Lecture 8: EDA

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

Lecture Outline

Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

Visualization

- Exploration (EDA)
- Communication

Lecture Outline

Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

Visualization

- Exploration (EDA)
- Communication

Example

Let's say that we are interested in the English Premier League (football/soccer) and want to build a model to predict a player's market value.

Question Does age affect one's market value?

Example

What do we do?

Example

What do we do?

- Ask an interesting question
- **Get the Data**
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

Dataset Considerations

- What data is necessary to answer our question?
- Is the source credible/authoritative? (.com, .net, .org, .gov, .name)
- How difficult is it to analyze the dataset? (photos, videos, text?)
- What is the allowed usage of data under its license?
- Who collected the data?
- When was the data collected?

Dataset Considerations (continued)

- How was the data collected?
- How is the data formatted?
- Confidentiality concerns
- Does your data collection procedures need to be approved by an IRB?
- Comprehensive data vs sampled data?
- Biases

Dataset Considerations (continued)

- How was the data collected?
- How is the data formatted?
- Confidentiality concerns
- Does your data collection procedures need to be approved by an IRB?
- Comprehensive data vs sampled data?
- Biases

Lecture Outline

Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

Visualization

- Exploration (EDA)
- Communication

Dataset Considerations: Comprehensive Data

- We have access to all the data points that exist, which is usually a lot
- Collected and digitized as part of generalized procedures of an institution

13 million articles

~500 million tweets per day

100,000s votes per year

Dataset Considerations: Sampled Data

- When collecting individual data is relatively expensive
- Only a portion of the population is sampled
- Not just restricted to polling or surveys

GALLUP

IMDb

1. Clover Food Lab

821 reviews

$$ · American (New), Sandwiches, Cafes

THE Q EVALUATIONS

nielsen

Lecture Outline

Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

Visualization

- Exploration (EDA)
- Communication

Dataset Considerations: Biases

- **A bias in sampled data occurs when a procedure causes the sample to overrepresent a subpopulation**
- Biases may not necessarily be intentional
- Even if you don't think over-representation of a subpopulation will bias the dataset with regard to your question, it's still a bias
- Always strive to minimize any biases in your data collection procedures

Dataset Considerations: Biases

Gallup Polls

- Randomly calls two groups of ~500 people a day by sampling among all possible phone numbers
- For landlines, asks for household member who has the next birthday
- Calls people living in all 50 states
- Tries to assure 70% cellphone, 30% landlines
- Weights data to reflect the demographics of the general population

Dataset Considerations: Biases

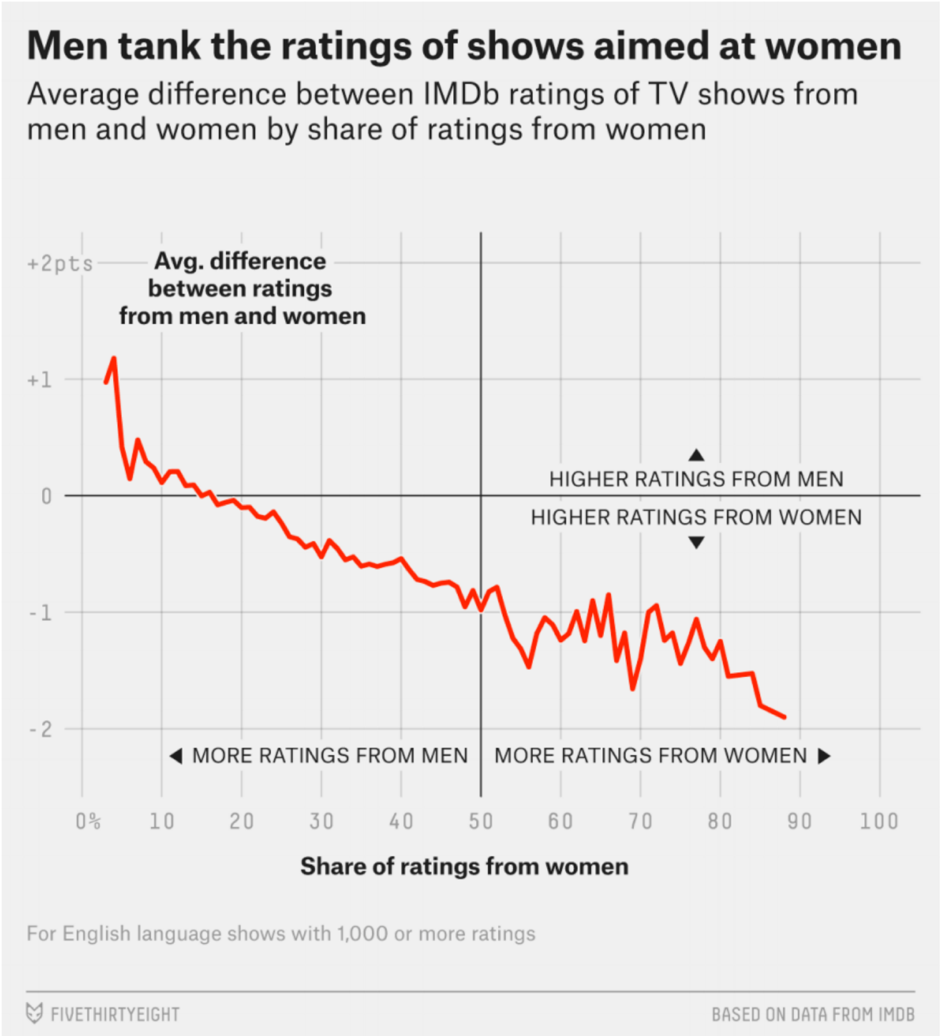
IMDb Movie Ratings

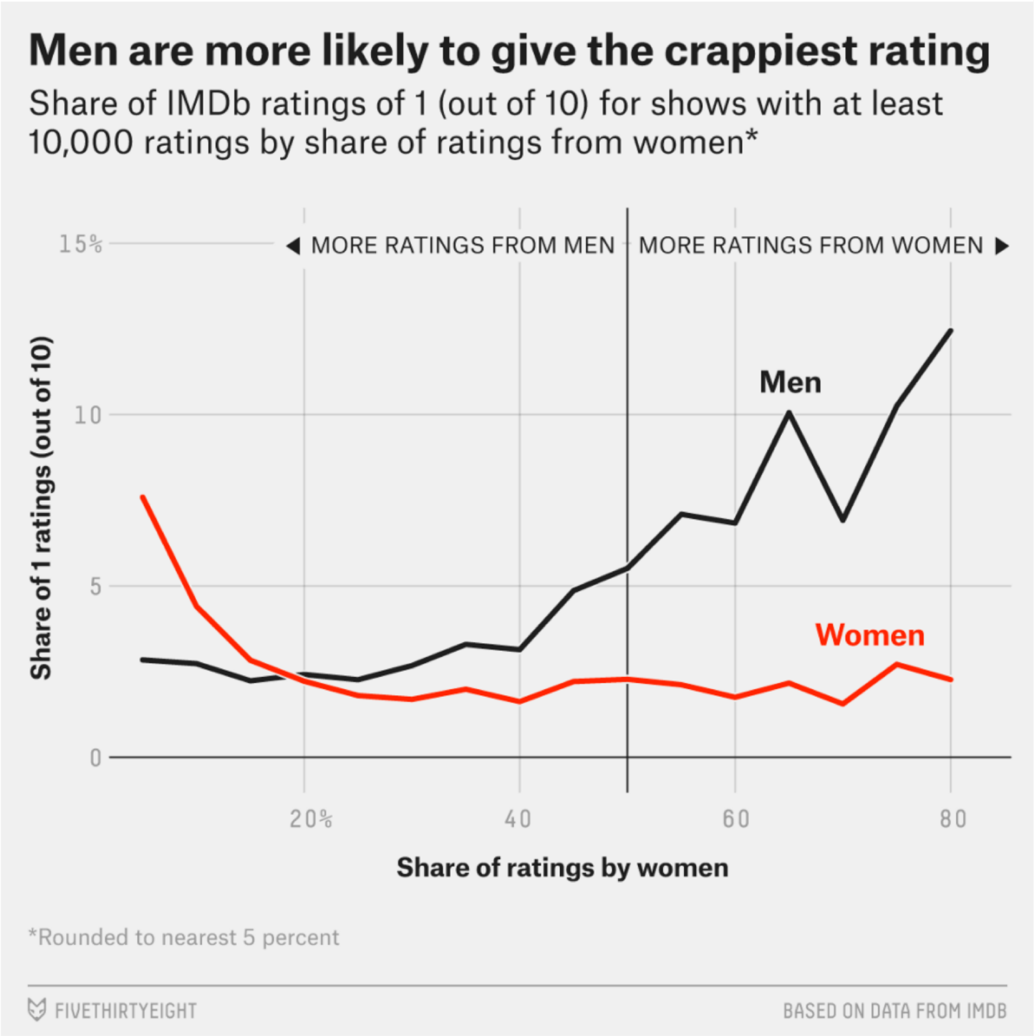
- Registered users rate films 1-10 stars; they are an overrepresented subpopulation relative to the general population
- Registered users who rate movies in their free time further over represents a specific segment of the general population
- *"Men Are Sabotaging The Online Reviews Of TV Shows Aimed At Women*[1]*"*
  - 60% who rated Sex in the City were women. Women gave it a 8.1, men gave it 5.8.

