Data Science: A Review

Professor Padhraic Smyth
Departments of Computer Science and Statistics
University of California, Irvine

Stats 5 Seminar, Winter 2016
## Schedule of Speakers

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 5th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>2</td>
<td>Jan 12th</td>
<td>Alex Ihler</td>
<td>Computer Science</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>3</td>
<td>Jan 19th</td>
<td>Jim Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>4</td>
<td>Jan 26th</td>
<td>Charless Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision using Machine Learning</td>
</tr>
<tr>
<td>5</td>
<td>Feb 2nd</td>
<td>Alfred Kobsa</td>
<td>Informatics</td>
<td>Data Privacy and Personalization</td>
</tr>
<tr>
<td>6</td>
<td>Feb 9th</td>
<td>Carter Butts</td>
<td>Social Sciences</td>
<td>Social Network Data Analysis</td>
</tr>
<tr>
<td>7</td>
<td>Feb 16th</td>
<td>Michael Carey</td>
<td>Computer Science</td>
<td>Systems for Big Data</td>
</tr>
<tr>
<td>8</td>
<td>Feb 23rd</td>
<td>Hal Stern</td>
<td>Statistics</td>
<td>Bayesian Statistics</td>
</tr>
<tr>
<td>9</td>
<td>Mar 1st</td>
<td>Stacey Hancock</td>
<td>Statistics</td>
<td>Randomization Methods and the R Language</td>
</tr>
<tr>
<td>10</td>
<td>Mar 8th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>The Future of Data Science</td>
</tr>
</tbody>
</table>
Themes in Data Science

Computing
- Databases (Carey)
- Software (Hancock)

Data Analysis Methods
- Machine learning (Ihler)
- Bayesian methods (Stern)
- Randomization (Hancock)

Applications
- Climate data (Randerson)
- Computer vision (Fowlkes)
- Social networks (Butts)

Human Interface
- Data Privacy (Kobsa)
- Interpretation (Stern)
Years 1 and 2: foundational courses in computer science, mathematics, statistics, including statistical computing

<table>
<thead>
<tr>
<th>Fall</th>
<th>12</th>
<th>Winter</th>
<th>13</th>
<th>Spring</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICS 31</td>
<td>4</td>
<td>ICS 32</td>
<td>4</td>
<td>ICS 33</td>
<td>4</td>
</tr>
<tr>
<td>Math 2A</td>
<td>4</td>
<td>Math 2B</td>
<td>4</td>
<td>Math 2D</td>
<td>4</td>
</tr>
<tr>
<td>Writing 39A</td>
<td>4</td>
<td>Writing 39B</td>
<td>4</td>
<td>Stats 7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stats 5</td>
<td>1</td>
<td>Writing 39C</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fall</th>
<th>16</th>
<th>Winter</th>
<th>14</th>
<th>Spring</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICS 6B</td>
<td>4</td>
<td>ICS 45C</td>
<td>4</td>
<td>Stats 68</td>
<td>4</td>
</tr>
<tr>
<td>Math 3A</td>
<td>4</td>
<td>ICS 51</td>
<td>6</td>
<td>Stats 120C</td>
<td>4</td>
</tr>
<tr>
<td>Stats 120A</td>
<td>4</td>
<td>Stats 120B</td>
<td>4</td>
<td>ICS 46</td>
<td>4</td>
</tr>
<tr>
<td>GE III</td>
<td>4</td>
<td></td>
<td>4</td>
<td>ICS 6D</td>
<td></td>
</tr>
</tbody>
</table>
Change of Major Requirements for Data Science

- Cumulative UC GPA: 2.7 or higher.

- 3.0 or higher average GPA and no grade lower than a C for ICS 31, ICS 32, and one of the following: Math 2A, Math 2B, Math 2D, ICS 6B, or ICS 6D.

- Students with more than 60 units will be reviewed on a case-by-case basis and may not be admitted to the major.

- Students will not be able to complete the degree in Data Science prior to Spring 2018.

