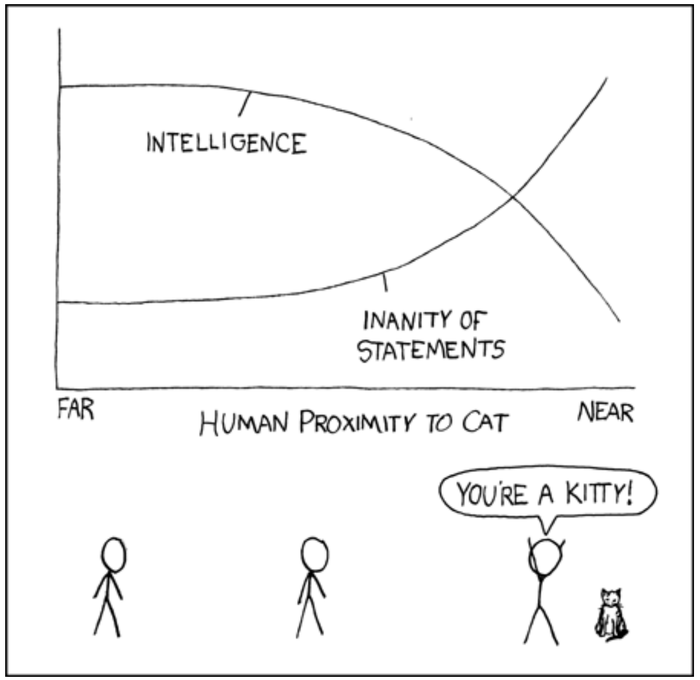
Introduction to Data Science
CS 5360 / Math 4100

Alexander Lex
alex@sci.utah.edu

Braxton Osting
osting@math.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

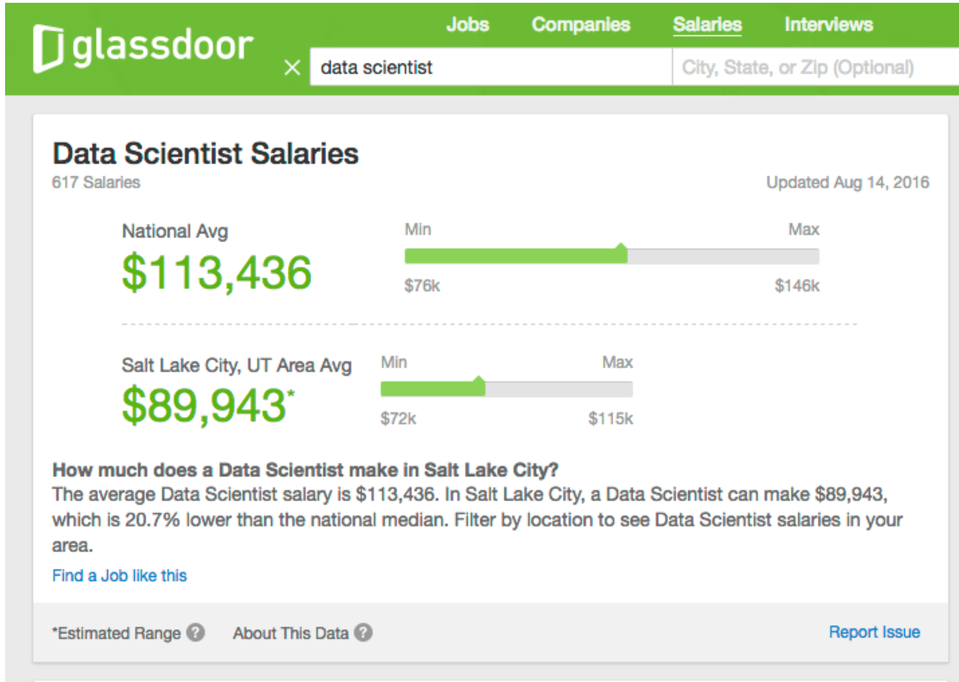
What is Data Science?

The sexiest job of the century —Harvard Buisness Review

A data scientist is a statistician who lives in San Fransisco

Data Science is statistics on a Mac

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

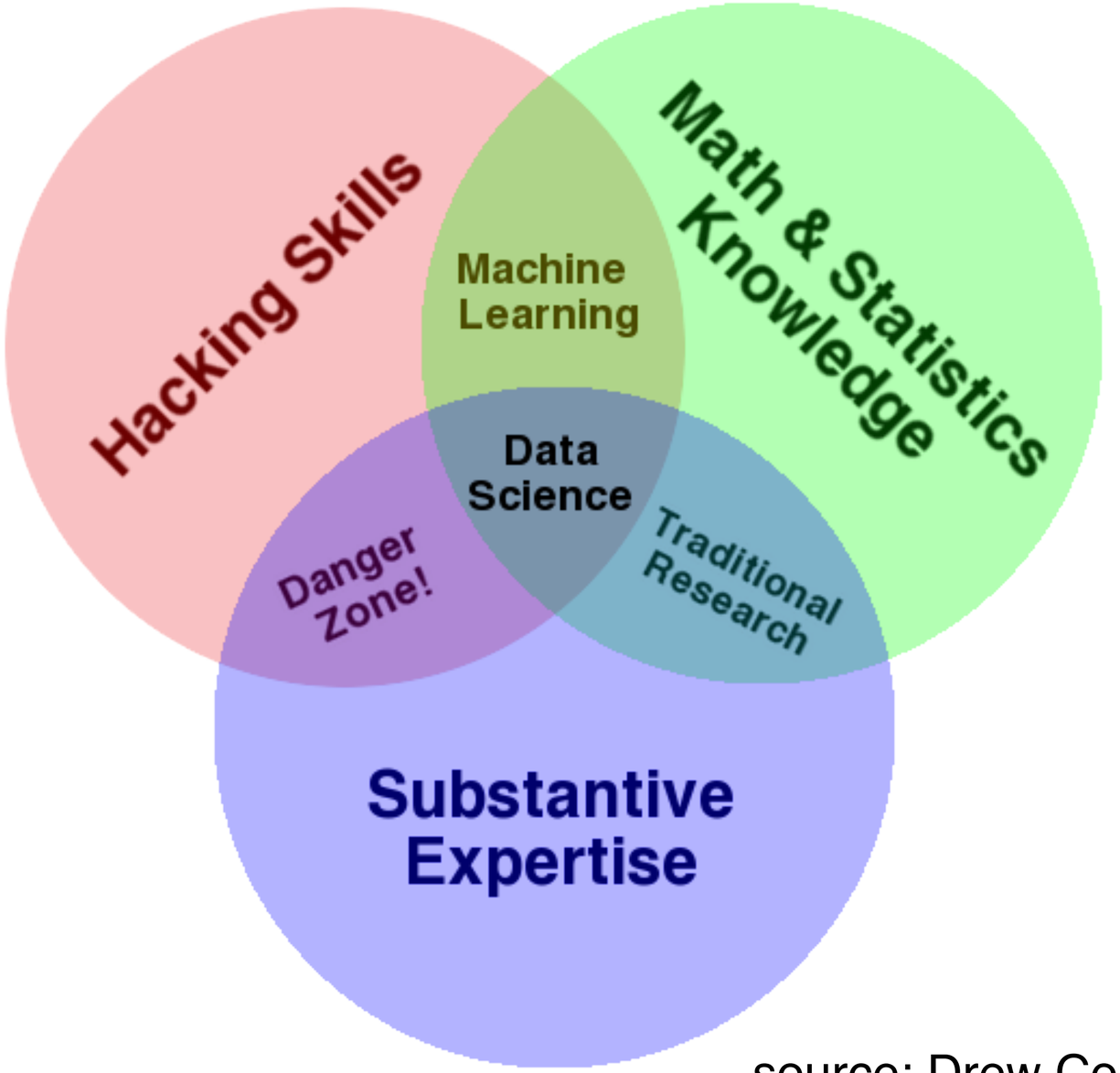


https://twitter.com/jeremyjarvis/status/428848527226437632/photo/1

What is Data Science?

Source: datascience.berkeley.edu

What is Data Science?



source: Drew Conway blog

What is Data Science?

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms. (Wikipedia)

Data Science closes the circle from collecting real-world data, to processing and analyzing it, to influence the real world again.

