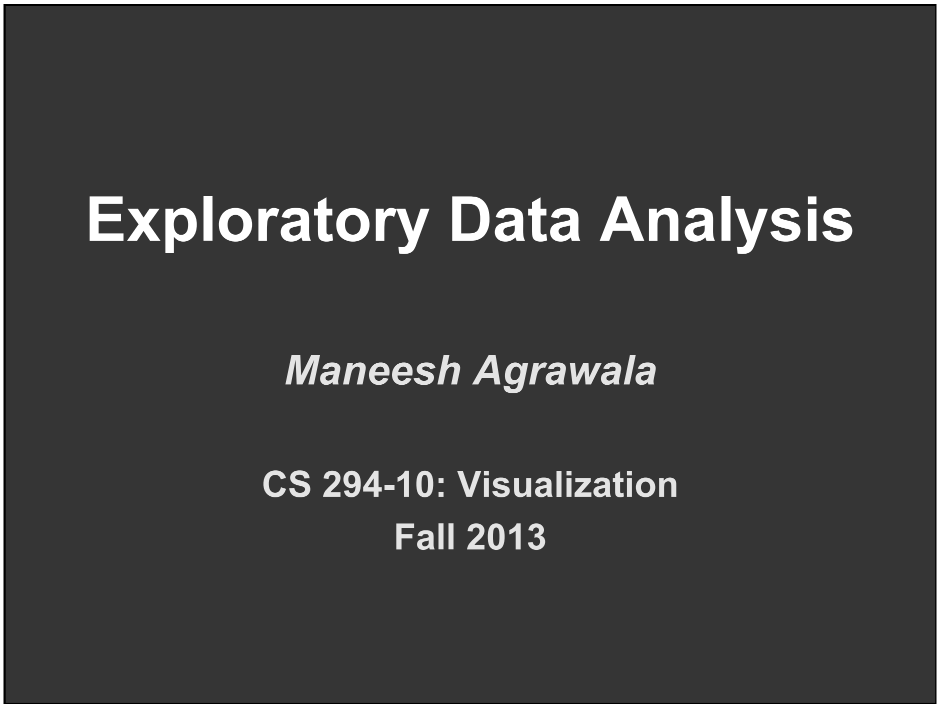
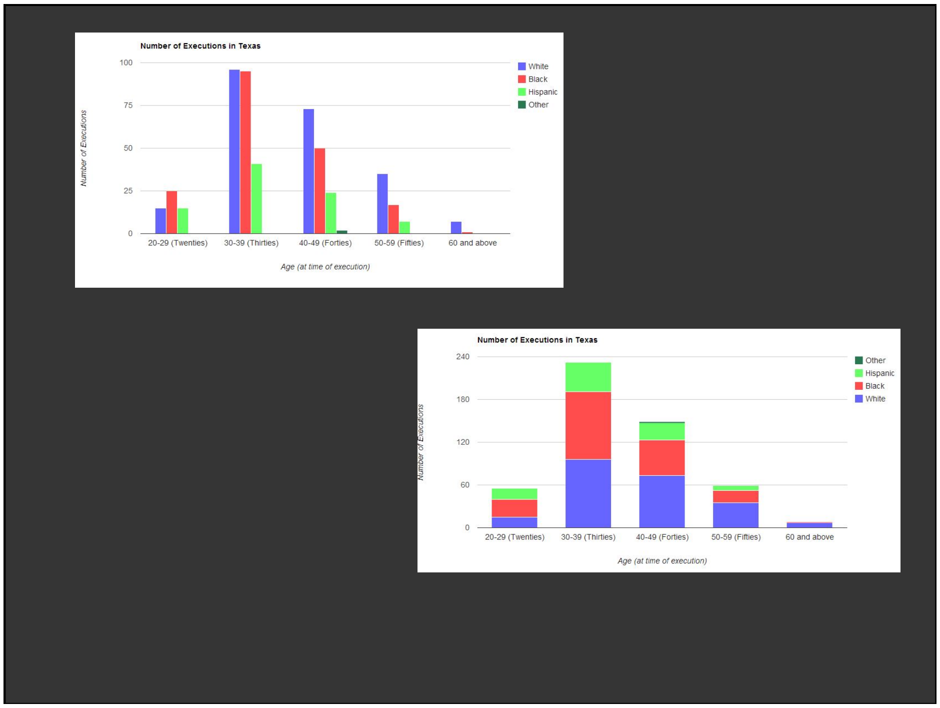


