Fitting Big Data into Business Statistics

Bob Stine
Department of Statistics, Wharton

Wharton
Department of Statistics

Issues from Big Data

Changes to the curriculum?

New skills for the four Vs of big data?

volume, variety, velocity, validity

Is it a zero sum game?

Changes to Curriculum

One semester course

Typical curriculum

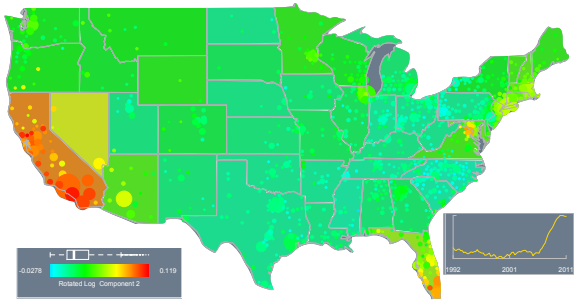
- Descriptive statistics
- Basic probability, random variables
- Sampling
- Inference
- **Regression**

Top level outline remains

- Retain major sequence
- Adapt what goes underneath

Offer a few suggestions ...

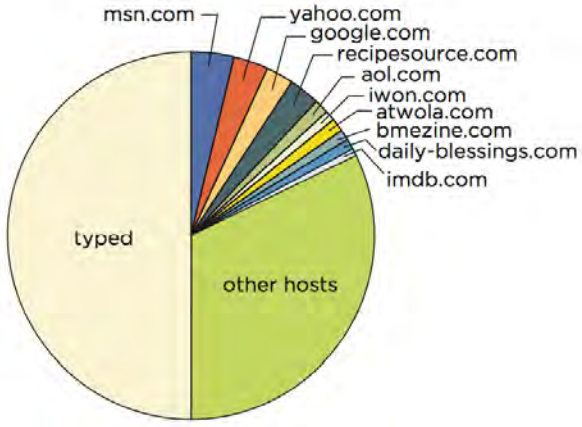
Changes: Descriptive



Greater variety

- Social networks, text, spatial
- More categorical variables, more bins

Amazon visitors referred by 11,142 sites. Same chart. Richer data.



Validity: lower data quality

- Eg. Missing data, coding errors, wrong labels

Changes: Probability

More coverage of dependence

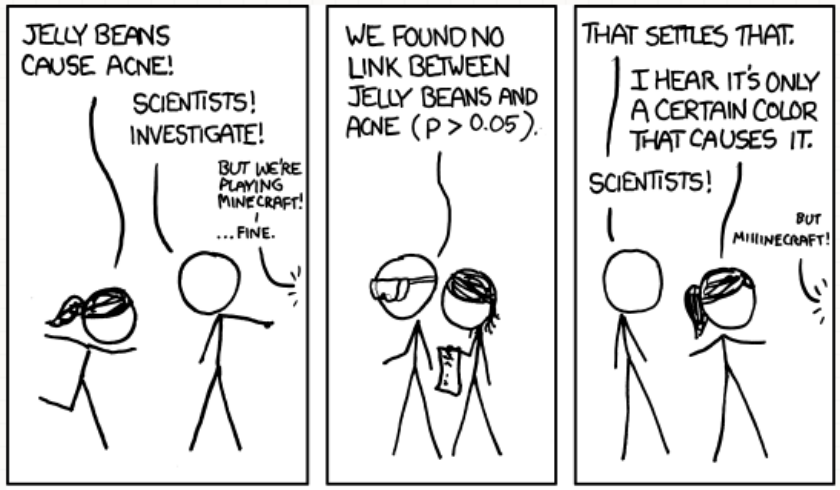
- That big sample might not be so big!
- Credit default recession of 2008
- Hurricane insurance ≠ Auto insurance

