

Exploratory Data Analysis

Maneesh Agrawala

CS 448B: Visualization
Winter 2020

1

A2: Exploratory Data Analysis

Use **Tableau** to formulate & answer questions

First steps

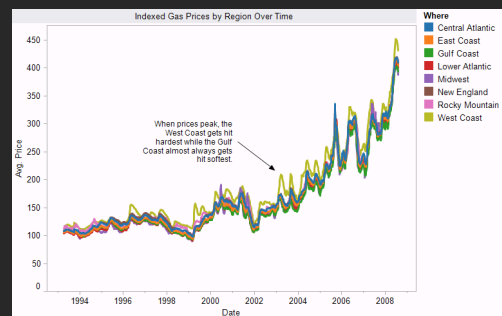
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile data
- Iterate as needed

Create visualizations

- Interact with data
- Refine questions

Author a report

- Screenshots of most insightful views (10+)
- Include titles and captions for each view



Due before class on Jan 27, 2020

2

Exploratory Data Analysis

3

The Rise of Statistics (1900-1950s)

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

Little innovation in graphical methods

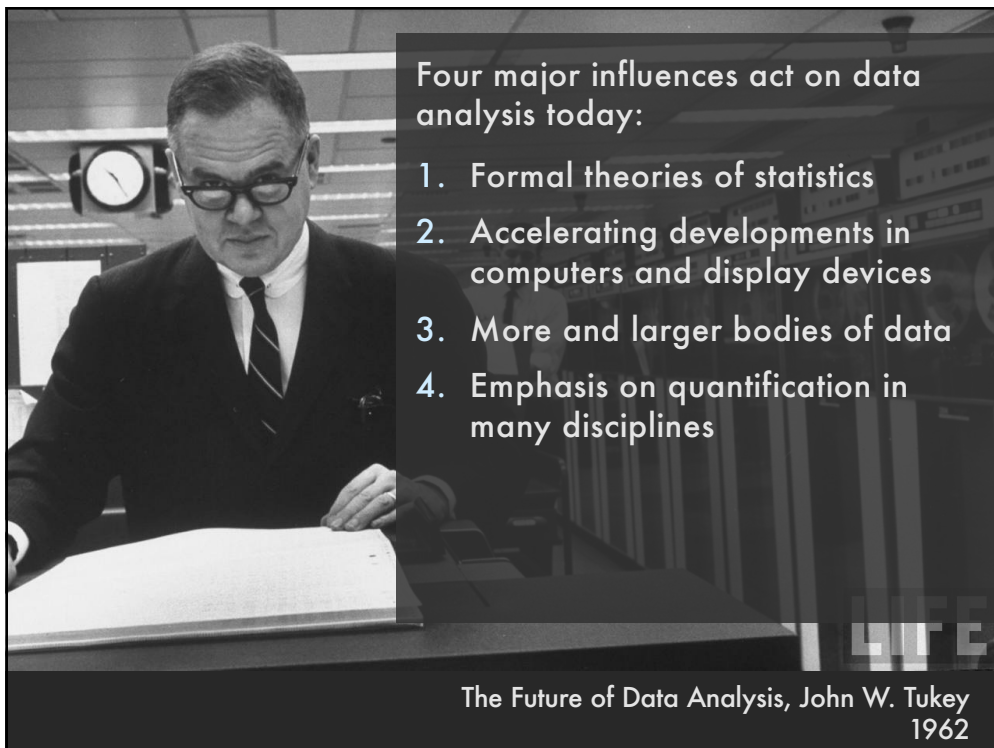
A period of **application** and **popularization**

