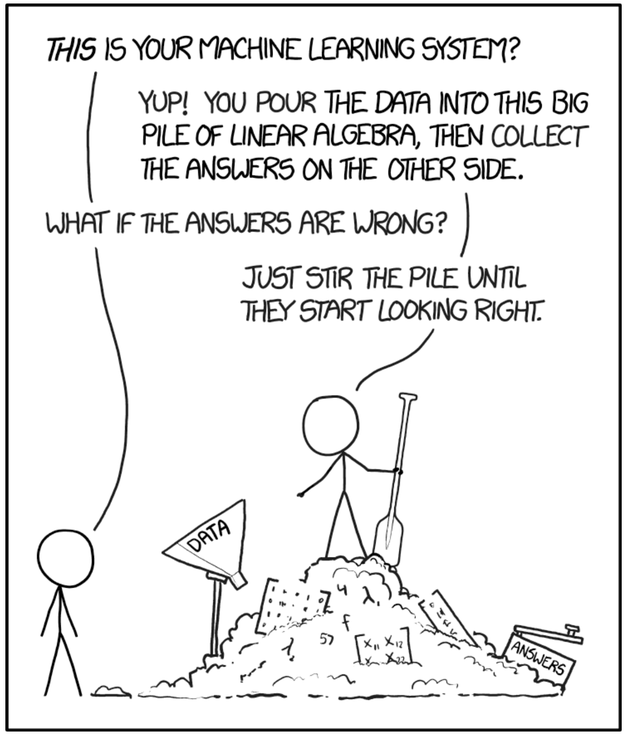
Introduction to Data Science
CS 5360 / Math 4100

Alexander Lex
alex@sci.utah.edu

Anna Little
little@math.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

Recording

Note that this and all future zoom lectures are recorded and live-streamed to YouTube.

Your voice and video may be visible when you participate in discussion.

If you want to participate in a way that is not publicly visible, please ask written questions.

This does not apply to breakout rooms.

If you have concerns, please get in touch.

Camera Policy

We encourage you to keep your camera on: more engaging for everyone. But not required.

We strongly advise that you have your **camera on in breakout rooms**: these are small groups that are not recorded and can serve to get to know other students.

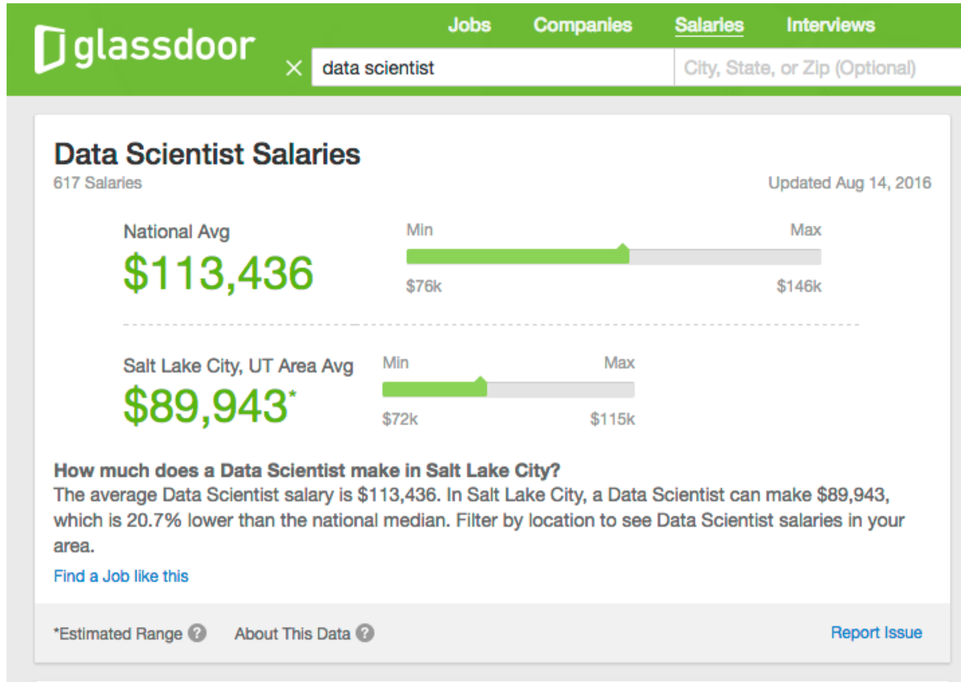
What is Data Science?

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.

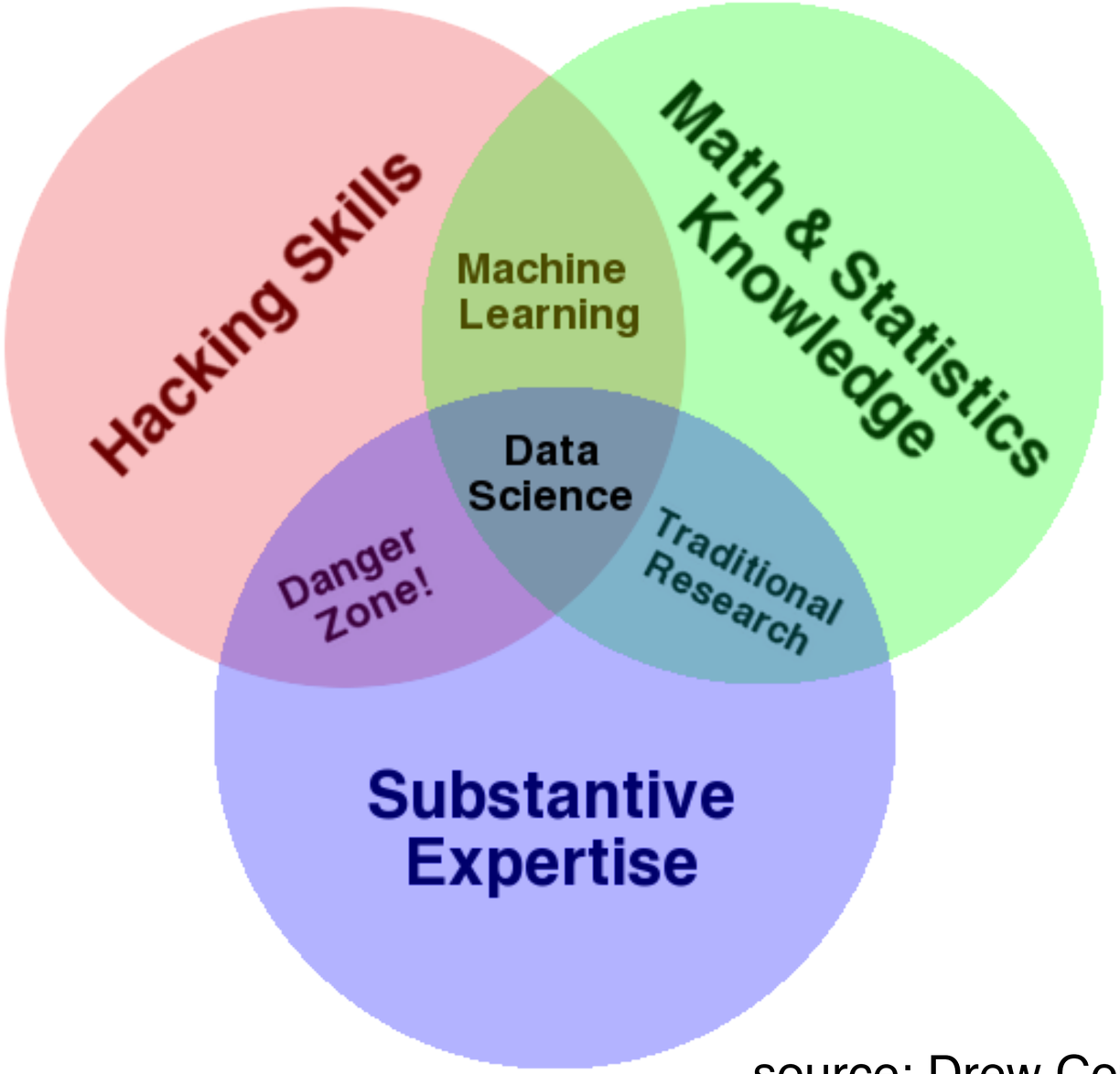


https://twitter.com/jeremyjarvis/status/428848527226437632/photo/1

What is Data Science?

Source: datascience.berkeley.edu

What is Data Science?



source: Drew Conway blog

What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. (Wikipedia)

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

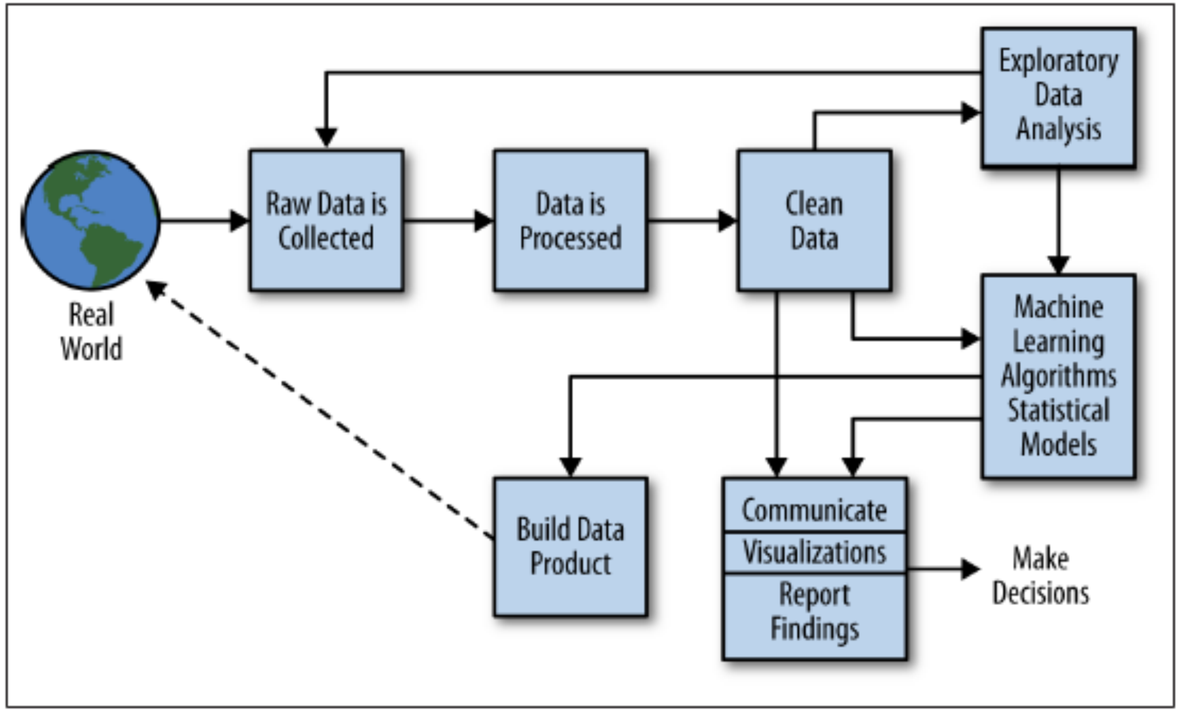


Figure 2-2. The data science process

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? Data is 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 non-majors.