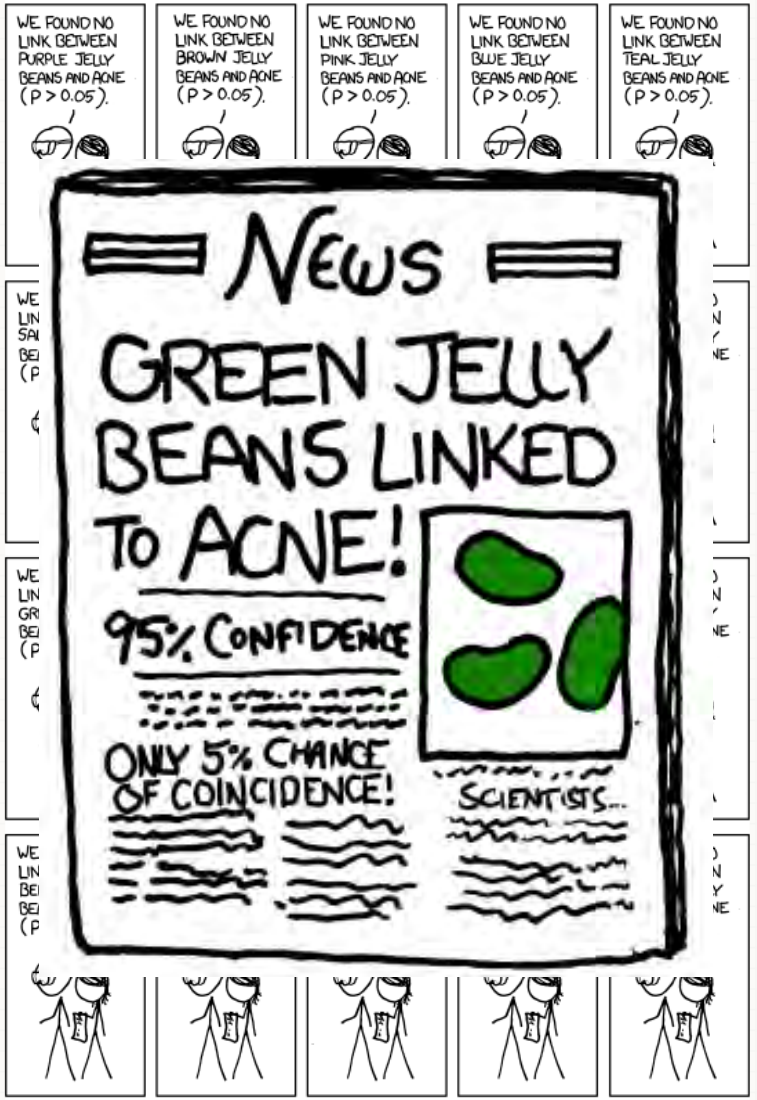
Greater awareness of multiplicity

- Boole's inequality, Bonferroni

p	P(max |z|>1.96)
1	0.05
5	0.23
25	0.72
100	0.99

Multiplicity Cartoon





Changes: Sampling

Return to designed experiments!

- Velocity: Detecting changes
- A/B testing in web design

The A/B Test: Inside the Technology That's Changing the Rules of Business

BY BRIAN CHRISTIAN 04.25.12 8:47 PM

Wired, 2012

Transactional vs sampled data

- Big data often derive from monitoring transactions, such as billing
- Don't forget dependence!
- Big ≠ good

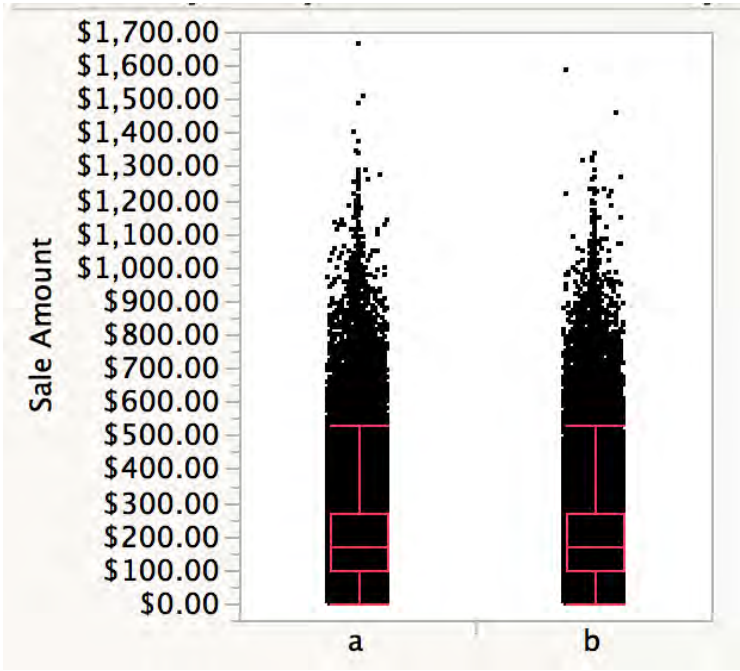
Changes: Inference

What's it mean to be statistically significant?