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

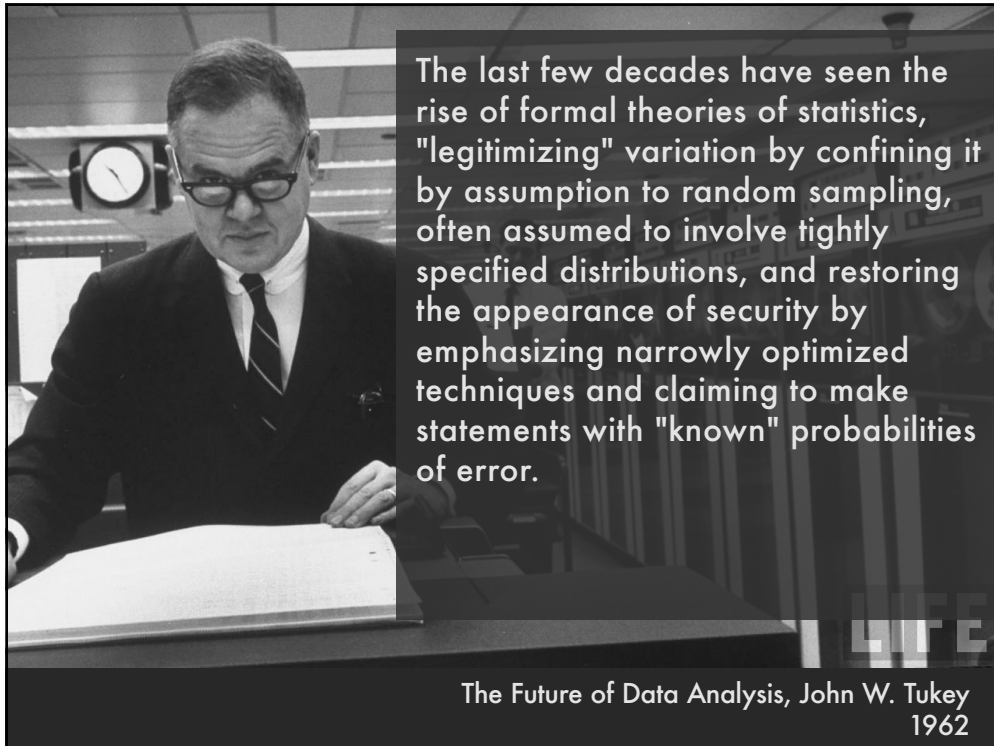
4



5



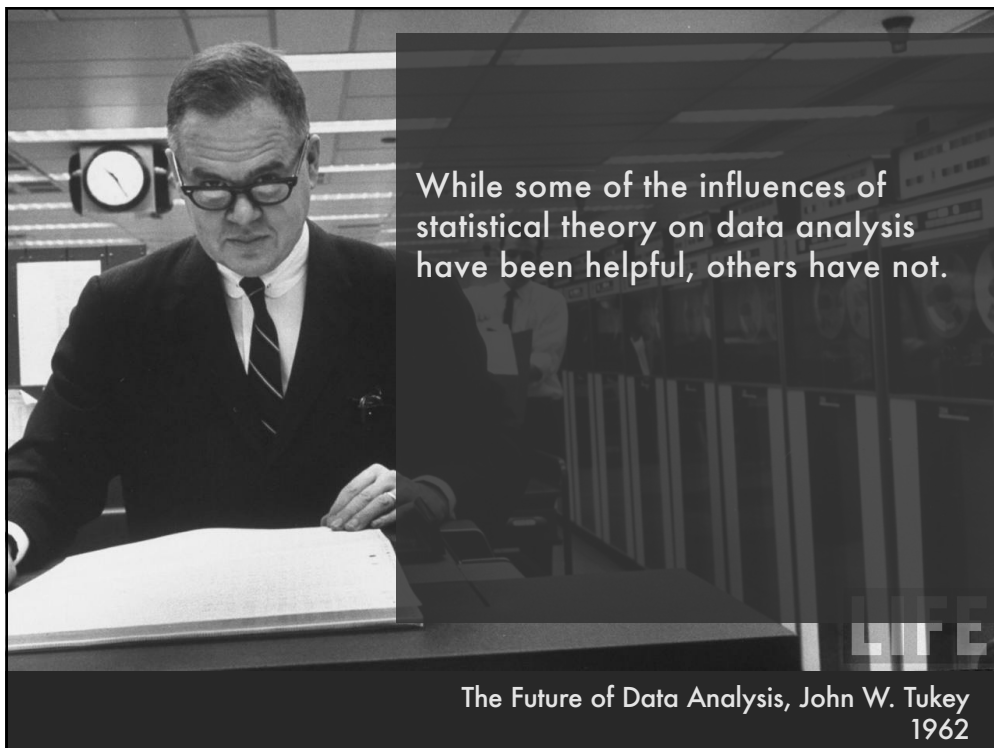
6



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.

The Future of Data Analysis, John W. Tukey
1962


7



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

The Future of Data Analysis, John W. Tukey
1962


8



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.

The Future of Data Analysis, John W. Tukey
1962

9



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**.

The Future of Data Analysis, John W. Tukey
1962

10

Topics

Data Diagnostics
Effectiveness of antibiotics
Intro to Tableau

13

Data Wrangling

14

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	987.7	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

15

DataWrangler

Transform Script Import Export

Split data repeatedly on newline into rows

Split split repeatedly on ','

Promote row 0 to header

Delete empty rows

Text Columns Rows Table Clear

Extract from Year after 'in'

Extract from Year after 'in'

Cut from Year after 'in'

Cut from Year after 'in'

Split Year after 'in'

Split Year after 'in'

Year	extract	Property_crime_rate
0	Reported crime in Alabama	
1	2004	4029.3
2	2005	3900
3	2006	3937
4	2007	3974.9
5	2008	4081.9
6	Reported crime in Alaska	
7	2004	3370.9
8	2005	3615
9	2006	3582
10	2007	3373.9
11	2008	2928.3
12	Reported crime in Arizona	
13	2004	5073.3
14	2005	4827
15	2006	4741.6
16	2007	4502.6
17	2008	4087.3
18	Reported crime in Arkansas	
19	2004	4033.1
20	2005	4068
21	2006	4021.6
22	2007	3945.5
23	2008	3843.7
24	Reported crime in California	
25	2004	3423.9
26	2005	3321
27	2006	3175.2
28	2007	3032.6
29	2008	2940.3
30	Reported crime in Colorado	

16

Data “Wrangling”

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

Some approaches:

Writing custom scripts

Manual manipulation in spreadsheets

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

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

17

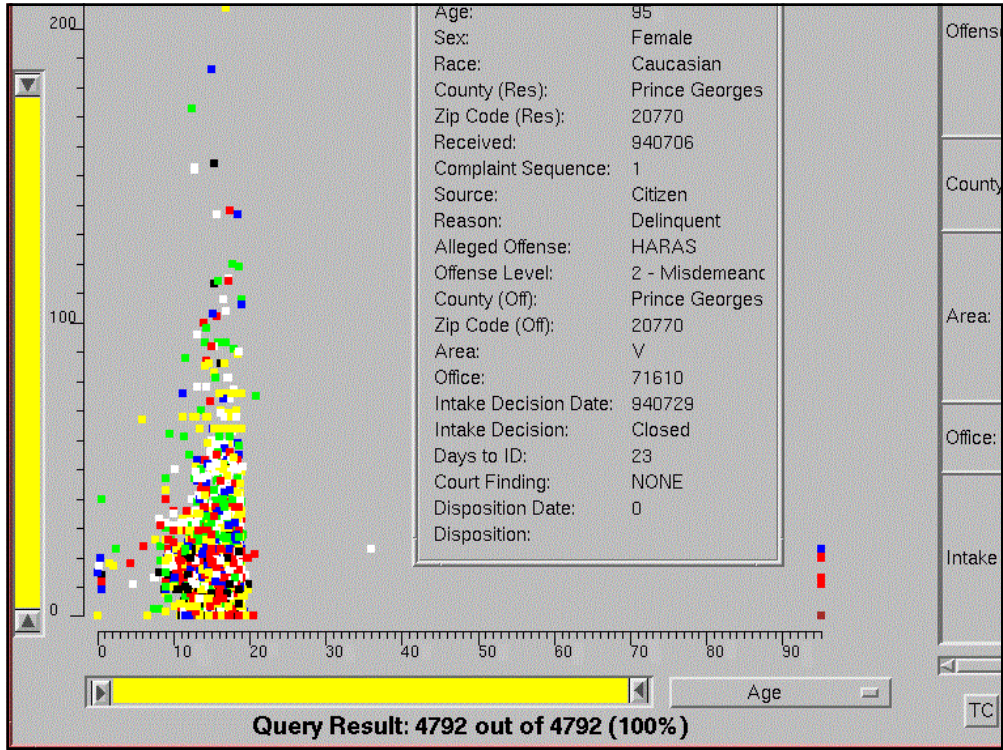
How to gauge the quality of a visualization?

