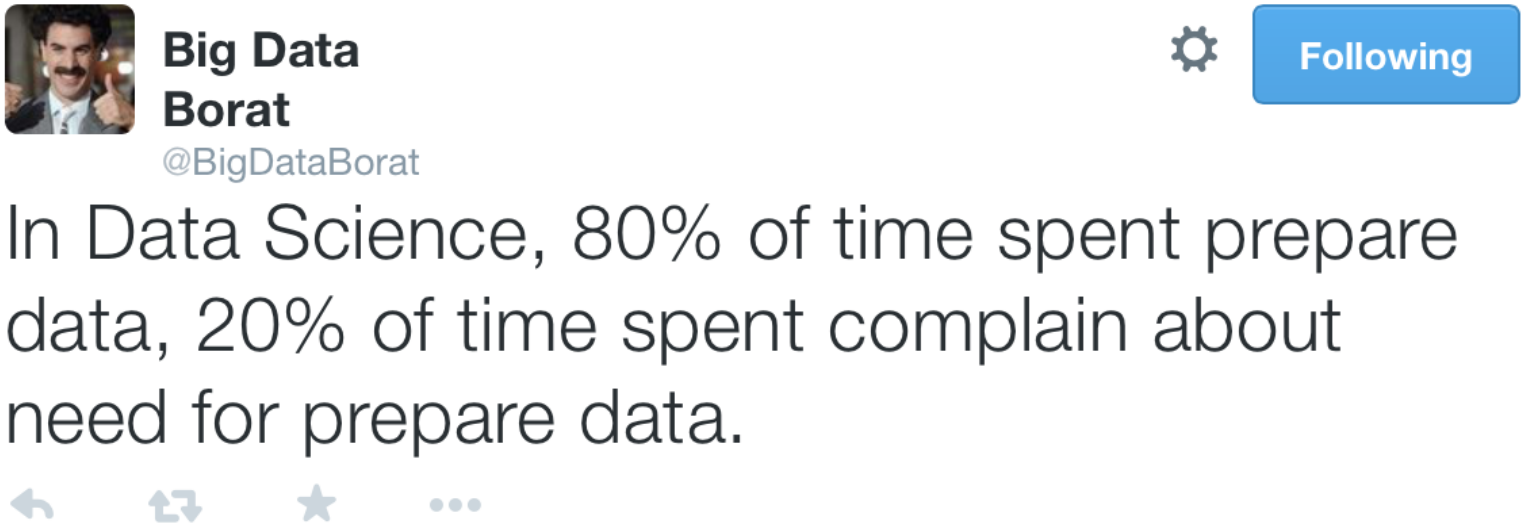
Análisis Exploratorio de Datos

Eduardo Fernández

Manipulación de Datos

- I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all [Kandel et al. 2012]



Primeras Visualizaciones de Datos

[¿Demasiado antiguo para ser un mapa?]

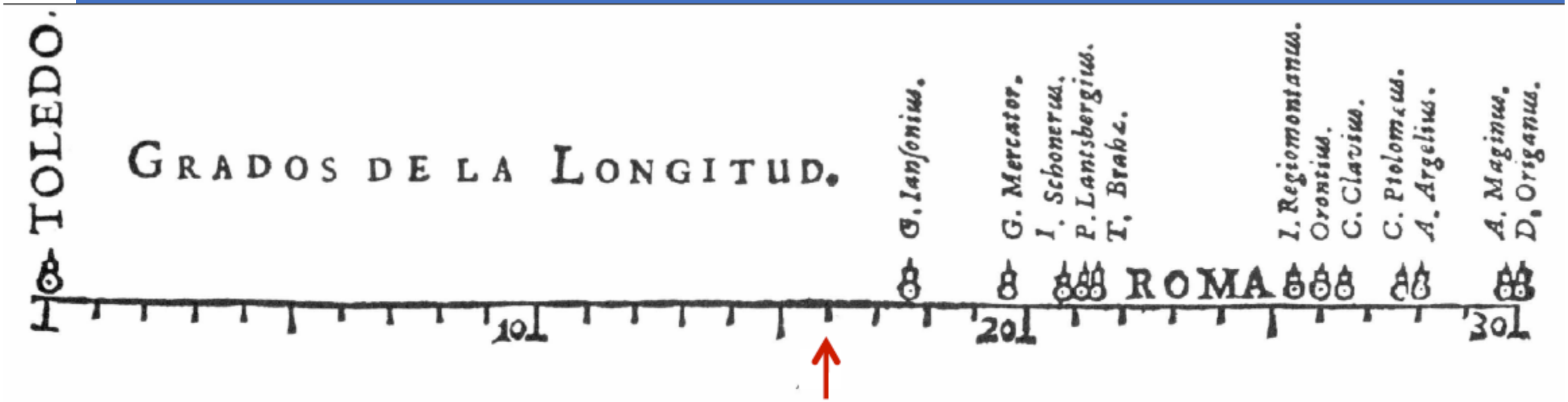
Primeras Visualizaciones de Datos

El mapa del mundo de Ptolomeo

Reconstruido a partir de la **Geografía de Ptolomeo** (150 DC) en el siglo XV, indica

- "Sinae" (China) en el extremo derecho,
- más allá de la isla de "Taprobane" (Ceilán o Sri Lanka, sobredimensionada)
- y la "Aurea Chersonesus" (península del sudeste asiático).

Primeras Visualizaciones de Datos



Distancia real en grados
de longitud

Michael Florent Van Langren (1644)

12 estimaciones conocidas de la distancia de longitud entre Roma y Toledo.

Podría haber realizado una tabla, pero prefirió una representación visual.

Primeras Visualizaciones de Datos

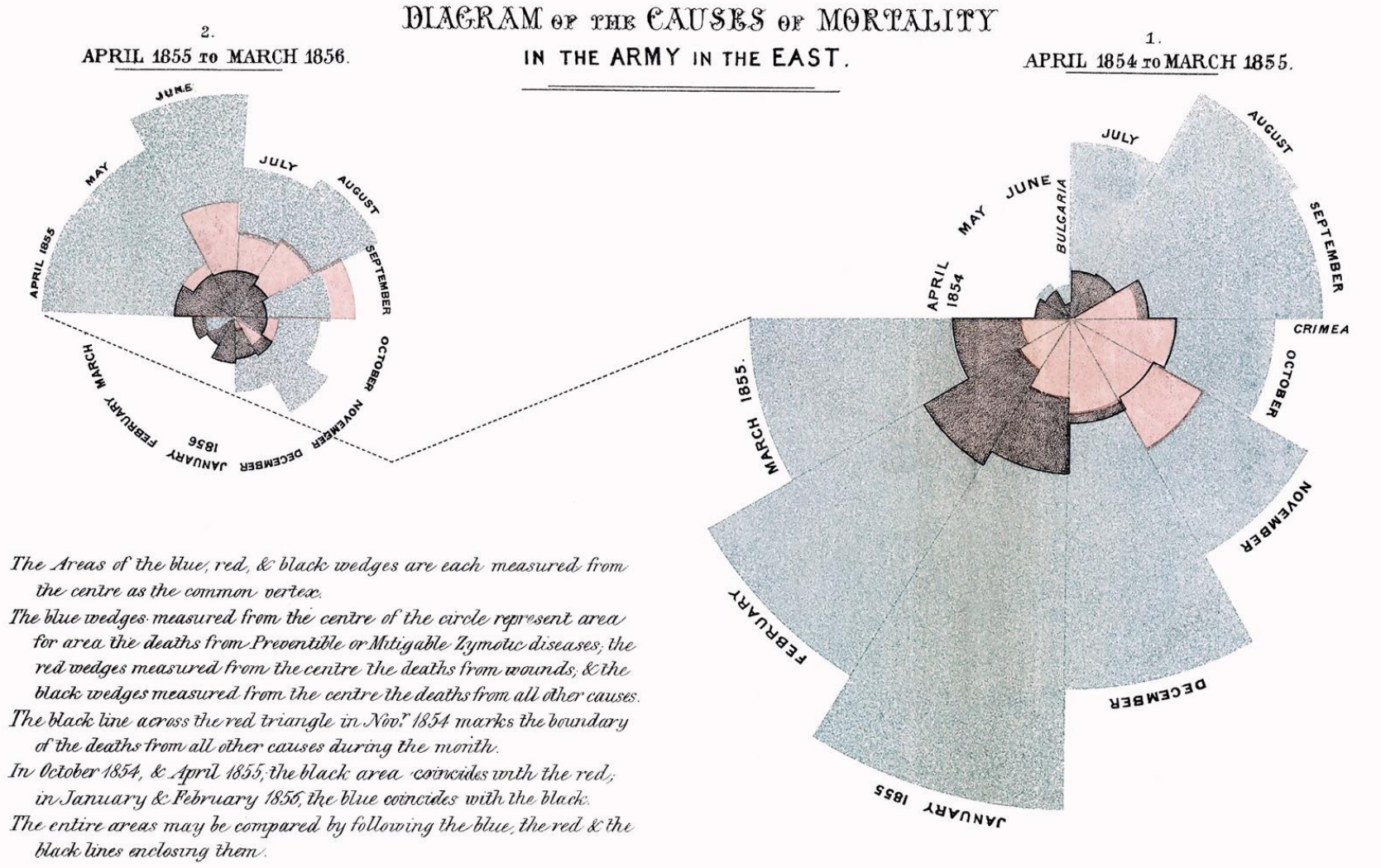


Gráfico de área polar de Florence Nightingale. Muertes en la Guerra de Crimea (1853-1856):

