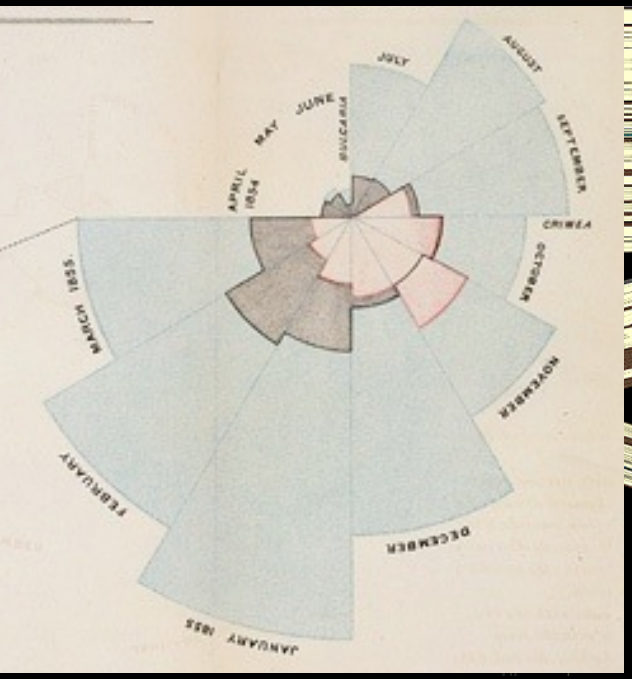
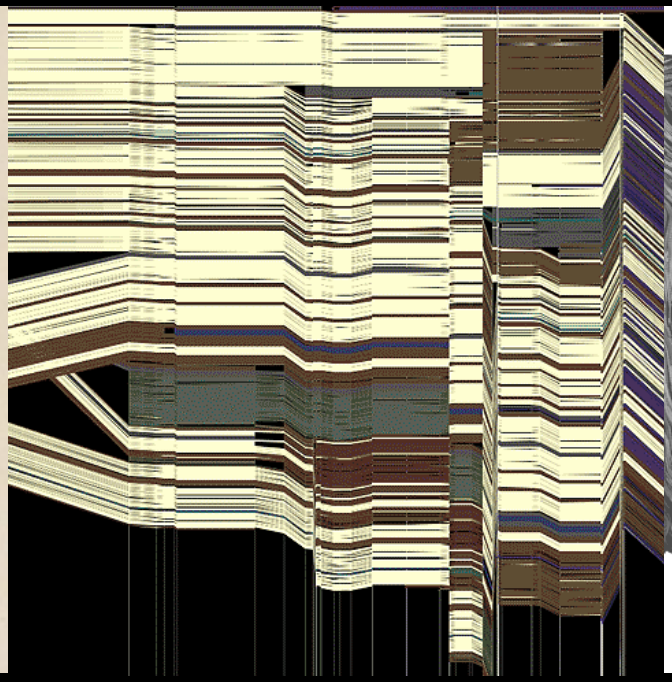
CSE 412 - Intro to Data Visualization