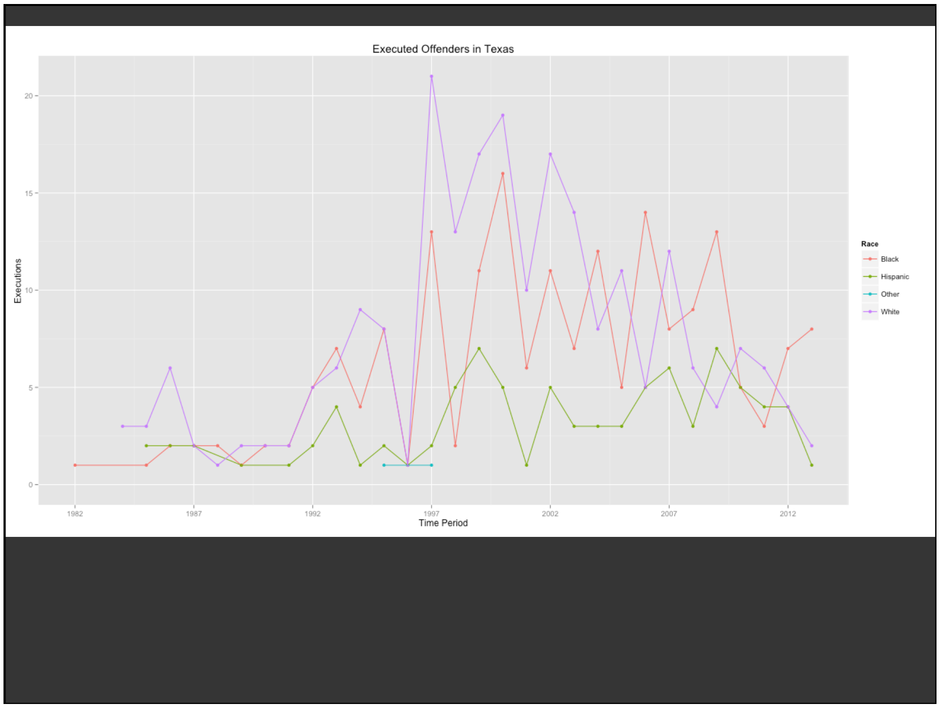
Exploratory Data Analysis

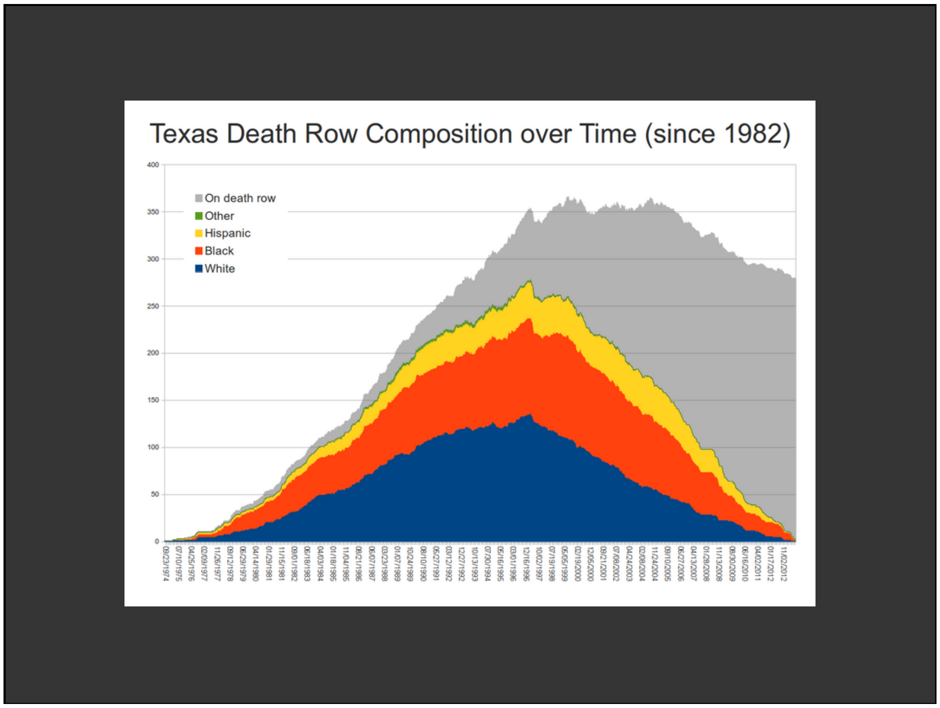
Maneesh Agrawala

CS 294-10: Visualization
Fall 2013

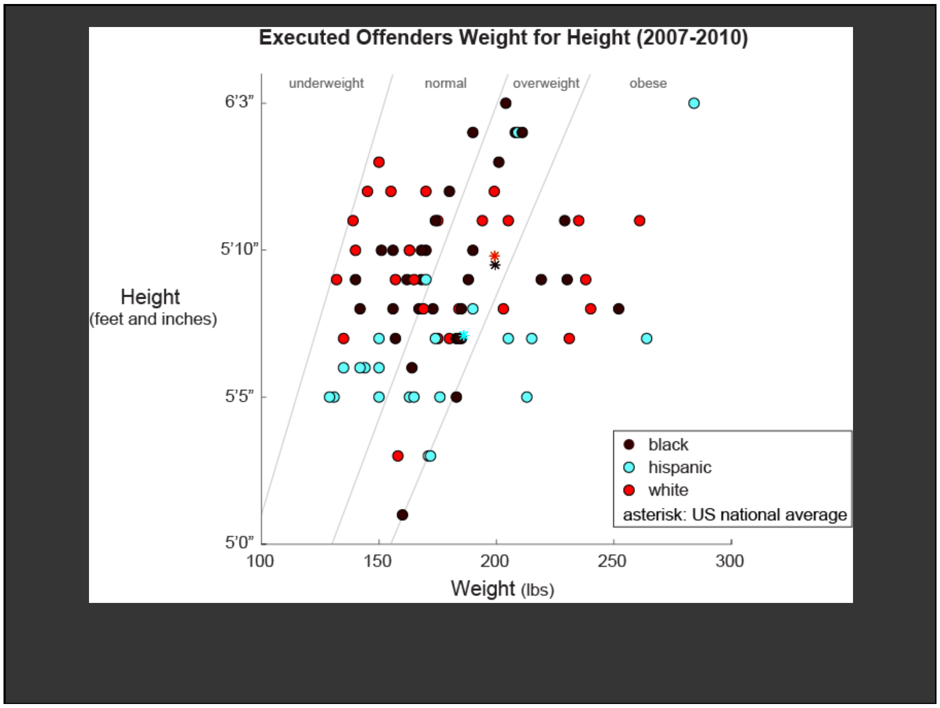
Last Time: Visualization Designs


Number of Executions in Texas
100
75
50
25
0
Number of Executions
20-29 (Twenties)
30-39 (Thirties)
40-49 (Forties)
50-59 (Fifties)
60 and above
Age (at time of execution)
White
Black
Hispanic
Other
Number of Executions in Texas
240
180
120
60
0
Number of Executions
20-29 (Twenties)
30-39 (Thirties)
40-49 (Forties)
50-59 (Fifties)
60 and above
Age (at time of execution)
Other
Hispanic
Black
White

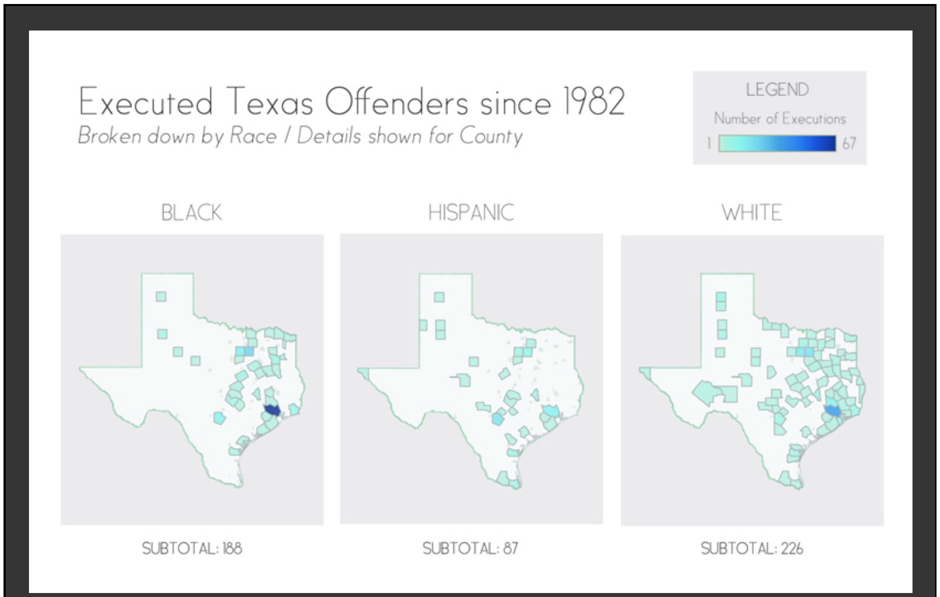

Executed Offenders in Texas
20
15
10
5
0
Executions
1982
1987
1992
1997
2002
2007
2012
Time Period
Race
Black
Hispanic
Other
White



Texas Death Row Composition over Time (since 1982)

