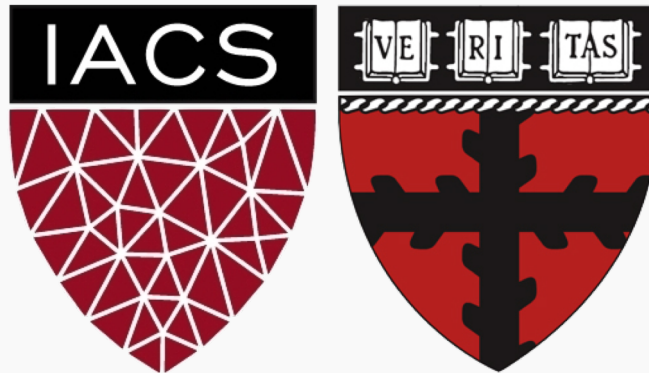


# Lecture 8: EDA

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

## 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

**The New York Times**

13 million articles



~500 million tweets per day

**CONGRESS.GOV**

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<sup>1</sup>"*
  - 60% who rated Sex in the City were women. Women gave it a 8.1, men gave it 5.8.

<sup>1</sup> [fivethirtyeight.com](https://www.fivethirtyeight.com)



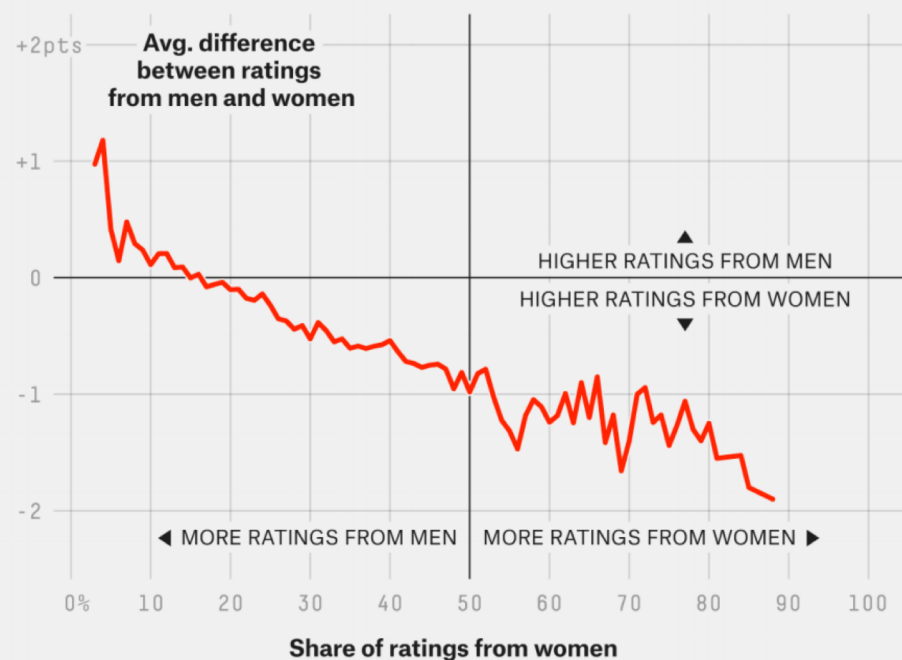


# Dataset Considerations: Biases

## IMDb Movie Ratings

### Men tank the ratings of shows aimed at women

Average difference between IMDb ratings of TV shows from men and women by share of ratings from women



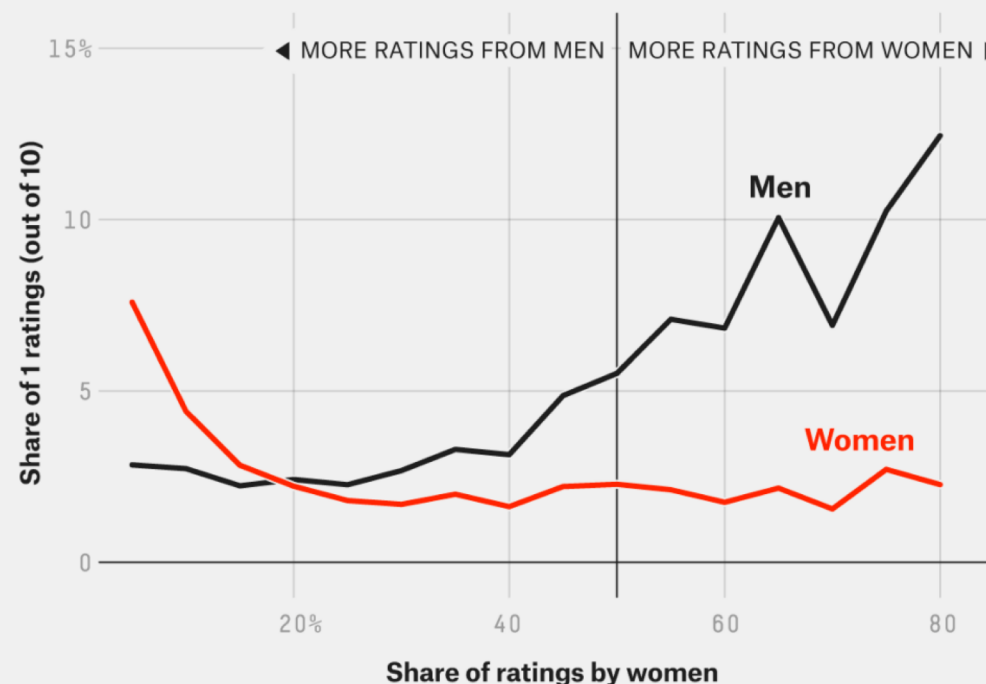
For English language shows with 1,000 or more ratings

FIVETHIRTYEIGHT

BASED ON DATA FROM IMDB

### Men are more likely to give the crappiest rating

Share of IMDb ratings of 1 (out of 10) for shows with at least 10,000 ratings by share of ratings from women\*



\*Rounded to nearest 5 percent

FIVETHIRTYEIGHT

BASED ON DATA FROM IMDB

# 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



### 6. Clover Food Lab

★ ★ ★ ★ ★ 104 reviews

\$ · Sandwiches, Cafes,  
American (New)



### 1. Clover Food Lab

★ ★ ★ ★ ★ 821 reviews

\$\$ · American (New),  
Sandwiches, Cafes

# Dataset Considerations: Biases

## Yelp Reviews



### 6. Clover Food Lab

★★★★☆ 104 reviews

\$ · Sandwiches, Cafes,  
American (New)



### 1. Clover Food Lab

★★★★☆ 821 reviews

\$\$ · American (New),  
Sandwiches, Cafes

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](http://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
...	...	...	GK	7
...	...	...	RW	20
...	...	...	CB	22

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

from [www.transfermarkt.us](http://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
<b>count</b>	461.000000	461.000000	461.000000	461.000000	461.000000
<b>mean</b>	26.804772	763.776573	5.447939	57.314534	11.012039
<b>std</b>	3.961892	931.805757	1.346695	53.113811	12.257403
<b>min</b>	17.000000	3.000000	4.000000	0.000000	0.050000
<b>25%</b>	24.000000	220.000000	4.500000	5.000000	3.000000
<b>50%</b>	27.000000	460.000000	5.000000	51.000000	7.000000
<b>75%</b>	30.000000	896.000000	5.500000	94.000000	15.000000
<b>max</b>	38.000000	7664.000000	12.500000	264.000000	75.000000

