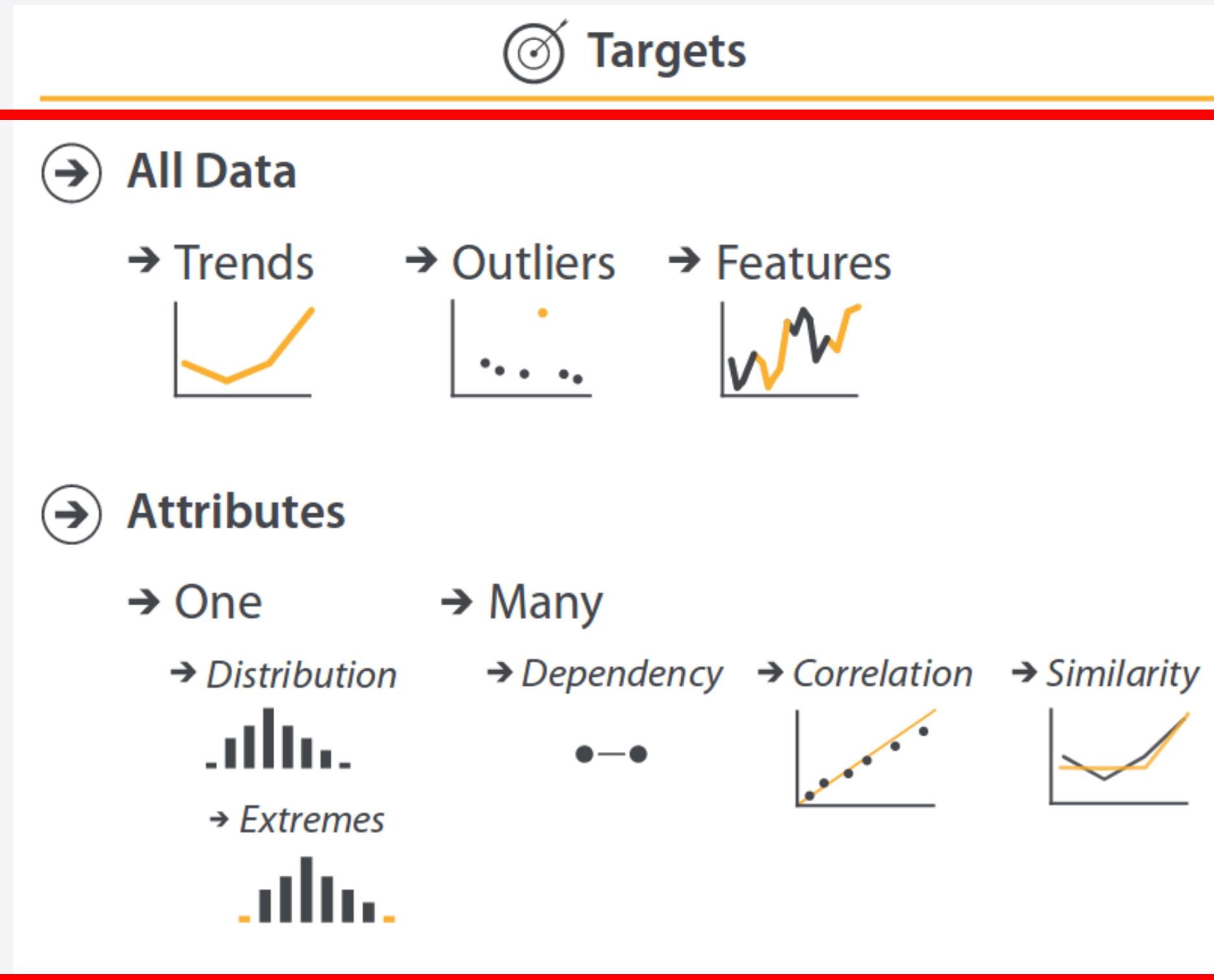
Targets

- **El dataset completo**

- *tendencias* (cambios o patrones consistentes en una misma dirección)
- *outliers* (observaciones que se salen del comportamiento habitual distorsionando los resultados del análisis)
- *características* (patrones comunes dentro del dataset, e.g., *clusters*)

- **Atributos (propiedades específicas)**

- *uno* (distribución, valores extremos)
- *varios* (relaciones de dependencia, correlación, o similitud)



→ Network Data

→ Topology



→ Paths



→ Spatial Data

→ Shape



What?

Why?

How?

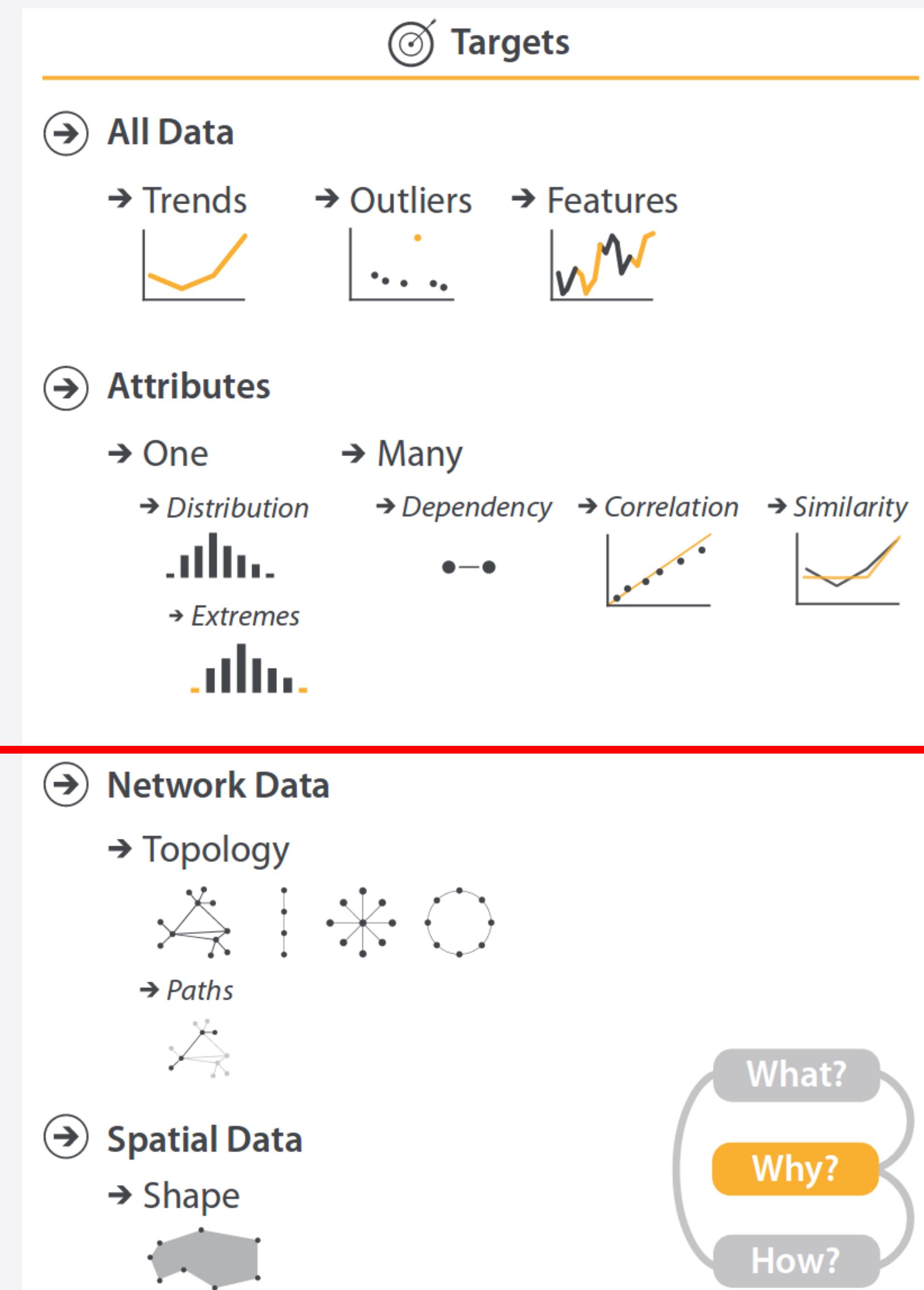
Targets

- **Datos de red**

- se analiza la *topología* (estructura de la red, o el proceso que la llevó a tener su forma actual), o *paths* dentro de ella (camino para llegar desde un nodo hasta otro).

- **Datos espaciales**

- se despliega la *forma* que tienen los elementos (o un conjunto de elementos) con características espaciales o geográficas

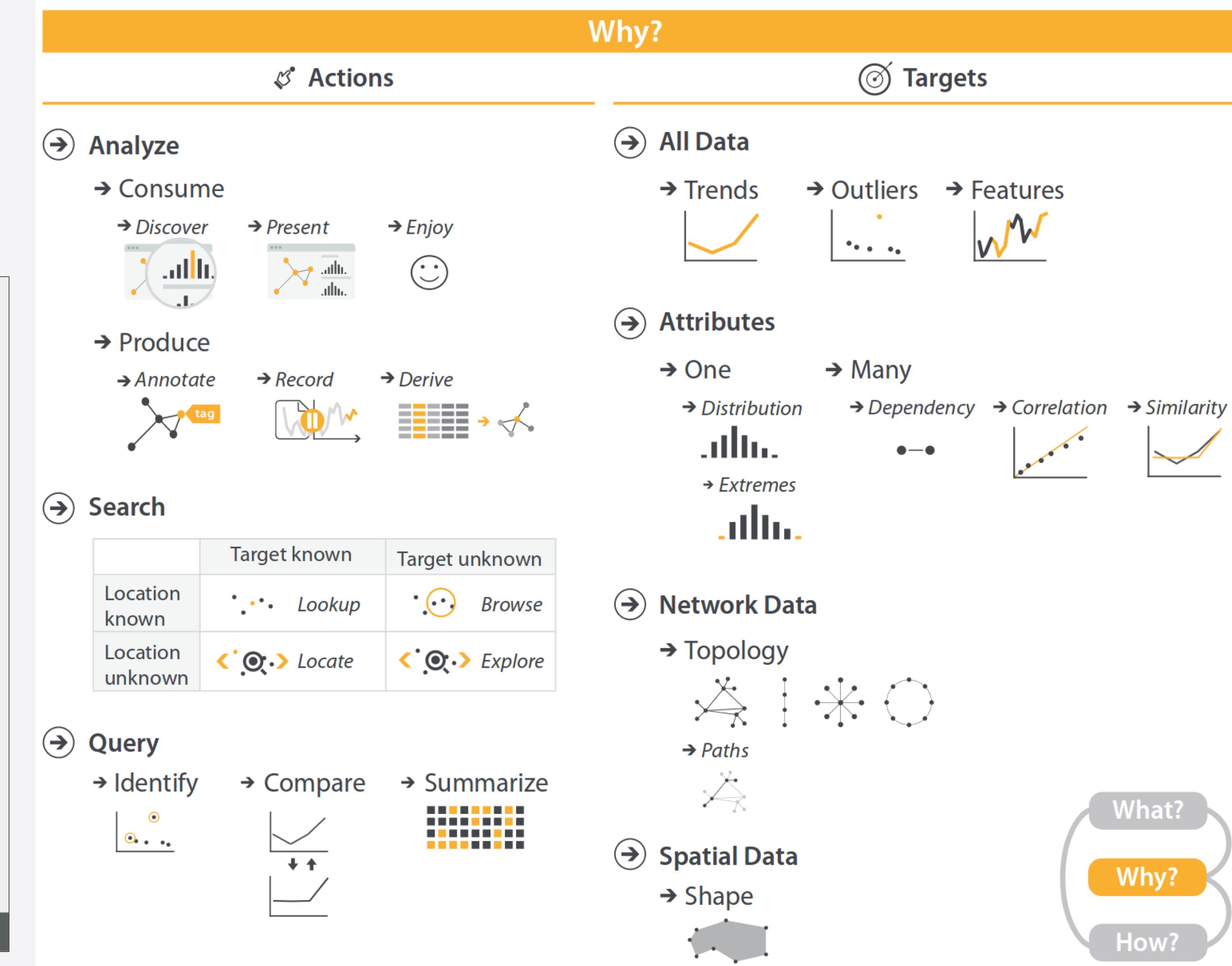
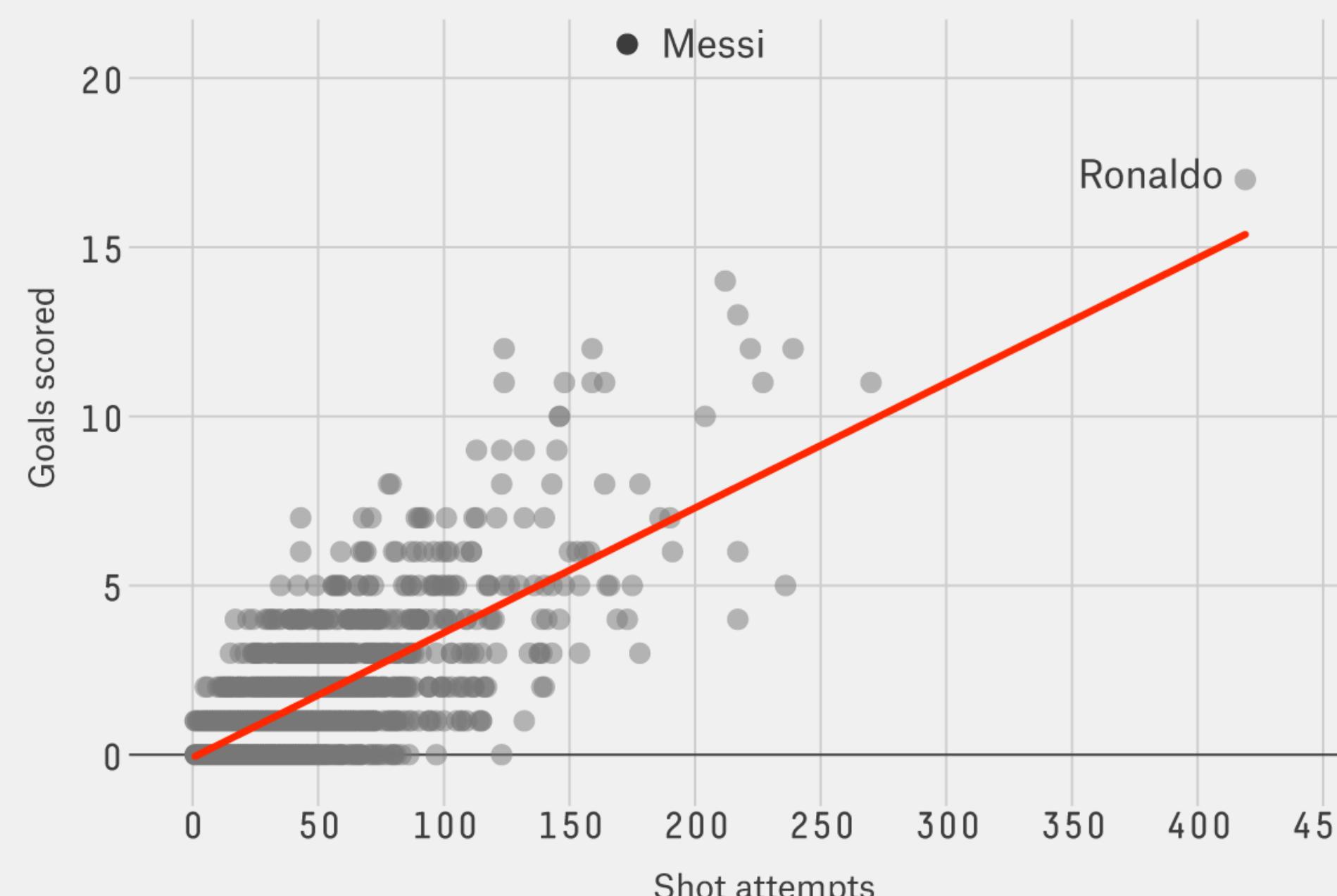


Tareas

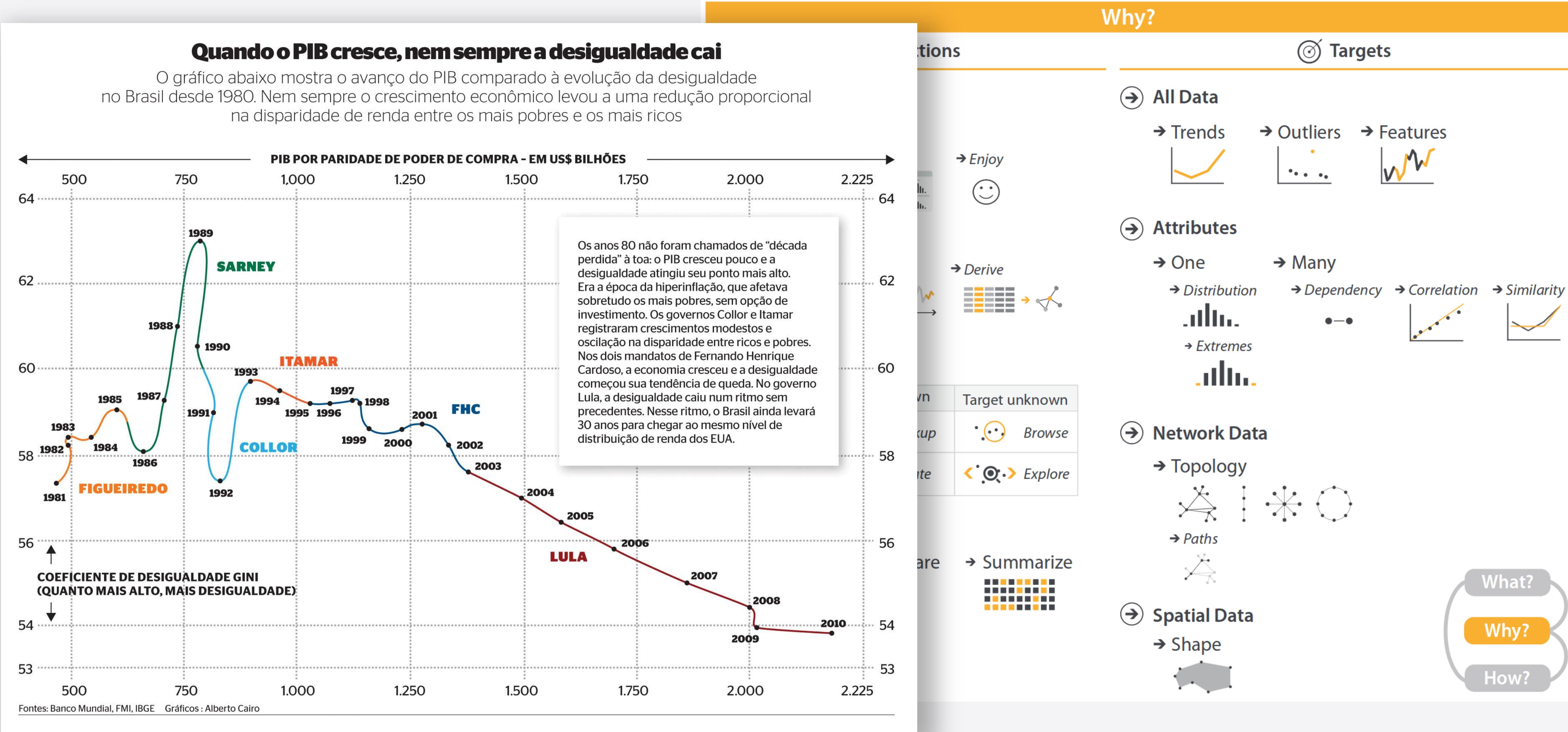
Presentar → Locate → Identificar + Trends & outliers

Deadly From Outside the Penalty Area

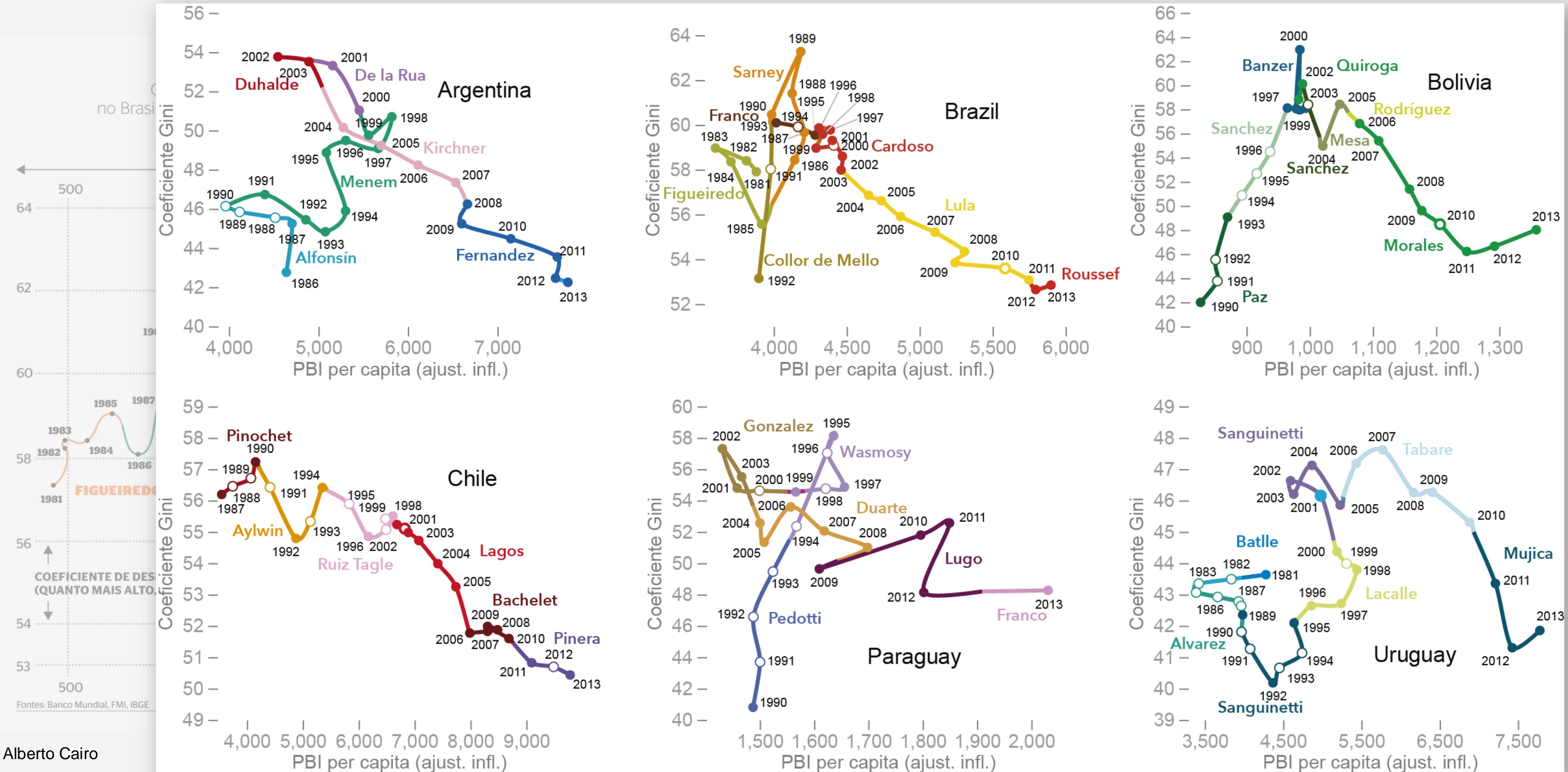
Goals scored vs. shot attempts



Presentar → Explore → Identify + Correlación entre Atributos



Present + Atributos/Correlación



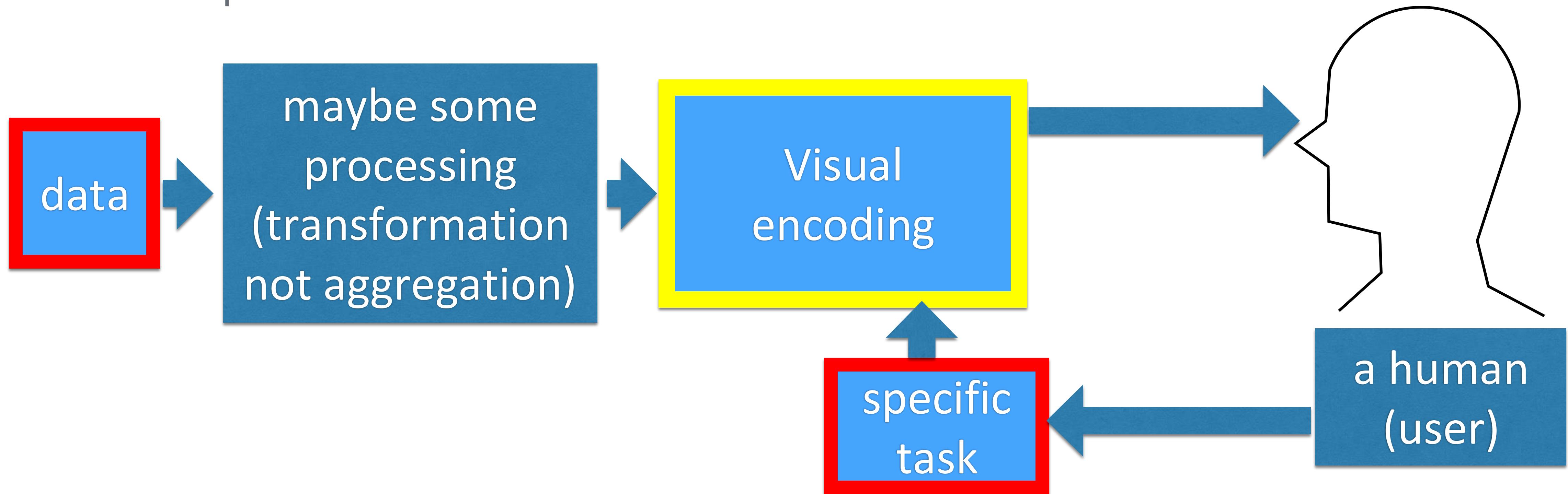
Alberto Cairo

Identificar + Dataset, Comparar + Atributos, Resumir + Atributos, otros?



Data Visualisation

- Datos. El proceso empieza con uno o más datasets. Conocemos el tipo y las características de sus atributos.
- Tareas. Definición de las tareas que podemos resolver, caracterizadas como acción + objetivo



EJERCICIO 1

- Descargar el archivo **filmdeathcounts.csv** (alternativa **.xlsx**)
- Abrir el archivo (Excel, Preview de mac, GoogleSheets, etc) y analizar el dataset.
 - Anota qué tipo de dataset es, qué tipos de datos contiene, y de qué tipo son los atributos.
- Abrir la web www.datawrapper.de /Click en “Start Creating”. Cargar el archivo
- [Opcional: Crear una cuenta para guardar las gráficas]

Round 2.

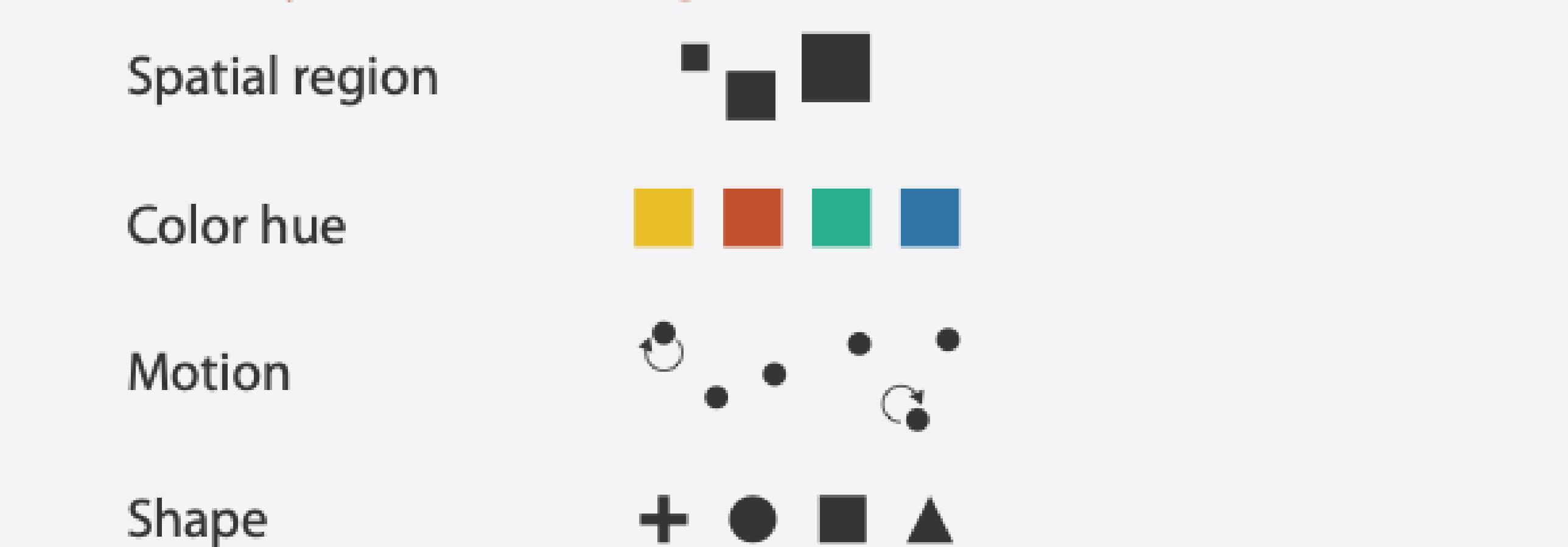
- ¿Cuál es la gráfica más adecuada para **analizar/Descubrir la correlación entre dos variables cuantitativas?** ¿Y entre tres?
- Experimenta con las opciones de customización (labels, sort, colores, swap de ejes, etc.)
- Tips: Activar Automatic labeling / Customize tooltip

Channels: Expressiveness Types And Effectiveness Ranks

→ Magnitude Channels: Ordered Attributes



→ Identity Channels: Categorical Attributes



Canales visuales separados en Magnitud e Identidad
Orden vertical según efectividad
Principio de expresividad:

- Canales de Identidad son la elección correcta para atributos categóricos sin orden o relación intrínseca
- Canales de magnitud para atributos con orden inherente: ordinales y cuantitativos

What?

Datos

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

Clusters,
Sets, Lists

Attributes

Links

Positions

Items

Positions

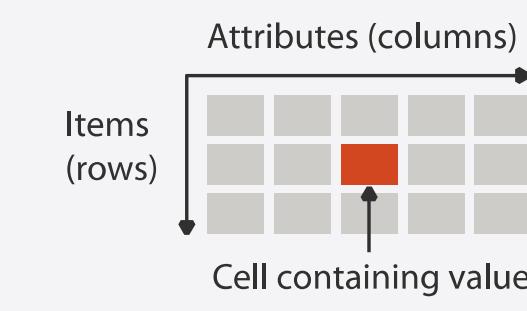
Attributes

Attributes

Attributes

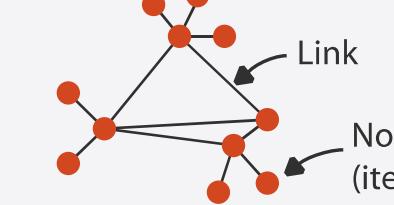
➔ Dataset Types

➔ Tables

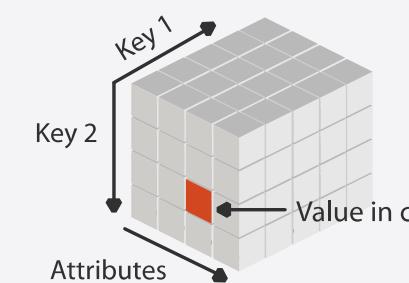


➔ Multidimensional Table

➔ Networks



➔ Trees



➔ Geometry (Spatial)



Attributes

➔ Attribute Types

➔ Categorical

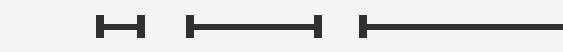


➔ Ordered

➔ Ordinal



➔ Quantitative



➔ Ordering Direction

➔ Sequential



➔ Diverging



➔ Cyclic



Bibliografia

- T. Munzner, **Visualization Analysis & Design**, CRC Press, 2014.

GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

Classes teòriques:
Teoria 3, Teoria 6, Teoria 7, Teoria 9.

Alvaro Corral Cano

alvaro.corral@uab.cat

Judit Chamorro Servent

judit.chamorro@uab.cat

Departament de Matemàtiques

Data processing for visualization

- Chapter 3 (today) - Data processing for visualization (I)
 - Uncertainty and error
 - Transformations and data massage (seminars)
- Chapter 6 (11/03) - Data processing for visualization (II)
 - Dimensionality reduction
 - Computation and important metrics selection
- Chapter 7 (21/03) - Advanced systems (I)
 - Múltiples variables i múltiples dimensions
 - Xarxes
 - Camps de vectors
- Chapter 9 (08/04) - Advanced systems (II)
 - Dades 3D
 - Visualització científica
 - Mapes

GRAU EN ENGINYERIA DE DADES
104365 Visualització de Dades

Teoria 3. Tractament de dades I

3. Data processing for visualization (I). Contents:

1. Introduction of Visualizing errors & uncertainty
2. Error
3. Uncertainty
4. Transformation and data massage (mostly in seminar 4 & practices)

3. Visualizing errors & uncertainty. Contents:

1. Introduction of Visualizing errors & uncertainty

2. Error

1. Introduction
2. Residual error (absolute error, square error, percentage error)
3. Error (graded) bars and confidence bands
4. Systematic errors & random errors
5. Statistical concepts for visualization errors (descriptive analysis/distributions)
6. Visualizing random errors

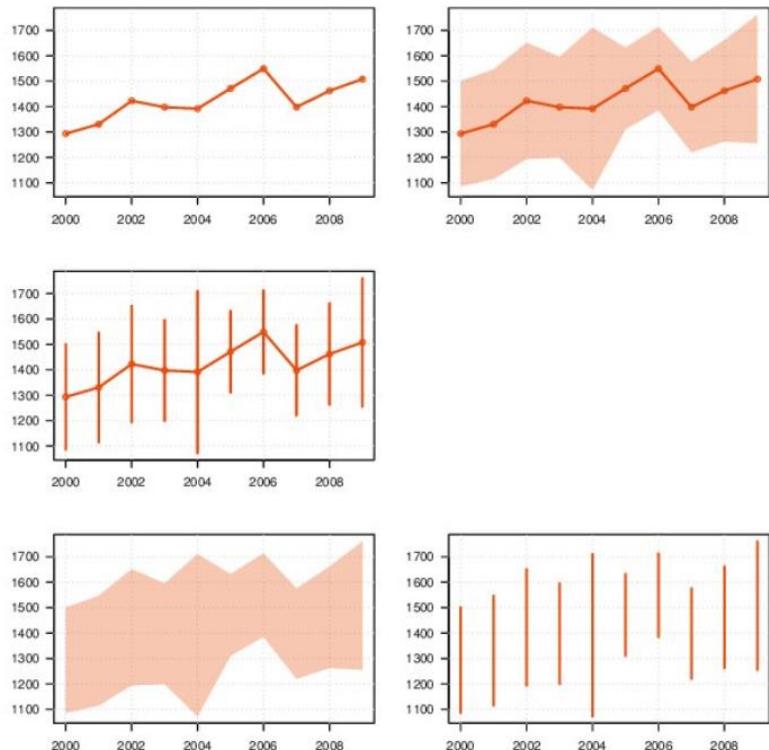
3. Uncertainty

1. Introduction
2. Uncertainty visualization: Confidence bands. Frequency framing. Standard Error.
3. Dynamic uncertainty visualization: Curve fits and Hypothetical outcome plots
4. Bayesian tools to determine distributions (Monte Carlo simulation). And to normalize them (Central Limit Theorem)

3.1 Introduction: Visualizing errors

Effect of Displaying Uncertainty in Line and Bar Charts – Presentation and Interpretation

Article - January 2015
DOI: 10.5220/0005300702250232



Five types of line charts: *line*,
ribbon+line, *error bar+line*,
ribbon, *error bar*

Edwin de Jonge

3.1 Introduction: Visualizing errors



(a)



(b)



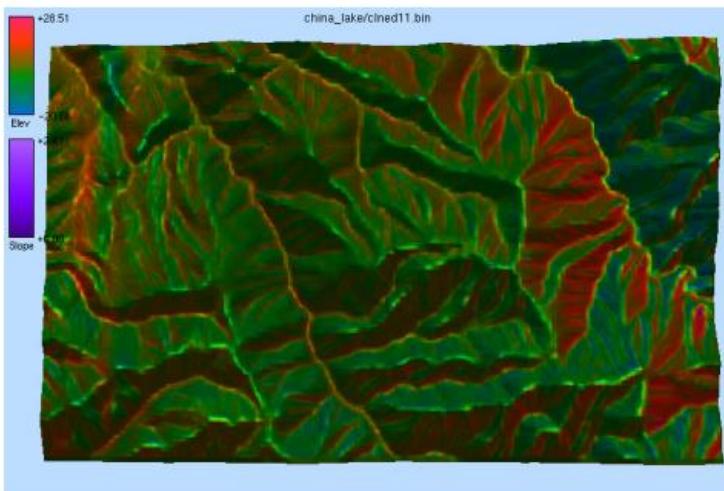
(c)

VisTRE: A Visualization Tool to Evaluate Errors in Terrain Representation

Christopher G. Healey
Computer Science Department
North Carolina State University
Raleigh, NC 27695-8206, USA

Jack Snoeyink
Computer Science Department
UNC Chapel Hill
Chapel Hill, NC 27599-3175, USA

Elevation error



(d)



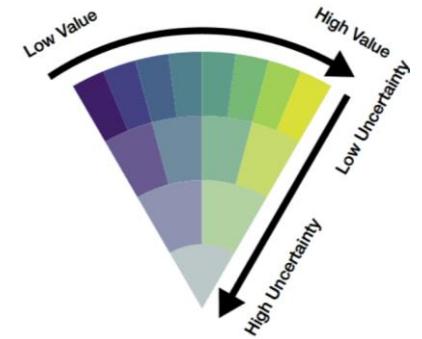
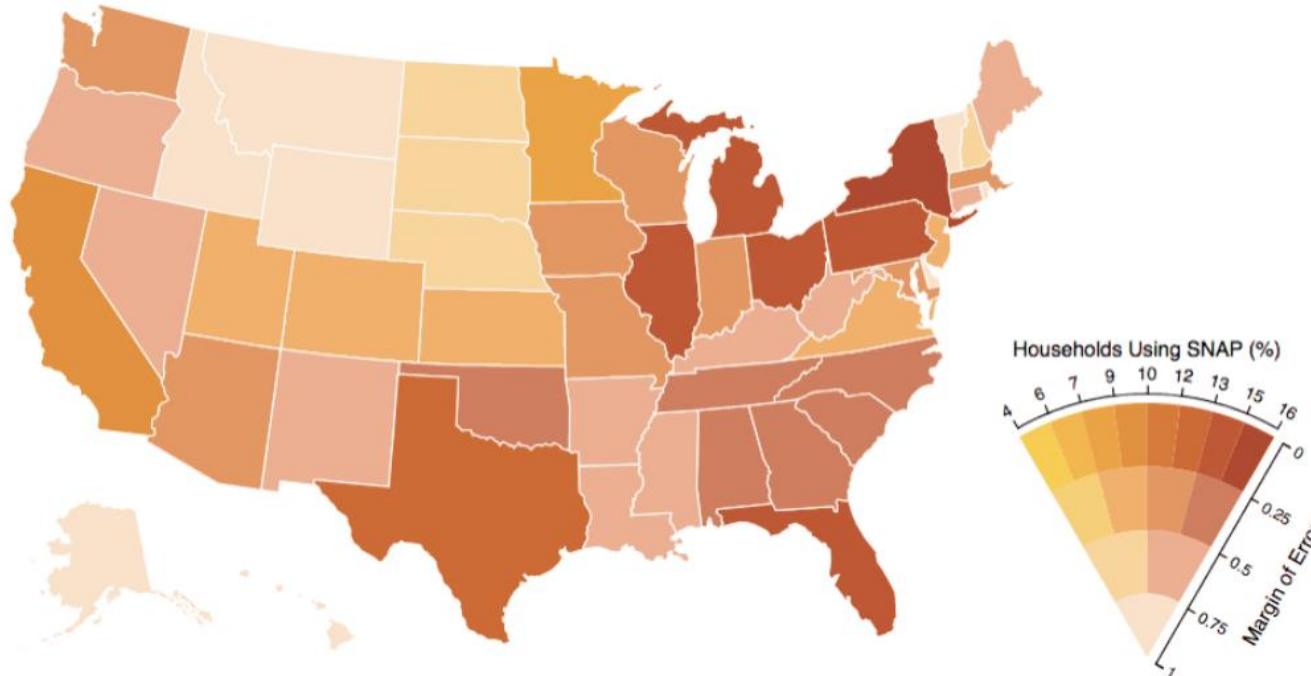
(e)

Elevation error
combined with
Slope error

Figure 2: Examples of different visual features used to represent error values: (a) elevation error from the national elevation dataset (NED) terrain model at 1° resolution visualized using hue; (b) elevation error visualized using luminance; (c) elevation error visualized using size; (d) NED terrain model with elevation error visualized using hue and slope error visualized using luminance; (e) elevation error visualized using hue and slope error visualized using size

Healey & Snoeyink

3.1 Introduction: Visualizing errors



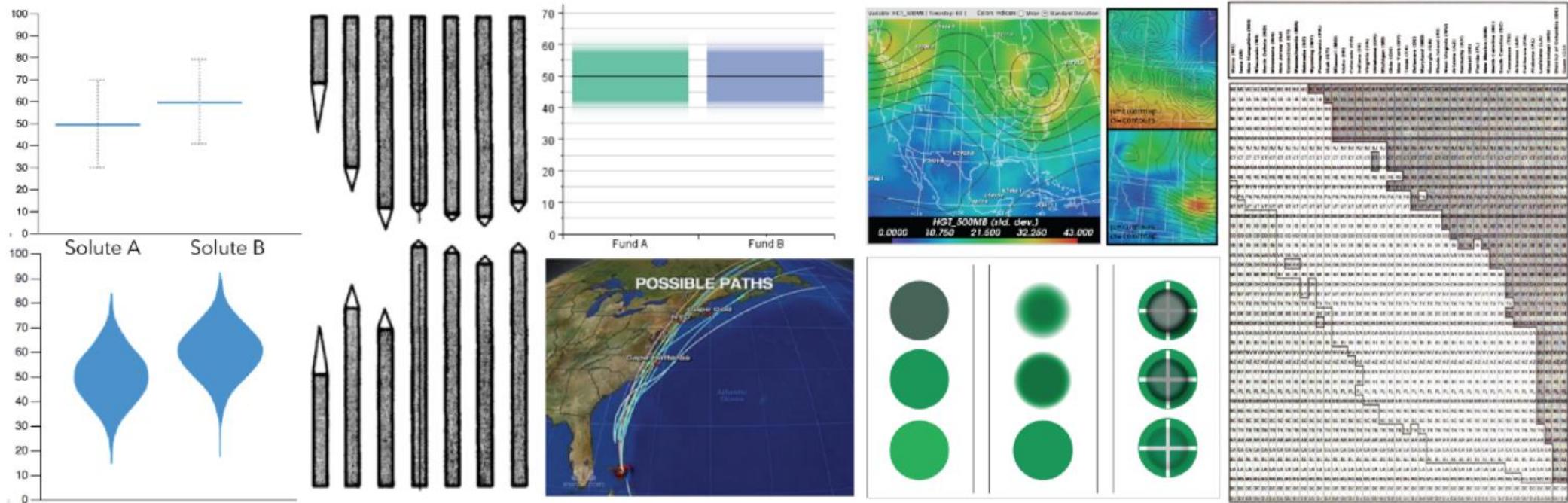
Also useful for uncertainty

Fig. 9: A map combining measurements by region and showing the margin of error in the measurement using ranges of color saturation.

Color saturation

Robert Falkowitz, 2020

3.1 Introduction: Visualizing uncertainty



Techniques for visualizing uncertainty (clockwise from top left): error bars, using negative space to convey confidence intervals on random variables, gradient plots, ensemble visualization, a matrix in which three different shades convey the reliability of precomputed comparisons, visual encodings like saturation, fuzziness, and transparency, possible hurricane paths on a news weather forecast, violin plots.

<https://medium.com/hci-design-at-uw/hypothetical-outcomes-plots-experiencing-the-uncertain-b9ea60d7c740>

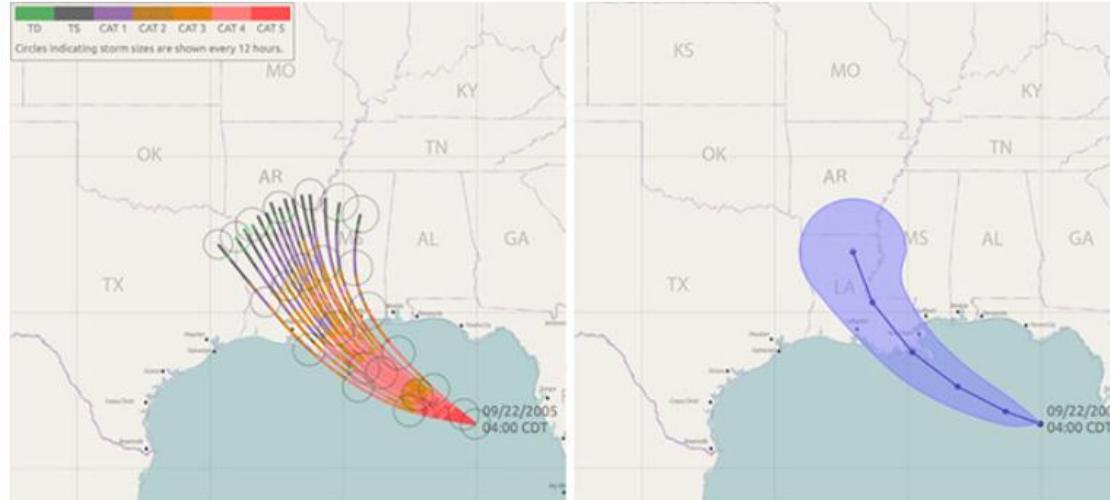
3.1 Introduction: Visualizing uncertainty

How data visualizations can clarify and confound uncertainty

Jessica Hullman's Scientific American article weighs pros and cons of common data visualizations

OCT 15, 2019

Spaghetti plot showing an ensemble of predictions



Two approaches to visualizing uncertainty in hurricane paths. From Liu, Padilla, Creem-Regehr, and House. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks (2019)

<https://www.mccormick.northwestern.edu/computer-science/news-events/news/articles/2019/how-data-visualizations-can-clarify-and-confound-uncertainty.html>

!!! Cone of uncertainty is NOT the area in which the hurricane may lie.

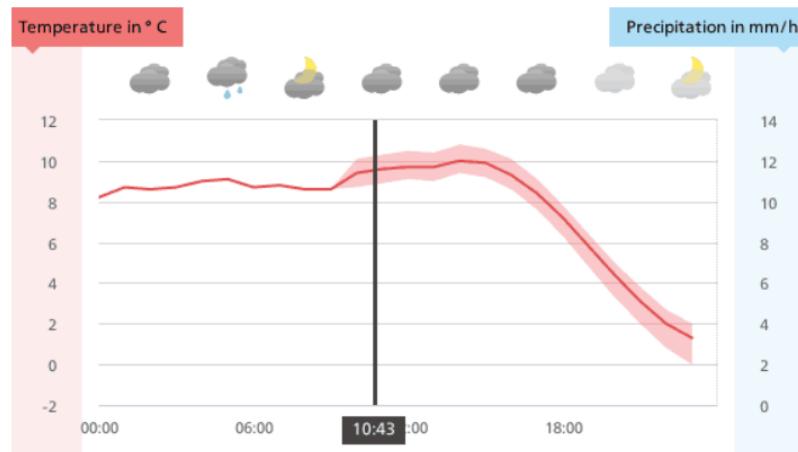
The cone is intended to indicate that the storm may take a multitude of paths, and that *it is harder to predict its location further into the future*.

3.1 Introduction: Visualizing uncertainty



A visual representation of temperature forecast uncertainty in the news. Visualizations for the general public (e.g., on TV or in newspapers) often provide **no explicit information on the uncertainty that is shown, and the viewer needs to adopt a model to interpret the underlying distribution.**

Alexander Toet, 2016



Robert Falkowitz, 2020

Fig. 13: Using shading to indicate a range of probable future measurements

3.1. Introduction: Revealing Error & Uncertainty

Revealing errors and uncertainty:

If you have information about the errors or uncertainty present in your data, whether it be

- **from a model**
- **or from distributional assumptions,**

it is a good idea display it.

3.2.1. Introduction: Potential sources of ERROR

In statistics, the entire set of raw data that you may have available for a test or experiment is known as the **population**.

Statistics allows us to take a **sample**, perform some computations on that set of data.

Potential sources of error

in estimating a population distribution using a sample

Sampling
error

Non-sampling error

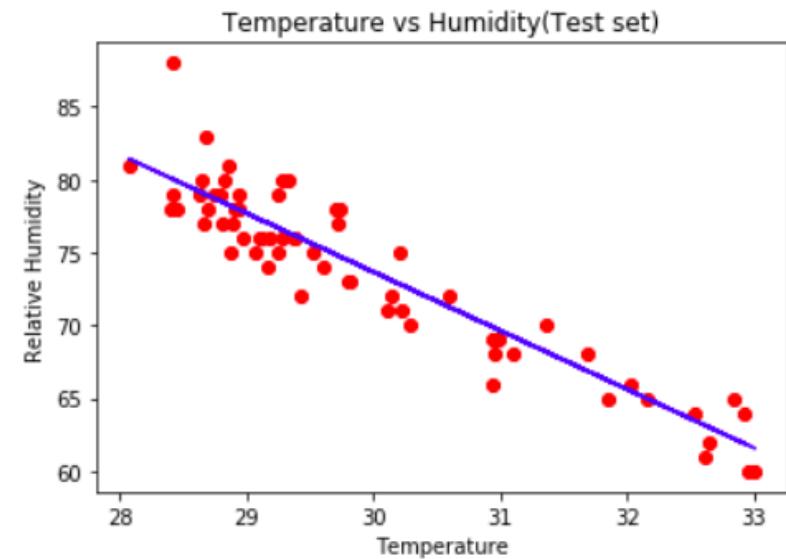
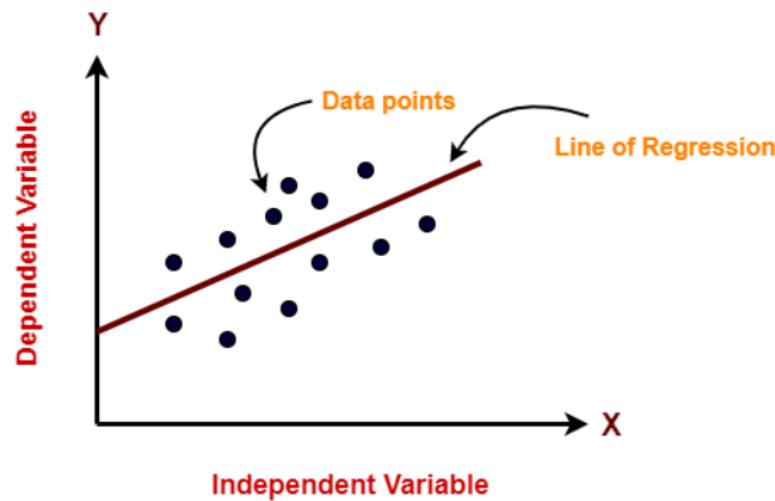
Image from Creative maths

Using probability and some assumptions we can **with a certain degree of certainty** understand trends for the entire population or predict future events.

3.2.1. Introduction: Example – quality of a model

Example: linear regression – used in supervised machine learning.

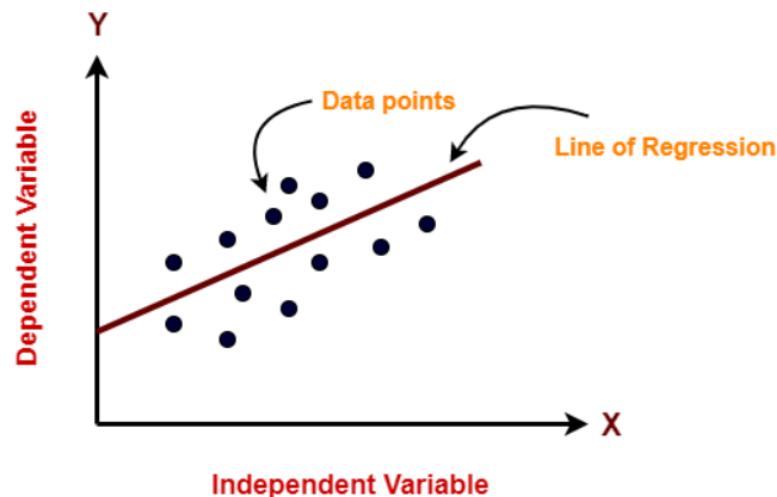
As the name suggests, linear regression follows the **linear mathematical model for determining the value of one dependent variable from value of one given independent variable**.



3.2.1. Introduction: Example – quality of a model

Example: linear regression – used in supervised machine learning.

As the name suggests, linear regression follows the linear mathematical model for determining the value of one dependent variable from value of one given independent variable. **Do you remember the linear equation from school?**



$$y = ax$$

where y is the dependent variable (usually quantitative), a is the slope, x is the independent variable.

If the line does not cross the origin $(0,0)$ we can have $y=c+ax$.

3.2.1. Introduction: Example – quality of a model

Example: to judge the quality of a model and enable us to compare regressions against other regressions with different parameters, error metrics can help.

Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output Coefficients Input Error

Linear
regression

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

3.2.1. Introduction: Error metrics Example - regression model

Error metrics will be able to judge the differences between predictions and actual values.

!!!But we cannot know how much the error has contributed to the discrepancy

The **fitted (or predicted)** values are the \hat{y} -values that you would expect for the given x -values according to the built regression model (or visually, the best-fitting straight regression line).

The **quality of a regression model** is how well its predictions (\hat{y}) match up against actual values (x).

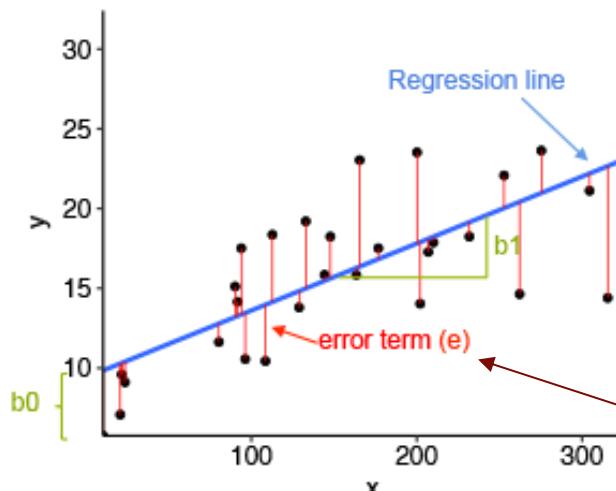
The image contains two diagrams illustrating linear regression equations. The top diagram, titled 'Linear Regression: Single Variable', shows the equation $\hat{y} = \beta_0 + \beta_1 x + \epsilon$. The terms are labeled: \hat{y} (Predicted output), β_0 and β_1 (Coefficients), x (Input), and ϵ (Error). The bottom diagram, titled 'Linear Regression: Multiple Variables', shows the equation $\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$. It uses a green bracket under $\beta_1 x_1 + \dots + \beta_p x_p$ to indicate multiple input variables, and a green bracket under β_1 and β_p to indicate multiple coefficients.

3.2.2. Residual Error. Example - regression model

Error metrics will be able to judge the differences between predictions and actual values. !!! But we cannot know how much the error has contributed to the discrepancy

The quality of a regression model is how well its predictions match up against actual values.

Residual error: the difference between the actual value and the model's estimate.



(e does not refer to the ϵ)

$$\widehat{y} = \underbrace{\beta_0 + \beta_1 x}_{\text{Predicted output}} + \underbrace{\epsilon}_{\text{Error}}$$

Linear Regression: Single Variable

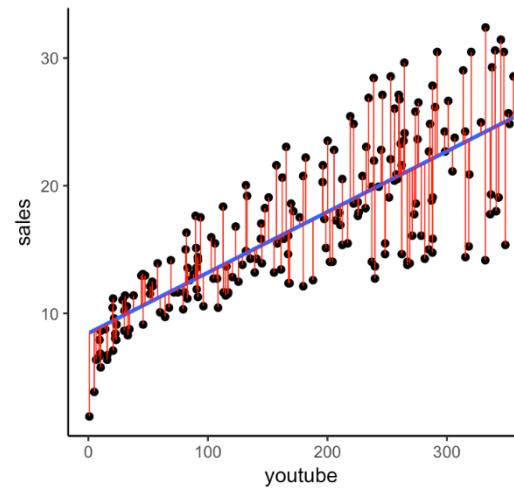
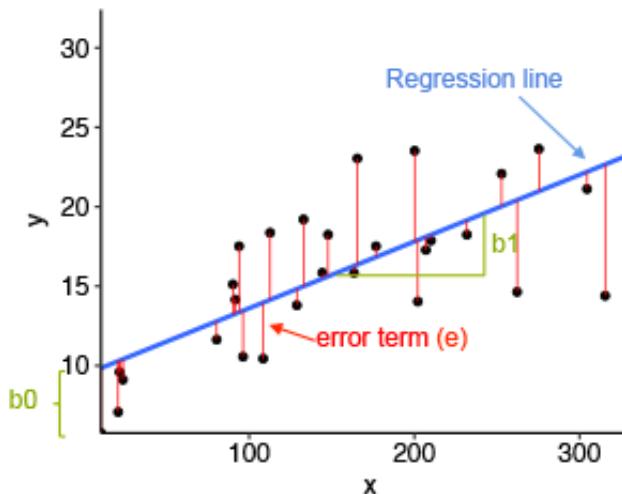
$$\widehat{y} = \underbrace{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}_{\text{Coefficients}} + \underbrace{\epsilon}_{\text{Input}}$$

Linear Regression: Multiple Variables

3.2.2. Introduction: Example – quality of a model

Error metrics – Example: to judge the quality of a model and enable us to compare regressions against other regressions with different parameters

The *quality* of a regression model is how well its predictions match up against actual values.



We build a model to predict sales on the basis of advertising budget spent in youtube medias

To check the regression assumptions, we'll examine the distribution residuals

3.2.2. Residual Error: Mean absolute error (MAE)

Mean absolute error (MAE): is the simplest regression error metric. It **describes** the typical **magnitude of the residuals**

1. We'll calculate the residual for every data point, taking only the absolute value of each (negative and positive residuals do not cancel out).
2. Take the average of all the residuals.

$$MAE = \frac{1}{n} \sum \left| \text{Actual output value} - \text{Predicted output value} \right|$$

Divide by the total number of data points

Predicted output value

Actual output value

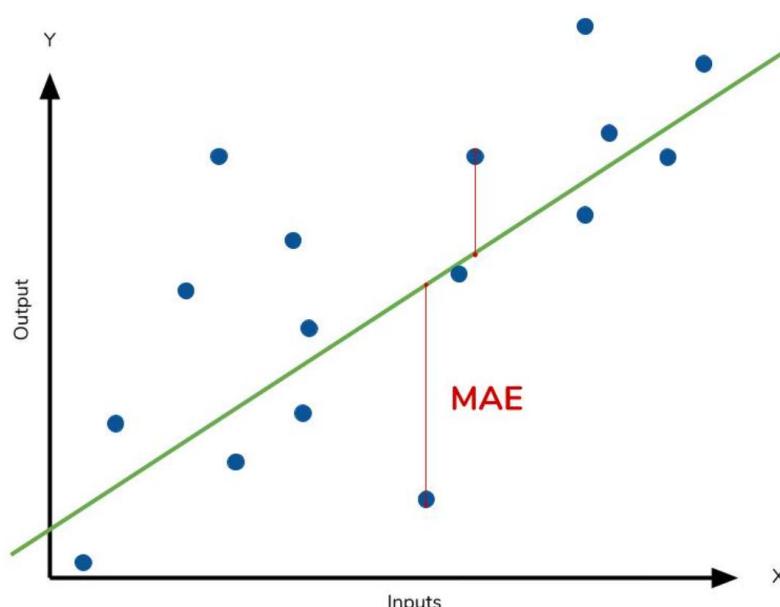
Sum of

The absolute value of the residual

3.2.2. Residual Error: Mean absolute error (MAE)

Limitations of MAE:

- It does not indicate underperformance or overperformance of the model/experiment.
- It does not bring attention to the outliers (or extrema values)



Green line represents the model's predictions

Blue points represent the data

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Divide by the total number of data points
Actual output value
Predicted output value
Sum of
The absolute value of the residual

3.2.2. Residual Error: Mean square error (MSE)

Mean square error (MSE): is just like MAE, but squares the difference before summing them all instead of using the absolute value

$$MSE = \frac{1}{n} \sum \underbrace{\left(y - \hat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

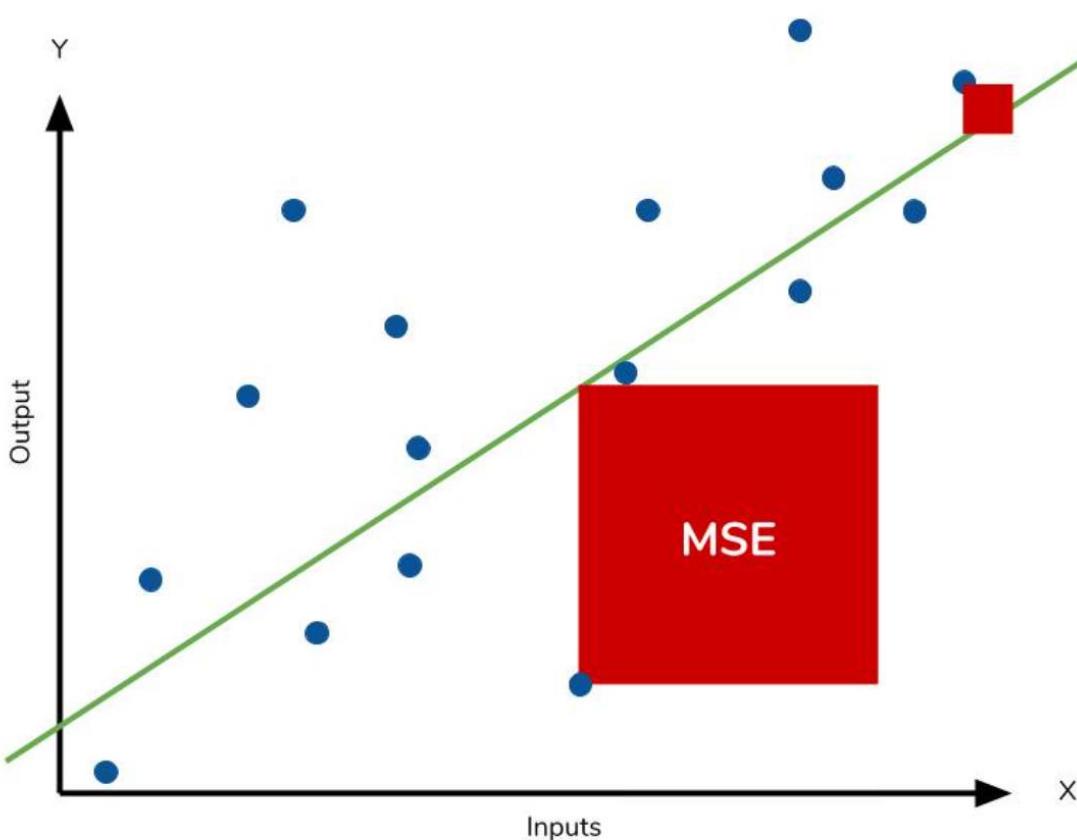
Because we are squaring the difference, the **MSE will almost always be bigger than the MAE** (we cannot compare MAE to the MSE)

The effect of the square term is most apparent with the presence of outliers in our data – **Outliers in our data will contribute to much higher total error in MSE than they would the MAE.**

3.2.2. Residual Error: Mean square error (MSE)

Mean square error (MSE):

!!! Outliers will produce these exponentially larger differences



Green line represents the model's predictions

Blue points represent the data

3.2.2. Residual Error: MAE vs MSE

MAE or MSE??

- To downplay **the outlier's significance**, we would use the MAE since the outliers' residuals won't contribute as much to the total error as MSE.
- **The choice between MSE and MAE is application-specific** and depends on how you want treat large errors.
- Both **MAE and MSE can range from 0 to positive infinity**, so as both measures get higher, it becomes harder to interpret how well your model is performing.

3.2.2. Residual Error: Root mean squared error (RMSE)

Root mean squared error (RMSE) is the square root of the MSE.

MSE vs RMSE:

- RMSE is often used to convert the error metric back into similar units, making interpretation easier.
- The effect of the outliers in MSE and RMSE is similar. They both square residual.
- The RMSE is analogous to the standard deviation (MSE variance) and measure of how large your residuals are spread out.

3.2.2. Residual Error: Mean percentatge error (MPE)

Mean percentage error (MPE): it lacks the absolute value operation

$$MPE = \frac{100\%}{n} \sum \left(\frac{y - \hat{y}}{y} \right)$$

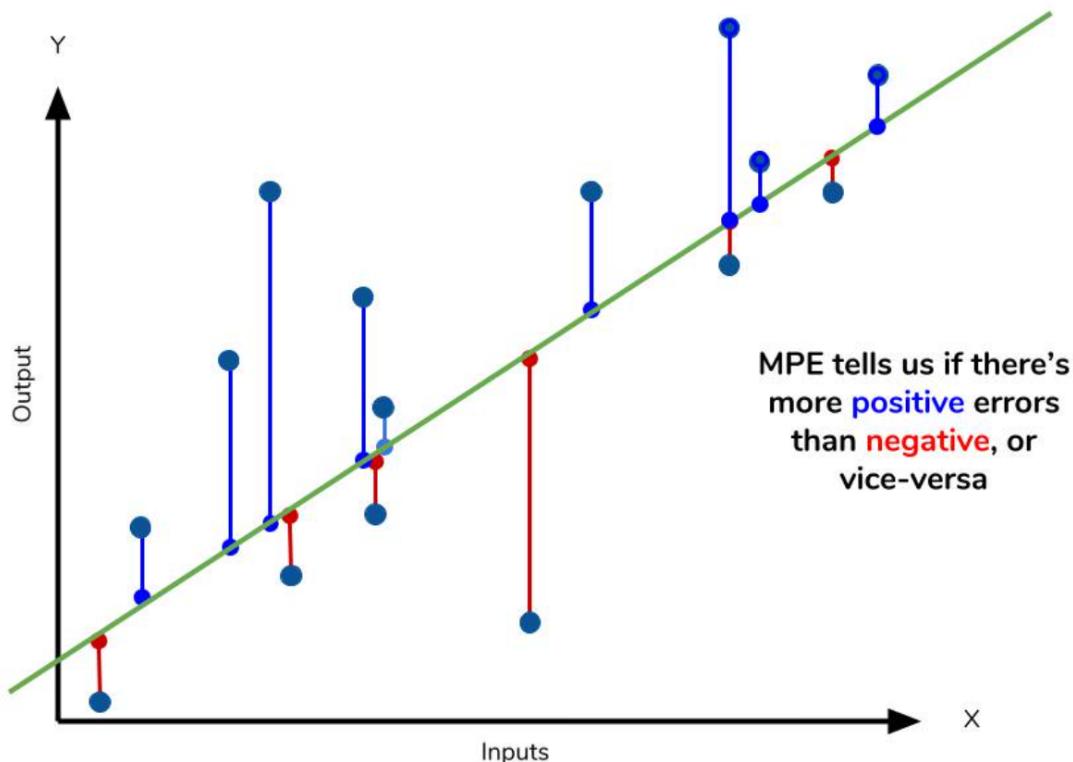
However, unlike MAE and MPE, MPE **allows us to see if our model systematically underestimates or overestimates** (bias).

Underestimates -> MPE more negative

Overestimates -> MPE positive

3.2.2. Residual Error: Mean percentage error (MPE)

Mean percentage error (MPE):

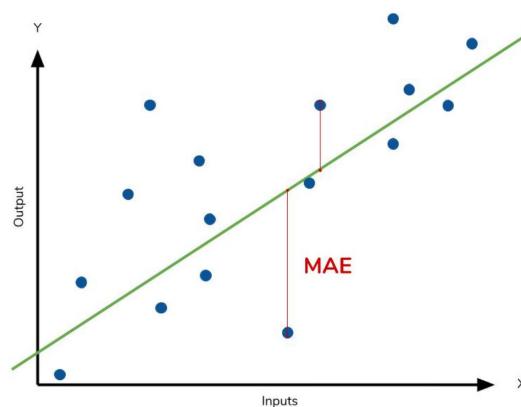


Green line represents the model's predictions

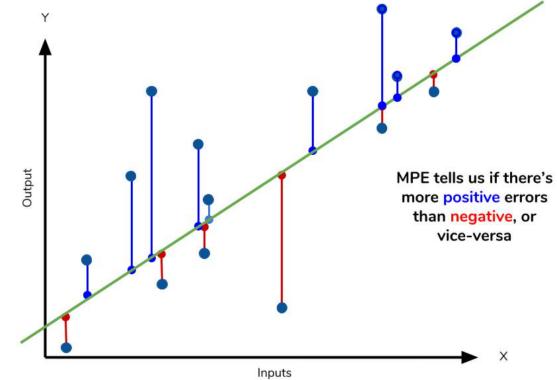
Blue points represent the data

3.2.2. Residual Error – Metric's summary

Acronym	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MPE	Mean Percentage Error	N/A	Yes

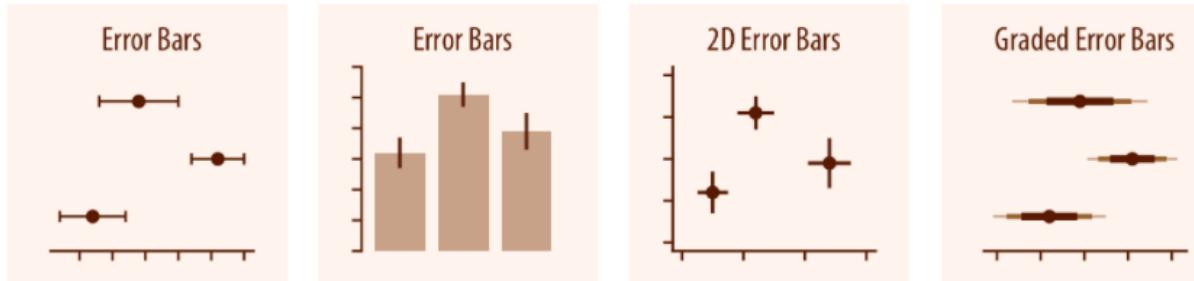


Be specific



To check the regression assumptions, we'll examine the distribution residuals.

5.3.2 Uncertainty visualization: Error bars

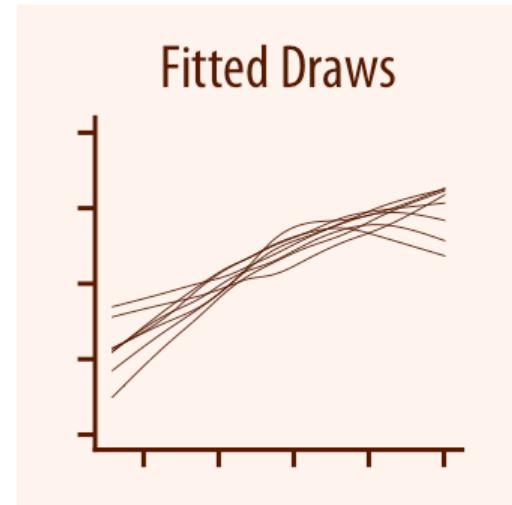
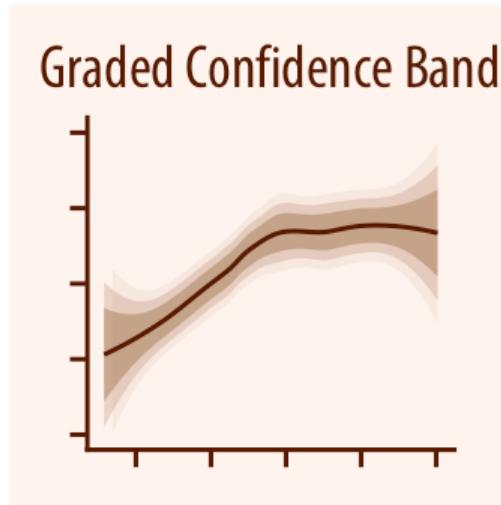
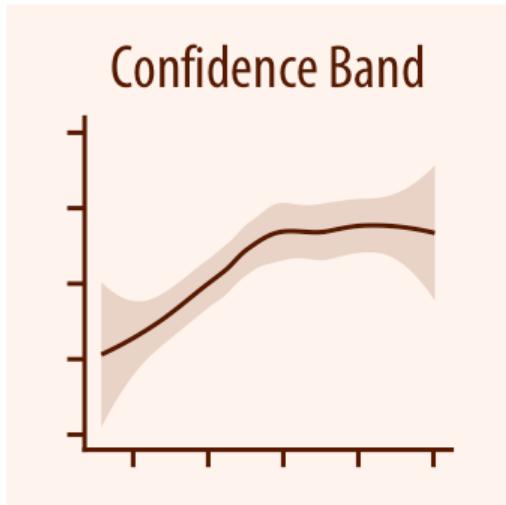


Claus Wilke

- **Error bars are meant to indicate the range of likely values for some estimate or measurement.** They extend horizontally and/or vertically from some reference point representing the estimate or measurement.
- **Graded error bars** show multiple ranges at the same time, where each range corresponds to a different degree of confidence. They are in effect multiple error bars with different line thicknesses plotted on top of each other.

3.2.3 Visualizing errors (confidence bands)

Confidence band: the equivalent of an error bar for smooth line graphs.

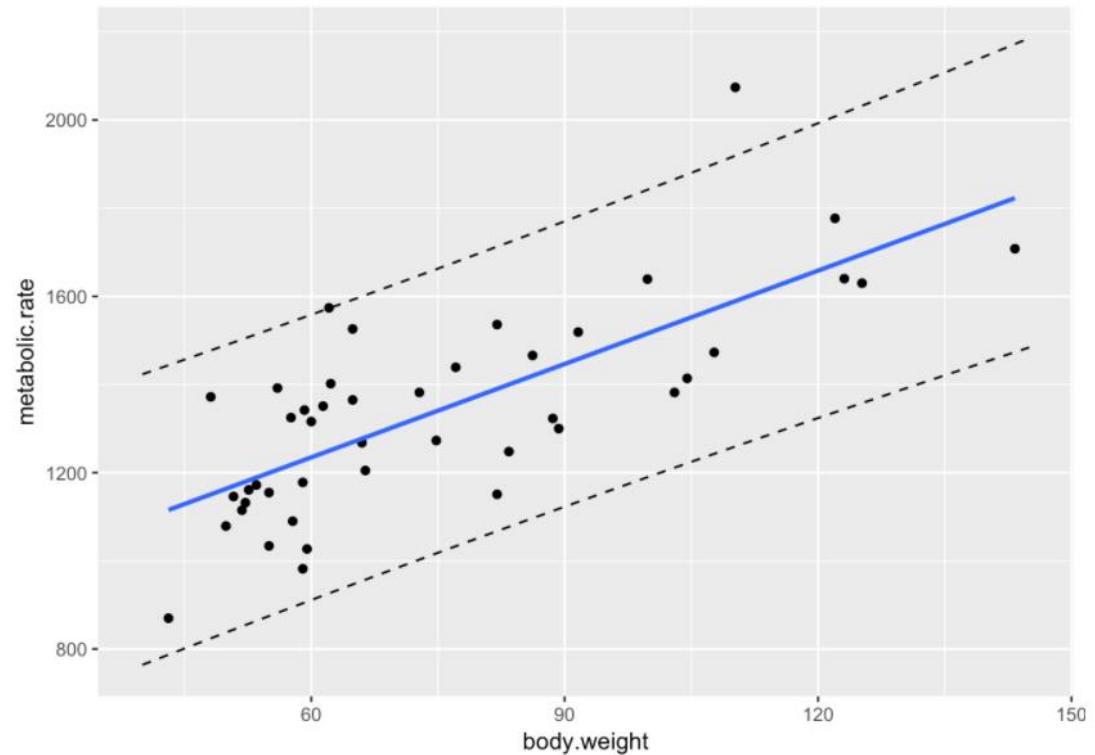


Claus O.Wilke

3.2.3. Confidence bands

Prediction Intervals - Suppose you want to predict the metabolic rate of a new patient, whose body weight is known. How large is your error?

Regression coefficients give you an estimate. However, **to know how large is the error, you need to use prediction intervals.**

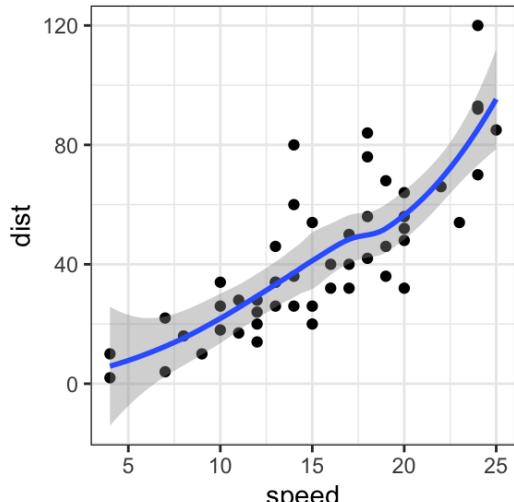


About 95% of the data points do fall within the bands

3.2.3. Confidence bands

- One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.
- (seminars) GGPlot: Adding the canal ‘geom_smooth’ or ‘stat_smooth()’: aids the eye in seeing patterns in the presence of **overplotting**. See ?geom_smooth or ?stat_smooth

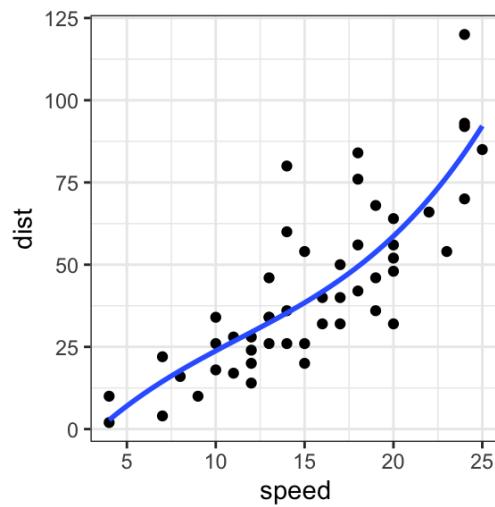
Example:



One can play with one of the arguments for stat-smooth which is level: level of confidence Interval to use (0.95 by default)

3.2.3. Confidence bands

- One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.
- (seminars) GGPlot: Adding the canal ‘**geom_smooth**’ or ‘**stat_smooth()**’: aids the eye in seeing patterns in the presence of **overplotting**. See ?geom_smooth or ?stat_smooth



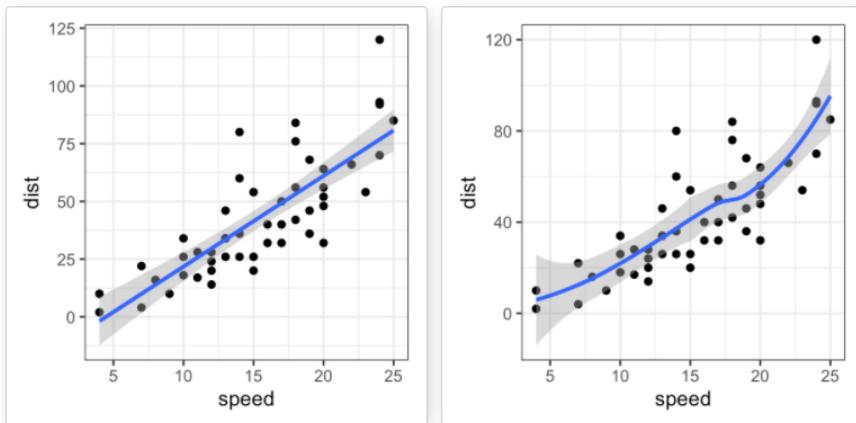
Polynomial interpolation

```
# Remove the confidence bande: se = FALSE  
p + geom_smooth(method = "lm", formula = y ~ poly(x, 3), se = FALSE)
```



3.2.3. Confidence bands

- One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.
- Adding the canal ‘geom_smooth’ or ‘stat_smooth()’: aids the eye in seeing patterns in the presence of overplotting.
By default, the method is: local regression fitting (“loess”)



```
p <- ggplot(cars, aes(speed, dist)) +  
  geom_point()  
# Add regression line  
p + geom_smooth(method = lm)  
  
# loess method: local regression fitting  
p + geom_smooth(method = "loess")
```

There is also the method “lm” (**linear regression**) & “glm” (**generalized linear model**)

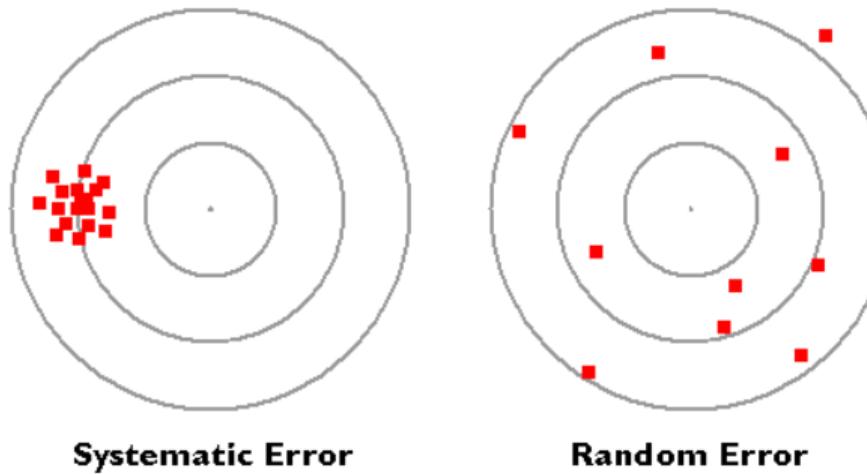
3.2.4 Systematic vs. Random Errors

- **Systematic errors:**
 - Errors that **affect all measurements in the same way**.
 - Systematic errors **have a determinate origin (there is a cause)**.

Example: Not calibrating the measuring instrument
- **Random errors:**
 - Errors that **occur randomly and affect measurements in an unpredictable manner**.
 - **They are undetermined in origin and cause.** Random errors may occur due to carelessness or lack of concentration

3.2.4 Systematic vs. Random Errors

- **Systematic errors:** Tend to be consistent in magnitude and/or direction (the recorded values differ from the “true” values to be measured in a way that is both consistent and predictable).
- **Random errors:** Vary in magnitude and direction



3.2.4 Graphing: Systematic vs. Random Errors

– **Systematic errors:** Tend to be consistent in magnitude and/or direction. They have an origin/cause. Two types:

- **Constant errors:** the size of the error is often **independent of measurement magnitude (not correlated)**.

It will be reflected in a *change in the ‘y-axis’ intercept* on the graph.

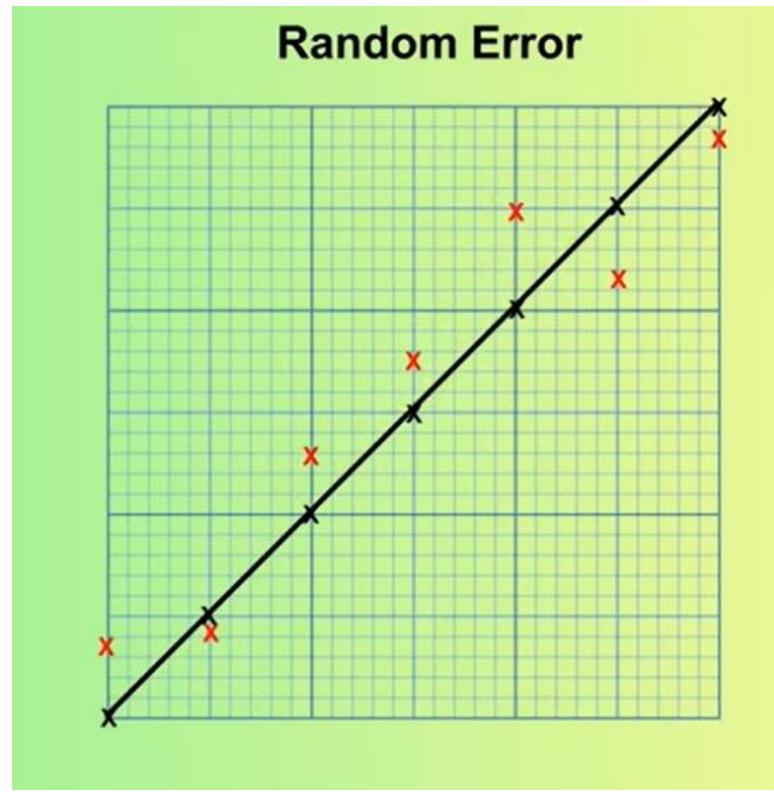
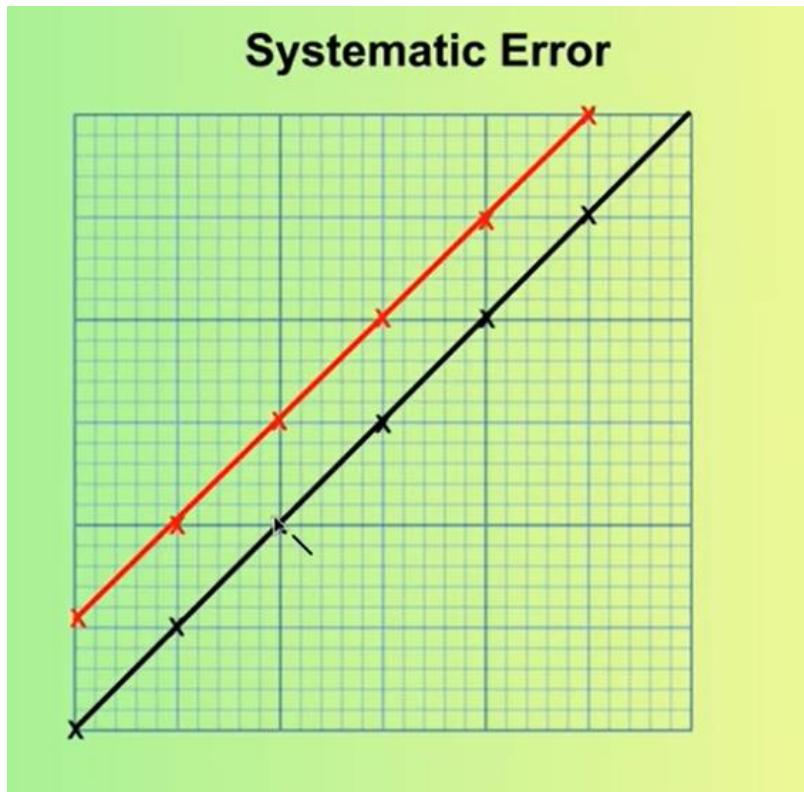
- **Proportional errors:** the error may **increase with the magnitude of the measurement**.

It will change *the slope* of the line of the graph.

– **Random errors:** **Vary in magnitude and direction.** Not origin/cause.

They will cause a *scatter plot effect* on the graph, *making the determination of the line of best fit impossible*.

3.2.4 Systematic vs. Random Errors



Scattered all around to the true value

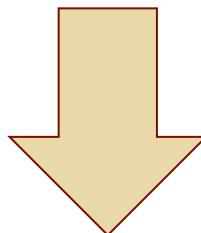
Black points represents the true data

Red points represent the experimental data

3.2.5 Visualization of Errors (statistical concepts)

– Before going depth, **let's do a summary about statistical concepts** that we will need:

1. Descriptive statistics
2. Distributions
3. Variability



Visualization of random errors

3.2.5 Visualization of Errors (statistical concepts)

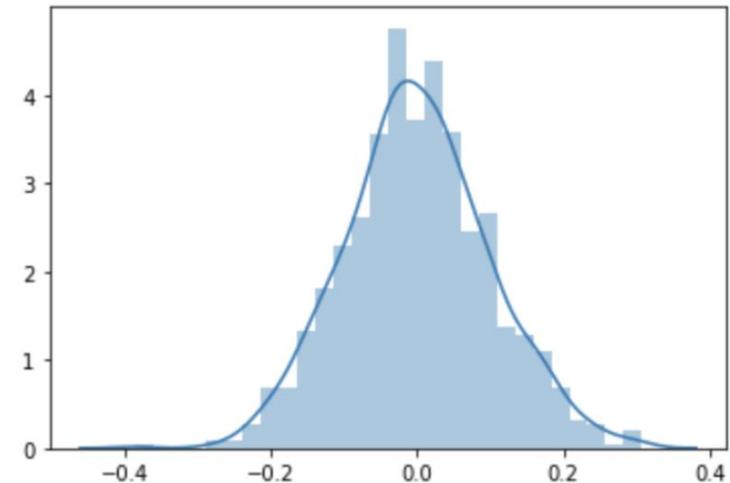
- Before going depth, **let's do a summary about statistical concepts** that we will need:
 1. **Descriptive statistics** (it simply provides a description of what the data sample we have looks like)
 - **Mean:** the **central value**, commonly called the average
 - **Median:** the **middle value** if we ordered the data from low to high and divide exactly in half
 - **Mode:** the value which occurs more often

Descriptive statistics are useful, but they **can often hide important information about the data set** (*you will see it with the Anscombe dataset on Friday*)

3.2.5 Visualization of Errors (statistical concepts)

2. Distributions

- A **distribution** is a chart, for example a histogram, that displays the frequency with which each value appears in a data set. This type of chart gives us information about the spread and skewness of the data.
- One of the most important distributions is the **normal distribution**. It is symmetrical in shape with most of the values clustering around the central peak and the further away values distributed equally on each side of the curve. Many variables in nature will form a normal distribution.



3.2.5 Visualization of Errors (statistical concepts)

3. Variability

- **Variance** measures **how far each value in the data set is from the mean**
- **Standard deviation (SD)** is a common measure of variation for data that has a normal distribution. It gives a value to represent **how widely distributed the values are.**
 - low SD: the values tend to lie quite close to the mean.
 - high SD: the values are more spread out.

3.2.5 Visualization of Errors (statistical concepts)

3. Variability

- **Variance:** how far each value in the data set is from the mean
- **Standard deviation:** how widely distributed the values are.

If the data does not follow a normal distribution, then other measures of variance are used:

– **The interquartile range:**

- Derived by first ordering the values by rank and then dividing the data points into four equal parts, called quartiles.
- Each quartile describes where 25% of the data points lie according to the median.
- The interquartile range is calculated by subtracting the median for the two central quarter (Q1 & Q3).

3.2.5 Visualization of Errors (statistical concepts)

3. Variability

- Variance: how far each value in the data set is from the mean
- Standard deviation: how widely distributed the values are.
- Interquartile (different distributions): Q1, median=Q2, Q3.

!!! Standard error versus standard deviation:

- The standard deviation is a property of the population. It tells us how much spread there is among individual observations we could make.
- The standard error tells us how precisely we have determined a parameter estimate.

The standard error is approximately given by the sample standard deviation divided by the square root of the sample size, and confidence intervals are calculated by multiplying the standard error with small, constant values.

3.2.5 Visualization of Random Errors

- **The form of the distribution of the random errors must be known.**
- Although the form of the probability distribution must be known, the parameters of **the distribution can be estimated from the data.**

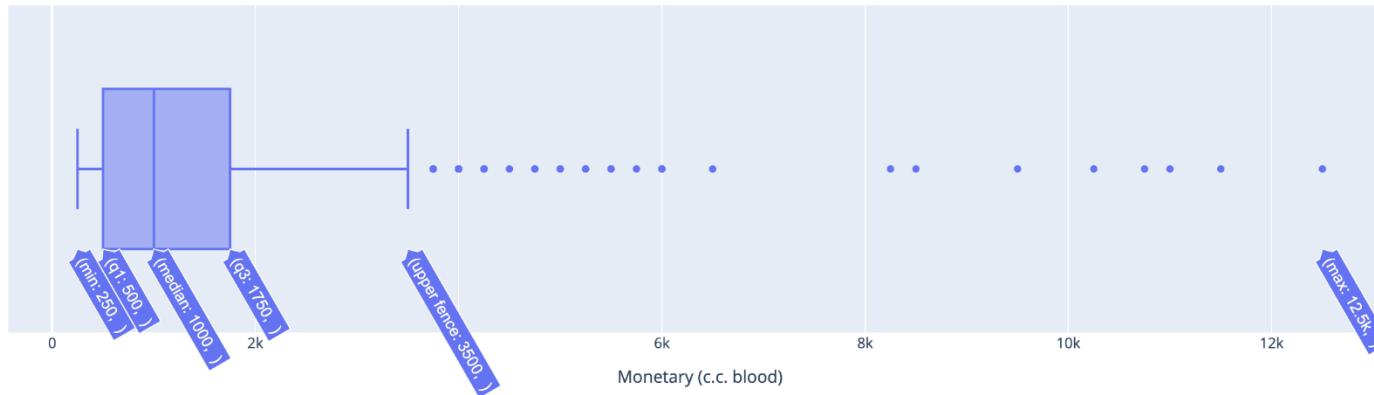
The random errors from different types of processes could be **described by any one of a wide range of different probability distributions in general**, including: the uniform, triangular, double exponential, binomial and Poisson distributions.

- **The normal distribution often describes the actual distribution of the random errors in real-world processes reasonably well** (This is related to the CLT, that we will see later today). With most process modelling methods - inferences are based on the idea that the random errors are drawn from a normal distribution.

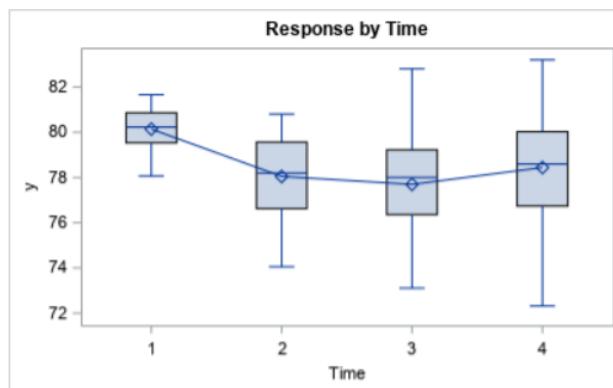
The normal distribution is also used because **the mathematical theory behind it is well-developed.**

3.2.6 Visualization of Random Errors

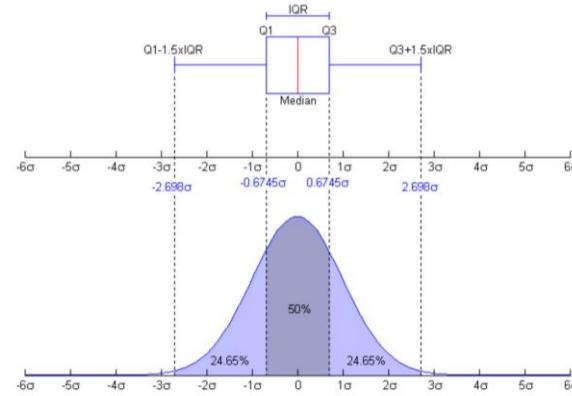
Statistical methods may be used to analyze the data & random errors



Example by Rebecca Vickery 2021: A boxplot provides a useful visualization of the interquartile range (IQR) .

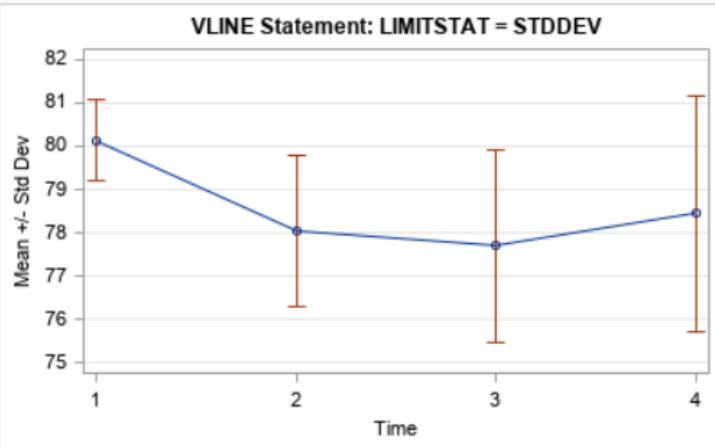


Example by Wicklin 2019: The boxplot shows the schematic distribution of the data at each point. The boxes use the **interquartile range and whiskers to indicate the spread of the data**. A line connects the means/medians of the response at each time point.

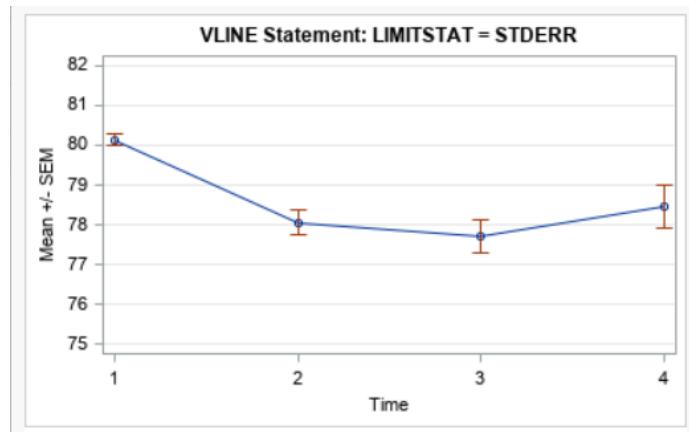


Example by Jhguch 2011: IQR of a normal distribution

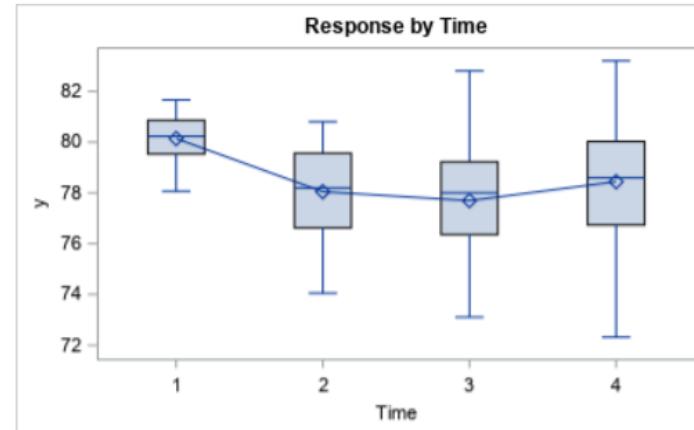
3.2.5 Visualization of Random Errors



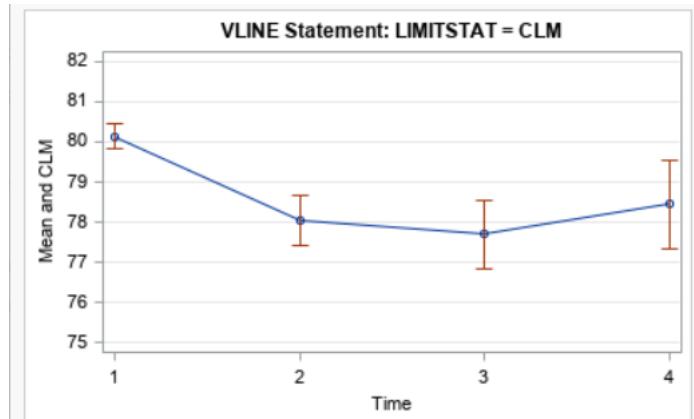
"standard deviation" is a term that is familiar to a lay audience



The exact meaning of the "standard error of the mean" might be difficult to explain to a lay audience, but the qualitative explanation is often sufficient



The boxes use the **interquartile range and whiskers to indicate the spread of the data**. A line connects the means/medians of the response at each time point.



The "confidence interval of the mean (CLM)" is hard to explain to a lay audience

3. Visualizing errors & uncertainty. Contents:

1. Introduction of Visualizing errors & uncertainty

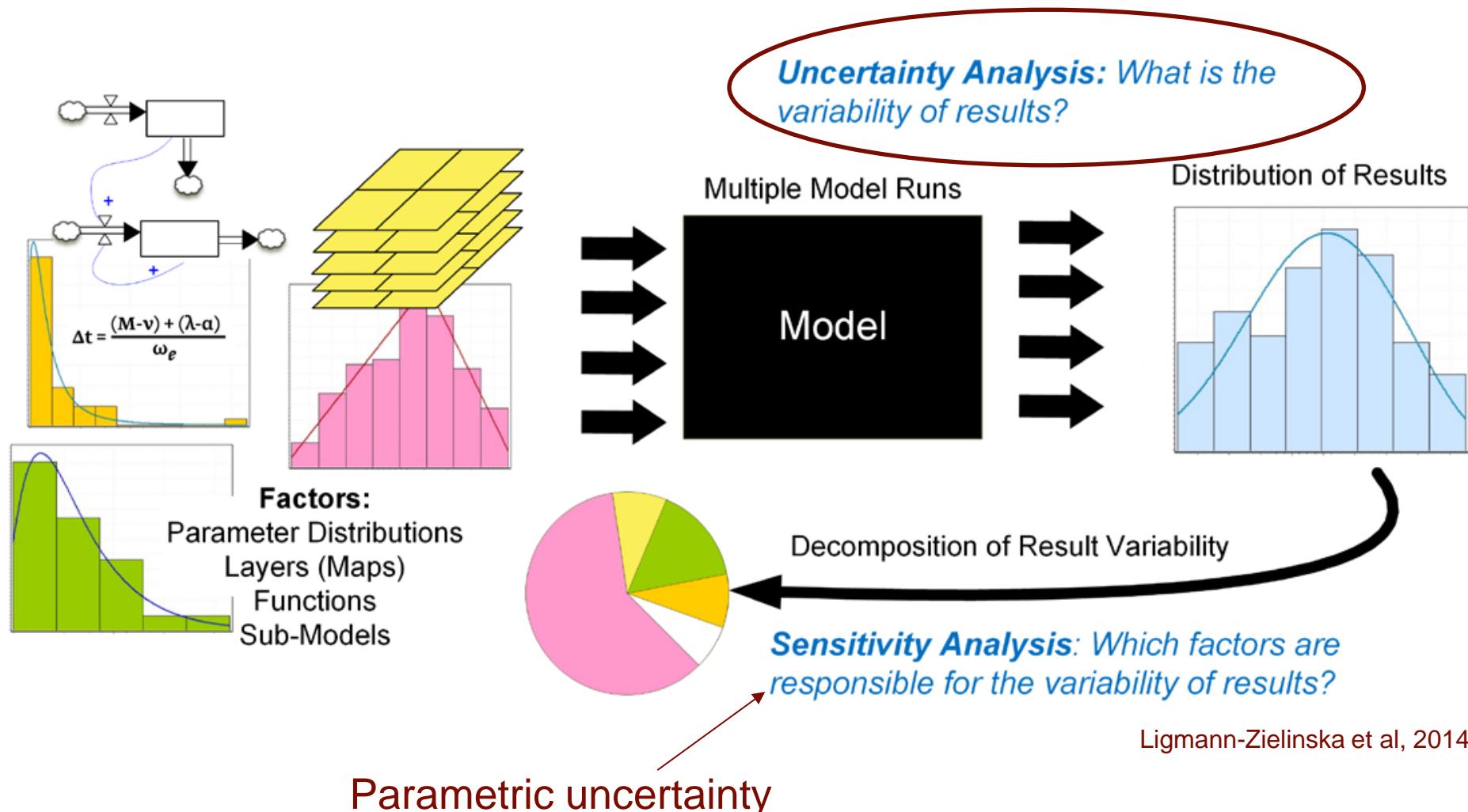
2. Error

1. Introduction
2. Residual error (absolute error, square error, percentage error)
3. Error (graded) bars, confidence strips, eyes, quantile dots and confidence bands
4. Systematic errors & random errors
5. Statistical concepts for visualization errors (descriptive analysis/distributions)
6. Visualizing random errors

3. Uncertainty

1. Introduction
2. Uncertainty visualization: Error bars. Confidence bands. Frequency framing. Standard Error.
3. Dynamic uncertainty visualization: Curve fits and Hypothetical outcome plots
4. Bayesian tools to determine distributions (Monte Carlo simulation). And to normalize them (Central Limit Theorem)

3.3.1 Introduction: Uncertainty vs sensitivity analysis

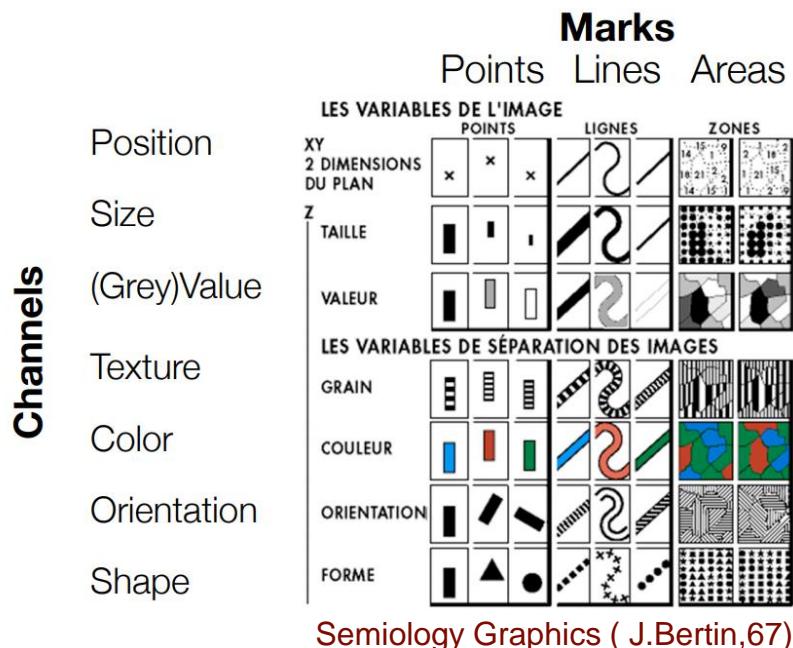


Ligmann-Zielinska et al, 2014

3.3.1 Introduction: Uncertainty

Limitations visualizing uncertainty?

- We are **limited by the number of visualization channels**.
- When **moving from quantified uncertainty to visualized uncertainty**, we often **simplify the uncertainty** to make it fit into the available visual representations.



- We need a **balance between the degree of complexity and the audience that we want to reach**.

+Other channels such as opacity, 3D positions, motion...
! Still limited
! Channels overwhelmed when increasing the amount and dimensionality of the data

3.3.1 Introduction: Uncertainty

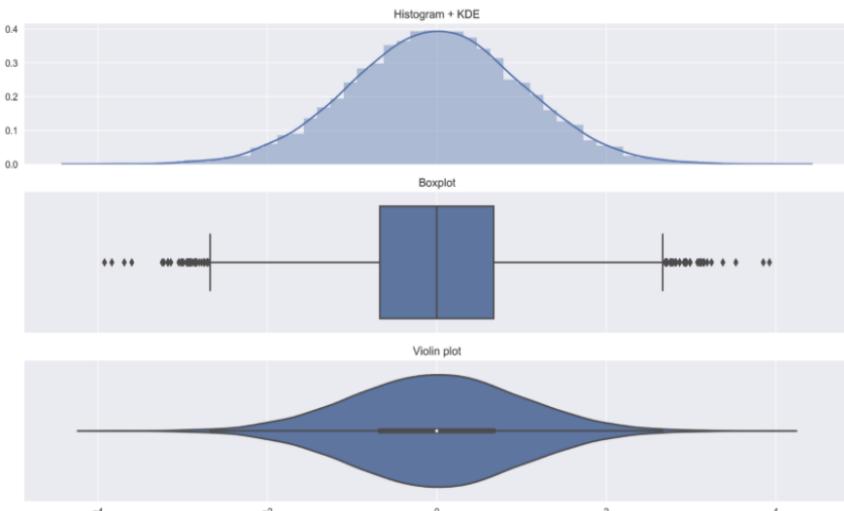
- **Epistemic uncertainties:** due to *lack of knowledge and limited data in practice* (deficient measurements, poor models, missing data)
- **Aleatoric uncertainty:** inherent random uncertainty (from running an experiment and getting slightly different results each time) -> *often represented as a probability density function (PDF)*

The most straightforward understanding of uncertainty is often the easiest to expose visually -> often thought that is entirely statistically defined

3.3.1 Introduction: Uncertainty

Non-spatial data uncertainty displays

- **Error bars**
- **Boxplots** to express variability by showing the quartiles.
- **Violin plots**, *additionally displaying the probability density* (kernel density estimation) of the data at each value.



3.3.1 Introduction: Uncertainty

Non-spatial data uncertainty displays

- **Error bars**
- **Boxplots** to express variability by showing the quartiles.
- **Violin plots**, additionally displaying the probability density (kernel density estimation) of the data at each value.

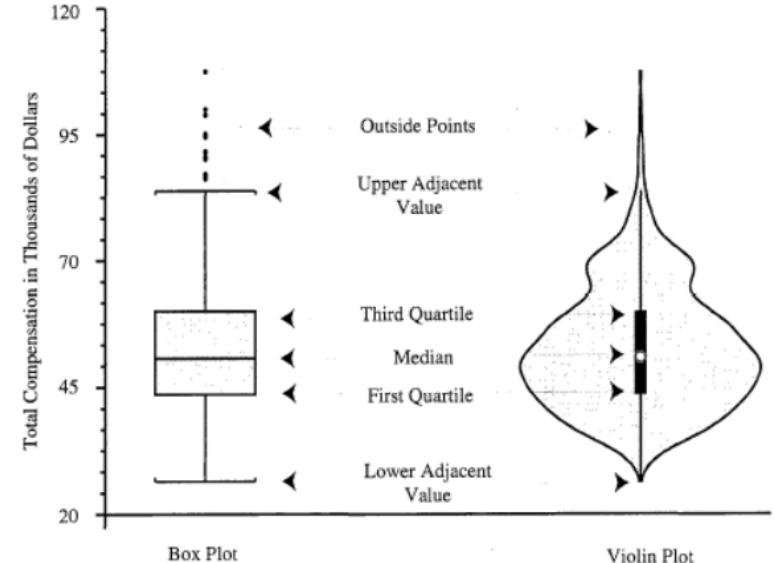


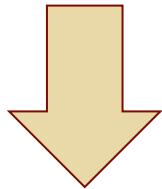
Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Hintze and Nelson 1998

3.3.1 Introduction: Uncertainty

Spatial data uncertainty metrics (like for errors)

- A measure of central tendency, such as the mean.
- An indicator of dispersion, for example the variance or standard deviation.
- Extrema (minimum and maximum, or extreme percentiles).



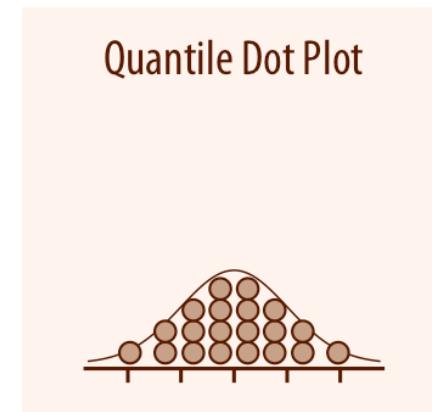
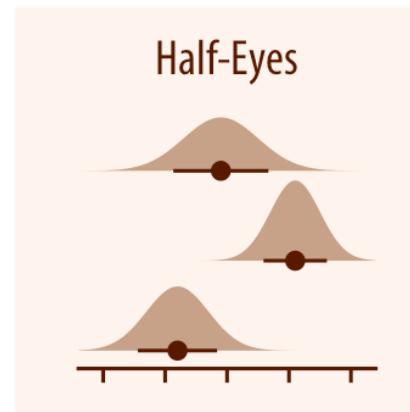
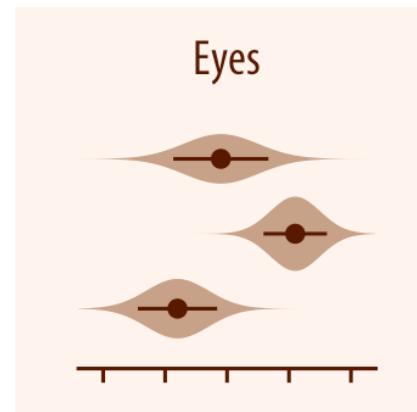
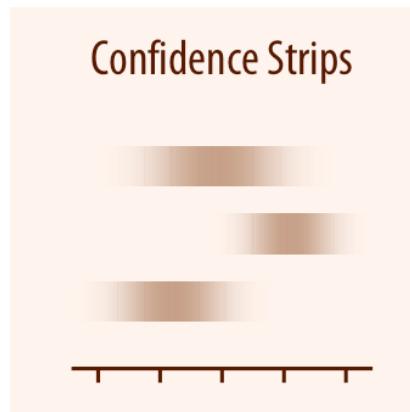
Two commonly used approaches to indicate uncertainty are error bars and confidence bands.

3.3.2 Visualizing uncertainty

Confidence strips: graduated.

Eyes and half eyes: combine error bars with methodologies to combine distribution (violins and ridgelines).

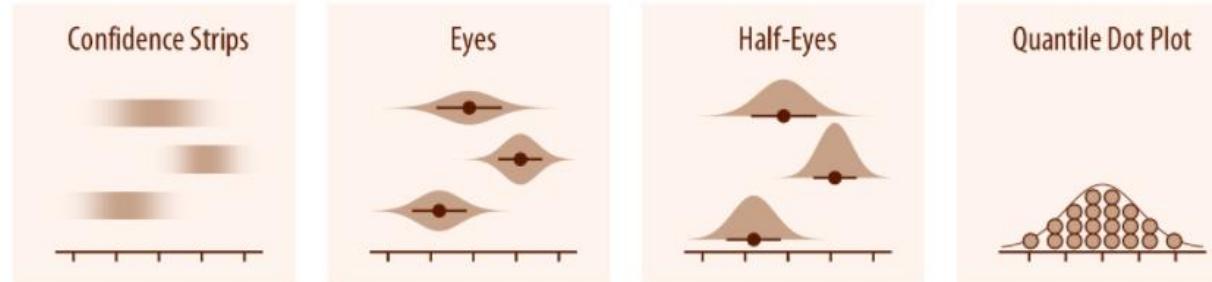
Quantile dots: the distribution in discrete units.



Claus O.Wilke

3.3.2 Visualizing uncertainty

To achieve a more detailed visualization than is possible with error bars or graded error bars, we can visualize the actual confidence or posterior distributions

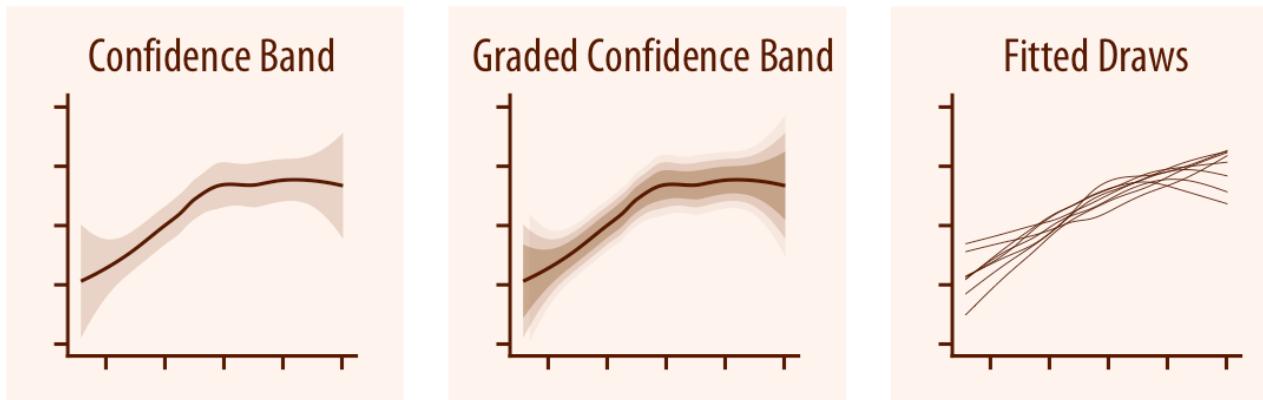


Claus Wilke

- **Confidence strips** provide a clear visual sense of uncertainty but are difficult to read accurately.
- **Eyes and half-eyes** combine error bars with approaches to visualize distributions (violins and ridgelines, respectively) -> show both precise ranges for some confidence levels and the overall uncertainty distribution.
- **A quantile dot plot** can serve as an alternative visualization of an uncertainty distribution (later)

3.3.2 Visualizing uncertainty

For smooth line graphs, the equivalent of an error bar is a confidence band (as we saw in the errors' subsection)



Claus Wilke

- It shows a range of values the line might pass through at a given confidence level.
- As in the case of error bars, we can draw graded confidence bands that show multiple confidence levels at once.
- We can also show individual fitted draws instead of / or in addition to the confidence bands.

3.3.2 Uncertainty visualization: for a lay audience

Then, two **commonly used approaches** to indicate uncertainty **are error bars and confidence bands**.

For a lay audience, however, **visualization strategies that create a strong intuitive impression of the uncertainty will be preferable**, even if they come at the cost of either reduced visualization accuracy or less data-dense displays.

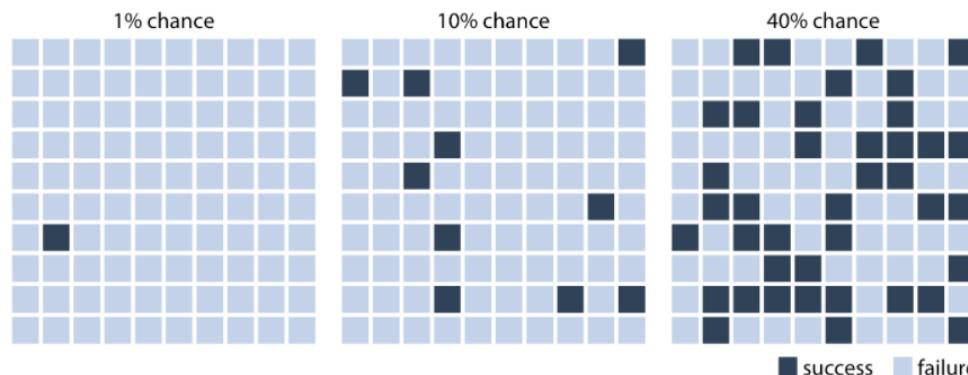
Options here include:

- **frequency framing**, where we explicitly draw different possible scenarios in approximate proportions
- **animations** that cycle through different possible scenarios.

3.3.2 Uncertainty visualization: Frequency framing

- Mathematically, we deal with uncertainty by employing the concept of probability. BUT **visualising a single probability is difficult**.
- For many problems of practical relevance, it is sufficient **to think about relative frequencies**.

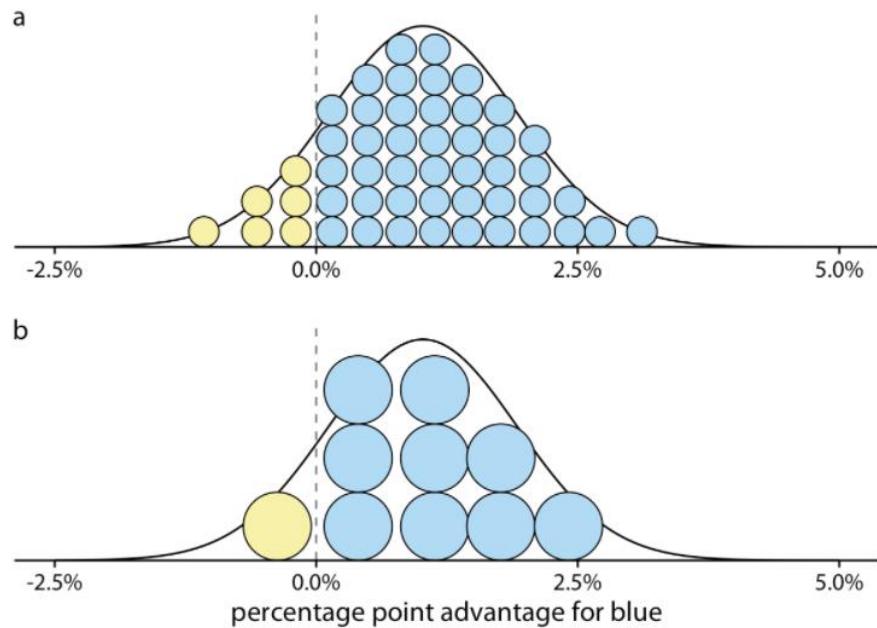
We can make the concept of probability tangible by creating a graph that emphasizes both the frequency aspect and the unpredictability of a random trial, for example by drawing squares of different colours in a random arrangement.



Claus Wilke

3.3.2 Uncertainty visualization: Frequency framing

- What happens if we have more than two discrete outcomes (success or failure)?



Quantile dotplot representations of an election outcome distribution.

The percentage chance in (b) is not accurate.

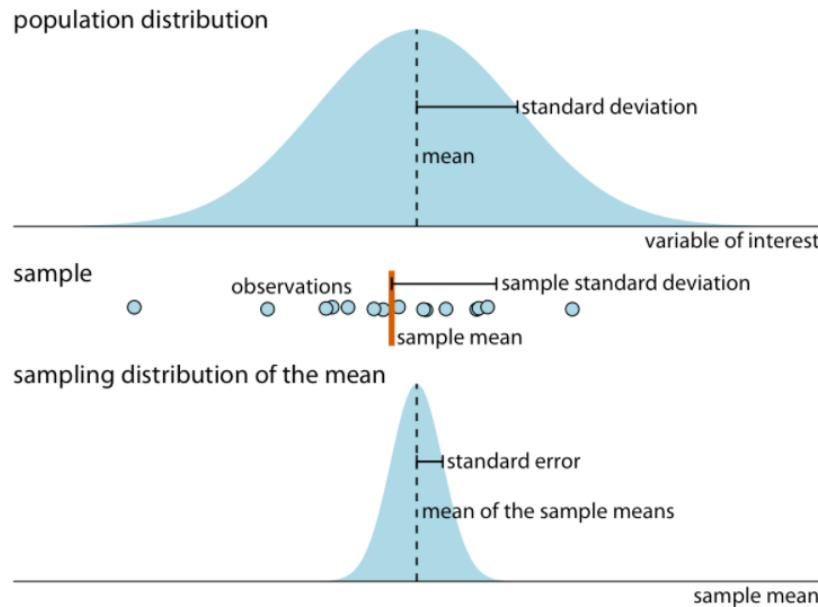
They serve to trade mathematical precision for more accurate human perception of the resulting visualization - communicating to a lay audience

Claus Wilke

As a general principle, **quantile dotplots** should use a small to moderate number of dots. If there are too many dots, then we tend to perceive them as a continuum rather than as individual, discrete units.

3.3.2 Uncertainty visualization: Standard Error

The standard error provides a measure of the uncertainty associated with our parameter estimate.



The variable of interest that we are studying has some true distribution in the population, with a true **population mean and standard deviation**.

Any finite **sample** of that variable will have a **sample mean and standard deviation that differ from the population parameters**.

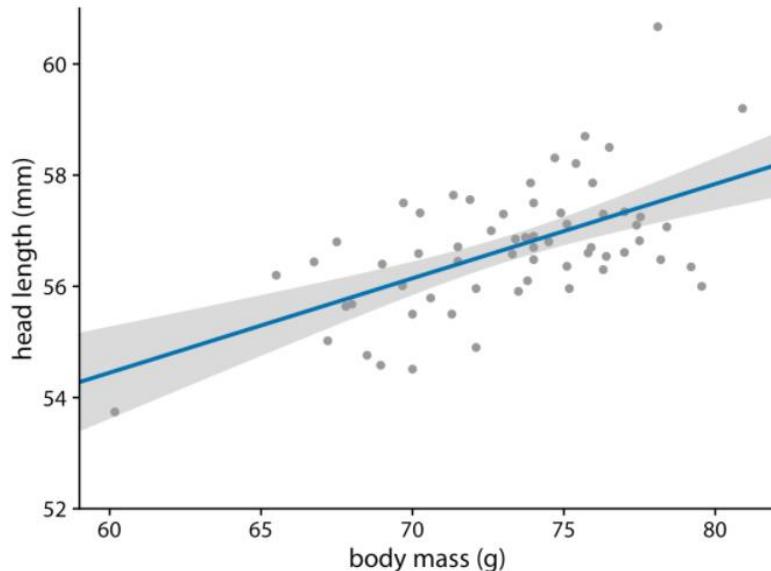
The standard error informs us about how precisely we are estimating the population mean (our parameter estimate)

Claus Wilke

The larger the sample size -> the smaller the standard error and thus the less uncertain the estimate

3.3.3 Dynamic uncertainty visualization: Curve fits

- We can show a trend in a dataset by fitting a straight line or curve to the data
- These trend estimates also have uncertainty, and it is customary to show the uncertainty in a trend line with a confidence band

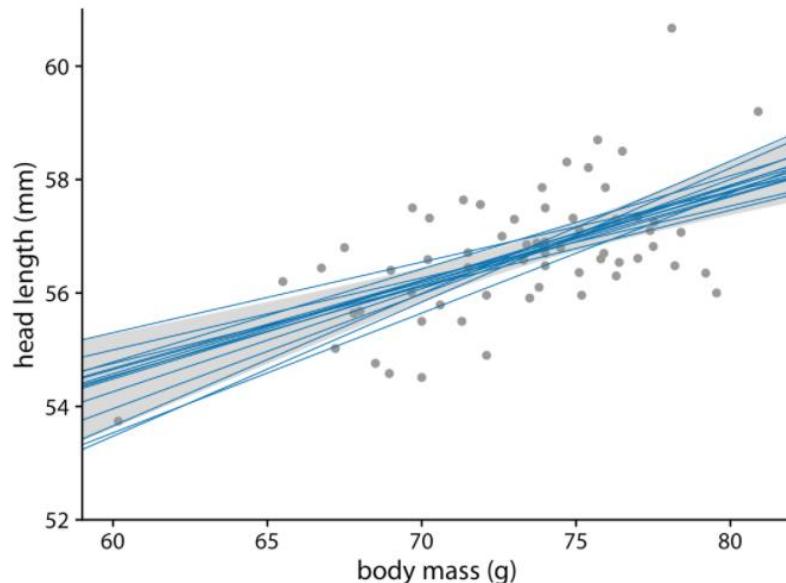


- The straight blue line represents the best linear fit to the data.
- The gray band around the line shows the uncertainty in the linear fit. The gray band represents a 95% confidence level.

Keith Tarvin, Oberlin College

3.3.3 Dynamic uncertainty visualization: Curve fits

- We can show a trend in a dataset by fitting a straight line or curve to the data
- These trend estimates also have uncertainty, and it is customary to **show the uncertainty in a trend line with a confidence band**

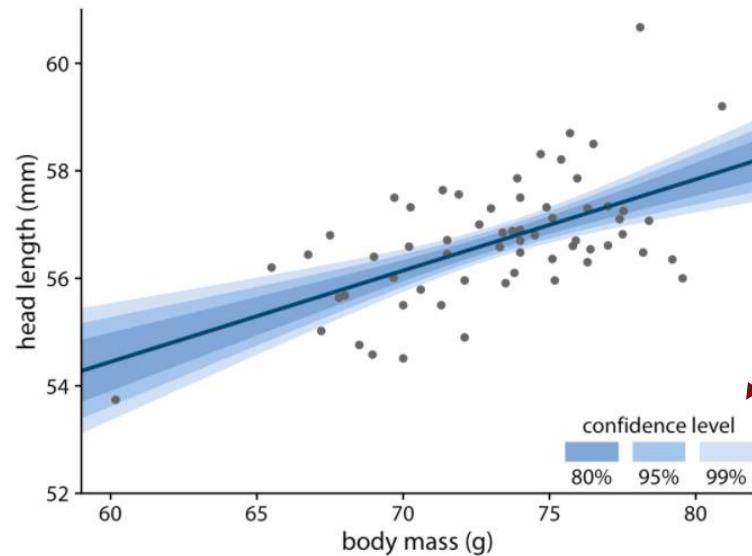


The straight blue lines now represent equally likely alternative fits randomly drawn from the posterior distribution.

Keith Tarvin, Oberlin College

3.3.3 Dynamic uncertainty visualization: Curve fits

- To draw a confidence band, we need to specify a confidence level, and it can be useful to highlight different levels of confidence.
- A graded confidence band enhances the sense of uncertainty in the reader, and it forces the reader to confront the possibility that the data might support different alternative trend lines.



We can draw
graded confidence
bands to highlight
the uncertainty in
the estimate.

Keith Tarvin, Oberlin College

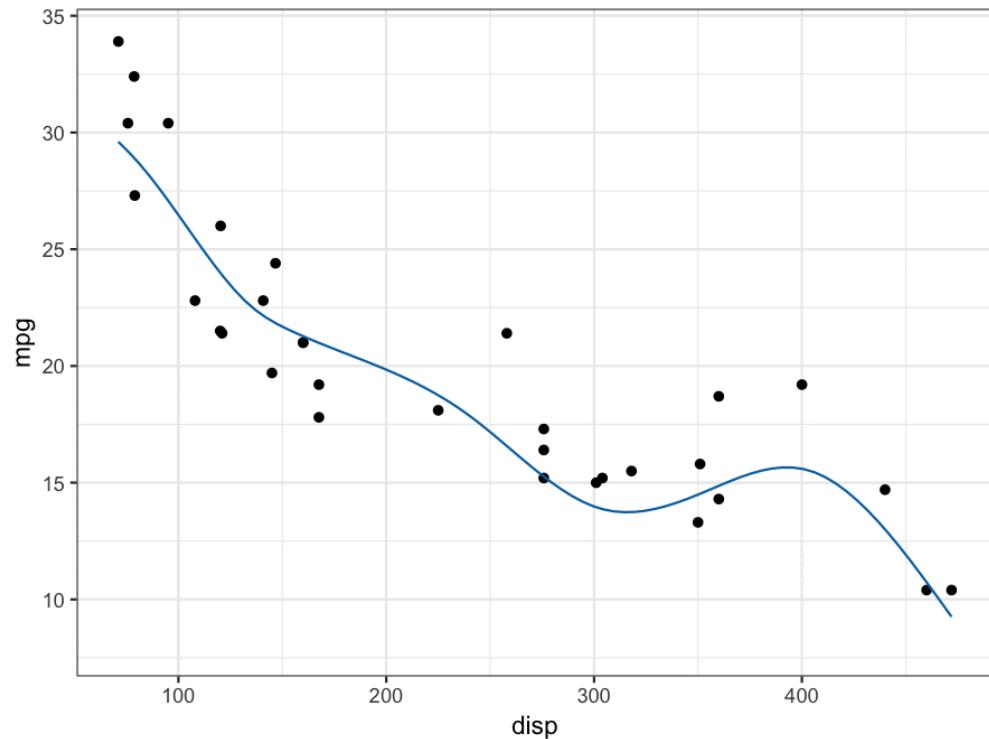
3.3.3 Dynamic uncertainty visualization: HOPs

Hypothetical outcomes plots (HOPs)

- HOPs visualize a set of draws from a distribution -> **each draw is shown as a new plot** in either a small multiples or animated form.
- Better than showing a continuous probability distribution.
- **HOPs require relatively little background knowledge to interpret.**

! Limitation: dynamically presenting draws introduces sampling error

3.3.3 Dynamic uncertainty visualization: HOPs



Hullman J. et al, 2015

Visualizing uncertainty with hypothetical outcomes plots (Claus Wilke) :
<https://www.youtube.com/watch?v=SjYwhku2si0>

3.3.4 Bayesian - Monte Carlo Simulation

When to use it?

- When it is impossible (or impractical) to determine a distribution theoretically (or deterministically)
- In many areas: engineering, finance, management of risks of a project

What involves?

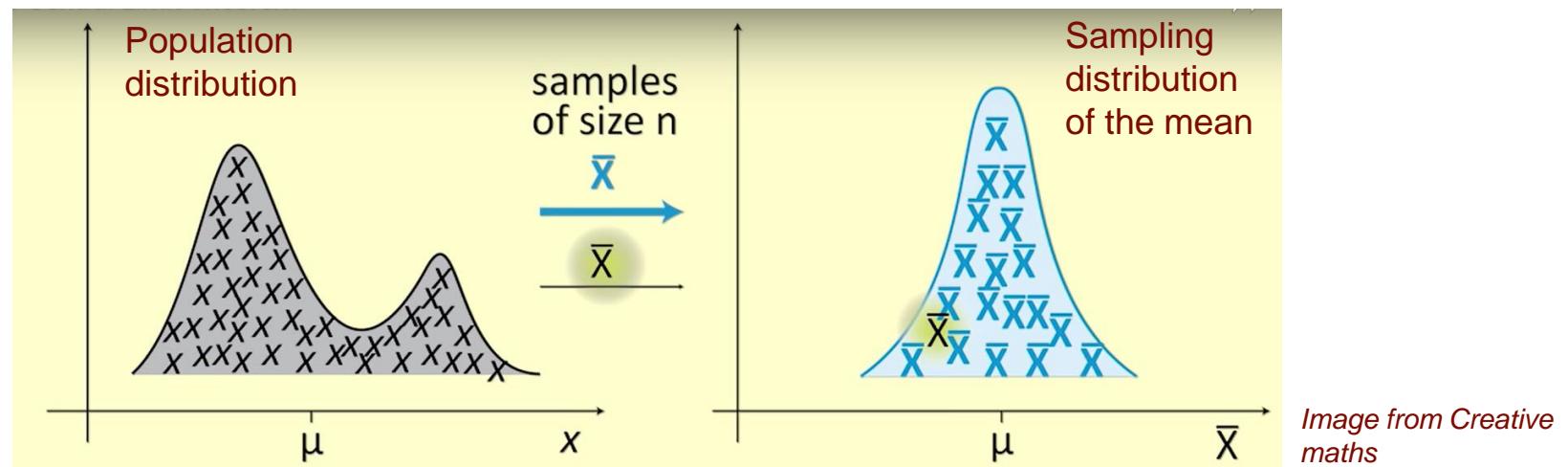
- One or more input variables X (some of which usually follow a probability distribution)
- One or more output variables Y (whose distribution is desired)
- A mathematical model coupling the X's and the Y's

3.3.4 Central limit theorem (CLT)

Idea : Given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

Why is it important?

It implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions -> the Gaussian distribution is used for hypothesis testing using confidence intervals, or to look at the statistical significance of experiment results.

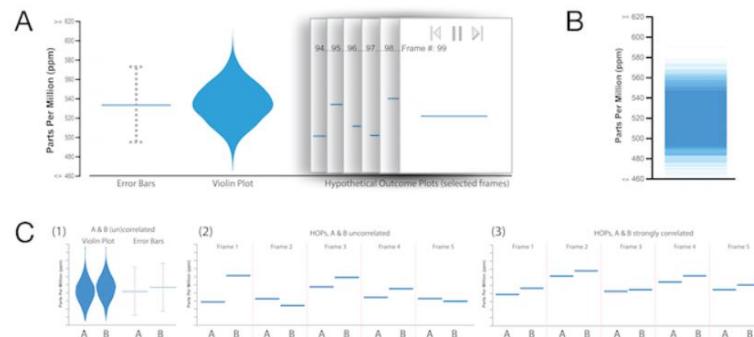


3.3.4 Assumptions behind the CLT

- **Randomization Condition:**
The data must be sampled randomly (or errors)
- **Independence Assumption:**
The sample values must be independent of each other
- **Sample Size:**
When the sample is drawn without replacement (usually the case), the sample size, n , should be no more than 10% of the population.
When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well.

3.3 More examples visualizing uncertainty

- Visualizing uncertainty by Robert Falkowitz
- Uncertainty Quantification and Visualization: Geo-Spatially Registered Terrains and Mobile Targets Suresh Lodha Computer Science, University of California, - ppt download (slideplayer.com)
- Visualizing Uncertainty About the Future:
spiegelhalter_visualizing.pdf (berkeley.edu)
- Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering (manuscript)



(A) Representations of uncertainty compared in our study, (B) HOPs limiting case, (C) HOPs can express properties of a joint distribution.

3.4 Transformation & Data massage. Contents:

- 1. Introduction of Visualizing errors & uncertainty**
- 2. Error**
- 3. Uncertainty**
- 4. Transformation and data massage**

- 1. Introduction**
- 2. Best practices**
- 3. Potential activity transforming your data**
- 4. Tidy and transform data in R (basics)**

3.4.1 Introduction: Transformation & Data massage.

What if the database is not formatted in the way you expect? Or the data is completely unstructured?

It is rare that you get the data in exactly the right form you need

- Before data is loaded to visualize it, **it must be transformed** to meet any format and structural requirements
- ***Data massaging*, also known as data *cleansing* or *scrubbing*, is a process that eliminates unnecessary information from data or cleans a dataset to make it useable.**

3.4.1 Introduction: Transformation & Data massage.

What does it mean for your data to be “tidy”?

The word “tidy” in data science **using R** means that your data follows a **standardized format**:

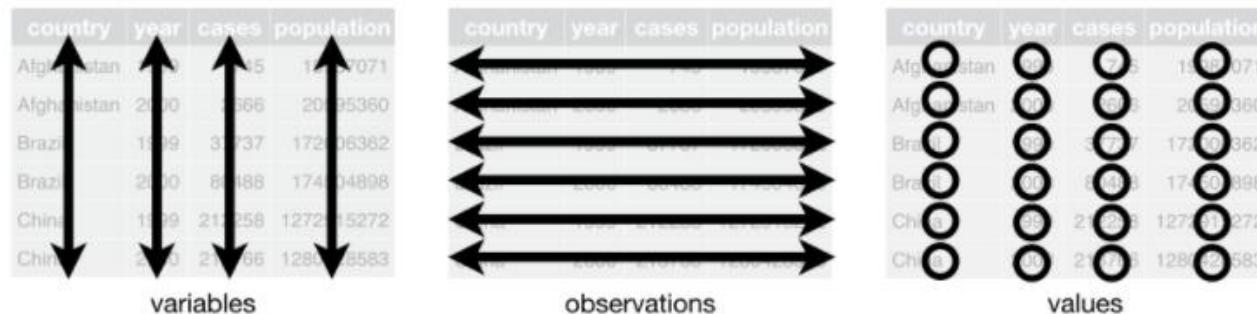
- A **dataset** is a collection of values, usually either numbers (if quantitative) or strings AKA text data (if qualitative/categorical). **Values are organised in two ways. Every value belongs to a variable and an observation.**
- **A variable contains all values that measure the same underlying attribute across units** (examples: weight, temperature, duration).
- **An observation contains all values measured on the same unit across attributes** (examples: person, day, village).
- **“Tidy” data is a standard way of mapping the meaning of a dataset to its structure.** A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

3.4.1 Introduction: Transformation & Data massage.

What does it mean for your data to be “tidy”?

In “tidy data”:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table



Tidy data graphic from R for Data Science

3.4.1 Introduction: Transformation & Data massage.

What does it mean for your data to be “tidy”?

Fecha	Nombre	Mate	Ingles
1-11-2015	Hernandez, Rodrigo	90	60

mes	año	primer	apellido	materia	puntos
11	2015	Rodrigo	Hernandez	mate	90
11	2015	Rodrigo	Hernandez	ingles	60



Edgar Ruiz
2018

country	year	cases	population
Afghanistan	2010	15	1507071
Afghanistan	2010	3666	2095380
Brazil	1999	37737	17206362
Brazil	2010	86488	17404898
China	1999	21258	1272515272
China	2010	21066	128022583

variables

country	year	cases	population
Afghanistan	2010	15	1507071
Afghanistan	2010	3666	2095380
Brazil	1999	37737	17206362
Brazil	2010	86488	17404898
China	1999	21258	1272515272
China	2010	21066	128022583

observations

country	year	cases	population
Afghanistan	2010	15	1507071
Afghanistan	2010	3666	2095380
Brazil	1999	37737	17206362
Brazil	2010	86488	17404898
China	1999	21258	1272515272
China	2010	21066	128022583

values

Tidy data graphic from R for Data Science

3.4.1 Introduction: Transformation & Data massage.

- Databases come in different shapes and sizes and each must be treated as unique.
- A few data massaging techniques are required to adapt the data to the algorithms we are working with.
- Common tasks include stripping unwanted characters and whitespace, converting number and date values into desired formats, and organising data into a meaningful structure.
- ***Massaging the data is usually the "transform" step.*** In most cases, one or more transformations are required.

3.4.1 Introduction: Transformation & Data massage.

Things we do to massage the data include:

- **Change formats** from the standard source system emissions to the target system requirements, e.g. change date format from m/d/y to d/m/y, or sort the data.
- **Replace missing values** with defaults, e.g. "0" when a quantity is not given.
- **Filter out data** that is not desired in the destination system. Sub setting or removing observations based on some condition.
- **Check validity of data and fixing records:** ignore or report on rows that would cause an error, remove unwanted characters and duplicates.
- **Splitting and resampling**
- **Normalise/standardizing data** to remove variations that should be the same, e.g. replace upper case with lower case, replace "01" with "1".

3.4.1 Introduction: Transformation & Data massage.

Reducing Items and Attributes

④ Filter

→ Items

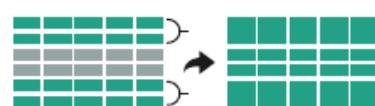


→ Attributes

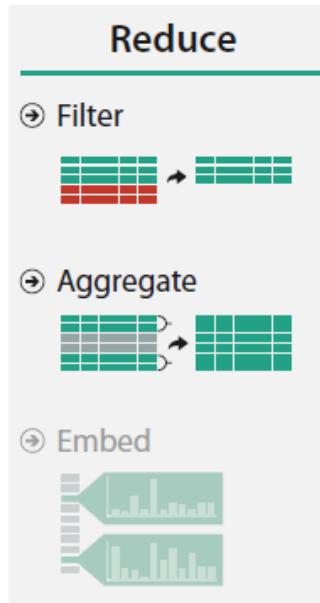
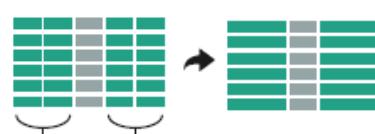


④ Aggregate

→ Items



→ Attributes

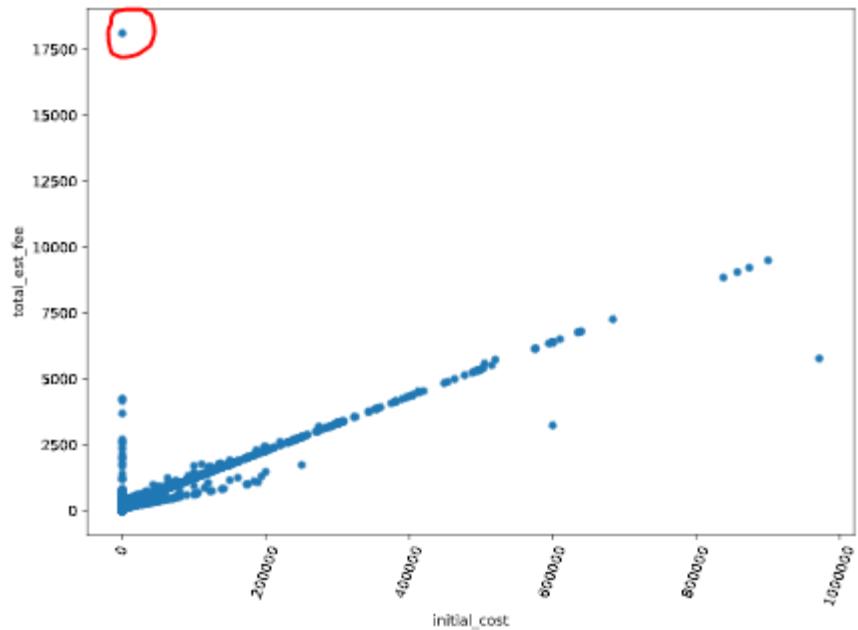


Design choices for reducing (or increasing) the amount of data items and attributes to show.

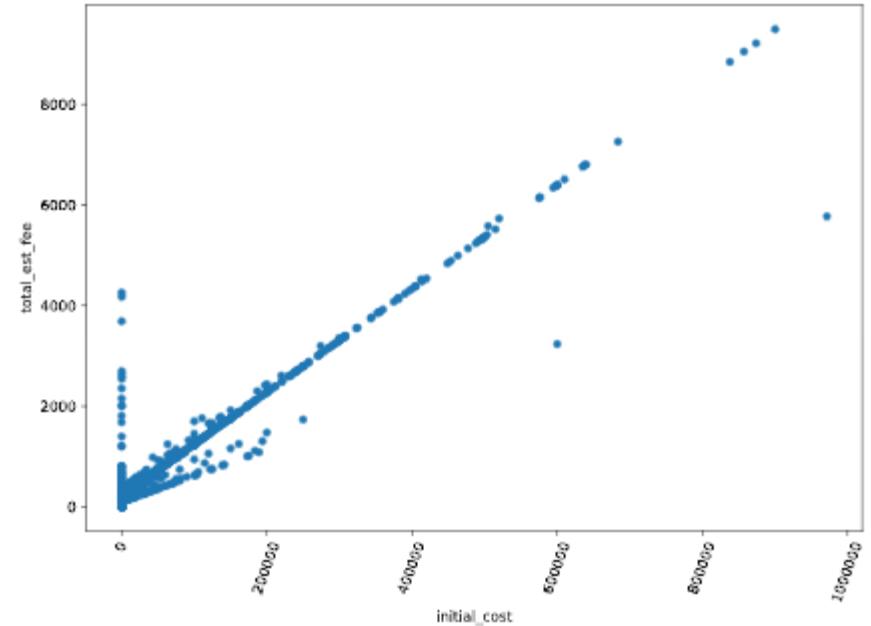
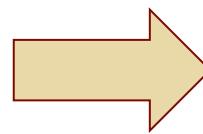
Tamara Munzner



3.4.1 Introduction: Transformation & Data massage.



Removing
the outlier



	Name	Height	Roll
0	A	5.2	55
1	A	5.2	55
2	C	5.6	15
3	D	5.5	80
4	E	5.3	12
5	E	5.3	12
6	G	5.6	47
7	H	5.5	104

Removing
duplicate
rows

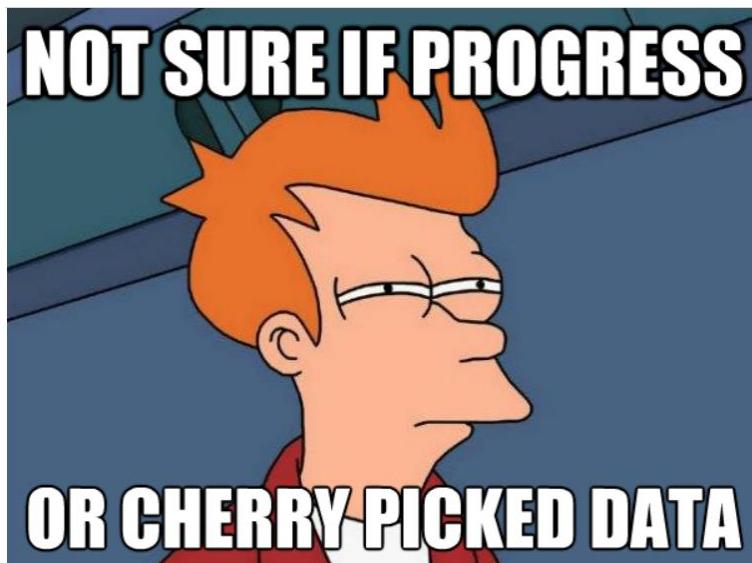


	Name	Height	Roll
0	A	5.2	55
2	C	5.6	15
3	D	5.5	80
4	E	5.3	12
6	G	5.6	47
7	H	5.5	104

3.4.2 Best practices

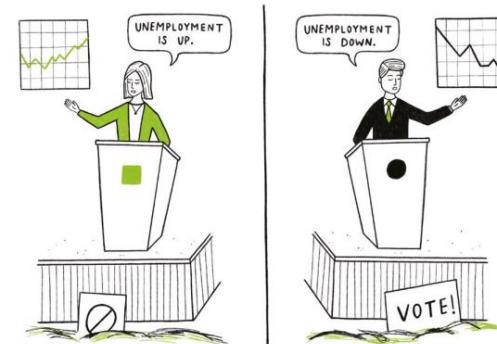
Unfortunately, there is such a thing as a bad message. **The term “data massaging” is also associated with the practice of “cherry-picking”,** selectively excluding or altering data based on what people want (or don’t want) it to reflect.

Cherry-picking changes the message that the final visualisation communicates to the audience.



!!! Be careful

CHERRY PICKING



The practice of selecting results that fit your claim, and excluding those that don't. The worst and most harmful example of being dishonest with data.

3.4.2 Best practices

The questions you need to ask of your data are:

- Does it represent genuine observations about a given phenomenon or is it influenced by the limitations of a collection method?
- Does your data reflect the entirety of a particular phenomenon, a recognised sample, or maybe even an obstructed view caused by hidden limitations in the availability of data about that phenomenon?

Once you complete your examination of your data you will have a good idea about what actions may be needed to transform your data.

In accordance with the desire for trustworthy design, any modifications or enhancements you apply to your data need to be noted and potentially explained to your audience.

3.4.3 Potential activity transforming your data

'Before you can plot or graph anything, you have to find the data, understand it, evaluate it, clean it, and perhaps restructure it.' (Marcia Gray, graphic designer)

Three different types of potential activity involved in transforming your data:

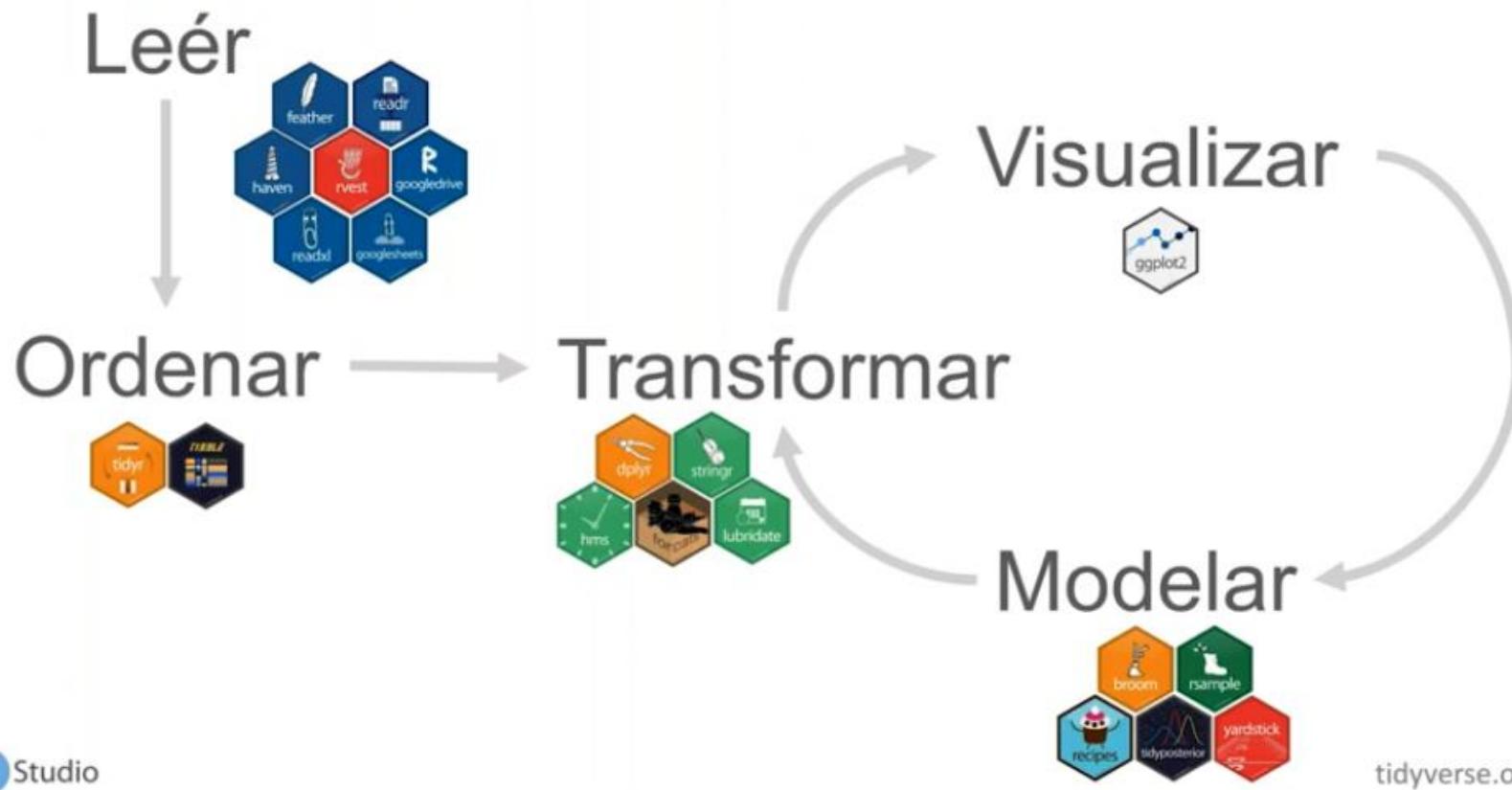
- **Cleaning:** resolve any data condition issues
- **Creating:** consider developing new calculations and value conversions
- **Consolidating:** think about introducing further data to expand or append to what you already have

3.4.3 Potential activity transforming your data

- **Cleaning:** There is no single approach for how best to conduct data cleaning- Issues may be resolved through manual intervention, sorting, filtering, isolating, modifying any problem values/characters.
- **Creating:** Expand your data to form new calculations and derive new groupings or any other mathematical treatments. This may include:
 - Creating percentage calculations based on existing quantities.
 - Using ‘start date’ and ‘end date’ values to calculate the duration in days.
 - Using logic-based formulae to create new categorical values out of quantities
 - To derive reasonable categorical or quantitative values from the original form.
- **Consolidating:** you may seek to source and introduce additional data to **expand** (more variables) or **append** (more items) your data further in order to enhance its analytical potential

3.4.4 Tidy and transform data with R (basics)

Paquetes del “tidyverse”



3.4.4 Data Transforming with R

Data Transformation with dplyr :: CHEAT SHEET

dplyr functions work with pipes and expect tidy data. In tidy data:



Each variable is in its own column



Each observation, or case, is in its own row



x %>% f(y) becomes f(x, y)

Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).



summary function

summarise(data, ...)
Compute stats of summaries.
summarise(mtcars, avg = mean(mpg))



count(x, ..., wt = NULL, sort = FALSE)
Count number of rows in each group defined by the variables in ... Also tally().
count(iris, Species)

VARIATIONS

summarise_all() - Apply funs to every column.
summarise_at() - Apply funs to specific columns.
summarise_if() - Apply funs to all cols of one type.

Group Cases

Use **group_by()** to create a "grouped" copy of a table. dplyr functions will manipulate each "group" separately and then combine the results.



mtcars %>%
group_by(cyl) %>%
summarise(avg = mean(mpg))

group_by(data, ..., add = FALSE)
Returns copy of table grouped by ...
g_iris <- group_by(iris, Species)

ungroup(x, ...)
Returns ungrouped copy of table.
ungroup(g_iris)



RStudio® is a trademark of RStudio, Inc. • CC BY SA RStudio • info@rstudio.com • 844-448-1212 • rstudio.com • Learn more with browseVignettes(package = c("dplyr", "tibble")) • dplyr 0.7.0 • tibble 1.2.0 • Updated: 2017-03

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.



filter(.data, ...) Extract rows that meet logical criteria. filter(iris, Sepal.Length > 7)



distinct(.data, ..., keep_all = FALSE) Remove rows with duplicate values. distinct(iris, Species)



sample_frac(tbl, size = 1, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select fraction of rows. sample_frac(iris, 0.5, replace = TRUE)



sample_n(tbl, size, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select size rows. sample_n(iris, 10, replace = TRUE)



slice(.data, ..., n) Select rows by position. slice(iris, 10:15)



top_n(x, n, wt) Select and order top n entries (by group if grouped data). top_n(iris, 5, Sepal.Width)



Logical and boolean operators to use with filter()

< <= is.na() %in% | xor()
> >= !is.na() ! &

See ?base::logic and ?Comparison for help.

ARRANGE CASES



arrange(.data, ...) Order rows by values of a column or columns (low to high), use with desc() to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))



ADD CASES



add_row(.data, ..., before = NULL, after = NULL)
Add one or more rows to a table.
add_row(faithful, eruptions = 1, waiting = 1)



Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.



pull(.data, var = -1) Extract column values as a vector. Choose by name or index.
pull(iris, Sepal.Length)



select(.data, ...) Extract columns as a table. Also select_if().
select(iris, Sepal.Length, Species)



Use these helpers with select(), e.g. select(iris, starts_with("Sepal"))

contains(match) num_range(prefix, range) : e.g. mpg:cyl
ends_with(match) one_of(...) -, e.g. -Species
matches(match) starts_with(match)

MAKE NEW VARIABLES

These apply **vectorized functions** to columns. Vectorized funs take vectors as input and return vectors of the same length as output (see back).



vectorized function



mutate(.data, ...) Compute new column(s).
mutate(mtcars, gpm = 1/mpg)



transmute(.data, ...) Compute new column(s), drop others.
transmute(mtcars, gpm = 1/mpg)



mutate_all(.tbl, .funs, ...) Apply funs to every column. Use with funs(). Also mutate_if().
mutate_all(faithful, funs(log10, log2))
mutate_if(iris, is.numeric, funs(log(.)))



mutate_at(.tbl, .cols, .funs, ...) Apply funs to specific columns. Use with funs(), vars() and the helper functions for select().
mutate_at(iris, vars(-Species), funs(log(.)))



add_column(.data, ..., before = NULL, after = NULL) Add new column(s). Also add_count(), add_tally().
add_column(mtcars, new = 1:32)



rename(.data, ...) Rename columns.
rename(iris, Length = Sepal.Length)



We will see some examples during seminars



3.4.4 Tidy Data with R

Tibbles - an enhanced data frame

The **tibble** package provides a new S3 class for storing tabular data, the **tibble**. Tibbles inherit the data frame class, but improve three behaviors:

- **Subsetting** - [always returns a new tibble, [[and \$ always return a vector.
- **No partial matching** - You must use full column names when subsetting
- **Display** - When you print a tibble, R provides a concise view of the data that fits on one screen

A large table to display → tibble display → data frame display

A tibble: 234 x 6
manufacture... model: chr
... with 234 more variables: cyl <dbl>, trans <chr>,...

- Control the default appearance with options:
`options(tibble.print_max = n,
tibble.print_min = m, tibble.width = Inf)`
- View full data set with `View()` or `glimpse()`
- Revert to data frame with `as.data.frame()`

CONSTRUCT A TIBBLE IN TWO WAYS

tibble(...)
Construct by columns.
`tibble(x=1:3, y=c("a", "b", "c"))`

Both make this tibble

tribble(...)
Construct by rows.
`tribble(~x, ~y, ~z, 1, "a", "a", 2, "b", "b", 3, "c", "c")`

as_tibble(x, ...) Convert data frame to tibble.
`enframe(x, name = "name", value = "value")` Convert named vector to a tibble
`is_tibble(x)` Test whether x is a tibble.



Tidy Data with tidyverse

Tidy data is a way to organize tabular data. It provides a consistent data structure across packages.

A table is tidy if:

Each variable is in its own column

&

Each observation, or case, is in its own row

Tidy data:
Makes variables easy to access as vectors
Preserves cases during vectorized operations

Reshape Data - change the layout of values in a table

Use `gather()` and `spread()` to reorganize the values of a table into a new layout.

`gather(data, key, value, ..., na.rm = FALSE, convert = FALSE, factor_key = FALSE)`

`gather()` moves column names into a key column, gathering the column values into a single value column.

table4a

country	1990	2000
A	0.7K	2K
B	37K	37K
C	212K	213K

→

country	year	cases
A	1999	0.7K
A	1999	pop
C	1999	212K

key value

table2

country	year	type	count
A	1999	cases	0.7K
A	1999	pop	19M
C	1999	212K	2K
A	2000	cases	2K
B	1999	cases	37K
B	2000	cases	80K
C	2000	cases	213K

key value

`gather(table4a, `1999`, `2000`, key = "year", value = "cases")`

`gather(x, year, value, ..., na.rm = TRUE)`

`spread(table2, type, count)`

`fill(data, ..., direction = c("down", "up"))`
Fill in NAs in ... columns with most recent non-NA values.

`replace_na(data, replace = list(...))`
Replace NA's by column.

x1	x2	x1	x2
A	NA	A	1
B	NA	B	1
C	NA	C	1
D	3	D	3
E	NA	E	3

`drop_na(x, x2)`

`fill(x, x2)`

`replace_na(x, list(x2 = 2))`

`Handle Missing Values`

`drop_na(data, ...)`

Drop rows containing NAs in ... columns.

x1	x2	x1	x2
A	NA	A	1
B	NA	B	1
C	NA	C	1
D	3	D	3
E	NA	E	3

`drop_na(x, x2)`

`fill(x, x2)`

`replace_na(x, list(x2 = 2))`

`complete(data, ..., fill = list())`

Adds to the data missing combinations of the values of the variables listed in ...

`complete(mtcars, cyl, gear, carb)`

x1	x2	x1	x2
A	1	A	1
B	2	B	2
C	3	C	3
D	4	D	4
E	5	E	5

`expand(data, ...)`

Create new tibble with all possible combinations of the values of the variables listed in ...

`expand(mtcars, cyl, gear, carb)`

Split Cells

Use these functions to split or combine cells into individual, isolated values.



`separate(data, col, into, sep = "[^[:alnum:]]+", remove = TRUE, convert = FALSE, extra = "warn", fill = "warn", ...)`

Separate each cell in a column to make several columns.

country	year	rate
A	1999	0.7K/19M
A	2000	2K/20M
B	1999	37K/172M
B	2000	80K/174M
C	1999	212K/1T
C	2000	213K/1T

`separate(table3, rate, sep = "/", into = c("cases", "pop"))`

`separate_rows(data, ..., sep = "[^[:alnum:]]+", convert = FALSE)`

Separate each cell in a column to make several rows.

country	year	rate
A	1999	0.7K/19M
A	2000	2K/20M
B	1999	37K/172M
B	2000	80K/174M
C	1999	212K/1T
C	2000	213K/1T

`separate_rows(table3, rate, sep = "/")`

`unite(data, col, ..., sep = "_", remove = TRUE)`

Collapse cells across several columns to make a single column.

country	century	year
Afghan	19	99
Afghan	20	00
Brazil	19	99
Brazil	20	00
China	19	99
China	20	00

`unite(table5, century, year, col = "year", sep = "")`

We will see some functions during seminars too

Thanks for your attention!

Judit Chamorro Servent

Departament de Matemàtiques

judit.chamorro@uab.cat



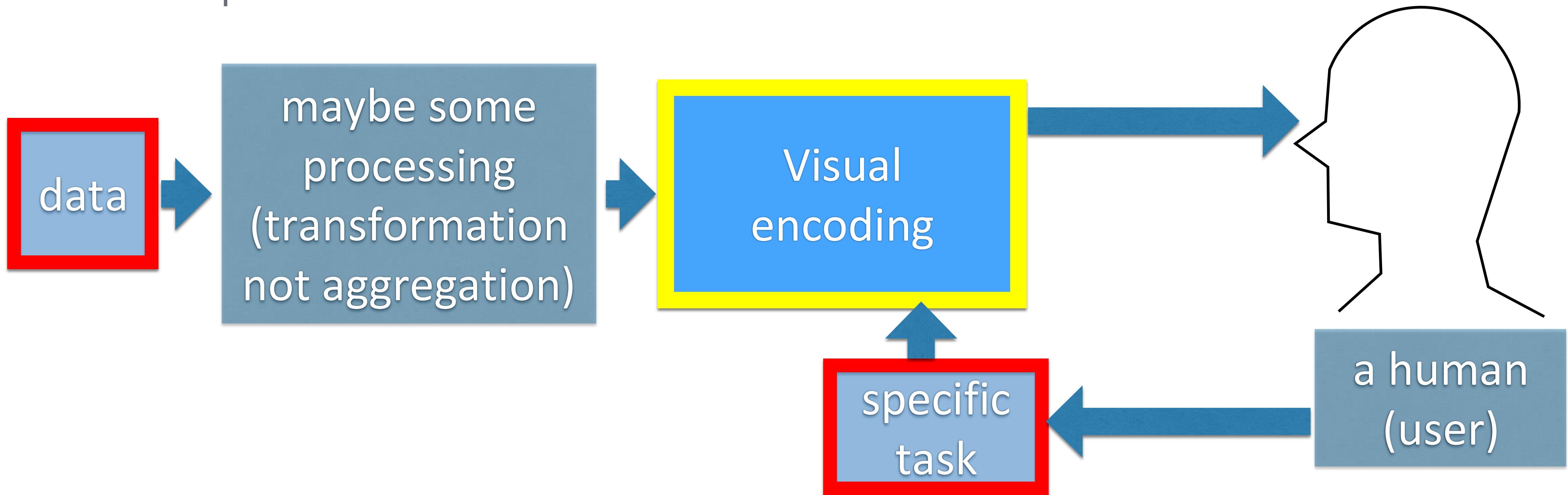
3. Codificación Visual

Sol Bucalo
sol.bucalo@uab.cat

Guillermo Marin
guillermo.marin@uab.cat

Data Visualisation

- Datos. El proceso empieza con uno o más datasets. Conocemos el tipo y las características de sus atributos.
- Tareas. Definición de las tareas que podemos resolver, caracterizadas como acción + objetivo



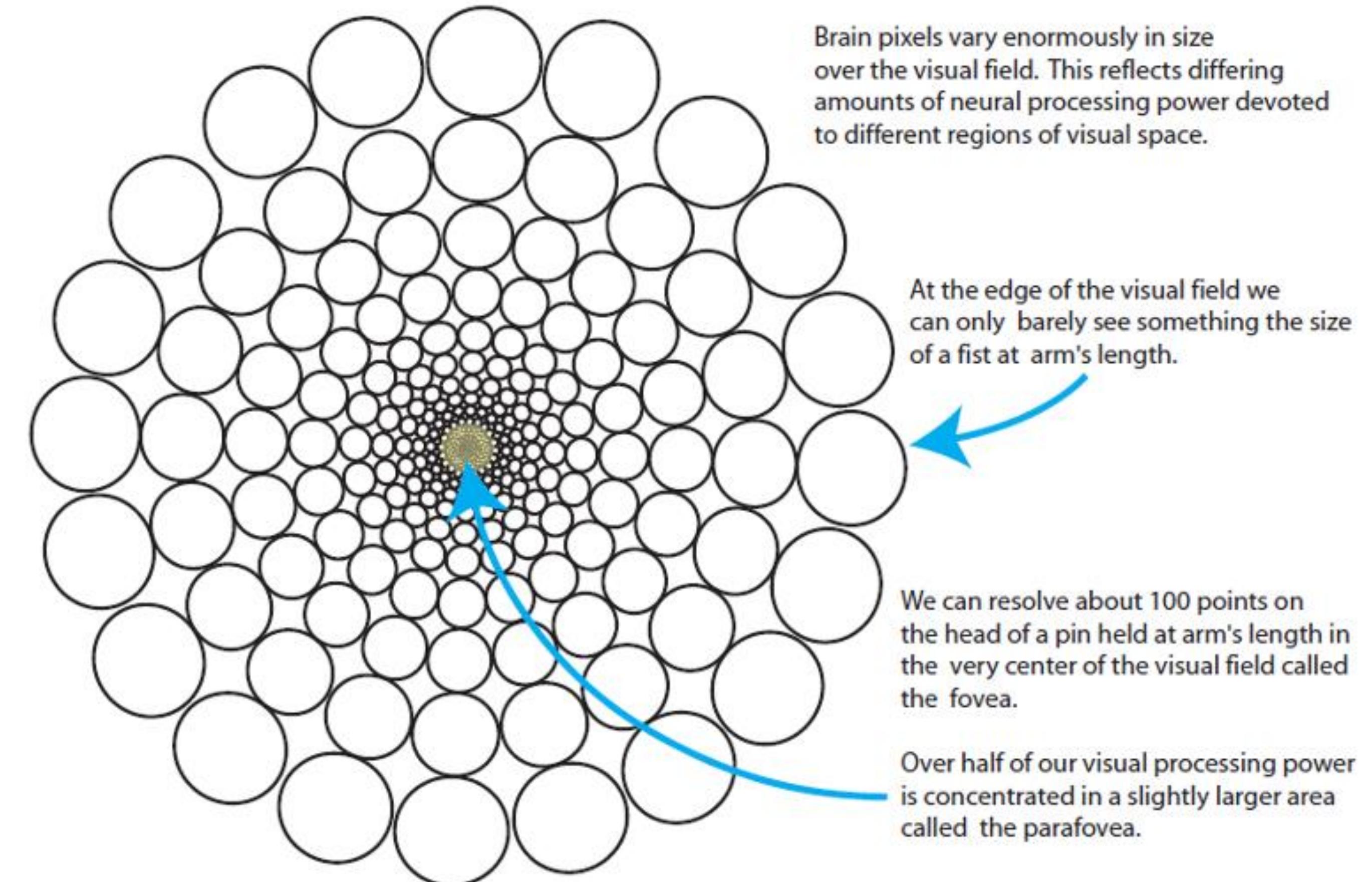
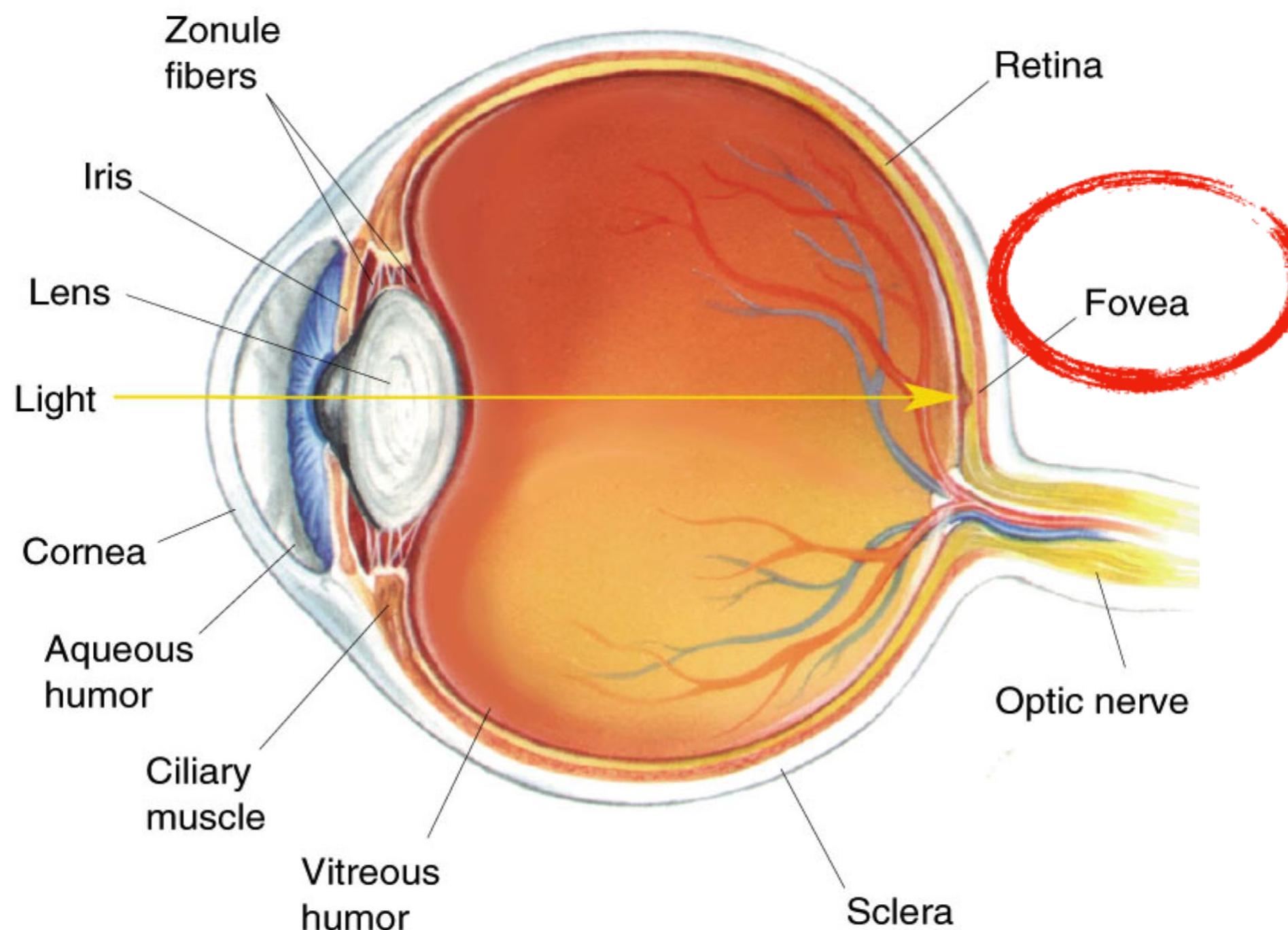


A group of people in Victorian-era clothing are gathered in a parlor. In the center, a man in a dark suit lies dead on the floor. A woman in a blue dress stands over him, looking shocked. To her left, a man in a green jacket and hat is crouching near a large red rose bush. Another man in a dark uniform and hat stands behind the woman. In the background, a man in a top hat and a woman in a white blouse and apron stand near a piano. A painting hangs on the wall, and deer antlers are mounted on the wall above the piano. The scene is set on a patterned rug.

WHODUNNIT?

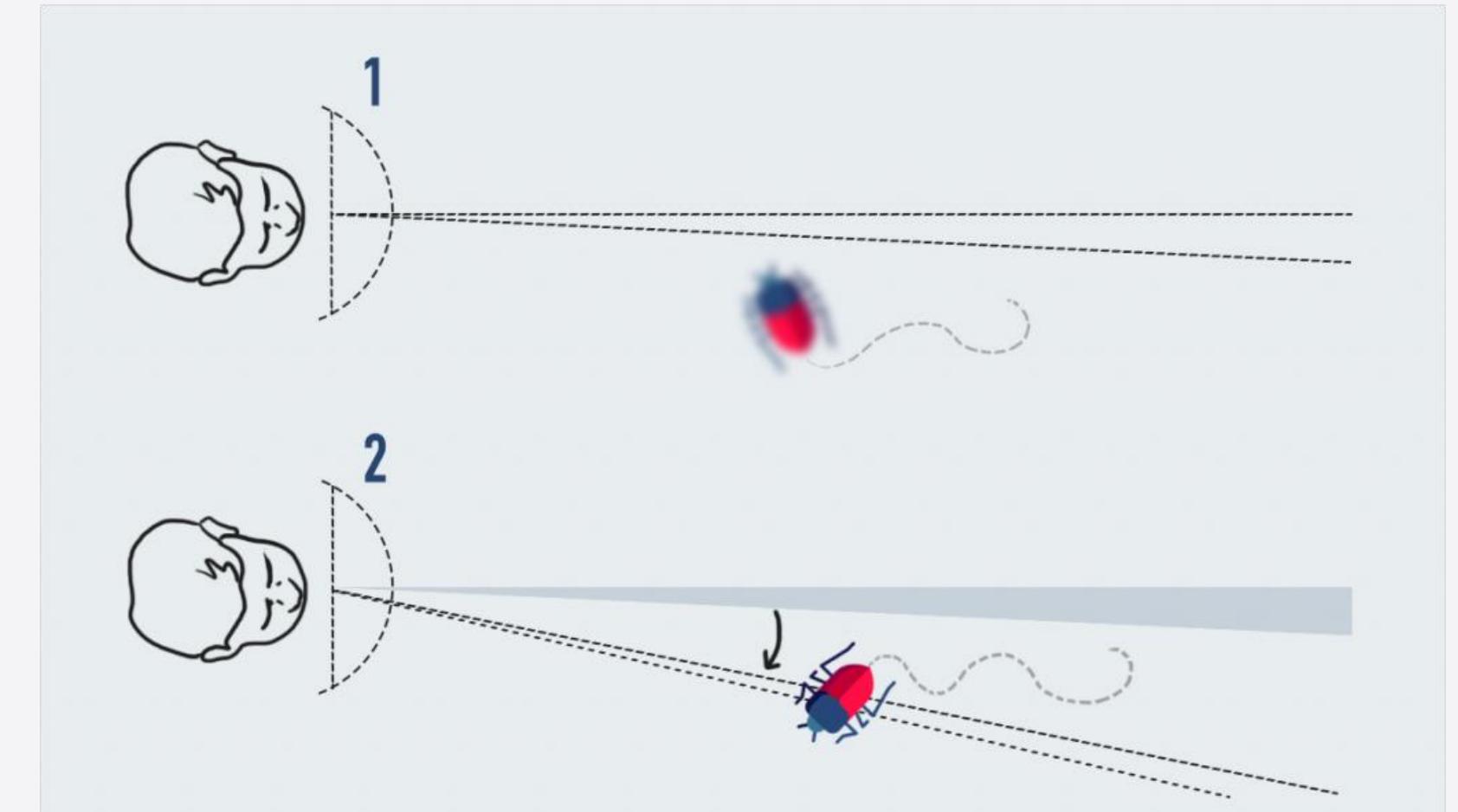
4.1 Percepción Visual

We see what we need: limited peripheral vision

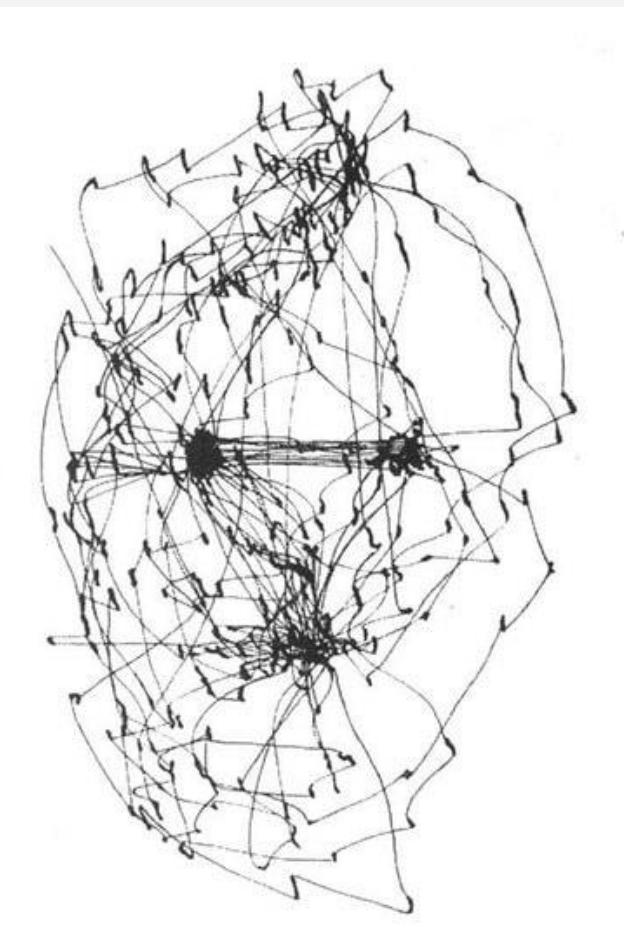




A. Cairo



Nayomi Chibana



Eye movements and vision. Alfred Yarbus

From: yarbus.eu

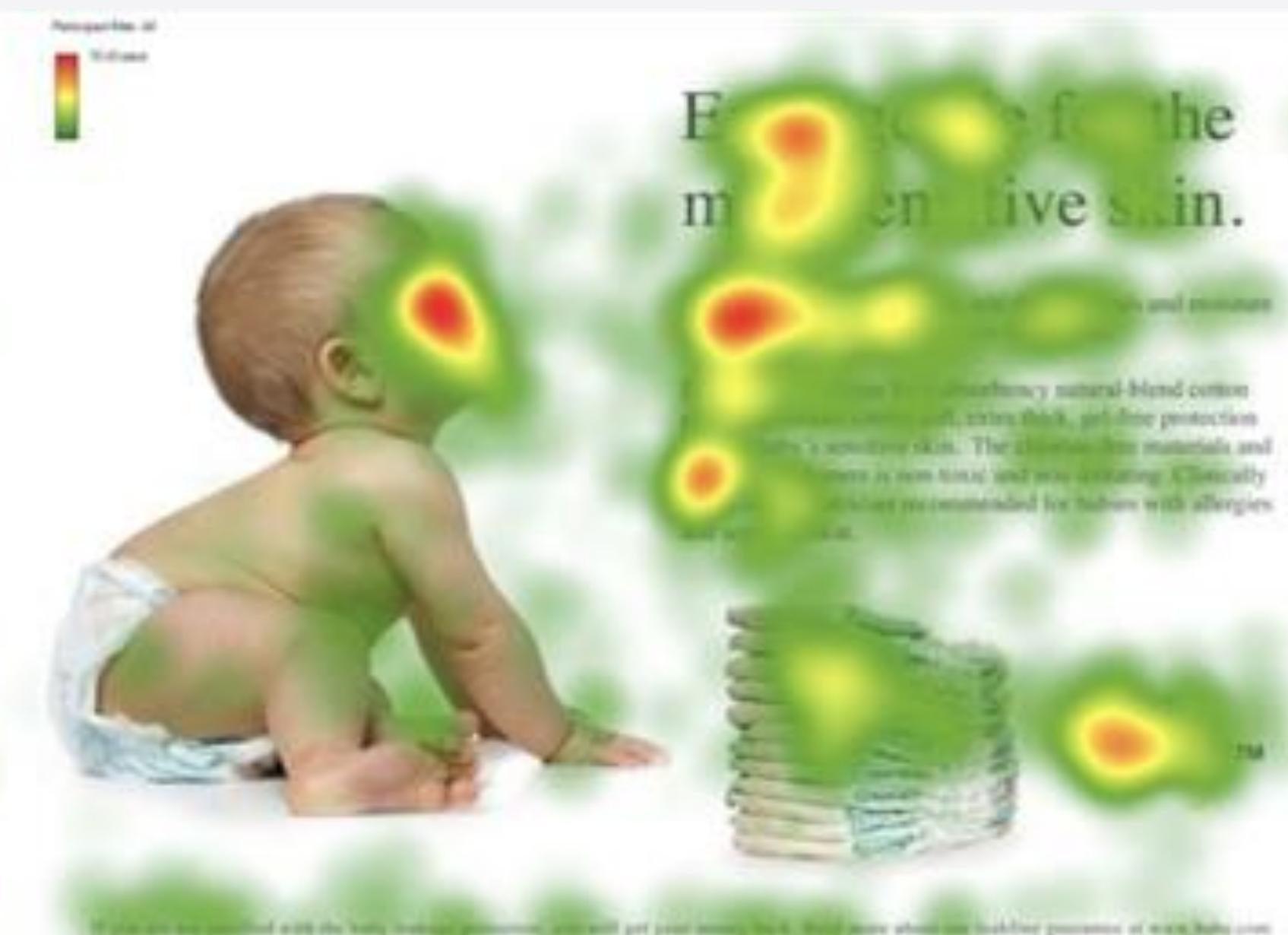
- Los movimientos sacádicos y fijaciones son inconscientes pero no aleatorios.
- El ojo se ve atraido por ciertos atributos y se fija en ellos.
- Movimiento, parches de colores vivos, formas extrañas

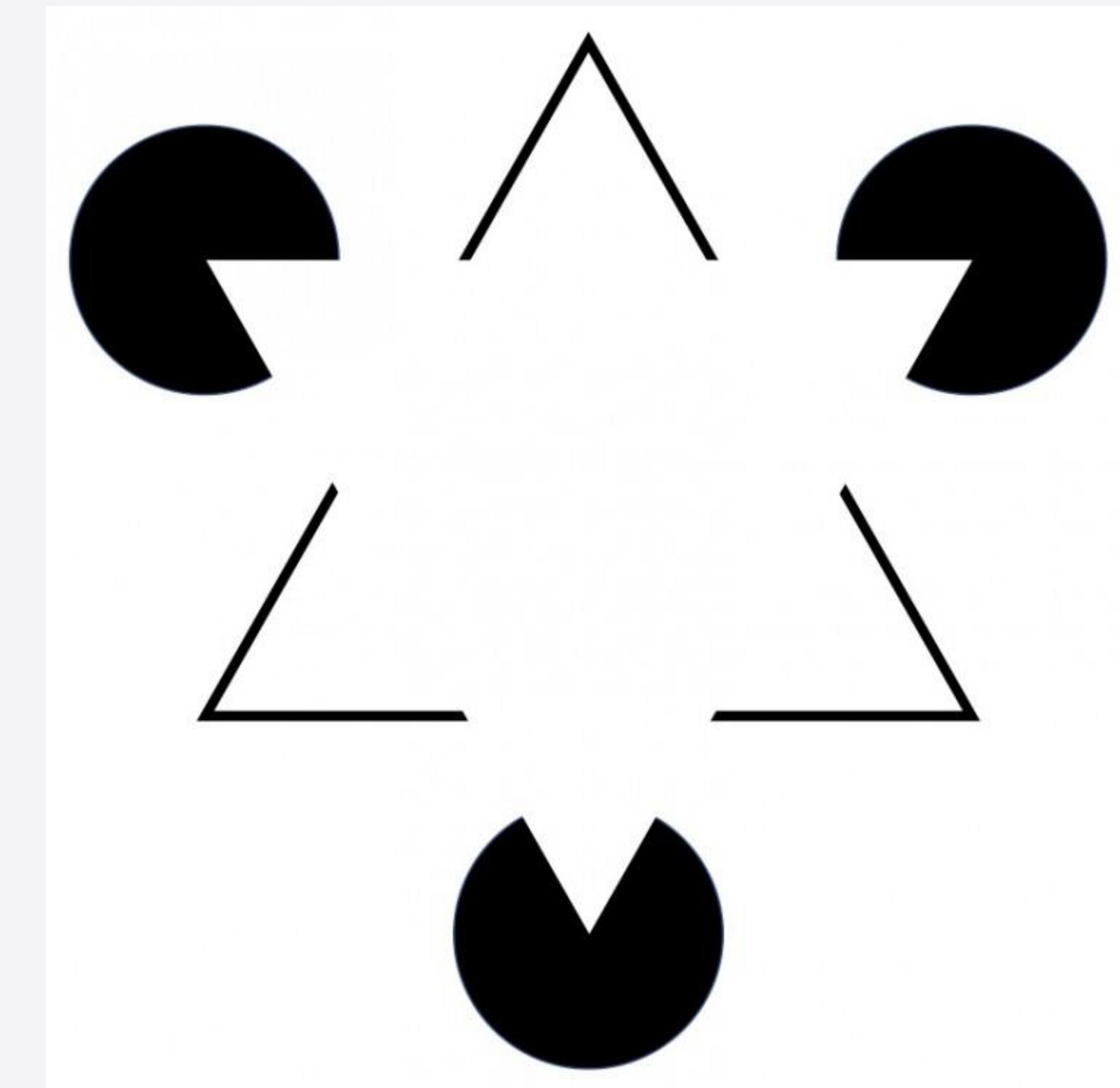
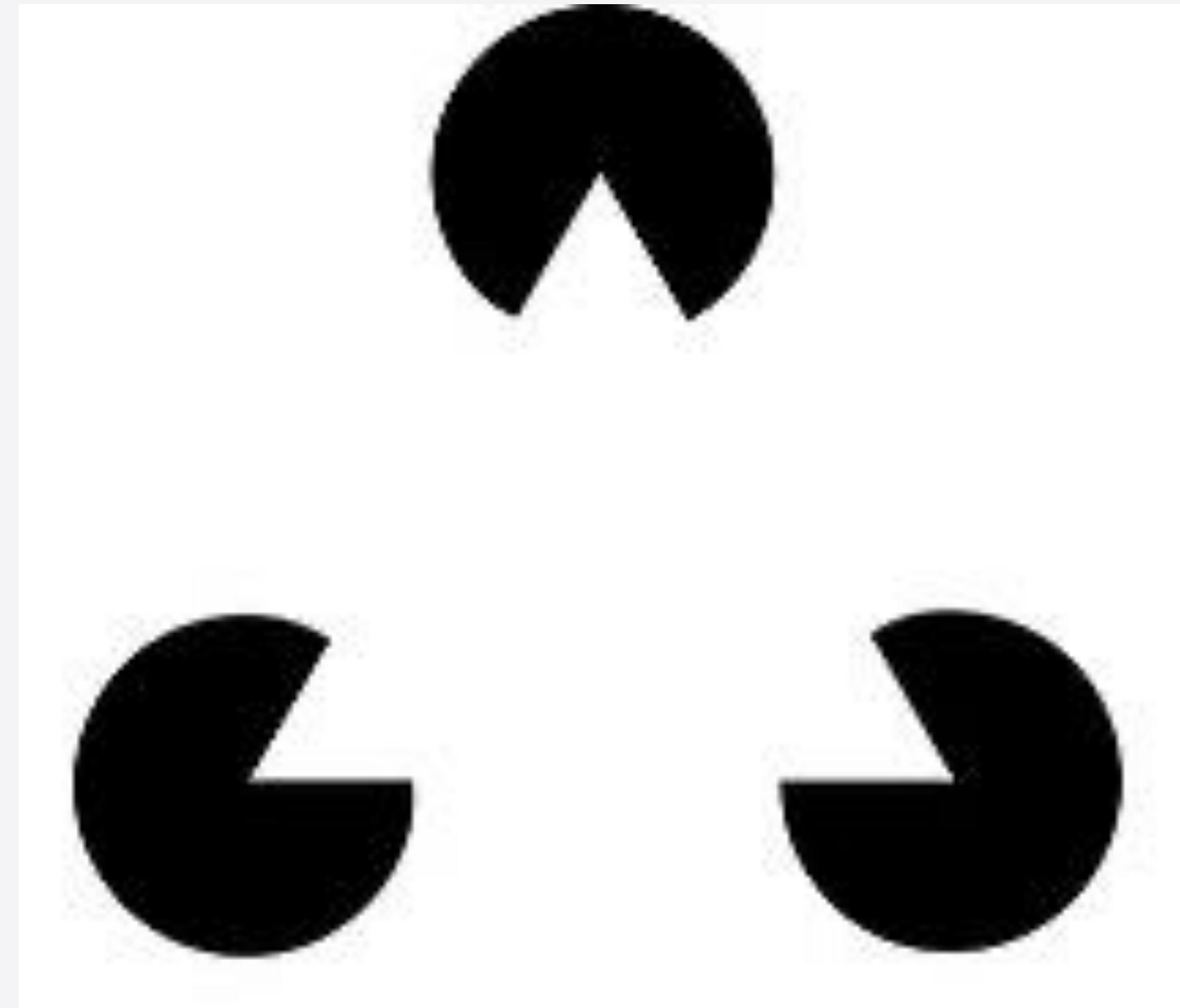


first here

then
down
▼

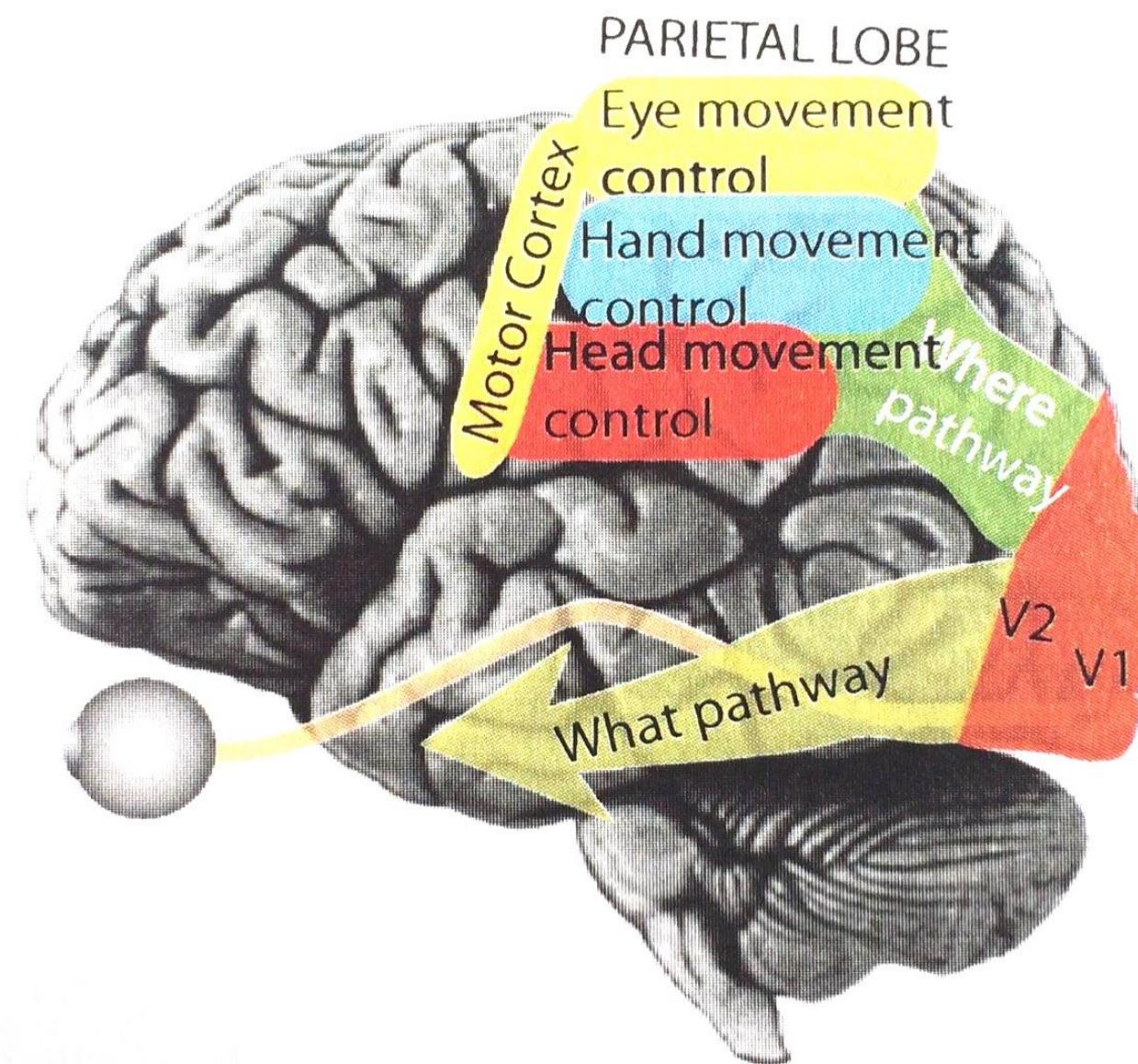
end here



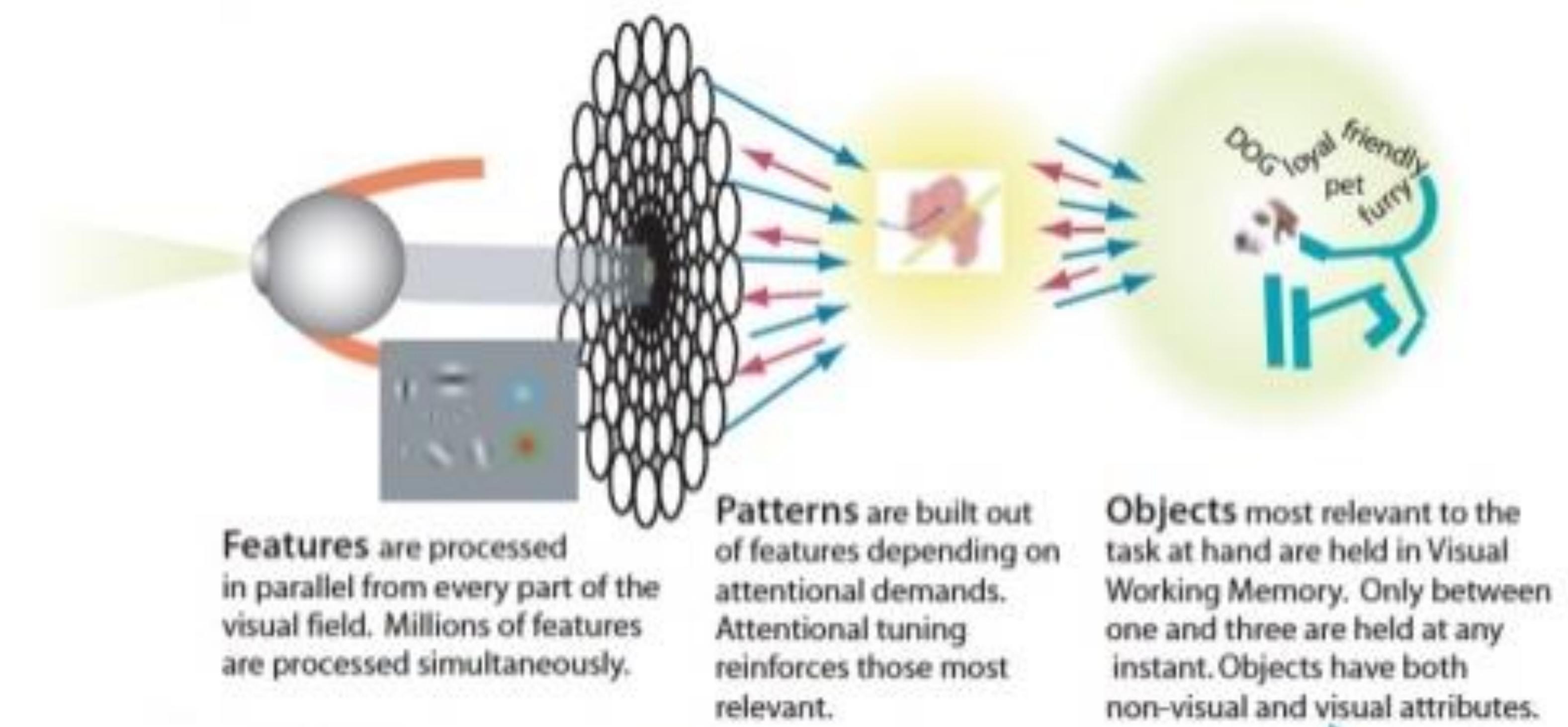


Percepción Visual

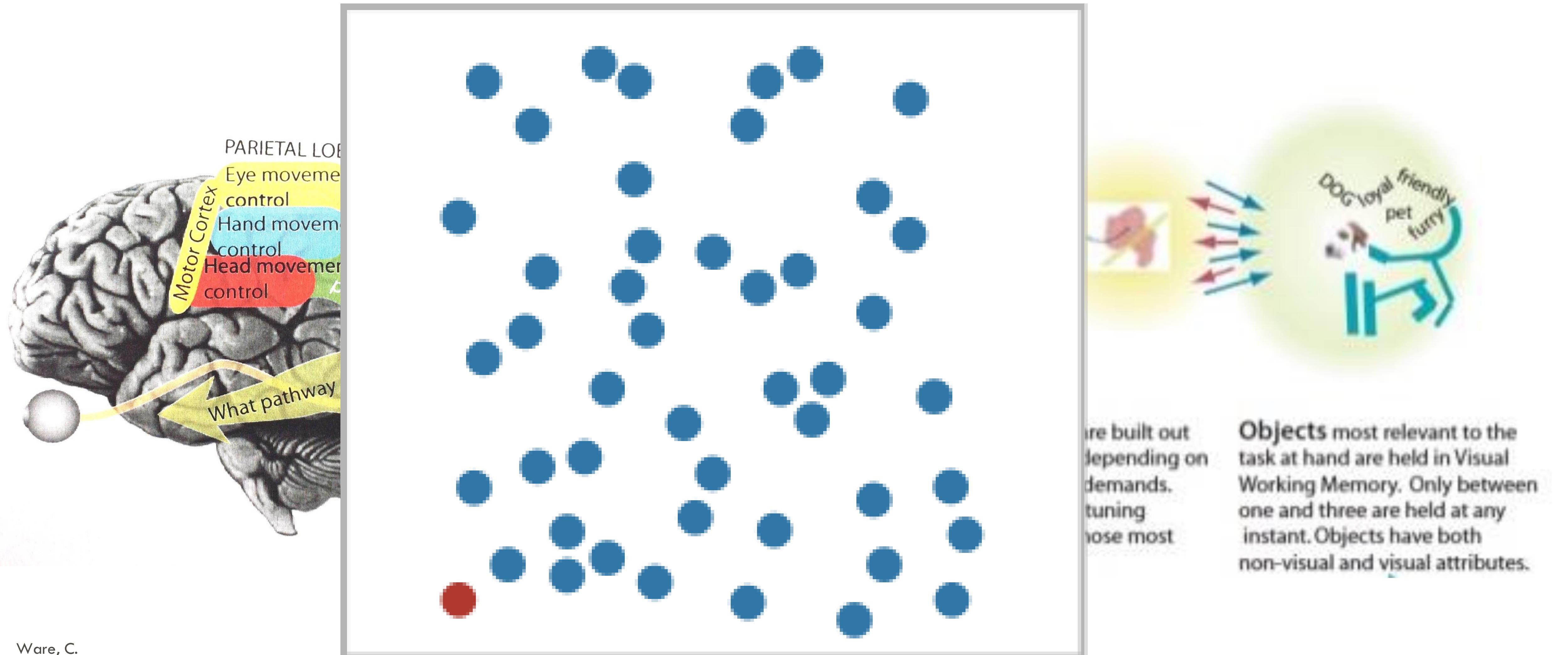
- El procesado de la imagen se da en fases.
- La percepción visual esta distribuida. Partes del procesado ocurren antes que otras a lo largo de la cadena:
 - La retina convierte los patrones de luz en señales eléctricas, se procesan características en paralelo
 - La mente empieza a diferenciar aspectos básicos de los objetos: color, formas básicas.
 - Solo más adelante se da un análisis más profundo de qué está viendo y procede a asignarle un contexto.



