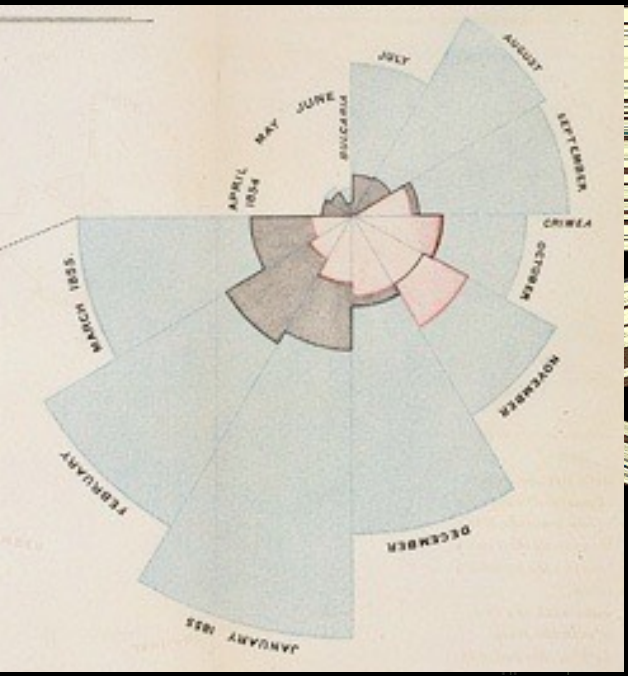
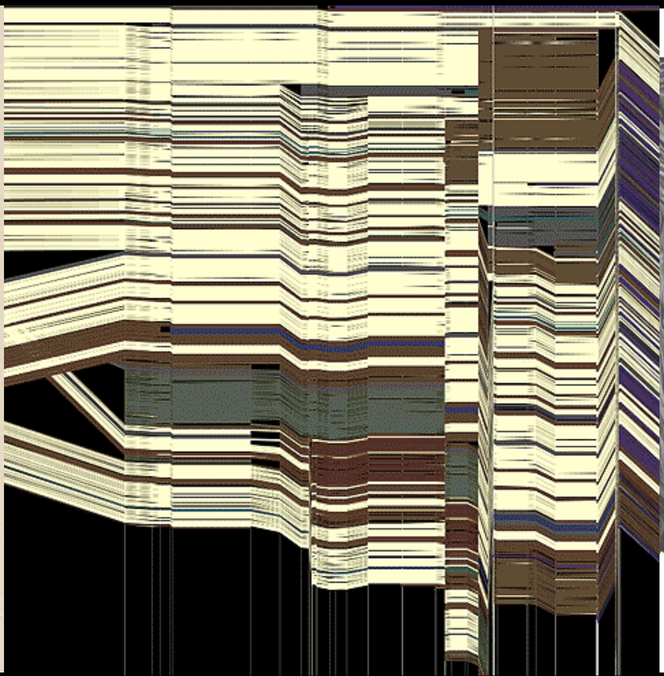
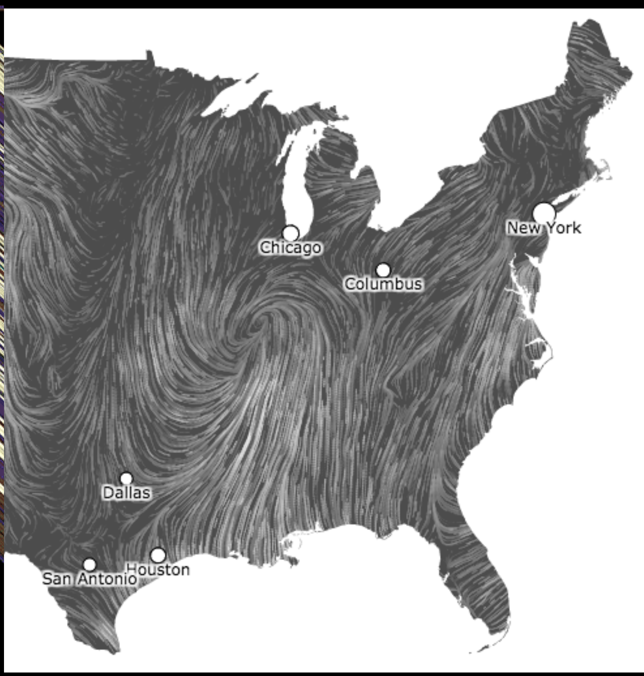
CSE 442 - Data Visualization

Exploratory Data Analysis



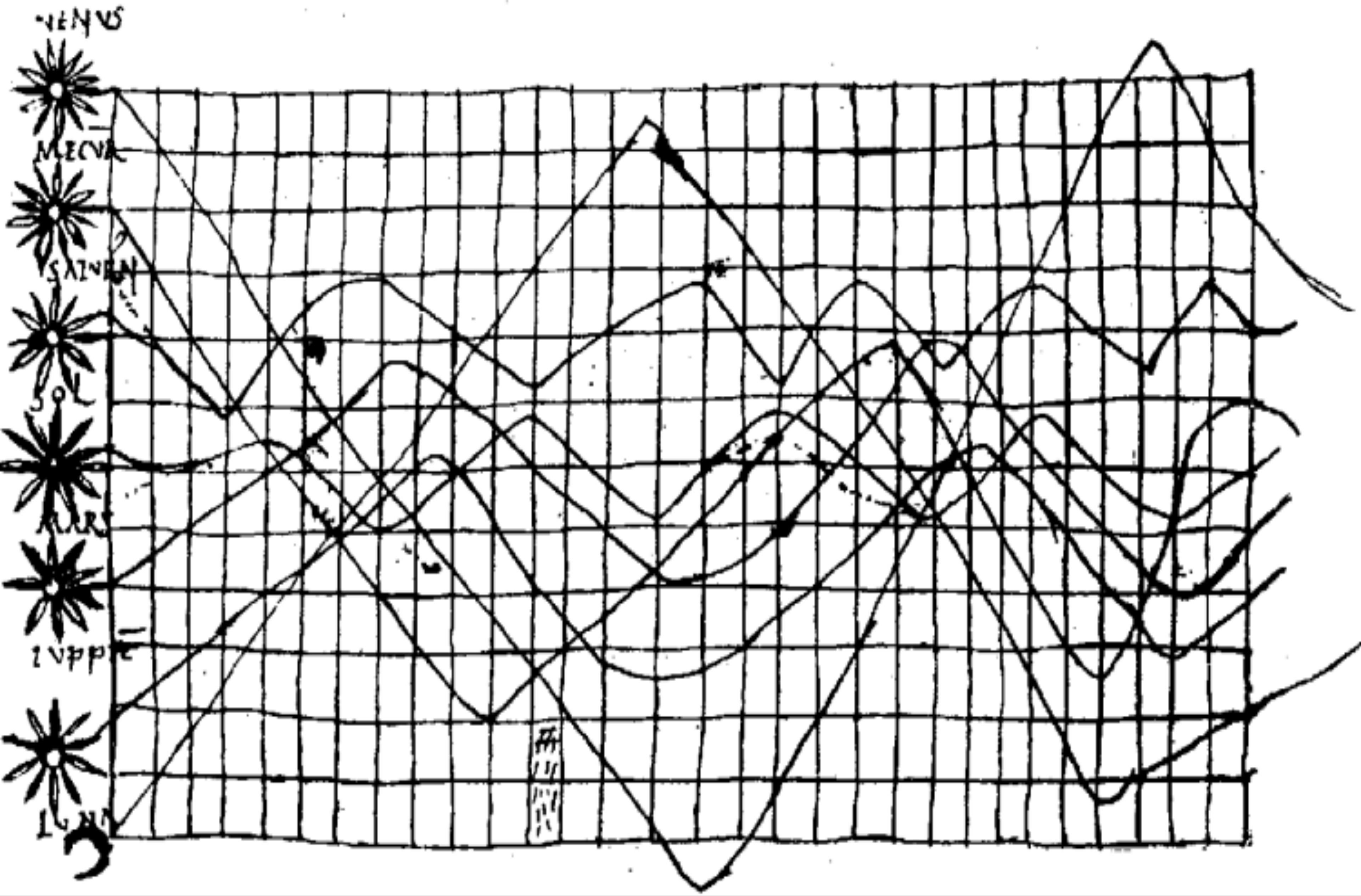
Jeffrey Heer University of Washington

What was the **first** data visualization?

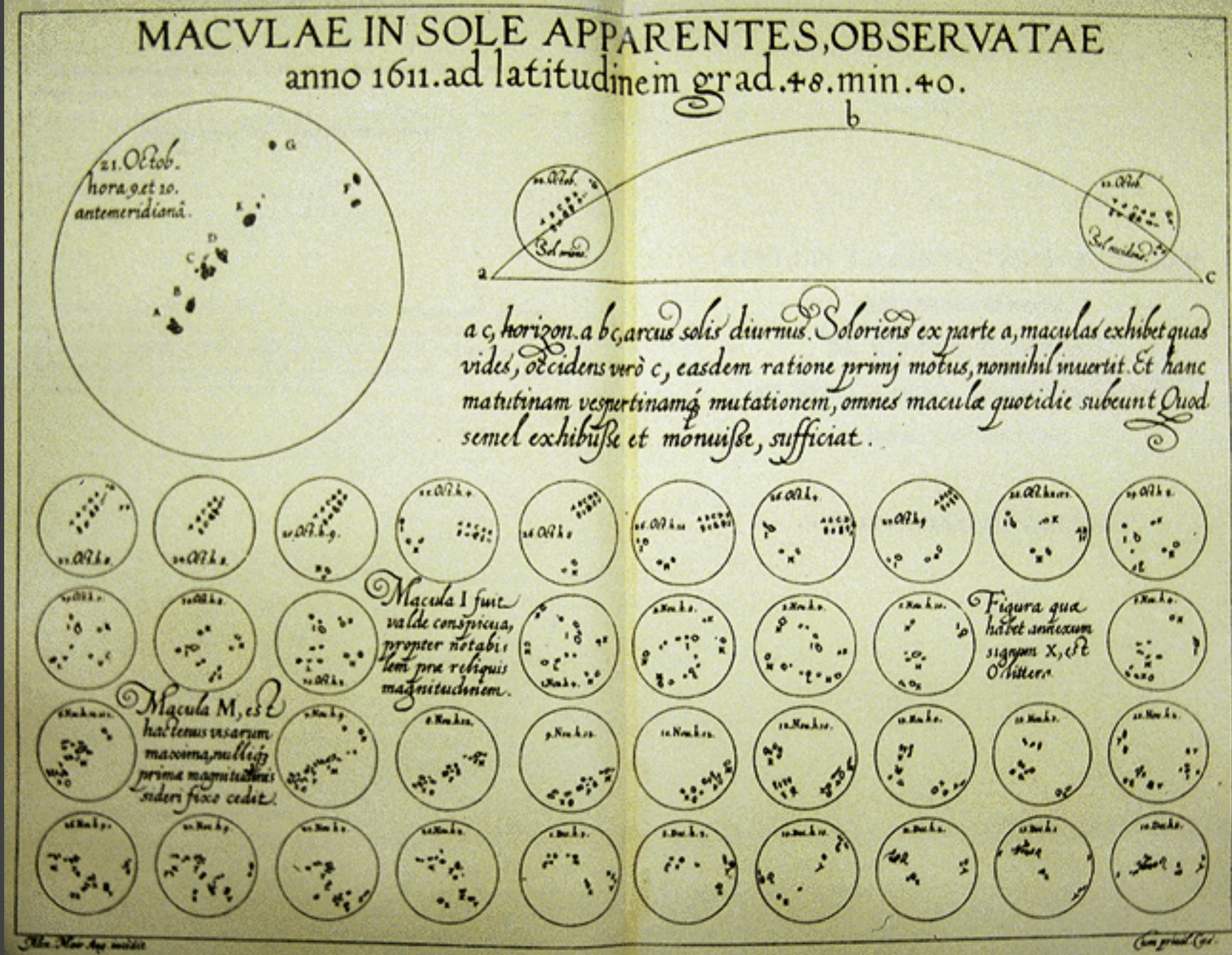
0 BC

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

0 BC

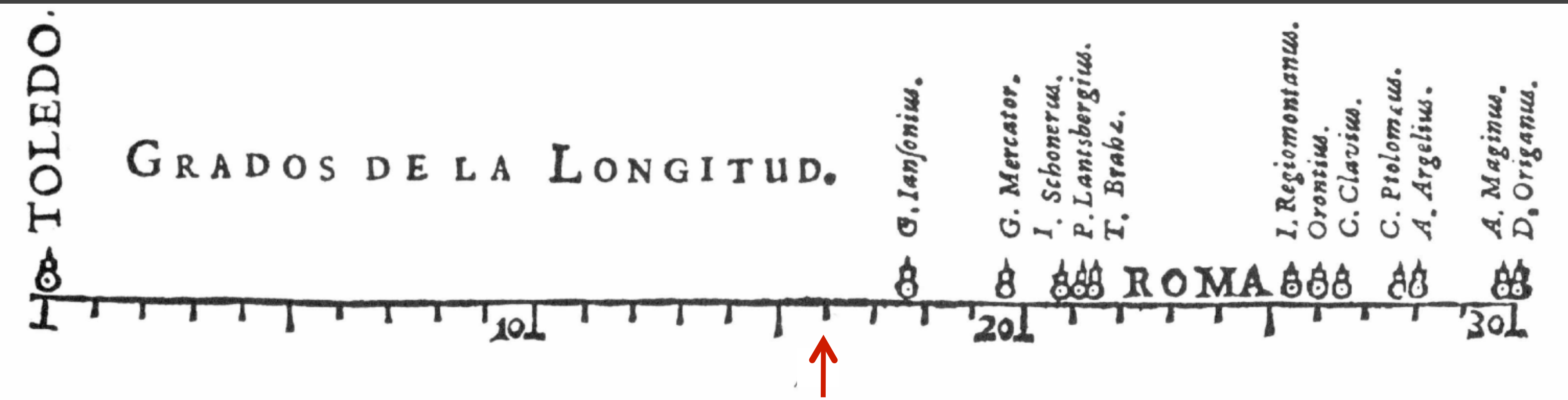


~950 AD Position of Sun, Moon and Planets

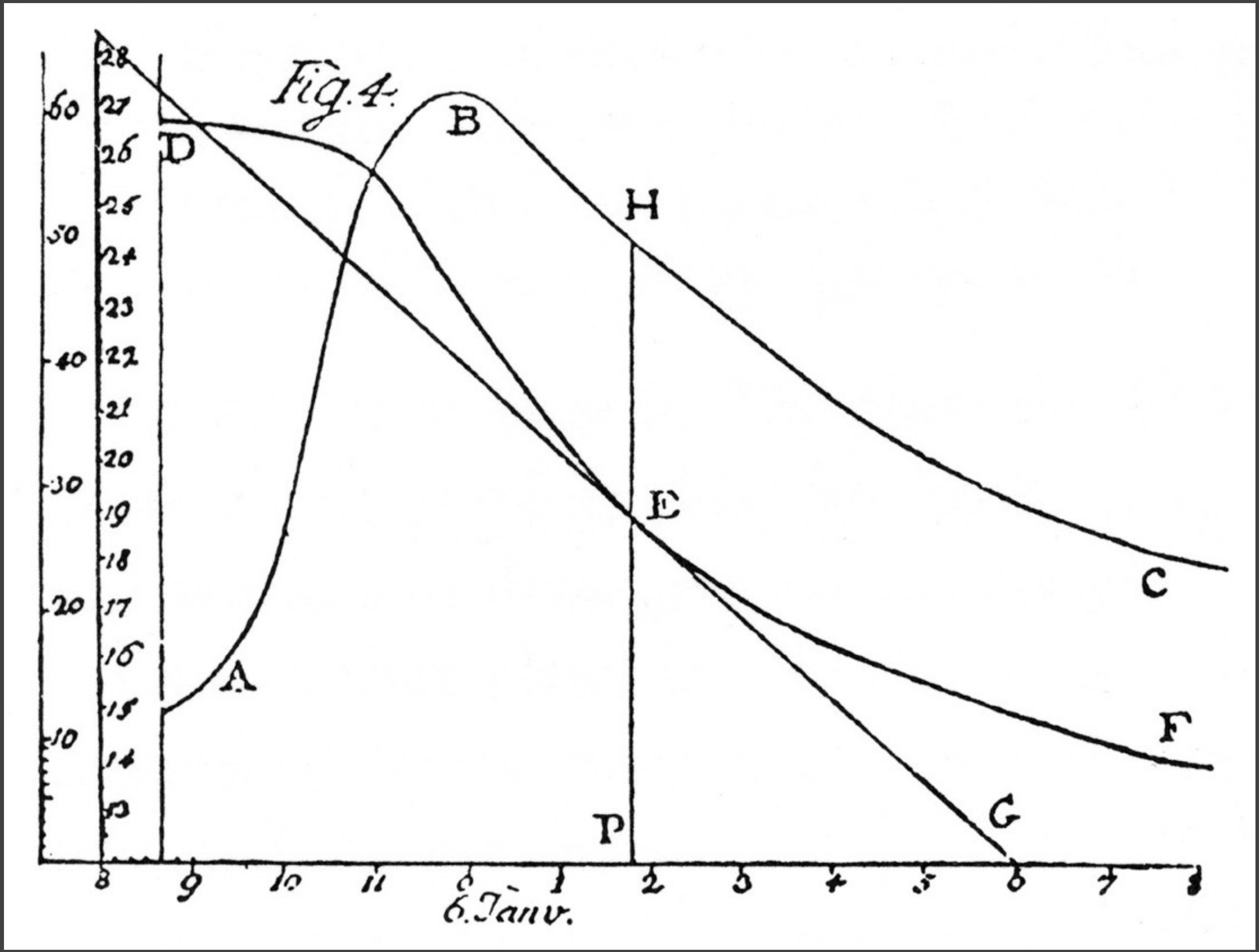


0 BC

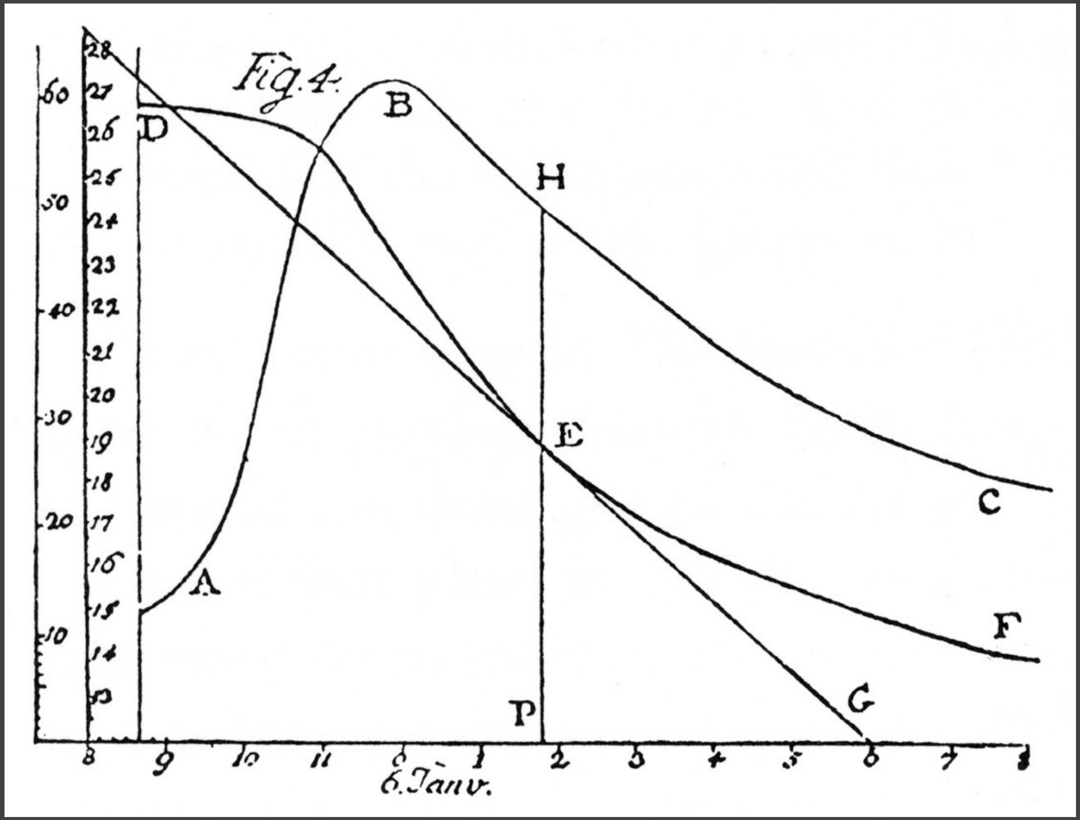
Sunspots over time, Scheiner 1626

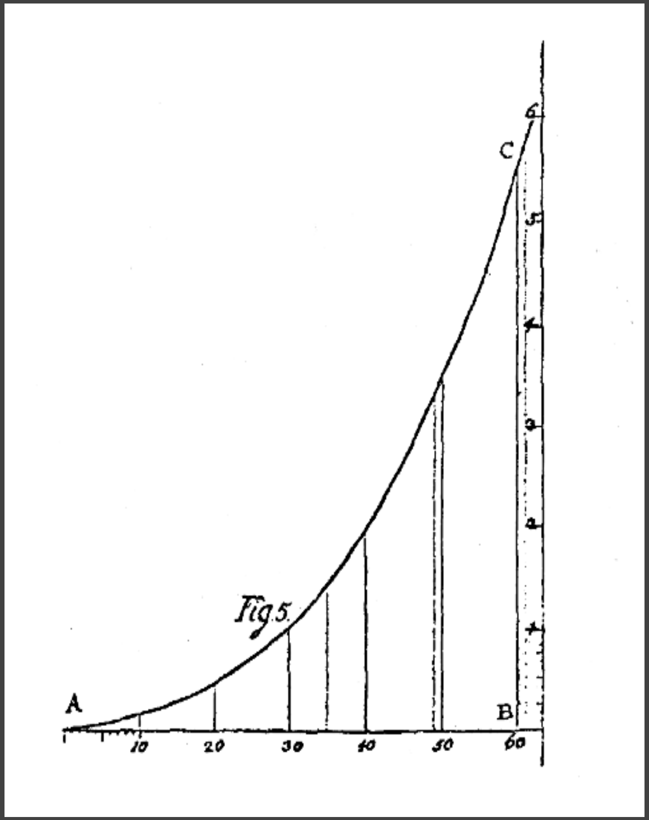


Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765

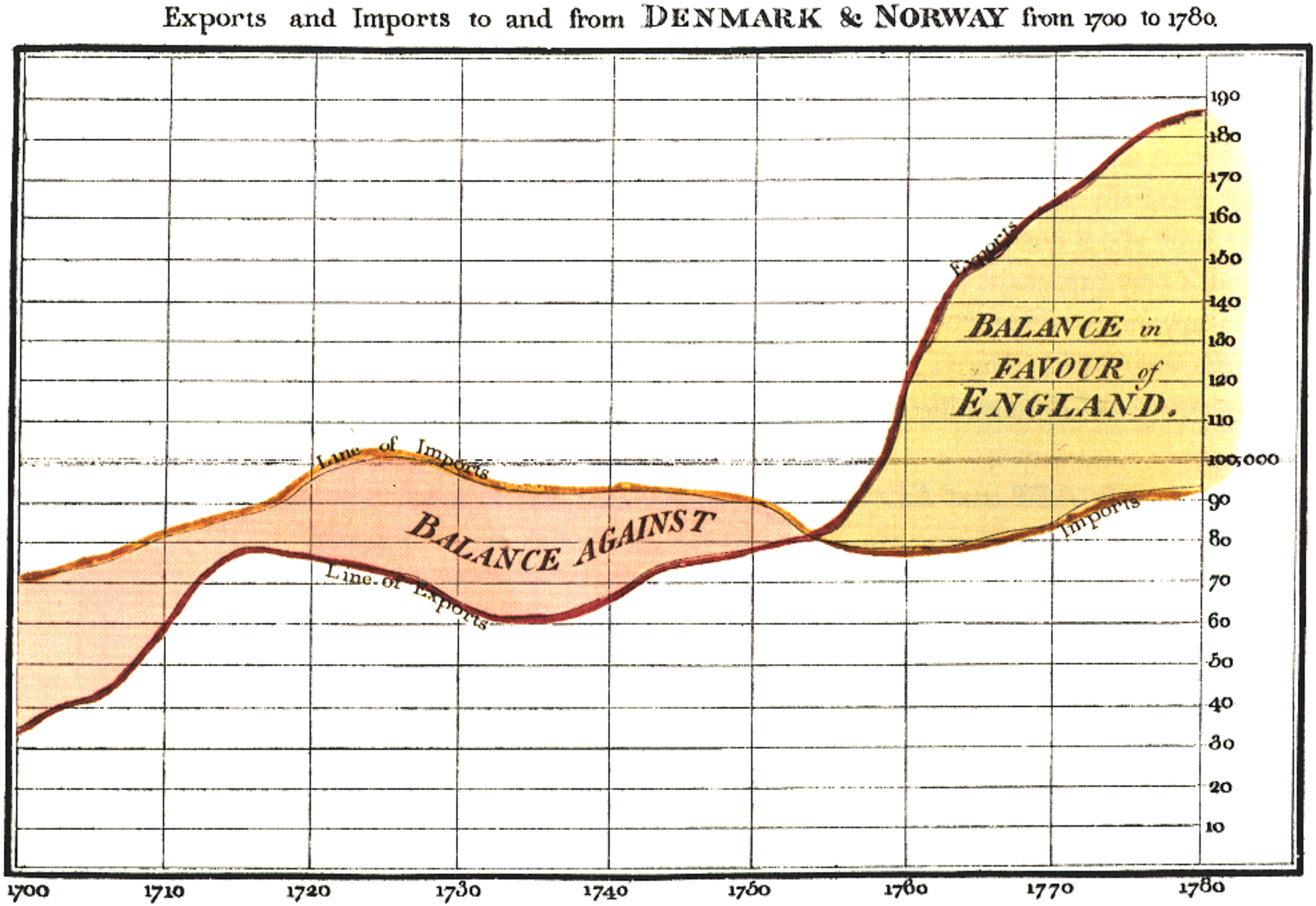




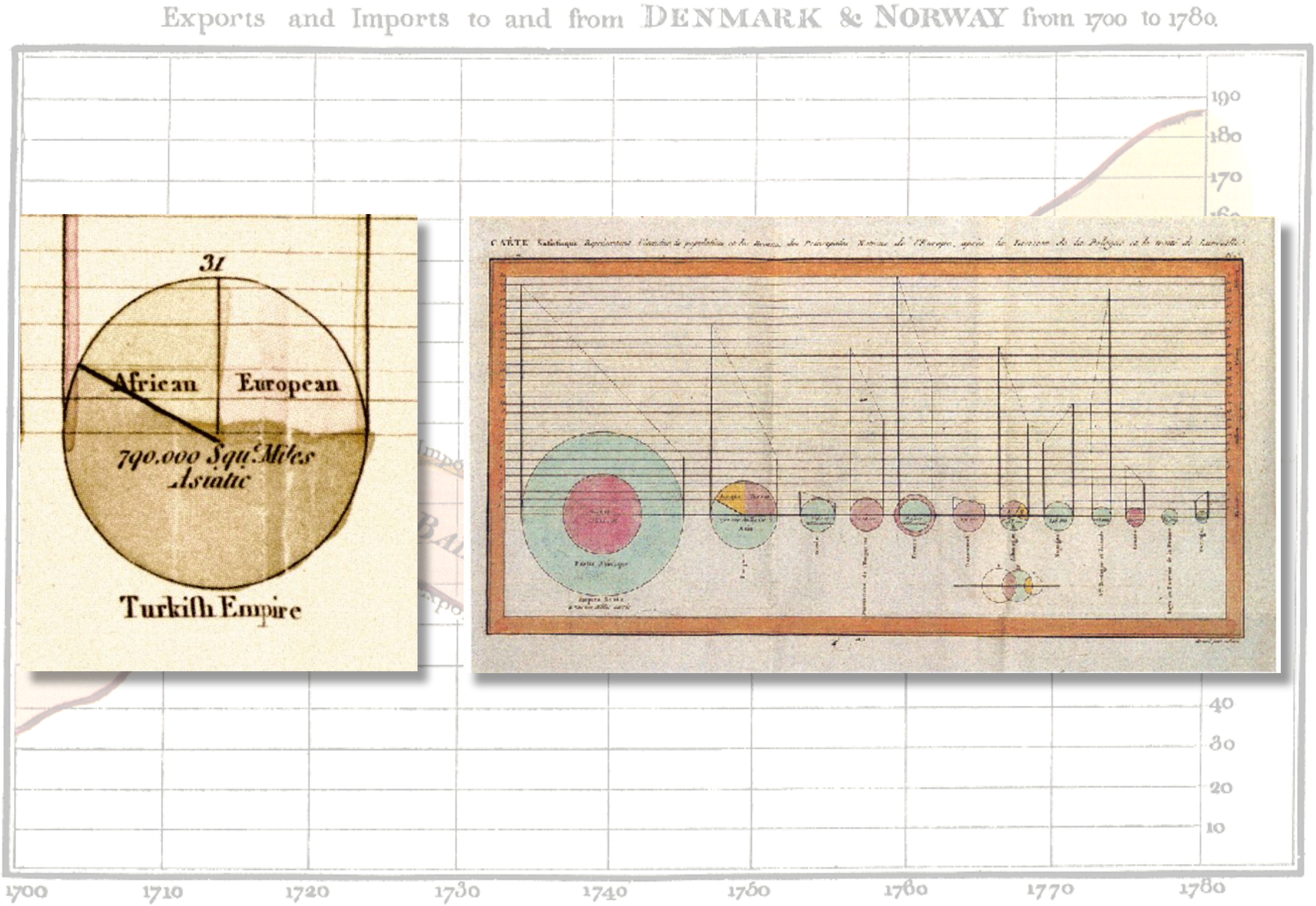
The Rate of Water Evaporation, Lambert 1765