[1] **fivethirtyeight.com**

Dataset Considerations: Biases

IMDb Movie Ratings





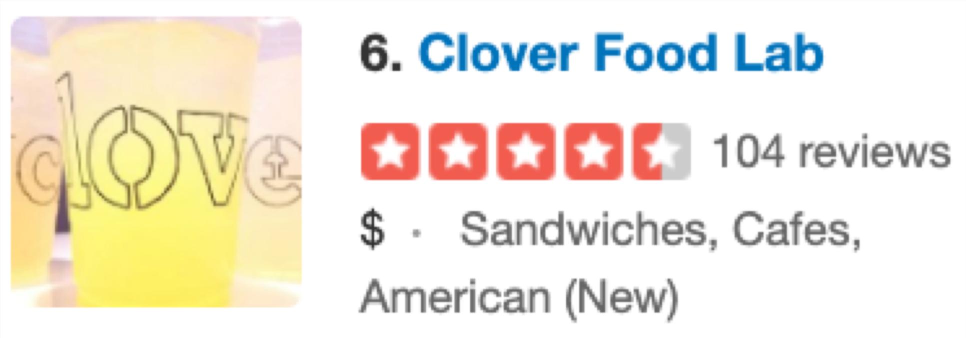
Dataset Considerations: Biases

Yelp Reviews

- Registered users rate businesses on a 1-5 star scale
- Registered users tend to represent a certain subset of the population (those who are more social media inclined and opinionated)
- Customers with extreme experiences are more likely to voice their opinions

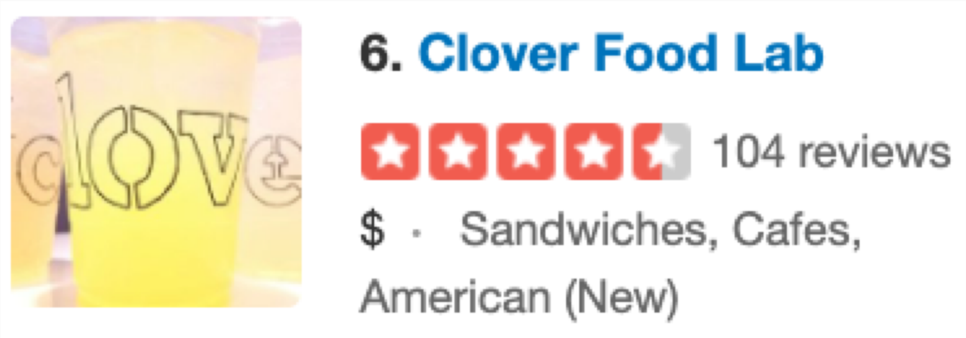
Dataset Considerations: Biases

Yelp Reviews



Dataset Considerations: Biases

Yelp Reviews



Longwood Medical

Harvard Square

Back to our example…

Let’s say that we are interested in the English Premier League (football/soccer) and want to build a model to predict a player’s market value.

Question Does age affect one’s market value?

Example: Get the data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

from www.transfermarkt.us

Example: Get the data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
[illegible]	[illegible]	[illegible]	GK	7
[illegible]	[illegible]	[illegible]	RW	20
[illegible]	[illegible]	[illegible]	CB	22

- Credible/Trustworthy?
- Possibly subjective market values?
- Sampled data

from www.transfermarkt.us

Example

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

Example: Explore the Data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny				22

Does it contain the necessary information?

Example: Explore the Data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

Missing data? Imputation needed?

Example: Explore the Data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

Are the data types okay (`df.dtypes`)? Should be casted?

Example: Explore the Data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

Are the values reasonable? `DataFrame.describe()` …

Example: Explore the Data

	age	page_views	fpl_value	fpl_points	market_value
count	461.000000	461.000000	461.000000	461.000000	461.000000
mean	26.804772	763.776573	5.447939	57.314534	11.012039
std	3.961892	931.805757	1.346695	53.113811	12.257403
min	17.000000	3.000000	4.000000	0.000000	0.050000
25%	24.000000	220.000000	4.500000	5.000000	3.000000
50%	27.000000	460.000000	5.000000	51.000000	7.000000
75%	30.000000	896.000000	5.500000	94.000000	15.000000
max	38.000000	7664.000000	12.500000	264.000000	75.000000

Are the values reasonable? `DataFrame.describe()` …

Example: Explore the Data

	age	page_views	fpl_value	fpl_points	market_value
count	461.000000	461.000000	461.000000	461.000000	461.000000
mean	26.804772	763.776573	5.447939	57.314534	11.012039
std	3.961892	931.805757	1.346695	53.113811	12.257403
min	17.000000	3.000000	4.000000	0.000000	0.050000
25%	24.000000	220.000000	4.500000	5.000000	3.000000
50%	27.000000	460.000000	5.000000	51.000000	7.000000
75%	30.000000	896.000000	5.500000	94.000000	15.000000
max	38.000000	7664.000000	12.500000	264.000000	75.000000

Summary statistics can only reveal so much

Lecture Outline

Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

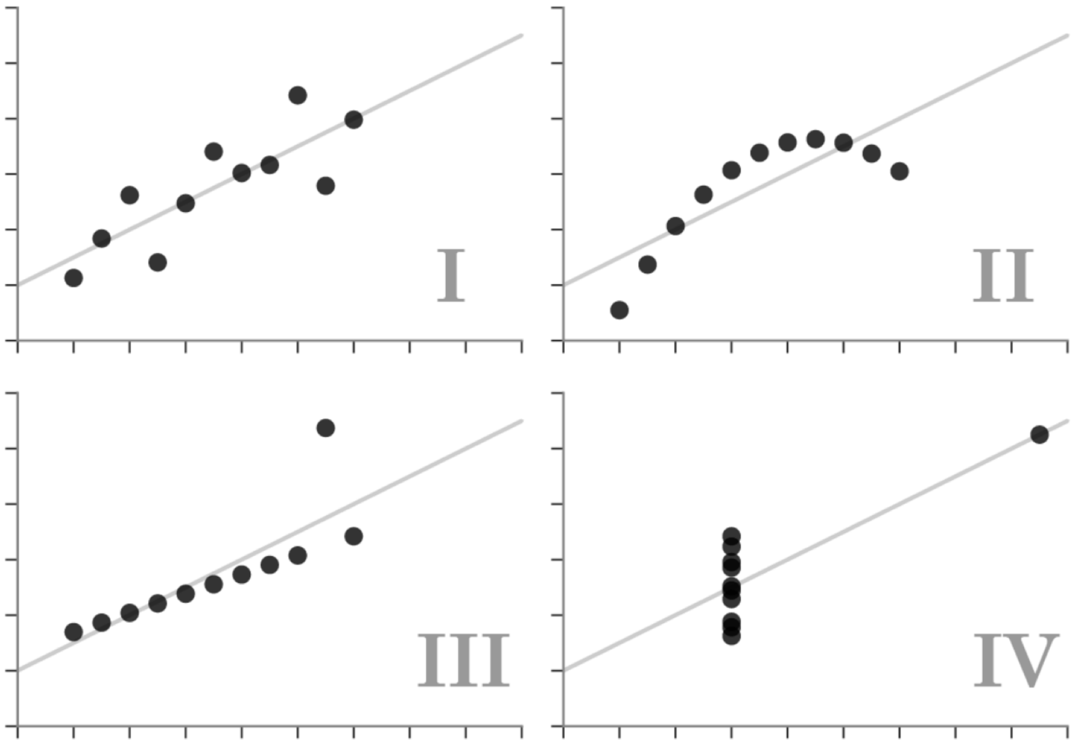
Visualization

- Exploration (EDA)
- Communication

Visualization

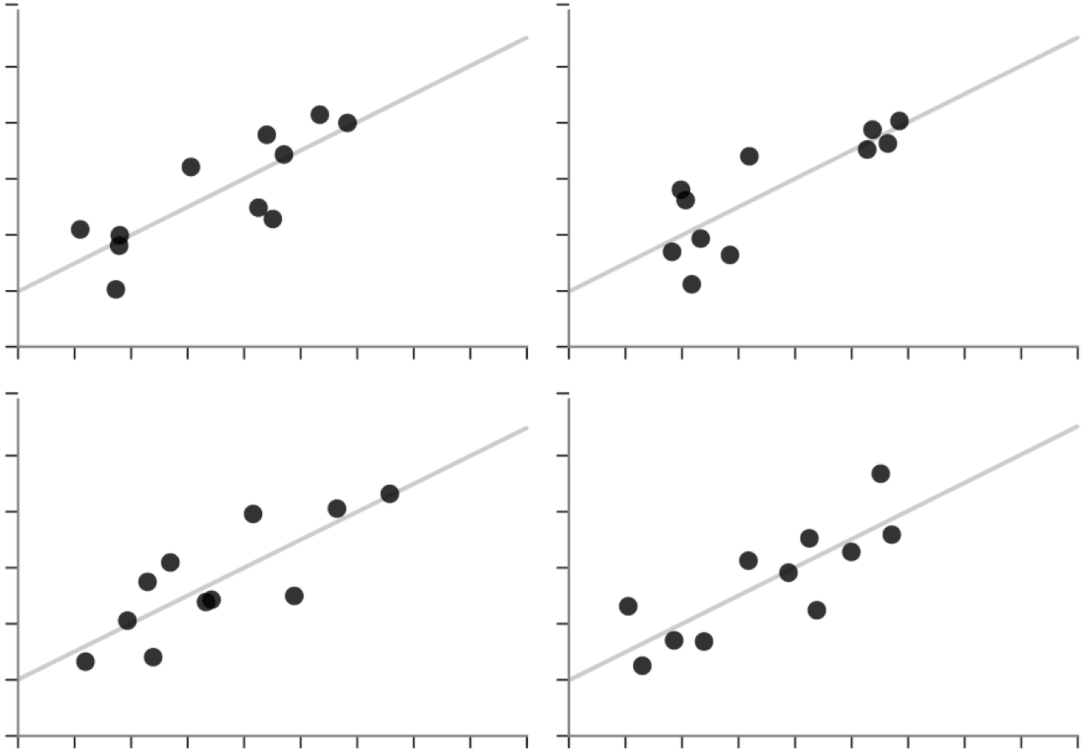
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



