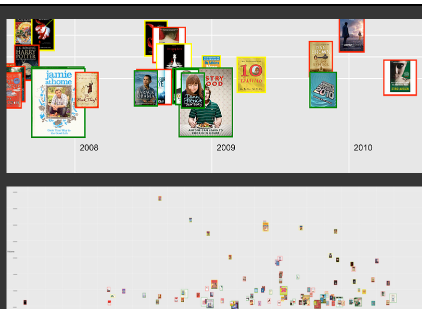
Exploratory Data Analysis

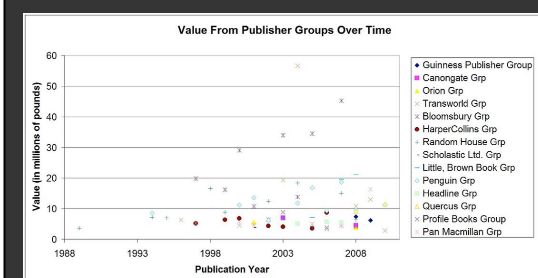
Maneesh Agrawala

CS 294-10: Visualization
Spring 2011

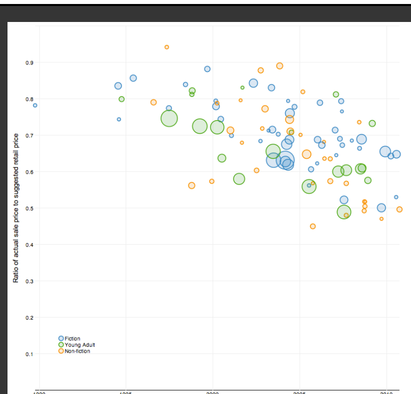
Last Time: Visualization Designs



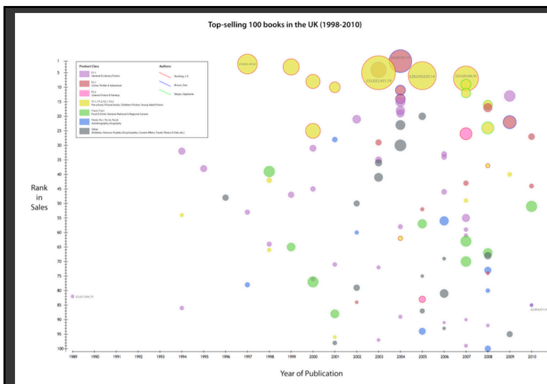
Siamak Faridani



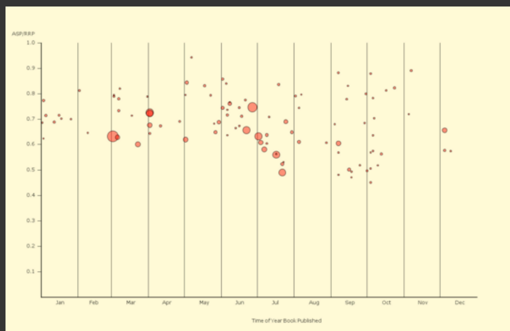
Saung Li



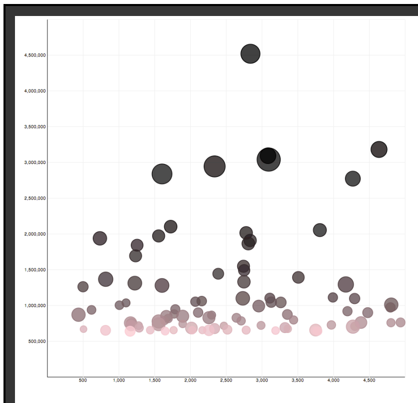
Michael Cohen



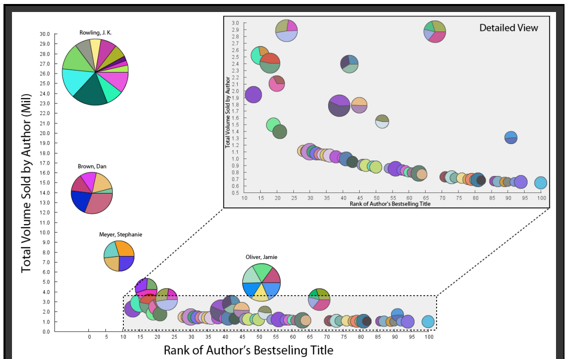
Sally Ahn



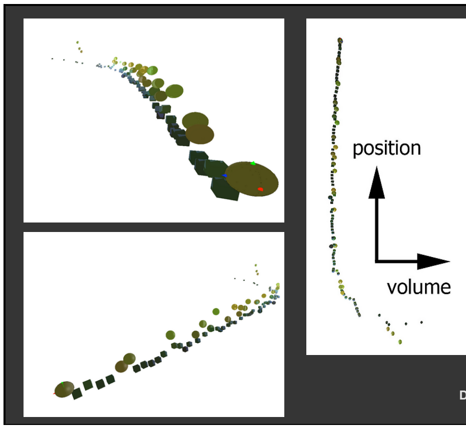
Matthew Can



Thomas Schluchter

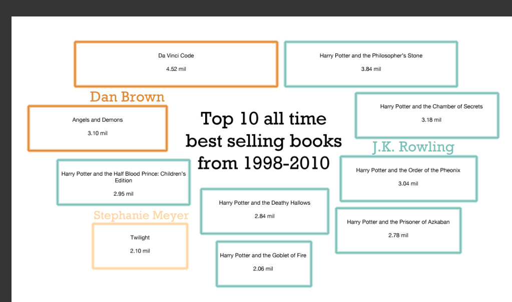


Michael Hsueh



Dan Lynch

Others



Bo Hee Kim

Manas Mittal

Last Last Time: Jock Mackinlay's APT

Combinatorics of encodings

Challenge:

Pick the best encoding from the exponential number of possibilities $(n+1)^8$

Principle of Consistency:

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

Principle of Importance Ordering:

Encode the most important information in the most effective way.

Mackinlay's expressiveness criteria

Expressiveness

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

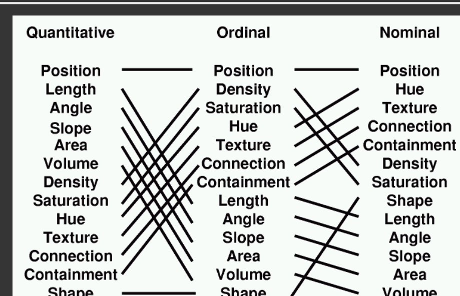
Mackinlay's effectiveness criteria