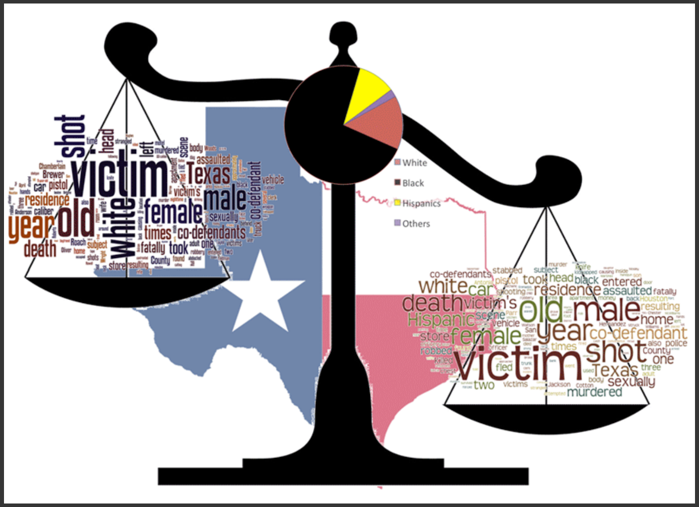


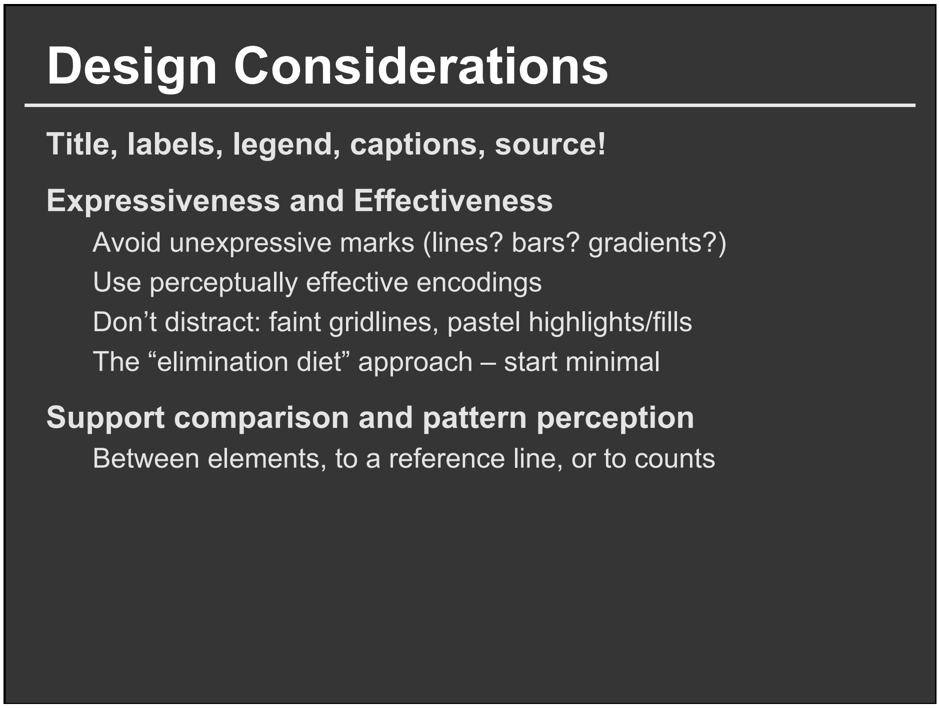
Executed Offenders Weight for Height (2007-2010)



Executed Texas Offenders since 1982

Broken down by Race / Details shown for County





Design Considerations

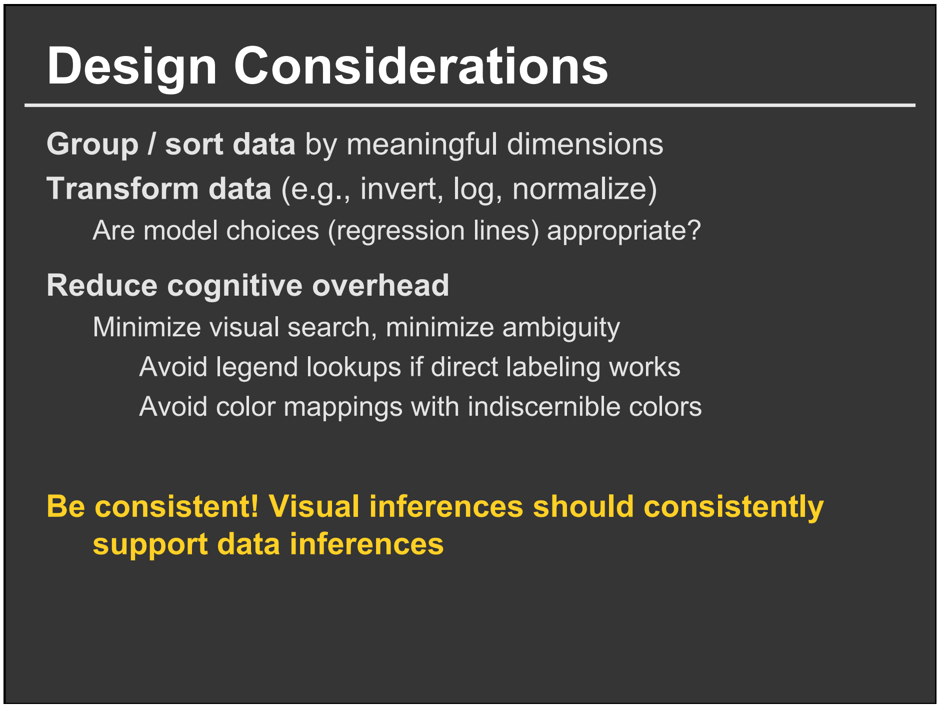
Title, labels, legend, captions, source!

Expressiveness and Effectiveness

- Avoid unexpressive marks (lines? bars? gradients?)
- Use perceptually effective encodings
- Don’t distract: faint gridlines, pastel highlights/fills
- The “elimination diet” approach – start minimal

Support comparison and pattern perception

- Between elements, to a reference line, or to counts



Design Considerations

Group / sort data by meaningful dimensions

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

- Are model choices (regression lines) appropriate?

Reduce cognitive overhead

- Minimize visual search, minimize ambiguity
  - Avoid legend lookups if direct labeling works
  - Avoid color mappings with indiscernible colors

Be consistent! Visual inferences should consistently support data inferences

In-Class Review

Procedure

Break into groups of 4 (assigned by me)

Appoint a time keeper

Take turns showing your visualization – present findings (~3 min each)

Then critique – rubric on next slide (~5 min each)

- Get feedback from everyone in group
- Author must take notes

Post writeup to assignment 1 page after class

- Include feedback
- Briefly describe how you would re-design the visualization

Write-up of critique will be used in grading

In-Class Review Rubric

Expressiveness

- Prioritizes important information / Avoids false inferences
- Consistent visual mappings (e.g., respect color mappings)
- Make encodings *meaningful* rather than arbitrary