Same stats do not imply same graphs

Same graphs do not imply same stats

Visualization

	Antibiotic			
Bacteria	Penicillin	Streptomycin	Neomycin	Gram Staining
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

Visualization

	Antibiotic			
Bacteria	**Penicillin**	**Streptomycin**	**Neomycin**	**Gram Staining**
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

Visualization

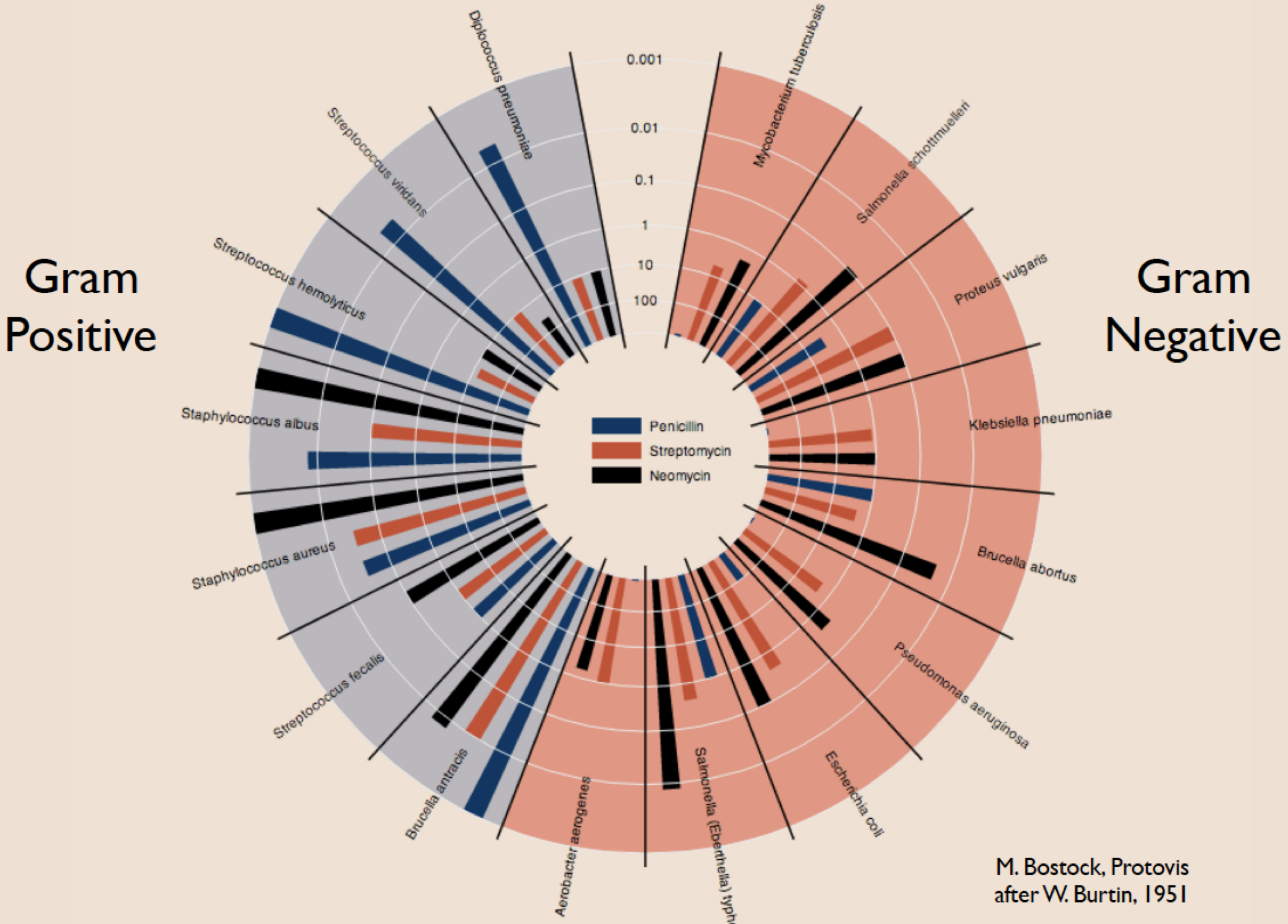
		Antibiotic		
Bacteria	**Penicillin**	**Streptomycin**	**Neomycin**	**Gram Staining**
Aerobacter aerogenes	870	1	1.6	negative
Brucella abortus	1	2	0.02	negative
Brucella anthracis	0.001	0.01	0.007	positive
Diplococcus pneumoniae	0.005	11	10	positive
Escherichia coli	100	0.4	0.1	negative
Klebsiella pneumoniae	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

What are some questions we could ask?

Visualization

	Antibiotic			
Bacteria	**Penicillin**	**Streptomycin**	**Neomycin**	**Gram Staining**
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 viridans	0.005	10	40	positive

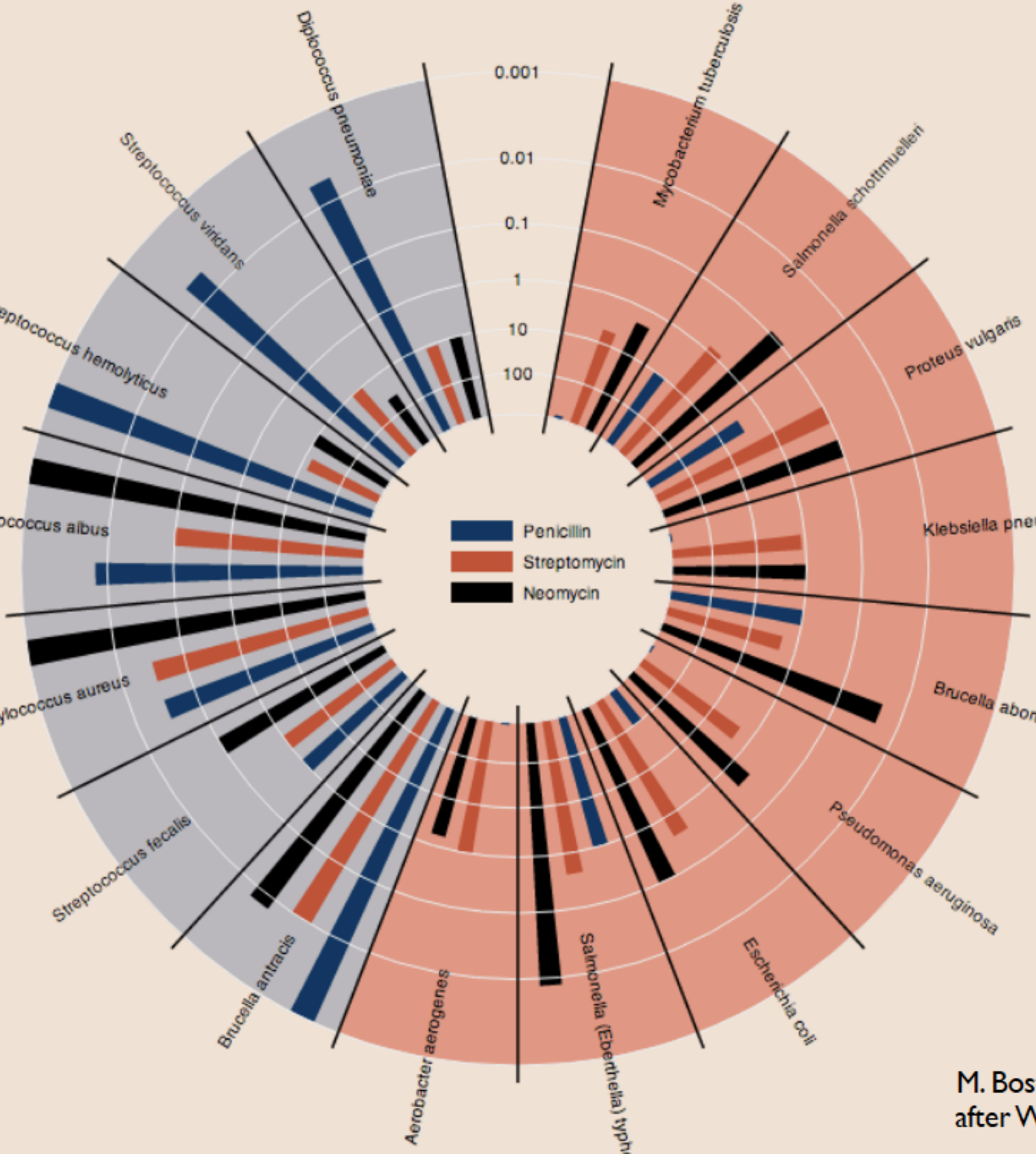
Q: How effective are the antibiotics?