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

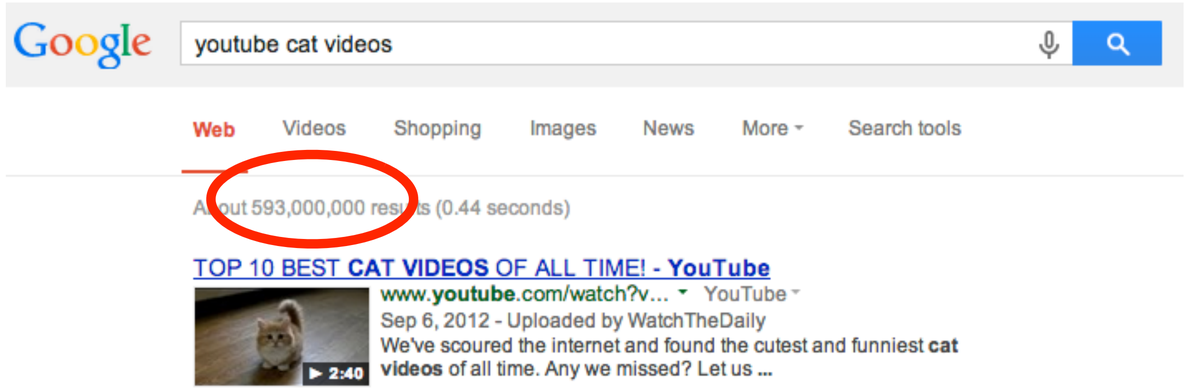
Big Data

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

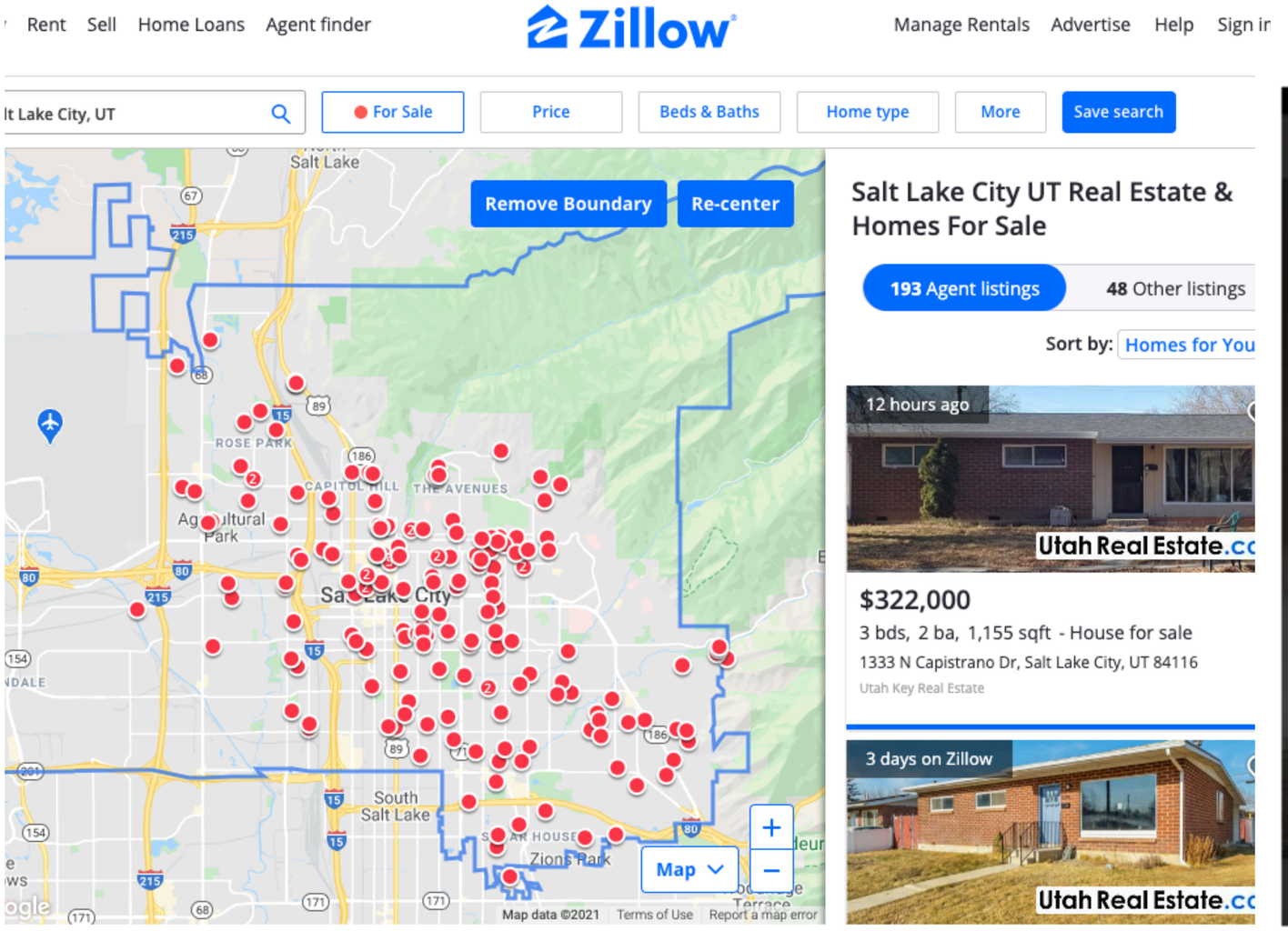
Make better decisions

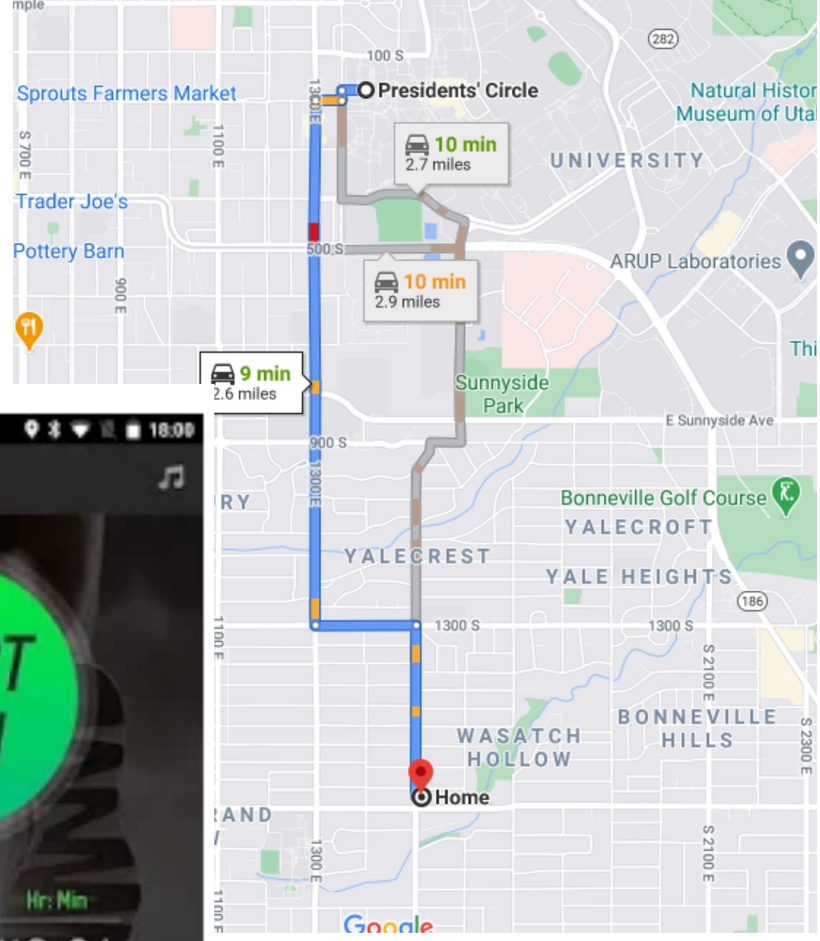
- exact diagnosis, choose right medication, pick good restaurant

Predict elections, events, crowd behavior, etc.

… and many more applications

Example: Personal Data





Big Data in Science and Engineering

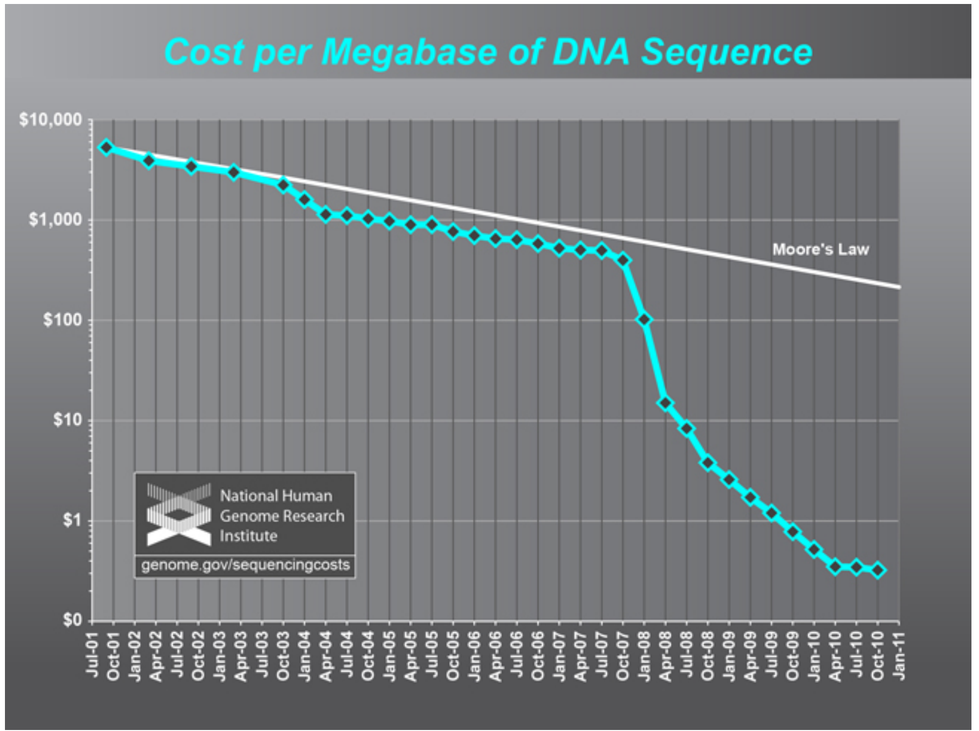
Big Data transformed science and engineering.

