Introduction to Data Science

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MSAN 601 - Linear Regression Analysis

Outline

- Course Overview
- What is Data Science?
  - Where is Data Science?
  - A brief history

A Little About Me

- Ph.D. Statistics and Operations Research (UNC Chapel Hill, ’15)
  - Research focused on statistical analysis of networks
  - Explore, model, and analyze network data (e.g., social networks)
- M.S. Mathematical Sciences (Clemson University, ’10)
- B.S. Mathematics and Chemistry (Campbell University ’08)

A Little About Me

Classes I teach:

- BSDS 100 - Intro to Data Science with R
- MATH 106 - Business Statistics
- MATH 370 - Probability with Applications
- MATH 373 - Statistical Learning
- MSAN 601 - Linear Regression Analysis
- MSAN 630 - Advanced Computational Statistics
- MSAN 700 - Social Network Analysis

A Little About Me

- Born and raised in NC (near Raleigh)
- Live in Rockridge, Berkeley.
- A huge college basketball fan! (Go Heels!)
- Have loved college football since 2008 (Go Tigers!)
- Enjoy tasting beers (bourbon-barrel stouts are my favorite).

Course Description and Syllabus

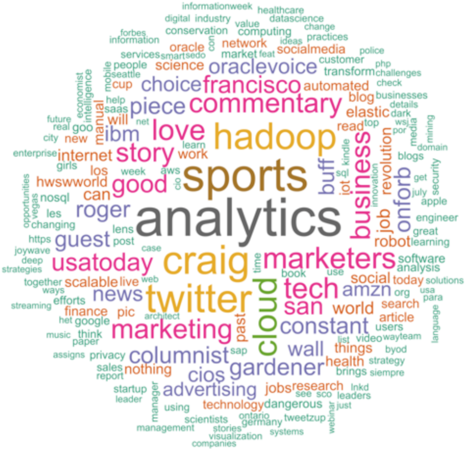
All lecture notes, the syllabus, assignments, and course description are available at this course website:

https://github.com/jdwilson4/Regression-Analysis

What is Data Science?

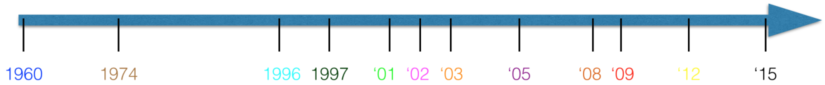
- **Wikipedia**: “the extraction of knowledge from data."
- A precise definition is a bit unclear and has faced much controversy... (we’ll see more on this in a moment)
- Practitioners tend to agree on the *components* of data science:
  - gathering and cleaning data
  - database management
  - exploratory analysis
  - predictive modeling
  - data summary and visualization

Where is Data Science?



- Twitter feed, December 2014

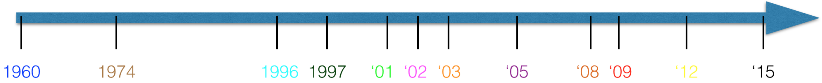
The Evolution of Data Science



- 1960: Peter Naur (CS Ph.D.) published *Datalogy: the science of data and its place in education.*

- 1974: Peter Naur published *Concise Survey of Computer Methods*.
  - defines data science as "the science of dealing with data, once they have been established."
  - continues to say that "... the relation of the data to what they represent is delegated to other fields and sciences."

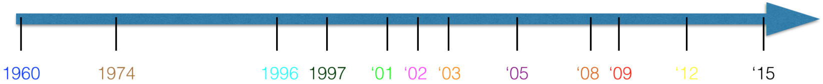
The Evolution of Data Science



1960 1974 1996 1997 '01 '02 '03 '05 '08 '09 '12 '15

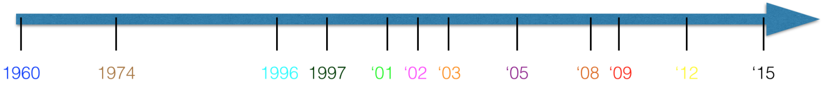
- 1996: International Federation of Classification Societies meet in Tokyo and for the first time include “data science" in the conference title: “Data science, classification, and related methods."
- 1997: C.F. Jeff Wu gave the inaugural lecture “Statistics = Data Science?" for appointment to the H. C. Carver Professorship at the University of Michigan.

The Evolution of Data Science



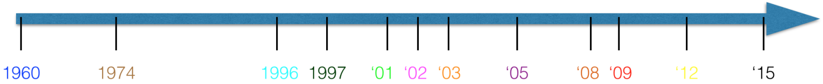
- 2001: William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*.
  - Sets forth 6 areas for a university department involving statistics.
- 2002: *Data Science Journal* is launched
  - Focus on data systems, publications on internet, and applications
- 2003: *Journal of Data Science* is launched
  - Focus on application of statistical and quantitative methods

The Evolution of Data Science



- 2005: National Science board redefines data scientists:
  - “The information and computer scientists, data and software programmers, disciplinary experts, ... who are crucial to successful management of a digital data collection whose primary activity is to conduct creative inquiry and analysis"
- 2008: DJ Patil (LinkedIn) and Jeff Hammerbacher (Facebook) coined the term "data scientist" to define their jobs

The Evolution of Data Science



- January, 2009: Hal Varian (chief economist at Google) writes that "... the sexy job in the next 10 years will be statisticians."
- October, 2012: Harvard Business Review publishes "Data Scientist: The Sexiest Job of the 21st Century."
- February 5th, 2015: DJ Patil appointed as the first Chief Data Scientist in the White House.

