Data Science: A Review

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

Stats 5 Seminar, Winter 2017
# Schedule of Speakers

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 10</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>Jan 17</td>
<td>Sameer Singh</td>
<td>Computer Science</td>
<td>Machine Learning for Text</td>
</tr>
<tr>
<td>Jan 24</td>
<td>Charless Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision with Machine Learning</td>
</tr>
<tr>
<td>Jan 31</td>
<td>Pierre Baldi</td>
<td>Computer Science</td>
<td>Deep Learning and Neural Networks</td>
</tr>
<tr>
<td>Feb 7</td>
<td>Hernando Ombao</td>
<td>Statistics</td>
<td>Brain Signal Analysis</td>
</tr>
<tr>
<td>Feb 14</td>
<td>Zhoaxia Yu</td>
<td>Statistics</td>
<td>Multivariate Data Analysis</td>
</tr>
<tr>
<td>Feb 21</td>
<td>Ramesh Jain</td>
<td>Computer Science</td>
<td>Event and Web Data</td>
</tr>
<tr>
<td>Feb 28</td>
<td>James Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>Mar 7</td>
<td>Mark Steyvers</td>
<td>Cognitive Sciences</td>
<td>Data from Crowdsourcing</td>
</tr>
<tr>
<td>Mar 14</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Review of Data Science</td>
</tr>
</tbody>
</table>
Data Science

Statistical Methods
Computing
Domain Knowledge
Human Interface
Years 1 and 2: foundational courses in computer science, mathematics, statistics, including statistical computing

<table>
<thead>
<tr>
<th>2015-16, First Year: 41 units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
</tr>
<tr>
<td>ICS 31</td>
</tr>
<tr>
<td>Math 2A</td>
</tr>
<tr>
<td>Writing 39A</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2016-17, Second Year: 46 units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
</tr>
<tr>
<td>ICS 6B</td>
</tr>
<tr>
<td>Math 3A</td>
</tr>
<tr>
<td>Stats 120A</td>
</tr>
<tr>
<td>GE III</td>
</tr>
</tbody>
</table>
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
Change of Major Requirements for Data Science Major

• 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 (100 in practice) 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.
• For questions, contact
  – Professor Zhaoxia Yu at zhaoxia@ics.uci.edu
  – Counselors in ICS Student Affairs Office
Data Science encompasses the full spectrum of theories and methods that use data to understand and make predictions about the world around us. This includes fundamental research on statistical methods, prediction algorithms, data management techniques, and policy issues; as well as a broad range of domain-specific data-driven research problems in the sciences, engineering, humanities, education, medicine, and business.
ASA DataFest 2017 at Chapman University

ASA DataFest™ 2017 at Chapman University will happen on April 21-23. Registration will begin soon.

What Is ASA DataFest™?
ASA DataFest™ is a data hackathon for undergraduate students, sponsored by the American Statistical Association and founded at UCLA, in 2011. ASA DataFest™ at Chapman is hosted by the Schmid College of Science and Technology and Office of Residential Life.

Analyze ASA DataFest™ will introduce you to what is likely the richest, most complex dataset you’ve seen so far in your undergraduate career. The dataset is provided by a real-life organization and is chosen to provide many avenues of discovery. Students at any stage of their data science education will find something of interest and will have the opportunity to make an original finding. Students from any major are welcome.

Network Mingle with data science professionals who visit DataFest™ to offer their advice and answer your questions. You’ll also get to meet students from other colleges and universities in southern California.

Experience Past participants of the ASA DataFest™ have gone to job interviews able to describe technical challenges overcome, explain how they work under time-pressure, and talk about their thoughts on solving real-life data problems.

ABOUT

ABOUT DATAFEST
expedition: hackathon
Los Angeles, CA

Come change the way we use geospatial information to better understand, model, visualize and monitor the nexus between Food Security and Regional Stability. These efforts are important in understanding humanitarian and military crisis.

Here are some key ideas for the challenge:

1. Validate prediction models for crops
2. Identify Food Logistics Points
3. Build a Crop forecasting model

Recruiters will be on hand from the National Geospatial Intelligence Agency - Bring a Resume!

Social Media Contests for a Leap Motion, 2 TB Portable HD and High Capacity Power Bank!

Scan this code to get your tickets NOW!

**WHEN**
March 25 & March 26, 2017
3/25 - 9:00 AM - Overnight (Saturday)
3/26 - 8:00 AM - 4:00 PM (Sunday)

**WHERE**
CTRL COLLECTIVE
12575 Beatrice St,
Los Angeles, CA 90066

**PRIZES**
Grand Prize: $3000
1st Runner-Up: $1000

Register And Get Your Tickets ASAP! THIS WILL SELLOUT!

TO REGISTER AND JOIN US ON SLACK GO TO: http://expeditionhacks.com/la
The Life Cycle of Data Analysis

1. Data Measurement
   - Acquiring data

2. Data Organization
   - Storing, cleaning, sorting data

3. Data Exploration
   - Exploratory data analysis

4. Data Analysis
   - Statistics, machine learning

5. Interpretation
   - Analyzing, understanding, evaluating

6. Decisions and Actions
   - Taking action
The Life Cycle of Data Analysis

Measure

Organize

Explore

Analyze

Interpret

Decide
The Life Cycle of Data Analysis

Measure

What are different ways we can collect data?

Organize

Explore

Analyze

Interpret

Decide
### 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></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
NASA’s MODIS satellite

entire planet
250m resolution
37 spectral bands
every 2 days
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
Daily Report: At WWDC, Apple Expected to Expand Into Health and Home Monitoring

By THE NEW YORK TIMES  JUNE 2, 2014 7:14 AM  Comment

Apple is unlikely to introduce new devices this week, the things that most excite customers and investors these days. But the company is expected to dive deeper into two new areas: connected health and the so-called smart home, Brian X. Chen reports.

