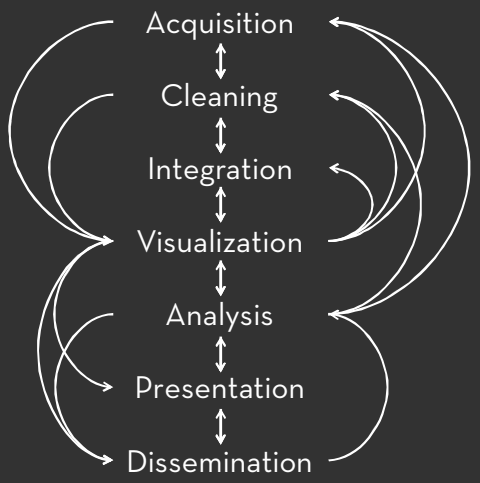
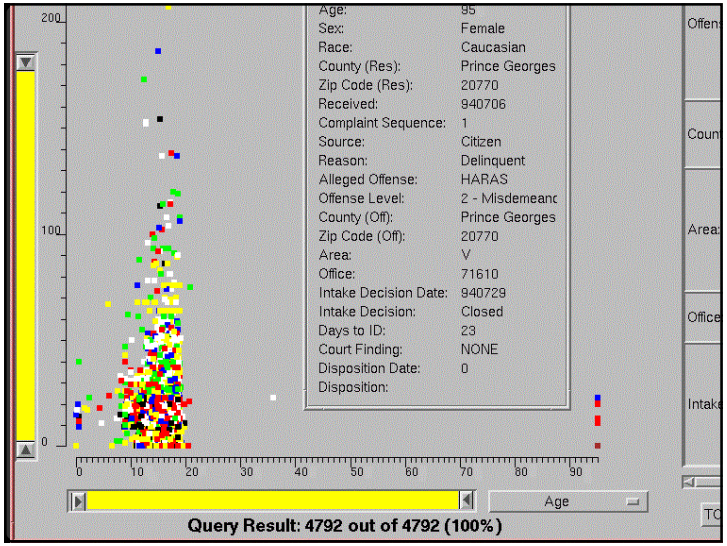
CS448G :: 11 Apr 2011

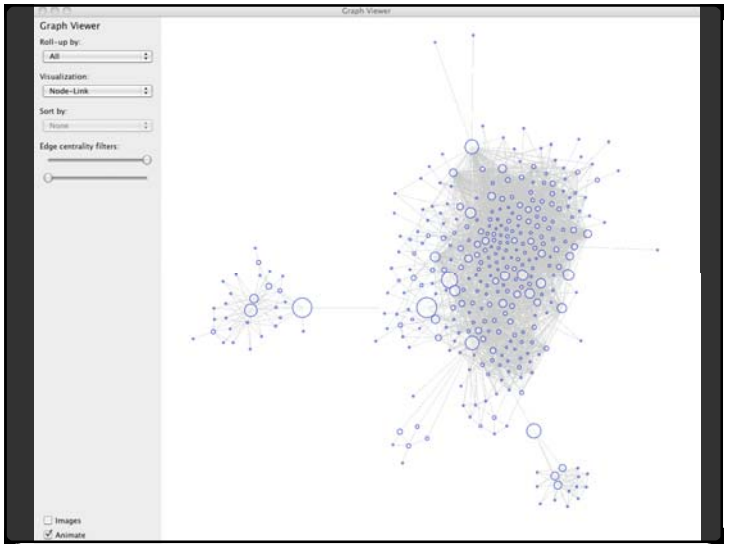
Data "Cleaning"

