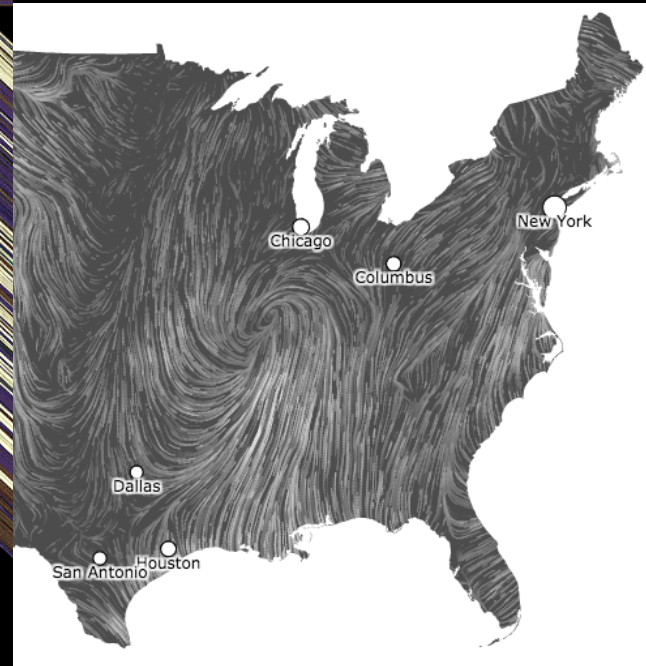
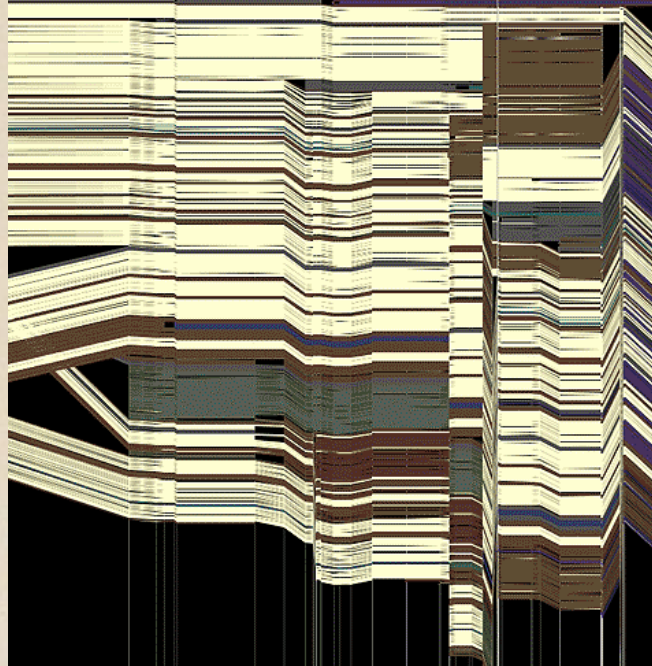
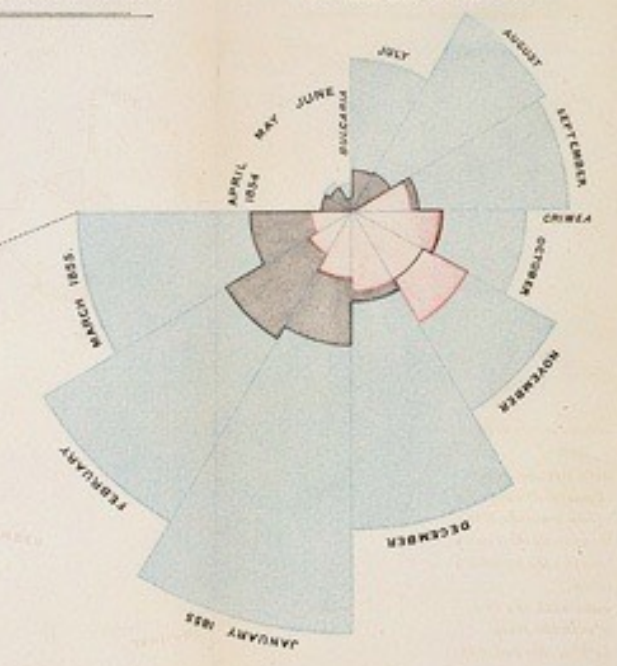


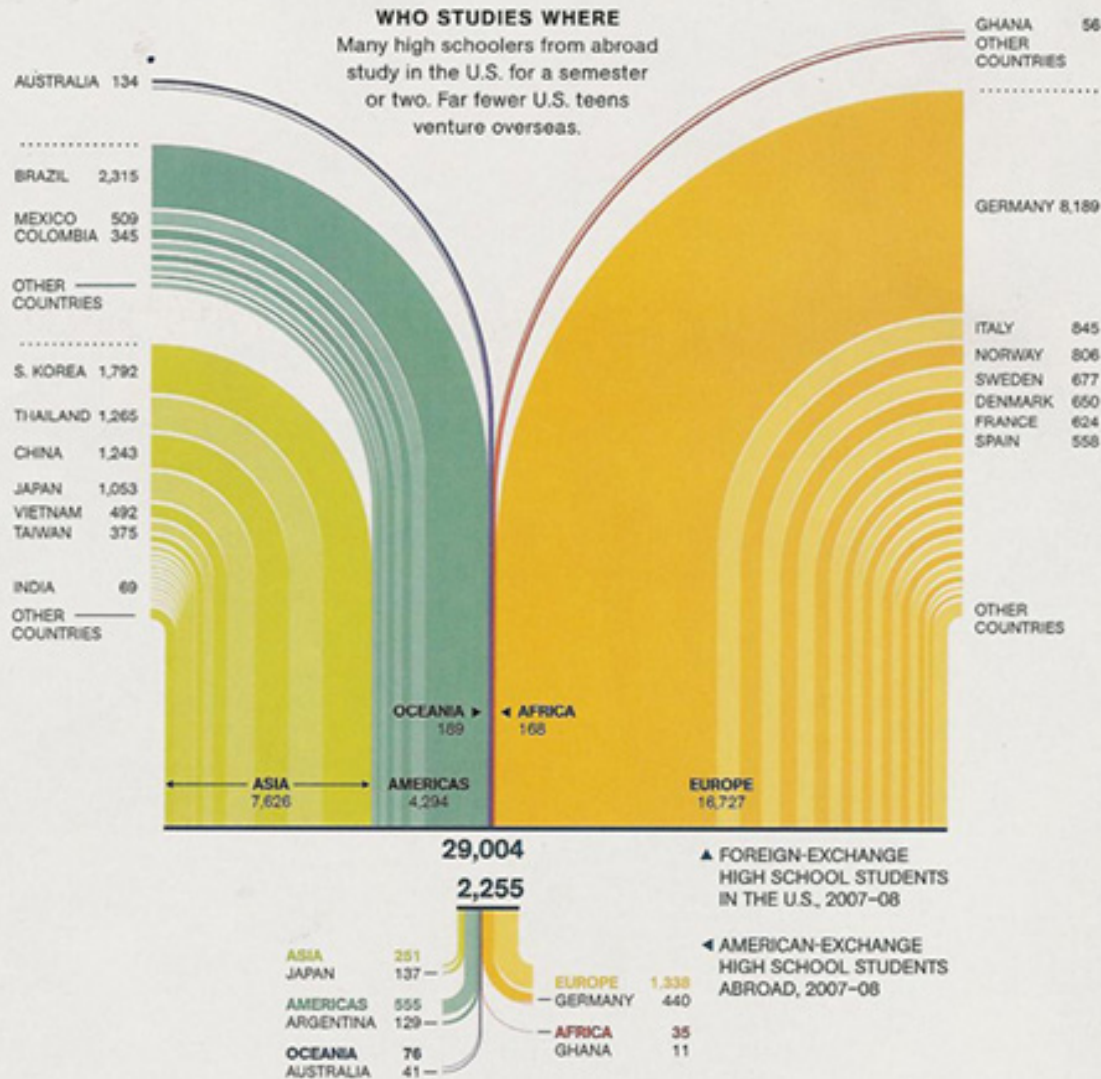
CSE 512 - Data Visualization

Exploratory Data Analysis



Jeffrey Heer University of Washington

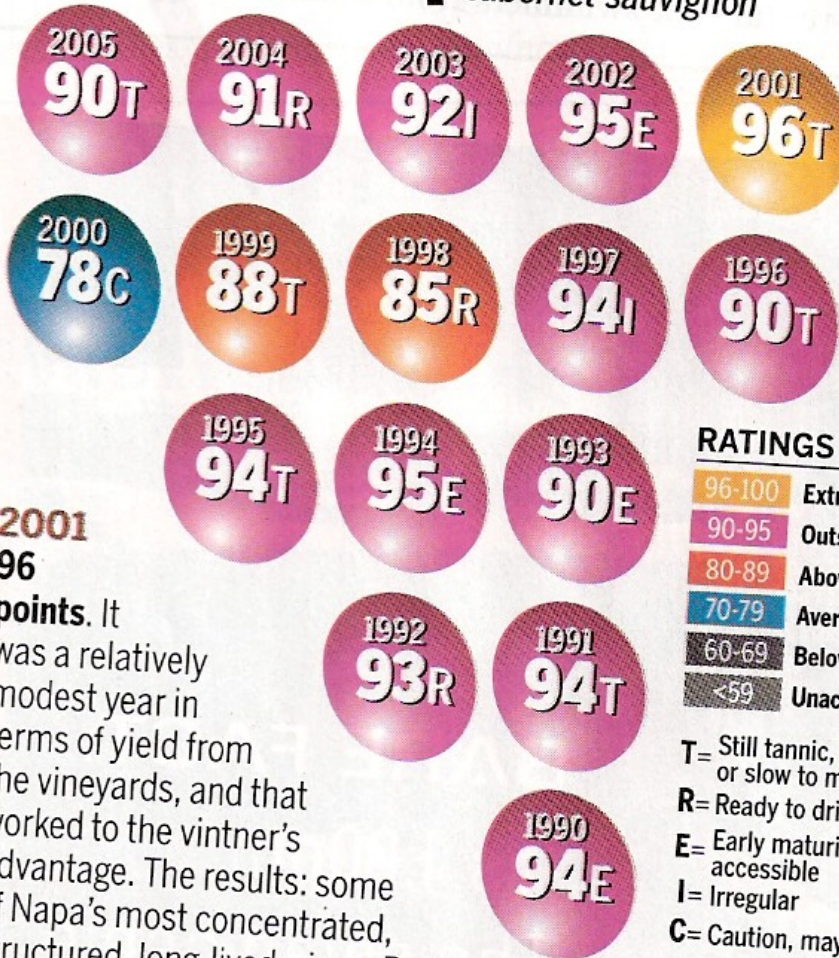
Re-Design Exercise



Source: *National Geographic*, September, 2008, p. 22.
 Silver, Mark. "High School Give-and-Take."

IT WAS A VERY GOOD YEAR?

Robert Parker's ratings for vintages of Napa Valley cabernet sauvignon



2001
96

points. It was a relatively modest year in terms of yield from the vineyards, and that worked to the vintner's advantage. The results: some of Napa's most concentrated, structured, long-lived wines. Built for aging, they are rich, densely colored,

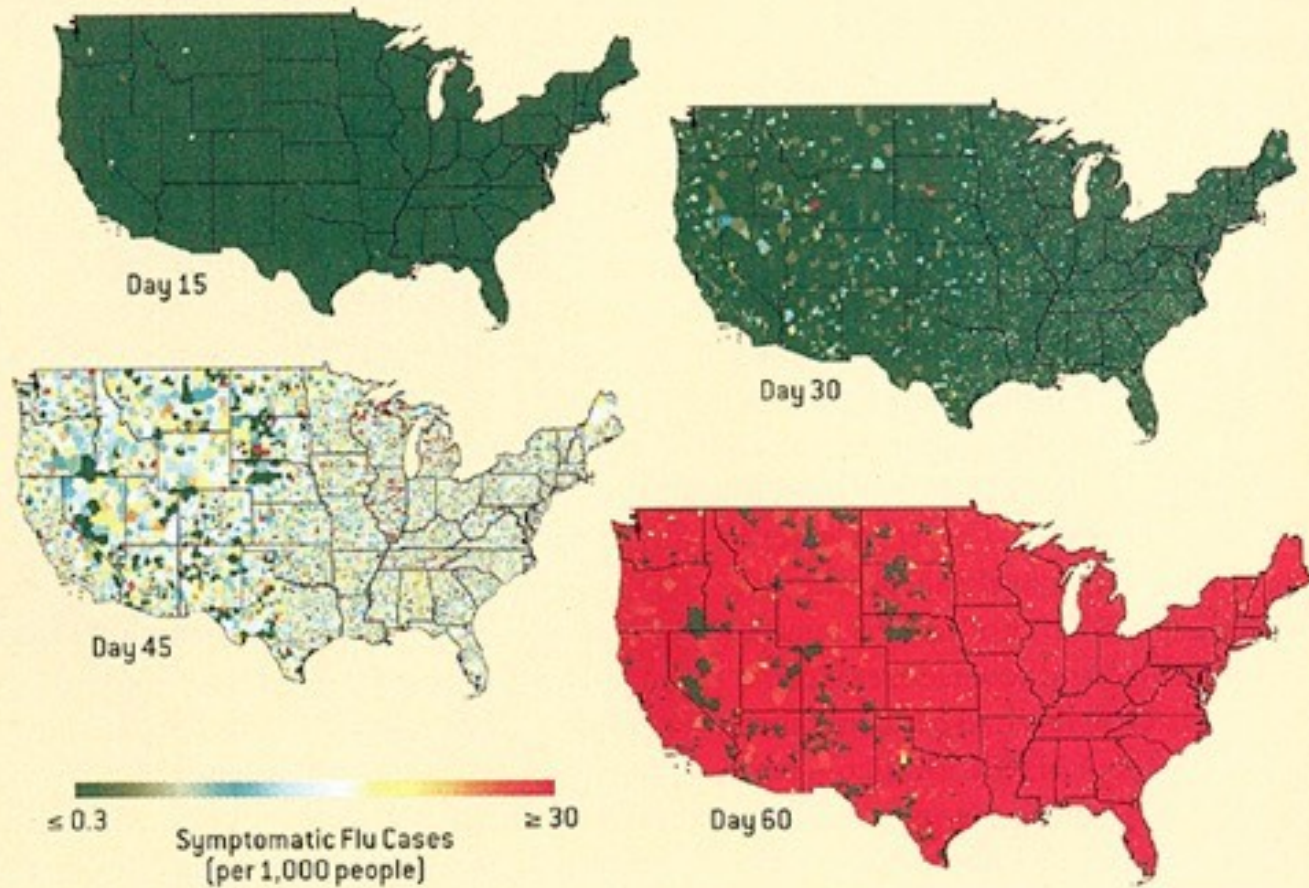
RATINGS

96-100	Extraordinary
90-95	Outstanding
80-89	Above average
70-79	Average
60-69	Below average
<59	Unacceptable

T= Still tannic, youthful, or slow to mature
 R= Ready to drink
 E= Early maturing and accessible
 I= Irregular
 C= Caution, may be too old

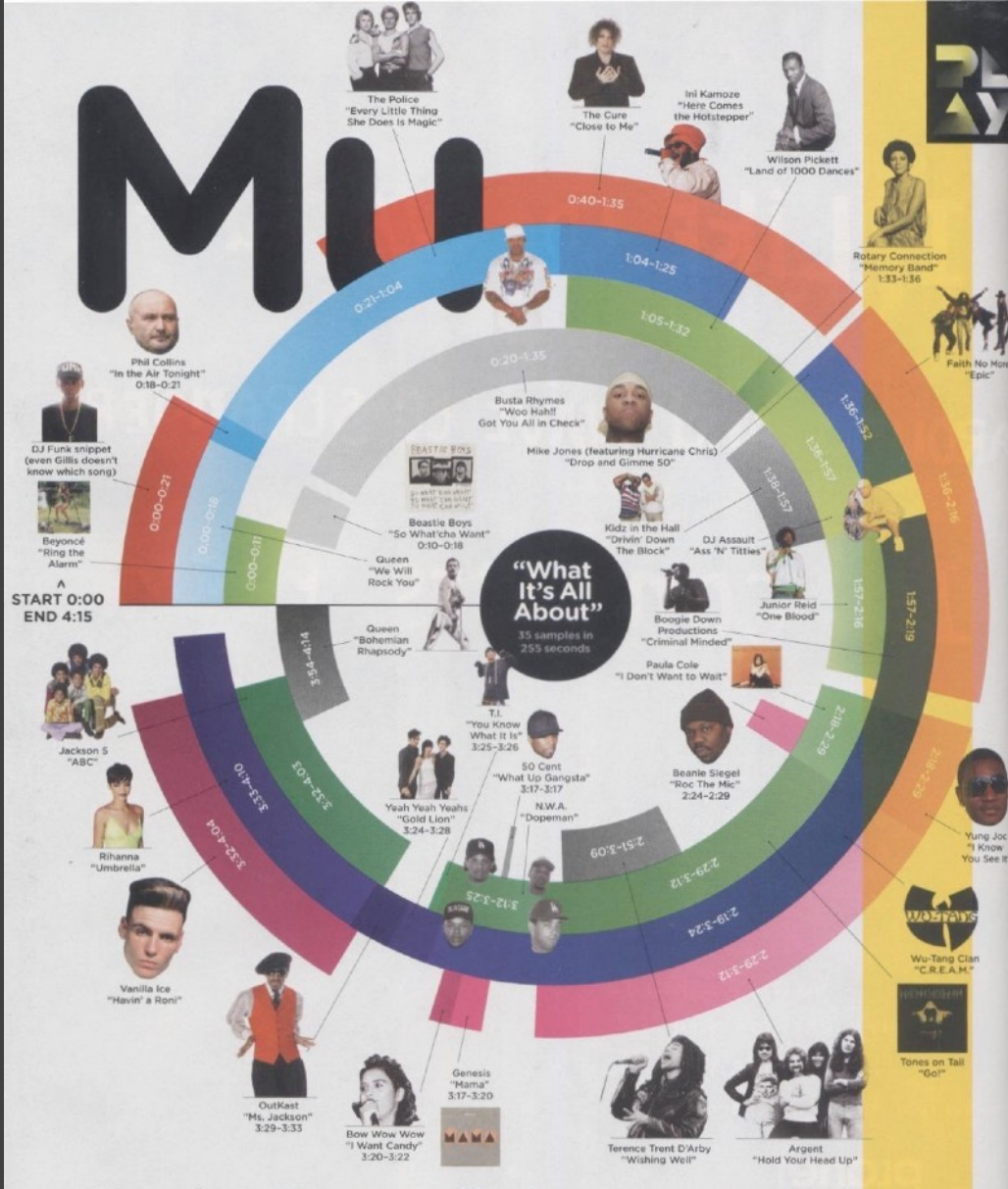
Pandemic Flu Hits the U.S.

A simulation created by researchers from Los Alamos National Laboratory and Emory University shows the first wave of a pandemic spreading rapidly with no vaccine or antiviral drugs employed to slow it down. Colors represent the number of symptomatic flu cases per 1,000 people (see scale). Starting with 40 infected people on the first day, nationwide cases peak around day 60, and the wave subsides after four months with 33 percent of the population having become sick. The scientists are also modeling potential interventions with drugs and vaccines to learn if travel restrictions, quarantines and other disruptive disease-control strategies could be avoided.



Preparing for a Pandemic

Source: *Scientific American*, 293(5). November, 2005, p. 50



Source: *Wired Magazine*, September 2008 Edition
 Music: Super Cuts (page 92)

What was the **first**
data visualization?

0 BC

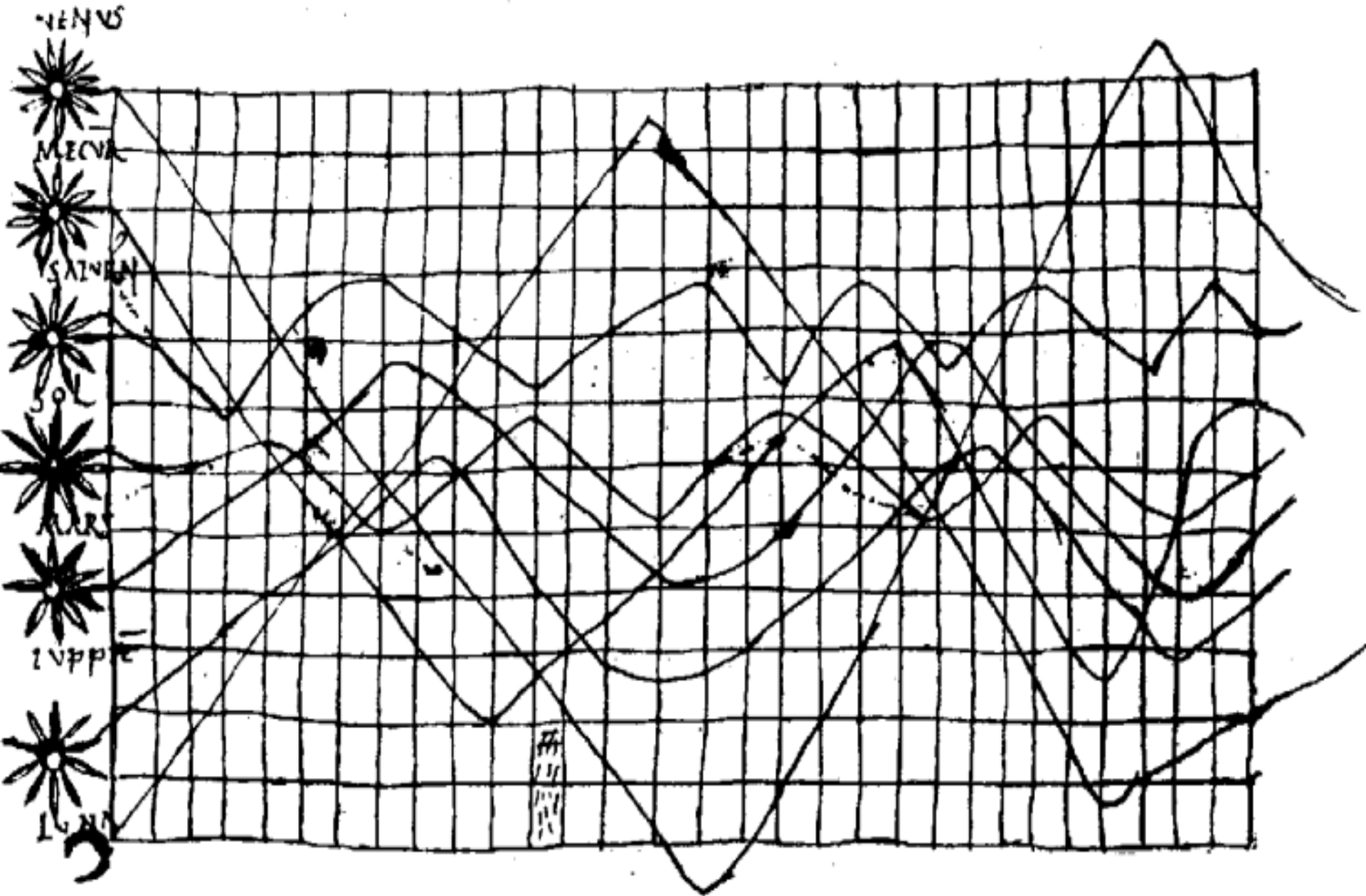




~6200 BC Town Map (?) of Catal Hyük, Konya Plain, Turkey

0 BC

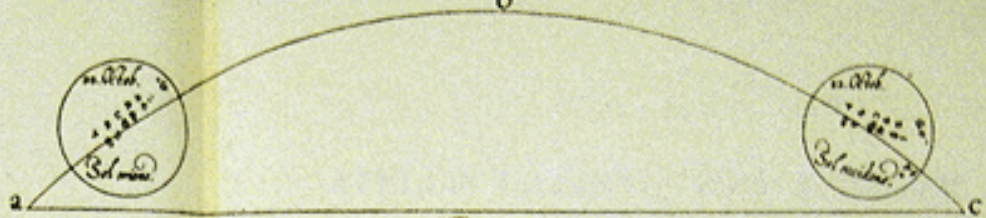




~950 AD Position of Sun, Moon and Planets



MACVLAE IN SOLE APPARENTES, OBSERVATAE
 anno 1611. ad latitudinem grad. 48. min. 40.



a c, horizon. a b c, arcus solis diurnus. Sol oriens ex parte a, maculas exhibet quas vides, occidens vero c, easdem ratione primj motus, nonnihil inuertit. Et hanc matutinam vespertinamq; mutationem, omnes maculae quotidie subeunt. Quod semel exhibuisse et mouisse, sufficiat.



Macula M, es
 hac tenus usarum
 maxima, nulliq;
 prima magnitudinis
 sideri fixo cedit.

Macula I fuit
 valde conspicua,
 propter notabi-
 lem pra reliquis
 magnitudinem.

Figura qua
 habet sinuatum
 signum X, est
 Omittere.

Alm. New. Ang. m. 1611.

