Lecture 10: Real-World Applications of Data Science

“B****es be yearning my earnings concerning machine learning,
Your girl started flirting when she saw my code churning”

Young’s Modulus
Agenda

- Data-Driven Thinking
- Data Science in the Real World
- An Important Note on Ethics
- Ideating Side Projects
- Next Steps
- Courses at Cornell
- Careers in Data Science
Data-Driven Thinking
Going beyond traditional problem-solving

Problem
How can we use data to solve it?
Use Available Data
Collect Data
(or both!)

Available Data
What can we find out?
Solve problems
Generate additional value

How can we use data to solve it?
Use Available Data
Collect Data
(or both!)

What can we find out?
Solve problems
Generate additional value

Going beyond traditional problem-solving
Data-Driven Thinking

Traditional Approach

Problem

How can we use data to solve it?

Use Available Data

Collect Data

(or both!)

Sample Problems

1. Who will win the 2020 Elections?
   FiveThirtyEight

2. Does a patient have lung cancer?
   Data Science Bowl '17

3. Roads are unsafe with increasing traffic.
   DataKind & Vision Zero
Data-Driven Thinking
The New Approach

Available Data

What can we find out?

Solve problems Generate additional value

Sample Data

1. What are the interests of internet user X?
   Advertising

2. All Traffic Data in a city
   Optimizing signals, opening up a new business, traffic sign placement

3. All hip-hop music lyrics ever
   RapStats, Rap Analysis Project
Let’s think data!
Exploring Real-World Applications

1. Advertising
   - Case Study - Cambridge Analytica: Data Science in Political Campaigning

2. Healthcare
   - Case Study – BiliScreen: A Selfie to Diagnose Pancreatic Cancer

3. Media
   - Case Study – How Netflix Keeps You Hooked

4. Social Impact
   - Case Study – Fighting Human Trafficking with Data
Advertising

Machine Learning: The Modern Mad Men

**Context**

Some Big Tech giants earn the bulk of their revenue through ads. One usually earns money when the ad is ‘clicked’ by the user (this differs!). Users are most likely to click on ads when the ads are relevant to them. Ads could be tailored to users only when there is data on the users.

### Sample Data (Extremely small slice): What can you interpret?

<table>
<thead>
<tr>
<th>c_id</th>
<th>ip</th>
<th>loc</th>
<th>city</th>
<th>state</th>
<th>link</th>
<th>time</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3d5wf31</td>
<td>128.83.126</td>
<td>(68.3, 98.5)</td>
<td>Hoboken</td>
<td>NJ</td>
<td>../cutefallskirts</td>
<td>143s</td>
<td>07:56:31</td>
</tr>
<tr>
<td>6d1wd34</td>
<td>128.45.313</td>
<td>(62.3, 89.5)</td>
<td>SYR</td>
<td>NY</td>
<td>../shoestobuy</td>
<td>9s</td>
<td>07:56:35</td>
</tr>
<tr>
<td>3d5wf31</td>
<td>341.34.345</td>
<td>(68.5, 98.6)</td>
<td>NYC</td>
<td>NY</td>
<td>../excelhelp</td>
<td>552s</td>
<td>14:42:23</td>
</tr>
</tbody>
</table>
## Advertising

<table>
<thead>
<tr>
<th>c_id</th>
<th>ip</th>
<th>loc</th>
<th>city</th>
<th>state</th>
<th>link</th>
<th>time</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3d5wf31</td>
<td>128.83.126</td>
<td>(68.3, 98.5)</td>
<td>Hoboken</td>
<td>NJ</td>
<td>../cutefallskirts</td>
<td>143s</td>
<td>07:56:31</td>
</tr>
<tr>
<td>6d1wd34</td>
<td>128.45.313</td>
<td>(62.3, 89.5)</td>
<td>SYR</td>
<td>NY</td>
<td>../shoestobuy</td>
<td>9s</td>
<td>07:56:35</td>
</tr>
<tr>
<td>3d5wf31</td>
<td>341.34.345</td>
<td>(68.5, 98.6)</td>
<td>NYC</td>
<td>NY</td>
<td>../excelhelp</td>
<td>552s</td>
<td>14:42:23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_id</th>
<th>ip</th>
<th>loc</th>
<th>city</th>
<th>state</th>
<th>link</th>
<th>time</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3d5wf31</td>
<td>128.83.126</td>
<td>(68.3, 98.5)</td>
<td>Hoboken</td>
<td>NJ</td>
<td>../cutefallskirts</td>
<td>143s</td>
<td>07:56:31</td>
</tr>
<tr>
<td>3d5wf31</td>
<td>341.34.345</td>
<td>(68.5, 98.6)</td>
<td>NYC</td>
<td>NY</td>
<td>../excelhelp</td>
<td>552s</td>
<td>14:42:23</td>
</tr>
<tr>
<td>6d1wd34</td>
<td>128.45.313</td>
<td>(62.3, 89.5)</td>
<td>SYR</td>
<td>NY</td>
<td>../shoestobuy</td>
<td>9s</td>
<td>07:56:35</td>
</tr>
</tbody>
</table>

