ECOTS 2018
DATA SCIENCE FOR ALL
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 1 – 2 pm    | The future of data science education is plain text (Keynote)  
Jeff Leek (Johns Hopkins University) |
| 2:15 – 2:45 pm | Undergraduate Data Science Pathways: What is Needed for Entry and Success?  
Nick Horton, Roxy Peck, Rebecca Hartzler (Amherst College) |
| 3 – 3:30pm  | Data Science Education across the US and the National Academies  
Eric Kolacyk (Boston University) |
| 3:45 – 4:15 pm | Introducing Data Science Elements through Parallel Courses in Statistics and Computing  
Eric Reyes & Megan Heyman (Rose-Hulman Institute of Technology) |
| 4:30 – 5 pm | Integrating Programming into Statistics Curricula  
Jonathan Duggins & Justin Post (North Carolina State University) |
| 5:30 – 6:30 pm | Posters and Beyond 1B  
Various Speakers |
| 6:30 – 7:15 pm | Dinner |
| 7:15 – 8 pm  | Incorporating Data Science with a Fixed Curriculum  
Michael Posner (Villanova University) |
Incorporating Data Science with a Fixed Curriculum

eCOTS Regional Conference – Villanova University
May 21, 2018

Michael A. Posner, PhD, PStat®, Villanova University
http://homepage.villanova.edu/michael.posner
Overview

• Data Science – Definition & The Craze!
• Topics – Past and Present
• Topics – Present…Re-structured
Overview

• Data Science – Definition & The Craze!
• Topics – Past and Present
• Topics – Present…Re-structured
Statistics & Data Science

• The “sexy” job of the 21st Century
  - Google’s Chief Economist, Hal Varian (2009): “I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?”

• …with a growing and unmet job demand
  - McKinsey Report on Big Data (2011): “The US alone faces a shortage of 140,000 to 190,000 people with deep analytic skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings”
  - Information Week (2013): “Data Scientist: The Sexiest Job No One Has”

• Additional links
  - Six analytics and data science jobs are included in Glassdoor’s 50 best jobs In America for 2018
  - Statistician Projected As Top 10 Fastest-Growing Job for 2014-2024, New Projections for 2016-2026
Data Science

• What is data science?
  – ?
  – ?
What is Data Science?

- From Modern Data Science with R (Baumer, Kaplan, Horton):
  - The science of extracting meaningful information from data
- Michael Jordan:
  - Computer Science is more than just programming; it is the creation of appropriate abstractions to express computational structures and the development of algorithms that operate on those abstractions.
  - Statistics is more than just collections of estimators and tests; it is the interplay of general notions of sampling, models, distributions, and decision making.
  - Data Science is based on the idea that these styles of thinking support each other.
- Unknown:
  - Turning data into knowledge and knowledge into insight.
The Venn Diagram
Yet Another Venn Diagram

Data Science Venn Diagram v2.0

- Computer Science
- Machine Learning
- Math and Statistics
- Traditional Software
- Traditional Research
- Subject Matter Expertise
- Unicorn

Copyright © 2014 by Steven Geringer Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact.
Why Such Growth?

- Data are readily available
  - From data entry...to extraction...to sensors
  - In 2011, the world had created a total of 1.8 zettabytes of data
  - In 2012, that rose to 2.8 zb
  - In 2020, estimated 40 zb

1 zb = 10^21 bytes = 1 trillion gigabytes
From IBM -http://www.ibmbigdatahub.com/infographic/four-vs-big-data
Your Data are Valuable!
Name That Company…

• Largest transportation company in the world
  – How many vehicles do they own?

• Largest room rental company in the world
  – How many properties do they own?

Target Figures Out Teen Girl Is Pregnant Before Her Father Does, Sends Helpful Coupons

By Mary Beth Quirk  February 17, 2012

We didn’t really believe it when we were told in 7th grade that math could unlock the secrets of the universe, but after reading about the coupon-wielding power of a Target statistician, which resulted in a mighty surprise for one father of a teenage girl, we might be converts. Doesn’t make math any better though.

The New York Times (via Forbes) had some time chatting with Target’s statistician royale, Andrew, before he was told to zip his lips by the company. He discussed how retailers figure out how to sort out your purchases — from what you need, what you will use a coupon for and your personal preferences. Oh yeah, and they can decode if you’re pregnant even before you buy diapers.

