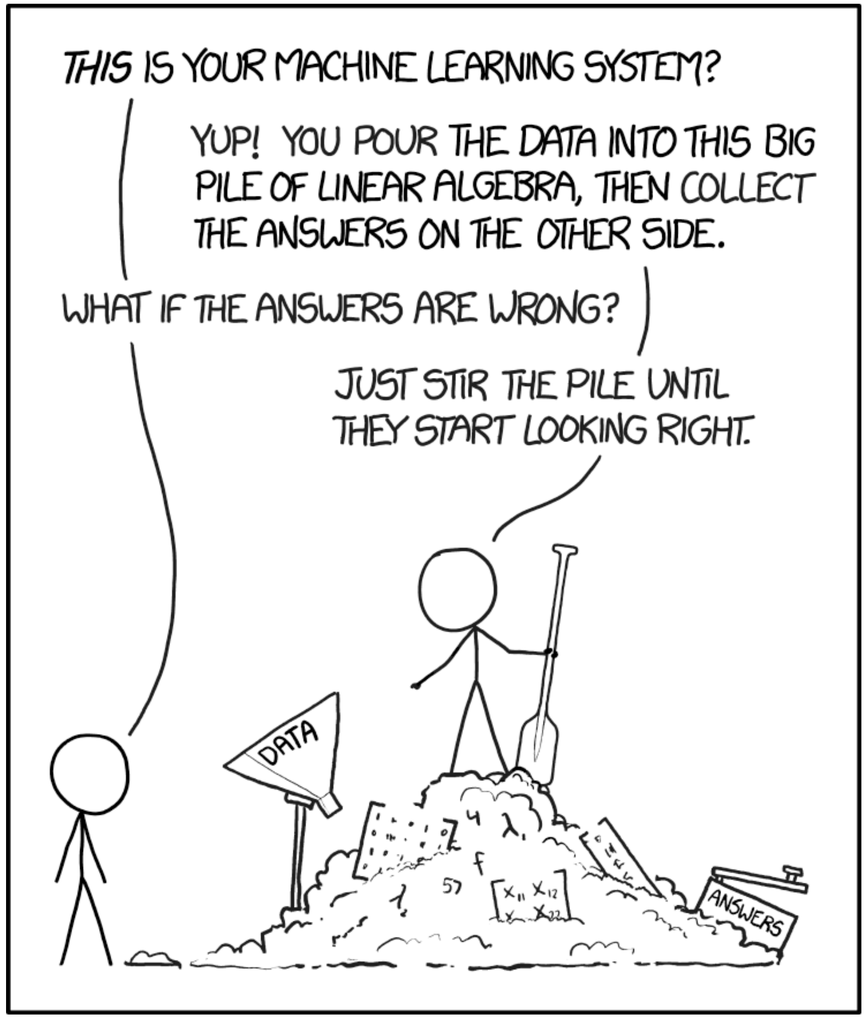
# Introduction to Data Science
# CS 5360 / Math 4100


Kate Isaacs
u6044649@gcloud.utah.edu

THE UNIVERSITY OF UTAH

Anna Little
little@math.utah.edu


[xkcd]

# In-Person Format

Course meets in person in WEB L101.

Lectures are not recorded. Lectures from previous years are available on the course websites, for example last year's https://datasciencecourse.net/2022 (old site)

In person attendance typically* not required, but you are highly encouraged to attend in real time due to class activities.

# Logistics

See CANVAS Module “Important Online Links” for links to Slack and office hours.

We will use SLACK for course discussions and Q&A, so please sign up. Invite link is on CANVAS.

All assignments will be submitted and graded in CANVAS, but see the course webpage for syllabus, schedule, and detailed course information.

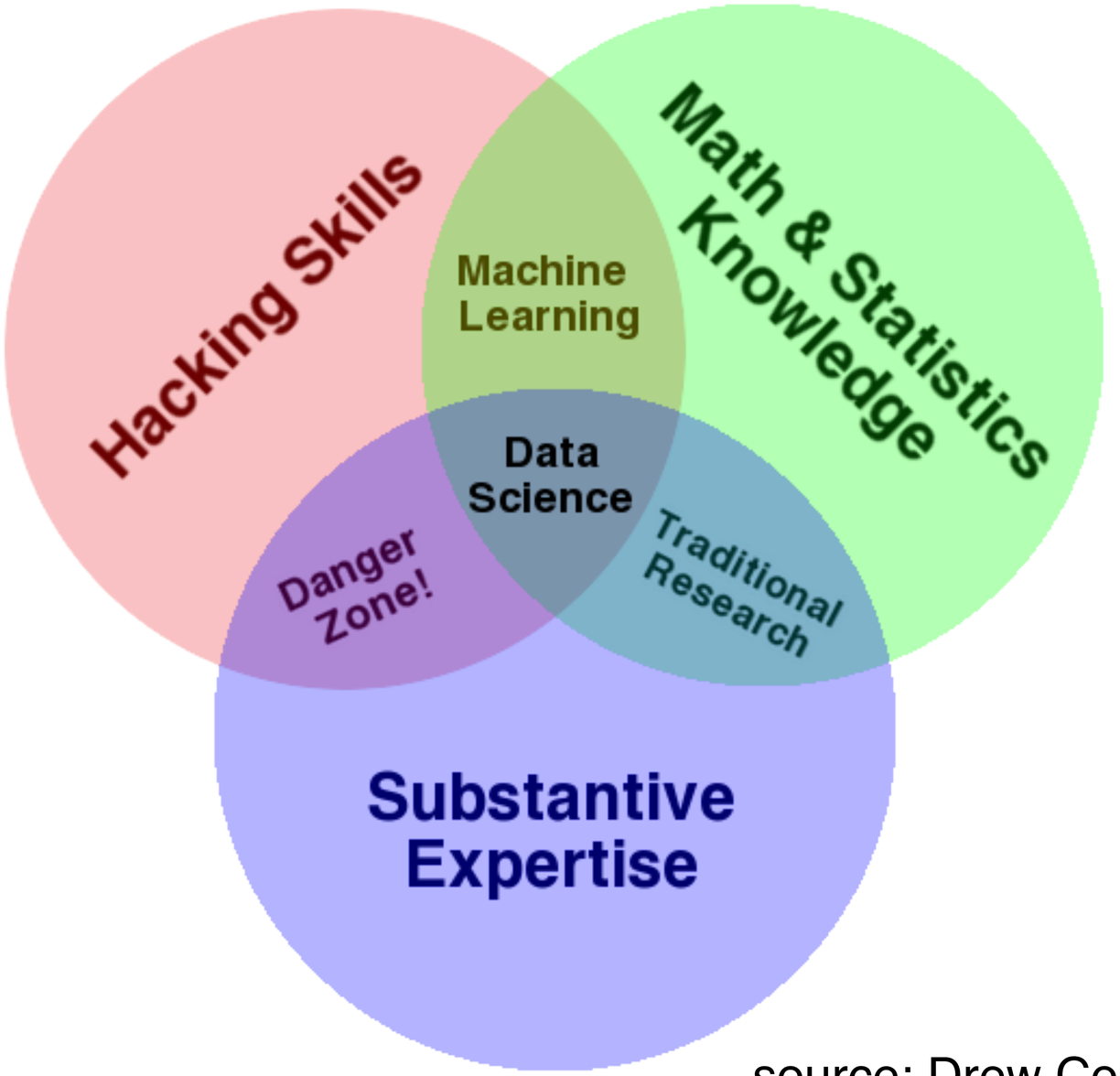
# What is Data Science?

- 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.

# 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.

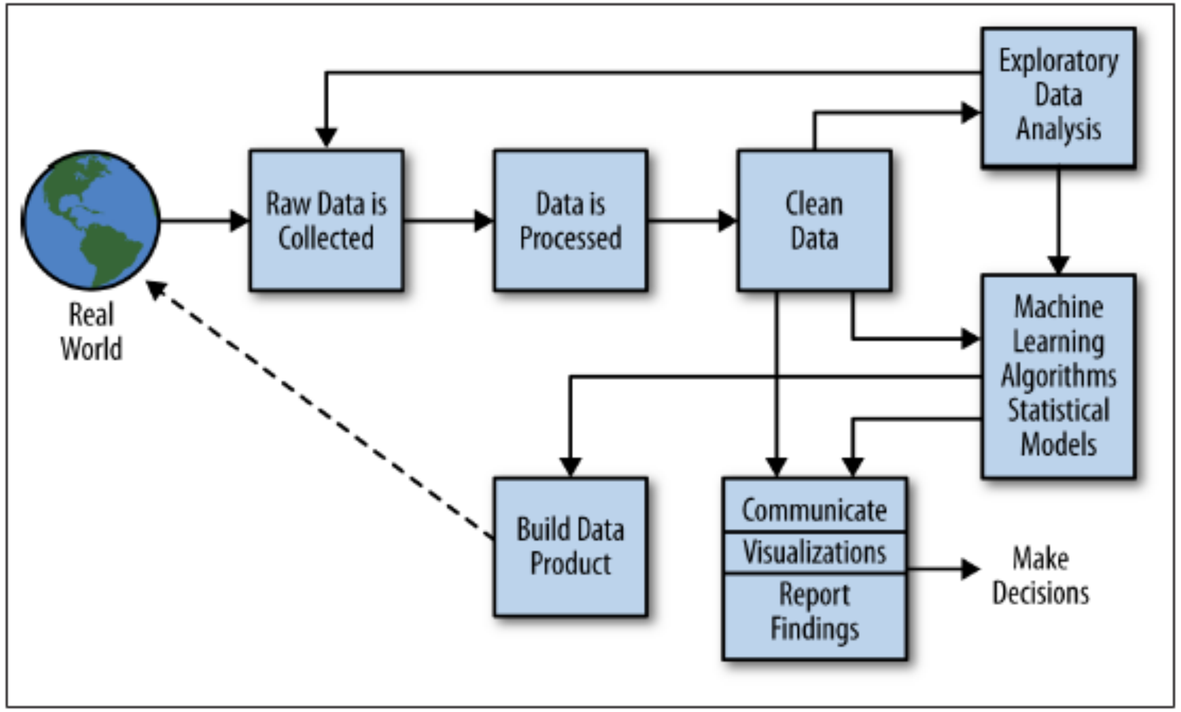



*Figure 2-2. The data science process*

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!

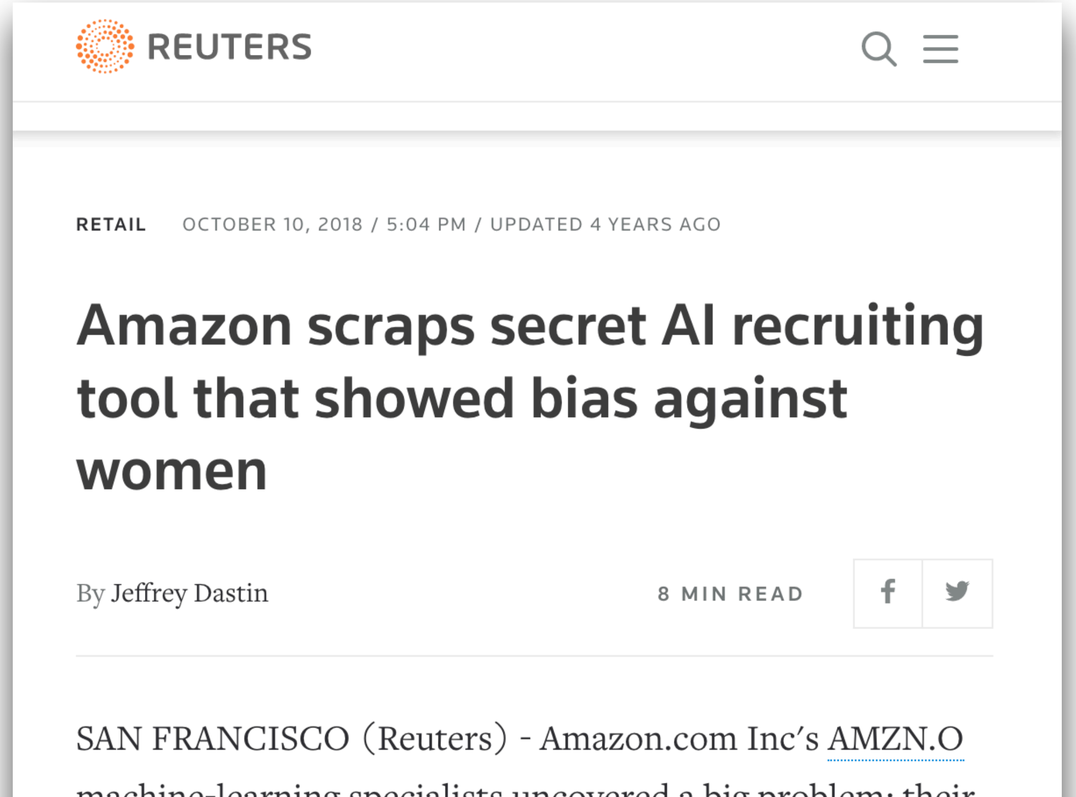
Medicine? Data-centered & computational!