- **Azul:** Muertes por enfermedades prevenibles, por malas condiciones sanitarias.
- **Rojo:** Muertes por heridas de guerra, como las sufridas en combate.
- **Negro:** Todas las demás causas de muerte.

Manipulación de Datos

- A menudo es necesario manipular los datos antes de realizar un análisis/visualización propiamente dicha. Estas tareas incluyen reformatear, limpiar, evaluación de la calidad e integración de datos

- La manipulación se puede realizar:
  - con planillas de cálculo (MS Excel, LibreOffice, etc.)
  - Code: arquero (JS), dplyr (R), pandas (Python).
  - Open Refine http://openrefine.org/

Datos ordenados

- Cómo las filas, columnas y tablas se corresponden con las observaciones, las variables y los tipos?
- Según [Wickham 2014] los datos están ordenados si:

1. Cada variable forma una columna
2. Cada observación forma una fila
3. Cada tipo de unidad observacional forma una Tabla.

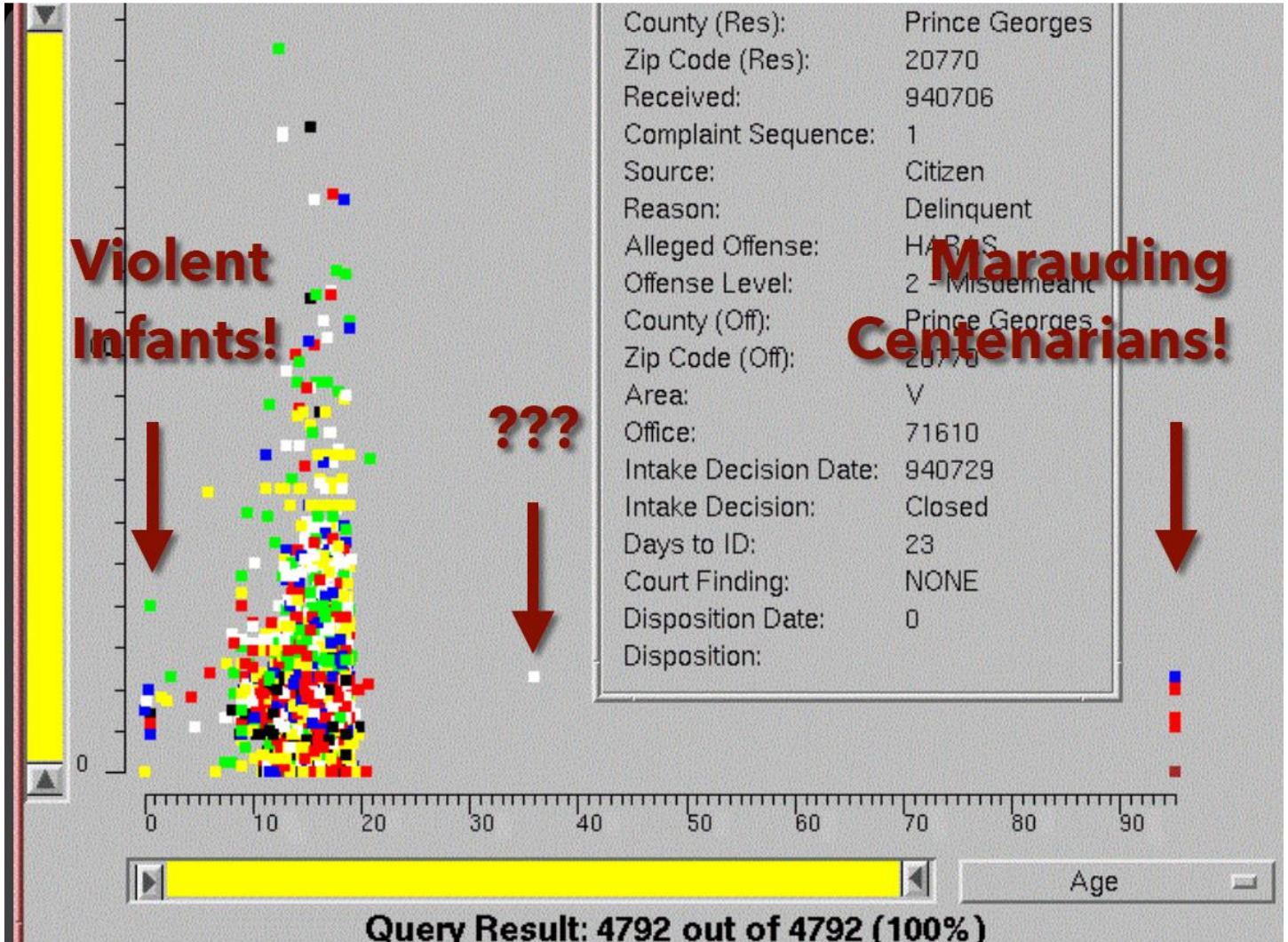
Esto provee un punto de partida para el análisis, la transformación y la visualización de los datos.

Datos ordenados

- “**La primera señal de que una visualización es buena ocurre cuando muestra un problema en sus datos.** Cada visualización exitosa en la que he estado involucrado ha tenido esta etapa en la que te das cuenta que, "¡Dios mío, estos datos no son lo que pensé que serían!" Así que ya has descubierto algo ".

  Martín Wattenberg

Problemas en los datos?



- Edades erroneas de delincuentes?

Obstáculos en la calidad de los datos

- Datos faltantes.
  - Faltan medidas, los datos están redactados sin ninguna estructura.
- Valores erróneos.
  - ¿Mal escritos o datos atípicos pero verdaderos?
- Conversión de tipo.
  - p. ej., pasar de código postal a lat-lon
- Resolver la entidad.
  - El mismo dato con diferentes valores/nombres.
- Integración de datos.
  - Esfuerzo/errores surgidos al combinar datos

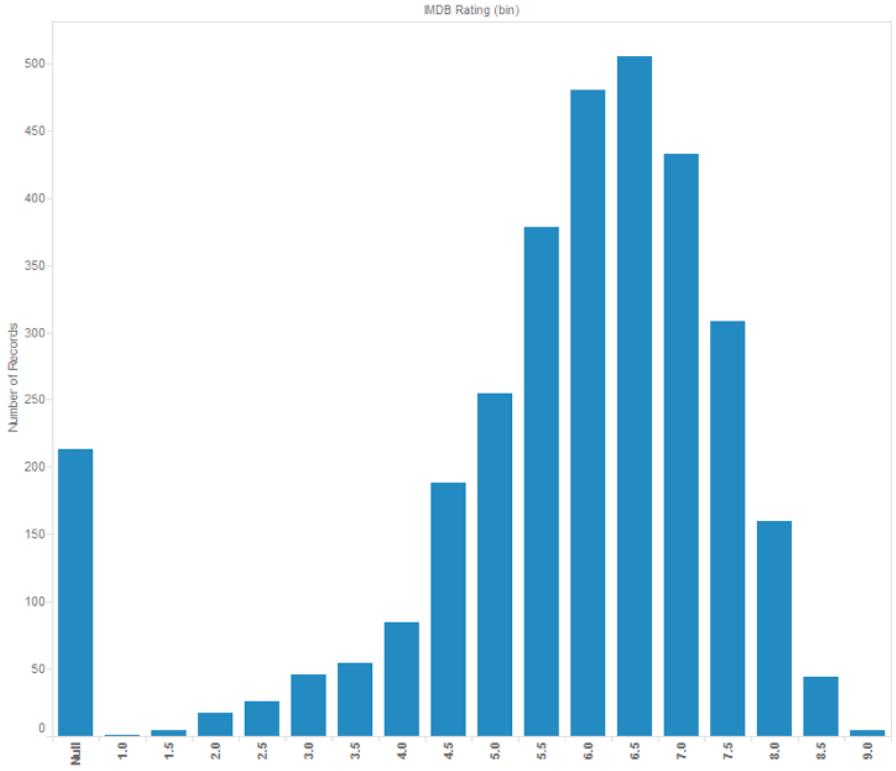
LECCIÓN: Anticípese a problemas con sus datos.