Exploratory Data Analysis



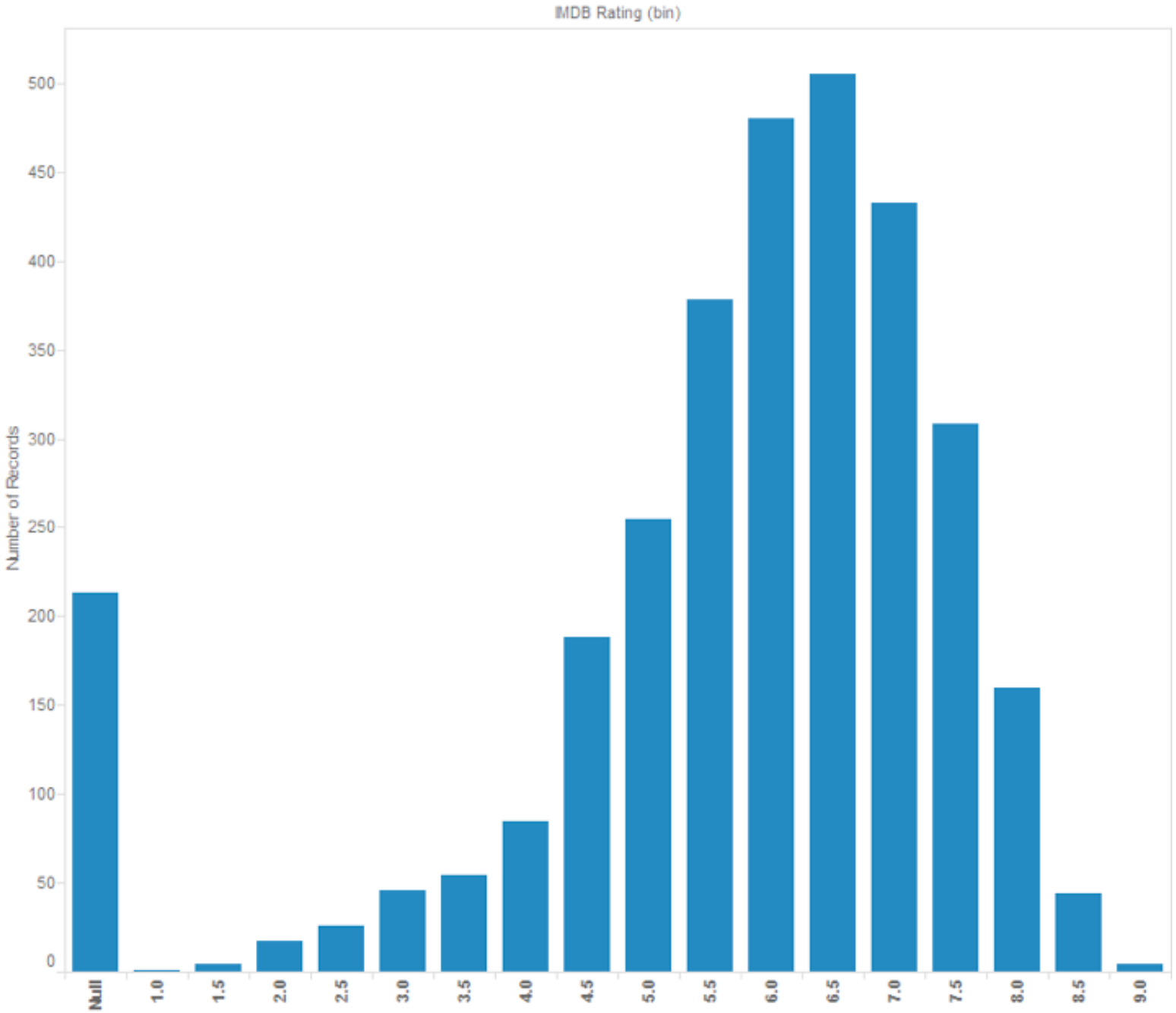


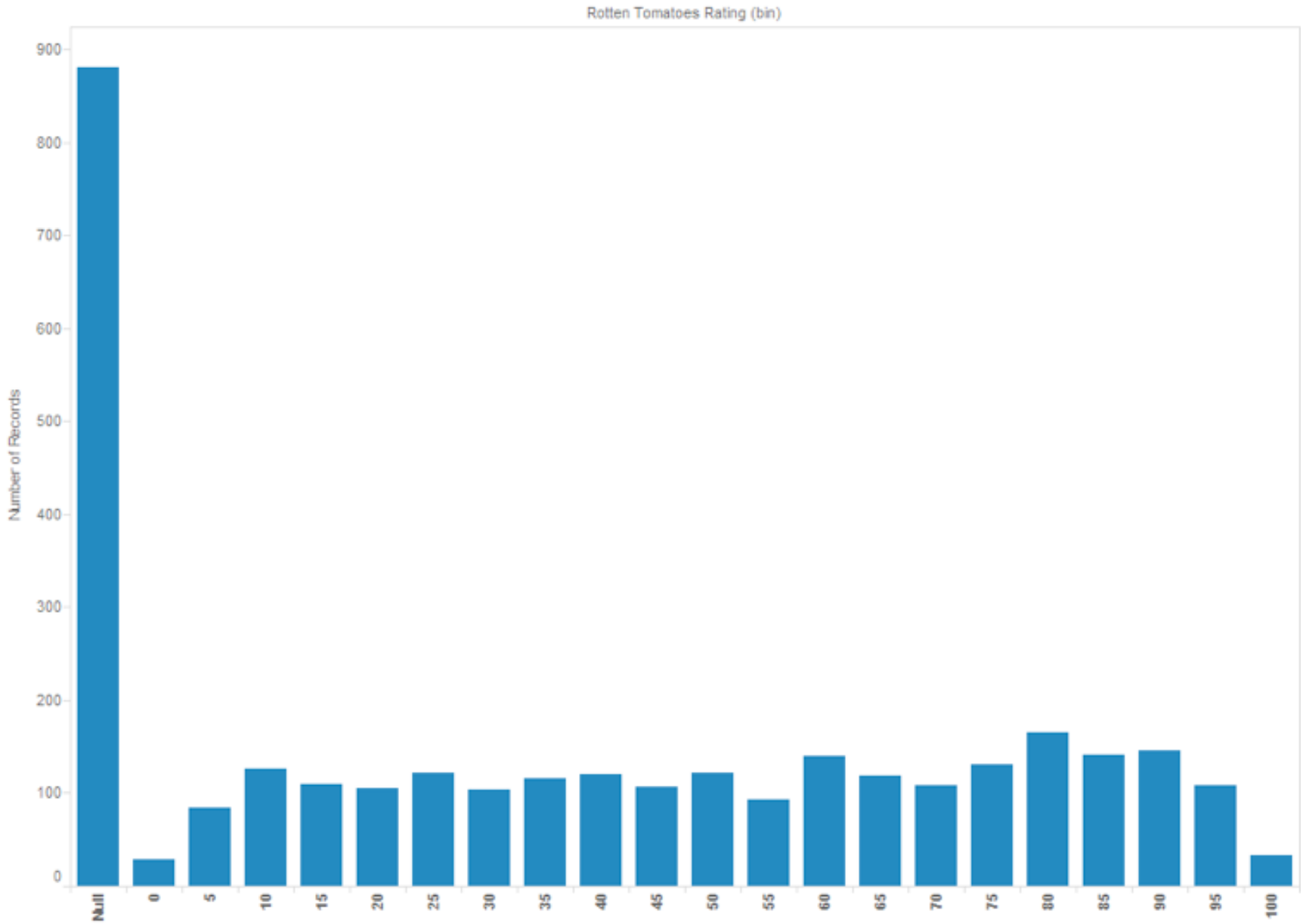
Jane Hoffswell University of Washington

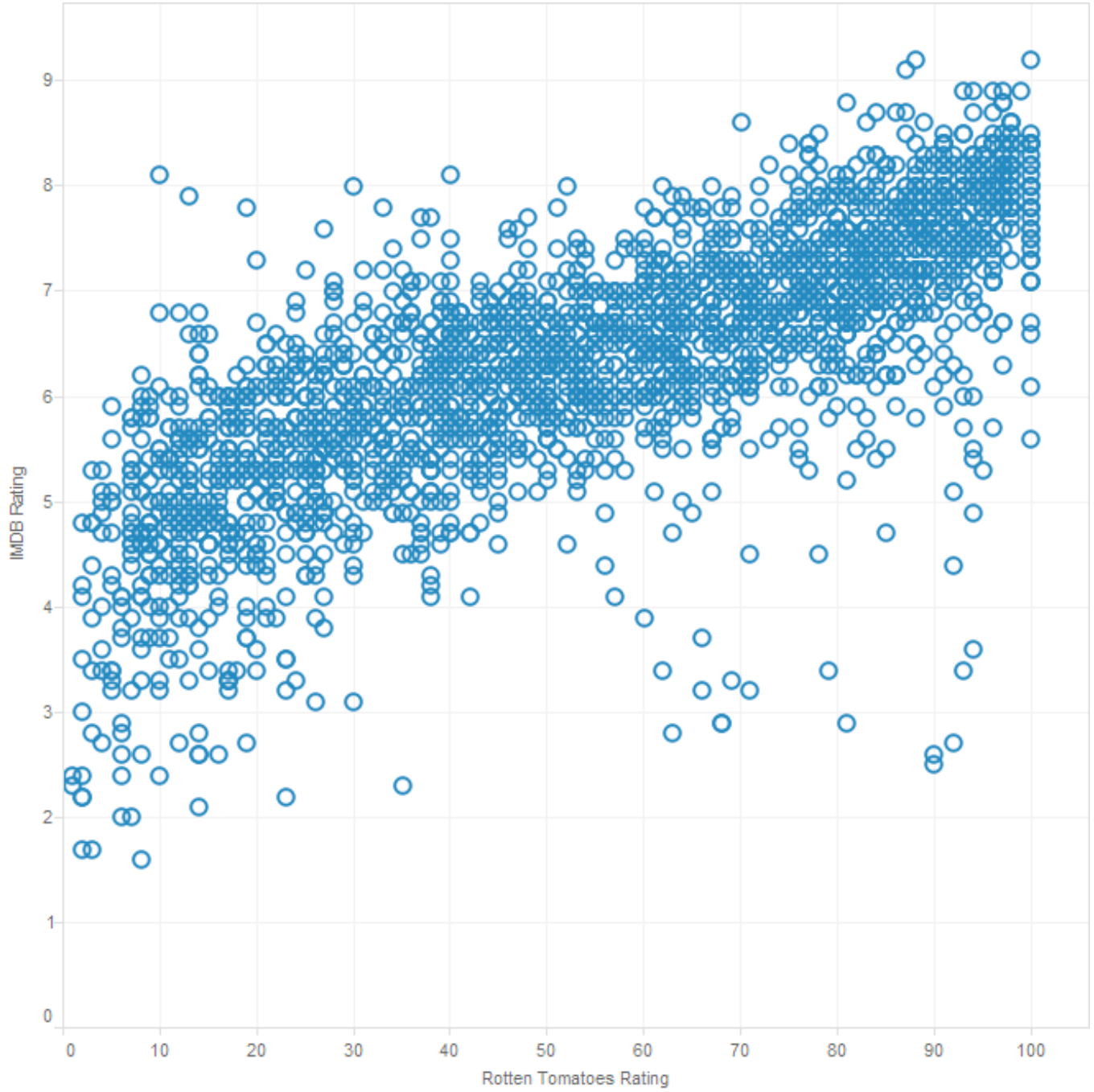
Analysis Example: Motion Pictures Data

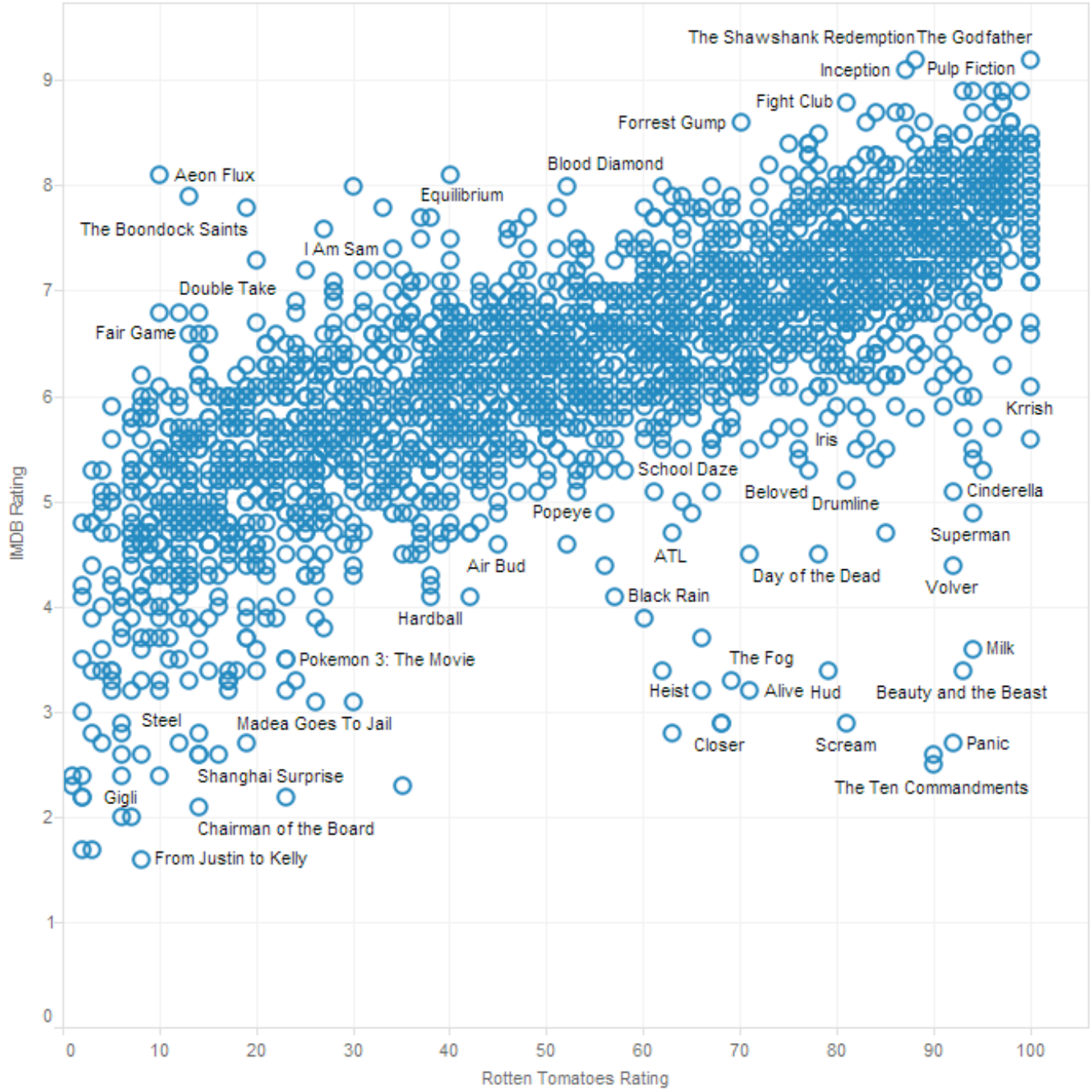
Motion Pictures Data

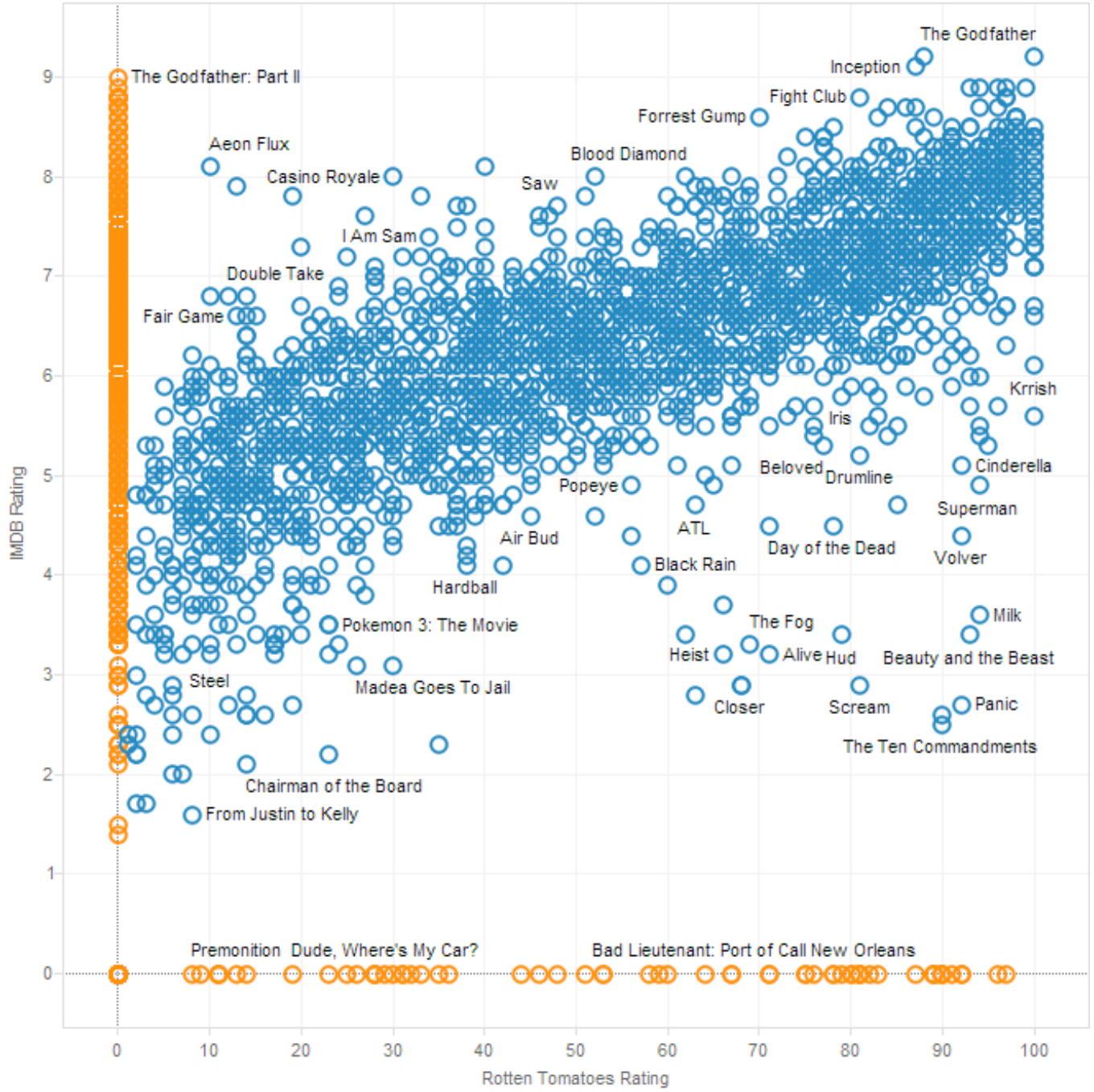
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)

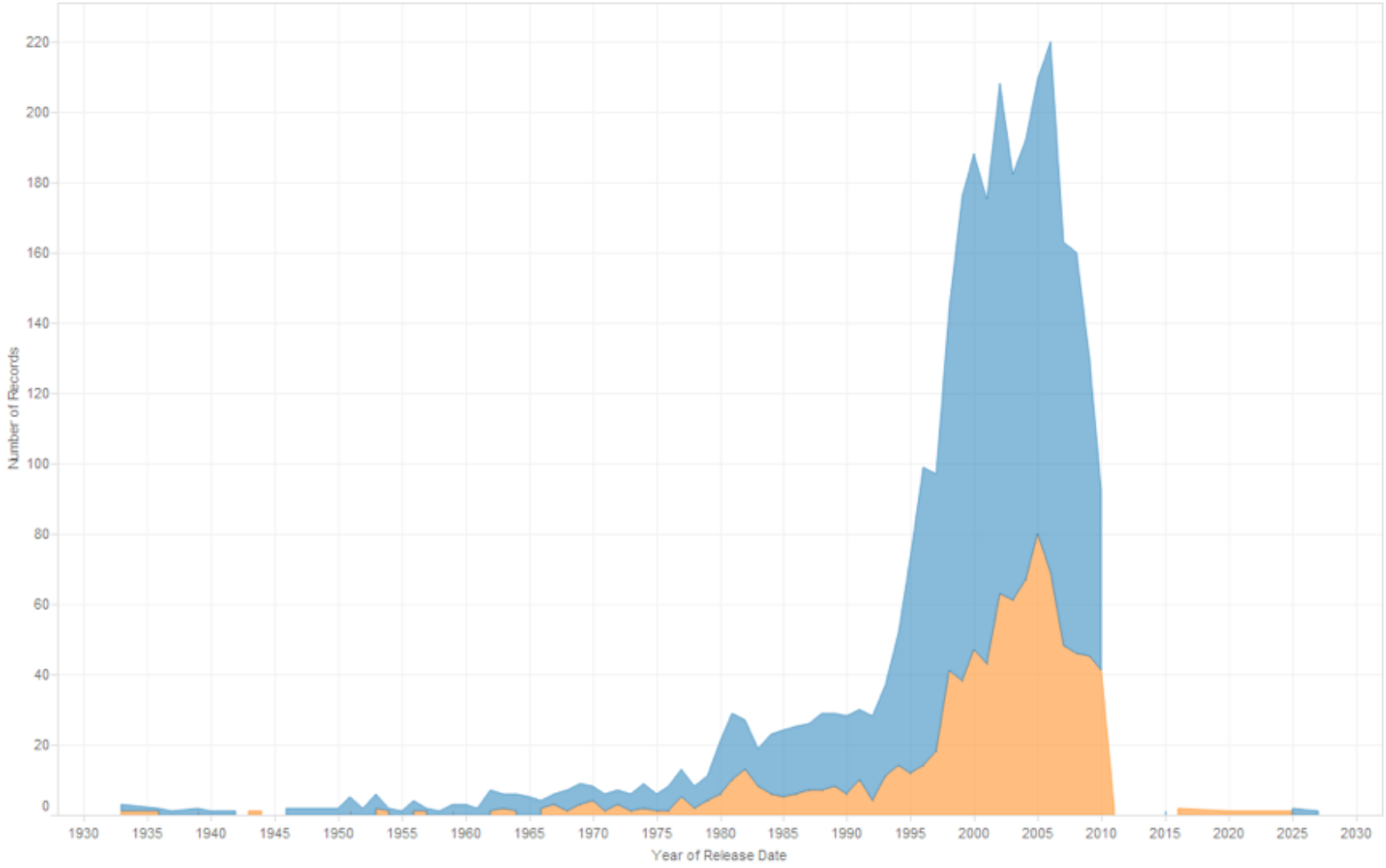

IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0

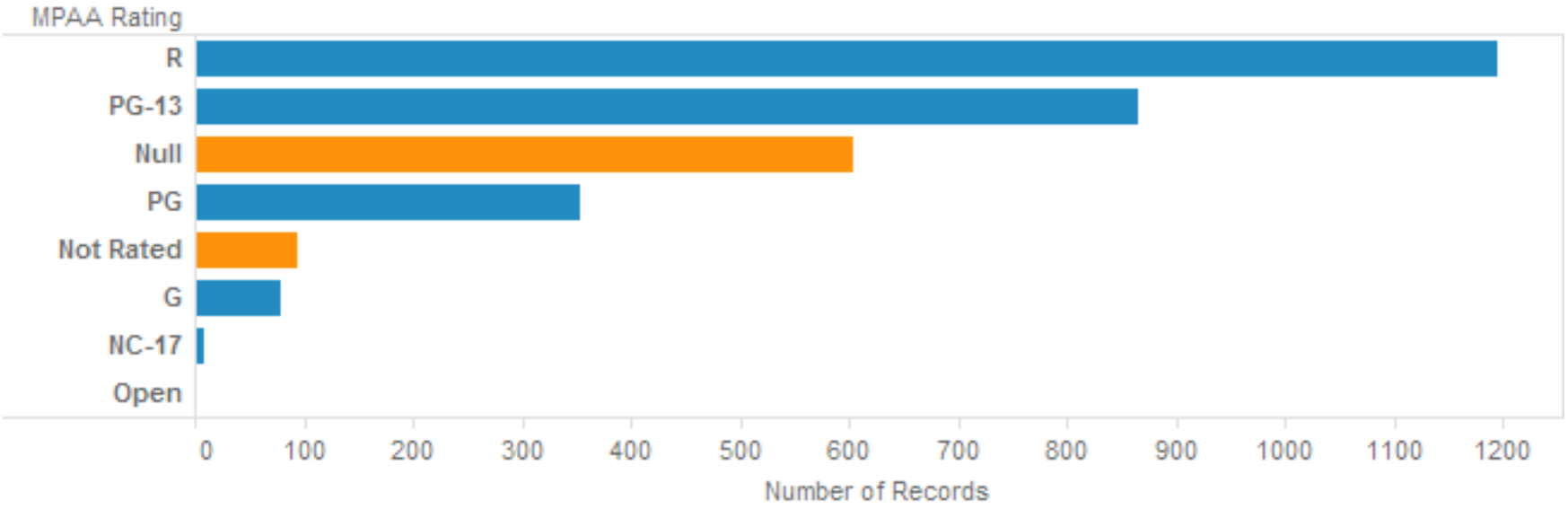

Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

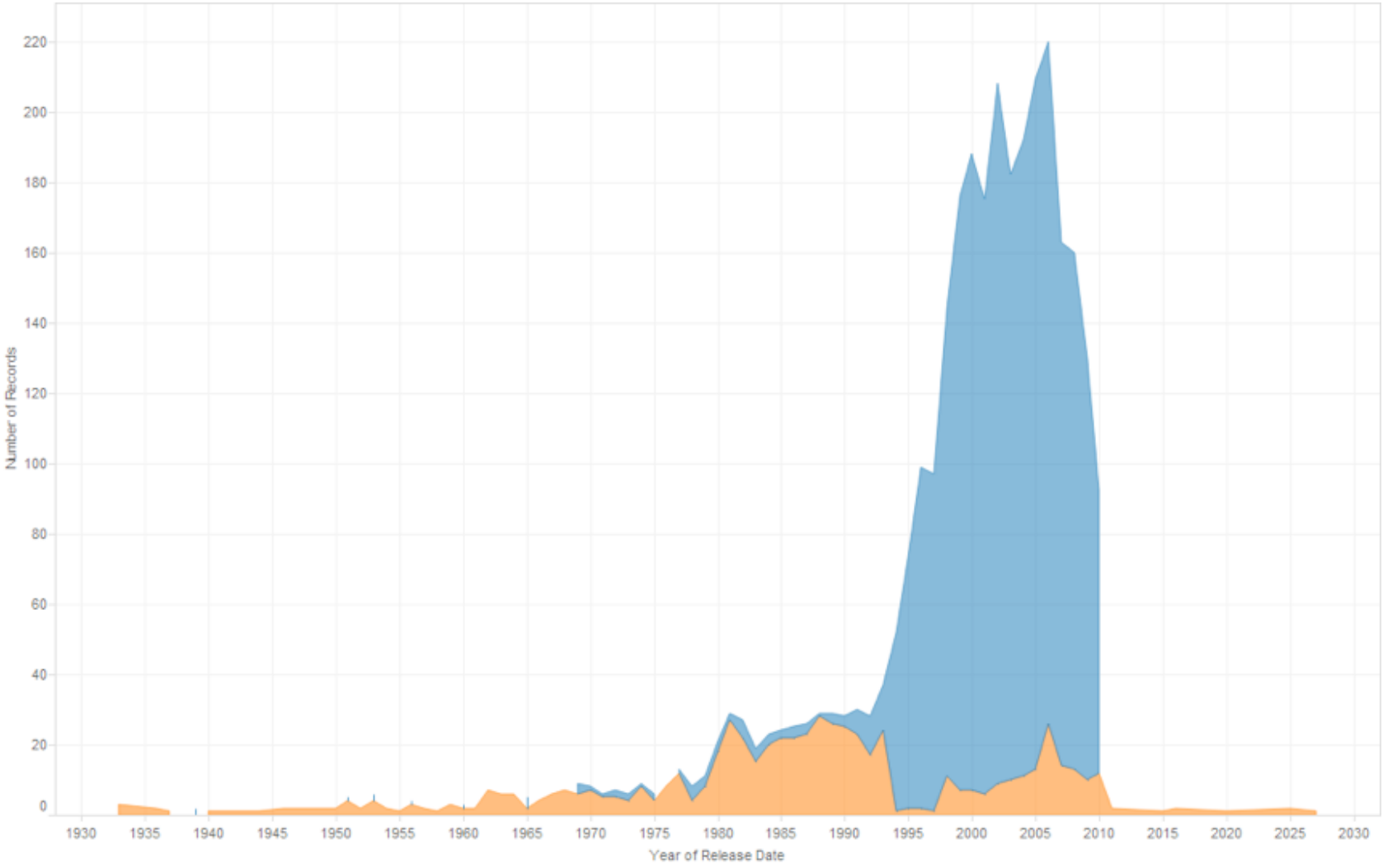

IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100


The Shawshank Redemption
The Godfather
Inception
Pulp Fiction
Fight Club
Forrest Gump
Aeon Flux
Blood Diamond
Equilibrium
The Boondock Saints
I Am Sam
Double Take
Fair Game
Krrish
Iris
School Daze
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Milk
Pokemon 3: The Movie
The Fog
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
Shanghai Surprise
The Ten Commandments
Gigli
Chairman of the Board
From Justin to Kelly
IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100


The Godfather
Inception
Fight Club
Forrest Gump
Blood Diamond
Saw
Aeon Flux
Casino Royale
I Am Sam
Double Take
Fair Game
The Godfather: Part II
Krrish
Iris
Beloved
Drumline
Cinderella
Superman
Volver
Popeye
ATL
Day of the Dead
Air Bud
Black Rain
Hardball
Pokemon 3: The Movie
The Fog
Milk
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
The Ten Commandments
Chairman of the Board
From Justin to Kelly
Premonition
Dude, Where's My Car?
Bad Lieutenant: Port of Call New Orleans
IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100


Number of Records
0
20
40
60
80
100
120
140
160
180
200
220
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date


MPAA Rating
R
PG-13
Null
PG
Not Rated
G
NC-17
Open
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Number of Records


Number of Records
220
200
180
160
140
120
100
80
60
40
20
0
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date

Lesson: Exercise Skepticism

Check **data quality** and your **assumptions**.

Start with **univariate summaries**, then start to consider **relationships among variables**.

Avoid premature fixation!