Effect size, economic value

$H_0: \mu_a = \mu_b$

t=3.2 with p-val ≈ 0.002



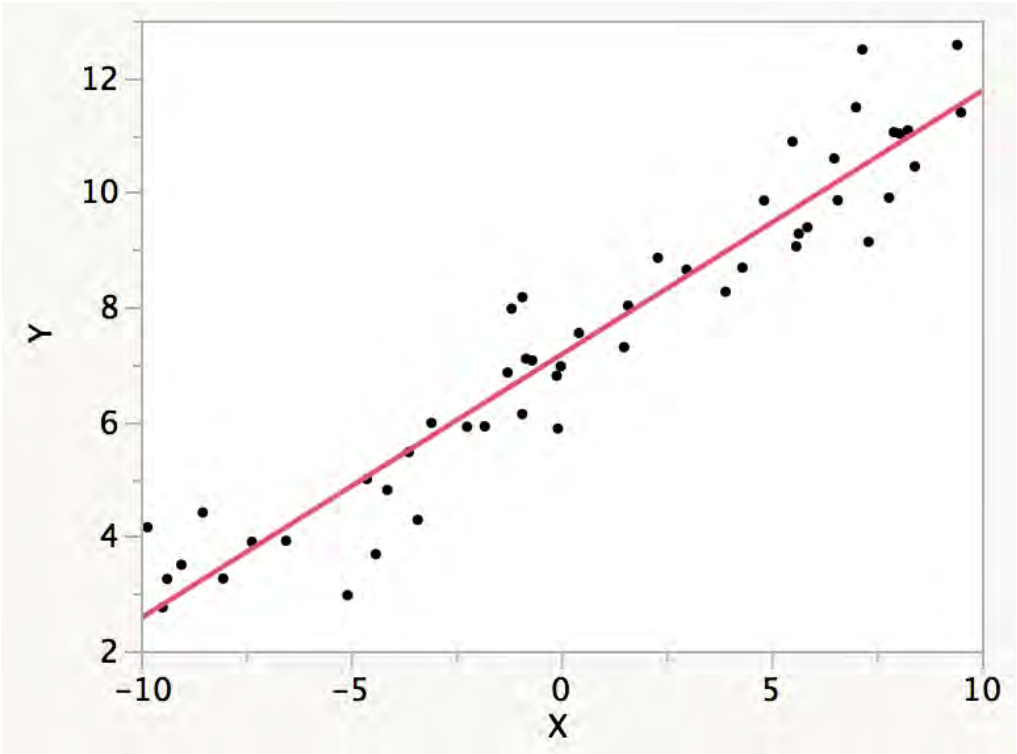
Another reason to prefer CIs over tests?

Upper CL Dif	-0.4575
Lower CL Dif	-2.0299

Changes: Regression

What will happen to this regression if the sample size increases from 50 to 100,000?

Idealized sampling from population, just more.



Linear Fit

Y = 7.1690493 + 0.4599771*X

Summary of Fit

RSquare	0.916
RSquare Adj	0.914
Root Mean Square Error	0.817
Mean of Response	7.460
Observations (or Sum Wgts)	50.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.169	0.116189	61.70	<.0001*
X	0.460	0.020097	22.89	<.0001*
X	0.49999	0.000549	910.28	<.0001*