Ejemplo: Datos de películas

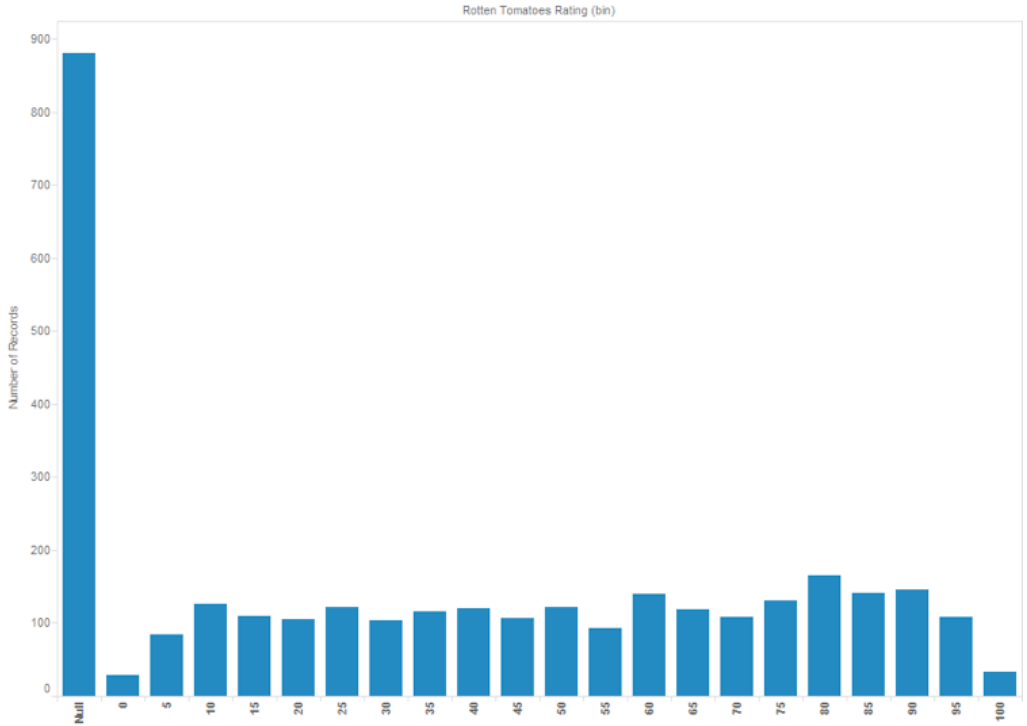
- Title — Cadena de caracteres (N)
- IMDB Rating — Número (C)
- Rotten Tomatoes Rating — Número (C)
- MPAA Rating — Cadena de caracteres (O) **{G, PG, PG-13, R, NC-17}**
- Release Date — Fecha (C)