Effectiveness

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

Subject of perception lecture

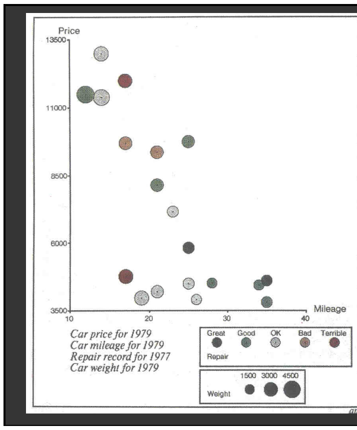
Mackinlay's ranking



Conjectured *effectiveness* of the encoding

Mackinlay’s design algorithm

- User formally specifies data model and type
- APT searches over design space
  - Tests expressiveness of each visual encoding
  - Generates image for encodings that pass test
  - Tests perceptual effectiveness of resulting image
- Outputs most effective visualization



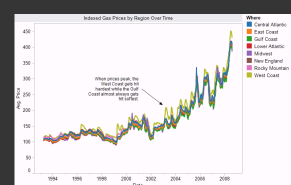
[Mackinlay, APT, 1986]

Announcements

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



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 Feb 14, 2011

Exploratory Data Analysis

The Future of Data Analysis, John W. Tukey 1962

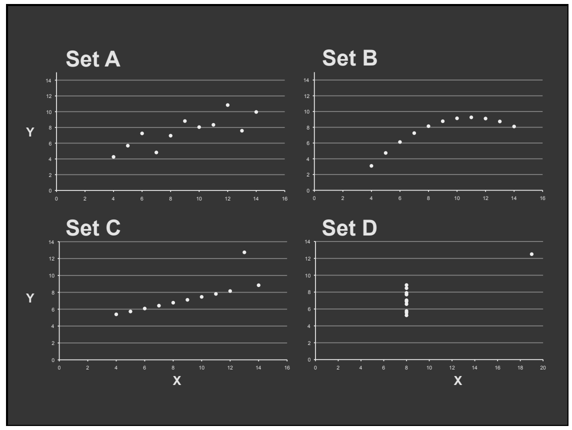
Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15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

$u_X = 9.0$ $\sigma_X = 3.317$ $Y^2 = 3 + 0.5\,X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

Anscombe 1973



Topics

Exploratory Data Analysis
- **Data Diagnostics**
- **Graphical Methods**
- **Data Transformation**

