Stats5 Seminar: Introduction to Data Science

Winter 2018

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

• Class organization and topics

• History of data analysis

• Data science and real-world applications

• The Data Science Major

• Limitations of what we can do with data
Class Organization

- Meet weekly for 40 minute seminar with 5-10 minute discussion

- 8 topics (with guest speakers), weeks 2 through 9
  - You are encouraged to ask questions during and after the talks

- Intro and wrap-up talks in weeks 1 and 10

- Class Web site is at [www.ics.uci.edu/~smyth/courses/stats5](http://www.ics.uci.edu/~smyth/courses/stats5)
  - Slides and related materials will be posted during the quarter
## 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>
Submission of Review Forms (Weeks 2 to 10)

• Submit Review forms for Lectures 2 through 10

• Review forms will be available online at the start of each class
  – A few relatively short questions based on the lecture that day
  – Needs to be submitted to EEE by noon for each lecture
  – Bring your laptop or other device

• Requirements to pass the class
  – Attend and submit review form for least 8 lectures for weeks 2 through 10
    (allowed to miss one if you need to for some reason)

• No final exam: pass/fail based on attendance and review forms
Academic Integrity

- The review form you submit each week must be
  (a) written by you, and
  (b) written during the lecture that week

- Failure to adhere to this policy may result in failing the class

- It is the responsibility of each student to be familiar with UCI's Academic Integrity Policies and UCI's definitions and examples of academic misconduct. See the class Web site for additional info.
A BRIEF HISTORY OF DATA ANALYSIS AND COMPUTING
Computers and Data

The historical meaning of the term “computer”:
“one who computes” (i.e., a person)

Since the 1700’s, statisticians have been using “computers” to analyze data – so it’s not a new idea.
Computers and Data

The historical meaning of the term “computer”: “one who computes” (i.e., a person)

Since the 1700’s, statisticians have been using “computers” to analyze data – so its not a new idea

For example, Karl Pearson, one of the founders of statistics, directed a team of “computers” in his lab in London around the early 1900’s

…..but for many years, “computers” could only work on relatively small problems
Statistics and Modern Computing

• Post World War II
  – Increasing use of computing to solve algorithmic aspects of statistical analyses

• 1960’s
  – Development of statistical computing and exploratory data analysis

• 1980’s
  – Computing allowed statisticians to explore more flexible models
  – Increase in use of “non-parametric” techniques and simulation methods

• 1990’s
  – Development of “machine learning” – very flexible predictive modeling techniques developed in computer science

• Today
  – Data science = computing + statistics + applications
1985: ~ $100k per gigabyte

2015: ~ $0.3 cents per gigabyte
From http://exploringbigdata.blogspot.com/
Modeling Human Behavior using Social Media

From Lichman and Smyth, ACM SIGKDD 2014
Geolocated Tweets around UC Irvine
Scientific Data: Large Hadron Collider at CERN

60 Terabytes/day
20 Petabytes/year

1 Terabyte = $10^{12}$ bytes
1 Petabyte = $10^{15}$ bytes
A Paradigm Shift in Data Analysis

• Technological drivers
  – Sensors (cheap and ubiquitous, e.g., GPS on your phone)
  – Data storage (we are all “data owners”)
  – Computational power
  – Data analysis methods (statistics and machine learning)
  – Internet and wireless communication (can collect and share data)

• Convergence.....tremendous demand for data analysis
  – In business, in sciences, in medicine, in engineering, and more......

• In the past, this demand was met by statistics
  – Does not scale up – there are not nearly enough statisticians
  – Need more tools than just statistics....need databases, algorithms, machine learning,....
DATA SCIENCE IN THE REAL WORLD
What is Data Science?

Data science involves the full lifecycle of data:
from real-world unstructured data.....to predictions and decisions

Data science is broader than just databases, statistics, ML, algorithms
.....but these are all critical components

Key aspects of data science include
– Domain knowledge and problem definition
– Data preparation/organization/management
– Understanding of uncertainty (statistics)
– Computing, algorithms, fitting models, machine learning
– Iterative exploration and experimentation
– Human judgement and interpretation
How is Data Science used in each of these Organizations?
How is Data Science used in each of these Organizations?
Organizations

- Facebook
- Google
- Amazon
- Spotify
- Disney
- Kaiser Permanente
- Blizzard
- Honda

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support
Organizations

- Facebook
- Google
- Amazon
- Spotify
- Disney
- Kaiser Permanente
- Blizzard Entertainment
- Honda

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support
How does Facebook predict what content to show you?

MONTHLY USERS ON FACEBOOK 2004-2017

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
Web Search: How do search engines rank search results?
How do ad companies decide what online ads to show you?
How does Amazon forecast how many items for its warehouses?

From dailymail.co.uk

From www.formaspace.com

From linkedin.com
How do autonomous cars recognize objects in image data?
How can we use wearable data to improve our health?

Images from community.fitbit.com
How can we make personalized recommendations in medicine?

Data Matrix:
Rows = genes
Columns = patients
Astronomy: How can we process terabytes/day of telescope data?

Large Synoptic Telescope (LST)
15 Terabytes/day
100+ Petabytes in 10 years

From Raddick et al, Astronomy Education Review, 2009
Physics: What is required to search for new physics particles?

Large Hadron Collider:
700 Mbytes/second
60 Terabytes/day
20 Petabytes/year
How can we detect land changes in NASA satellite images?

From www.spot-7.com

From http://cimss.ssec.wisc.edu/
How can algorithms interpret and summarize sports data?
Politics: How can we reliably predict events like elections?

Data Pipelines

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

Unstructured Data

Extracted Data

Transformed Data

Data for Modeling

Predictive Model

Predictions/Decisions
Hidden Technical Debt in Machine Learning Systems

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Scullley et al, NIPS 2015 Conference
THE DATA SCIENCE MAJOR
All of the applications we discussed are built on ideas from...

- Database systems
- Algorithms
- Software engineering
- Machine learning
- Probabilistic and statistical models
- Quantification of uncertainty
- Data visualization
- and more...
Components of Data Science

- **Statistics**
  (Mathematical and Probabilistic Foundations)

- **Computing**
  (Algorithms and Software)

- **Applications**
  (Analyzing Real Data)
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)
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
Research at UC Irvine in Data Science
LIMITATIONS OF WHAT WE CAN LEARN FROM DATA
Today's Random Medical News

Today's Random Medical News from the New England Journal of Panic-Inducing Grabology

- Smoking
- Coffee
- Computer terminals
- Exercise
- Fatty foods
- Stress
- Red wine
- Daycare

Can cause:
- Hypothermia
- Heart disease
- Breast cancer
- Spontaneous remission
- Depression
- Migraines
- A feeling of well-being

In:
- Twins
- Arthritis sufferers
- 7 out of 10 women

Two-income families
- Men 25-40
- Overweight smokers

According to a report released today...
Kidney Cancer Death Rates by County in the US

Lowest Rates

Highest Rates

From A. Gelman and D. Nolan
Oxford University Press, 2002
(see also response letters at http://faculty.washington.edu/kenrice/natureletter.pdf)
How Much Climate Data Do We Actually Have?

Image from http://cimss.ssec.wisc.edu/

Image from ipcc.ch
## 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>Classification Algorithms in 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>Zhoxia 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</td>
<td>Microsoft Research and NYU</td>
<td>Bias in Machine Learning</td>
</tr>
<tr>
<td></td>
<td>(Kate Crawford)</td>
<td></td>
<td></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>