Ware, C.



Preattentive attributes



El contraste es importante

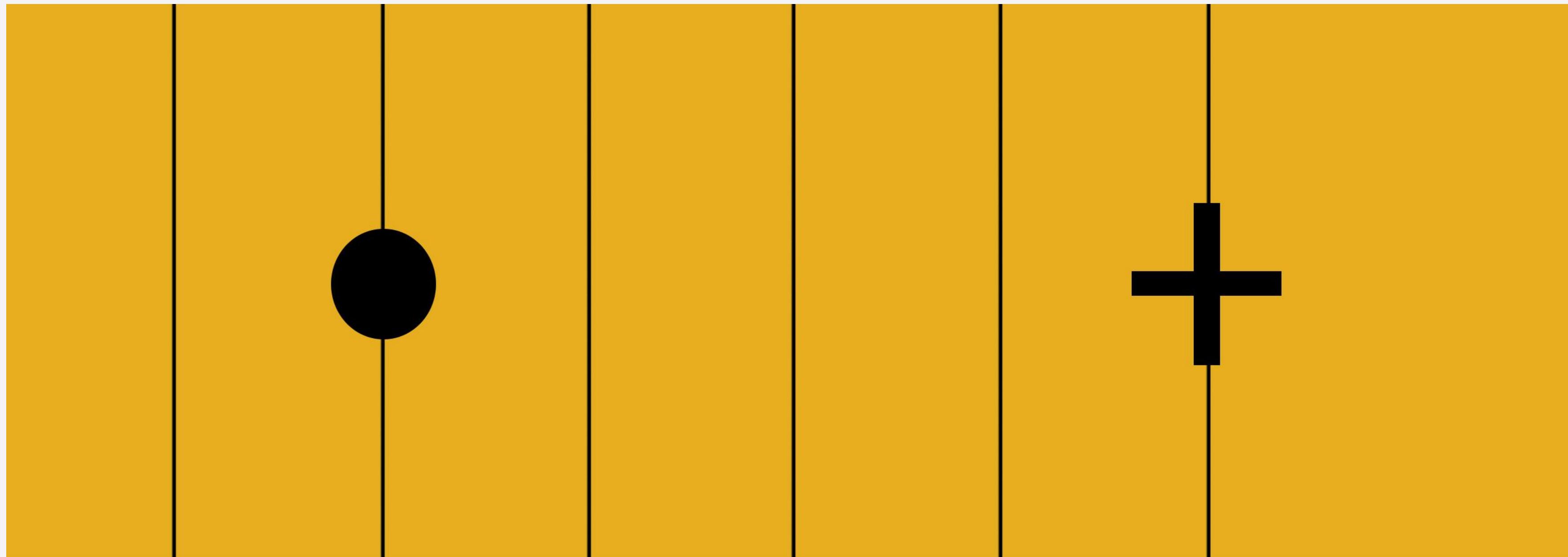
Fácil



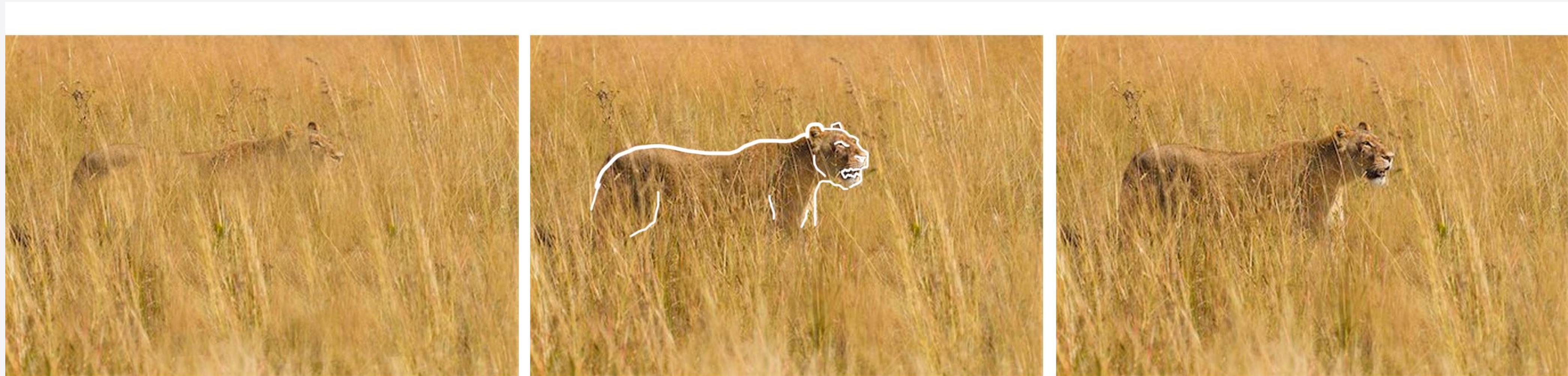
Difícil







Colocarse a unos 20 cm del ordenador
Cerrar el ojo derecho, mirar la cruz con el izquierdo
y acercarse lentamente.



Inspirado en The Functional Art

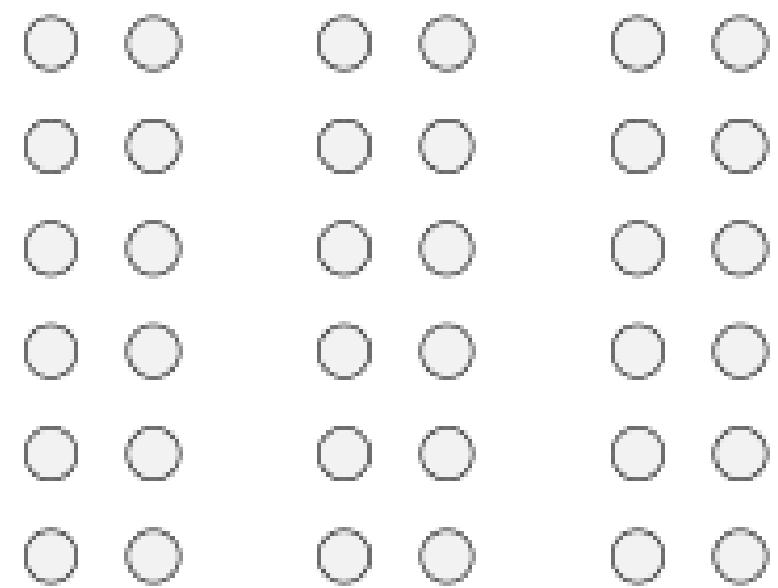
Background/Foreground



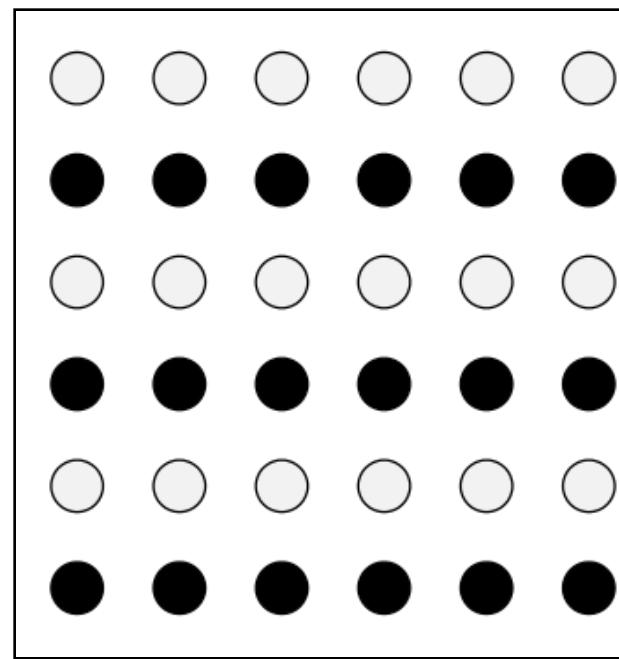
- La mente busca patrones de distinción entre objetos y fondo
- Detección de bordes se basa en diferencias de luz, colores, y definición de los bordes
- Contraste de luz permite la detección más rápida, contraste en tono también rápida, pero algo menos. Sin contraste requiere identificar a la criatura por la forma.

Gestalt principles

Proximity



Similarity



Continuity

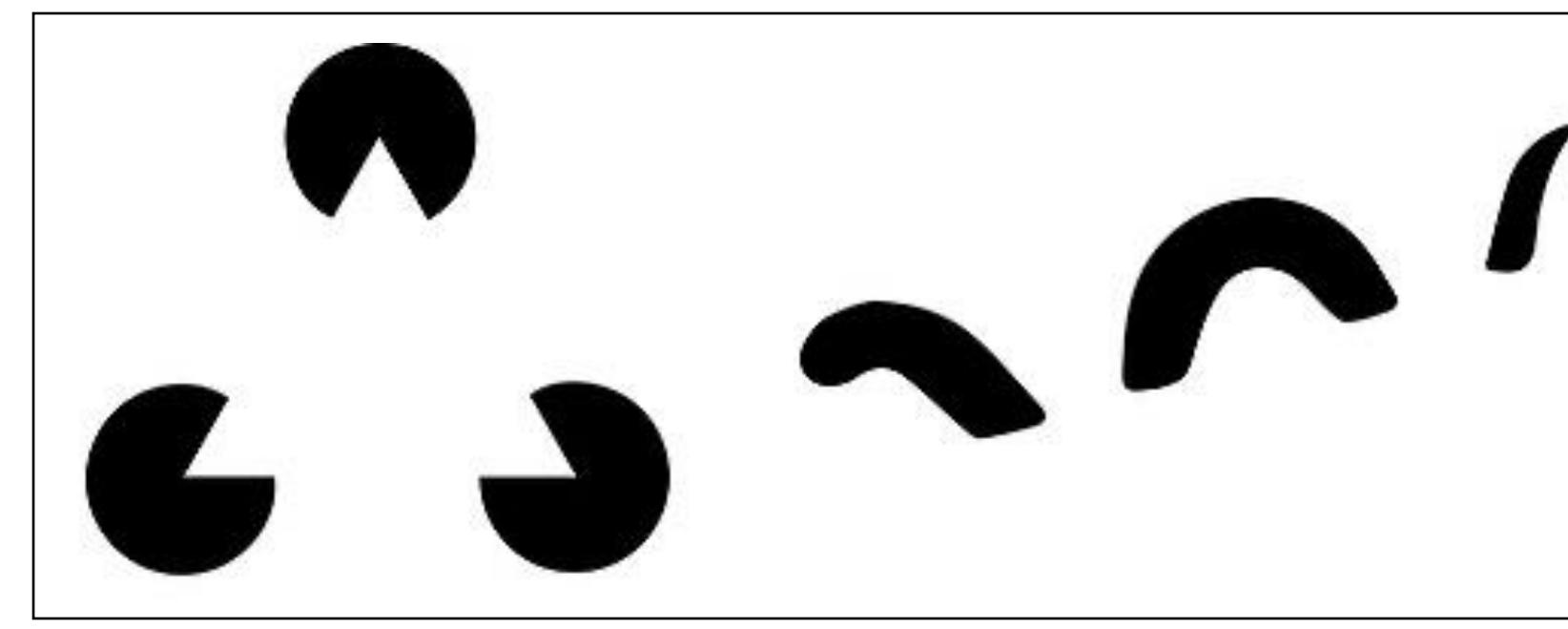


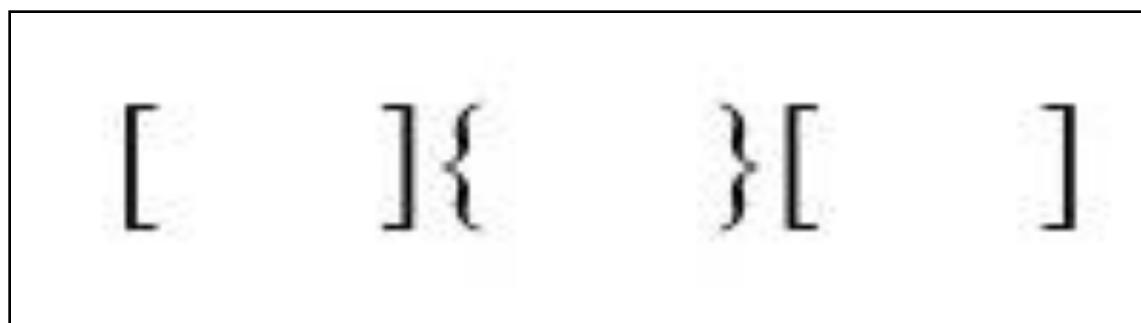
Figure and ground



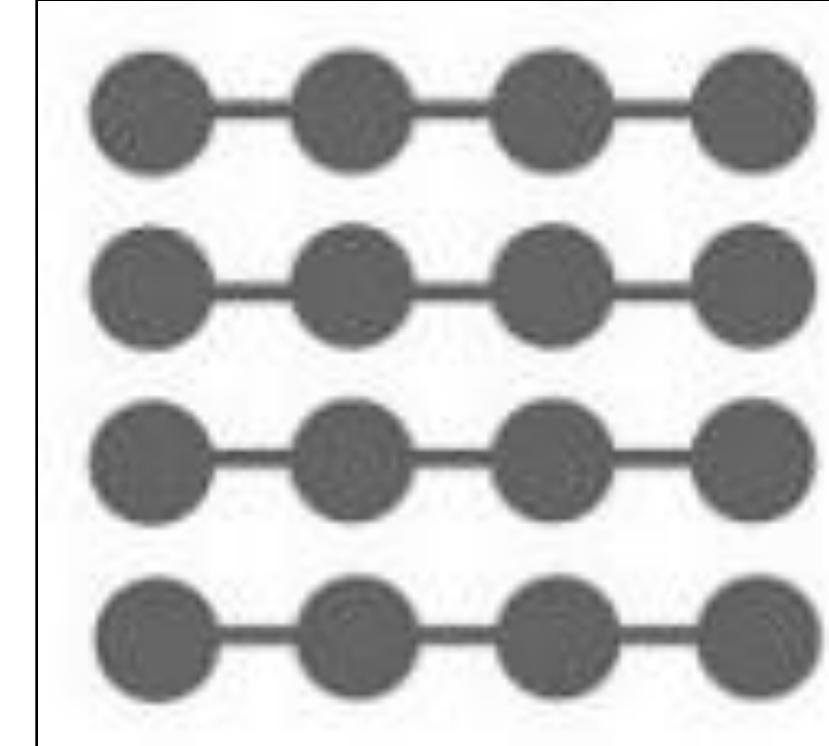
Closure



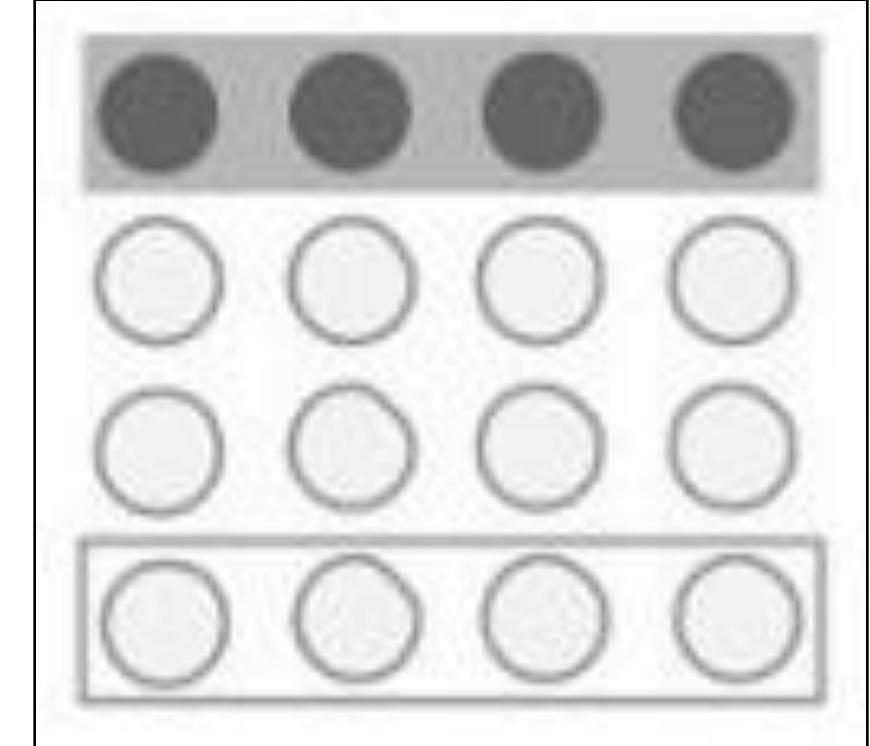
Symmetry



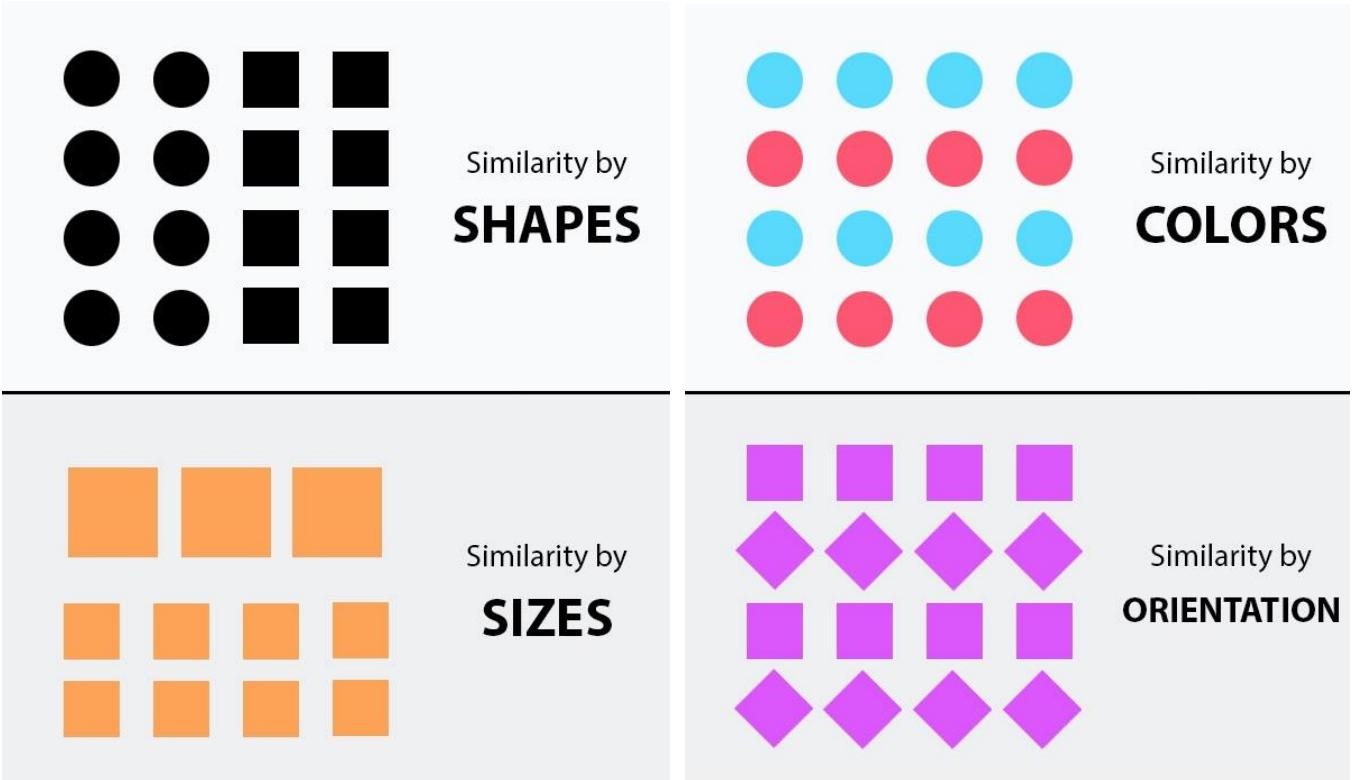
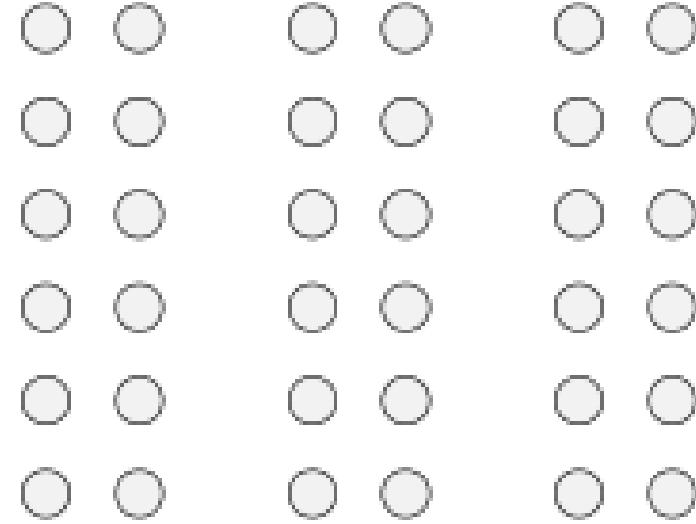
Connection



Enclosure



Psicología Gestalt



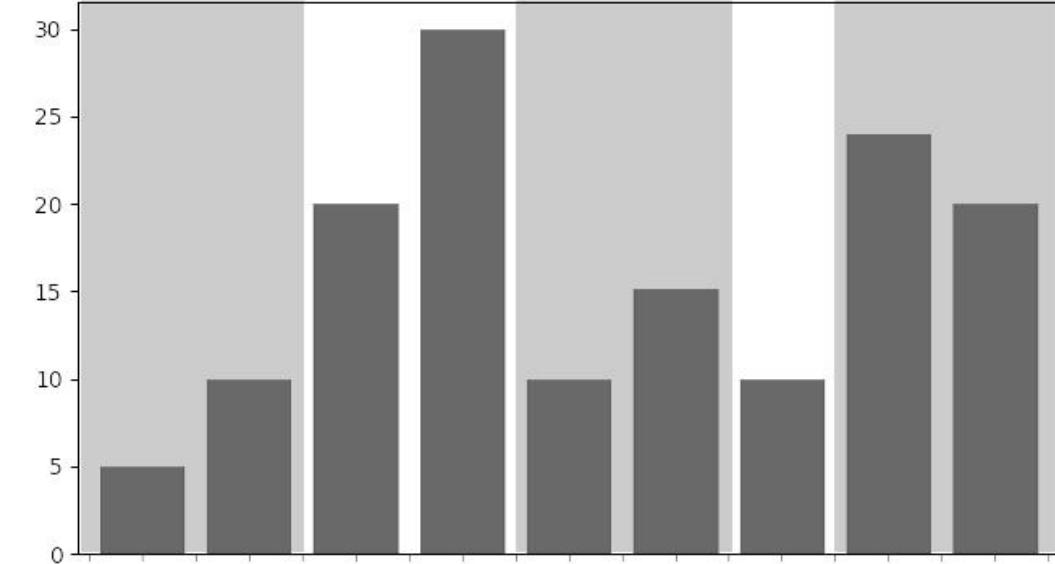
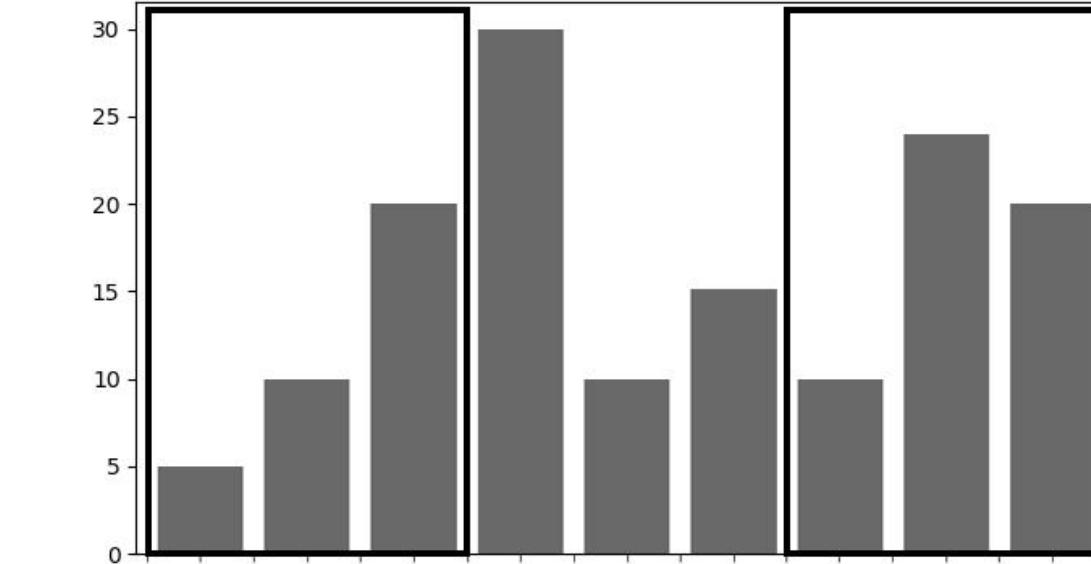
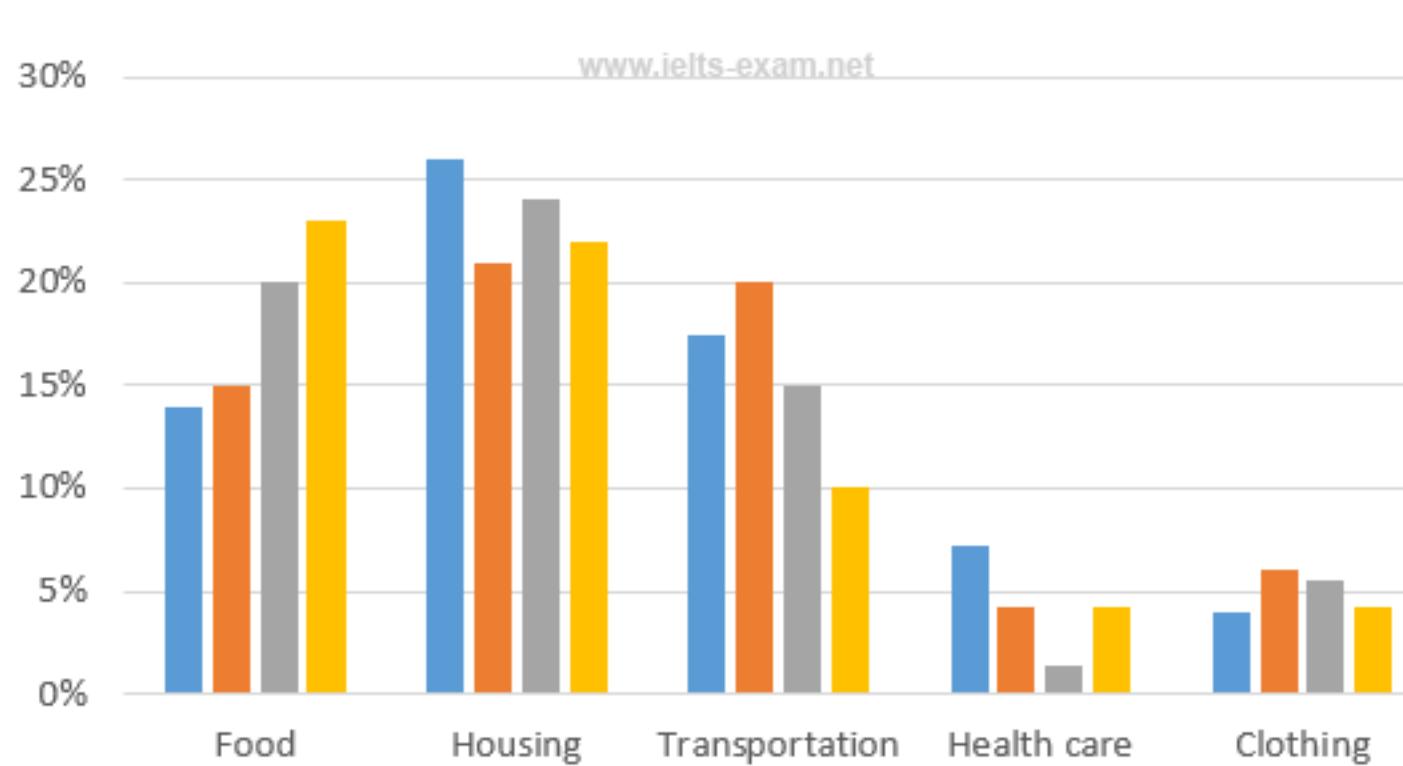
Proximity

Objetos cercanos se perciben como pertenecientes a un mismo grupo

1234 56789 1234
1234 56789 1234
123 2567 8912341
1234 6789 12 234

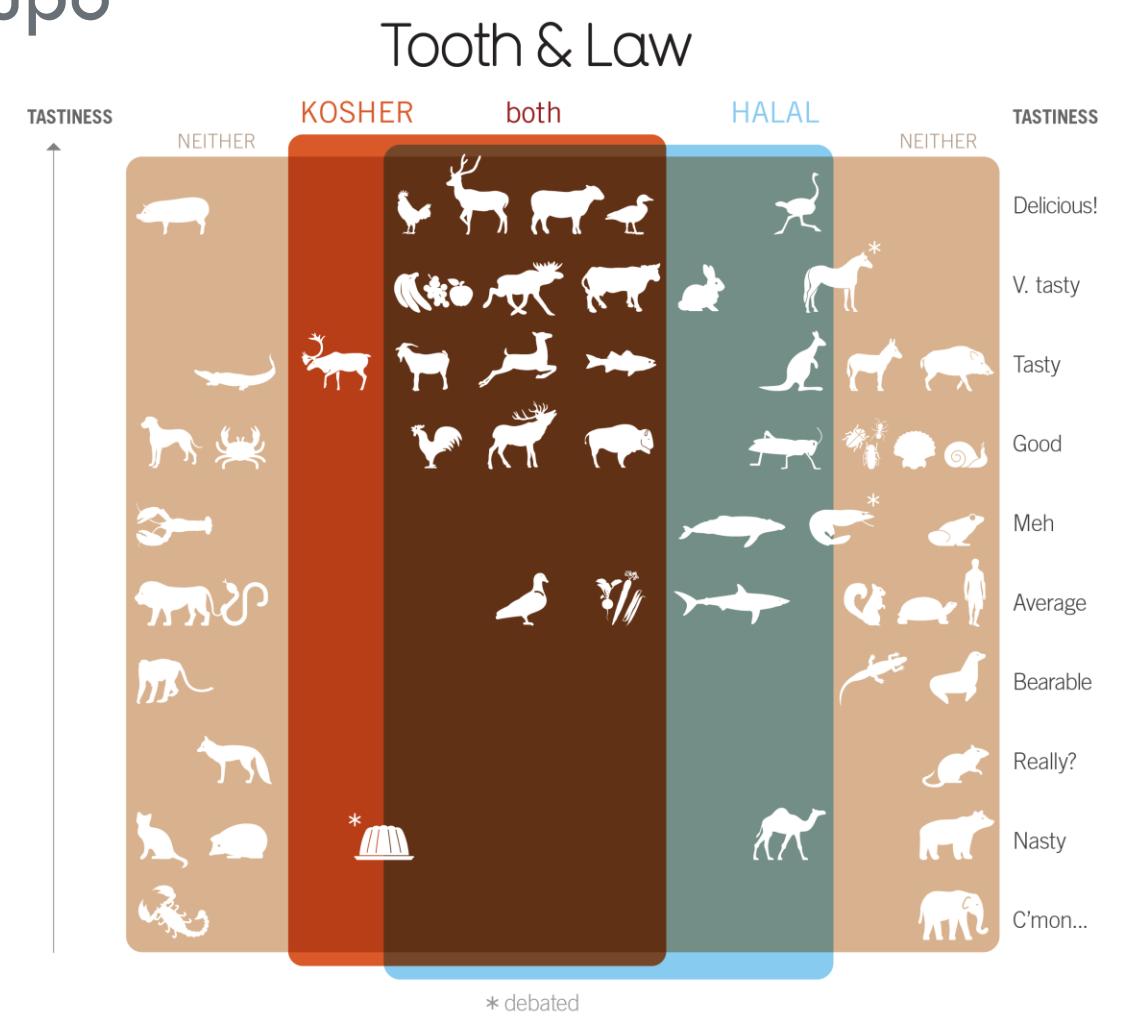
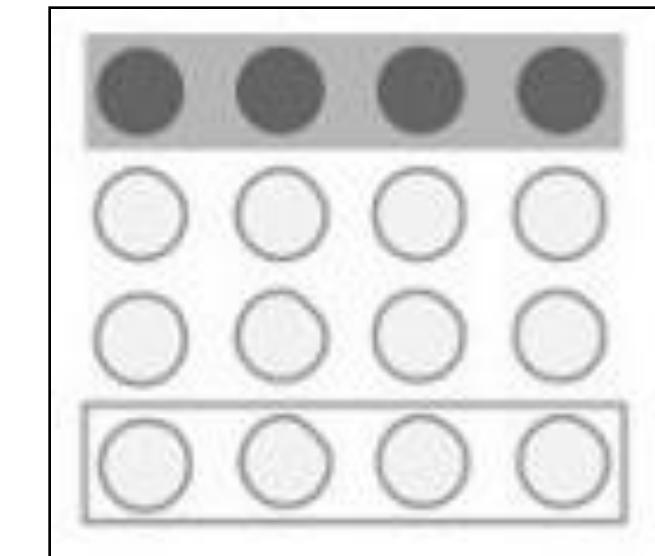
Similarity

Objetos idénticos se perciben como pertenecientes a un mismo grupo



Closure

Los objetos dentro de un área con bordes bien definidos se perciben como del mismo grupo



From informationisbeautiful.net

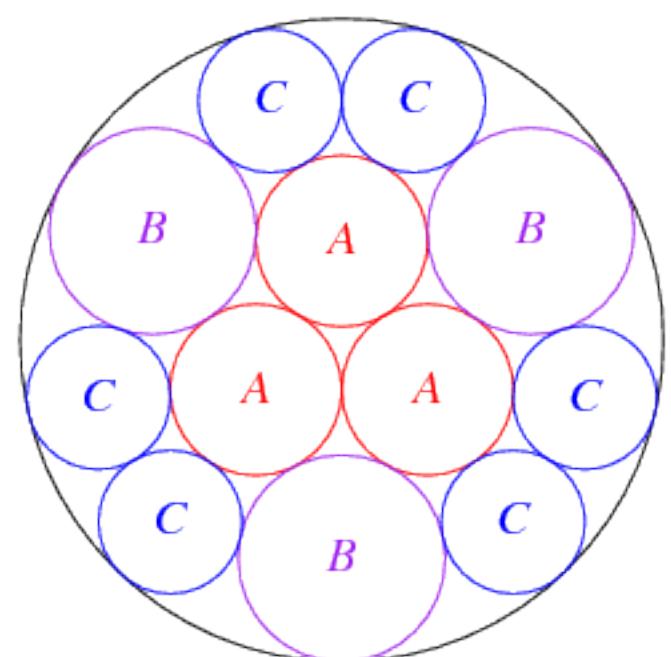
Grouping

Atributos Visuales que sirven para crear relaciones y categorías
a través de la posición en el espacio

Connectedness



Containment



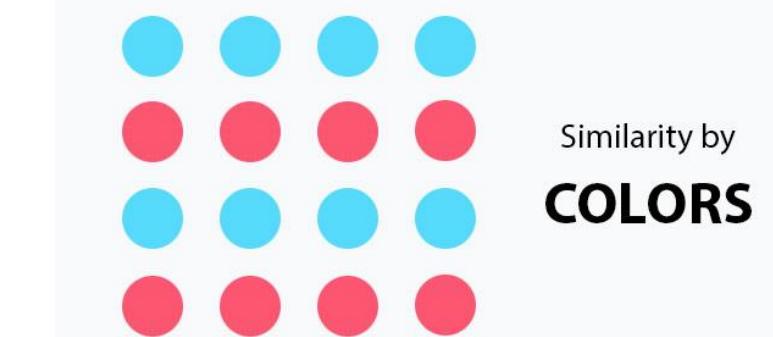
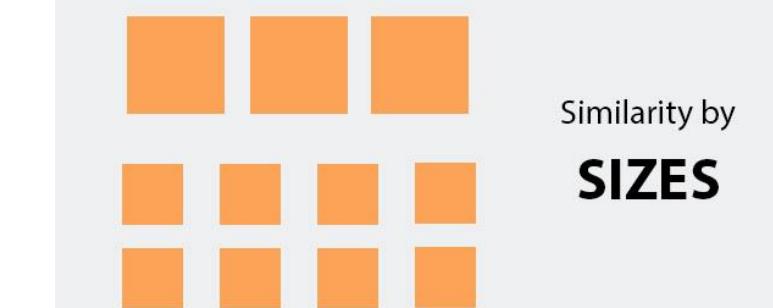
Proximity

1234 56789 1234

1234 56789 1234
123 2567 8912341

1234 6789 12 234

Similarity



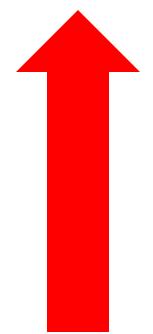
Contraste

34620191094

18496507394

47690759752

94709268921



Busca los números 6

Contraste

34620191094

18496507394

47690759752

94709268921

34620191094

18496507394

47690759752

94709268921

- Captamos diferencias de luminancia muy eficientemente
- Si las tablas de números son visualizaciones con la **función** de identificar el 6, la de la derecha es más efectiva porque está diseñada para explotar los mecanismos cognitivos que intervienen en la tarea
- La equivalencia de esta función cognitiva en un fundamento de diseño es: usar el máximo contraste posible para ayudar al usuario a llevar a cabo la tarea.

Contraste

34620191094
18496507394
47690759752
94709268921

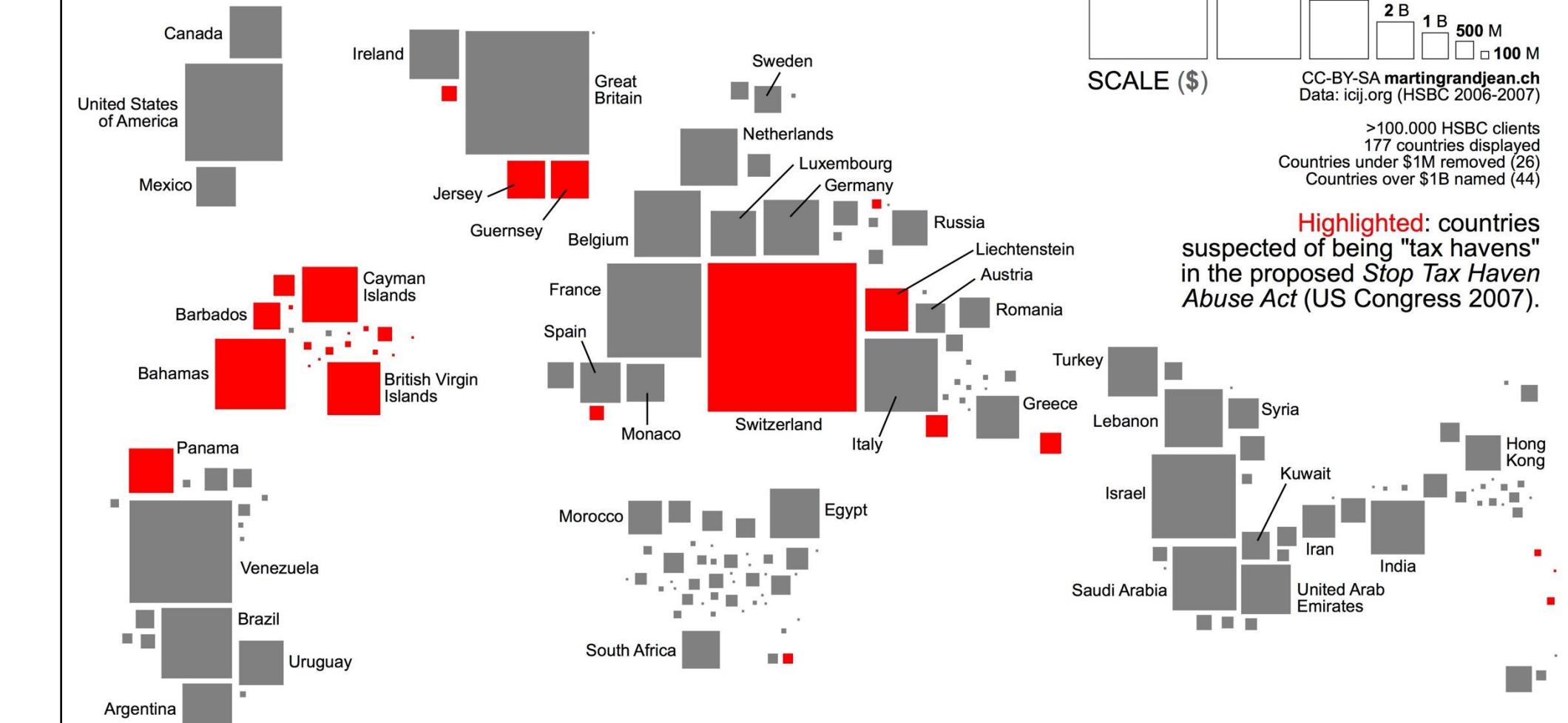
34620191094
18496507394
47690759752
94709268921

Fácil

▪ Algo menos, pero también

SWISS LEAKS | Globalized finance

Mapping leaked HSBC amounts by country.



- Captamos diferencias de luminancia muy eficientemente
- Si las tablas de números son visualizaciones con la **función** de identificar el 6, la de la derecha es más efectiva porque está diseñada para explotar los mecanismos cognitivos que intervienen en la tarea
- La equivalencia de esta función cognitiva en un fundamento de diseño es: usar el máximo contraste posible para ayudar al usuario a llevar a cabo la tarea.

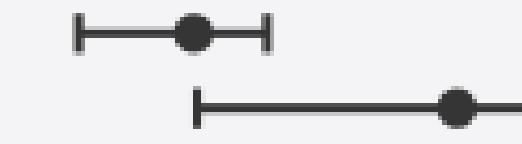
Channels: Expressiveness Types And Effectiveness Ranks

→ **Magnitude** Channels: **Ordered Attributes**

Position on common scale



Position on unaligned scale



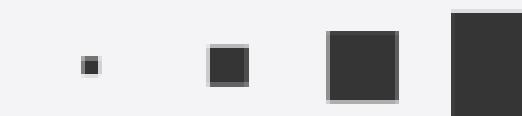
Length (1D size)



Tilt/angle



Area (2D size)



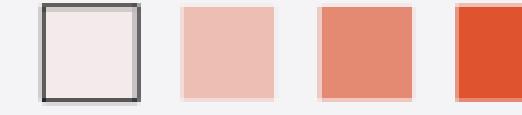
Depth (3D position)



Color luminance



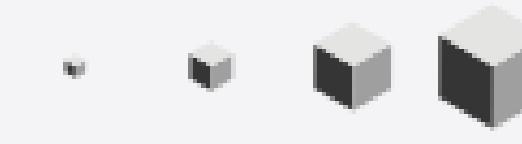
Color saturation



Curvature

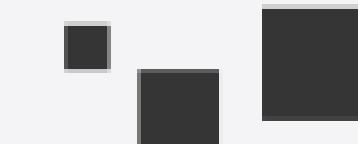


Volume (3D size)



→ **Identity** Channels: **Categorical Attributes**

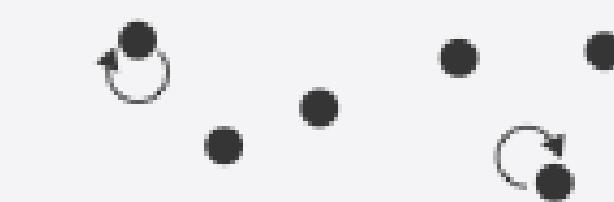
Spatial region



Color hue



Motion



Shape



Usamos estos mecanismos perceptuales en visualización.

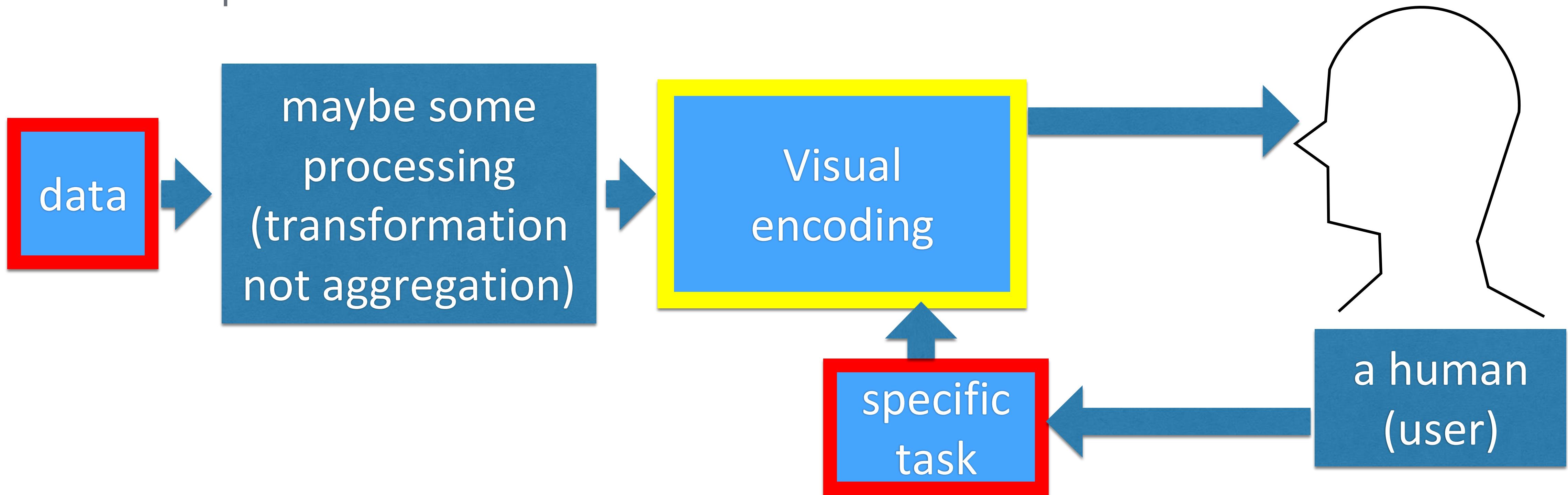
- Hacemos representaciones externas que utilizan nuestras habilidades inherentes de detección de patrones.
- Reducimos la carga cognitiva usando elementos gráficos que la mente puede detectar más rápidamente;
- utilizamos técnicas de codificación que maximicen el contraste entre elementos, dirijan la atención eficazmente, y guíen el razonamiento

Bibliografia

- **The Functional Art . Alberto Cairo, 2012 (Cap. 5-6)**
- Designing with the Mind in Mind. Jeff Johnson, 2010
- Visual Thinking for Design. Colin Ware, 2008
- Information Visualization. Colin Ware, 2013

Data Visualisation

- Datos. El proceso empieza con uno o más datasets. Conocemos el tipo y las características de sus atributos.
- Tareas. Definición de las tareas que podemos resolver, caracterizadas como acción + objetivo

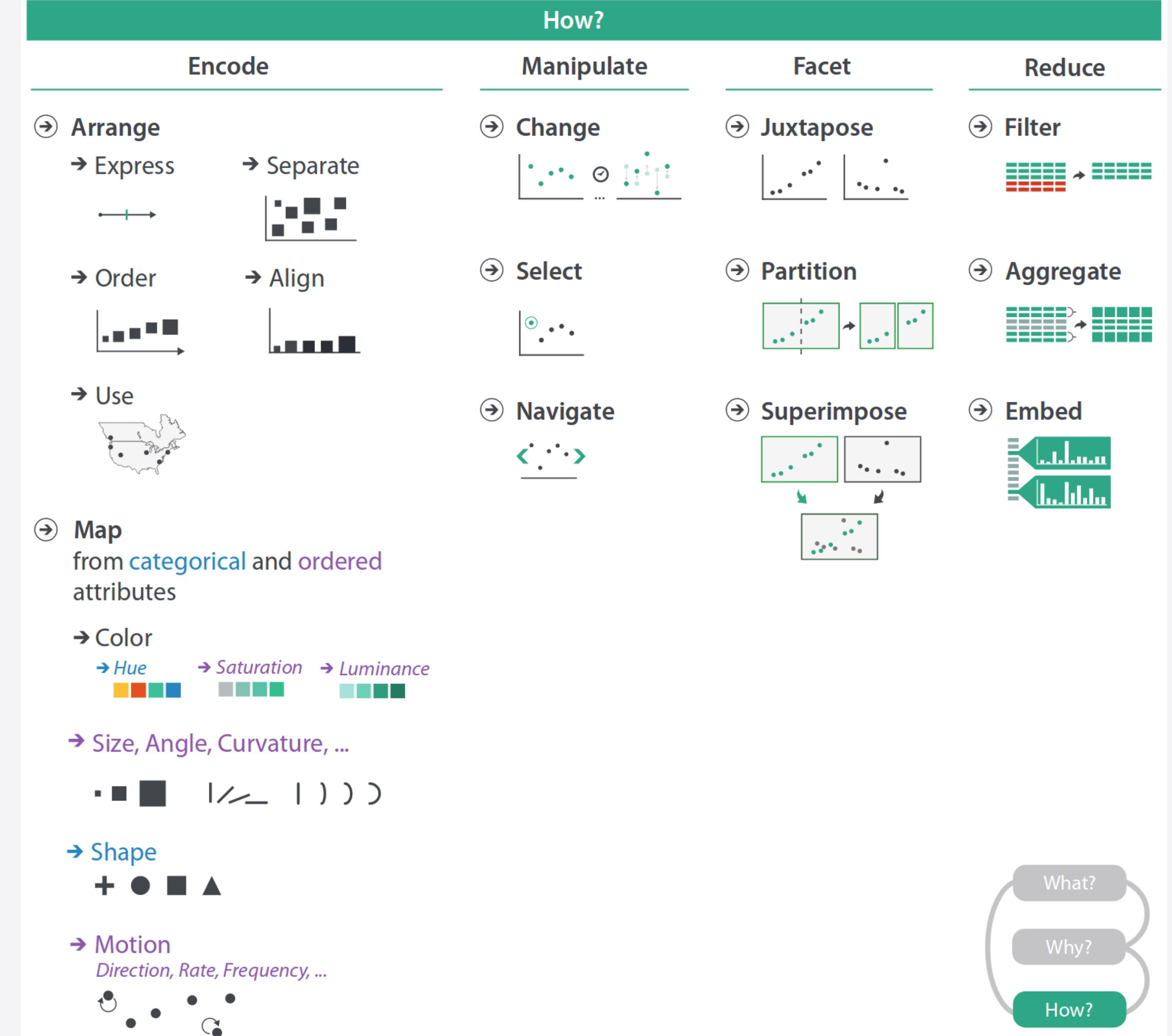


How?

Codificación

4 categorías

- Encode (Representar)
- Manipulate
- Facet (Separar)
- Reduce



Manipulate

- Change (highlight, update, animate...)
- Select
- Navigate /change viewpoint (zoom, pan, etc)

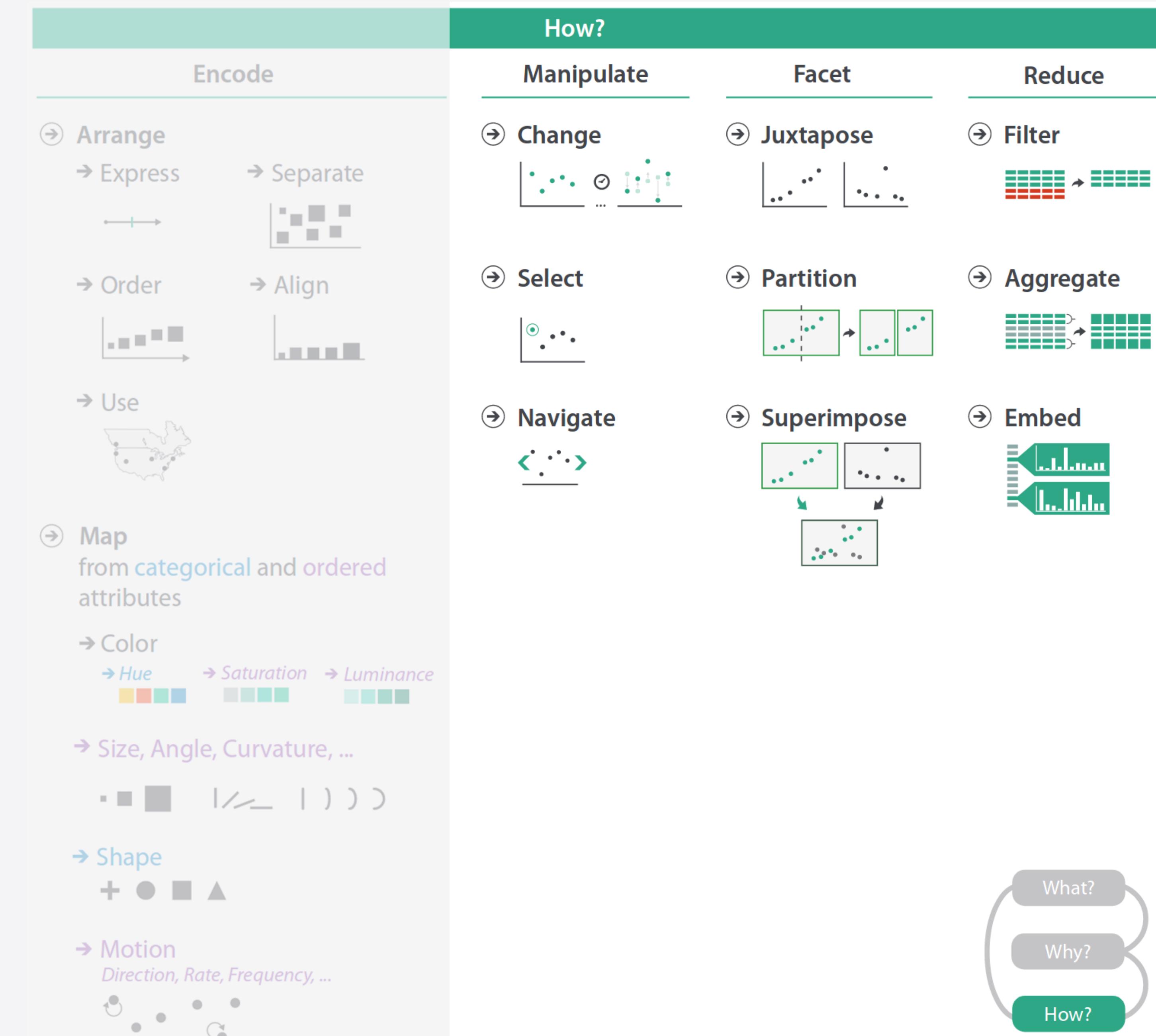
Facet

Partir los datos en varias vistas organizadas/ coordinadas

- Juxtapose
- Partition
- Superimpose

Reduce

- Filter
- Aggregate
- Embed (Incorporar)



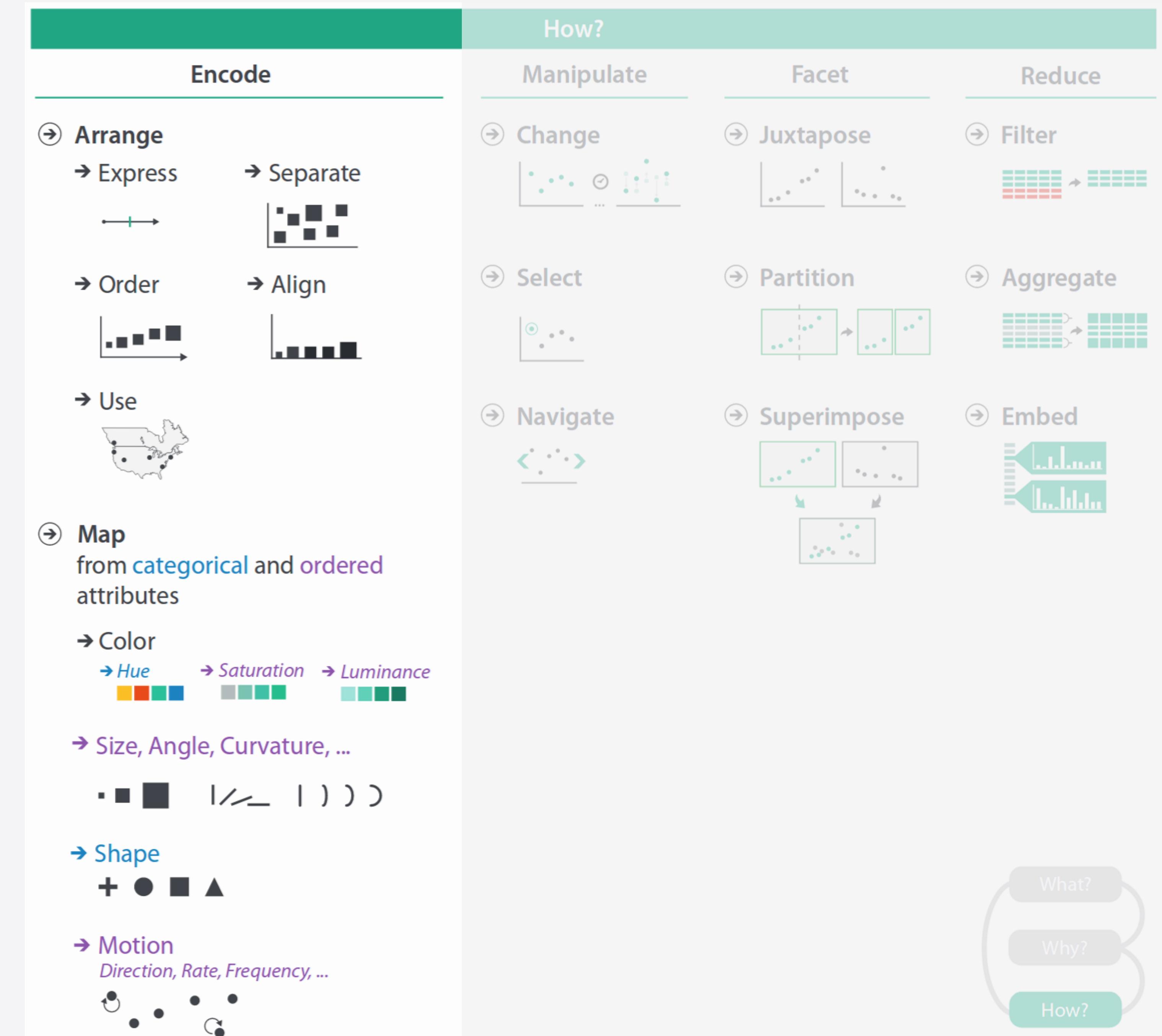
Codificación

Arrange

- Spatial arrangement
- Express values
- Separate, order and align regions
- Use spatial data

Map data into nonspatial visual channels

Color, shape, size, Angle, Curvature, Transparency, Motion, etc.



Ejercicio

Representa 14 y 33

Marcas y Canales

Asignación de propiedades gráficas a los datos

Marcas

Elementos geométricos básicos clasificadas por dimensiones

Canales

Los canales visuales que modulan la apariencia de las marcas



Position

→ Horizontal



→ Vertical



→ Both



Color



Shape



Tilt



Size

→ Length



→ Area



→ Volume



Marcas y Canales

Asignación de propiedades gráficas a los datos

Marcas: Barras, círculos // Canales: X, Y, tono, tamaño

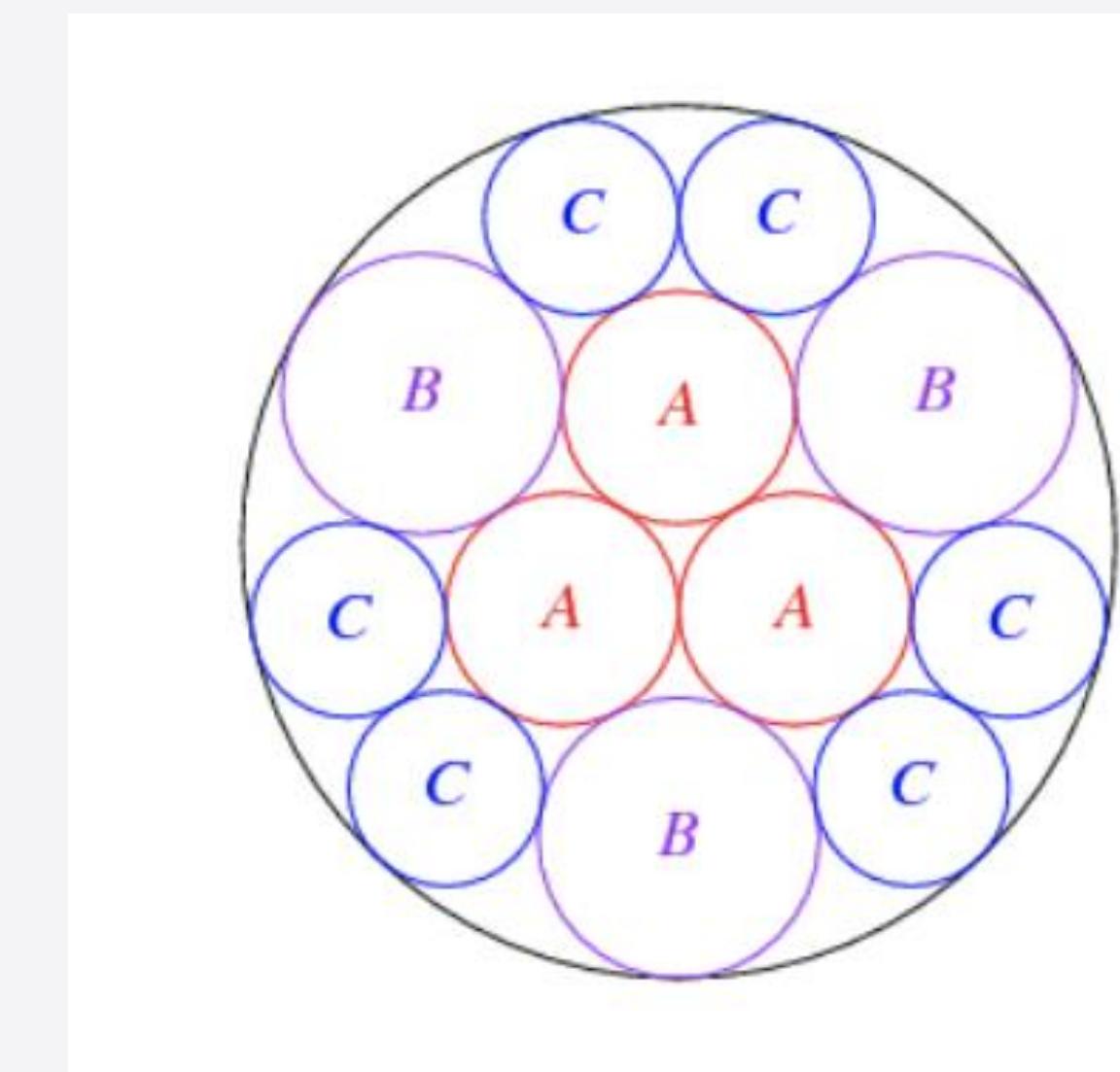
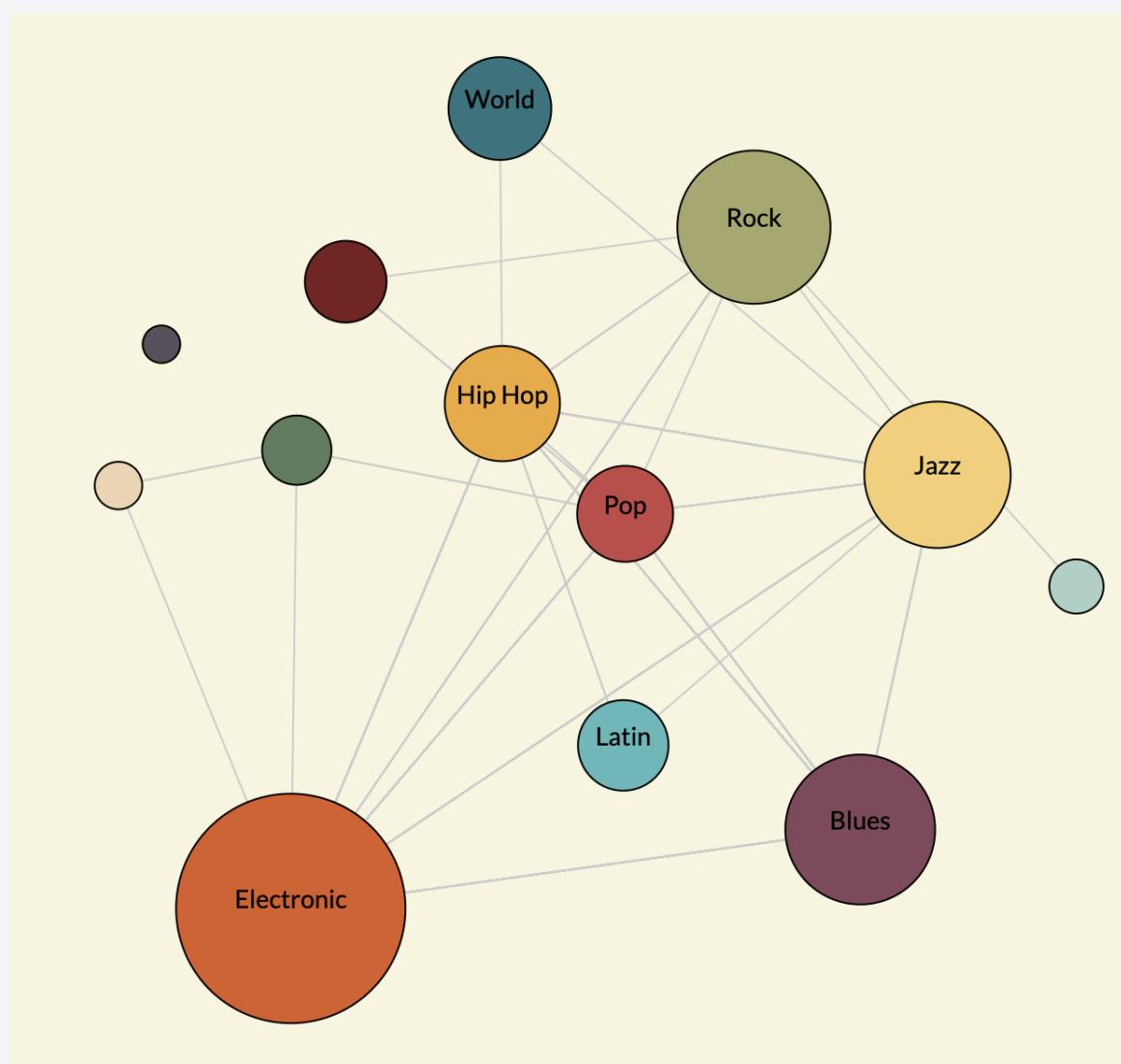
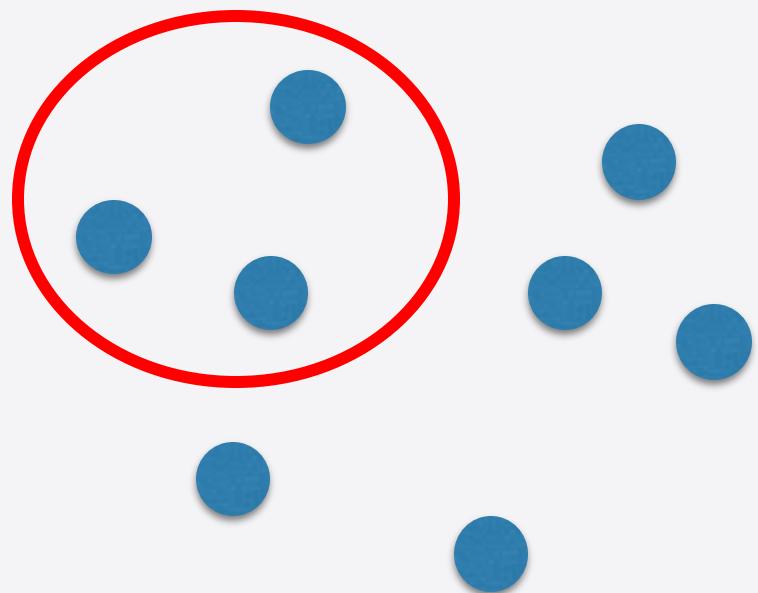
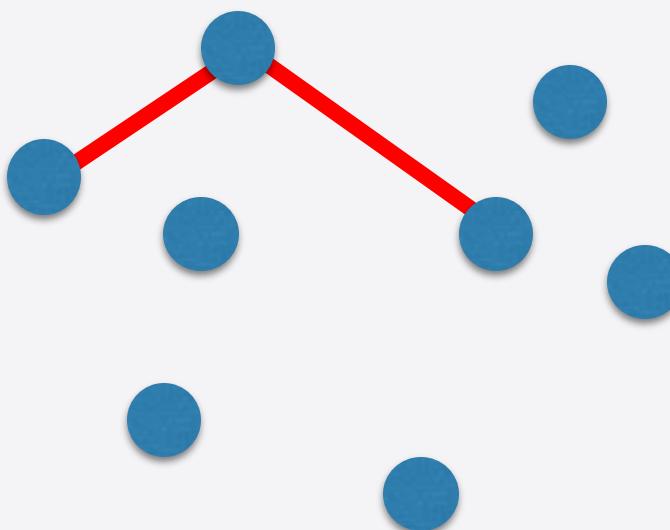


Marcas: Lineas
Canales: Longitud

Marcas: Puntos
Canales: Pos.

Marcas: Puntos
Canales: Pos. +
Color

Marcas: Puntos
Canales: Pos. +
Color +
Tamaño



Tipos de Marcas

- En **tablas** las **marcas** representan observaciones
- En **redes** usamos marcas para nodos (observaciones) y uniones.
- Marcas de **conexión o contenedoras**

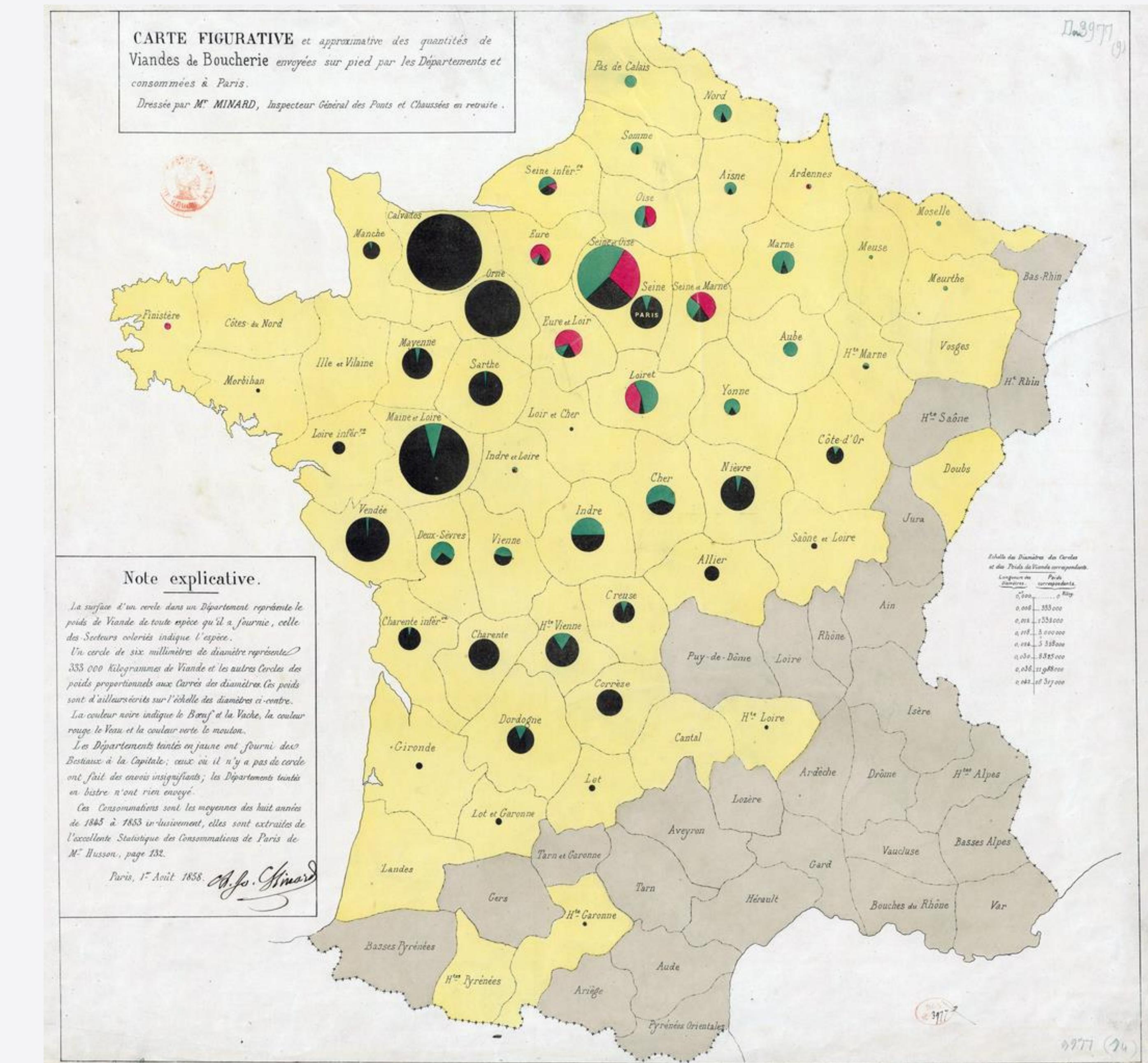
Marcas compuestas: Glifos

Distribución geográfica de cantidades de carne enviadas a París.

Ejemplo de 1858 de Charles Minard

Utiliza múltiples piecharts para categorizar envíos de carne desde Paris.

- Provincias en gris no tienen envíos
- Tamaño = volumen del envío, el más grande es 330 T
- Color es tipo de carne (negro=vaca, verde=cordero, rojo=ternera).



Marcas compuestas:

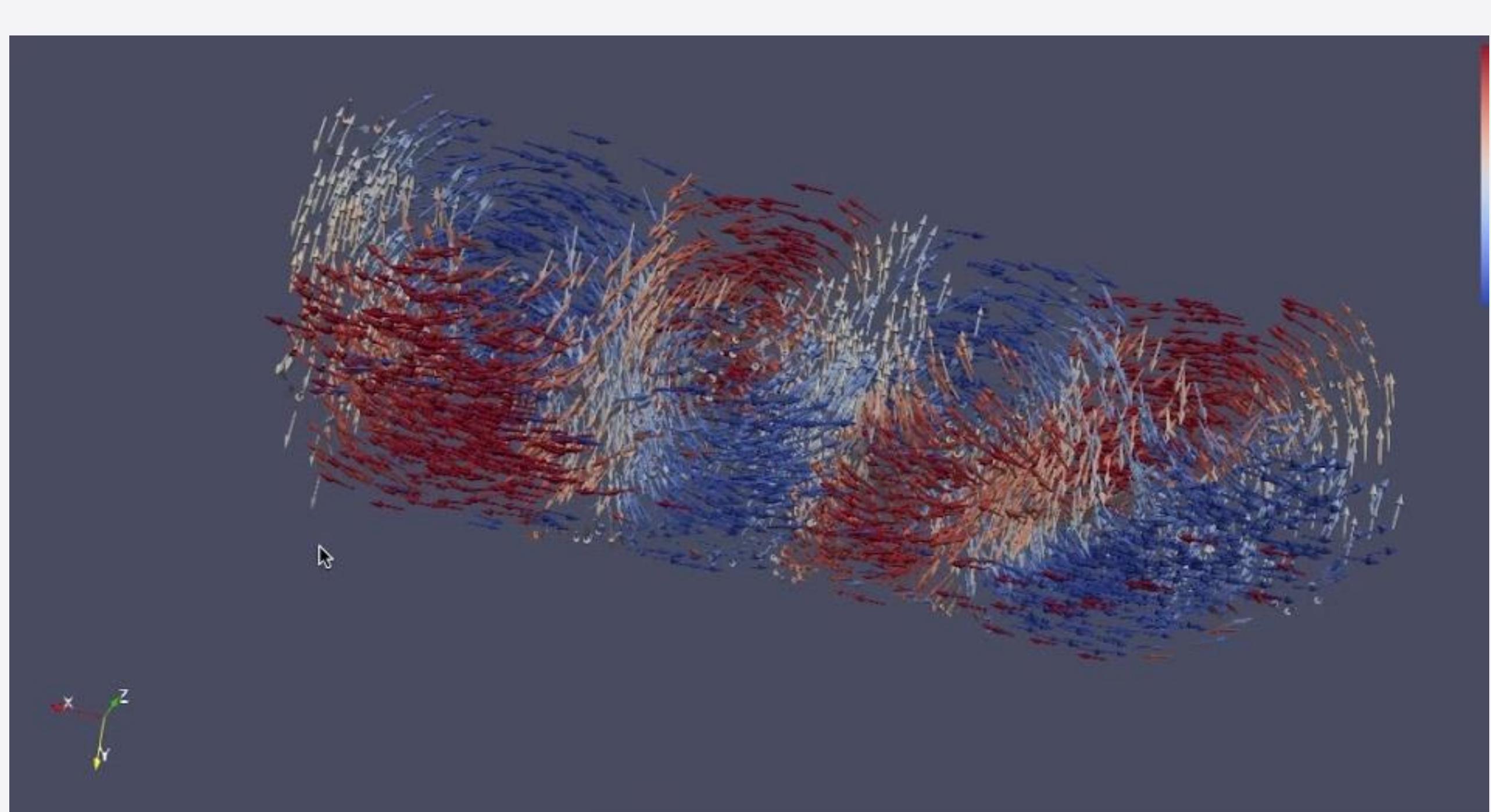
Glifos

Glyphs show linear trends of USHCN stations, 1950–2010.

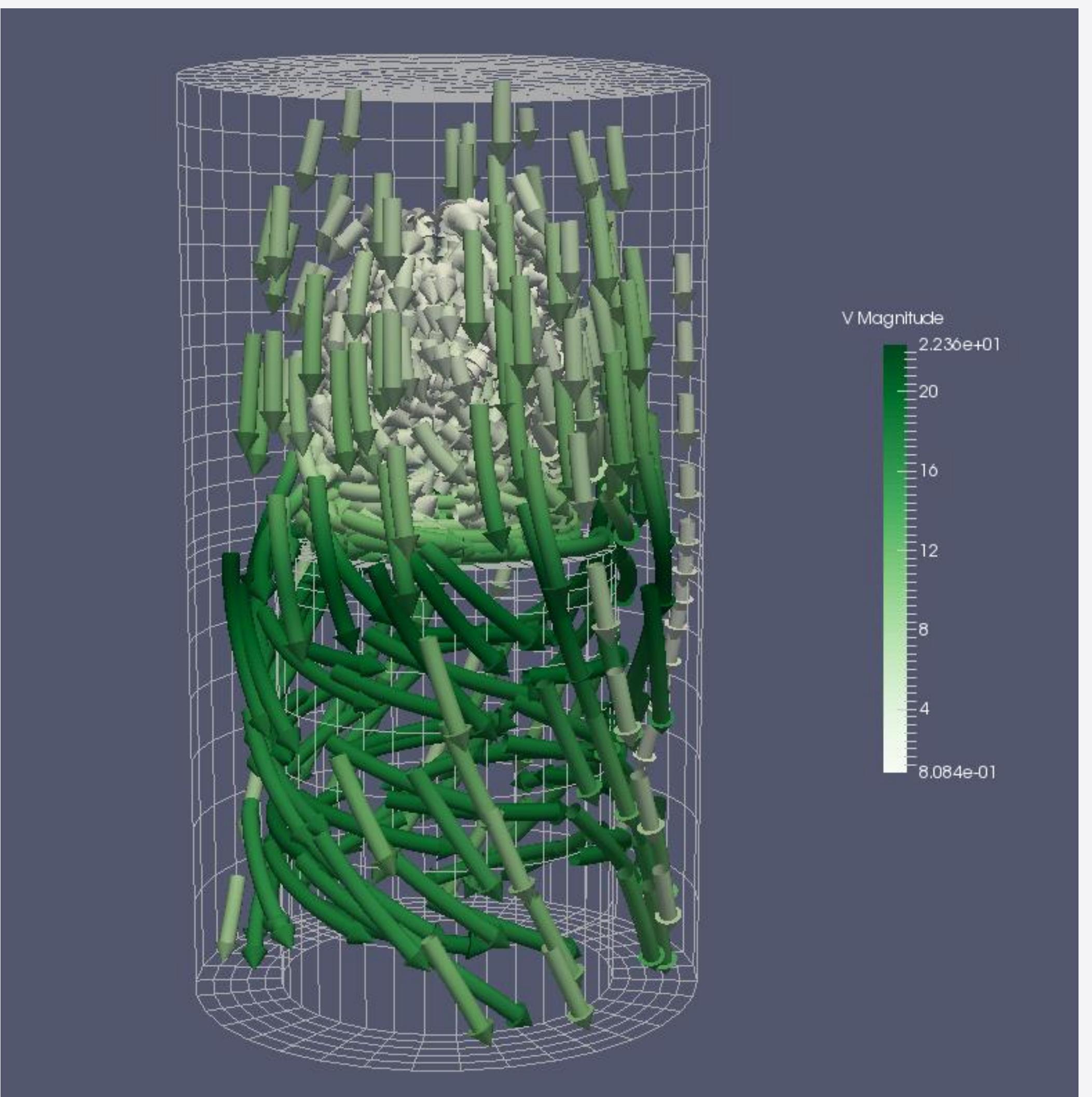


Wickham, H., Hofmann, H., Wickham, C., & Cook, D. (2012). Glyph-maps for visually exploring temporal patterns in climate data and models. *Environmetrics*, 23(5), 382-393.

Marcas compuestas: Glifos

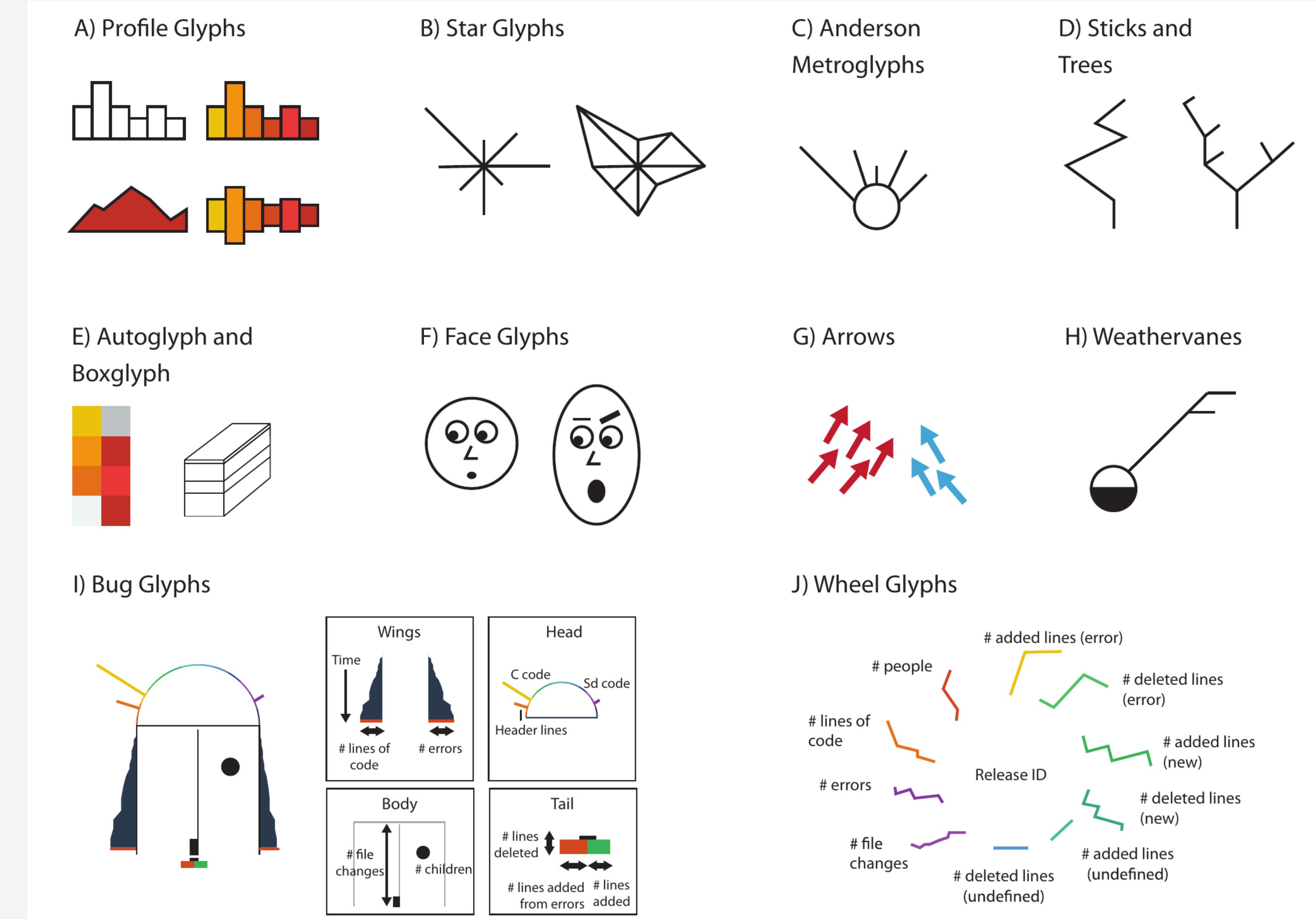


From Youtube / Liam Cooper



blog.kitware.com

Marcas compuestas: Glifos



Systematising Glyph Design for Visualization, de E.J aguire
From Eduardo Graells

Tipos de Canales

- No todos los canales son iguales
- Percibimos dos tipos generales de modalidades sensoriales:
- Canales de Identidad – Dan información de *Qué es algo o Donde está*)
- Canales de Magnitud – Nos dice *Cuanto hay de algo*.

Identidad

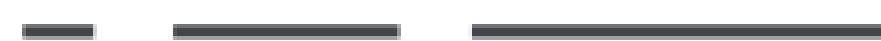
→ Shape



e.g., podemos decir *Qué forma vemos*
¿Tiene sentido hacer preguntas de magnitud?

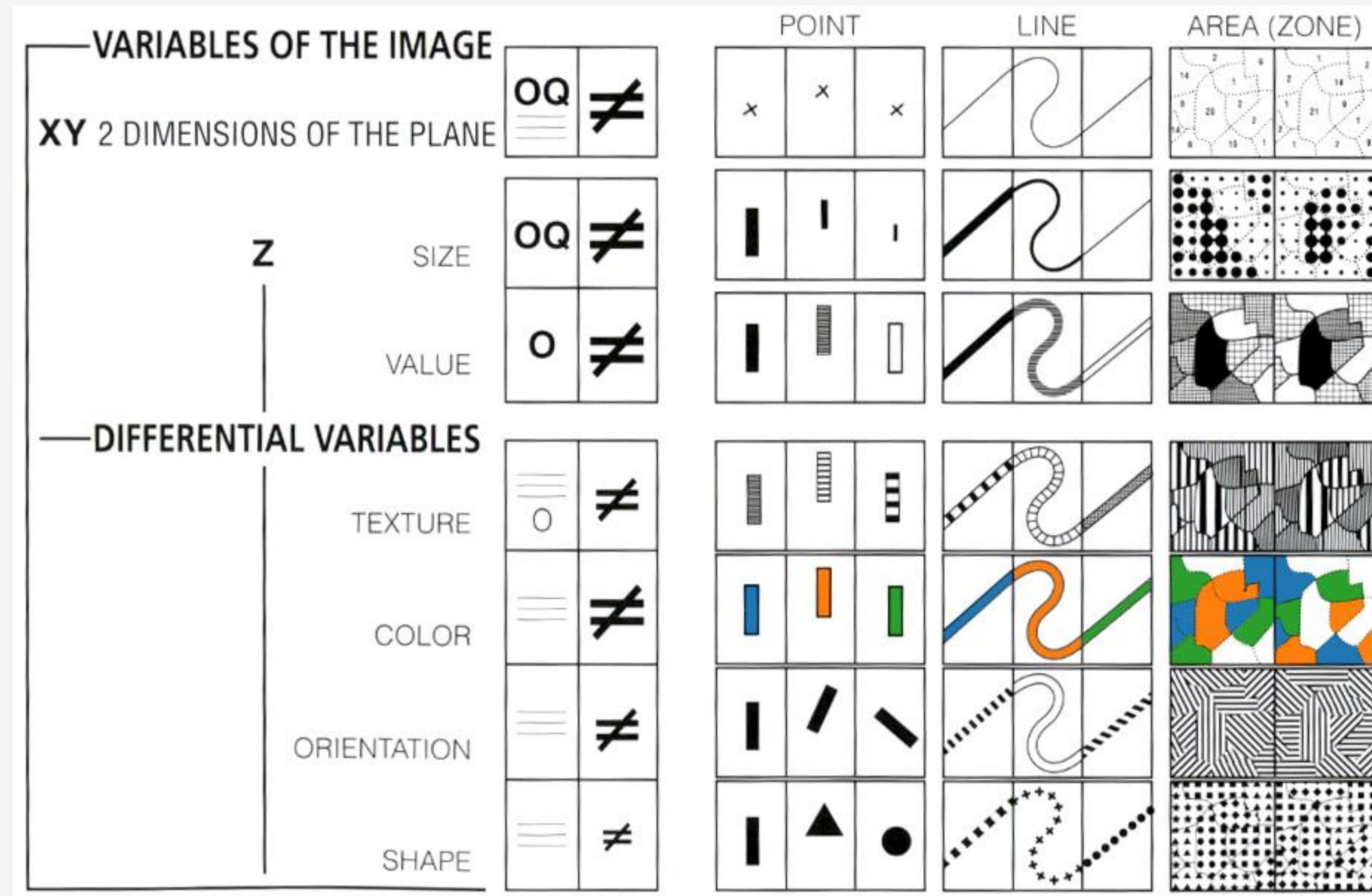
Magnitud

→ Length



e.g., podemos hacer preguntas de magnitud para la
longitud. ¿Preguntar identidad?

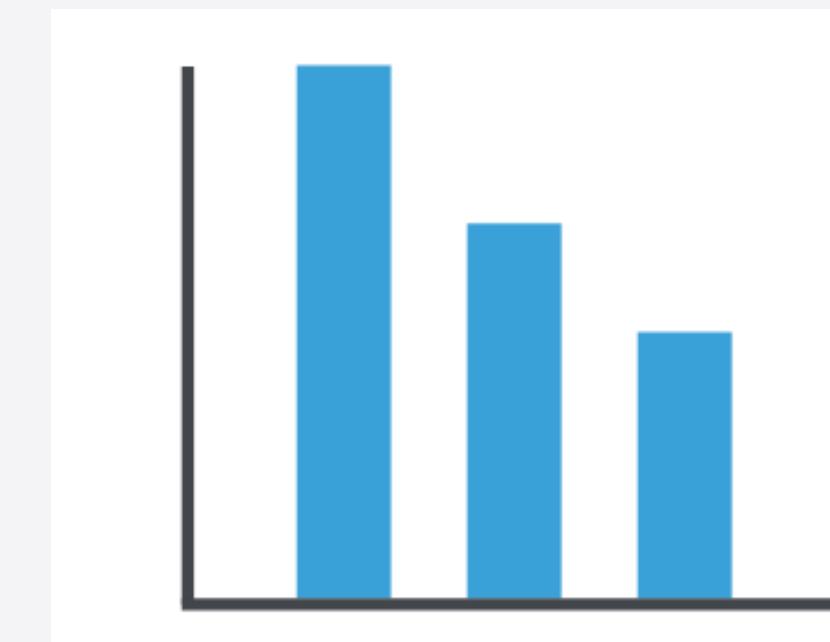
Marcas y Canales



Marcas y Canales

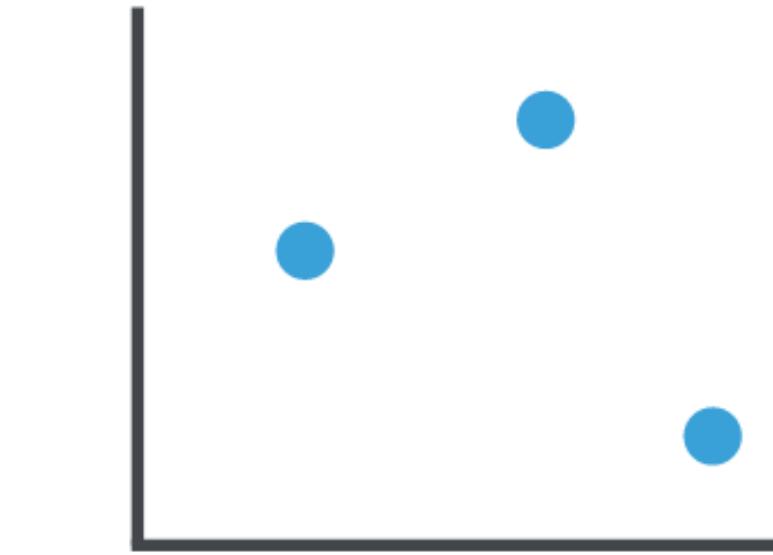
- La importancia del tipo de atributo está en la codificación en canales de **Magnitud o Identidad**
- Estas gráficas incorporan variables en canales de identidad o de magnitud, según el tipo de variable

1 cuantitativo



Marcas: Lineas
Canales: Longitud

2 cuantitativos



Marcas: Puntos
Canales: Pos.

**2 cuantitativos
+ 1 categorico**



Marcas: Puntos
Canales: Pos. +
Color

**3 cuantitativos
+ 1 categorico**



Marcas: Puntos
Canales: Pos. +
Color +
Tamaño

Principios de Expresividad y Efectividad

Expresividad:

A visualization is said to be expressive if, and only if, it encodes all the data relations intended and no other data relations.

[Card, 2008, p. 523]

Traducido:

- Mostrar la verdad y nada más que la verdad
- No usar codificaciones que impliquen relaciones que no están en los datos.

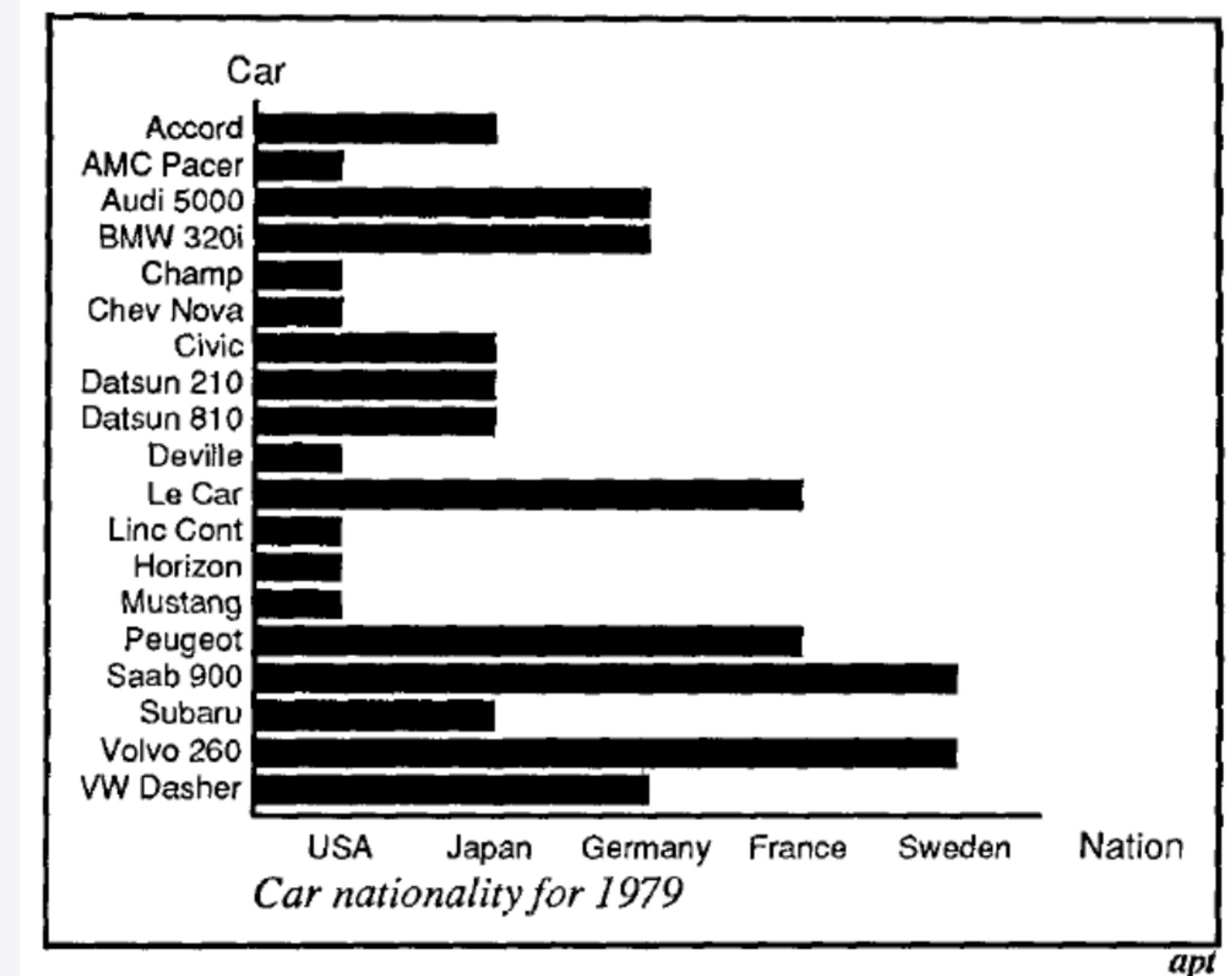
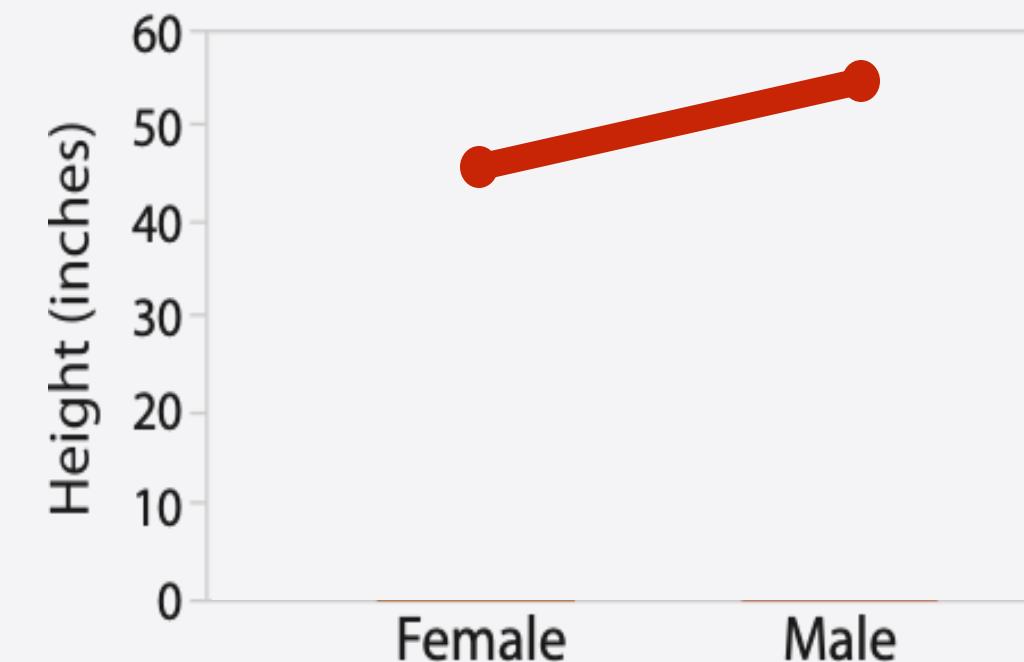
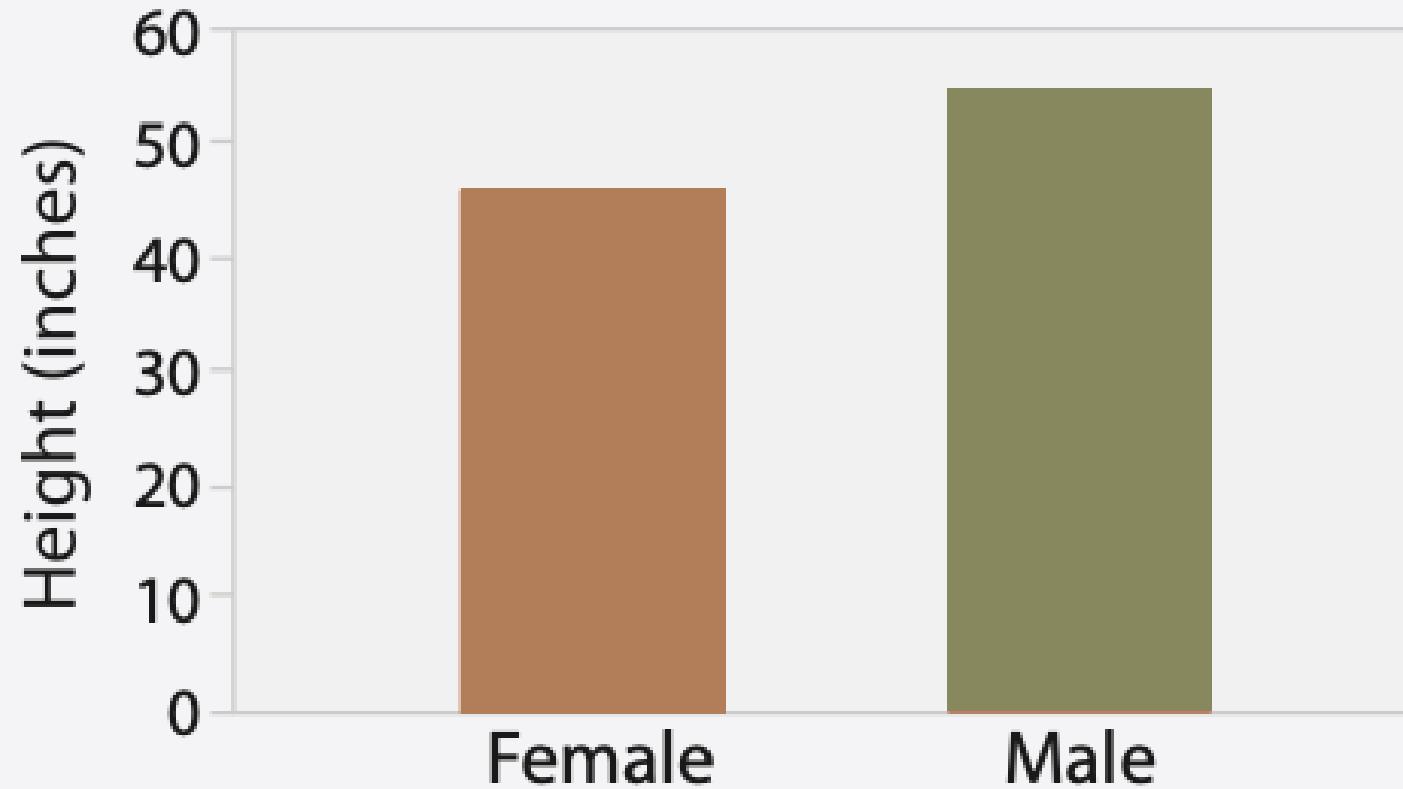


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

Automating the Design of Graphical Presentations of Relational Information, Mackinlay

Principios de Efectividad y Expresividad

Efectividad:

La importancia de un atributo debe ser proporcional a la **saliencia** del canal visual que lo representa.

Traducido:

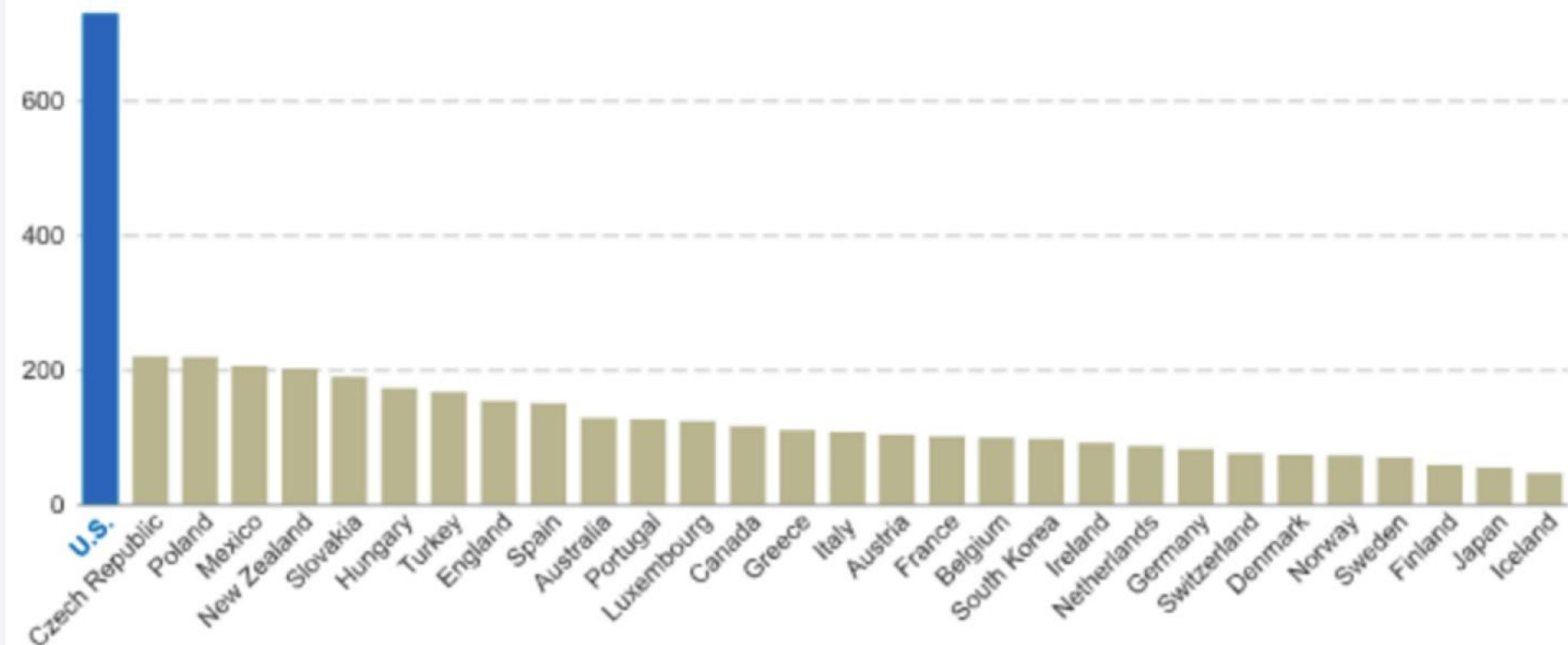
Utiliza los canales más efectivos para los atributos más importantes y viceversa.

The U.S. has the highest incarceration rate of any country in the world, imprisoning about 730 out of every 100,000 citizens.

Incarceration Rates for Countries in the OECD

800 prisoners per every 100,000 citizens

Source: International Centre for Prison Studies



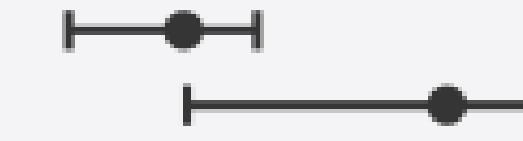
Channels: Expressiveness Types And Effectiveness Ranks

→ **Magnitude** Channels: **Ordered Attributes**

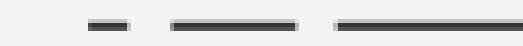
Position on common scale



Position on unaligned scale



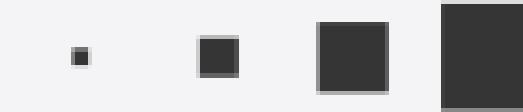
Length (1D size)



Tilt/angle



Area (2D size)



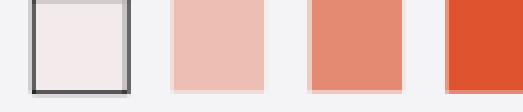
Depth (3D position)



Color luminance



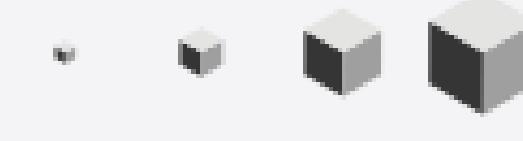
Color saturation



Curvature



Volume (3D size)



→ **Identity** Channels: **Categorical Attributes**

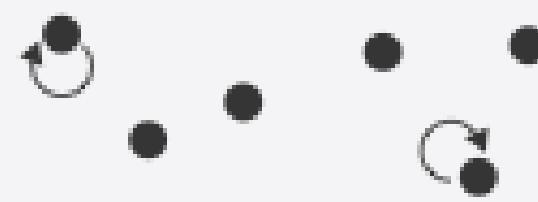
Spatial region



Color hue



Motion



Shape



Canales visuales separados en Magnitud e Identidad
Orden vertical según efectividad

Principio de expresividad:

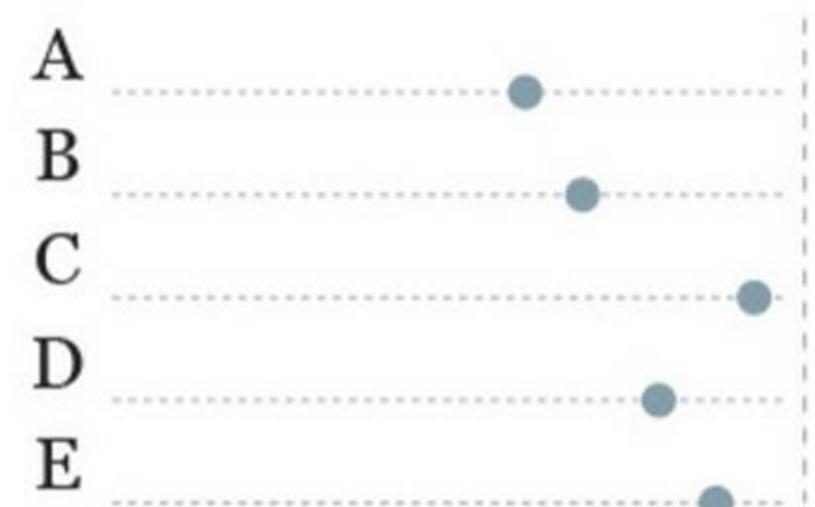
- Canales de Identidad son la elección correcta para atributos categóricos sin orden o relación intrínseca
- Canales de magnitud para atributos con orden inherente: ordinales y cuantitativos

Figures represented
in all these graphics:
22%, 25%, 34%, 29%, 32%

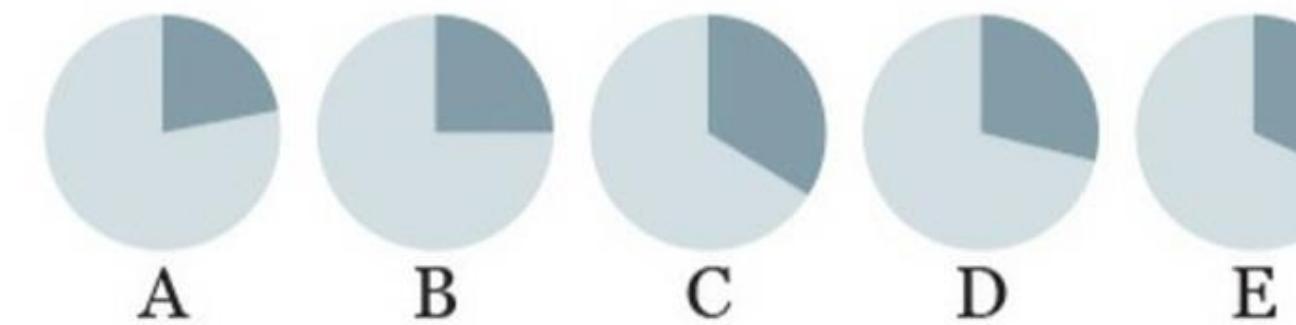
Length or height



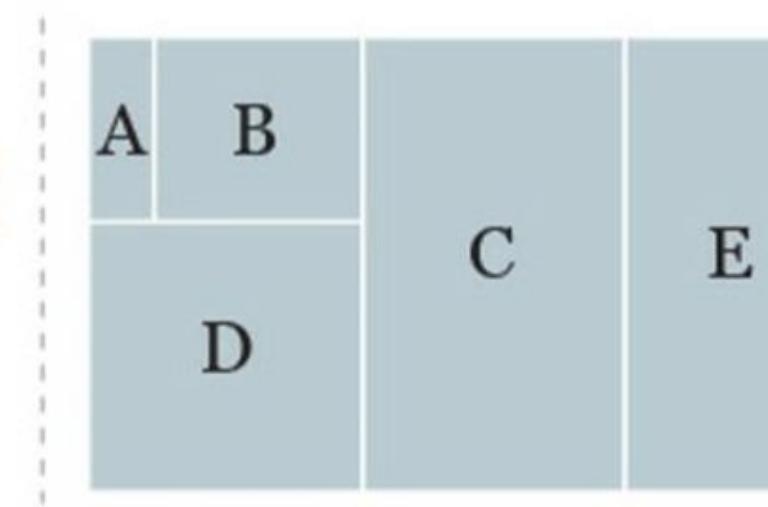
Position



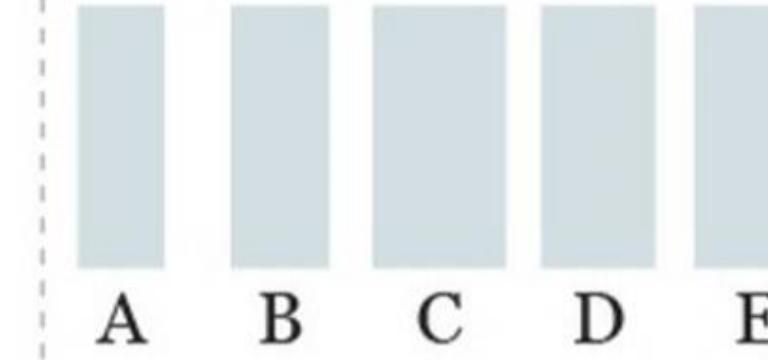
Angle/area



Area



Line weight



Efectividad

Estudios experimentales que evalúa características de una visualización:

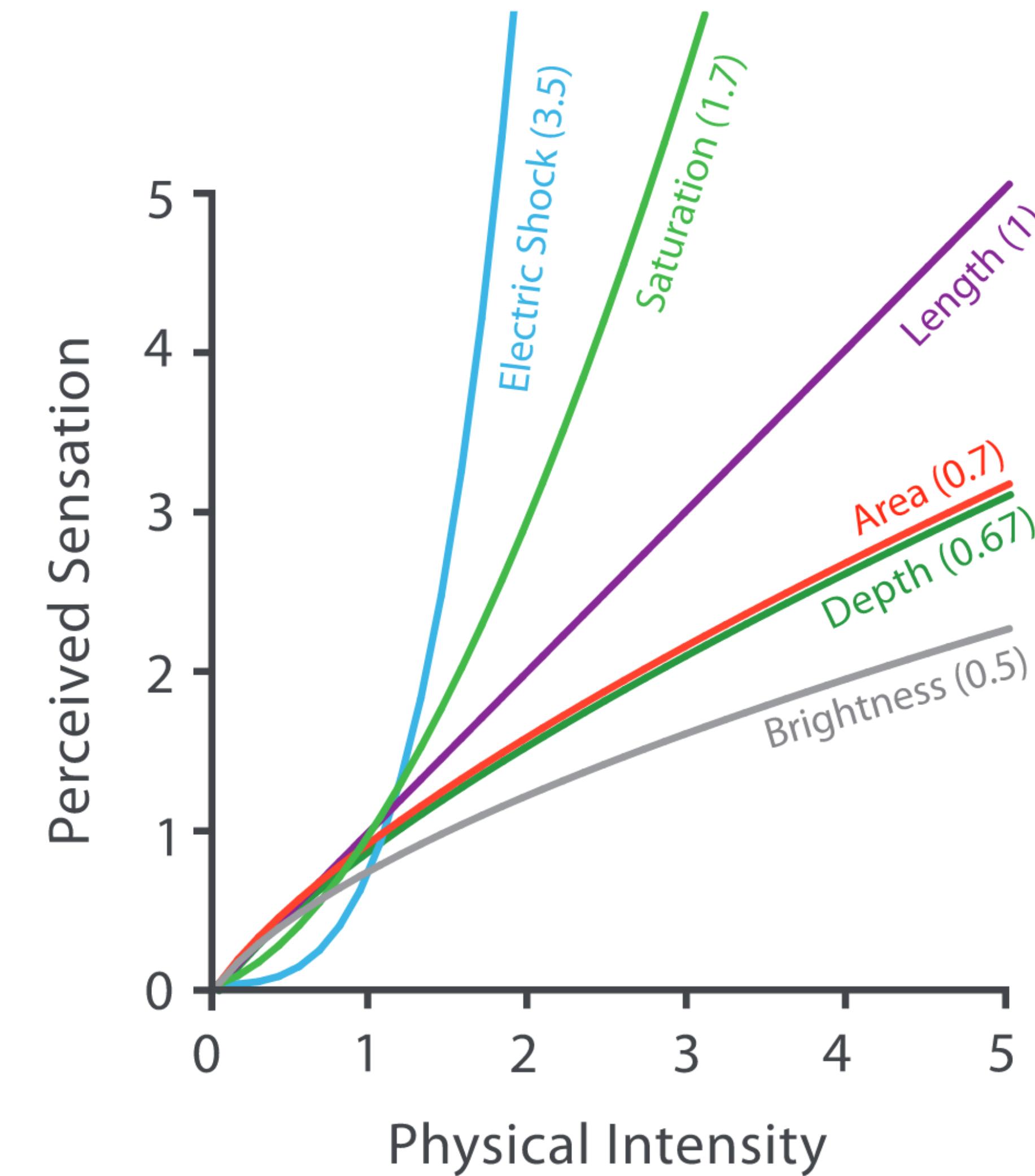
- **Precisión**
- **Discriminabilidad**
- **Separabilidad**
- **Salienteza**
- **Relatividad**

Precisión

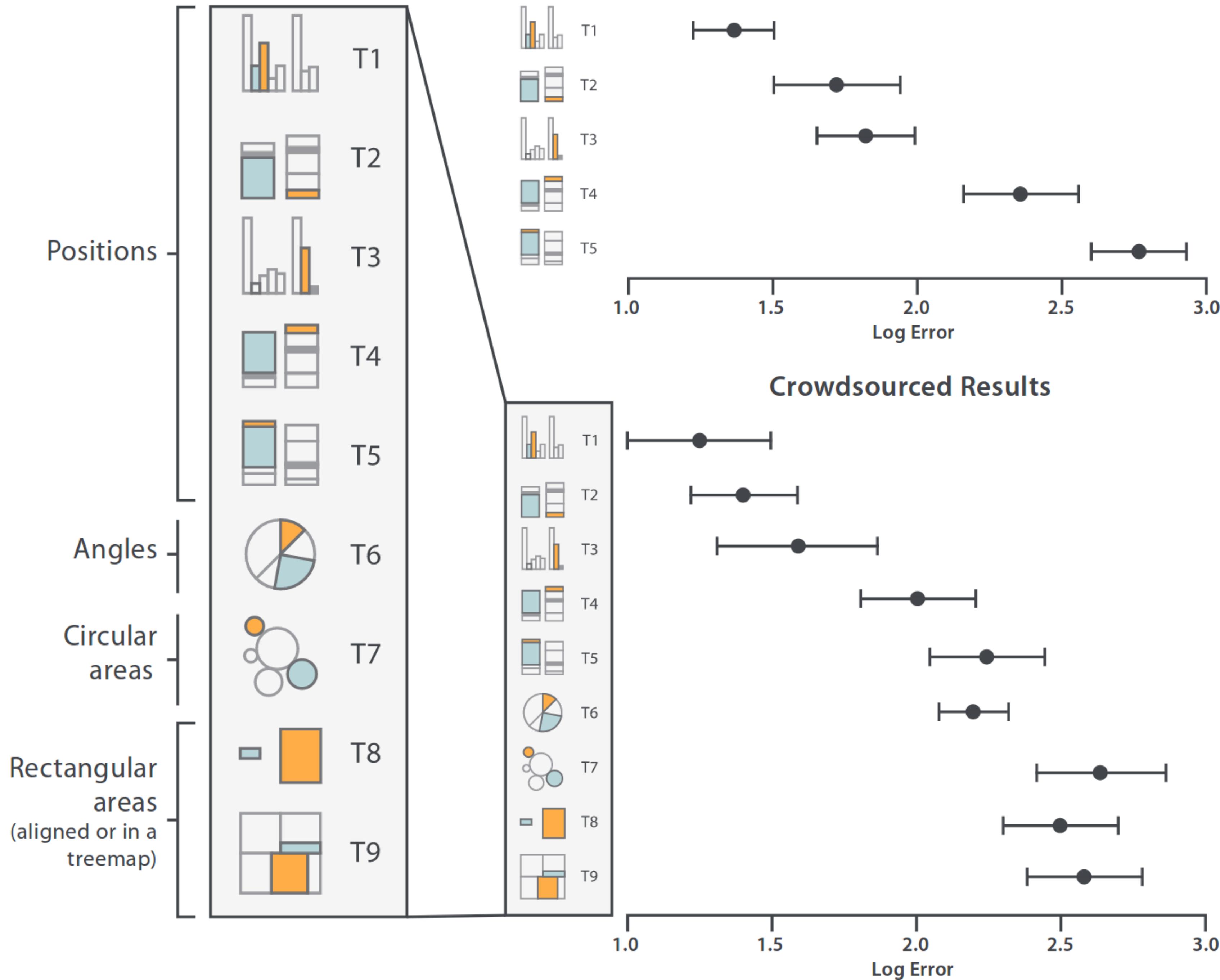
Ley de Stevens (1957)

- Psicofísica – Área de psicología que estudia la percepción humana.
- Experiencia sensorial de Magnitud se puede caracterizar por una ley de potencias donde el exponente depende de la modalidad.
- La mayoría de estímulos son magnificados o reducidos.

Steven's Psychophysical Power Law: $S = I^n$



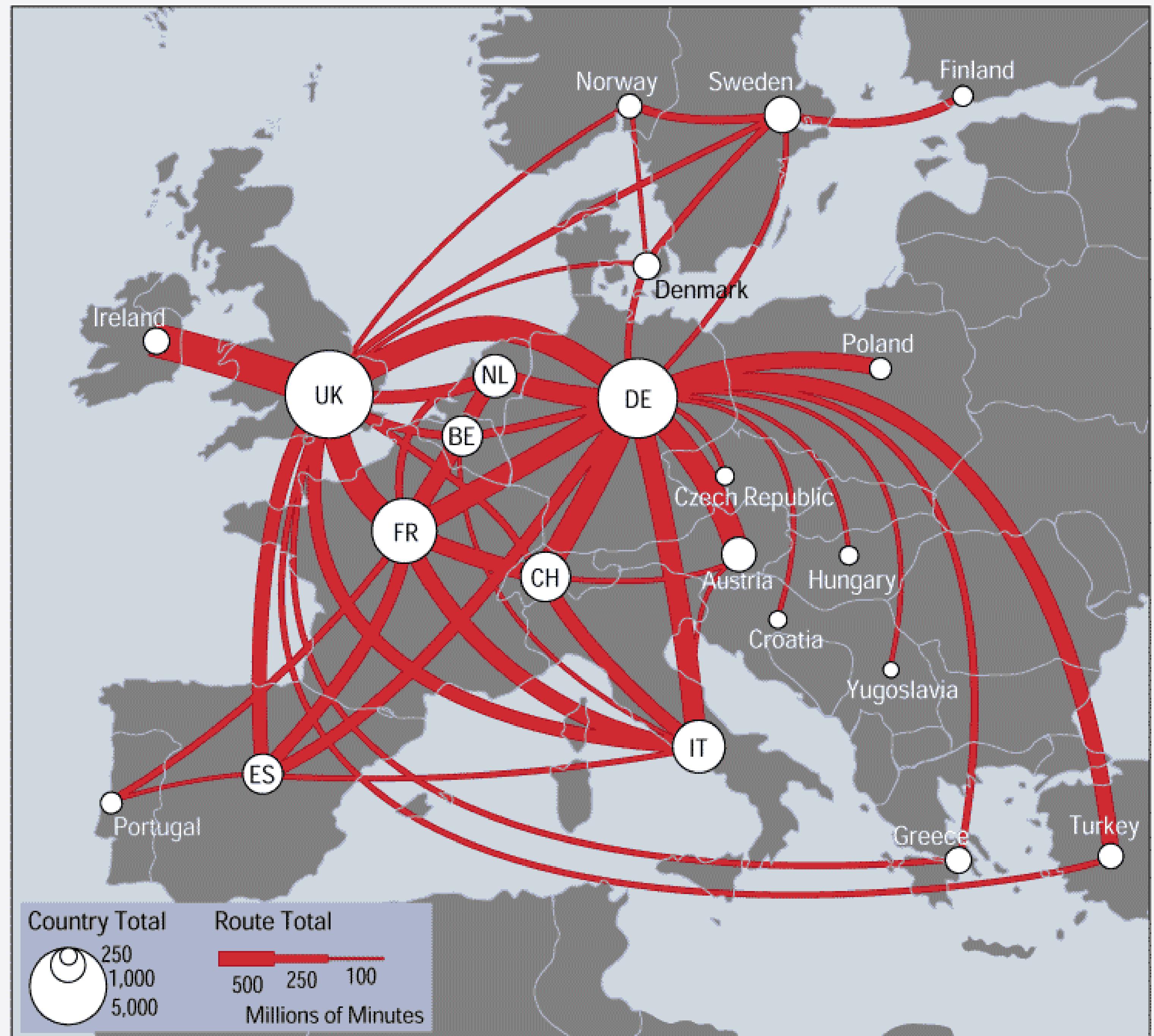
Precisión



Efectividad

Discriminabilidad

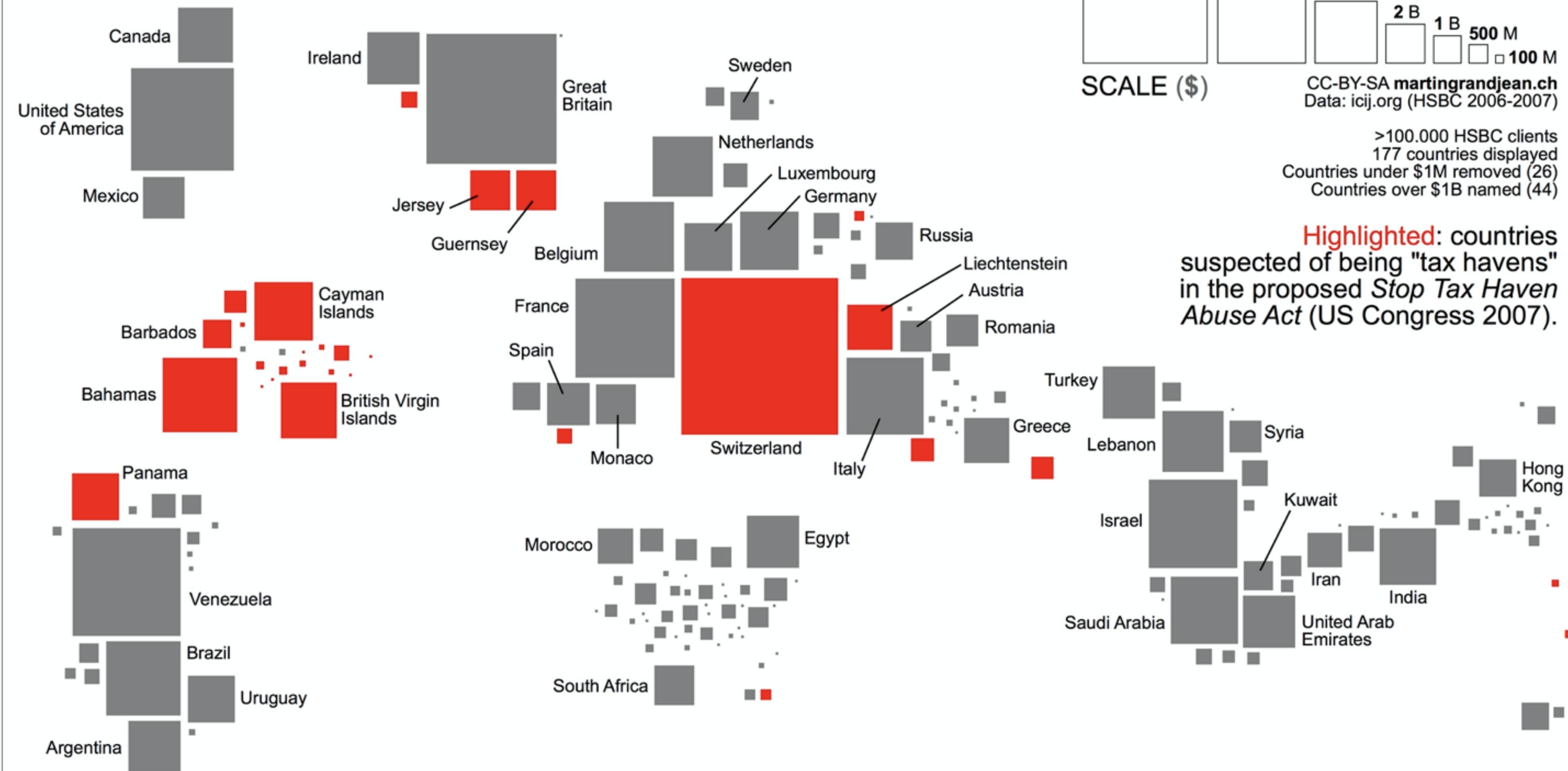
- En el canal visual utilizado se deben poder discernir las diferencias entre items.
- Para ello asignamos un número de Bins disponible para usar en el canal
- Un Bin es cada nivel distingible de los demás
- Muchas veces implica discretizar la escala



Discriminabilidad

SWISS LEAKS | Globalized finance

Mapping leaked HSBC amounts by country.

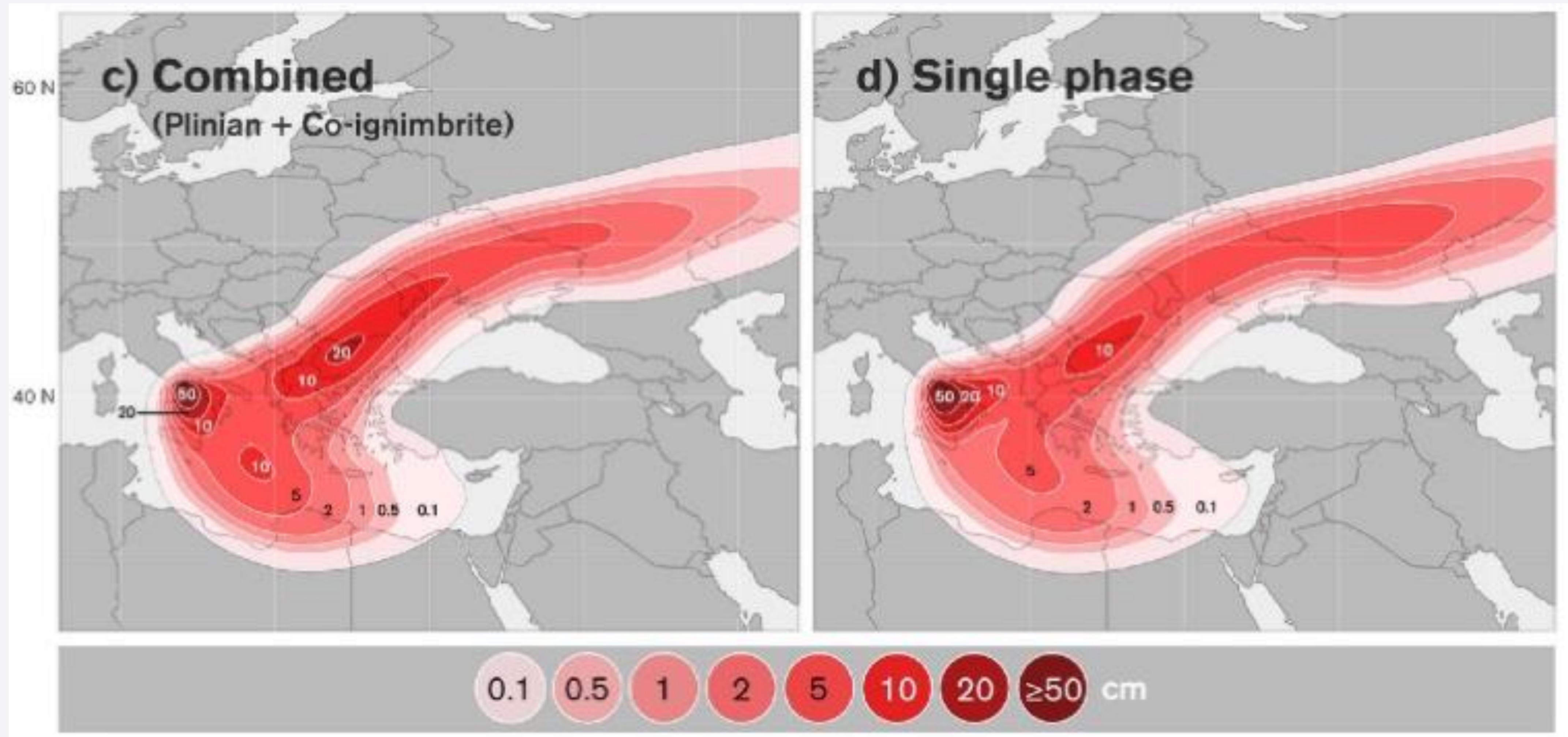


Efectividad

Discriminabilidad

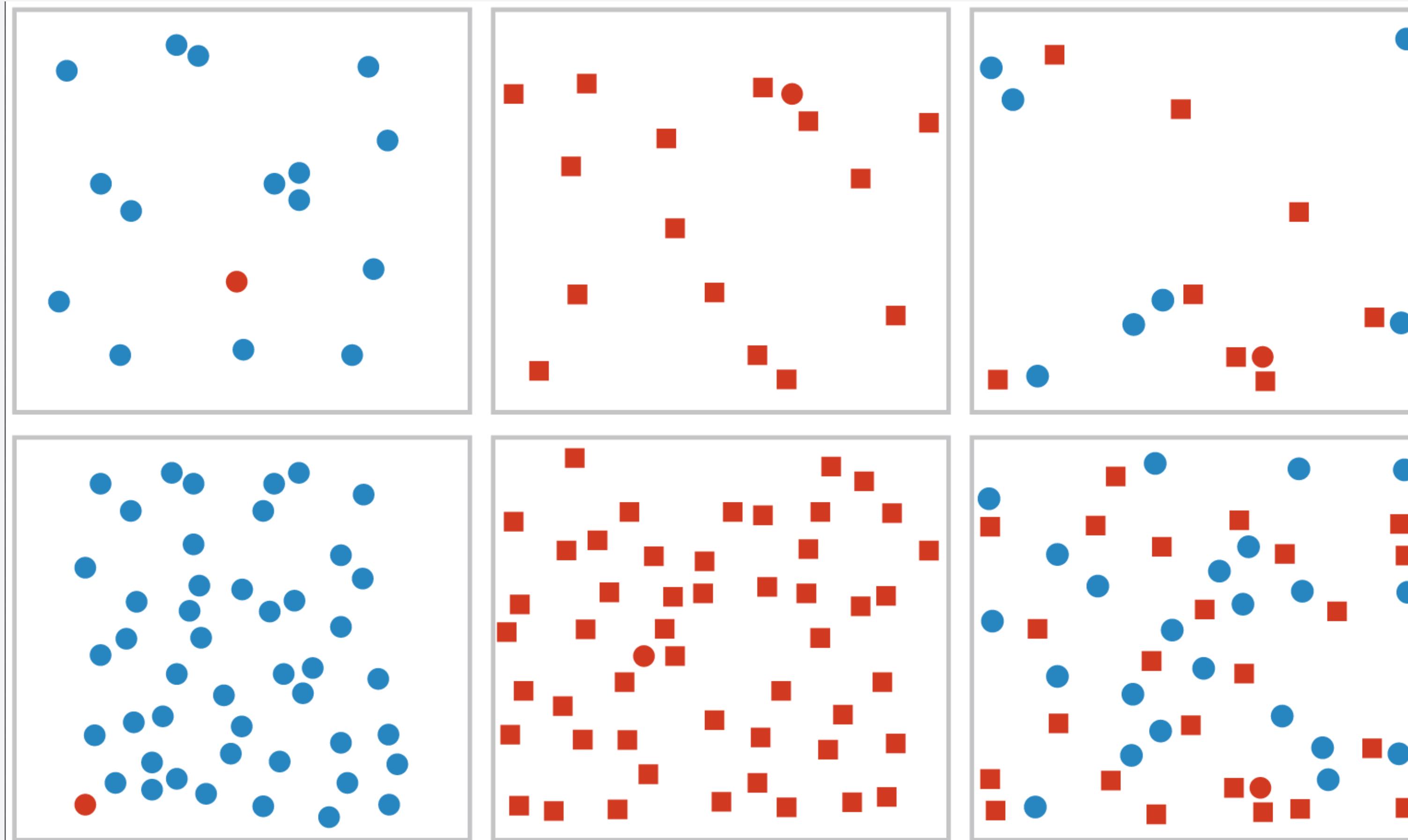
Se aplica también a canales de magnitud no-espaciales como saturación y luminancia

La suele ser más fácilmente discernible en un número limitado y bajo de bins



Efectividad

Saliente (Popout)

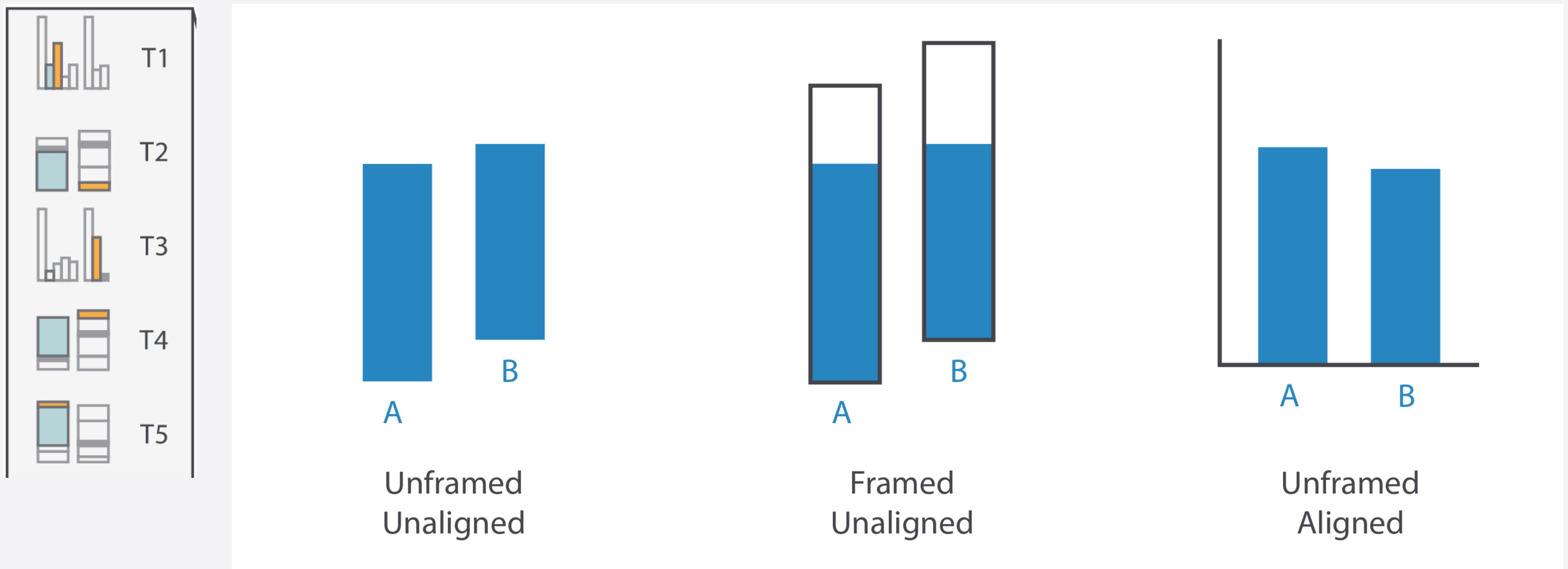


- Un ítem sobresale entre otros inmediatamente.
- El tiempo que nos lleva detectarlos no depende del número de distractores, sino del tipo.
- El popout está modulado por el canal en si mismo y cuan diferente sea el ítem de lo que lo rodea.

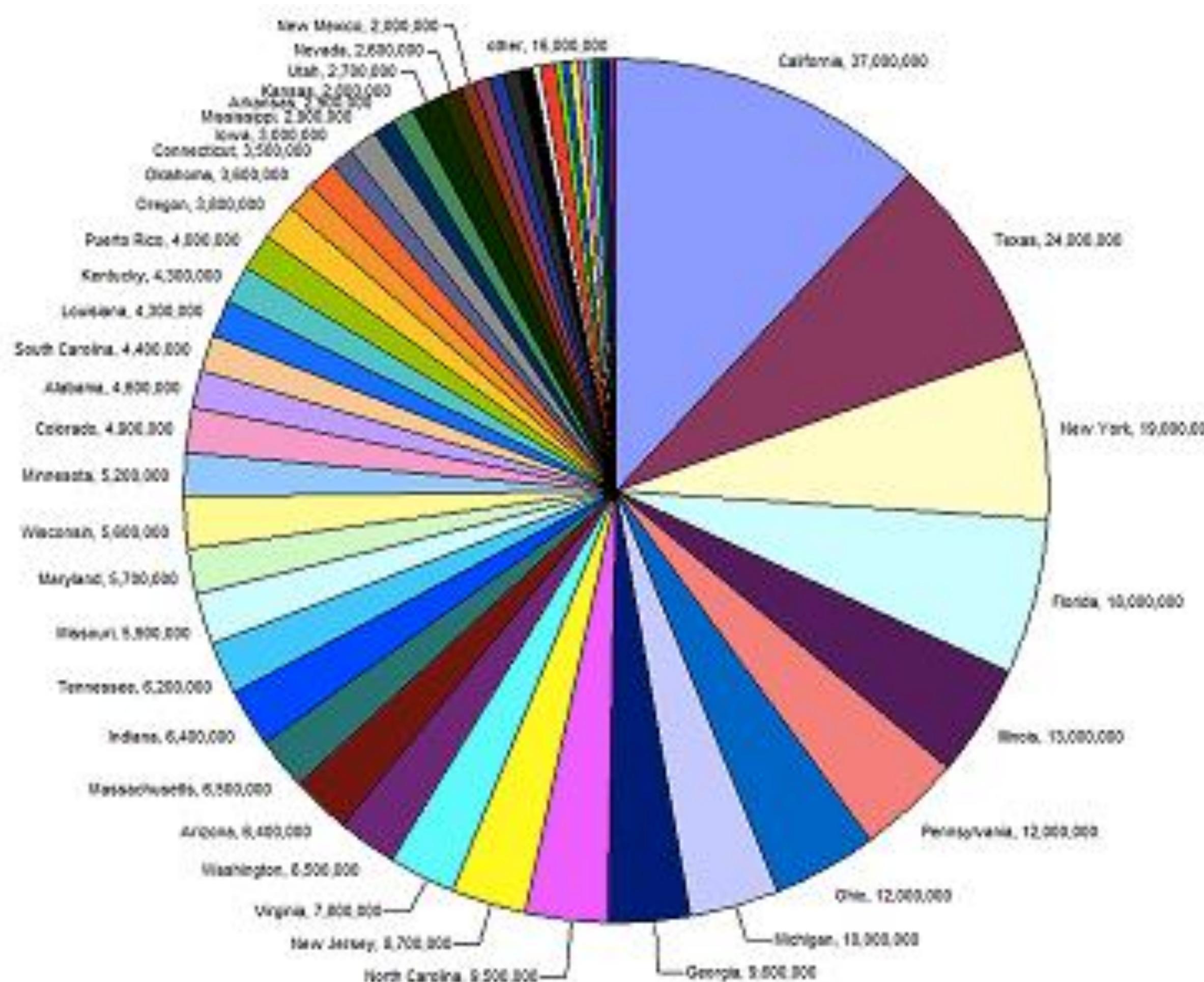
Efectividad

Relatividad

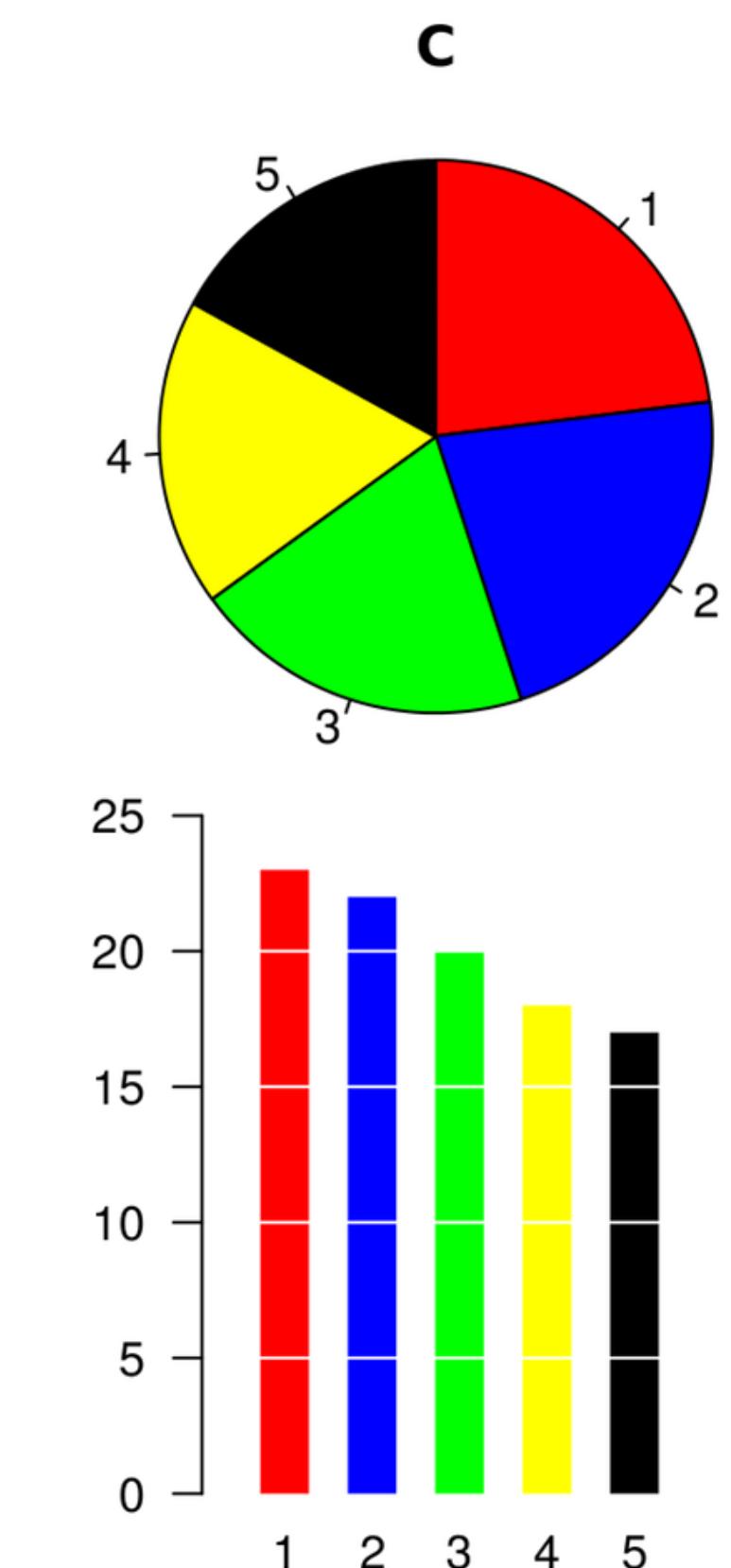
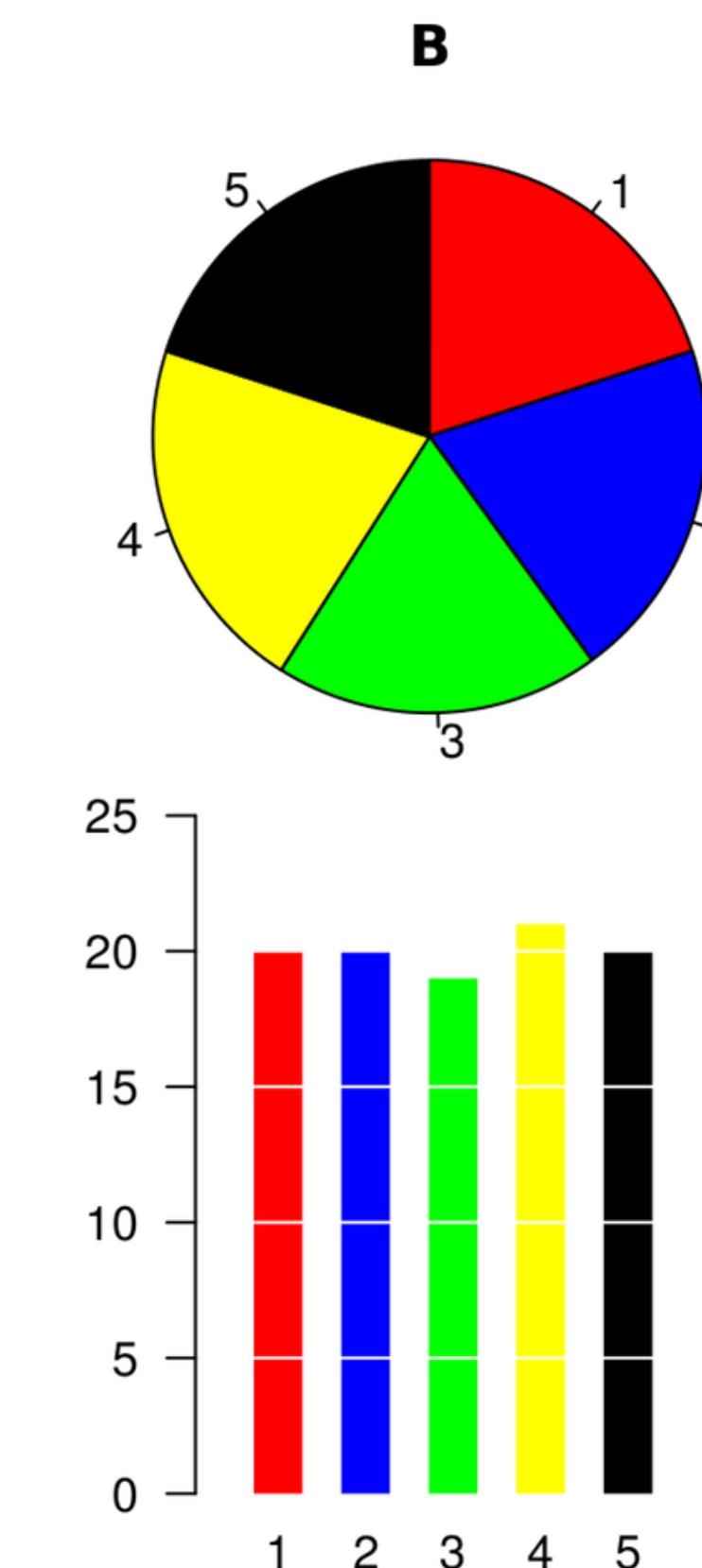
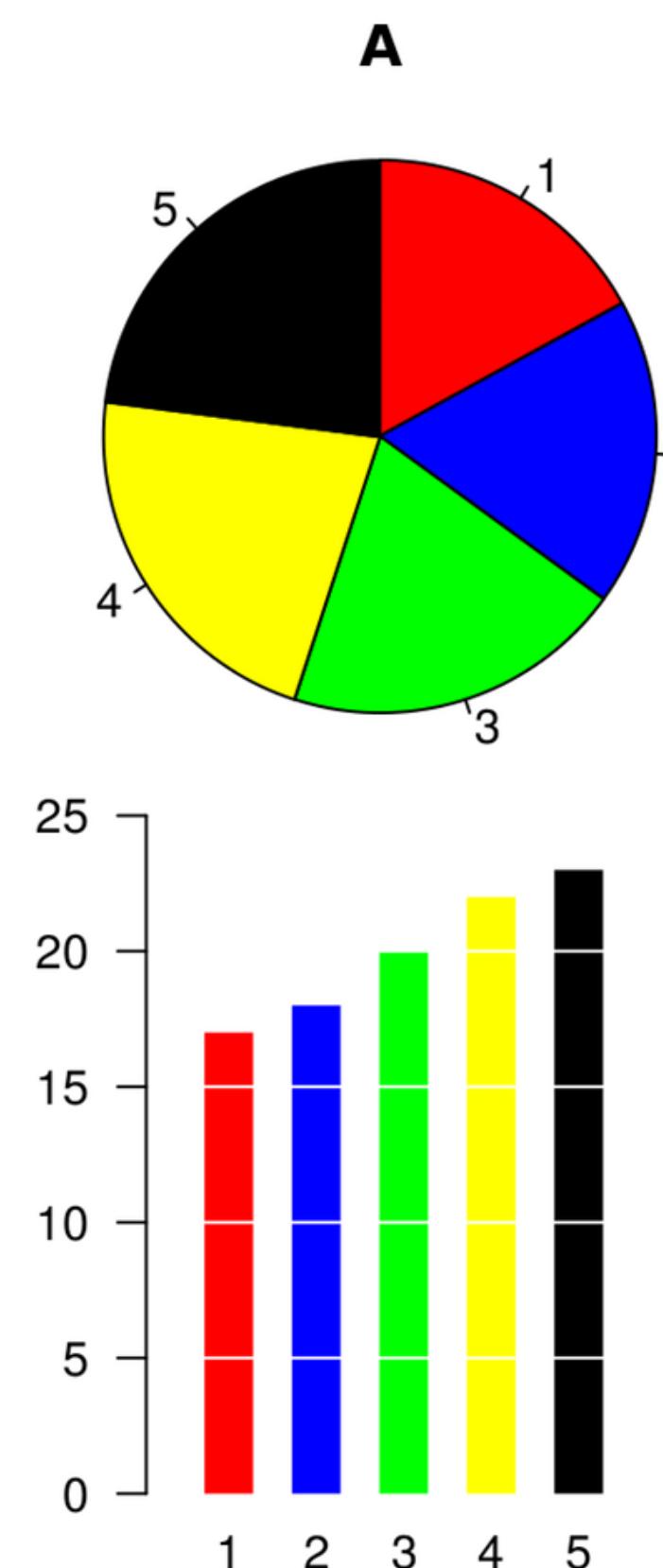
- Nuestro Sistema perceptual se basa fundamentalmente en juicios relativos, no absolutos (Ley de Weber)
- El contexto que rodea a los elementos modula la aplicación de **Discriminabilidad y Precisión**



Ejemplo: piecharts

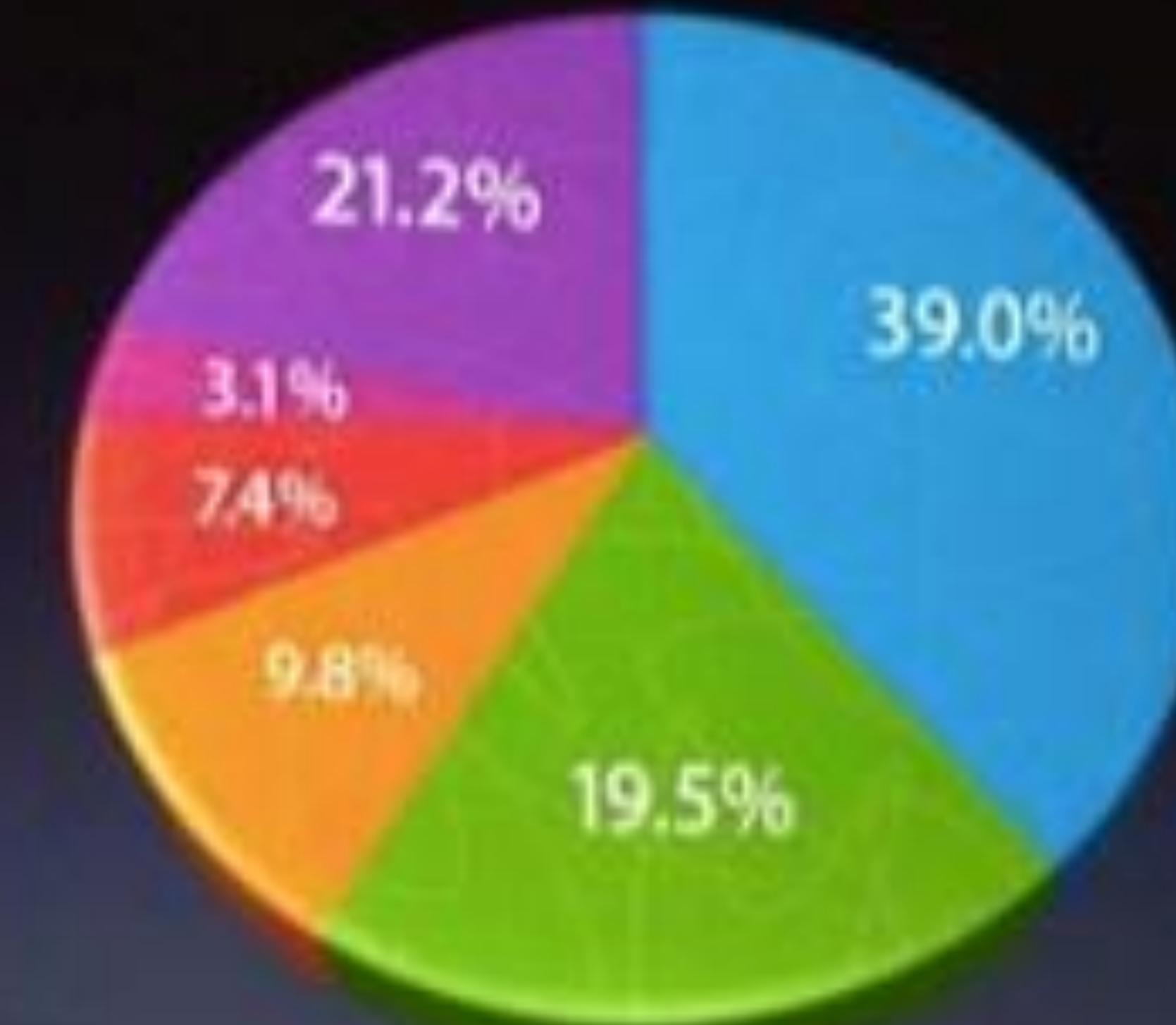


No funciona bien con muchas categorías,
ni para comparar valores con precisión



U.S. SmartPhone Marketshare

- RIM
- Apple
- Palm
- Motorola
- Nokia
- Other



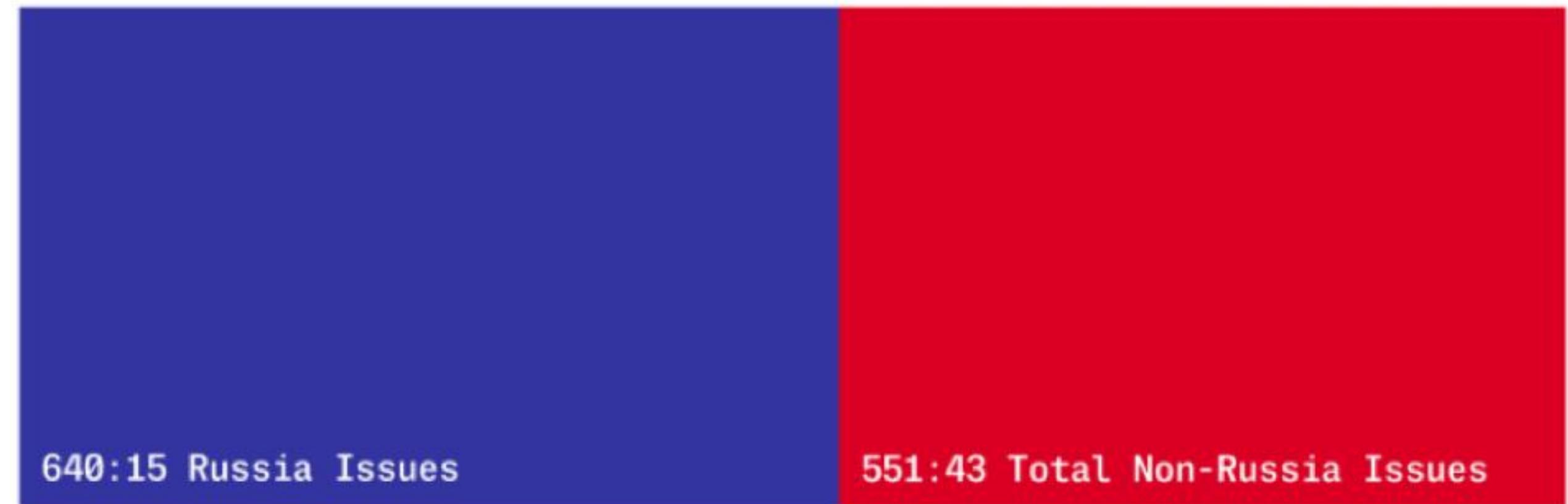
Gartner Inc.

No usar piecharts **siempre**

RUSSIA ISSUES VS. NON-RUSSIA ISSUES

February 20–March 31, 2017

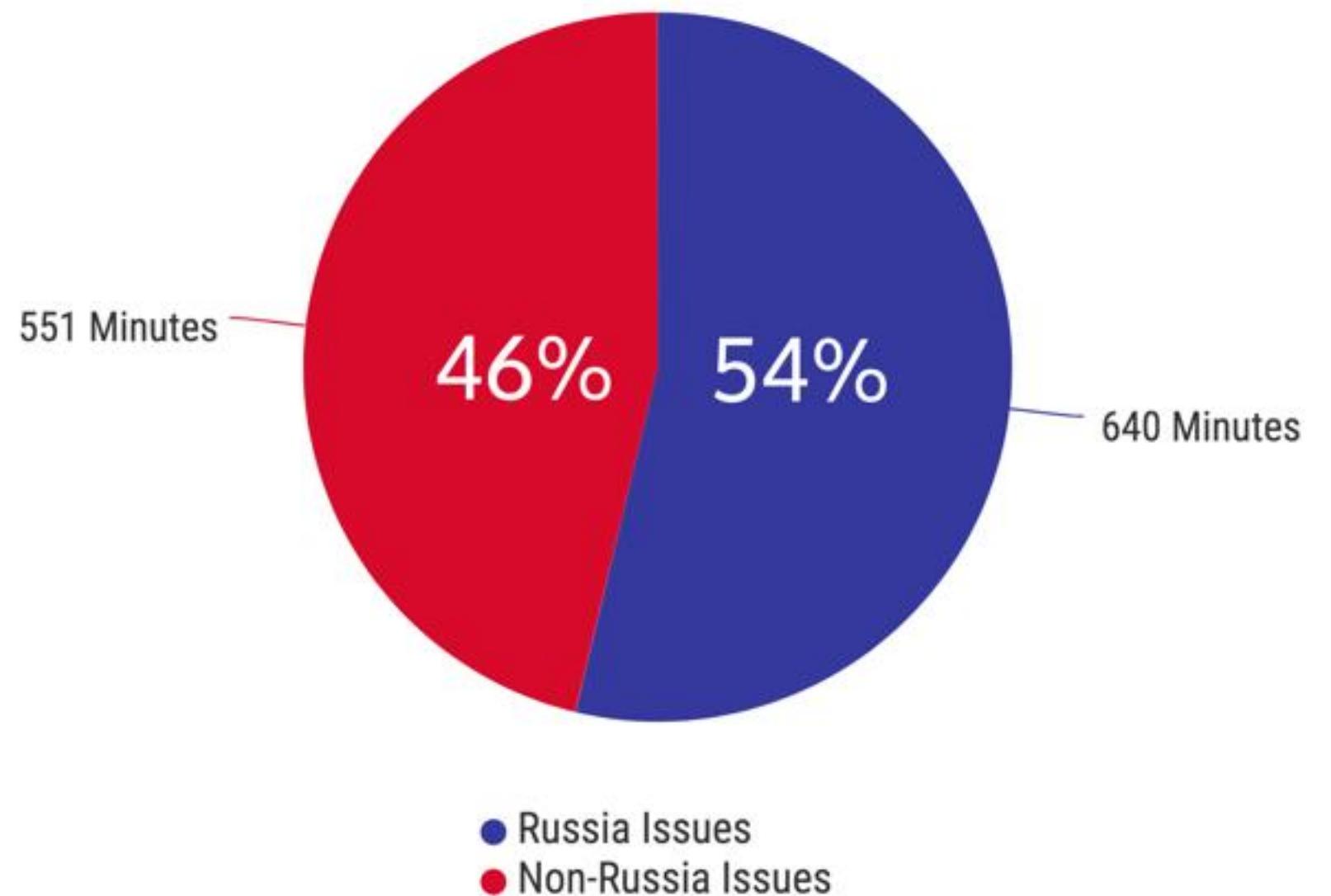
1191:58 (min:sec) Total Show Minutes



Russia issues vs. Non-Russia issues. Chart: The Intercept

from venngage.com

Russia Issues Vs Non-Russia Issues



How Music Preferences Have Changed in Two Decades

Music styles preferred by University of Miami students.

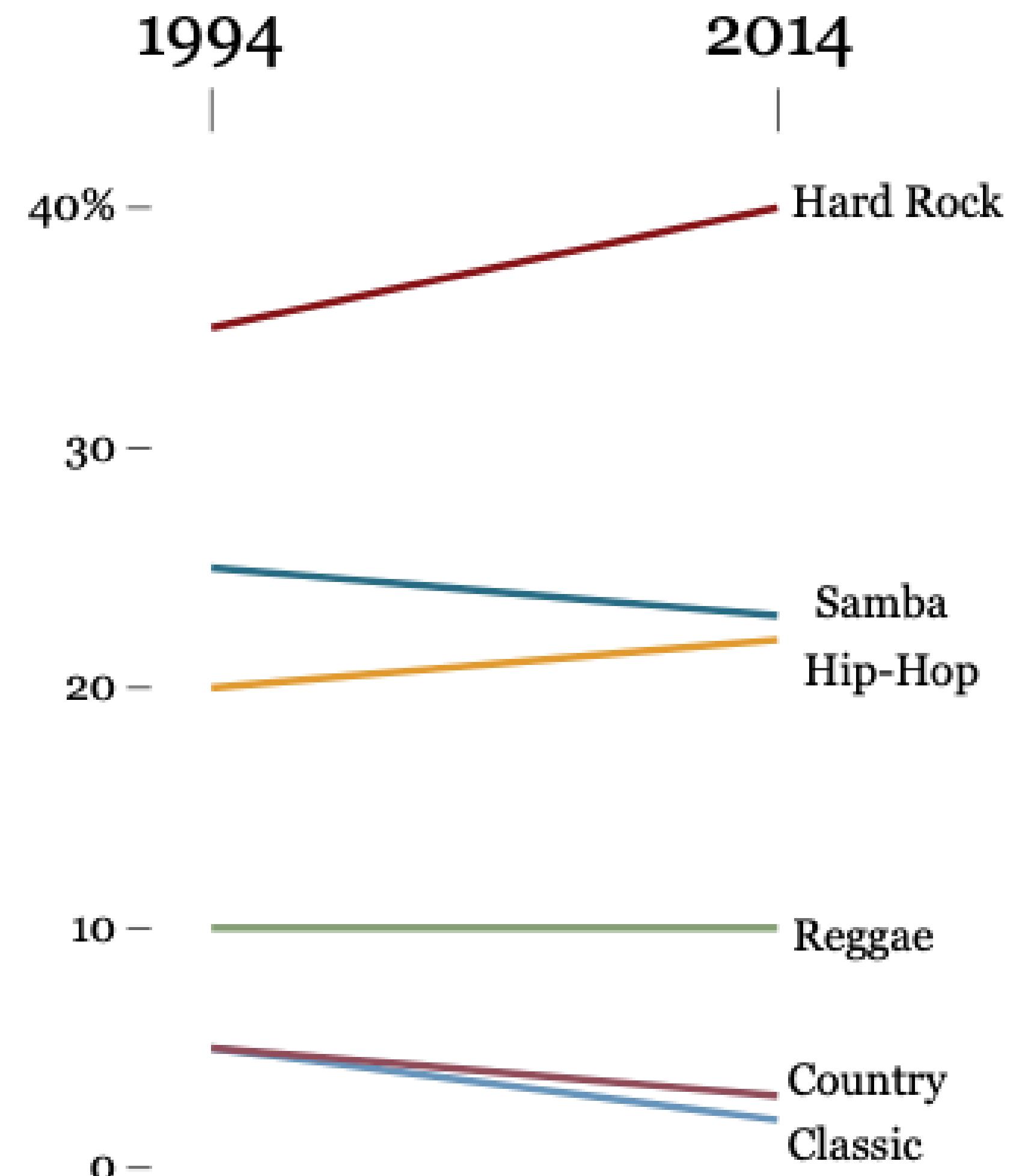
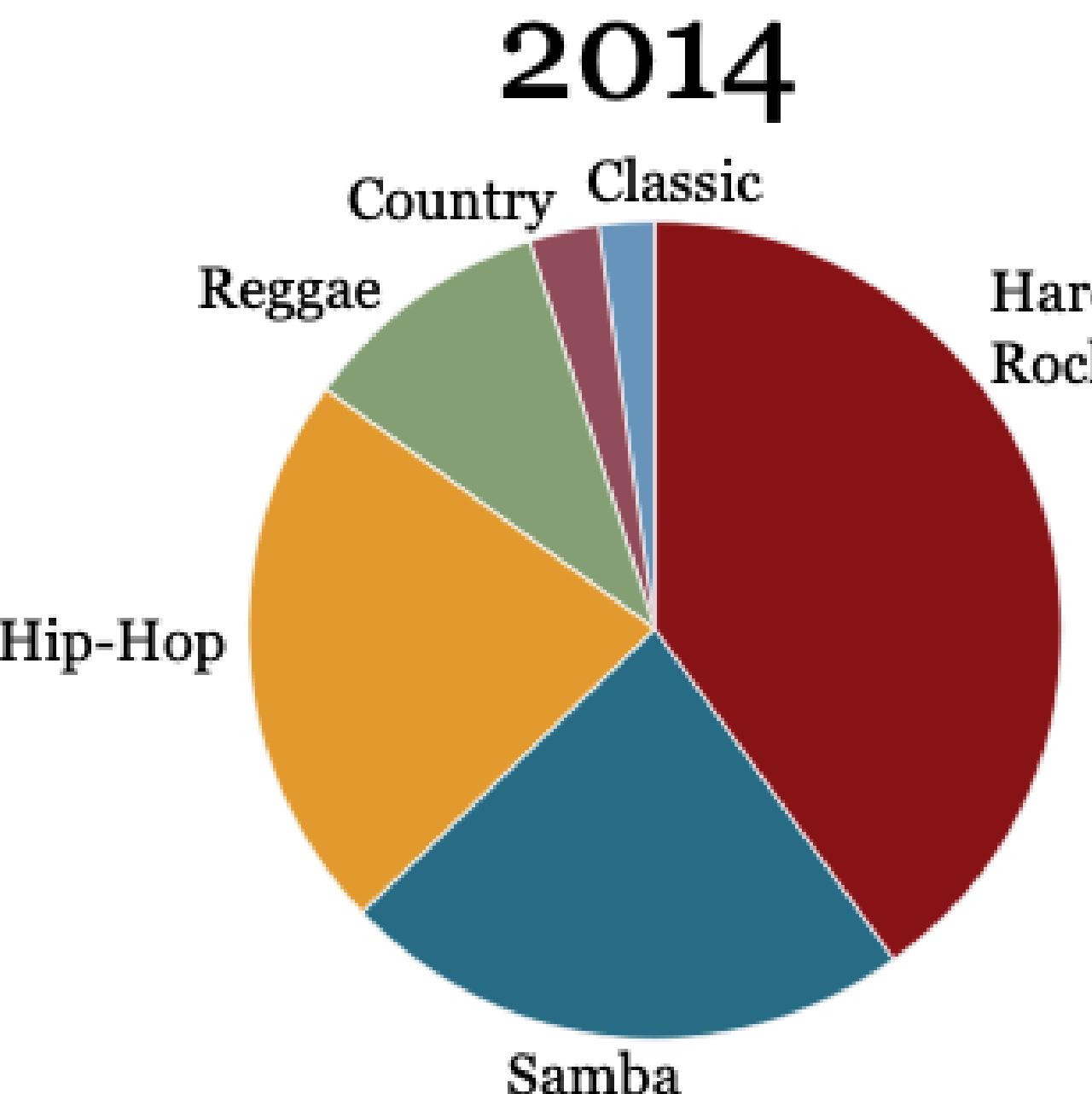
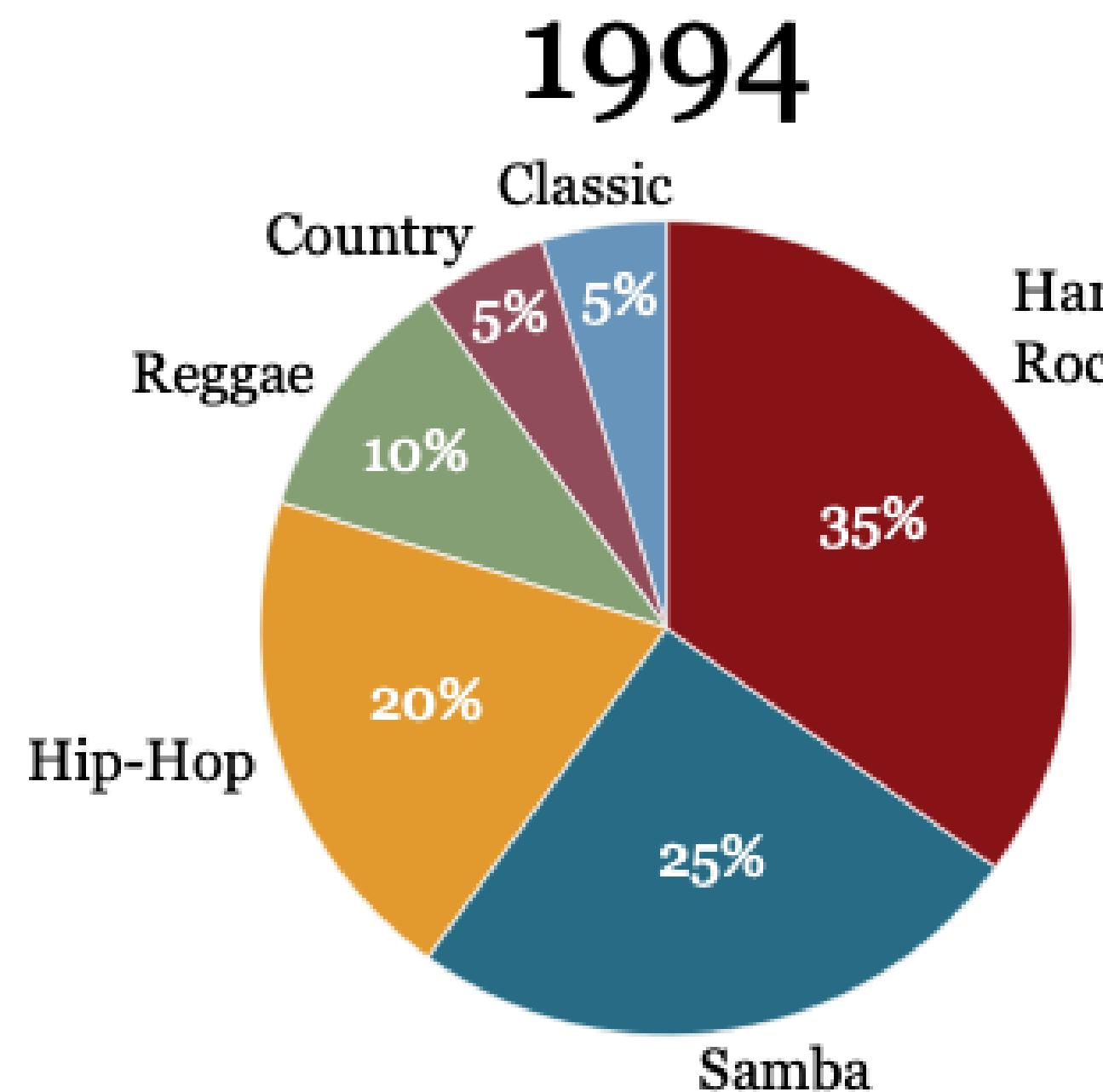
Survey based on interviews with 1,000 students.

SOURCE: WishfulThinkingData Inc.

