Introduction to Data Science

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

Stats 5 Seminar, Winter 2016
Outline

• What is Data Science?

• The UCI Data Science Major

• Data Science Activities at UCI
  – The UCI Data Science Initiative

• Data Science in the Real World
  – Data sets and applications
Class Organization

• Meet weekly for 50 minute seminar

• 8 guest speakers, weeks 2 through 9
  – You are encouraged to ask questions during the talks
  – You will be asked to fill out a short form to hand in, in class, by the end of each talk

• Required to attend at least 7 of the 8 guest speakers to pass the class

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

• Class Web site is at www.ics.uci.edu/~smyth/courses/stats5
  – Slides and related materials will be posted during the quarter
# Schedule of Speakers (from class Website)

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 5th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>2</td>
<td>Jan 12th</td>
<td>Alex Ihler</td>
<td>Computer Science</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>3</td>
<td>Jan 19th</td>
<td>Jim Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>4</td>
<td>Jan 26th</td>
<td>Charless Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision using Machine Learning</td>
</tr>
<tr>
<td>5</td>
<td>Feb 2nd</td>
<td>Alfred Kobsa</td>
<td>Informatics</td>
<td>Data Privacy and Personalization</td>
</tr>
<tr>
<td>6</td>
<td>Feb 9th</td>
<td>Carter Butts</td>
<td>Social Sciences</td>
<td>Social Network Data Analysis</td>
</tr>
<tr>
<td>7</td>
<td>Feb 16th</td>
<td>Michael Carey</td>
<td>Computer Science</td>
<td>Systems for Big Data</td>
</tr>
<tr>
<td>8</td>
<td>Feb 23rd</td>
<td>Hal Stern</td>
<td>Statistics</td>
<td>Bayesian Statistics</td>
</tr>
<tr>
<td>9</td>
<td>Mar 1st</td>
<td>Stacey Hancock</td>
<td>Statistics</td>
<td>Randomization Methods and the R Language</td>
</tr>
<tr>
<td>10</td>
<td>Mar 8th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>The Future of Data Science</td>
</tr>
</tbody>
</table>
What is Data Science?
A Revolution in Data Technology

Magnetic Data Storage
(Bits Per Dollar, constant 2000 dollars)

Graphic from Ray Kurzweil, singularity.com
Image from en.wikipedia.org/wiki/Technological_singularity
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 the sciences, in medicine, in engineering, in business, and more......

• In the past, this demand was met by statisticians
  – Does not scale up – there are way too few statisticians
  – And even statisticians need computers to analyze complex data
Human Computers

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
Human Computers

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

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**
  - Interface of statistics and computing is very active area in research and applications
What is Data Science?

No precise definition.....
What is Data Science?

No precise definition.....

Understand ing the full lifecycle of data,

from data measurement,

to organization and management of data,

to exploration and understanding of data,

to modeling and analysis of data,

to making predictions and decisions
What is Data Science?

No precise definition.....

Understanding the full lifecycle of data,

from data measurement,

to organization and management of data,

to exploration and understanding of data,

to modeling and analysis of data,

to making predictions and decisions

Note: this is not a 1-way process, typically iterative
What are the Components of Data Science?

- Statistics + Computing are a powerful combination

  - Statistics:
    - the original “data science”: the basic principles of how to generalize from data

  - Computing:
    - the “engine” driving modern data analysis: algorithms and software

- Also important
  - Machine learning: make accurate predictions from data
  - Database principles: efficiently organize, store, and access data
  - Computing systems: large-scale computing
  - Domain knowledge: data analysis does not happen in a vacuum
  - Privacy and policy: understanding data and individual rights
The Science of Data Analysis

Applications of Data Analysis in Business, Medicine, Engineering, Science, Humanities, Computing, Algorithms, Databases + Statistics, Mathematics, Optimization + Privacy, Policy, Decisions
The UCI Data Science Major
New Undergraduate Major in Data Science

Motivation

Increasing demand for graduates with data analysis skills, e.g.,

“*The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings. The shortage of talent is just beginning.*”

(McKinsey Global Institute Study on Big Data, 2011)
Job Trends from Indeed.com

- "data science"
The UCI Major in Data Science

Goal of this Major

Provide students with skills at the intersection of computing and statistics

• Databases, algorithms, machine learning, visualization (computer science)
• Probabilistic modeling, statistical analysis, prediction (statistics)

