INTRODUCTION TO DATA SCIENCE

Introduction and Administration
Plan

Why data science is important?
- “Why are you here”

What is data science?
- Mashup of disciplines

What this course is about?
- Hopefully right mix of theory and practical skills

Course requirements
- Syllabus
- Grade, exam, homework assignments
- Homepage, contact details
1. Why are you here?

Introduction: Media Buzz
Data Scientists are in high demand

Data Scientist: The Sexiest Job of the 21st Century
by Thomas H. Davenport and D.J. Patil

Why your kids will want to be data scientists
John Phillips | @J_Phillips_IV
Tuesday, 3 June 2014 | 7:05 PM ET

Big data skills: Should data scientist be your next job?
Also in Academia

WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A $37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS

BY NEAL UNGERLEIDER

It’s official: America needs more data scientists. This week, a $37.8 million project
Demand will outpace the supply

Over 2/3 believe demand for talent will outpace the supply of data scientists

**Over the next five years, demand for data scientists will:**

- Be significantly less than the talent available: 1%
- Be less than the talent available: 5%
- Be met by the available talent: 31%
- Significantly outpace the supply of talent: 31%
- Somewhat outpace the supply of talent: 32%
Israel
Pays well

Big Data, Big Paycheck
Median salary for analytics professionals and those specifically within data science, by level of experience.

- **Up to 3 years**
  - Analytics professionals: $65,000
  - Data scientists: $80,000

- **4 to 8 years**
  - $85,000

- **9+ years**
  - $115,000

Note: Data do not include managers
Source: Burtch Works
The Wall Street Journal
2. What is data science?

Technology and raising expectations
Data Science

- New Discipline

- Very little/none textbooks/courses covering the discipline as a whole
  - Compare to Software Engineering/Computer Science during 70-80’s of the last century
  - Data Science is what data scientists do

- Why data science and data scientists are needed?
  - Development of enabling technology
  - Raising Expectations from customers
2. What is data science?

Technological developments
Declining cost of storage

STORAGE: FROM HIGHWAY ROBBERY TO RUNAWAY BARGAIN

$ per megabyte

1983: Iomega appears on the storage scene with its Bernoulli Box, the box itself costing over $2,000 and each 10MB removable disk priced at $89.
1995: Seagate introduces its 1GB drive for $849.
2007: Hitachi introduces the first terabyte drive, the Deskstar 7K1000, $399.

*Projected. No data is available for 1986.

Sources: Ars Technica, Little Tech Shoppe, Steve Gilheany, ExtremeTech
Declining cost of computing

Cost Per Transistor (Logarithmic Plot - 2014 US$)

2010: One transistor costs less than the ink for one letter of newsprint.

18 month halving

© 2014 Millennia Media, FPC
Surpassing the brain
More data can be stored and processed.
Value of Big Data

- Retail: 49% increase, $9.6B
- Consulting: 39% increase, $5.0B
- Air transportation: 21% increase, $4.3B
- Construction: 20% increase, $4.2B
- Food products: 20% increase, $3.4B
- Steel: 20% increase, $3.4B
- Automobile: 19% increase, $2.0B
- Industrial instruments: 18% increase, $1.2B
- Publishing: 18% increase, $0.8B
- Telecommunications: 17% increase, $0.4B

*Source: University of Texas (2011)*
Devices vs. People

Growth in Internet-Connected Devices by 2020

- Orange: World population (in billions)
- Blue: Internet-connected devices (in billions)
- Light blue: Internet-connected devices per person

Source: Cisco IBSG, April 2011
Internet of Things

Source: Ericsson AB, "Infrastructure Innovation - Can the Challenge be met?," Sept 2010
Next frontier: IoT

Information from the Internet of Things:
We have gone beyond the decimal system

Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)

- **Yottabyte**: This is our digital universe today = 250 trillion of DVDs
- **Zettabyte**: 1.3 ZB of network traffic by 2016
- **Exabyte**: The CERN Large Hadron Collider generates 1 PB per second
- **Petabyte**: 500TB of new data per day are ingested in Facebook databases
- **Terabyte**: 1 EB of data is created on the internet each day = 250 million DVDs worth of information.
- The proposed Square Kilometer Array telescope will generate an EB of data per day
2. What is data science?