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

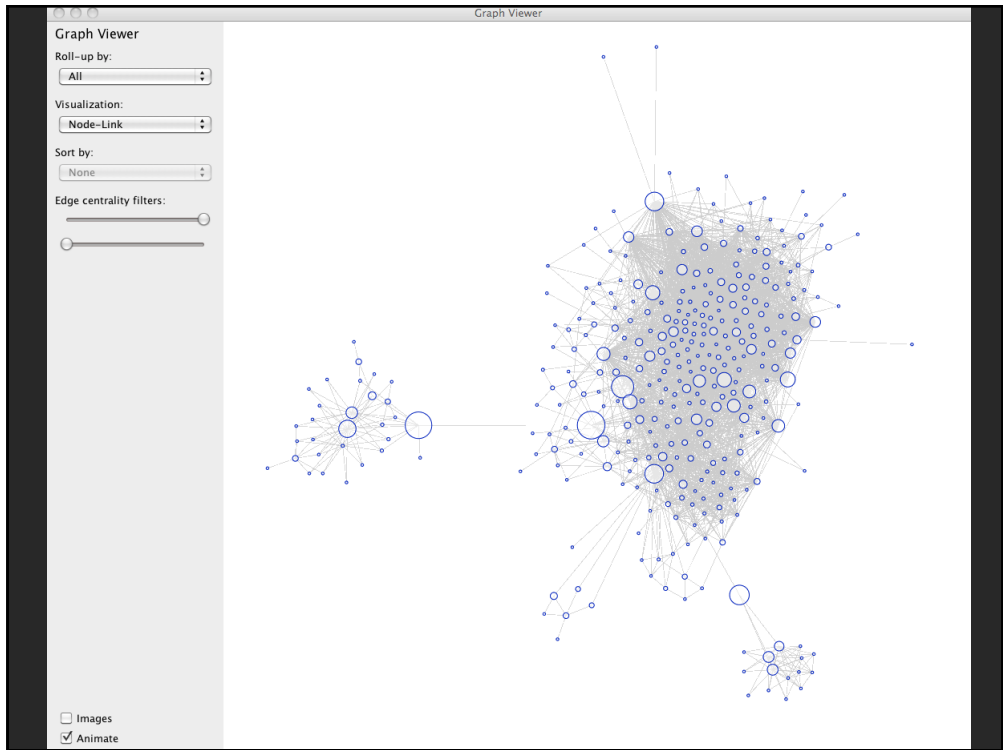
...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

