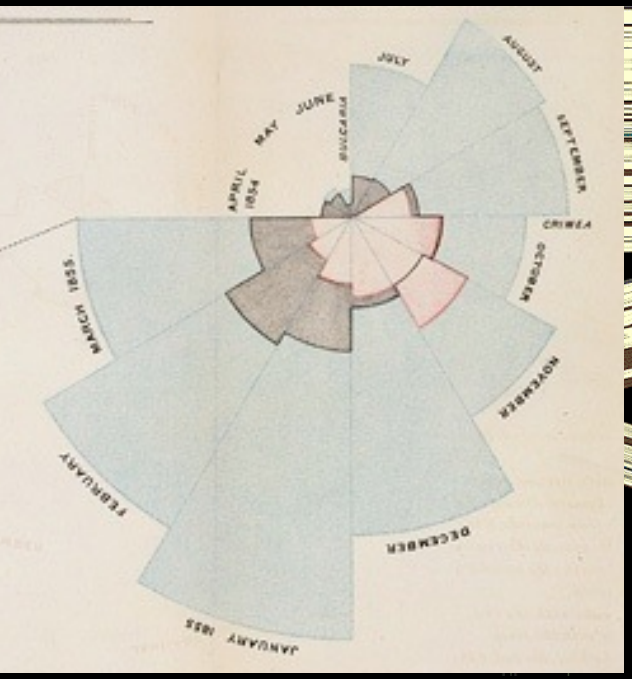
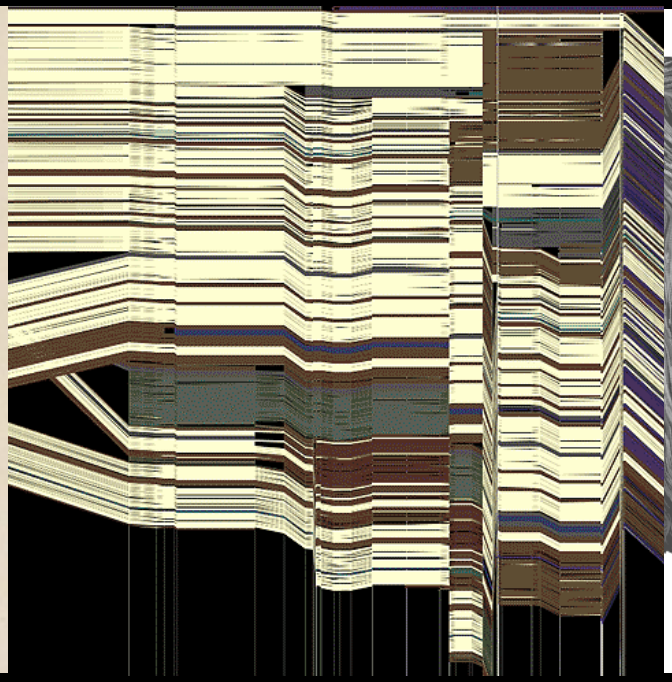
CSE 442 - Data Visualization

Exploratory Data Analysis

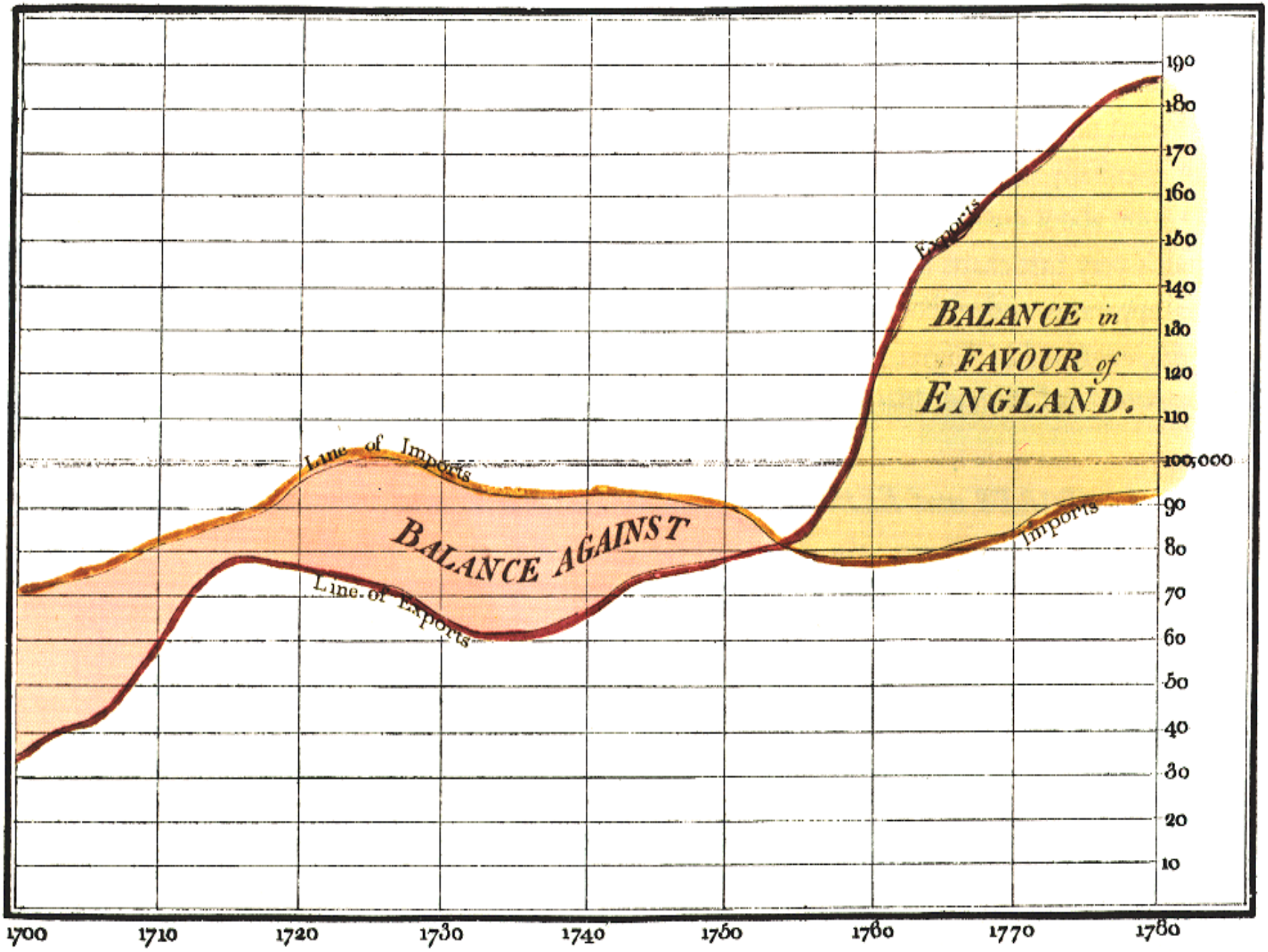




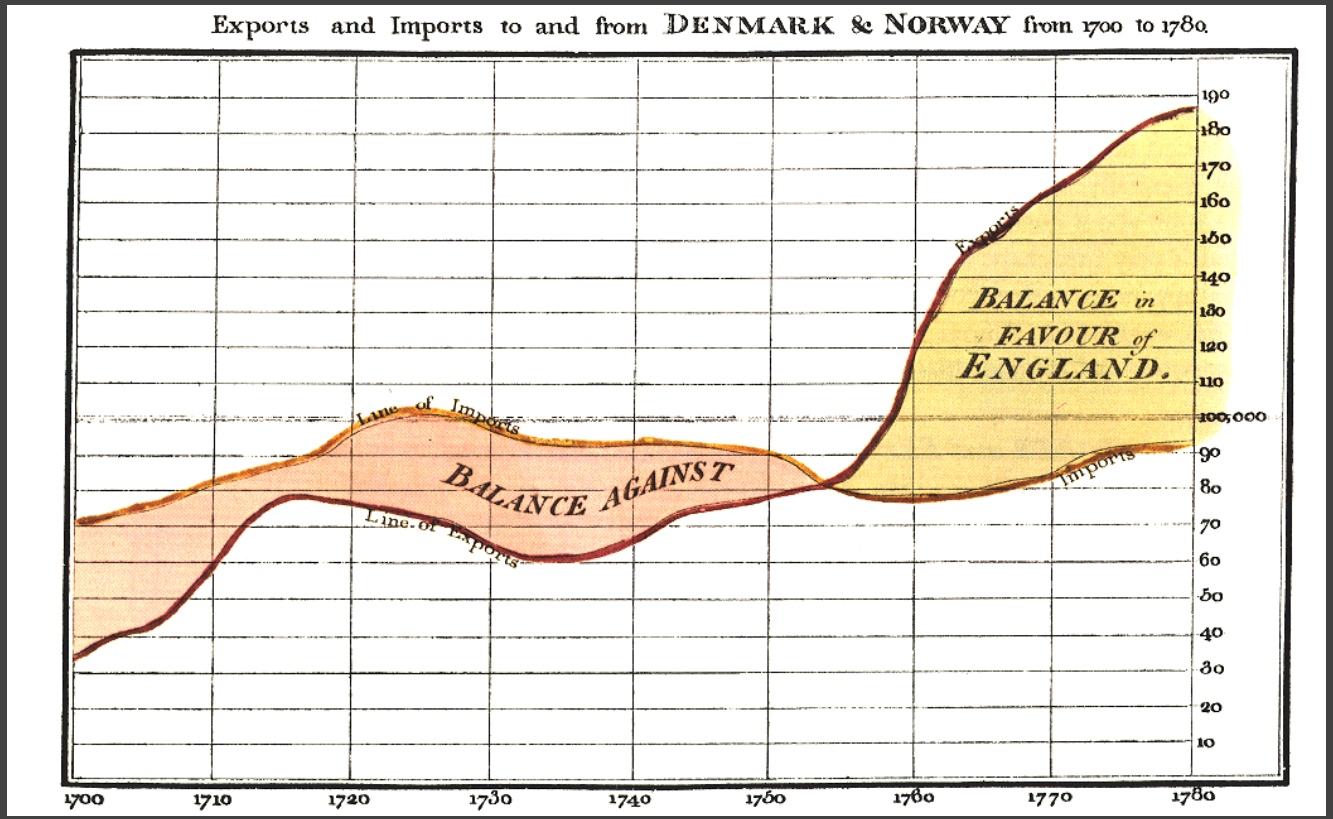
Matthew Conlen University of Washington

Deconstructions

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



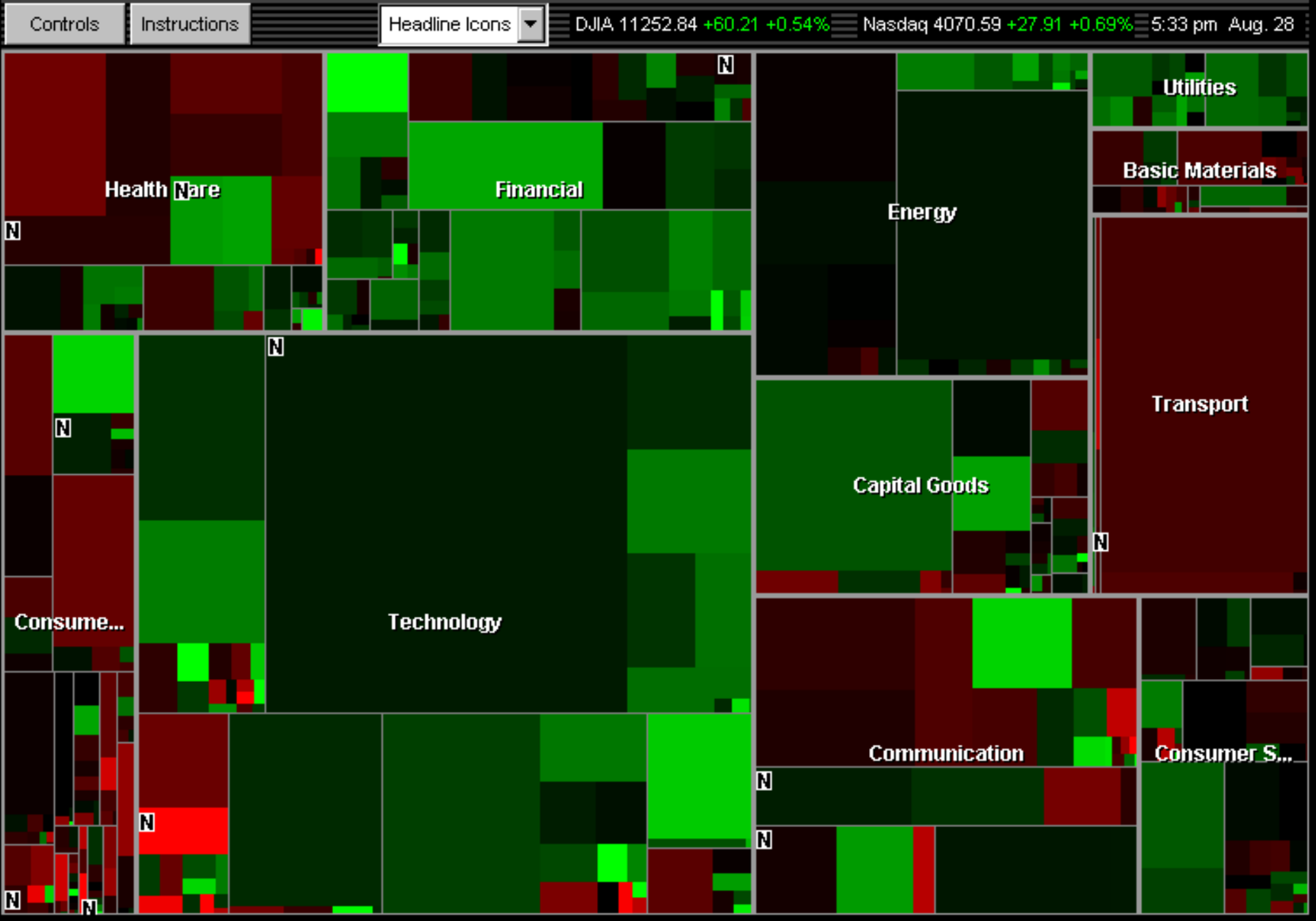
William Playfair, 1786



X-axis: year (Q)

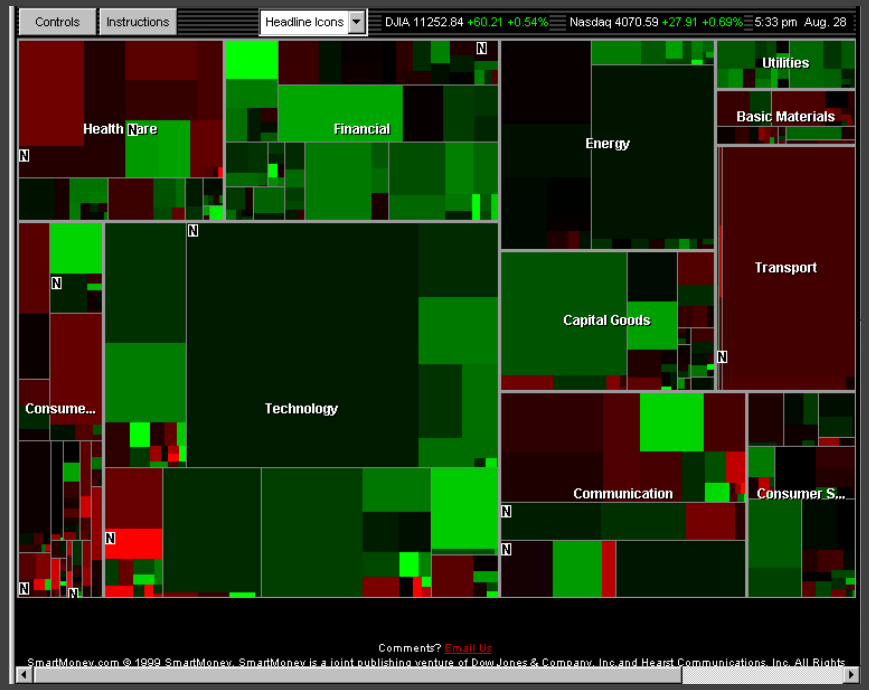
Y-axis: currency (Q)

Color: imports/exports (N, O)



http://www.smartmoney.com/marketmap/

Wattenberg's Map of the Market



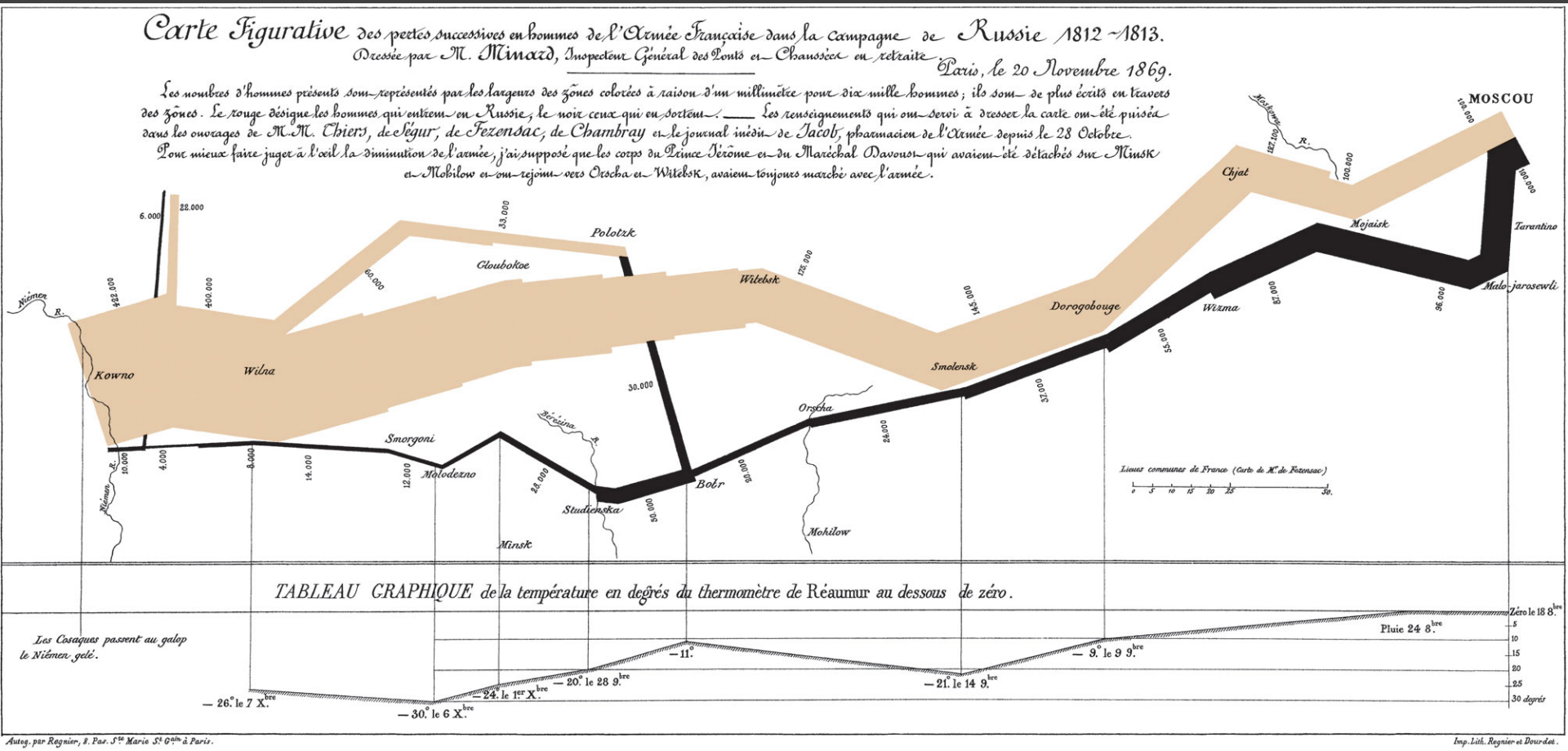
Rectangle Area: market cap (Q)

Rectangle Position: market sector (N), market cap (Q)

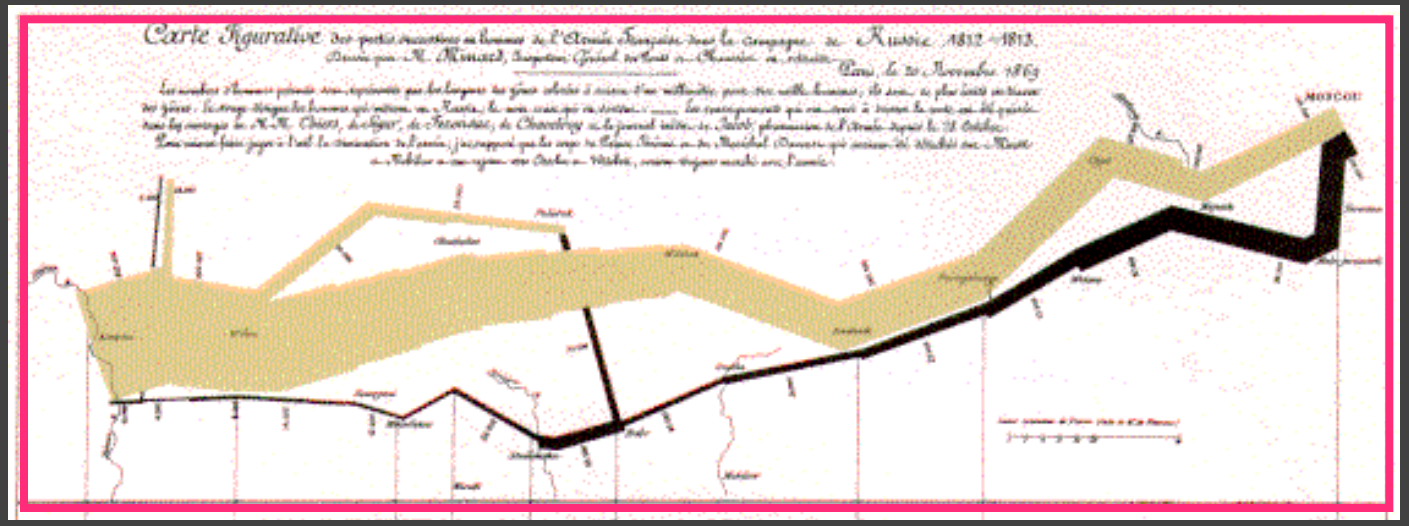
Color Hue: loss vs. gain (N, O)

Color Value: magnitude of loss or gain (Q)

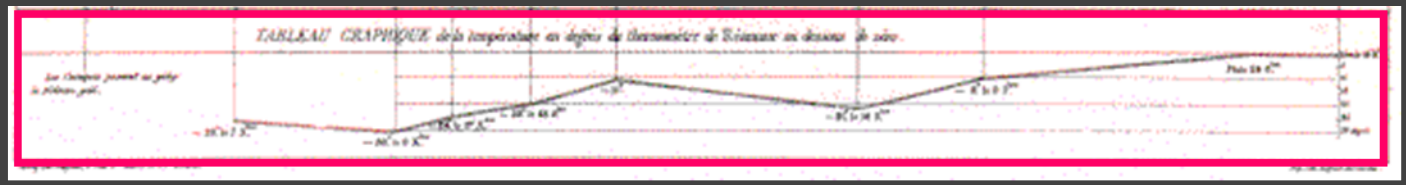
Minard 1869: Napoleon's March



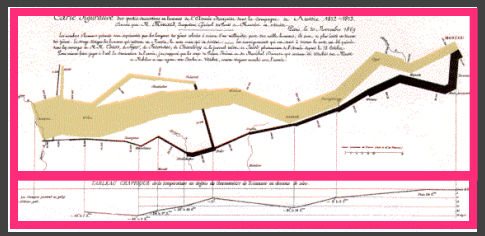
Single-Axis Composition



+



=

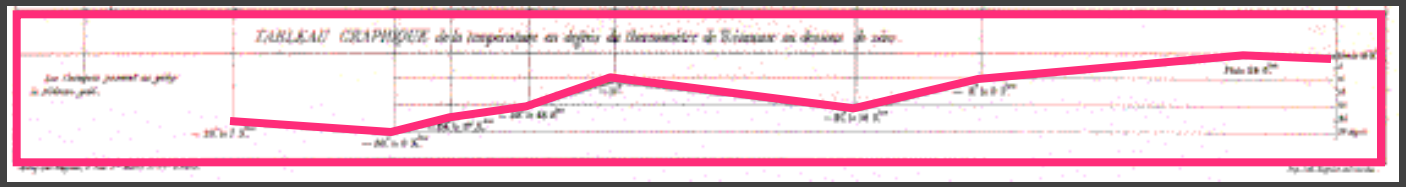


Mark Composition

Y-axis: temperature (Q)

\+ X-axis: longitude (Q) / time (O)

=



Temp over space/time (Q x Q)

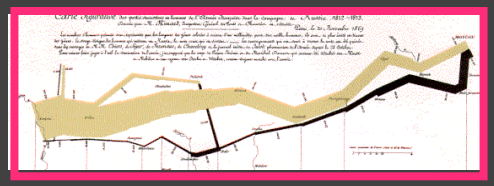
Mark Composition

Y-axis: longitude (Q)

\+ **X-axis**: latitude (Q)

\+ **Width**: army size (Q)

=

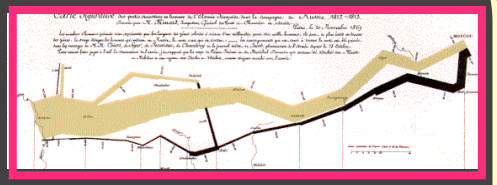


Army position (Q x Q) and army size (Q)

longitude (Q)

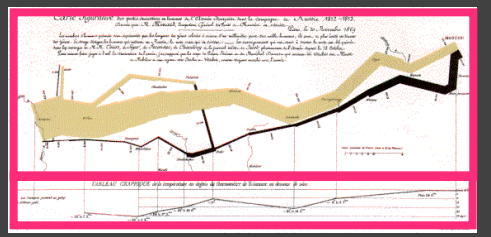
latitude (Q)

army size (Q)

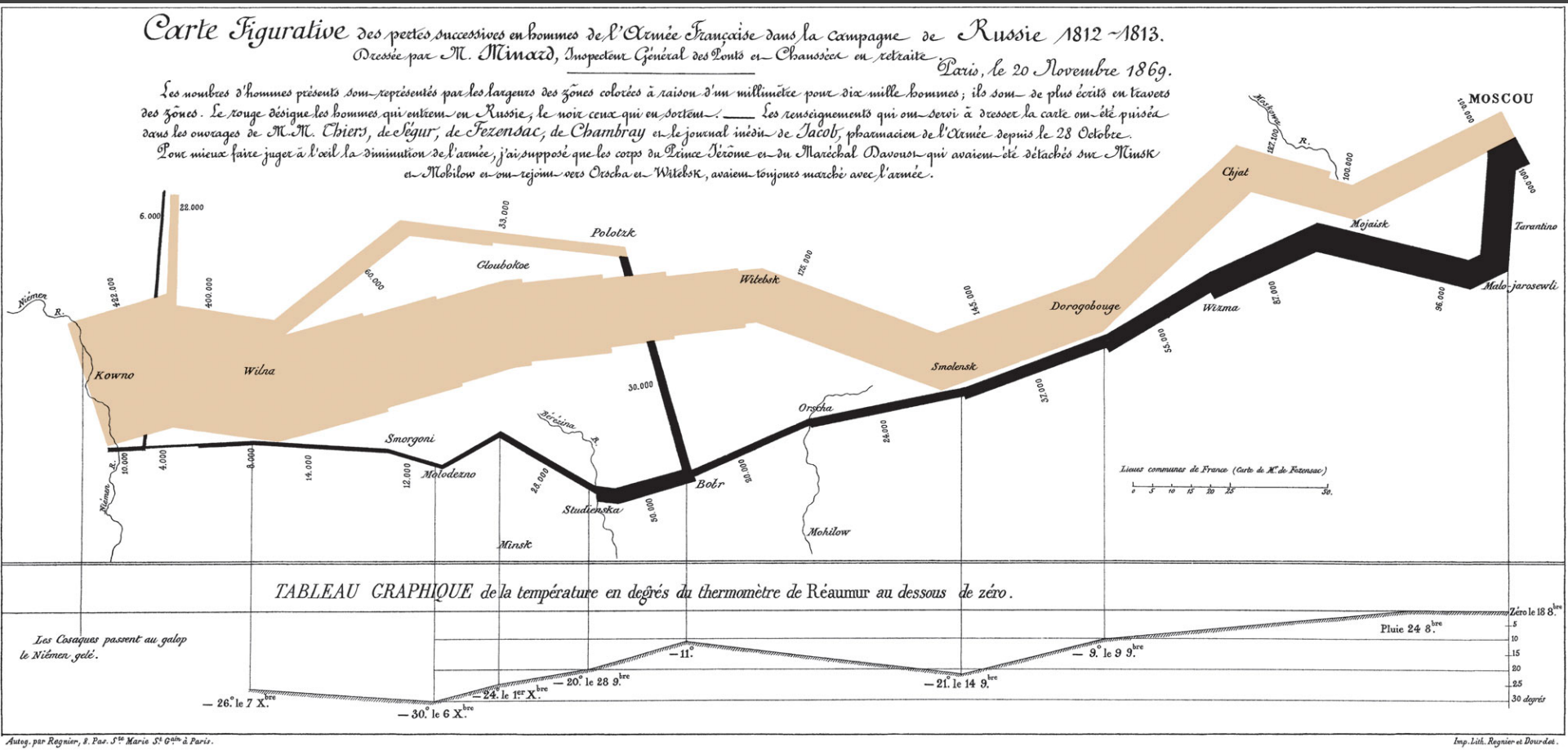


temperature (Q)

latitude (Q) / time (O)



Minard 1869: Napoleon's March



Depicts at least 5 quantitative variables. Any others?

