

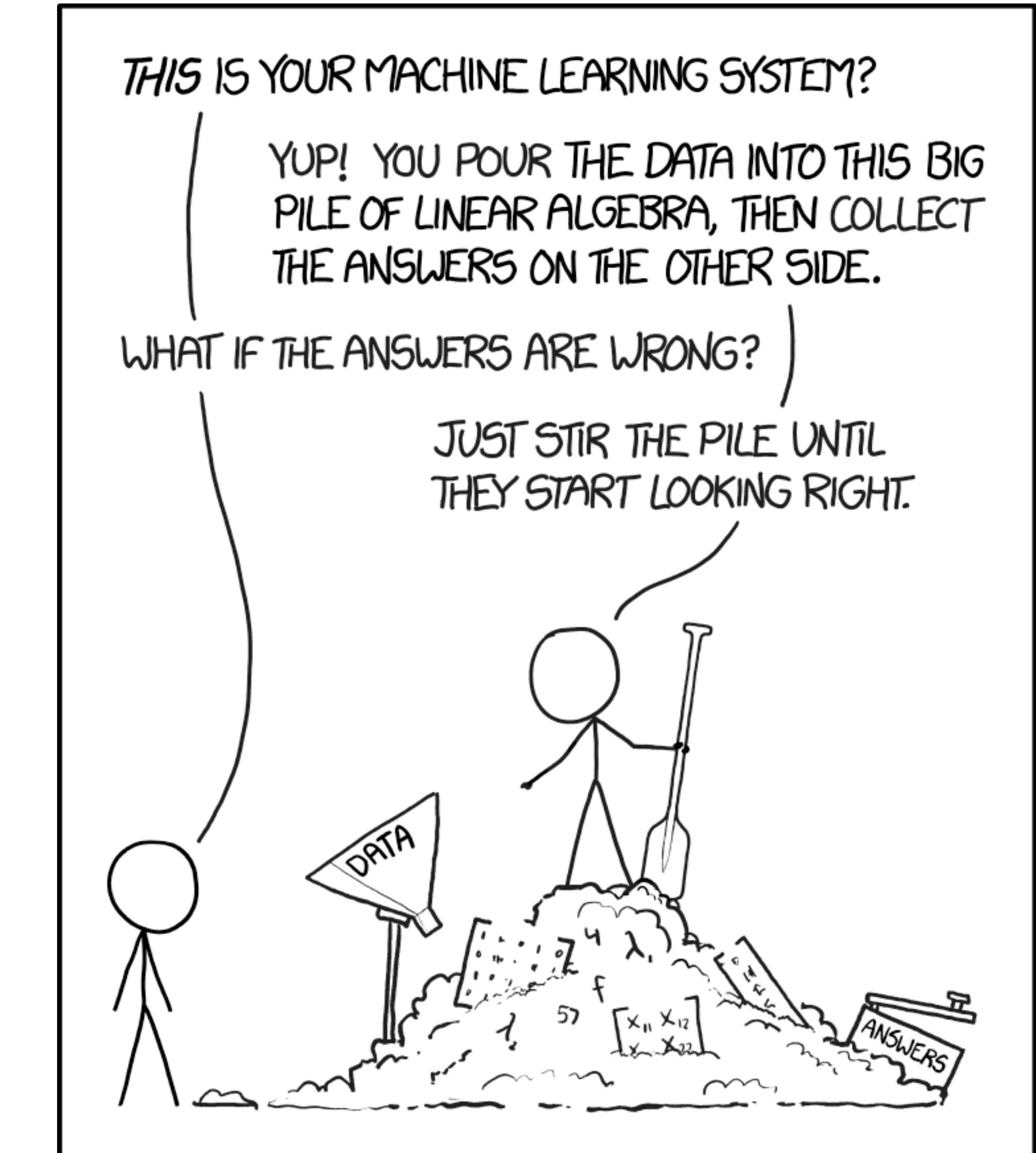
Introduction to Data Science

CS 5360 / Math 4100



Kate Isaacs
u6044649@gcloud.utah.edu

Bao Wang
bwang@math.utah.edu



Logistics

See CANVAS Module “Important Online Links” for links to the zoom meeting, YouTube channel, and office hours.

We will use PIAZZA for course discussions and Q&A, so please sign up.

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

(Old) Recordings

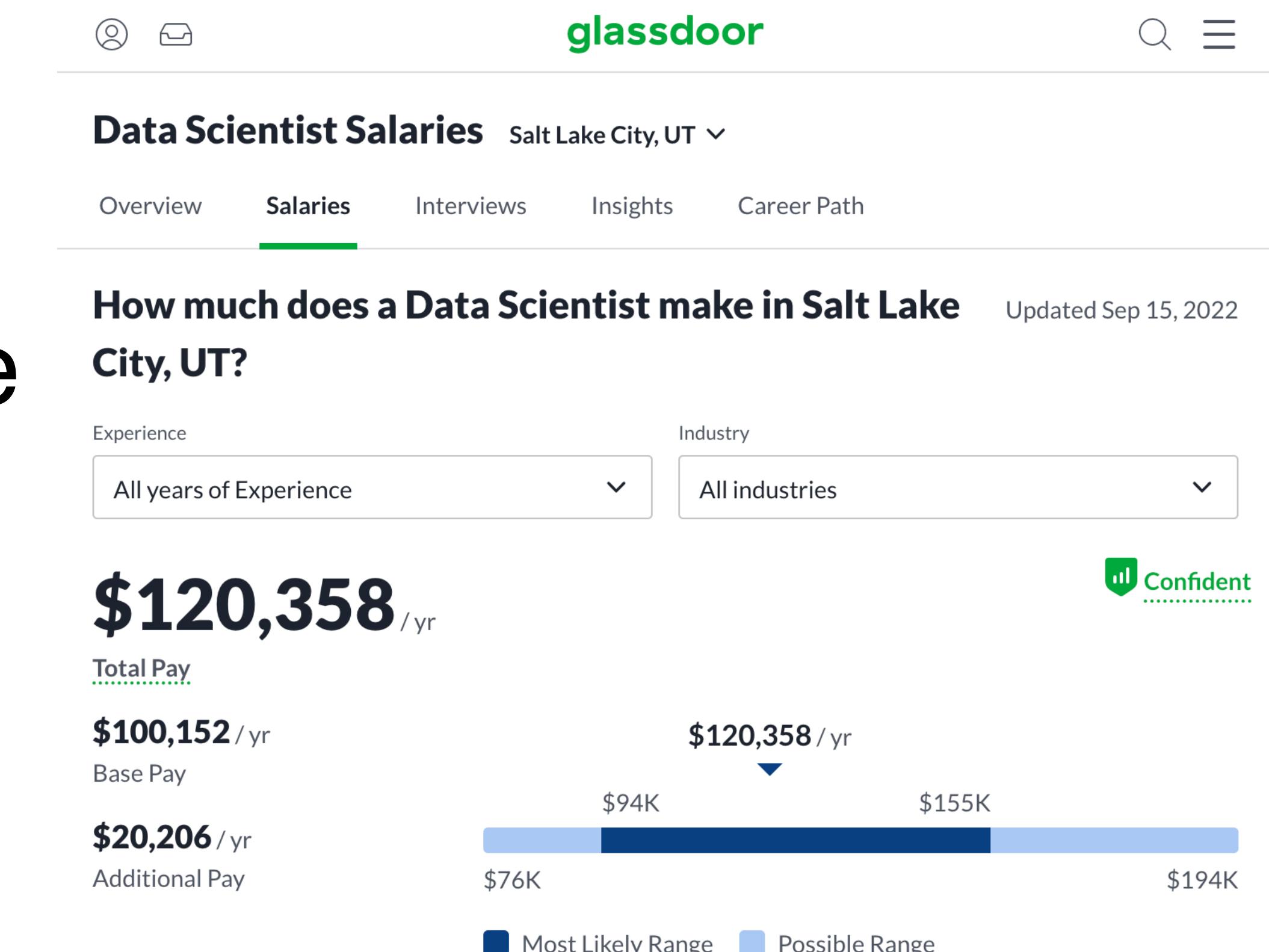
We are **not** recording this class or doing hybrid lectures per guidance from our colleges.

There are lectures from 2022 which are similar, can serve for review or backup <https://datasciencecourse.net/2022/>

If there is severe weather, we may do a hybrid or Zoom lecture. We will announce on Canvas. Please reach out on days you're concerned about if we haven't made a call.

What is Data Science?

- The sexiest job of the century – Harvard Business Review
- A data scientist is a statistician who lives in San Francisco
- 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.

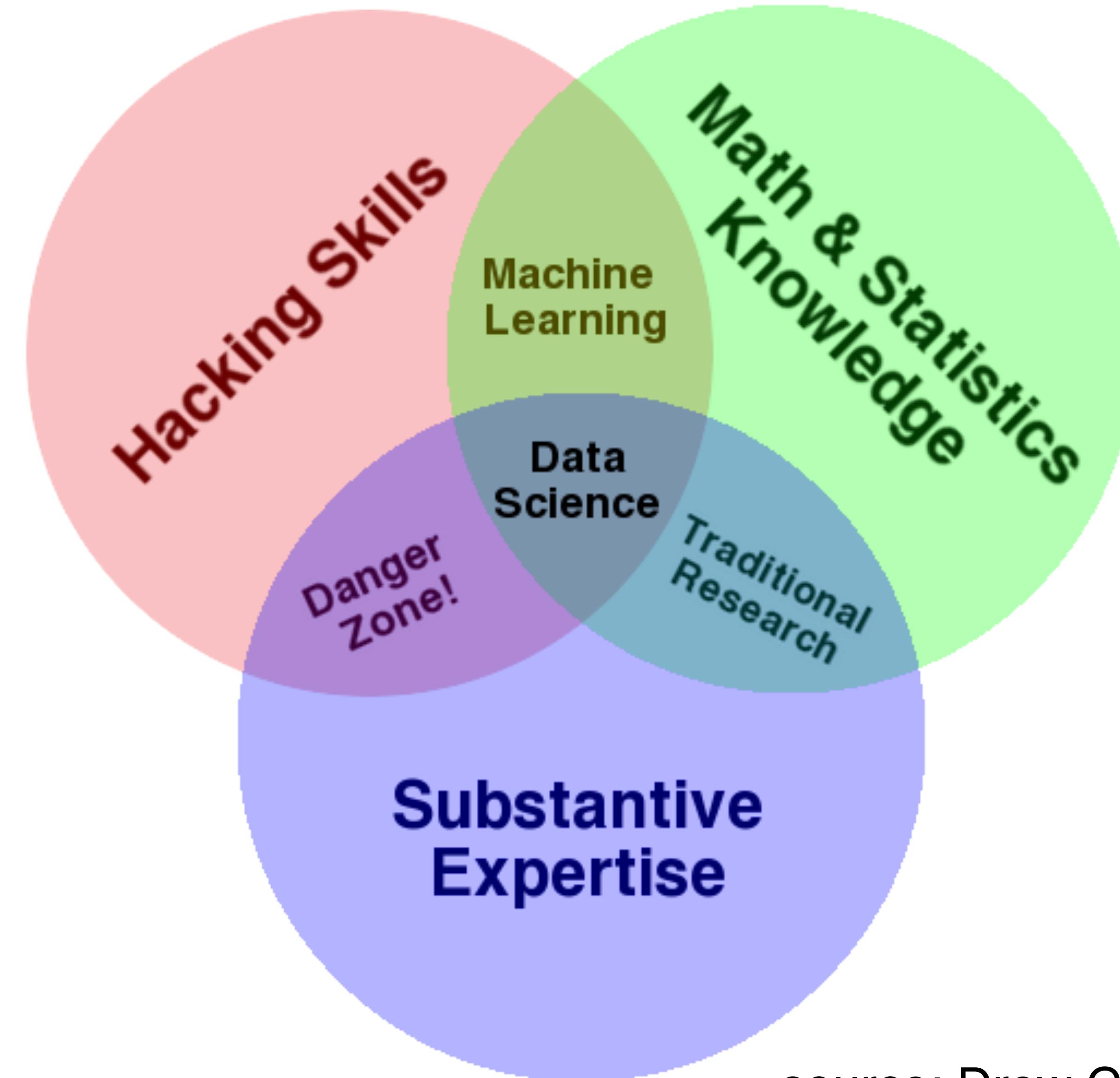


The estimated total pay for a Data Scientist is \$120,358 per year in the Salt Lake City, UT area, with an average salary of \$100,152 per year. These numbers represent the median, which is the midpoint of the

What is Data Science?

evidence collect approach
warehouse related derived system describes much
scale small new system whether system
term enough continues insights things storage anything
often velocity scope computational reality made
refers diverse really machine think world like right
means technological actually understand particular
information society datasets complex
changes way used standard advances capturing
decisions organization behavior people lot essential
organization tool gather better fit tb
making store databases storing
patterns single make organizations number
buzzword traditional processes products
time internet methods memory also issues
mean

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.

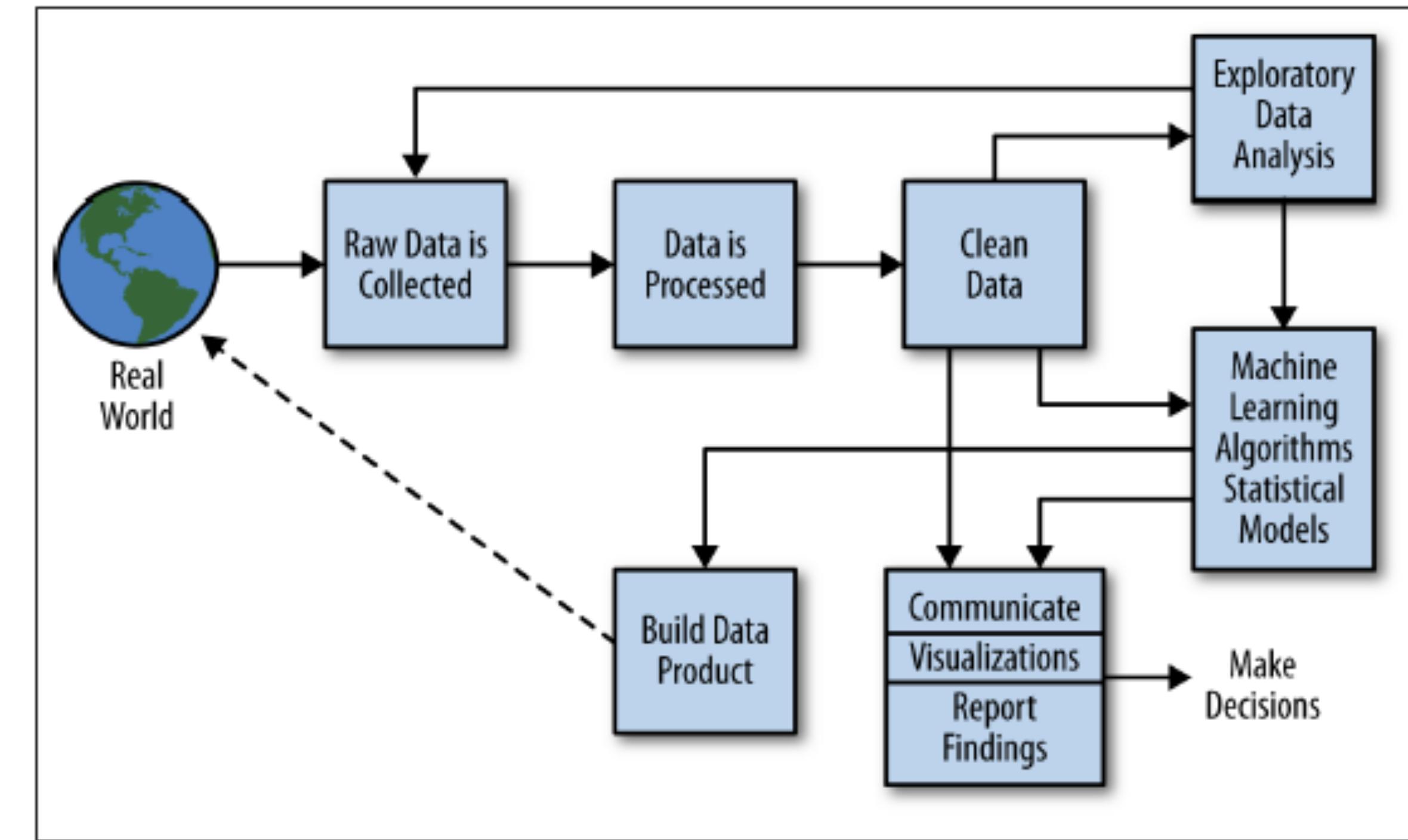


Figure 2-2. The data science process

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.

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

A screenshot of a Google search results page. The search query "youtube cat videos" is entered in the search bar. Below the search bar, there are navigation links for "Web", "Videos", "Shopping", "Images", "News", "More", and "Search tools". A red oval highlights the text "About 593,000,000 results (0.44 seconds)" which is displayed below the search bar. The first result is a link to "TOP 10 BEST CAT VIDEOS OF ALL TIME! - YouTube" with a thumbnail image of a cat.

15 Exabytes in Punch Cards:
4.5 km over New England



In one second on the Internet there are...



How can we leverage data?