Status

Proposal for the major was approved by UCI in April 2015
Major officially opened for student enrollment in Fall 2016

8 students currently enrolled in the major
First students will graduate in summer 2018

Program is run jointly by Statistics and Computer Science departments
Web Sites and Resources for the Major

- General information about the major: http://www.ics.uci.edu/ugrad/degrees/degree_datascience.php

- UCI Catalog Description of the major (courses, etc) http://catalogue.uci.edu/donaldbrenschoolofinformationandcomputerscience/departmentofstatistics/#majortext

Department of Statistics

On This Page:
- Admissions
- Requirements for the Bachelor's Degree in Data Science
- Career Opportunities

Undergraduate Major in Data Science

The Data Science Major prepares students for a career in data analysis, combining foundational statistical concepts with computational principles from computer science. In the first two years of the program students will take core courses in both the Statistics and Computer Science Departments, providing a strong foundation in the principles of each field. In the 3rd and 4th years of the program, students will take more specialized courses, on topics such as design of algorithms, machine learning, information visualization, and Bayesian statistics. A major component of this degree is the final year capstone project course, a 2-quarter course that teaches students how to apply statistical and computational principles to solve large-scale real-world data analysis problems.

Admissions

Freshman Applicants: See the Undergraduate Admissions section.

Transfer Applicants: Junior-level applicants who satisfactorily complete course requirements will be given preference for admission. Applicants must satisfy the following requirements:

1. Completion of one year of college level mathematics (calculus or discrete math) and one semester of college level statistics.
2. Completion of one year of transferable Computer Science courses*; at least one of these should involve concepts such as those found in the Python and C++ programming languages, or another high-level programming language.
## Data Science Major Requirements

<table>
<thead>
<tr>
<th>Lower-division:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WRITING 39A-39B-39C</strong></td>
<td>Introduction to Writing and Rhetoric and Critical Reading and Rhetoric and Argument and Research</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 6B</strong></td>
<td>Boolean Algebra and Logic</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 6D</strong></td>
<td>Discrete Mathematics for Computer Science</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 31</strong></td>
<td>Introduction to Programming</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 32</strong></td>
<td>Programming with Software Libraries</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 33</strong></td>
<td>Intermediate Programming</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 45C</strong></td>
<td>Programming in C/C++ as a Second Language</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 46</strong></td>
<td>Data Structure Implementation and Analysis</td>
</tr>
<tr>
<td><strong>I&amp;C SCI 51</strong></td>
<td>Introductory Computer Organization</td>
</tr>
<tr>
<td><strong>IN4MATX 43</strong></td>
<td>Introduction to Software Engineering</td>
</tr>
<tr>
<td><strong>MATH 2A</strong></td>
<td>Single-Variable Calculus</td>
</tr>
<tr>
<td><strong>MATH 2B</strong></td>
<td>Single-Variable Calculus</td>
</tr>
<tr>
<td><strong>MATH 2D</strong></td>
<td>Multivariable Calculus</td>
</tr>
<tr>
<td><strong>MATH 3A</strong></td>
<td>Introduction to Linear Algebra</td>
</tr>
<tr>
<td>or <strong>I&amp;C SCI 6N</strong></td>
<td>Computational Linear Algebra</td>
</tr>
<tr>
<td><strong>STATS 5</strong></td>
<td>Seminar in Data Science</td>
</tr>
<tr>
<td><strong>STATS 7</strong></td>
<td>Basic Statistics</td>
</tr>
<tr>
<td><strong>STATS 68</strong></td>
<td>Exploratory Data Analysis</td>
</tr>
</tbody>
</table>
## Upper-division:

### A. Data Science core requirements:

<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATS 110</td>
<td>Statistical Methods for Data Analysis I</td>
</tr>
<tr>
<td>STATS 111</td>
<td>Statistical Methods for Data Analysis II</td>
</tr>
<tr>
<td>STATS 112</td>
<td>Statistical Methods for Data Analysis III</td>
</tr>
<tr>
<td>STATS 115</td>
<td>Introduction to Bayesian Data Analysis</td>
</tr>
<tr>
<td>STATS 120A</td>
<td>Introduction to Probability and Statistics</td>
</tr>
<tr>
<td>STATS 120B</td>
<td>Introduction to Probability and Statistics</td>
</tr>
<tr>
<td>STATS 120C</td>
<td>Introduction to Probability and Statistics</td>
</tr>
<tr>
<td>I&amp;C SCI 139W</td>
<td>Critical Writing on Information Technology</td>
</tr>
<tr>
<td>COMPSCI 122A</td>
<td>Introduction to Data Management</td>
</tr>
<tr>
<td>COMPSCI 161</td>
<td>Design and Analysis of Algorithms</td>
</tr>
<tr>
<td>COMPSCI 178</td>
<td>Machine Learning and Data-Mining</td>
</tr>
<tr>
<td>IN4MATX 143</td>
<td>Information Visualization</td>
</tr>
</tbody>
</table>

### B. Three elective courses from the list below:

<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATH 130B</td>
<td>Probability and Stochastic Processes</td>
</tr>
<tr>
<td>MATH 130C</td>
<td>Probability and Stochastic Processes</td>
</tr>
<tr>
<td>STATS 140</td>
<td>Multivariate Statistical Methods</td>
</tr>
<tr>
<td>I&amp;C SCI 53</td>
<td>Principles in System Design</td>
</tr>
<tr>
<td>COMPSCI 111</td>
<td>Digital Image Processing</td>
</tr>
<tr>
<td>COMPSCI 115</td>
<td>Computer Simulation</td>
</tr>
<tr>
<td>COMPSCI 121</td>
<td>Information Retrieval</td>
</tr>
<tr>
<td>COMPSCI 122B</td>
<td>Project in Databases and Web Applications</td>
</tr>
<tr>
<td>COMPSCI 122C</td>
<td>Principles of Data Management</td>
</tr>
</tbody>
</table>
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
Change of Major Requirements for Data Science

- Cumulative UC GPA: 2.7 or higher.

- 3.0 or higher average GPA and no grade lower than a C for ICS 31, ICS 32, and one of the following: Math 2A, Math 2B, Math 2D, ICS 6B, or ICS 6D.

- Students with more than 60 units will be reviewed on a case-by-case basis and may not be admitted to the major.

- Students will not be able to complete the degree in Data Science prior to Spring 2018.

If you are a freshman, contact Dr. Stacey Hancock, stacey.hancock@uci.edu, to inquire about getting a waiver to change into the major.

More generally, you can talk to ICS Student Affairs Office Counselor if you are interested in changing your major to Data Science.
Careers in Machine Learning and Data Science

- Internet/Computing
  - Google, Facebook, Microsoft, Apple, Intel, IBM, + many more
  - Ecommerce: Walmart, Amazon, eBay, etc

- Banking and Finance
  - Banks, credit card companies (Experian), insurance companies
  - Investment firms, Wall Street

- Health
  - Health insurers, personalized medicine/genomics

- Engineering
  - Automotive companies, sensor analytics, systems monitoring

- Government Labs
  - NASA/JPL, Department of Defense, intelligence agencies.
Data Scientist jobs

My recent searches
Data Scientist - Los Angeles, CA
» clear searches

Sort by: relevance - date

Salary Estimate
$45,000+ (17246)
$60,000+ (13735)
$75,000+ (10312)
$90,000+ (6687)
$105,000+ (3972)

Company
Location
Job Type

what: Data Scientist

where: 

Find Jobs

Tip: Enter your city or zip code in the "where" box to show results in your area.

New! Join Indeed Prime - Get offers from great tech companies

Jobs 1 to 10 of 19,855

Senior Data Scientist/Applied Researcher
eBay Inc. ★★★★★ 597 reviews - San Jose, CA
Conceptualize, code, deploy, and iterate on designs from prototypes all the way through to production systems. Analyze petabytes of real-world performance data.
21 days ago - email
Sponsored

Data Scientist
Tremor Video - Boston, MA
Passion for “playing” with tons of data and supporting scientific experiments to improve and validate the performance of algorithms.
12 days ago - email
Sponsored