In Target’s case, it all comes down to your Guest ID number tied your credit card, name, and other info, which saves all kinds of data about what you buy. Statistician Andrew mined that data and saw patterns in it, for example — women on baby registries buy larger amounts of unscented lotion around the beginning of their second trimester. Bam! Send ’em some coupons for other baby items. More Andrew magic!

As his computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Freaky! Those scores lead Target to send coupons for baby items to their customers with certain
Looks like Nate Silver’s predictions were 100% accurate (50 for 50 states) as of polls right now (11:58pm):

Election Forecasts – FiveThirtyEight Blog – NYTimes.com (nytimes.com)
Sports Statisticians
## Data Competitions

Based on your rating, we think you'll enjoy these titles

### Active Competitions

<table>
<thead>
<tr>
<th>Competition</th>
<th>Remaining Days</th>
<th>Teams</th>
<th>Prize</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second Annual Data Science Bowl</strong></td>
<td>36</td>
<td>347</td>
<td>$200,000</td>
</tr>
<tr>
<td>Transforming How We Diagnose Heart Disease</td>
<td>36</td>
<td>347</td>
<td>$200,000</td>
</tr>
<tr>
<td><strong>Cervical Cancer Screening</strong></td>
<td>14</td>
<td>30</td>
<td>$100,000</td>
</tr>
<tr>
<td>Help prevent cervical cancer by identifying at-risk populations</td>
<td>14</td>
<td>30</td>
<td>$100,000</td>
</tr>
<tr>
<td><strong>The Allen AI Science Challenge</strong></td>
<td>26</td>
<td>668</td>
<td>$80,000</td>
</tr>
<tr>
<td>Is your model smarter than an 8th grader?</td>
<td>26</td>
<td>668</td>
<td>$80,000</td>
</tr>
<tr>
<td><strong>The Winton Stock Market Challenge</strong></td>
<td>8.3</td>
<td>672</td>
<td>$50,000</td>
</tr>
<tr>
<td>Join a multi-disciplinary team of research scientists</td>
<td>8.3</td>
<td>672</td>
<td>$50,000</td>
</tr>
<tr>
<td><strong>Home Depot Product Search Relevance</strong></td>
<td>3 months</td>
<td>4</td>
<td>$40,000</td>
</tr>
<tr>
<td>Predict the relevance of search results on homedepot.com</td>
<td>3 months</td>
<td>4</td>
<td>$40,000</td>
</tr>
<tr>
<td><strong>Prudential Life Insurance Assessment</strong></td>
<td>28 days</td>
<td>1902</td>
<td>1121 scripts</td>
</tr>
<tr>
<td>Can you make buying life insurance easier?</td>
<td>28 days</td>
<td>1902</td>
<td>1121 scripts</td>
</tr>
</tbody>
</table>

*In 2009, after three years and submissions by more than 40,000 teams from 180 countries, Netflix awarded the $1 million Netflix Prize to a team of engineers, statisticians and mathematicians from the BandLab in India.*
Save the Date
(Tentative)
April 12-14, 2019
Data Science Foible Example
Google Flu

- http://www.nature.com/news/when-google-got-flu-wrong-1.12413
Overview

• Data Science – Definition & The Craze!
• Topics – Past and Present
• Topics – Present…Re-structured
<table>
<thead>
<tr>
<th>Topic in Traditional Statistics Course</th>
<th>Triola</th>
<th>AP Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro to Statistics</td>
<td>1</td>
<td>Exploring Data</td>
</tr>
<tr>
<td>Summarizing and Graphing</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Statistics for Describing, Exploring, Comparing</td>
<td>3</td>
<td>Sampling &amp; Experiments</td>
</tr>
<tr>
<td>Sampling and Experimentation</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td>4</td>
<td>Anticipating Patterns</td>
</tr>
<tr>
<td>Discrete Probability Distributions</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Normal Probably Distributions</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Estimates and Sample Size (one sample)</td>
<td>7</td>
<td>Statistical Inference</td>
</tr>
<tr>
<td>Hypothesis Testing (one sample)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Inference for Two Samples</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Correlation and Regression</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Goodness of Fit and Contingency Tables</td>
<td>(11)</td>
<td></td>
</tr>
<tr>
<td>Analysis of Variance</td>
<td>(12)</td>
<td>no</td>
</tr>
<tr>
<td>Nonparametrics Tests</td>
<td>(13)</td>
<td>no</td>
</tr>
<tr>
<td>Statistics Process Control</td>
<td>(14)</td>
<td>no</td>
</tr>
<tr>
<td>Ethics, Projects, Procedures, Perspectives</td>
<td>(15)</td>
<td>?</td>
</tr>
</tbody>
</table>
Overview

