



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Introducción al Análisis y Visualización de Datos con Python

Principios de visualización de información

Felipe Restrepo Calle

ferestrepoca@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial

Facultad de Ingeniería

Universidad Nacional de Colombia

Sede Bogotá

Libro

Visualization Analysis & Design

Tamara Munzner

AK Peters Visualization Series

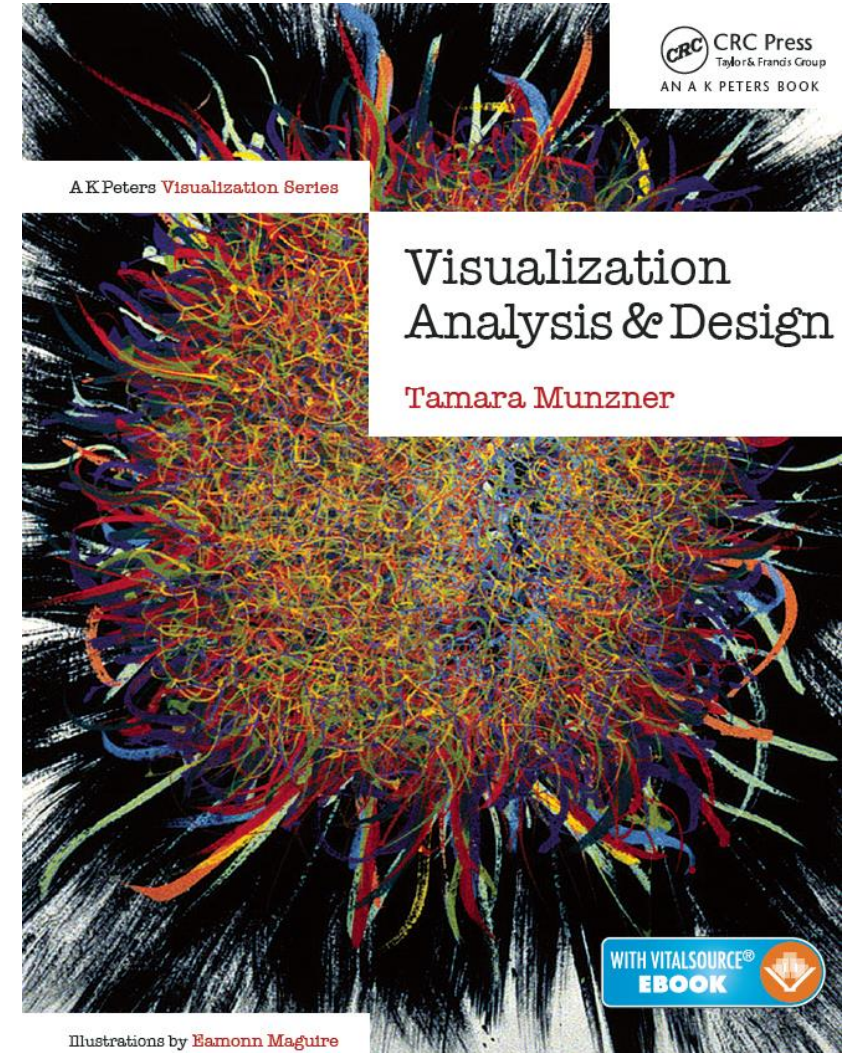
CRC Pres

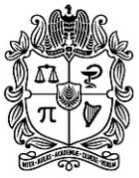
2014

Presentaciones:

<http://www.cs.ubc.ca/~tmm/talks.html>

http://johnguerra.co/lectures/visualAnalytics_fall2018/





Agenda

Introducción

Análisis:

¿Qué?

¿Por qué?

¿Cómo?

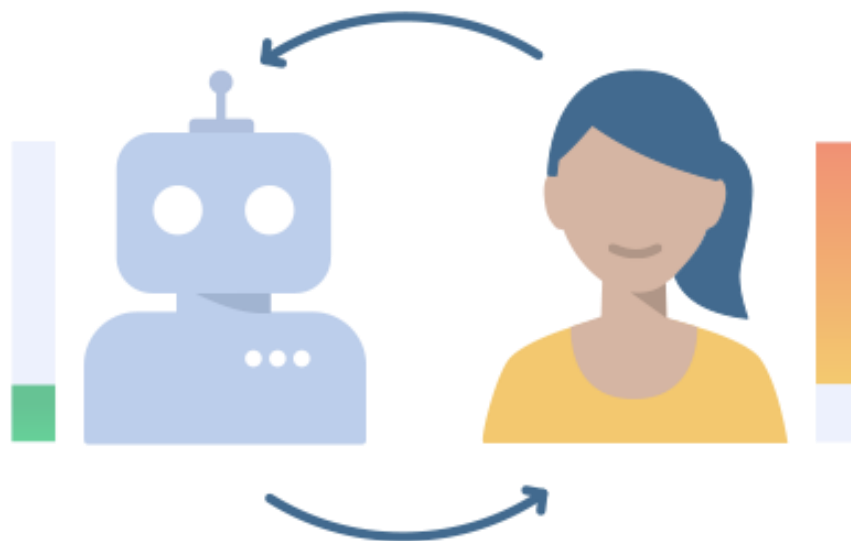
Marcadores y canales

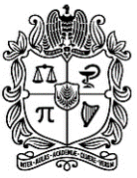
Ejemplos

Introducción: Visualización (vis)

Los sistemas de visualización basados en computador proporcionan representaciones visuales de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

La visualización es adecuada cuando existe la necesidad de aumentar las capacidades humanas en lugar de reemplazar a las personas con métodos computacionales de toma de decisiones.

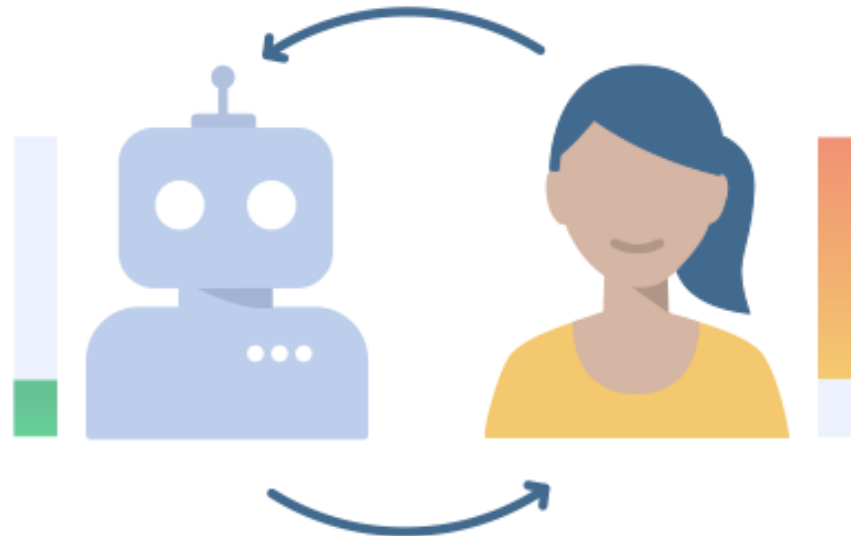




Introducción: Visualización (vis)

Los sistemas de visualización basados en computador proporcionan representaciones visuales de conjuntos de datos diseñadas para **ayudar a las personas** a realizar tareas de manera más efectiva.

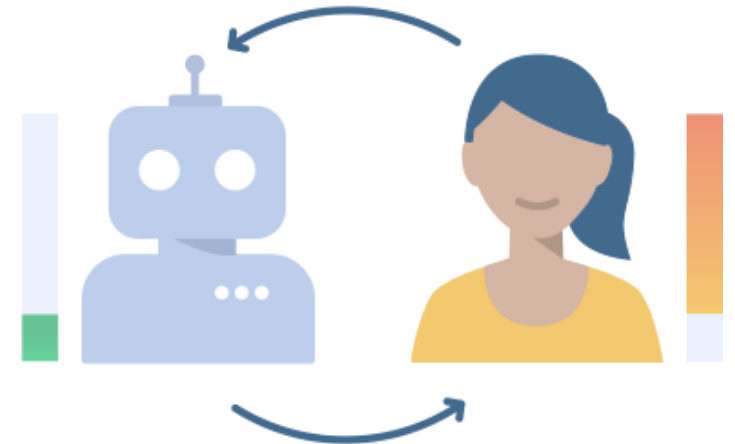
La visualización es adecuada cuando existe la necesidad de **aumentar las capacidades humanas** en lugar de reemplazar a las personas con métodos computacionales de toma de decisiones.

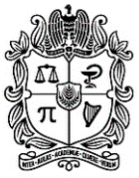


Introducción: Visualización (vis)

¿Por qué involucrar al humano?

- No se necesita visualización si existe una solución completamente automática y de confianza
- Muchos problemas de analítica están mal especificados
 - ✓ No se sabe exactamente qué preguntas hacer con anticipación
- Posibilidades
 - ✓ Los usuarios finales usarán las visualización a largo plazo (ejemplo: análisis exploratorio detallado)
 - ✓ Presentación de resultados conocidos
 - ✓ Sirven para una mejor comprensión de los requisitos antes de desarrollar modelos
 - ✓ Ayudar a desarrolladores de soluciones automáticas a refinar/depurar y determinar parámetros
 - ✓ Ayudar a los usuarios finales de soluciones automáticas a verificar





Introducción: Visualización (vis)

Los sistemas de **visualización basados en computador** proporcionan representaciones visuales de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

¿Por qué usar computadores?