M. Bostock, Protovis
after W. Burtin, 1951

Gram Positive

Gram Negative

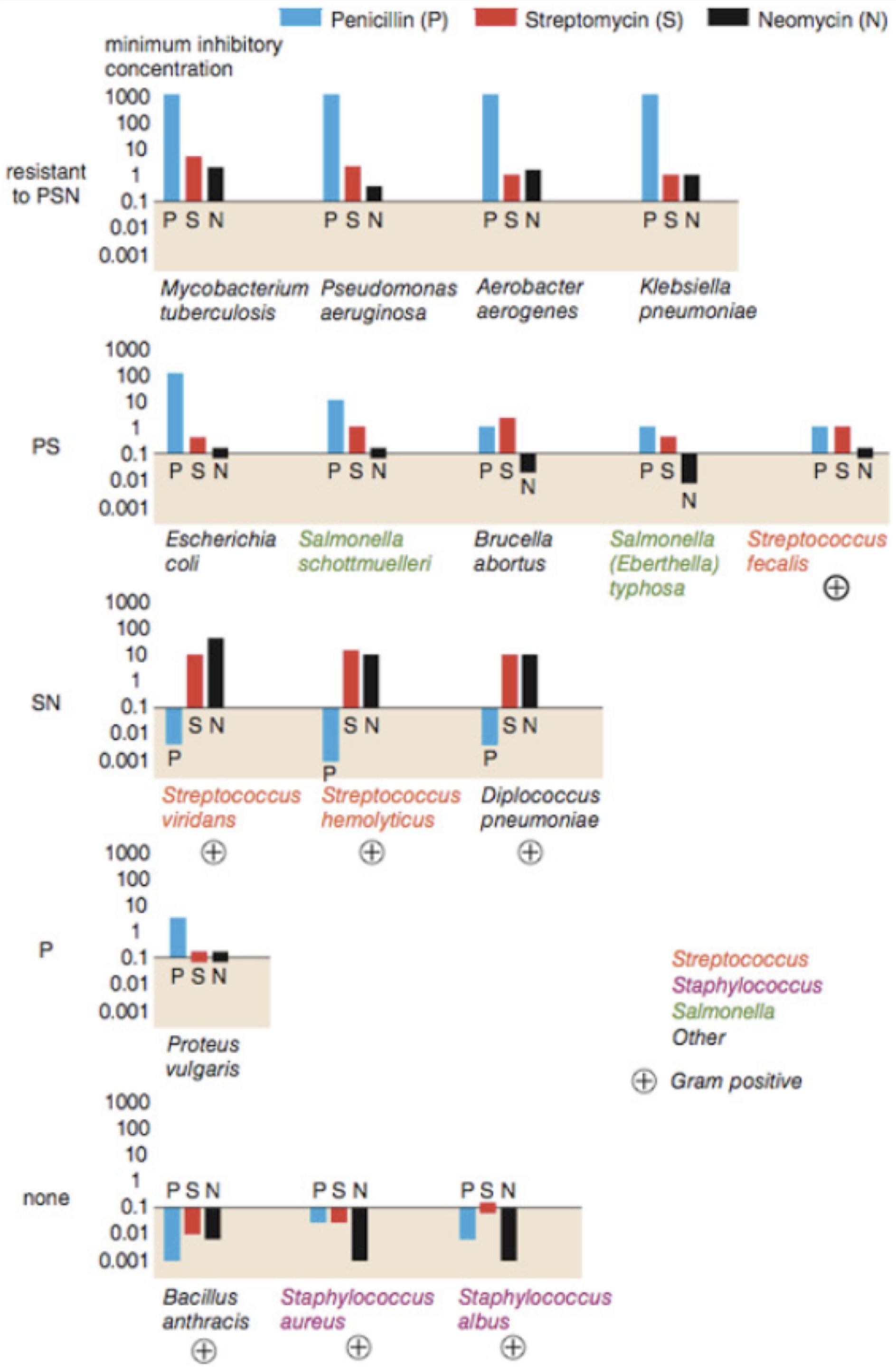


If bacteria is gram positive, Penicillin & Neomycin are most effective

If bacteria is gram negative, Neomycin is most effective

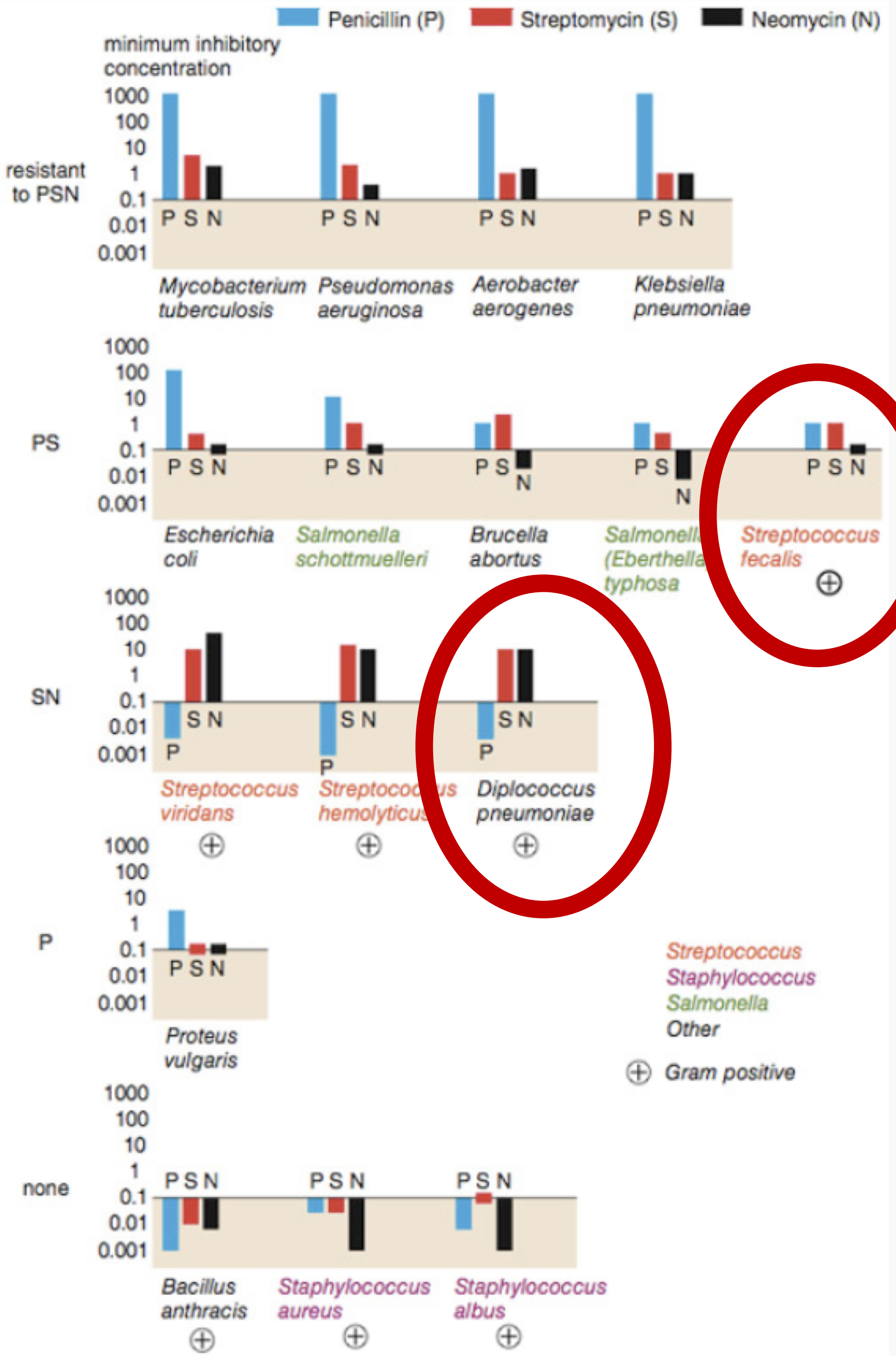
M. Bostock, Protovis after W. Burtin, 1951

How do the bacteria compare?