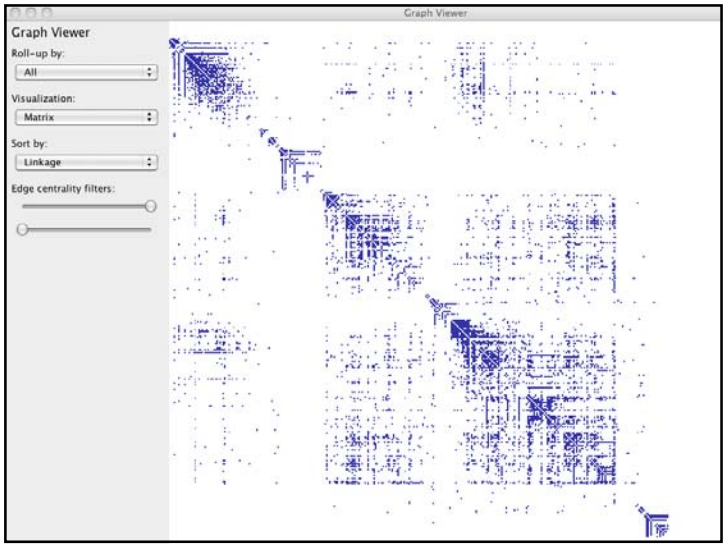
Jeffrey Heer Stanford University

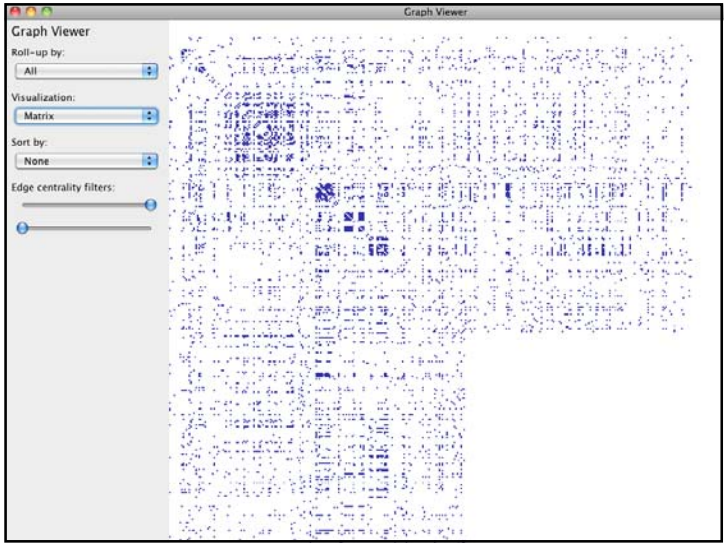


What is "dirty" data?



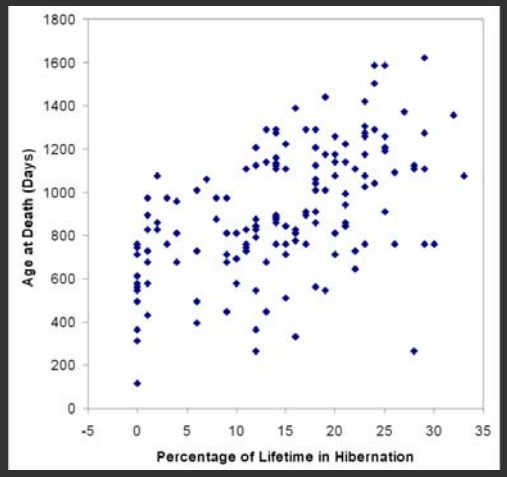




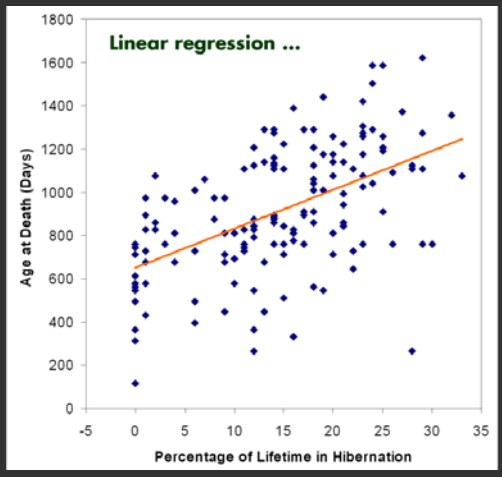


Visualize Friends by School

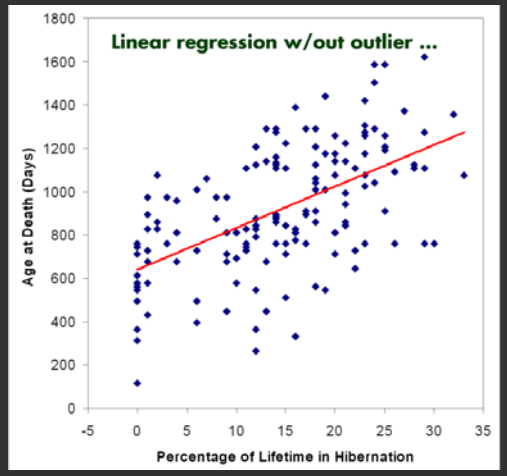
Berkeley	||||||||||||||||||||||||||||||
Cornell	||||
Harvard	||||||||
Harvard University	|||||||
Stanford	||||||||||||||||||
Stanford University	|||||||||
UC Berkeley	||||||||||||||||||||
UC Davis	|||||||||
University of California at Berkeley	||||||||||||||
University of California, Berkeley	||||||||||||||||||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



[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]

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	2780.3	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 3943.5	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	36134147	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 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!