- Martin Wattenberg

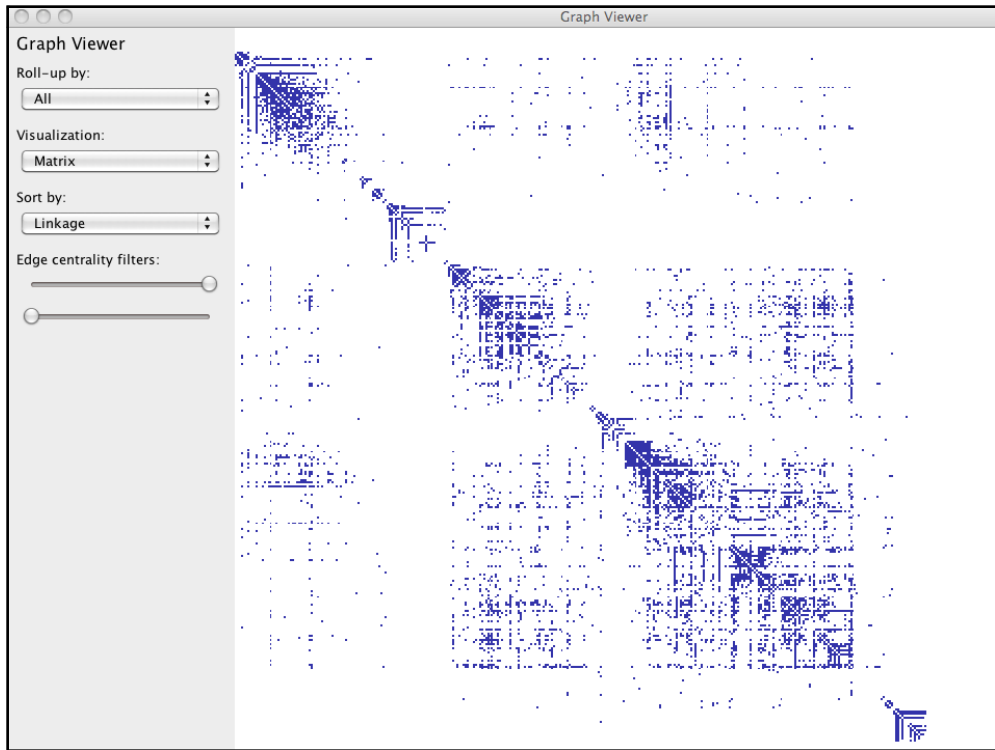
18



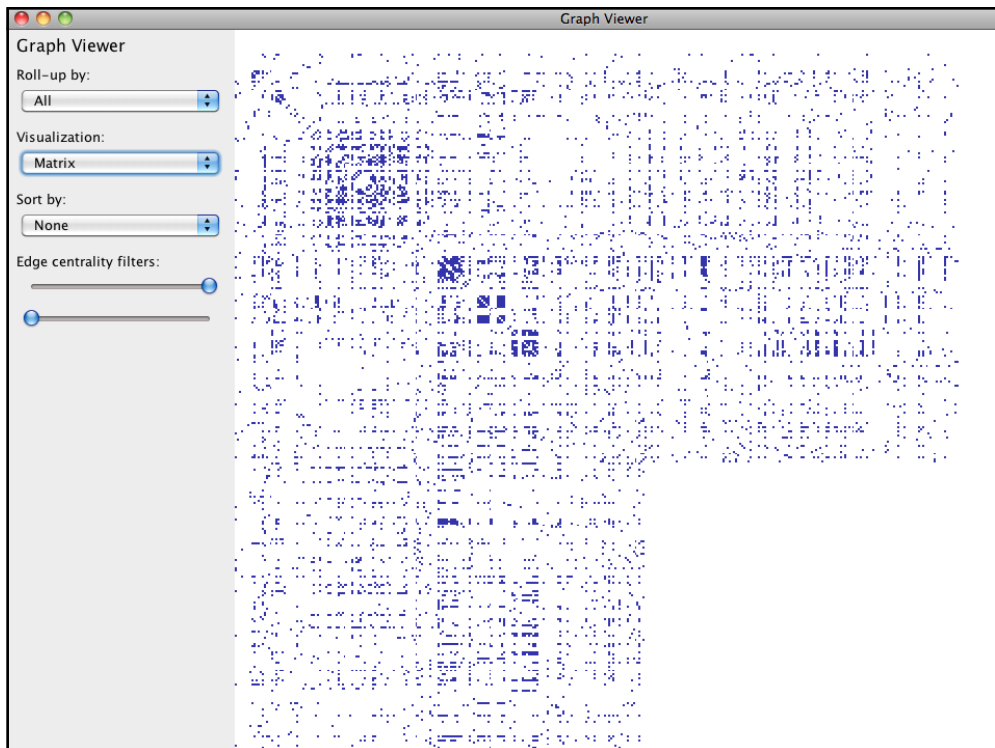
19



21



22



23

Visualize Friends by School?

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

24

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!**

25

Analysis Example: Effectiveness of Antibiotics

35

Antibiotic Effectiveness: The Data

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

Collected prior to 1951

36

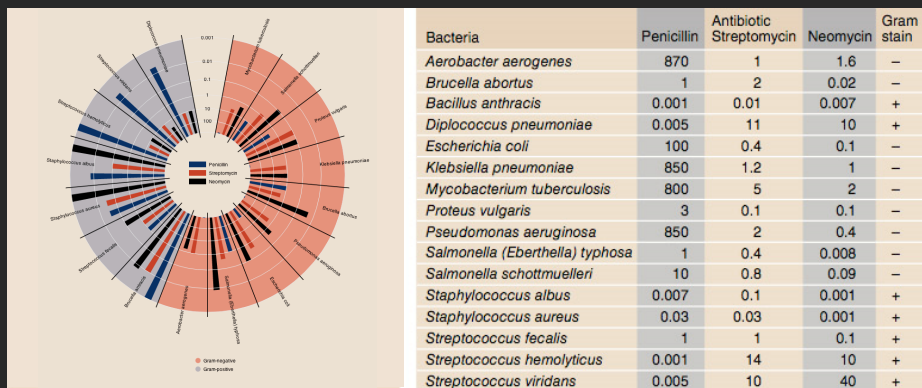
What questions might we ask?

Table 1: Burtin's data.

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

37

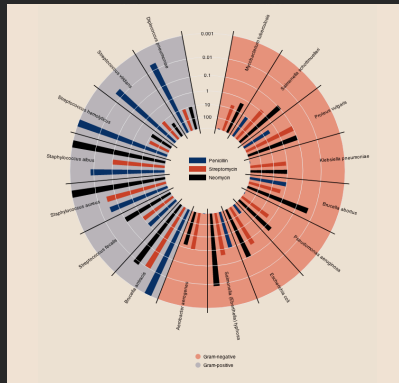
Will Burtin, 1951



How do the drugs compare?

38

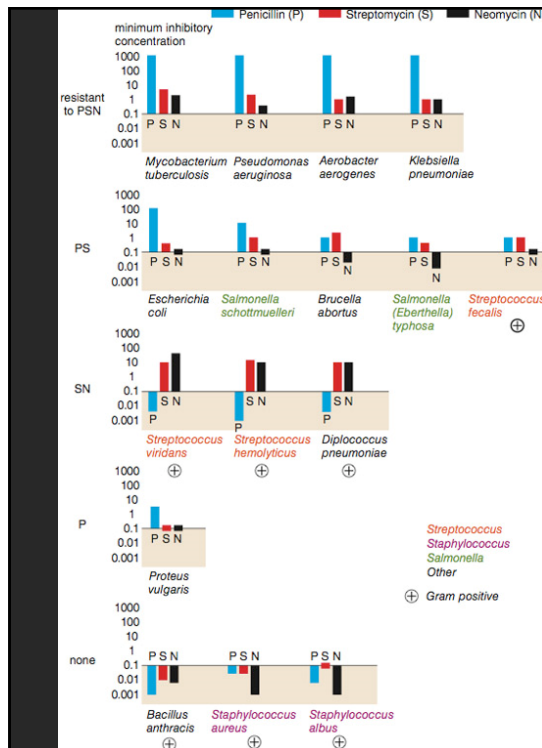
Will Burtin, 1951



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

Radius: $1/\log(\text{MIC})$
 Bar Color: Antibiotic
 Background Color: Gram Staining