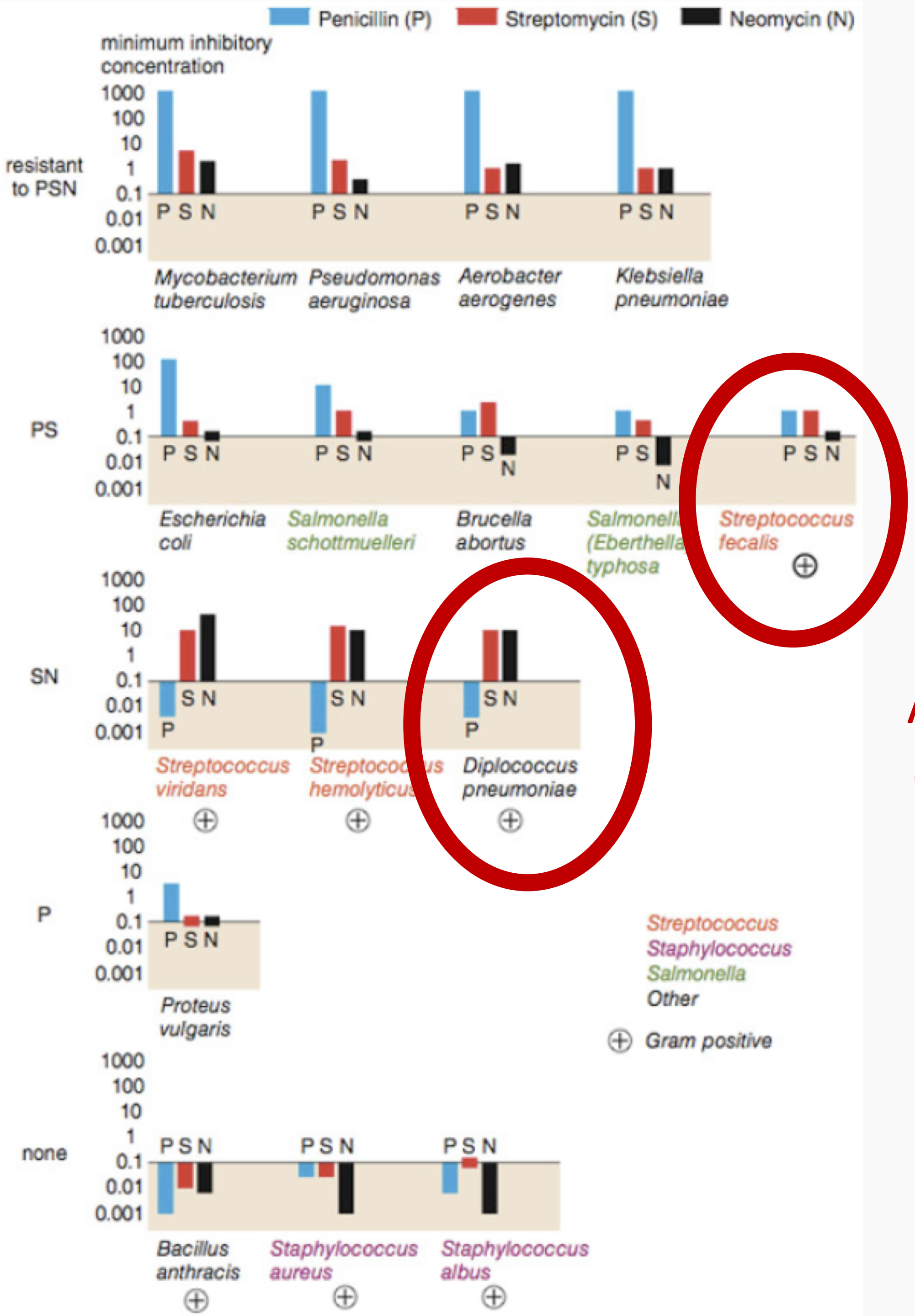
Wainer & Lysen, “That’s funny...”
American Scientist, 2009
Adapted from Brian Schmotzer

How do the bacteria compare?



Wainer & Lysen, “That’s funny...”
American Scientist, 2009
Adapted from Brian Schmotzer

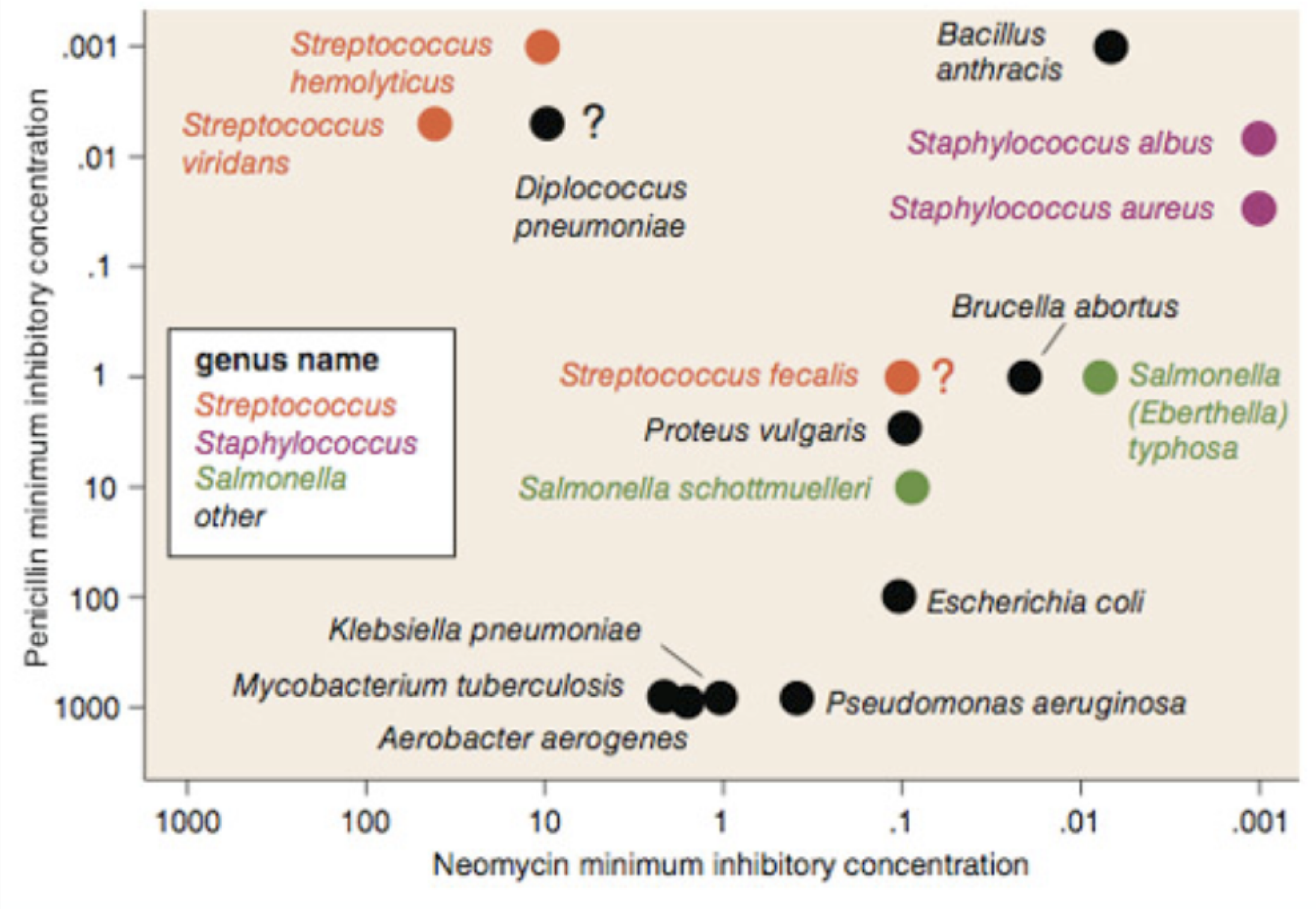
How do the bacteria compare?



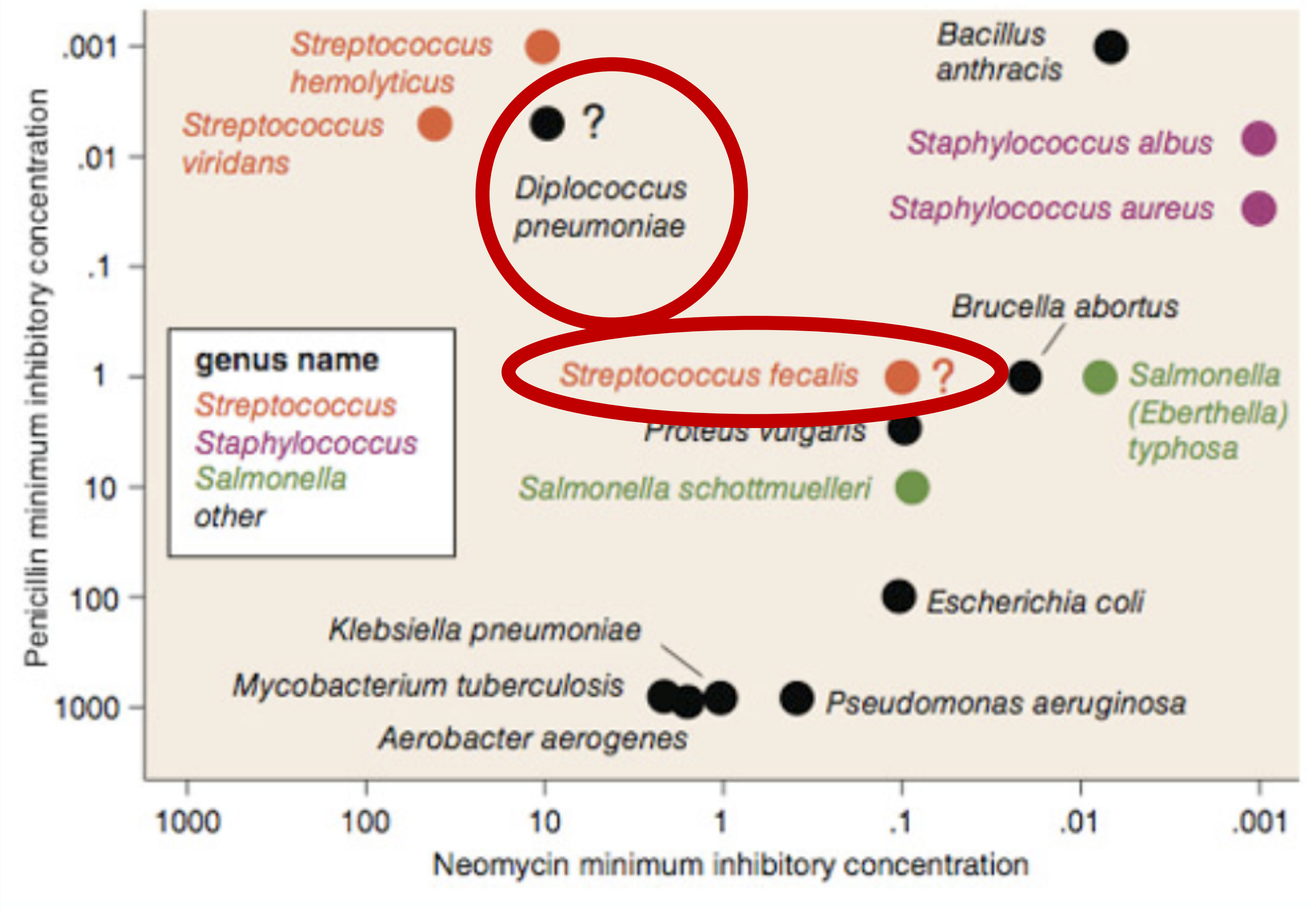
Not a streptococcus!
(realized ~30 years later)