Introducción: Visualización (vis)

Los sistemas de **visualización basados en computador** proporcionan representaciones visuales de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

¿Por qué usar computadores?

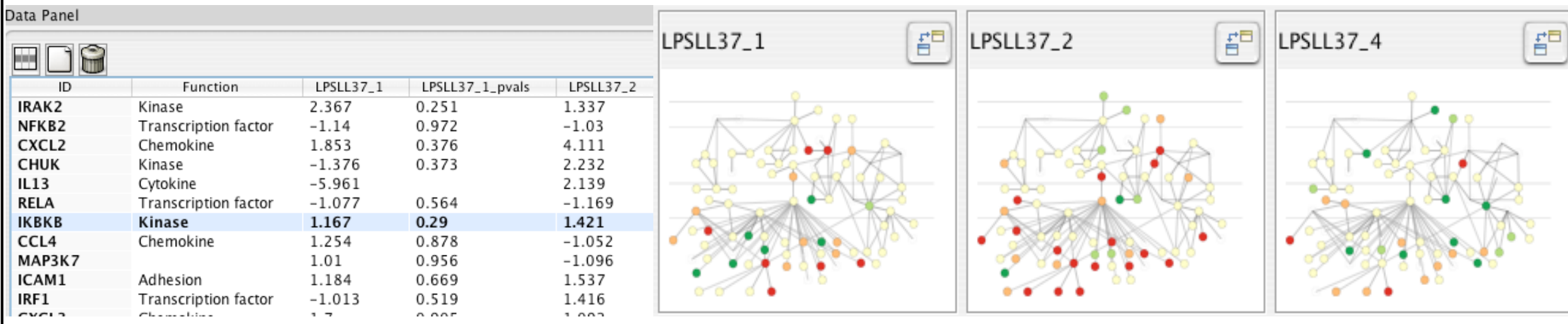
- ✓ Más allá de la paciencia humana
- ✓ Escalabilidad a grandes conjuntos de datos
- ✓ Soportan interactividad

Introducción: Visualización (vis)

Los sistemas de visualización basados en computador proporcionan **representaciones visuales** de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

¿Por qué usar una representación externa?

Representación externa: reemplaza la cognición con la percepción

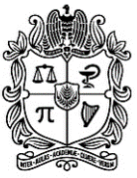


Introducción: Visualización (vis)

Los sistemas de visualización basados en computador proporcionan representaciones **visuales** de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

¿Por qué depender de la visión?

- ✓ El sistema visual humano es un canal de ancho de banda alto para el cerebro
 - Visión general - procesamiento en segundo plano
 - Experiencia subjetiva de ver todo simultáneamente
 - Un procesamiento significativo ocurre en paralelo
- ✓ Sonido: menor ancho de banda y diferente semántica
 - Audición general - no admitida
- ✓ Táctil / háptica: poca capacidad de grabación/reproducción
 - ✓ Solo comunicación de ancho de banda muy bajo hasta el momento
- ✓ Sabor, olor: no hay dispositivos de grabación/reproducción viables



Introducción: Visualización (vis)

Los sistemas de visualización basados en computador proporcionan representaciones visuales de conjuntos de datos diseñadas para ayudar a las personas a realizar tareas de manera más efectiva.

¿Por qué visualizar los datos en detalle?

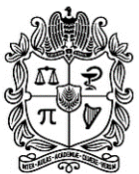
- ✓ Los resúmenes pierden información, los detalles importan
- ✓ Sirve para confirmar lo esperado y encontrar patrones inesperados
- ✓ Útil para evaluar la validez del modelo estadístico

Introducción: Visualización (vis)

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

¿Por qué visualizar los datos en detalle?

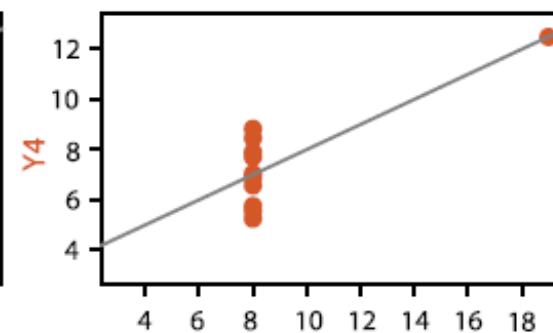
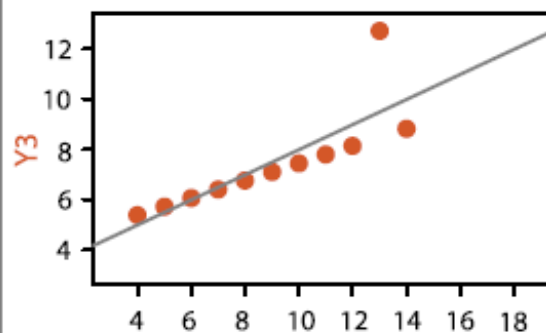
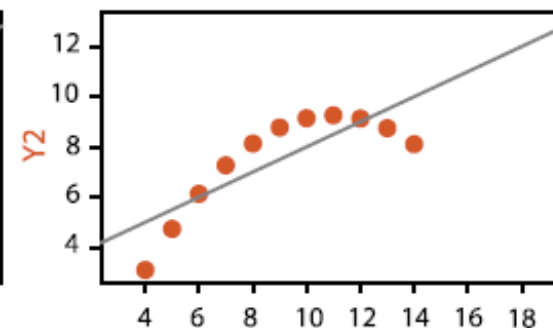
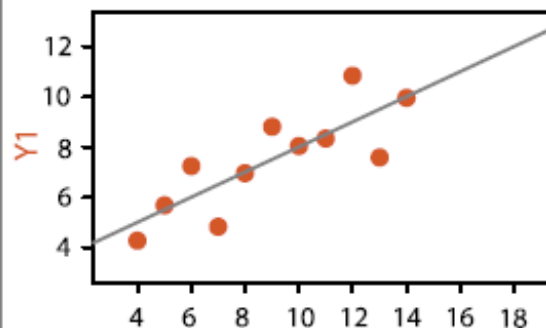


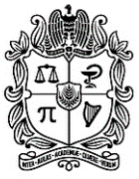
Introducción: Visualización (vis)

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

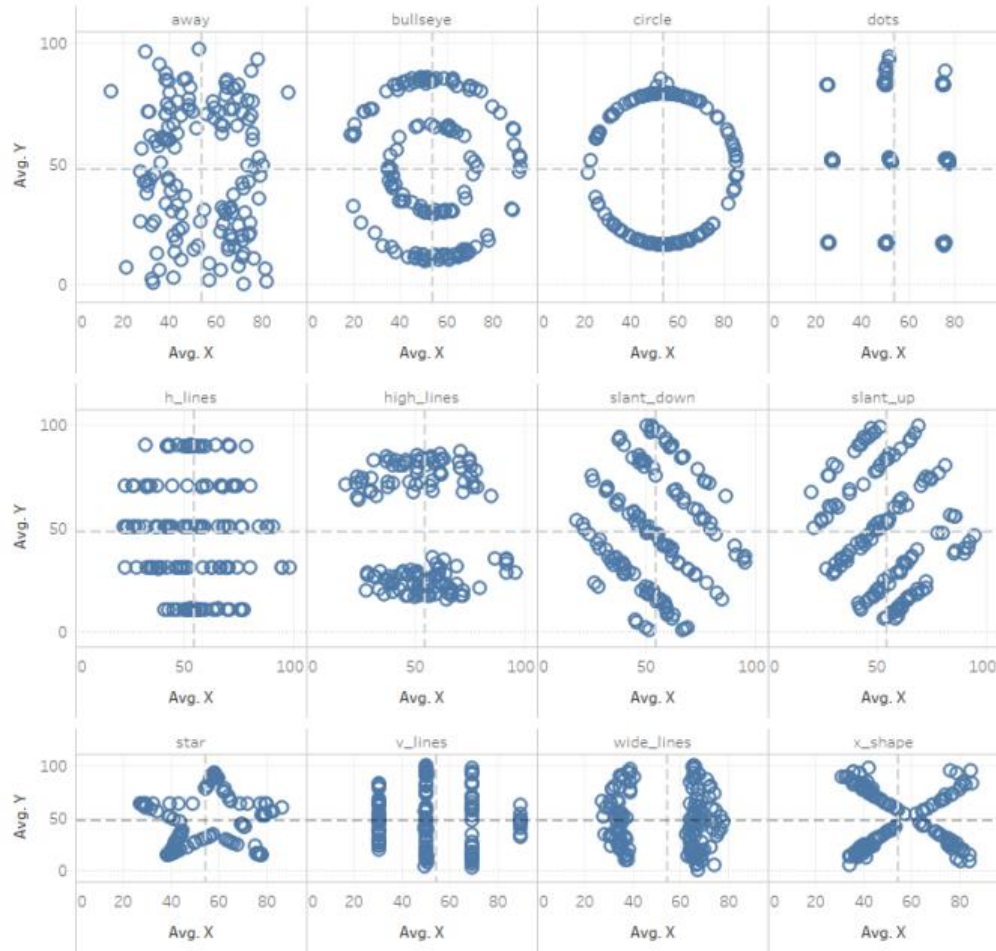
¿Por qué visualizar los datos en detalle?