If you are a freshman, contact Dr. Stacey Hancock, stacey.hancock@uci.edu, to inquire about getting a waiver to change into the major

More generally, you can talk to ICS Student Affairs Office Counselor if you are interested in changing your major to Data Science
**Computing**

- ICS 46: Data Structures
- IFMTX 43: Intro to Software Engineering
- CS 122A: Intro to Data Management
- CS 161: Design and Analysis of Algorithms
  (CS 131: Parallel and Distributed Computing)

**Data Analysis Methods**

- Stats 120 ABC: Intro to Prob and Stats
- Stats 68: Exploratory Data Analysis
- Stats 110-112: Statistical Methods
- CS 178: Machine Learning
  (Stats 140: Multivariate Statistics)
  (CS 172: Neural Networks/Deep Learning)

**Applications**

- Stats 170AB: Data Science Capstone Project
  (CS 121: Information Retrieval)
  (CS 122B: Project in Databases/Web Applications)
  (Summer internships, e.g., junior year)

**Human Interface**

- INF 143: Information Visualization
  (INF 131: Human Computer Interaction)
  (INF 161: Social Analysis of Computing)

(Electives shown in parentheses)
Years 3 and 4: more emphasis and specialization in data science topics such as machine learning, databases, visualization, advanced statistics

Year 3: sample program

<table>
<thead>
<tr>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stats 110, Statistical Methods for Data Analysis I</td>
<td>Stats 111, Statistical Methods for Data Analysis II</td>
<td>Stats 112, Statistical Methods for Data Analysis III</td>
</tr>
<tr>
<td>CS 161, Design and Analysis of Algorithms</td>
<td>CS 178, Machine Learning and Data-Mining</td>
<td>CS 122A, Introduction to Data Management</td>
</tr>
<tr>
<td>In4matx 43, Introduction to Software Engineering</td>
<td>ICS 139W, Critical Writing on Information Technology</td>
<td>In4matx 143, Information Visualization</td>
</tr>
<tr>
<td>GE IV/VIII,</td>
<td>GE III/VII,</td>
<td>GE VI,</td>
</tr>
</tbody>
</table>

Year 4: two-quarter capstone “data-intensive” project, + statistics and CS electives
The Life Cycle of Data Analysis

Data Measurement
Acquiring data

Data Organization
Storing, cleaning, sorting data

Data Exploration
Exploratory data analysis

Data Analysis
Statistics, machine learning

Interpretation
Analyzing, understanding, evaluating

Decisions and Actions
Taking action
The Life Cycle of Data Analysis

Measure
   ↓
Organize
   ↓
Explore
   ↓
Analyze
   ↓
Interpret
   ↓
Decide
The Life Cycle of Data Analysis

What are different ways we can collect data?
### Data in Matrix Form

#### Measurements

<table>
<thead>
<tr>
<th>ID</th>
<th>Income</th>
<th>Age</th>
<th>Monthly Debt</th>
<th>Good Risk?</th>
</tr>
</thead>
<tbody>
<tr>
<td>18276</td>
<td>65,000</td>
<td>55</td>
<td>2200</td>
<td>Yes</td>
</tr>
<tr>
<td>72514</td>
<td>28,000</td>
<td>19</td>
<td>1500</td>
<td>No</td>
</tr>
<tr>
<td>28163</td>
<td>120,000</td>
<td>62</td>
<td>1800</td>
<td>Yes</td>
</tr>
<tr>
<td>17265</td>
<td>90,000</td>
<td>35</td>
<td>4500</td>
<td>No</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>61524</td>
<td>35,000</td>
<td>22</td>
<td>900</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Text Collections

NYT
330,000 articles

Enron
250,000 emails

NSF/ NIH
100,000 grants

16 million Medline articles

Pennsylvania Gazette
80,000 articles
1728-1800
500 million 30-day active users

The Friendship graph

500M users each connect to an average of 130 other users = ~ 60 Billion Edges

Over 30 billion pieces of content shared every month

Over 3 billion photos uploaded each month

Graphics from Lars Backstrom, ESWC 2011
Eye-Tracking: The Golden Triangle for Search
from Hotchkiss, Alston, Edwards, 2005; EnquiroResearch
Ebird.org

Over 1.5 million submissions per month

From Wood et al, PLOS Biology, 2011
NASA’s MODIS satellite

entire planet
250m resolution
37 spectral bands
every 2 days
Real-Time Sports Statistics

Other ideas on how we can collect data?
The Life Cycle of Data Analysis

Measure

Organize

Explore

Analyze

Interpret

Decide

What are different ways we can collect data?