• Data Science – Definition & The Craze!
• Topics – Past and Present
• Topics – Present…Re-structured
Topic Reduction for Data Science

- Less probability, less probability distributions
- Simulation-based inference
  - Computational thinking
  - Pedagogical advantages
- Start broad, develop deep (program-level)
  - “Understanding by Design” (Wiggins & McTighe)
  - BA, MS, PhD in Statistics
  - Ballroom Dance!
  - Even Tiger Woods…
Topic Foci for Data Science I

- Data Visualization (know what you have)
  - additional aesthetics, beyond traditional
- Bias (know what you don’t have)
  - Where do my data come from?
  - Missing/Messy Data – choices about processing
- “Disaggregation” / Confounding
  - Correlation ≠ Causation
- Simulation-based inference (Hesterberg)
  - Bootstrap
- Sampling – example about subsetting data
Topic Foci for Data Science II

• Data Ethics
• Model Scientific Process
  – Spend time to talk about data (fewer datasets?)
  – Discuss limitations and alternate approaches
• Communication, collaboration
• “Build up” your analysis
  – Univariate / Bivariate comparisons
  – Not “one and done” or only one right approach
• Data manipulation is okay, even if not perfect
Multivariable Thinking / Thinking “Slow”

- We have an inherent desire to assume causal relationships
  - ...even when there is evidence or knowledge to the contrary
- Examples
  - Magic: Teller’s article: “magic goes straight to the brain; its essence is intellectual.”
  - Brain Games (National Geographic series focusing on cognitive science behind illusions, psychological experiments, and counterintuitive thinking)
- Thinking Fast and Slow by Kahneman – 2012 Nobel Prize in economics
  - Jessica Utts 2016 ASA Presidential Address (28:04 -> 42:51 or 46:15)
  - Causal thinking in statistics and using “Data With Stories”
  - Targeted for statistical audience, but very accessible
Example of Disaggregation

Boxplot of Mortgage Rejection Rates by Race

Mortgage Rejection Rates

Minority

White
Example of Disaggregation

Boxplot of Mortgage Rejection Rates by Race/ SES

Mortgage Rejection Rates

Minority | White | HighMinority | HighWhite
A Case of Discrimination in California

- **Discrimination in Services to the Disabled (JSE)**
- **Data Set:** Random sample of 1,000 clients of California Department of Developmental Services
- **Variables:**
  - Annual expenditure for support to individual and family
  - Gender
  - Ethnicity
  - Age Cohort (based on amount of financial support typically required during a particular life phase) (0 – 5 years, 6 – 12 years, 13 – 17 years, 18 – 21 years, 22 – 50 years, over 50 years)
- **Is there discrimination by gender or by ethnicity?**
## Average Expenditure by Gender/Ethnicity

<table>
<thead>
<tr>
<th>Ethnicity of Consumers</th>
<th>Average of Expenditures ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian</td>
<td>$ 36,438</td>
</tr>
<tr>
<td>Asian</td>
<td>$ 18,392</td>
</tr>
<tr>
<td>Black</td>
<td>$ 20,885</td>
</tr>
<tr>
<td>Hispanic</td>
<td>$ 11,066</td>
</tr>
<tr>
<td>Multi Race</td>
<td>$ 4,457</td>
</tr>
<tr>
<td>Native Hawaiian</td>
<td>$ 42,782</td>
</tr>
<tr>
<td>Other</td>
<td>$ 3,317</td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>$ 24,698</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$ 18,066</td>
</tr>
</tbody>
</table>

### Table 1. Average Expenditures by Ethnicity

<table>
<thead>
<tr>
<th>Gender</th>
<th>Average of Expenditures ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>$ 18,130</td>
</tr>
<tr>
<td>Male</td>
<td>$ 18,001</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$ 18,066</td>
</tr>
</tbody>
</table>

### Table 2. Average Expenditures by Gender
Limiting the Analysis…

- Sample sizes were very small for most groups, so focus on comparing two groups—Hispanic and White non-Hispanic.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Average of Expenditures ($)</th>
<th>% of Consumers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>$11,066</td>
<td>38%</td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>$24,698</td>
<td>40%</td>
</tr>
</tbody>
</table>

Table 5.1 Average Expenditures and # of Consumers by Ethnicity

Is there evidence of discrimination?
What else might explain difference in mean expenditure?
### (Joint) Expenditure by Ethnicity & Age