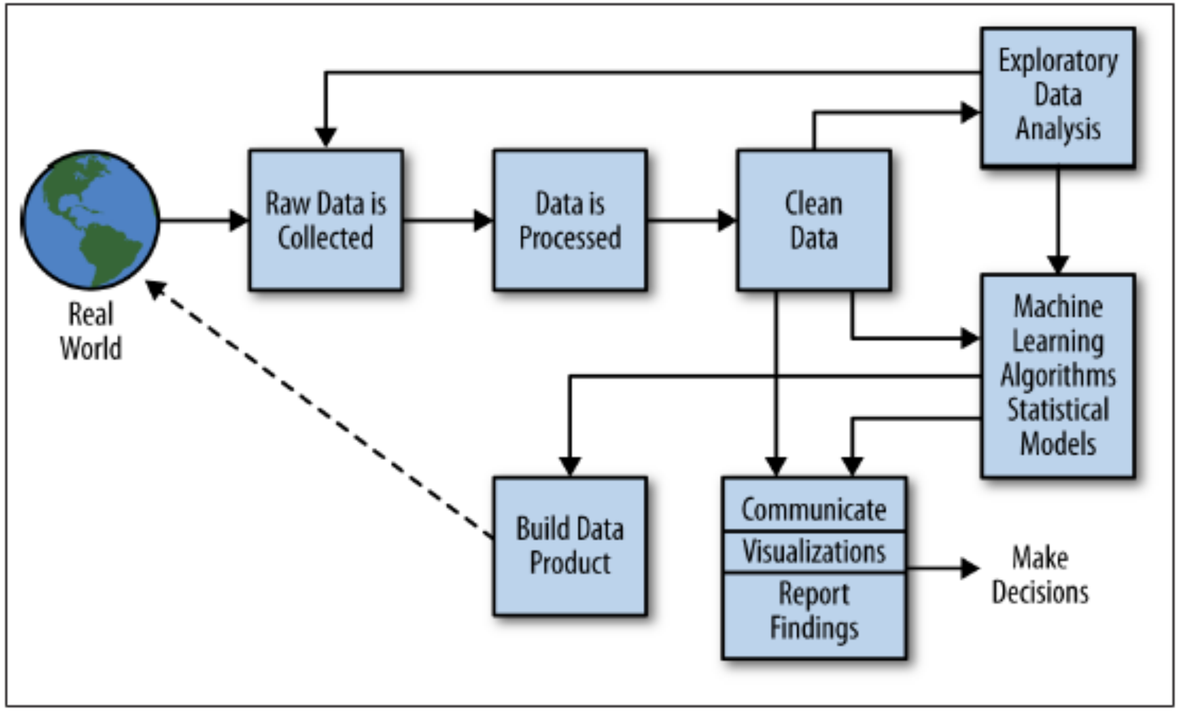


Figure 2-2. The data science process

DDS, p.41

Data Science vs. Machine Learning vs. Statistics ?!?

-> read 50 years of Data Science by David Donoho

What is Data Science?

"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**."

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

Why do we care? It’s everywhere!

Biology? Data-centered & computational!

Physics? Data-centered & computational!

Medicine? Data-centered & computational!

Social Sciences? Data-centered & computational!

Business? Data-centered & computational!

Why do we care? Jobs!

CS enrollments are exploding with both a growing number of majors and a growing non-majors population.

The non-majors are wise in their choices. The recent "Rebooting Jobs" report from Burning Glass and Oracle Academy shows that CS skills are the most rapidly growing skills requested in job ads, but only 18% of those job ads ask for a CS degree.

https://cacm.acm.org/blogs/blog-cacm/224105-learning-computer-science-is-different-than-learning-other-stem-disciplines/fulltext

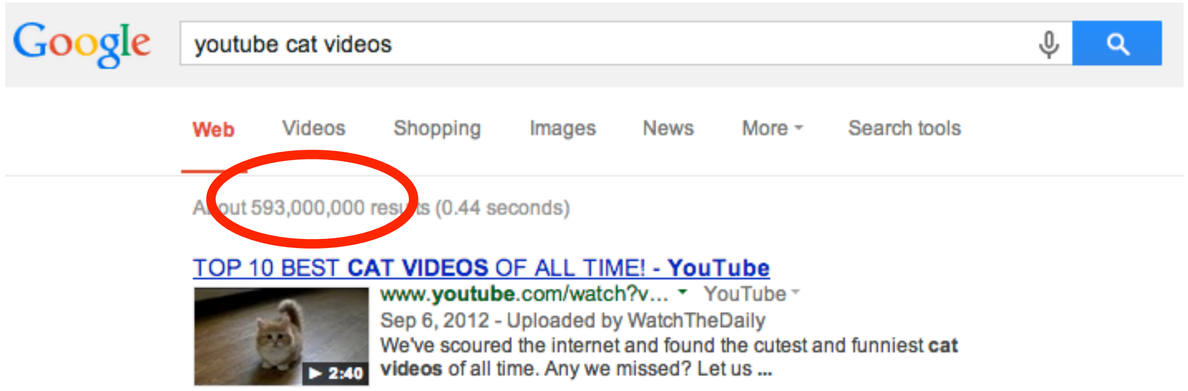
Big Data

2010: 1,200 exabytes, largely unstructured

Google stores ~10 exabytes (2013)

Hard disk industry ships ~8 exabytes/year

2.5 exabytes (2.5 billion gigabytes) generated every day in 2012



15 Exabytes in Punch Cards:
4.5 km over New England

In one second on the Internet there are...

http://onesecond.designly.com/

How can we leverage data?

Improve your fitness by targeted training

Improve your product

- by targeting your audience
- by considering semantics

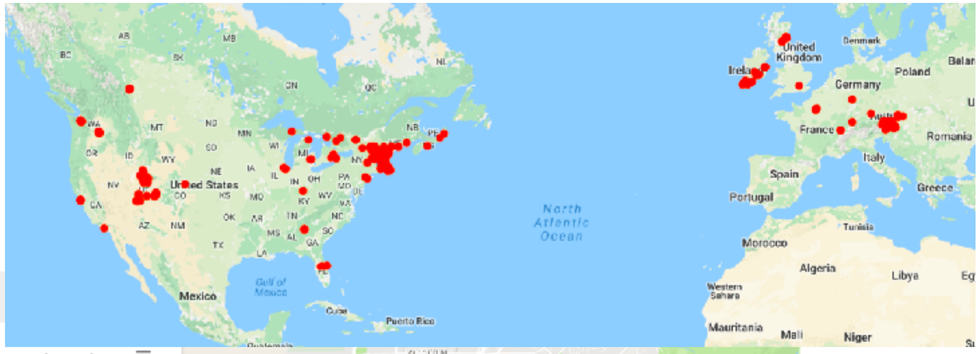
Make better decisions

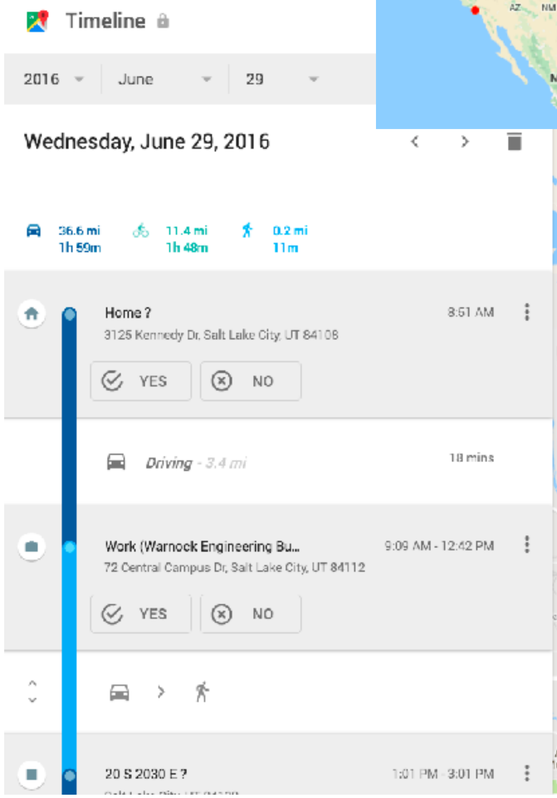
- exact diagnosis, choose right medication, pick good restaurant

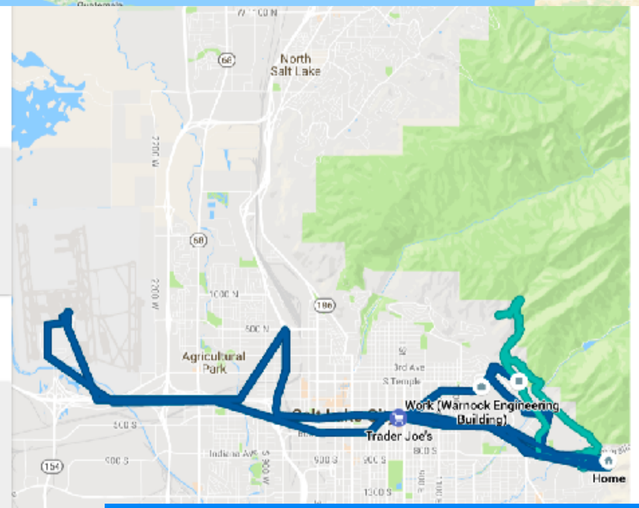
Predict elections, events, crowd behavior, etc.