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 its still small compared to the amount of data that the National Security Agency (NSA) 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

Example TCGA (Cancer Genome Atlas): 1 Petabyte

“As a single human genome takes up 100 gigabytes of storage space, and more and more genomes are sequenced, storage needs will grow from gigabytes to petabytes to exabytes. By 2025, an estimated 40 exabytes of storage capacity will be required for human genomic data.”

Source: medicalfuturist.com

NSA Utah Data Center (Bluffdale, Utah)

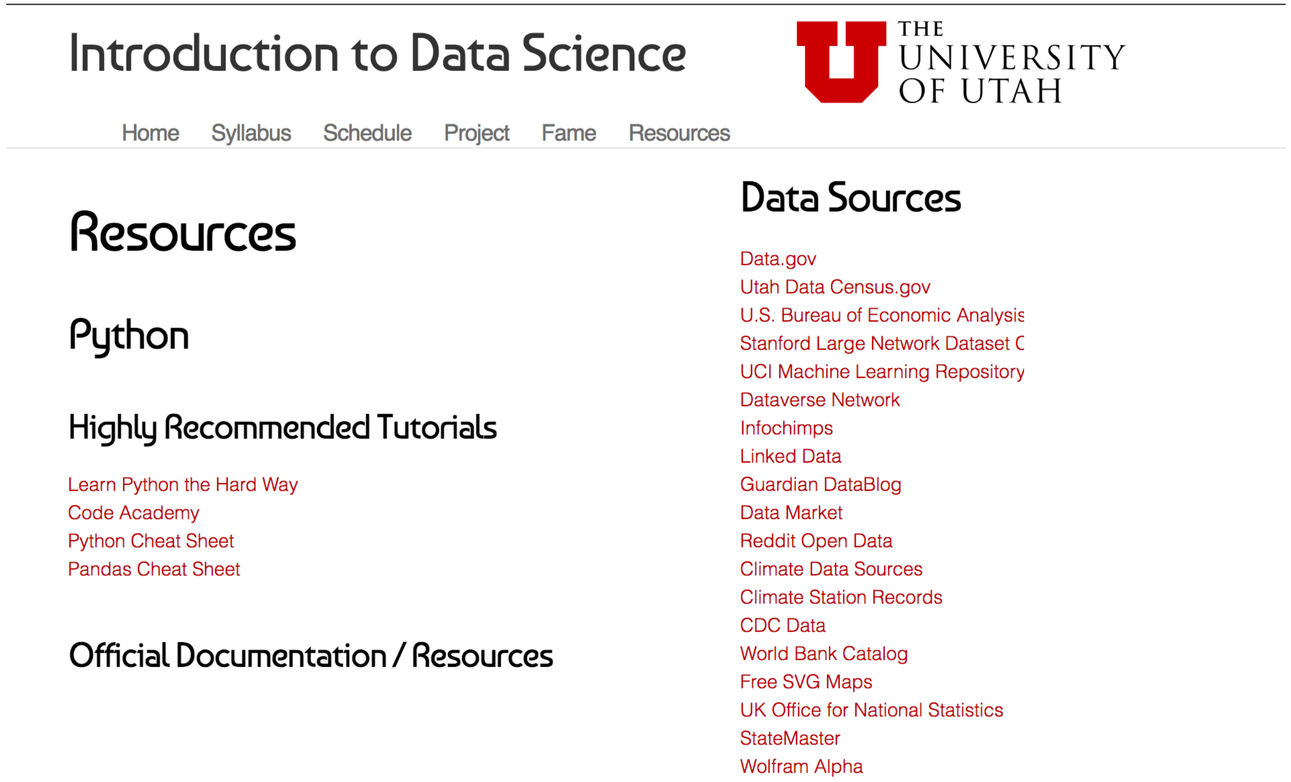
Storage Capacity?

estimates vary, but NPR estimates the center will be able to handle 5 zettabytes (5 billion terabytes)

By User:ParkerHiggins (Electronic Frontier Foundation) - Own work; https://www.eff.org/deeplinks/2014/07/releasing-public-domain-image-nsas-utah-data-center, CC0, https://commons.wikimedia.org/w/index.php?curid=33871035

Where can you find data?

Today, a lot of data is publicly available. You probably have access to data that 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 Project Fame Resources

Resources

Python

Highly Recommended Tutorials

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

Official Documentation / Resources

Data Sources

Data.gov
Utah Data Census.gov
U.S. Bureau of Economic Analysis
Stanford Large Network Dataset C
UCI Machine Learning Repository
Dataverse Network
Infochimps
Linked Data
Guardian DataBlog
Data Market
Reddit Open Data
Climate Data Sources
Climate Station Records
CDC Data
World Bank Catalog
Free SVG Maps
UK Office for National Statistics
StateMaster
Wolfram Alpha

Course 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

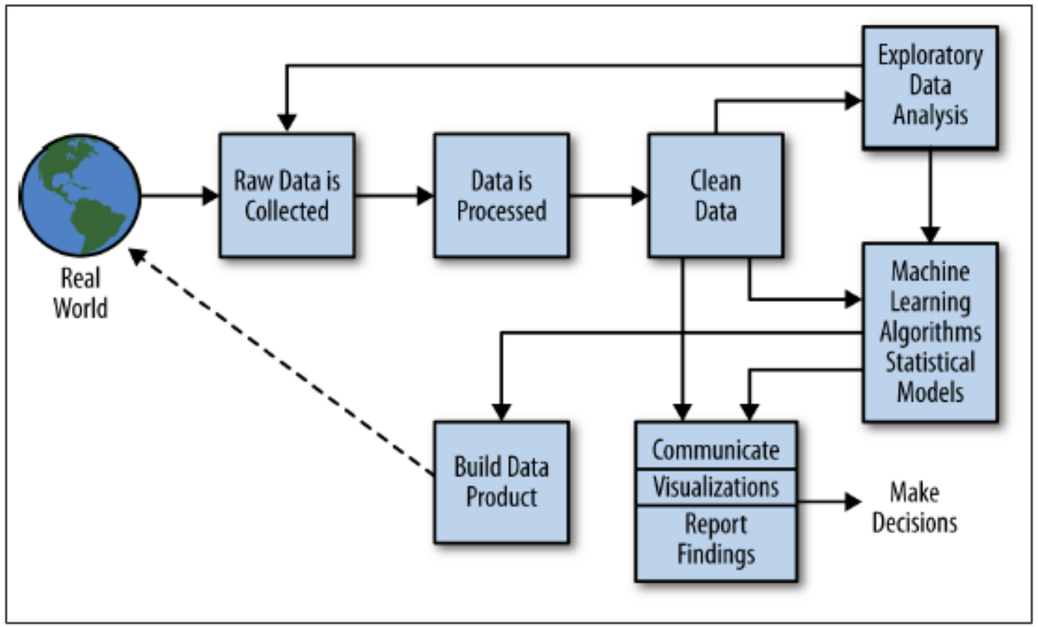


Figure 2-2. The data science process

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 Networks
- Clustering
- Dimensionality Reduction
- Network Analysis
- Natural Language Processing
- Ethics

Who is
CS 5360 / Math 4100?

Anna Little

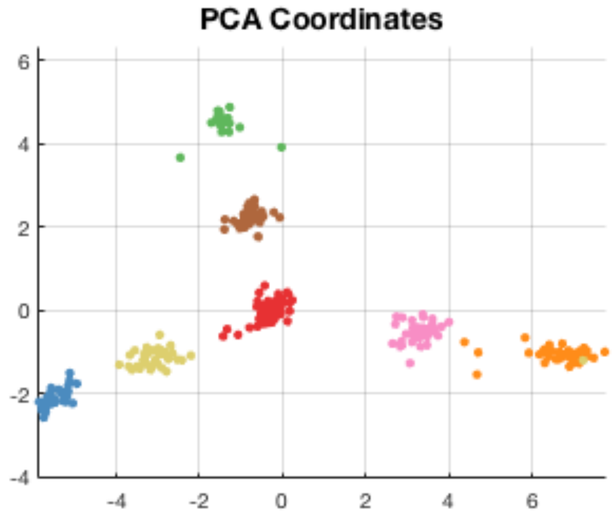
Assistant Professor, Mathematics

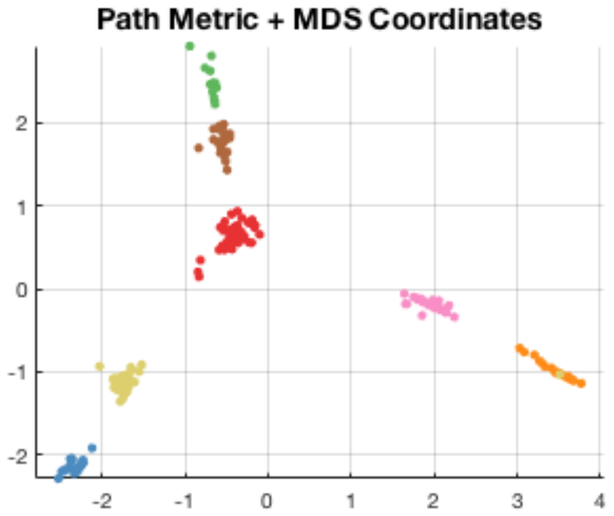
Before that: Postdoc, Michigan State University