**Objective:** Get data on the users
Advertising

<table>
<thead>
<tr>
<th>c_id</th>
<th>ip</th>
<th>loc</th>
<th>city</th>
<th>state</th>
<th>link</th>
<th>time</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>3d5wf31</td>
<td>128.83.126</td>
<td>(68.3, 98.5)</td>
<td>Hoboken</td>
<td>NJ</td>
<td>../cutefallskirts</td>
<td>143s</td>
<td>07:56:31</td>
</tr>
<tr>
<td></td>
<td>341.34.345</td>
<td>(68.5, 98.6)</td>
<td>NYC</td>
<td>NY</td>
<td>../excelhelp</td>
<td>552s</td>
<td>14:42:23</td>
</tr>
</tbody>
</table>

Hypotheses:

- Lives in NJ and works in NYC
- Lives in area with average rent: $r
- Lives in area with average income: $i
- Works in area with average salary: $s
- Falls in k income bracket (Estimated)
- Takes NJTransit to work
- Takes the 67 Train at 8:05am
- Works at XYZ Company
- Works in Business/Data Analytics
- Is a Female
- Is interested in topics A, B, C

With **enough data** and **testing**, the hypotheses could be affirmed or rejected.
Cambridge Analytica: Data Science in Political Campaigning

Case Study

Overview
Cambridge Analytica combined *data analytics, behavioral sciences*, and *innovative ad tech* to influence voters. Widely regarded as instrumental in the result of the 2016 Elections, and many more across the globe.

Methodology
- Facebook activity
- Surveys
- Misc. external data

Data on Voters → Behavioral Analyses → Personalized Ads

Example
- Likes, Comments, Surveys, etc.

+ Life Stage + Political Leaning + Location + Educational Status + …

Source: towardsdatascience.com/effect-of-cambridge-analyticas-facebook-ads-on-the-2016-us-presidential-election-dac5462155d
Healthcare

All-round betterment in the healthcare industry

- Automated Prescriptions
  - Patient Analytics
  - Case Prioritization
  - Assisted follow-through
  - Personalized Care

- Drug Discovery
- Gene Analytics and Editing
- Drug Comparative Effectiveness

Source: https://blog.appliedai.com/healthcare-ai/
BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

Overview
A smartphone app that captures pictures of the eye and produces an estimate of a person’s bilirubin level
Uses: (1) A 3D-printed box that controls the eyes’ exposure to light
(2) Paper glasses with colored squares for calibration

Methodology

BiliScreen: A Selfie to Diagnose Pancreatic Cancer

Case Study

Overview
A smartphone app that captures pictures of the eye and produces an estimate of a person’s bilirubin level

Uses: (1) A 3D-printed box that controls the eyes’ exposure to light
(2) Paper glasses with colored squares for calibration

Methodology

Random Forest
with 10-fold Cross Validation

89.7% Sensitivity
96.8% Specificity

Media: Recommender Systems
How Netflix keeps you hooked

Overview
Most of Netflix’s views (~80%) come through recommendations
The famous Netflix Challenge offered $1m to the participant that could do better than Netflix’s recommender system
These algorithms are relatively simple and intuitive, but extremely effective

<table>
<thead>
<tr>
<th>c_id</th>
<th>movie</th>
<th>tags</th>
<th>time</th>
<th>duration</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Avengers</td>
<td>Action, Superhero</td>
<td>07:56:31</td>
<td>112m</td>
<td>5/5</td>
</tr>
<tr>
<td>A</td>
<td>Mr. Bean</td>
<td>Comedy</td>
<td>07:36:35</td>
<td>3m</td>
<td>2/5</td>
</tr>
<tr>
<td>B</td>
<td>Batman</td>
<td>Superhero</td>
<td>14:42:23</td>
<td>59m</td>
<td>4/5</td>
</tr>
<tr>
<td>B</td>
<td>Black Mirror</td>
<td>Sci-Fi</td>
<td>07:56:34</td>
<td>142m</td>
<td>5/5</td>
</tr>
</tbody>
</table>

Sample: What would you recommend A next?
Usually, many other features and tags for the movies/shows would exist in the database as well
Media: Recommender Systems

How Netflix keeps you hooked

<table>
<thead>
<tr>
<th>c_id</th>
<th>movie</th>
<th>tags</th>
<th>time</th>
<th>duration</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Avengers</td>
<td>Action, Superhero</td>
<td>07:56:31</td>
<td>112m</td>
<td>5/5</td>
</tr>
<tr>
<td></td>
<td>Mr. Bean</td>
<td>Comedy</td>
<td>07:36:35</td>
<td>3m</td>
<td>2/5</td>
</tr>
<tr>
<td>B</td>
<td>Batman</td>
<td>Superhero</td>
<td>14:42:23</td>
<td>59m</td>
<td>4/5</td>
</tr>
<tr>
<td></td>
<td>Black Mirror</td>
<td>Sci-Fi</td>
<td>07:56:34</td>
<td>142m</td>
<td>5/5</td>
</tr>
</tbody>
</table>

Sample: What would you recommend A next?