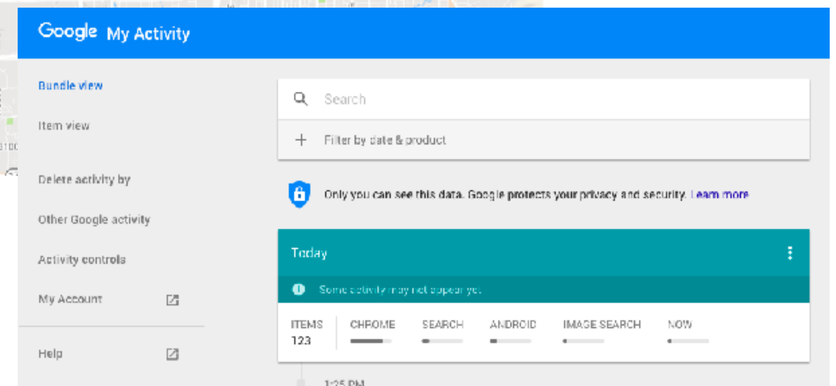
… and many more applications

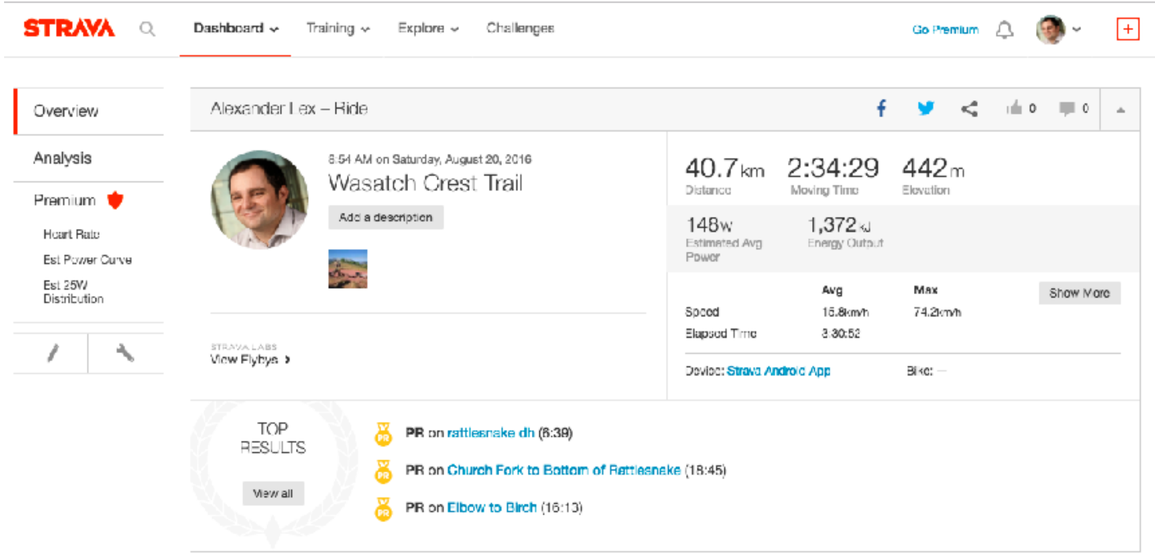
Example: Personal Data











Big Data in Science and Engineering

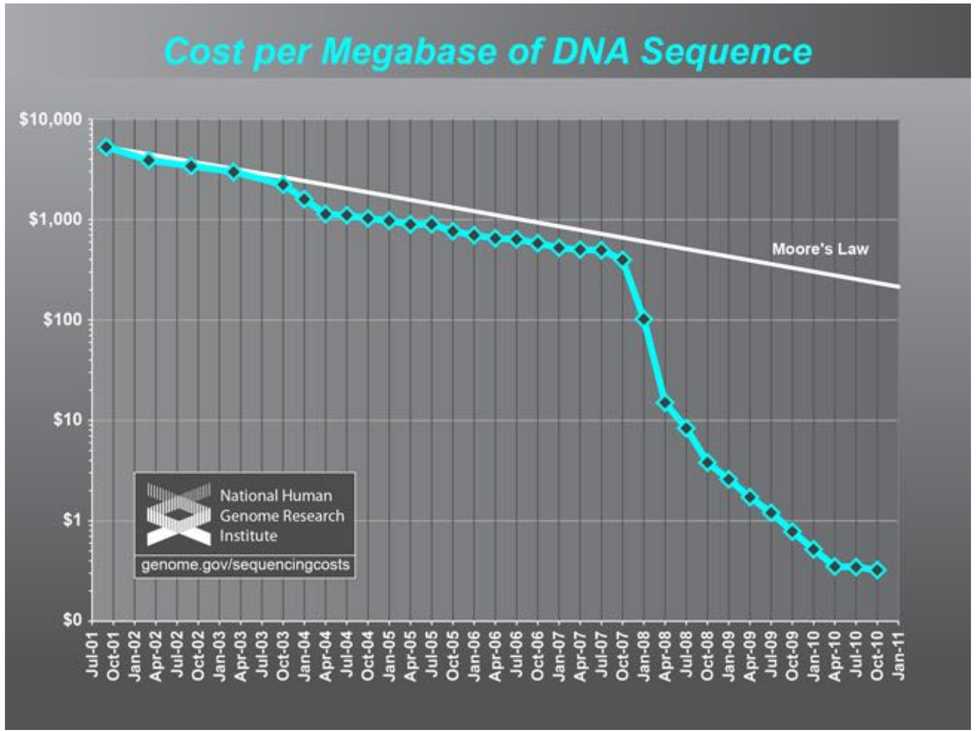
Big Data transformed science and engineering.

Cheap sensors (e.g., imaging) have changed the way science and engineering are done.

Examples:

- Large physics experiments and observations
- Cheaper and automated genome sequencing
- Smart buildings / cities (blyncsy)
- Geophysical imaging

Controversy: Hypothesis or data driven methods



Example: CERN Large Hadron Collider Data

CERN has publicly released over 300TB of data: CERN Open Data Portal

How much is that?

- At 15 GB of storage a piece, you'd need **20,000 Gmail** accounts. As attachments (25 MB), it would take you 12 million emails.
- A DVD-R holds 4.7 GB. You'd need **63,830 DVD-Rs, or 6,000 Blu-ray disks**.
- It takes Pandora about a day and a half to burn through a gig of mobile data. So if the CERN data was an album, you could **stream it in just over 1,230 years**.
- But it ain't no thing compared to what the National Security Agency works with. Going by 2013 figures the agency released, the NSA's various activities "touch" 300 TB of data every 15 minutes or so.

(Popular Mechanics Article)

Example: Genomics

23andMe

All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

Example TCGA: 1 Petabyte

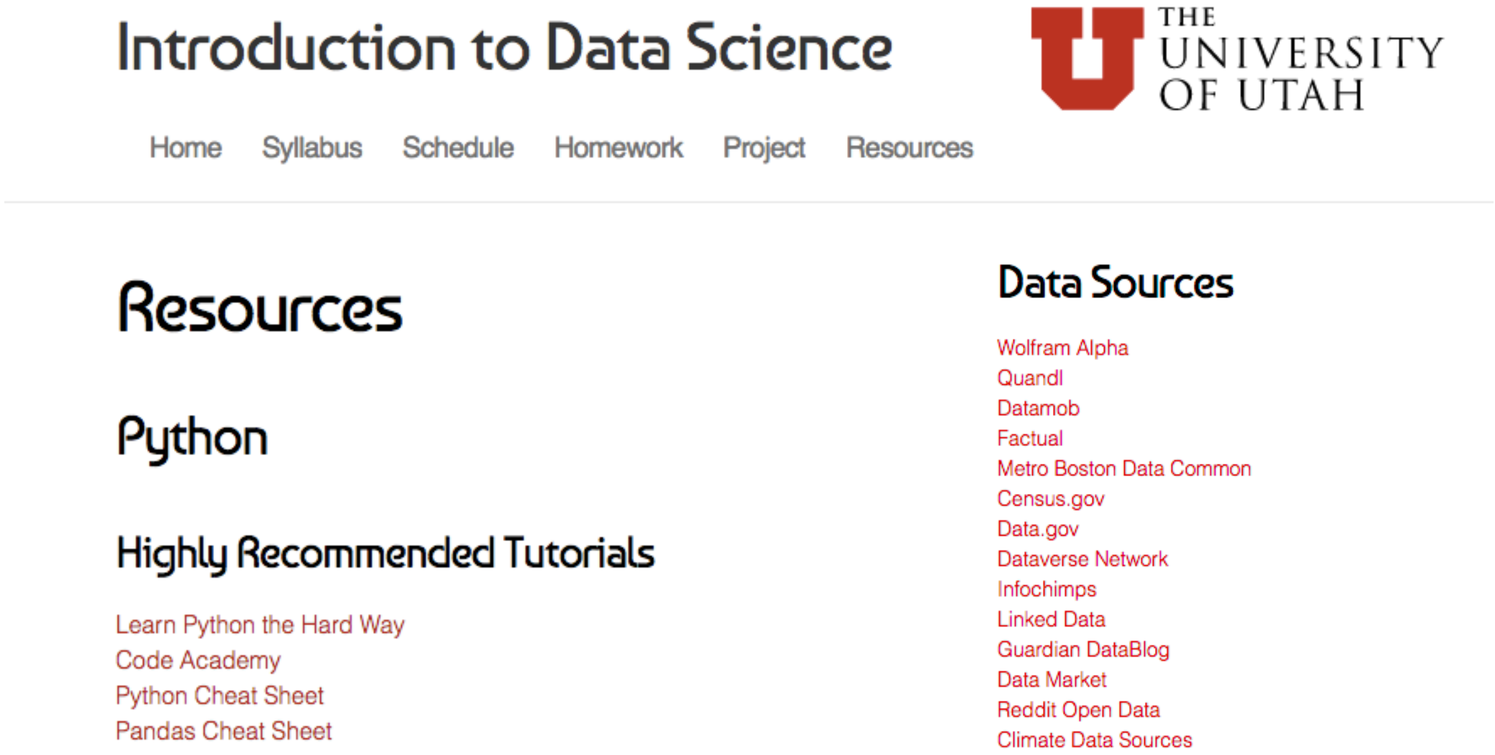
NSA Utah Data Center (Bluffdale, Utah)

Storage Capacity?

estimates vary, but Forbes magazine estimates 12 exabytes (12,000 petabytes or 12 million terabytes)

Where to find data?