Changes: Regression

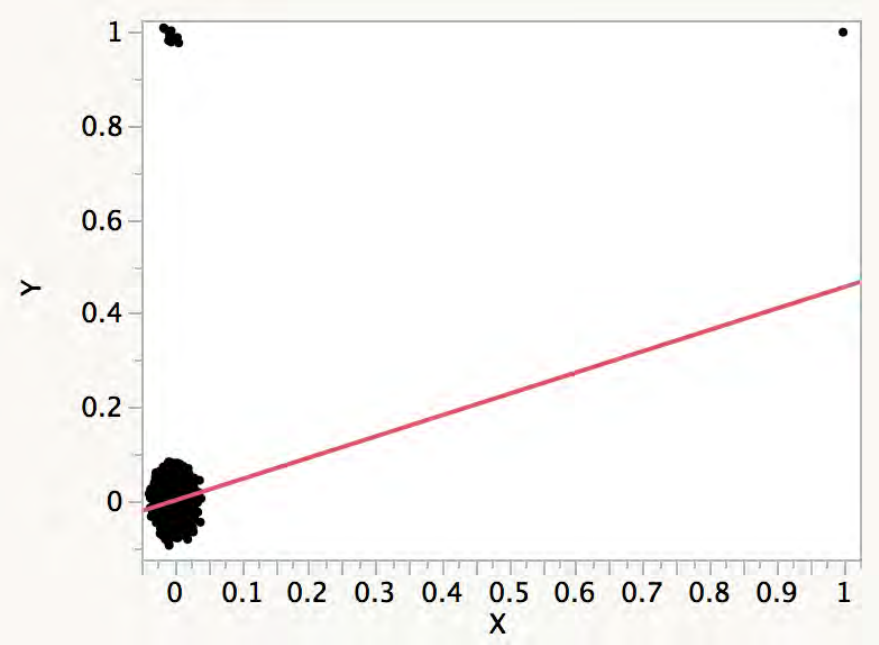
Outliers less important with more data?

CLT has to work with millions!

Example

Estimating standard errors

n=10,000 with 9,999 at x≈0 and one at x = 1.

