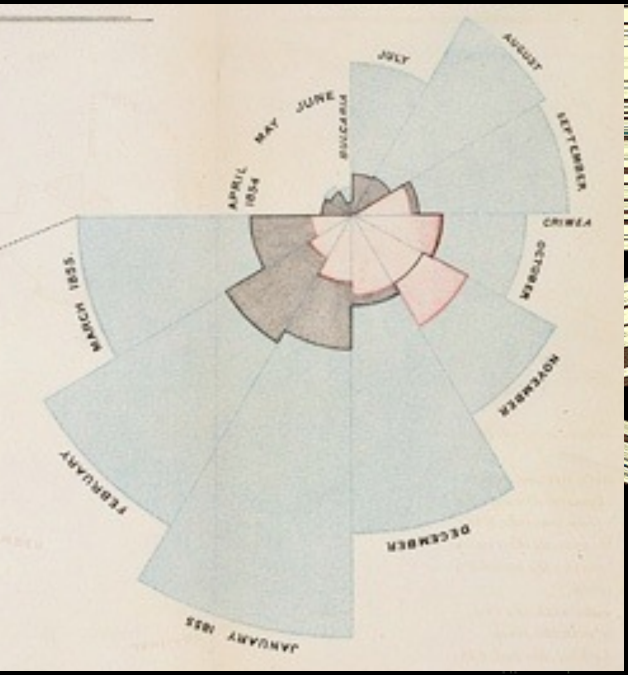
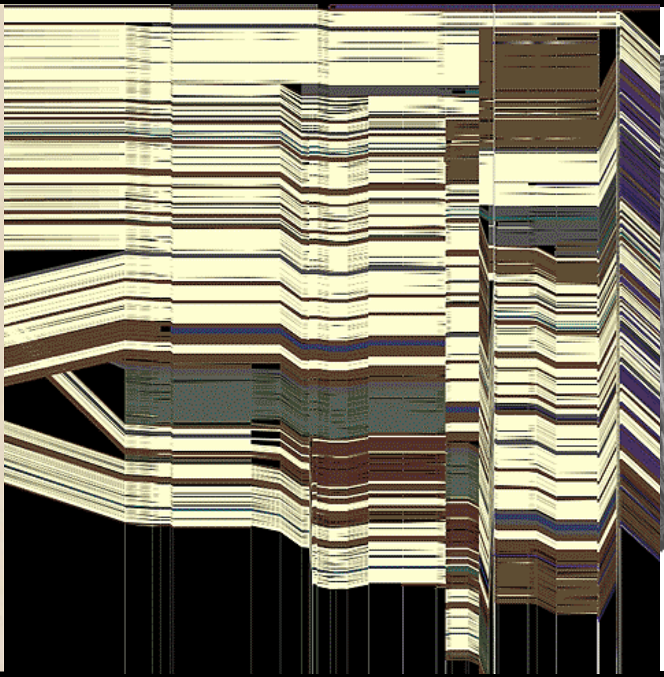
CSE 512 - Data Visualization

Exploratory Data Analysis





Leilani Battle University of Washington

Learning Goals

What is exploratory data analysis and why is it important?

What factors should we consider when exploring a dataset?

How do visualization researchers design tools to support exploratory data analysis? (one example)

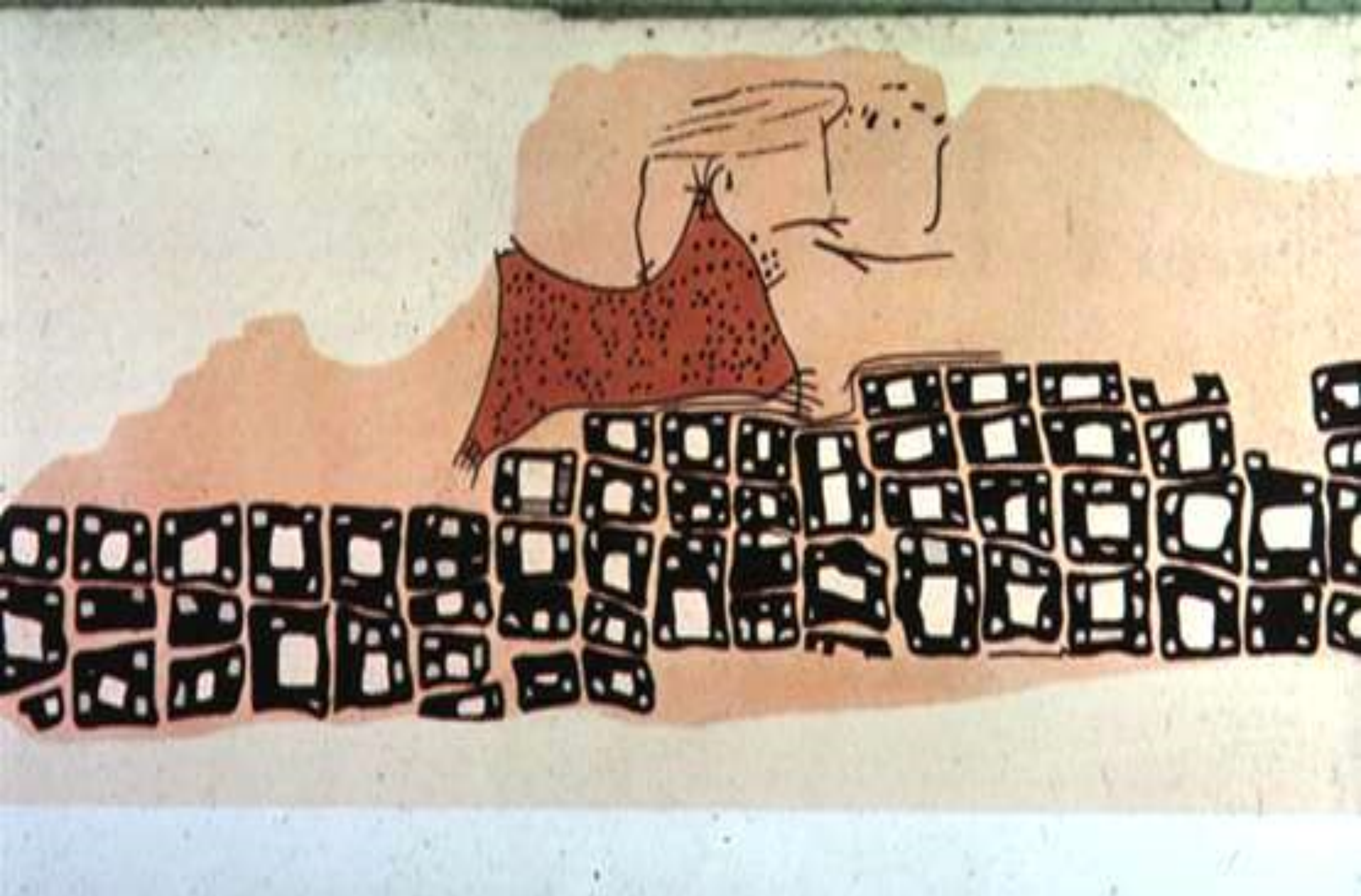
Topics

- Exploratory Data Analysis
  - Historical Context
  - Visualizations vs Statistical Models
- Data Wrangling
- Exploratory Analysis Examples

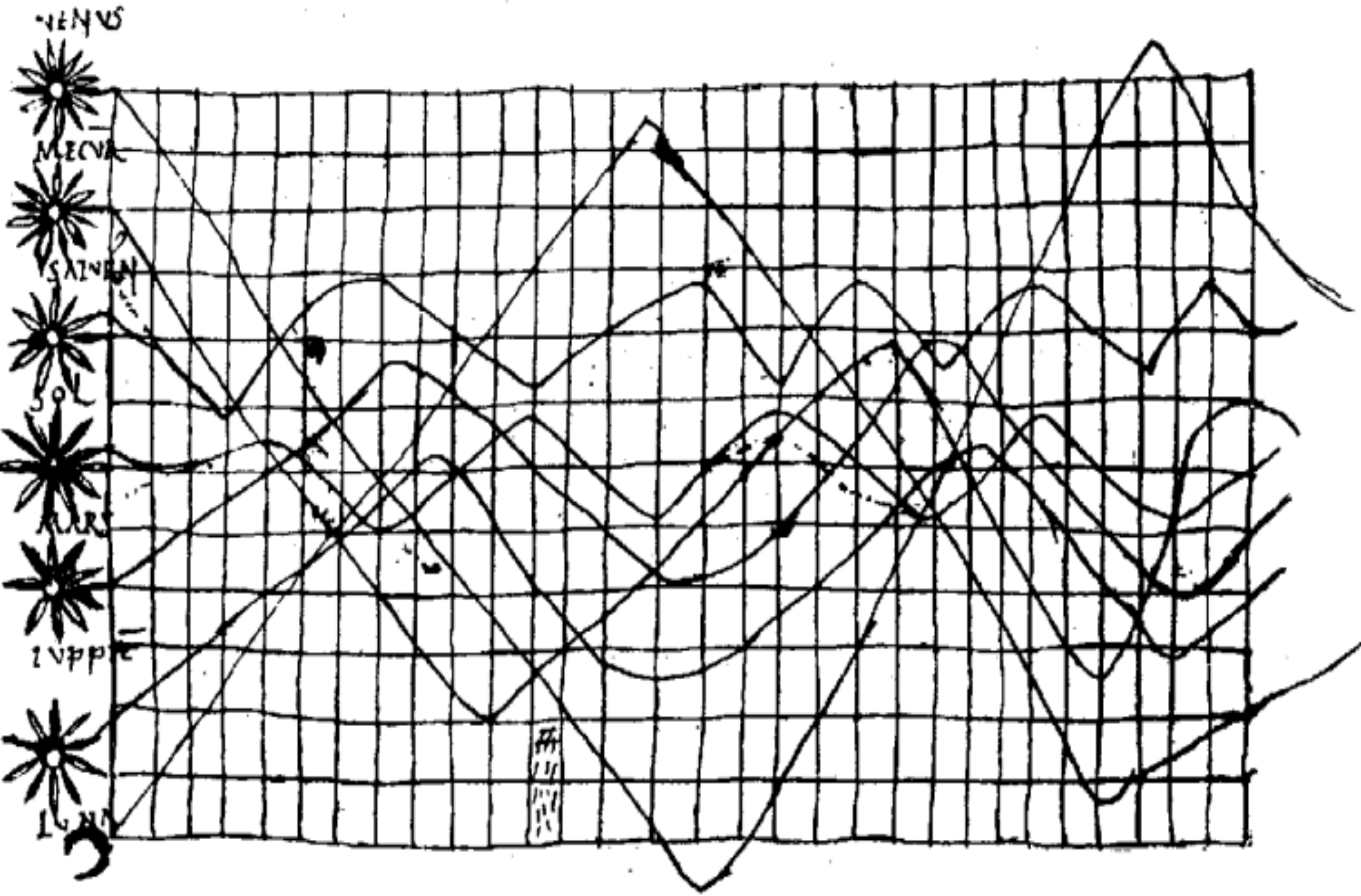
- Tableau / Polaris

What was the **first** data visualization?

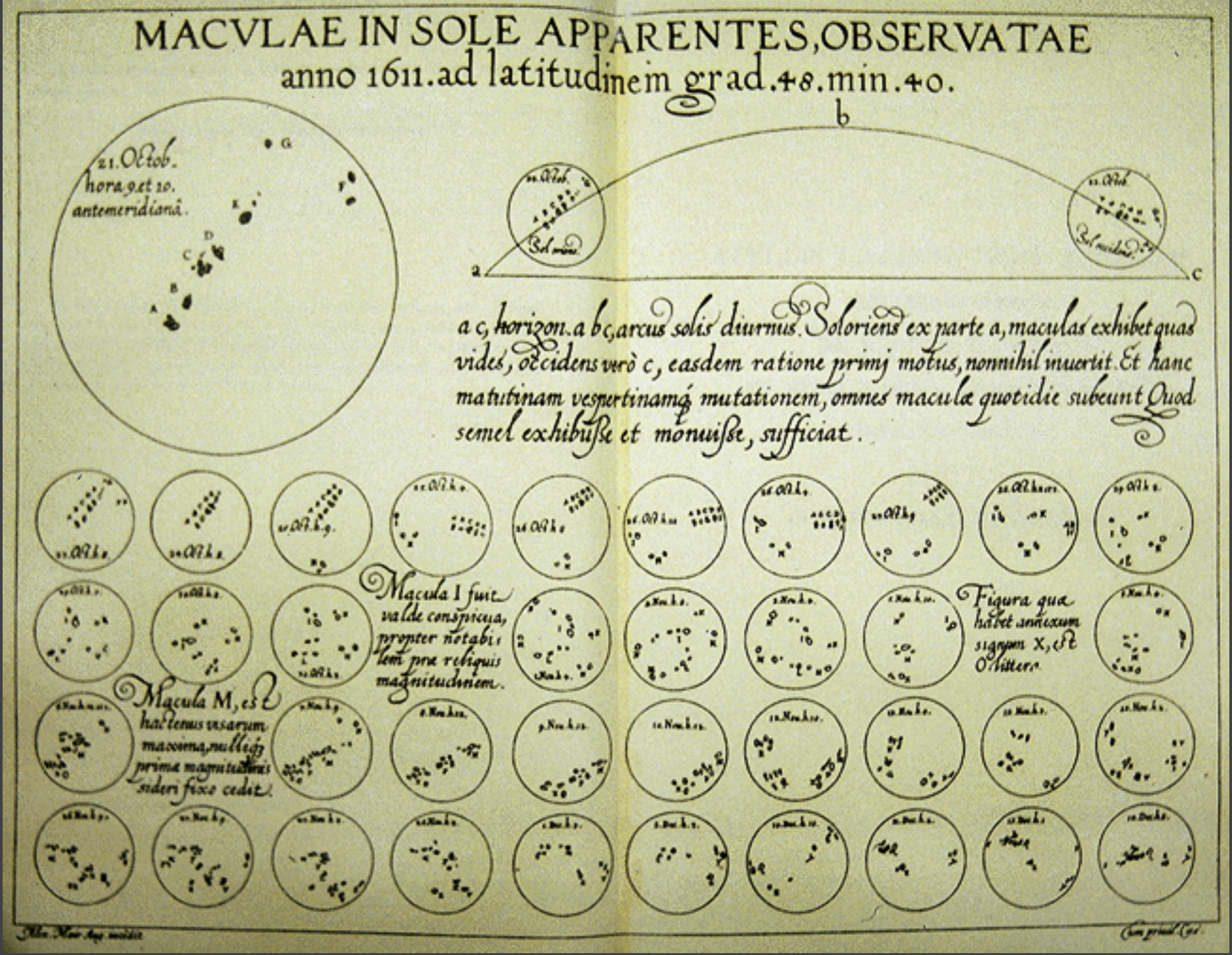
0 BC



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



~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 verò c, easdem ratione primj motus, nonnihil inuertit. Et hanc matutinam vespertinamq; mutationem, omnes maculæ quotidie subeunt. Quod semel exhibuiſse et monuiſse, sufficiat.

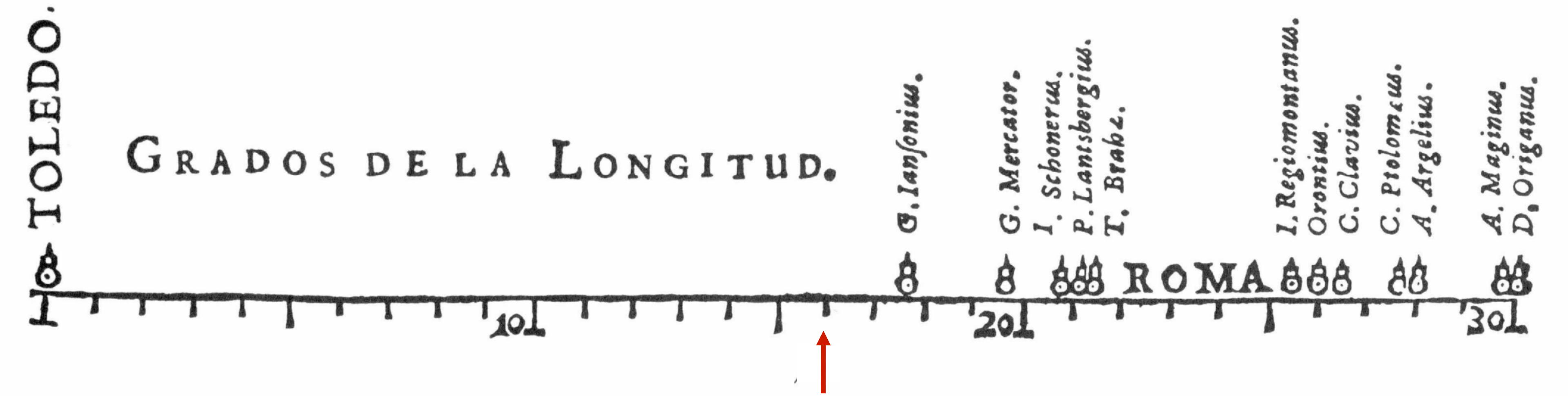
Macula I fuit valde conspicua, propter notabilem præ reliquis magnitudinem.

Macula M, est hactenus visarum maxima, nullique primæ magnitudinis sideri fixo cedit.

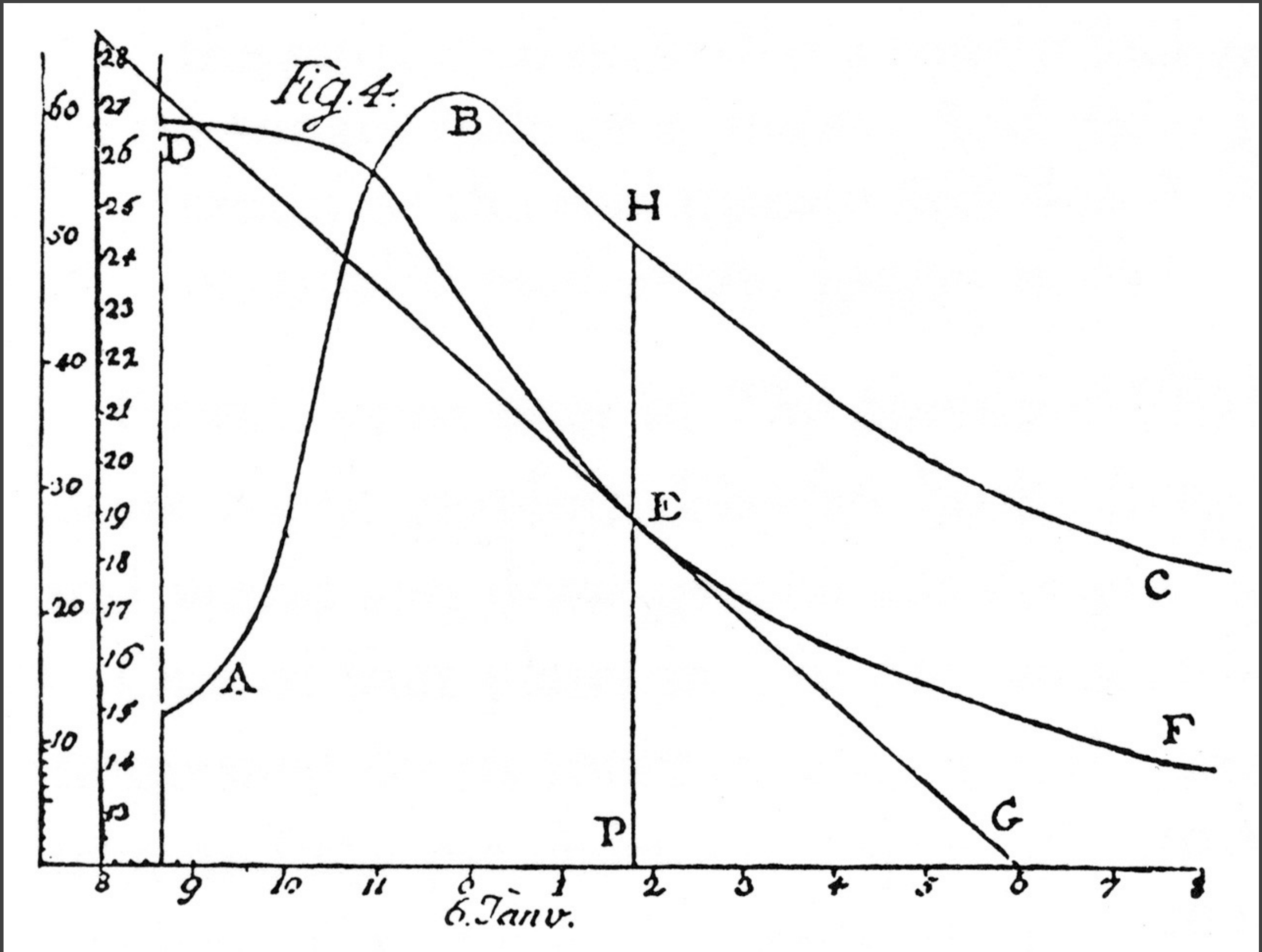
Figura quæ habet annexum signum X, est O litteræ.