Effectiveness

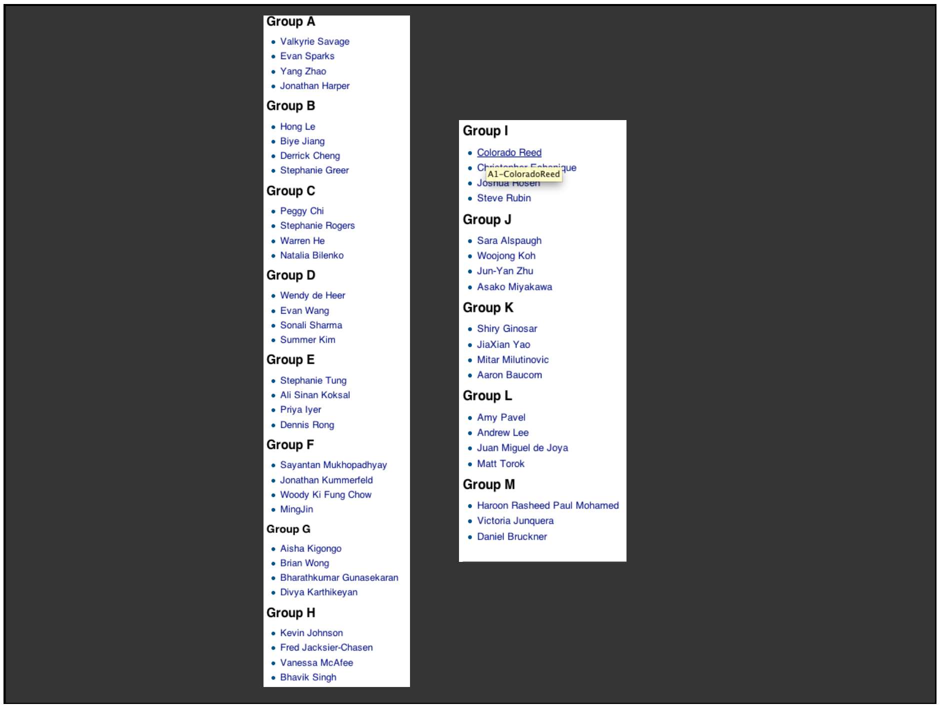
- Facilitates accurate decoding / Minimizes cognitive overhead
- Highlight elements of primary interest

Grouping / Sorting

Data Transformation

Non-Data Elements

- Descriptive: Title, Label, Caption, Data Source, Annotations
- Reference: Gridlines, Legend



Group A

- Valkyrie Savage
- Evan Sparks
- Yang Zhao
- Jonathan Harper

Group B

- Hong Le
- Biye Jiang
- Derrick Cheng
- Stephanie Greer

Group C

- Peggy Chi
- Stephanie Rogers
- Warren He
- Natalia Bilenko

Group D

- Wendy de Heer
- Evan Wang
- Sonali Sharma
- Summer Kim

Group E

- Stephanie Tung
- Ali Sinan Koksal
- Priya Iyer
- Dennis Rong

Group F

- Sayantan Mukhopadhyay
- Jonathan Kummerfeld
- Woody Ki Fung Chow
- MingJin

Group G

- Aisha Kigongo
- Brian Wong
- Bharathkumar Gunasekaran
- Divya Karthikeyan

Group H

- Kevin Johnson
- Fred Jacksier-Chasen
- Vanessa McAfee
- Bhavik Singh

Group I

- Colorado Reed
- Christopher Echenique
- Joshua Rosen
- Steve Rubin

Group J

- Sara Alspaugh
- Woojong Koh
- Jun-Yan Zhu
- Asako Miyakawa

Group K

- Shiry Ginosar
- JiaXian Yao
- Mitar Milutinovic
- Aaron Baucom

Group L

- Amy Pavel
- Andrew Lee
- Juan Miguel de Joya
- Matt Torok

Group M

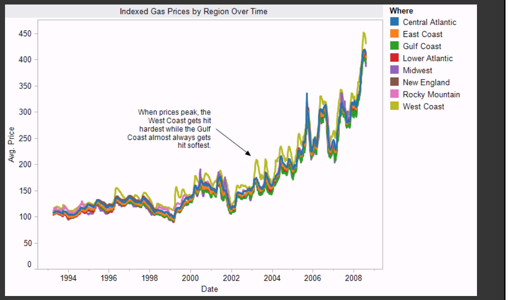
- Haroon Rasheed Paul Mohamed
- Victoria Junquera
- Daniel Bruckner

Assignment 2: Exploratory Data Analysis

Use existing software to formulate & answer questions

First steps

- Step 1: Pick a domain
- Step 2: Pose questions
- Step 3: Find data
- Iterate



Indexed Gas Prices by Region Over Time

When prices peak, the West Coast gets hit hardest while the Gulf Coast almost always gets hit softest.