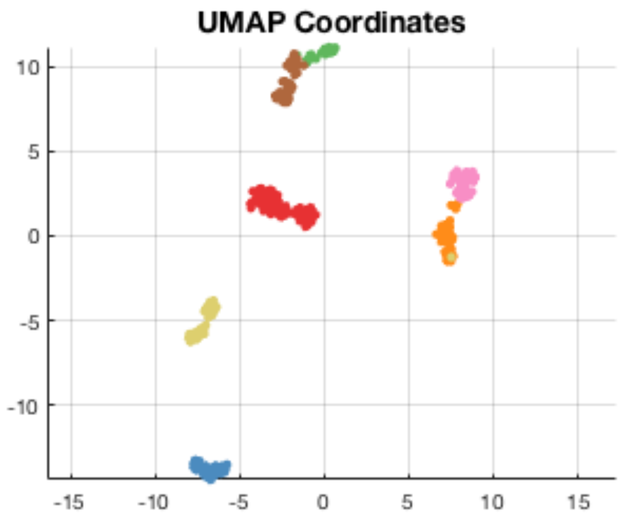
PhD in Mathematics, Duke University

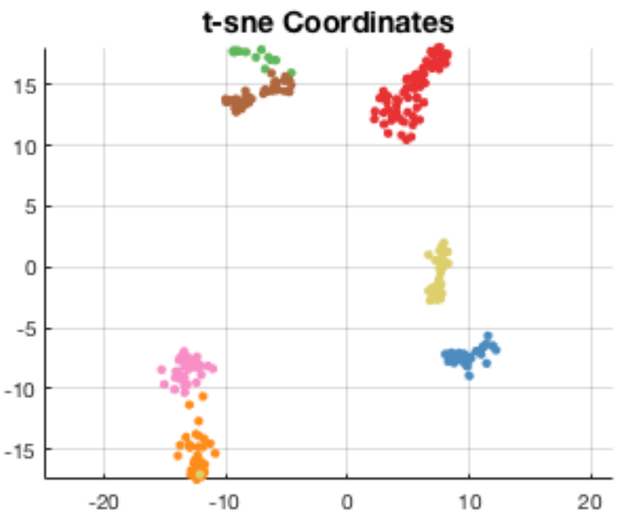
https://www.anna-little.com/

Data Analysis, Dimension Reduction & Estimation

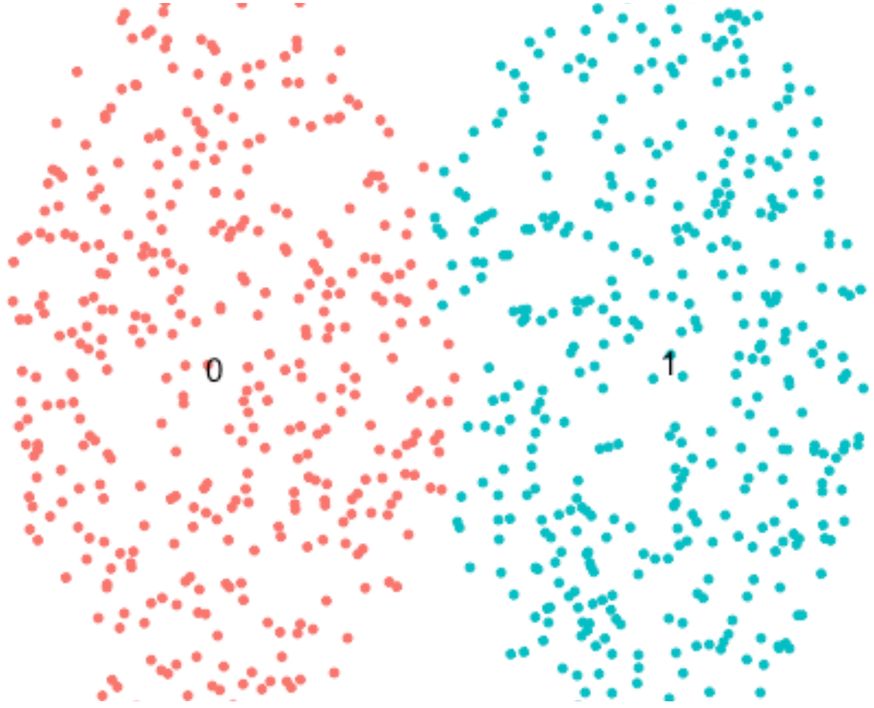


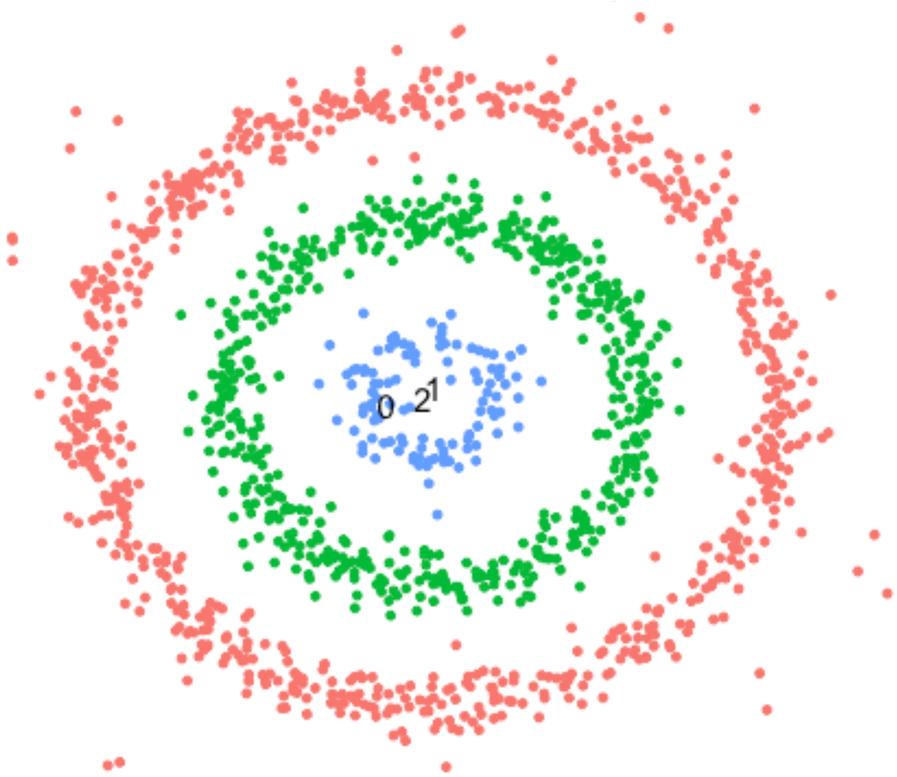






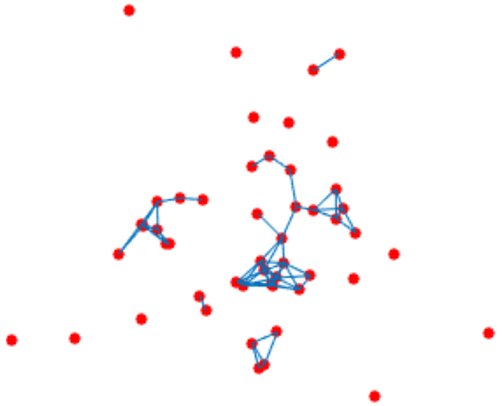
Clustering and Notions of Similarity



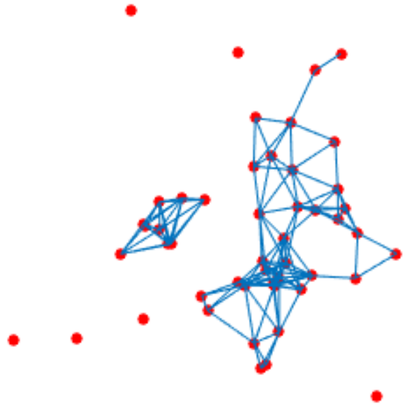


Graph-based Methods, Fast Computation

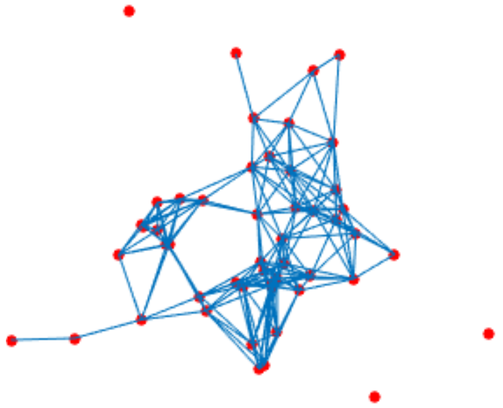