Formalizing Design

Choosing Visual Encodings

Assume *k* visual encodings and *n* data attributes. We would like to pick the "best" encoding among a combinatorial set of possibilities of size $(n+1)^k$

Principle of Consistency
The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering
Encode the most important information in the most effective way.

Design Criteria [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria [Mackinlay 86]

Expressiveness

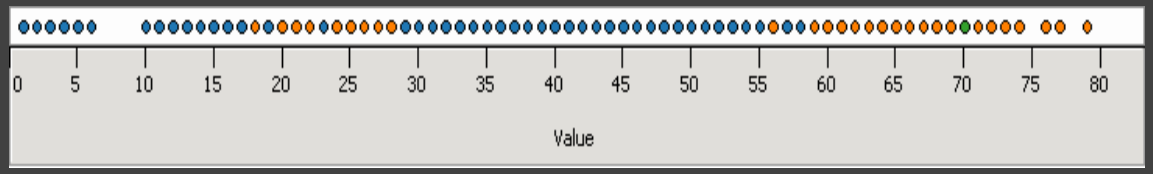
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

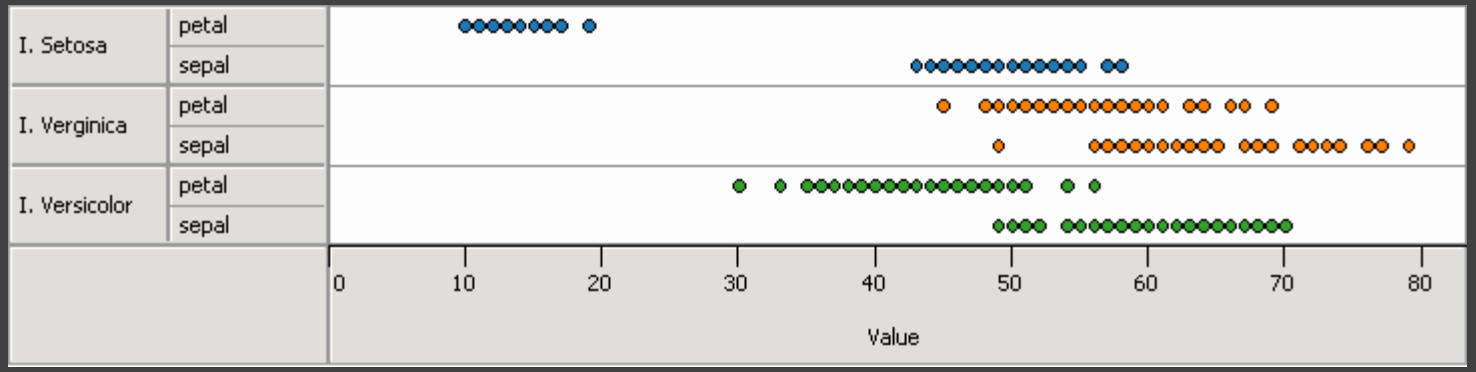
Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Can not express the facts

A multivariate relation may be *inexpressive* in a single horizontal dot plot because multiple records are mapped to the same position.





Expresses facts not in the data

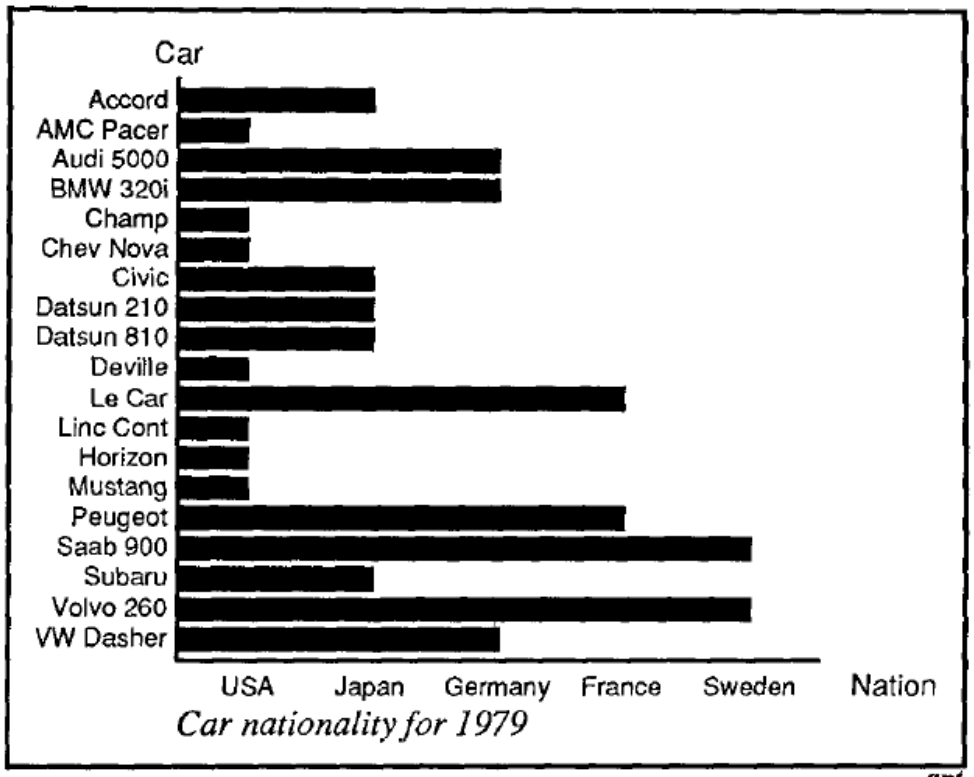


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

A length is interpreted as a quantitative value.

Design Criteria [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria [Mackinlay 86]

Expressiveness
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

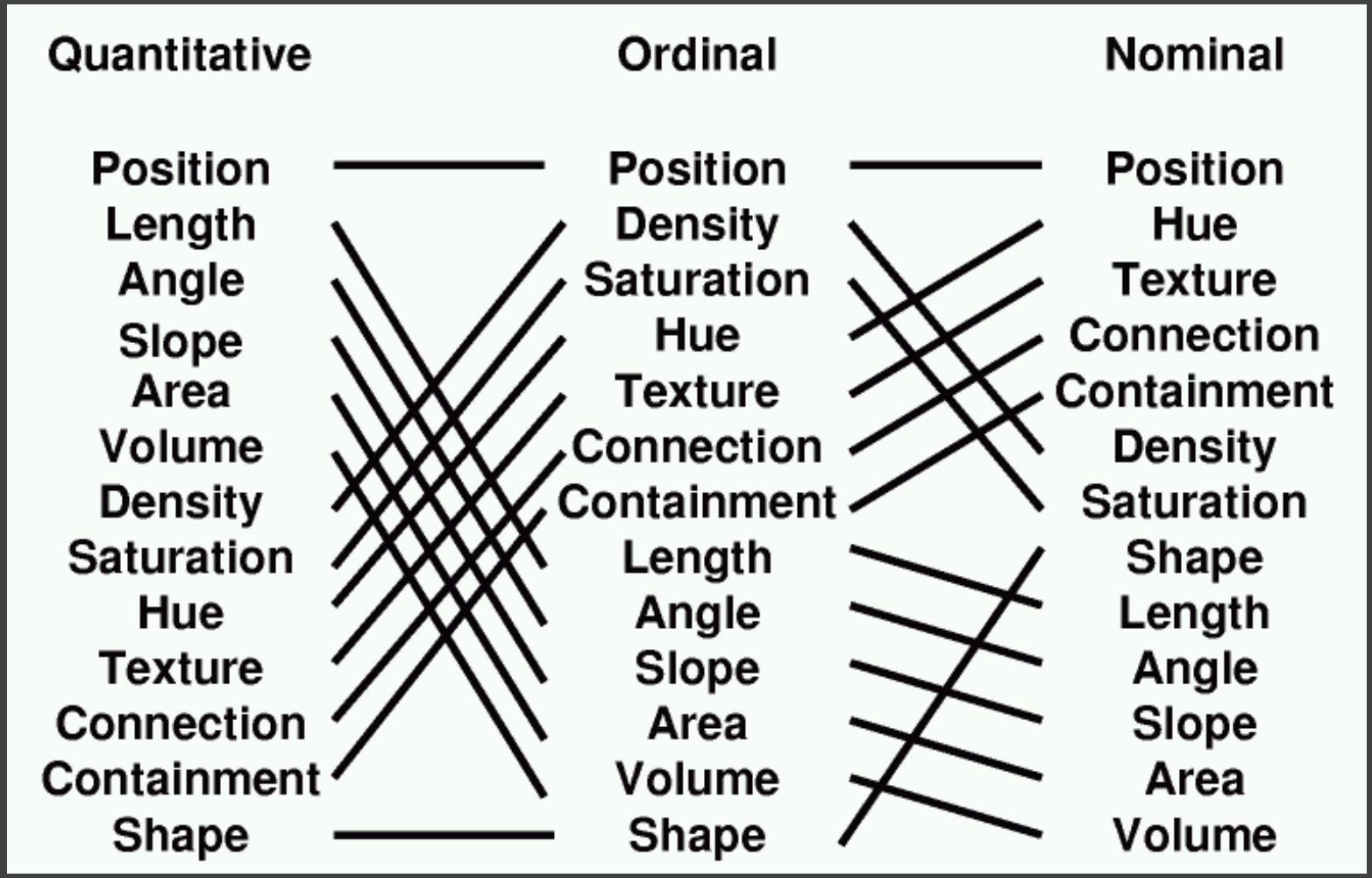
Effectiveness
A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria *Translated*

Tell the truth and nothing but the truth
(don't lie, and don't lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)

Mackinlay's Ranking



Conjectured *effectiveness* of encodings by data type

Mackinlay's Design Algorithm

APT - "A Presentation Tool", 1986

User formally specifies data model and type
Input: ordered list of data variables to show

APT searches over design space
Test expressiveness of each visual encoding
Generate encodings that pass test
Rank by perceptual effectiveness criteria

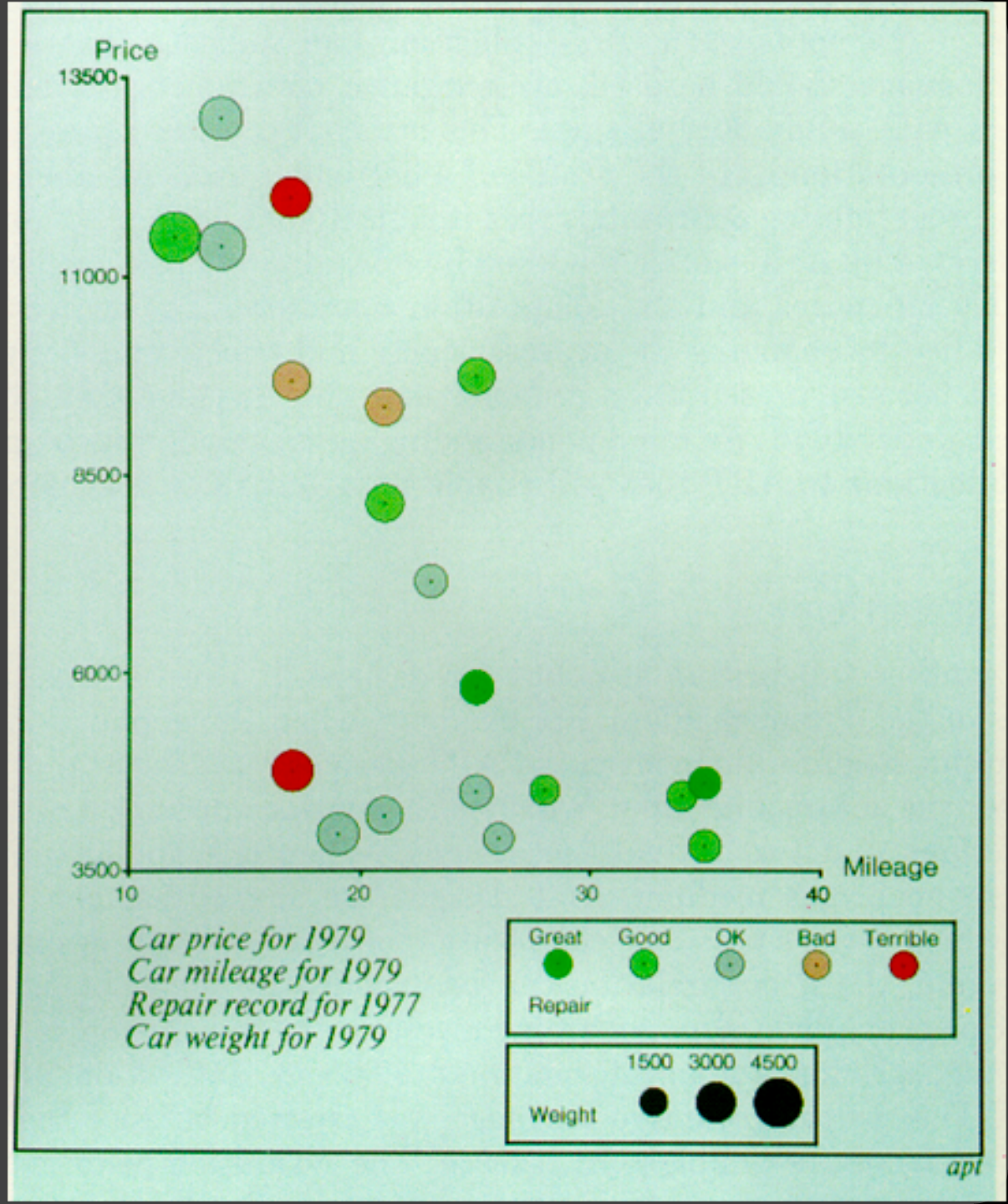
Output the "most effective" visualization

APT

Automatically generate chart for car data

Input variables:

1. Price
2. Mileage
3. Repair
4. Weight



Limitations of APT?

Limitations of APT

Does not cover many visualization techniques
Networks, hierarchies, maps, diagrams
Also: 3D structure, animation, illustration, …

Does not consider interaction

Does not consider semantics / conventions

Assumes single visualization as output

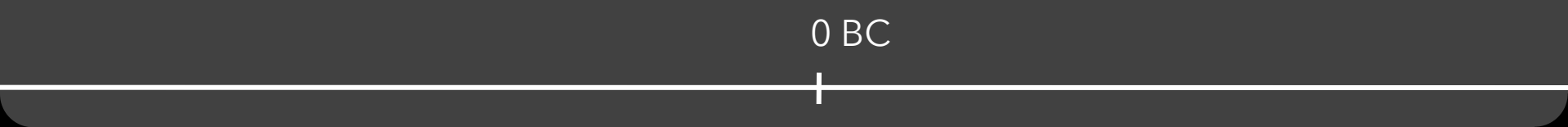
Summary: Data & Image Models

Formal specification
Data model: relational data; N,O,Q types
Image model: visual encoding channels
Encodings map data to visual variables

Choose expressive and effective encodings
Rule-based tests of expressiveness
Perceptual effectiveness rankings

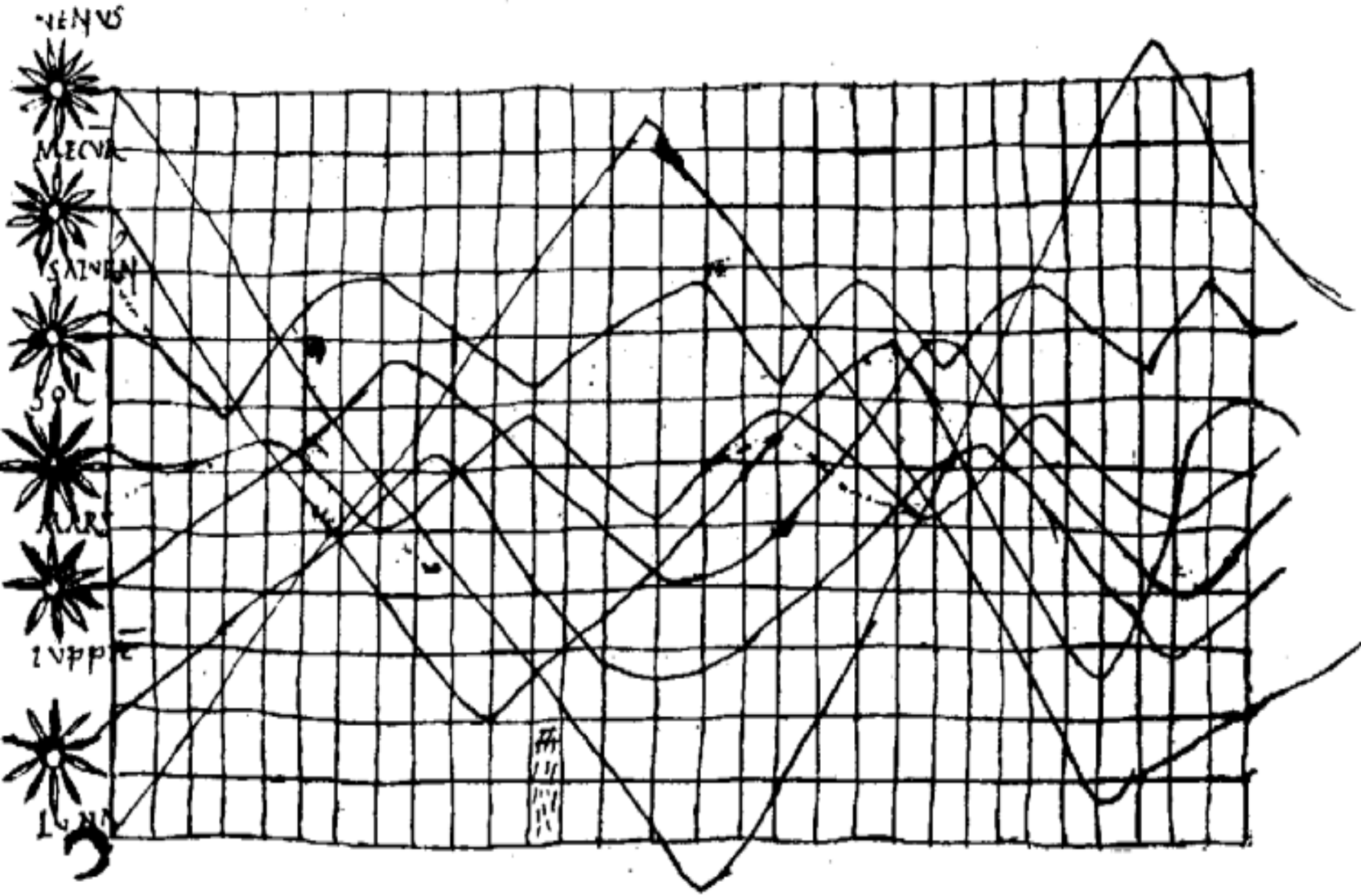
Question: how do we establish effectiveness criteria? *Subject of perception lectures…*

What was the **first** data visualization?