How might data collection influence our results?
Email network from 500 HP researchers
Geolocated Tweets in Southern California over 6 months
Geolocated Tweets around UC Irvine
Typical Challenges with “Large Data”

- Observational/secondary
  - Collected for some other purposes, e.g., from social media

- Noisy, Biased
  - Measurement mechanisms are often unclear, subject to whims of data owners

- Size
  - Size brings complexity: in data management, in interactive analysis, etc

- Complex and Multisource
  - e.g., text data, location data, demographic data: poses challenge in analysis

- Non-Stationary
  - Changing over time: trends, seasonality, etc
Other ideas on how data collection might influence our analysis?
The Life Cycle of Data Analysis

1. Measure
2. Organize
3. Explore
4. Analyze
5. Interpret
6. Decide

Why is data organization important?
## Data in Matrix Form

### Measurements

<table>
<thead>
<tr>
<th>ID</th>
<th>Income</th>
<th>Age</th>
<th>…..</th>
<th>Monthly Debt</th>
<th>Good Risk?</th>
</tr>
</thead>
<tbody>
<tr>
<td>18276</td>
<td>65,000</td>
<td>55</td>
<td>…..</td>
<td>2200</td>
<td>Yes</td>
</tr>
<tr>
<td>72514</td>
<td>28,000</td>
<td>19</td>
<td>…..</td>
<td>1500</td>
<td>No</td>
</tr>
<tr>
<td>28163</td>
<td>120,000</td>
<td>62</td>
<td>…..</td>
<td>1800</td>
<td>Yes</td>
</tr>
<tr>
<td>17265</td>
<td>90,000</td>
<td>35</td>
<td>…..</td>
<td>4500</td>
<td>No</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…..</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…..</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>61524</td>
<td>35,000</td>
<td>22</td>
<td>…..</td>
<td>900</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Data Engineering at Web Scale
How Far Away are the Data?

CPU

RAM

Disk

$10^{-8}$ seconds

$10^{-3}$ seconds

Random Access Times
How Far Away are the Data?

Effective Distances

CPU

RAM

Disk

1 meter

100 km
Why Use a DBMS?

- Data independence
- Efficient (automatic) data access
- Reduced development time
- Data integrity and security
- Uniform data administration
- Concurrent access and recovery from crashes

From Professor Mike Carey’s Stats 5 presentation
Other ideas on why data organization is important?
The Life Cycle of Data Analysis

Measure
─
Organize
─
Explore
─
Analyze
─
Interpret
─
Decide

What are ways we can explore data?
Histogram of Unimodal Data

1000 data points simulated from a Normal distribution, mean 10, variance 1, 30 bins
Histograms: Unimodal Data

100 data points from a Normal, mean 10, variance 1, with 5, 10, 30 bins
Histogram of Multimodal Data

15000 data points simulated from a mixture of 3 Normal distributions, 300 bins
What will the mean or median tell us about this data?
The Life Cycle of Data Analysis

- Measure
- Organize
- Explore
- Analyze
- Interpret
- Decide

What are ways we can explore data?

How can exploration help us?
Histogram with Outliers

Pima Indians Diabetes Data, From UC Irvine Machine Learning Repository

Number of Individuals

X values

UC Irvine
Histogram with Outliers

Number of Individuals

blood pressure = 0?

Diastolic Blood Pressure

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Repository
Linear Correlation Coefficient

- Measures the degree of linear dependence of two variables
- Linear correlation coefficient is defined as:

\[
\rho(X, Y) = \frac{\sum_{i=1}^{n} (x(i) - \bar{x})(y(i) - \bar{y})}{\left(\sum_{i=1}^{n} (x(i) - \bar{x})^2 \sum_{i=1}^{n} (y(i) - \bar{y})^2\right)^{1/2}}
\]

- Ranges between -1 and +1
- Note: lack of linear correlation does not imply lack of dependence
Examples of X-Y plots and linear correlation values
# Data Set on Housing Prices in Boston

(widely used data set in research on regression models)