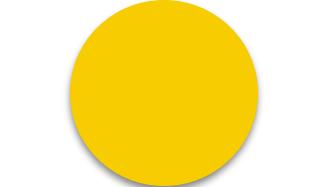
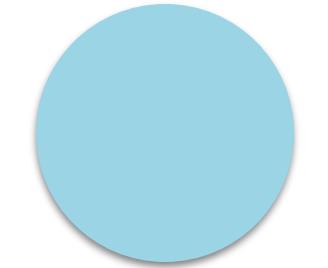
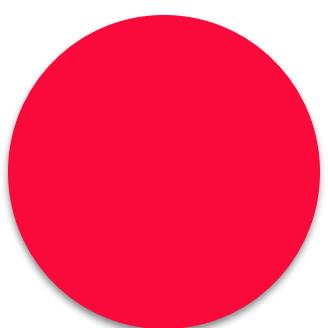
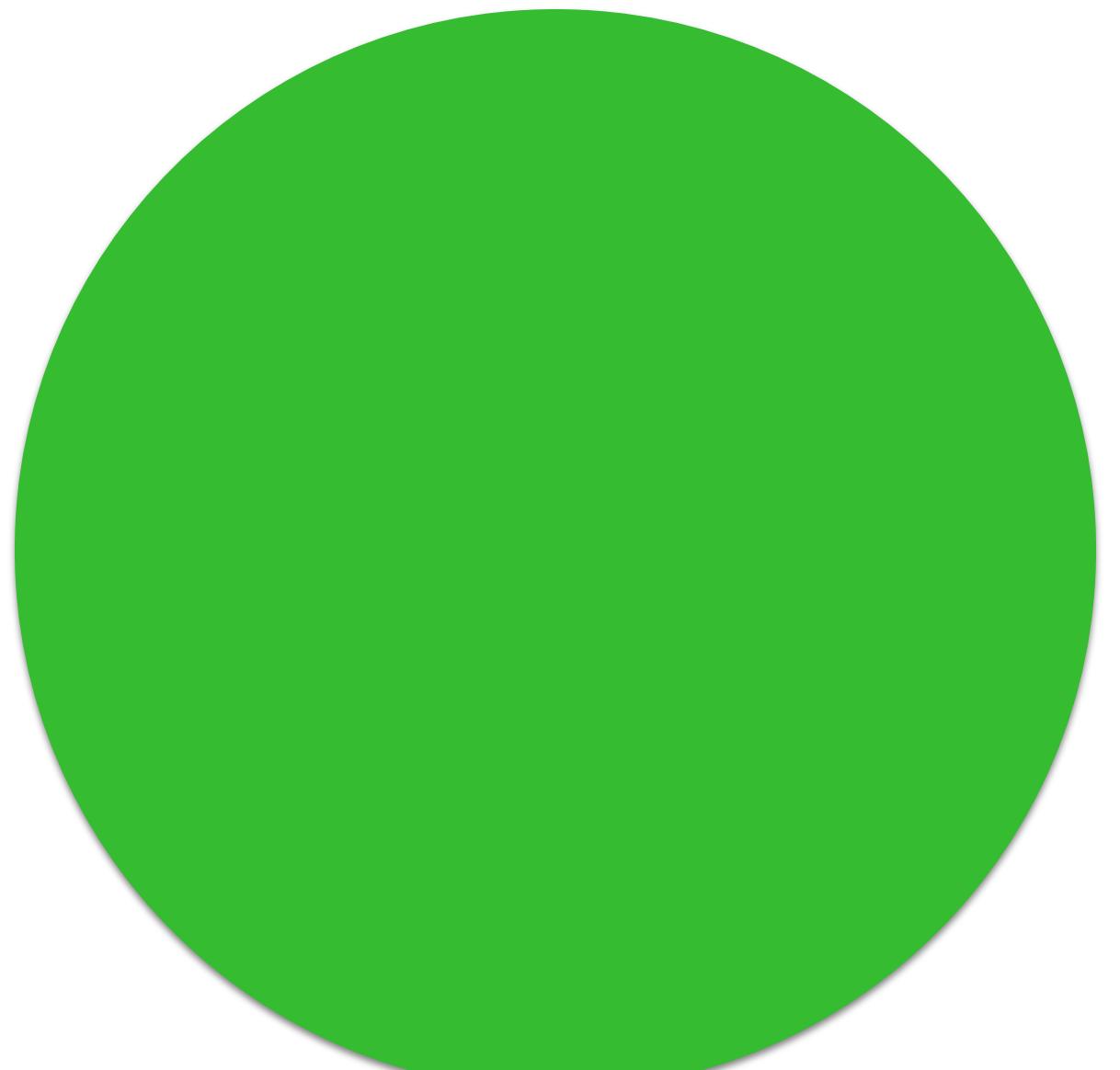
How Music Preferences Have Changed in Two Decades

Music styles preferred by University of Miami students. Survey based on interviews with 1,000 students.

SOURCE: WishfulThinkingData Inc.



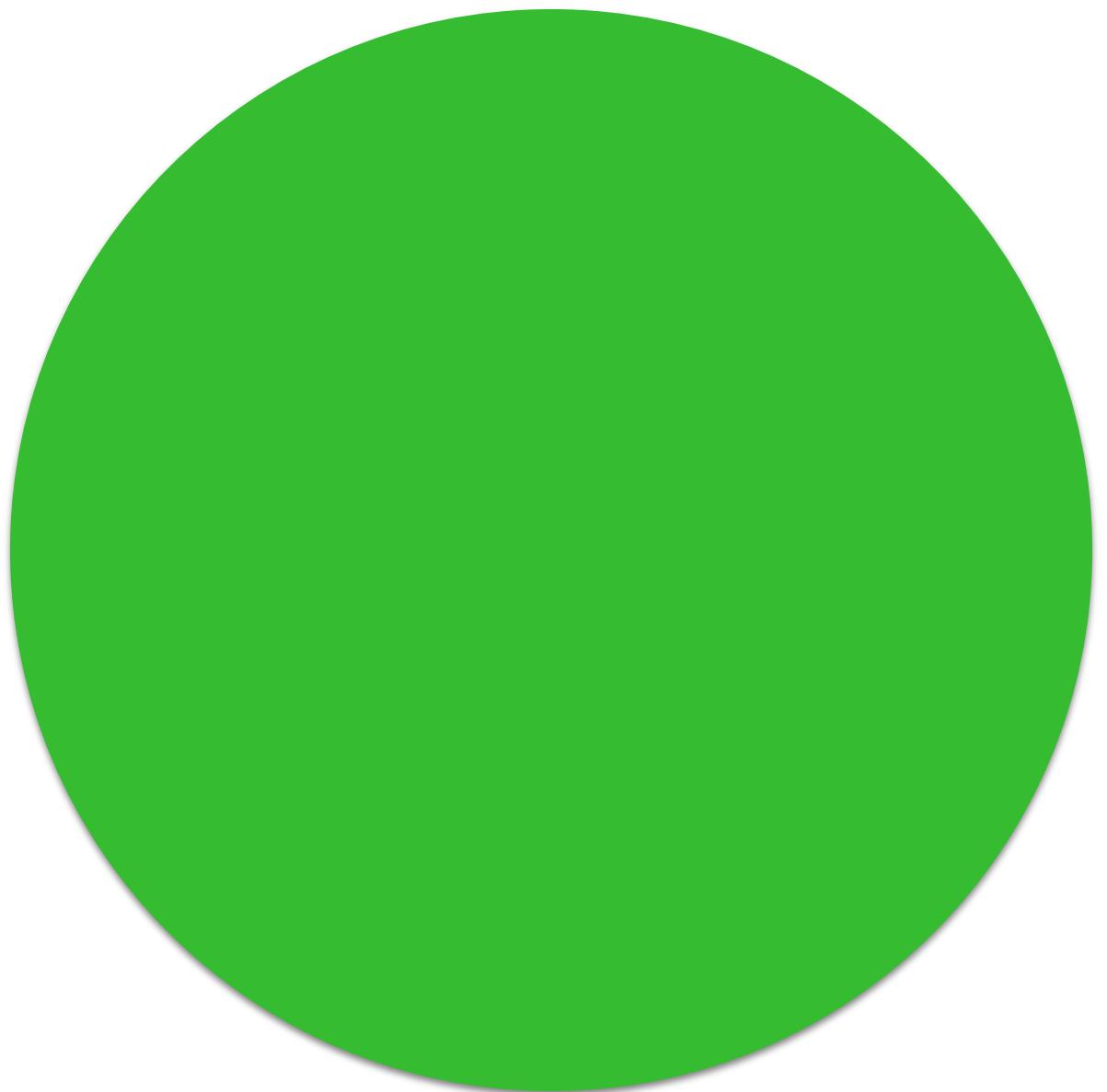
Ejemplo: Bubble charts



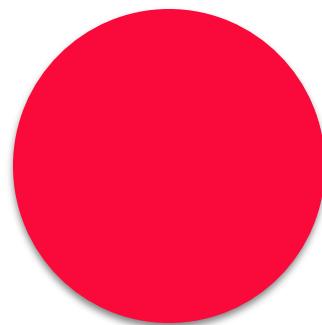
juntas. Así lo ilustra el gráfico informativo de Marcus Lu, de Visual Capitalist, que recoge las reservas probadas de cada empresa.



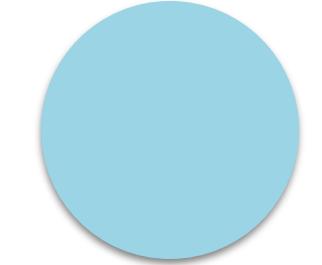
Ejemplo: Bubble charts



$$A=259 \rightarrow r = 9,08$$

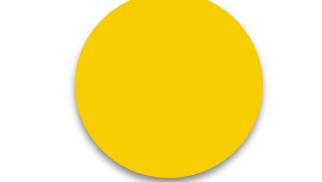


$$A = 18 \rightarrow r = 2,4$$



11

$$A = 7 \rightarrow r = 1,5$$

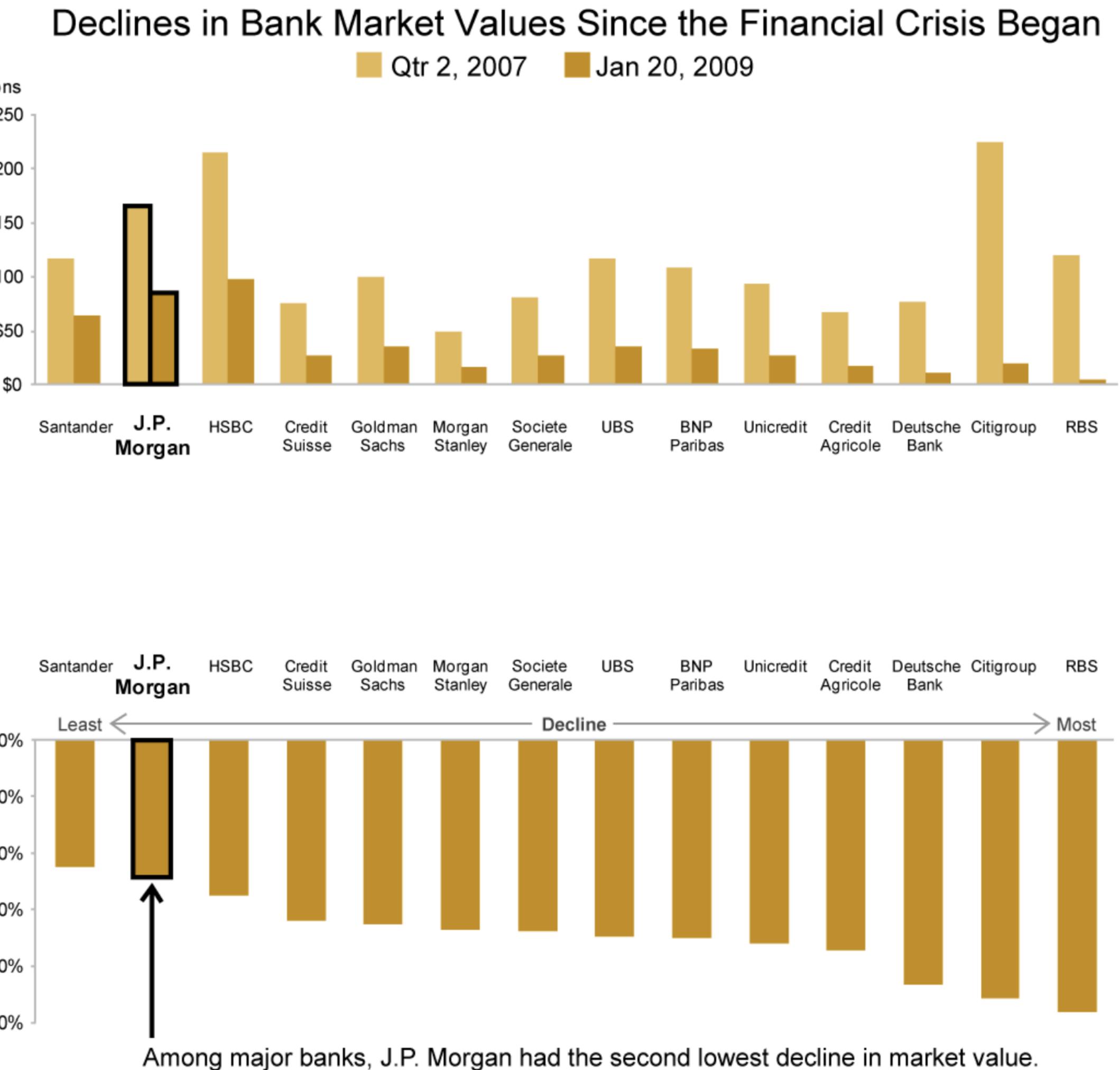
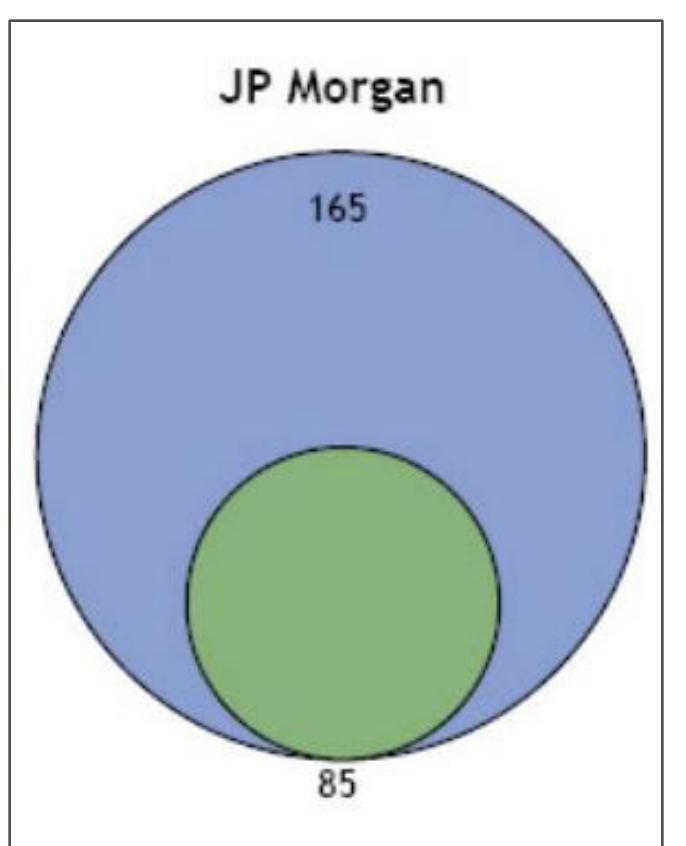
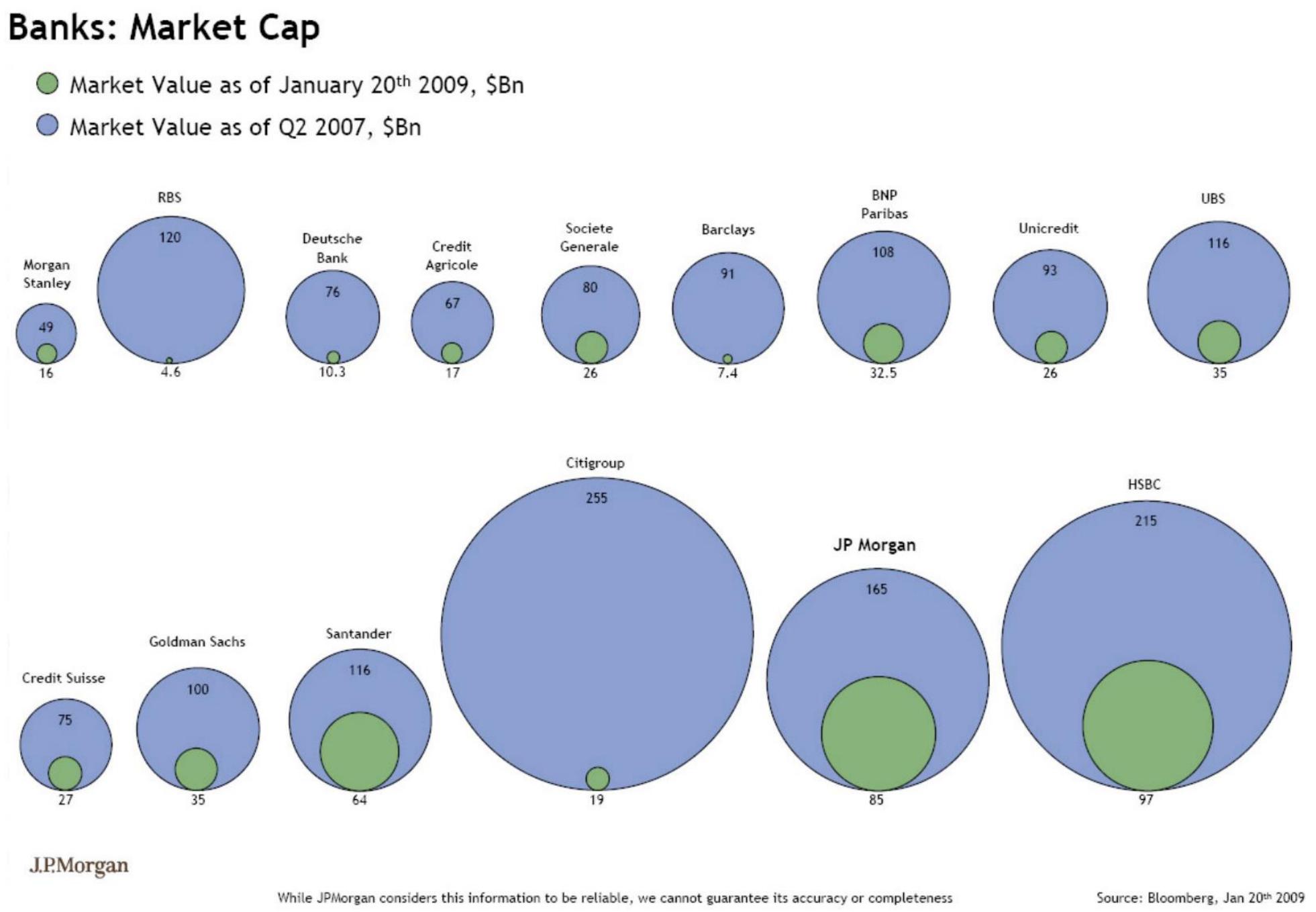


7

juntas. Así lo ilustra el gráfico informativo de Marcus Lu, de Visual Capitalist, que recoge las reservas probadas de cada empresa.

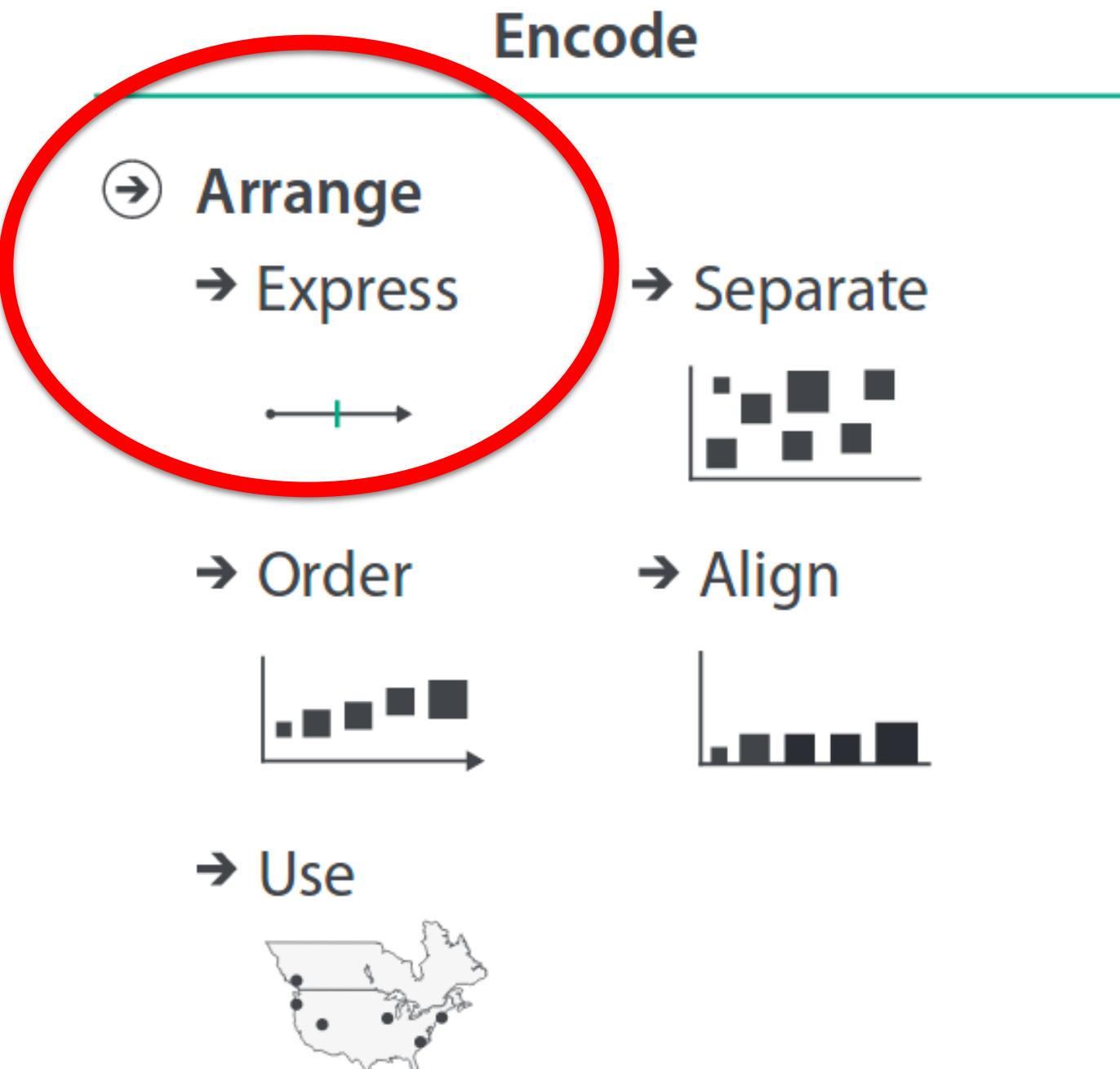
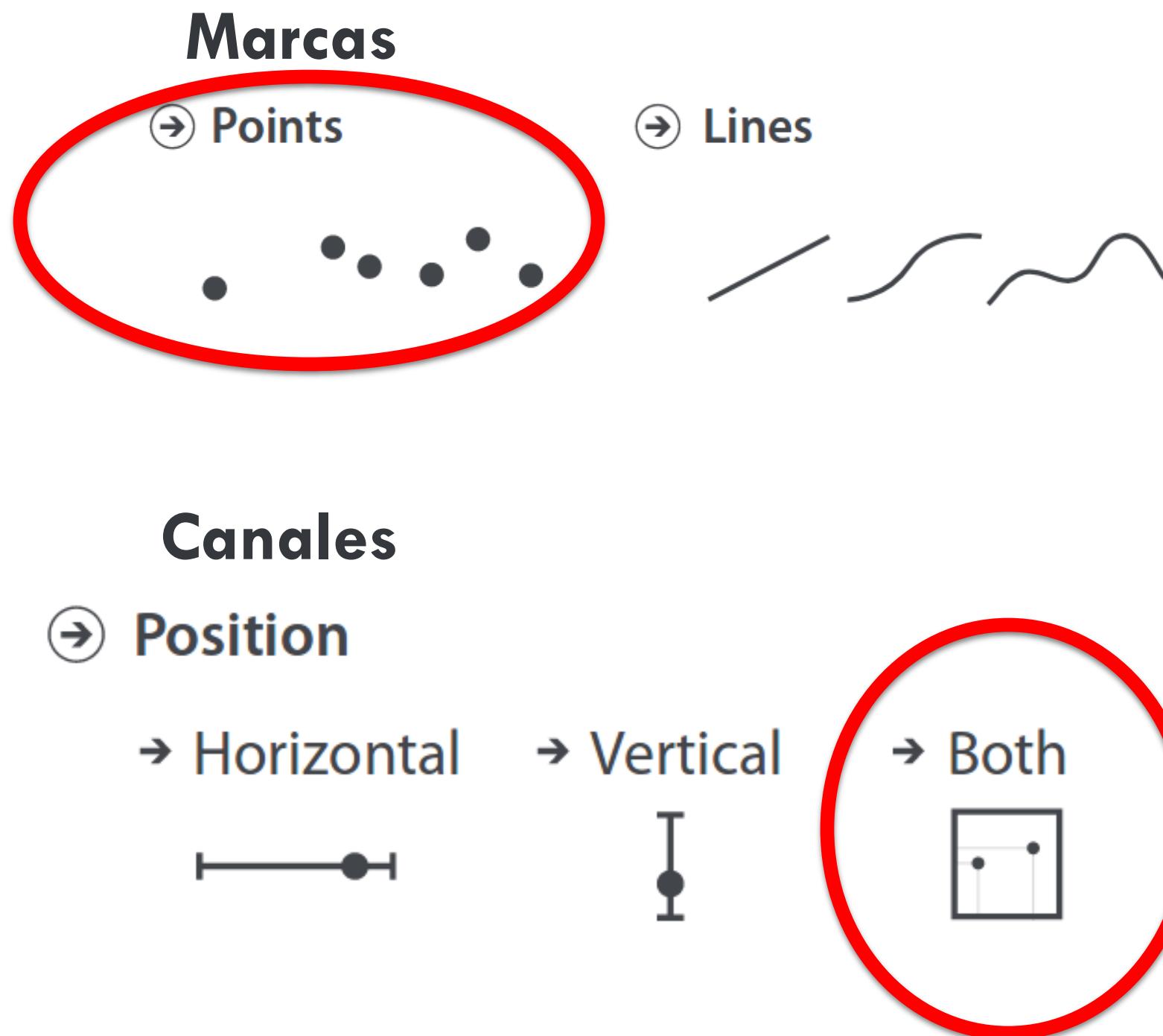
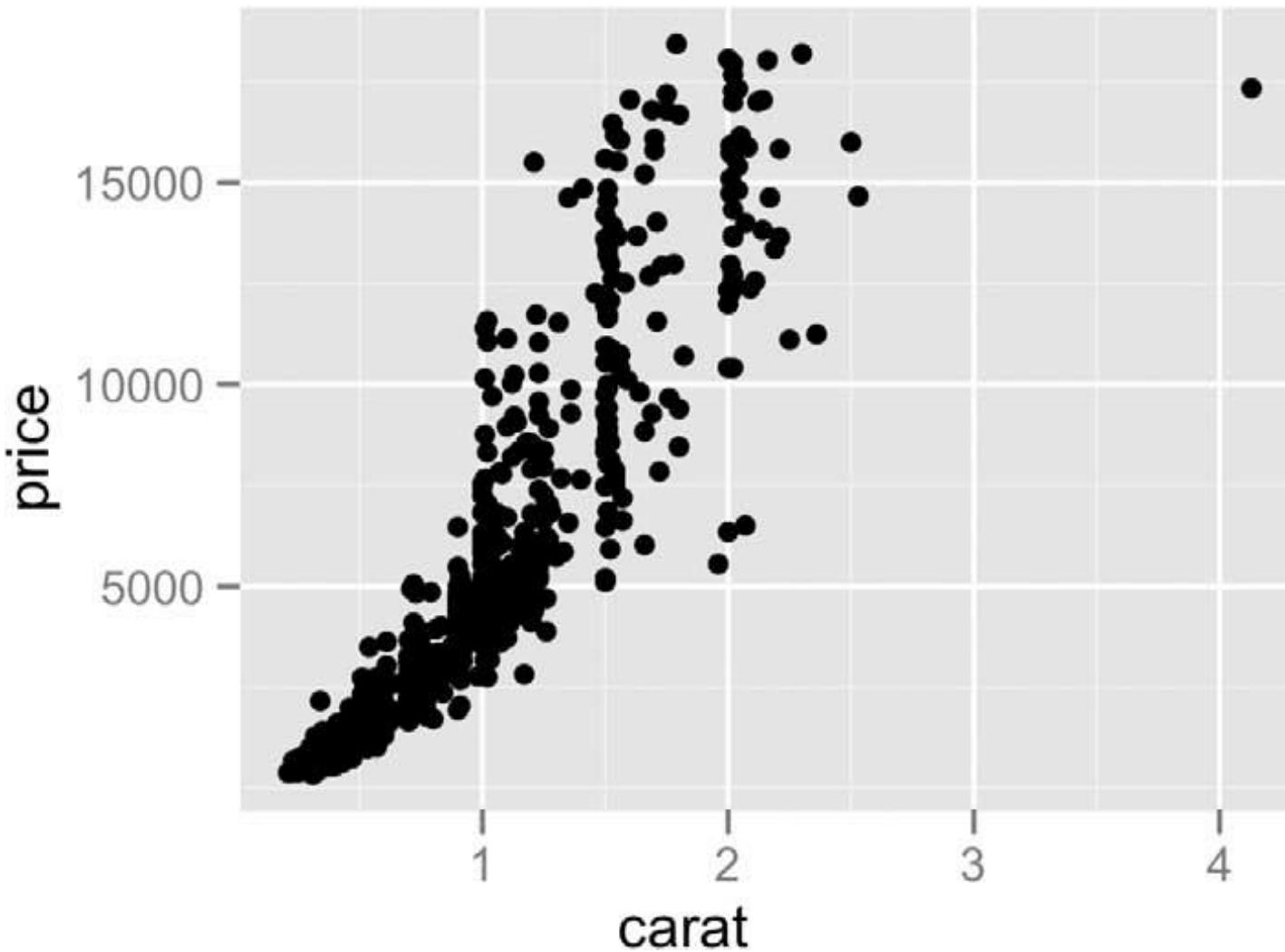


Ejemplo: Bubble charts



Our Irresistible Fascination with All Things Circular. Few, S. 2010

Scatterplot



- **Expresa** 2 atributos cuantitativos
- **Marcas:** Puntos // **Canales:** Posición X Y
- **Tareas:**
 - Encontrar patrones y tendencias
 - Analizar distribución
 - Identificar outliers, clusters, etc.

EJERCICIO 1

- Abrir *life_expectancy.csv* o *.xlsx* (Excel, Preview de mac, GoogleSheets, Python, etc).
- Es necesario **formatear** antes de usarlo.
- Derivar dataset filtrando **un solo año**.

- Hacer gráficas para visualizar las preguntas:
 - ¿Top 20 países con mayor EV media?
 - ¿Top 20 países con mayor diferencia por sexo?

- Identificar tipos de datos: cuantitativos, ordinales o categóricos.
- Abrir la web www.datawrapper.de
- Hacer gráficas según el planteamiento anterior de Datos, Tareas y Codificación.
 - Ranking ordenado de Both Sexes por país.
 - ¿Qué gráfica mostraría más claramente la diferencia entre ambos sexos por país?

EJERCICIO 2

- Abrir *life_expectancy.csv* o *.xlsx* (Excel, Preview de mac, GoogleSheets, Python, etc). **Necesario formatear antes de usar.**
- **Queremos visualizar la evolución temporal de Both sexes por país, para todos los países a la vez.**
- ¿Qué gráfica usamos?
- ¿Cómo formatear los datos para hacer esa visualización?
- **Haz las operaciones necesarias para transformar los datos**
- Abrir www.datawrapper.de
 - ¿Qué gráfica sería más adecuada?
 - Modifica los parámetros necesarios para hacerla más comprensible y fácil de entender.

Bibliografía

- Visualization Analysis and Design. Munzner. Cap. 7
- **The Functional Art . Alberto Cairo, 2012 (Cap. 5-6)**

Inspiración

- <http://www.thefunctionalart.com/>
- <https://eagereyes.org/>
- <http://flowingdata.com/>
- <http://fivethirtyeight.com/>
- <http://truth-and-beauty.net/>



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



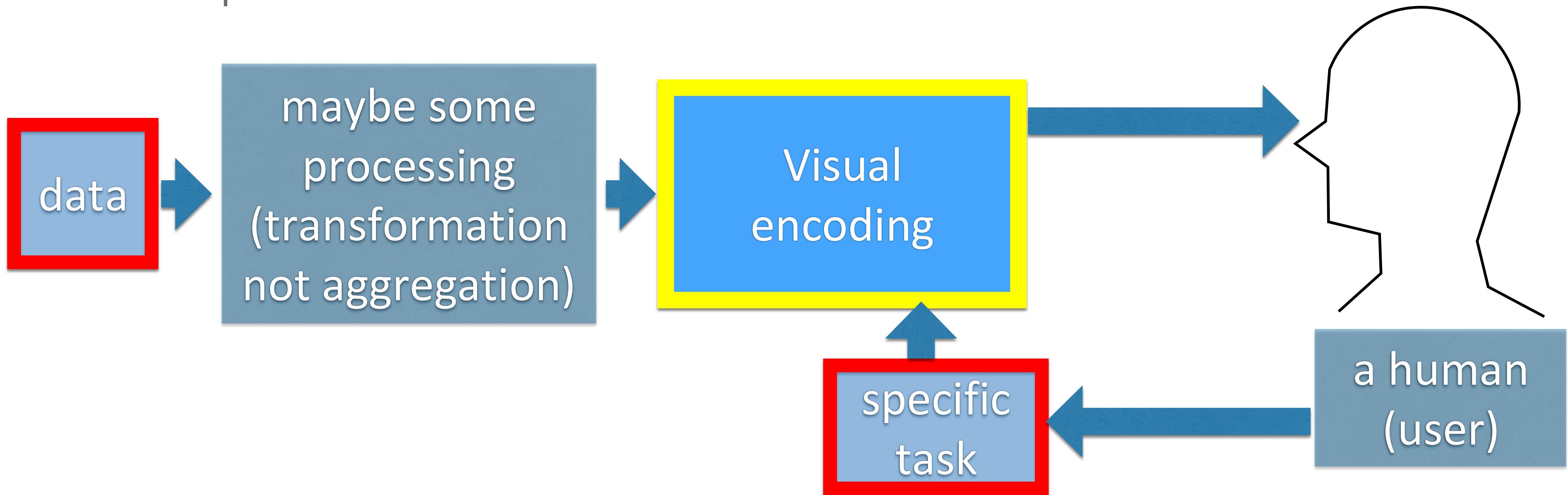
4. Visualización de Tablas

Sol Bucalo
sol.bucalo@uab.cat

Guillermo Marin
guillermo.marin@uab.cat

Data Visualisation

- Datos. El proceso empieza con uno o más datasets. Conocemos el tipo y las características de sus atributos.
- Tareas. Definición de las tareas que podemos resolver, caracterizadas como acción + objetivo



Tablas

What?

Datasets

Attributes

→ Data Types

→ Items → Attributes → Links → Positions → Grids

→ Attribute Types

→ Categorical



→ Data and Dataset Types

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

→ Ordered

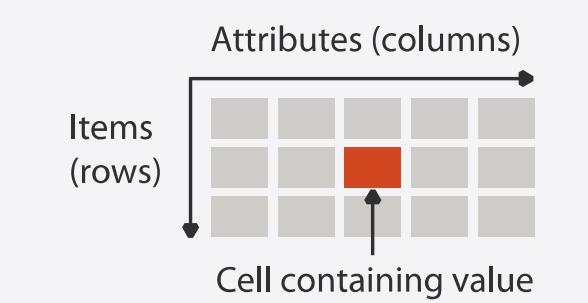


→ Quantitative

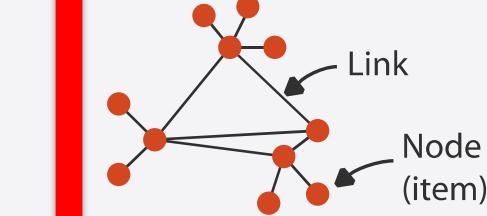


→ Dataset Types

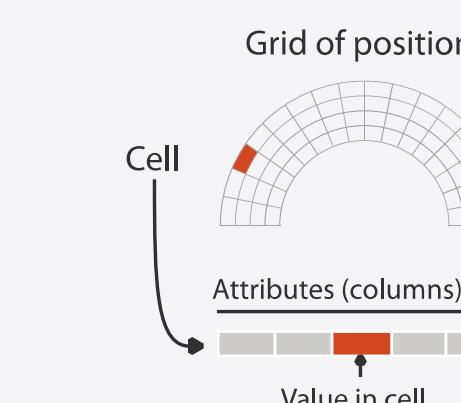
→ Tables



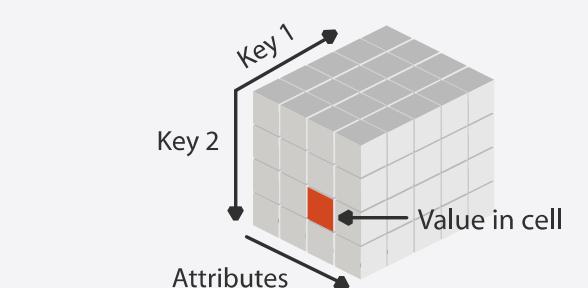
→ Networks



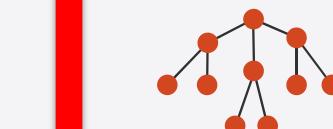
→ Fields (Continuous)



→ Multidimensional Table



→ Trees



→ Geometry (Spatial)



How?

Encode

→ Arrange

→ Express



→ Separate



→ Order



→ Align

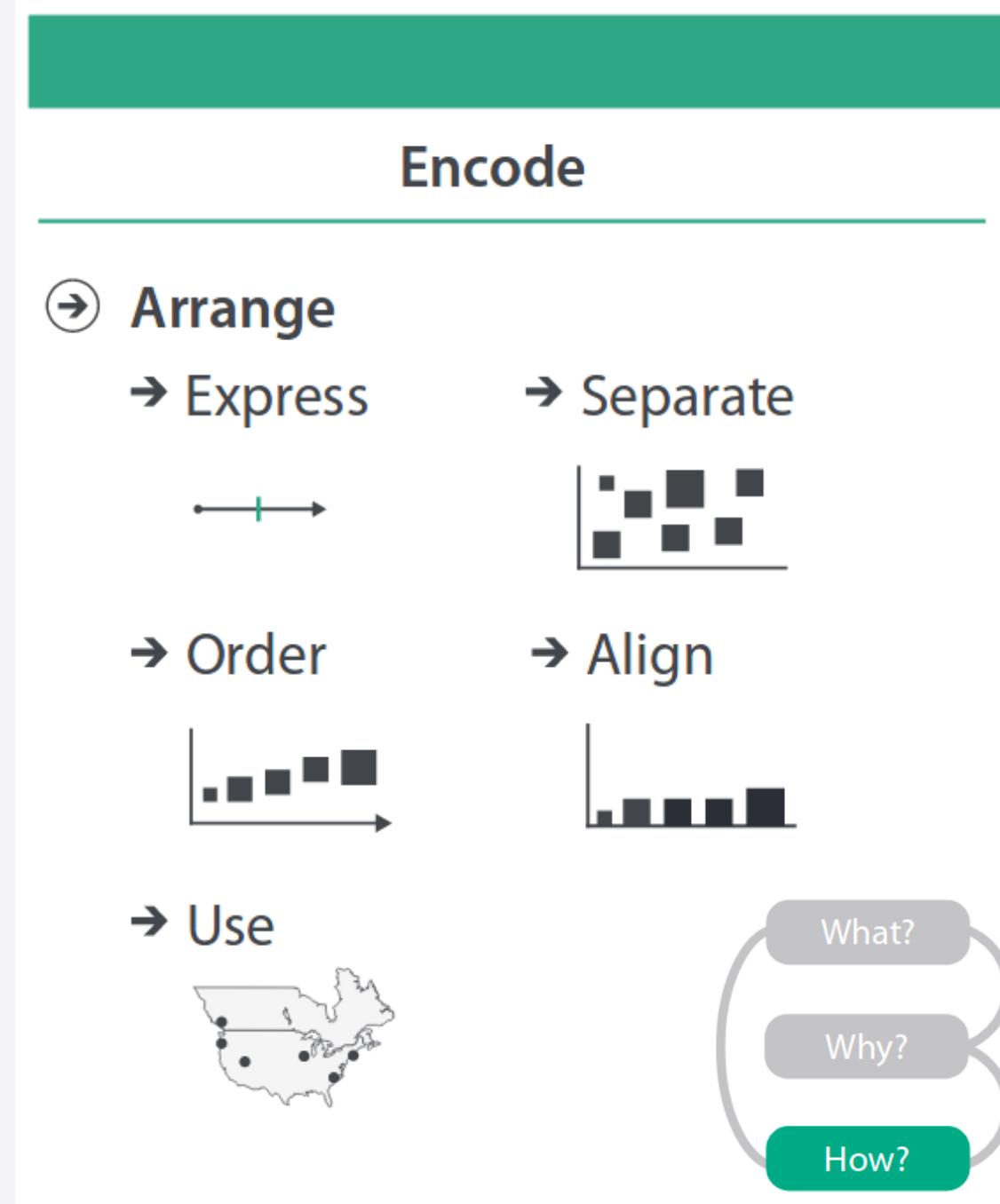


→ Use



Tablas

How?



Channels: Expressiveness Types And Effectiveness Ranks

→ Magnitude Channels: Ordered Attributes

Position on common scale



Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



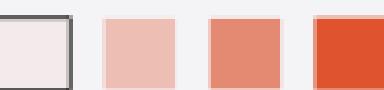
Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



→ Identity Channels: Categorical Attributes

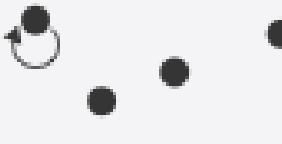
Spatial region



Color hue



Motion



Shape



Best ↑

Effectiveness

Least ↓

Same] [Same] [Least

- Codificación visual para tablas suele centrarse en **expresar el contenido**
- Las decisiones más importantes son la de **Arrange**
- El espacio domina el modelo mental del dataset del usuario.
- Los canales más efectivos son de posición en el espacio.

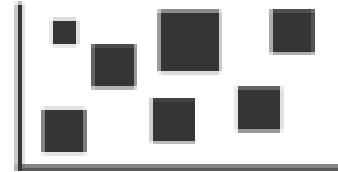
Arrange tables

④ Express Values

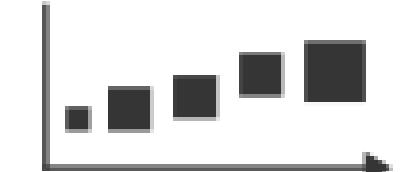


⑤ Separate, Order, Align Regions

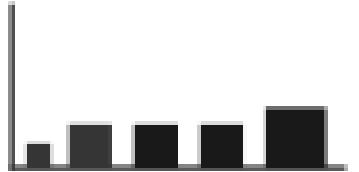
→ Separate



→ Order



→ Align



- Existen distintos tipos de tablas según el número de claves
- Key=Clave primaria = identificador único de cada fila/observación

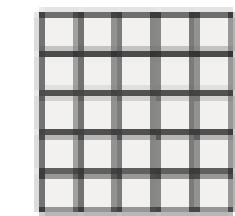
→ 1 Key

List



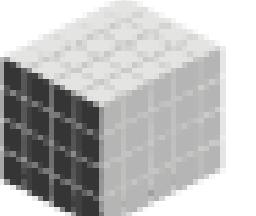
→ 2 Keys

Matrix



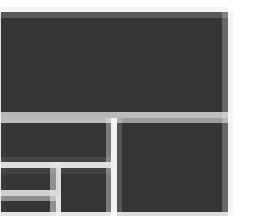
→ 3 Keys

Volume



→ Many Keys

Recursive Subdivision

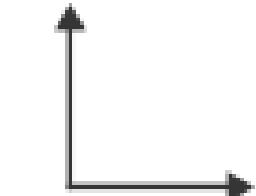


Para ordenar los datos también se tiene en cuenta:

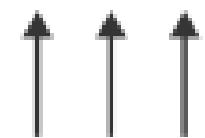
- Orientación de los ejes
- Densidad del Layout: Densa o exhaustiva

⑥ Axis Orientation

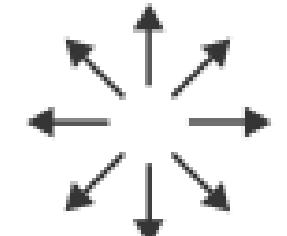
→ Rectilinear



→ Parallel

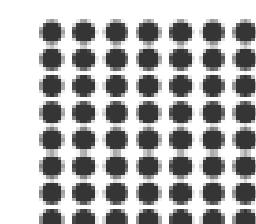


→ Radial

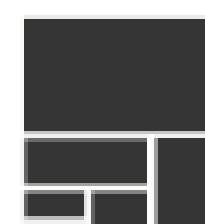


⑦ Layout Density

→ Dense



→ Space-Filling



Keys / Values

Key

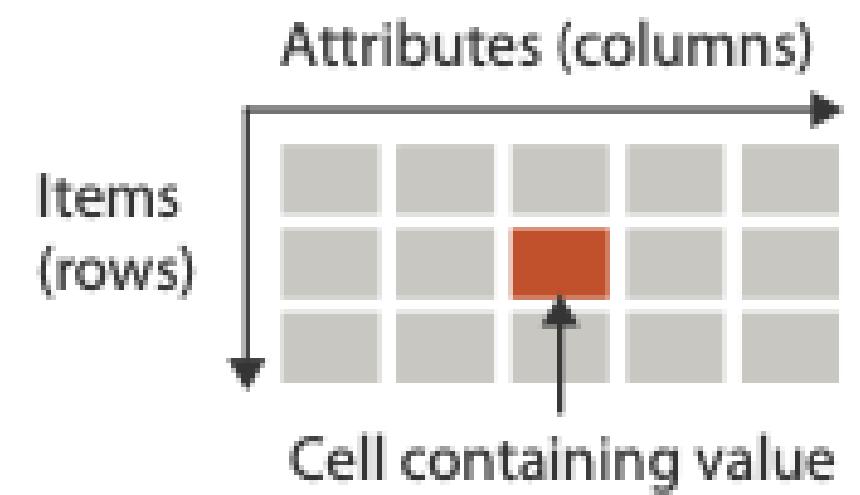
- Variable independiente
- Se puede usar como índice único para identificar ítems
- Tipo **categórico** u **ordinal**



Value

- Variable dependiente, valor de la celda
- Categórico, ordinal o cuantitativo

→ Tables



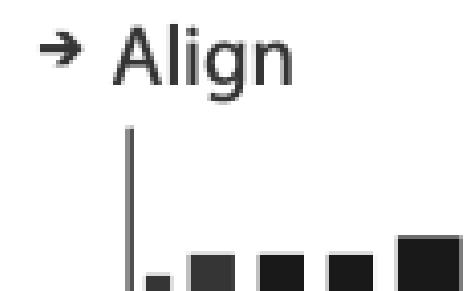
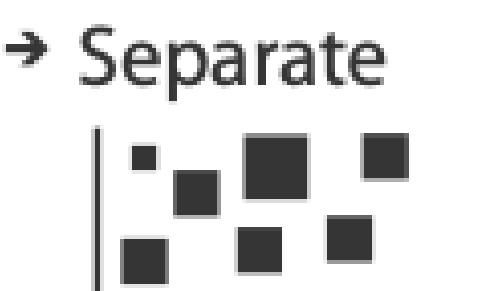
Decisiones básicas de diseño

- ¿Cuantos Keys? ¿cuantos Valores?
- ¿De qué tipo son (Categórico, ordinal, cuantitativo)?

→ Express Values



→ Separate, Order, Align Regions



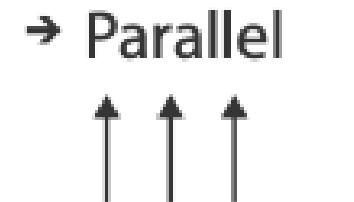
Scatterplot

- *Expresa 2 atributos cuantitativos*
- *Sin keys, solo values*
- *Marcas: Puntos*
- *Canales: Posición X Y*
- *Tareas:*
 - *Acciones: Encontrar patrones y tendencias // Analizar distribución // Identificar outliers, clusters.*
 - *Objetivos: Todo el dataset.*
- *Escalable a cientos de elementos*

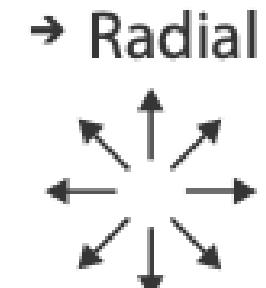
④ Axis Orientation

→ Rectilinear

→ Parallel

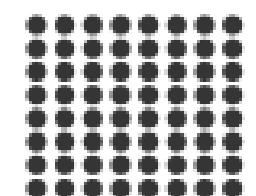


→ Radial

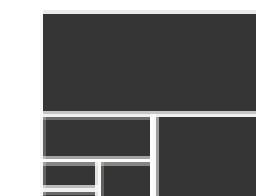


④ Layout Density

→ Dense



→ Space-Filling



Marcas

④ Points



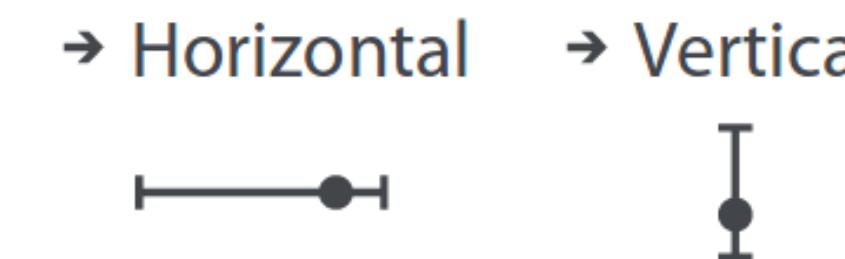
④ Lines



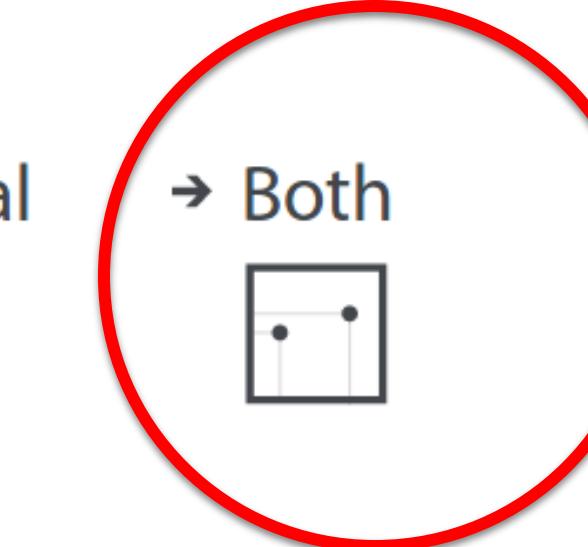
④ Areas



④ Position



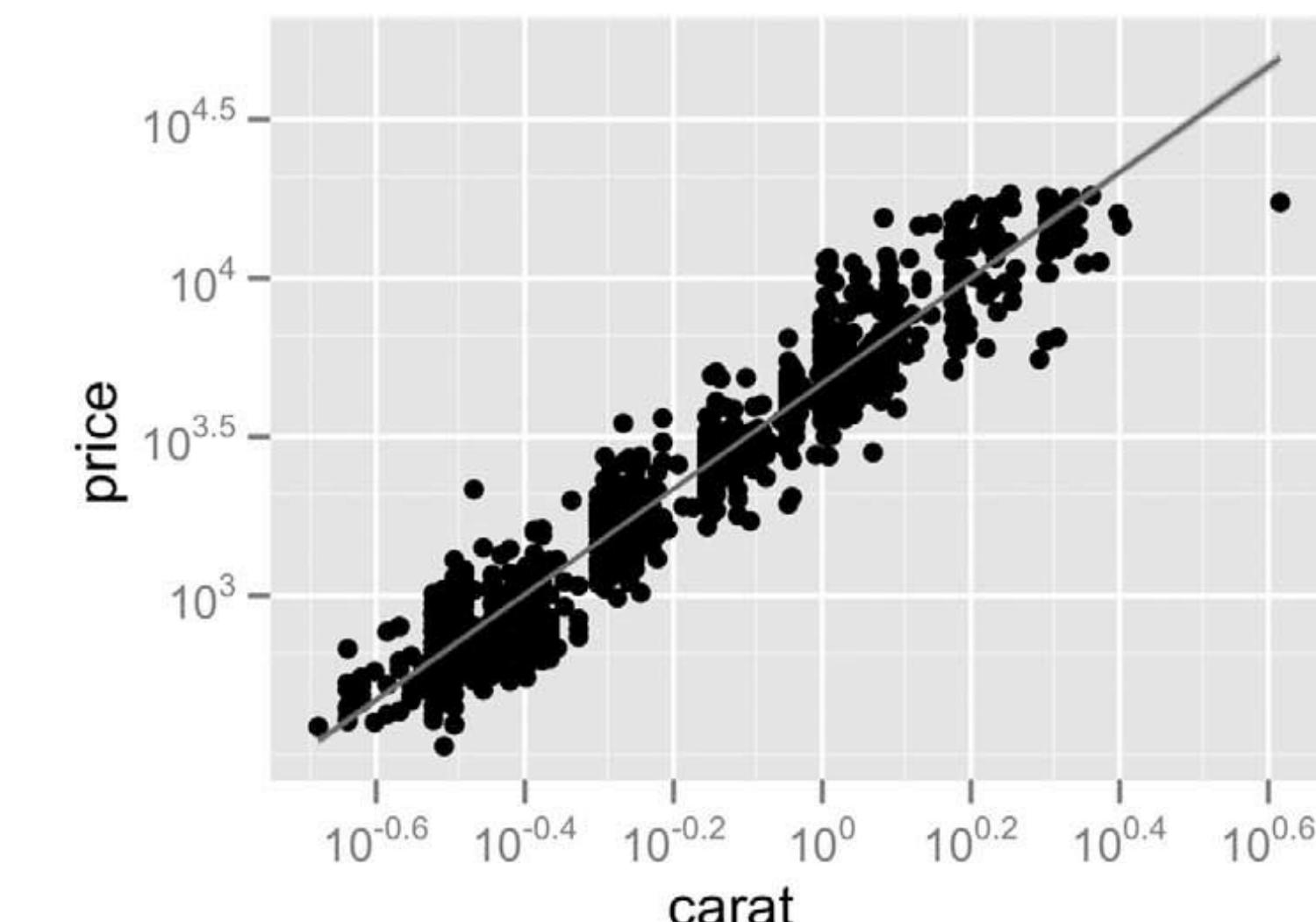
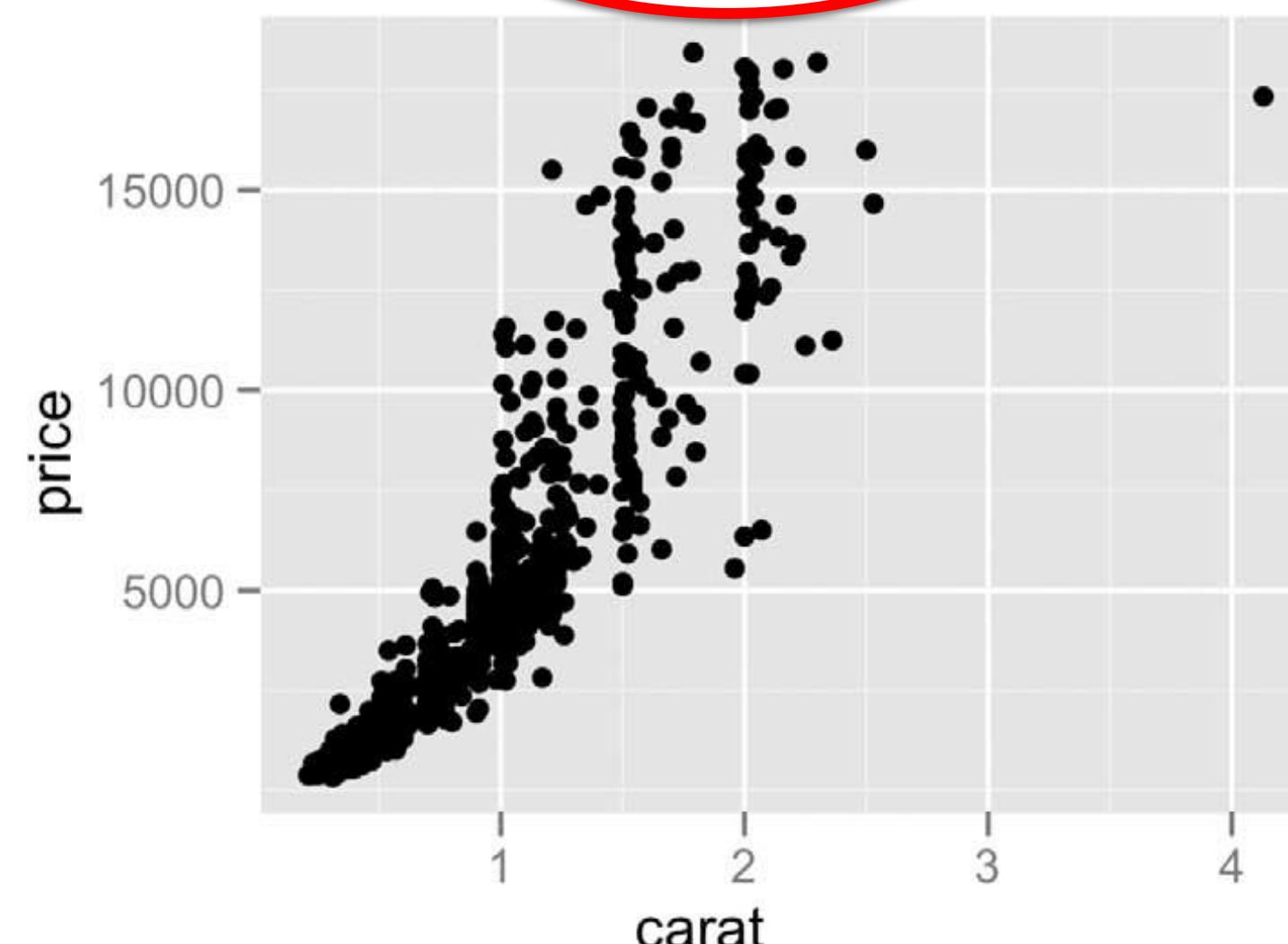
④ Both



Canales

→ Horizontal → Vertical

④ Express Values

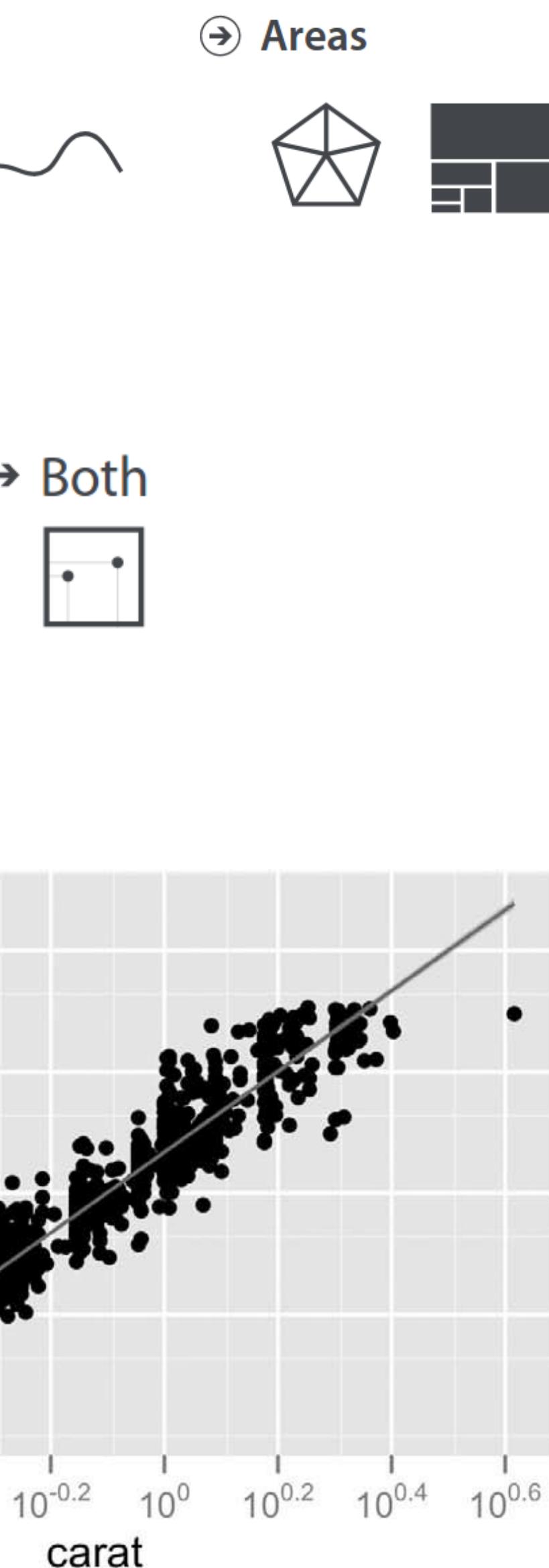
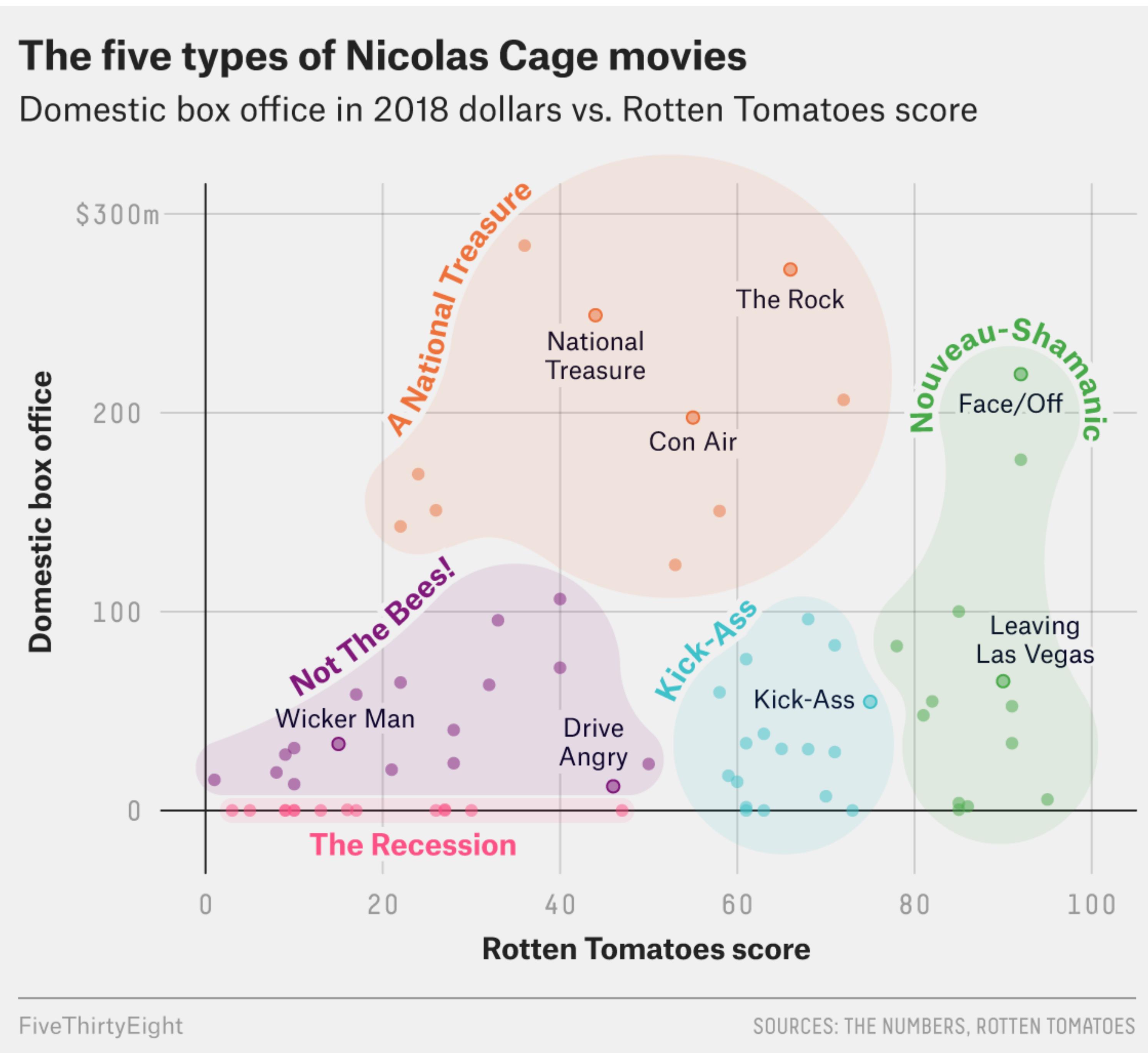


[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.] (From Visualization Analysis and Design)

Scatterplot

- Expresa 2 atributos
- Sin keys, solo valores
- Marcas: Puntos
- Canales: Posición
- Tareas:
 - Encontrar patrones
 - Analizar distribuciones
 - Identificar outliers
- Escalable a cientos de miles

⇒ Express Value



ational and Graphical Statistics
sign)

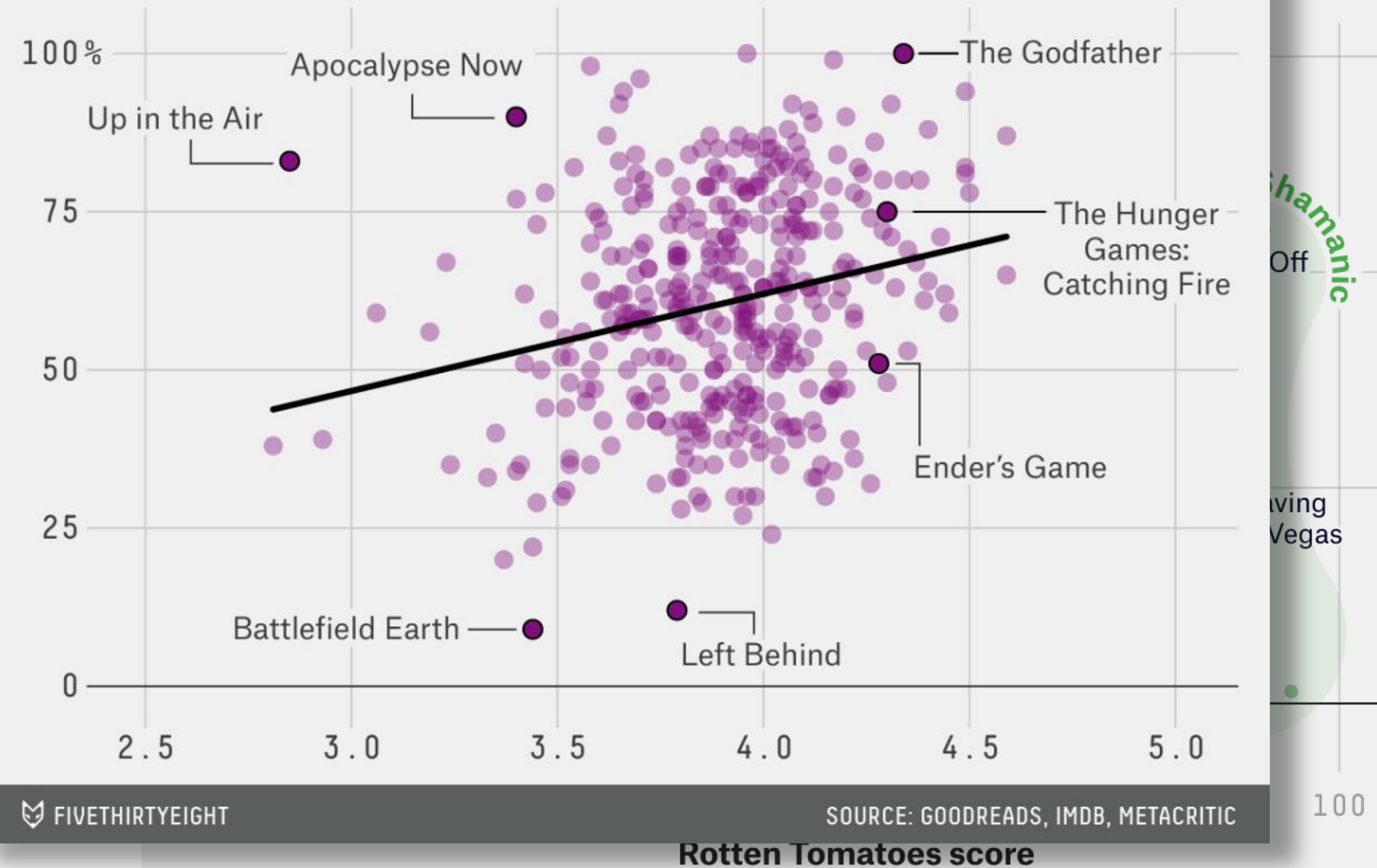
Scatterplot

- Expressive
- Sin keys,
- Marcas:
- Canales:
- Tareas:
- Encon...
- Analiza...
- Identifi...
- Escalable

→ Ex

When Books Become Movies

Metacritic score of films vs. Goodreads score of source novel



FIVETHIRTYEIGHT

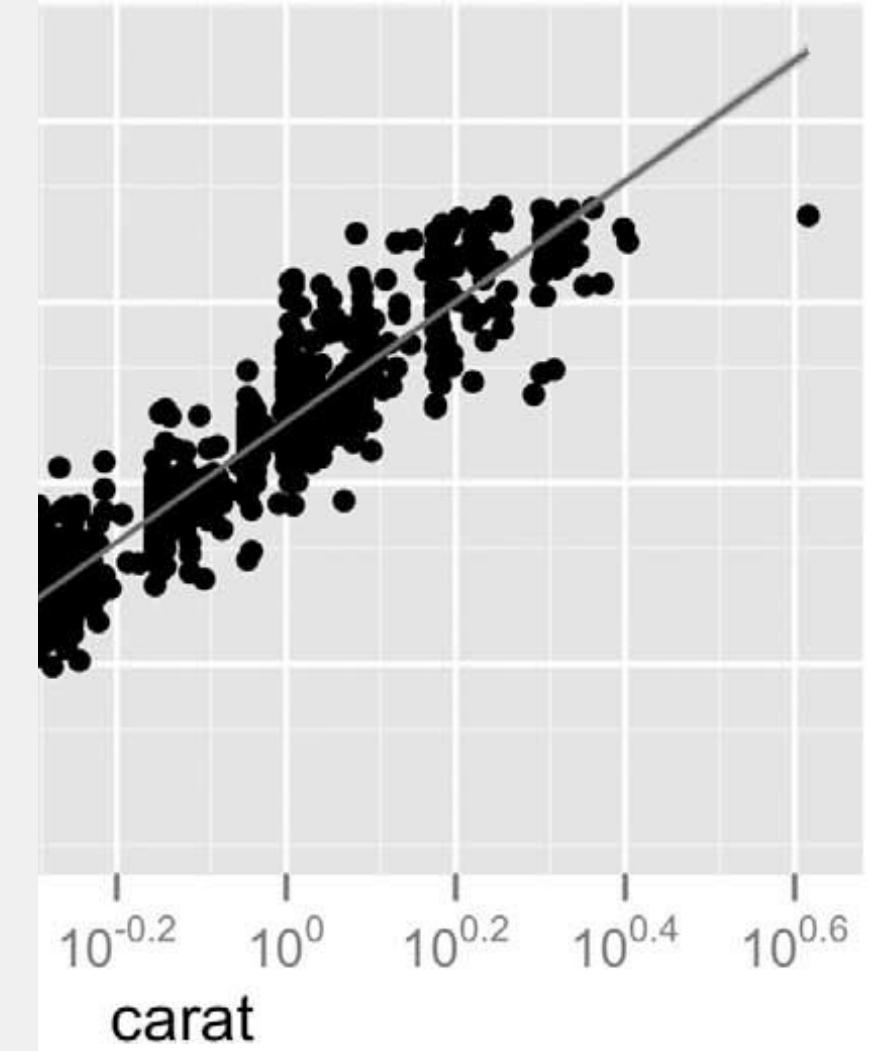
FiveThirtyEight

SOURCES: THE NUMBERS, ROTTEN TOMATOES

→ Areas



→ Both

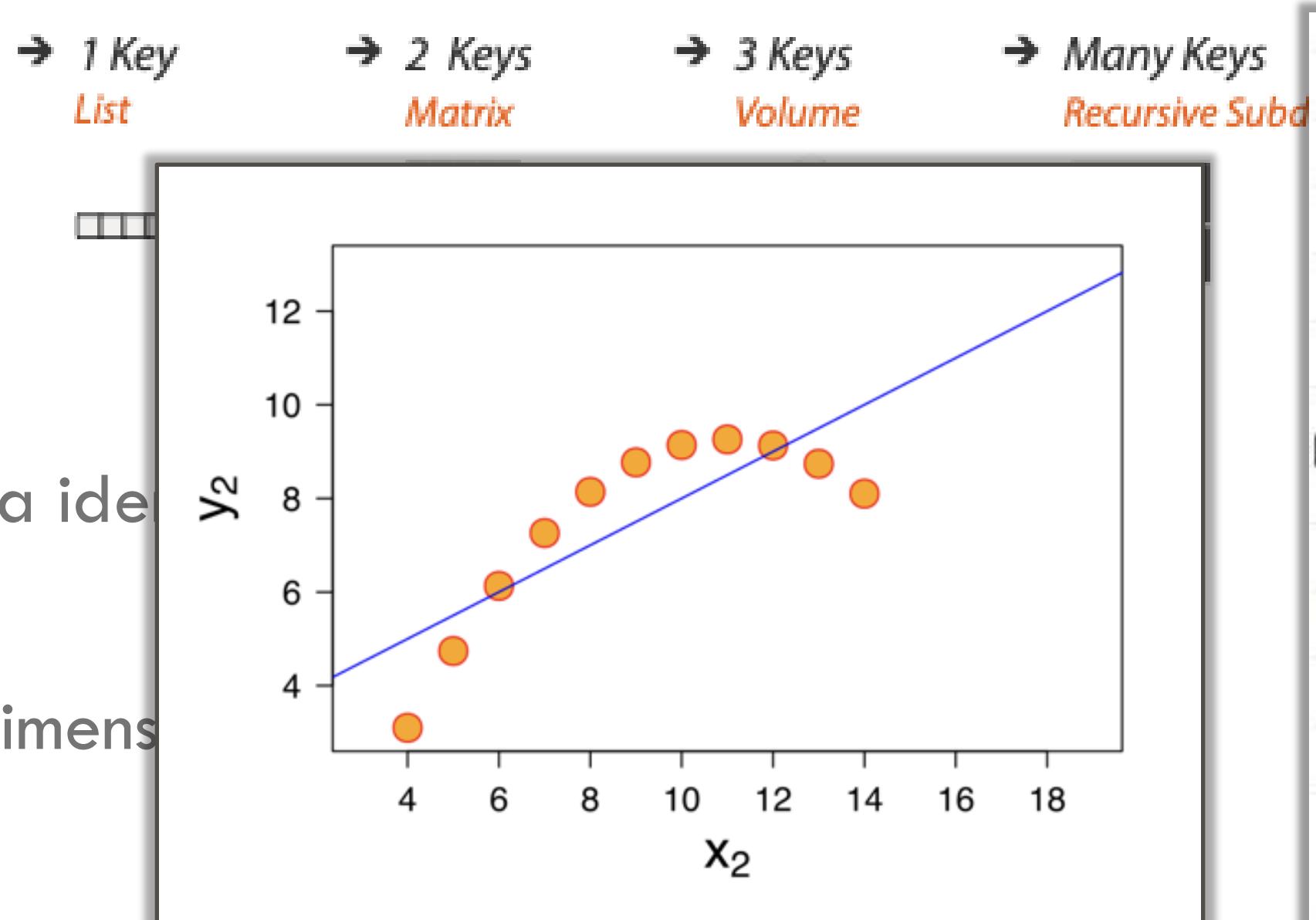


ational and Graphical Statistics
sign)

Keys / Values

Key

- Variable independiente
- Se puede usar como índice único para identificar filas y columnas
- Tipo categórico u ordinal
- Tablas simples: 1 key - Tablas multidimensionales

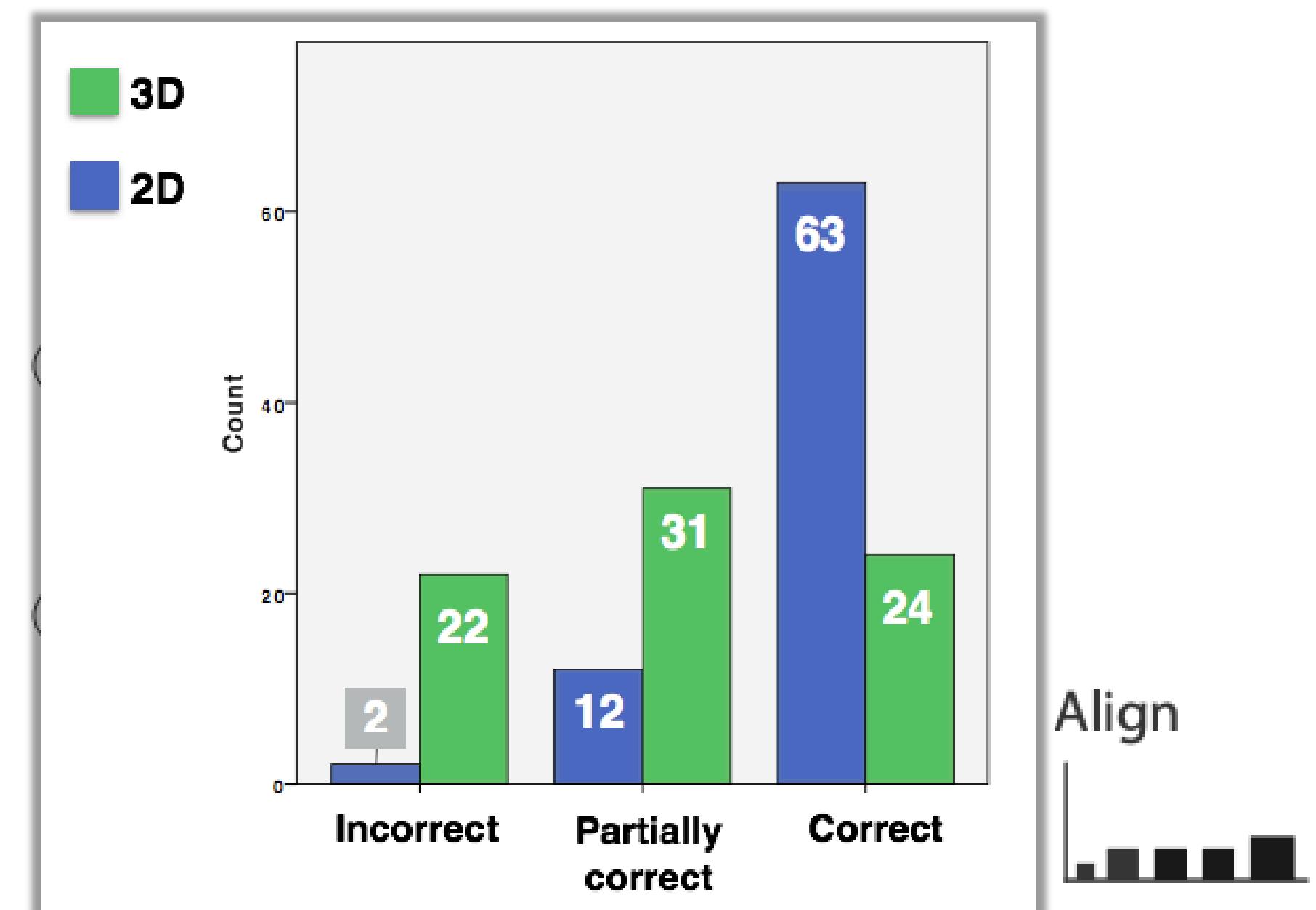


Value

- Variable dependiente, valor de la celda
- Categórico, ordinal o cuantitativo
- Los valores únicos de ordinales y categóricos se llaman **niveles**

Decisiones básicas de diseño

- ¿Cuántos Keys? ¿cuántos Valores?
- ¿De qué tipo son? (categorico, ordinal, cuantitativo)
- Una viz puede mostrar dos valores y ninguna clave (scatter), un valor y una clave (barchart), un valor y dos claves (heatmap), etc.

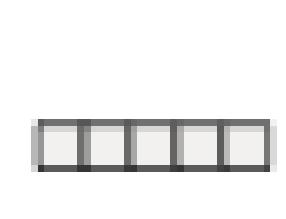


Keys

→ Express Values

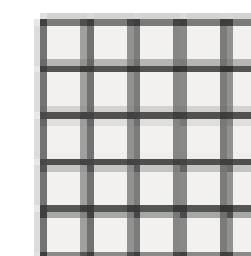
→ 1 Key

List



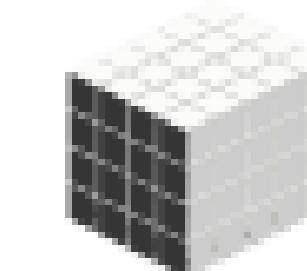
→ 2 Keys

Matrix



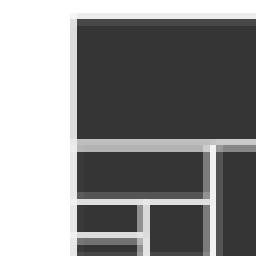
→ 3 Keys

Volume

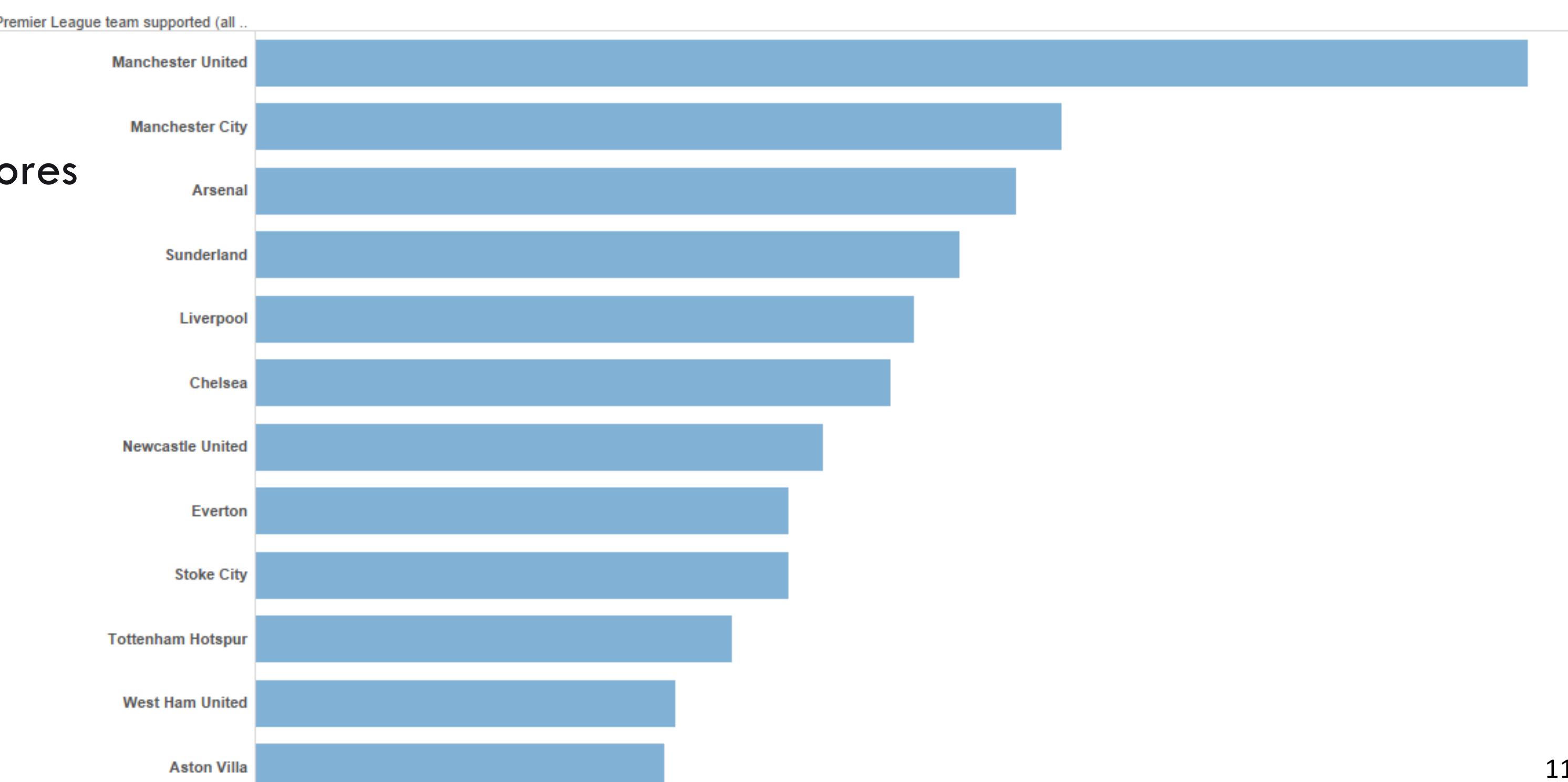


→ Many Keys

Recursive Subdivision



- Estrategias de visualización que involucran valores y **claves**
- Con **una Clave** normalmente se organizan las marcas en una lista unidimensional alineada
- Las claves **NO** son cuantitativas



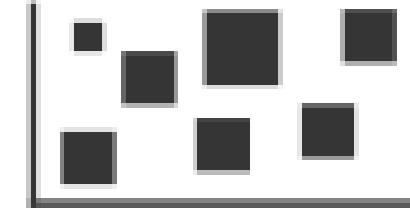
Keys - Variables categóricas

→ Express Values

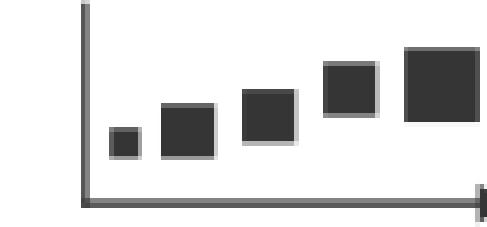


→ Separate, Order, Align Regions

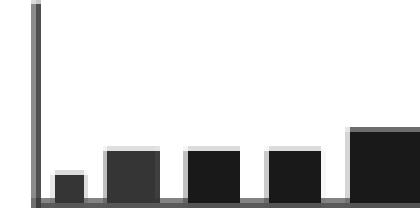
→ Separate



→ Order



→ Align



Channels: Expressiveness Types And Effectiveness Ranks

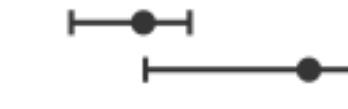
→ Magnitude Channels: Ordered Attributes

Position on common scale



↑
Best

Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



→ Identity Channels: Categorical Attributes

Spatial region



Color hue



Motion



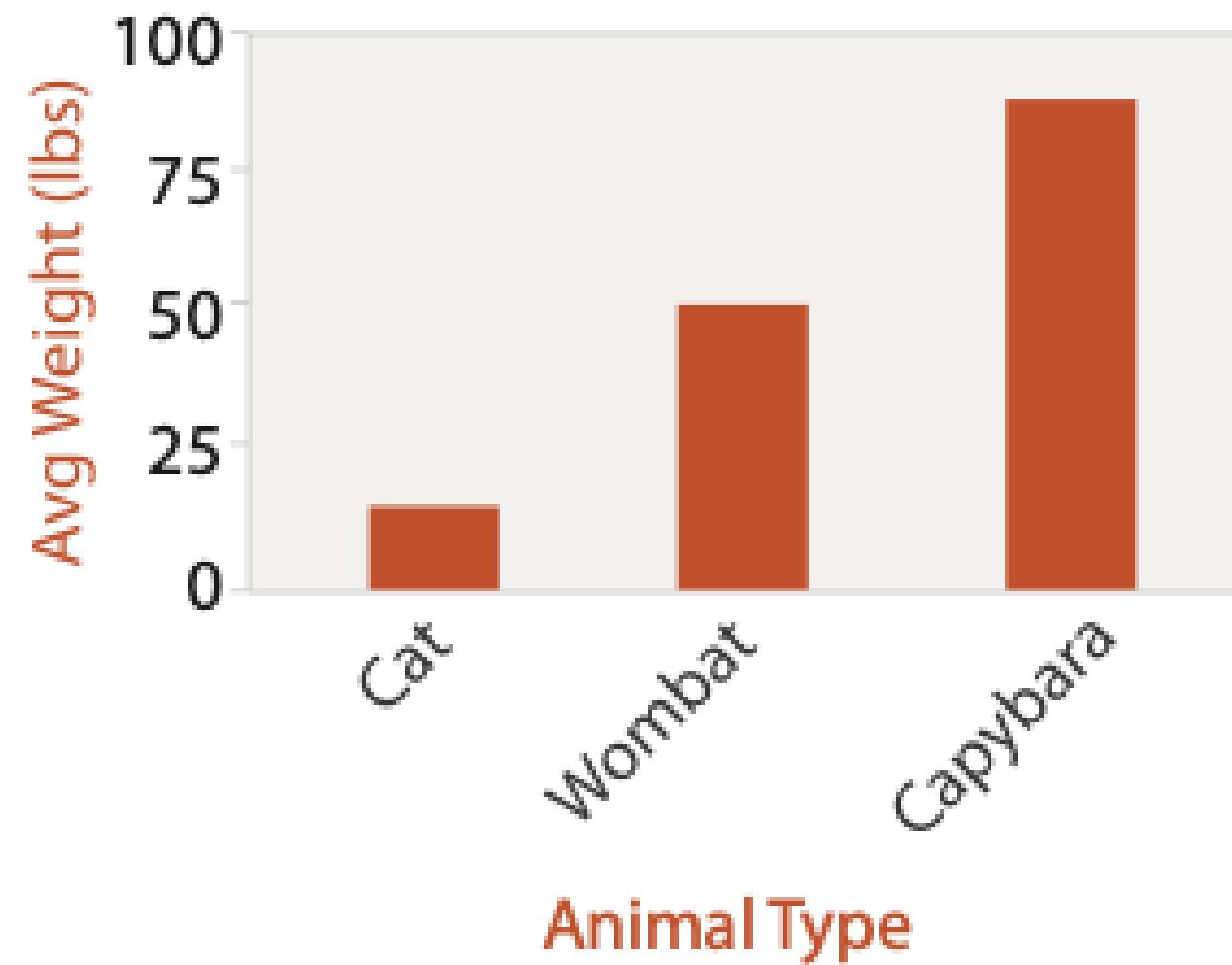
Shape



- Expresar Valores se usa con cuantitativos
- Atributos Categóricos en posición espacial es más complejo - Posición es un canal de **magnitud**
- Atributos categóricos encajan mejor con la idea de **Región**- áreas separadas y distinguibles.
- Tres operaciones:
 - **Separar** - Definir regiones y proximidad. Usar el atributo categórico
 - **Ordenar** – Se debe usar un atributo de semántica ordenable (ordinal/cuantitativo).
 - **Alinear** – Establecer un eje común.

Bar chart

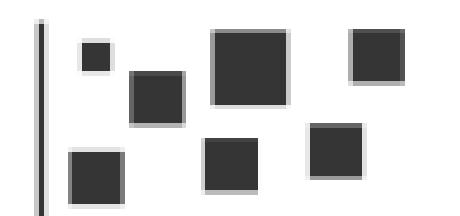
- 1 key, 1 value
- Datos: 1 categórico y 1 cuantitativo
- Marcas: Lineas
- Canales:
 - Longitud para cuantitativo
 - Regiones para categórico:
 - Una por marca
 - Separadas en un eje (X),
 - Alineadas en el otro (Y)
 - Ordenadas por cuantitativo
 - Label (alfabético) – tamaño (data-driven)
- Tareas:
 - Comparar, ver + valores
 - Escalable de docenas a cientos. Limitada por tamaño en pantalla



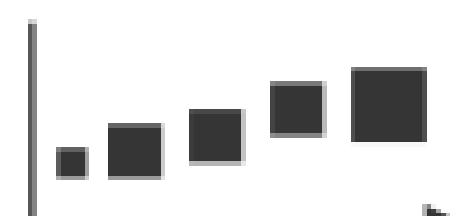
→ Express Values
→ →

→ Separate, Order, Align Regions

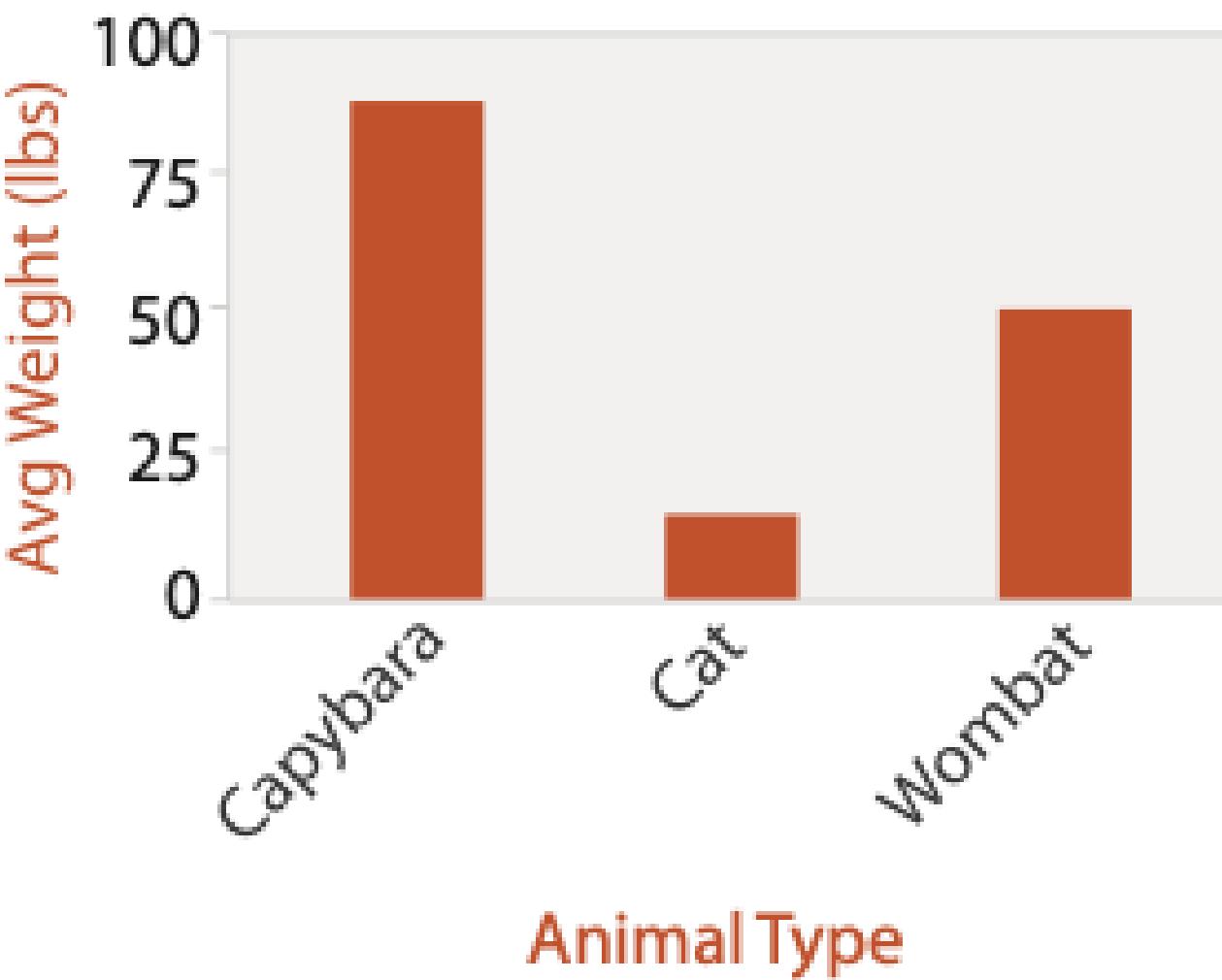
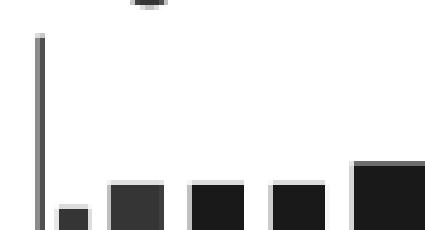
→ Separate



→ Order



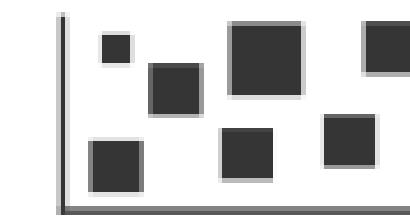
→ Align



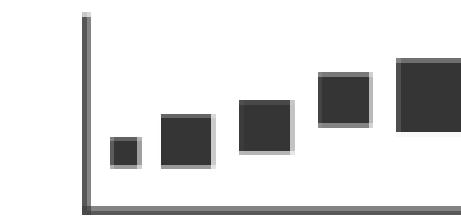
Separar - Alinear - Ordenar

→ Separate, Order, Align Regions

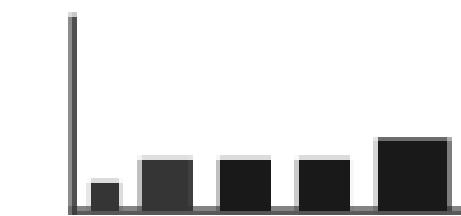
→ Separate



→ Order



→ Align

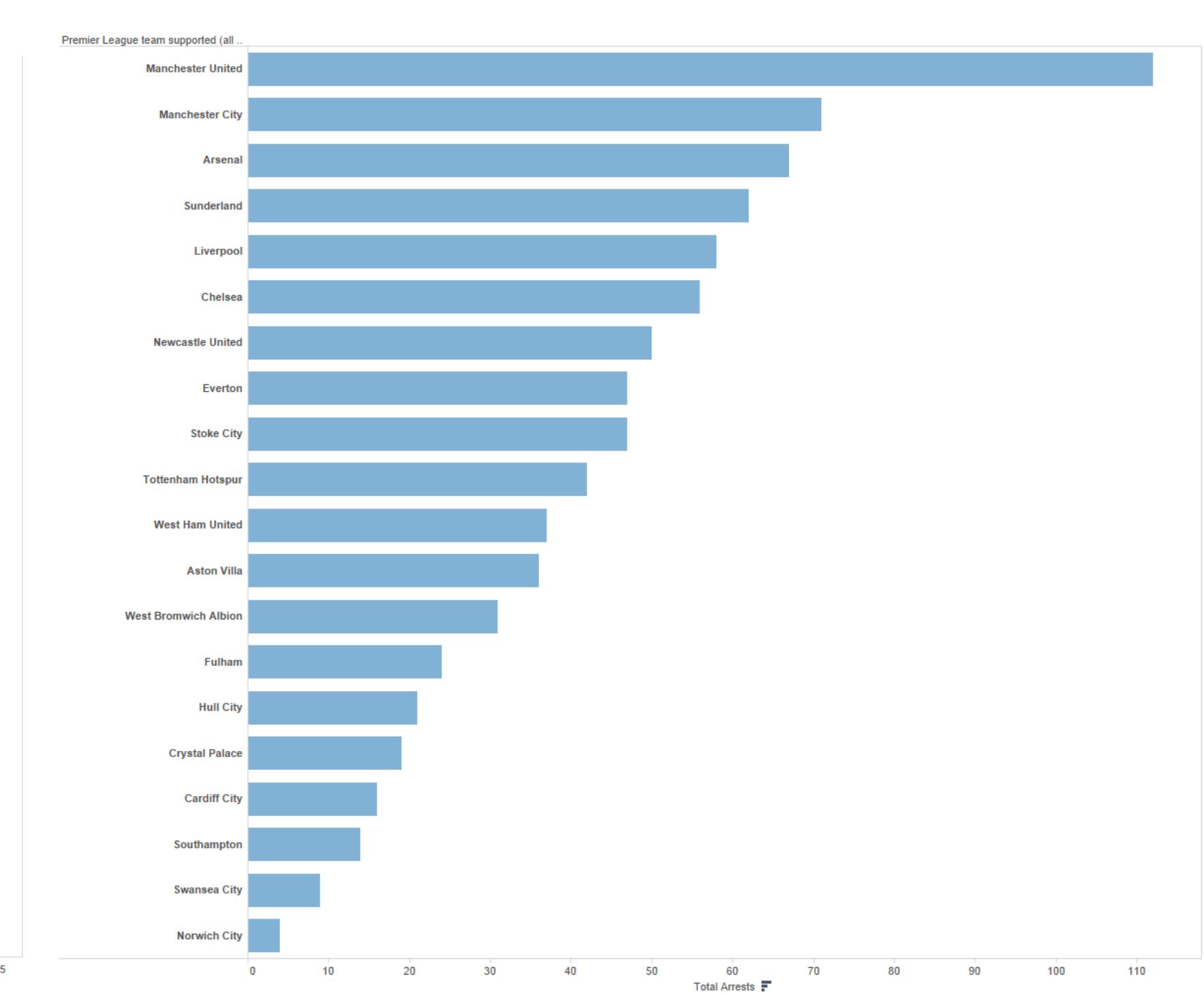
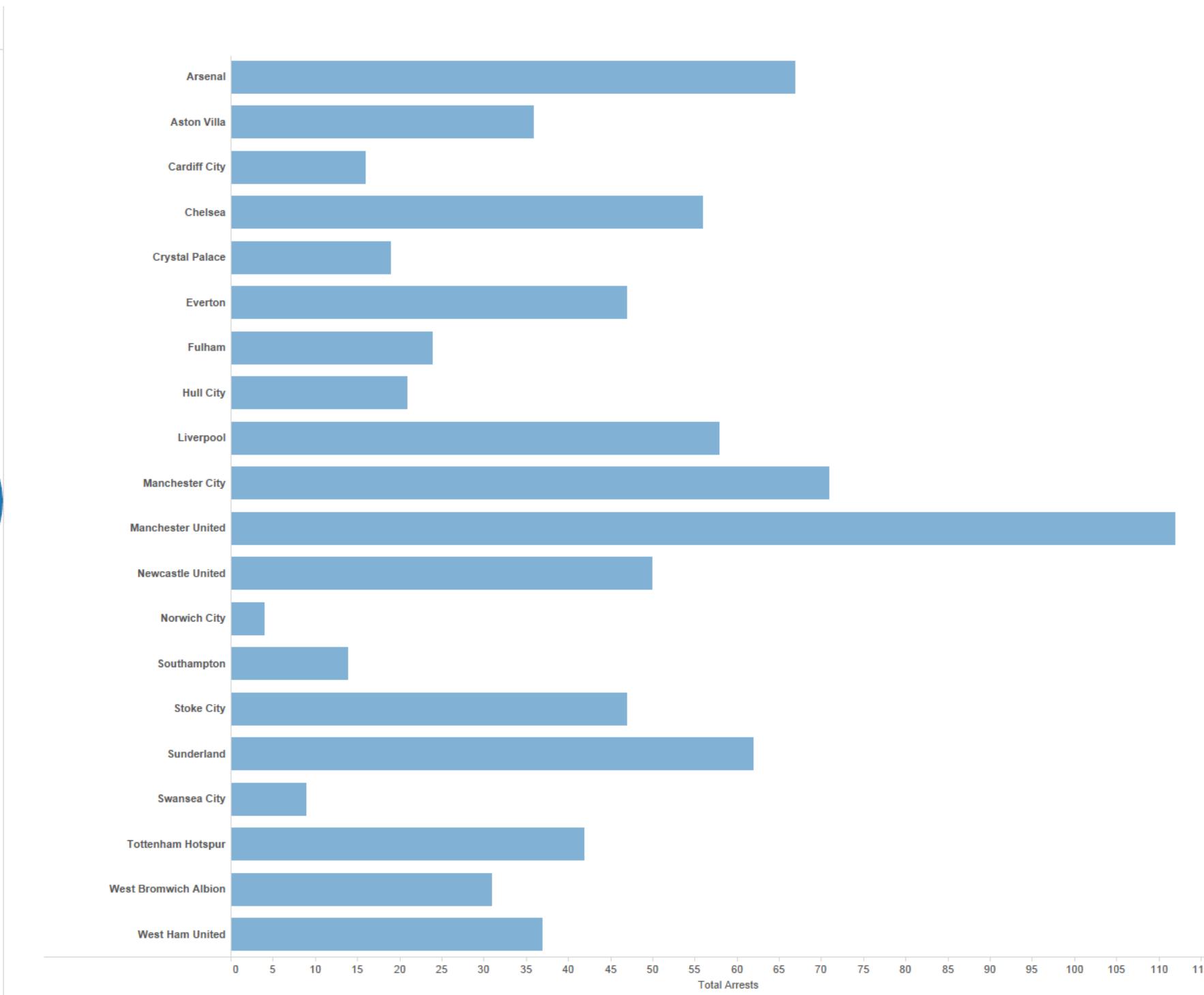
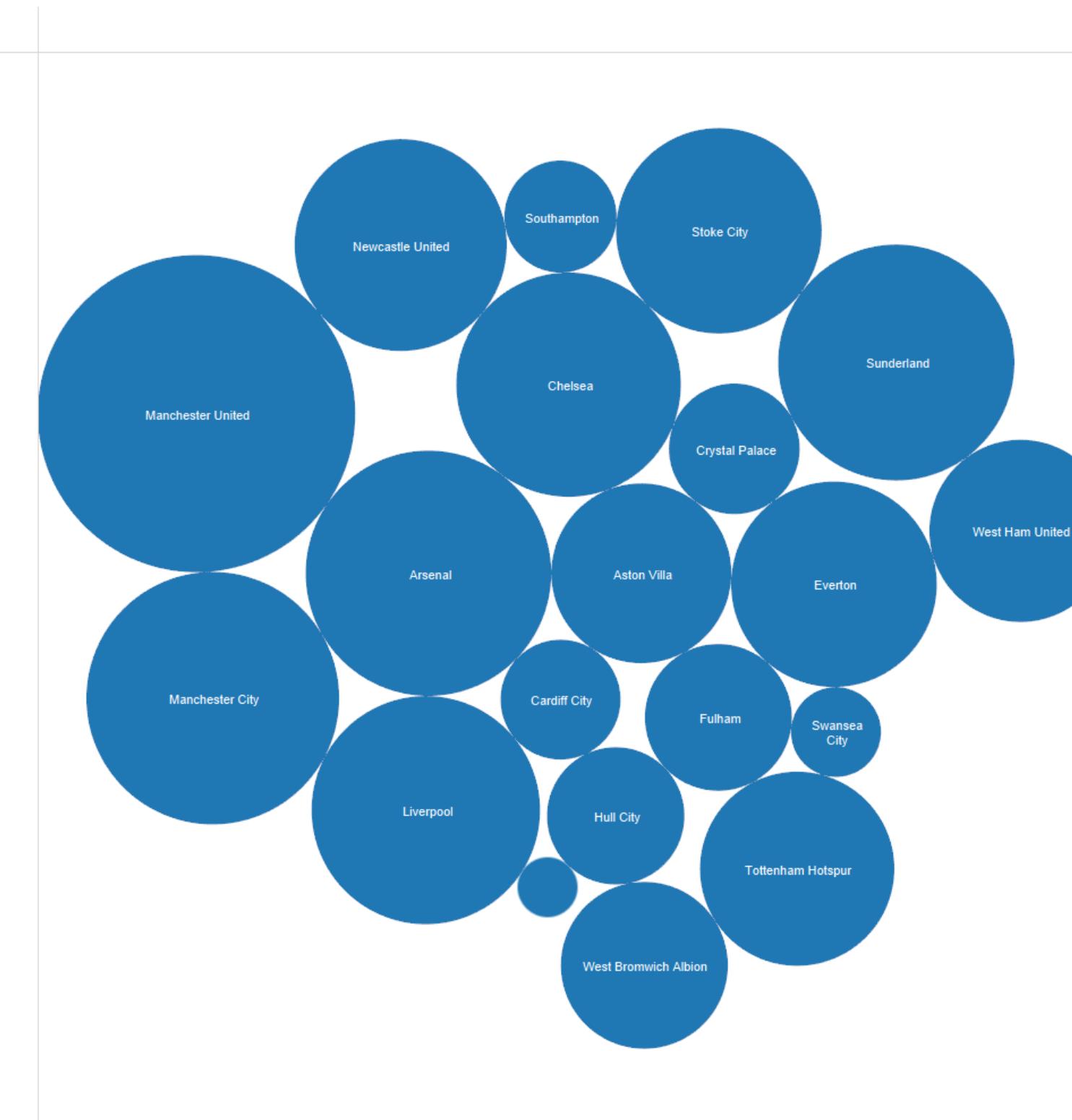


Limitaciones:

Separar- Difícil diferenciar y comparar

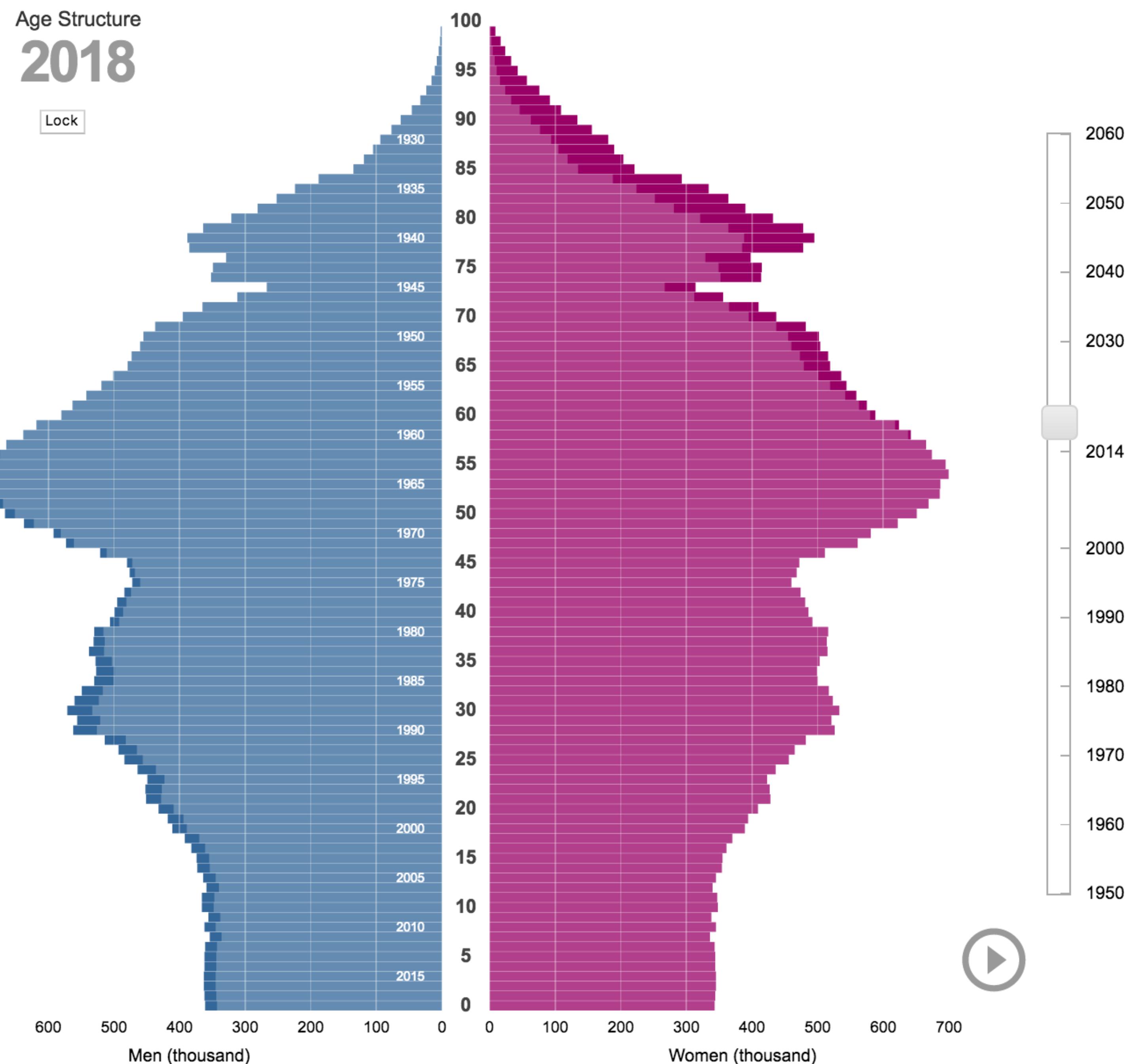
Separar + Alinear- Difícil establecer ranking visualmente. ¿Cual es el tercero más grande? ¿El séptimo?

Separar + Alinear + Ordenar- Ranking. Fácil comparar e identificar extremos



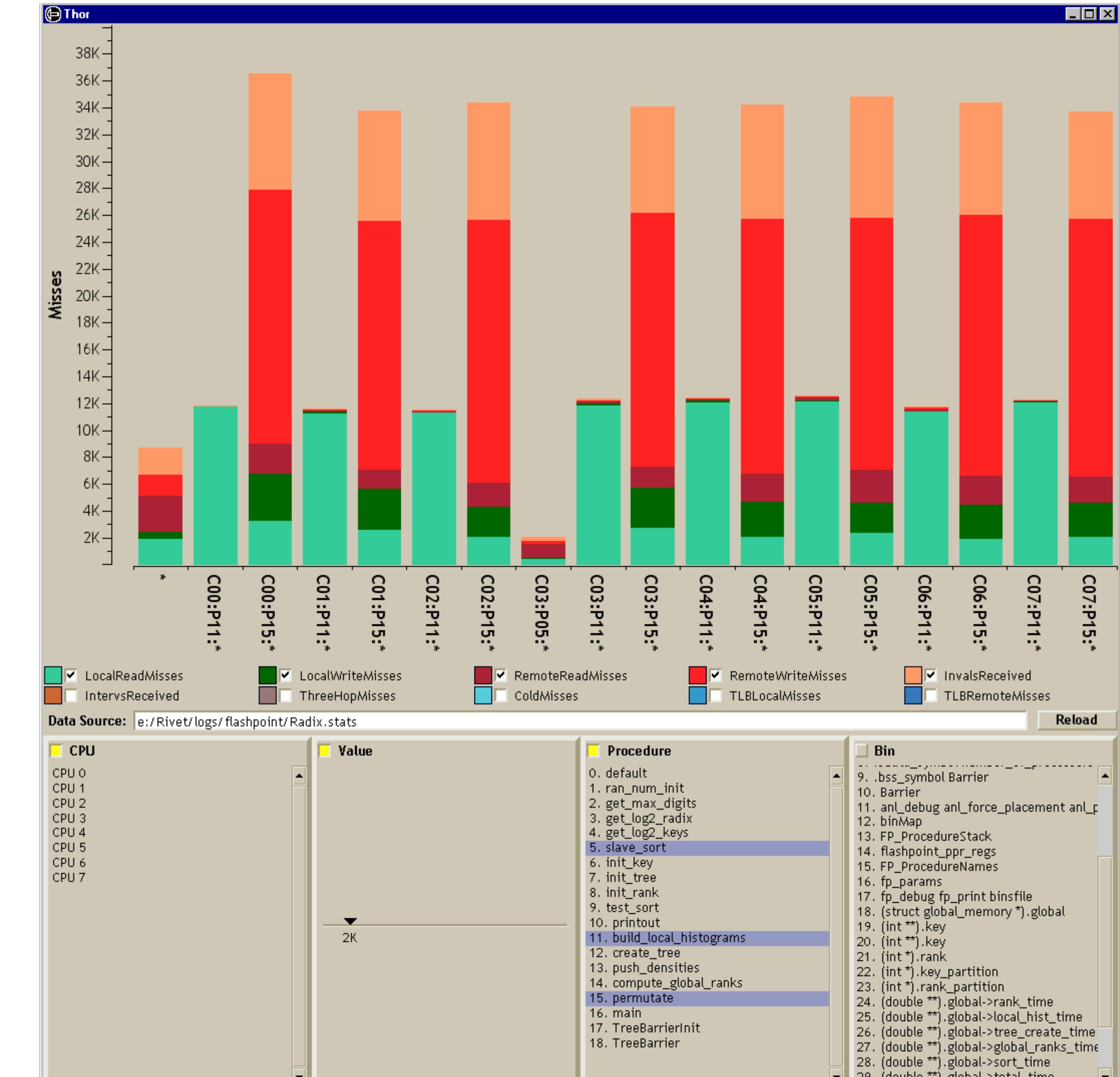
Variación: Barras divergentes

- Pirámide poblacional ($Y = \text{edad}/X = \text{count}$)
- Datos: 1 categórico, 1 ordinal, 1 cuantitativo (+cuantitativo interactivo)
- Marcas: Líneas
- Canales: Longitud y color
- Regiones: Una por marca; ordenadas por atr. Ordinal (grupo de edad)
- Tarea:
 - Analizar forma de la distribución
 - Comparar distribuciones
- Limitaciones: Difícil comparar entre lados
- Solución: marcar la diferencia en un color más oscuro



Stacked bar chart

- **2 keys, 1 value**
- Datos: 2 categórico y 1 cuantitativo
- Marcas: Lineas apiladas verticalmente
- **Glifos:** Objetos compuestos por múltiples marcas:
 - Canales: longitud y Tono
- Regiones:
 - Una por glifo
 - Alineadas:
 - Componente más bajo
 - Otros componentes del glifo
- Tareas:
 - Relaciones parte-todo
 - Escalable a <docena de niveles para el atributo apilado; igual a barras para regiones



[Using Visualization to Understand the Behavior of Computer Systems. Bosch. Ph.D. thesis, Stanford Computer Science, 2001.]

From: Visualization Analysis and Design

Stacked bar chart

Ejemplo:

- Aporte al GDP global/anual en puntos porcentuales de países más ricos
- Eje vertical divergente para negativos
- Fácil identificar caídas y tendencias globales
- Difícil ver evolución por país y compararla entre países. Capas inferiores se comparan mejor que superiores.

Channels: Expressiveness Types And Effectiveness Ranks

④ **Magnitude Channels: Ordered Attributes**

Position on common scale



▲
Effectiveness

Position on unaligned scale



Length (1D size)



▲
Effectiveness

Tilt/angle



▲
Effectiveness

Area (2D size)



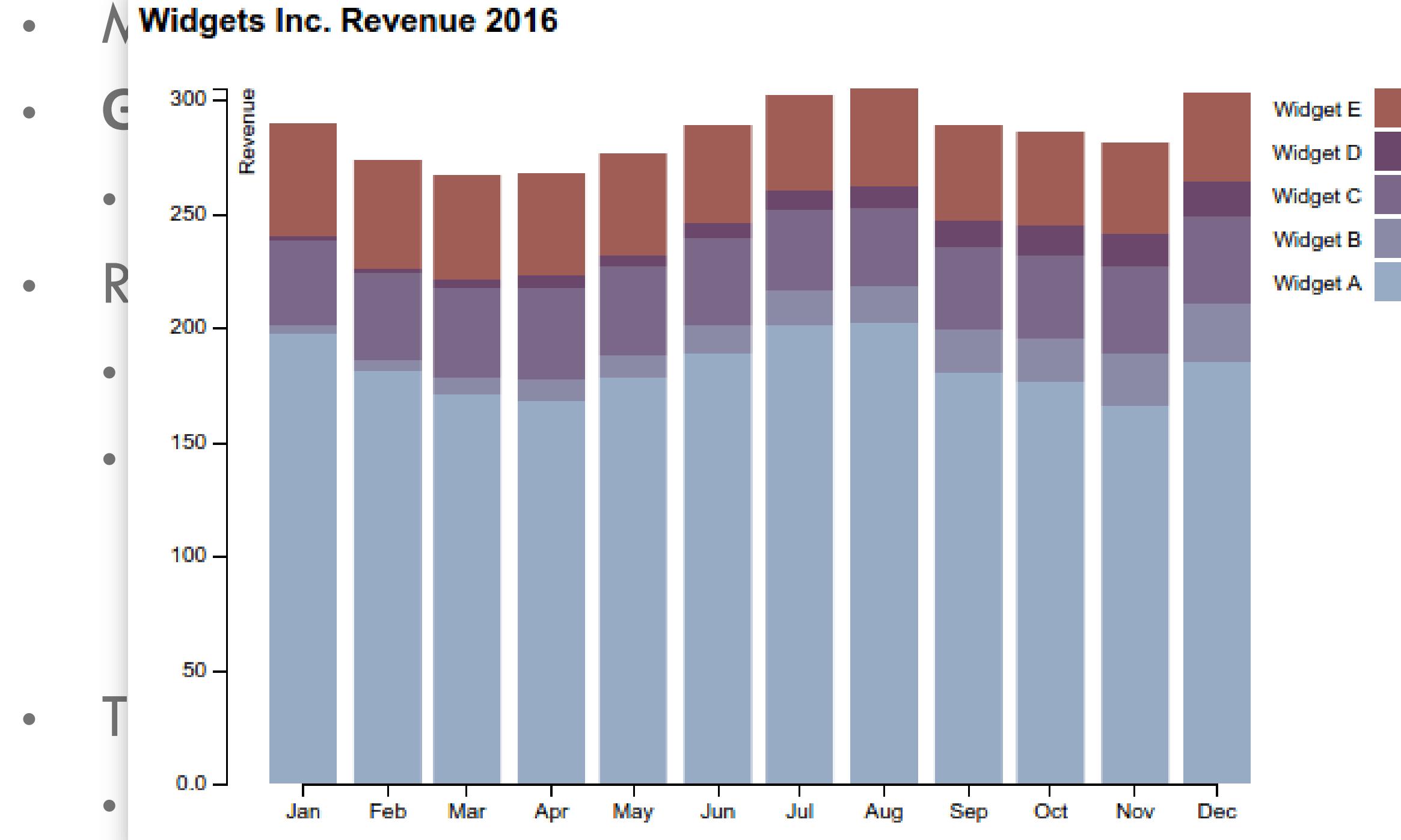
▲
Effectiveness



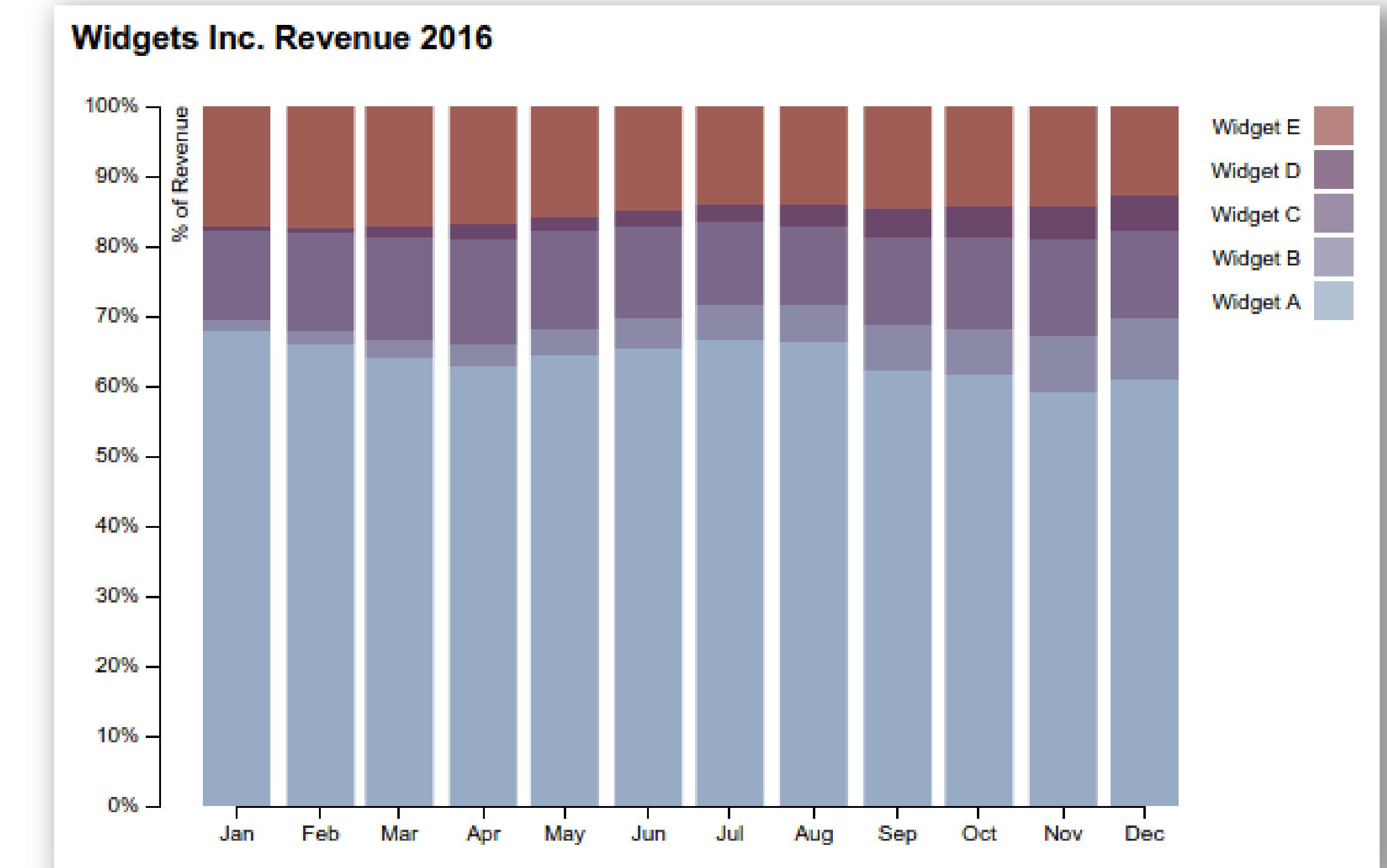
The Economist

Normalized stacked bar chart

- 2 keys, 1 value
- Datos: 2 categórico y 1 cuantitativo



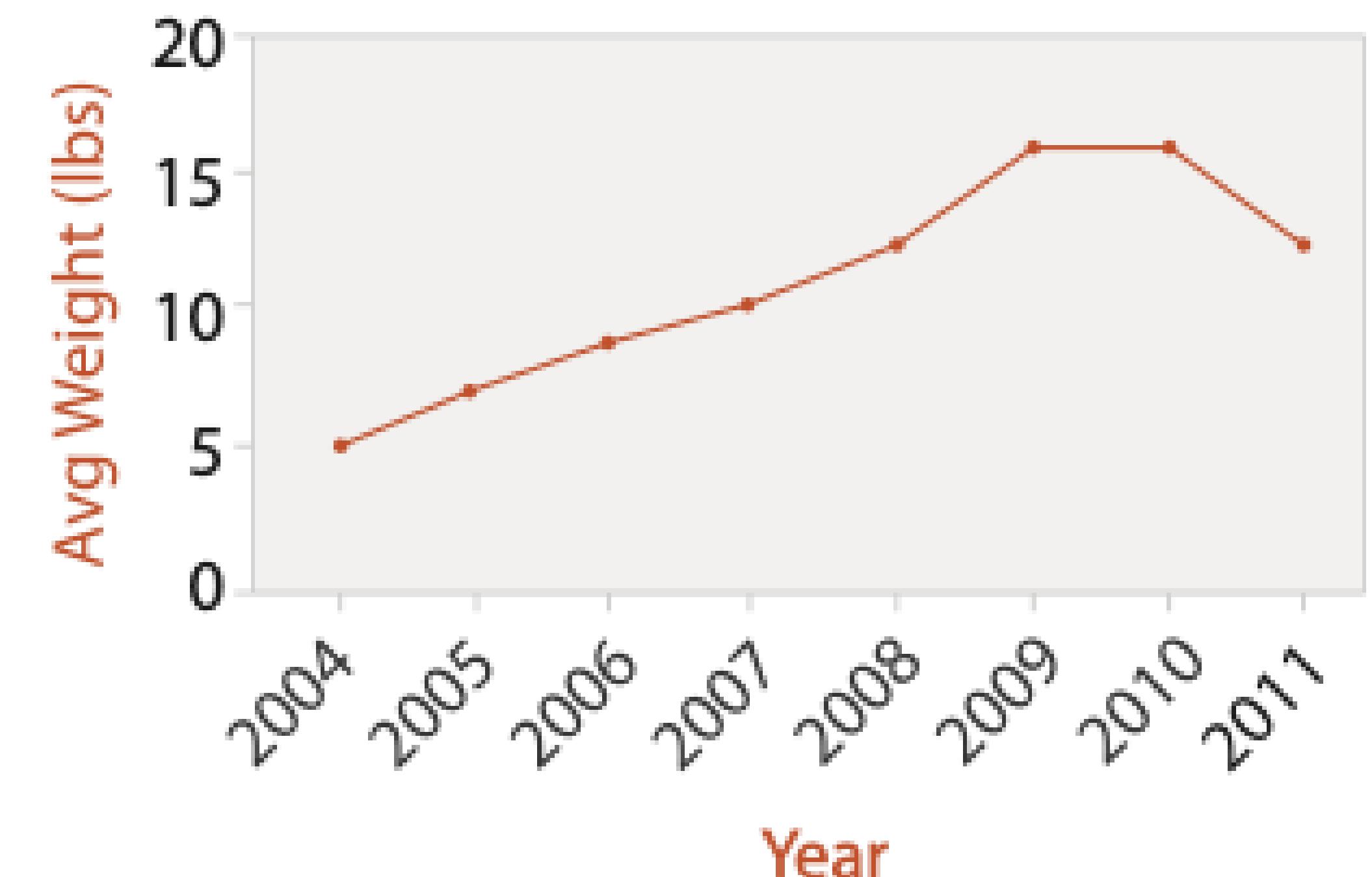
- Escalable a $< \text{docena}$ de niveles para el atributo apilado



Chris Maness, From: medium.com

Line chart / dot plot

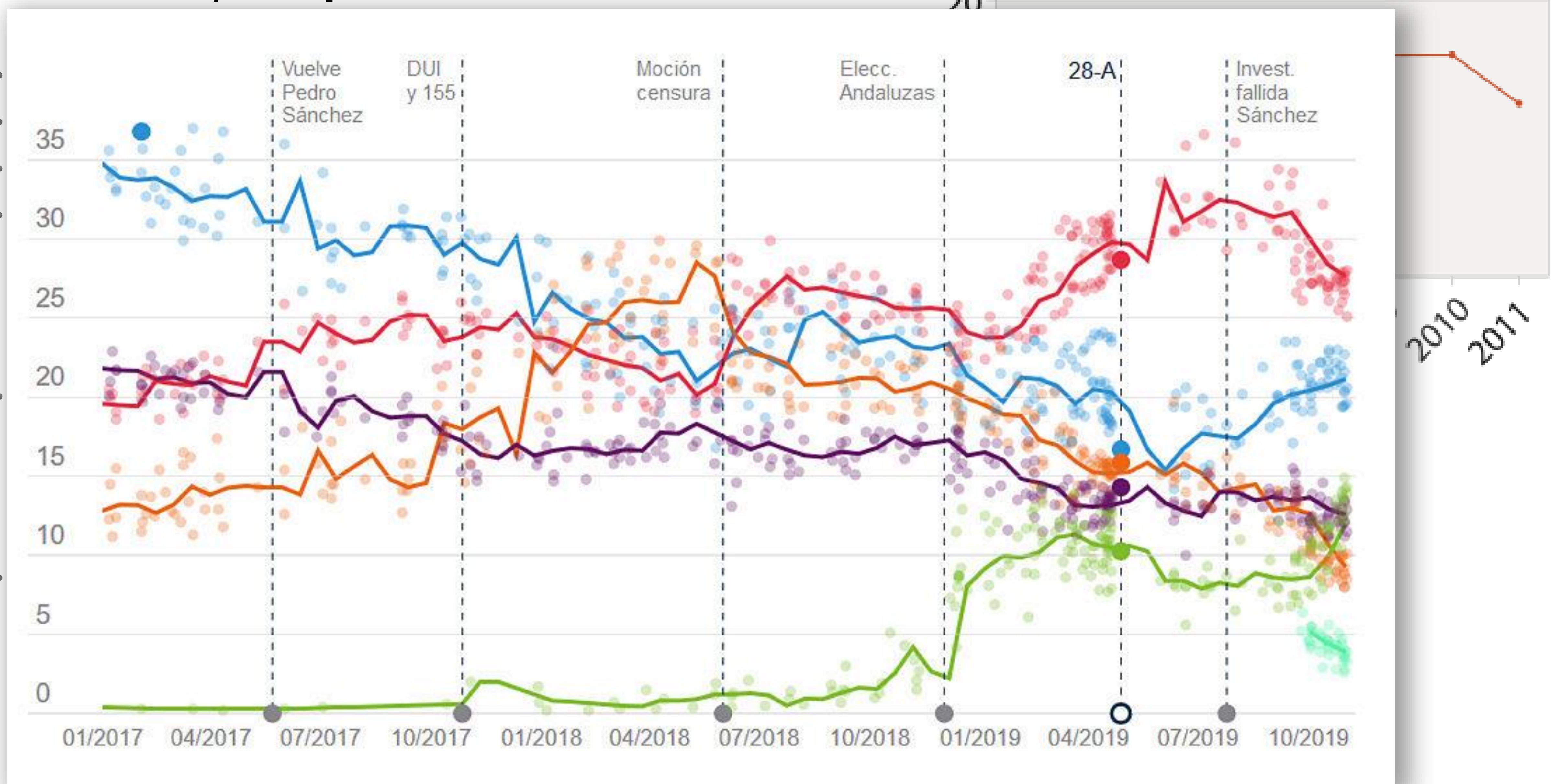
- 1 key, 1 value
- Datos: 2 cuantitativos
- Marcas: Puntos- Línea que los conecta
- Canales:
 - Posición alineada para cuantitativo
 - Separados y ordenados por key en regiones horizontales
- Tareas:
 - Identificar tendencia
 - La línea resalta el orden- Relación explícita entre un ítem y el siguiente
- Escalable a cientos de keys



	A	B	C	D	E	F	G	H
1	Row Labels	Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and	Argentina
2	2000	54,8	72,6	71,3		45,3	73,6	
3	2001	55,3	73,6	71,4		45,7	73,8	
4	2002	56,2	73,3	71,6		46,5	74	
5	2003	56,7	72,8	71,7		46,8	74,2	
6	2004	57	73	72,3		47,1	74,4	
7	2005	57,3	73,5	72,9		47,4	74,6	
8	2006	57,3	74,2	73,4		47,7	74,8	
9	2007	57,5	75,9	73,8		48,2	75	
10	2008	58,1	75,3	74,1		48,7	75,2	
11	2009	58,6	76,1	74,4		49,1	75,4	
12	2010	58,8	76,2	74,7		49,6	75,6	
13	2011	59,2	76,6	74,9		50,1	75,7	
14	2012	59,5	76,9	75,1		50,6	75,9	
15	2013	59,9	77,2	75,3		51,1	76,1	

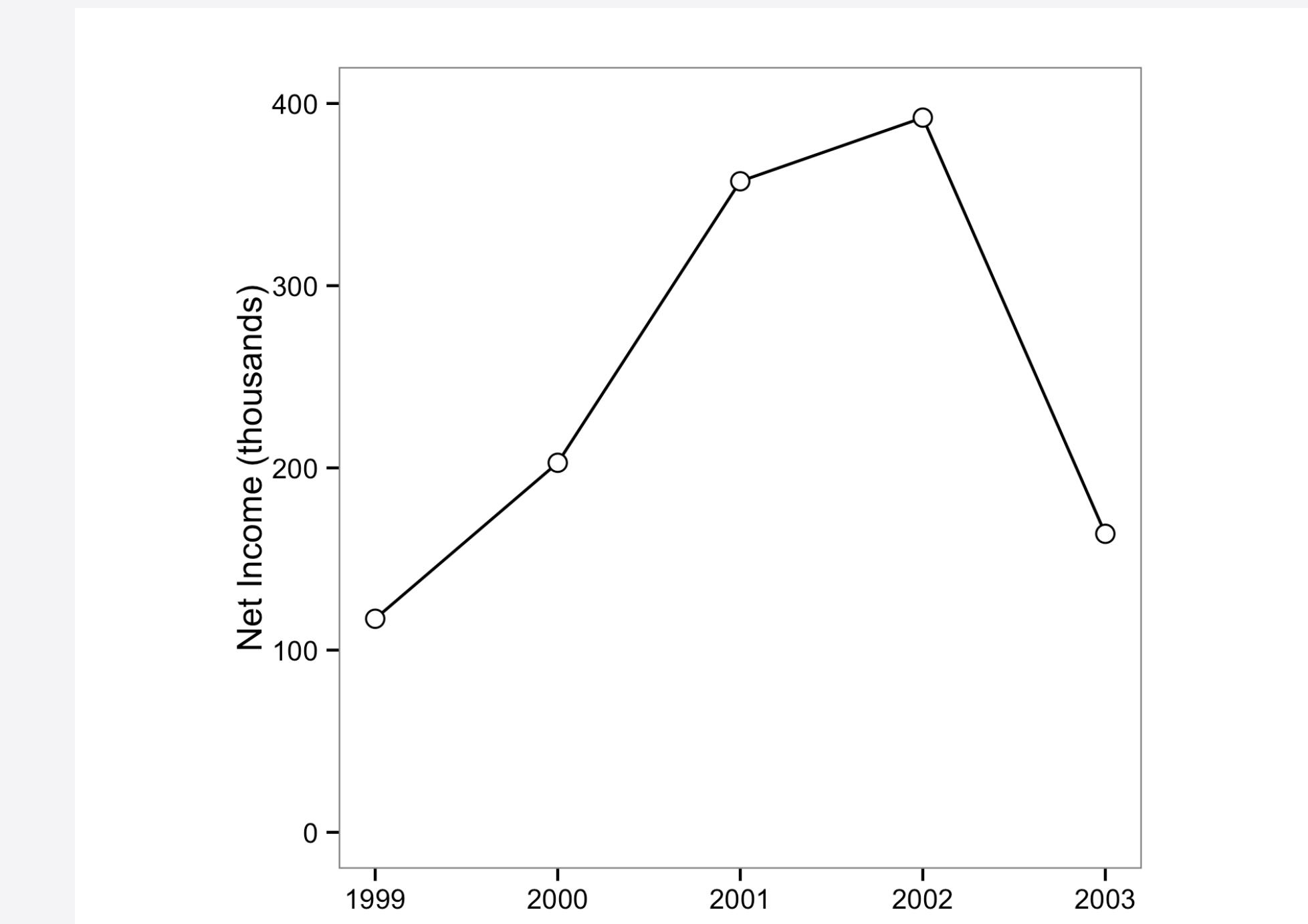
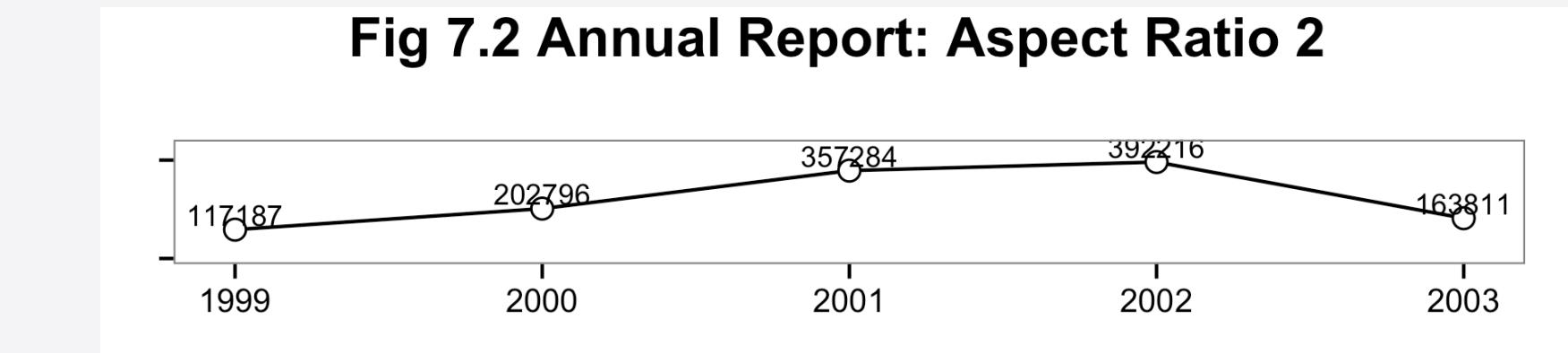
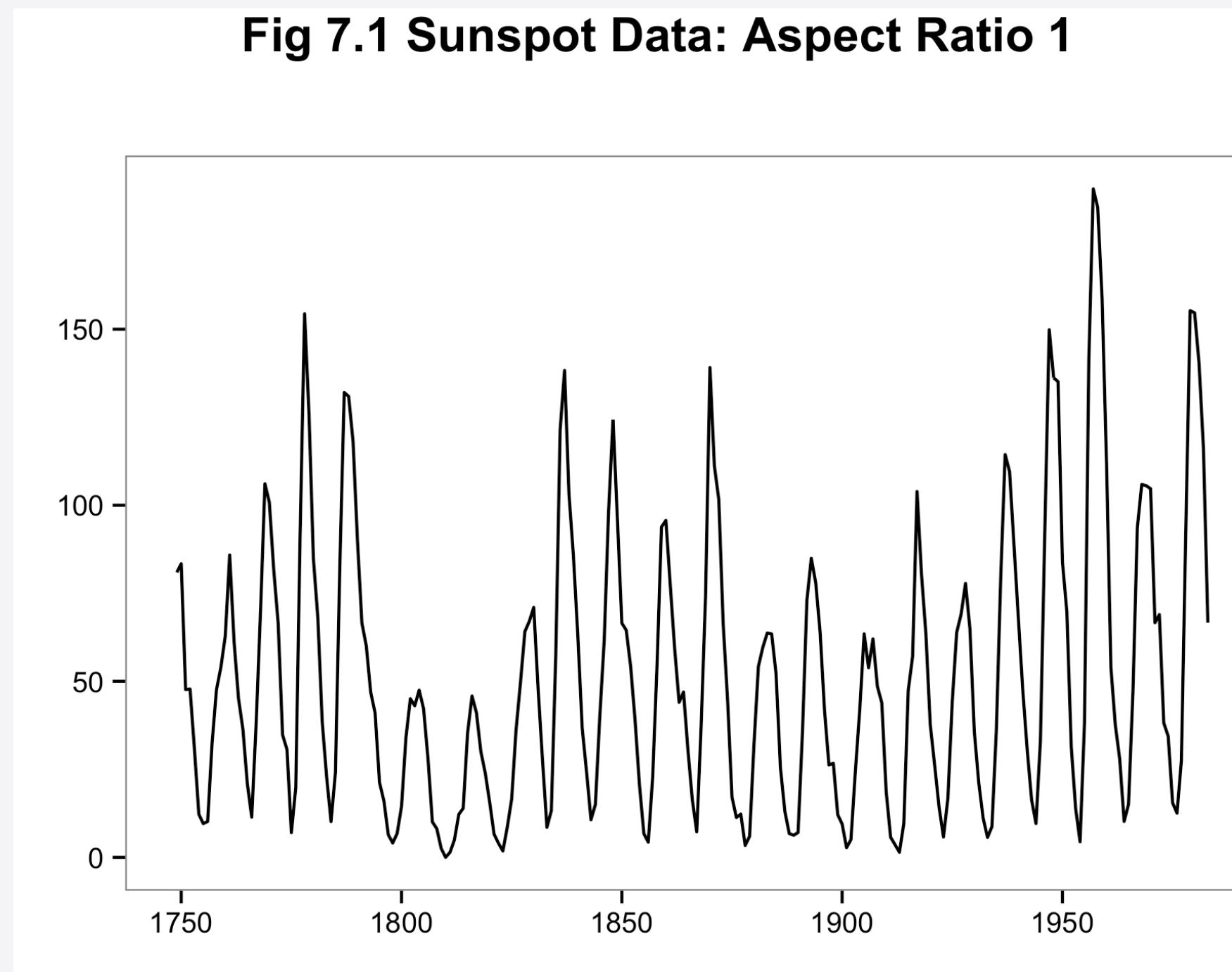
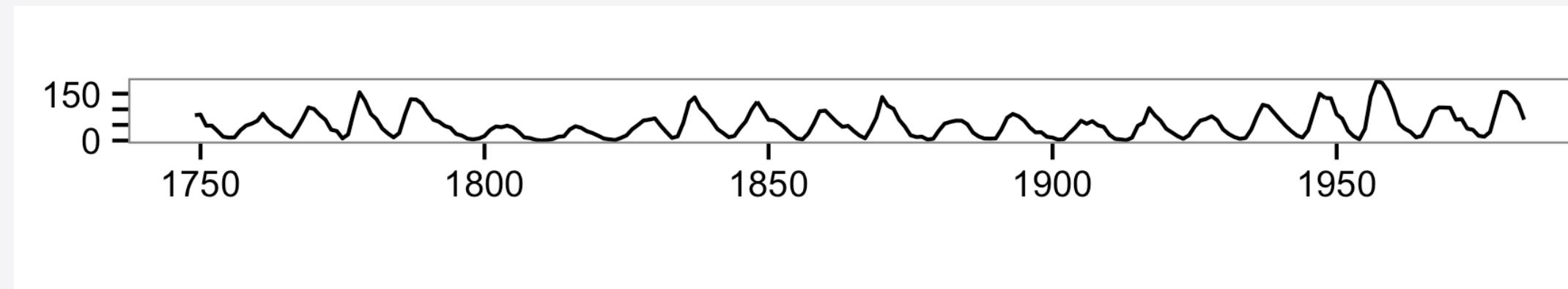
Line chart / dot plot

20



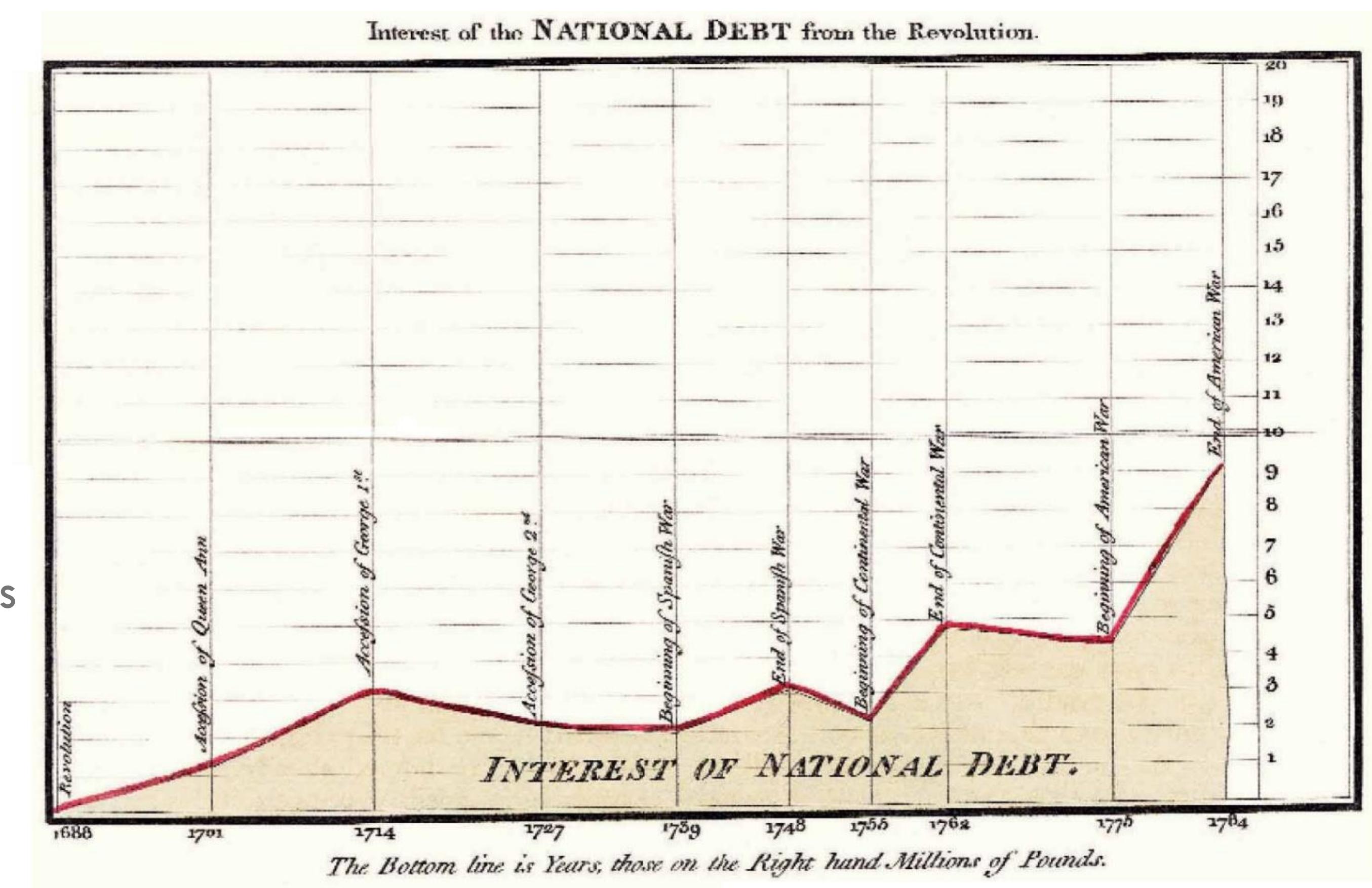
Line chart: Aspect ratio

- 1: Inclinación a 45 (1980s)
 - Cleveland percepción: Diferenciaciones más precisas en ángulos a 45
 - Métodos automáticos para encontrar ratio óptimo.



Area charts

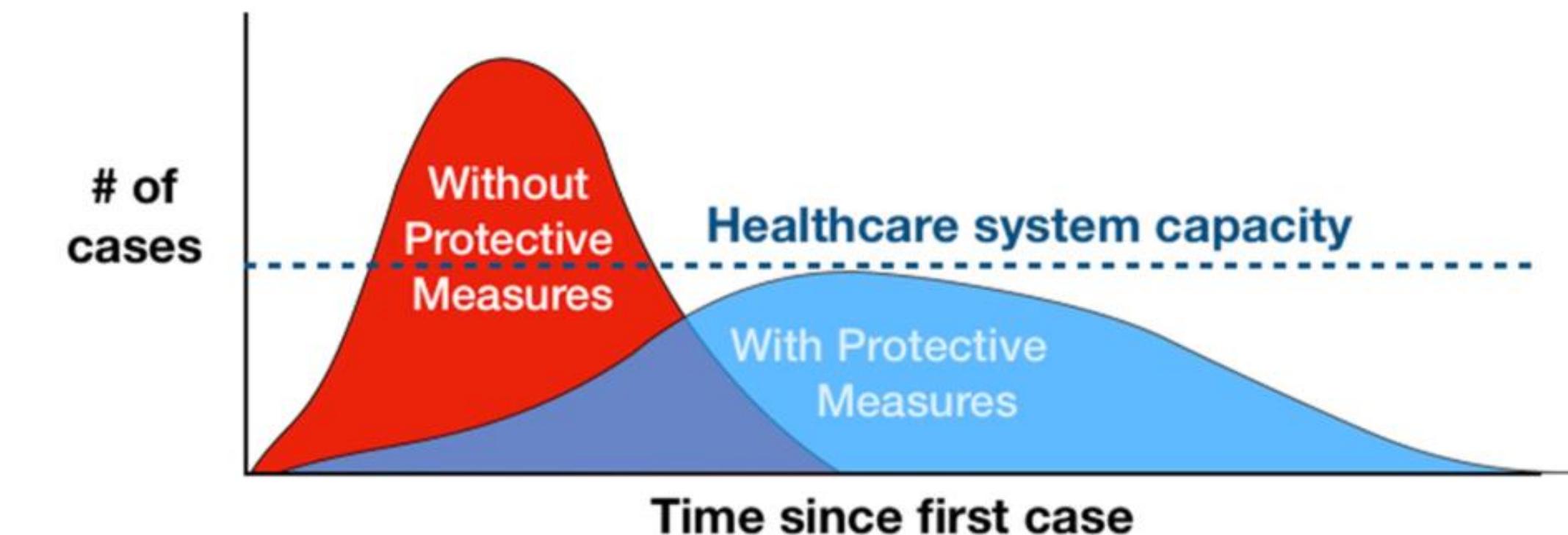
- 1 key, 1 value
- Datos: 2 cuantitativos y un categórico
- Datos derivados: geometrías
- Marcas: Línea, Área
- Canales:
 - Color
 - Posición alineada para cuantitativo
 - Separados y ordenados por key en regiones horizontales
- Tareas:
 - Mostrar variación en cantidades
 - El área resalta el orden y la asociación con “tamaño” del valor
- Escalable a cientos de keys y values



W. Playfair. From: Wikipedia

Area charts

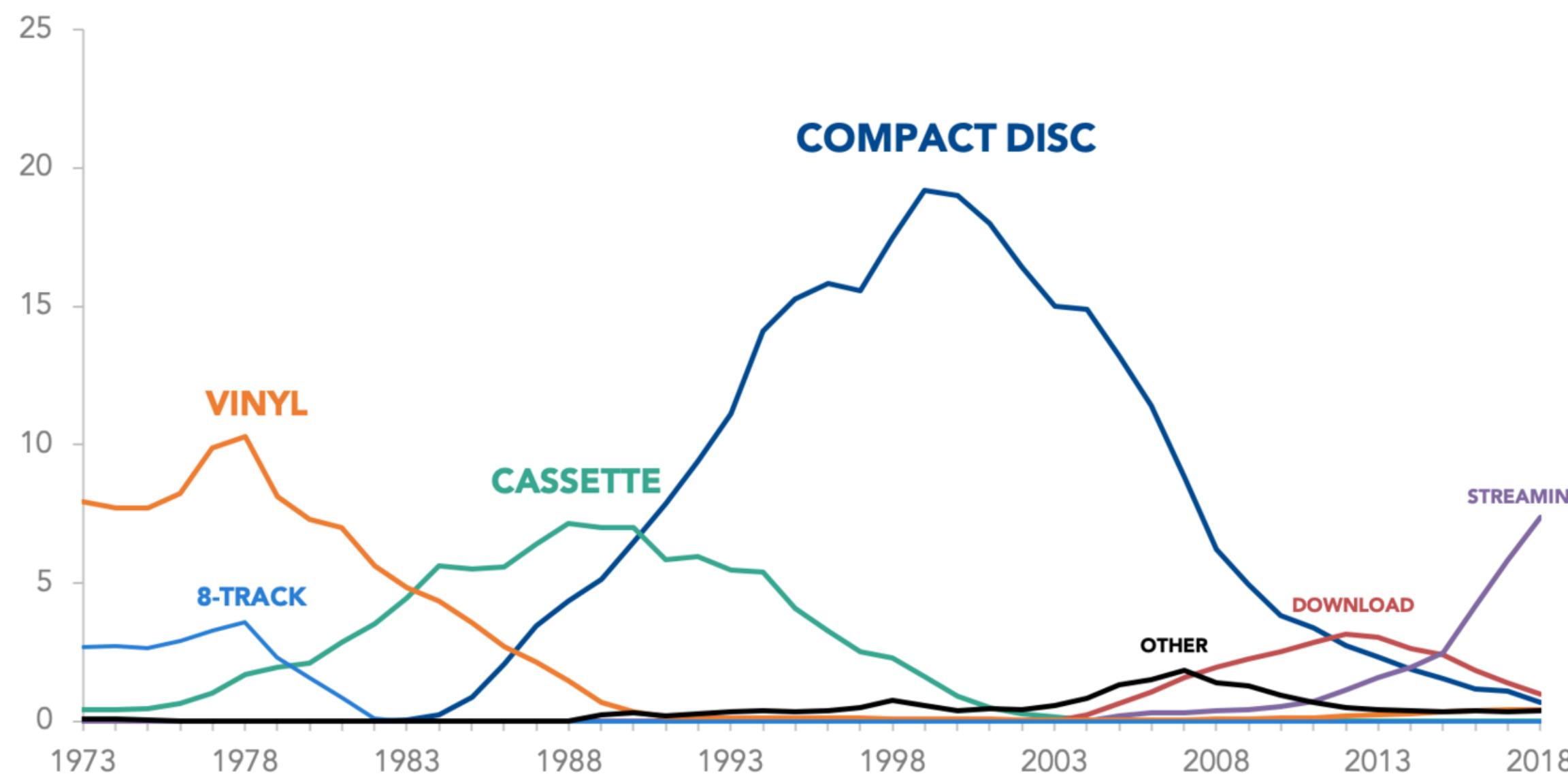
- Buenas para líneas que se cruzan o tienen subidas y bajadas pronunciadas y cerca del 0



SOURCE: nytimes.com/2020/03/11/science/coronavirus-curve-mitigation-infection.html

US music sales by format (inflation-adjusted)

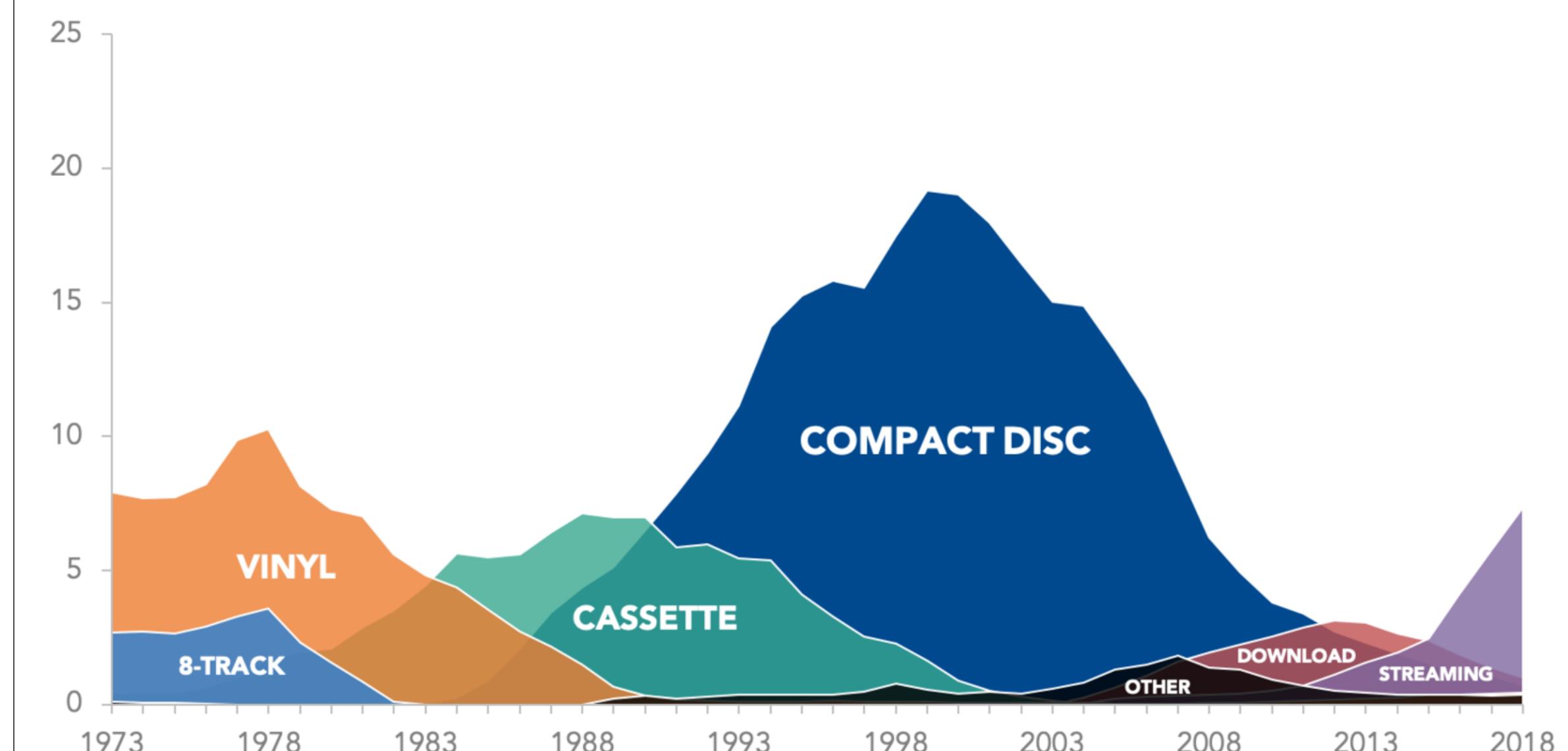
IN BILLIONS (USD)



SOURCE: Recording Industry Association of America

US music sales by format (inflation-adjusted)

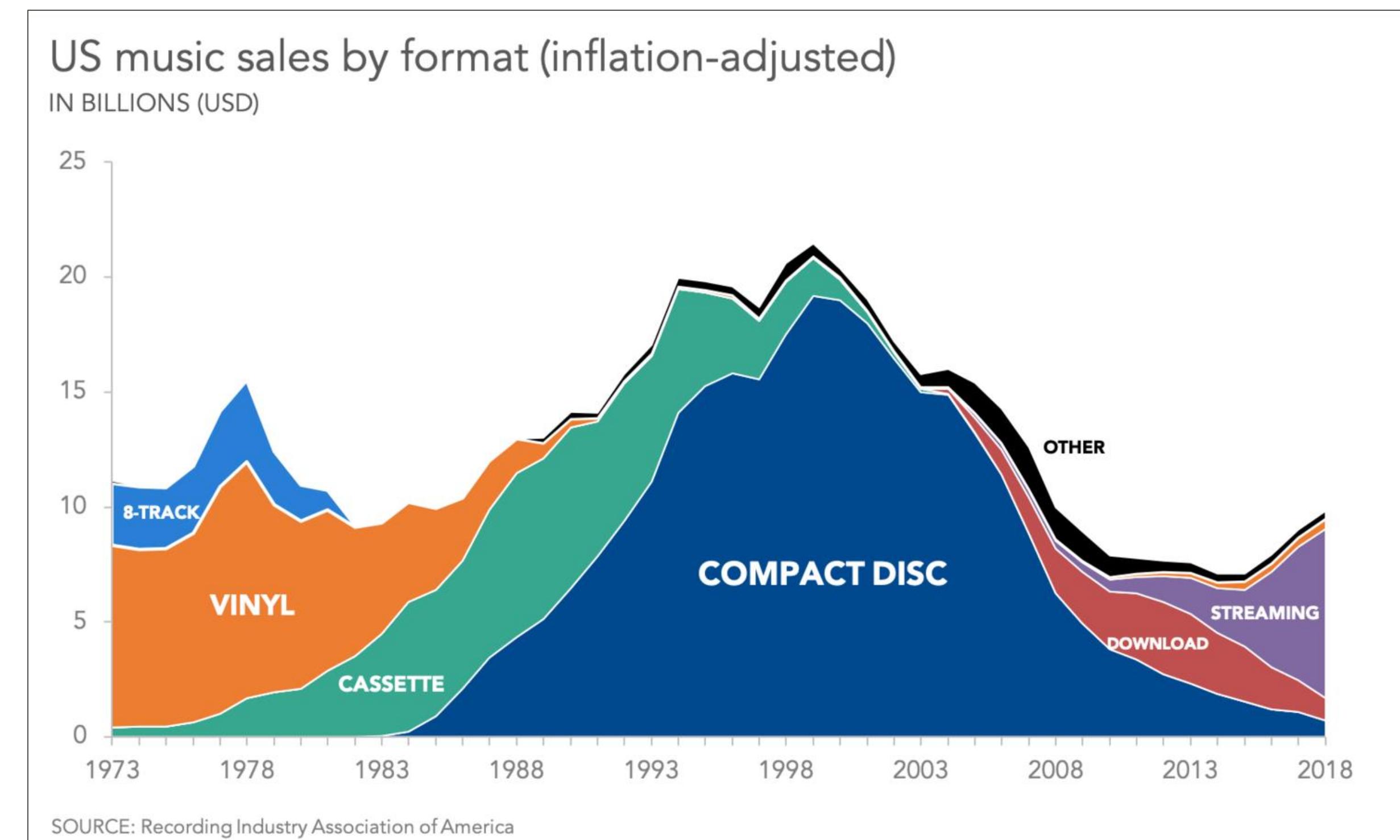
IN BILLIONS (USD)



SOURCE: Recording Industry Association of America

Stacked area charts

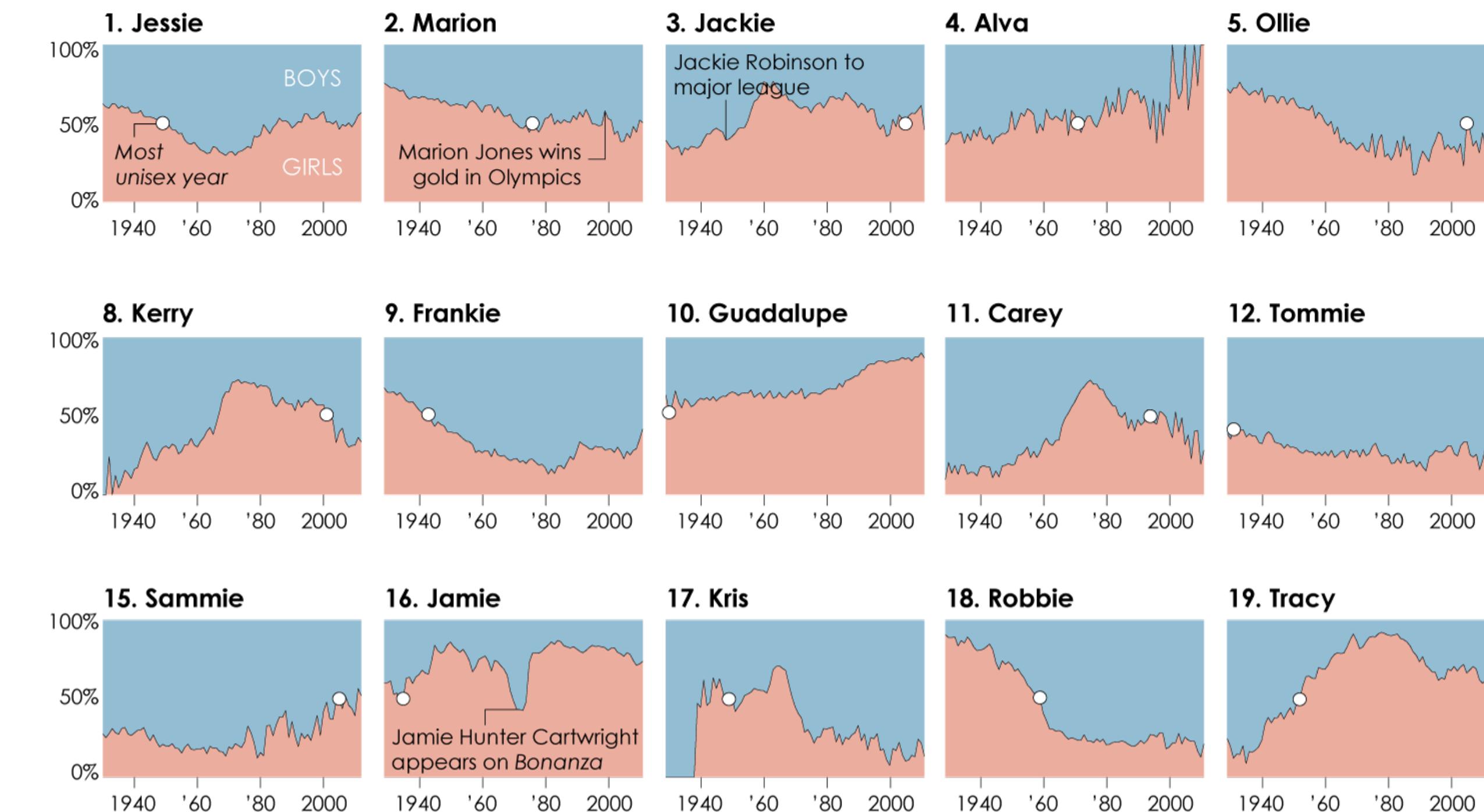
- Enfatiza continuidad horizontal vs. ítems verticales
- datos
 - 1 categ (soporte)
 - 1 ordinal key (tiempo)
 - 1 cuantitativo (USD)
- Datos derivados
 - geometría: capas, altura codifica el cuantitativo
 - 1 cuantitativo (orden de capas)
- Escalabilidad
 - Cientos de keys temporales
 - Docenas a cientos de keys verticales
 - Más que barras apiladas: No requiere espacio intermedio y muchas capas no ocupan todo el gráfico



Normalized stacked area charts

- Enfatiza continuidad horizontal vs. ítems verticales
- datos
 - 1 categ (sexo)
 - 1 ordinal key (tiempo)
 - 1 cuantitativo (%)
- Datos derivados
 - geometría: capas, altura codifica el cuantitativo
- Escalabilidad
 - Cientos de keys temporales
 - Docenas a cientos de keys verticales
 - Más que barras apiladas: No requiere espacio intermedio y muchas capas no ocupan todo el gráfico

The Most Unisex Names in US History



Nathan Yau

Normalized stacked area charts

- Permite comparaciones
- Óptimo cuando además resalta patrones específicos de los datos

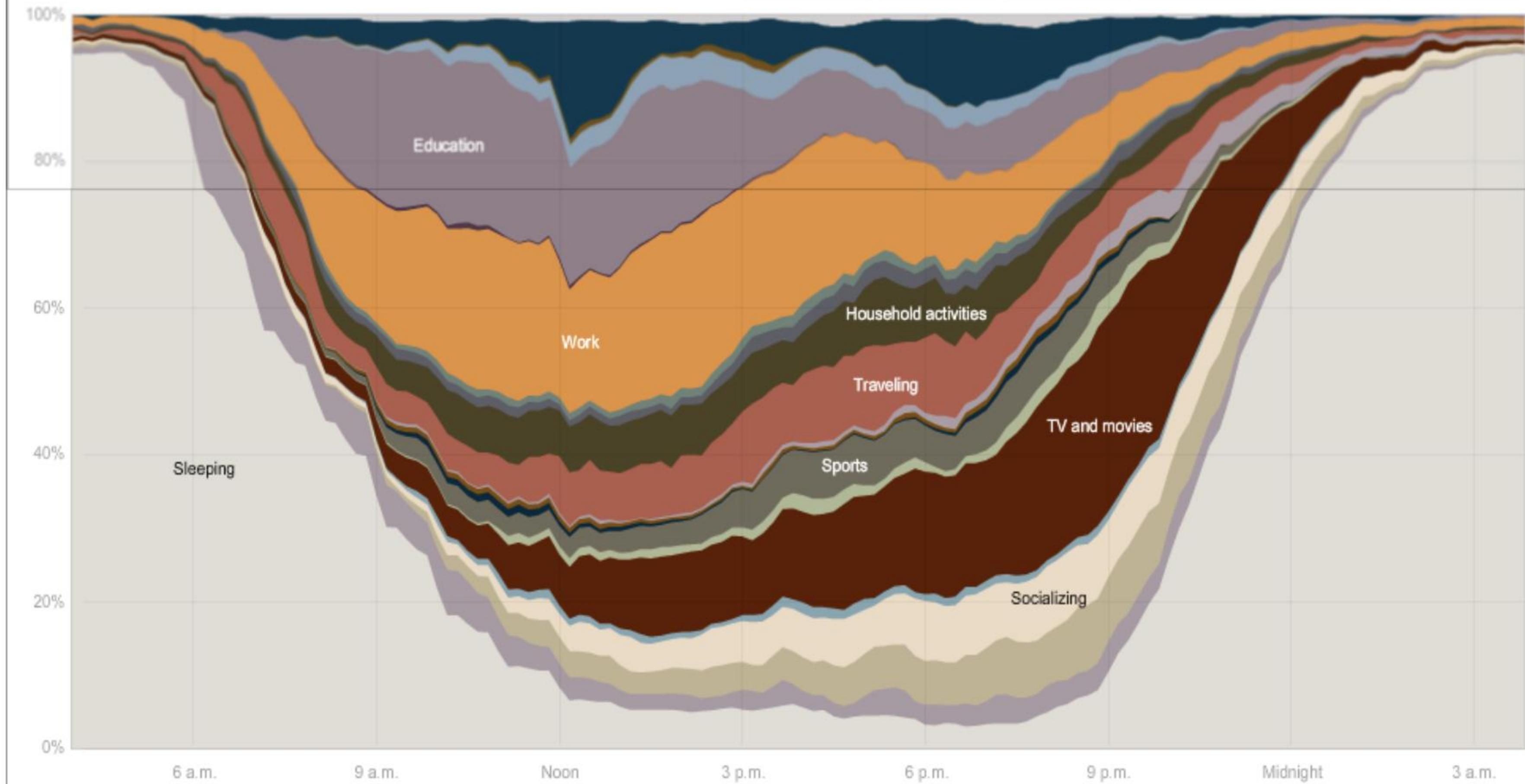
How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. [Related article](#)

People ages 15 to 24

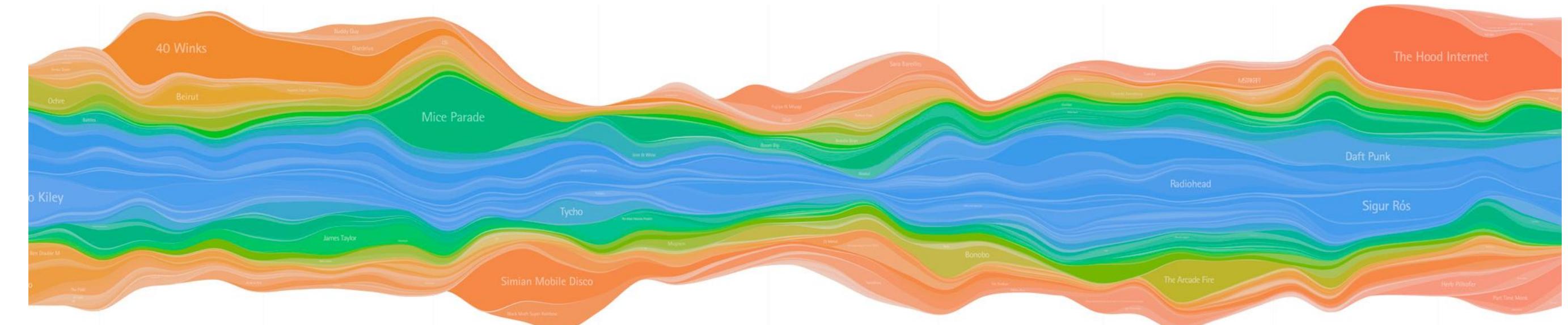
About half of this group is enrolled in school. While the young spend the most time on the telephone, they spend the least time on calls to family members.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab...	Hispanic	Age 65+	Advanced	Two+ children



Streamgraph

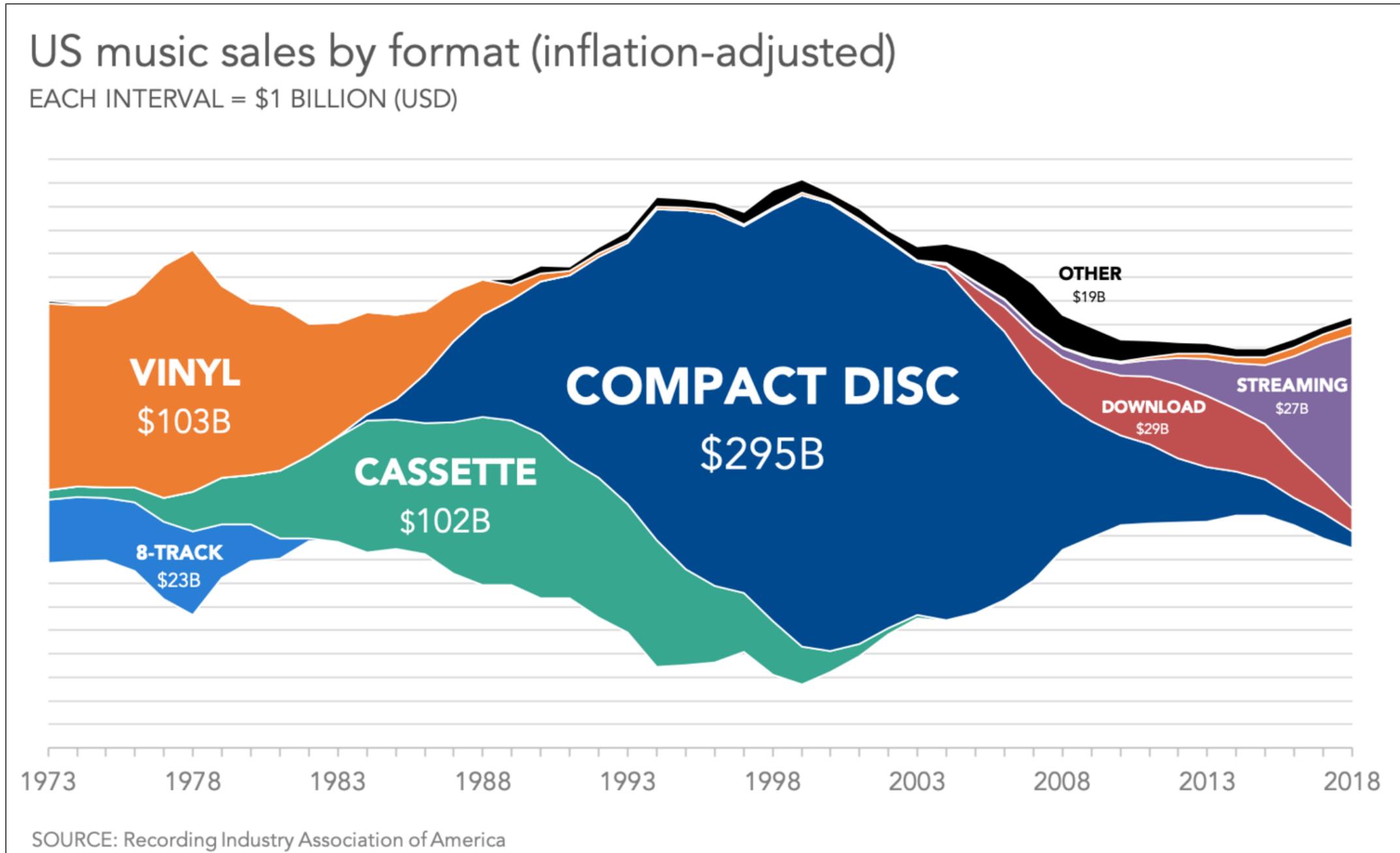
- Generalized stacked graph
 - Enfatiza continuidad horizontal vs ítems verticales
 - datos
 - 1 categ key (artista)
 - 1 ordenado key (tiempo)
 - 1 cuantitativo (counts)
 - Datos derivados
 - Geometría por capas en el tiempo
 - Altura de capas codifica el atr. cuant.
 - 1 cuantitativo (orden de capas)
 - Escalabilidad
 - hundreds of time keys
 - Docenas a cientos de keys
 - Más que barras apiladas: No requiere espacio intermedio y muchas capas no ocupan todo el gráfico



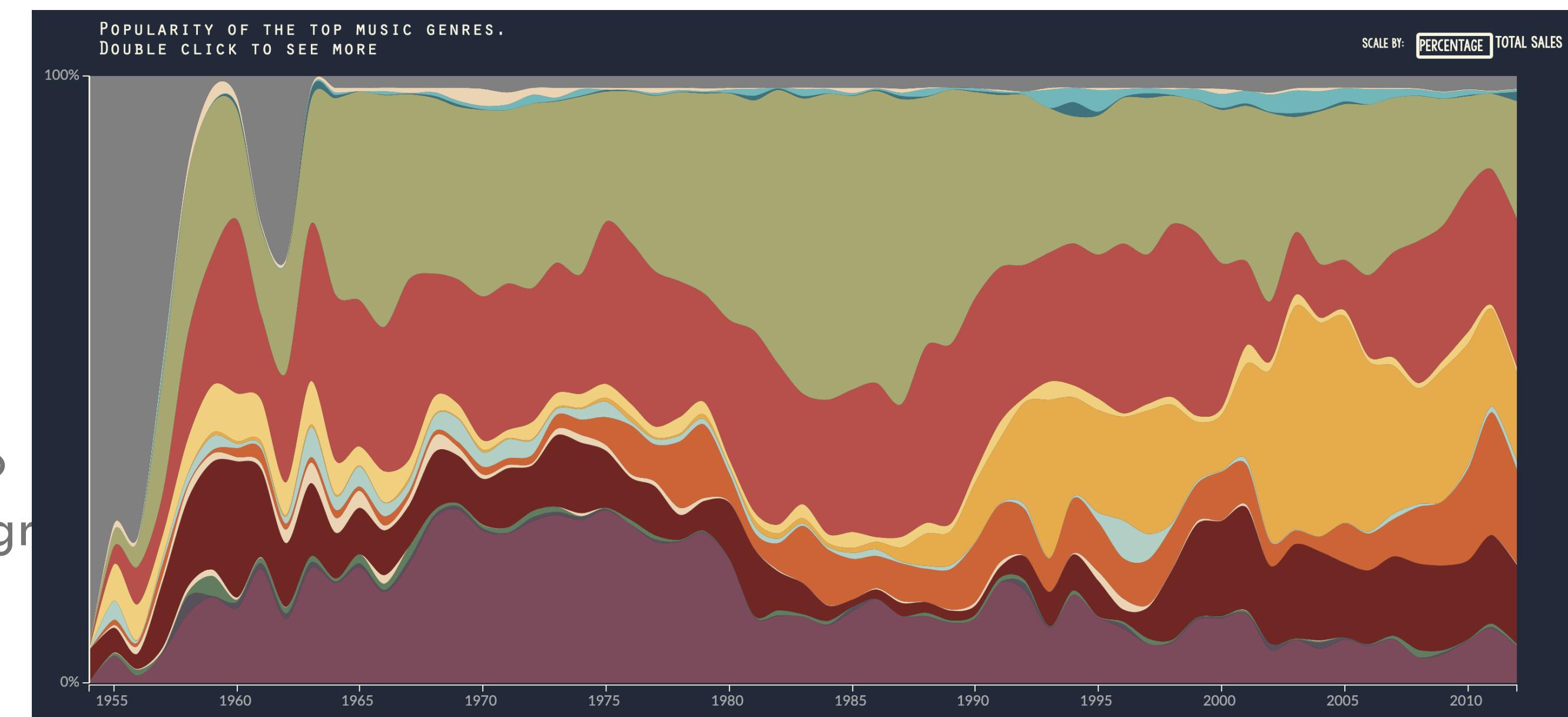
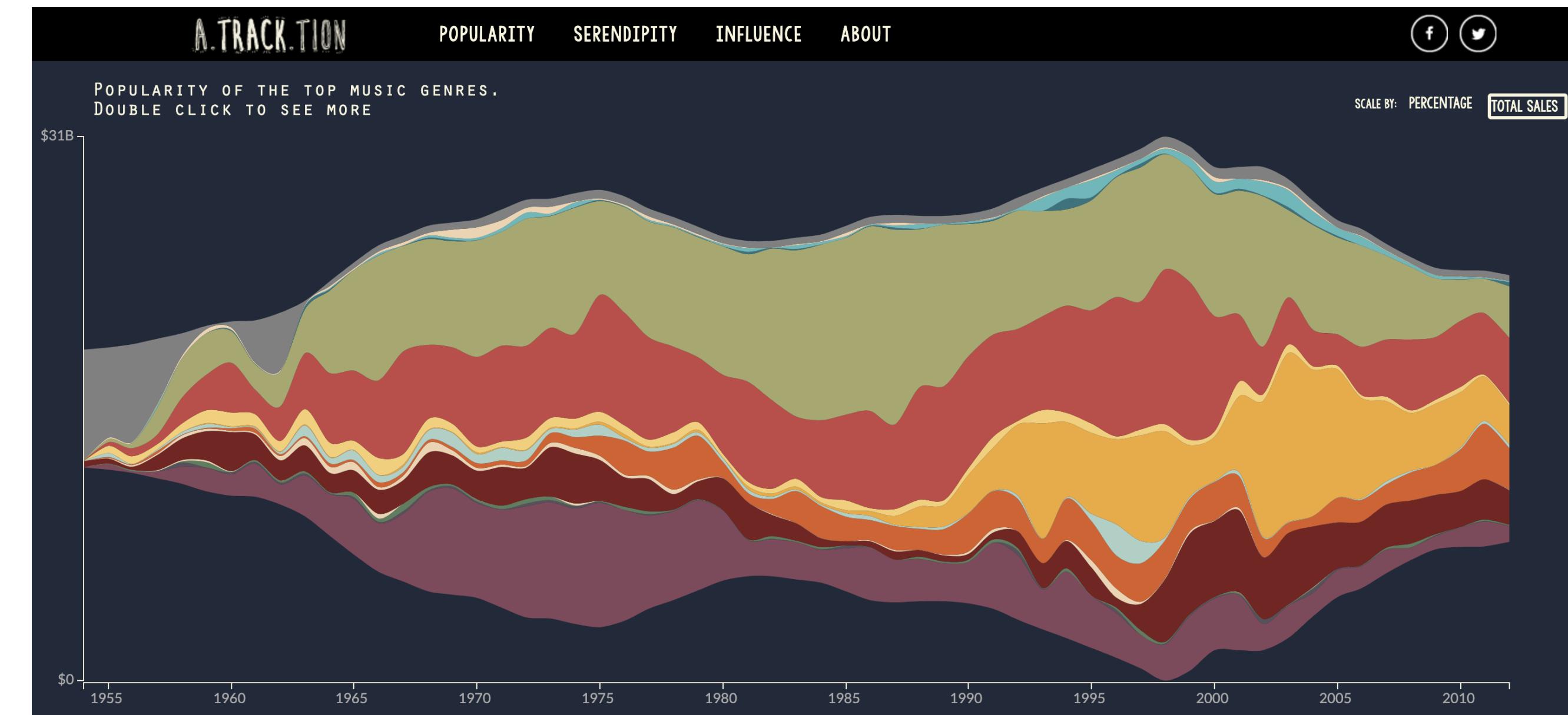
[Stacked Graphs Geometry & Aesthetics. Byron and Wattenberg. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2008) 14(6): 1245–1252, (2008).]