Create visualizations

- Interact with data
- Question will evolve
- Tableau

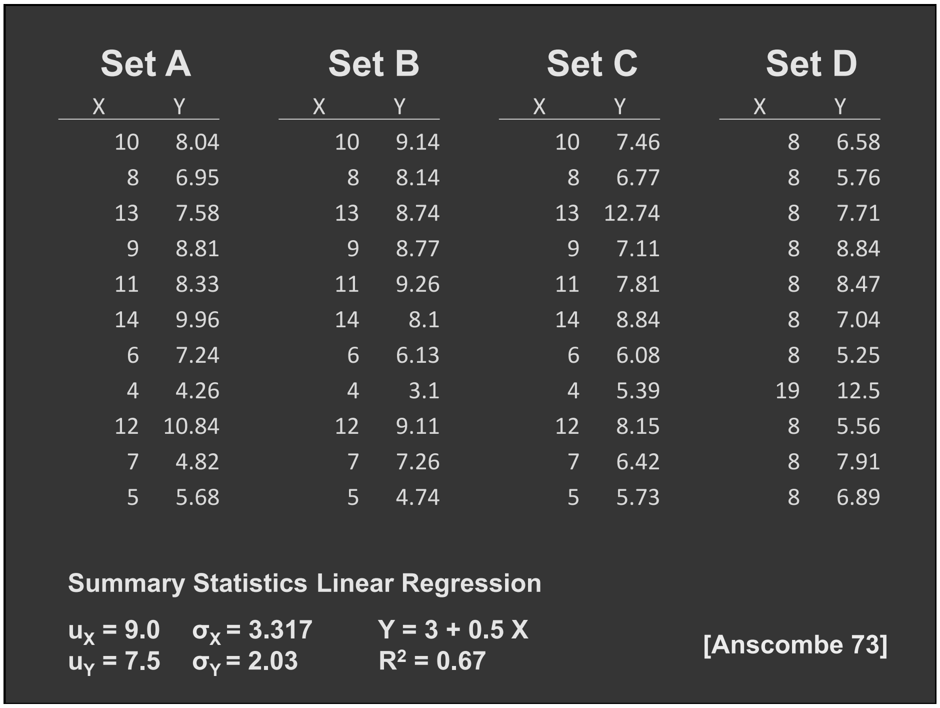
Make wiki notebook

- Keep record of all steps you took to answer the questions

Due before class on Sep 30, 2013

Exploratory Data Analysis

The Future of Data Analysis, John W. Tukey 1962

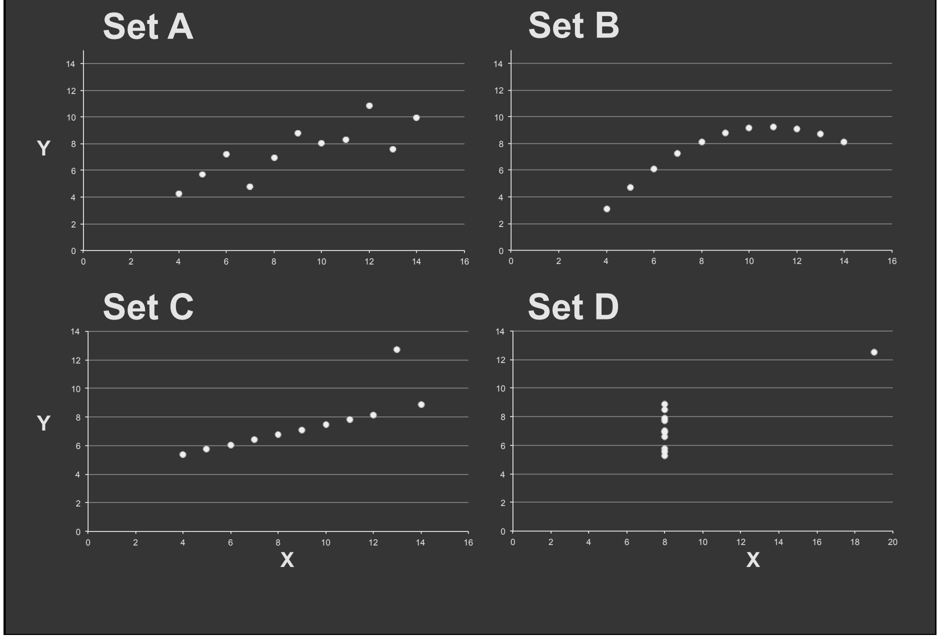


Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	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 Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\,X$
$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]