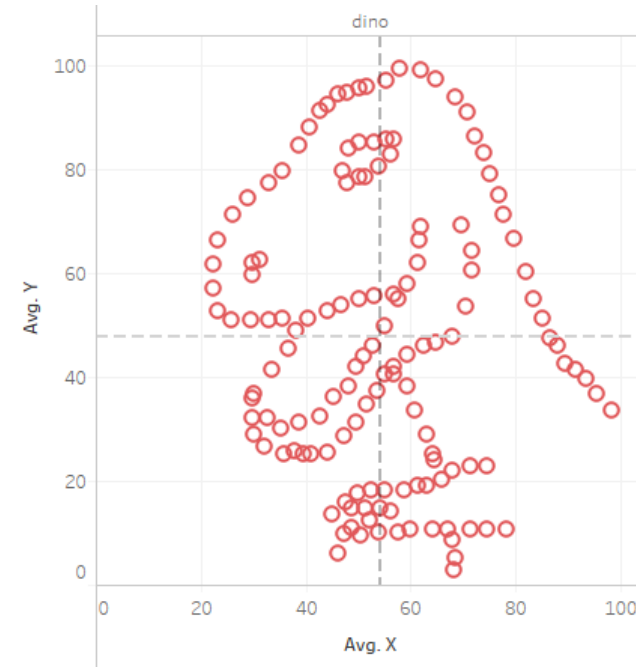


Introducción: Visualización (vis)

Datasaurus Dozen



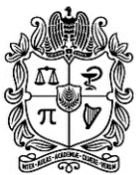
¿Por qué visualizar los datos en detalle?



<https://dabblingwithdata.wordpress.com/2017/05/03/the-datasaurus-a-monstrous-anscombe-for-the-21st-century/>

Introducción: Limitaciones de recursos

- **Límites computacionales**
 - ✓ Tiempo de procesamiento
 - ✓ Memoria del sistema
- **Límites humanos**
 - ✓ Atención humana
 - ✓ Memoria
 - ✓ Retención
- **Límites de las pantallas**
 - ✓ Los píxeles son recursos preciosos, el recurso más limitado
 - ✓ Densidad de información: tasa de espacio utilizado para codificar información vs. espacios en blanco no utilizados. Compromiso entre desorden y desperdicio de espacio. Necesario encontrar un punto óptimo entre denso y disperso.



Agenda

Introducción

Análisis:

¿Qué?

¿Por qué?

¿Cómo?

Marcadores y canales

Ejemplos

Análisis: Marco de trabajo para análisis visual

Dominio: ¿Quiénes son los usuarios objetivo?

Abstracción: Traducción de los detalles del dominio al vocabulario de visualización

What? ¿Qué se muestra? Abstracción de **datos**

No solo dibujar lo que le dan: transformar a nueva forma

Why? ¿Por qué le interesa al usuario? Abstracción de **tareas**

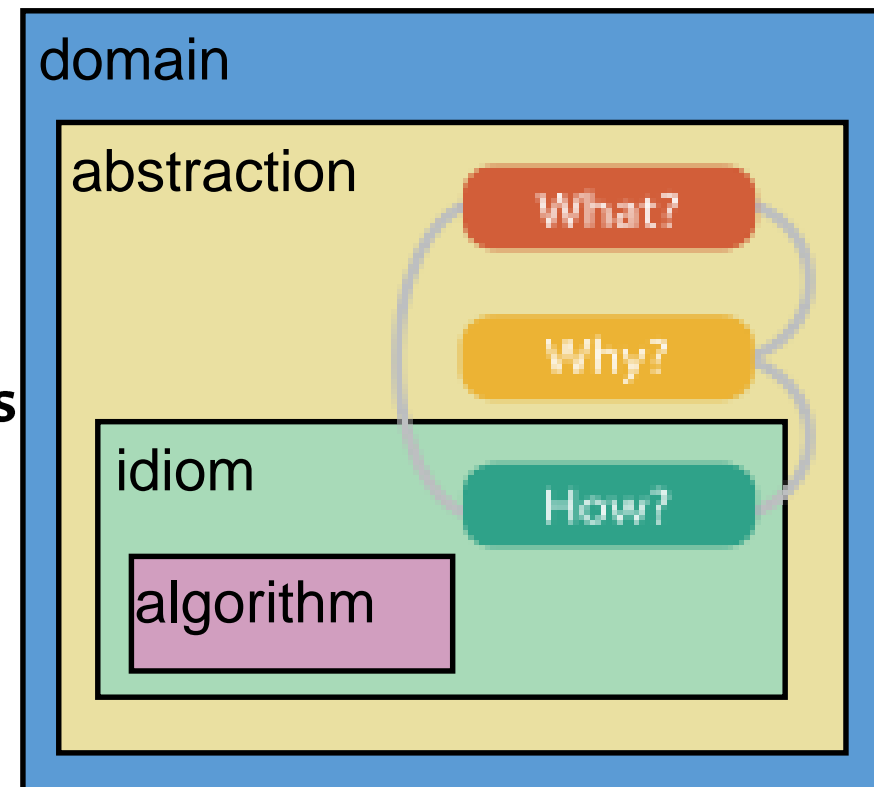
Representación (idiom)

How? ¿Cómo se muestra?

- Codificación visual: ¿cómo dibujarlo?
- Interacción: ¿cómo manipularlo?

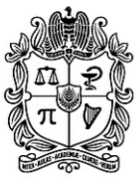
Algoritmo

- Computación eficiente

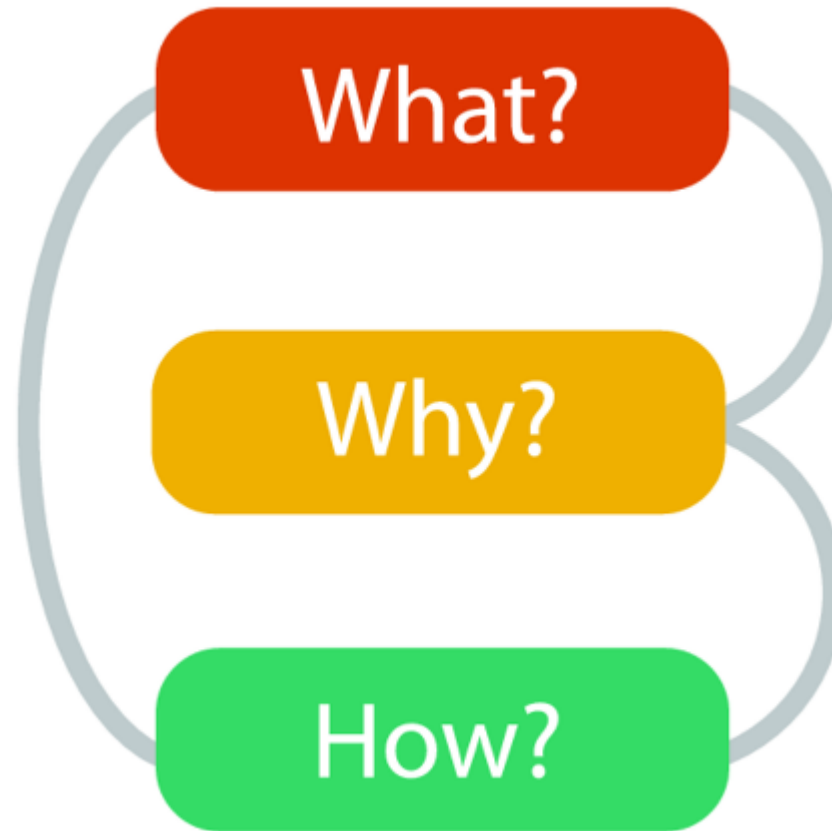


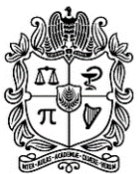
A Multi-Level Typology of Abstract Visualization Tasks

Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).



Análisis: Marco de trabajo para análisis visual





Análisis: ¿Qué?



What?