<table>
<thead>
<tr>
<th></th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CRIM</td>
<td>per capita crime rate by town</td>
</tr>
<tr>
<td>2</td>
<td>ZN</td>
<td>proportion of residential land zoned for lots over 25,000 ft²</td>
</tr>
<tr>
<td>3</td>
<td>INDUS</td>
<td>proportion of non-retail business acres per town</td>
</tr>
<tr>
<td>4</td>
<td>NOX</td>
<td>Nitrogen oxide concentration (parts per 10 million)</td>
</tr>
<tr>
<td>5</td>
<td>RM</td>
<td>average number of rooms per dwelling</td>
</tr>
<tr>
<td>6</td>
<td>AGE</td>
<td>proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>7</td>
<td>DIS</td>
<td>weighted distances to five Boston employment centres</td>
</tr>
<tr>
<td>8</td>
<td>RAD</td>
<td>index of accessibility to radial highways</td>
</tr>
<tr>
<td>9</td>
<td>TAX</td>
<td>full-value property-tax rate per $10,000</td>
</tr>
<tr>
<td>10</td>
<td>PTRATIO</td>
<td>pupil-teacher ratio by town</td>
</tr>
<tr>
<td>11</td>
<td>MEDV</td>
<td>Median value of owner-occupied homes in $1000's</td>
</tr>
</tbody>
</table>
Matrix of Pairwise Linear Correlations

Data on characteristics of Boston housing

- Crime Rate
- Industry
- Nitrous oxide
- Average # rooms
- Proportion of old houses
- Highway accessibility
- Property tax rate
- Student-teacher ratio

- Percentage of large residential lots
- Distance to employment centers
- Median house value
Examples of X-Y plots and linear correlation values
Examples of X-Y plots and linear correlation values

Non-Linear Dependence

Lack of linear correlation does not imply lack of dependence

Linear Dependence
Guess the Linear Correlation Values for each Data Set

DATA SET 1

DATA SET 2

DATA SET 3

DATA SET 4

Actual Correlation Values

DATA SET 1

Correlation = 0.82

DATA SET 2

Correlation = 0.82

DATA SET 3

Correlation = 0.82

DATA SET 4

Correlation = 0.82

## Summary Statistics for each Data Set

**Summary Statistics of Data Set 1**  
\( N = 11 \)  
Mean of \( X = 9.0 \)  
Mean of \( Y = 7.5 \)  
Intercept = 3  
Slope = 0.5  
Correlation = 0.82

**Summary Statistics of Data Set 2**  
\( N = 11 \)  
Mean of \( X = 9.0 \)  
Mean of \( Y = 7.5 \)  
Intercept = 3  
Slope = 0.5  
Correlation = 0.82

**Summary Statistics of Data Set 3**  
\( N = 11 \)  
Mean of \( X = 9.0 \)  
Mean of \( Y = 7.5 \)  
Intercept = 3  
Slope = 0.5  
Correlation = 0.82

**Summary Statistics of Data Set 4**  
\( N = 11 \)  
Mean of \( X = 9.0 \)  
Mean of \( Y = 7.5 \)  
Intercept = 3  
Slope = 0.5  
Correlation = 0.82

---

Other ideas on how we can explore data?
The Life Cycle of Data Analysis

Measure
  ↓
Organize
  ↓
Explore
  ↓
Analyze
  ↓
Interpret
  ↓
Decide

What are ways we can explore data?

How can exploration help us?

What are the limits of manual exploration?
Using Color to Show Group Information in Scatter Plots

Iris classification data set, 3 classes

Figure from www.originlab.com
Chernoff Faces

- Variable values associated with facial characteristic parameters, e.g., head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length and degree of mouth opening

- Limitations?
  - Only up to 10 or so dimensions
  - Overemphasizes certain variables because of our perceptual biases
# Chernoff Faces