Raising expectations
Cognitive Computing

- People expect systems to behave like humans
  - Be Adaptive
    - Learn as information and goals change
  - Be Interactive
    - Interact easily with people and other systems
  - Be Contextual
    - Understand meaning, exploit additional sources of information

- Need to process large quantities of uncertain data of different types (text, speech, sensors, images etc.)
Cognitive Computing
IBM: Computers Will See, Hear, Taste, Smell and Touch in 5 Years
People want their systems/devices to behave smarter
- Personal devices
- Industrial systems

More data to acquire and analyze using more complex algorithms and technologies
3. What is data science

Some examples
Example I: Marketing

- Predicting Lifetime Value (LTV)
  - **what for:** if you can predict the characteristics of high LTV customers, this supports customer segmentation, identifies upsell opportunities and supports other marketing initiatives

- **usage:** can be both an online algorithm and a static report showing the characteristics of high LTV customers
Example II: Logistics

- Demand forecasting
  - How many of what thing do you need and where will we need them? (Enables lean inventory and prevents out of stock situations.)
  - revenue impact: supports growth and militates against revenue leakage
  - usage: online algorithm and static report
Example III: Healthcare

- **Survival analysis**
  - Analyze survival statistics for different patient attributes (age, blood type, gender, etc) and treatments

- **Medication (dosage) effectiveness**
  - Analyze effects of admitting different types and dosage of medication for a disease

- **Readmission risk**
  - Predict risk of re-admittance based on patient attributes, medical history, diagnose & treatment
Example IV: Wearable Health and Fitness
Example V: Brain Computer Interface
2. What is data science?

A Mashup of disciplines
# A mashup of disciplines

<table>
<thead>
<tr>
<th>Category</th>
<th>Disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math and Theory</td>
<td>Statistics, Linear Algebra, Optimization, Time Series, etc.</td>
</tr>
<tr>
<td>Applied Algorithms</td>
<td>Machine Learning, Data Structures, Parallel Algorithms, etc.</td>
</tr>
<tr>
<td>Engineering and Technologies</td>
<td>Storage and computing platforms, statistical tools, etc.</td>
</tr>
<tr>
<td>Domain Expertise</td>
<td>Text, Finance, Images, Econometrics etc.</td>
</tr>
<tr>
<td>Art</td>
<td>Visualization, Infographics</td>
</tr>
<tr>
<td>Best practices and hacks</td>
<td>Handle missed values in data, transform and represent data, etc.</td>
</tr>
</tbody>
</table>
Yet Another View

Data Science Venn Diagram v2.0

- Data Science
- Computer Science
- Machine Learning
- Math and Statistics
- Unicorn
- Traditional Software
- Traditional Research
- Subject Matter Expertise

Copyright © 2014 by Steven Geringer Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact.
Types of Data Scientists
Roles and Paycheck

Salary Trend – Data Scientist Experts

- Data Scientist: $111,000
- Data Analyst: $70,000
- Data Engineer: $99,000
- Data Developer: $95,000
- Other IT Jobs: $61,000
3. About this course

A mix of theory and practice
General

- Introductory course
  - But for advanced undergrads

- Broad overview of subjects
  - But deep enough to have an exam

- Focus on practical aspects
  - But not on ever-changing technology and tools
Tentative content (subject to change)

- 70% Statistical Machine Learning (7 weeks)
  - Focus on practical aspects
  - Classes
    - Necessary theoretical background
    - Basic R programming lab

- 20% Big Data Algorithms (2 weeks)
  - Focus on algorithms not on big data technologies

- 10% Data Visualization (1 weeks)
  - Grammar of graphics in R
This course is not

- About big data tools or technologies
  - No: Hadoop technical details
  - Yes: Basic R programming

- About statistical learning theory
  - No: Theoretical low bounds or other proofs
  - Yes: Some theory is necessary

- About a specific domain
  - No: Deep discussions on Text, Finance, BI etc.
  - Yes: Some examples will be presented
Some case studies we will cover

<table>
<thead>
<tr>
<th>Area</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction of future movements in the stock market</td>
<td>• What is the next move of S&amp;P 500?</td>
</tr>
<tr>
<td>Predicting insurance purchase</td>
<td>• Will a potential customer purchase?</td>
</tr>
<tr>
<td>Direct marketing</td>
<td>• Who will respond?</td>
</tr>
<tr>
<td>Housing valuations</td>
<td>• What affect the price of a house?</td>
</tr>
<tr>
<td>Marketing of orange juice</td>
<td>• What brand a customer will buy?</td>
</tr>
<tr>
<td>Email spam</td>
<td>• Is this a spam message?</td>
</tr>
</tbody>
</table>
The course’s language of choice: R