Datasets

➔ Data Types

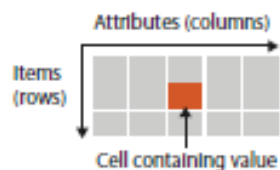
→ Items → Attributes → Links → Positions → Grids

➔ Data and Dataset Types

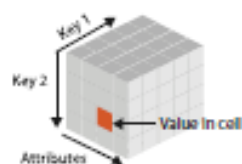
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

➔ Dataset Types

→ Tables



→ Multidimensional Table



→ Geometry (Spatial)



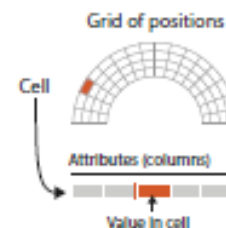
→ Networks



→ Trees



→ Fields (Continuous)



➔ Dataset Availability

→ Static



➔ Attribute Types

→ Categorical



→ Ordered

→ Ordinal



→ Quantitative



➔ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



Análisis: ¿Qué? – Tipos de datos y conjuntos de datos



➔ 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 Availability

➔ Static



➔ Dynamic

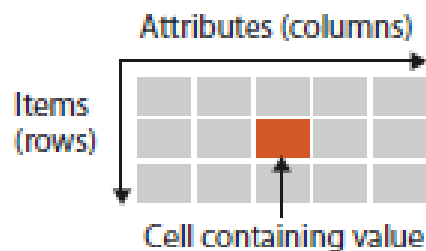


Análisis: ¿Qué? – Tipos de conjuntos de datos

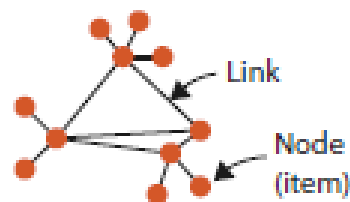
→ Dataset Types



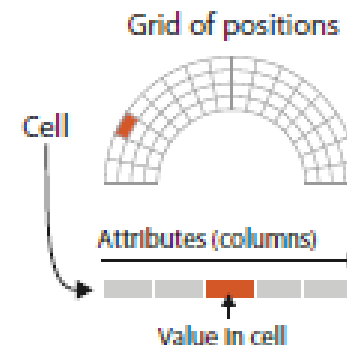
→ Tables



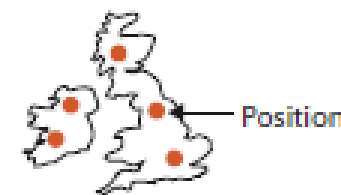
→ Networks



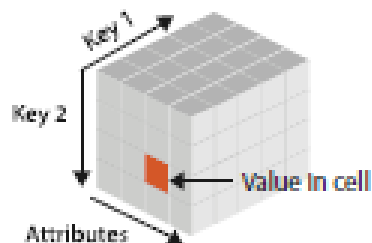
→ Fields (Continuous)



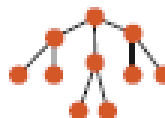
→ Geometry (Spatial)



→ Multidimensional Table



→ Trees



Análisis: ¿Qué? – Tipos de atributos



Attributes

➔ Attribute Types

➔ Categorical

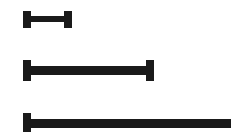


➔ Ordered

➔ Ordinal



➔ Quantitative



➔ Ordering Direction

➔ Sequential

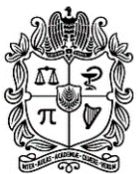


➔ Diverging



➔ Cyclic





Análisis: ¿Qué?



What?

Datasets

➔ Data Types

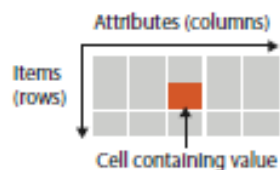
➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

➔ Data and Dataset Types

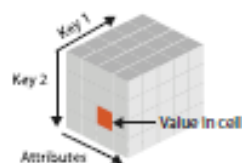
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

➔ Dataset Types

➔ Tables



➔ Multidimensional Table



➔ Geometry (Spatial)



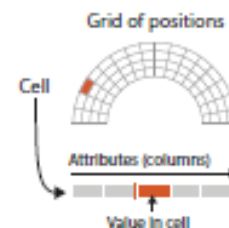
➔ Networks



➔ Trees



➔ Fields (Continuous)



➔ Dataset Availability

➔ Static



➔ Ordering Direction

➔ Sequential



➔ Diverging



➔ Cyclic



➔ Attribute Types

➔ Categorical



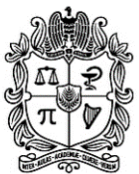
➔ Ordered

➔ Ordinal



➔ Quantitative





Análisis: ¿Por qué?



{action, target} pairs

- discover distribution
- compare trends
- locate outliers
- browse topology

Why?

Actions

Targets

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



All Data

→ Trends



→ Outliers



→ Features



Attributes

→ One

→ Distribution



→ Extremes

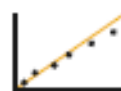


→ Many

→ Dependency



→ Correlation

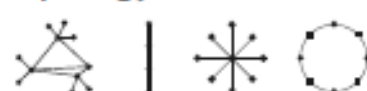


→ Similarity



Network Data

→ Topology



→ Paths



Spatial Data

→ Shape



Análisis: ¿Por qué? – Acciones: analizar

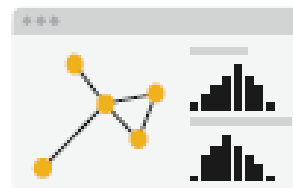
➔ Analyze

➔ Consume

➔ Discover



➔ Present



➔ Enjoy



➔ Produce

➔ Annotate

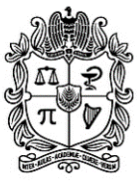


➔ Record



➔ Derive





Análisis: ¿Por qué? – Acciones: ejemplo “disfrutar”

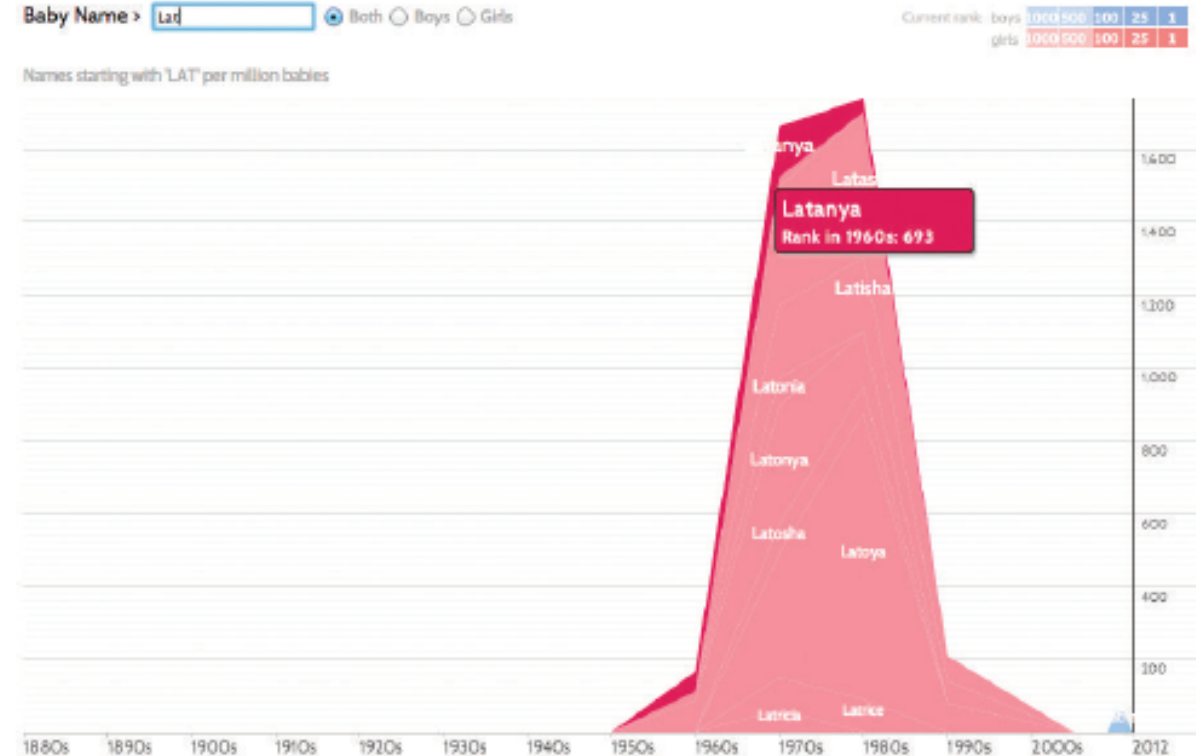
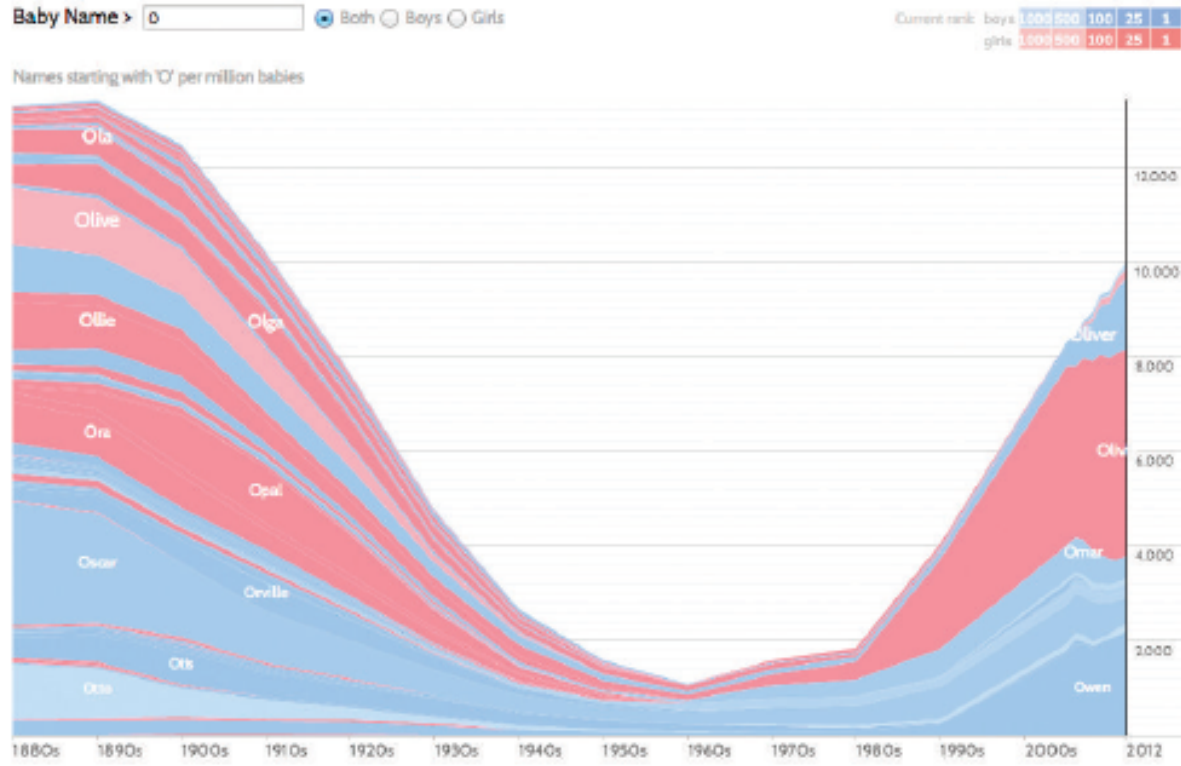
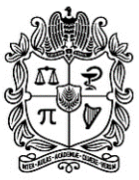