The **Golden Age** of Data Visualization

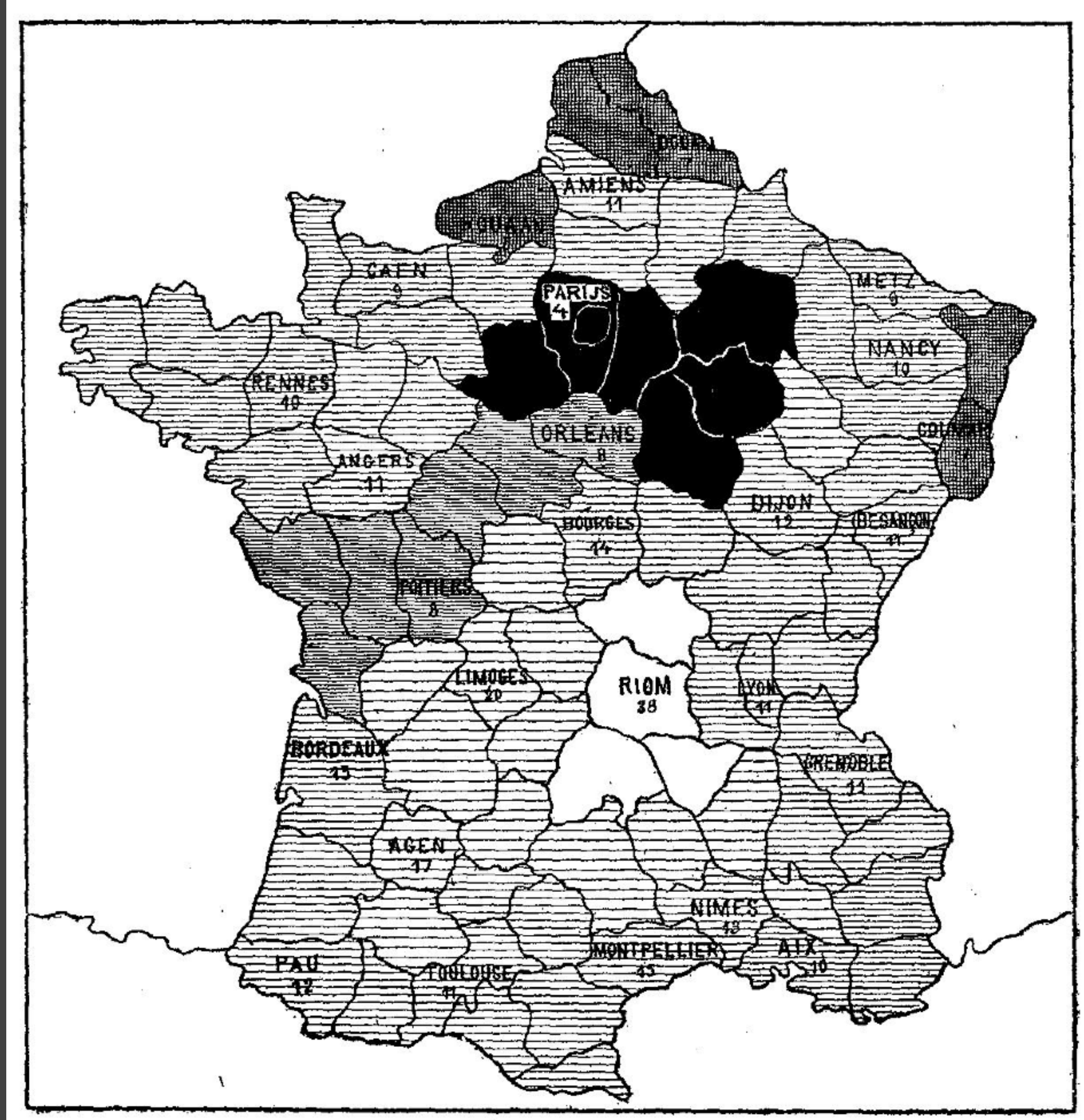
1786 1900



The Commercial and Political Atlas, William Playfair 1786

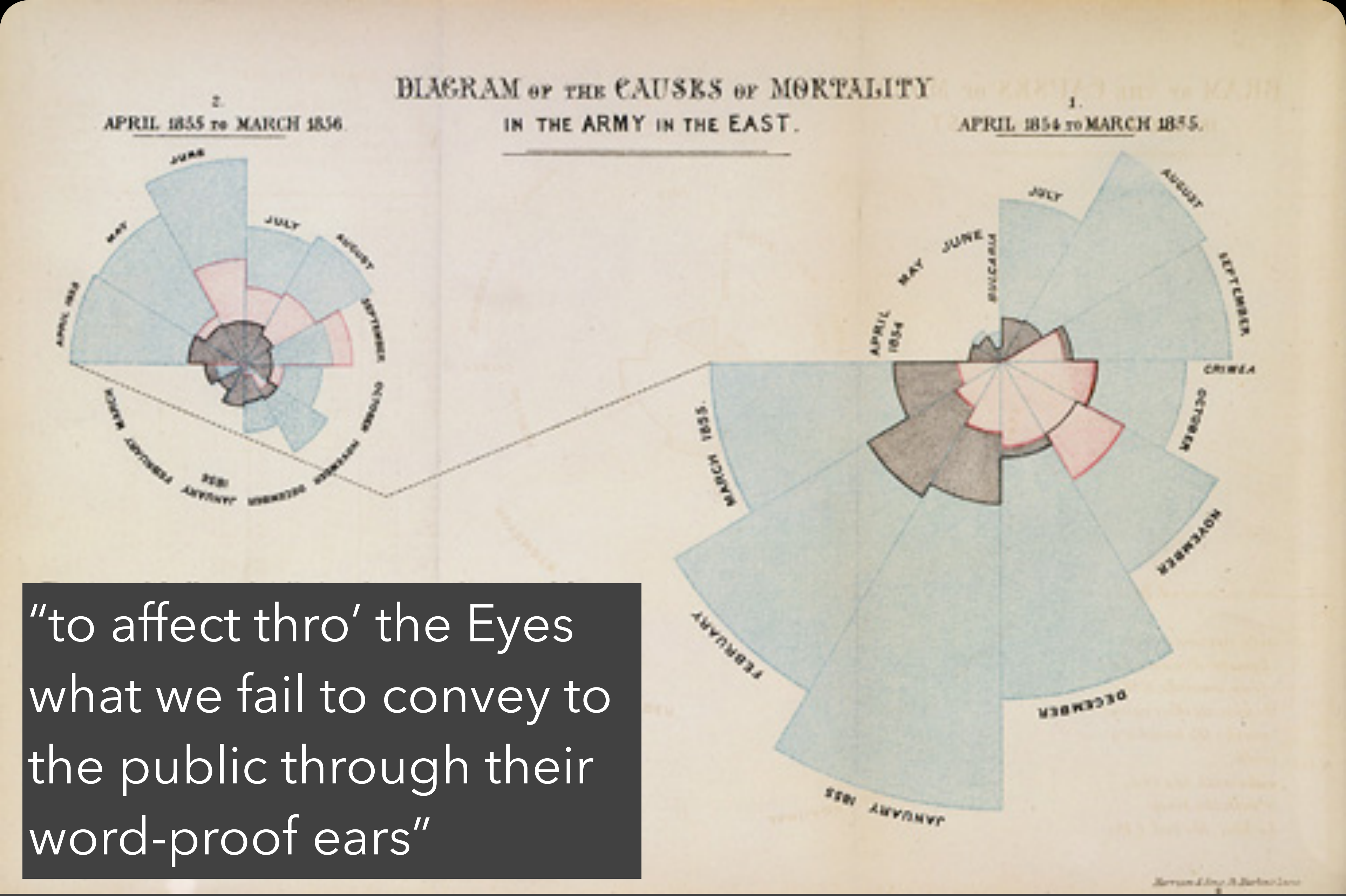


Statistical Breviary, William Playfair 1801



1786

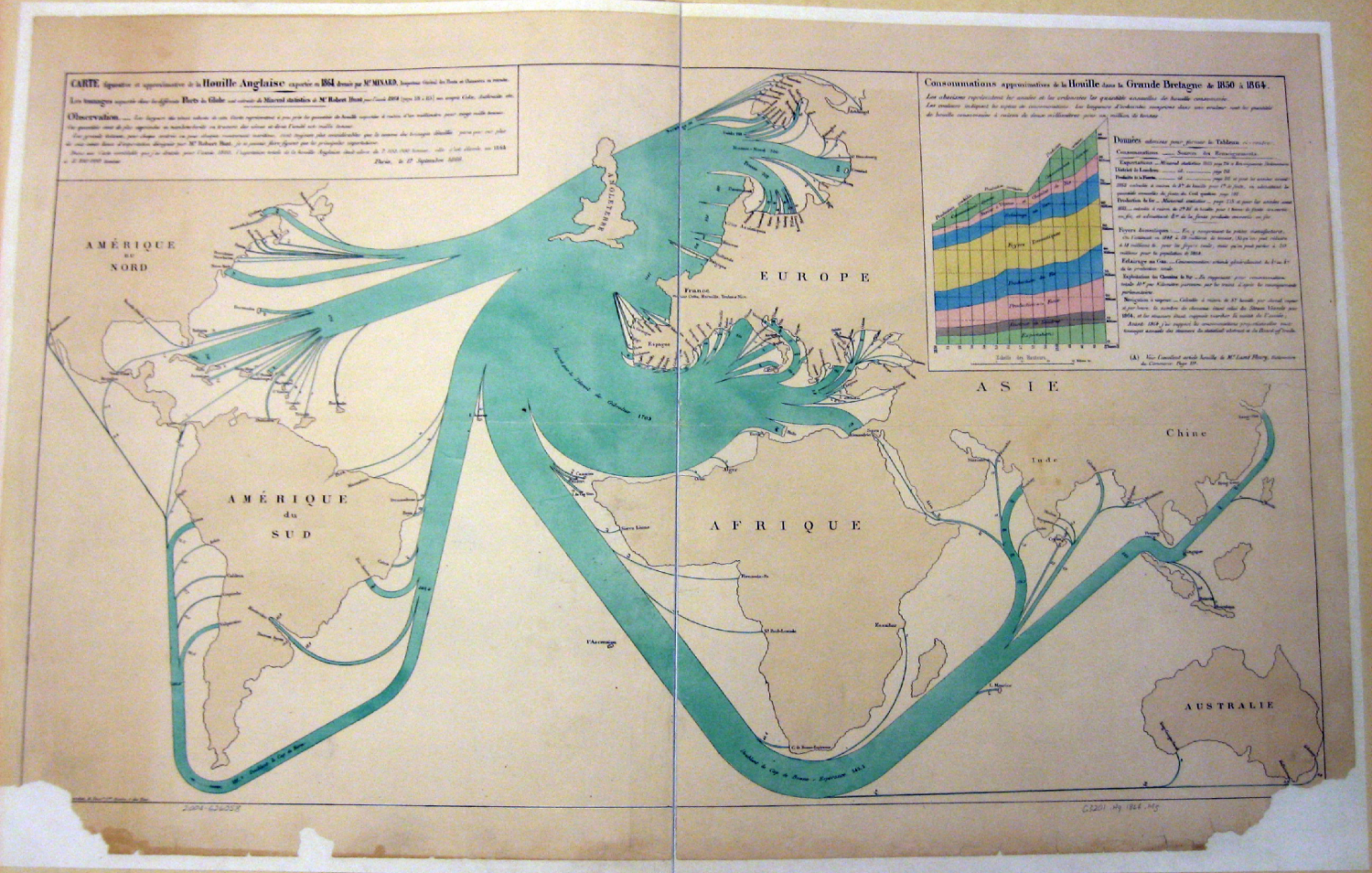
1826(?) Illiteracy in France, Pierre Charles Dupin



“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

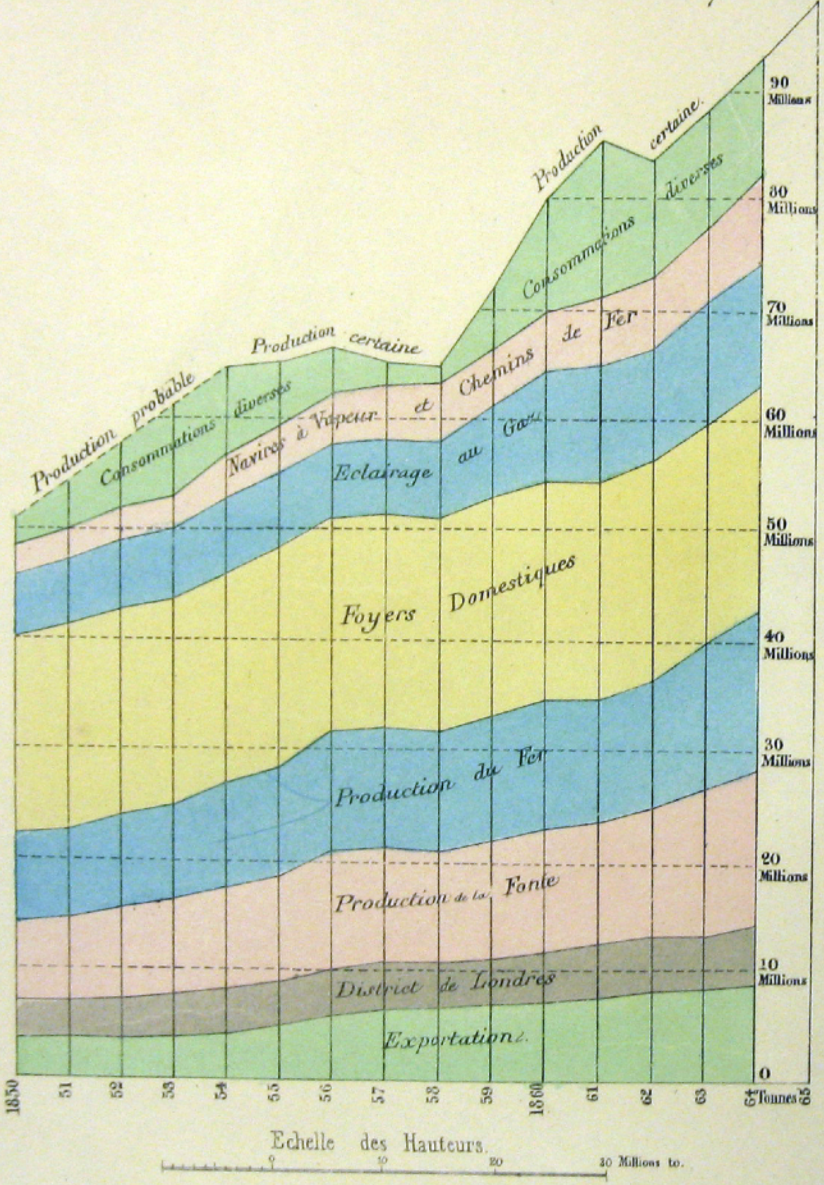


1786

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.*

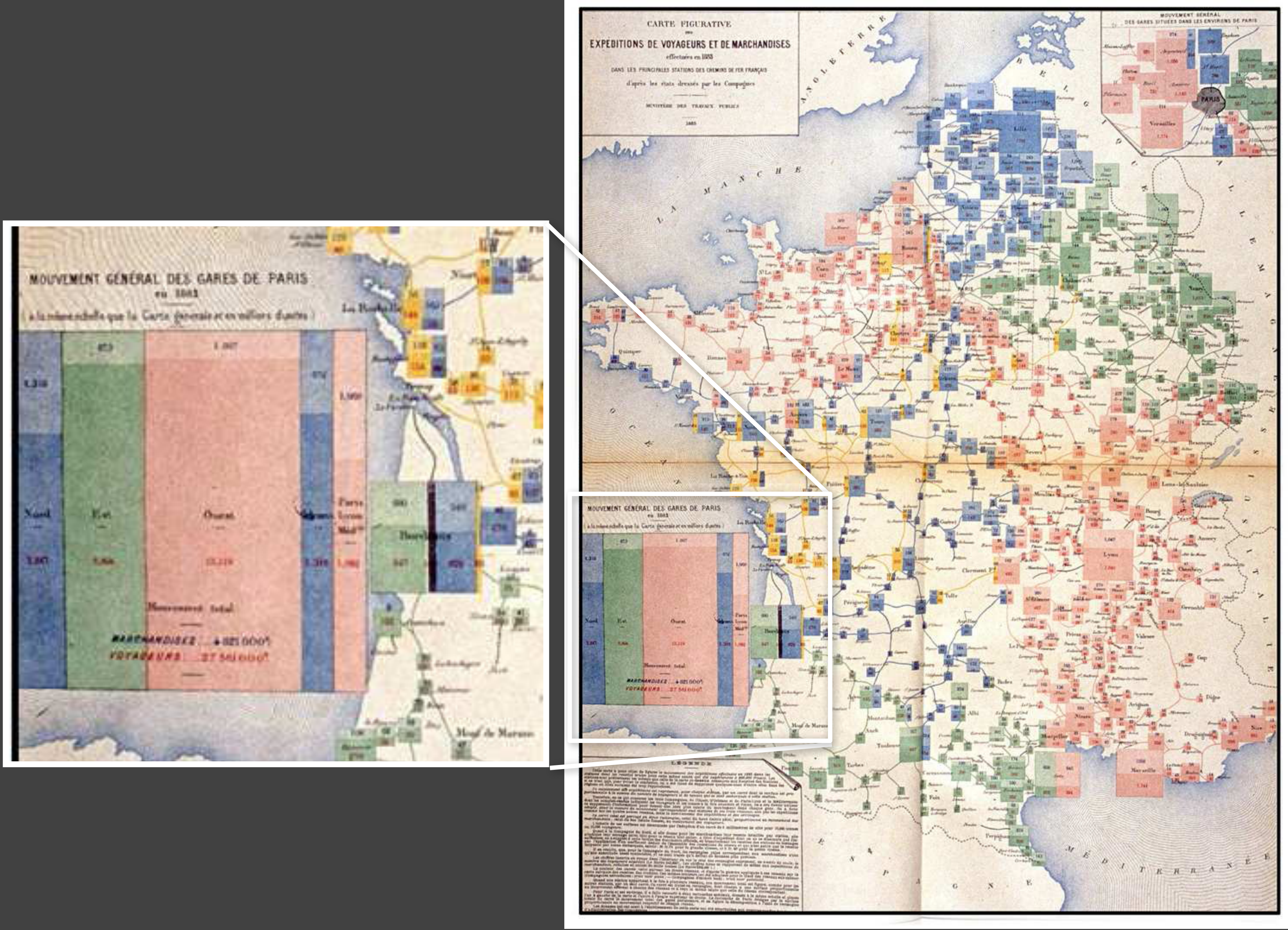
Foyers domestiques. — *En y comprenant les petites manufactures. On l'estimait en 1848 à 19 millions de tonnes,* (A) *qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.*

Eclairage au Gaz. — *Consommation estimée généralement du $\frac{1}{5}$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.*

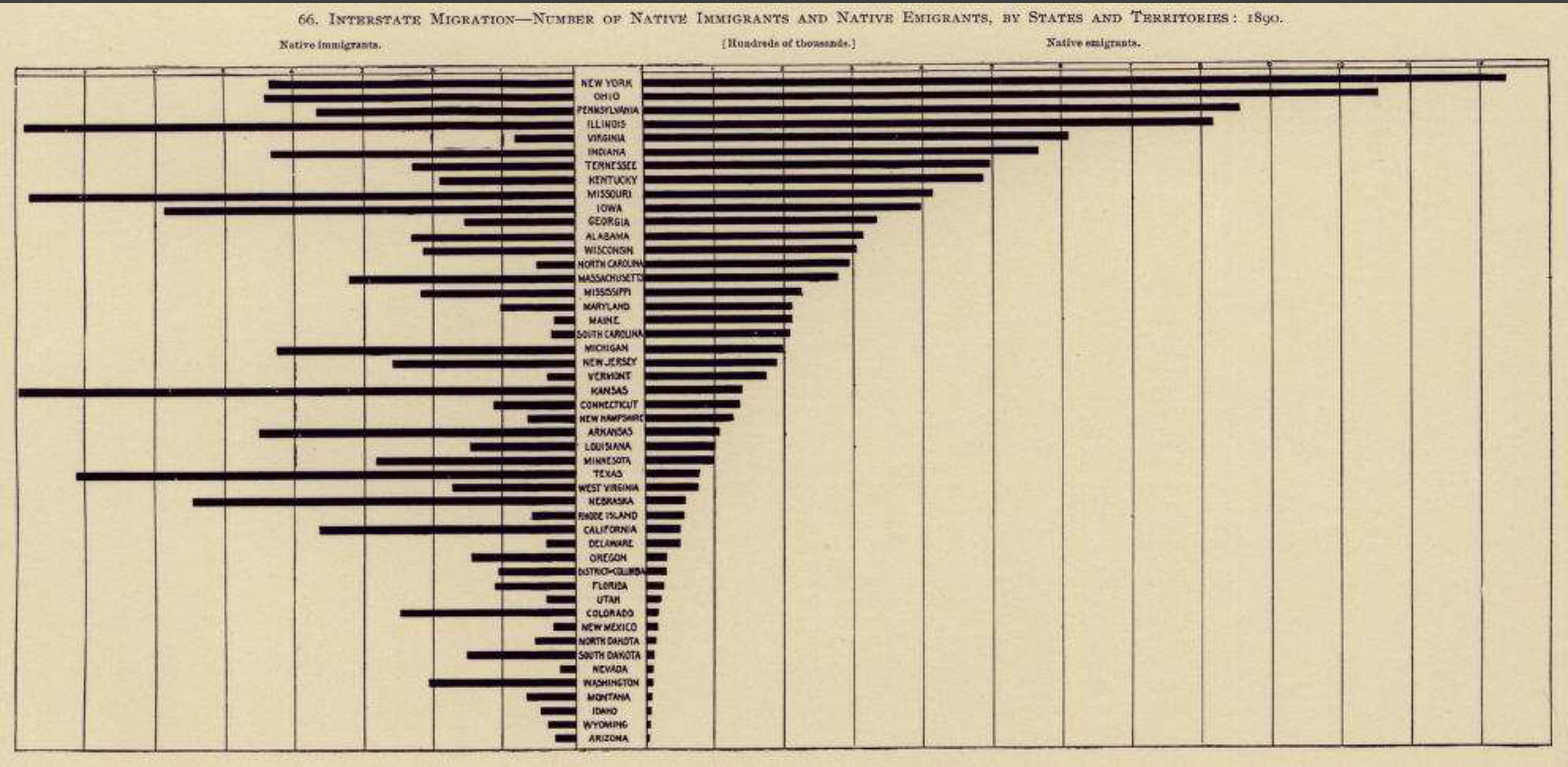
Navigation à vapeur. — *Calculée à raison de 5^{K} houille par cheval vapeur et par heure, le nombre de chevaux étant celui du Steam Vessels pour* 1864, *et les steamers étant supposés marcher la moitié de l'année; Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers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

(A) *Voir l'excellent article houille de Mr Lamé Fleury, Dictionnaire du Commerce Page 111.*



1786

1884 Rail Passengers and Freight from Paris



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**

1786 1900 1950

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

$u_X = 9.0 \quad \sigma_X = 3.317$

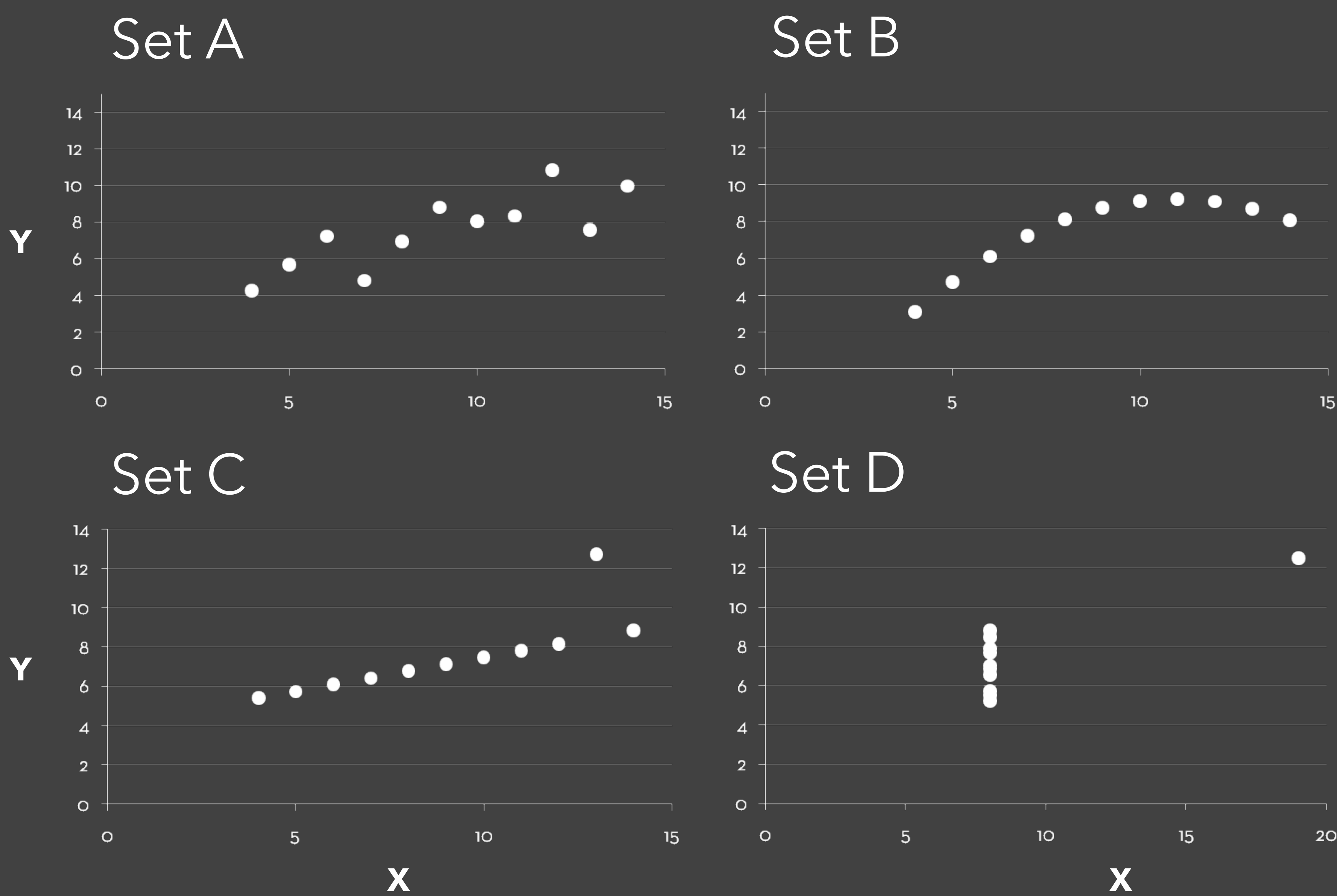
$u_Y = 7.5 \quad \sigma_Y = 2.03$

Linear Regression

$Y = 3 + 0.5\,X$

$R^2 = 0.67$

[Anscombe 1973]