Streamgraph

- Generalized stacked graph
 - Enfatiza continuidad horizontal vs ítems verticales

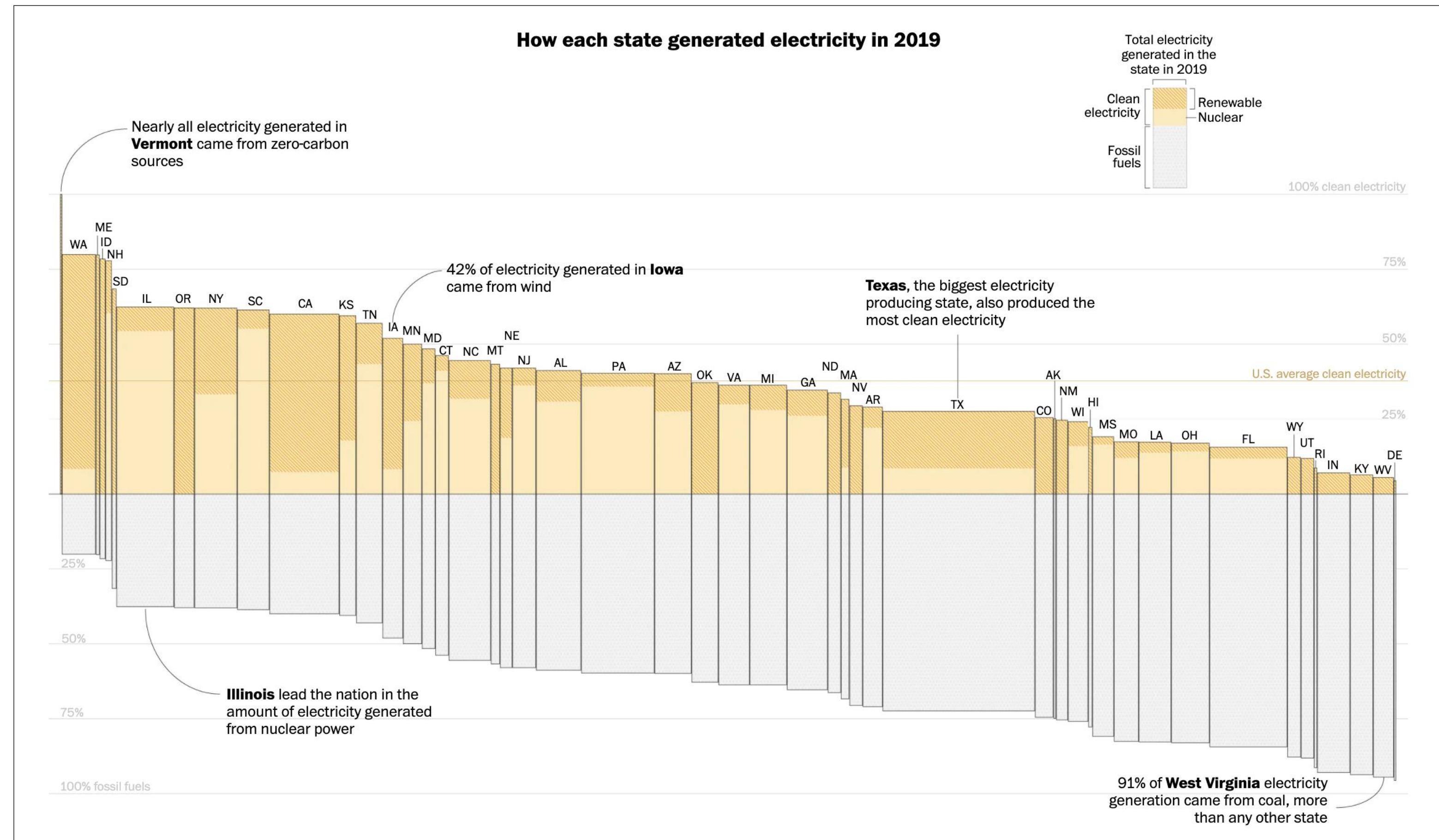


- Docenas a cientos de keys
- Más que barras apiladas: No requiere espacio intermedio y muchas capas no ocupan todo el gr

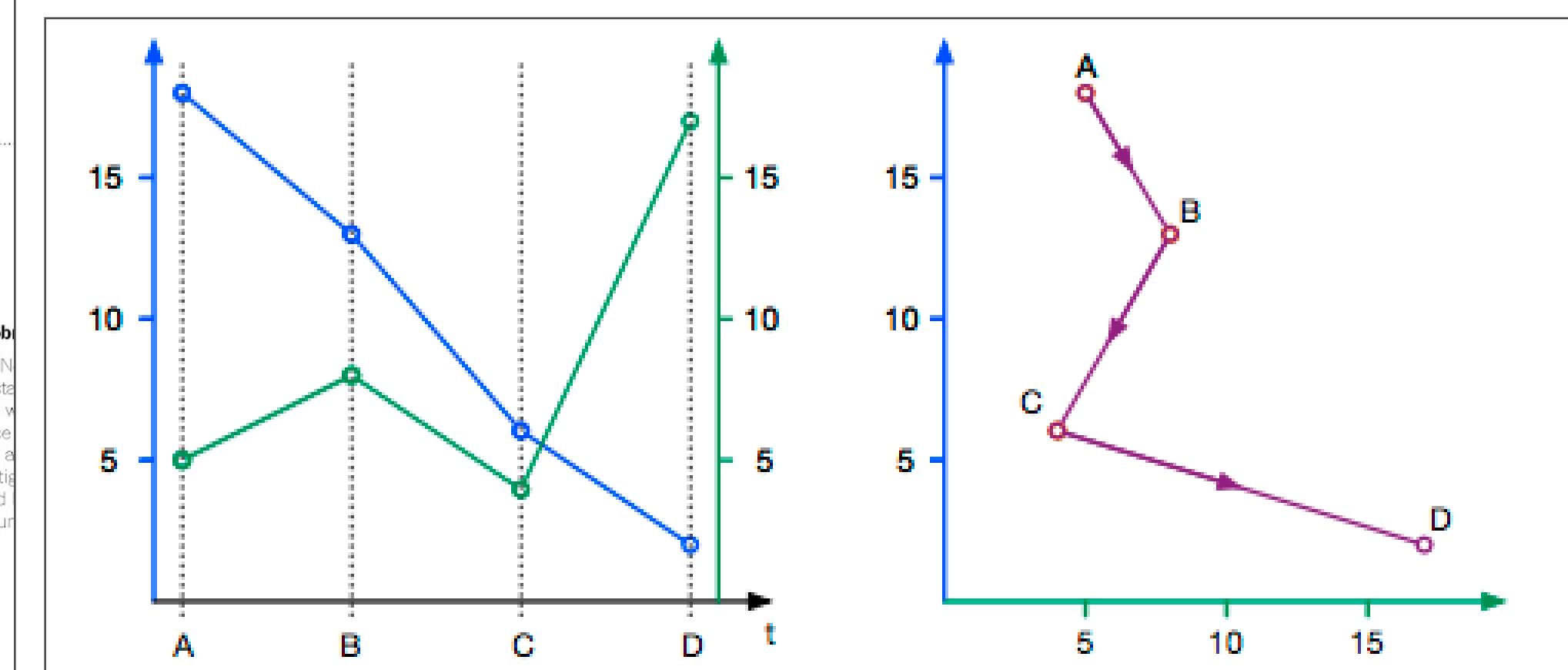
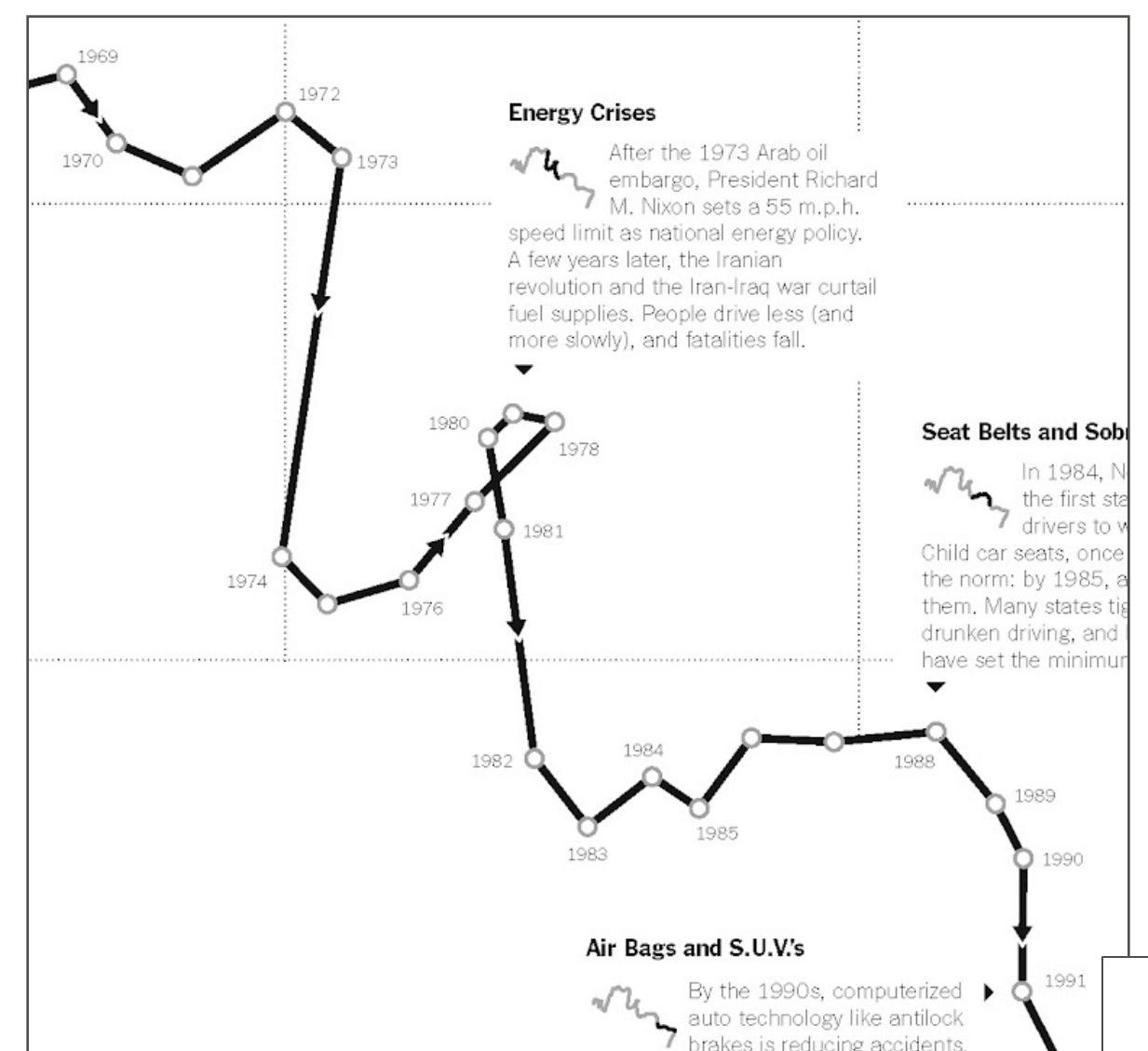


Mosaic plot

- Similar a stacked bar chart, pero puede incluir una variable más en el ancho de cada barra
- 2 categóricas: división horizontal y división dentro de cada glifo
- 2 cuantitativas: ancho de cada barra y altura (subdividida dentro de cada barra)
- Válido para relaciones parte-todo
- Requieren esfuerzo para interpretarlas, pero pueden mostrar relaciones difíciles de ver en otras gráficas



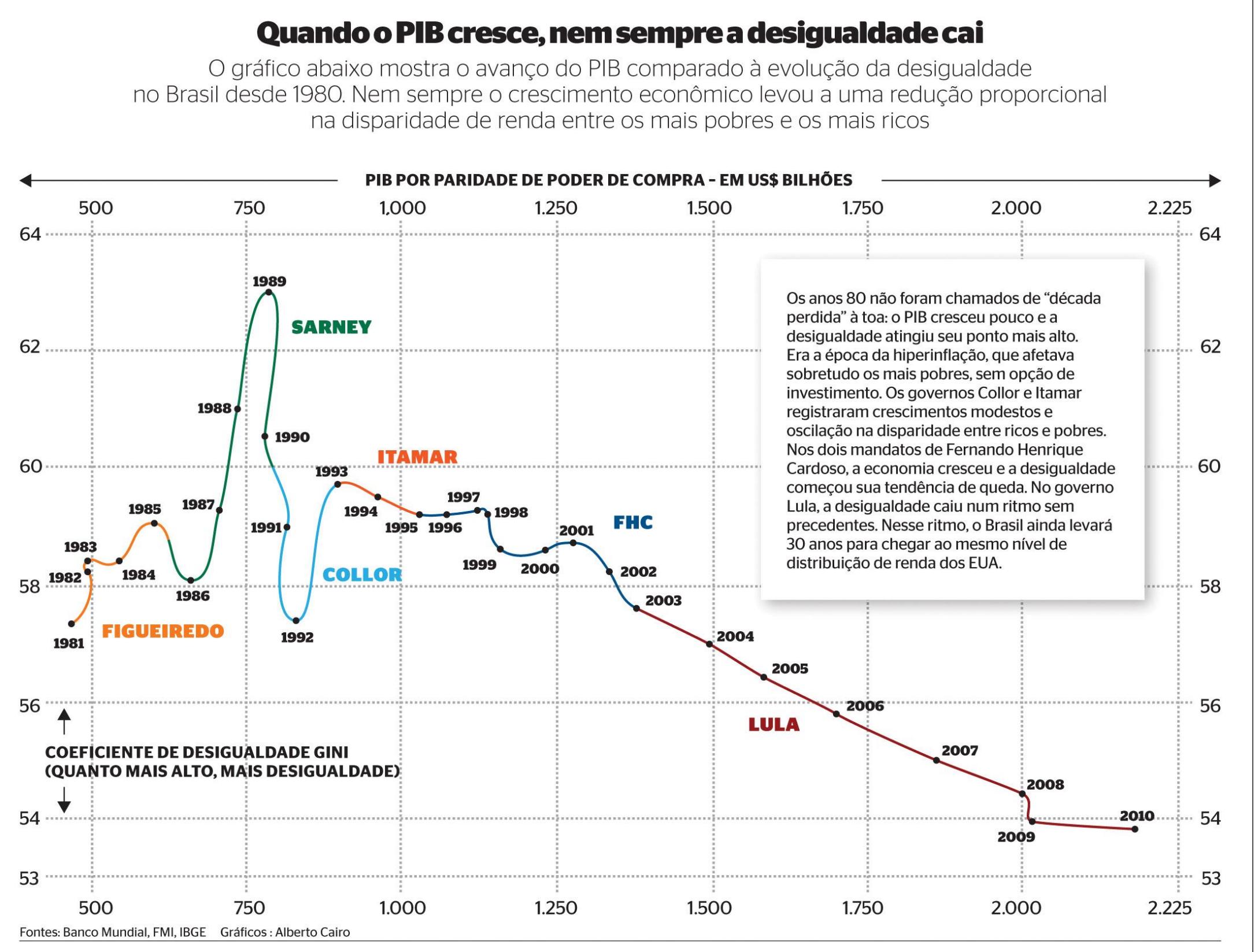
Connected scatterplots

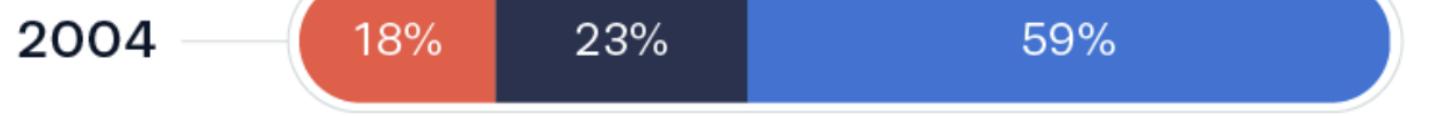


Scatterplot con marcas de conexión (líneas)

- Popular en periodismo
- Ejes horiz + vert: Values
- line connection marks: Orden temporal
- Estudio experimental:
 - Gráfica atractiva pero la correlación no queda clara

[The Connected Scatterplot for Presenting Paired Time Series.
Haroz, Kosara and Franconeri. IEEE TVCG 22(9):2174-86,
2016.]



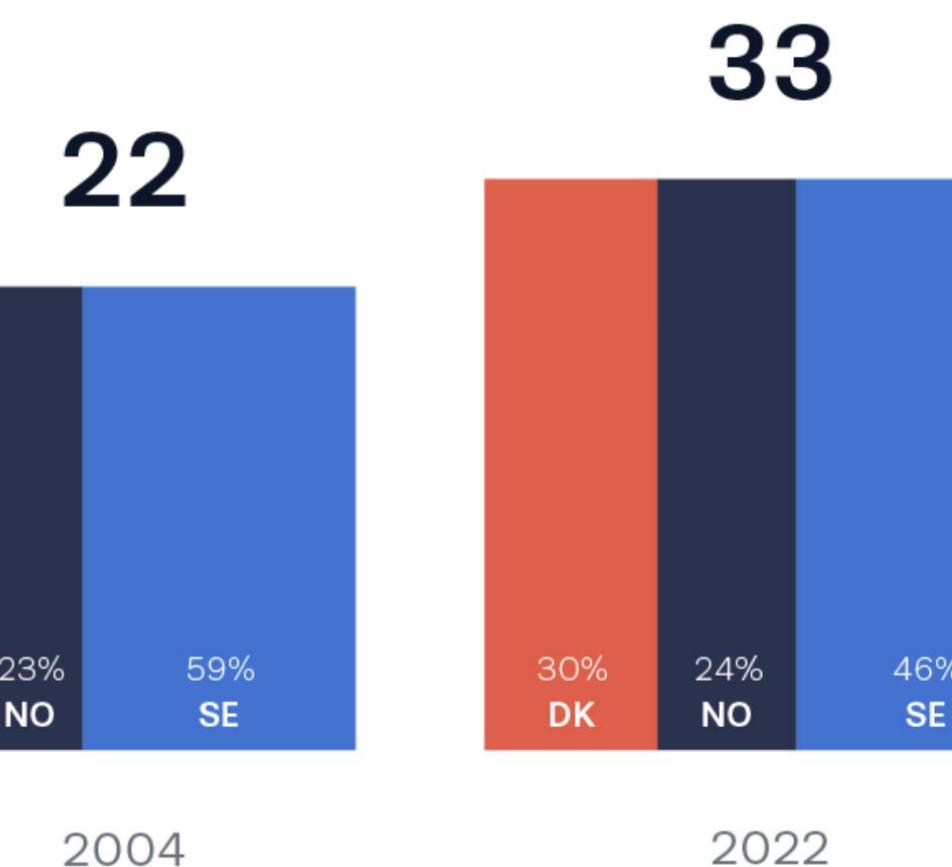
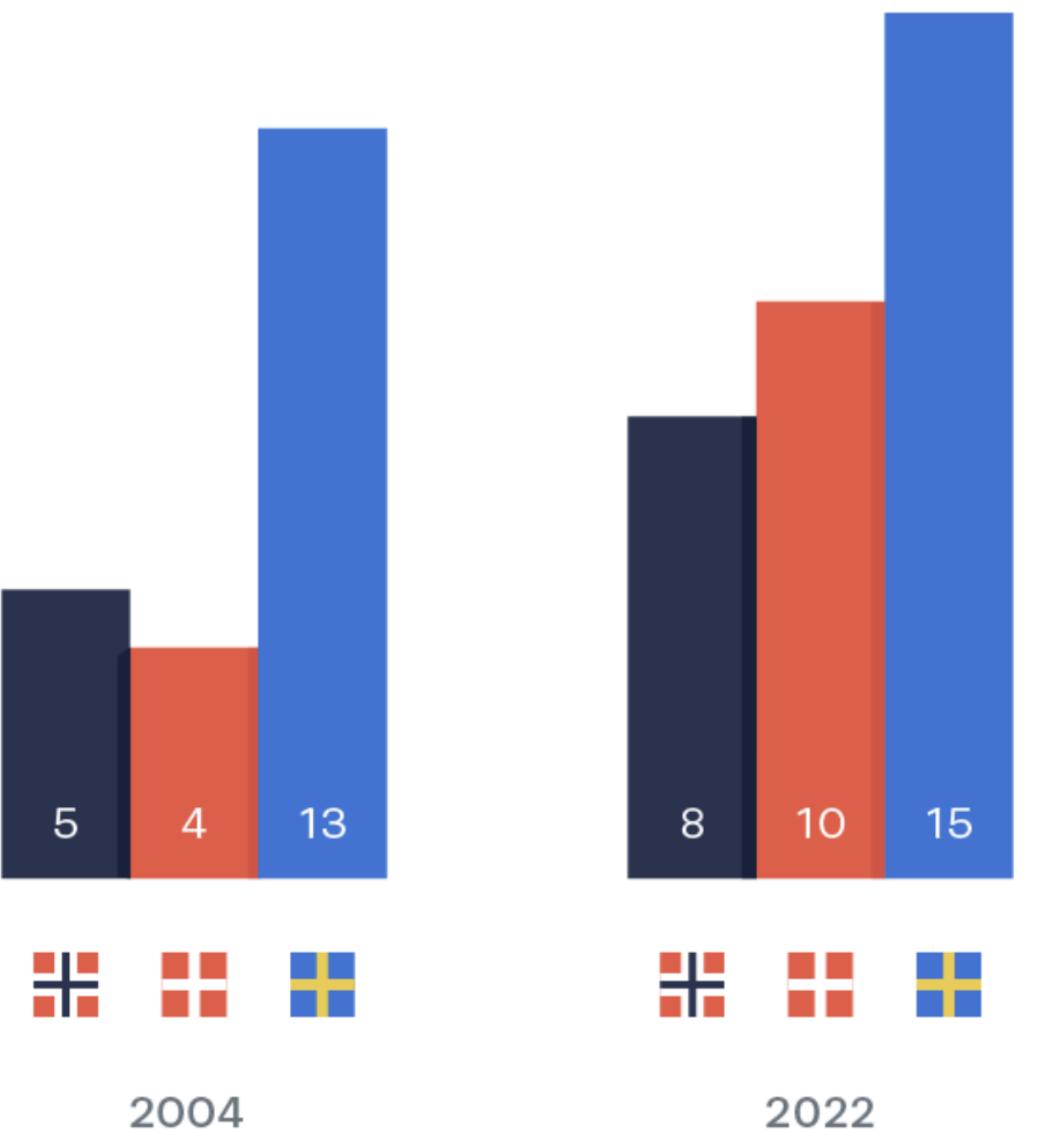


1 dataset 100 visualizations

Can we come up with 100 visualizations from one simple dataset?

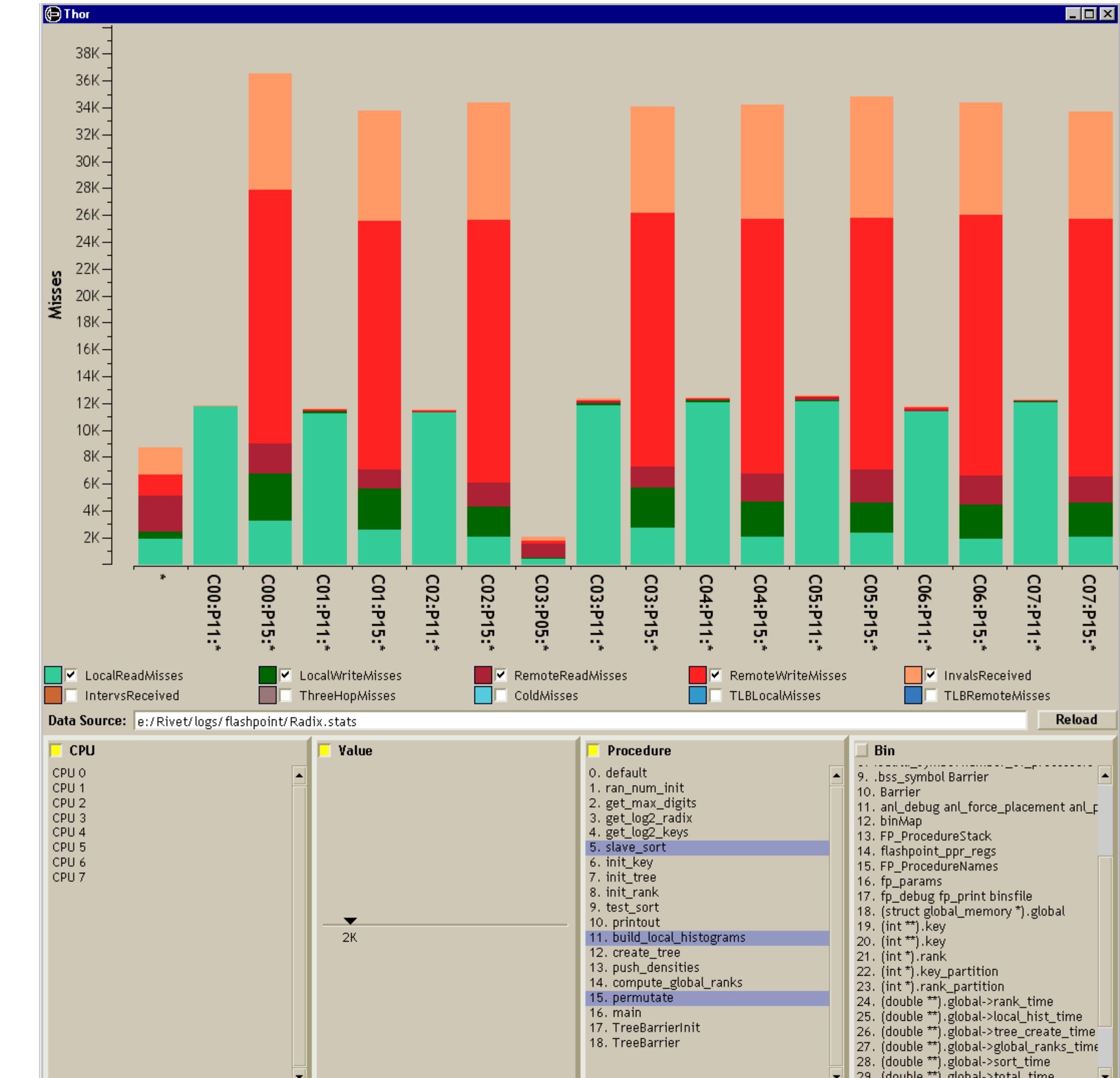
As an information design agency working with data visualization every day, we challenged ourselves to accomplish this using insightful and visually appealing visualizations.

We wanted to show the diversity and complexity of data visualization and how we can tell different stories using limited visual properties and assets.



Stacked bar chart

- 2 keys, 1 value
- **Datos: 2 categórico y 1 cuantitativo**
- Marcas: Lineas apiladas verticalmente
- **Glifos:** Objetos compuestos por múltiples marcas:
 - Canales: longitud y Tono
- Regiones:
 - Una por glifo
 - Alineadas:
 - Componente más bajo
 - Otros componentes del glifo
- Tareas:
 - Relaciones parte-todo
 - Escalable a <docena de niveles para el atributo apilado; igual a barras para regiones



[Using Visualization to Understand the Behavior of Computer Systems. Bosch. Ph.D. thesis,
Stanford Computer Science, 2001.]

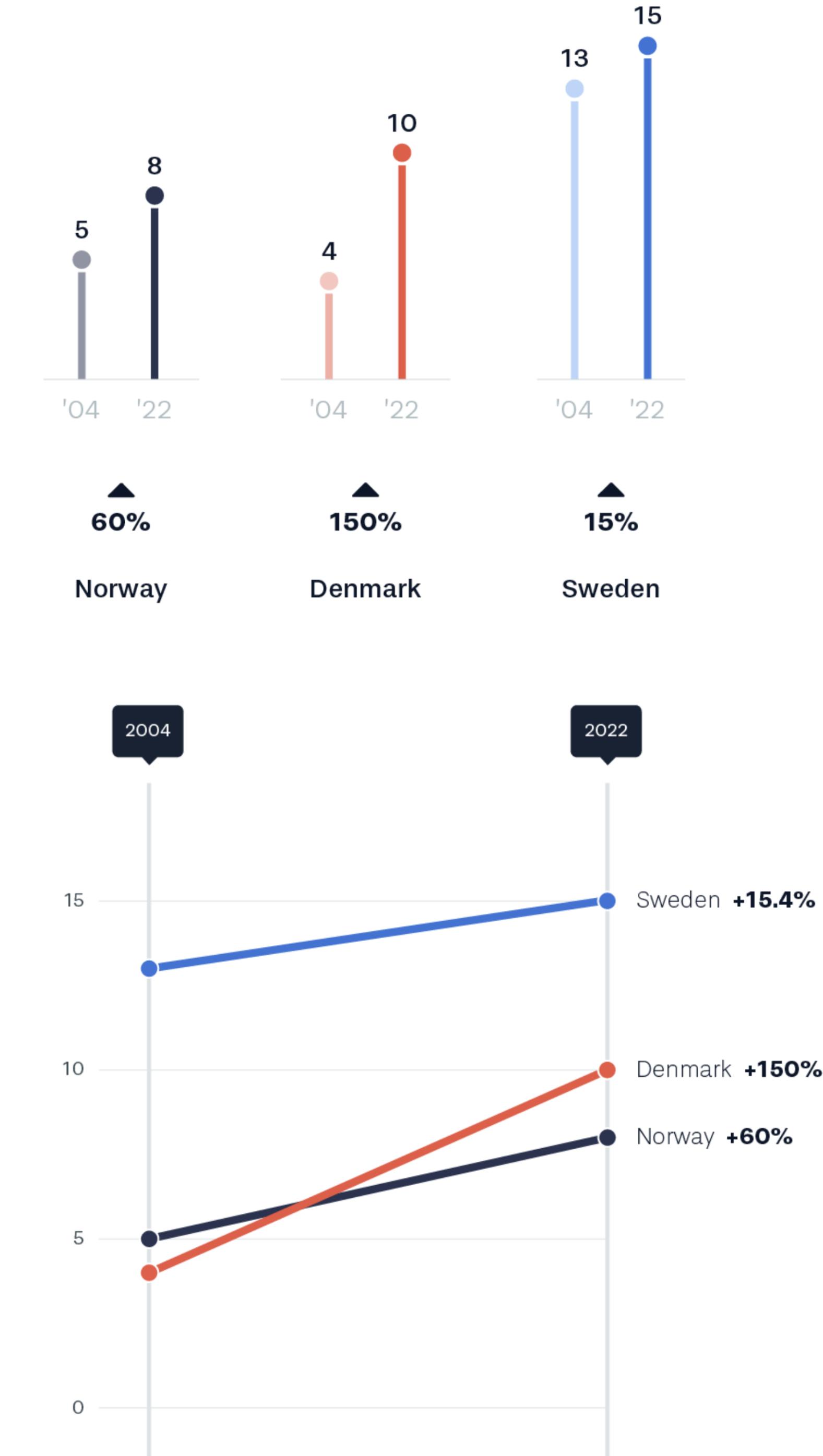
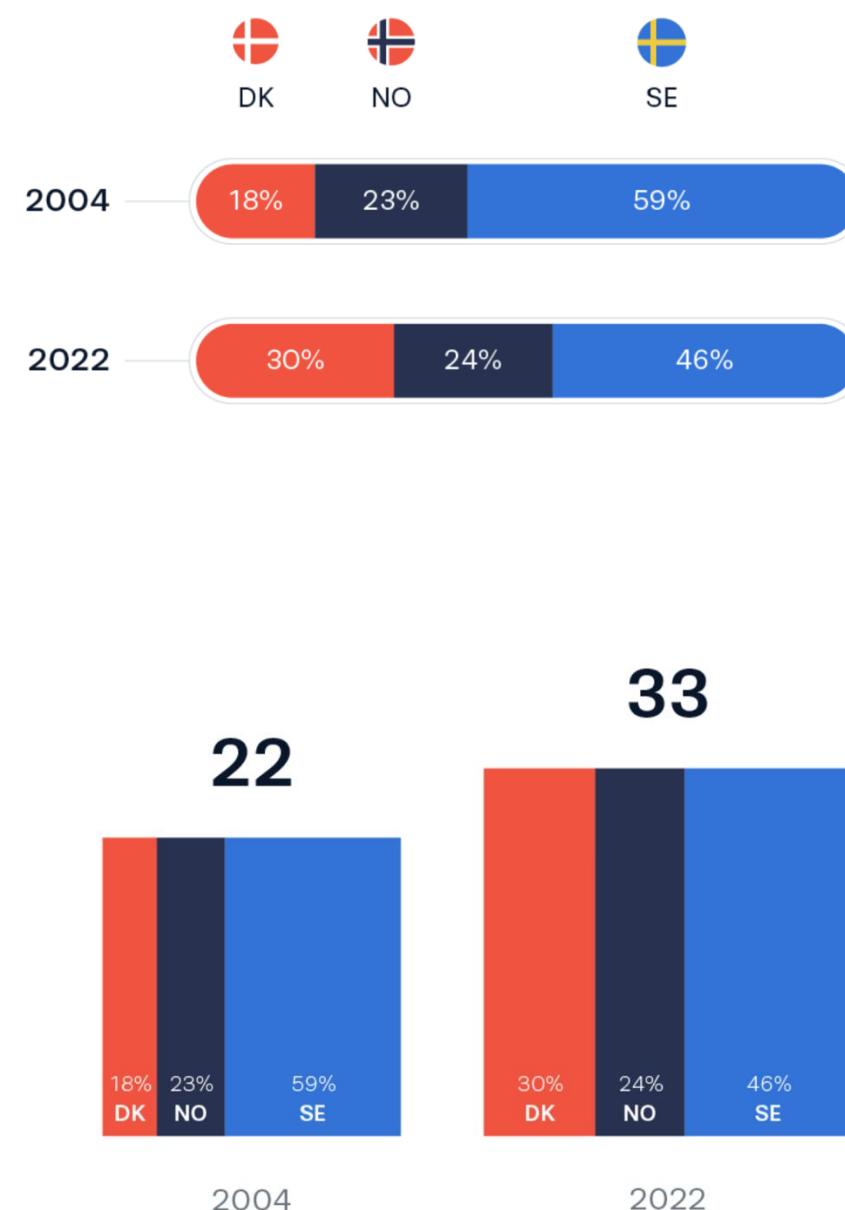
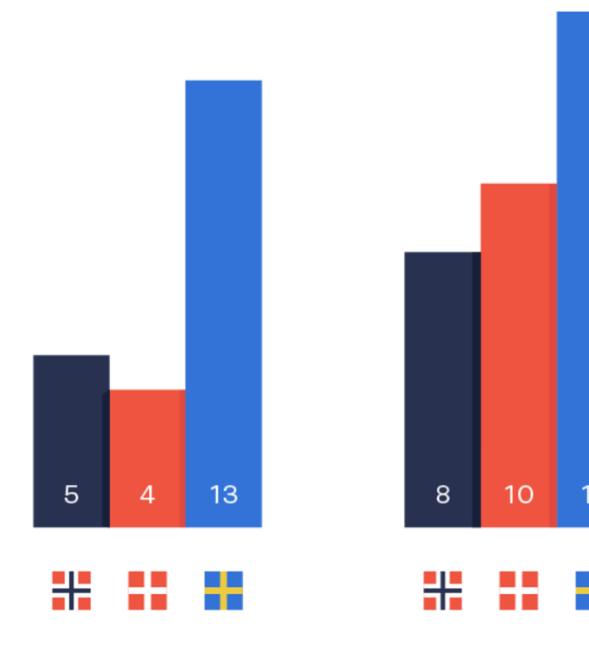
From: Visualization Analysis and Design

1 dataset 100 visualizations

Can we come up with 100 visualizations from one simple dataset?

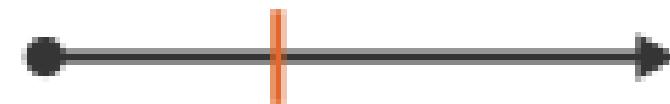
As an information design agency working with data visualization every day, we challenged ourselves to accomplish this using insightful and visually appealing visualizations.

We wanted to show the diversity and complexity of data visualization and how we can tell different stories using limited visual properties and assets.



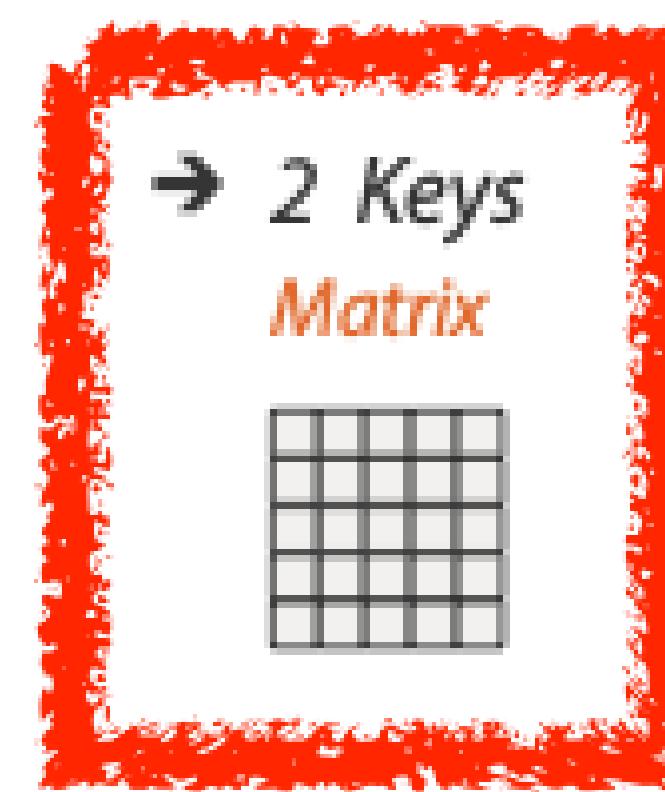
2 Keys

→ Express Values



→ 1 Key

List

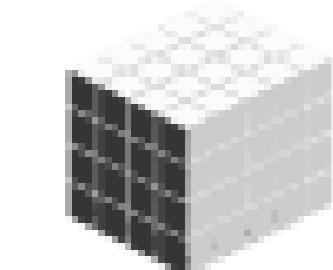


→ 2 Keys

Matrix

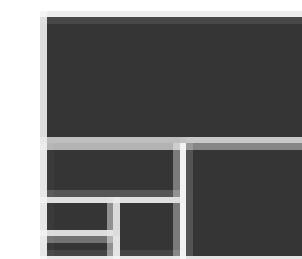
→ 3 Keys

Volume



→ Many Keys

Recursive Subdivision

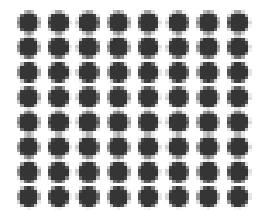


Heatmap

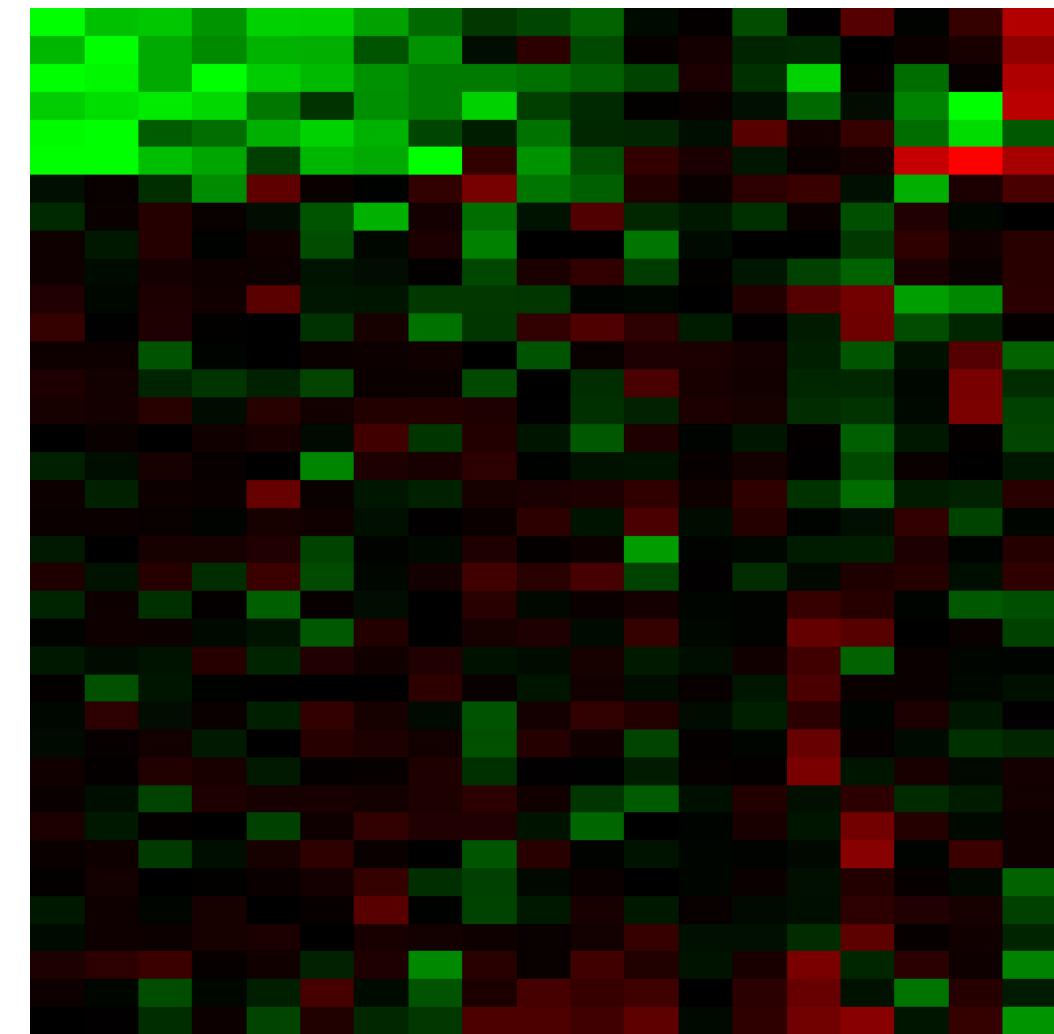
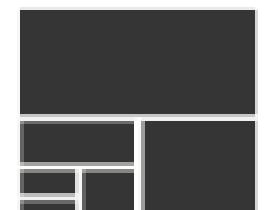
- Dos keys, un valor
- Datos
 - 2 categóricos (gen, condición experimental)
 - 1 cuantitativo (niveles de expresión)
- Marcas: Áreas
 - Separar y alinear en matriz 2D
 - Indexado por los 2 atributos categóricos
- Canales
 - color por cuant.
 - Mapa de color divergente ordenado
- Tarea
 - Identificar + clusters y outliers
- Escalabilidad
 - 1M items, 100s de niveles categ., ~10 niveles en el attr. quant

④ Layout Density

→ Dense



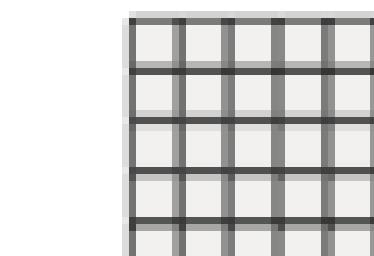
→ Space-Filling



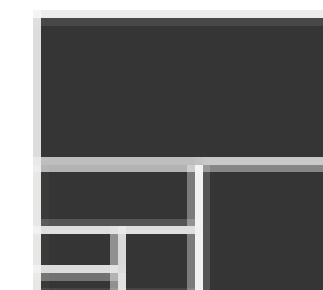
→ 1 Key
List



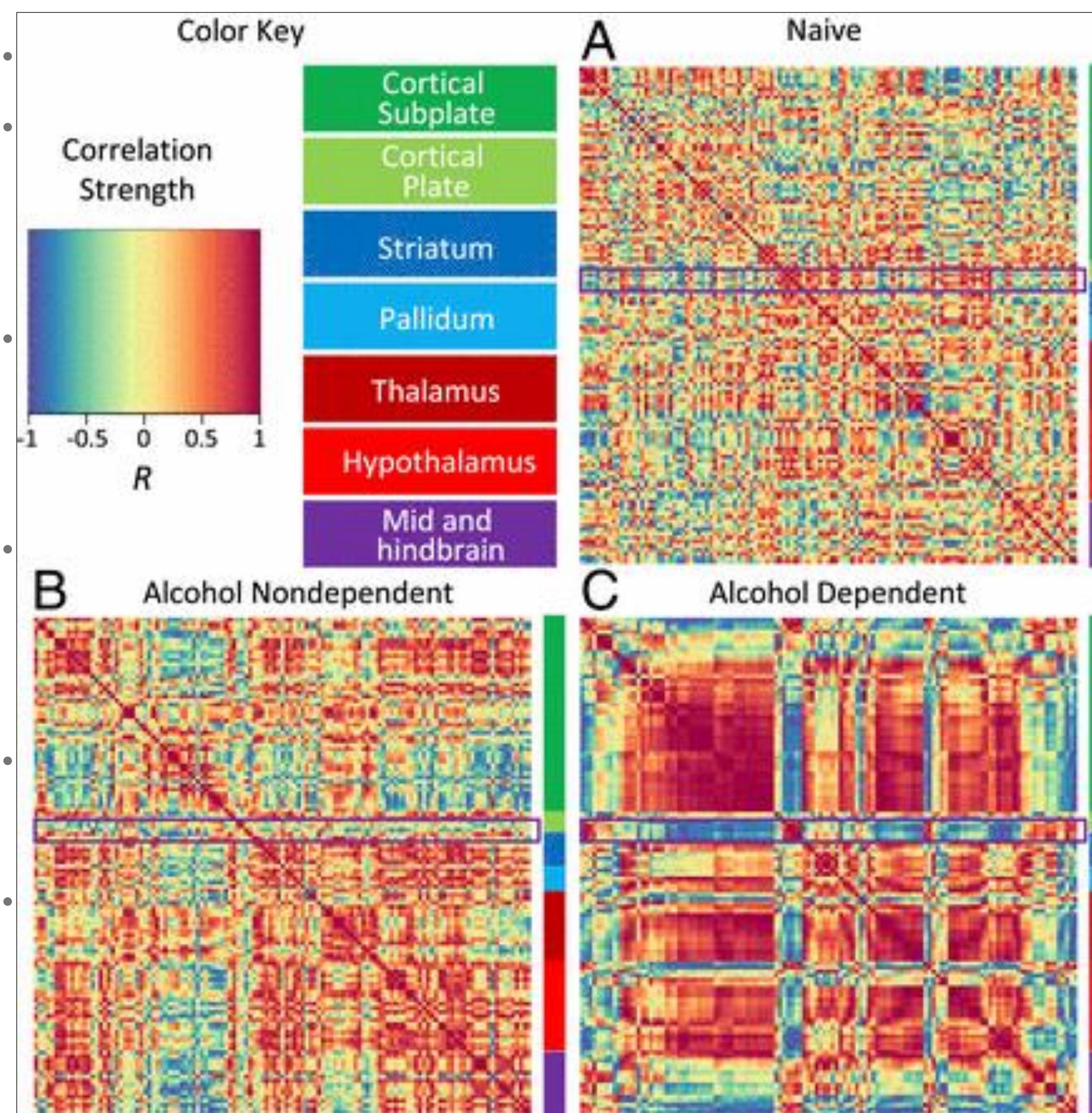
→ 2 Keys
Matrix



→ Many Keys
Recursive Subdivision

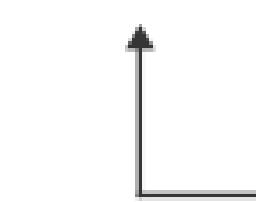


Heatmap

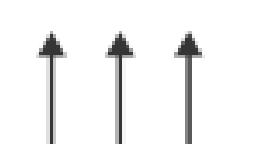


④ Axis Orientation

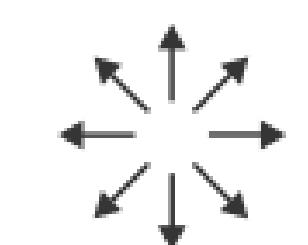
→ Rectilinear



→ Parallel

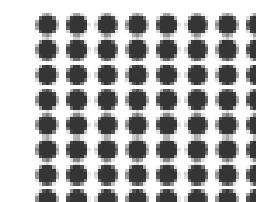


→ Radial

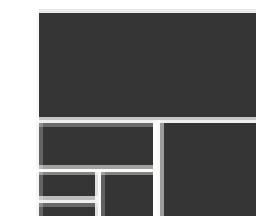


④ Layout Density

→ Dense



→ Space-Filling



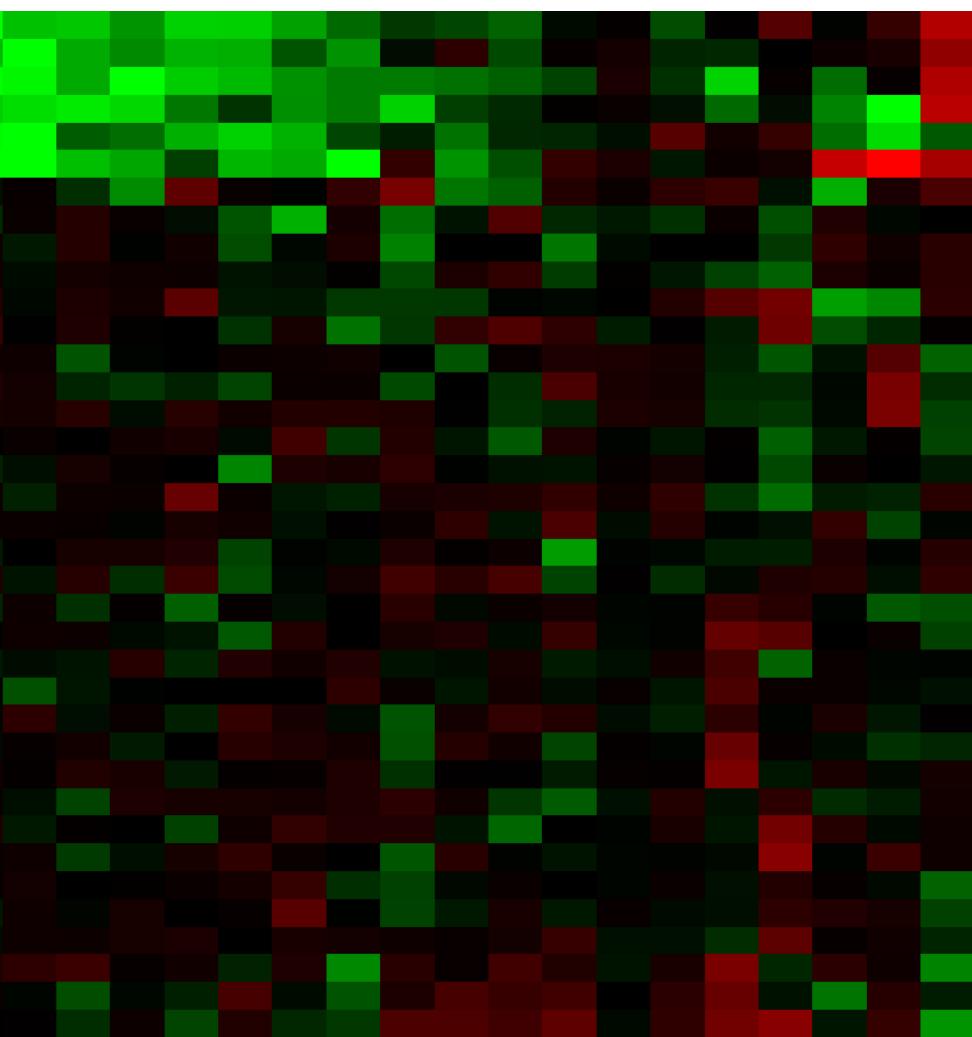
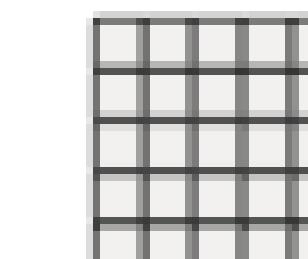
→ 1 Key

List



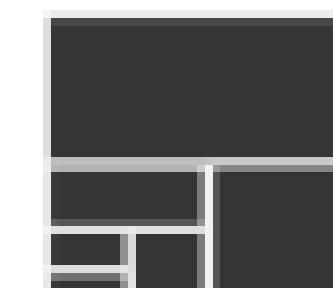
→ 2 Keys

Matrix



→ Many Keys

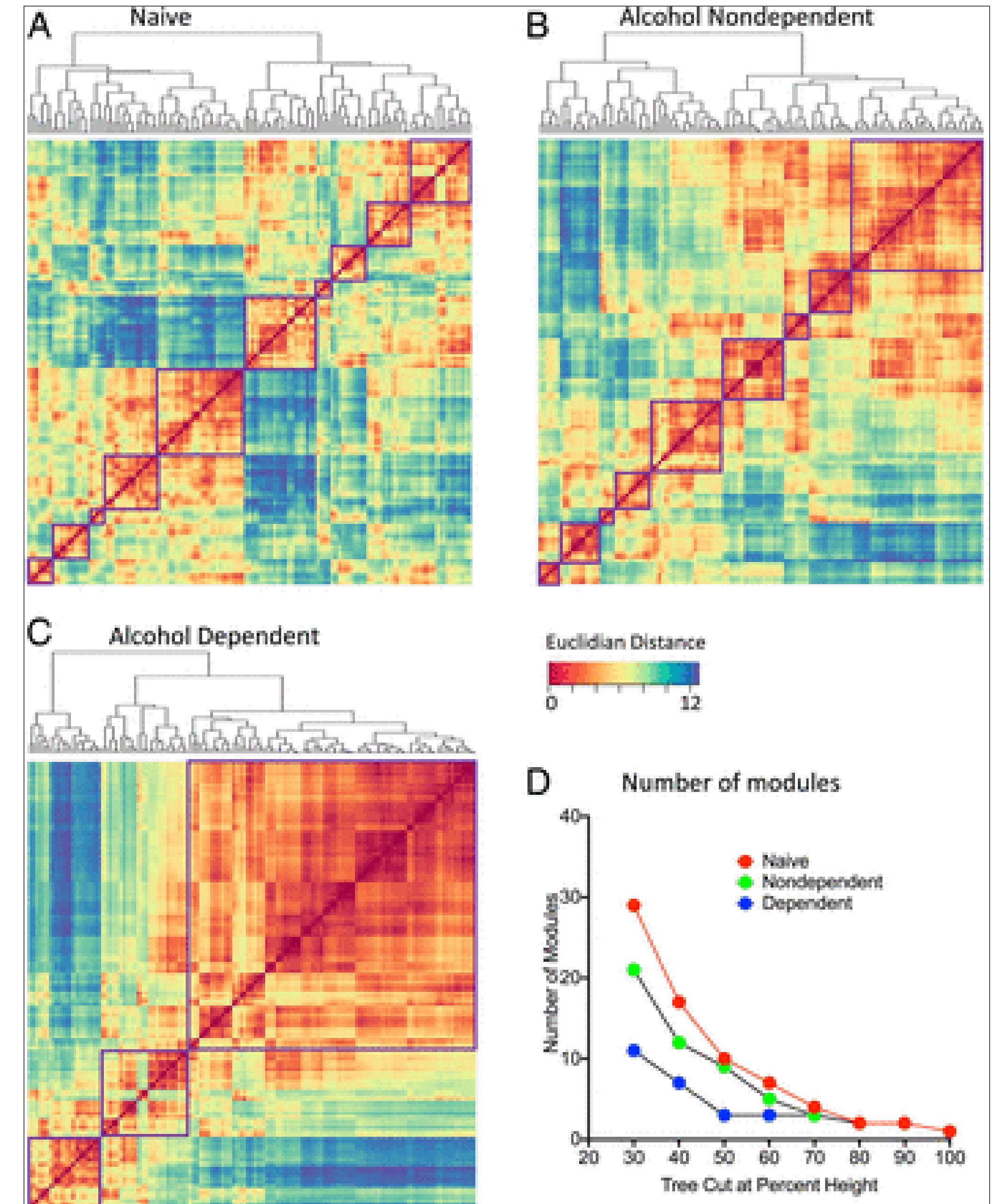
Recursive Subdivision



Cluster heatmap

Además de lo anterior:

- Datos derivados: Clusters jerárquicos
- Dendrogramas
 - Relaciones jerárquicas en árbol conectadas por líneas
 - Alineados por las hojas del árbol para comparar mejor la longitud de las ramas
- Heatmap
 - Marcas (re-)ordenadas según clusters
- Tareas:
 - Evaluar + calidad de clusters automáticos.
 - Detectar + nuevos patrones en la correlación



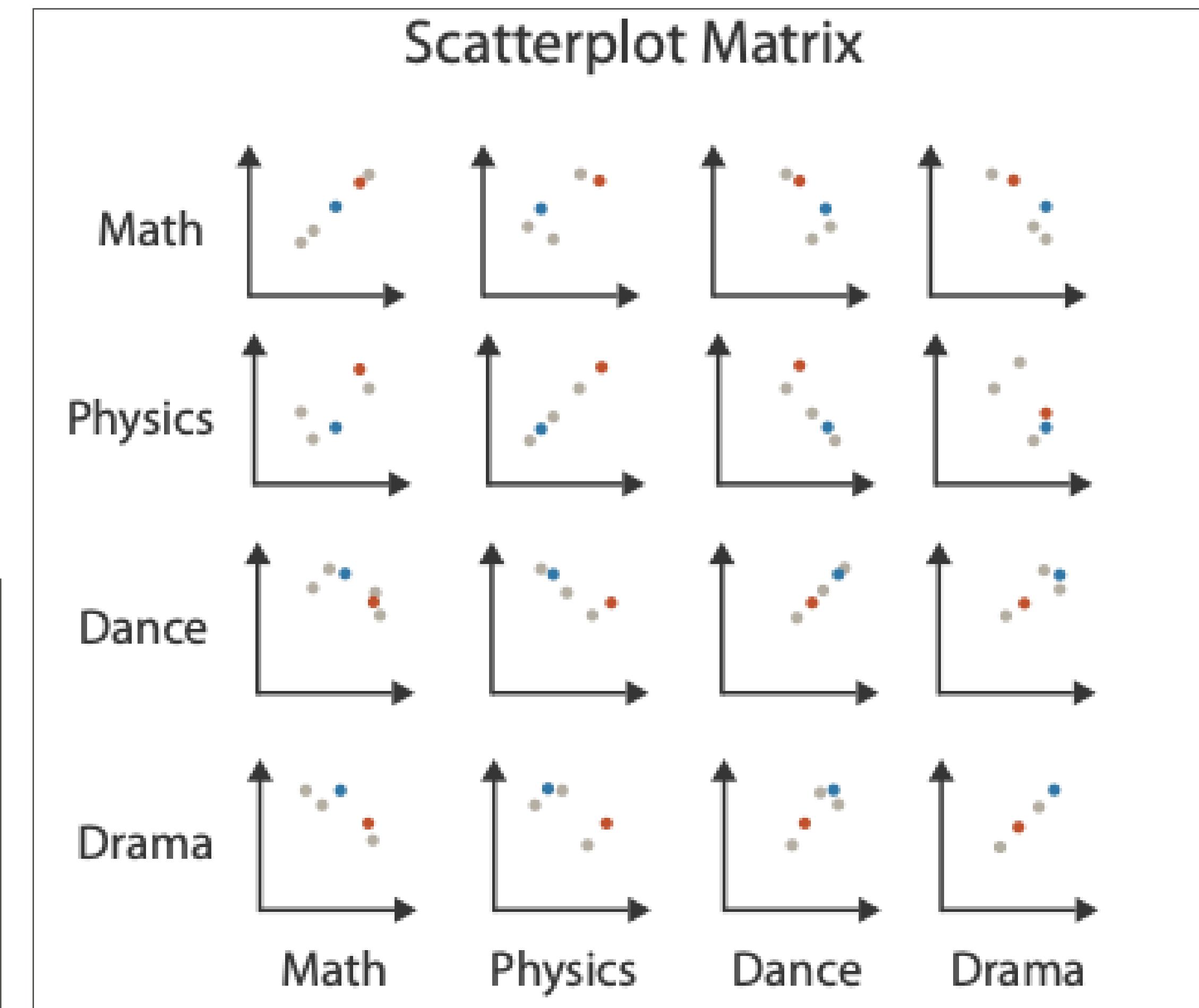
Brain-wide functional architecture remodeling by alcohol dependence and abstinence. Kimbrough, A., et al. (2020)

Scatterplot matrix

Scatterplot matrix (SPLOM)

- Matriz densa donde cada celda contiene una gráfica entera.
- Muestra todas las combinaciones posibles de pares de atributos
- Ejes rectilíneos, marcas: puntos
- Tareas:
 - Detectar + tendencias, correlaciones, outliers
- Escalabilidad:
 - Una docena de atributos
 - Docenas a cientos de items

Table				
	Math	Physics	Dance	Drama
	85	95	70	65
	90	80	60	50
	65	50	90	90
	50	40	95	80
	40	60	80	90



after [Visualization Course Figures. McGuffin, 2014. <http://www.michaelmcguffin.com/courses/vis/>]

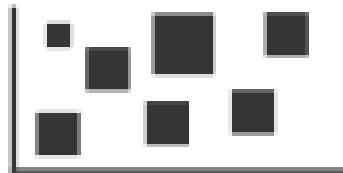
Arrange tables

④ Express Values

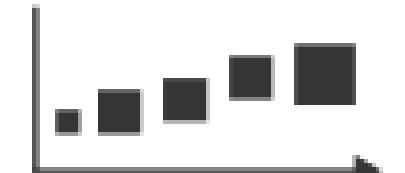


④ Separate, Order, Align Regions

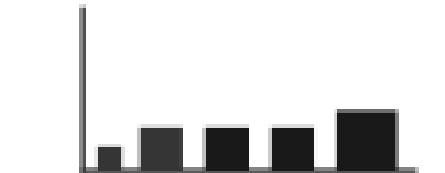
→ Separate



→ Order



→ Align



- Existen distintos tipos de tablas según el número de claves
- Key=Clave primaria- identificador único de cada observación

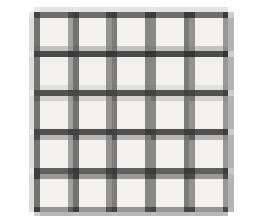
→ 1 Key

List



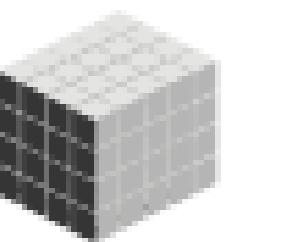
→ 2 Keys

Matrix



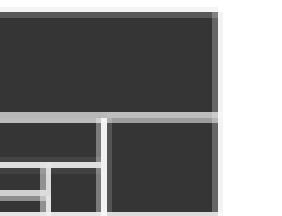
→ 3 Keys

Volume



→ Many Keys

Recursive Subdivision

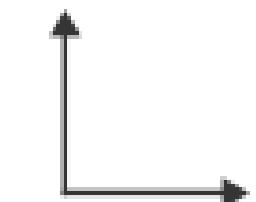


Para ordenar los datos también se tiene en cuenta:

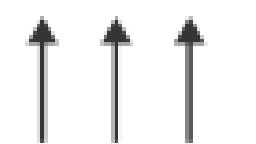
- Orientación de los ejes
- Densidad del Layout: Densa o exhaustiva

④ Axis Orientation

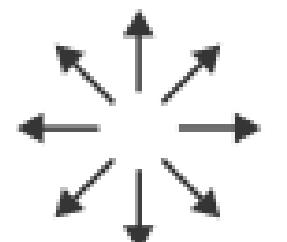
→ Rectilinear



→ Parallel

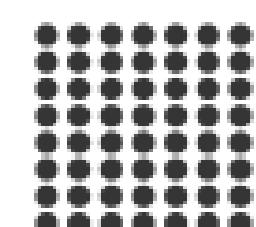


→ Radial

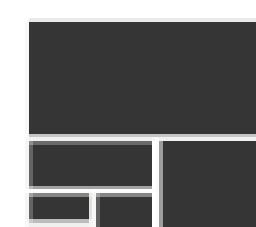


④ Layout Density

→ Dense



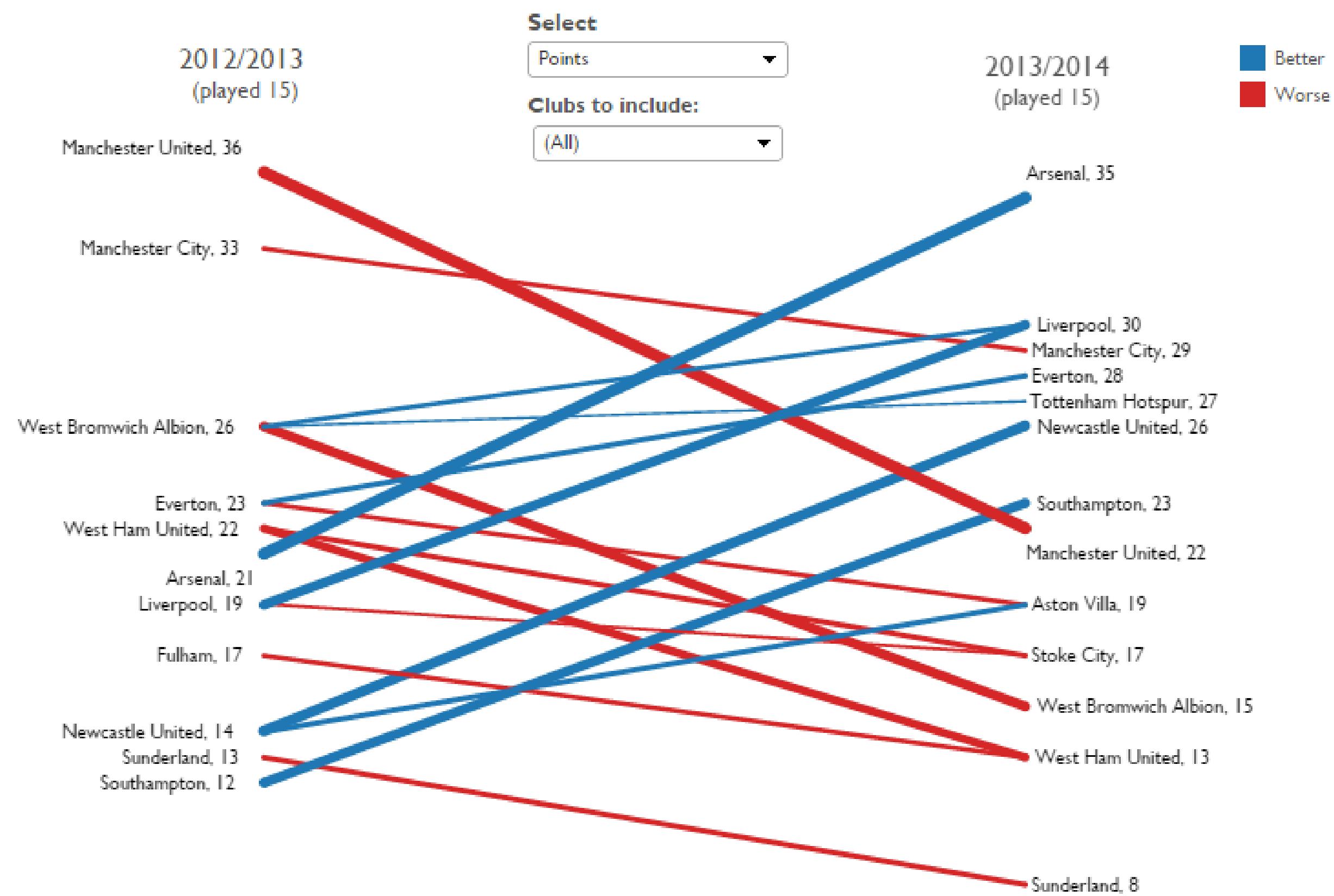
→ Space-Filling



Slopegraphs

- Resalta el cambio entre valores de un ranking
- Datos
 - 2 valores cuantitativos
 - (1 atributo derivado: cambio de magnitud)
- Marcas: puntos y líneas de conexión
- Canales:
 - 2 pos. vertical: Expresar valores
 - linewidth/tamaño, color
- Tarea:
 - Resaltar cambios en un ranking
- Escalabilidad:
 - Cientos de niveles

Barclay's Premier League Tables: Comparing 2012/2013 Starts to 2013/2014 Starts



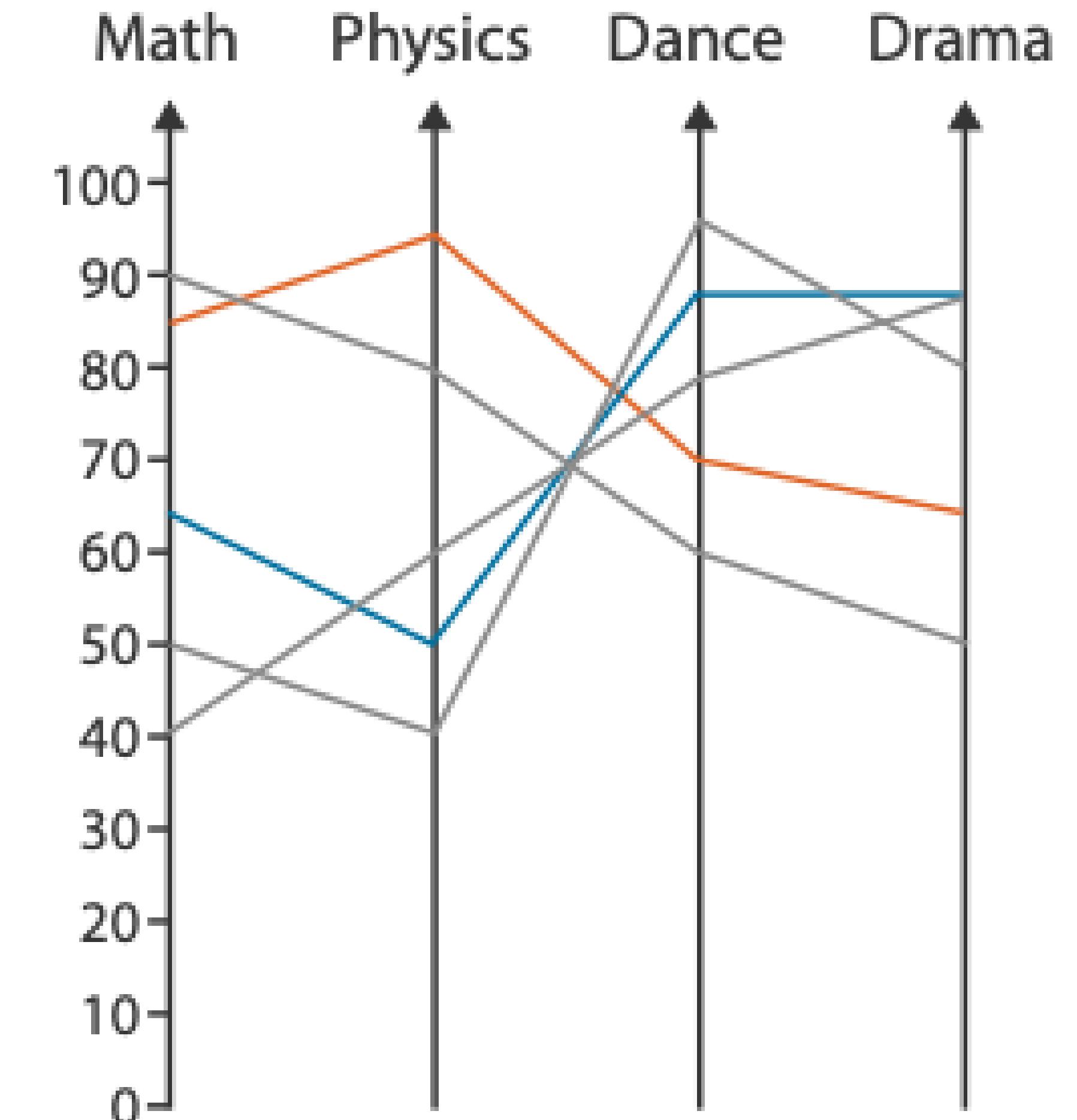
<https://public.tableau.com/profile/ben.jones#!/vizhome/Slopegraphs/Slopegraphs>

Coordenadas paralelas

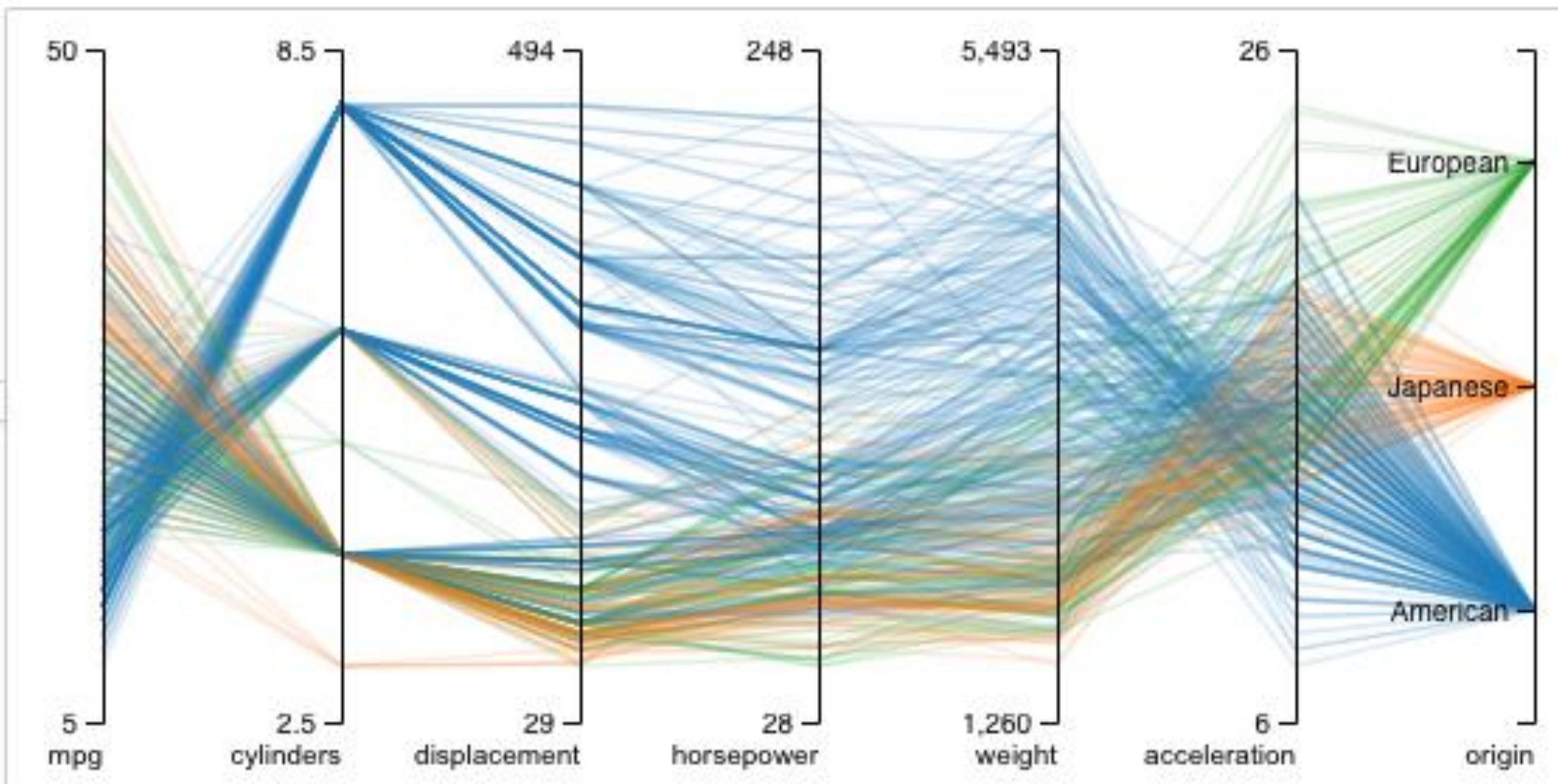
- Muestra un número X de atributos cuantitativos sobre ejes en paralelo.
- Marcas: puntos sobre cada eje según su valor
- Líneas zig-zag representan ítems a través de los ejes
- Tareas:
 - Identificar correlaciones
 - Overview de todos los atributos
 - outliers
- **Ordenar los ejes no es trivial**
- Escalabilidad:
 - Docenas de atributos.
 - Cientos de ítems.

Table				
	Math	Physics	Dance	Drama
	85	95	70	65
	90	80	60	50
	65	50	90	90
	50	40	95	80
	40	60	80	90

Parallel Coordinates



Coordenadas paralelas



Visflow.org

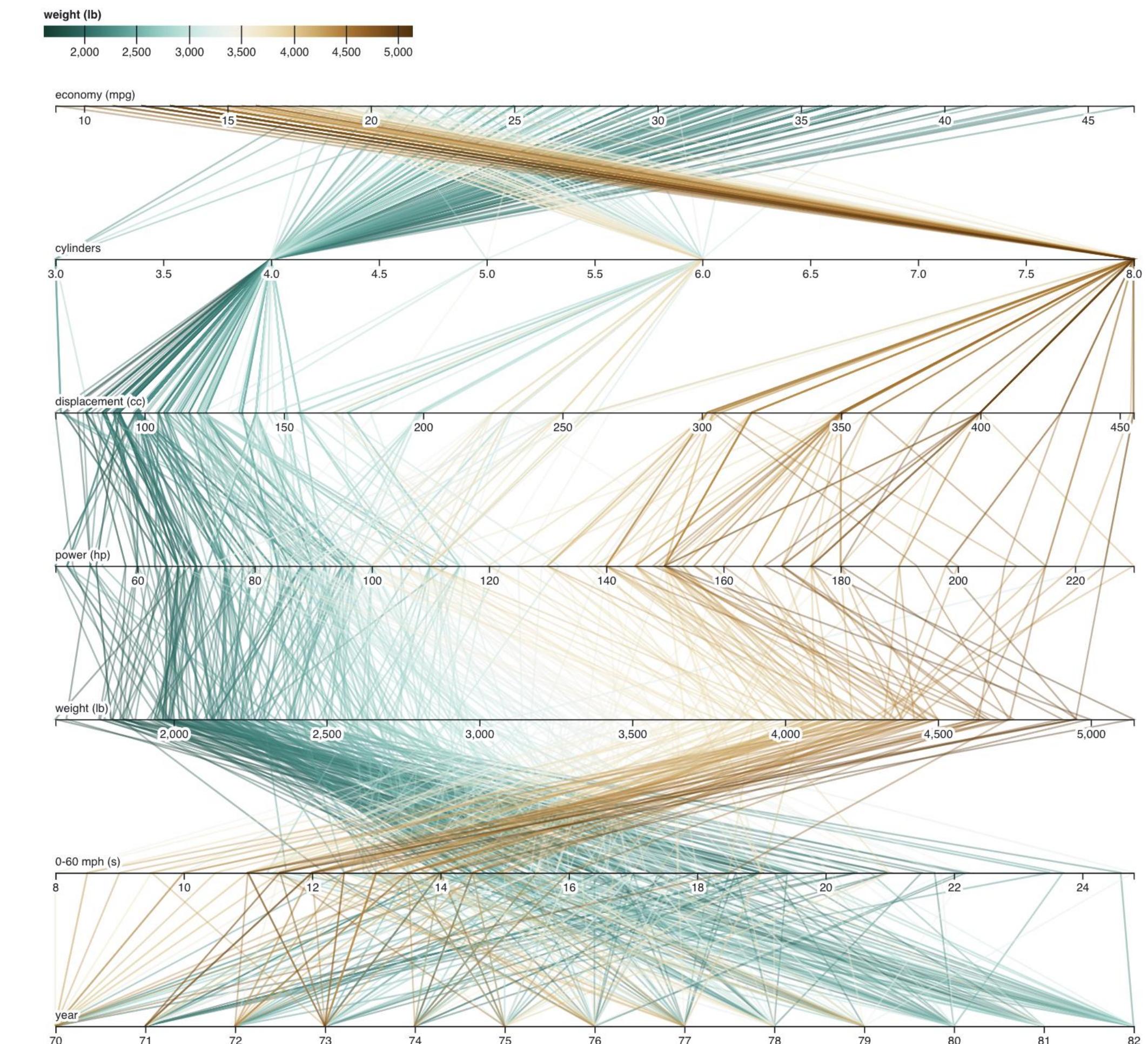
Correlaciones negativas entre eficiencia y cilindros

Positiva entre cilindros, ccs y potencia

Negativa para peso y aceleración

El último eje y el color muestran país de procedencia

Difícil escoger el orden de los ejes (opciones automáticas e interactivas).



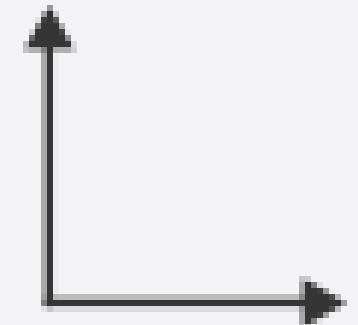
Mike Bostock

Orientación: Limitaciones

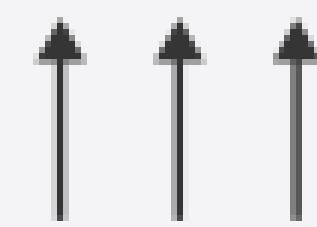
- Rectilinear:
 - No escalable en número de ejes
 - 2 ejes óptimo
 - 3 problemático
 - 4+ imposible
- Paralelo:
 - No intuitivo, tiempo de entrenamiento

→ Axis Orientation

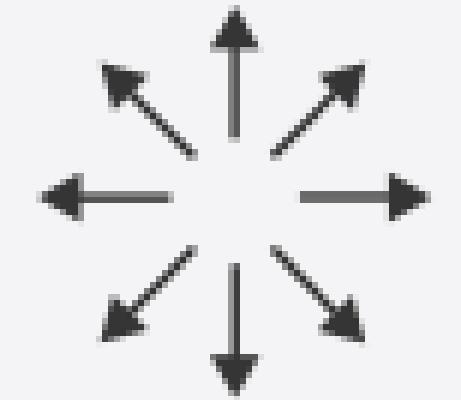
→ Rectilinear



→ Parallel



→ Radial

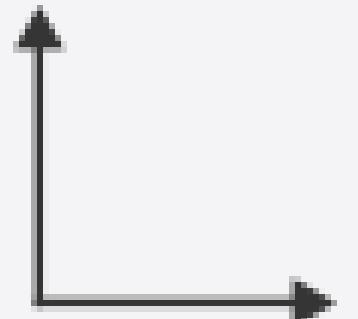


Orientación: Limitaciones

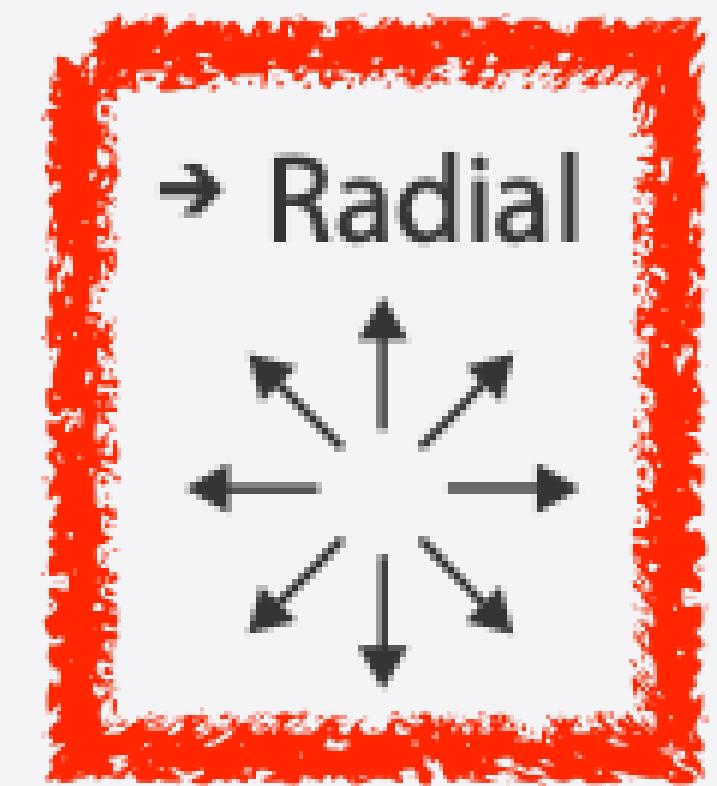
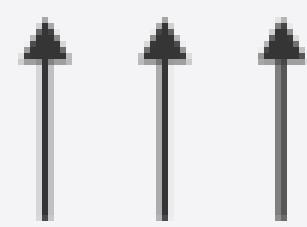
- Rectilinear:
 - No escalable en número de ejes
 - 2 ejes óptimo
 - 3 problemático
 - 4+ imposible
- Paralelo:
 - No intuitivo, tiempo de entrenamiento

→ Axis Orientation

→ Rectilinear

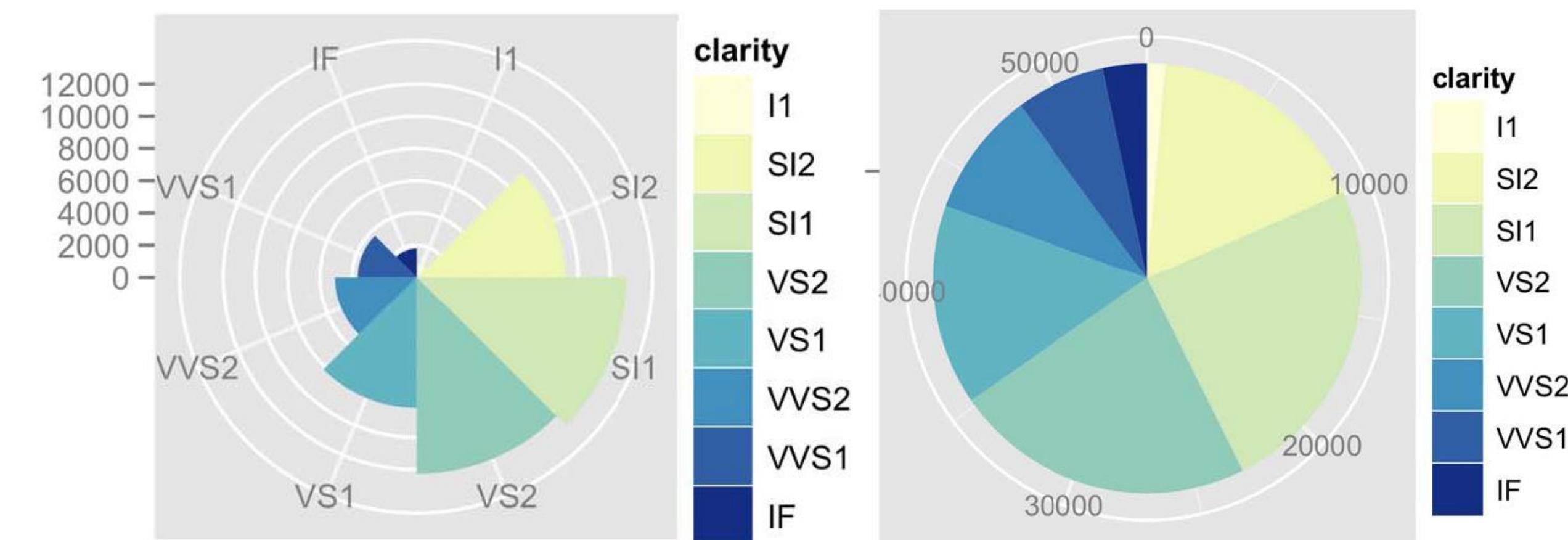
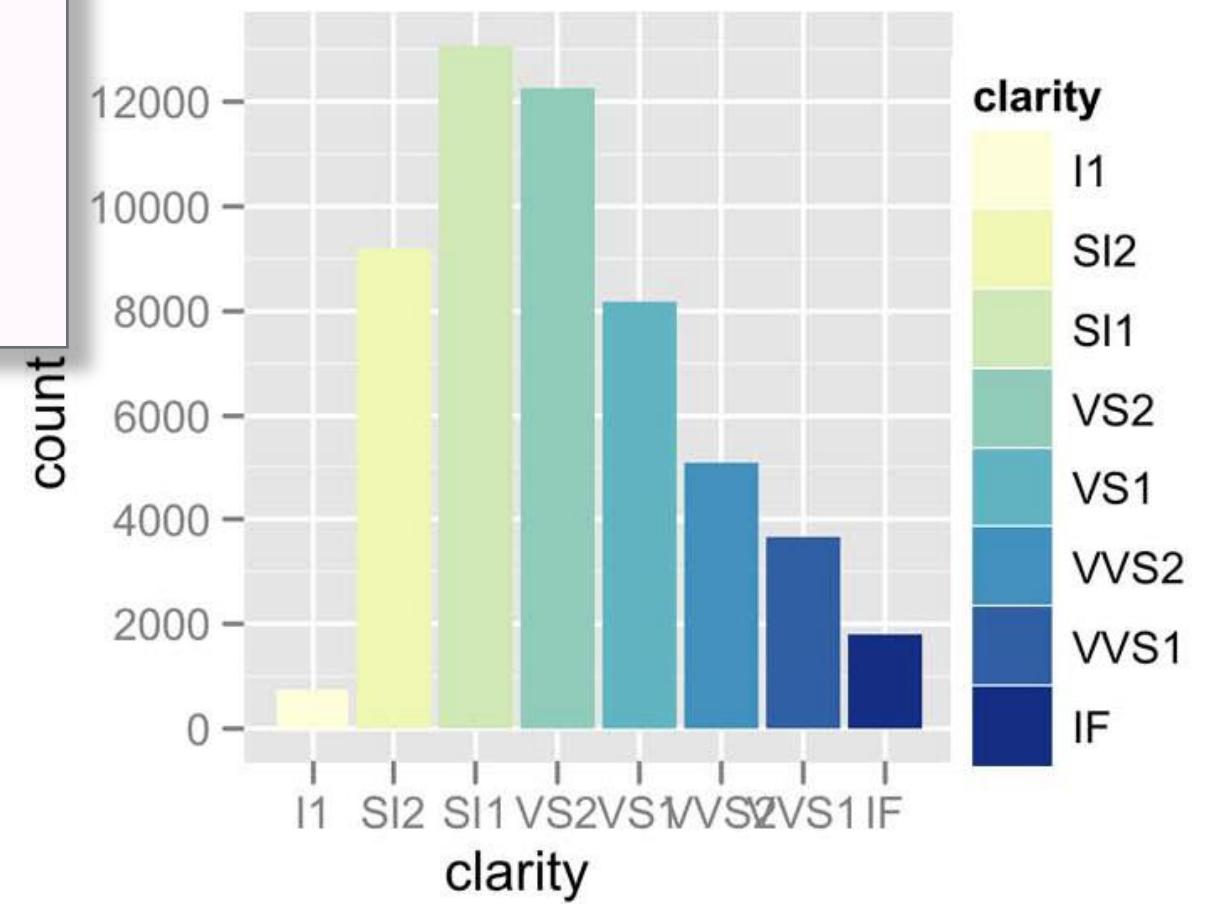
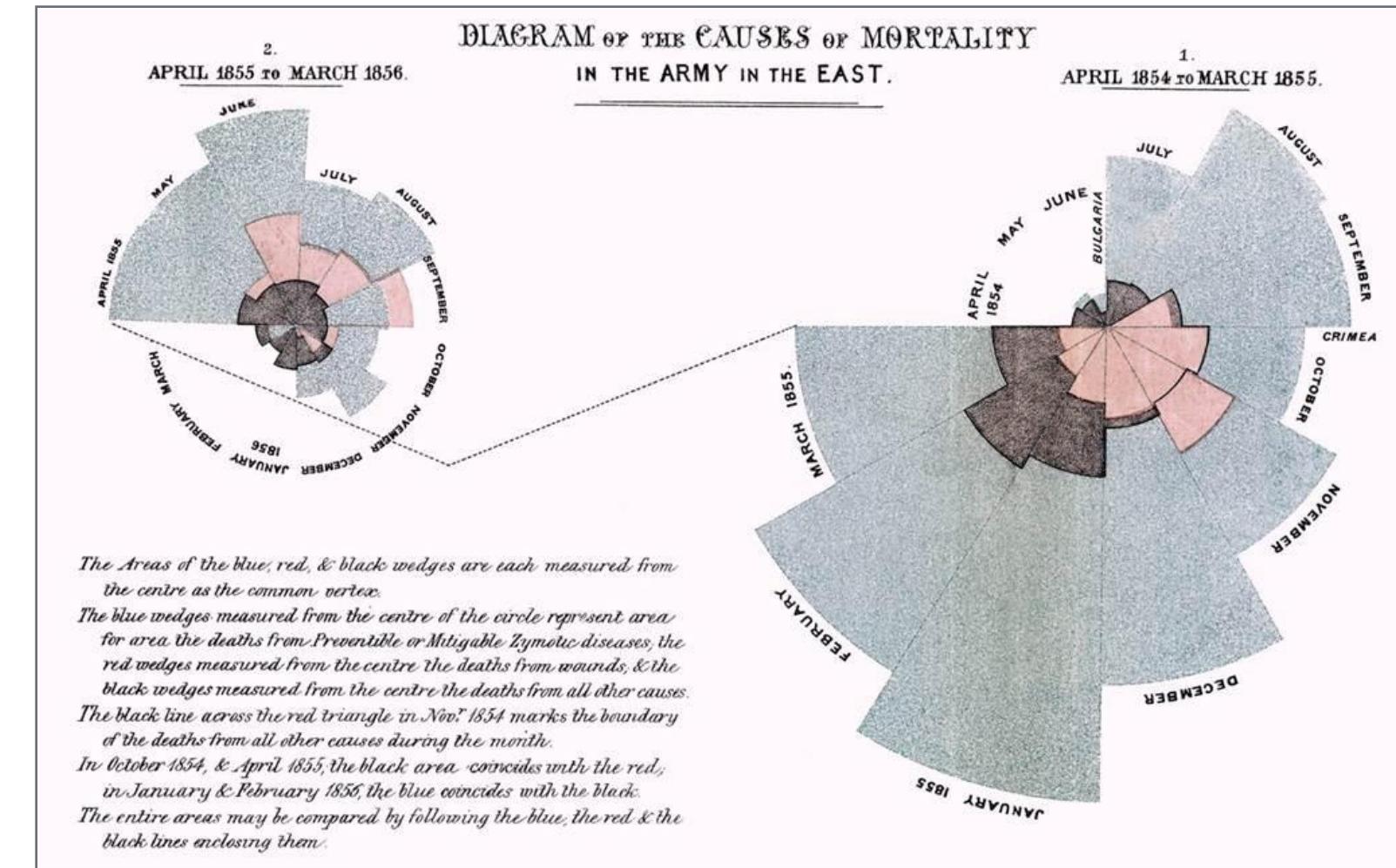


→ Parallel



Pie chart - polar area chart

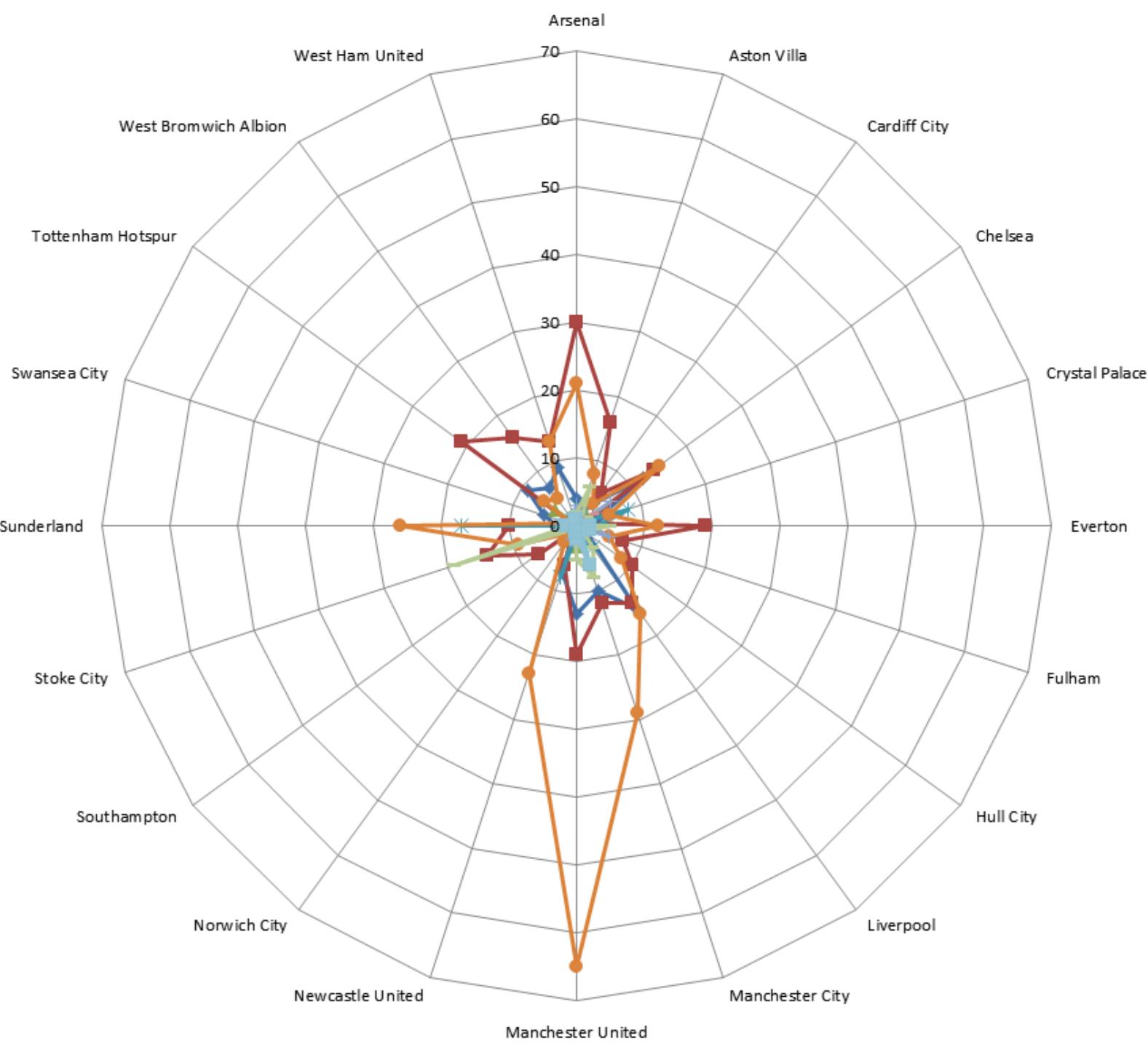
- Pie chart
 - Marcas de área con canal de ángulo
 - Precisión: Menor que length (bars)
- Polar area chart
 - Marcas de área con canal de longitud
 - Más similitud con barras
- Datos:
 - 1 key categórico, 1 value cuantitativo
- Tarea:
 - Relaciones parte-todo



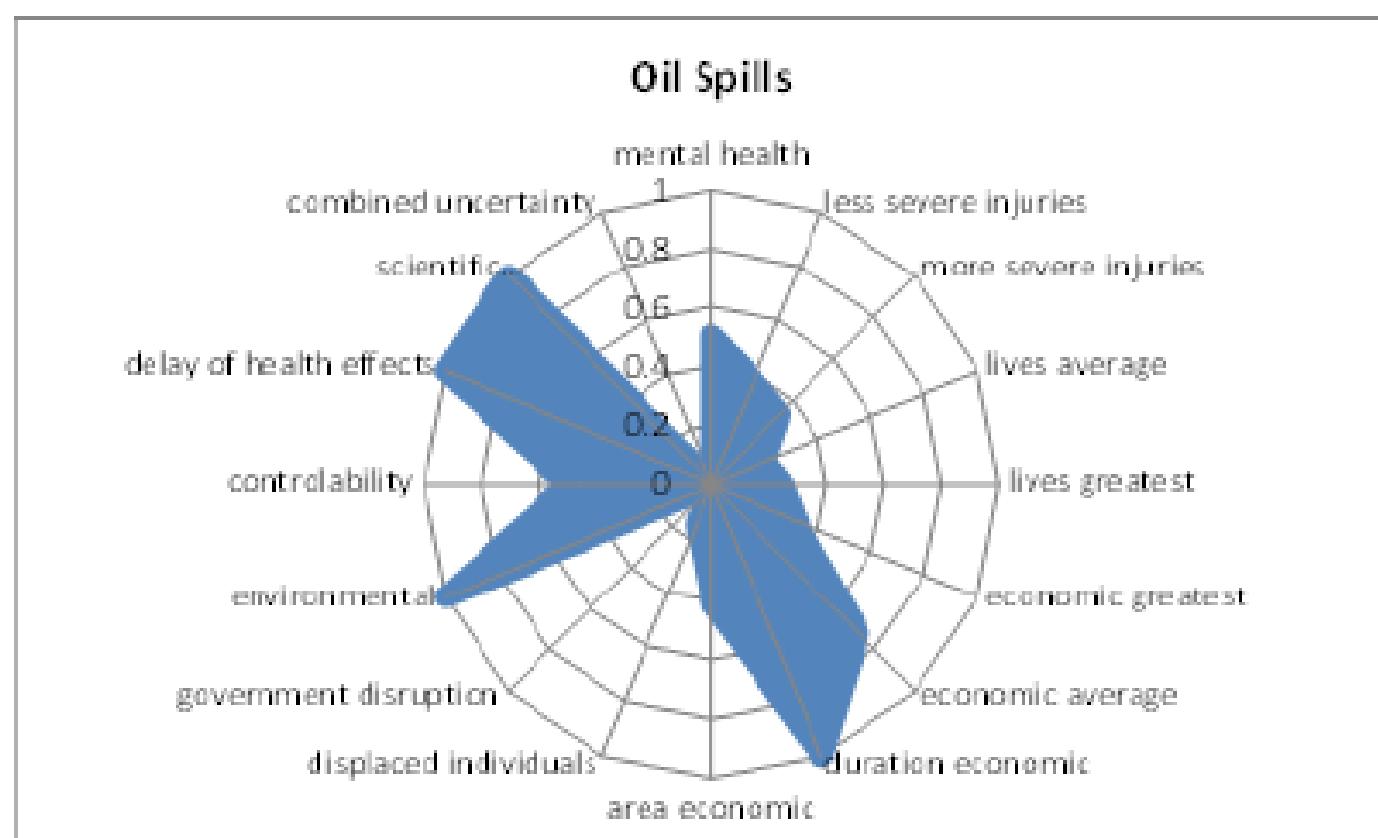
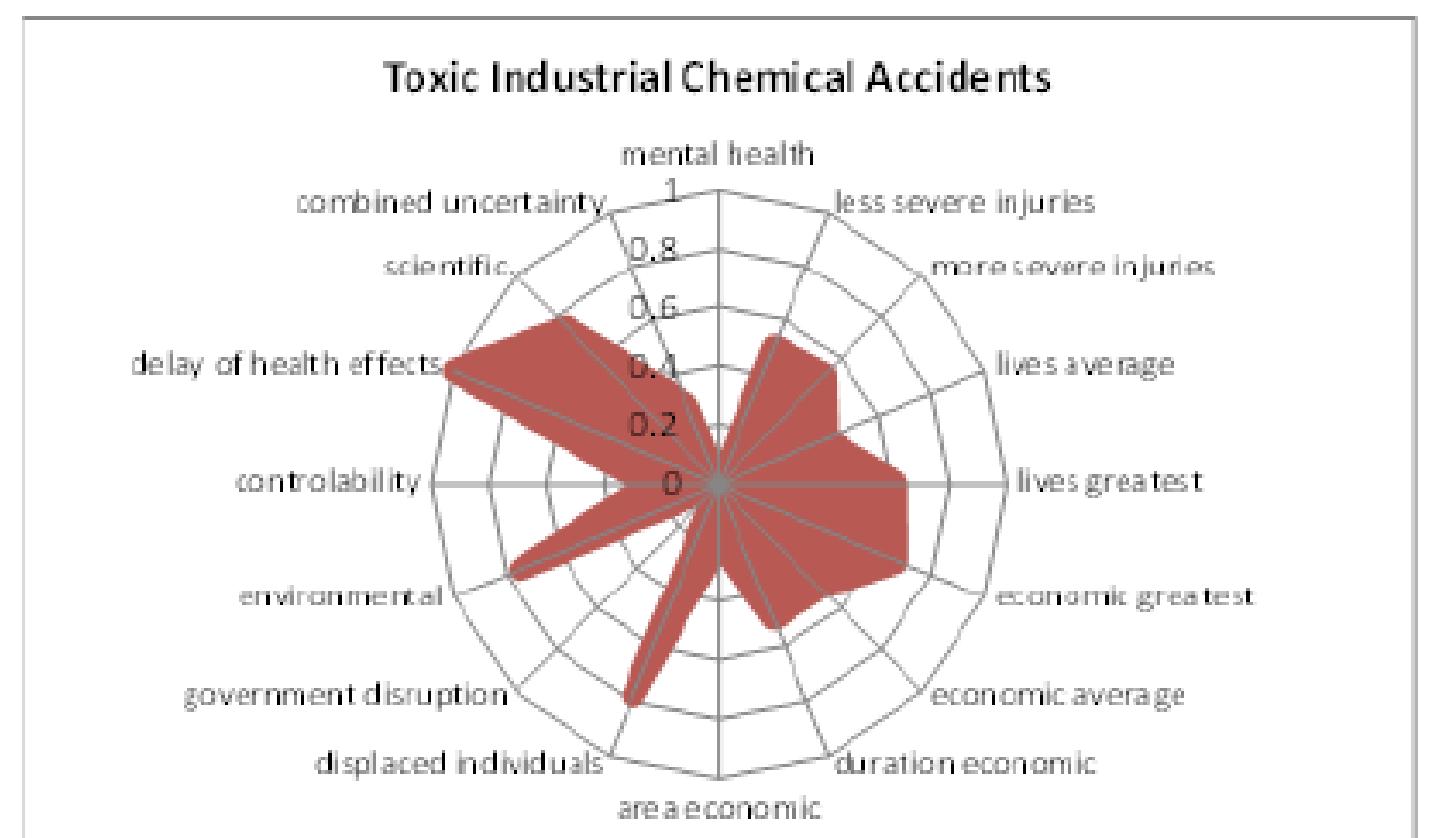
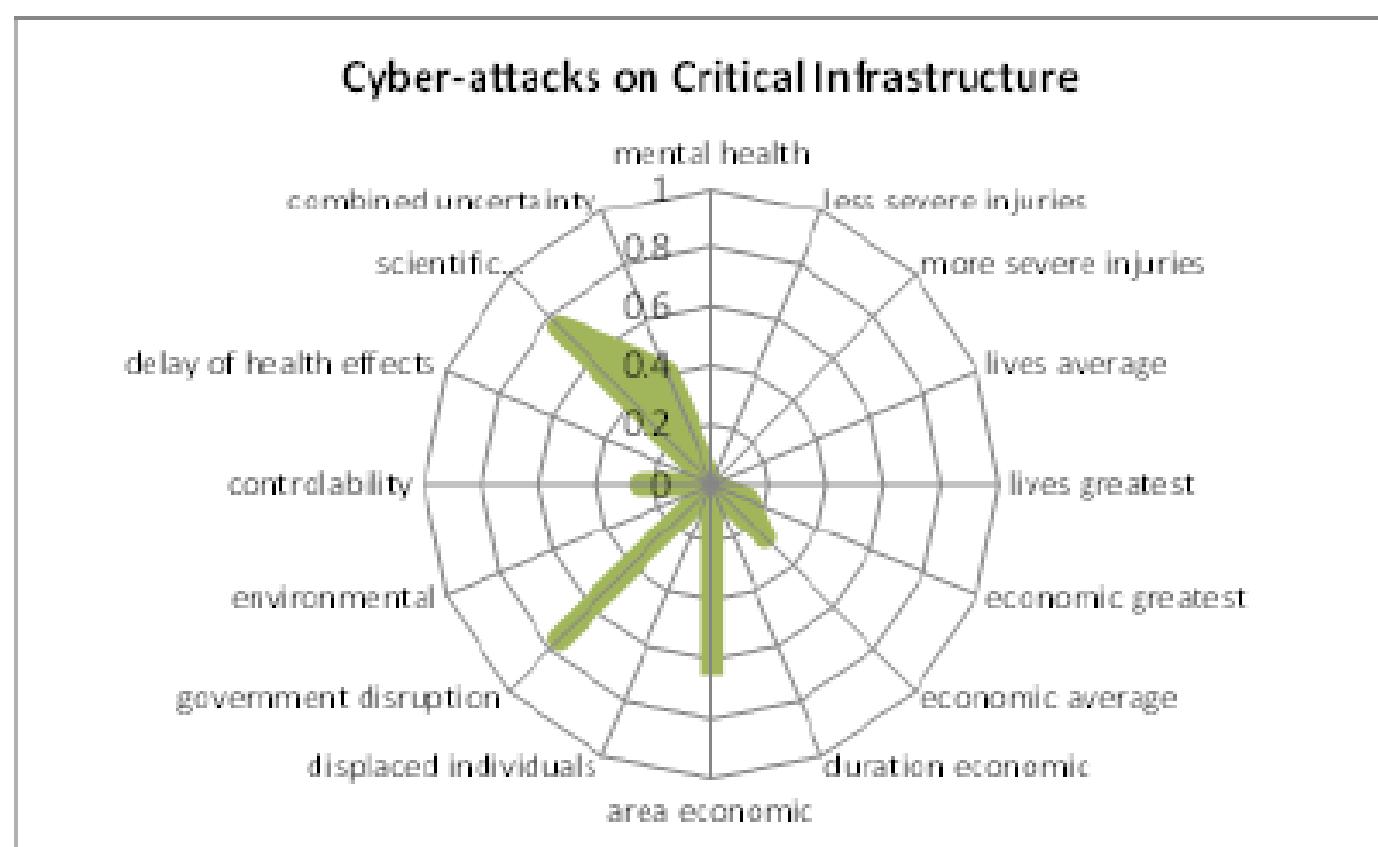
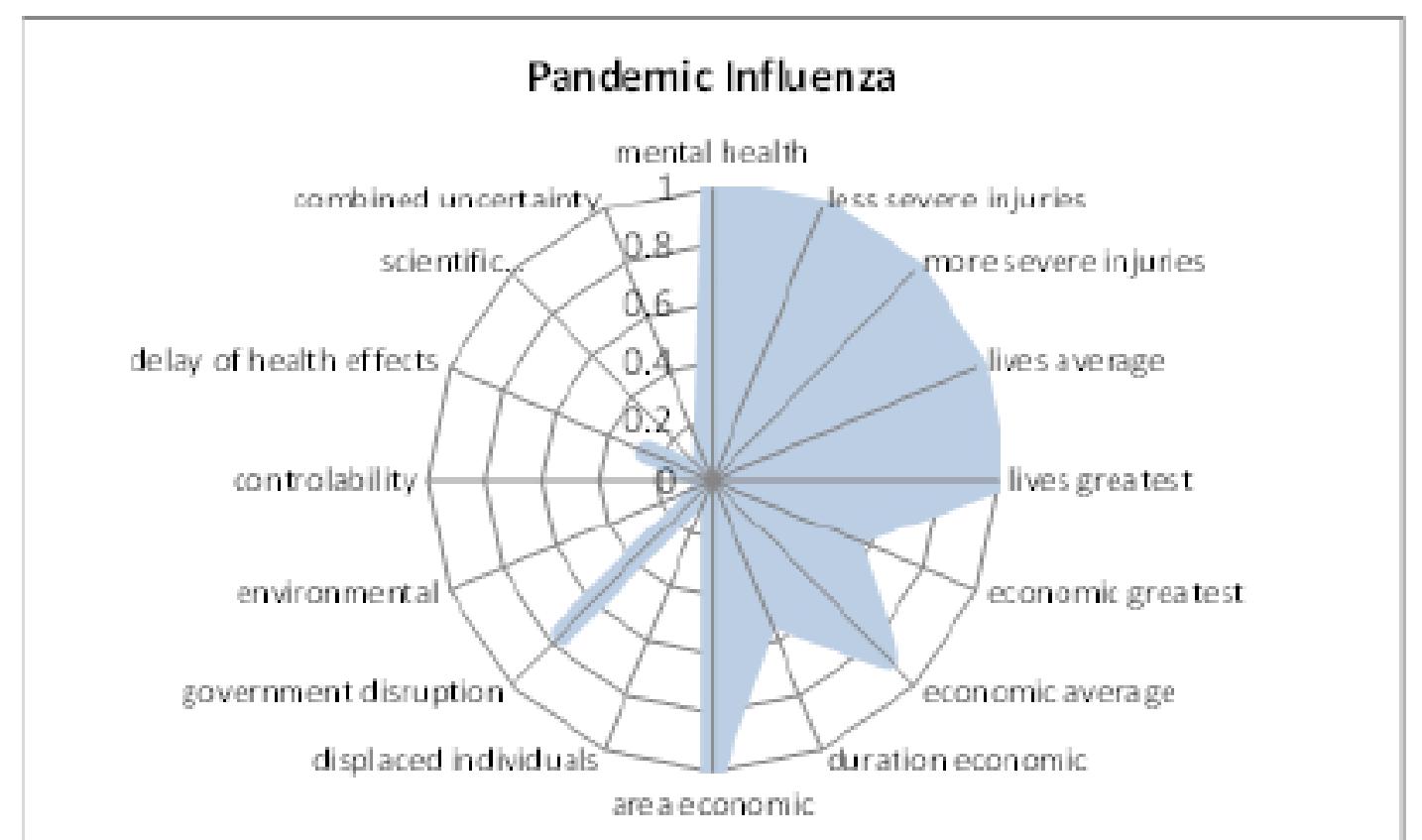
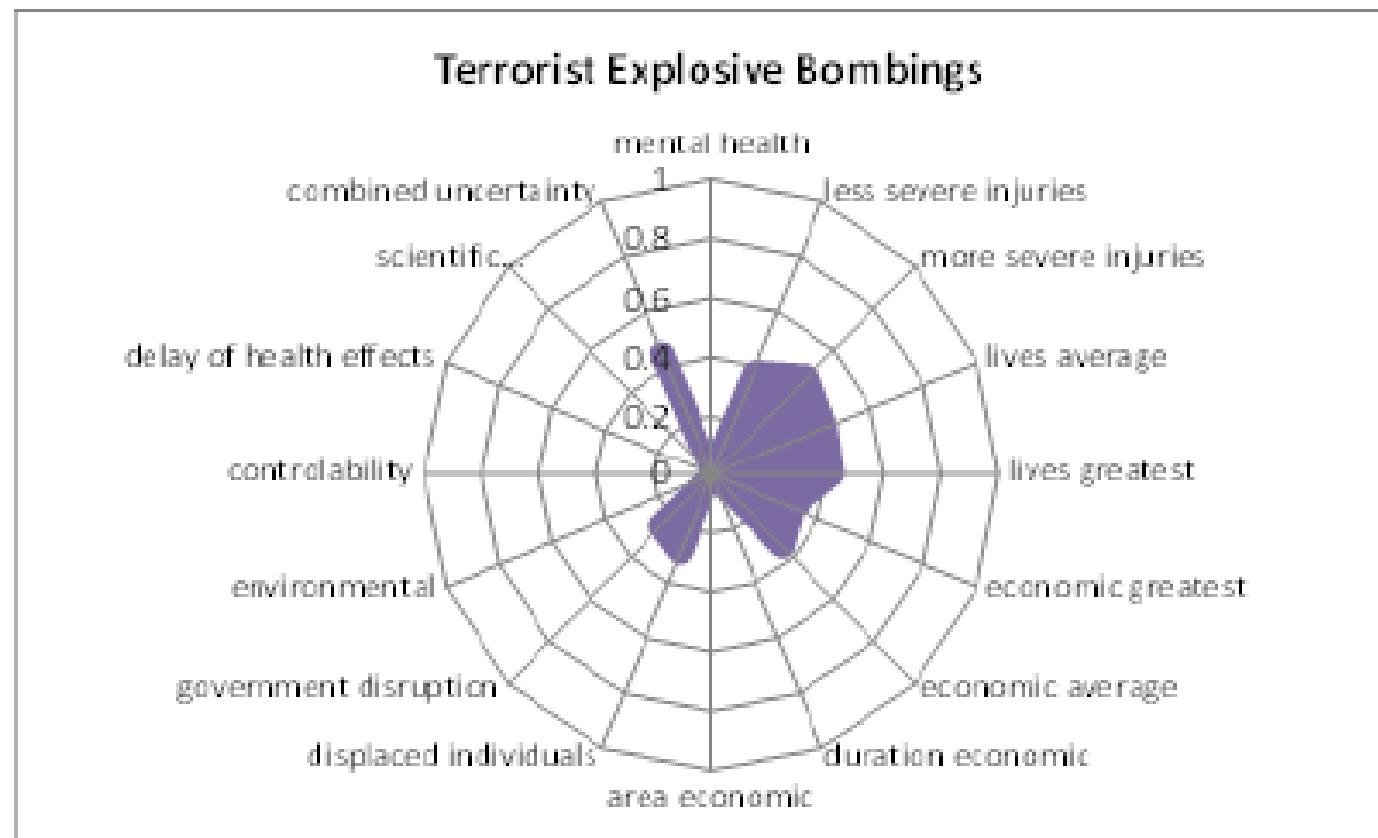
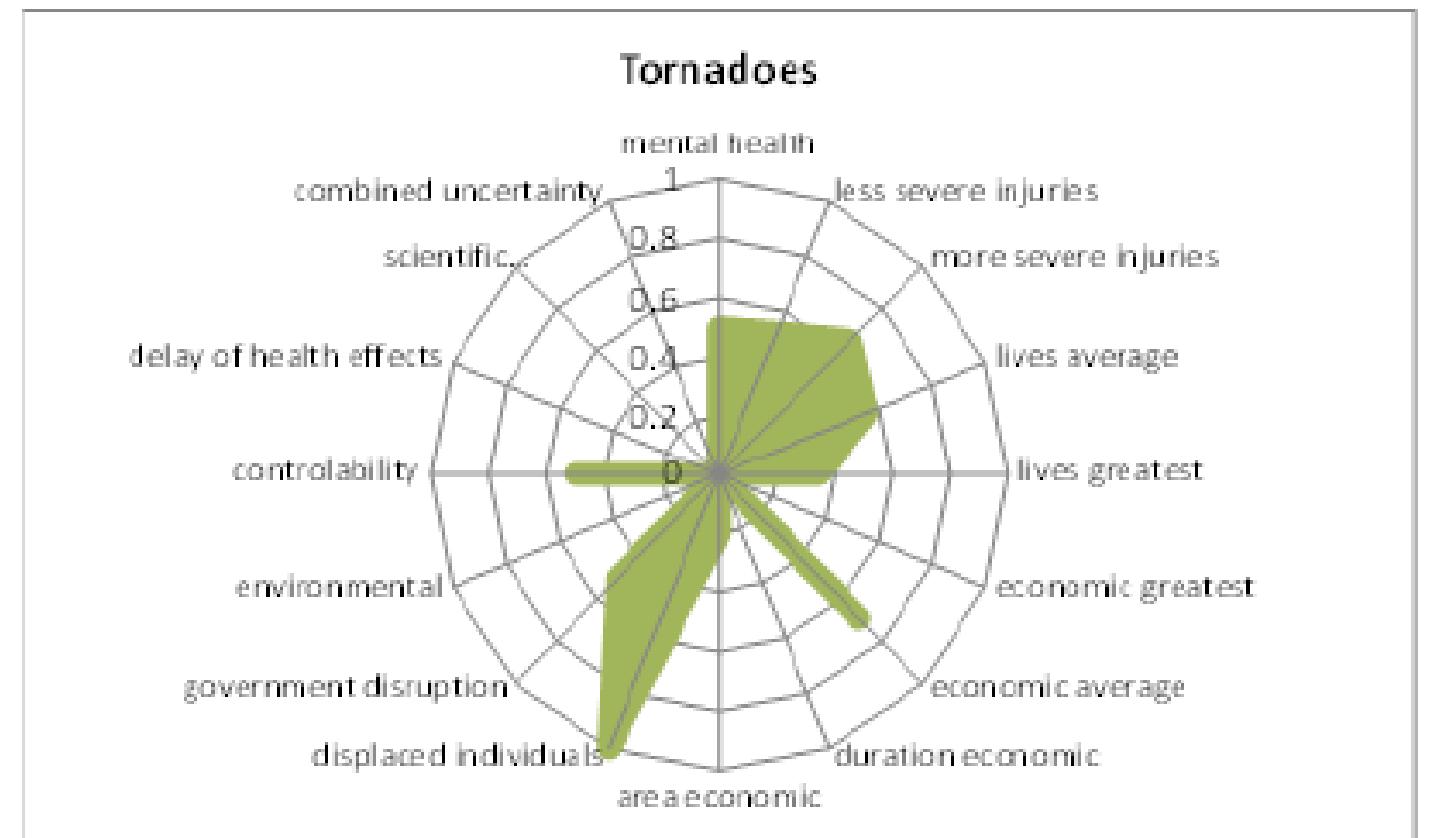
Radar Plots

Limitación:

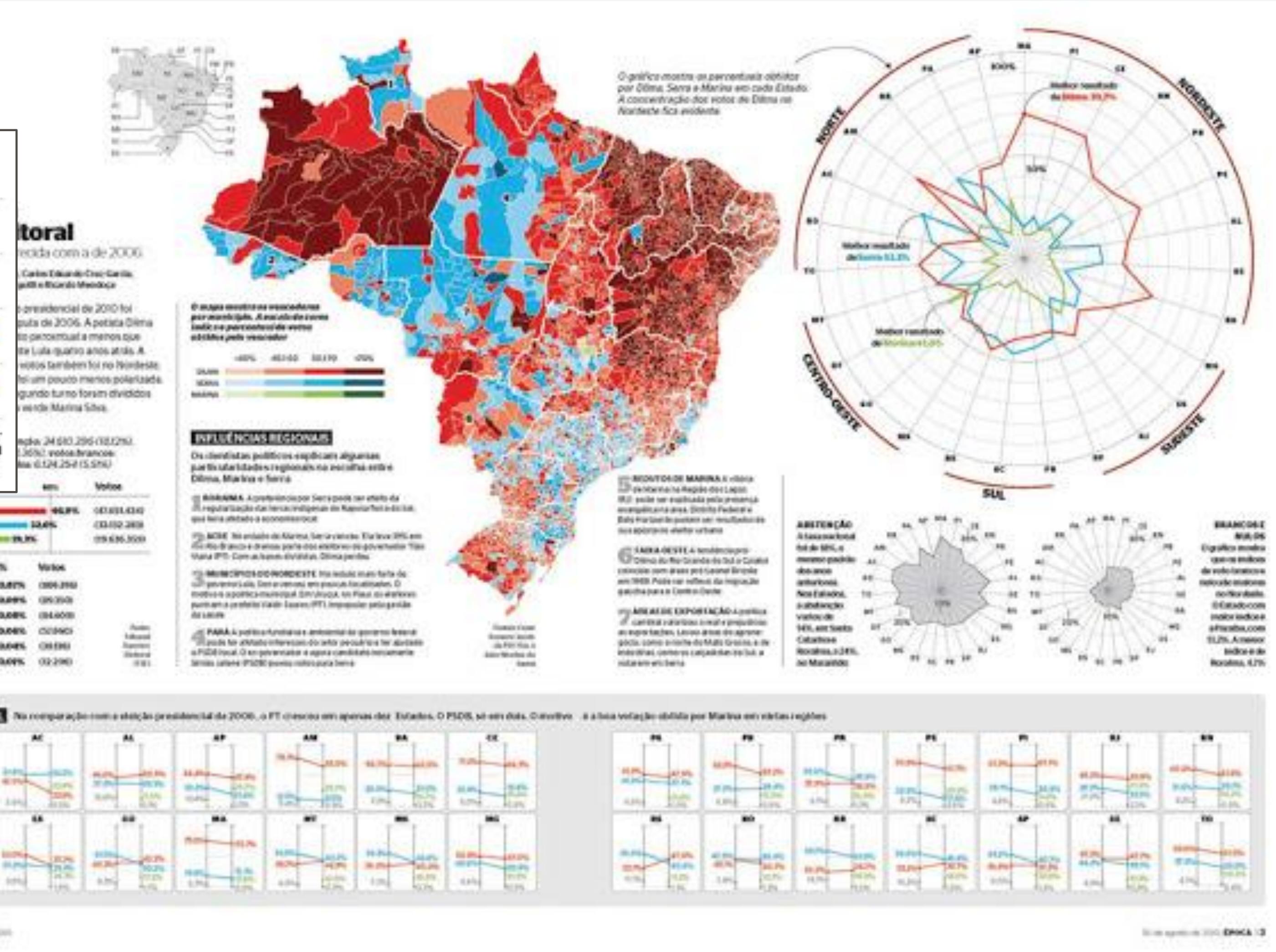
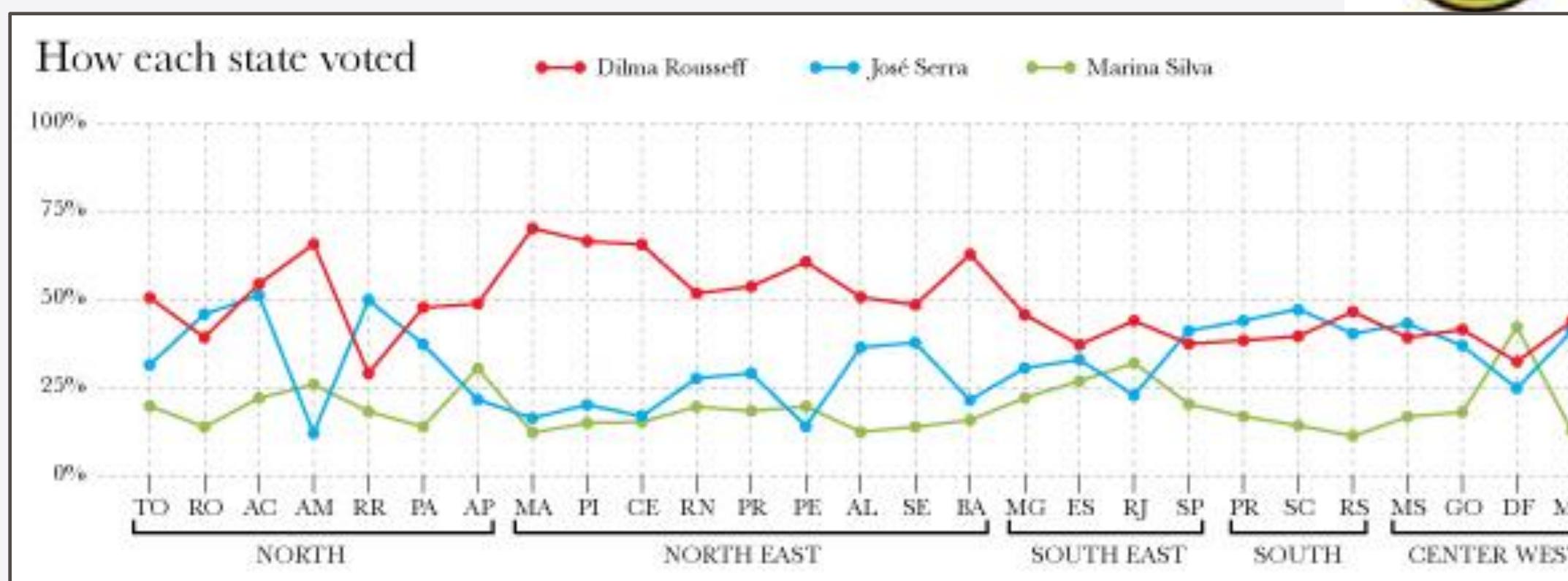
- Mejor para categorías no-cílicas



[Slide courtesy of Ben Jones]. From: Visualization analysis and design



“Radar graphs: Avoid them (99.9% of the time)”



The functional art: Radar graphs: Avoid them (99.9% of the time)

Bibliografia

- Visualization Analysis and Design. Munzner. Cap. 7
- **The Functional Art . Alberto Cairo, 2012 (Cap. 5-6)**

Coordenadas paralelas: <https://eagereyes.org/techniques/parallel-coordinates>

Coordenadas paralelas en D3: [@d3/parallel-coordinates](https://observablehq.com/@d3/parallel-coordinates)

Mosaic plots: <http://www.pmean.com/definitions/mosaic.htm>

Stacked barcharts: <https://blog.datawrapper.de/divergingbars/>

Radar graphs: <http://www.thefunctionalart.com/2012/11/radar-graphs-avoid-them-999-of-time.html>

Inspiración

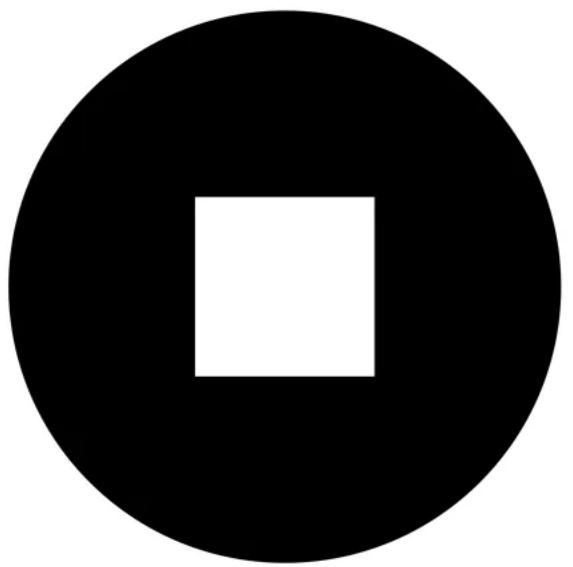
- <http://www.thefunctionalart.com/>
- <https://eagereyes.org/>
- <http://flowingdata.com/>
- <http://fivethirtyeight.com/>
- <http://truth-and-beauty.net/>



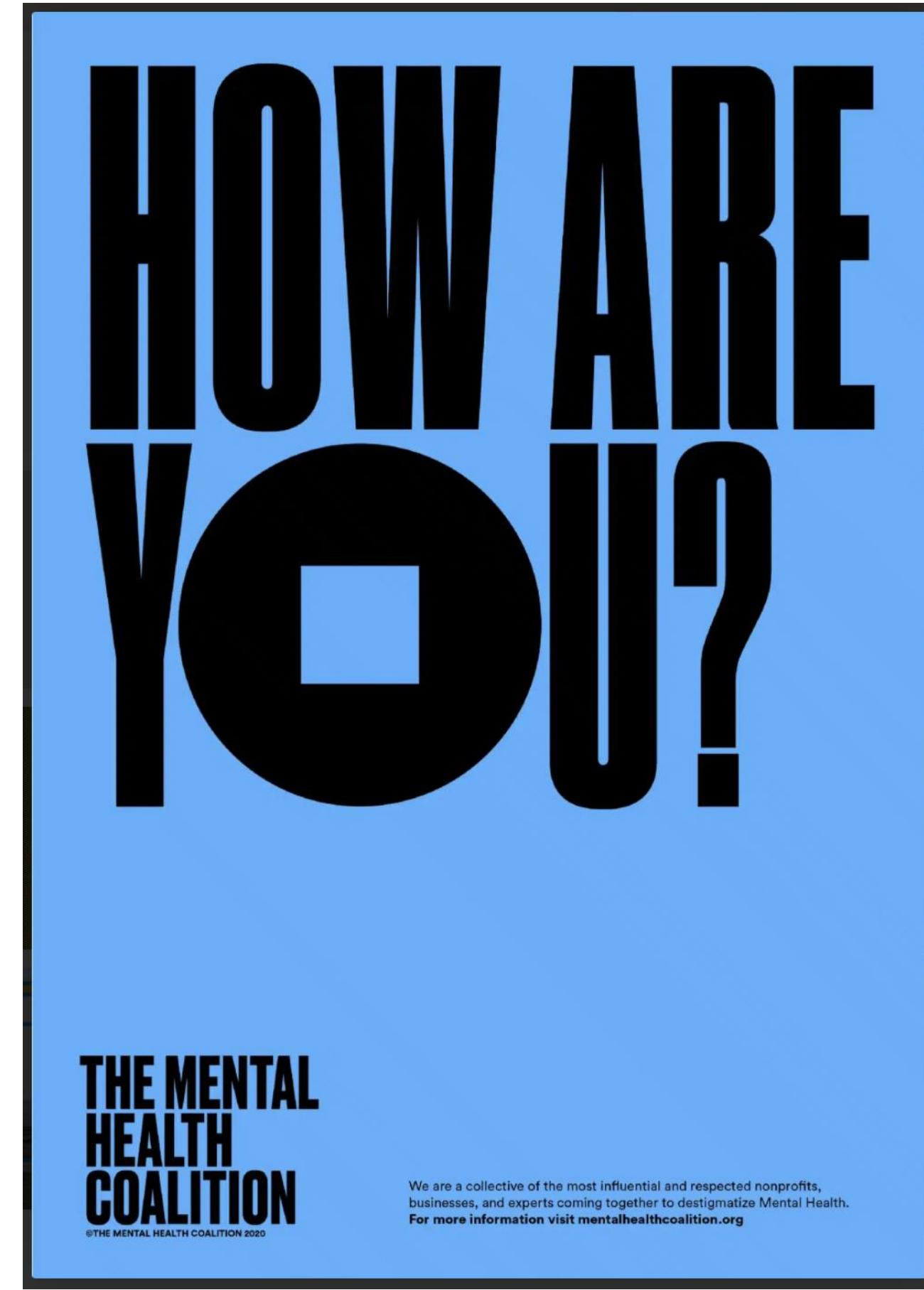
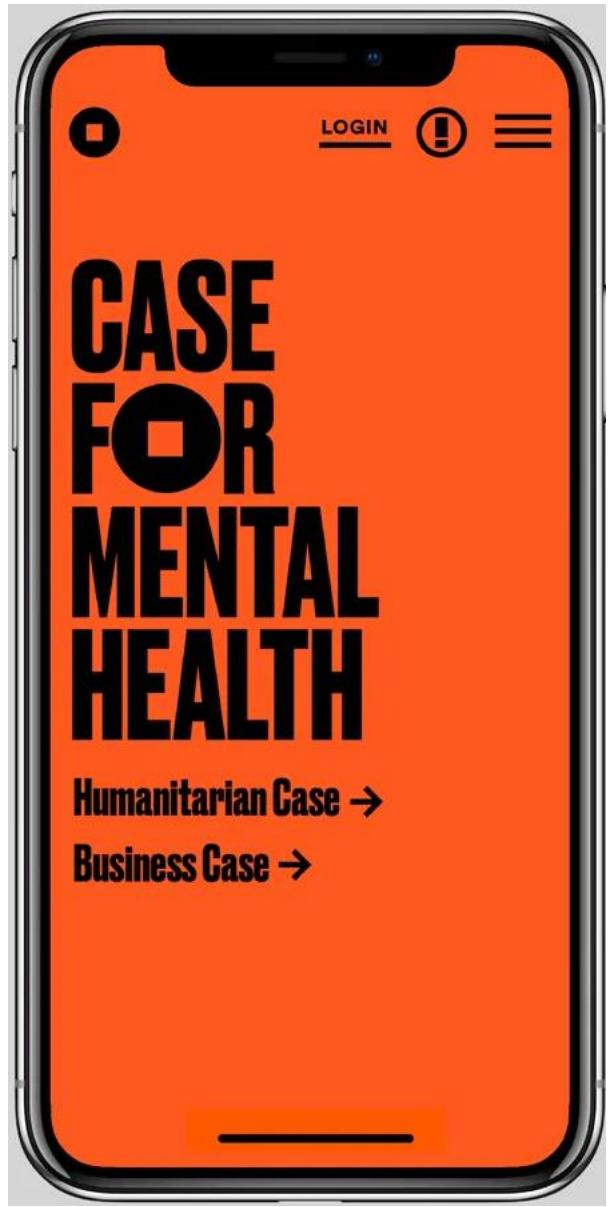
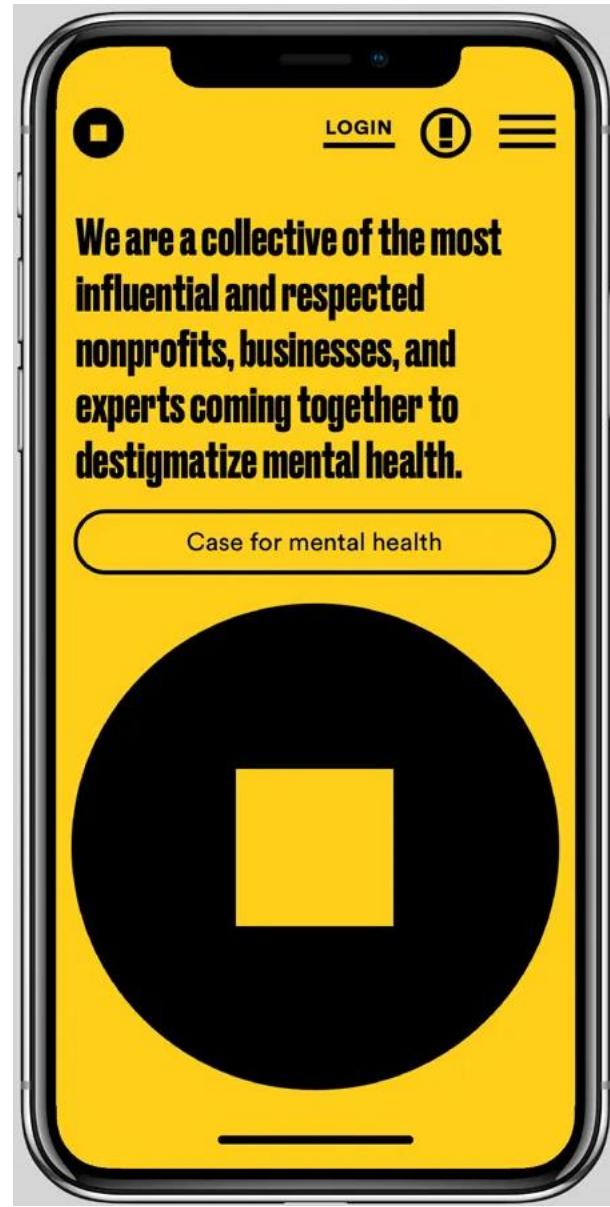
4. Diseño de la Información

Sol Bucalo
sol.bucalo@uab.cat

Guillermo Marin
guillermo.marin@uab.cat

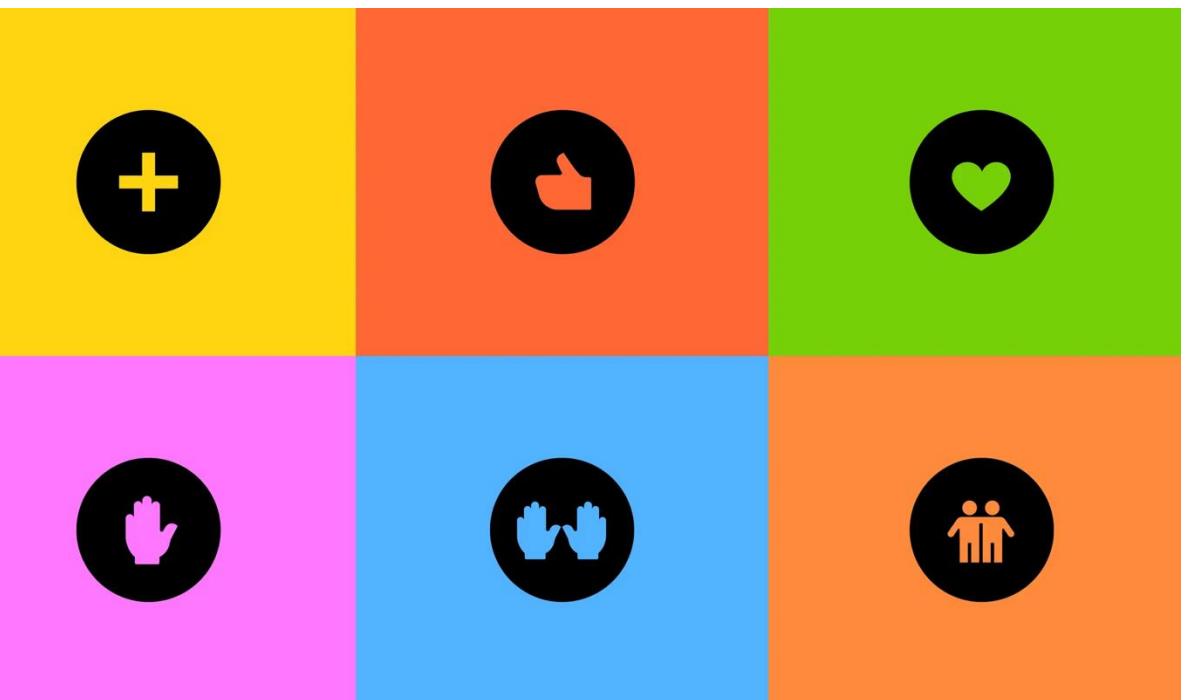


THE MENTAL
HEALTH
COALITION



HOW
ARE YOU
REALLY?

HOW
ARE YOU
REALLY?



Design is not just about aesthetics



Instituto de Estudos Orientais.



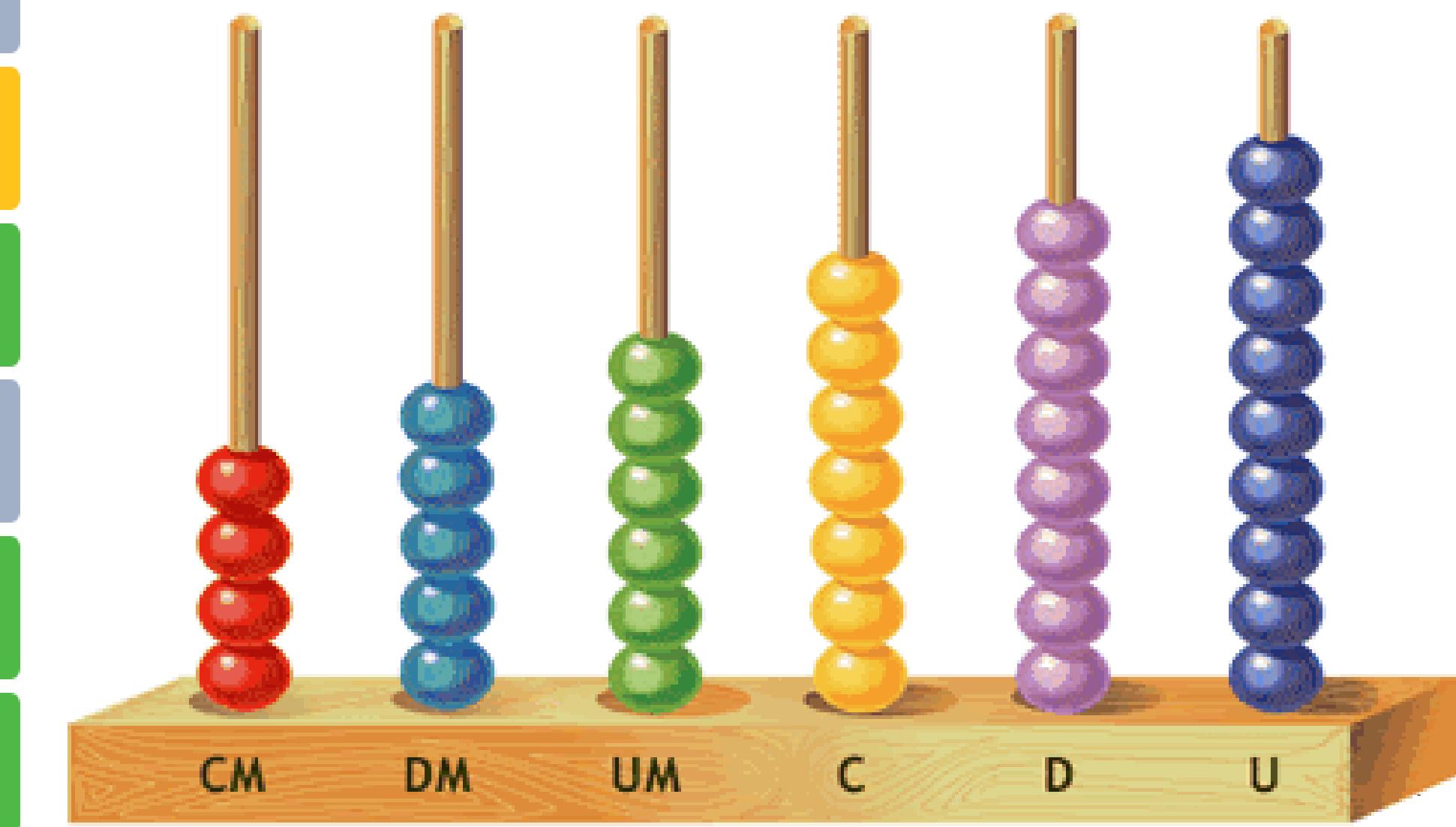
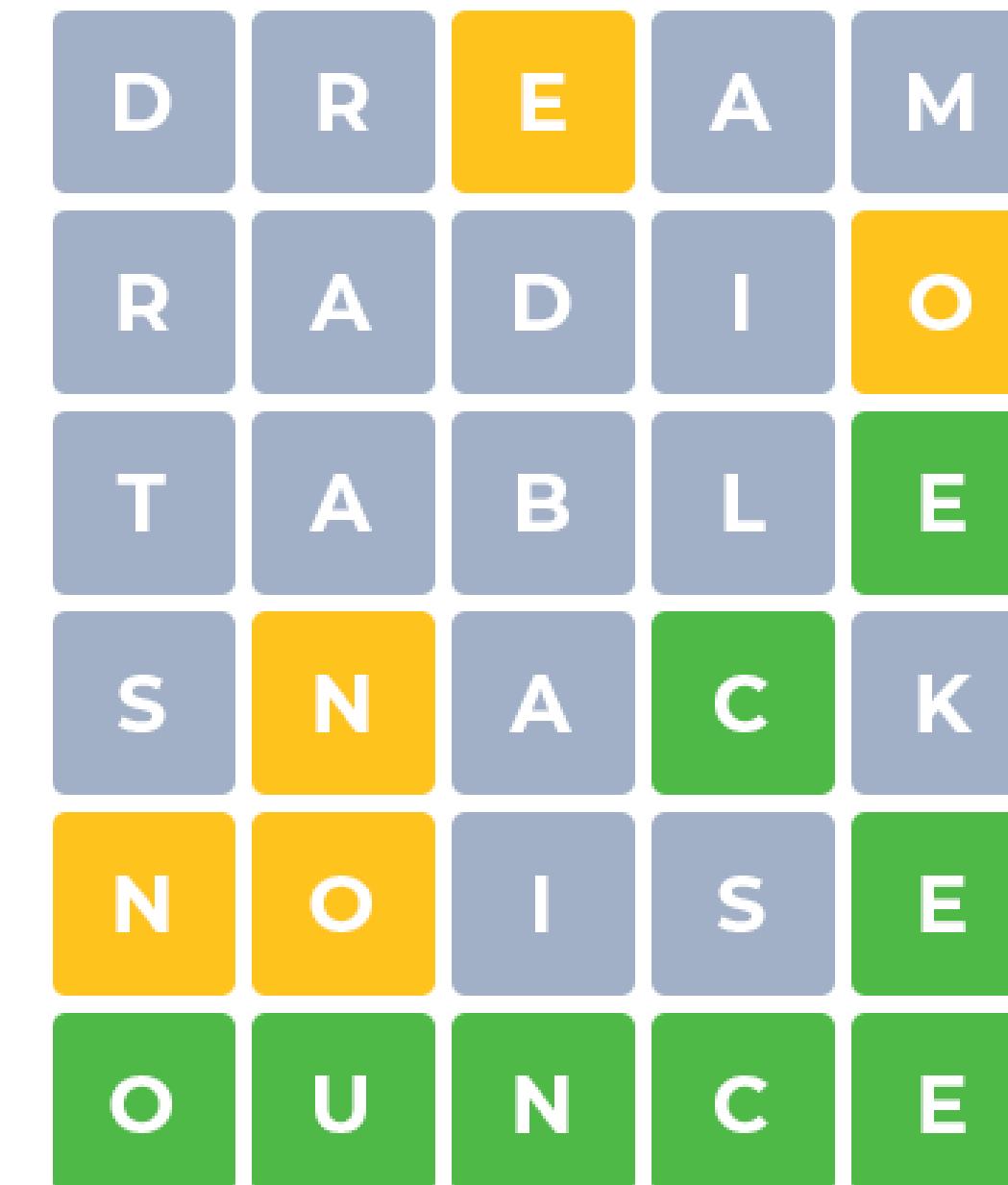
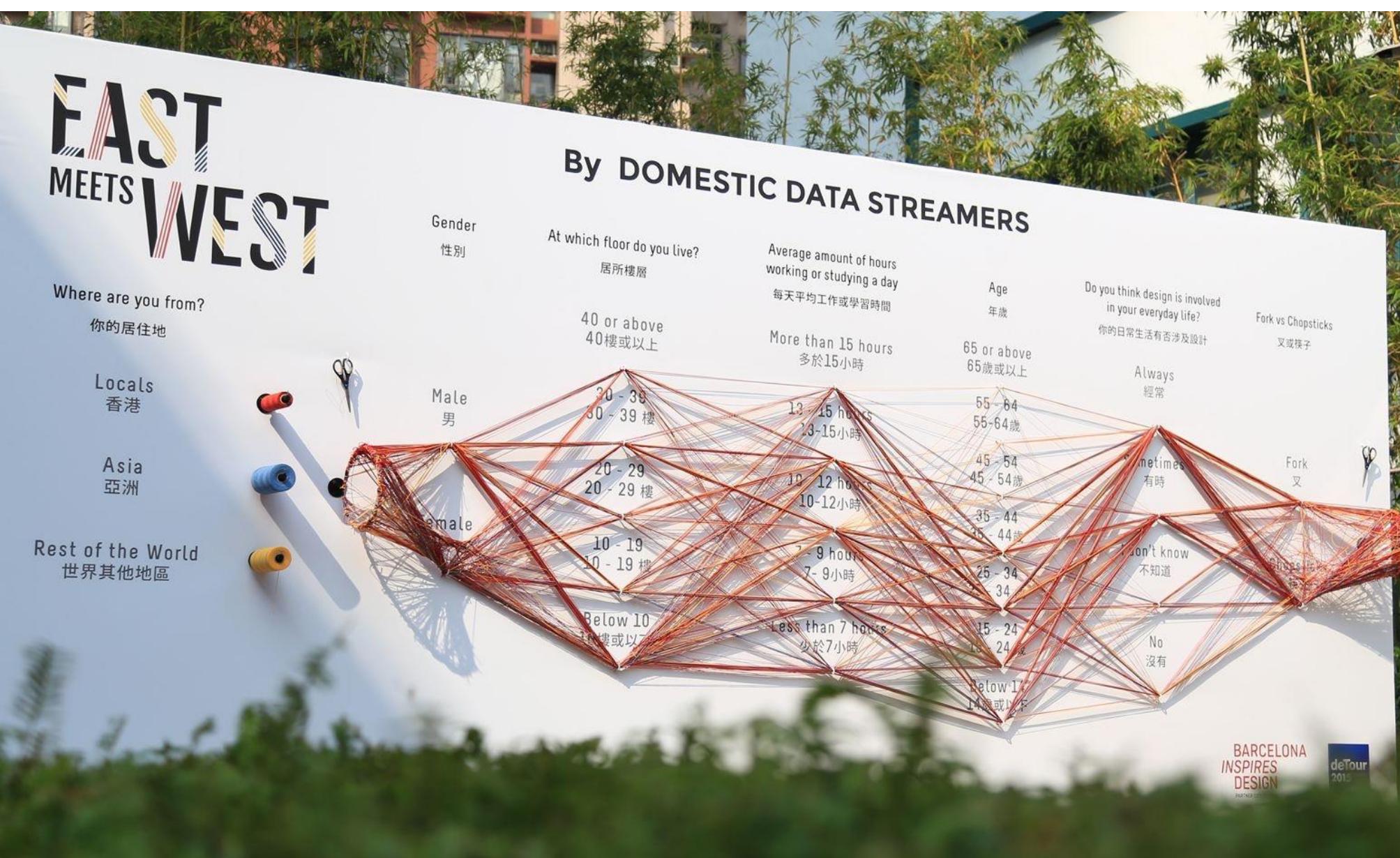
Design is not just about aesthetics



Design is not just about aesthetics

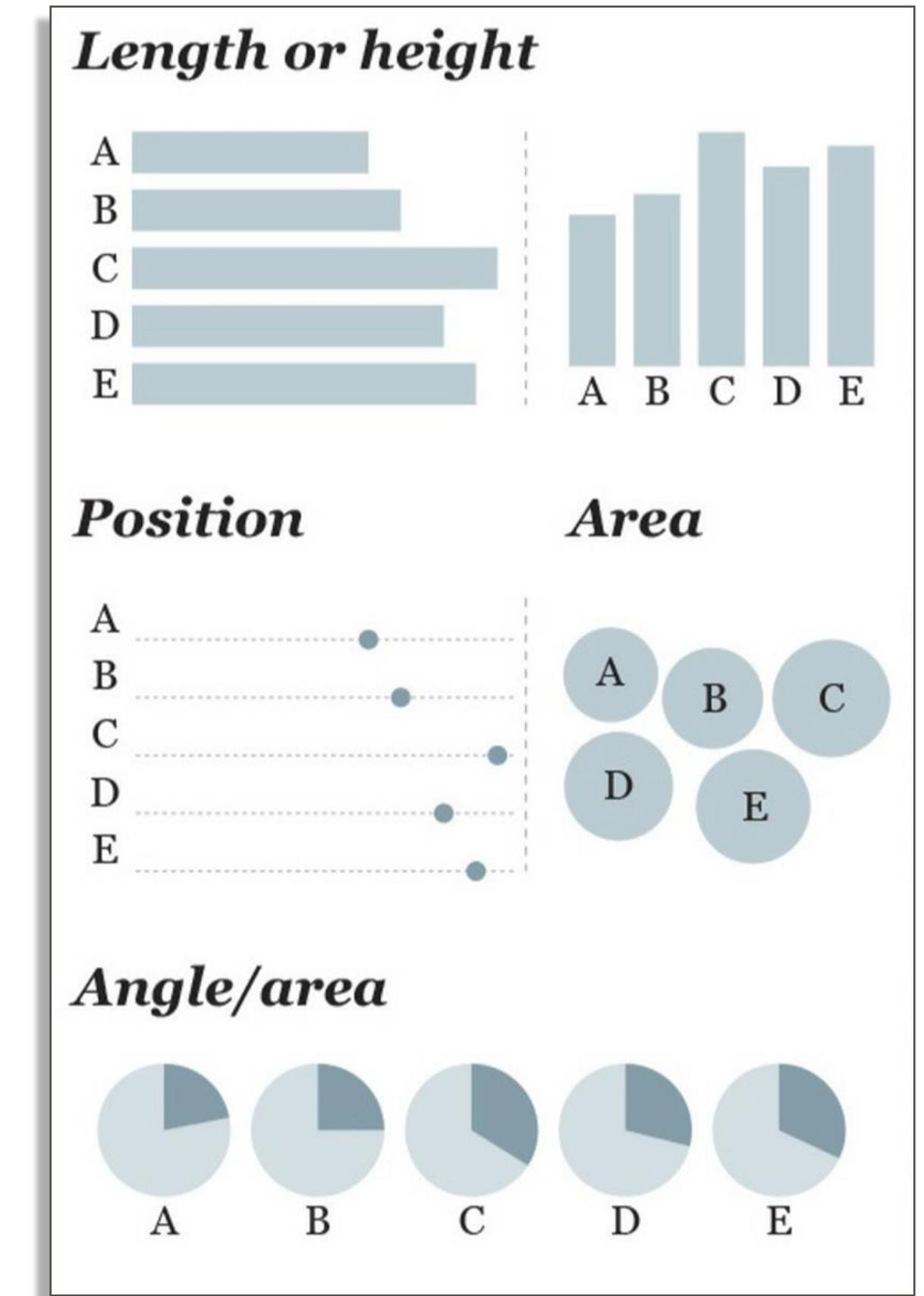
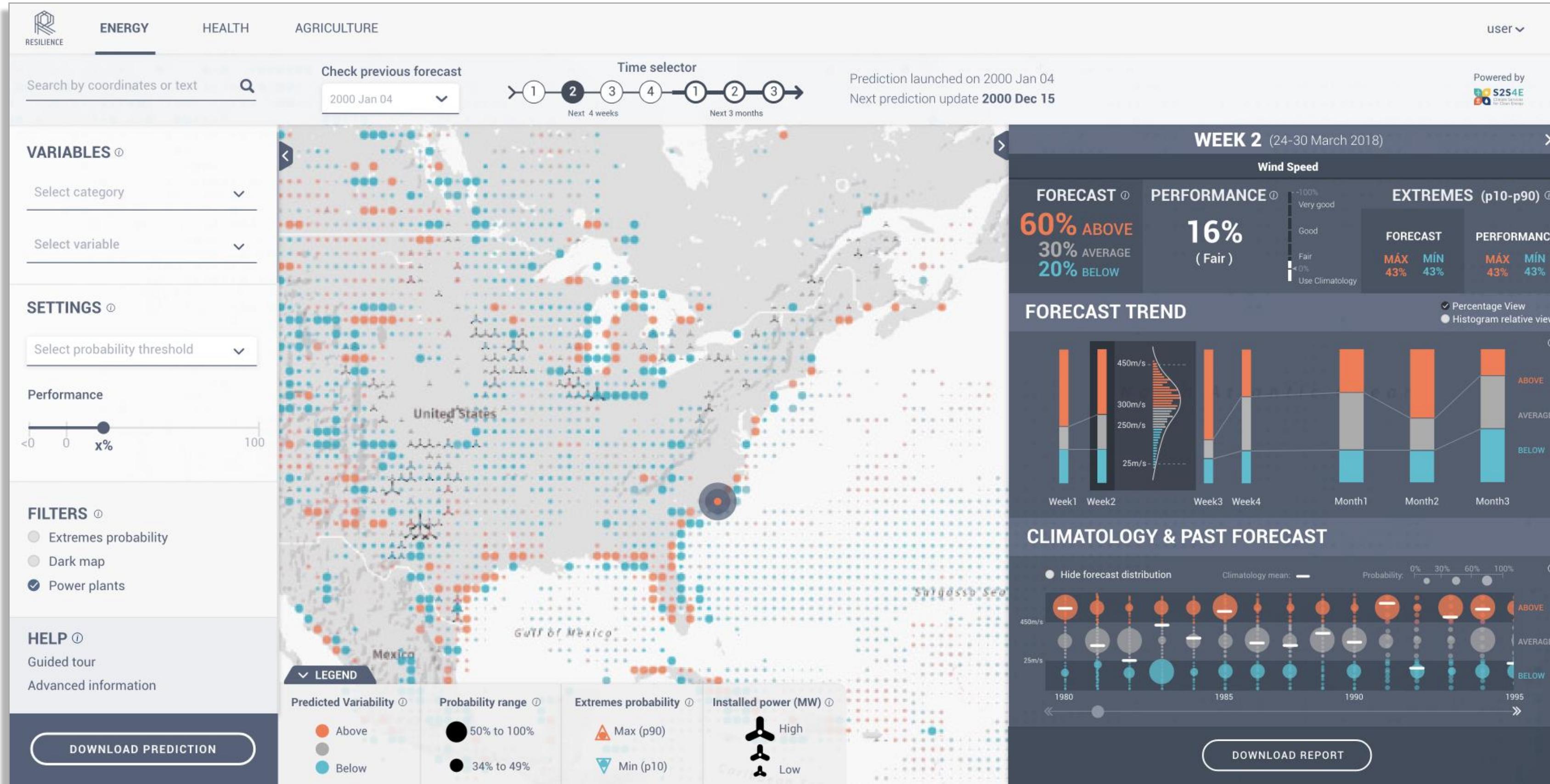
Data Visualization

- Son DISEÑOS que contienen INFORMACIÓN
- Su función es AUMENTAR NUESTRAS CAPACIDADES para superar nuestras limitaciones cognitivas
- Herramientas cognitivas que AYUDEN A PENSAR



Percepción Visual

- Mecanismos cognitivos.
- Como afectan al diseño y consumo de visualizaciones de datos.
- Su relación con la efectividad de los canales visuales
- **Diseñar teniendo en cuenta cómo funciona nuestra percepción visual**



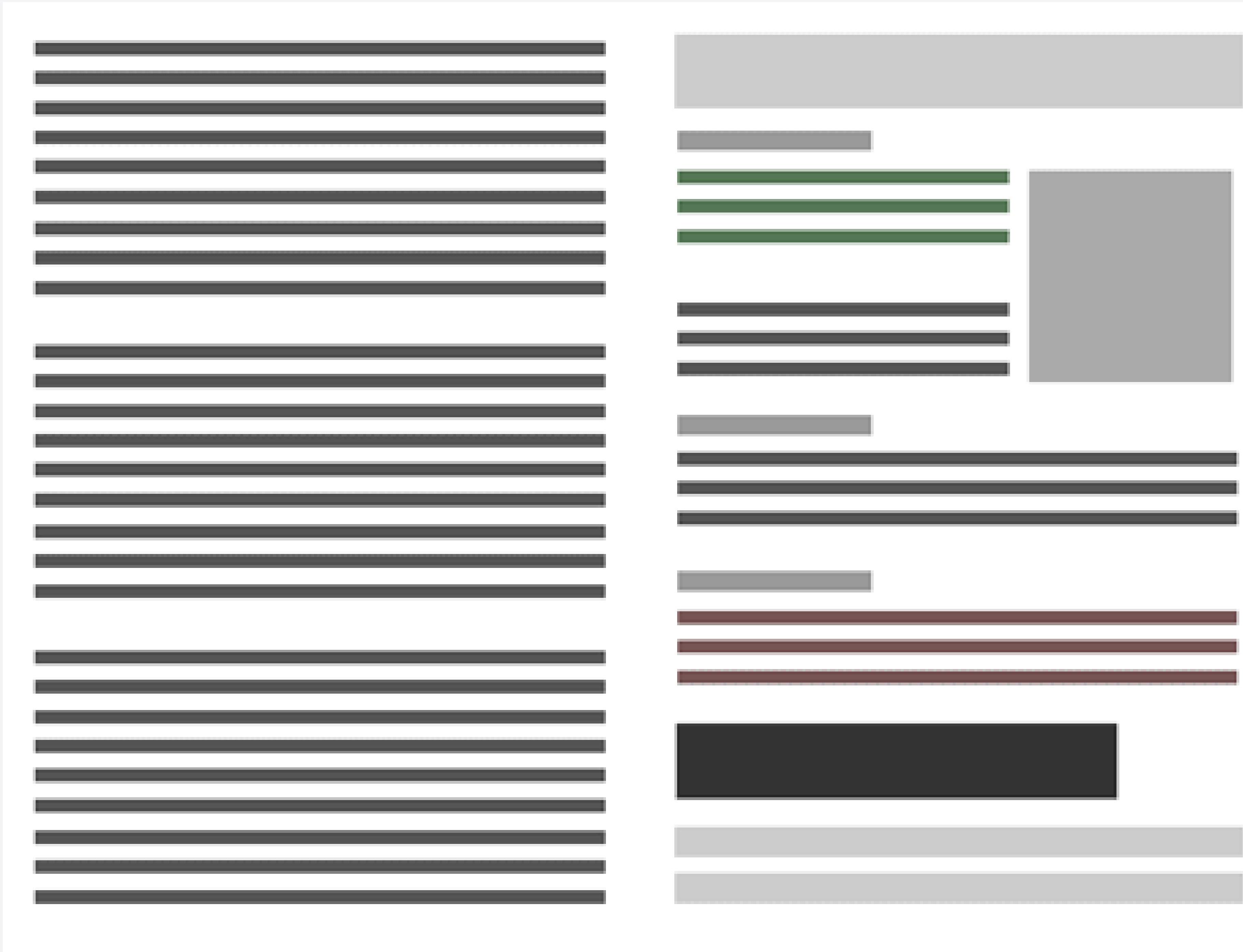


4.2. Jerarquía visual

Para diseñar una visualización de datos
necesitamos organizar la información.

Jerarquía visual

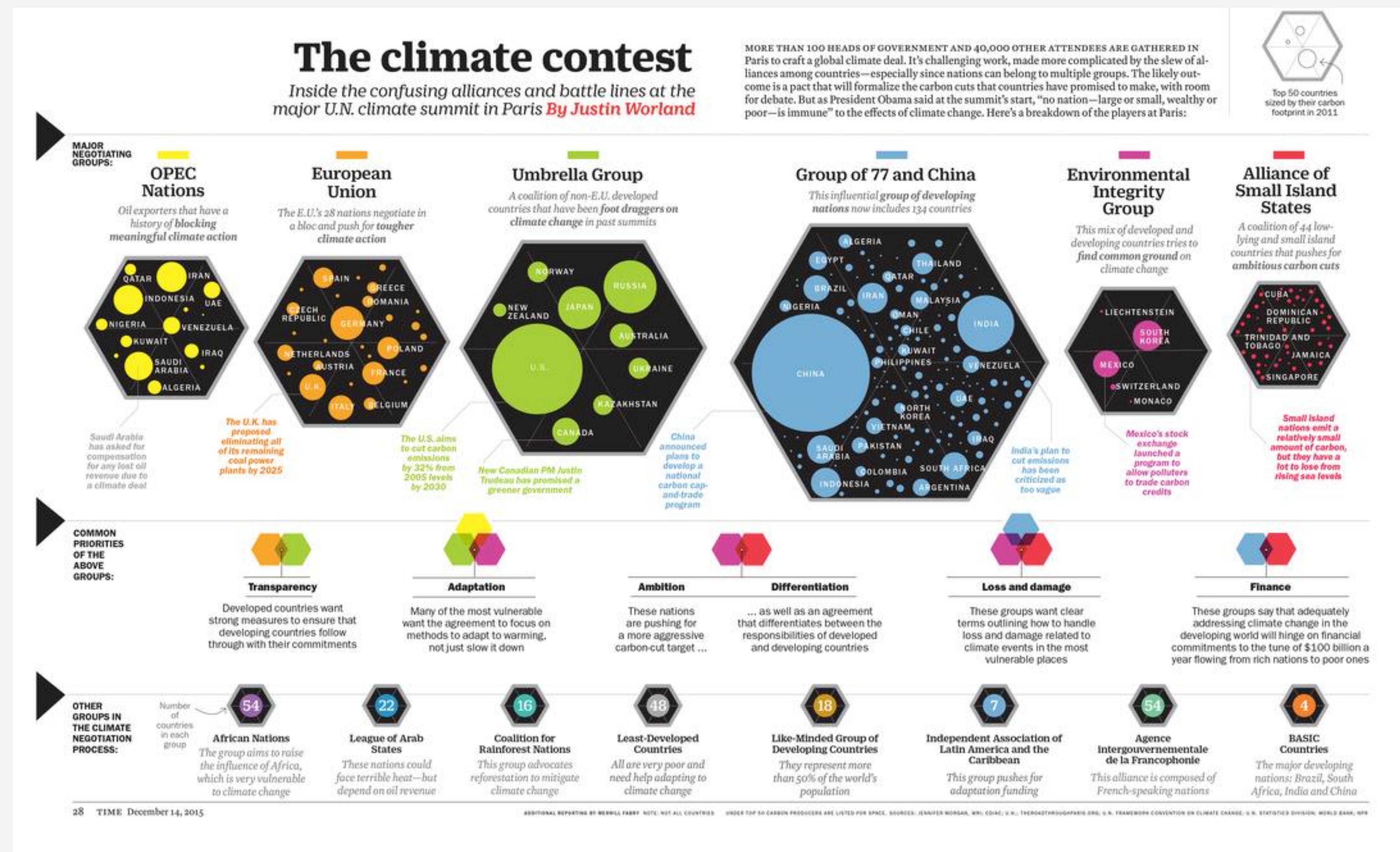
- ¿Lo que estoy buscando esta en esta visualización?
- ¿Dónde?
- ¿Cómo completo mi tarea?



La importancia visual no se puede aplicar
a demasiados elementos de diseño,
de lo contrario todo se vuelve igual.

Los 5 pilares de la jerarquía visual

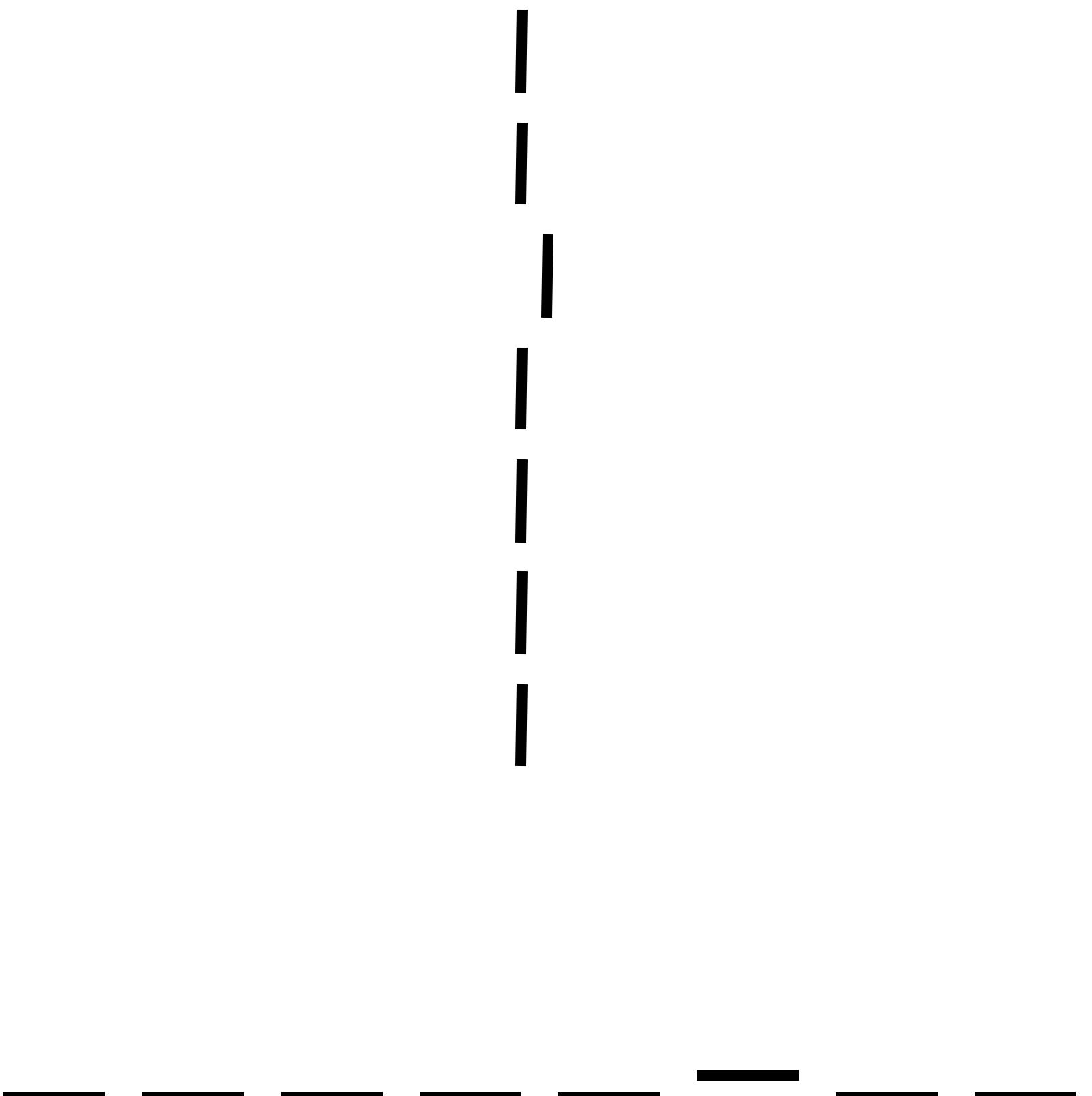
1. Tamaño: Cuanto más grande, más se nota (pero no siempre es mejor)



El contraste muestra la importancia relativa

2. Composición

- Una parte esencial de organizar los elementos pasa por **alinearlos**
- Somos muy sensibles a pequeñas diferencias en alineación
- Alinear los elementos ayuda a crear orden
- Estructura la información y ayuda a predecir donde encontrar información -> donde seguir mirando



AIR POLLUTION LEVELS IN ITALY ARE DECREASING



AIR POLLUTION LEVELS IN ITALY ARE DECREASING

JAN 20 FEB 2 FEB 16 MAR 1 MAR 15 MAR 20

PM 2.5
(MG / M³)

20

10

NO₂
(PPB)

30

20

10



Columnas o sistemas modulares

The Grid System

The ultimate resource in grid systems.

"The grid system is an aid, not a guarantee. It permits a number of possible uses and each designer can look for a solution appropriate to his personal style. But one must learn how to use the grid; it is an art that requires practice."

Josef Müller-Brockmann

Search

Articles	Tools	Books	Templates	Blog	Inspiration
Compose to a Vertical Rhythm On the Web, vertical rhythm is contributed to by three factors: font size, line height and margin or padding. All of these factors must be calculated with care in order that the rhythm is maintained. 04.Dec.2008	960 Grid System An effort to streamline web development workflow by providing commonly used dimensions, based on a width of 960 pixels. There are two variants: 12 and 16 columns, which can be used separately or in tandem. 04.Dec.2008	Geometry of Design The book focuses on the classic systems of proportioning, such as the golden section and root rectangles, as well as systems such as the Fibonacci Series. 04.Dec.2008	InDesign 8.5x11 Grid System (12) Adobe InDesign file with a grid system for an 8.5"x11" page that is divided into 12 columns and rows using the Rule of Thirds (Golden Ratio). Includes a 12pt baseline grid. 29.Nov.2008	UX Magazine A well designed collaborative site, with a very nice grid structure, that focuses on user experience. 02.Dec.2008	Ace Jet 170 AisleOne Athletics BBDK Blanka Build Corporate Risk Watch David Airey Dirty Mouse Experimenta Experimental Jetset Form Fifty Five Grafik Magazine Grain Edit Graphic Hug Helvetica Film I Love Typography Lamosca magCulture Mark Boulton Minimal Sites Monocle Neubau NewWork OK-RM Original Linkage Robin Uleman SampsonMay Schmid Today September Industry Sonifier Souellis Subtraction Swiss Legacy Thinking for a Living This Studio Toko Visuelle Xavier Encinas Year of the Sheep
Incremental leading In editorial design, there is a technique used for sidenotes and boxouts that aligns to the baseline grid, or vertical rhythm. It's called incremental leading. 03.Dec.2008	Graph Paper by Konigi This graph paper is made for visual designers, interaction designers, and information architects. You'll find styles for wireframing, story boarding, plotting values and for drafting sitemaps. 03.Dec.2008	The Typographic Grid We consider this to be the academic part two to "Grid Systems." Hans Rudolf Bosshard tackles a deeper understanding of the complex grid. 30.Nov.2008	InDesign 11x17 Grid System (12) Adobe InDesign file with a grid system for an 11"x17" page that is divided into 12 columns and rows using the Rule of Thirds (Golden Ratio). Includes a 12pt baseline grid. 29.Nov.2008	Doane Paper Utility Notebook A portable notebook featuring a patent pending Grid+Lines stationery design that combines the benefits of grid and ruled lines onto a single sheet of paper. 28.Nov.2008	
Applying Divine Proportion to Your Web Designs This article explains what is the Divine proportion and what is the Rule of Thirds and describes how you can apply both of them effectively to your designs. 01.Dec.2008	Syncotype Syncotype is a simple tool to help align your text to a baseline grid. Enter your line height and offset in pixels in the Syncotype control box and click "Syncotype it" to overlay a baseline grid in red. 01.Dec.2008	Grid Systems Grid Systems provides a rich, easy-to-understand overview and demonstrates a step-by-step approach to typographic composition. 21.Nov.2008	Photoshop 975px Grid System (12) Adobe Photoshop file with a grid system for a 975px wide page that is divided into 12 columns and rows using the Rule of Thirds (Golden Ratio). Includes a 16px baseline grid. 29.Nov.2008	Replica Typeface Replica is a new typeface by Norm that was designed on a strict grid system. Available in the following weights: Regular, Italic, Light, Light Italic, Bold and Bold Italic. 21.Nov.2008	

[View All Articles →](#) [View All Tools →](#) [View All Books →](#) [View All Templates →](#) [View All Blog Posts →](#) [View All Grid Systems →](#)

Our Design Project



Lorem Ipsum
Lorem ipsum dolor sit amet, consectetur adipiscing elit. In eu diam non ante condimentum malesuada lacinia eu sit amet ligula.

Lorem Ipsum
Lorem ipsum dolor sit amet, consectetur adipiscing elit. In eu diam non ante condimentum malesuada lacinia eu sit amet ligula. Suspendisse posuere dolor vitae laoreet varius. Vivamus eget felis rutrum, venenatis felis.

Lorem Ipsum
Lorem ipsum dolor sit amet, consectetur adipiscing elit. In eu diam non ante condimentum malesuada lacinia eu sit amet ligula.

Canva

Las cuadrículas proporcionan consistencia y flexibilidad

1 column vertical grid



2 column vertical grid



2 column vertical grid



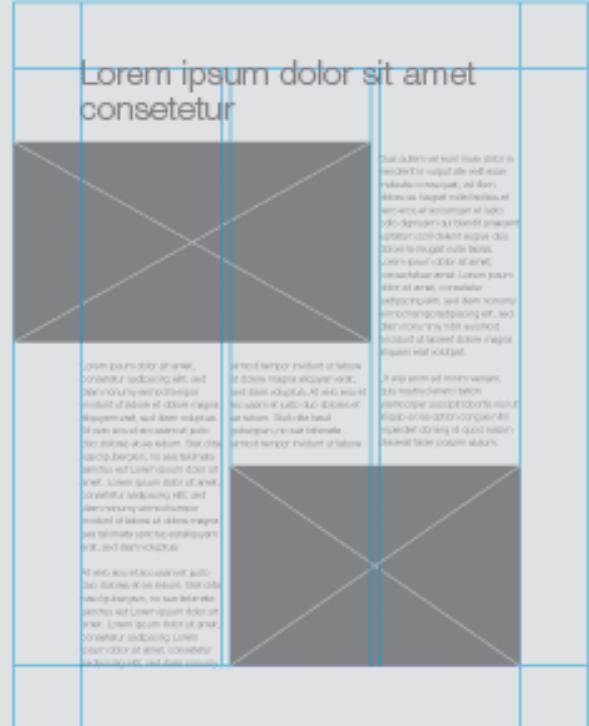
1 column landscape grid



2 column landscape grid



3 column vertical grid



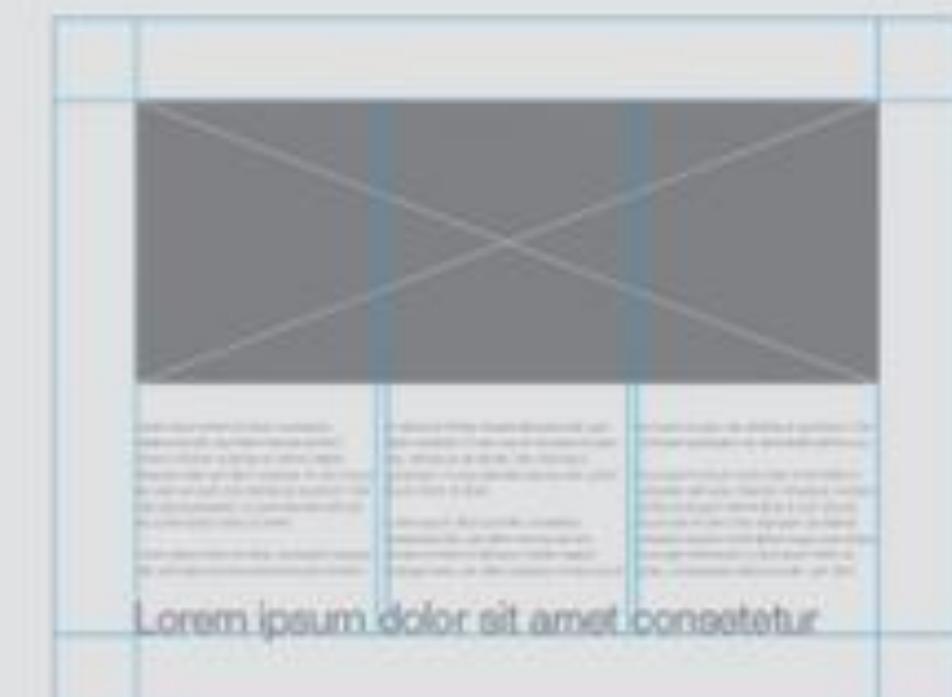
3 column vertical grid



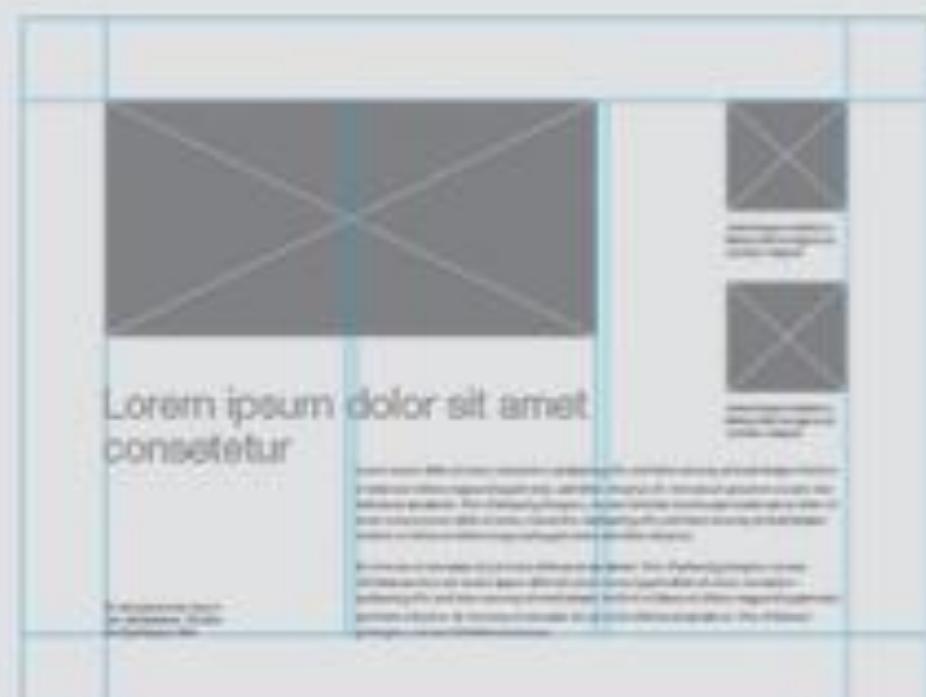
3 column vertical grid



3 column landscape grid



3 column landscape grid



3. Espacio

*Gestalt:

Proximidad / Similaridad

agrupa elementos dentro de una jerarquía y crea nuevas subjerarquías.

*Ritmo:

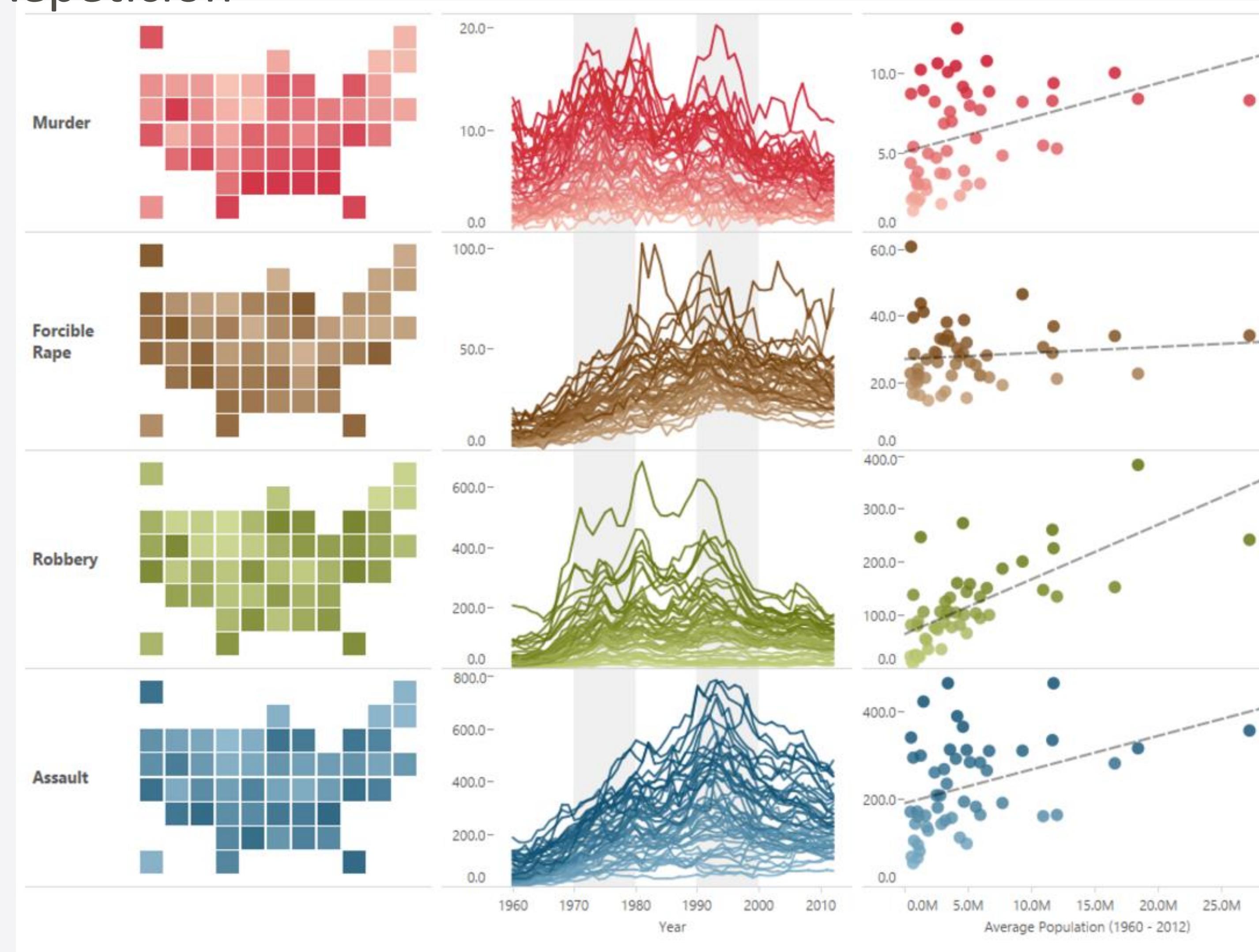
Repetición genera un ritmo visual que puede ser continuo o discontinuo, creciente o descendiente.

*Espacio negativo o vacío:

Cuento menos elementos tengas, más potente.