Figure 3.3. Name Voyager, a vis tool originally intended for parents focused deciding on what to name their expected baby, ended up being used by many nonparents to analyze historical trends for their own enjoyment. Left: Names starting with ‘O’ had a notable dip in popularity in the middle of the century. Right: Names starting with ‘LAT’ show a trend of the 1970s. After [Wattenberg 05, Figures 2 and 3], using <http://www.babynamewizard.com>.

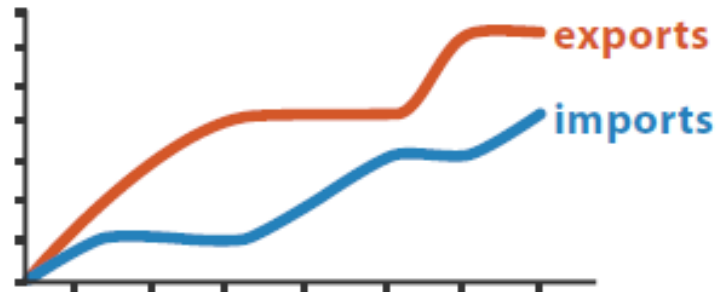
Análisis: ¿Por qué? – Acciones: ejemplo “grabar”



Figure 3.4. Graphical history recorded during an analysis session with Tableau. From [Heer et al. 08, Figure 1].

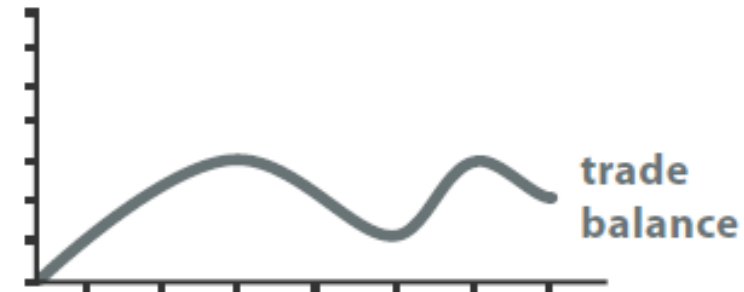


Análisis: ¿Por qué? – Acciones: ejemplo “derivar”



Original Data

(a)



$$\text{trade balance} = \text{exports} - \text{imports}$$

Derived Data


(b)

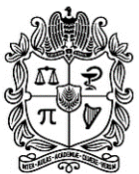
Figure 3.5. Derived attributes can be directly visually encoded. (a) Two original data attributes are plotted, imports and exports. (b) The quantitative derived attribute of trade balance, the difference between the two originals, can be plotted directly.

Análisis: ¿Por qué? – Acciones: buscar



➔ Search

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

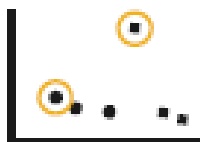


Análisis: ¿Por qué? – Acciones: consultar

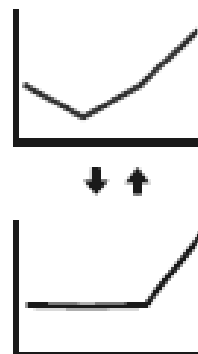


➔ Query

➔ Identify



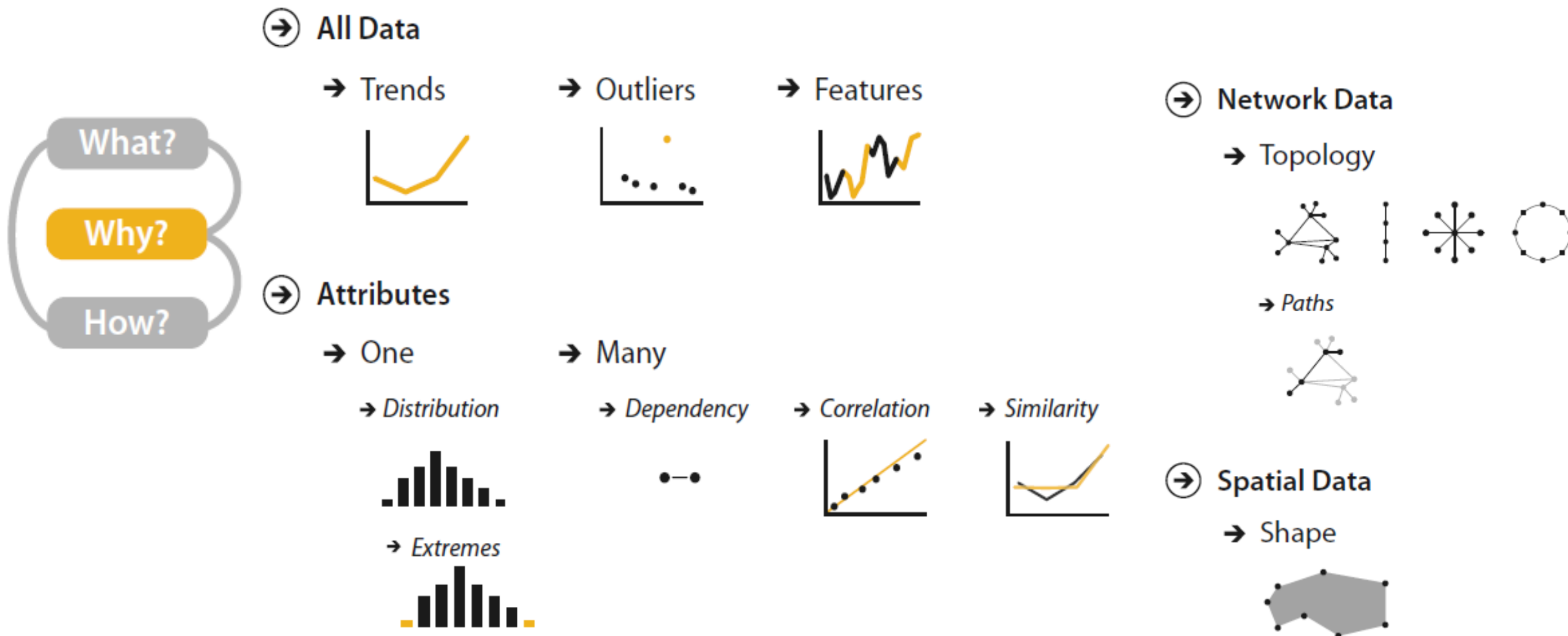
➔ Compare

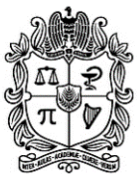


➔ Summarize



Análisis: ¿Por qué? – Objetivos (targets)





Análisis: ¿Por qué?



{action, target} pairs

- discover distribution
- compare trends
- locate outliers
- browse topology

Why?

Actions

Targets

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



All Data

→ Trends



→ Outliers



→ Features



Attributes

→ One

→ Distribution



→ Extremes



→ Many

→ Dependency



→ Correlation

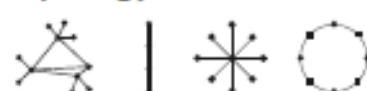


→ Similarity



Network Data

→ Topology



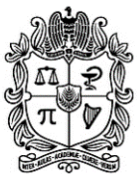
→ Paths



Spatial Data

→ Shape





Análisis: ¿Cómo?

How?

Encode

➔ Arrange

➔ Express



➔ Separate



➔ Order



➔ Align



➔ Use



➔ Map

from **categorical** and **ordered** attributes

➔ Color

➔ Hue



➔ Saturation



➔ Luminance



➔ Size, Angle, Curvature, ...



➔ Shape



➔ Motion

Direction, Rate, Frequency, ...