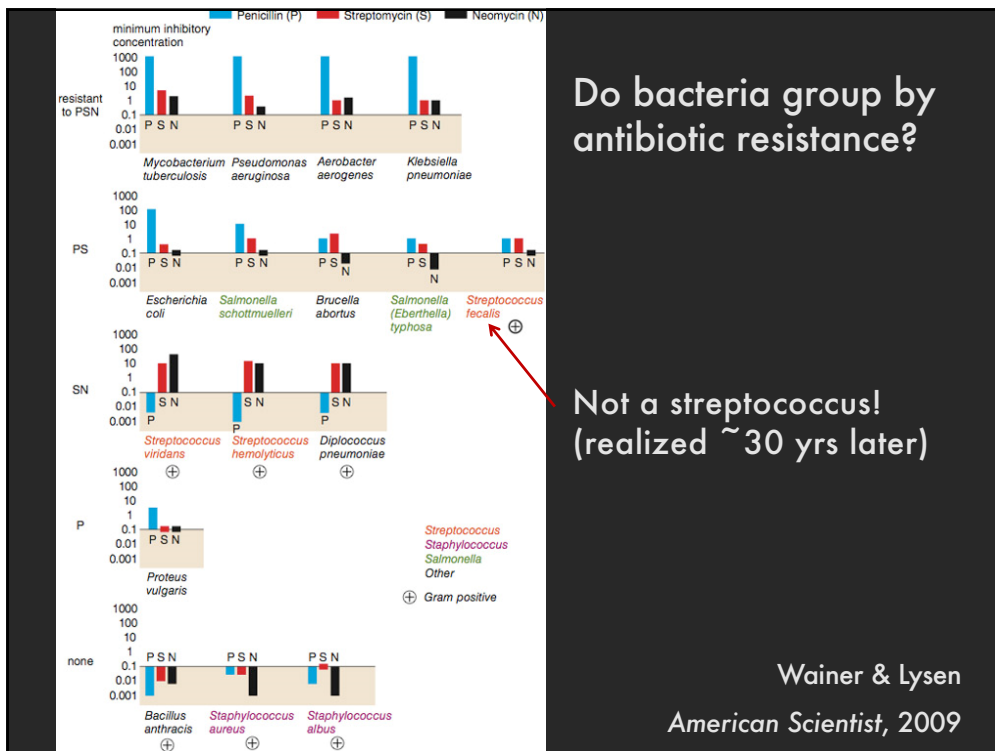
39



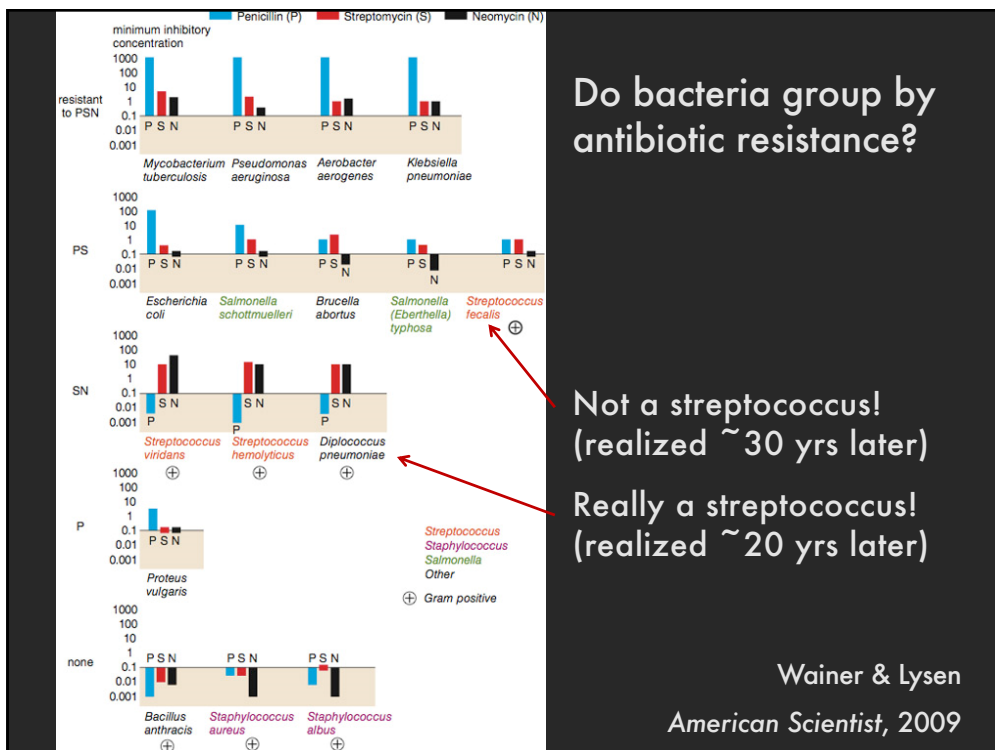
Do bacteria group by antibiotic resistance?

Wainer & Lysen
 American Scientist, 2009

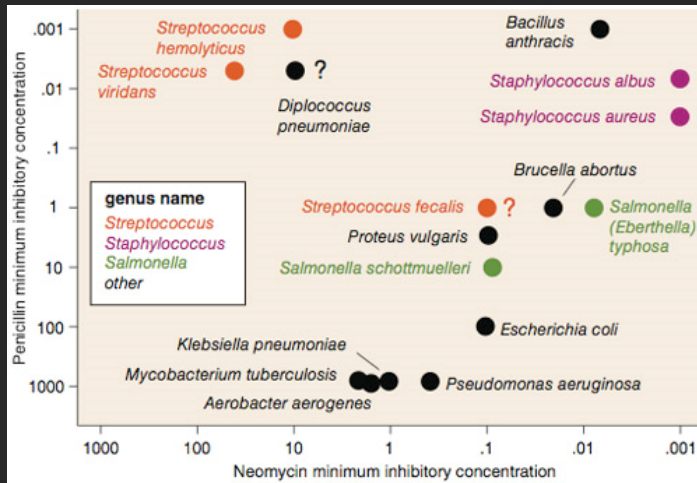
40



41



42



**How do the bacteria group w.r.t. resistance?
Do different drugs correlate?**

Wainer & Lysen
American Scientist, 2009

43

Lessons

Exploratory Process

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

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

“Show data variation, not design variation”