Además, el espacio positivo pesa más que el espacio negativo

3. Espacio: Repetición



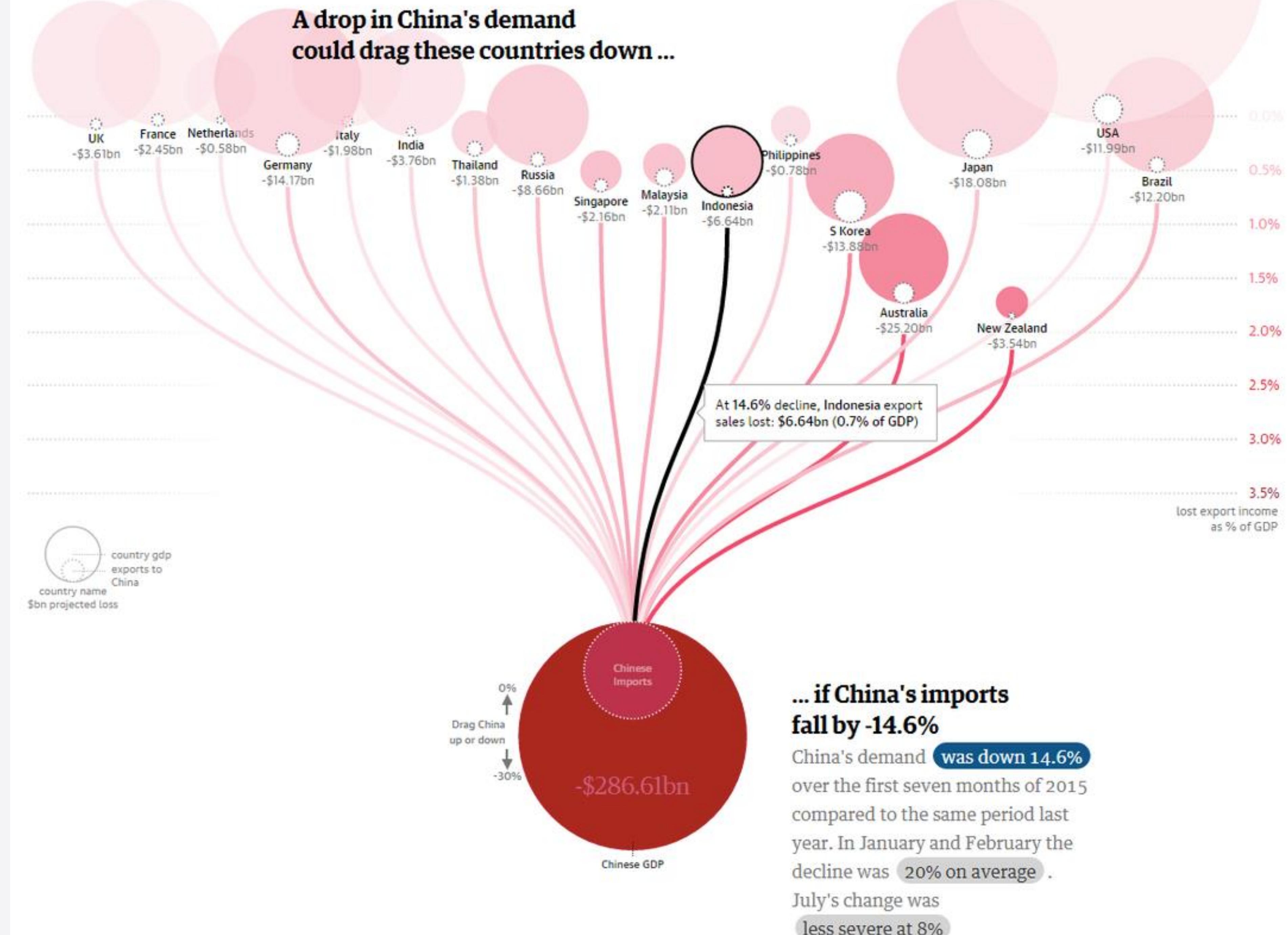
La repetición otorga significado a los nuevos elementos: están en el mismo nivel en la jerarquía
La alineación crea orden. Permite conectar rápidamente elementos en toda la página
Un solo elemento que rompe la alineación llama la atención sobre sí mismo y su importancia

3. Espacio: Elemento central



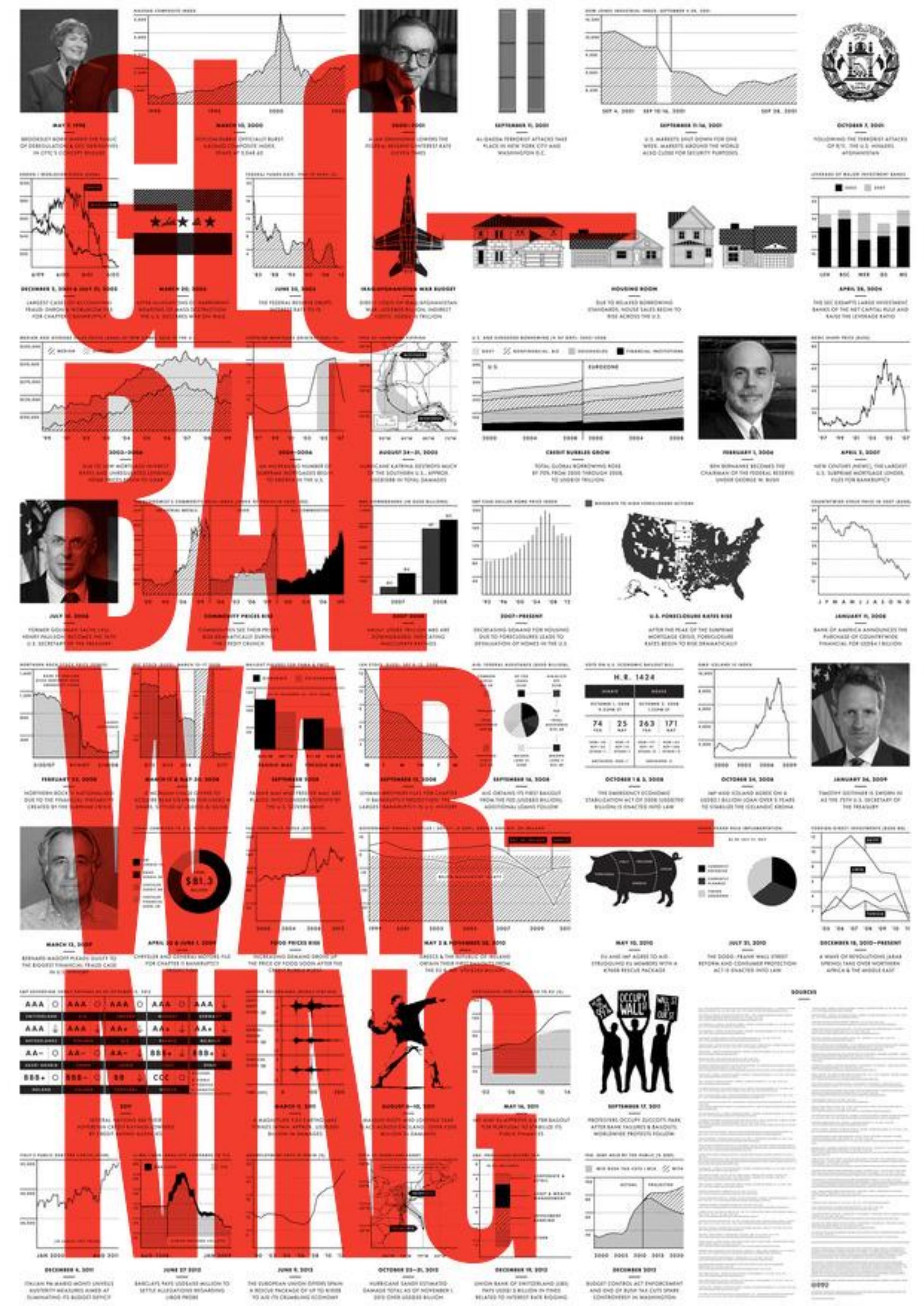
3. Espacio:

Espacio negativo



4. Color

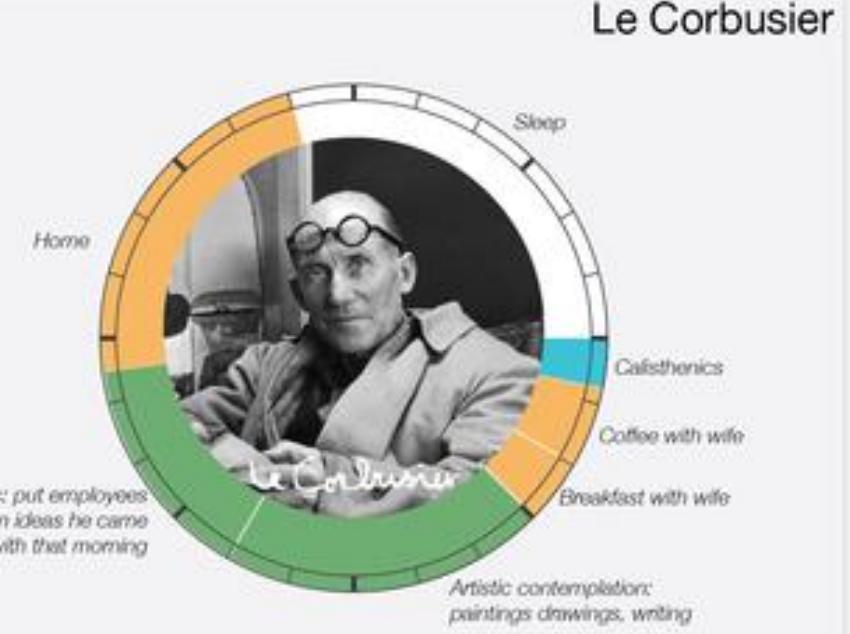
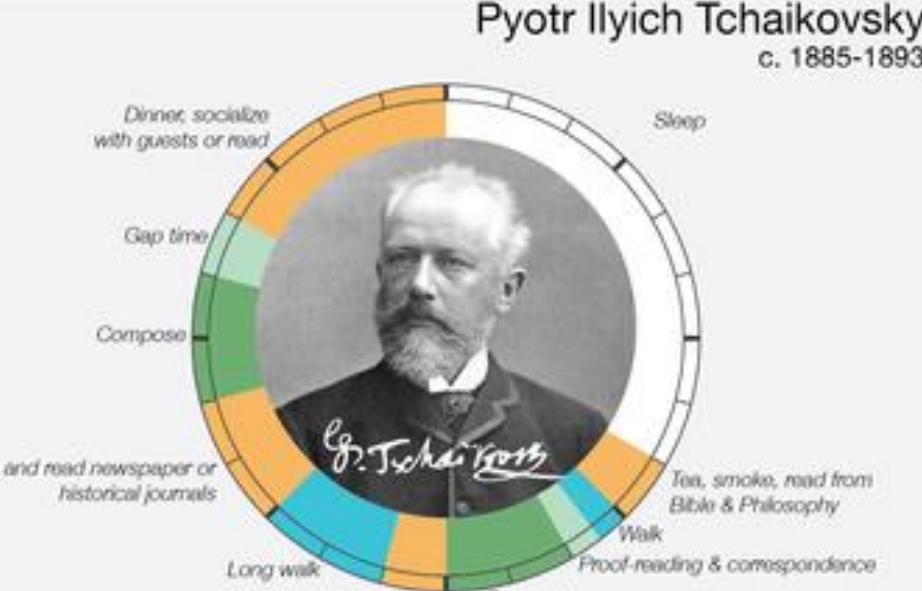
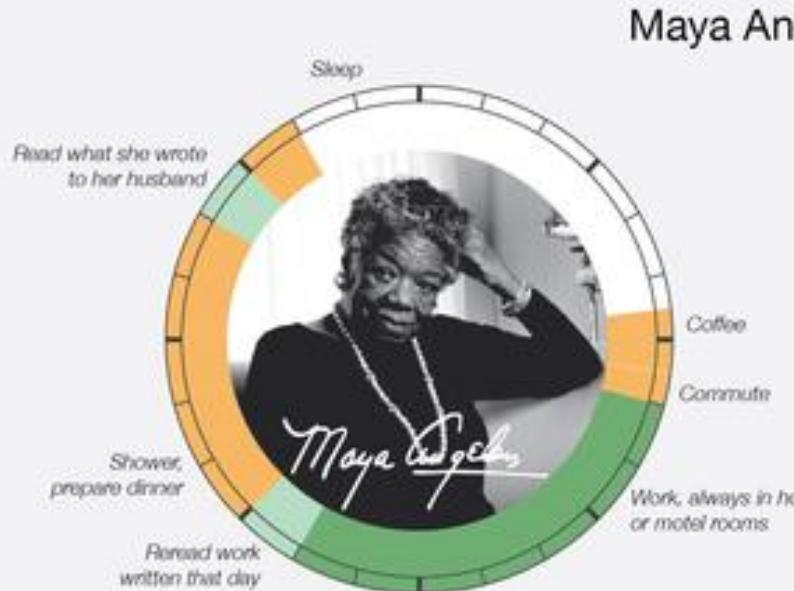
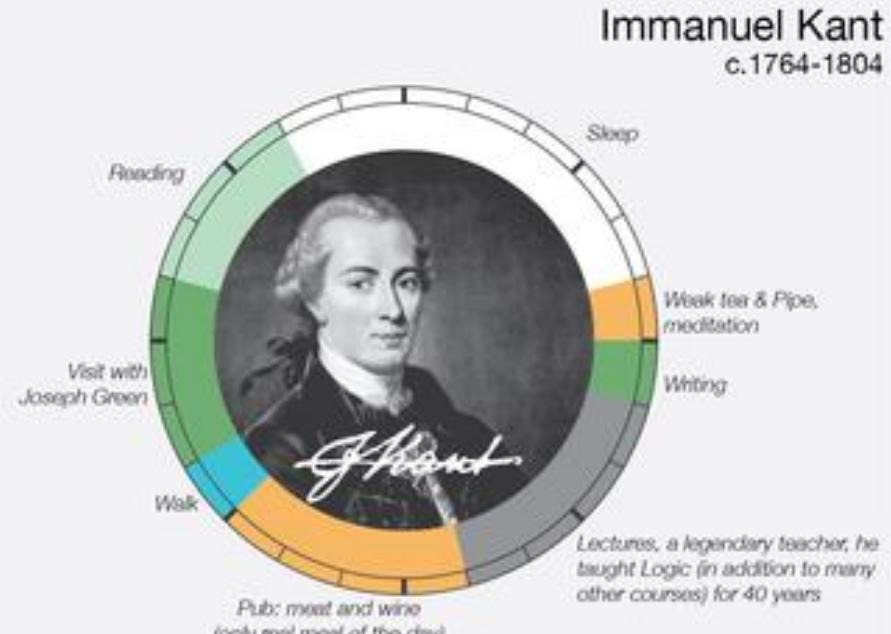
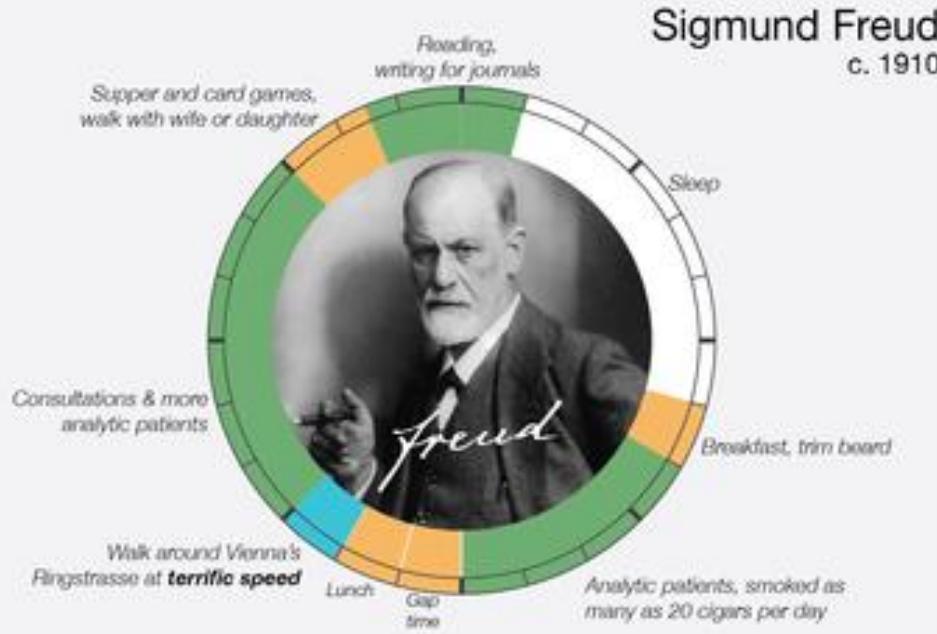
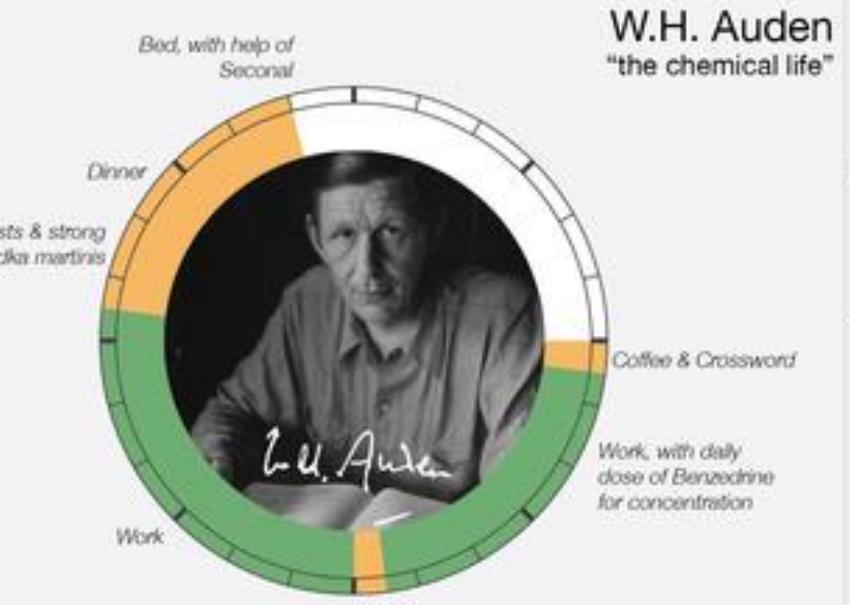
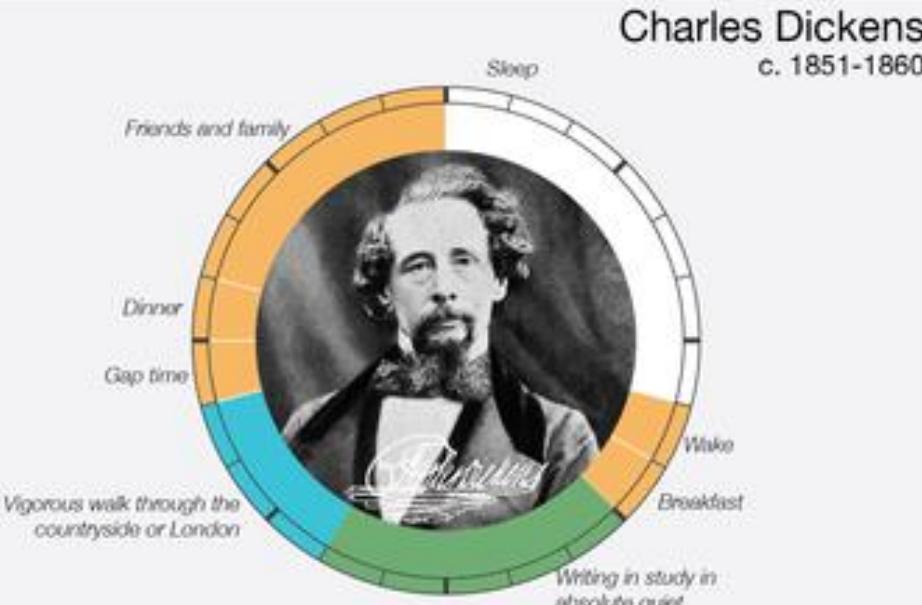
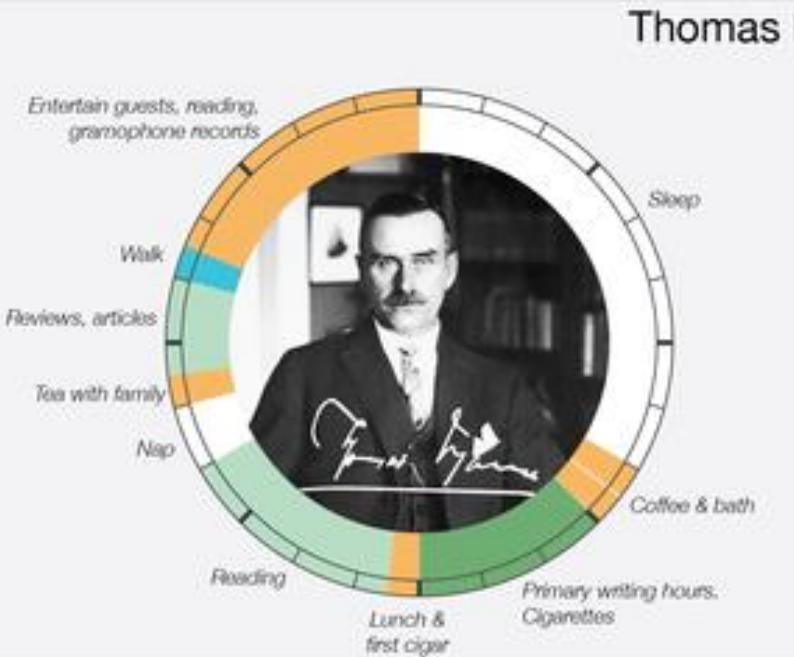
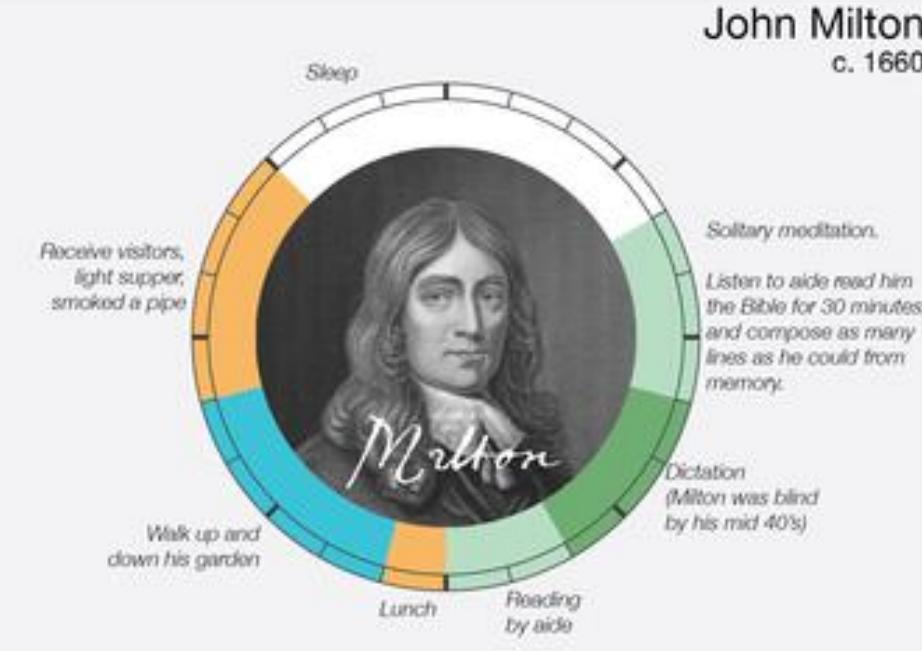
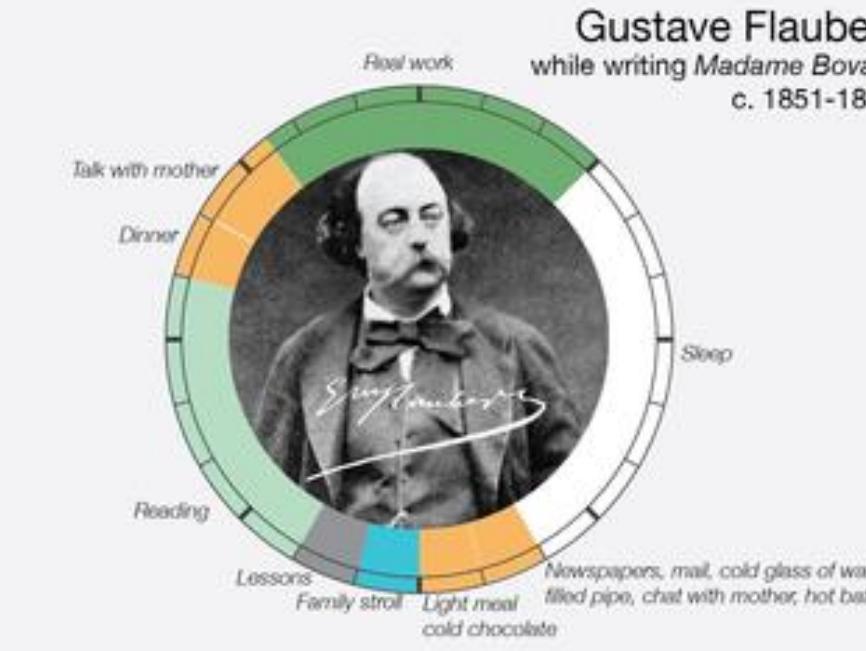
CONTRASTE



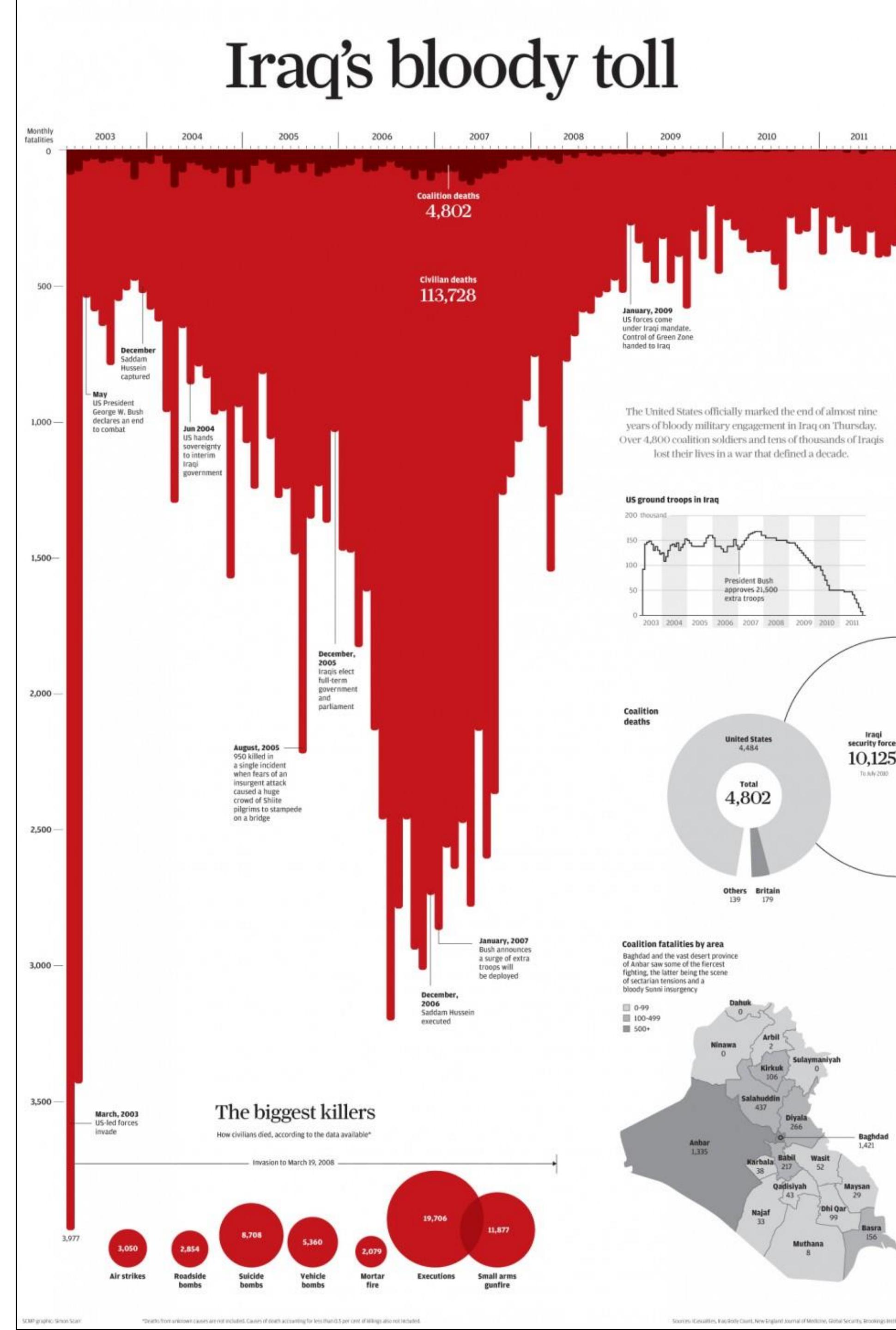
4. Color: SIMILITUD

CREATIVE ROUTINES

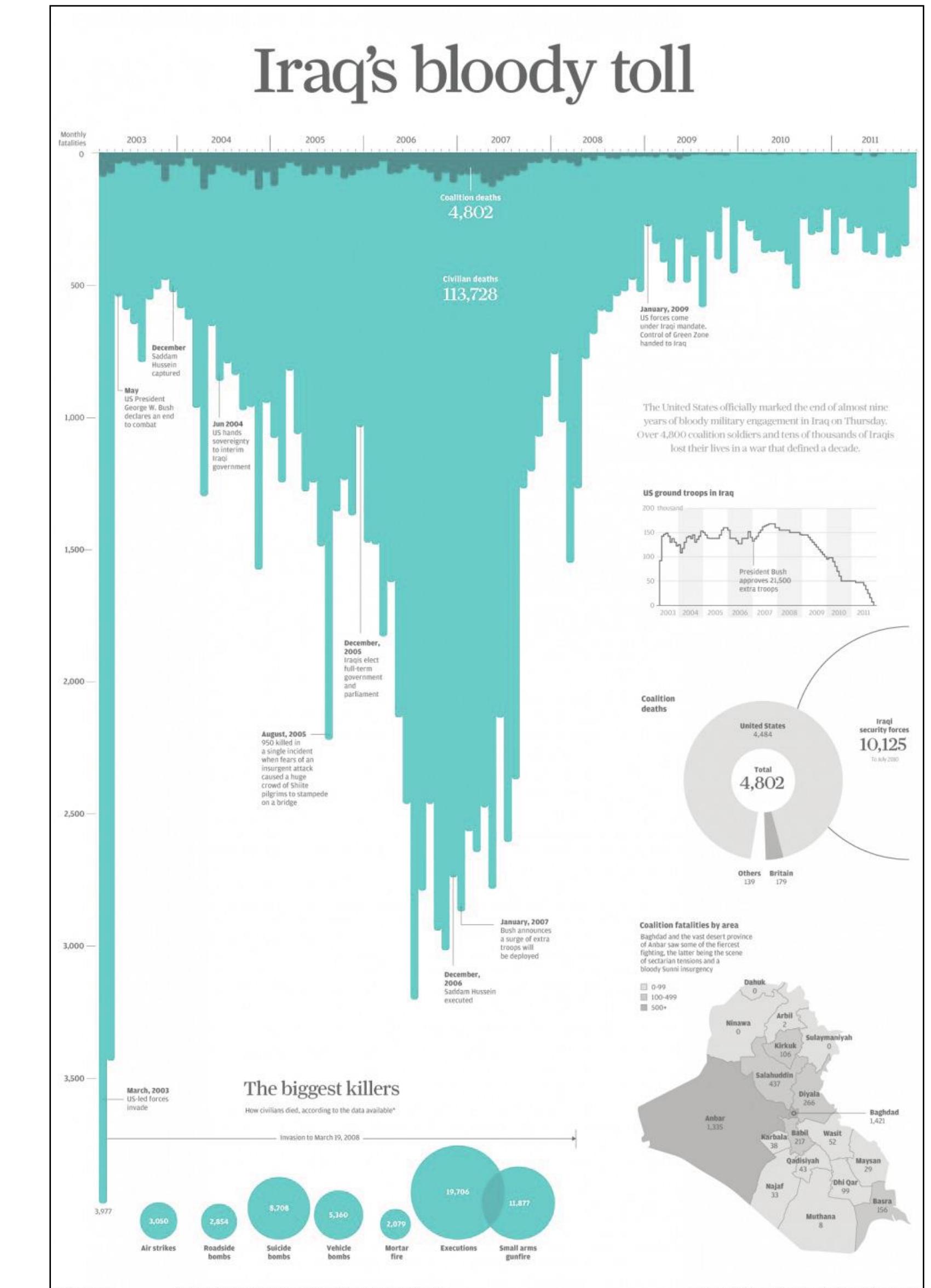
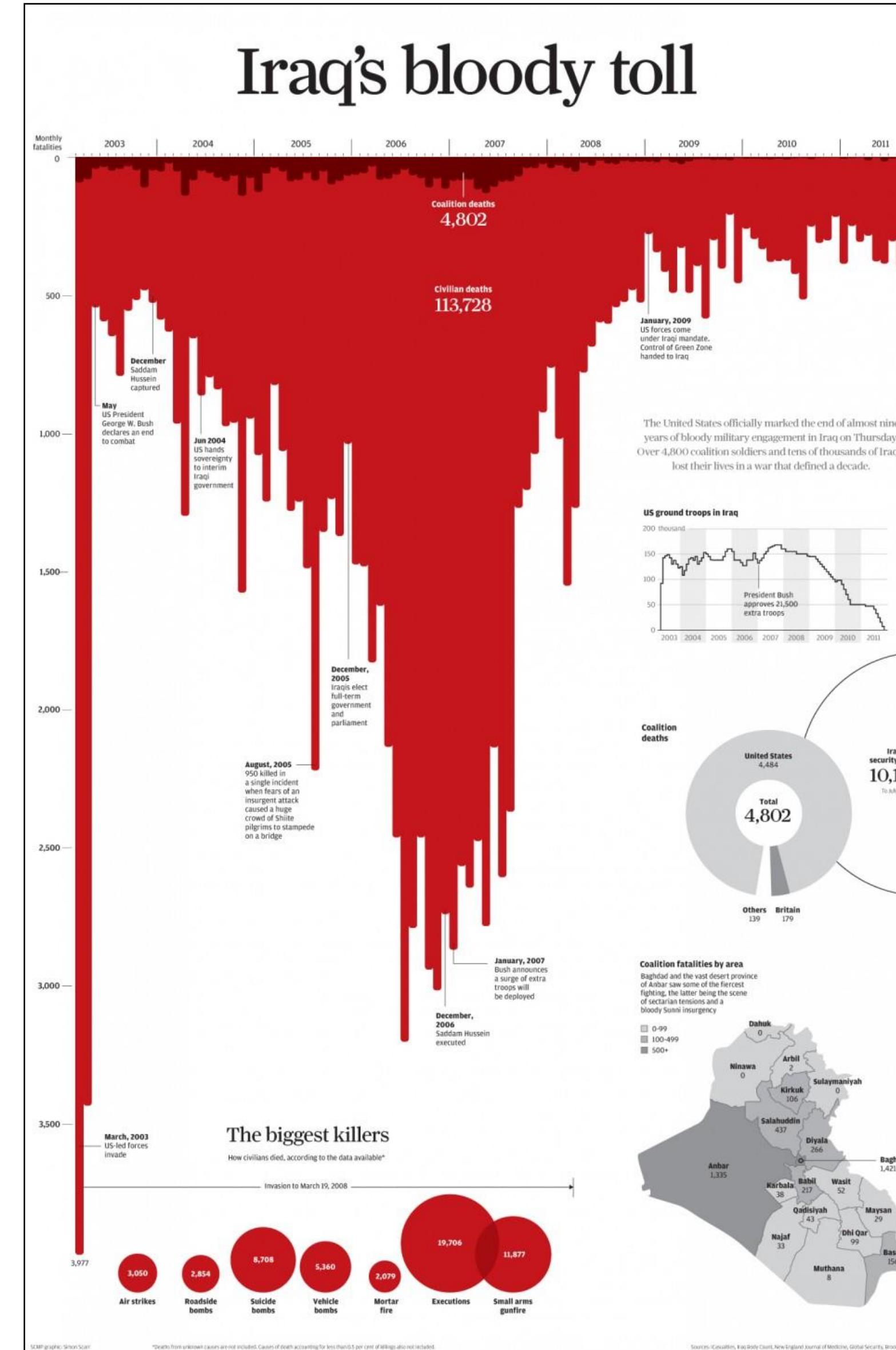
"In the right hands, it can be a finely calibrated mechanism for taking advantage of limited resources... a solid routine fosters a well-worn groove for one's mental energies...." - Mason Currey, author of the inspiring book, DAILY RITUALS



4. Color: SEMÁNTICA



4. Color: SEMÁNTICA



5. Flujo Visual

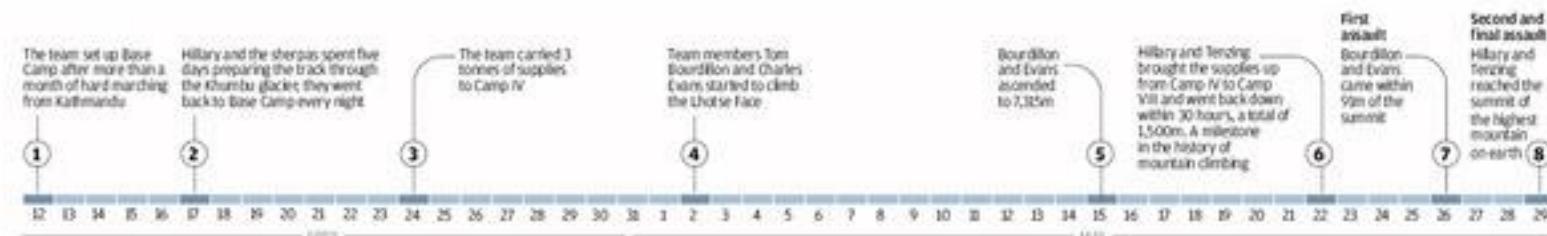
Podemos crear flujo visual con cualquiera de los elementos anteriores, y también con ojos, manos, líneas, triángulos y flechas sutiles o explícitas.

5. Flujo Visual

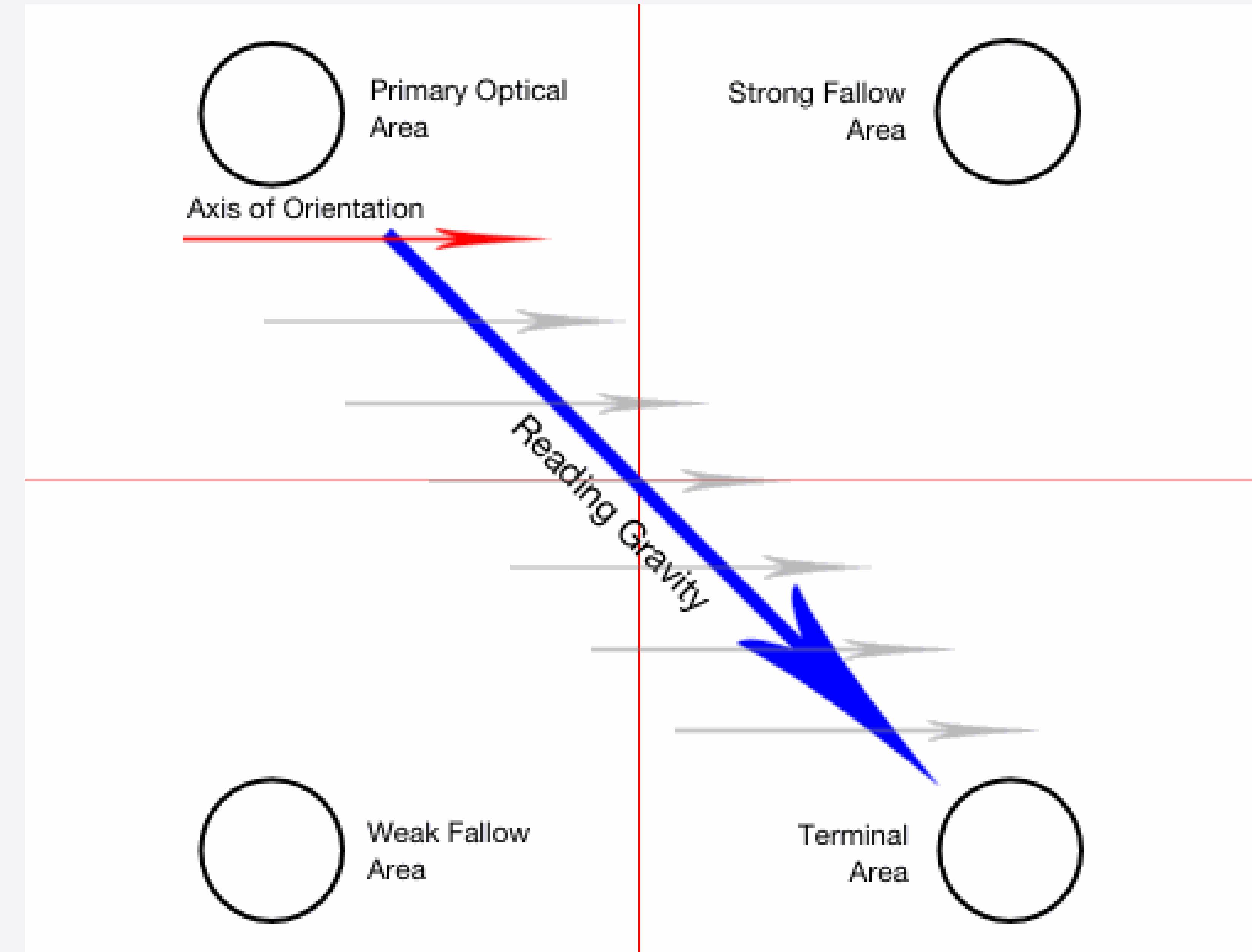
Mount Everest

Sixty years ago, Edmund Hillary and Tenzing Norgay became the first climbers to reach the summit of Mount Everest, known as Qomolangma in China and Chomolungma in Tibet. Their bravery and resilience set an example for many other adventurers in the following decades.

The challenge of scaling Everest still attracts thousands of people from all walks of life who want to push themselves to the limit.



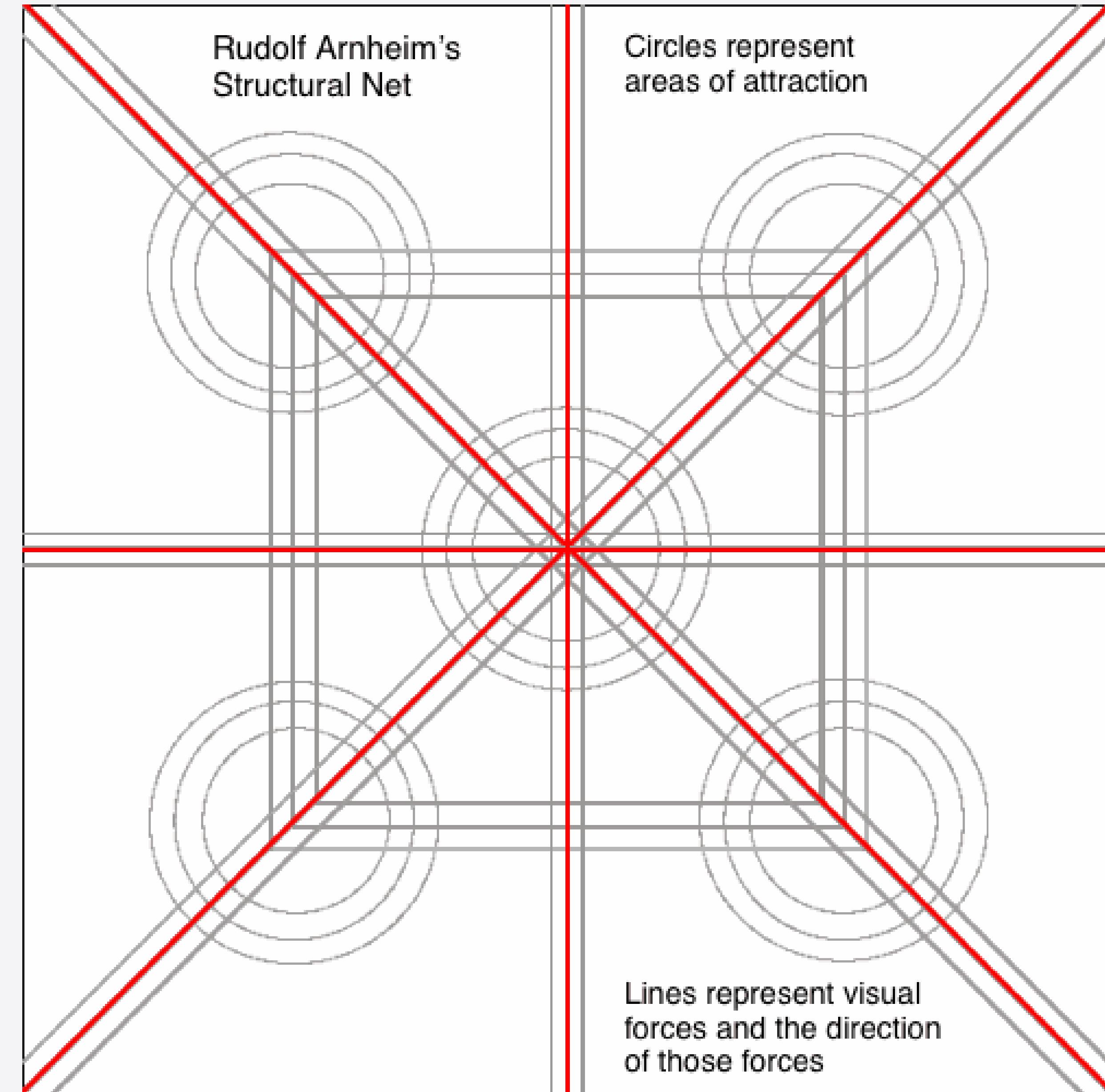
5. Flujo Visual



Cuando el diseño no está presente

6. Flujo Visual

Puntos naturales de atracción
Y líneas de flujo



6. Estilo Gráfico



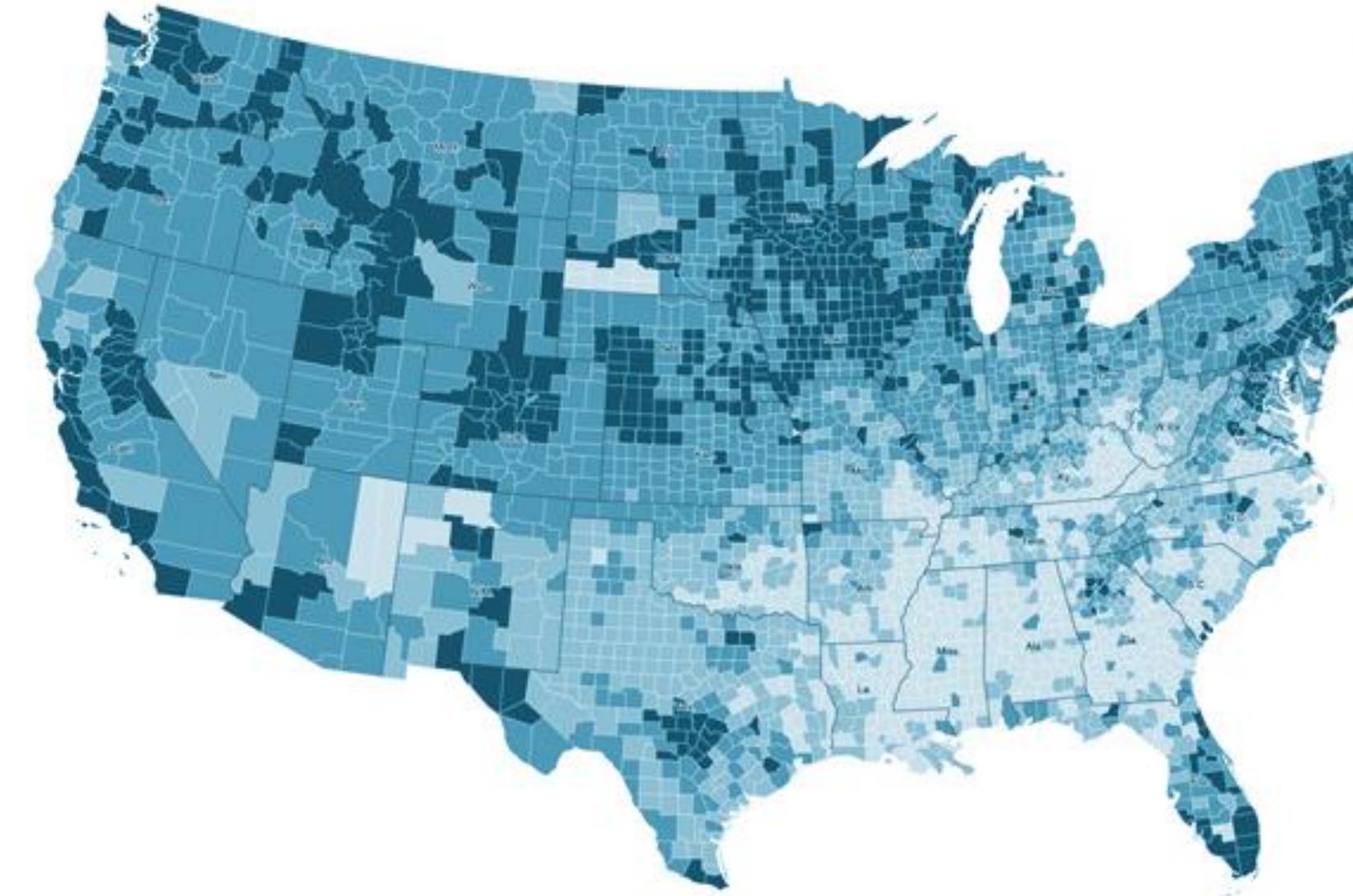
Recomendación: no usar más de 3 niveles de jerarquía tipográfica
Valor: un objeto más oscuro tendrá más peso que un objeto más claro

Life Expectancy in the US

Comparing life expectancy across our nation and the State of Virginia

Created by Boost Labs using US Census Bureau Data available as of June, 2011

US Male Life Expectancy by County



Legend

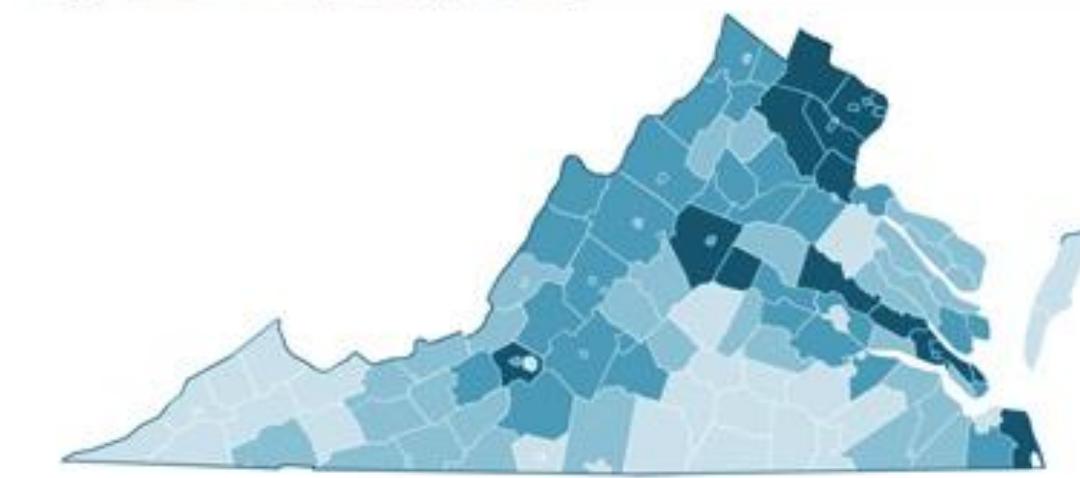
Male Life Expectancy

- 70–72 yrs
- 72–74 yrs
- 74–76 yrs
- 76–78 yrs

Female Life Expectancy

- 78–79 yrs
- 79–80 yrs
- 80–82 yrs
- 82–84 yrs

Virginia Male Life Expectancy by County



US Quick Stats

US Average Life Expectancy in Years

78.7

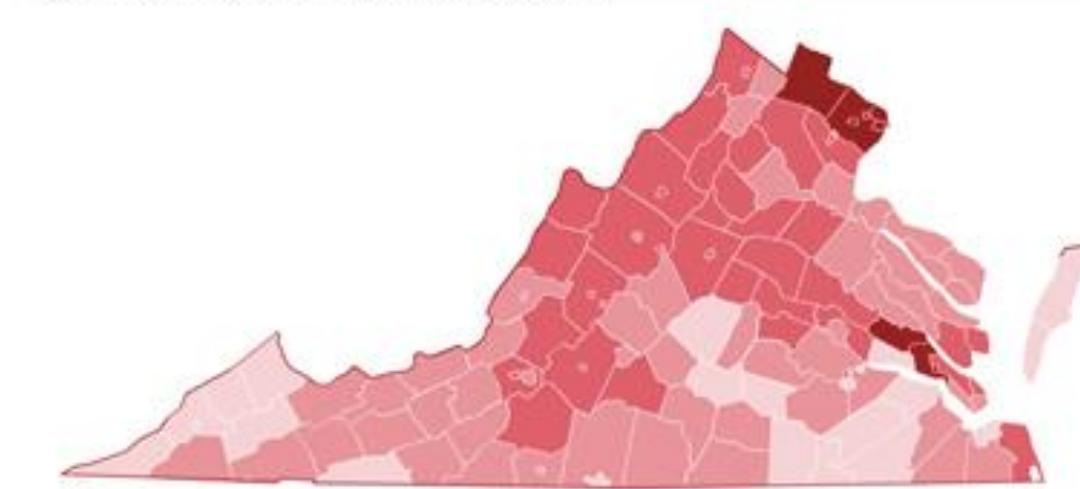
Males

75.9

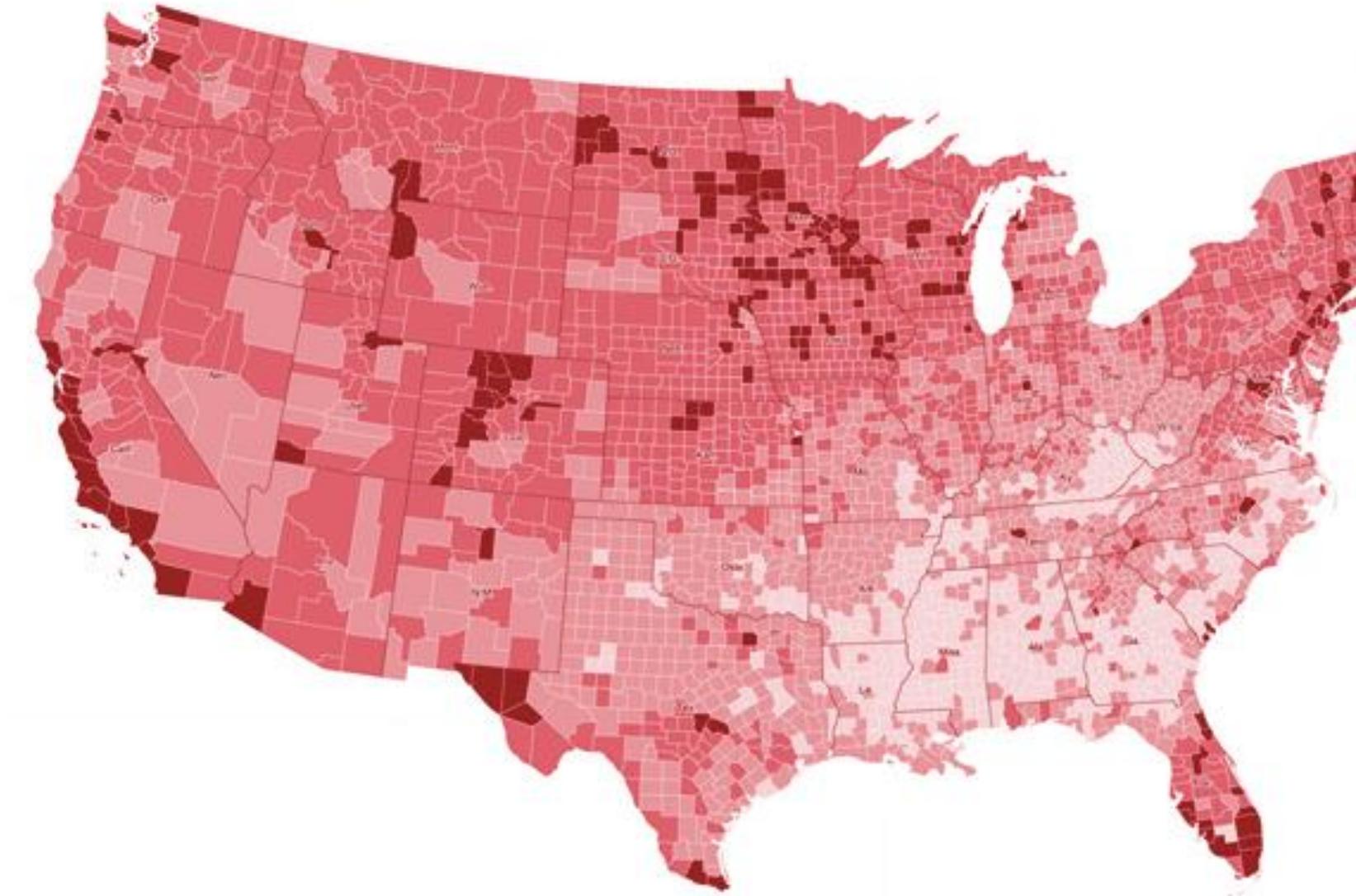
Females

81.1

Virginia Female Life Expectancy by County



US Male Life Expectancy by County



Virginia Quick Stats

US Average Life Expectancy in Years

78.6

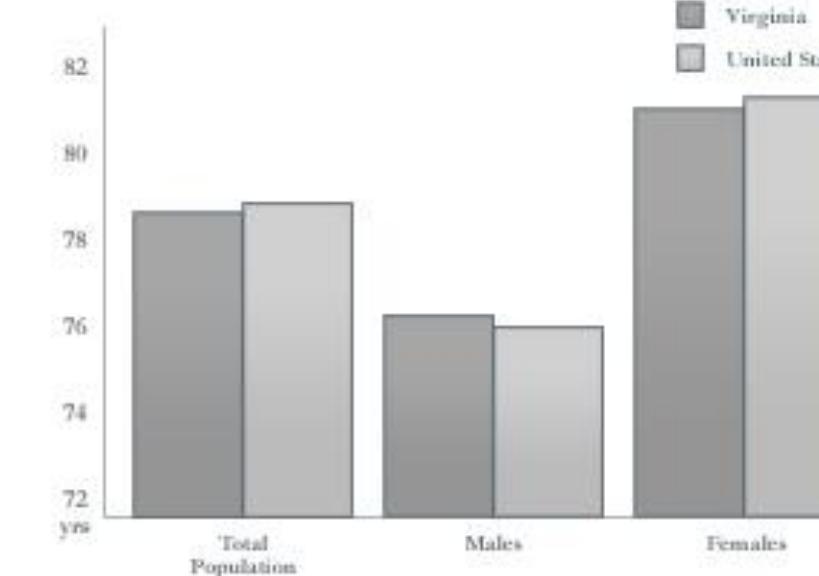
Males

76.1

Females

80.9

Estimated Life Expectancy at Birth by Gender



US Population Stats

US Total Population

308,745,538

Female persons

50.7%

White persons

72.4%

Persons under 18 years old

24.3%

Persons 65 years old and over

12.9%

Virginia Population Stats

Virginia Total Population

8,001,024

Female persons

50.8%

White persons

68.6%

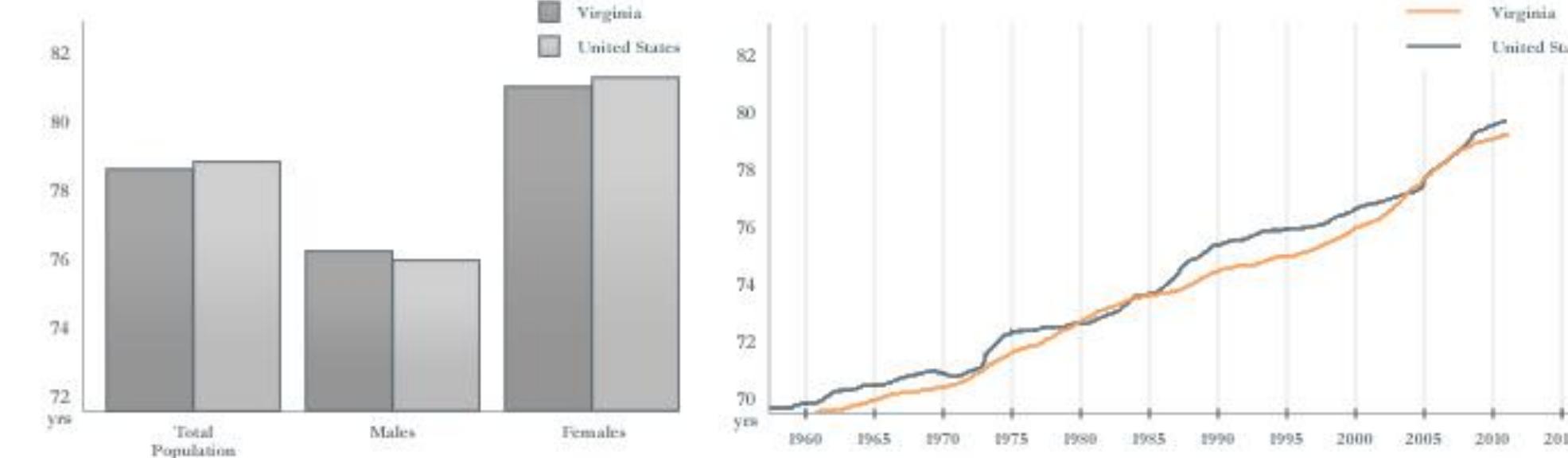
Persons under 18 years old

23.4%

Persons 65 years old and over

12.2%

Estimated Life Expectancy at Birth by Year



WAKE UP

Mohamad Waked

Forward

Textbox

Forward

Backward

Forward

Backward

Textbox

Wake Up

The Tragedy of US School Shootings

By Mohamad Waked

For the best experience, view this project on a laptop or large screen with a 16:10 aspect ratio. Use the buttons on the top left and right to navigate and interact.

www.alhadaqa.com

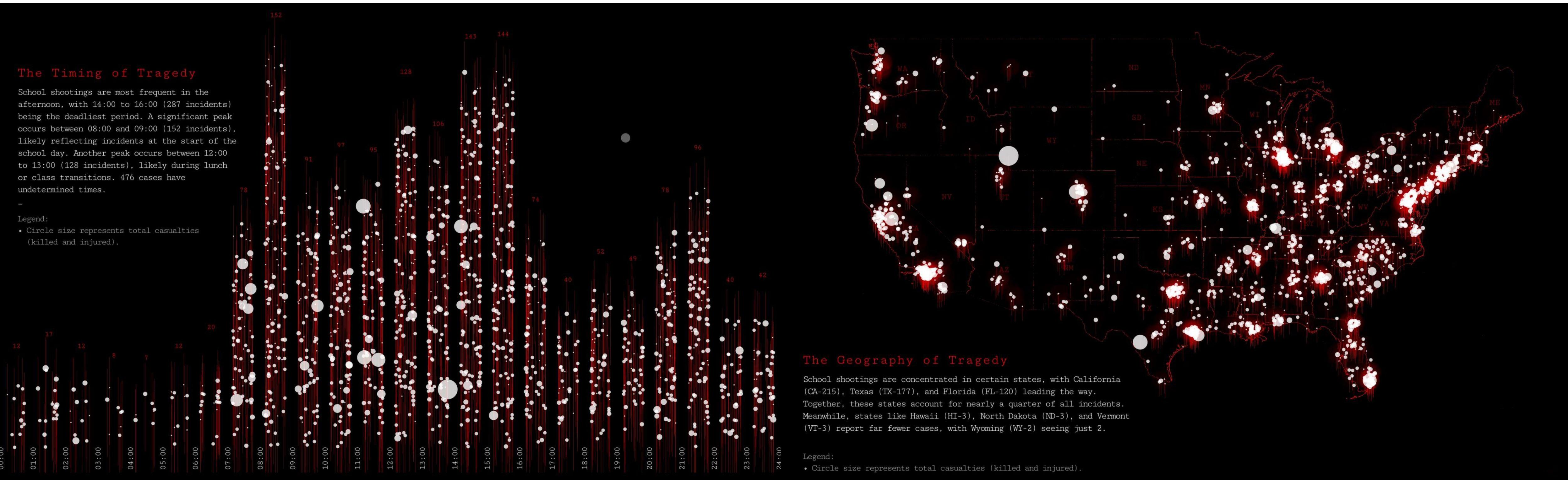
The Last Morning Light

"My boy, wake up."
The sunlight spilled over his face, soft and warm.
"Five more minutes," he murmured, pulling the blanket tight.
"Not today," I said, brushing his messy hair.
He sat up slowly, rubbing sleep from his eyes.
"Do I really have to?" he asked, still half-dreaming.
"Yes," I replied. "Every day is a gift, my boy."

<https://alhadaqa.github.io/wakeup/>

WAKE UP

Mohamad Waked



<https://alhadaqa.github.io/wakeup/>

EJERCICIO 1

- Abrir *life_expectancy.csv* o *.xlsx* (Excel, Preview de mac, GoogleSheets, Python, etc).
- Es necesario **formatear** antes de usarlo.
- Derivar dataset filtrando **un solo año**.

- Hacer gráficas para visualizar las preguntas:
 - ¿Top 20 países con mayor EV media?
 - ¿Top 20 países con mayor diferencia por sexo?

- Identificar tipos de datos: cuantitativos, ordinales o categóricos.
- Abrir la web www.datawrapper.de
- Hacer gráficas según el planteamiento anterior de Datos, Tareas y Codificación.
 - Ranking ordenado de Both Sexes por país.
 - ¿Qué gráfica mostraría más claramente la diferencia entre ambos sexos por país?

EJERCICIO 2

- Abrir *life_expectancy.csv* o *.xlsx*. **Necesario formatear antes de usar.**
- Queremos visualizar la **evolución temporal** de **Both Sexes** por país, para todos los países a la vez.
- ¿Cómo formatear los datos para hacer esa visualización? **Haz las operaciones necesarias**
- Se puede hacer en www.datawrapper.de
- ¿Qué gráfica sería más adecuada?
 - Cuantos atributos y de qué tipo?
 - Keys? Values?
 - Marcas: ?
 - Canales: ?
 - Tareas:
 - Acción: Identificar tendencias en el tiempo
 - Objetivo: Variable Both Sexes, país y año
 - Escalable a ~200 países

EJERCICIO 2

- Abrir *life_expectancy.csv* o *.xlsx* (Excel, Preview de mac, GoogleSheets, Python, etc). **Necesario formatear antes de usar.**
 - Queremos visualizar la evolución temporal de Both sexes por país, para todos los países.
 - ¿Cómo formatear los datos para hacer esa visualización?
 - **Haz las operaciones necesarias para transformar los datos**

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Row Labels	Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and	Argentina	Armenia	Australia	Austria	Azerbaijan	Bahamas
2	2000	54,8	72,6	71,3		45,3	73,6	74,1	72	79,5	78,1	66,6	72,6
3	2001	55,3	73,6	71,4		45,7	73,8	74	72,6	79,9	78,6	67,5	72,9
4	2002	56,2	73,3	71,6		46,5	74	74,1	72,6	79,9	78,7	67,8	73,1
5	2003	56,7	72,8	71,7		46,8	74,2	74,1	72,7	80,3	78,8	67,8	73,2
6	2004	57	73	72,3		47,1	74,4	74,7	73	80,6	79,3	68,4	73,8
7	2005	57,3	73,5	72,9		47,4	74,6	74,9	73	81	79,4	68,4	74,1
8	2006	57,3	74,2	73,4		47,7	74,8	75,2	72,9	81,2	79,8	69,2	74,2
9	2007	57,5	75,9	73,8		48,2	75	74,8	73,5	81,3	80,1	70,3	74,4
10	2008	58,1	75,3	74,1		48,7	75,2	75,4	73,2	81,3	80,4	70,3	74,5
11	2009	58,6	76,1	74,4		49,1	75,4	75,6	73,3	81,7	80,2	70,8	74,6
12	2010	58,8	76,2	74,7		49,6	75,6	75,5	73,5	81,9	80,4	71,1	75
13	2011	59,2	76,6	74,9		50,1	75,7	75,7	73,9	82	80,8	71,6	75
14	2012	59,5	76,9	75,1		50,6	75,9	75,9	74,4	82,3	80,8	71,9	74,9
15	2013	59,9	77,2	75,3		51,1	76,1	76	74,4	82,5	81,1	72,2	74,8
16	2014	59,9	77,5	75,4		51,7	76,2	76,2	74,6	82,7	81,4	72,5	75,4
17	2015	60,5	77,8	75,6		52,4	76,4	76,3	74,8	82,8	81,5	72,7	76,1

EJERCICIO 3

- Abrir *life_expectancy.csv* o *.xlsx* (Excel, Preview de mac, GoogleSheets, Python, etc).
- Queremos visualizar la **correlación** entre Male y Female para todos los países en el año 2015.
- Haz las operaciones necesarias para transformar los datos
- ¿Qué gráfica sería más adecuada?
- Abrir www.datawrapper.de
 - Cuantos atributos y de qué tipo?
 - Keys? Values?
 - Marcas: ?
 - Canales: ?
 - Tareas:
 - Encontrar patrones y tendencias, Analizar distribución e identificar outliers
 - Objetivos: Todo el dataset.
 - Debe ser escalable a ~200 paises

GRAU EN ENGINYERIA DE DADES
104365 Visualització de Dades

Teoria 6. Tractament de dades II

Departament de Matemàtiques

Data processing for visualization

➤ Chapter 5 - Data processing for visualization (I)

- Uncertainty and error
- Transformations and data massage (+ seminars & PRT1)

➤ Chapter 6 (today) - Data processing for visualization (II)

- Dimensionality reduction
- Computation and important metrics selection

6. Data processing for visualization (II). Contents:

1. Dimensionality reduction

1. Introduction
2. Correlograms
3. Feature projection – PCA
4. Discriminant analysis (linear) - LDA
5. T-Distributed stochastic neighbour embedding (t-SNE)
6. Tomography- Slice along a plane, 2D isosurfaces for a 3D field, isocontours

2. Computation and important metrics selection

1. Quality metrics: Noise reduction, clutter reduction, search outliers

6.1.1. Introduction: Dimensionality reduction

Dimensionality reduction (DR) means: the process of **transformation of data from high dimensional space to low dimensional space** while maintaining most of the meaningful insights from the original data.

The goal is to preserve the meaningful structure of a dataset while using fewer attributes to represent the items.

For example: We have a dataset contains hundred columns (i.e features) or it could be an array of points that make up a large sphere in the 3D space. DR?

DR entails lowering the number of columns to a smaller number, such as 2D.

6.1.1. Introduction: Dimensionality reduction

Dimensionality reduction (DR) has two primary use cases:

- data exploration
- machine learning.

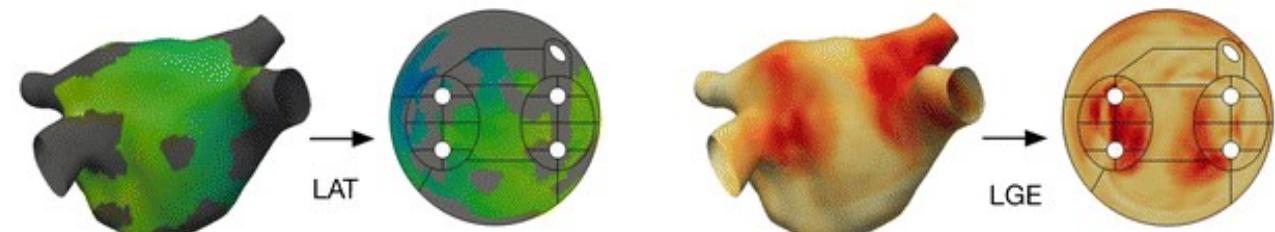
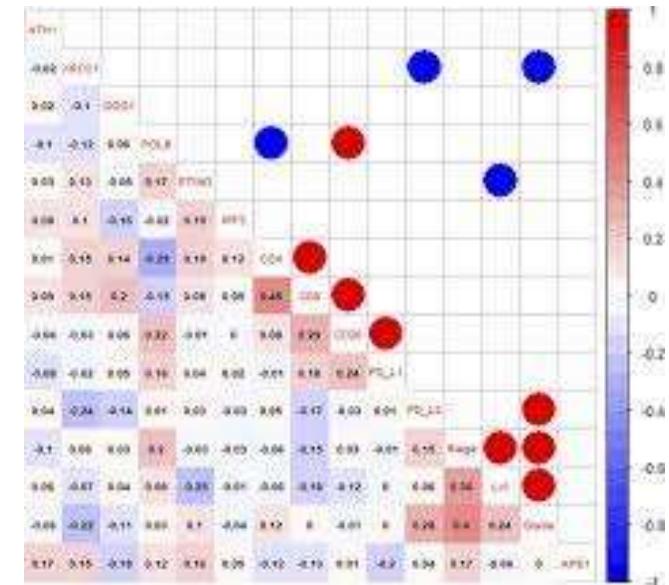
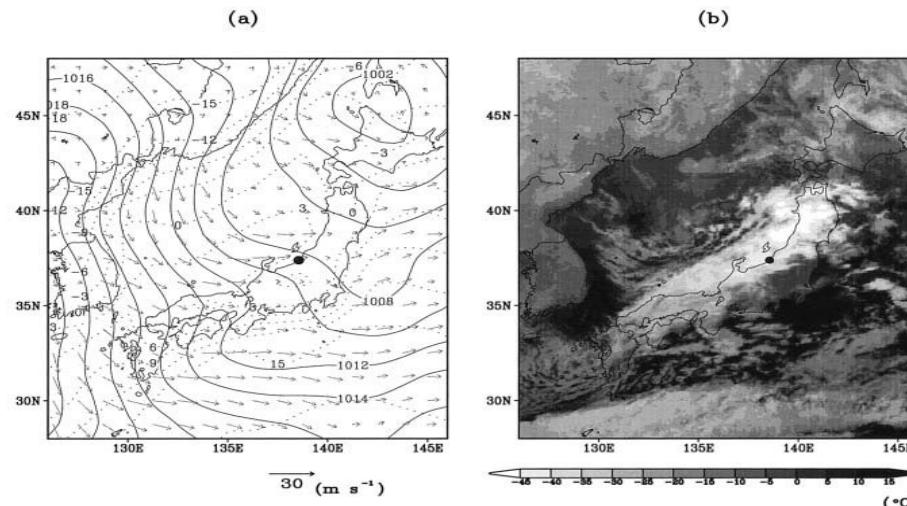
DR is a strategy for managing complexity in visualization:

It is useful *for data exploration* because **dimensionality reduction to few dimensions** (e.g., 2D or 3D) **allows for visualizing the samples**.

Such a visualization can then be used to obtain insights from the data (e.g., detect clusters and identify outliers).

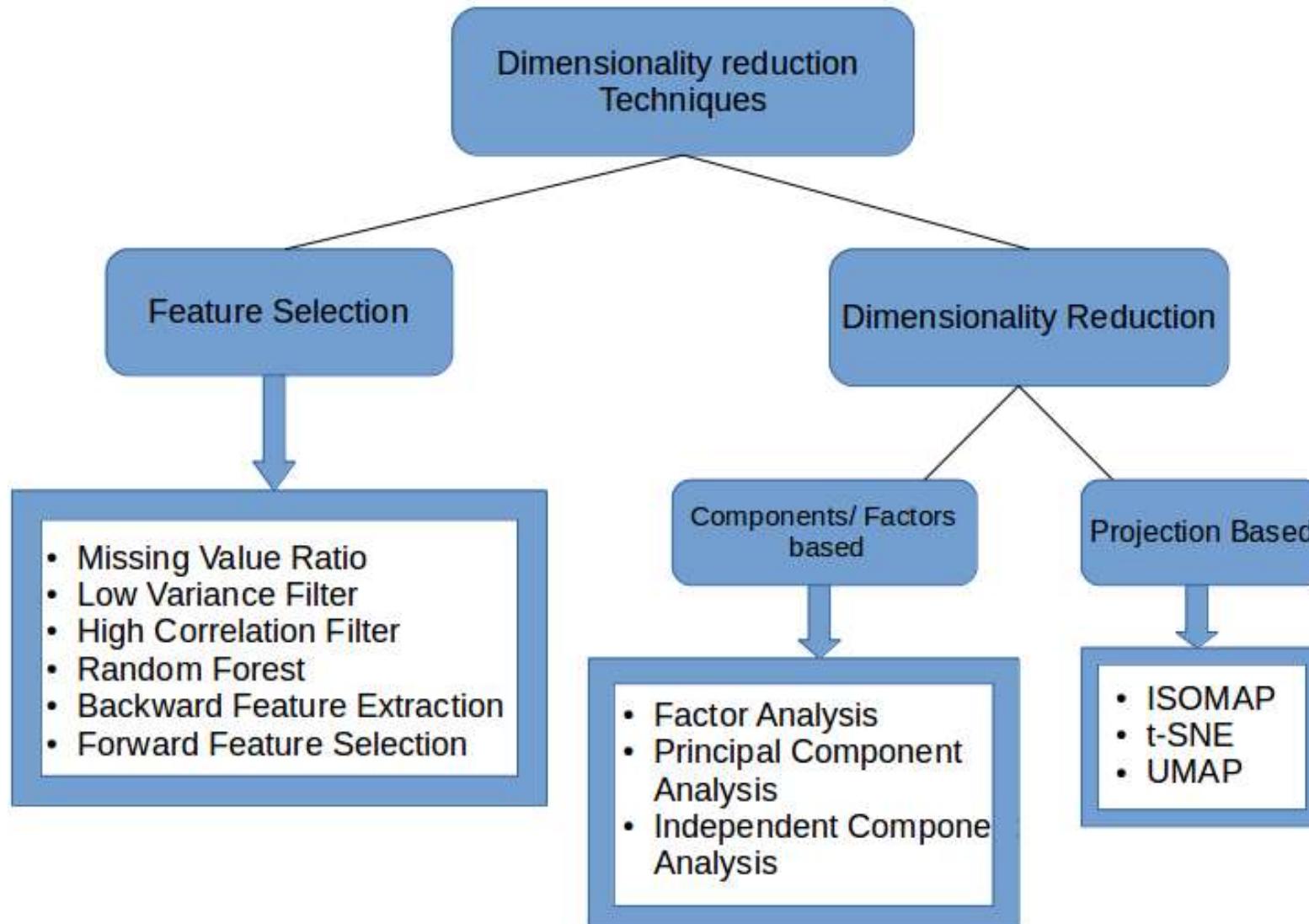
6.1.1. Introduction: Dimensionality reduction

Examples of dimension reduction:



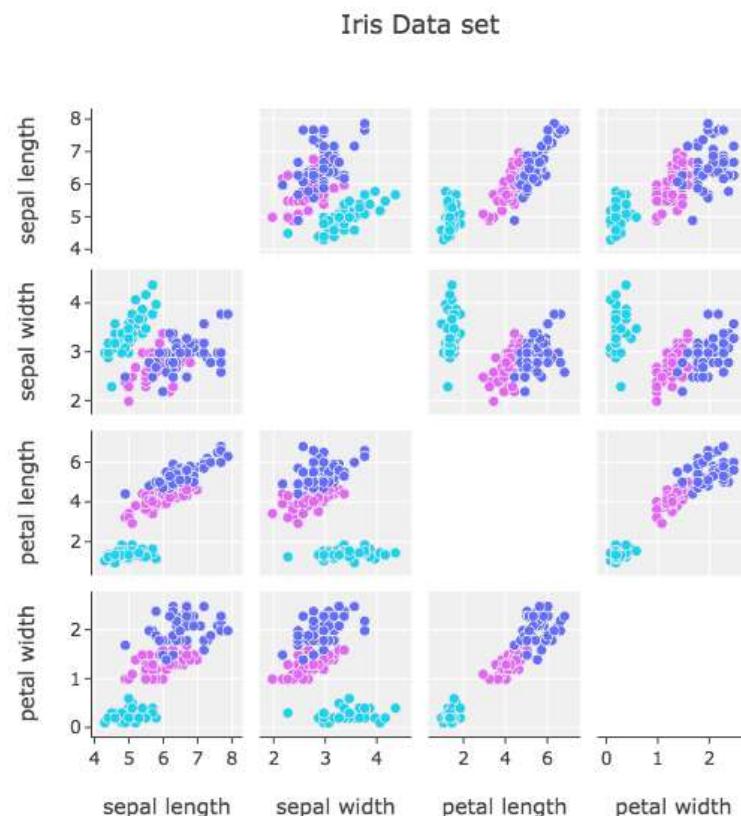
DOI: [10.1007/s10840-017-0281-3](https://doi.org/10.1007/s10840-017-0281-3)

6.1.1. Introduction: Dimensionality reduction



6.1.1. Remember: Scatterplot matrix limitation

We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots *to determine the correlation (if any) between a series of variables.*



!! When we have >3 or 4 quantitative variables – scatterplot matrices quickly become unwieldy

6.1.1. Correlation coefficients

We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots to determine *the correlation (if any) between a series of variables.*

!! When we have >3 or 4 quantitative variables – scatterplot matrices quickly become unwieldy

In this case, it is **more useful to quantify the amount of association between pairs of variables and visualize these quantities rather than the raw data.**

One common way to do this is to calculate **correlation coefficients.**

6.1.1. Correlation coefficient

- Having two sets of observations: x_i and y_i
- And: \bar{x} and \bar{y} the corresponding sample means

The correlation coefficient is:

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

The **correlation coefficient R** is a number between -1 and 1 that measures to what extent two variables are correlated

6.1.1. Correlation coefficients

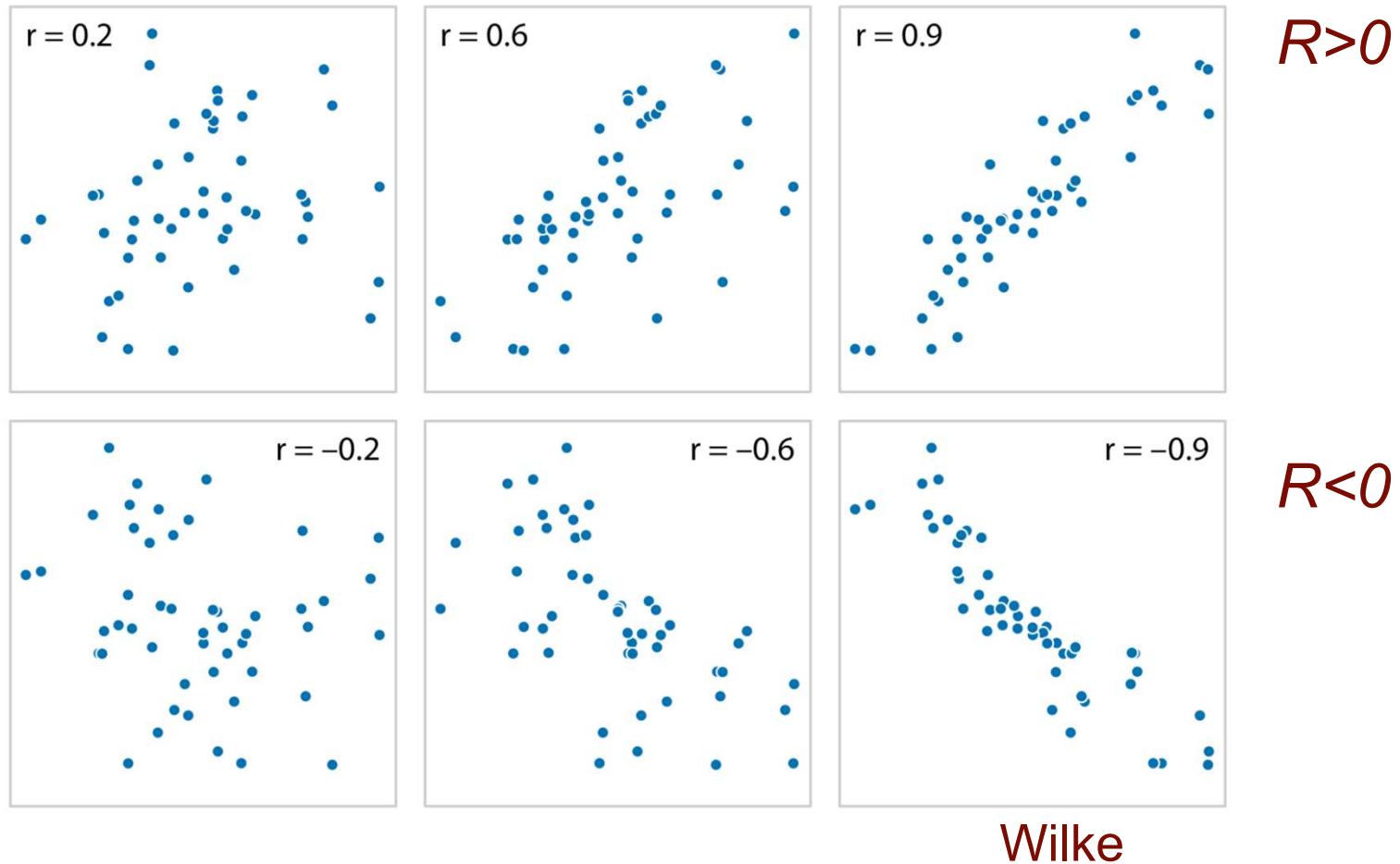
$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}, \quad -1 < R < 1$$

- $R = 0$ means there is **no association** whatsoever
- $R = 1$ or -1 indicates a **perfect association**

The sign of the correlation coefficient R indicates :

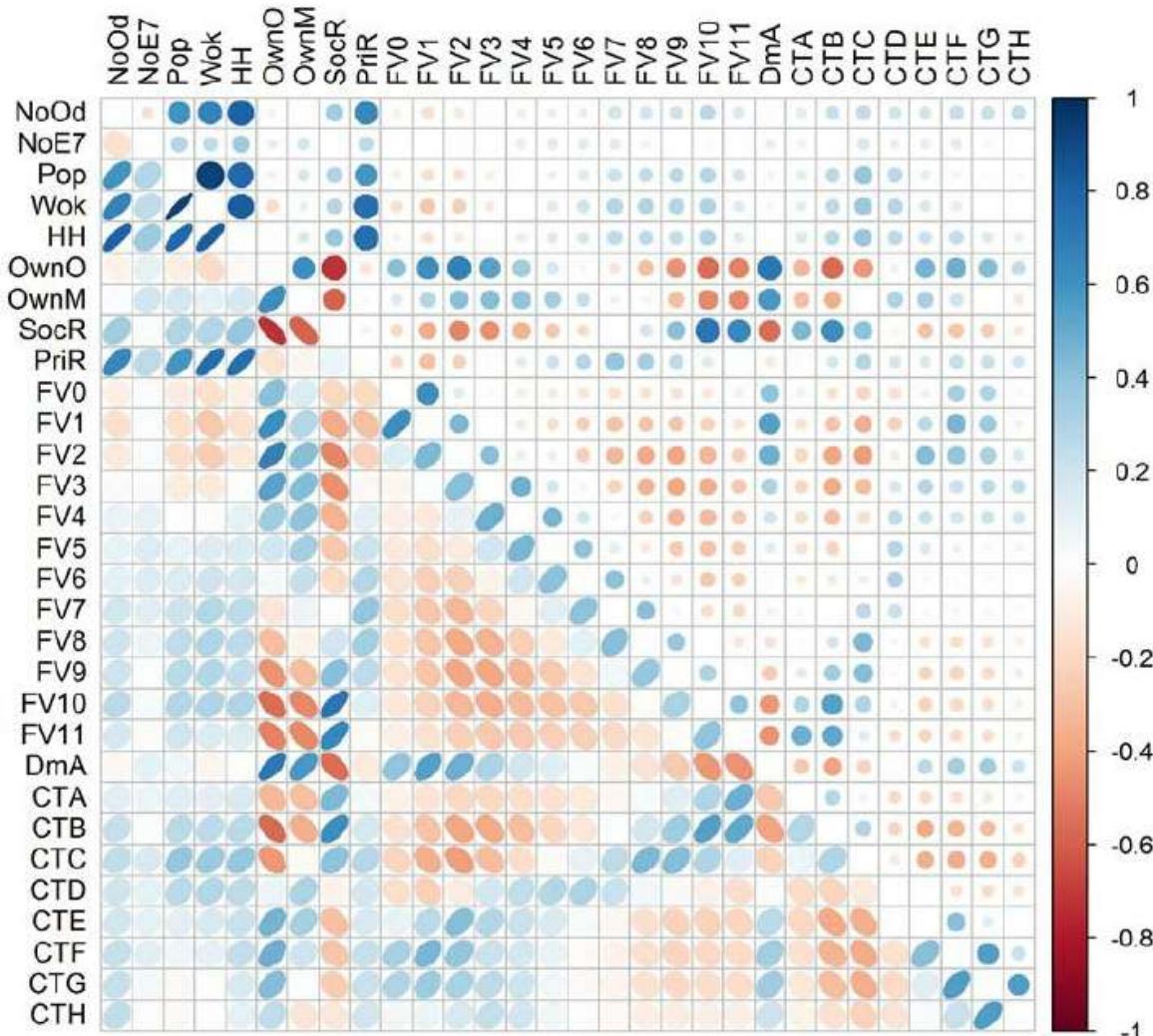
- $R > 0$: variables are **correlated** (larger values in one variable coincide with larger values in the other)
- $R < 0$: **anticorrelated** (larger values in one variable coincide with smaller values in the other)

6.1.1. Correlation coefficients



*Examples of correlations of different magnitude and direction,
with associated correlation coefficient R*

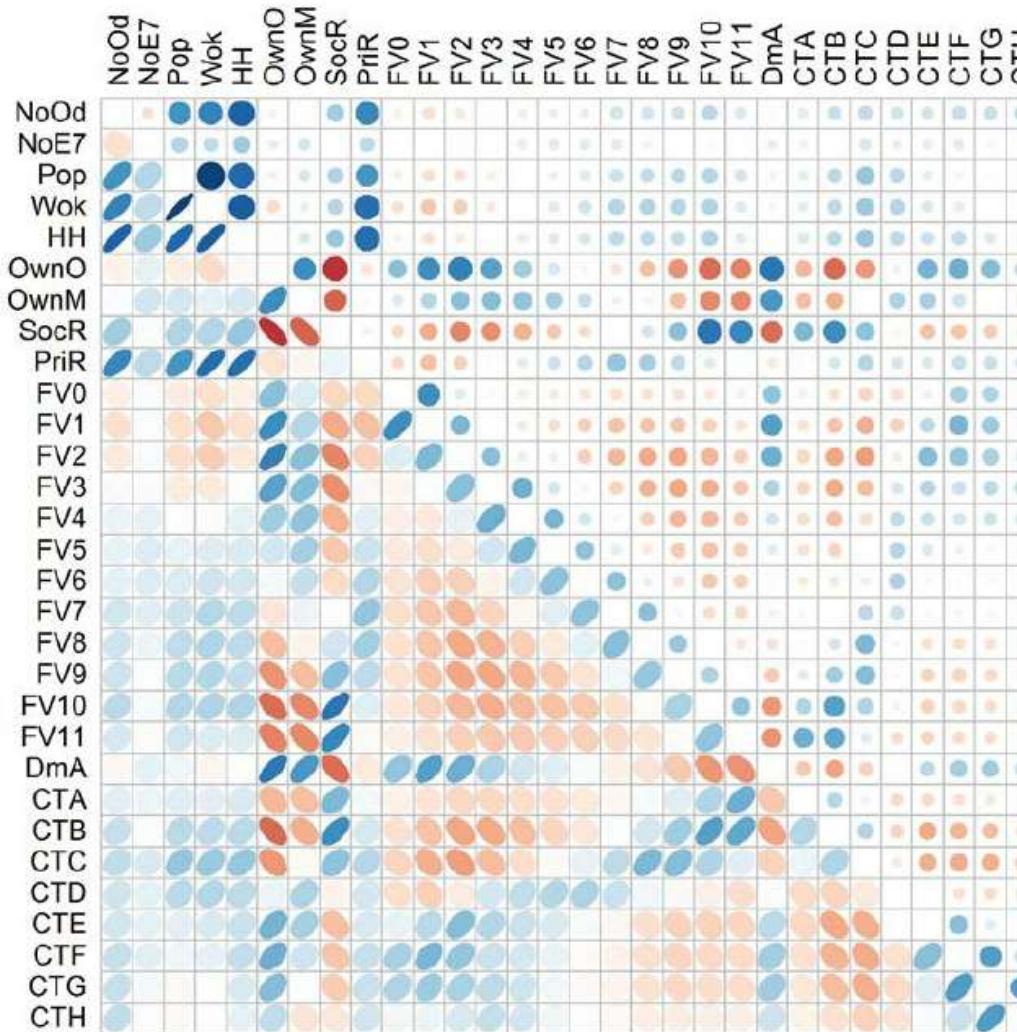
6.1.2. Correlogram



Correlation structure
of the input variables.
**Colour intensity and
ellipse shape** are
directly linked to
correlation coefficient

DOI: [10.1016/j.proeng.2015.08.1069](https://doi.org/10.1016/j.proeng.2015.08.1069)

6.1.2. Correlogram

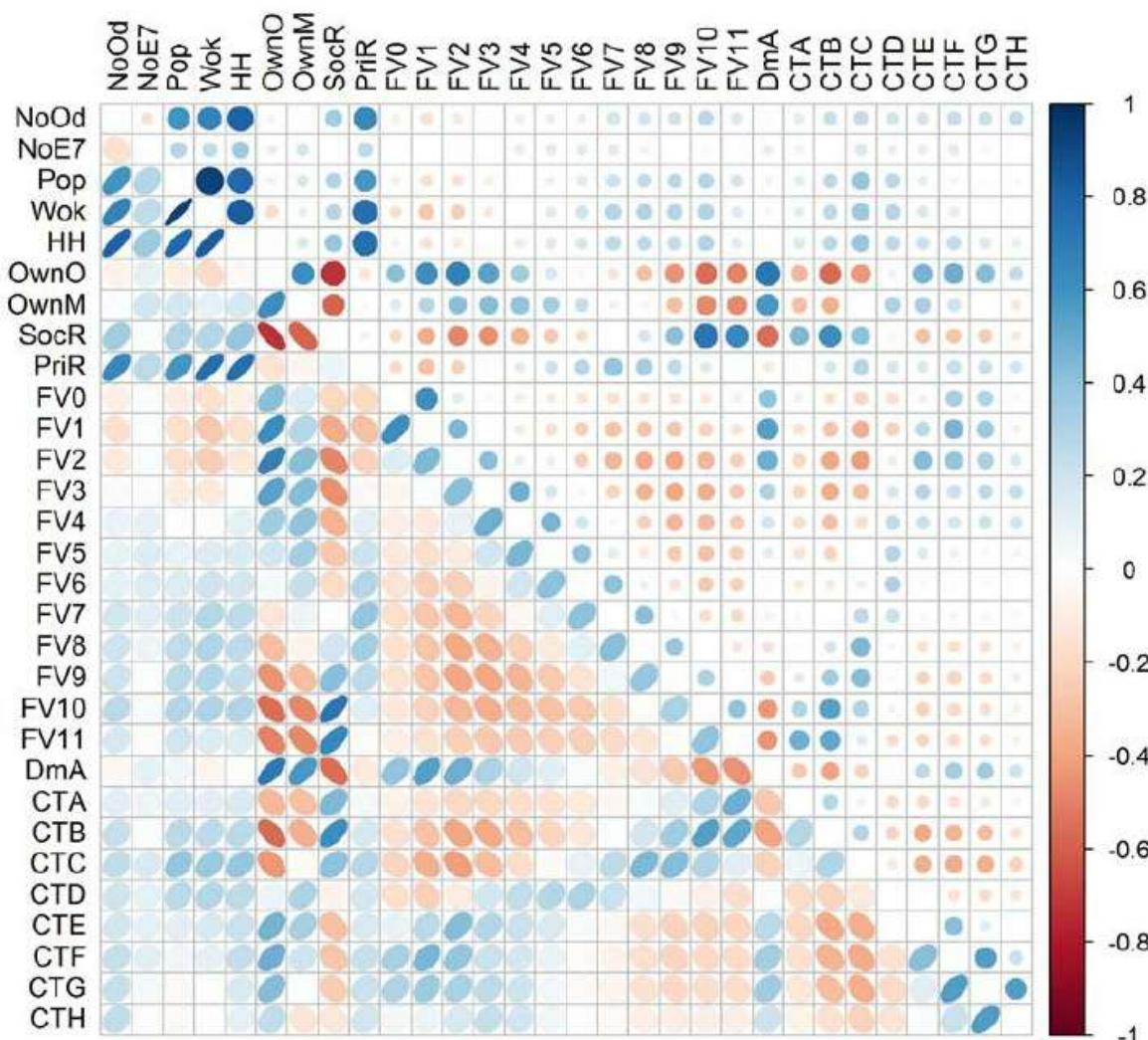


Correlation structure
of the input variables.
Colour intensity and
ellipse shape are
directly linked to
correlation coefficient

$R > 0$ is displayed in
blue (correlated)

$R < 0$ is shown in red
(anticorrelated).

6.1.2 Correlogram



- : there is a **higher negative correlation** coefficient between two variables.
- : indicates a **weak correlation** for two factors
- : there is a **higher positive correlation** coefficient between two variables

6.1. Correlation & Dimensionality reduction

DR relies on the key insight that **most high-dimensional datasets consist of multiple correlated variables** that convey overlapping information

Such datasets can be reduced to a smaller number of key dimensions without loss of much critical information.

DR can be achieved by:

- **Feature elimination** – we reduce the feature space by elimination feature
- **Feature selection** – process of selecting required features from all the features available in data. *Goal: to choose features that represent the dataset perfectly*
- **Feature Engineering** – process of **transforming raw data into feature**, which represent the dataset well

6.1. Correlation & Dimensionality reduction

DR relies on the key insight that **most high-dimensional datasets consist of multiple correlated variables that convey overlapping information**

Such datasets can be reduced to a smaller number of key dimensions without loss of much critical information.

There are many techniques for dimension reduction. We will see:

- Principal Components Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- T-Distributed Stochastic Neighbour Embedding (t-SNE)
(next day)

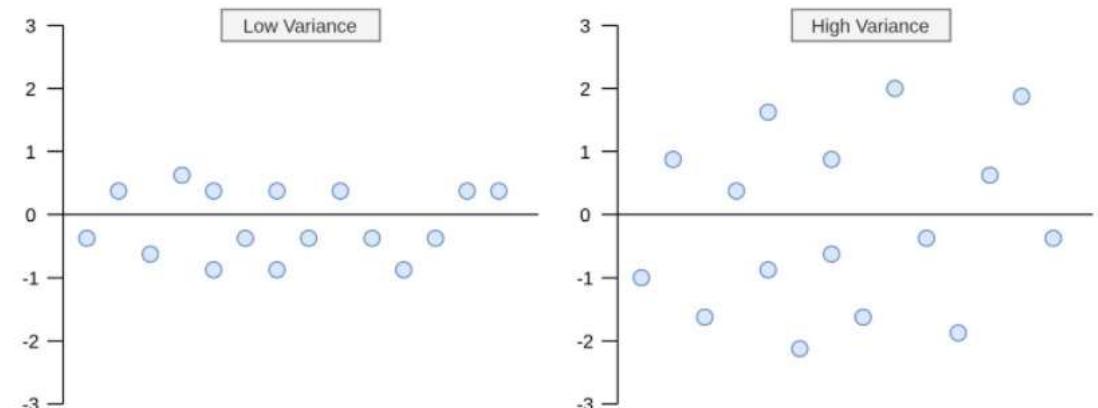
6.1. Remember: Variance

Before going further, let's clarify some concepts:

- We know that variance represents the variation of values in a single variable. *It depends on how the values far from each other.*

Having a set of observations: x_i and \bar{x} the corresponding sample mean:

$$var(x) = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}$$



Sergen Cansiz

6.1. Remember: Covariance

Before going further, let's clarify some concepts:

- Unlike the variance, **covariance** is calculated **between two different variables**. Its purpose is to find the value that indicates **how these two variables vary together**.
- Having two sets of observations: x_i and y_i
- And: \bar{x} and \bar{y} the corresponding sample means

$$cov(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

6.1. Remember: Covariance vs correlation

Before going further, let's clarify some concepts:

- **Covariance vs correlation**
 - Having two sets of observations: x_i and y_i
 - And: \bar{x} and \bar{y} the corresponding sample means

$$cov(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$$var(x) = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}$$

$$R(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

6.1. Covariance vs correlation

Before going further, let's clarify some concepts:

- **Covariance vs correlation**

Covariance	Correlation
Covariance is a measure to indicate the extent to which two random variables change in tandem	Correlation is a measure used to represent how strongly two random variables are related to each other
Covariance indicates the direction of the linear relationship between variables	Correlation measures both the strength and direction of the linear relationship between variables
Covariance can vary between $-\infty$ and $+\infty$	Correlation ranges between -1 and 1

6.1. Covariance vs correlation

Before going further, let's clarify some concepts:

- **Covariance vs correlation**

Covariance	Correlation
<p>Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed</p>	<p>Correlation is NOT influenced by the change in scale</p>
<p>Covariance of two <i>dependent variables</i> measures how much in real quantity (i.e., e.g., cm, km, liters) on average they covary.</p>	<p>Correlation of two <i>dependent variables</i> measures the proportion of how much on average these variables vary with respect to one another</p>
<p>Covariance is zero for independent variables</p>	<p>Completely independent variables have a zero correlation</p>

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**

Because covariance can only be calculated between two variables, **covariance matrices** stand for representing **covariance values of each pair of variables in multivariate data**. Also, the covariance between the same variables equals variance, so, **the diagonal shows the variance of each variable**

Symmetric matrix

$$\begin{matrix} & \begin{matrix} x & y \end{matrix} \\ \begin{matrix} x \\ y \end{matrix} & \begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix} \end{matrix} \quad \begin{matrix} & \begin{matrix} x & y & z \end{matrix} \\ \begin{matrix} x \\ y \\ z \end{matrix} & \begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix} \end{matrix}$$

2 and 3- dimensional covariance matrices

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- **Covariance matrix**

These *values* in the covariance matrix **show the distribution magnitude and direction** of multivariate data in multidimensional space.

By controlling these values we can have *information about how data spread among two dimensions*.

Symmetric matrix

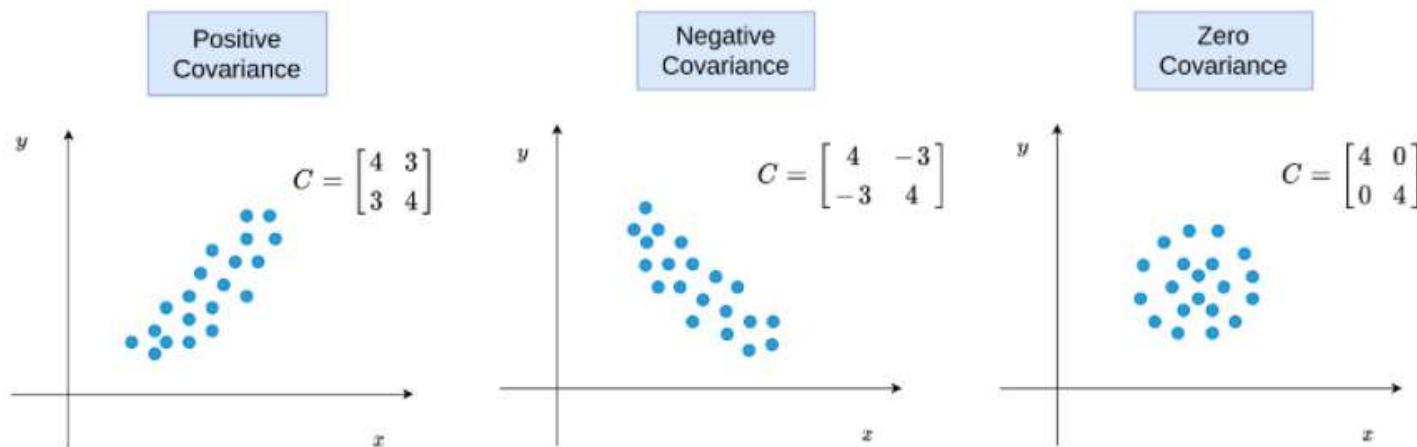
$$\begin{matrix} & \begin{matrix} x & y \end{matrix} \\ \begin{matrix} x \\ y \end{matrix} & \begin{bmatrix} var(x) & cov(x, y) \\ cov(x, y) & var(y) \end{bmatrix} \end{matrix} \quad \begin{matrix} x & y & z \\ \begin{matrix} x \\ y \\ z \end{matrix} & \begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix} \end{matrix}$$

2 and 3- dimensional covariance matrices

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- Covariance matrix

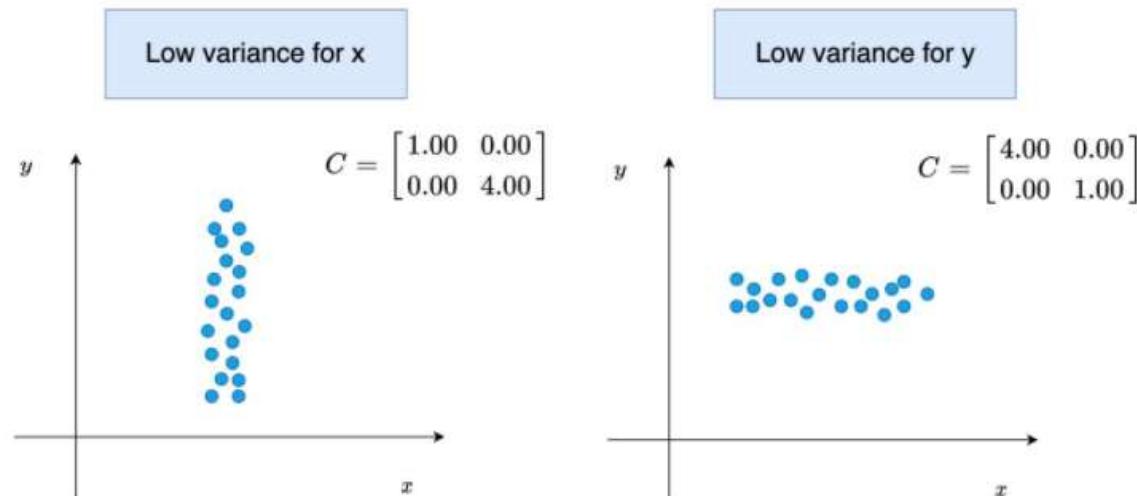


Sergen Cansiz

6.1. Covariance matrix

Before going further, let's clarify some concepts:

- Covariance matrix



Covariance
near zero
and different
variances

Sergen Cansiz

6.1.3. Principal Components Analysis (PCA)

- When should I use PCA?
1. Do you want to reduce the number of variables, but ***you are not able to identify variables*** to completely remove from consideration?
 2. Do you want ***to ensure your variables are independent*** of one another?
 3. Are you ***comfortable making your independent variable less interpretable***?

6.1.3. Principal Components Analysis (PCA)

- PCA introduces a **new set of variables (smaller number of variables)**, called **principal components (PCs)**, by linear combination of the original variables in the data, standardized to zero mean and unit variance.
- **The axes or new variables are the PCs and are ordered by variance:**
 - The first component, *PC 1*, represents the *direction of the highest variance of the data*.
 - The direction of the *PC 2*, represents the *highest of the remaining variance orthogonal to the PC 1*.

This can be naturally **extended to obtain the required number of components, which together span a component space covering the desired amount of variance**.

6.1.3. Covariance matrix – relation with PCA

Before going further, let's clarify some concepts:

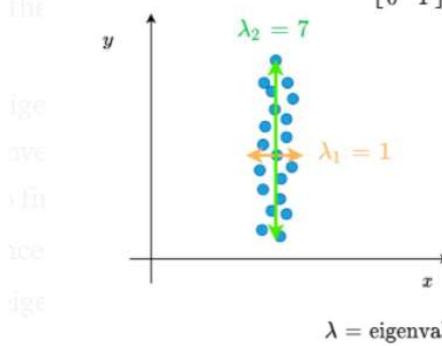
- **Eigenvalues and eigenvectors of covariance matrix:**
 - The **eigenvalues represent the magnitude of the spread** in the direction of the principal components in PCA.
 - The **eigenvectors show the direction**.

6.1.3 Covariance matrix – relation with PCA

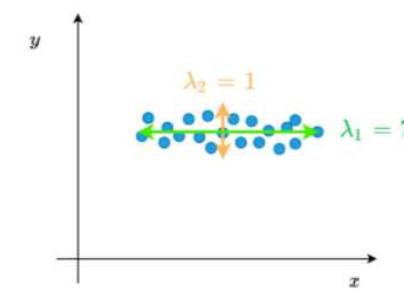
Before going further, let's clarify some concepts:

- Eigenvalues and eigenvectors of covariance matrix

1 $C = \begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix}$ $\lambda_{1,2} = [1 \ 7]$
 $V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



2 $C = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix}$ $\lambda_{1,2} = [7 \ 1]$
 $V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



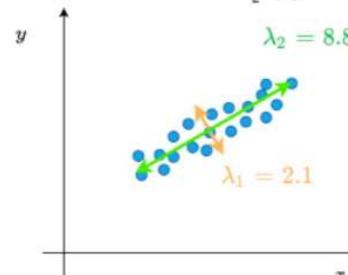
- The first and second plots show the distribution of points when the covariance is near zero (independent variables).

Note: when the covariance is zero the eigenvalues=variance values

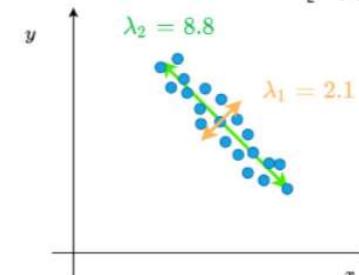
- The third and fourth plots represent the distribution of points when the covariance is different from zero.

Note: here we need to calculate the eigenvalues and eigenvectors

3 $C = \begin{bmatrix} 4 & 3 \\ 3 & 7 \end{bmatrix}$ $\lambda_{1,2} = [2.1 \ 8.8]$
 $V = \begin{bmatrix} -0.8 & -0.5 \\ 0.5 & -0.8 \end{bmatrix}$



4 $C = \begin{bmatrix} 4 & -3 \\ -3 & 7 \end{bmatrix}$ $\lambda_{1,2} = [2.1 \ 8.8]$
 $V = \begin{bmatrix} -0.8 & 0.5 \\ -0.5 & -0.8 \end{bmatrix}$



Sergen Cansiz

6.1.3 Principal Components Analysis (PCA)

PCA is maybe the most popular technique to examine high-dimensional data (unsupervised learning)

PCA computes a rotation matrix : $W \in \mathbb{R}^{P \times P}$ **from**
the matrix of features $X \in \mathbb{R}^{N \times P}$

W can be understood as a mapping function that transforms the observations in X to a rotated space

The coordinates of observations in X are transformed to their new form, Z , via: $Z = XW$

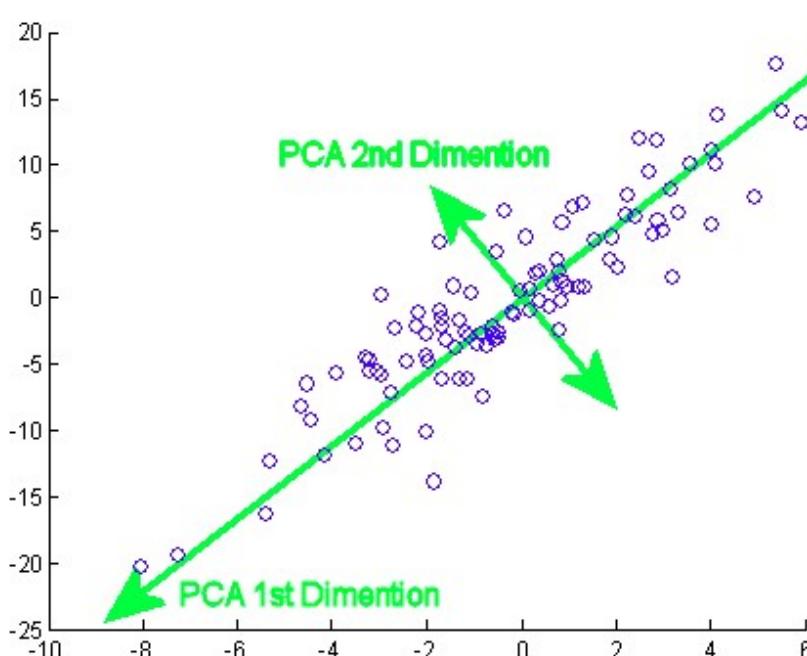
The rotation matrix, W , is constructed through orthogonal linear transformations. Each of these transformations is performed in order to maximize the variance on the data

6.1.3 Principal Components Analysis (PCA)

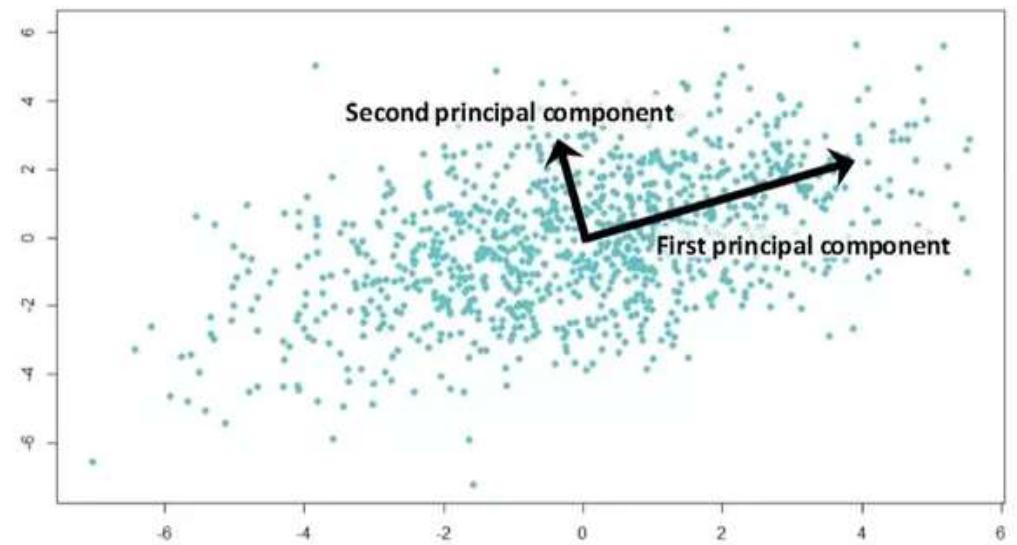
Steps:

1. Take the **matrix of features** $X \in \mathbb{R}^{N \times P}$, $N > P$
2. Compute the **mean vector for each dimension**
3. Compute the **covariance matrix**
4. Compute the eigenvectors and corresponding eigenvalues for each dimension
5. Sort the eigenvectors by decreasing eigenvalues and choose P eigenvectors with the largest eigenvalues to form a new matrix $W \in \mathbb{R}^{P \times P}$
6. Use this eigenvector matrix to transform the samples onto the new subspace: $Z = XW$

6.1.3 Principal Components Analysis (PCA)



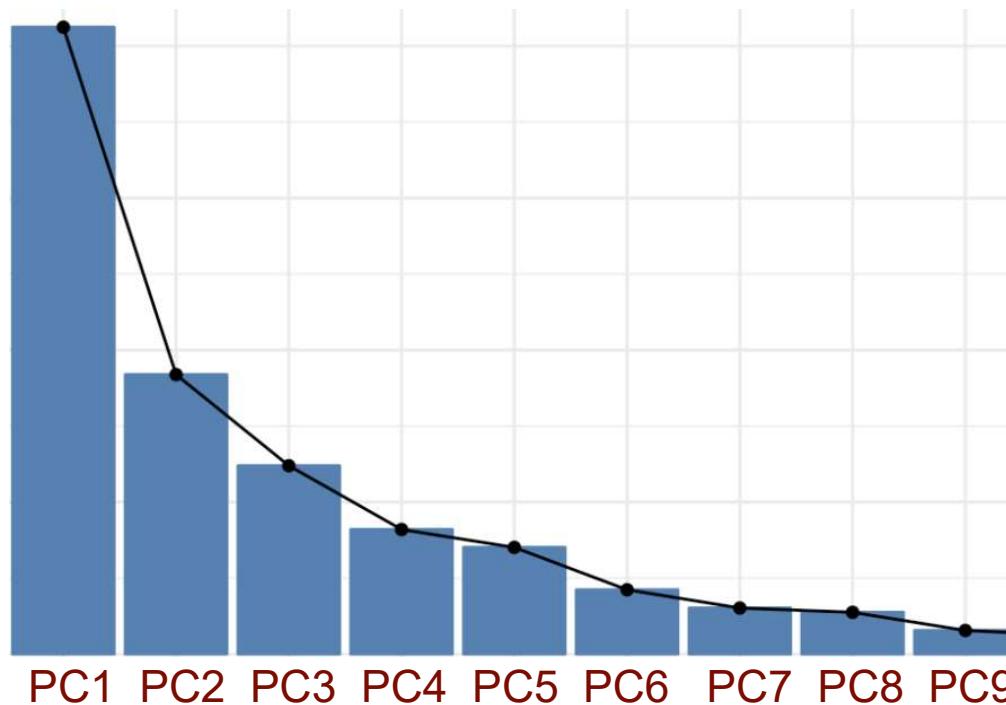
weigend.com



Wavy AI Research
Foundation

6.1.3 Principal Components Analysis (PCA)

- In highly dimensional datasets, **the vast majority of the variance in the data is often captured by a small number of principal components.**
- A plot of the distribution of the variance across principal components may look like this:

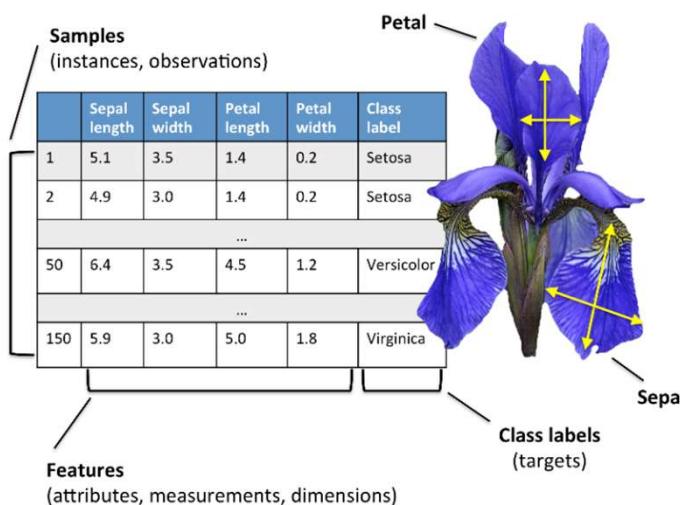


6.1.3 Principal Components Analysis (PCA)

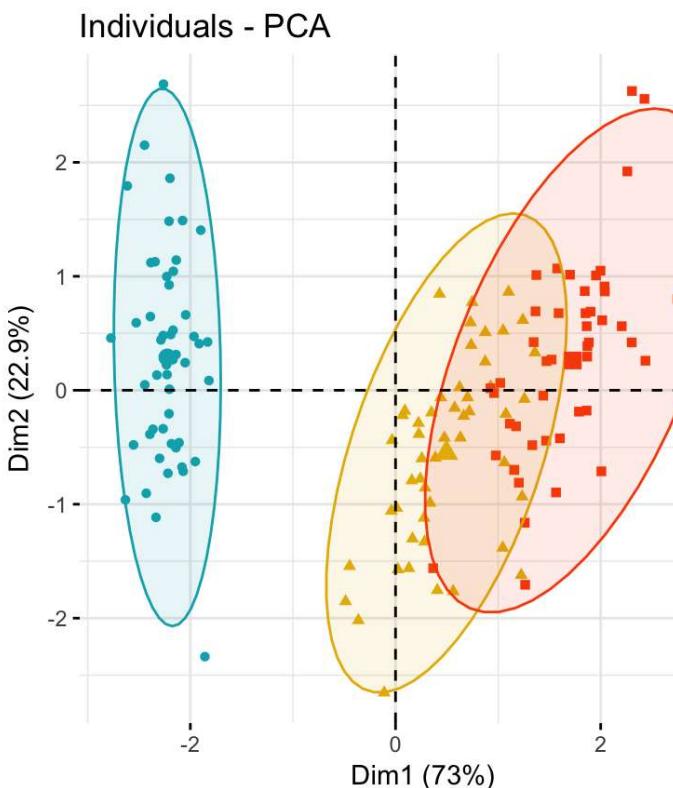


```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa
```

3 kind of Iris flowers with 4 attributes:
sepal length, sepal width, petal length and
petal width



PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the most variance in the data.



Here we plot the different samples on the 2 first PCs.

6.1.3 Principal Components Analysis (PCA)

In the theoretical class, we saw:



```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

How to do PCA Visualization using R:

The following functions, from factoextra package can be used:

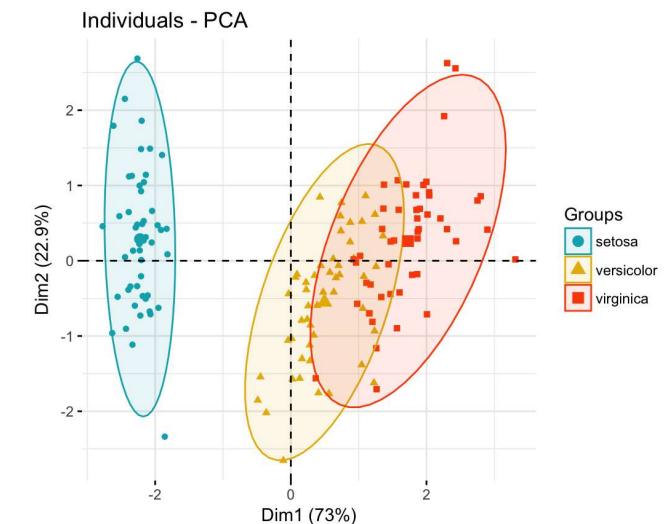
`fviz_pca_ind()`: Graph of individuals

`fviz_pca_var()`: Graph of variables

`fviz_pca_biplot()` (or `fviz_pca()`): Biplot of individuals and variables

PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the most variance in the data.

In practice:



6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:



3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
```

First, we need to install the packages and load the libraries:

```
> install.packages("devtools")
> library("devtools")
> install.packages("factoextra")
> library("factoextra")
```

Afterwards, we prepare the dataframe if needed:

The variable Species (index = 5) is removed (not numerical)

We use 'center=TRUE' to center the variables to 0 and we scale them to have variance 1 by using 'scale.=TRUE'

```
> iris_pca<-prcomp(iris[,-5], center=TRUE, scale.=TRUE)
```

6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:



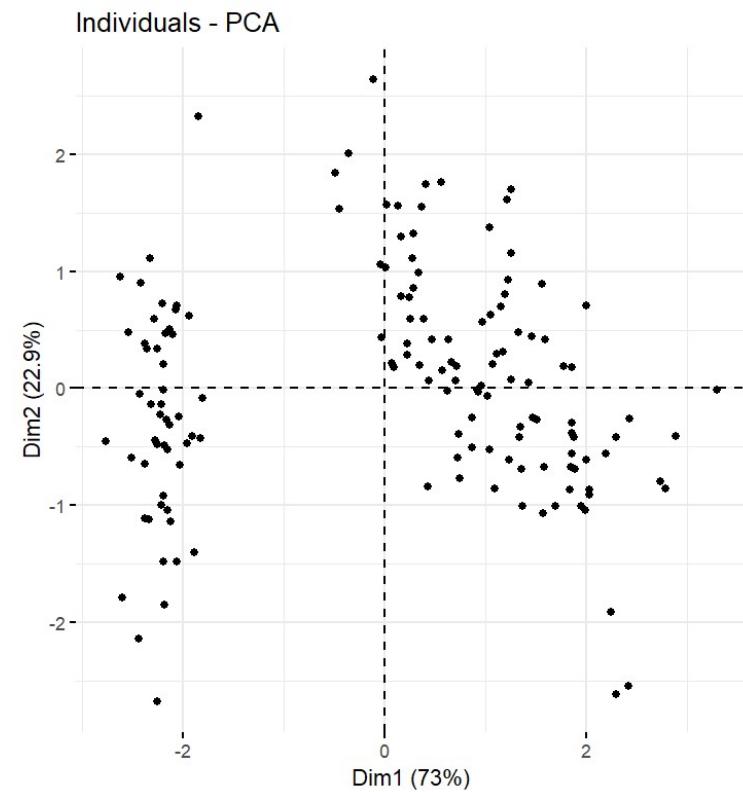
3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
```

How to do PCA Visualization using R:

```
fviz_pca_ind(iris_pca, geom="point")
```

Graph of individuals using only points



6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:



3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
```

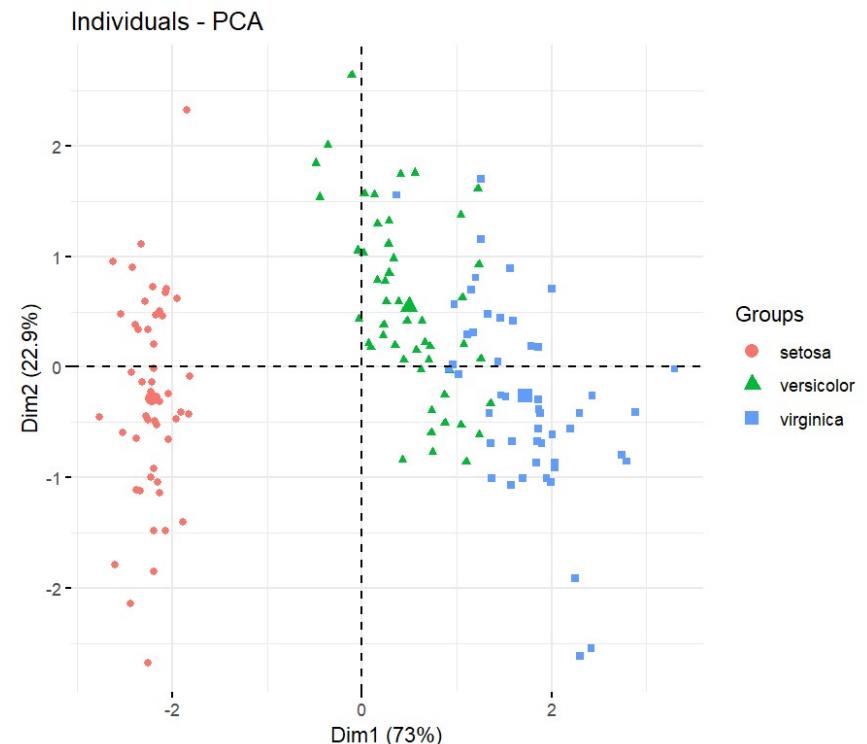
How to do PCA Visualization using R:

```
fviz_pca_ind(iris_pca, geom="point")
```

Graph of individuals using only points

```
fviz_pca_ind(iris_pca, label="none",
habillage=iris$Species)
```

To Color individuals by groups



6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:

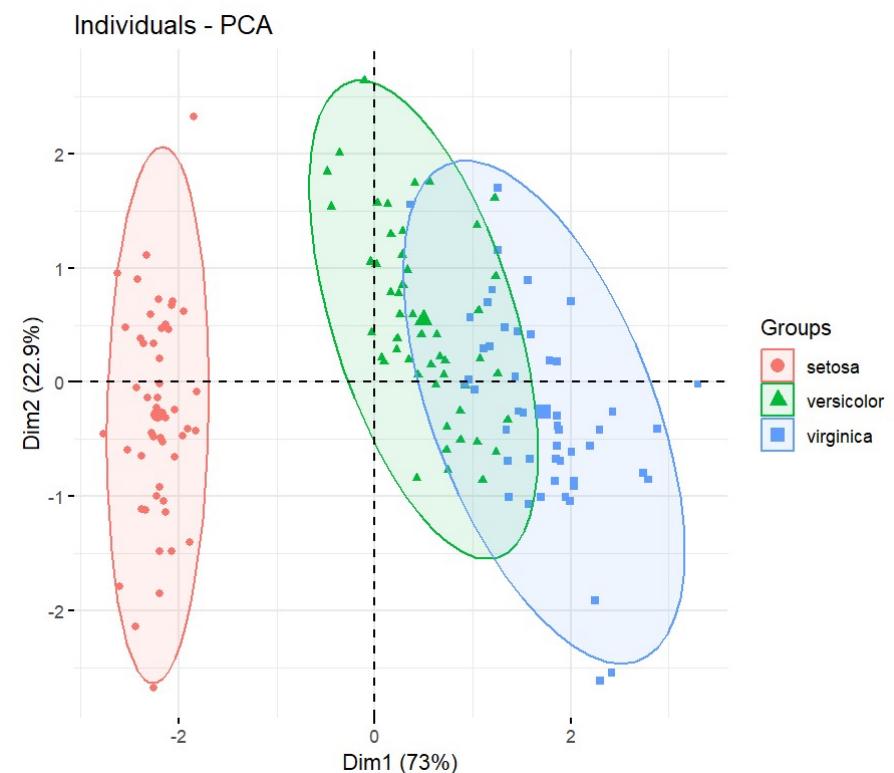


3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
```

How to do PCA Visualization using R:

```
fviz_pca_ind(iris_pca, geom="point")
# Graph of individuals using only points
fviz_pca_ind(iris_pca, label="none",
habillage=iris$Species)
# To Color individuals by groups
fviz_pca_ind(iris_pca, label="none",
habillage=iris$Species,
addEllipses=TRUE, ellipse.level=0.95)
# To add ellipses
```



6.1.3. Dimension reduction – PCA

In the theoretical class, we saw:



3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1      5.1       3.5      1.4       0.2   setosa  
## 2      4.9       3.0      1.4       0.2   setosa  
## 3      4.7       3.2      1.3       0.2   setosa
```

How to do PCA Visualization using R:

```
summary(iris_pca) # Give us the importance of the components
```

```
> summary(iris_pca)
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4
Standard deviation	1.7084	0.9560	0.38309	0.14393
Proportion of Variance	0.7296	0.2285	0.03669	0.00518
Cumulative Proportion	0.7296	0.9581	0.99482	1.00000

More options in:

<http://www.sthda.com/english/wiki/fviz-pca-quick-principal-component-analysis-data-visualization-r-software-and-data-mining>

6.1.3 Principal Components Analysis (PCA)

Summarizing:

- PCA is a very interpretable method.
- **Each PC is well-defined as we know that it is orthogonal to the other dimensions.**
- **We can obtain the variance that is explained by each PC to select an appropriate number of dimensions**

Weakness of PCA:

It tends to be highly affected by outliers in the data

To overcome this issue many robust versions of PCA has been developed: RandomizedPCA, sparsePCA, etc

PCA works best only with continuous data

6.1.3 Principal Components Analysis (PCA)

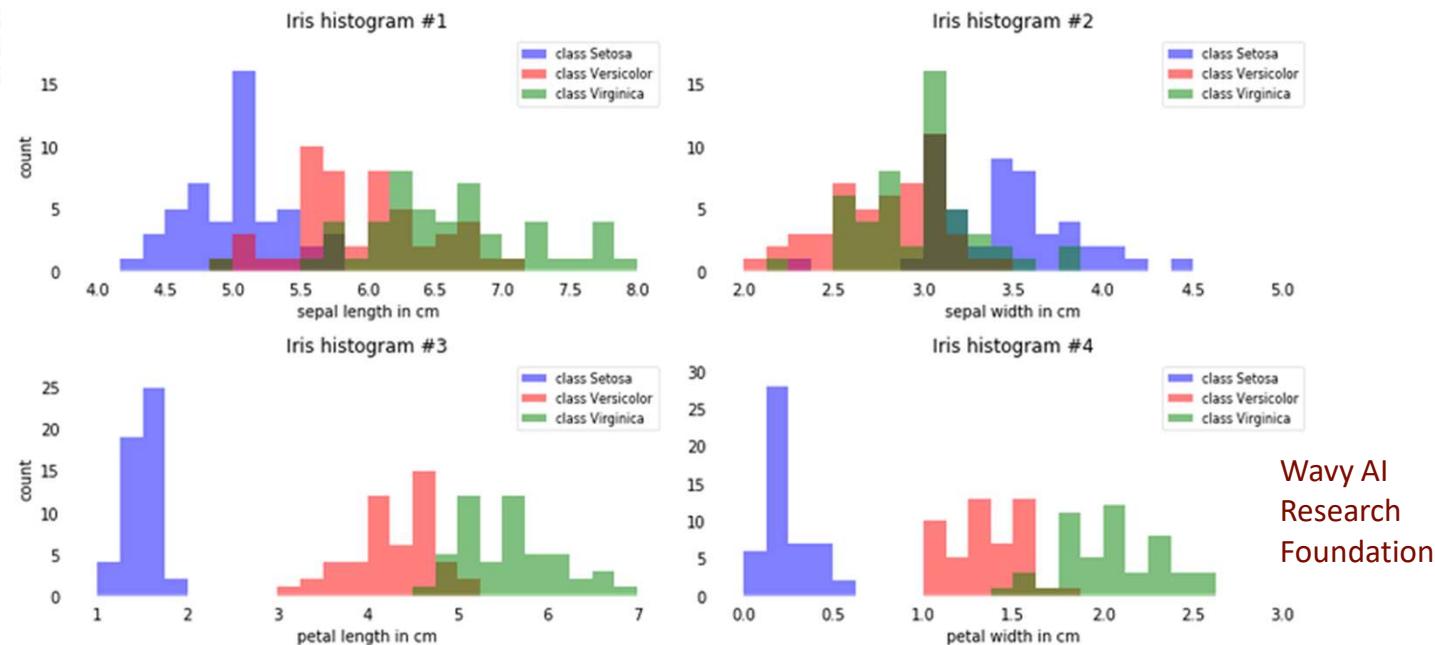
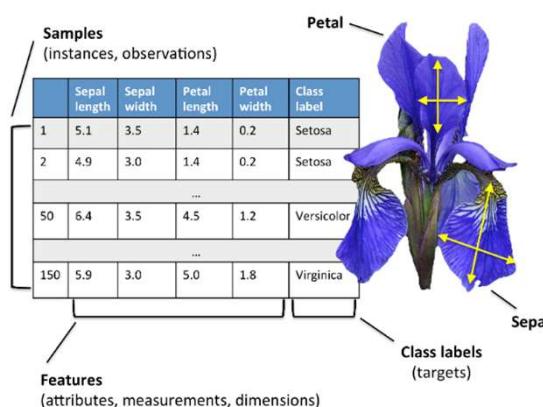


$$\mathbf{X} = \begin{bmatrix} x_{1\text{sepal length}} & x_{1\text{sepal width}} & x_{1\text{petal length}} & x_{1\text{petal width}} \\ \dots & & & \\ x_{2\text{sepal length}} & x_{2\text{sepal width}} & x_{2\text{petal length}} & x_{2\text{petal width}} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \omega_{\text{iris-setosa}} \\ \dots \\ \omega_{\text{iris-virginica}} \end{bmatrix}$$

En numériques

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width



! Remark: For low-dimensional datasets like Iris, those histograms would already be very informative.

6.1.4 Linear Discriminant Analysis (LDA)

LDA seeks to best **separate** (or discriminate) **the samples** in the training dataset **by their class value**.

The fundamental idea of linear combinations goes back as far as the 1960s

- ✓ The idea behind LDA: to find a new feature space to project the data in order to *maximize classes separability*

In 1988, the statistician Ronald Fisher proposed :

- Maximize the function that represents the difference between the means, normalized by a measure of the within-class variability

6.1.4 Linear Discriminant Analysis (LDA)

The Fisher's model seeks to **find a linear combination of input variables** that:

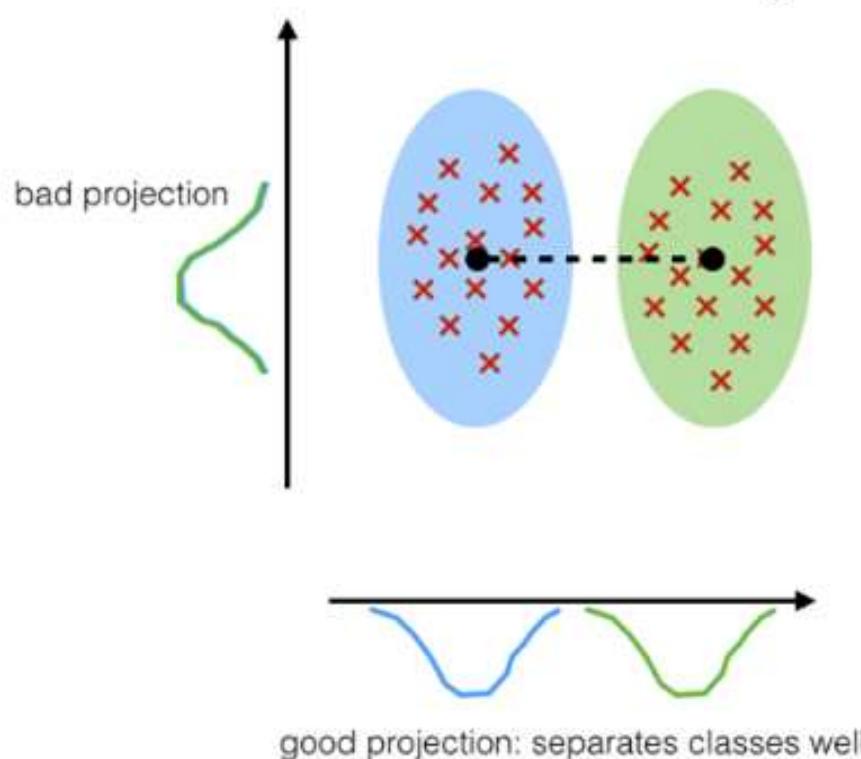
- achieves the **maximum separation** for samples **between classes** (class centroids or means),
- and the **minimum separation** of samples **within each class**.

The LDA takes the mean value for each class and considers variants to make predictions assuming a Gaussian distribution

6.1.4 Linear Discriminant Analysis (LDA)

LDA seeks to best **separate** (or discriminate) the **samples** in the training dataset **by their class value**.

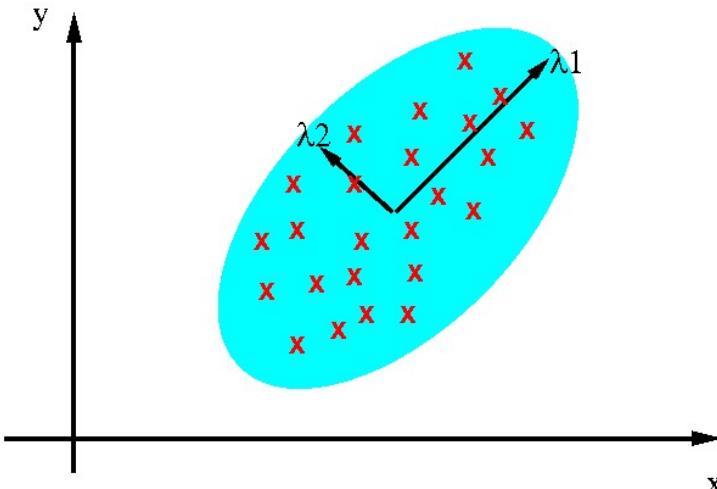
Maximizing the component axes for class-separation:



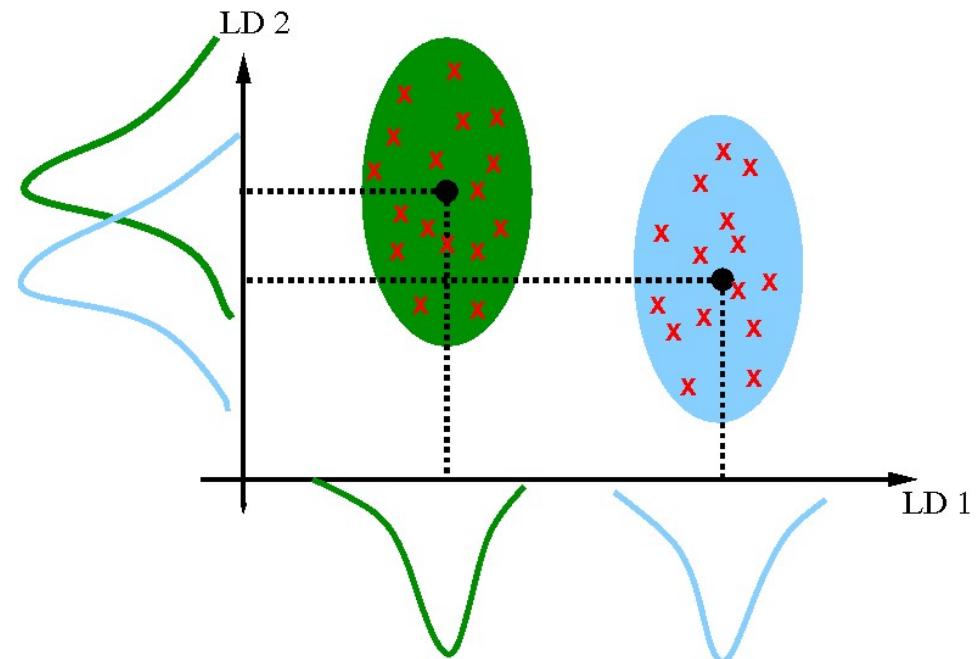
Wavy AI Research
Foundation

6.1.4 PCA versus LDA

PCA: component axes that maximize the variance



LDA: maximizing the component axes for class-separation

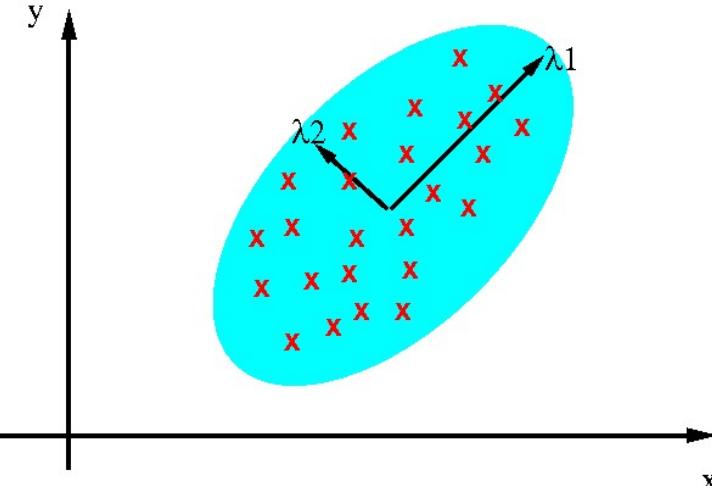


Wavy AI
Research
Foundation

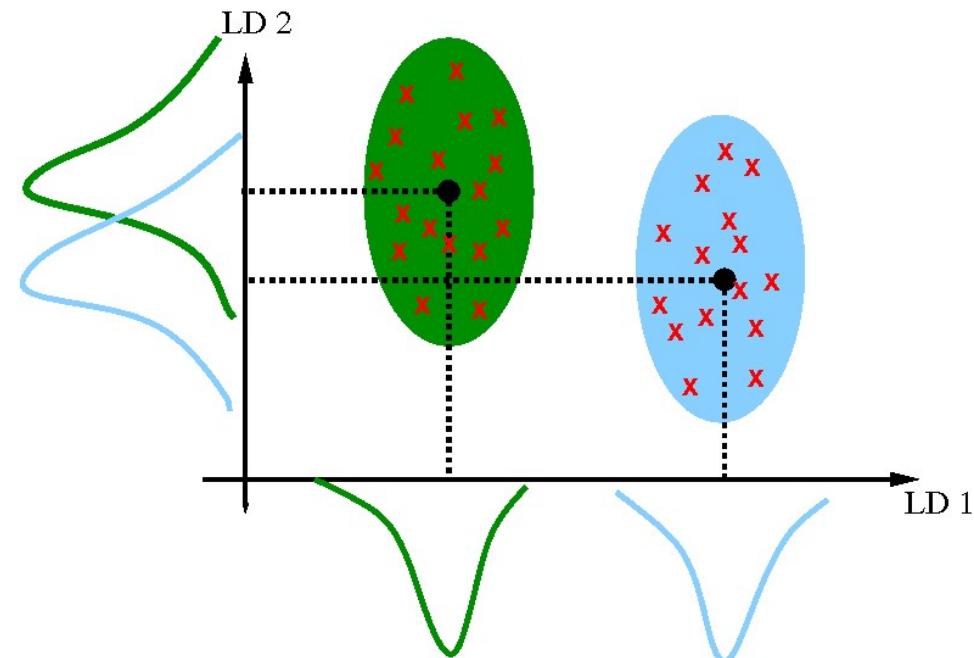
- Both, LDA and PCA are linear transformation techniques that are commonly used for dimensionality reduction (both are techniques for the data Matrix Factorization)

6.1.4 PCA versus LDA

PCA: component axes that maximize the variance



LDA: maximizing the component axes for class-separation



Wavy AI
Research
Foundation

- **PCA** is unsupervised algorithm that **attempts to find the orthogonal component axes of maximum variance in a dataset**
- while the goal of LDA as supervised algorithm **is to find the feature subspace that optimizes class separability.**

6.1.4. PCA versus LDA

Samples

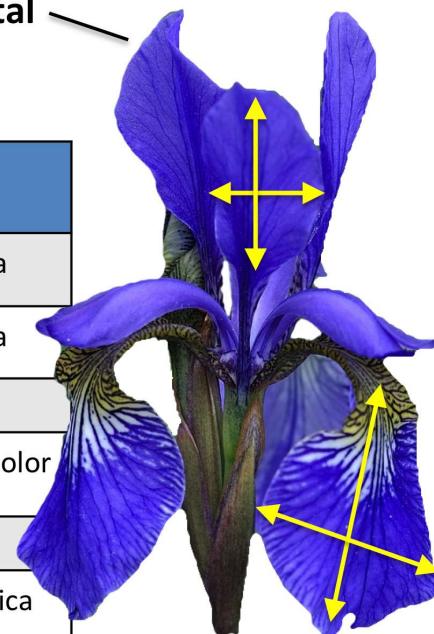
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features

(attributes, measurements, dimensions)

Petal



Sepal

Class labels
(targets)

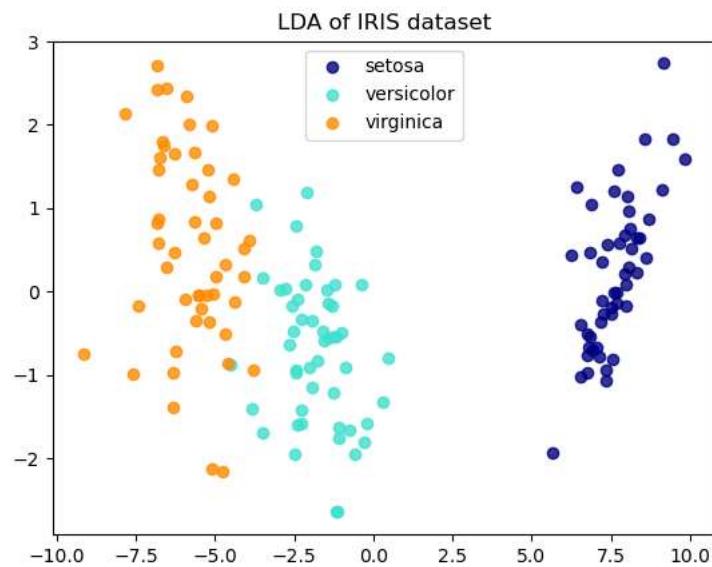


```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1       5.1      3.5       1.4      0.2    setosa
## 2       4.9      3.0       1.4      0.2    setosa
## 3       4.7      3.2       1.3      0.2    setosa
```

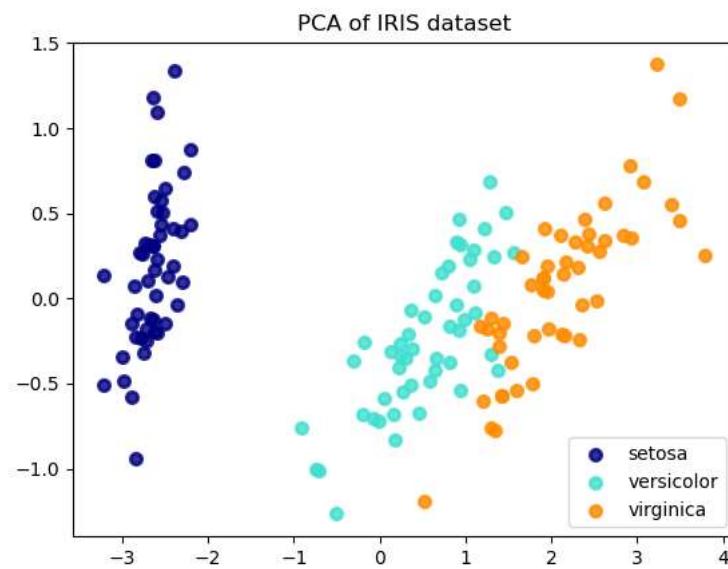
6.1.4. PCA versus LDA



3 kind of Iris flowers with 4 attributes: sepal length, sepal width, petal length and petal width



PCA identifies the combination of attributes (PCs, or directions in the feature space) that account for the most variance in the data.



LDA: tries to identify attributes that account for the most variance between classes

6.1.4. Linear Discriminant Analysis (LDA)

Steps: (see this link for an example with iris dataframe)

1. Compute the **d-dimensional mean vector for the different classes** from the dataset. (in PCA was for each direction)
2. Compute the Scatter matrix (in between class and within the class scatter matrix)
3. Sort the Eigen Vector by decreasing Eigen Value order and choose k eigenvector with the largest eigenvalue to form a dxk dimensional matrix W (where every column represent an eigenvector)
4. Used dxk eigenvector matrix to transform the sample onto the new subspace. This can be summarised by the matrix multiplication:

$$Y = XW$$

where X is a $n \times d$ dimension matrix representing the n samples and you are transformed $n \times k$ dimensional samples in the new subspace.

6.1.4. Linear Discriminant Analysis (LDA)

LDA can be useful in areas like image recognition and predictive analysis in marketing

Weakness of LDA:

- **LDA does not work well if the design is not balanced** (i.e. the number of objects in various classes are (highly) different)
- If the **distribution of your data is significantly non-Gaussian**, the LDA might not perform very well.
- It is **sensitive to overfit**
- **LDA is not applicable (inferior) for non-linear problems**

Thanks for your attention!

Judit Chamorro Servent

Departament de Matemàtiques

judit.chamorro@uab.cat

GRAU EN ENGINYERIA DE DADES

104365

TEMA - 7. Sistemes Avançats I

Departament de Matemàtiques

7. Advanced systems of visualization (I)

7.1 Multiple variables and dimensions

7.2 Networks

7.3 3D Data

7.4 Vector Fields

7.1 Multiple variables and dimensions. Contents

1. Introduction

Multiple variables and dimensions

2. Visualizing many distributions at once

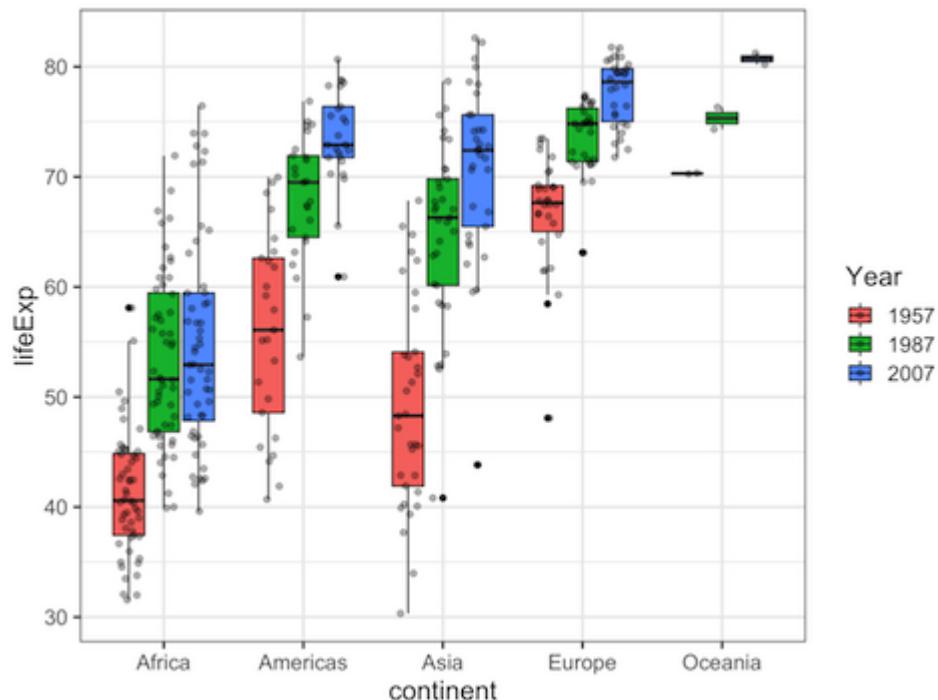
3. Visualizing many proportions at once

4. Visualizing many relations (correlations) at once : Bubble plot and Scatter plot matrices

7.1.1 Multiple variables and dimensions. Introduction

There are large datasets, containing much more information than can be shown in a plot.

- Some datasets can be shown in a single figure panel by grouping variables.



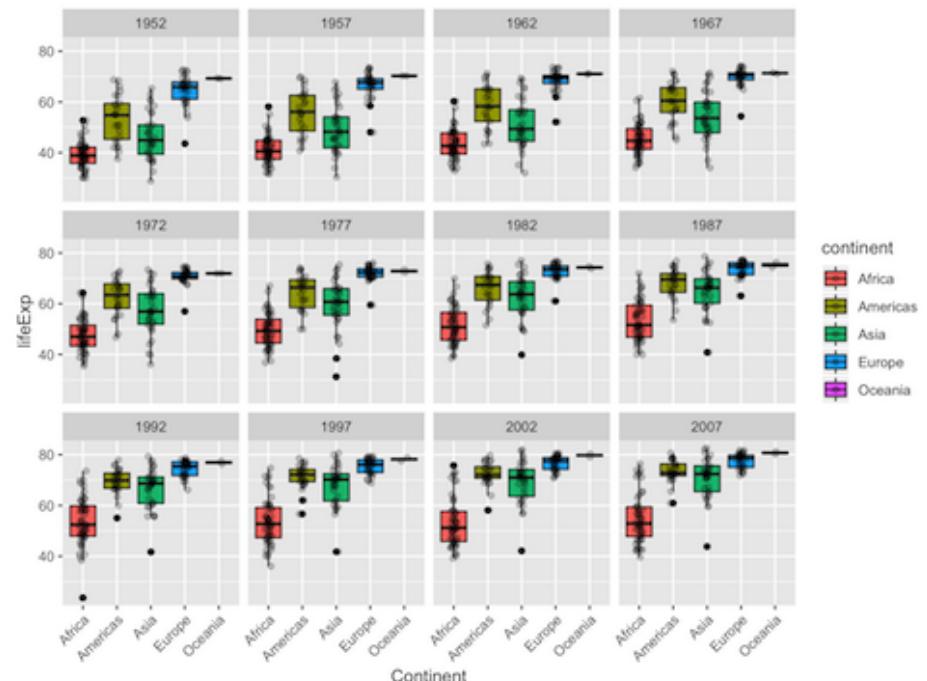
7.1.1 Multiple variables and dimensions. Introduction

There are large datasets, containing much more information than can be shown in a plot.

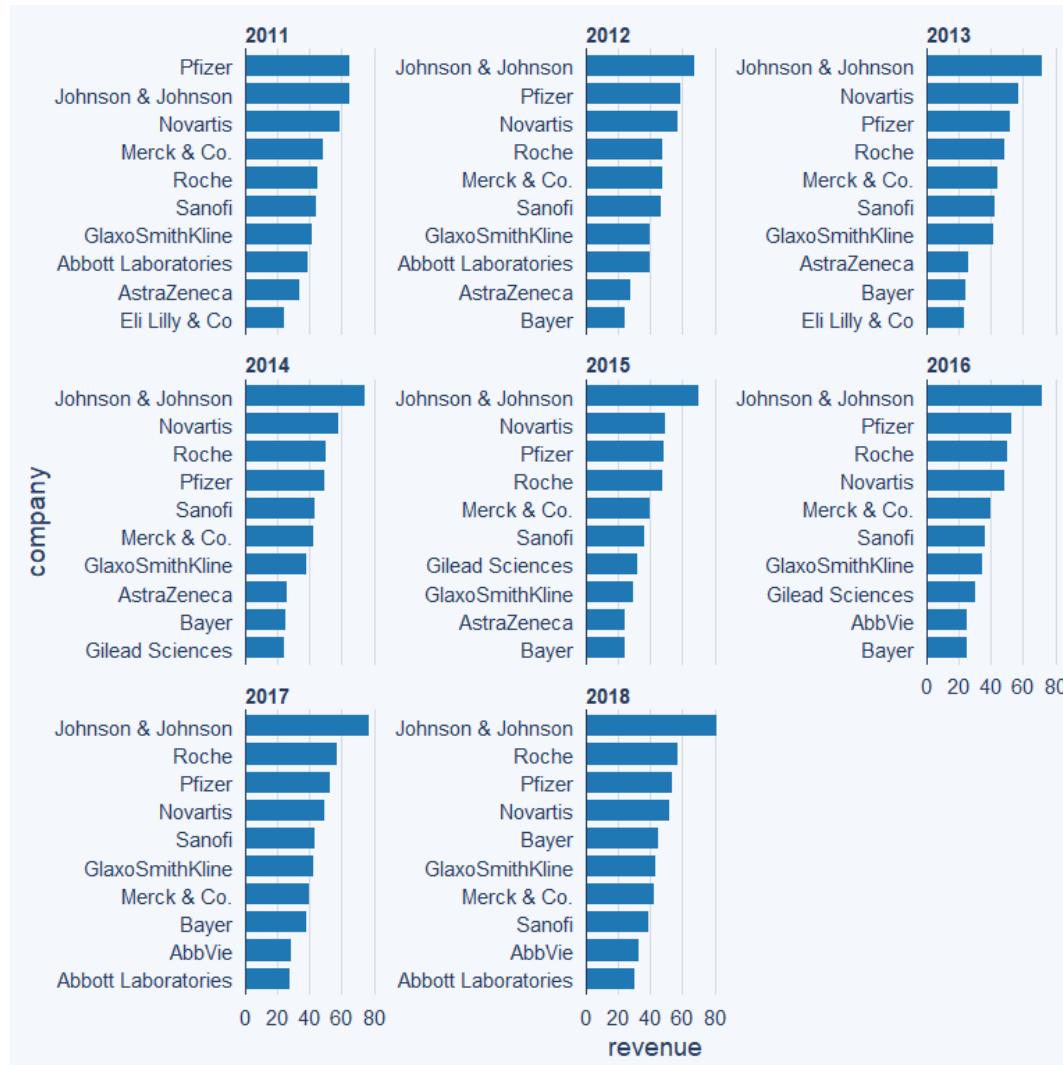
- However, for more complex datasets, it can be helpful to create multi-panel figures.

These are figures that consist of *multiple figure panels where each panel shows some subset of the data.*

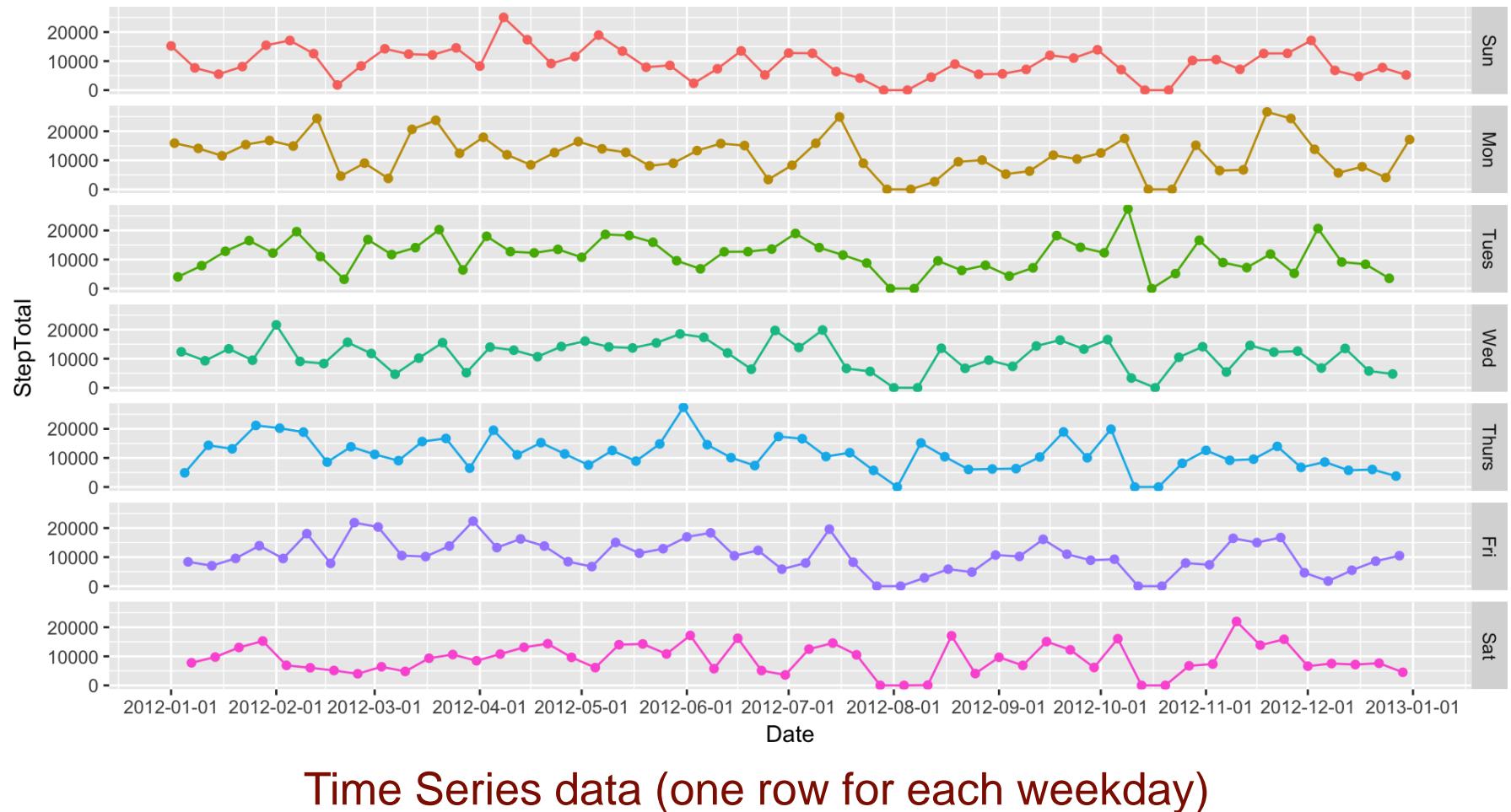
Exemple: Facets en R



7.1.1 Multiple variables and dimensions. Introduction



7.1.1 Multiple variables and dimensions. Introduction



7.1.2 Visualizing many distributions at once

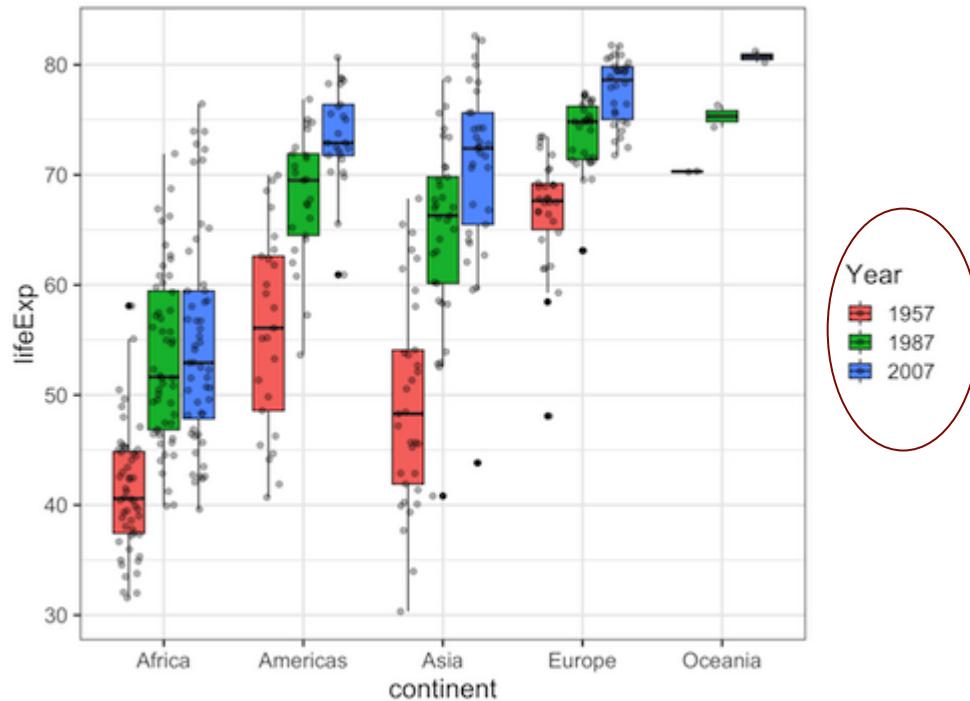
There are many scenarios in which we want to **visualize multiple distributions at the same time**.

For this, it is helpful to think in terms of the response variable and one or more grouping variables.

- The **response variable** is the variable *whose distributions we want to show*.
- The **grouping variables** define *subsets of the data with distinct distributions of the response variable*.

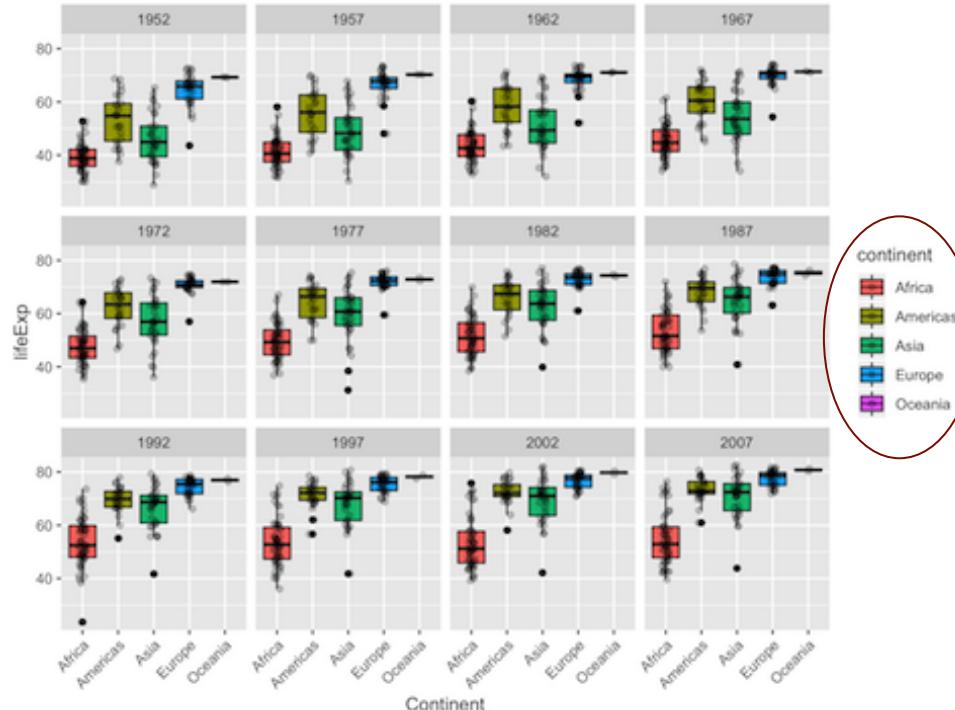
7.1.2 Visualizing many distributions at once

- The **response variable**: is the variable whose distributions we want to show (*lifeExp*).
- The **grouping variables**: define subsets of the data with distinct distributions of the response variable (*continent / year*)



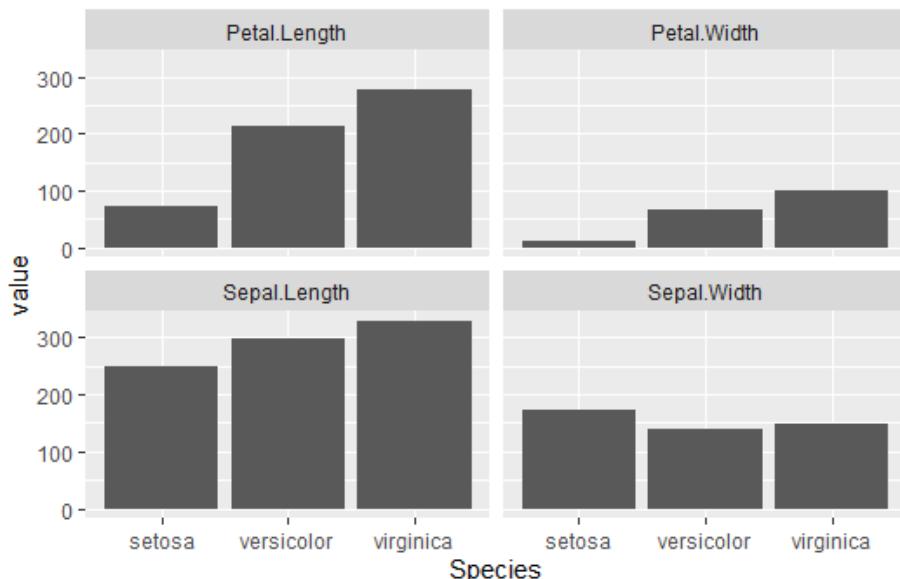
7.1.2 Visualizing many distributions at once

- The **response variable**: is the variable whose distributions we want to show (*lifeExp*).
- The **grouping variables**: define subsets of the data with distinct distributions of the response variable (*continent / year*)



7.1.2 Visualizing many distributions at once

- The **response variable**: is the variable whose distributions we want to show.
- The **grouping variables**: define subsets of the data with distinct distributions of the response variable.

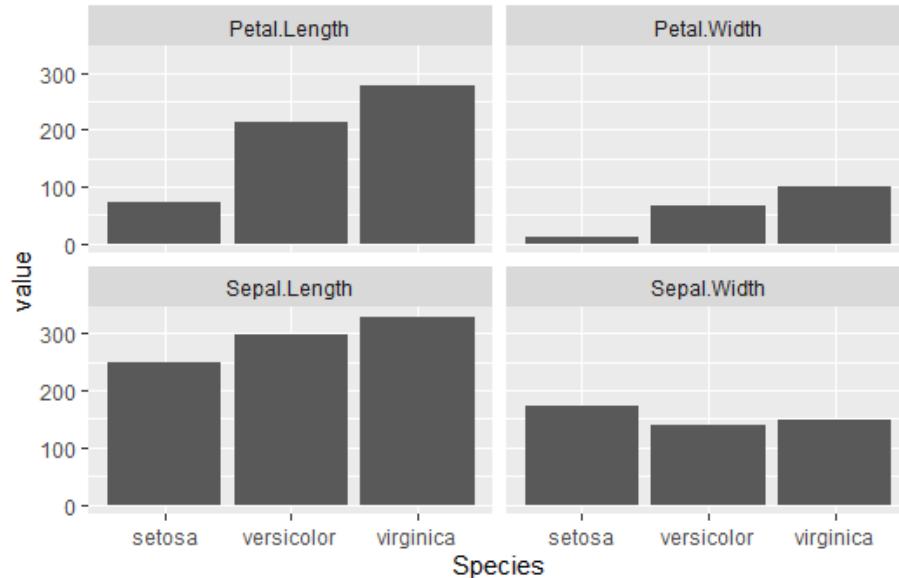


Seminar 4

```
>iris_long<-gather(iris, metric,  
value, -Species)  
>ggplot(iris_long)+aes(Species,  
value)+geom_bar(stat='identity')+  
facet_wrap(~ metric)
```

7.1.2 Visualizing many distributions at once

- The **response variable**: is the variable whose distributions we want to show (*value of the metric in cm—length or width*).
- The **grouping variables**: define subsets of the data with distinct distributions of the response variable (*species / petal/sepal*)

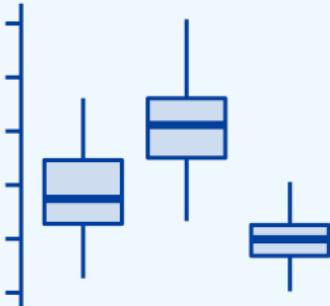


Seminar 4

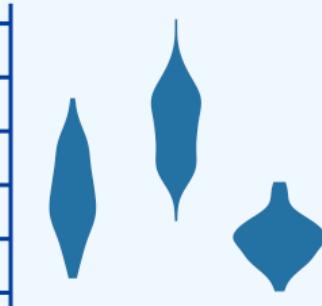
```
>iris_long<-gather(iris, metric,  
value, -Species)  
>ggplot(iris_long)+aes(Species,  
value)+geom_bar(stat='identity')+  
facet_wrap(~ metric)
```

7.1.2 Visualizing many distributions at once

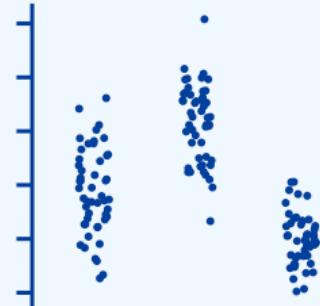
Boxplots



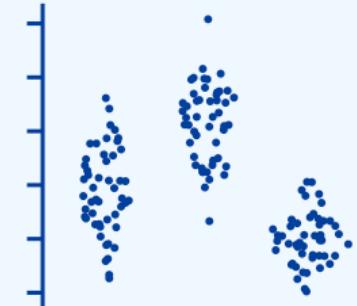
Violins



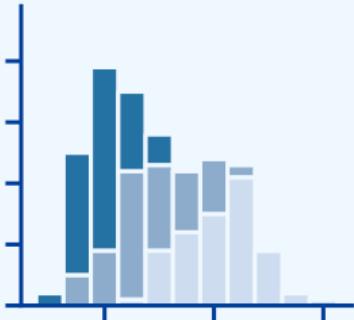
Strip Charts



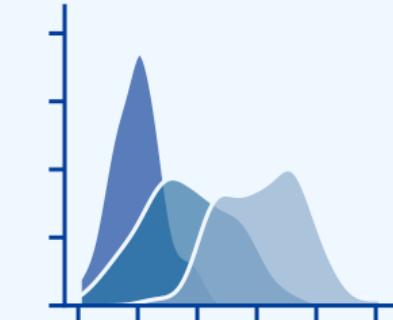
Sina Plots



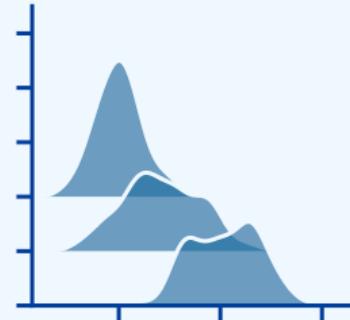
Stacked Histograms



Overlapping Densities



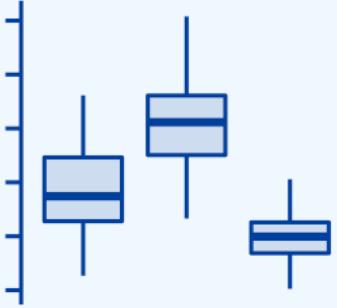
Ridgeline Plot



Claus Wilke

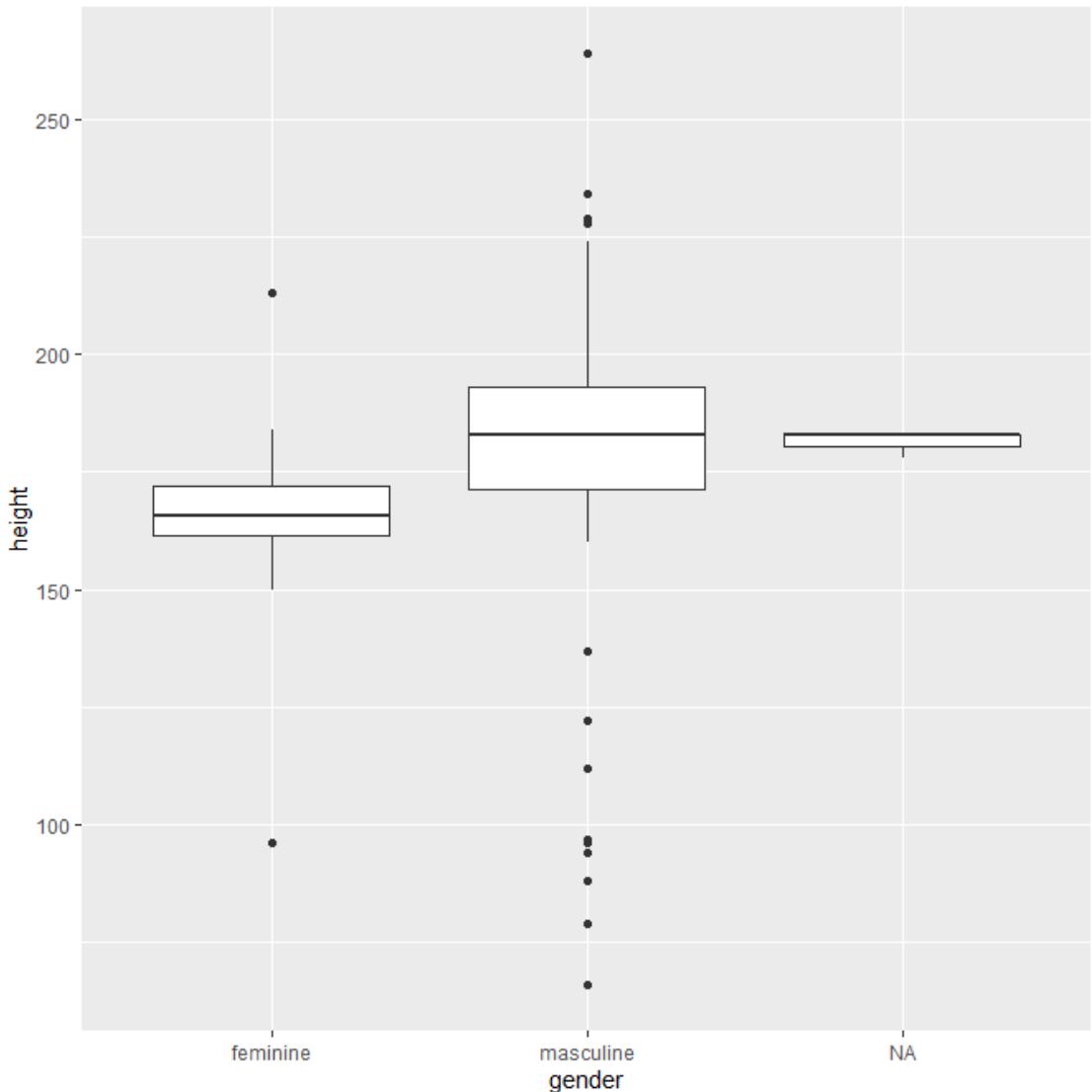
7.1.2 Visualizing many distributions at once

Boxplots



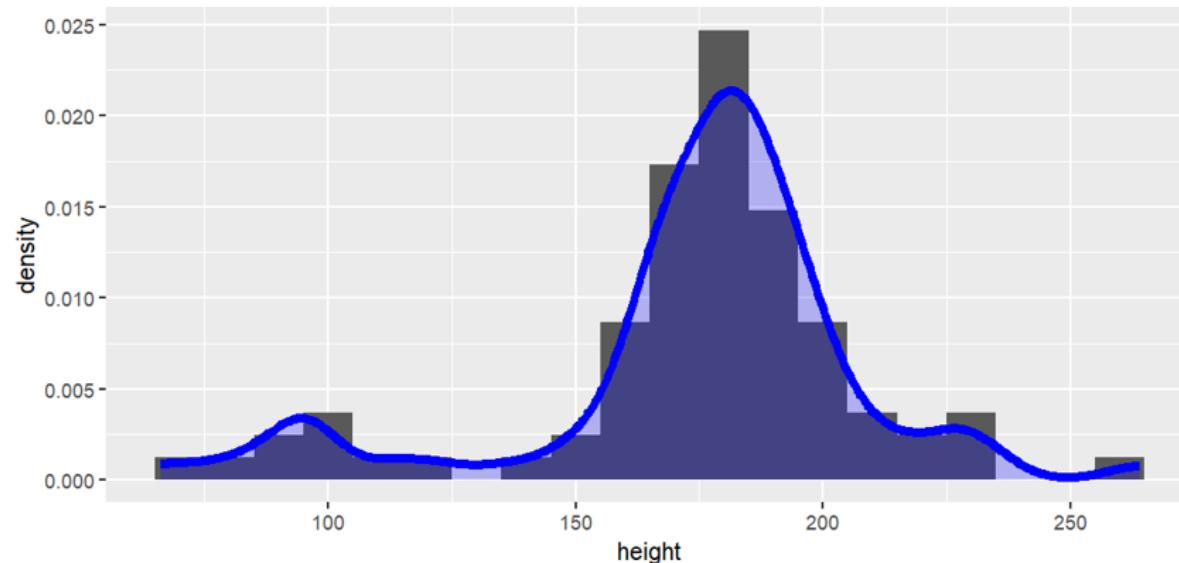
Seminar 3

```
>ggplot(starwars,aes(x=gender,  
y=height))+geom_boxplot()
```



7.1.2 Visualizing many distributions at once

Una altra opció és adjuntar ambdues gràfiques en una fent servir una transparència. Per això podeu posar `aes(y=..density..)` en el `geom_histogram`.

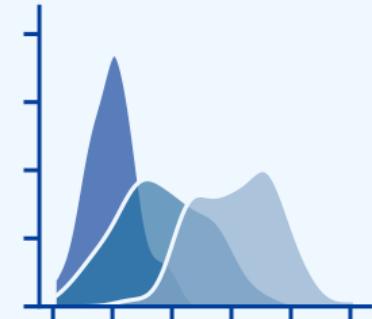


```
>ggplot(starwars,aes(x=height))+geom_histogram(binwidth=10, aes(y=..de  
nsity..))+geom_density(lwd = 2, colour = 'blue', fill = 'blue', alpha  
= 0.25)
```

Nota: `lwd` només marca el gruix de la línia de `geom_density`.

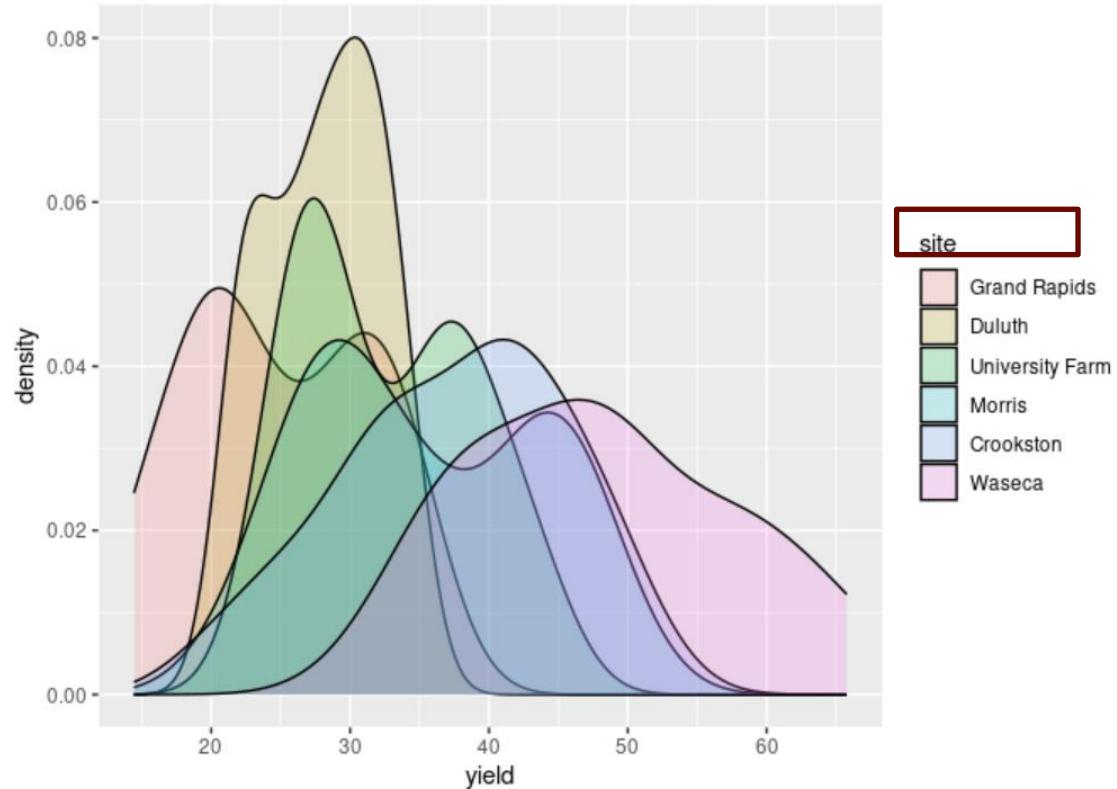
Seminar 3 Overlapping histogram and density

Overlapping Densities



7.1.2 Visualizing many distributions at once

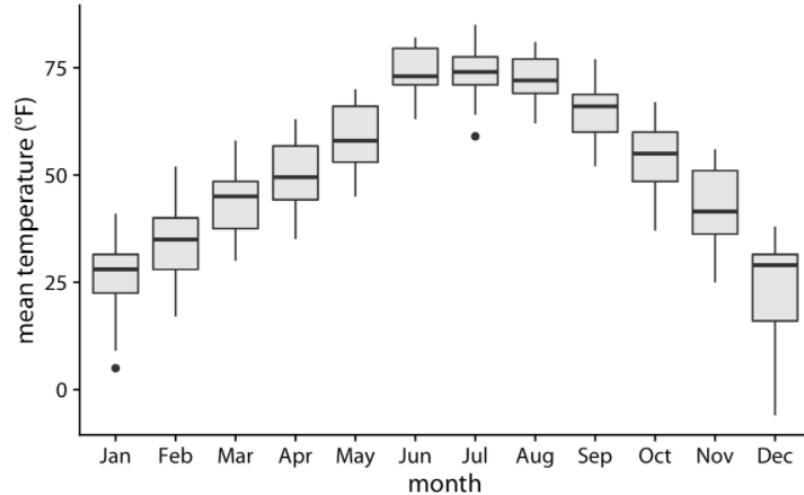
```
ggplot(barley) + geom_density(aes(x = yield, fill = site), alpha = 0.2)
```



<https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/histdens.html>

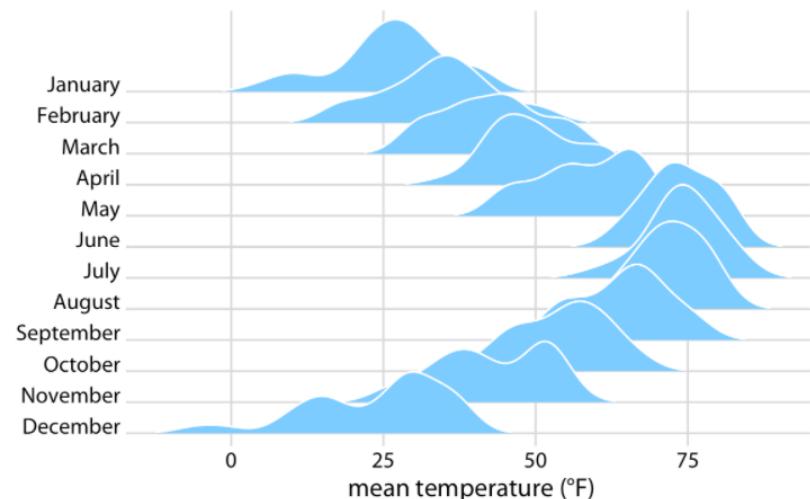
7.1.2 Visualizing many distributions at once

- **Along the vertical axis**

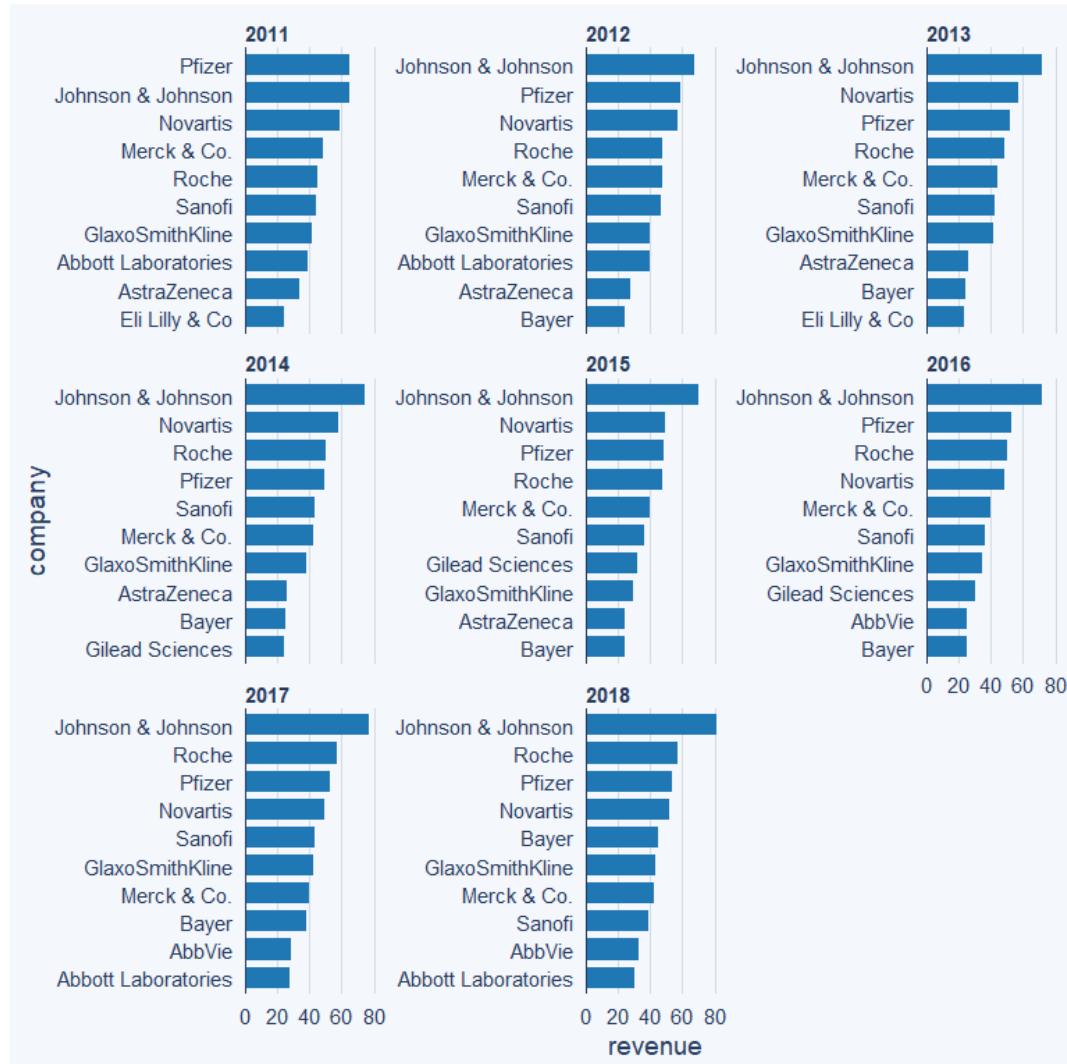


Claus Wilke

- **Along the horizontal axis**

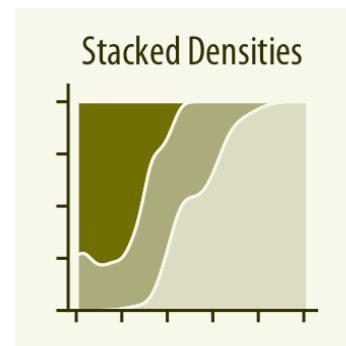
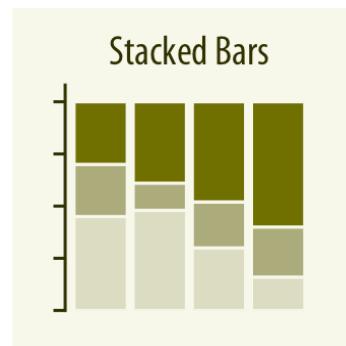
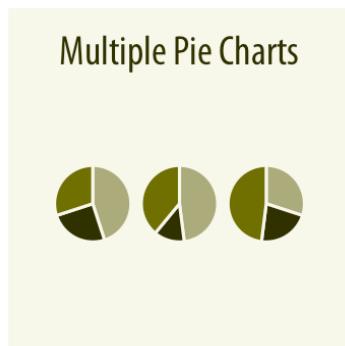


7.1.3 Visualizing many proportions at once



7.1.3 Visualizing many proportions at once

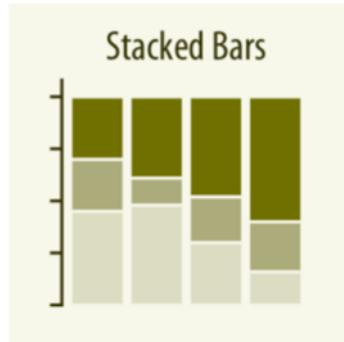
- **Pie charts** tend to be space-inefficient and often obscure relationships. (*Example seminari 3*)
- **Grouped bars** work well as long as the number of conditions compared is moderate.
- **Stacked bars** can work for large numbers of conditions.
- **Stacked densities** are appropriate when the proportions change along a continuous variable.



Claus Wilke

7.1.3 Visualizing many proportions at once

- **Stacked bars** can work for large numbers of conditions.

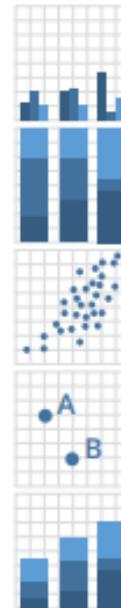


Claus Wilke



Position Adjustments

Position adjustments determine how to arrange geoms that would otherwise occupy the same space.



`s <- ggplot(mpg, aes(fl, fill = drv))`

`s + geom_bar(position = "dodge")`

Arrange elements side by side

`s + geom_bar(position = "fill")`

Stack elements on top of one another, normalize height

`e + geom_point(position = "jitter")`

Add random noise to X and Y position of each element to avoid overplotting

`e + geom_label(position = "nudge")`

Nudge labels away from points

`s + geom_bar(position = "stack")`

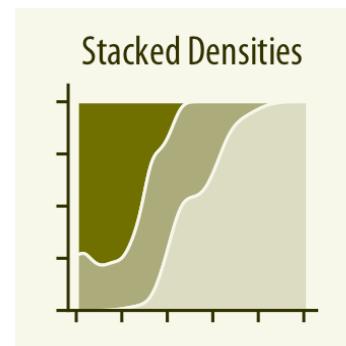
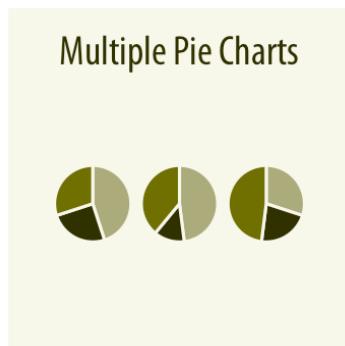
Stack elements on top of one another

Each position adjustment can be recast as a function with manual **width** and **height** arguments

`s + geom_bar(position = position_dodge(width = 1))`

7.1.3 Visualizing many proportions at once

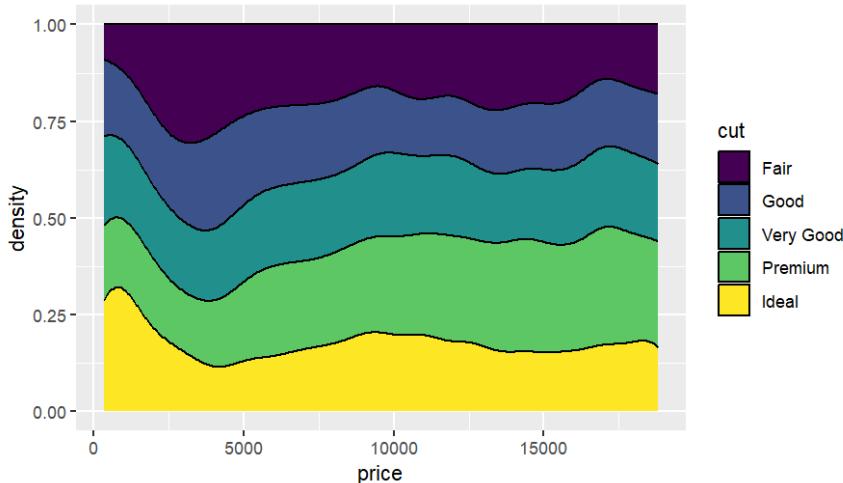
- **Pie charts** tend to be space-inefficient and often obscure relationships. (*Example seminari 3*)
- **Grouped bars** work well as long as the number of conditions compared is moderate.
- **Stacked bars** can work for large numbers of conditions.
- **Stacked densities** are appropriate when the proportions change along a continuous variable.



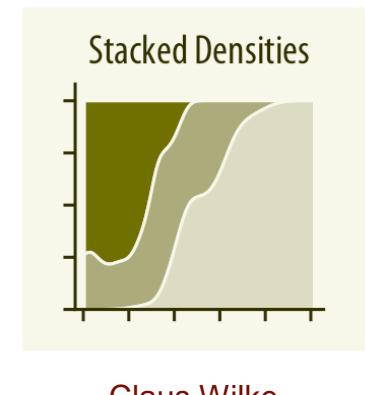
Claus Wilke

7.1.3 Visualizing many proportions at once

- **Stacked densities** are appropriate when the proportions change along a continuous variable.
- Stacking is a process where a chart is broken up across more than one categoric variables which make up the whole.

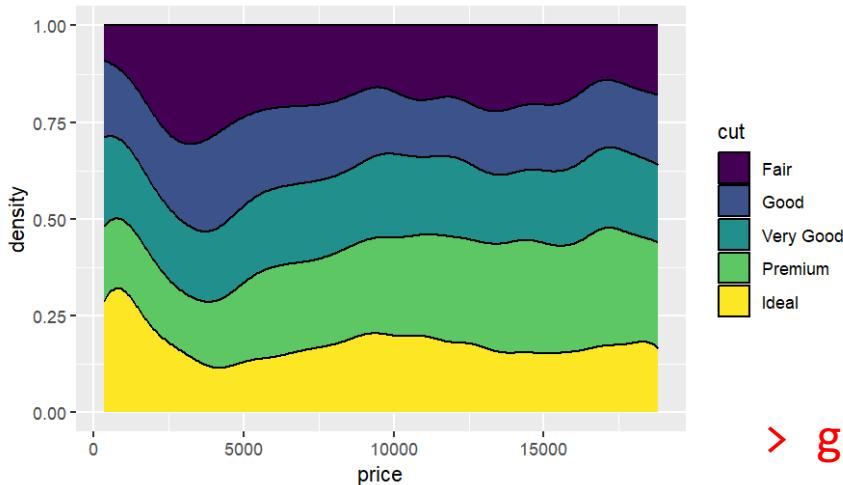


Diamonds contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds

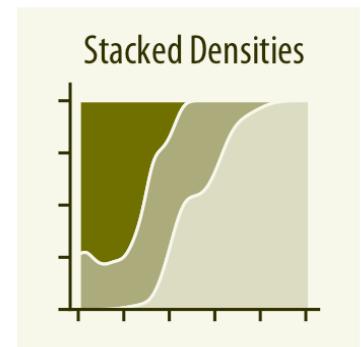


7.1.3 Visualizing many proportions at once

- **Stacked densities** are appropriate when the proportions change along a continuous variable.
- Stacking is a process where a chart is broken up across more than one categoric variables which make up the whole.



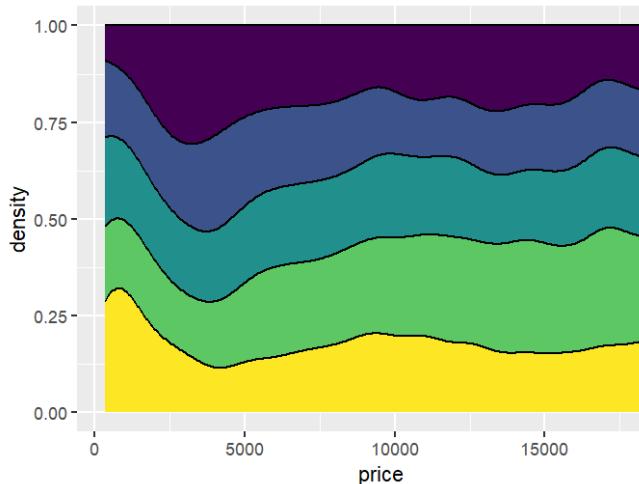
Diamonds contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds



Claus Wilke

