Outline

• Class organization and topics

• Data science and real-world applications

• Examples of data science algorithms

• The UCI Data Science Major
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 submit answers to some brief questions for 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](http://www.ics.uci.edu/~smyth/courses/stats5)
  - Slides and related materials will be posted during the quarter
## Schedule of Speakers

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 10</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>Jan 17</td>
<td>Sameer Singh</td>
<td>Computer Science</td>
<td>Machine Learning for Text</td>
</tr>
<tr>
<td>Jan 24</td>
<td>Charless Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision using Machine Learning</td>
</tr>
<tr>
<td>Jan 31</td>
<td>Pierre Baldi</td>
<td>Computer Science</td>
<td>Deep Learning and Neural Networks</td>
</tr>
<tr>
<td>Feb 7</td>
<td>Hernando Ombao</td>
<td>Statistics</td>
<td>Brain Signal Analysis: Statistics, Computing and Visualization</td>
</tr>
<tr>
<td>Feb 14</td>
<td>Zhoaxia Yu</td>
<td>Statistics</td>
<td>Multivariate Data Analysis</td>
</tr>
<tr>
<td>Feb 21</td>
<td>Ramesh Jain</td>
<td>Computer Science</td>
<td>Event and Web Data</td>
</tr>
<tr>
<td>Feb 28</td>
<td>James Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>Mar 7</td>
<td>Matthew Harding</td>
<td>Economics</td>
<td>Big Data and Economics</td>
</tr>
<tr>
<td>Mar 14</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>The Future of Data Science</td>
</tr>
</tbody>
</table>
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
60 Terabytes/day
20 Petabytes/year

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

Scientific Data: Large Hadron Collider at CERN
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 not nearly enough statisticians
  – And even statisticians need computers to analyze complex data
Web Search: How do search engines like Google or Bing rank search results?
Social media: How does Facebook recognize people in images?

From Le Cun and Ranzato 2013
Shopping: How does Amazon forecast how many items it needs to store in its warehouses?

From dailymail.co.uk

From www.formaspace.com

From linkedin.com
Climate: How can NASA automatically detect land changes using satellite image data?

From www.spot-7.com

From http://cimss.ssec.wisc.edu/
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: How do you write software to search for new physics particles?

Large Hadron Collider:
700 Mbytes/second
60 Terabytes/day
20 Petabytes/year
Medicine: How can researchers use genomics to make personalized medical recommendations?

Data Matrix:
Rows = genes
Columns = patients

From www.originlab.com
Politics: How can we reliably predict events like elections?

Sports: How can we visualize and understand massive amounts of game sensor data?

All of these applications use Data Science......

....these applications are built on combinations of ideas from

- Database systems
- Algorithms
- Machine learning
- Probabilistic models
- Statistical forecasting
- Data visualization
- and more...
What is Data Science?

Statistics (Mathematical and Probabilistic Foundations)

Computing (Algorithms and Software)

Data Science

Applications (Analyzing Real Data)
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
Examples of Data Science Algorithms and Applications
Recommender Systems

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Items

Users
### Recommender Systems

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

<table>
<thead>
<tr>
<th>Users</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>⬝ 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Items</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬝ 3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* e.g., for Web sites that recommend books, movies, etc
Matrix Factorization of Ratings Data

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

- m users
- n movies

"user weights" $\sim$

- m users
- f

$\times$

- n movies
- "movie weights"
Figure from Koren, Bell, Volinksy, IEEE Computer, 2009
Matrix Factorization of Ratings Data

Rating of unseen movie = user weights * movie weights
The $1 Million Question
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.
Good work, but I think we might need just a little more detail right here.
Figure from Krizhevsky, Sutskever, Hinton, 2012
Figure from Krizhevsky, Sutskever, Hinton, 2012
ImageNet Error Rate

<table>
<thead>
<tr>
<th>Year</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>28.2</td>
</tr>
<tr>
<td>2011</td>
<td>25.8</td>
</tr>
<tr>
<td>2012</td>
<td>16.4</td>
</tr>
<tr>
<td>2013</td>
<td>11.1</td>
</tr>
<tr>
<td>2014</td>
<td>6.6</td>
</tr>
<tr>
<td>2015</td>
<td>3.57</td>
</tr>
</tbody>
</table>