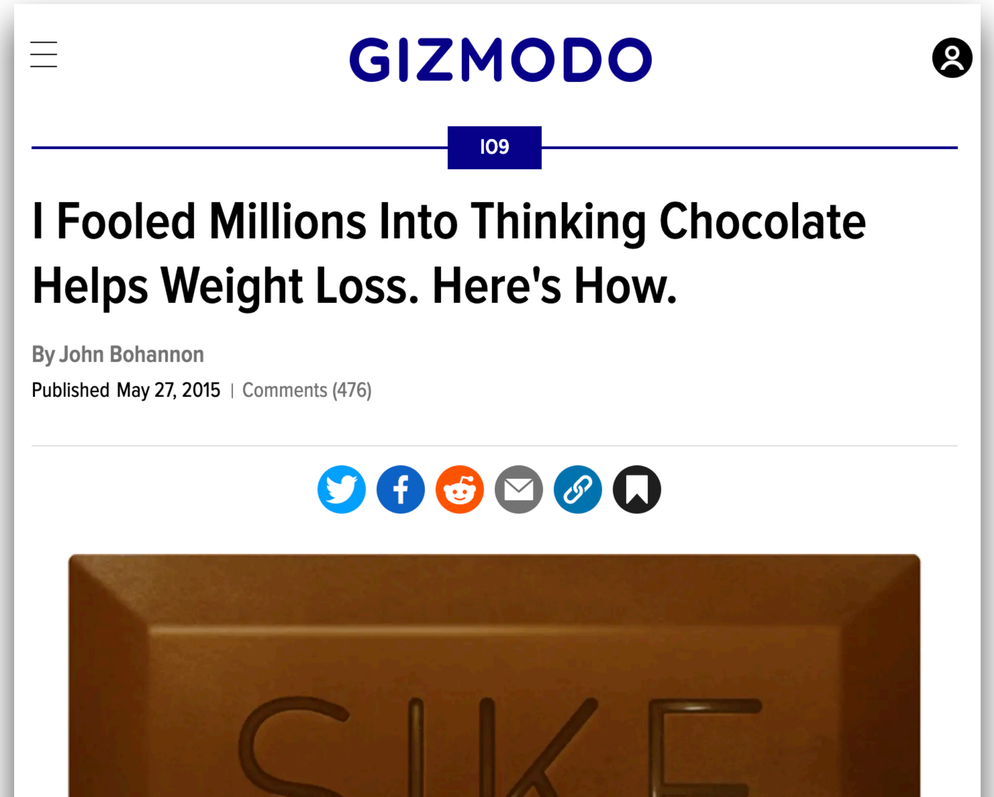
Social Sciences? Data-centered & computational!

Business? Data-centered & computational!

# Why do we care? Data science is everywhere!

…and it's not always used right. We need to understand the validity of our results.

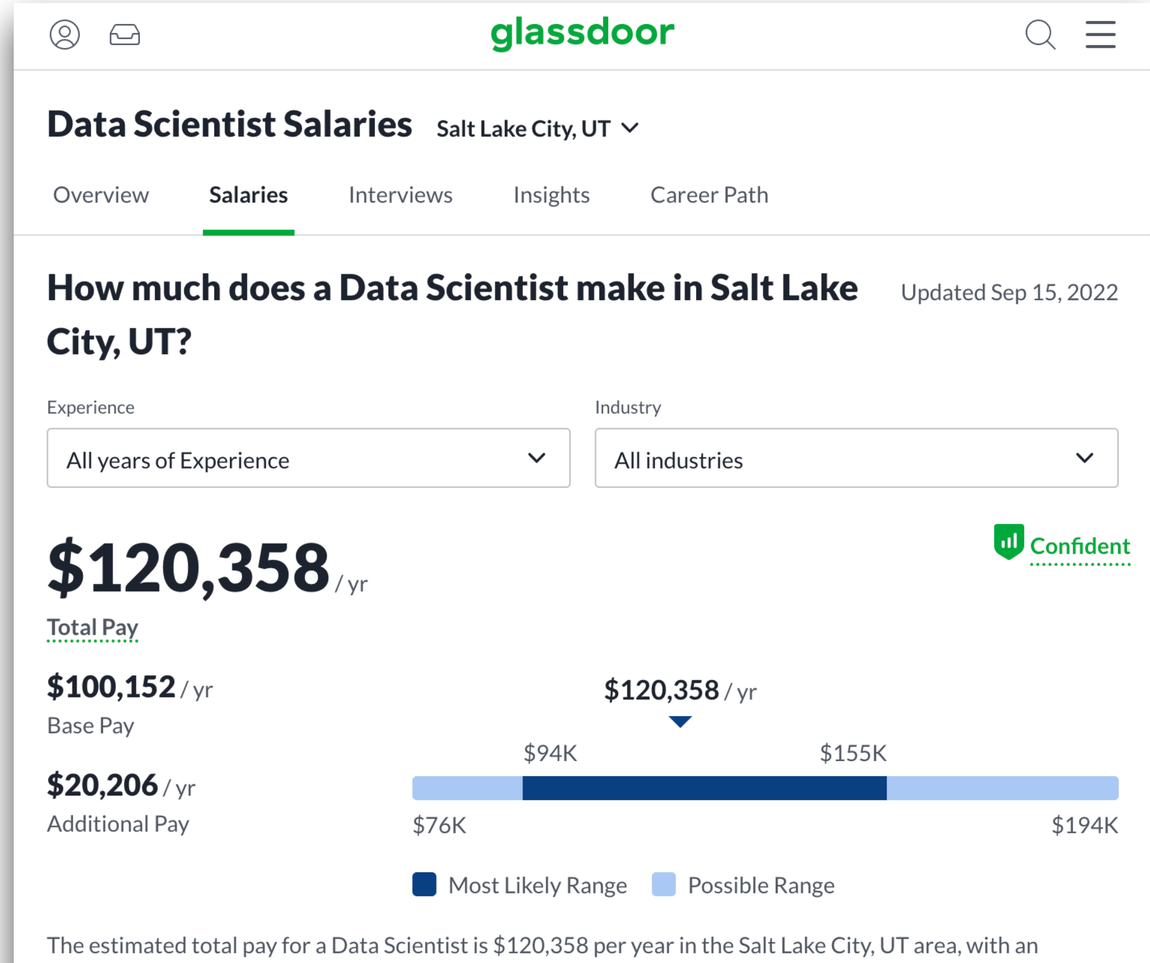


# Why do we care? Jobs!

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

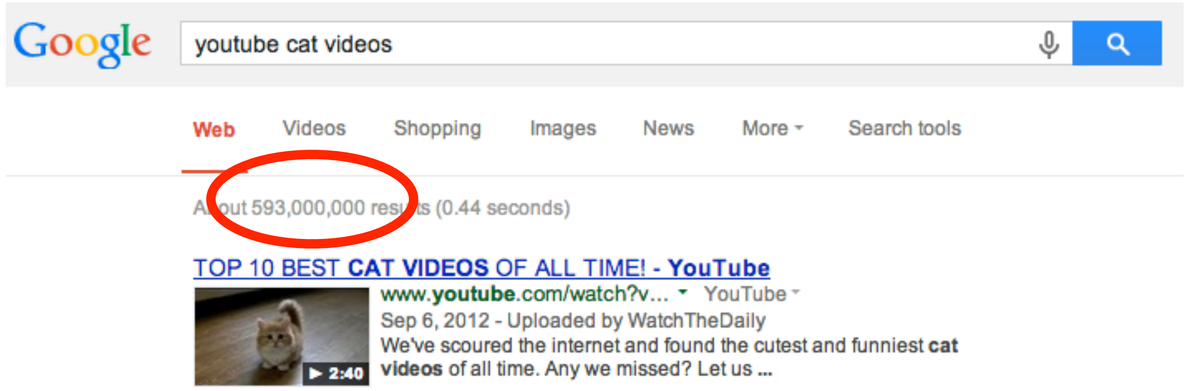
# Big Data

2010: 1,200 exabytes, largely unstructured

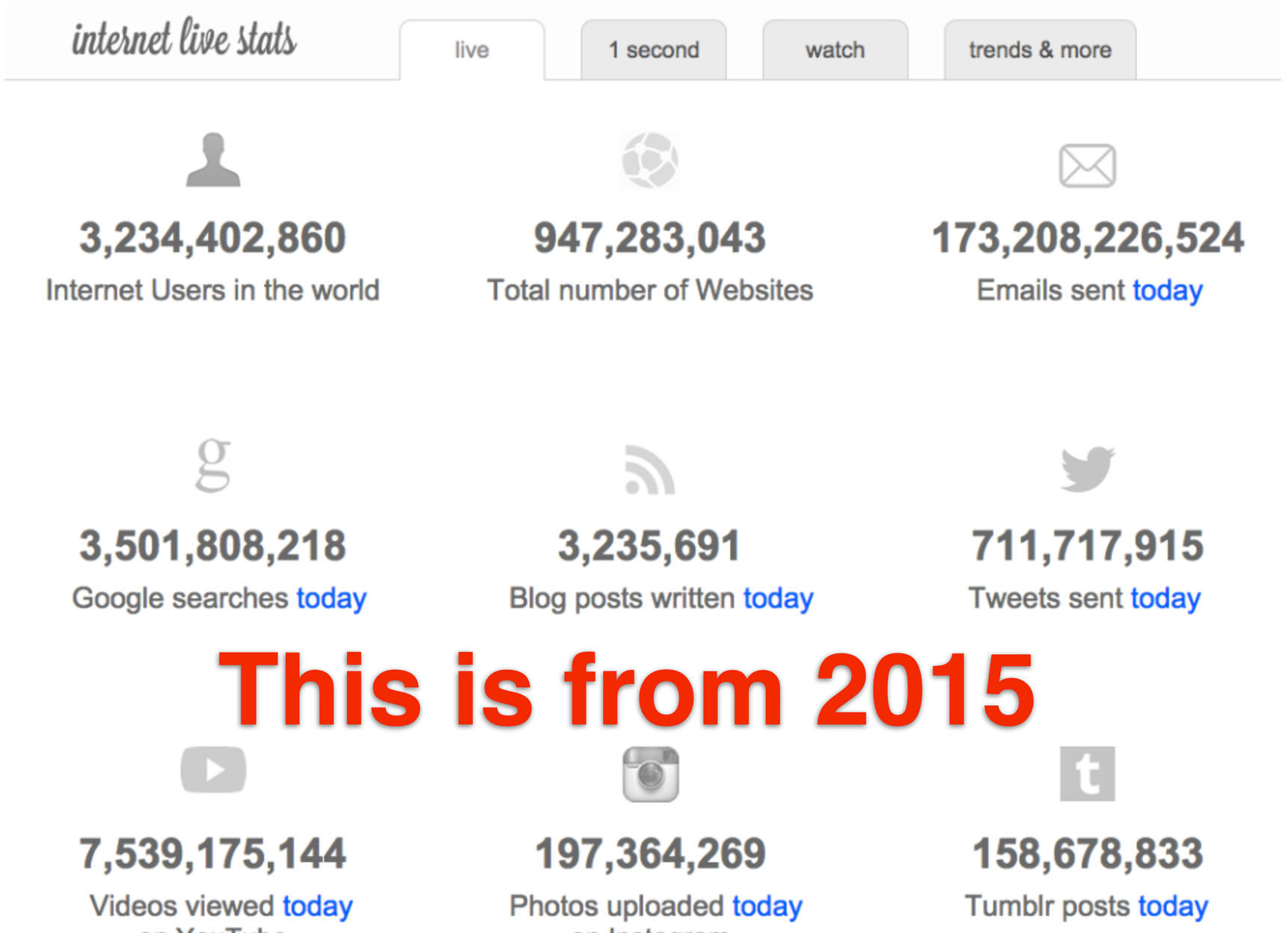
Google stores ~10 exabytes (2013)