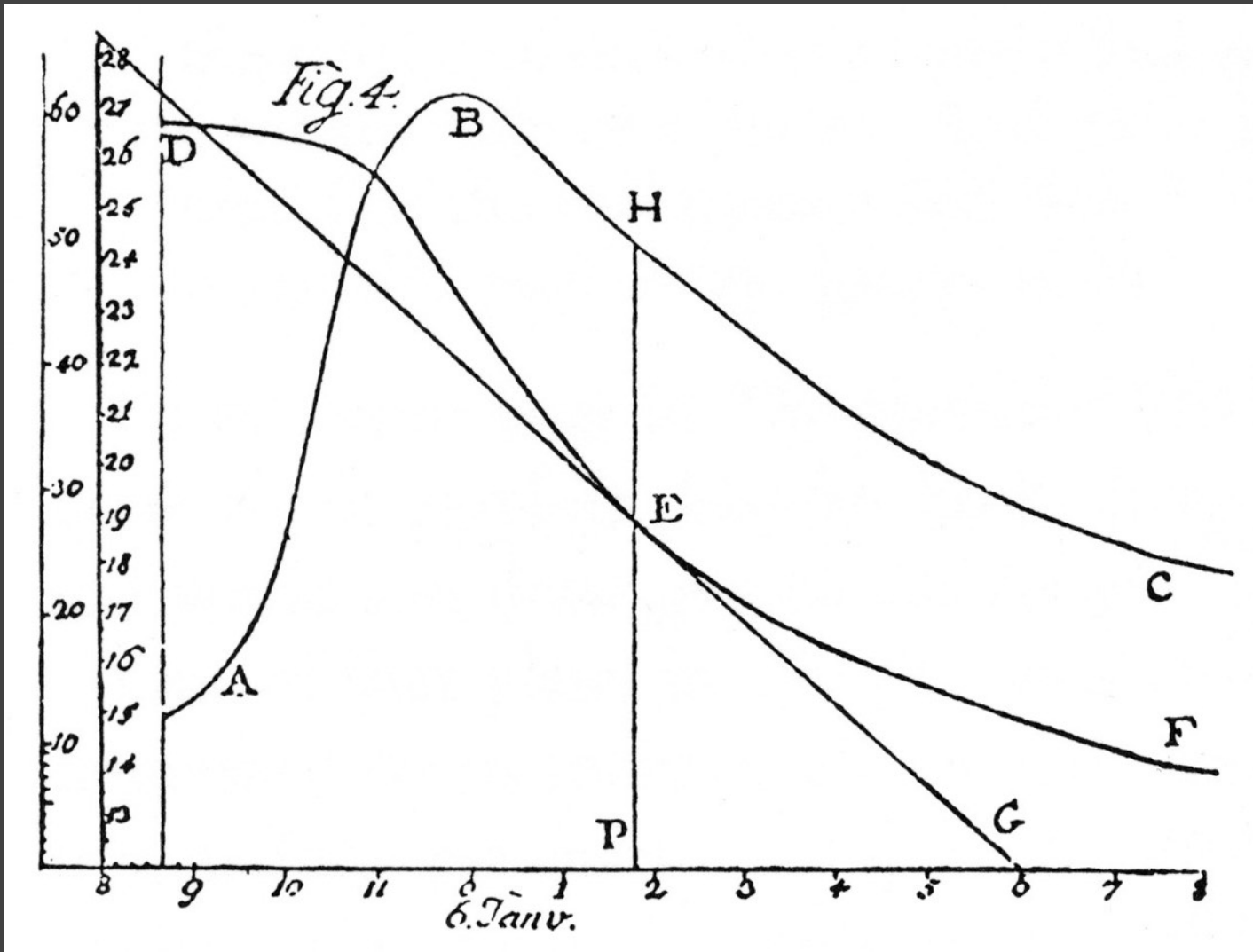
Com. p. 1611. (1611)

TOLEDO.

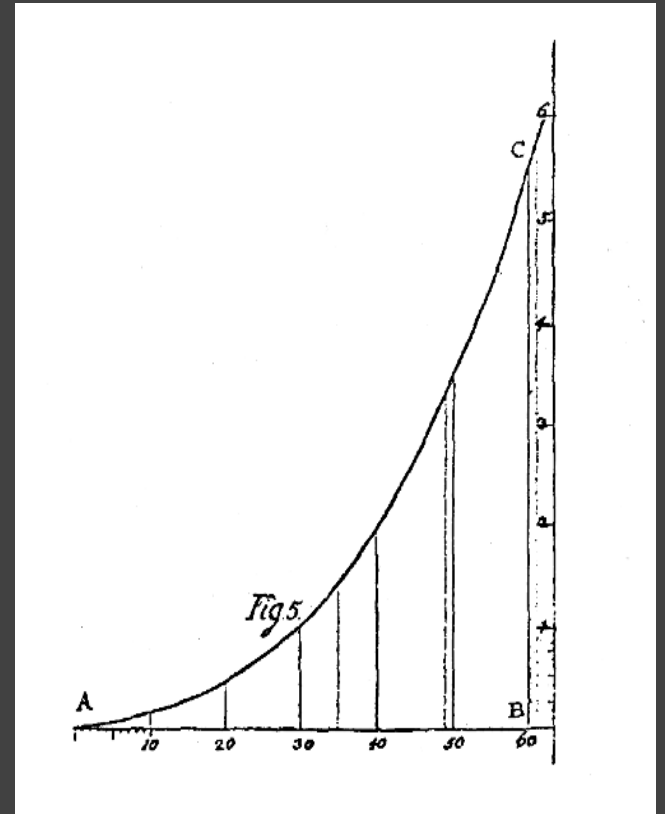
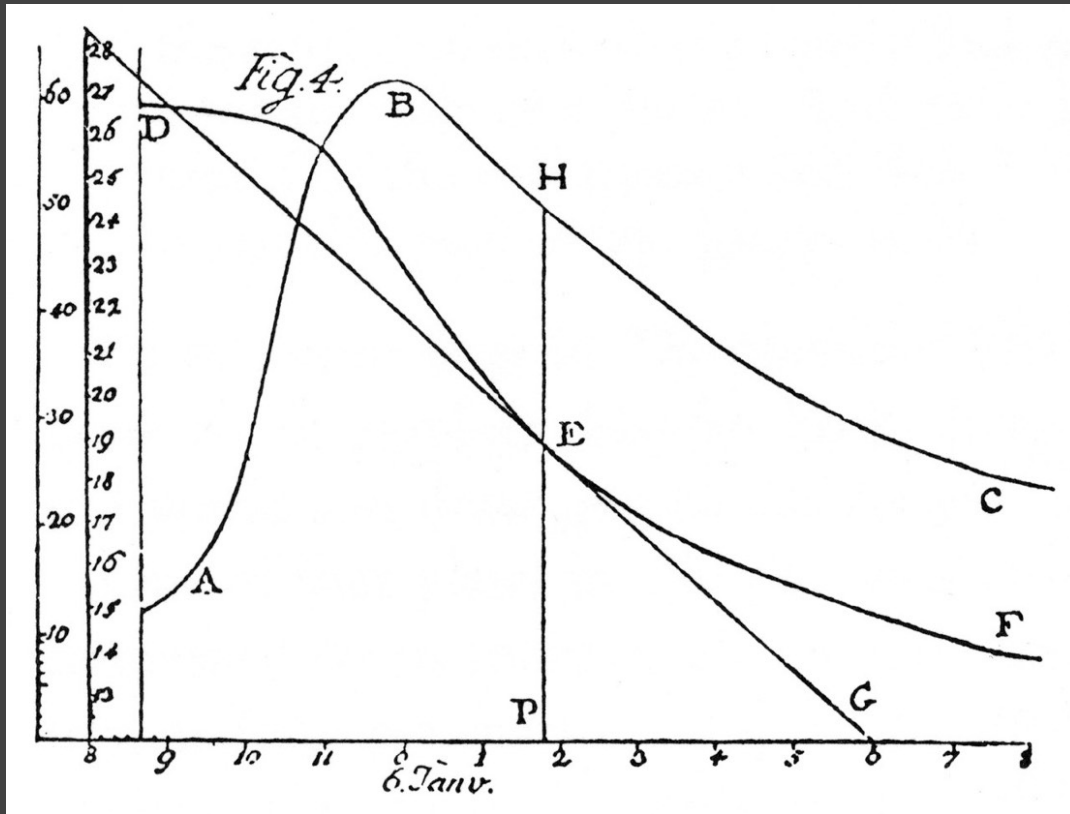
GRADOS DE LA LONGITUD.



Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765



The Rate of Water Evaporation, Lambert 1765

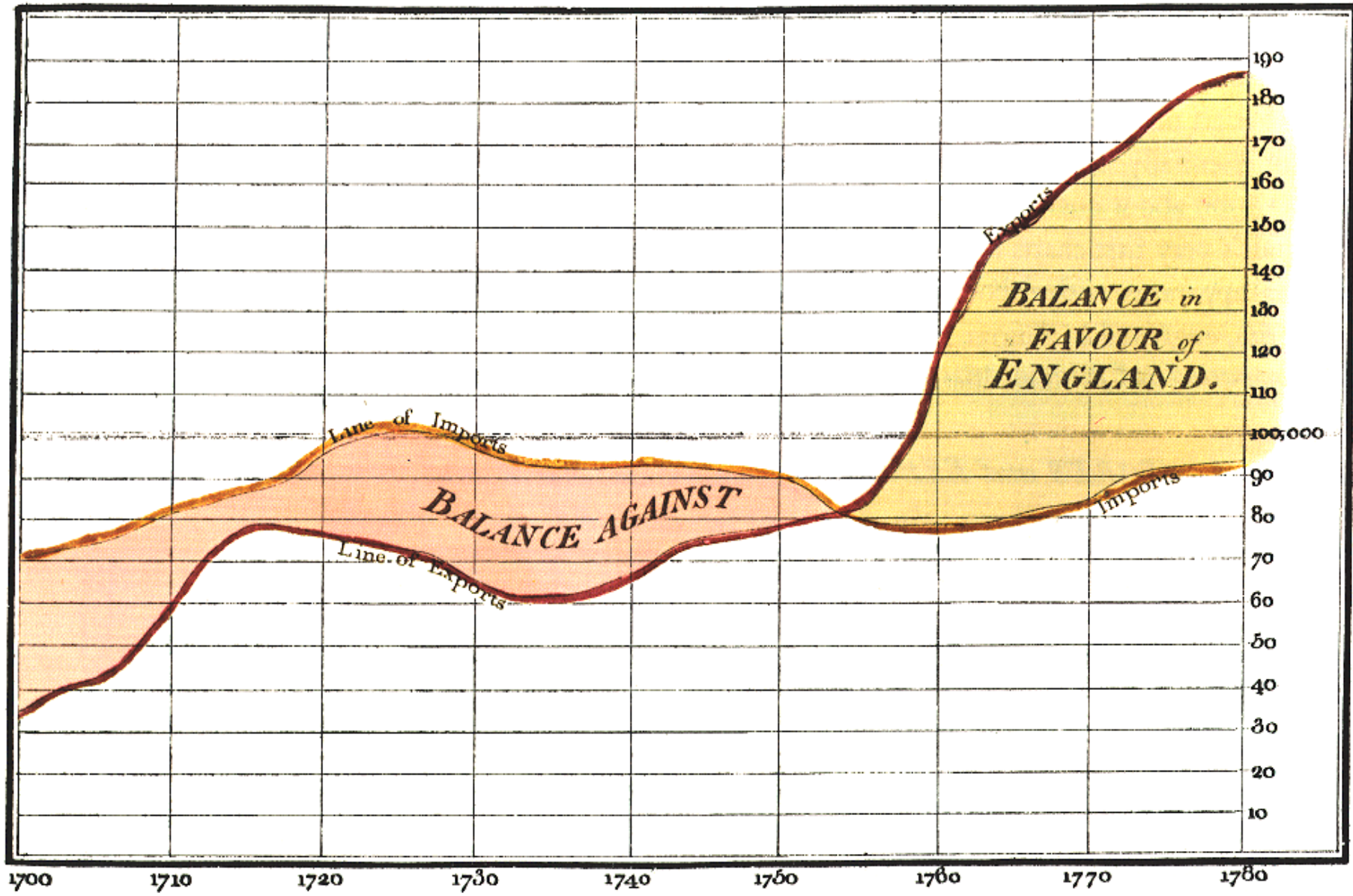


The **Golden Age** of Data Visualization

1786 1900

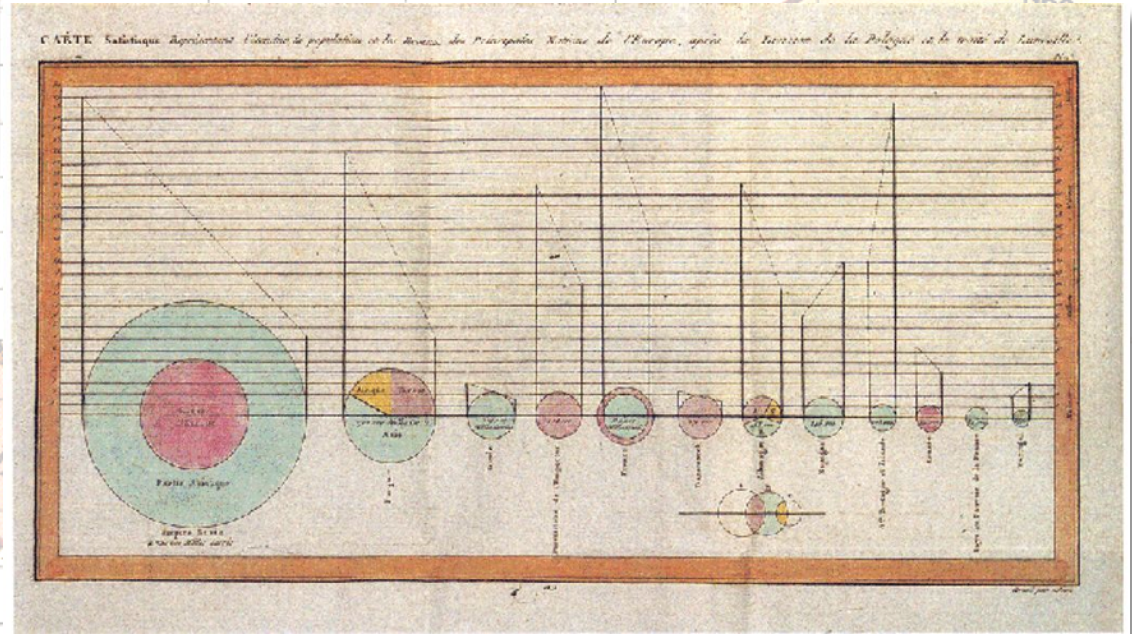
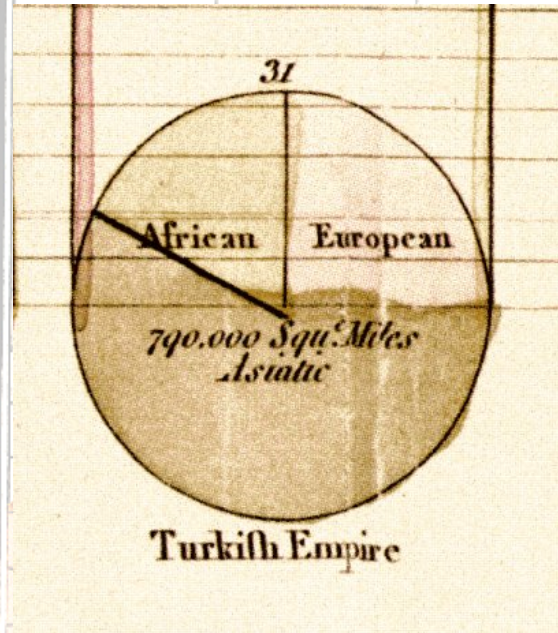
A horizontal white line at the bottom of the slide serves as a timeline. A small vertical tick mark is on the left. A red segment is located on the right side of the line, positioned between the years 1786 and 1900.

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



The Commercial and Political Atlas, William Playfair 1786

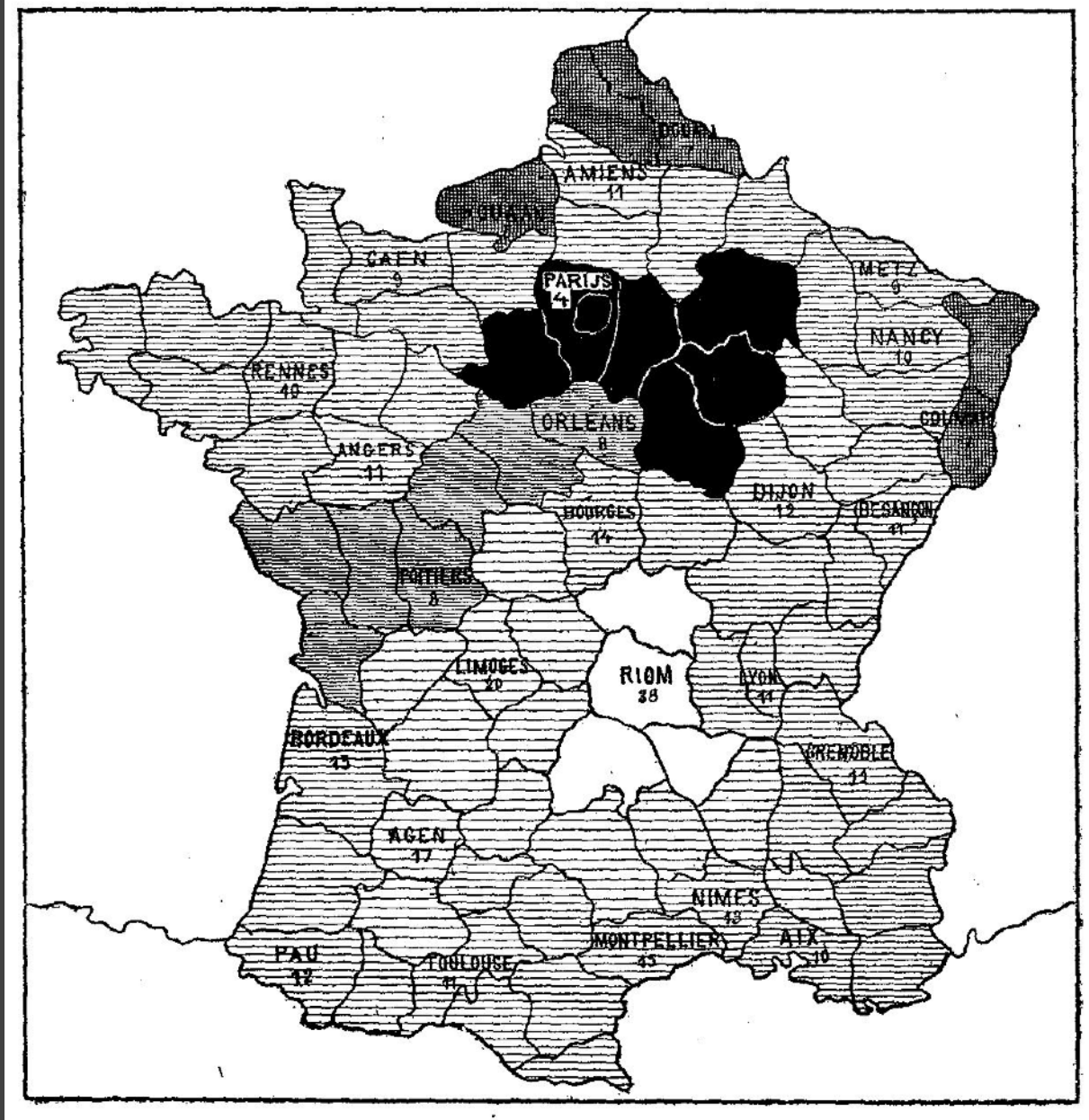
Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



1700 1710 1720 1730 1740 1750 1760 1770 1780

Statistical Breviary, William Playfair 1801





1786 1826(?) Illiteracy in France, Pierre Charles Dupin

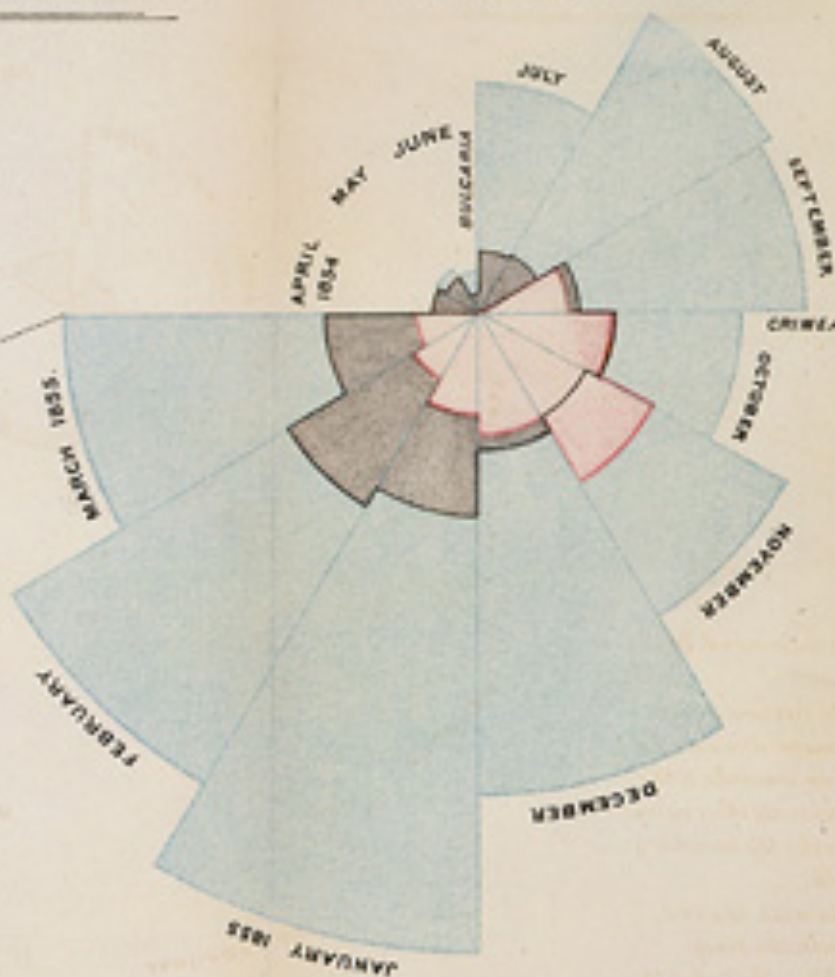


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.



1.
APRIL 1854 TO MARCH 1855.



“to affect thro’ the Eyes
what we fail to convey to
the public through their
word-proof ears”



CARTE répartition et approximations de la **houille Anglaise** exportée en 1864 dessinée par M^r MINARD, Ingénieur civil des Ponts et Chaussées en retraite.

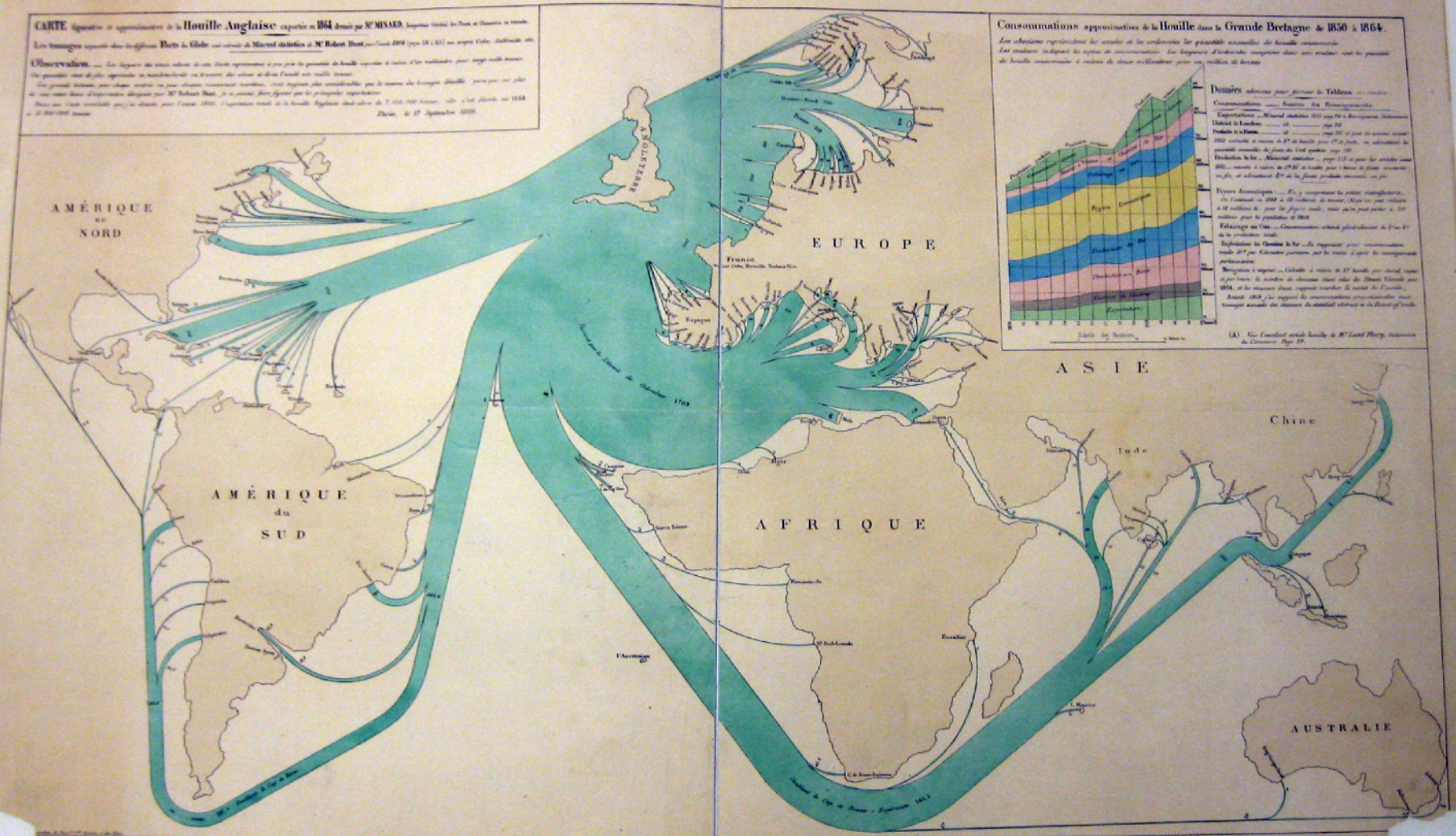
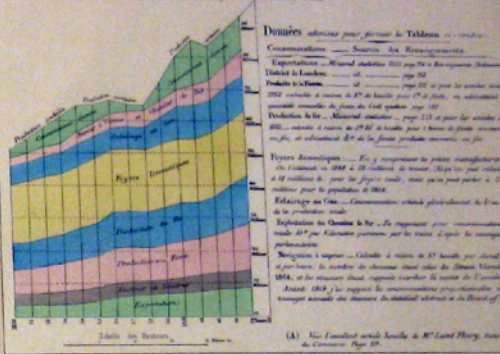
Les branches représentent dans les chiffres l'Écart au total de la houille anglaise à M^r Robert Hunt pour l'année 1864 (page 18 - 19) au moyen d'un tableau de M^r Robert Hunt.