Improve your fitness by targeted training

Improve your product

- by targeting your audience

- by considering semantics

Make better decisions

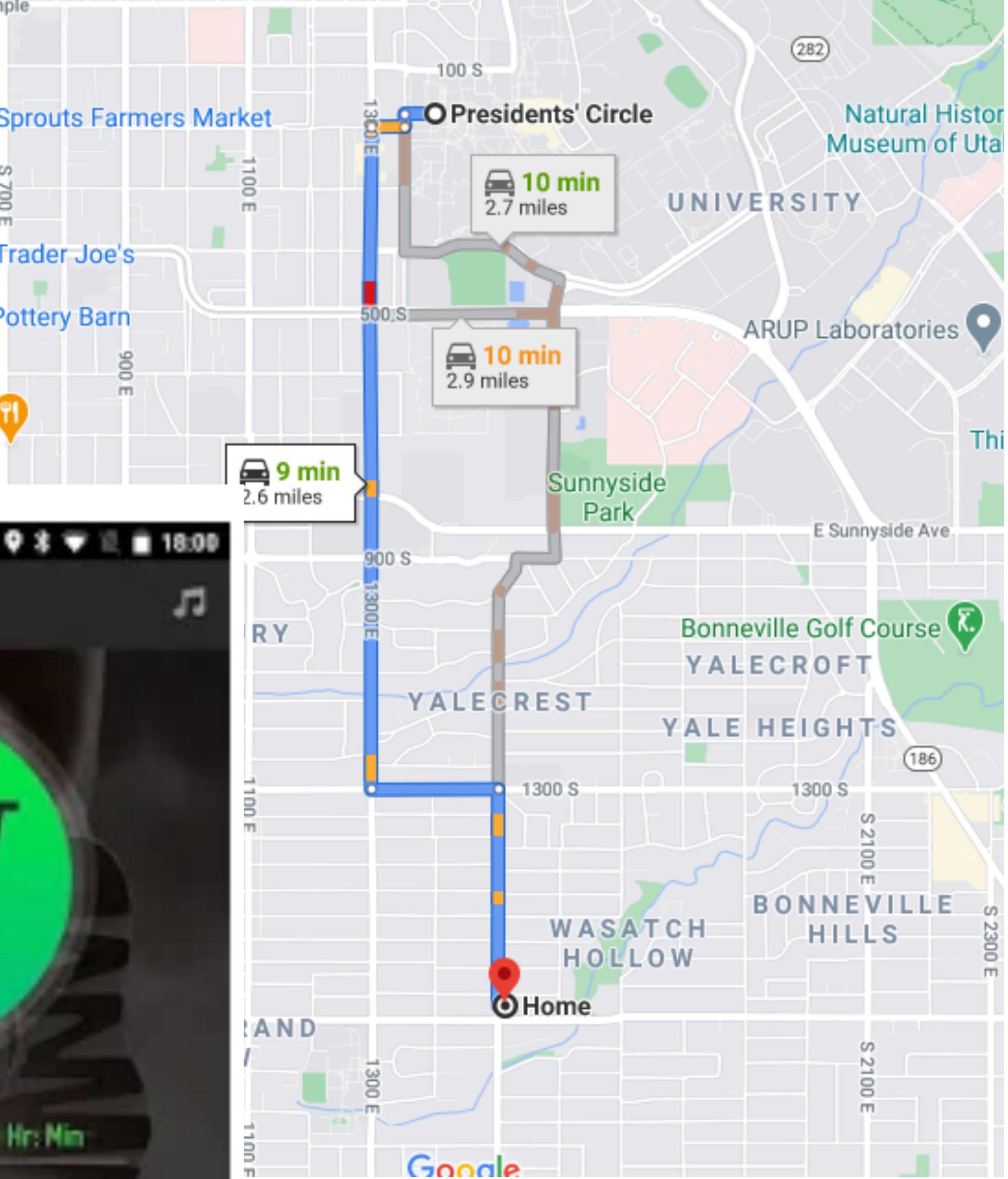
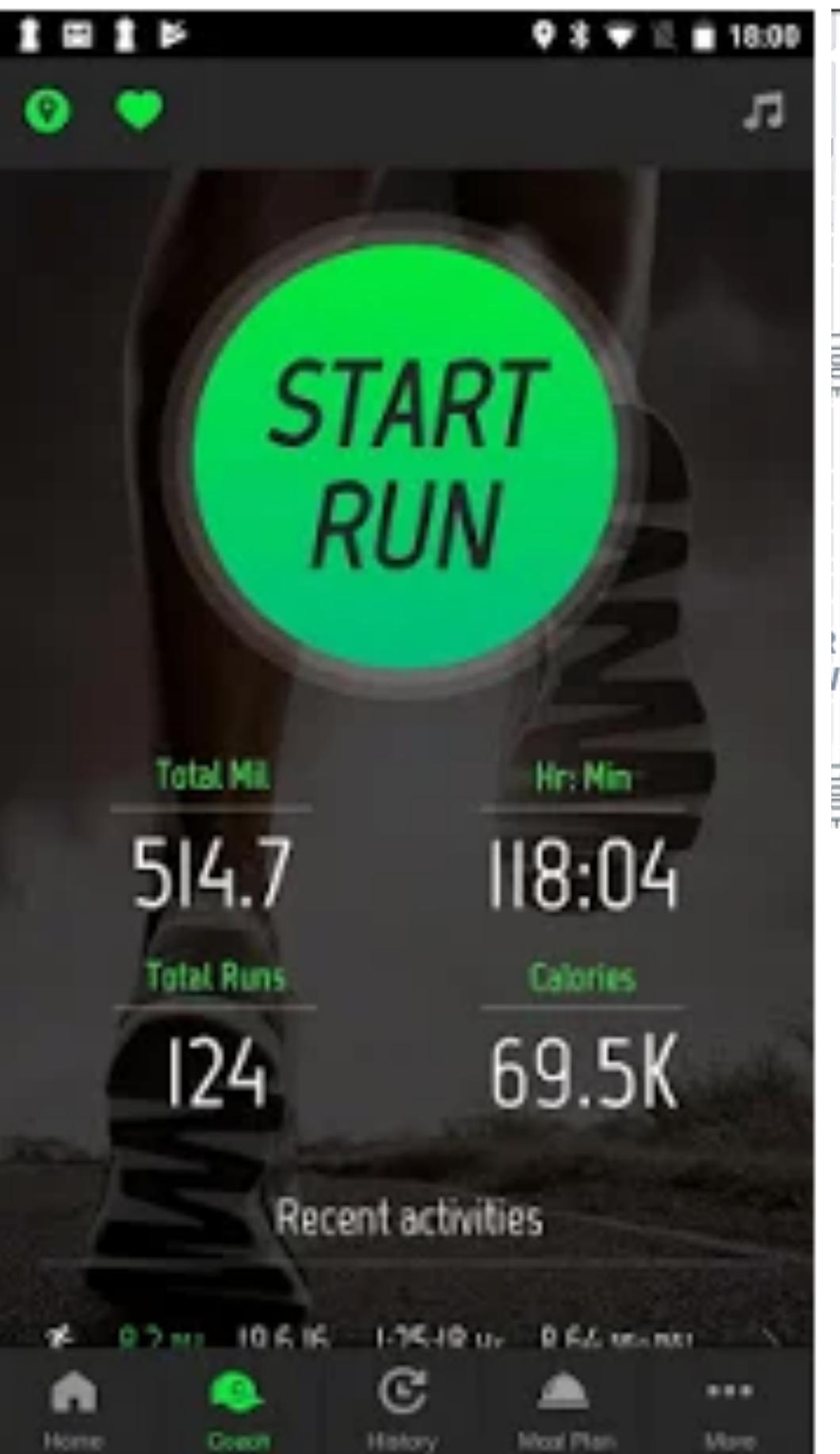
- exact diagnosis, choose right medication, pick good restaurant

Predict elections, events, crowd behavior, etc.

... and many more applications

Example: Personal Data

The Zillow search interface for Salt Lake City, UT, displays a map of the city with numerous red dots representing homes for sale. A blue boundary box highlights a specific area in the central business district. Buttons for "Remove Boundary" and "Re-center" are visible. The search filters include "For Sale", "Price", "Beds & Baths", "Home type", and "More". The results page shows 193 Agent listings and 48 Other listings, sorted by "Homes for You". One listing is highlighted: a house at 1333 N Capistrano Dr, UT 84116, listed for \$322,000. The listing includes a photo, the price, and the agent's name, Utah Key Real Estate. The Zillow logo is at the top right.



Big Data in Science and Engineering

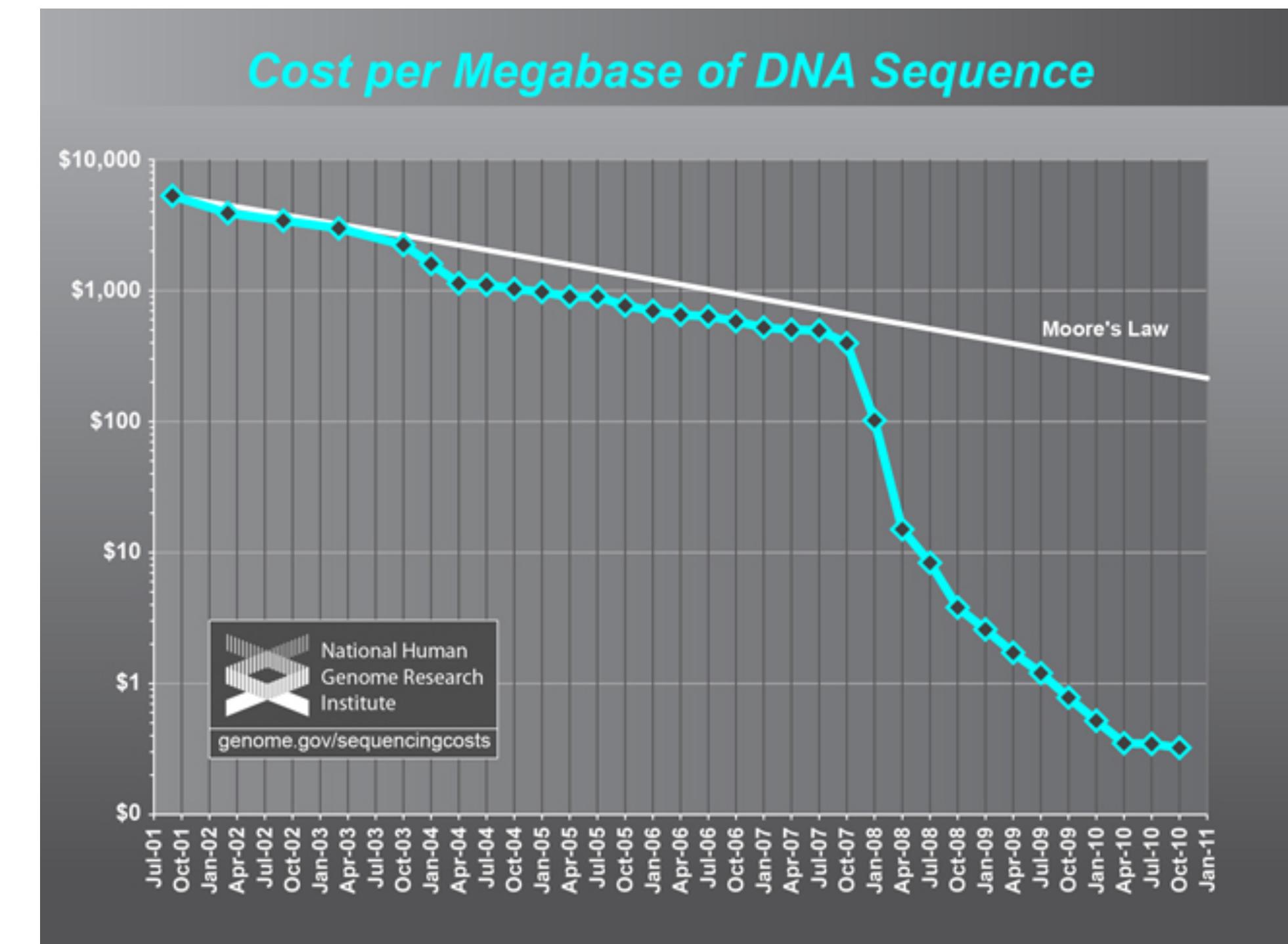
Big Data transformed science and engineering.

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 (blynksy)
- 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 “BumbleHive” 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



[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

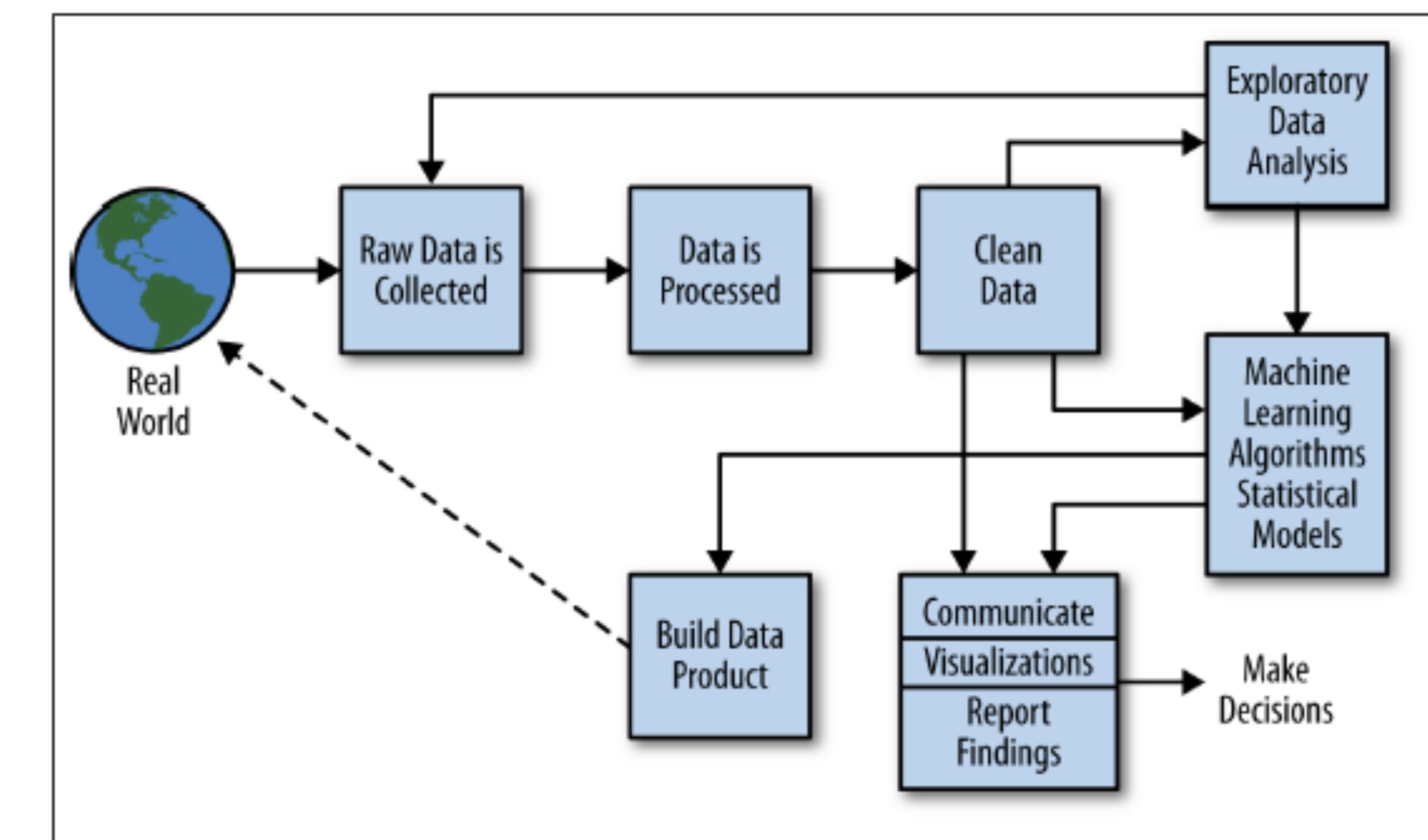


Figure 2-2. The data science process

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**

Sayef Azad Sakin - TA



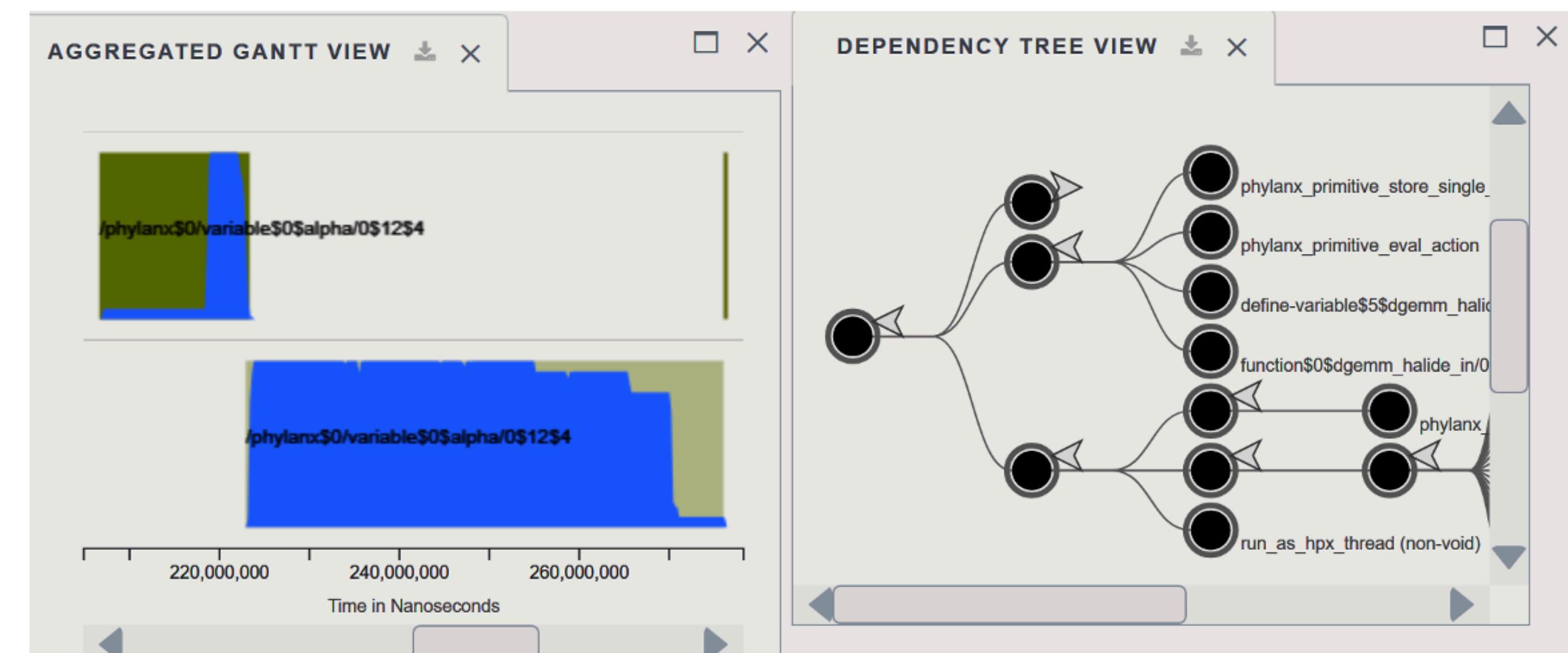
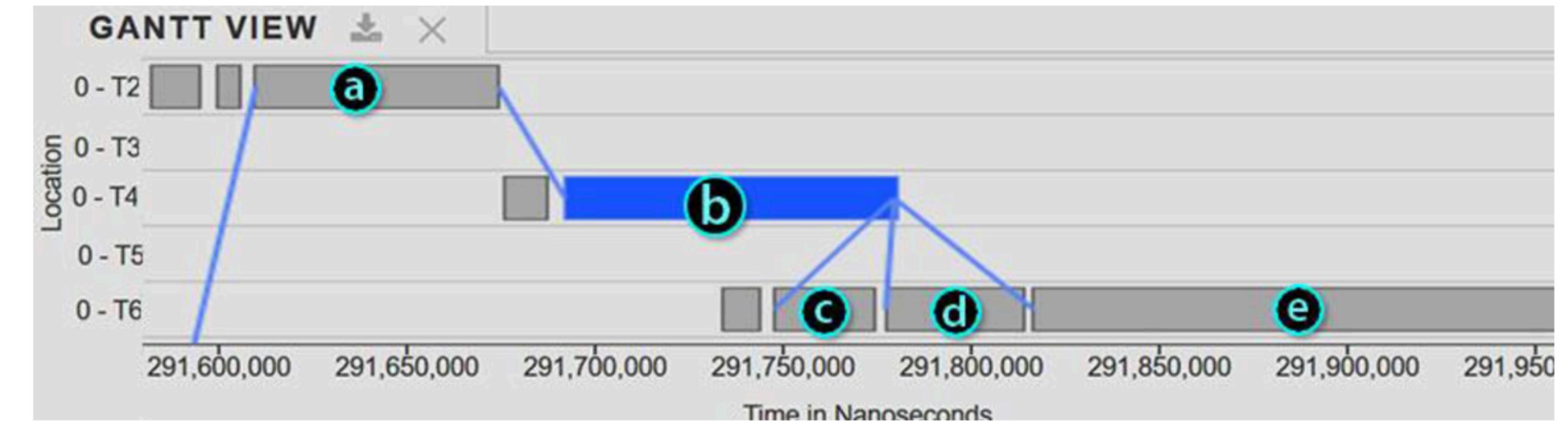
Email: sayefsakin@sci.utah.edu
Website: sayefsakin.blogspot.com
Office: WEB 3760