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


internet live stats
live
1 second
watch
trends & more
3,234,402,860
Internet Users in the world
947,283,043
Total number of Websites
173,208,226,524
Emails sent today
3,501,808,218
Google searches today
3,235,691
Blog posts written today
711,717,915
Tweets sent today
This is from 2015
7,539,175,144
Videos viewed today
on YouTube
197,364,269
Photos uploaded today
on Instagram
158,678,833
Tumblr posts today

# How can we leverage data?

Improve your fitness by targeted training

Improve your product

- by targeting your audience
- by considering semantics

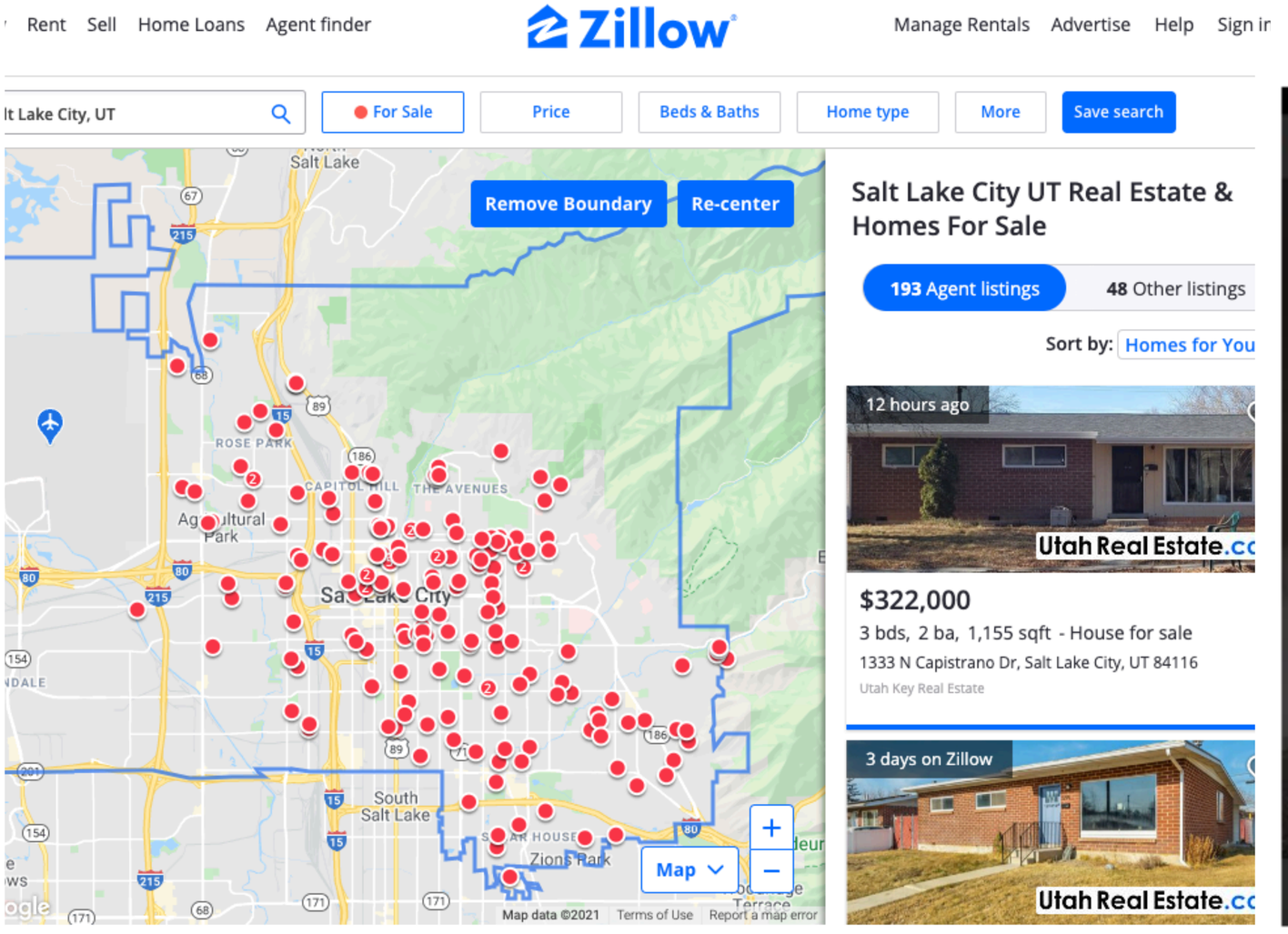
Make better decisions

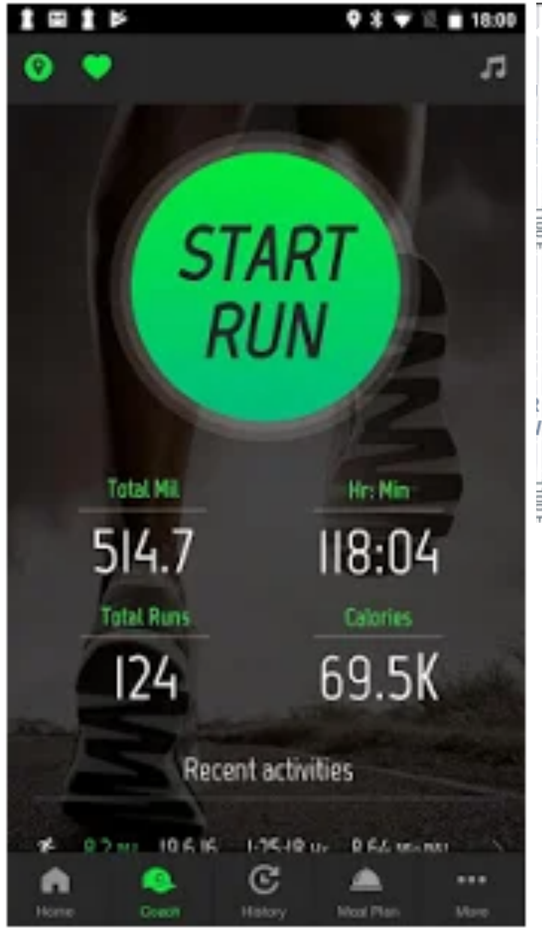
- exact diagnosis, choose right medication, pick good restaurant

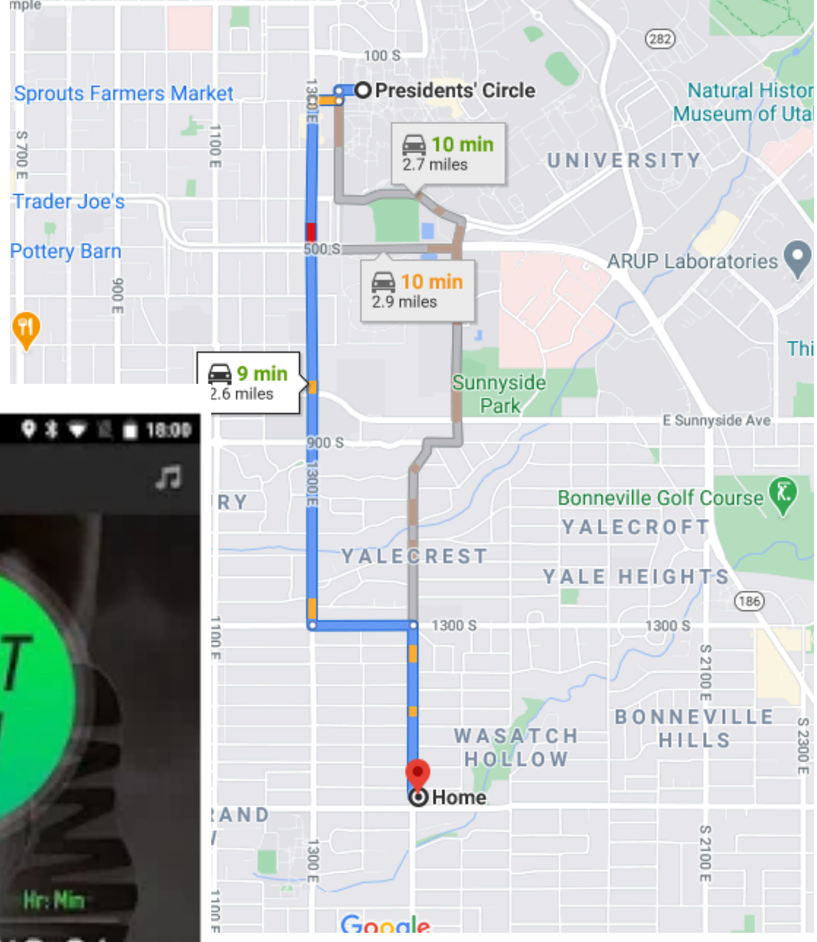
Predict elections, events, crowd behavior, etc.

… and many more applications

# Example: Personal Data



# Big Data in Science and Engineering

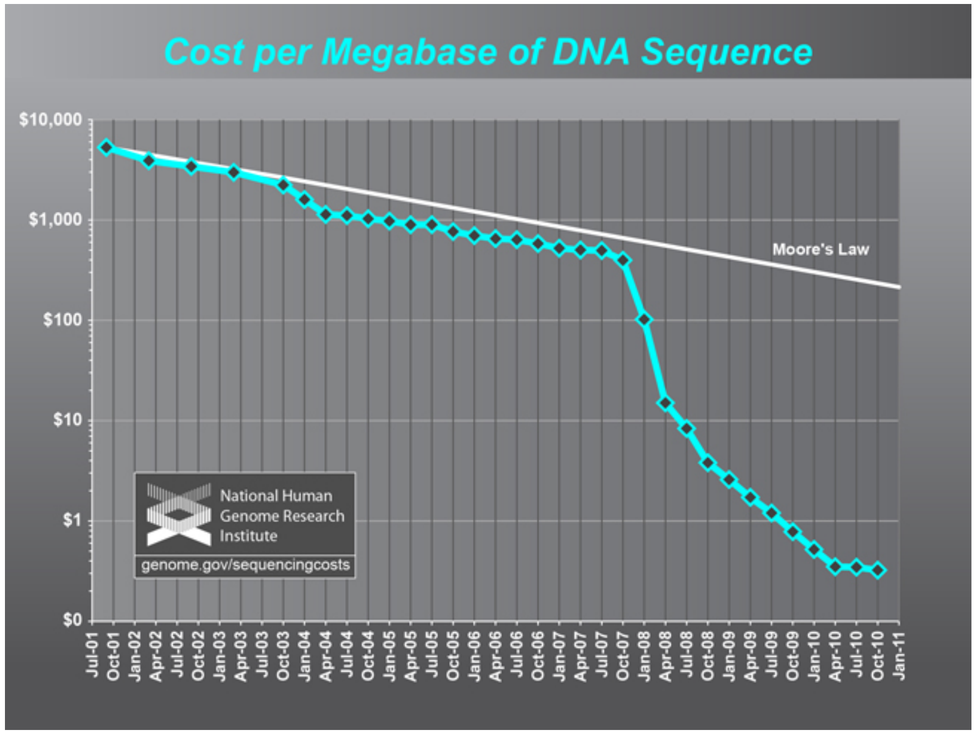
Big Data transformed science and engineering.