Observation — Les lignes de même couleur de cette carte représentent à peu près la quantité de houille exportée et celle de son utilisation pour usage domestique. Les quantités sont de plus indiquées au maximum de la houille anglaise en France dans les chiffres au-dessus de l'île. Les chiffres indiqués au-dessous de l'île sont les quantités de houille anglaise en France dans les chiffres au-dessous de l'île. Les chiffres indiqués au-dessous de l'île sont les quantités de houille anglaise en France dans les chiffres au-dessous de l'île.

Paris, le 27 Septembre 1864.

Consommations approximatives de la Houille dans la Grande Bretagne & 1850 à 1864.

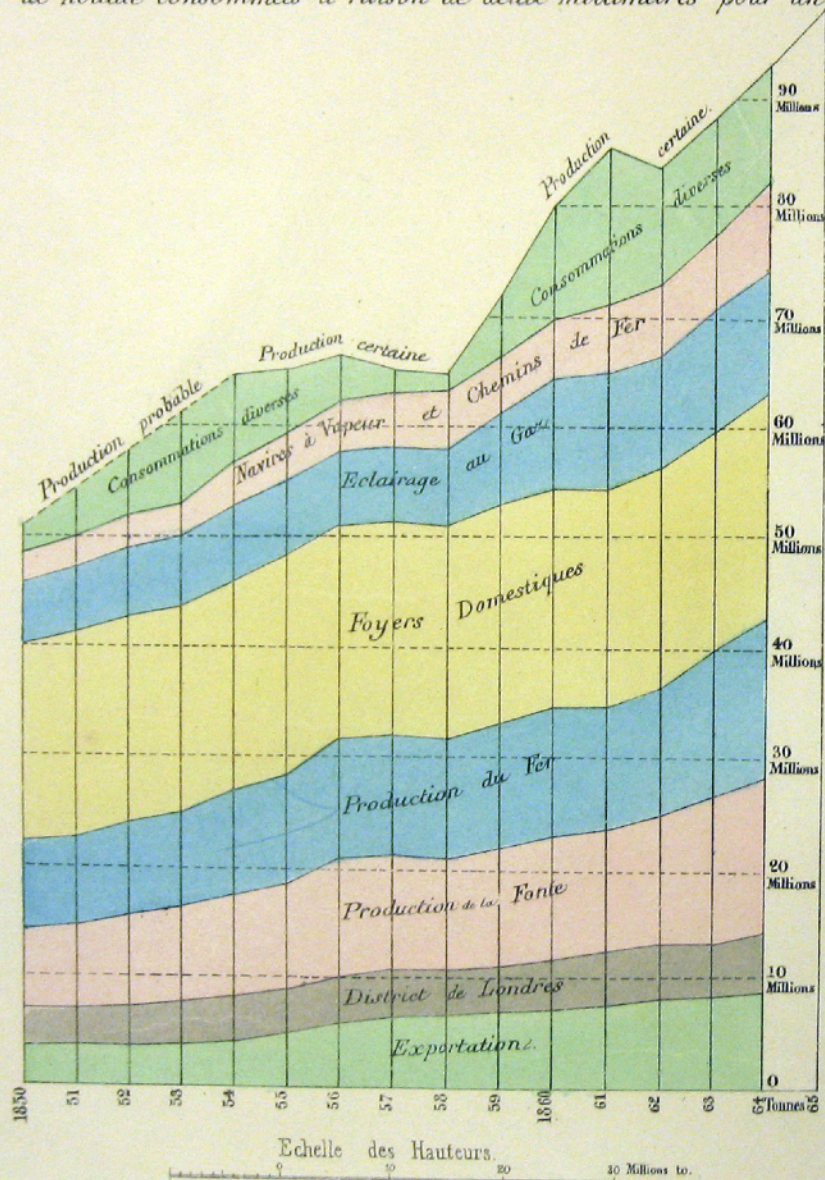
Les chiffres représentent les années et les lettres les quantités annuelles de houille consommée. Les chiffres indiquent le type de consommation. Les longueurs d'indicateurs comparés dans une échelle avec les quantités de houille consommée à raison de deux millions pour un million de tonnes.



Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.

Les abscisses représentent les années et les ordonnées les quantités annuelles de houille consommée.

Les couleurs indiquent les espèces de consommations. Les longueurs d'ordonnées comprises dans une couleur sont les quantités de houille consommées à raison de deux millimètres pour un million de tonnes.



Données admises pour former le Tableau ci-contre.

Consommations. — Sources des Renseignements.

Exportations. — *Mineral statistics 1865 page 214 et Renseignements Parlementaires.*

District de Londres. — *id.* — page 213

Produits de la Fonte. — *id.* — page 215 et pour les années avant 1855 calculée à raison de 3^{tes} de houille pour 1^{re} de fonte, en admettant les quantités annuelles de fonte du Coal question page 192.

Production du fer — *Mineral statistics* — page 215 et pour les années avant 1855 — calculée à raison de 3^{tes} 35 de houille pour 1 tonne de fonte convertie en fer, et admettant $\frac{2}{10}$ de la fonte produite convertis en fer.

Foyers domestiques. — En y comprenant les petites manufactures.

On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.

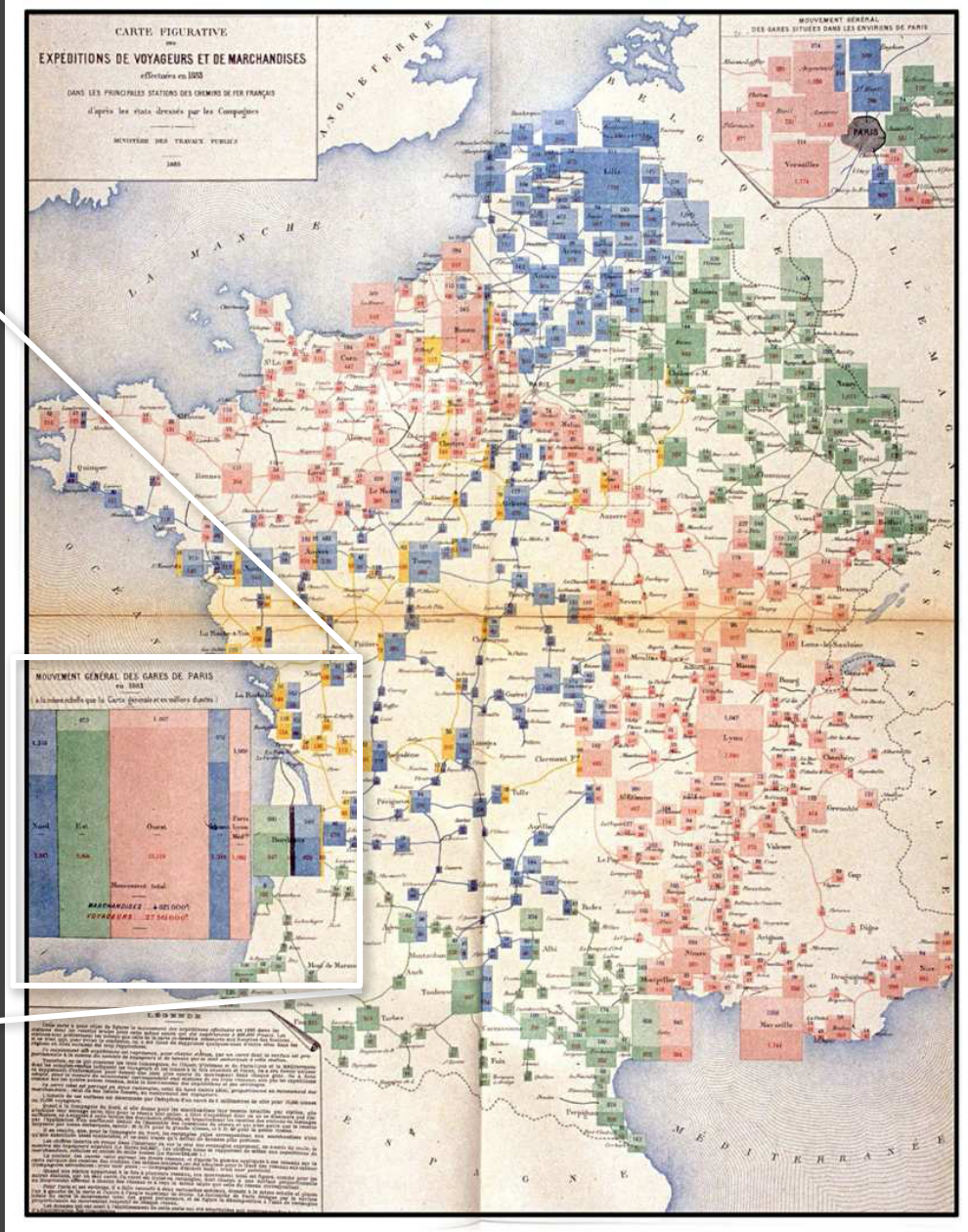
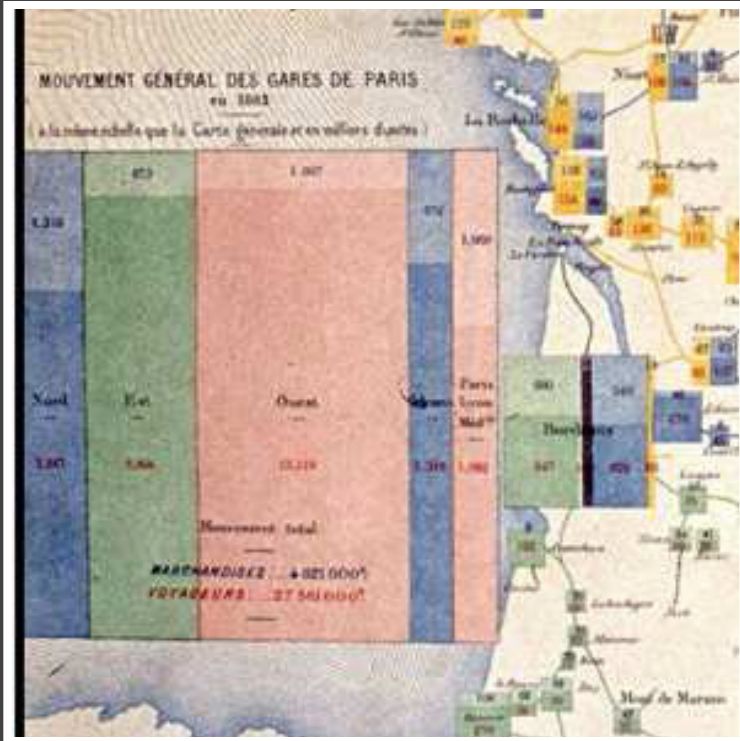
Eclairage au Gaz. — Consommation estimée généralement du $\frac{1}{3}$ au $\frac{1}{8}$ de la production totale.

Exploitation des Chemins de Fer. — En supposant pour consommation totale 10^{tes} par Kilomètre parcouru par les trains d'après les renseignements parlementaires.

Navigation à vapeur. — Calculée à raison de 5^{tes} houille par cheval vapeur et par heure, le nombre de chevaux étant celui du Steam Vessels pour 1864, et les steamers étant supposés marcher la moitié de l'année;

Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of trade.

(A) Voir l'excellent article houille de M.^r Lamé Fleury, Dictionnaire du Commerce Page III.



1786

1884 Rail Passengers and Freight from Paris

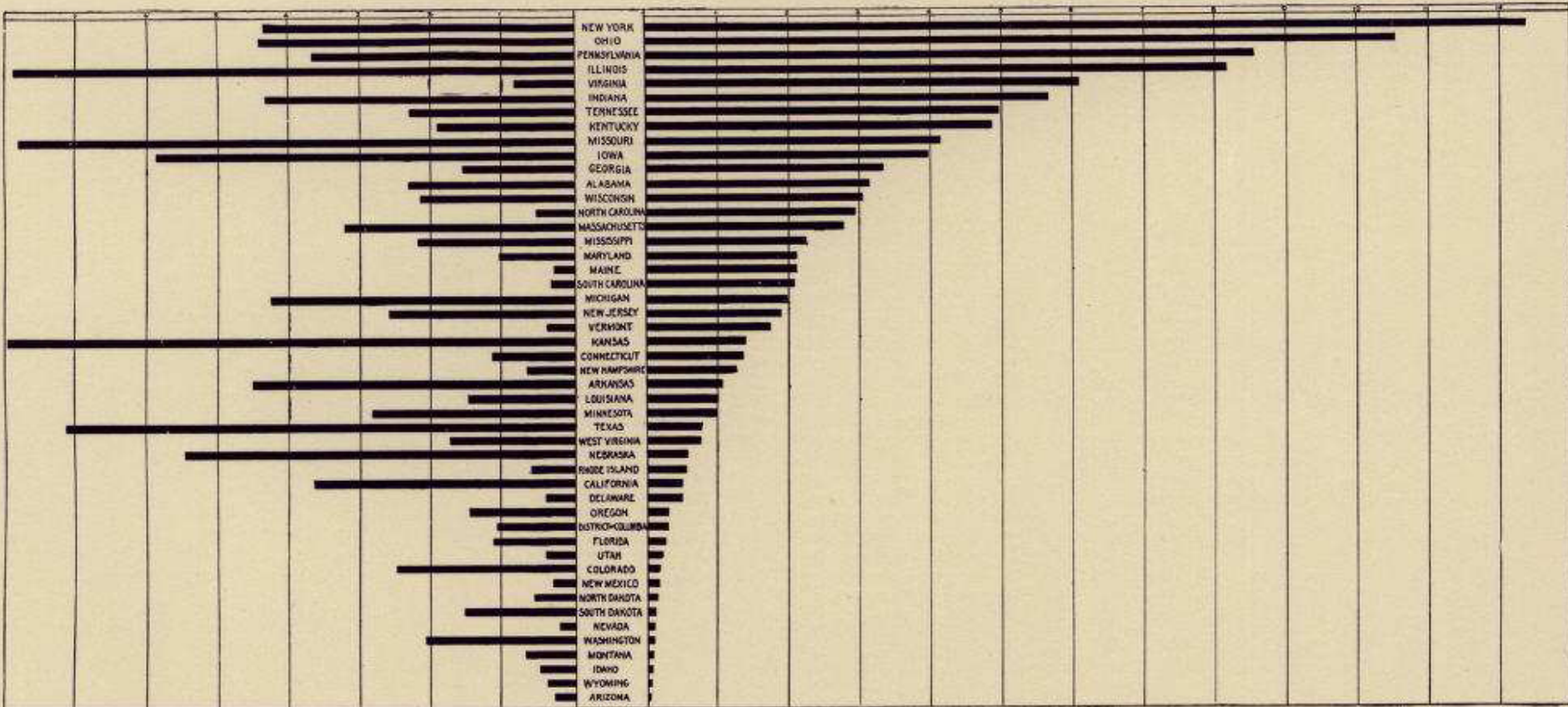


66. INTERSTATE MIGRATION—NUMBER OF NATIVE IMMIGRANTS AND NATIVE EMIGRANTS, BY STATES AND TERRITORIES: 1890.

Native immigrants.

[Hundreds of thousands.]

Native emigrants.



The Rise of Statistics

1786



1900



1950

Rise of **formal methods** in statistics and social science – Fisher, Pearson, ...

Little innovation in graphical methods

A period of **application and popularization**

Graphical methods enter textbooks, curricula, and **mainstream use**

1786

1900

1950





LIFE

1786

Data Analysis & Statistics, Tukey 1962





Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.

1786

Data Analysis & Statistics, Tukey 1962





The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.

1786

Data Analysis & Statistics, Tukey 1962





While some of the influences of statistical theory on data analysis have been helpful, others have not.

1786

Data Analysis & Statistics, Tukey 1962





Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.





Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind.**

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention.**

1786

Data Analysis & Statistics, Tukey 1962



Set A

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$$u_X = 9.0 \quad \sigma_X = 3.317$$

$$u_Y = 7.5 \quad \sigma_Y = 2.03$$

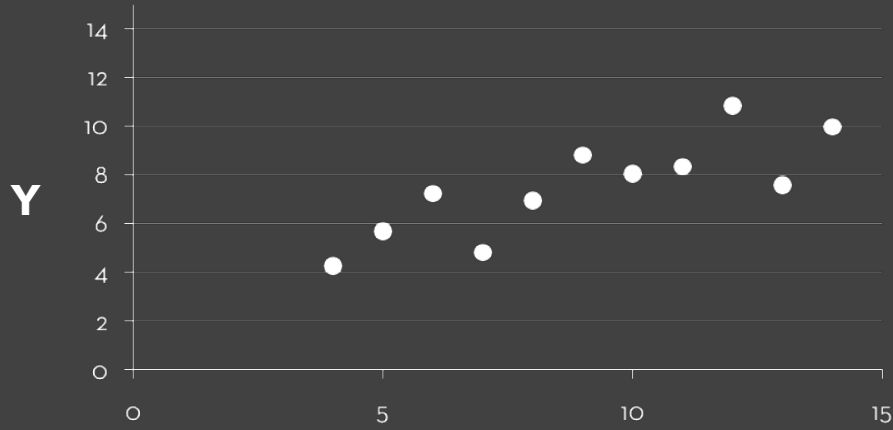
Linear Regression

$$Y = 3 + 0.5 X$$

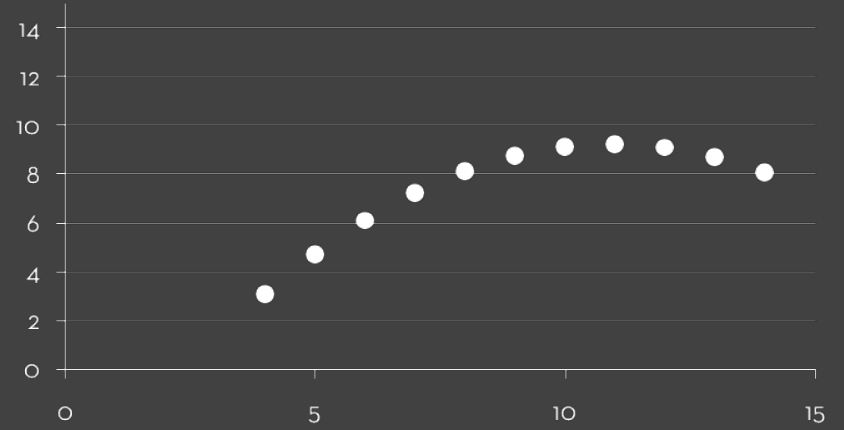
$$R^2 = 0.67$$

[Anscombe 1973]

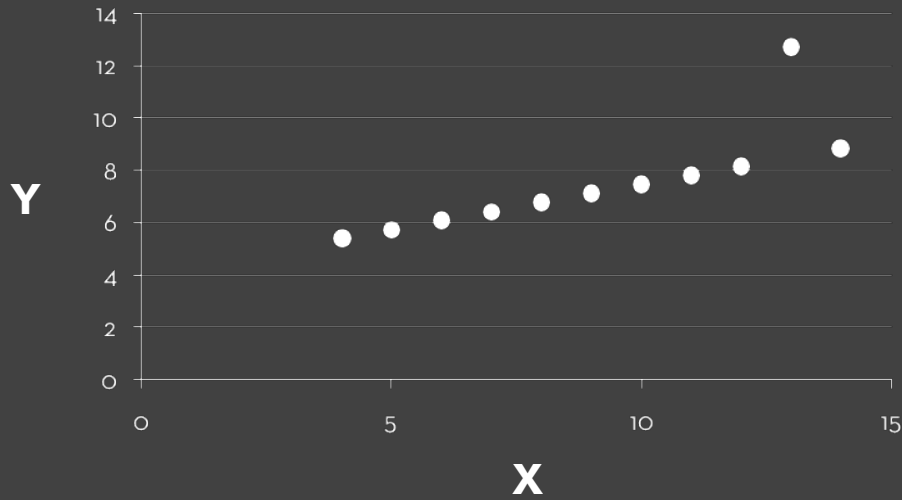
Set A



Set B



Set C



Set D



Topics: Exploratory Data Analysis

A Brief History of Data Visualization

Data Wrangling

Exploratory Analysis Examples

Polaris / Tableau

Data Wrangling

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.

Anonymous Data Scientist

[Kandel et al. VAST '12]





**Big Data
Borat**

@BigDataBorat



Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.



Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

DataWrangler

The screenshot displays the DataWrangler interface. On the left, a 'Suggestions' panel lists four data transformation options: 'Delete rows 8,10', 'Delete empty rows', 'Delete rows where Property_crime_rate is null', and 'Delete rows where Year is null'. Below this is a 'Script' panel with two suggestions: 'Split data repeatedly on newline into rows' and 'Split data repeatedly on \','. An 'Export' button is visible next to the Script header. On the right, a data table is shown with 408 rows. The table has two columns: '# Year' and '# Property_crime_rate'. The data is grouped into sections: 'Reported crime in Alabama' (rows 1-7), 'Reported crime in Alaska' (rows 9-14), and a section for 'Reported crime in California' (rows 11-14) which is partially obscured by the suggestion list. The table shows years from 2004 to 2007 and corresponding property crime rates.

#	Year	#	Property_crime_rate
1	Reported crime in Alabama		
2			
3	2004		4029.3
4	2005		3900
5	2006		3937
6	2007		3974.9
7	2008		4081.9
8			
9	Reported crime in Alaska		
10			
11	2004		3370.9
12	2005		3615
13	2006		3582
14	2007		3373.9

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel et al. *CHI'11*

Data Wrangling

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration.

Approaches include:

Manual manipulation in spreadsheets

Code (e.g., dplyr in R, Pandas in Python)

Trifacta Wrangler <http://www.trifacta.com/products/wrangler/>

Open Refine <http://openrefine.org/>

Data Quality

“The first sign that a visualization is good is that it shows you a problem in your data...