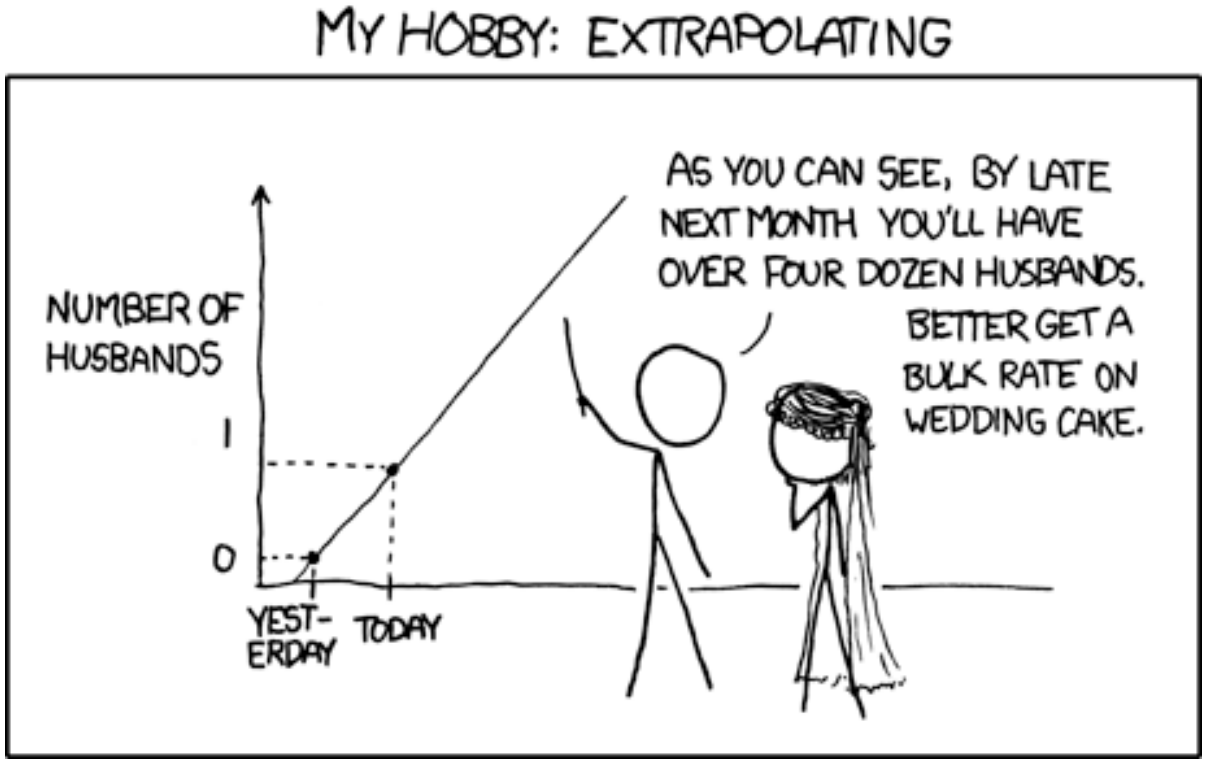


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0007143	0.000407	1.75	0.0795
X	0.4545881	0.028639	15.87	<.0001*

Changes: Regression

Extrapolation gets a lot easier when you have more unfamiliar variables



xkcd

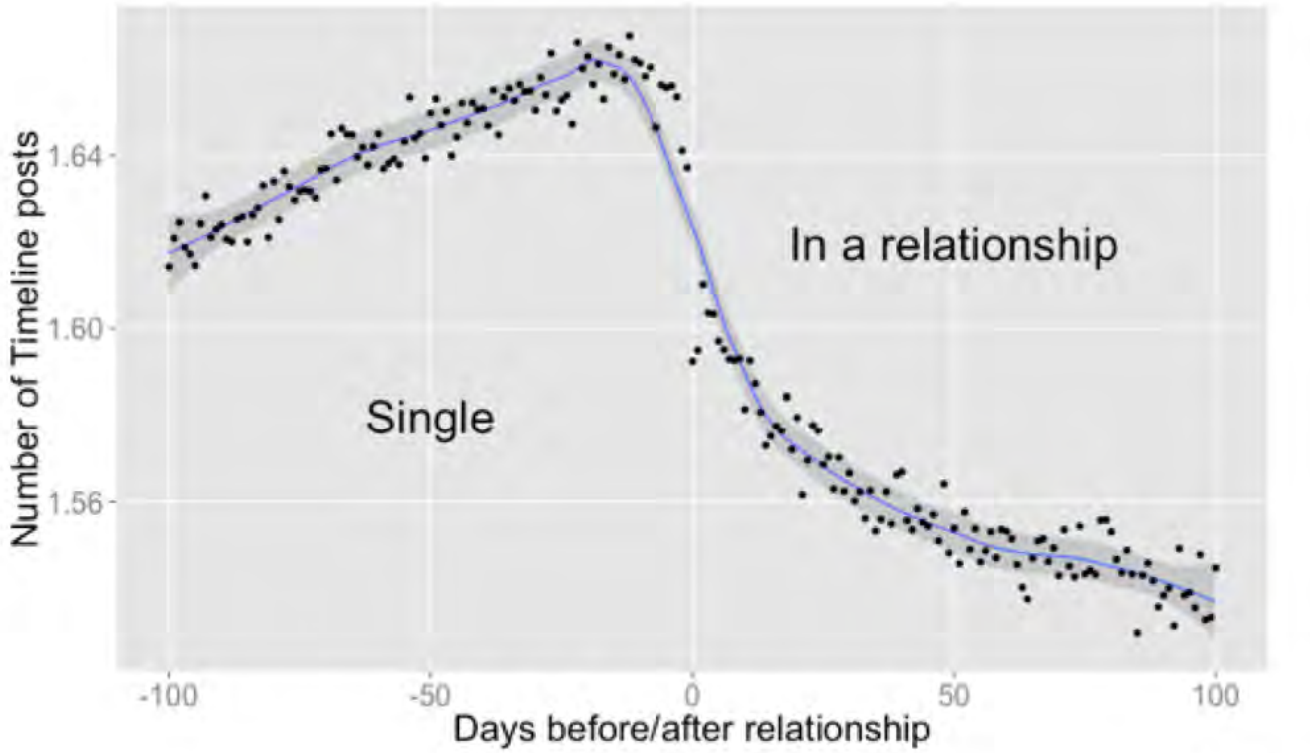
Changes: Regression

Why pretend everything is linear?