Set A

Set B

Set C

Set D

Topics

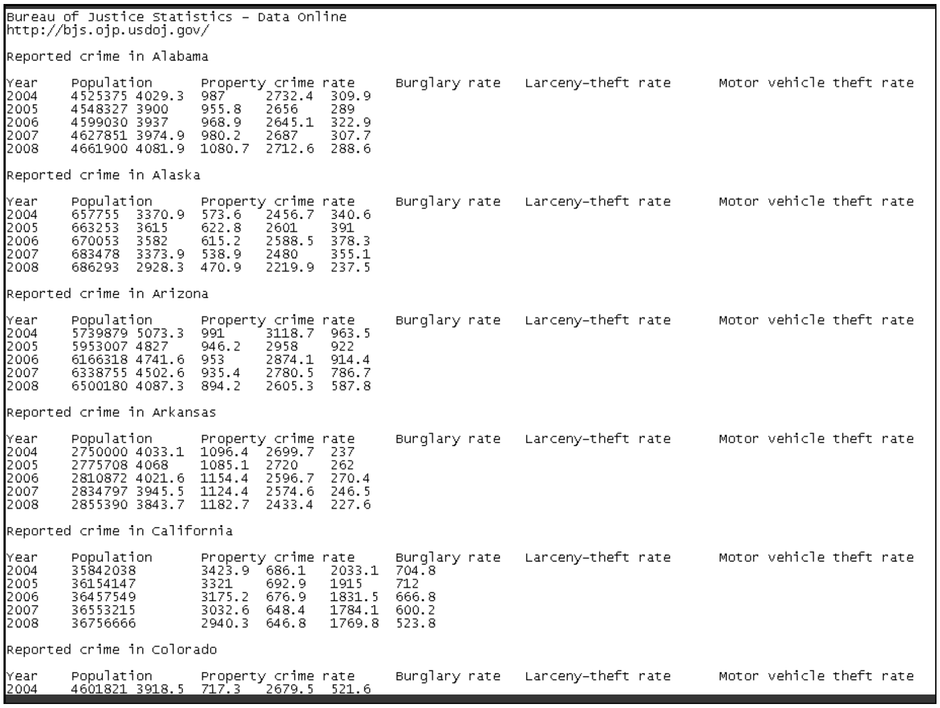
Exploratory Data Analysis

- Data Diagnostics
- Graphical Methods
- Data Transformation

Confirmatory Data Analysis

- Statistical Hypothesis Testing
- Graphical Inference

Data Diagnostics



```
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

Some approaches:

Writing custom scripts

Manual manipulation in spreadsheets

Data Wrangler: http://vis.stanford.edu/wrangler

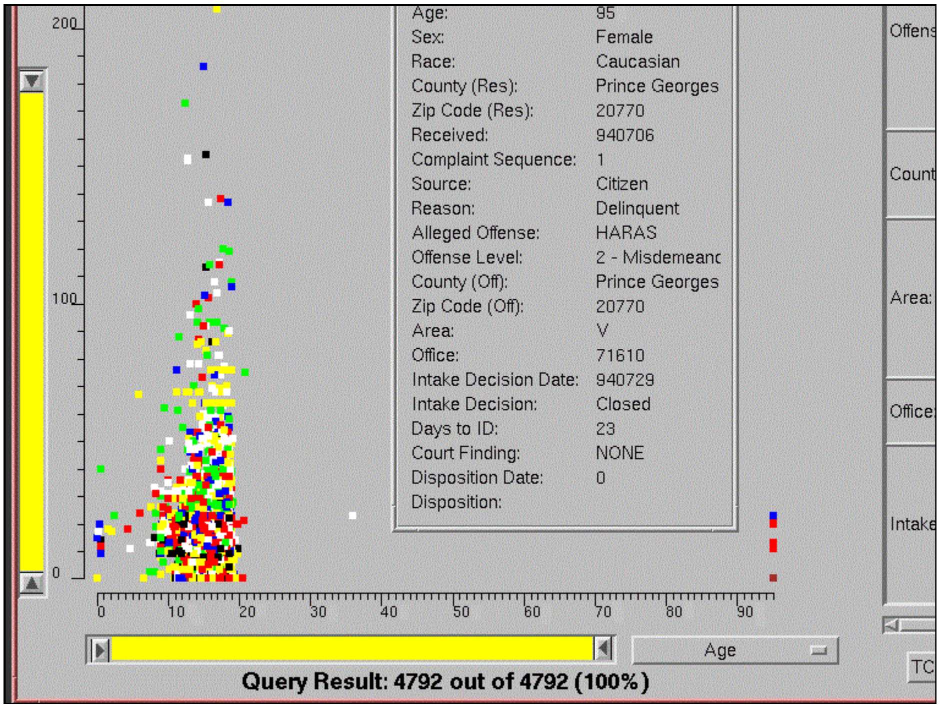
Google Refine: http://code.google.com/p/google-refine

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…

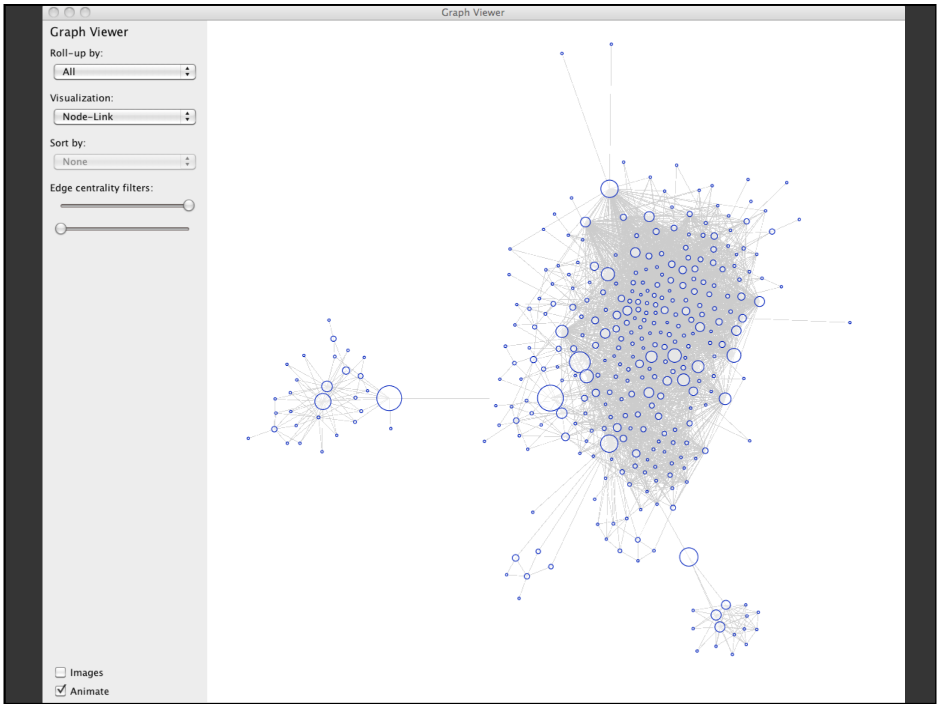
…every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

- Martin Wattenberg



Age:	95
Sex:	Female
Race:	Caucasian
County (Res):	Prince Georges
Zip Code (Res):	20770
Received:	940706
Complaint Sequence:	1
Source:	Citizen