<table>
<thead>
<tr>
<th>Skill</th>
<th>2013</th>
<th>YR/YR CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>$115,531</td>
<td>n/a</td>
</tr>
<tr>
<td>NoSQL</td>
<td>$114,796</td>
<td>1.6%</td>
</tr>
<tr>
<td>MapReduce</td>
<td>$114,396</td>
<td>n/a</td>
</tr>
<tr>
<td>PMBok</td>
<td>$112,382</td>
<td>1.3%</td>
</tr>
<tr>
<td>Cassandra</td>
<td>$112,382</td>
<td>n/a</td>
</tr>
<tr>
<td>Omnigraffle</td>
<td>$111,039</td>
<td>0.3%</td>
</tr>
<tr>
<td>Pig</td>
<td>$109,561</td>
<td>n/a</td>
</tr>
<tr>
<td>SOA (Service Oriented Architecture)</td>
<td>$108,997</td>
<td>-0.5%</td>
</tr>
<tr>
<td>Hadoop</td>
<td>$108,669</td>
<td>-5.6%</td>
</tr>
<tr>
<td>Mongo DB</td>
<td>$107,825</td>
<td>-0.4%</td>
</tr>
</tbody>
</table>
What you are expected to know

- Data is represented as a matrix
  - Basic linear algebra

- Most problems are not well-defined/uncertain
  - Basic probability and statistics

- Big data requires non-trivial data structures and algorithms
  - Basic data structures and algorithms concepts

- Practical means programming
  - Basic Programming skills
Textbooks are available online

**Machine Learning and R**

*An Introduction to Statistical Learning*
with Applications in R

**Big Data Algorithms**

*Mining of Massive Datasets*
Visualization

Introduction from

Hadley Wickham

ggplot2

Elegant Graphics for Data Analysis

On-going examples
For curious minds

More on Machine Learning

The Elements of Statistical Learning
Data Mining, Inference, and Prediction

Second Edition

More on R Programming

THE ART OF R PROGRAMMING
A TOUR OF STATISTICAL SOFTWARE DESIGN

NORMAN MATLOFF
Becoming a data scientist

Data Scientist Skills

Practical Data Science with R
Nina Zumel, John Mount

Quick Hacks/Examples

Machine Learning for Hackers
Drew Conway & John Myles White
4. Course requirements
Requirements

- **Grade**
  - 100% closed material exam

- **No previous year exams**
  - Both textbooks have after chapter exercises
  - Exam questions (and HW assignments) will be very similar to these questions

- See course homepage for HW submission guidelines
Contacts

- **Lecturer:** Dr. Sasha Apartsin ([apartisn@gmail.com](mailto:apartisn@gmail.com))

- **Course homepage:**

- **Office hours:** By appointment

- **Course forum:**
  - [groups.google.com/d/forum/tau-data-science-course-2015s](https://groups.google.com/d/forum/tau-data-science-course-2015s)
## Plan

<table>
<thead>
<tr>
<th>Topic</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why data science is important?</td>
<td>“Why are you here”</td>
</tr>
<tr>
<td>What is data science?</td>
<td>Mashup of disciplines</td>
</tr>
<tr>
<td>What this course is about?</td>
<td>Hopefully right mix of theory and practical skills</td>
</tr>
<tr>
<td>Course requirements</td>
<td>Syllabus, Grade, exam, homework assignments, Homepage, contact details</td>
</tr>
</tbody>
</table>
Few More Disclaimers
Very inaccurate explanation

- **Statistics**: take a sample (data), answer questions about the **process** that produced this sample
  - Is it a normal distribution? **Estimate** its mean.

- **Machine Learning**: take a sample (data), build a model to answer questions about **future samples**
  - Given a sample of named faces, **design a model** for naming a new unseen face.

- **Data Mining**: mine huge data store for interesting patterns or relationships
  - Given DB of transactions, **apply** tools and algorithms to find frequent product bundles

- **Data Science**: do whatever necessary to **extract value** from the data
  - Use data to improve book sales: mine patterns, engineer recommender systems, suggest improvements, estimate impact

---

**No clear-cut boundaries!**
Disclaimer: Math in the course

- All the computation are performed by computer

- You are in charge for interpretation of numbers

- So you’ll have to understand the logic behind the number
  - You’ll see significant amount formulas during the course
  - Mostly arithmetic, matrices and probability

- You are not expected to memorize or derive each formula (with exceptions), but you are expected to
  - Understand its meaning and use