TA Office Hour

Time: Friday 3:30pm - 4:30pm

Location: MEB 3105

- ✓ From Dhaka, Bangladesh
- ✓ 5th Year PhD Student. Advisor: Dr. Kate Isaacs
- ✓ Data Visualization, High Performance Computing, Data Management



Estelle Trieu

P.hD. student, Computer Science

Graduate RA, Scientific Computing and Imaging (SCI)
Institute

Gynecology Medical Scribe, Huntsman Cancer
Institute

Before that: B.S. in Mathematics and Computer
Science, University of Oregon

Research: Developing novel MR guided Focused
Ultrasound Breast Cancer treatment using Deep
Learning



TA Office Hour

Time: Thursday 2:30 - 3:30pm
Location: MEB 3105

Bao Wang

Assistant Professor, Mathematics

Faculty, Scientific Computing and Imaging (SCI) Institute

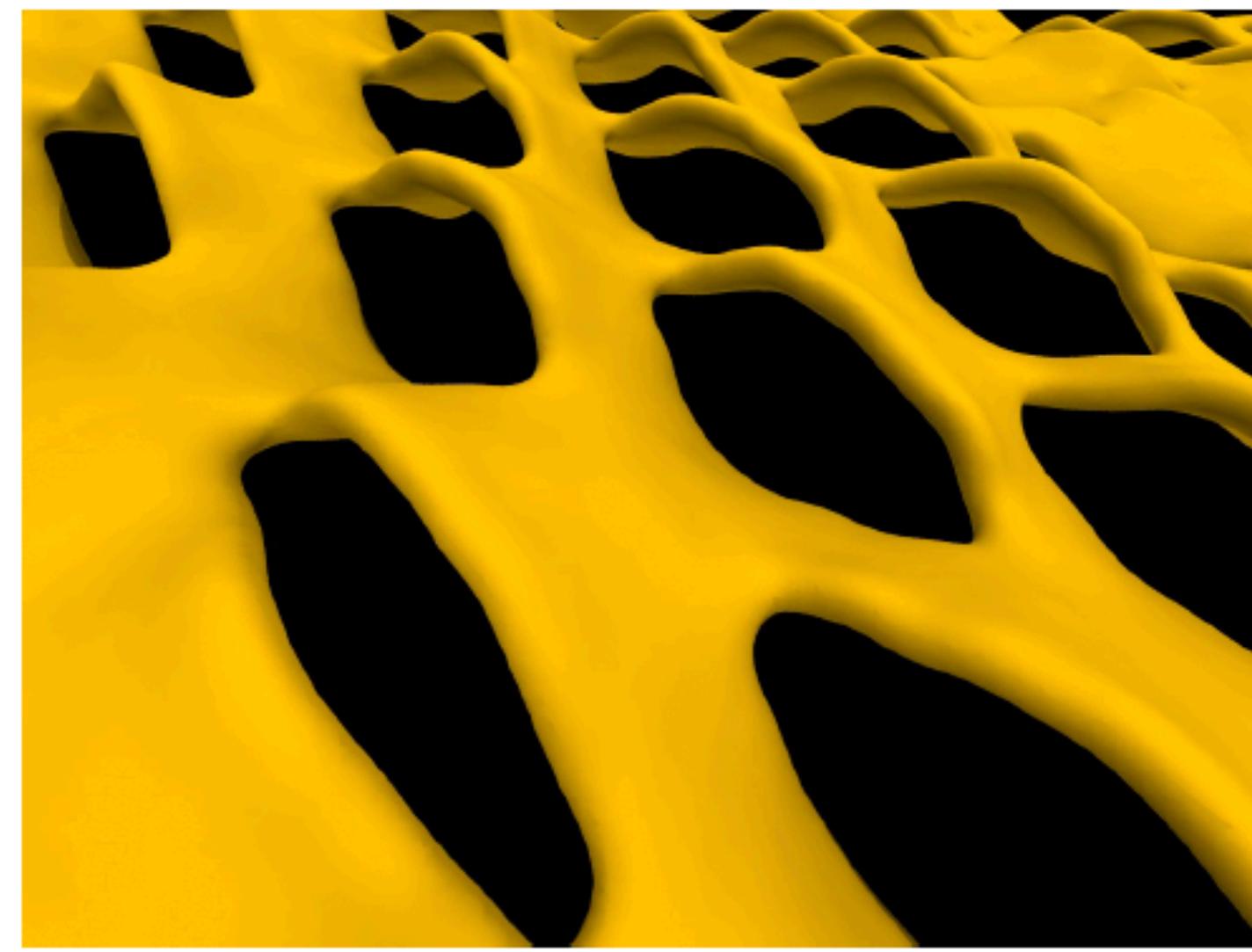
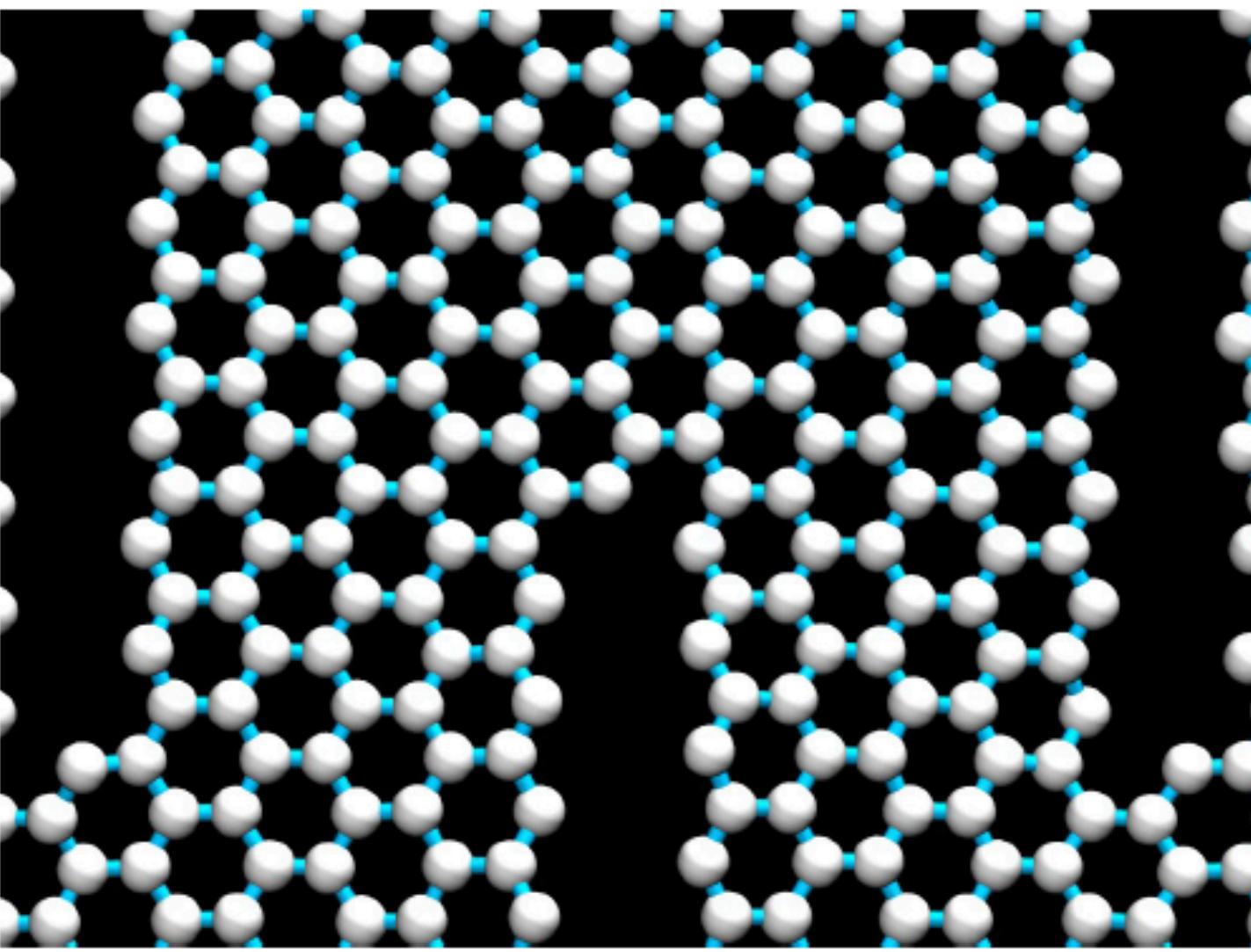
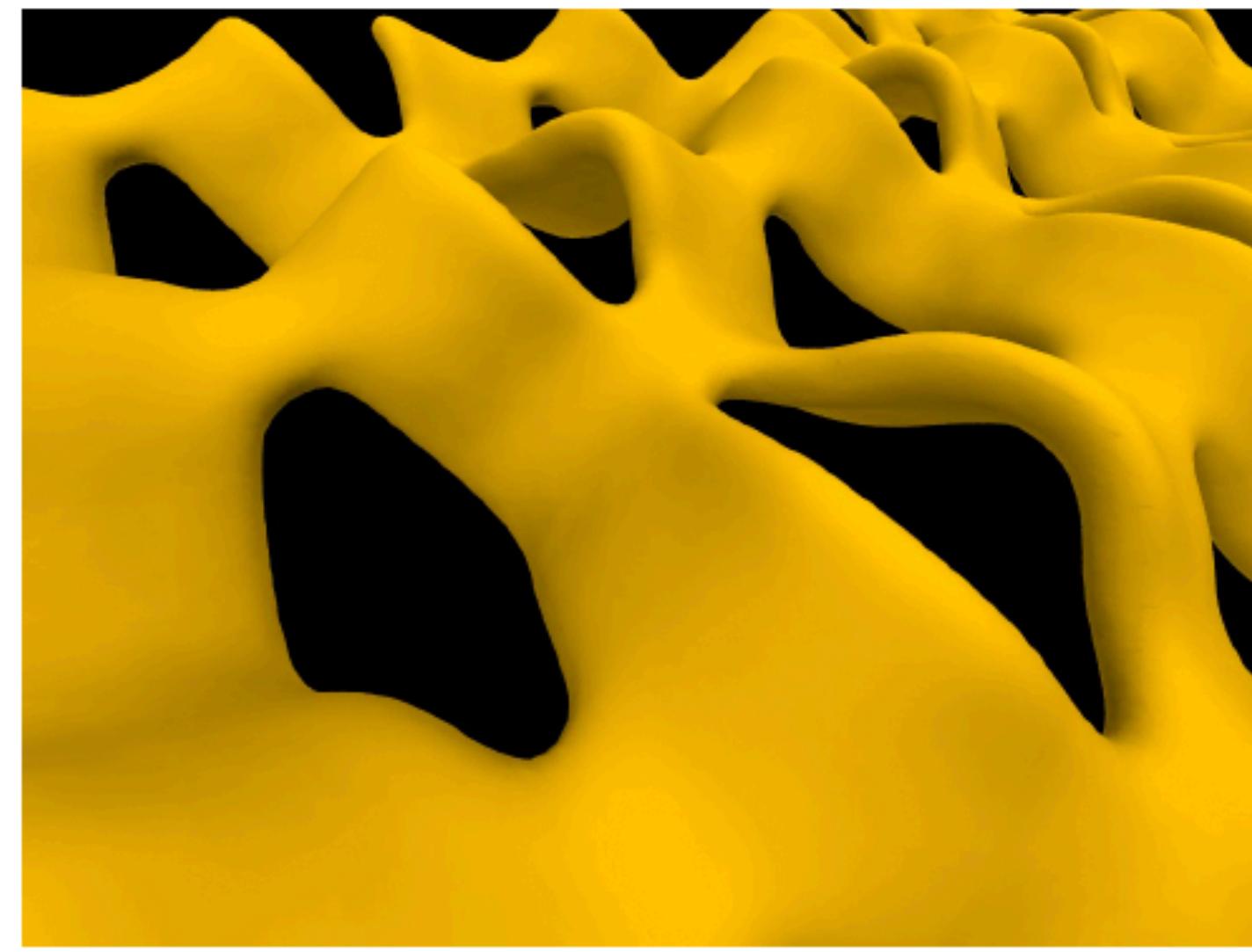
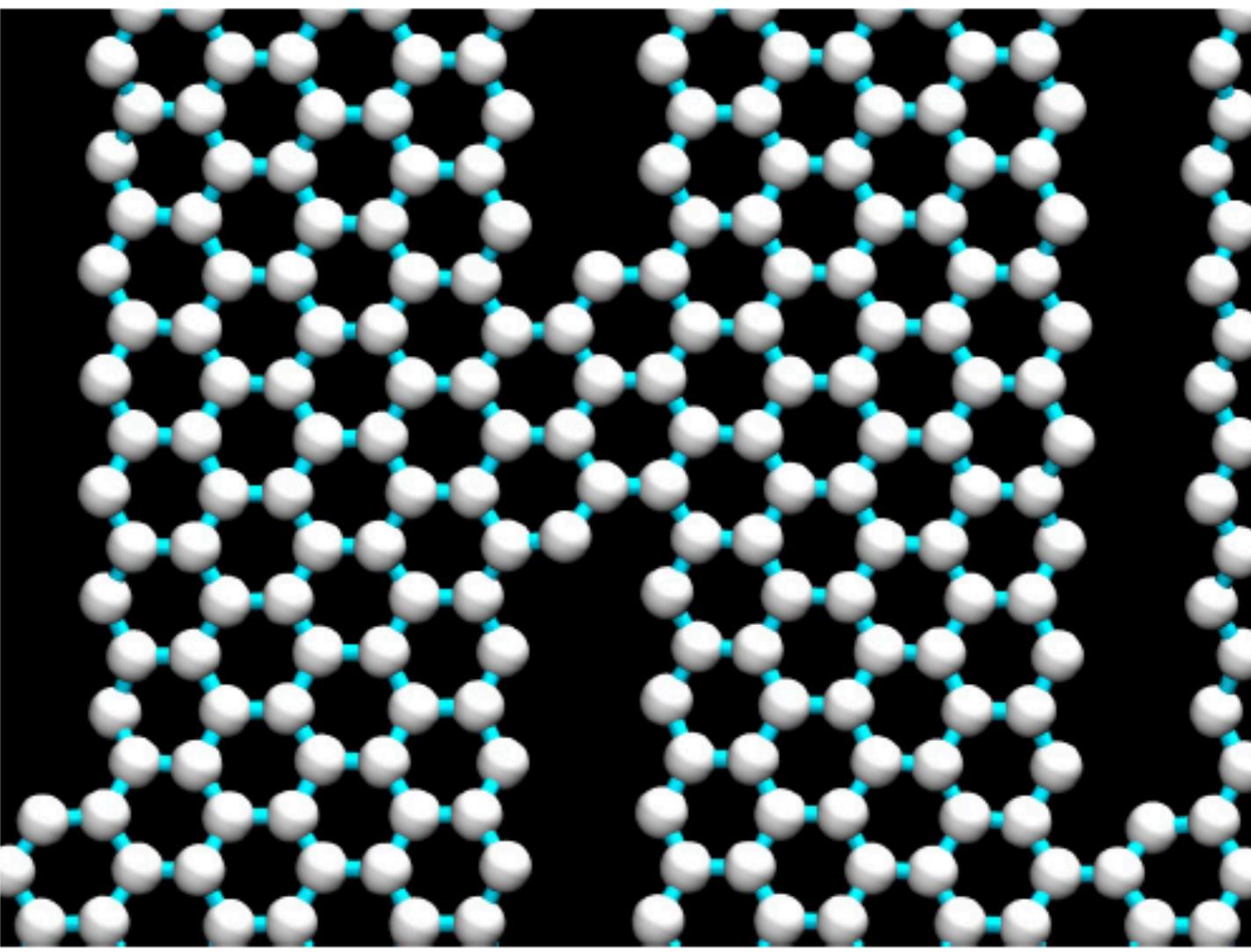
Office: WEB 4652