Ratings

IMDB Rating

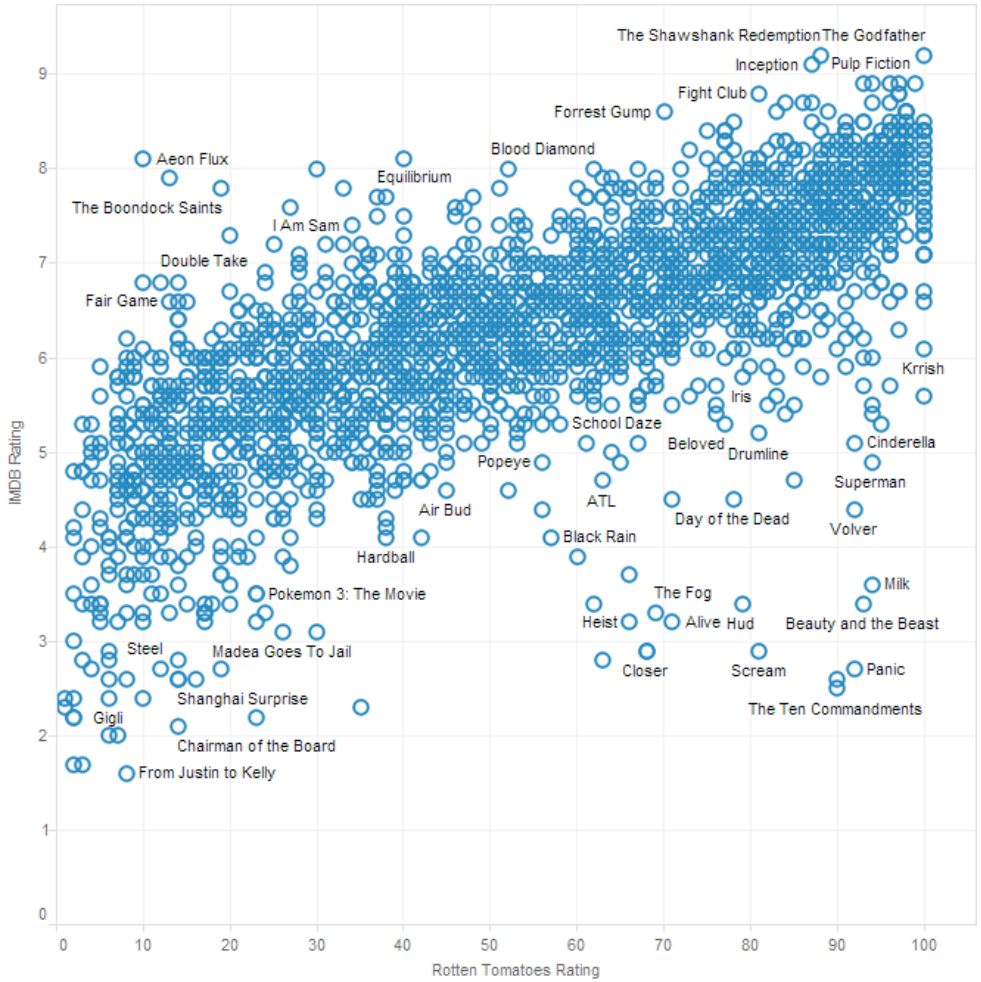


Rotten Tomatoes Rating

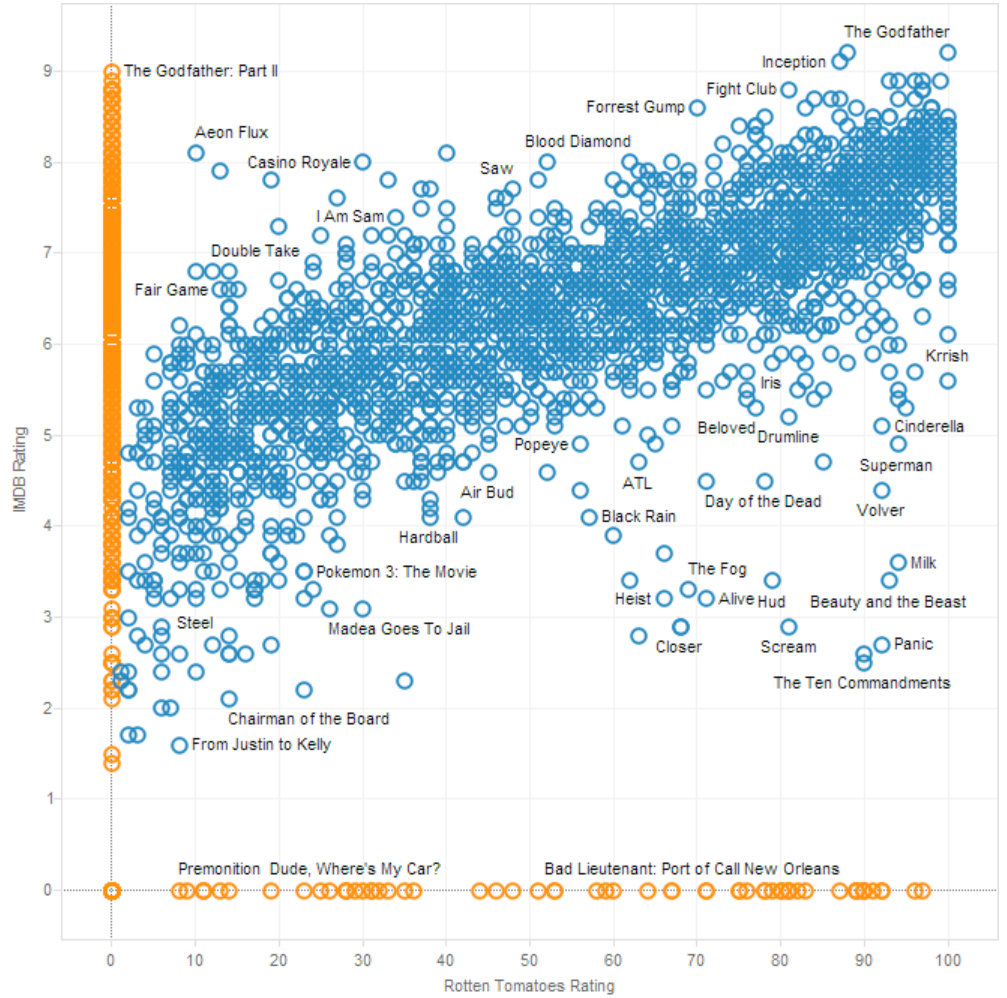


Comparación de Ratings

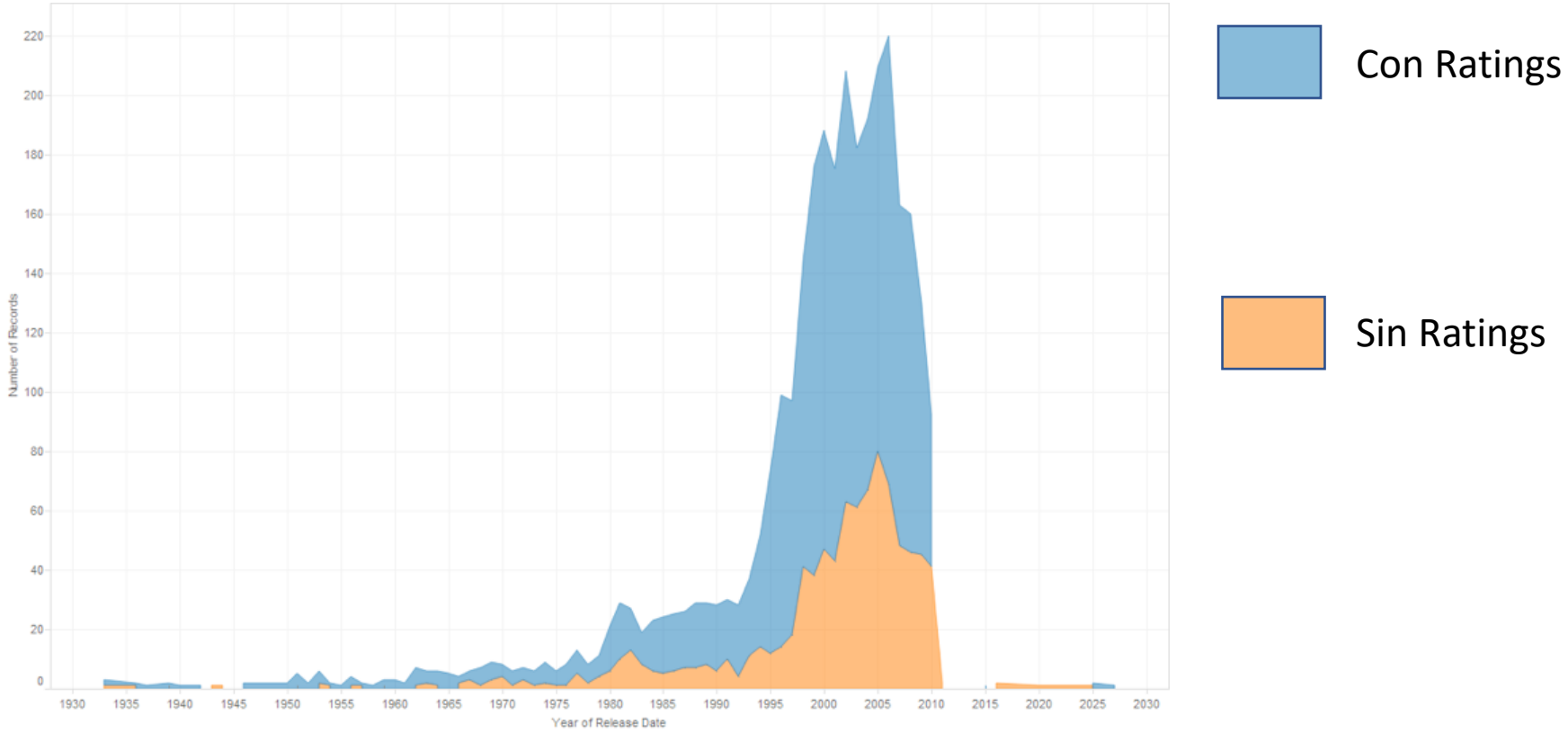
Sin considerar los “sin datos”



Considerando los “sin datos”



Distribución en el Tiempo de las observaciones



Lección: Ejercitar el Escepticismo

- Verifique la calidad de los datos y sus suposiciones.
- Comience con resúmenes de una variable por vez y luego comience a considerar las relaciones entre las variables.

- ¡Evitar conclusiones prematuramente!
  - Realizar múltiples análisis de las variables, diseños de los gráficos.
  - Seleccionar aquellas mejores, que muestren más datos interesantes, que expresen mejor los aspectos importantes que han descubierto de los datos.
  - Luego recién vaya a un diseño y visualización más detallado.

Ejemplo: Datos de efectividad de antibióticos

• Género de bacterias.	Cadena de caracteres (N)
• Especies de bacterias.	Cadena de caracteres (N)
• Aplicación de antibiótico.	Cadena de caracteres (N)
• Tinción de Gram ?.	Positivo / Negativo (N)
• Min. Concentrado inhibidor. (g)	Número (C)

Tabla recopilada antes de 1951.

