Data Science: A Review

Stats 5, Winter 2018

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Department Or Organization</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 9</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>Jan 16</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>Jan 23</td>
<td>Michael Carey</td>
<td>Computer Science</td>
<td>Databases and Data Management</td>
</tr>
<tr>
<td>Jan 30</td>
<td>Sameer Singh</td>
<td>Computer Science</td>
<td>Statistical Natural Language Processing</td>
</tr>
<tr>
<td>Feb 6</td>
<td>Zhaoxia Yu</td>
<td>Statistics</td>
<td>An Introduction to Cluster Analysis</td>
</tr>
<tr>
<td>Feb 13</td>
<td>Erik Sudderth</td>
<td>Computer Science</td>
<td>Computer Vision and Machine Learning</td>
</tr>
<tr>
<td>Feb 20</td>
<td>John Brock</td>
<td>Cylance, Inc</td>
<td>Data Science and CyberSecurity</td>
</tr>
<tr>
<td>Feb 27</td>
<td>Video Lecture (Kate Crawford)</td>
<td>Microsoft Research and NYU</td>
<td>Bias in Machine Learning</td>
</tr>
<tr>
<td>Mar 6</td>
<td>Matt Harding</td>
<td>Economics</td>
<td>Data Science in Economics and Finance</td>
</tr>
<tr>
<td>Mar 13</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Review: Past and Future of Data Science</td>
</tr>
</tbody>
</table>
Components of Data Science

- **Statistics** (Mathematical and Probabilistic Foundations)
- **Computing** (Algorithms and Software)
- **Applications** (Analyzing Real Data)
Core Data Science Skills

- Database systems
- Programming (Python, R, C, etc)
- Software engineering
- Algorithms
- Matrix-vector algebra and calculus
- Probability
- Machine learning
- Statistical modeling
- Communication and writing skills
What Classes will you take in the DS Major?

**Statistics**
- 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)

**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)
  (CS 172: Neural Networks/Deep Learning)

**Applications**
- Stats 170AB: Data Science Capstone Project
- INF 143: Information Visualization
  (INF 131: Human Computer Interaction)
  (CS 121: Information Retrieval)
  (CS 122B: Project in Databases/Web Applications)
  (Summer internships, e.g., junior year)

(Sample electives shown in parentheses)
Stats 170AB: Project in Data Science

Unstructured Data → Extracted Data → Transformed Data → Data for Modeling → Predictive Model → Predictions/Decisions

- Extracted Data
- Transformed Data
- Data for Modeling
- Predictive Model
- Predictions/Decisions

- Unstructured Data
- Extracted Data
- Transformed Data
- Data for Modeling
- Predictive Model
- Predictions/Decisions
Stats 170AB: Project in Data Science

$y_t = \beta_1 x_{1t} + \mu_t + \epsilon_t$

Python for Data Analysis

scikit-learn

Classification
Identifying to which set of categories a new observation belongs.
Applications: Credit risk assessment, image recognition.
Algorithms: SVM, random forest, decision tree.

Regression
Predicting a continuous value for a new example.
Applications: Stock price prediction, house price estimation.
Algorithms: SVM, ridge regression, Lasso.

Dimensionality reduction
Reducing the number of variables to consider.

Model selection
Choosing, validating, and choosing parameters and models.

Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, gene expression data analysis.
Algorithms: k-Means, hierarchical clustering, DBSCAN.

Preprocessing
Feature extraction and normalization.
Applications: Transforming input data such as text for use with machine learning algorithms.

Jupyter

Who uses scikit-learn?
Stats 170AB: Project in Data Science

Text from 4 million Wikipedia articles
Data Science Skills: the Spectrum of Data Analysis

- **Raw Data**
  - Databases, Algorithms, Software Engineering
- **Data Wrangling**
- **Data Management**
  - Predictive Modeling
  - Exploratory Data Analysis
- **Consumers**
  - External Business Customers
  - Internal Business Customers
- **Government**
- **Scientists**
  - Business, scientific, medical, and other domain knowledge

- **Machine Learning, Statistics**
## Sample Course of Study in the Major

Years 1 and 2: foundational courses in computer science, mathematics, statistics, including statistical computing

### 2015-16, First Year: 41 units

<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>4</td>
</tr>
</tbody>
</table>

### 2016-17, Second Year: 46 units

<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>4</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
Topics from Lectures this Quarter

Core methodologies

Databases
Machine Learning
Clustering Algorithms

Technologies

Natural Language Processing
Computer Vision

Applications

Economics And Business
Cybersecurity
Fairness and Bias
Final Assignment

- Write a ½ to 1 page short essay that takes any two of the topics from lectures 2 to 9, and describes how you think the two topics could “intersect” going forward,
e.g.,
  - What aspects of each method could be combined to produce new ideas?
  - What new applications might be enabled by combining these methods?
  - What are the potential challenges in these areas?