0 BC

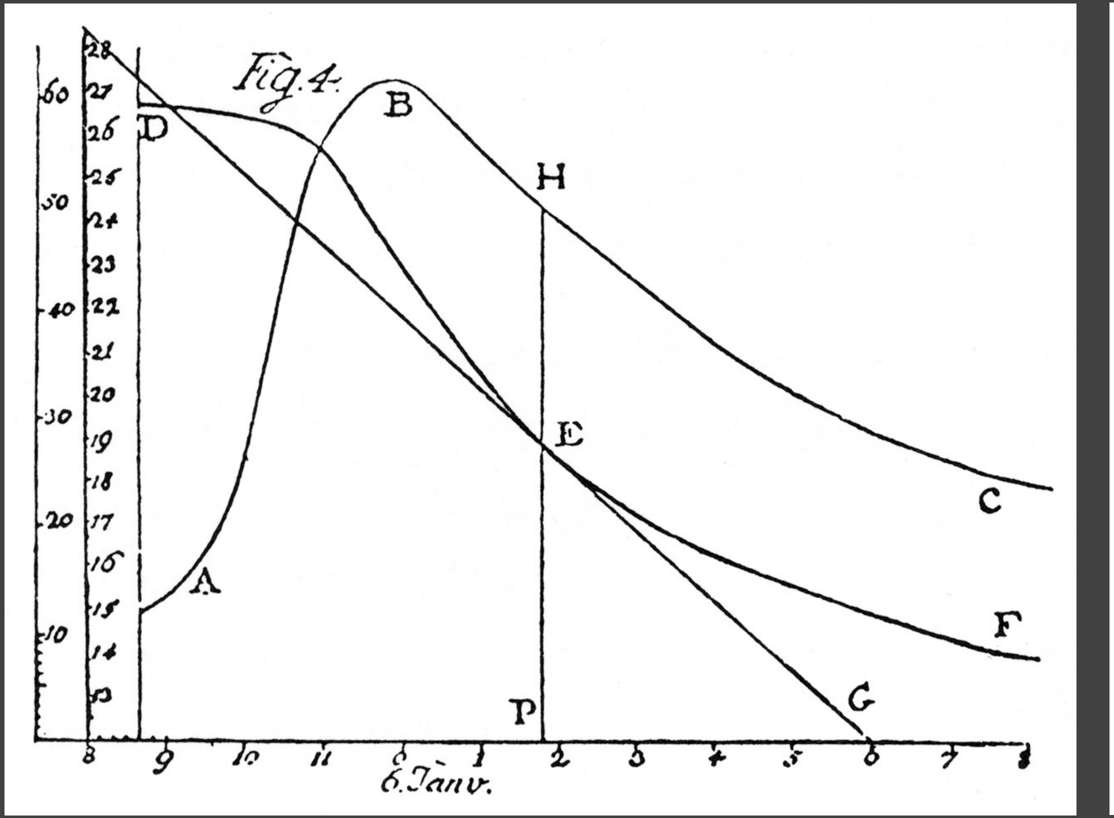
Sunspots over time, Scheiner 1626

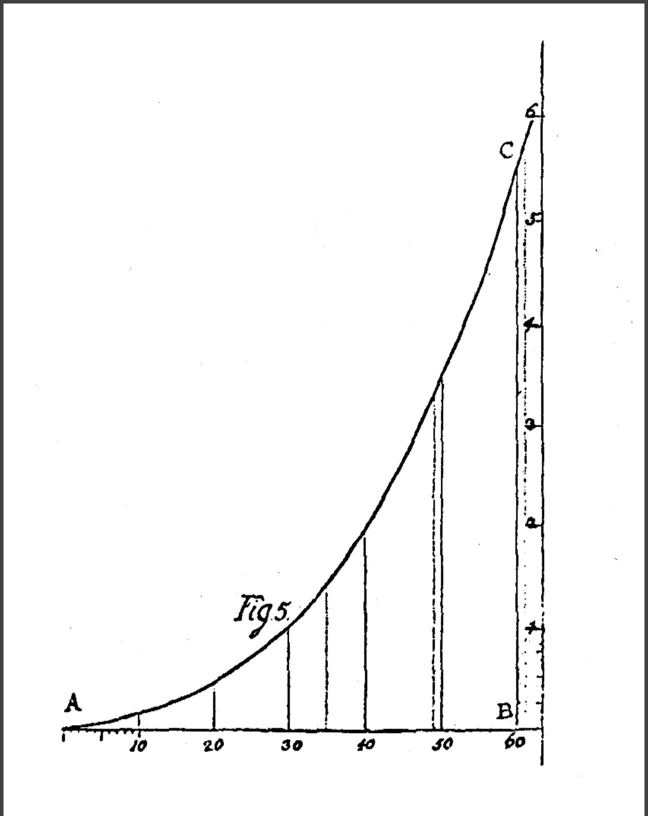


Longitudinal distance between Toledo and Rome, van Langren
1644



The Rate of Water Evaporation, Lambert 1765

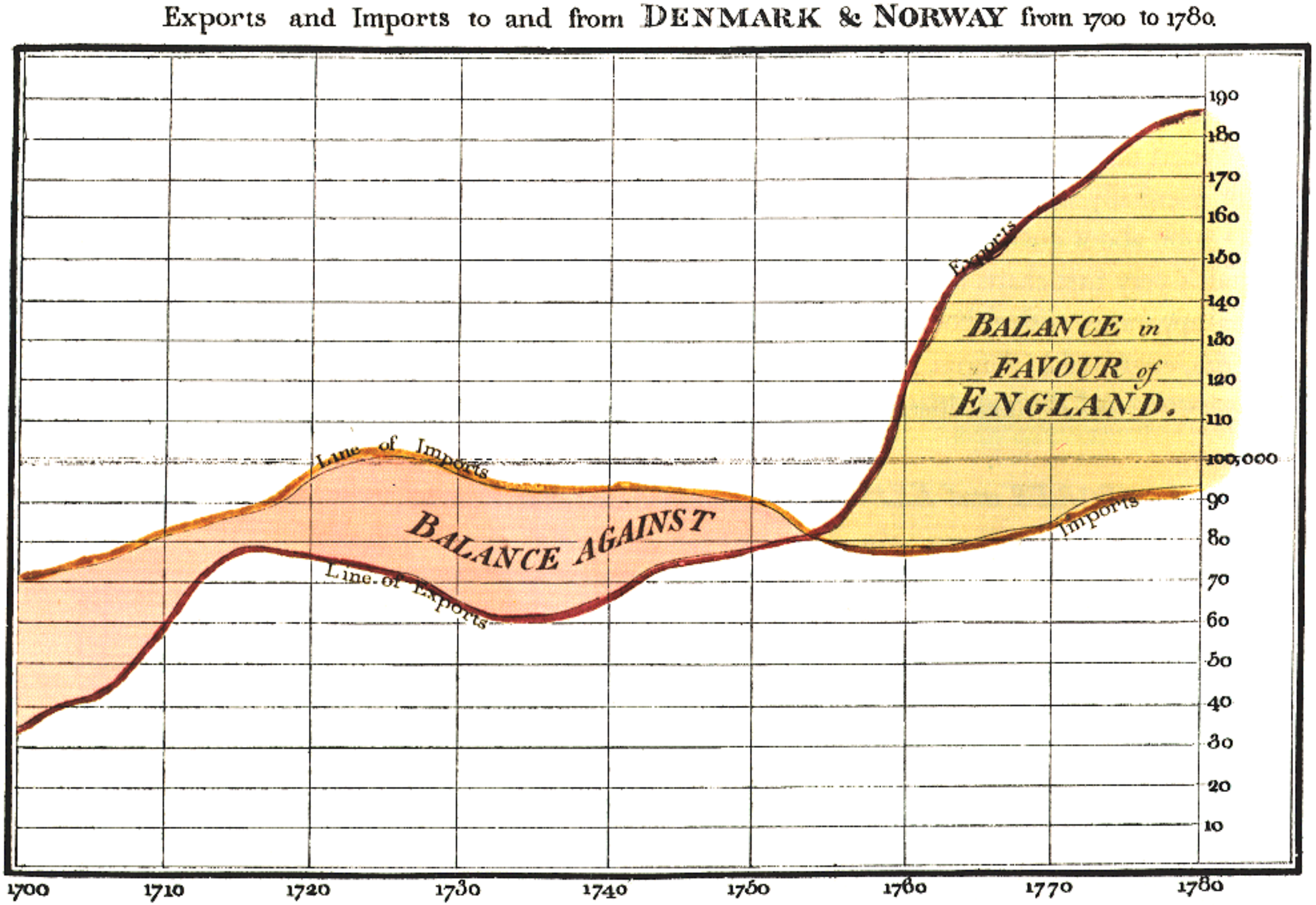




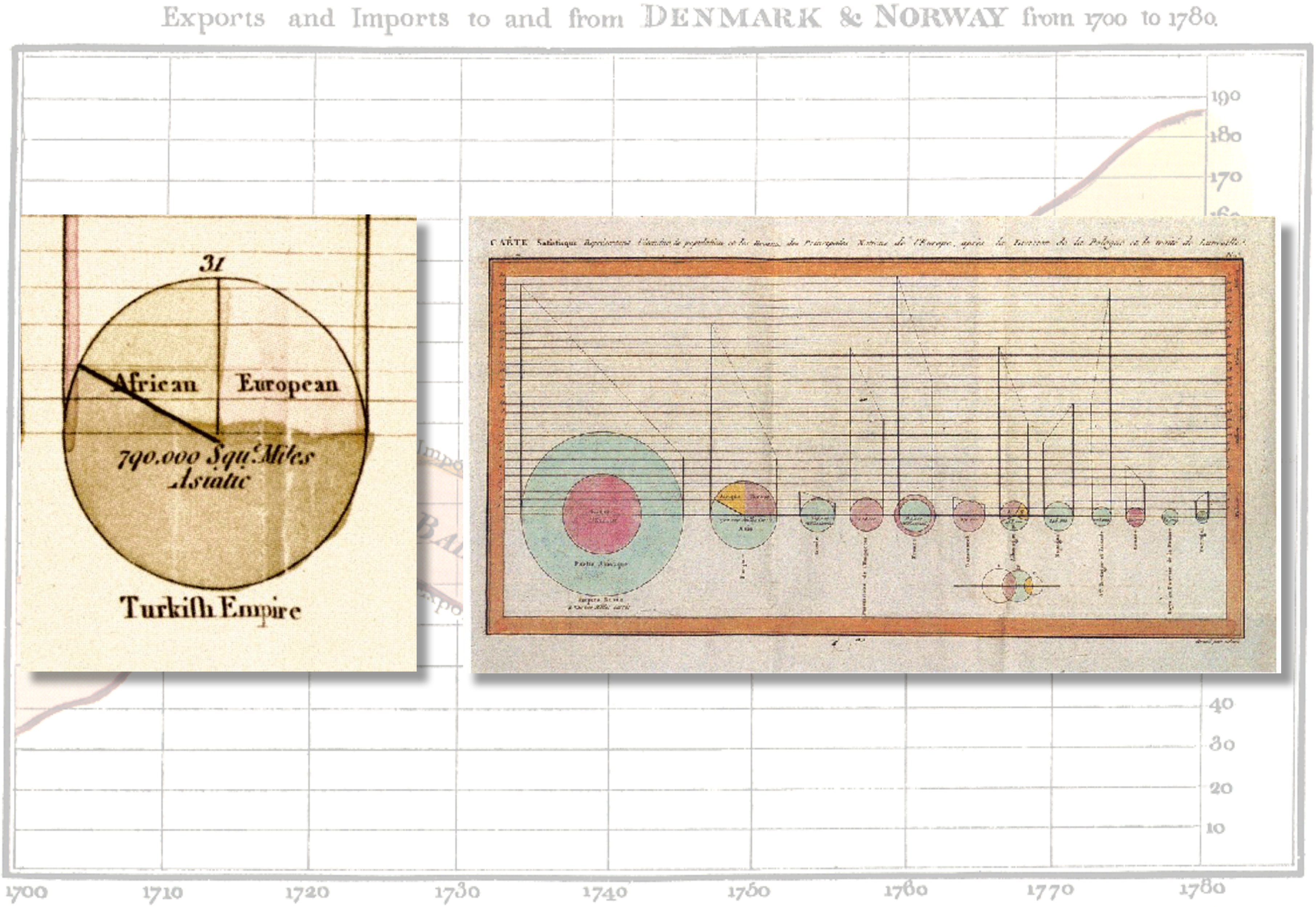
The Rate of Water Evaporation, Lambert 1765

The **Golden Age** of Data Visualization

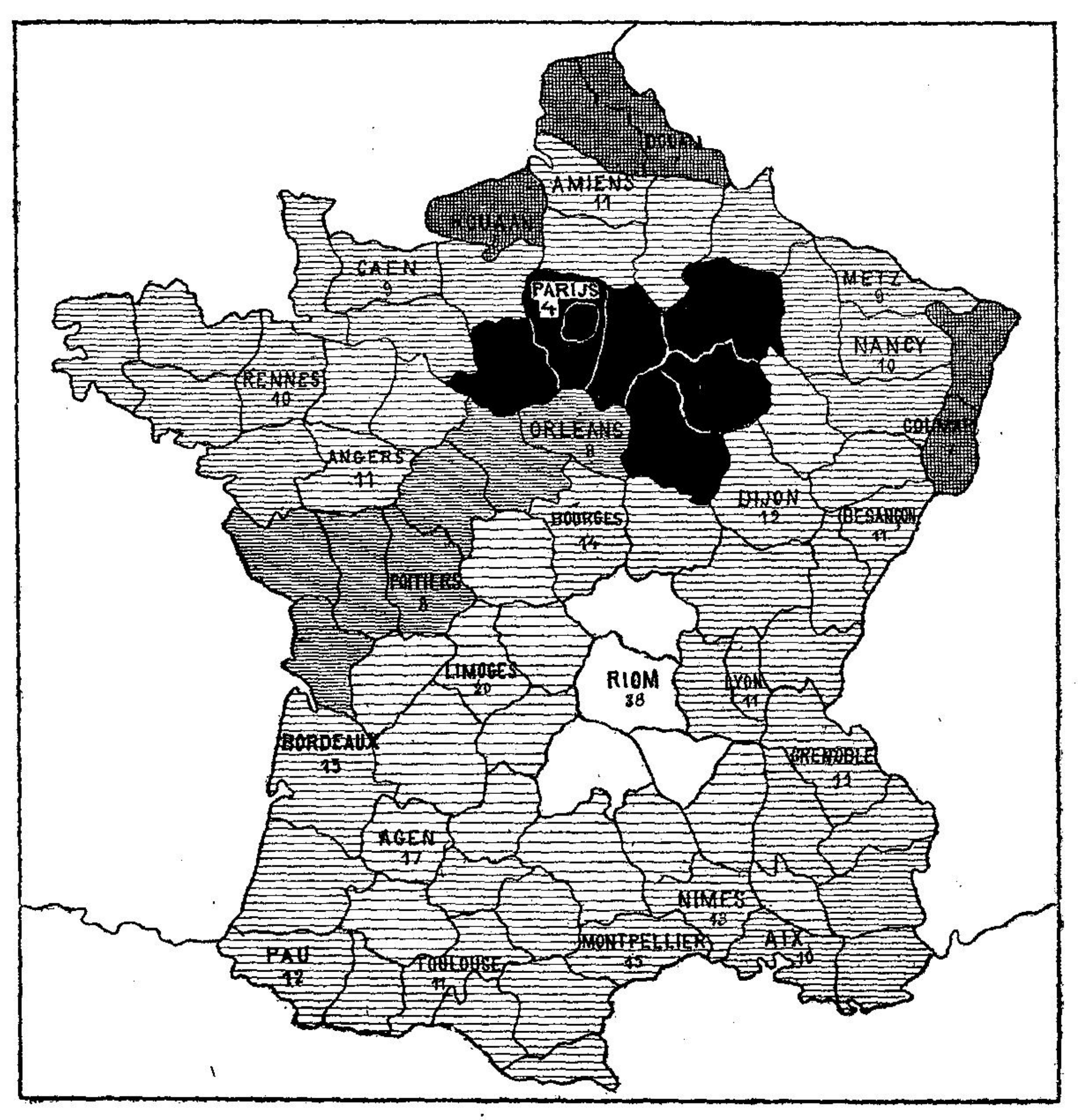
1786 1900



The Commercial and Political Atlas, William Playfair 1786

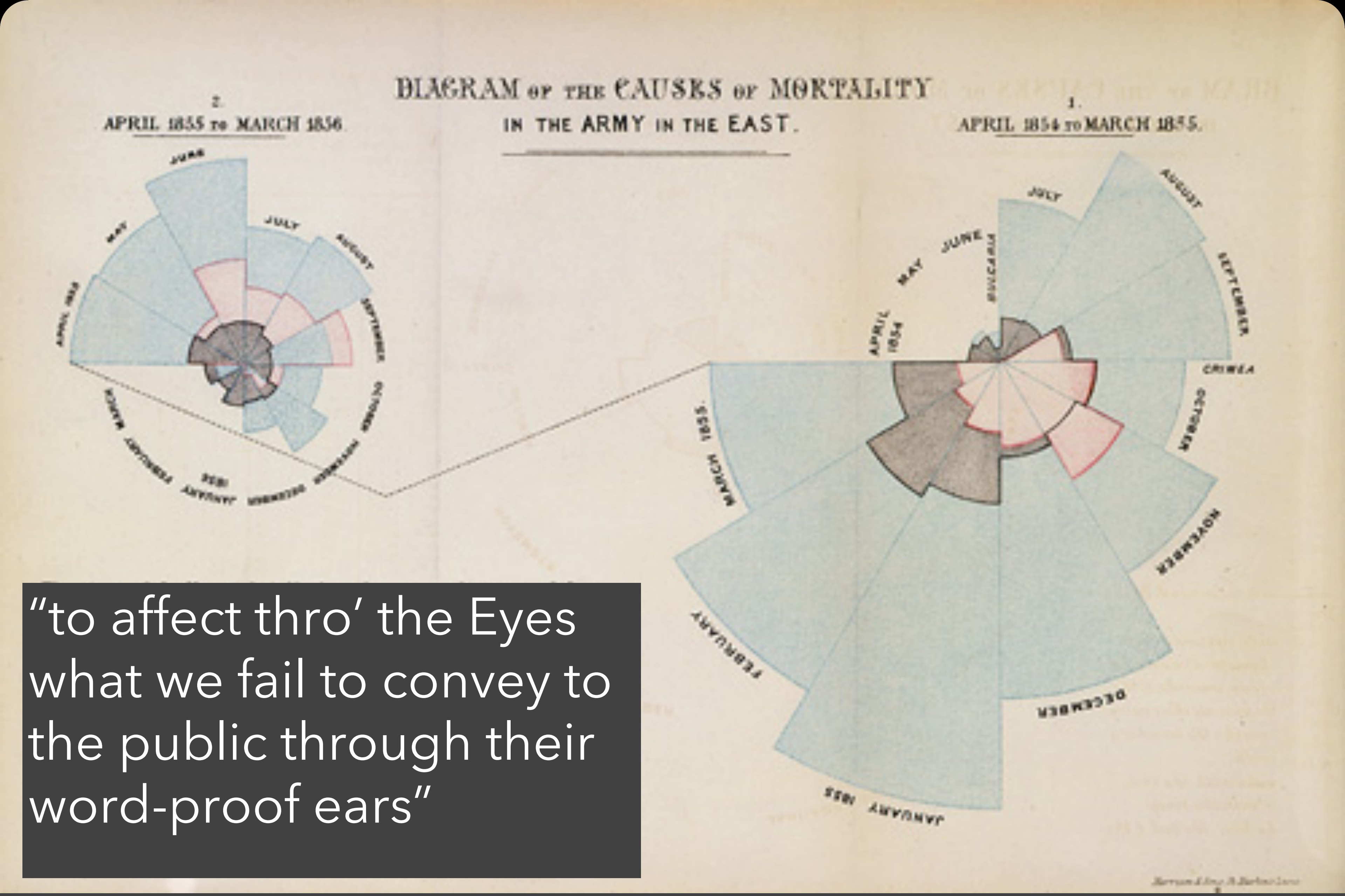


Statistical Breviary, William Playfair 1801



1786

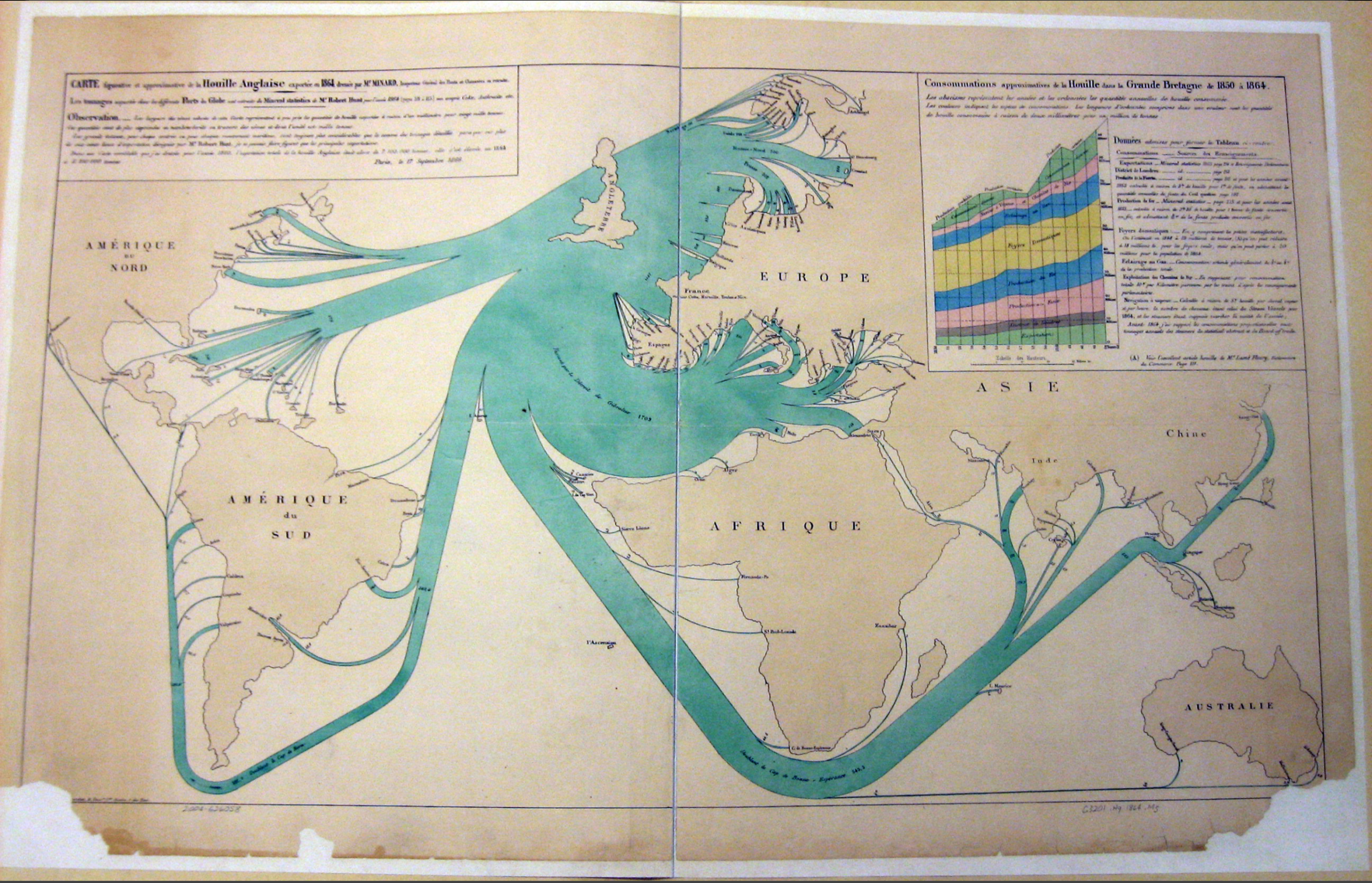
1826(?) Illiteracy in France, Pierre Charles Dupin



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

1786

1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

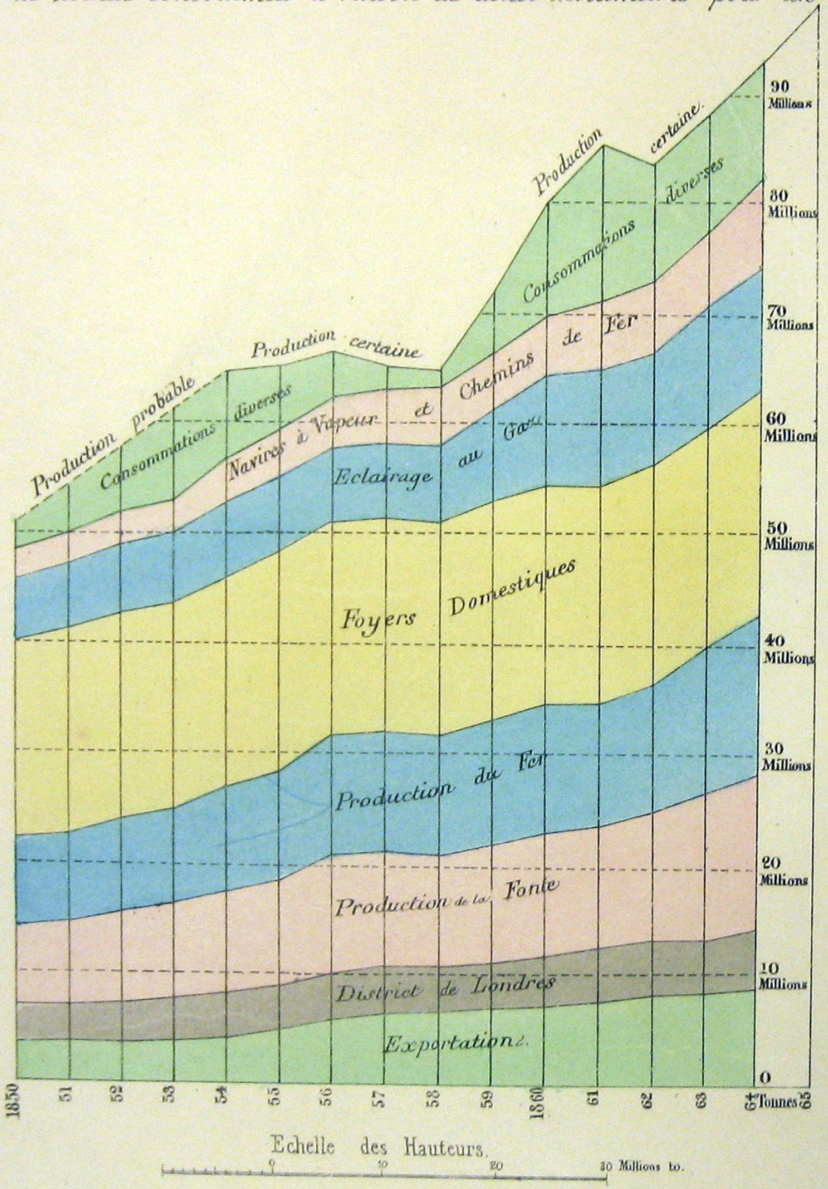

CARTE figurative et approximative de la Houille Anglaise exportée en 1864 dressée par Mr MINARD
Les tonnages exportés dans les différents Ports du Globe sont extraits de Mineral statistics de Mr Robert Hunt
Observation
Paris, le 17 Septembre 1866
Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.
Données admises pour former le Tableau ci-contre
Consommations — Sources des Renseignements
Foyers Domestiques
Production du Fer
Exportation
AMÉRIQUE DU NORD
AMÉRIQUE du SUD
ANGLETERRE
EUROPE
France
Espagne
ASIE
Chine
Inde
AFRIQUE
AUSTRALIE
Passant par le Détroit de Gibraltar 1703
Doublant le Cap de Bonne-Espérance
Doublant le Cap Horn

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 3to de houille pour 1to 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 3to35 de houille pour 1 tonne de fonte convertie en fer; et admettant $\frac{9}{10}$^{es} de la fonte produite convertis en fer.*

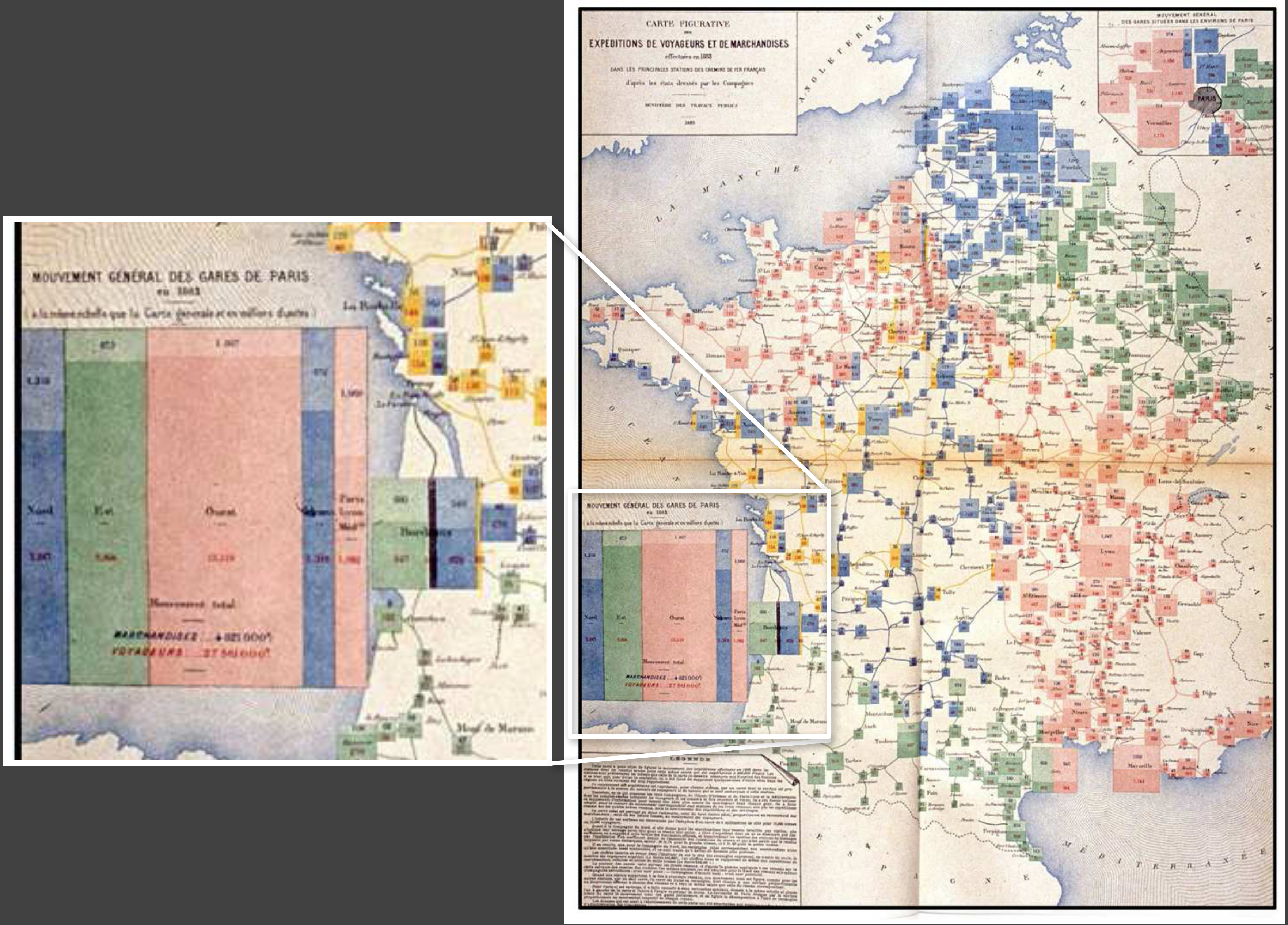
Foyers domestiques : — *En y comprenant les petites manufactures.*
On l'estimait en 1848 à 19 millions de tonnes, (A) *qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.*

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.*

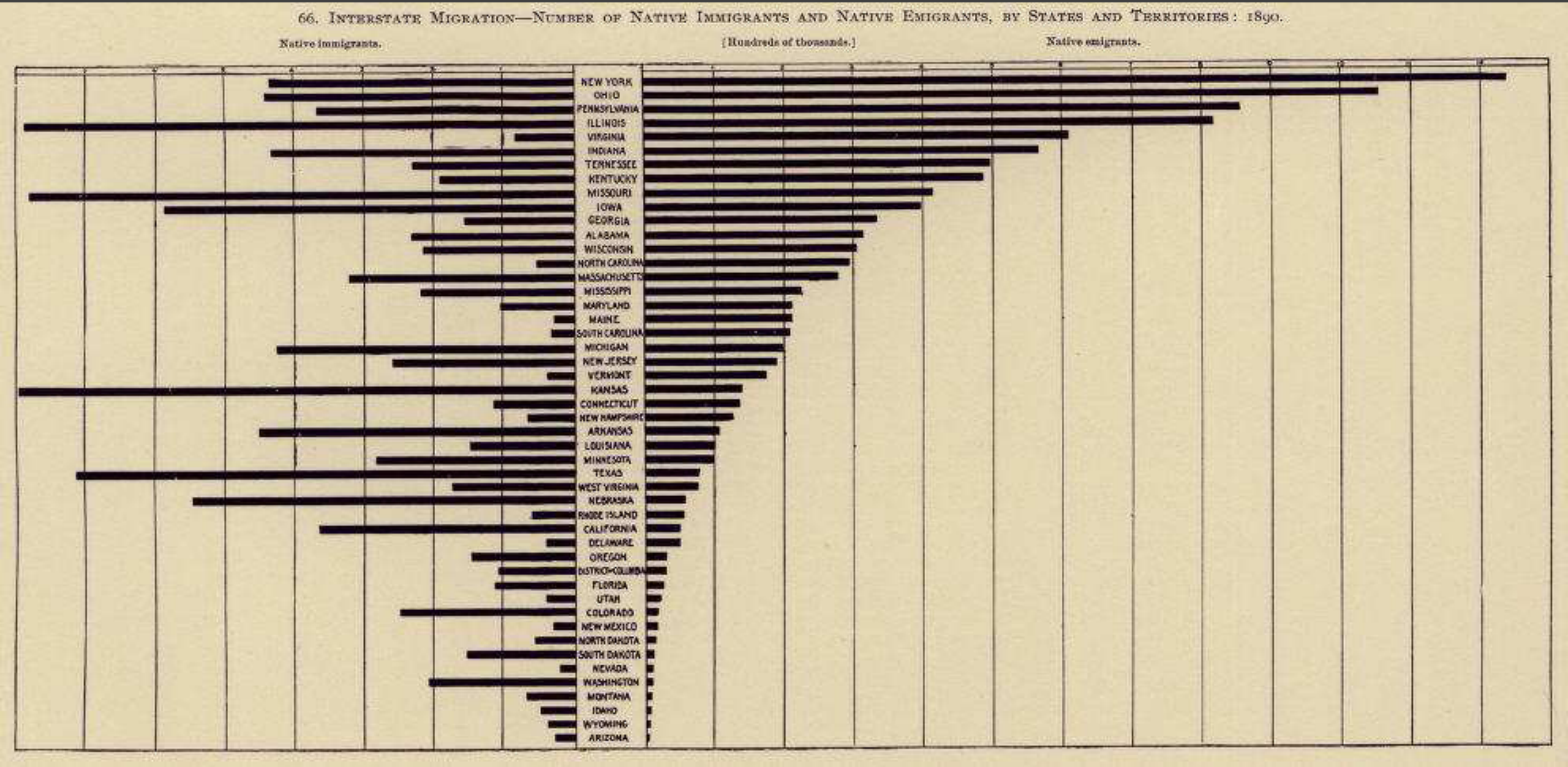
Navigation à vapeur. — *Calculée à raison de 5^{K} houille par cheval vapeur et par heure, le nombre de chevaux étant celui du Steam Vessels pour* 1864, *et les steamers étant supposés marcher la moitié de l'année;*
Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of trade.