Today, a lot of data is publicly available. You probably have access to data you're interested in. If not, to get you started, we've provided some links to repositories on the course website.



Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Homework Project Resources

Resources

Python

Highly Recommended Tutorials

Learn Python the Hard Way
Code Academy
Python Cheat Sheet
Pandas Cheat Sheet

Data Sources

Wolfram Alpha
Quandl
Datamob
Factual
Metro Boston Data Common
Census.gov
Data.gov
Dataverse Network
Infochimps
Linked Data
Guardian DataBlog
Data Market
Reddit Open Data
Climate Data Sources

Who is CS-5360 / Math-4100?

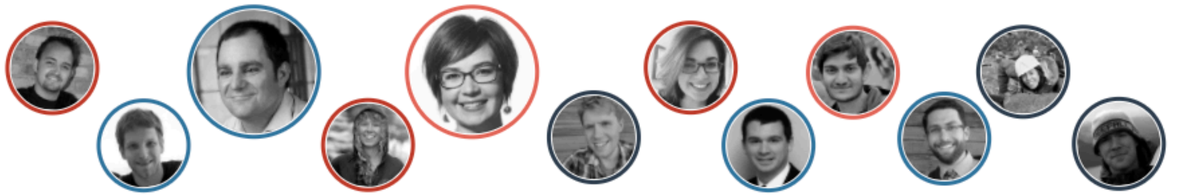
Alexander Lex

@alexander_lex
http://alexander-lex.net
http://vdl.sci.utah.edu

Assistant Professor, Computer Science

Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology



Faculty

Miriah Meyer	Visualization Design Models, Design Studies, User-centered Design
Alexander Lex	Biology Visualization, Multivariate Graphs, Visualization Tools, Exploratory Visualization for Scientists

Post-Doctoral Fellows

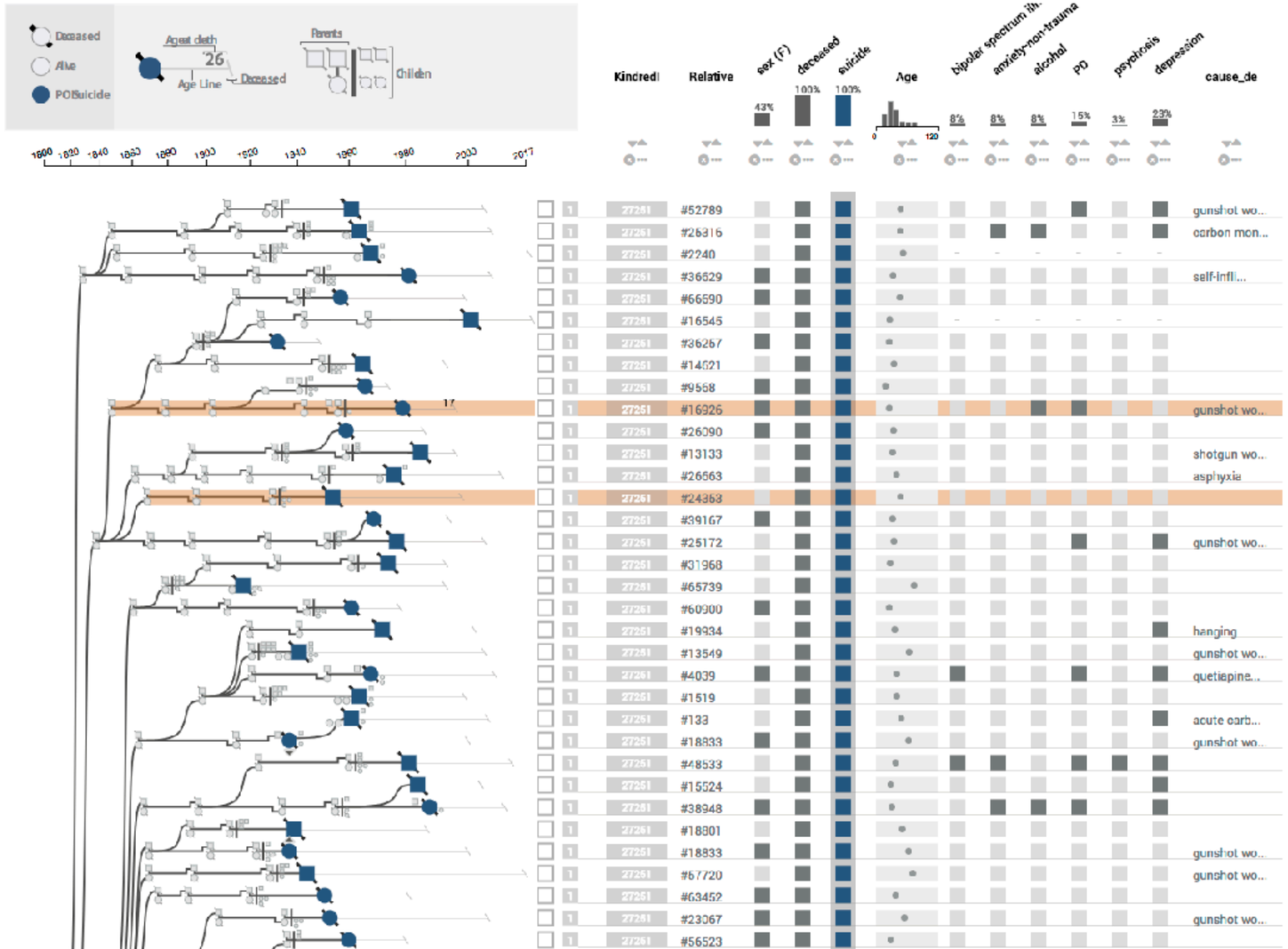
Pascal Goffin	User-Centered Visualization Design, Text Visualization, Word-Scale Visualization, Visualization Tools

PhD Students

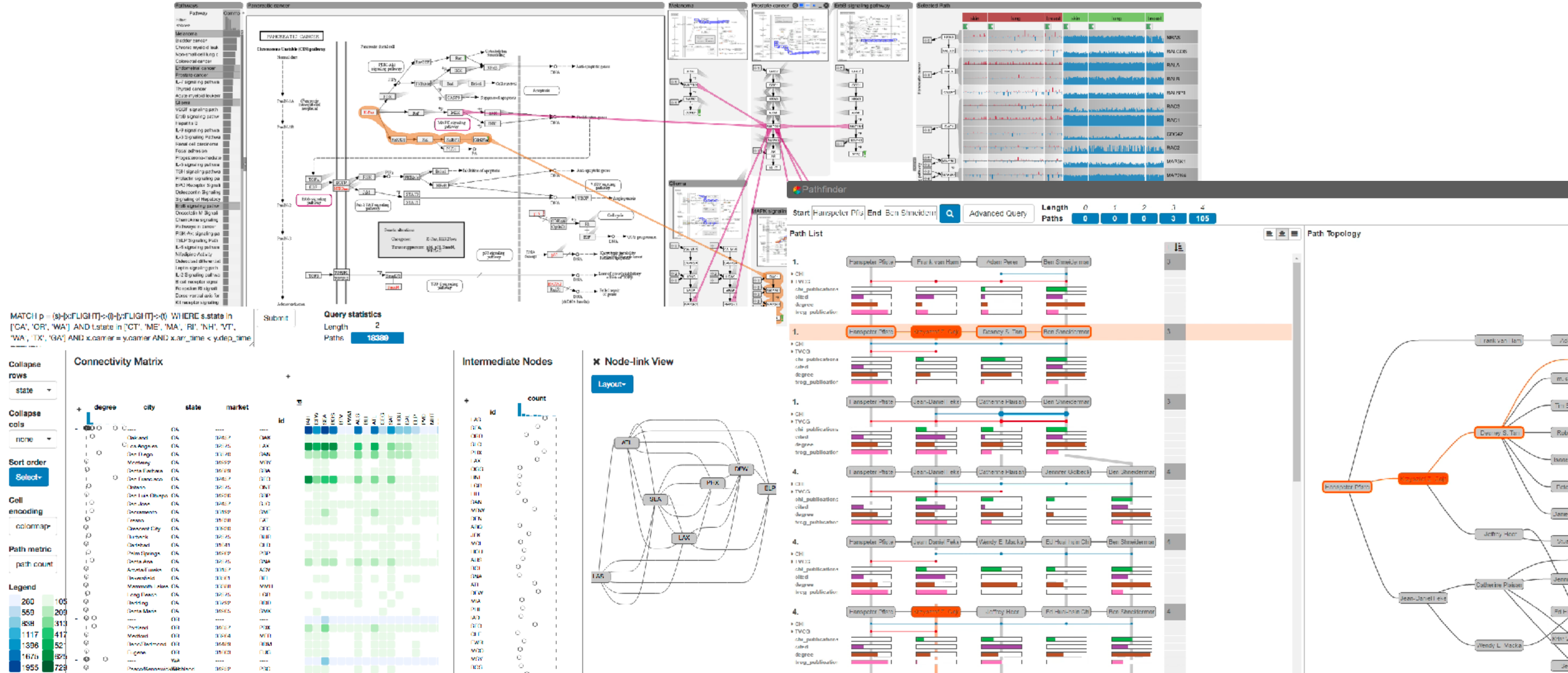
Alex Bigelow	Visualization Design Toolkits, Graph Data, User-centered Design
Ethan Kerzner	Visualization for Scientists and Engineers
Nina McCurdy	Visualization of multivariate, spatial + abstract data, Visualization in the Digital Humanities, Computational Creativity.
Jimmy Moore	Time Series Visualization, User Interface Design, Web Programming, Human-Computer Interaction
Carolina Nobre	Multivariate Graphs, Genealogies, Social Networks
P. Samuel Quinan	Uncertainty Visualization, User-Centered Design, Cognition / Decision-Making, Color
Jen Rogers	Interactive visualization for biological and medical data.
T Cameron Waller	Visualization of Biological Networks, Metabolism