Human Error Rate: 5.1%

Figure from Kevin Murphy, Google, 2016
Automated Face Detection

From http://cs.brown.edu/courses/cs143/2011/proj4/images/class_photo_detections.jpg

Automated Vehicle Detection

From https://i.ytimg.com/vi/xVwsr9p3irA/maxresdefault.jpg
The Good

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

Figure courtesy of Kevin Murphy, Google, 2016
The Not So Good

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

Figure courtesy of Kevin Murphy, Google, 2016
The UCI Data Science Major
Department of Statistics

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.
Is the Data Science Major a good match for you?

• Are you interested in computing?
  – Enjoy working with algorithms, programming, machine learning,…

• Do you have a good mathematics background?
  – Comfortable with mathematical ideas and concepts?
  – Interested in applying mathematical ideas to real-world problems?

• Enthusiastic about analyzing data?
  – Enjoy working with data? exploring, visualizing, modeling, understanding

• Seeking a career that has broad and flexible options?

If your answers are YES, the DS Major is for you!
What Classes will I take in the DS Major? ...here are examples

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>Course</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICS 31</td>
<td>12</td>
<td>ICS 32</td>
<td>ICS 33</td>
</tr>
<tr>
<td>Math 2A</td>
<td>4</td>
<td>Math 2B</td>
<td>Math 2D</td>
</tr>
<tr>
<td>Writing 39A</td>
<td>4</td>
<td>Writing 39B</td>
<td>Stats 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stats 5</td>
<td>Writing 39C</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Course</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICS 6B</td>
<td>16</td>
<td>ICS 45C</td>
<td>Stats 68</td>
</tr>
<tr>
<td>Math 3A</td>
<td>4</td>
<td>ICS 51</td>
<td>Stats 120C</td>
</tr>
<tr>
<td>Stats 120A</td>
<td>4</td>
<td>Stats 120B</td>
<td>ICS 46</td>
</tr>
<tr>
<td>GE III</td>
<td>4</td>
<td></td>
<td>ICS 6D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P. Smyth: Stats 5: Data Science Seminar, Winter 2017: 48
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>
</table>
| Stats 110, Statistical Methods for Data Analysis I  
CS 161, Design and Analysis of Algorithms  
In4matx 43, Introduction to Software Engineering  
GE IV/VIII, | Stats 111, Statistical Methods for Data Analysis II  
CS 178, Machine Learning and Data-Mining  
ICS 139W, Critical Writing on Information Technology  
GE III/VII, | Stats 112, Statistical Methods for Data Analysis III  
CS 122A, Introduction to Data Management  
In4matx 143, Information Visualization  
GE VI, |

**Year 4: two-quarter capstone “data-intensive” project, + statistics and CS electives**
Change of Major Requirements for Data Science Major

- 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 Professor Zhaoxia Yu at zhaoxia@ics.uci.edu to inquire about getting a waiver to change into the major.
What can I do with a Data Science Major?

- **Careers in “Data-Oriented” Companies and Organizations**
  - Computing/internet companies: Google, Amazon, Facebook, IBM,….
  - Engineering companies: Intel, Samsung, Boeing, ….
  - Finance/insurance companies
  - Medical/pharmaceutical companies
  - Government/national labs: NASA, NIST, DoD, ….
  - Many many more…….

- **Option to specialize with a Graduate Degrees (MS or PhD)**
  - Computer Science: specialize in a topic such as machine learning, databases, etc
  - Statistics: specialize in a statistical topic, e.g., computational statistics
  - MS/PhD degrees lead to a wide variety of careers
Jobs for Data Scientists?

From indeed.com
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...
Data Scientist
6sense - San Francisco, CA
Our data scientists are not optimizing software; We work with big data at scale, advanced machine learning and predictive modeling to find buyers and predict...

Easily apply
3 days ago - save job - email - more...