Set A
Set B
Set C
Set D
Y
X
Y
X
0
2
4
6
8
10
12
14
0
5
10
15
20

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

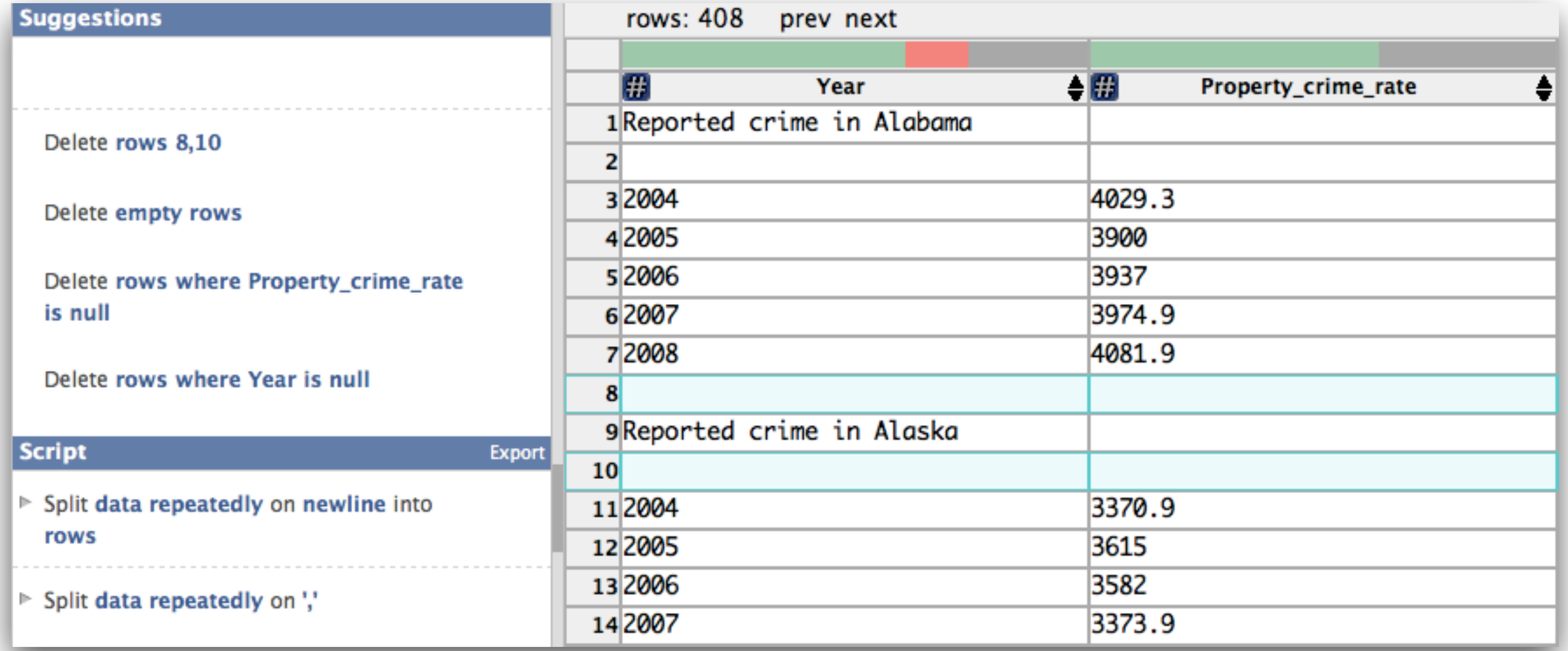
Reported crime in California

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

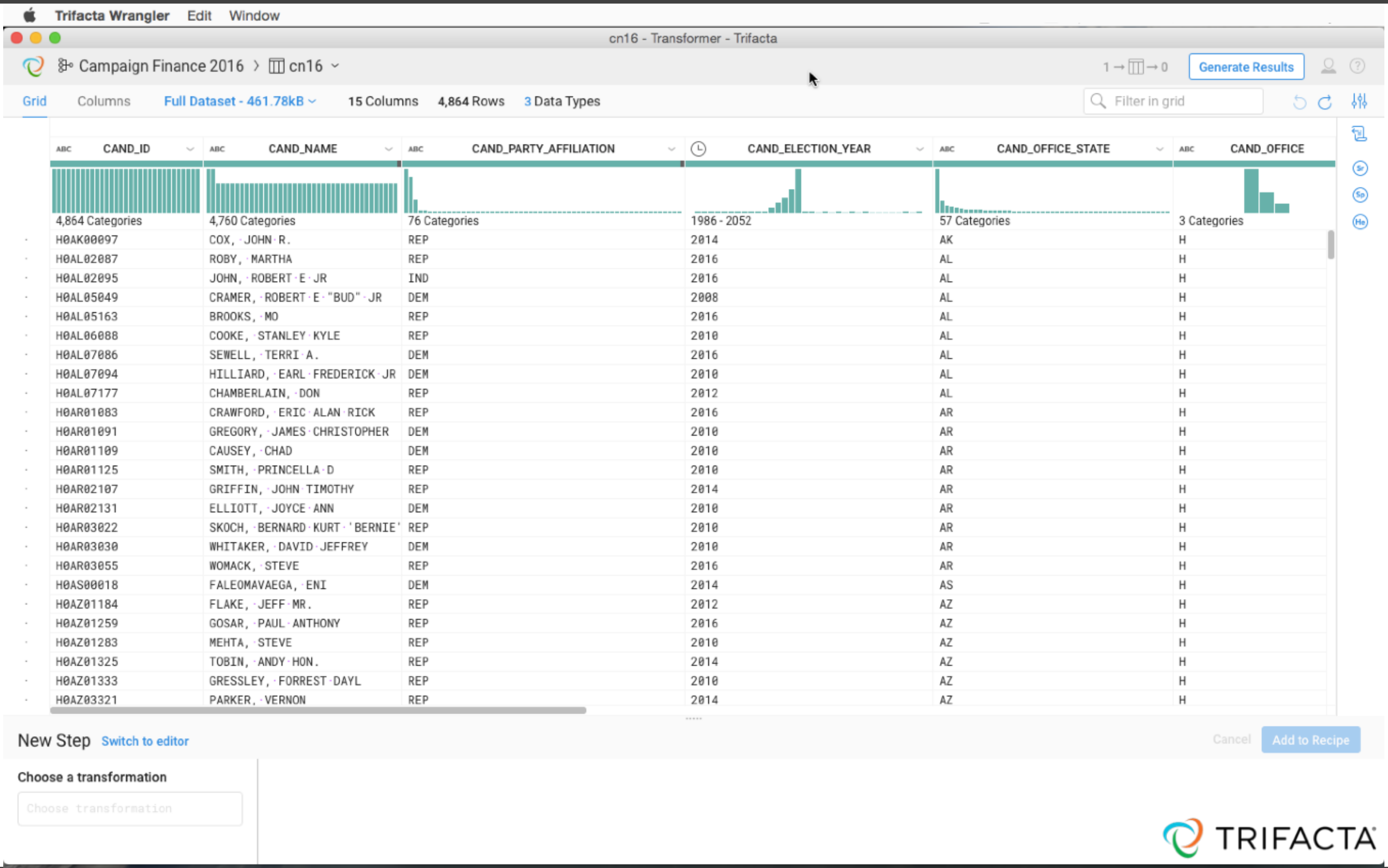
Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

DataWrangler

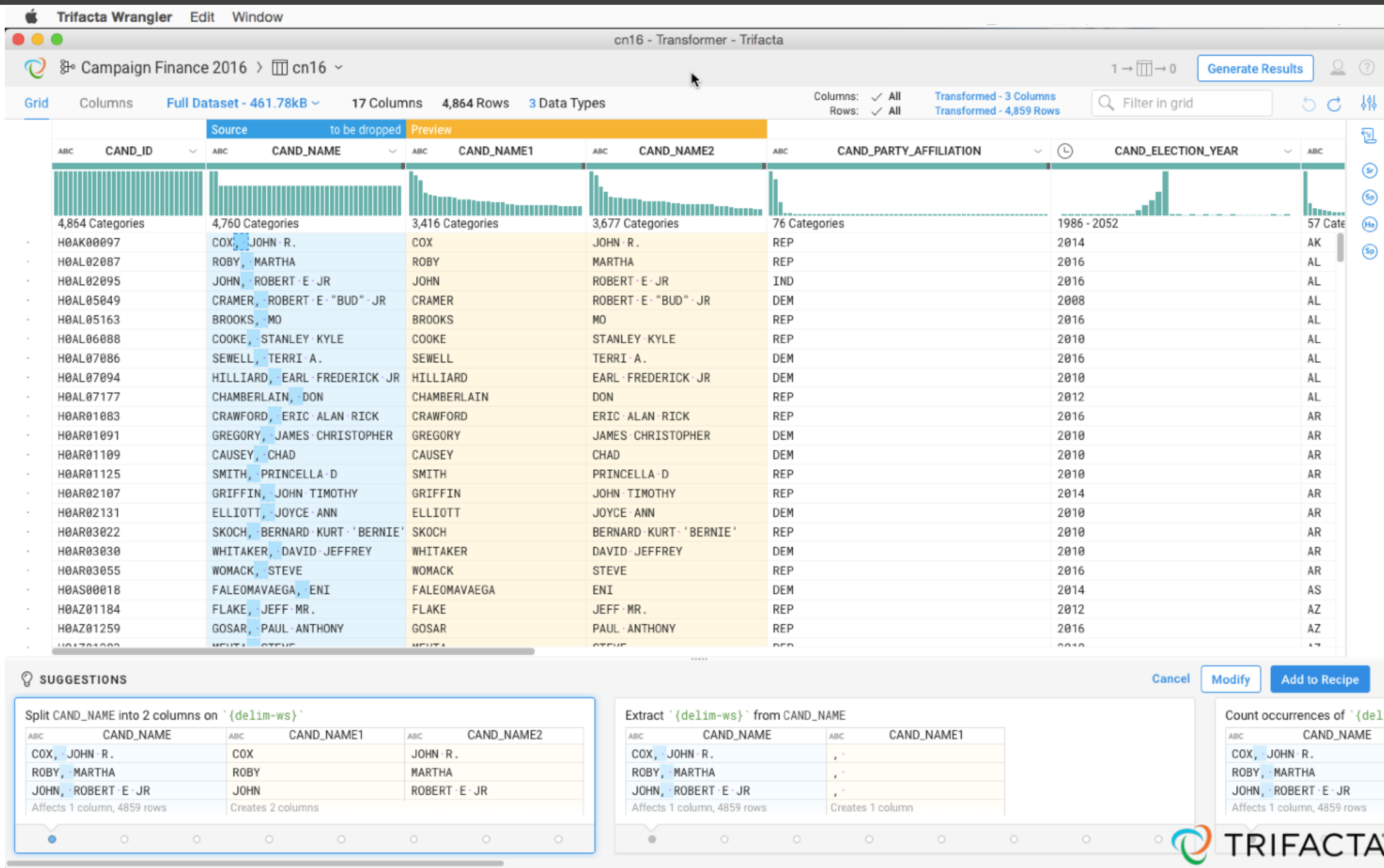


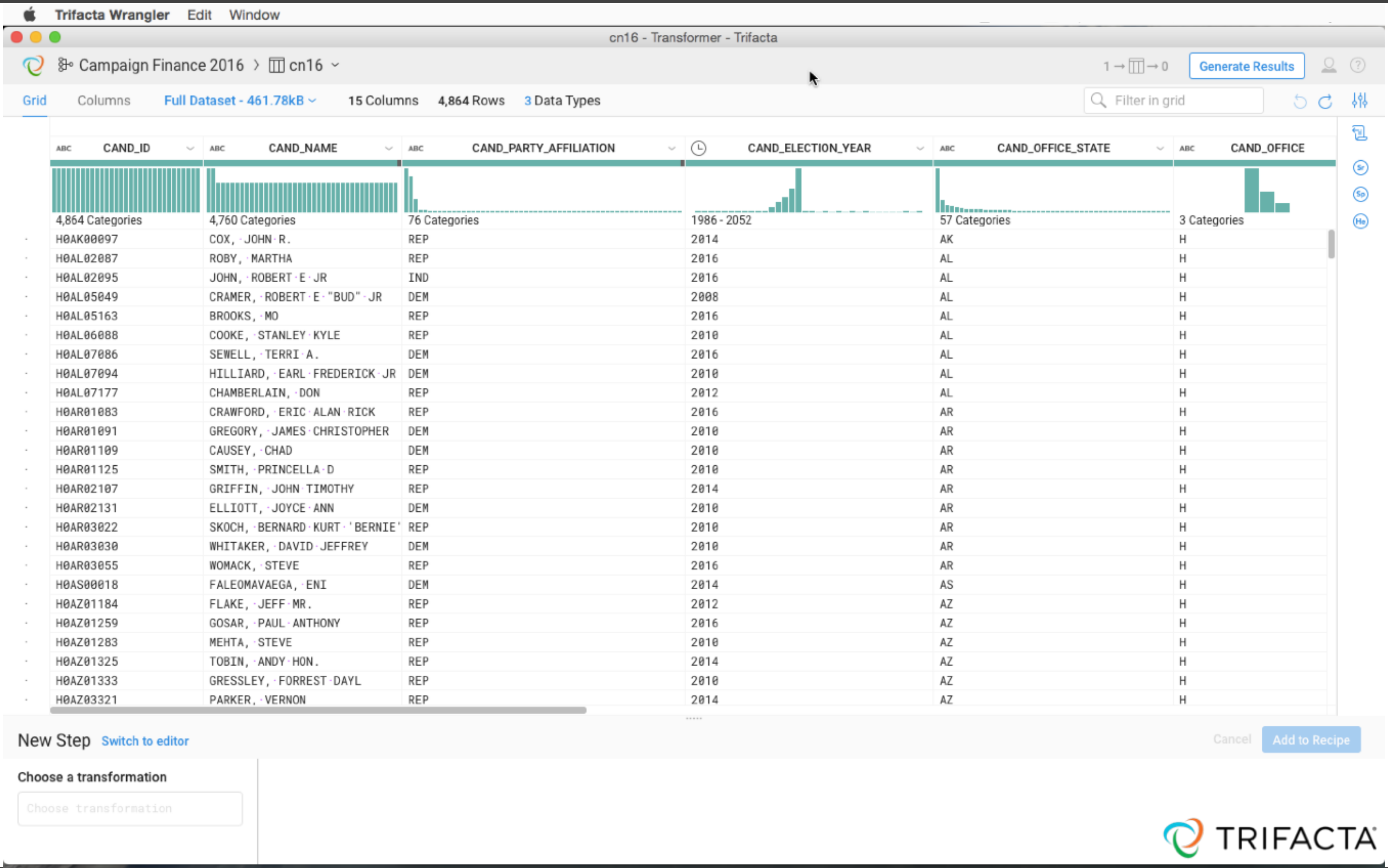
Wrangler: Interactive Visual Specification of Data Transformation Scripts

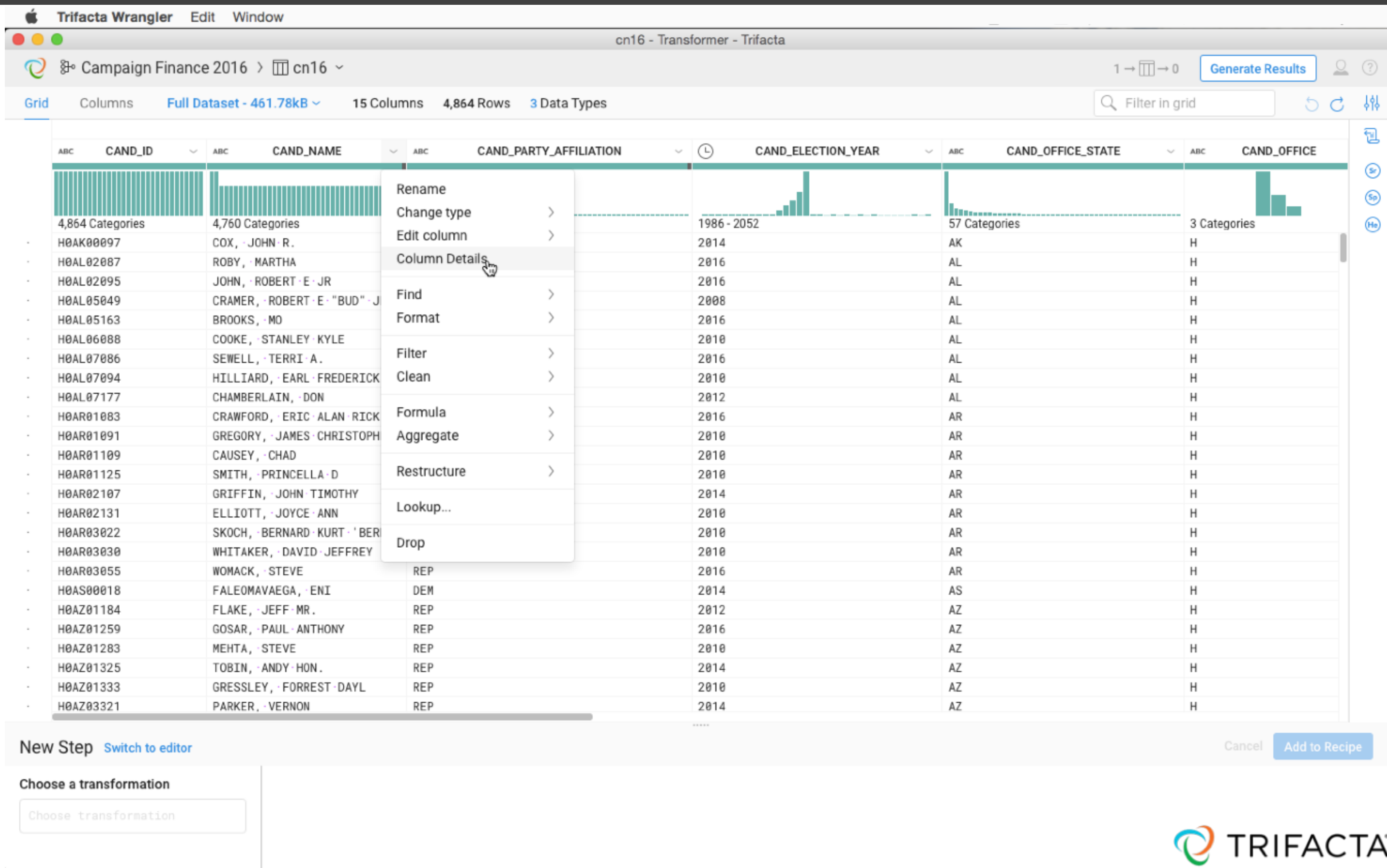
Sean Kandel et al. *CHI'11*



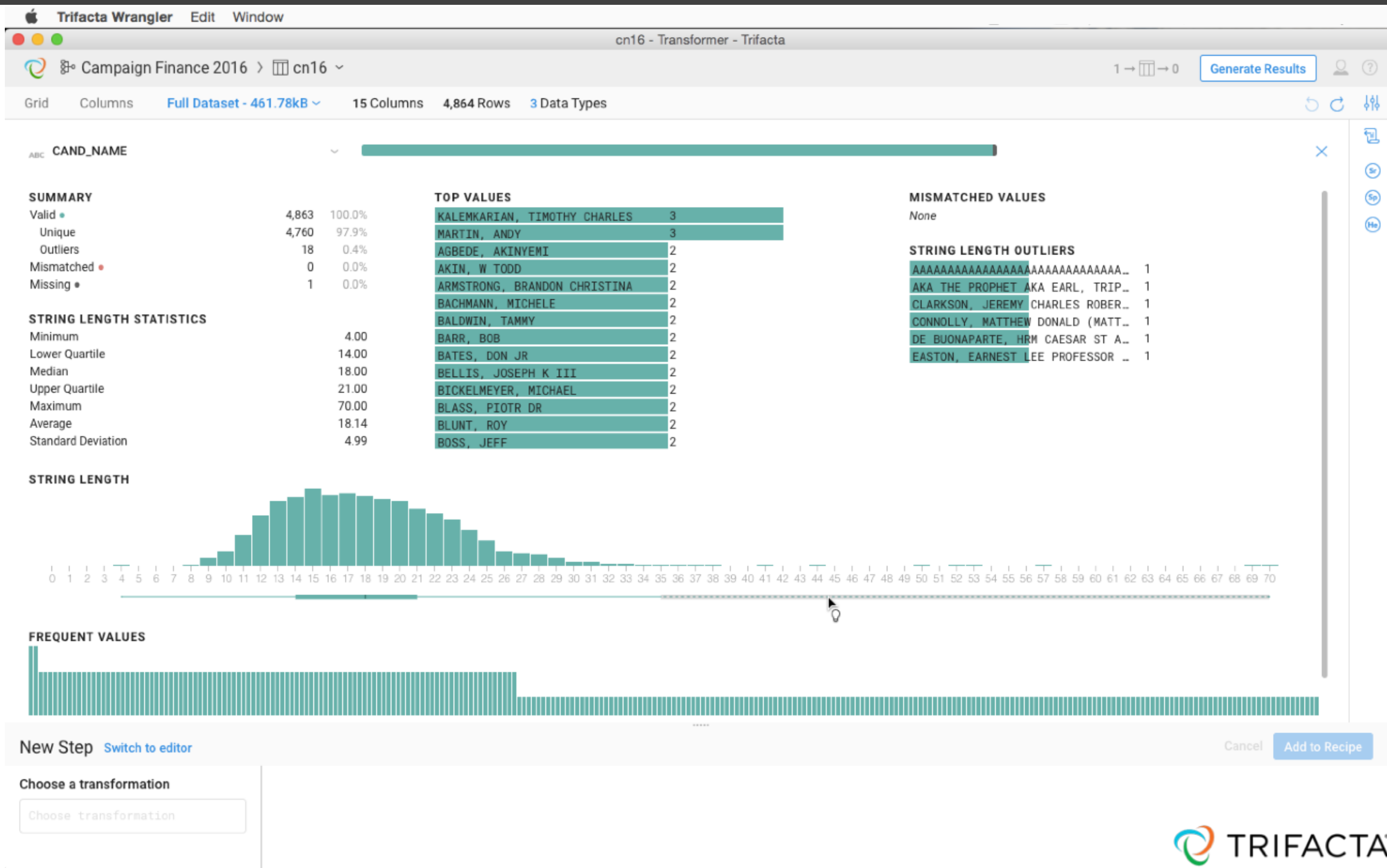
CAND_ID	CAND_NAME	CAND_PARTY_AFFILIATION	CAND_ELECTION_YEAR	CAND_OFFICE_STATE	CAND_OFFICE
4,864 Categories	4,760 Categories	76 Categories	1986 - 2052	57 Categories	3 Categories
H0AK00097	COX, JOHN R.	REP	2014	AK	H
H0AL02087	ROBY, MARTHA	REP	2016	AL	H
H0AL02095	JOHN, ROBERT E JR	IND	2016	AL	H
H0AL05049	CRAMER, ROBERT E "BUD" JR	DEM	2008	AL	H
H0AL05163	BROOKS, MO	REP	2016	AL	H
H0AL06088	COOKE, STANLEY KYLE	REP	2010	AL	H
H0AL07086	SEWELL, TERRI A.	DEM	2016	AL	H
H0AL07094	HILLIARD, EARL FREDERICK JR	DEM	2010	AL	H
H0AL07177	CHAMBERLAIN, DON	REP	2012	AL	H
H0AR01083	CRAWFORD, ERIC ALAN RICK	REP	2016	AR	H
H0AR01091	GREGORY, JAMES CHRISTOPHER	DEM	2010	AR	H
H0AR01109	CAUSEY, CHAD	DEM	2010	AR	H
H0AR01125	SMITH, PRINCELLA D	REP	2010	AR	H
H0AR02107	GRIFFIN, JOHN TIMOTHY	REP	2014	AR	H
H0AR02131	ELLIOTT, JOYCE ANN	DEM	2010	AR	H
H0AR03022	SKOCH, BERNARD KURT 'BERNIE'	REP	2010	AR	H
H0AR03030	WHITAKER, DAVID JEFFREY	DEM	2010	AR	H
H0AR03055	WOMACK, STEVE	REP	2016	AR	H
H0AS00018	FALEOMAVAEGA, ENI	DEM	2014	AS	H
H0AZ01184	FLAKE, JEFF MR.	REP	2012	AZ	H
H0AZ01259	GOSAR, PAUL ANTHONY	REP	2016	AZ	H
H0AZ01283	MEHTA, STEVE	REP	2010	AZ	H
H0AZ01325	TOBIN, ANDY HON.	REP	2014	AZ	H
H0AZ01333	GRESSLEY, FORREST DAYL	REP	2010	AZ	H
H0AZ03321	PARKER, VERNON	REP	2014	AZ	H


Trifacta Wrangler Edit Window
cn16 - Transformer - Trifacta
Campaign Finance 2016 > cn16
Generate Results
Grid Columns Full Dataset - 461.78kB 17 Columns 4,864 Rows 3 Data Types
Columns: All
Rows: All
Transformed - 3 Columns
Transformed - 4,859 Rows
Filter in grid
Source to be dropped
Preview
CAND_ID CAND_NAME CAND_NAME1 CAND_NAME2 CAND_PARTY_AFFILIATION CAND_ELECTION_YEAR
4,864 Categories 4,760 Categories 3,416 Categories 3,677 Categories 76 Categories 1986 - 2052 57 Cate
H0AK00097 COX, JOHN R. COX JOHN R. REP 2014 AK
H0AL02087 ROBY, MARTHA ROBY MARTHA REP 2016 AL
H0AL02095 JOHN, ROBERT E JR JOHN ROBERT E JR IND 2016 AL
H0AL05049 CRAMER, ROBERT E "BUD" JR CRAMER ROBERT E "BUD" JR DEM 2008 AL
H0AL05163 BROOKS, MO BROOKS MO REP 2016 AL
H0AL06088 COOKE, STANLEY KYLE COOKE STANLEY KYLE REP 2010 AL
H0AL07086 SEWELL, TERRI A. SEWELL TERRI A. DEM 2016 AL
H0AL07094 HILLIARD, EARL FREDERICK JR HILLIARD EARL FREDERICK JR DEM 2010 AL
H0AL07177 CHAMBERLAIN, DON CHAMBERLAIN DON REP 2012 AL
H0AR01083 CRAWFORD, ERIC ALAN RICK CRAWFORD ERIC ALAN RICK REP 2016 AR
H0AR01091 GREGORY, JAMES CHRISTOPHER GREGORY JAMES CHRISTOPHER DEM 2010 AR
H0AR01109 CAUSEY, CHAD CAUSEY CHAD DEM 2010 AR
H0AR01125 SMITH, PRINCELLA D SMITH PRINCELLA D REP 2010 AR
H0AR02107 GRIFFIN, JOHN TIMOTHY GRIFFIN JOHN TIMOTHY REP 2014 AR
H0AR02131 ELLIOTT, JOYCE ANN ELLIOTT JOYCE ANN DEM 2010 AR
H0AR03022 SKOCH, BERNARD KURT 'BERNIE' SKOCH BERNARD KURT 'BERNIE' REP 2010 AR
H0AR03030 WHITAKER, DAVID JEFFREY WHITAKER DAVID JEFFREY DEM 2010 AR
H0AR03055 WOMACK, STEVE WOMACK STEVE REP 2016 AR
H0AS00018 FALEOMAVAEGA, ENI FALEOMAVAEGA ENI DEM 2014 AS
H0AZ01184 FLAKE, JEFF MR. FLAKE JEFF MR. REP 2012 AZ
H0AZ01259 GOSAR, PAUL ANTHONY GOSAR PAUL ANTHONY REP 2016 AZ
SUGGESTIONS
Cancel Modify Add to Recipe
Split CAND_NAME into 2 columns on `{delim-ws}`
CAND_NAME CAND_NAME1 CAND_NAME2
COX, JOHN R. COX JOHN R.
ROBY, MARTHA ROBY MARTHA
JOHN, ROBERT E JR JOHN ROBERT E JR
Affects 1 column, 4859 rows Creates 2 columns
Extract `{delim-ws}` from CAND_NAME
CAND_NAME CAND_NAME1
COX, JOHN R. ,
ROBY, MARTHA ,
JOHN, ROBERT E JR ,
Affects 1 column, 4859 rows Creates 1 column
Count occurrences of `{delim
CAND_NAME
COX, JOHN R.
ROBY, MARTHA
JOHN, ROBERT E JR
Affects 1 column, 4859 rows
TRIFACTA


