Data Analytics
CS301
Introduction to Data Analytics

Week 1: 1st Sept
Fall 2020
Oliver BONHAM-CARTER
Have you ever wondered about the secrets in your data?
Links To Our Class

- Course web site: https://www.cs.allegheny.edu/sites/obonhamcarter/cs301.html
  - Syllabus
  - “Planning-Your-Time”, class schedule
- Course calendar
  - https://calendar.google.com/calendar/b/1?cid=Y184bXN0dDg2cW5oaWNjb3NxBWdibHNlNzFva0Bncm91cC5jYWxlbmRhci5nb29nbGUuY29t
- Zoom meetings for class and lab
  - https://allegheny.zoom.us/j/95834628670
  - Also see calendar for Zoom link
Flow in Our Class

Tuesday class
- Tuesday group: In-person
- Thursday group: Online

Thursday class
- Tuesday group: Online
- Thursday group: In-person

Friday Lab
- Tuesday group: Online
- Thursday group: Online
Two Class Groups

- Your group’s day determines the weekday of class when are physically present.
- Tuesday group: Physically in class on Tuesdays
- Thursday group: Physcially in class on Thursdays
- When you are not in class, it is expected that you will be coming to class via Zoom, or watching the recorded class videos.
Computers and Information
Computers and Information

• In this class, you will learn how to use machines to understand *trends* in data.

• (Making decisions by data)
Analytics in Action

• The Jeopardy Challenge of February 2011
• IBM’s Watson beat the show’s greatest champions: Ken Jennings and Brad Rutter.
Machines, Data and Information
Is Watson magic??

http://watson2016.com/
(The Electronic Frontier Foundation)
Surrounded by DATA!

- We live in the “Information age”
- Actually, we live in the “Data age” since there is more data available than information
- Data != Information
Surrounded by DATA!

• It is cheap (and free or even lucrative) for businesses to collect data concerning:
  − in e-commerce,
  − customer behaviors,
  − purchase interests,
  − *health and medical data.*
We Voluntarily Give Away Our Data
Our Phones Create Data

- Smart phones constantly monitor us and keep data.
- Q: How does the iPhone decide whether we are actually getting enough sleep?
- Who keeps the data?
And So, Data is Increasing

- **$65 billion**: Location-tagged payments made in the U.S. annually
- **154 billion**: E-mails sent per day
- **87%**: U.S. adults whose location is known via their mobile phone

Digital Information Created Each Year, Globally

- **2,000%**: Expected increase in global data by 2020

III Megabytes

- **Video and photos stored by Facebook, per user**
- **75%**: Percentage of all digital data created by consumers

Sources: IDC, Radicati Group, Facebook, TR research, Pew Internet
Data, Data, Data, Data!

- How much data is there?

If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the moon. By 2020, there would be 6.6 stacks from the Earth to the Moon.
Data of User Ages

Age distribution by social networking site platform

% of social networking site users on each site who are in each age group. For instance, 29% of MySpace users are 18-22 years old.

![Age distribution chart](http://www.vincos.it/wp-content/uploads/2011/06/PEW_sns_breakdown_age.jpg)

Source: Pew Research Center’s Internet & American Life Social Network Site survey conducted on landline and cell phone between October 20-November 28, 2010. N for full sample is 2,255 and margin of error is +/- 2.3 percentage points. N for social network site and Twitter users is 975 and margin of error is +/- 3.5 percentage points.

By the way: These last slides visually describe trends ...

• Graphics have informed us:
  - Which apps are popular
  - Number of people in age groups for social networking sites
  - How much data is created each year, in relation to other years
  - Twitter “fast-facts”
  - Monthly users of services
  - Increases in Linked-In membership

• How did we learn this information to make these previous visualizations?

Seriously, where did this information come from???
From Raw DATA!!

- Algorithms processed seemingly unconnected data to filter out unimportant material.
How Do We Know?

- The previous graphs came to us via raw Big Data from sites like Google, Facebook, Twitter and others.

- **Raw Data**: Seemingly meaningless clutter-like gibberish in which patterns are masked.

*Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*

-- Gartner
So, It Looks Like We Need **Data** to Live Intelligently

- Making *smart (?)* decisions:
  - **Can we make reliable decisions without data?**
  - **Is the quality of our society diminished by bad or missing data?**
  - **How can we improve commerce, trade without knowledge from data?**
  - **How can we make better health decisions without knowledge from data?**

- You could give surveys to gather ideas from people but few are likely to respond...