Manipulate

➔ Change



➔ Select



➔ Navigate



Facet

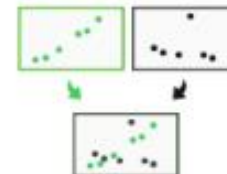
➔ Juxtapose



➔ Partition



➔ Superimpose



Reduce

➔ Filter



➔ Aggregate



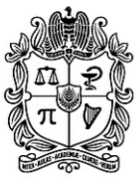
➔ Embed



What?

Why?

How?



Agenda

Introducción

Análisis:

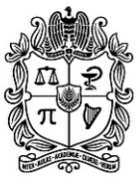
¿Qué?

¿Por qué?

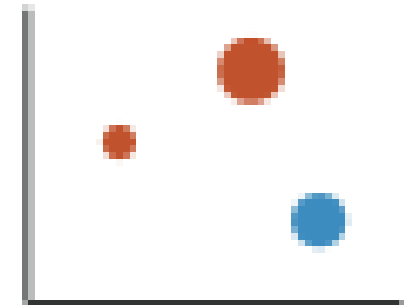
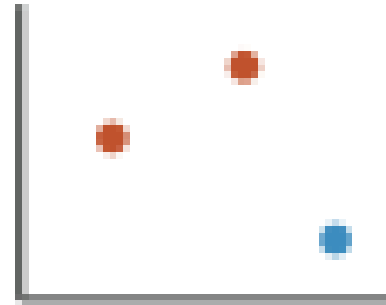
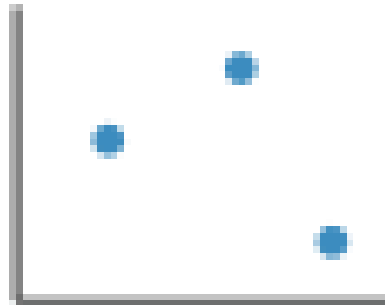
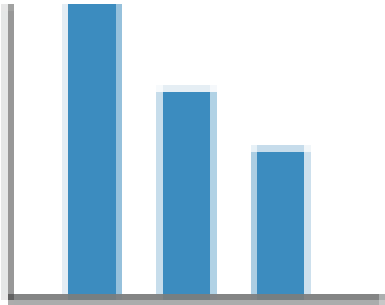
¿Cómo?

Marcadores y canales

Ejemplos



Marcadores y canales: Codificación visual



Marcadores y canales

- **Marcadores:** primitivas geométricas
- **Canales:**
 - ✓ Aspecto de los marcadores
 - ✓ Se puede codificar info redundante con varios canales
- **Consideraciones**
 - Los marcadores de puntos sólo transmiten la posición; sin restricciones de área. Se pueden codificar en tamaño y forma
 - Los marcadores de línea transmiten la posición y la longitud. Sólo se puede codificar por tamaño en 1D (ancho)
 - Los marcadores de área son muy restringidos. No se pueden codificar por tamaño o forma

Marcadores

→ Points



→ Lines



→ Areas



Canales

→ Position

→ Horizontal



→ Vertical



→ Both



→ Color



→ Shape



→ Tilt



→ Size

→ Length

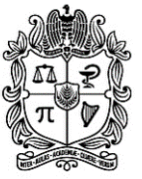


→ Area

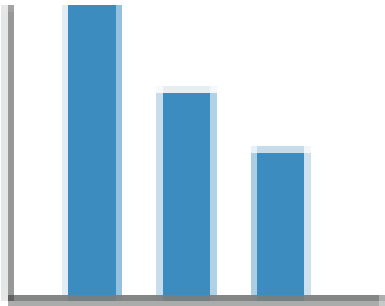


→ Volume



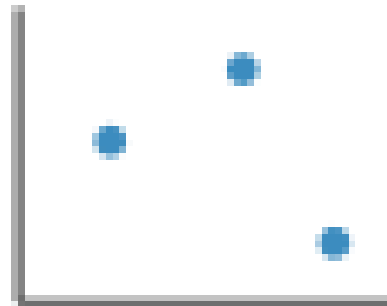


Marcadores y canales: Codificación visual como combinación de marcadores y canales



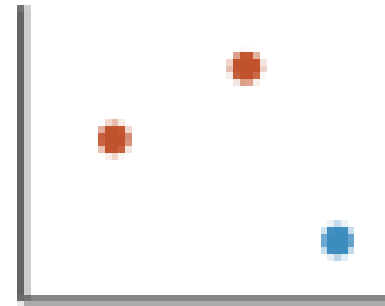
1:
vertical position

mark: line



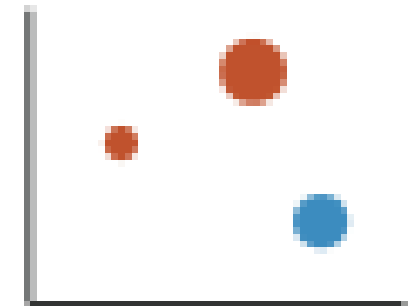
2:
vertical position
horizontal position

mark: point



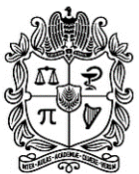
3:
vertical position
horizontal position
color hue

mark: point



4:
vertical position
horizontal position
color hue
size (area)

mark: point

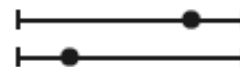


Marcadores y canales:

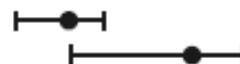
Canales: Tipos de expresividad y rankings de efectividad

➔ 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



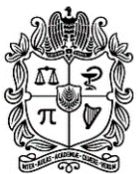
Most

Effectiveness

Least

Same

Same

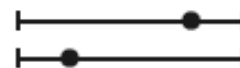


Marcadores y canales:

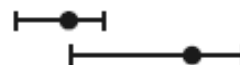
Canales: Tipos de expresividad y rankings de efectividad

➔ Magnitude Channels: Ordered Attributes

Position on common scale



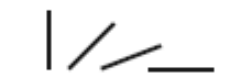
Position on unaligned scale



Length (1D size)



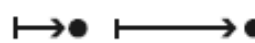
Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



➔ Identity Channels: Categorical Attributes

Spatial region



Color hue



Motion



Shape



Most

Effectiveness

Least

Same

Same

Principio de efectividad

Codificar los atributos más importantes con los canales mejor clasificados

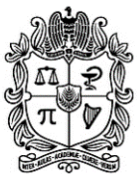
Principio de expresividad

Hacer coincidir las características del canal con las de los datos

Marcadores y canales: Expresividad y efectividad

Factores a tener en cuenta:

- Precisión
- Discriminabilidad
- Posibilidad de separación
- Popout



Marcadores y canales:

Factores a tener en cuenta:

- **Precisión** →
- Discriminabilidad
- Posibilidad de separación
- Popout

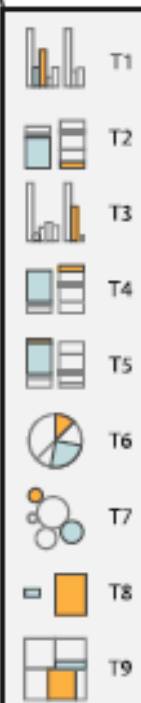
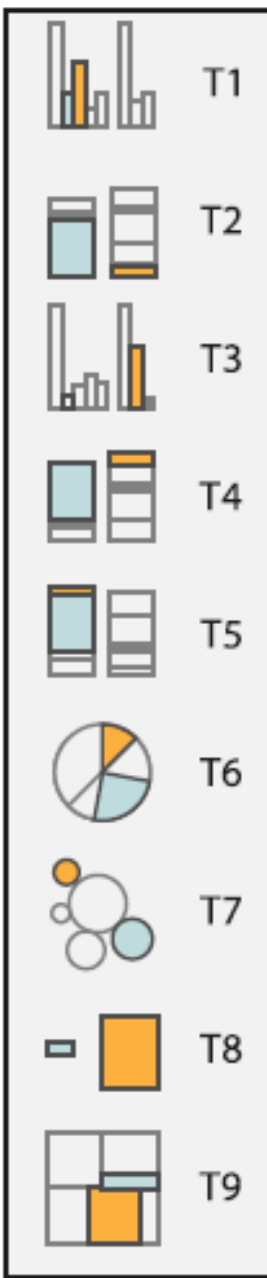
[Cleveland and McGill 84a].
After [Heer and Bostock 10, Figure 4]