Analysis Example: Antibiotic Effectiveness

Data Set: Antibiotic Effectiveness

Genus of Bacteria	String (N)
Species of Bacteria	String (N)
Antibiotic Applied	String (N)
Gram-Staining?	Pos / Neg (N)
Min. Inhibitory Concent. (g)	Number (Q)

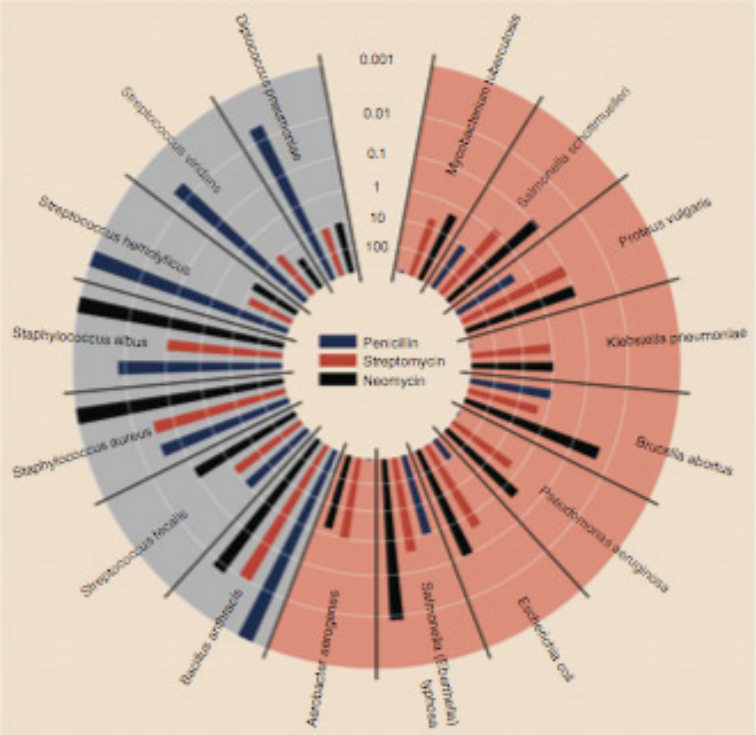
Collected prior to 1951.

What questions might we ask?

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

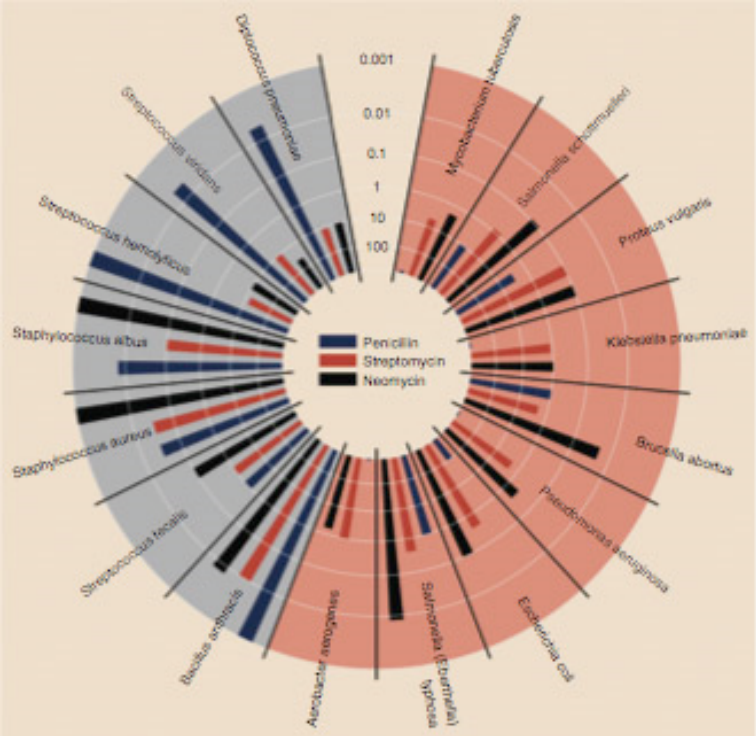
How do the drugs compare?



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

Original graphic by Will Burtin, 1951

How do the drugs compare?



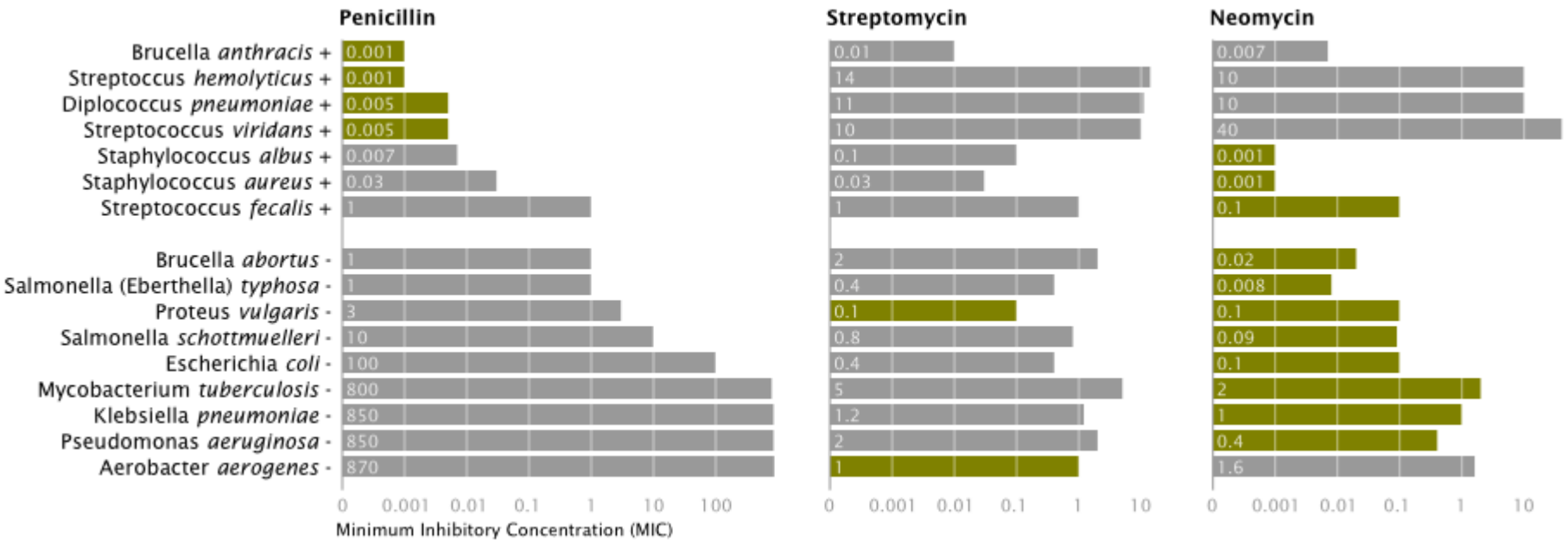
Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–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

Radius: 1 / log(MIC)

Bar Color: Antibiotic

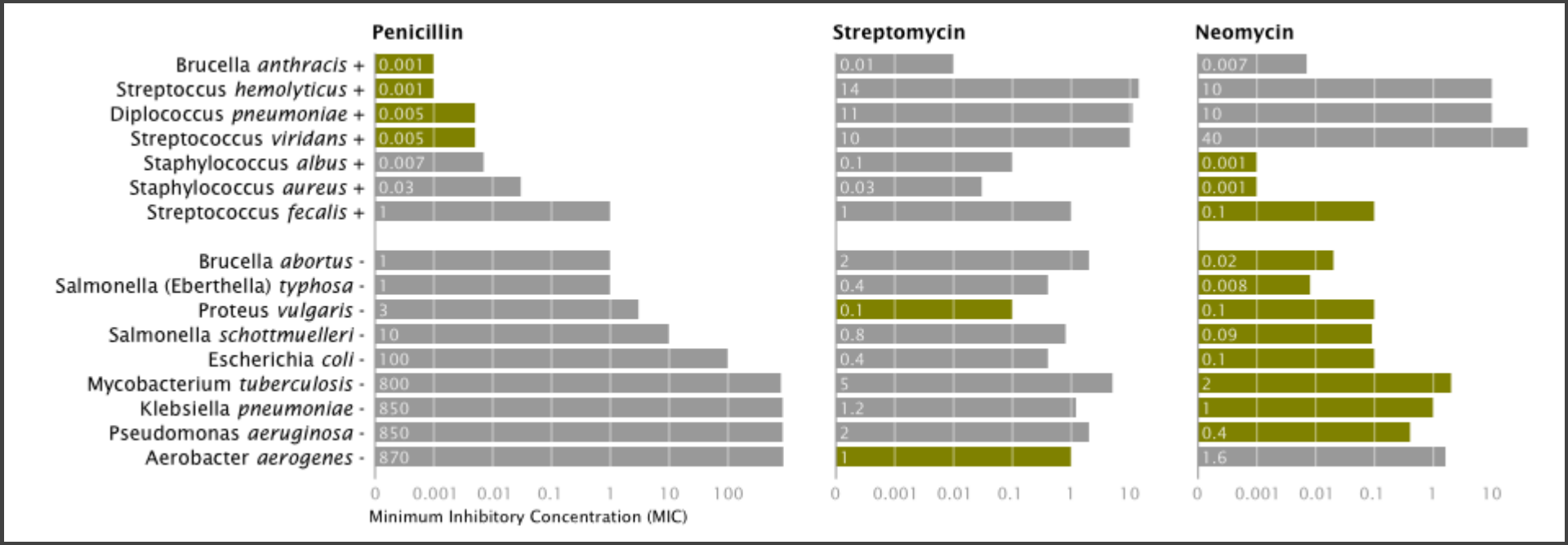
Background Color: Gram Staining

How do the drugs compare?



Mike Bostock
Stanford CS448B, Winter 2009

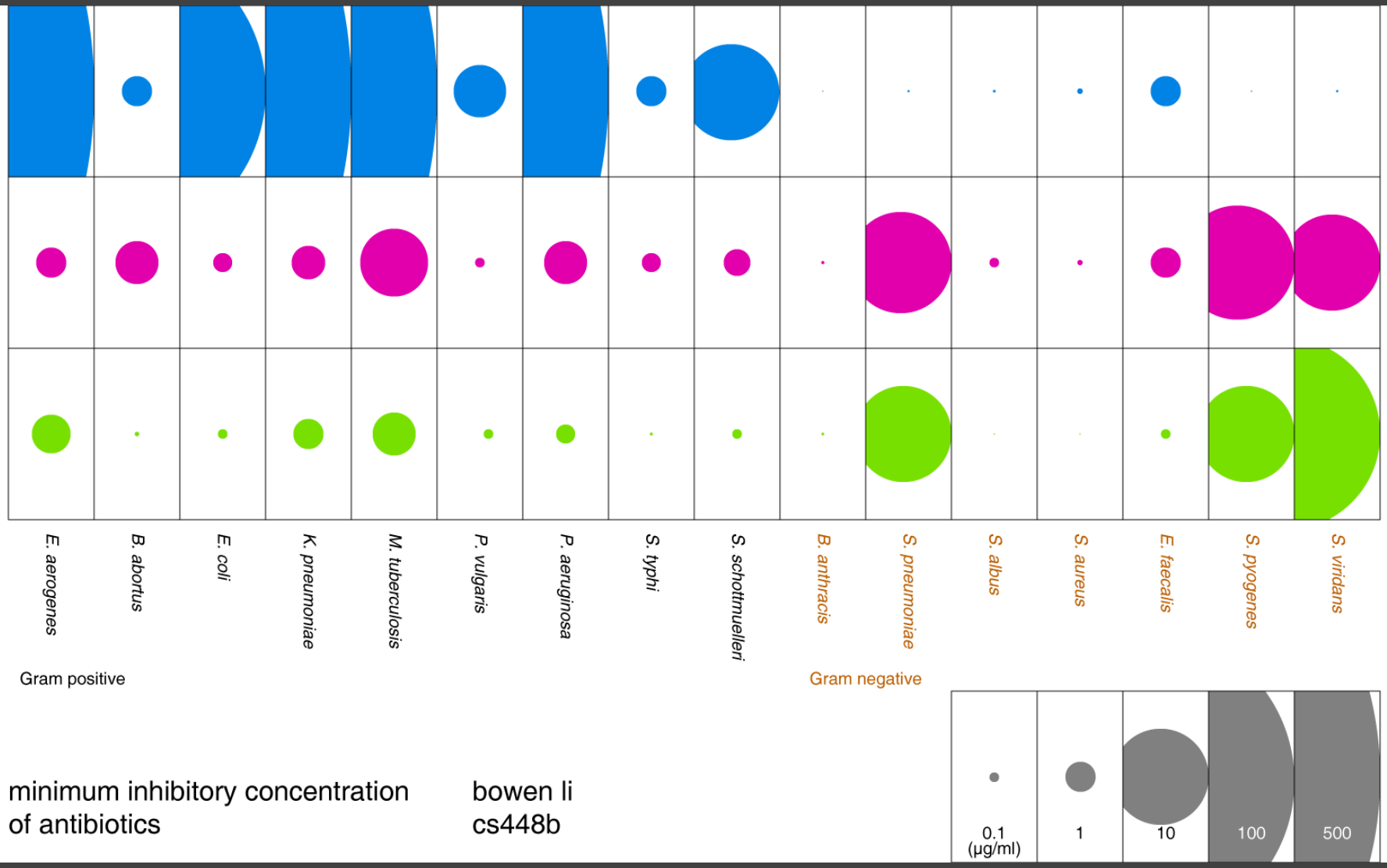
How do the drugs compare?



X-axis: Antibiotic | log(MIC)

Y-axis: Gram-Staining | Species

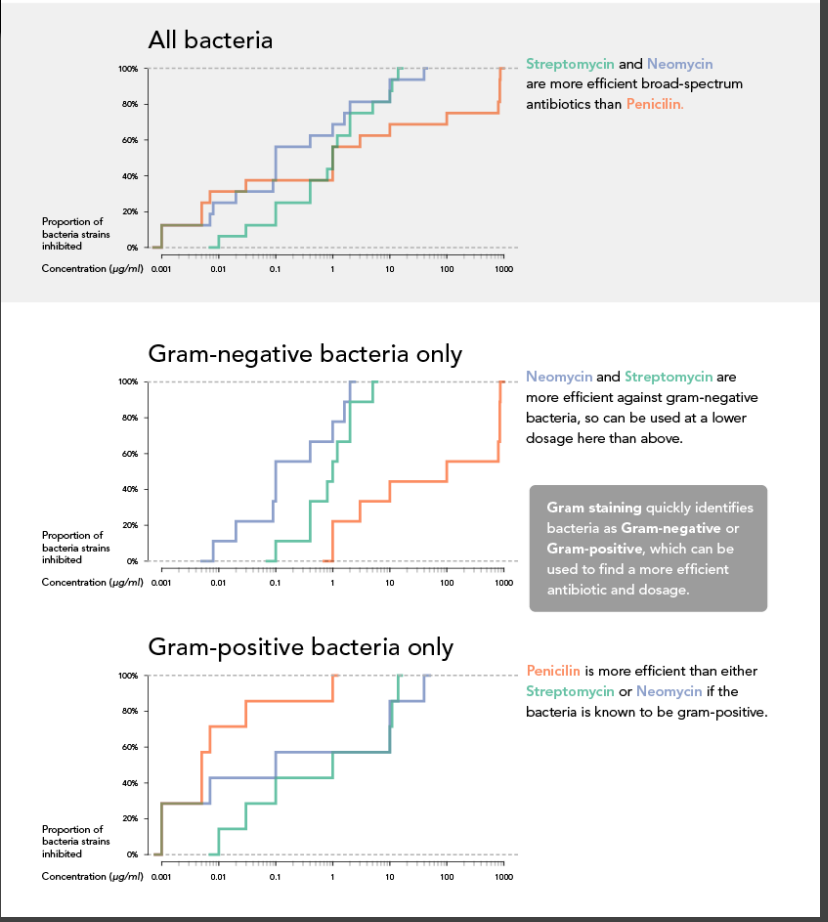
Color: Most-Effective?