<table>
<thead>
<tr>
<th>Age Cohort</th>
<th>Hispanic (avg. of expenditures)</th>
<th>White non-Hispanic (avg. of expenditures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 5</td>
<td>$1,393</td>
<td>$1,367</td>
</tr>
<tr>
<td>6-12</td>
<td>$2,312</td>
<td>$2,052</td>
</tr>
<tr>
<td>13-17</td>
<td>$3,955</td>
<td>$3,904</td>
</tr>
<tr>
<td>18-21</td>
<td>$9,960</td>
<td>$10,133</td>
</tr>
<tr>
<td>22-50</td>
<td>$40,924</td>
<td>$40,188</td>
</tr>
<tr>
<td>51 +</td>
<td>$55,585</td>
<td>$52,670</td>
</tr>
<tr>
<td>All Consumers</td>
<td>$11,066</td>
<td>$24,698</td>
</tr>
</tbody>
</table>

Table 6. Average Expenditures by Ethnicity and Age Cohort
Halderman: The presidential election was 'probably not' hacked – but the votes should be checked

Effect of paper ballots in Wisconsin goes from 7 pts, like NY article, to 0 if you control for race education, density (true w&w/o weights,)

```
formula = d.shift ~ paper + colplus + black_pct + data
  data = wi, weights = total_reported

R-squared: 0.3603, Adjusted R-squared: 0.3314

Estimate Std. Error t value Pr(>|t|)
intercept) -0.52344  0.03691 -14.180  < 2e-16 ***
  perpaper -0.04009  0.02785  -1.439  0.154678
  plus -0.57272  0.15205  -3.767  0.000351 ***
  black_pct -1.82717  0.23883  -7.651  1.03e-10 ***
sity.rescale  0.48369  0.09689   4.992  4.51e-06 ***

Residuals:    Min  1Q  Median   3Q  Max
  -9.440 -2.090   3.381   93.054

```

Individual standard error: 17.31 on 67 degrees of freedom
Multiple R-squared: 0.5336, Adjusted R-squared: 0.5271
test statistic: 19.16 on 4 and 67 DF, p-value: 1.514e-10
Plot of SAT Score by Teacher Salary
Plot of SAT Score by Teacher Salary
Sat Score & Teacher Salary

Plot of SAT Score by Teacher Salary

State Avg SAT Score vs Average Teacher Salary
Sat Score & Teacher Salary

Plot of SAT Score by Teacher Salary

State Avg SAT Score vs Average Teacher Salary
Sat Score & Teacher Salary

Plot of SAT Score by Teacher Salary by Fraction

State Avg SAT Score

Average Teacher Salary
PA School Funding - Aesthetics

District Per-Student BEF vs. Poverty. Yellow: >92% White; Brown: < 92% White

From the Notebook: http://thenotebook.org/blog/147895/pa-state-education-funding-has-racially-discriminatory-impact
Confounder Activity

For each of the following, identify the confounder:

• Shoe size and vocabulary among children?
• As crime rates increase, so does ice cream sales, so let’s stop selling ice cream!
• The number of speeding tickets that a student has is less than that of his/her parents.
• Typically, students do worse on exam problems that they spend a lot of time on.
• Strong positive relationship between # of TVs per house and life expectancy (by country)
• The mortality rate in the US is higher than most other North/South American countries
• The more children a woman has, the lower her risk of breast cancer.
• Do storks really bring babies?
• Bonjour Paris L’école (https://www.youtube.com/watch?v=tS55WeYwPfA)
Activity on Bias - Population Estimation

• Do you think the population of the Philippines is above 20 million?
• Do you think the population of the Philippines is above 90 million?
• What is your best guess for the population of the Philippines?
• Record and review these
• This is an example of bias!
Exploring Bias in Leading Questions for Philippines Population Size

- True Mean = 104

Group (Suggestion):
- A (20)
- B (90)
Imagine the Data Pipeline...
The Real Data Pipeline

From B Fiore-Gartland
Ethics in Data Science

4R approach

– **Reuse** – using data again for a similar purpose
  • Blood samples
  • Bank of America – sold data for statistical analysis but not contact/action
  • Twitter, Facebook, etc. (Twitter terms for fair use)

– **Repurpose** – use data for unrelated purpose
  • Education data for research

– **Recombine** – merging data with other sources
  • Russian use of Facebook to meddle in the 2016 US election?