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*

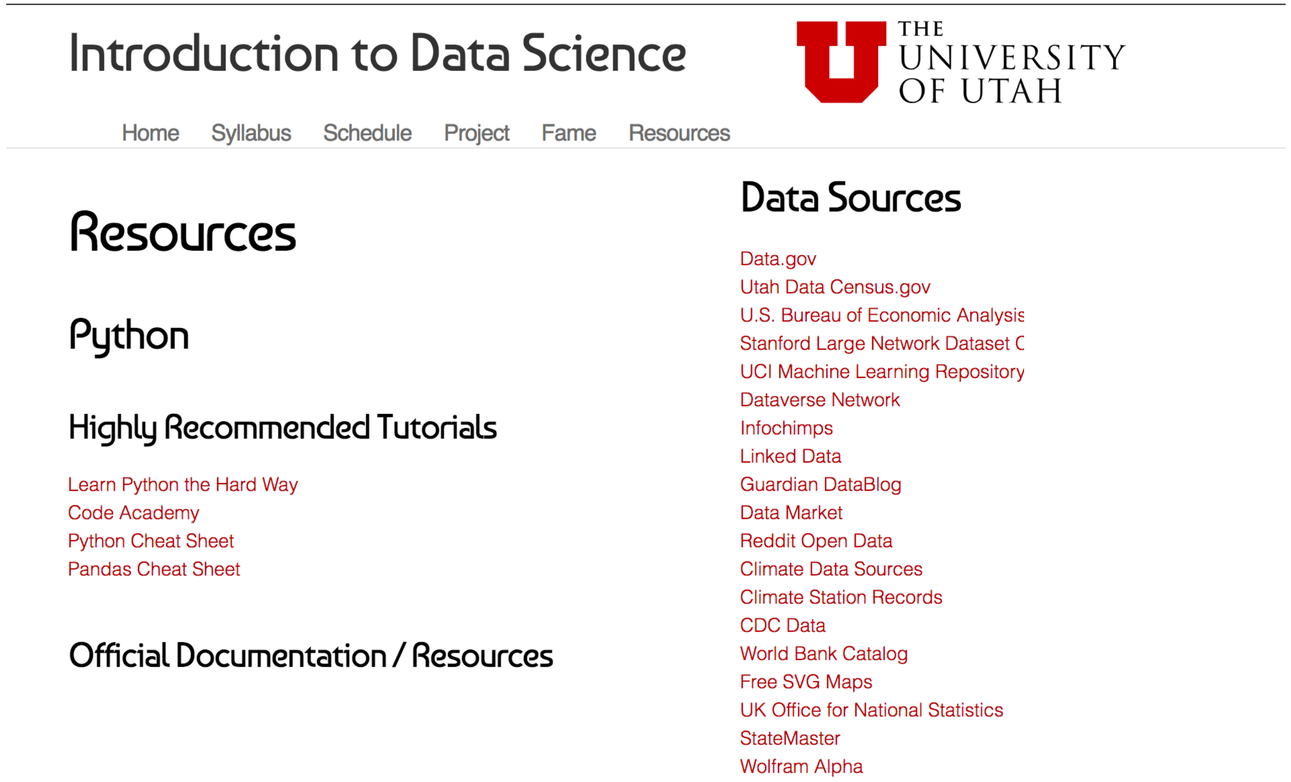

# NSA Utah Data Center (Bluffdale, Utah)

Storage Capacity?

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

# 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**

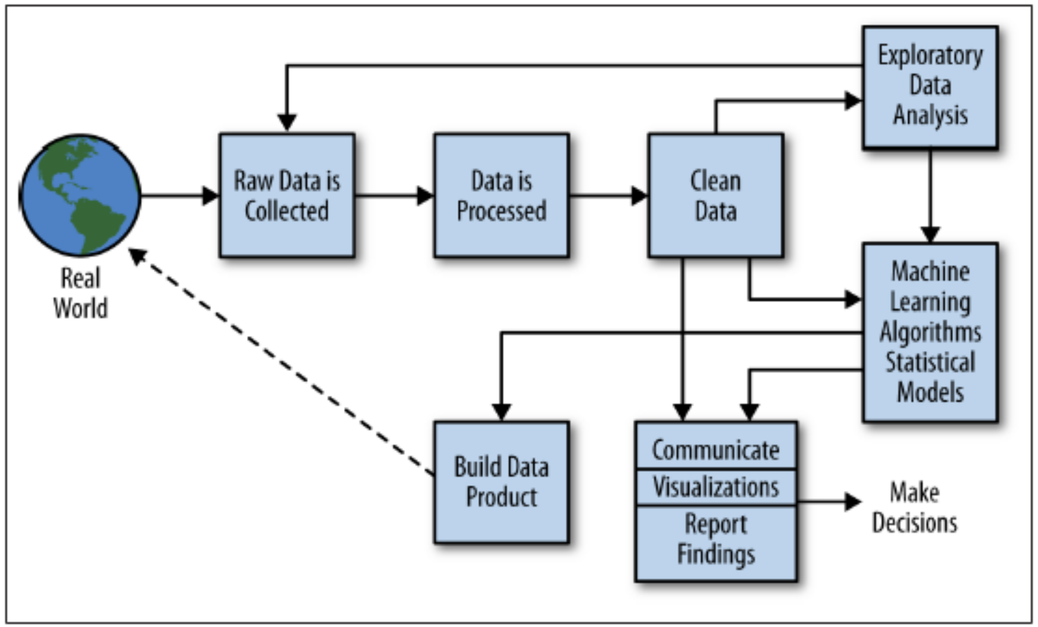



*Figure 2-2. The data science process*

# Topics

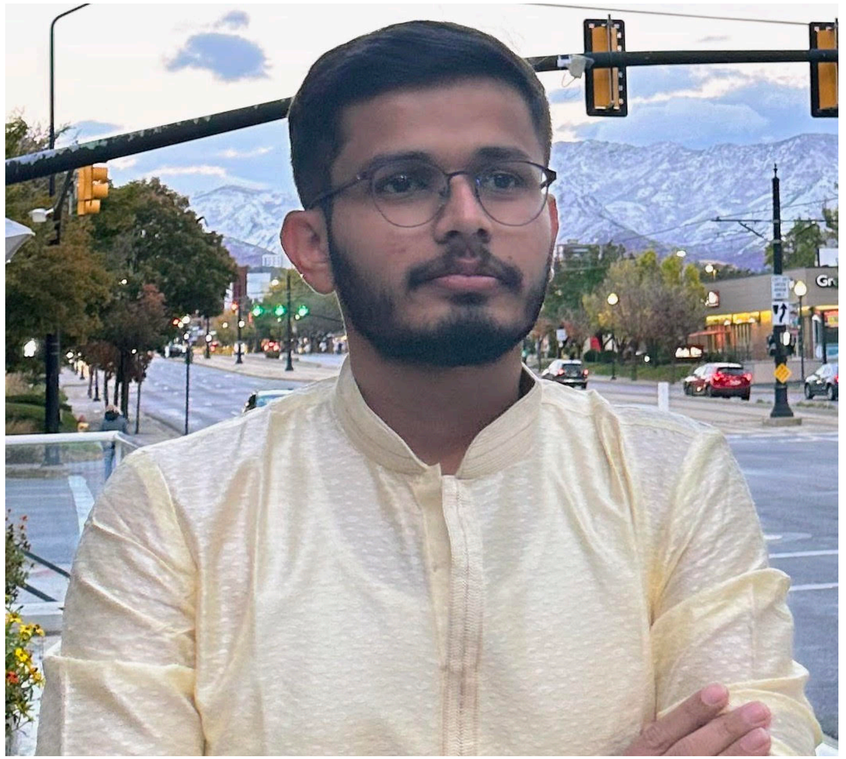
- Programming
- Version Control
- Data Wrangling (Pandas)
- Data Acquisition
  - Web Scraping
  - Web APIs
  - Databases
- Basic Stats
- Hypothesis Testing
- Visualization
- Regression
- Classification
  - Logistic Regression, K-Nearest Neighbors, SVM, Decision Trees, Neural Networks
- Clustering
- Dimensionality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

# Teaching Staff
# CS 5360 / Math 4100

# Teaching Assistants

CS MS

Arpit Patil



CS MS

Harsh Mahajan

# Anna Little

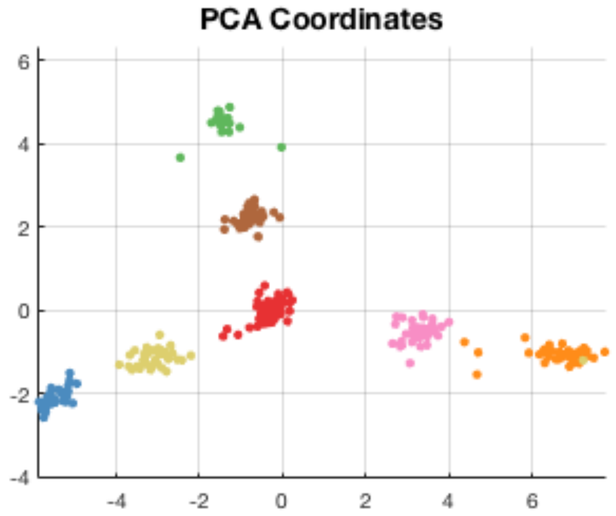
Assistant Professor, Mathematics

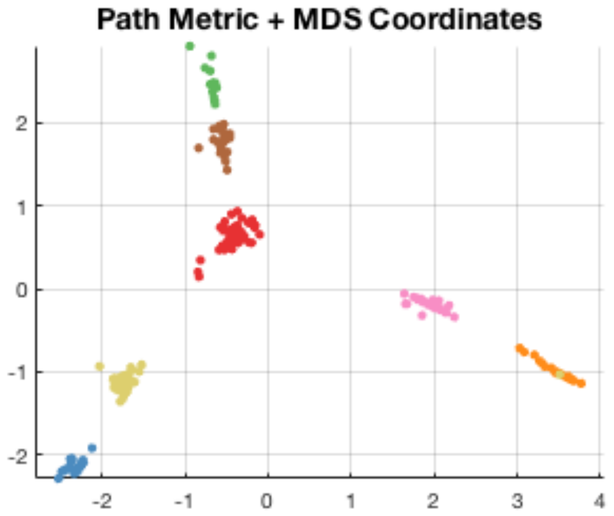
Before that: Postdoc, Michigan State University