Trifacta Wrangler Edit Window
cn16 - Transformer - Trifacta
Campaign Finance 2016 > cn16
1 → 0
Generate Results
Grid Columns Full Dataset - 461.78kB 15 Columns 4,864 Rows 3 Data Types
Filter in grid
CAND_ID CAND_NAME CAND_PARTY_AFFILIATION CAND_ELECTION_YEAR CAND_OFFICE_STATE CAND_OFFICE
4,864 Categories 4,760 Categories 76 Categories 1986 - 2052 57 Categories 3 Categories
H0AK00097 COX, JOHN R. REP 2014 AK H
H0AL02087 ROBY, MARTHA REP 2016 AL H
H0AL02095 JOHN, ROBERT E JR IND 2016 AL H
H0AL05049 CRAMER, ROBERT E "BUD" JR DEM 2008 AL H
H0AL05163 BROOKS, MO REP 2016 AL H
H0AL06088 COOKE, STANLEY KYLE REP 2010 AL H
H0AL07086 SEWELL, TERRI A. DEM 2016 AL H
H0AL07094 HILLIARD, EARL FREDERICK JR DEM 2010 AL H
H0AL07177 CHAMBERLAIN, DON REP 2012 AL H
H0AR01083 CRAWFORD, ERIC ALAN RICK REP 2016 AR H
H0AR01091 GREGORY, JAMES CHRISTOPHER DEM 2010 AR H
H0AR01109 CAUSEY, CHAD DEM 2010 AR H
H0AR01125 SMITH, PRINCELLA D REP 2010 AR H
H0AR02107 GRIFFIN, JOHN TIMOTHY REP 2014 AR H
H0AR02131 ELLIOTT, JOYCE ANN DEM 2010 AR H
H0AR03022 SKOCH, BERNARD KURT 'BERNIE' REP 2010 AR H
H0AR03030 WHITAKER, DAVID JEFFREY DEM 2010 AR H
H0AR03055 WOMACK, STEVE REP 2016 AR H
H0AS00018 FALEOMAVAEGA, ENI DEM 2014 AS H
H0AZ01184 FLAKE, JEFF MR. REP 2012 AZ H
H0AZ01259 GOSAR, PAUL ANTHONY REP 2016 AZ H
H0AZ01283 MEHTA, STEVE REP 2010 AZ H
H0AZ01325 TOBIN, ANDY HON. REP 2014 AZ H
H0AZ01333 GRESSLEY, FORREST DAYL REP 2010 AZ H
H0AZ03321 PARKER, VERNON REP 2014 AZ H
New Step Switch to editor
Cancel Add to Recipe
Choose a transformation
Choose transformation
TRIFACTA



Trifacta Wrangler Edit Window
cn16 - Transformer - Trifacta
Campaign Finance 2016 › cn16
1 → 0
Generate Results
Grid Columns Full Dataset - 461.78kB 15 Columns 4,864 Rows 3 Data Types
Filter in grid
CAND_ID CAND_NAME CAND_PARTY_AFFILIATION CAND_ELECTION_YEAR CAND_OFFICE_STATE CAND_OFFICE
4,864 Categories 4,760 Categories 1986 - 2052 57 Categories 3 Categories
H0AK00097 COX, JOHN R. 2014 AK H
H0AL02087 ROBY, MARTHA 2016 AL H
H0AL02095 JOHN, ROBERT E JR 2016 AL H
H0AL05049 CRAMER, ROBERT E "BUD" J 2008 AL H
H0AL05163 BROOKS, MO 2016 AL H
H0AL06088 COOKE, STANLEY KYLE 2010 AL H
H0AL07086 SEWELL, TERRI A. 2016 AL H
H0AL07094 HILLIARD, EARL FREDERICK 2010 AL H
H0AL07177 CHAMBERLAIN, DON 2012 AL H
H0AR01083 CRAWFORD, ERIC ALAN RICK 2016 AR H
H0AR01091 GREGORY, JAMES CHRISTOPH 2010 AR H
H0AR01109 CAUSEY, CHAD 2010 AR H
H0AR01125 SMITH, PRINCELLA D 2010 AR H
H0AR02107 GRIFFIN, JOHN TIMOTHY 2014 AR H
H0AR02131 ELLIOTT, JOYCE ANN 2010 AR H
H0AR03022 SKOCH, BERNARD KURT 'BER 2010 AR H
H0AR03030 WHITAKER, DAVID JEFFREY 2010 AR H
H0AR03055 WOMACK, STEVE REP 2016 AR H
H0AS00018 FALEOMAVAEGA, ENI DEM 2014 AS H
H0AZ01184 FLAKE, JEFF MR. REP 2012 AZ H
H0AZ01259 GOSAR, PAUL ANTHONY REP 2016 AZ H
H0AZ01283 MEHTA, STEVE REP 2010 AZ H
H0AZ01325 TOBIN, ANDY HON. REP 2014 AZ H
H0AZ01333 GRESSLEY, FORREST DAYL REP 2010 AZ H
H0AZ03321 PARKER, VERNON REP 2014 AZ H
Rename
Change type
Edit column
Column Details
Find
Format
Filter
Clean
Formula
Aggregate
Restructure
Lookup...
Drop
New Step Switch to editor
Cancel Add to Recipe
Choose a transformation
Choose transformation
TRIFACTA


Trifacta Wrangler Edit Window
cn16 - Transformer - Trifacta
Campaign Finance 2016 › cn16
Generate Results
Grid Columns Full Dataset - 461.78kB 15 Columns 4,864 Rows 3 Data Types
CAND_NAME
SUMMARY
Valid 4,863 100.0%
Unique 4,760 97.9%
Outliers 18 0.4%
Mismatched 0 0.0%
Missing 1 0.0%
STRING LENGTH STATISTICS
Minimum 4.00
Lower Quartile 14.00
Median 18.00
Upper Quartile 21.00
Maximum 70.00
Average 18.14
Standard Deviation 4.99
TOP VALUES
KALEMKARIAN, TIMOTHY CHARLES 3
MARTIN, ANDY 3
AGBEDE, AKINYEMI 2
AKIN, W TODD 2
ARMSTRONG, BRANDON CHRISTINA 2
BACHMANN, MICHELE 2
BALDWIN, TAMMY 2
BARR, BOB 2
BATES, DON JR 2
BELLIS, JOSEPH K III 2
BICKELMEYER, MICHAEL 2
BLASS, PIOTR DR 2
BLUNT, ROY 2
BOSS, JEFF 2
MISMATCHED VALUES
None
STRING LENGTH OUTLIERS
AAAAAAAAAAAAAAAAAAAAAAAAAAAAA… 1
AKA THE PROPHET AKA EARL, TRIP… 1
CLARKSON, JEREMY CHARLES ROBER… 1
CONNOLLY, MATTHEW DONALD (MATT… 1
DE BUONAPARTE, HRM CAESAR ST A… 1
EASTON, EARNEST LEE PROFESSOR … 1
STRING LENGTH
FREQUENT VALUES
New Step Switch to editor
Cancel Add to Recipe
Choose a transformation
Choose transformation
TRIFACTA

Data Wrangling

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

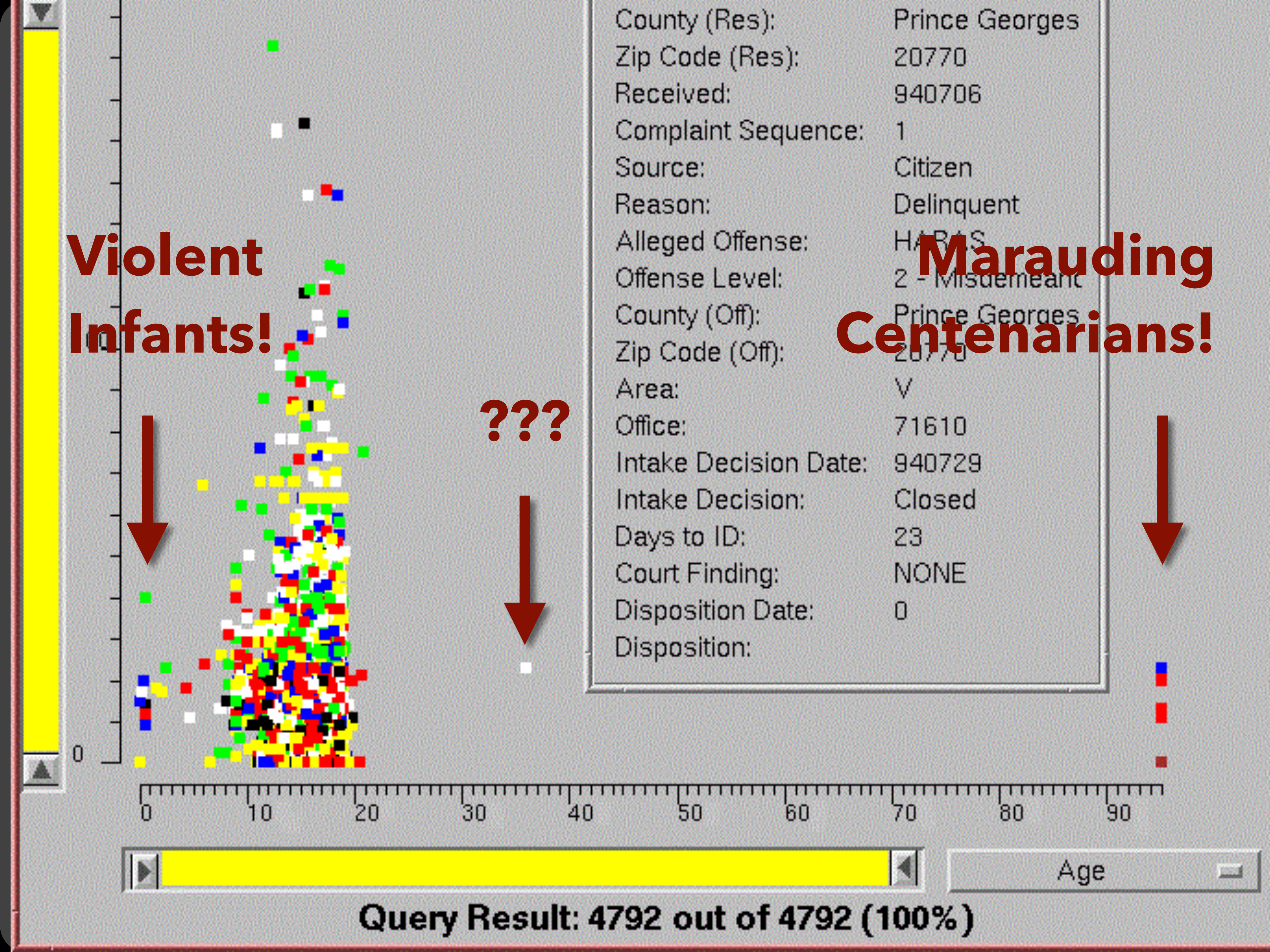
Approaches include:
Manual manipulation in spreadsheets
Custom code (e.g., dplyr in R, Pandas in Python)
Trifacta Wrangler http://www.trifacta.com/products/wrangler/
Open Refine http://openrefine.org/

Data Quality

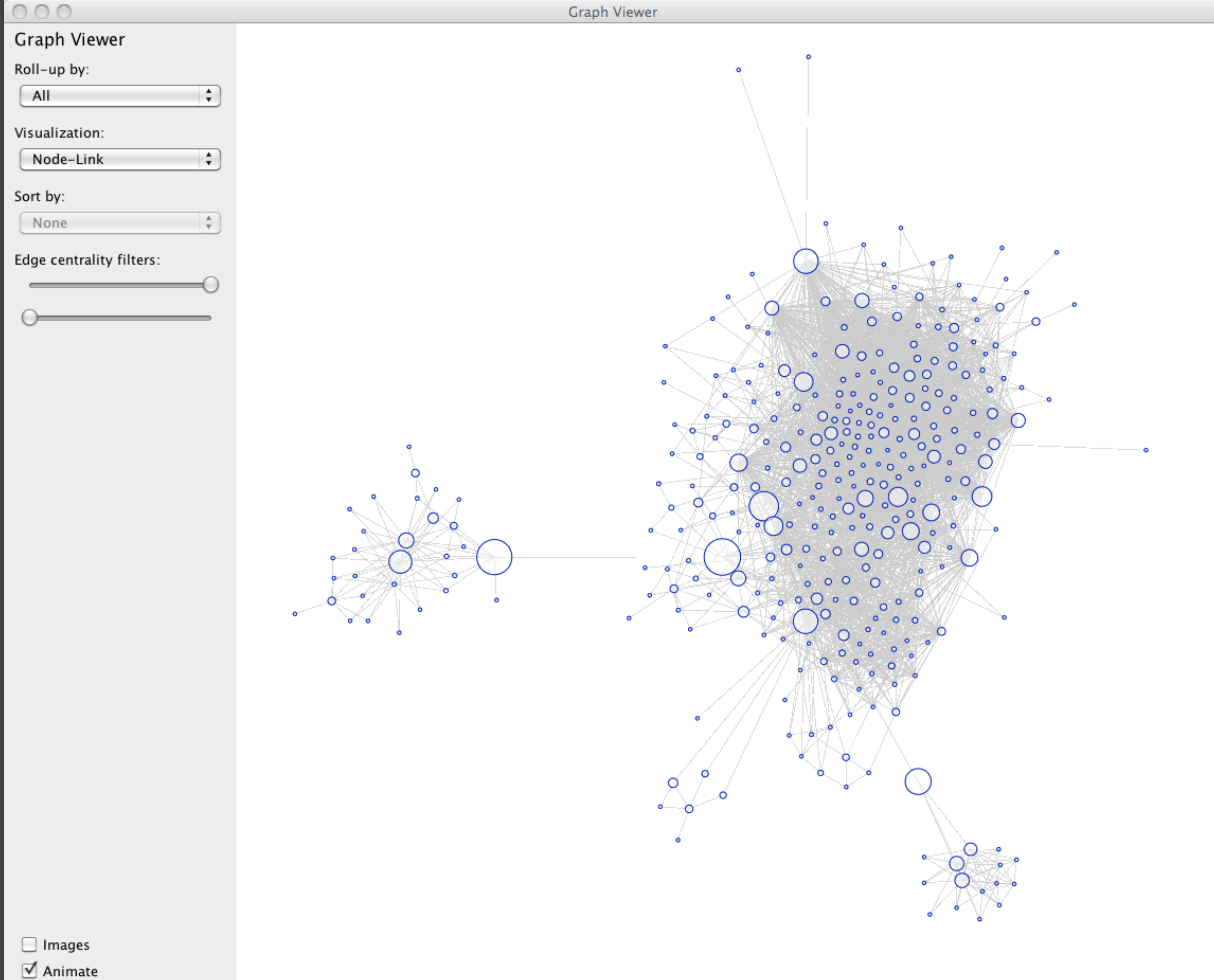
“The first sign that a visualization is good is that it shows you a problem in your data…

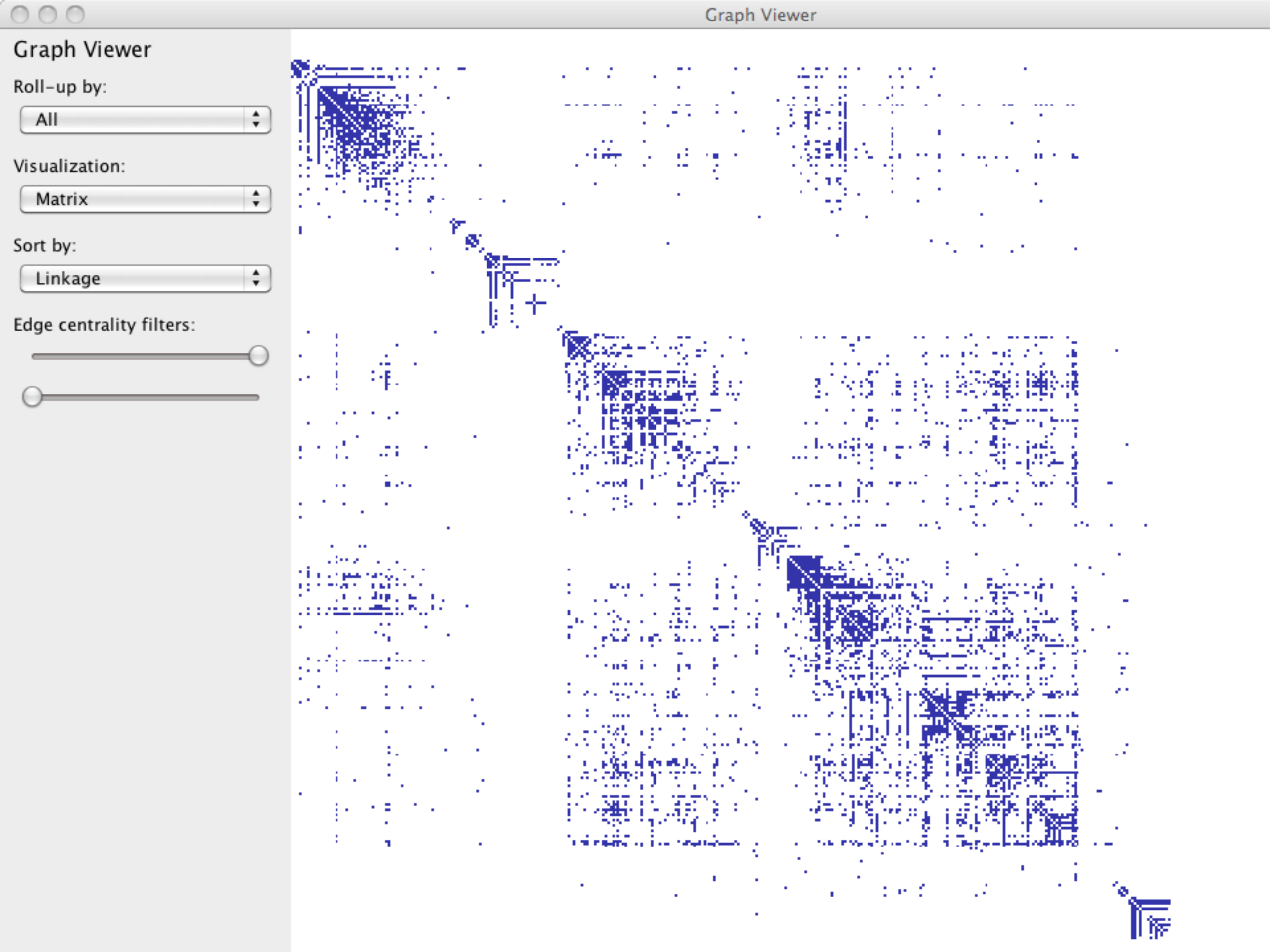
…every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

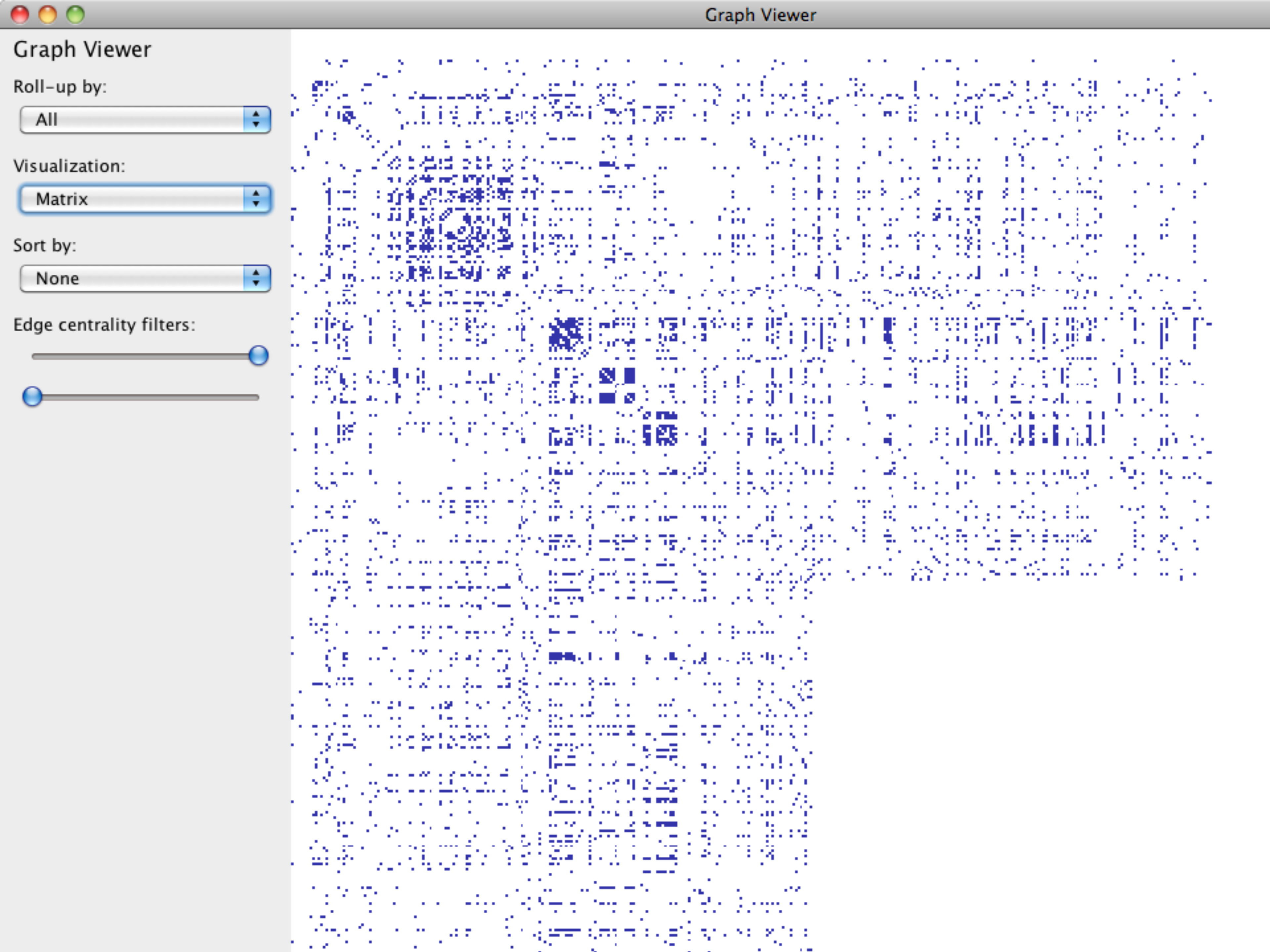
Martin Wattenberg



Violent Infants!
???
Marauding Centenarians!
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense:
Offense Level:
County (Off):
Zip Code (Off):
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%)

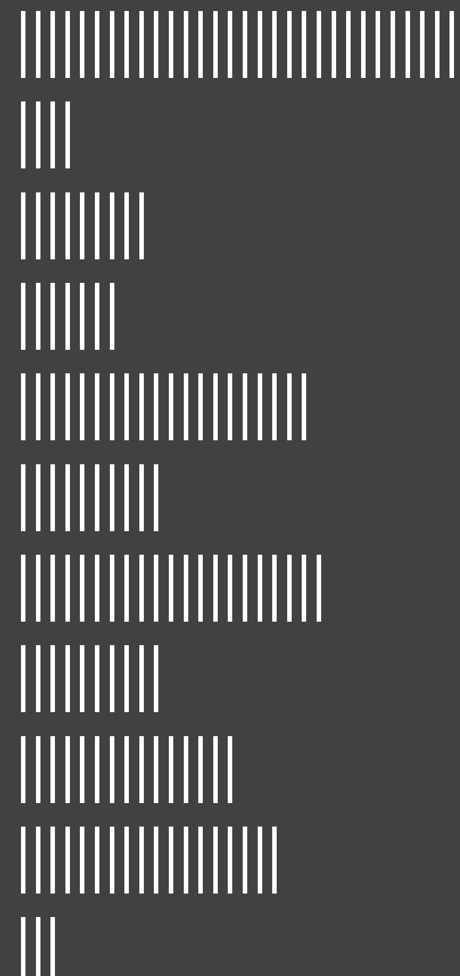

Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


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?

Berkeley
Cornell
Harvard
Harvard University
Stanford
Stanford University
UC Berkeley
UC Davis
University of California at Berkeley
University of California, Berkeley
University of California, Davis



Data Quality Hurdles

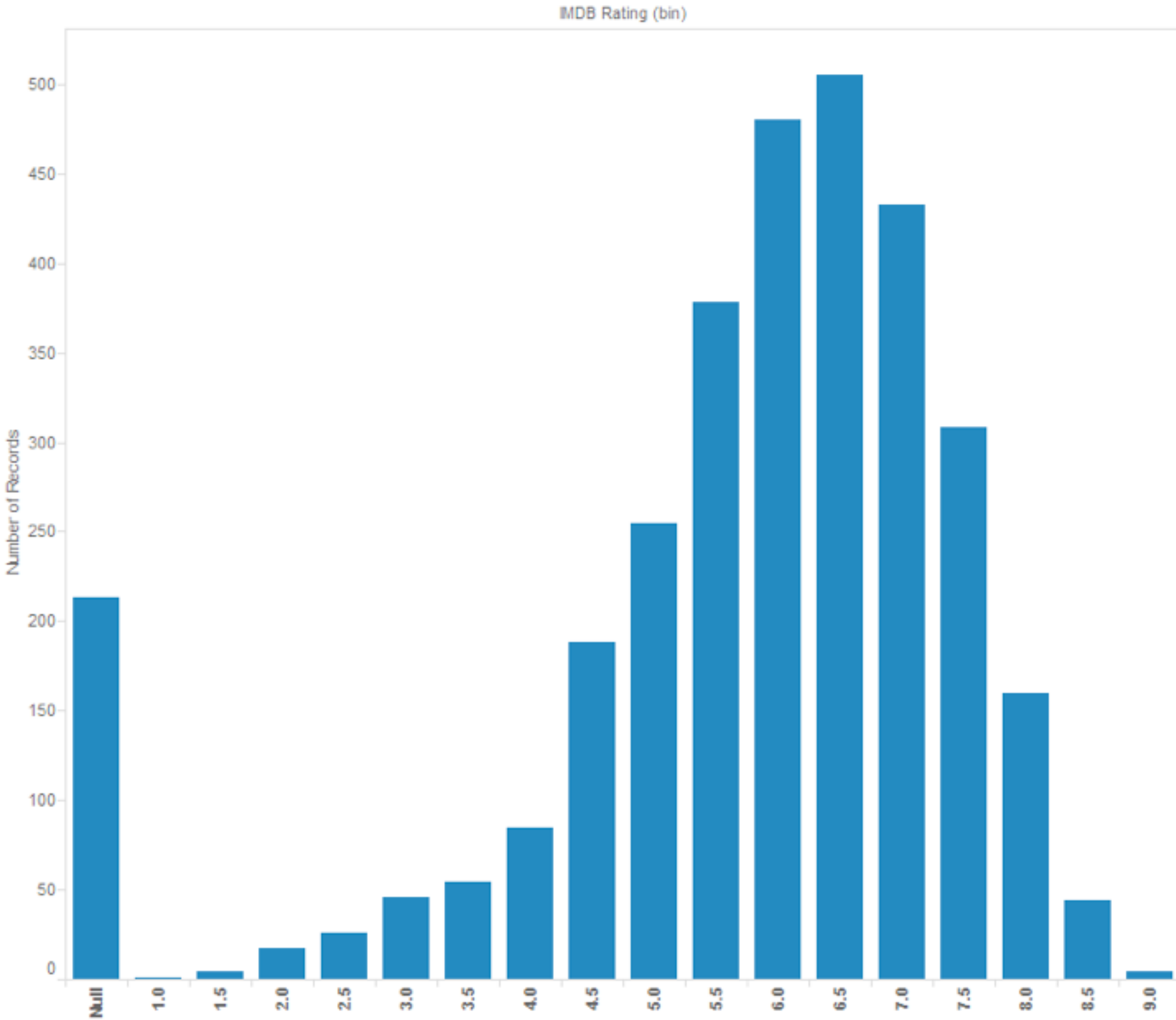
Missing Data	no measurements, redacted, …?
Erroneous Values	misspelling, outliers, …?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

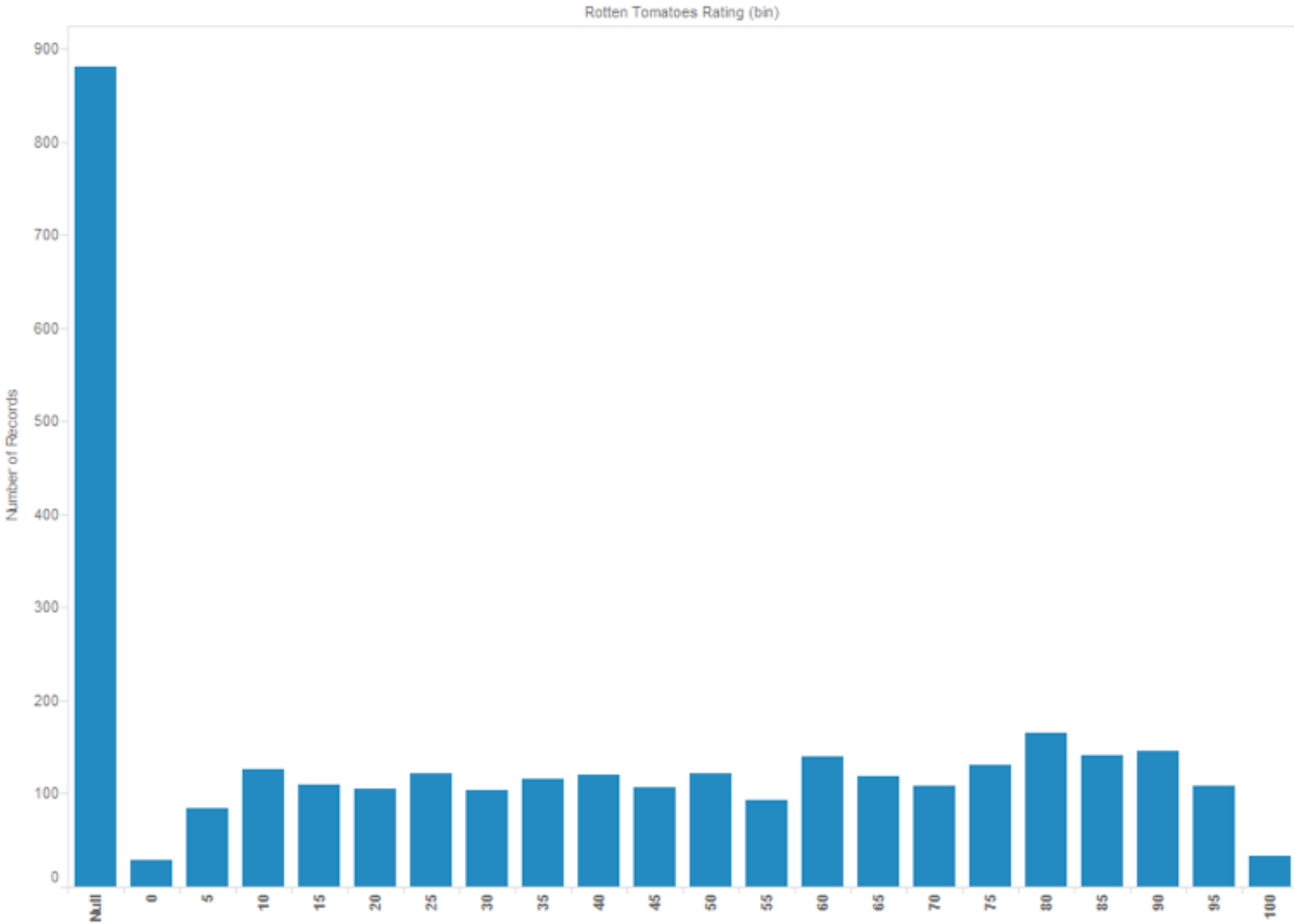
LESSON: Anticipate problems with your data.
Many research problems around these issues!