GBS Entry Level Data Scientist Analytics Co-Op
IBM ★★★★★ 6,979 reviews - Columbus, OH 43228 (Westland area)
Growth Play Analytics (BigData). You must reside within a reasonable commuting distance - generally 50 miles or less from Columbus, Ohio OR be available to work...
4 hours ago - save job - email - more...

Data Scientist
General Motors ★★★★★ 1,276 reviews - Austin, TX
Expert level knowledge, development expertise & commiserate experience with Python, Matlab, R, Java and SQL. Data Engineering Team....
11 days ago - email

Sponsored

Data Scientist
National Security Agency ★★★★★ 63 reviews - Fort George G Meade, MD
$64,923 - $83,774 a year
Applied statistics, calculus, quantitative or statistical methods and techniques, data mining, informatics, data science, programming, computational algorithms,...
7 days ago - email

Sponsored

10 hours ago - save job - email - more...
Do I need a Data Science degree to do Data Science?

• Technically no......many people currently are “data scientists” with backgrounds in quantitative degrees that are not data science
  – Some with statistics, some with computer science, some with a combination
  – Some with other quantitative degrees

• Advantages of the DS major
  – Puts you on the “fast track” to becoming a Data Scientist
  – Ensures that you will know the fundamentals of both
    • Computing
    • Statistics
  – Provides you with skills that are likely to have lasting value (as technology changes)
What are other degree options?

- **Computer Science with a Statistics minor?**
  - More classes in “systems” aspects of computer science
  - Fewer classes in statistics
  - No capstone data science project class

- **Another degree like Math or Economics with a Statistics minor?**
  - Far fewer classes in computer science
  - Fewer classes in statistics
  - No capstone data science project class

- **Statistics undergraduate degree (e.g., at another UC)?**
  - More classes in mathematics and statistics
  - Far fewer classes in computer science
  - No capstone data science project class
# Schedule of Speakers (from class Website)

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Department</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 10</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>Introduction to Data Science</td>
</tr>
<tr>
<td>Jan 17</td>
<td>Sameer Singh</td>
<td>Computer Science</td>
<td>Machine Learning for Text</td>
</tr>
<tr>
<td>Jan 24</td>
<td>Charless Fowlkes</td>
<td>Computer Science</td>
<td>Computer Vision using Machine Learning</td>
</tr>
<tr>
<td>Jan 31</td>
<td>Pierre Baldi</td>
<td>Computer Science</td>
<td>Deep Learning and Neural Networks</td>
</tr>
<tr>
<td>Feb 7</td>
<td>Hernando Ombao</td>
<td>Statistics</td>
<td>Brain Signal Analysis: Statistics, Computing and Visualization</td>
</tr>
<tr>
<td>Feb 14</td>
<td>Zhoaxia Yu</td>
<td>Statistics</td>
<td>Multivariate Data Analysis</td>
</tr>
<tr>
<td>Feb 21</td>
<td>Ramesh Jain</td>
<td>Computer Science</td>
<td>Event and Web Data</td>
</tr>
<tr>
<td>Feb 28</td>
<td>James Randerson</td>
<td>Earth Systems Science</td>
<td>Data-Driven Climate Science</td>
</tr>
<tr>
<td>Mar 7</td>
<td>Matthew Harding</td>
<td>Economics</td>
<td>Big Data and Economics</td>
</tr>
<tr>
<td>Mar 14</td>
<td>Padhraic Smyth</td>
<td>Computer Science</td>
<td>The Future of Data Science</td>
</tr>
</tbody>
</table>
BACKUP SLIDES
Research at UC Irvine in Data Science

[Image: Photos of researchers and data visualizations related to data science research.]

[Image: UCI Data Science Initiative logo and website banner for the Machine Learning Repository.]

[Image: Center for Machine Learning and Intelligent Systems logo.]

Welcome to the UC Irvine Machine Learning Repository!
We currently maintain 350 data sets as a service to the machine learning community. You may view all data sets through our searchable interface old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please refer consult our donation policy. For any other questions, feel free to contact the Repository Manager. We have also set up a mirror site for the Repository.

Supported By: [Logo]
In Collaboration With: [Logo]

P. Smyth: Stats 5: Data Science Seminar, Winter 2017: 59
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 those 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