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

# Example: Explore the Data

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

Summary statistics can only reveal so much

## Data Science Process Example

- Dataset considerations
  - Comprehensive vs Sampled
  - Biases

## Visualization

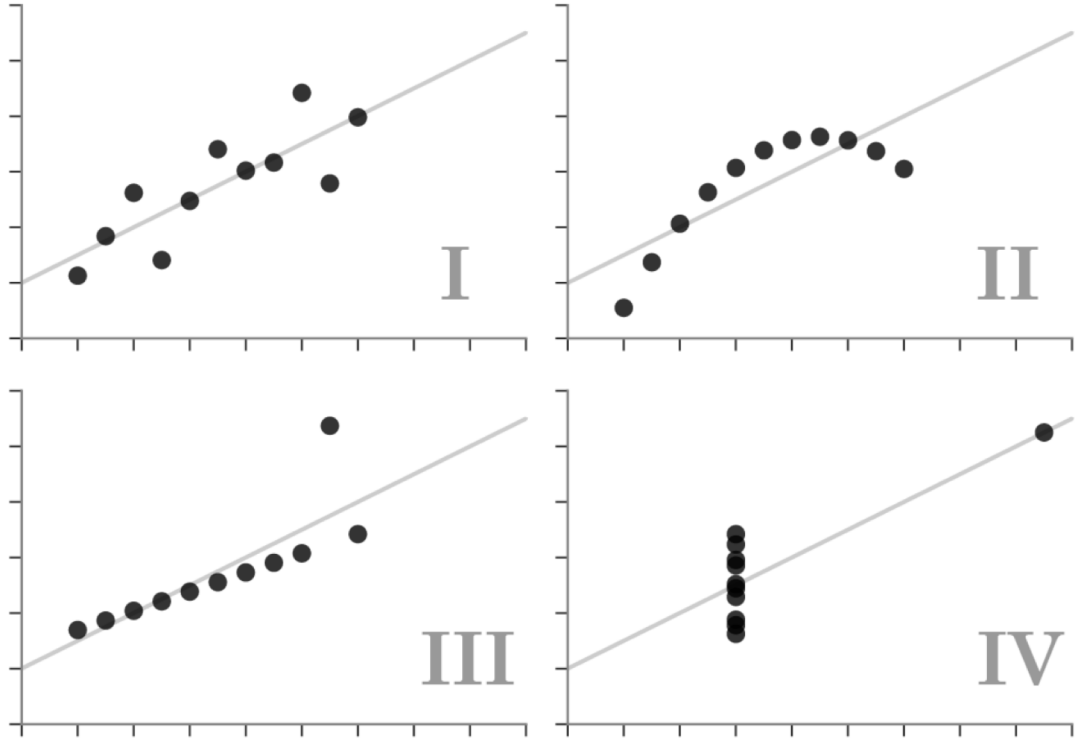
- Exploration (EDA)
- Communication

# Visualization



## Anscombe's Quartet

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

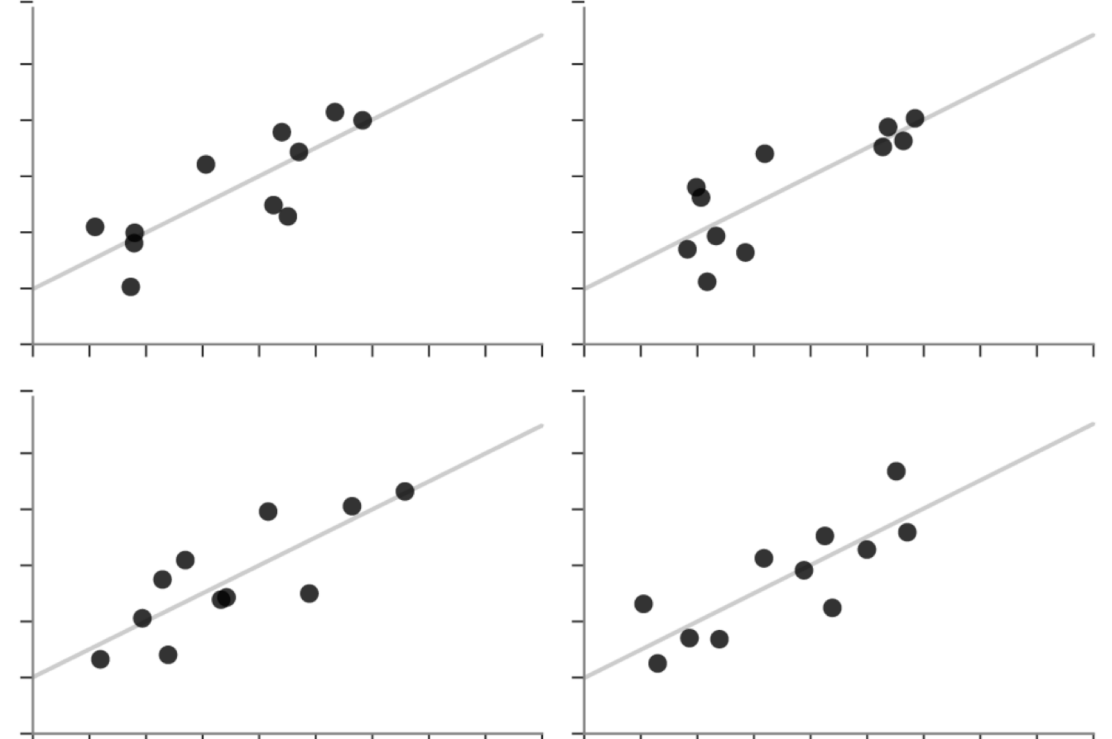


Same stats do not imply same graphs



## Unstructured Quartet

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



Same graphs do not imply same stats



# Visualization

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

# Visualization

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

# Visualization

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

What are some questions we could ask?

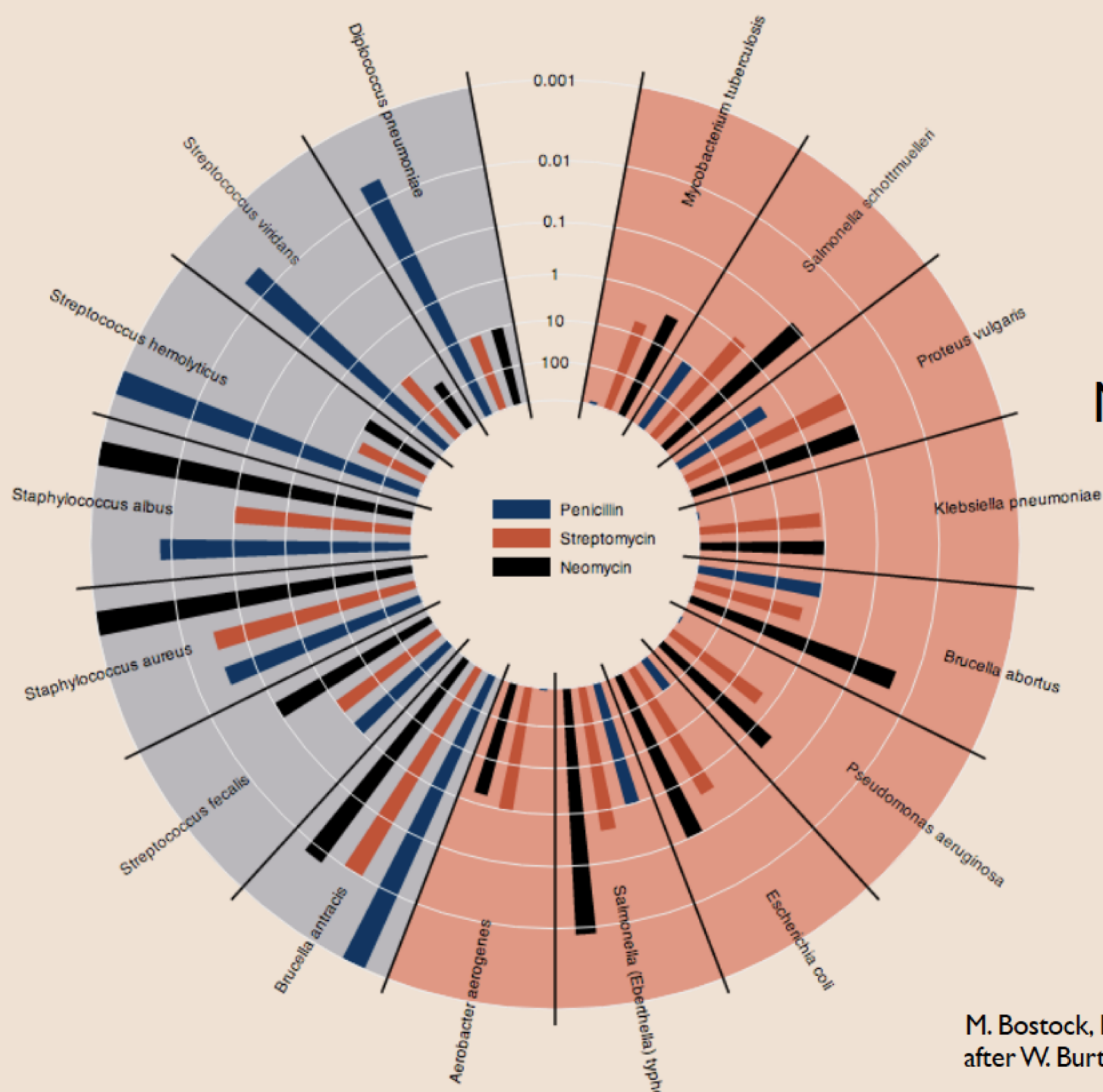
# Visualization

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

Q: How effective are the antibiotics?