(A) *Voir l'excellent article houille de M^{r} Lamé Fleury, Dictionnaire du Commerce Page 111.*


MOUVEMENT GÉNÉRAL DES GARES DE PARIS
en 1883
(à la même échelle que la Carte générale et en milliers d'unités)
Nord
Est
Ouest
Bordeaux
Mouvement total
MARCHANDISES 4 821 000t
VOYAGEURS 27 561 000
CARTE FIGURATIVE
des
EXPÉDITIONS DE VOYAGEURS ET DE MARCHANDISES
effectuées en 1883
DANS LES PRINCIPALES STATIONS DES CHEMINS DE FER FRANÇAIS
d'après les états dressés par les Compagnies
MINISTÈRE DES TRAVAUX PUBLICS
1885
MOUVEMENT GÉNÉRAL DES GARES SITUÉES DANS LES ENVIRONS DE PARIS
PARIS
Versailles
Lille
Lyon
Marseille
Nancy
LA MANCHE
ANGLETERRE
BELGIQUE
ALLEMAGNE
SUISSE
ITALIE
OCÉAN
ESPAGNE
MÉDITERRANÉE
LÉGENDE

1786

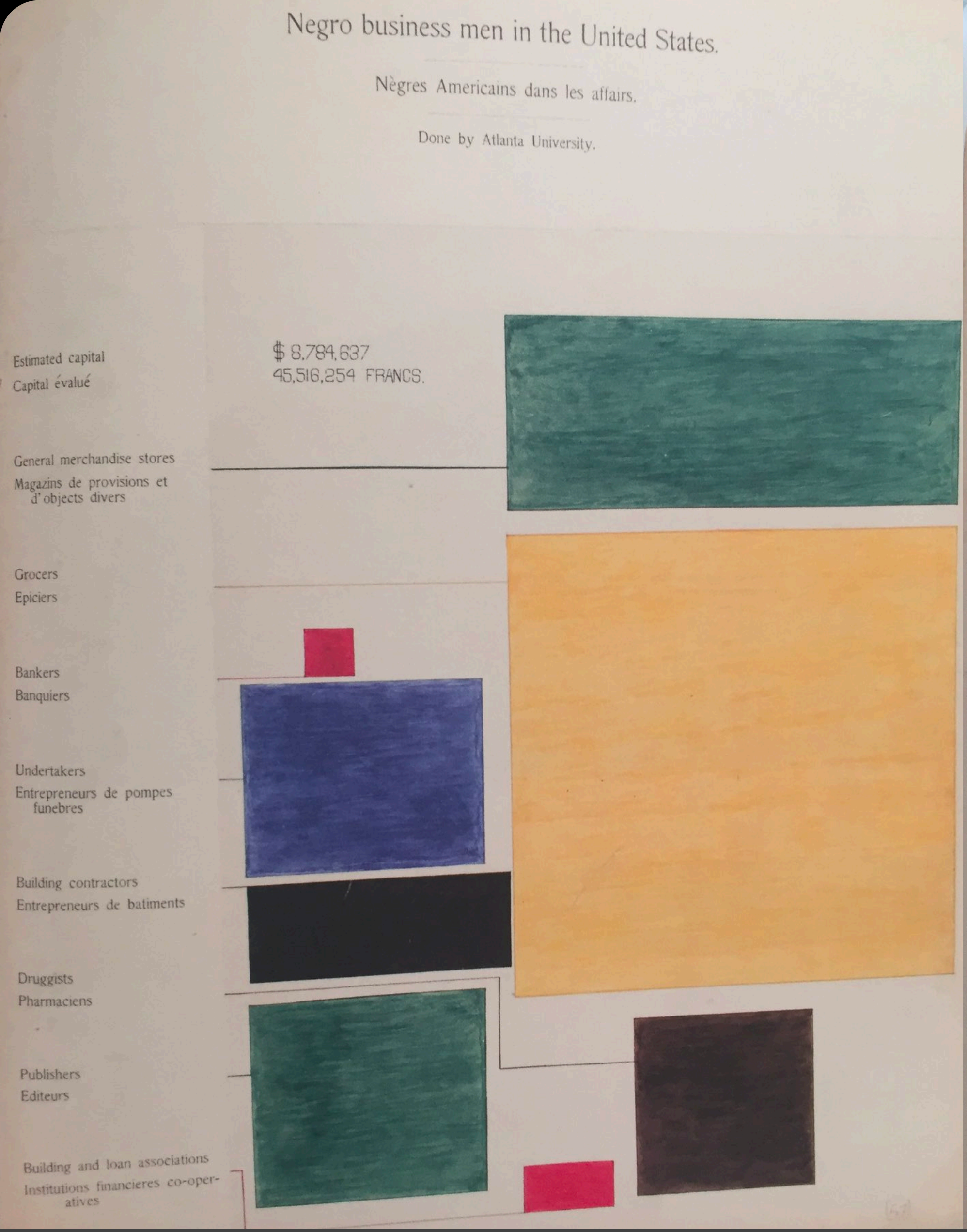
1884 Rail Passengers and Freight from Paris

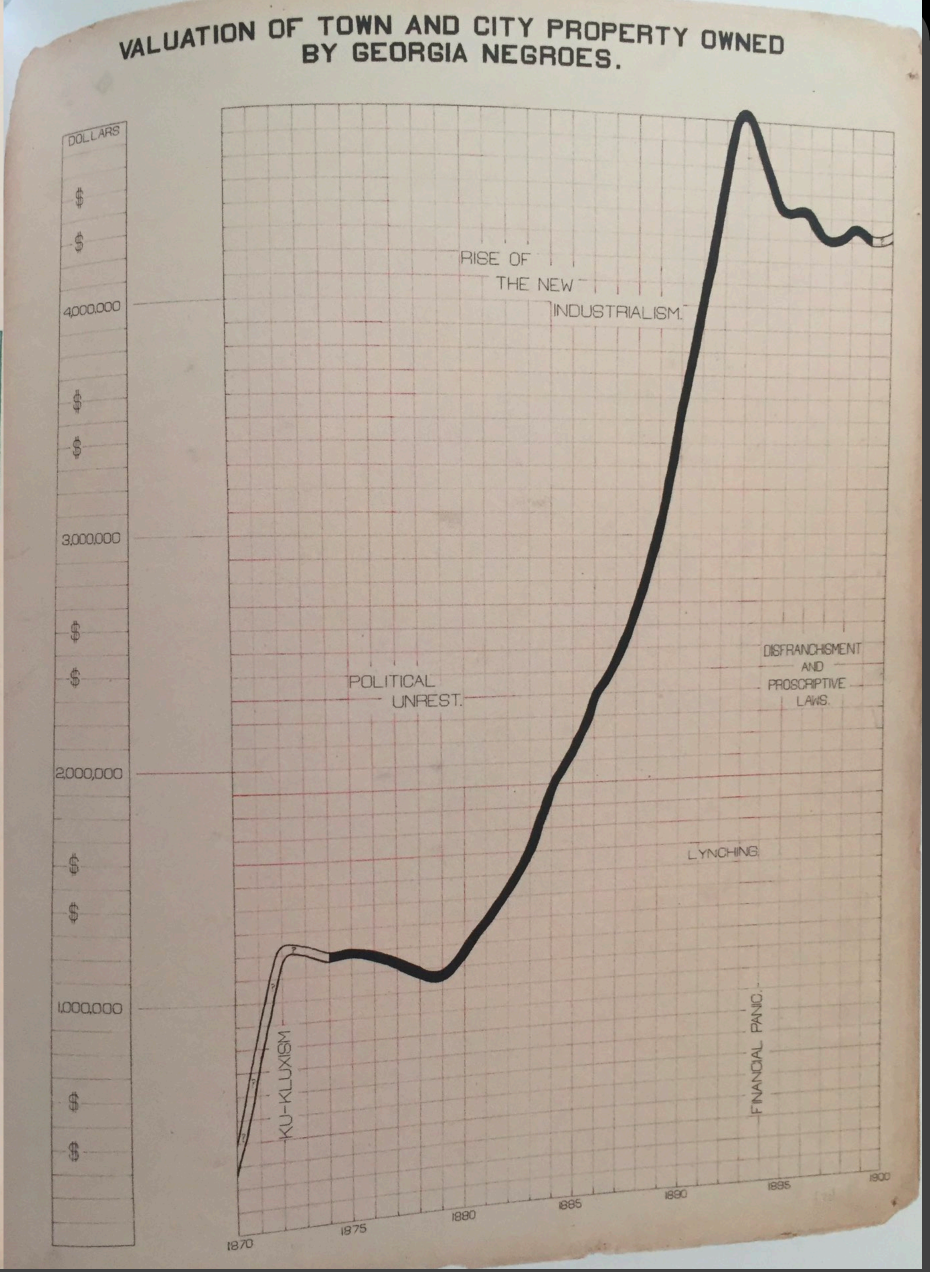


66. INTERSTATE MIGRATION—NUMBER OF NATIVE IMMIGRANTS AND NATIVE EMIGRANTS, BY STATES AND TERRITORIES: 1890.

1786

1890 Statistical Atlas of the Eleventh U.S. Census

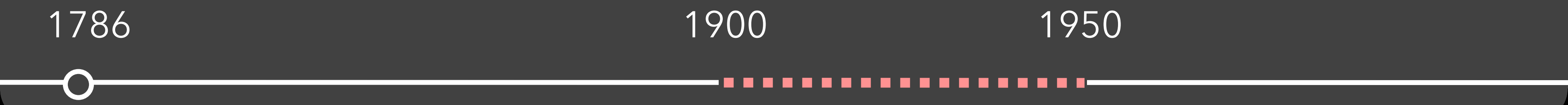




1786

1900 Visualizing Black America, W. E. B. DuBois et al.

The Rise of Statistics



1786

1900

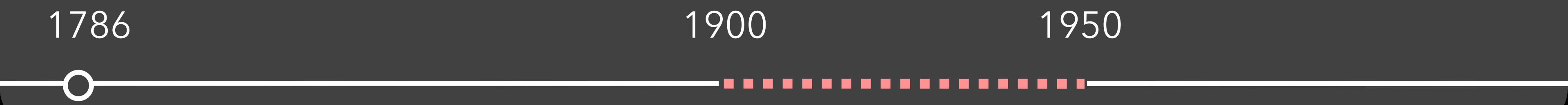
1950

Rise of **formal statistical methods** in the physical and social sciences

Little innovation in graphical methods

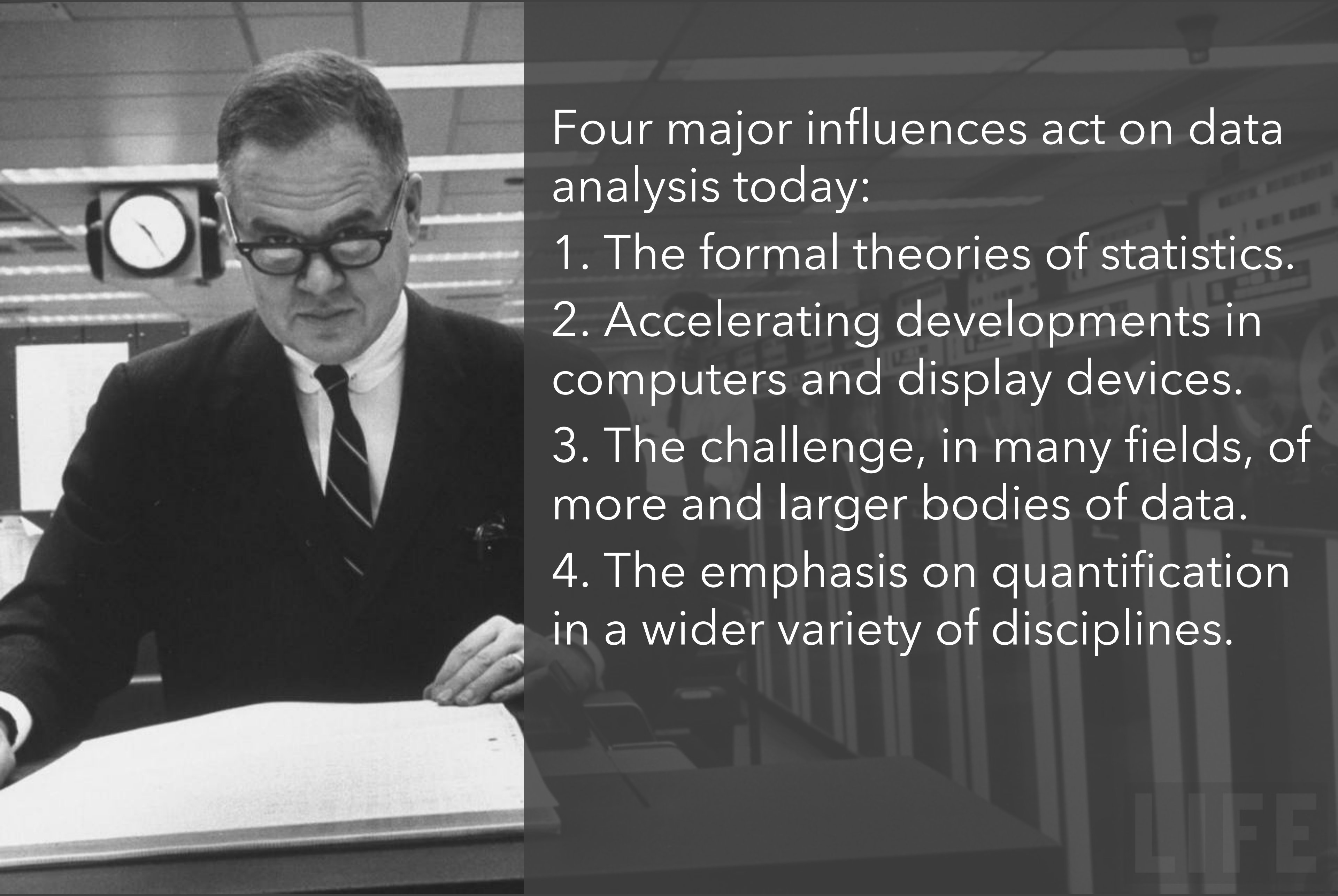
A period of **application and popularization**

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



1786

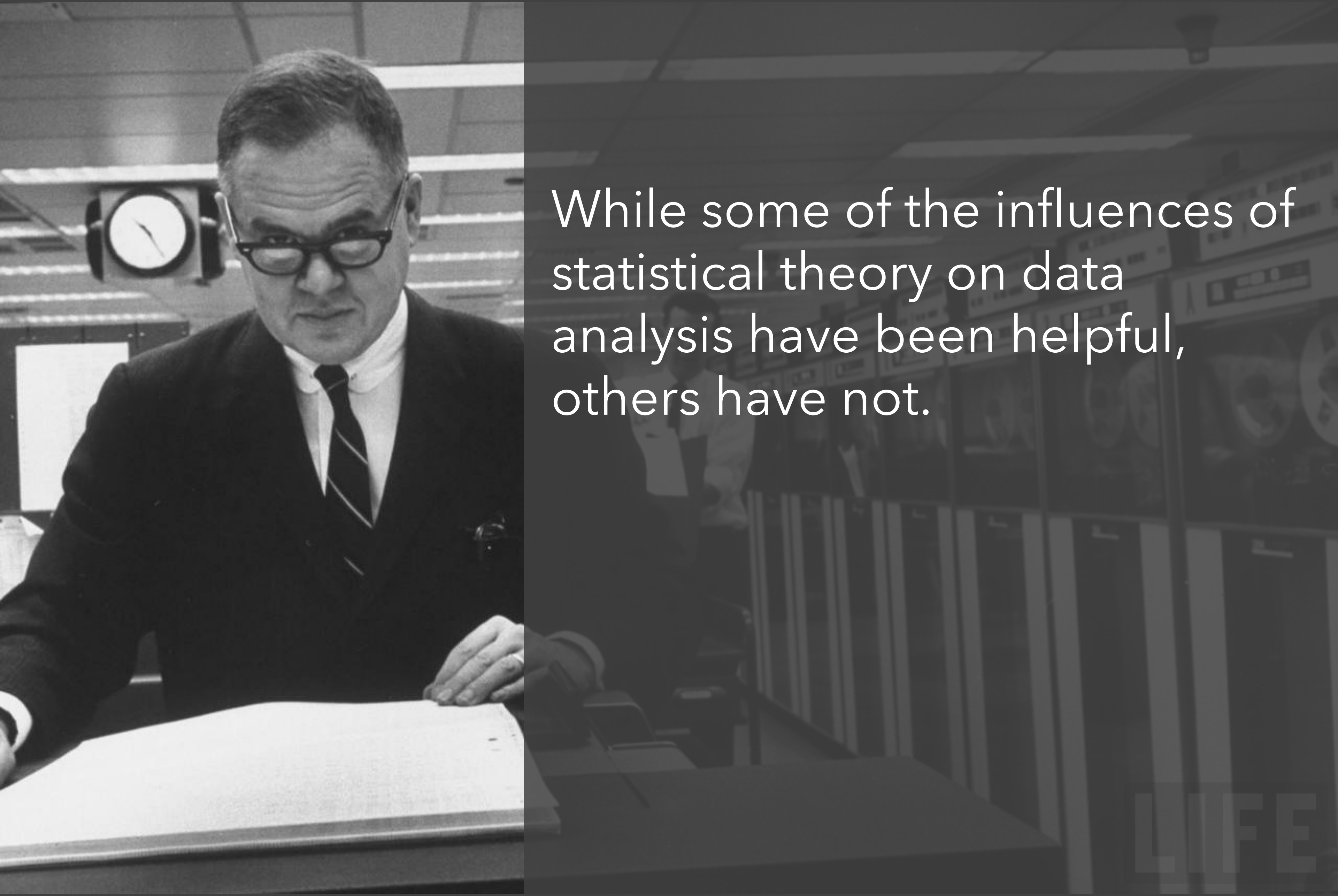
Data Analysis & Statistics, Tukey 1962



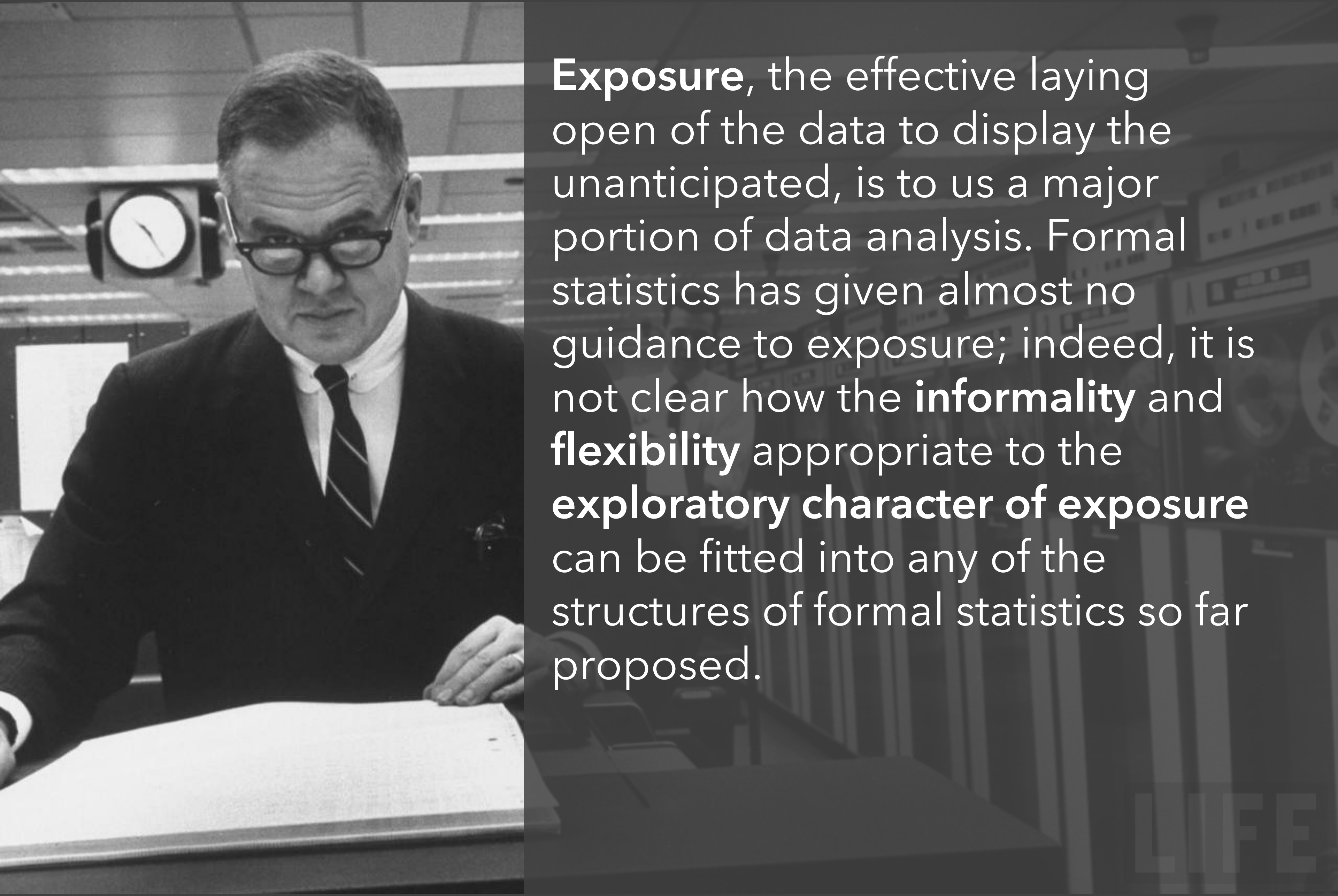
Four major influences act on data analysis today:

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.

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

$u_X = 9.0$ $\sigma_X = 3.317$

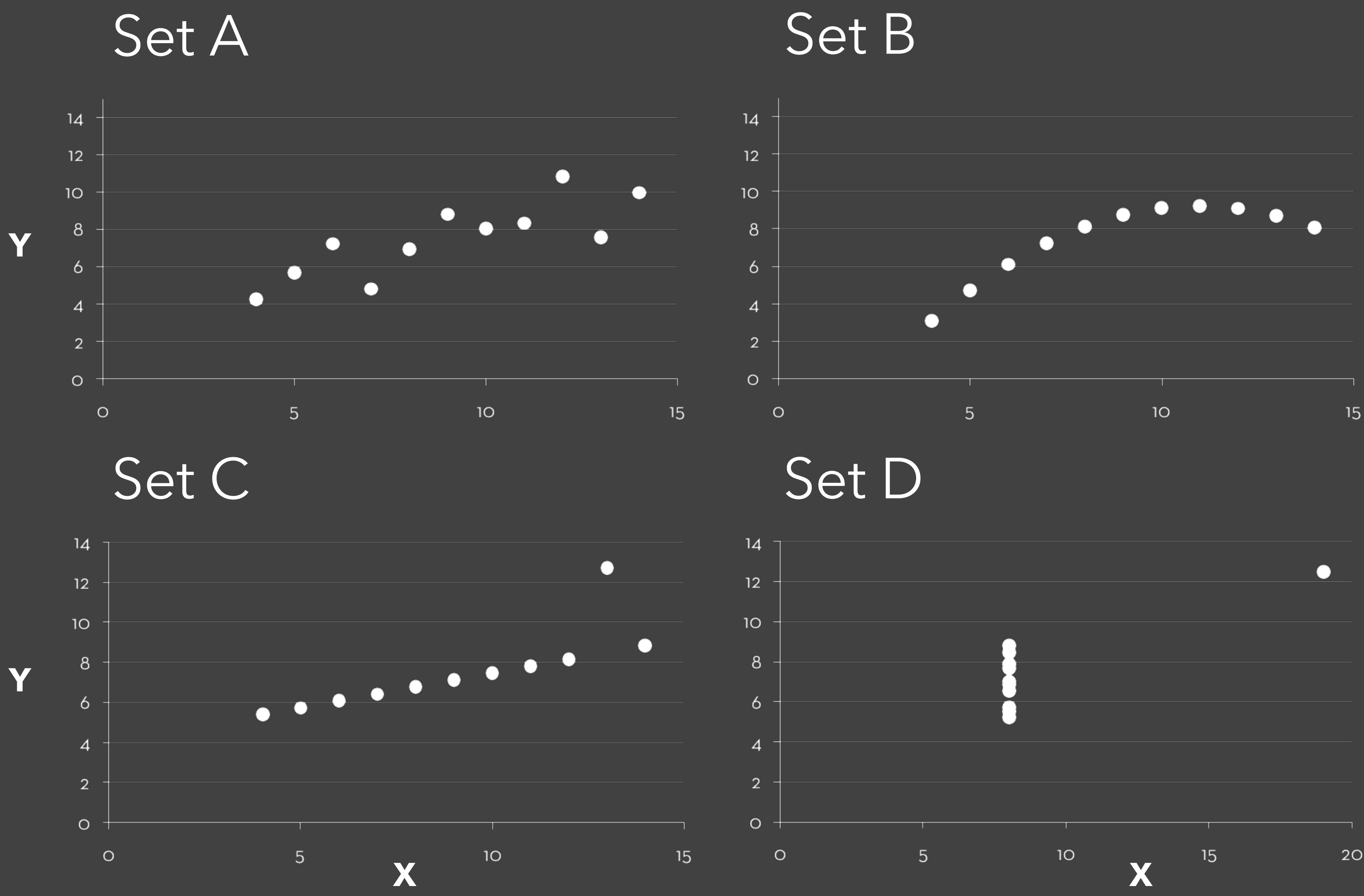
$u_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression

$Y^2 = 3 + 0.5 X$

$R^2 = 0.67$

[Anscombe 1973]



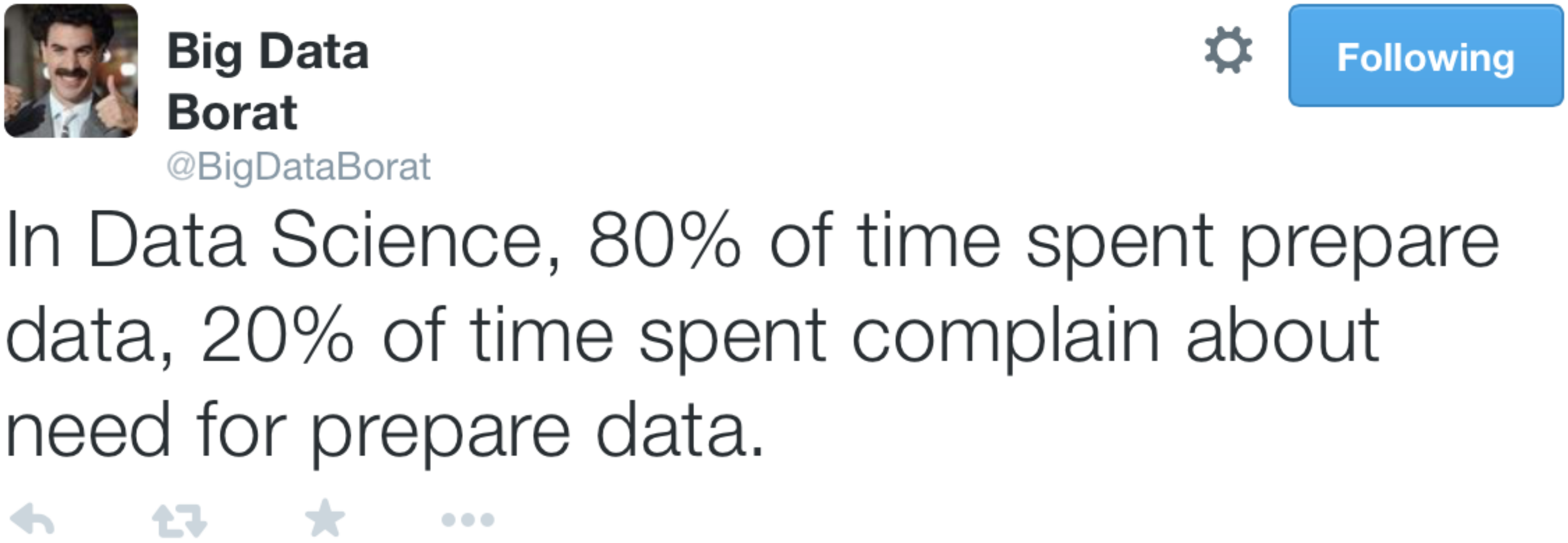
[Anscombe 1973]

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.