G_1: All Edges < 0.36



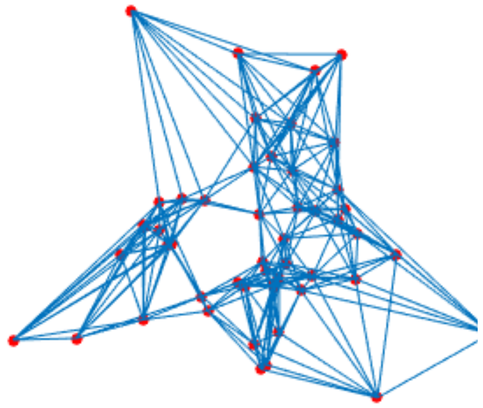
G_2: All Edges < 0.56



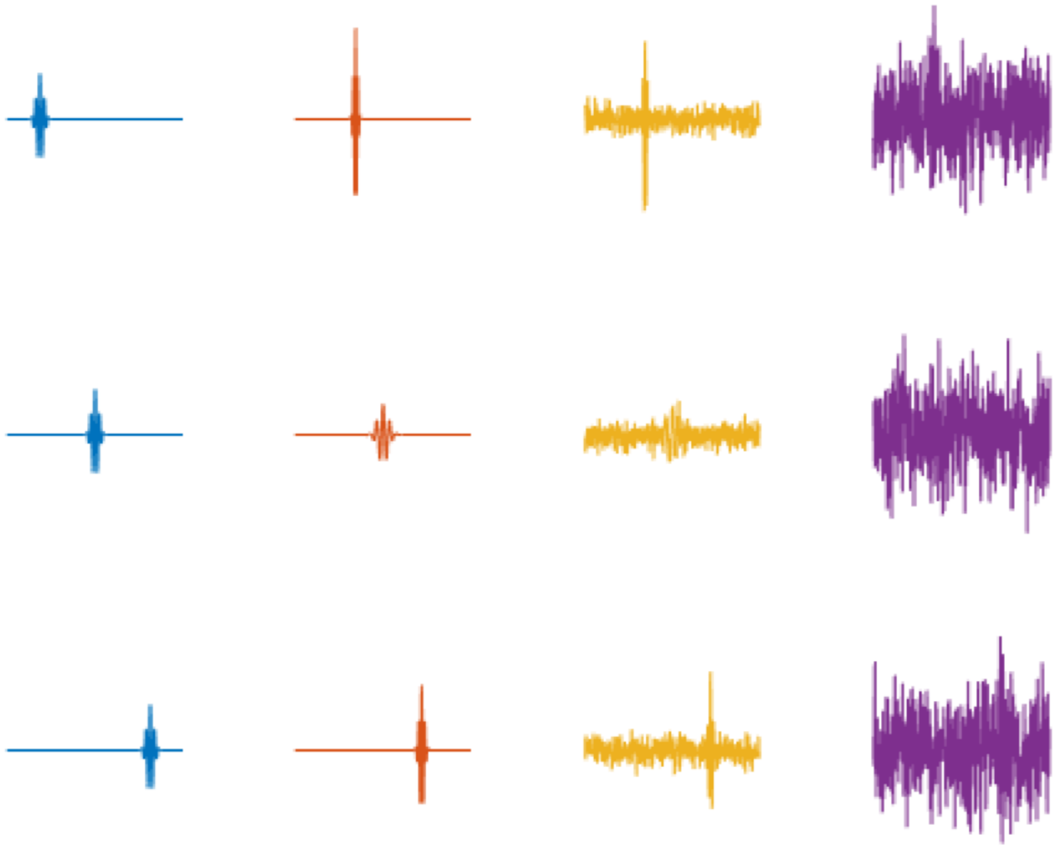
G_3: All Edges < 0.8

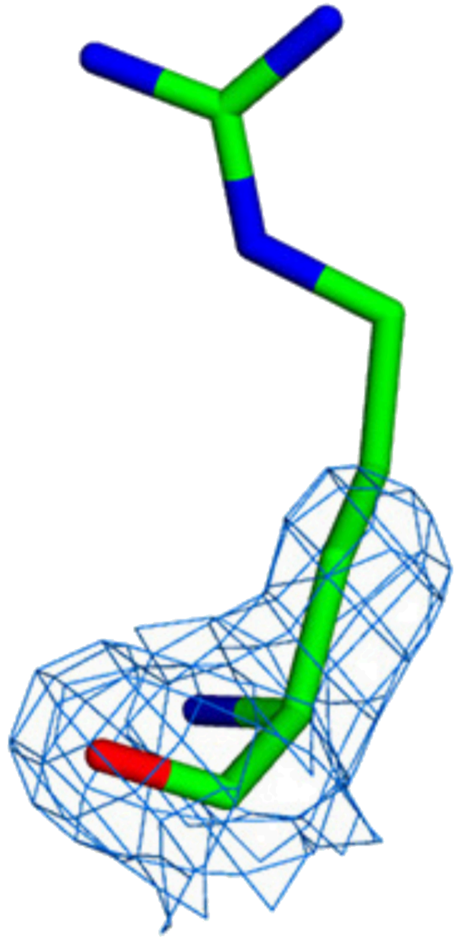


G_4: All Edges



Multi-reference Alignment, Signal Processing





Alexander Lex

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

Associate Professor, Computer Science

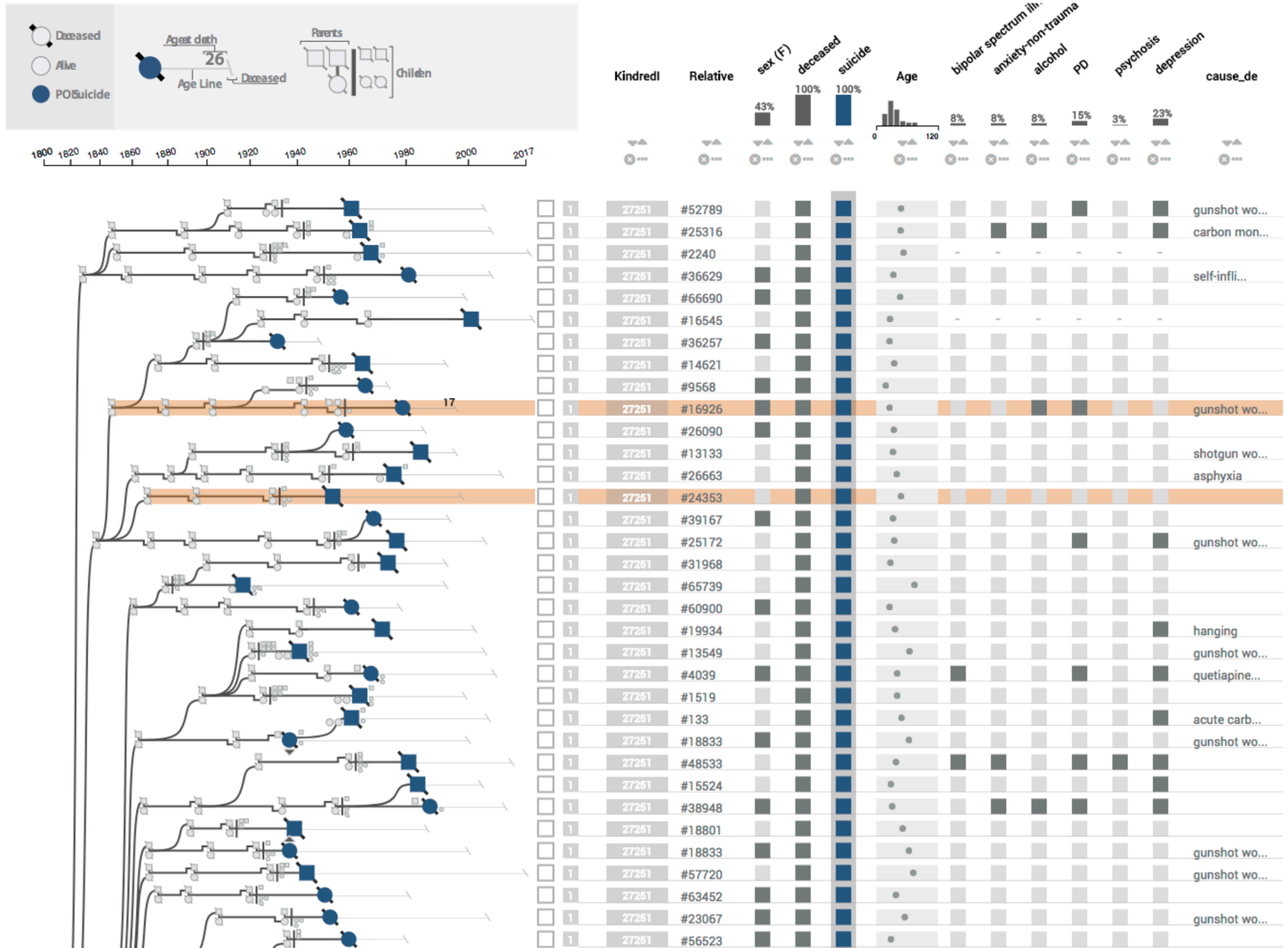
Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology

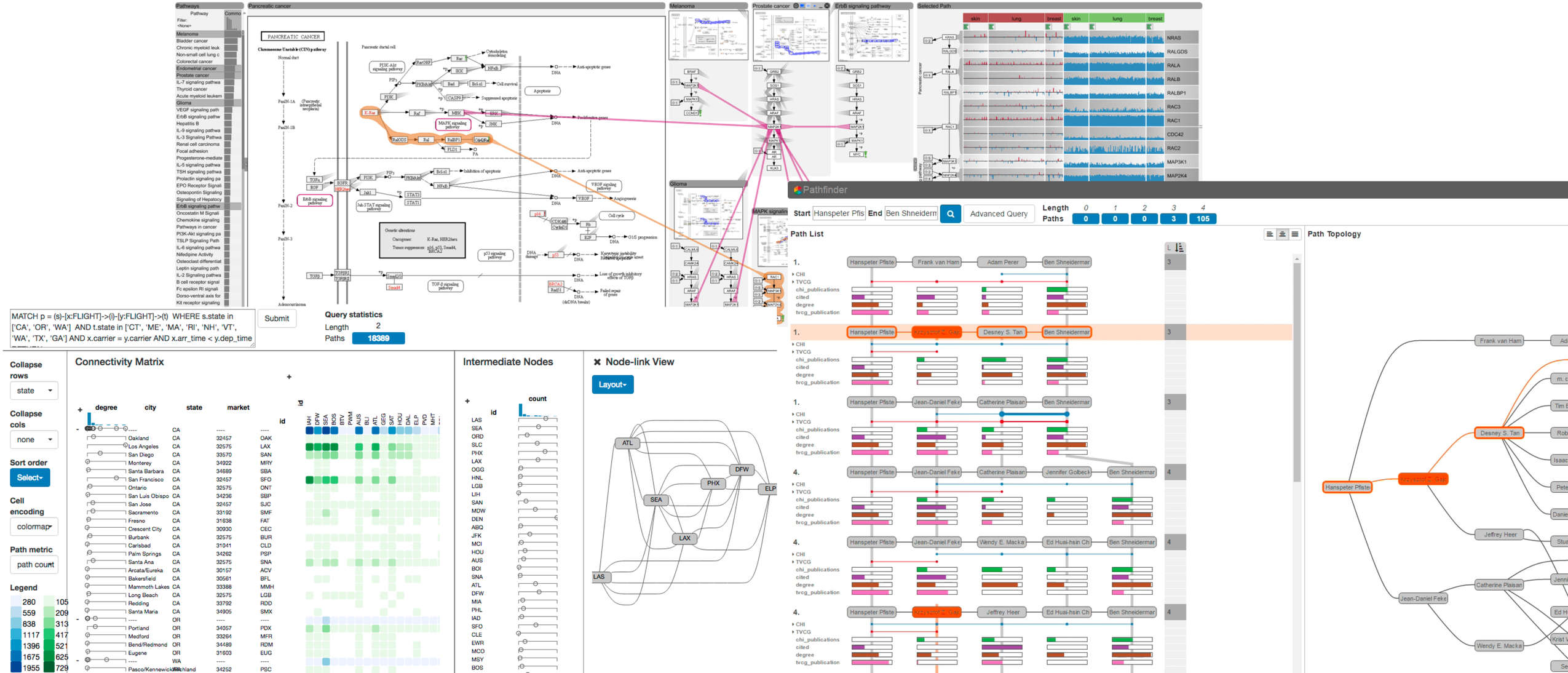
visualization
design lab

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

Clinical Genealogies

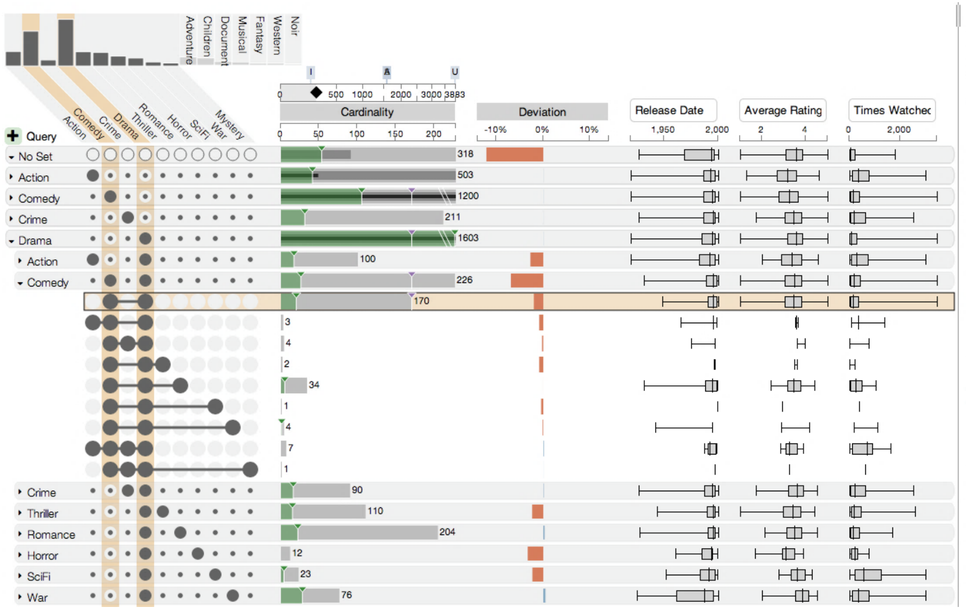


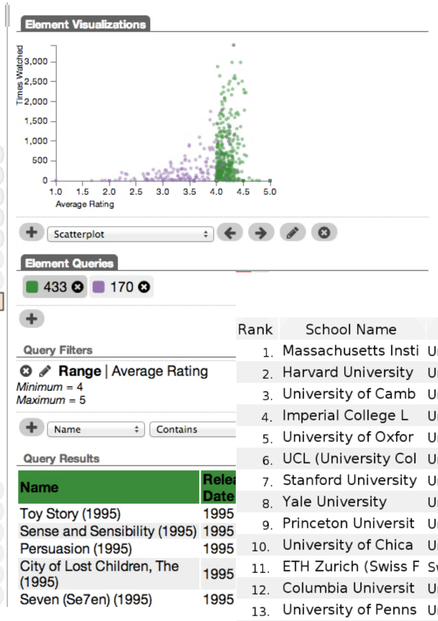
Large, Multivariate (Biological) Networks