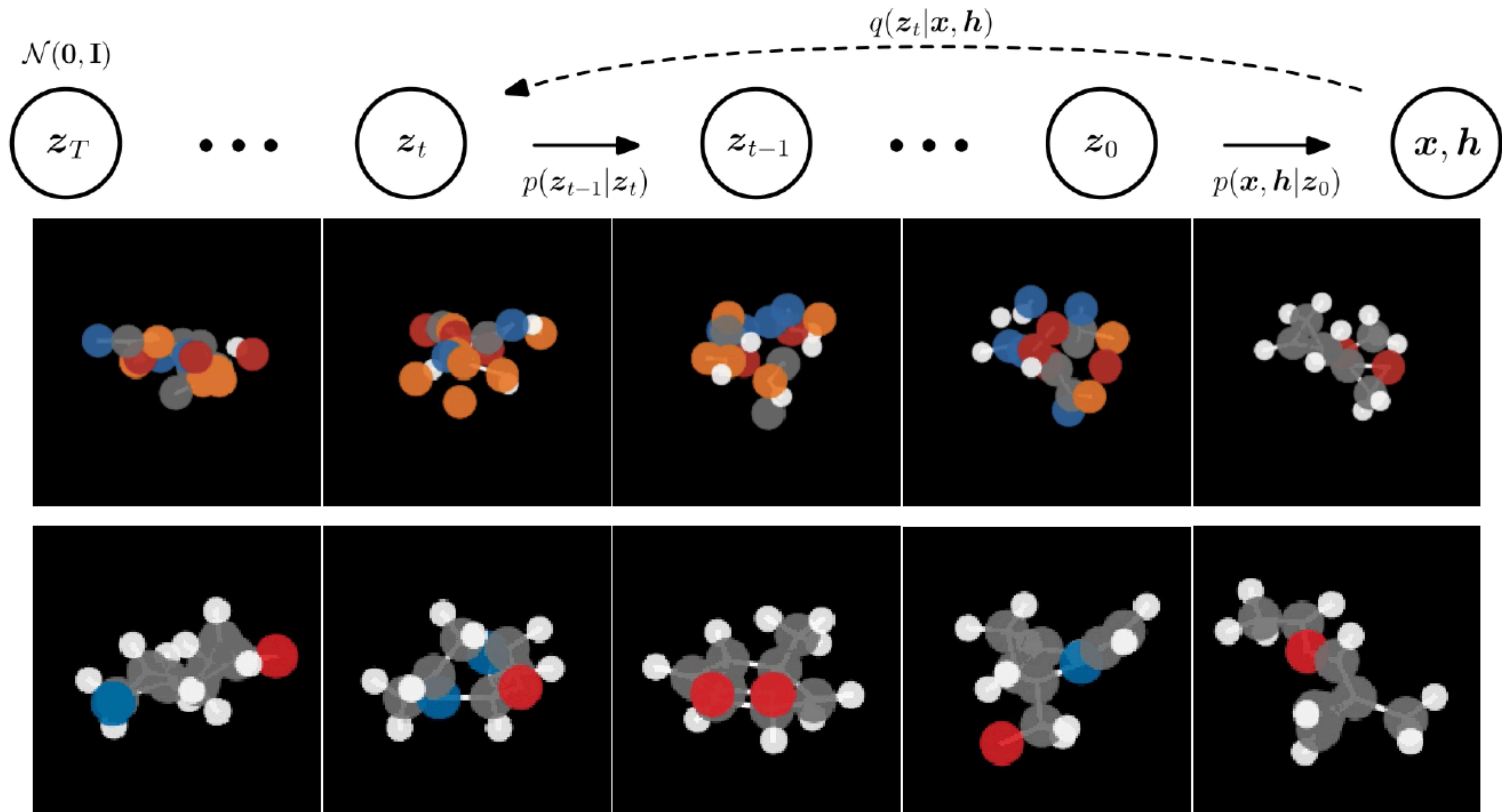
<https://www.math.utah.edu/~bwang/>

Deep Learning, Mathematics and Data Science, and AI for Science

AI for Materials Design: Graphene Kirigami

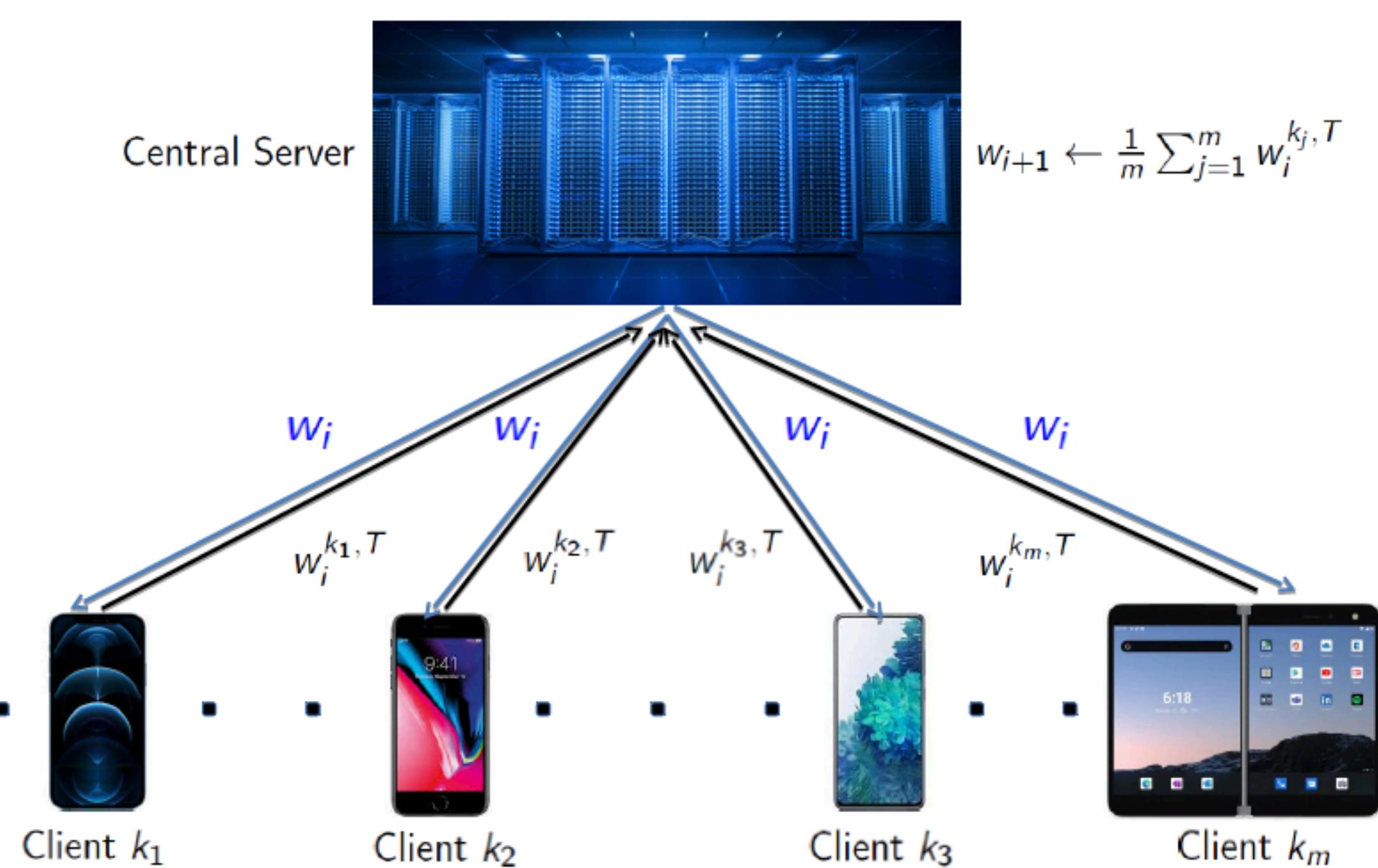


AI for Drug Discovery



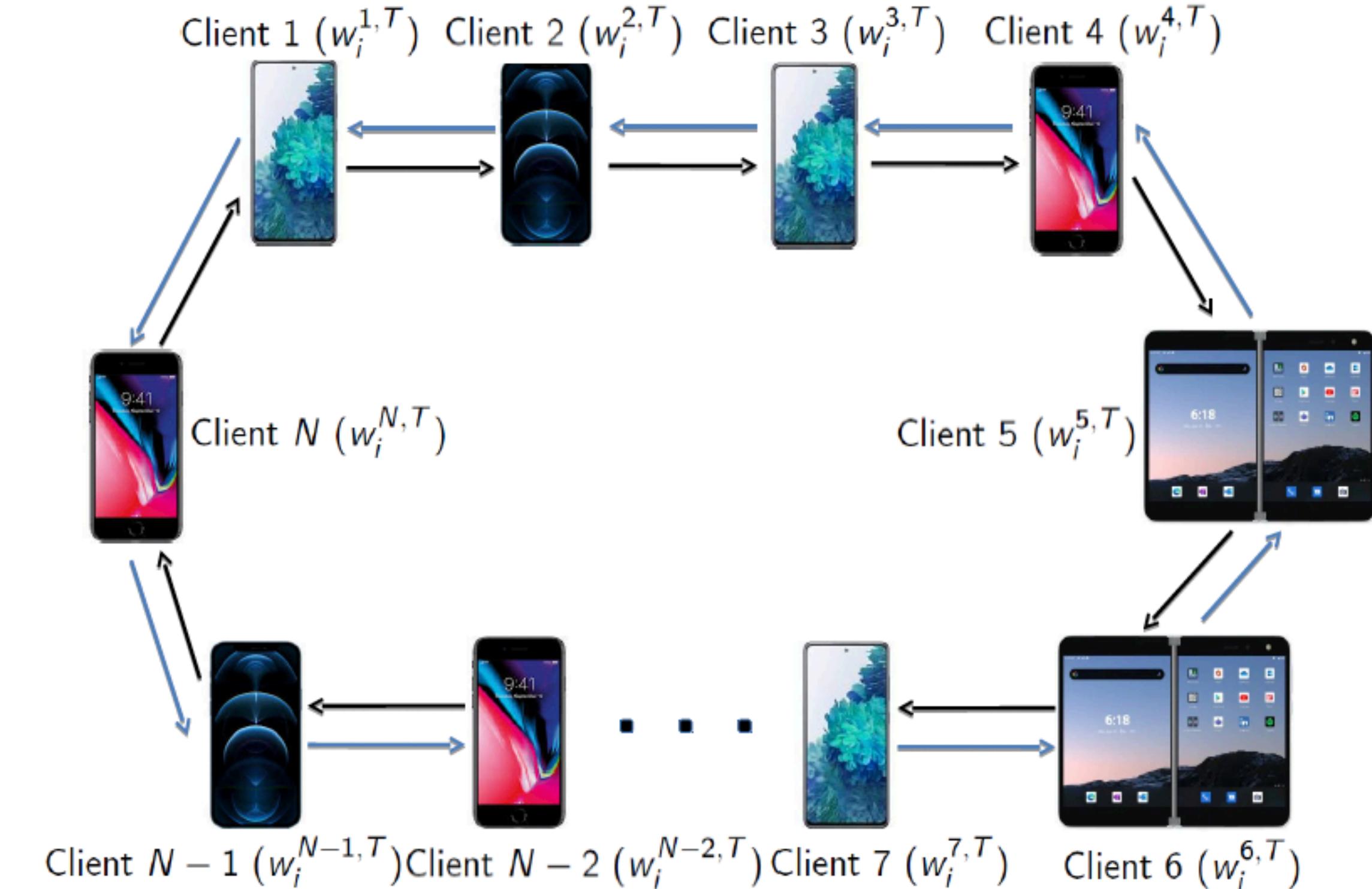
Trustworthy Machine Learning

Communication Round i



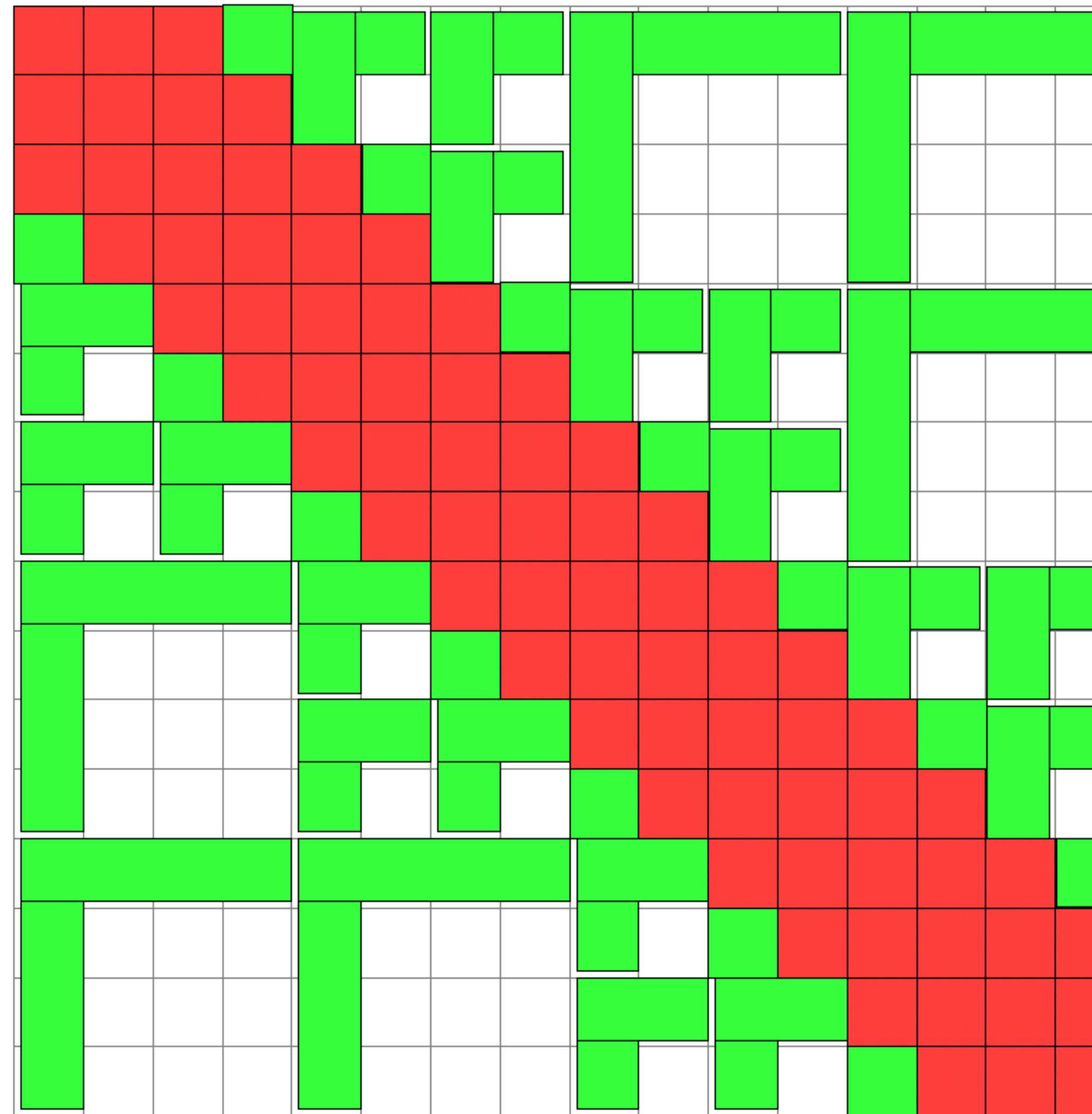
(a) Federated learning

Communication Round i



(b) Decentralized federated learning

Low-rank and Sparse Approximation of LLM



Kate Isaacs

Associate Professor, Computer Science

Faculty, Scientific Computing and Imaging (SCI) Institute

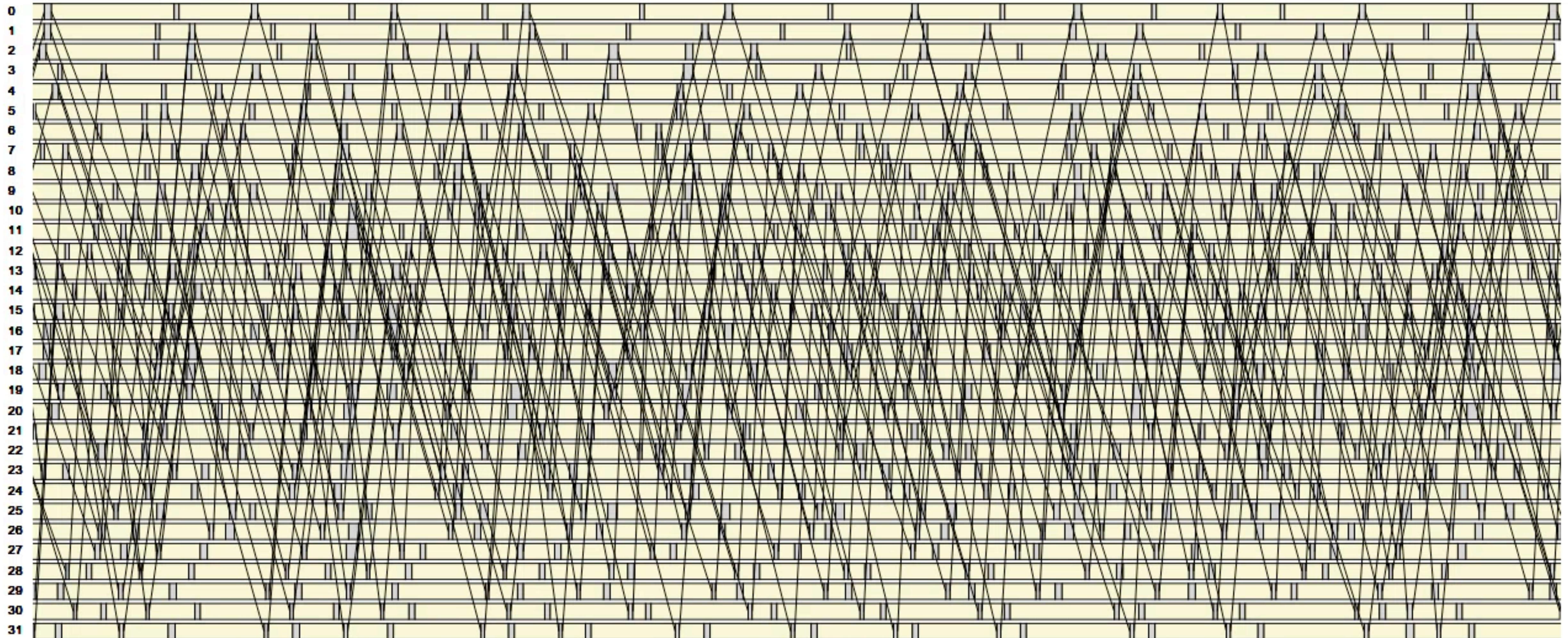
PhD in CS, UC Davis



<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]:

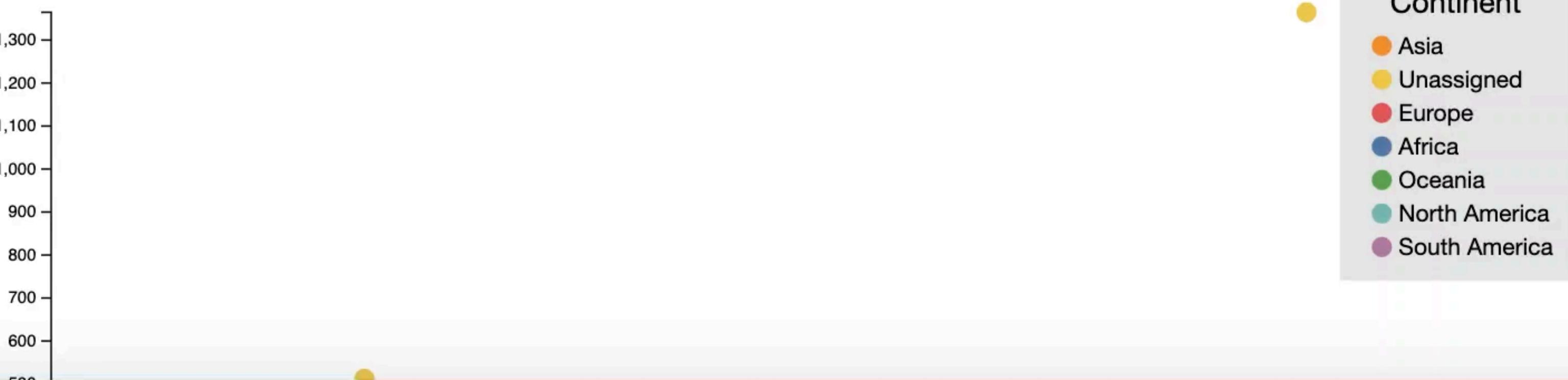
	Entity	Code	Year	Electricity from coal (TWh)	Electricity from gas (TWh)	Electricity from hydro (TWh)	Electricity from other renewables (TWh)	Electricity from solar (TWh)	Electricity from oil (TWh)	Electricity from wind (TWh)	Electricity from nuclear (TWh)	Continent
0	Afghanistan	AFG	2019	0.000000	0.344000	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

X Metric: Electricity from gas (TWh) Y Metric: Electricity from oil (TWh) Brush



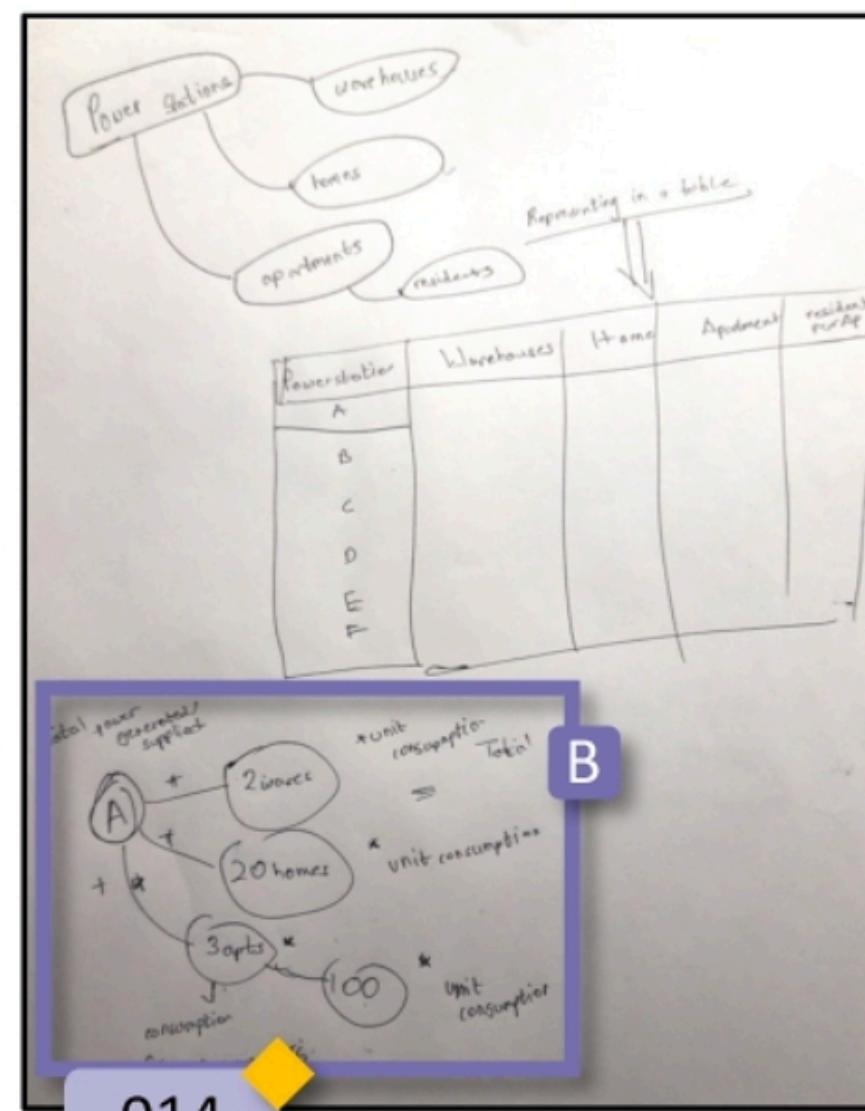
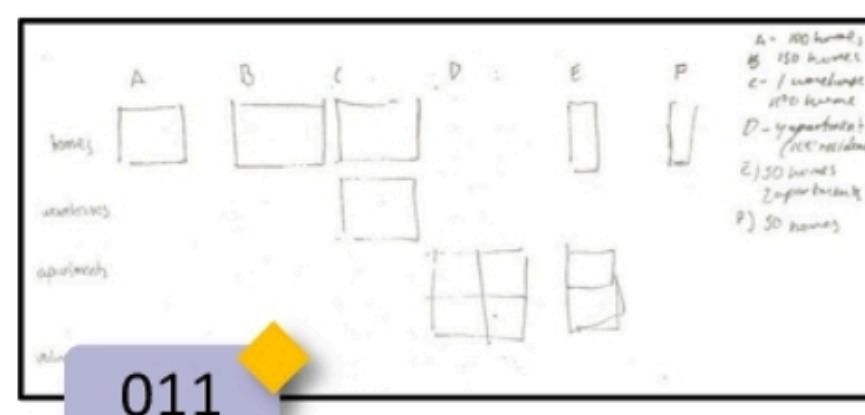
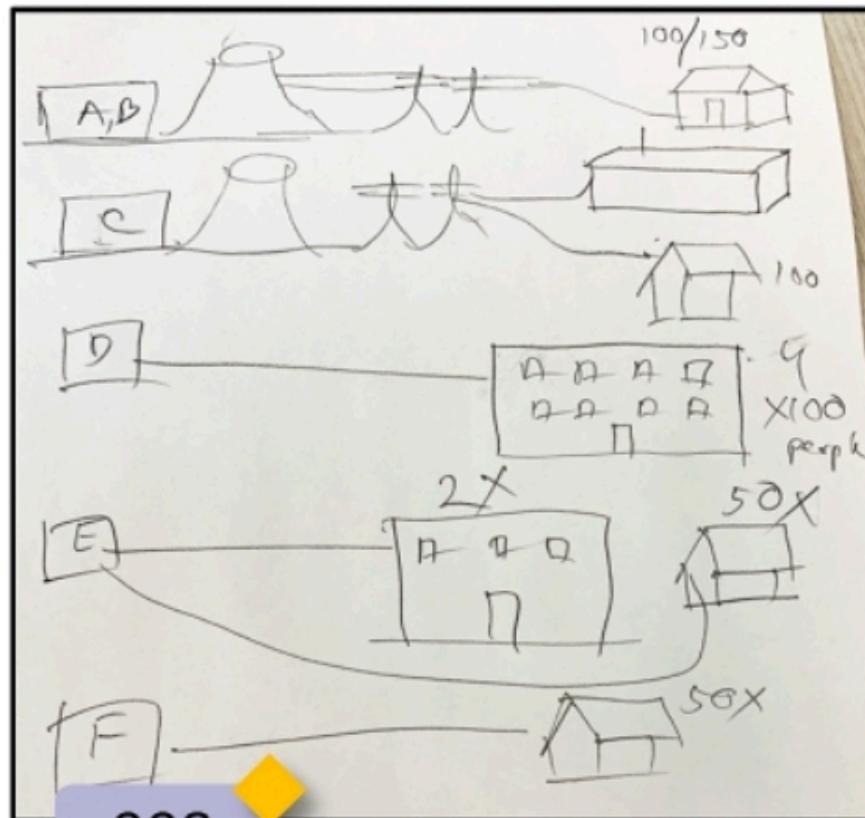
Participant ID

Second Sketch

Computing Participant

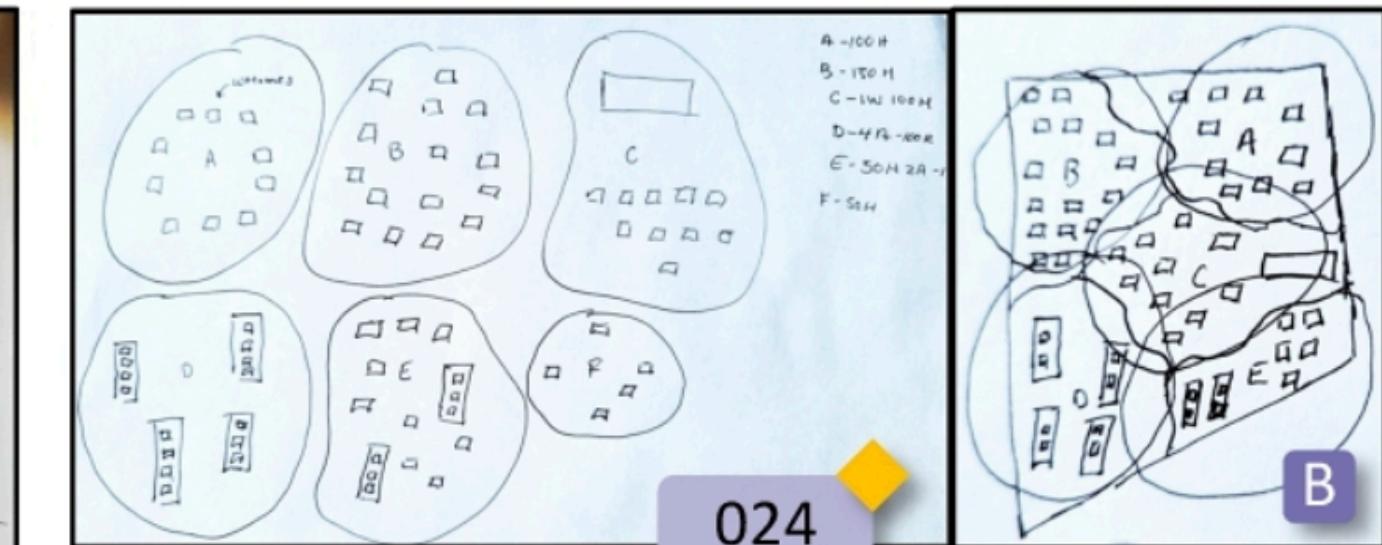
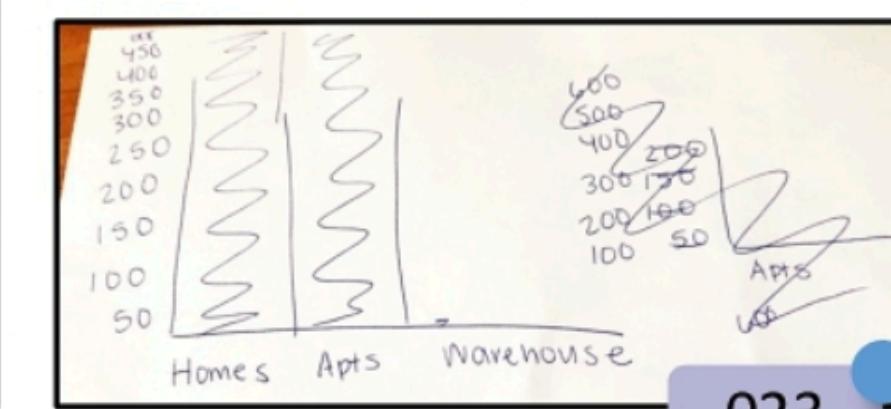
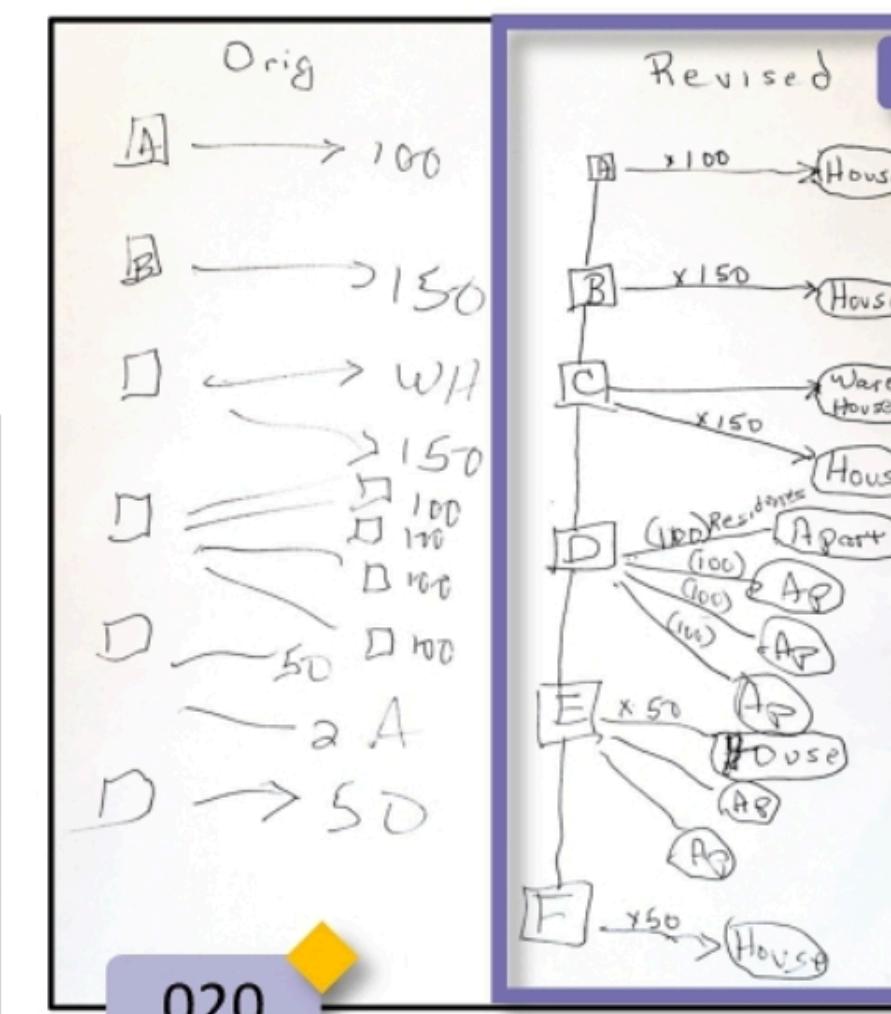
Non-Computing Participant

How do people conceptualize data?



	Homes	Apt	Whse
A	100		
B	150		
C	100		
D			4x100
E			1x100
F	50		

017



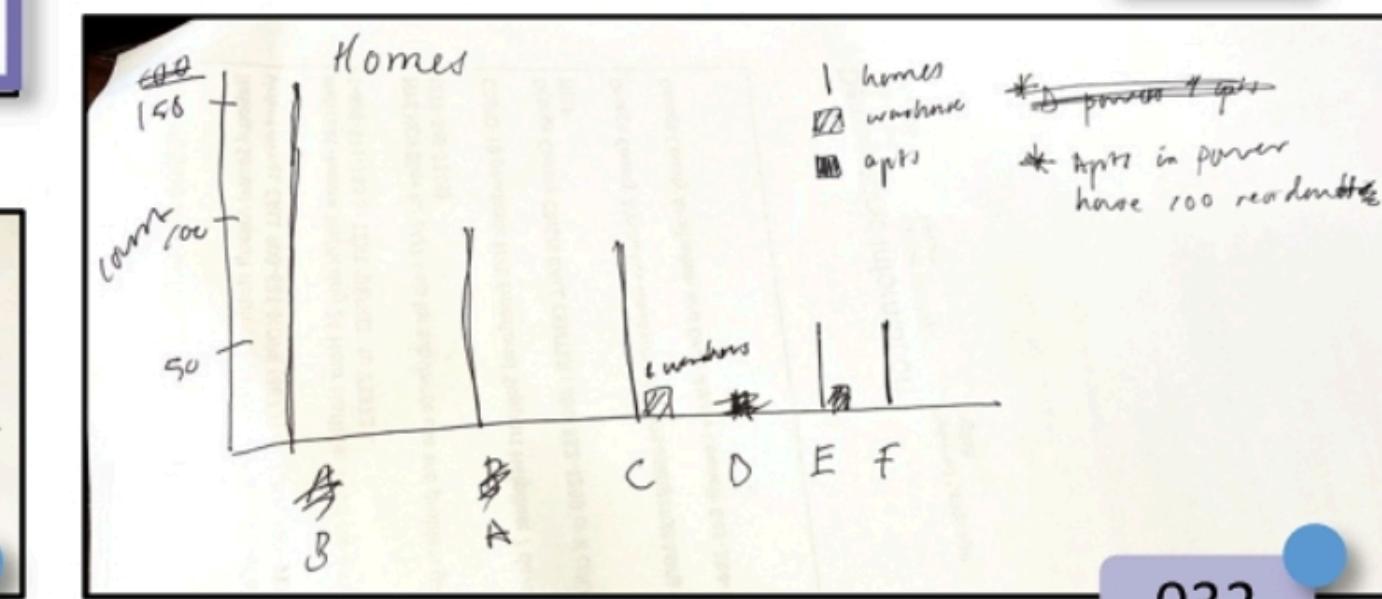
Power station	# of homes	# residents	# Apts	# warehouse
	100			
A			150	
B				100
C			4	
D				
E				
F		50		