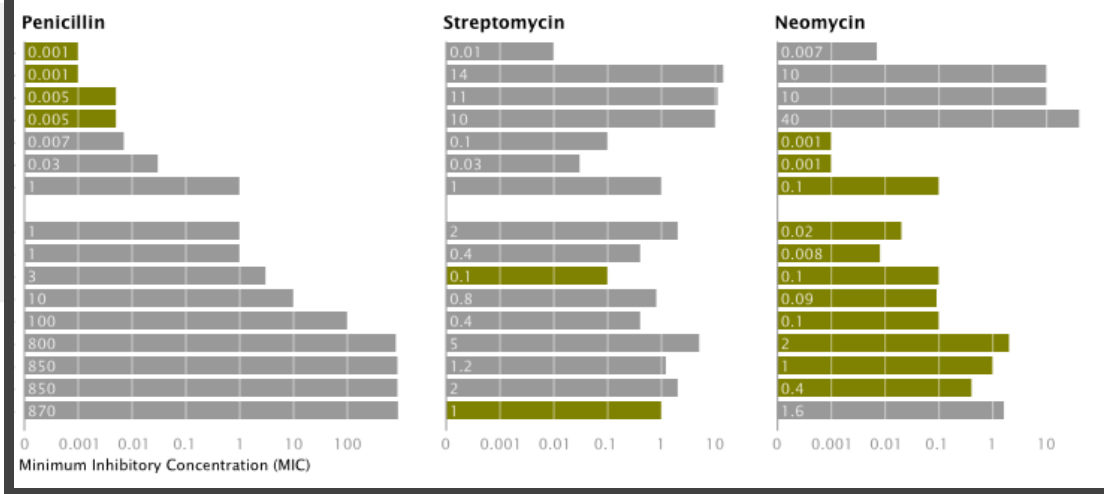


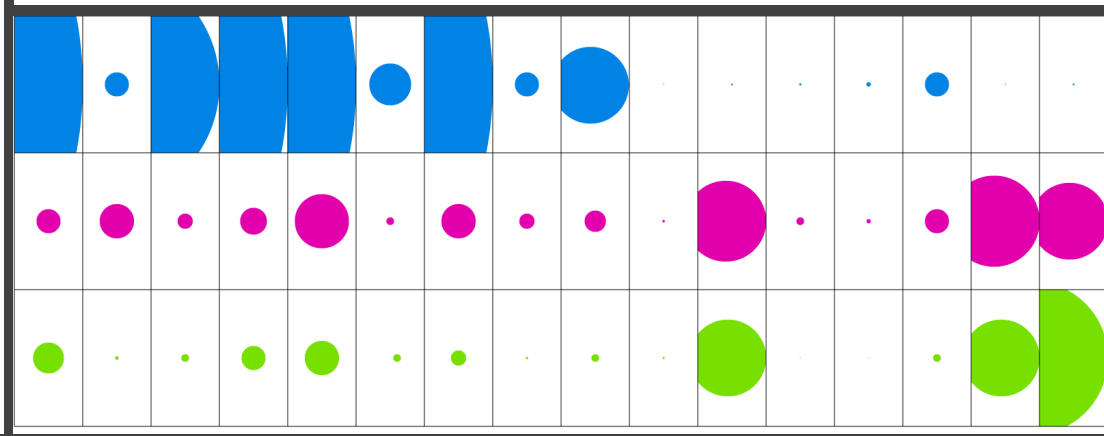
minimum inhibitory concentration of antibiotics

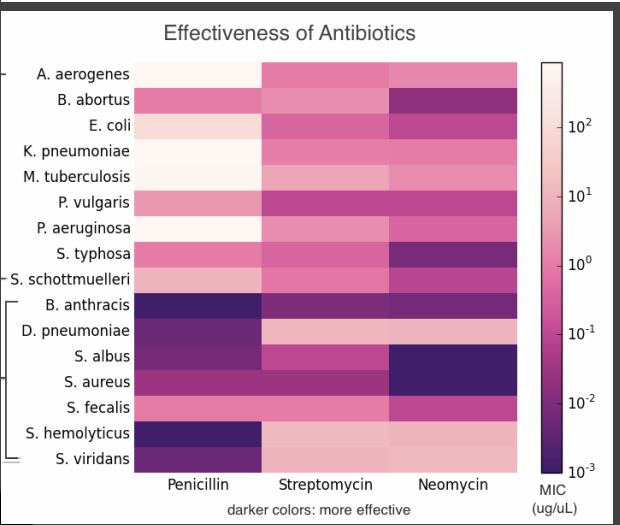
bowen li
cs448b

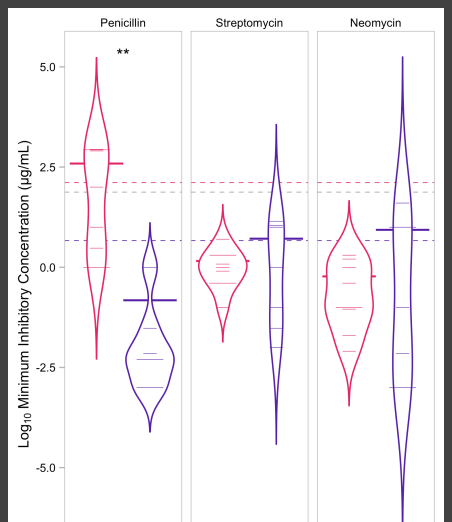
Bowen Li
Stanford CS448B, Fall 2009

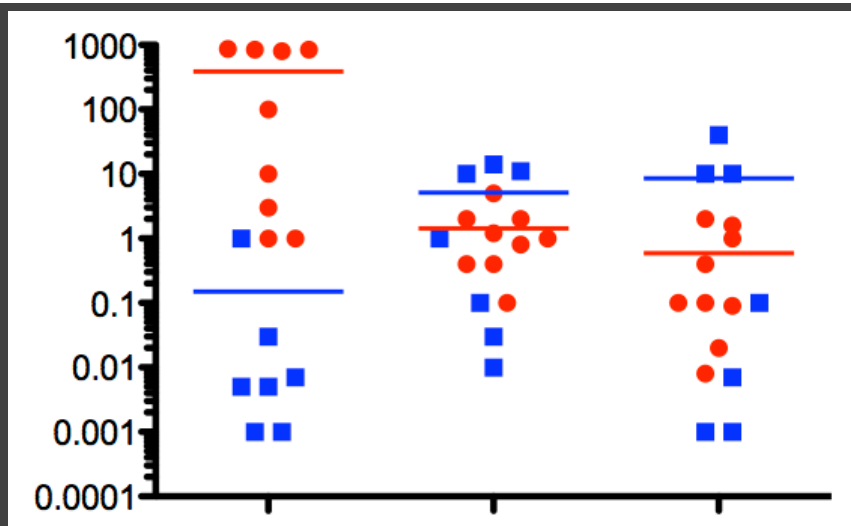

All bacteria
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.
Proportion of bacteria strains inhibited
Concentration (μg/ml)
Gram-negative bacteria only
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.
Gram staining quickly identifies bacteria as Gram-negative or Gram-positive, which can be used to find a more efficient antibiotic and dosage.
Gram-positive bacteria only
Penicilin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.

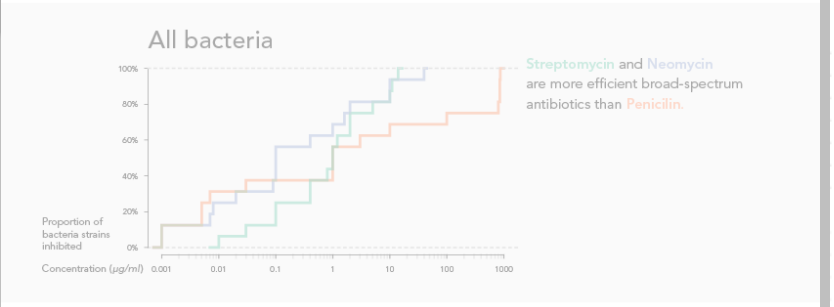

Penicillin
Streptomycin
Neomycin
Minimum Inhibitory Concentration (MIC)

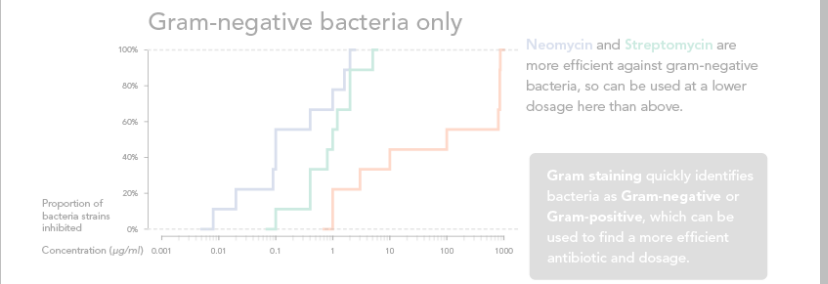


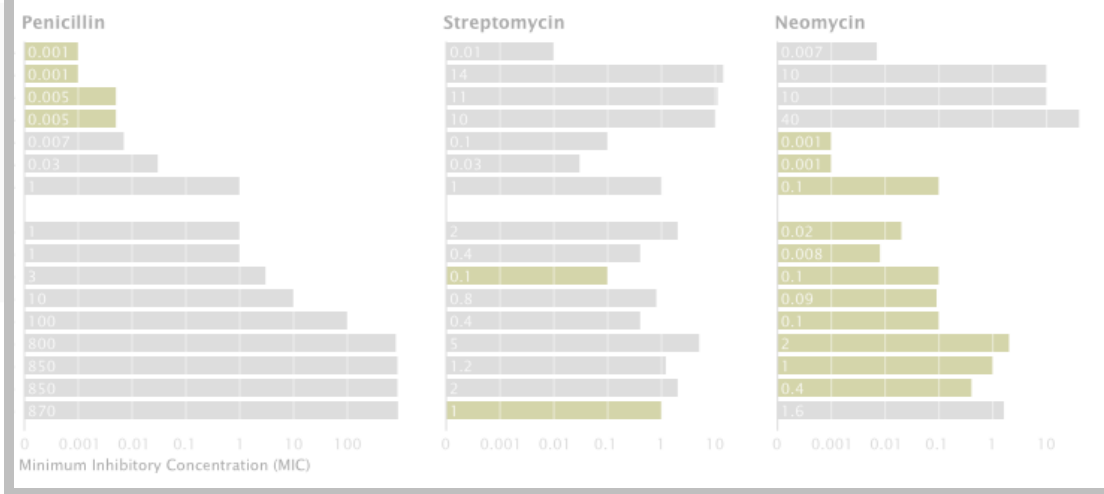

Effectiveness of Antibiotics
A. aerogenes
B. abortus
E. coli
K. pneumoniae
M. tuberculosis
P. vulgaris
P. aeruginosa
S. typhosa
S. schottmuelleri
B. anthracis
D. pneumoniae
S. albus
S. aureus
S. fecalis
S. hemolyticus
S. viridans
Penicillin
Streptomycin
Neomycin
darker colors: more effective
MIC (ug/uL)

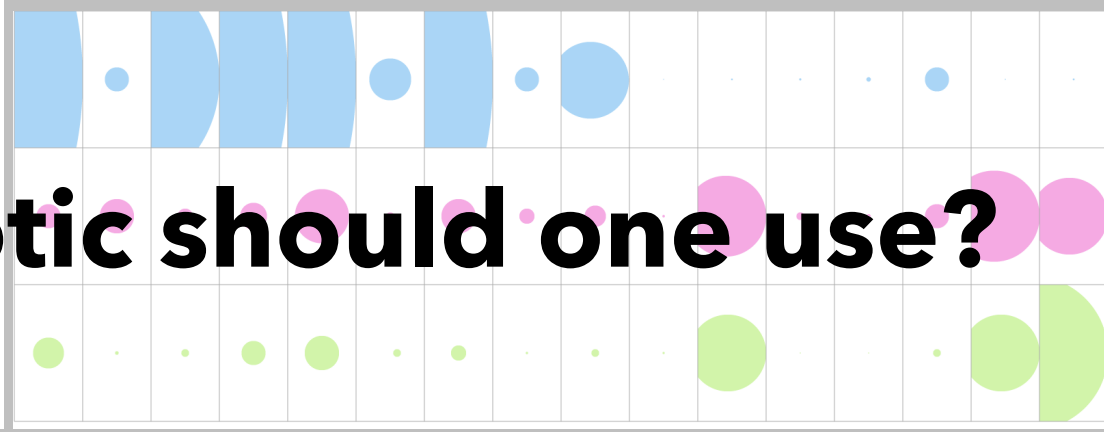

Penicillin
Streptomycin
Neomycin
**
Log10 Minimum Inhibitory Concentration (μg/mL)


1000
100
10
1
0.1
0.01
0.001
0.0001

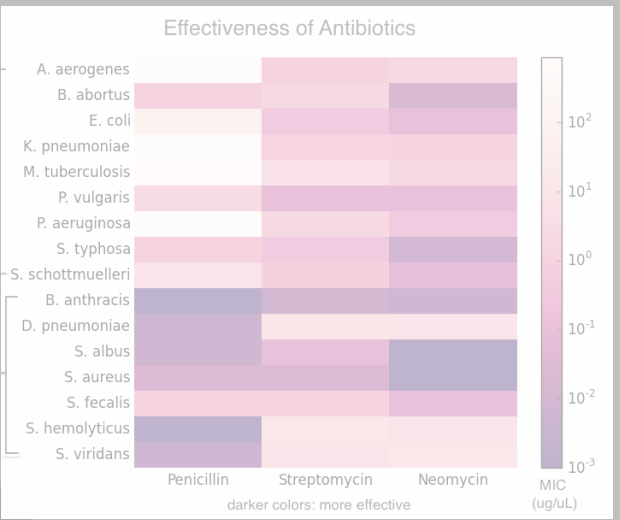


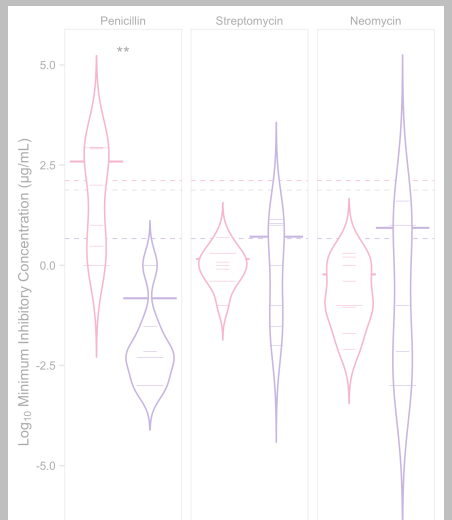


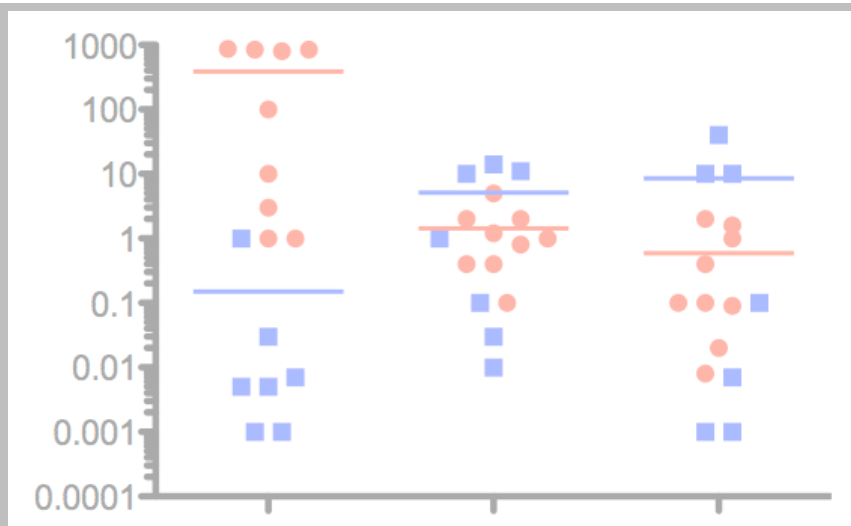




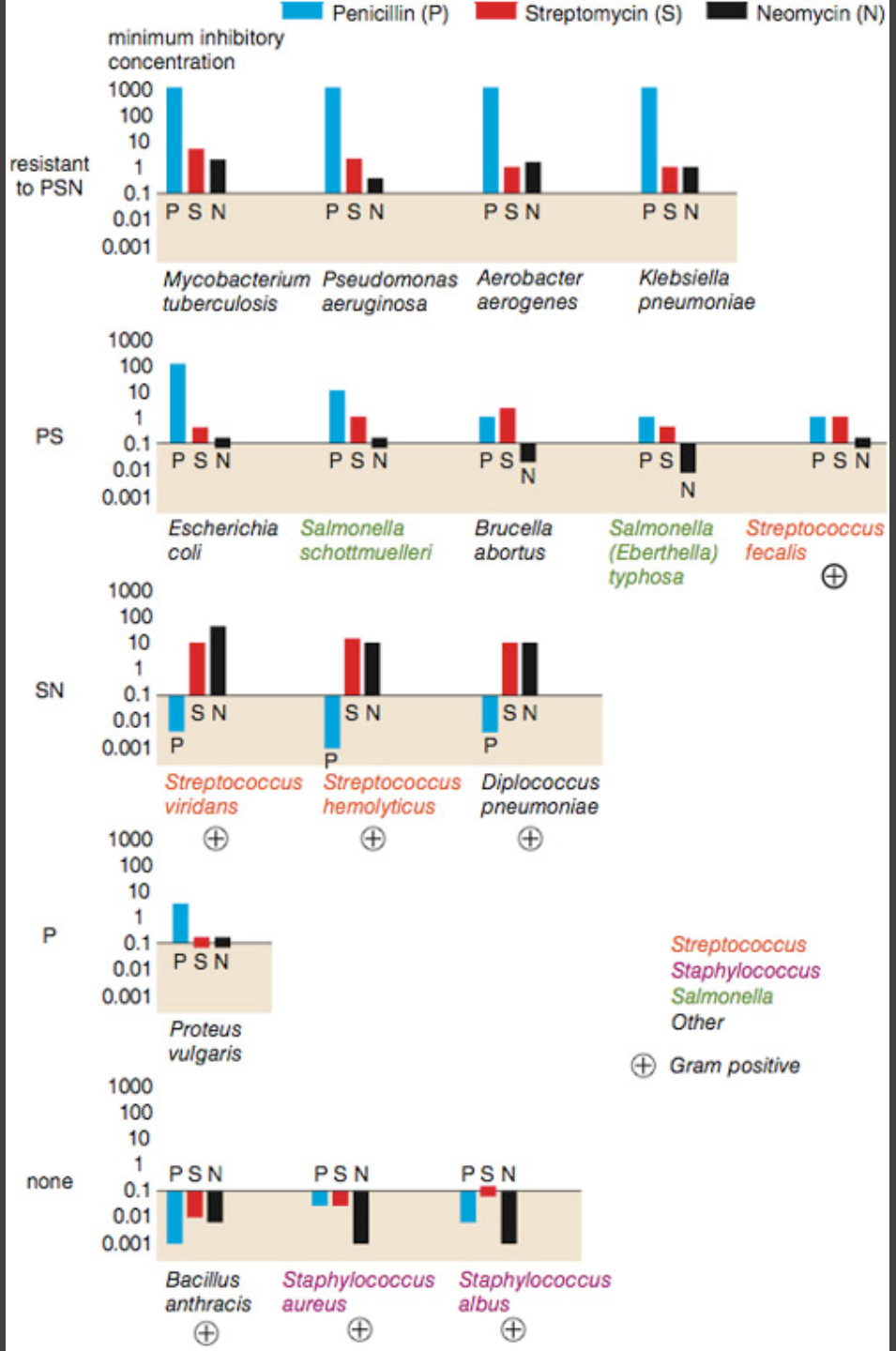
Which antibiotic should one use?