*But, when was the last time YOU took a survey?*
Policy Creation by Analytics

Data -> Analysis -> Decision

Policy A
Policy B
Thus, Much Interest in Data Analytics

- The present and future are information-driven
- Some of the decisions made after studying trends in a population
  - **Commerce**: what have customers already bought?
  - **Media**: What themes of films, music make money?
  - **Industry**: What products should we make to build, satisfy a market? Which market?
  - **Life Sciences and Medicine**: Reasons for sickness? Bad types of foods? Exposures to toxins?
Your Career Could Be Here!

“Big Data & Analytics Is The Most Wanted Expertise By 75% Of IoT (Internet of Things) Providers”


“75% of IoT providers are prioritizing big data and analytics expertise in their hiring decisions.”


“68% of vendors developing IoT solutions are struggling to find and recruit employees with development expertise.”

“75% of firms are prioritizing big data and analytics expertise in their hiring decisions, stating that having these skills is critical for any candidate to be considered an IoT (Internet of Things) expert.”

<table>
<thead>
<tr>
<th>Skill</th>
<th>% Technological skills necessary for IoT experts</th>
<th>% Difficult areas in hiring for IoT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data analytics and big data</td>
<td>75%</td>
<td>35%</td>
</tr>
<tr>
<td>Embedded software development</td>
<td>71%</td>
<td>33%</td>
</tr>
<tr>
<td>Embedded electronics</td>
<td>64%</td>
<td>32%</td>
</tr>
<tr>
<td>IT security</td>
<td>68%</td>
<td>31%</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>Cloud software development</td>
<td>61%</td>
<td>17%</td>
</tr>
<tr>
<td>Automation</td>
<td>45%</td>
<td>14%</td>
</tr>
<tr>
<td>Robotics</td>
<td>21%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Source: Internet of Things Business Report, Defining IoT Business Models
Glassdoor Informs of Careers

**Prodigy Game 4.3 ★
Game Data Analyst**

**Bonus Points For:**
- Degree in Engineering, Computer Science, Stats, Mathematics
- Demonstrated ability to solve hard mathematical, algorithmic, and statistical problems
- Expertise in advanced game analytics - user segmentation, player modelling
- Knowledge of Python/R languages
- Experience working with cloud platforms like AWS (Redshift, Athena, S3)
- Experience working with A/B testing and experimentation
- Significant accomplishments that required both technical and strategic capabilities, such as research projects, open source software contributions, and entrepreneurship

**ADURO 3.9 ★
Data Analyst**

**Why (Glassdoor)](https://www.glassdoor.com)

- Demonstrated ability to maintain absolute confidentiality
- Proficiency in BI Tools (Sisense, Tableau, PowerPivot, DOMO or other comparable tools)
- Aptitude with SQL, R, and other languages supporting data analysis
- Experience with C based object-oriented programming
- Experience working within a large reporting Data Warehouse
- Experience working with web and application analytics tools (Firebase, Google Analytics)
- Experience using analytics to support product development
- Familiar with source-control repositories and associated practices (Git, GitHub)
- Proven attention to detail and accuracy

- **An Analytics Expert**

  - To apply data analysis skills to help development teams better understand users by applying analytics

  - Find and integrate data from multiple sources to provide analysis

  - Develop tools & methods to ensure data accuracy

  - Collaborate with Data & Analytics team members

  - R skills
Consider This ...

- You are given the lists of words from several mainstream-news articles.
- Pick a list to work on with a group of your peers.
- Although the article text cannot be read directly, can you determine the general sense of the article from a list of its words?
- What is the general subject of your article?
  - Are there names of people you recognize in your list? What can you infer about the article from the name(s)?
  - Do the listed nouns support your conclusions?
  - What type of media source would contain such a story?

Find the data at:
https://www.cs.allegheny.edu/sites/obonhamcarter/cs301_resources.html
Please Read for Next Class

- Come prepared to discuss
- *Twelve Million Phones, One Dataset, Zero Privacy*, A New York Times opinion piece

*Opinion* | **THE PRIVACY PROJECT**

*Twelve Million Phones, One Dataset, Zero Privacy*

By Stuart A. Thompson and Charlie Warzel

DEC. 19, 2019