026

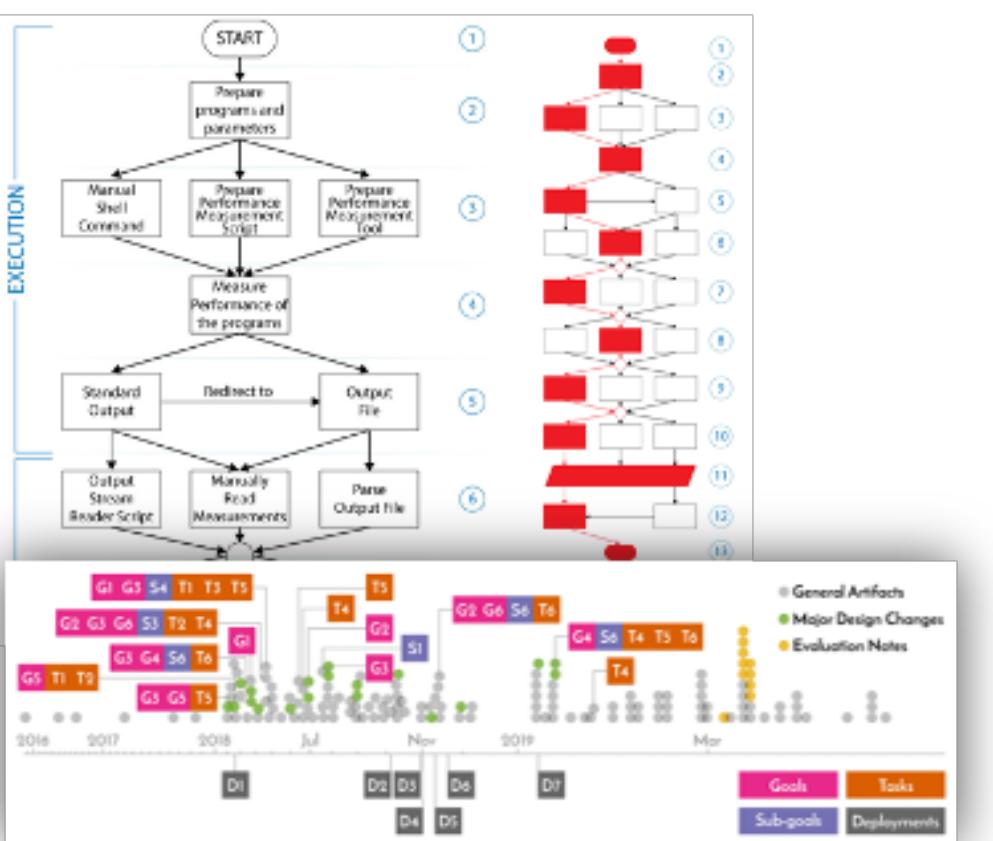
POWER STATIONS						
X	A	B	C	D	E	F
Home	100	150	100		50	50
Apt*				4	2	
Warehouse					1	

* 100 residents per complex

029



Vis Research Methodology

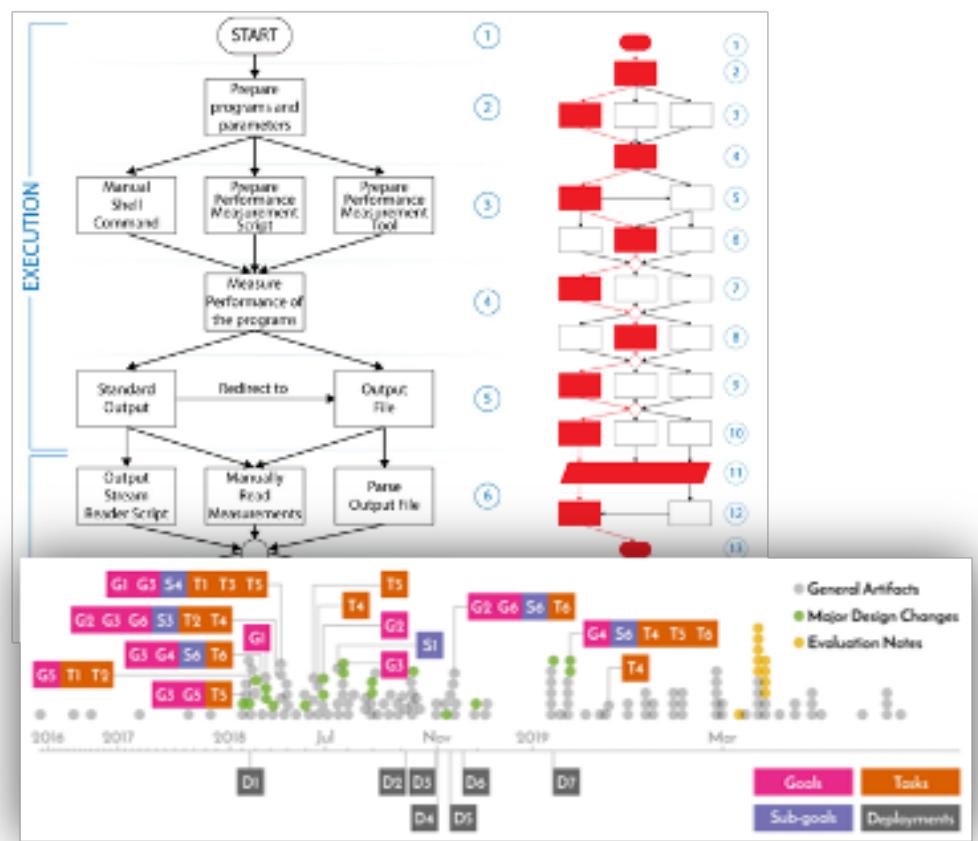


Vis Design Studies (InfoVis 2019, Ongoing)

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

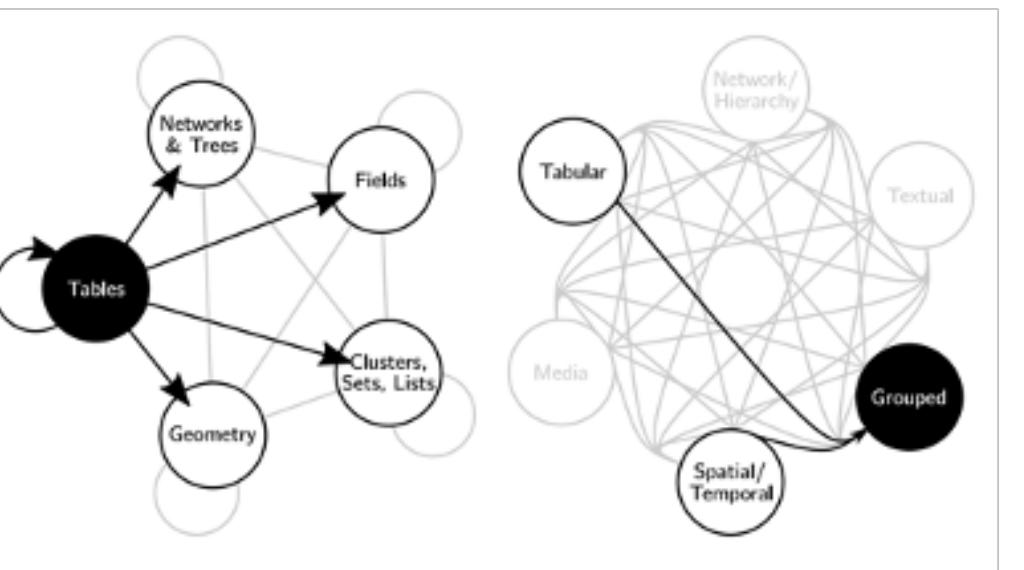


Vis Design Studies (InfoVis 2019, Ongoing)

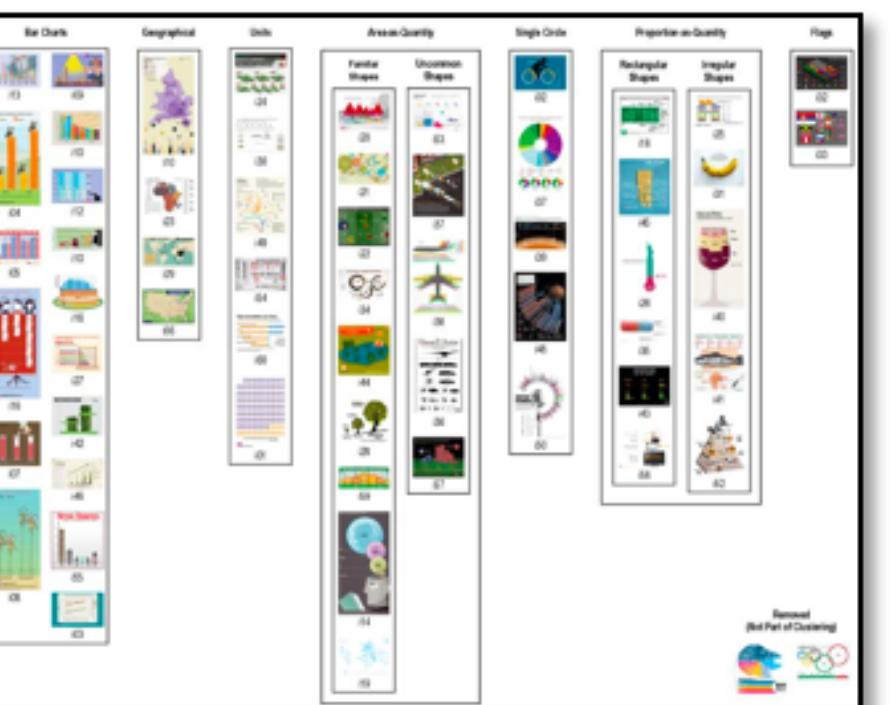
Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem 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 concepts

Contribution Types (CG&A 2019)

How People Conceptualize Data

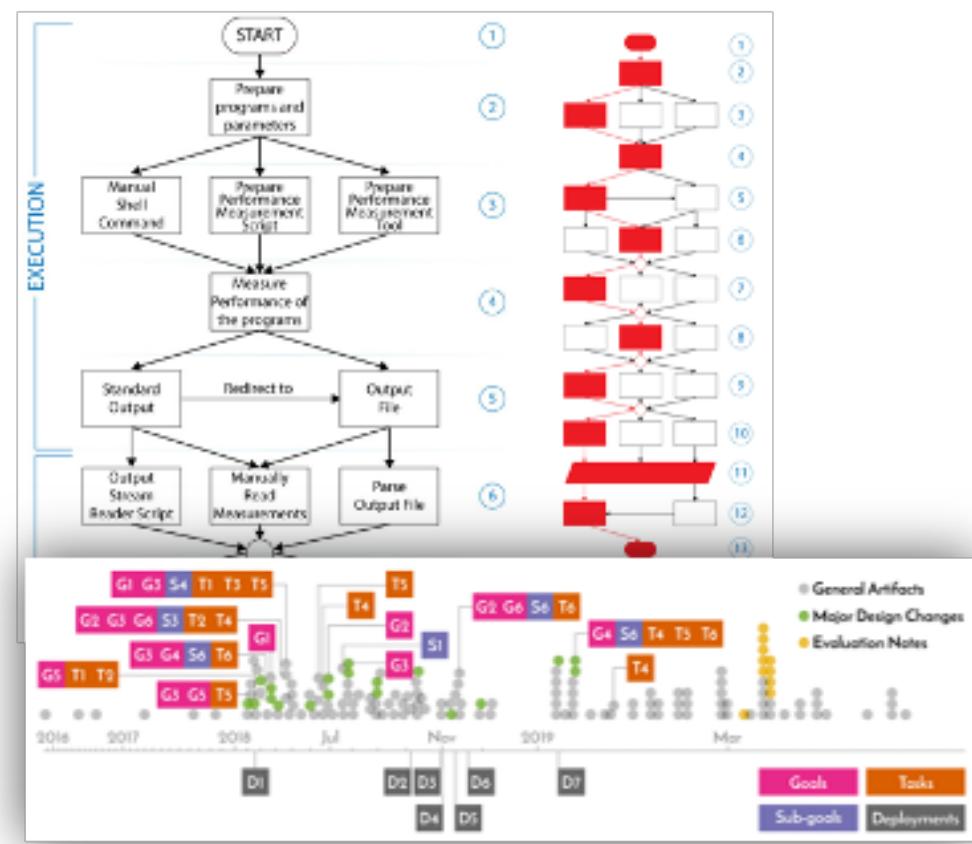


Latent Data Abstractions (InfoVis 2020, Ongoing)



Classification of Infographics (Diagrams 2018)

Vis Research Methodology

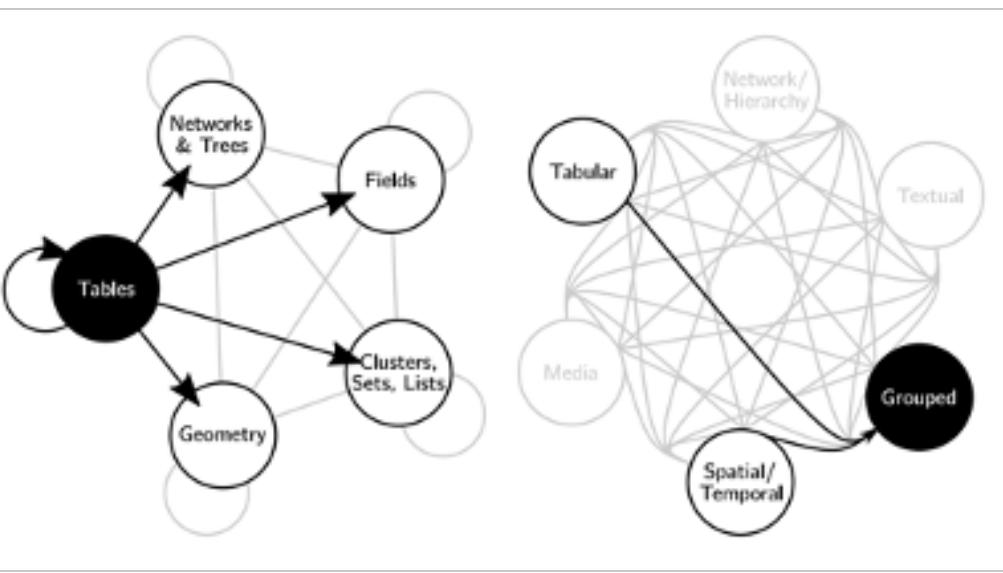


Vis Design Studies (InfoVis 2019, Ongoing)

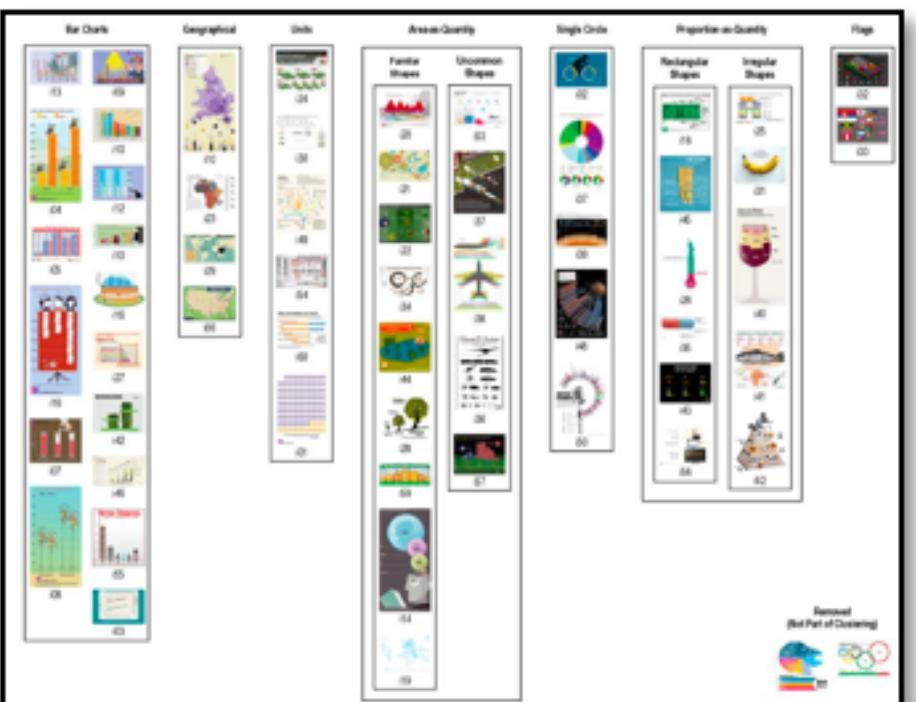
Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem 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