Regression estimates E(Y|X)

Smoothing via local averages so simple

Effect Size?



Facebook

Changes: Regression

Deciding what to use in a model

- Wide data tables
- Substantive choices… Business analytics?
- Data table does not always have what you need.

Automated search

- Modern versions of stepwise regression

Concern

- Over-fitting, building on multiplicity foundation
- Some notion of cross validation

Where to find the time?

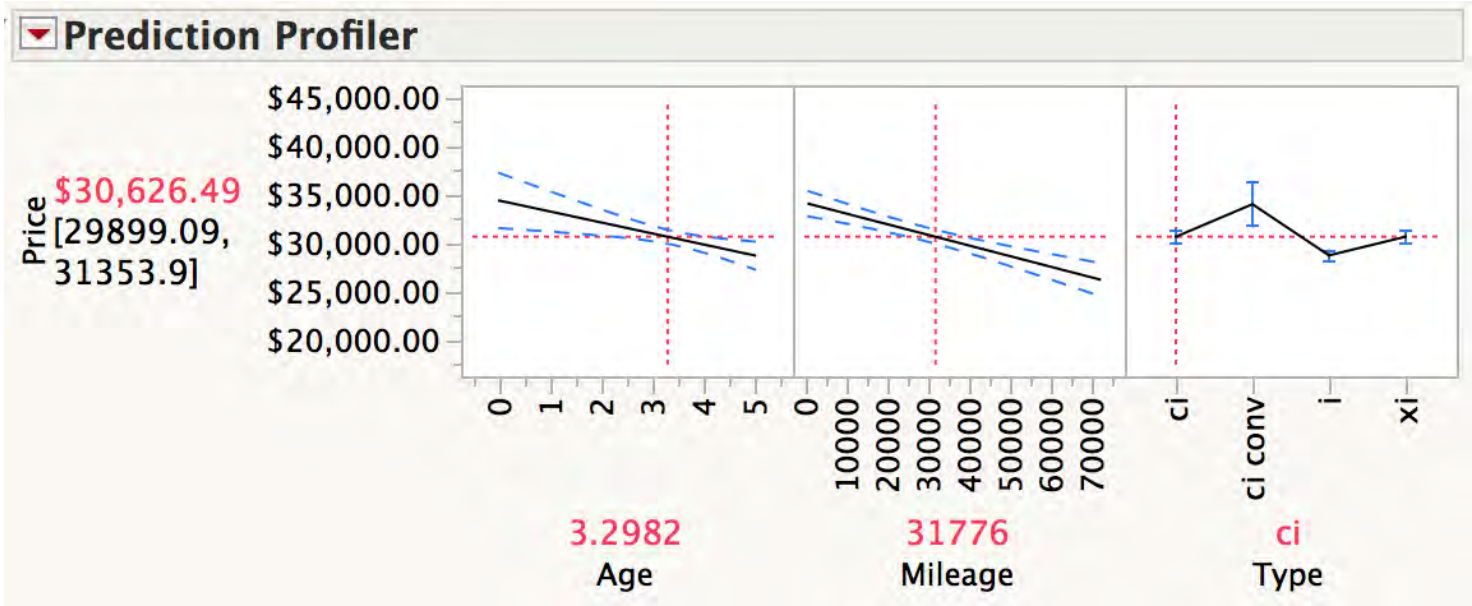
Exploit Technology

Greater reliance on software

Still need basic examples, but only illustrative

Automated tools

Animating regression models using profile tools



Issues from Big Data

Changes to the curriculum?

Not at the top level

New skills for the four Vs of big data?

Yes. Examples include
More categorical, multiplicity, effect sizes, model choice

Does it have to be a zero sum game?

Key ideas remain.
Opportunity to revise and update
Rich examples that span semester

What about...

- Host of other data issues
  - Database systems, layout
  - Information technology
  - Computer science
    - Hadoop, MapReduce,...
- Opportunity to collaborate
  - Change management, implementation
  - Information systems
  - Supply change
  - Marketing

Closing Remarks

Graphics remain important

- Maybe more important than ever
- Communication remains essential

Having clear sense of where an analysis is going is essential with big data.

- Easy to get lost
- Business analytics

Thanks!