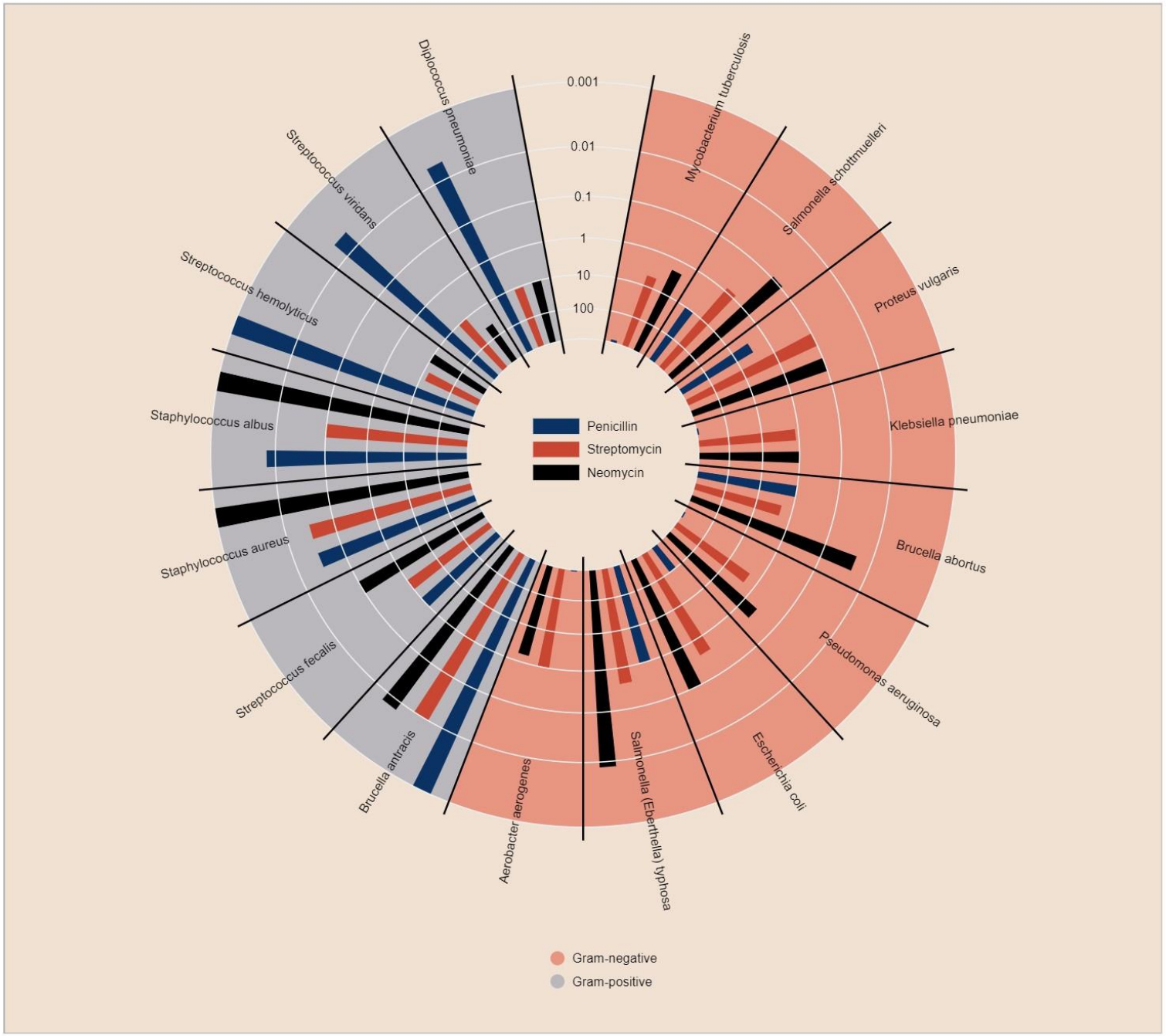
las bacterias Gram positivas y Gram negativas a menudo difieren en sus respuestas a los antibióticos, por lo que conocer la clasificación de Gram puede ayudar a guiar el tratamiento de las infecciones bacterianas.

¿Qué preguntas podríamos hacer?

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

Burtin's Antibiotics



¿Cómo se comparan las drogas?

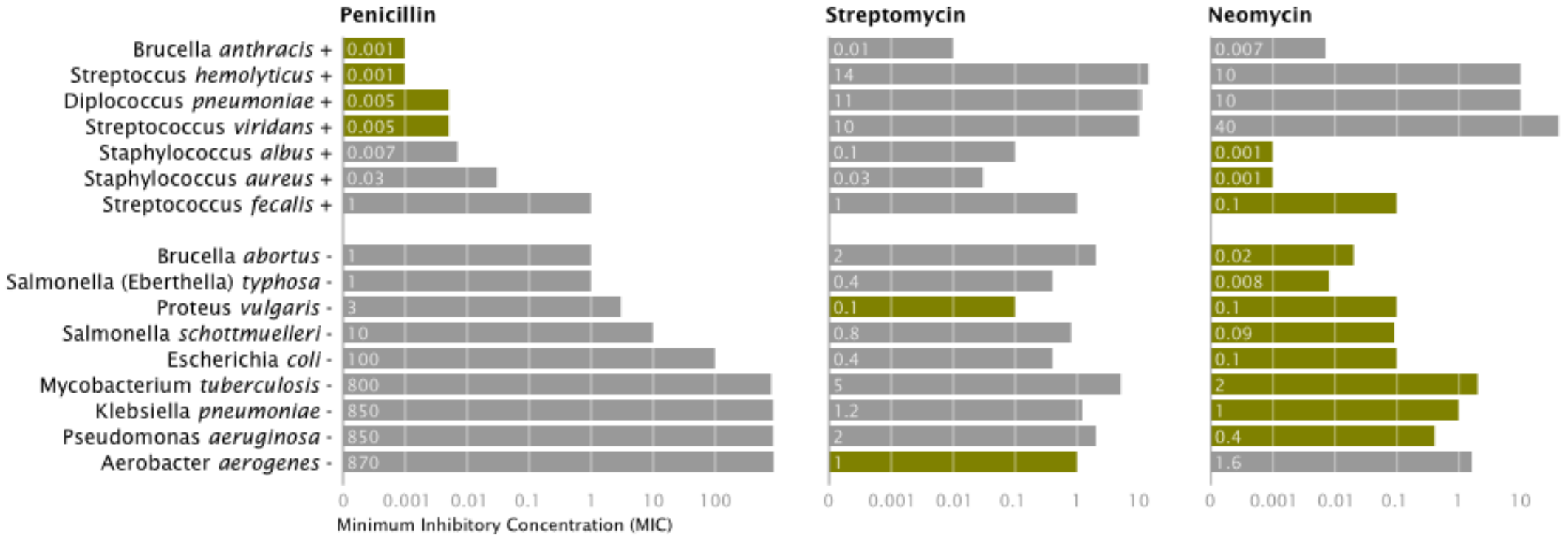
Diagrama original de Will Burtin 1951

Radius: 1 / log(Min.Conc.Inhibidor)

Bar Color: Antibiótico

Color de Fondo: Gram Staining

¿Cómo se comparan las drogas?

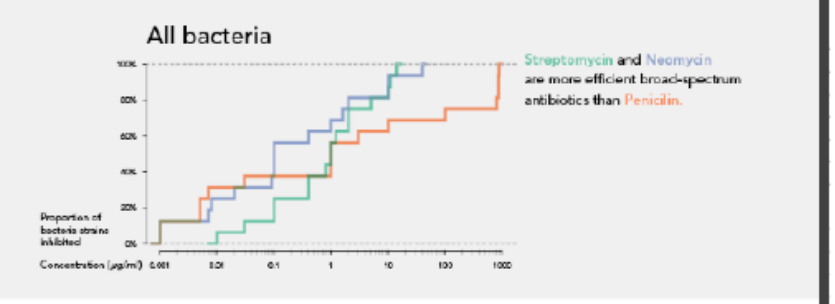


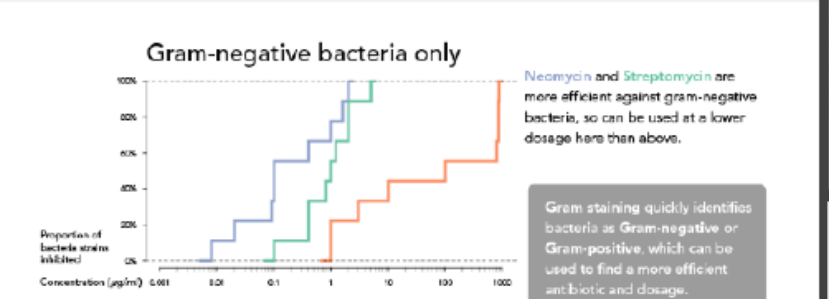
Eje X: Antibióticos, log(MIC)