Reason:	Delinquent
Alleged Offense:	HARAS
Offense Level:	2 - Misdemeanc
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%)



Graph Viewer

Roll-up by: All

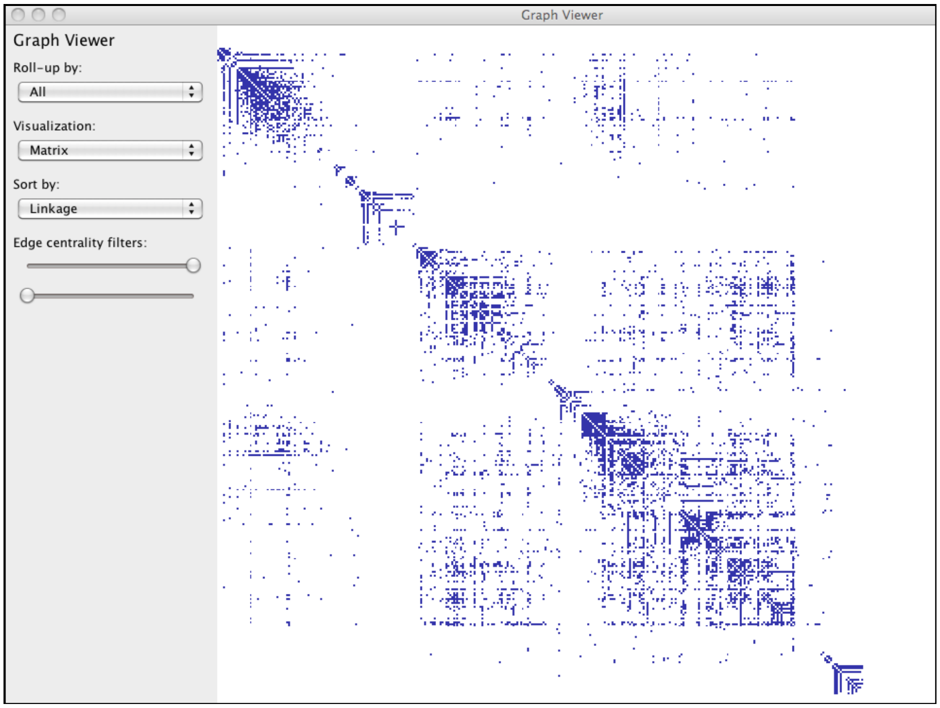
Visualization: Node-Link

Sort by: None

Edge centrality filters:

☐ Images

☑ Animate



Graph Viewer

Graph Viewer

Roll-up by:

All

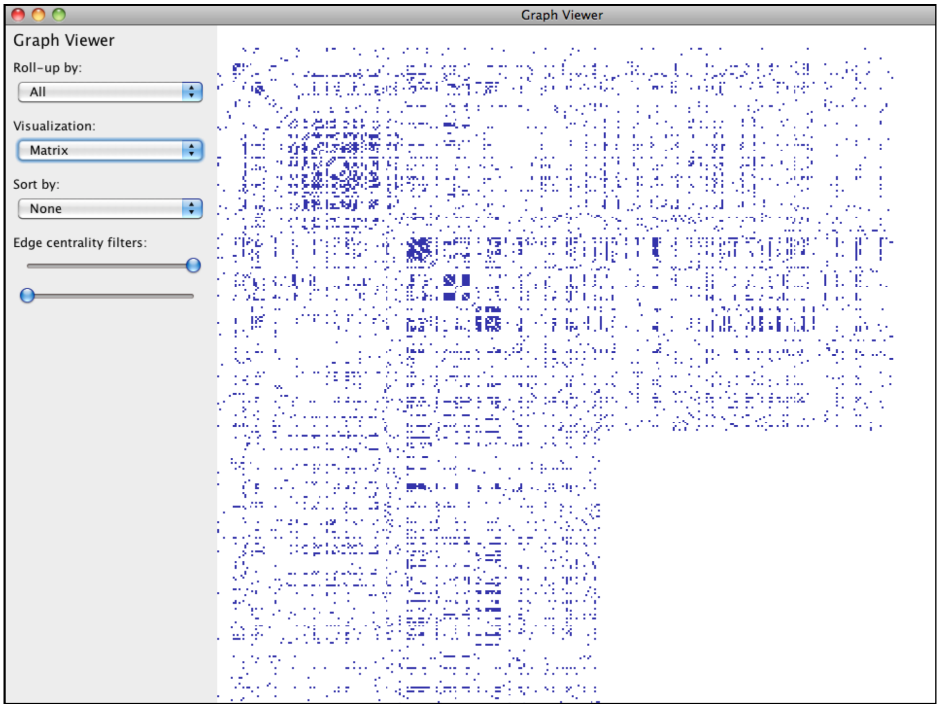
Visualization:

Matrix

Sort by:

Linkage

Edge centrality filters:



Graph Viewer

Graph Viewer

Roll-up by:

All

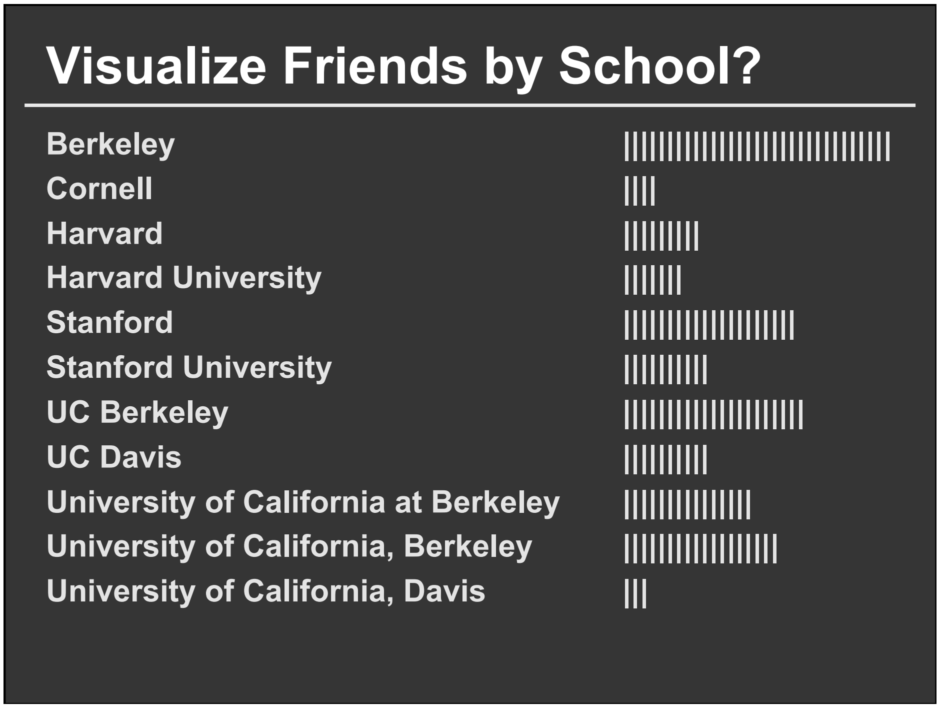
Visualization:

Matrix

Sort by:

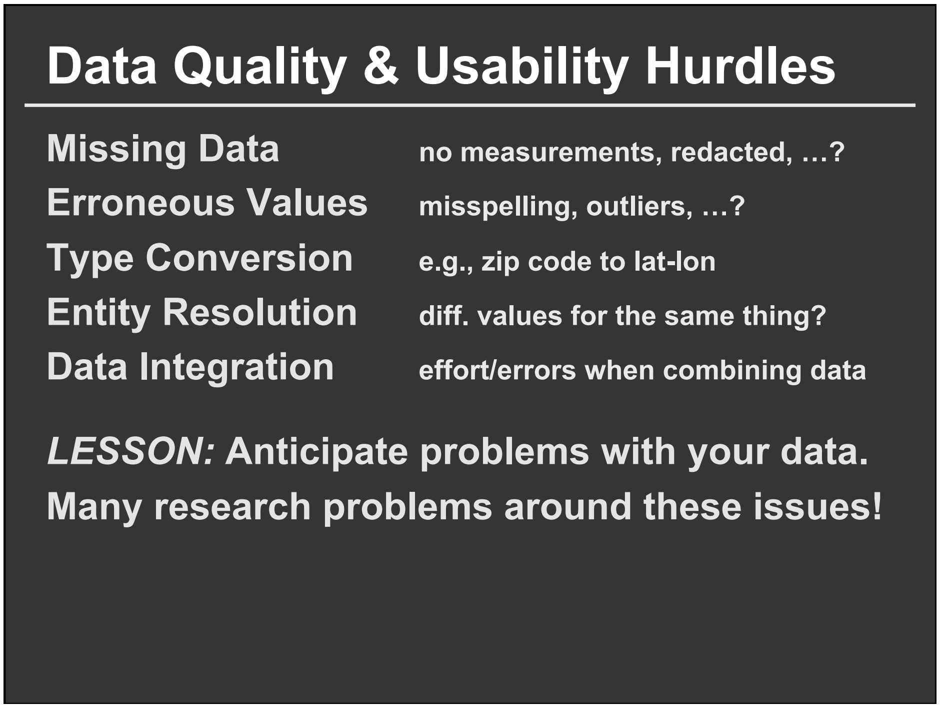
None

Edge centrality filters:



Visualize Friends by School?

<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 & Usability 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!

Exploratory Analysis: Effectiveness of Antibiotics

The Data Set