...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

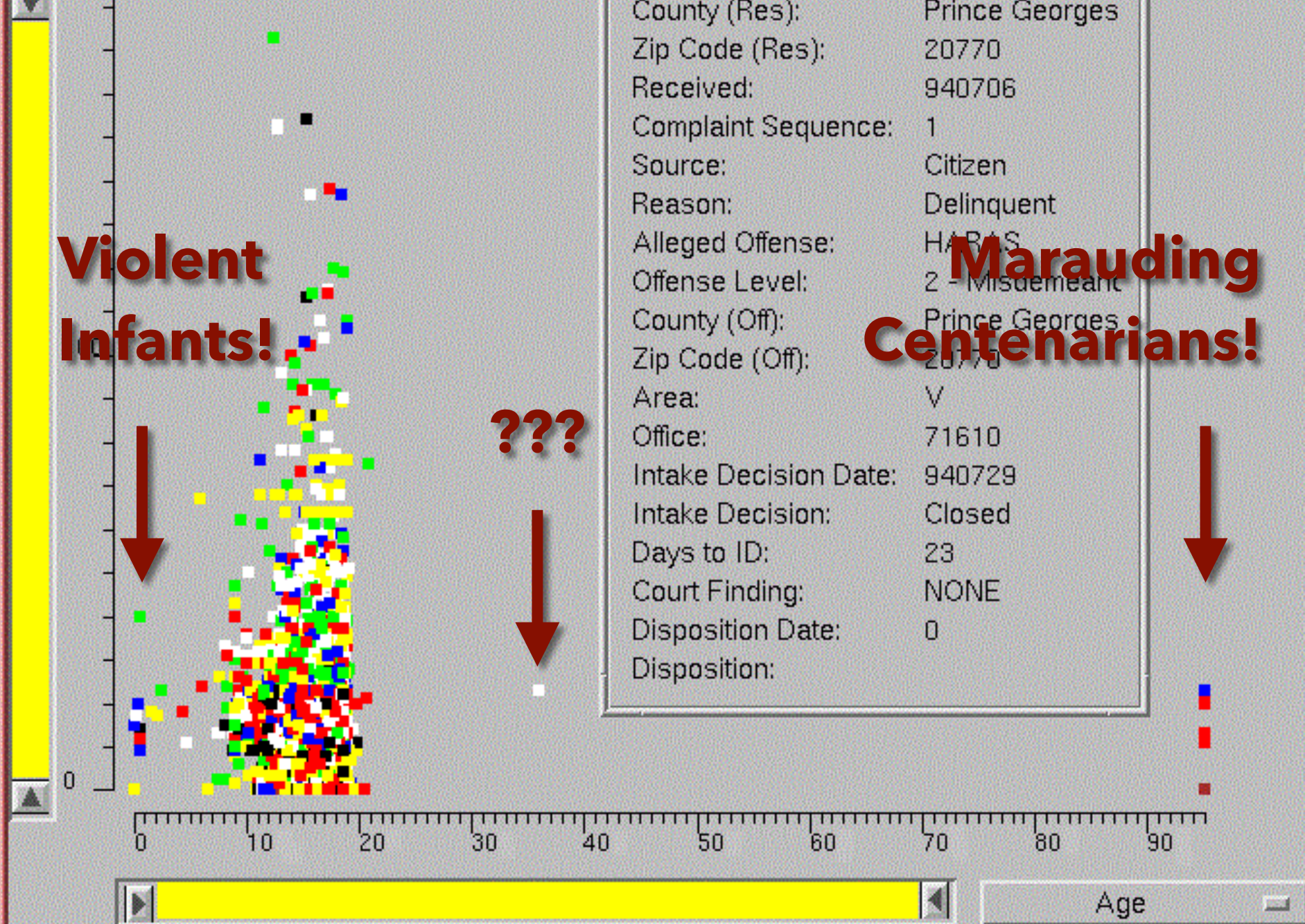
Martin Wattenberg

County (Res):	Prince Georges
Zip Code (Res):	20770
Received:	940706
Complaint Sequence:	1
Source:	Citizen
Reason:	Delinquent
Alleged Offense:	HARAS
Offense Level:	2 - Misdemeanor
County (Off):	Prince Georges
Zip Code (Off):	20770
Area:	V
Office:	71610
Intake Decision Date:	940729
Intake Decision:	Closed
Days to ID:	23
Court Finding:	NONE
Disposition Date:	0
Disposition:	

**Violent
Infants!**

**Marauding
Centenarians!**

???



Query Result: 4792 out of 4792 (100%)

Graph Viewer

Roll-up by:

All

Visualization:

Node-Link

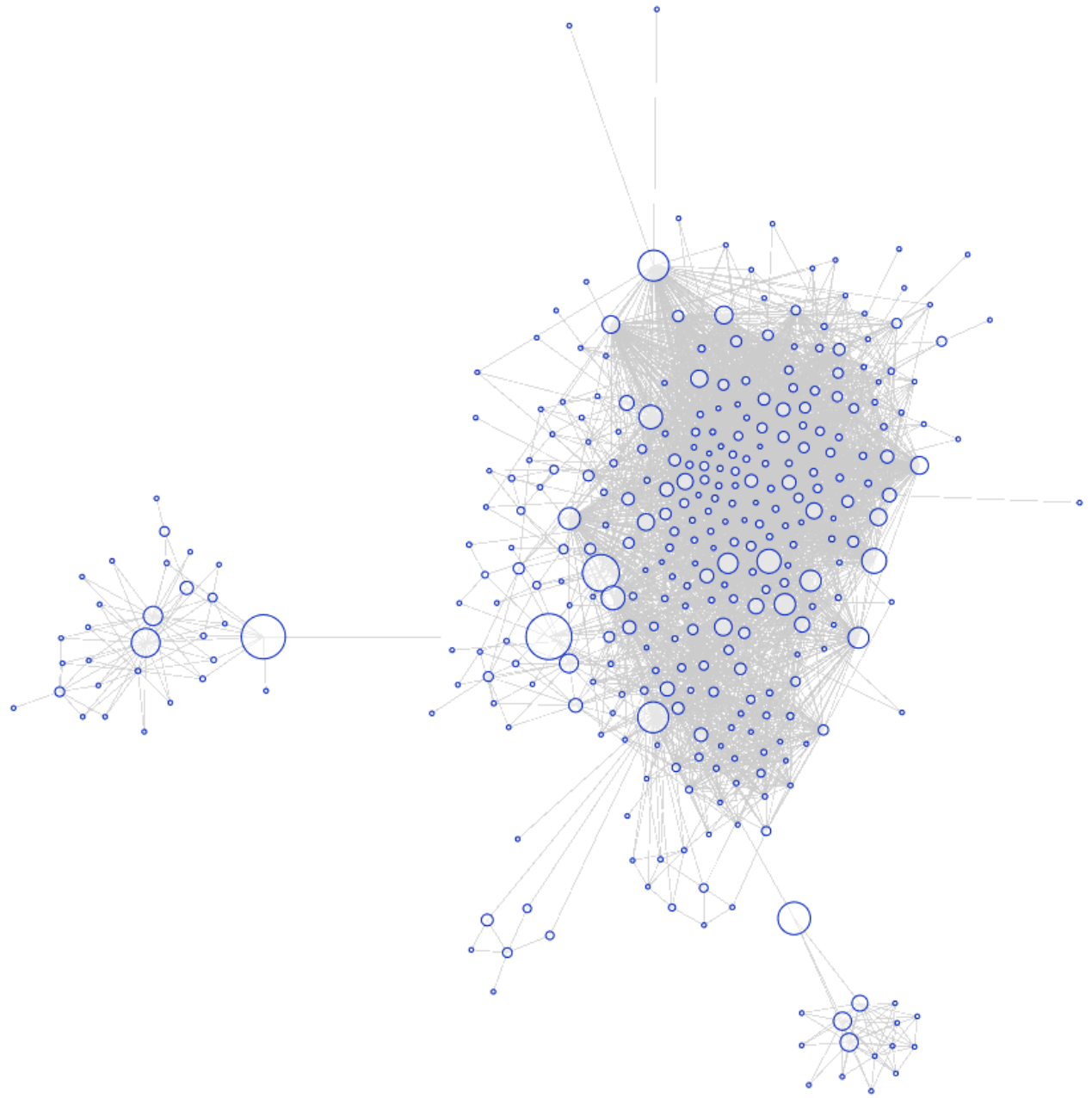
Sort by:

None

Edge centrality filters:

Two horizontal sliders for edge centrality filtering, both currently at the minimum value.

- Images
- Animate



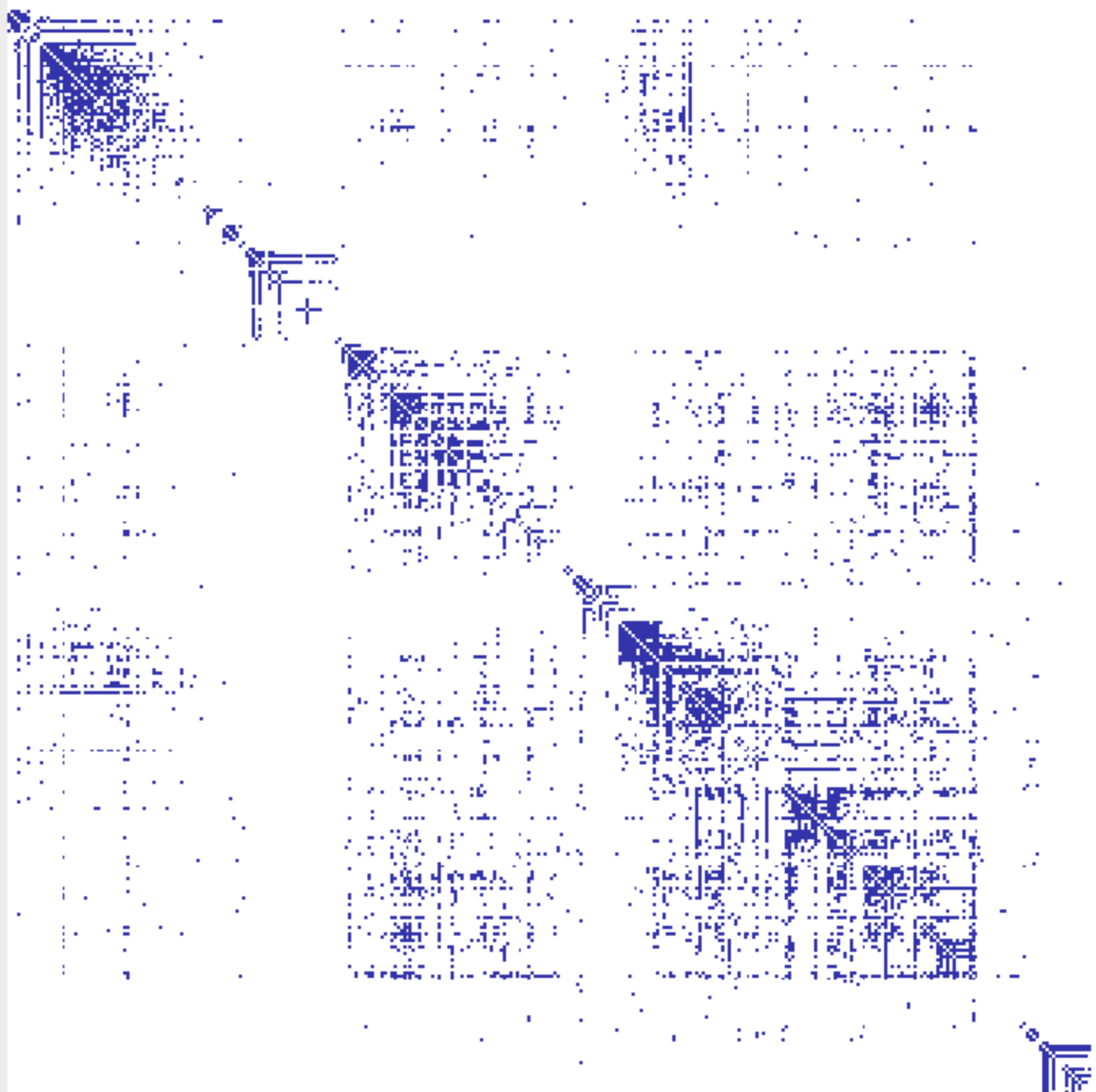
Graph Viewer

Roll-up by:

Visualization:

Sort by:

Edge centrality filters:



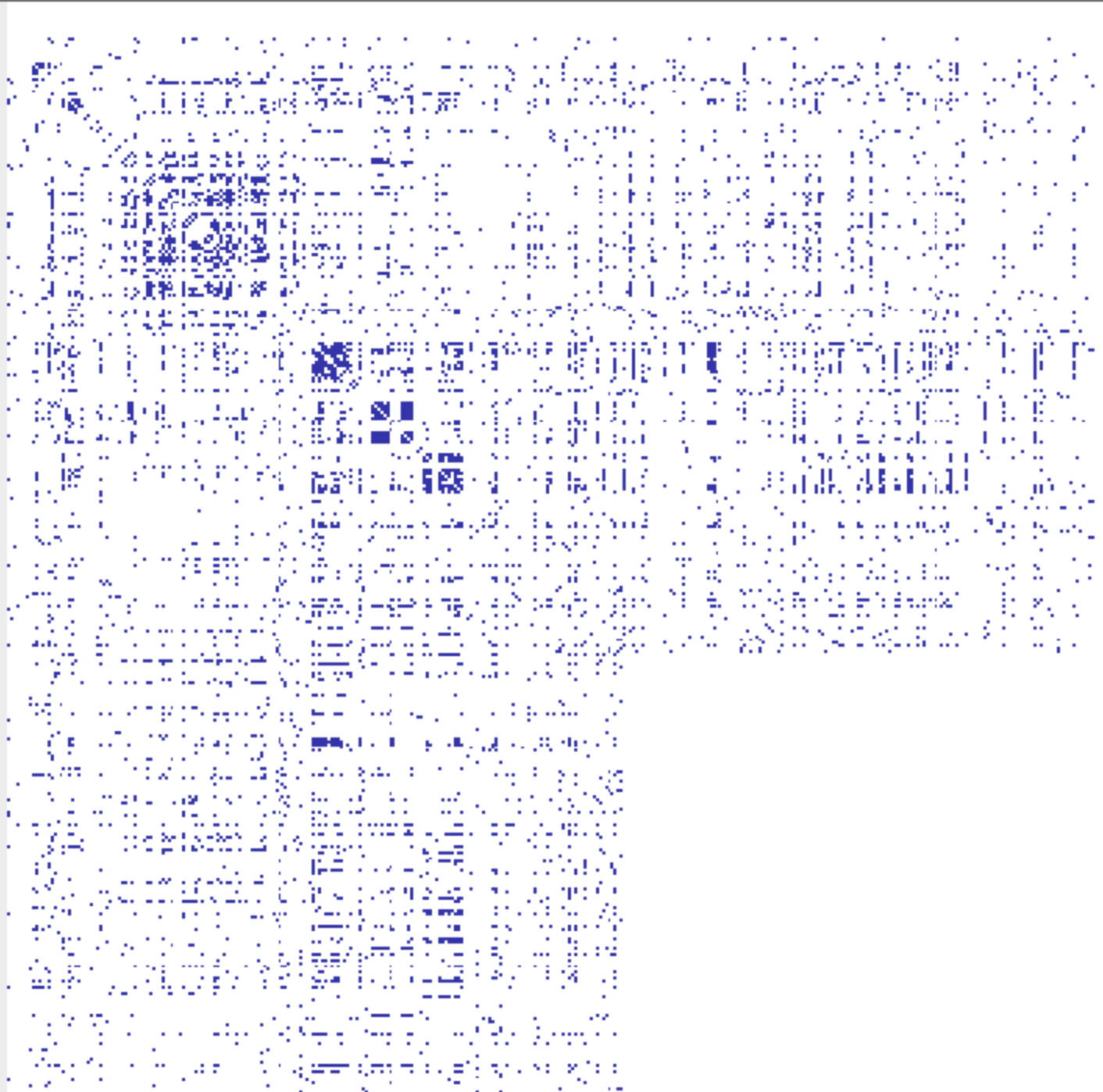
Graph Viewer

Roll-up by:

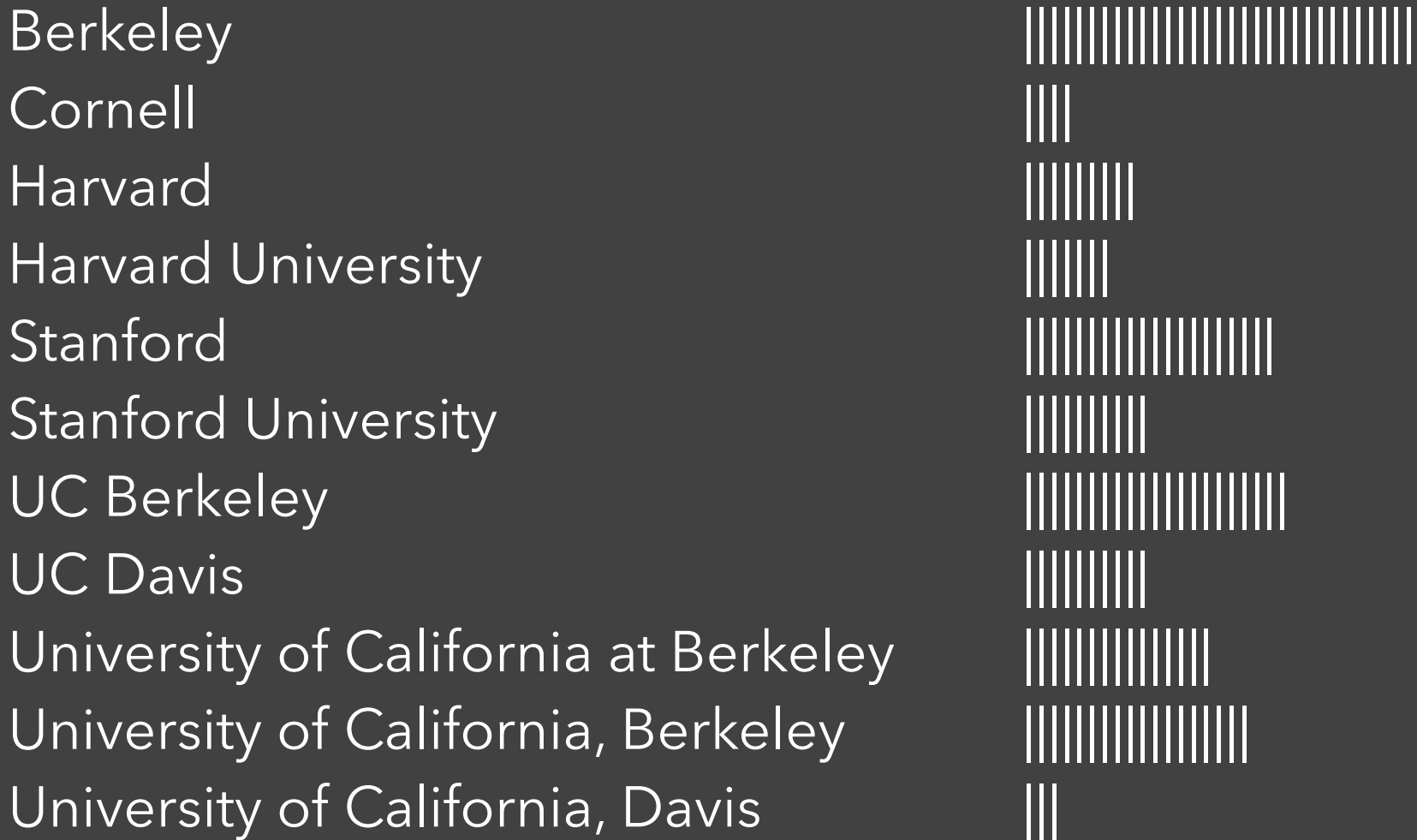
Visualization:

Sort by:

Edge centrality filters:



Visualize Friends by School?



Data Quality Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

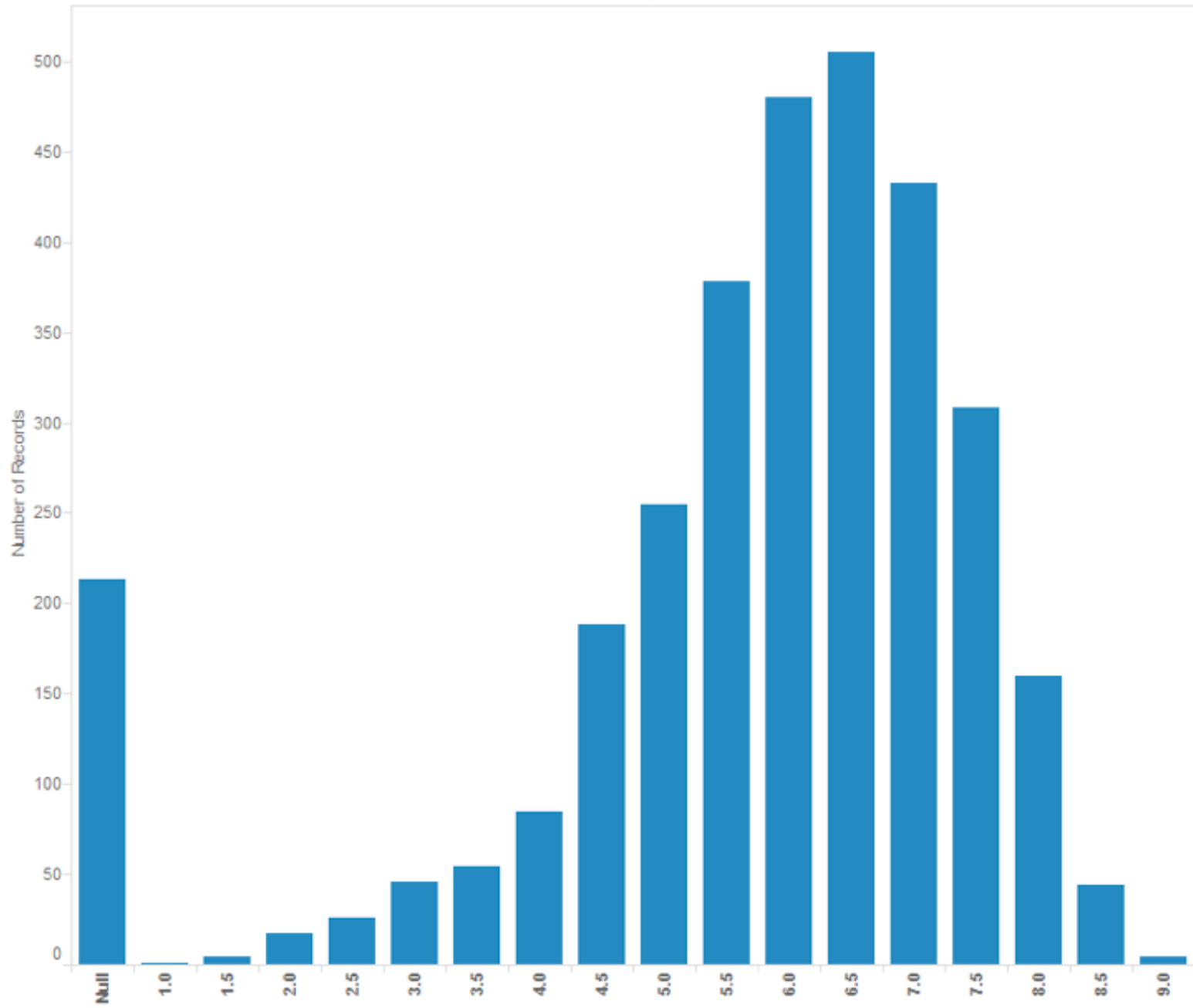
LESSON: Anticipate problems with your data.
Many research problems around these issues!