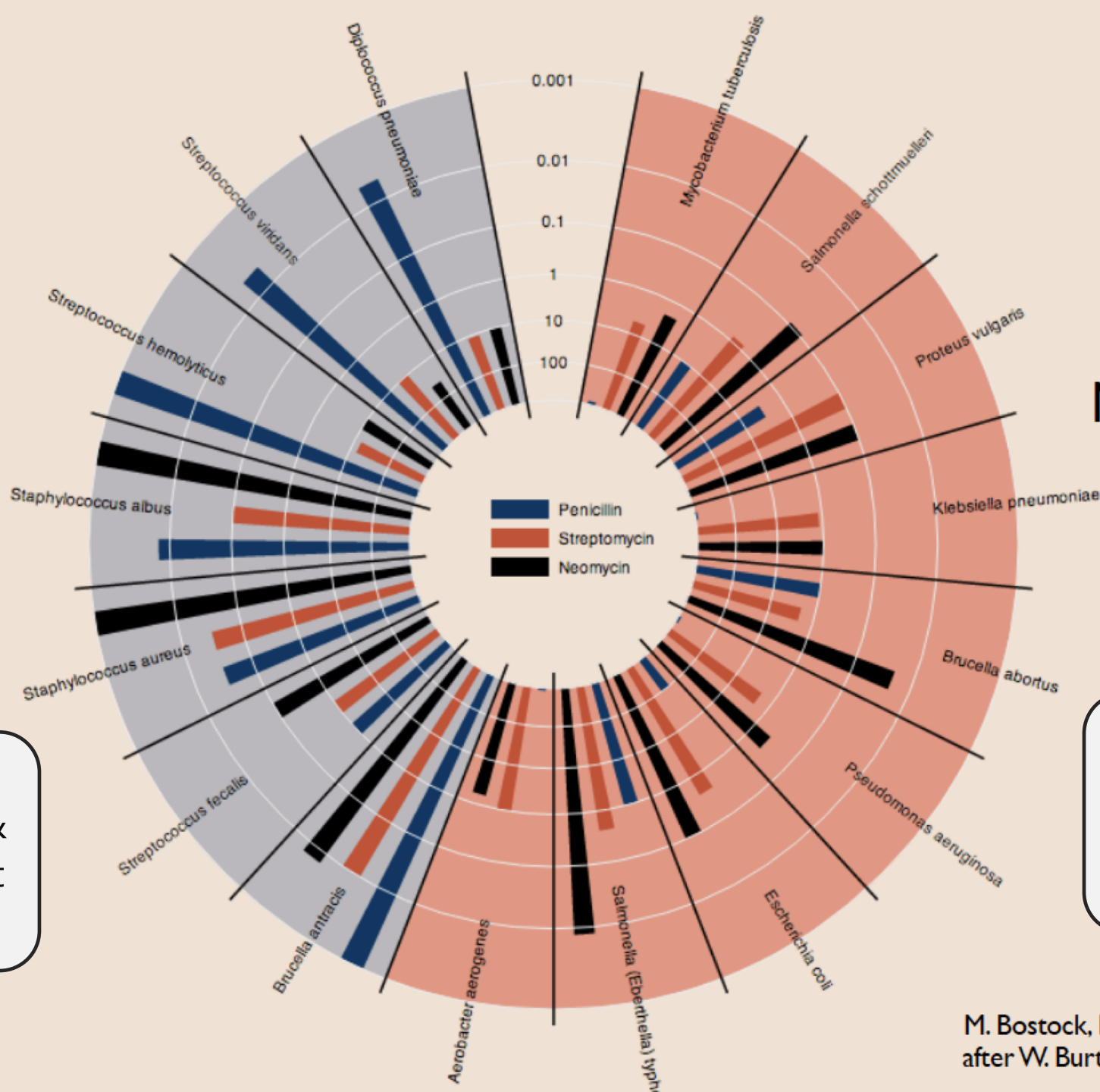
Gram Positive

Gram Negative



Gram Positive

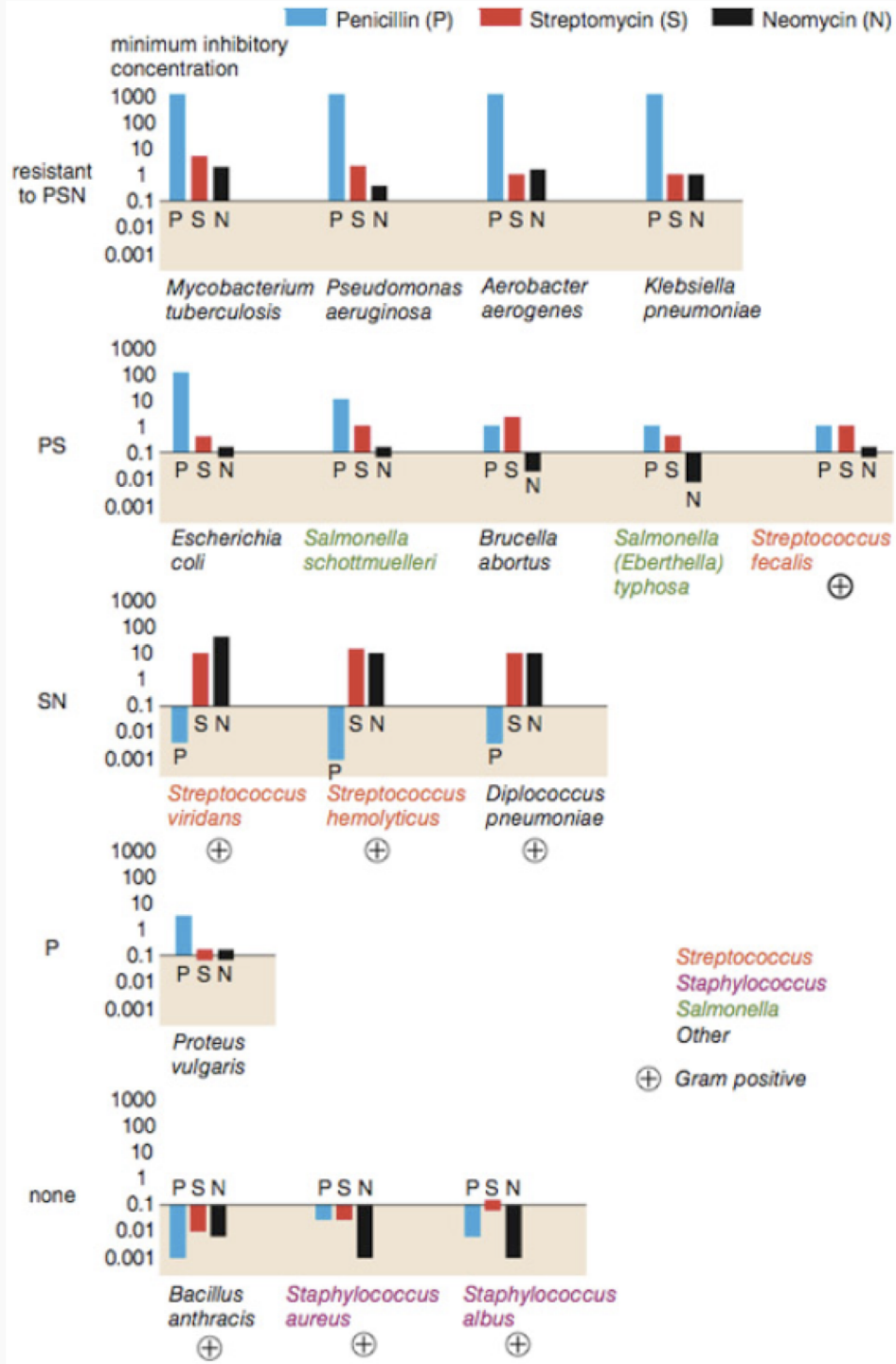
Gram Negative



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

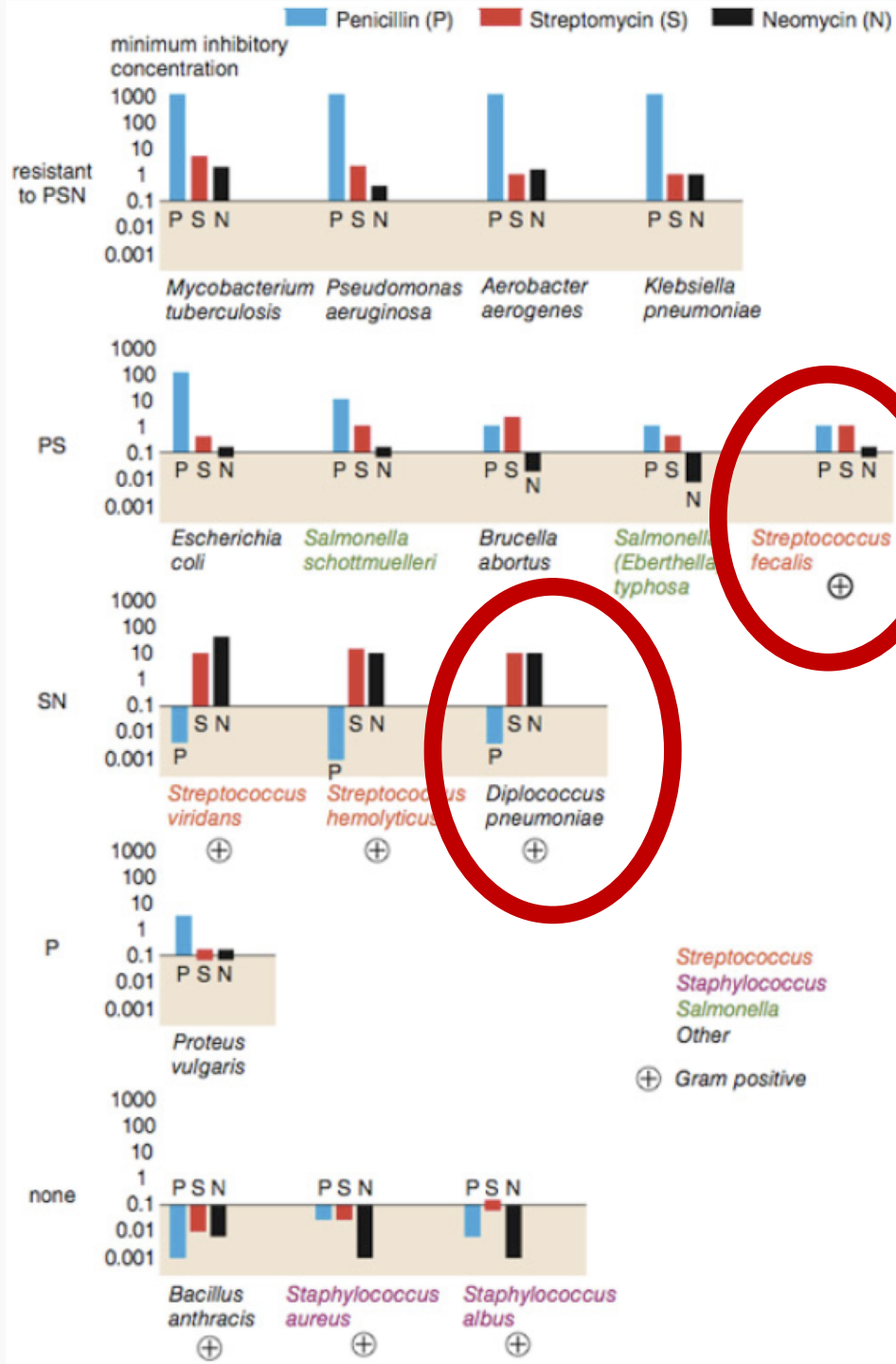