## 2005 National League

alexreisner.com/baseball/stats/chernoff

<table>
<thead>
<tr>
<th></th>
<th>PCT</th>
<th>H</th>
<th>HR</th>
<th>BB</th>
<th>SB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>0.475</td>
<td>1419</td>
<td>191</td>
<td>606</td>
<td>67</td>
</tr>
<tr>
<td>ATL</td>
<td>0.556</td>
<td>1453</td>
<td>184</td>
<td>534</td>
<td>92</td>
</tr>
<tr>
<td>CHI</td>
<td>0.488</td>
<td>1506</td>
<td>194</td>
<td>419</td>
<td>65</td>
</tr>
<tr>
<td>CIN</td>
<td>0.451</td>
<td>1453</td>
<td>222</td>
<td>611</td>
<td>72</td>
</tr>
<tr>
<td>COL</td>
<td>0.414</td>
<td>1477</td>
<td>150</td>
<td>509</td>
<td>65</td>
</tr>
<tr>
<td>FLO</td>
<td>0.512</td>
<td>1499</td>
<td>128</td>
<td>512</td>
<td>96</td>
</tr>
<tr>
<td>HOU</td>
<td>0.549</td>
<td>1400</td>
<td>161</td>
<td>481</td>
<td>115</td>
</tr>
<tr>
<td>LAD</td>
<td>0.438</td>
<td>1374</td>
<td>149</td>
<td>541</td>
<td>58</td>
</tr>
<tr>
<td>MIL</td>
<td>0.500</td>
<td>1413</td>
<td>175</td>
<td>531</td>
<td>79</td>
</tr>
<tr>
<td>NYM</td>
<td>0.512</td>
<td>1421</td>
<td>175</td>
<td>486</td>
<td>153</td>
</tr>
<tr>
<td>PHI</td>
<td>0.543</td>
<td>1494</td>
<td>167</td>
<td>639</td>
<td>116</td>
</tr>
<tr>
<td>PIT</td>
<td>0.414</td>
<td>1445</td>
<td>139</td>
<td>471</td>
<td>73</td>
</tr>
<tr>
<td>SDP</td>
<td>0.506</td>
<td>1416</td>
<td>130</td>
<td>600</td>
<td>99</td>
</tr>
<tr>
<td>SFG</td>
<td>0.463</td>
<td>1427</td>
<td>128</td>
<td>431</td>
<td>71</td>
</tr>
<tr>
<td>STL</td>
<td>0.617</td>
<td>1494</td>
<td>170</td>
<td>534</td>
<td>83</td>
</tr>
<tr>
<td>WAS</td>
<td>0.500</td>
<td>1367</td>
<td>117</td>
<td>491</td>
<td>45</td>
</tr>
</tbody>
</table>

Arizona | Atlanta | Chicago | Cincinnati

Colorado | Florida | Houston | Los Angeles

Milwaukee | New York | Philadelphia | Pittsburgh

San Diego | San Francisco | St. Louis | Washington
The Life Cycle of Data Analysis

Measure → Organize → Explore → Analyze → Interpret → Decide

What data analysis techniques are available?
Learning to Predict with Weighted Sums

This is known as a logistic regression model

\[
f(x) = \hat{P}(Y = 1 | x) = \frac{1}{1 + e^{(-\sum_{j=1}^{d} \alpha_j x_j)}}
\]

Each “edge” has a weight or parameter, \( \alpha_j \)
A Neural Network with 1 Hidden Layer

Can recursively create more complex prediction models

Many more weights now....requires more data to estimate
Deep Learning: Models with 2 or More Hidden Layers

We can build on this idea to create “deep models” with many hidden layers.

The model is now a very flexible highly non-linear function.
Figure from Krizhevsky, Sutskever, Hinton, 2012
Other general approaches for analyzing data?
The Life Cycle of Data Analysis

Measure

Organize

Explore

Analyze

Interpret

Why is interpretation important?

Decide
Today's Random Medical News

According to a report released today...

Can cause in...

Hypothermia
Breast cancer
Spontaneous remission
Glaucoma
A feeling of well-being
Depression

In...

Two-income families
Men 25-40
Overweight smokers
Rats

From the New England Journal of Panic-Inducing Gobbledygook
Example: a data set with

- 100 independent variables
- Simulate 50 data vectors
- Compute the correlation of all pairs of variables from the data
- This gives us 50*49/2 correlation values

What do you think these correlation values will look like if we plot them as a histogram?
Conclusion: even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.