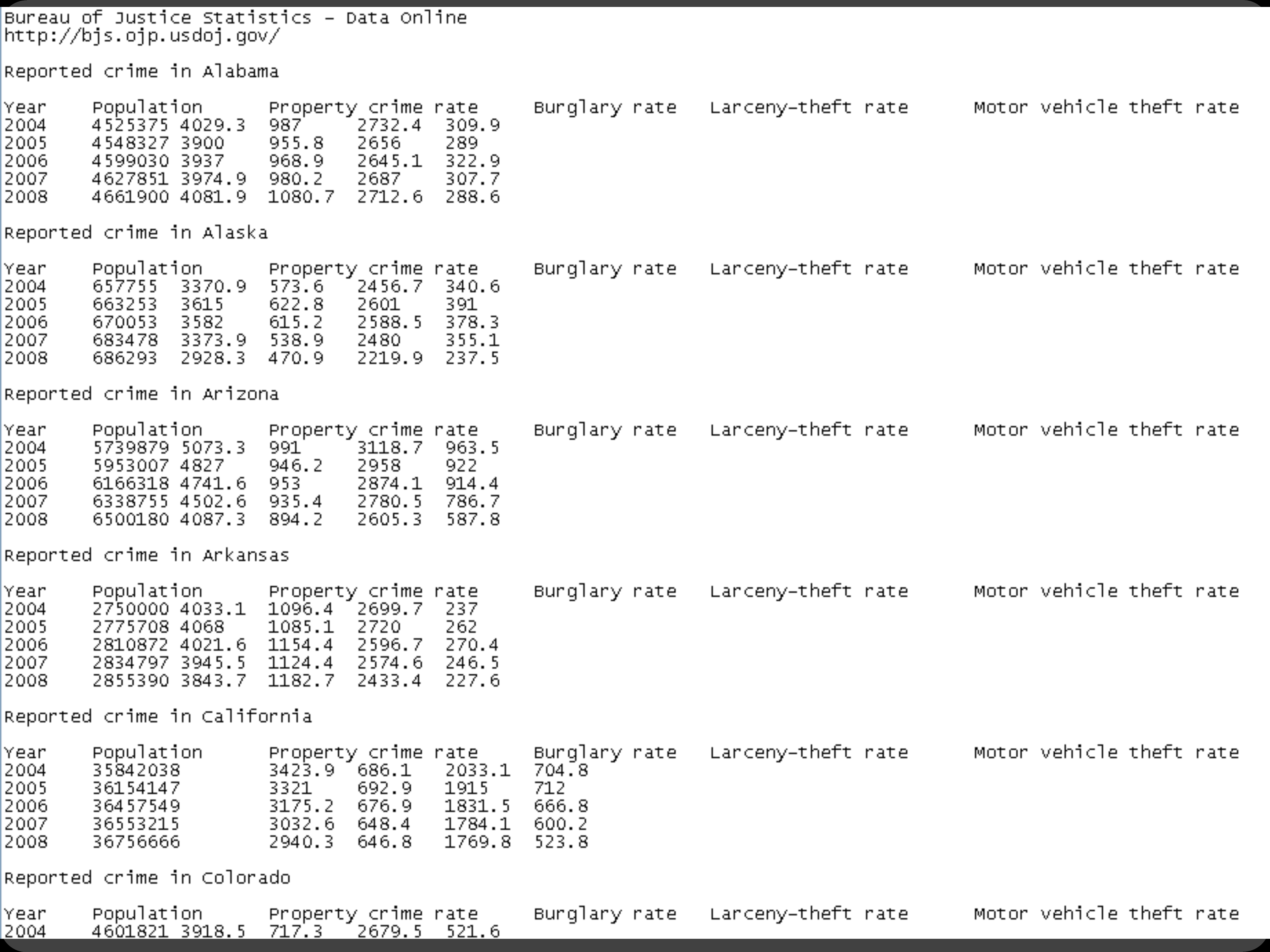
Anonymous Data Scientist

[Kandel et al. '12]