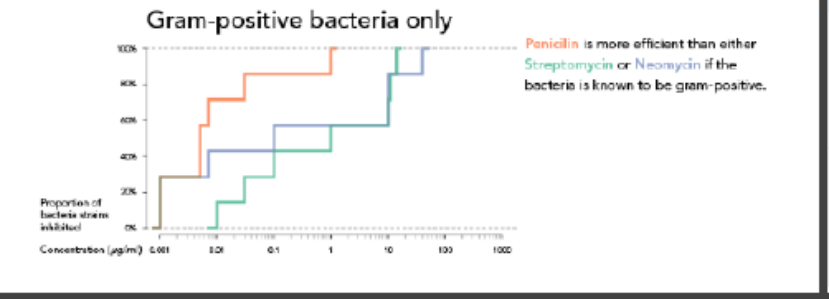
Eje Y: Gram-Staining | Especies

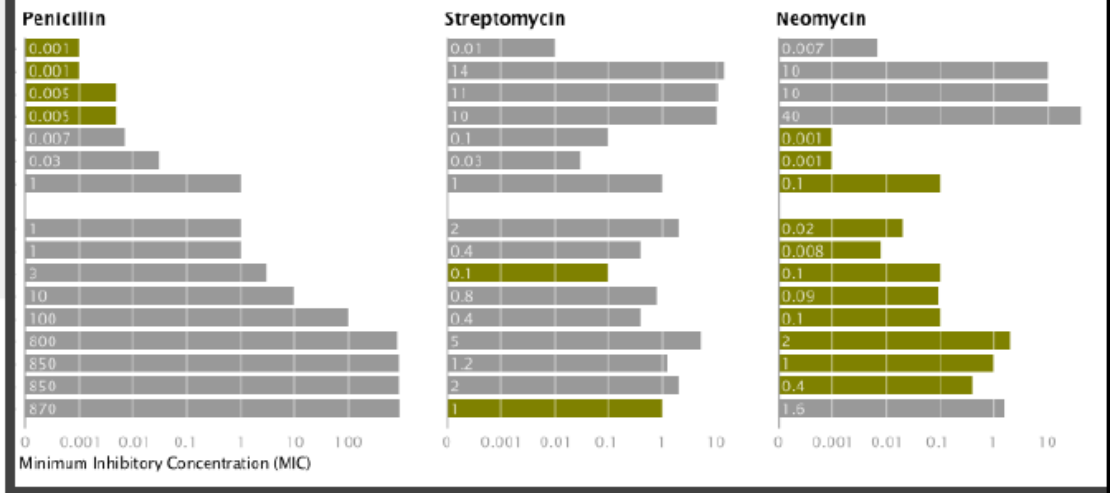
Color: ¿La más efectiva?

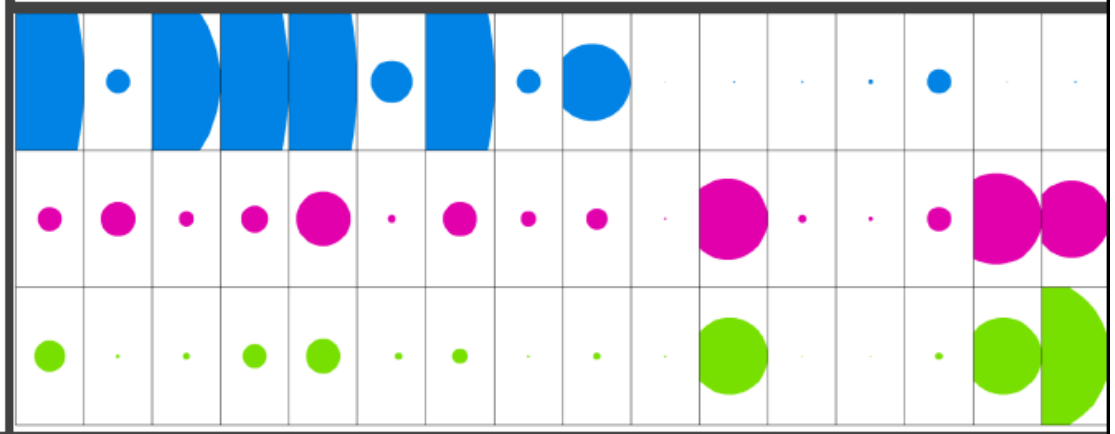
Mike Bostock
Stanford CS448B,
Winter 2009