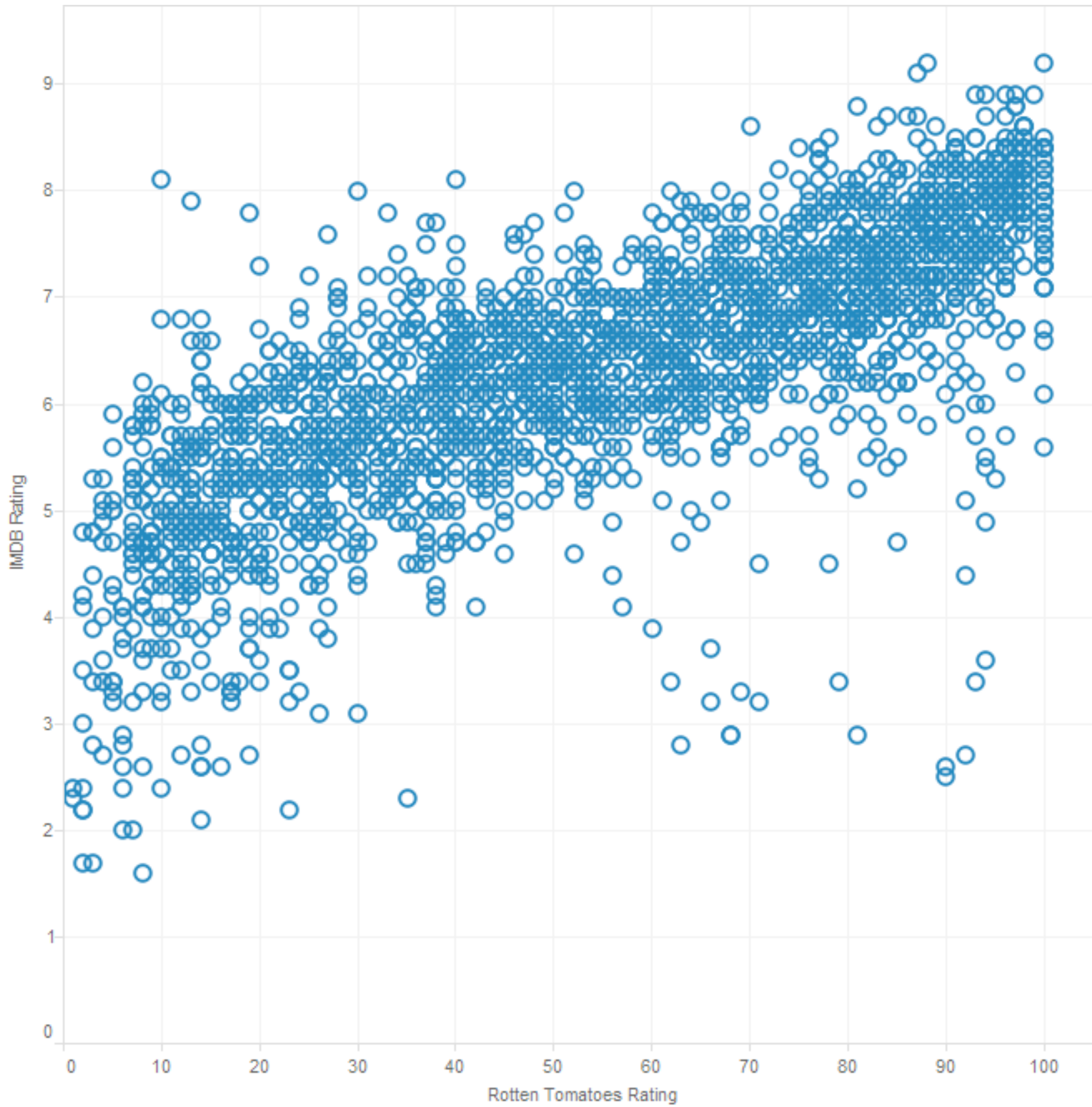
Analysis Example: Motion Pictures Data

Motion Pictures Data

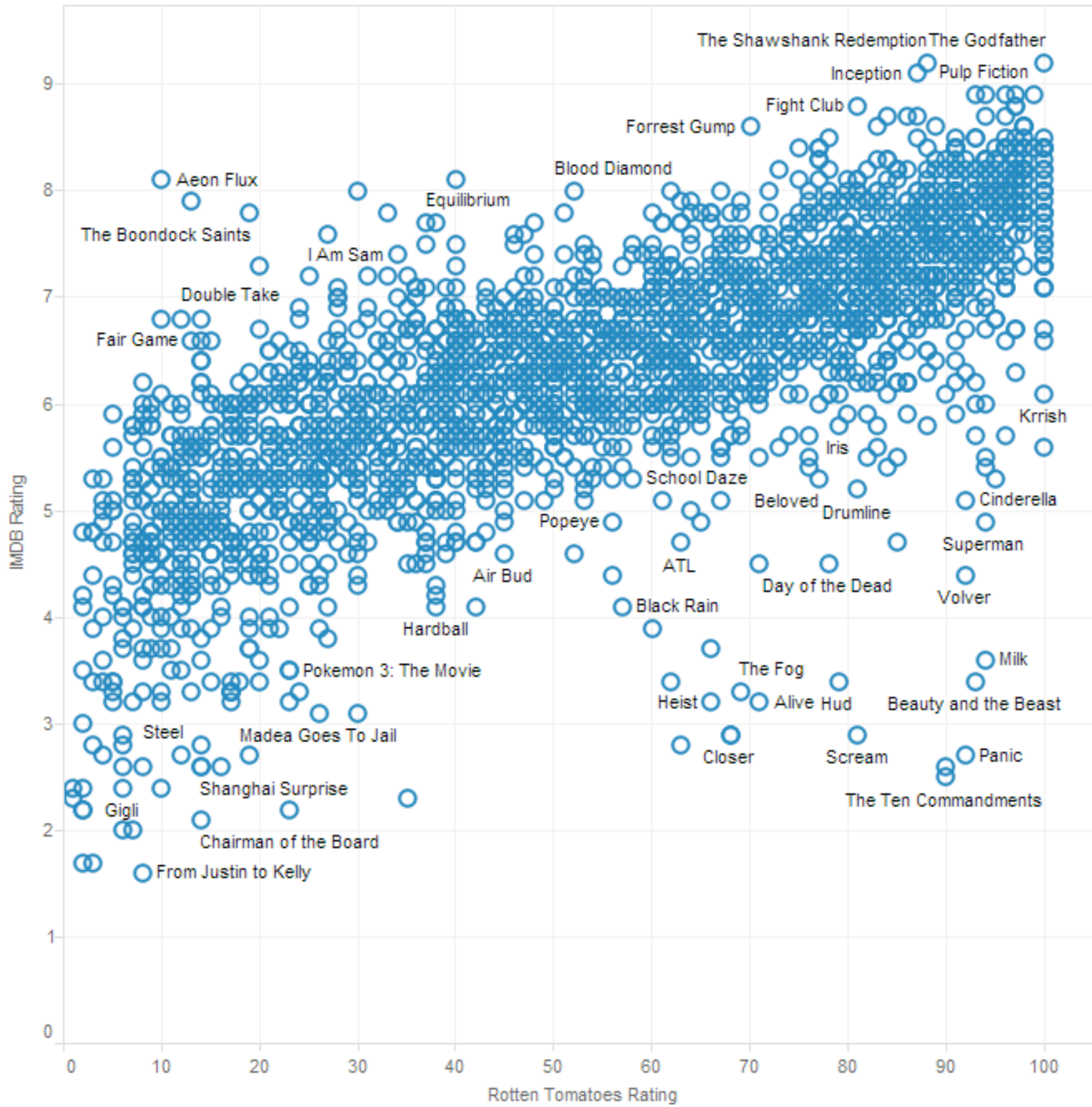
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)


IMDB Rating (bin)
Number of Records
0
50
100
150
200
250
300
350
400
450
500
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

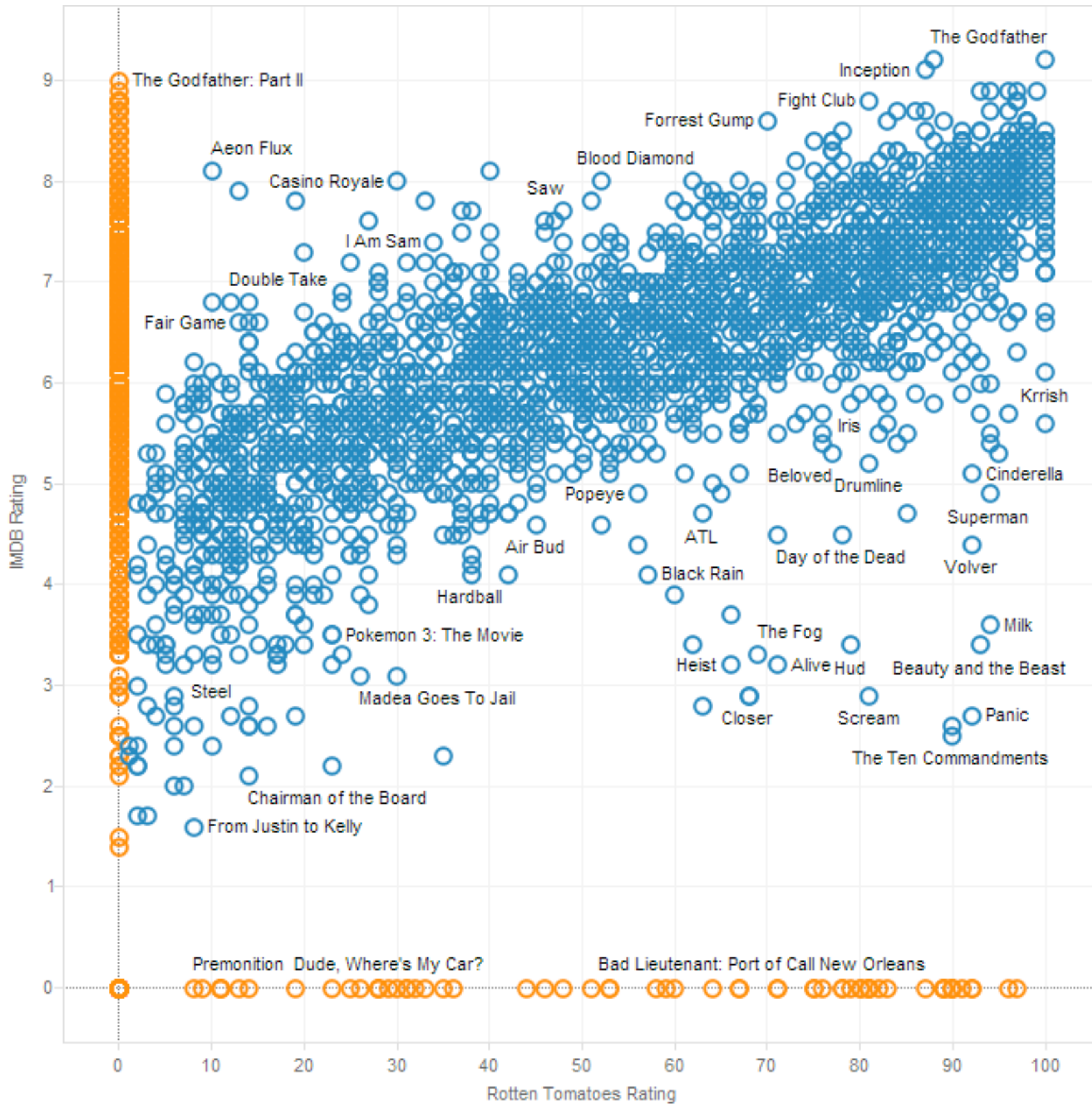

Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

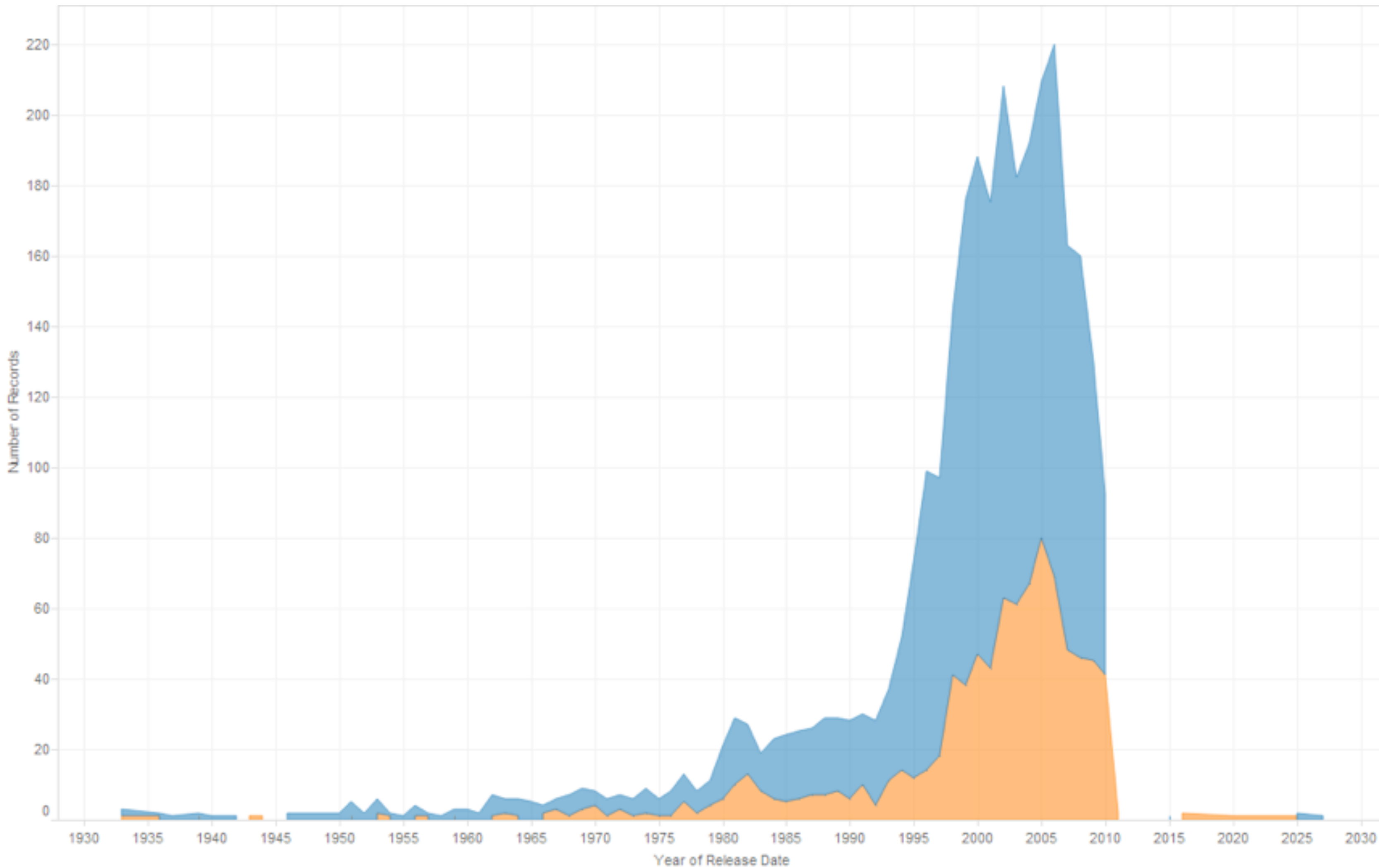


IMDB Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100
Rotten Tomatoes Rating



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
ATL
Air Bud
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
Rotten Tomatoes Rating


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
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
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)

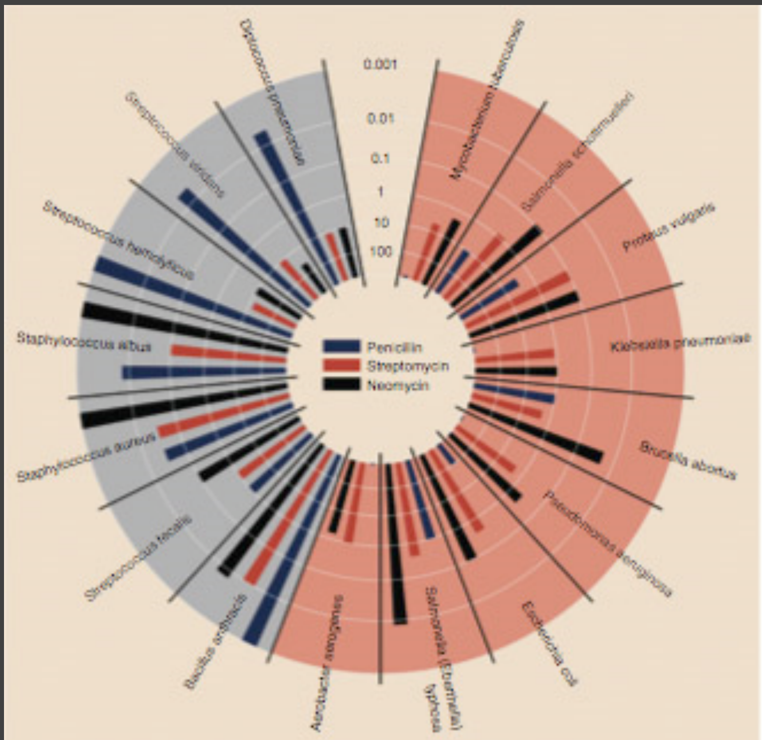
Collected prior to 1951.

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

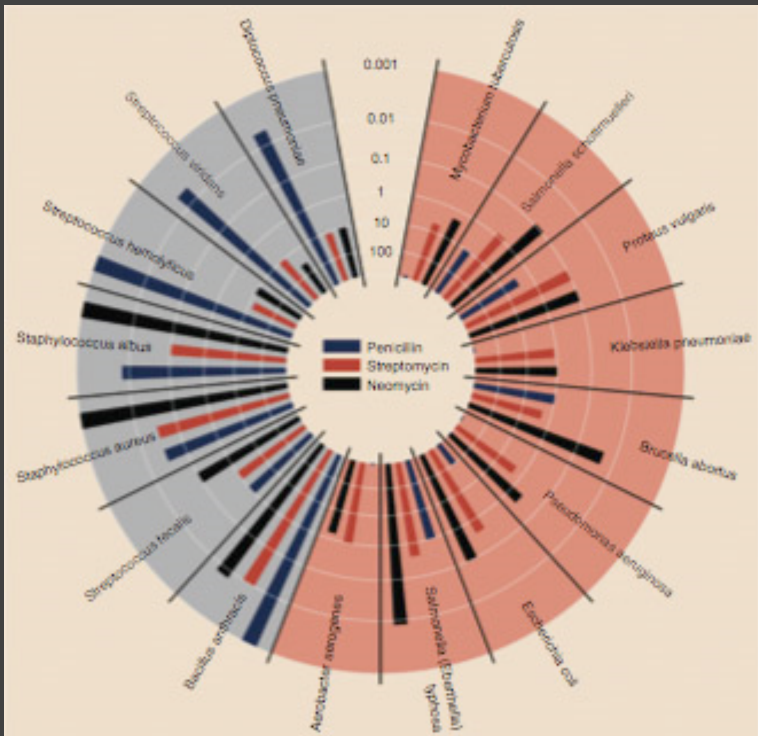
How do the drugs compare?



Bacteria	Antibiotic			Gram stain
	Penicillin	Streptomycin	Neomycin	
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

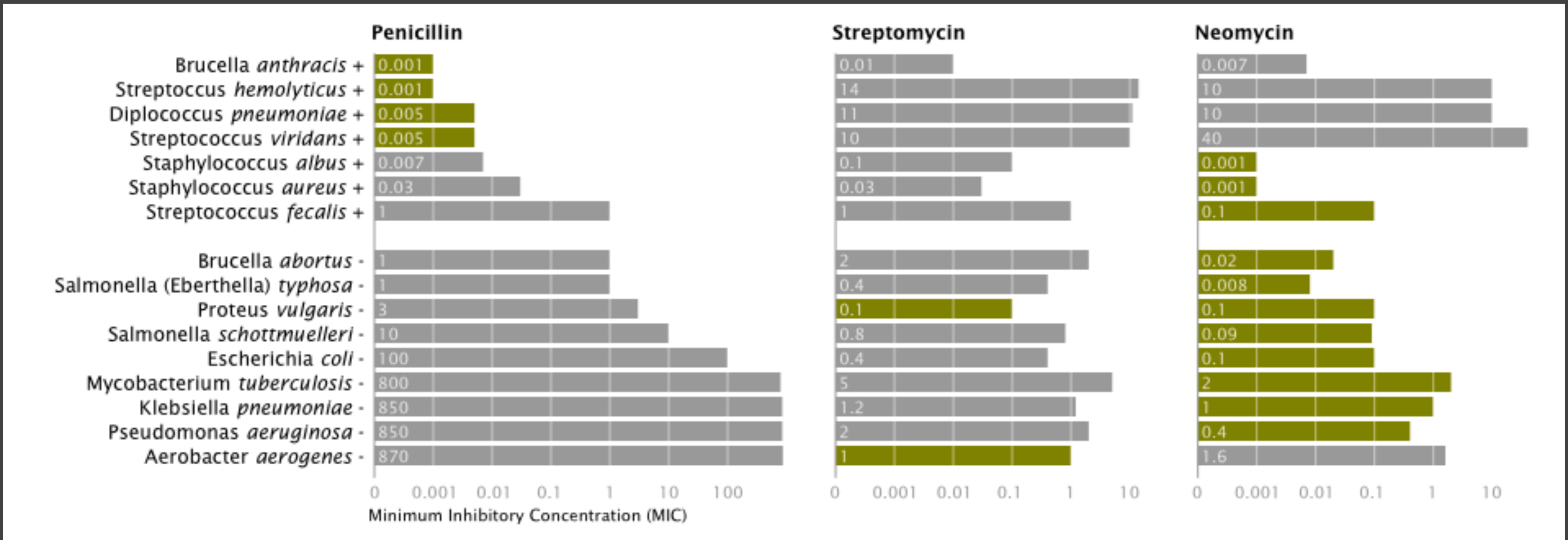
How do the drugs compare?



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	+

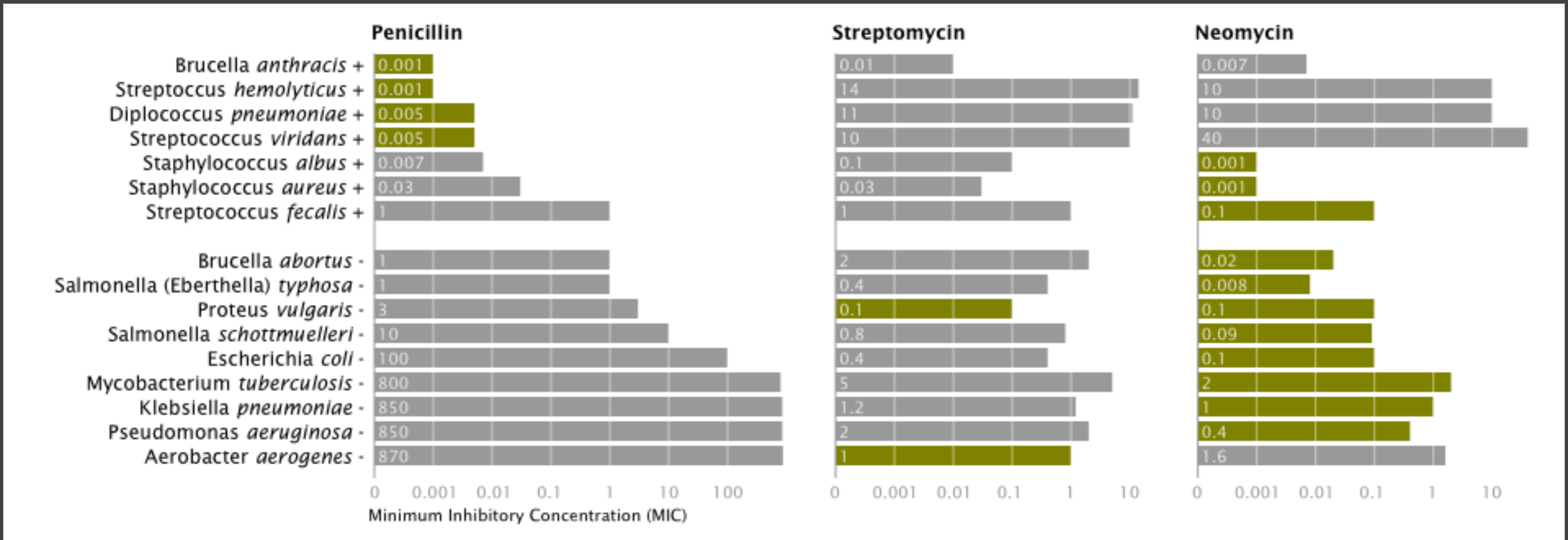
Radius: 1 / log(MIC)
Bar Color: Antibiotic
Background Color: Gram Staining

How do the drugs compare?