If bacteria is gram negative, Neomycin is most effective

# How do the bacteria compare?



Wainer & Lysen, "That's funny..."  
 American Scientist, 2009  
 Adapted from Brian Schmotzer

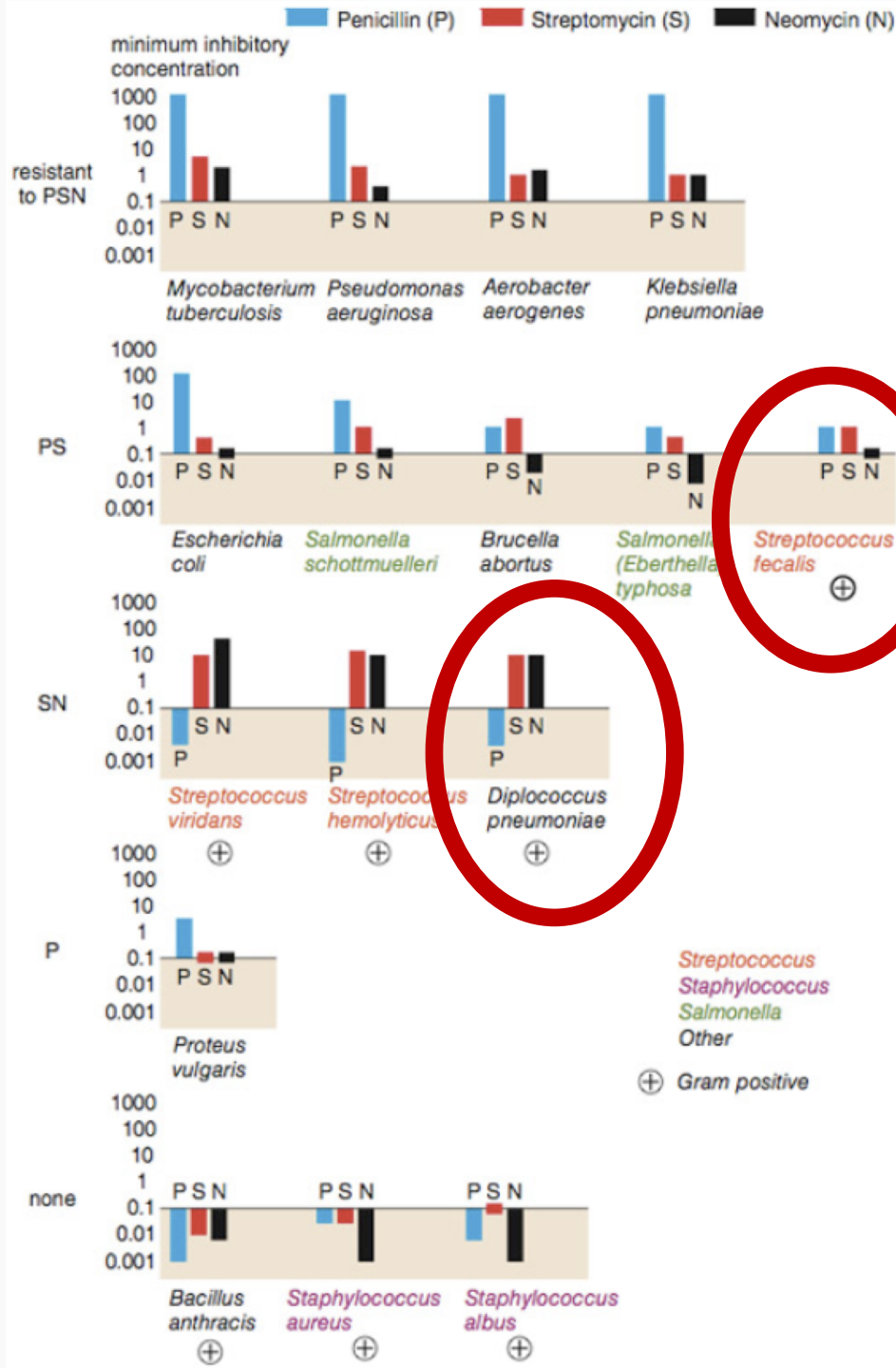
How do the bacteria compare?