Confirmatory Data Analysis
- **Statistical Hypothesis Testing**

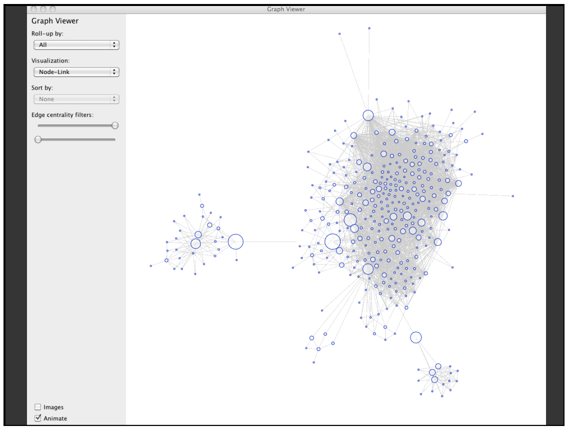
Data Diagnostics

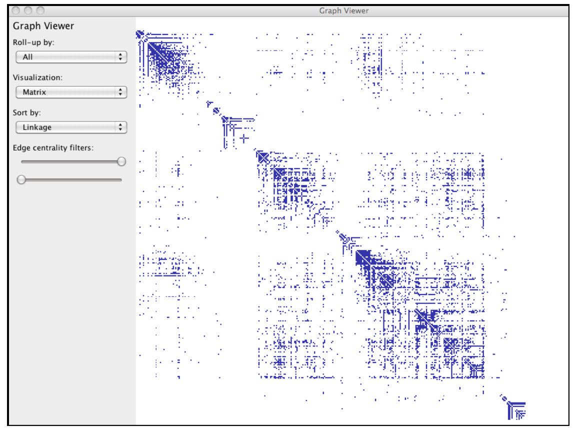
How to gauge the quality of a visualization?

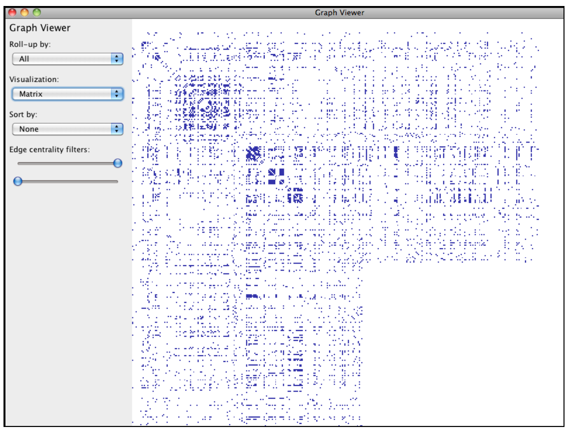
"The first sign that a visualization is good is that it shows you a problem in your data…

…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







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 & 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

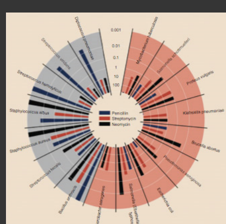
Collected prior to 1951

What questions might we ask?

Table 1: Burtin's data.

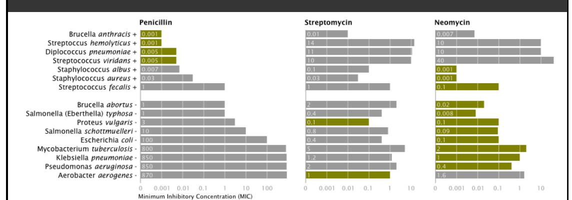
Bacteria	Antibiotic: Penicillin	Antibiotic: Streptomycin	Antibiotic: Neomycin	Gram Staining
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
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

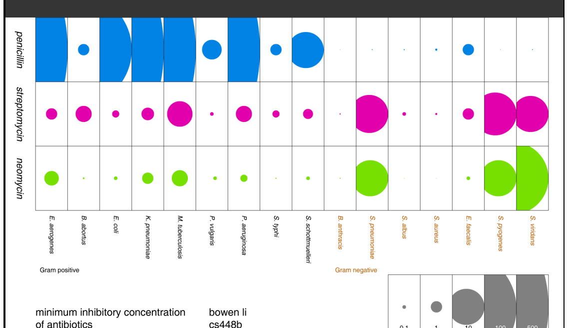


Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–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

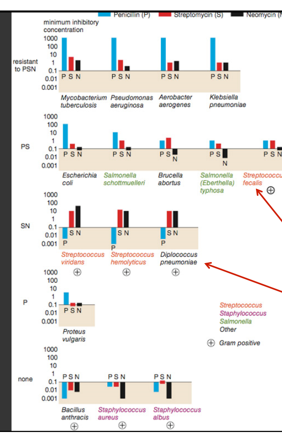
How do the drugs compare?



Mike Bostock, CS448B Winter 2009



Bowen Li, CS448B Fall 2009



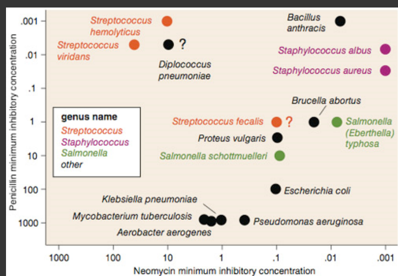
How do the bacteria group with respect to antibiotic resistance?

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

Common Data Transformations

Normalize	$y_i / \Sigma_i\, y_i$	(among others)
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

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