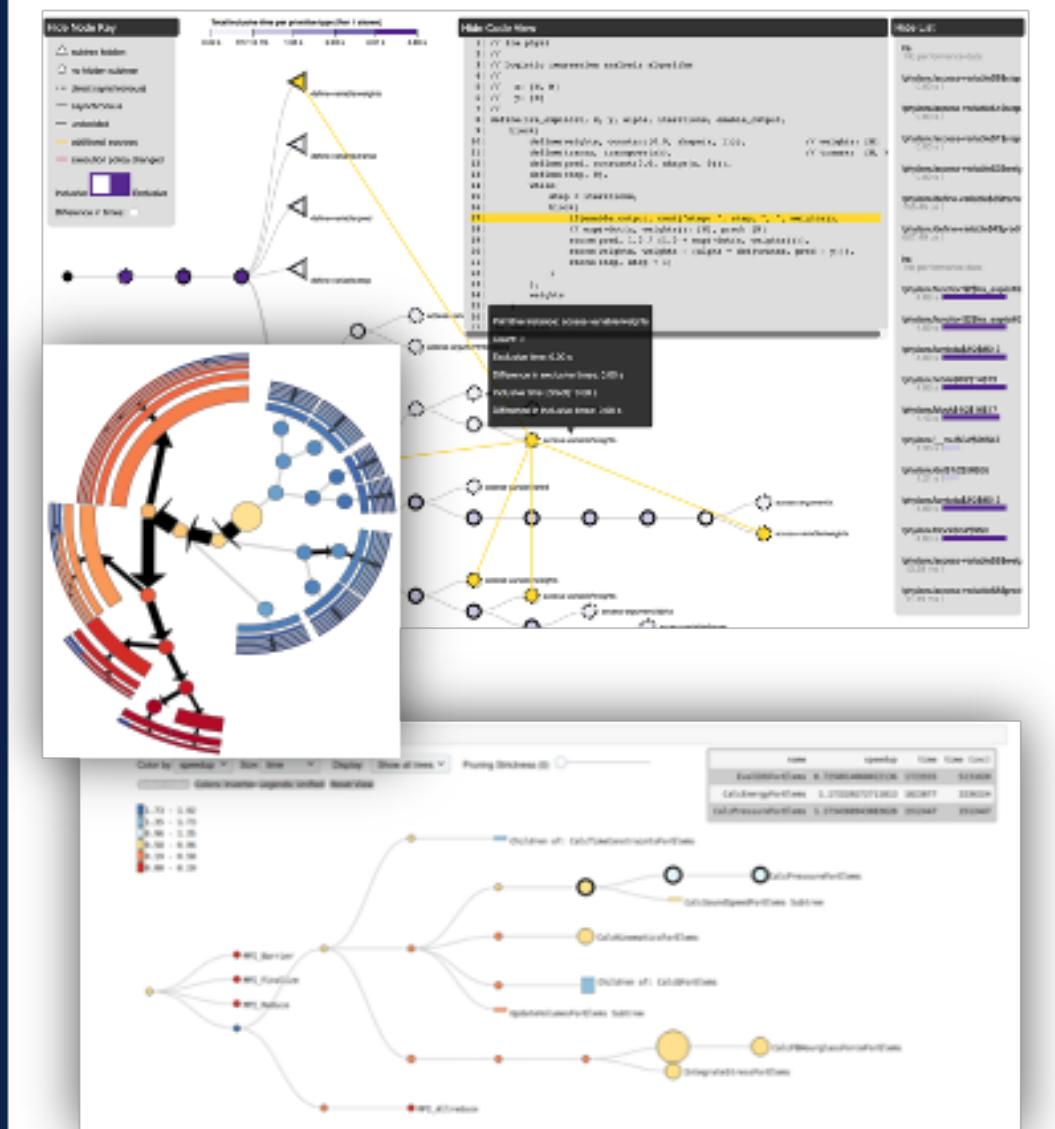


Latent Data Abstractions (InfoVis 2020, Ongoing)

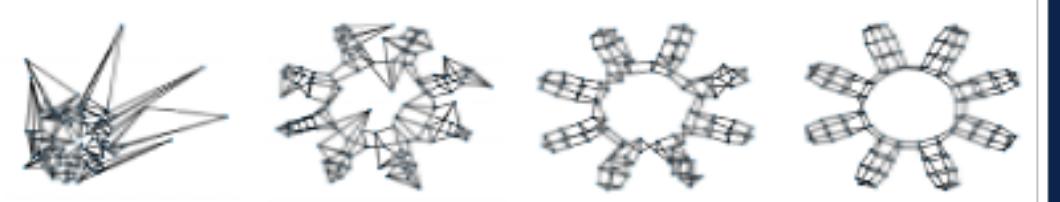


Classification of Infographics (Diagrams 2018)

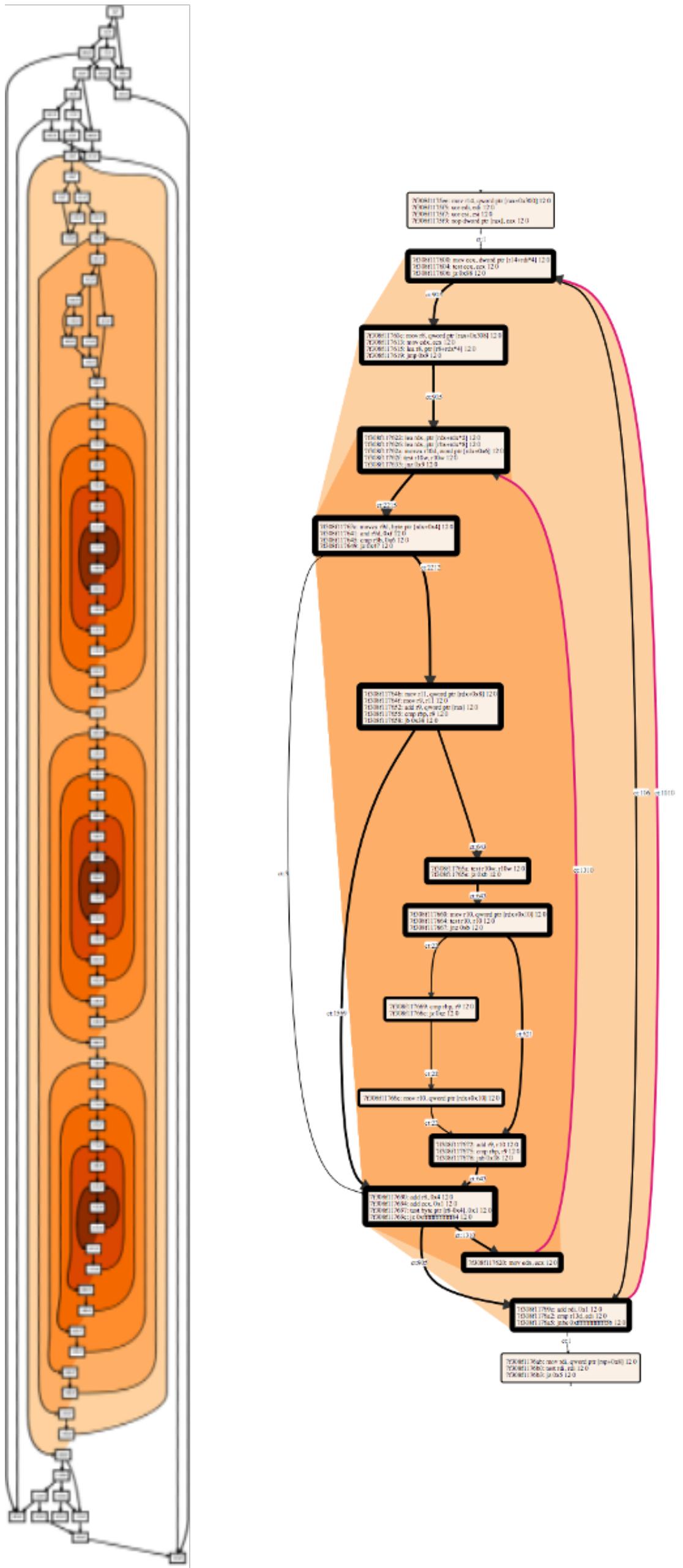
Trees & Graphs



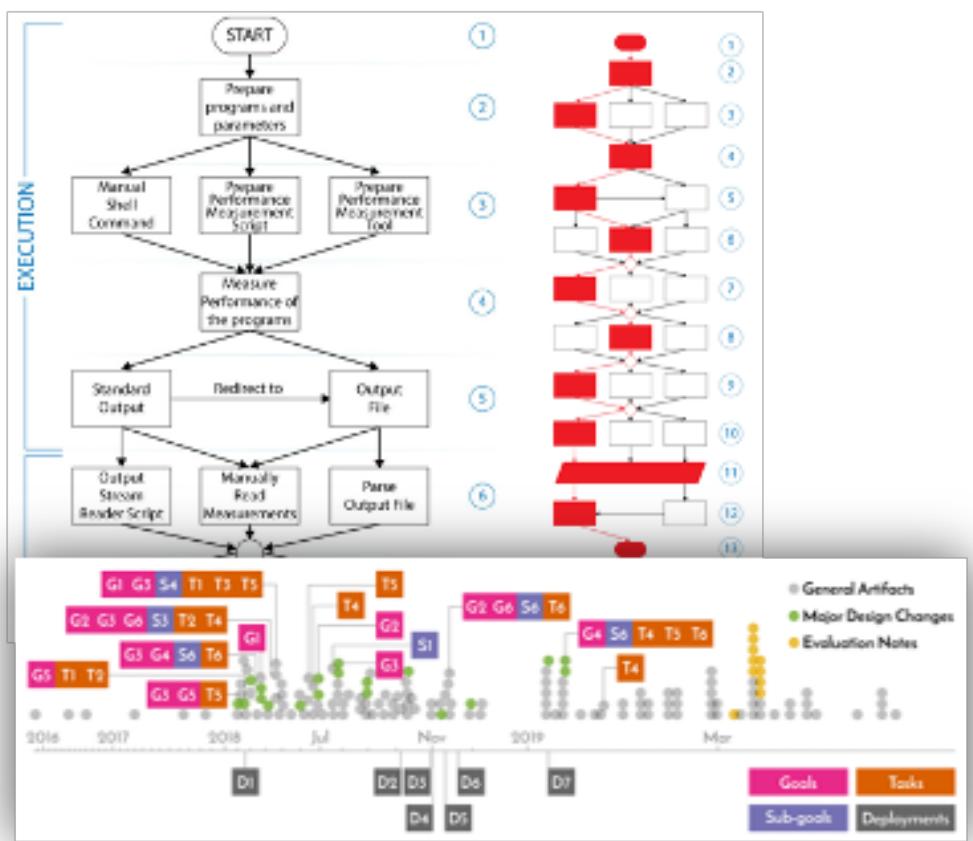
Trees in Computing (SC12, ProTools, Ongoing)



Stress-Plus-X Layout (Graph Drawing 2019)



Vis Research Methodology

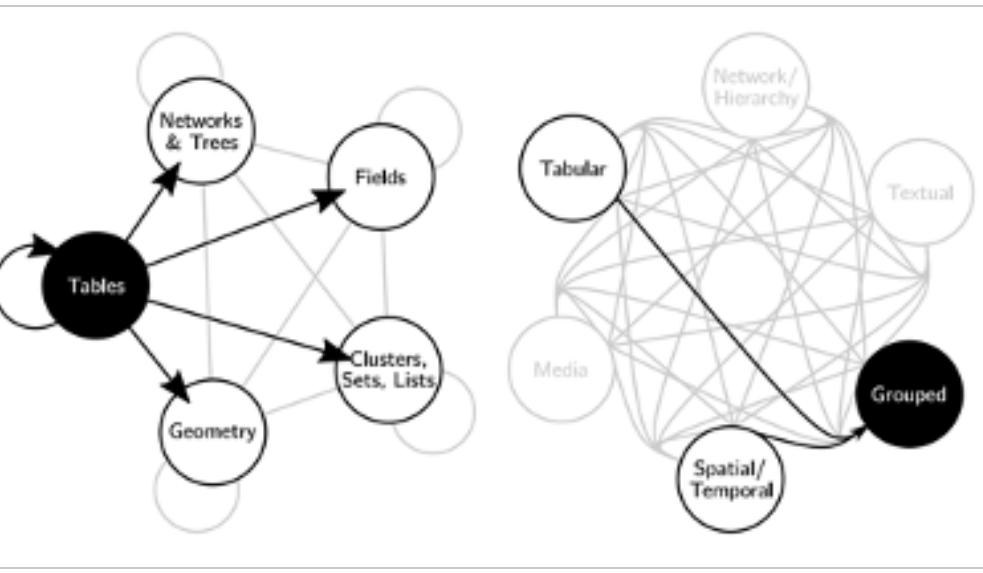


Vis Design Studies (InfoVis 2019, Ongoing)

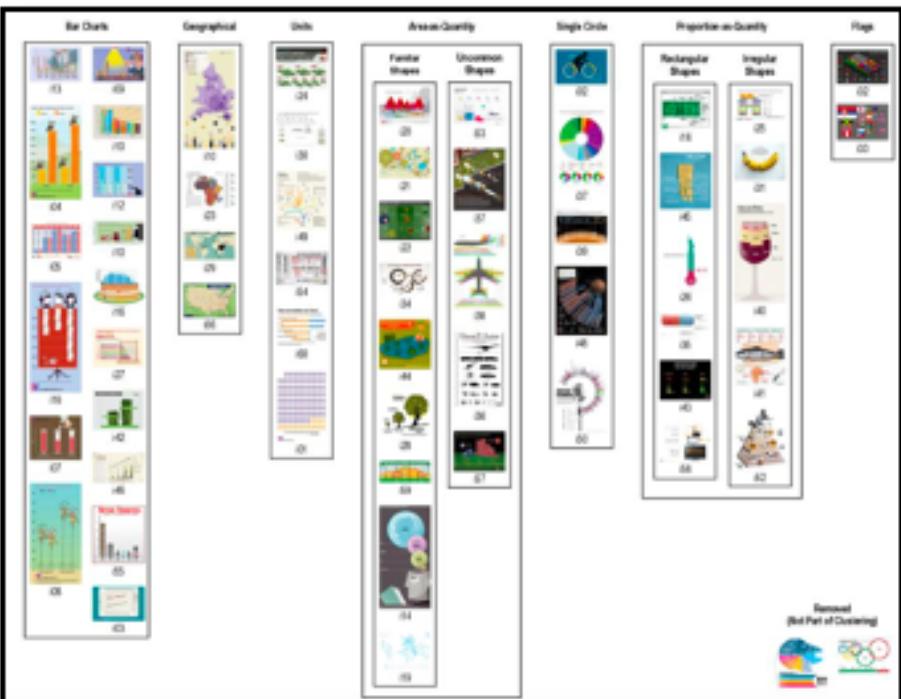
Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem 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

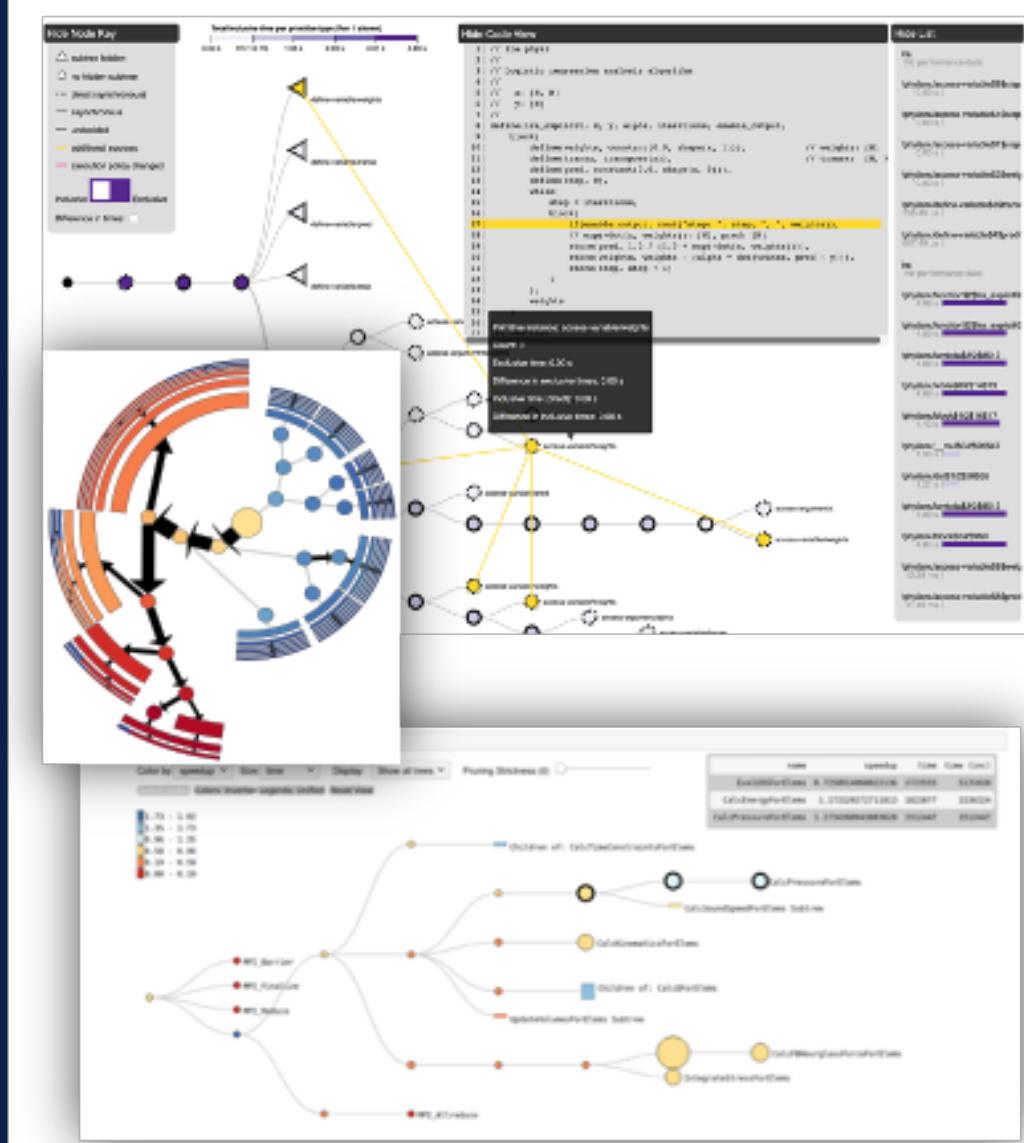


Latent Data Abstractions (InfoVis 2020, Ongoing)



Classification of Infographics (Diagrams 2018)

Trees & Graphs

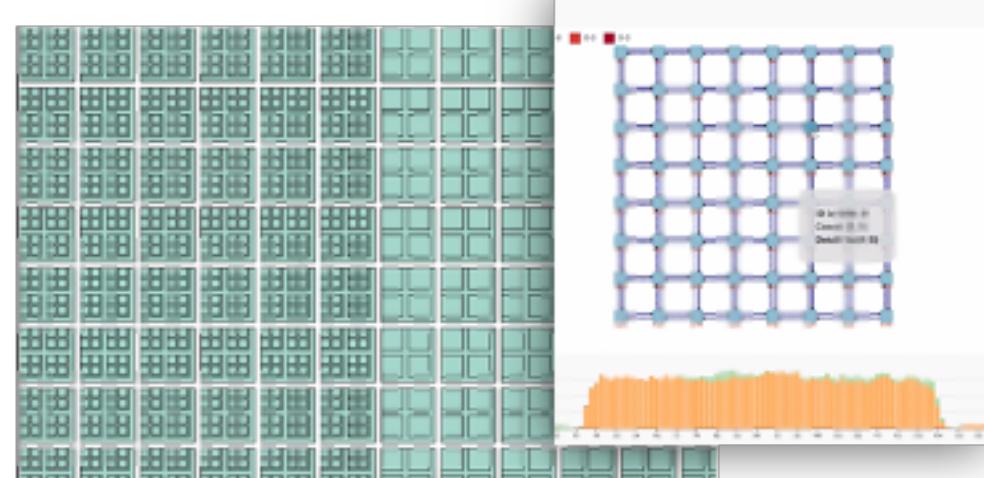


Trees in Computing (SC12, ProTools, Ongoing)