Wainer & Lysen, "That's funny..."  
 American Scientist, 2009  
 Adapted from Brian Schmotzer



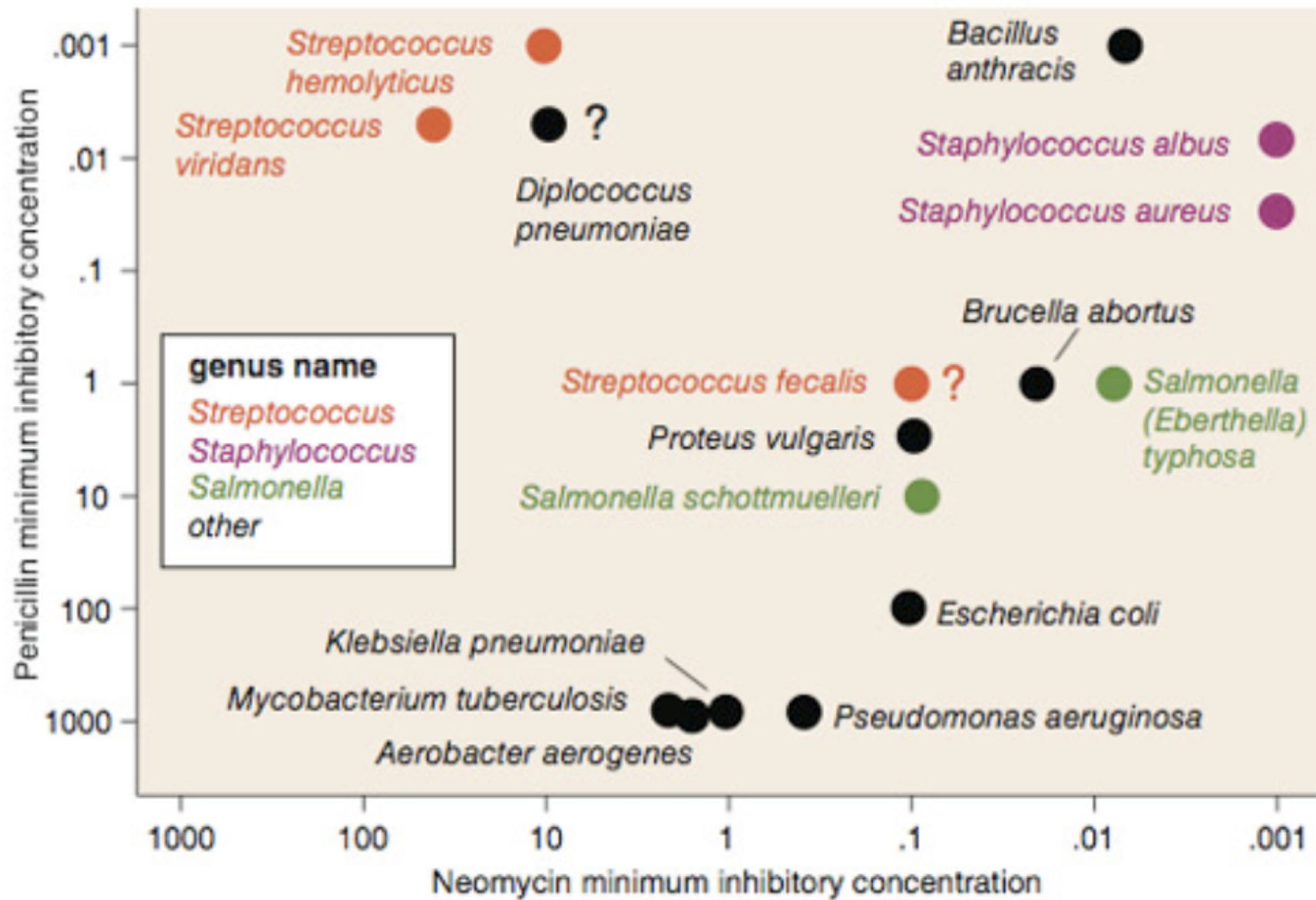
How do the bacteria compare?



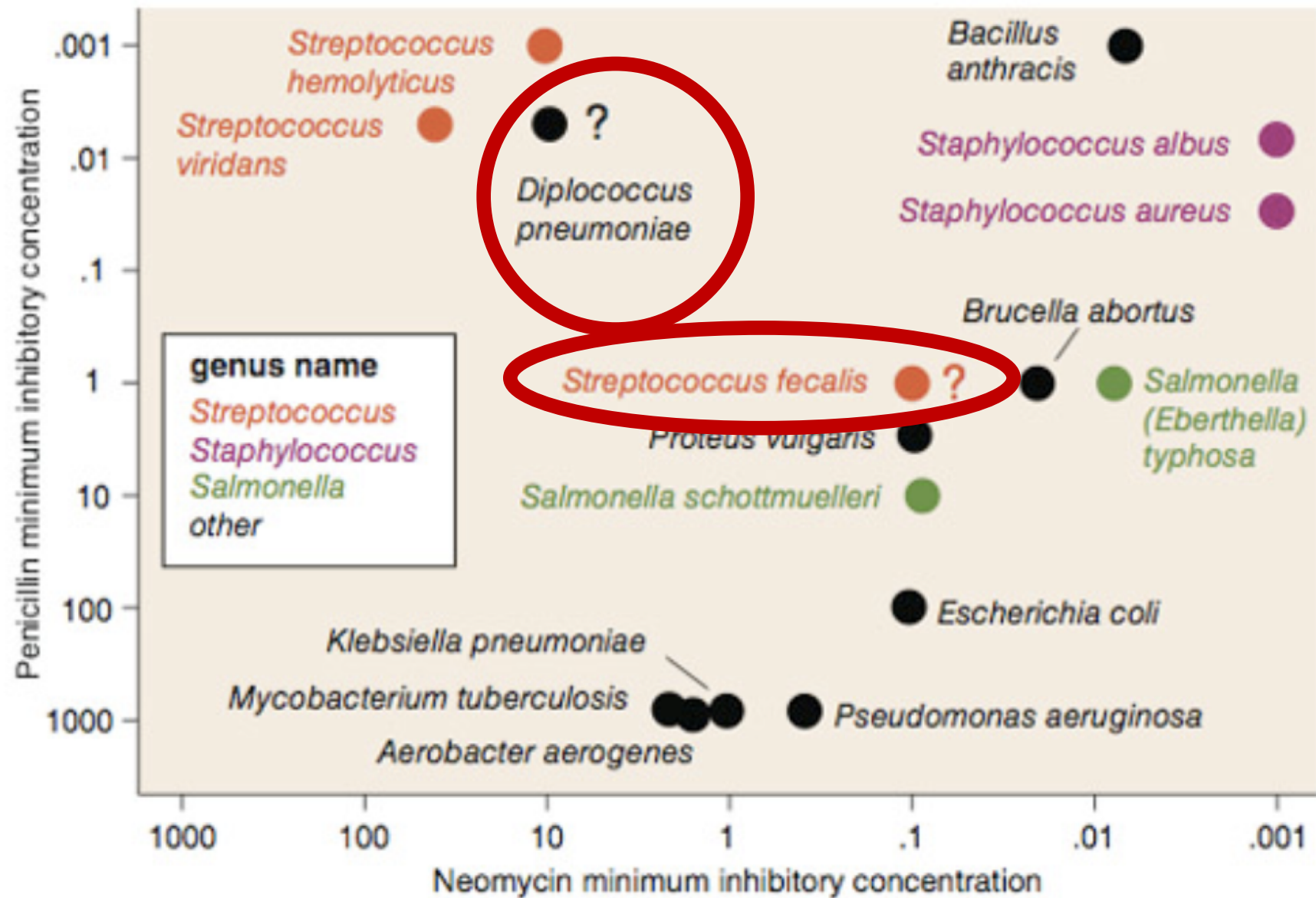
Not a streptococcus!  
(realized ~30 years later)

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

# 755



## Steroids or Not, the Pursuit Is On

Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.

Lines are cumulative home runs.