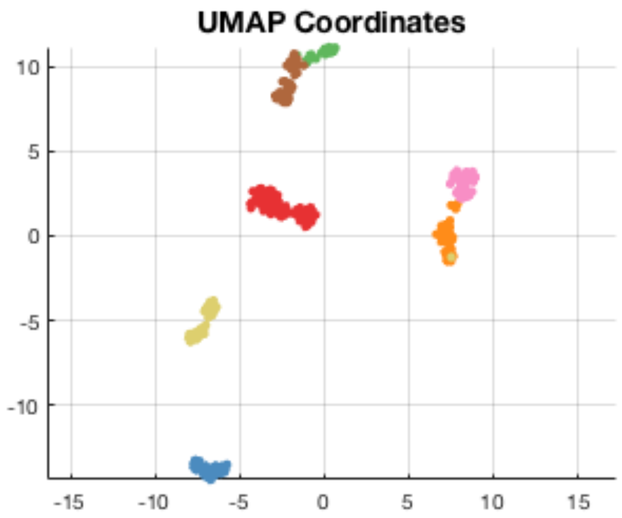
PhD in Mathematics, Duke University

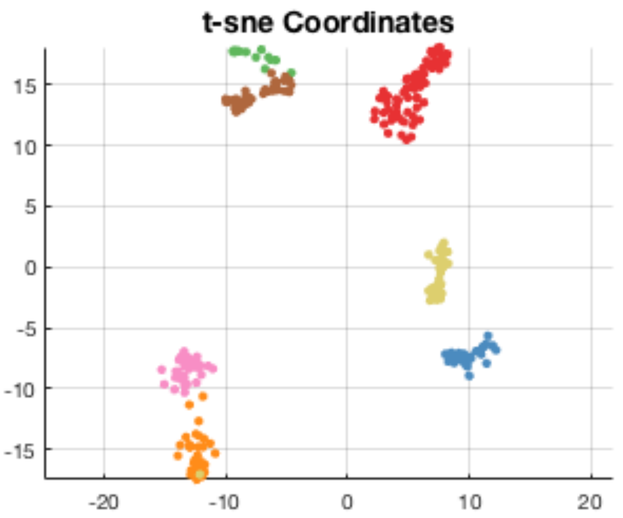
https://www.anna-little.com/

# Data Analysis, Dimension Reduction & Estimation

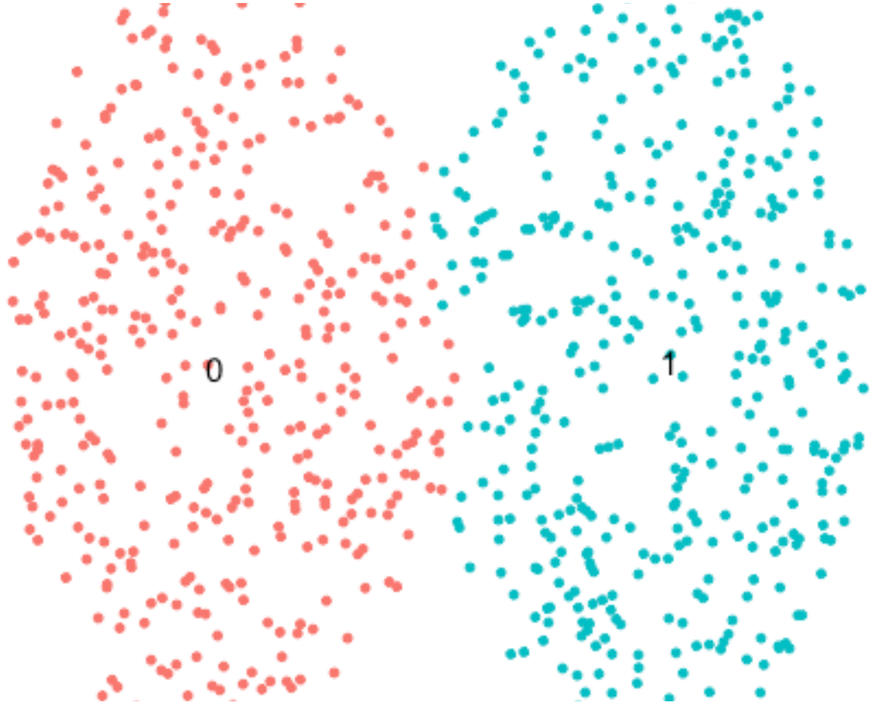


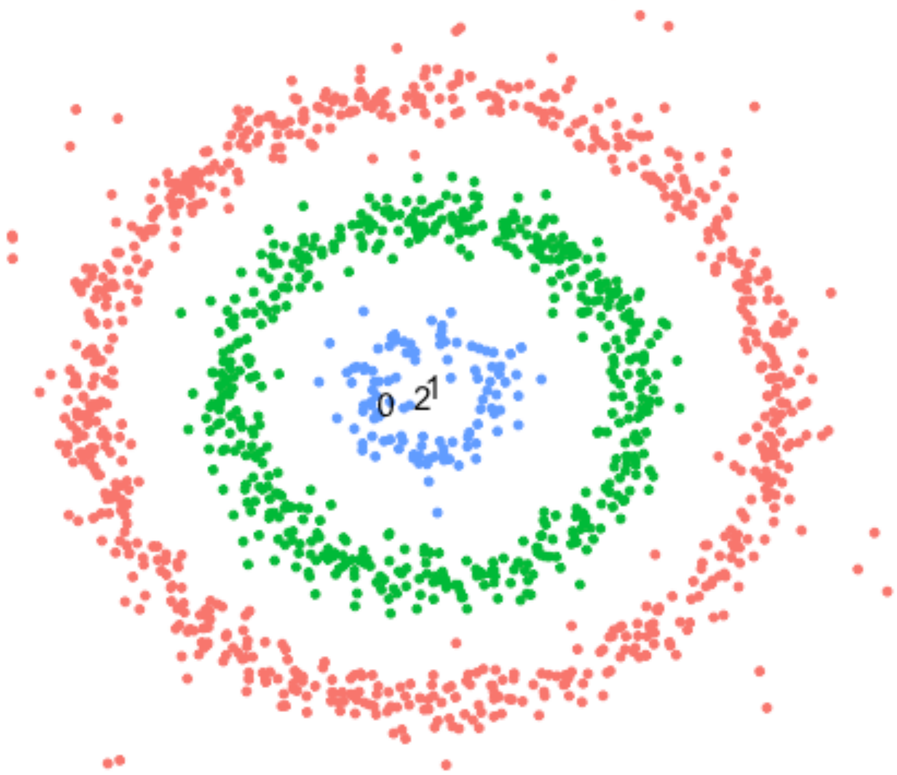



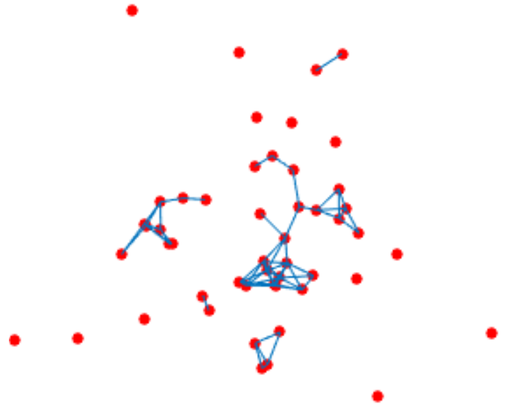
# Clustering and Notions of Similarity



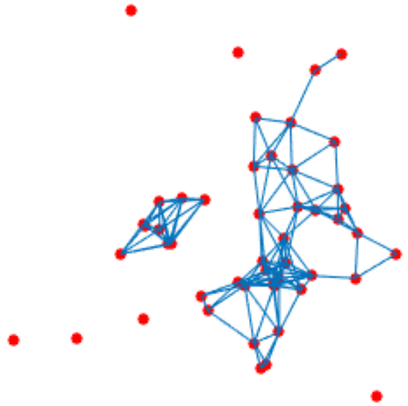



# Graph-based Methods, Fast Computation

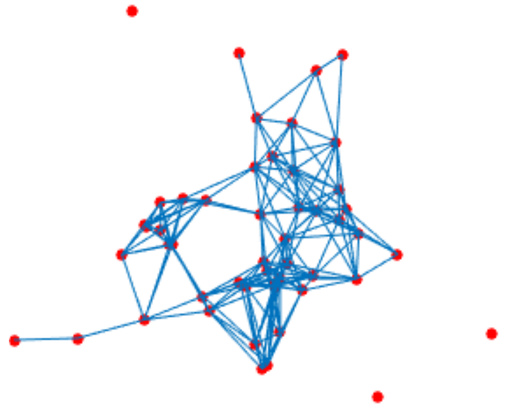
$G_1$: All Edges $< 0.36$



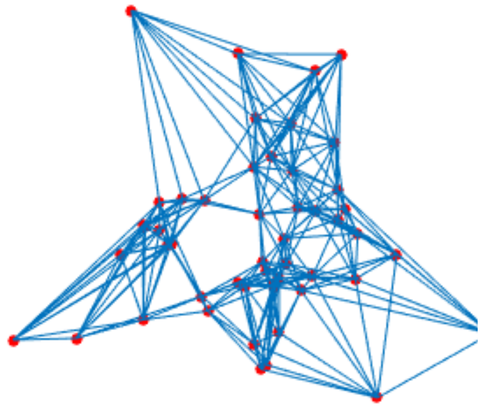
$G_2$: All Edges $< 0.56$



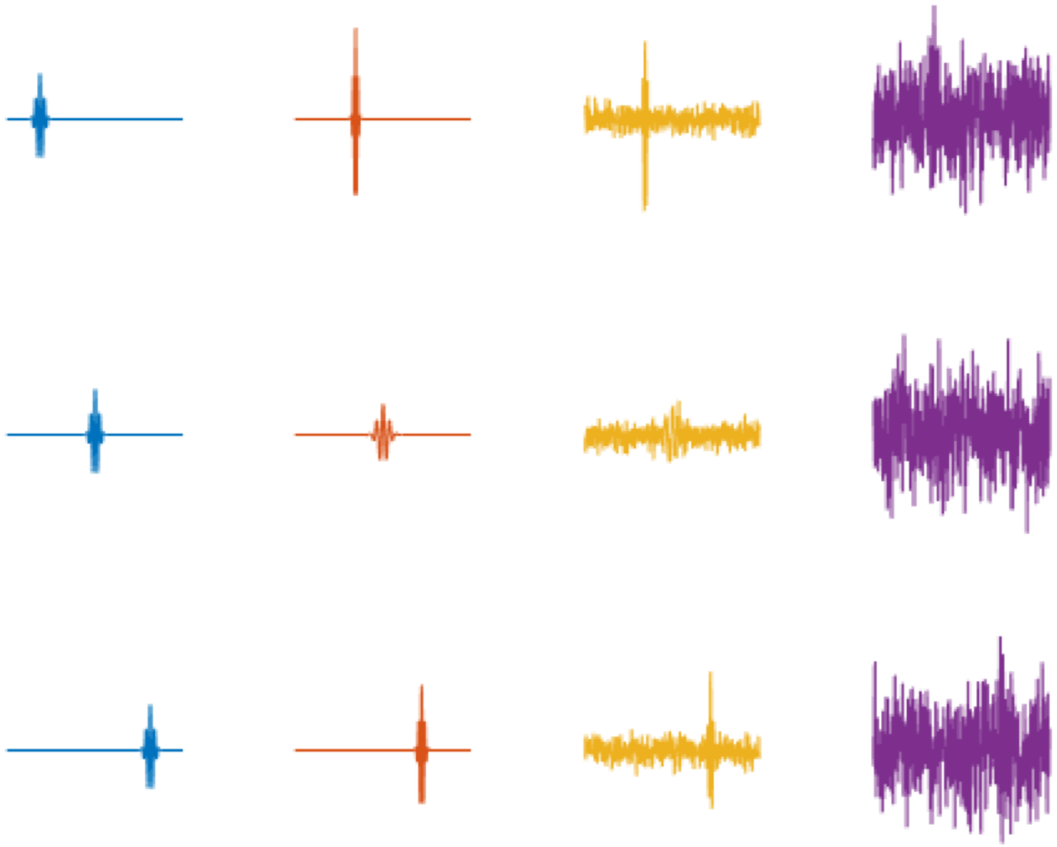
$G_3$: All Edges $< 0.8$

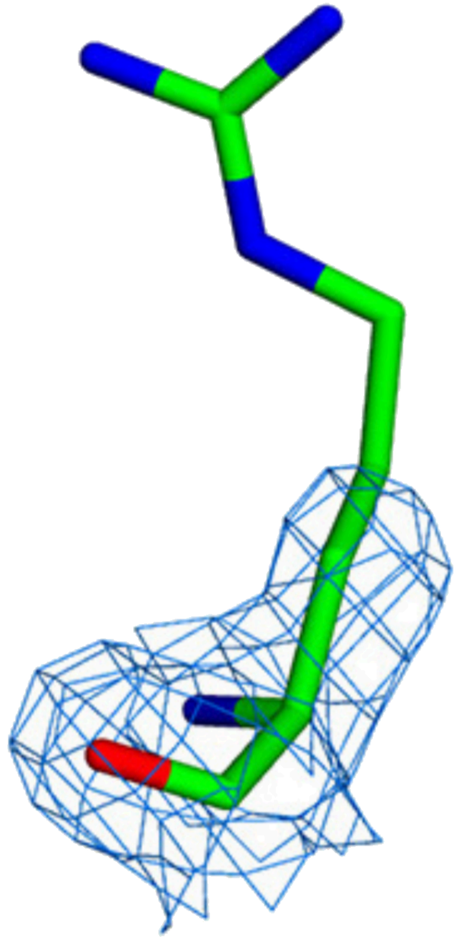


$G_4$: All Edges

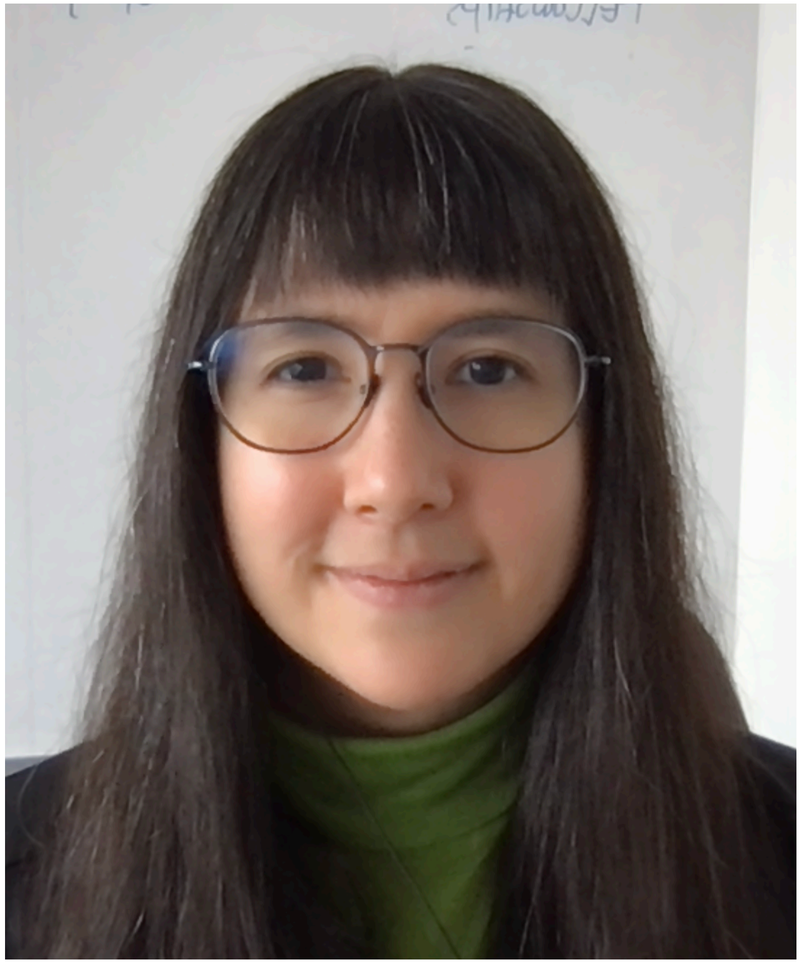


# Multi-reference Alignment, Signal Processing





# Kate Isaacs



Associate Professor, Computer Science

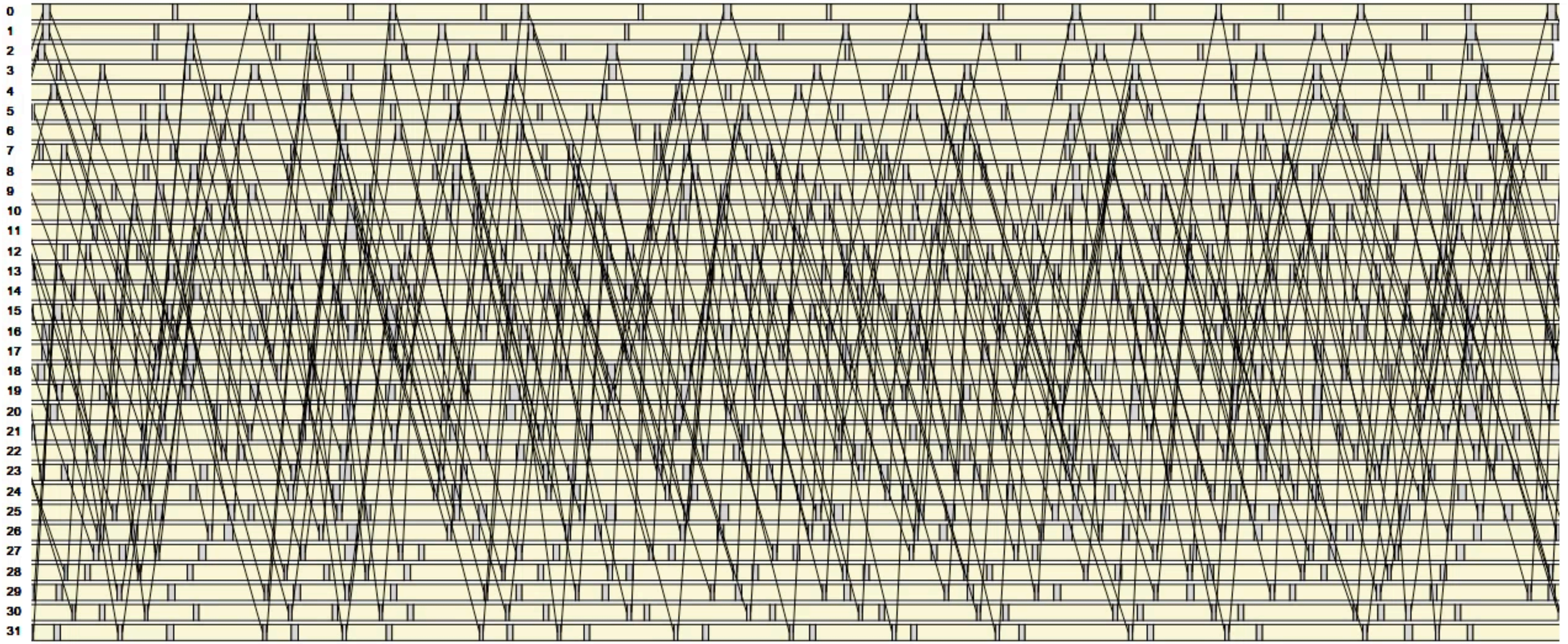
Faculty, Scientific Computing and Imaging (SCI) Institute

PhD in CS, UC Davis

http://hdc.cs.arizona.edu/people/kisaacs/

Data visualization, high performance computing

# Interactive data visualization — parallel timelines



# Interactive data visualization: Designing for notebooks

```
In [3]: vis_in
```

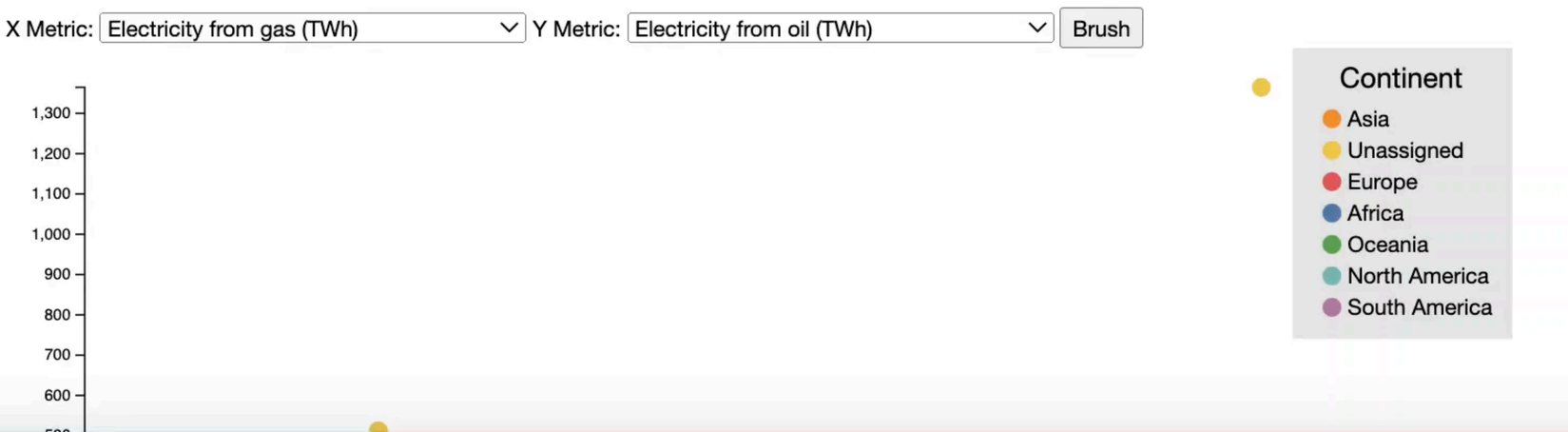
Out[3]:

| | Entity | Code | Year | Electricity from coal (TWh) | Electricity from gas (TWh) | Electricity from hydro (TWh) | Electricity from other renewables (TWh) | Electricity from solar (TWh) | Electricity from oil (TWh) | Electricity from wind (TWh) | Electricity from nuclear (TWh) | Continent |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | AFG | 2019 | 0.000000 | 0.344000 | 1.050000 | 0.000000 | 0.040000 | 0.000000 | 0.000100 | 0.000000 | AS |