```
> ggplot(data=diamonds, aes(x=price,  
group=cut, fill=cut)) +  
geom_density(adjust=1.5, position="fill")
```

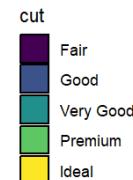
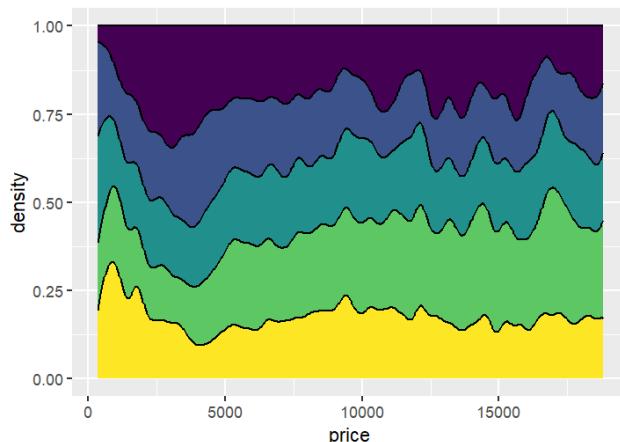
7.1.3 Visualizing many proportions at once



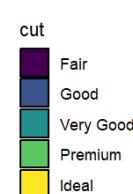
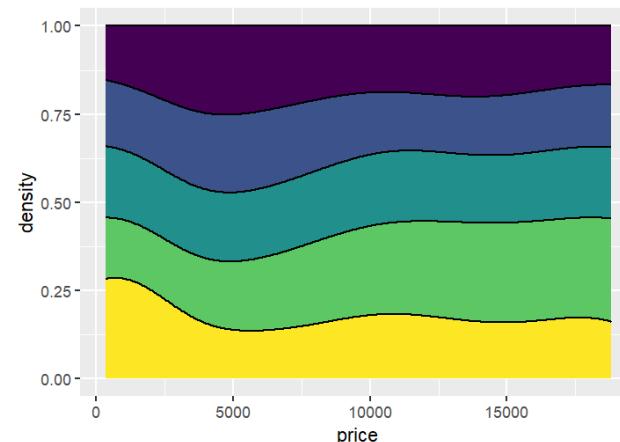
It contains information on price, cut characteristics, color, carats, etc... of nearly 54,000 diamonds

```
> ggplot(data=diamonds, aes(x=price,  
group=cut, fill=cut)) +  
geom_density(adjust=1.5, position="fill")
```

A multiplicative bandwidth adjustment. This makes it possible to adjust the bandwidth while still using a bandwidth estimator



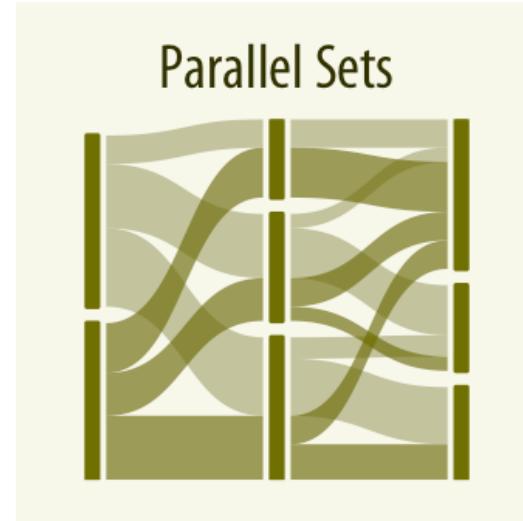
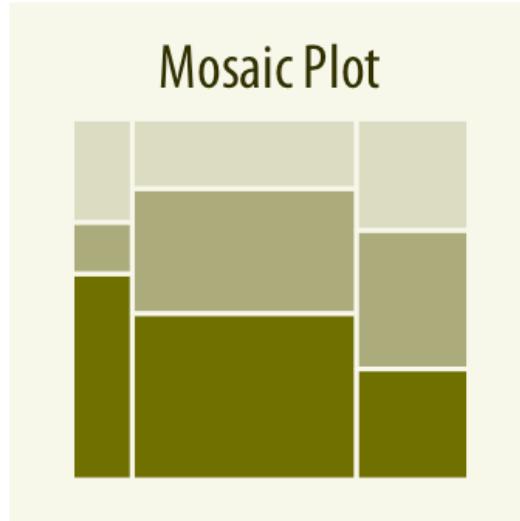
adjust=0.5



adjust=4

7.1.3 Visualizing many proportions at once

When proportions are specified according to multiple grouping variables:



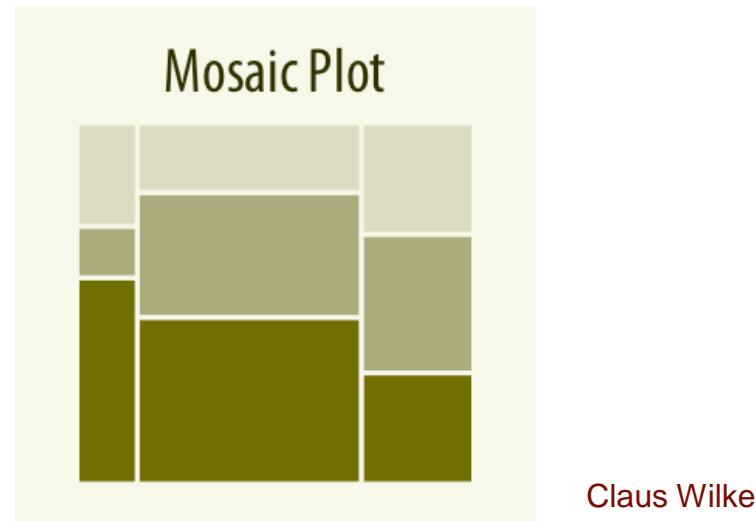
Claus Wilke

Mosaic plots, treemaps, or parallel sets are useful visualization approaches

7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

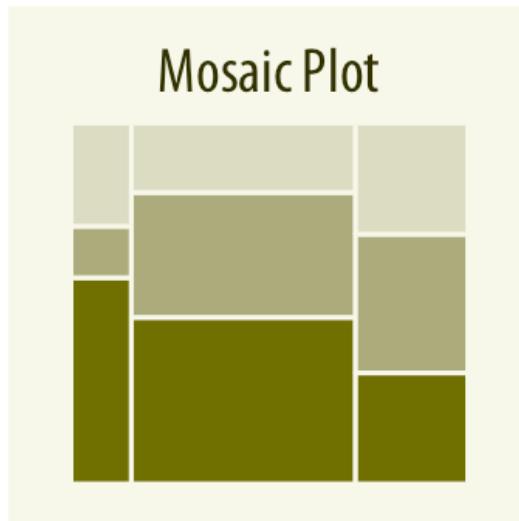
Mosaic Plot: assumes that each level of one grouping variable can be combined with each level of another grouping variable.



7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

- **Mosaic plots** looks similar to the stacked bar plot. However, in a mosaic plot **both the heights and the width of individual shaded areas vary**



Claus Wilke

7.1.3 Visualizing many proportions at once

Whenever we have categories that overlap, it is best to show clearly how they relate to each other

- **Mosaic plots example:**

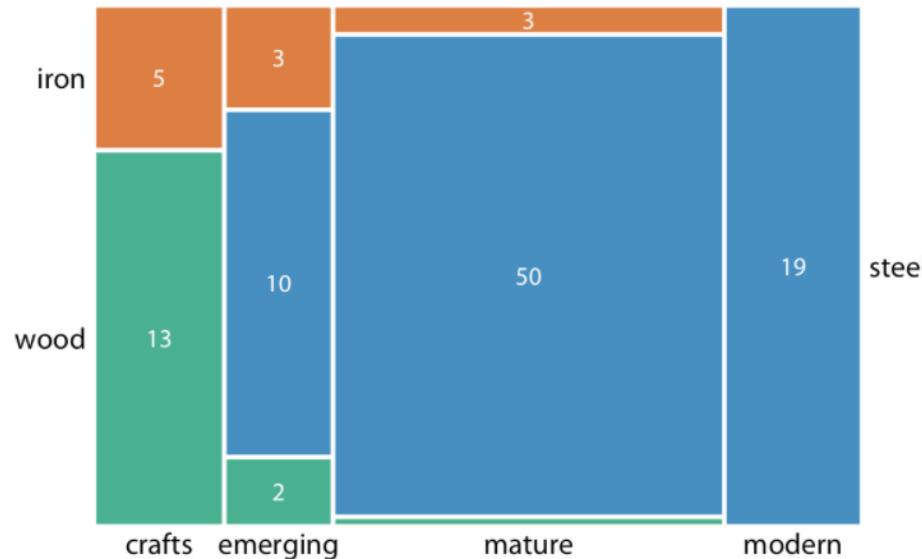


Figure 11.3: Breakdown of bridges in Pittsburgh [by construction material \(steel, wood, iron\)](#) and [by era of construction \(crafts, emerging, mature, modern\)](#), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fennes, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

Claus Wilke

7.1.3 Visualizing many proportions at once

- Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable, whereas treemaps do not make such an assumption.

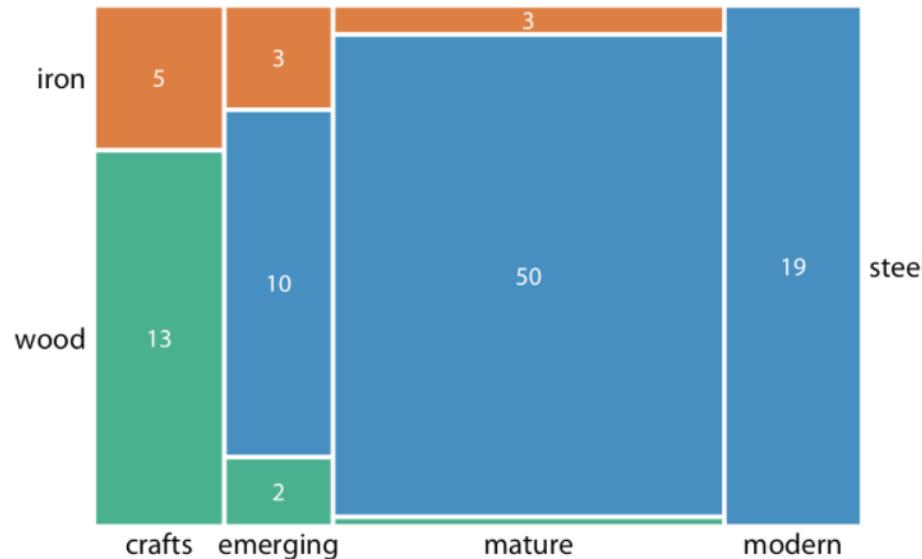


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fennes, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)



7.1.3 Visualizing many proportions at once

- Mosaic plots

We begin by **placing one categorical variable along the x axis** (here, era of bridge construction) **and subdivide the x axis by the relative proportions that make up the categories.**

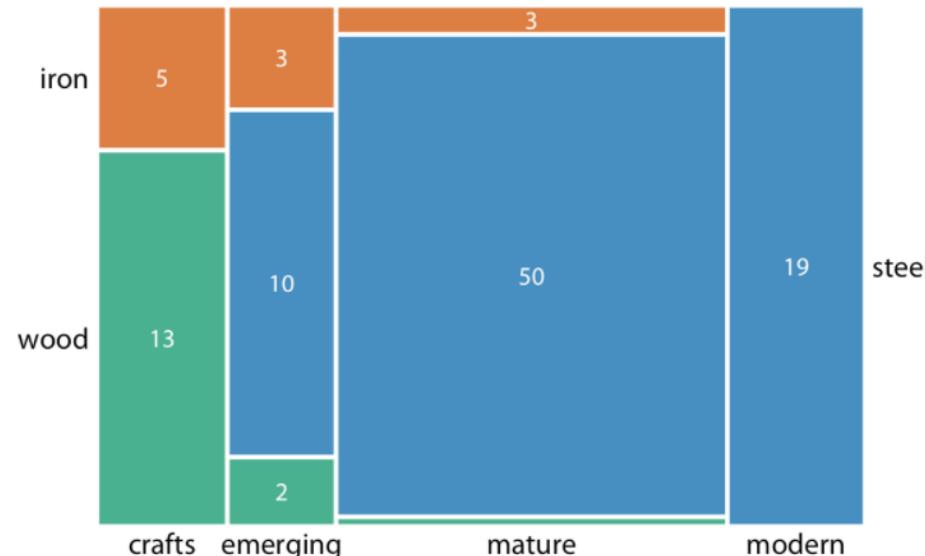


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fennes, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- **Mosaic plots**

We begin by **placing one categorical variable along the x axis** (here, era of bridge construction) **and subdivide the x axis by the relative proportions that make up the categories.**

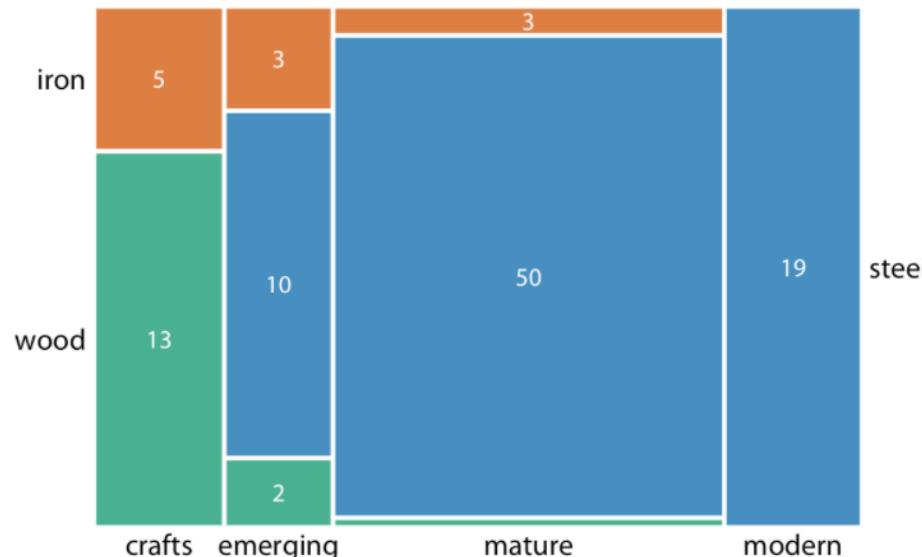


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fennes, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

We then place the other categorical variable along the y axis (here, *building material*) and, within each category along the x axis, subdivide the y axis by the relative proportions that make up the categories of the y variable.

7.1.3 Visualizing many proportions at once

- Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable, whereas treemaps do not make such an assumption.

The result is a set of rectangles whose areas are proportional to the number of cases representing each possible combination of the two categorical variables.

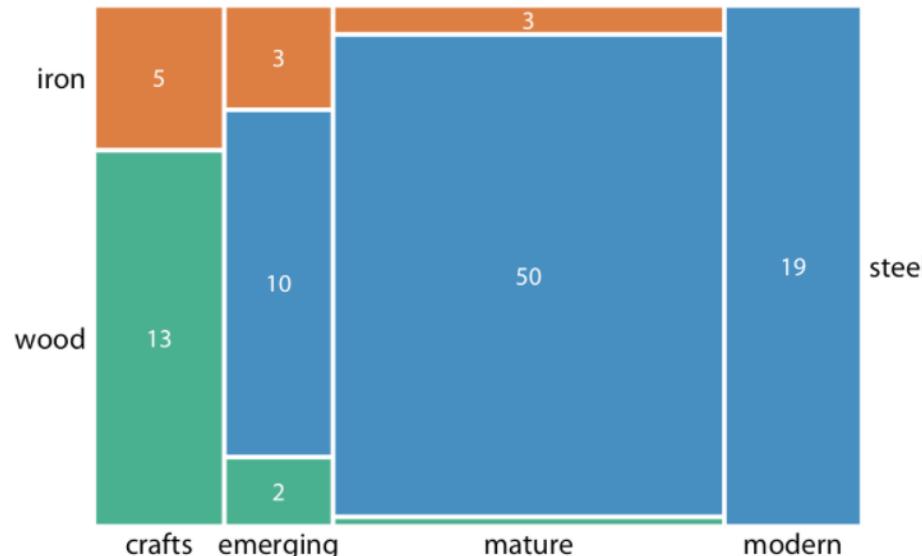


Figure 11.3: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a mosaic plot. The widths of each rectangle are proportional to the number of bridges constructed in that era, and the heights are proportional to the number of bridges constructed from that material. Numbers represent the counts of bridges within each category. Data source: Yoram Reich and Steven J. Fennes, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)



7.1.2 Visualizing many proportions at once

- **Treemaps** work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

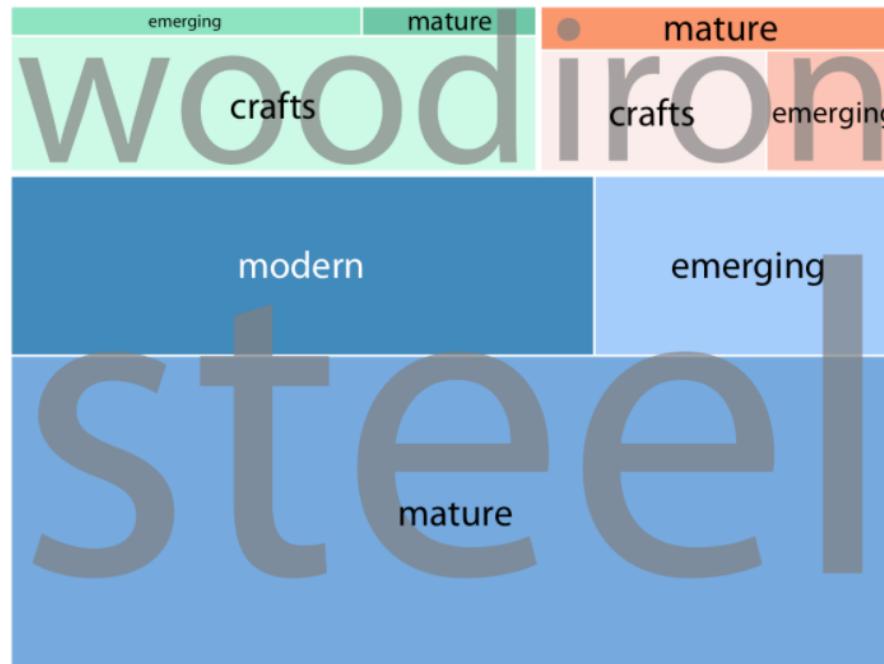


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)



7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

In a treemap, we recursively nest rectangles inside each other.

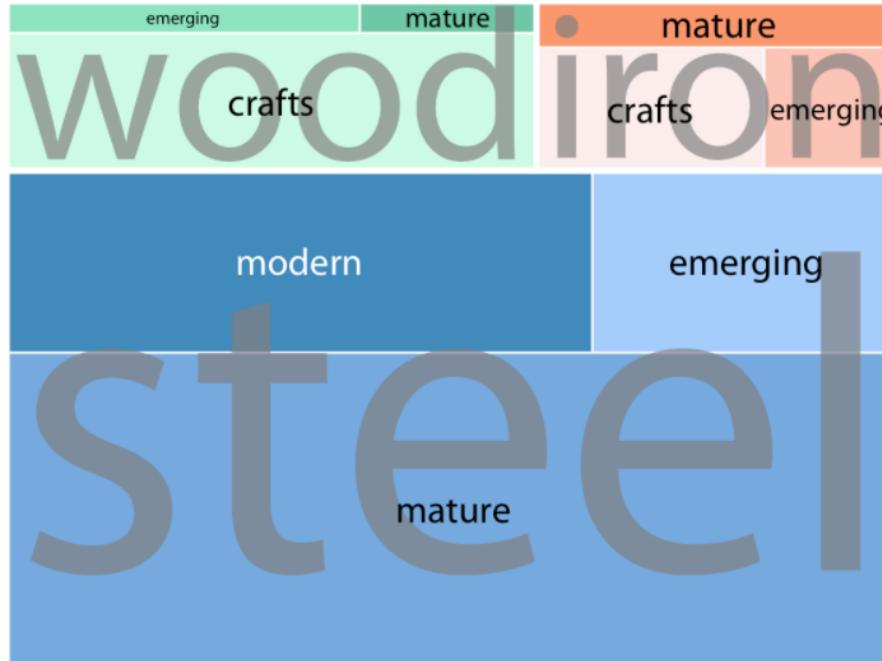


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)



7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

For example, in the case of the Pittsburgh bridges, we can **first subdivide the total area into three parts representing the three building materials wood, iron, and steel.**

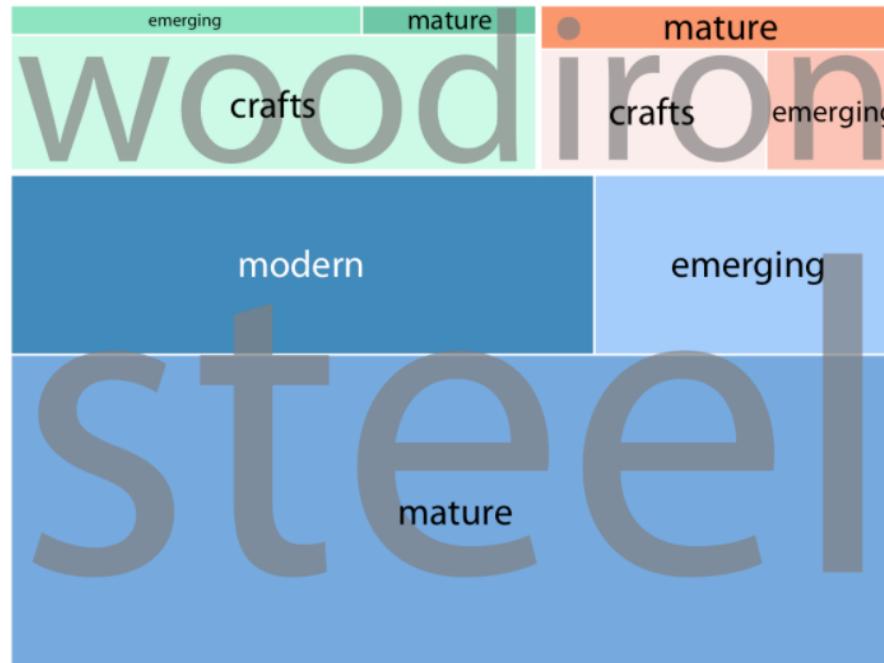


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)



7.1.3 Visualizing many proportions at once

- Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.

For example, in the case of the Pittsburgh bridges, we can first subdivide the total area into three parts representing the three building materials wood, iron, and steel.

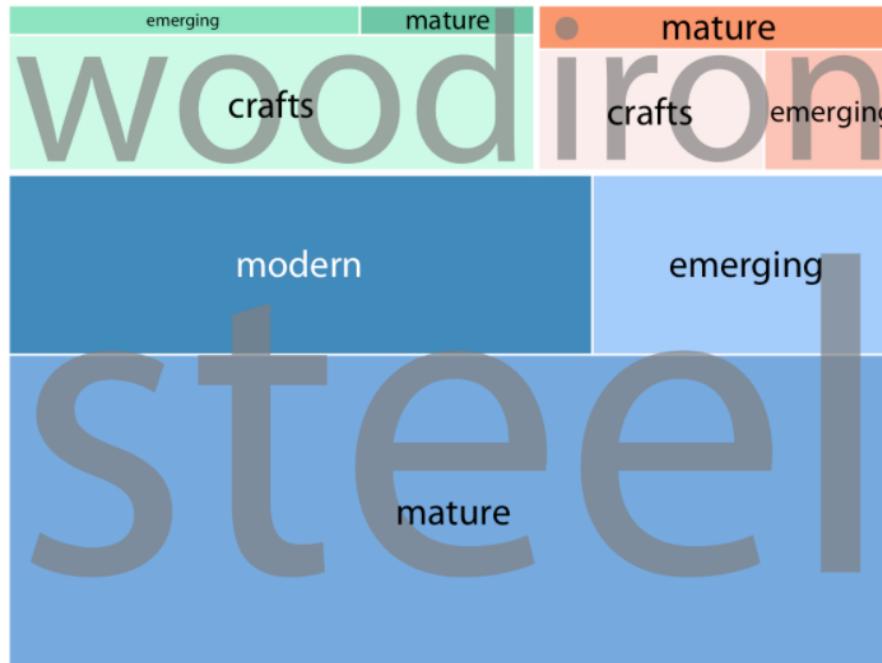


Figure 11.4: Breakdown of bridges in Pittsburgh by construction material (steel, wood, iron) and by era of construction (crafts, emerging, mature, modern), shown as a treemap. The area of each rectangle is proportional to the number of bridges of that type. Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

7.1.3 Visualizing many proportions at once

- **Mosaic plots** assume that every level of one grouping variable can be combined with every level of another grouping variable, whereas treemaps do not make such an assumption. (width/height)
- **Treemaps** work well even if the subdivisions of one group are entirely distinct from the subdivisions of another. (area)
- **Parallel sets** work better than either mosaic plots or treemaps when there are more than two grouping variables. (next class)

7.1.3 Visualizing many proportions at once

Remember: A treemap is a rectangular plot divided into tiles, each of which represents a single observation. It is a nice way of displaying hierarchical data by using nested rectangles.

The relative area of each tile expresses a continuous variable.

- treemapify provides ggplot2 geoms for drawing [treemaps](#)
- Install the release version of treemapify from CRAN:

```
>install.packages("treemapify")
> library(treemapify)
```
- geom_treemap () – A ‘ggplot2’ geom to draw a treemap

Material of interest (extra):

<https://cran.r-project.org/web/packages/treemapify/treemapify.pdf>

<https://cran.r-project.org/web/packages/treemapify/vignettes/introduction-to-treemapify.html>

7.1.3 Visualizing many proportions at once

One simple example that you can test (first remember to install the package and load the library):

1. Let's read the .csv dataframe:

<https://raw.githubusercontent.com/selva86/datasets/master/proglanguages.csv>

```
> Proglangs <-  
read.csv("https://raw.githubusercontent.com/selva86/datasets/master/proglanguages.csv")
```

This dataframe contains a hierarchical list of programming languages:

```
'data.frame':      40 obs. of  4 variables:  
 $ id   : chr "Java (general)" "PHP (general)" "dotNet (general)" "Python (general)" ...  
 $ value : int 423 253 220 219 185 121 91 89 83 79 ...  
 $ parent: chr "Java" "PHP" "dotNet" "Python" ...  
 $ rank  : num 40 39 38 37 36 35 34 33 32 31 ...
```

7.1.3 Visualizing many proportions at once

This dataframe contains a hierarchical list of programming languages:

```
'data.frame': 40 obs. of 4 variables:  
 $ id    : chr "Java (general)" "PHP (general)" "dotNet (general)" "Python (general)" ...  
 $ value : int 423 253 220 219 185 121 91 89 83 79 ...  
 $ parent: chr "Java" "PHP" "dotNet" "Python" ...  
 $ rank  : num 40 39 38 37 36 35 34 33 32 31 ...
```

2. In order to create a treemap, the data must be converted to desired format using `treemapify()`. The important requirement is, you need to identify in your data:
 - One numerical continuous variable each describes the area of the tiles ('`value`')
 - One variable for fill color ('`parent`')
 - One variable that has the tile's label ('`id`')
 - The parent group ('`parent`')

7.1.3 Visualizing many proportions at once

2. In order to create a treemap, the data must be converted to desired format using `treemapify()`. The important requirement is, you need to identify in your data:
 - One variable each describes the area of the tiles ('value')
 - One variable for fill color ('parent')
 - One variable that has the tile's label ('id')
 - The parent group ('parent')

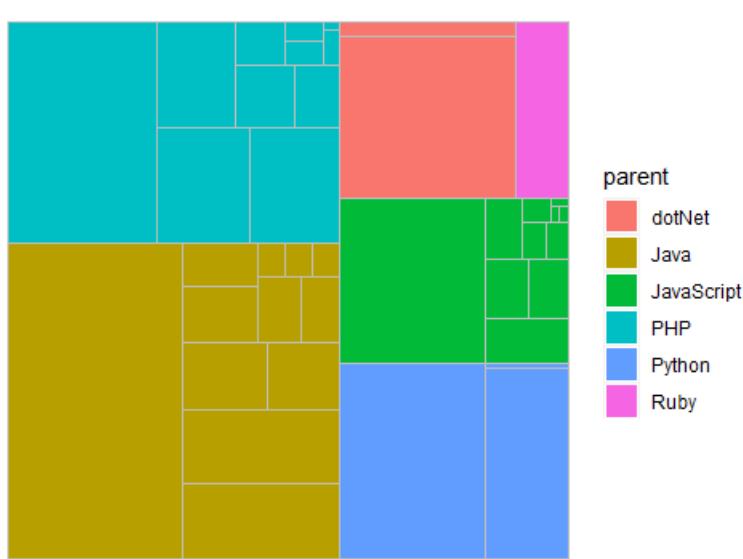
And map them by using `aes`:

```
> ggplot(Proglangs, aes(area=value, fill=parent,  
label=id, subgroup=parent)) + geom_treemap()
```

3. Use `geom_treemap`

7.1.3 Visualizing many proportions at once

```
> ggplot(Proglangs, aes(area=value, fill=parent,  
label=id, subgroup=parent)) + geom_treemap()
```

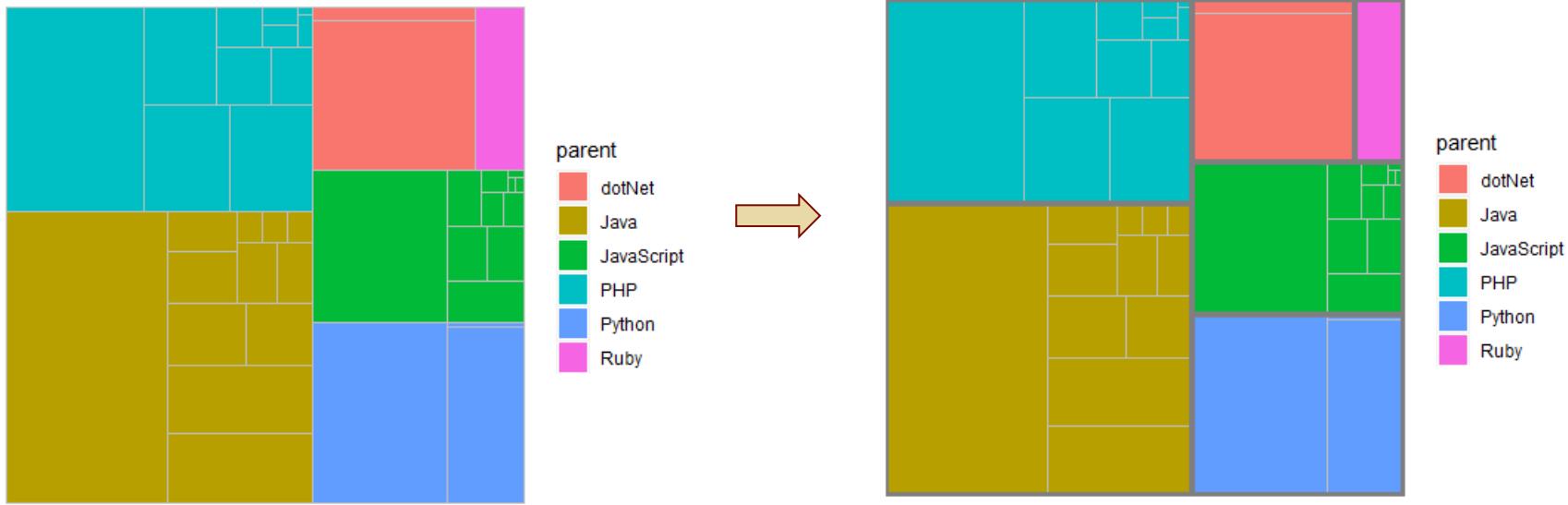


Remember that you can assign this to a variable and add the other optional layers afterwards

4. You can now add the main group bordering

```
> ggplot(Proglangs, aes(area=value, fill=parent,  
label=id, subgroup=parent)) + geom_treemap() +  
geom_treemap_subgroup_border()
```

7.1.3 Visualizing many proportions at once

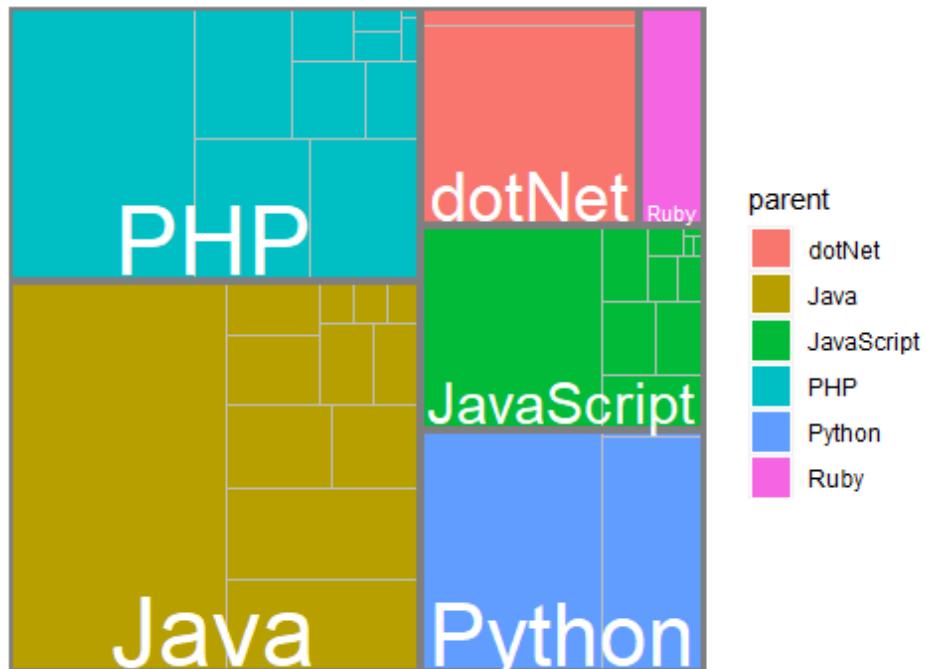


Thanks to:
`geom_treemap_subgroup_border()`

7.1.3 Visualizing many proportions at once

5. We can add subgroup heading in white

```
> ggplot(Proglangs, aes(area=value, fill=parent,  
label=id, subgroup=parent)) + geom_treemap()  
+geom_treemap_subgroup_border()  
+geom_treemap_subgroup_text(color='white')
```



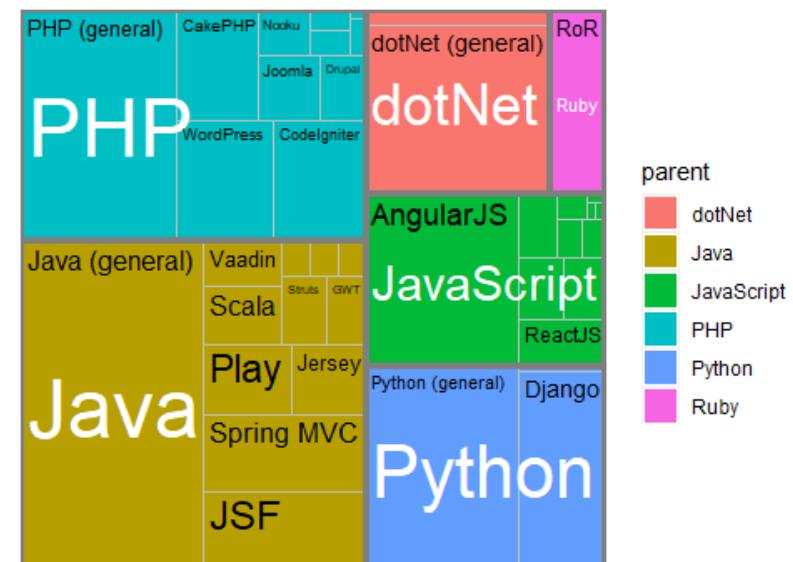
7.1.3 Visualizing many proportions at once

6. Add all other text in black

```
> ggplot(Proglangs, aes(area=value,  
fill=parent,label=id,subgroup=parent))+geom_treemap()  
+geom_treemap_subgroup_border()  
+geom_treemap_subgroup_text(color="white",place="left")  
+geom_treemap_text (aes(label=id))
```

More examples:

<https://cran.r-project.org/web/packages/treemapify/vignettes/introduction-to-treemapify.html>



7.1.4 Visualizing many relations at once

Many datasets contain two or more quantitative variables, and we may be interested in how these variables relate to each other.

Scatterplot matrix (SPLOM) uses multiple scatterplots to determine the **correlation** (if any) between a series of variables.

These scatterplots are then ***organized into a matrix***, making it easy to look at all the potential correlations in one place.

7.1.4 Correlation between series variables

Bubble plot

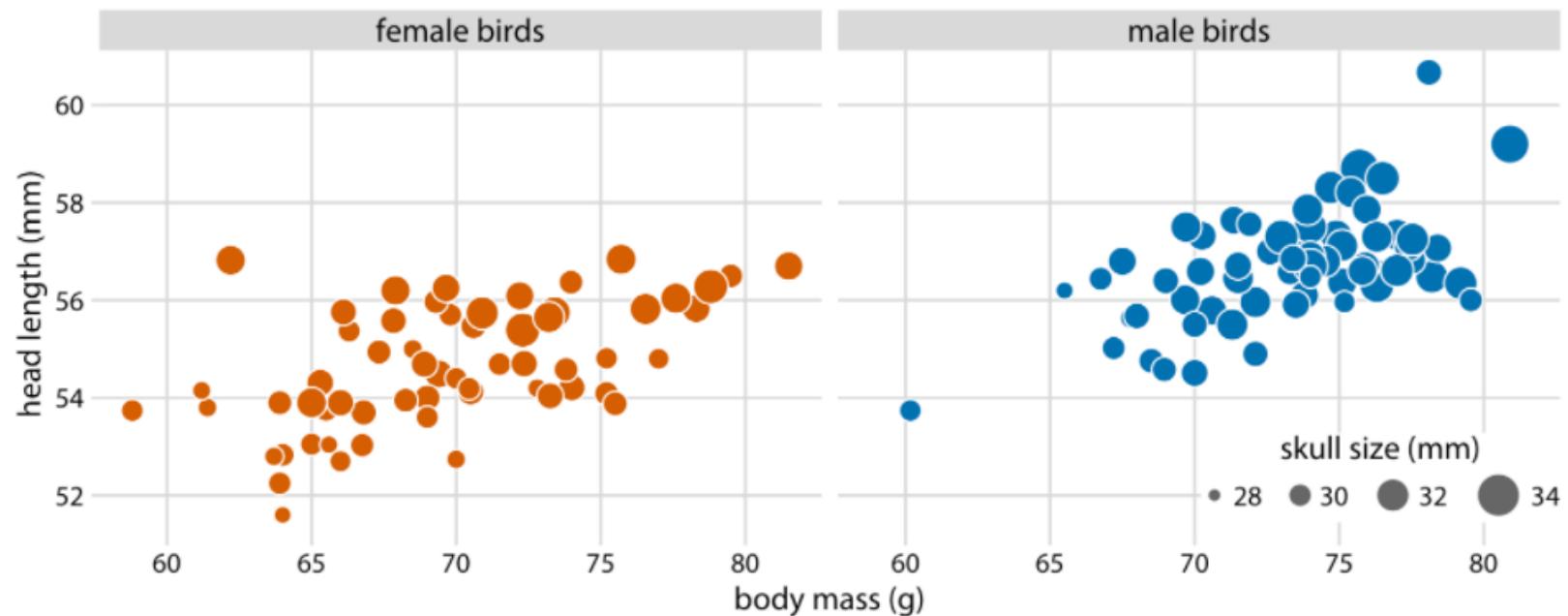


Figure 12.3: Head length versus body mass for 123 blue jays. The birds' sex is indicated by color, and the birds' skull size by symbol size. Head-length measurements include the length of the bill while skull-size measurements do not. Head length and skull size tend to be correlated, but there are some birds with unusually long or short bills given their skull size. Data source: Keith Tarvin, Oberlin College

Claus Wilke

7.1.4 Correlation between series variables

Scatterplot matrix (SPLOM) can provide answers to the following questions:

- Are there **pair wise relationships** between the variables?
- If there are relationships, what is **the nature of these relationships**?
- Are there **outliers** in the data?
- Is there **clustering by groups** in the data?

7.1.4 Scatterplot Matrix



More SPLOM examples

Thanks for your attention!



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



8. Color y Precisión Visual

Sol Bucalo
sol.bucalo@uab.cat

Guillermo Marin
guillermo.marin@uab.cat



Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

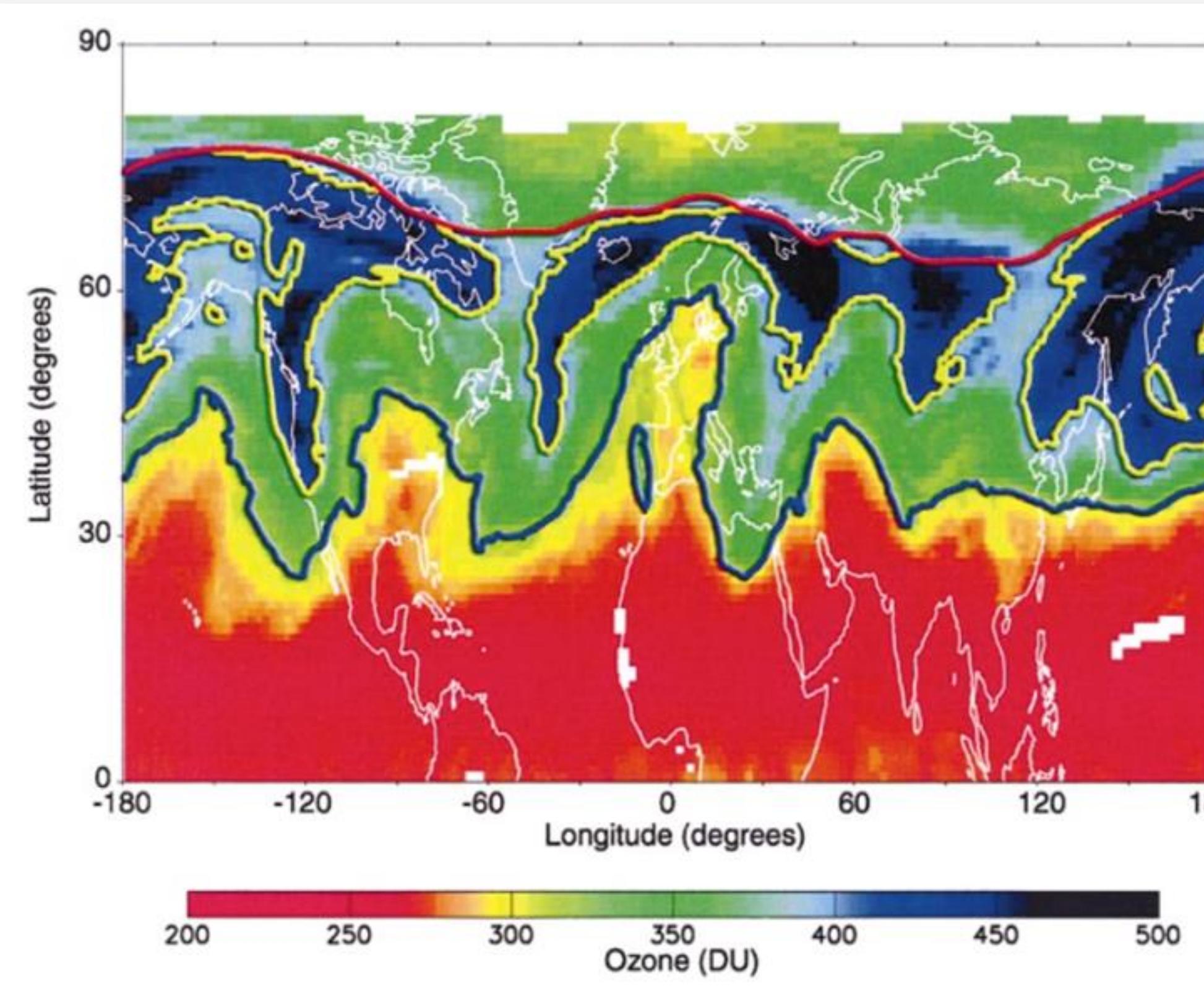
UAB
Universitat Autònoma
de Barcelona

8.1

C O I O R

The word 'COIOR' is rendered in a large, bold, sans-serif font. Each letter is defined by multiple concentric circles of pink and teal, creating a 3D effect. The 'I' is represented by two vertical rectangles with the same pink and teal border style. A horizontal line is positioned below the letters.

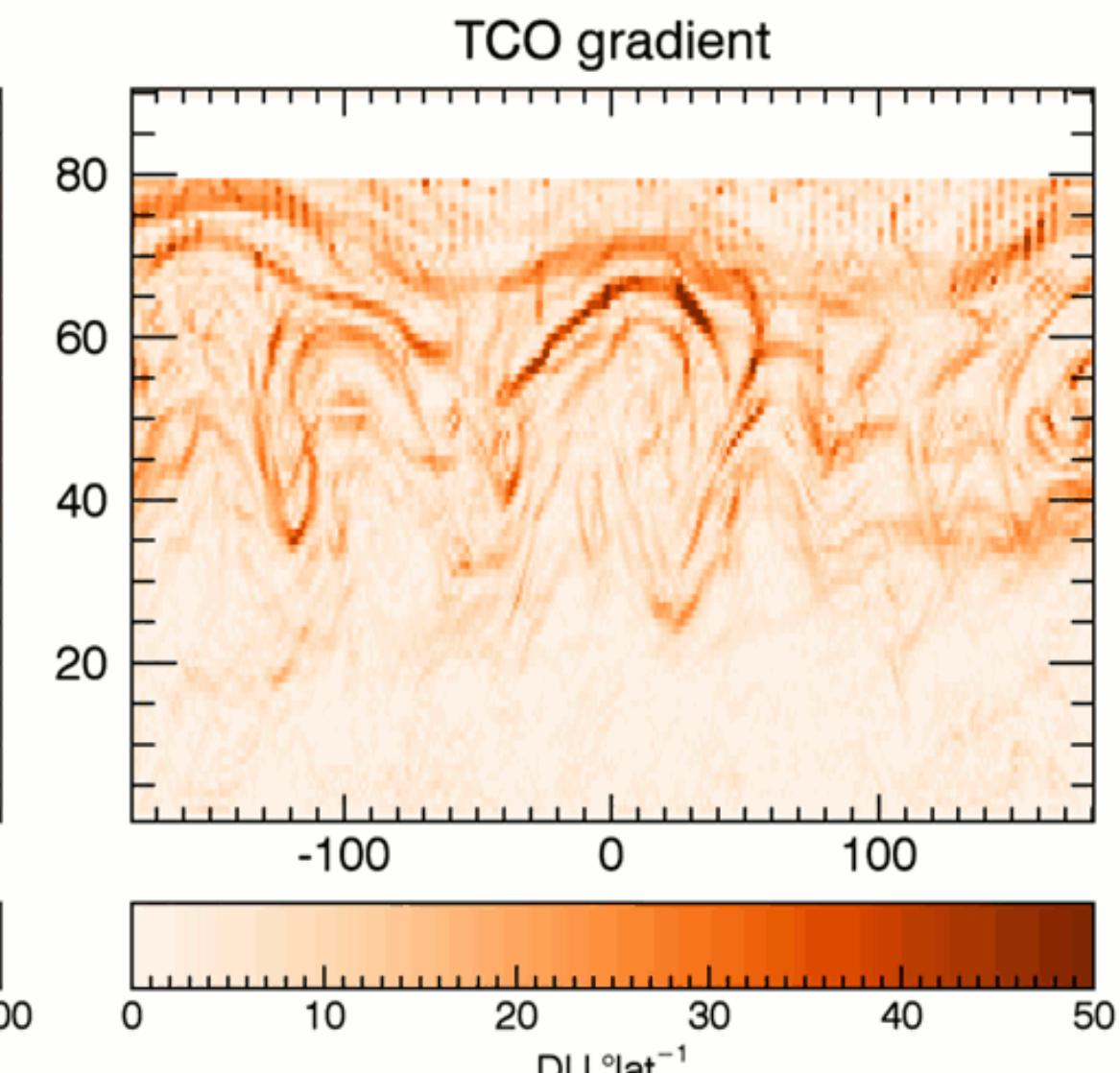
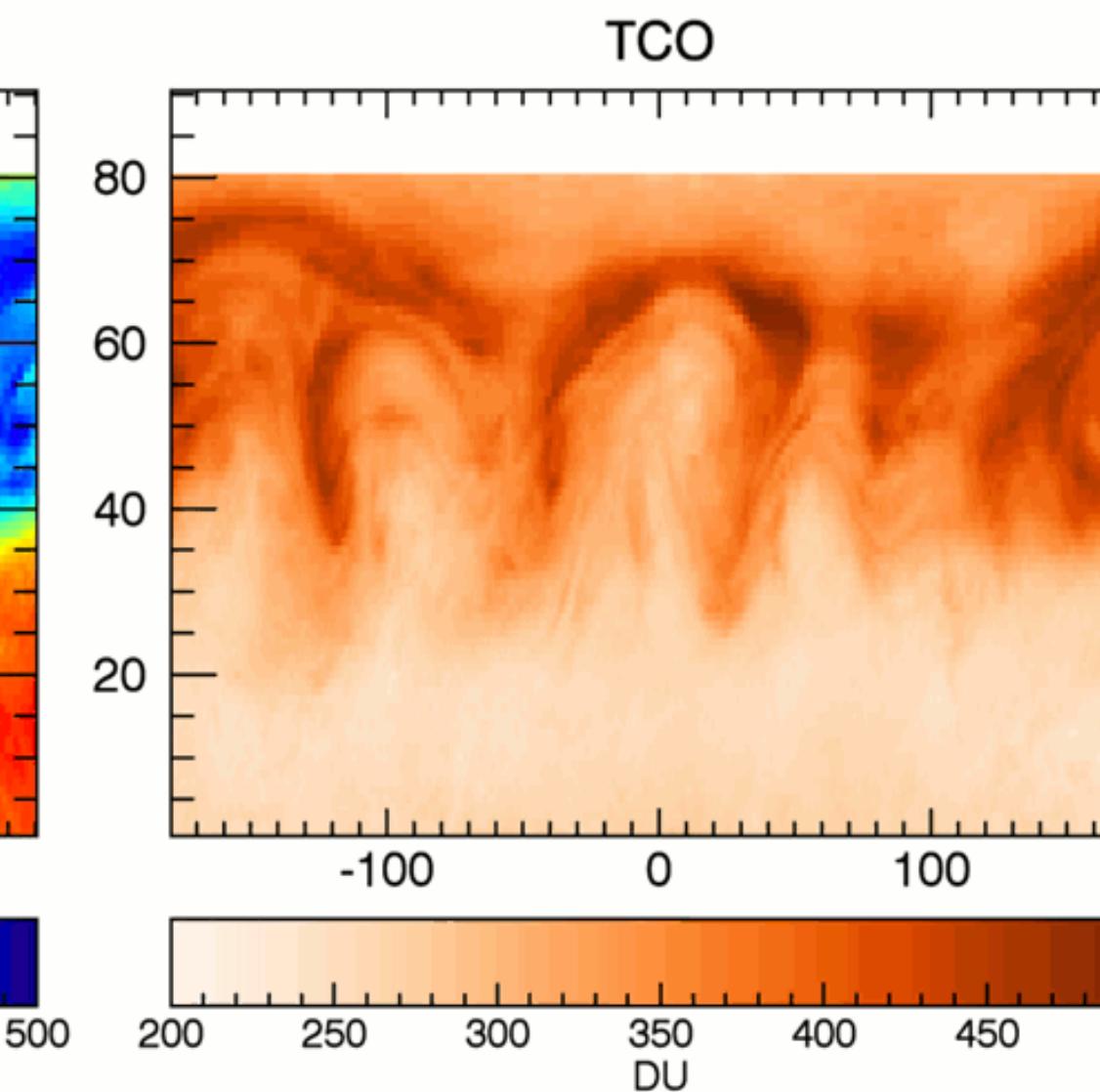
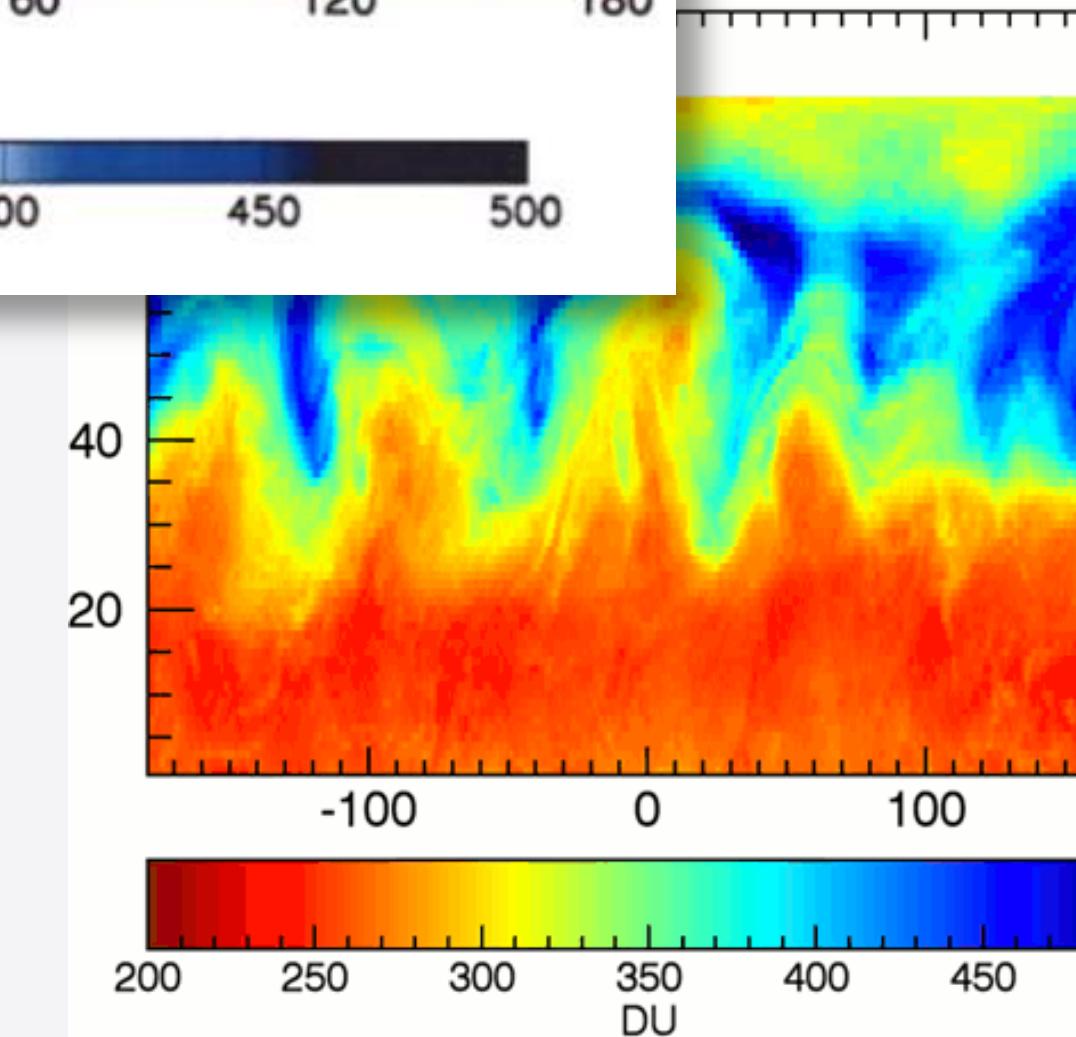
BEWARE OF THE RAINBOW



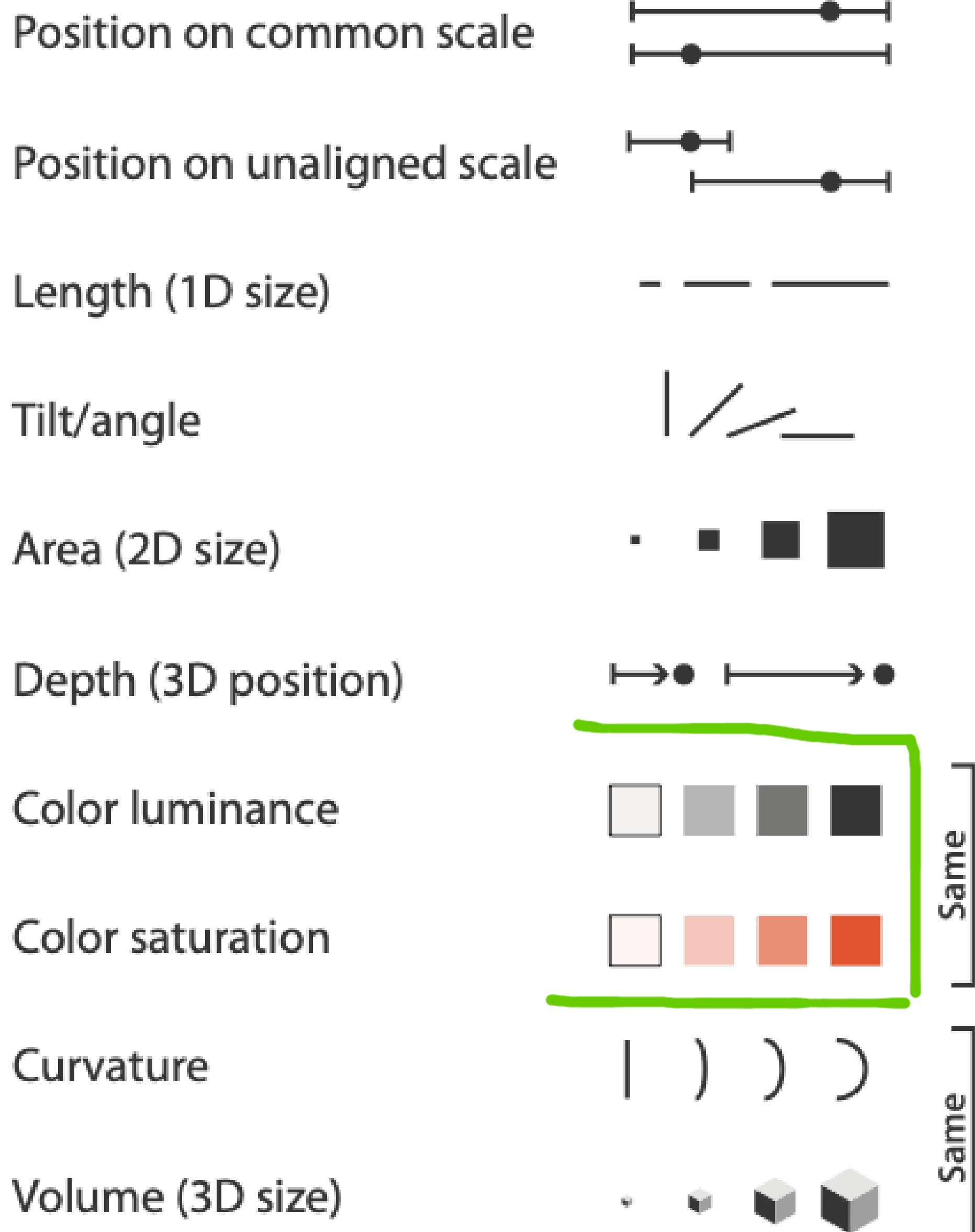
Hudson, R. D., Andrade, M. F., Follette, M. B., and Frolov, A. D.:

The total ozone field separated into meteorological regimes – Part II: Northern Hemisphere mid-latitude total ozone trends, Atmos. Chem. Phys., 6, 5183-5191, <https://doi.org/10.5194/acp-6-5183-2006>, 2006.

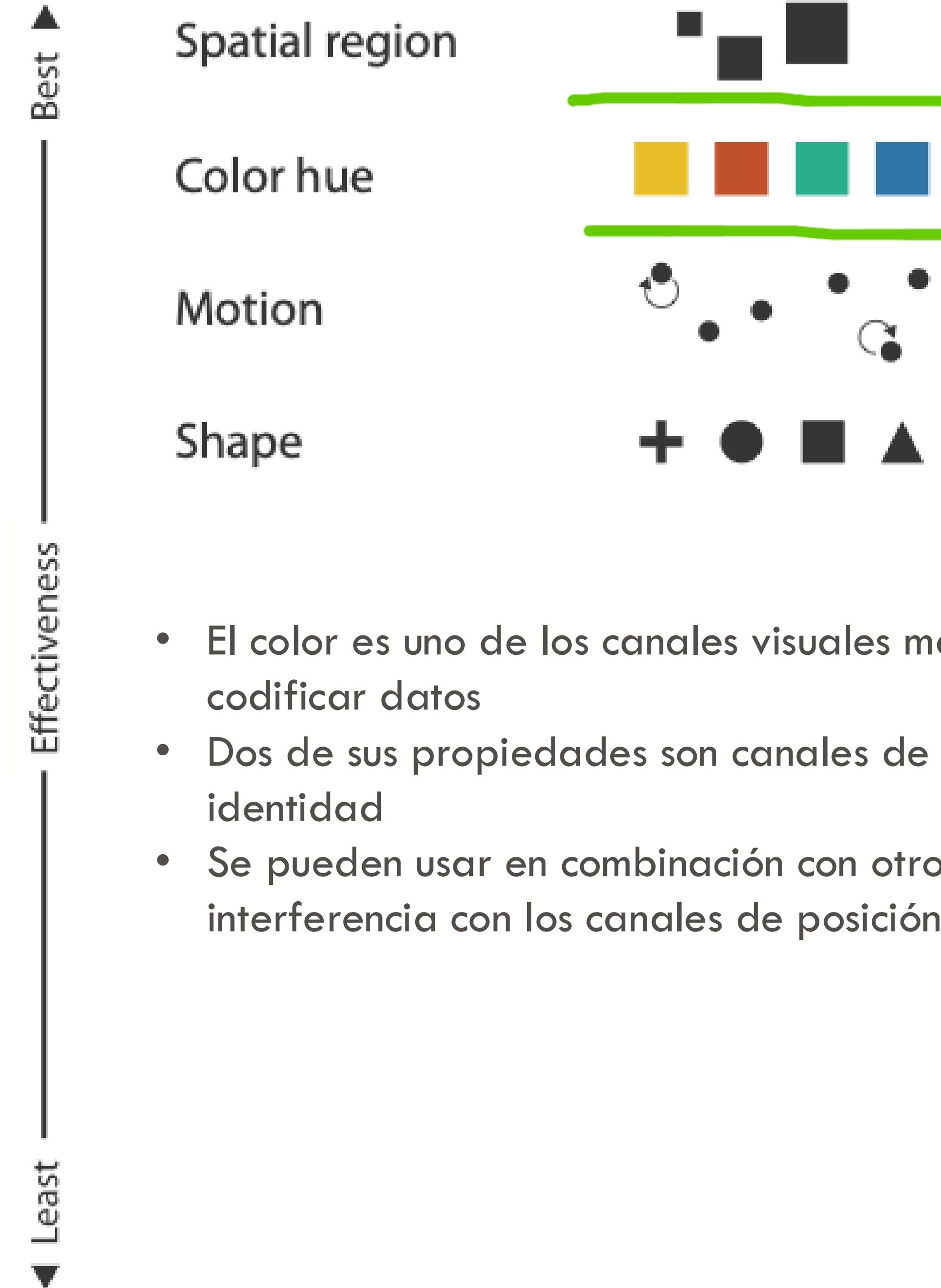
The wrong color palette can show effects in the visualization not present in the data



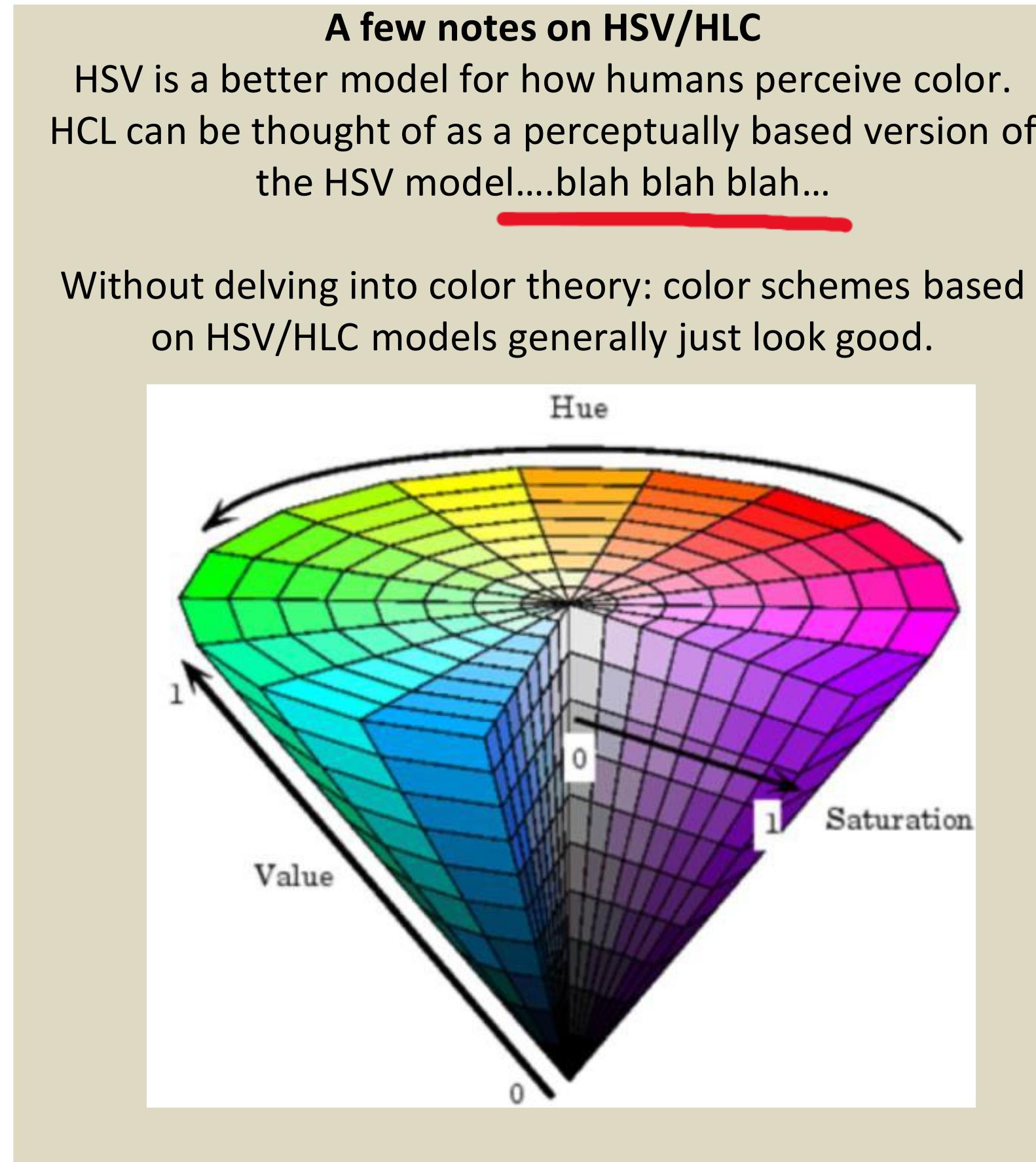
→ Magnitude Channels: Ordered Attributes



→ Identity Channels: Categorical Attributes



R cheatsheet



R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggb). Each character has 16 possible symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F:

"00" can be interpreted as 0.0 and "FF" as 1.0 i.e., red = #FF0000, black = #000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

R has 657 built in color names

Example:
To see a list of names:
`colors()`

peachpuff4

These colors are displayed on P. 3.

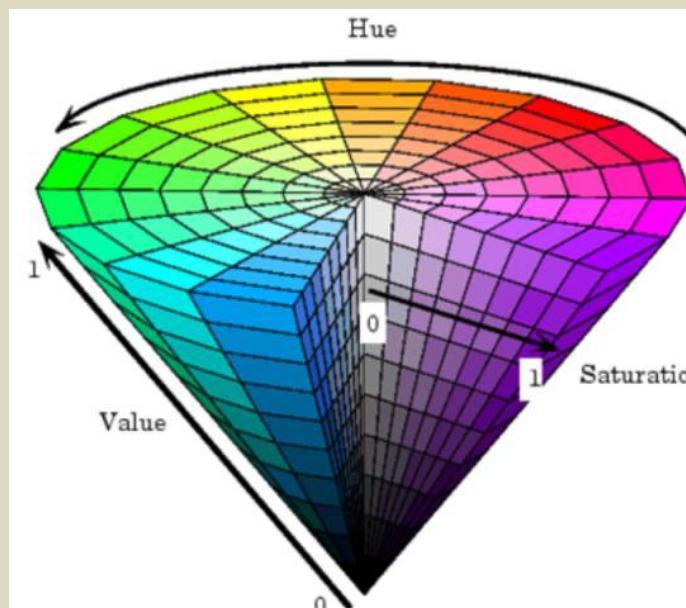
R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using `maxColorValue=255`. `alpha` is an optional argument for transparency, with the same intensity scale.
`rgb(r, g, b, maxColorValue=255, alpha=255)`
- HSV (hue, saturation, value): values range from 0-1, with optional `alpha` argument
`hsv(h, s, v, alpha)`
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
`hcl(h, c, l, alpha)`

A few notes on HSV/HLC

HSV is a better model for how humans perceive color.
HCL can be thought of as a perceptually based version of
the HSV model....blah blah blah...

Without delving into color theory: color schemes based
on HSV/HLC models generally just look good.



R can translate colors to `rgb` (this is handy for matching colors in other programs)
`col2rgb(c("#FF0000", "blue"))`

R Color Palettes

This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure....NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: `grDevices` and `colorRamps`

`grDevices` comes with the base installation and `colorRamps` must be installed. Each palette's function has an argument for the number of colors and transparency (`alpha`):

`heat.colors(4, alpha=1)`

> "#FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFF800FF"

For the `rainbow` palette you can also select start/end color (red = 0, yellow = 1/6, green = 2/6, cyan = 3/6, blue = 4/6 and magenta = 5/6) and saturation (s) and value (v):
`rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)`

`grDevices`
`palettes`
`cm.colors`
`topo.colors`
`terrain.colors`
`heat.colors`
`rainbow`
see P. 4 for options

Package: `RcolorBrewer`

This function has an argument for the number of colors and the color palette (see P. 4 for options).
`brewer.pal(4, "Set3")`

> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: `display.brewer.all(5)`
There is also a very nice interactive viewer:
<http://colorbrewer2.org/>

My Recommendation

Package: `colorspace`

These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

`rainbow_hcl(4)`

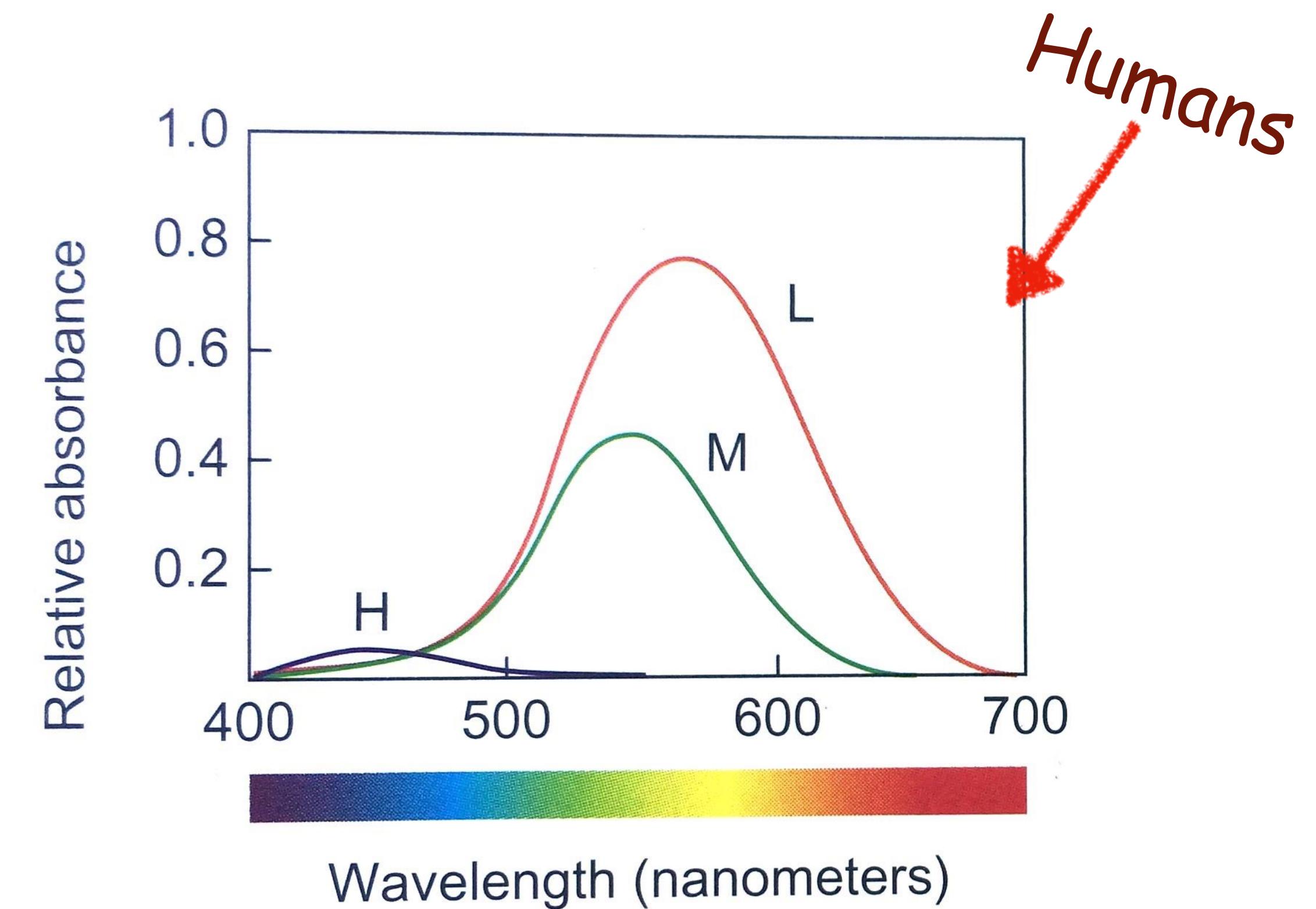
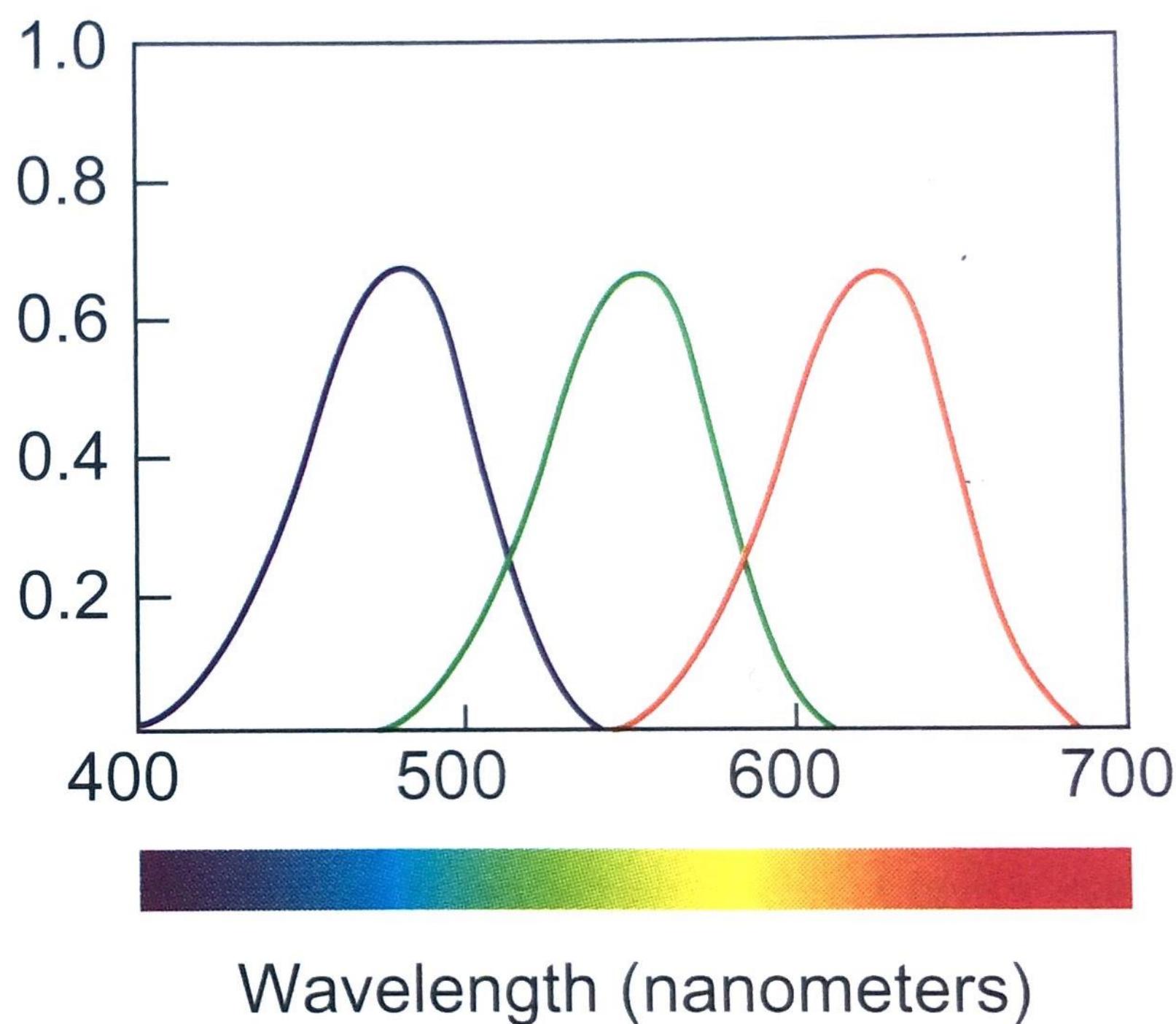
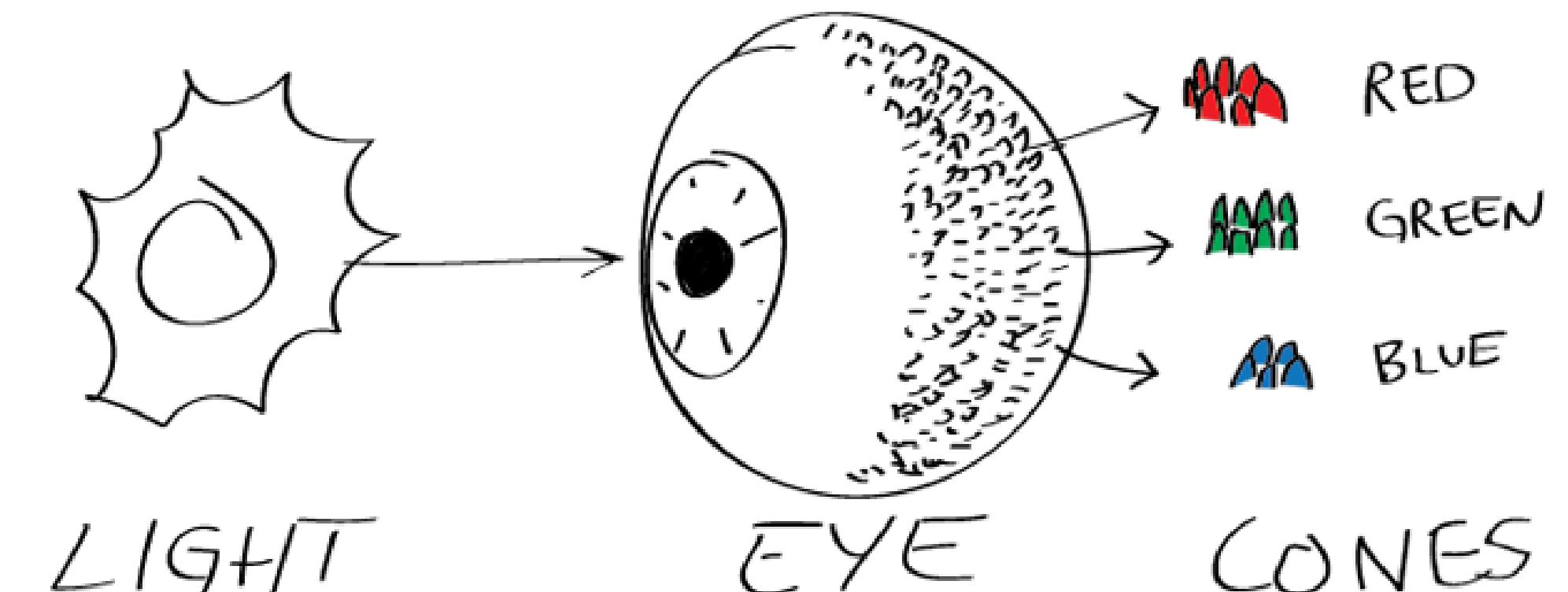
> "#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"

`colorspace`
`default_palettes`
`diverge_hcl`
`diverge_hsl`
`terrain_hcl`
`sequential_hcl`
`rainbow_hcl`

However, all palettes are fully customizable:
`diverge_hcl(7, h = c(246, 40), c = 96, l = c(65, 90))`
Choosing the values would be daunting. But there are some recommended palettes in the `colorspace` documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
`pal <- choose_palette()`

Percepción

- Tenemos tres tipos de conos que reaccionan a espectros distintos y de forma desigual
- No son como receptores de una cámara, frecuencias más largas generan estímulos más fuertes

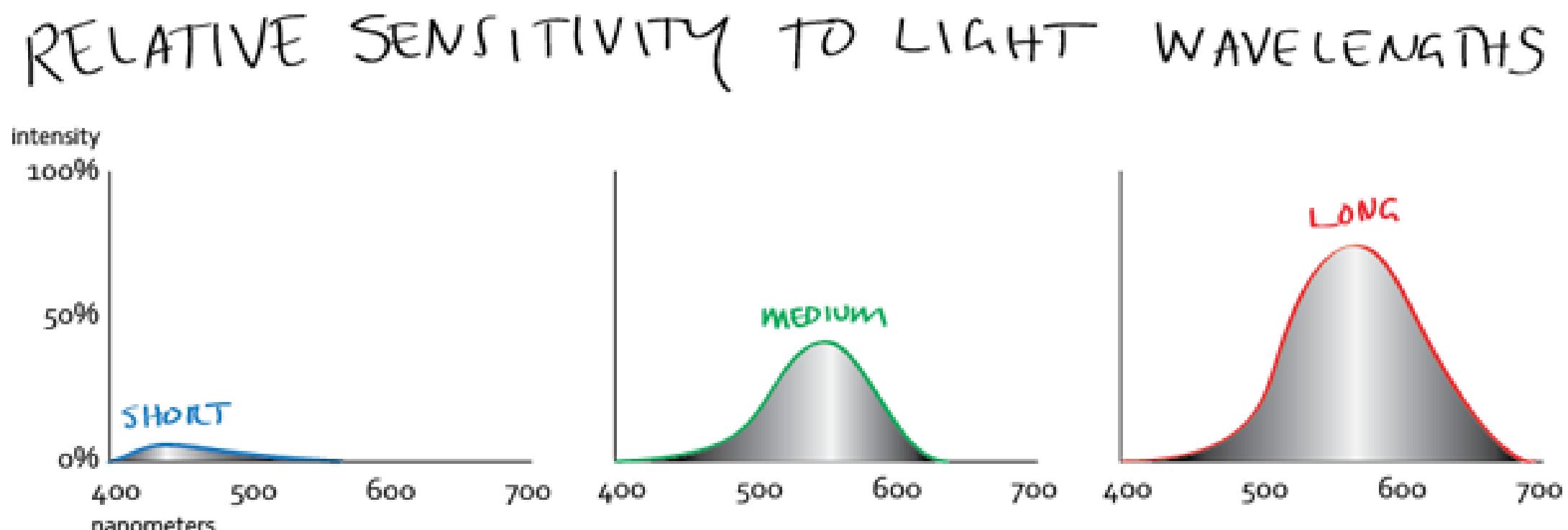
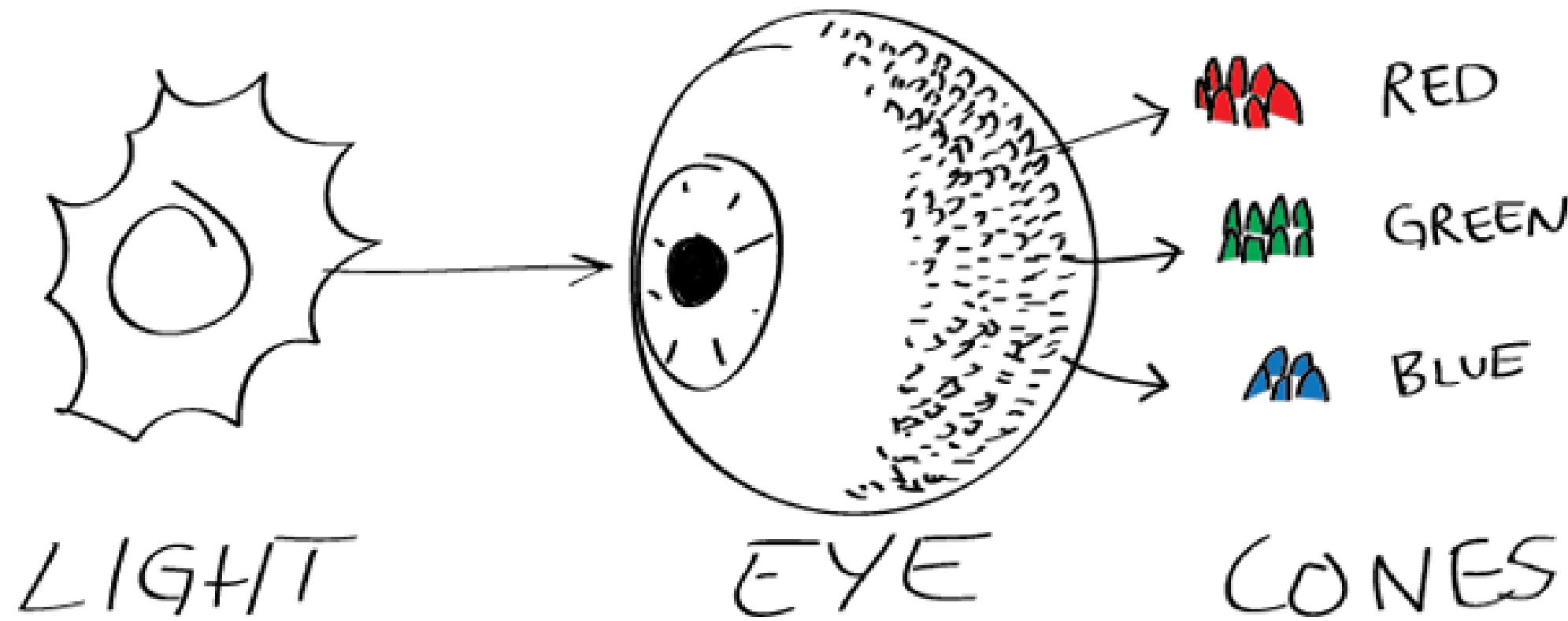


Opponent Process

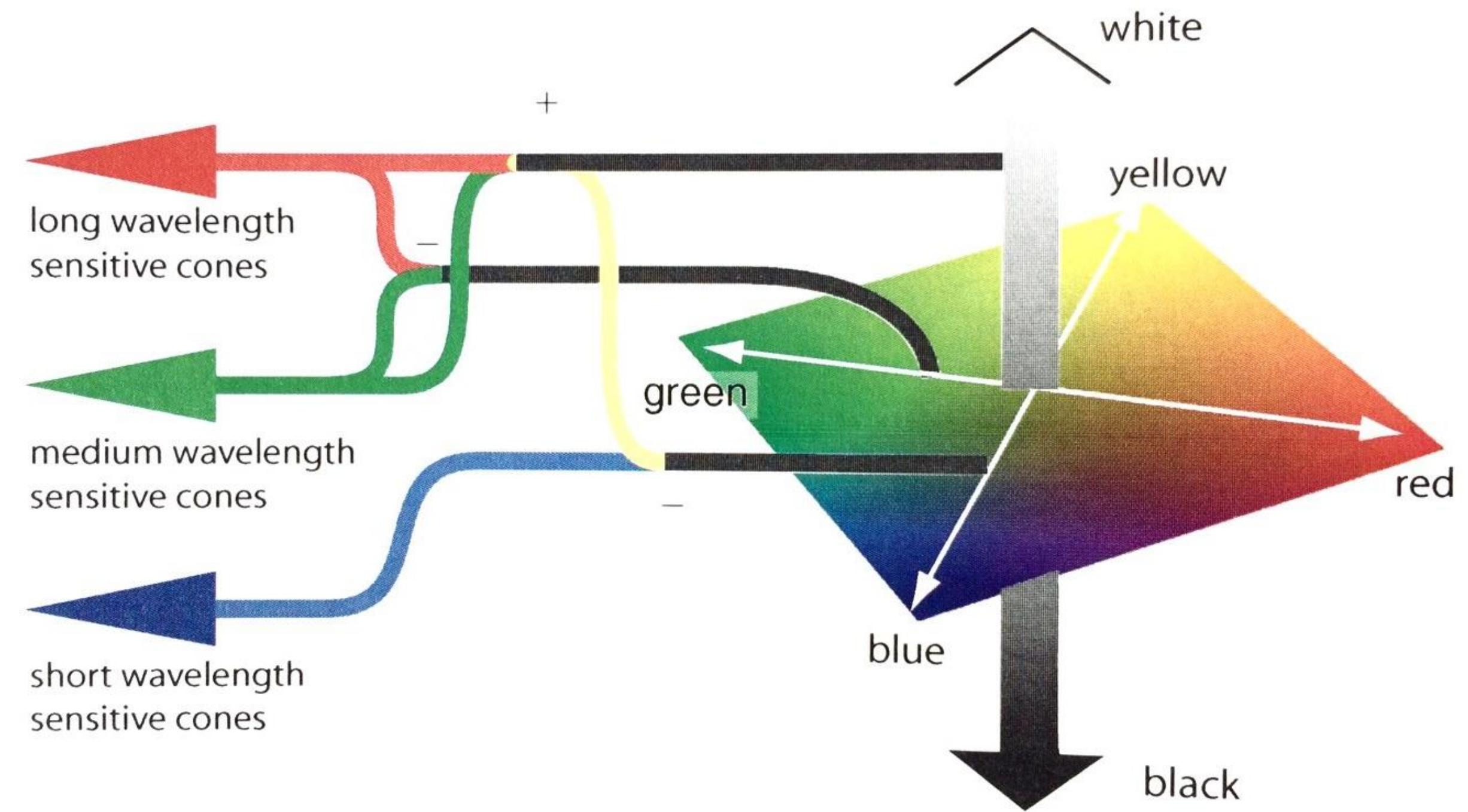
Theory of Color

Ewald Hering, 1920

- The brain combines the signals by subtraction: receptors subtract med. and long freq. making a red-green-difference signal channel.
- Other neurones subtract long and short yielding yellow-blue-diff signal channel;
- Third group of neurones ADD long and medium to create luminance (B&W) channel



difference signal channels



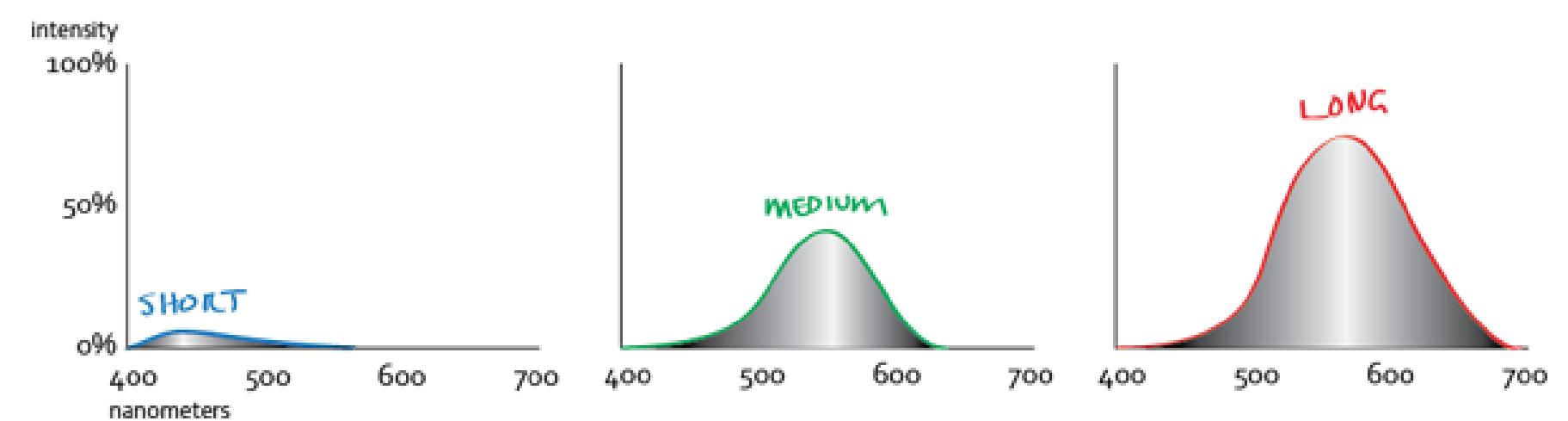
Ware, 2008

Opponent Process

Theory of Color

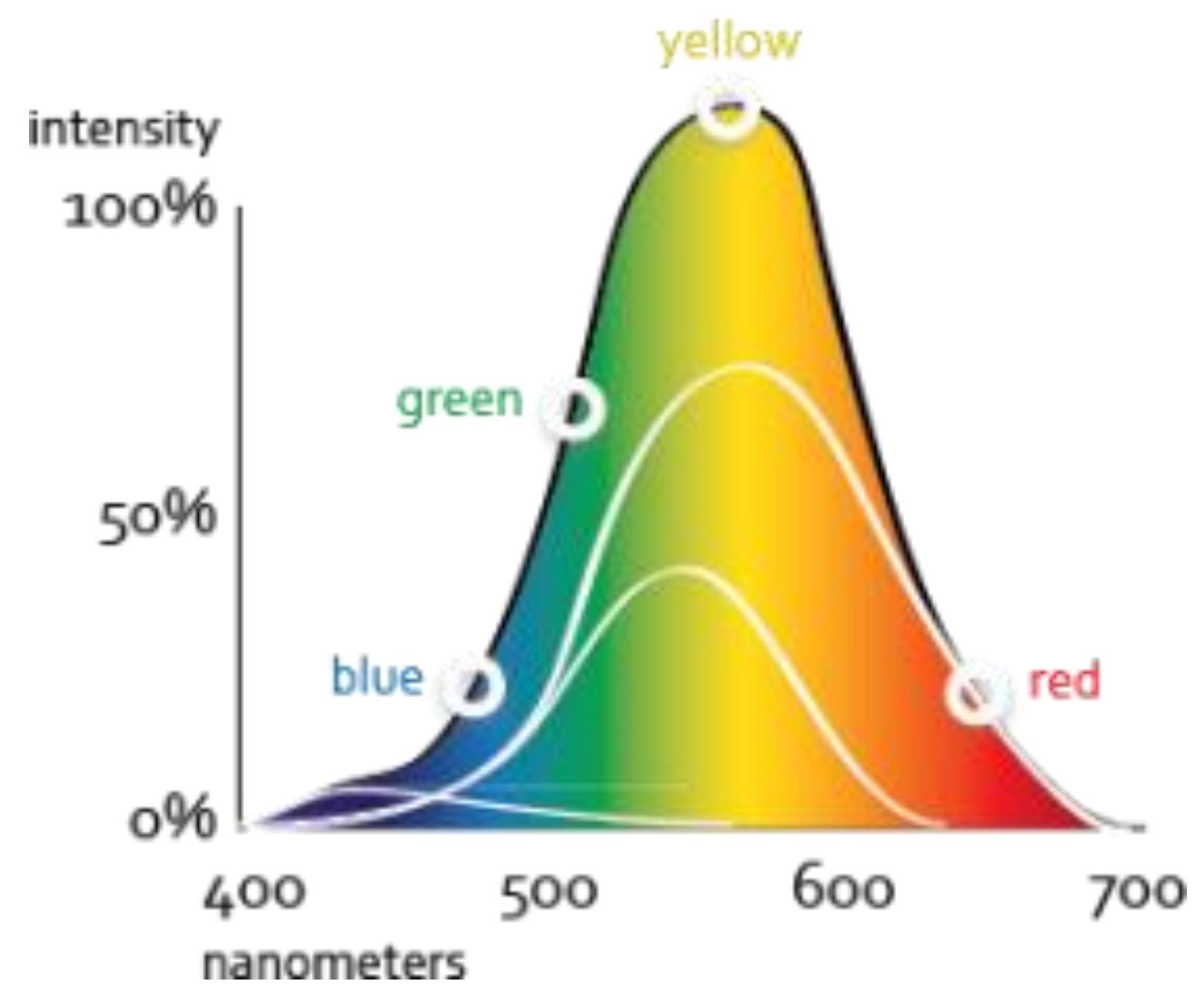
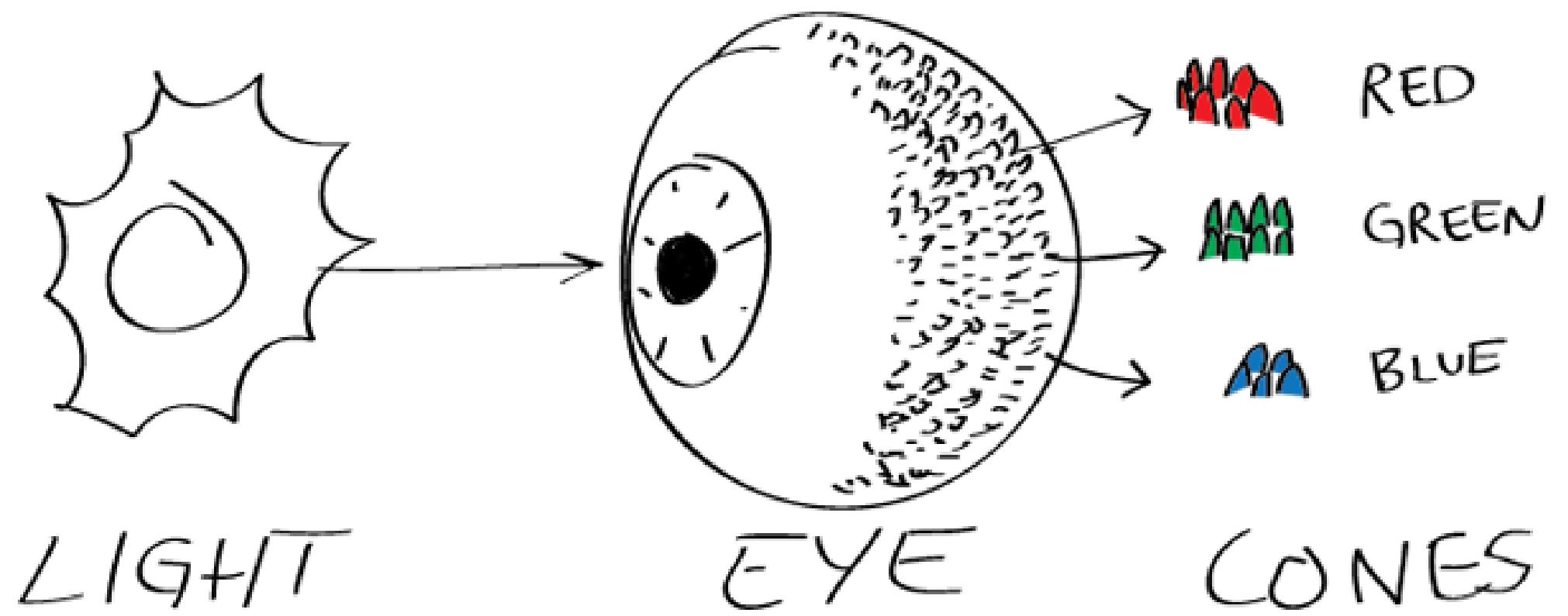
Ewald Hering, 1920

RELATIVE SENSITIVITY TO LIGHT WAVELENGTHS



- Cuando vemos Amarillo hay una sobre excitación de muchos receptores que hacen que lo percibamos más intensamente que otros colores

PUTTING IT ALL TOGETHER

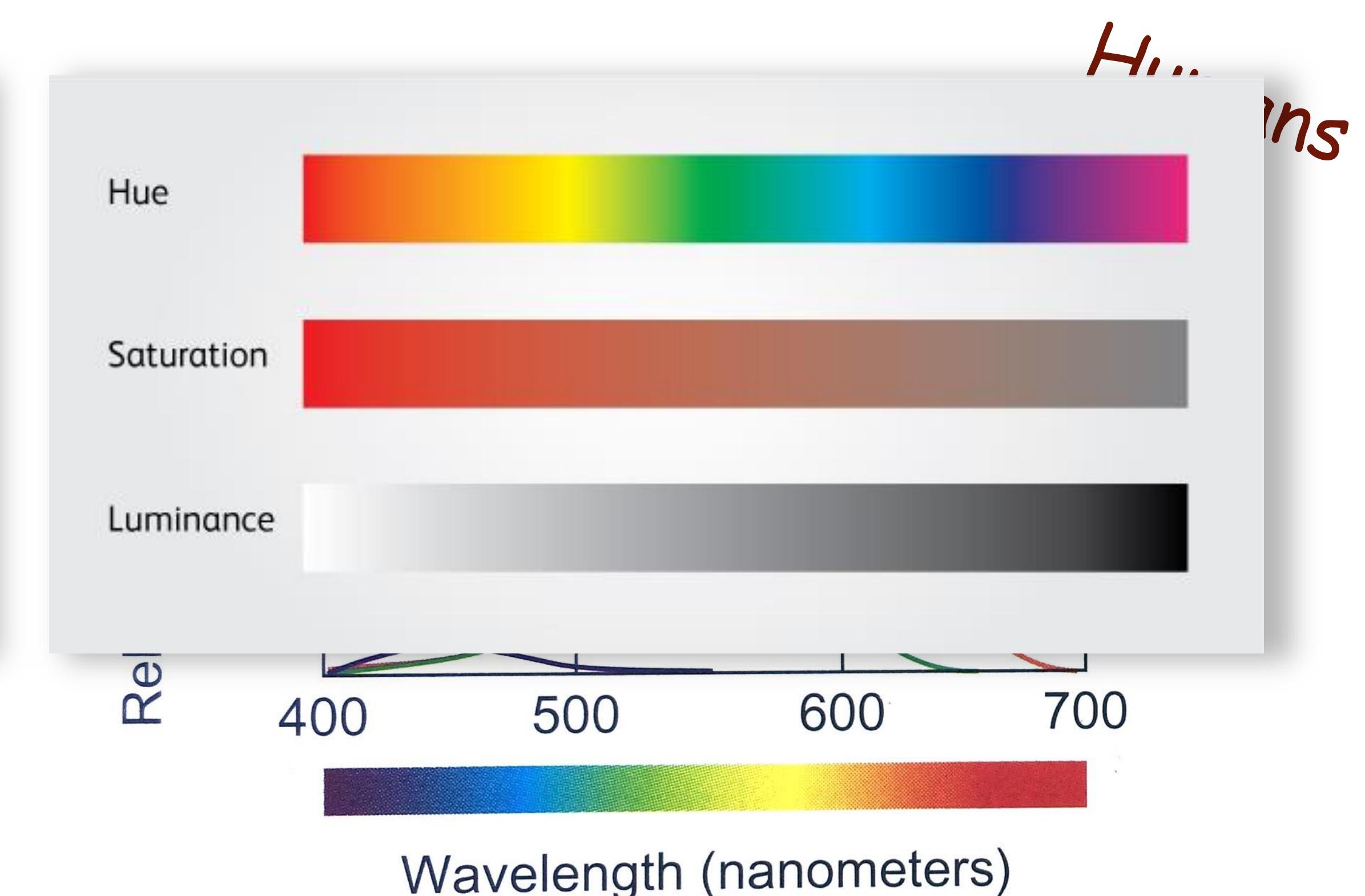
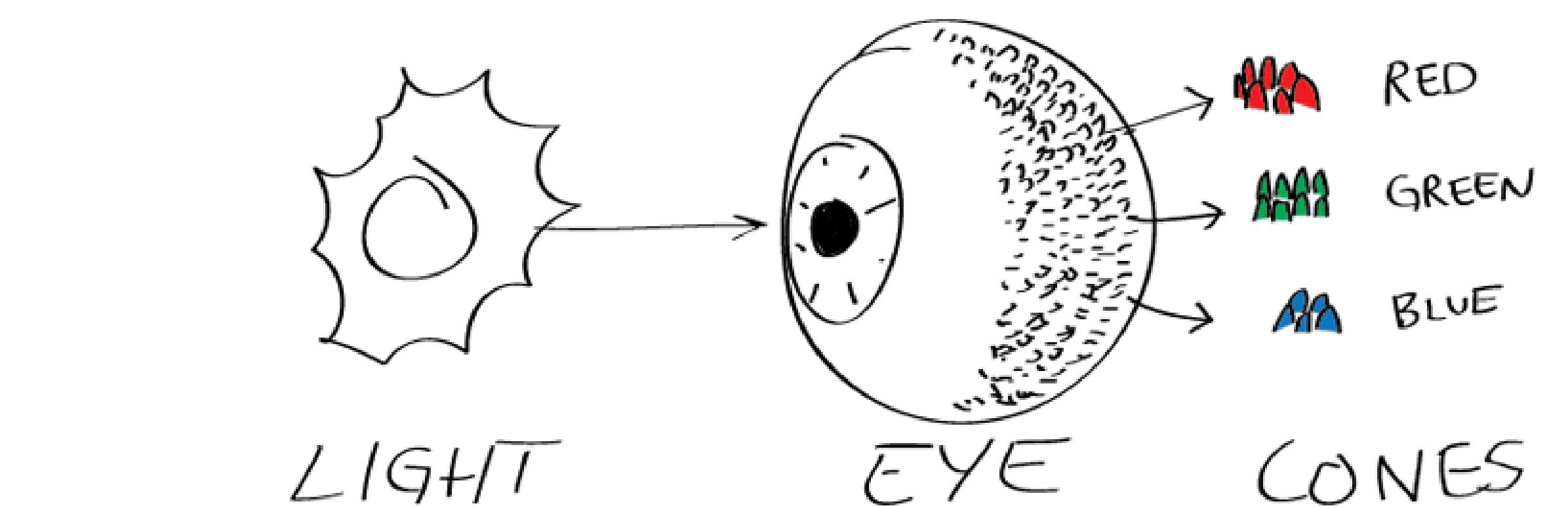
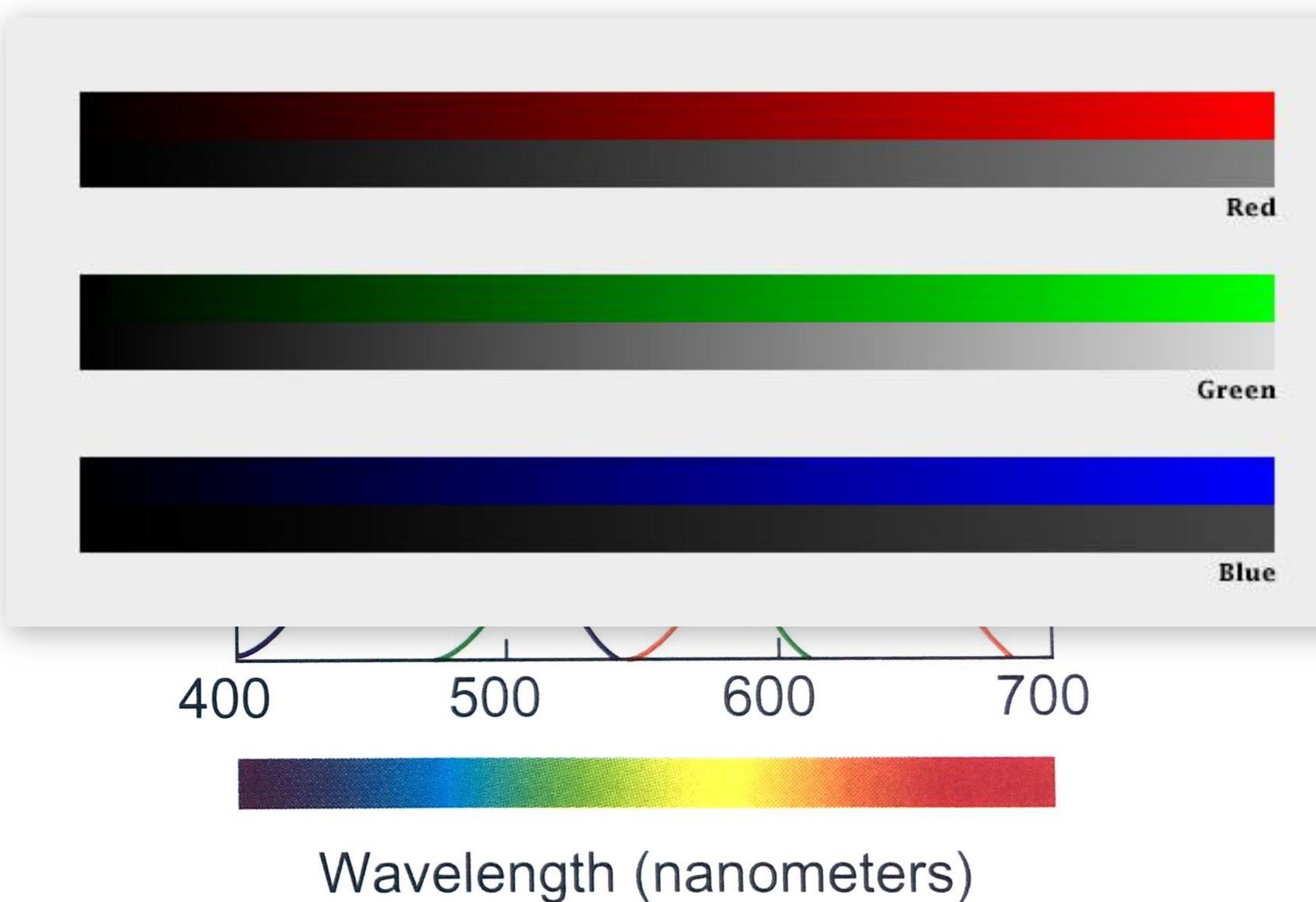


Percepción

Computers process a combination of very narrow frequency bands of red, green, and blue

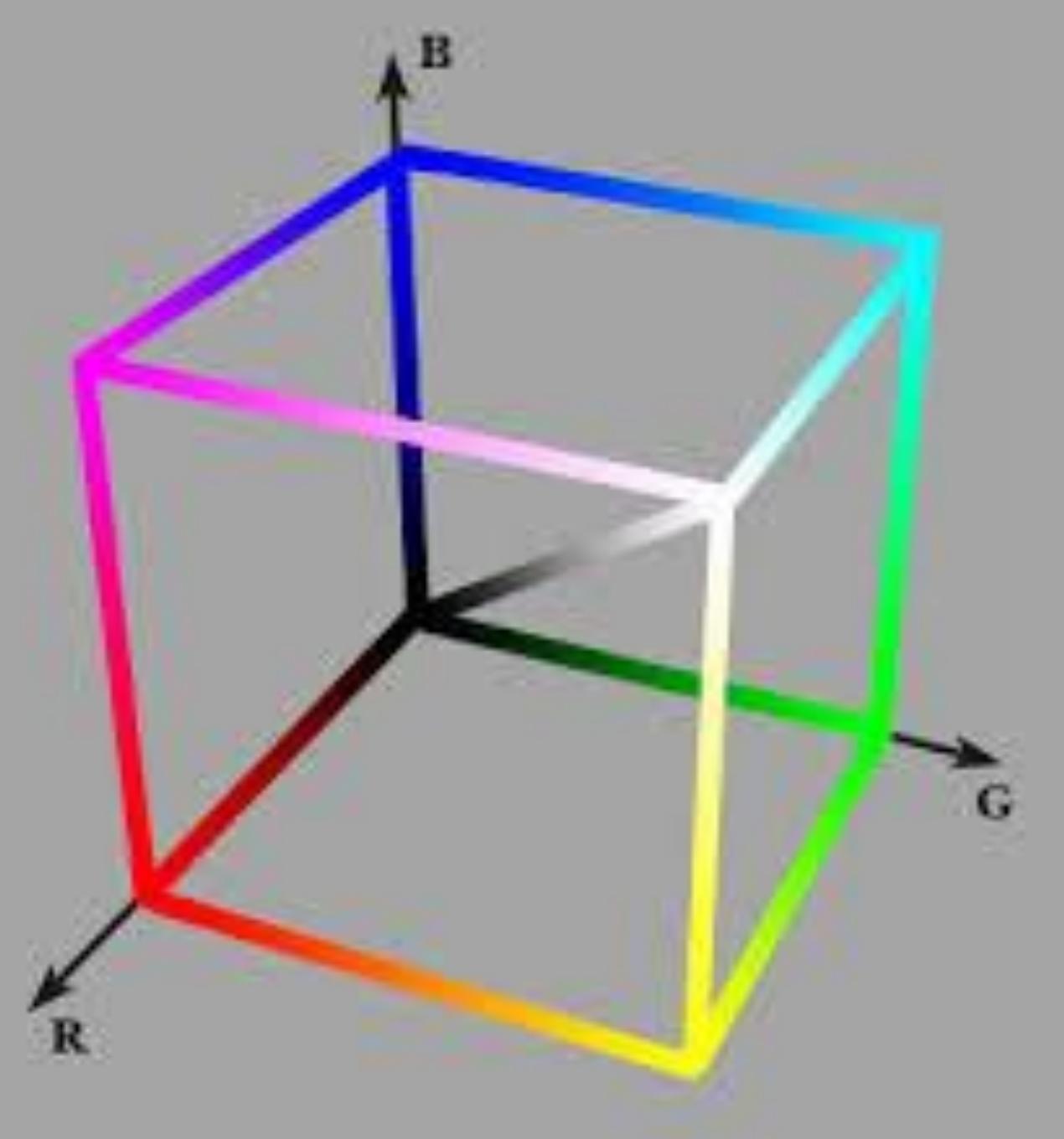
Computers calculate light linearly

Our eyes detect red, green, and blue light but we think about colour in terms of lightness, hue, and saturation

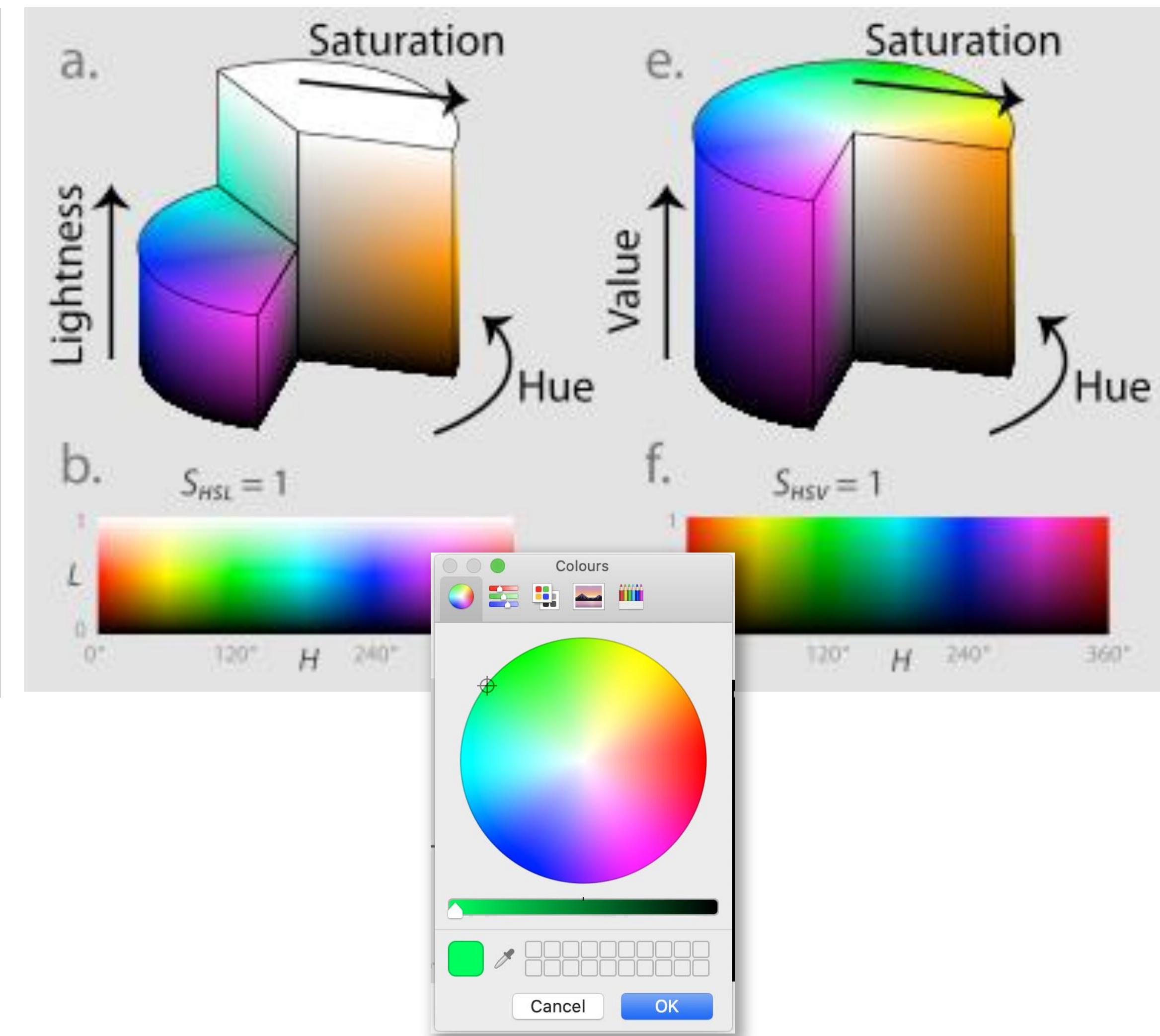


Color spaces

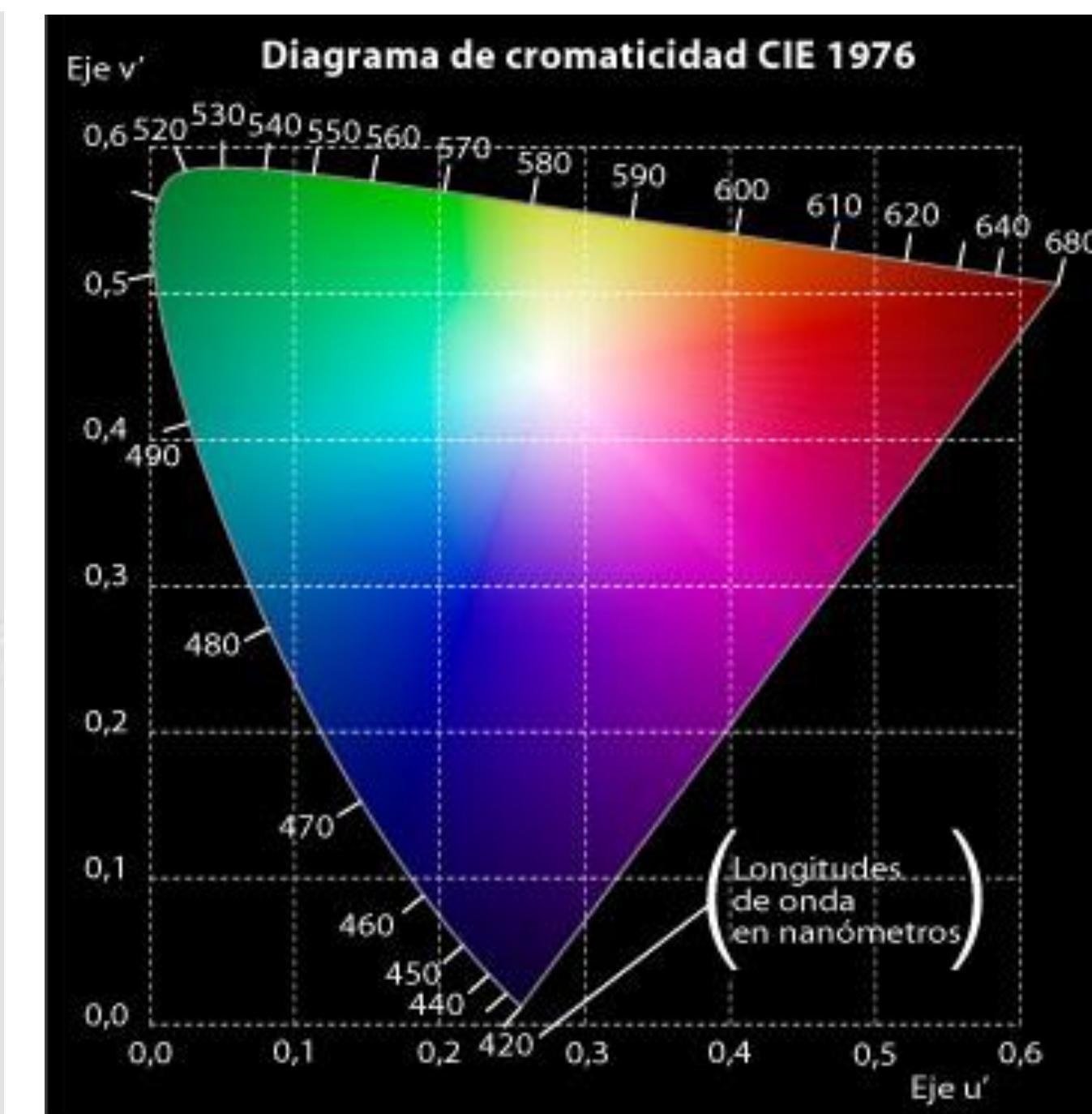
RGB



HSL



HSV



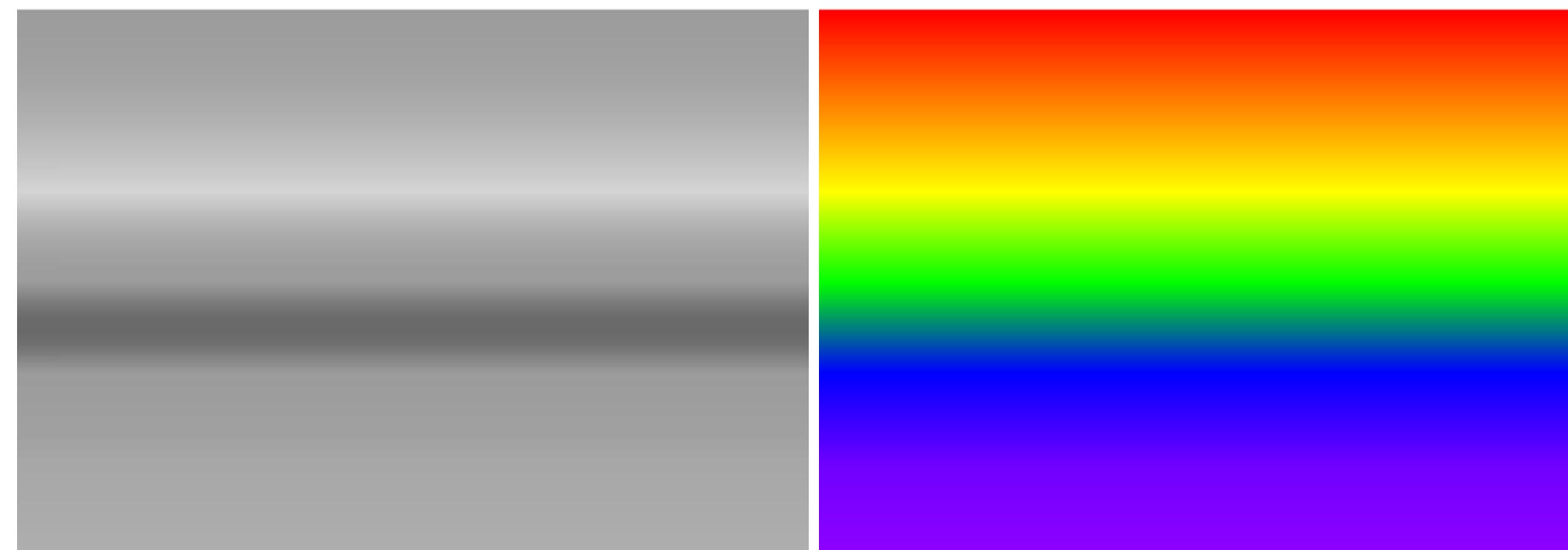
CIEL

Escala arcoíris

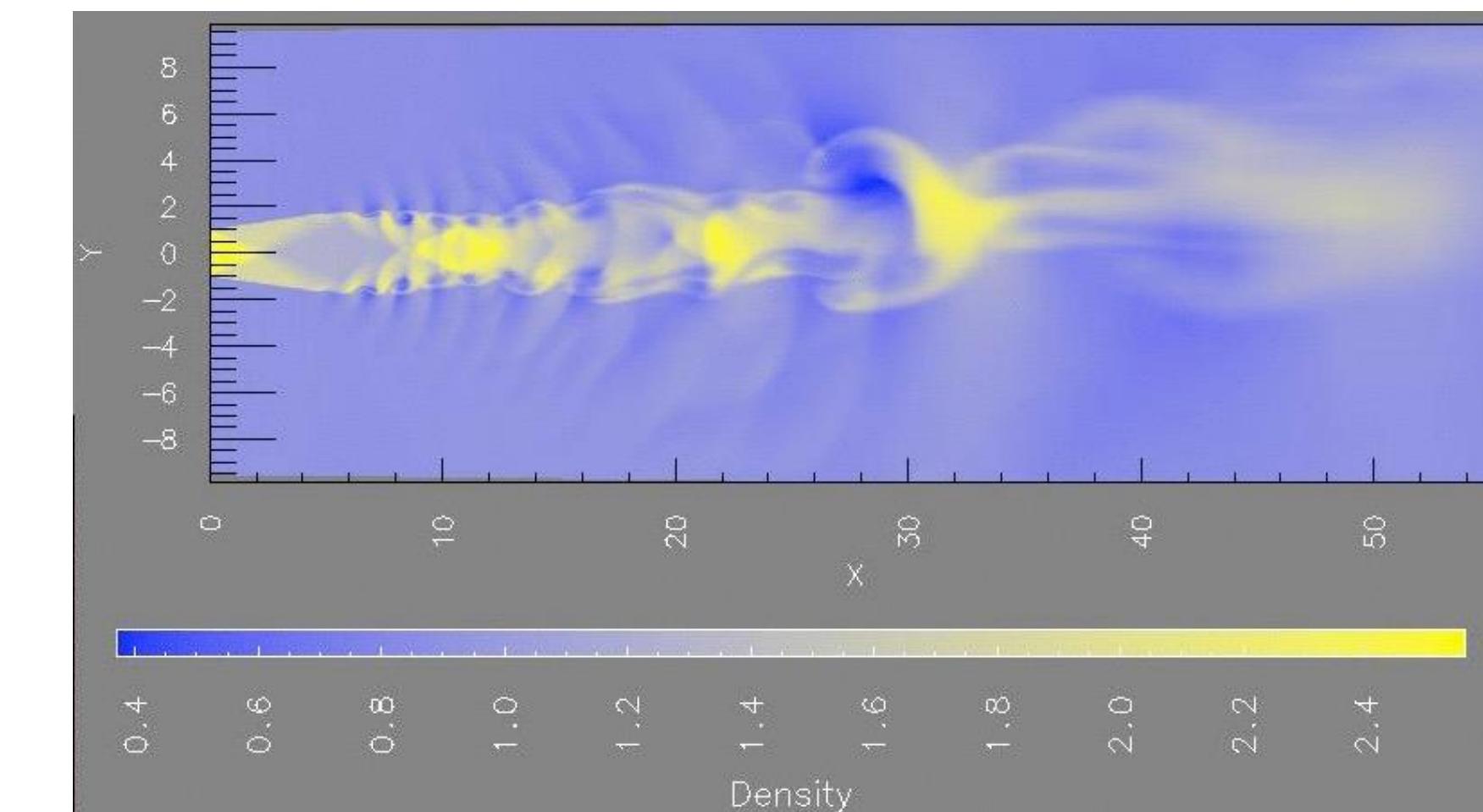
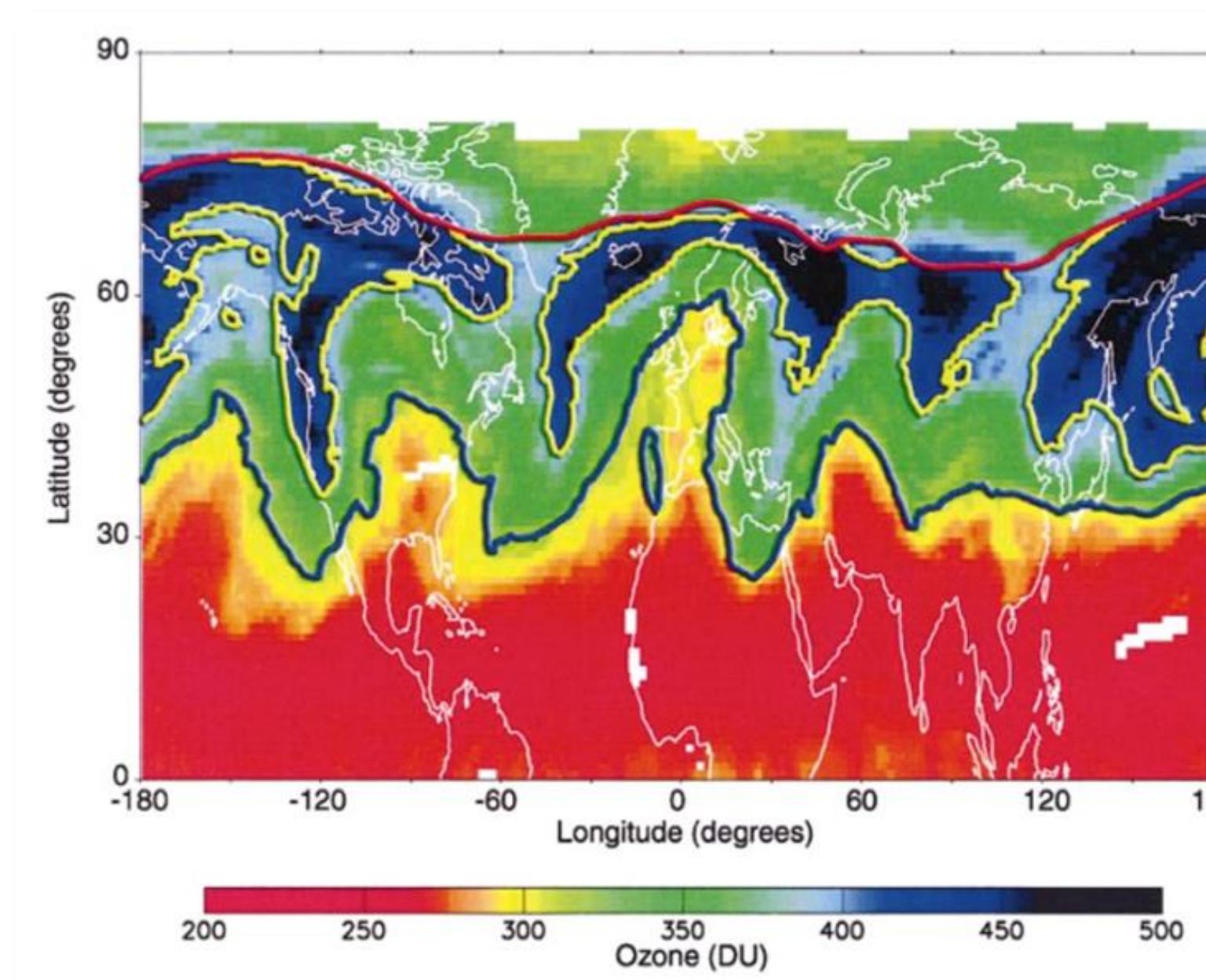
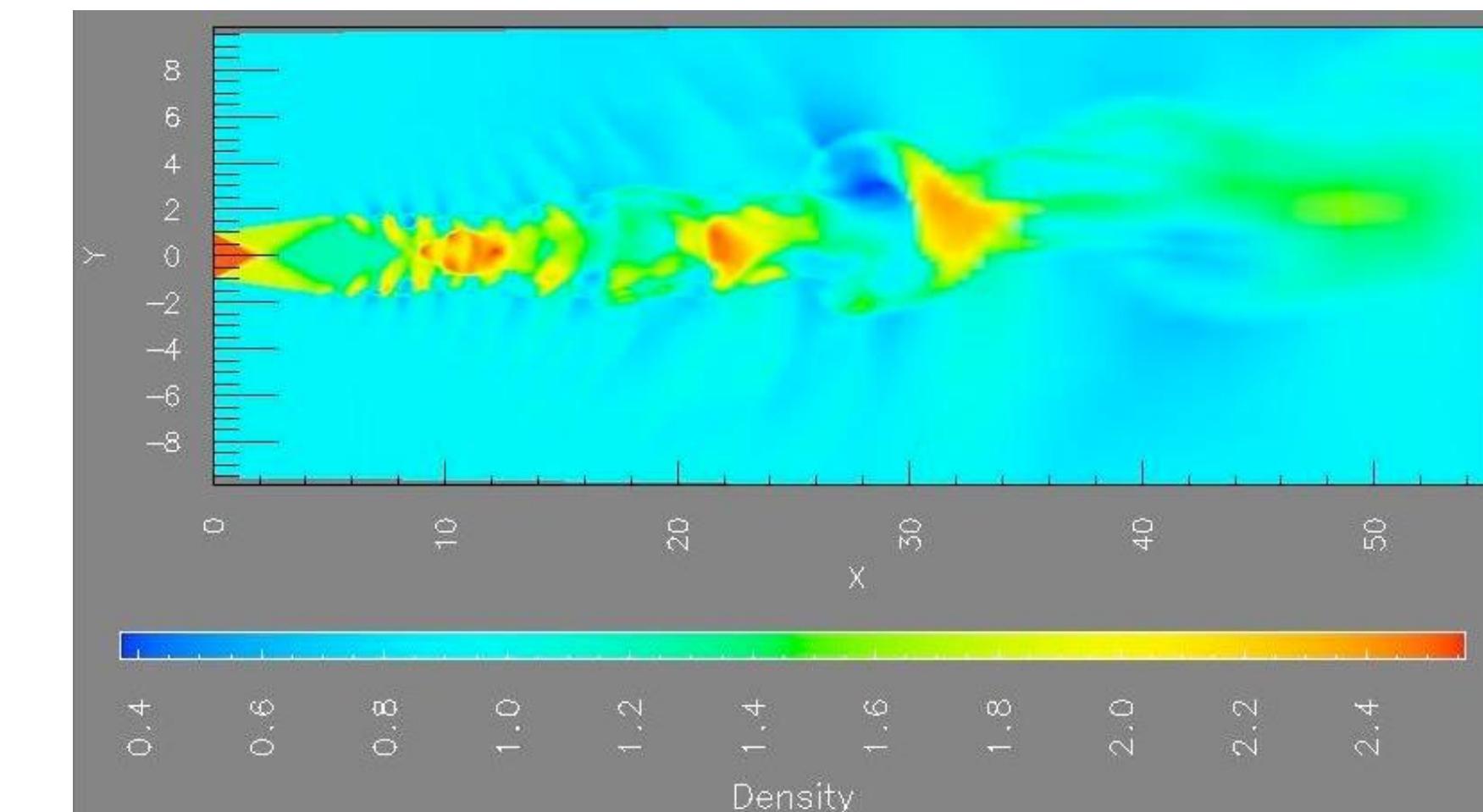
Default en varios softwares de visualización



- Los tonos no tienen un orden inherente



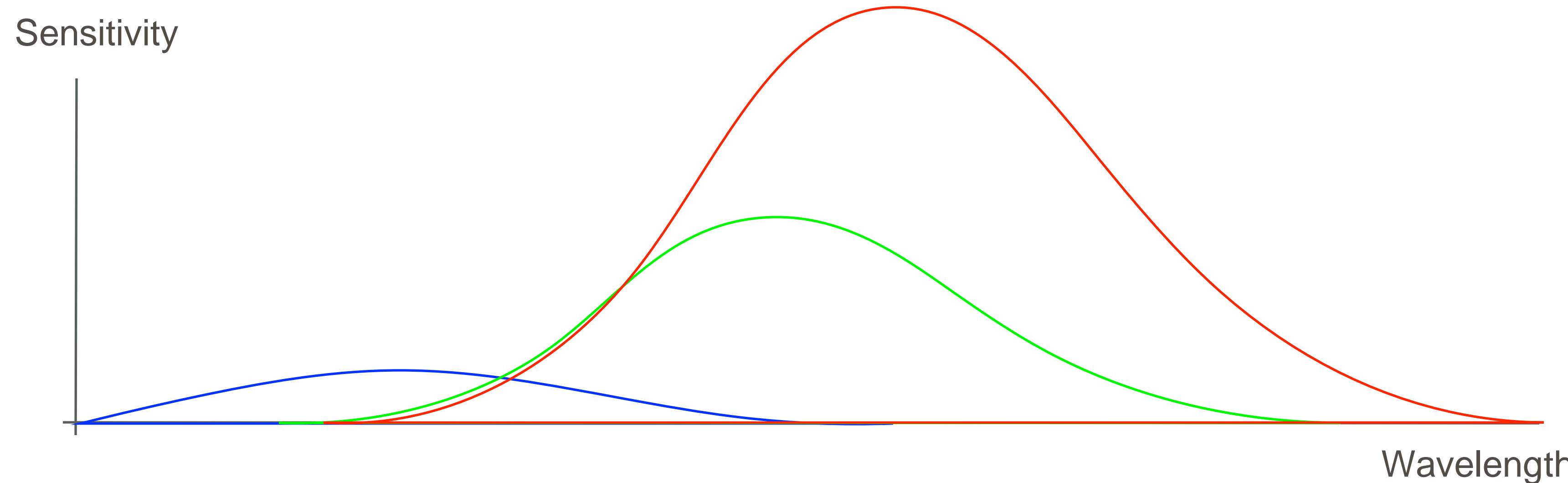
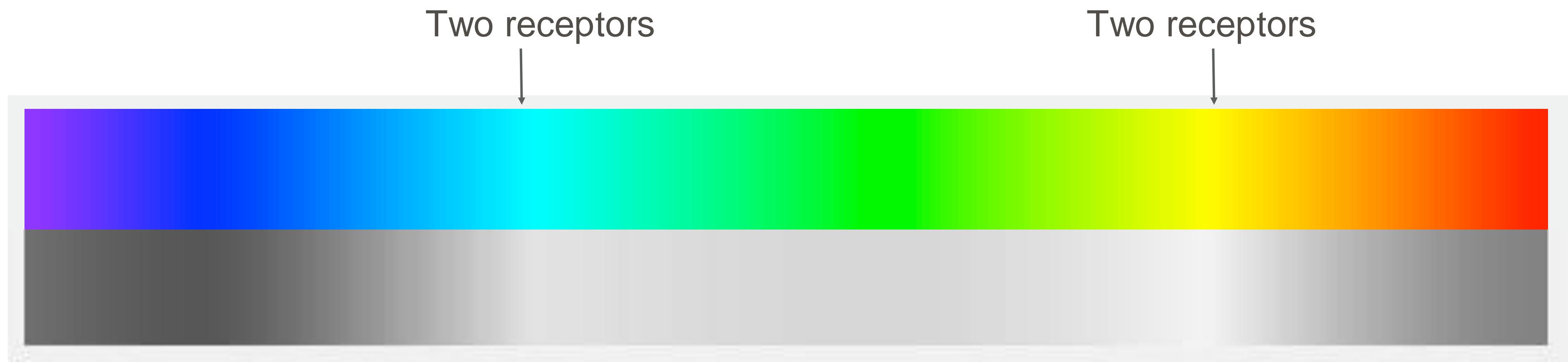
- Es más difícil ver algunos detalles
- Aparecen boundaries que no existen en realidad



[A Rule-based Tool for Assisting Colormap Selection. Bergman,., Rogowitz, and. Treinish. Proc. IEEE Visualization (Vis), pp. 118–125, 1995.]

Problemas de la escala arcoíris, variación no-lineal y la dificultad para distinguir regiones:

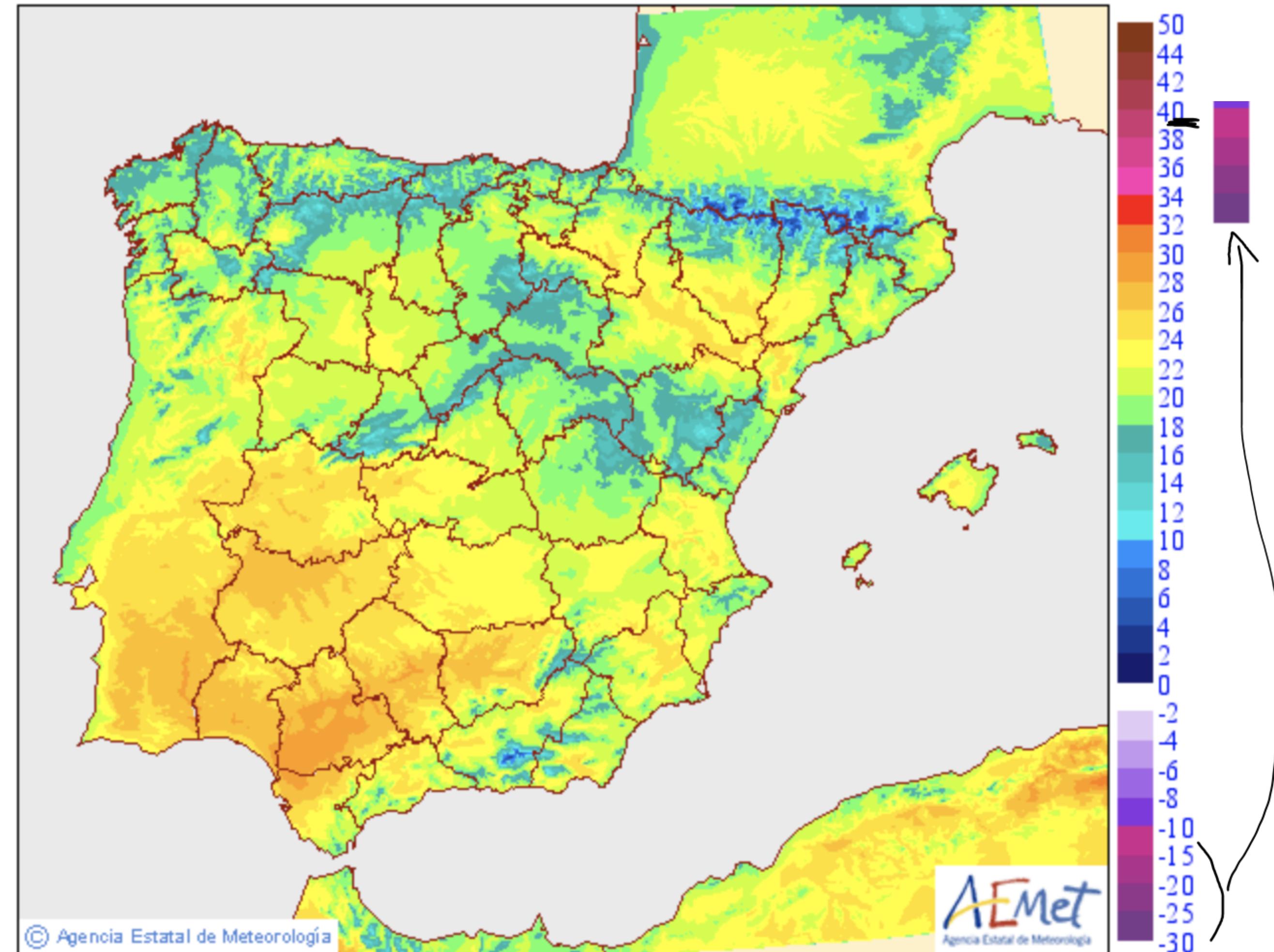
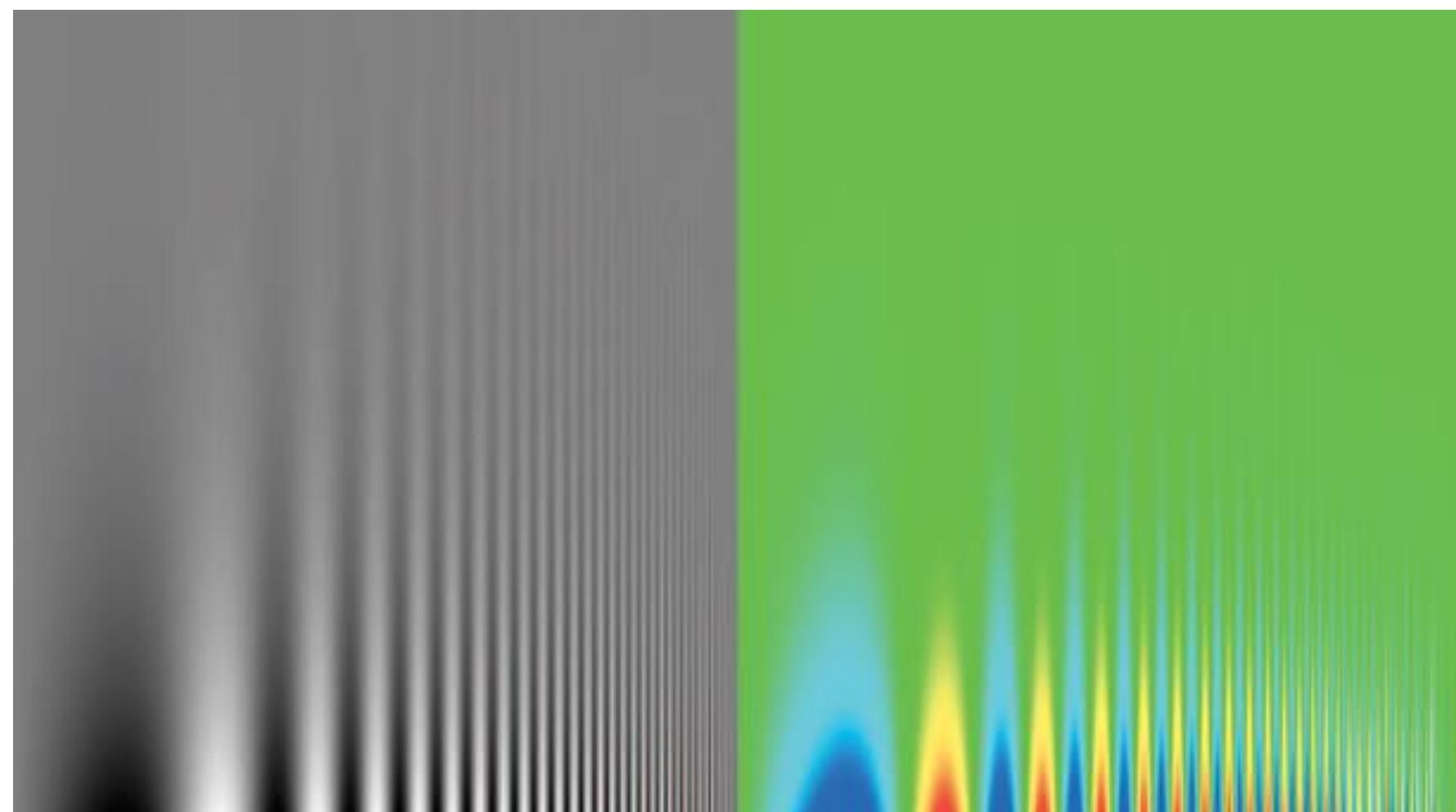
- Sobreexcitación de receptores en algunos puntos, que rompen la linealidad de la escala.



Escala arcoíris

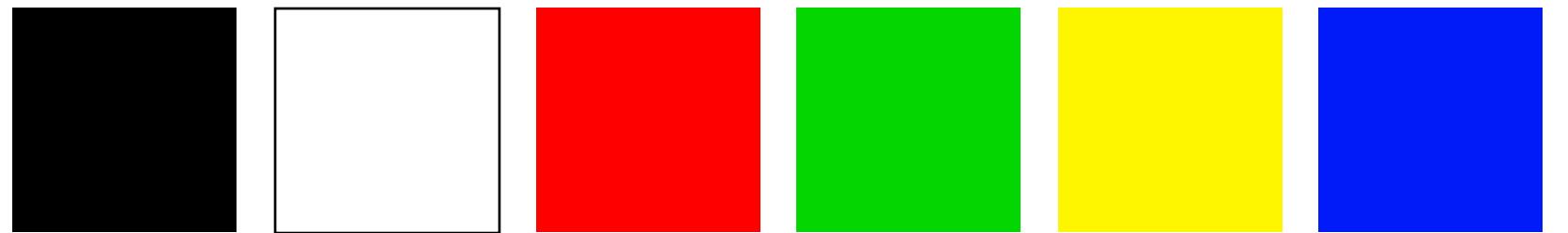
No tiene en cuenta cómo percibimos el color

- No tiene un orden inherente
- Hay colores que percibimos más intensamente que otros
- “Bandas” perceptuales
- Pérdida de detalle

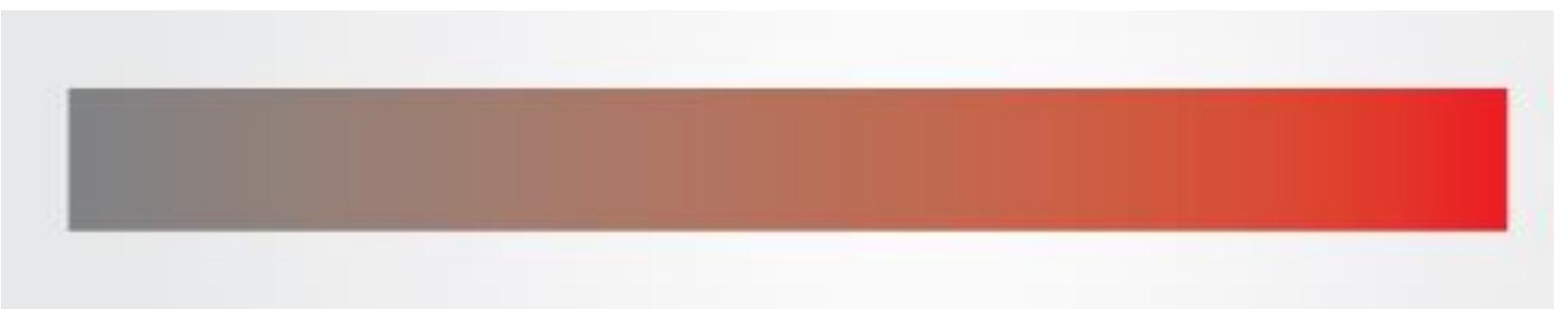


Channel Properties

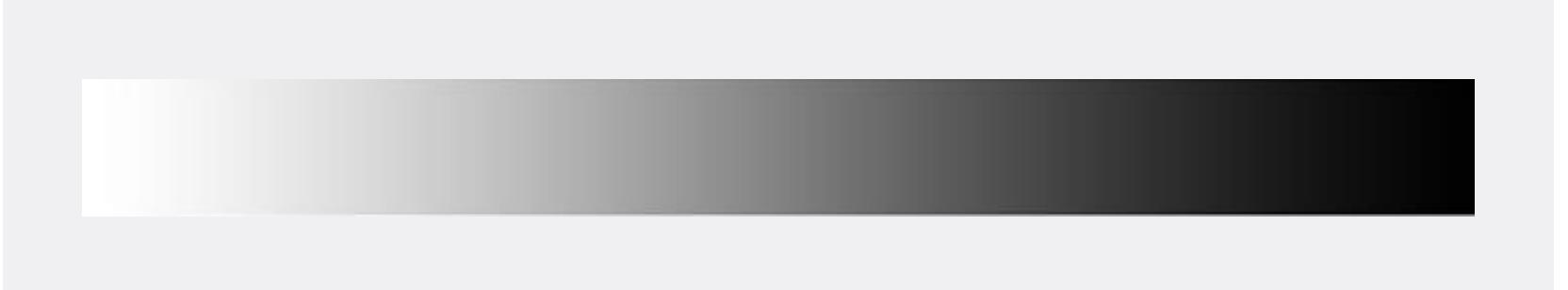
Tono (Hue)



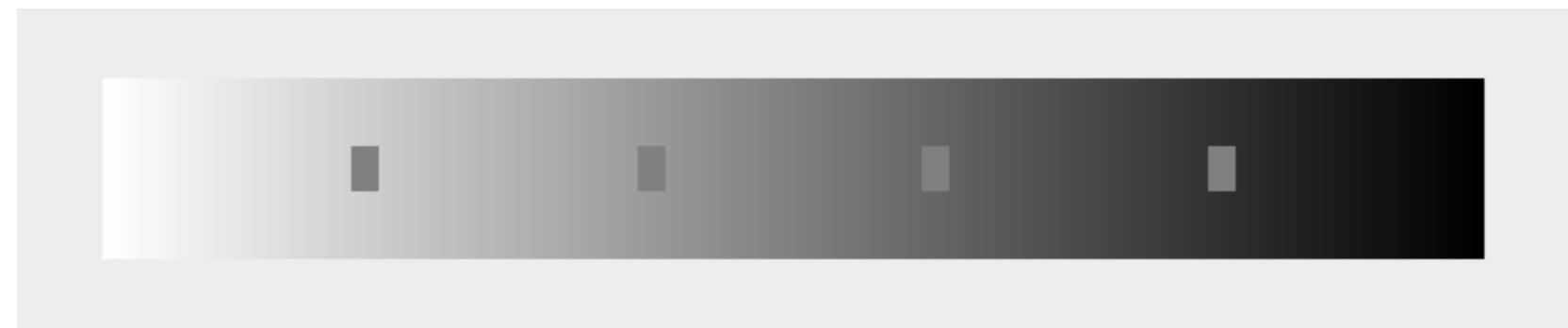
Saturation



Luminance/Lightness/Brightness



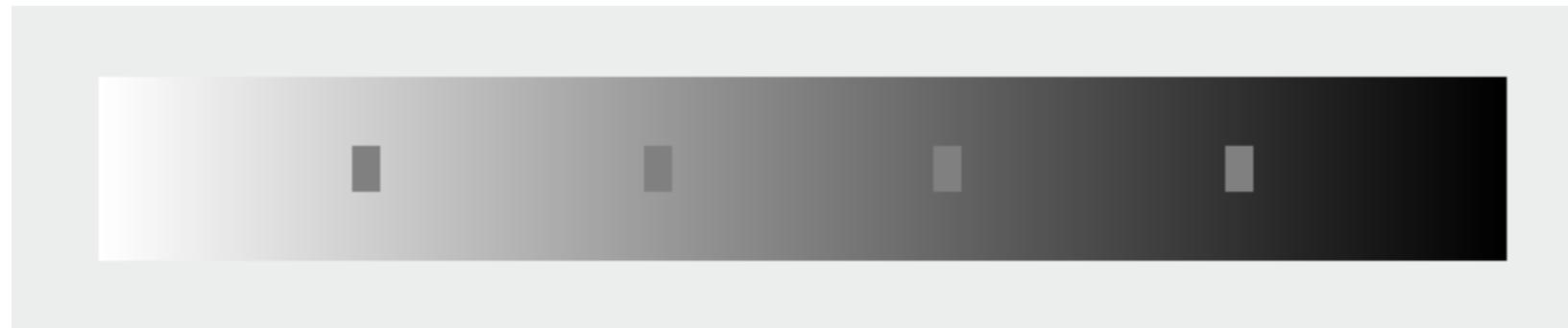
Contrast



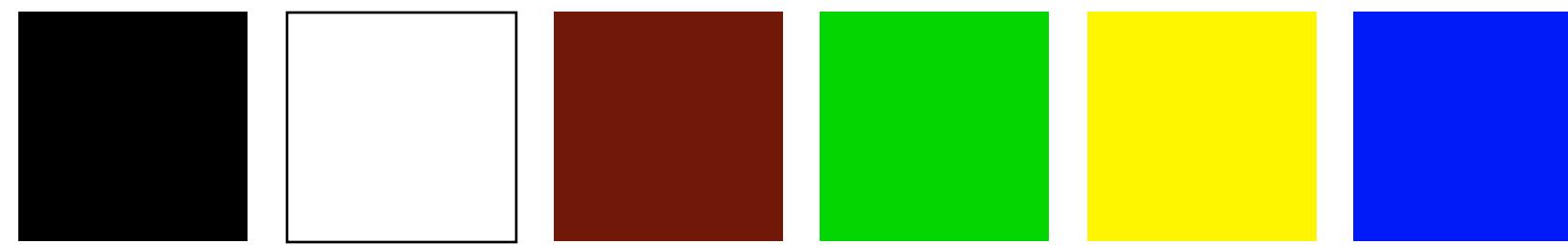


Channel Properties

Contrast



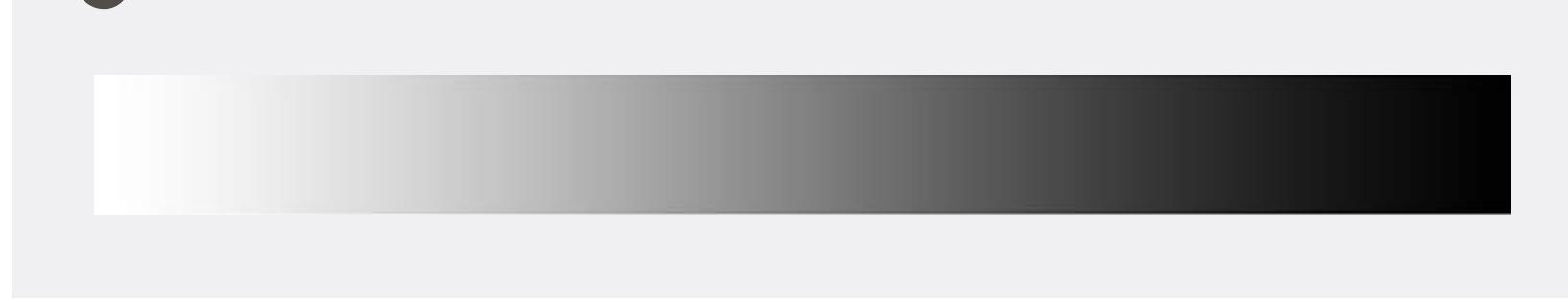
Unique hues



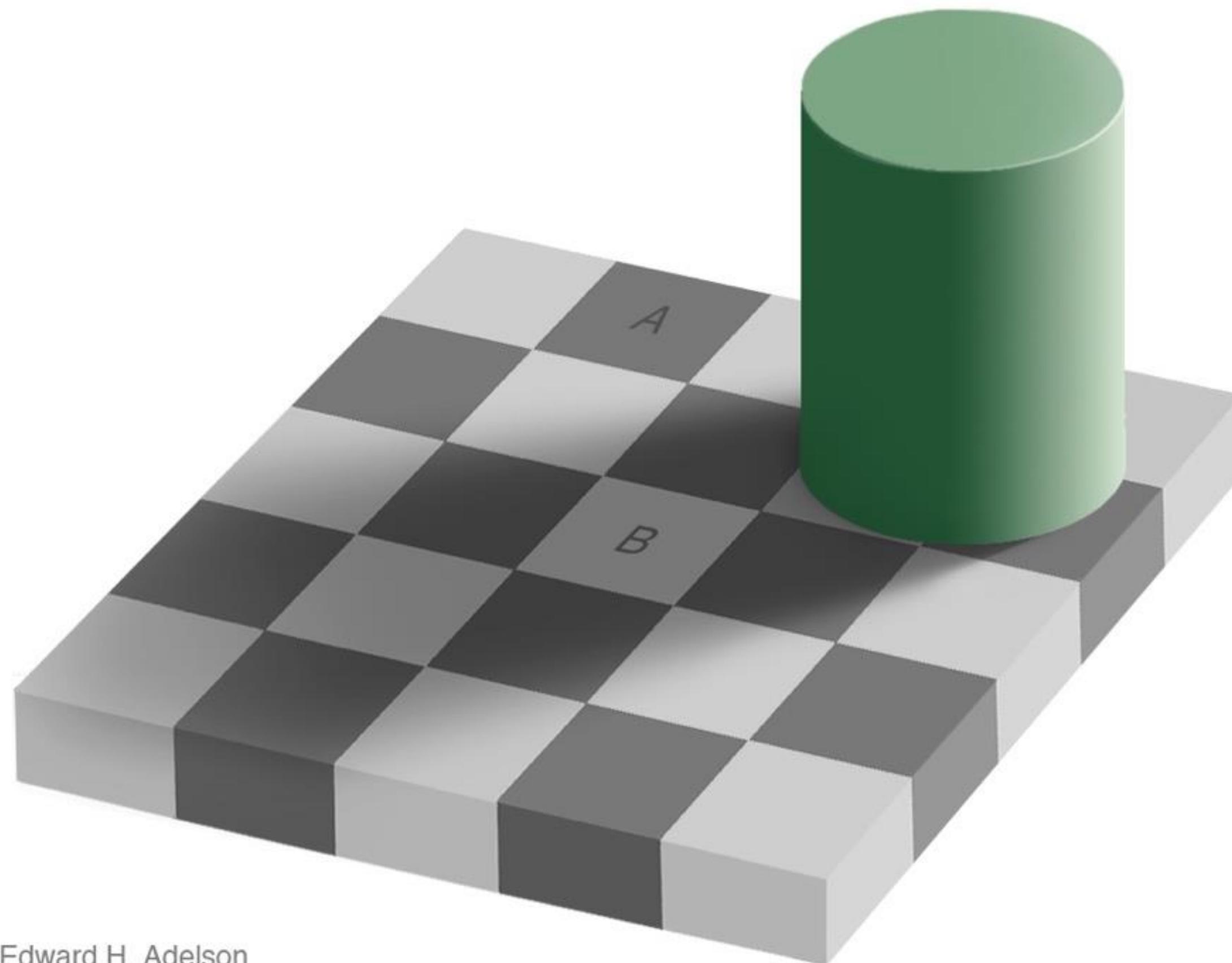
Saturation



Lightness



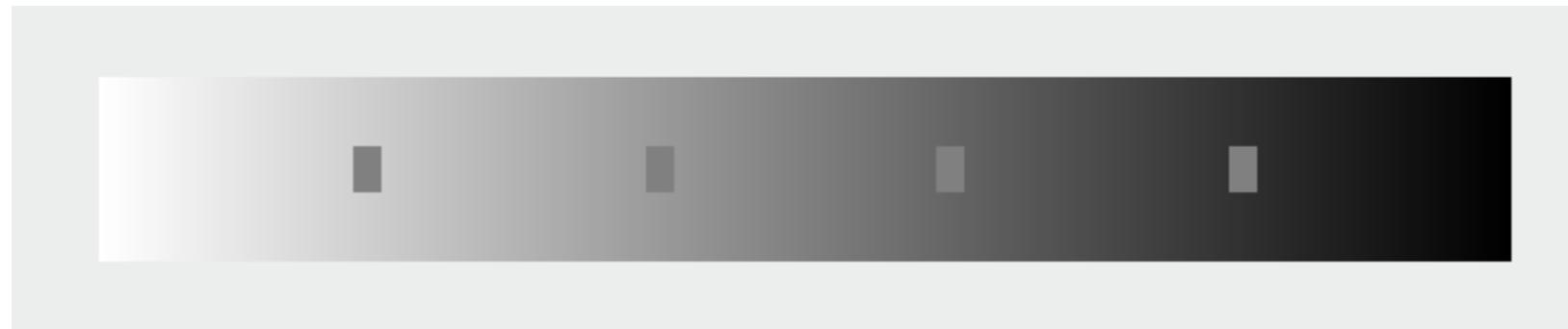
Simultaneous Contrast



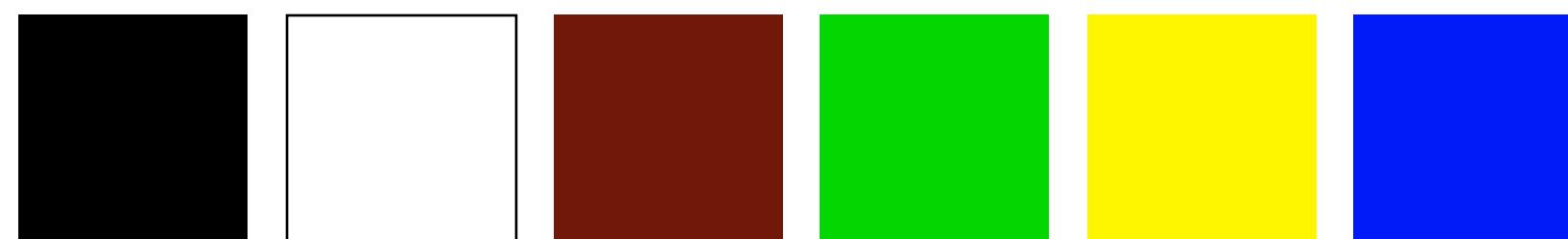
Edward H. Adelson

Channel Properties

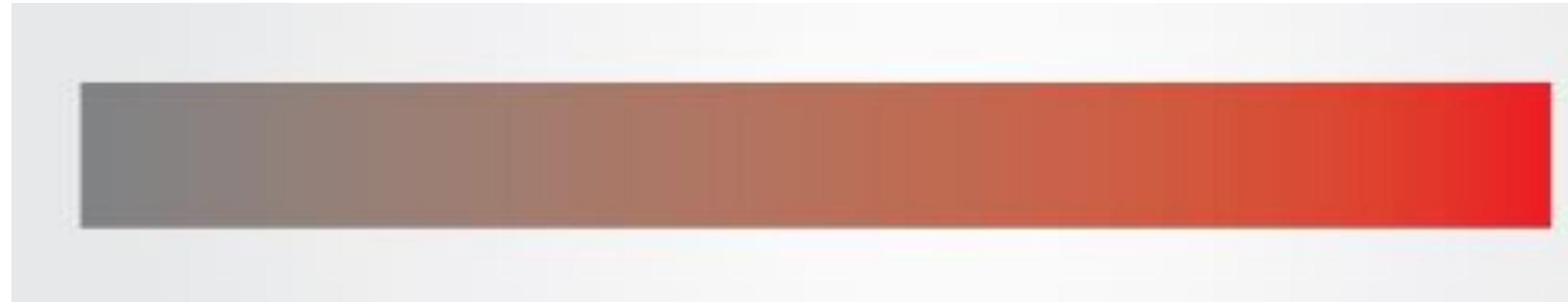
Contrast



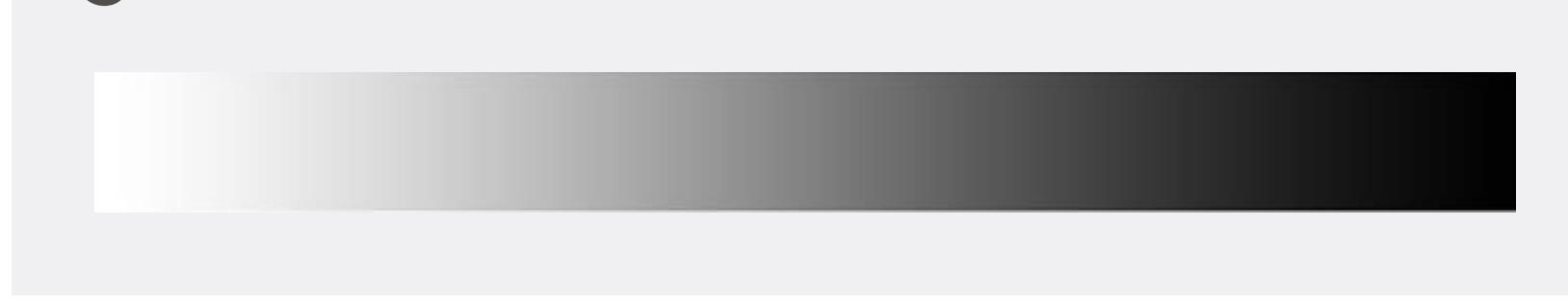
Unique hues



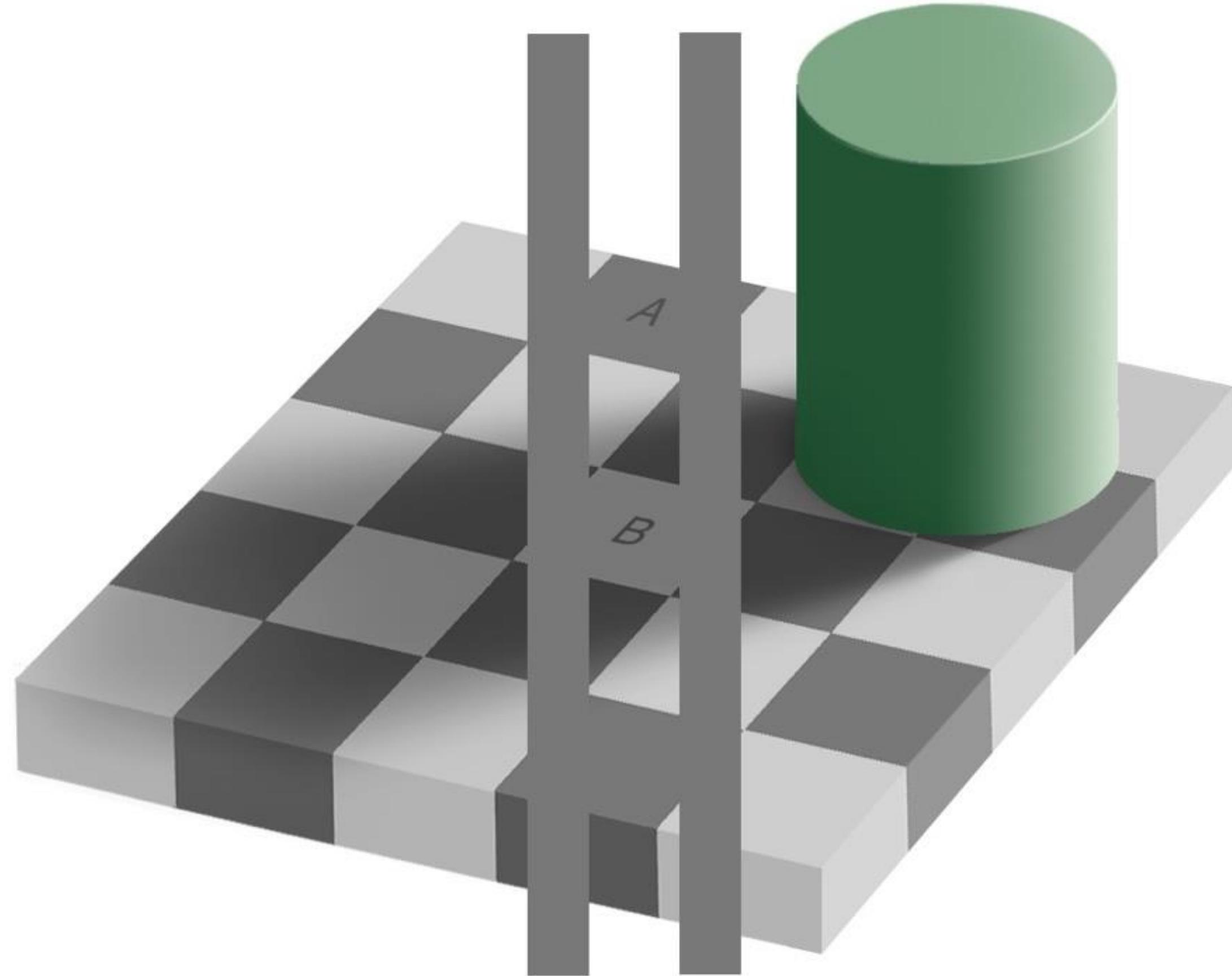
Saturation



Lightness



Simultaneous Contrast

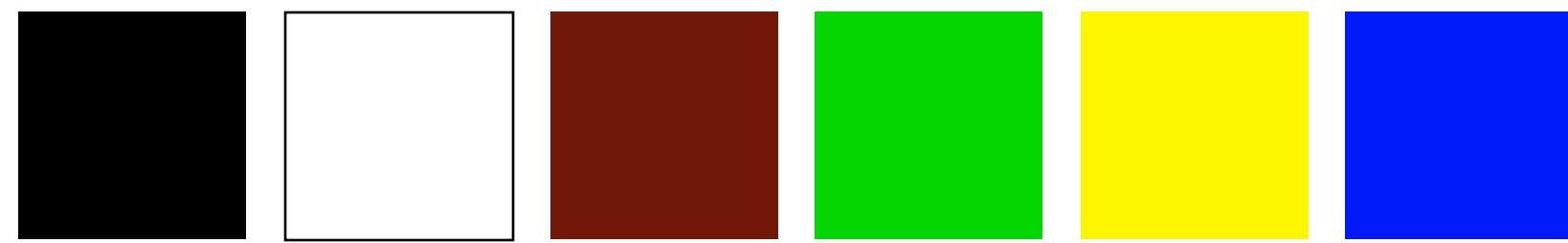


Channel Properties

Contrast



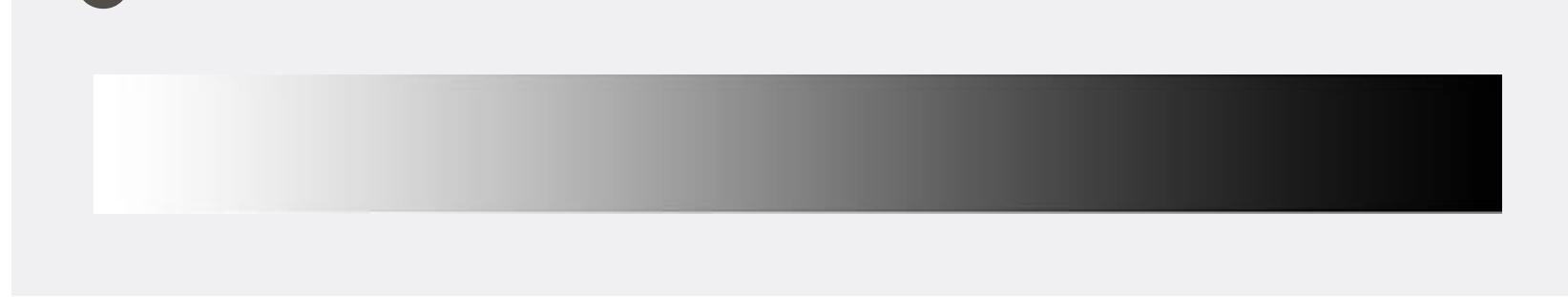
Unique hues



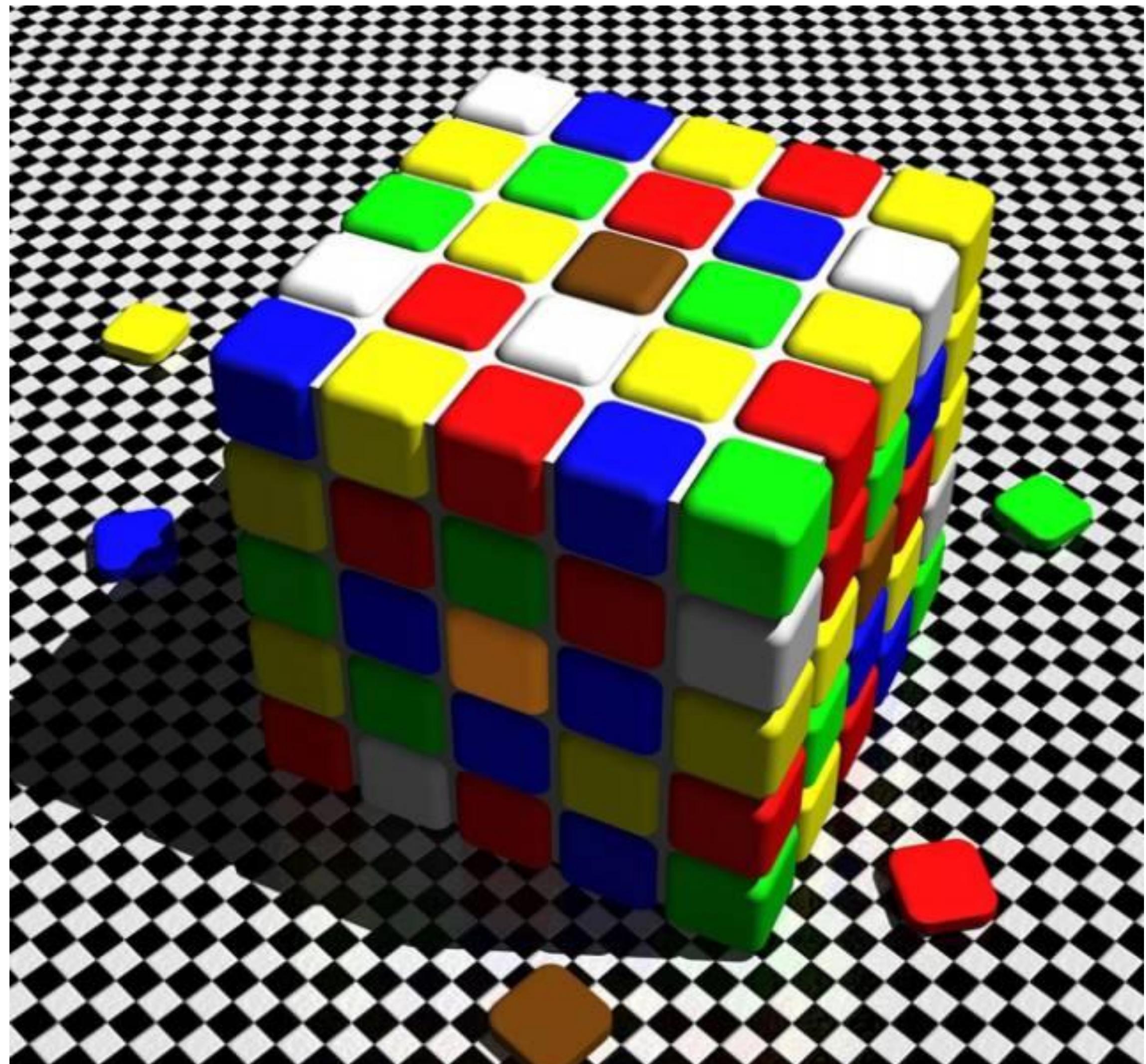
Saturation



Lightness

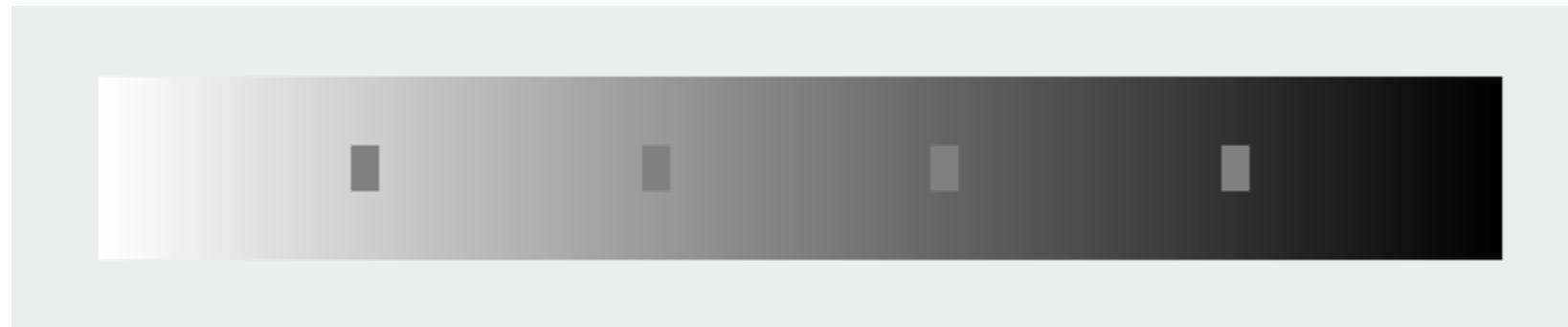


Color appearance

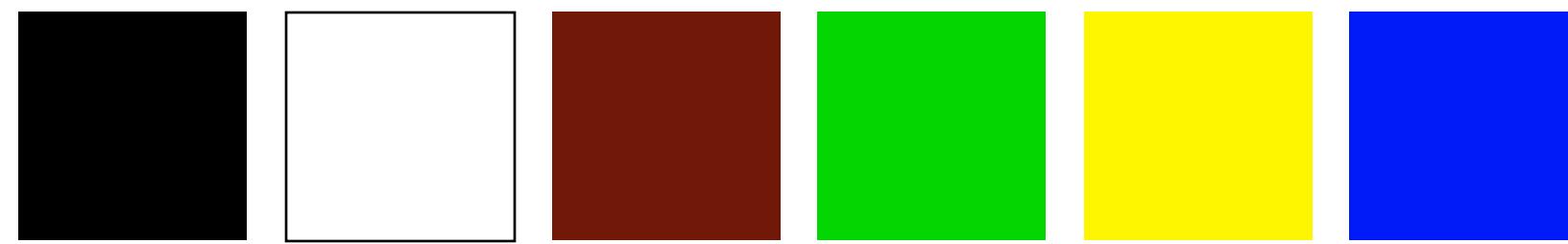


Channel Properties

Contrast



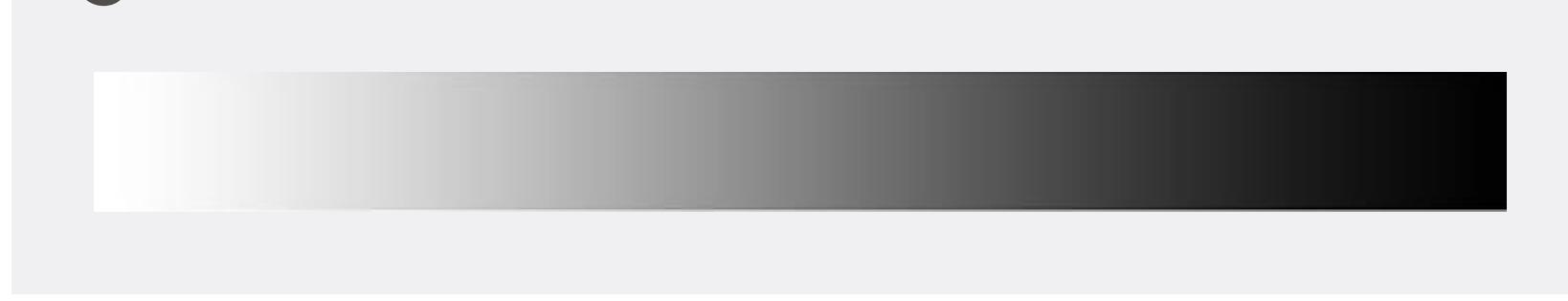
Unique hues



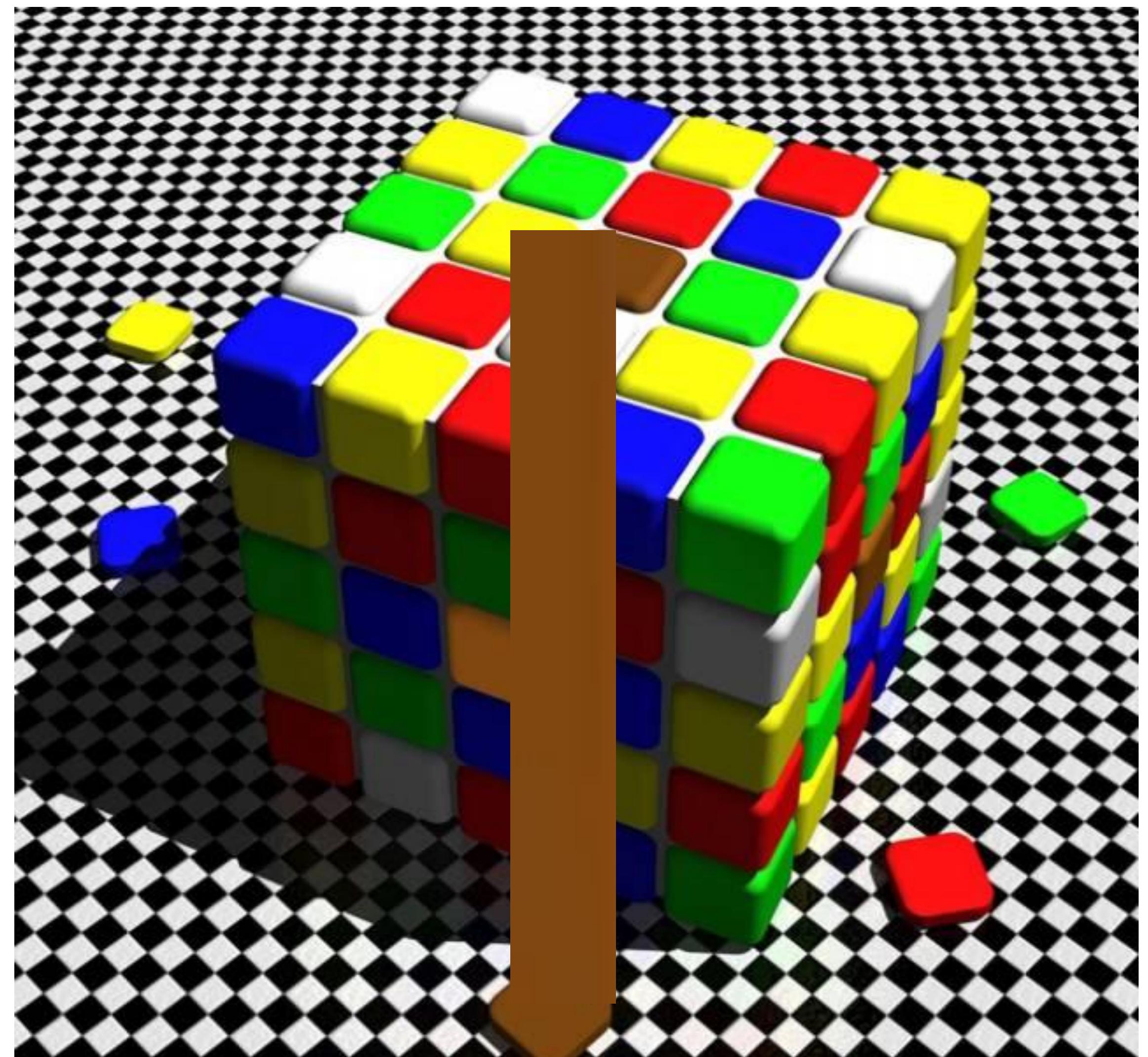
Saturation



Lightness



Color appearance



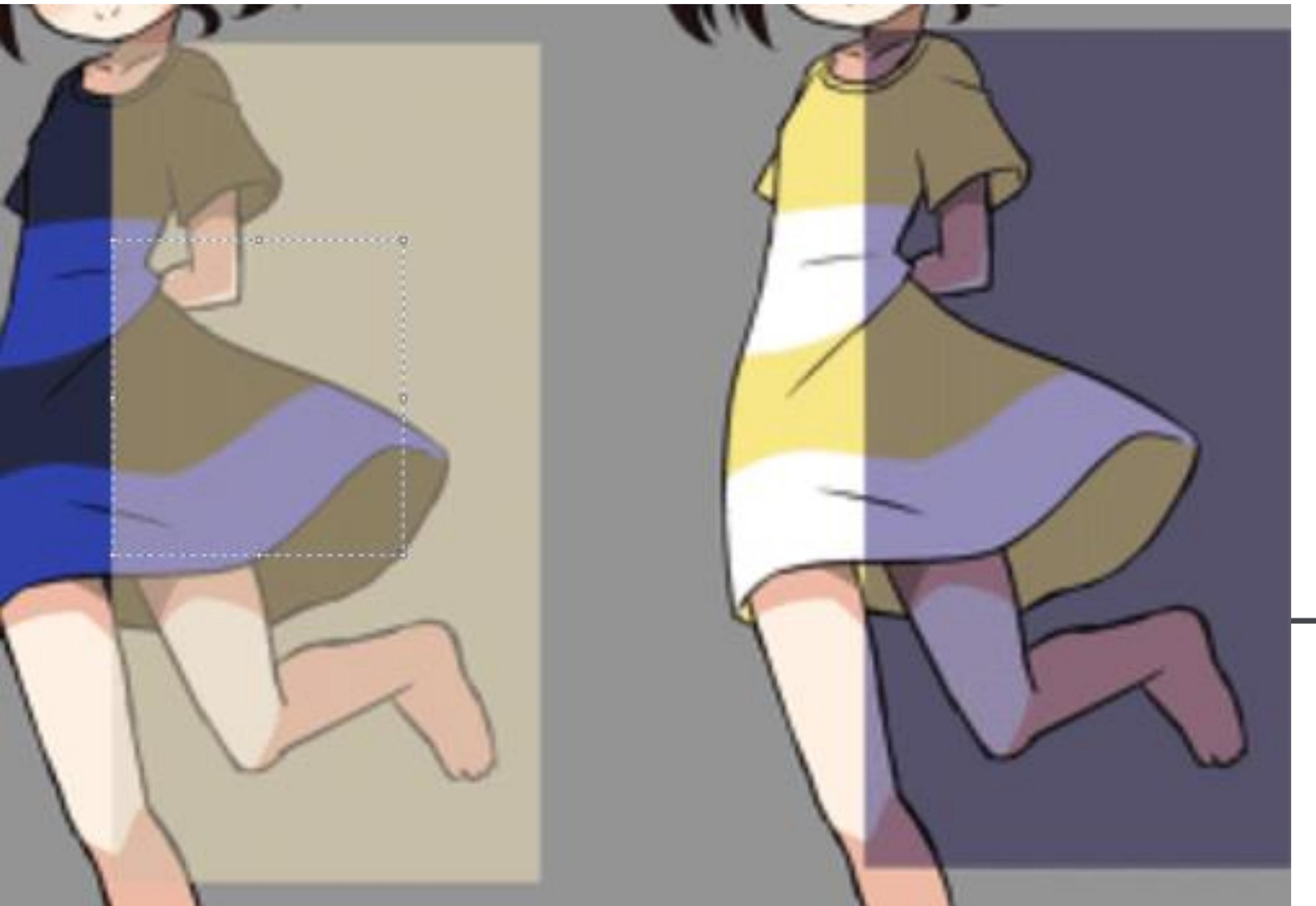
Relatividad

- Nuestro Sistema perceptual se basa fundamentalmente en juicios relativos, no absolutos
- El contexto que rodea a los elementos modula la aplicación de **Discriminabilidad** y **Precisión**



A

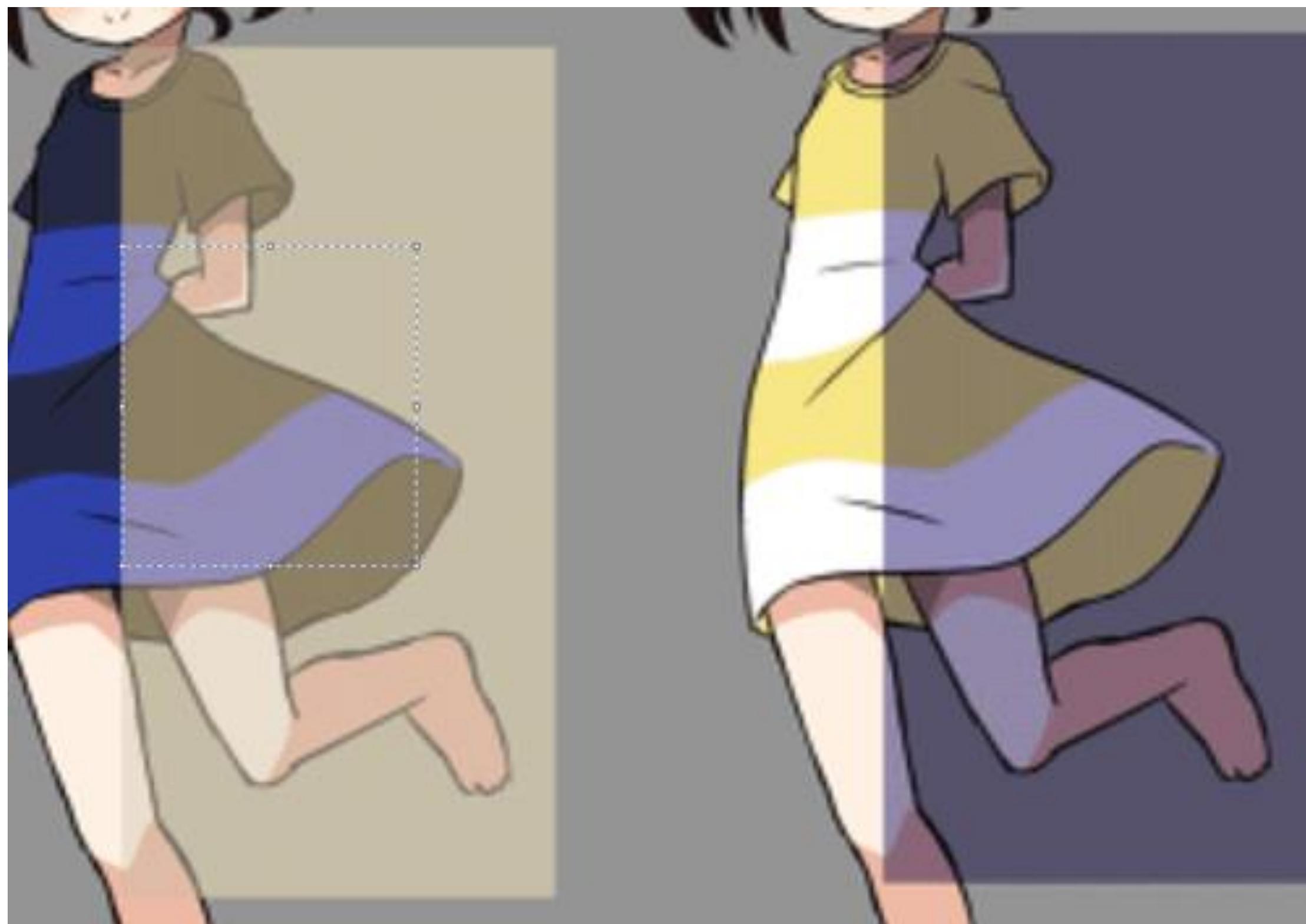
Unfram
Unalign





R177, G160, B164

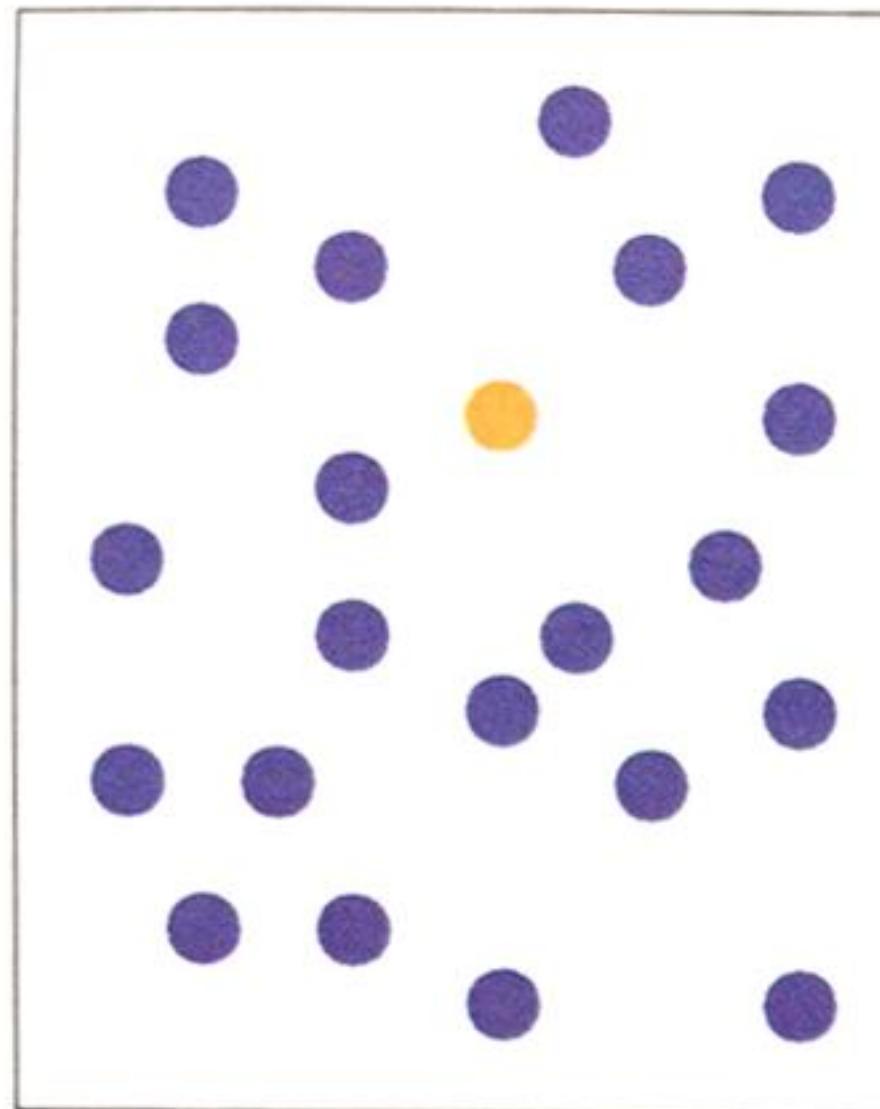
Akiyoshi kitaoka



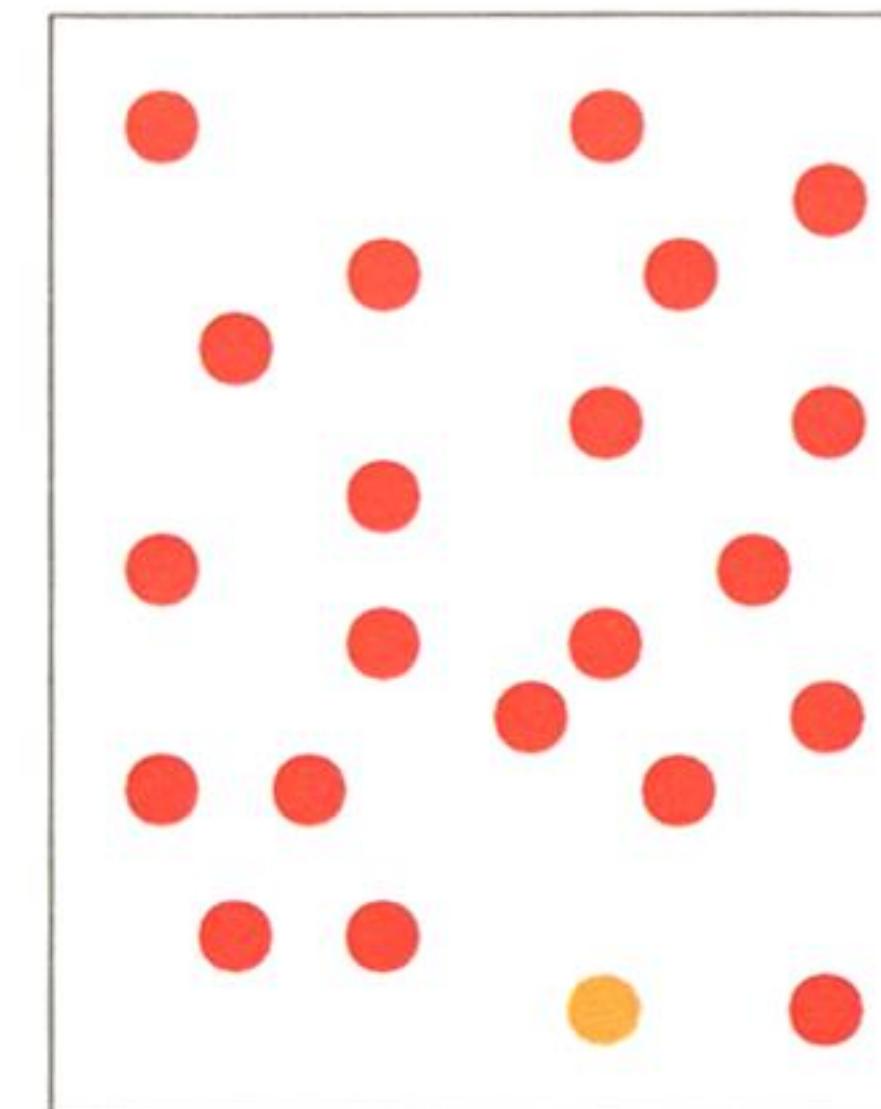


Akiyoshi kitaoka

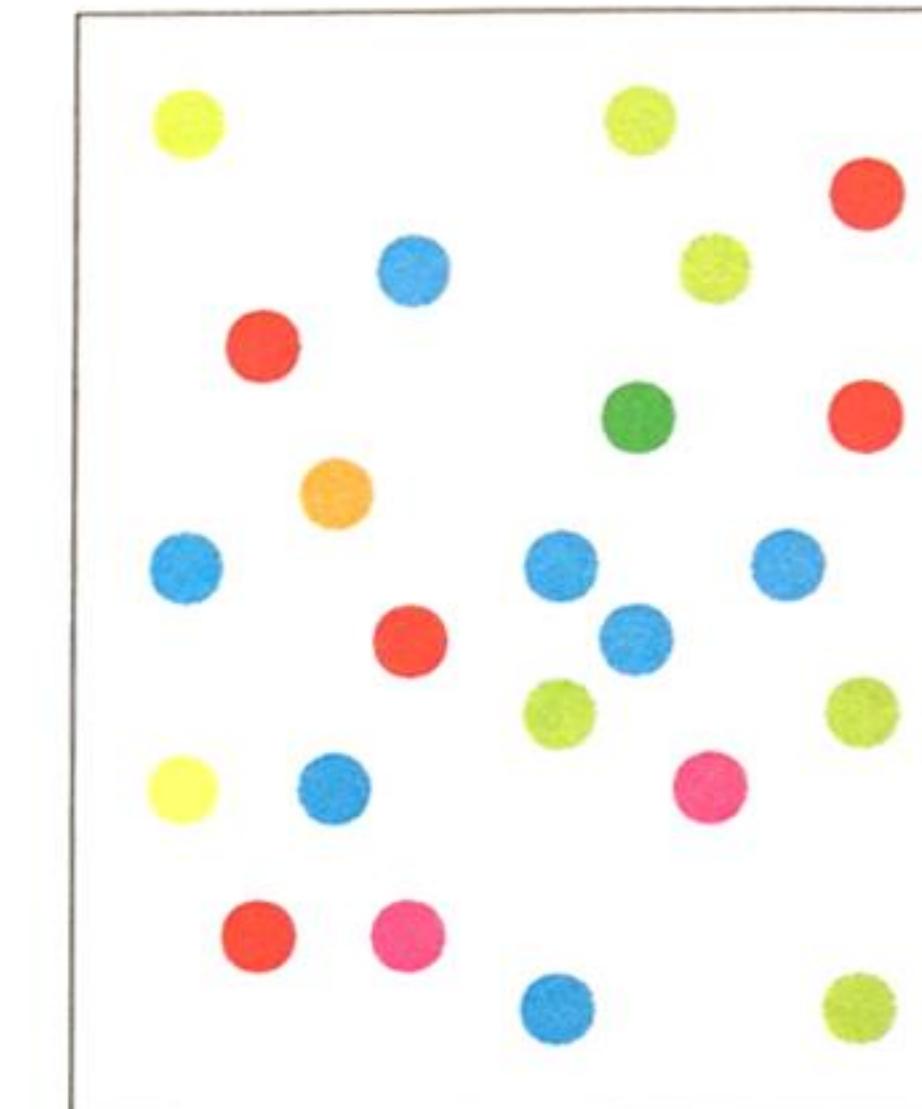
Percepción del Color - Contraste



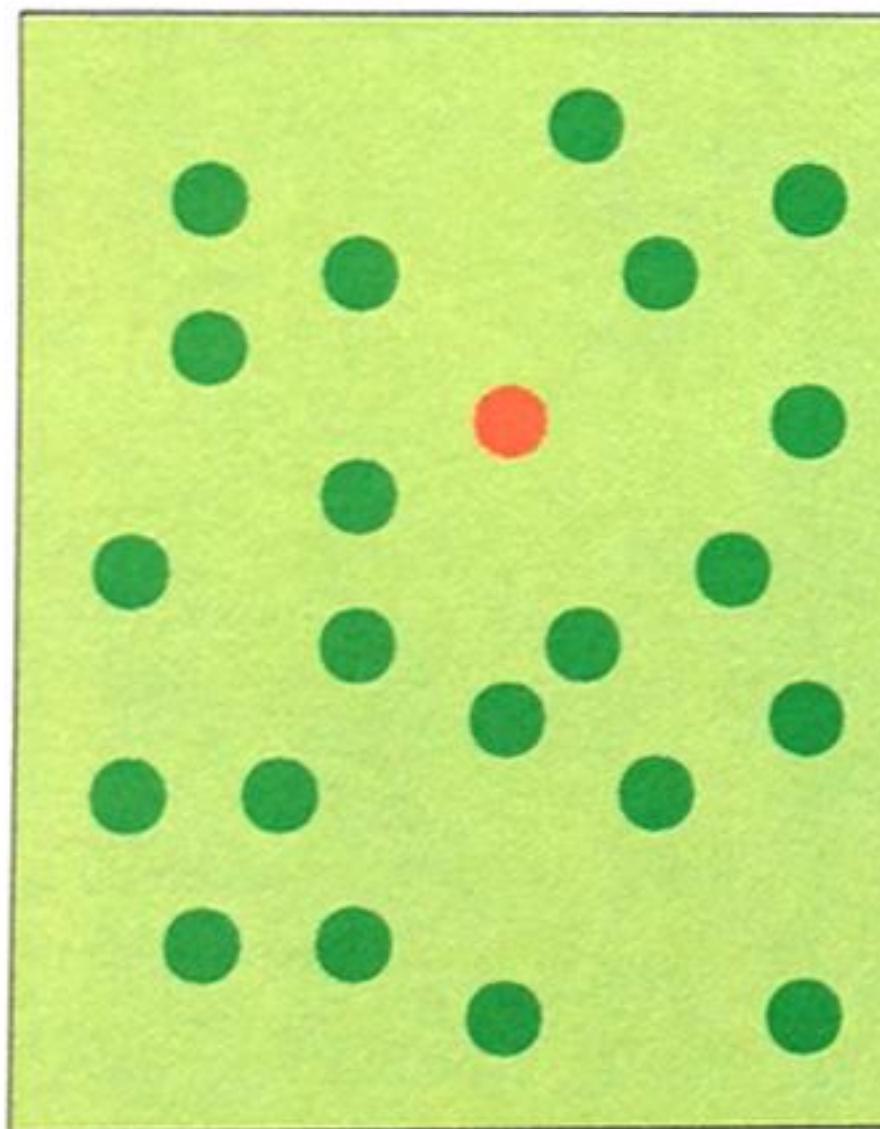
More contrast=easier



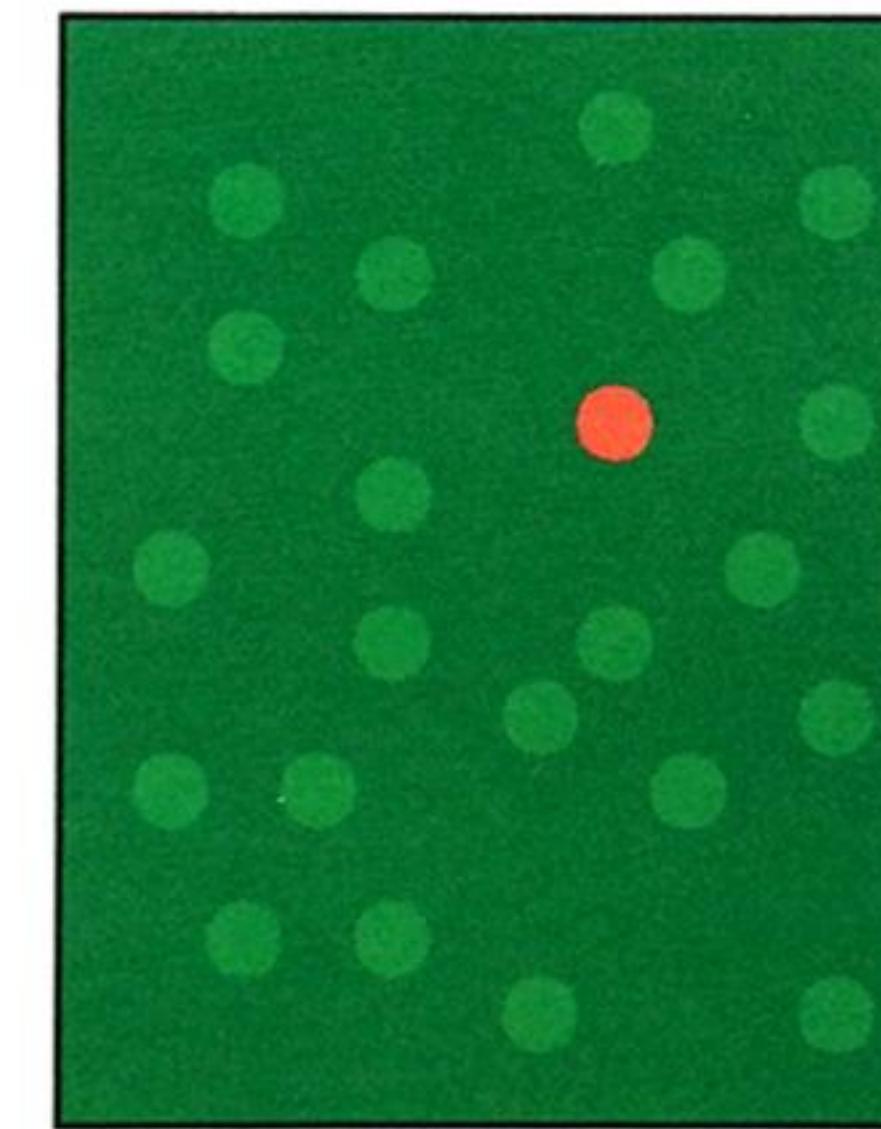
Less contrast=difficult



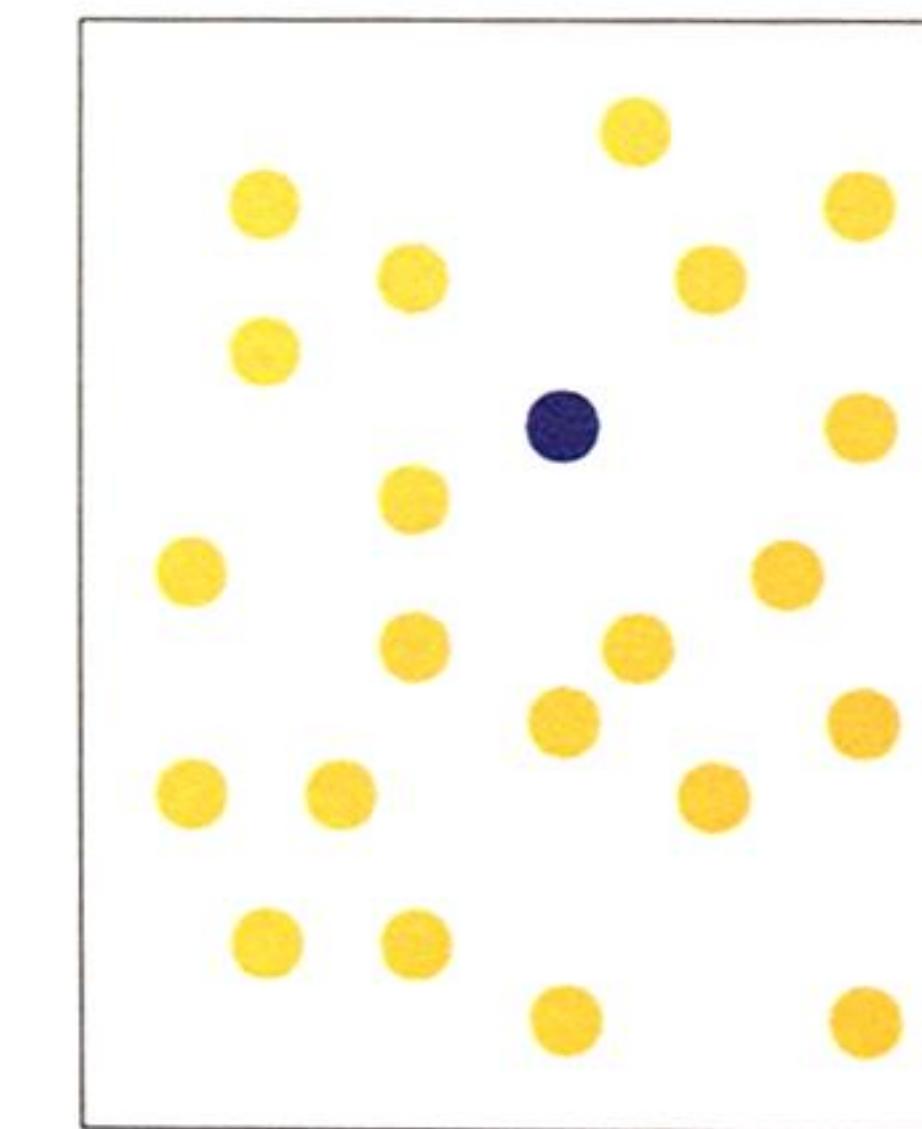
More difficult



Similarity=Easy to exclude



L+H = Easiest search



Opposite = Easy too

- El contraste facilita la detección
- Hue homogéneo
- Variación en Hue y Lightness= Más fácil

R cheatsheet

Todo esto nos sirve para saber que:

- Hay varios modelos de color, no todos son perceptualmente correctos.
- Para escalas de color necesitamos modelos de interpolación no lineal.
- Algunos colores tienen mayor contraste entre si. Podemos usarlo para dirigir la atención.
- El tamaño modula la percepción de los colores
- Evitar artefactos (contraste simultaneo, apariencia de color)

R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0 i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

R has 657 built in color names

Example: To see a list of names:

[colors\(\)](#)

These colors are displayed on P. 3.

peachpuff4

R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.

[rgb\(r, g, b, maxColorValue=255, alpha=255\)](#)

- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument

[hsv\(h, s, v, alpha\)](#)

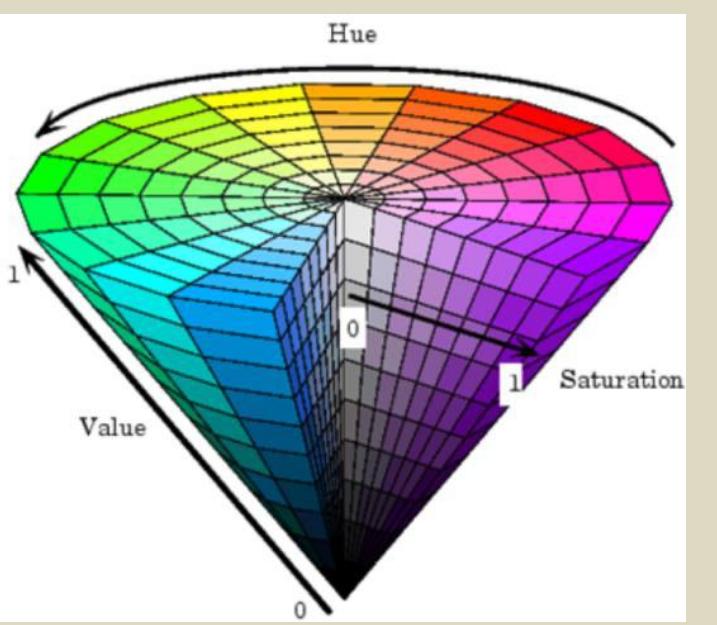
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other

[hcl\(h, c, l, alpha\)](#)

A few notes on HSV/HLC

HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model....blah blah blah...

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.



R can translate colors to **rgb** (this is handy for matching colors in other programs)
[col2rgb\(c\("#FF0000", "blue"\)\)](#)

R Color Palettes

This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure....NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps

grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

[heat.colors\(4, alpha=1\)](#)

> "#FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFF800FF"

For the **rainbow** palette you can also select start/end color (red = 0, yellow = 1/6, green = 2/6, cyan = 3/6, blue = 4/6 and magenta = 5/6) and saturation (s) and value (v):
[rainbow\(n, s = 1, v = 1, start = 0, end = max\(1, n - 1\)/n, alpha = 1\)](#)

[grDevices palettes](#)
[cm.colors](#)
[topo.colors](#)
[terrain.colors](#)
[heat.colors](#)
[rainbow](#)
see P. 4 for options

Package: RcolorBrewer

This function has an argument for the number of colors and the color palette (see P. 4 for options).
[brewer.pal\(4, "Set3"\)](#)

> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: [display.brewer.all\(5\)](#)

There is also a very nice interactive viewer:

<http://colorbrewer2.org/>

My Recommendation

Package: colorspace

These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

[rainbow_hcl\(4\)](#)

"#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"

[colorspace default_palettes](#)
[diverge_hcl](#)
[diverge_hsl](#)
[terrain_hcl](#)
[sequential_hcl](#)
[rainbow_hcl](#)

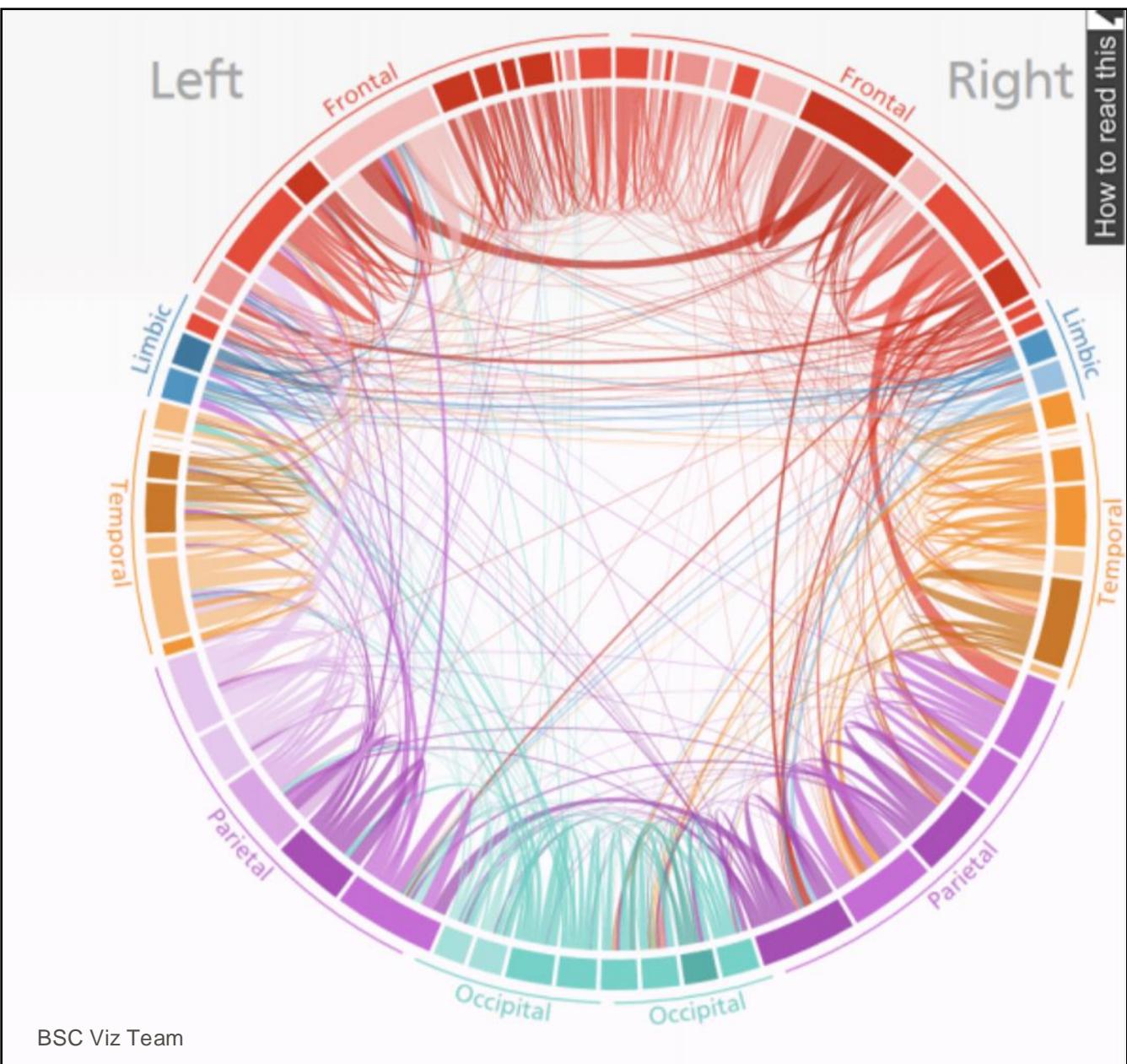
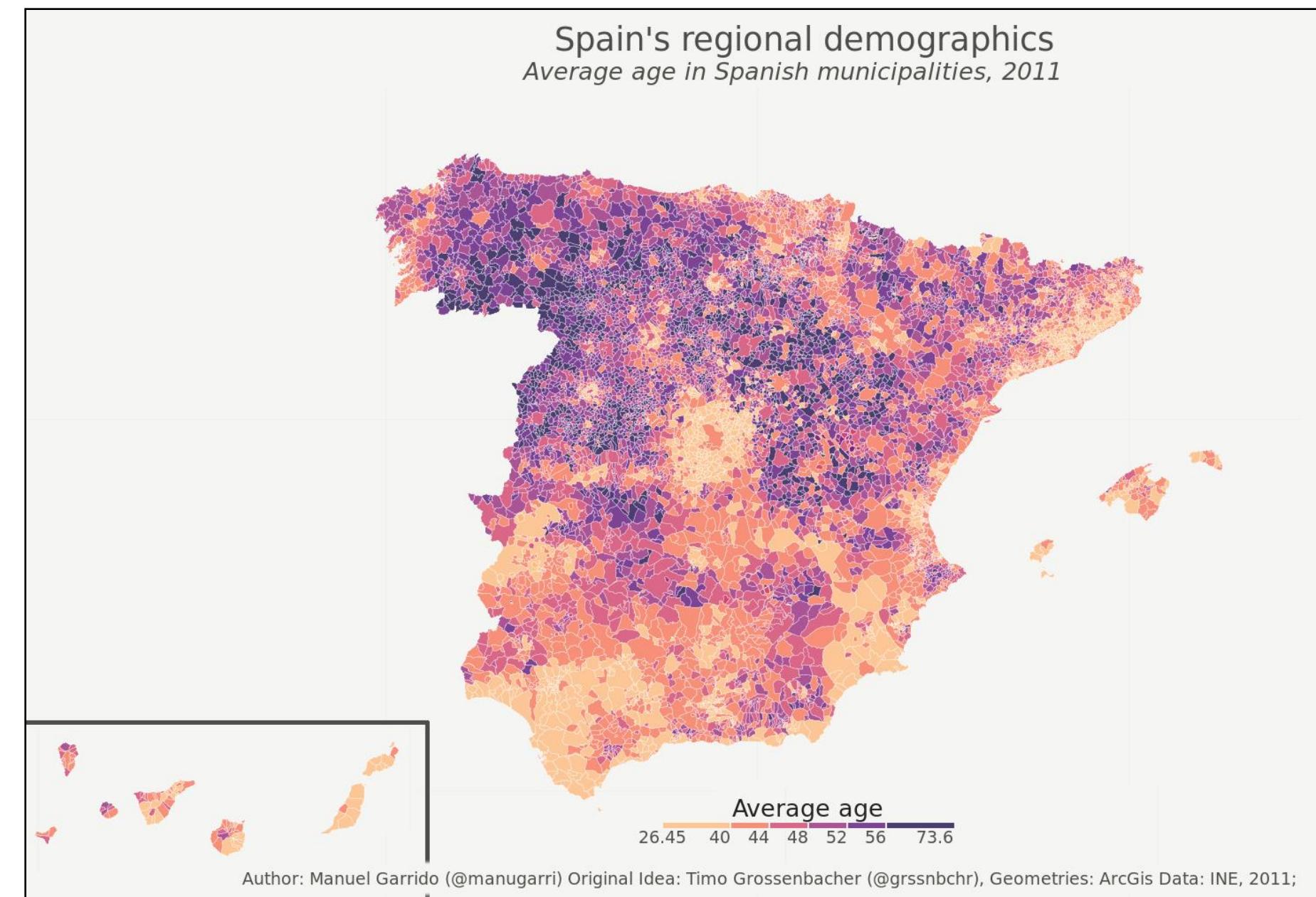
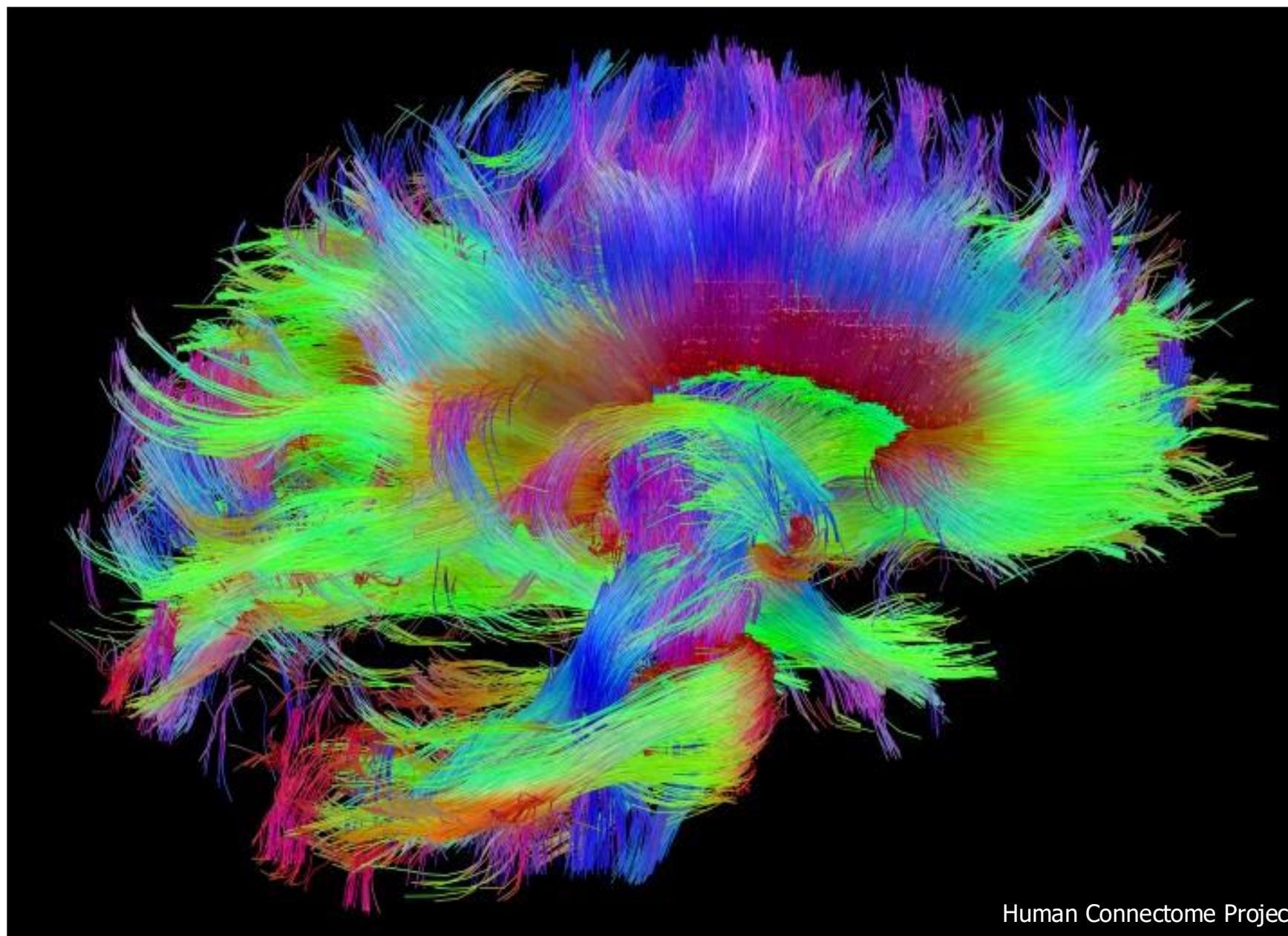
However, all palettes are fully customizable:

[diverge_hcl\(7, h = c\(246, 40\), c = 96, l = c\(65, 90\)\)](#)

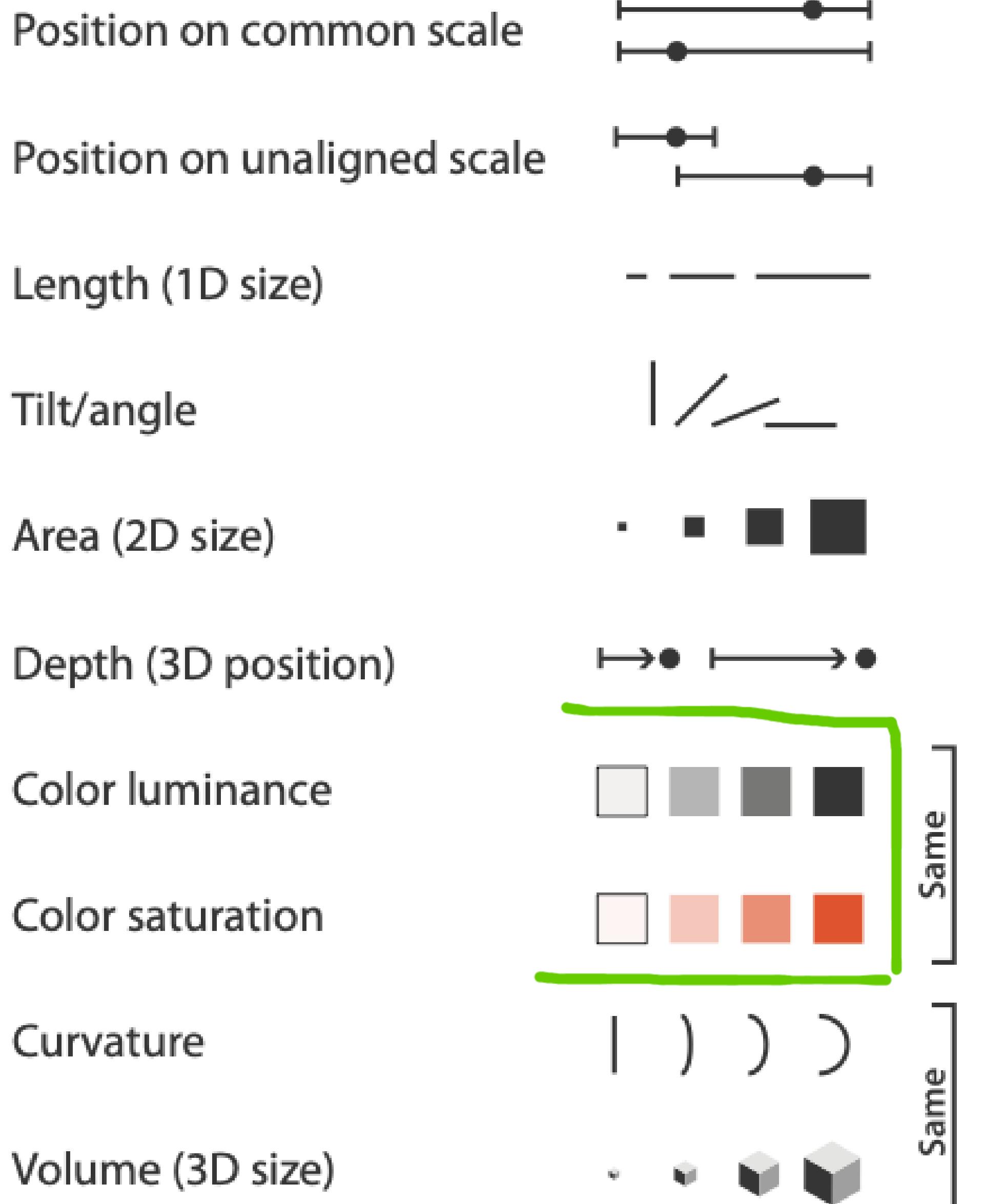
Choosing the values would be daunting. But there are some recommended palettes in the colorspace documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
[pal <- choose_palette\(\)](#)

Escalas y paletas de Color

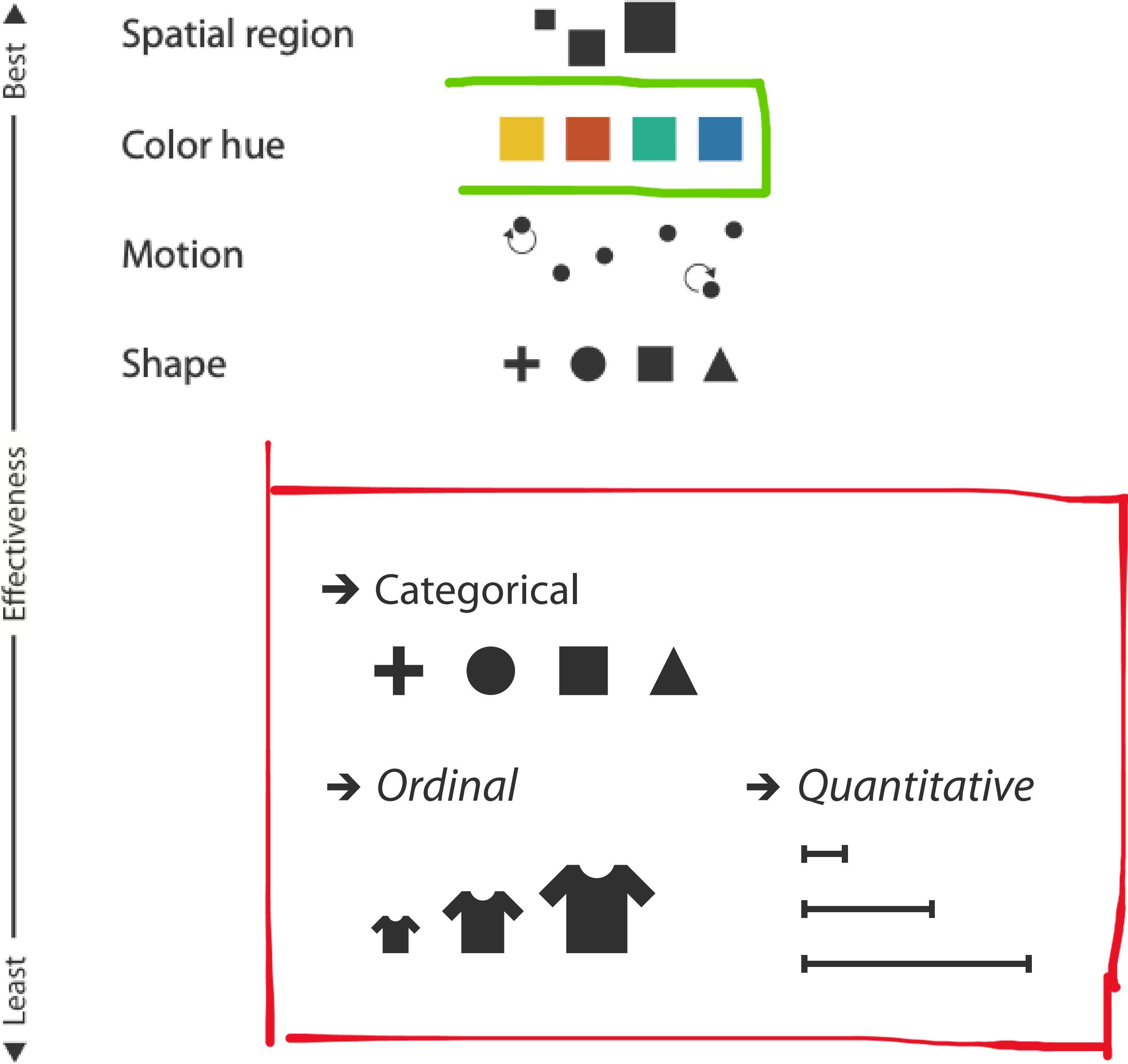
Colorear datos



→ **Magnitude Channels: Ordered Attributes**

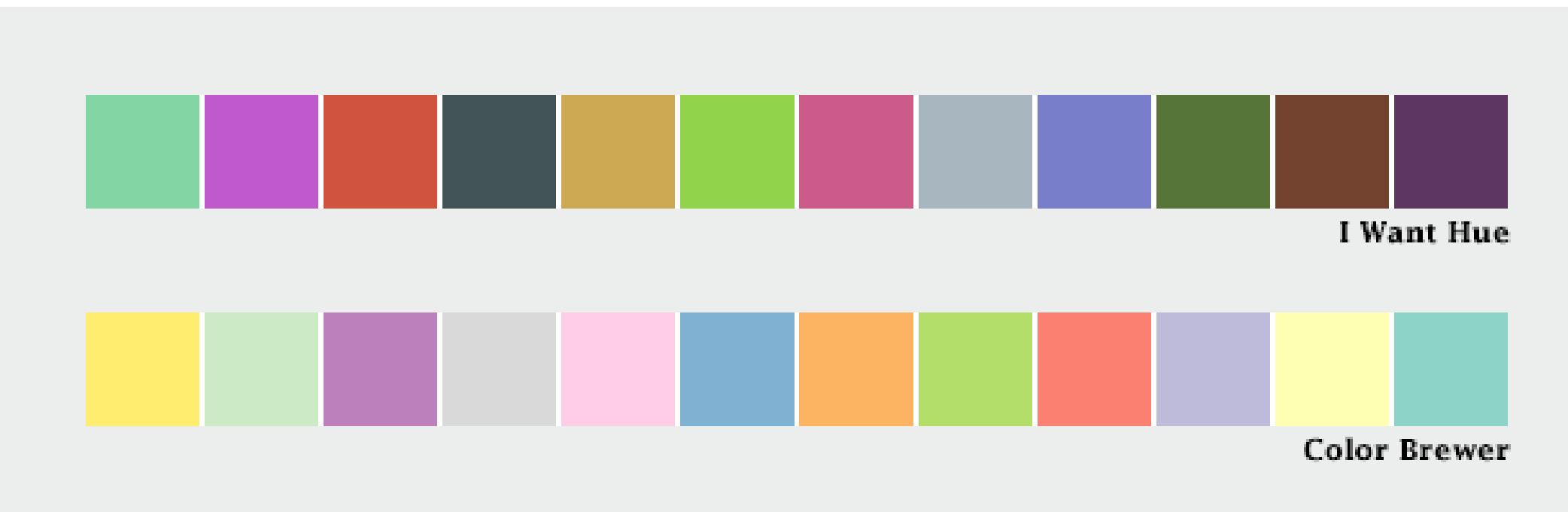


→ **Identity Channels: Categorical Attributes**



Colorear datos -> Color palette/scheme/scale

Categórica/Cualitativa



➡ Ordering Direction

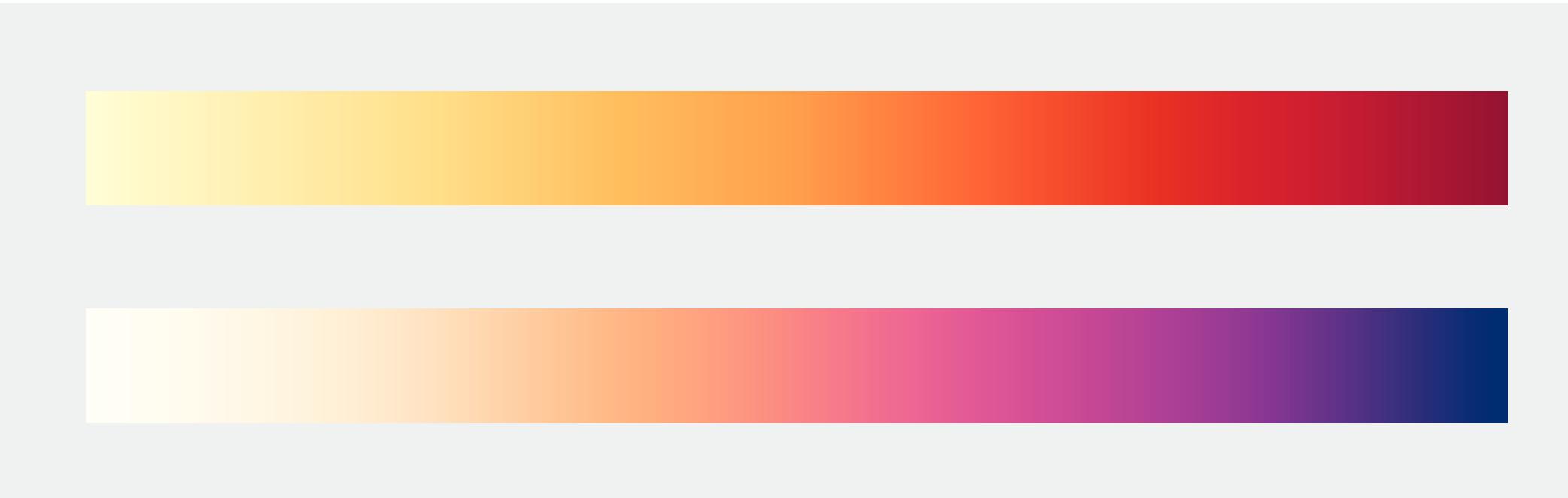
→ Sequential



→ Diverging



Secuencial

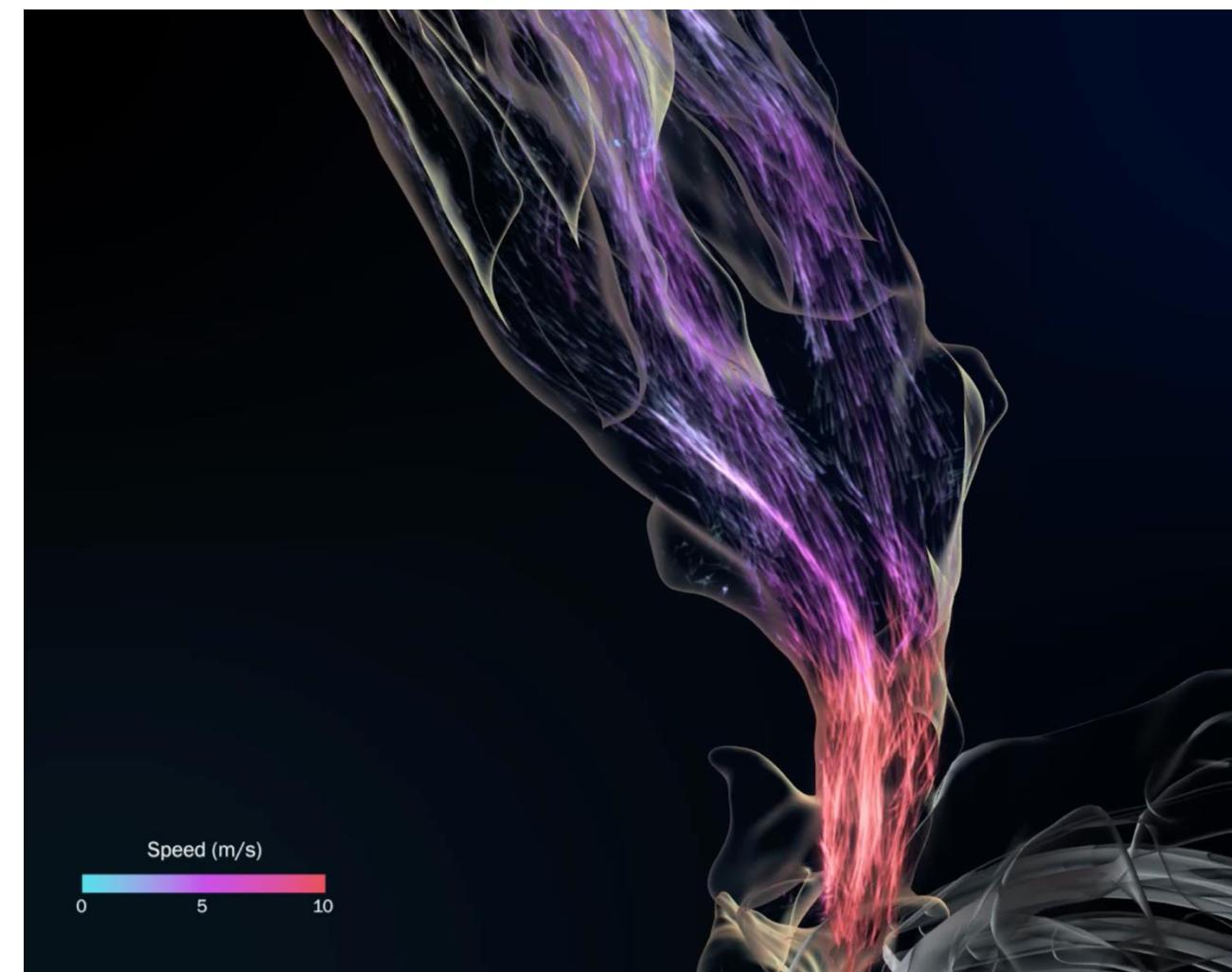


Divergente

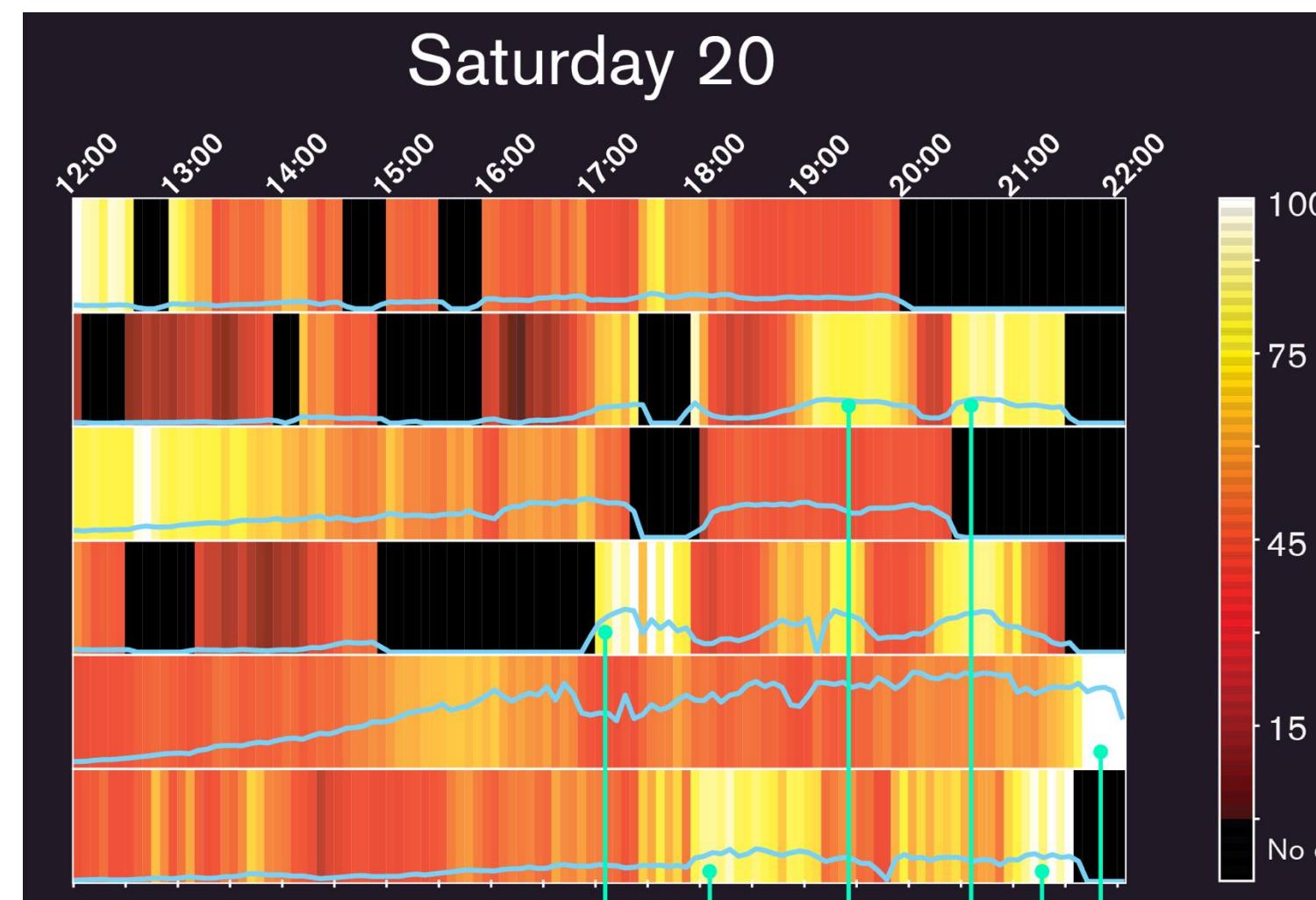


Cuantitativo Secuencial

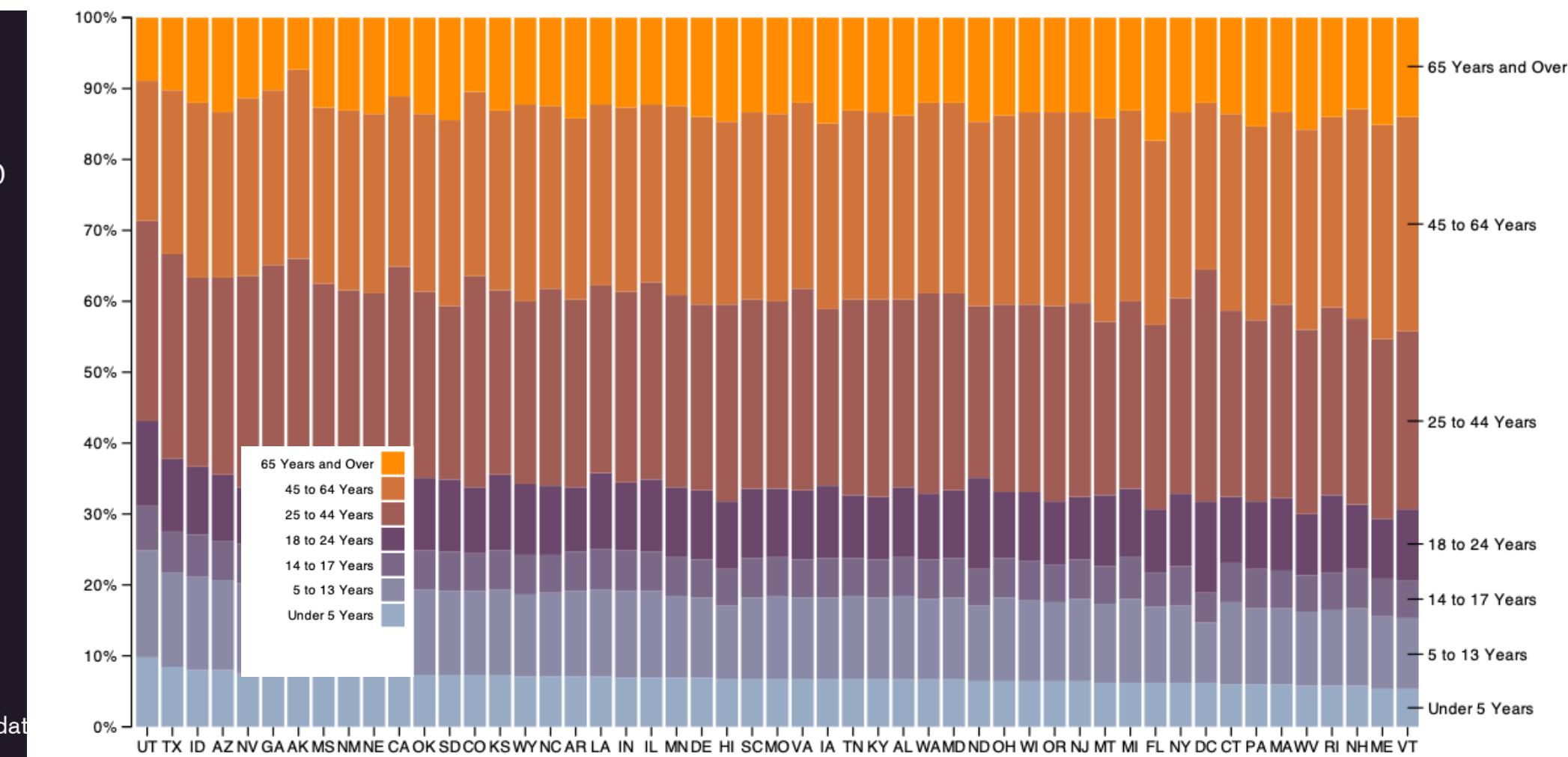
Continua



Discretizada

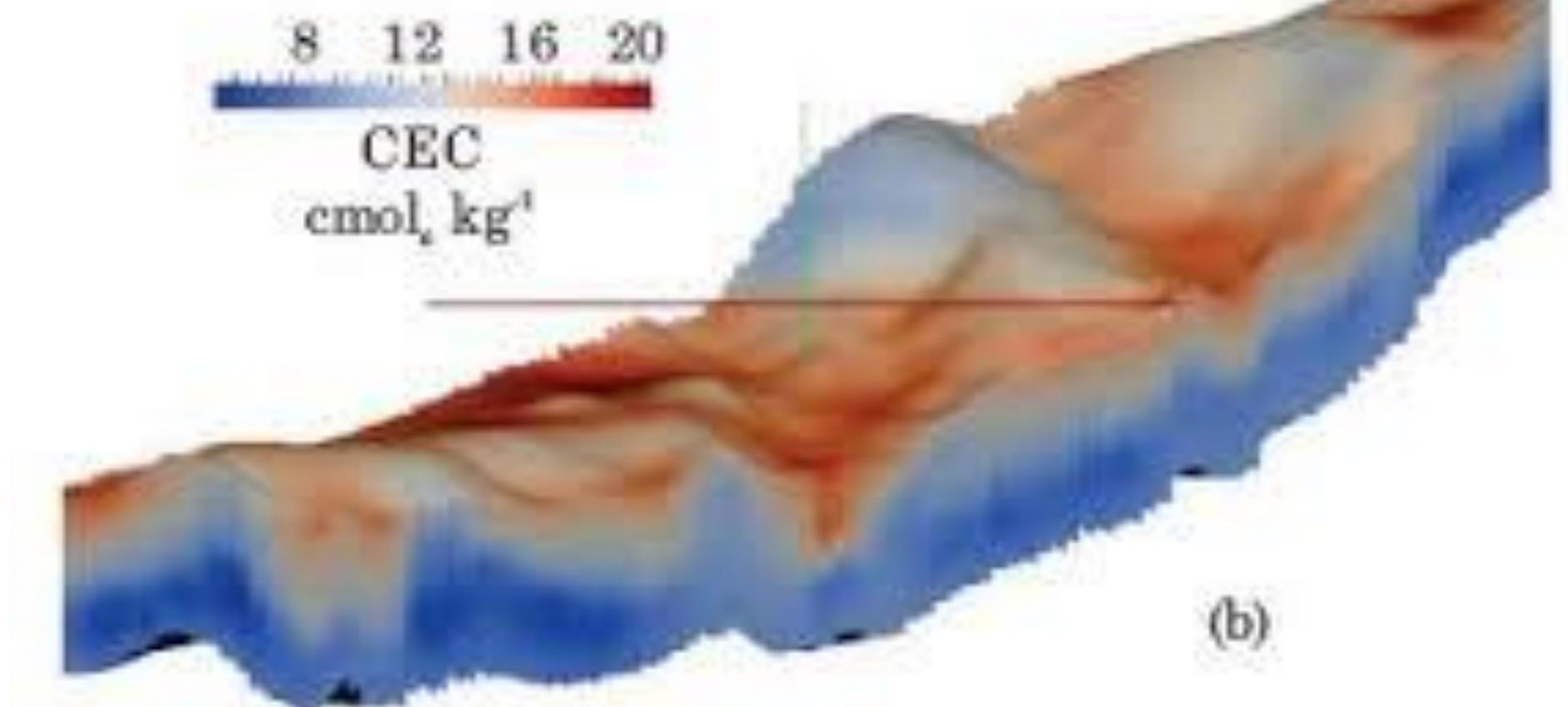


Ordinal = Continua discretizada

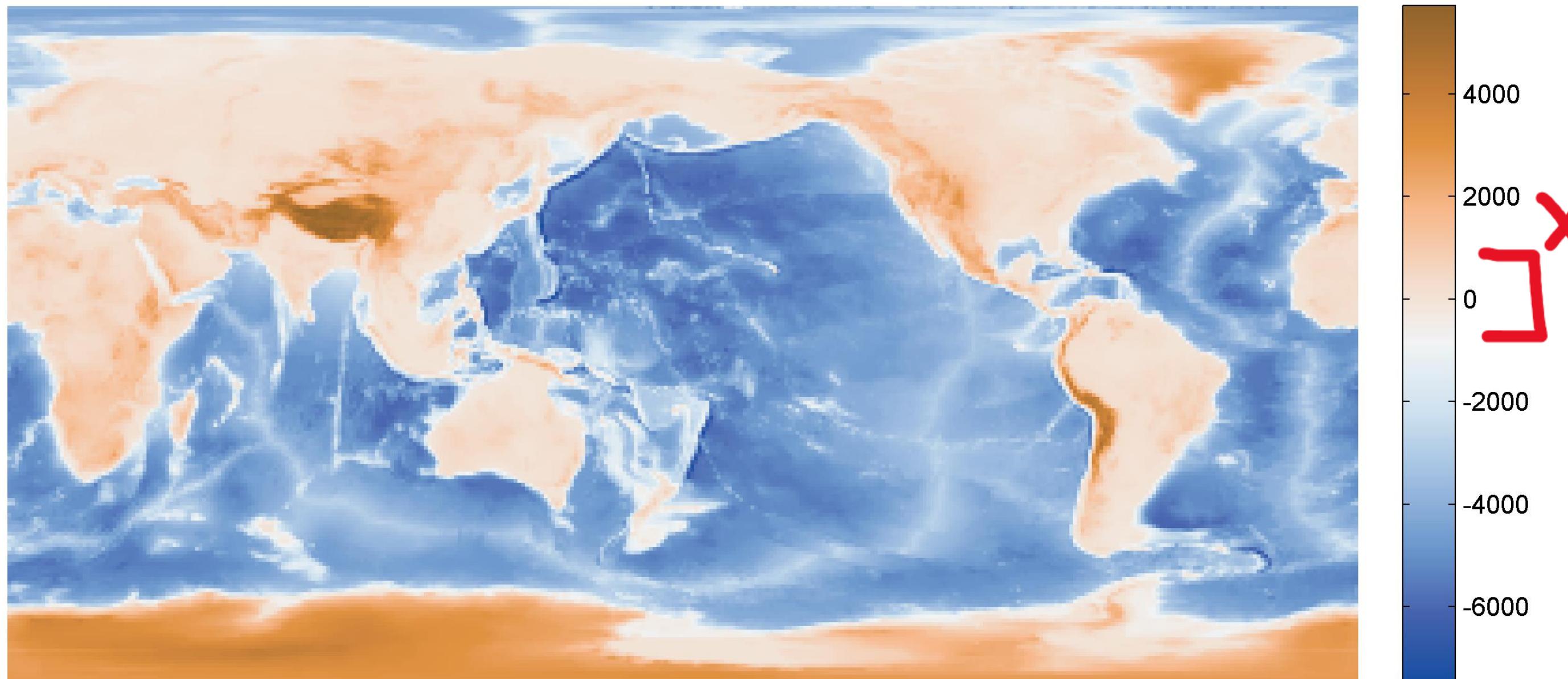


Cuantitativo Divergente

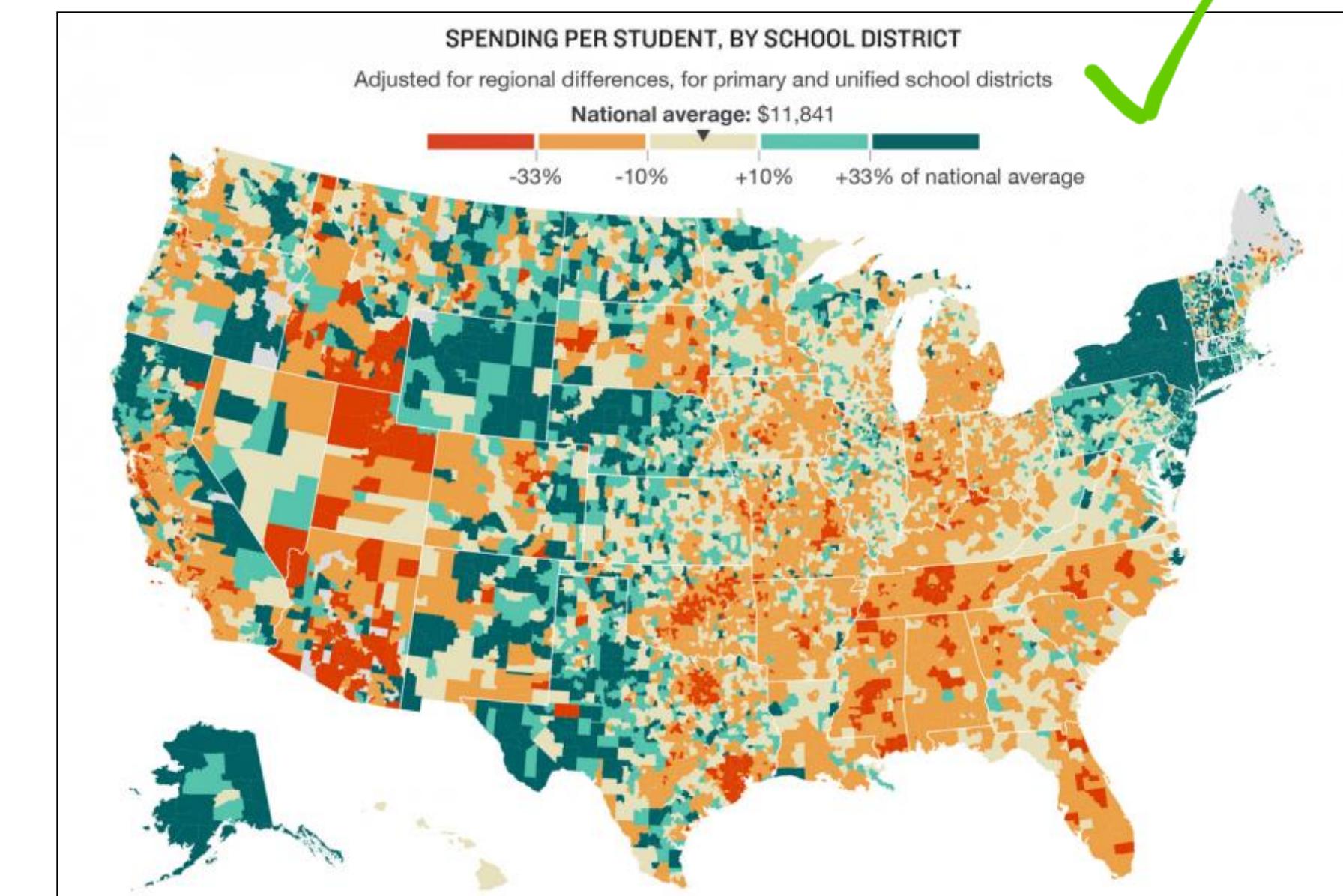
Es importante que el color central (normalmente blanco) represente el valor central en los datos (normalmente 0)



Continua

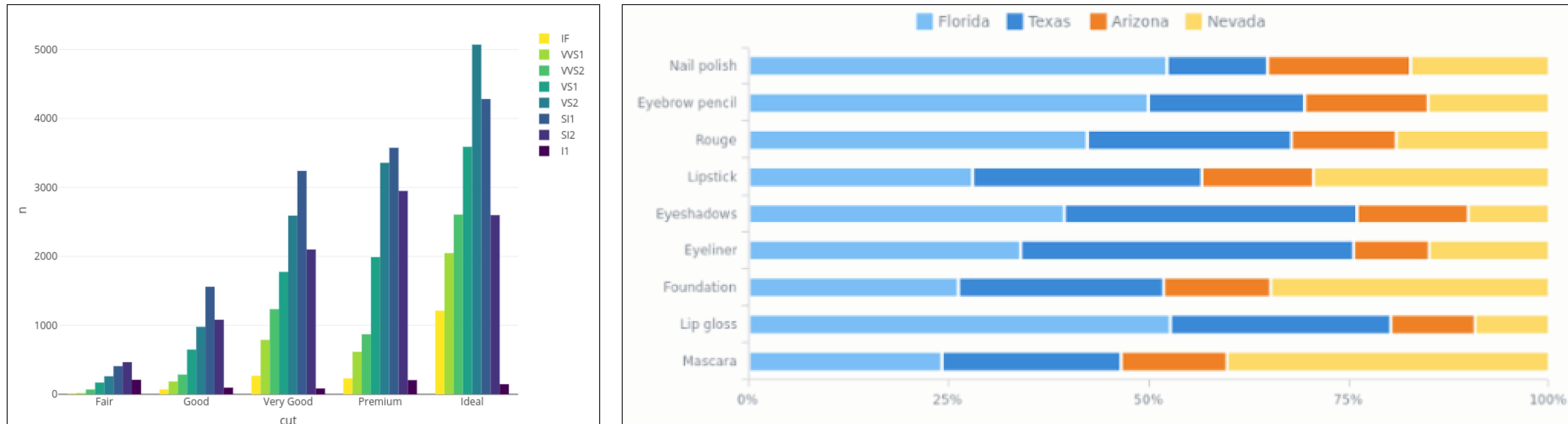


Discretizada



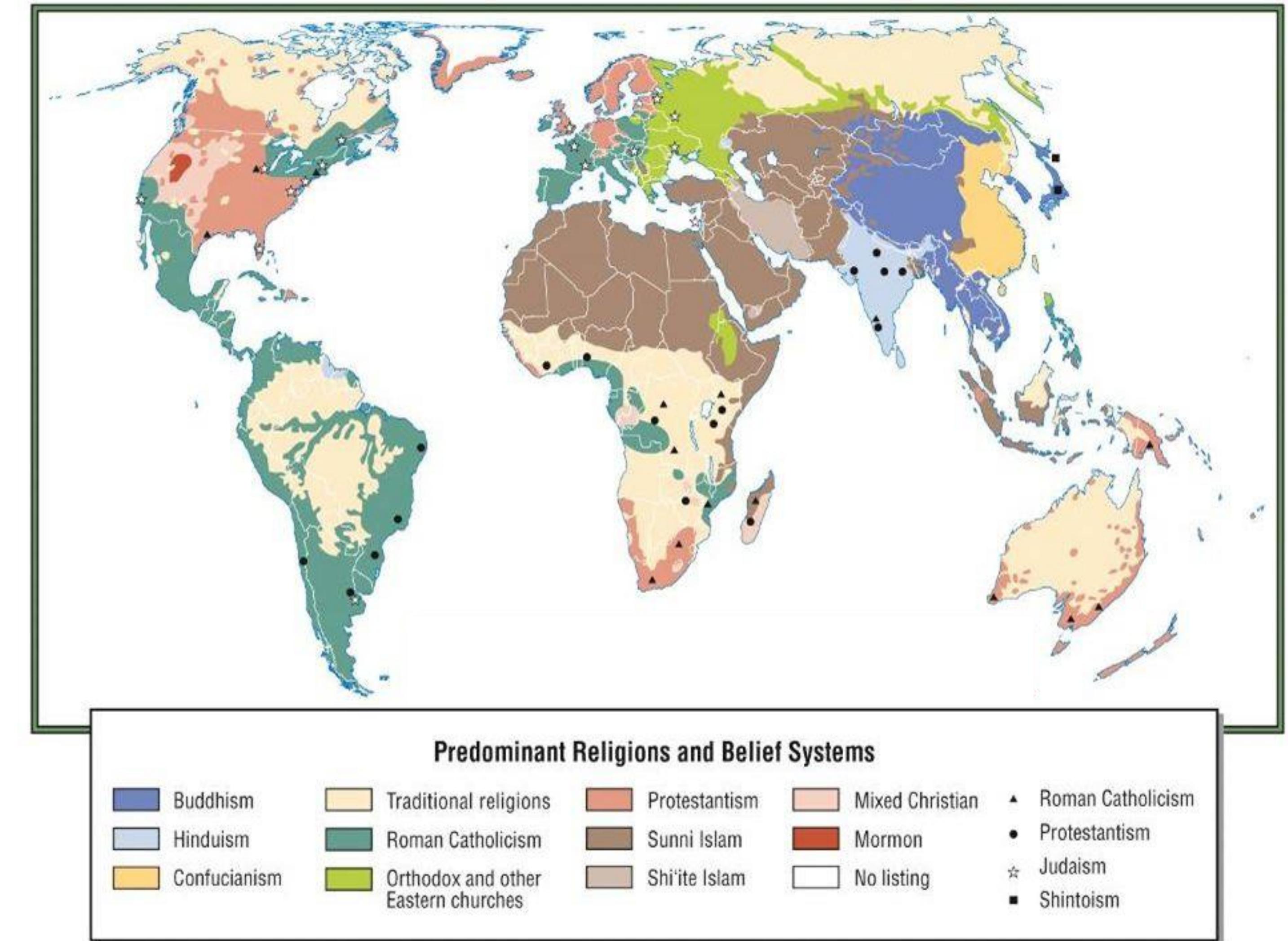
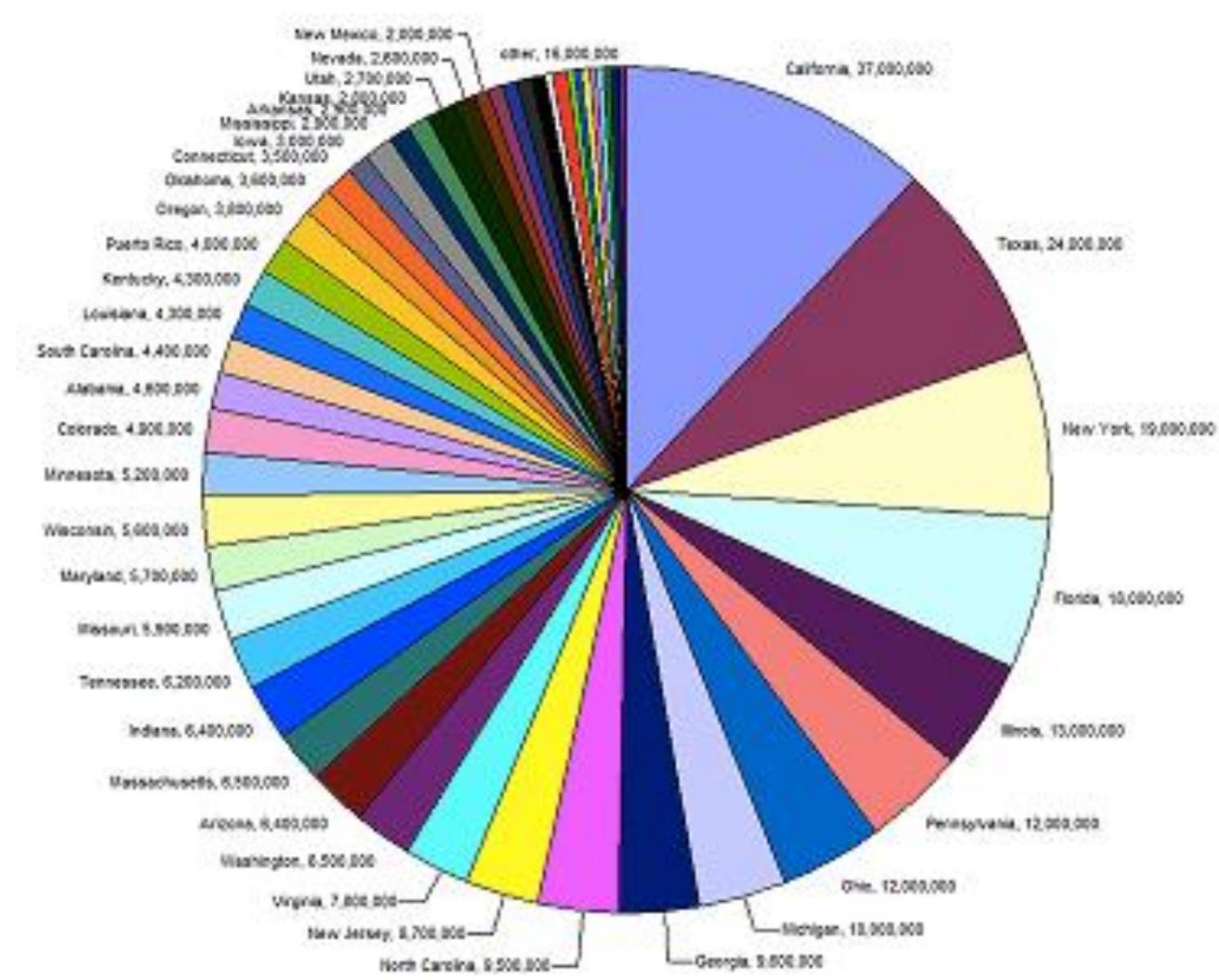
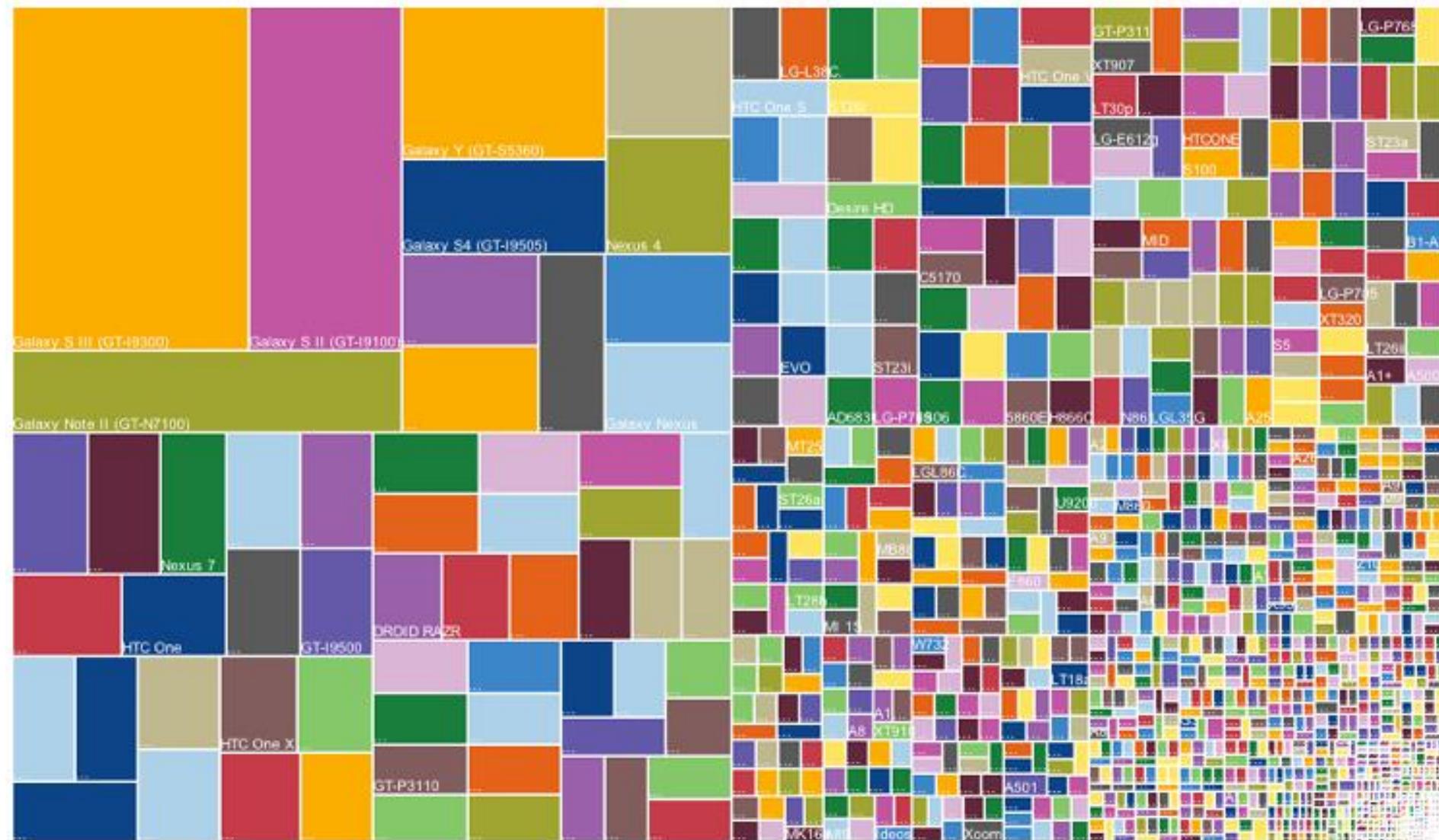
Categoría

Cuidado con colores que formen un gradiente continuo, pueden confundirse con valores ordinales
Casi siempre es mejor diferencias grandes en tono, saturación, y/o luminosidad



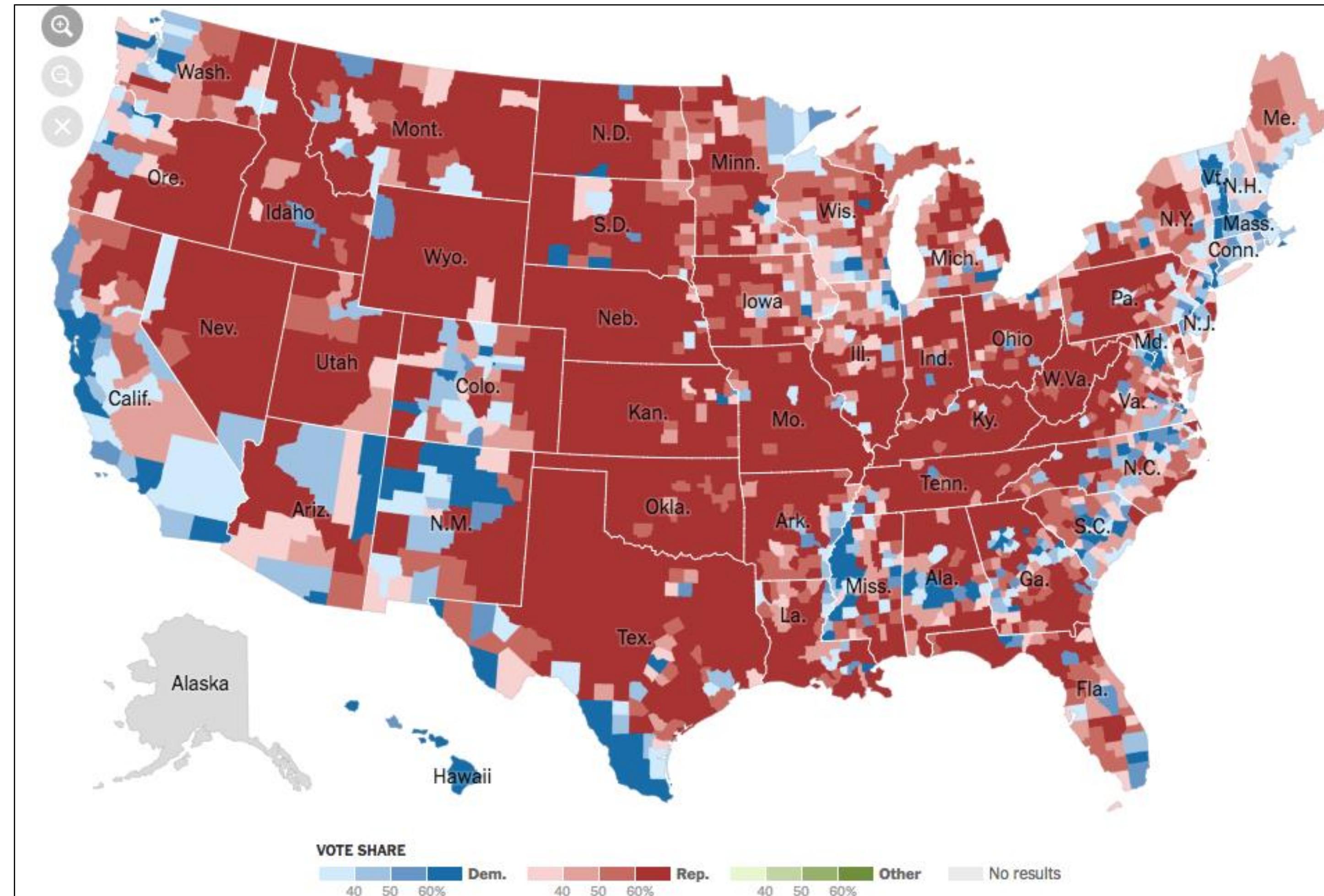
Categórica

No demasiados colores (max. 12)



Agrupado

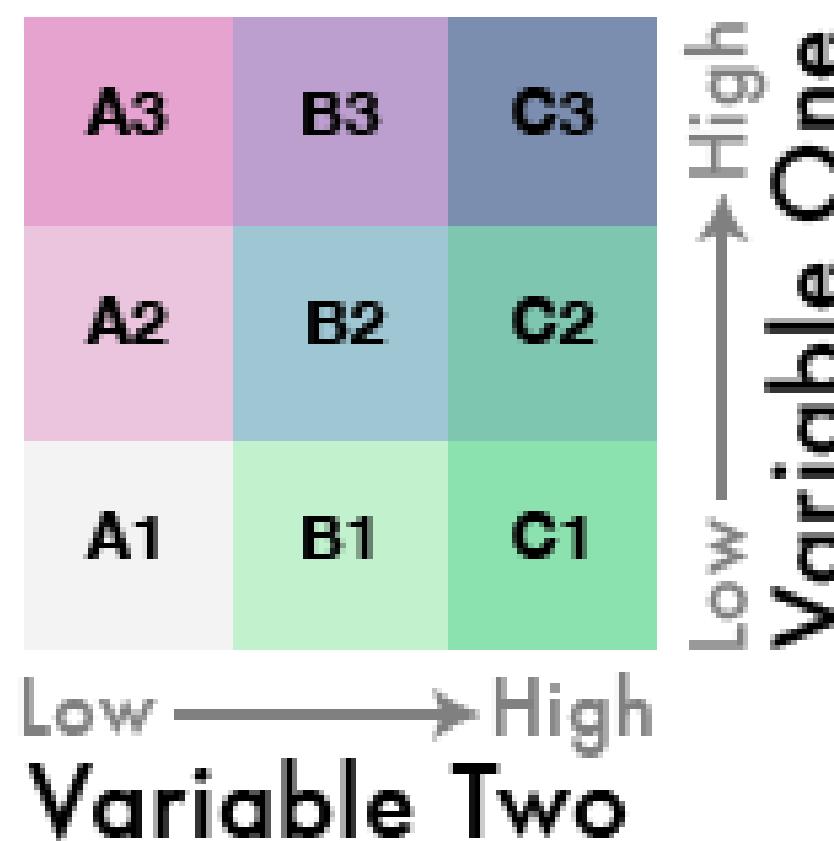
Esquema de color agrupado de 4 categorías (4 tonos), con 4 pasos de saturación y luminosidad cada uno



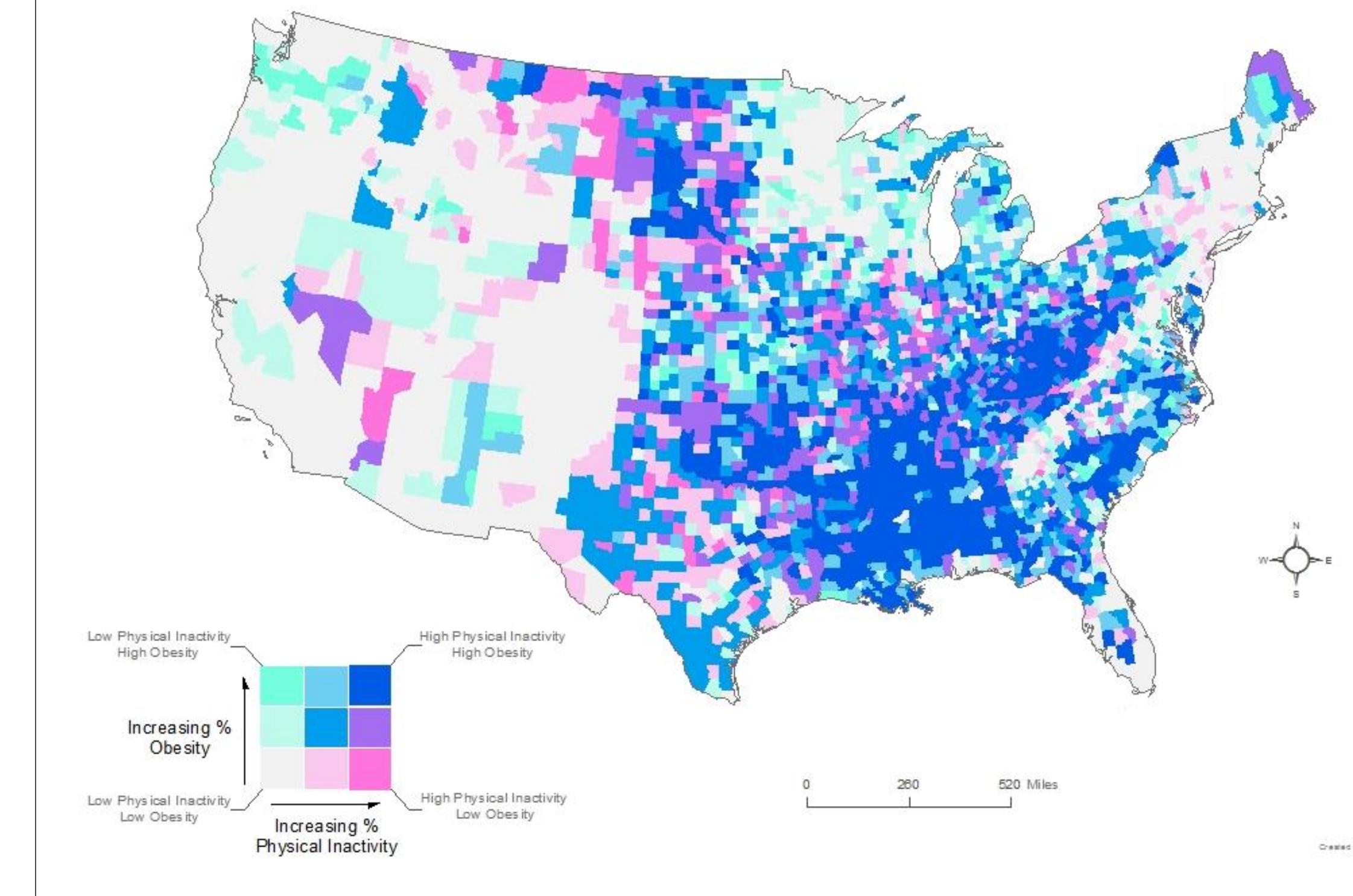
Bi-variate choropleth maps

Esquema de color con dos escalas sobreuestas en dos direcciones.

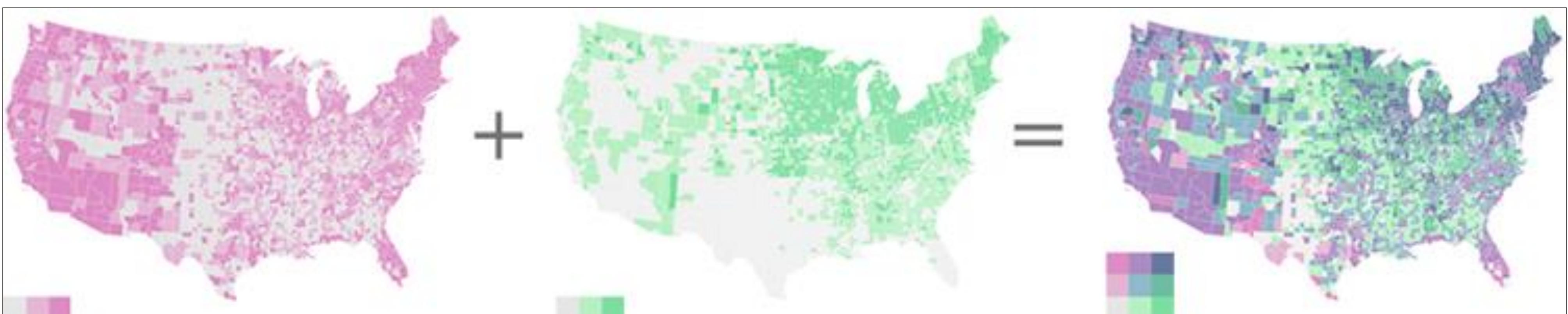
Normalmente resulta una matriz de 3x3



The Effects of Physical Inactivity on Obesity

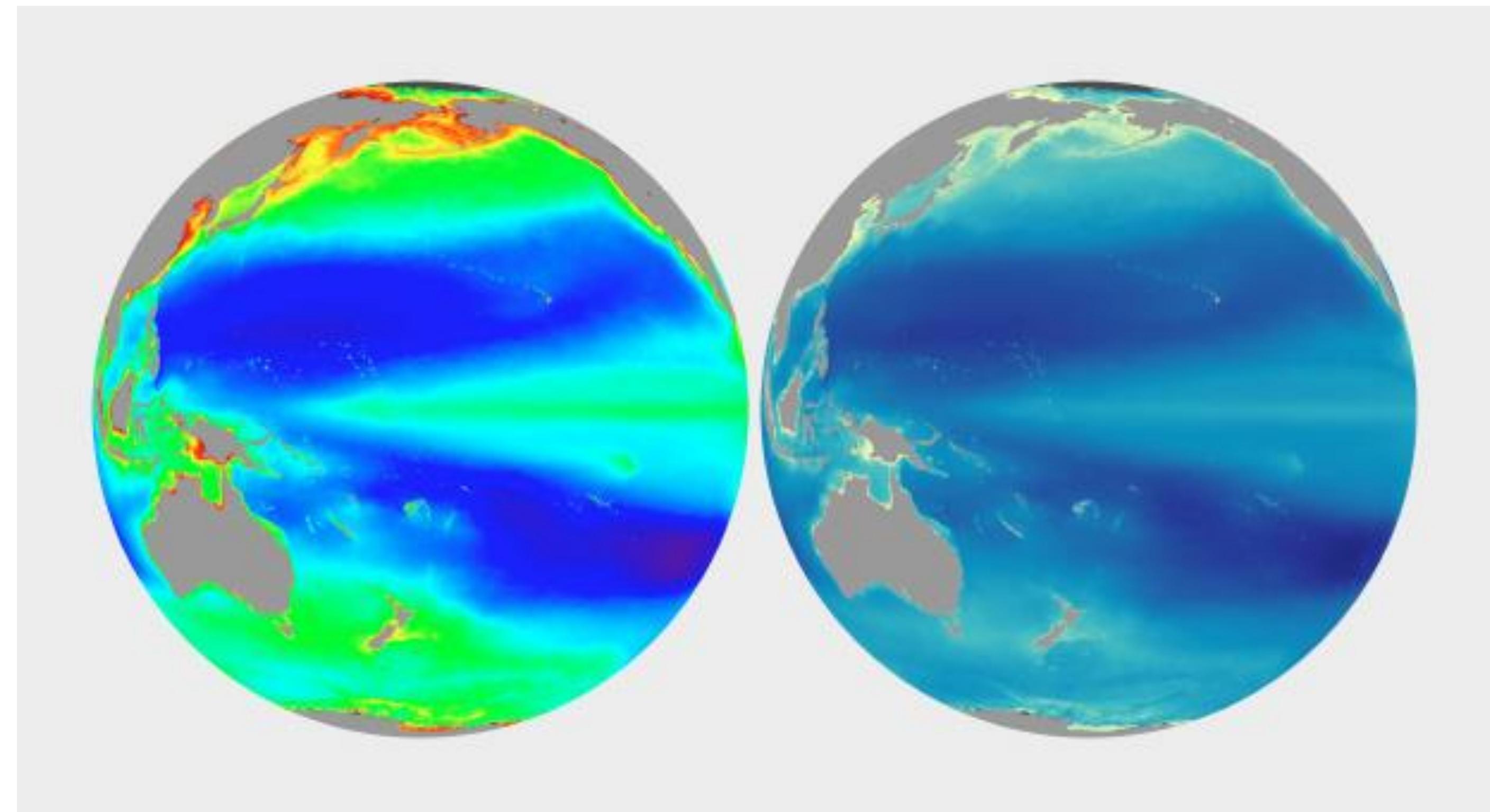
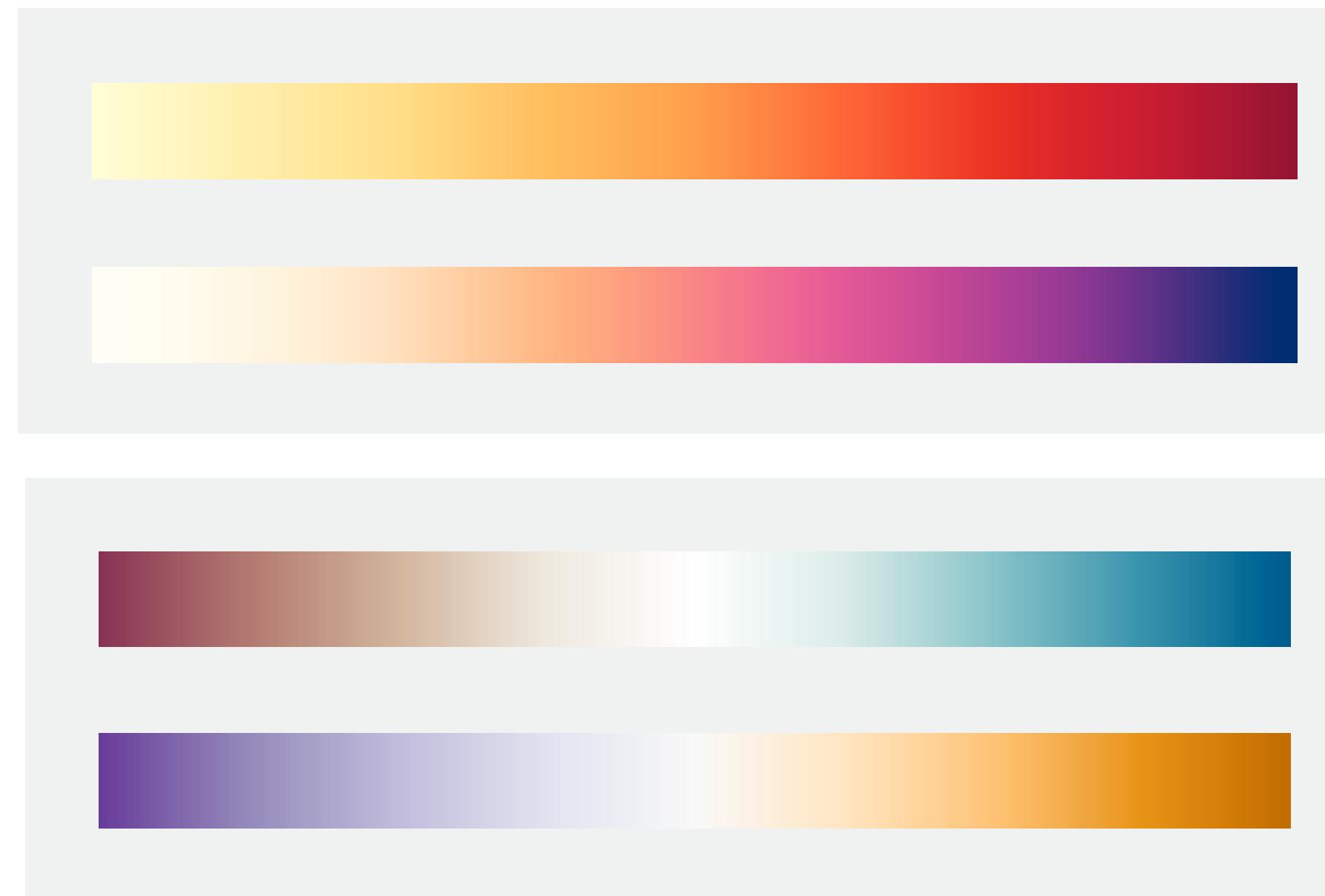


Joshua Stevens, from: www.joshuastevens.net



Perceptually linear scales

- Cuidado al usar presets
- Una escala de color adecuada debe variar de forma consistente a través del rango de valores
- Usar escalas **perceptualmente correctas**

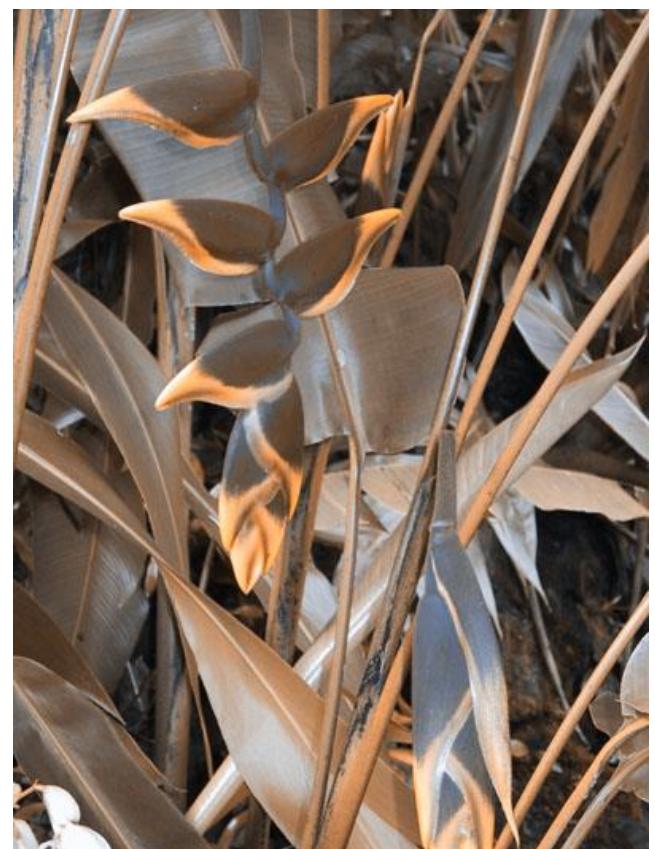


Simmon, 2013

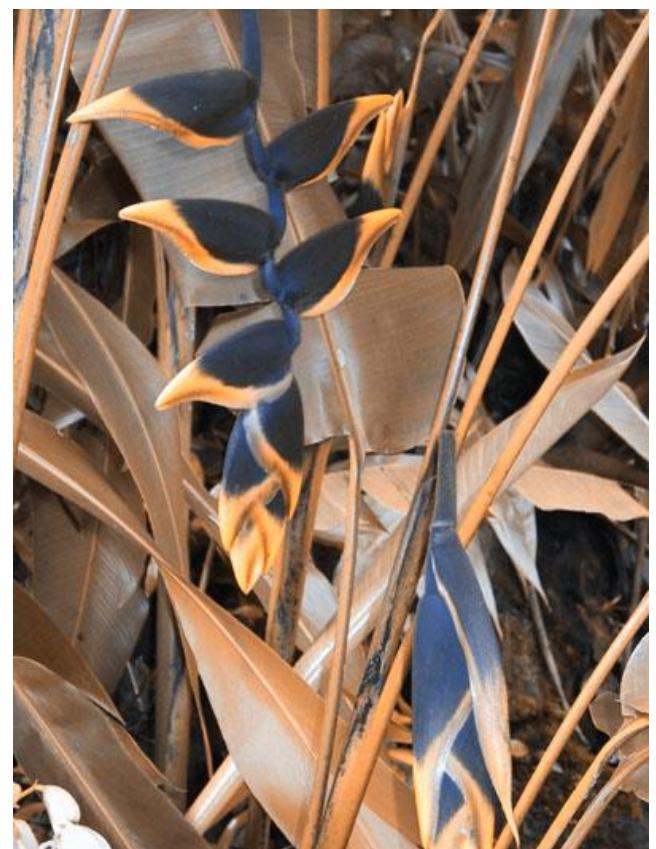
Diseñar para color deficiency: Usar simuladores



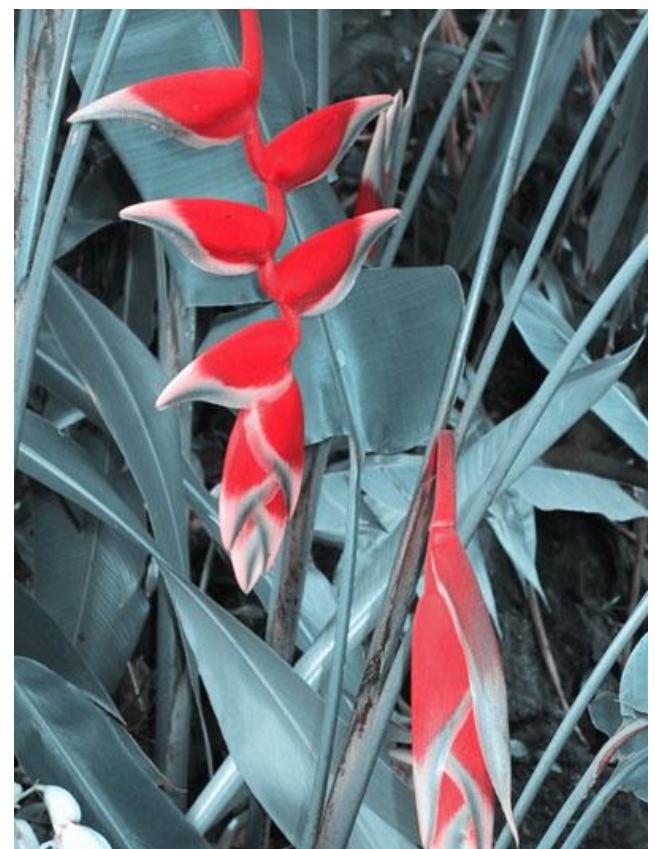
Normal vision



Deutanope



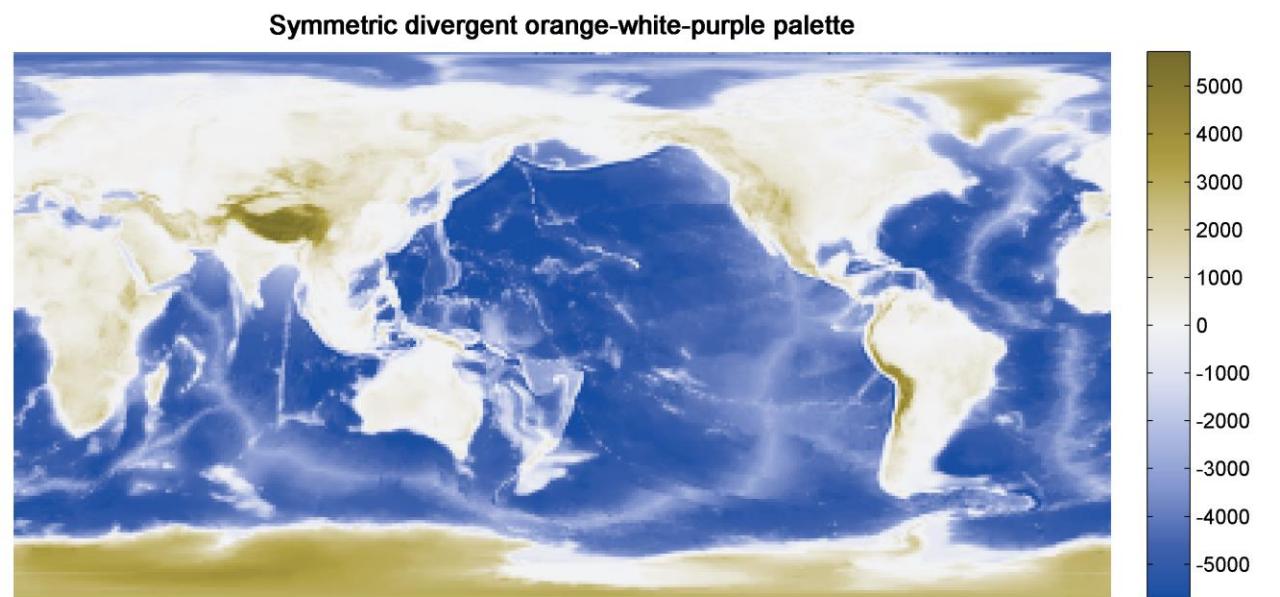
Protanope



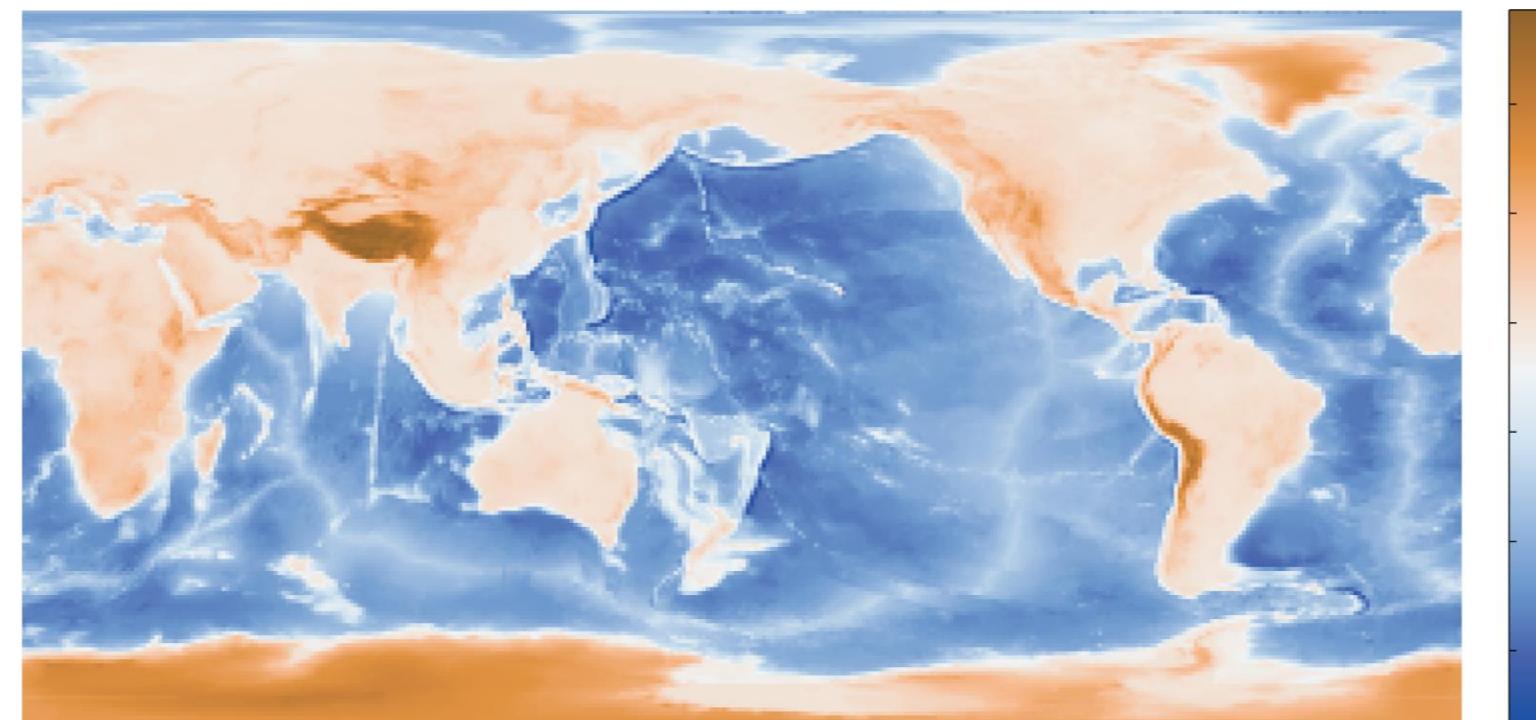
Tritanope

[Seriously Colorful: Advanced Color Principles & Practices. Stone.Tableau Customer Conference 2014.]

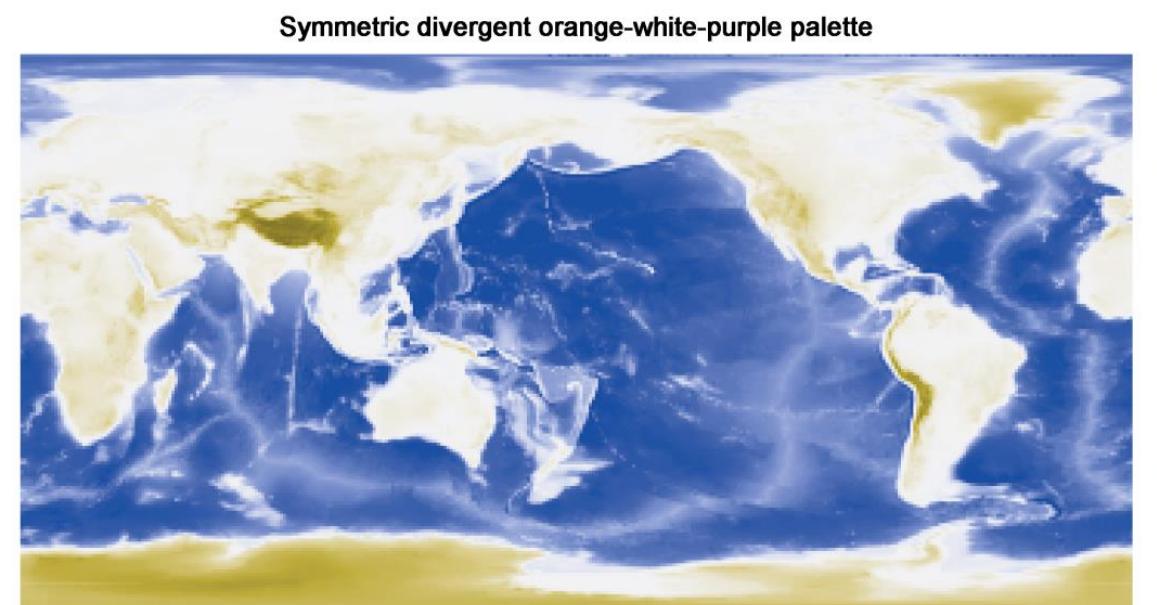
Protanope



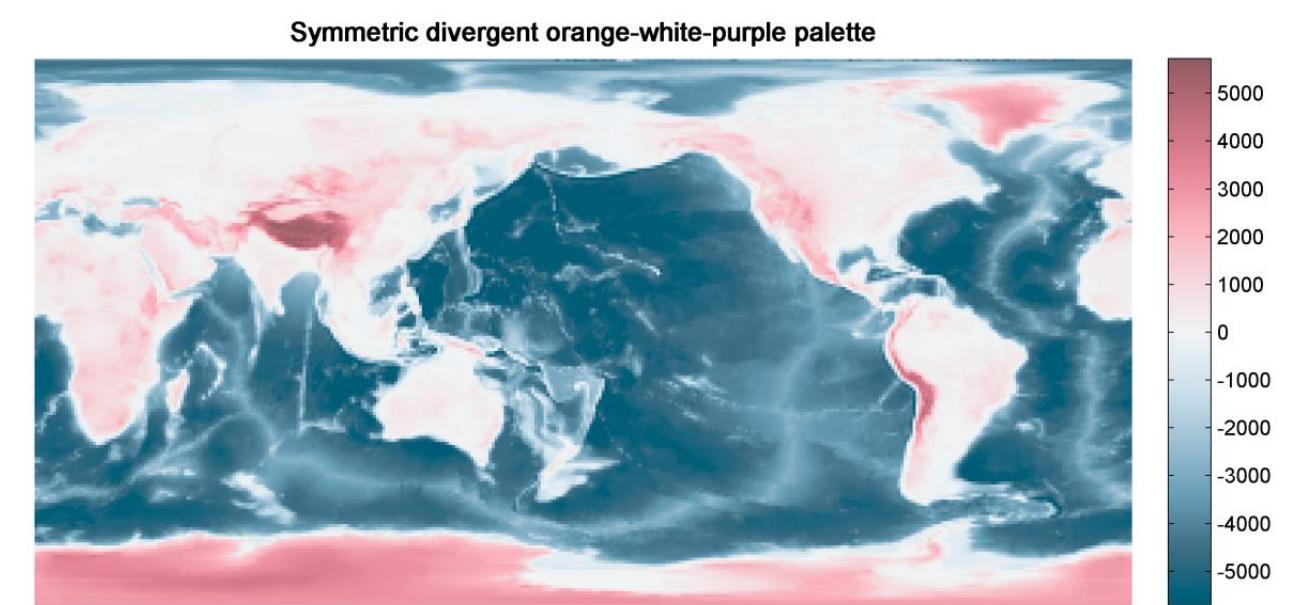
Non-symmetric divergent orange-white-purple palette



Deutanope

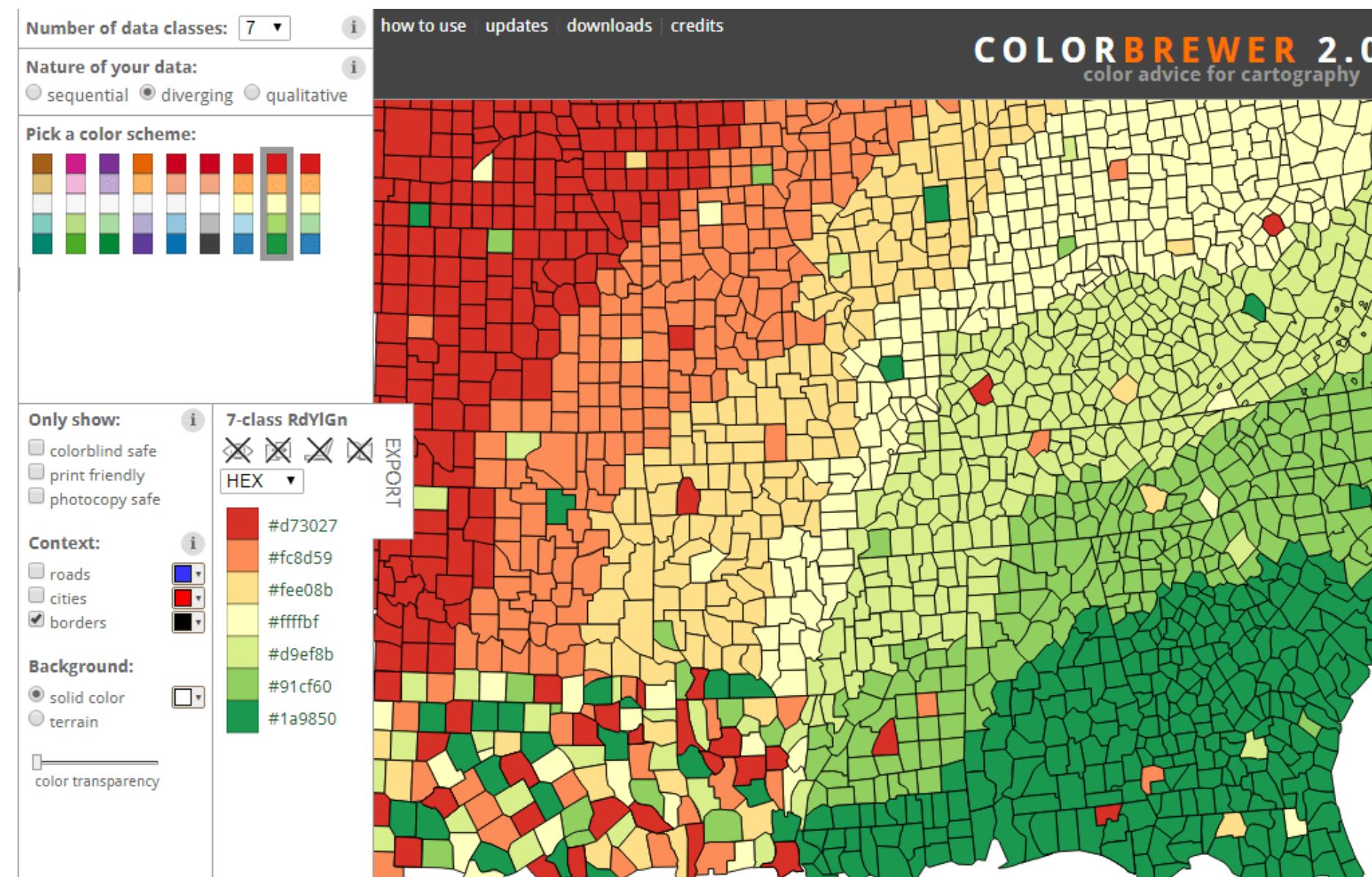


Tritanope



Simuladores de color deficiency

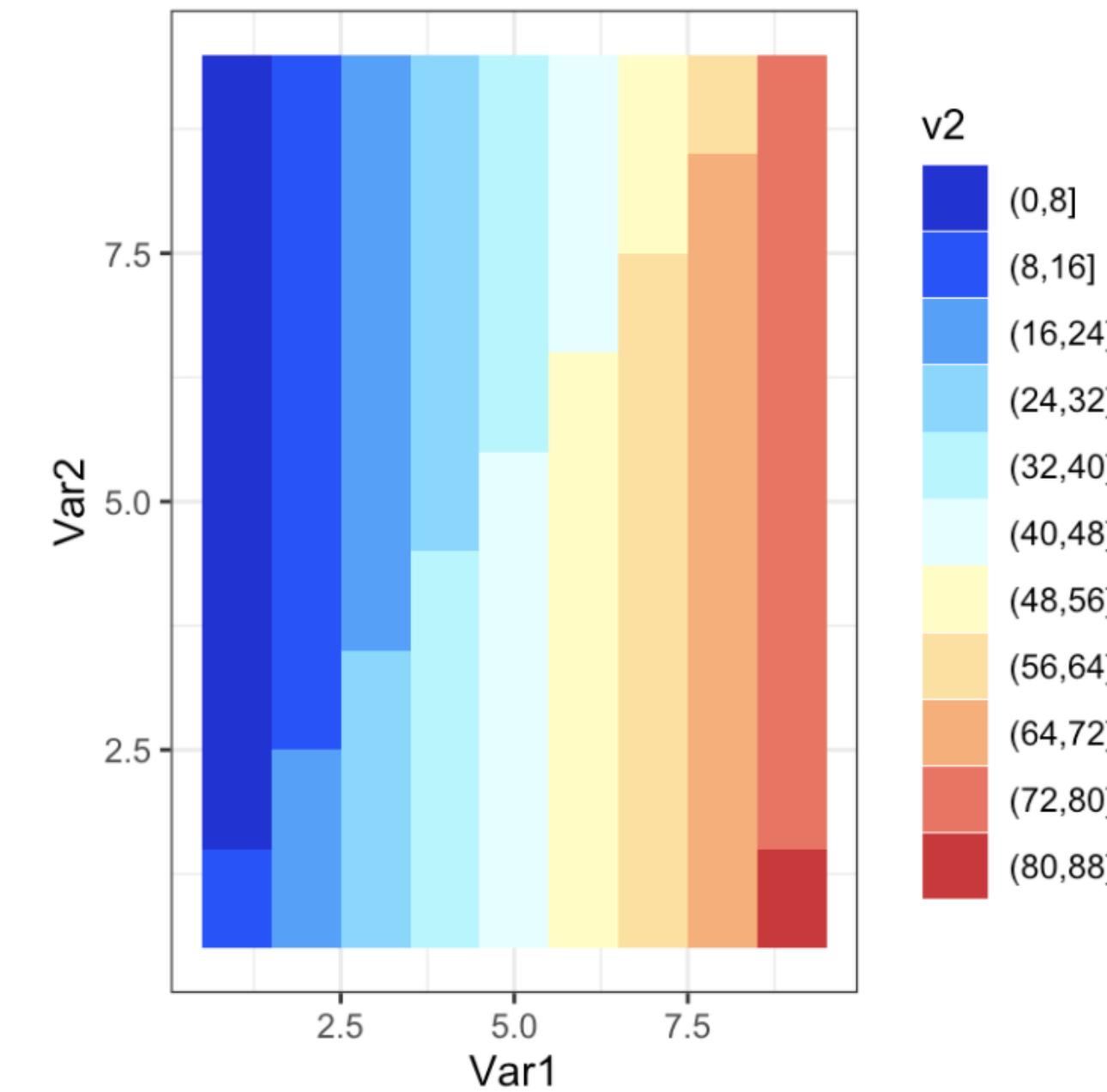
- Colorblindness para R
- Coblis www.color-blindness.com
- Color Brewer 2.0



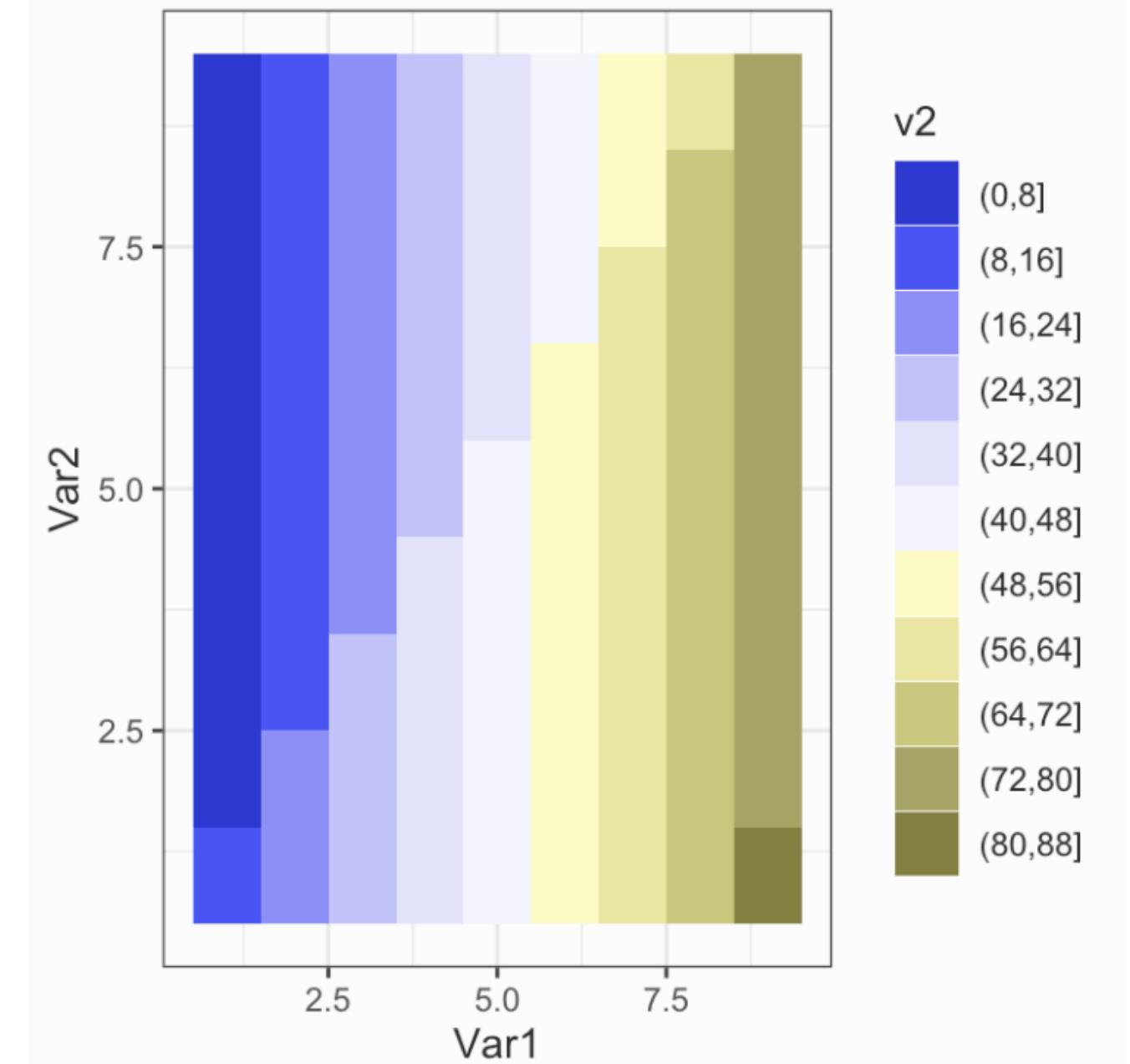
```
library(colorBlindness)
mat <- matrix(1:81, nrow = 9, ncol = 9)