- Possible combinations
  - Natural language and cybersecurity
  - Clustering algorithms and computer vision
  - Computer vision and fairness/bias

- ...feel free to pick any 2 topics that interest you
Final Assignment Instructions

• Put your name and student ID at the top of the page

• Submit as a PDF file

• Due to EEE dropbox by 9am on Monday March 19th (next week)

• Note: there is no final exam in this class
How is data measured and collected?
## 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>
Text Collections

NYT
330,000 articles

Enron
250,000 emails

NSF/ NIH
100,000 grants

16 million Medline articles
Examples of Student Clickstream Data
Yelp Dataset Challenge

Discover what insights lie hidden in our data.

The Challenge
We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification
Maybe you’ve heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?

Natural Language Processing & Sentiment Analysis
What’s in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining
We recently launched our Local Graph but can you take the graph further? How do user’s relationships define their usage through? Who are the trend setters eating before it becomes popular?

5.2 million reviews
174k businesses
11 metropolitan areas

Round 11
Our dataset has been updated for this iteration of the challenge - we’re sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 11 metropolitan areas across 4 countries. Round 11 has kicked off starting January 18, 2018 and will run through June 30, 2018.

Download Dataset
Yelp Challenge Data Set
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
NASA’s MODIS satellite

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

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

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.

Source: UC Davis

Graphic from Orange County Register
Sensors Measuring Human Activity

Optical people counter at a building entrance on campus

Loop sensors on Southern California freeways
BASEBALL GAME EVENTS

COUNTS

SUN MON TUE WED THU FRI SAT
Ebird.org

Over 1.5 million submissions per month

From Wood et al, PLOS Biology, 2011
What are potential issues with data collection?
Geolocated Tweets in Southern California
Geolocated Tweets around UC Irvine
Sensors Measuring Human Activity

Optical people counter at a building entrance on campus

Loop sensors on Southern California freeways
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
Why is data management and organization important?
Computer Architecture 101
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

RAM → Disk

1 meter

100 km
Data Engineering at Web Scale
Why is it important to explore and understand data before analysis?
Histogram of Unimodal Data

1000 data points simulated from a Normal distribution, mean 10, variance 1, 30 bins
Histogram of Age at Death of 68,000 individuals

Notice anything unusual?
Summary Statistics

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
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
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

![Graphs showing actual correlation values](image)

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

## Data Set on Housing Prices in Boston

(widely used data set in research on prediction 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

- Percentage of large residential lots
- Distance to employment centers
- Median house value

- Crime Rate
- Industry
- Nitrous oxide
- Average # rooms
- Proportion of old houses
- Highway accessibility
- Property tax rate
- Student-teacher ratio
Human judgement is important in data analysis
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

Correlation: 95%
Sources: CDC & US Census tylervigen.com

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 (CDC)</th>
<th>Total revenue generated by skiing facilities (US) (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

Today's Random Medical News

According to a report released today...

- Exercise
- Fatty foods
- Stress
- Red wine
- Coffee
- Computer terminals
- Daycare
- Smoking
- Heart disease
- Hypothermia
- A feeling of well-being
- Depression
- Migraines
- Twins
- Spontaneous regression
- Glaucoma
- Arthritis
- Twins
- Seven out of ten
- Men 25-40
- Overweight smokers
- Two-income families
- Women

from the New England Journal of Panic-Inducing Gobbledygook
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.
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.
What are the legal and ethical aspects of data analysis?
Who Owns Your Data?
Collection of Individual-Level Data

1960’s

1980’s

2000’s

2020’s
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 analysis techniques might be developed?

What new application areas might emerge?

What are the societal implications of data science?
Final Assignment

• Write a ½ to 1 page short essay that takes any two of the topics from lectures 2 to 9, and describes how you think the two topics could “intersect” going forward, e.g.,
  – What aspects of each method could be combined to produce new ideas?
  – What new applications might be enabled by combining these methods?
  – What are the potential challenges in these areas?

• Possible combinations
  – Natural language and cybersecurity
  – Clustering algorithms and computer vision
  – Computer vision and fairness/bias
  – ...feel free to pick any 2 topics that interest you
Final Assignment Instructions

• Put your name and student ID at the top of the page

• Submit as a PDF file

• Due to EEE dropbox by 9am on Monday March 19th (next week)

• Note: there is **no final exam** in this class