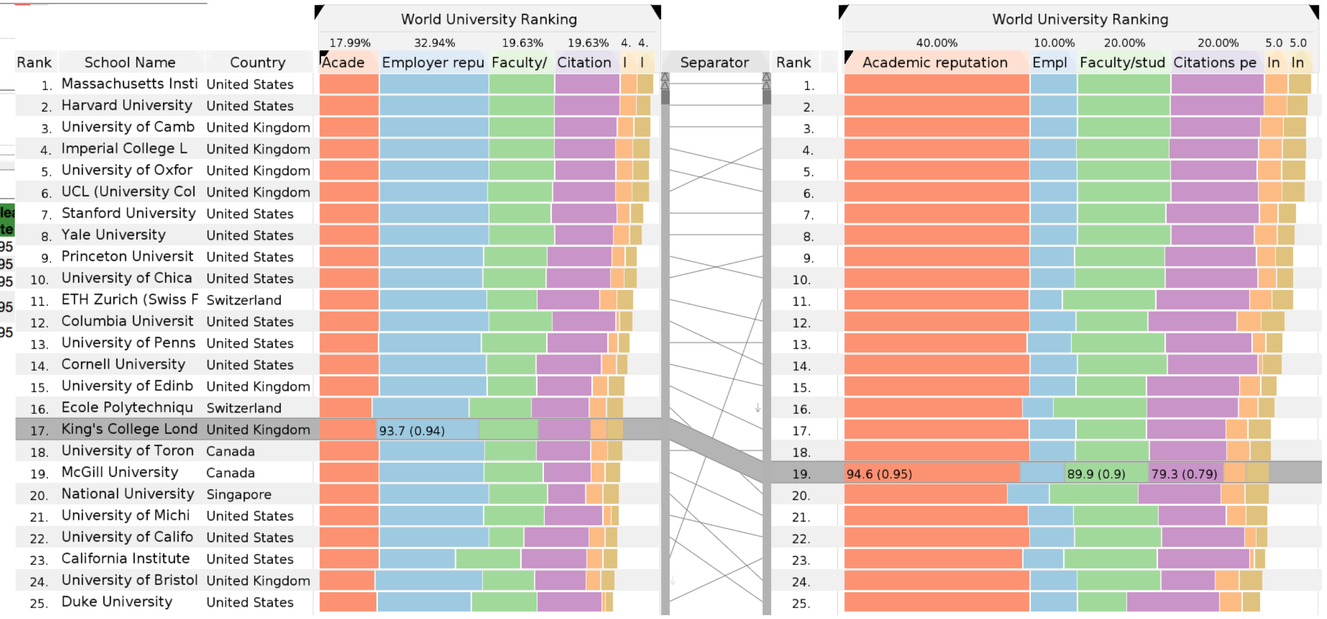
Multidimensional Data

Set Visualization





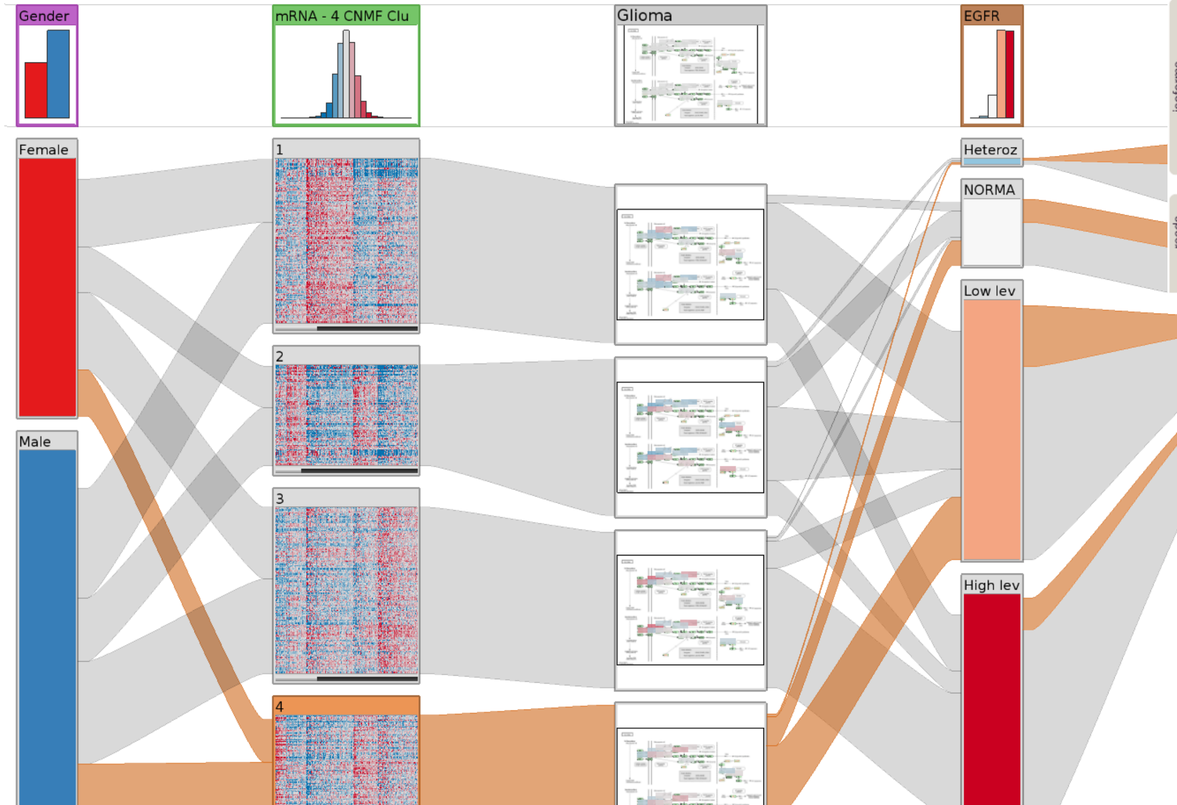
Multivariate Rankings

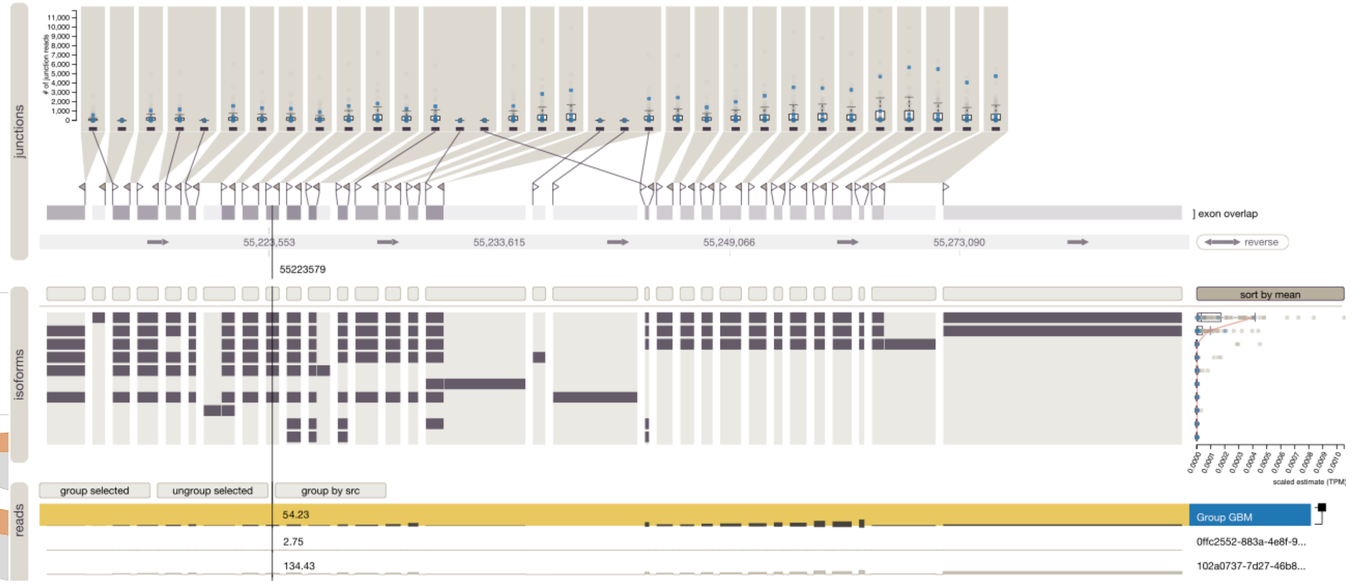


Genomic Data

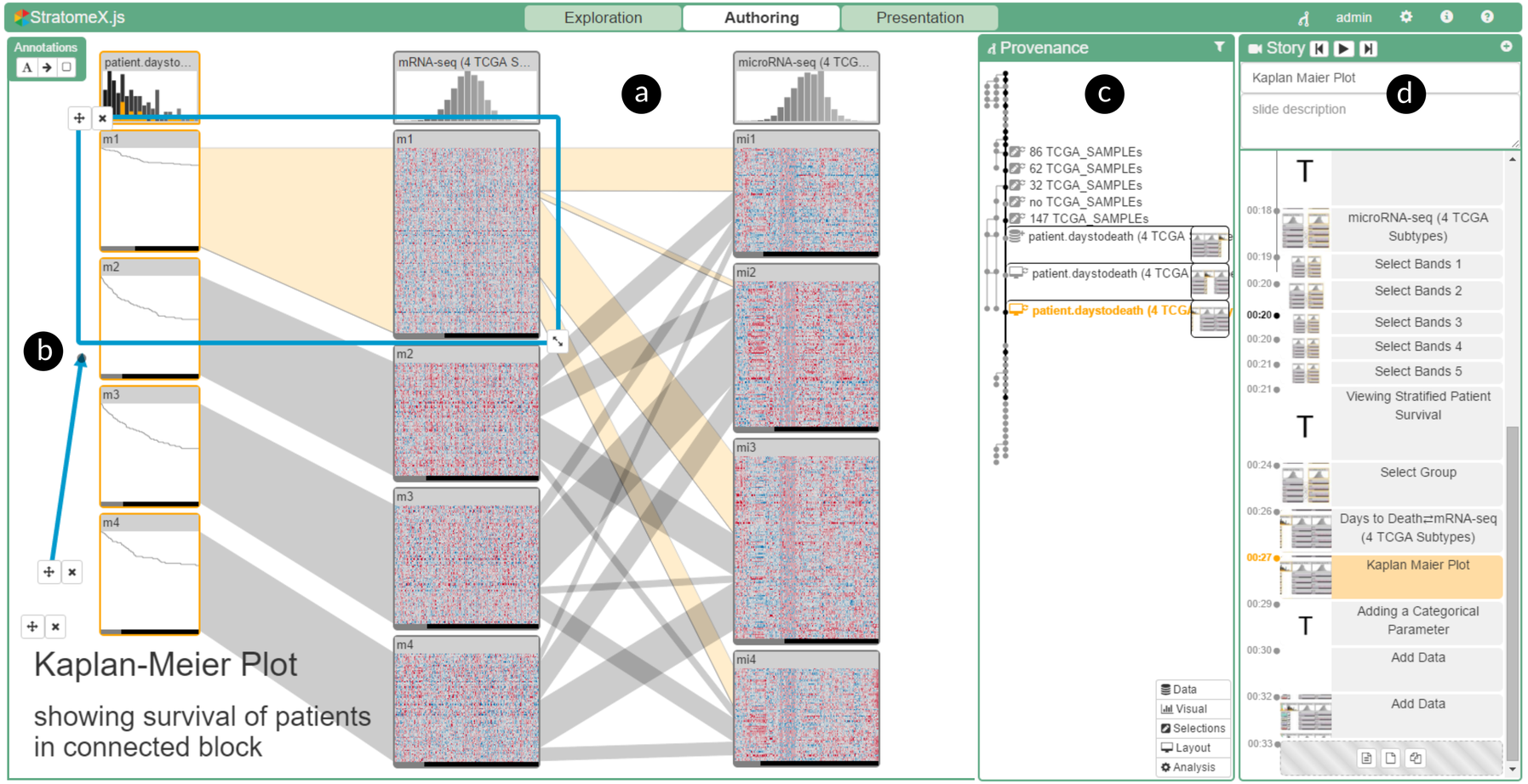
Alternative Splicing / mRNA-seq

Cancer Subtypes / Omics Clustering and Stratification





Reproducibility, Storytelling, Annotation, and Integration in Computational Workflows



Teaching Assistants

Daniel Hallman

Devin Lange

Shaurya Sahai

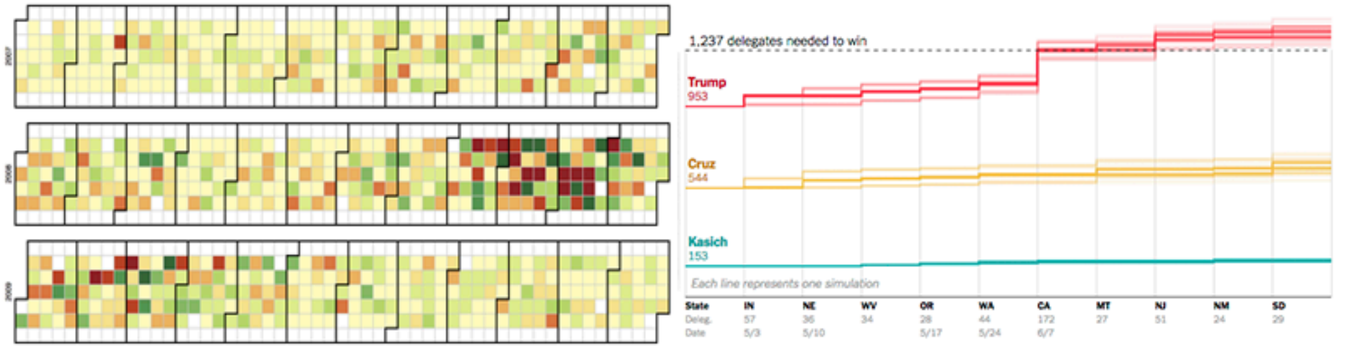
Course Structure

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 2020** 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!

The course begins by **bootstrapping your coding skills** (we will be using Python), and will move through a series of data science methods via real-life, project-based, lectures and computer labs. The goal of this course is to develop your skills in:

- **data wrangling**: how to acquire, clean, reshape, or sample data so that it's ready for further processing?

Communicate

Slack Team

https://datasciencecourse2021.slack.com/

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

Canvas

https://utah.instructure.com/courses/667313

Used only for homework submissions/grading

Github

https://github.com/datascience-course/2021-datascience-lectures

https://github.com/datascience-course/2021-datascience-homeworks

Used to post lectures and homework

Office Hours

See calendar on website

Tuesday/Thursday after class

Friday Morning

E-Mail

contact@datasciencecourse.net

alex@sci.utah.edu

little@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 GitHub
- 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

Online Lectures & COVID

No official spring break, but no lectures on March 9 & 11.

- Two week period (Feb 26–March 12) to complete 1-week homework.

Added dropped homework and activities (next slide).

Note that CR/NC designation can be used for general education credit. Might depend on your program.

Get in touch if you are sick or falling behind due to responsibilities related to COVID for extensions and accommodations.

5-Minute “Bio Break” each lecture.

How are you graded?

Homework Assignments: 45%

- Varying value, depending on length/difficulty
- Start early!
- Due on Fridays, late days: -1 point (10%) per day, up to two days.
- Lowest score will be dropped.

Final Project: 50%

- Teams, proposal & two milestones

Group Activities: 5%

- In class

Activities

In-class, 5-15 minutes mini-activities in breakout rooms.

Typically a coding problem.

1 person shares screen, all work together on the problem. Note names. That person shares solution with everyone to submit.

Submit the day of lecture by 6pm.