Big Data Borat
@BigDataBorat

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
```

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: arquero (JS), dplyr (R), pandas (Python)
Trifacta Wrangler http://www.trifacta.com/products/wrangler/
Open Refine http://openrefine.org/

Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In "tidy" data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not "tidy"!

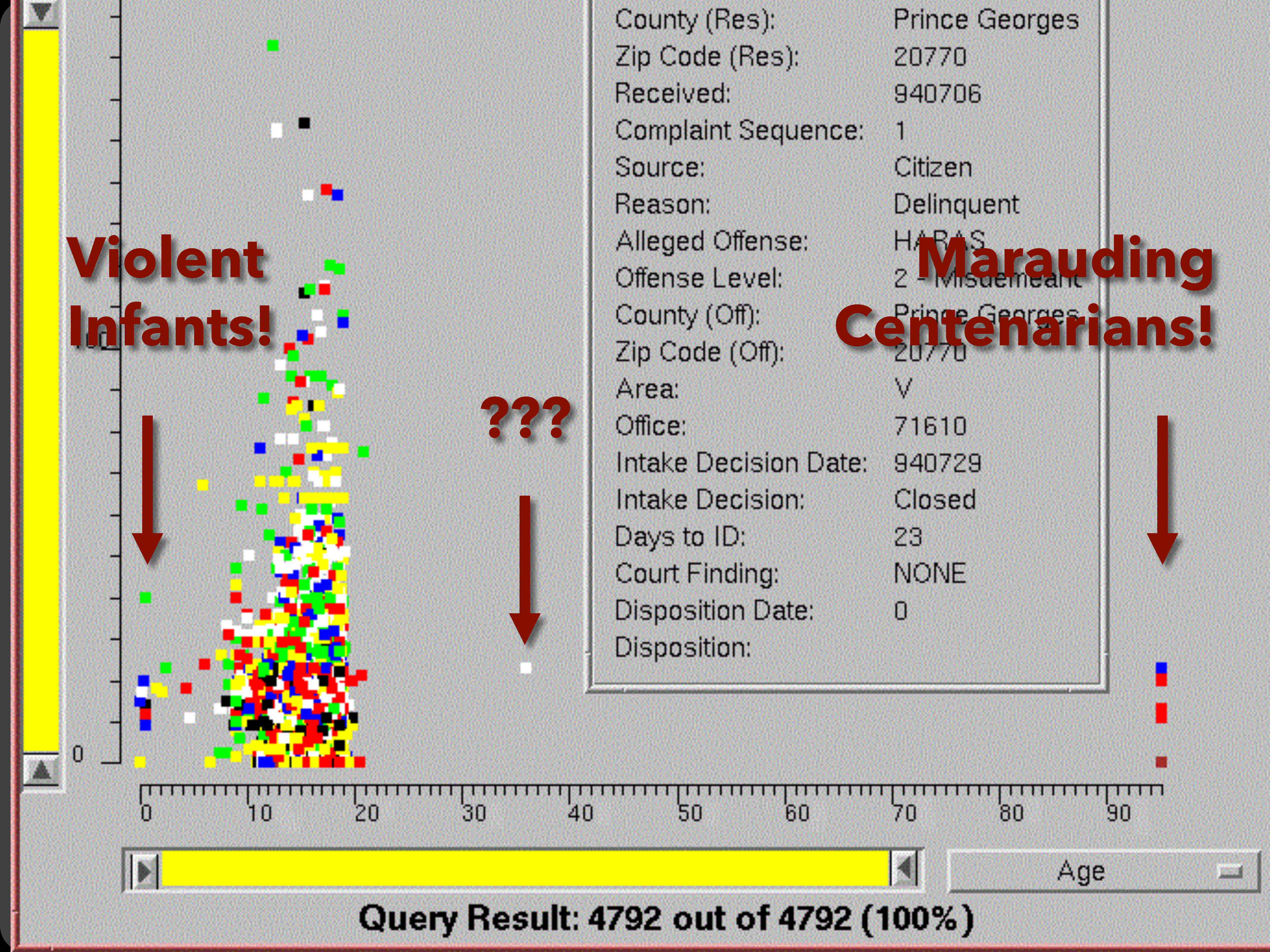
(This is a variant of normalized forms in DB theory)

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 - Misdemeant
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:	

Query Result: 4792 out of 4792 (100%)

Violent Infants!

???

Marauding Centenarians!

Visualize Degrees by School?

<table>
<tr><td>Berkeley</td><td>|||||||||||||||||||||||||||||||</td></tr>
<tr><td>Cornell</td><td>||||</td></tr>
<tr><td>Harvard</td><td>|||||||||</td></tr>
<tr><td>Harvard University</td><td>|||||||</td></tr>
<tr><td>Stanford</td><td>||||||||||||||||||||</td></tr>
<tr><td>Stanford University</td><td>||||||||||</td></tr>
<tr><td>UC Berkeley</td><td>|||||||||||||||||||||</td></tr>
<tr><td>UC Davis</td><td>||||||||||</td></tr>
<tr><td>University of California at Berkeley</td><td>|||||||||||||||</td></tr>
<tr><td>University of California, Berkeley</td><td>||||||||||||||||||</td></tr>
<tr><td>University of California, Davis</td><td>|||</td></tr>
</table>

Data Quality Hurdles

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

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

Administrivia

A1: Visualization Design

Pick a **guiding question**, use it to title your vis.
Design a **static visualization** for that question.
You are free to **use any tools** (inc. pen & paper).

Deliverables (upload via Canvas; see A1 page)
Image of your visualization (PNG or JPG format)
Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm, Wednesday April 6**.

Tableau Tutorial (Optional)

Friday April 8, 1-2pm

Zoom link available on Canvas

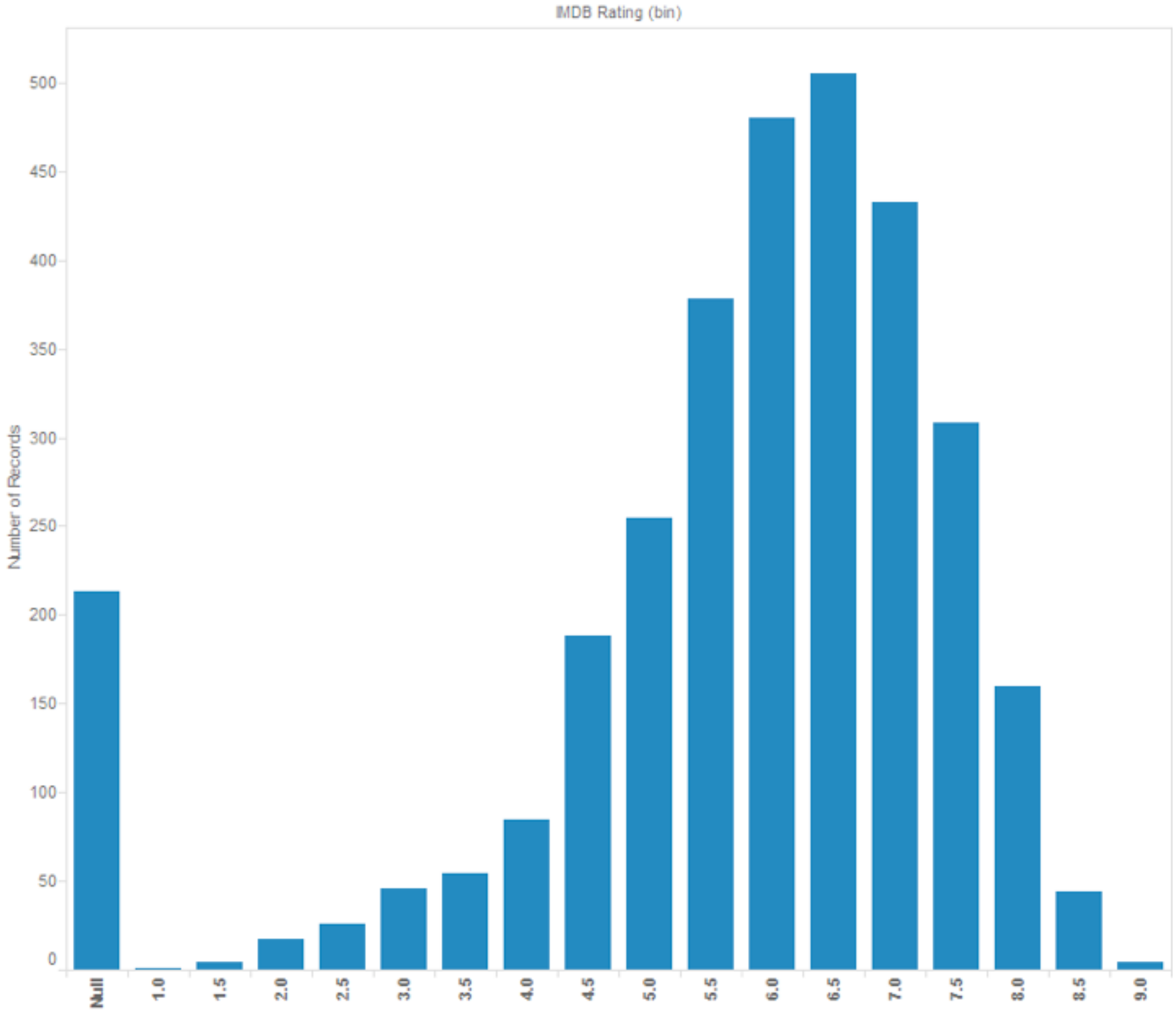
Session will be recorded.

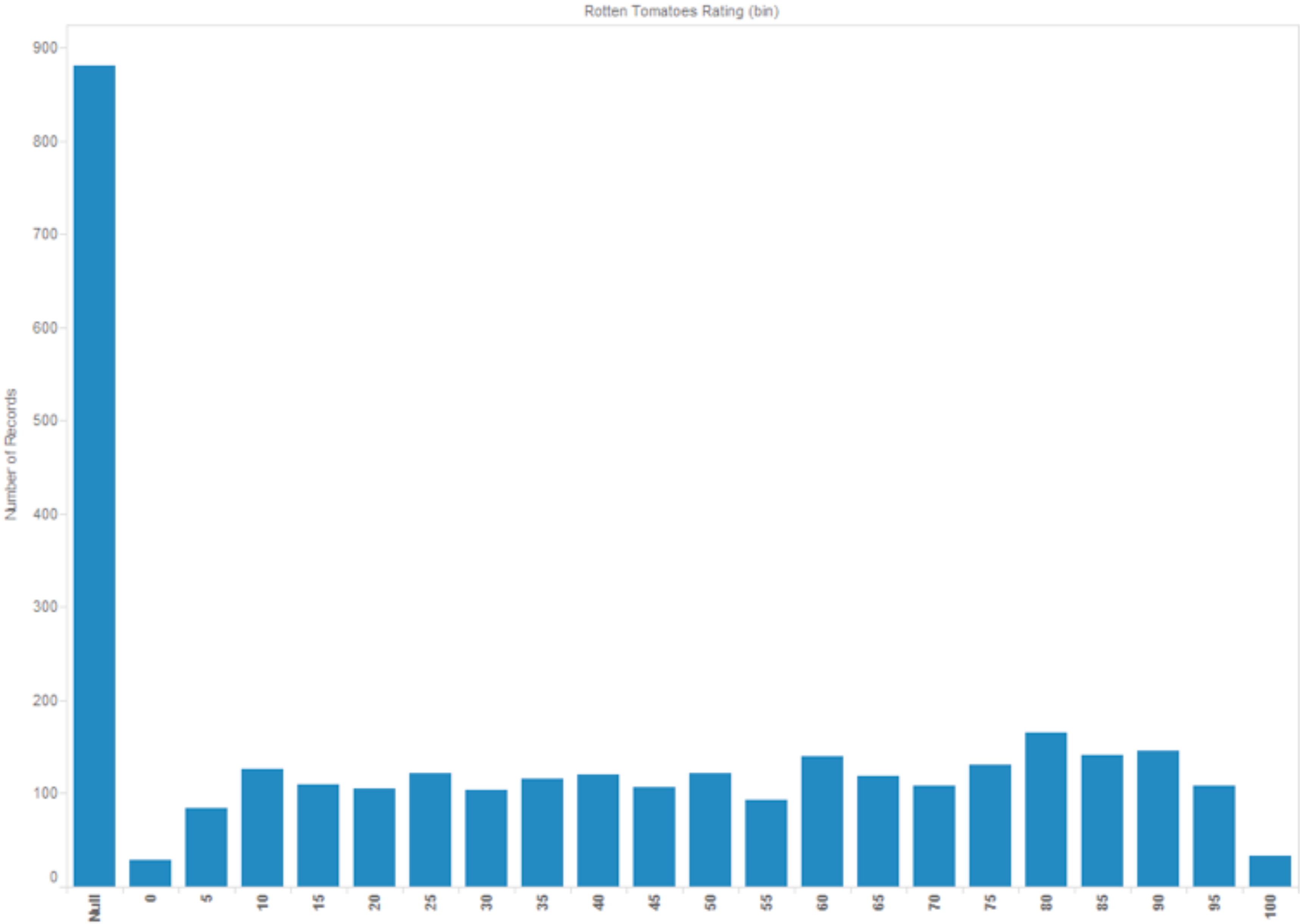
Break Time!

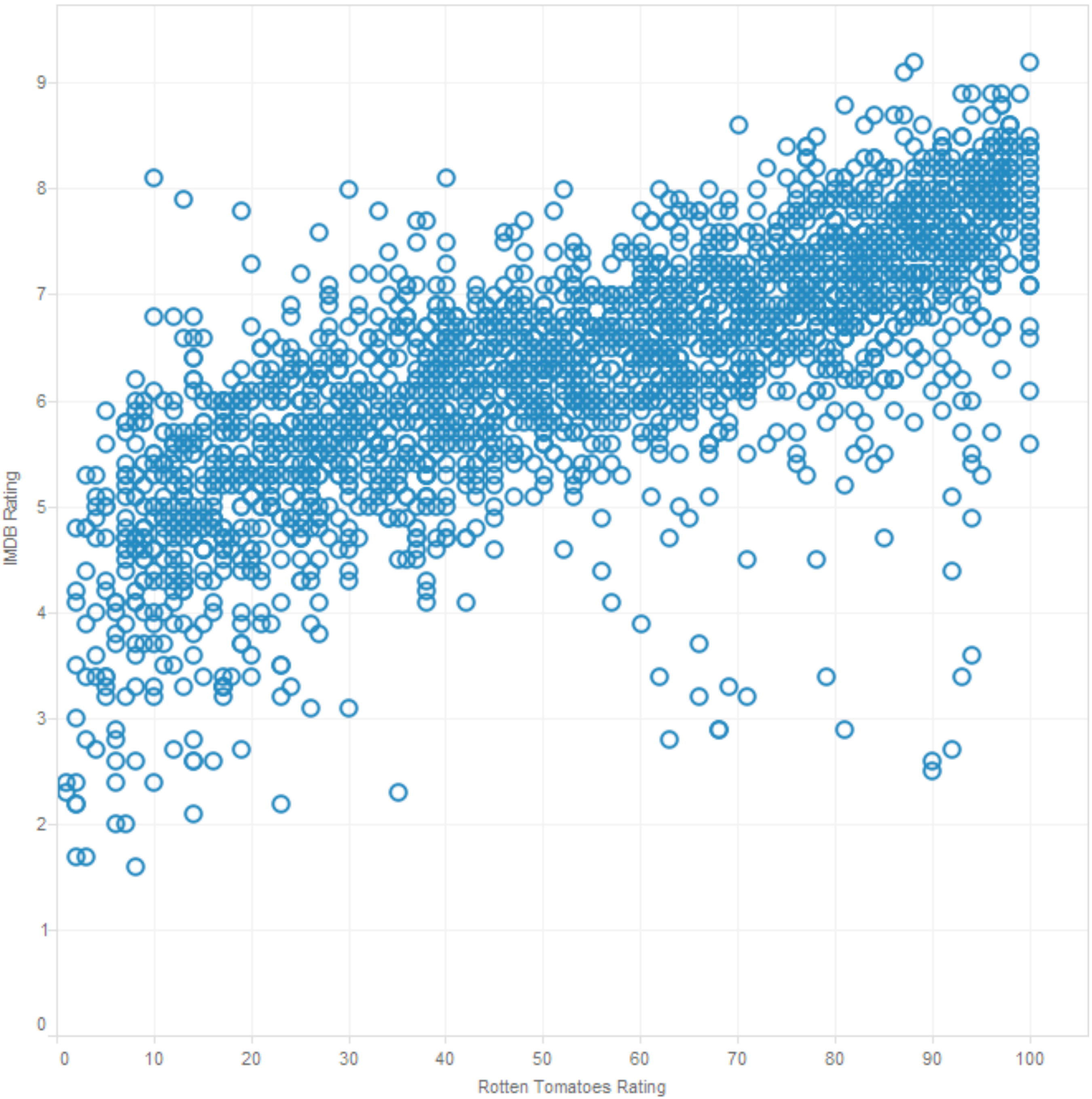
Analysis Example: Motion Pictures Data

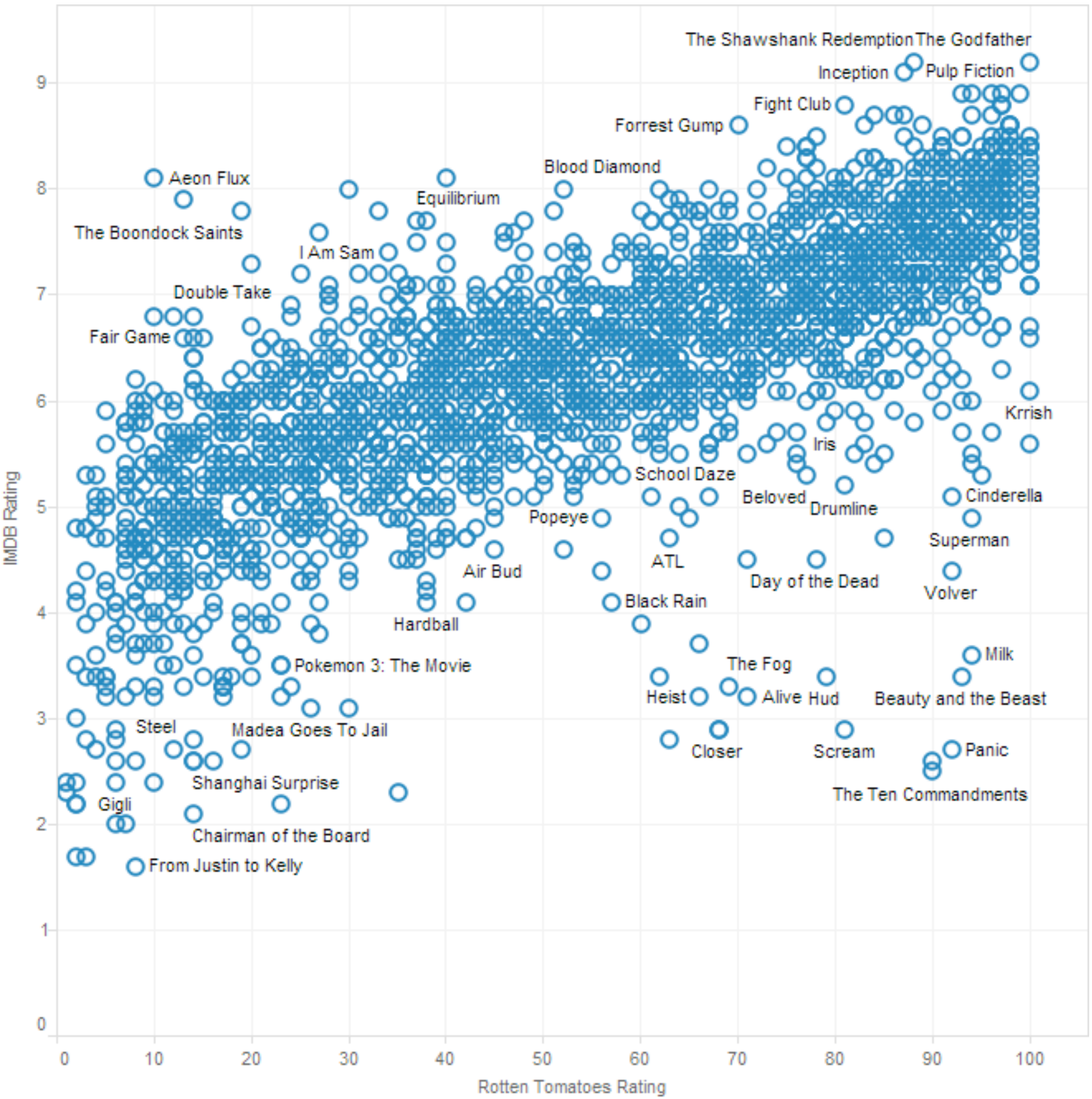
Motion Pictures Data

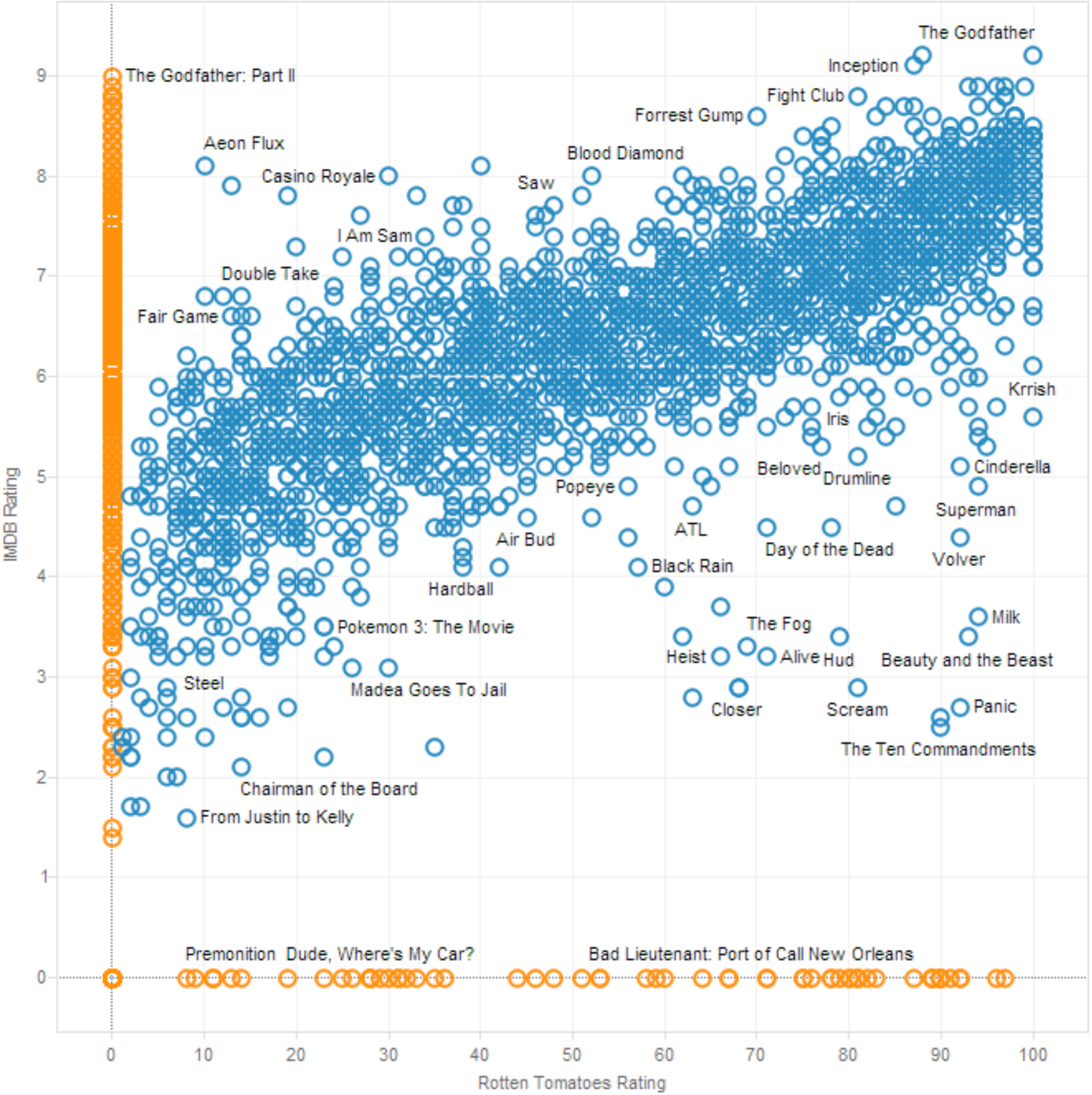
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)


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

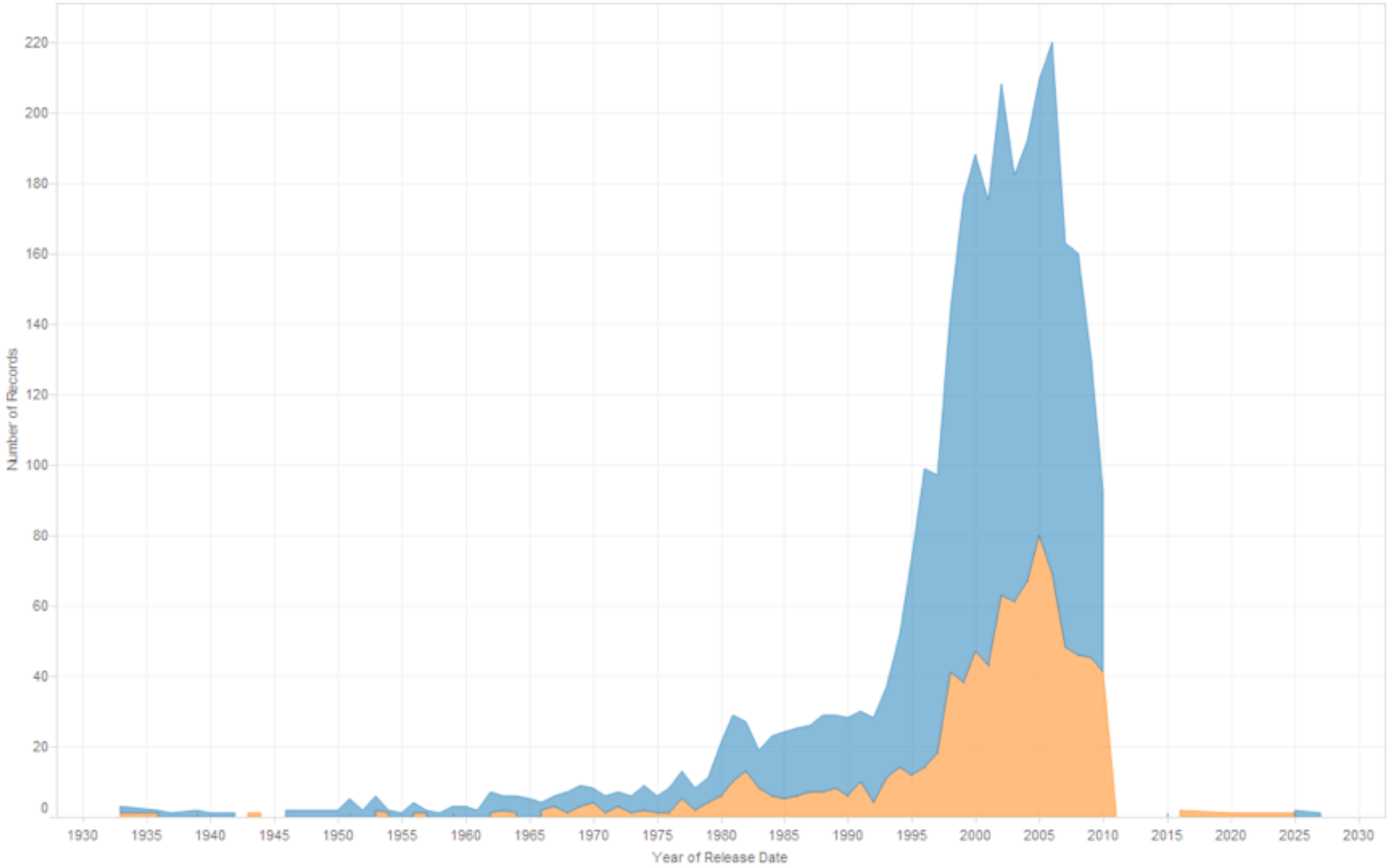

Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100


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