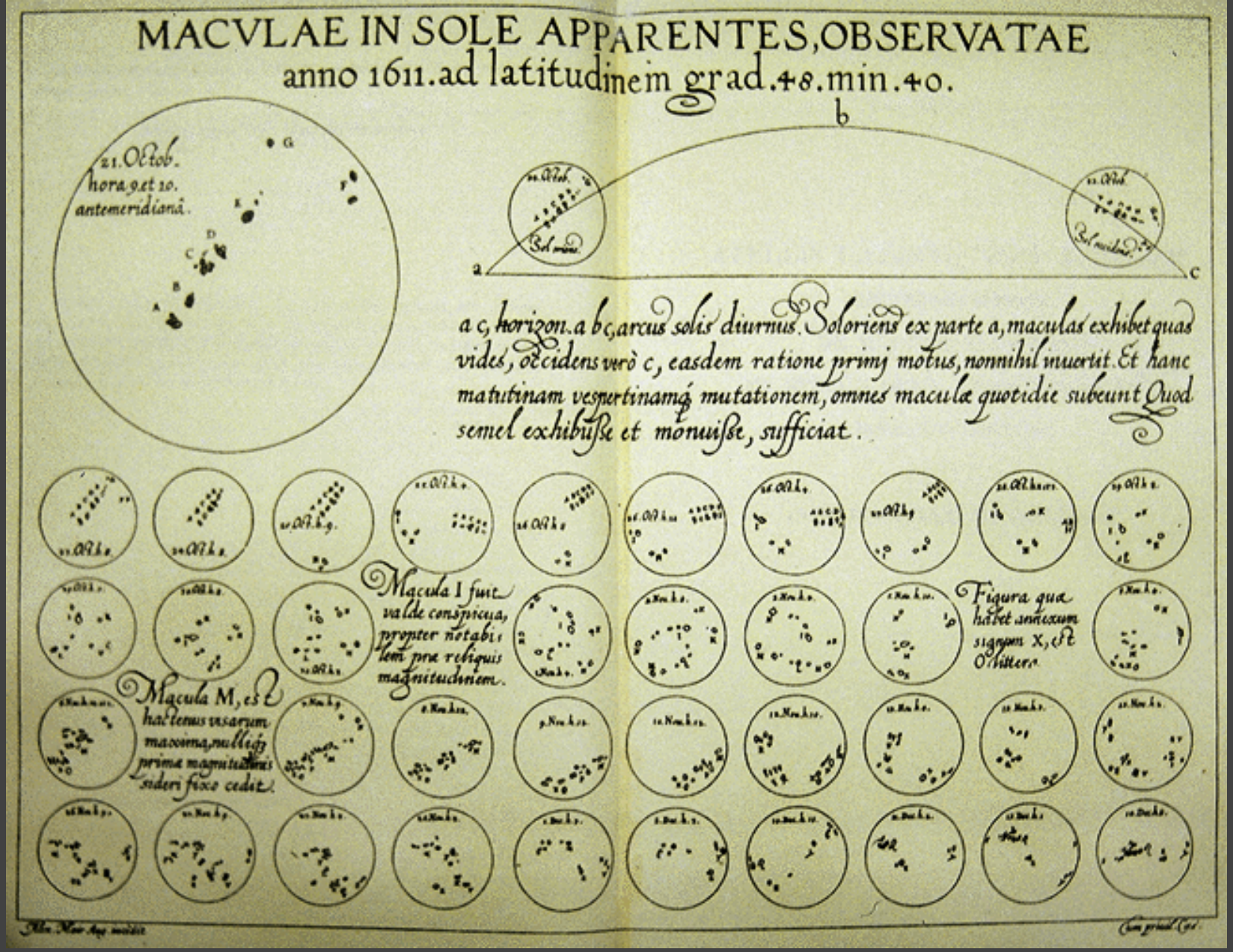


~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey

0 BC

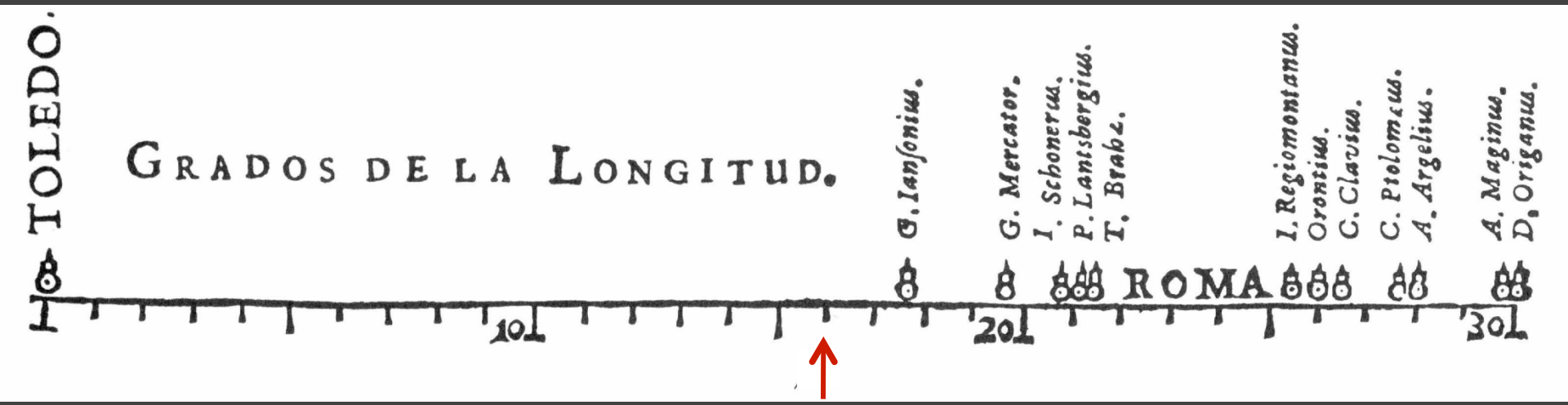


~950 AD Position of Sun, Moon and Planets

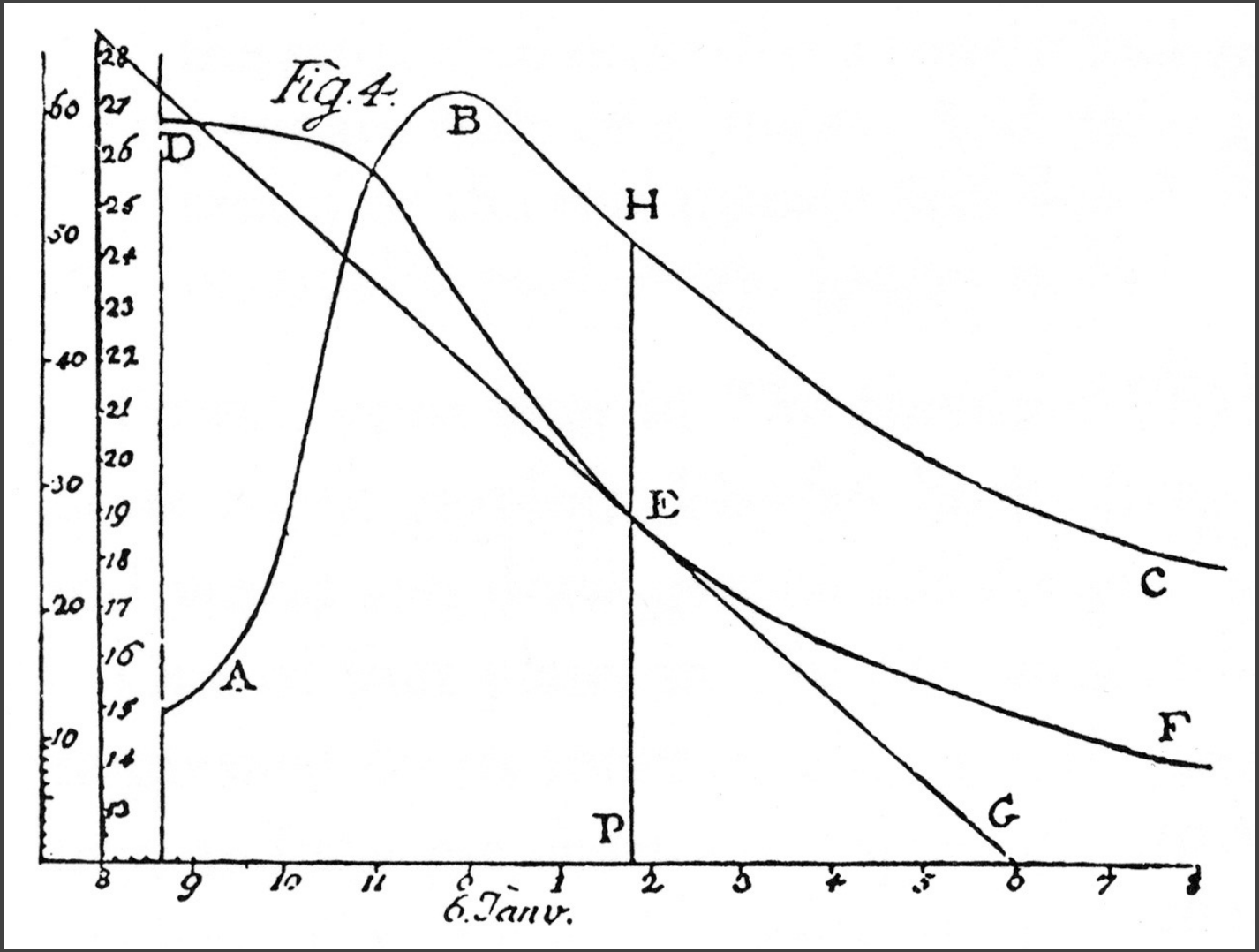


0 BC

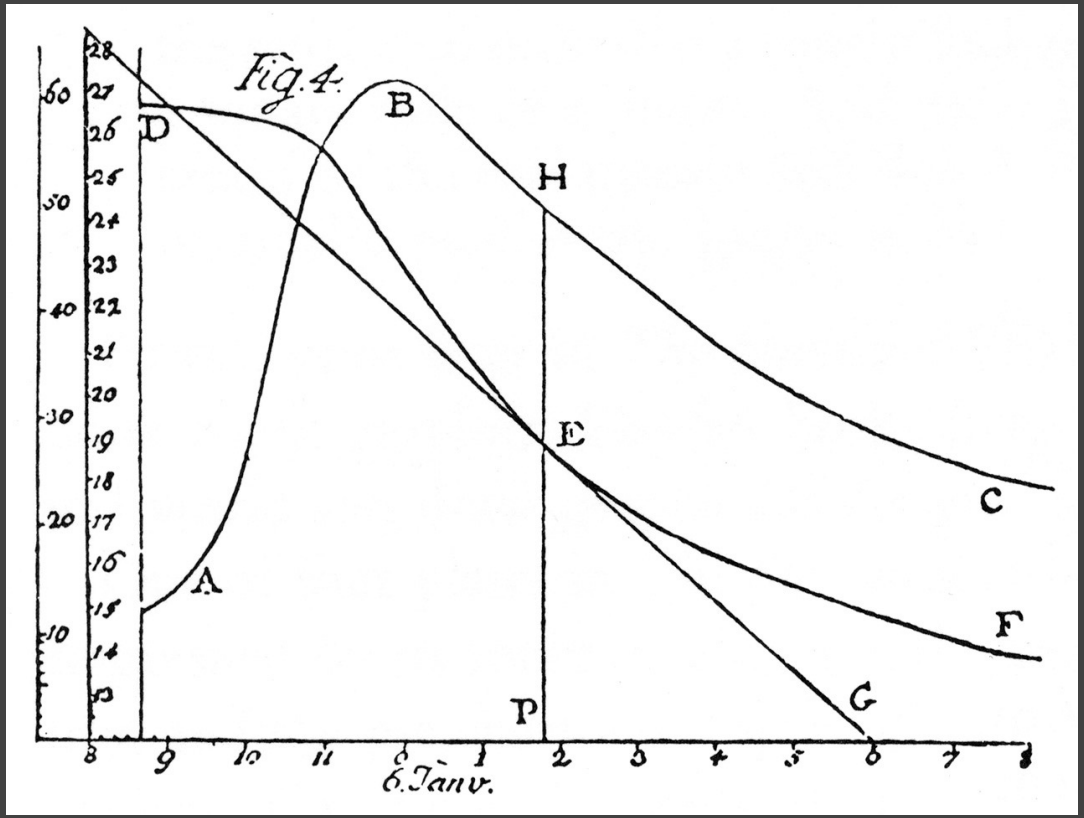
Sunspots over time, Scheiner 1626

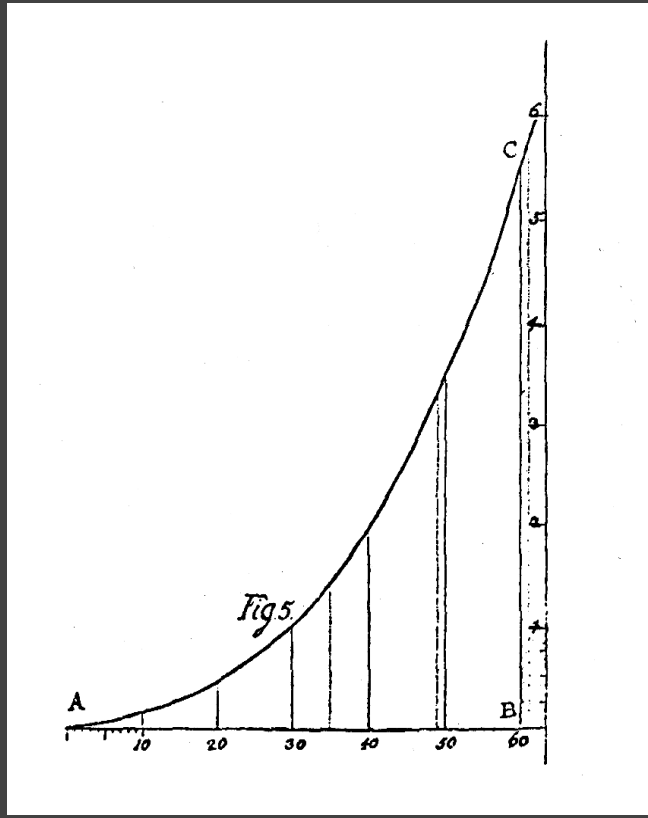


Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765

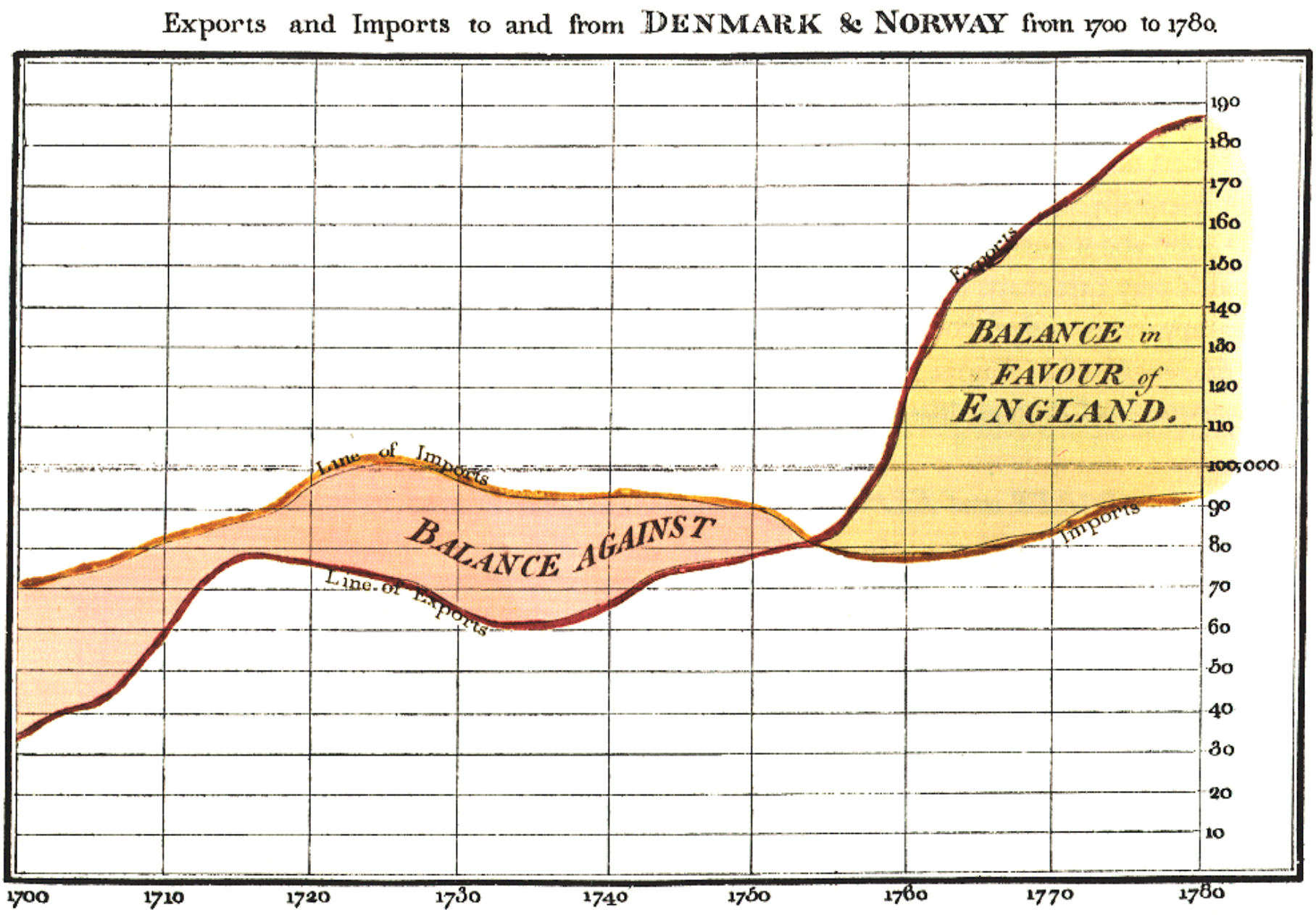




The Rate of Water Evaporation, Lambert 1765

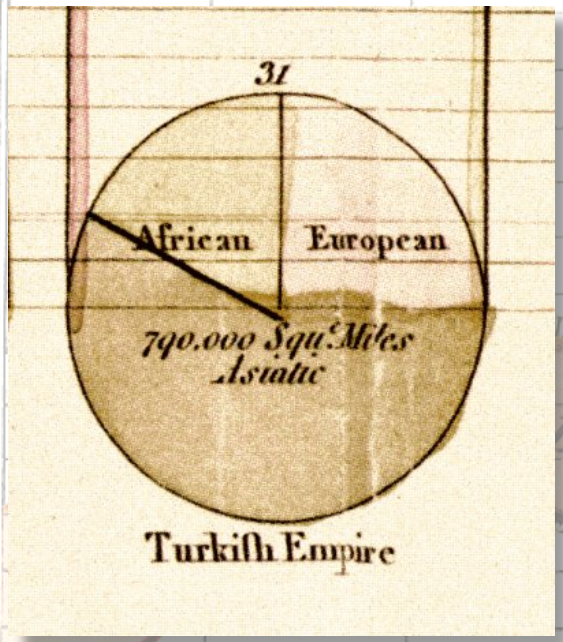
The **Golden Age** of Data Visualization

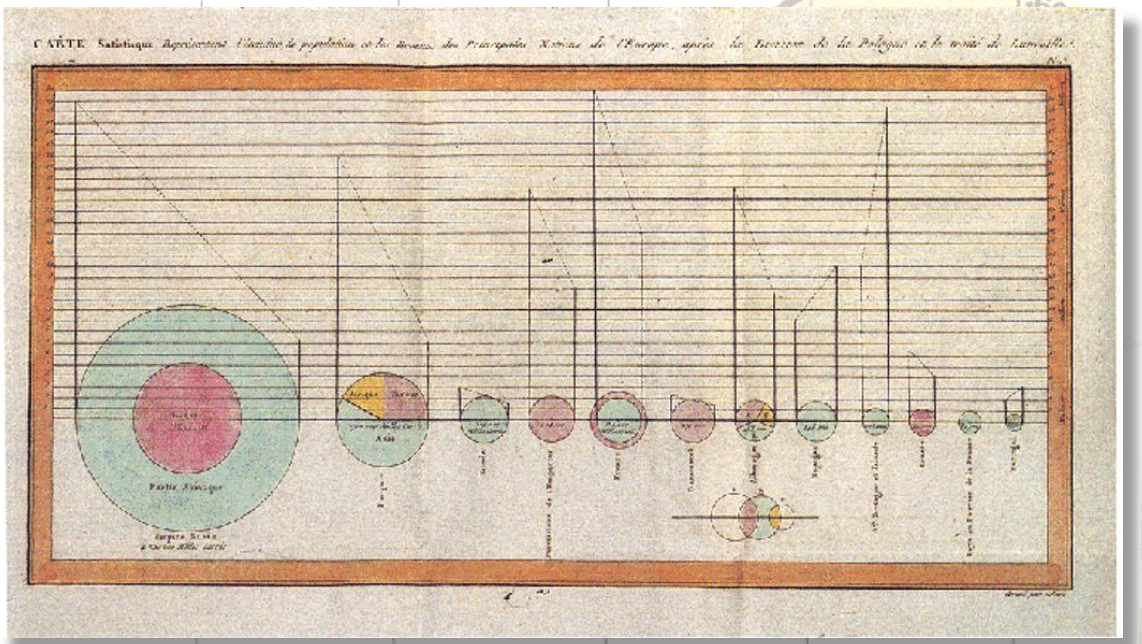
1786 1900



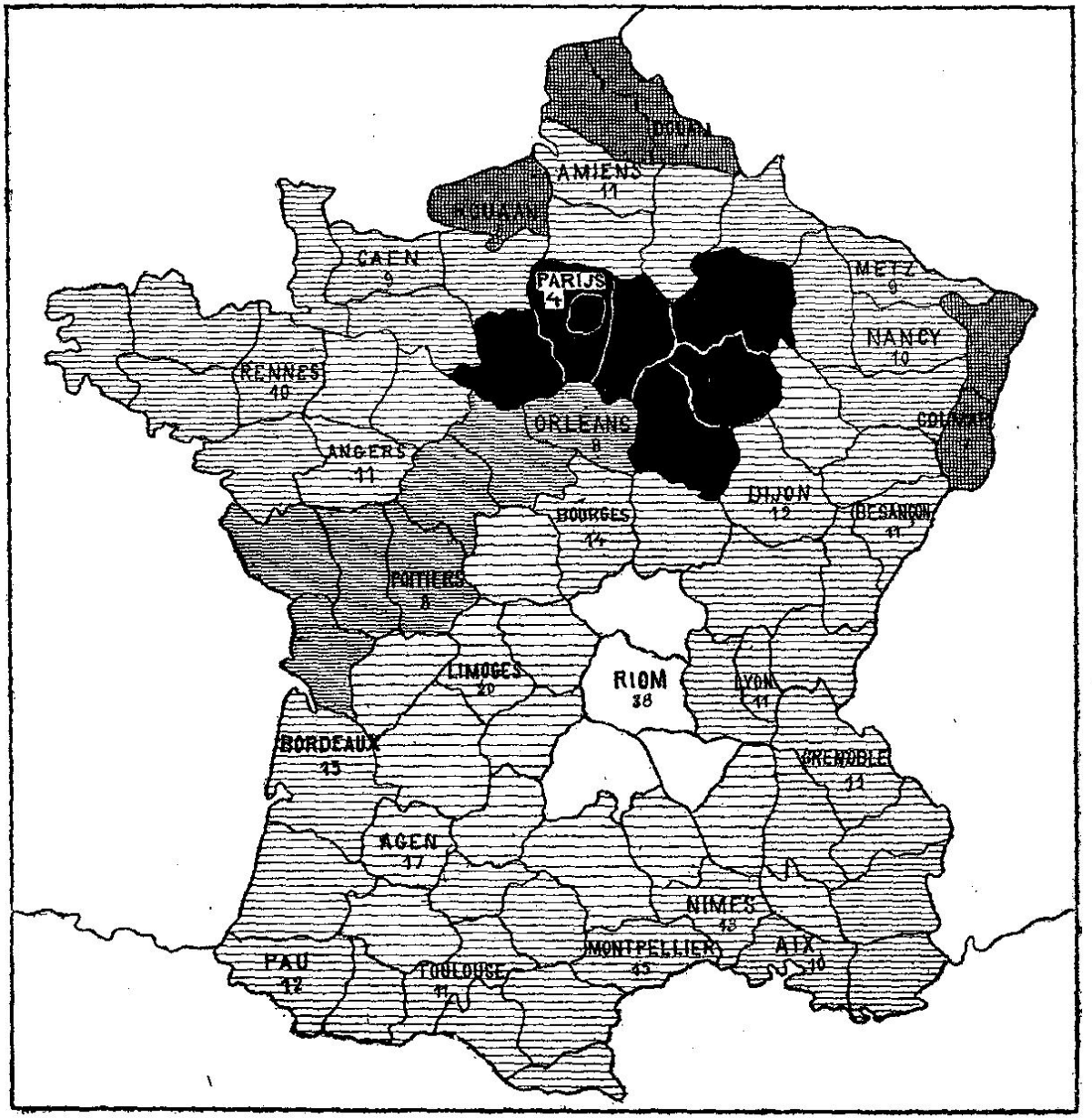
The Commercial and Political Atlas, William Playfair 1786