Applications

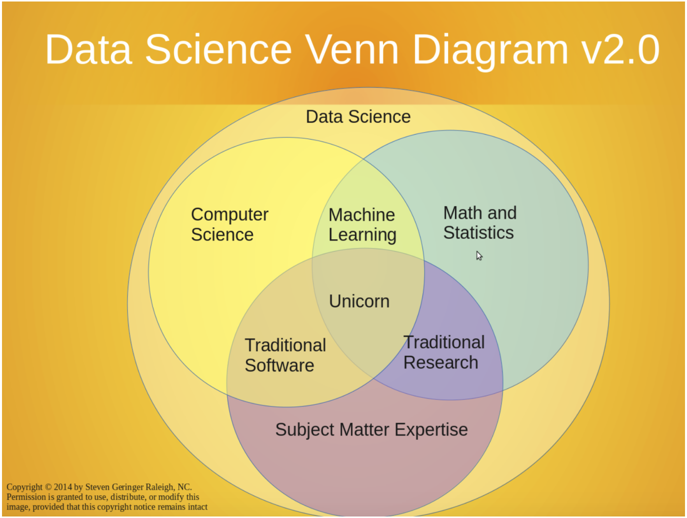
indeed
one search. all jobs.
IBM
Anheuser-Busch
THE WHITE HOUSE
WASHINGTON
Genentech
IN BUSINESS FOR LIFE
BOSTON RED SOX
CAT
SAN ANTONIO
SPURS
f
shutterfly
in

Marketing analytics, sports analytics, biotechnology, social experiments, e-commerce, government analysis, ...

Why Data Science?

- Size, complexity, and amount of data
  - Predicted ≈ 40 trillion gigabytes of data in 2020; up from 130 billion in 2005!
  - Big data requires innovative techniques for analysis
- *McKinsey*: "The U.S. faces a shortage of 140K - 190K people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data." (May, 2011)
- *Harvard Business Review*: "Data Scientist: The Sexiest Job of the 21st Century." (October, 2012)

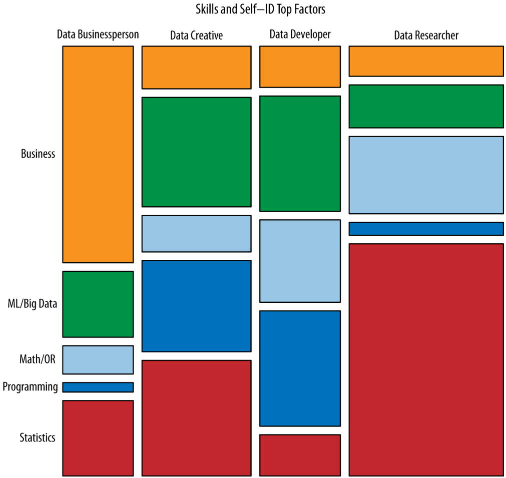
Data Scientists: The unicorn industries want?



Data Scientists: The unicorn industries want?

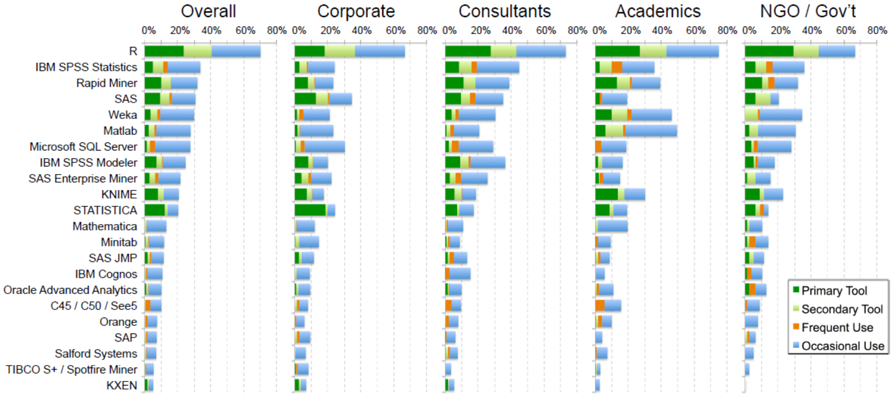
- The field is inherently interdisciplinary
  - mathematical statistics
  - computer science
  - domain expertise
- The magical Unicorn: having all three skills
  - In 2014, these jobs go unfilled for 6 months or longer on average
- Has lead to the development of data science *teams*
  - hope is to merge skills of analysts

The Analysts of Data Science



"Analyzing the Analyzers (2013) by Harry, Murphy, and Vaisman."

Software: A Data Scientist's first weapon



-www.datasciencecentral.com

A Data Scientist's Toolkit

Harvard's data science toolkit:

1. **Wrangle the data**: gather, clean, and sample data
2. **Manage the data**: access big data quickly and reliably
3. **Explore the data**: to make a hypothesis
4. **Make predictions**: statistical methods
5. **Communicate the results**: visualization, presentations, summaries

Get Involved! Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- Coursera.org
  - Free online courses in data science and machine learning
  - 972 courses. Great resource for coding, data analysis, etc.
  - Recent notable course: "The Data Scientist's Toolbox."