Analysis Example: Motion Pictures Data

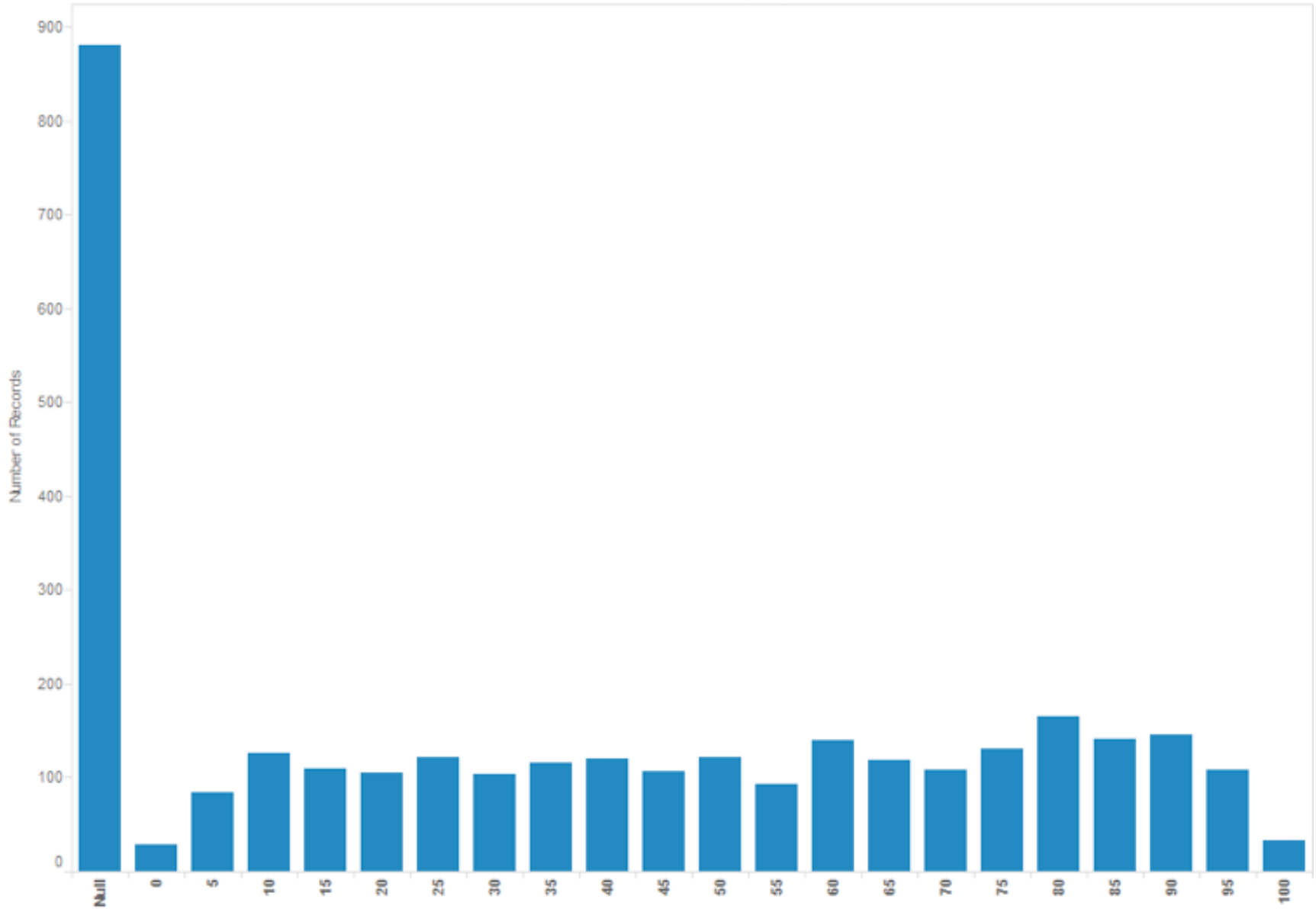
Motion Pictures Data

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

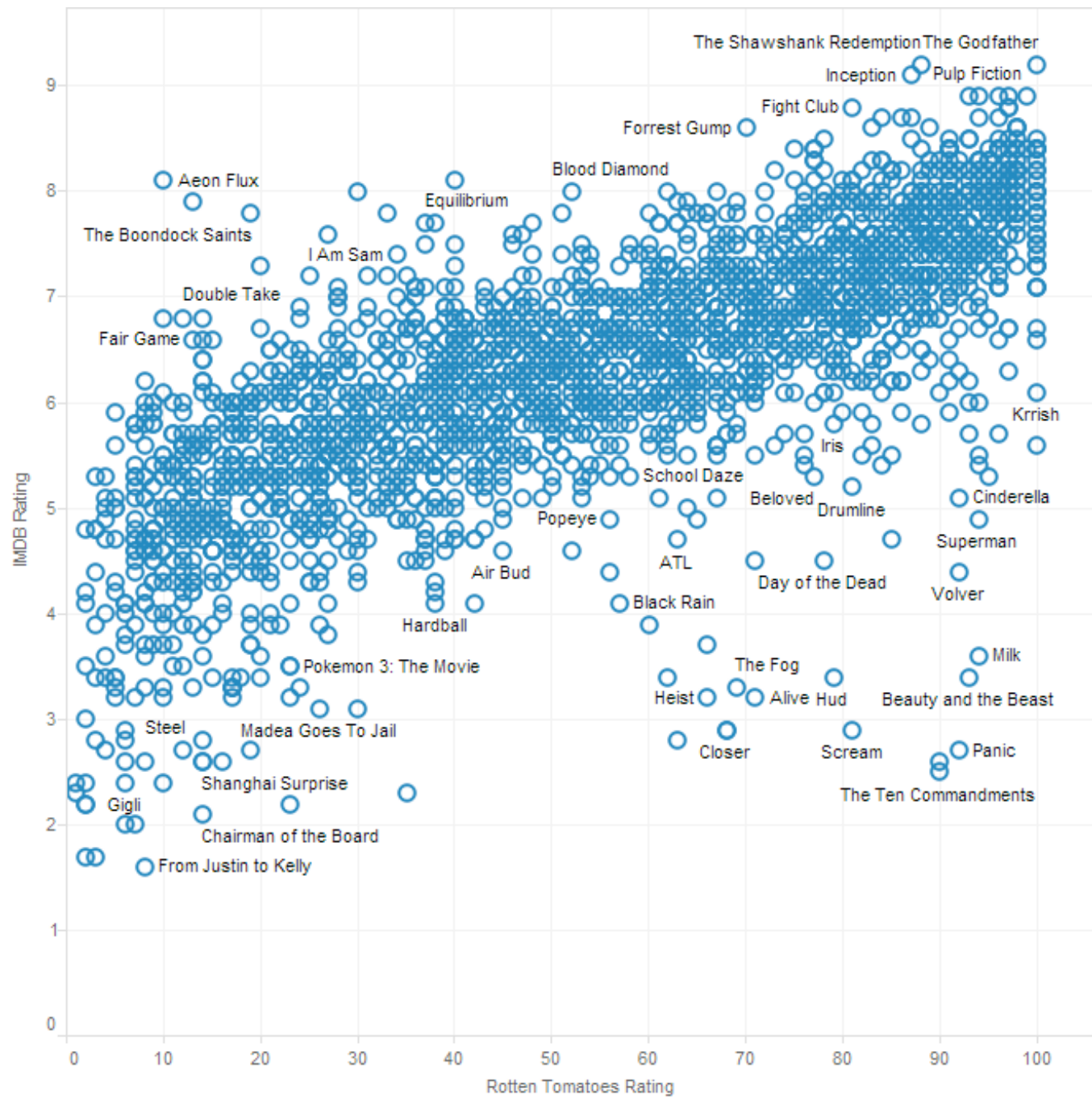
IMDB Rating (bin)

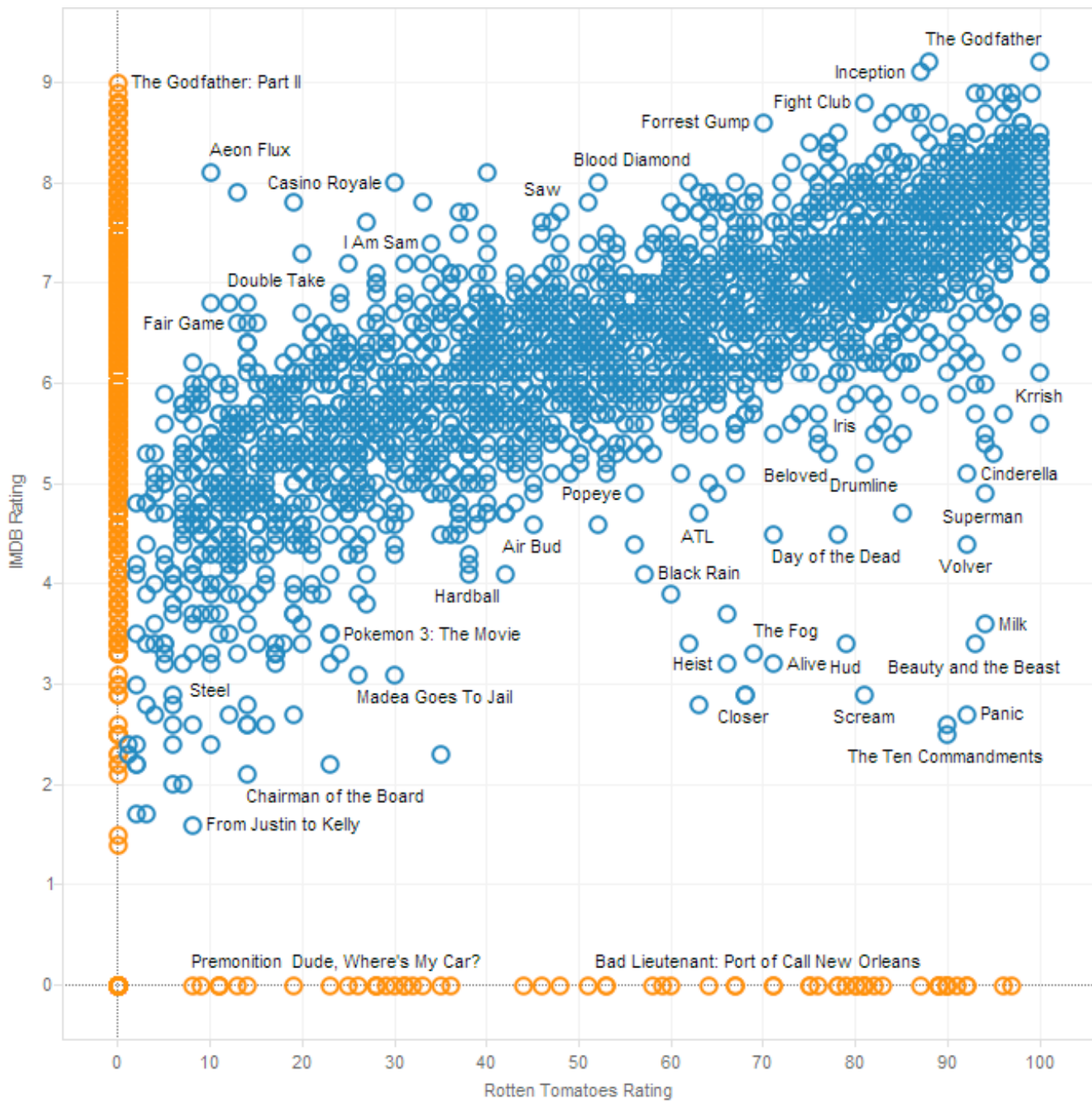


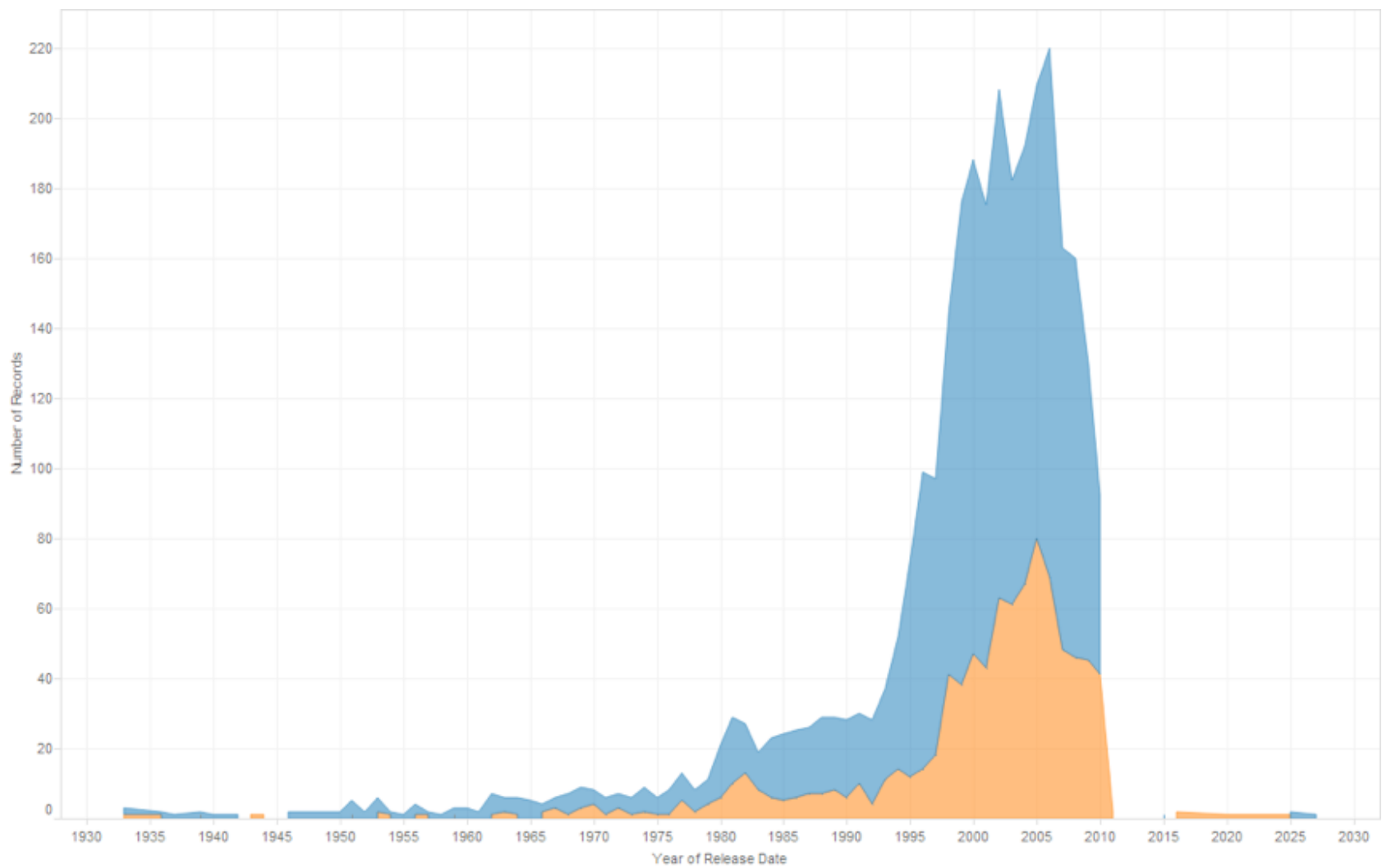
Rotten Tomatoes Rating (bin)











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)

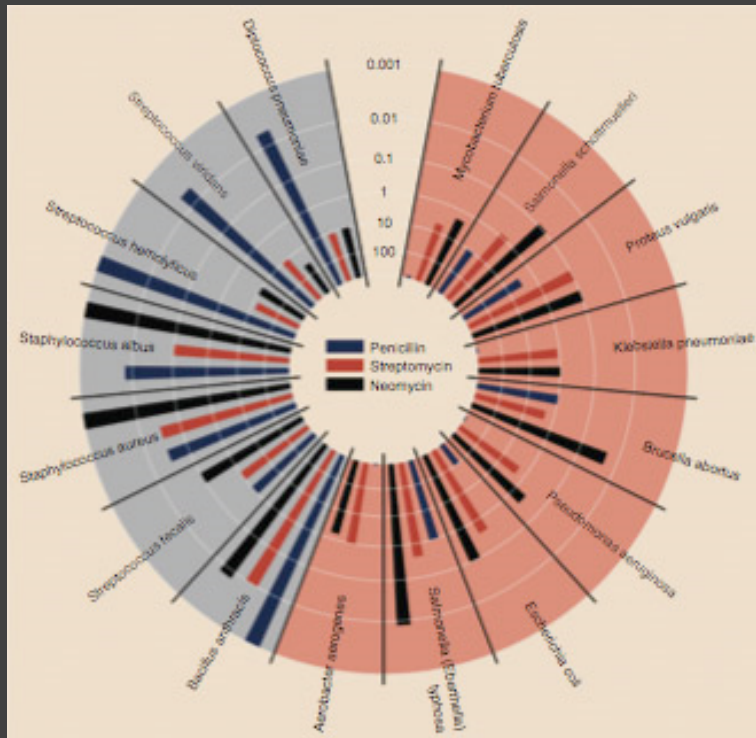
Collected prior to 1951.

What questions might we ask?

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
<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

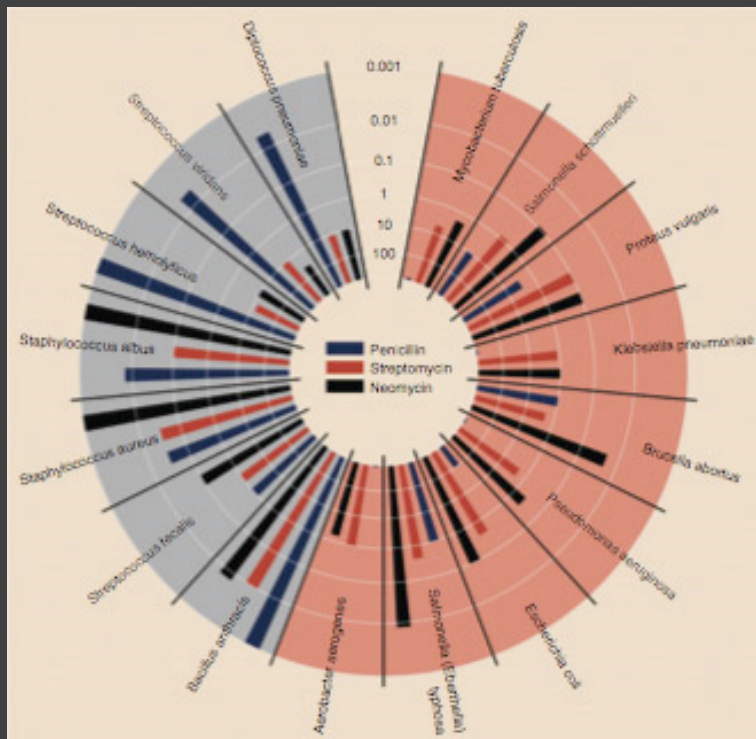
How do the drugs compare?



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

Original graphic by Will Burtin, 1951

How do the drugs compare?



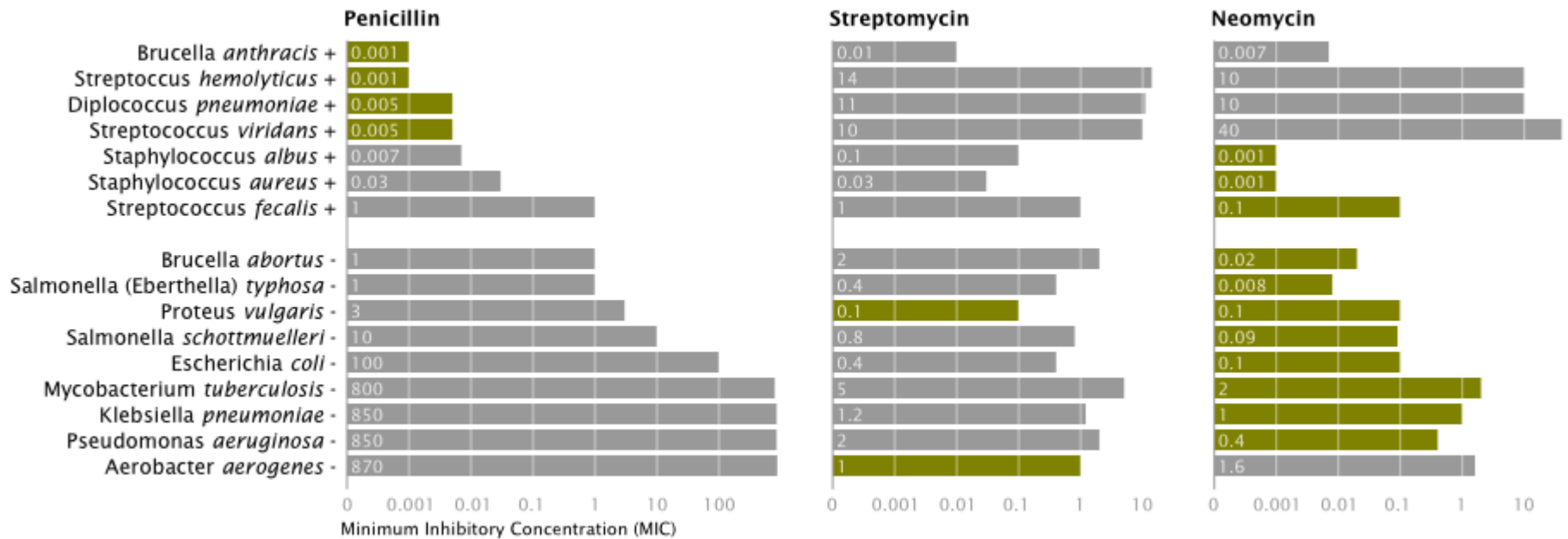
Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
<i>Aerobacter aerogenes</i>	870	1	1.6	-
<i>Brucella abortus</i>	1	2	0.02	-
<i>Bacillus anthracis</i>	0.001	0.01	0.007	+
<i>Diplococcus pneumoniae</i>	0.005	11	10	+
<i>Escherichia coli</i>	100	0.4	0.1	-
<i>Klebsiella pneumoniae</i>	850	1.2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

Radius: $1 / \log(\text{MIC})$

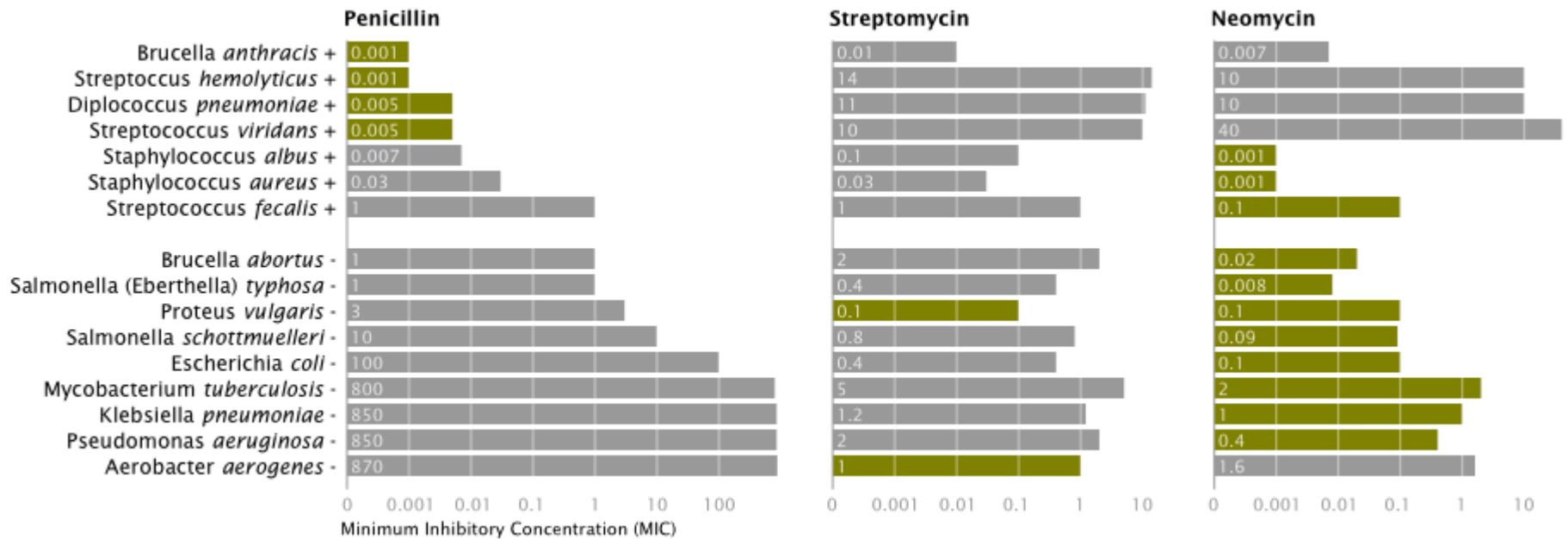
Bar Color: Antibiotic

Background Color: Gram Staining

How do the drugs compare?



How do the drugs compare?



X-axis: Antibiotic | $\log(\text{MIC})$

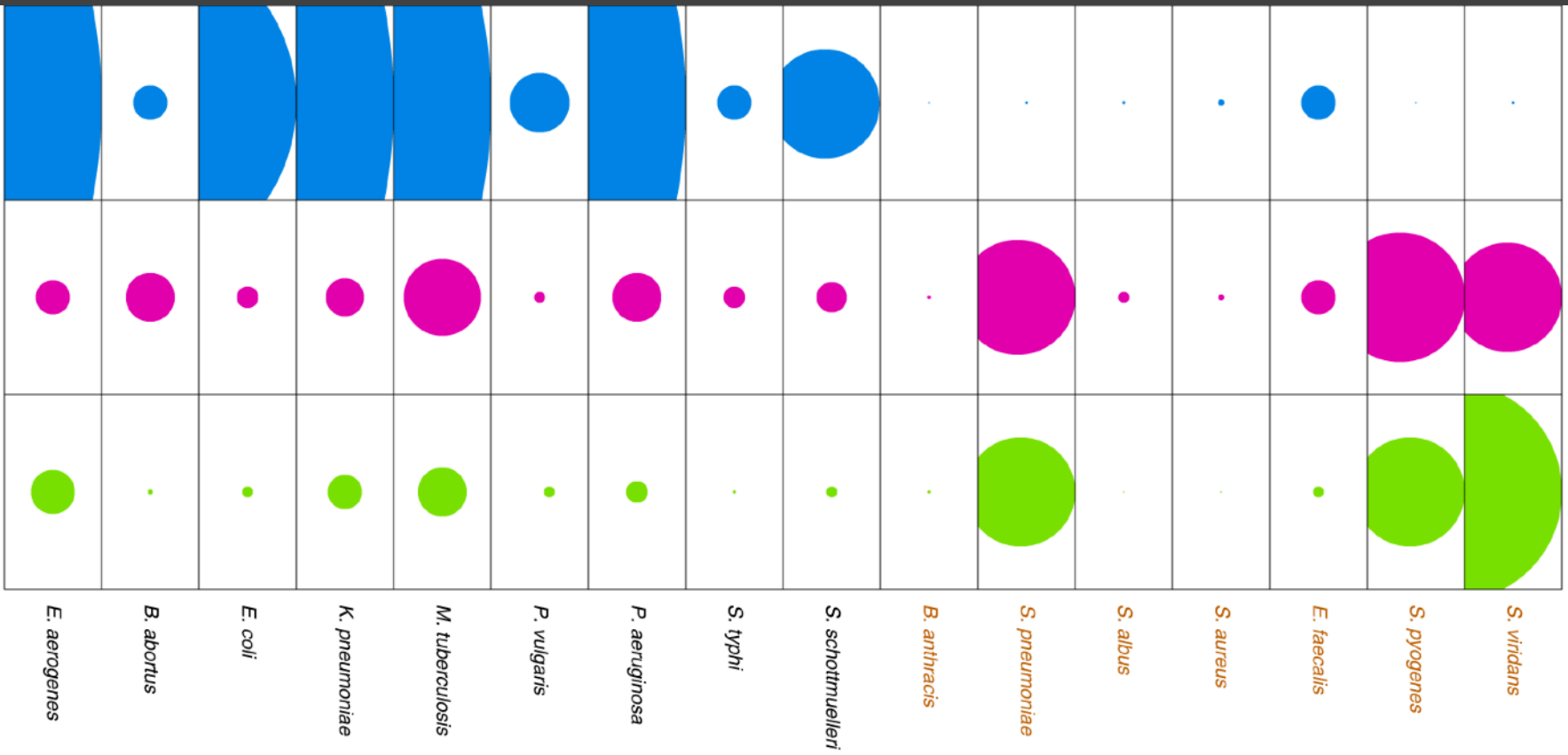
Y-axis: Gram-Staining | Species

Color: Most-Effective?

penicillin

streptomycin

neomycin

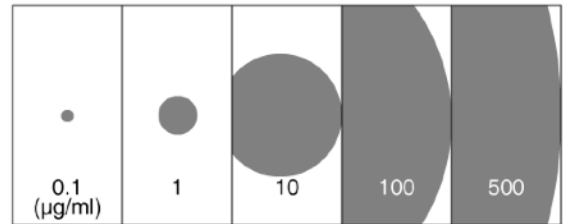


Gram positive

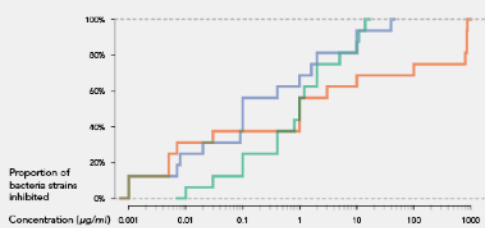
Gram negative

minimum inhibitory concentration of antibiotics

bowen li cs448b

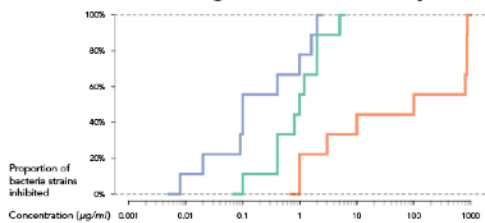


All bacteria



Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicillin.