The Shawshank Redemption
The Godfather
Inception
Pulp Fiction
Fight Club
Forrest Gump
Blood Diamond
Aeon Flux
Equilibrium
The Boondock Saints
I Am Sam
Double Take
Fair Game
Krrish
Iris
School Daze
Beloved
Drumline
Cinderella
Popeye
Superman
Air Bud
ATL
Day of the Dead
Volver
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
Shanghai Surprise
The Ten Commandments
Gigli
Chairman of the Board
From Justin to Kelly
IMDB Rating
0
1
2
3
4
5
6
7
8
9
Rotten Tomatoes Rating
0
10
20
30
40
50
60
70
80
90
100



The Godfather
Inception
Fight Club
Forrest Gump
The Godfather: Part II
Aeon Flux
Casino Royale
Blood Diamond
Saw
I Am Sam
Double Take
Fair Game
Krrish
Iris
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Black Rain
Volver
Hardball
Milk
Pokemon 3: The Movie
The Fog
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
9
8
7
6
5
4
3
2
1
0
Rotten Tomatoes Rating
0
10
20
30
40
50
60
70
80
90
100


220
200
180
160
140
120
100
80
60
40
20
0
Number of Records
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?

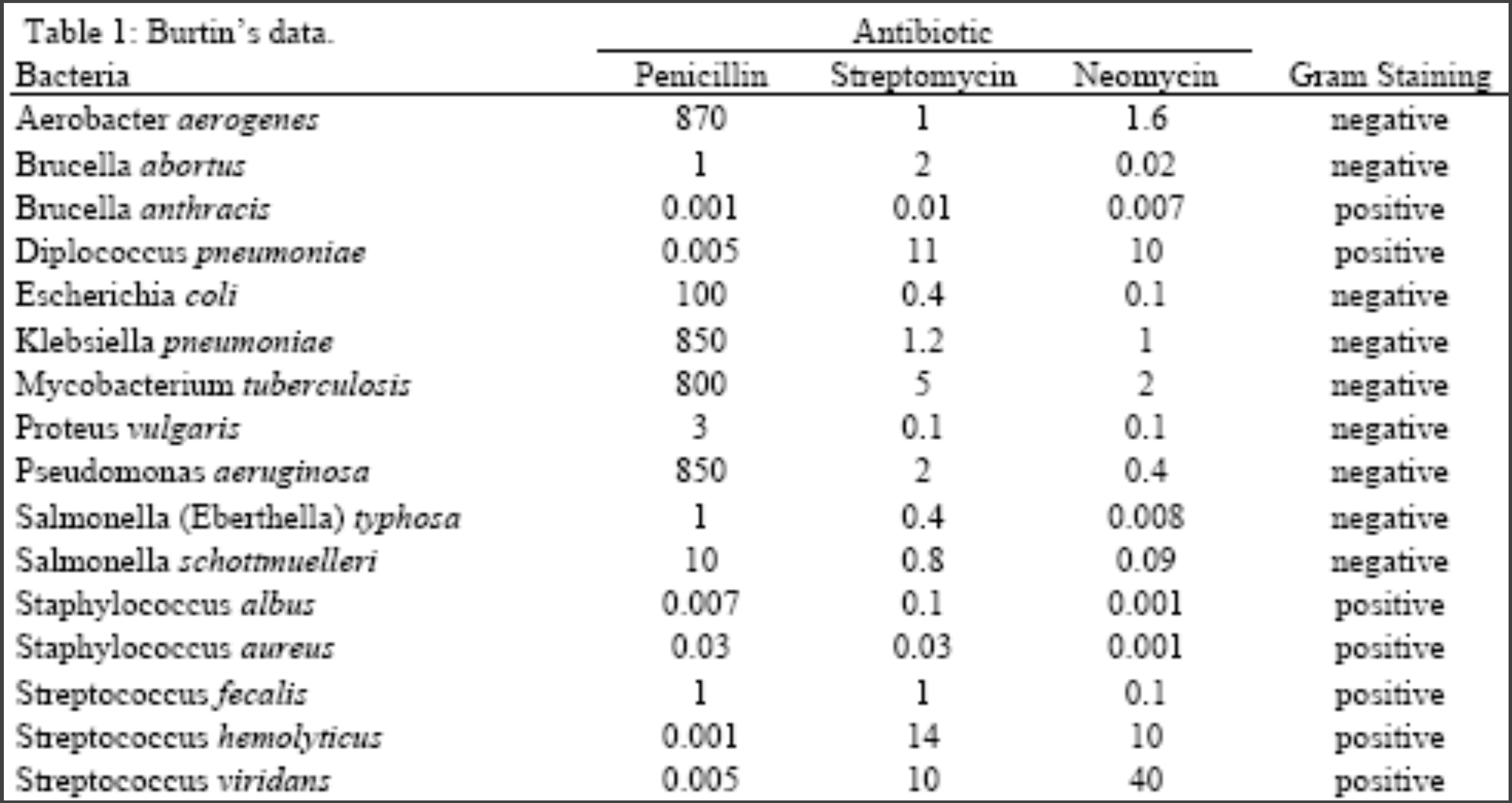
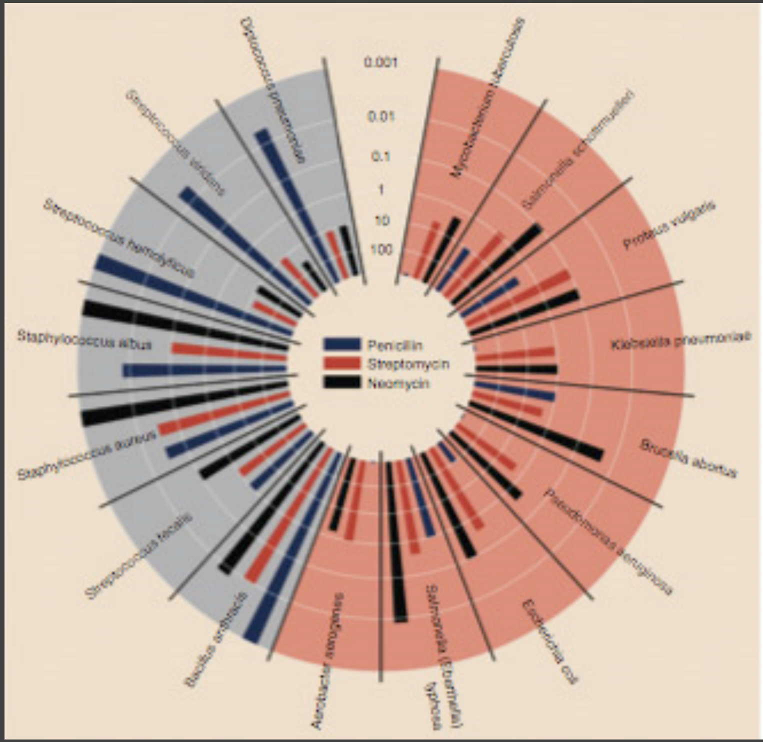
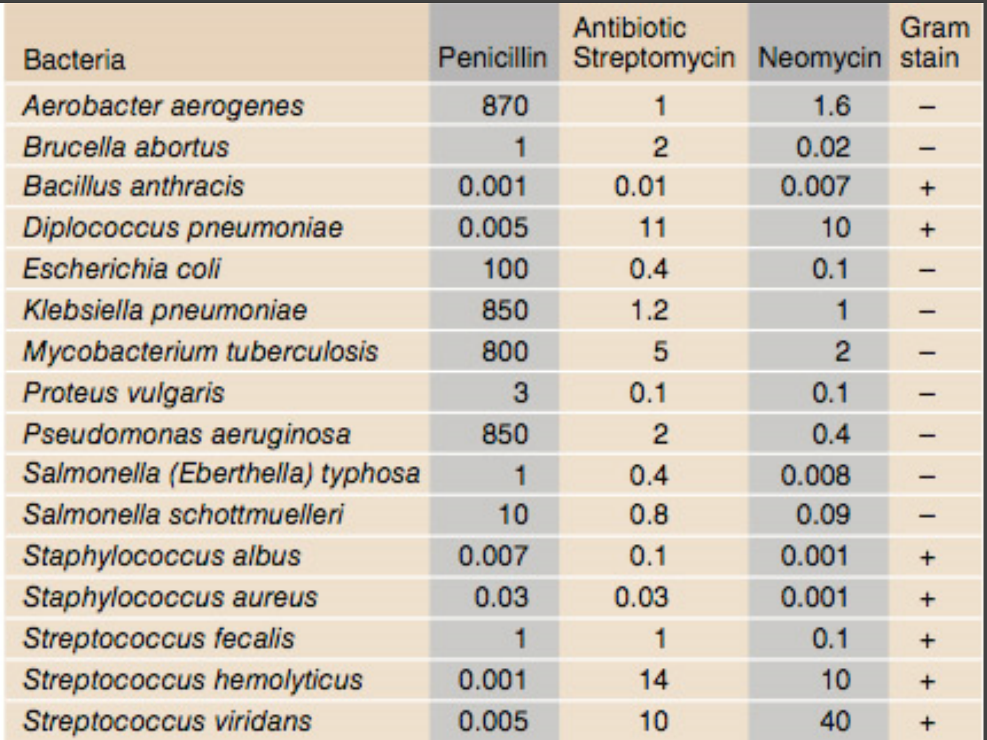


Table 1: Burtin's data.

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?

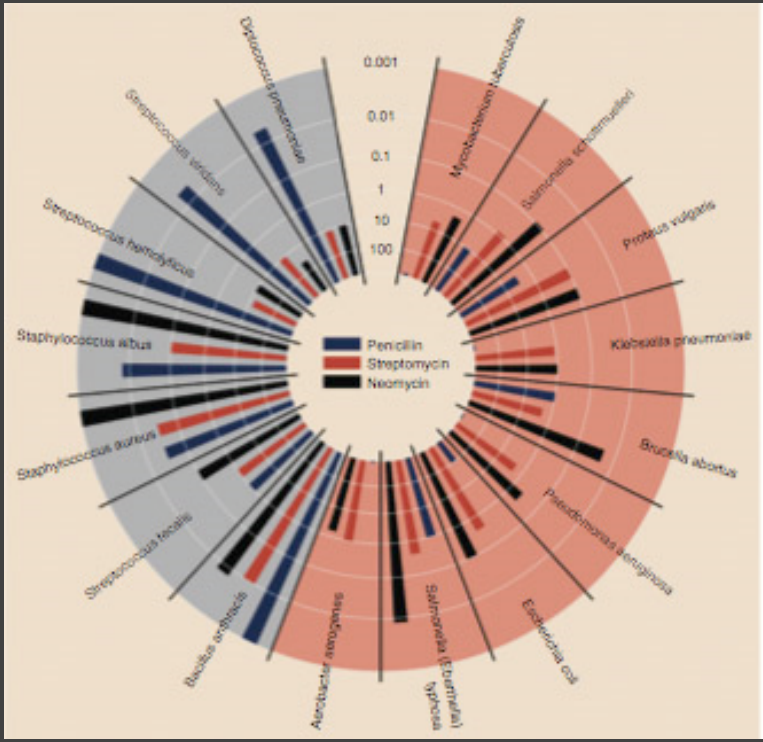


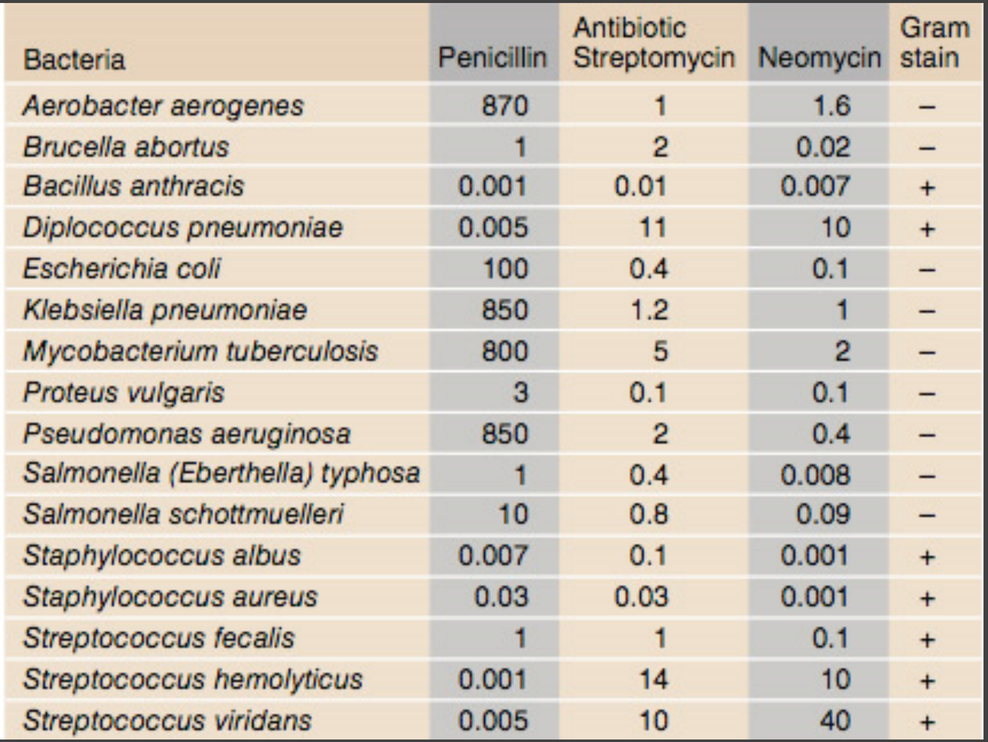


		Antibiotic		Gram
Bacteria	Penicillin	Streptomycin	Neomycin	stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2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

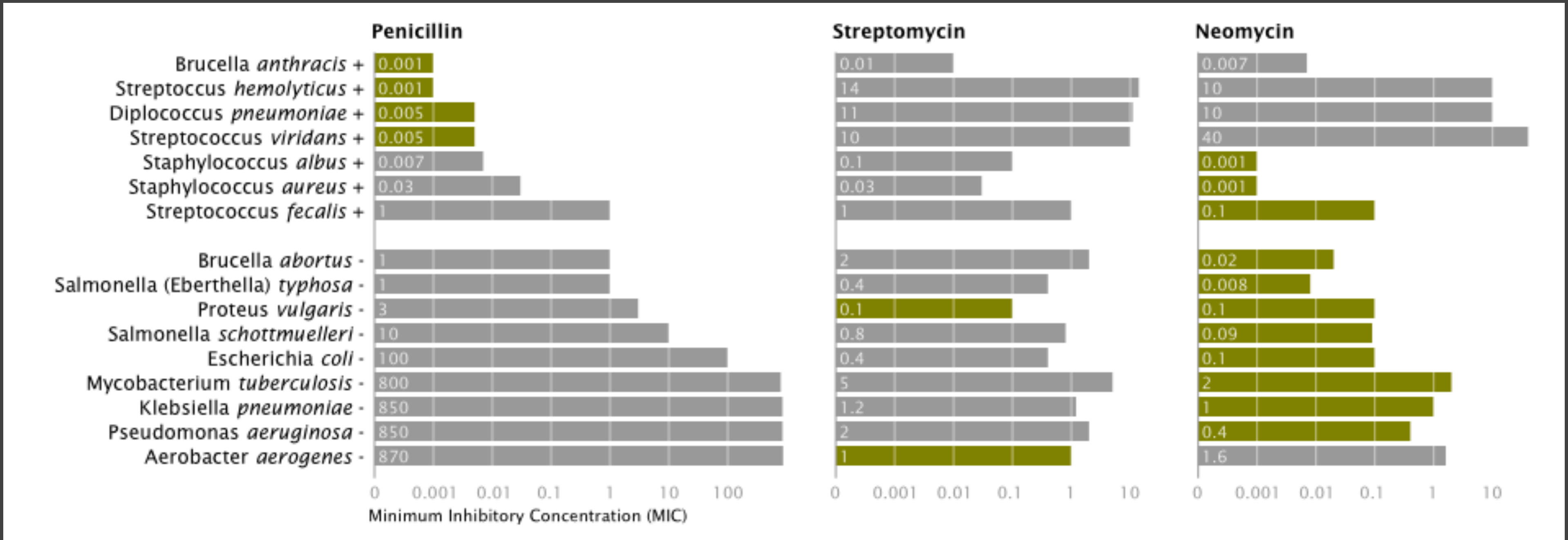




Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–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

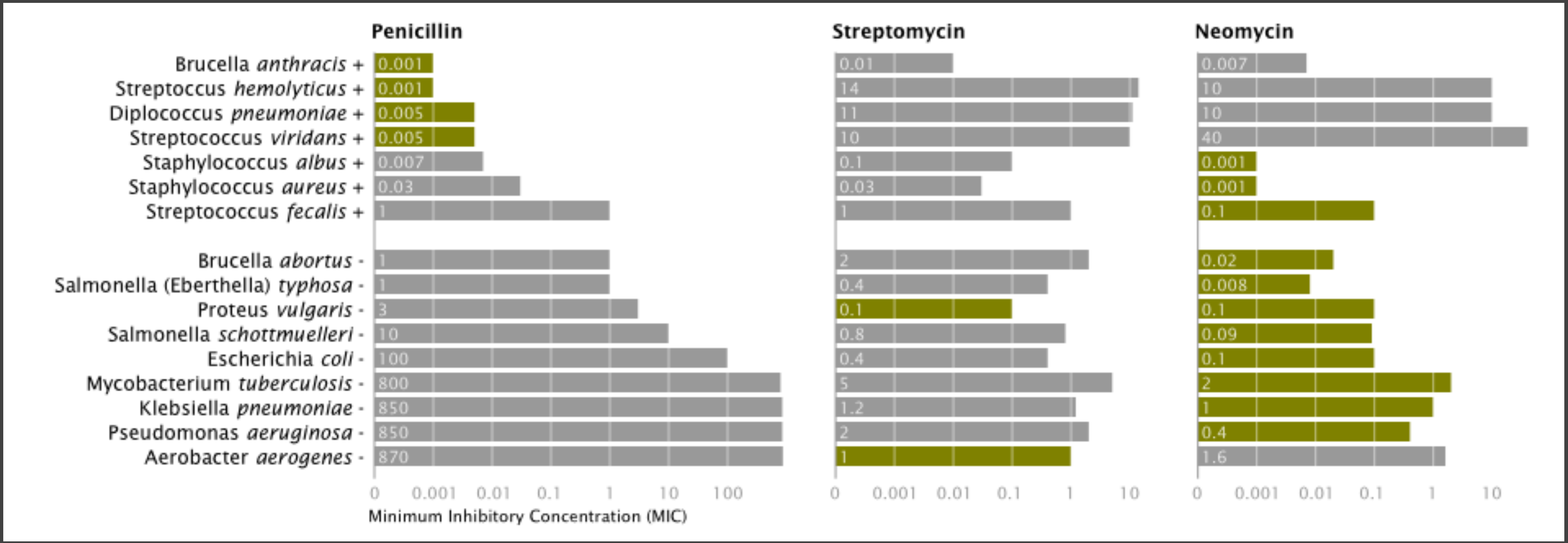
Radius: 1 / log(MIC)
Bar Color: Antibiotic
Background Color: Gram Staining

How do the drugs compare?

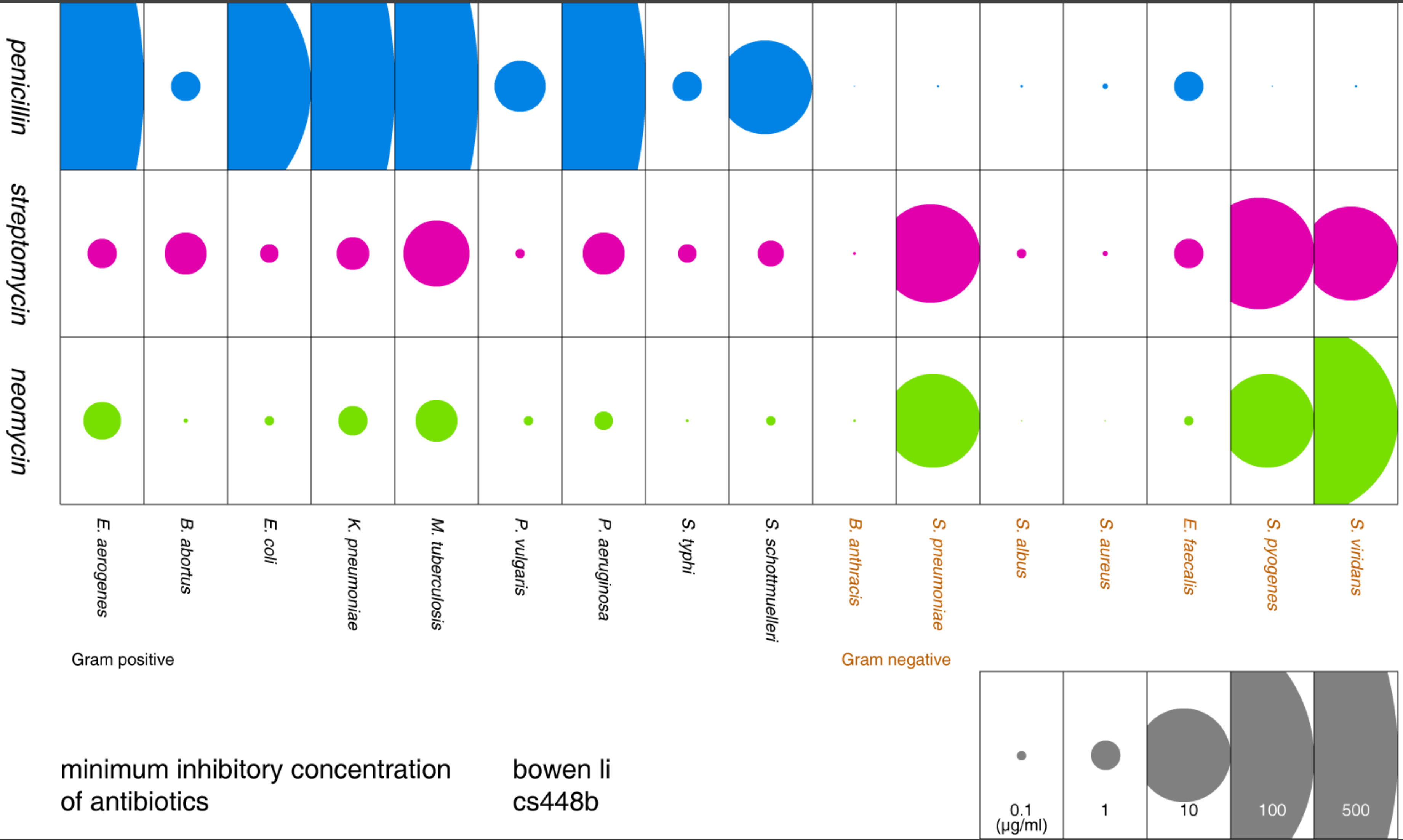


Mike Bostock
Stanford CS448B, Winter 2009

How do the drugs compare?



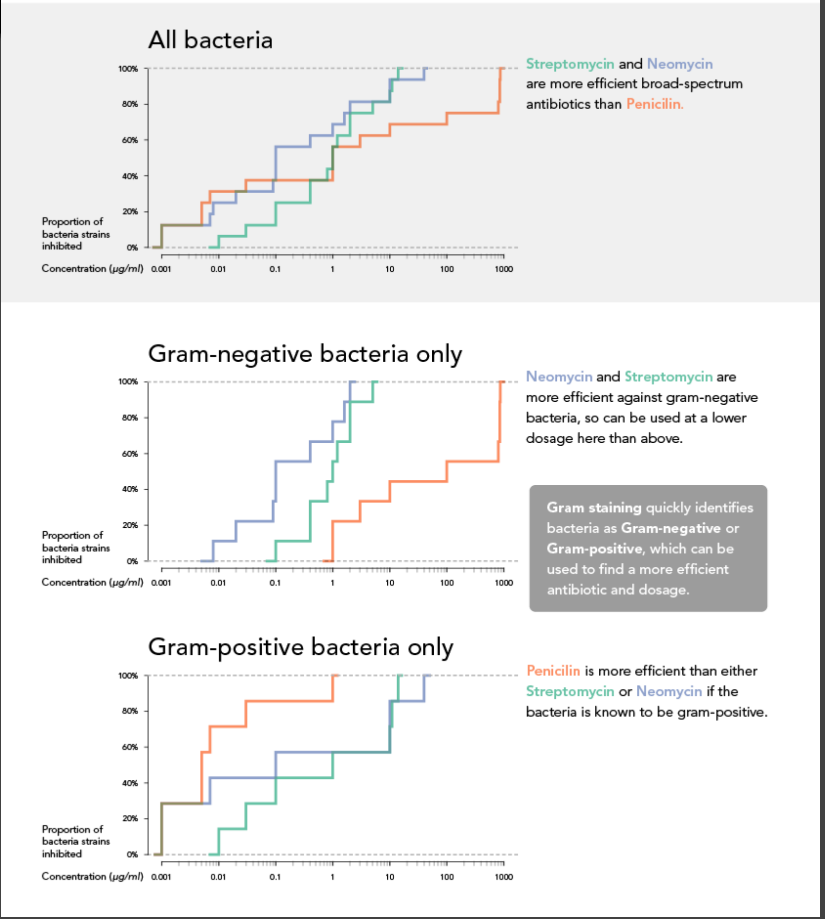
X-axis: Antibiotic | log(MIC)
Y-axis: Gram-Staining | Species
Color: Most-Effective?



minimum inhibitory concentration
of antibiotics

bowen li
cs448b

Bowen Li
Stanford CS448B, Fall 2009

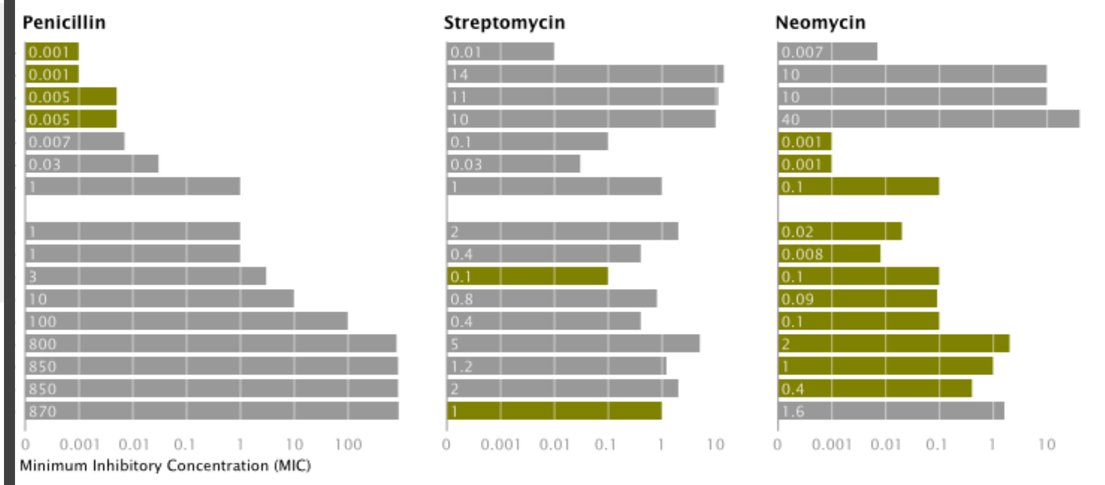


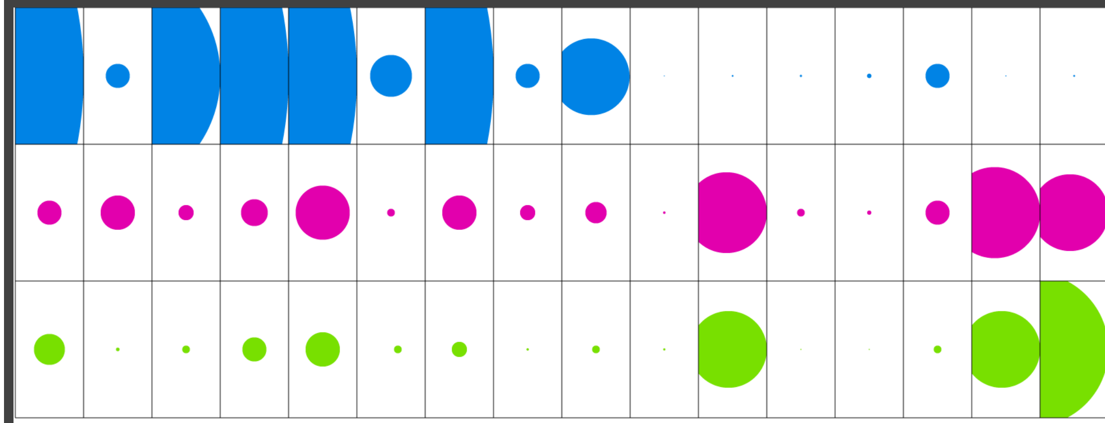
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.

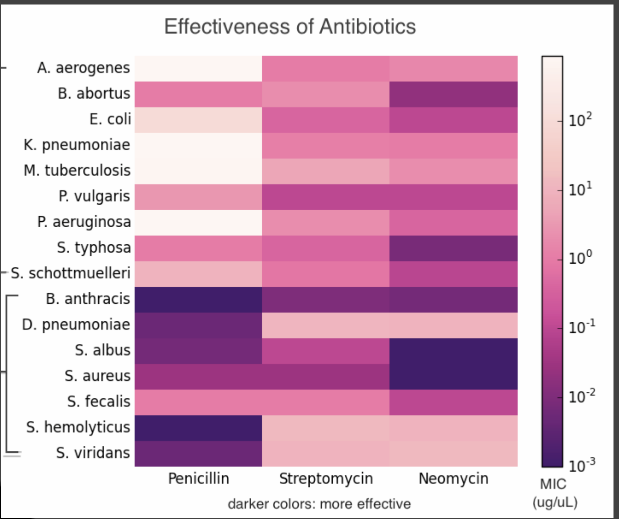
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

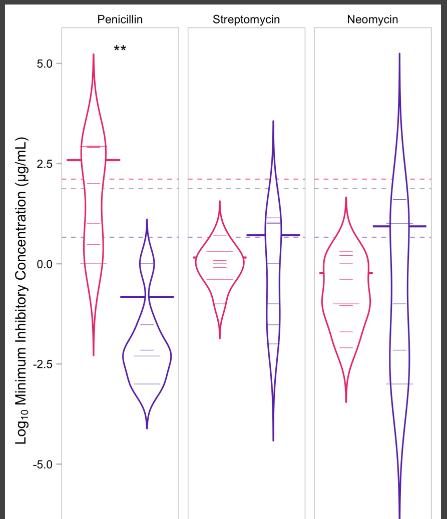
Gram staining quickly identifies bacteria as **Gram-negative** or **Gram-positive**, which can be used to find a more efficient antibiotic and dosage.