– **Reanalyze** – secondary data analysis
  • NAEP, NHANES, etc.
  • Genetic data – rights to the data (can kids know about parents)?
A Case of Recombining Data

• Latanya Sweeney’s de-identification of public data
  – The case of Governor Weld
  – Sweeney says 87% of US people can be identified by their zip code, birthdate, and gender – citation
  – Golle says 63% - citation
    • Gender, zip code & year of birth = 0.2%
    • Gender, zip code & year and month of birth = 4.2%
    • Gender, zip code & year, date, and month of birth = 63.3%

• She’s not done (2013 report)…

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study
A Great Resource

http://callingbullshit.org/syllabus.html
Implicit assumption and expectation that big data has the quality of seeming or being felt to be true, even if not necessarily true.

(Adapted from Stephen Colbert’s “Truthiness”)
Messy Data – Example

How many hours of exercise do you get in a typical week? (What answers do you expect student to give?)
Why I Love Teaching Data Science!

• Allows interaction with new data structures
  – Volume
  – Variety – structure
  – Variety – web scraping
  – Variety – text mining

• Enables me to explore visualizations
  – Motivate/understand multiple regression

• Excites students about career potential
Scatterplot of Price and Proportion Black
Further Investigation!

Scatter plot of Total Price and Proportion black by Chain Outlet

factor(chain)
1 BK
2 KFC
3 RR
4 Wendy's
TYCDSS – Minor/Certif

1) Intro to Data Science (book “The Art of Data Science” Michael Stadzel?)
   a) Focus on range of possible decisions to be made with data
   b) Some programming, but no prerequisites
   c) Data life cycle, cleaning, ethics - embedded and role-modeled throughout
   d) Descriptive Statistics
   e) Know what you know, what you don’t know, and when to consult others

2) Intro to Statistical Reasoning
   a) Time spent on each topic different than stereotypical / traditional intro stat course
   b) Focus on bias, multivariable thinking, philosophy of inference
   c) Simulation-based inference is preferred

3) Intro to Programming
   a) “Applied data programming” - Applications of statistical software for data management and reporting (data processing, linear and logistic regression)
   b) Database management?

4) Communicating with Data (including Visualization)
   1) Discussion about GUI-interface (Tableau, Excel, etc.) - Is it enough? Can be done early?

5) Capstone experience
   1) Application-area based
   2) Revisit intro data science topics
TYCDSS - Skills

1) Summarize, critique, and communicate findings of complex analyses in a concise way for a target audience using graphical and descriptive measures.

2) Understand the data life cycle including obtaining, storing, managing, cleaning, and analyzing. Recognize the implications to the problem at hand.

3) Capable of reviewing common modelling techniques and the quality of analysis (significance/fit) in the context of making decisions.

4) Weave ethics into the curricula at relevant points and considering consequences. Balance privacy concerns with creating social good.

5) Computational skills
   1) Programming – understand at least one programming language with enough depth to apply statistical methods and visualization techniques
   2) Data management – apply data management techniques to real world data sets
   3) Application – use a package or program to analyze and visualize data using appropriate models in a specific domain

6) Develop statistical reasoning which includes data collection methods, sampling, variability, multivariable thinking, bias, observational versus experimental data, modern descriptive methods, basic inference and an awareness of model building with cross-validation.
FOLLOW THE MONEY

GOAL:
To find trends in candidate spending, look at where spending during campaigning occurred most, and find interesting disbursements unique to this dataset
Processing the Data

1. Read in the data and clean
   - Removing “error” dates
   - Looking at negative disbursements

2. Univariate plots of variables

3. Creation of graphics
   - Text processing
   - Spatial analysis
   - Ggplot
   - Data wrangling
   - Tables
   - Bar charts
   - Line graphs
   - US maps
   - Wordclouds
Clinton had the most disbursements and spent the most total amount.

Trump had the highest mean disbursement amount per transaction.

Sanders fell surprisingly close to Trump in both disbursement frequency and total amount spent.
Trump Disbursement Overview

Trump’s Frequently Used Words in Disbursement Descriptions

Trump Spending Trend:
Large expenses with low frequency of use
Clinton Disbursement Overview

- Clinton’s top words were travel, payroll, and phone.
- Media Buy was the most expensive disbursement.
- Payroll and Travel were high expenses and frequent disbursements.

Clinton’s Frequently Used Words in Disbursement Descriptions

- Event, services, catering, equipment, technology, software, strategic.
- Advertising, email, production, credit, tax.
- Venue, beverages, fees, rentals, subsistence.