| 1 | Africa | NaN | 2019 | 262.470414 | 340.472820 | 132.734791 | 8.184247 | 19.377978 | 92.816685 | 17.508052 | 15.913439 | OTH |
| 2 | Albania | ALB | 2019 | 0.000000 | 0.000000 | 8.466480 | 0.000000 | 0.022000 | 0.163560 | 0.000000 | 0.000000 | EU |
| 3 | Algeria | DZA | 2019 | 0.000000 | 73.615787 | 0.646000 | 0.000000 | 0.615000 | 0.000000 | 0.029000 | 0.000000 | AF |
| 4 | American Samoa | ASM | 2019 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.187000 | 0.000000 | 0.000000 | OC |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 230 | Western Sahara | ESH | 2019 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.090000 | 0.000000 | 0.000000 | OTH |
| 231 | World | OWID_WRL | 2020 | 9345.340997 | 5943.112329 | 4355.041636 | 702.887552 | 844.385951 | 1364.578206 | 1590.189440 | 2720.674471 | OTH |
| 232 | Yemen | YEM | 2019 | 0.000000 | 3.058733 | 0.000000 | 0.000000 | 0.500000 | 6.339360 | 0.000000 | 0.000000 | AS |
| 233 | Zambia | ZMB | 2019 | 1.998501 | 0.000000 | 13.902000 | 0.080073 | 0.002432 | 1.003232 | 0.000000 | 0.000000 | AF |
| 234 | Zimbabwe | ZWE | 2019 | 4.614000 | 0.000000 | 5.741000 | 0.206667 | 0.014000 | 0.000000 | 0.000000 | 0.000000 | AF |

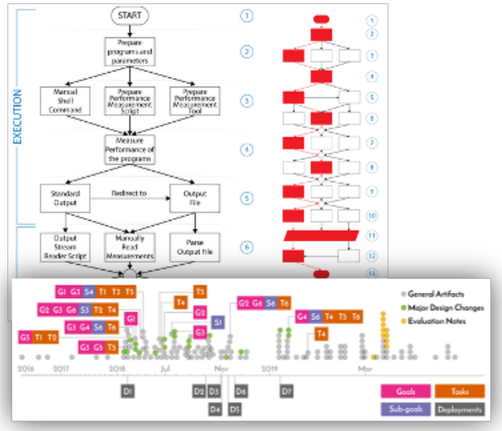
235 rows × 12 columns

We can use the basic functionality in roundtrip to pass this data into our scatterplot as a DataFrame (or other complex datatype). The visualization developer -- whoever built scatterplot -- will manage the necessary conversions.

```
In [4]: %scatter_plt vis_in
```