http://vdl.sci.utah.edu/team/

Clinical Genealogies

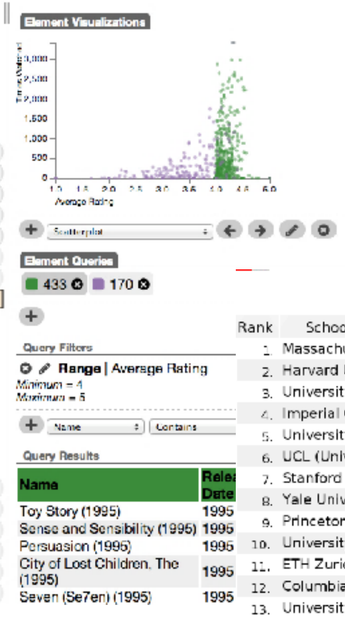


Large, Multivariate (Biological) Networks

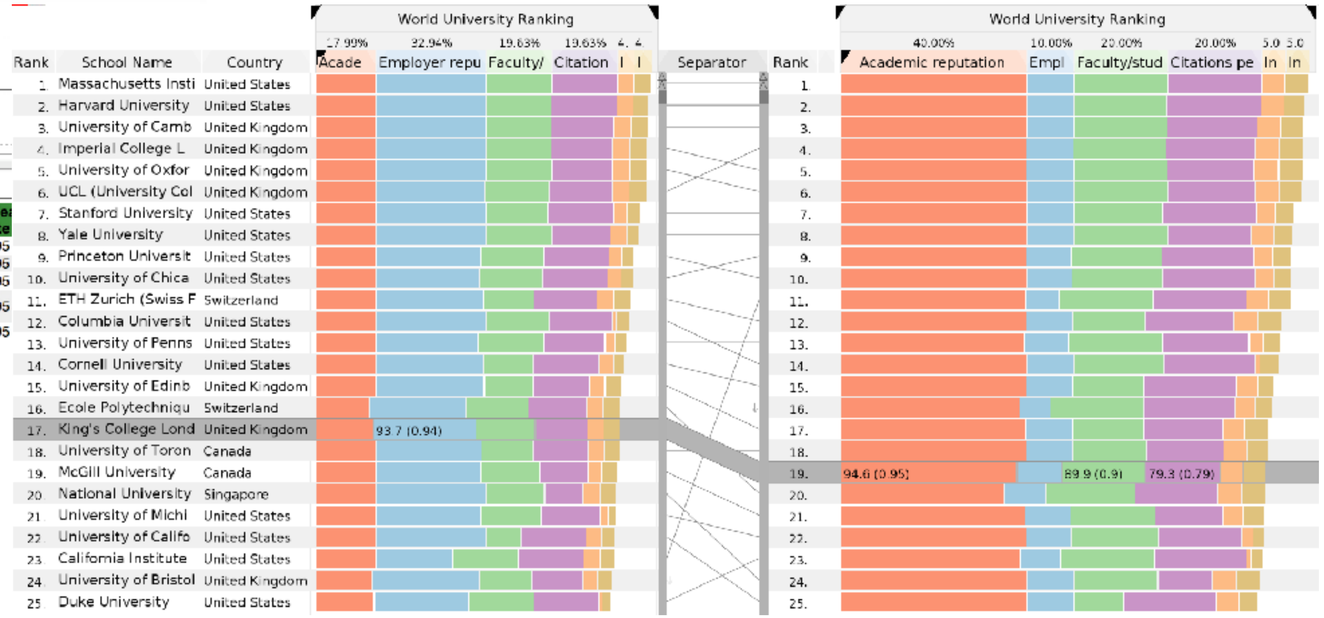


Multidimensional Data

Set Visualization

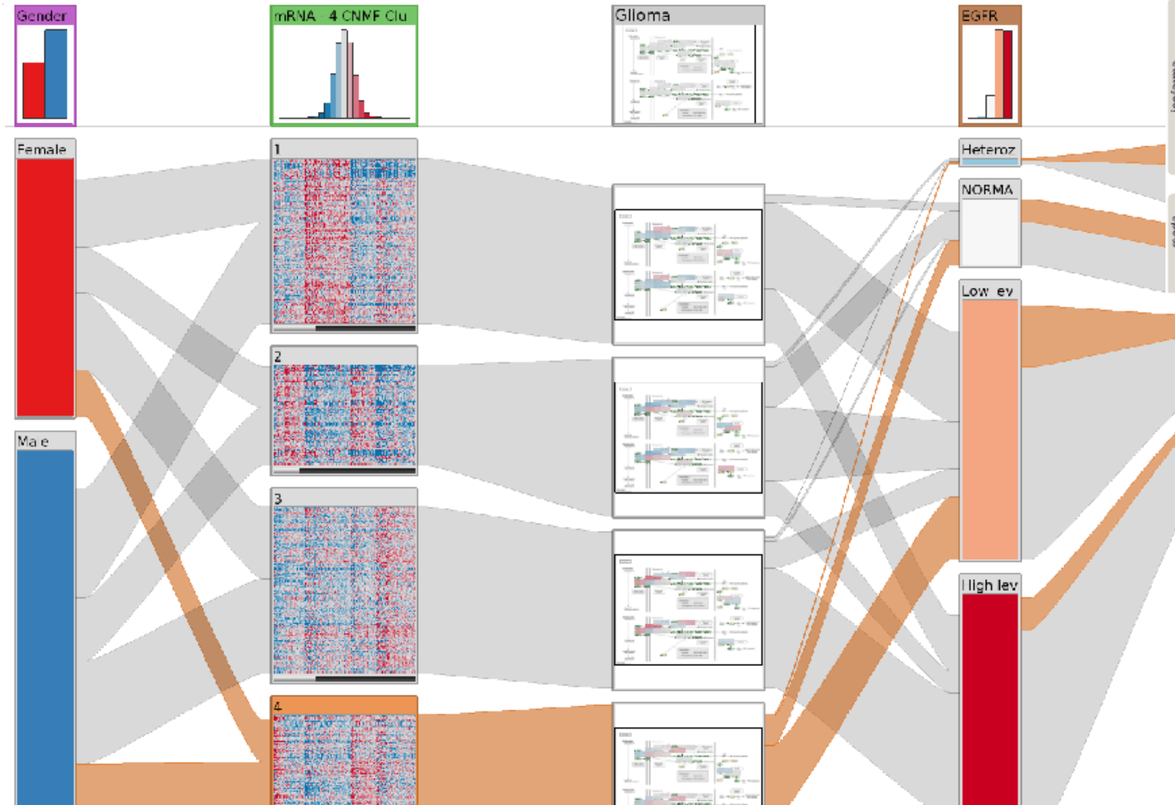


Multivariate Rankings

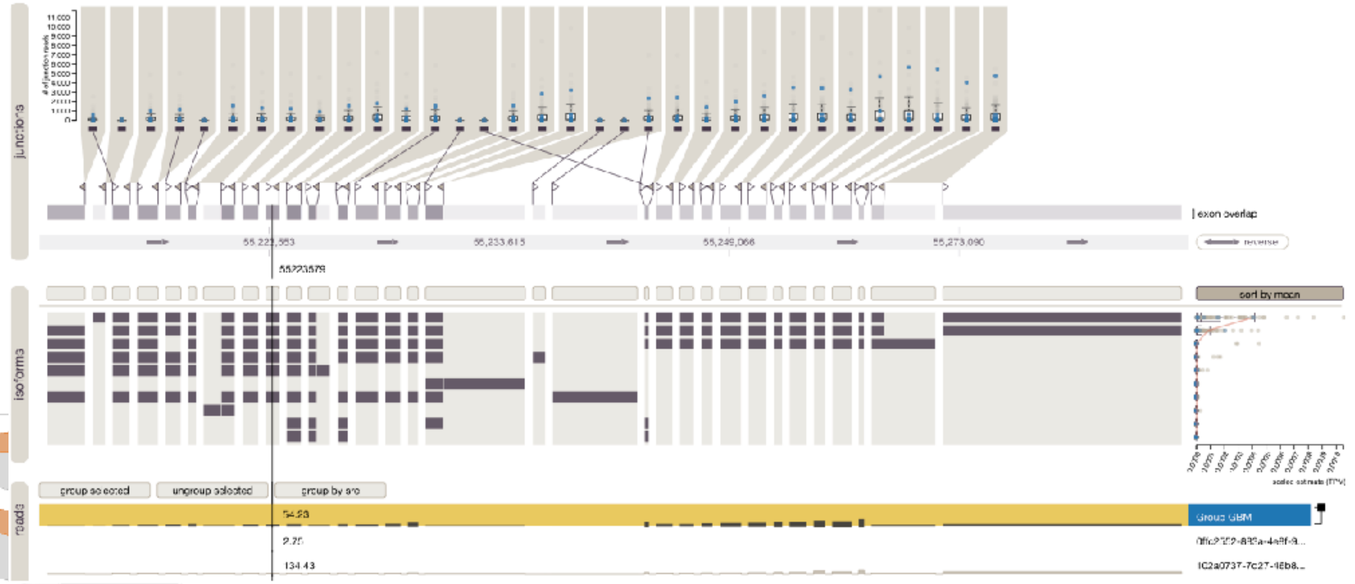


Genomic Data

Cancer Subtypes / Omics
Clustering and Stratification



Alternative Splicing / mRNA-seq



Braxton Osting

Assistant Professor, Mathematics

Before that: Postdoctoral Fellow, UCLA

PhD in Applied Mathematics, Columbia University

http://math.utah.edu/~osting

Teaching Assistants

RK Yoon

Shuvrajit Mukherjee

Structure & Goals

Course Goals

Convey basic skills about each step in the data science process

data wrangling: acquire, clean, reshape, sample data
data exploration and analysis: get a feeling for the dataset, describe dataset
prediction: inferences and decisions based on data
communication