Clinton’s Spending Trend:
Many smaller expenses totaling to a large overall cost.

frequent disbursements.
General Election: US Maps

- Redder states: Trump outspent Clinton
- Bluer states: Clinton outspent Trump
- Can easily see relative difference in spending between candidates
- Binary Map: Red is Trump and Blue is Clinton
- All states are now clearly colored

The orange circles indicate large areas of the country that did not reflect the same as spending.
Orange stars are “swing states” that were reverse of the spending map.
GA, NC, IA, WI, OH, AZ, and CO did follow who spent more.

https://www.27towin.com/maps/2016-election-state-winners
Are Candidates Paying Each Other or Themselves?

- Trump disbursed a concerning amount to himself and family members (Eric, Laura, Donald Trump Jr.)
- Clinton paid Trump only $260 for Trump Hotel
- Clinton only disbursed to herself in expenses to the Clinton Executive Services Corp. and Tyler Cassidy Clinton.

### To Clinton Disbursements

<table>
<thead>
<tr>
<th>cand_nm</th>
<th>recipient_nm</th>
<th>disb/Tot</th>
<th>numDisb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>CLINTON EXECUTIVE SERVICES CORPORATION</td>
<td>882482.74</td>
<td>26</td>
</tr>
<tr>
<td>Sanders, Bernard</td>
<td>PETSON, CLINTON J</td>
<td>4099.94</td>
<td>6</td>
</tr>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>CLINTON IOWA 4, LLC</td>
<td>4250.00</td>
<td>6</td>
</tr>
<tr>
<td>Clinton, Hilary Rodham</td>
<td>CLINTON SCHOOL DISTRICT</td>
<td>2365.50</td>
<td>2</td>
</tr>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>CLINTON, TYLER CASSIDY</td>
<td>2193.09</td>
<td>7</td>
</tr>
<tr>
<td>Webb, James Henry Jr.</td>
<td>HAMPTON INN - CLINTON IA</td>
<td>1789.93</td>
<td>6</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>CLINTON COMMUNITY SCHOOL DISTRICT</td>
<td>1076.85</td>
<td>3</td>
</tr>
<tr>
<td>Sanders, Bernard</td>
<td>CLINTON, HILLARY ROODHAM</td>
<td>755.00</td>
<td>1</td>
</tr>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>MCCCLINTON, SANDRA</td>
<td>529.65</td>
<td>1</td>
</tr>
<tr>
<td>Bush, Jeb</td>
<td>CLINTON SCHOOL DISTRICT</td>
<td>438.75</td>
<td>1</td>
</tr>
<tr>
<td>Sanders, Bernard</td>
<td>CLINTON MASONIC CENTER ASSOCIATION</td>
<td>200.00</td>
<td>1</td>
</tr>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>CLINTON DENTAL ASSOCIATION</td>
<td>96.81</td>
<td>1</td>
</tr>
<tr>
<td>Sanders, Bernard</td>
<td>CLINTON, HILARY ROODHAM</td>
<td>79.66</td>
<td>2</td>
</tr>
<tr>
<td>Sanders, Bernard</td>
<td>CLINTON MASONIC CENTER ASSOCIATION</td>
<td>25.81</td>
<td>1</td>
</tr>
</tbody>
</table>

### To Trump Disbursements

<table>
<thead>
<tr>
<th>cand_nm</th>
<th>recipient_nm</th>
<th>disb/Tot</th>
<th>numDisb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP TOWER COMMERCIAL LLC</td>
<td>301725.14</td>
<td>91</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL LAS VEGAS</td>
<td>263397.60</td>
<td>60</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL - LAS VEGAS</td>
<td>218346.72</td>
<td>37</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP PAYROLL CORP</td>
<td>302035.85</td>
<td>37</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP PLAZA LLC</td>
<td>120000.00</td>
<td>17</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>THE TRUMP CORPORATION</td>
<td>6003.87</td>
<td>3</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL GOLF CLUB</td>
<td>4327.86</td>
<td>5</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL GOLF CLUB - LAS VEGAS</td>
<td>77915.16</td>
<td>6</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL</td>
<td>720000.00</td>
<td>12</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL - LAS VEGAS</td>
<td>45062.35</td>
<td>13</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL GOLF CLUB - WESTCHESTER</td>
<td>13927.70</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL Doral</td>
<td>47985.76</td>
<td>6</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL GOLF CLUB - DRESDEN</td>
<td>45045.34</td>
<td>6</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL</td>
<td>29745.42</td>
<td>7</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, ERIC</td>
<td>22416.02</td>
<td>7</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL - WASHINGTON</td>
<td>21185.32</td>
<td>7</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL GOLF CLUB - WASHINGTON</td>
<td>2104.00</td>
<td>2</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP NATIONAL HOTEL &amp; TOWER CHICAGO</td>
<td>18242.36</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP DONALD JR.</td>
<td>10389.11</td>
<td>13</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, HOTEL</td>
<td>11380.11</td>
<td>13</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>10470.78</td>
<td>2</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP SOHO</td>
<td>1271.70</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL GOLF CLUB - WESTCHESTER</td>
<td>12001.17</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL GOLF CLUB - CHARLOTTE</td>
<td>4948.90</td>
<td>10</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP INTERNATIONAL HOTEL &amp; TOWER CHICAGO</td>
<td>8593.85</td>
<td>3</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP OLD POST OFFICE LLC</td>
<td>5000.00</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>3946.66</td>
<td>6</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>3306.78</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>3314.23</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>2368.69</td>
<td>8</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>7033.45</td>
<td>11</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>9582.70</td>
<td>12</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>1325.94</td>
<td>1</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>TRUMP, DONALD J.</td>
<td>30645.77</td>
<td>37</td>
</tr>
</tbody>
</table>