# Vis Research Methodology





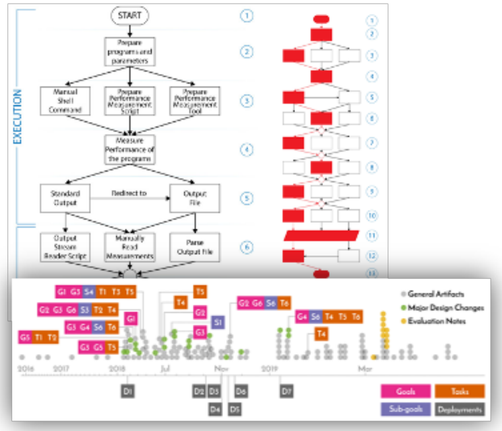


## Vis Design Studies
**(InfoVis 2019, Ongoing)**

| Contribution Type | Description |
|---|---|
| Algorithmic simplification | simplification or explanation of techniques and algorithms to make them easier to understand and reproduce |
| Artistic design | practices and evaluative reflection on expressiveness of visualization |
| Data abstraction | mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature |
| Data structure | improvements or new uses of existing ones to support visualization |
| Dataset | publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation |
| Deployment | discussion of insights gained from real world deployment of a tool or technique |
| Design methodology | methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches |
| Evaluation methodology | methodologies that enable new ways in evaluating visualization solutions |
| Formalism | generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts |

## Contribution Types
**(CG&A 2019)**

# Vis Research Methodology




## Vis Design Studies

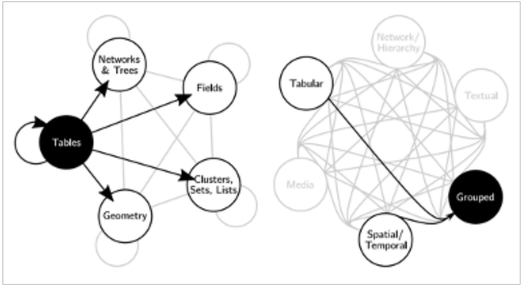
**(InfoVis 2019, Ongoing)**

| Contribution Type | Description |
|---|---|
| Algorithmic simplification | simplification or explanation of techniques and algorithms to make them easier to understand and reproduce |
| Artistic design | practices and evaluative reflection on expressiveness of visualization |
| Data abstraction | mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature |
| Data structure | improvements or new uses of existing ones to support visualization |
| Dataset | publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation |
| Deployment | discussion of insights gained from real world deployment of a tool or technique |
| Design methodology | methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches |
| Evaluation methodology | methodologies that enable new ways in evaluating visualization solutions |
| Formalism | generalized theoretical, algorithmic, or mathematical formalisms of visualization |

## Contribution Types

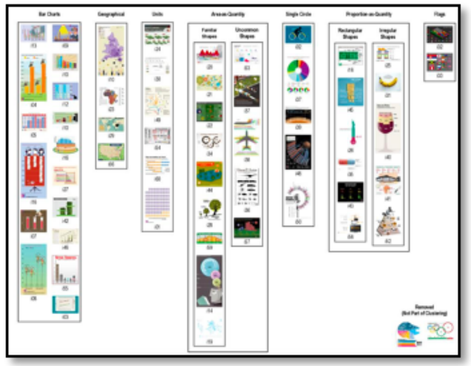
**(CG&A 2019)**

# How People Conceptualize Data




## Latent Data Abstractions

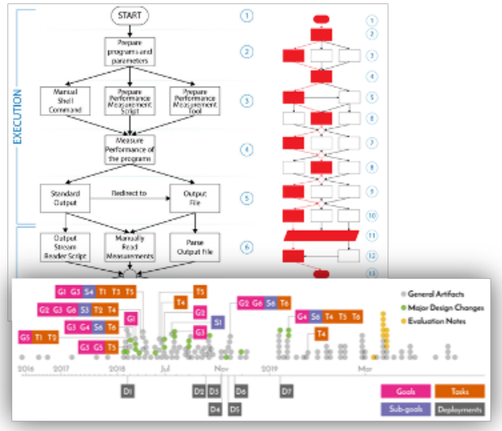
**(InfoVis 2020, Ongoing)**



## Classification of Infographics

**(Diagrams 2018)**

## Vis Research Methodology

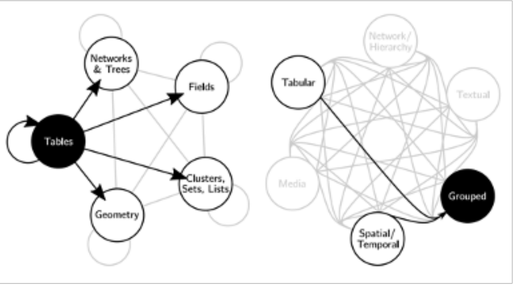



**Vis Design Studies**
**(InfoVis 2019, Ongoing)**

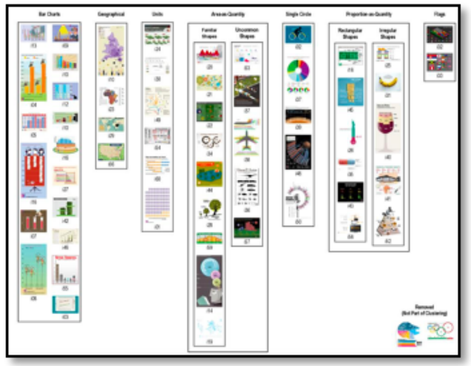
| Contribution Type | Description |
|---|---|
| Algorithmic simplification | simplification or explanation of techniques and algorithms to make them easier to understand and reproduce |
| Artistic design | practices and evaluative reflection on expressiveness of visualization |
| Data abstraction | mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature |
| Data structure | improvements or new uses of existing ones to support visualization |
| Dataset | publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation |
| Deployment | discussion of insights gained from real world deployment of a tool or technique |
| Design methodology | methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches |
| Evaluation methodology | methodologies that enable new ways in evaluating visualization solutions |
| Formalism | generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts |

**Contribution Types**
**(CG&A 2019)**

## How People Conceptualize Data

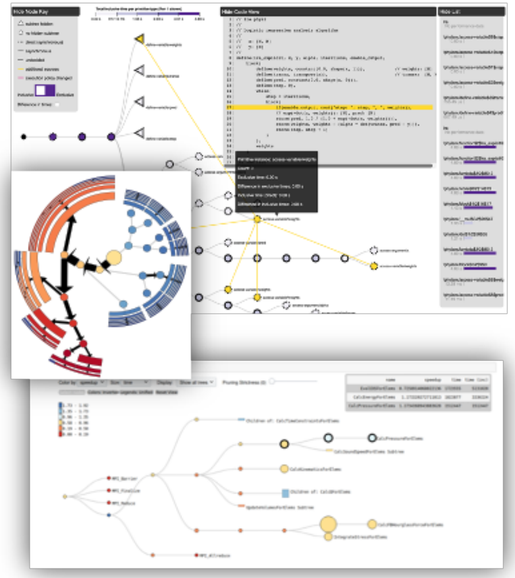



**Latent Data Abstractions**
**(InfoVis 2020, Ongoing)**



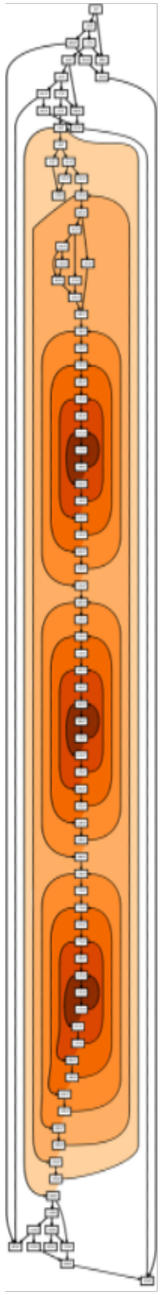
**Classification of Infographics**
**(Diagrams 2018)**

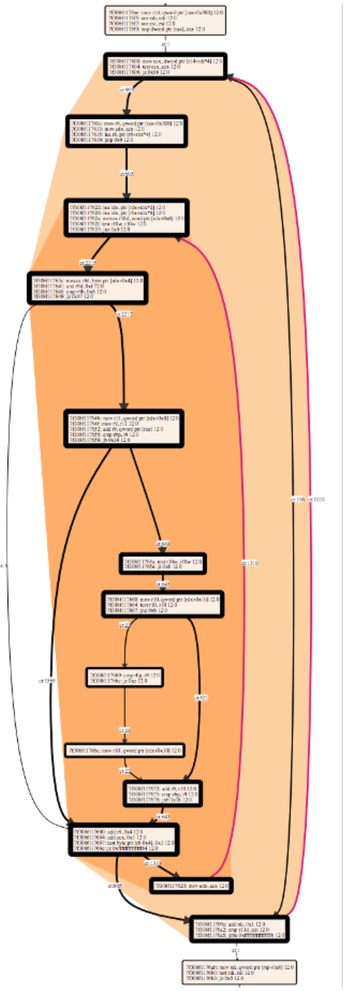
## Trees & Graphs



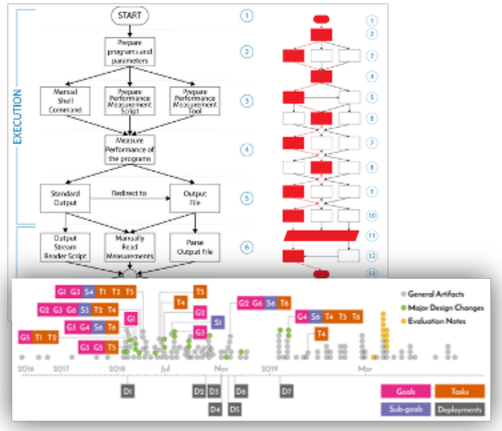
**Trees in Computing**
**(SC12, ProTools, Ongoing)**

**Stress-Plus-X Layout**
**(Graph Drawing 2019)**





## Vis Research Methodology

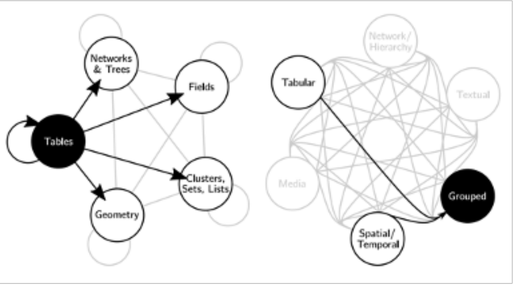



**Vis Design Studies**
**(InfoVis 2019, Ongoing)**

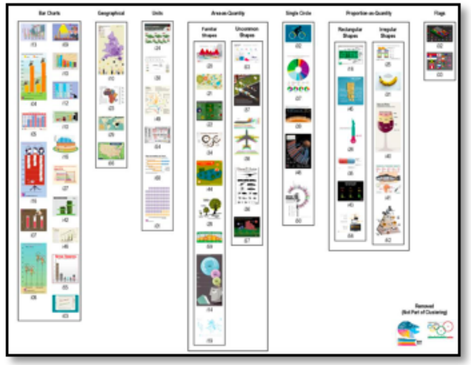
| Contribution Type | Description |
|---|---|
| Algorithmic simplification | simplification or explanation of techniques and algorithms to make them easier to understand and reproduce |
| Artistic design | practices and evaluative reflection on expressiveness of visualization |
| Data abstraction | mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature |
| Data structure | improvements or new uses of existing ones to support visualization |
| Dataset | publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation |
| Deployment | discussion of insights gained from real world deployment of a tool or technique |
| Design methodology | methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches |
| Evaluation methodology | methodologies that enable new ways in evaluating visualization solutions |
| Formalism | generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts |

**Contribution Types**
**(CG&A 2019)**

## How People Conceptualize Data

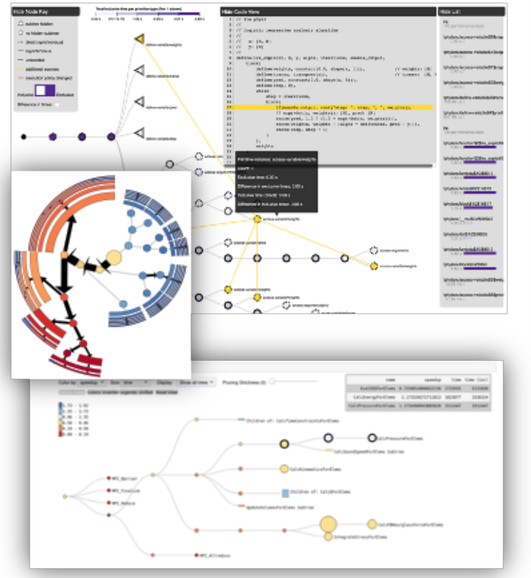



**Latent Data Abstractions**
**(InfoVis 2020, Ongoing)**



**Classification of Infographics**
**(Diagrams 2018)**

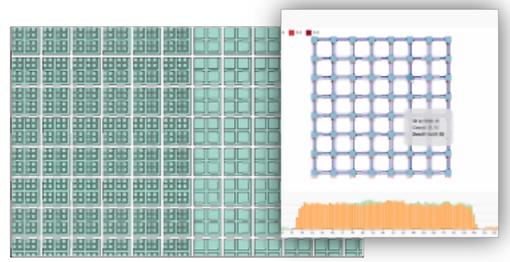
## Trees & Graphs



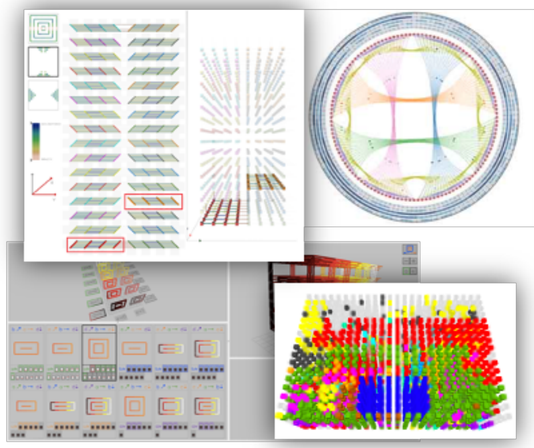
**Trees in Computing**
**(SC12, ProTools, Ongoing)**

**Stress-Plus-X Layout**
**(Graph Drawing 2019)**

## Hardware



**Data Movement on GPUs**
**(Ongoing)**



**Supercomputing Interconnects**
**(InfoVis 2012, SC12(b), SC13, +)**

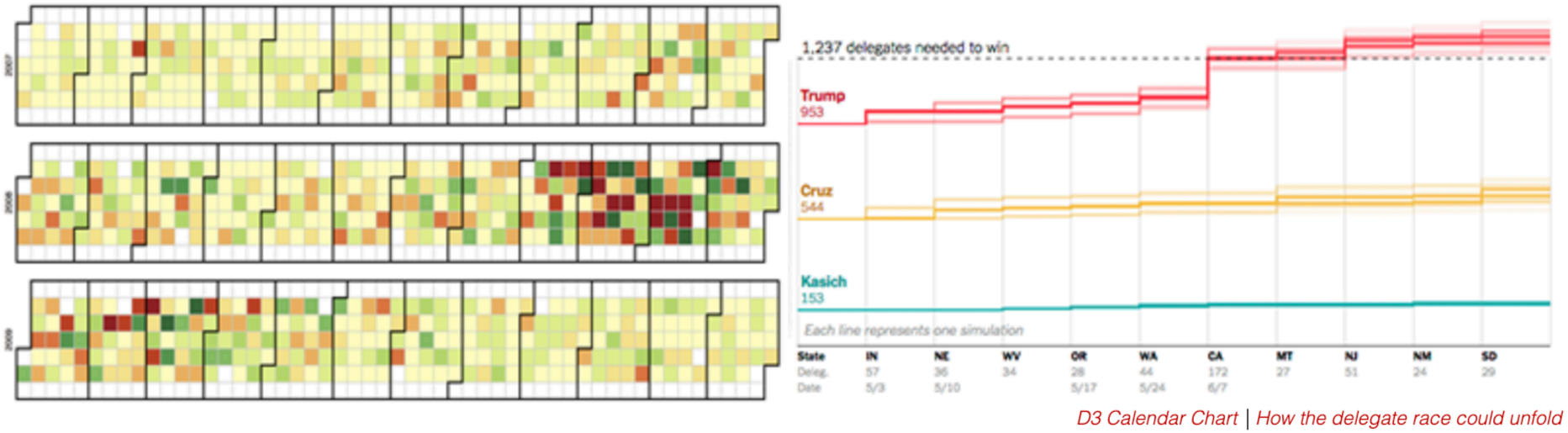
# Course Structure

# Information: datasciencecourse.net

DS 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 2023** semester at the University of Utah, cross-listed between **Mathematics (MATH 4100)** and **Computing (COMP 5360)**.

**This class is taught in person. We will use Slack for asynchronous communication and Canvas for announcement and submissions of assignments. All classes and sections are also archived online. Students are highly encouraged to attend class.**

# Communications

**Canvas**
Announcements.
Assignment submission and grading.
**Slack**
For discussions and two-way communications.
Sign up with your utah.edu e-mail address.
Accessible via Canvas.
**Github**
Used to post lectures and homework.
**Office Hours**
In person, virtual, or hybrid

To contact the teaching staff:

- Use Slack for questions related to course materials
- Private Slack messages or emails for personal issues unrelated to course materials (such as medical exceptions)
- Please **do not** use Canvas Message

# 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!

**Homework Assignments** and **Group Activities** help practice specific skills

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

# COVID

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

Drop the lowest scoring homework assignment.

5-Minute “Bio Break” each lecture.

# How are you graded?

Homework Assignments: 45%

- Equally weighted
- 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 small groups.

Typically a coding problem.

All work together on the problem. It is okay to note names and share common solution file with group members, but each person needs to upload the file to their CANVAS page.

Submit the day of lecture.

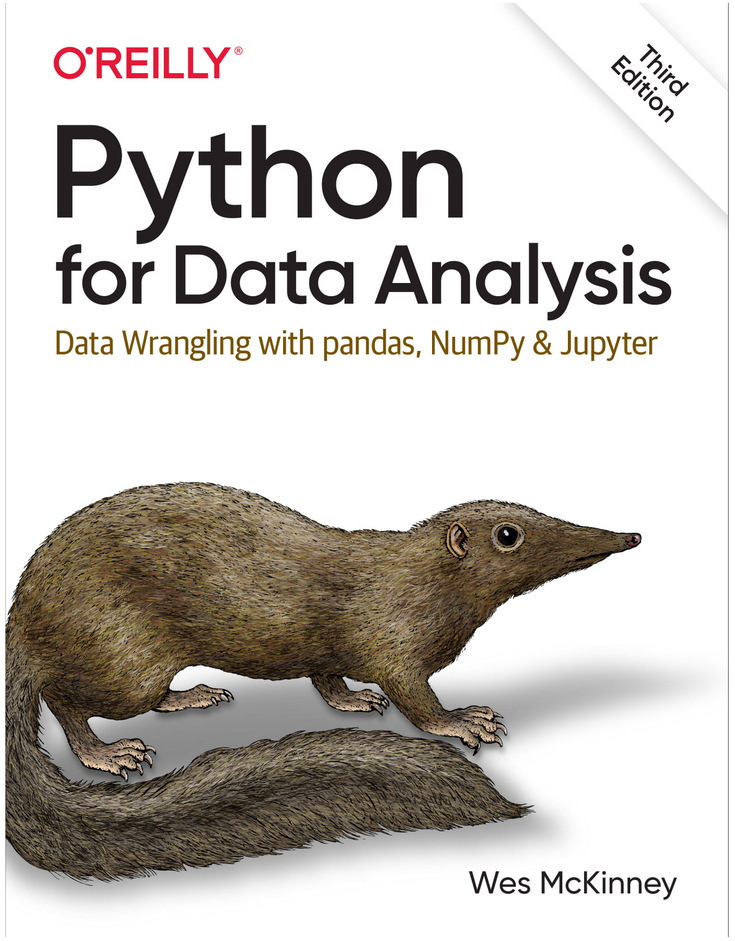
Complete 5 activities (out of ~10-15)

# Schedule

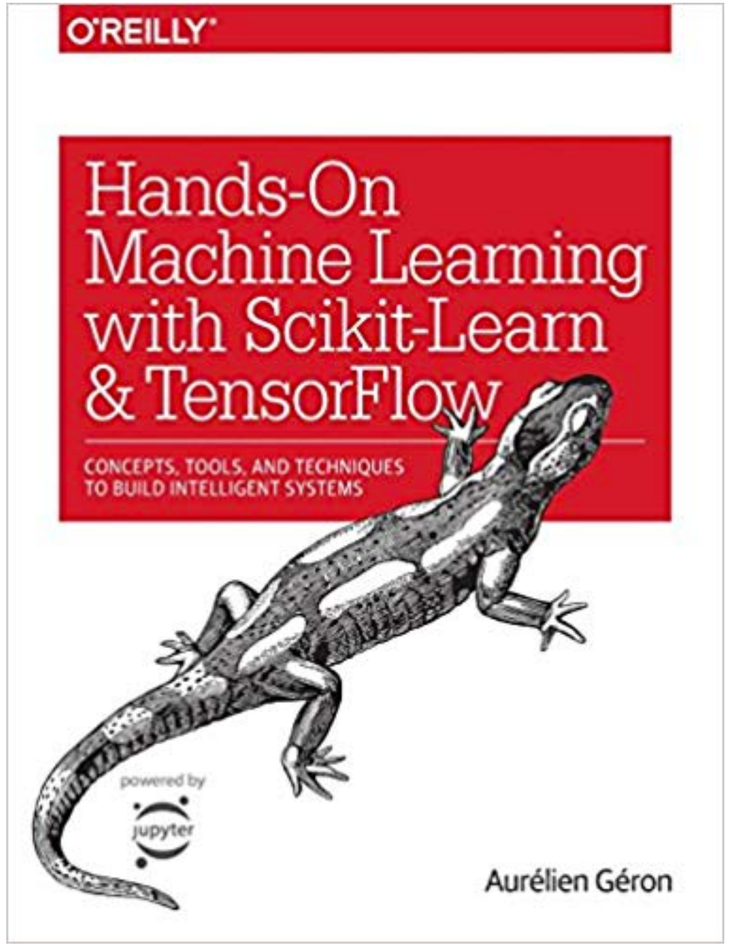
**Lectures:**
Tue / Th 3:40 - 5:00 PM MDT (in person in WEB L101).

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

# Books




Primary Text for Readings
Available for free here



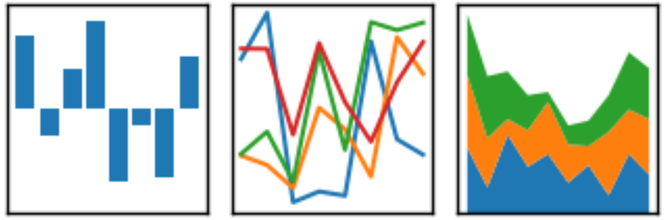

Supplementary Text
Available for free from Campus Library

# 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 in writing.

We reserve the right to **check all submissions for plagiarism**.

# 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

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/2023-datascience-homework/tree/main/HW0

## Homework 0

**Introduction to Data Science - MATH 4100 / COMP 5360.**
*This homework is due before class on Thursday, January 12th.*

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. Setup

We'll often work on practical skills related to data science. That means we'll write code, and we'll do that in a programming language called Python.

Python has three advantages for this class: it's pretty easy to learn, it's the language of choice for many data scientists,

# Github

https://github.com/datascience-course/2023-datascience-homework/blob/main/README.md

**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.**

# 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/2023-datascience-homework
```

This will create a folder `2023-datascience-homework` on your computer, with the individual homeworks in subdirectories.

# Next Week

Data Structures and Pandas

Introduction to Descriptive Statistics

HW1 due

# Enrollment

Math 4100: 41

COMP 5360: 49

For permission code of MATH 4100:

https://www.math.utah.edu/undergraduate/registration.php

For permission code of COMP 5360:

https://www.cs.utah.edu/undergraduate-advising/permcodes/
(Fill out the form under “undergraduate" even if you are a graduate student).

Trouble enrolling? send an email to instructor.

Please check your own department degree requirement.

# Small Groups!

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?