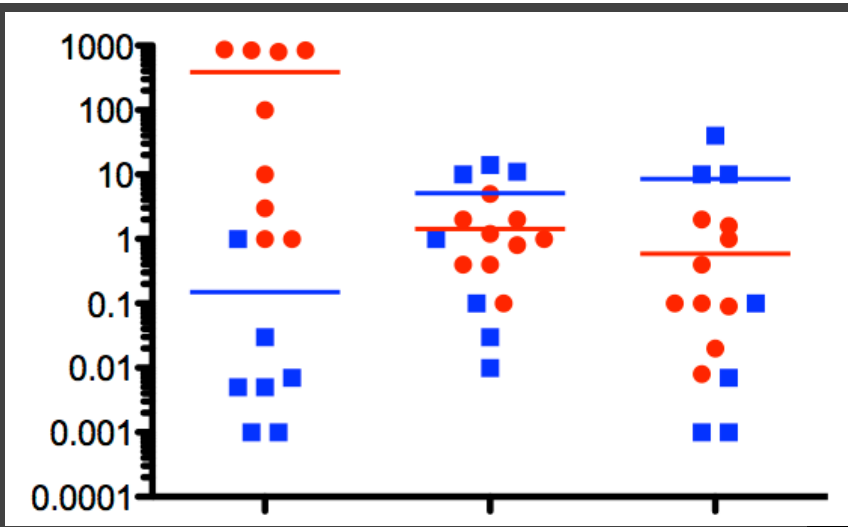
Penicilin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.



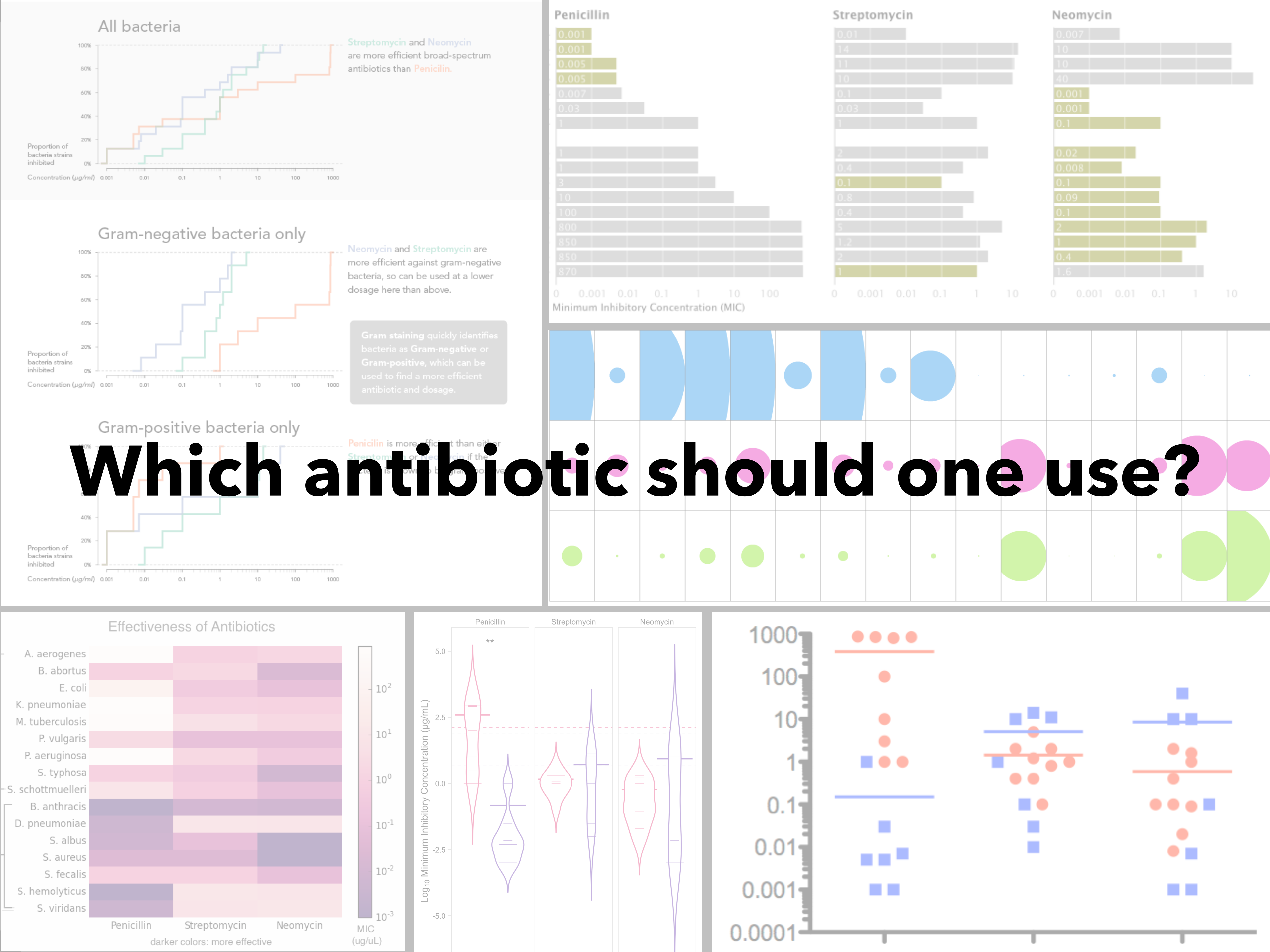




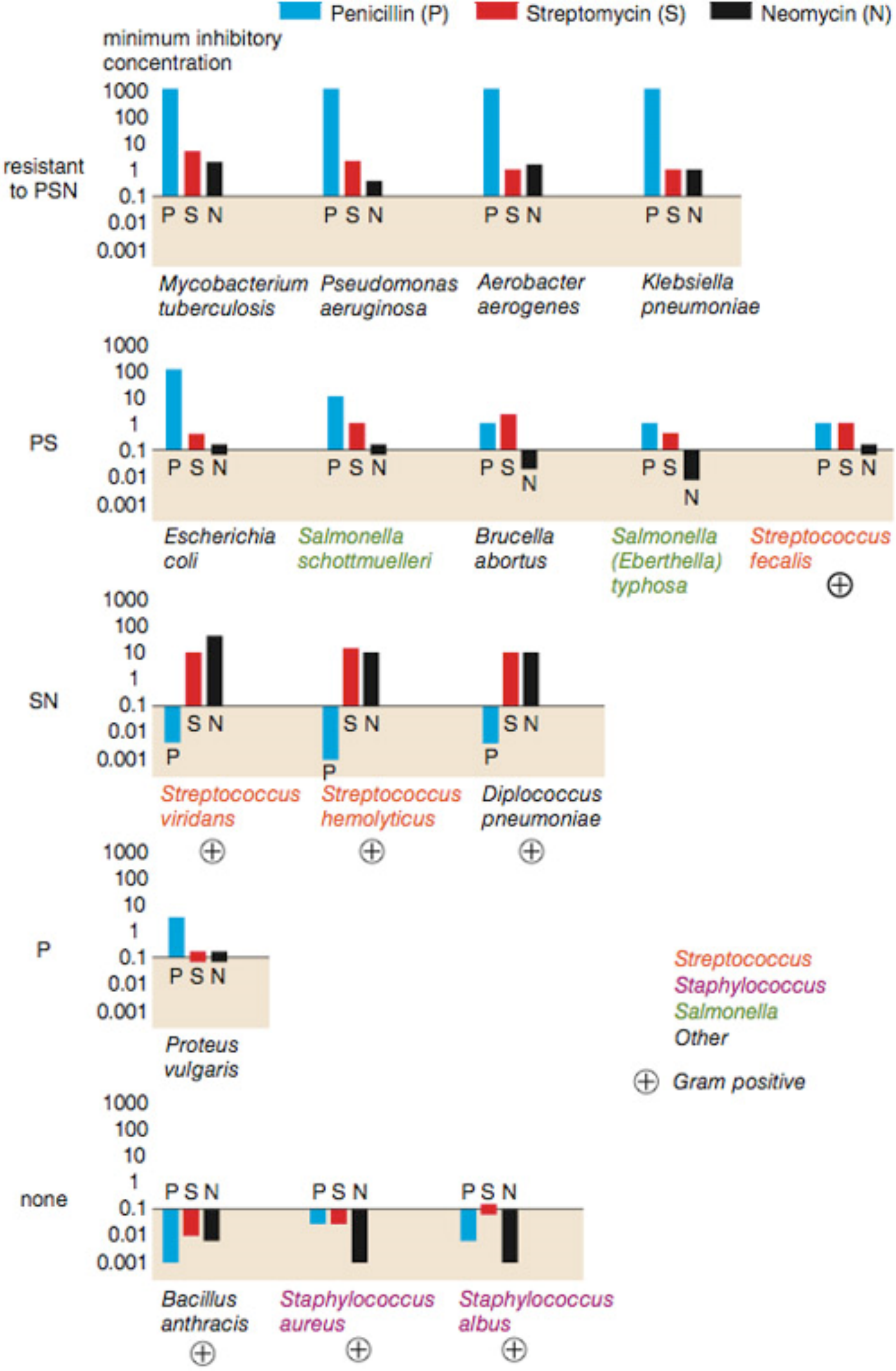




Which antibiotic should one use?

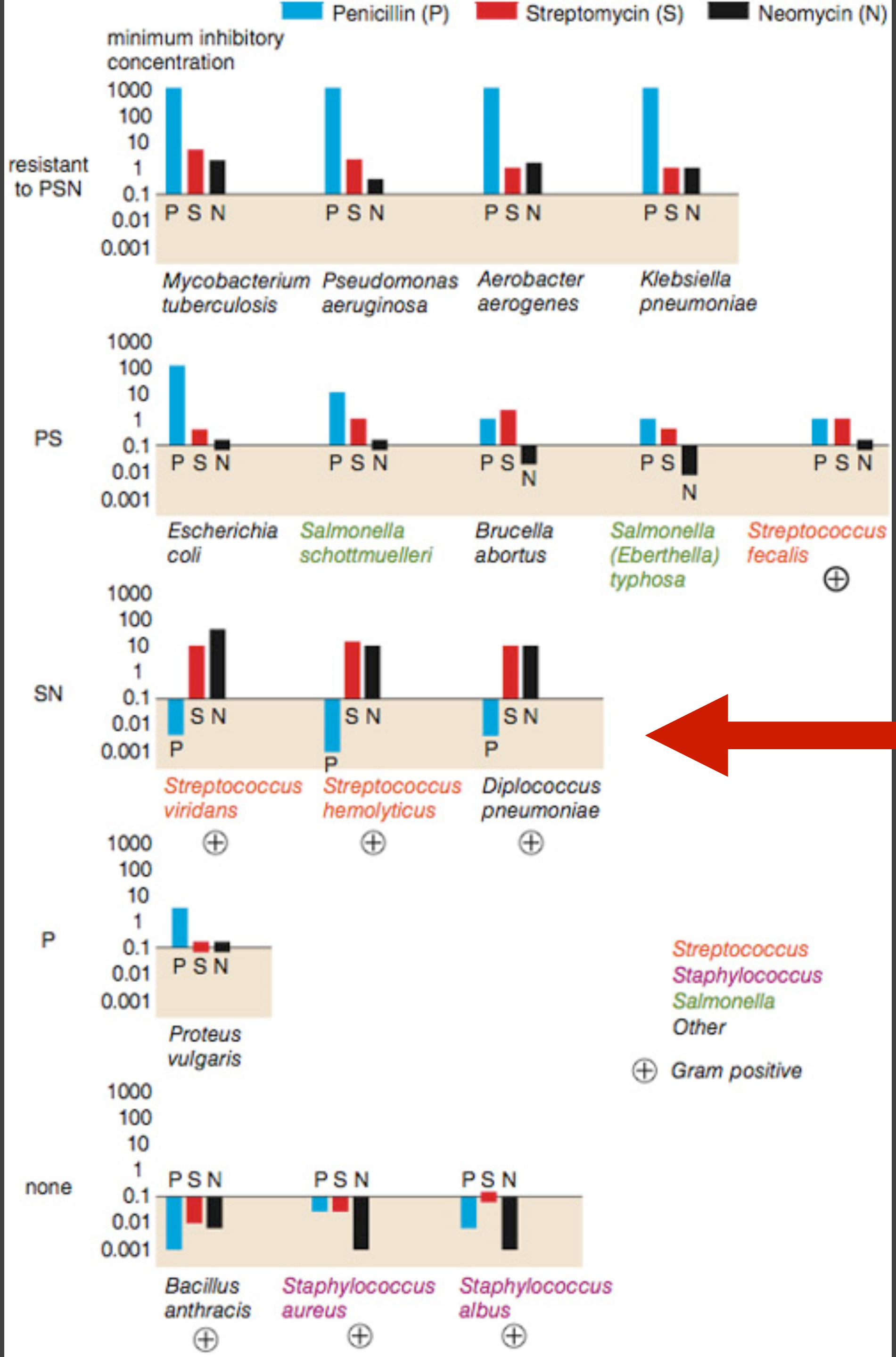


Do the bacteria group by antibiotic resistance?



Do the bacteria group by antibiotic resistance?

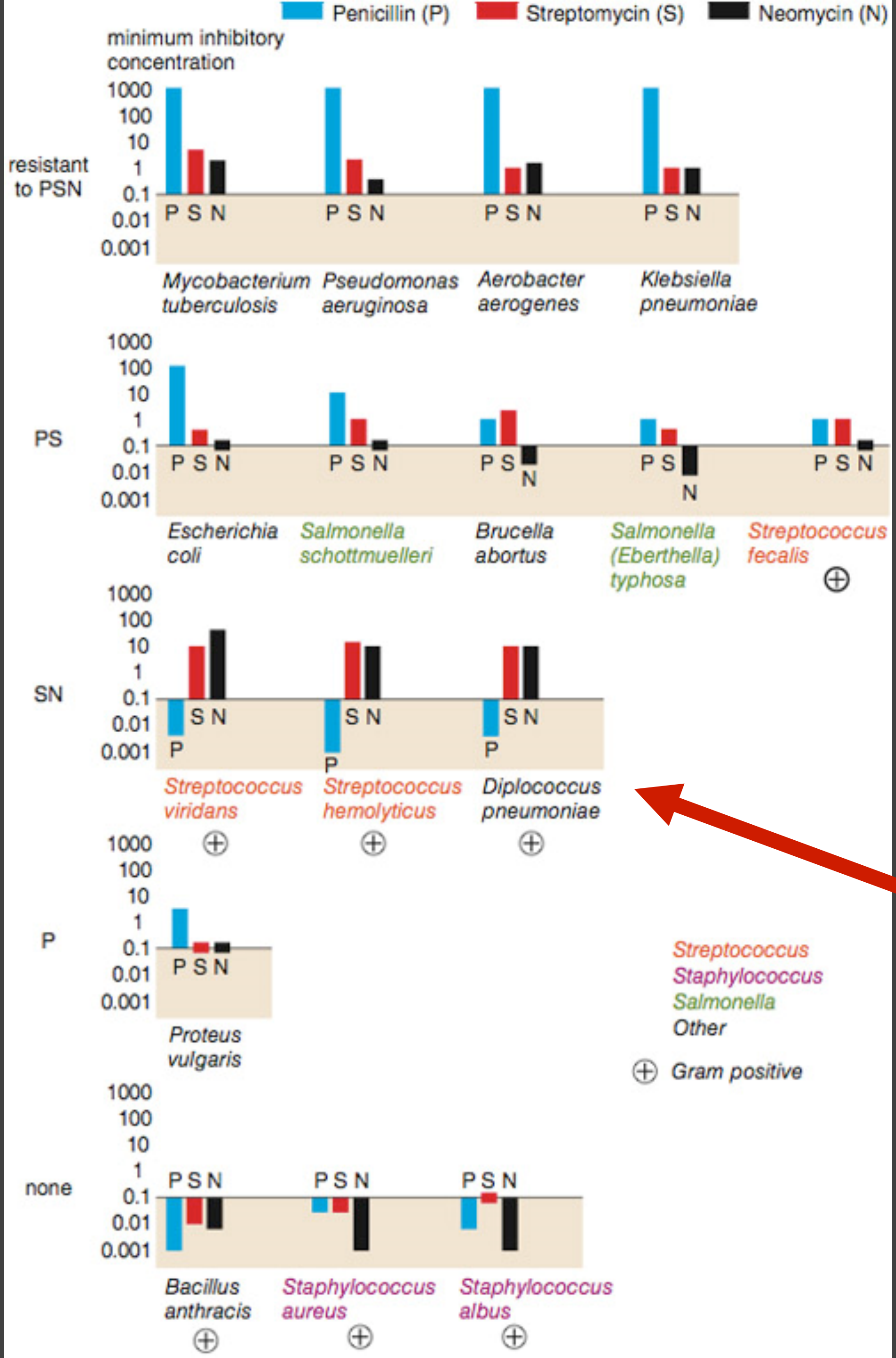
Wainer & Lysen
American Scientist, 2009



Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009

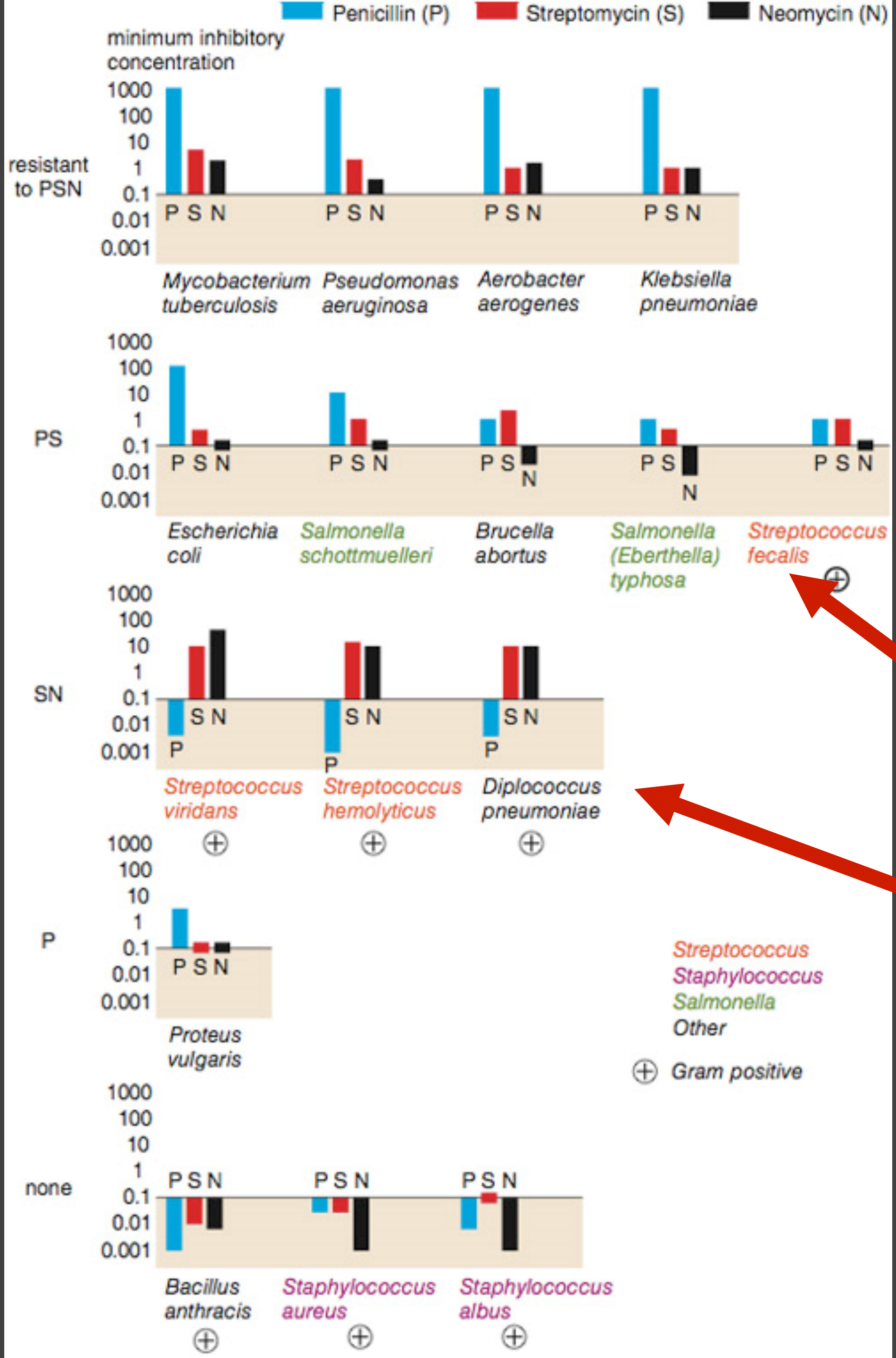
Do the bacteria group by antibiotic resistance?



Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

Do the bacteria group by antibiotic resistance?

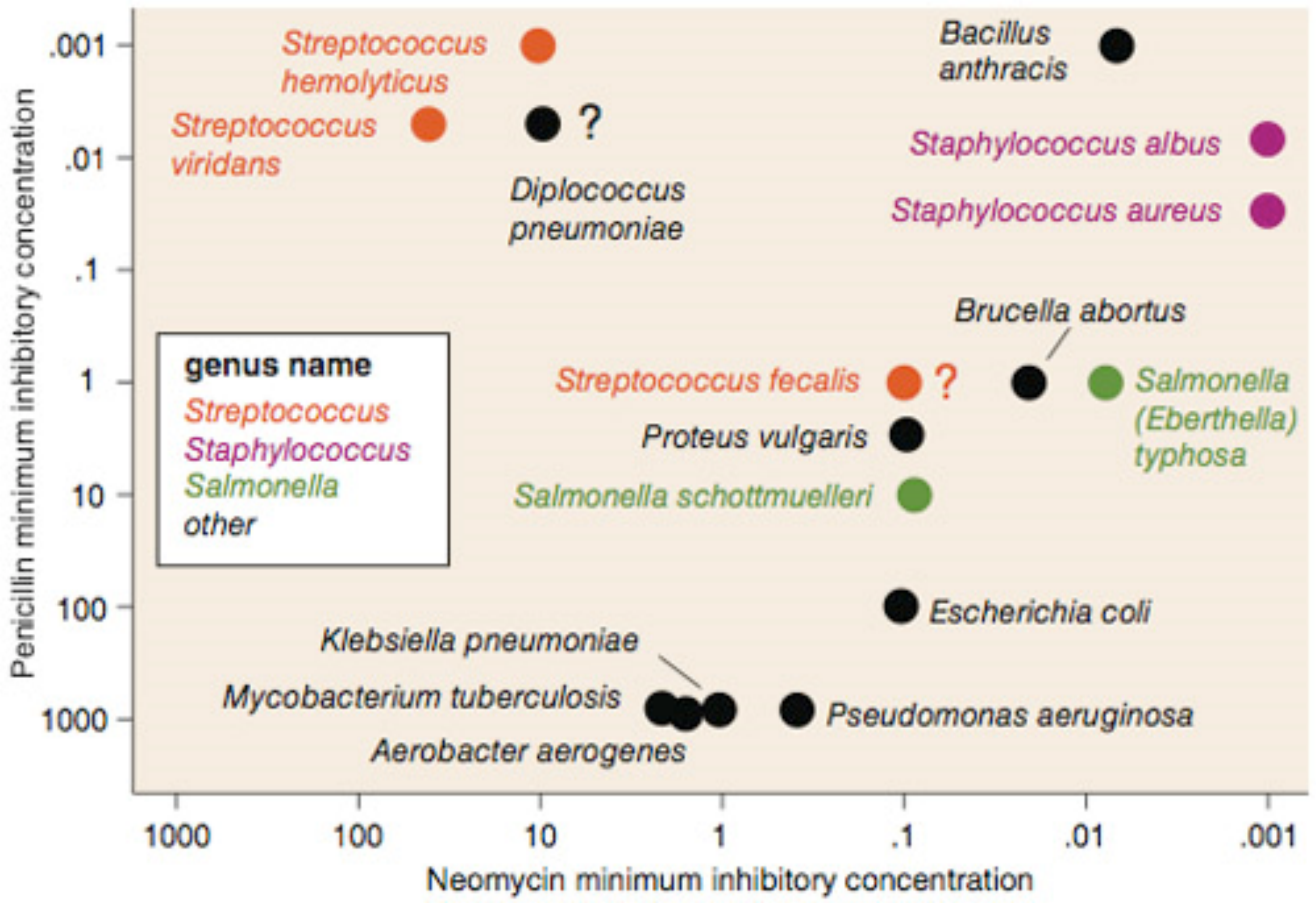


Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

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



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

Wainer & Lysen
American Scientist, 2009

Lesson: Iterative Exploration

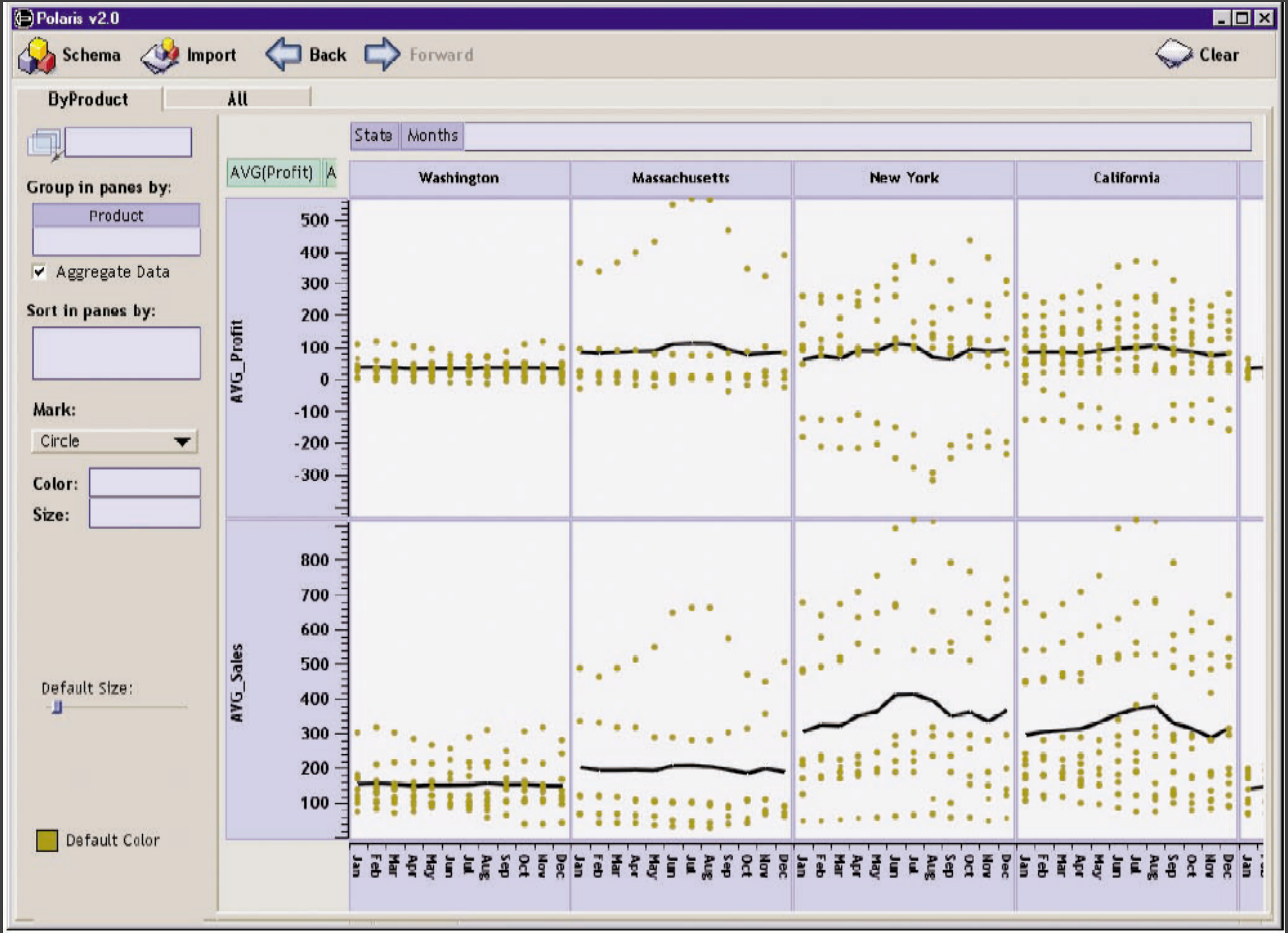
Exploratory Process
1 Construct graphics to address questions
2 Inspect "answer" and assess new questions
3 Repeat…

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

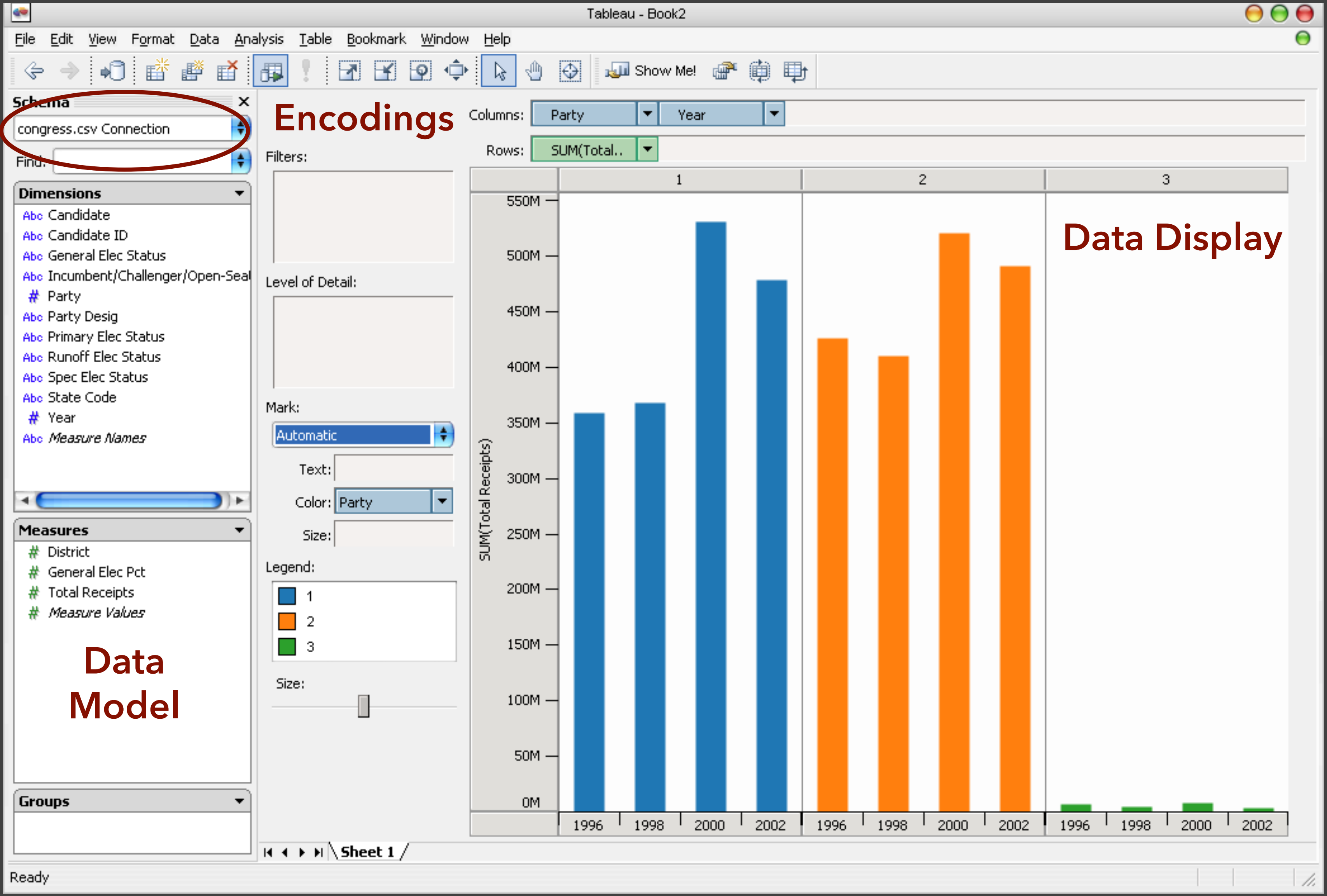
Show data variation, not design variation [Tufte]

Tableau / Polaris

Polaris [Stolte et al.]



Tableau



Columns: Party, Year

Rows: SUM(Total..

Mark: Automatic

Color: Party

Encodings

Data Display

Data Model

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

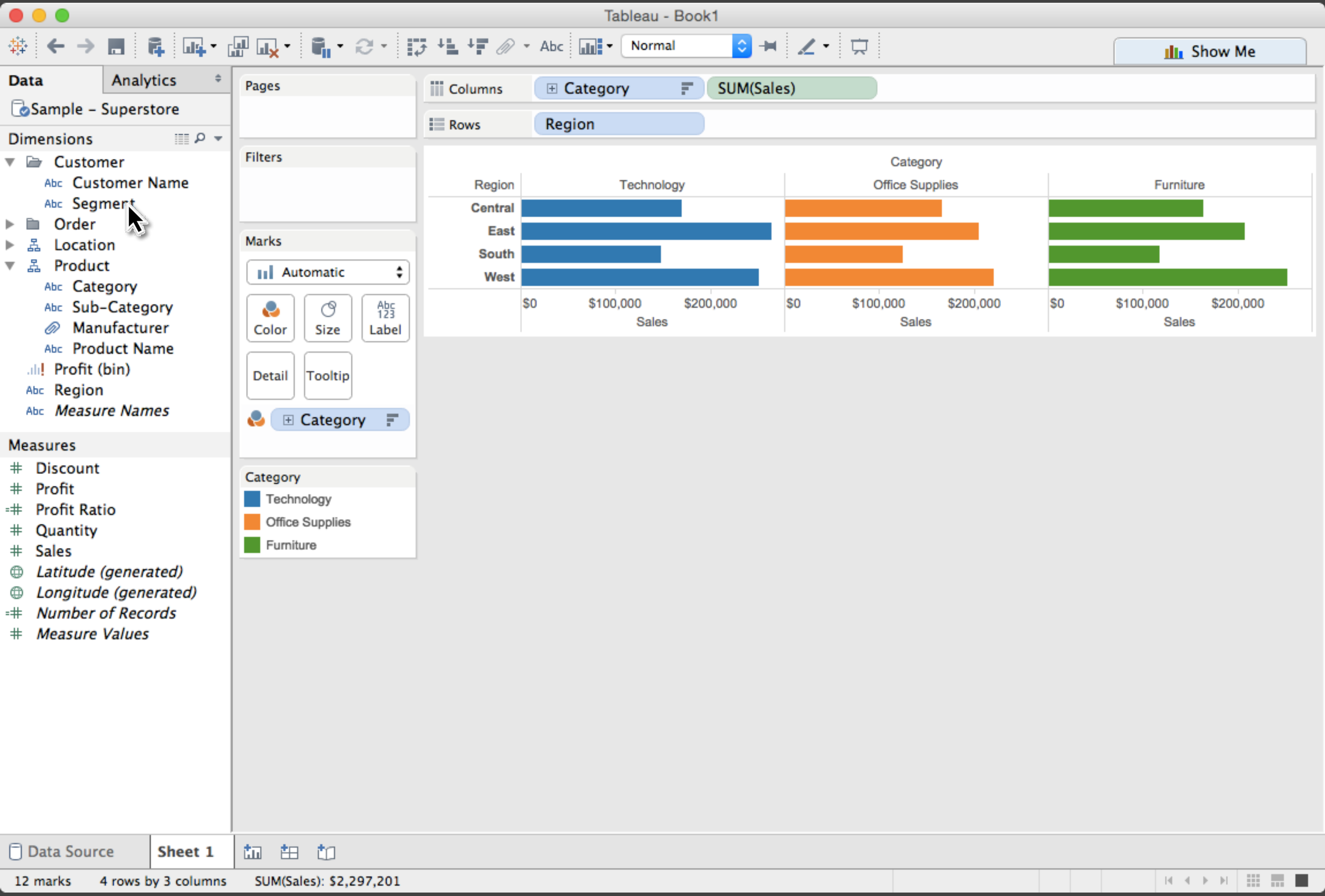
Can also suggest encodings upon request

Tableau Demo

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 (/)



Columns: Category, SUM(Sales)

Rows: Region

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

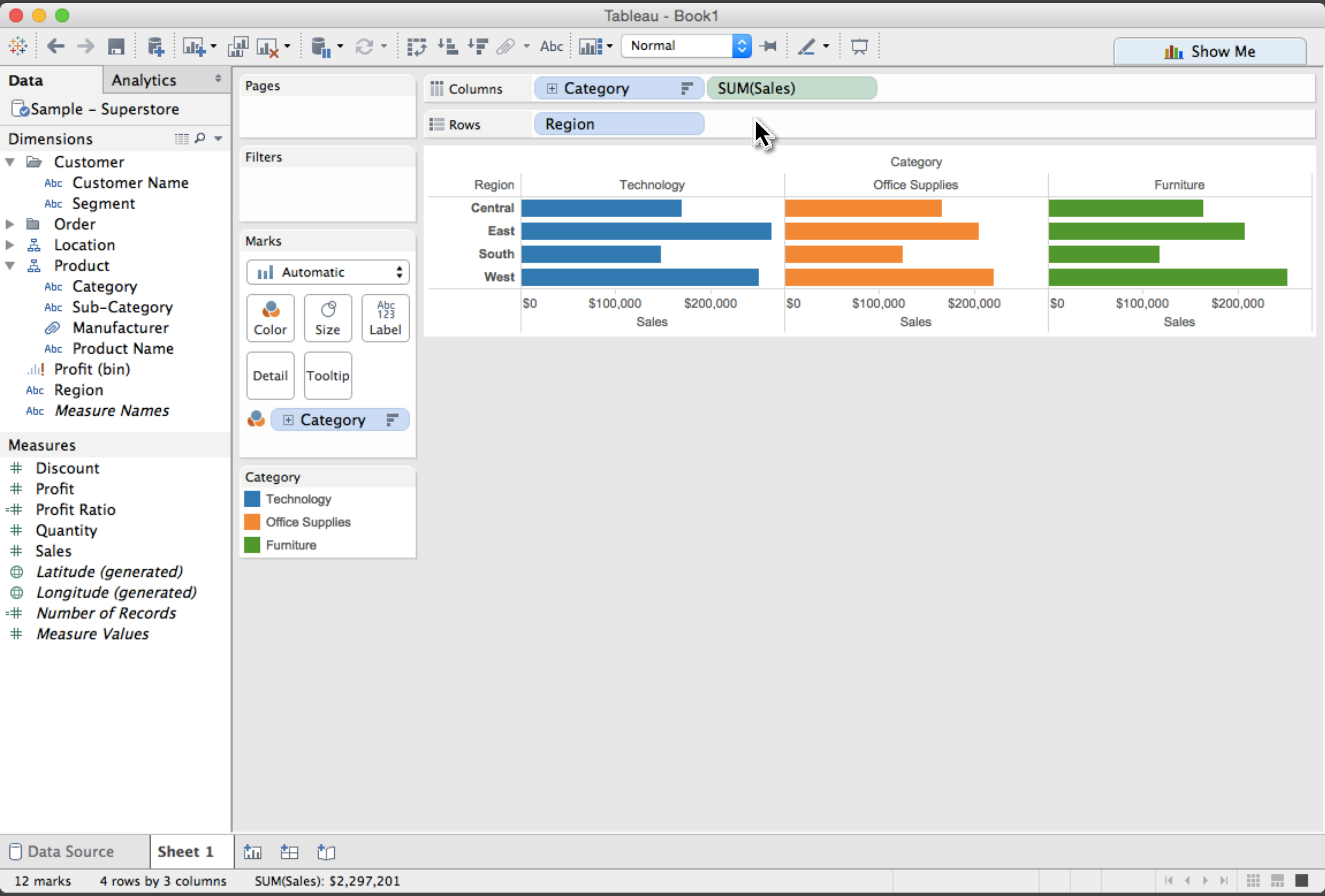


Tableau - Book1

Columns: Category, SUM(Sales)

Rows: Region

Marks: Automatic

Category
- 🟦 Technology
- 🟧 Office Supplies
- 🟩 Furniture

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

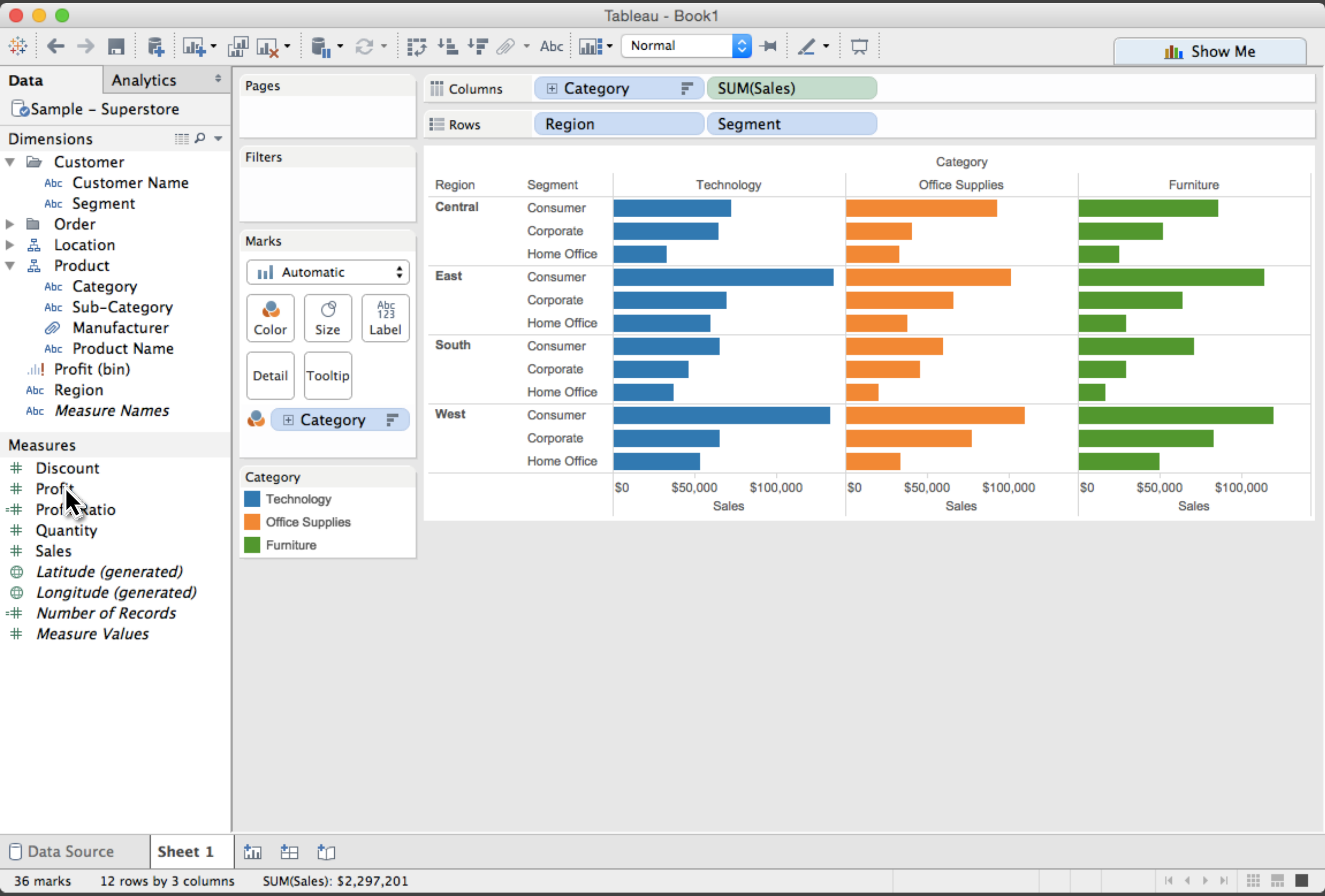


Tableau - Book1

Columns: Category, SUM(Sales)

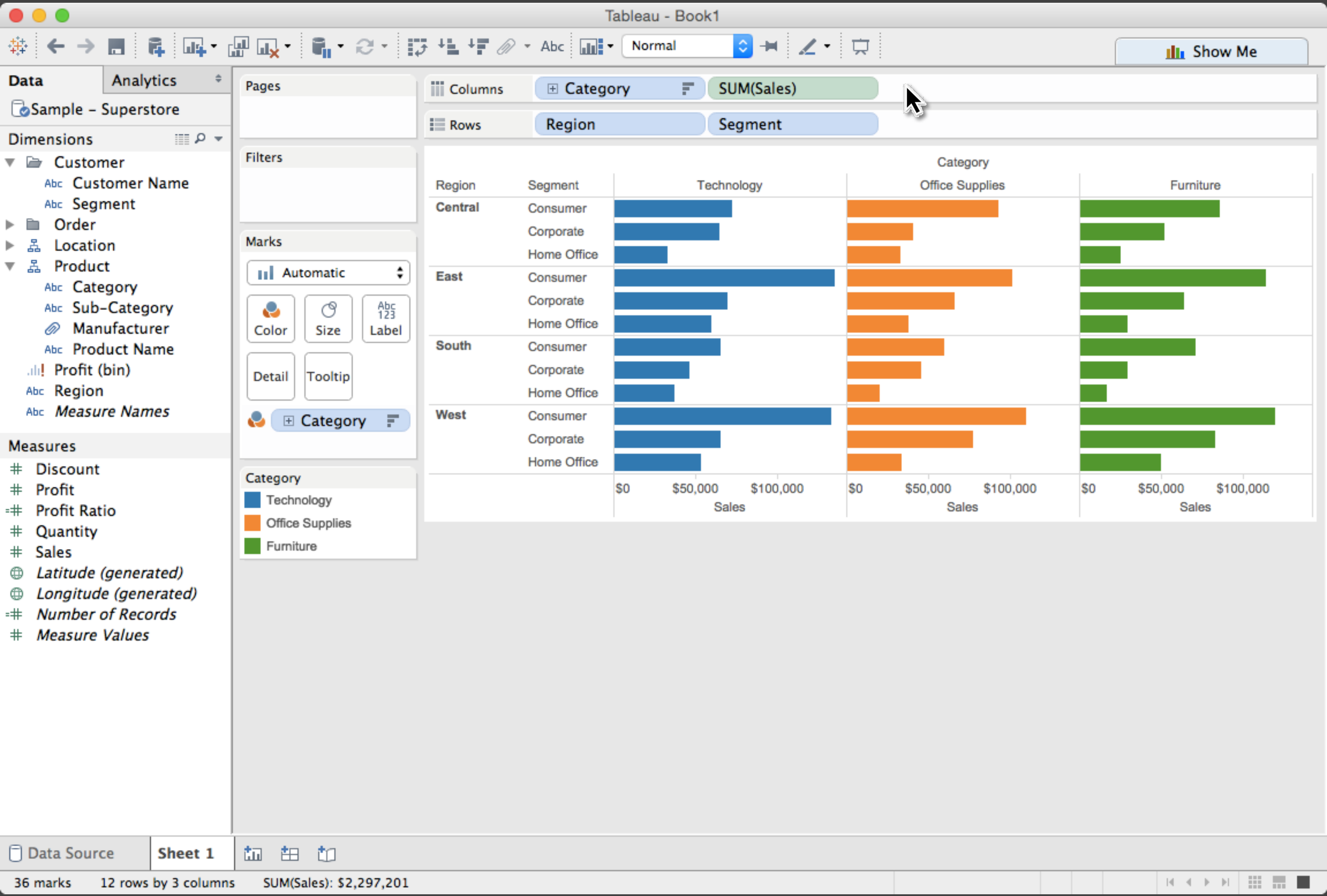
Rows: Region, Segment

Marks: Automatic

Color: Category

Category
- 🟦 Technology
- 🟧 Office Supplies
- 🟩 Furniture

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



Columns: Category, SUM(Sales)

Rows: Region, Segment

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

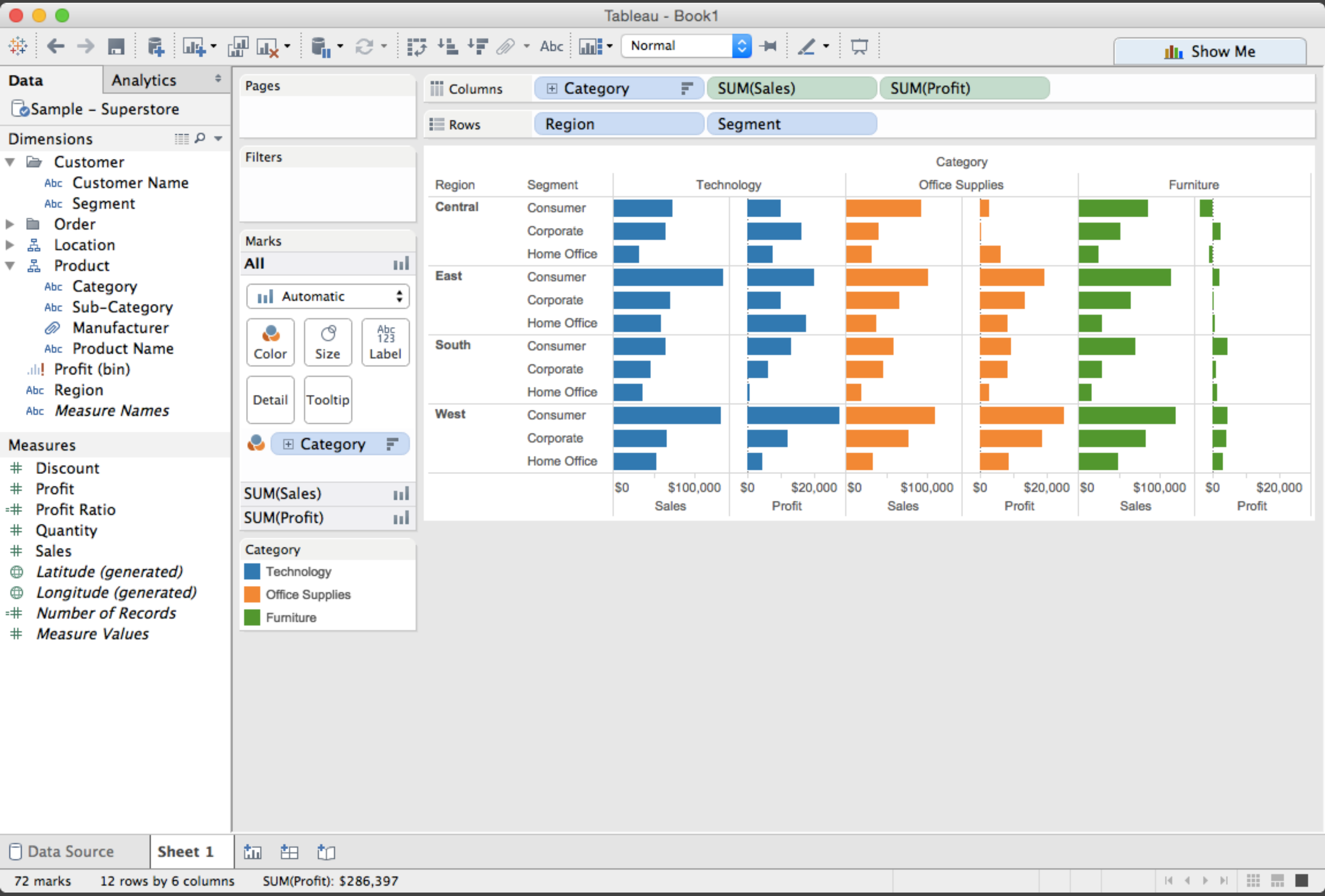


Tableau - Book1

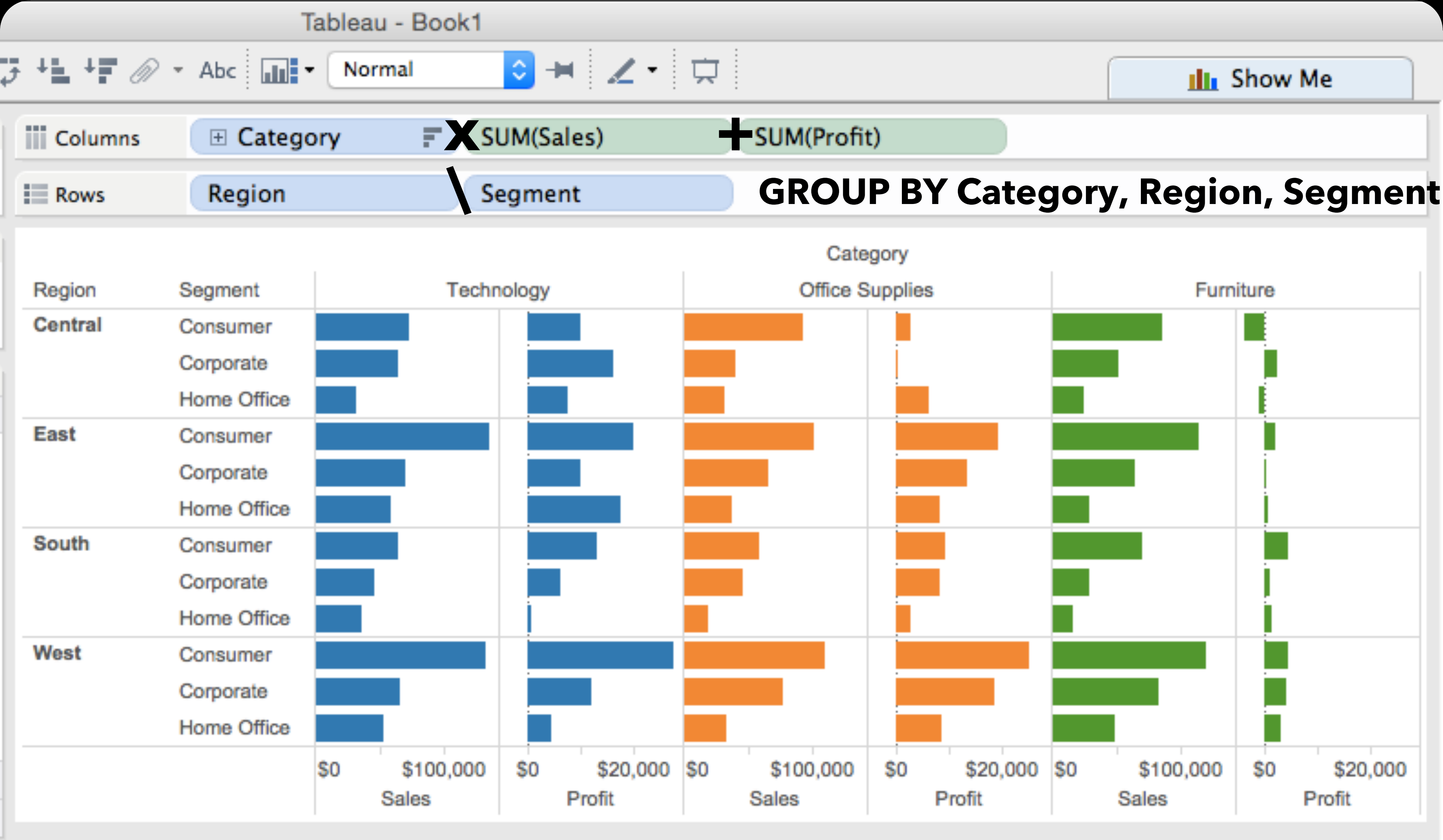
Columns: Category, SUM(Sales), SUM(Profit)

Rows: Region, Segment