Mike Bostock
Stanford CS448B, Winter 2009

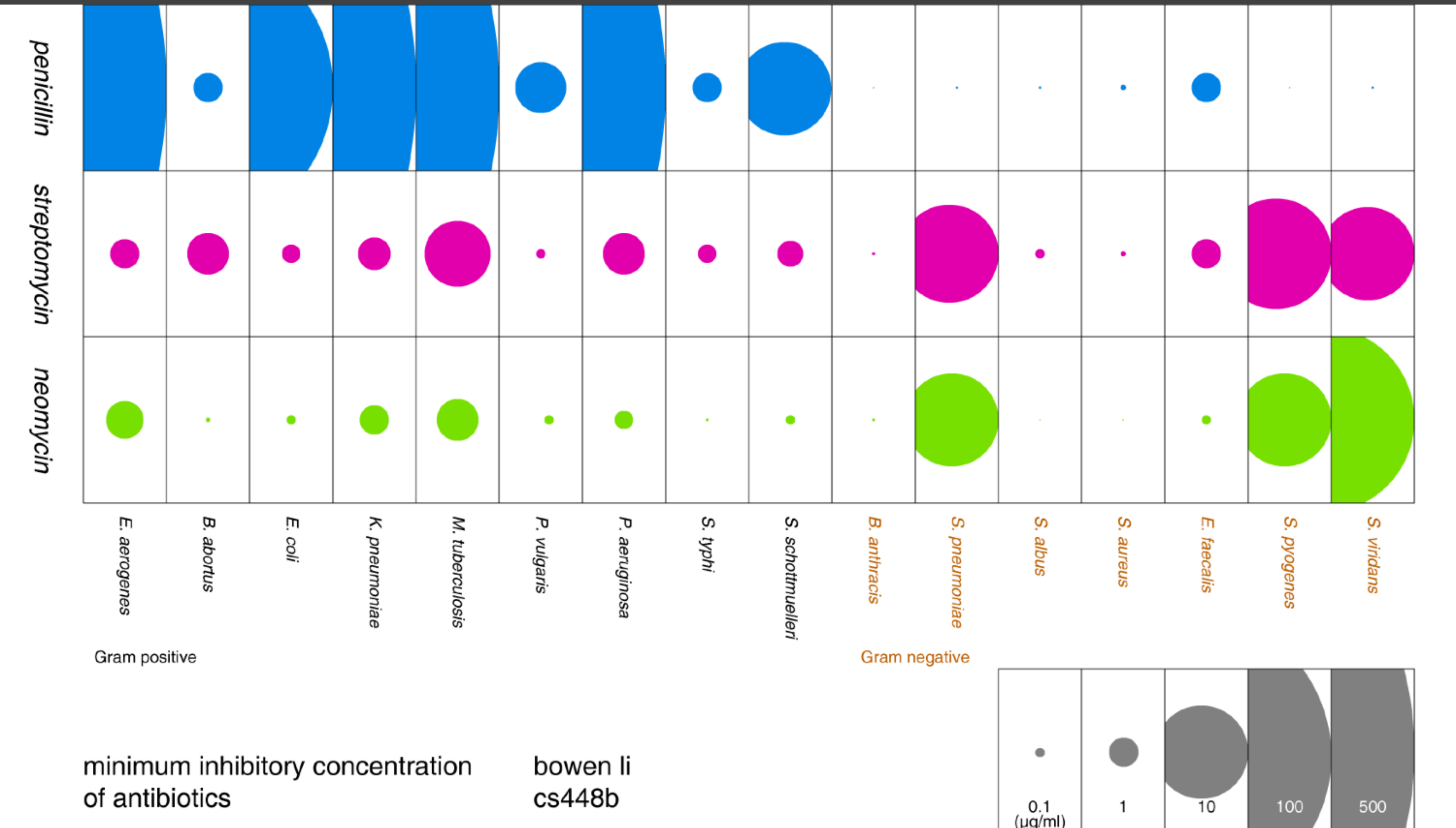
How do the drugs compare?



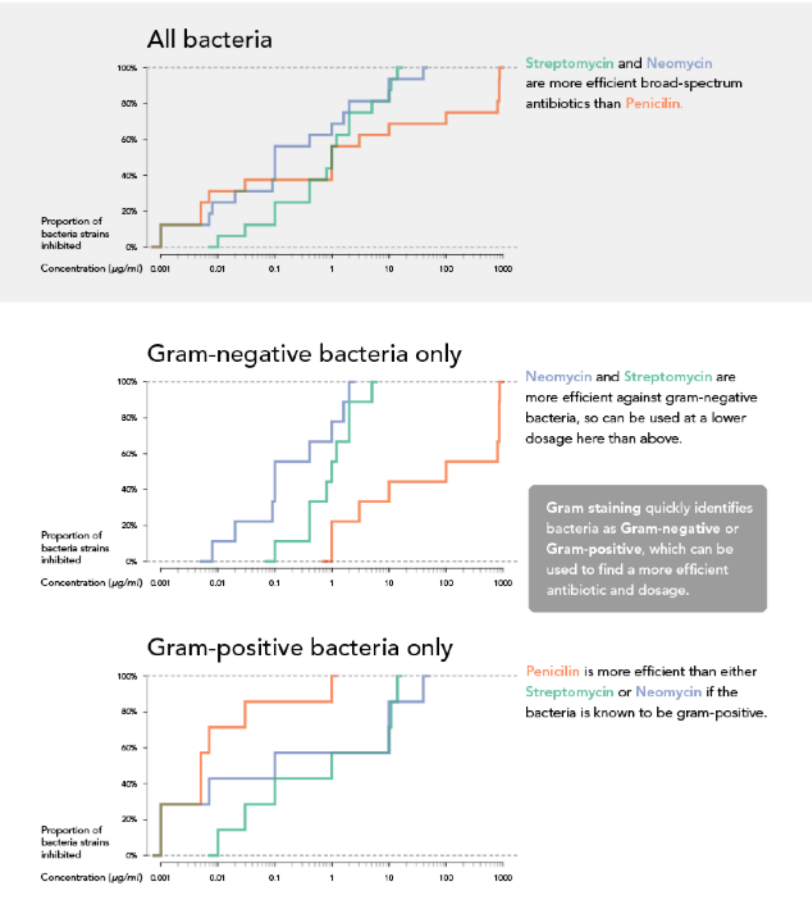
X-axis: Antibiotic | log(MIC)

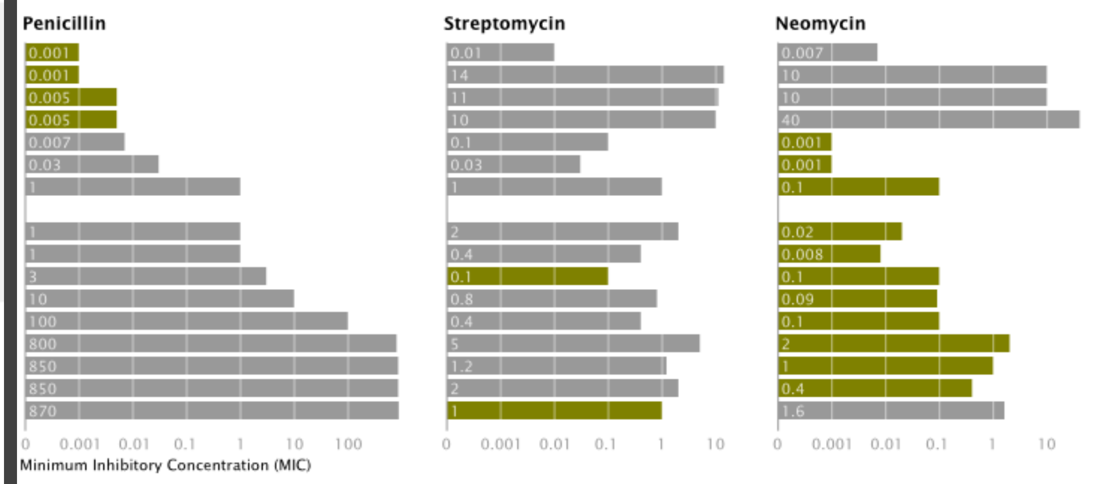
Y-axis: Gram-Staining | Species

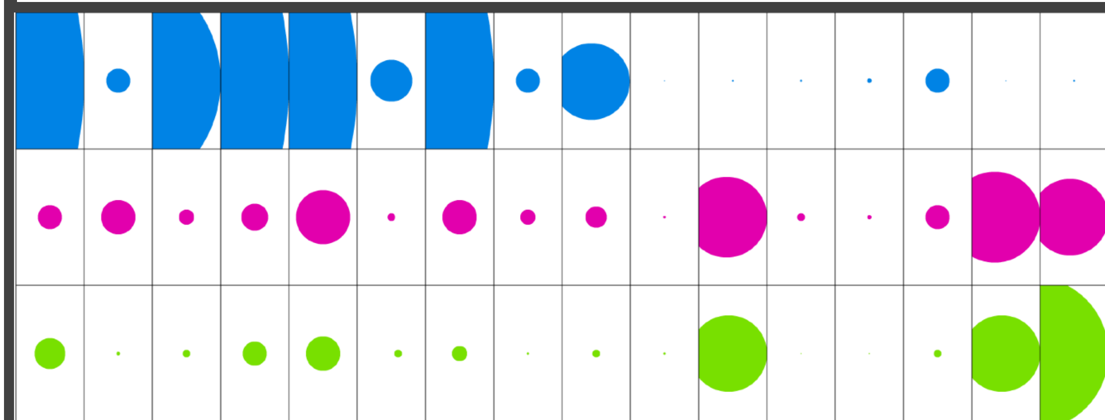
Color: Most-Effective?

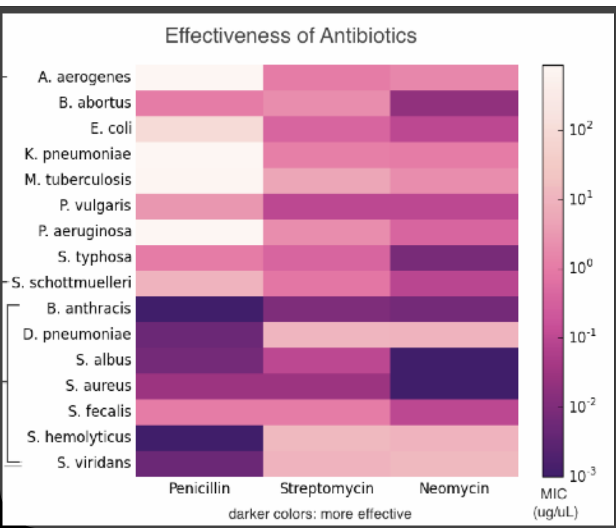


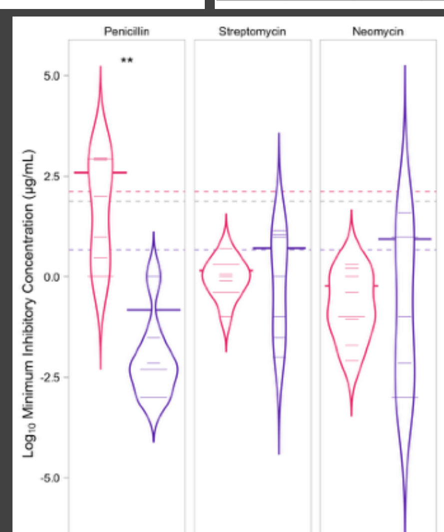
Bowen Li
Stanford CS448B, Fall 2009

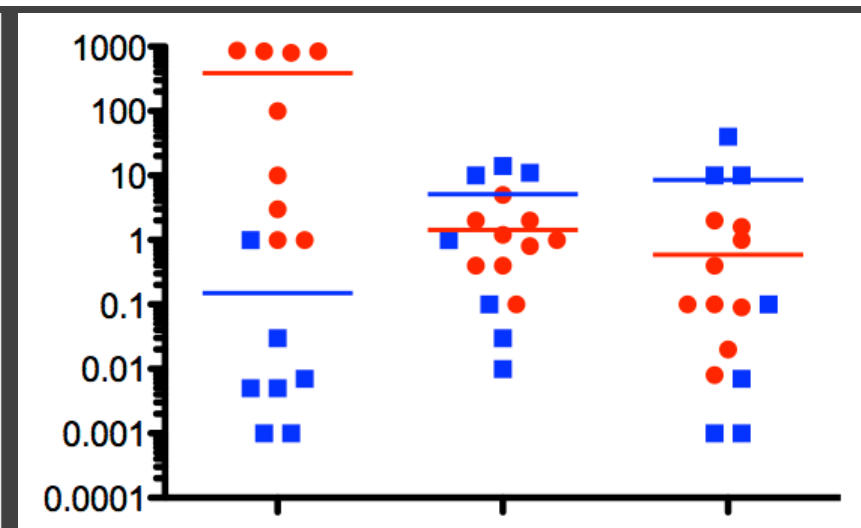

All bacteria
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.
Gram-negative bacteria only
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.
Gram staining quickly identifies bacteria as 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.
Proportion of bacteria strains inhibited
Concentration (µg/ml)
0.001
0.01
0.1
1
10
100
1000
100%
80%
60%
40%
20%
0%


Penicillin
Streptomycin
Neomycin
0.001
0.001
0.005
0.005
0.007
0.03
1
1
1
3
10
100
800
850
850
870
0.01
14
11
10
0.1
0.03
1
2
0.4
0.1
0.8
0.4
5
1.2
2
1
0.007
10
10
40
0.001
0.001
0.1
0.02
0.008
0.1
0.09
0.1
2
1
0.4
1.6
Minimum Inhibitory Concentration (MIC)

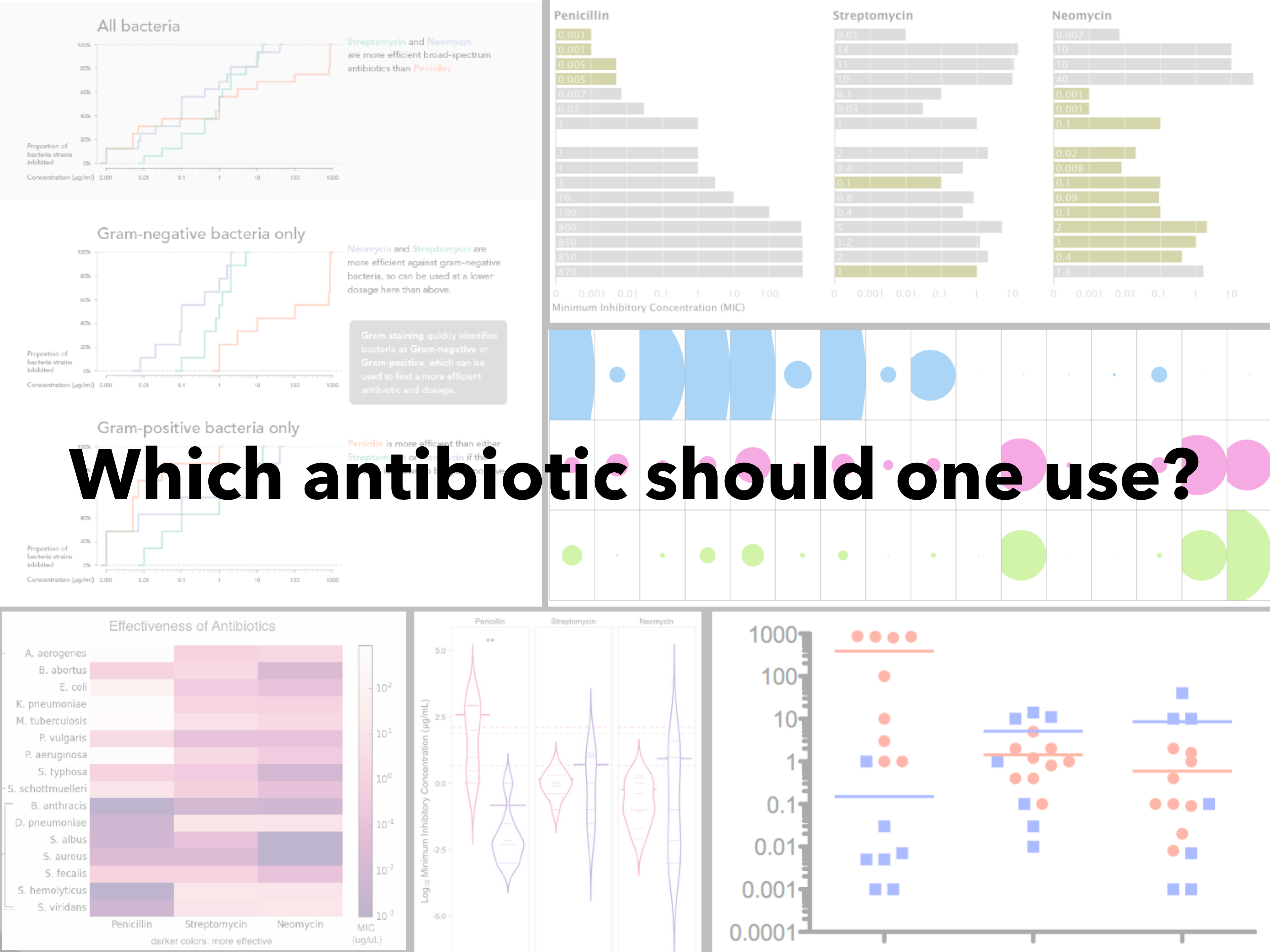



Effectiveness of Antibiotics
A. aerogenes
B. abortus
E. coli
K. pneumoniae
M. tuberculosis
P. vulgaris
P. aeruginosa
S. typhosa
S. schottmuelleri
B. anthracis
D. pneumoniae
S. albus
S. aureus
S. fecalis
S. hemolyticus
S. viridans
Penicillin
Streptomycin
Neomycin
darker colors: more effective
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
MIC (ug/uL)

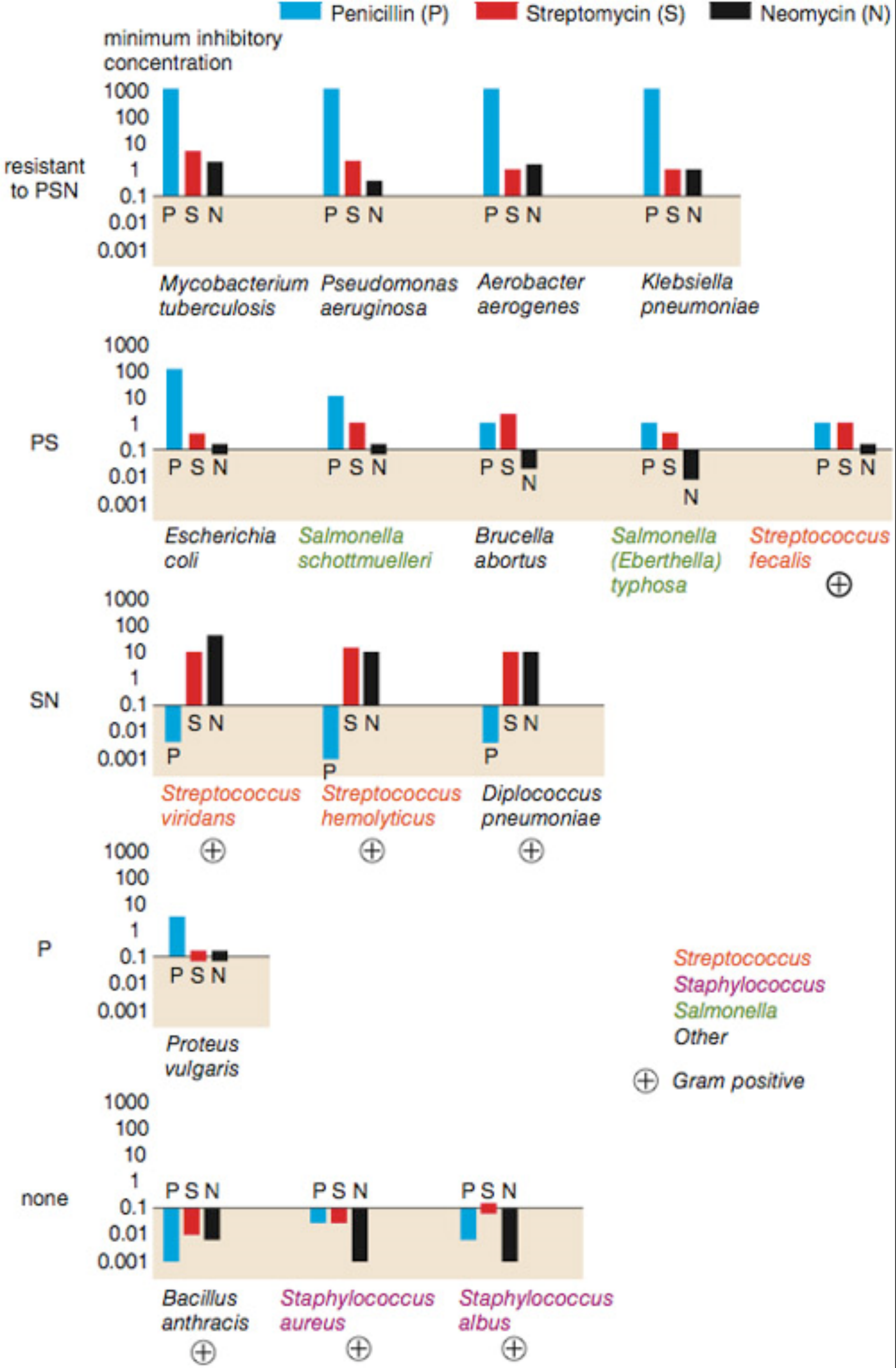

Penicillin
Streptomycin
Neomycin
**
Log_{10} Minimum Inhibitory Concentration (µg/mL)
5.0
2.5
0.0
-2.5
-5.0


1000
100
10
1
0.1
0.01
0.001
0.0001

Which antibiotic should one use?

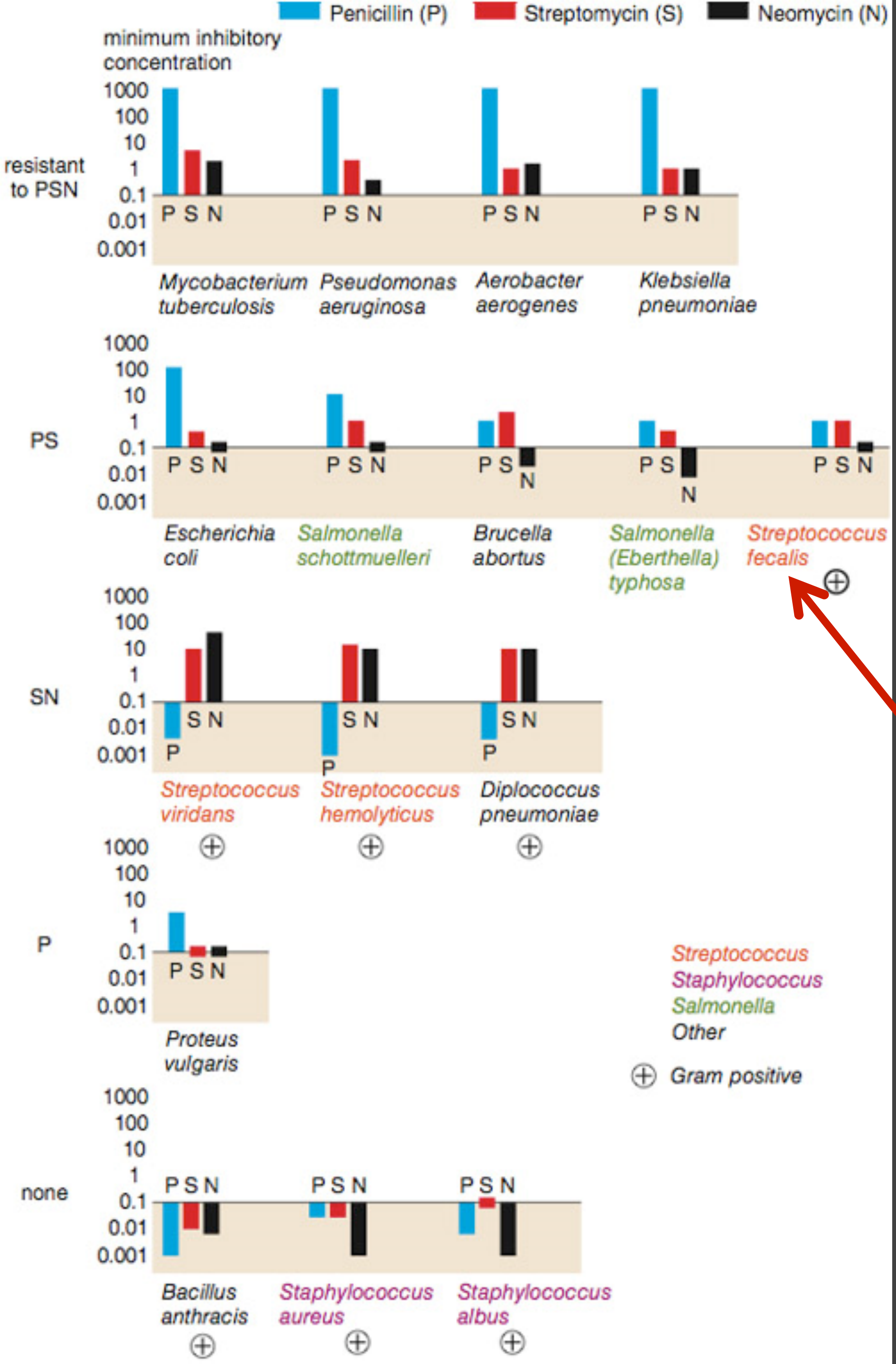


Do the bacteria group by antibiotic resistance?



Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009

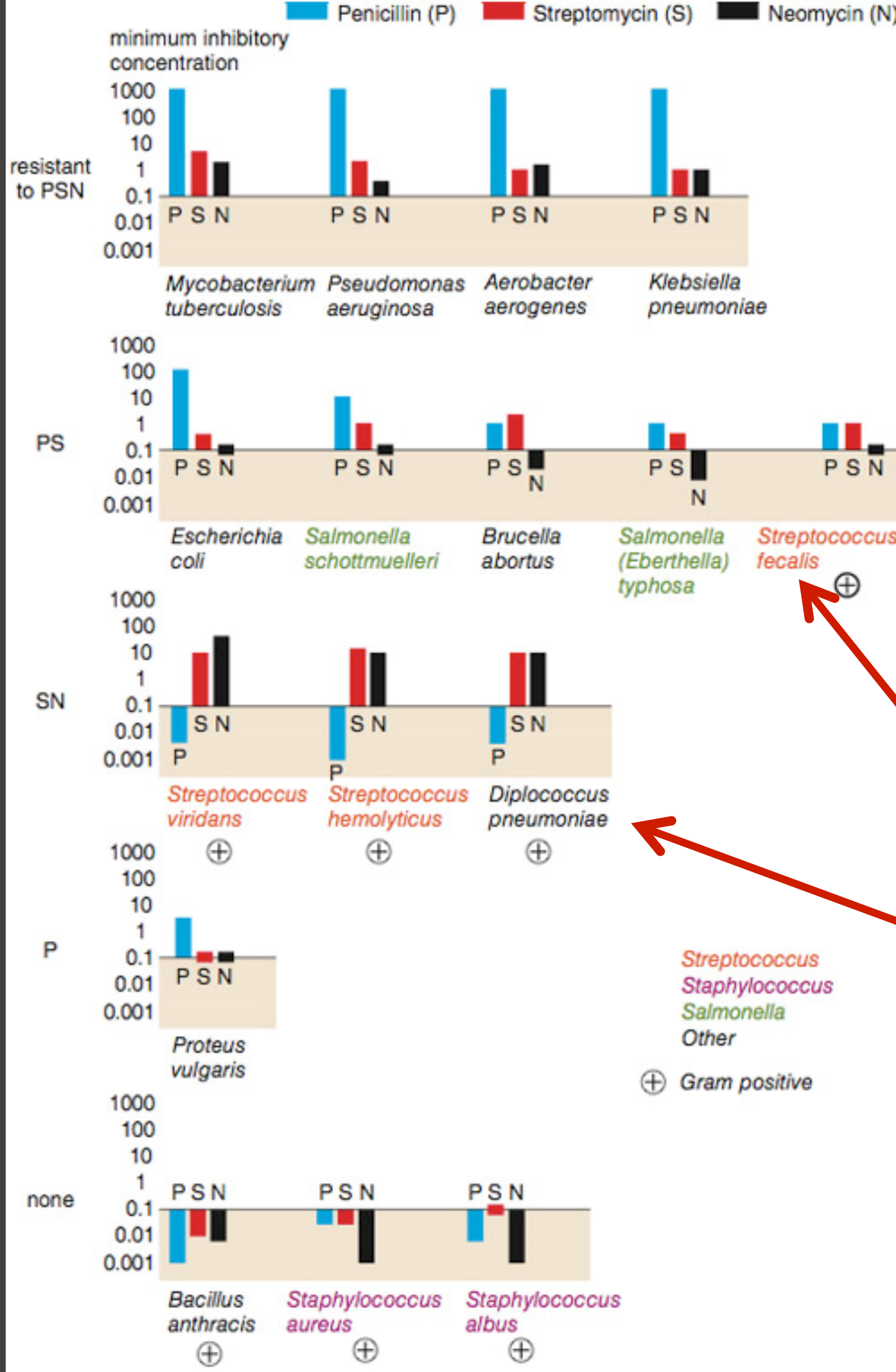


Do the bacteria group by antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Wainer & Lysen
American Scientist, 2009

Do the bacteria group by antibiotic resistance?