-Tufte

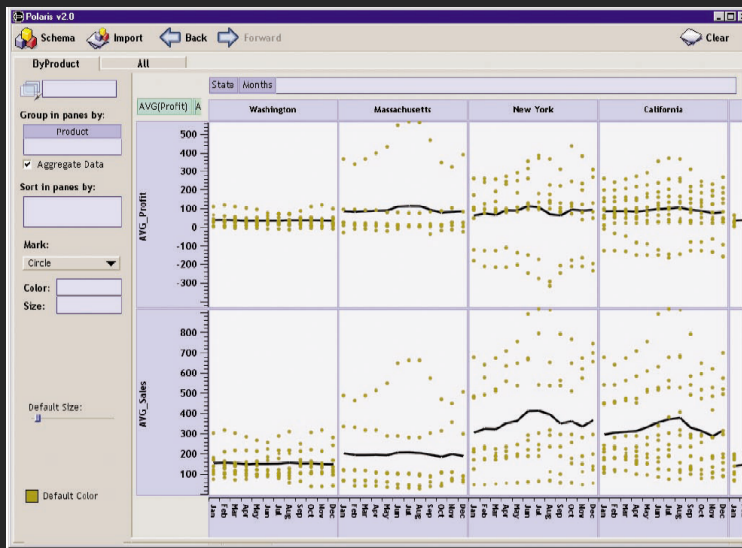
44

Tableau / Polaris

79

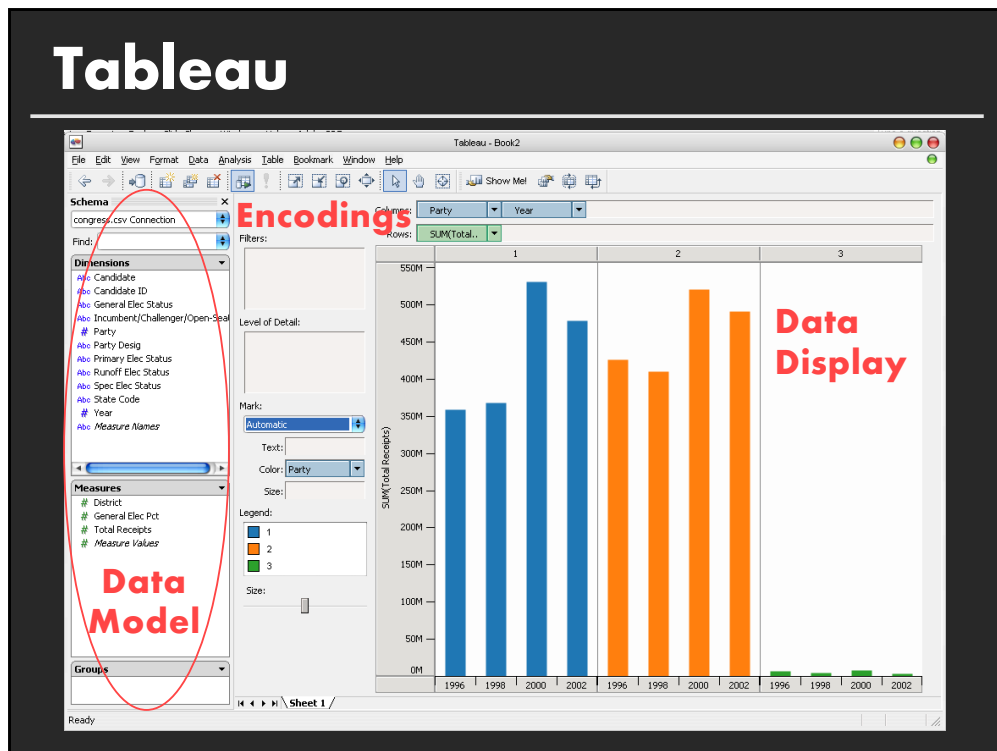
Tableau

Research at Stanford: "Polaris" by Stolte, Tang & Hanrahan.



80

Tableau



81

Polaris/Tableau Approach

Insight: simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest more encodings upon request
(ShowMe - Like APT)

82

Dataset

- **Federal Elections Commission Receipts**
- **Every Congressional Candidate from 1996 to 2002**
- **4 Election Cycles**
- **9216 Candidacies**

83

Data Set 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, but should be sufficient for the demo**

84

Hypotheses?

What might we learn from this data?

85

Hypotheses?

What might we learn from this data?

- Have receipts increased over time?
- Do democrats or republicans spend more?
- Candidates from which state spend the most money?