Marks: All — Automatic

Color: Category

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



Columns: Category x 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

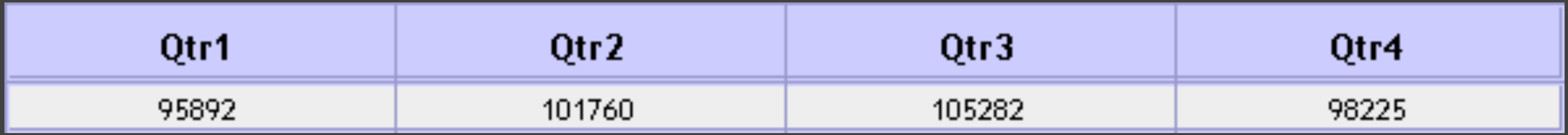
In Tableau, users make statements via drag-and-drop

- 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.

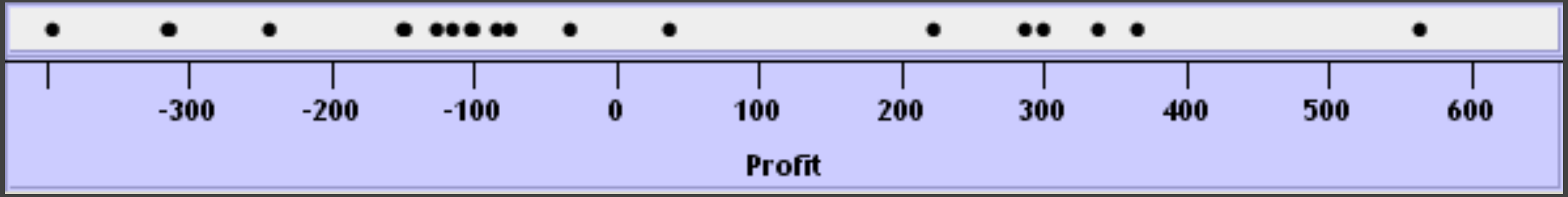
Quarter = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} ->



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])} ->



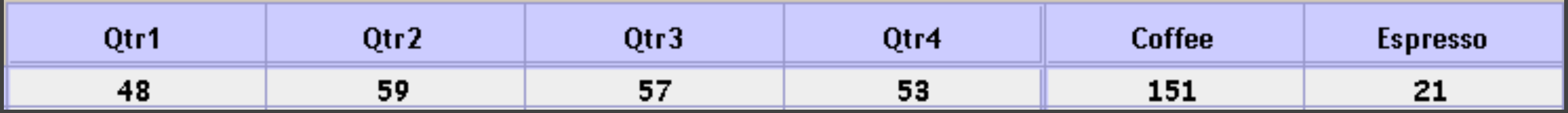
Concatenation (+) Operator

Ordered union of set interpretations

Quarter + Product Type

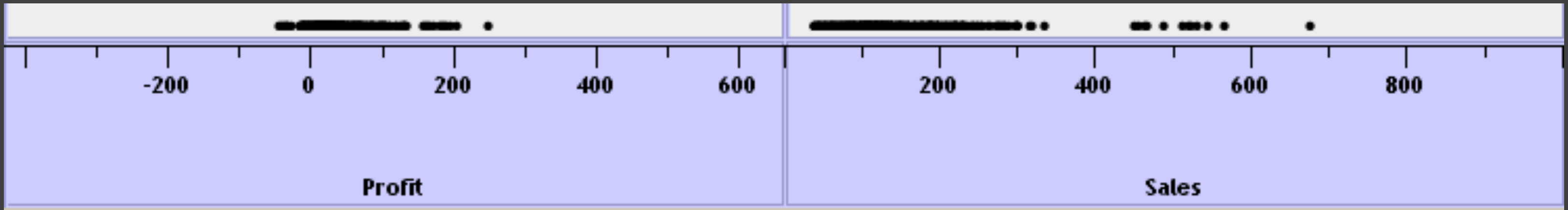
= {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} + {(Coffee), (Espresso)}

= {(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])}

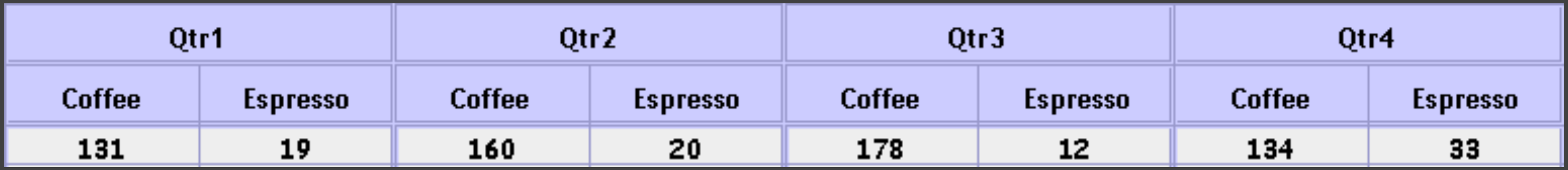


Cross (x) Operator

Cross-product of set interpretations

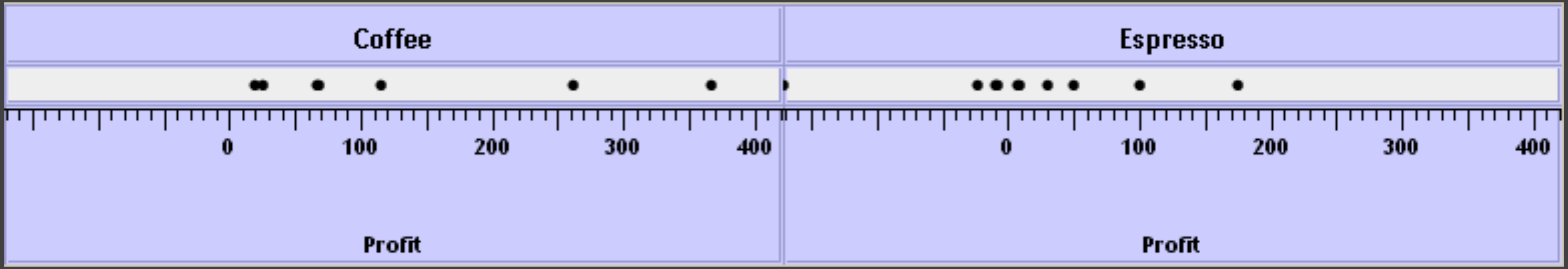
Quarter x Product Type =

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



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

Product Type x Profit =



Nest (/) Operator

Cross-product filtered by existing records

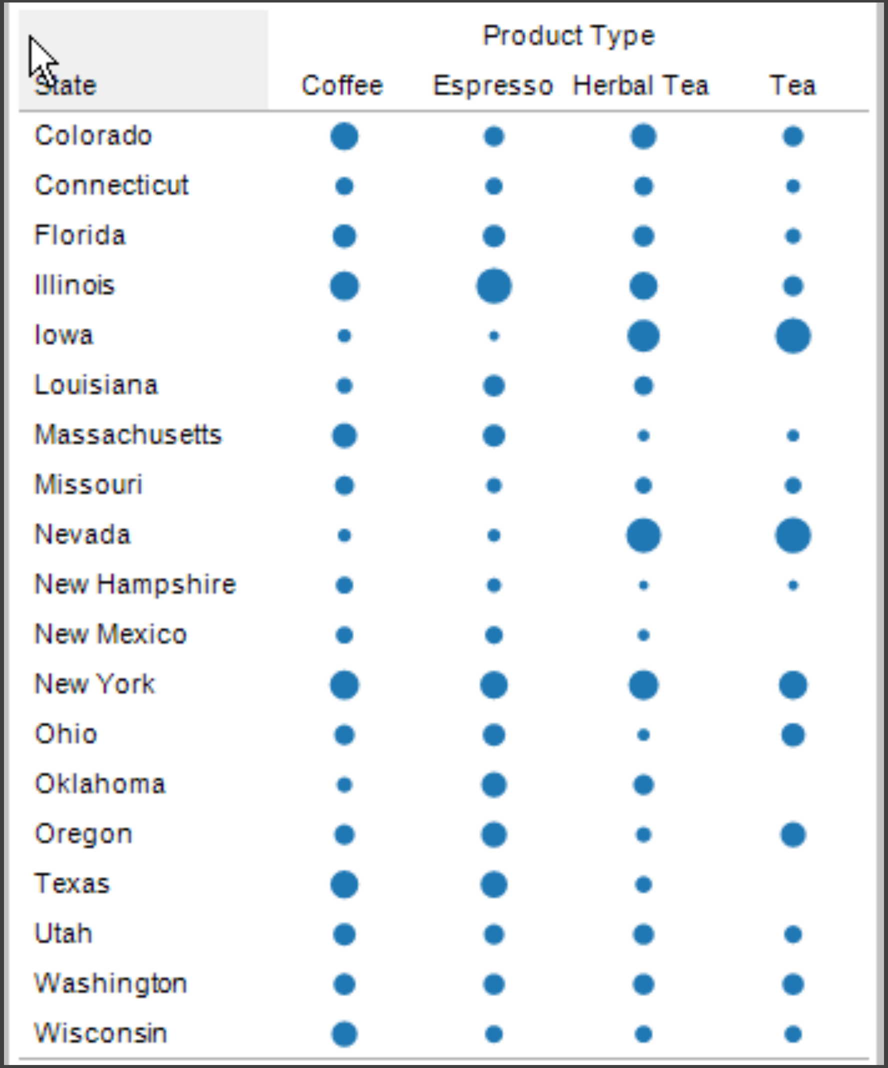
Quarter x Month ->

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

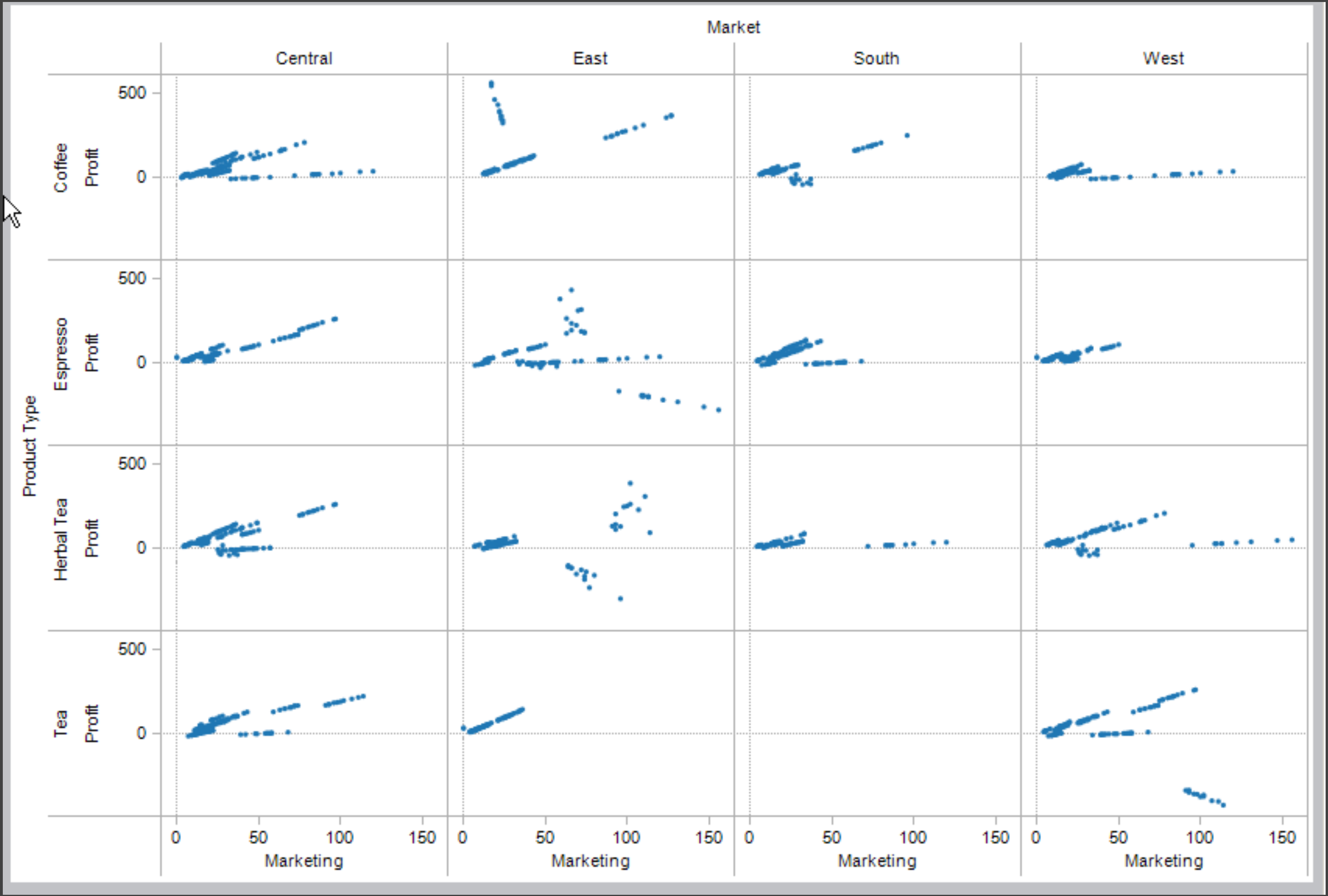
Quarter / Month ->

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

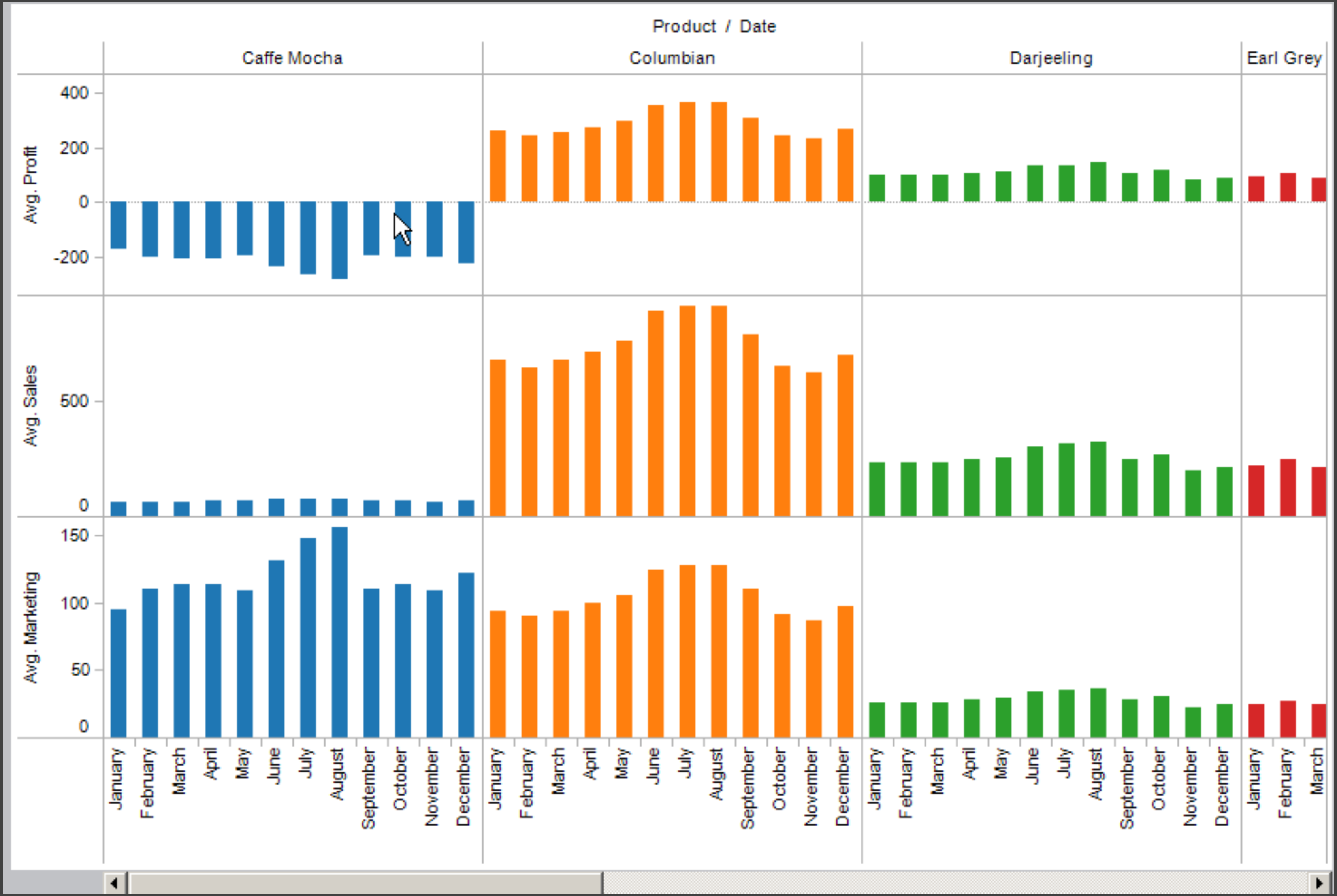
Ordinal-Ordinal



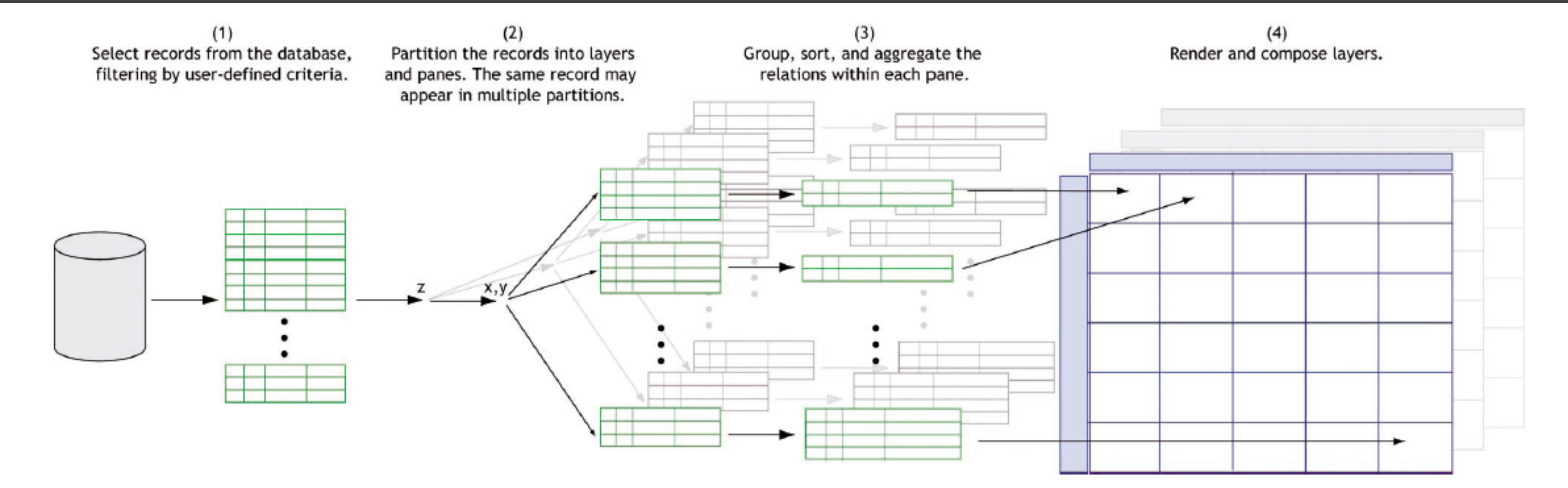
Quantitative-Quantitative



Ordinal-Quantitative



Querying the Database



(1)
Select records from the database, filtering by user-defined criteria.

(2)
Partition the records into layers and panes. The same record may appear in multiple partitions.

(3)
Group, sort, and aggregate the relations within each pane.

(4)
Render and compose layers.

Summary: Connecting Queries and Visualizations in Tableau

Tableau maintains a **joint representation** of analysis operations as both data queries and visualizations using a **table algebra**.

This allows Tableau to support a graphical user interface for expressing data queries.

This also enables Tableau to automatically map queries to visualizations and vice versa.

Common Data Transformations

Normalize	$y_i / \Sigma_i y_i$	
Log	$\log y$	
Power	$y^{1/k}$	
Box-Cox Transform	$(y^\lambda - 1) / \lambda$	if $\lambda \neq 0$
	$\log y$	if $\lambda = 0$
Binning	e.g., histograms	
Grouping	e.g., merge categories	

Often performed to aid comparison (% or scale difference) or better approx. normal distribution