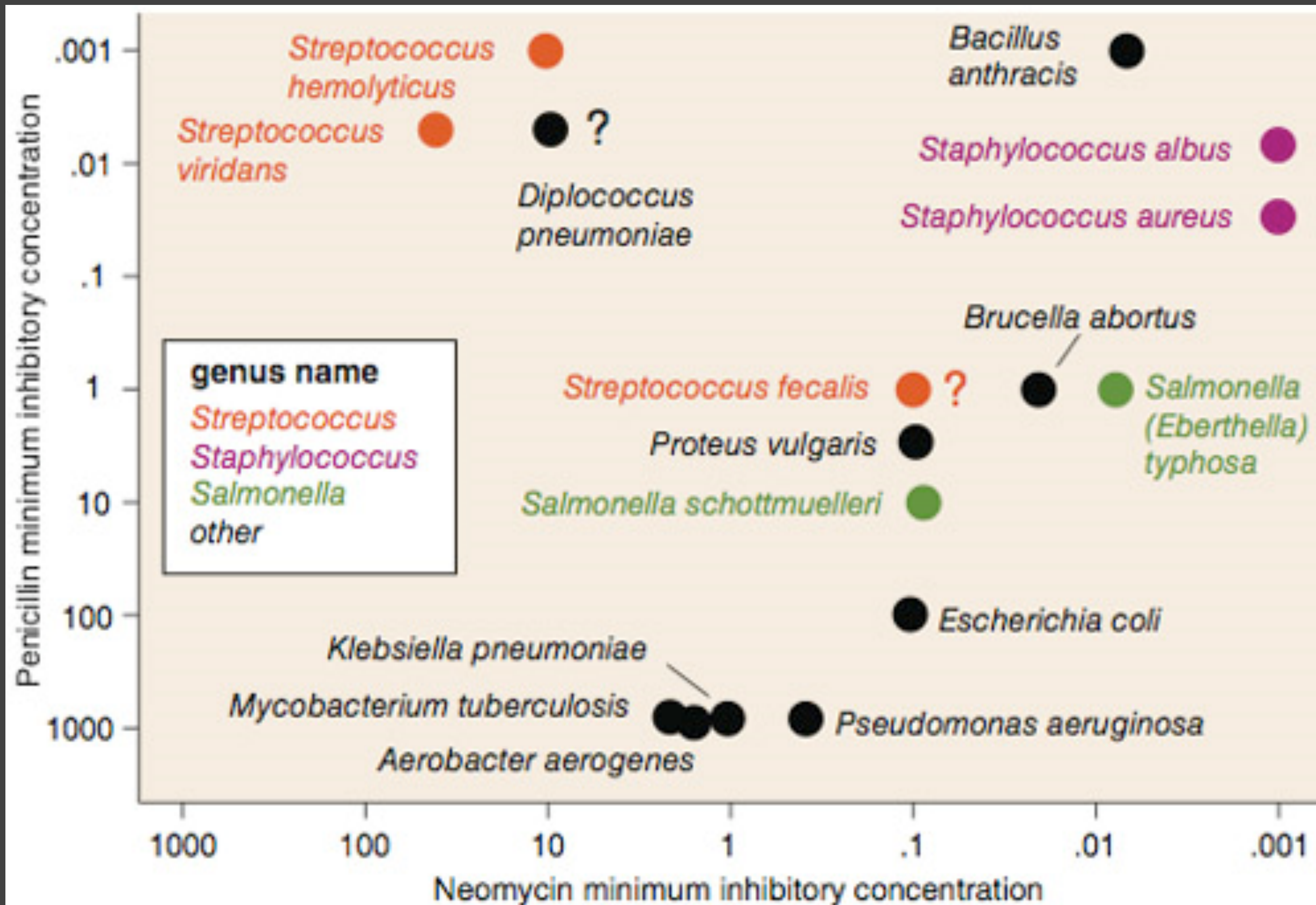


Not a streptococcus!
(realized ~30 yrs later)

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

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

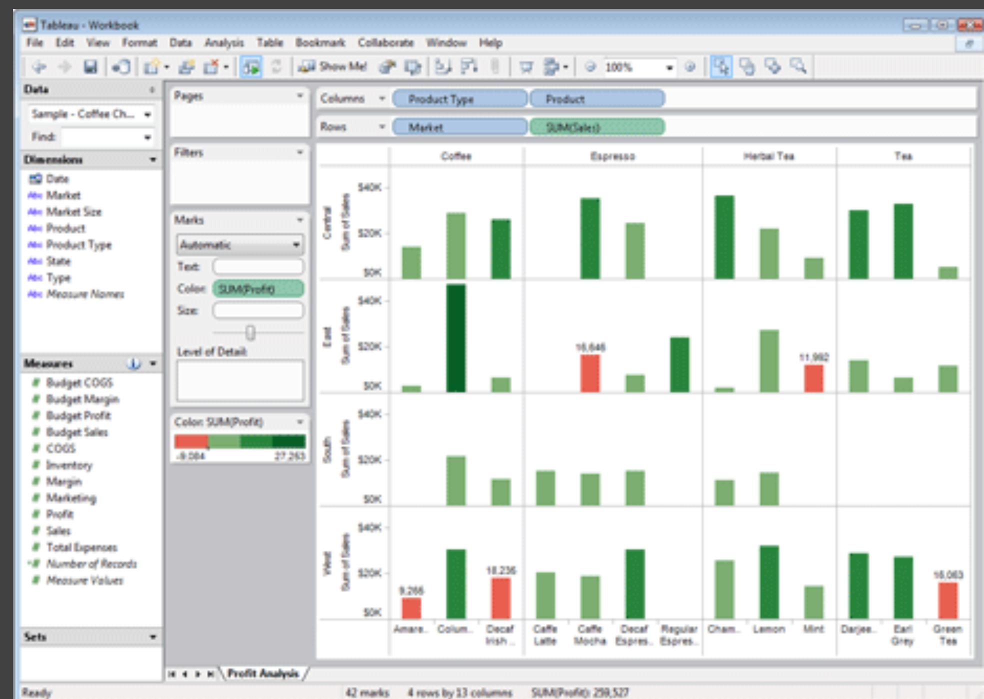
First steps:

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+)*

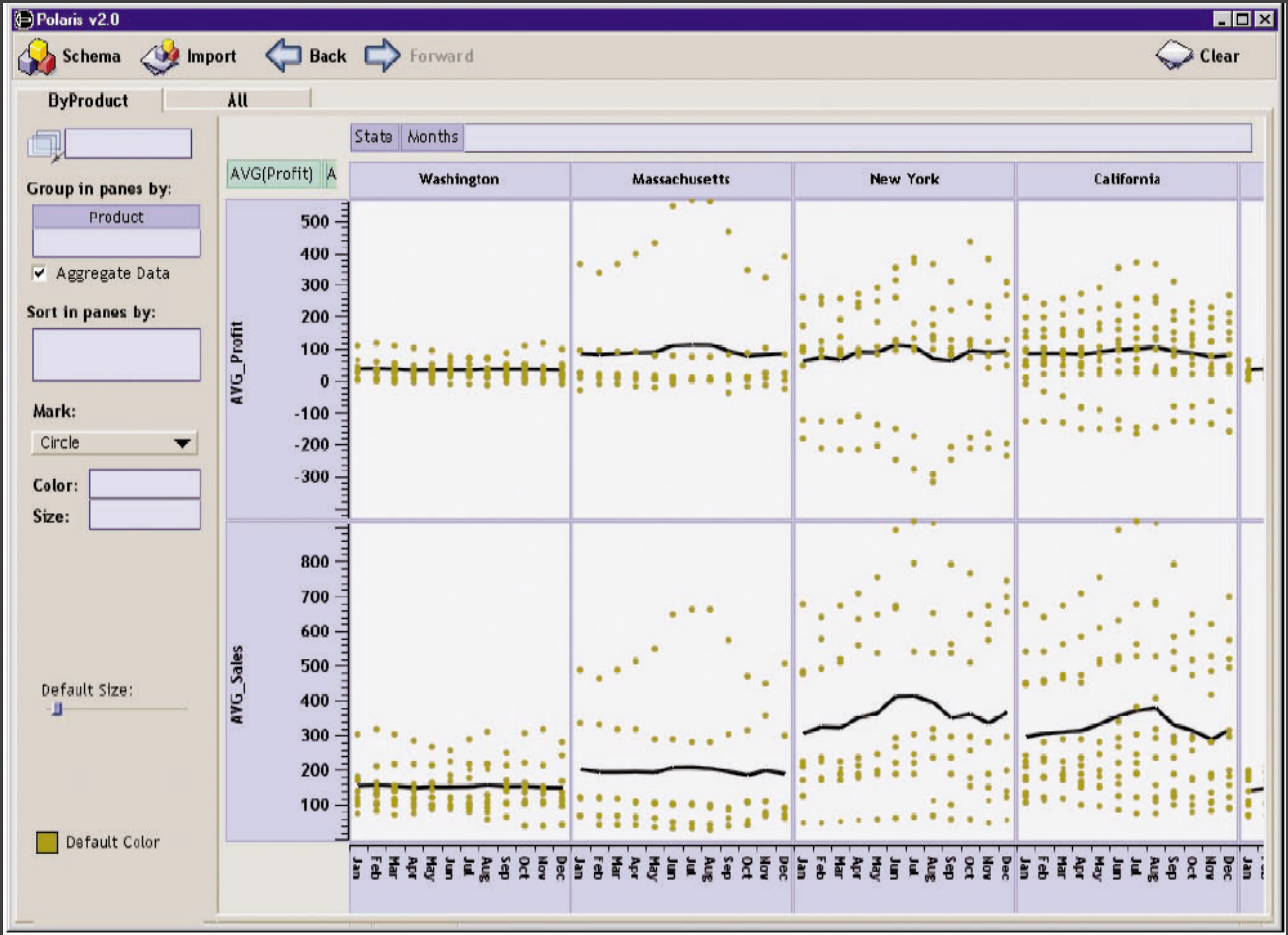
Include titles and captions for each view

Due by 5:00pm

Monday, Oct 16

Tableau / Polaris

Polaris [Stolte et al.]



Tableau

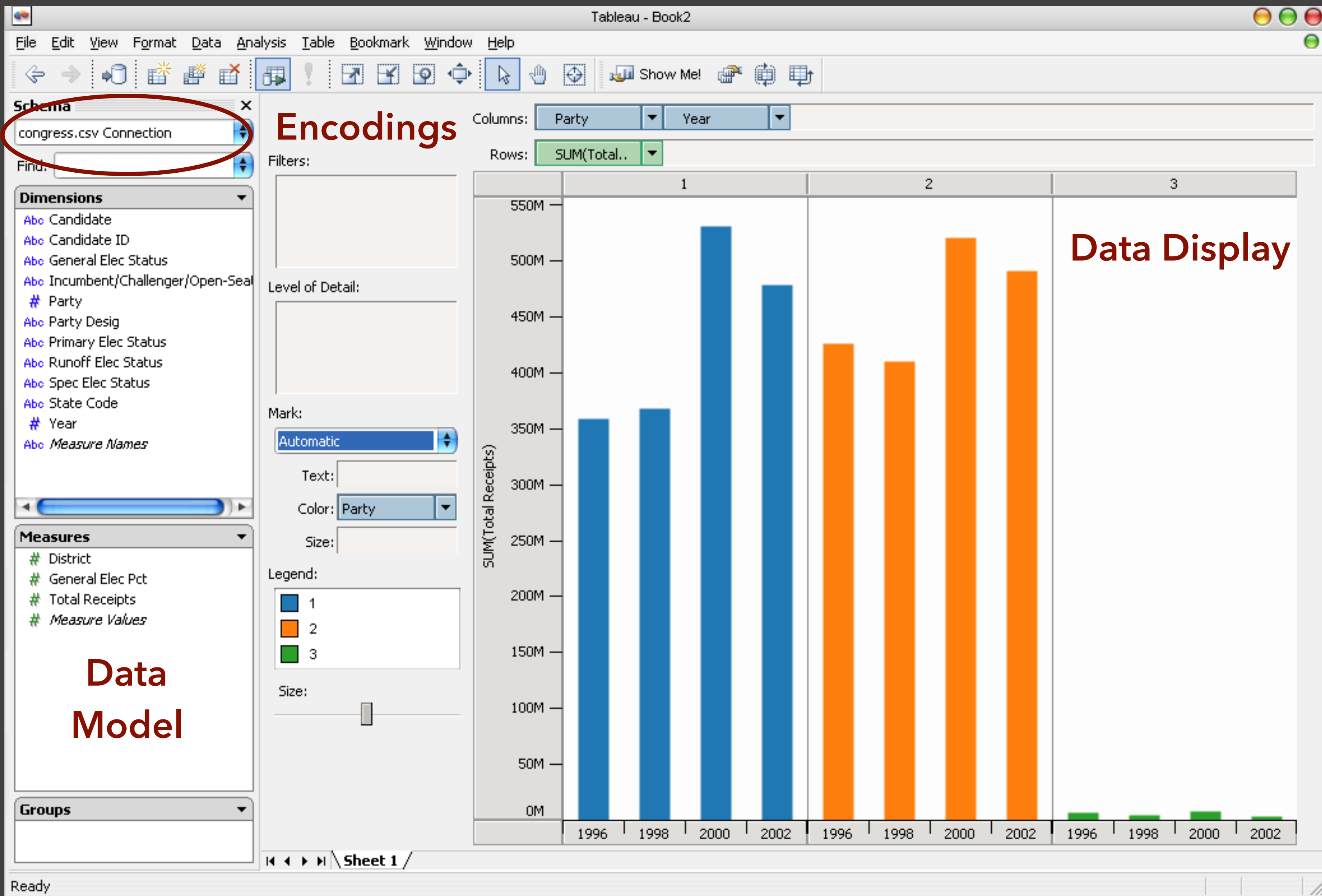


Tableau - Book2
File Edit View Format Data Analysis Table Bookmark Window Help
Show Me!
Schema
congress.csv Connection
Find:
Dimensions
Candidate
Candidate ID
General Elec Status
Incumbent/Challenger/Open-Seat
Party
Party Desig
Primary Elec Status
Runoff Elec Status
Spec Elec Status
State Code
Year
Measure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

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

Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest encodings upon request

Specifying Table Configurations

Operands are the database fields

Each operand interpreted as a set {…}

Quantitative and Ordinal fields treated differently

Three operators:

concatenation (+)

cross product (x)

nest (/)

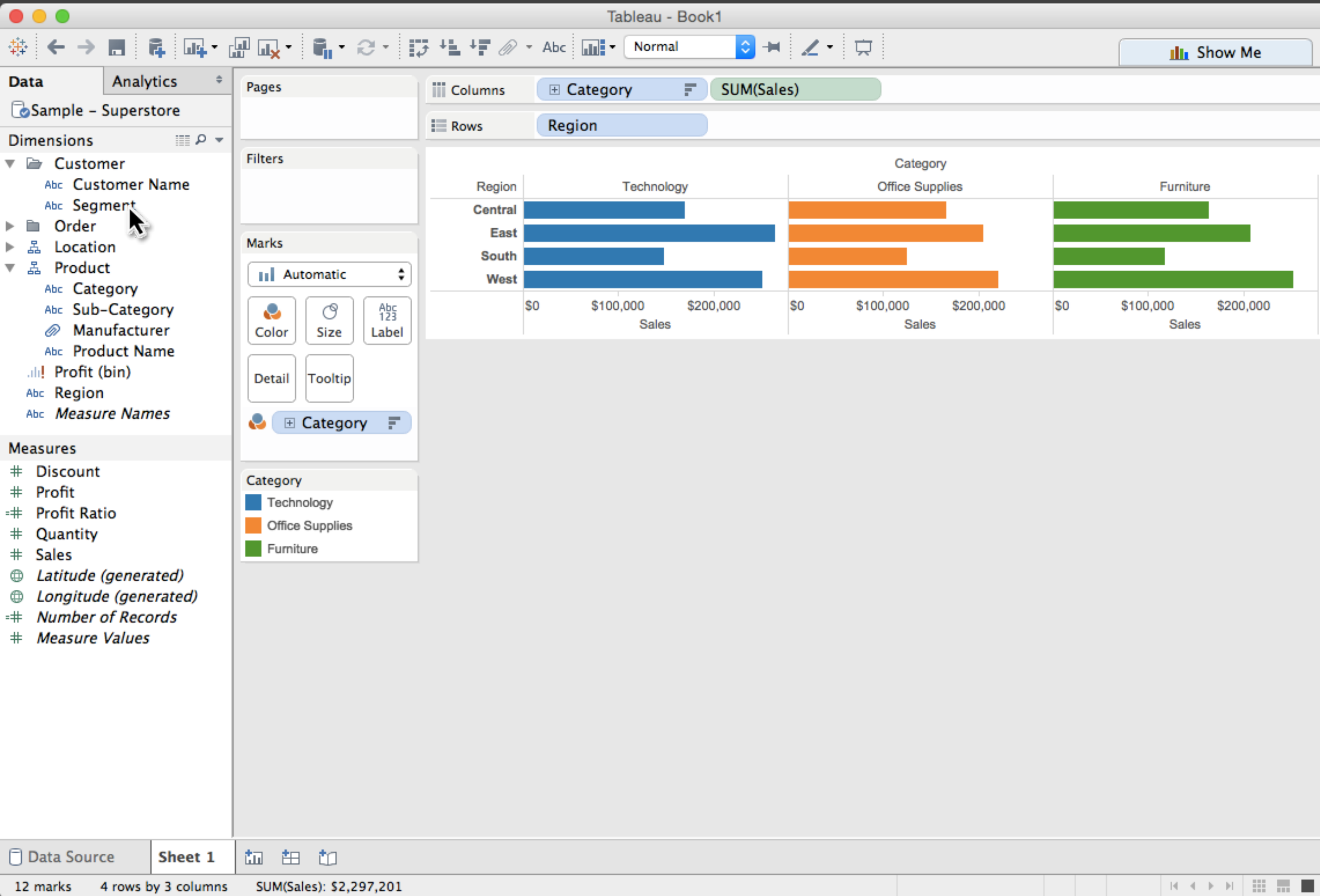


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Category
Region
Technology
Office Supplies
Furniture
Central
East
South
West
$0
$100,000
$200,000
Sales
Data Source
Sheet 1
12 marks
4 rows by 3 columns
SUM(Sales): $2,297,201

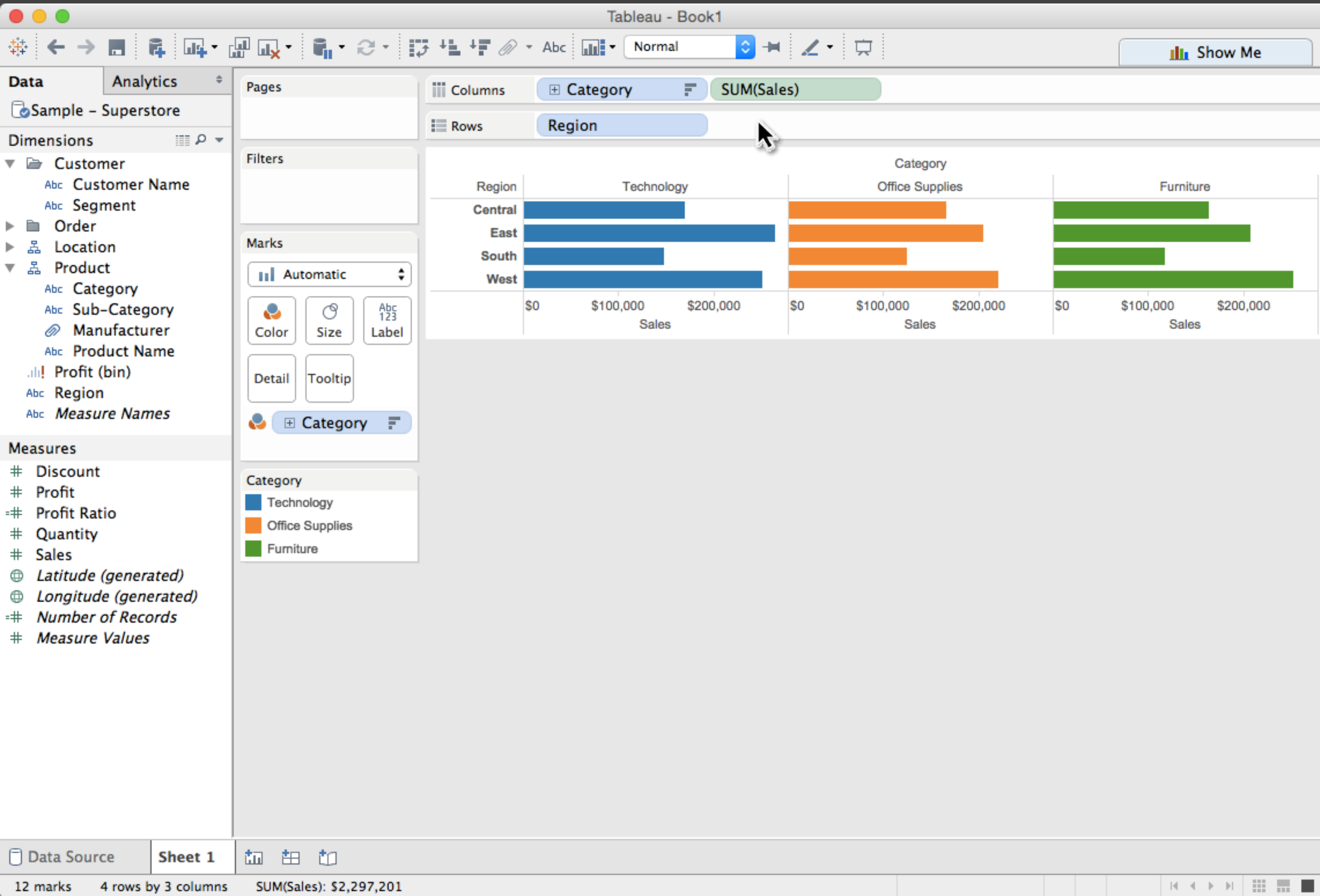


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample - Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
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
Category
Region
Technology
Office Supplies
Furniture
Central
East
South
West
$0
$100,000
$200,000
Sales
Data Source
Sheet 1
12 marks
4 rows by 3 columns
SUM(Sales): $2,297,201

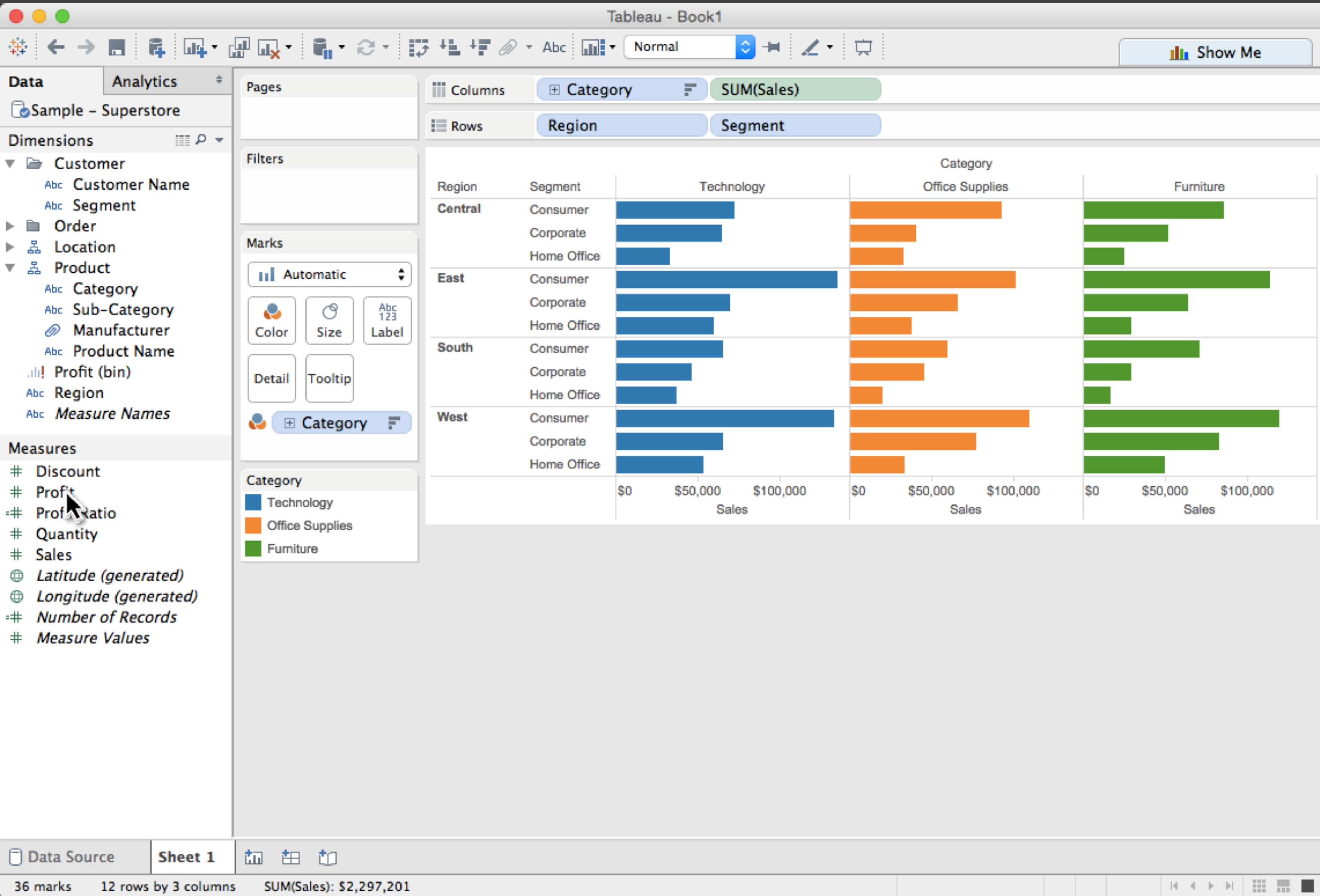


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample - Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
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
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$50,000
$100,000
Sales
Data Source
Sheet 1
36 marks
12 rows by 3 columns
SUM(Sales): $2,297,201

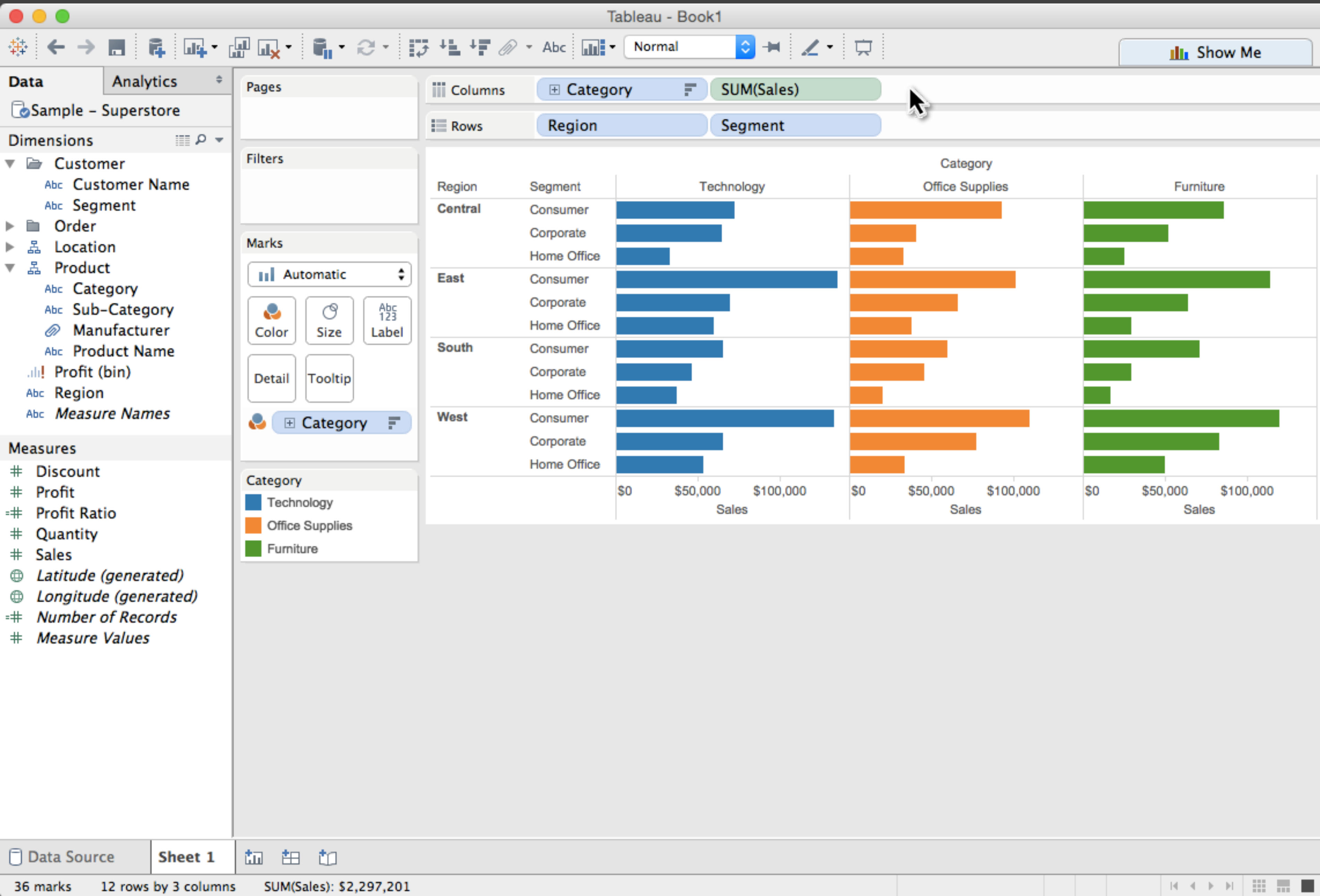


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
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
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$50,000
$100,000
Sales
Data Source
Sheet 1
36 marks
12 rows by 3 columns
SUM(Sales): $2,297,201