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



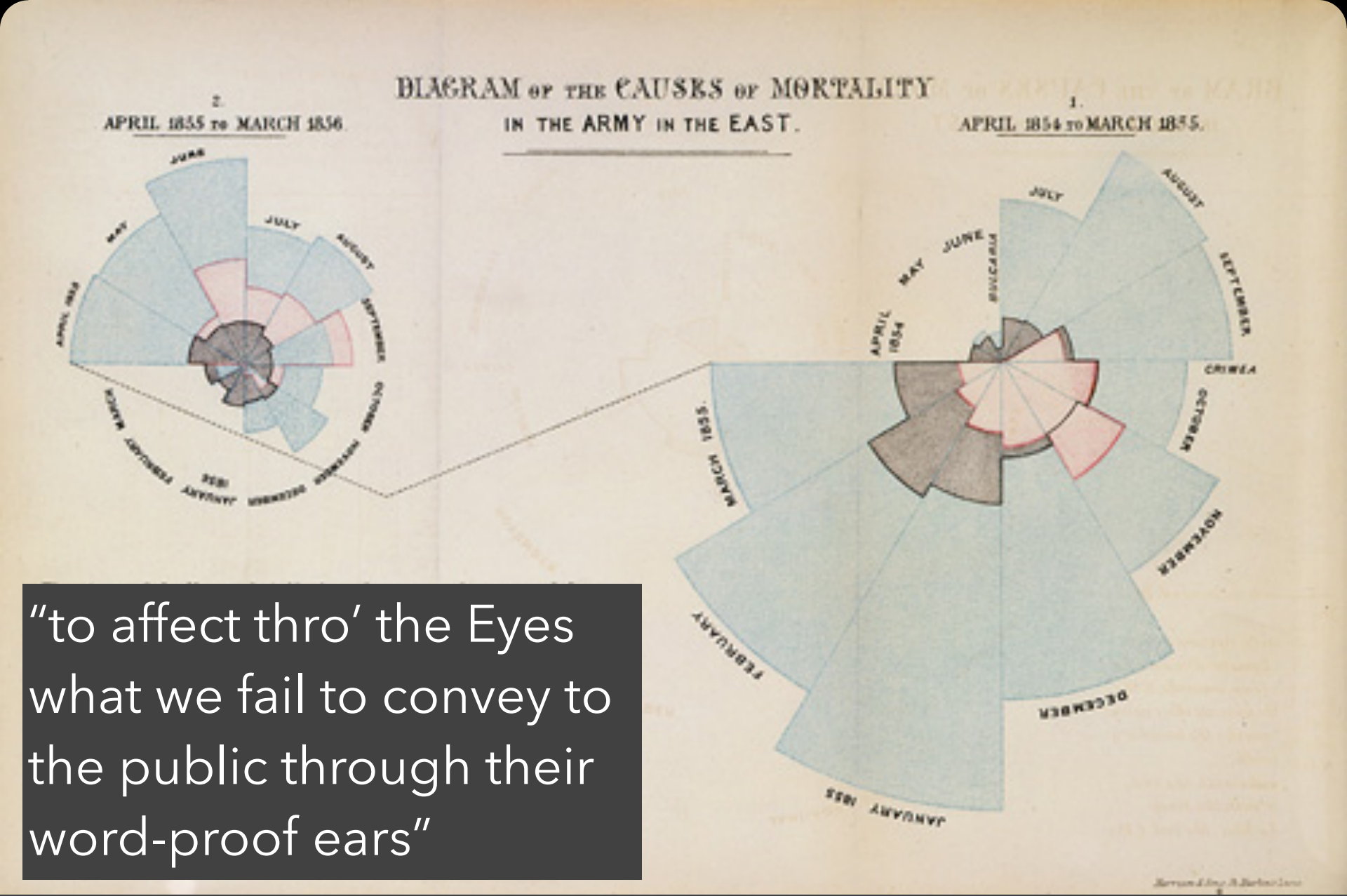


Statistical Breviary, William Playfair 1801

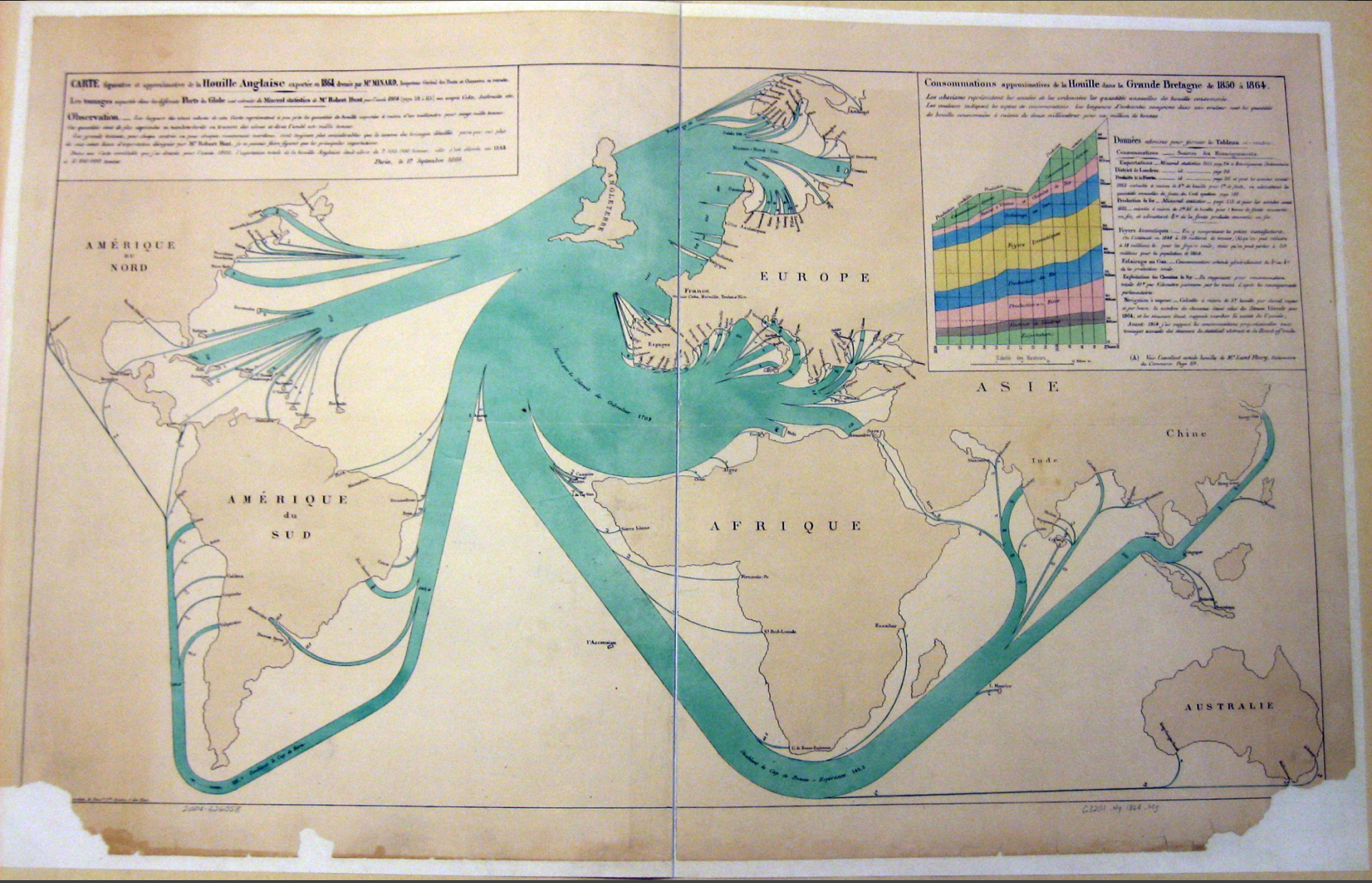


1786

1826(?) Illiteracy in France, Pierre Charles Dupin



"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

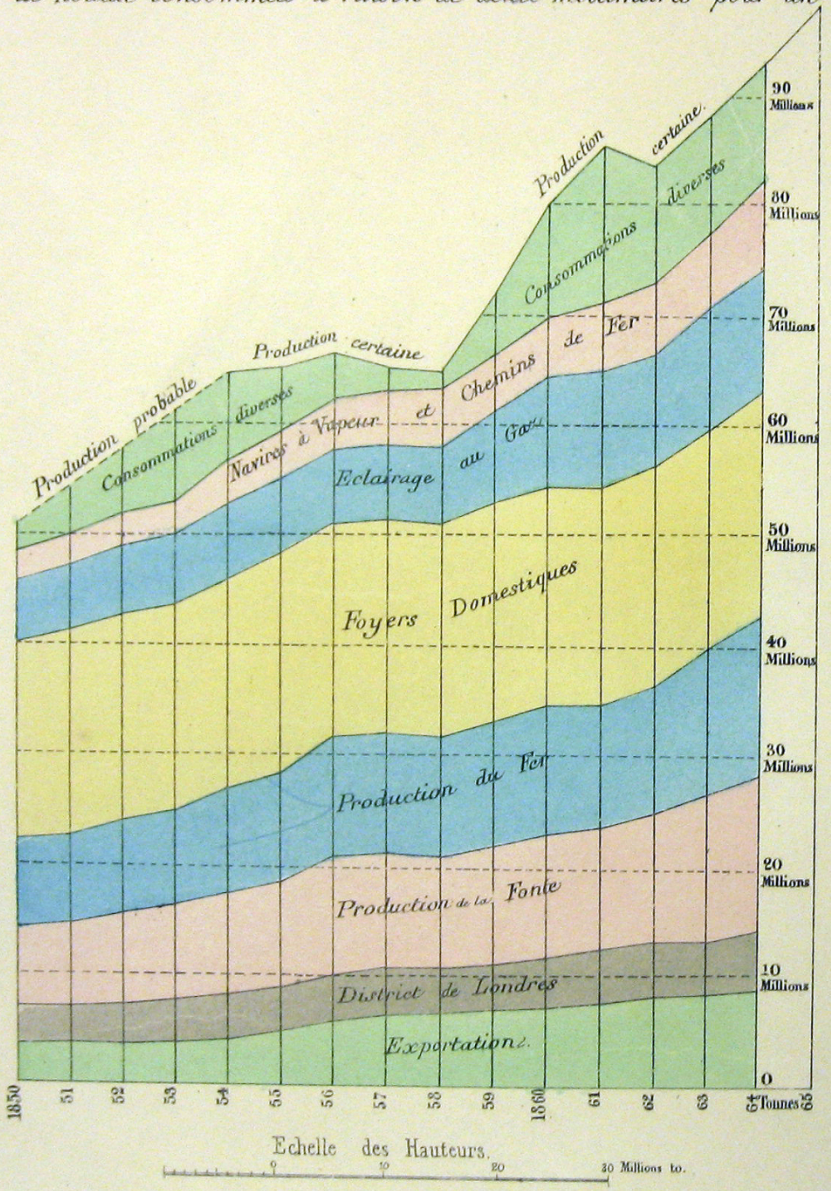


1786

1864 British Coal Exports, Charles Minard

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 to de houille pour 1 to 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 to 35 de houille pour 1 tonne de fonte convertie en fer, et admettant $\frac{2}{10}$ es de la fonte produite convertis en fer.*

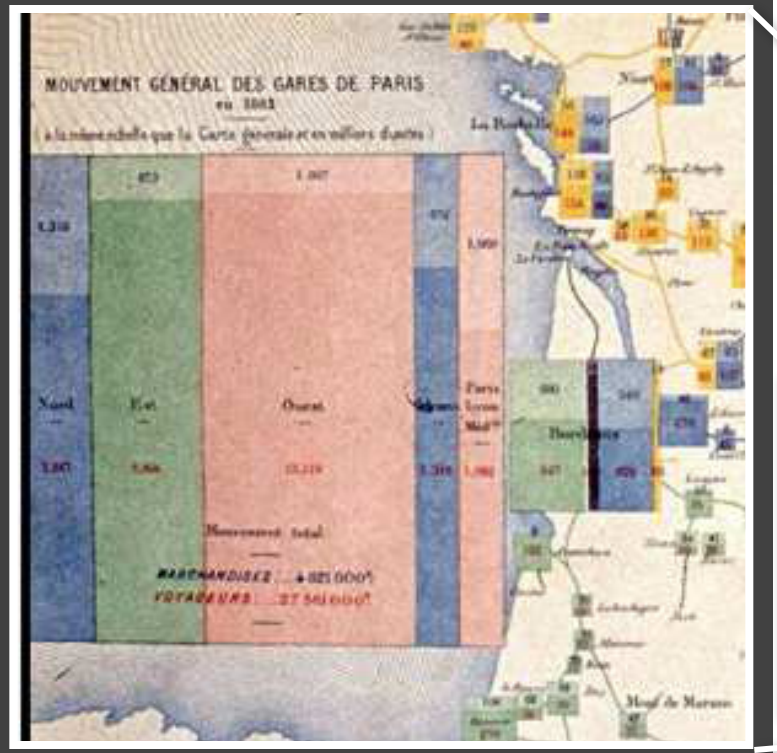
Foyers domestiques. — *En y comprenant les petites manufactures. On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.*

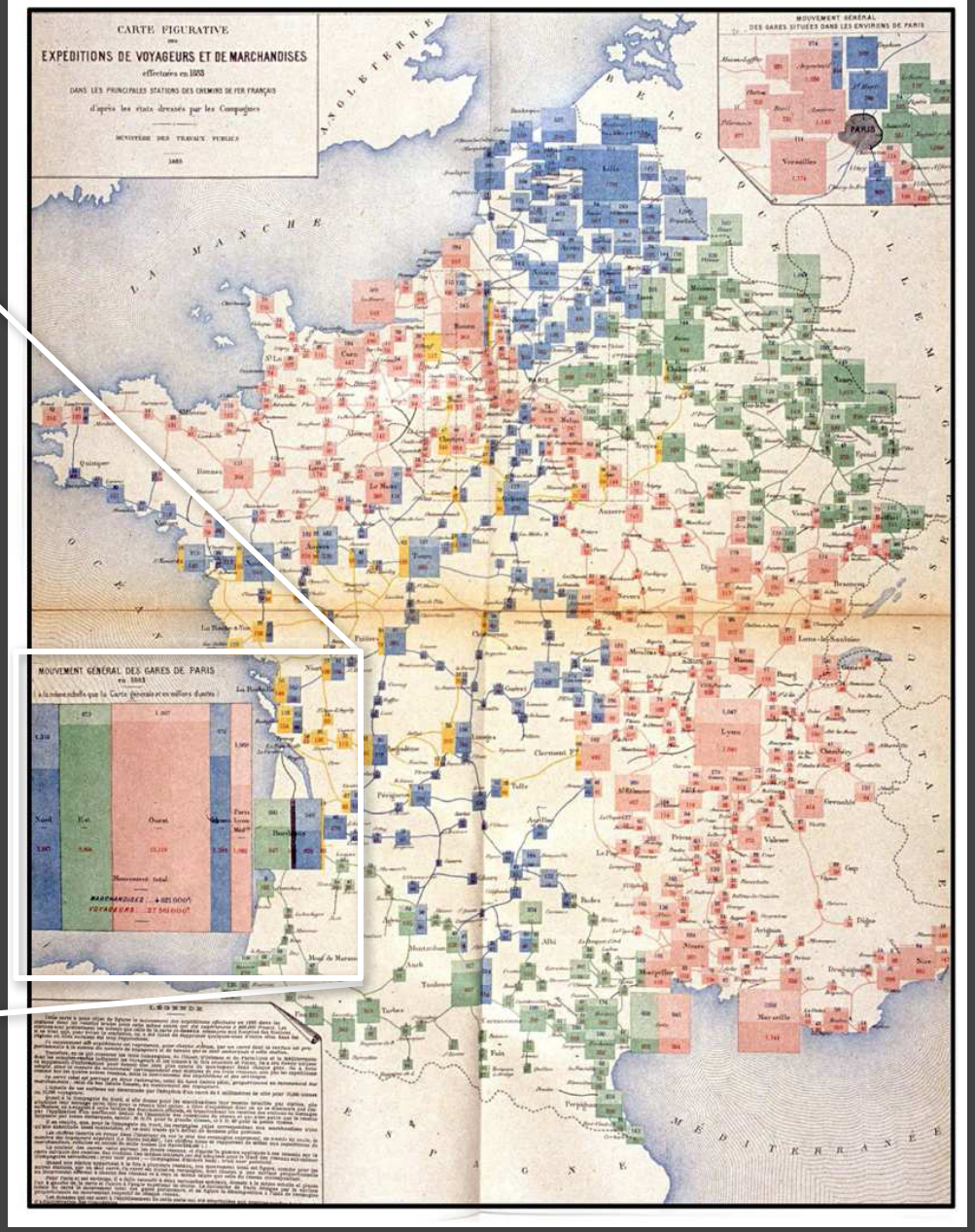
Eclairage au Gaz. — *Consommation estimée généralement du $\frac{1}{3}$e au $\frac{1}{8}$e de la production totale.*

Exploitation des Chemins de Fer. — *En supposant pour consommation totale 10 K par Kilomètre parcouru par les trains d'après les renseignements parlementaires.*

Navigation à vapeur. — *Calculée à raison de 5 K 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 111.*

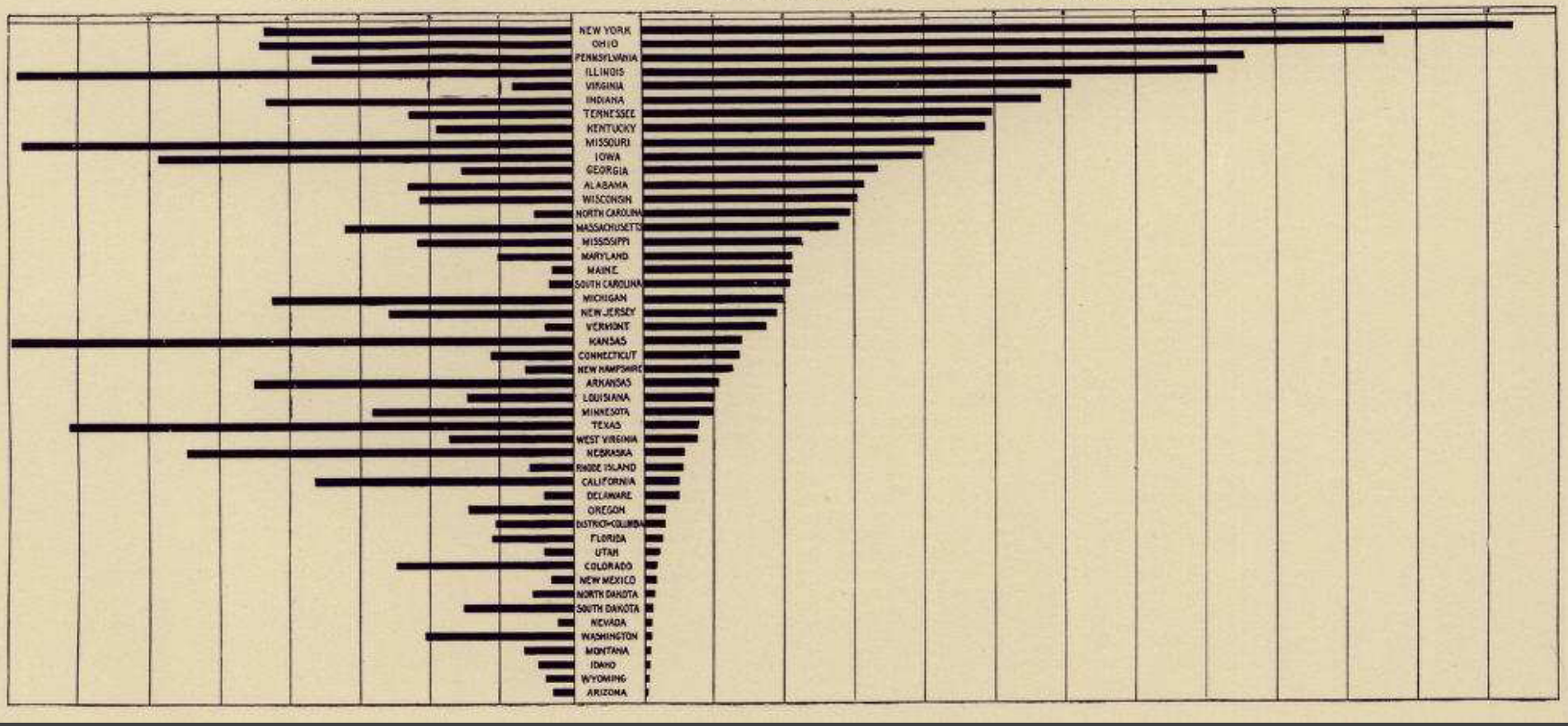




1786

1884 Rail Passengers and Freight from Paris

66. Interstate Migration—Number of Native Immigrants and Native Emigrants, by States and Territories: 1890.



1786

1890 Statistical Atlas of the Eleventh U.S. Census

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**


LIFE

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.

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.

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

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**.

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

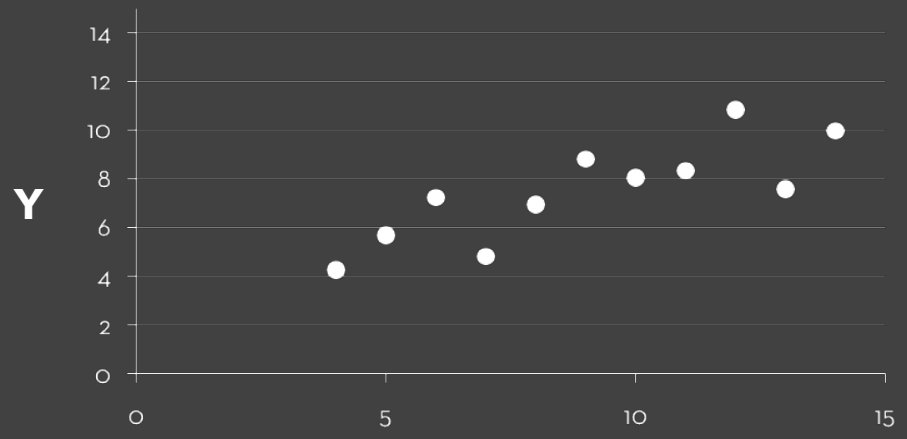
Linear Regression

$Y = 3 + 0.5\ X$

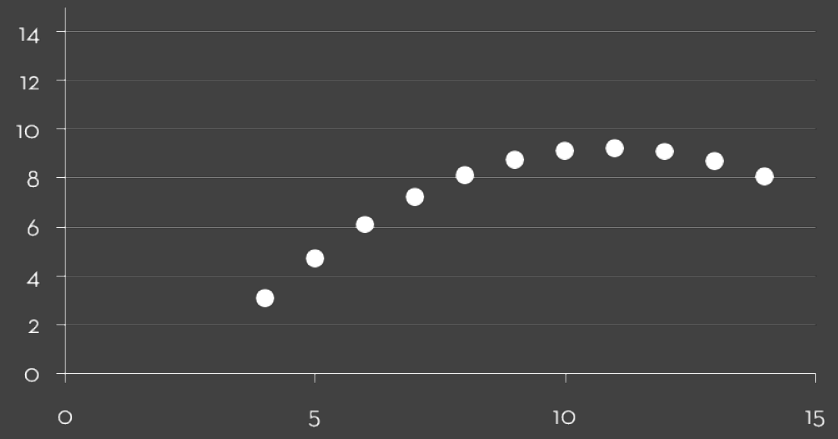
$R^2 = 0.67$

[Anscombe 1973]

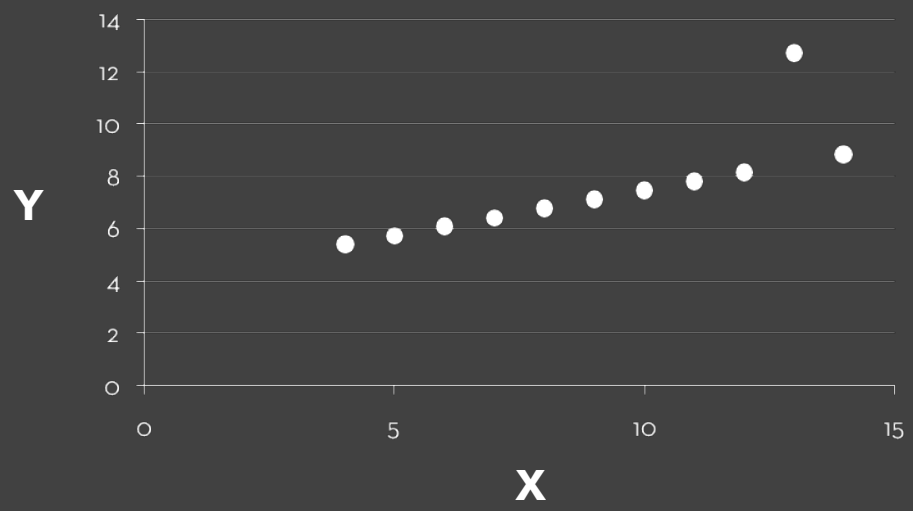
Set A



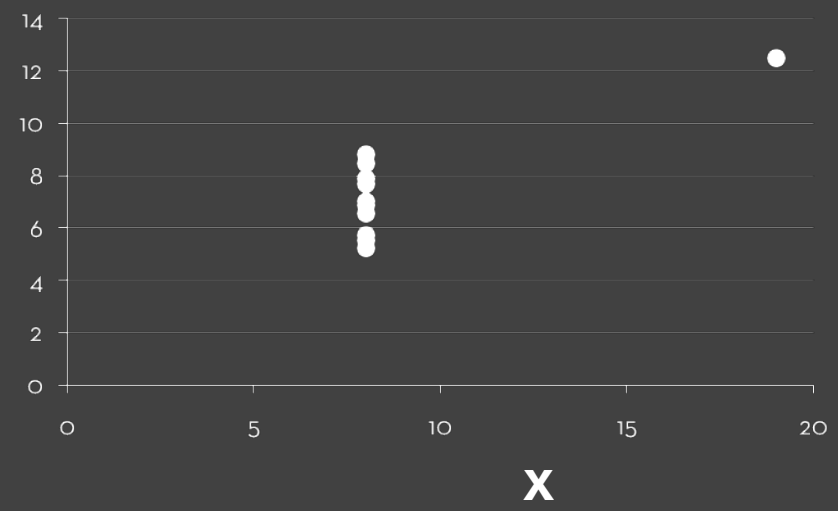
Set B



Set C



Set D



Topics

Exploratory Data Analysis
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. '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.