### Clinton Executive Services Corp. and Tyler Cassidy Clinton

<table>
<thead>
<tr>
<th>cand_nm</th>
<th>totalDisbursements</th>
<th>totTrump</th>
<th>percent/Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>579714658</td>
<td>260</td>
<td>0.00000448</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>361184316</td>
<td>3064577</td>
<td>0.9979882</td>
</tr>
</tbody>
</table>

### Table: Clinton Executive Services Corp. and Tyler Cassidy Clinton

<table>
<thead>
<tr>
<th>cand_nm</th>
<th>totalDisbursements</th>
<th>totTrump</th>
<th>percent/Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton, Hillary Rodham</td>
<td>579714658</td>
<td>260</td>
<td>0.00000448</td>
</tr>
<tr>
<td>Trump, Donald J.</td>
<td>361184316</td>
<td>3064577</td>
<td>0.9979882</td>
</tr>
</tbody>
</table>
Weather Wrangling Wizards
Processing the Data

- Condense the data to only include information we needed to the analysis
  - 22 columns related to dates, times, locations, storm types, deaths/injuries, damage, etc.
  - Break down “YEARMONTH” variable into “year” and “month

- Further condense the event type (type of storm) to include similar types of storms in 6 major categories:

<table>
<thead>
<tr>
<th>TYPE</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter Storm</td>
<td>494818</td>
</tr>
<tr>
<td>Thunderstorm</td>
<td>468690</td>
</tr>
<tr>
<td>Flood</td>
<td>122201</td>
</tr>
<tr>
<td>Wind</td>
<td>80439</td>
</tr>
<tr>
<td>Tornado</td>
<td>67424</td>
</tr>
<tr>
<td>Hurricane</td>
<td>1811</td>
</tr>
</tbody>
</table>

6 rows
Deaths by Storm

Grouped by type and summarised the Direct and Indirect deaths with a sum. Gathered the data so that it could be split by Direct or Indirect.

Grouped by type and summarised the Direct and Indirect deaths with a mean. Gathered the data so that it could be split by Direct or Indirect.
Injuries by Storm

Grouped by type and summarised the Direct and Indirect injuries with a **sum**. Gathered the data so that it could be split by Direct or Indirect.

Grouped by type and summarised the Direct and Indirect injuries with a **mean**. Gathered the data so that it could be split by Direct or Indirect.
Average Casualties per State
Tornado Locations

- Average latitude and longitude of each episode ID
- High density in the midwest, a few in California
Challenges/Surprises

- Condensing the data and making sure we were using data points that made sense for our findings
- Condensing the event types into smaller categories
  - Determining which types went with which major category
- Computing the difference in the time because time was a numeric variable. It needed to be separated into hours and minutes and then converted into all minutes.
- Changing from character to numeric for certain fields (like Damage to Property)
- Using summarise function
The Path to Success in the MLB: Pitching or Hitting?
Batting Average vs OBP

● Until recently teams thought batting average was more important than OBP and tended to ignore OBP
● However; OBP has a much stronger correlation to WAR than AVG
WAR Compared to Pitching Statistics

WAR Compared to Pitching Statistics

- **colour**
  - Red: Home Runs Allowed per 9 Innings
  - Green: Strikeouts per 9 Innings
  - Blue: Walks Allowed per 9 Innings
History of the MLB, told by WAR

- 1920 - End of the “Dead Ball Era”
- 1947 - End of Segregation
- 1961 - Westward Expansion
- 1969 - Mound lowered
- 1973 - Introduction of the DH
- 1981 - Strike
- 1990 - Beginning of Steroid Era
- 1994 - Strike
- 2005 - End of Steroid Era
WAR v. Wins, AL
How can these findings be used?