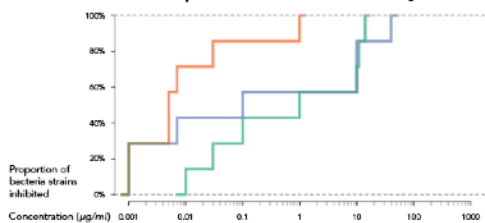
Gram-negative bacteria only



Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

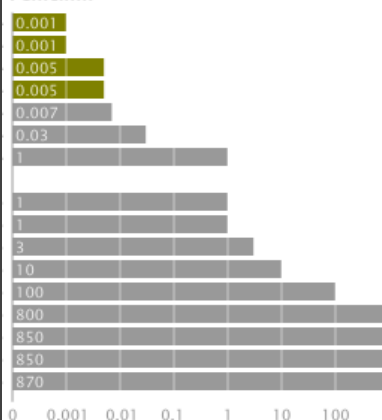
Gram staining quickly identifies bacteria as Gram-negative or Gram-positive, which can be used to find a more efficient antibiotic and dosage.

Gram-positive bacteria only

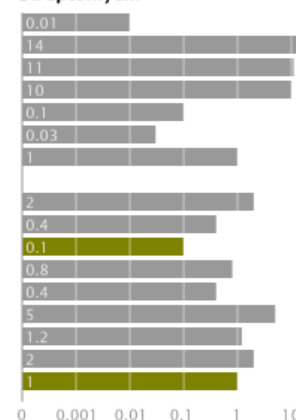


Penicillin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.

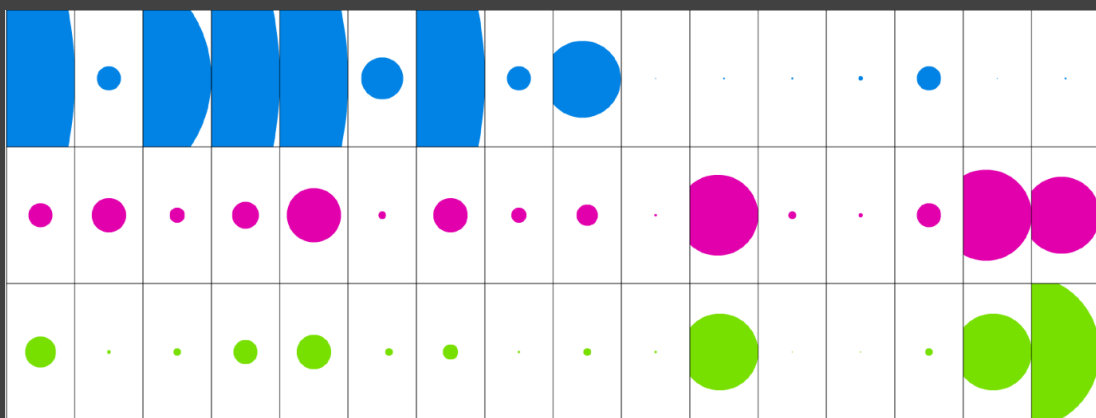
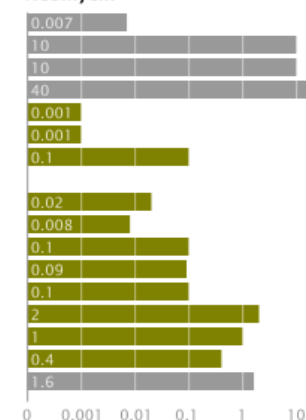
Penicillin



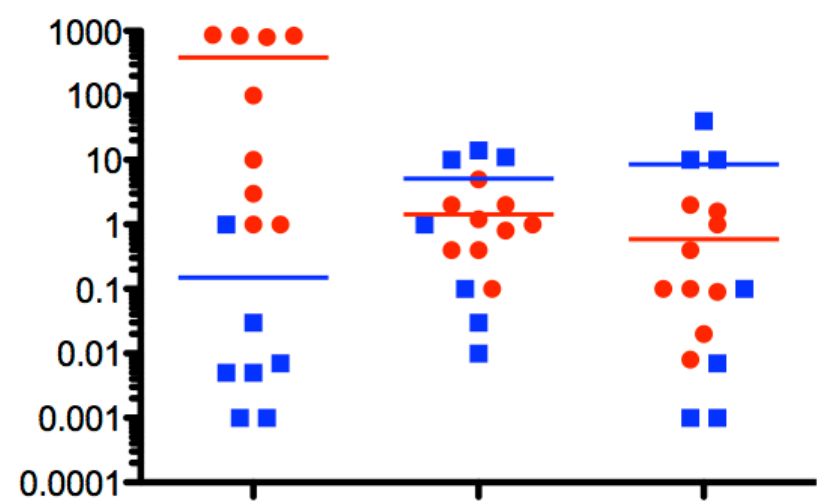
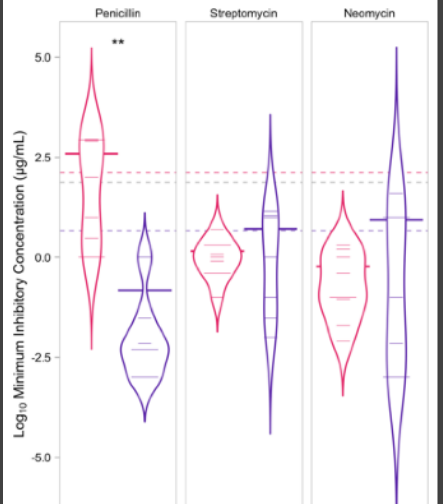
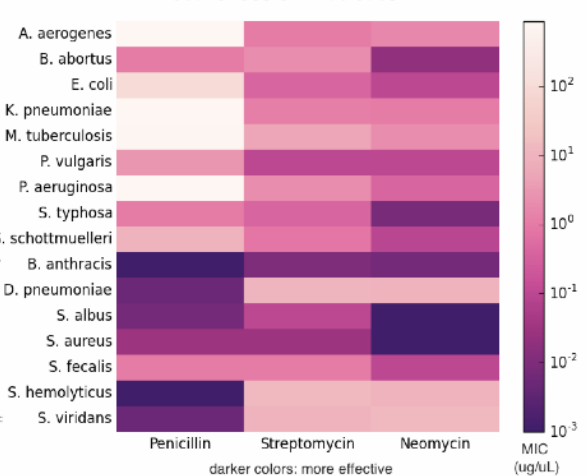
Streptomycin

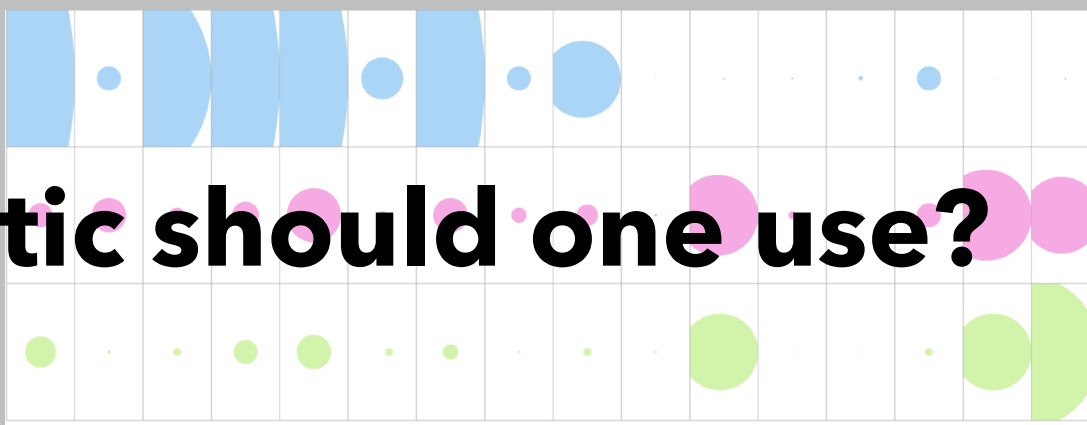
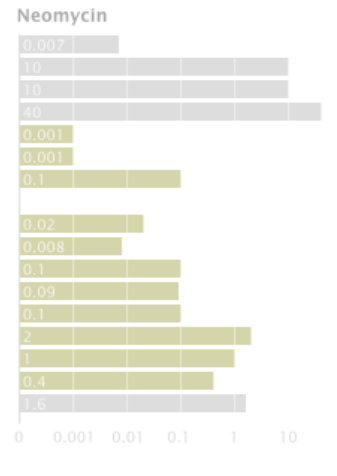
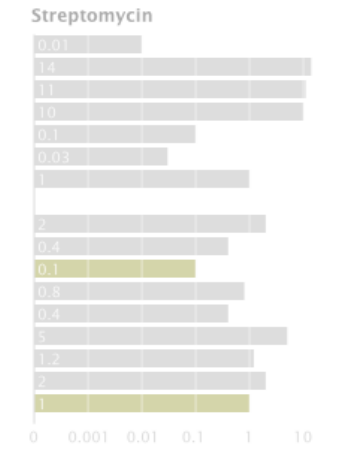
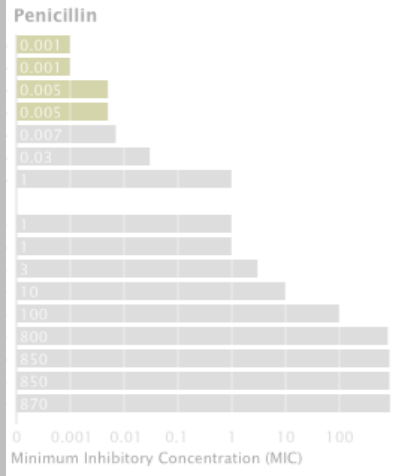
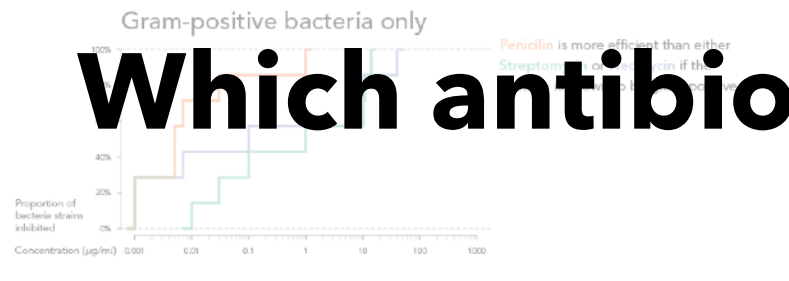
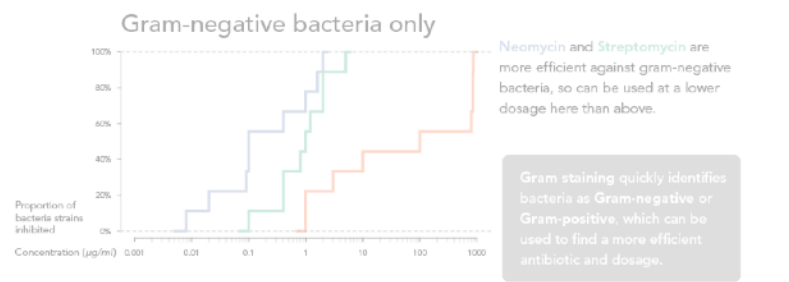
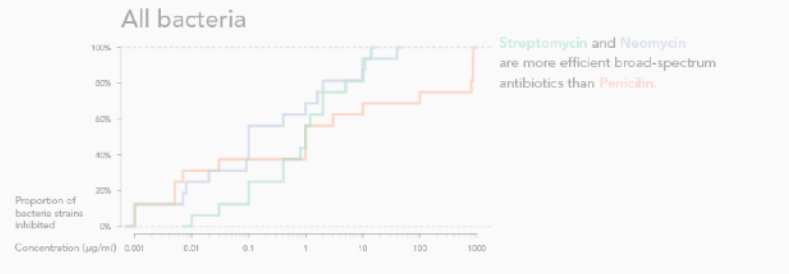


Neomycin

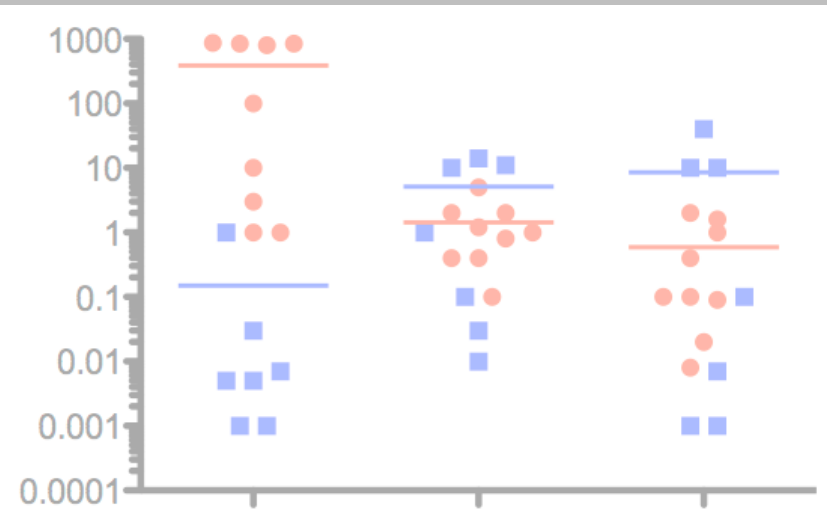
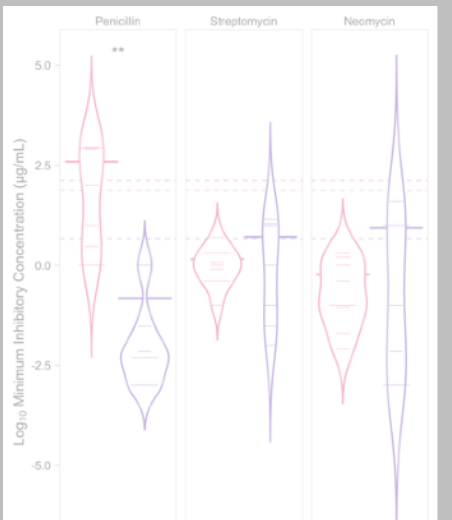
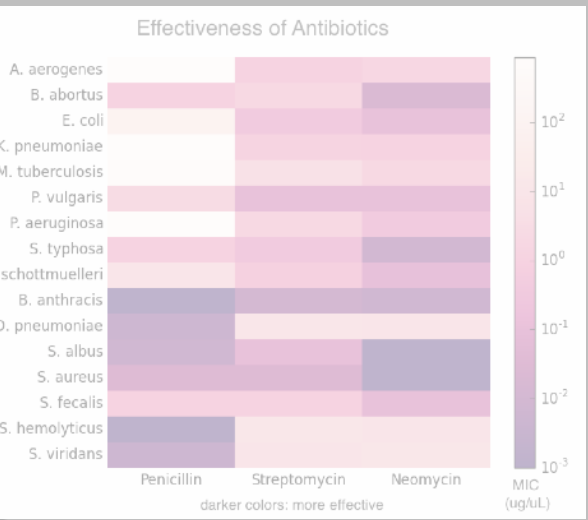


Effectiveness of Antibiotics



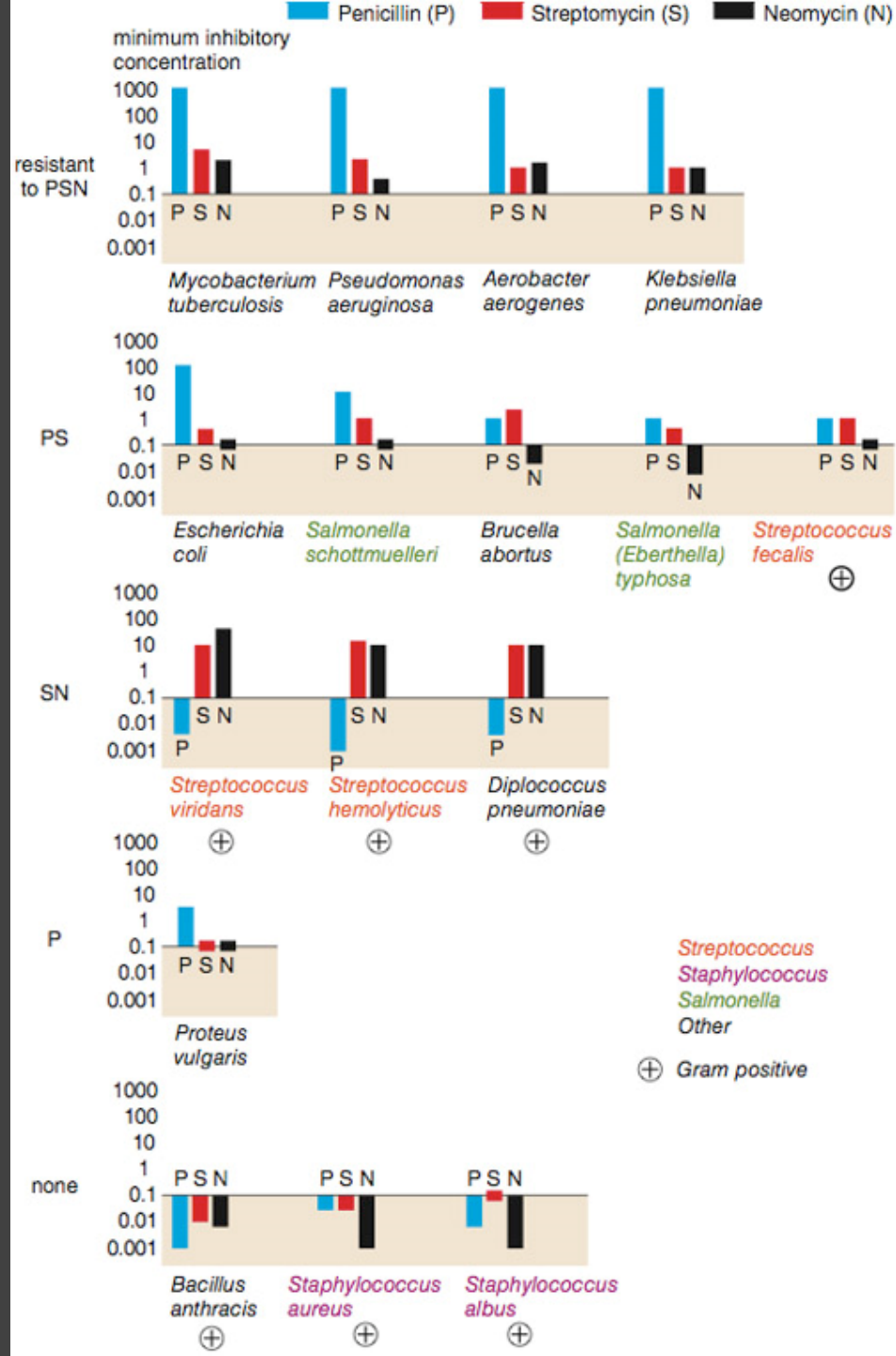


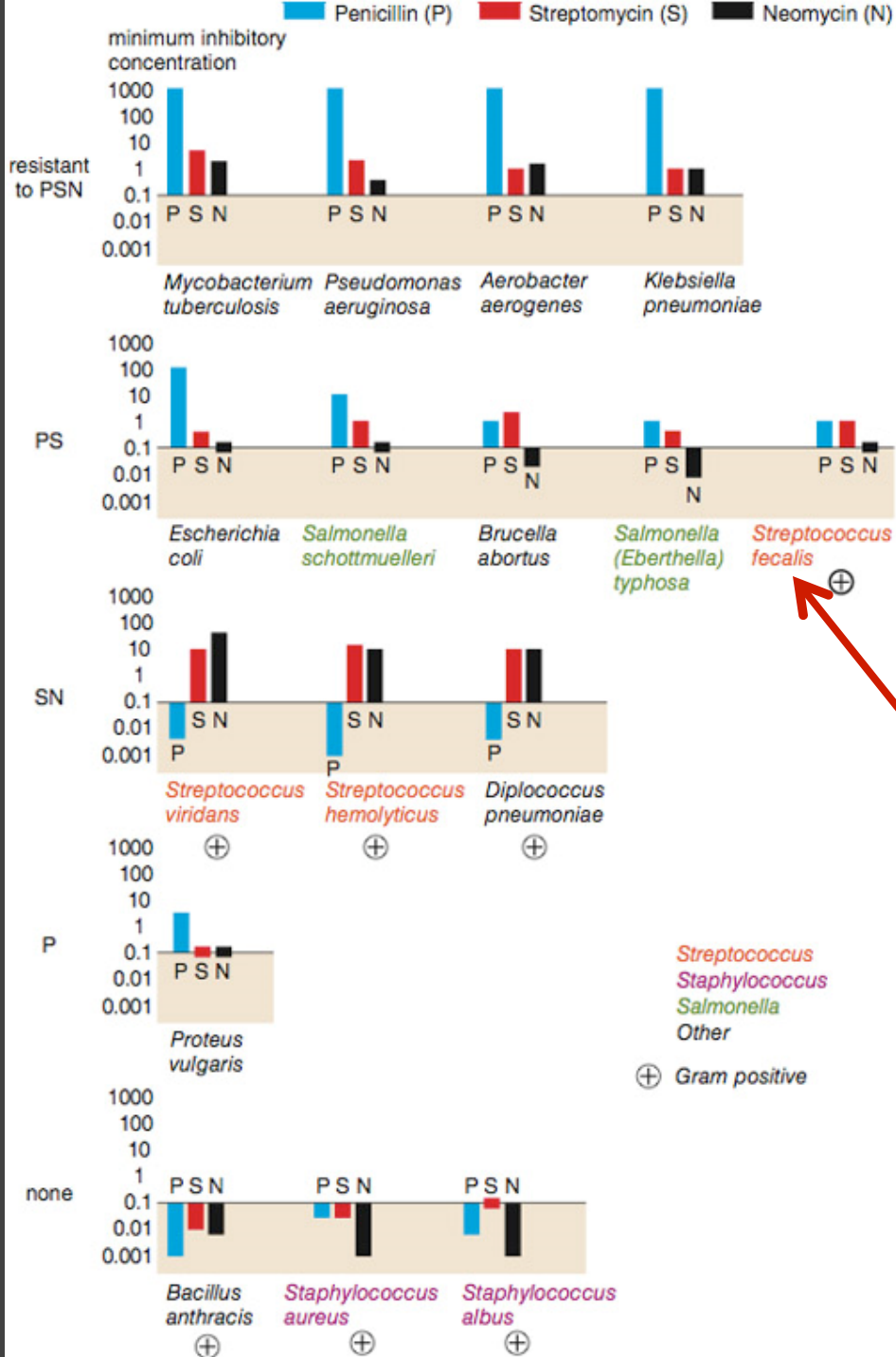
Which antibiotic should one use?