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

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

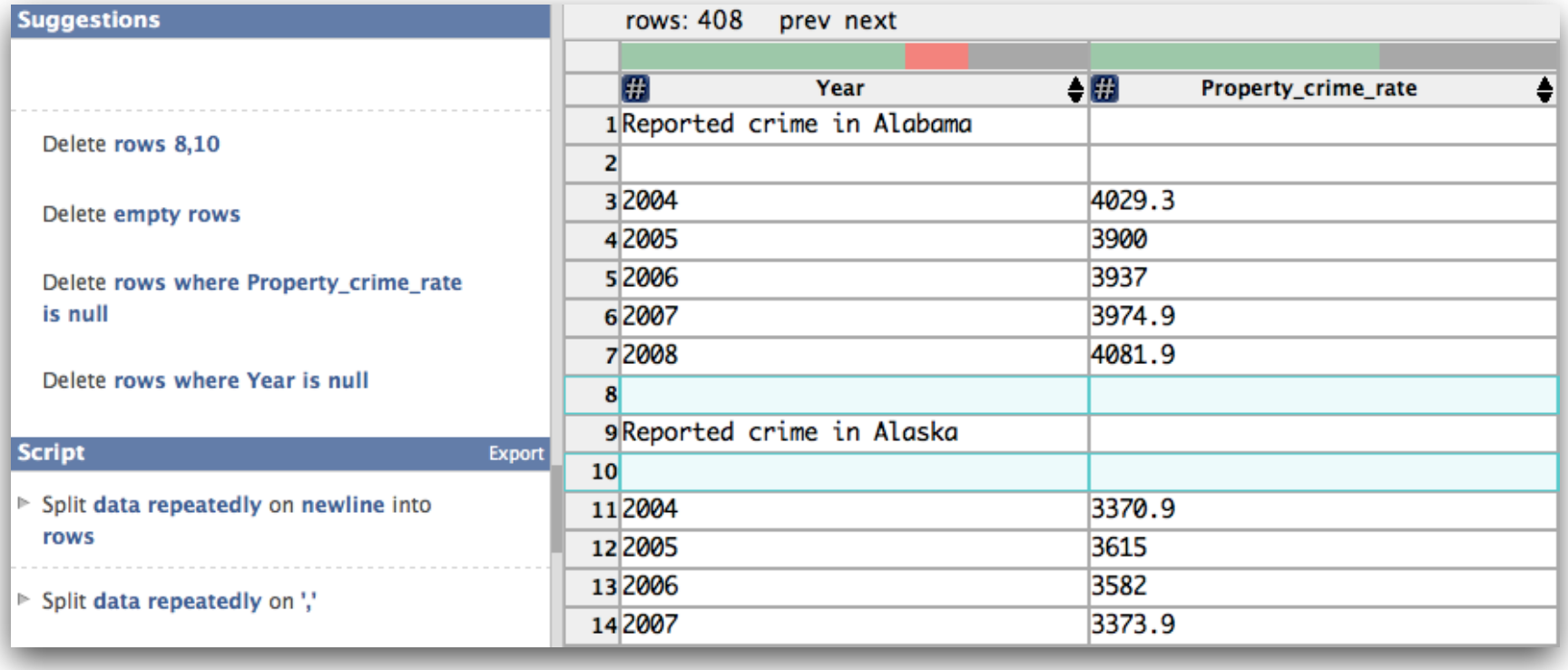
Reported crime in California

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



Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel et al. *CHI'11*

Demo:
https://vimeo.com/19185801

Data Wrangling

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

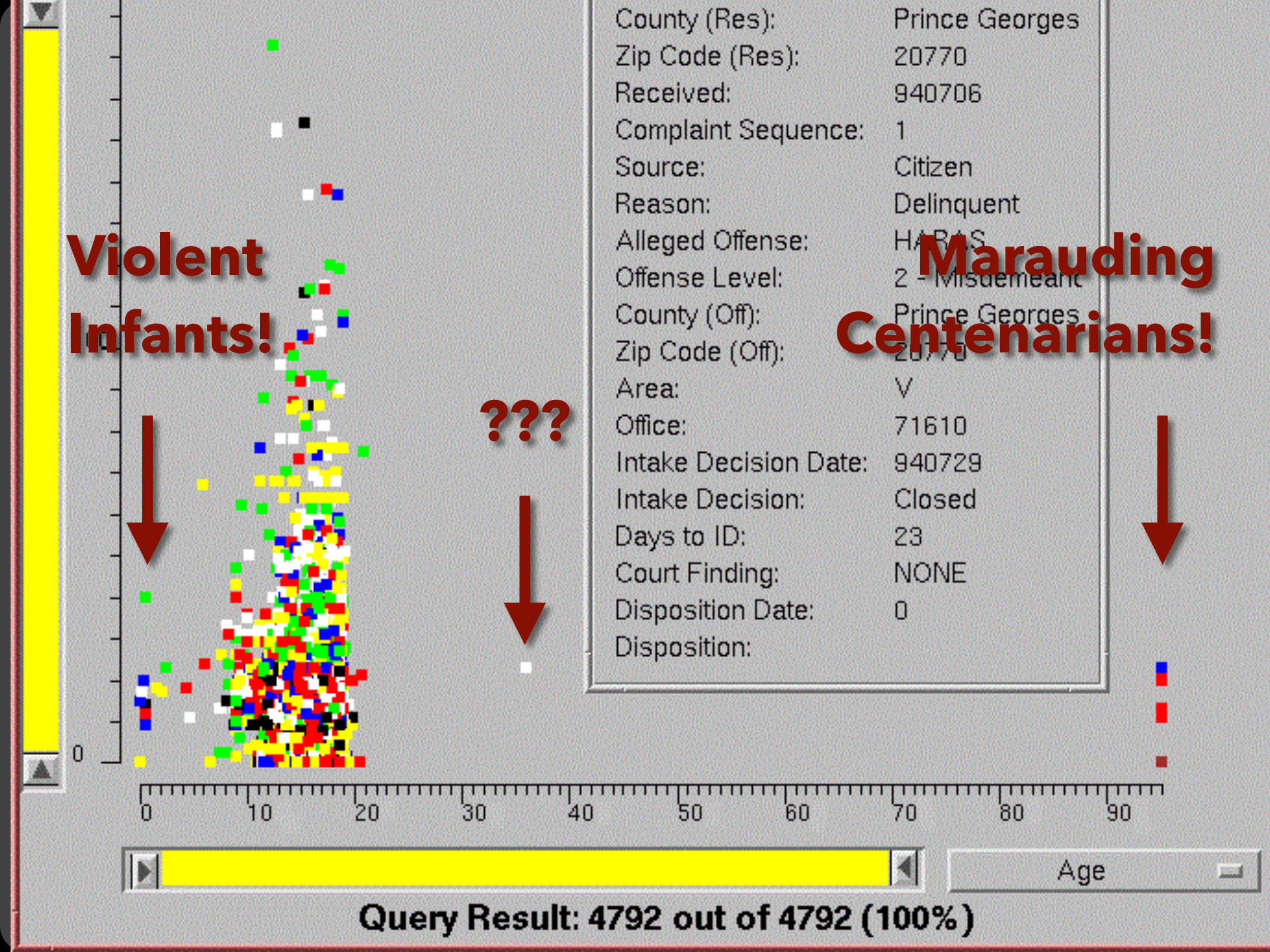
Approaches include:
Manual manipulation in spreadsheets
Custom code (e.g., dplyr in R, Pandas in Python)
Trifacta Wrangler http://www.trifacta.com/products/wrangler/
Open Refine http://openrefine.org/

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


Violent Infants!
???
Marauding Centenarians!
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanr
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:
0
0 10 20 30 40 50 60 70 80 90
Age
Query Result: 4792 out of 4792 (100%)

Graph Viewer

Graph Viewer

Roll-up by:

All

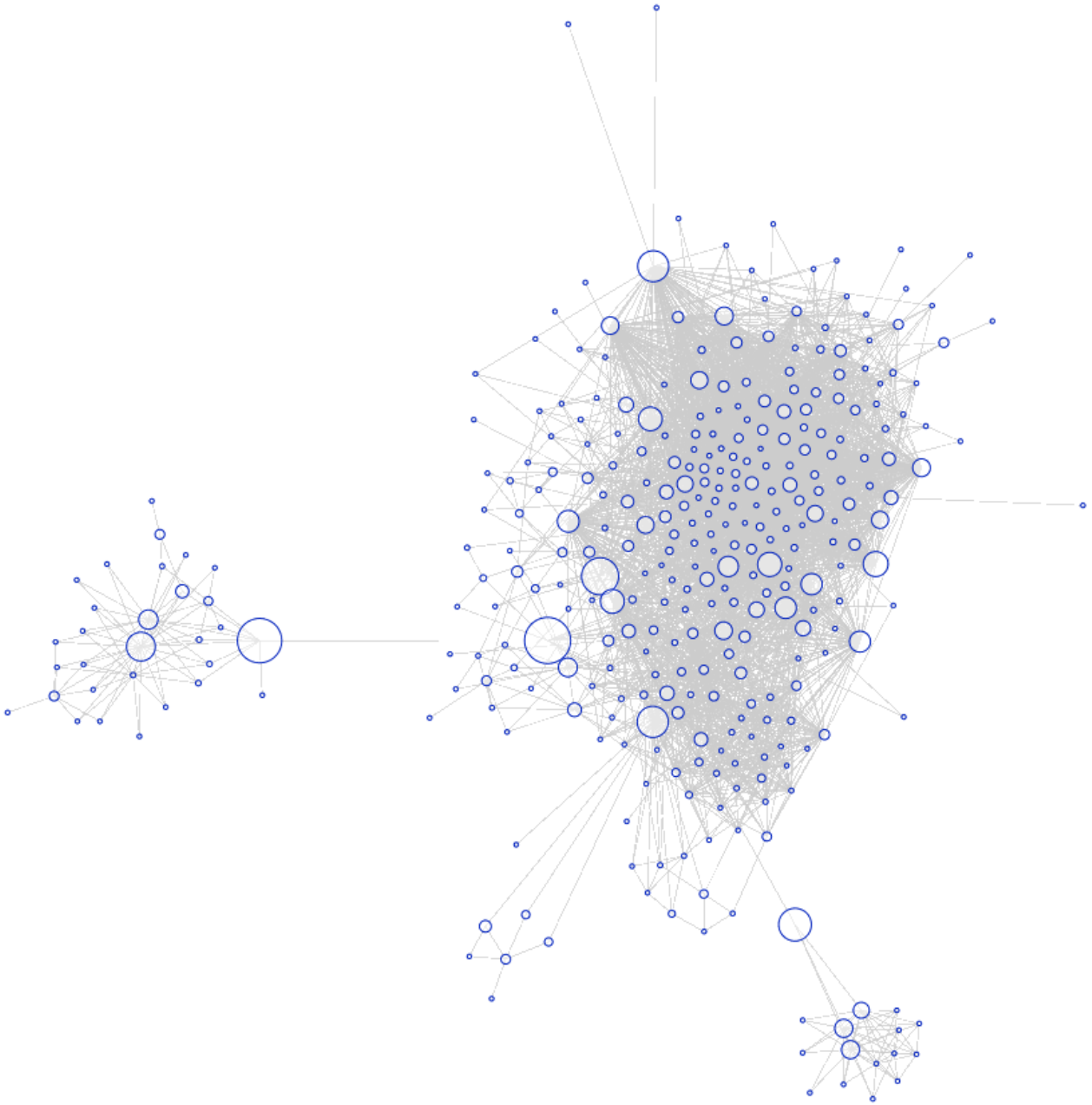
Visualization:

Node-Link

Sort by:

None

Edge centrality filters:



Images

Animate

Graph Viewer

Graph Viewer

Roll-up by:

All

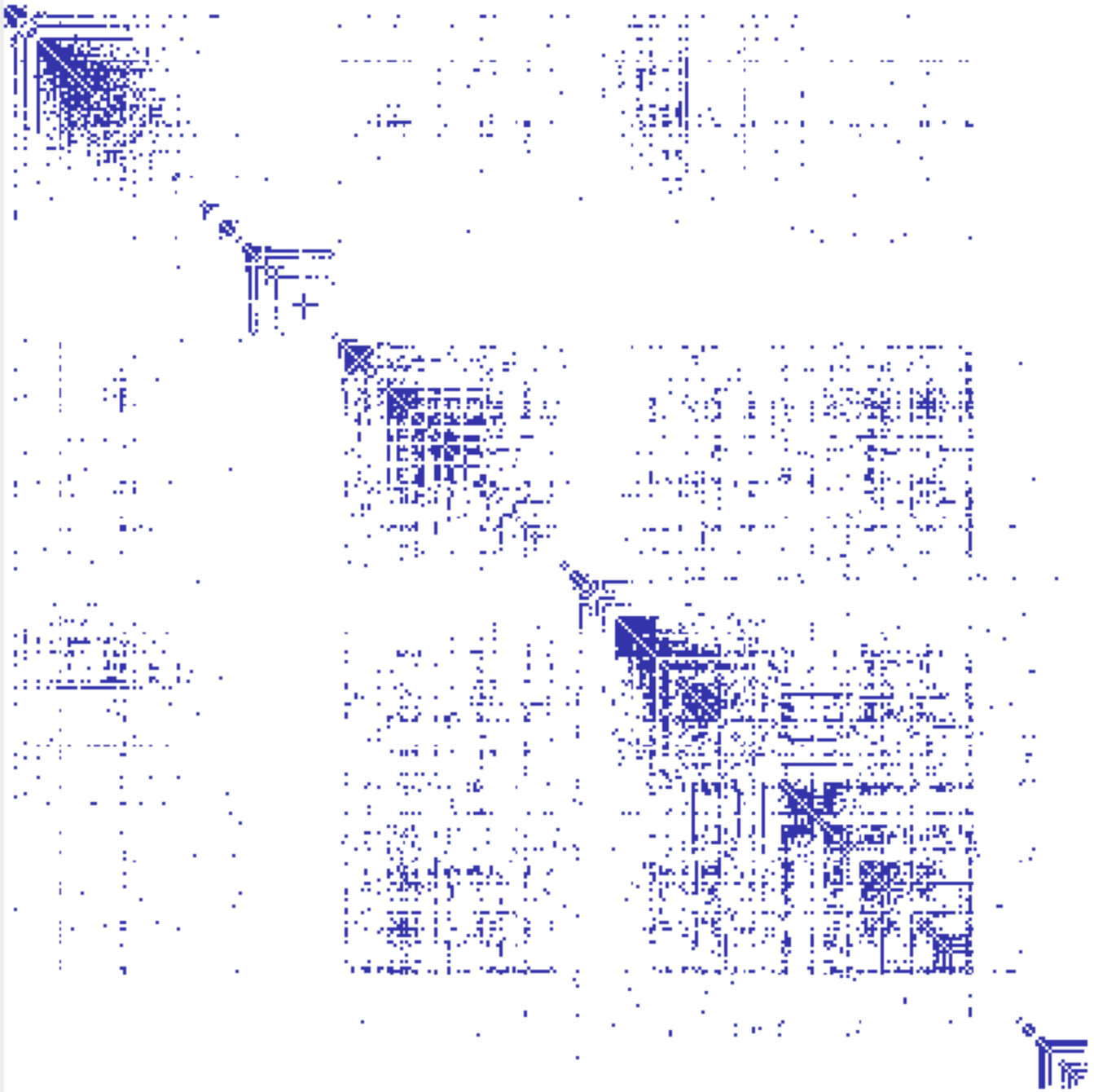
Visualization:

Matrix

Sort by:

Linkage

Edge centrality filters:



Graph Viewer

Graph Viewer

Roll-up by:

All

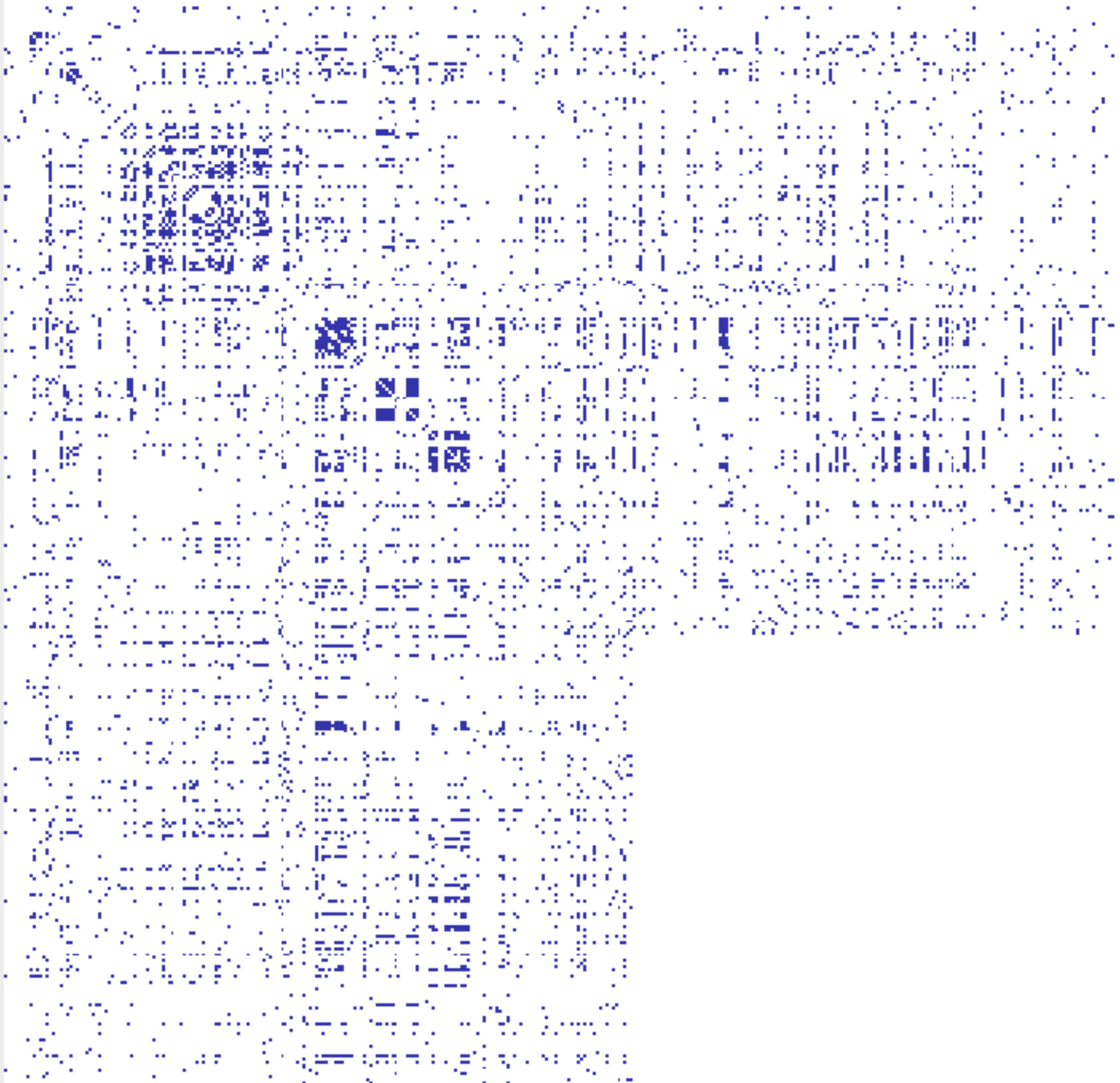
Visualization:

Matrix

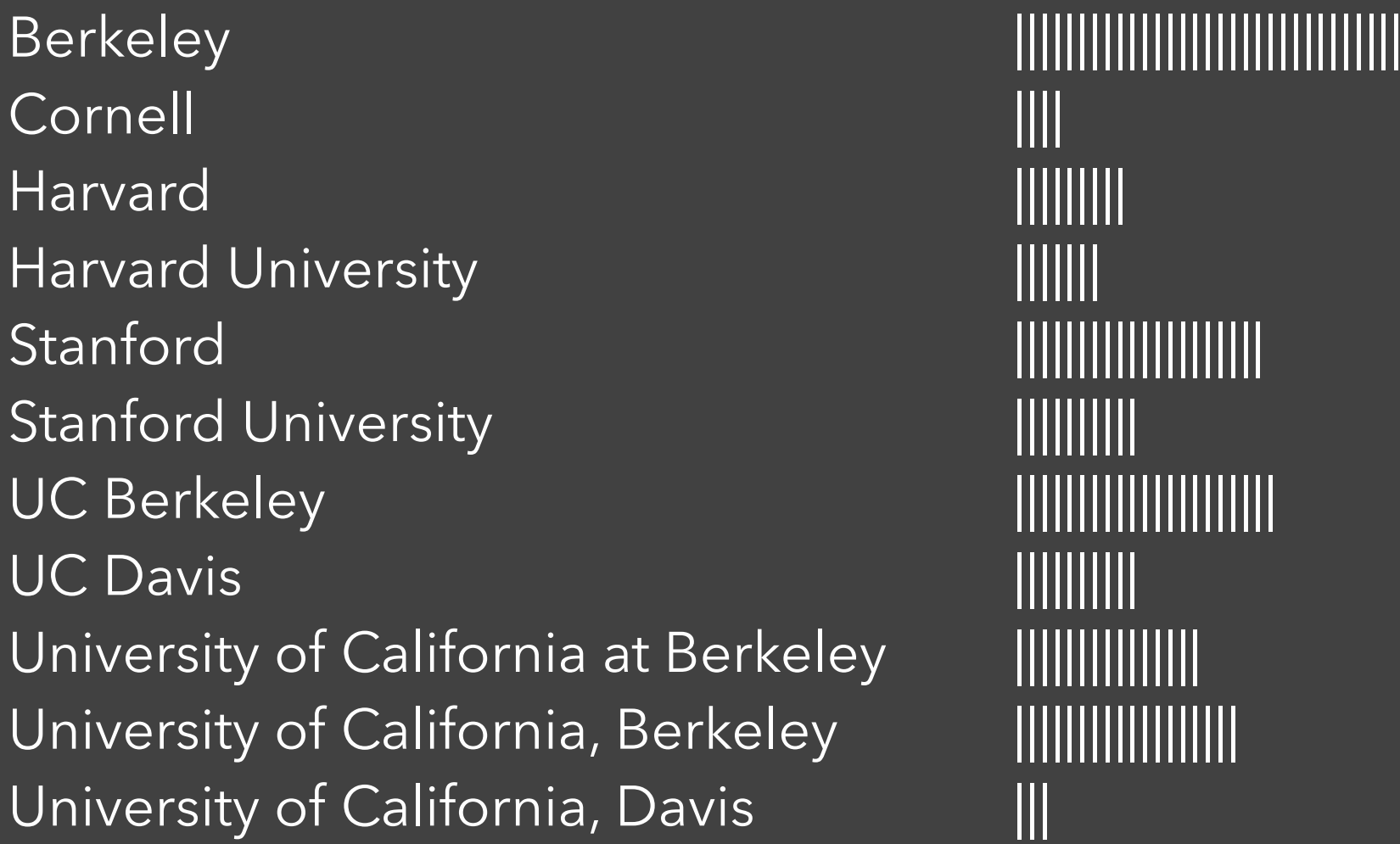
Sort by:

None

Edge centrality filters:



Visualize Friends by School?



Data Quality Hurdles

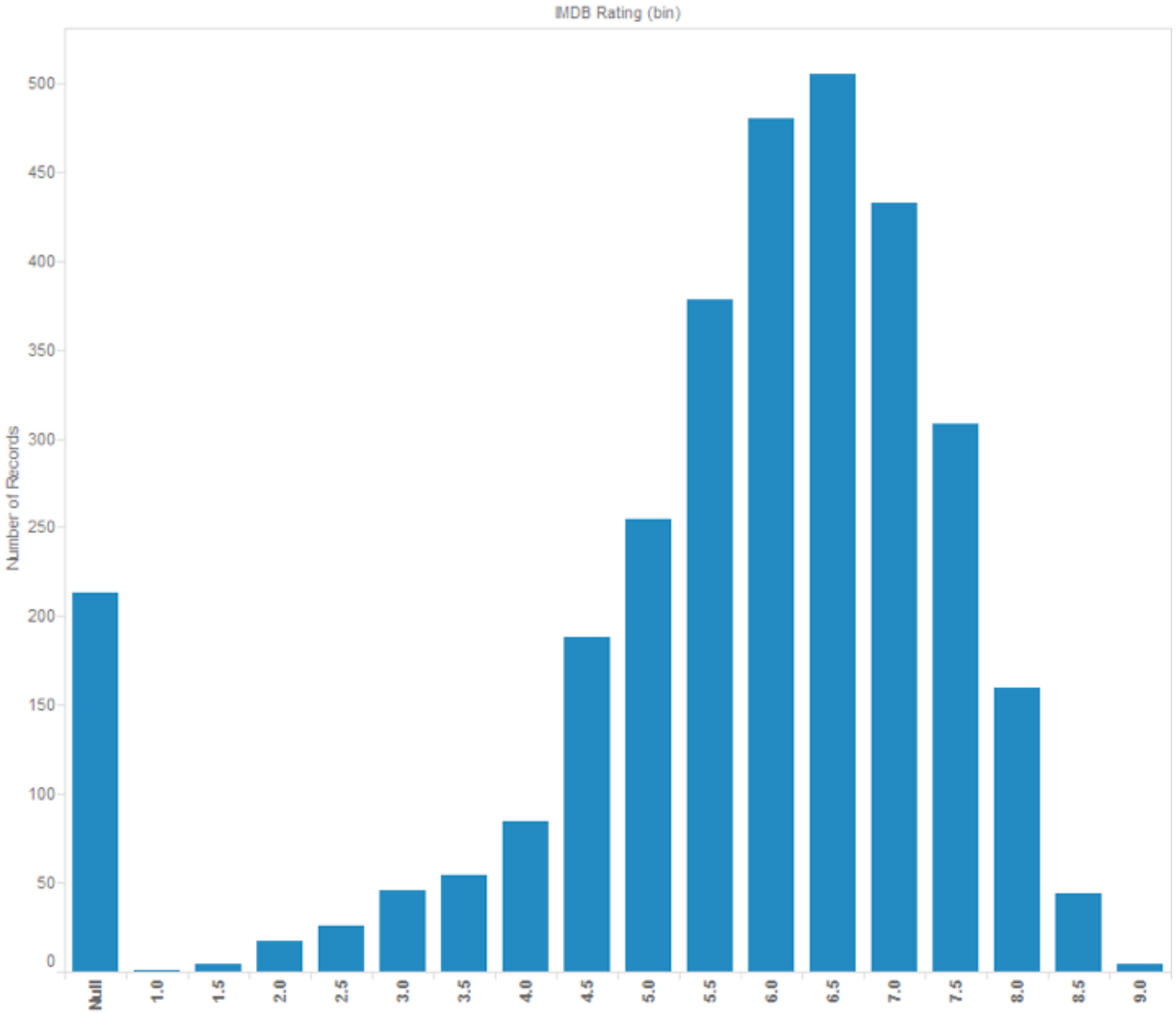
Missing Data	no measurements, redacted, …?
Erroneous Values	misspelling, outliers, …?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

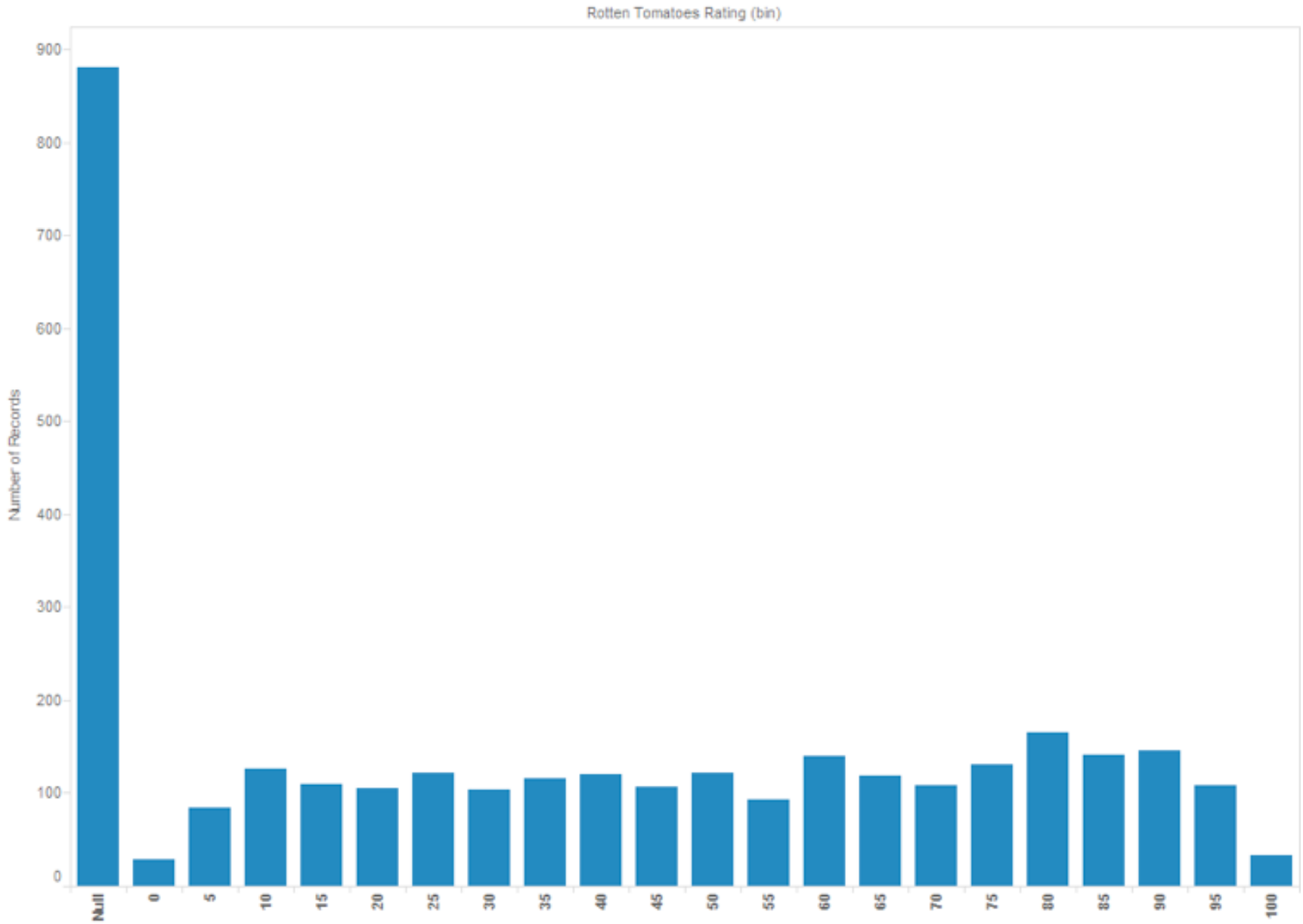
LESSON: Anticipate problems with your data.
Many research problems around these issues!