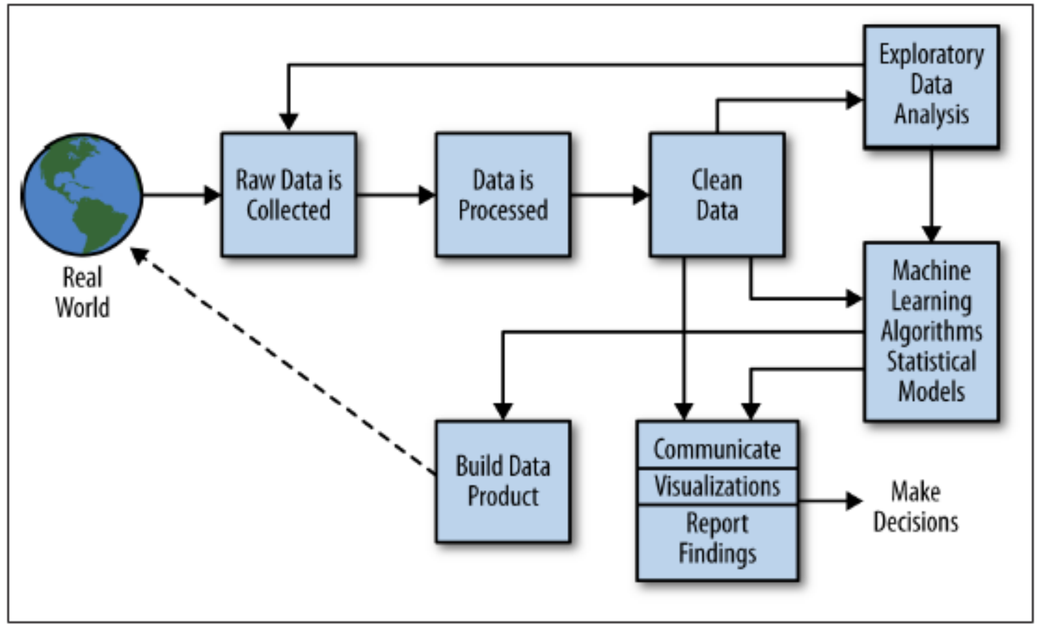


Figure 2-2. The data science process

Topics

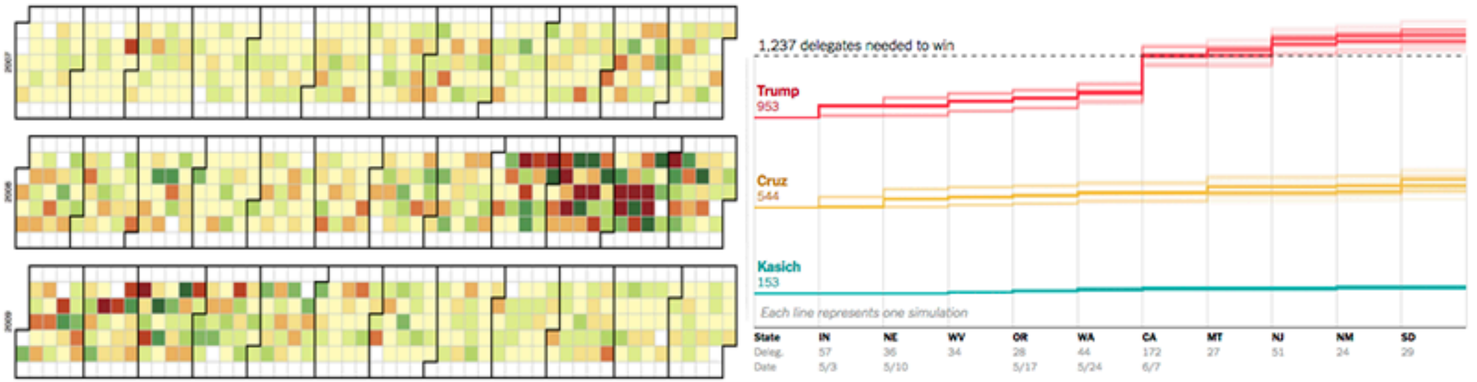
- Programming
- Version Control
- Data Wrangling (Pandas)
- Data Acquisition
  - Web Scraping
  - Web APIs
  - Databases
- Basic Stats
- Hypothesis Testing
- Visualization
- Regression
- Classification
  - Logistic Regression, K-Nearest Neighbors, SVM, Decision Trees, Neural Nets
- Clustering
- Dimensionality Reduction
- Network Analysis
- Natural Language Processing
- Ethics

Information datasciencecourse.net

Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Project Fame Resources



D3 Calendar Chart | How the delegate race could unfold

Introduction to Data Science is a **three-credit course**, offered in the **Spring 2018** semester at the University of Utah, cross-listed between **Mathematics (MATH 4100)** and **Computing (COMP 5360)**.

The amount and complexity of information produced in science, engineering, business, and everyday human activity are increasing at a staggering rate. **The goal of this course is to expose you to methods and techniques for analyzing and understanding complex data.** Data Science lies at the intersection of statistics, computer science, and, of course, the domain from which the data comes from. This course will provide an introduction to the former two: statistics and computer science and provide you with a toolset to conquer problems in your domain!

Communicate

Slack Team

https://datasciencecourse2018.slack.com/

Used for announcements and discussions. Sign up with your utah.edu e-mail address.

Canvas

https://utah.instructure.com/courses/476427

Used only for grading

Office Hours

Braxton: Tuesdays, 1:00-2:00, LCB 118

Alex: Thursdays, 3:30 - 4:30, WEB 3887

TAs:

Wednesdays, 3:00 - 4:30, room TBA

Thursdays, 12:30 - 2:00, room TBA

E-Mail

alex@sci.utah.edu

osting@math.utah.edu

Course Components

Lectures introduce theory and coding

- includes both short, hands-on coding exercises and longer, in-depth coding examples
- Based on a published Jupyter notebook on website
- Strongly related to homework assignments
- Applications!

Homeworks help practice specific skills

Final Project gives you a chance to go through the complete data science process

How are you graded?

Homework Assignments: 60%

- Varying value, depending on length/difficult
- Start early!
- Due on Fridays, late days: -10% per day, up to two days.