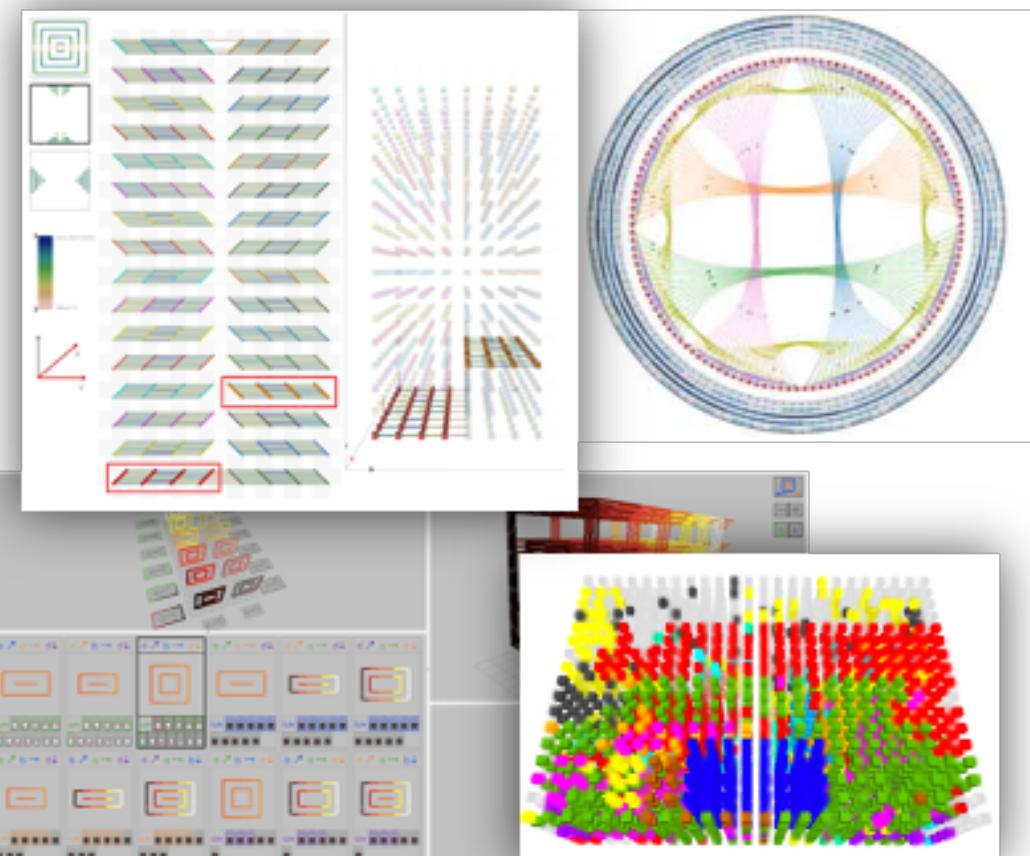


Stress-Plus-X Layout (Graph Drawing 2019)

Hardware



Data Movement on GPUs (Ongoing)



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

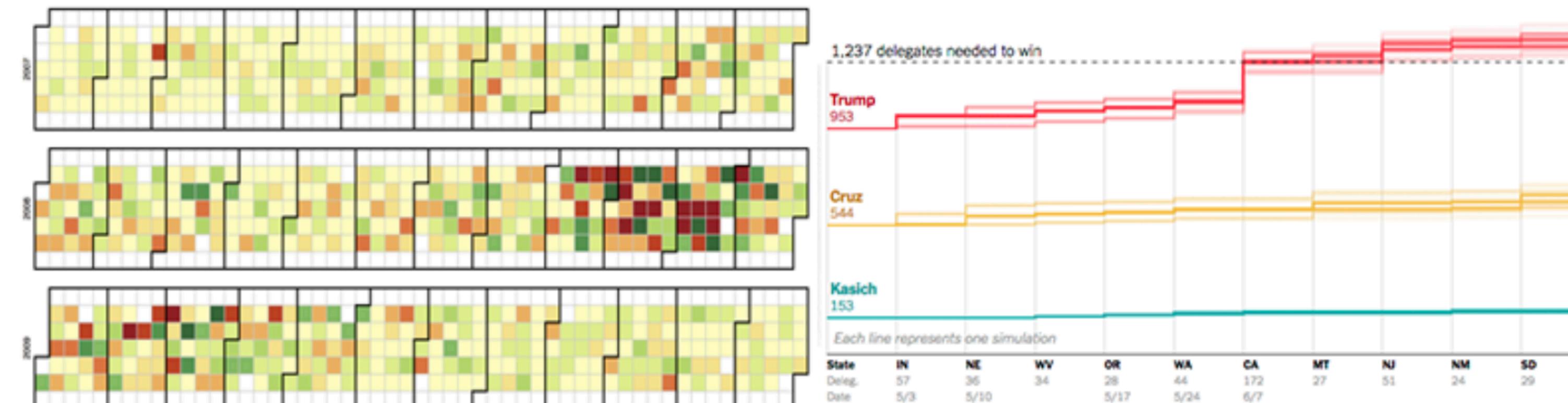
Course Structure

Information: datasciencecourse.net

DS Introduction to Data Science



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 2022** semester at the University of Utah, cross-listed between **Mathematics (MATH 4100)** and **Computing (COMP 5360)**.

This class is taught in person. We will also stream the class via Zoom (and Youtube). We will use Piazza for asynchronous communication, and Canvas for announcement and submission of assignments. All classes and sections are also archived online. Students are highly encouraged to attend the class live (either in person or via Zoom) due to in-class activities.

You should also own a notebook computer that you can bring to class.

Communications

Canvas

Announcements.
Assignment submission and grading.

Piazza

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 **Piazza** for questions related to course materials
- Private Piazza 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

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

Bonus 5-15 minutes mini-activities, in class if time.

Typically a coding problem.

Meant to be done the same day as lecture.

Complete 5 activities (out of ~10-15)

Lectures

Tue / Th 3:40 - 5:00 PM
MDT in person in L101.

Bring your computer with
Python/Conda installed
to follow along with
lectures.
(see HWO)

Python Basics

Functions

In math, functions transform an input to an output as defined by the property of the function, like this:

$$f(x) = x^2 + 3$$

In programming, functions can do exactly this, but are also used to execute “subroutines”, i.e., to execute pieces of code in various order and under various conditions. Functions in programming are very important for structuring and modularizing code.

In computer science, functions are also called “procedures” and “methods” (there are subtle distinctions, but nothing we need to worry about at this time).

The following Python function, for example, provides the output of the above defined function for every valid input:

```
In [ ]: def f(x):
           result = x ** 2 + 3
           return result
```

We can now run this function with multiple input values:

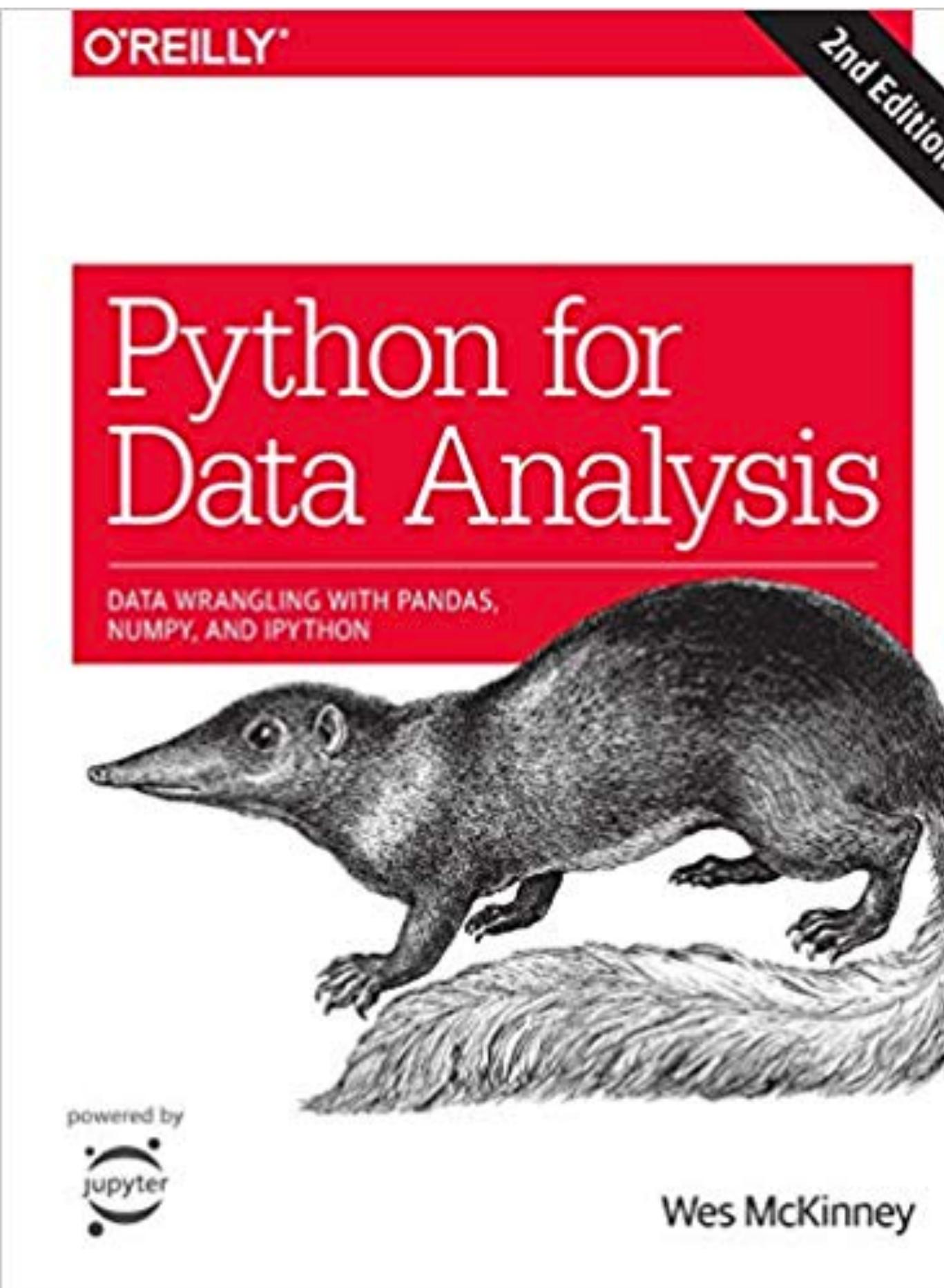
```
In [ ]: print(f(2))
           print(f(3))
           f(5)
```

Let's take a look at this function. The first line

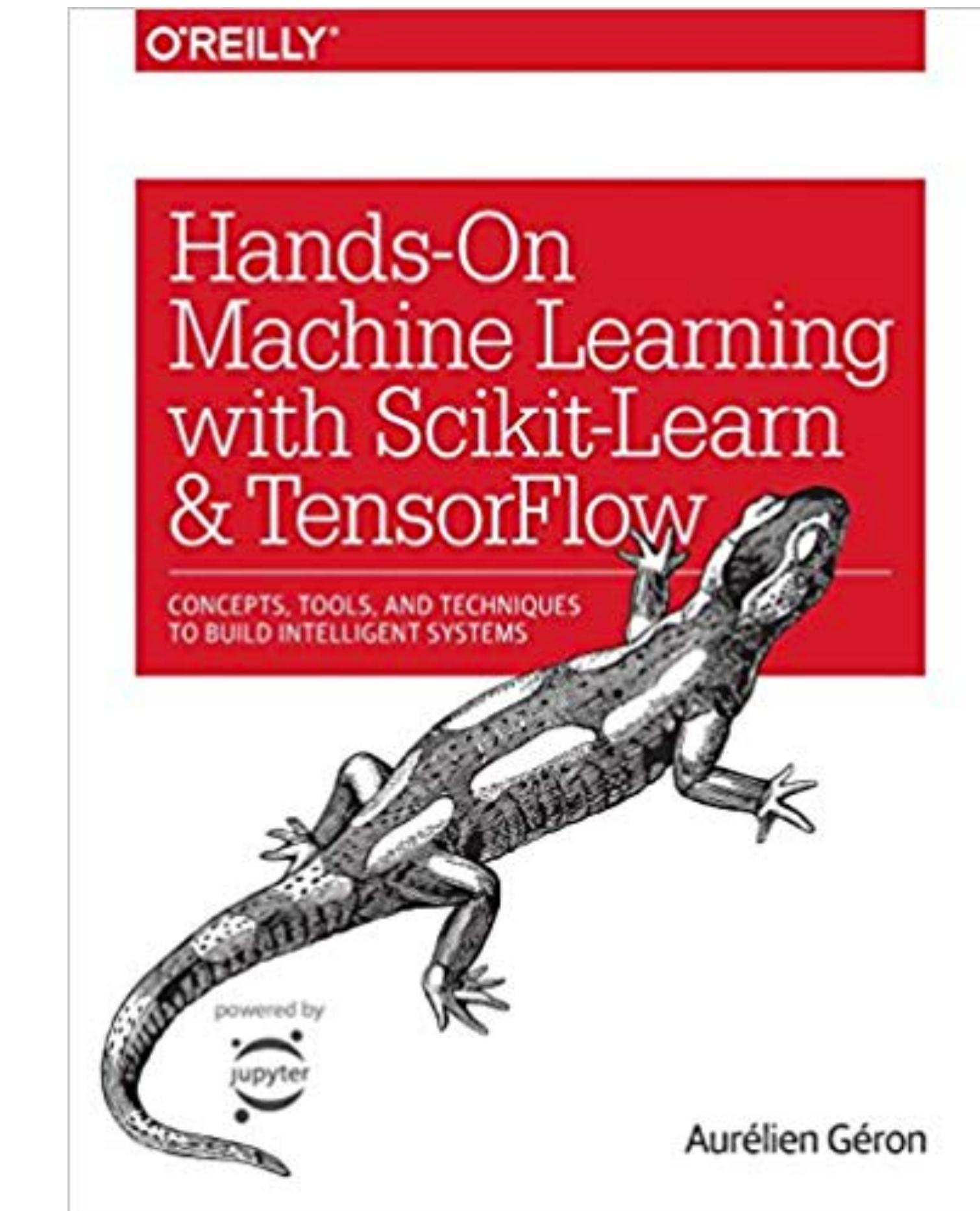
```
def f(x):
```

defines the function of name `f` using the `def` keyword. The name we use (`f` here) is largely arbitrary, but following good software engineering practices it should be something meaningful. So instead of `f`, `square_plus_three` would be a better function name in this case.

Books



Primary Text for Readings
Available for free on Campus: [link](#)



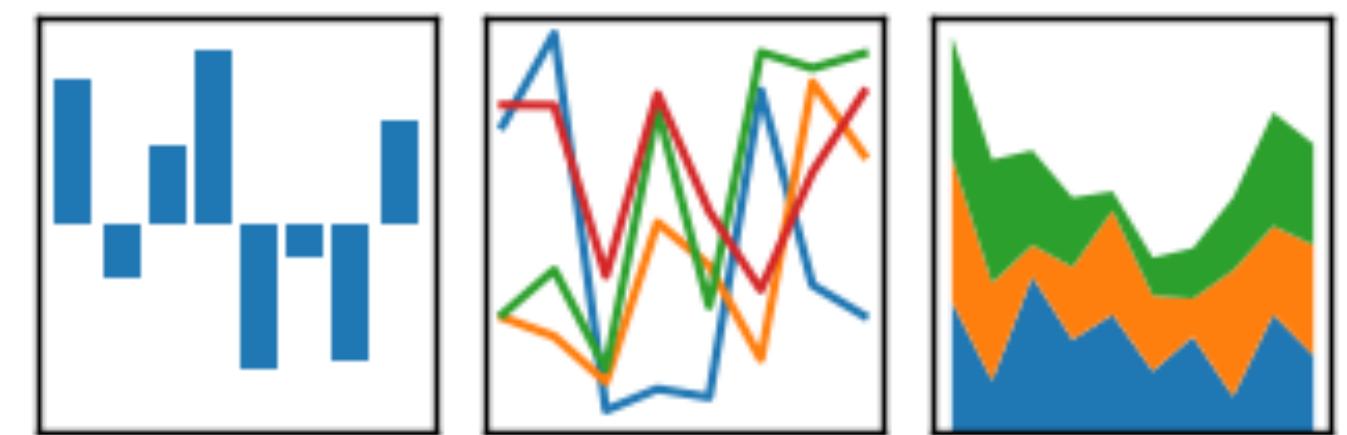
Supplementary Text
Available for free on Campus: [link](#)

Programming



pandas

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



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>

Academic Misconduct / 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 teammates'). For example, you must **create 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 or online course materials sites like Chegg or CourseHero.

See also the SoC Academic Misconduct Policy:

http://www.cs.utah.edu/wp-content/uploads/2014/12/cheating_policy.pdf

The misconduct will be reported to your home department in writing.

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

Coding Help Online

The homework are designed to need only the lectures, readings, and documentation for the libraries we use.

However, you may seek help from other online sites **with the exception** of course material aggregators like Chegg and Course Hero. If you refer to or copy code from these sites, **it must be fully cited** (see Syllabus).

You may use an AI coding assistant **BUT you must provide the assistant and the prompts** used for each code cell.

Note... sometimes these AI assistants give you harder-to-adapt code, be prepared to debug!

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

Homework 0

Introduction to Data Science - MATH 4100 / COMP 5360.

This homework is due by January 12th, 11:59pm

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, and it can be used inside of Jupyter Notebooks – more on the latter will follow later.

We also assume that you know the basics of how to work with a terminal / console. If you don't check out an introduction like [this](#).

First, we'll need to install some things:

1.1 Installing Python

Chances are, if you're on a mac, you already have Python installed. You can simply try to run python from a console by running

```
$ python
```



However, as most software, Python comes in different versions and is packaged differently depending on your needs. In this

Github

<https://github.com/datascience-course/2024-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/2024-datascience-homework
```

This will create a folder `2024-datascience-homework` 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 `2024-datascience-homework` directory and executing:

```
git pull
```

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

Next Week

Data Structures and Pandas

Introduction to Descriptive Statistics

HW1 due

Enrollment

Math 4100: 34

COMP 5360: 31

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.

Ice-breakers

Get to know the folks around you!

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 has programmed for 1 year? 5 years? More? Which languages?