Definitional Issues

What is *"clean"* data? What is *"clean enough"*?
Better yet, is the data ***"fit for a purpose"***?

Can I work with the data? (Is it *usable*)

Do I trust the data? (Is it *credible*)

Can I learn from it? (Is it *useful*)

Usability, Credibility, Usefulness

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

Data is ***credible*** if, according to one's subjective assessment, it is suitably representative of a phenomenon to enable productive analysis.

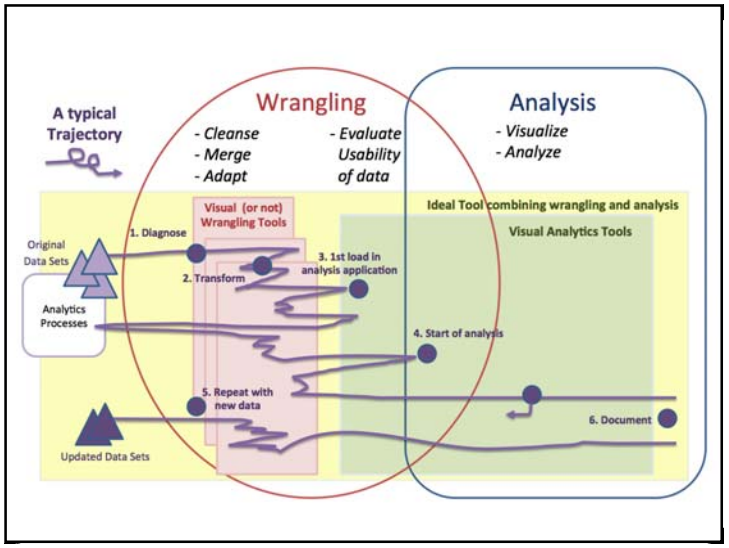
Data is ***useful*** if it is *usable*, *credible*, and responsive to one's inquiry.

Data Wrangling (n):

A process of iterative data exploration and transformation that enables analysis.

The goal of wrangling is to make data *useful*:

- Map data to a form readable by downstream tools (database, stats, visualization, ...)
- Identify, document, and (where possible) address data quality issues.



Data Wrangling Hypotheses

Data triage, exploration, cleaning and integration should be **integrated** and **iterative**.

Visual representations:
- Allow us to **see data quality issues**
- Can be an **input device for transformations**

The output of wrangling is a **transformation**; transformed data is only a by-product

Wrangling can be **amortized** via **collaboration**

Addressing Data Quality

Research Opportunities

Novel tools for data transformation
Focus of readings, discussion & guest lecture

Improve identification of data anomalies
Combine statistical and interactive techniques
Enable rapid correction / transformation

A Detective Story

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

Firm A		***Firm B***	
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

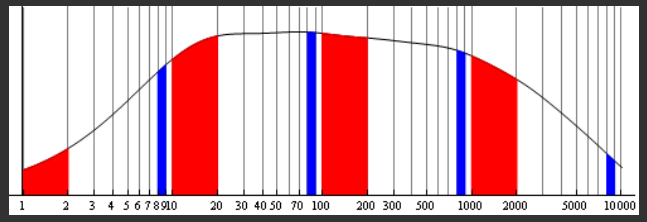
A Detective Story

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!

Model-Driven Data Validation

Deviations from the model *may* represent errors

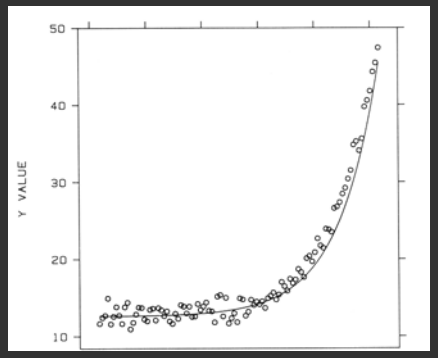
Find Statistical Outliers
- # std dev, Mahalanobis dist, nearest-neighbor, non-parametric methods, time-series models
- *Robust statistics* to combat noise, masking

Data Entry Errors
- Product codes: PZV, PZV, PZR, PZC, PZV
- Which of the above is most likely in error?