**Do the bacteria
group by antibiotic
resistance?**

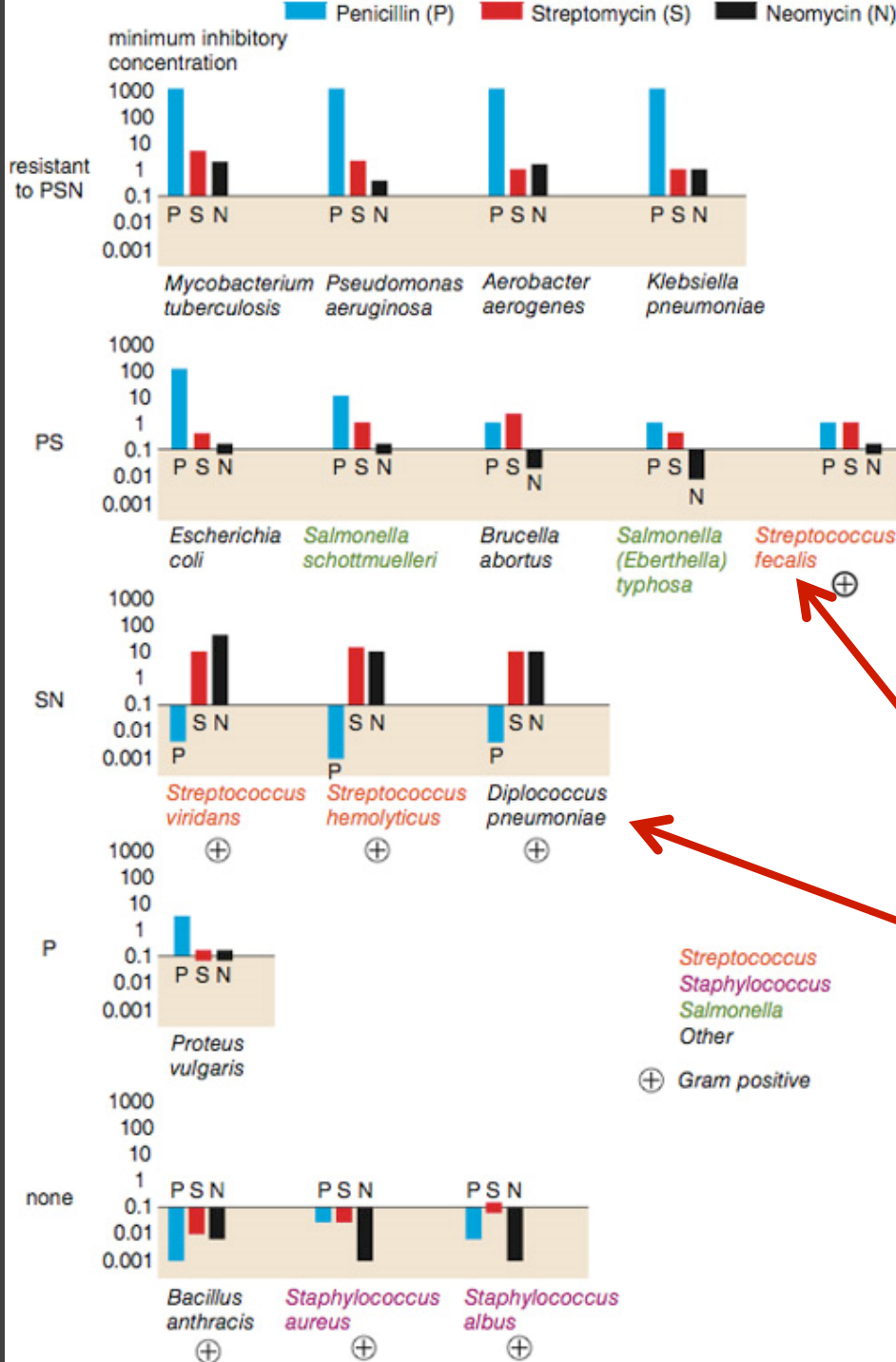
Do the bacteria group by antibiotic resistance?





Do the bacteria group by antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

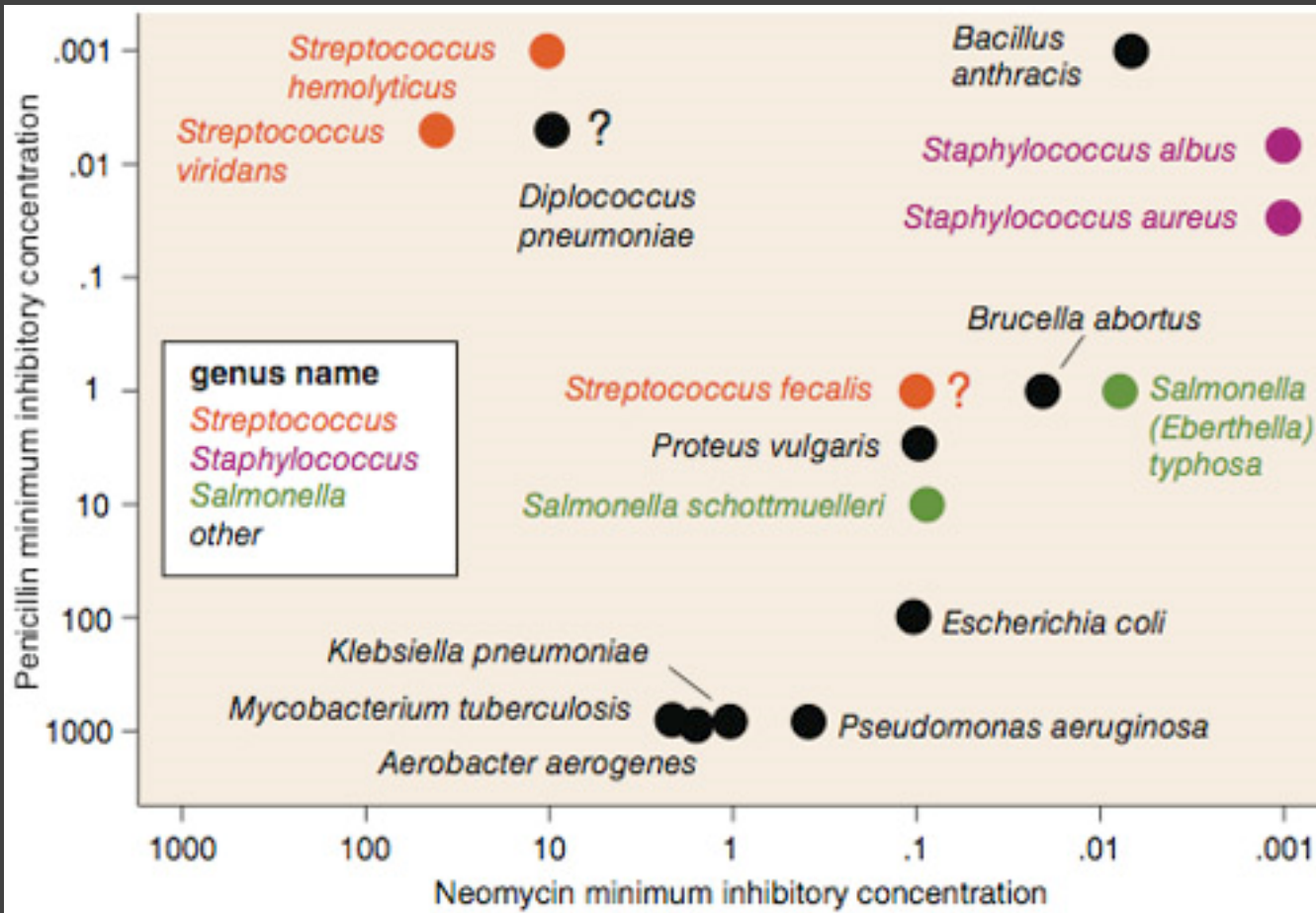


Do the bacteria group by antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

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



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

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

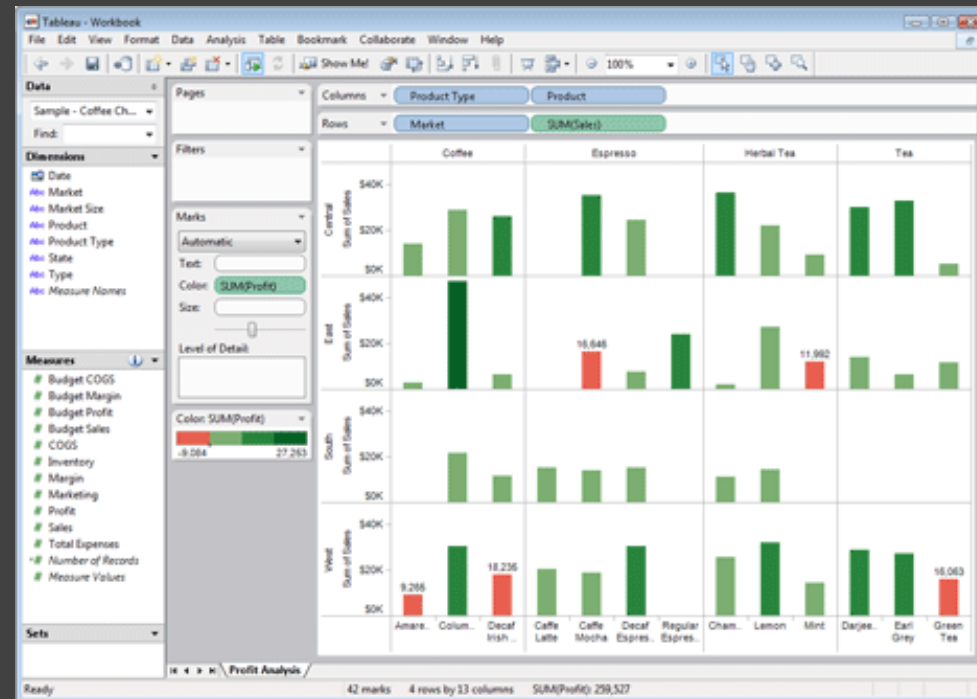
Interact with data

Refine your questions

Author a report

Screenshots of most insightful views (10+)

Include titles and captions for each view

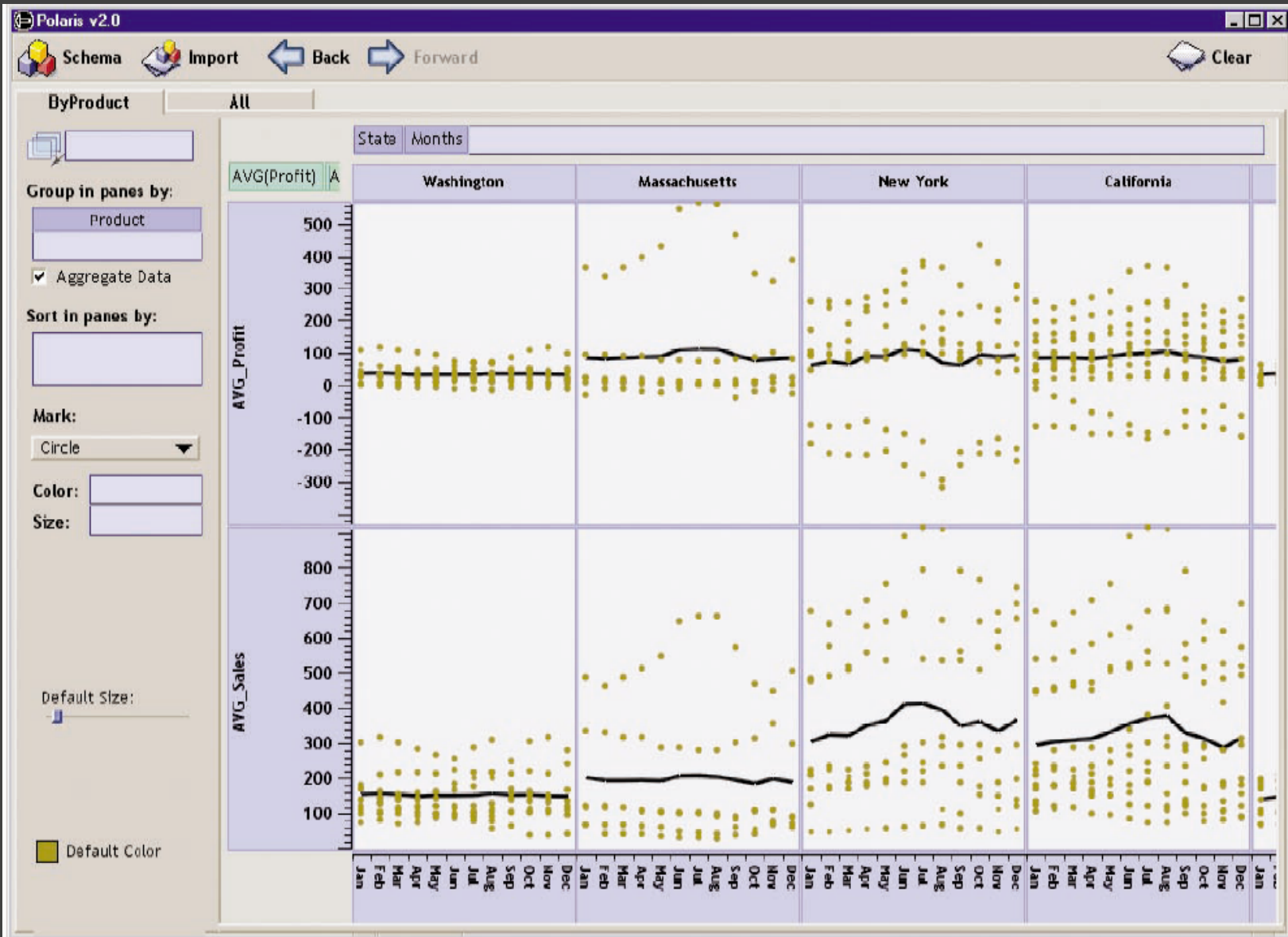


Due by 11:59pm

Monday, Apr 22

Tableau / Polaris

Polaris [Stolte, Tang & Hanrahan '00]



Tableau

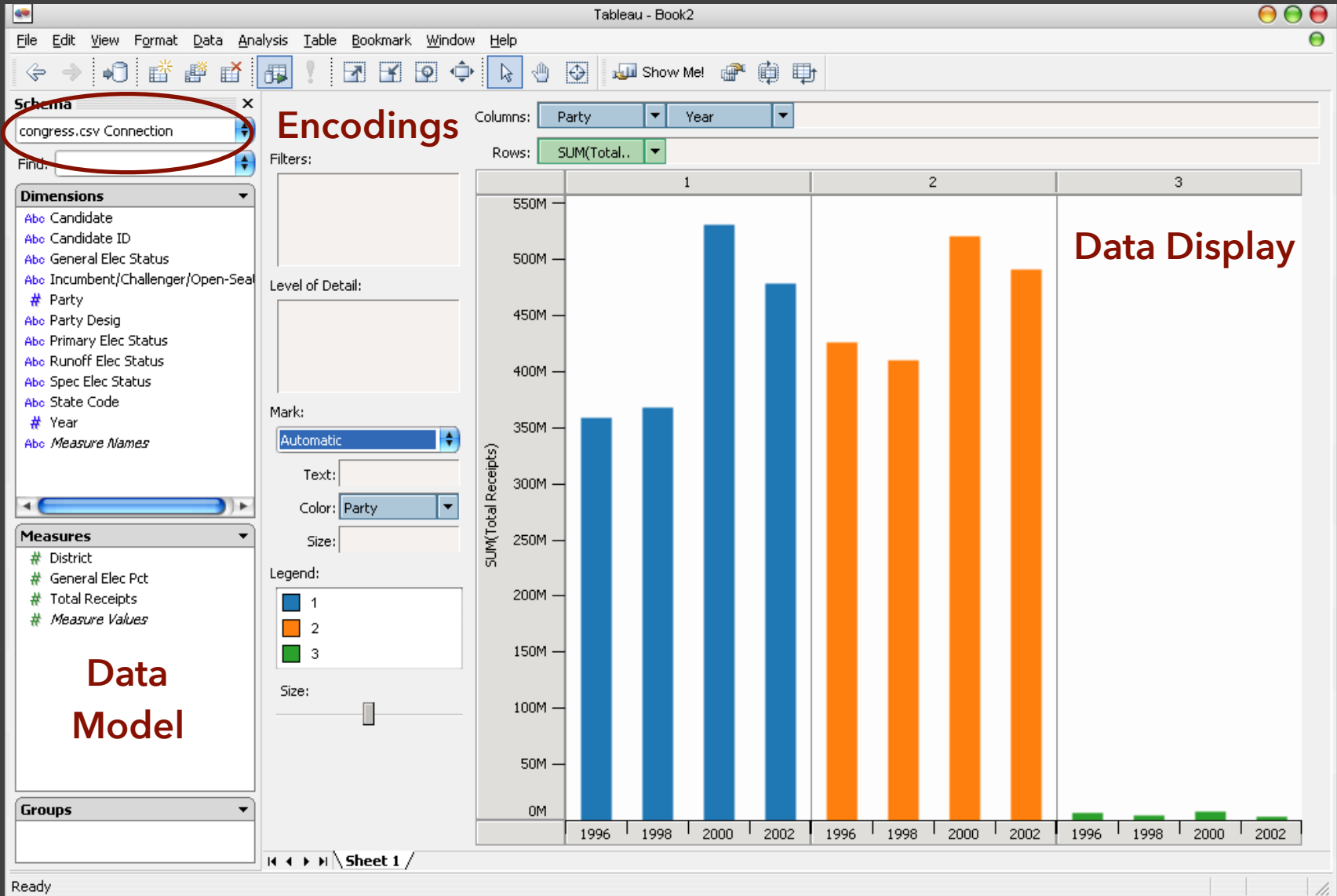


Tableau / Polaris Approach

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

Tableau Demo

The dataset:

Federal Elections Commission Receipts

Every Congressional Candidate from 1996 to 2002

4 Election Cycles

9216 Candidacies

Dataset Schema

Year (Qi)

Candidate Code (N)

Candidate Name (N)

Incumbent / Challenger / Open-Seat (N)

Party Code (N) [1=Dem,2=Rep,3=Other]

Party Name (N)

Total Receipts (Qr)

State (N)

District (N)

This is a subset of the larger data set available from the FEC.

Hypotheses?

What might we learn from this data?

Hypotheses?

What might we learn from this data?

Correlation between receipts and winners?

Do receipts increase over time?

Which states spend the most?

Which party spends the most?

Margin of victory vs. amount spent?

Amount spent between competitors?

Tableau Demo

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 (/)