- Depending on a team’s current composition, this data could help make more informed decisions on the marginal impact of getting new hitting or pitching.
- Finding players who favor one statistic, i.e. K/9, and seeing how their “specialty” affects potential WAR and overall success.
- Players would gravitate towards the AL due to higher salaries on average.
- Looking at trends in hitting statistics and finding pitching that “counters” them - “groundball pitchers”.
- Historically, it seems that teams have more overall success from higher hitting WAR.
Movie Study

Jonah Hill From Moneyball
Text Mining

- Used the positive negative word list and did a sentiment analysis on the overview of the movies
- Compared positive, neutral, and negative analyses with a log(revenue)
Sentiment analysis

Density of average rating based on sentiment

Density of log(revenue) based on sentiment
Wordcloud of the most common words used in the overview of movies

- More or less expected
Film revenue vs Number of men in top 5 credits - Notable outliers
Budget and Revenue Adjusted

CPI Index Adjusted Values by Release Year of Top Grossing Movies

Release Year

Log(Value/Normal)


variable

- budget
- revenue
Mean and Median

-Potential issues include that we have less data for older movies
-More new "indy" movies
Billion Dollar Statistics
## Correlations to W-L Record (All teams 1992-2016)

<table>
<thead>
<tr>
<th>Stat</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Shooting</td>
<td>0.64</td>
</tr>
<tr>
<td>Effective Field Goal %</td>
<td>0.60</td>
</tr>
<tr>
<td>Total Rebound %</td>
<td>0.54</td>
</tr>
<tr>
<td>Turnover %</td>
<td>-0.40</td>
</tr>
<tr>
<td>Free Throws Per FG Attempt</td>
<td>0.33</td>
</tr>
<tr>
<td>Offensive Rebound %</td>
<td>0.29</td>
</tr>
<tr>
<td>Block %</td>
<td>0.29</td>
</tr>
<tr>
<td>Free Throw Rate</td>
<td>0.25</td>
</tr>
<tr>
<td>Assist %</td>
<td>0.23</td>
</tr>
<tr>
<td>Steal %</td>
<td>0.22</td>
</tr>
<tr>
<td>3PT Attempt Rate</td>
<td>0.02</td>
</tr>
</tbody>
</table>

- **Strong Correlation**: True Shooting, Effective Field Goal %, Total Rebound %
- **Moderate Correlation**: Turnover %, Free Throws Per FG Attempt
- **Weak Correlation**: Offensive Rebound %, Block %, Free Throw Rate, Assist %, Steal %, 3PT Attempt Rate
Variable Importance Plot for a Random Forest of the Advanced Statistics

```
data4$`TS%`
data4$`TRB%`
data4$`eFG%`
data4$`TOV%`
data4$`BLK%`
data4$`FT/FGA`
data4$`AST%`
FTr
data4$`3PA%`
```

```
data4$`IncNodePurity`
Win-Loss Percentage vs True Shooting Percentage

\[ TS\% = \frac{PTS \times 100}{2 \times (FGA + (0.44 \times FTA))} \]
The US Government: Can A House Divided Stand?
More bills passed in House and Senate on average with Democratic party majorities.

Generally negative trend over time.
Bills Passed vs. Years

Take Away: As time has passed, irrespective of presidential party, bills passed has declined.

*Color denotes

Senate
Mean(Dem): 1108.4
Mean(Rep): 1181.4

House
Mean(Dem): 1189.31
Mean(Rep): 1174.72

*Color denotes
Presidential Vetoes vs. Years (Senate)

*Bills per Veto minimizes the time bias which exists because of the downward trend of bills.

A value of 90.8 for example means that Democratic Senates have passed, on average 90.8 bills per veto.

- Democratic Presidents have vetoed, on average less bills from Democratic Senates.
- Republican Presidents have vetoed, on average, less bills from Democratic Senates than Republican.
GDP %CHANGE - Exec

Mean(Dem): 3.63%
Mean(Rep): 2.80%
GDP %CHANGE - 4 Year Lag Exec

Mean (Dem): 3.70%
Mean (Rep): 2.84%