Actually a streptococcus!
(realized ~20 years later)

Wainer & Lysen, “That’s funny...”
American Scientist, 2009
Adapted from Brian Schmotzer



Wainer & Lysen, “That’s funny...”
American Scientist, 2009



Wainer & Lysen, "That's funny..."
American Scientist, 2009

Visualization

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey

Visualization Goals

Communicate (explanatory)

- Present data and ideas
- Explain and inform
- Provide evidence and support
- Influence and persuade

Analyze (exploratory)

- Explore the data
- Assess a situation
- Determine how to proceed
- Decide what to do

Visualization Goals

Communicate (explanatory)

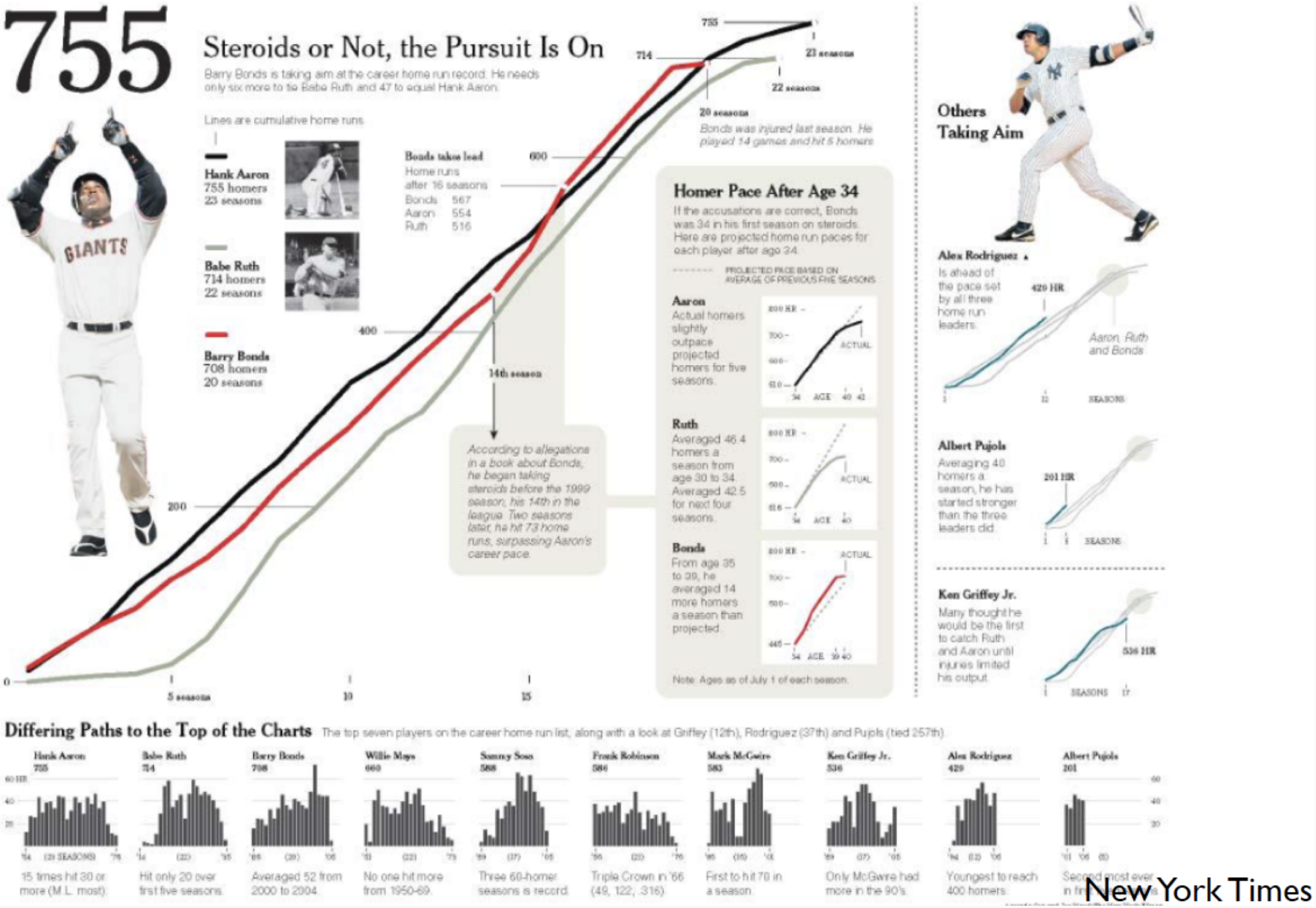
- Present data and ideas
- Explain and inform
- Provide evidence and support
- Influence and persuade

Analyze (exploratory)

- Explore the data
- Assess a situation
- Determine how to proceed
- Decide what to do

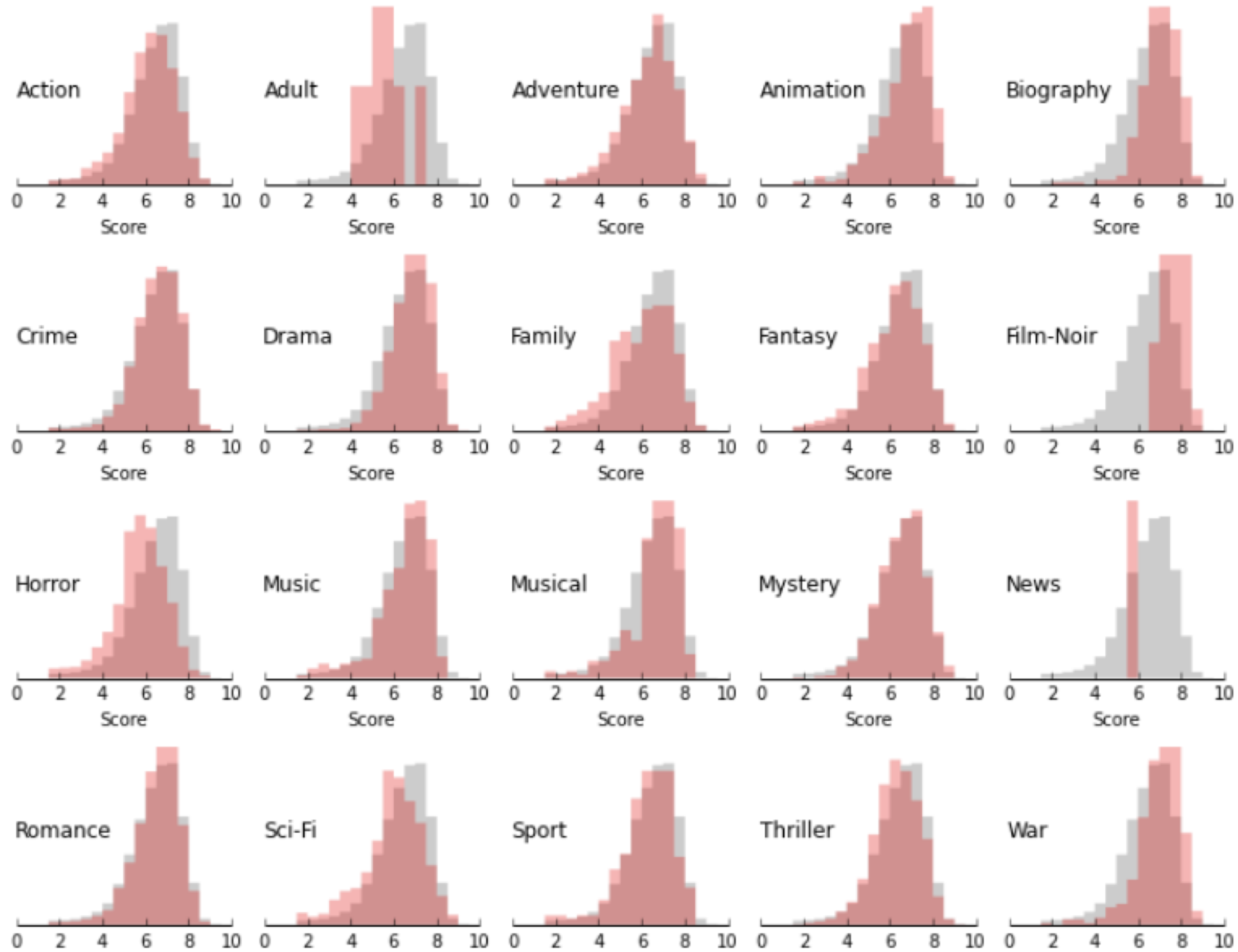
You're essentially communicating drafts to yourself