Machine Learning Algorithm Developer
Lucidyne Technologies, Inc. - Corvallis, OR
The position requires either a PhD degree, Masters degree or equivalent work experience in machine learning, with a focus on machine learning or a related field.
Easily apply
30+ days ago - email
Sponsored

Machine Learning Scientist/Architect - Polygraph required
Resolute Technologies, LLC - Hanover, MD
$190,000 a year
Develop on-sensor machine learning analytics to support multiple operational scenarios. An understanding of SIGINT processing systems, data flows, data formats,
Easily apply
10 hours ago - save job - email - more...
The UCI Data Science Initiative
Kickoff Meeting

Friday, October 24, 2014

1:30-2:00 PM  Introduction
              What is the UCI Data Science Initiative?

2:00-3:20 PM  Data Science Seminar

2:00  Bayesian Statistics
      Hal Stern

2:20  Deep Learning
      Pierre Neiswanger

2:40  Platforms for Data Science
      Michael Franklin

3:00  Me and My Data
      Geof Bock

3:20  Questions and Answers

3:30-3:45 PM  Break

3:45-4:45 PM  Data Science Seminar
Activities

• Research Symposia
  – Algorithms for analyzing social network data (March 2015)
  – Text mining and education data (May 2015)

  – Digital Humanities Symposium (Feb 5\textsuperscript{th} 2016)
  – Machine Learning Symposium (May 20\textsuperscript{th} 2016)
[See UCI Data Science Initiative Website for details]

• Graduate Student Education
  – 1 and 2-day short courses on “hands-on” data science topics
    • Introduction to R, Predictive modeling in Python, etc
  – Over 400 graduate students have participated in 15 course offerings in the past 12 months
  – Summer fellowship program for PhD students on interdisciplinary data science research
More Information about Data Science at UCI

UCI Data Science Initiative
http://datascience.uci.edu
Seminars, short courses, workshops – join the mailing list for more info

Center for Machine Learning and Intelligent Systems
http://cml.ics.uci.edu
Seminars every Monday at 1pm – join the mailing list

Statistics Department
Research seminars on Thursdays, 4pm, during academic quarters

Computer Science Department
Research seminars on Fridays, 11am, during academic quarters
Examples of Large Data and Real-World Applications
What Happens in an Internet Minute?

- 200 million emails sent
- 350,000 new tweets
- 2.5 million search queries issued

Figures from Lars Backstrom, Facebook, 2011

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
Example: Detecting Faces and Pedestrians in Images

Figures from Le Cun and Ranzato, ICML 2013 Tutorial
Example: Gene Expression Data

Data Matrix:
Rows = genes
Columns = patients

Source: Clin Breast Cancer © 2010 CIG Media, LP
Particle Physics: Large Hadron Collider at CERN

700 Mbytes/second
60 Terabytes/day
20 Petabytes/year

1 Terabyte = $10^{12}$ bytes
1 Petabyte = $10^{15}$ bytes

Detecting new types of particles = “Needle in a haystack”

New algorithms for searching massive amounts of data to find unusual patterns

Professor Daniel Whiteson
Department of Physics, UC Irvine
Real-Time Sports Statistics

Inside the Secret World of the Data Crunchers Who Helped Obama Win

Data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle

By Michael Scherer @michaelscherer | Nov. 07, 2012 | 273 Comments

"The cave" at President Obama's campaign headquarters in Chicago
Daily Report: At WWDC, Apple Expected to Expand Into Health and Home Monitoring

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

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.
Modeling Human Behavior using Social Media

From Lichman and Smyth, ACM SIGKDD 2014
Welcome to Livehoods!

Each dot on the map represents a check-in location. Groups of nearby dots of the same color form a Livehood.

The shapes of Livehoods are determined by the patterns of people that check-in to them. If many of the same people check-in to two nearby locations, then these locations will likely be part of the same Livehood.

Livehoods reveal how the people and places of a city come together to form the dynamic character of local urban areas.