Opportunity: combine with visualization methods

Transforming data

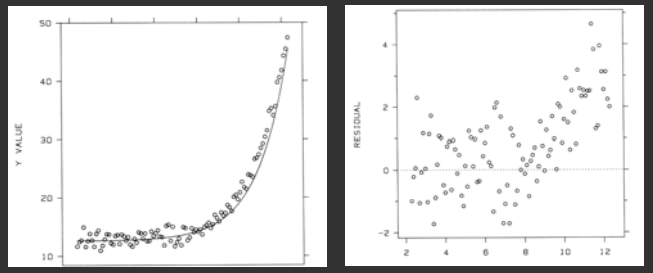
How well does curve fit data?



[Cleveland 85]

Plot the Residuals

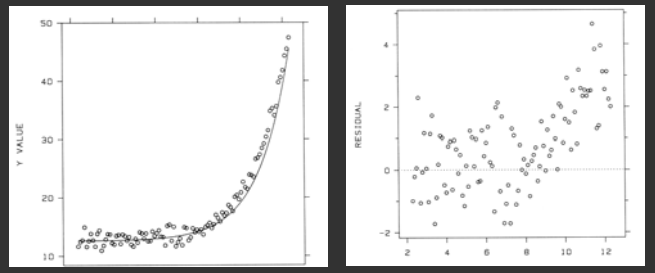
Plot vertical distance from best fit curve
Residual graph shows accuracy of fit



[Cleveland 85]

Multiple Plotting Options

Plot model in data space **Plot data in model space**



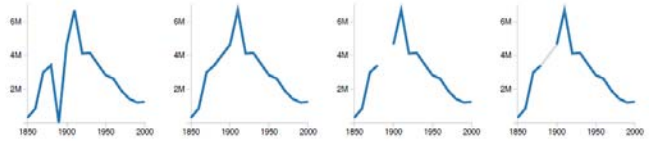
[Cleveland 85]

Research Opportunities

Novel tools for data transformation
- Focus of readings, discussion & guest lecture

Improve identification of data anomalies
- Combine statistical and interactive techniques
- Enable rapid correction / transformation

New visualization methods for data profiling
- Handle anomalies, scale & uncertainty
- Study the impact on perception & reasoning



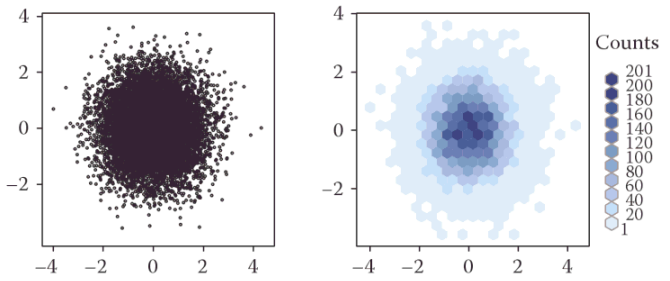
Plot the Data: US Farm Laborers

Year	*People*	*Year*	*People*
1850	0.4M	1930	4.0M
1860	0.8M	1940	3.5M
1870	3.2M	1950	3.0M
1880	3.5M	1960	2.8M
1890	**?**	1970	2.0M
1900	4.3M	1980	1.5M
1910	6.3M	1990	1.3M
1920	4.0M	2000	1.4M

Plot the Data: Sensor Readings

Schema:
- U - Number
- V - Number

Scatter plot! OK.
...but what if you have 3,141,590 points?



Research Opportunities

Novel tools for data transformation
- Focus of readings, discussion & guest lecture

Improve identification of data anomalies
- Combine statistical and interactive techniques
- Enable rapid correction / transformation

New visualization methods for data profiling
- Handle anomalies, scale & uncertainty
- Study the impact on perception & reasoning

A2 Part 2 – Due Mon 4/18

Devise your own hypotheses to test using MapReduce / Amazon EC2.

You may use the Wikipedia data, but we also encourage you to find your own (big) data set.

Example hypotheses:

- The distribution of first-letters in Wikipedia is uniform
- Most Twitter users have more "followees" than "followers"
- The words most associated with "democracy" on conservative blogs is different from those on liberal blogs

Discussants

Sean Kandel
Adrian Albert