**Hank Aaron**  
755 homers  
23 seasons

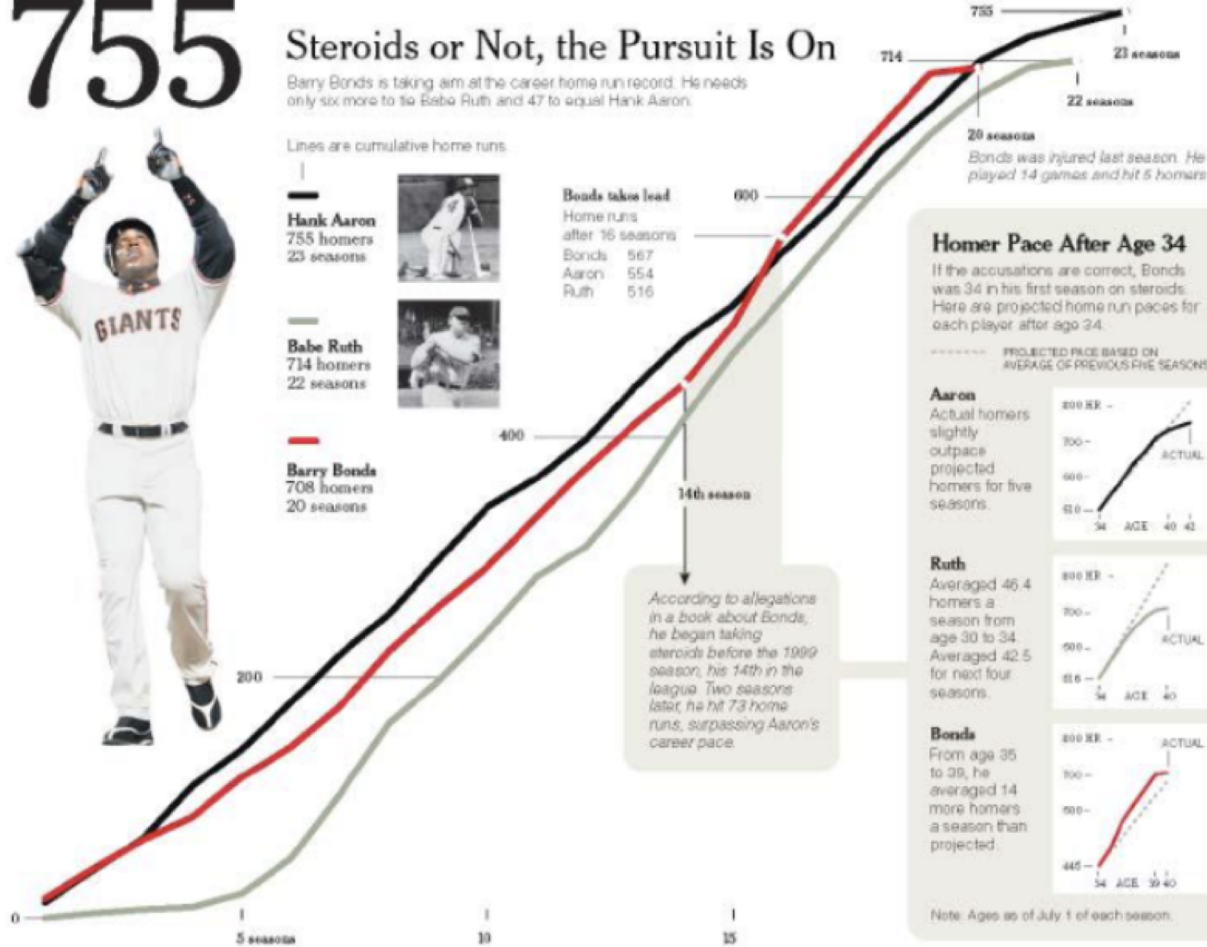


**Babe Ruth**  
714 homers  
22 seasons



**Barry Bonds**  
708 homers  
20 seasons

**Bonds takes lead**  
Home runs after 16 seasons:  
Bonds 567  
Aaron 554  
Ruth 516



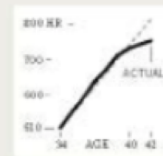
### Homer Pace After Age 34

If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

----- PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

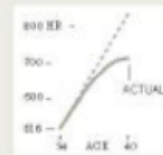
#### Aaron

Actual homers slightly outpace projected homers for five seasons.



#### Ruth

Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.



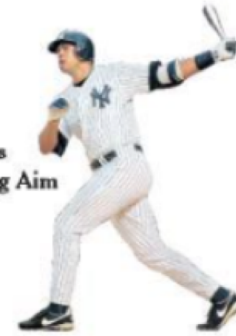
#### Bonds

From age 35 to 39, he averaged 14 more homers a season than projected.



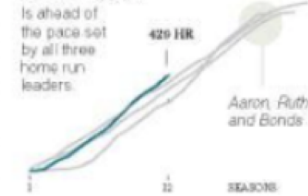
Note: Ages as of July 1 of each season.

### Others Taking Aim



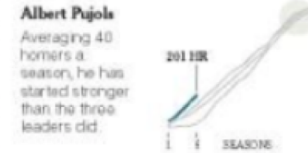
#### Alex Rodriguez

Is ahead of the pace set by all three home run leaders.



#### Albert Pujols

Averaging 40 homers a season, he has started stronger than the three leaders did.



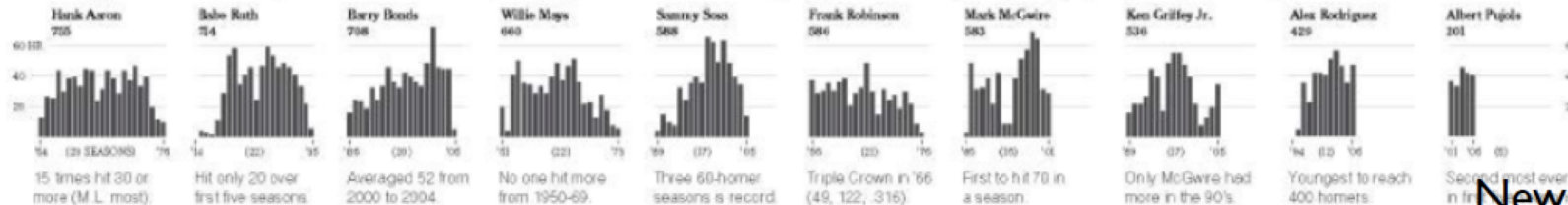
#### Ken Griffey Jr.

Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.