Tableau Demo

86

Specifying Table Configurations

Operands are names of database fields

Each operand interpreted as a set {...}

Data is either O or Q and treated differently

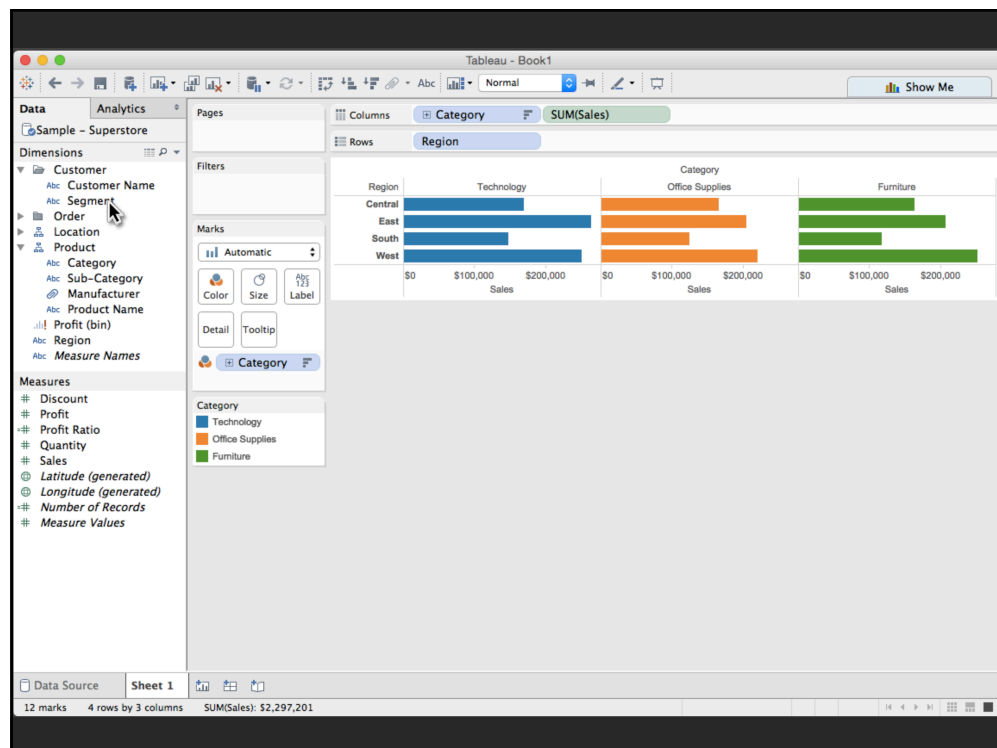
Three operators:

concatenation (+)

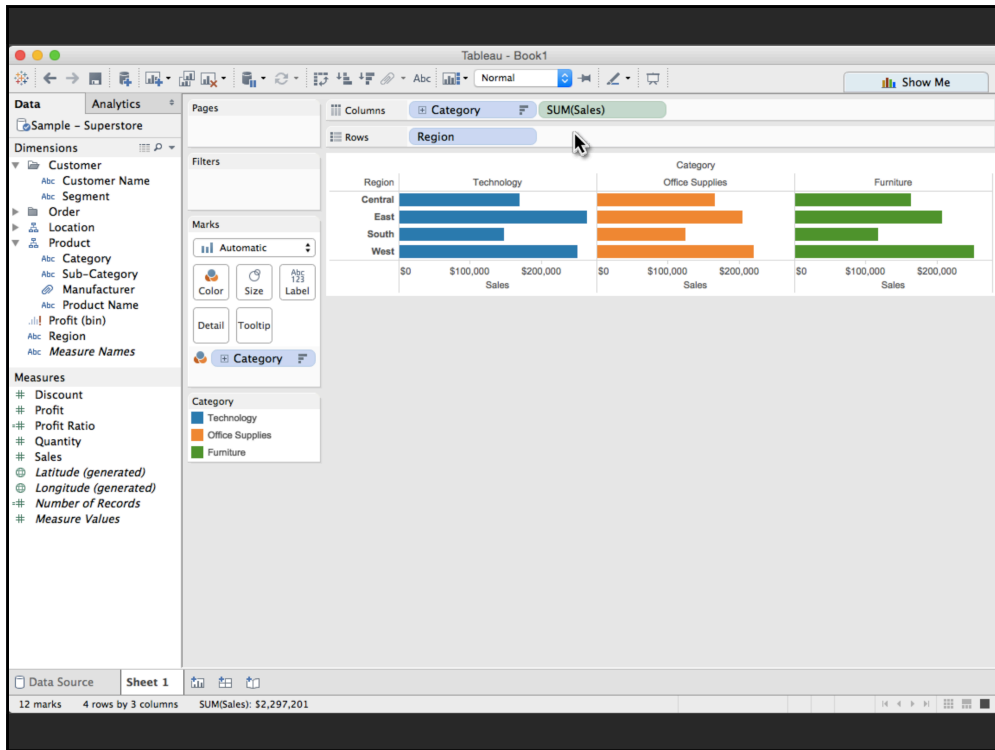
cross product (x)

nest (/)