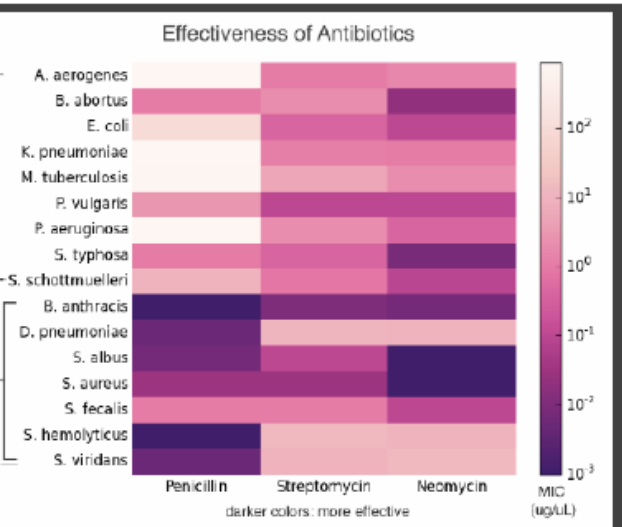


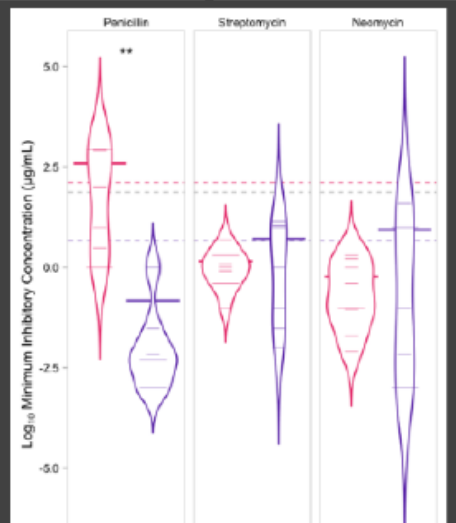


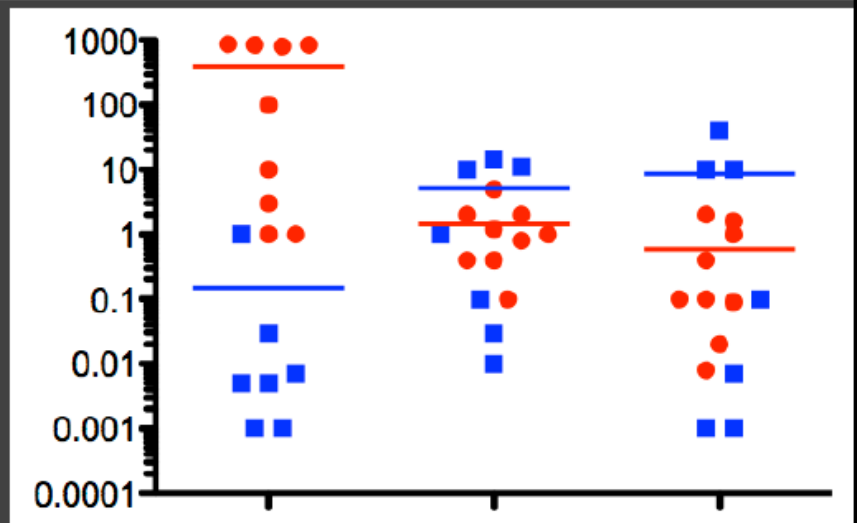






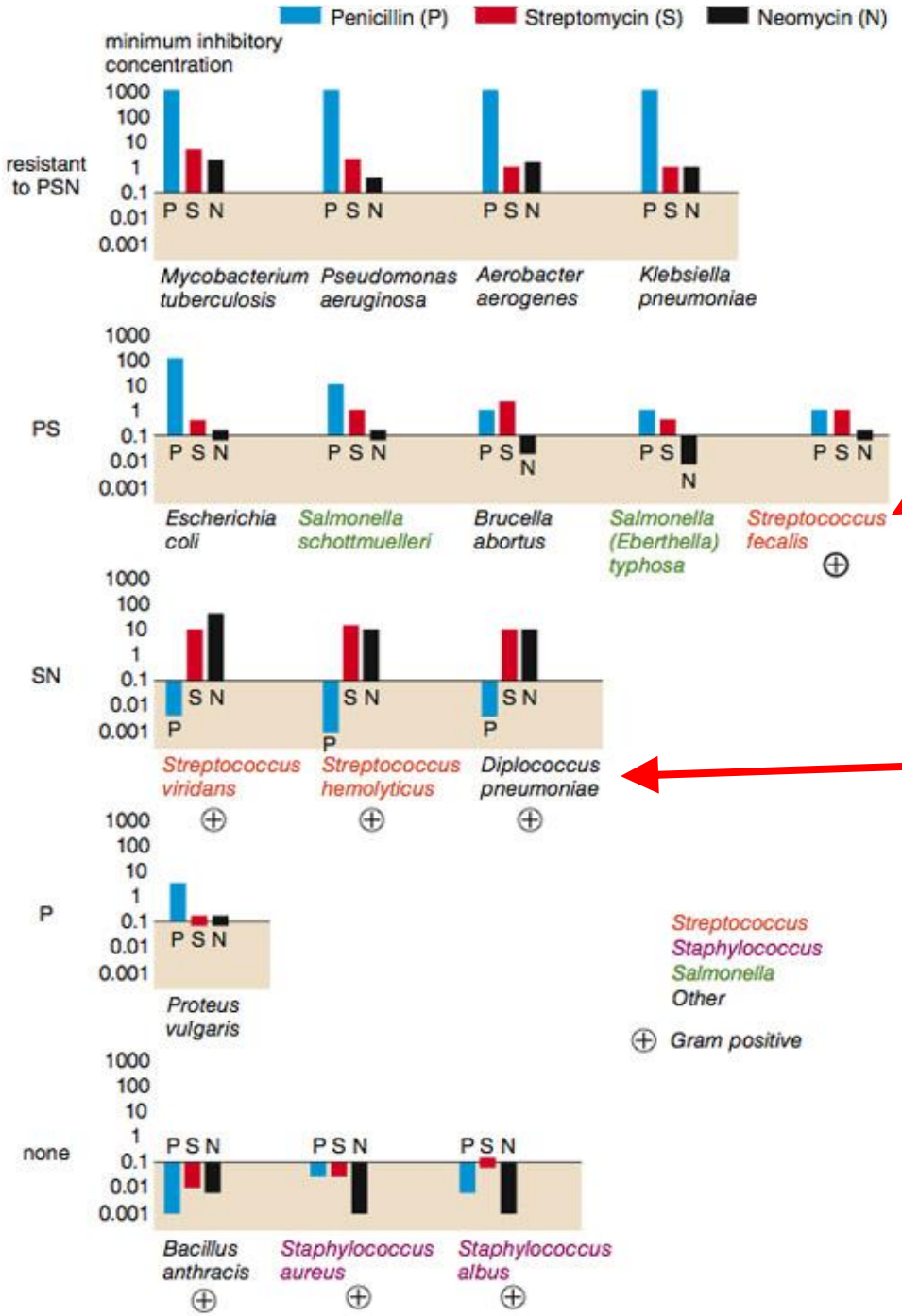






Ejemplos de gráficos a la pregunta:

¿Qué droga debería usar?

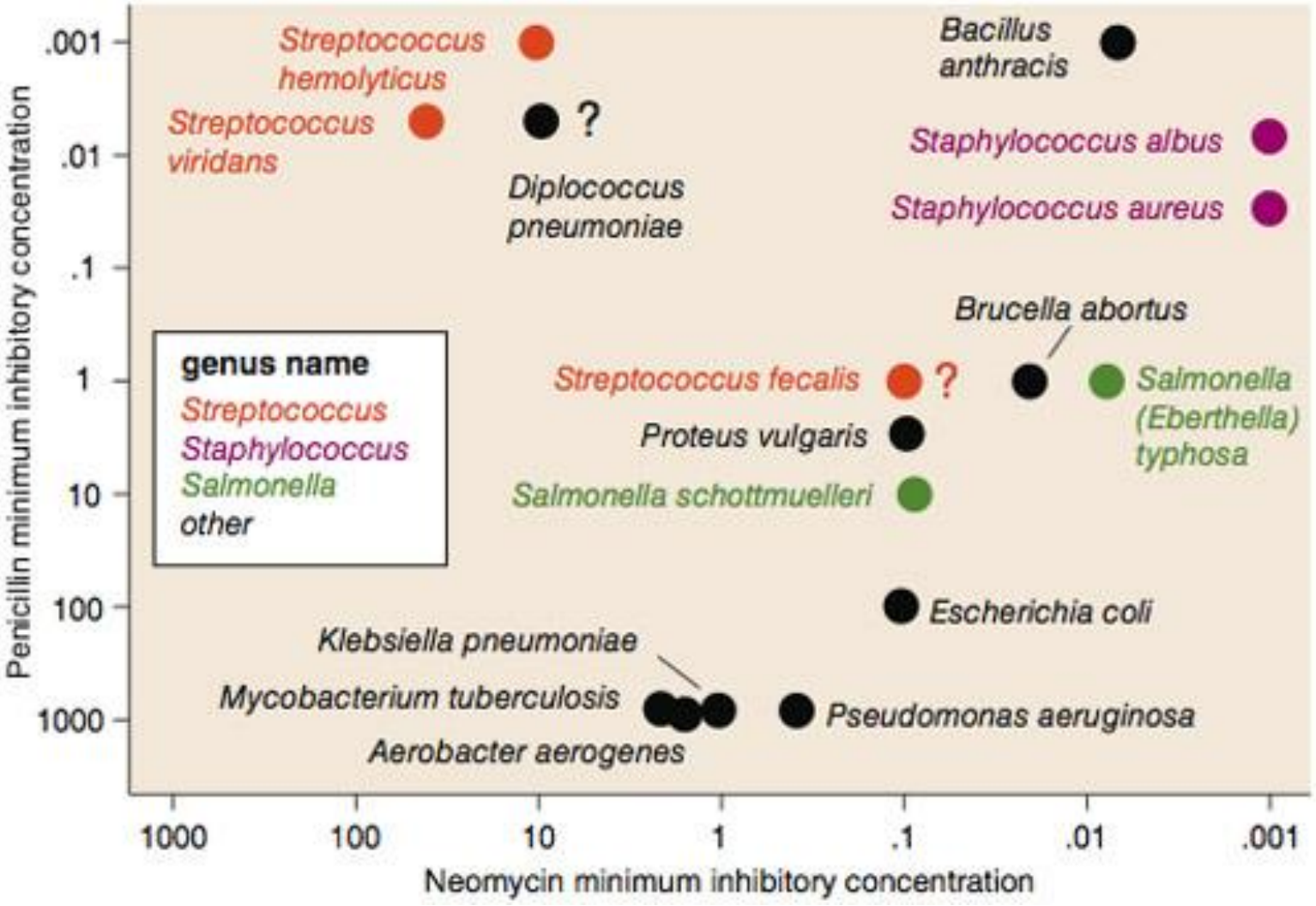


¿Las bacterias se pueden agrupar por resistencia?

¡No es un estreptococo! (se dieron cuenta 30 años más tarde)

¡Es un estreptococo! (se dieron cuenta 20 años más tarde)

¿Las bacterias se pueden agrupar por resistencia?



¿Existe alguna correlación entre las resistencias a las drogas?

Lección: Exploración iterativa

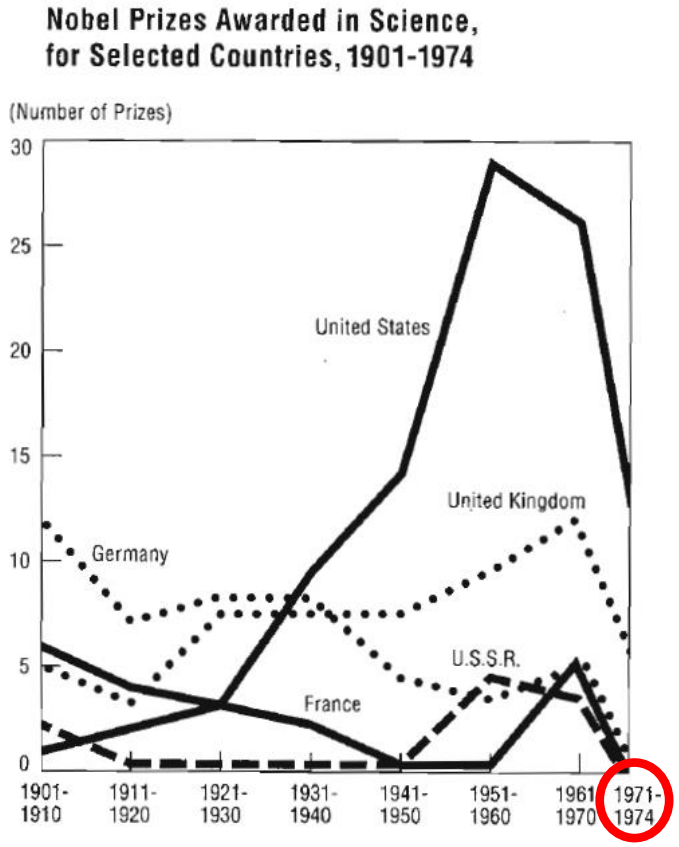
- Proceso exploratorio

1. Construya gráficos para abordar preguntas
2. Inspeccione la "respuesta" y evalúe las nuevas preguntas
3. Repita ...Transformar los datos de forma adecuada (por ejemplo, invertir, logaritmo, etc.)