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

Collected prior to 1951

What questions might we ask?

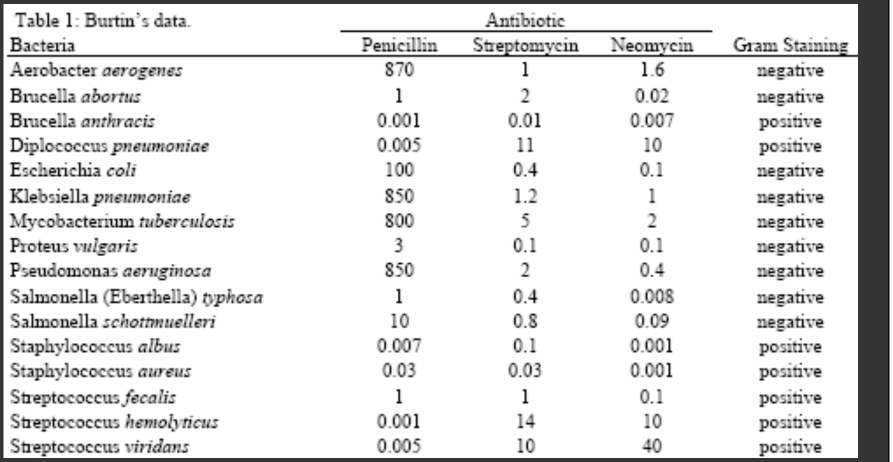
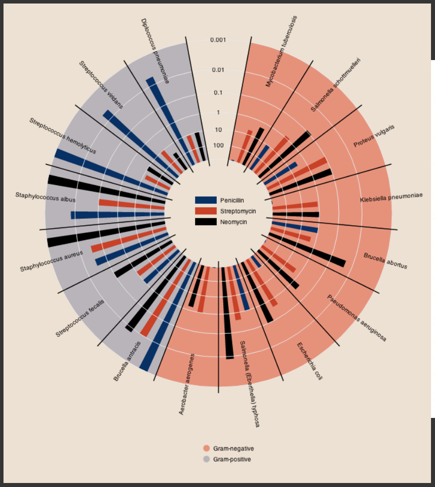
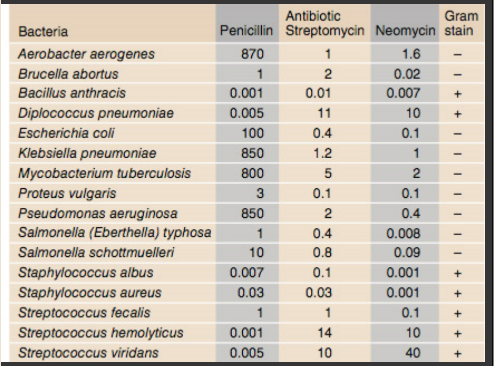


Table 1: Burtin's data.

Bacteria	Antibiotic Penicillin	 Streptomycin	 Neomycin	Gram Staining
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

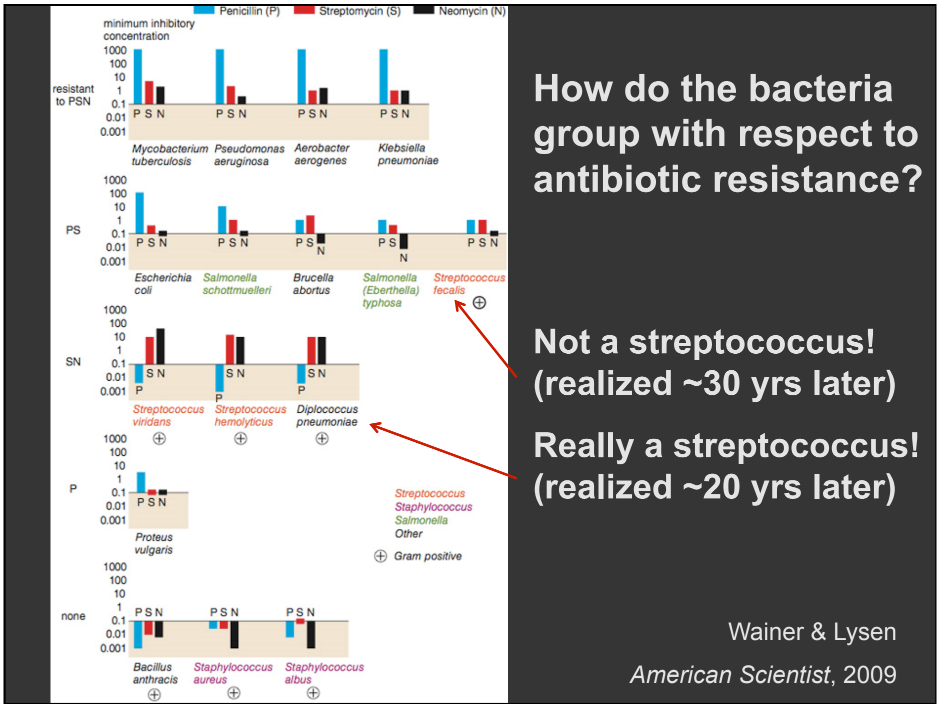
Will Burtin, 1951





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	+

How do the drugs compare?



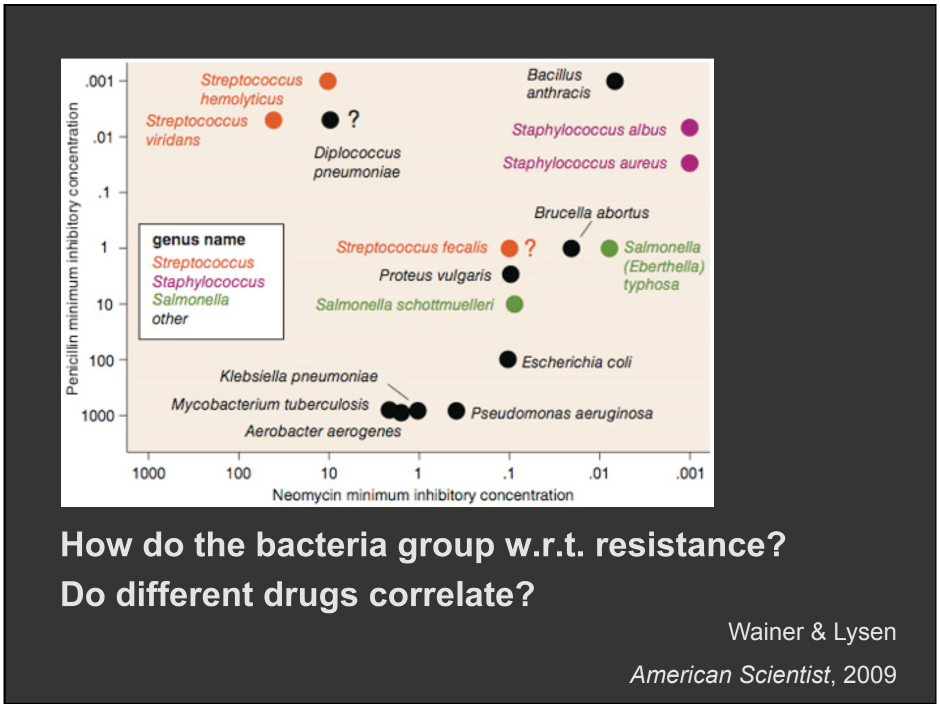
How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen

American Scientist, 2009

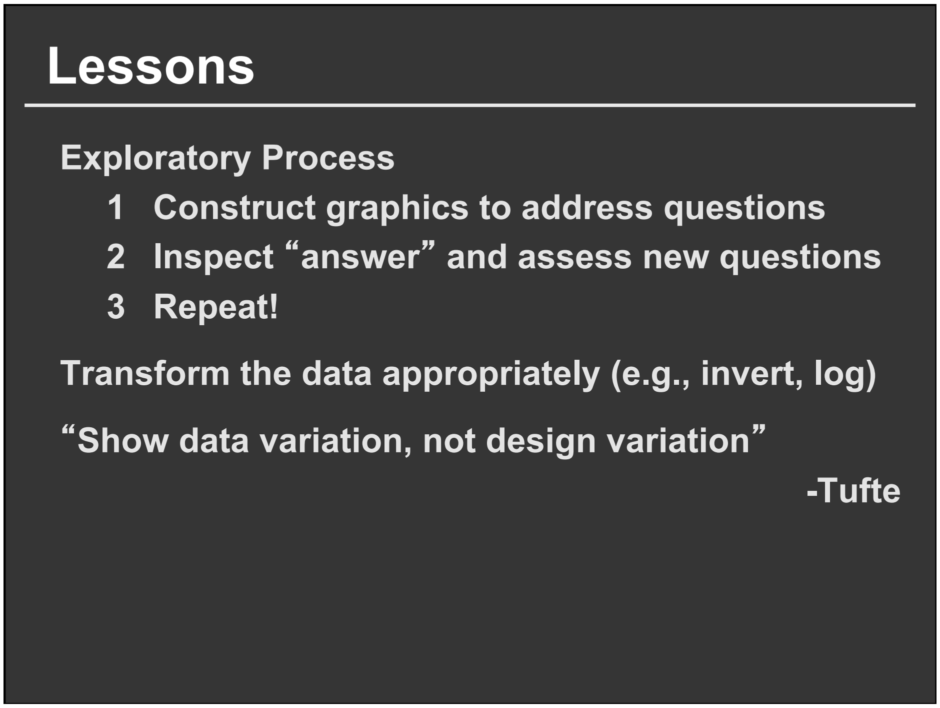


How do the bacteria group w.r.t. resistance?

Do different drugs correlate?

Wainer & Lysen

American Scientist, 2009



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