Communicate

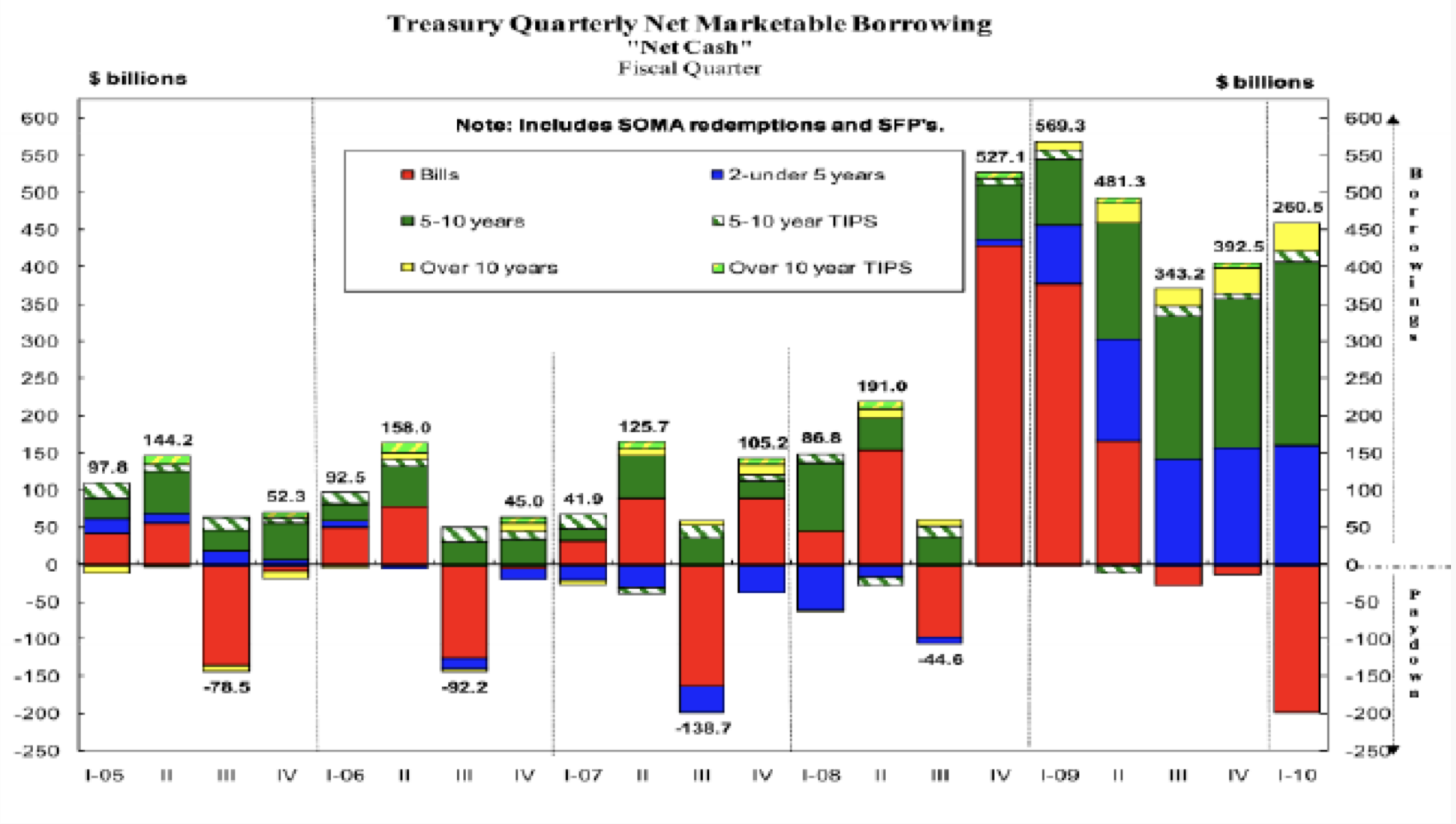


New York Times

Explore


Action
Adult
Adventure
Animation
Biography
Crime
Drama
Family
Fantasy
Film-Noir
Horror
Music
Musical
Mystery
News
Romance
Sci-Fi
Sport
Thriller
War
0 2 4 6 8 10
Score