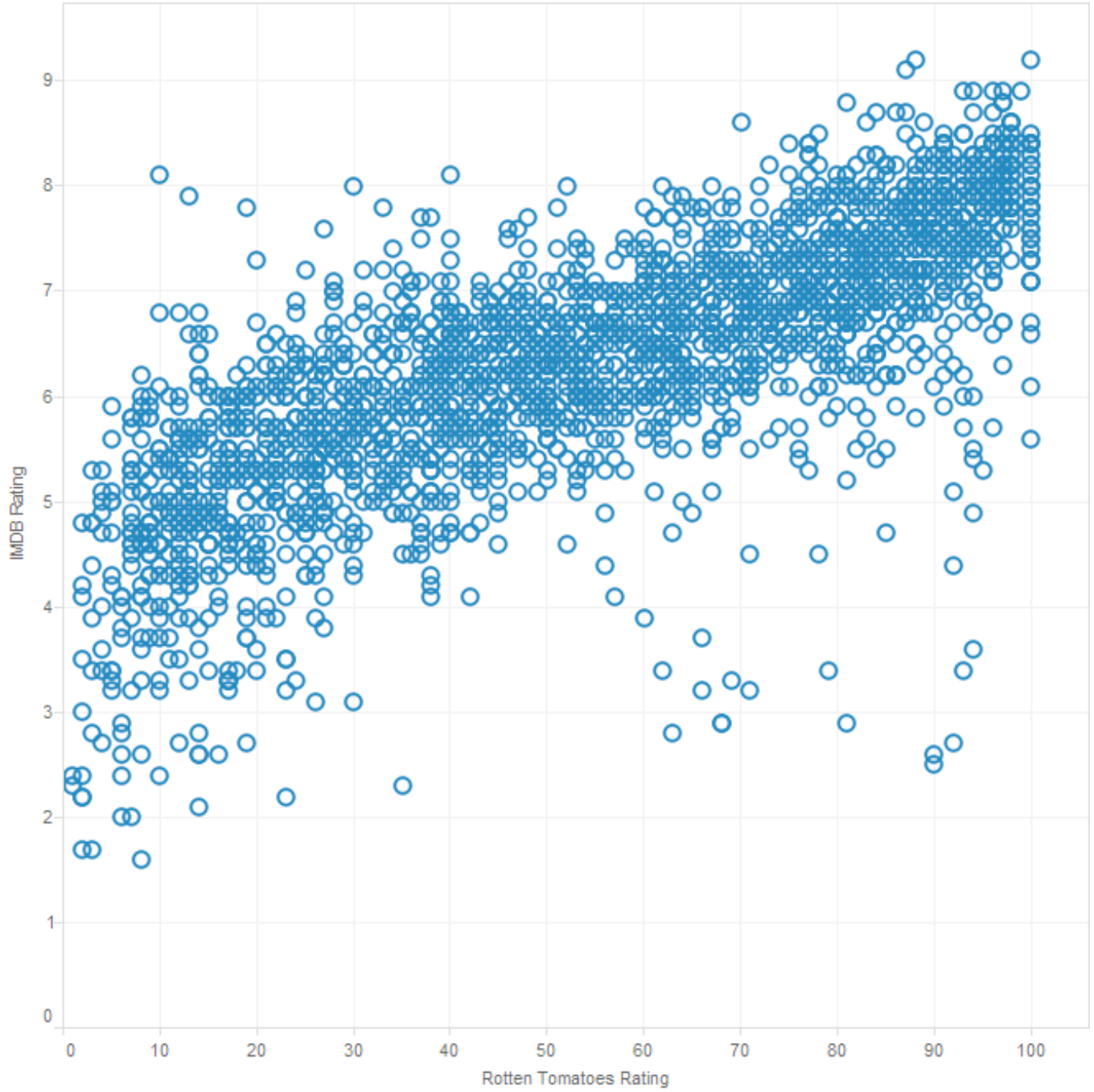
Analysis Example: Motion Pictures Data

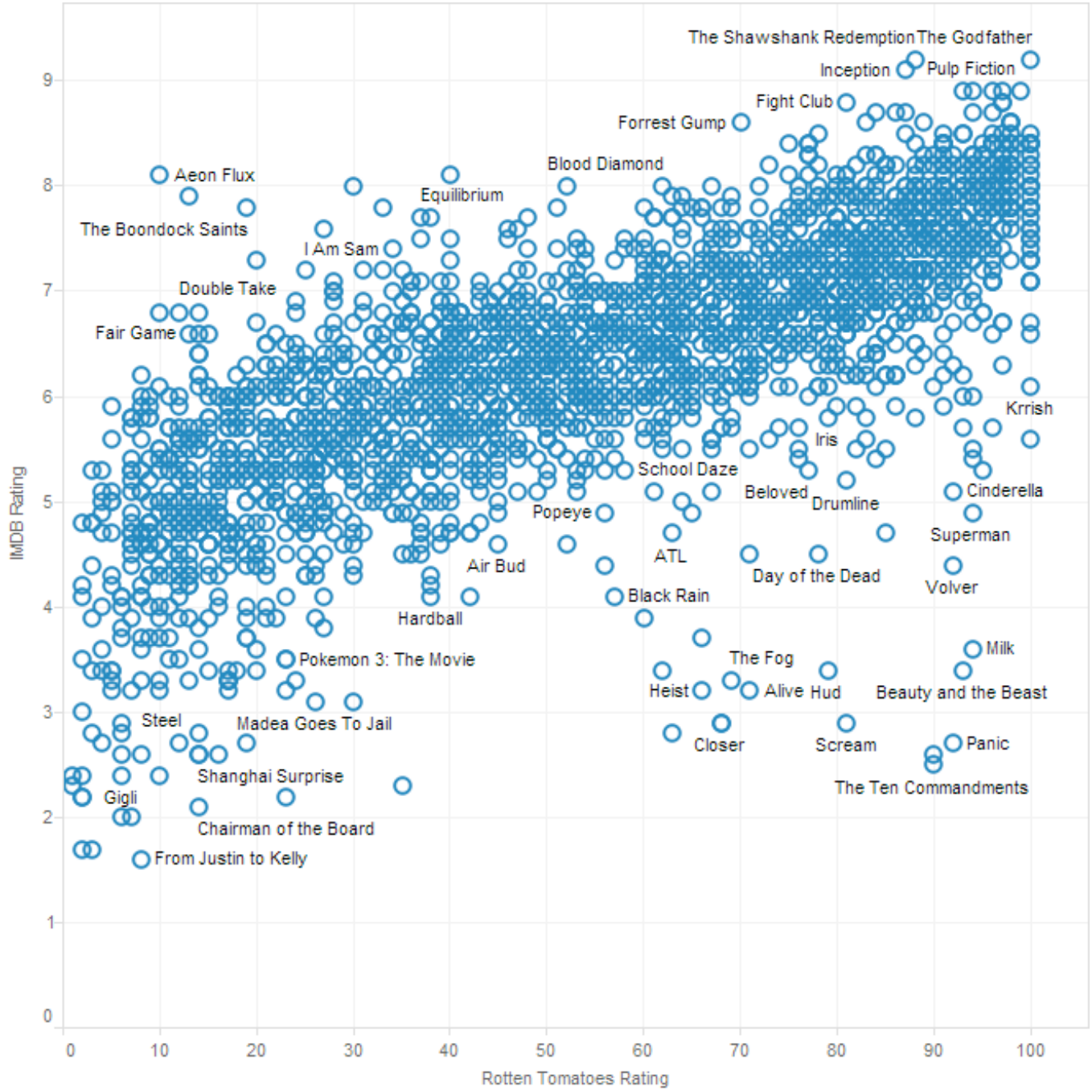
Motion Pictures Data

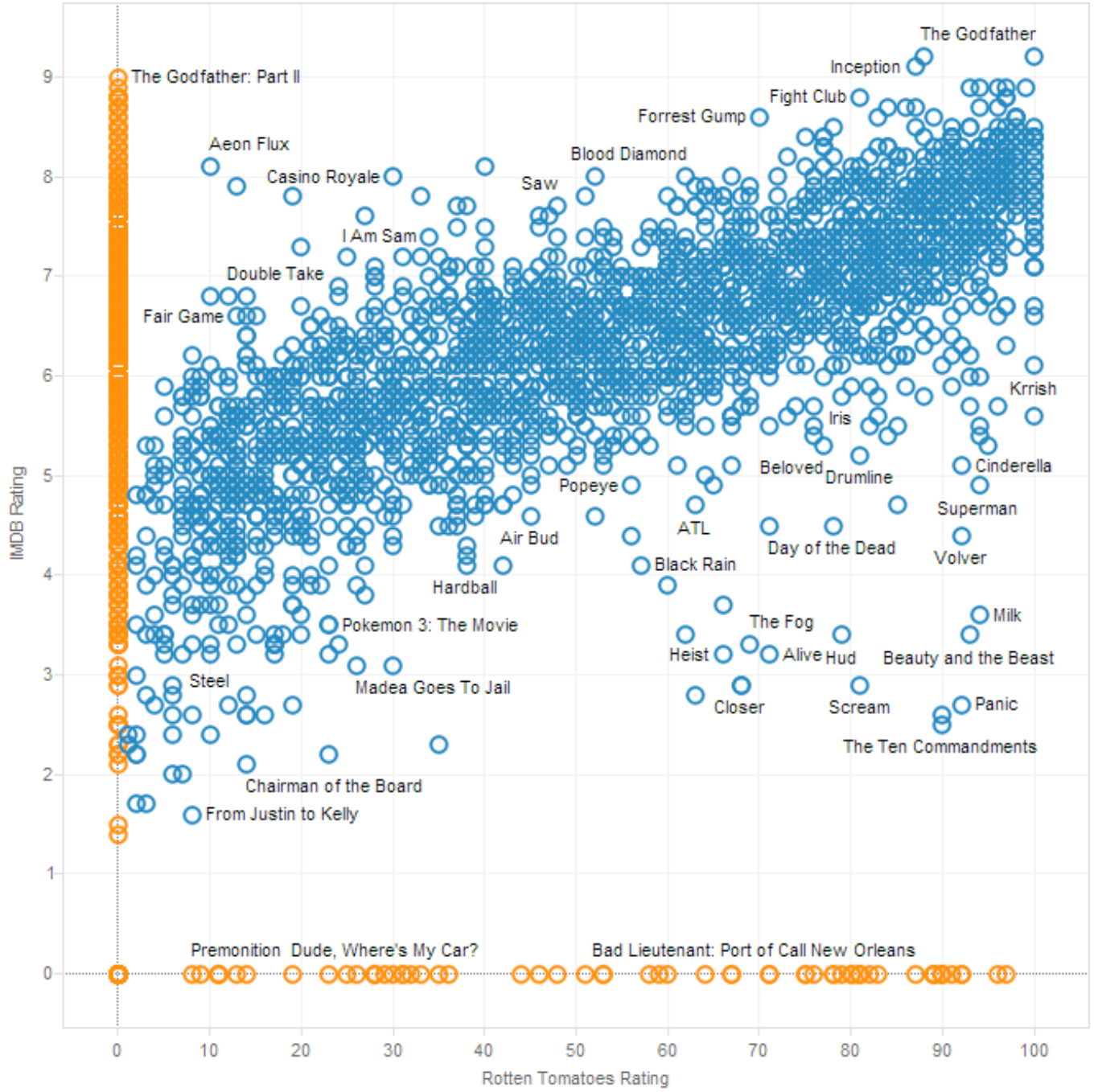
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)

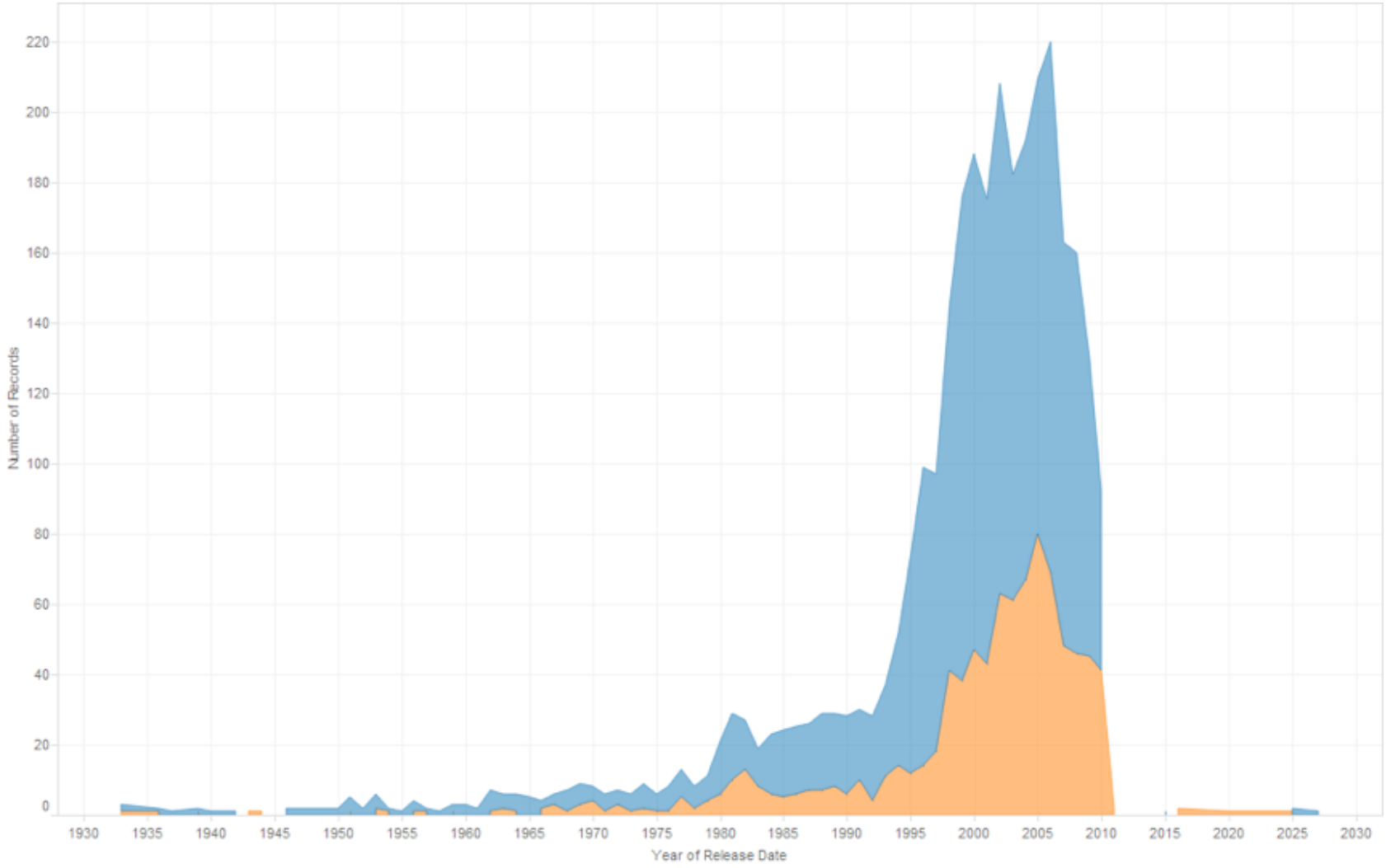

IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0


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

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)

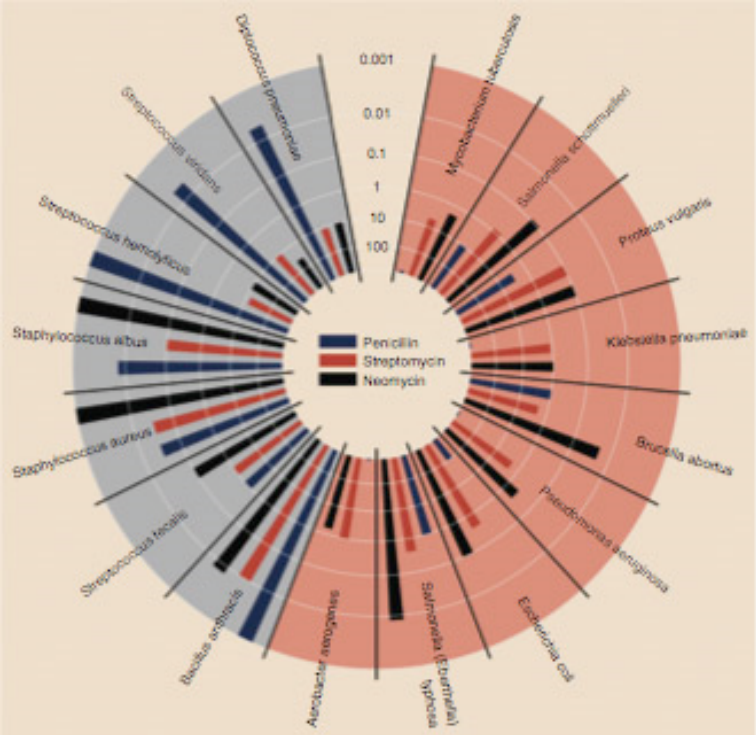
Collected prior to 1951.

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

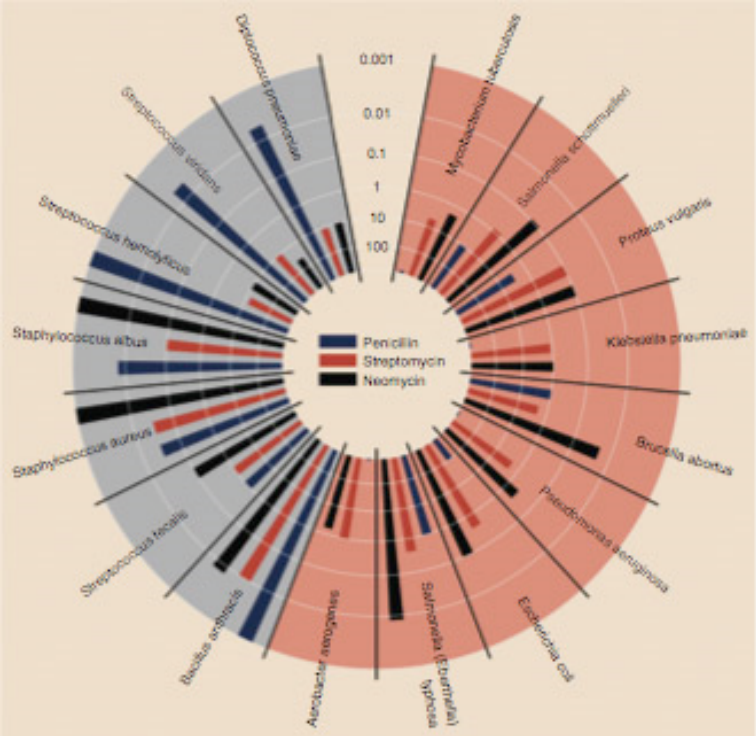
How do the drugs compare?



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?



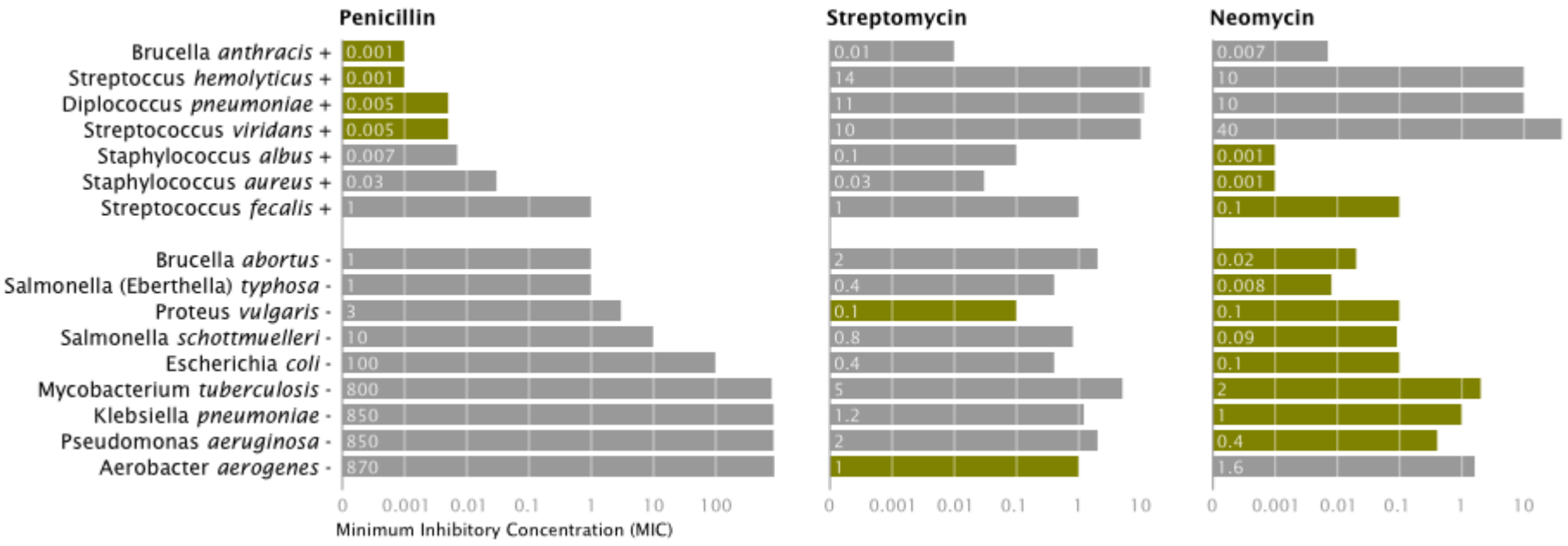
Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
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

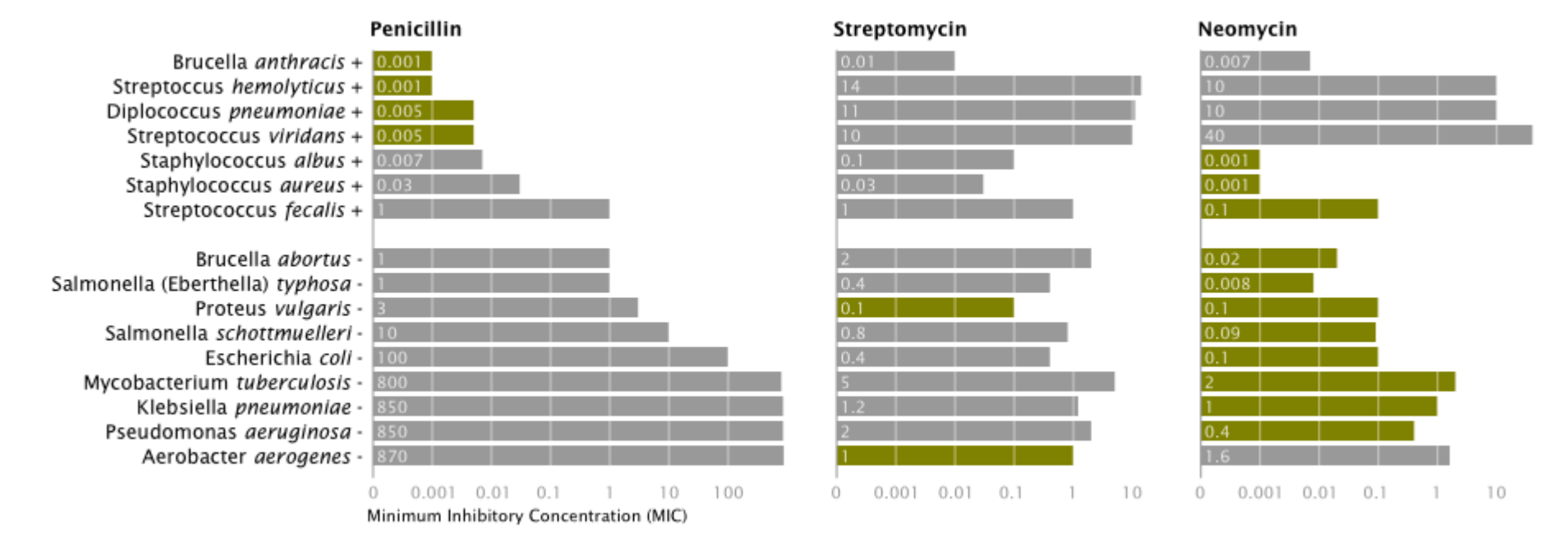
Background Color: Gram Staining

How do the drugs compare?