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

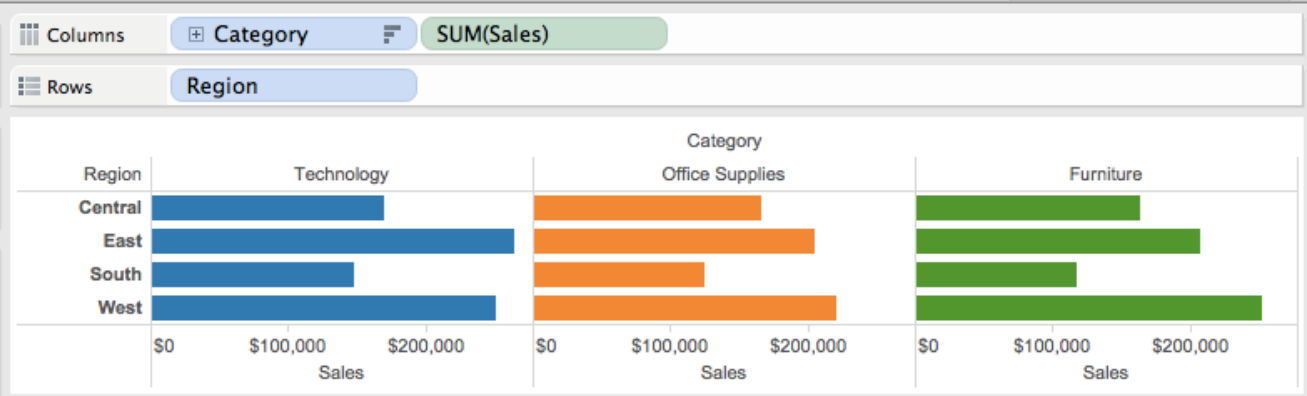
Automatic

Color Size Label

Detail Tooltip

Category

- Technology
- Office Supplies
- Furniture



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

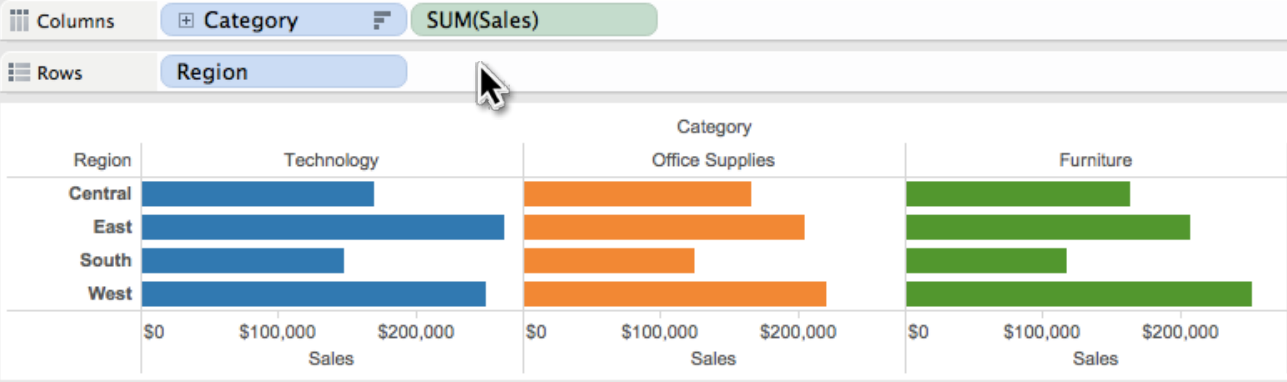
Color Size Label

Detail Tooltip

Category

Category

- Technology
- Office Supplies
- Furniture



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

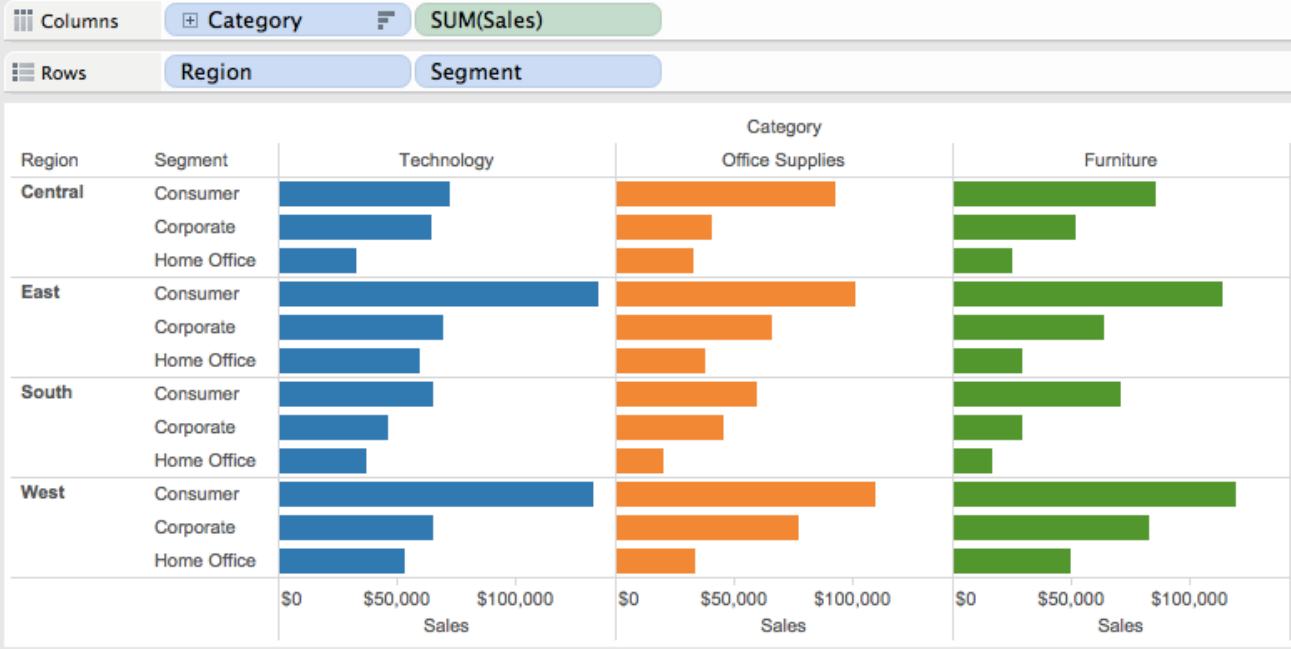
Automatic

Color Size Label

Detail Tooltip

Category

Technology
Office Supplies
Furniture



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

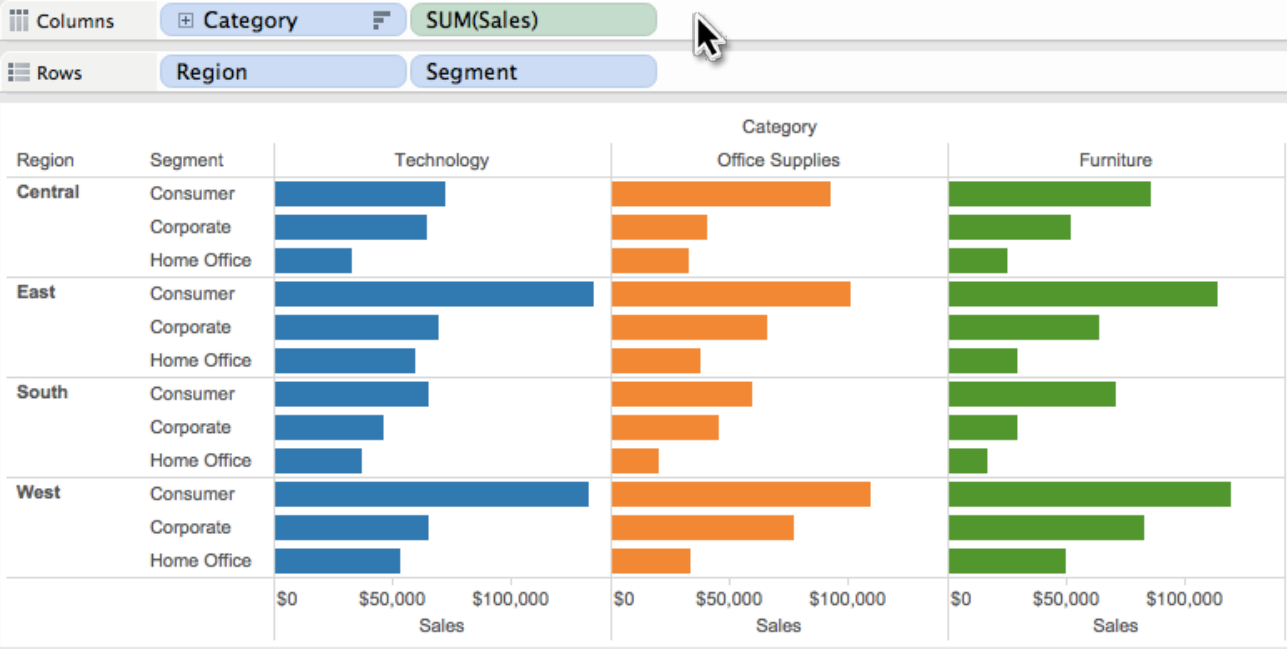
Color Size Label

Detail Tooltip

Category

Category

- Technology
- Office Supplies
- Furniture



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

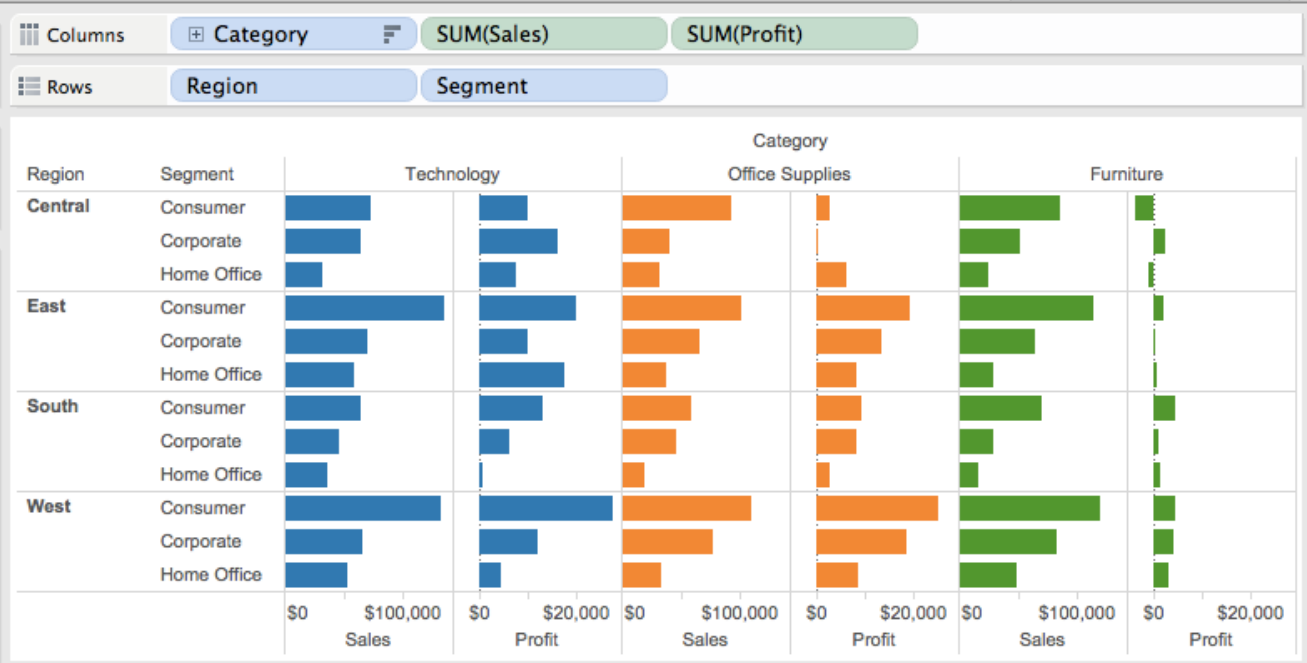
Category

SUM(Sales)

SUM(Profit)

Category

- Technology
- Office Supplies
- Furniture



Columns: Category, SUM(Sales), SUM(Profit)
 Rows: Region, Segment
GROUP BY Category, Region, Segment

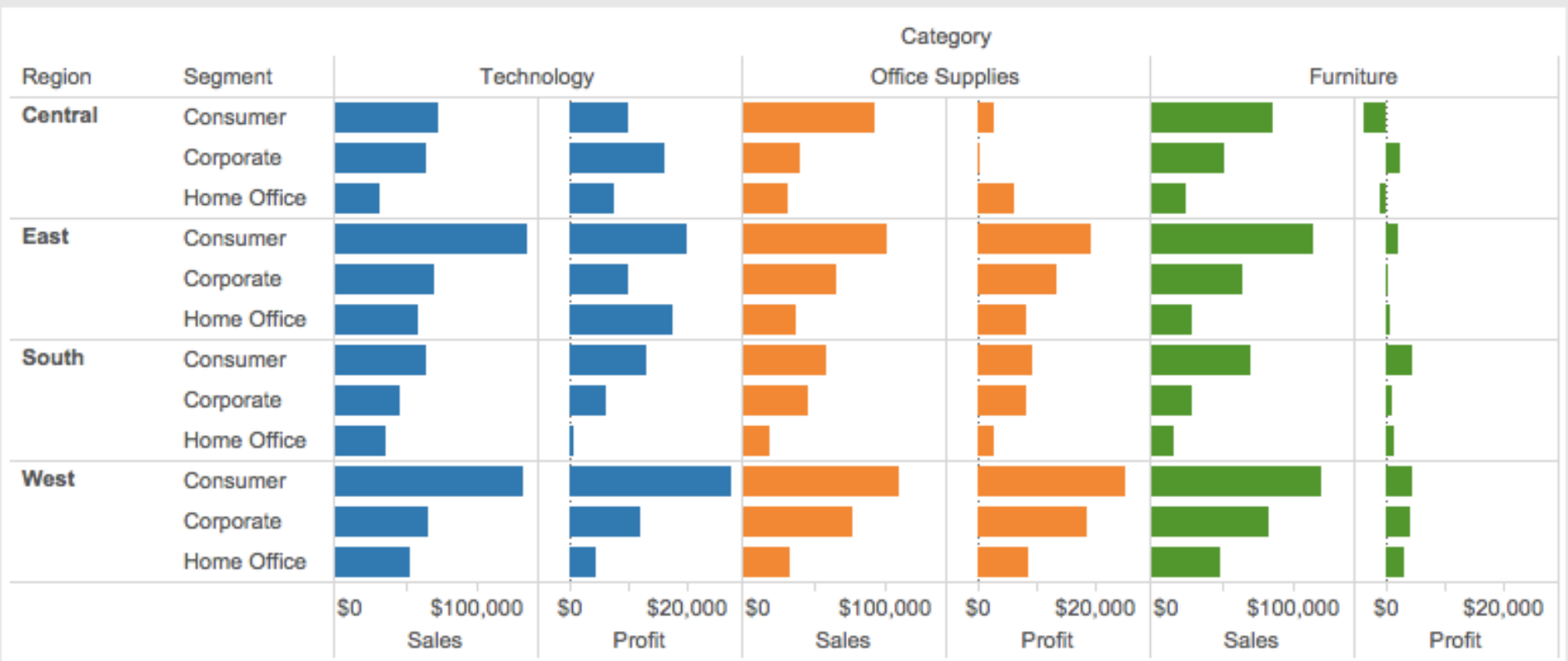


Table Algebra

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

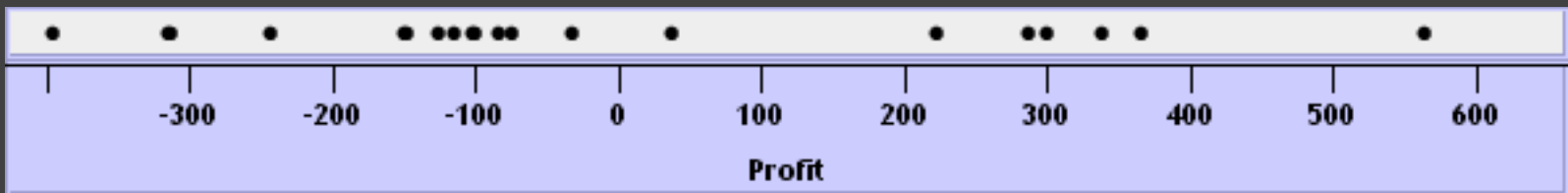
Ordinal fields: interpret domain as a set that partitions table into rows and columns.

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

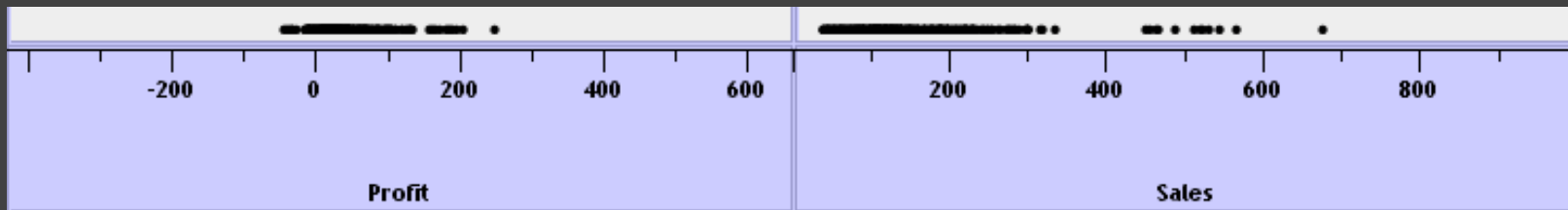
Quarter + Product Type

= {(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

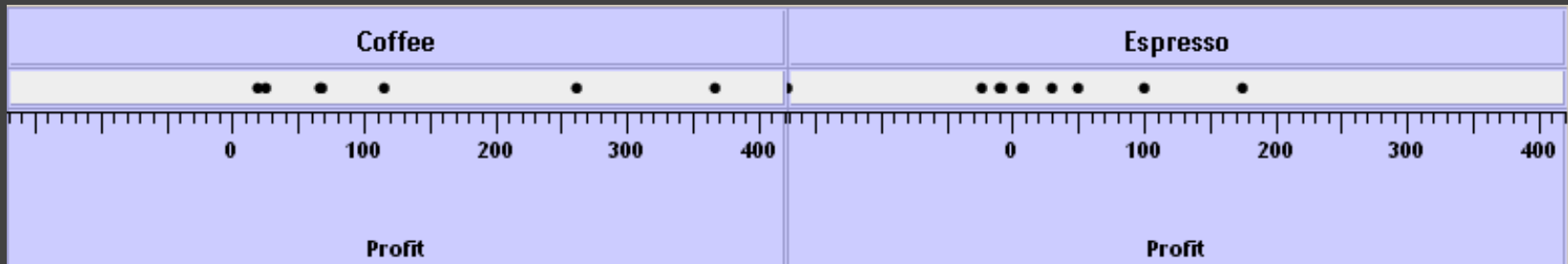
Cross-product of set interpretations

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)

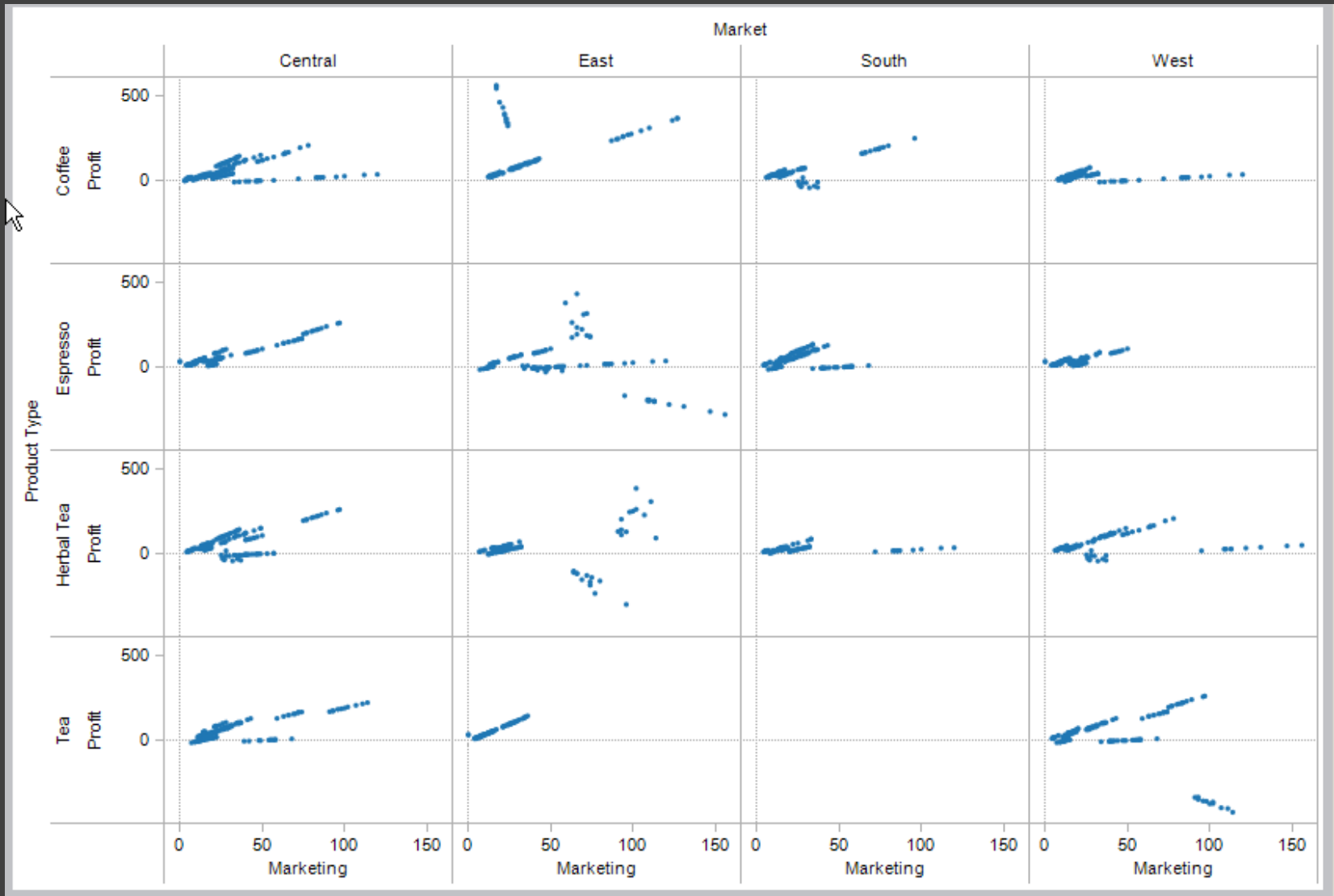
Quarter / Month ->

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

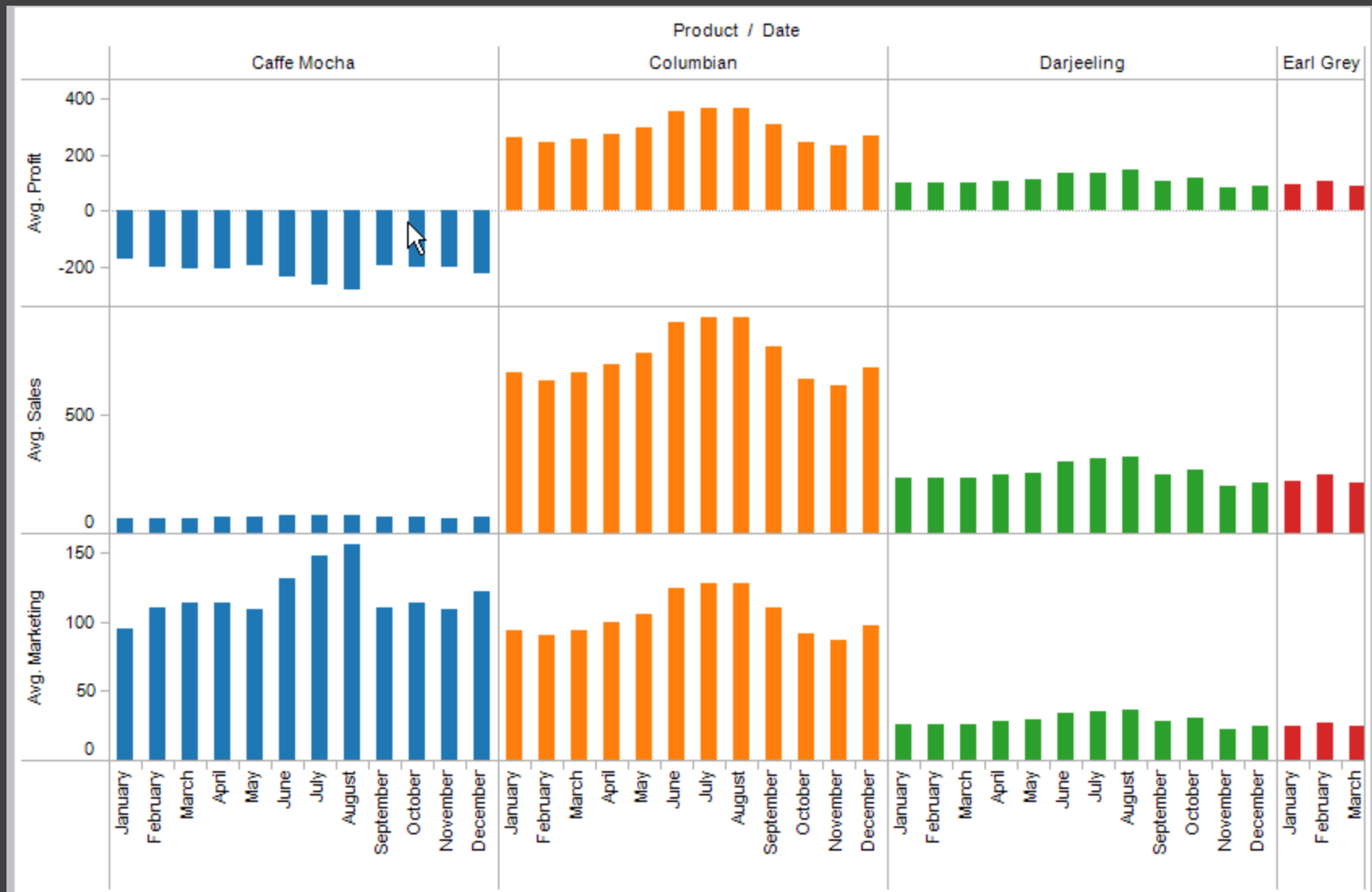
Ordinal-Ordinal

State	Product Type			
	Coffee	Espresso	Herbal Tea	Tea
Colorado	●	●	●	●
Connecticut	●	●	●	●
Florida	●	●	●	●
Illinois	●	●	●	●
Iowa	●	●	●	●
Louisiana	●	●	●	●
Massachusetts	●	●	●	●
Missouri	●	●	●	●
Nevada	●	●	●	●
New Hampshire	●	●	●	●
New Mexico	●	●	●	●
New York	●	●	●	●
Ohio	●	●	●	●
Oklahoma	●	●	●	●
Oregon	●	●	●	●
Texas	●	●	●	●
Utah	●	●	●	●
Washington	●	●	●	●
Wisconsin	●	●	●	●

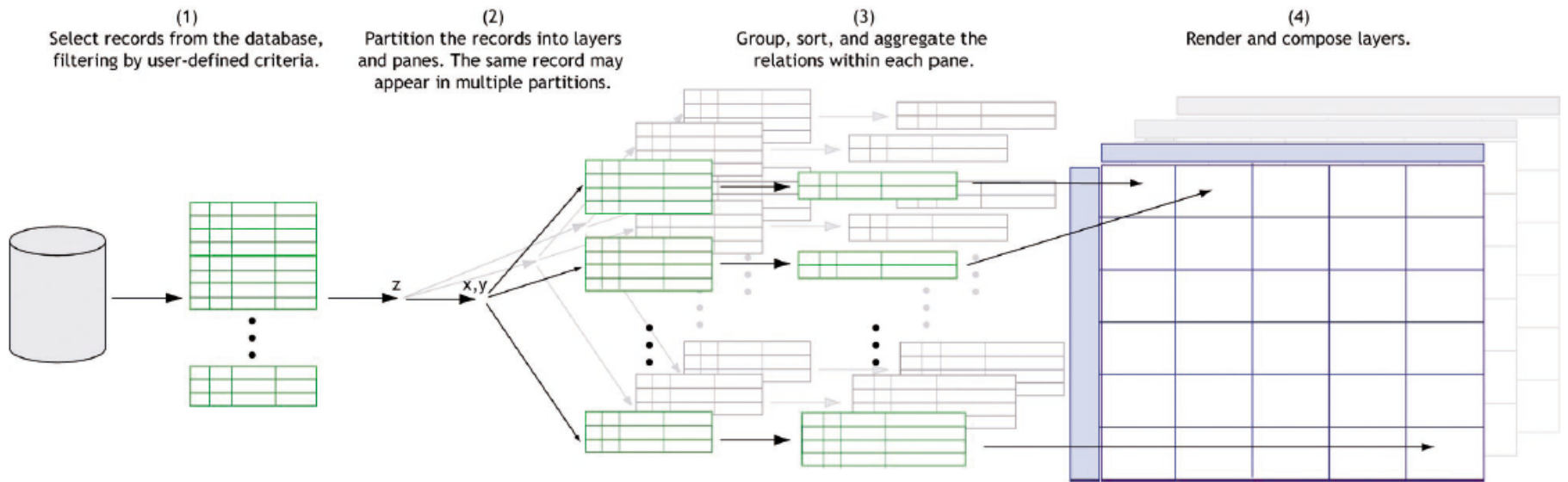
Quantitative-Quantitative



Ordinal-Quantitative



Querying the Database



BONUS TOPIC

Data Fraud

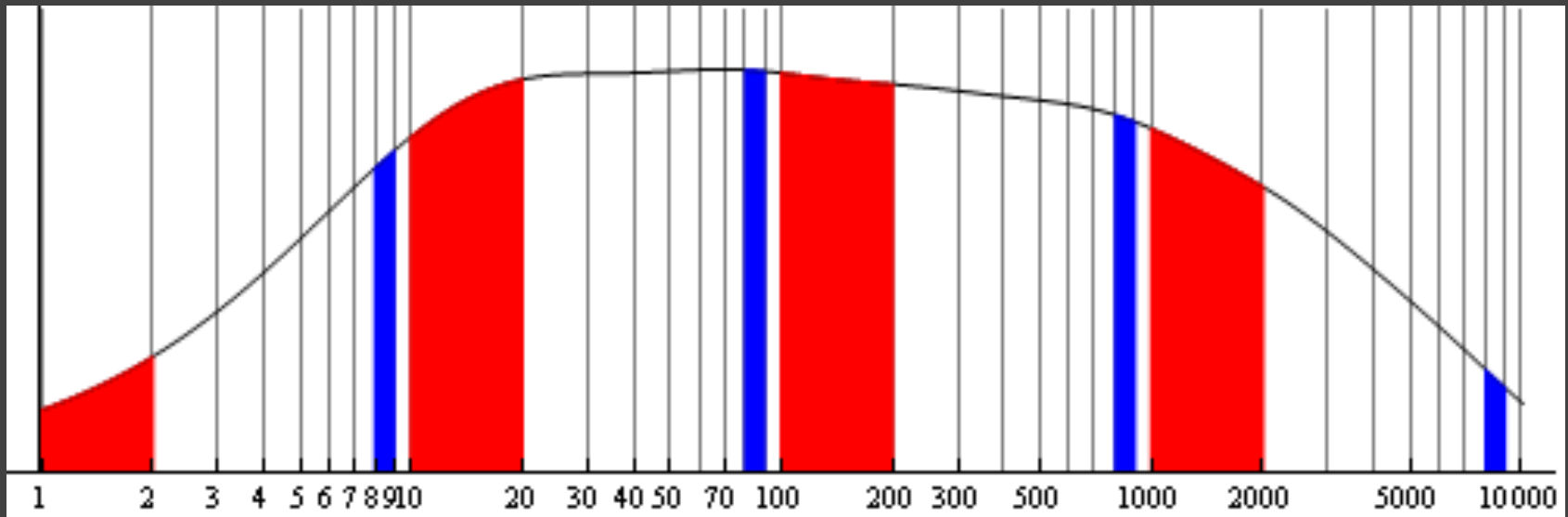
A Detective Story

You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

<i>Firm A</i>		<i>Firm B</i>	LIARS!
283.08	25.23	283.08	75.23
153.86	385.62	353.86	185.25
1448.97	12371.32	5322.79	9971.42
18595.91	1280.76	8795.64	4802.43
21.33	257.64	61.33	57.64
Amt. Paid: \$34823.72		Amt. Rec'd: \$29908.67	

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.



Hence the leading digit **1** has a ~30% likelihood. Larger digits are increasingly less likely.

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Data must span multiple orders of magnitude.

Evidence that records do not follow Benford's Law is admissible in a court of law!