- **Sci-Fi Movie**
  - Eg. Black Mirror
  - Collaborative Filtering

- **Action Movie**
  - Eg. The Terminator
  - Content-Based Filtering

Read More: towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada
Where else are recommender systems applicable?

- Amazon
- Spotify
- OkCupid
- Best Buy
Social Impact
Data Science for Social Good

Overview
Advanced analytics for social impact is becoming increasingly popular due to innumerable low-cost and high-impact applications

- Marine Data Science
- Data Science in Agriculture
- Big Data for Refugee Resettlement
- Saving Water in Drought-Stricken California
- Expanding Economic Opportunity for low-income people
- Data Science to Combat Trafficking
Predicting End Location: Tackling Human Trafficking

Case Study

Overview

Human trafficking is a great cause of concern, especially in developing countries. ML could be leveraged to aid ground rescue operations for trafficking victims.

Rescued Victims Data

Native Location, End Location, End Industry, Age, Sex, etc.

?  Probable End Locations

Probable End Industries

Social Impact
Predicting End Location: Tackling Human Trafficking

Case Study

Overview
Human trafficking is a great cause of concern, especially in developing countries. ML could be leveraged to aid ground rescue operations for trafficking victims.

Rescued Victims Data

Native Location, End Location, End Industry, Age, Sex, etc.

Classification Model

SVM, Decision Trees, kNN

Probable End Locations

Probable End Industries

Social Impact
Other Applications

Education
Adaptive-learning technology that could recommend material based on student’s success and engagement

Public Sector
Identifying tax-fraud using alternate data such as browsing history, retail data, or payments history.

Crisis
Predicting the progression of wildfires to optimize the response of firefighters.

An Important Note on Ethics

The ACM Code of Ethics and the Ethical Guidelines for Statistical Practice (American Statistical Association) are good places to start. It’s easy to get caught up in the technical challenge, but it is important to know that your work may affect other people directly or indirectly, now or in the future. Ask yourself the following questions often:

- Does your data or analysis impede on anyone’s privacy?
- Did the people give consent for their data to be used?
- Were the people given the option to opt out?
- Who has the right of access to your data?
- Who owns the data?
- Was the data anonymized sufficiently?
- Was there any bias in your dataset against certain sections of the society?
- Are you introducing any bias?
- Should you include any features that may be discriminatory?
- Is your analysis transparent?
- Are the end users aware of shortcomings?

‘Anonymous’ Data? Think again.
Looking Forward
Ideating Side Projects

Towards Data Science is a good place to start for quick reads. You could also follow pages and personalities on your preferred social media.

1. Dig into your own data – Health, Messages, Spotify, etc.
2. Make something you’d use.
3. Look at issues from a social/economic/political lens.
4. ...There’s always Kaggle and data.gov

I recommend Cassie Kozyro's articles!
Next Steps
Path to becoming a data scientist

Math
- Linear Algebra
- Prob/Stats

Data Analysis
- Data Wrangling
- Data Visualization

Machine Learning

Text Analysis

Big Data
- Data Engineering

Gathering, EDA, Deployment

Software Engineering Skills

Business Acumen
Courses @ Cornell

Examples of (some) relevant courses!

Math
- Linear Algebra
- Prob/Stats

Data Analysis
- Data Wrangling
- Data Visualization

Machine Learning
- Data Engineering

Text Analysis

Big Data

Gathering, EDA, Deployment
- INFO 2950

Software Engineering Skills
- CS 1110
- CS 2110
- CS 5150

Business Acumen & Domain Knowledge
- Read!

Other: CS 4700, CS 4670, CS 4787, etc.

Note: This is not an official list, and does not represent the views of Cornell Data Science.
Common roles and their meanings

**Data Analyst**
These are typically the roles right out of undergrad. You’ll likely be working with SQL/Excel (and maybe a little bit of Python/R).

**Data Scientist**
This role typically covers responsibilities additional to those that data analysts have. You’ll be expected to have a strong understanding of math fundamentals, and machine learning models. It’s also a good idea to be well-versed in programming.

**Data Engineer**
As a data engineer, you’ll be managing the data infrastructure – building data pipelines, pushing code into production, etc. You would ideally like to be well-versed in software development and have exposure to other software and tools your target companies use.

**Machine Learning Engineer**
This is similar to the data scientist role, but is more specific to building machine learning models. You would like be required to have a robust knowledge of applied math and software development.
Product Analytics
Focused on a certain product and the behaviors of the user’s product. For example, you may be working on boosting customer engagement using clickstream data.

Business Intelligence
Focused on creating business insights from your products/services and informing internal decisions. For example, you may be generating reports of number of users on your platform.

Source: Business Broadway
That’s all folks!

Just Kidding

• **Final Project Due**: May 13, 2020
• **Course Feedback Form** out soon!
• **Course Staff Invitations** out in summer
• **Office Hours go on** until May 13, 2020
• Stay tuned for **CDS Recruitment** next semester!
• **Get in touch**: tb444@cornell.edu

Thank you all for taking this class, and for an incredible semester. Good luck on finals, and stay safe!