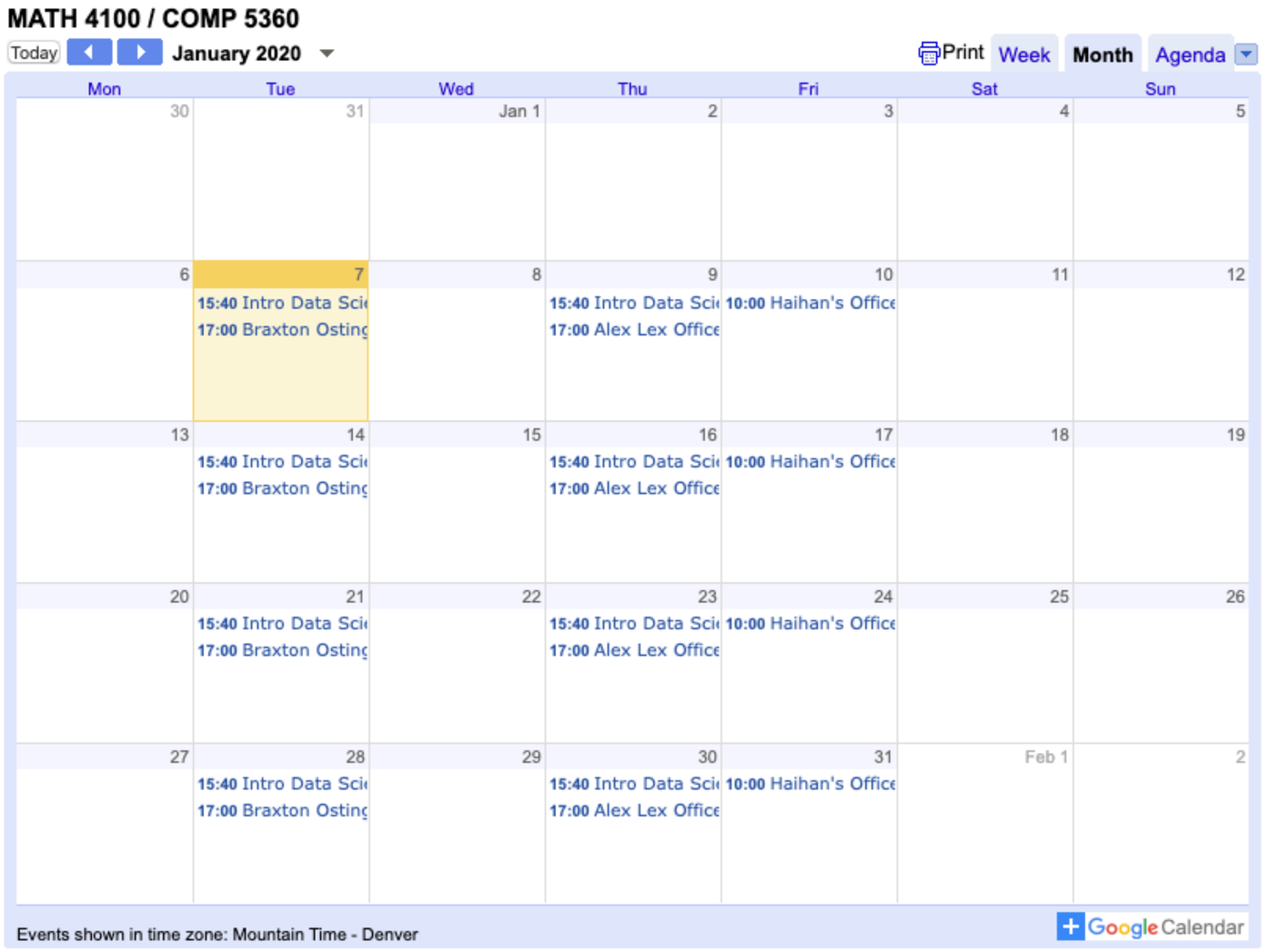
Complete 5 activities (out of ~10-15)

Schedule

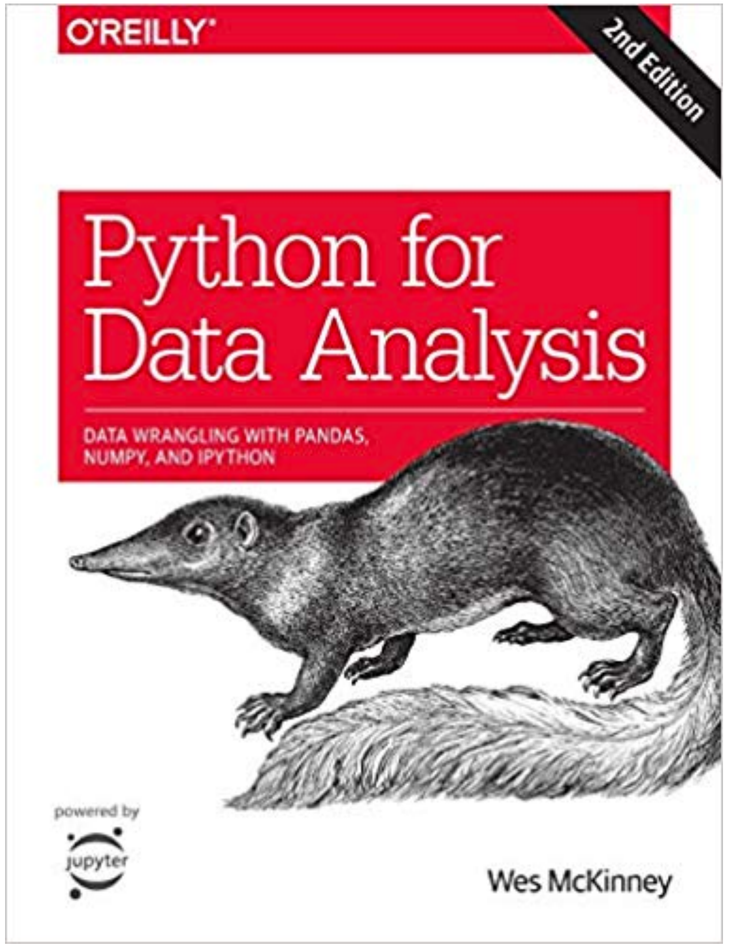
Lectures:
Tue / Th 12:25-1:45pm
Online via Zoom

Calendar
Link

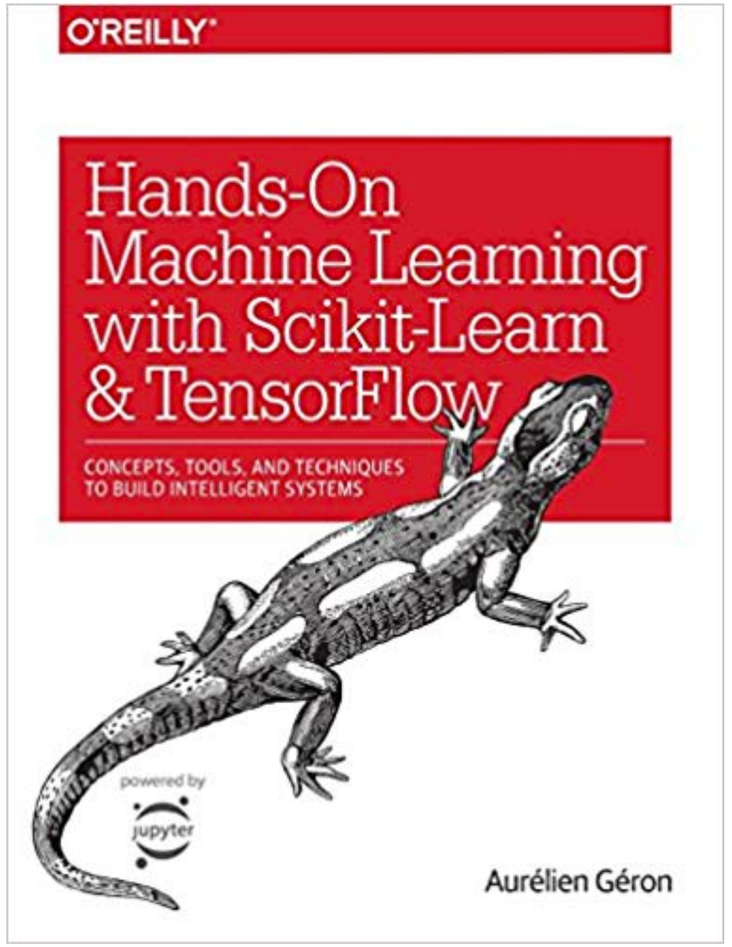
Lectures frequently involve computer activities.
Bring your own computer!
Have Python, etc installed (see HW0)



Books



Primary Text for Readings
Available for free on Campus: link

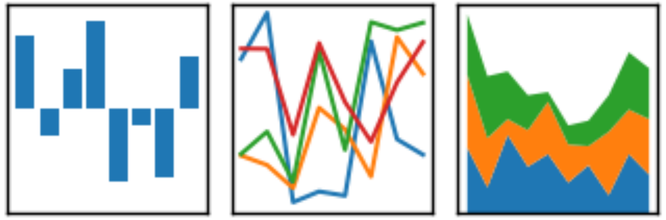


Supplementary Text
Available for free on Campus: link

Programming

pandas

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



TensorFlow

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 vs Math vs Sciences vs …

If in doubt, ask one of the instructors.

Code of Conduct

- We are committed to providing an inclusive and harassment-free environment in all interactions regardless of gender, sexual orientation, disability, physical appearance, race, or religion.
- We do not tolerate harassment in any form.
- Please report any harassment to us or the appropriate university office, which you can find at https://safeu.utah.edu/
- Please review the syllabus on these issues and the student code of conduct at https://regulations.utah.edu/academics/6-400.php

Cheating

You are welcome to **discuss** the course's ideas, material, and homework with others in order to better understand it, but **the work you turn in must be your own** (or for the project, yours and your teammate's). For example, you must **write your own code**, design your own visualizations, and critically evaluate the results in your own words.

You **may not submit the same or similar work** to this course that you have submitted or will submit to another. **Nor may you provide or make available solutions to homeworks to individuals** who take or may take this course in the future.

See also the SoC Academic Misconduct Policy:
http://www.cs.utah.edu/wp-content/uploads/2014/12/cheating_policy.pdf

You will fail the class if you cheat.

The misconduct will be reported to your home department.

We will **automatically check for plagiarism** in all your submissions.

Course Policies

Review Syllabus for:

Collaboration, Cheating and Plagiarism

Missed Activities and Assignment Deadline

Late Policies

Regrading Policies

Respect for Diversity

American with Disabilities Act

Sexual Misconduct

Student Name and Personal Pronoun

This Week

HW0, including course survey

- Make sure to complete this before class on Thursday. Use 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/2021-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 10th.

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.

Github

Github is a web-based hosting service for version control using git.
We'll discuss git and github in a later lecture.
The basics are described in the README.md file/

README.md

Introduction to Data Science - Homeworks

Course website: http://datasciencecourse.net

This repository will contain directories with all homeworks. You can manually download the files for each homework, but we recommend that you use git to clone and update this repository.

You can use GitHub Desktop to update this repository as new homeworks are published, or you can use the following commands:

Initial Step: Cloning

When you clone a repository you set up a copy on your computer. Run:

```
git clone https://github.com/datascience-course/2019-datascience-homeworks
```

This will create a folder `2019-datascience-homeworks` on your computer, with the individual homeworks in subdirectories.

Updating

As we release new homeworks, or if we discover mistakes and update an already released homework description, you'll have to update your repository. You can do this by changing into the `2019-datascience-homeworks` directory and executing:

```
git pull
```

That's it - you'll have the latest version of the homeworks.

Next Week

Data Structures and Pandas

Introduction to Descriptive Statistics

HW1 due

Current Enrollment

Math 4100: 51

COMP 5360: 63

Trouble enrolling? send an email

Breakout Rooms!

Get to know each other in small groups.

Some questions:

First year for anyone?
Who is an undergraduate?
Who is a MS student?
Who is a PhD student?

Math? Biology? Other Sciences? Engineering? Humanities? Business? Other?

Who knows Python?
R?
Matlab?
C / C++?
Java?
Other languages?

Who has programmed for 1 year? 5 years? More?