Final Project: 40%

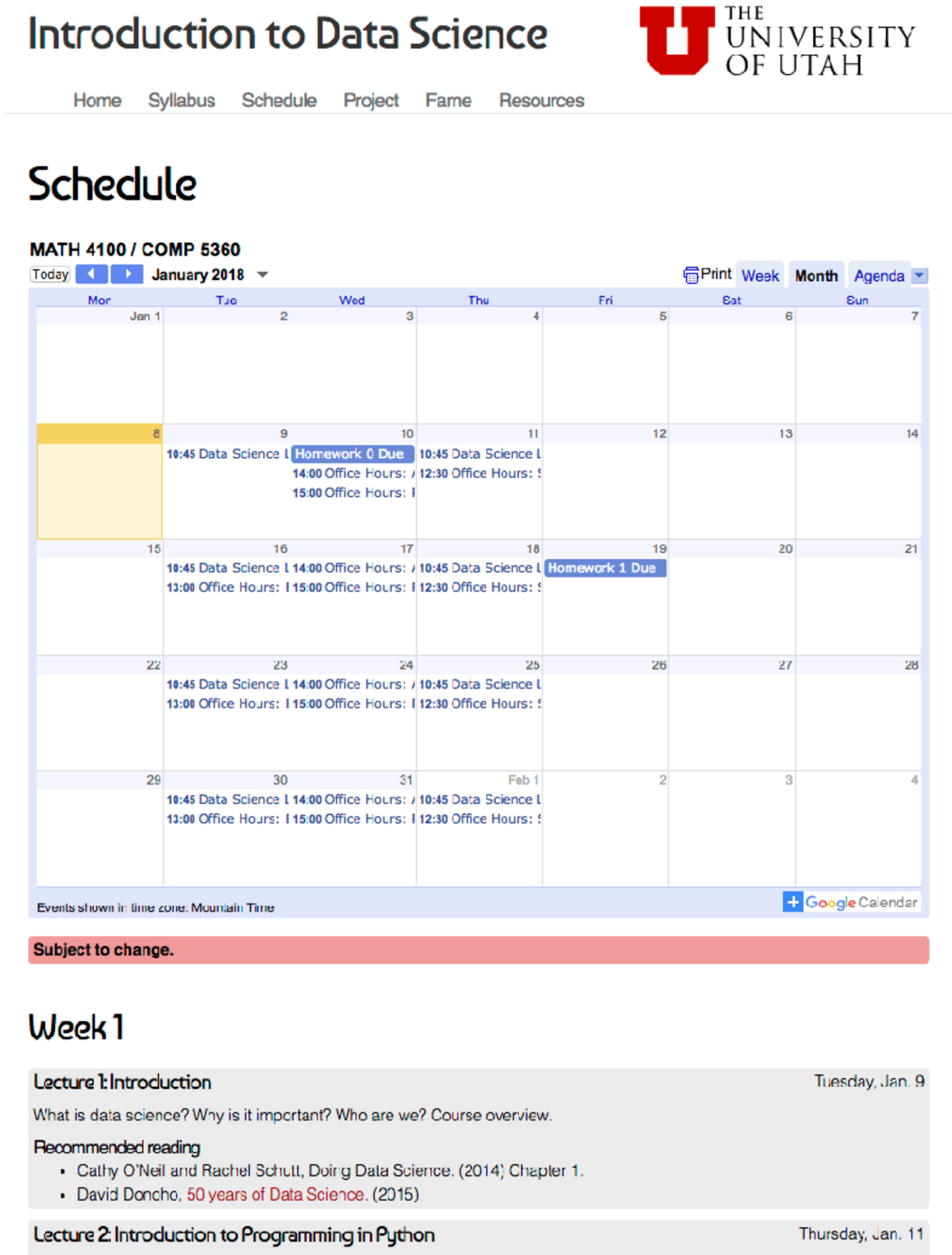
- Teams, two milestones

Schedule

Lectures:
Tue / Th 10:45 - 12:05
WEB 2250

Calendar
Link

Lectures frequently involve computer activities.
Bring your own computer!
Have Python, etc installed (see HW0)



Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Project Fame Resources

Schedule

MATH 4100 / COMP 5360

Today January 2018 Print Week Month Agenda

Mon	Tue	Wed	Thu	Fri	Sat	Sun
Jan 1	2	3	4	5	6	7
8	9 10:45 Data Science L	10 Homework 0 Due 14:00 Office Hours: 15:00 Office Hours:	11 10:45 Data Science L 12:30 Office Hours:	12	13	14
15	16 10:45 Data Science L 13:00 Office Hours:	17 14:00 Office Hours: 15:00 Office Hours:	18 10:45 Data Science L 12:30 Office Hours:	19 Homework 1 Due	20	21
22	23 10:45 Data Science L 13:00 Office Hours:	24 14:00 Office Hours: 15:00 Office Hours:	25 10:45 Data Science L 12:30 Office Hours:	26	27	28
29	30 10:45 Data Science L 13:00 Office Hours:	31 14:00 Office Hours: 15:00 Office Hours:	Feb 1 10:45 Data Science L 12:30 Office Hours:	2	3	4

Events shown in time zone: Mountain Time Google Calendar

Subject to change.

Week 1

Lecture 1: Introduction Tuesday, Jan. 9

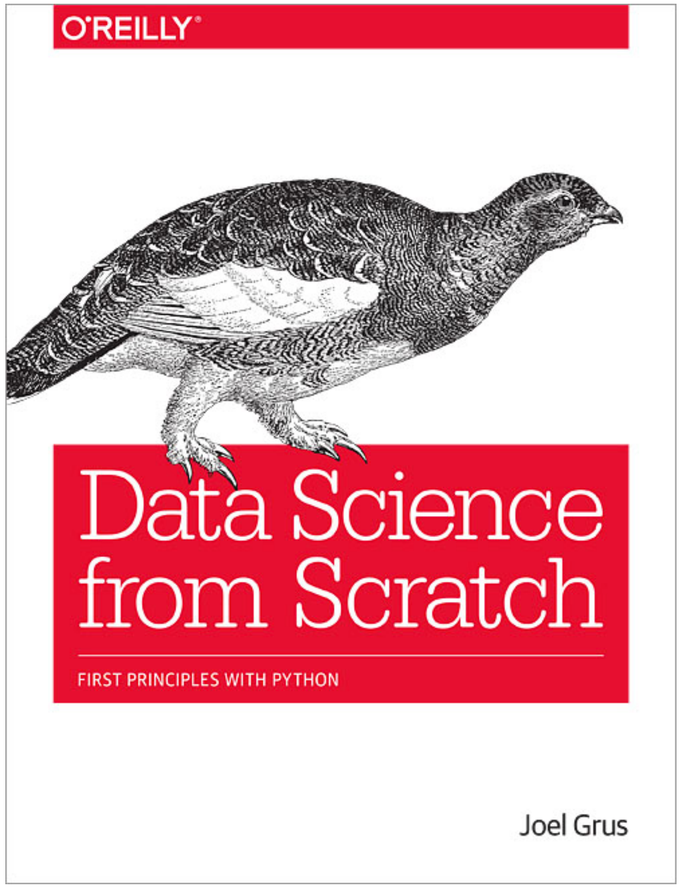
What is data science? Why is it important? Who are we? Course overview.

Recommended reading

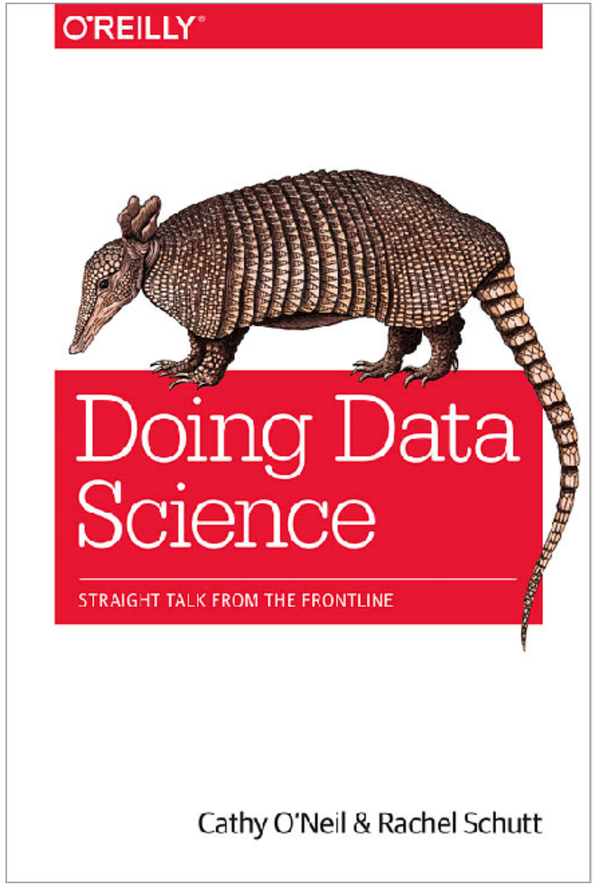
- Cathy O'Neil and Rachel Schutt, Doing Data Science. (2014) Chapter 1.
- David Doncho. 50 years of Data Science. (2015)

Lecture 2: Introduction to Programming in Python Thursday, Jan. 11

Books



Primary Text for Readings
Available for free on Campus:
http://proquest.safaribooksonline.com/9781491901410

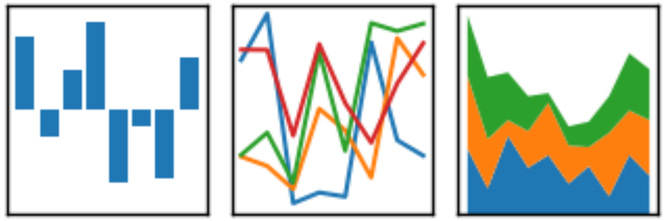


Supplementary Text

Programming

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Is this course for me ???

Prerequisites

Programming experience

- Python, C, C++, Java, etc.

Calculus 1

- UU Math 1170, 1210, 1250 1310, 1311 or equivalent

Willingness to learn new software & tools

- This can be time consuming

You will need to build skills by yourself!

- Engineering vs Computer Science

If in doubt, ask one of the instructors.

This Week

HW0, including course survey

- Make sure to get this working soon. Office Hours!

Introduction to programming in python

Readings:

- Cathy O'Neil and Rachel Schutt, Doing Data Science. (2014) Chapter 1.
- David Donoho, 50 years of Data Science. (2015).

HW 0

https://github.com/datascience-course/2018-datascience-homeworks/tree/master/HW0

README.md

Homework 0

Introduction to Data Science - MATH 4100 / COMP 5360.
This homework is due before class on Thursday, January 11th.

Welcome to MATH 4100 and Computing 5360 - Introduction to Data Science. In this class, we will be using a variety of tools that will require some initial configuration. To ensure everything goes smoothly moving forward, we will set up the majority of those tools in this homework. This homework will not be graded, but **it is essential that you complete it before the second lecture** as it sets up the tools that we will be using in class for exercises.

1. Survey

This is a class about data, so we also want to have some data about you! Please complete the course survey located here. It should only take a few moments of your time.

2. Introduction

Once you are signed up for the class and have access to Slack, introduce yourself to your classmates and course staff by introducing yourself in the #general channel. Include your name/nickname, your affiliation, why you are taking this course, and tell us something interesting about yourself (e.g., an unusual hobby, past travels, or a cool project you did, etc.). Also tell us whether you have experience with data science.