Click on a location to learn about its Livehood.
Geolocated Tweets around UC Irvine
Spatial and Temporal Characteristics of Tweets at Disneyland
The Google Books Project

- Google has digitized over 8 million books
  - Books from 40 university libraries around the world
  - 4.5 million in English, rest in other languages. 6% of all books ever published.
  - 500 billion words
  - Spans multiple centuries since 1500’s

- Reading the books manually is impossible
  - Reading only English-language entries since 2000, at the pace of 200 words/minute, with no sleep/food interruptions, would take 80 years!
The Google N-grams Corpus

• **N-grams**
  – 1-gram: a string of characters with no spaces, e.g., “dog”, “NFL”, “3.142”
    • In effect 1grams are “words”
  – N-gram: a sequence of N 1-grams, e.g.,
    • Bigram = sequence of pairs of 1-grams, e.g., “big dog”, “New York”
    • Trigram = sequence of triples of 1 grams, e.g., “New York City”

• Google N-grams corpus allows us to see how individual words (1-grams) and N-grams have changed in usage over the years

• Usage frequency =
  \[
  \frac{\text{(number of instances of an N-gram per year)}}{\text{(total number of words in the collection that year)}}
  \]
Detecting of Linguistic Usage and Change over Time

Political Analysis: Detection of Possible Censorship

- Suppression of mentions of political figures in Russian texts
- Differences in mentions of Tiananmen after 1976 and 1989 incidents, in Chinese (blue) and English (green)

Recommender Systems

e.g., for Web sites that recommend books, movies, etc

![Matrix Diagram](Image)
Recommender Systems

e.g., for Web sites that recommend books, movies, etc

Users

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
The $1 Million Question

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!
Matrix Factorization of Ratings Data

\[ \begin{align*}
\text{m users} & \quad \approx \quad \text{n movies} \\
\text{m users} & \quad \times \quad \text{f} \\
\text{f} & \quad \times \quad \text{n movies}
\end{align*} \]
Matrix Factorization of Ratings Data

$m$ users $\sim$ $n$ movies

"user weights"

$\times$

"movie weights"

$x$

$\mathbf{f}$
Figure from Koren, Bell, Volinsky, IEEE Computer, 2009
Matrix Factorization of Ratings Data

Rating of unseen movie = user weights * movie weights
A Big Cheque for $1 Million Prize
Learning to Predict with Weighted Sums

This is known as a logistic regression model

Each “edge” has a weight or parameter, $\alpha_j$

\[
f(x) = \hat{P}(Y = 1 | x) = \frac{1}{1 + e^{(-\sum_{j=1}^{d} \alpha_j x_j)}}
\]
A Neural Network with 1 Hidden Layer

Can recursively create more complex prediction models

Many more weights now....requires more data to estimate
Deep Learning: Models with 2 or More Hidden Layers

We can build on this idea to create “deep models” with many hidden layers.

The model is now a very flexible highly non-linear function.
Figure from Krizhevsky, Sutskever, Hinton, 2012
ILSVRC top-5 error on ImageNet
A Deep Neural Network for Image Recognition

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014
Poor Extrapolation

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014
Lack of Calibration

From Nguyen, Yosinski, Clune, ArXiv preprint, 2014
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 5th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>2</td>
<td>Jan 12th</td>
<td>Alex Ihler</td>
<td>Computer Science</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>3</td>
<td>Jan 19th</td>
<td>Jim Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>4</td>
<td>Jan 26th</td>
<td>Charlie Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision using Machine Learning</td>
</tr>
<tr>
<td>5</td>
<td>Feb 2nd</td>
<td>Alfred Kobsa</td>
<td>Informatics</td>
<td>Data Privacy and Personalization</td>
</tr>
<tr>
<td>6</td>
<td>Feb 9th</td>
<td>Carter Butts</td>
<td>Social Sciences</td>
<td>Social Network Data Analysis</td>
</tr>
<tr>
<td>7</td>
<td>Feb 16th</td>
<td>Michael Carey</td>
<td>Computer Science</td>
<td>Systems for Big Data</td>
</tr>
<tr>
<td>8</td>
<td>Feb 23rd</td>
<td>Hal Stern</td>
<td>Statistics</td>
<td>Bayesian Statistics</td>
</tr>
<tr>
<td>9</td>
<td>Mar 1st</td>
<td>Stacey Hancock</td>
<td>Statistics</td>
<td>Randomization Methods and the R Language</td>
</tr>
<tr>
<td>10</td>
<td>Mar 8th</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>The Future of Data Science</td>
</tr>
</tbody>
</table>