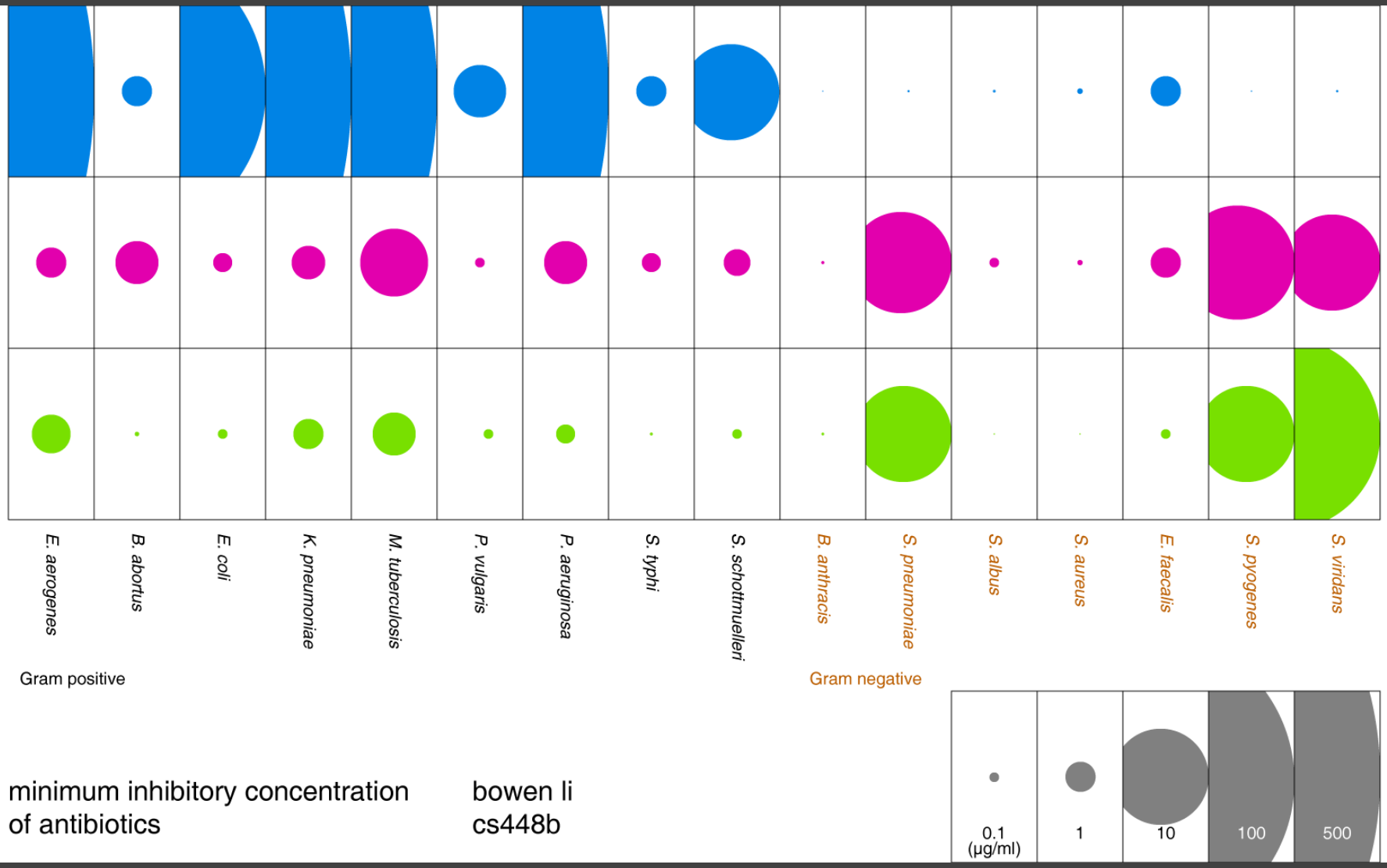


Mike Bostock
Stanford CS448B, Winter 2009

How do the drugs compare?



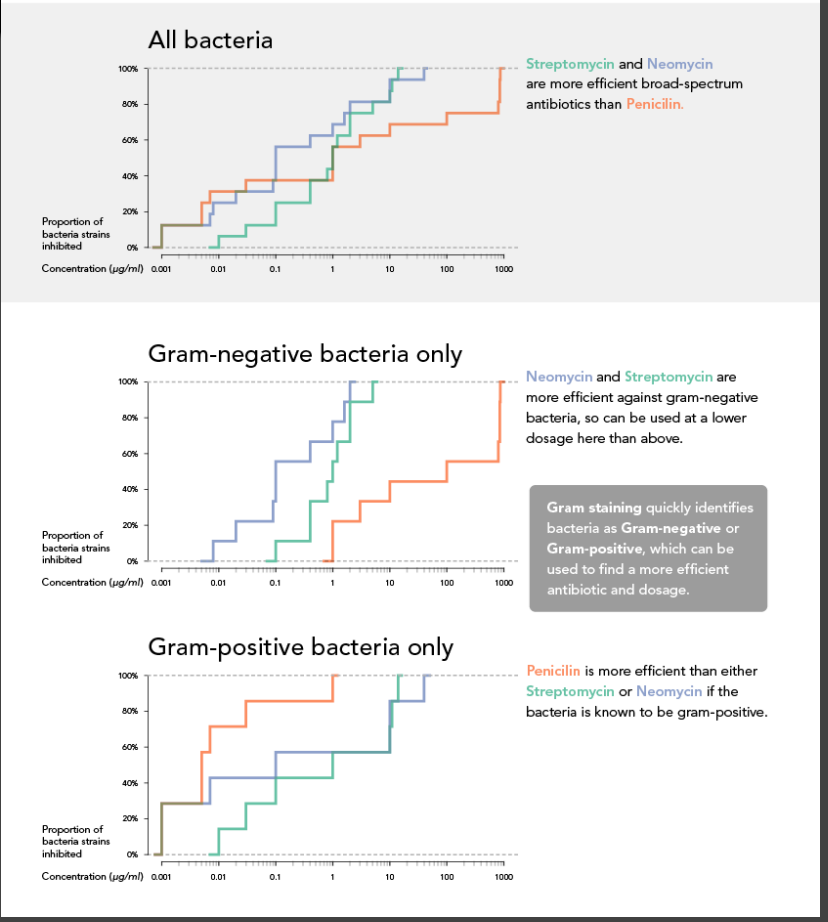
X-axis: Antibiotic | log(MIC)
Y-axis: Gram-Staining | Species
Color: Most-Effective?

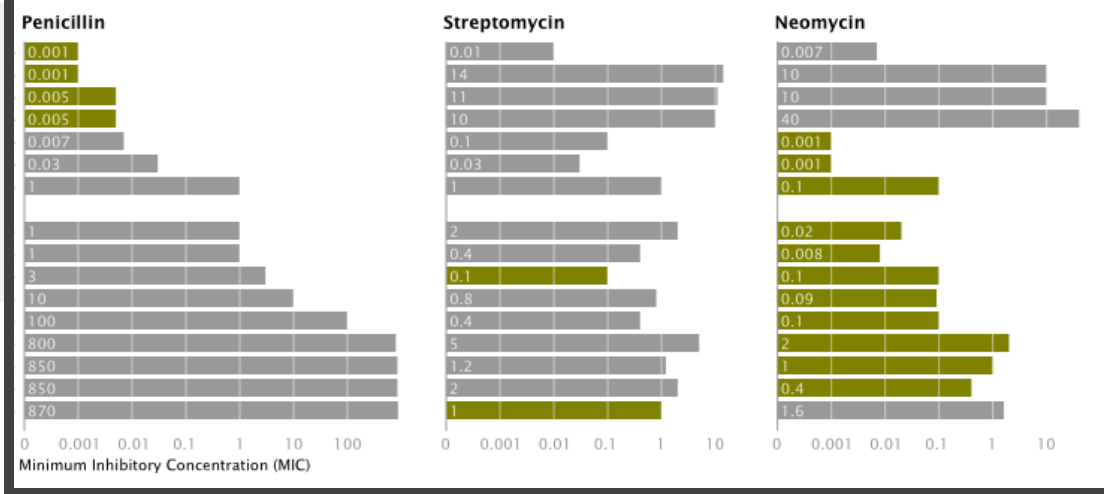


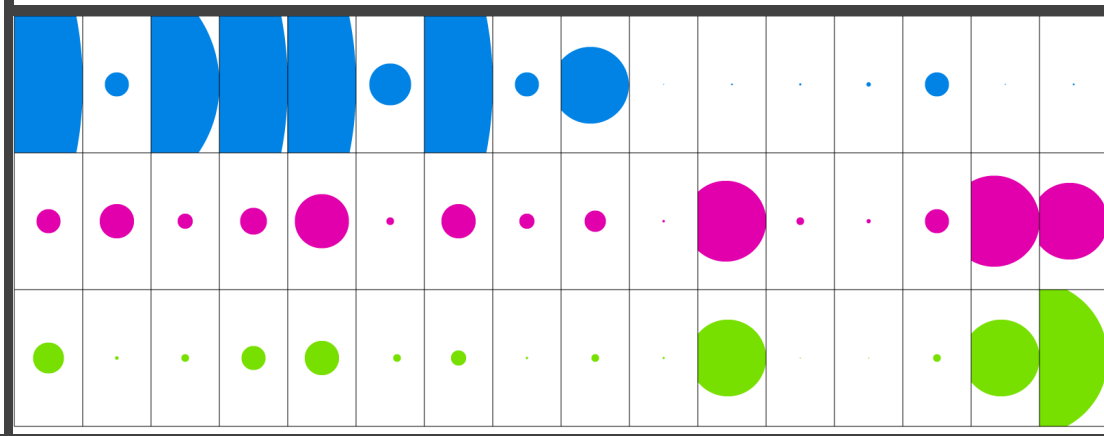
minimum inhibitory concentration of antibiotics

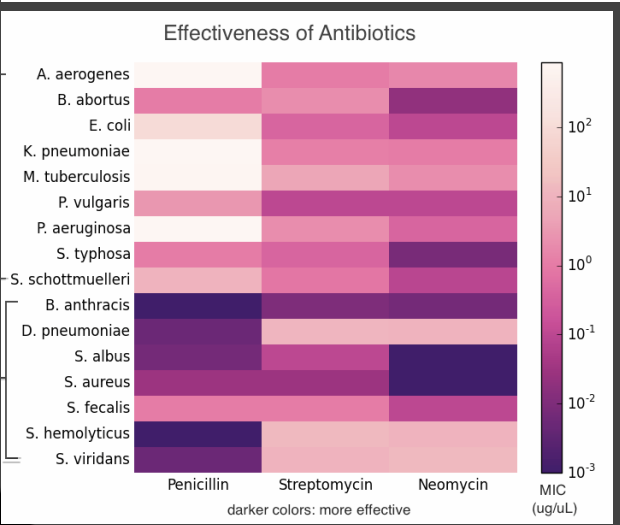
bowen li
cs448b

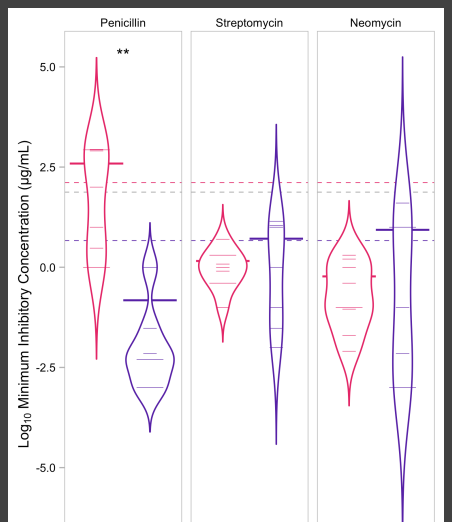
Bowen Li
Stanford CS448B, Fall 2009

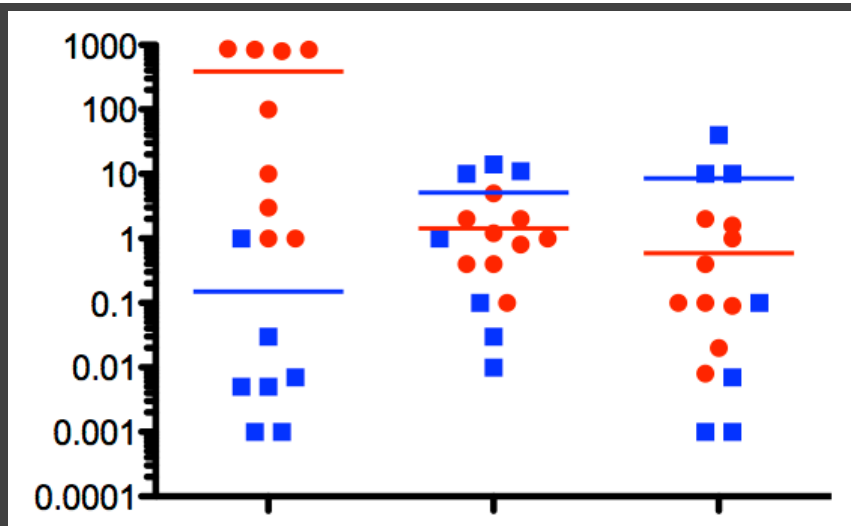

All bacteria
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.
Proportion of bacteria strains inhibited
Concentration (μg/ml)
Gram-negative bacteria only
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage 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.

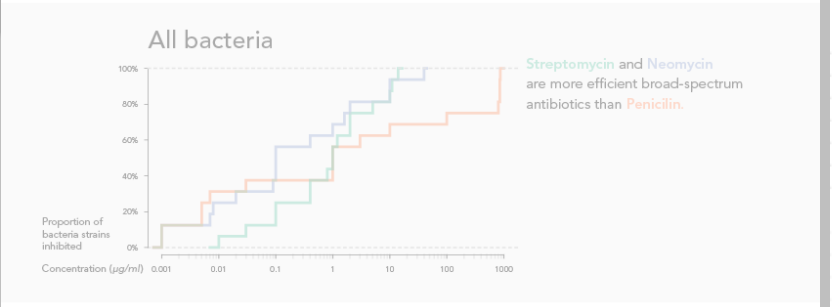

Penicillin
Streptomycin
Neomycin
Minimum Inhibitory Concentration (MIC)

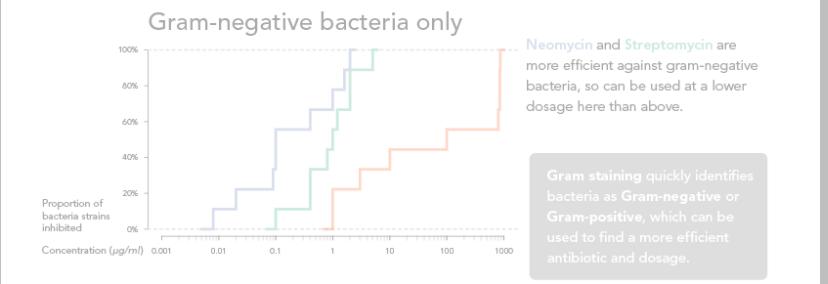


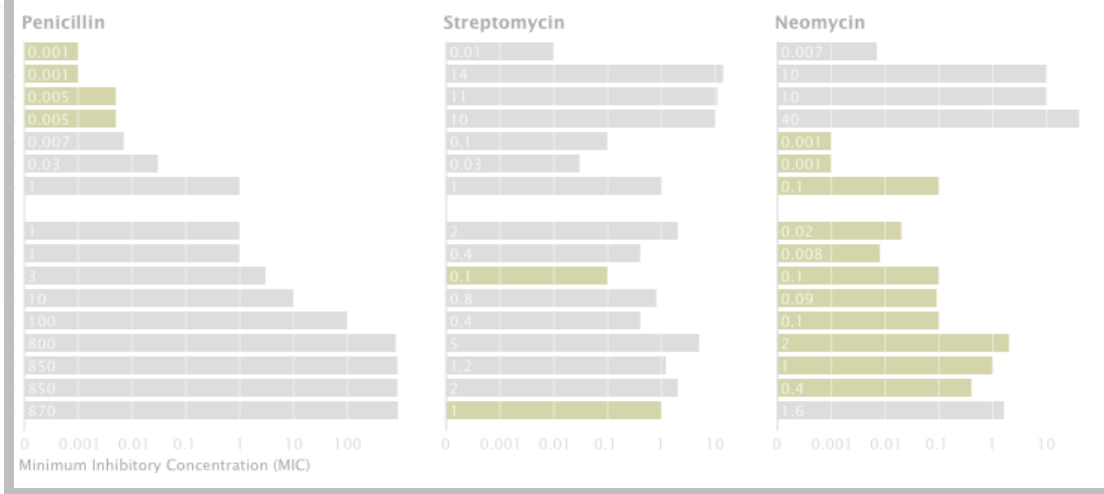

Effectiveness of Antibiotics
A. aerogenes
B. abortus
E. coli
K. pneumoniae
M. tuberculosis
P. vulgaris
P. aeruginosa
S. typhosa
S. schottmuelleri
B. anthracis
D. pneumoniae
S. albus
S. aureus
S. fecalis
S. hemolyticus
S. viridans
Penicillin
Streptomycin
Neomycin
darker colors: more effective
MIC (ug/uL)

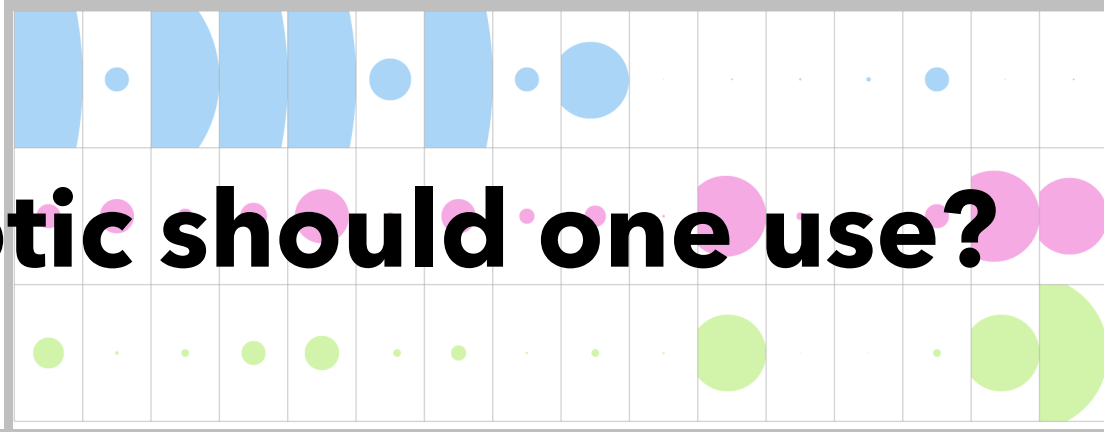

Penicillin
Streptomycin
Neomycin
**
Log10 Minimum Inhibitory Concentration (μg/mL)


1000
100
10
1
0.1
0.01
0.001
0.0001

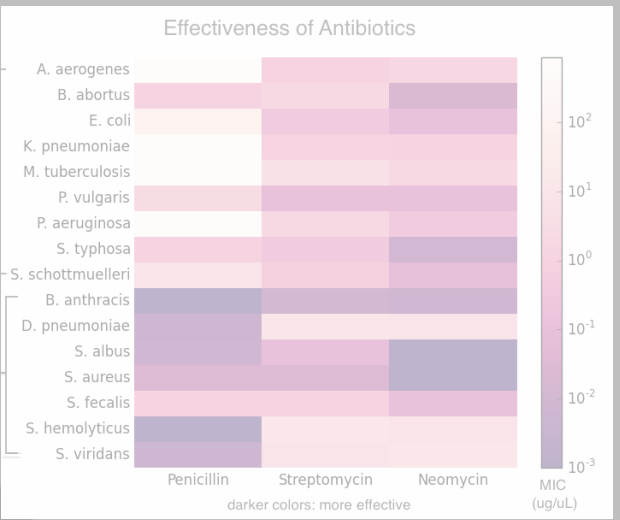


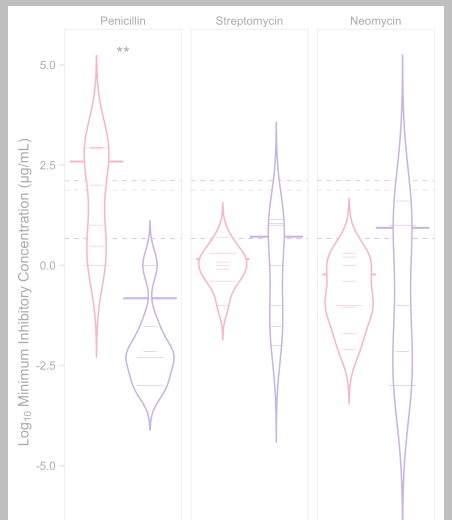


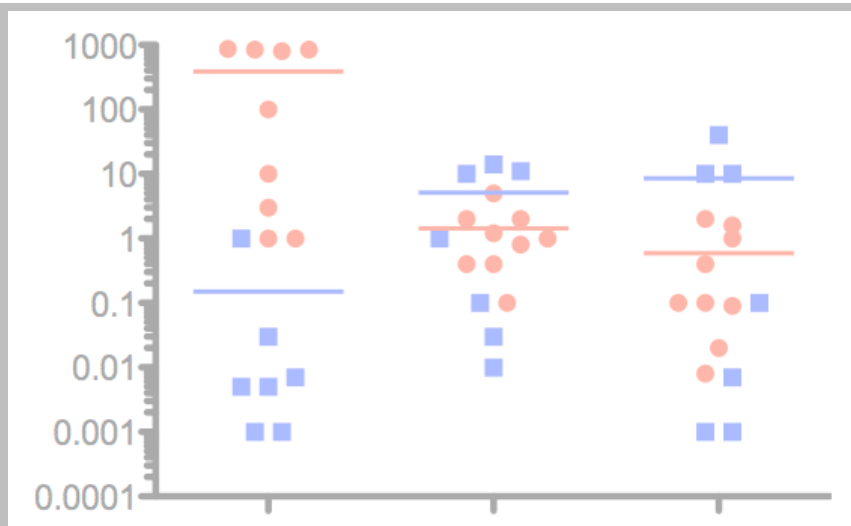




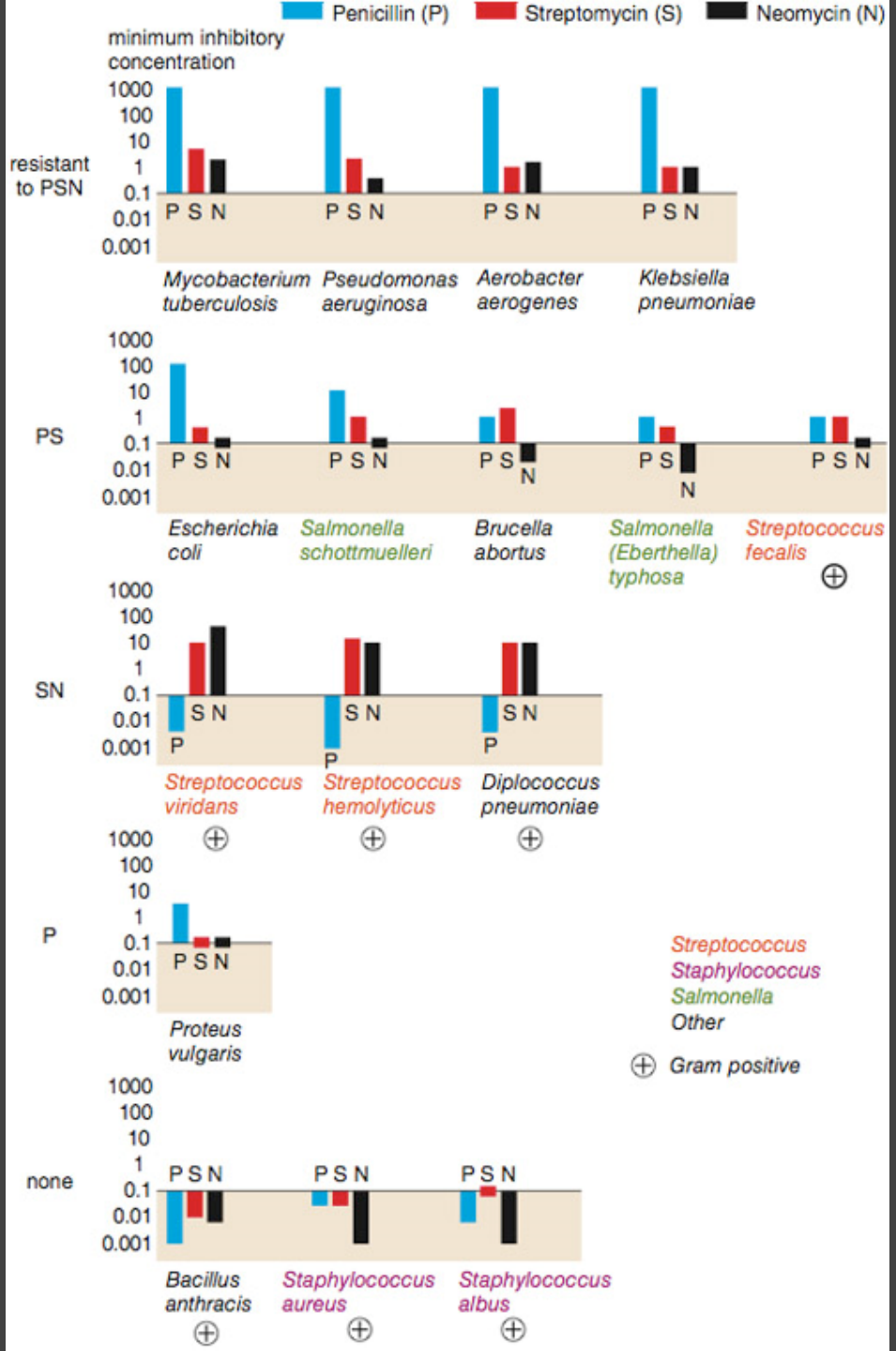
Which antibiotic should one use?