Positions

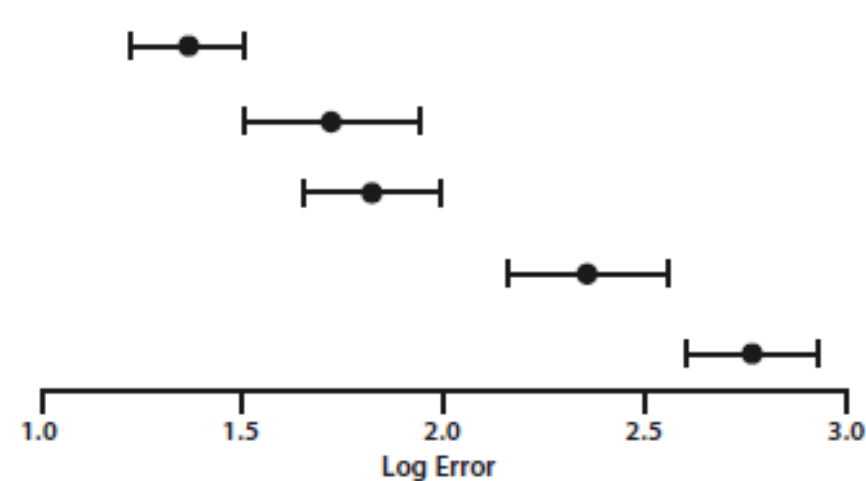
Angles

Circular
areas

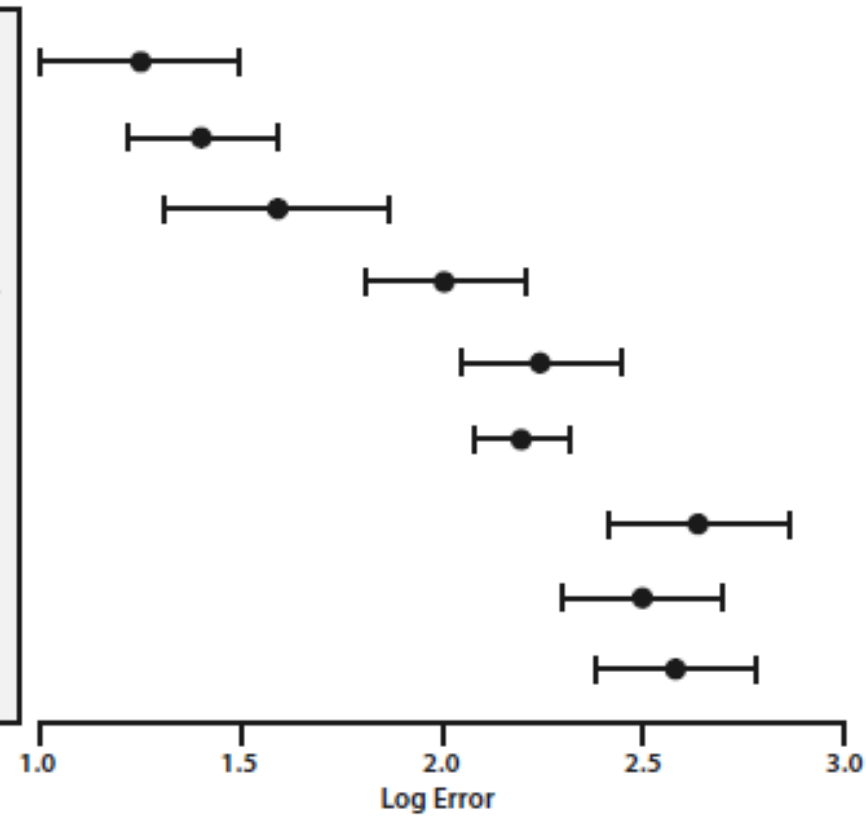
Rectangular
areas
(aligned or in a
treemap)

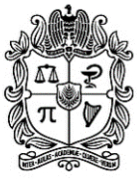


Cleveland & McGill's Results



Crowdsourced Results





Marcadores y canales:

Factores a tener en cuenta:

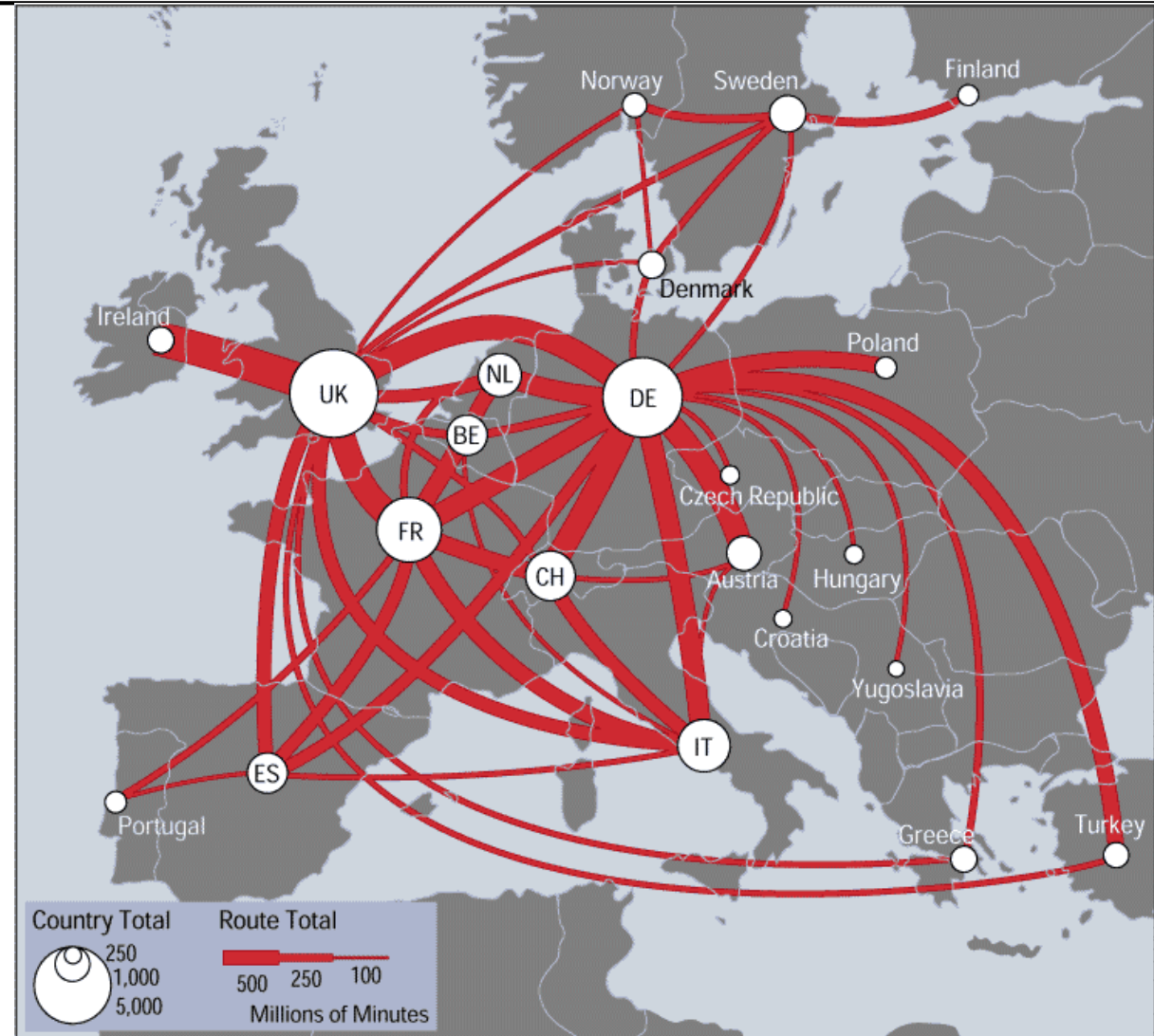
- Precisión
- **Discriminabilidad** →
- Posibilidad de separación
- Popout

¿Cuántos pasos diferenciables?

- Debe ser suficiente para el número de niveles de atributos para mostrar
- Ejemplo: ancho de línea - pocos pasos

[Cleveland and McGill 84a].

After [Heer and Bostock 10, Figure 4]



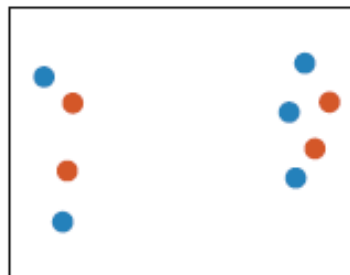
[\[mappa.mundi.net/maps/maps_014/telegeography.html\]](http://mappa.mundi.net/maps/maps_014/telegeography.html)

Marcadores y canales:

Factores a tener en cuenta:

- Precisión
- Discriminabilidad
- **Posibilidad de separación** →
- Popout

Position
+ Hue (Color)



Fully separable

Size
+ Hue (Color)



Some interference

Width
+ Height



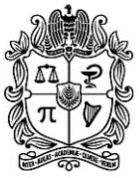
Some/significant
interference

Red
+ Green



Major interference

[Ware 13, Figure 5.23]