Mostrar variación de datos, no variación de diseño [Tufte]

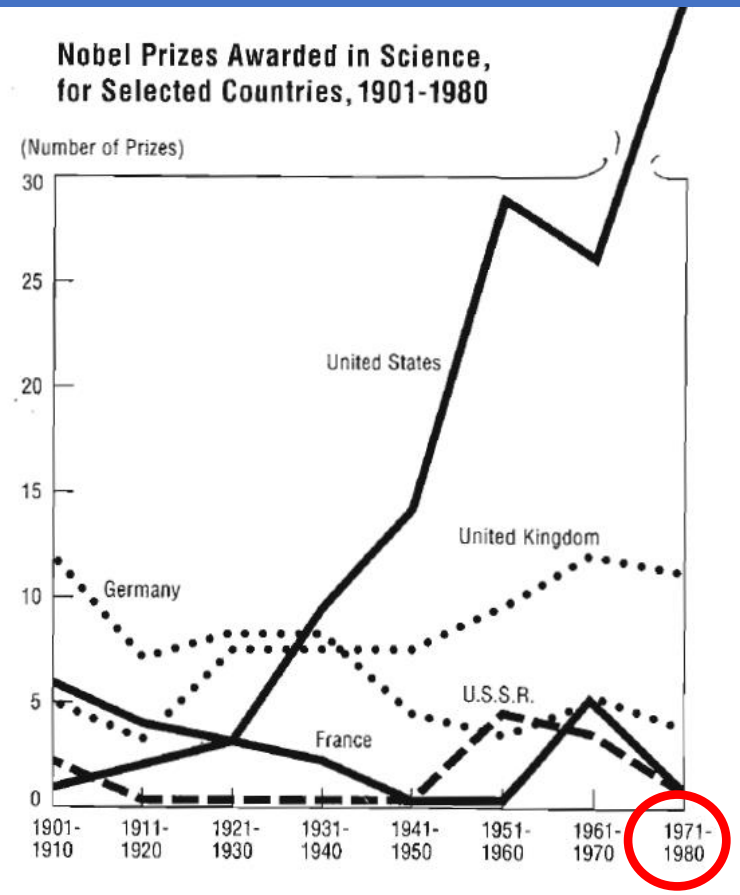
Mostrar variación de datos, no variación de diseño

Todos los datos están mostrados cada 10 años, salvo los datos de más a la derecha que corresponden a 4 años.



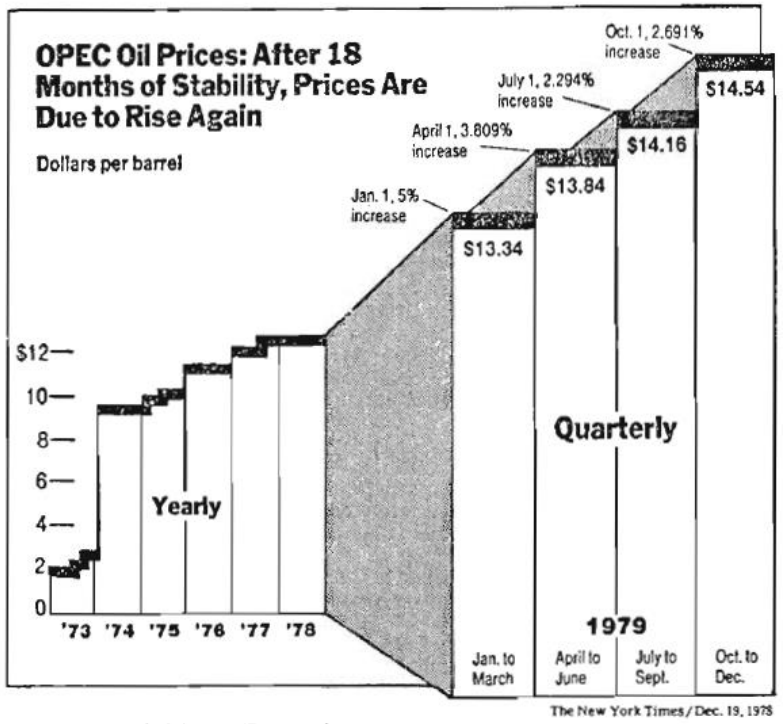
Lo publicado en 1974

Science Indicators, [NSF 1974]



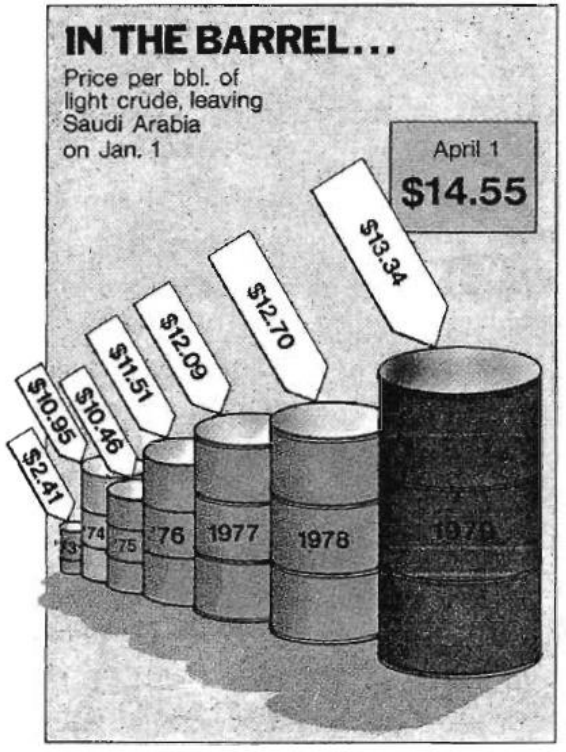
La corrección posterior con los datos correctos

Mostrar variación de datos, no variación de diseño



New York Times, December 19, 1978, p. D–7.

Los incrementos entre 1973 y 1978 están a una escala muy diferente que los incrementos en 1979.

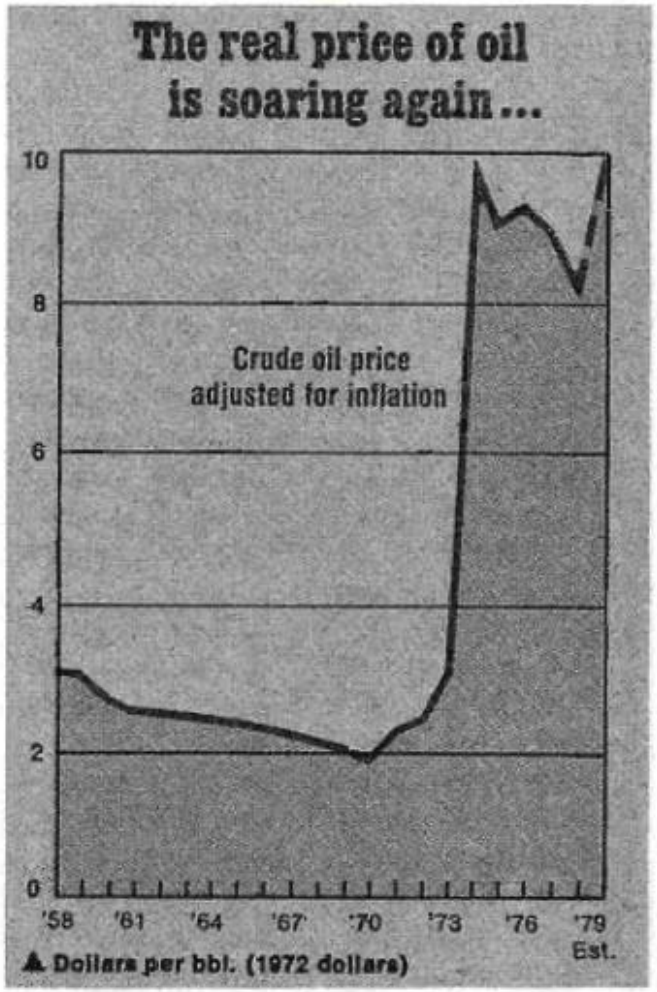


Time, April 9, 1979, p. 57.

Los incrementos están en relación con la altura del barril, pero el barril tiene volumen, entonces el incremento es al cubo. Además, la perspectiva es irreal (los barriles más alejados son demasiado pequeños).

Mostrar variación de datos, no variación de diseño

Business Week, April 9, 1979, p. 99.



En realidad, considerando la inflación, el precio del petróleo había bajado en los años anteriores a 1979.

Preocupados por hacer gráficos interesantes, se habían olvidado de mostrar la información verdadera.