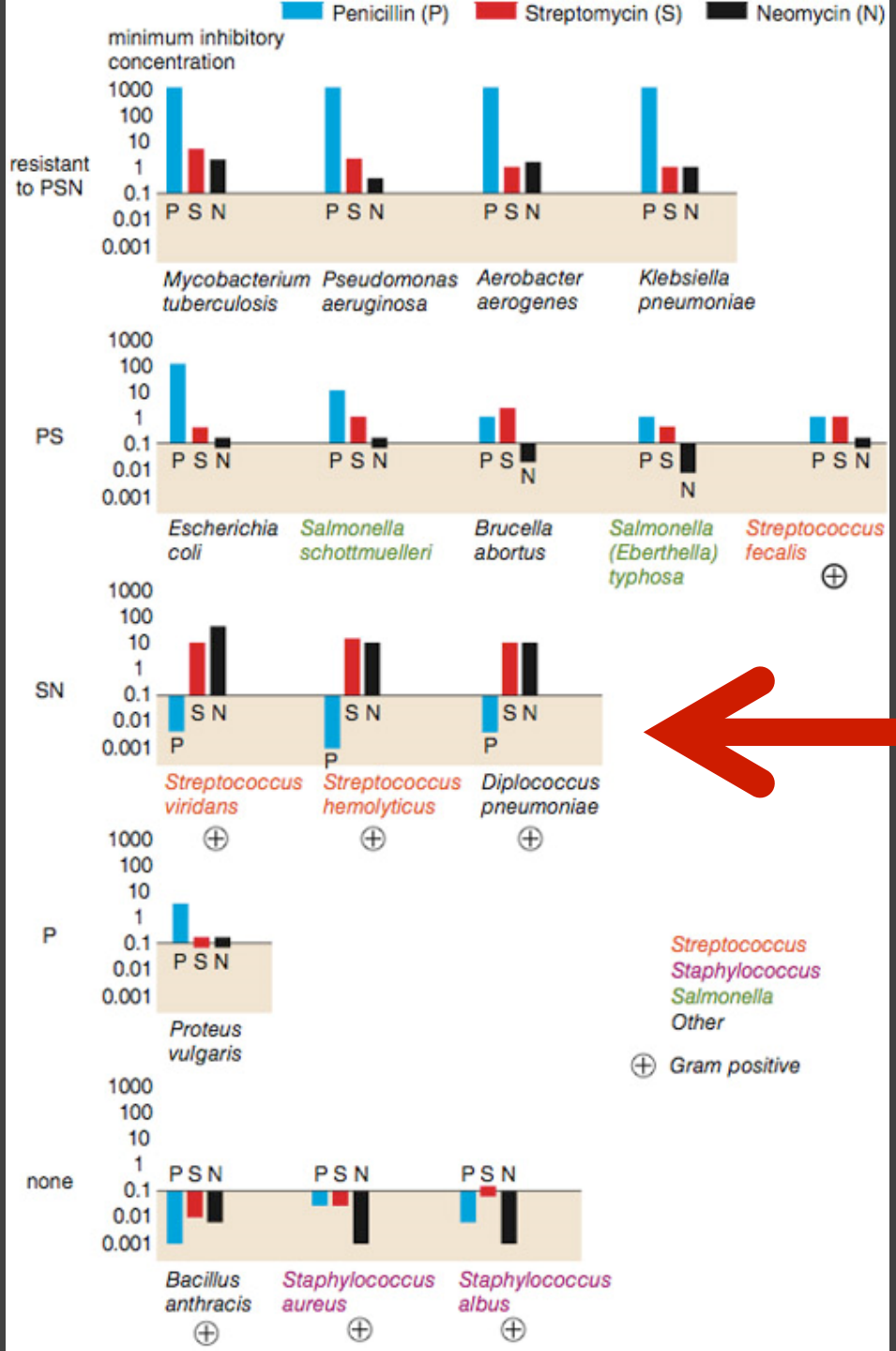
Do the bacteria group by antibiotic resistance?



Do the bacteria group by antibiotic resistance?

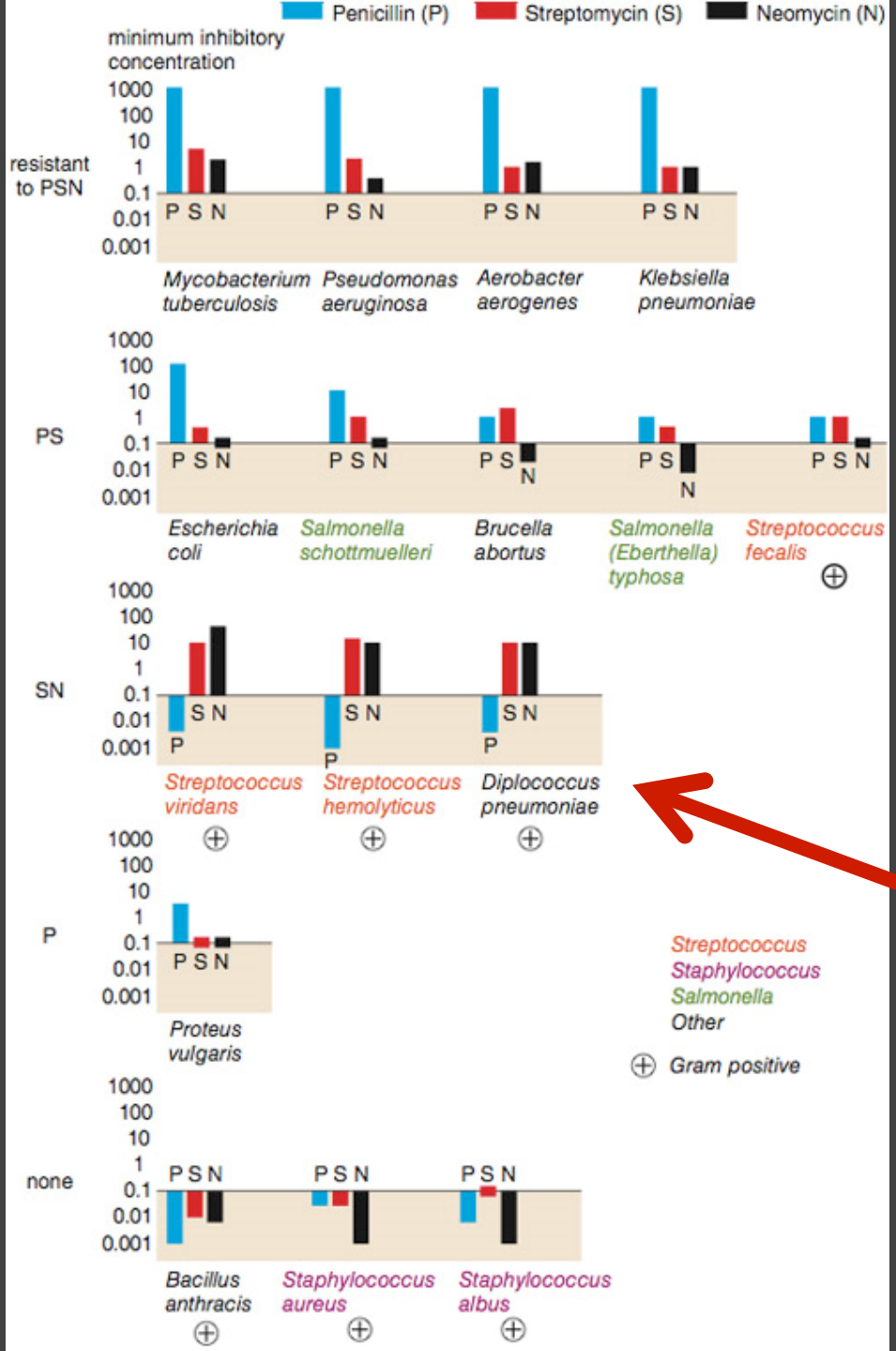
Wainer & Lysen

American Scientist, 2009



Do the bacteria group by antibiotic resistance?

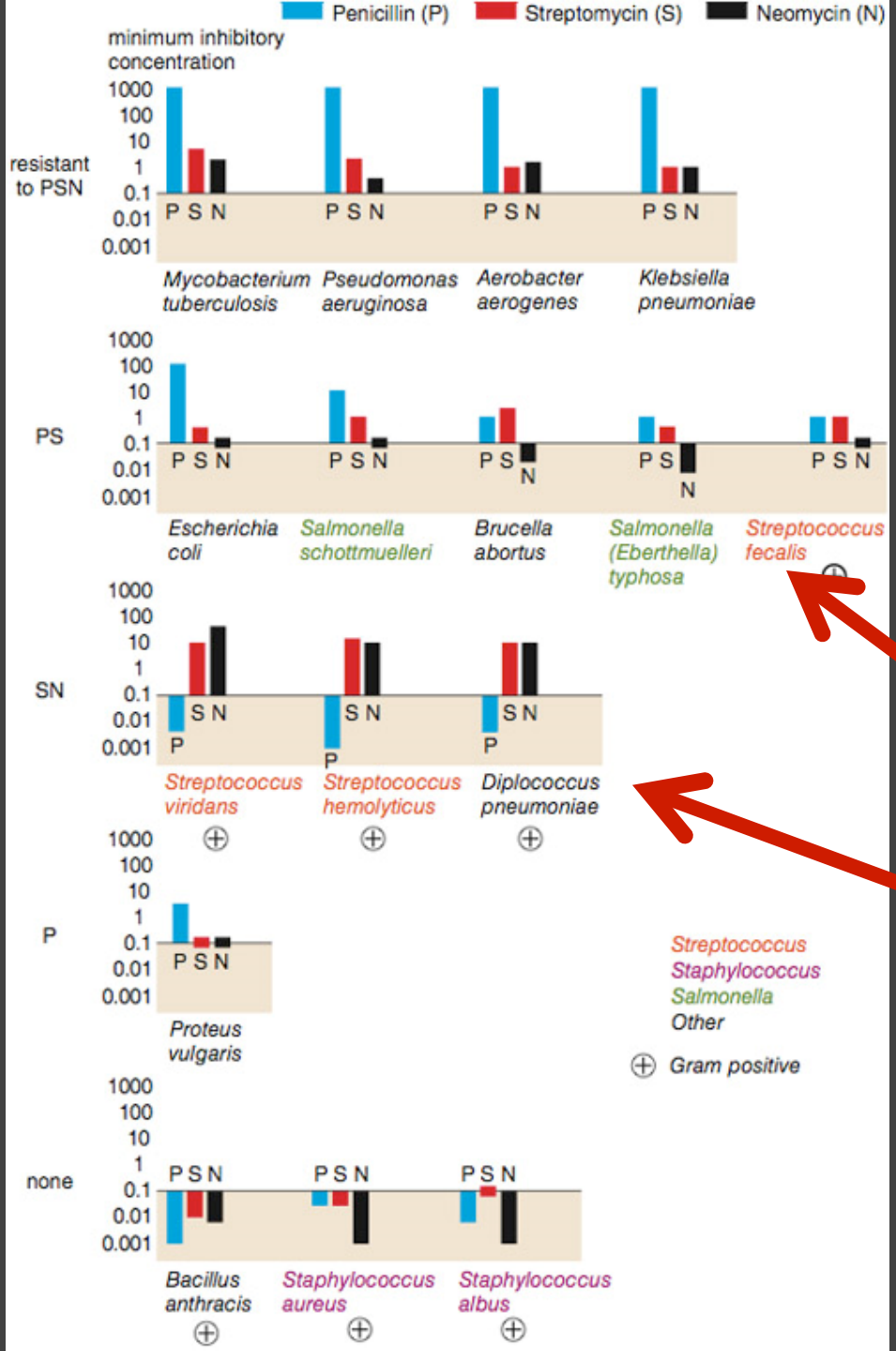
Wainer & Lysen
American Scientist, 2009



Do the bacteria group by antibiotic resistance?

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009



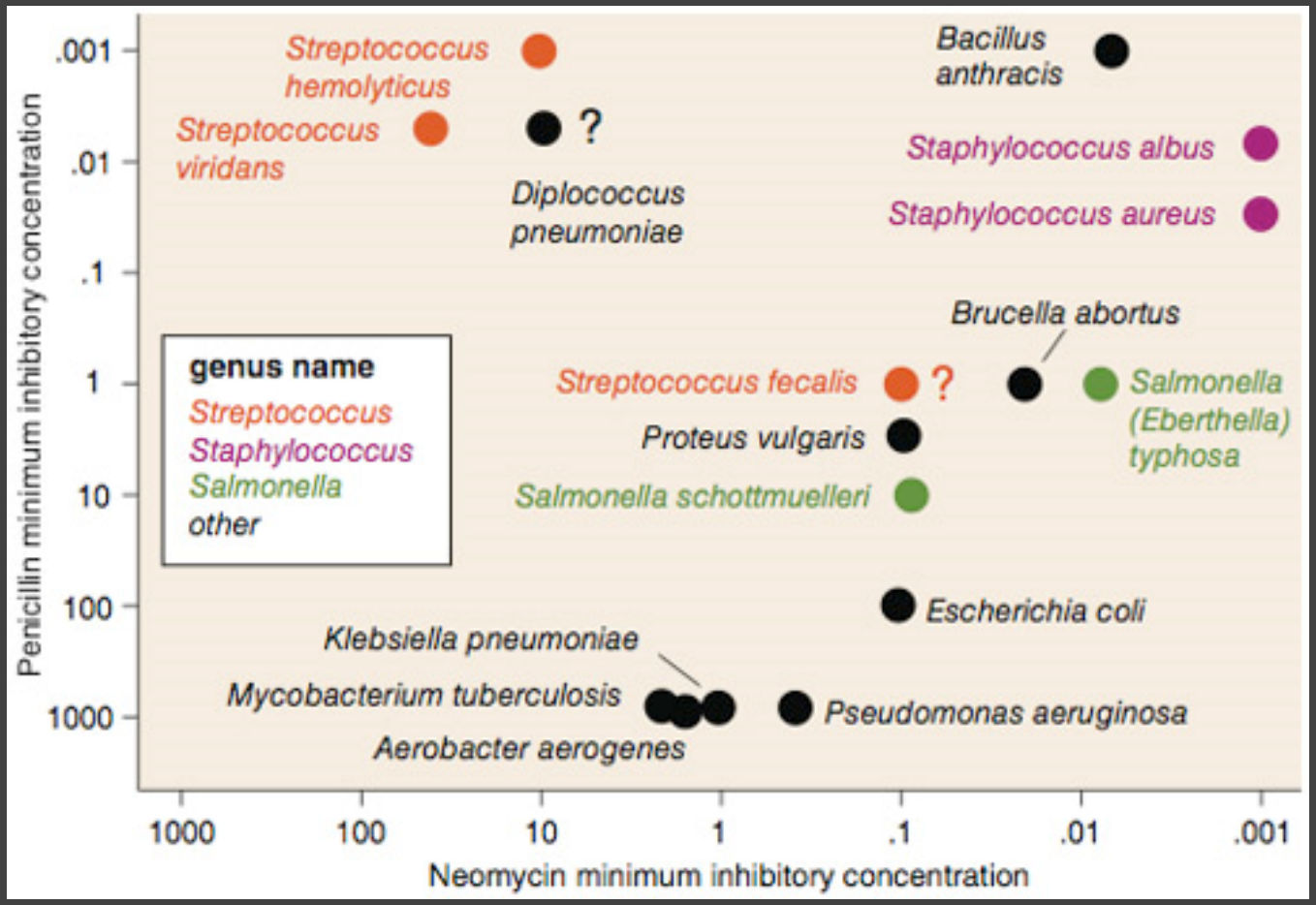
Do the bacteria group by antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

Do the bacteria group by resistance?
Do different drugs correlate?



Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009

Lesson: Iterative Exploration

Exploratory Process

1 Construct graphics to address questions
2 Inspect “answer” and assess new questions
3 Repeat…

Transform data appropriately (e.g., invert, log)

Show data variation, not design variation [Tufte]

Administrivia

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

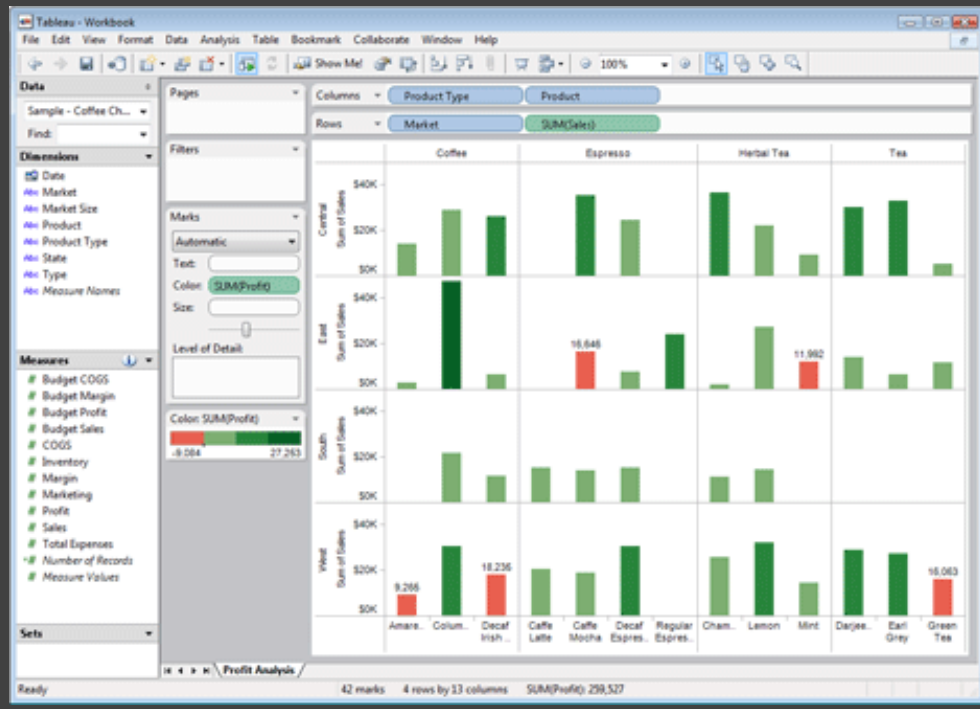
First steps:

Step 1: Pick domain & data

Step 2: Pose questions

Step 3: Profile the data

Iterate as needed



Create visualizations

Interact with data

Refine your questions

Author a report

Screenshots of most insightful views *(8+)*

Include titles and captions for each view

Due by 11:59pm

Monday, Apr 19

Final Project Theme

Data Visualization for Communicating Scientific Advancements or Social Phenomena

Goal: find data of social or scientific import, design visualizations to communicate it effectively to a general audience.

The specific data domain is open-ended. Possibilities include transportation, campaign finance, education, economics, chemical engineering, sociology, statistics, atmospheric science, molecular interactions, scientific research, and so on...

Use Assignment 2 and 3 to explore a data set of interest prior to committing to final project teams and topic!

Final Project

Produce **interactive web-based visualizations**

Initial **prototype** and **design review**

Final deliverables and **video presentation**

Submit and **publish online** (GitHub)

Projects from previous classes (442, 512) have been:

- Published as research papers
- Shared widely (some in the New York Times!)
- Released as successful open source projects

Final Project Teams

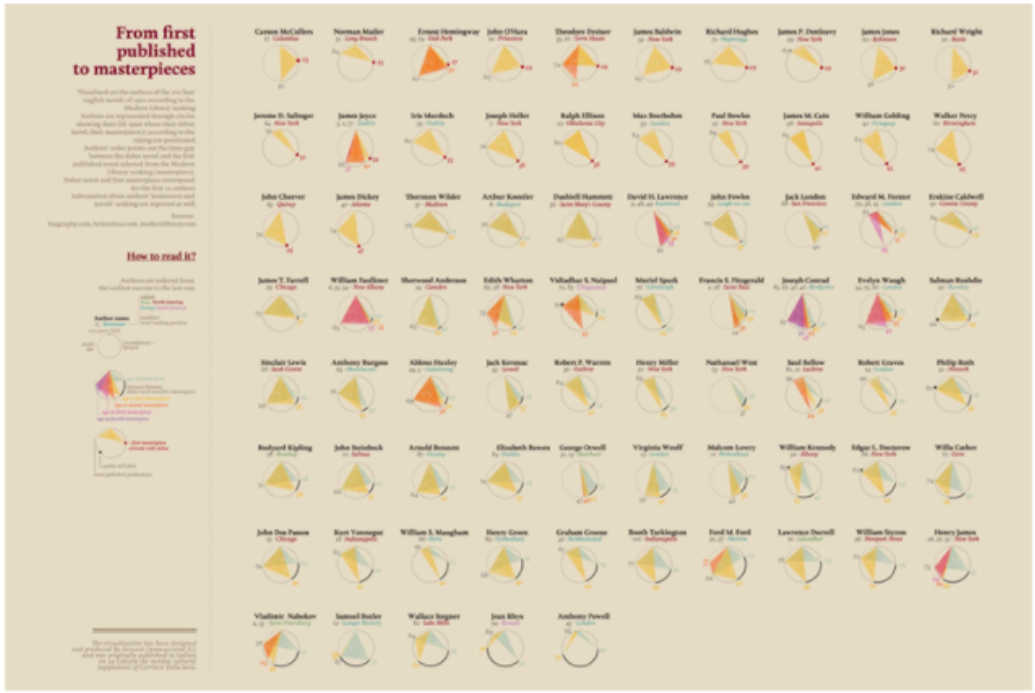
Work in groups of 3-5 people

Post your project ideas and interests on Ed,
or respond to classmates about their projects

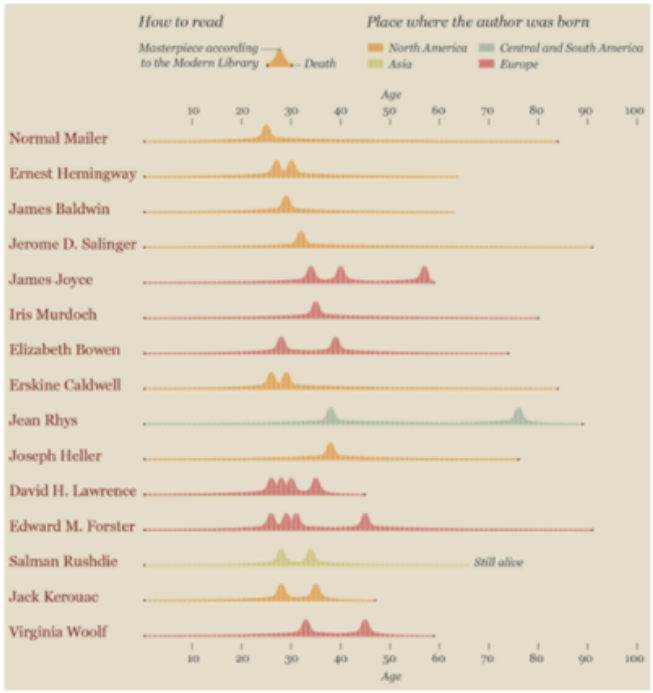
Mark thread as resolved when you are no longer looking for additional members

https://edstem.org/us/courses/4910/discussion/354324

Required Readings for Fri 4/9



Accurat's visualization of the output of famous authors.

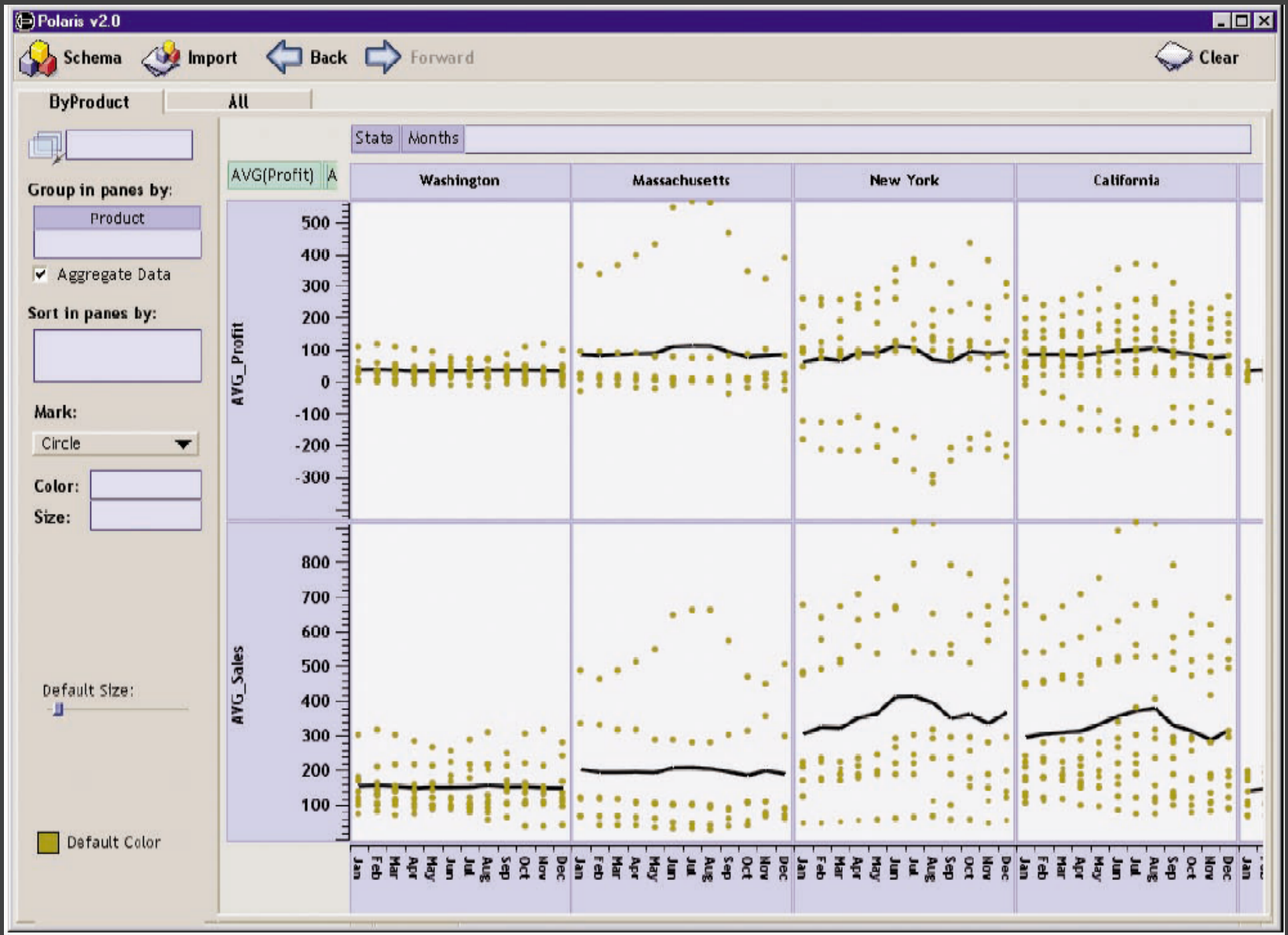


Alberto Cairo's redesign of the same graphic.

Design and Redesign in Data Visualization. Martin Wattenberg and Fernanda Viégas. 2015.

Tableau / Polaris

Polaris [Stolte et al.]



Tableau

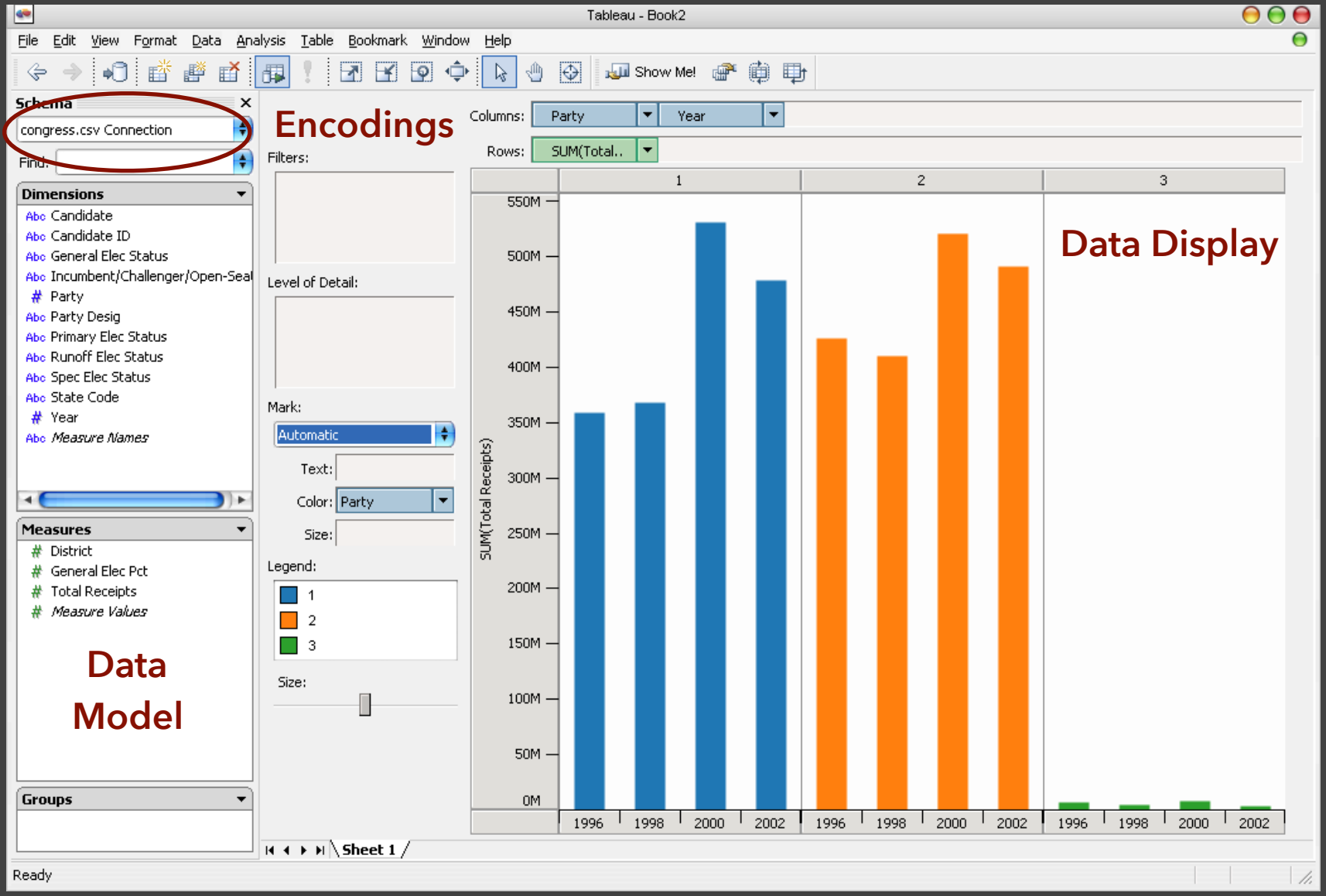


Tableau - Book2
File Edit View Format Data Analysis Table Bookmark Window Help
Show Me!
Schema
congress.csv Connection
Find:
Dimensions
Abc Candidate
Abc Candidate ID
Abc General Elec Status
Abc Incumbent/Challenger/Open-Seat
Party
Abc Party Desig
Abc Primary Elec Status
Abc Runoff Elec Status
Abc Spec Elec Status
Abc State Code
Year
Abc Measure Names
Measures
District
General Elec Pct
Total Receipts
Measure Values
Data Model
Groups
Encodings
Filters:
Level of Detail:
Mark:
Automatic
Text:
Color: Party
Size:
Legend:
1
2
3
Size:
Columns: Party Year
Rows: SUM(Total..
Data Display
SUM(Total Receipts)
550M
500M
450M
400M
350M
300M
250M
200M
150M
100M
50M
0M
1996 1998 2000 2002 1996 1998 2000 2002 1996 1998 2000 2002
Sheet 1
Ready

Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest encodings upon request

Specifying Table Configurations

Operands are the database fields
Each operand interpreted as a set {…}
Quantitative and Ordinal fields treated differently