Not Effective



Not Effective

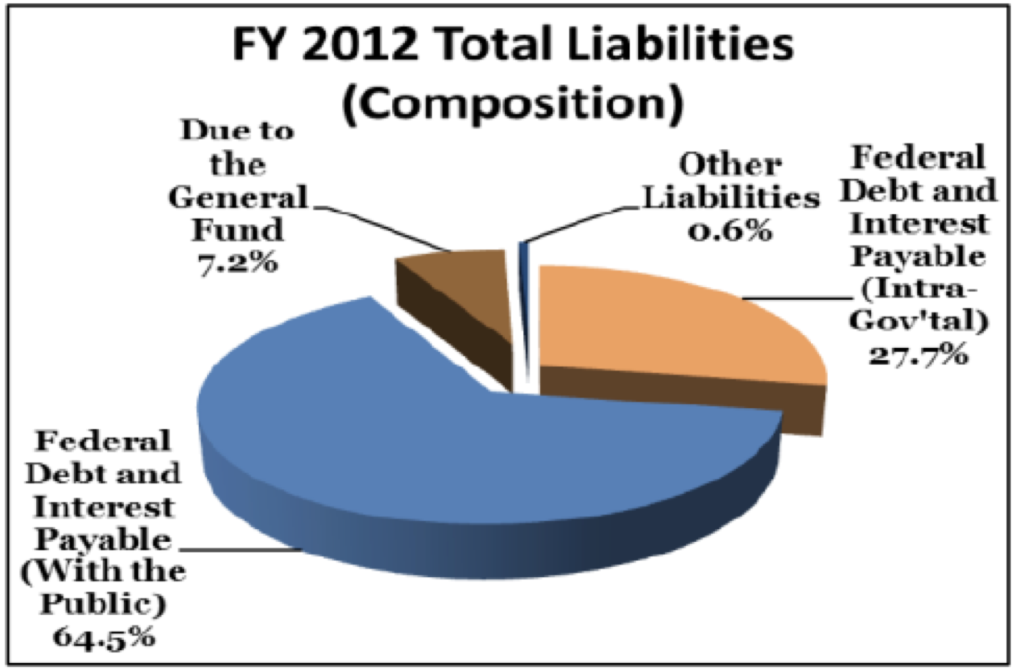
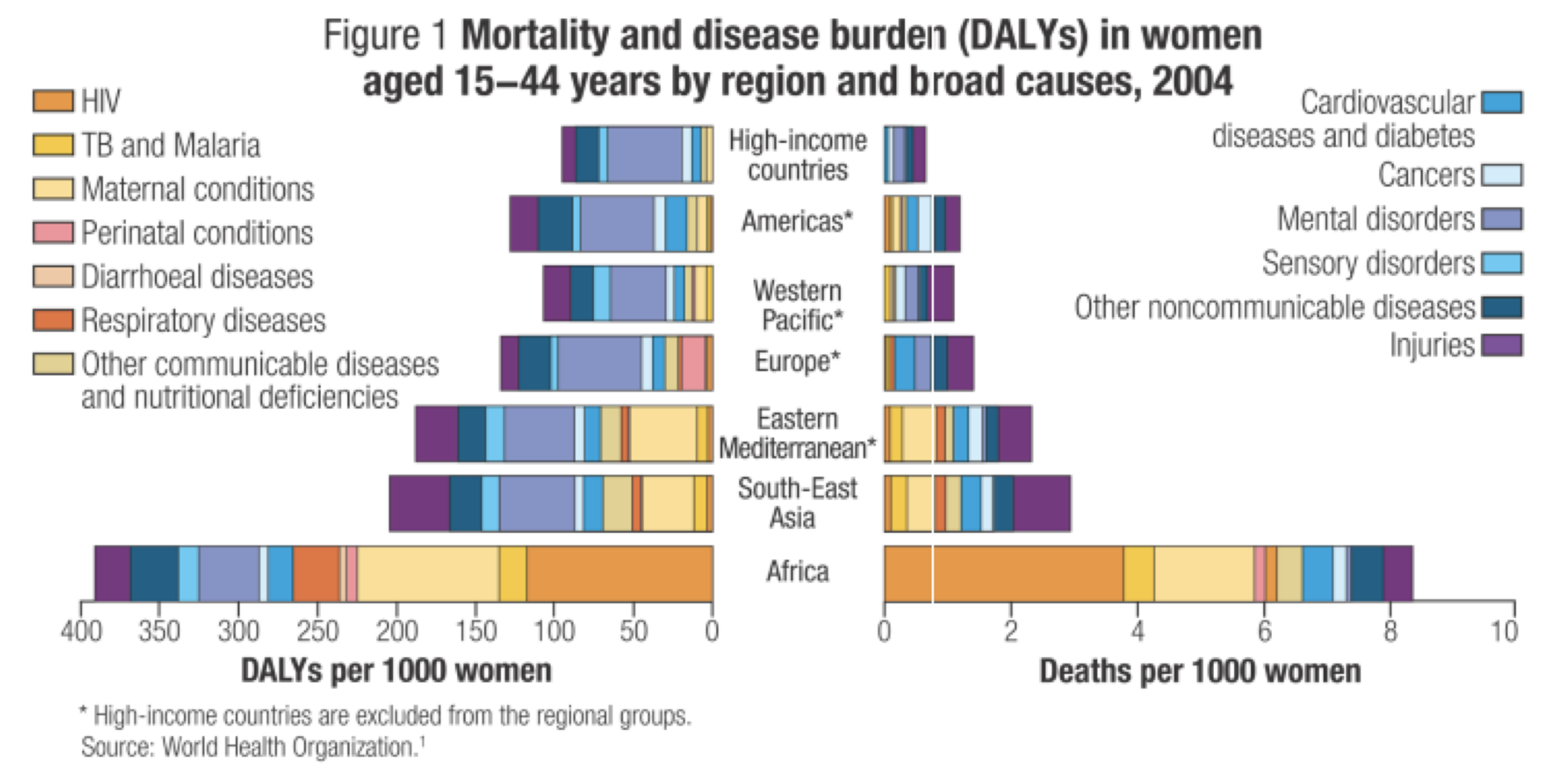
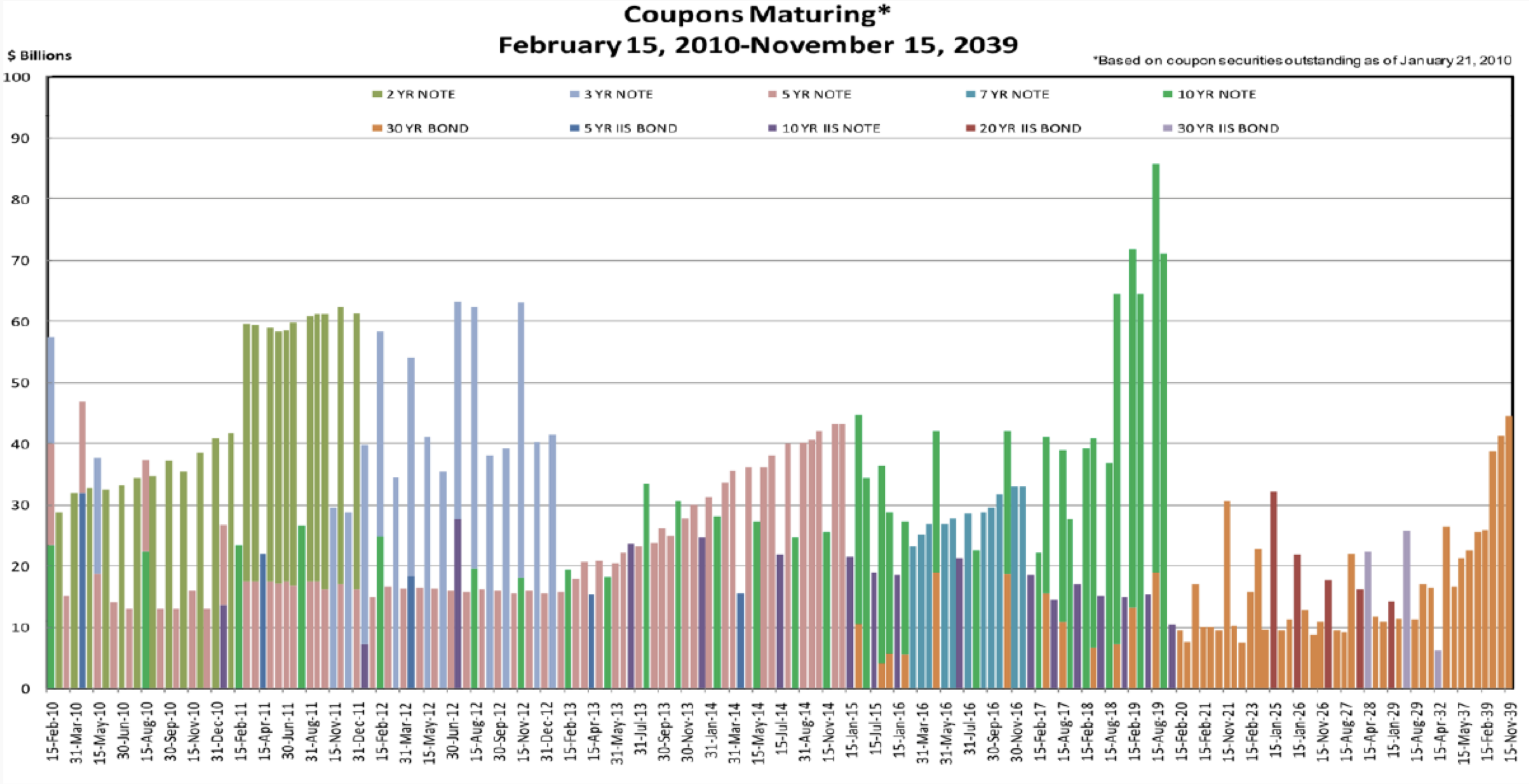


Figure 10

Not Effective



Not Effective



Not Effective

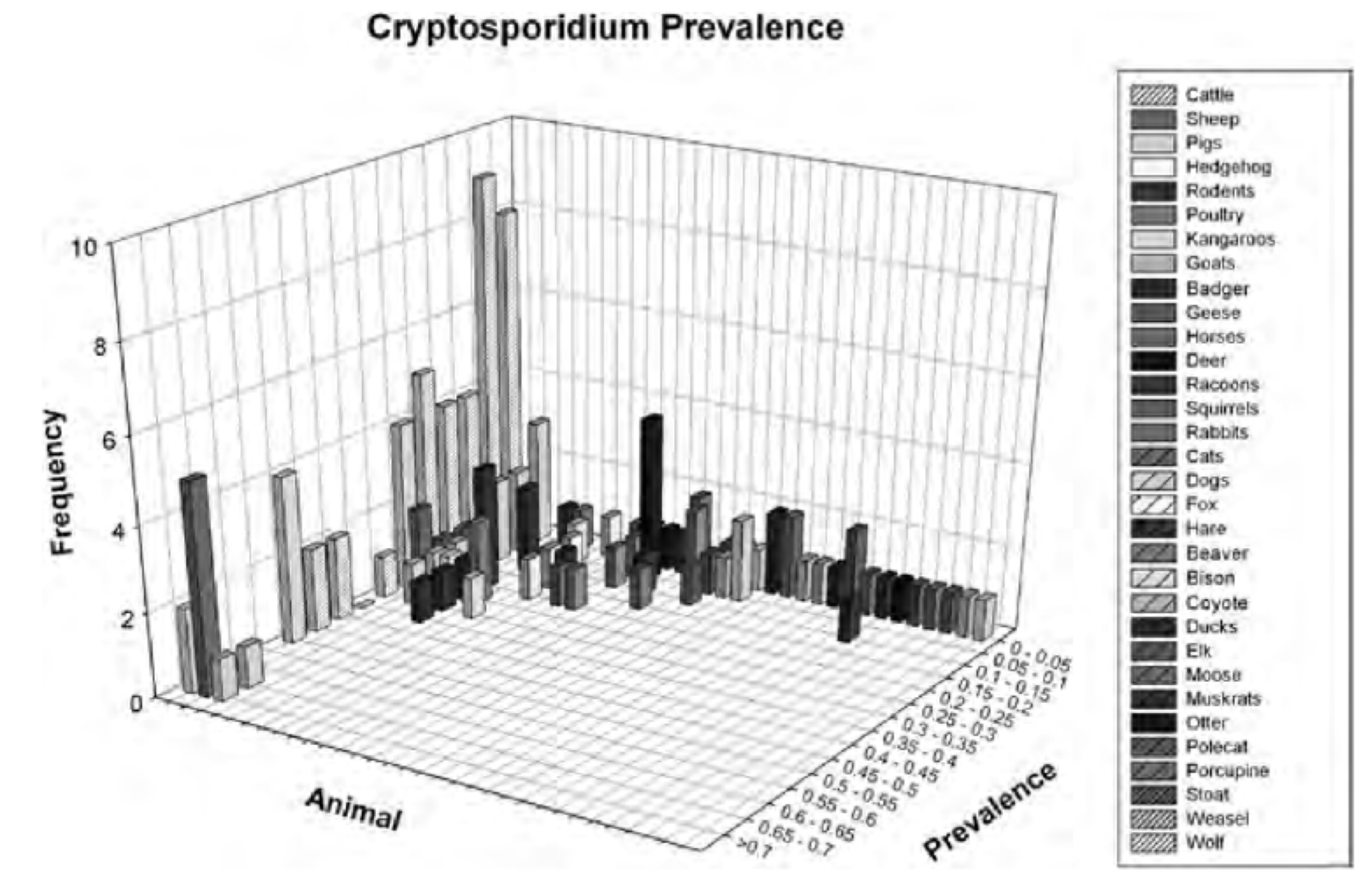


Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Visualization

Let’s say that we are interested in the English Premier League (football/soccer) and want to build a model to predict a player’s market value.

Question

Does age affect one’s market value?

What type of visualization would help us explore this question?