3. Setup

In the labs we'll work on practical skills related to what we discuss in the lectures. That means we'll write code, and we'll do that in a programming language called Python.

Next Week

HW1 due

Introduction to Descriptive Statistics

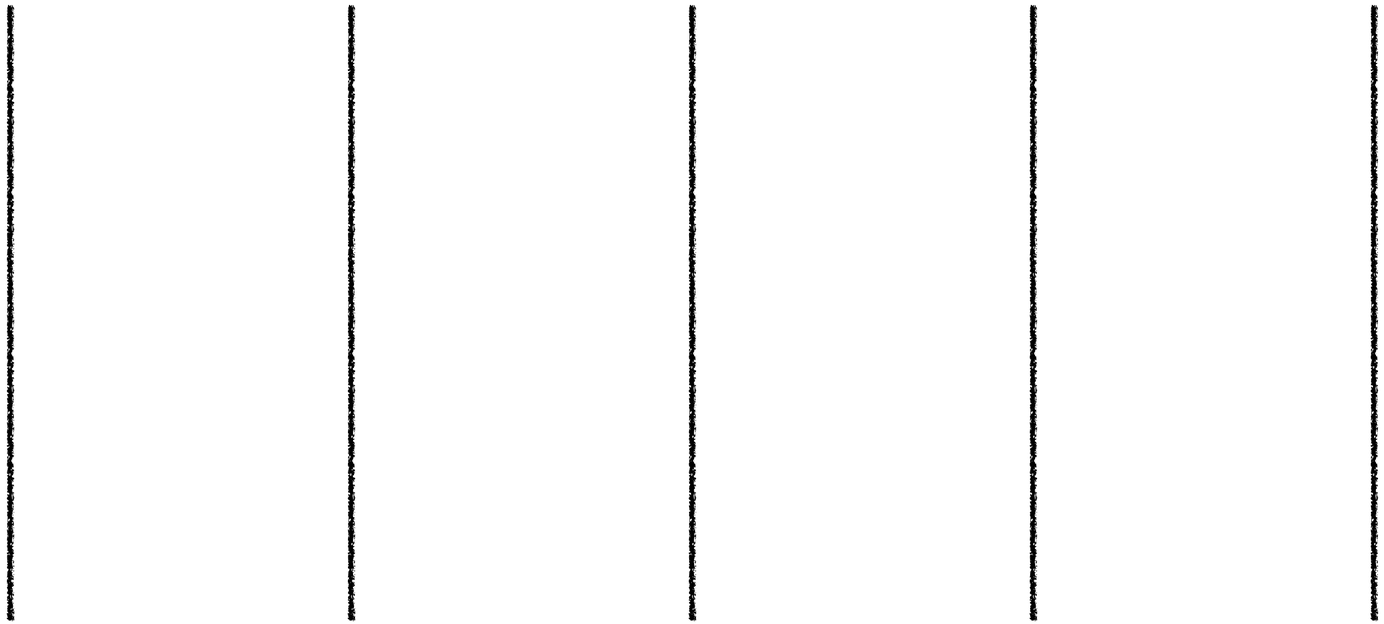
Data Structures and Pandas

About You

Enough about us! Please submit a “data science profile”

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise



1 - little knowledge

5 - Expert

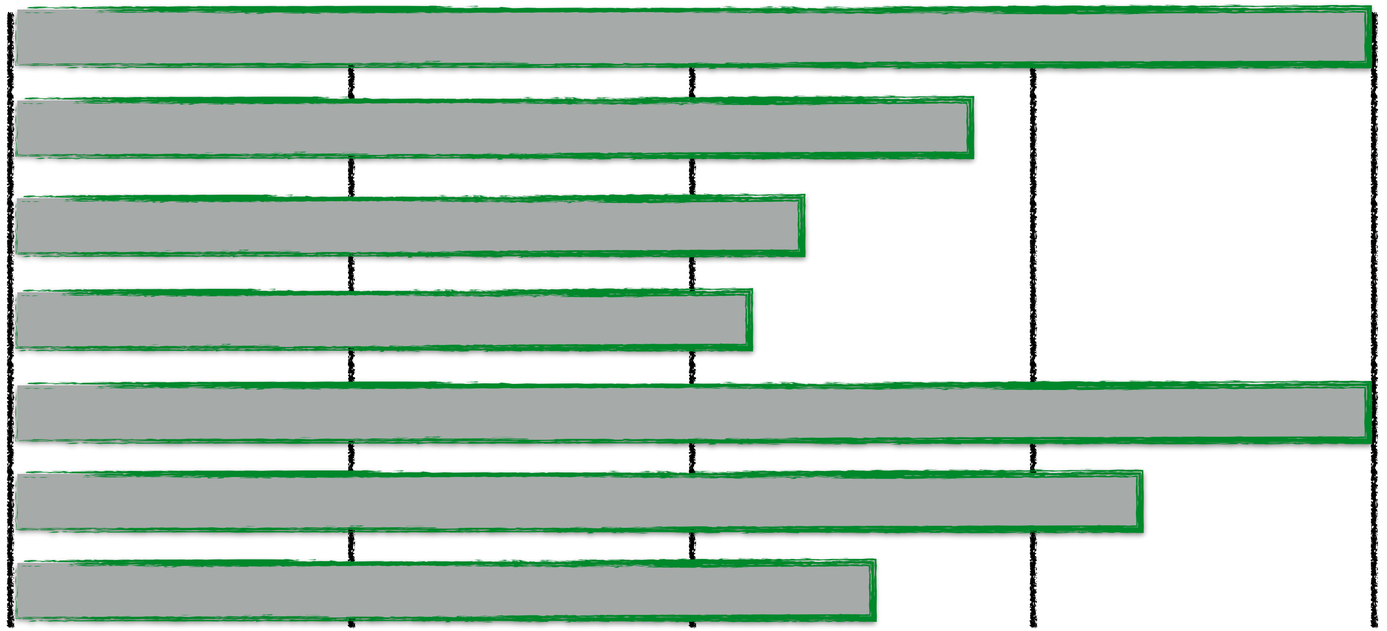
In addition, in the comments section, please write any particular subjects you'd like to see covered in class.

[O’Neil+Schutt (2013), p.10]

Alex's Data Science Profile

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise



1 - little knowledge

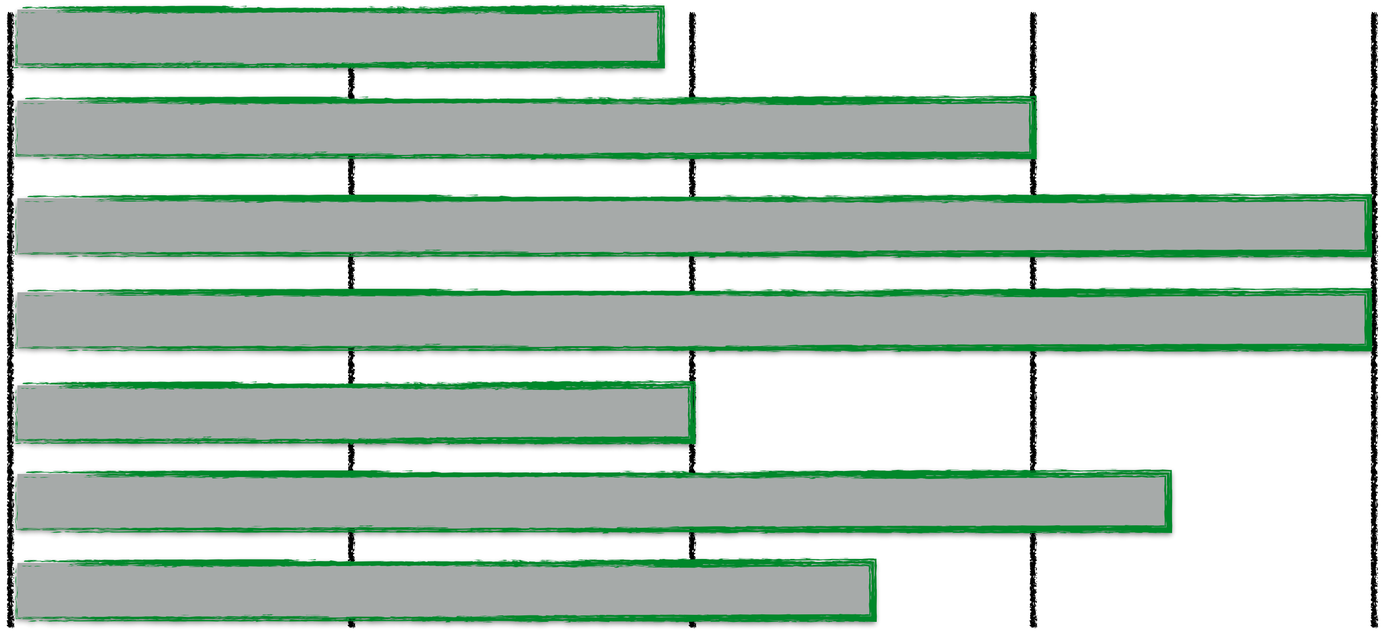
5 - Expert

[O'Neil+Schutt (2013), p.10]

Braxton's Data Science Profile

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise



1 - little knowledge

5 - Expert

[O'Neil+Schutt (2013), p.10]