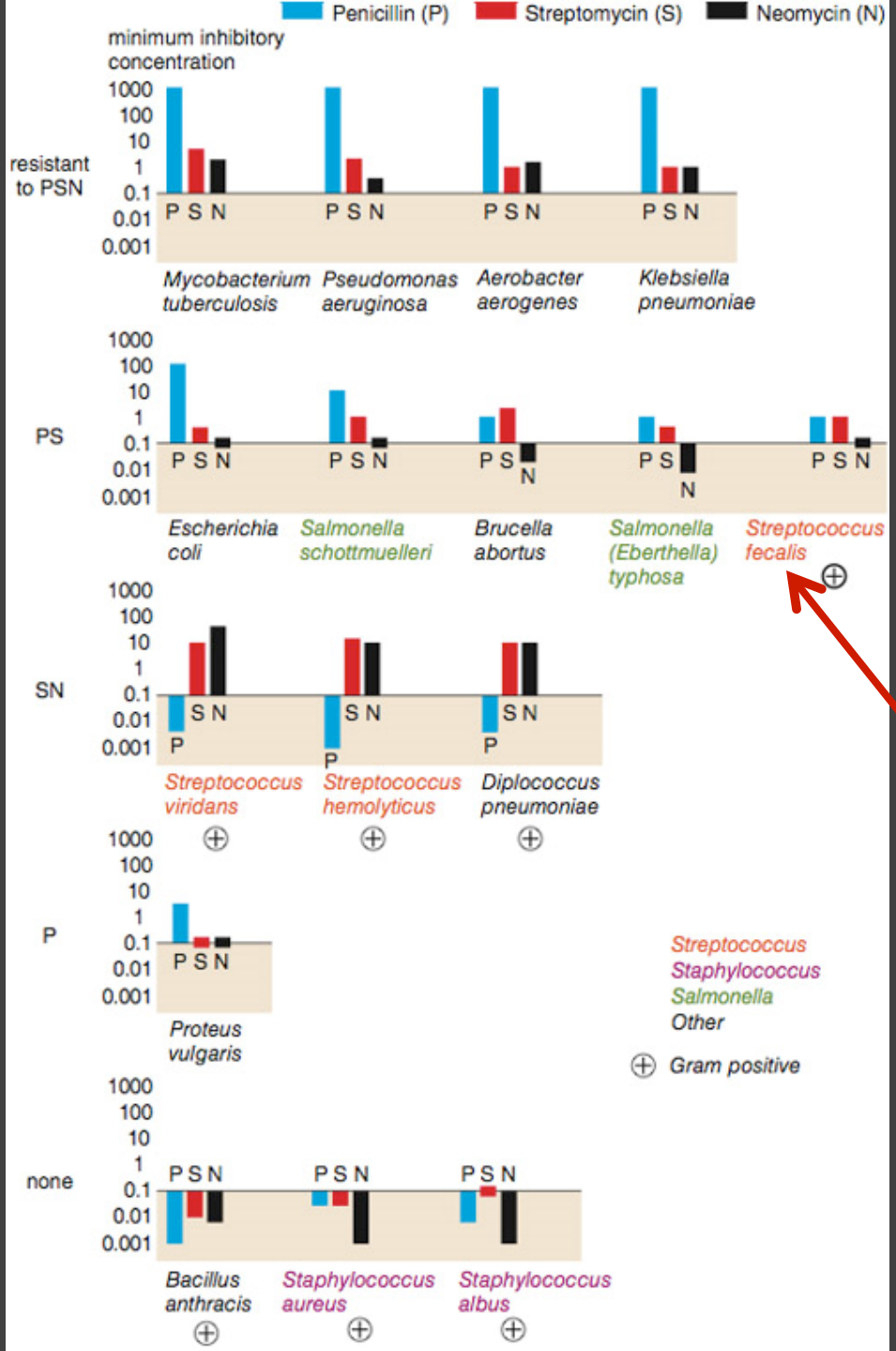
Do the bacteria group by antibiotic resistance?



Do the bacteria group by antibiotic resistance?

Wainer & Lysen

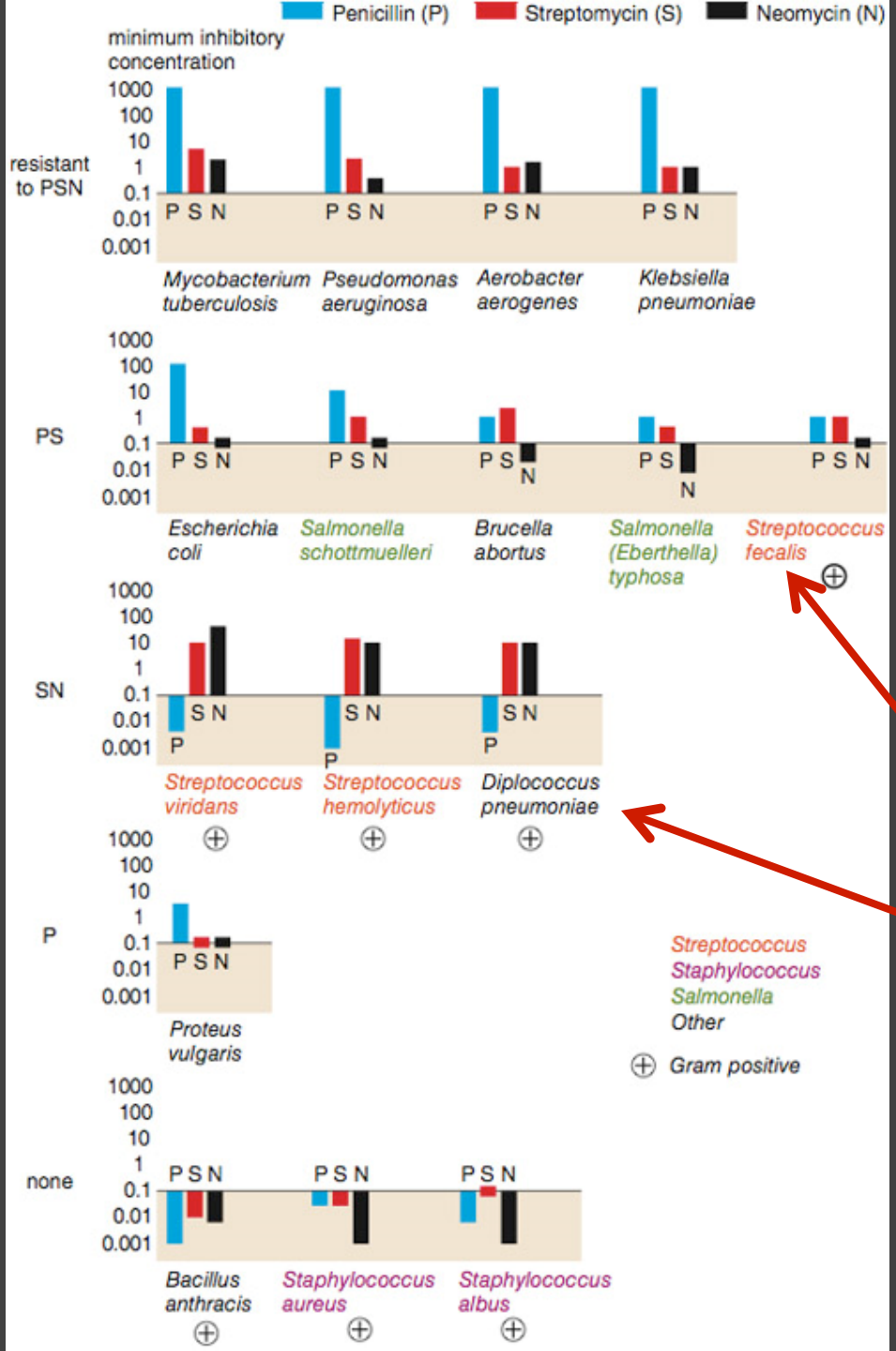
American Scientist, 2009



Do the bacteria group by antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

Wainer & Lysen
American Scientist, 2009



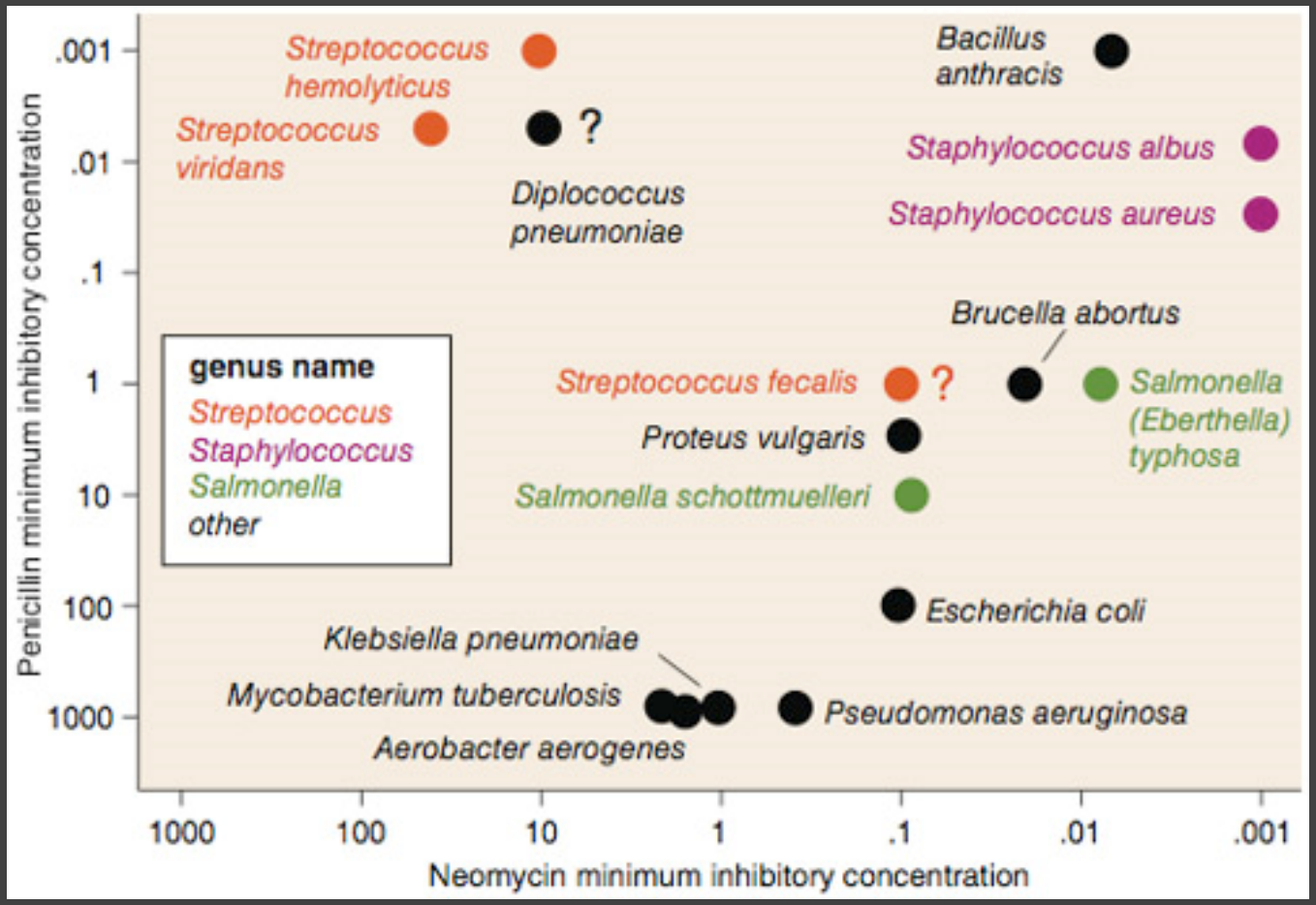
Do the bacteria group by antibiotic resistance?

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

Updates

A1 due tonight. We'll be discussing the submission in class on Thursday

A2 will be released on Thursday. This will require you to conduct an exploratory analysis and present your findings visually.

Lectures recorded starting with this one \o/

Tableau / Polaris

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

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

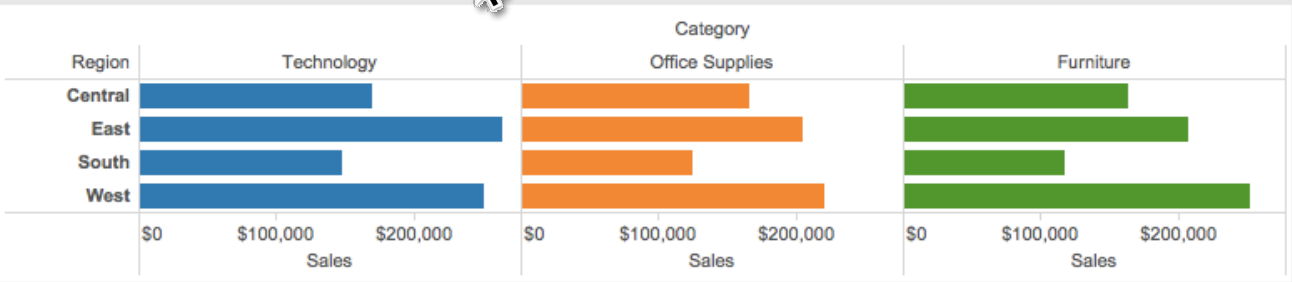
Category

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

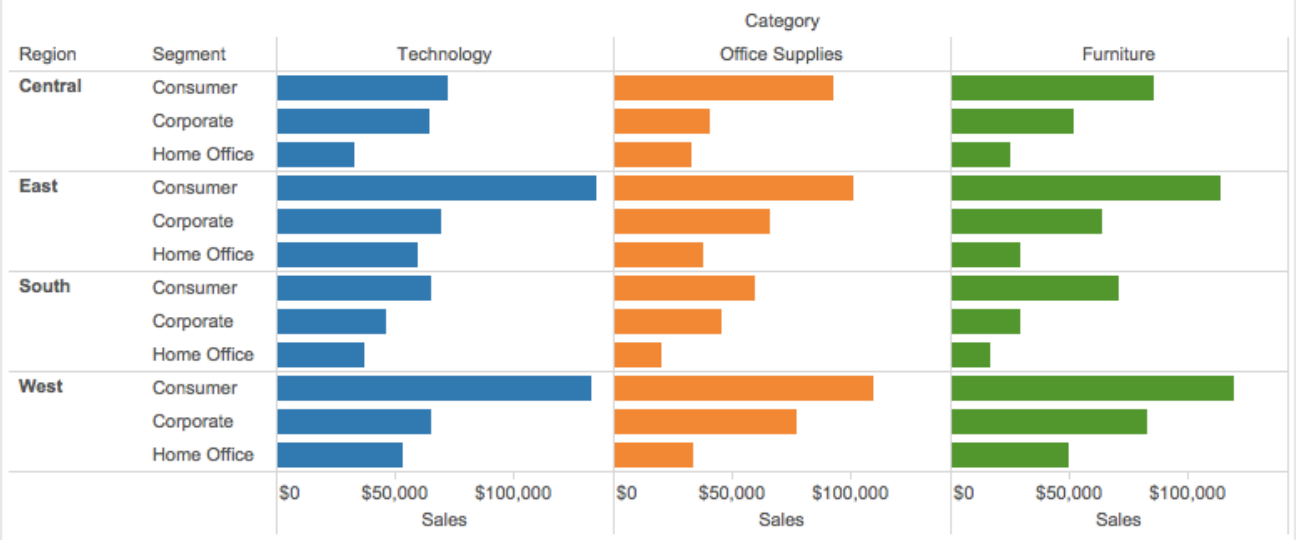
Filters

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

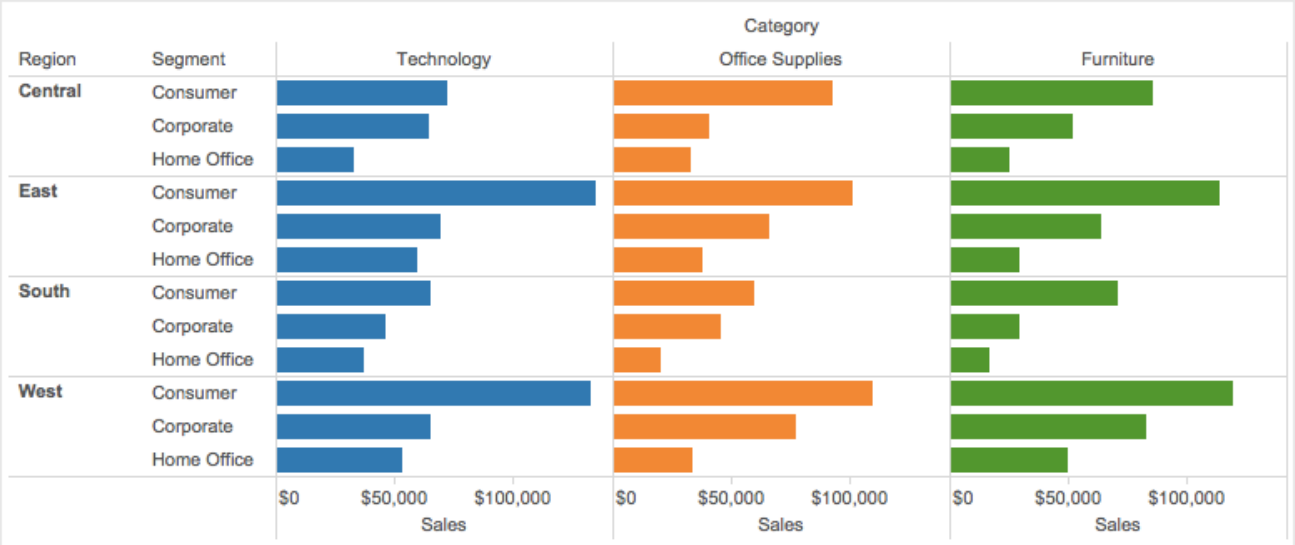
Category

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)

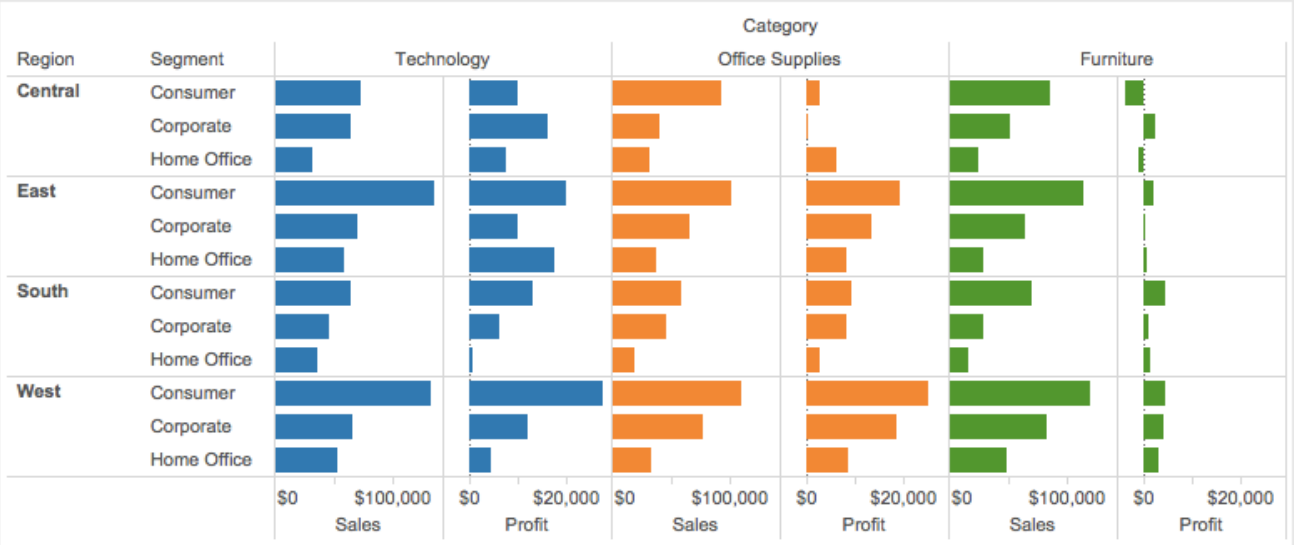
SUM(Profit)

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

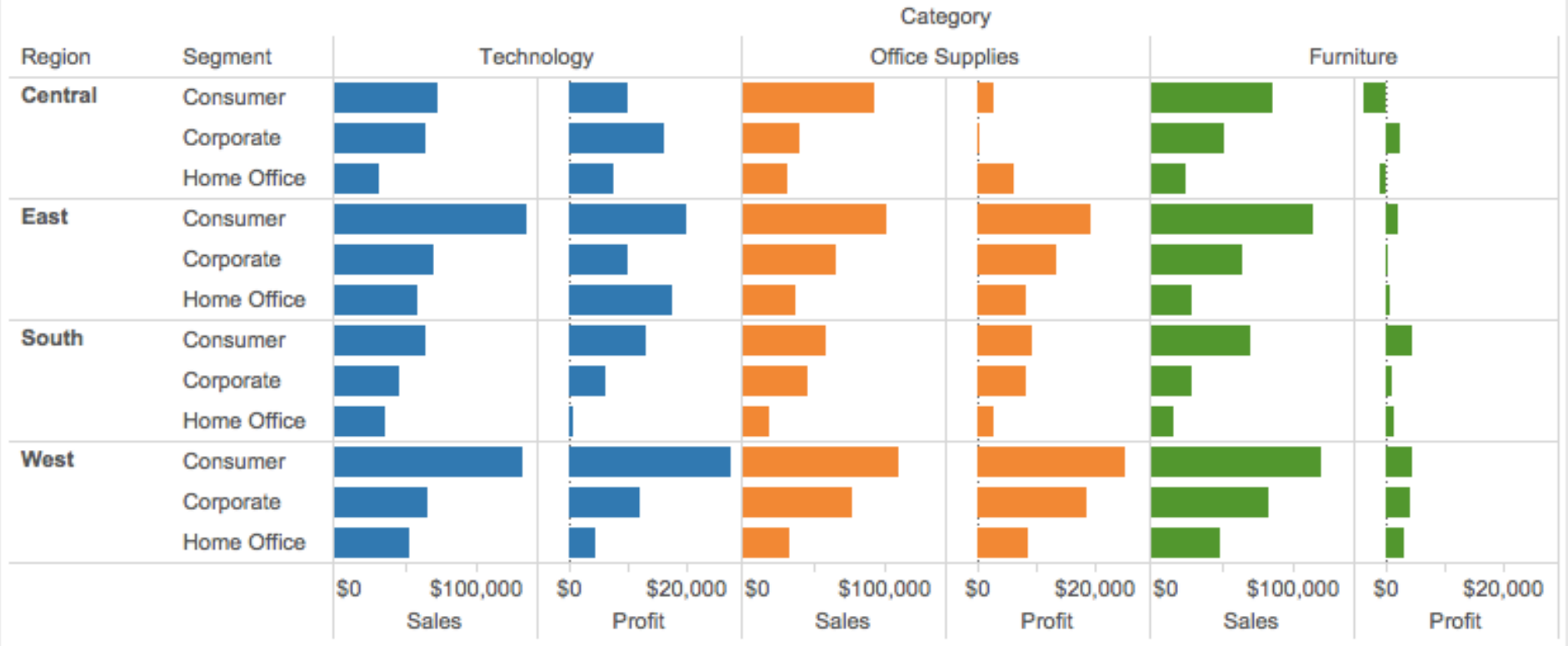
Normal

Show Me

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

Rows: Region, Segment

GROUP BY Category, Region, Segment



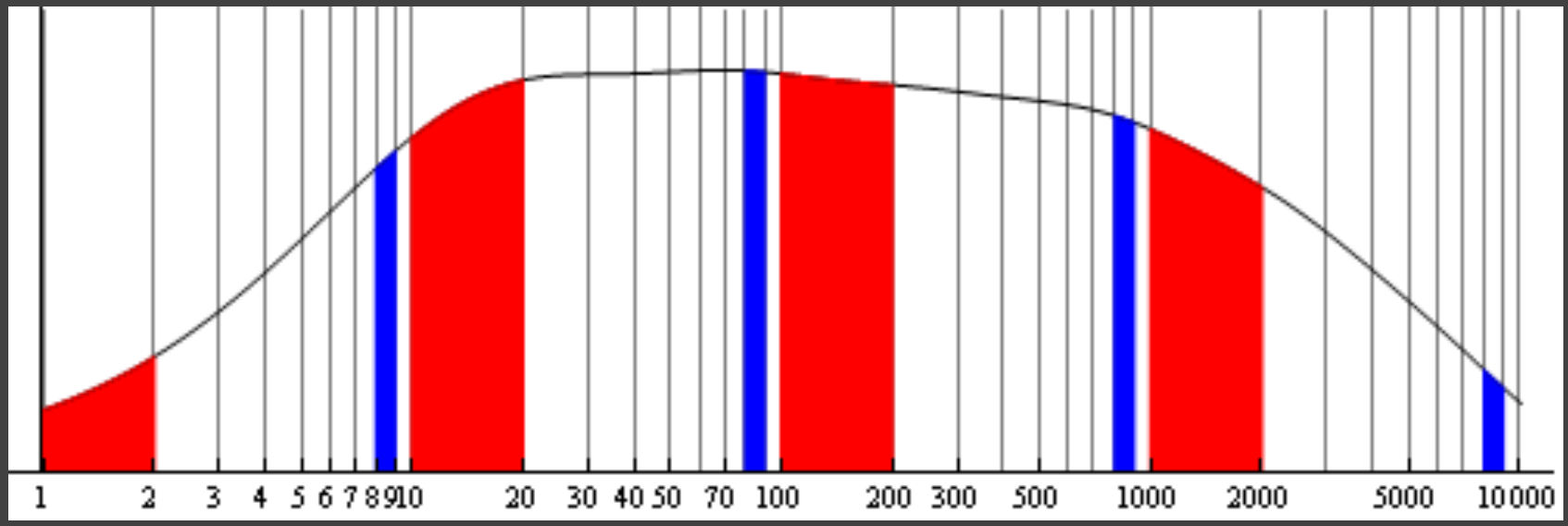
A Detective Story

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

Firm A		***Firm B***	**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!