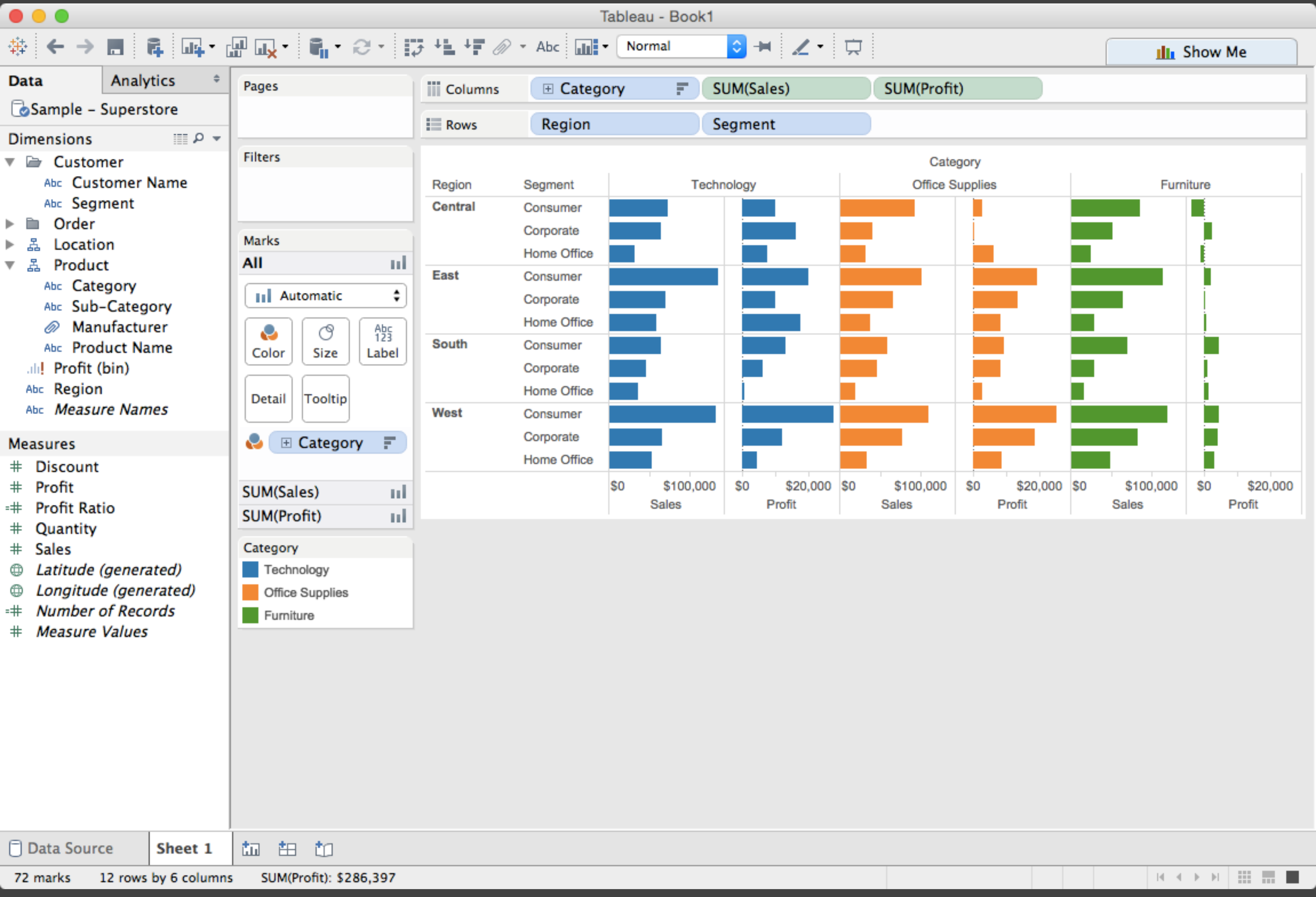


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
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
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$100,000
$20,000
Sales
Profit
Data Source
Sheet 1
72 marks
12 rows by 6 columns
SUM(Profit): $286,397

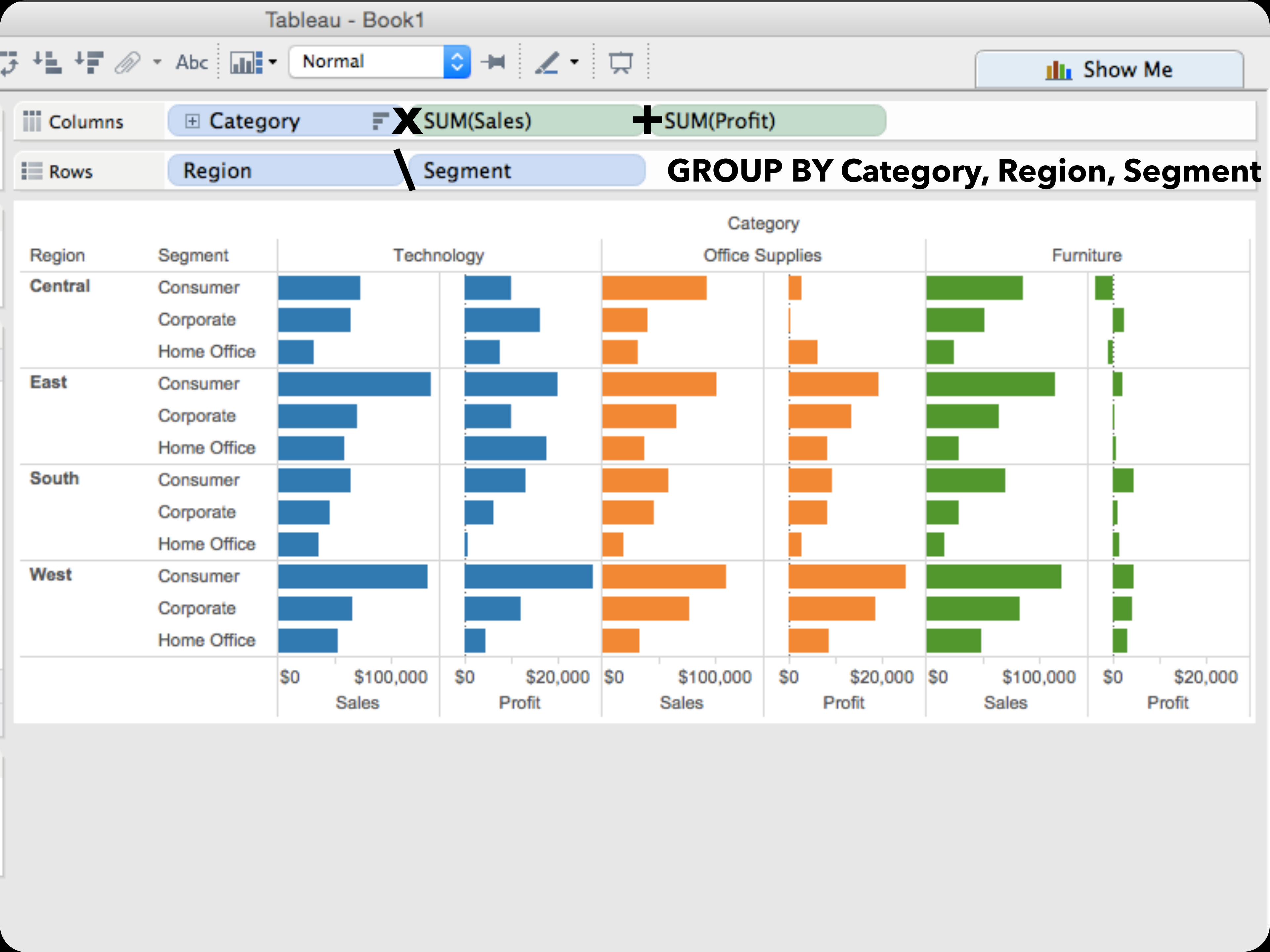


Tableau - Book1
Abc
Normal
Show Me
Columns
Category
X
SUM(Sales)
+
SUM(Profit)
Rows
Region
\
Segment
GROUP BY Category, Region, Segment
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
Consumer
Corporate
Home Office
East
Consumer
Corporate
Home Office
South
Consumer
Corporate
Home Office
West
Consumer
Corporate
Home Office
$0
$100,000
Sales
$0
$20,000
Profit

Table Algebra: Operands

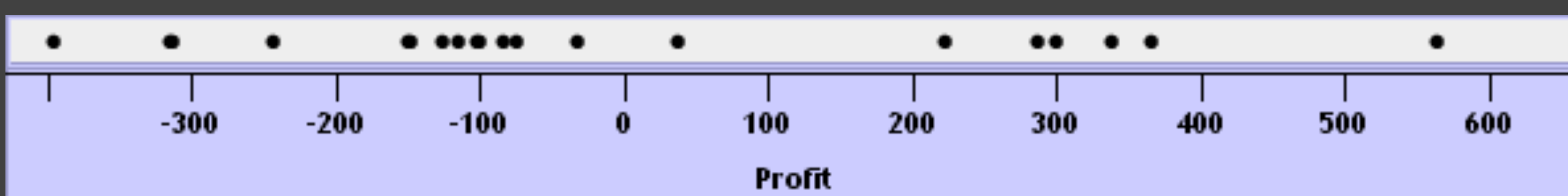
Ordinal fields: interpret domain as a set that partitions table into rows and columns.

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

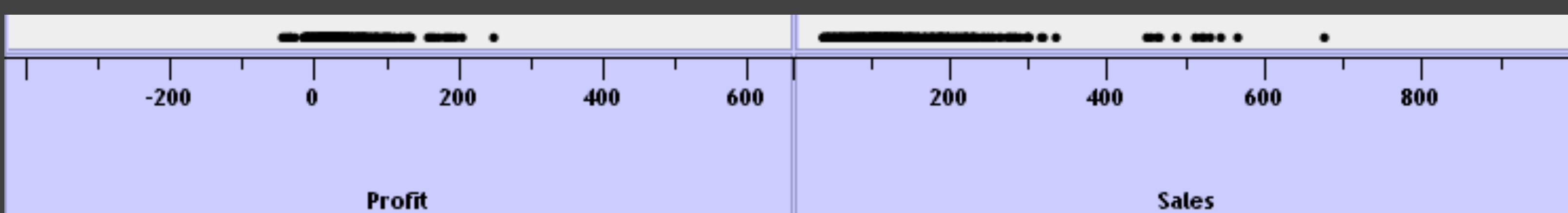
Quarter + Product Type

= {(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

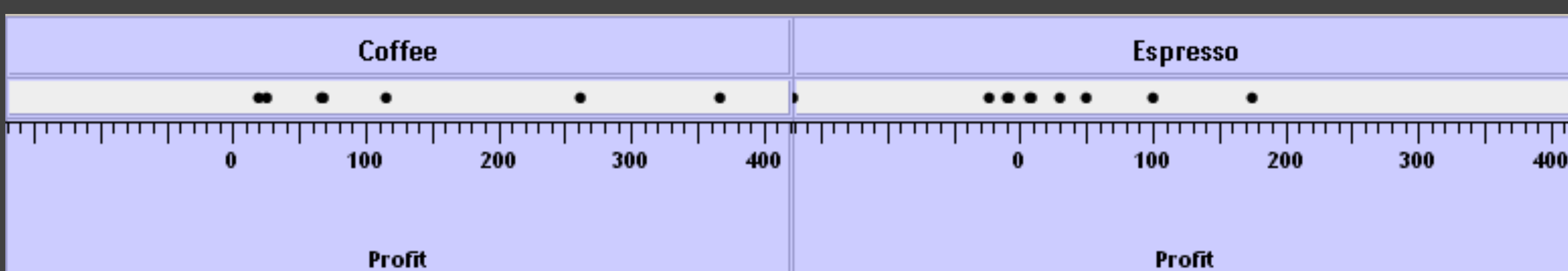
Cross-product of set interpretations

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)

Table Algebra

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

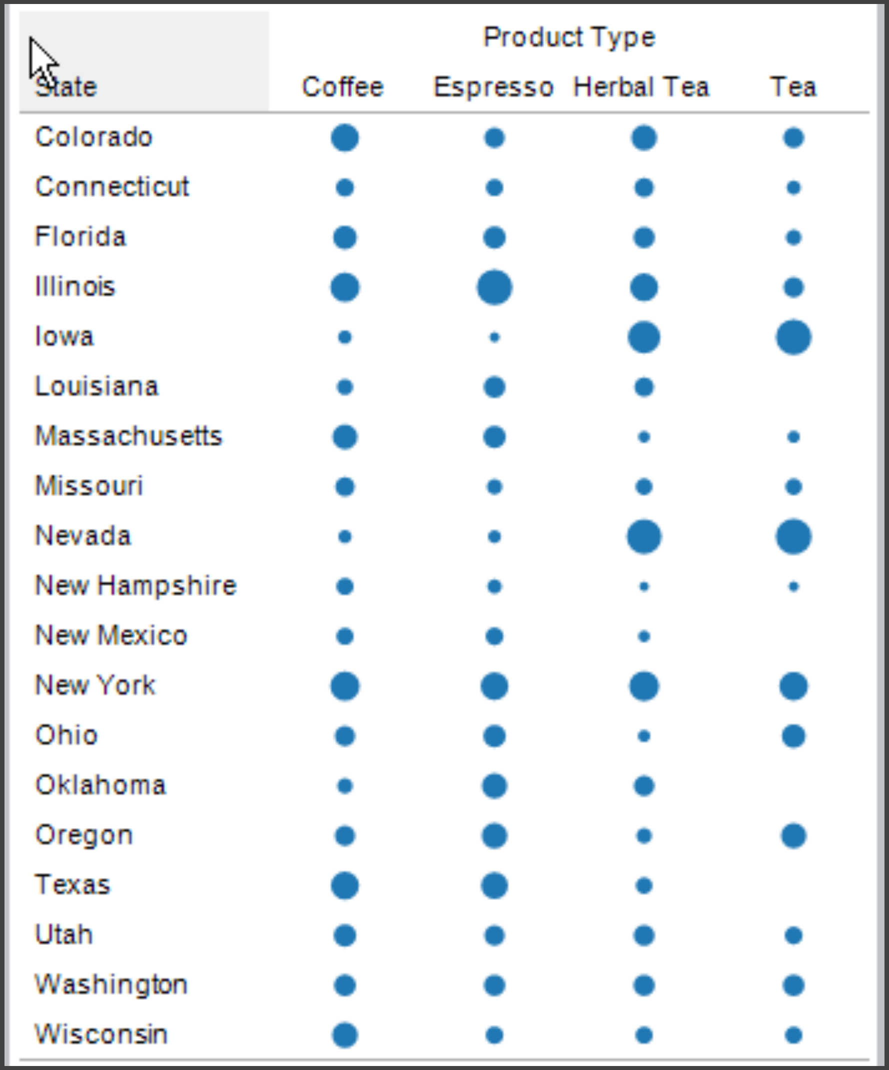
Algebraic statements are then mapped to:

- **Visualizations** - trellis plot partitions, visual encodings
- **Queries** - selection, projection, group-by aggregation

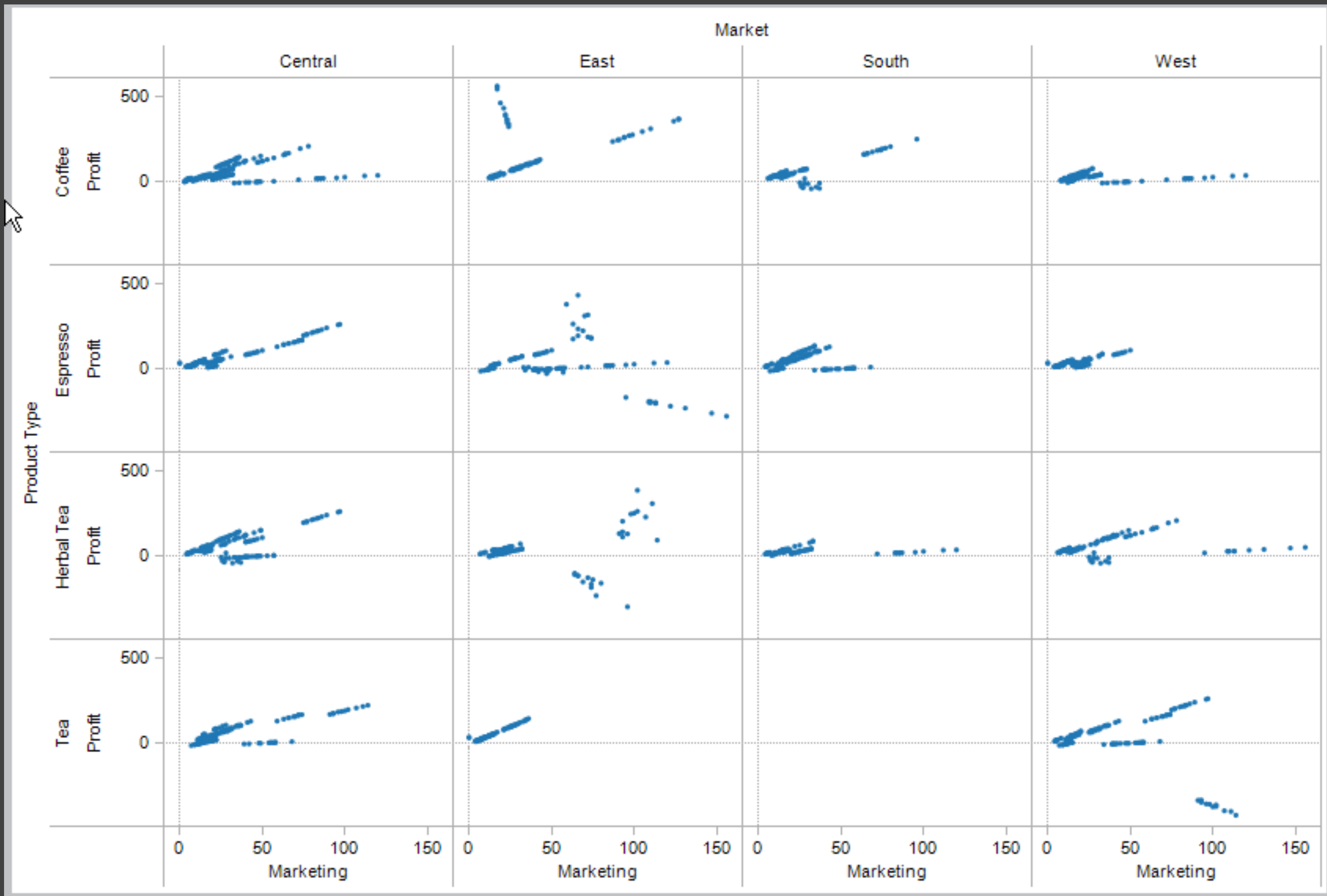
In Tableau, users make statements via drag-and-drop

- Note that this specifies operands *NOT* operators!
- Operators are inferred by data type (O, Q)

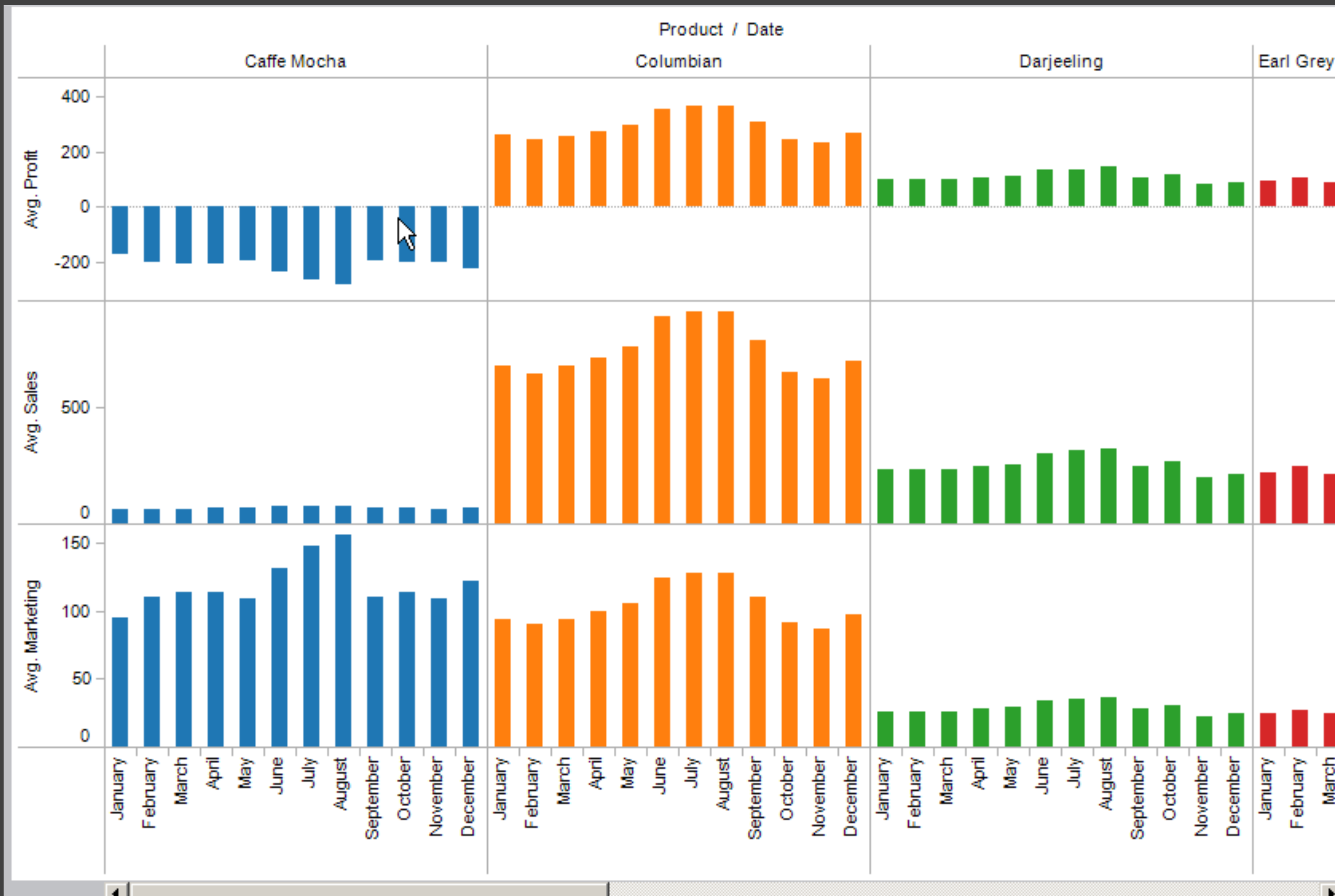
Ordinal-Ordinal



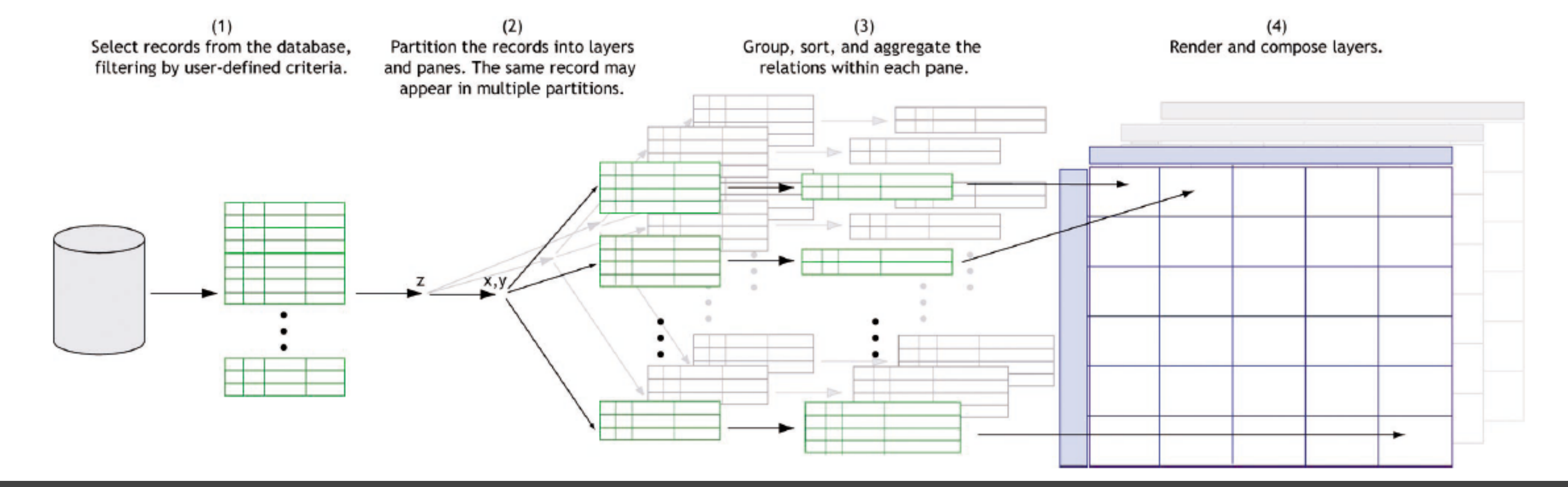
Quantitative-Quantitative



Ordinal-Quantitative



Querying the Database



BONUS TOPIC

Data Fraud

A Detective Story

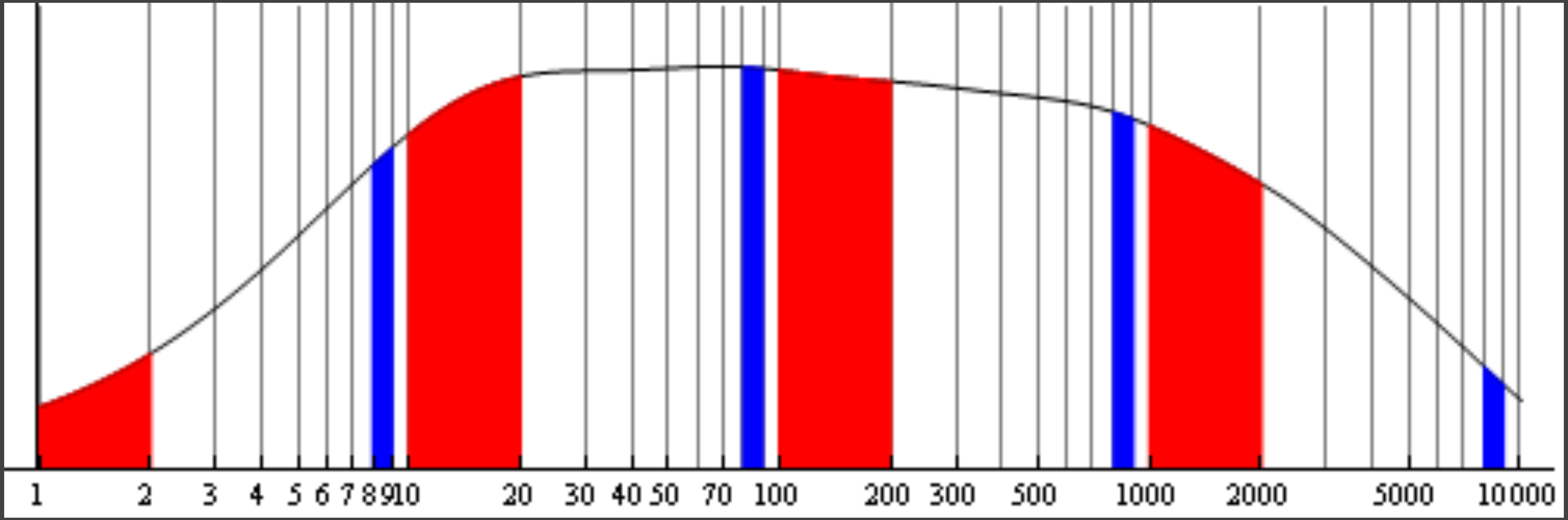
You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

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!