Along with operating system updates for mobile devices and desktop machines, Apple plans to introduce a new health-tracking app at its annual Worldwide Developers’ Conference on Monday, according to a person briefed on the product, who spoke on the condition of anonymity because the plans were confidential. The app for mobile devices will track statistics for health or fitness, like a user’s footsteps, heart rate and sleep activity.
Tracking pumas

From 2001 to 2013, scientists used GPS radio collars to track the pumas' movements in the Santa Ana Mountains and Eastern Peninsular Range in Orange and San Diego counties. Only one puma, M56, crossed between the mountains. Another, M53, moved out of the study area and into Mexico. The rest were hemmed in by highways and housing developments.
Sensors Measuring Human Activity

Optical people counter at a building entrance on campus

Loop sensors on Southern California freeways
Suburban Area
P. Smyth: Stats 5: Data Science Seminar, Winter 2017:

6:00am - 12:00pm - 6:00pm

Vehicle count / 5 min

Industrial Area
Ebird.org

Over 1.5 million submissions per month

From Wood et al, PLOS Biology, 2011
Other ideas on how we can collect data?
The Life Cycle of Data Analysis

Measure

Different ways we can collect data?

What are some of the issues with data collection?

Organize

Explore

Analyze

Interpret

Decide
Geolocated Tweets in Southern California
Geolocated Tweets around UC Irvine
Tweets mentioning Coke (green) and Pepsi (red)

from chimpler.wordpress.com
What might be the issues with using data like this?
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
The Life Cycle of Data Analysis

Measure

Organize Why is data organization important?

Explore

Analyze

Interpret

Decide
### 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>
Computer Architecture 101
How Far Away are the Data?

Random Access Times

- CPU to RAM: $10^{-8}$ seconds
- RAM to Disk: $10^{-3}$ seconds
How Far Away are the Data?

**Effective Distances**

- CPU to RAM: 1 meter
- RAM to Disk: 100 km
Data Engineering at Web Scale
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
Summary Statistics of the Data:

- $N = 11$
- Mean of $X = 9.0$
- Mean of $Y = 7.5$
- Intercept = 3
- Slope = 0.5
What will the mean or median tell us about this data?
Histogram with Outliers

Pima Indians Diabetes Data, From UC Irvine Machine Learning Repository

Number of Individuals

X values
Histogram with Outliers

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Repository

Number of Individuals

blood pressure = 0?
Matrix of Scatter Plots with Color Overlays

Iris classification data set, 3 classes

Figure from www.originlab.com
### Chernoff Faces
#### 2005 National League

[alexreisner.com/baseball/stats/chernoff](alexreisner.com/baseball/stats/chernoff)

<table>
<thead>
<tr>
<th>Team</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>

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)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></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
Example: 4 Data Sets, Y versus X

Guess the Linear Correlation Values for each Data Set

DATA SET 1

DATA SET 2

DATA SET 3

DATA SET 4

Actual Correlation Values

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

How can algorithms extract information from data?
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.
ILSVRC top-5 error on ImageNet
Other general approaches for analyzing data?
The Life Cycle of Data Analysis

Measure → Organize → Explore → Analyze → Interpret → Decide

Human interpretation of results is important
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 the correlation between People who drowned after falling out of a fishing boat and Marriage rate in Kentucky from 1999 to 2010. The graph includes data from CDC & US Census, tylervigen.com.](http://www.tylervigen.com/)

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of people who died by becoming tangled in their bedsheets (Deaths US) (CDC)</th>
<th>Total revenue generated by skiing facilities (US) (Dollars in millions US Census)</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/
Today’s Random Medical News

from the New England Journal of Panic-Inducing Grubbedybook

Can Cause

According to a report released today...

- Hypothermia
- Heart Disease
- Breast Cancer
- Spontaneous Regression
- Glaucoma
- A Feeling of Well-Being
- Multinuclear Tines
- Depression

In

- Children
- Two-Income Families
- Men 25-40
- Overweight Smokers
- Rats

- Twins
- Arthritis Suffers
- 7 Out of 10 Women

- Computer Terminals
- Daycare
- Red Wine
- Fatty Foods
- Exercise
- Stress
- Coffee
- Smoking
Another Example: Automated Essay Grading

From Inside Higher Ed, April 2012

Report on a major study comparing automated essay-grading software with trained human readers, on 22,000 high-school essays.

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

Why is automated essay grading of interest?

Human graders: 20 to 30 essays an hour
Automated: millions per hour
Human Interpretation of 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
Human Interpretation of Automated Essay Grading

From New Statesman and New York Times, April 2012

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.
The Life Cycle of Data Analysis

- Measure
- Organize
- Explore
- Analyze
- Interpret
- Decide: Who owns your data?
Collection of Individual-Level Data

- **Credit Score**
- **Gold Card**
- **Cell Phone**
- **Image of Smart Home**

- **1960’s**
- **1980’s**
- **2000’s**
- **2020’s**
Who Owns Your Data?
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”? 
BACKUP SLIDES
Classifying Handwritten Digits

- inputs: 64 binary pixel values from 8 x 8 image
- target: 10 classes (0, 1, ..., 9)
- neural network classifiers are over 99% accurate in recognizing digits
Examples of Errors from the Neural Network Classifier

Examples of Errors from the Neural Network Classifier

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

Lack of linear correlation does not imply lack of dependence.
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)

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)

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

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

(Electives shown in parentheses)
Figure from Krizhevsky, Sutskever, Hinton, 2012
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?
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