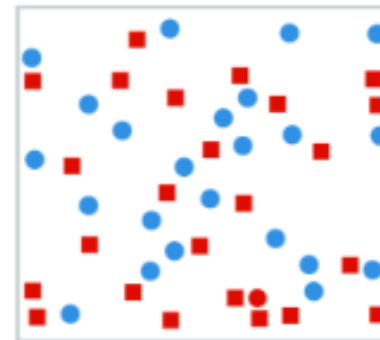
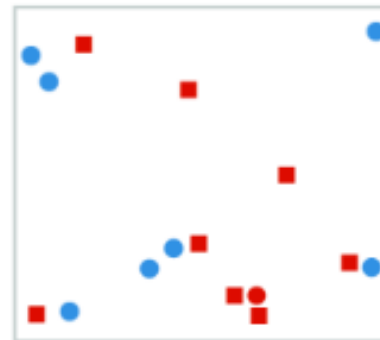
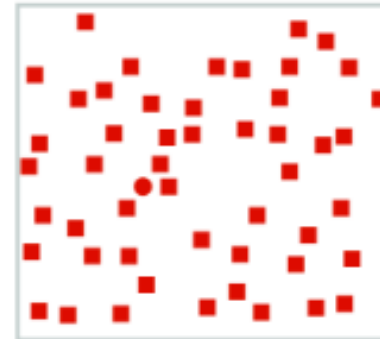
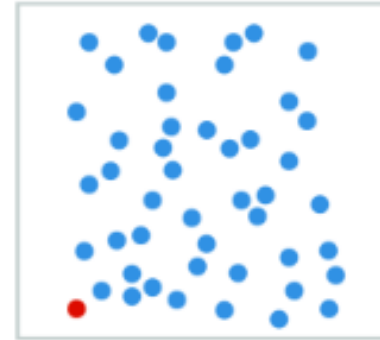
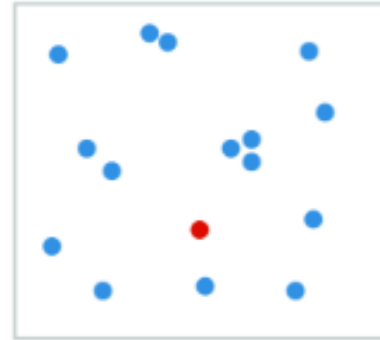


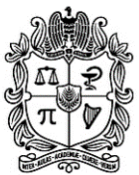
Marcadores y canales:

Factores a tener en cuenta:

- Precisión
- Discriminabilidad
- Posibilidad de separación
- **Popout** →

Encontrar el punto rojo
¿Cuánto tiempo tarda en cada figura?





Agenda

Introducción

Análisis:

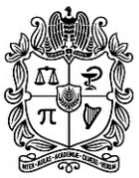
¿Qué?

¿Por qué?

¿Cómo?

Marcadores y canales

Ejemplos



Ejemplos:

<http://johnguerra.co>

John Alexis
Guerra Gómez

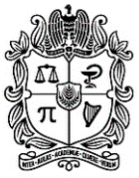
I do **Visual Analytics**, i.e., I include the **user** in the big data analysis/science loop.

john.guerra[~at~]gmail.com

johnguerra.co

PDF Version

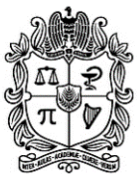
 Follow @duto_guerra



Ejemplos:



<http://johnguerra.co/viz/consultaAnticorrupcion2018/>



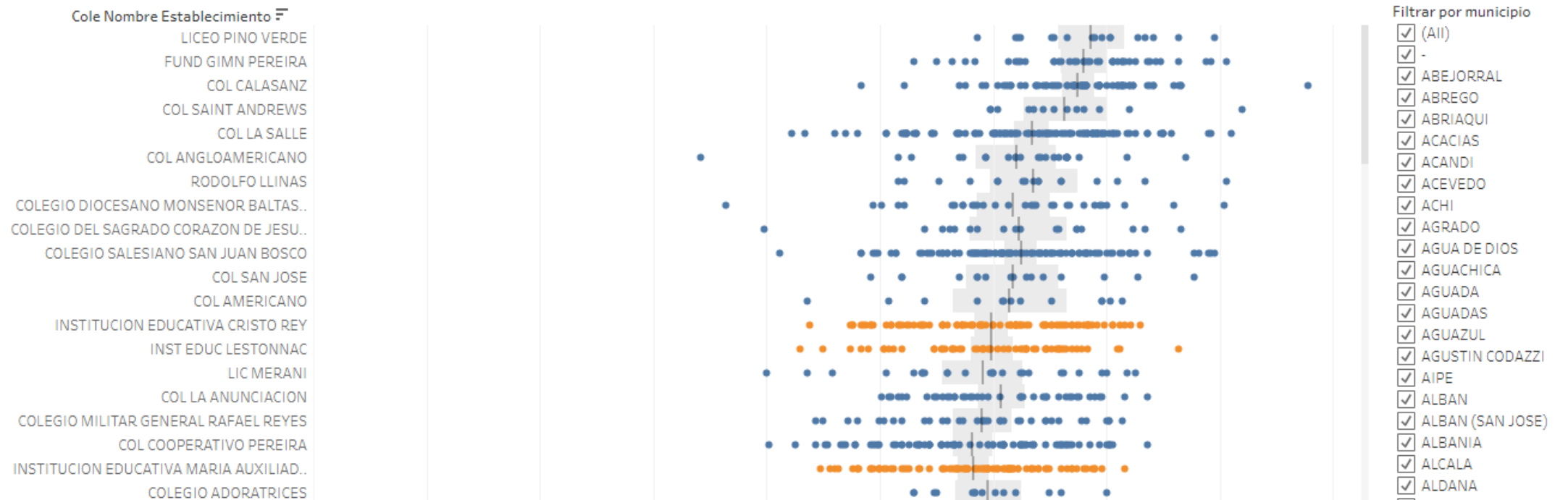
Ejemplos:

Ranking de los mejores colegios

Según los resultados de las pruebas Saber 11 2017-2. Ordenados por la mediana del puntaje general obtenido, y filtrando colegios que hayan presentado al menos 10 estudiantes

Por John Alexis Guerra Gómez (<http://johnguerra.co>)

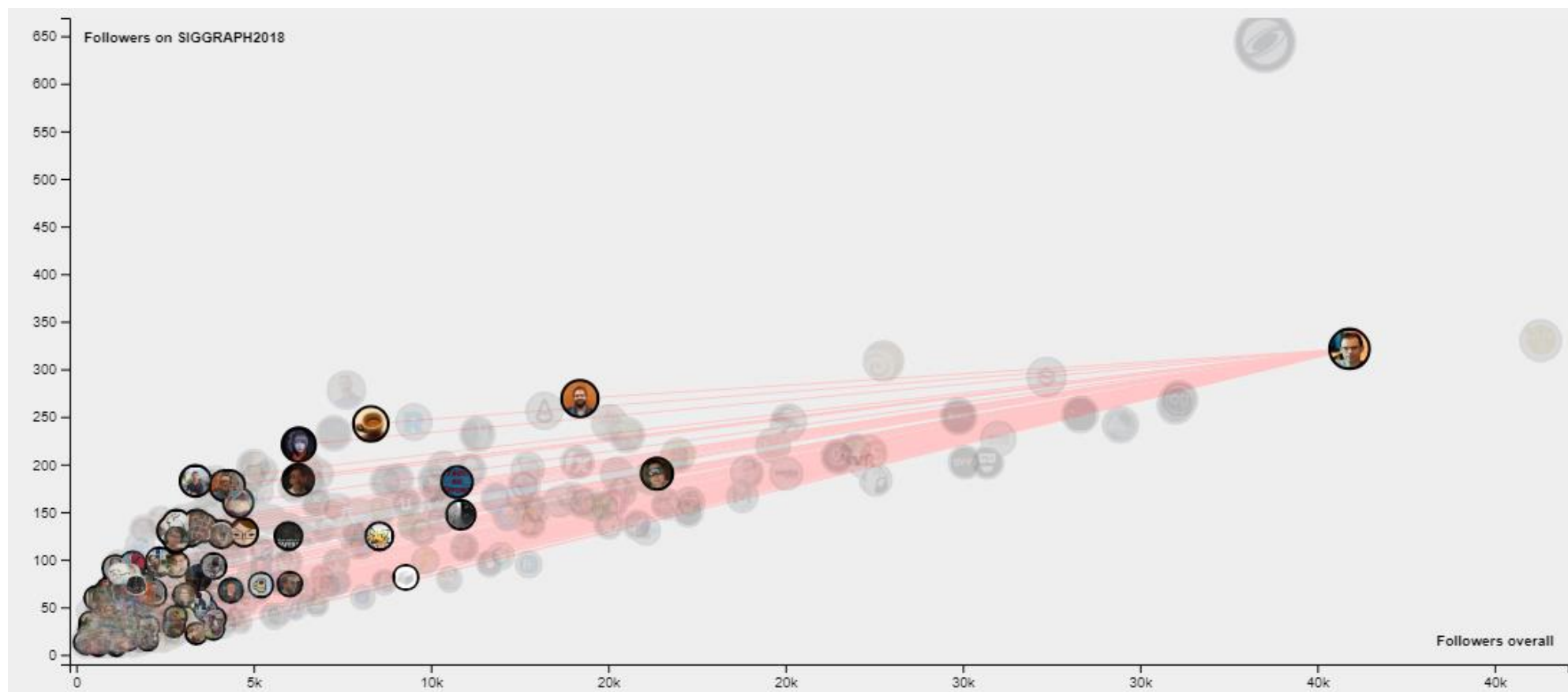
Datos de <https://www.datos.gov.co/Educaci-n/Saber-11-2017-2/s6qh-49yh>



<http://johnguerra.co/viz/saber11/>

Ejemplos:

Who to follow in SIGGRAPH2018



<http://johnguerra.co/viz/influentials/SIGGRAPH2018/>



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Introducción al Análisis y Visualización de Datos con Python

Principios de visualización de información

Felipe Restrepo Calle

ferestrepoca@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial

Facultad de Ingeniería

Universidad Nacional de Colombia

Sede Bogotá