Three operators:
concatenation (+)
cross product (x)
nest (/)

Tableau - Book1

Normal

Show Me

Data | Analytics

Sample – Superstore

Dimensions

- Customer
  - Customer Name
  - Segment
- Order
- Location
- Product
  - Category
  - Sub-Category
  - Manufacturer
  - Product Name
- Profit (bin)
- Region
- *Measure Names*

Measures

- Discount
- Profit
- Profit Ratio
- Quantity
- Sales
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

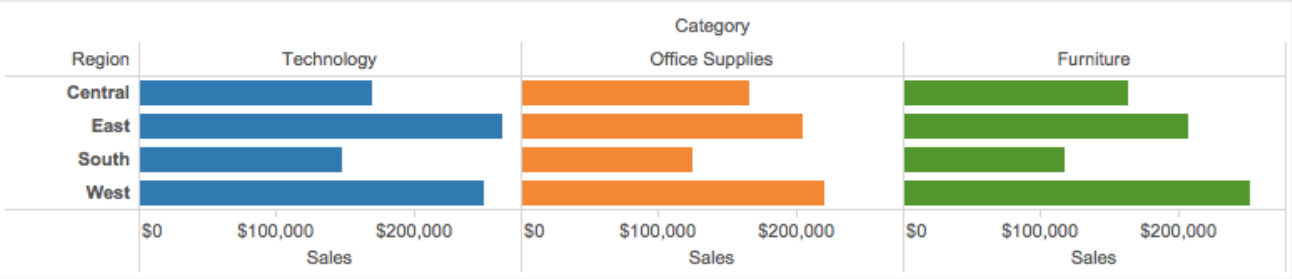
Color | Size | Label | Detail | Tooltip

Category

Category
- Technology
- Office Supplies
- Furniture

Columns: Category | SUM(Sales)

Rows: Region



Data Source | Sheet 1

12 marks | 4 rows by 3 columns | SUM(Sales): $2,297,201

Tableau - Book1

Normal

Show Me

Data | Analytics

Sample – Superstore

Dimensions

- Customer
  - Customer Name
  - Segment
- Order
- Location
- Product
  - Category
  - Sub-Category
  - Manufacturer
  - Product Name
- Profit (bin)
- Region
- *Measure Names*

Measures

- Discount
- Profit
- Profit Ratio
- Quantity
- Sales
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

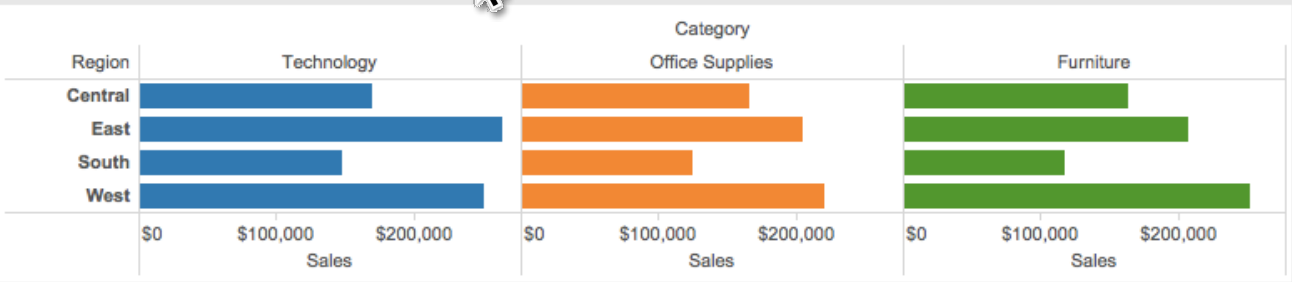
Color | Size | Label | Detail | Tooltip

Category

Category
- Technology
- Office Supplies
- Furniture

Columns: Category | SUM(Sales)

Rows: Region



Data Source | Sheet 1

12 marks | 4 rows by 3 columns | SUM(Sales): $2,297,201

Tableau - Book1

Normal | Show Me

Data | Analytics

Sample – Superstore

Dimensions

- Customer
  - Customer Name
  - Segment
- Order
- Location
- Product
  - Category
  - Sub-Category
  - Manufacturer
  - Product Name
- Profit (bin)
- Region
- *Measure Names*

Measures

- Discount
- Profit
- Profit Ratio
- Quantity
- Sales
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

Color | Size | Label | Detail | Tooltip

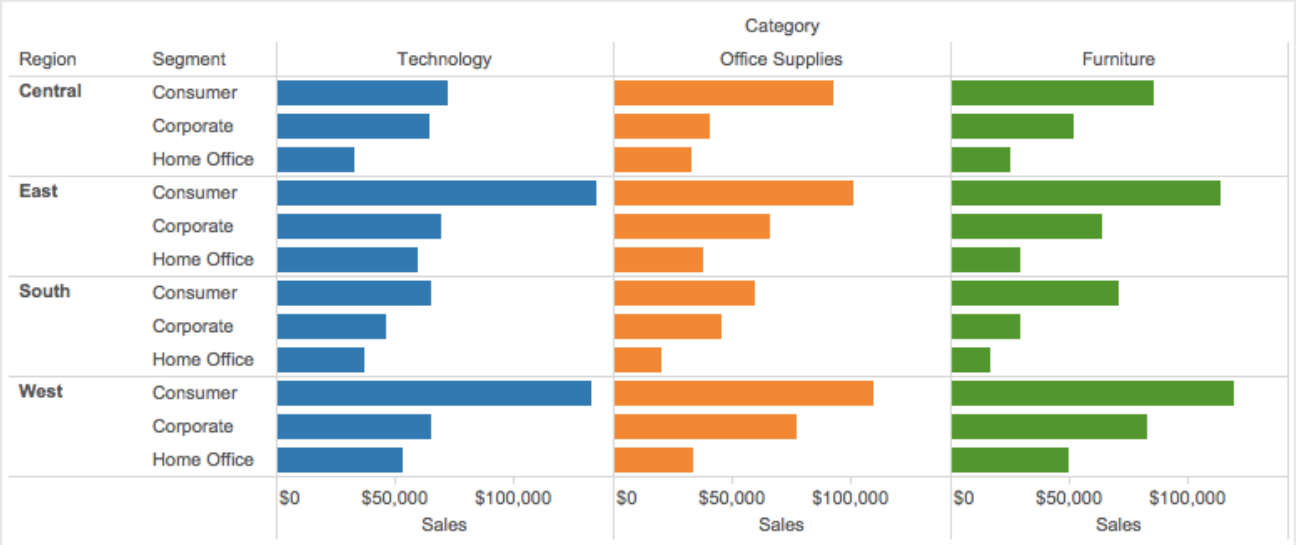
Category

Category

- Technology
- Office Supplies
- Furniture

Columns: Category | SUM(Sales)

Rows: Region | Segment



Data Source | Sheet 1

36 marks | 12 rows by 3 columns | SUM(Sales): $2,297,201

Tableau - Book1

Normal | Show Me

Data | Analytics

Sample – Superstore

Dimensions

- Customer
  - Customer Name
  - Segment
- Order
- Location
- Product
  - Category
  - Sub-Category
  - Manufacturer
  - Product Name
- Profit (bin)
- Region
- *Measure Names*

Measures

- Discount
- Profit
- Profit Ratio
- Quantity
- Sales
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

Filters

Marks

Automatic

Color | Size | Label | Detail | Tooltip

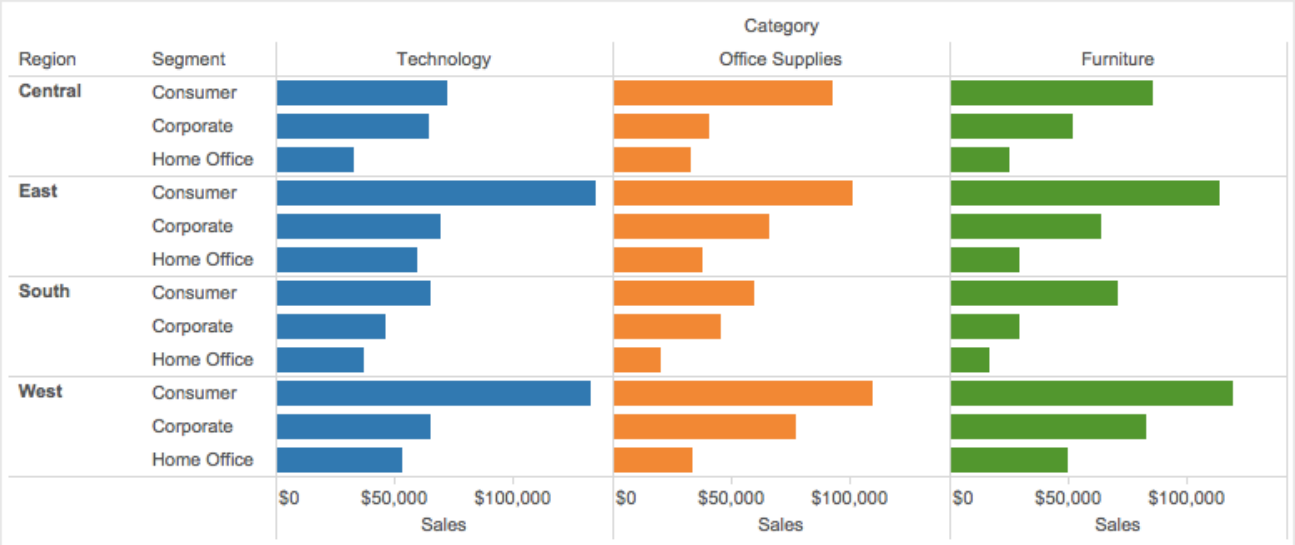
Category

Category

- Technology
- Office Supplies
- Furniture

Columns: Category | SUM(Sales)

Rows: Region | Segment



Data Source | Sheet 1

36 marks | 12 rows by 3 columns | SUM(Sales): $2,297,201

Tableau - Book1

Normal | Show Me

Data | Analytics

Sample – Superstore

Dimensions
- Customer
  - Customer Name
  - Segment
- Order
- Location
- Product
  - Category
  - Sub-Category
  - Manufacturer
  - Product Name
- Profit (bin)
- Region
- *Measure Names*

Measures
- Discount
- Profit
- Profit Ratio
- Quantity
- Sales
- *Latitude (generated)*
- *Longitude (generated)*
- *Number of Records*
- *Measure Values*

Pages

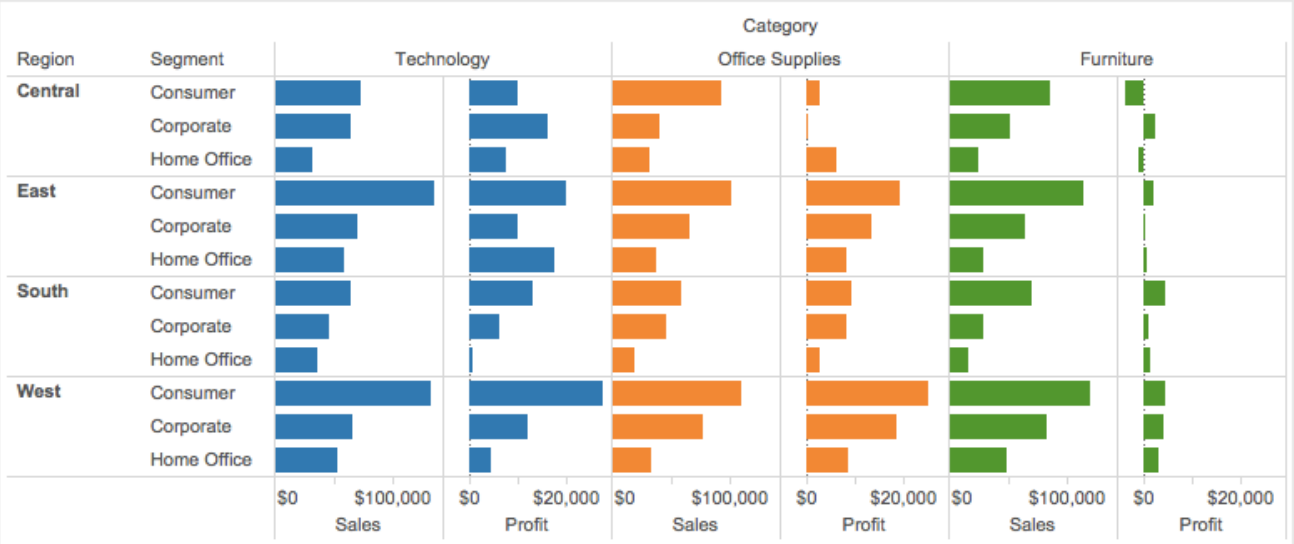
Filters

Marks
All
Automatic
Color | Size | Label | Detail | Tooltip
Category
SUM(Sales)
SUM(Profit)

Category
- Technology
- Office Supplies
- Furniture

Columns: Category | SUM(Sales) | SUM(Profit)

Rows: Region | Segment



Data Source | Sheet 1

72 marks | 12 rows by 6 columns | SUM(Profit): $286,397

Tableau - Book1

Normal

Show Me

Columns: Category, SUM(Sales), SUM(Profit)

Rows: Region, Segment

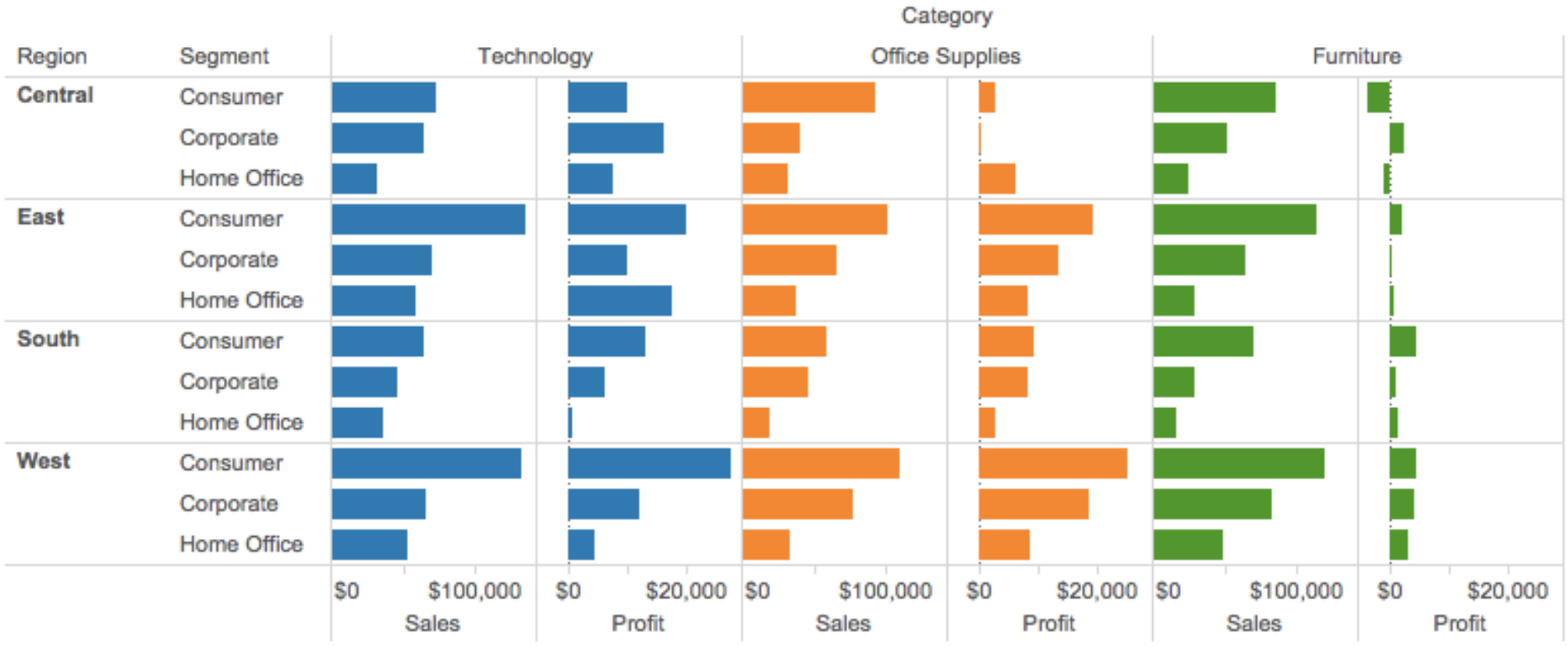


Table Algebra

The operators (+, x, /) and operands (O, Q) provide an *algebra* for tabular visualization.

Algebraic statements are then mapped to:
Visualizations - trellis plot partitions, visual encodings
Queries - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop
Note that this specifies operands *NOT* operators!
Operators are inferred by data type (O, Q)

Table Algebra: Operands

Ordinal fields: interpret domain as a set that partitions table into rows and columns.

Quarter = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} ->

Qtr1	Qtr2	Qtr3	Qtr4
95892	101760	105282	98225

Quantitative fields: treat domain as single element set and encode spatially as axes.