This is sometimes referred to as “data fishing”
People who drowned after falling out of a fishing boat correlates with Marriage rate in Kentucky

[Graph showing correlation between People who drowned and Marriage rate]

<table>
<thead>
<tr>
<th>People who drowned after falling out of a fishing boat</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths (US) (CDC)</td>
<td>19</td>
<td>16</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Marriage rate in Kentucky</td>
<td>10.9</td>
<td>9.8</td>
<td>9</td>
<td>9</td>
<td>9.1</td>
<td>8.8</td>
<td>8.7</td>
<td>8.4</td>
<td>7.8</td>
<td>7.9</td>
<td>7.6</td>
<td>7.4</td>
</tr>
</tbody>
</table>

Correlation: 0.952407

Graphics from http://www.tylervigen.com/
Number of people who died by becoming tangled in their bedsheets correlates with Total revenue generated by skiing facilities (US)

Correlation: 97%  Sources: CDC & US Census  tylervigen.com

<table>
<thead>
<tr>
<th>Year</th>
<th>Deaths (US)</th>
<th>Dollars in millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>327</td>
<td>1,551</td>
</tr>
<tr>
<td>2001</td>
<td>456</td>
<td>1,635</td>
</tr>
<tr>
<td>2002</td>
<td>509</td>
<td>1,801</td>
</tr>
<tr>
<td>2003</td>
<td>497</td>
<td>1,827</td>
</tr>
<tr>
<td>2004</td>
<td>596</td>
<td>1,956</td>
</tr>
<tr>
<td>2005</td>
<td>573</td>
<td>1,989</td>
</tr>
<tr>
<td>2006</td>
<td>661</td>
<td>2,178</td>
</tr>
<tr>
<td>2007</td>
<td>741</td>
<td>2,257</td>
</tr>
<tr>
<td>2008</td>
<td>809</td>
<td>2,476</td>
</tr>
<tr>
<td>2009</td>
<td>717</td>
<td>2,438</td>
</tr>
</tbody>
</table>

Correlation: 0.969724

Graphics from http://www.tylervigen.com/
Big Data Example: Analyzing Changes in Rainfall Patterns over Time

(joint work with Dr. Andy Robertson, IRI, Columbia University)

Billions of measurements about rainfall....

...but how much data do we really have for a problem like this?
Automated Essay Grading

From Inside Higher Ed, April 2012

“In the most comprehensive review to date of automated essay grading software...the study, funded by the William and Flora Hewlett Foundation, compared the software-generated ratings given to more than 22,000 short essays, written by students in junior high schools and high school sophomores, to the ratings given to the same essays by trained human readers.”

“The differences, across a number of different brands of automated essay scoring software (AES) and essay types, were minute. “

Human graders: 20 to 30 essays an hour
Automated: potentially millions per hour
Fooling Automated Essay Grading

From New Statesman and New York Times, April 2012

Les Perelman, MIT, experimented with different essays to test the Educational Testing Service (ETS)’s automated eRater program.

All of his essays received a perfect score.

SAT prompt:
"The rising cost of a college education is the fault of students who demand that colleges offer students luxuries unheard of by earlier generations of college students -- single dorm rooms, private bathrooms, gourmet meals, etc."

Discuss the extent to which you agree or disagree with this opinion. Support your views with specific reasons and examples from your own experience, observations, or reading.
Teaching assistants are paid an excessive amount of money. The average teaching assistant makes six times as much money as college presidents. In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, a starring roles in motion pictures.
Portions of a Perfect-Scoring Essay

In Heart of Darkness, Mr. Kurtz is a teaching assistant because of his connections, and he ruins all the universities that employ him. Finally, teaching assistants are able to exercise mind control over the rest of the university community. The last reason to write this way is the most important. Once you have it down, you can use it for practically anything. Does God exist? Well, you can say yes and give three reasons, or no and give three different reasons. It doesn't really matter.
Other ideas on why interpretation of our analysis is important?
The Life Cycle of Data Analysis

Measure

Organize

Explore

Analyze

Interpret

Decide

How are decisions influenced by privacy?
BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES

Source: US Department of Commerce and country specific legislation

Source: Forrester Research, Inc.
The Future of Data Science?

What types of new data might we collect?

What new application areas might emerge?

What new techniques might be developed?

What are the societal implications of “big data”? 