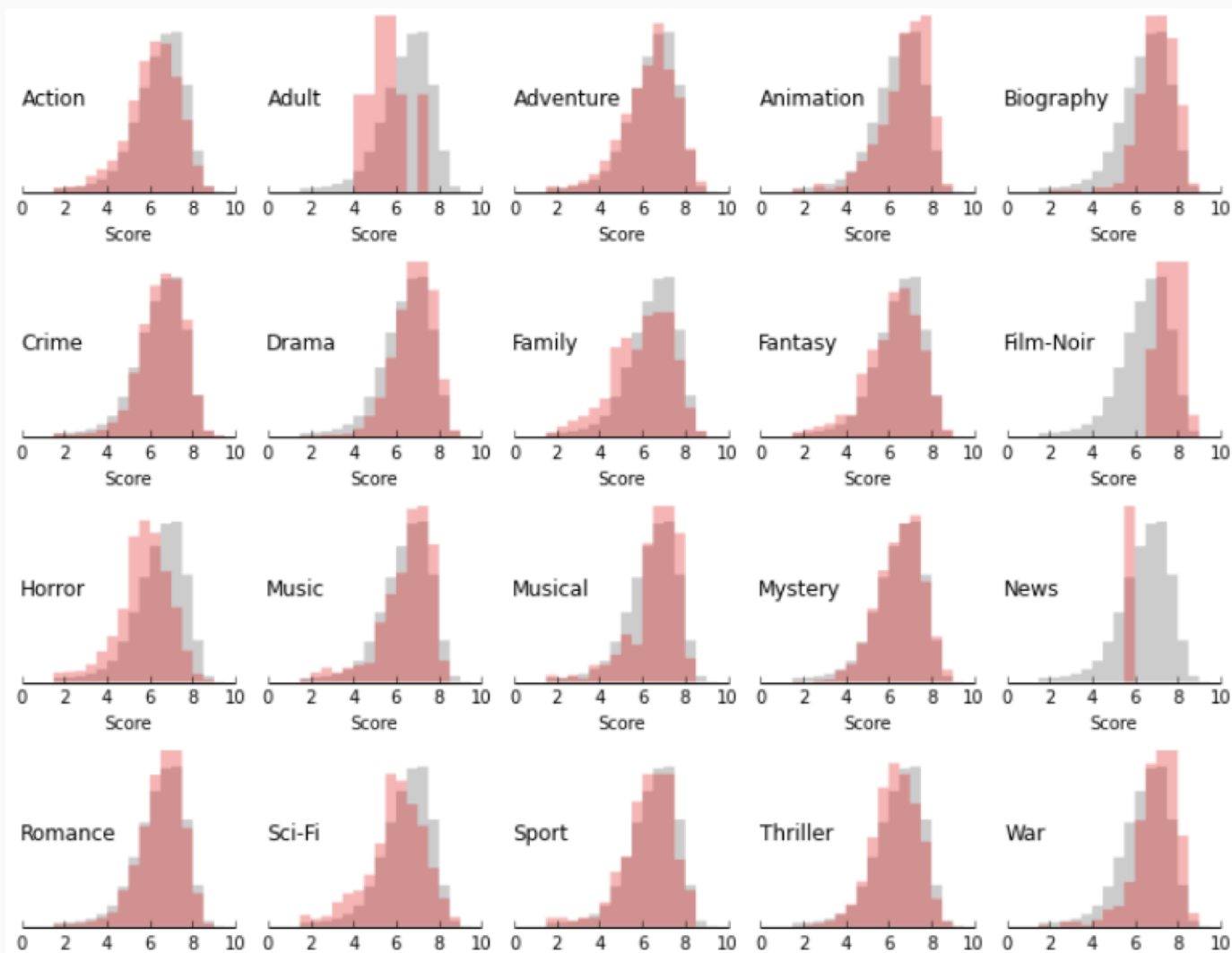


### Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).