Profit = {(Profit[-410,650])} ->

Concatenation (+) Operator

Ordered union of set interpretations

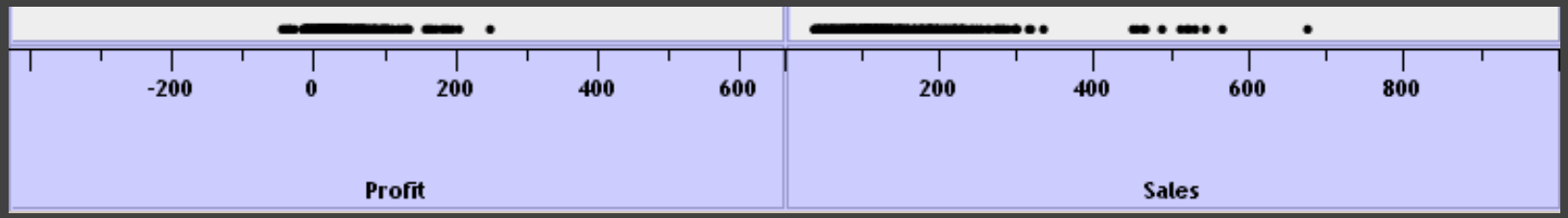
Quarter + Product Type

= {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} + {(Coffee), (Espresso)}

= {(Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso)}

Qtr1	Qtr2	Qtr3	Qtr4	Coffee	Espresso
48	59	57	53	151	21

Profit + Sales = {(Profit[-310,620]),(Sales[0,1000])}



Cross (x) Operator

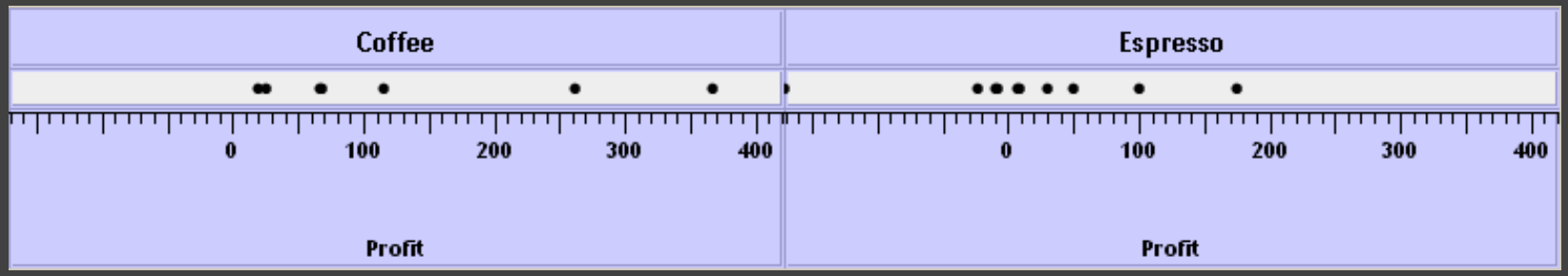
Cross-product of set interpretations

Quarter x Product Type =

{(Qtr1,Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4,Tea)}

Qtr1		Qtr2		Qtr3		Qtr4	
Coffee	Espresso	Coffee	Espresso	Coffee	Espresso	Coffee	Espresso
131	19	160	20	178	12	134	33

Product Type x Profit =



Nest (/) Operator

Cross-product filtered by existing records

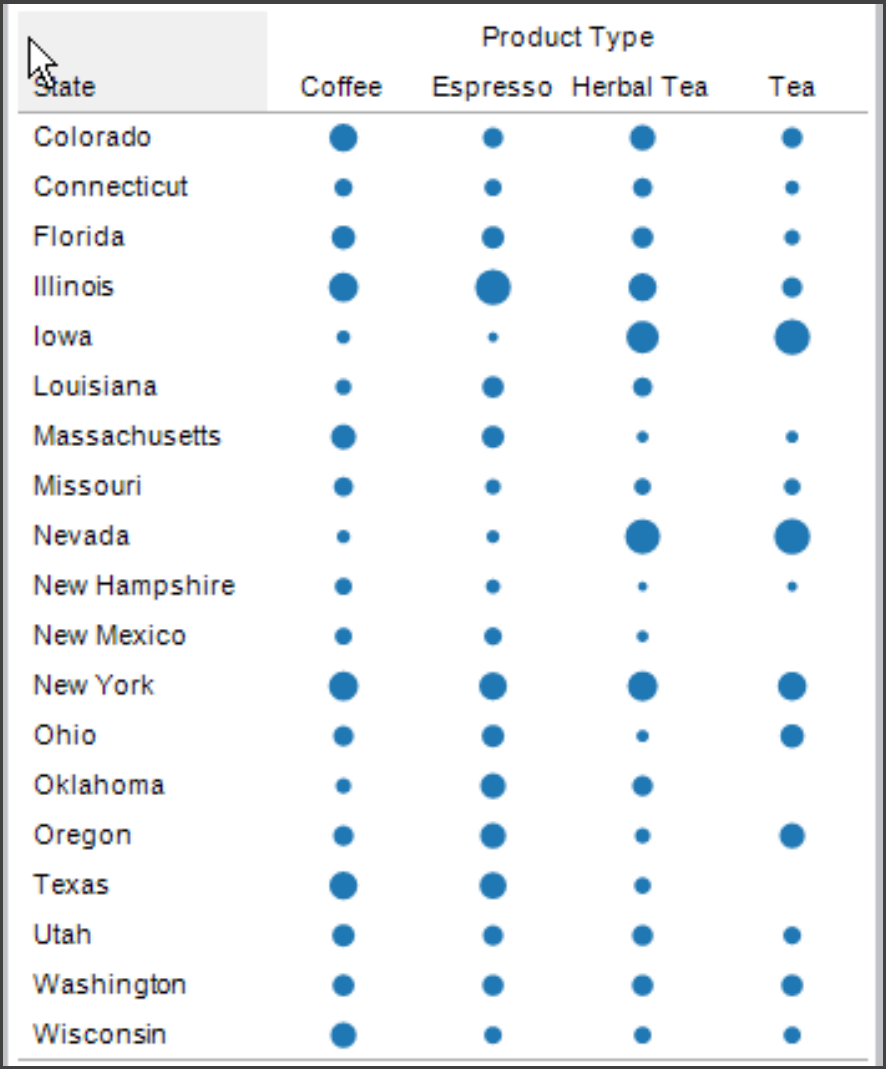
Quarter x Month ->

creates twelve entries for each quarter. i.e., (Qtr1, December)

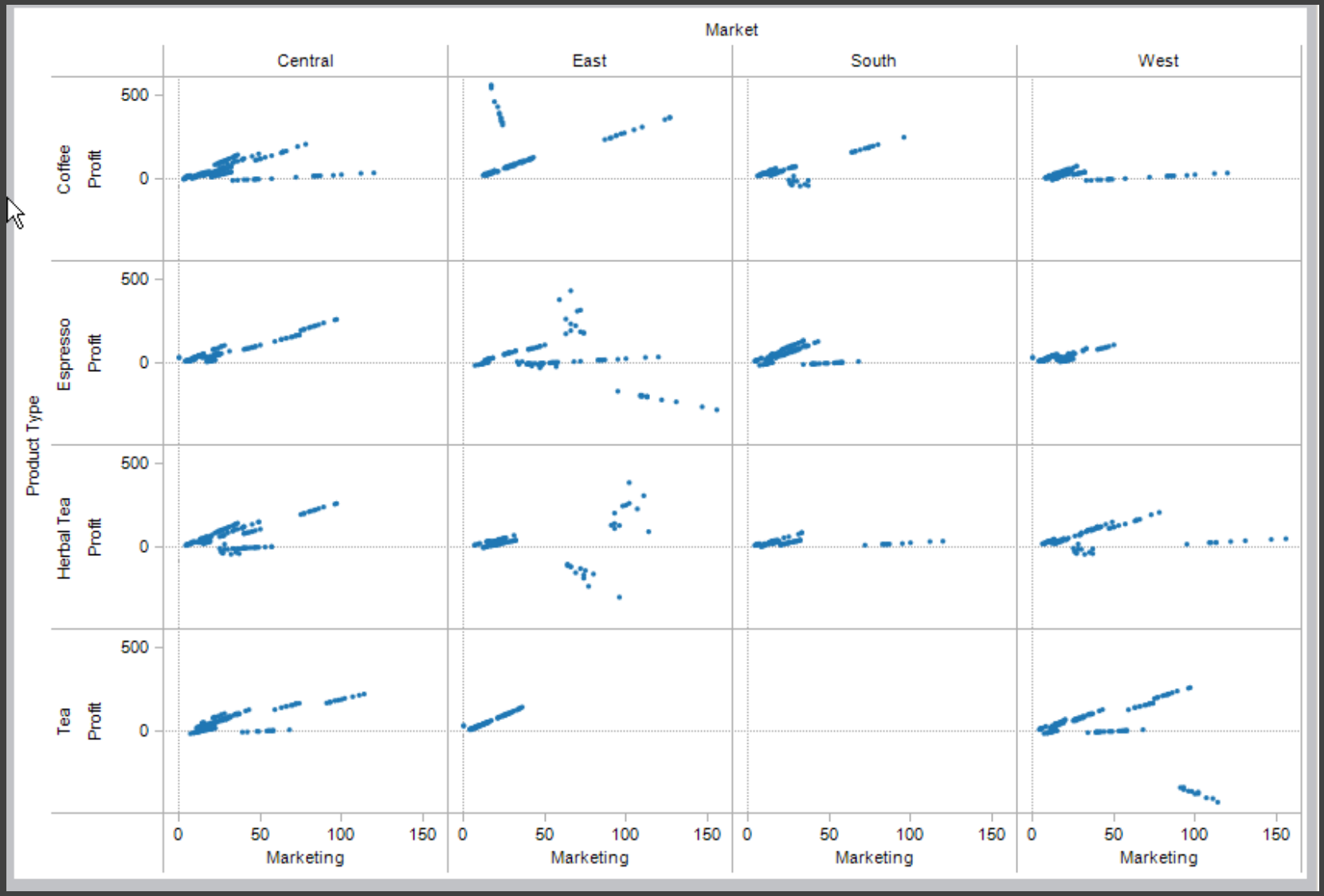
Quarter / Month ->

creates three entries per quarter based on tuples in database (not semantics)

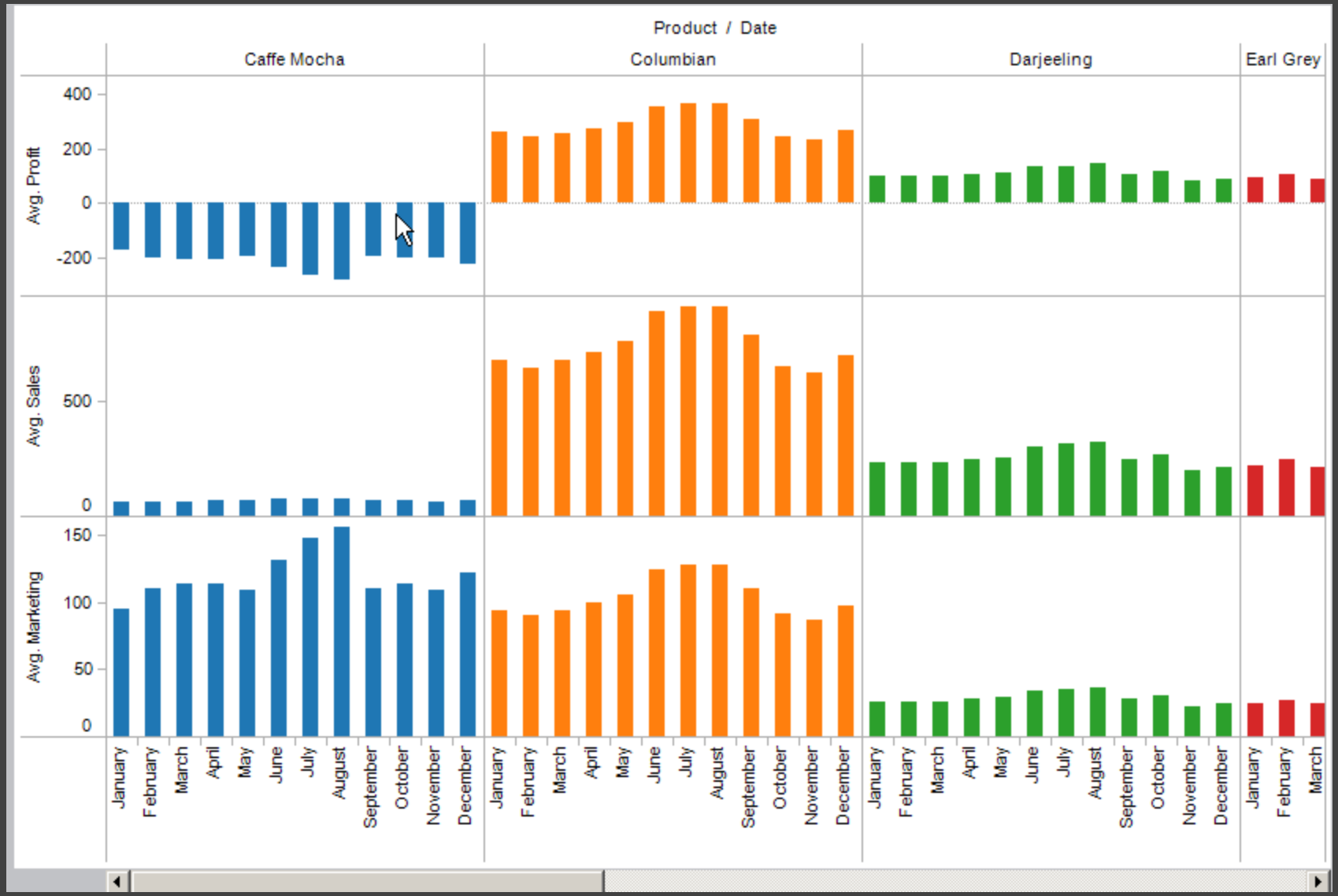
Ordinal-Ordinal



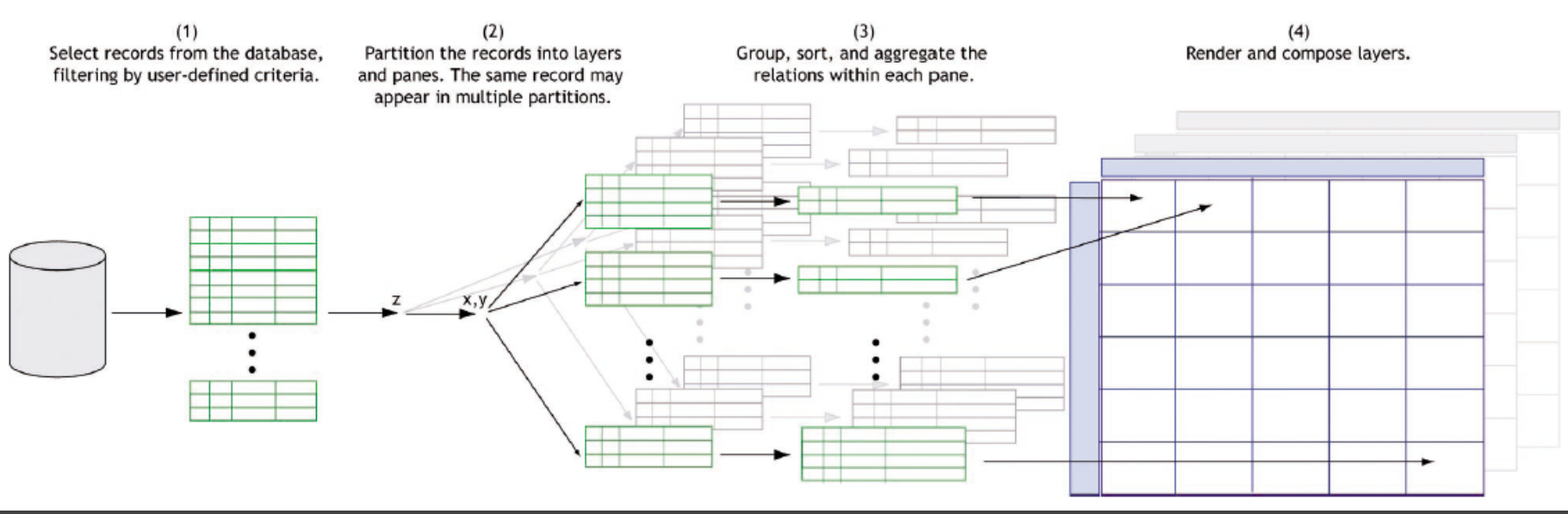
Quantitative-Quantitative



Ordinal-Quantitative



Querying the Database



Quiz Section: Tableau

Tomorrow, Thursday April 8th

Introduction and hands-on experience in Tableau
Come prepared with Tableau installed
See announcement on Ed for instructions

Up Next: Jane's Office Hour (link on Canvas)