library(ggplot2)
library(reshape2)
mat1 <- melt(t(mat[9:1, ]))
len <- length(Blue2DarkRed12Steps)-1
mat1$v2 <- cut(mat1$value,
                 breaks = seq(0,ceiling(81/len)*len,
                               length.out = len+1))
ht <- ggplot(mat1) +
  geom_tile(aes(x=Var1, y=Var2, fill=v2)) +
  scale_fill_manual(values=Blue2DarkRed12Steps) +
  theme_bw()
# check the plot by CVD simulator
cvdPlot(ht)
```

normal vision



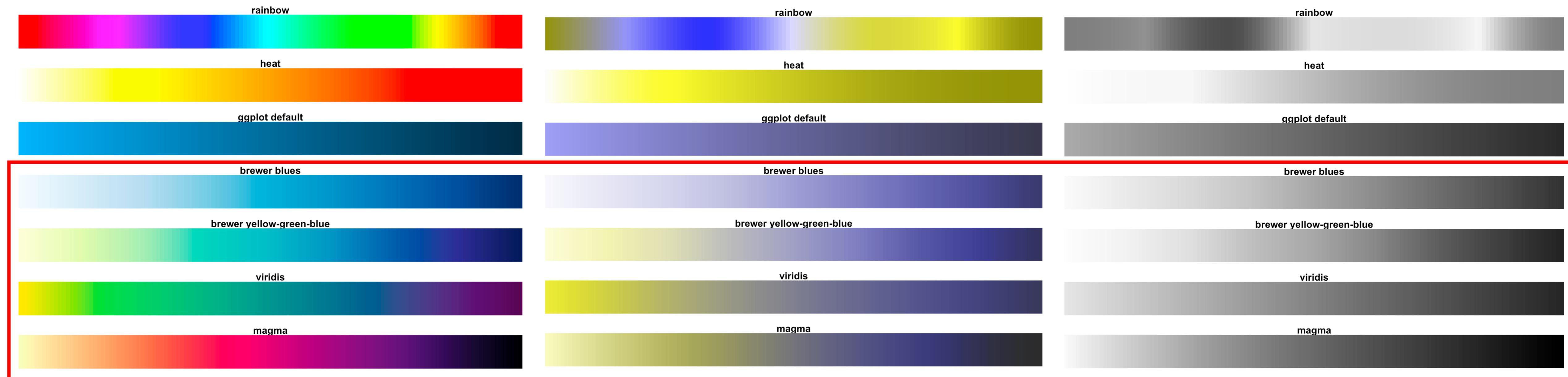
deuteranopia (6%)



Escalas perceptualmente correctas

Viridis. Perceptualmente correctas; corregidas para color deficiency y B/N

<https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html>



Buscar “Perceptually correct”

Viridis color scales (R, plotly, etc.)

The package contains eight color scales: “viridis”, the primary choice, and five alternatives with similar properties - “magma”, “plasma”, “inferno”, “cividis”, “mako”, and “rocket” -, and a rainbow color map - “turbo”.

The color maps `viridis`, `magma`, `inferno`, and `plasma` were created by Stéfan van der Walt ([@stefanv] (<https://github.com/stefanv>)) and Nathaniel Smith ([@njsmith] (<https://github.com/njsmith>)). If you want to know more about the science behind the creation of these color maps, you can watch this [presentation of `viridis`](#) by their authors at [SciPy 2015](#).

The color map `cividis` is a corrected version of 'viridis', developed by Jamie R. Nuñez, Christopher R. Anderton, and Ryan S. Renslow, and originally ported to R by Marco Sciaiani ([@msciain] (<https://github.com/marcoscii>)). More info about `cividis` can be found in [this paper](#).

The color maps `mako` and `rocket` were originally created for the Seaborn statistical data visualization package for Python. More info about `mako` and `rocket` can be found on the Seaborn [website](#).

The color map turbo was developed by Anton Mikhailov to address the shortcomings of the Jet rainbow color map such as false detail, banding and color blindness ambiguity. More info about turbo can be found [here](#).

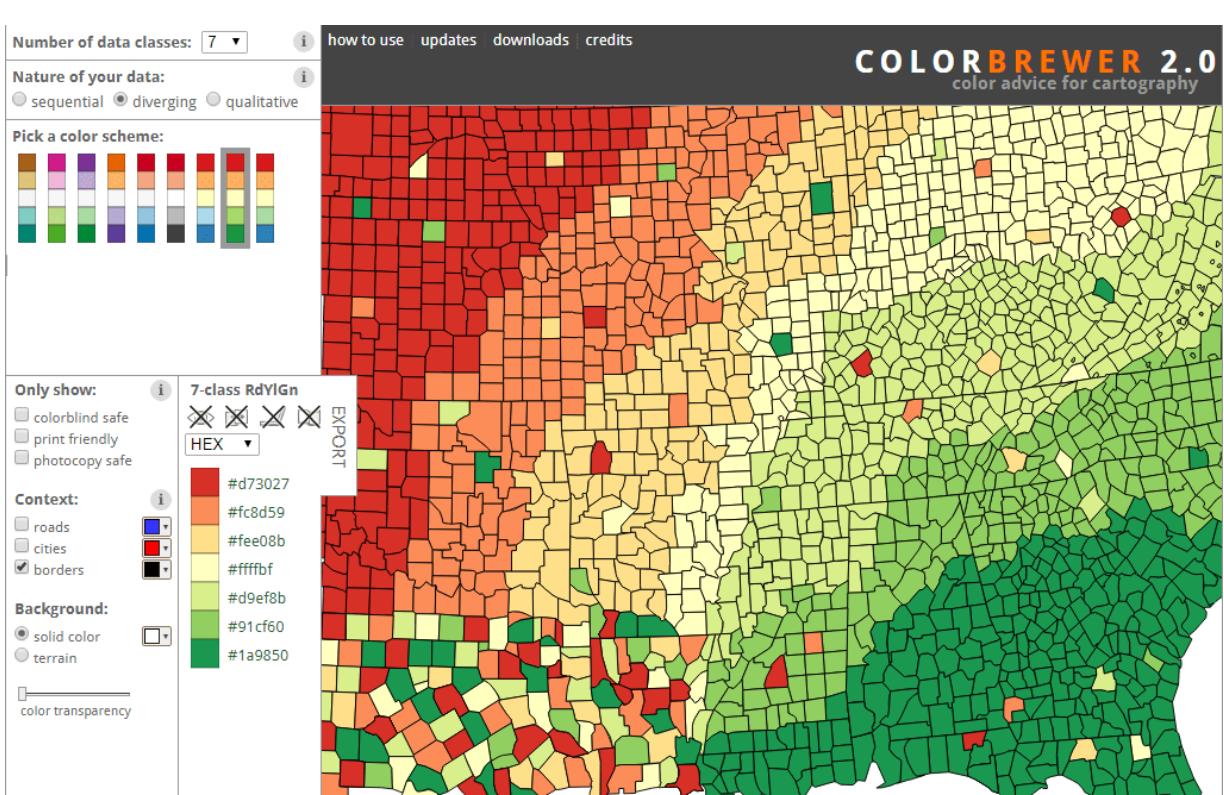
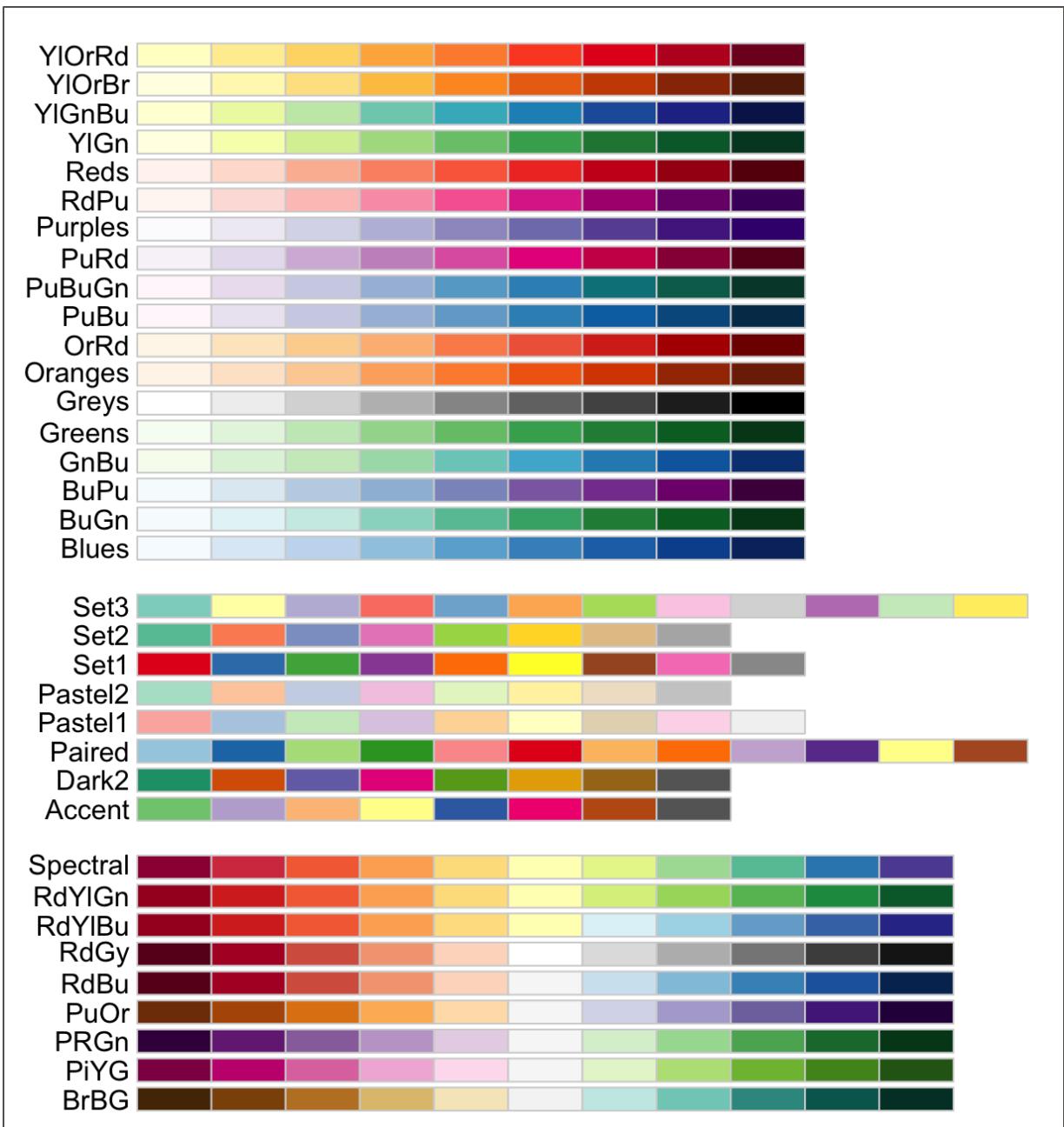


R-colorspace