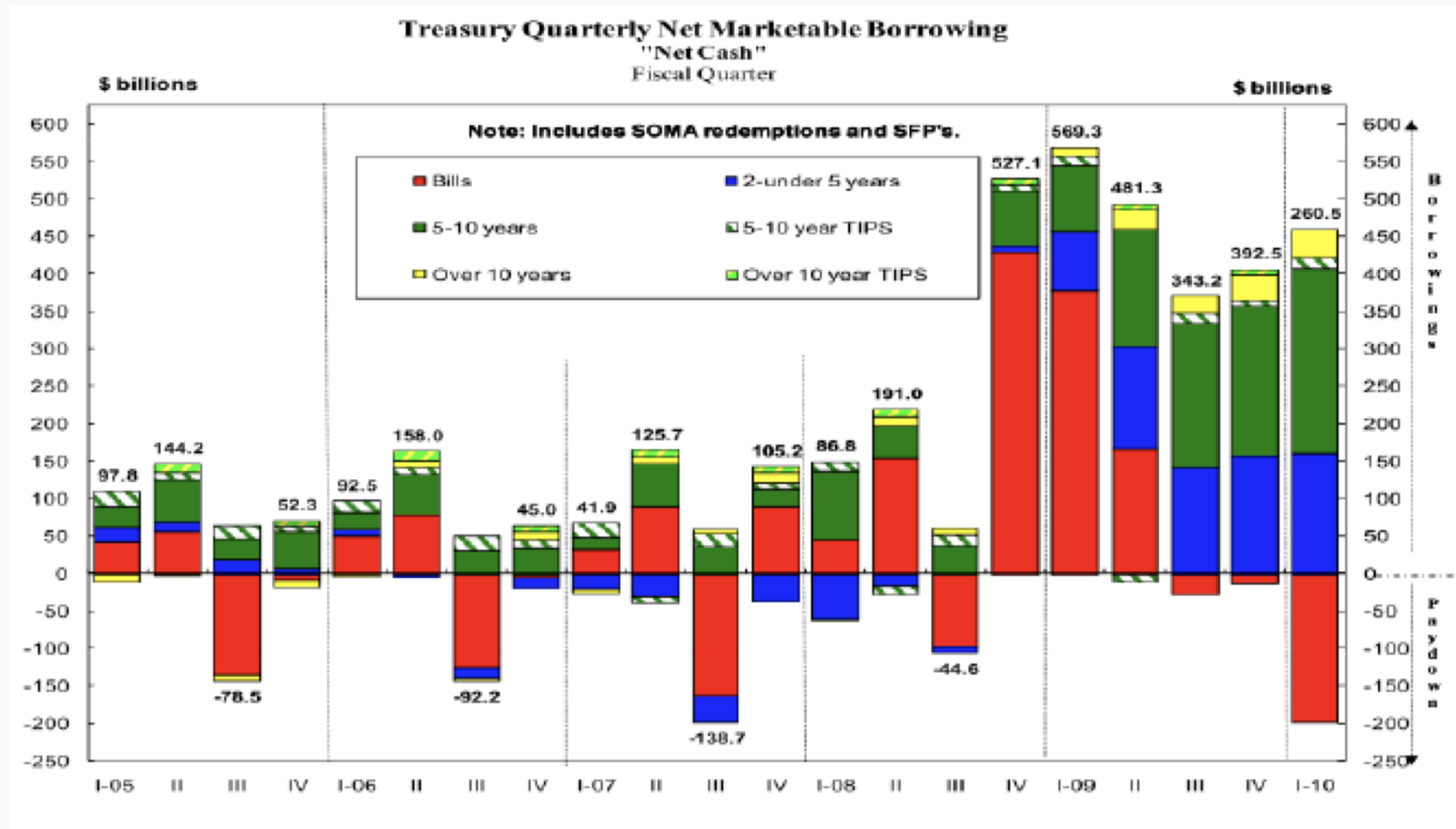


# Explore





# Not Effective



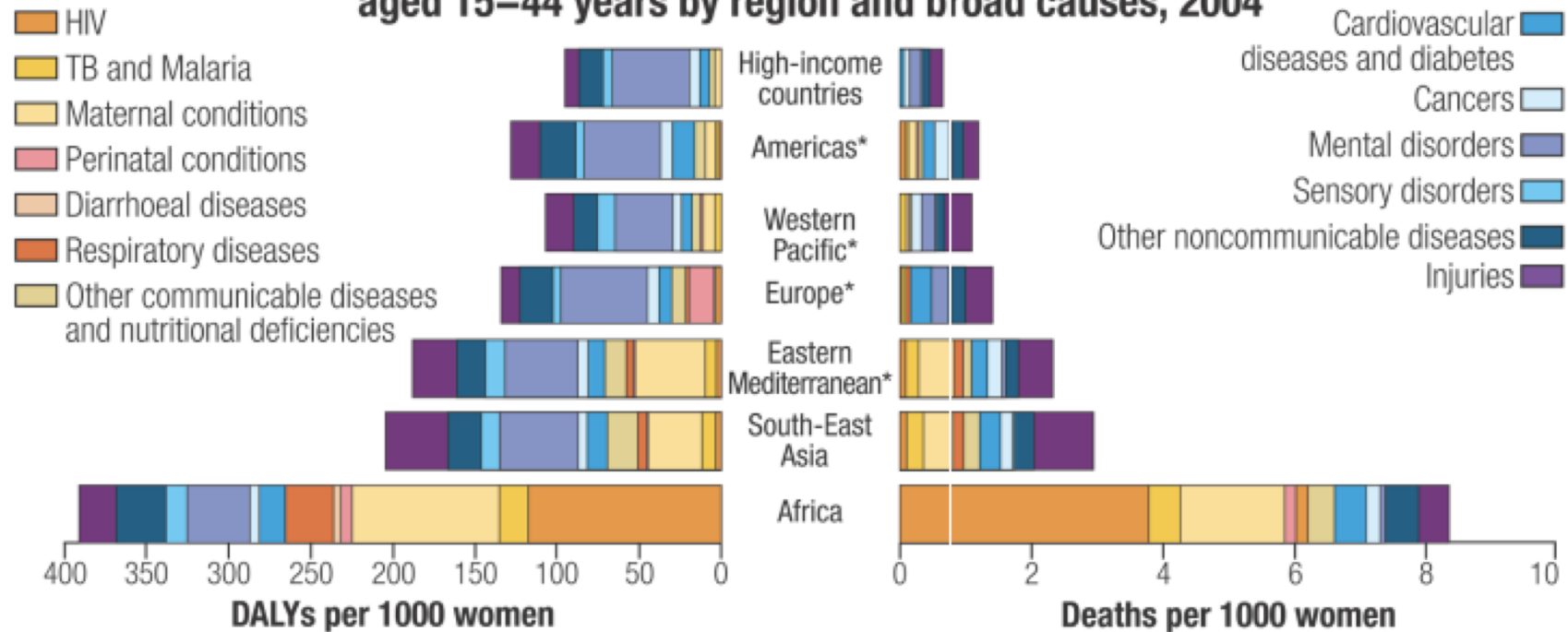
# Not Effective



Figure 10

# Not Effective

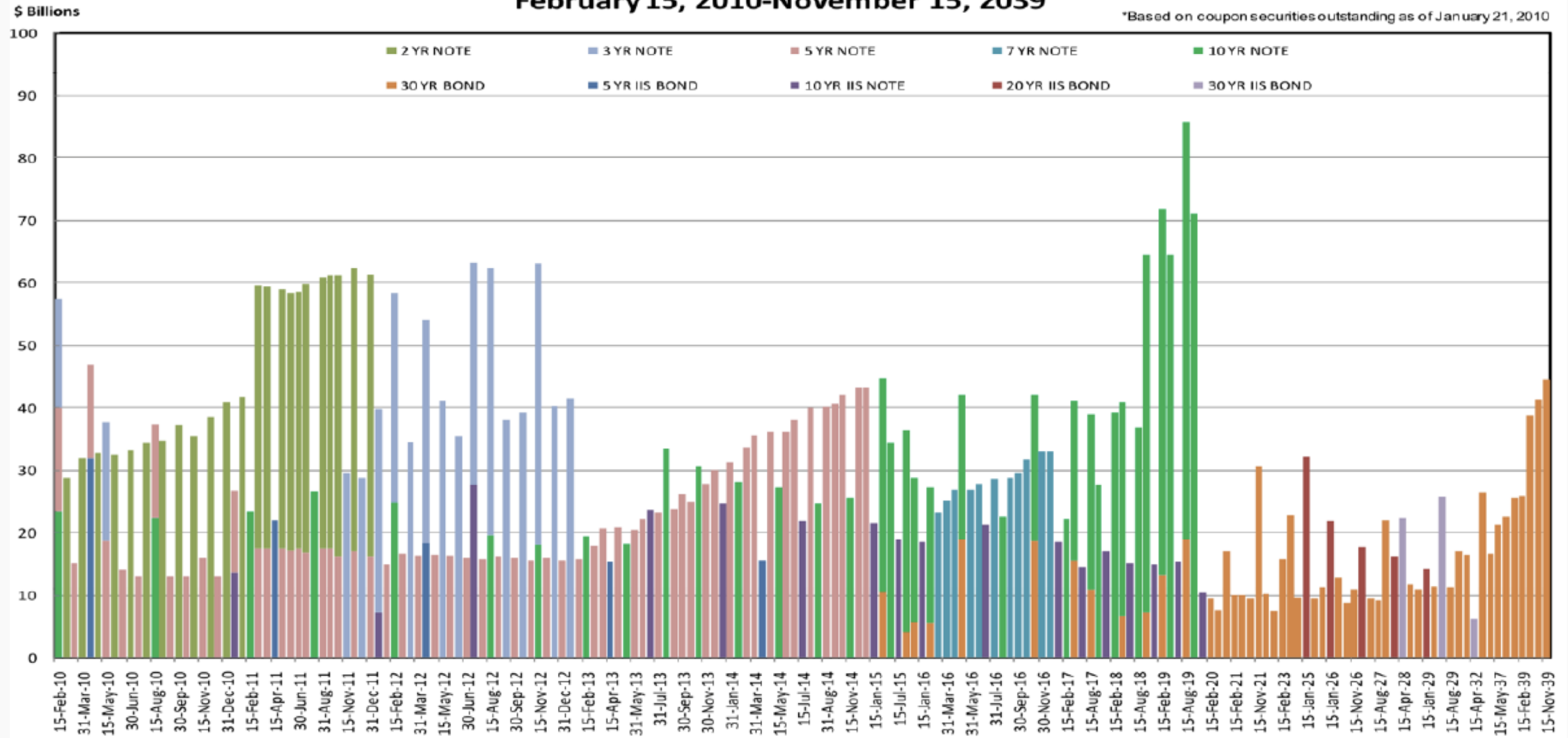
Figure 1 Mortality and disease burden (DALYs) in women aged 15–44 years by region and broad causes, 2004



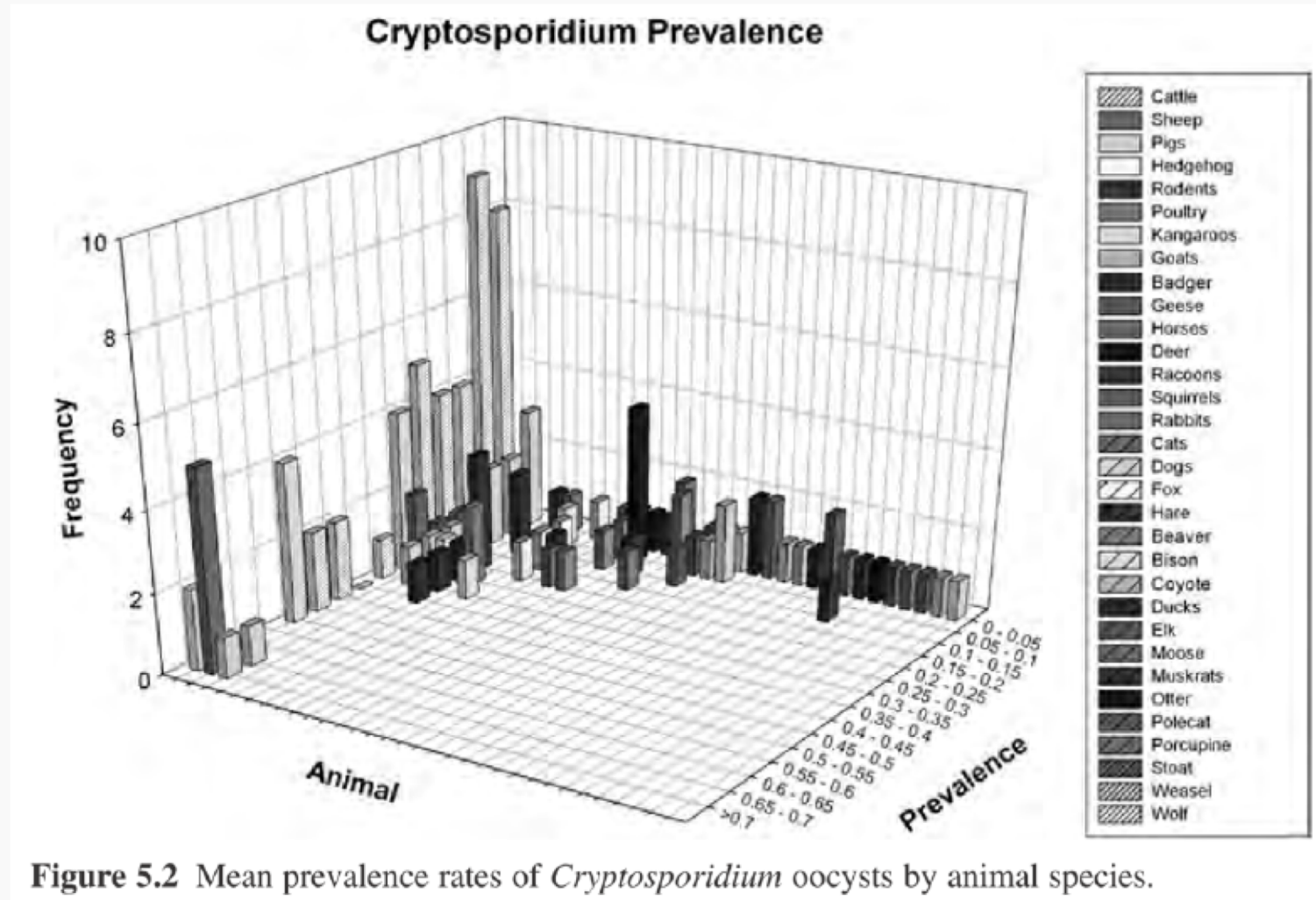
\* High-income countries are excluded from the regional groups.  
Source: World Health Organization.<sup>1</sup>

# Not Effective

## Coupons Maturing\* February 15, 2010-November 15, 2039



# Not Effective



# 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?**