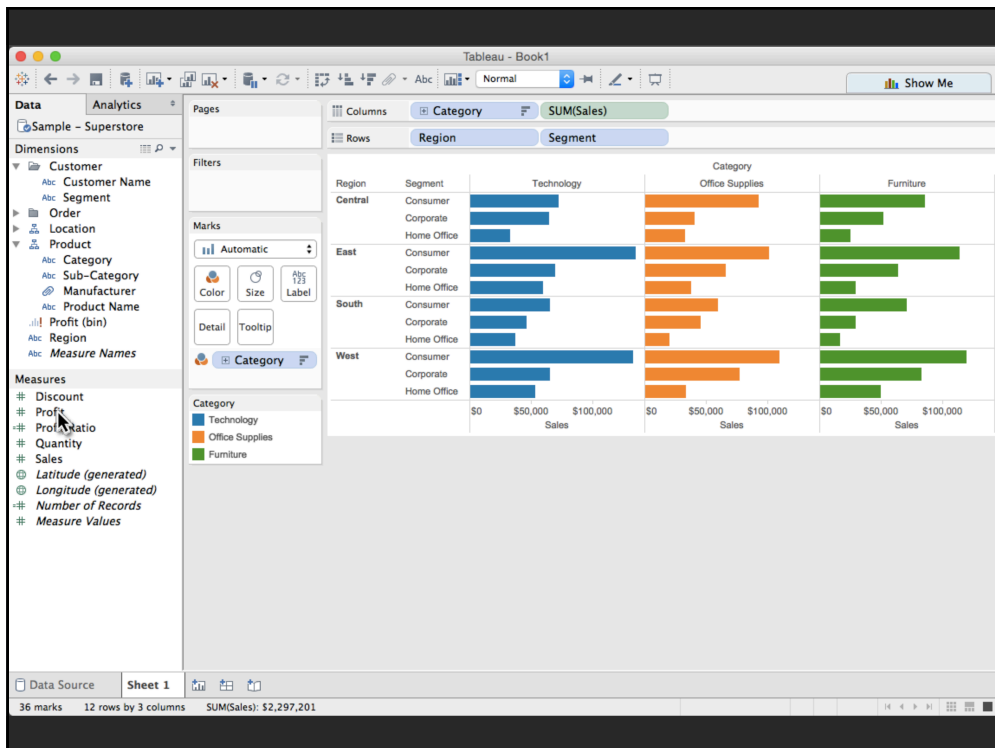
87



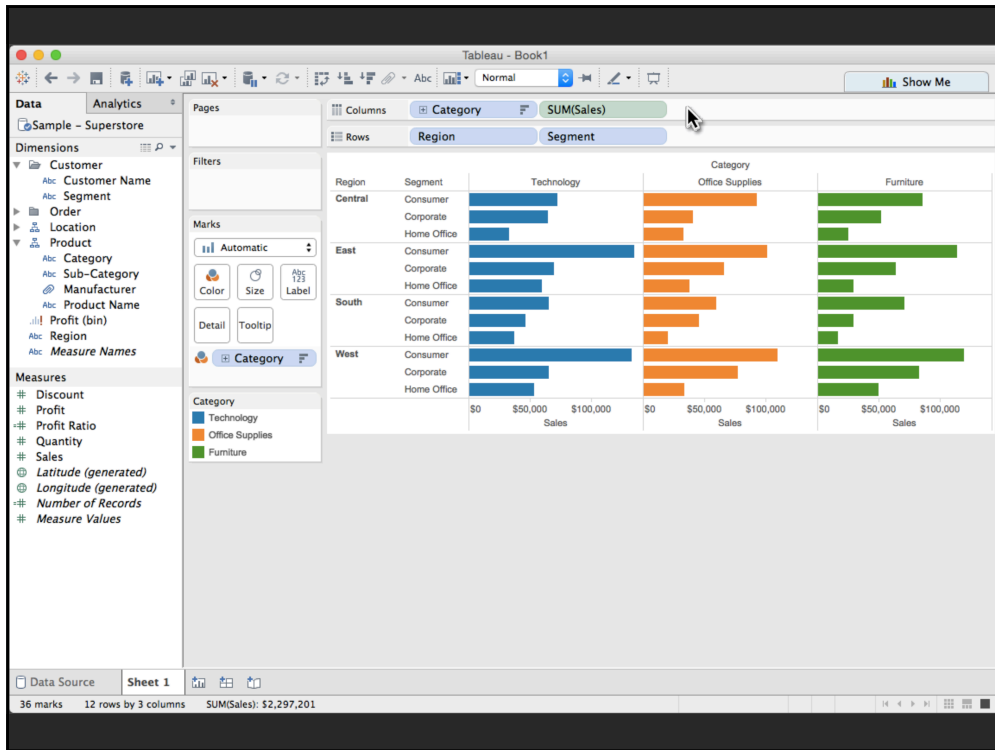
88



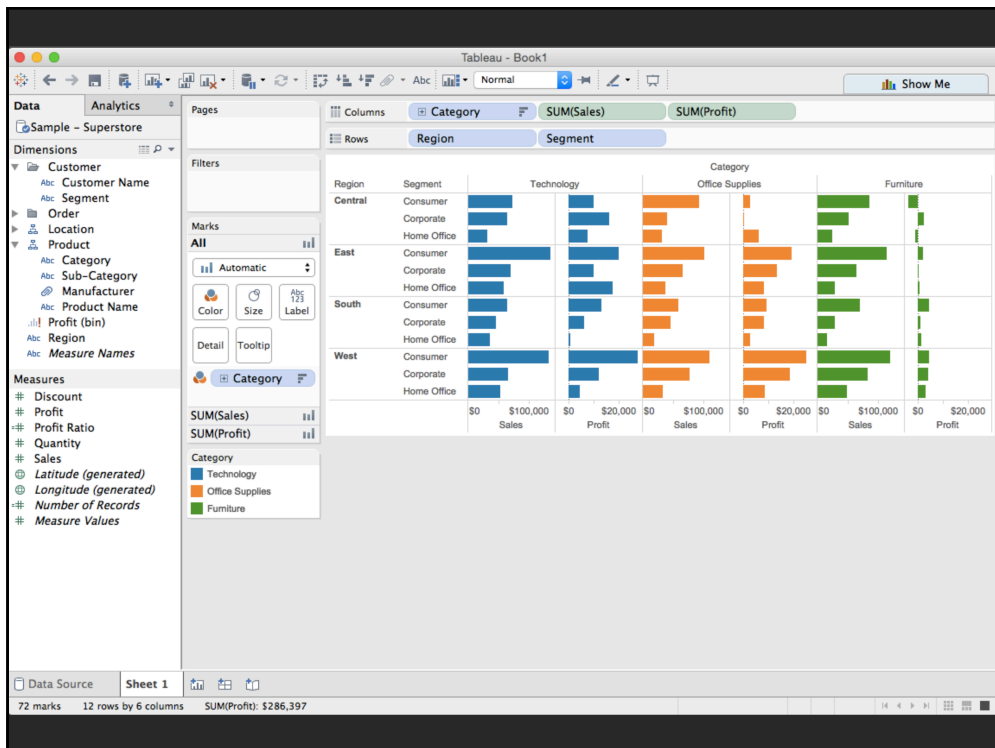
89



90



91



92



93

Table Algebra

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

Algebraic statements are mapped to

Visualizations – trellis partitions, visual encodings

Queries – selection, projection, group-by

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

Users specify operands NOT operators!

Operators are inferred by data type (O, Q)

94