```
library("colorspace")
hcl_palettes(plot = TRUE)
```



Color Brewer



R cheatsheet

- No se trata solo de estética
- Usar esquemas de color que funcionen sobre sistemas basados en percepción: HSV, HSL, CIEL
- Accesibilidad: tener en cuenta color-blindness
- Elegir paletas de color en función del tipo de datos (categórico, cuantitativo, ordinal // secuencial, divergente, cílico)

R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0
i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbaa)

R has 657 built in color names

Example:

To see a list of names:

`colors()`

`peachpuff4`

These colors are displayed on P. 3.

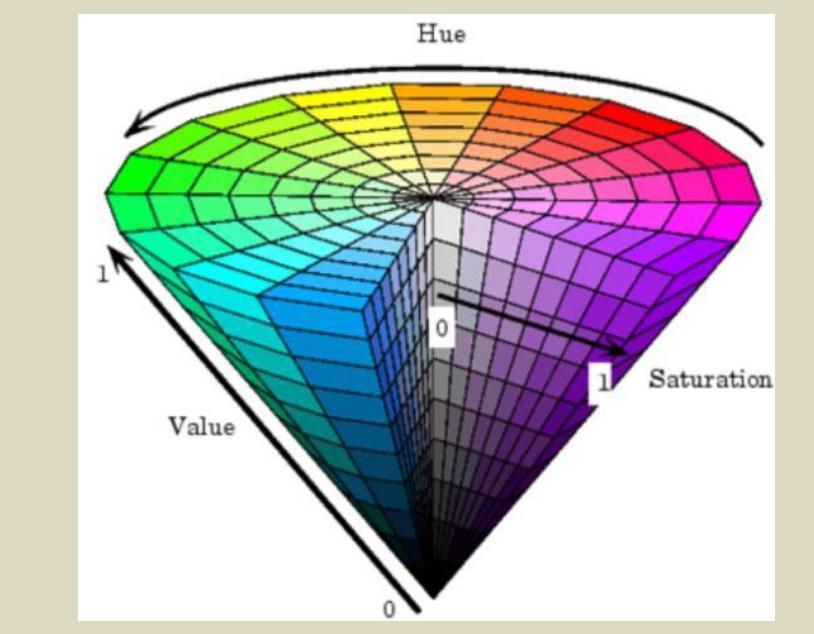
R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.
`rgb(r, g, b, maxColorValue=255, alpha=255)`
- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument
`hsv(h, s, v, alpha)`
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
`hcl(h, c, l, alpha)`

A few notes on HSV/HLC

HSV is a better model for how humans perceive color. HCL can be thought of as a perceptually based version of the HSV model....blah blah blah...

Without delving into color theory: color schemes based on HSV/HLC models generally just look good.



R can translate colors to rgb (this is handy for matching colors in other programs)
`col2rgb(c("#FF0000", "blue"))`

R Color Palettes

This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure....NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps

grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

`heat.colors(4, alpha=1)`

`> #FF0000FF "#FF8000FF" "#FFFF00FF" "#FFFF80FF"`

For the rainbow palette you can also select start/end color (red = 0, yellow = 1/6, green = 2/6, cyan = 3/6, blue = 4/6 and magenta = 5/6) and saturation (s) and value (v):
`rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)`

`grDevices palettes`
`cm.colors`
`topo.colors`
`terrain.colors`
`heat.colors`
`rainbow`
see P. 4 for options

Package: RcolorBrewer

This function has an argument for the number of colors and the color palette (see P. 4 for options).

`brewer.pal(4, "Set3")`

`> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"`

To view colorbrewer palettes in R: `display.brewer.all(5)`

There is also a very nice interactive viewer:
<http://colorbrewer2.org/>

My Recommendation

Package: colorspace

These color palettes are based on HCL and HSV color models. The results can be very aesthetically pleasing. There are some default palettes:

`rainbow_hcl(4)`

`"#E495A5" "#ABB065" "#39BEB1" "#ACA4E2"`

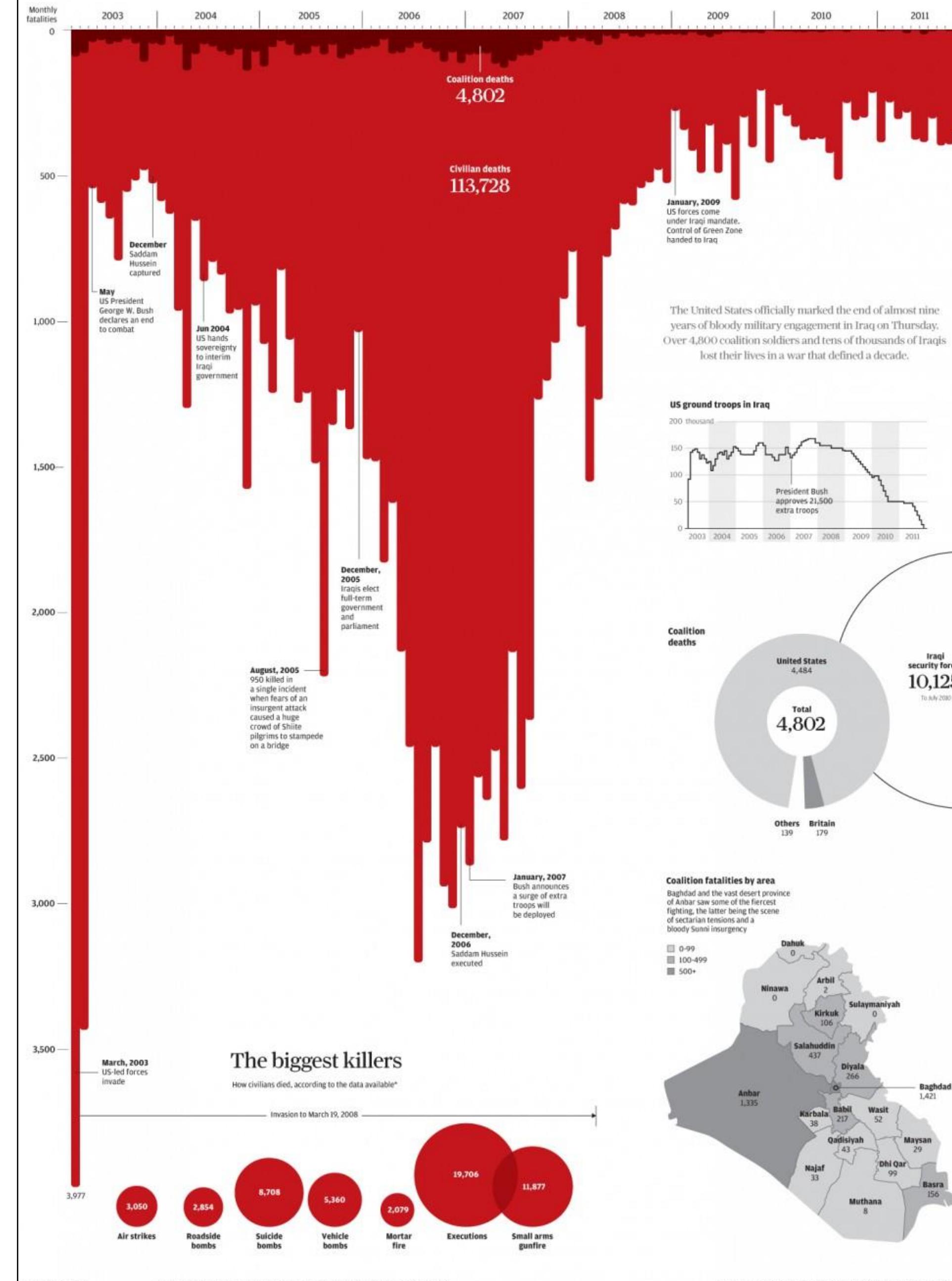
`colorspace`
`default palettes`
`diverge_hcl`
`diverge_hsl`
`terrain_hcl`
`sequential_hcl`
`rainbow_hcl`

However, all palettes are fully customizable:

`diverge_hcl(7, h = c(246, 40), c = 96, l = c(65, 90))`

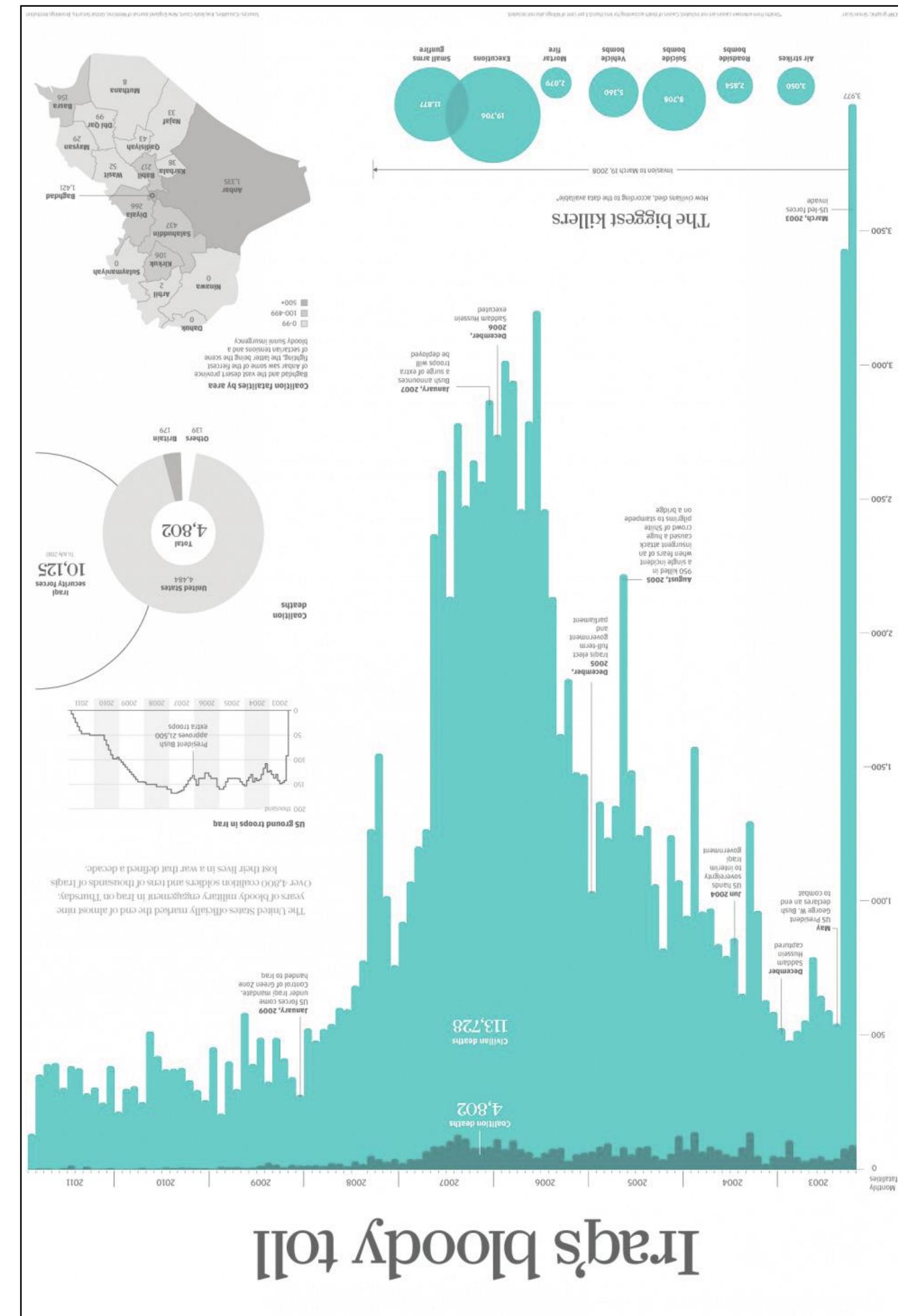
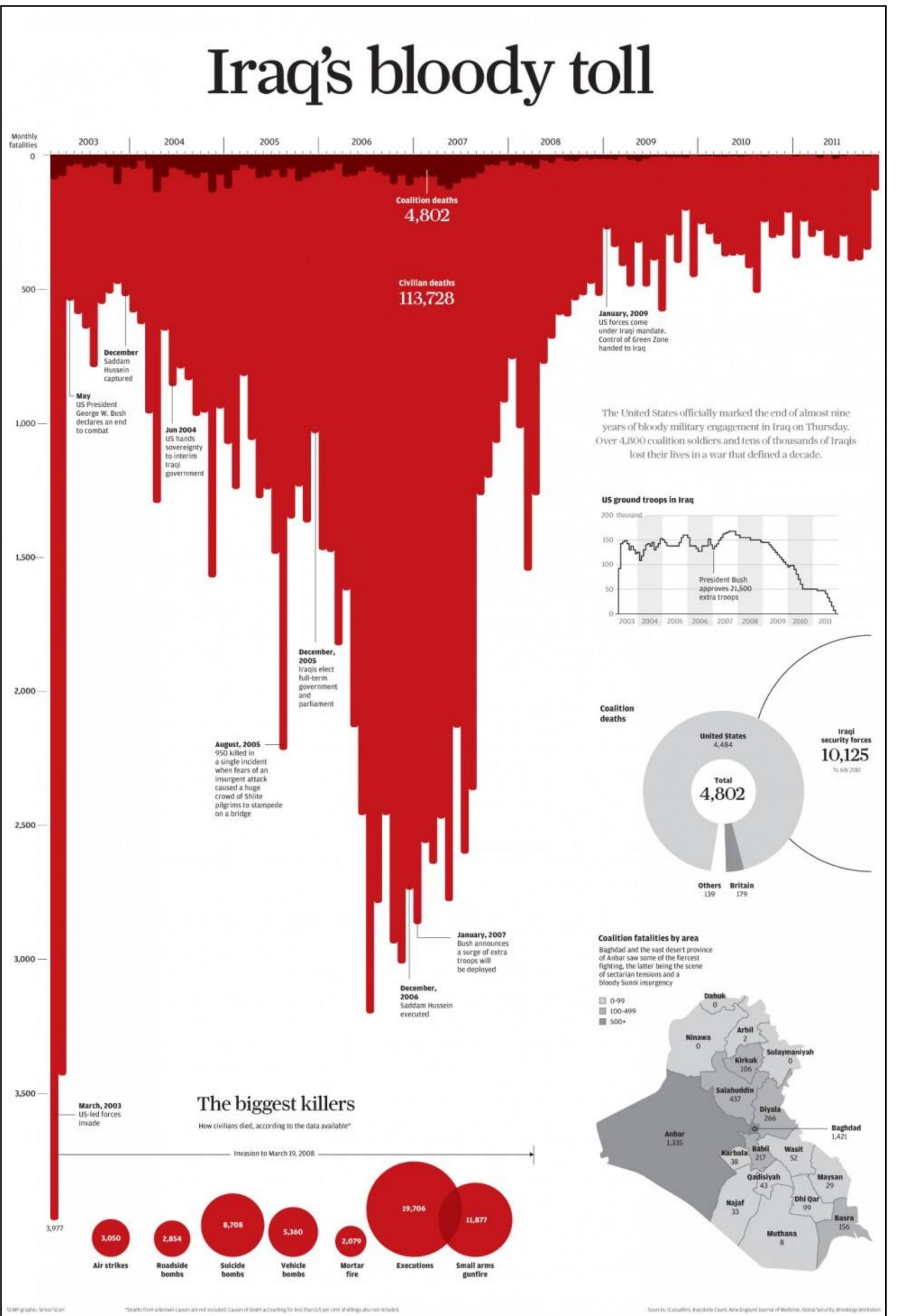
Choosing the values would be daunting. But there are some recommended palettes in the colorspace documentation. There is also an interactive tool that can be used to obtain a customized palette. To start the tool:
`pal <- choose_palette()`

Iraq's bloody toll

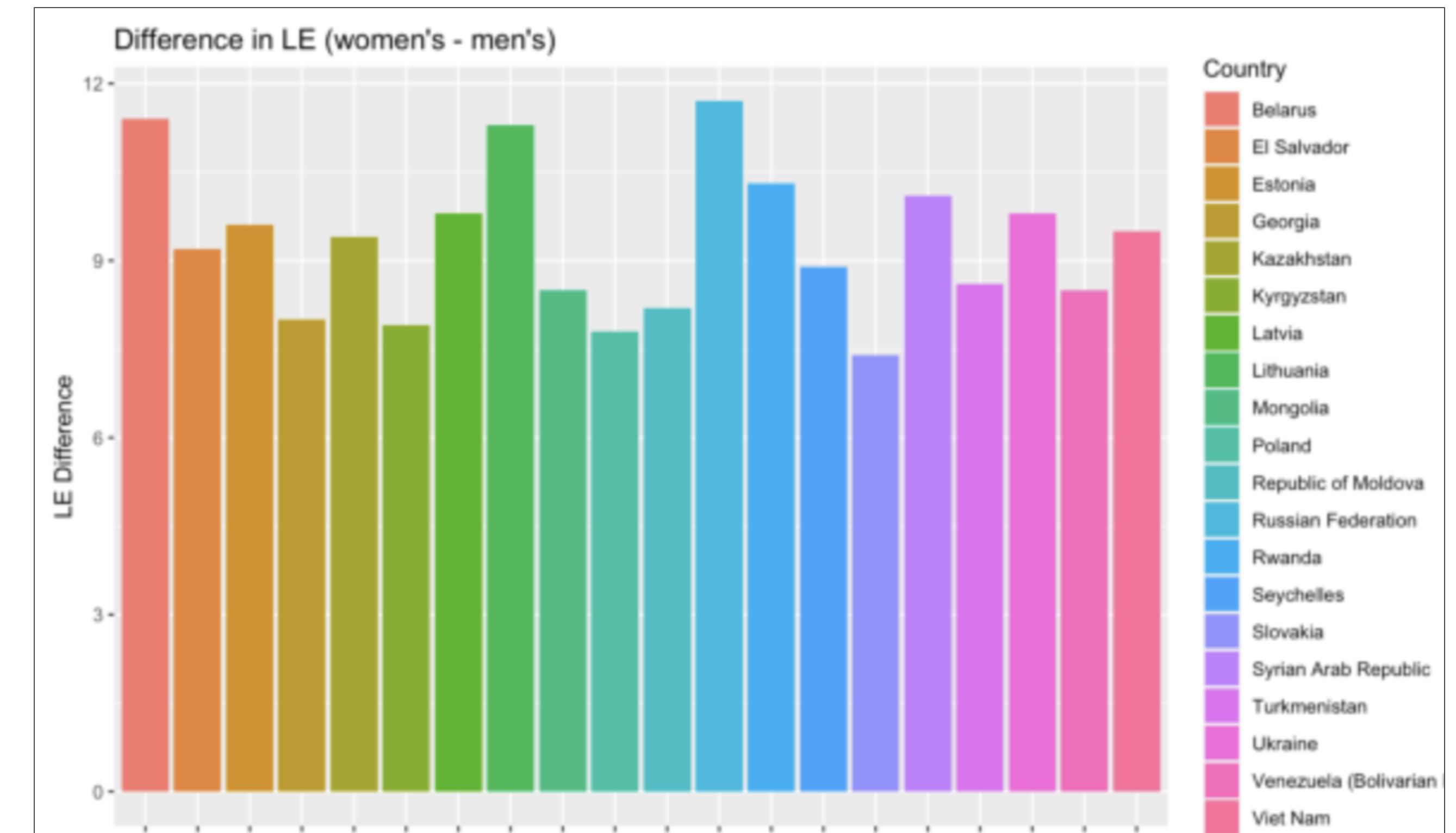


Semántica del color

Semántica del color



- Gráficas de diferencia en Esperanza de Vida al Nacer – Top 20
- ¿Uso del color correcto?
- ¿Qué fallos tienen?



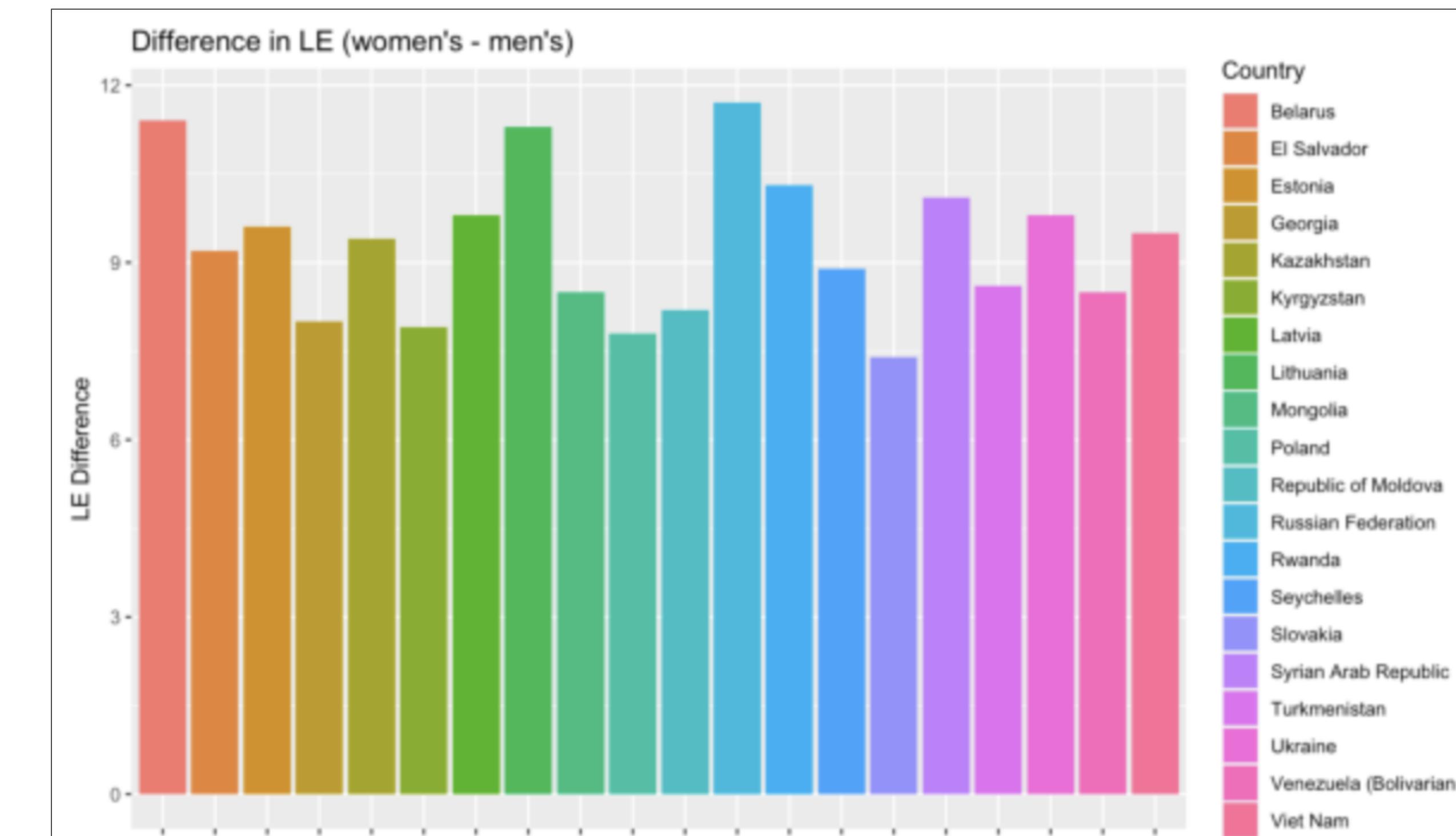
Do's ...

- Escala bicromática continua
- Variación en tono y luminosidad
- Escala **continua** usada para **attr. cuantitativo continuo**
- Doble redundancia: Ranking ordenado por tamaño de barras, color



... and dont's

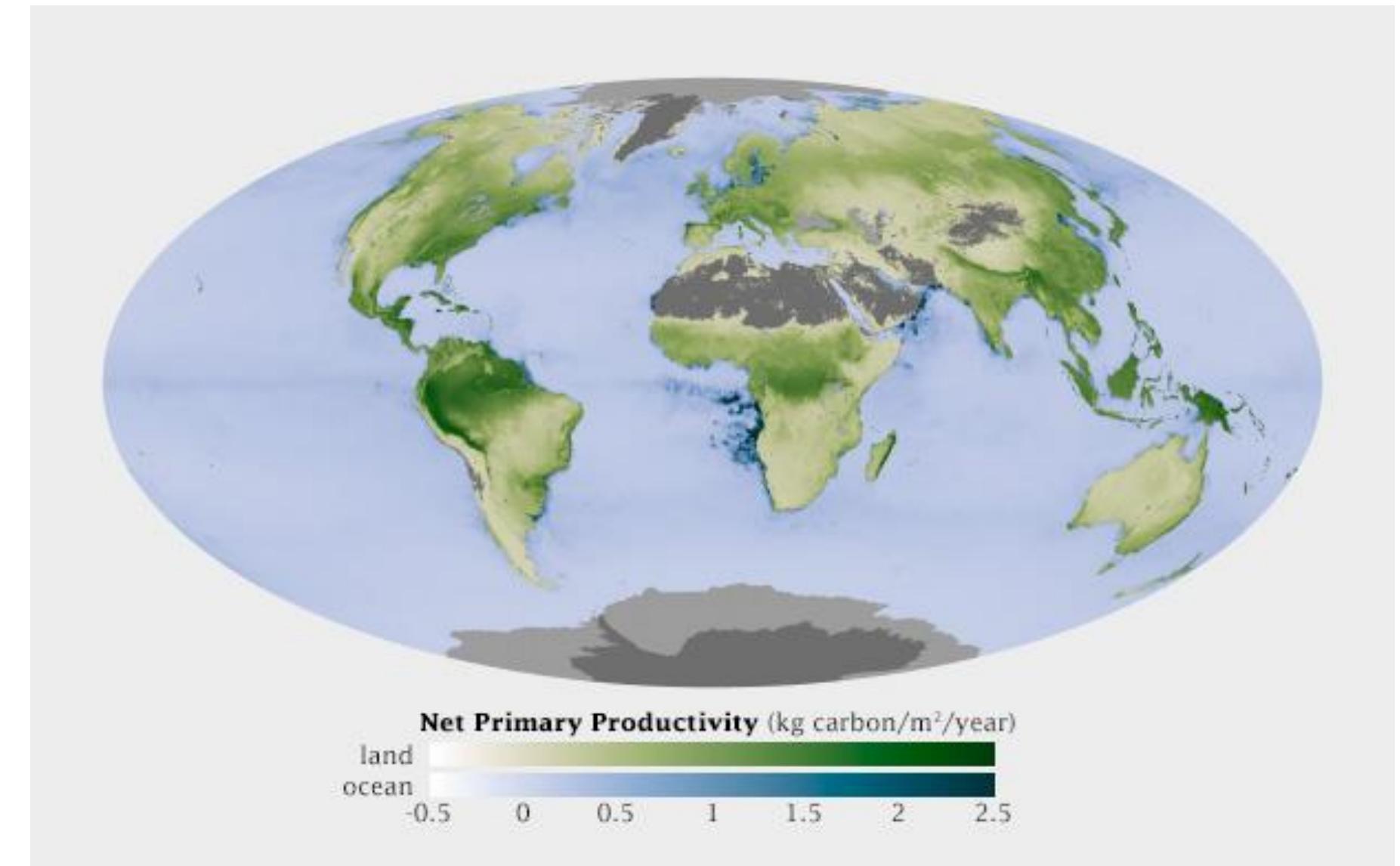
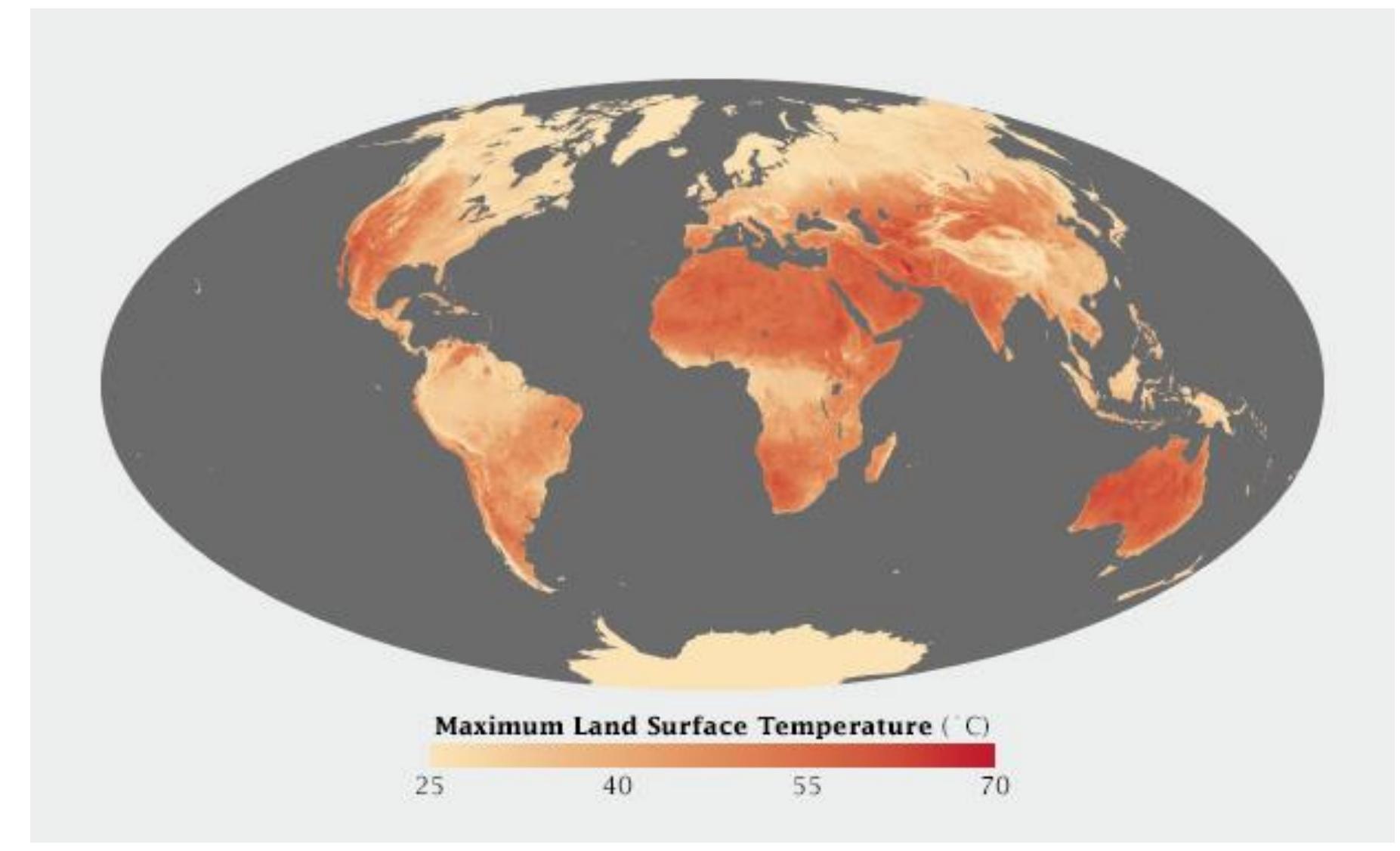
- Escala arcoíris, aunque con luminancia corregida
- Escala **continua** usada como **categórica**
- + de 12 categorías
- Ranking no ordenado
- Etiquetas del barchart fuera de la gráfica



Buenas prácticas

Robert Simmon, 2013

- Asignar color a significado (paletas cálidas para altas temperaturas; divergente azul-rojo para temp negativa a positiva, verde para datos de vegetación, etc)
- Usar la escala correcta para cada tipo de atributo:
 - Si es categórico o cuantitativo/ordinal y
 - si es divergente, secuencial, o cíclico
- Elegir paletas con contraste en luminancia para resaltar detalles.
- Tener en cuenta color-blindness y usos (imprimir, colores web, etc)
- No data, no color



Bibliografia

Color:

- *Subtleties of color:* <https://earthobservatory.nasa.gov/blogs/elegantfigures/2013/08/05/subtleties-of-color-part-1-of-6/>
- *Tableau color palettes:* <https://www.tableau.com/about/blog/2016/7/colors-upgrade-tableau-10-56782>
- *Bi-variate choropleth maps:* www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/
- *R colorspace package:* <https://cran.r-project.org/web/packages/colorspace/vignettes/colorspace.html>

Inspiración

<http://www.thefunctionalart.com/>

<https://eagereyes.org/>

<http://flowingdata.com/>

<http://fivethirtyeight.com/>

<http://truth-and-beauty.net/>



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



8.2 Honestidad y Precisión visual

Ejes

 National Review 
@NRO

The only **#climatechange** chart you need to see.
natl.re/wPKpro

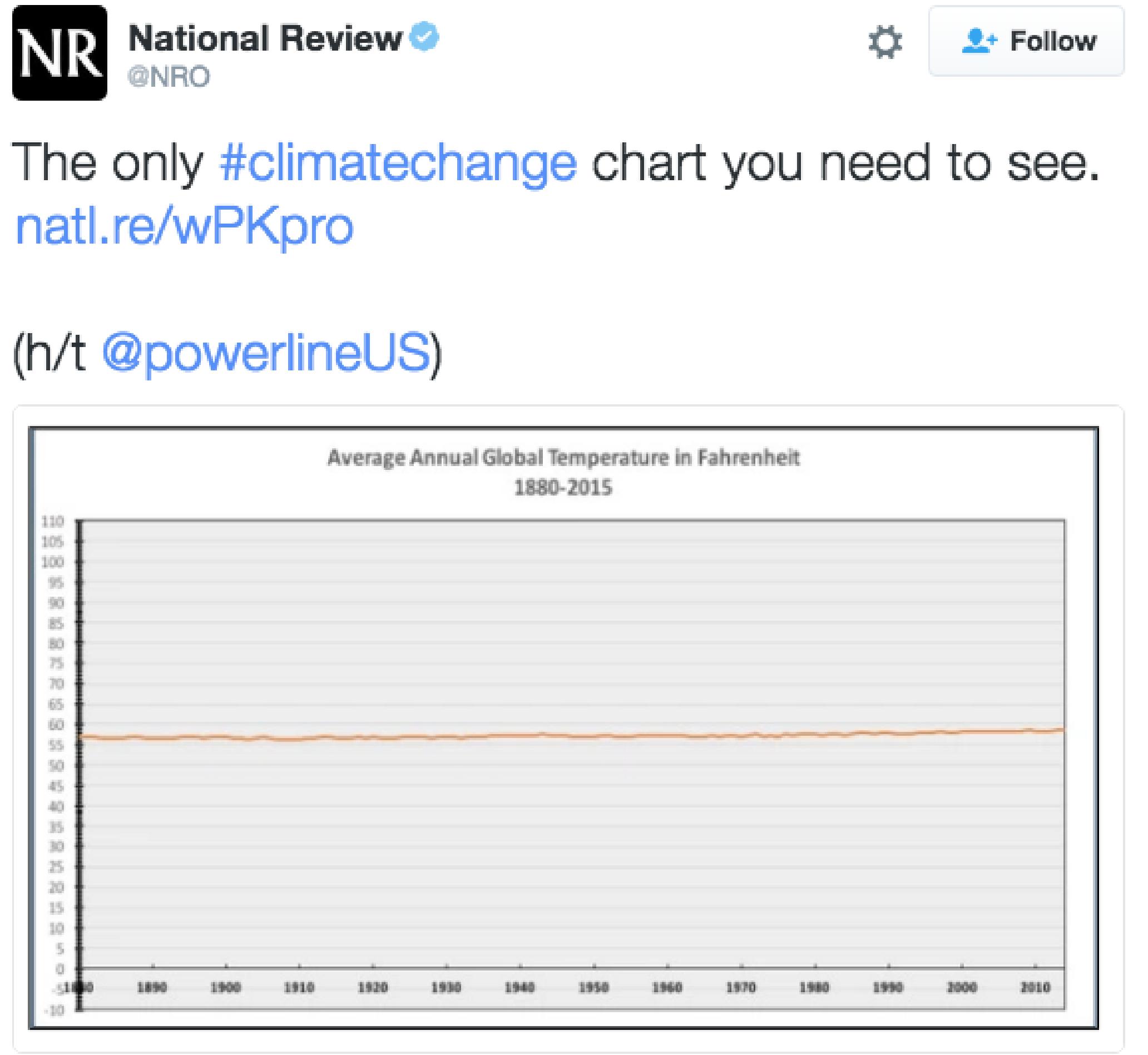
(h/t [@powerlineUS](#))

Average Annual Global Temperature in Fahrenheit
1880-2015



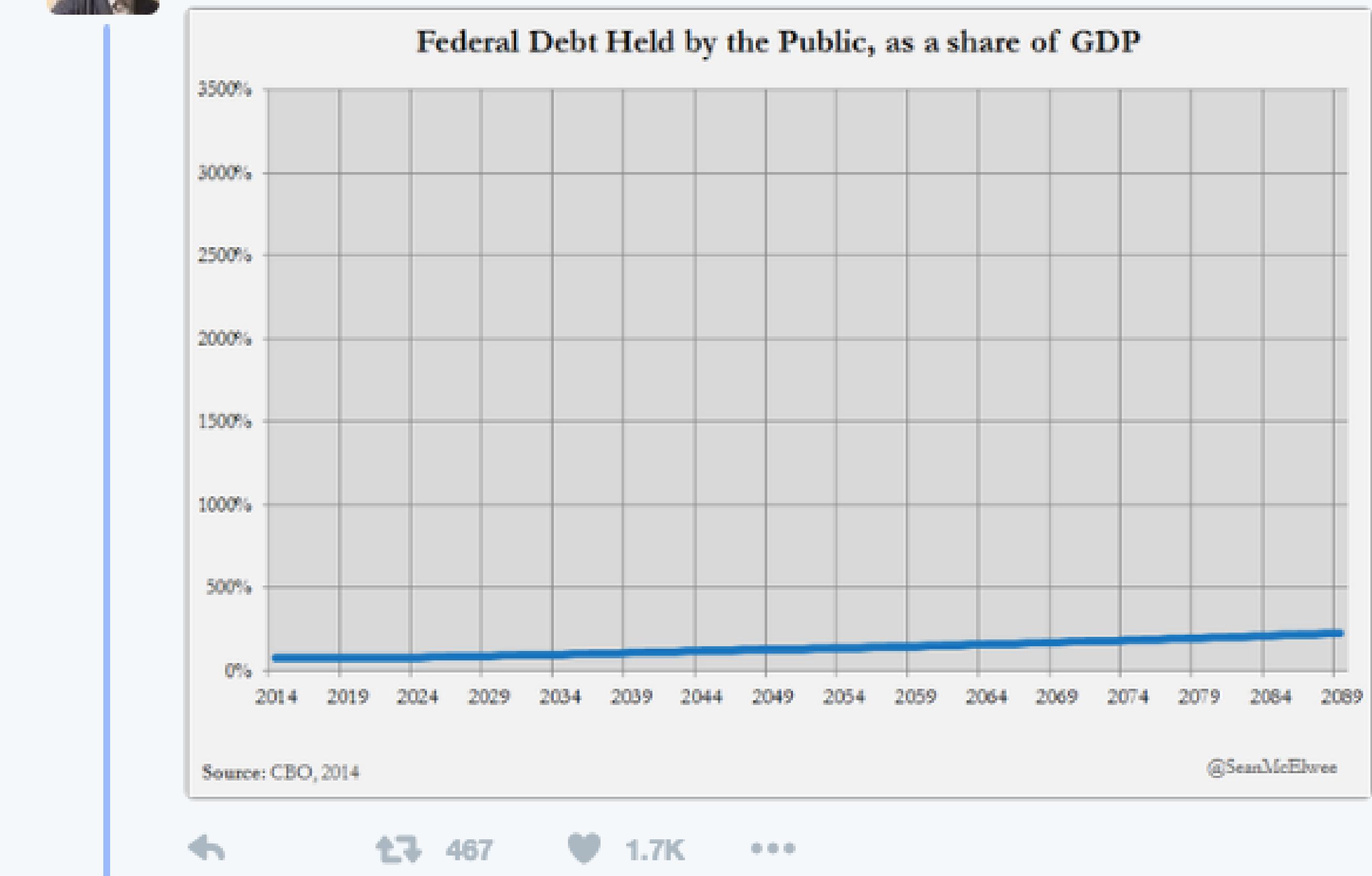
Year	Temperature (F)
1880	58
1900	58
1920	58
1940	58
1960	58
1980	58
2000	58
2010	58

Ejes



sean. @SeanMcElwee · 14 Dec 2015

@NRO @powerlineUS no need to worry about the national debt then either!



Ejes



National Review

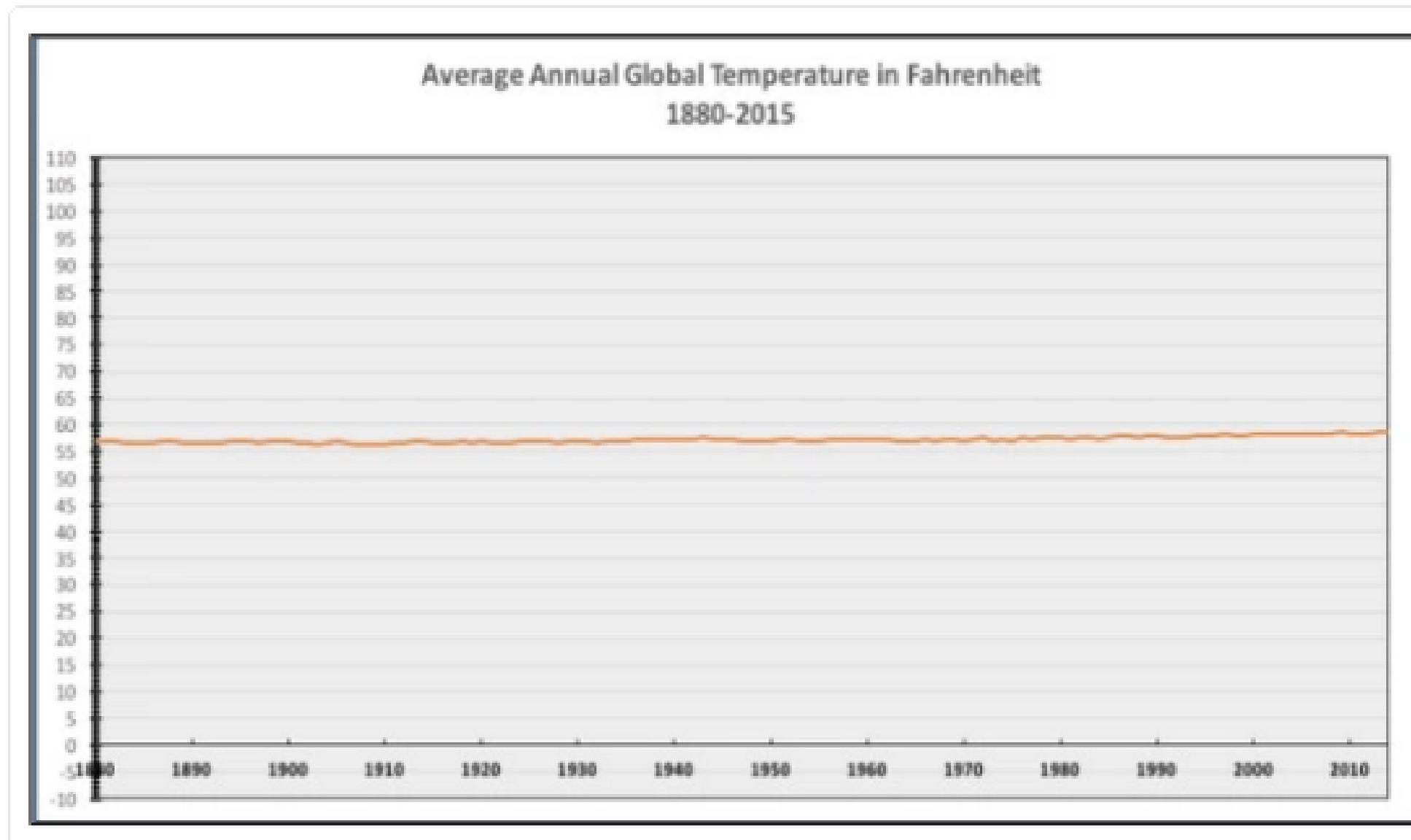
@NRO



Follow

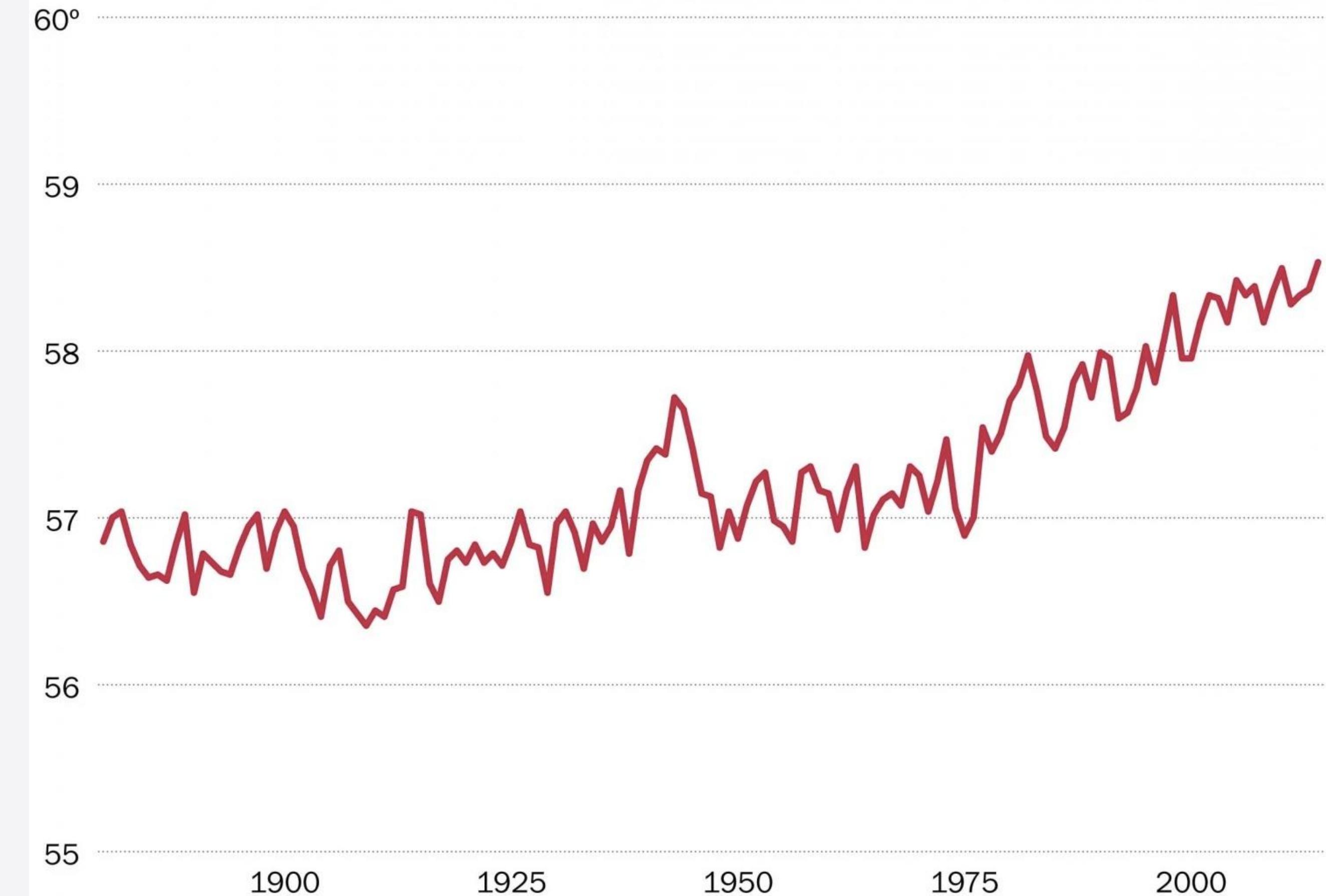
The only #climatechange chart you need to see.
natl.re/wPKpro

(h/t [@powerlineUS](#))

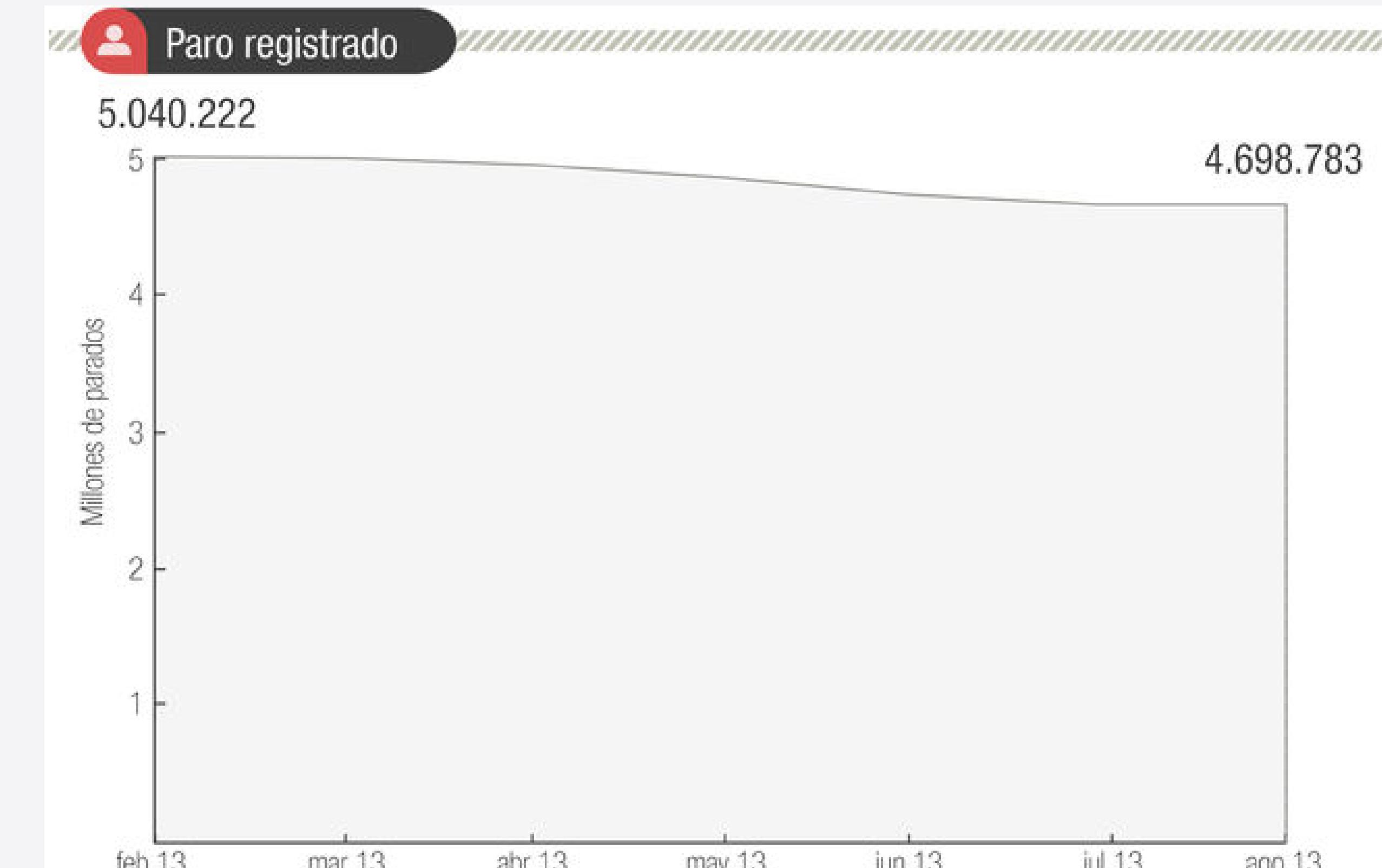


Average global temperature by year

Data from NASA/GISS.

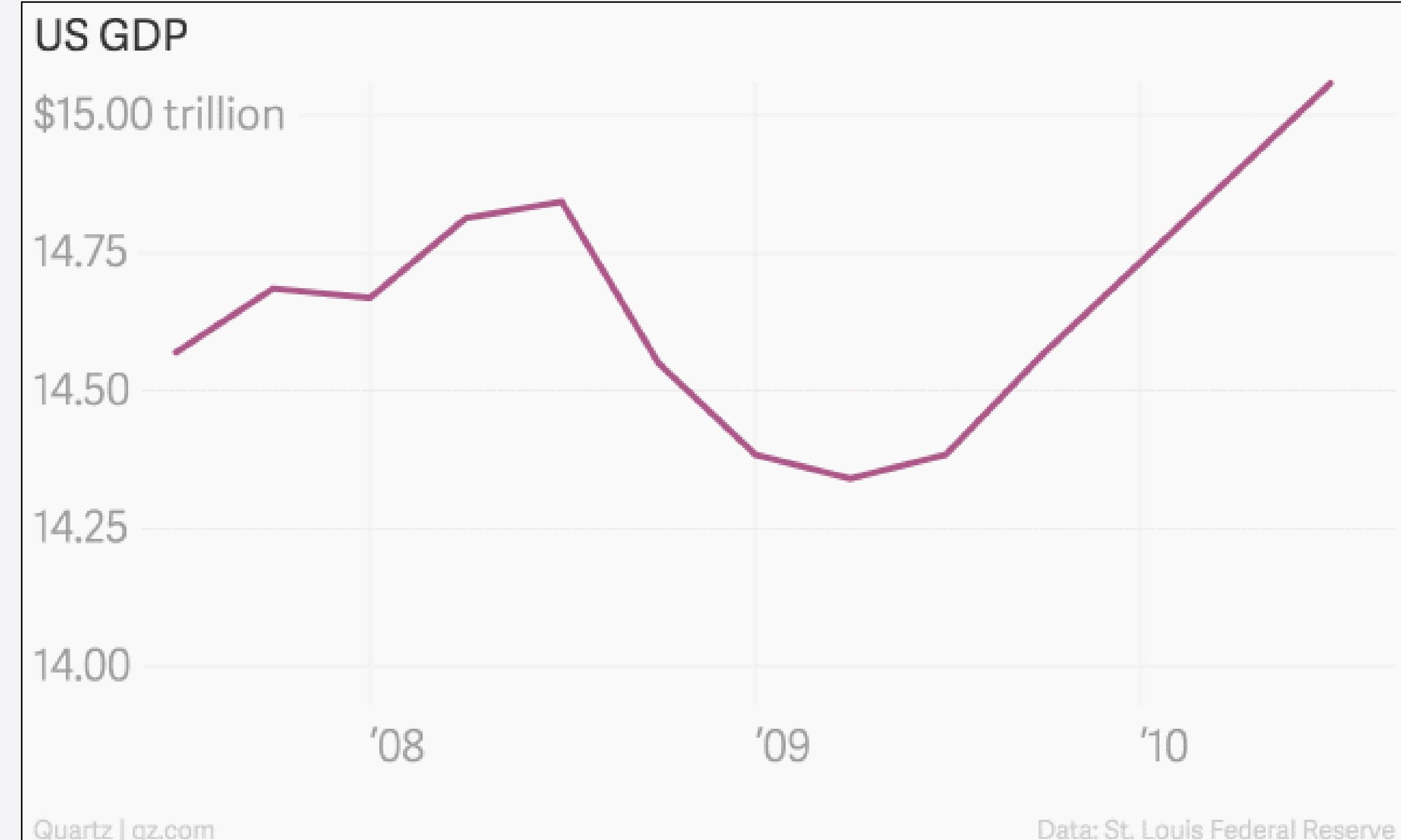
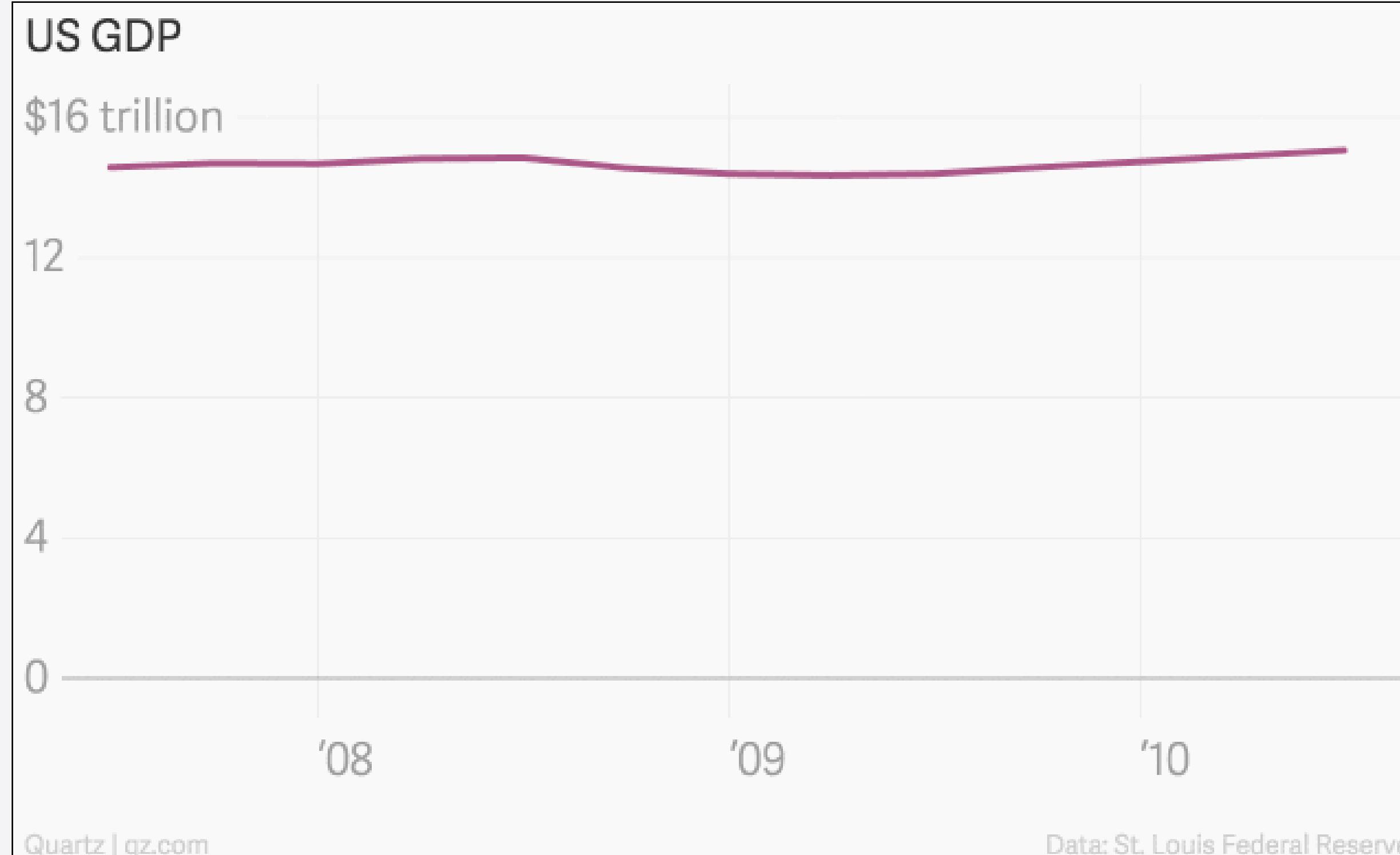


Ejes



Mostrar eje completo para no magnificar el descenso

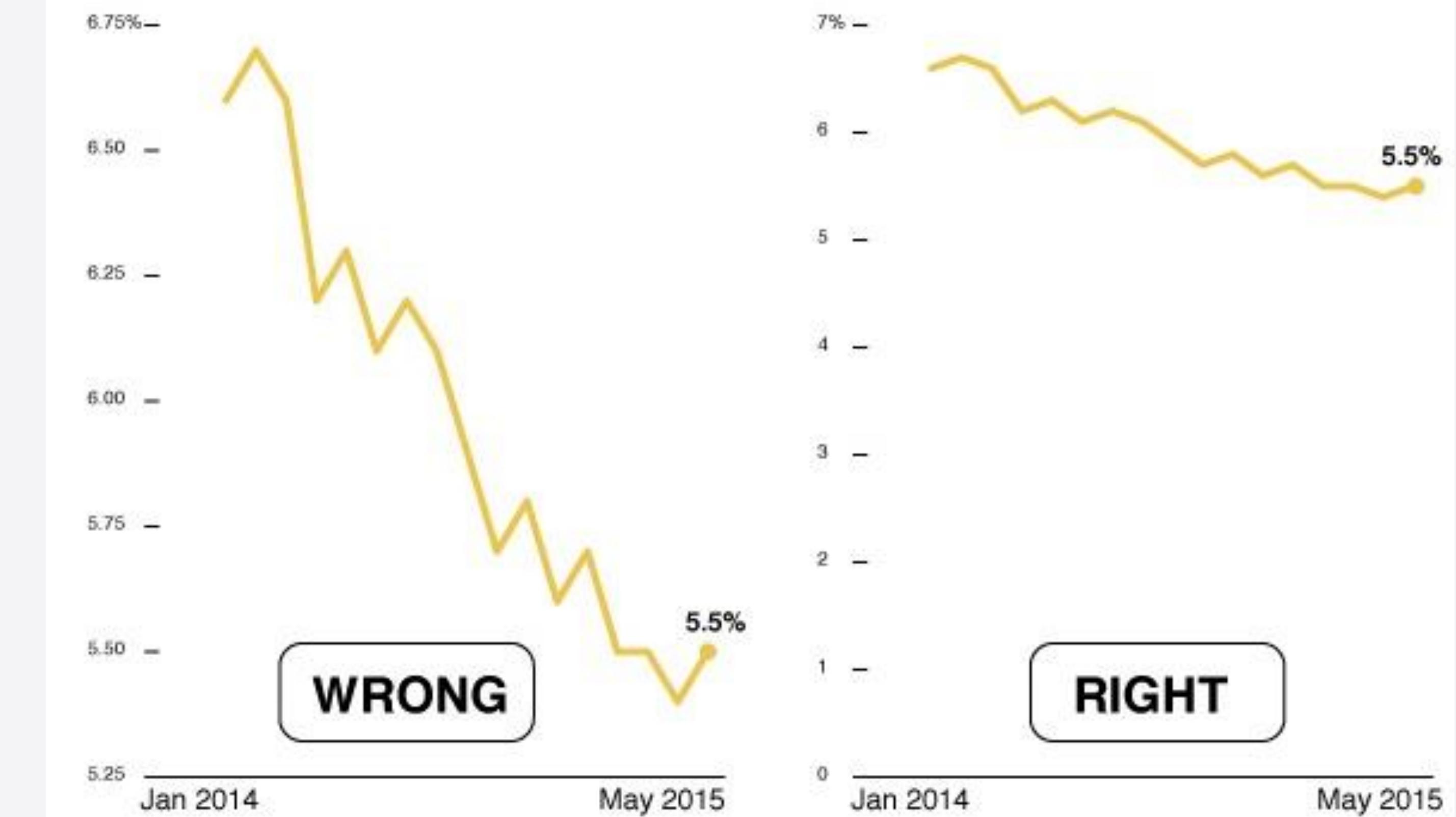
Ejes



La elección del eje puede enfatizar variación o tendencia

Ejes

- Linecharts codifican POSICIÓN de puntos
- Barras codifican TAMAÑO
- Las barras necesitan zero baseline, los puntos (linechart) no

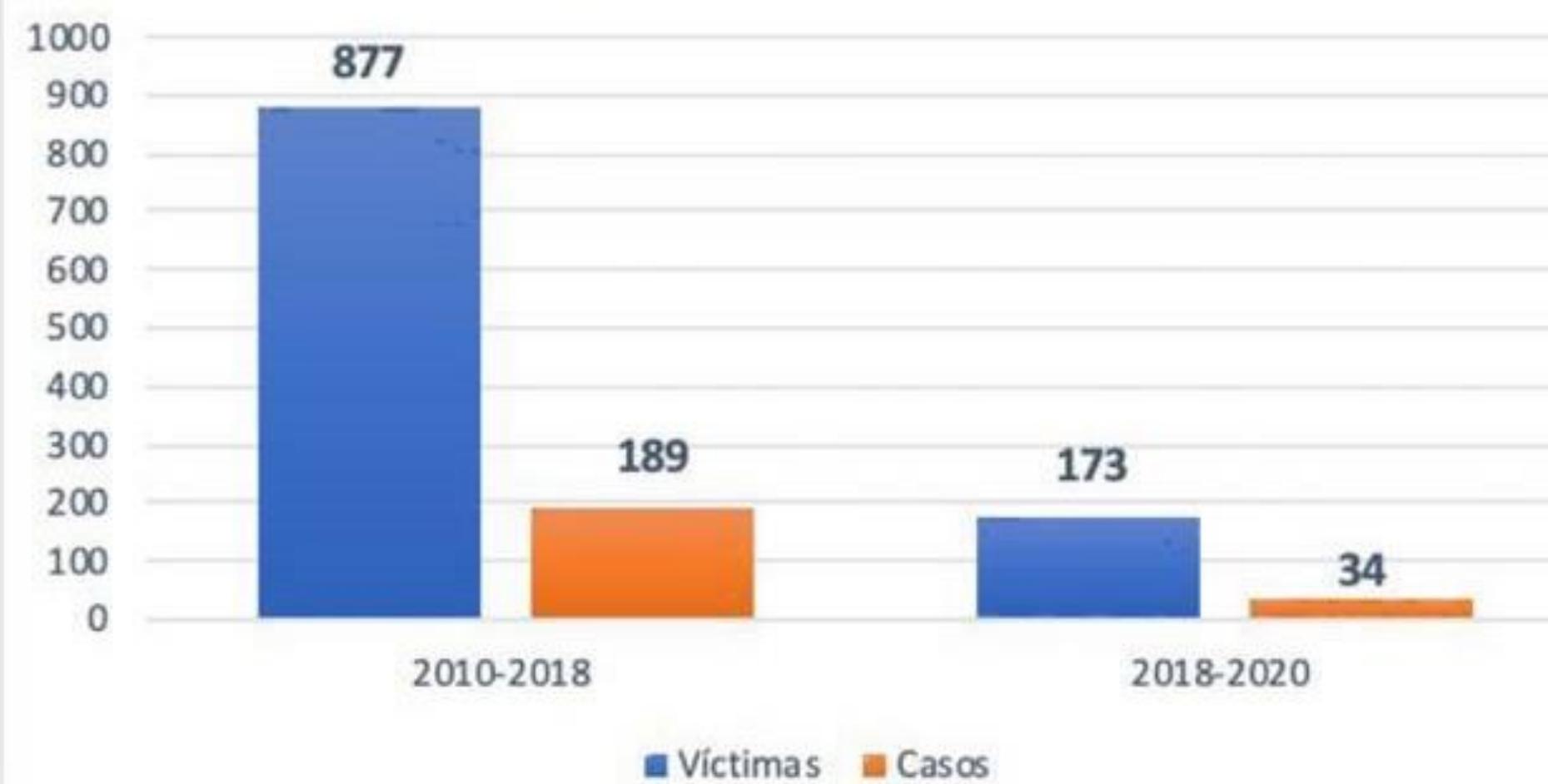


<http://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/>

Ejes

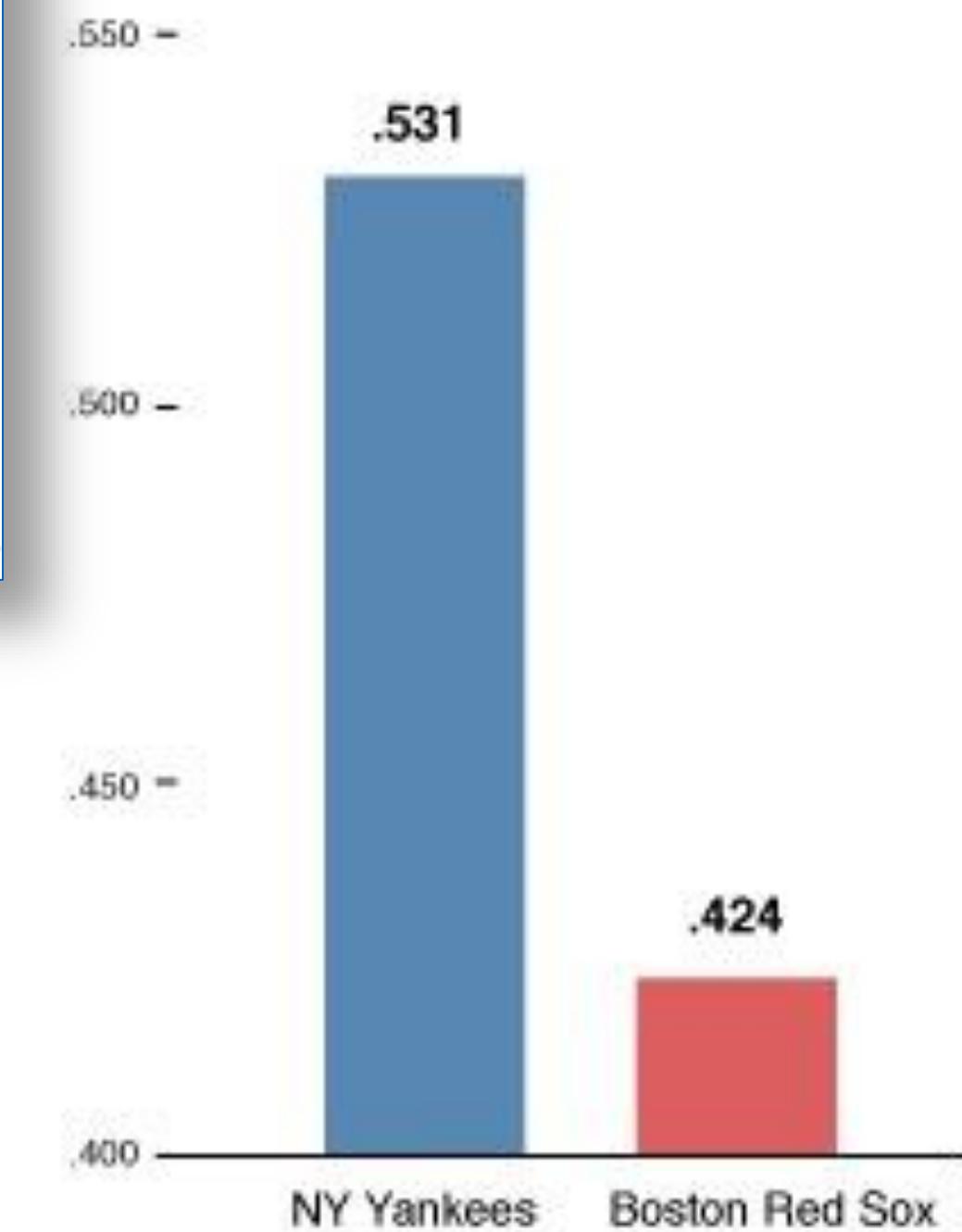
Translate Tweet

Homicidios Colectivos



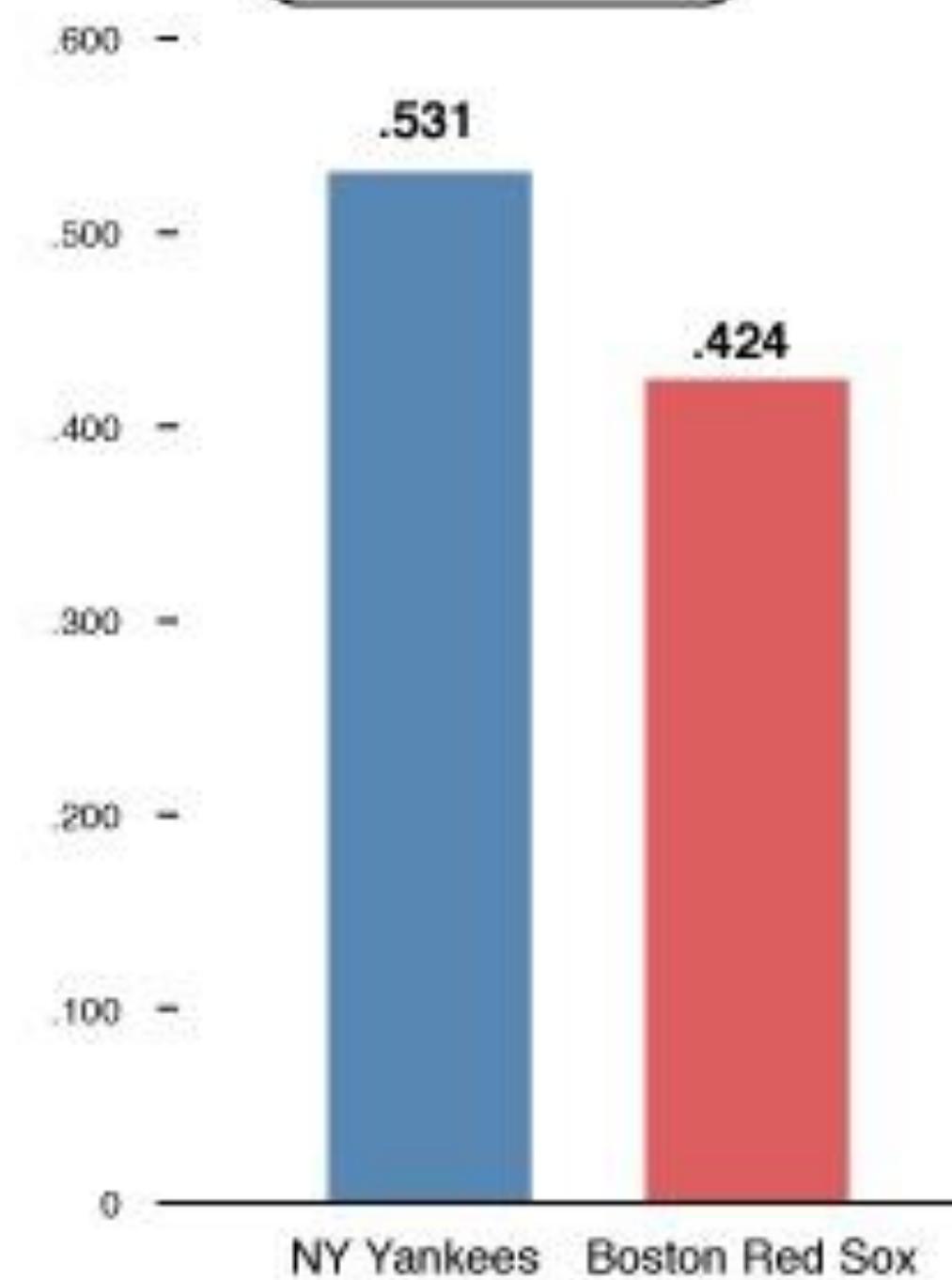
Percentage of victories

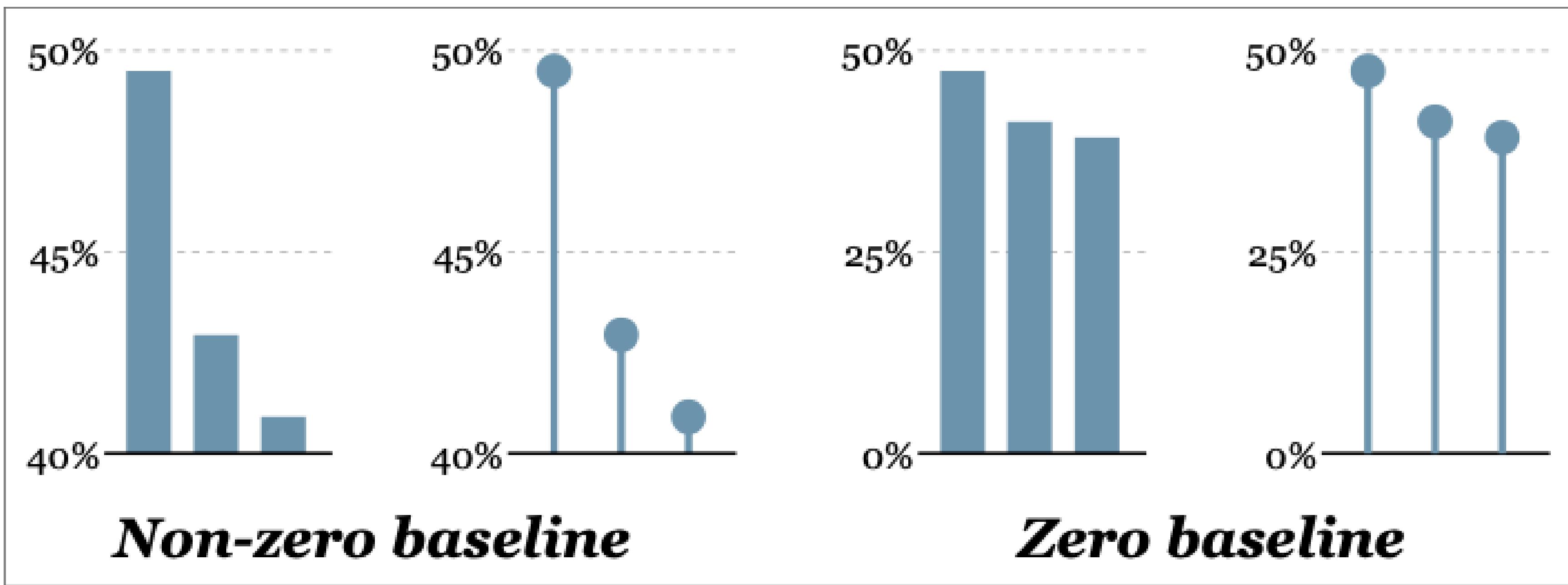
WRONG



Percentage of victories

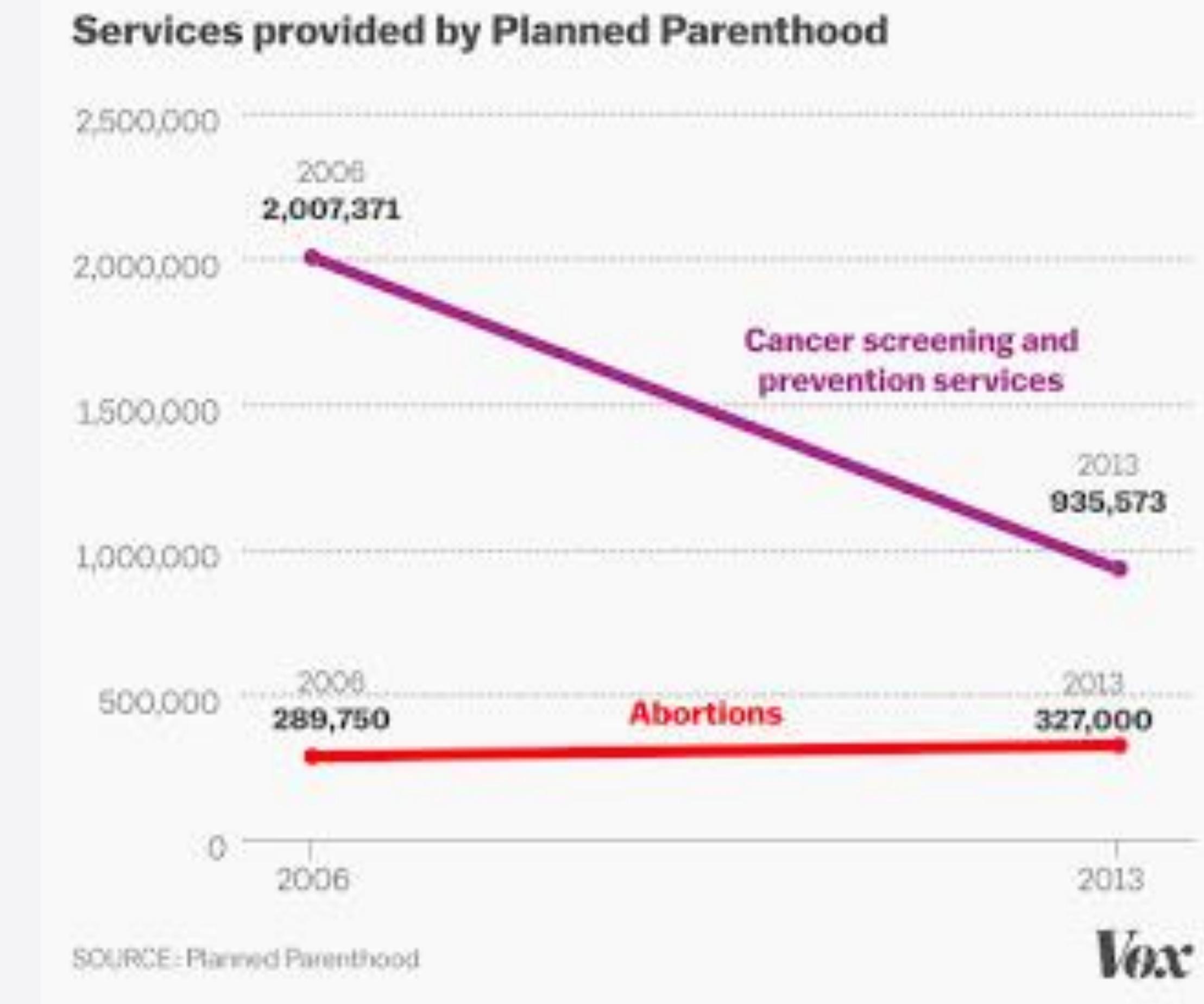
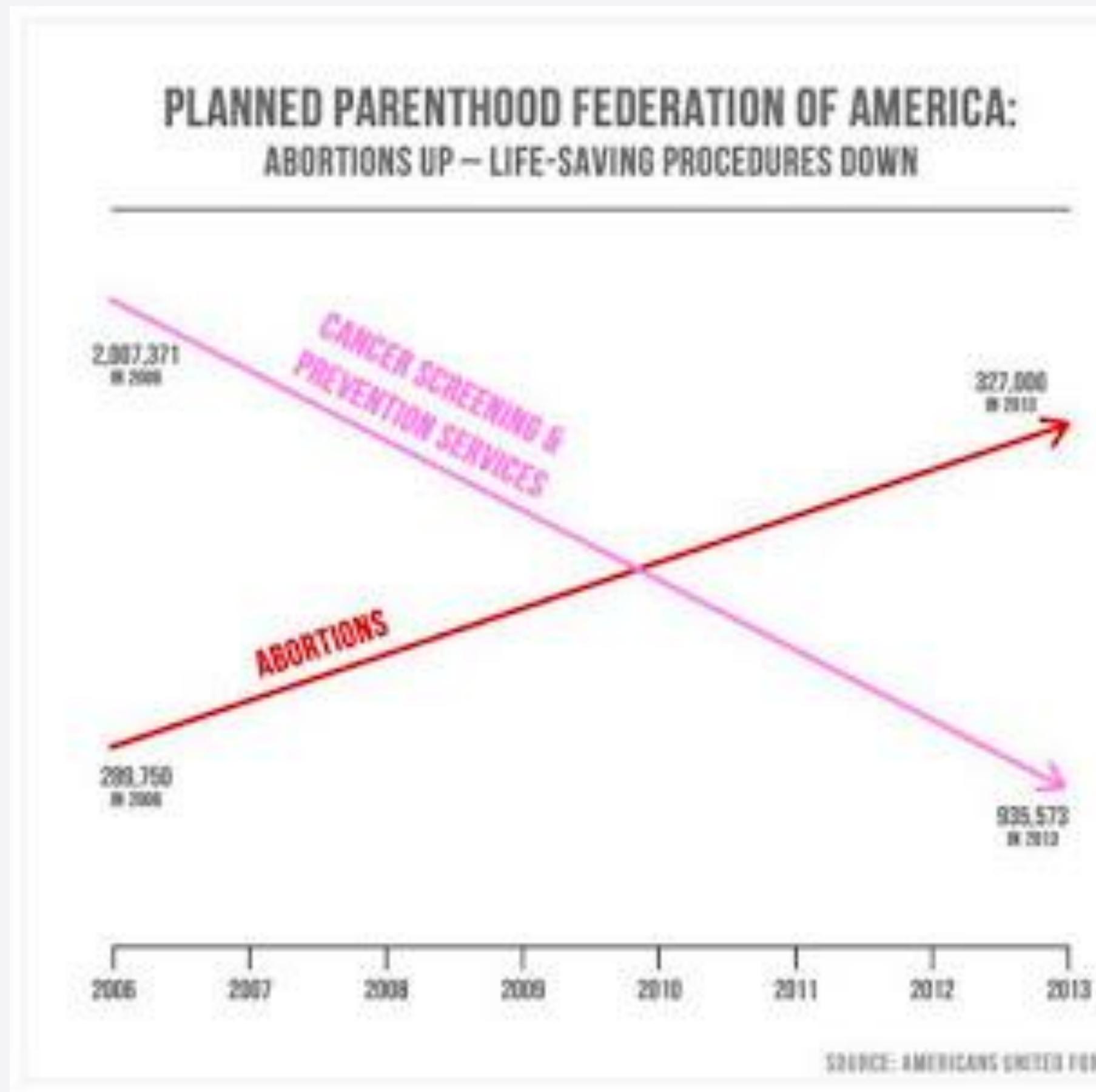
RIGHT





Ejes

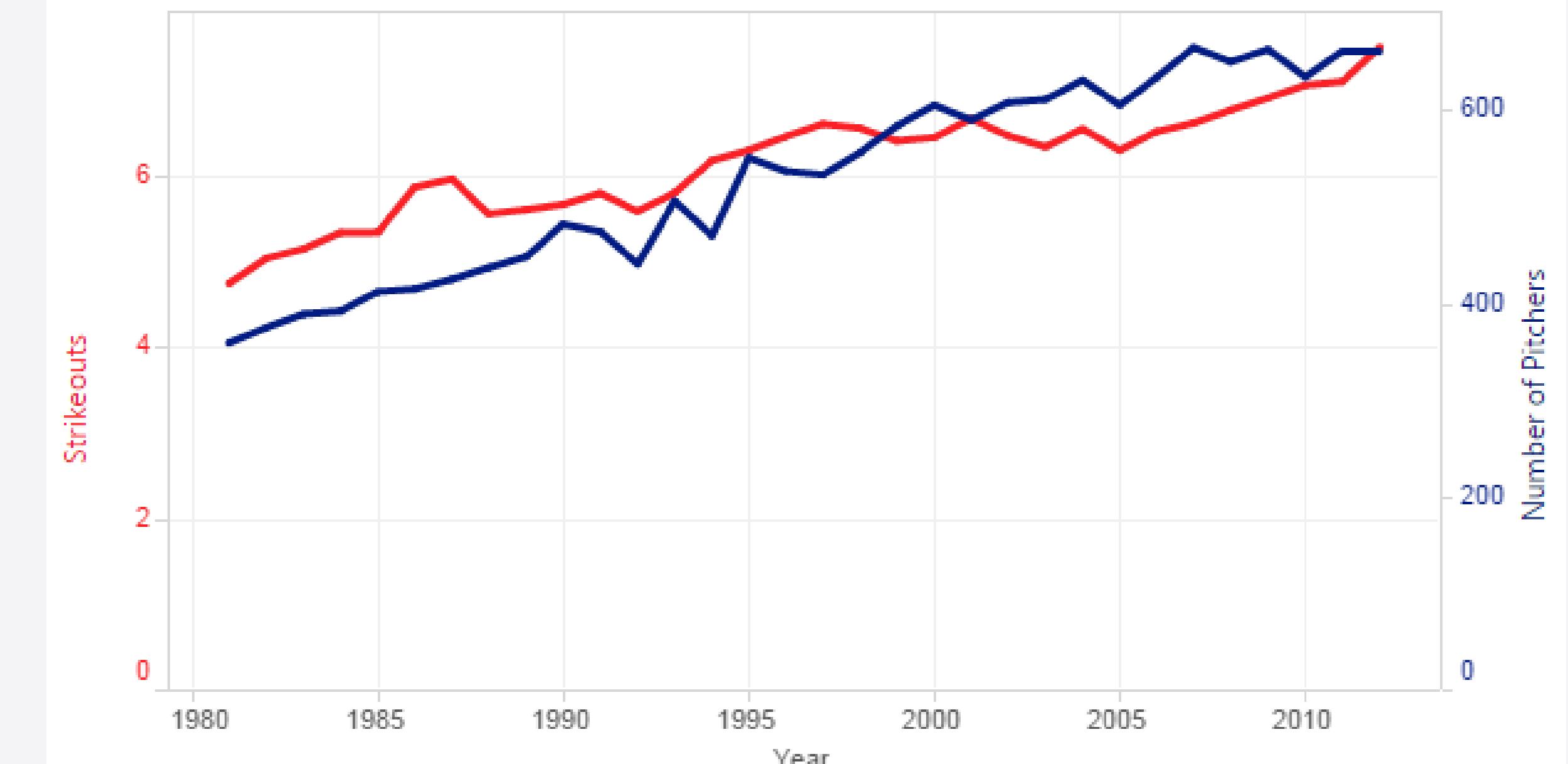
- Labels en los ejes son imprescindibles



Evitar ejes duales

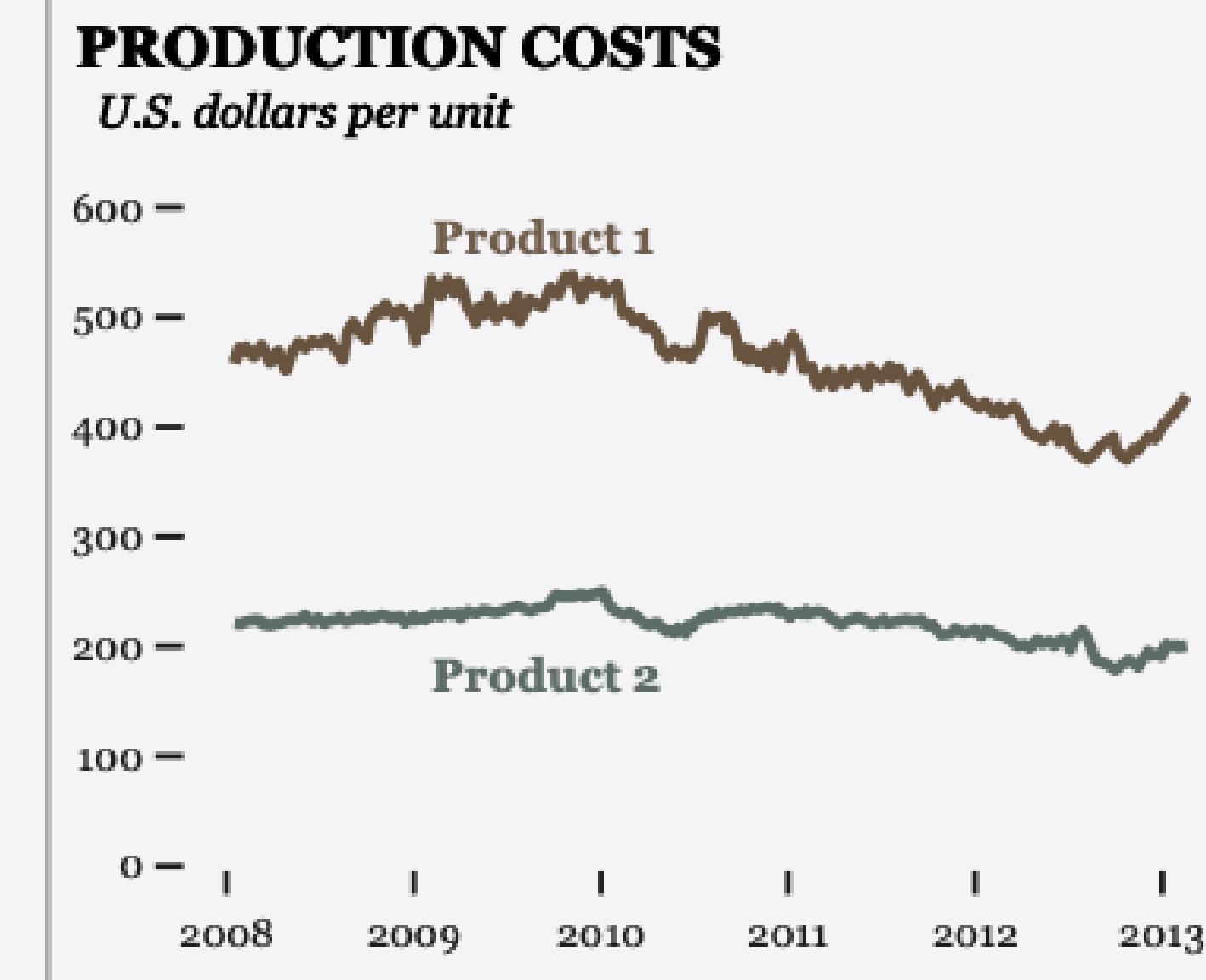
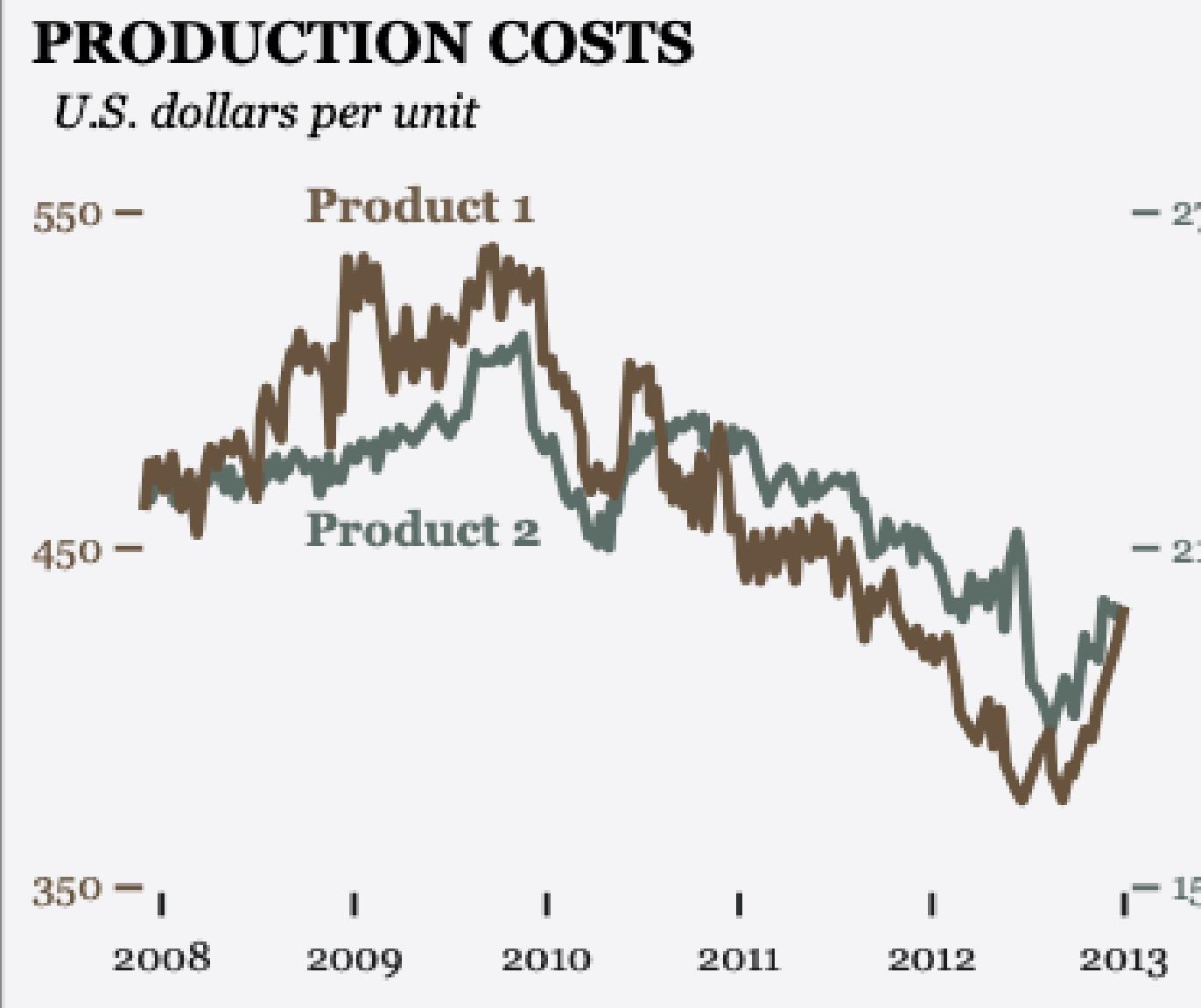
Problemáticos

- Aceptable en casos muy concretos
- Muy fácil llevar a conclusiones equivocadas:
 - Muestra implícita correlación entre líneas

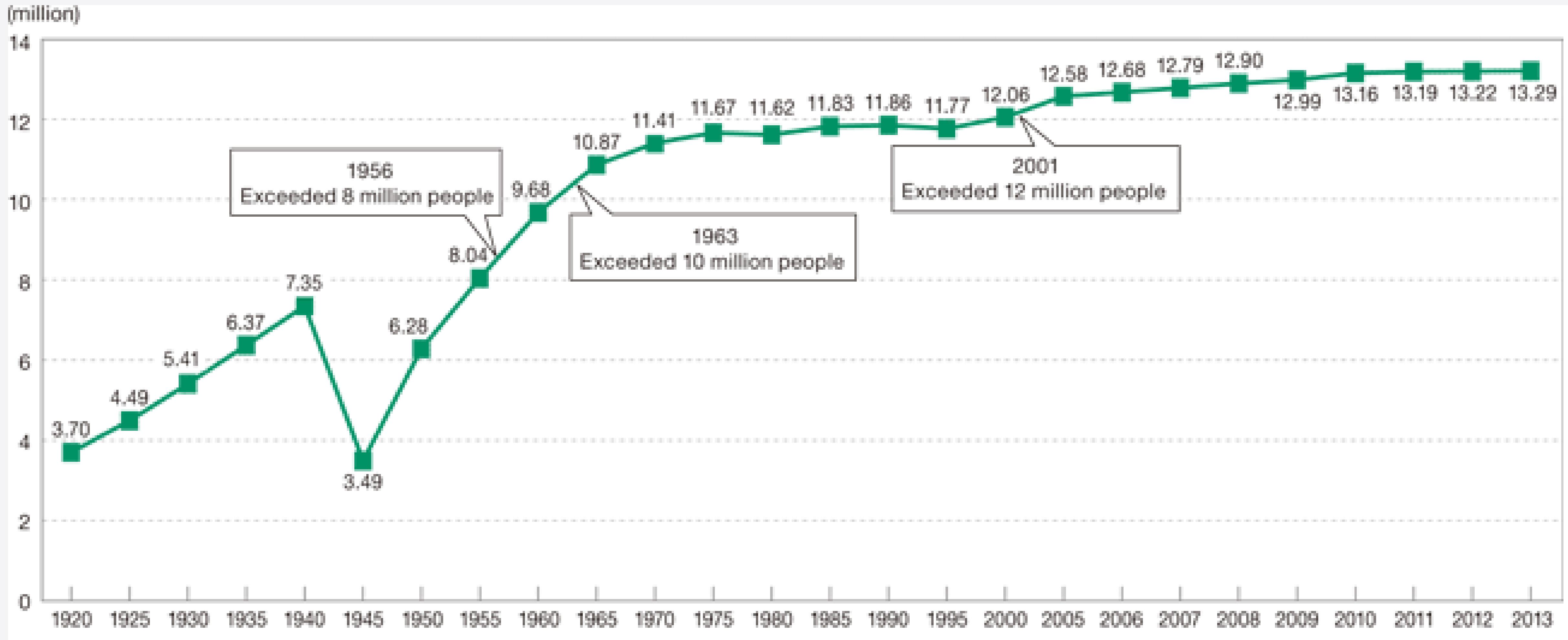


Source | <http://www.baseball-reference.com/leagues/MLB/pitch.shtml> Ben Jones (@DataRemixed) | 5/4/2013

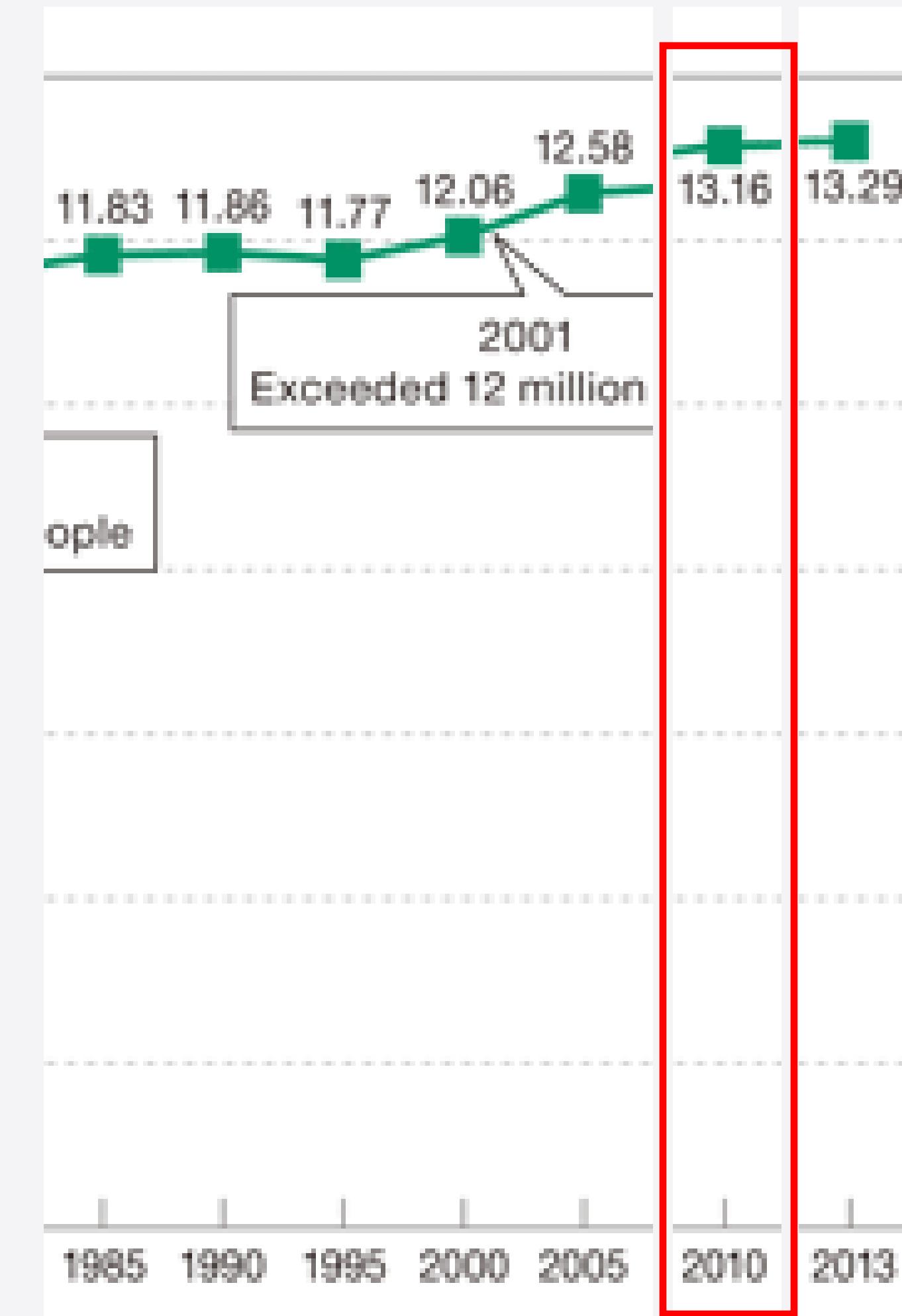
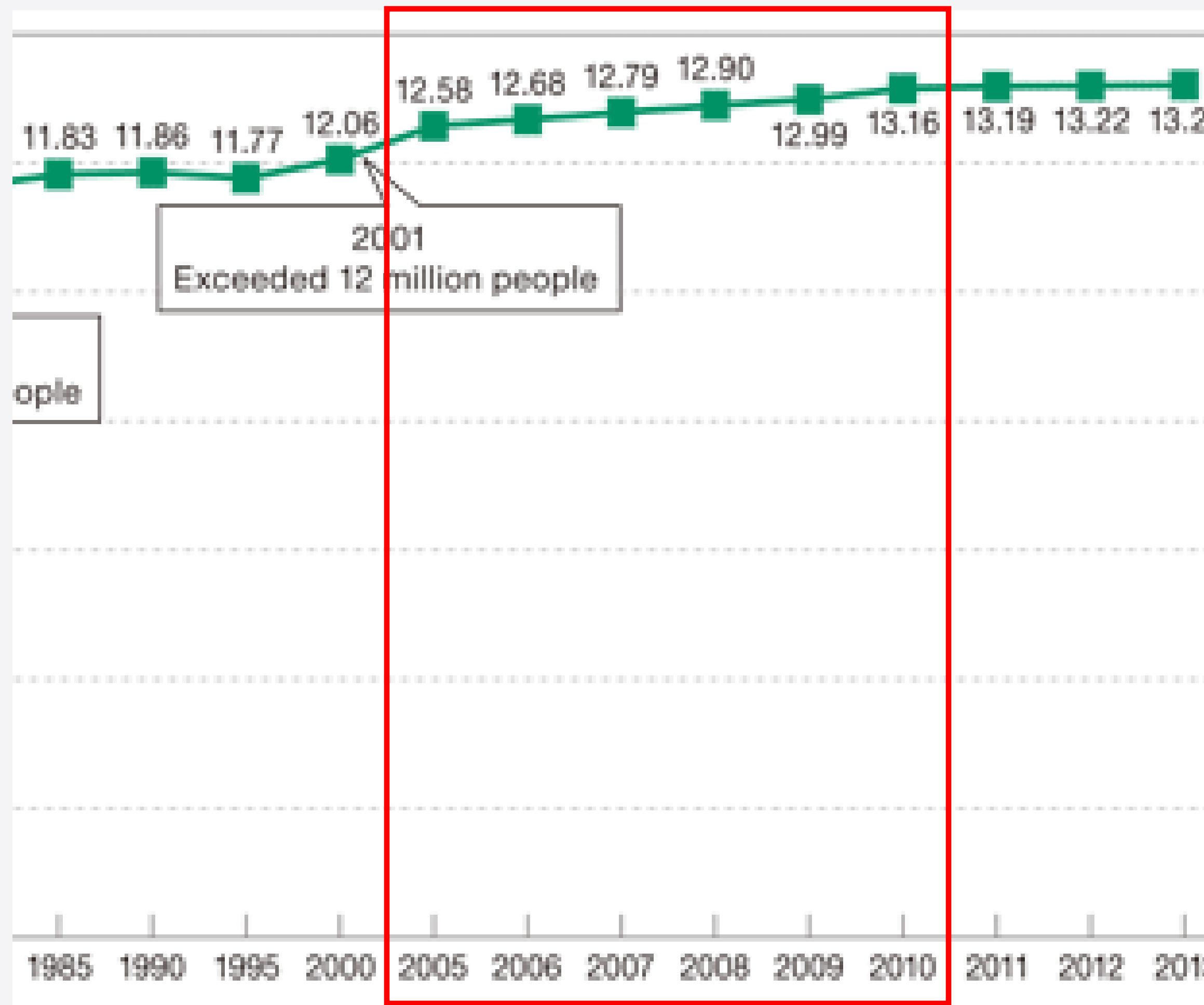
dataremixed.com



Steps

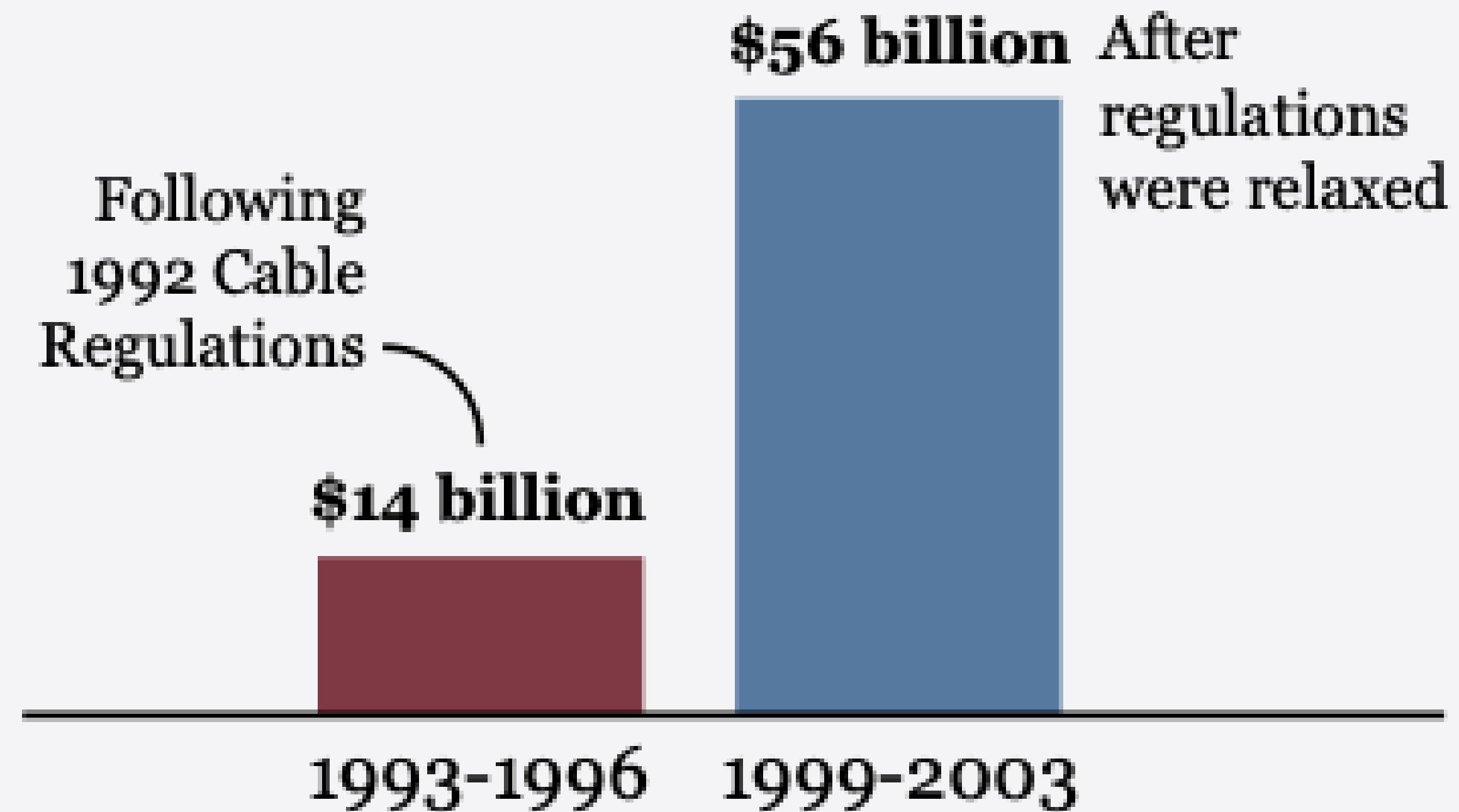


Steps



Hacer agregaciones equivalentes

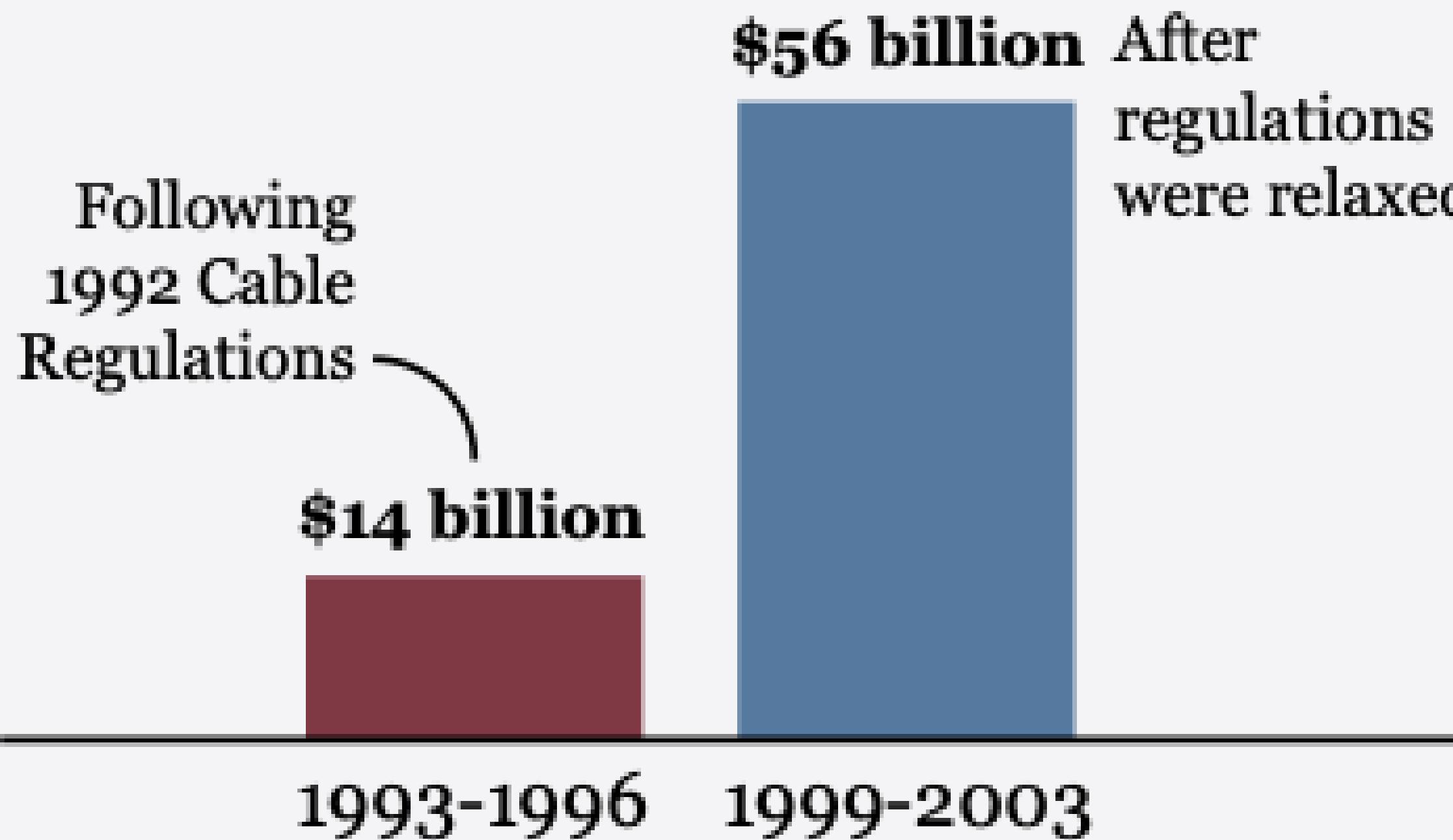
**Less regulation =
More industry investment**



Hacer agregaciones equivalentes

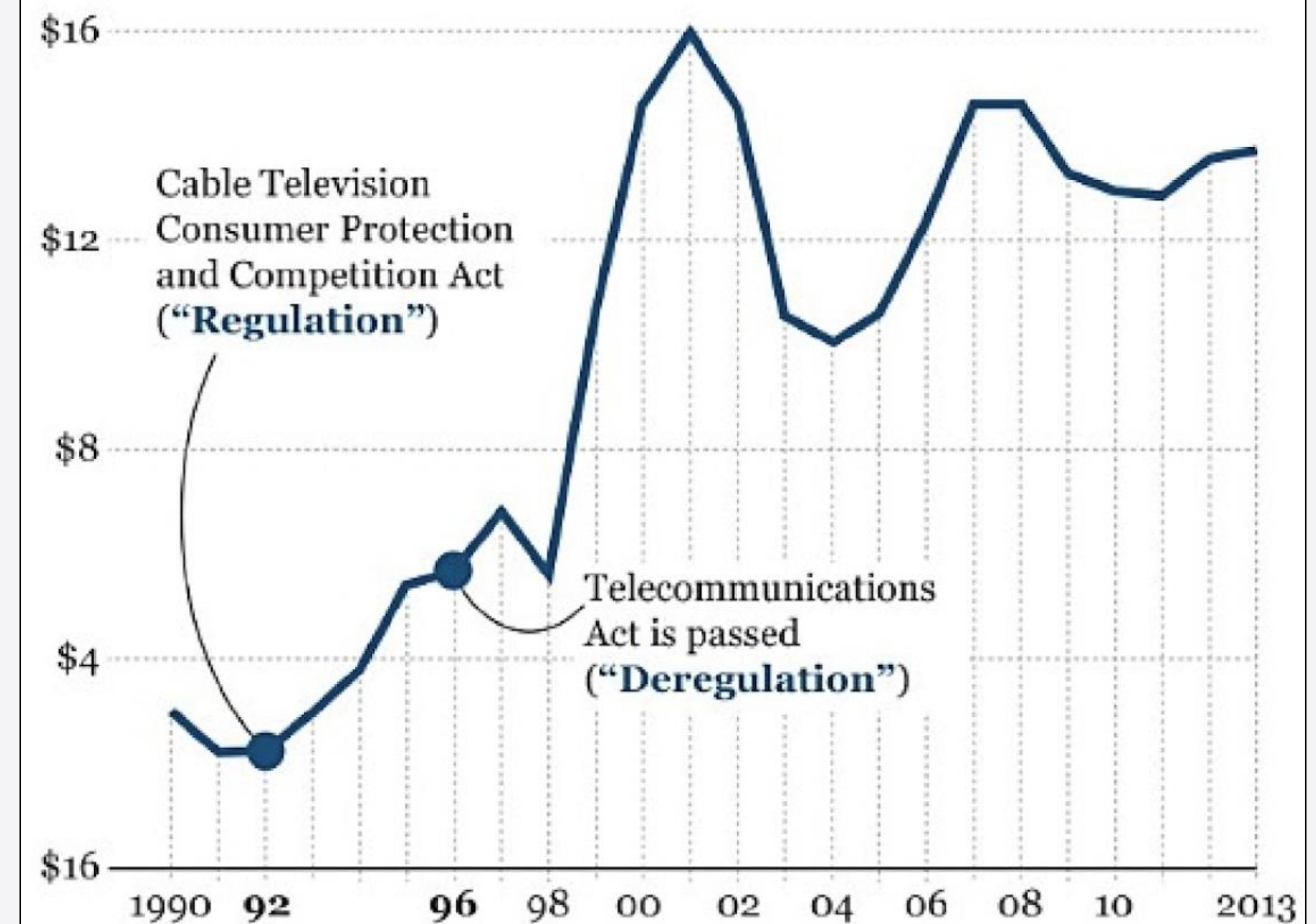
Less regulation =

More industry investment



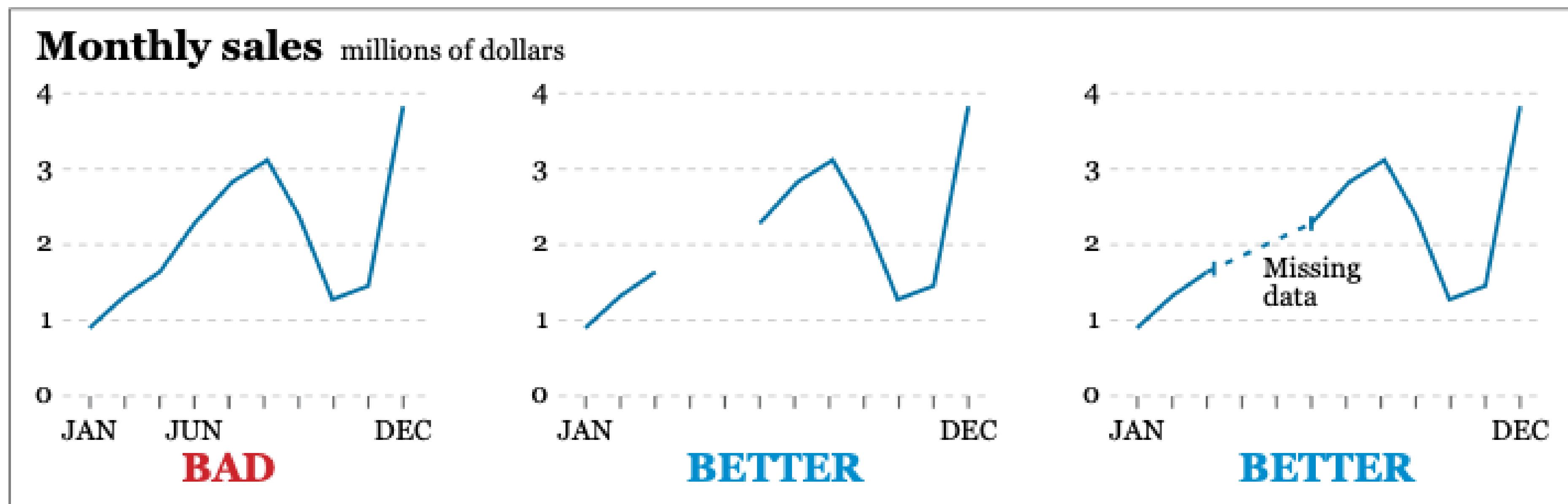
Cable Industry Infrastructure Expenditures

In billions



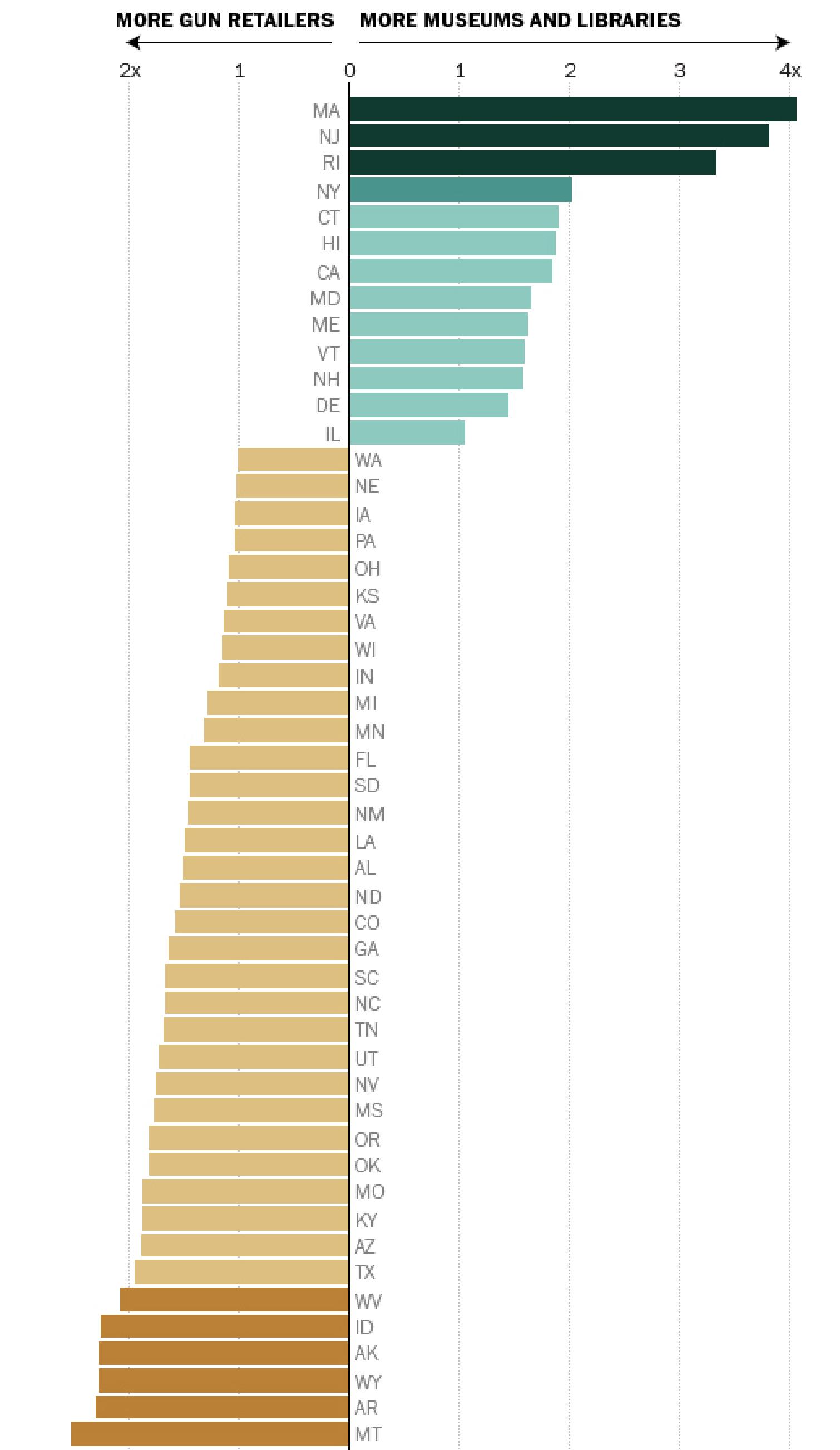
Missing data

- Marcar explícitamente donde no hay datos
- Mantener el espacio de las marcas / usar marcas distintas



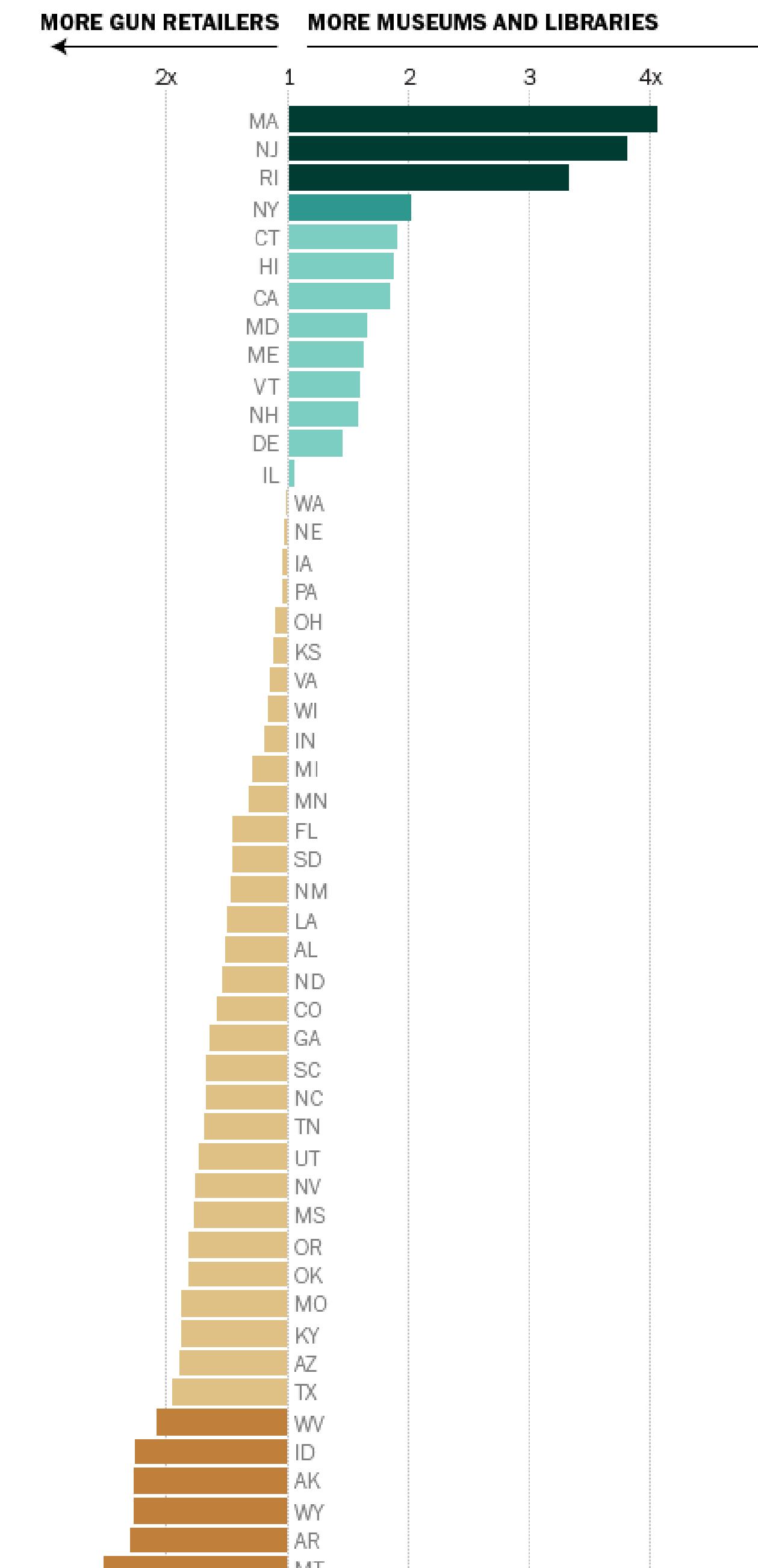
Baseline

In 37 states, gun dealers outnumber museums and libraries



SOURCE: Institute of Museum and Library Sciences; Bureau of Alcohol, Tobacco and Firearms.
GRAPHIC: The Washington Post. Published June 17, 2014

In 37 states, gun dealers outnumber museums and libraries



SOURCE: Institute of Museum and Library Sciences; Bureau of Alcohol, Tobacco and Firearms.
GRAPHIC: The Washington Post. Published June 17, 2014

Stacked bars no alineadas

- Problemático. Difícil comparar porcentajes
- Difícil identificar 50%

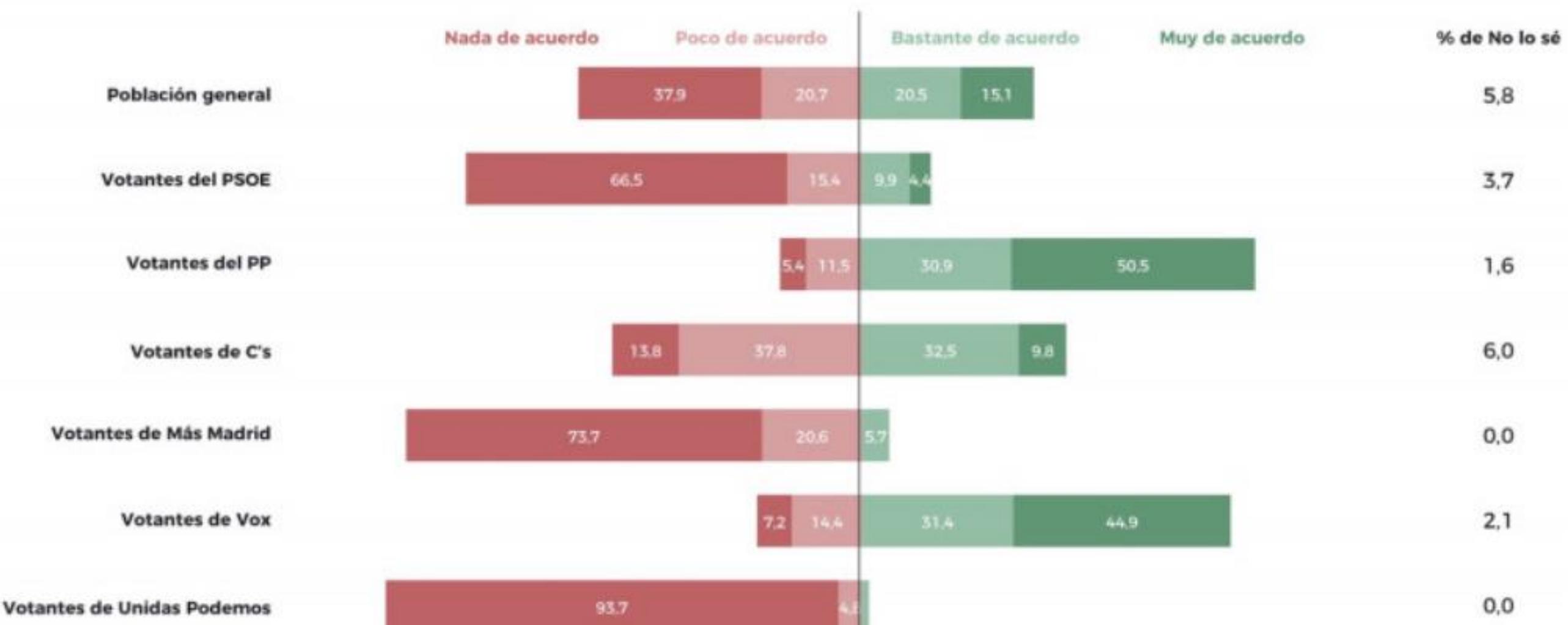
Más del 50% de los electores de C's suspenden a la presidenta de la Comunidad y al Gobierno regional de su propio partido. El 51,1% de los madrileños, a favor de un cambio. Sánchez sale peor parado que la líder popular

ctxt 6/02/2021

Sobre Isabel Díaz Ayuso

¿En qué medida estás de acuerdo con las siguientes afirmaciones sobre la Presidenta de la Comunidad de Madrid, Isabel Díaz Ayuso?

Está capacitada para presidir la Comunidad de Madrid

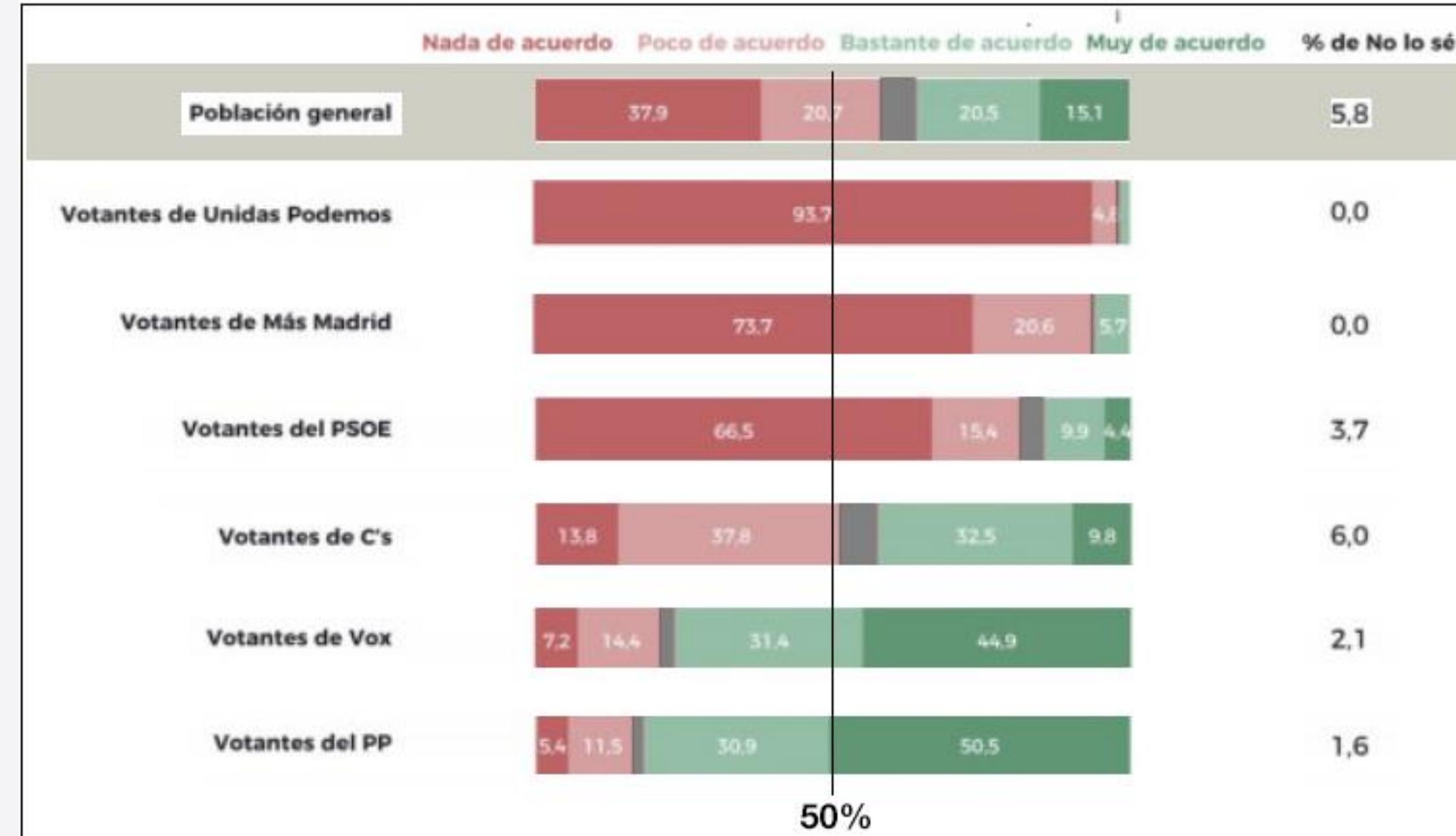


40dB.

ctxt
REVISTA CONTEXTO

Sobre Isabel Díaz Ayuso

Está capacitada para presidir la Comunidad de Madrid

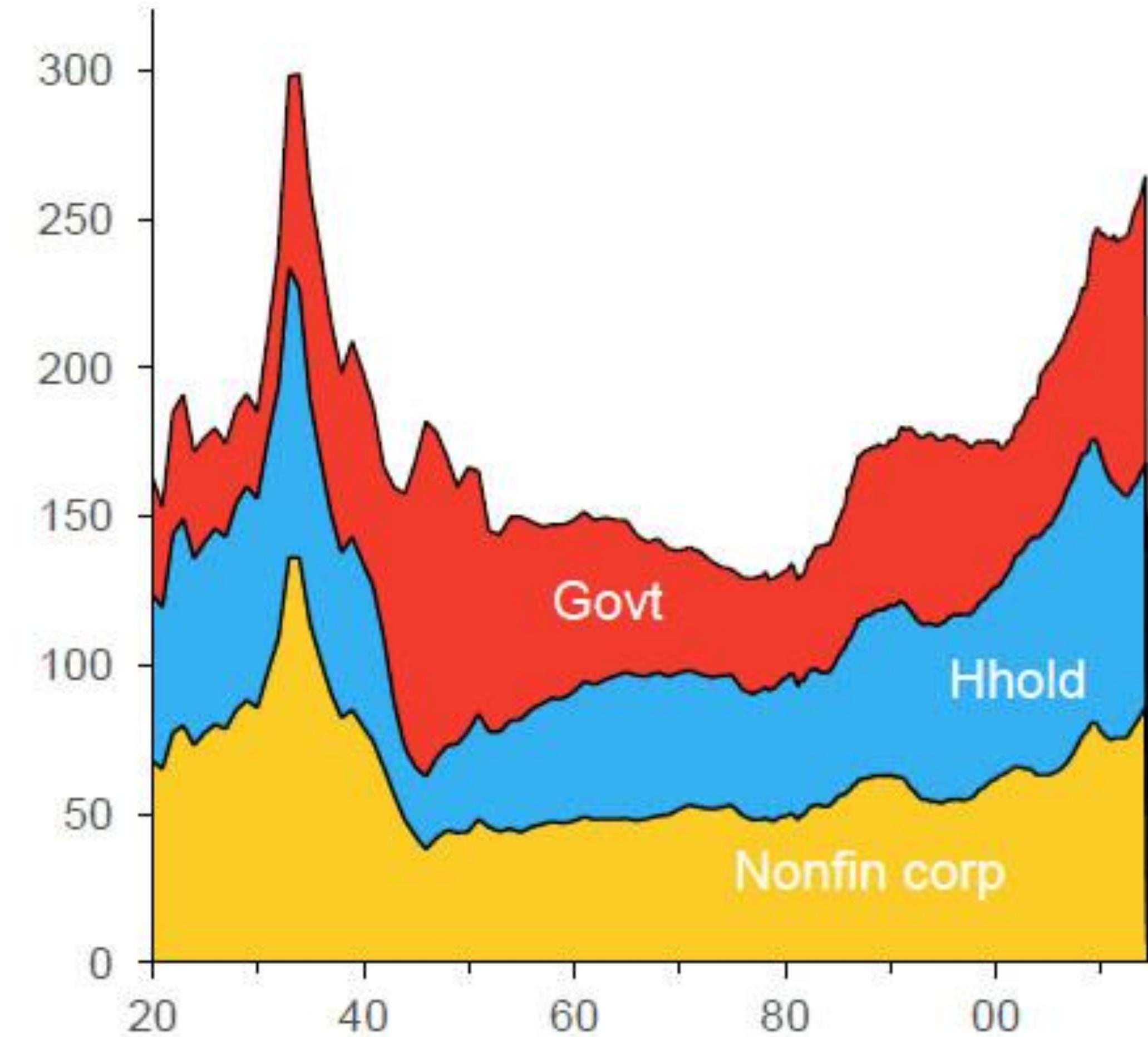


Precisión en area charts

- Mensaje: Niveles de deuda actuales similares a 1930s pero...
- Muy difícil identificar que categoría es responsable del aumento.
- ¿Otras relaciones entre ellas?

More debt = stronger wealth effects

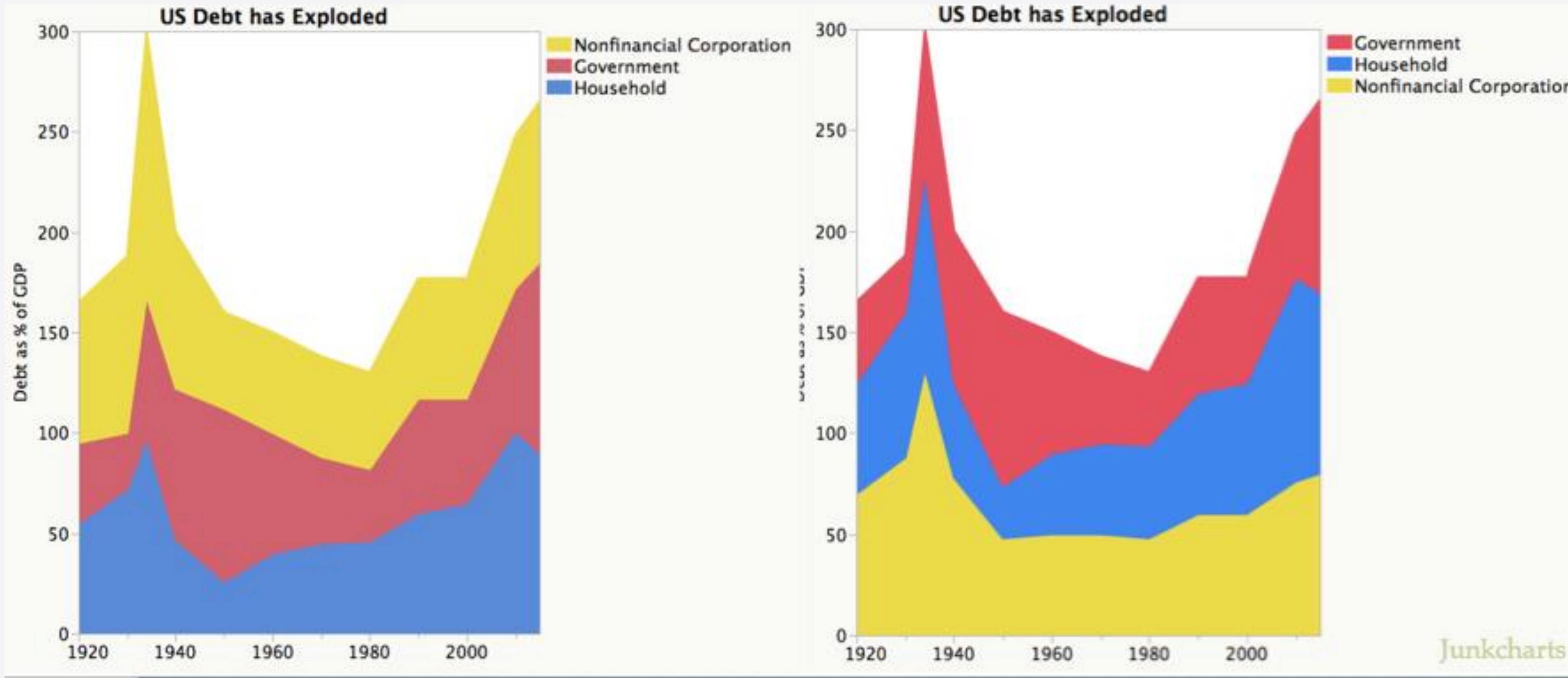
US debt across non-fin sectors, % GDP



Source: Federal Reserve.

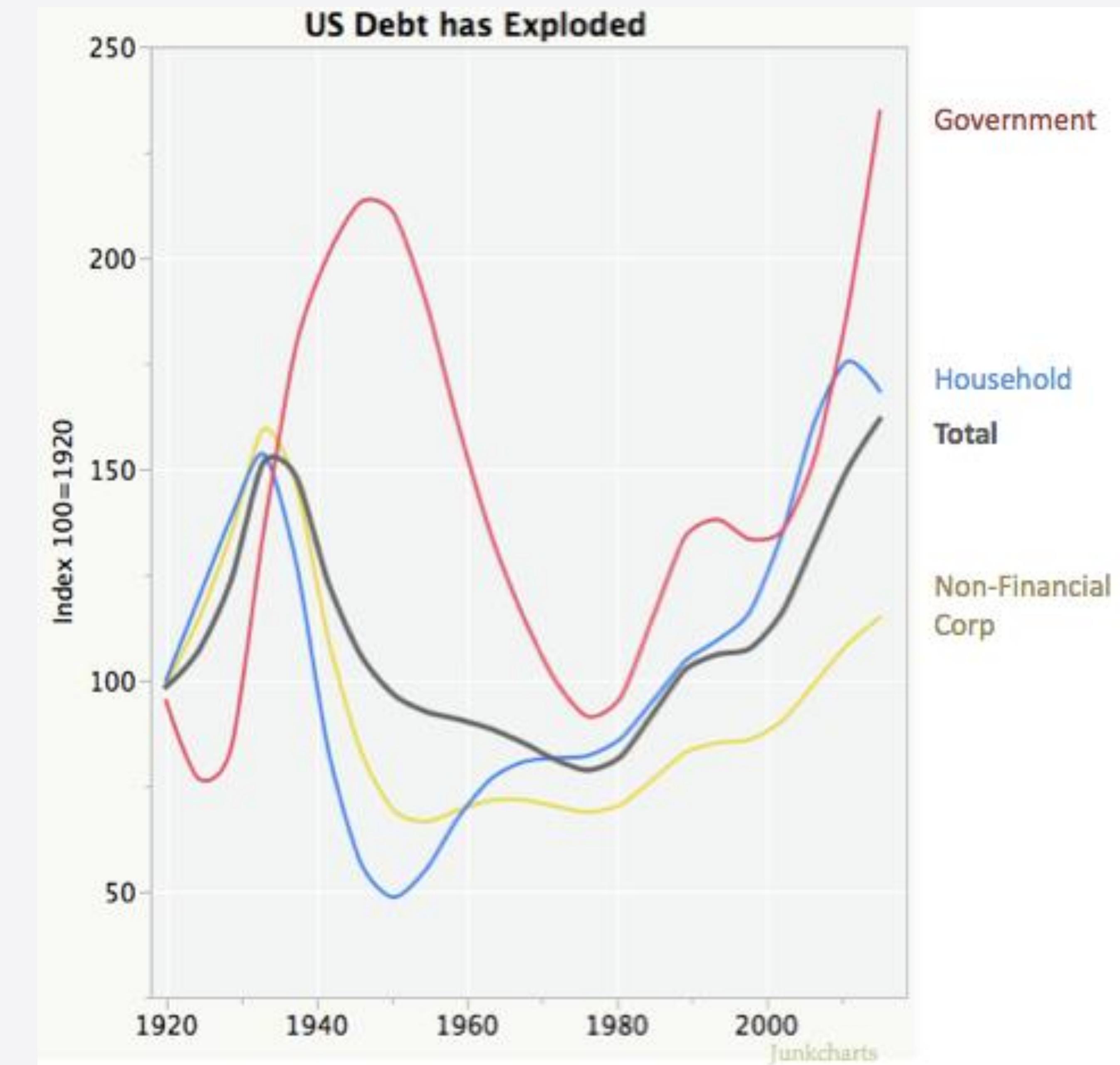
Precisión en area charts

- Dos opciones de 6 posibles
- Picos en capas superiores tienden a magnificarse



Precisión en area charts

- Solución: no usar área chart
- Calcular índices usando la media de 1920 como referencia

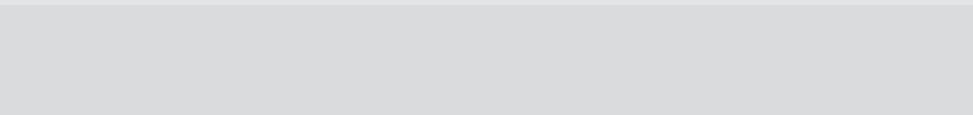


Elegir métricas relevantes

Most dangerous cities

Total murders in 2014

Chicago



407

New York

328

Detroit

304

Los Angeles

259

Philadelphia

248

Elegir métricas relevantes

Most dangerous cities

Total murders in 2014

Chicago

407

New York

328

Detroit

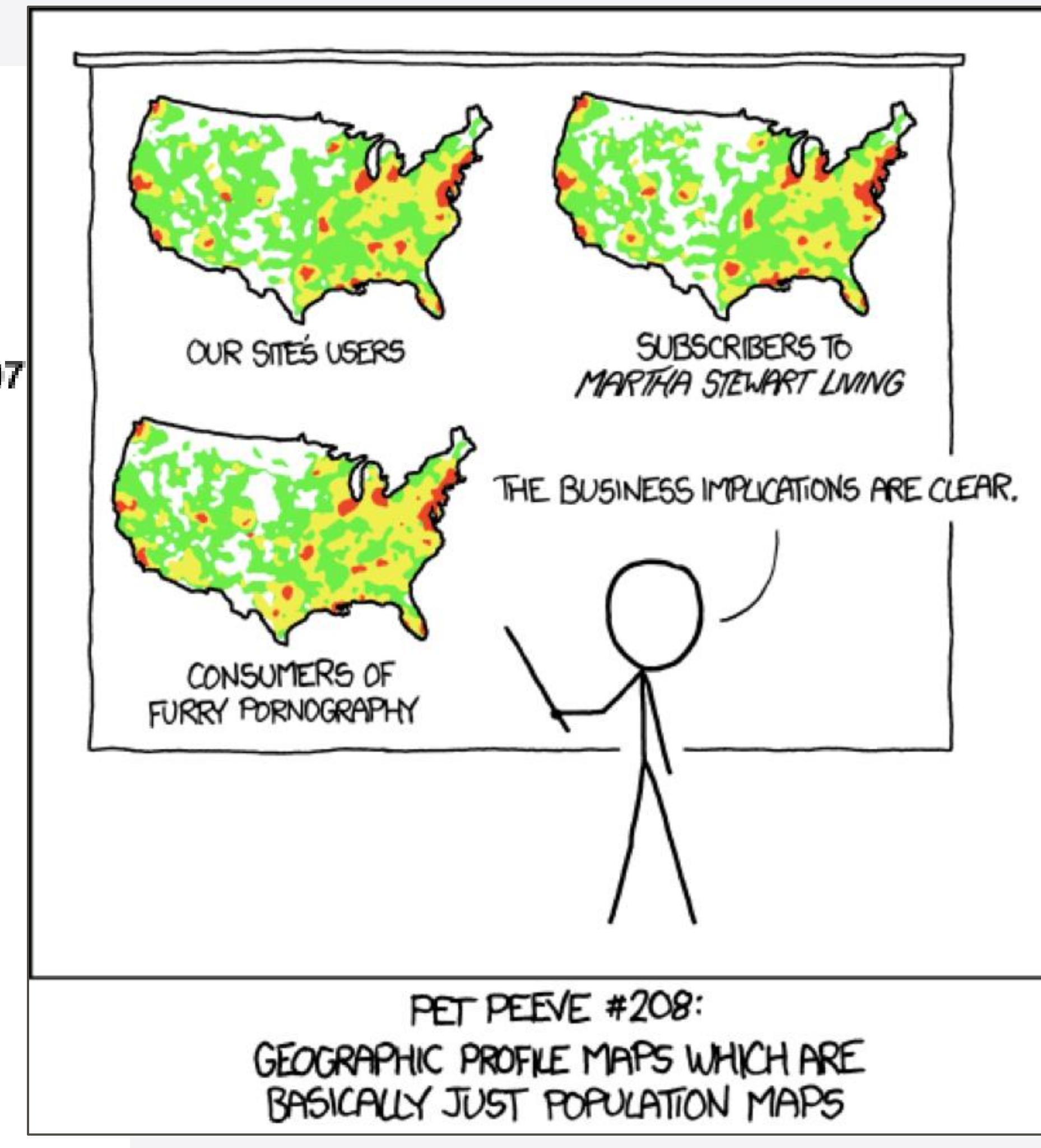
304

Los Angeles

259

Philadelphia

248



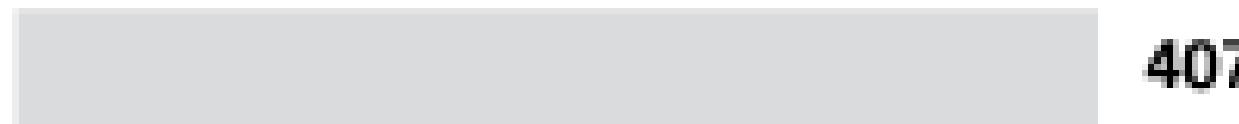
[<https://xkcd.com/1138>]

Elegir métricas relevantes

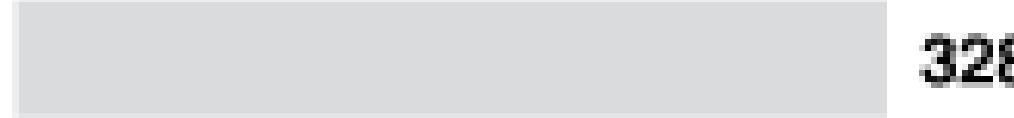
Most dangerous cities

Total murders in 2014

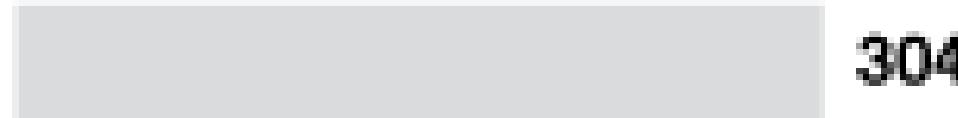
Chicago



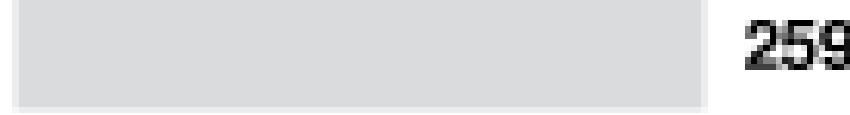
New York



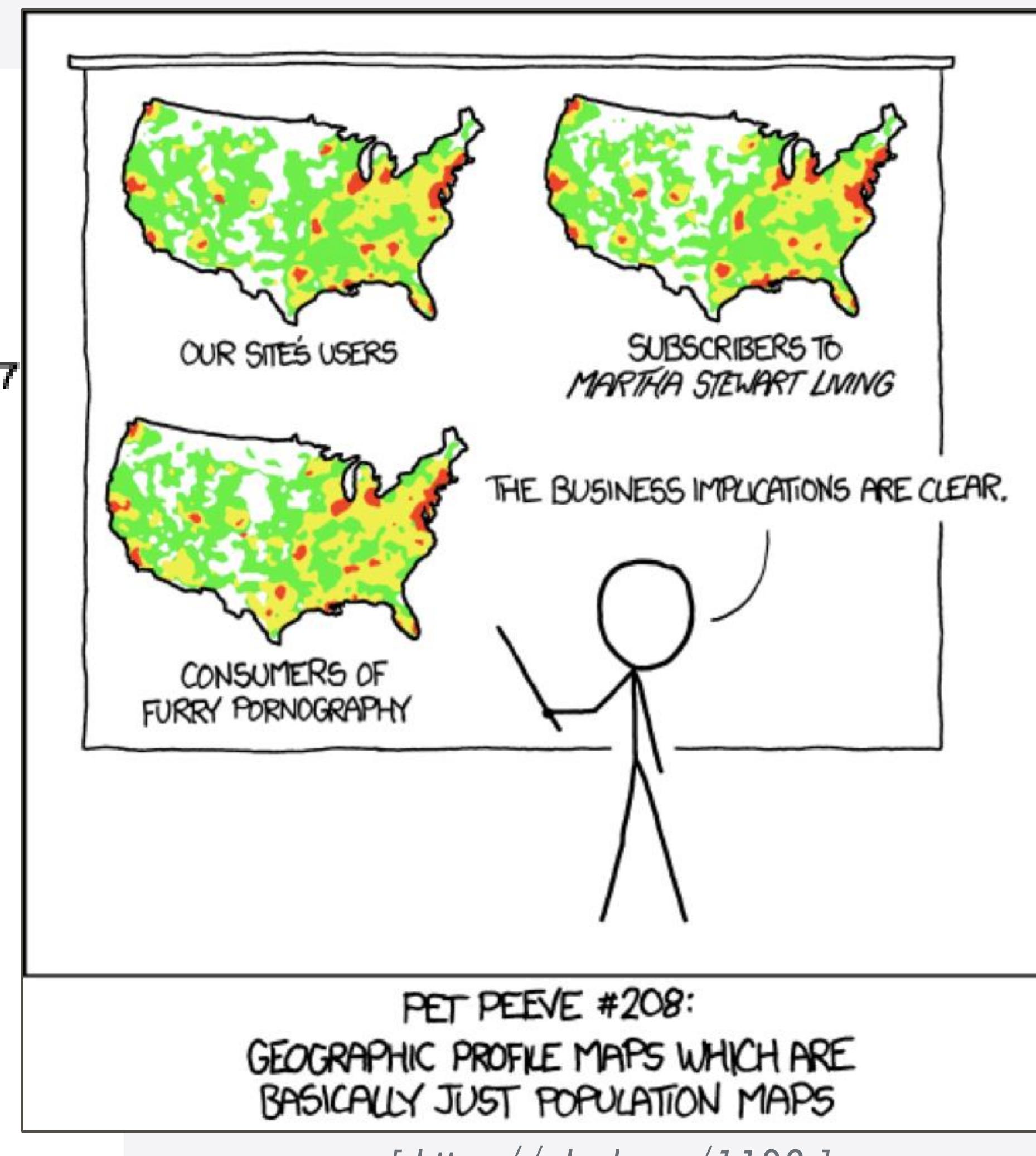
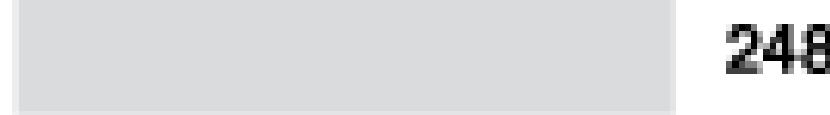
Detroit



Los Angeles



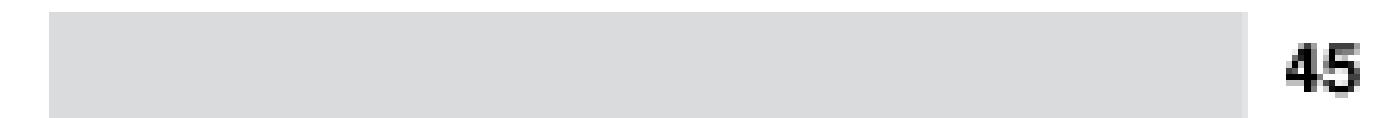
Philadelphia



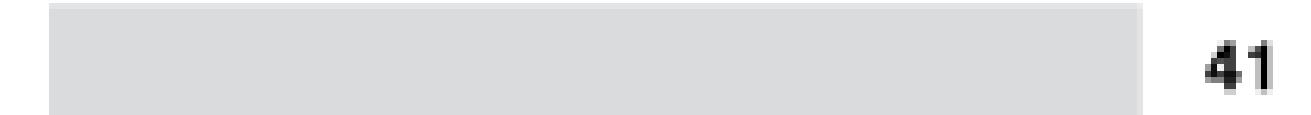
Most dangerous cities

Murder rate in major US cities in 2014, per 100,000 people

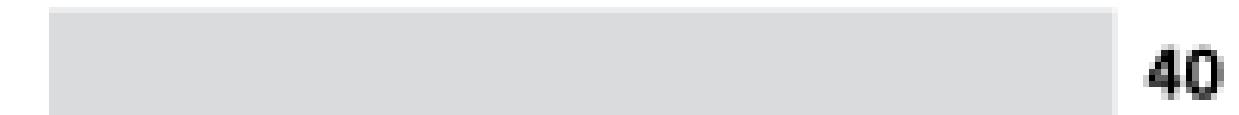
Detroit



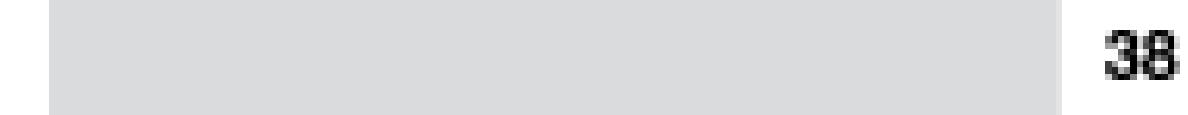
New Orleans



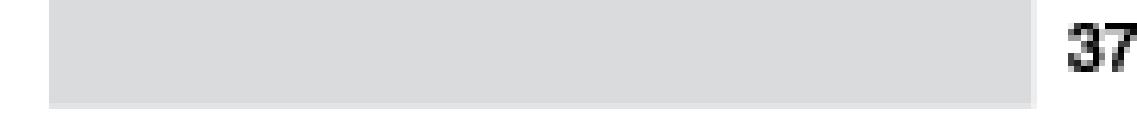
Newark



St. Louis



Baltimore

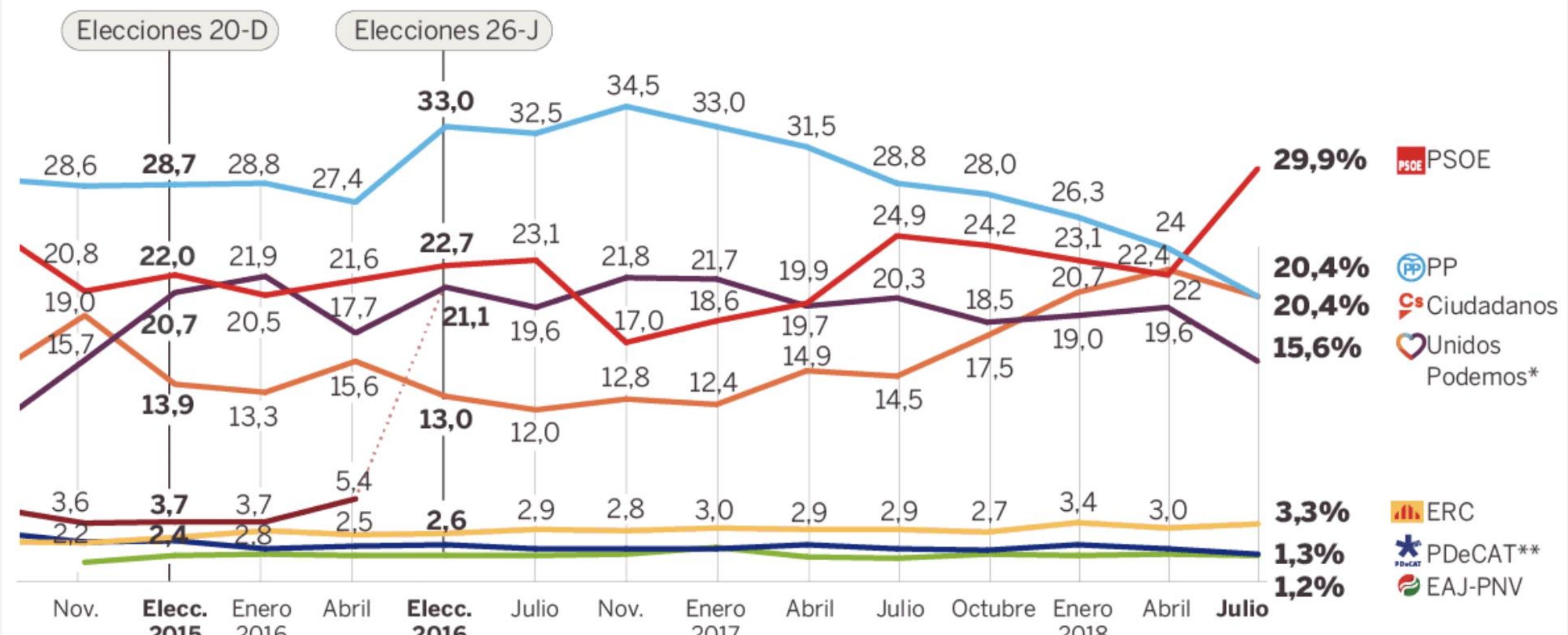


FICHA TÉCNICA

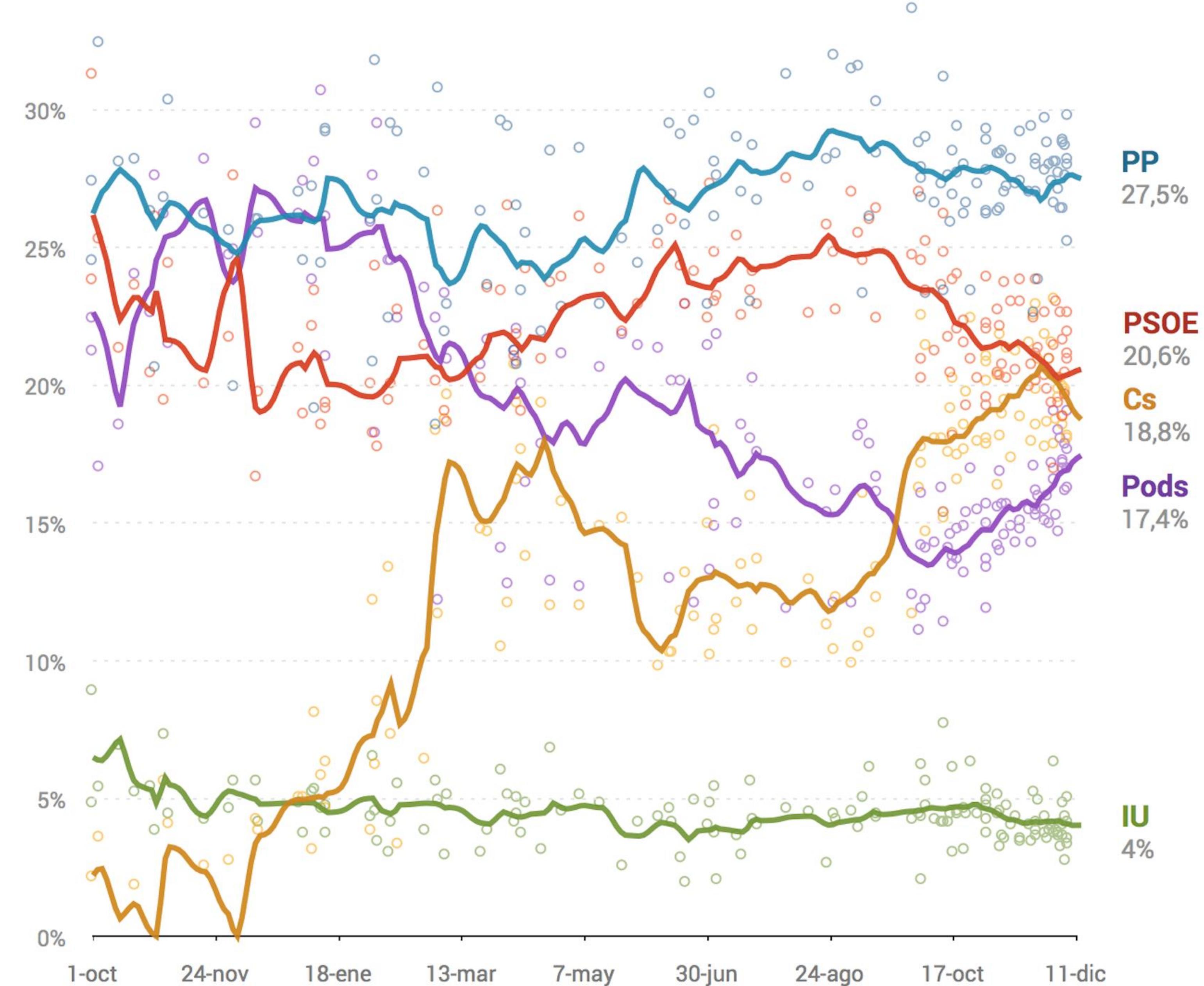
Sondeo efectuado mediante entrevista personal a 2.485 personas mayores de 18 años de ambos性os en 256 municipios de 47 provincias entre el 1 y el 10 de julio. Nivel de confianza: 95,5%. Margen de error: ± 2,0 puntos.

ESTIMACIÓN DE VOTO

En % sobre voto válido



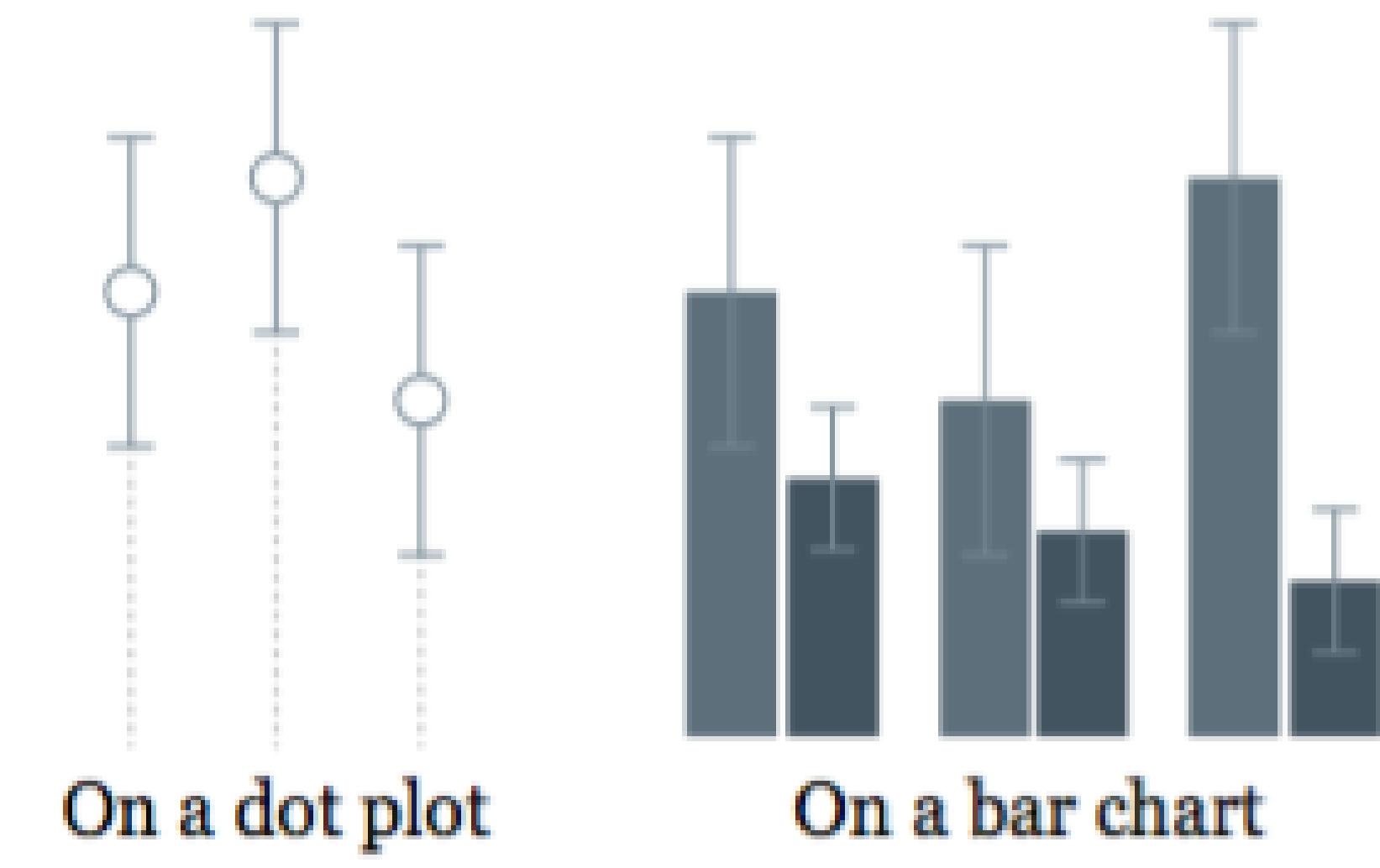
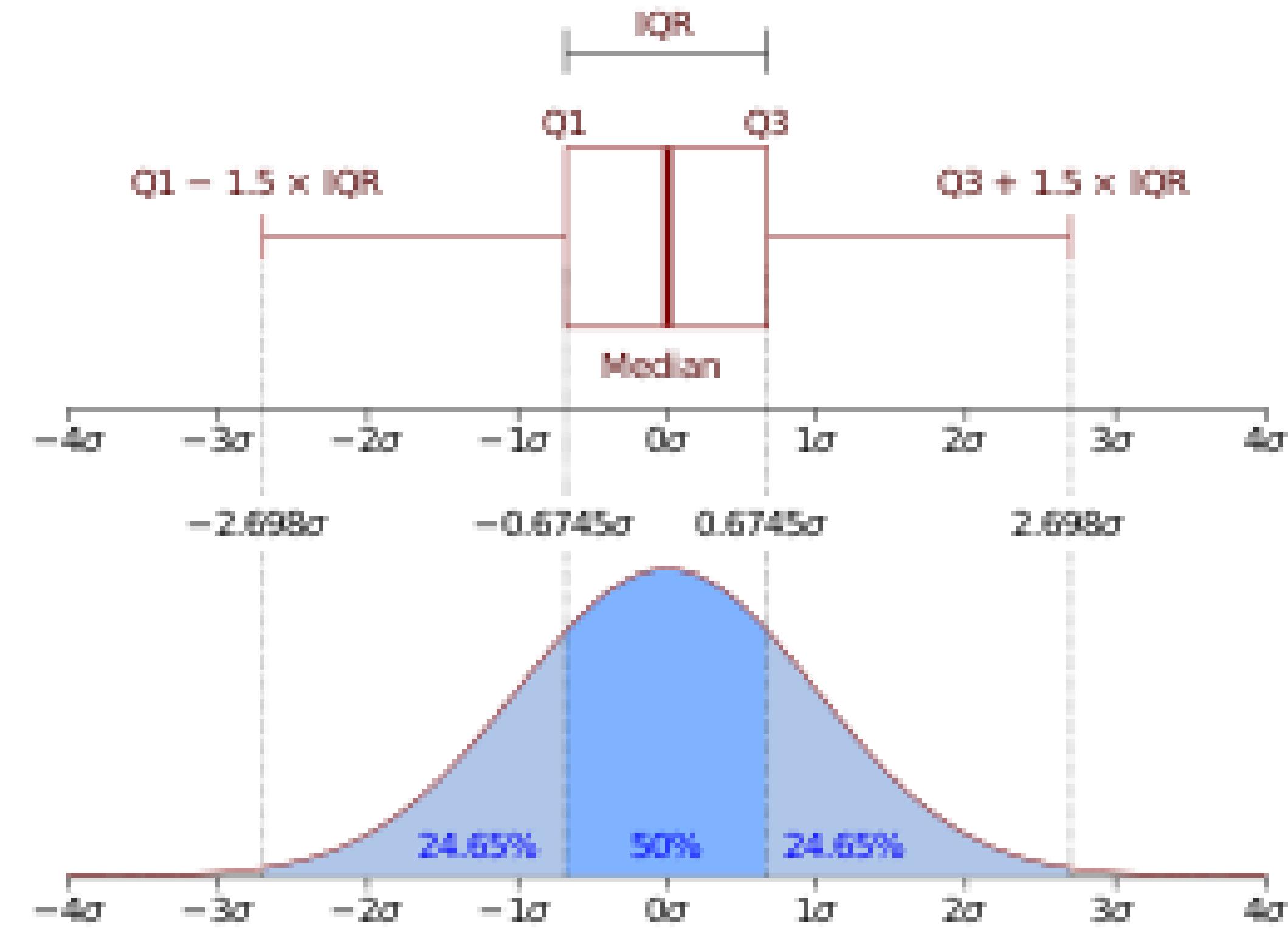
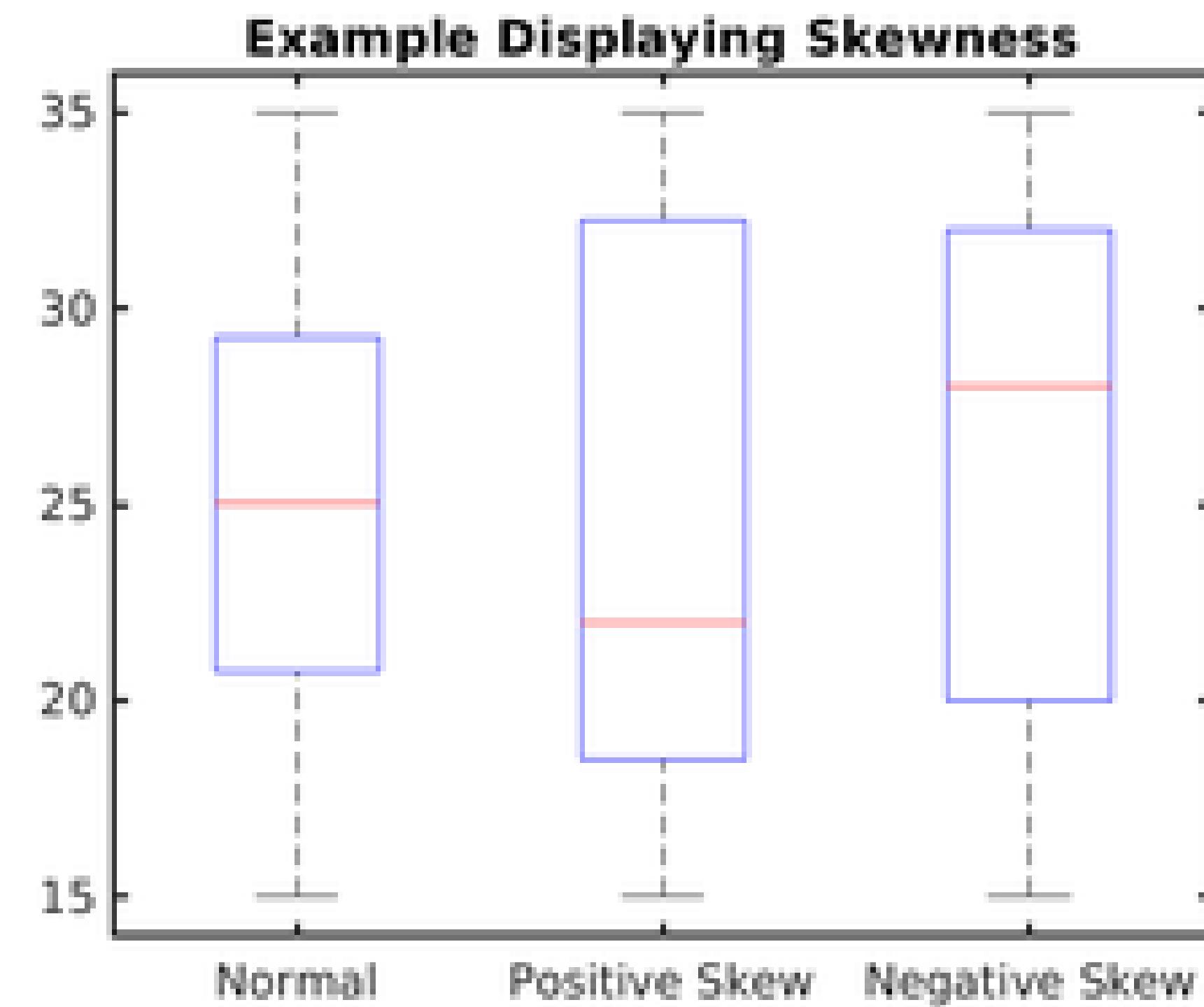
Porcentaje de voto según las encuestas. Las líneas representan un promedio ponderado por fecha, tamaño de muestra y empresa encuestadora.



Box and whisker plots

Representación visual de Distribución o:

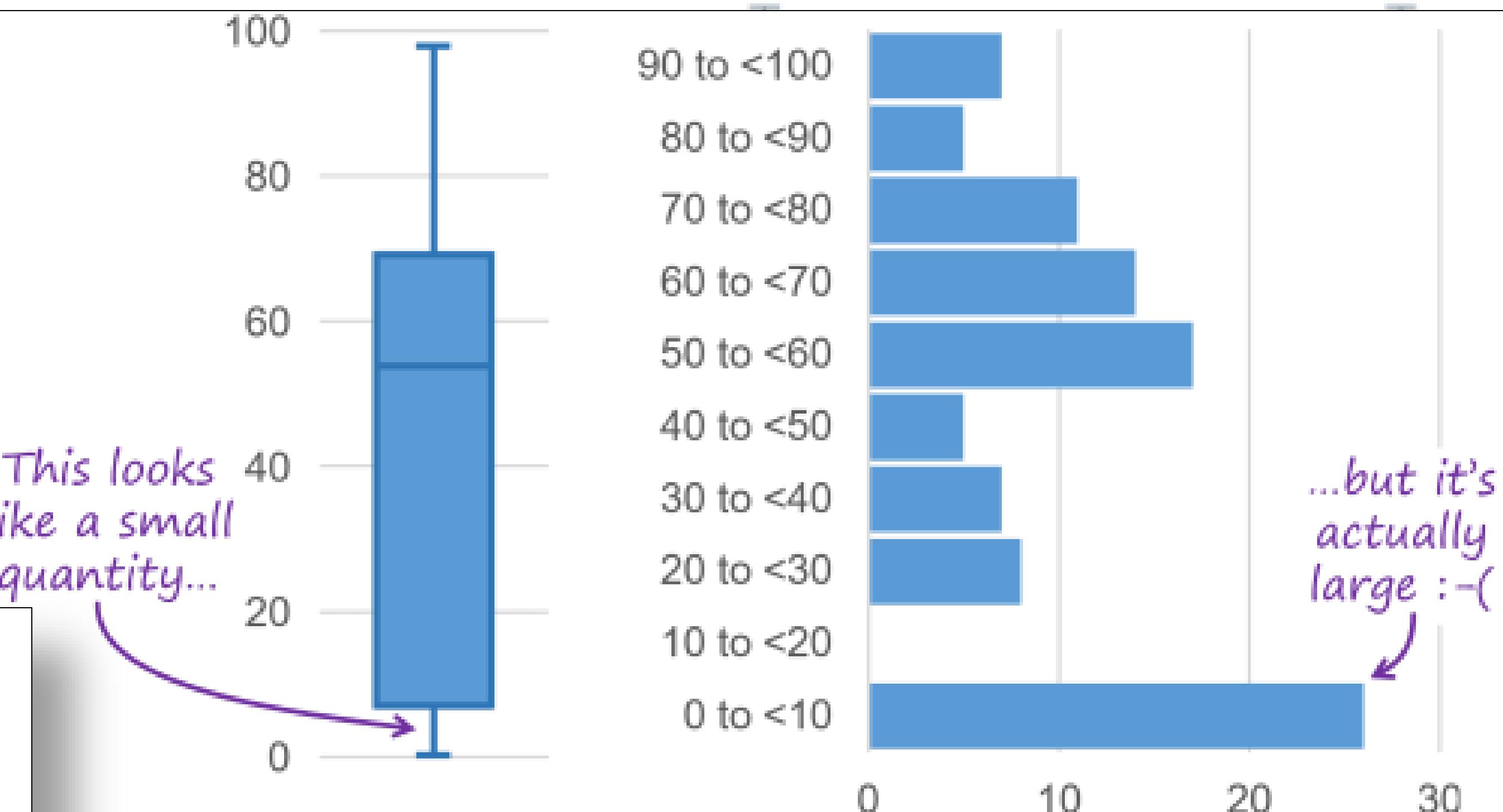
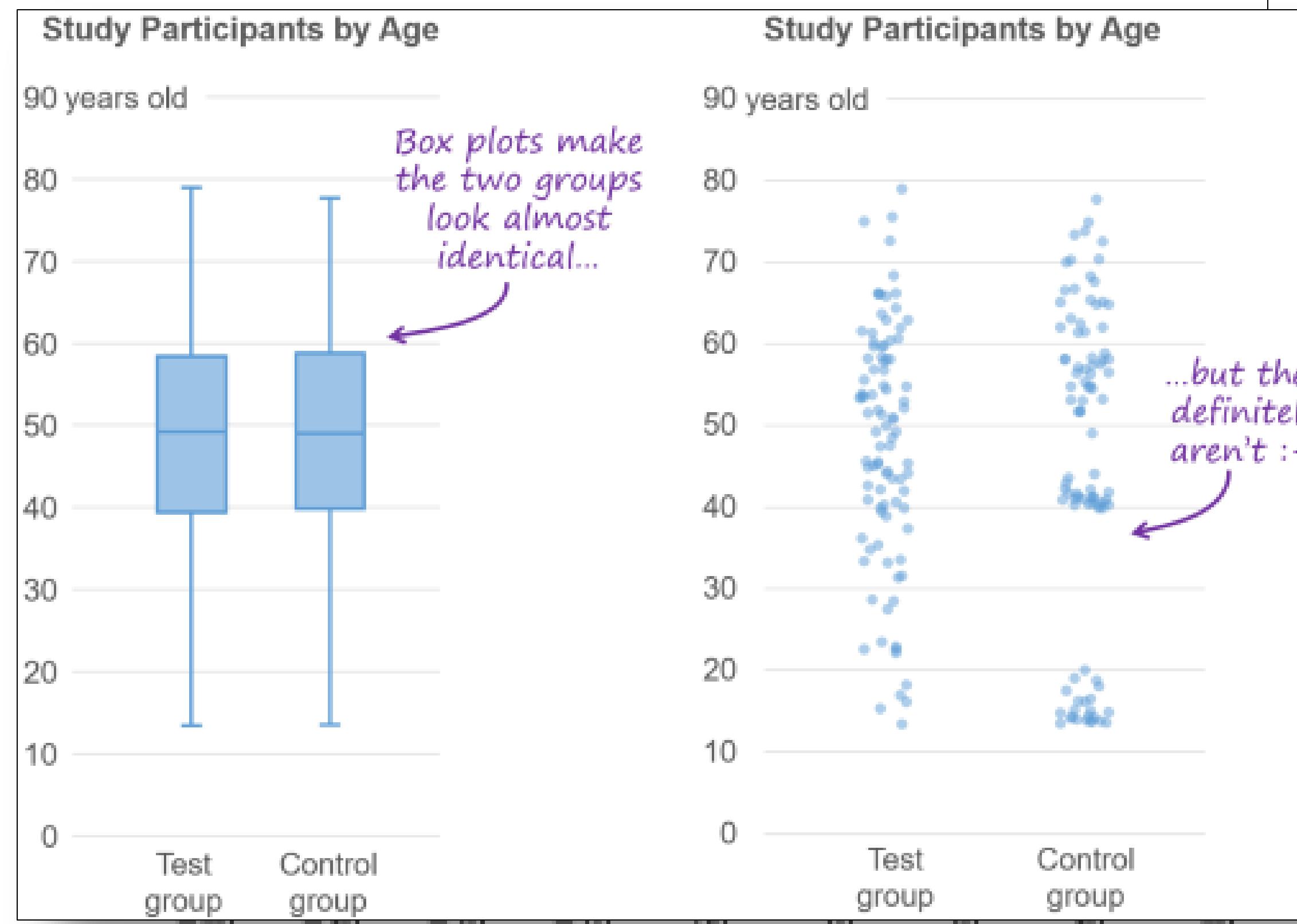
- error
- intervalos de confianza
- Desviacion standard
- Uncertainty
- ...



Box and whisker plots

Limitaciones:

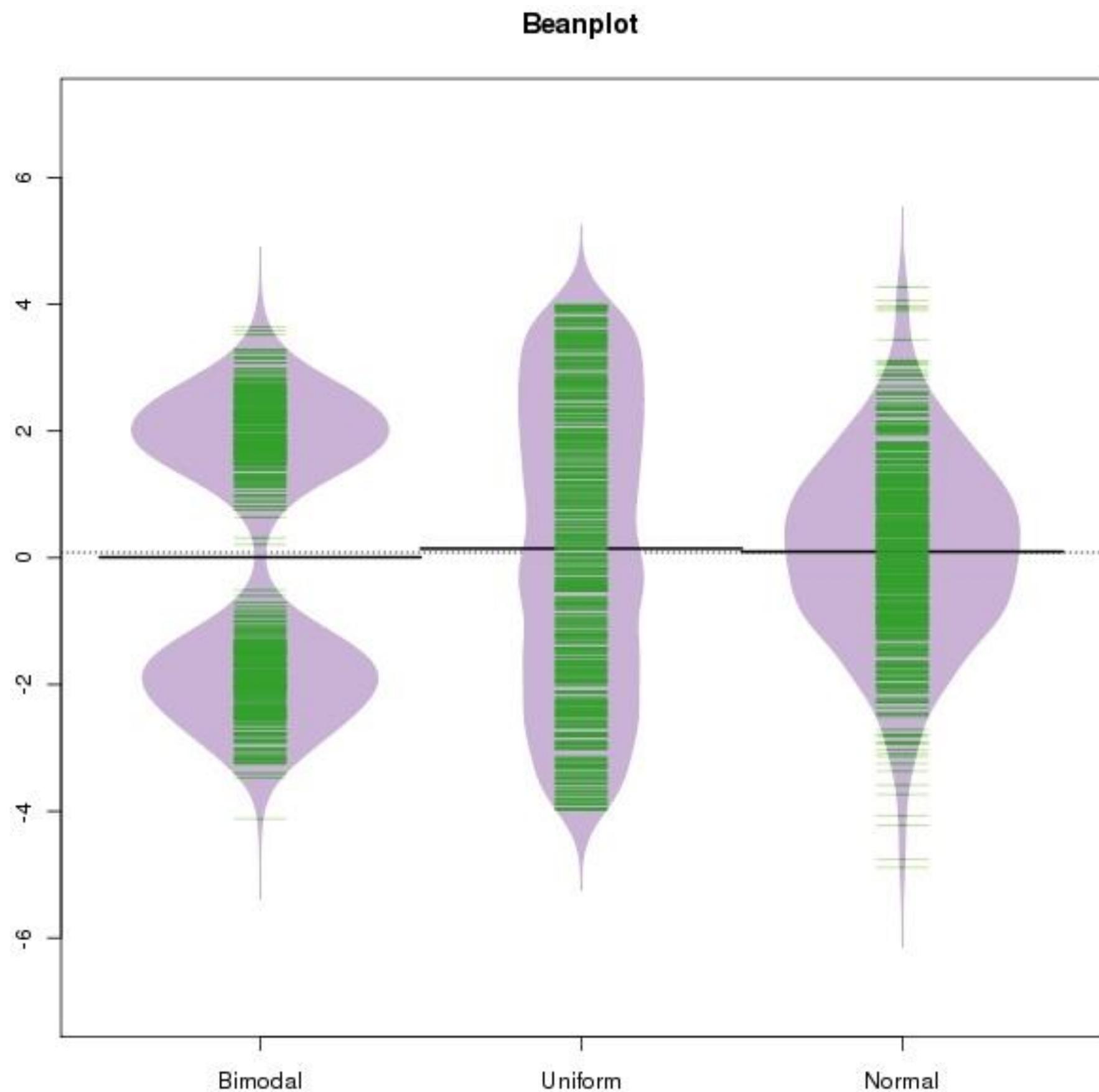
- Box se percibe como unidad
- No muestra toda la distribución
- Asociamos tamaño a cantidad



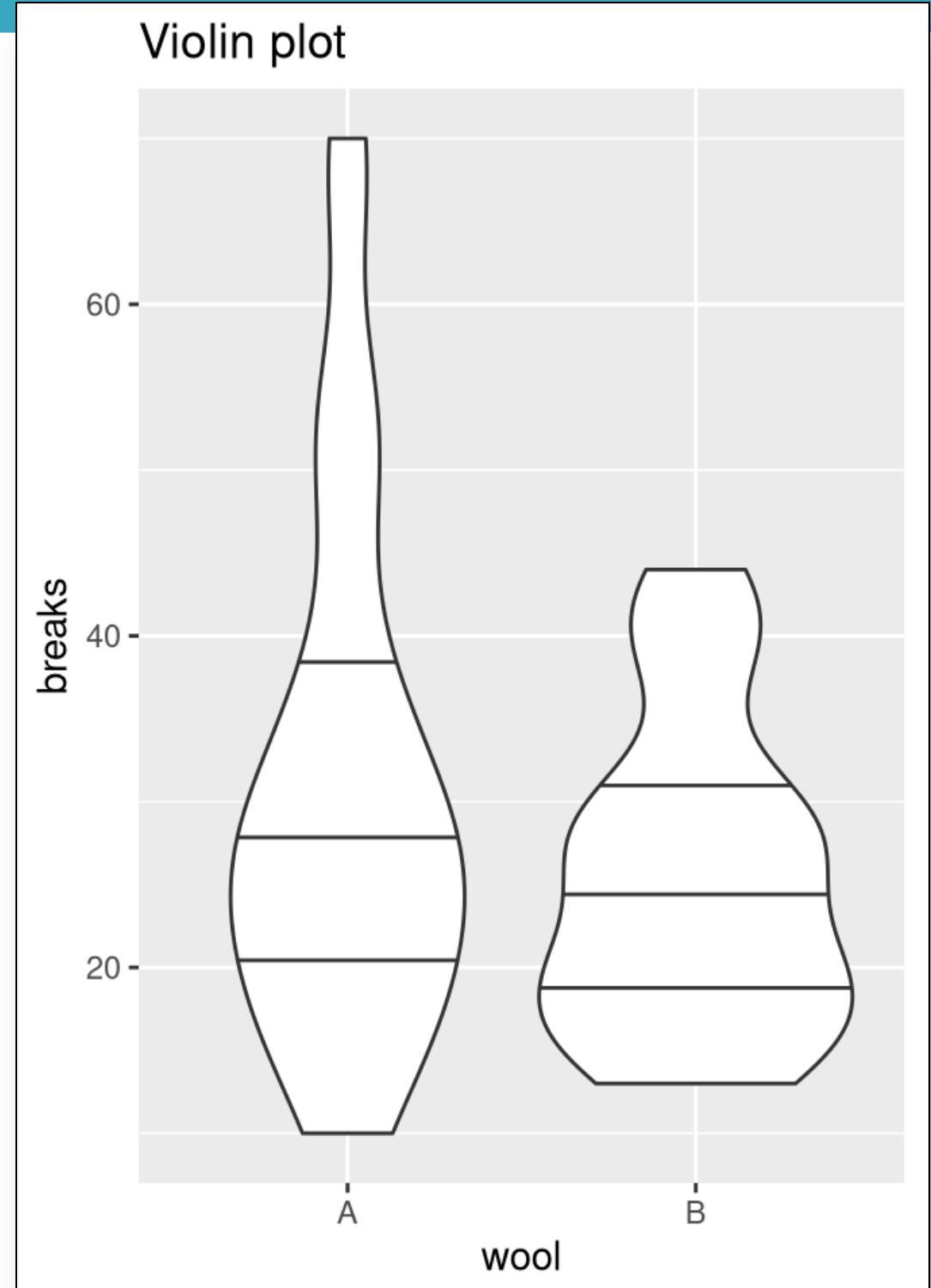
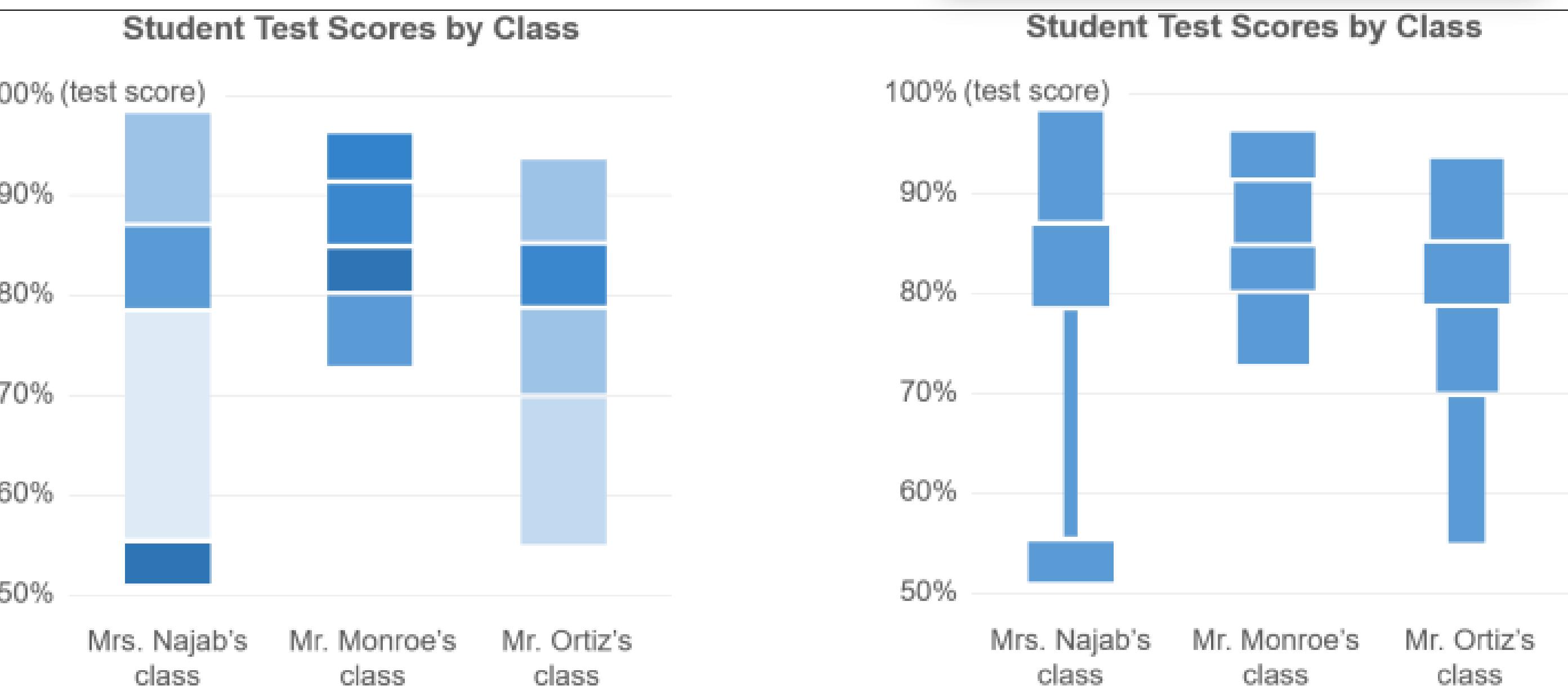
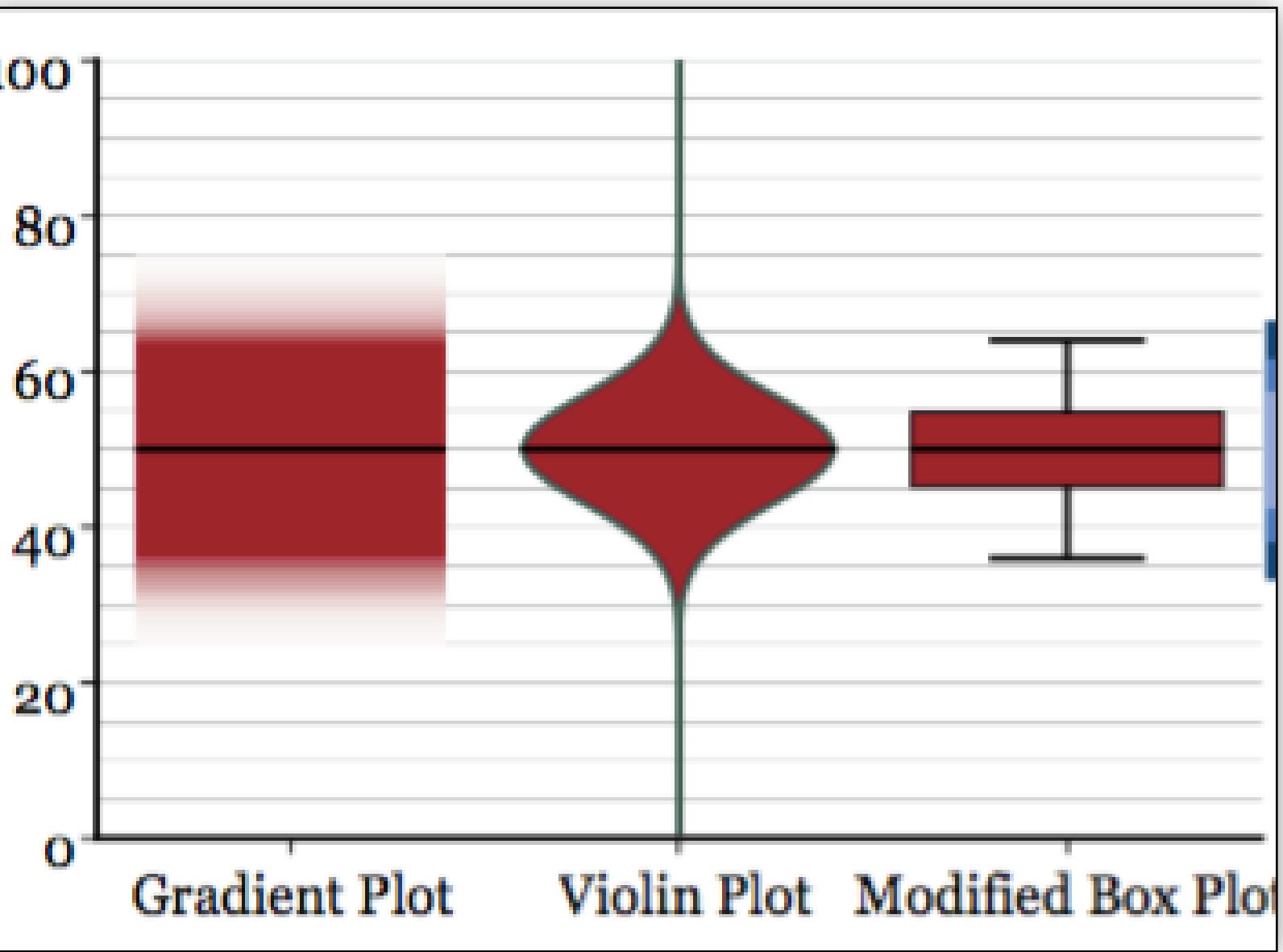
Box and whisker plots

Alternativas:

- Violin, gradient, Bean plots, ...



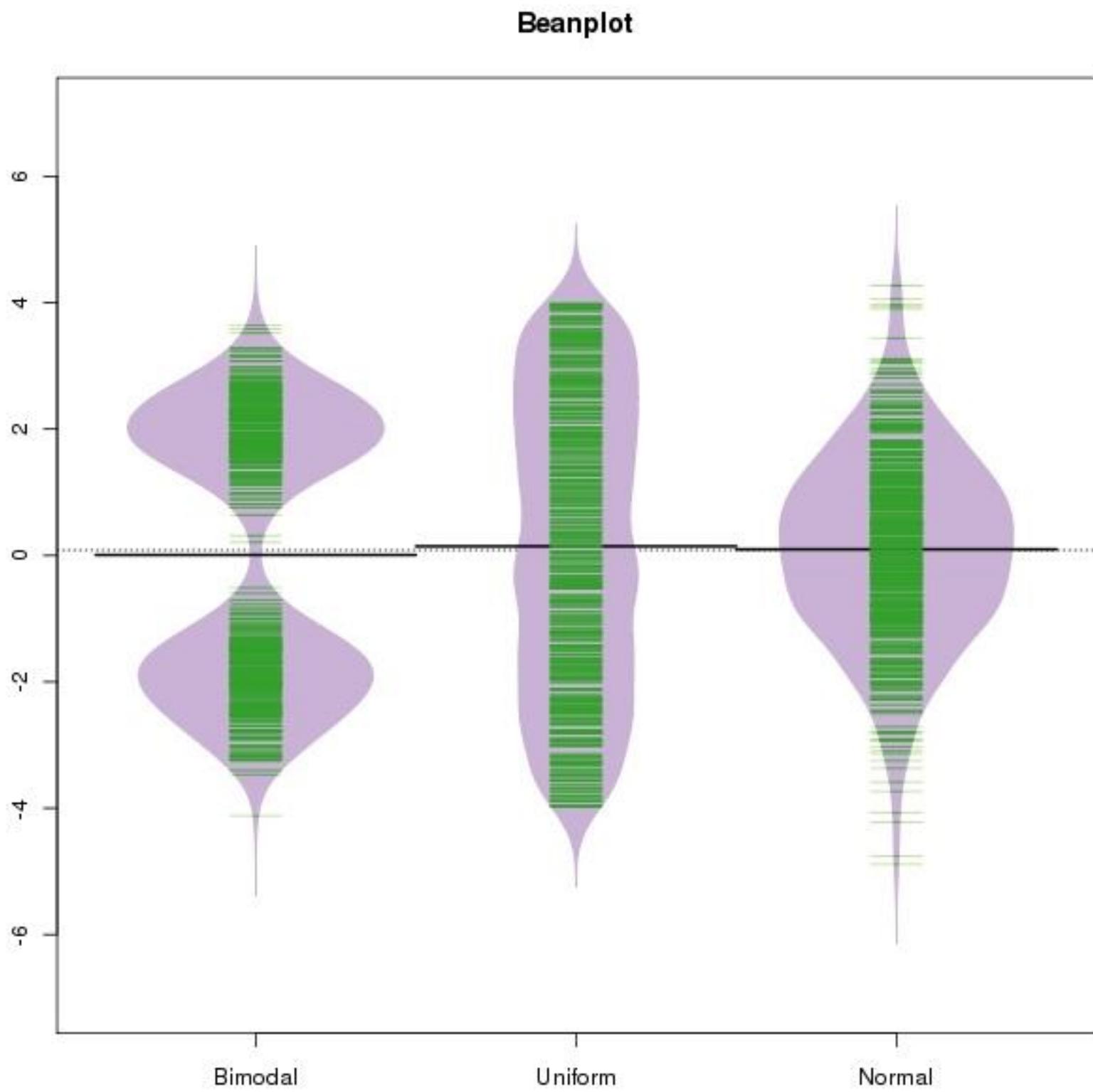
Peter Kampstra, 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, 28



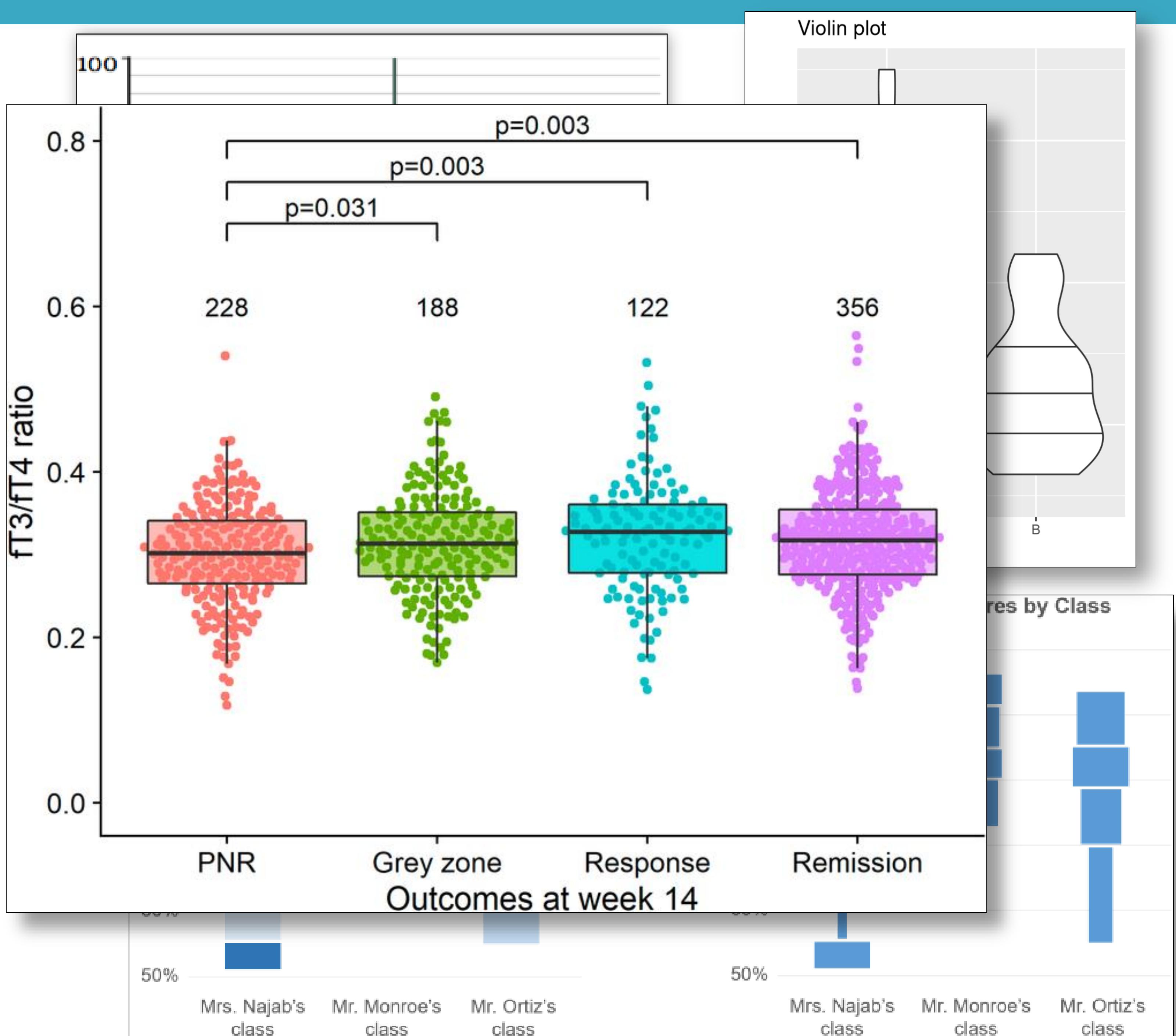
Box and whisker plots

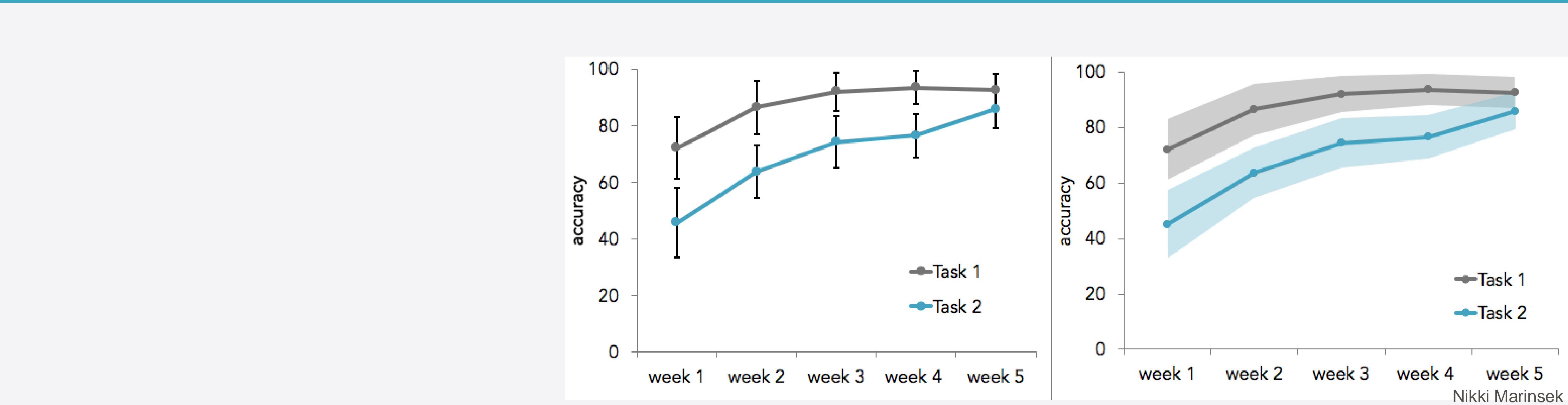
Alternativas:

- Violin, gradient, Bean plots, ...
- Beeswarm



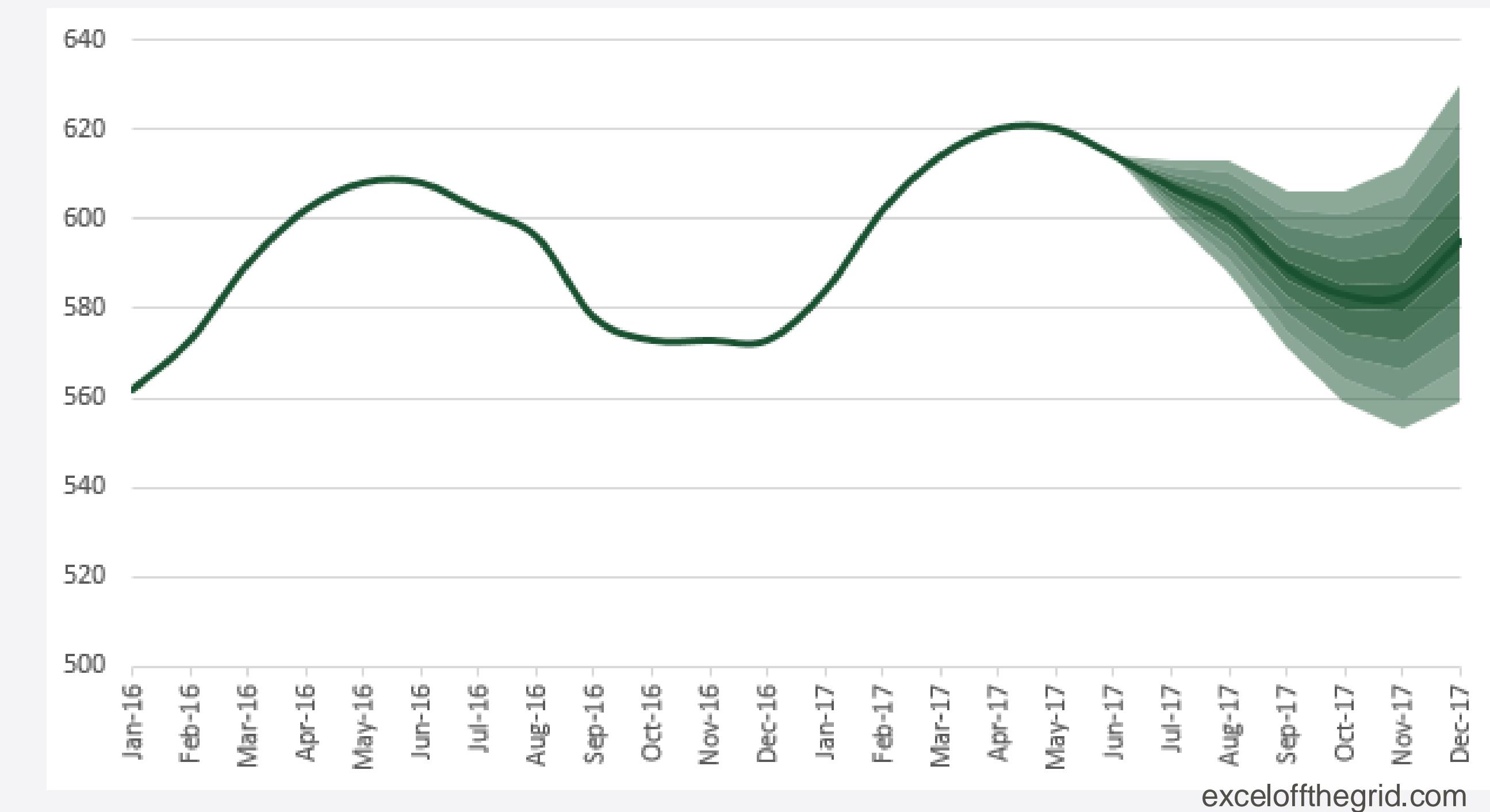
Peter Kampstra, 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, 28



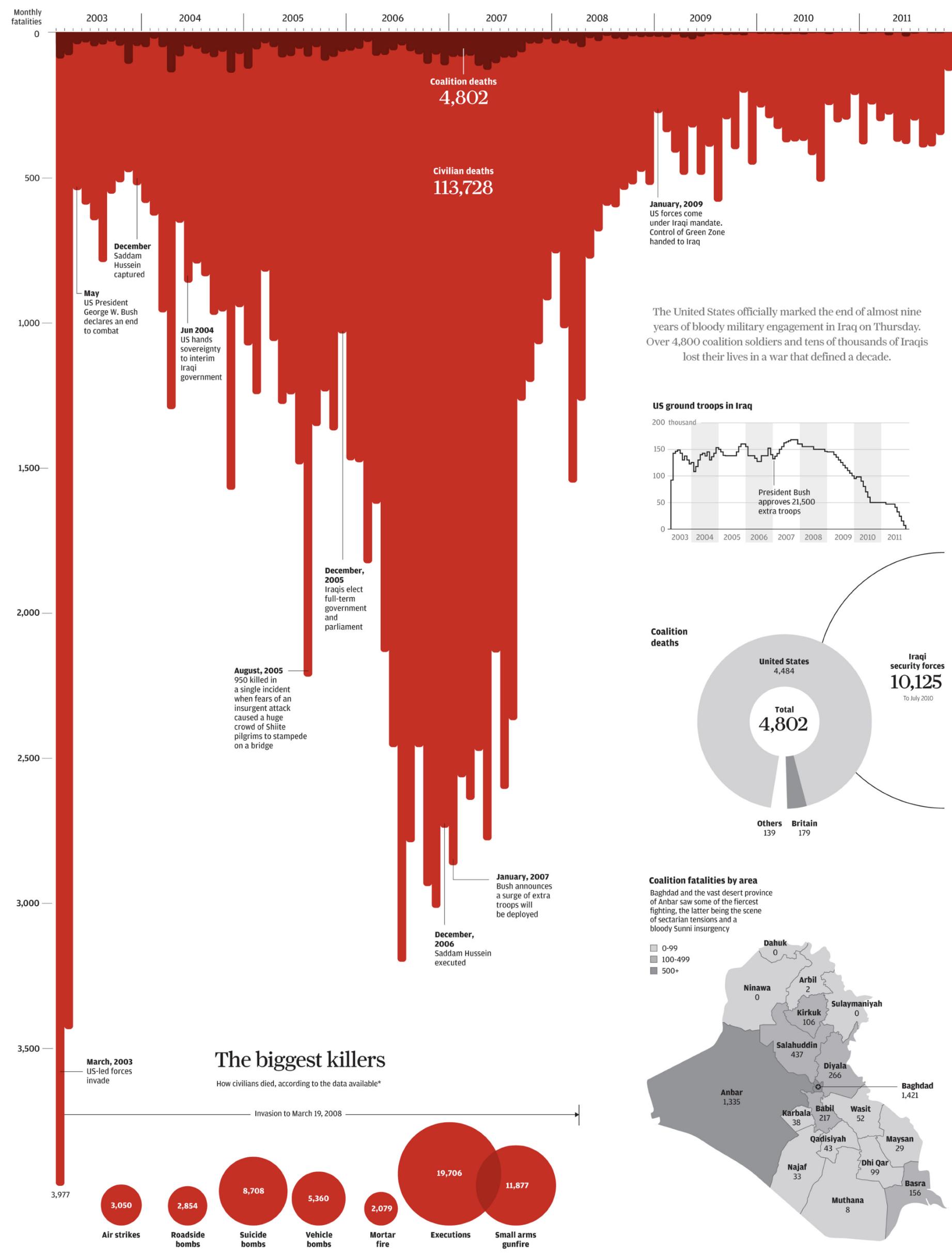


Alternativas a whiskers para
error/uncertainty en linecharts:

- Error bands
- Fancharts

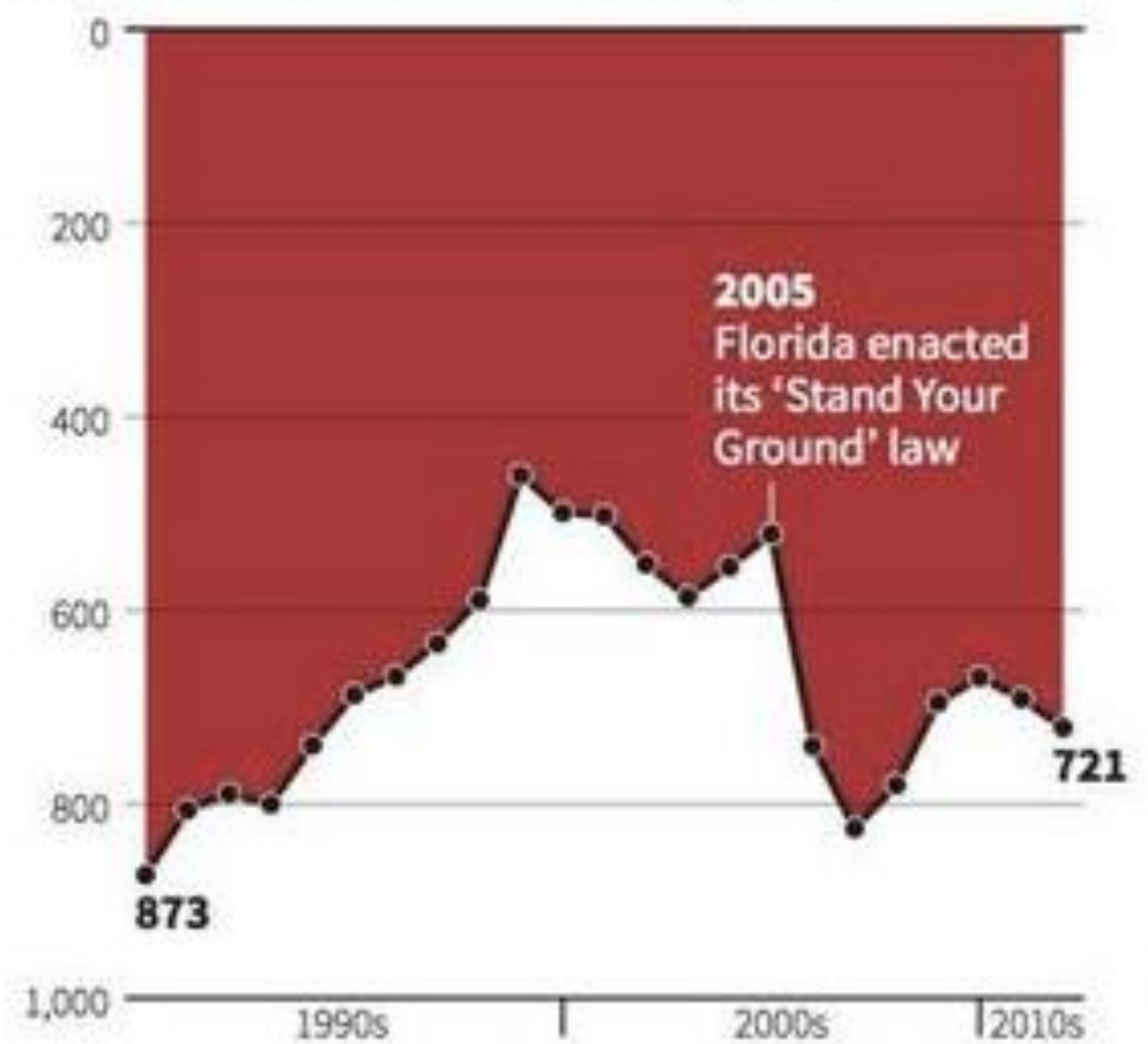


Iraq's bloody toll



Gun deaths in Florida

Number of murders committed using firearms



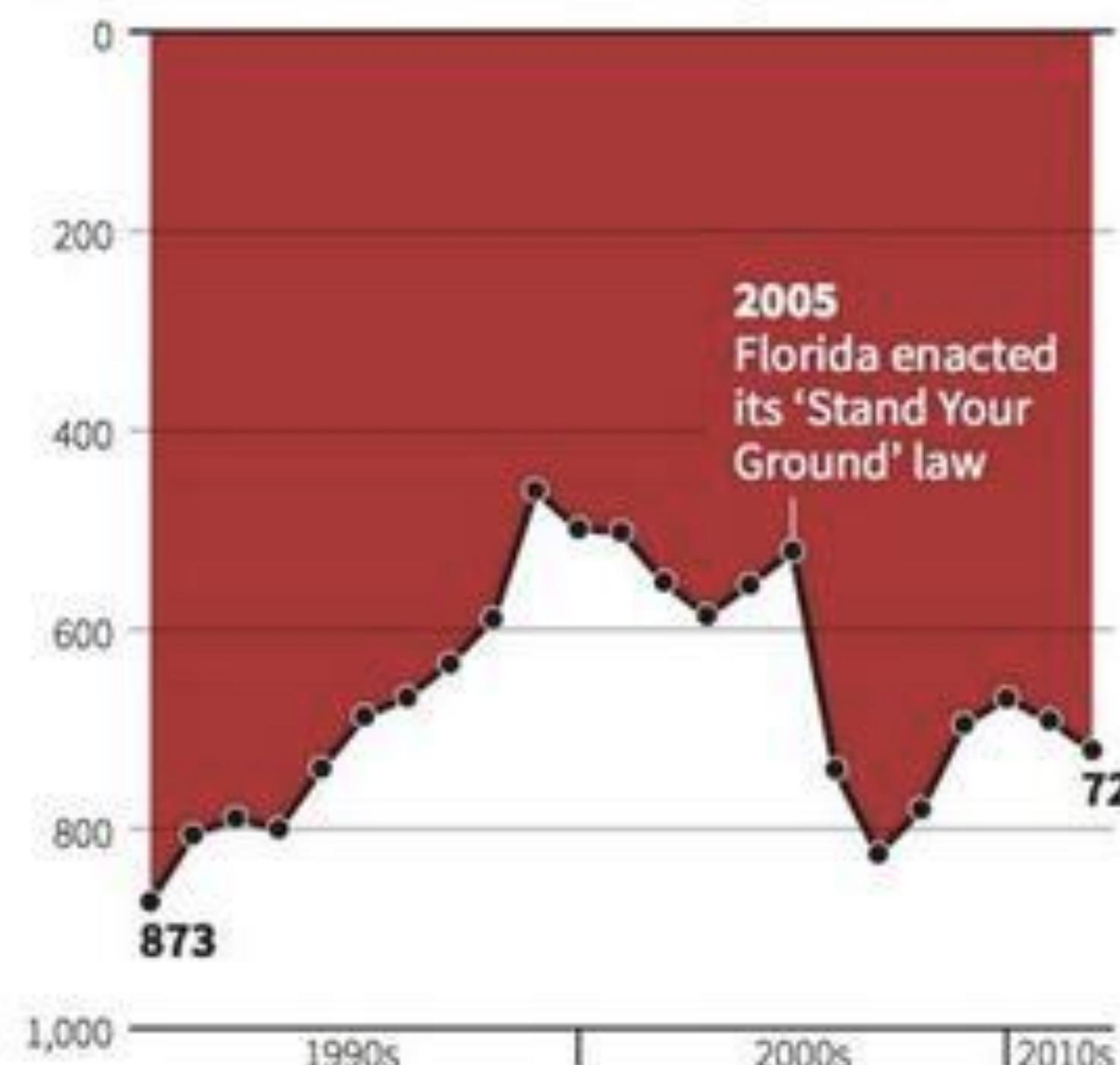
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

Gun deaths in Florida

Number of murders committed using firearms

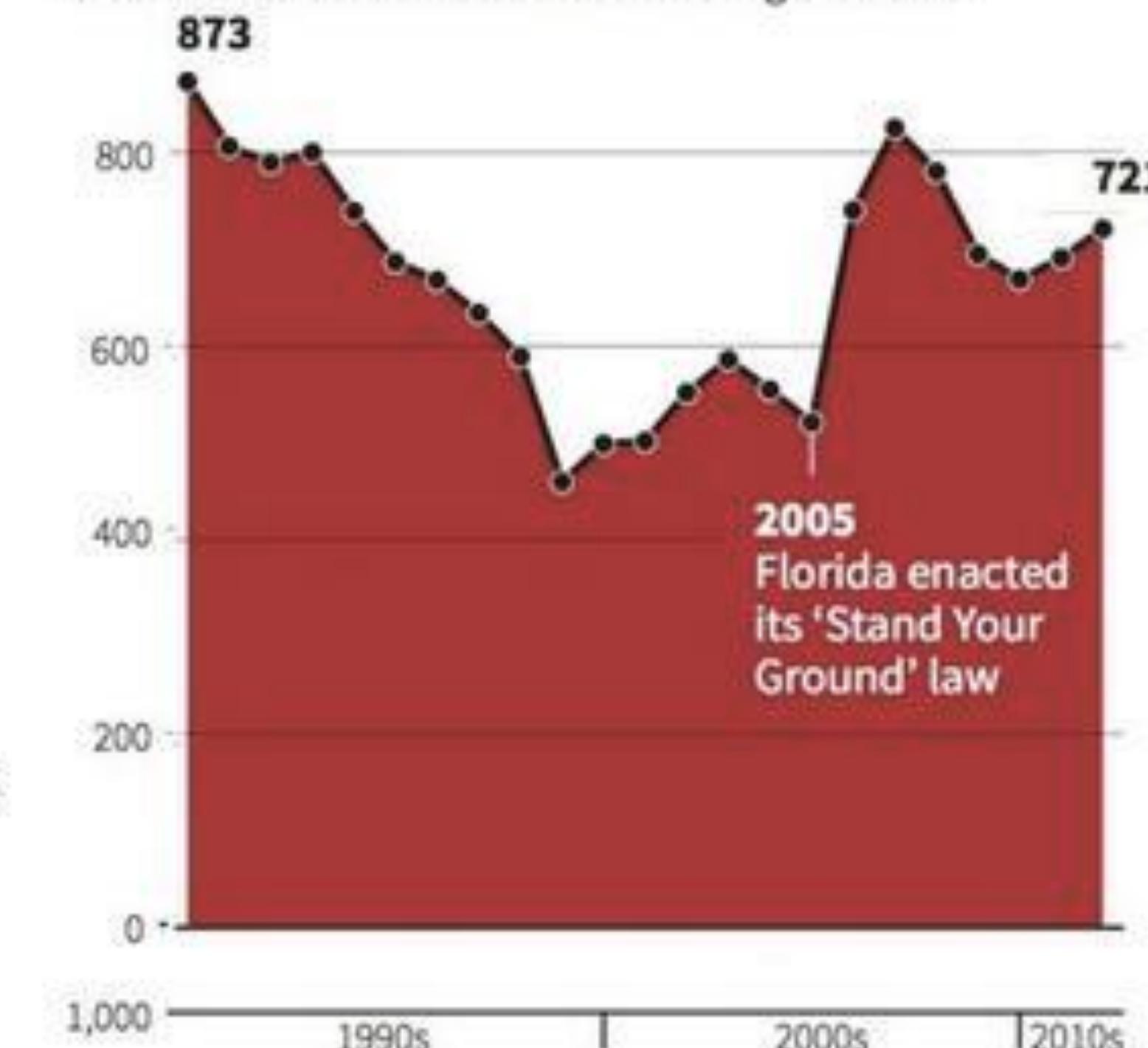


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

BEFORE

AFTER

Referencias

Cairo, A. (2019). How Charts Lie: Getting Smarter about Visual Information. W.W. Norton & Company.

How Charts Lie by Alberto Cairo (video)

<https://youtu.be/oX74Nge8Wkw>