

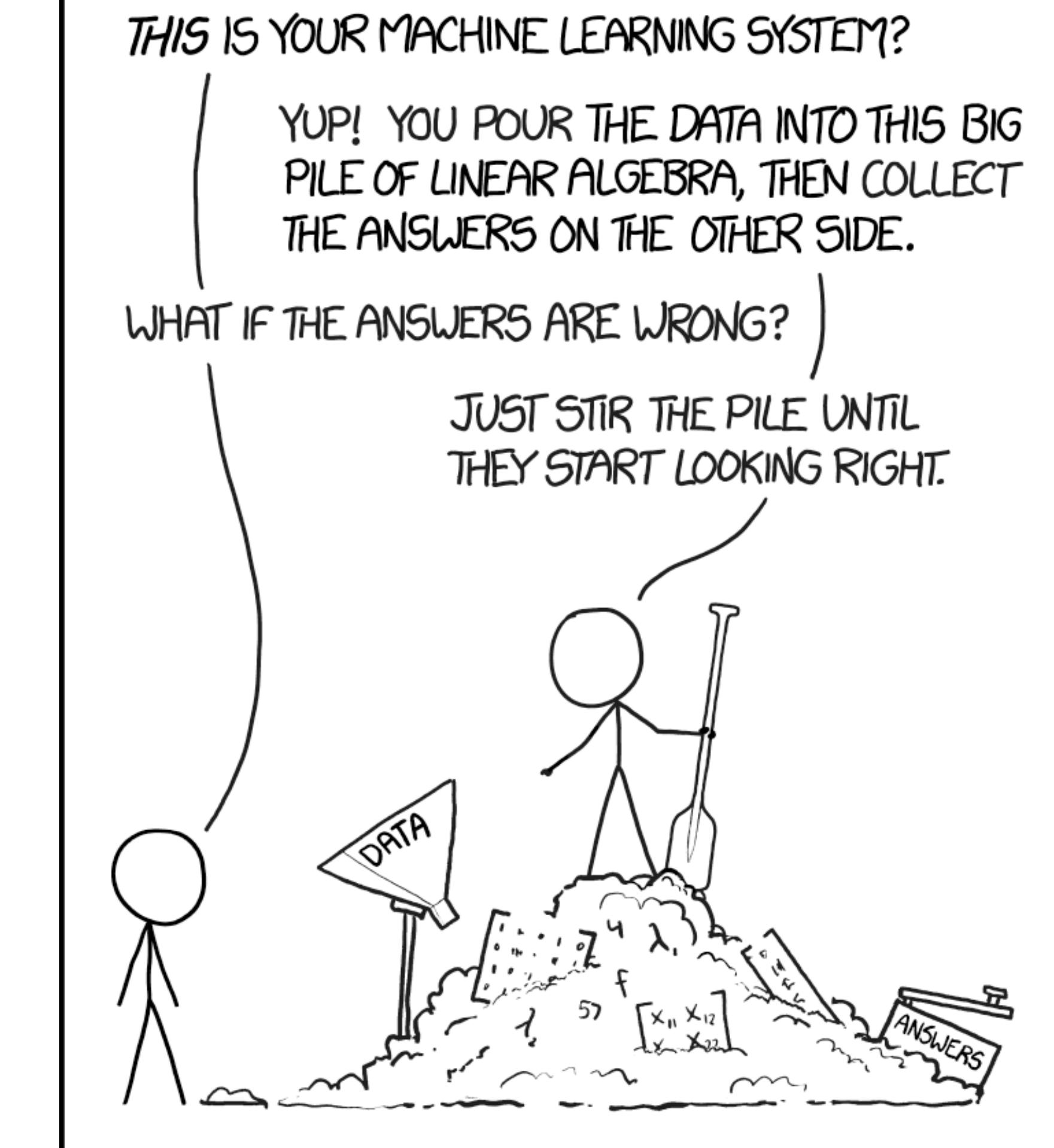
Introduction to Data Science

CS 5360 / Math 4100



Bei Wang Phillips
beiwang@cs.utah.edu

Anna Little
little@math.utah.edu



Hybrid Format

Course meets in person in WEB L101.

But lectures will also be live streamed via Zoom to a YouTube channel.

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

For activities we plan to have both in person groups and zoom break out rooms.

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.

Recording

Note that this and all future lectures are recorded in zoom 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 with the teaching staff prior to the lectures to be recorded.

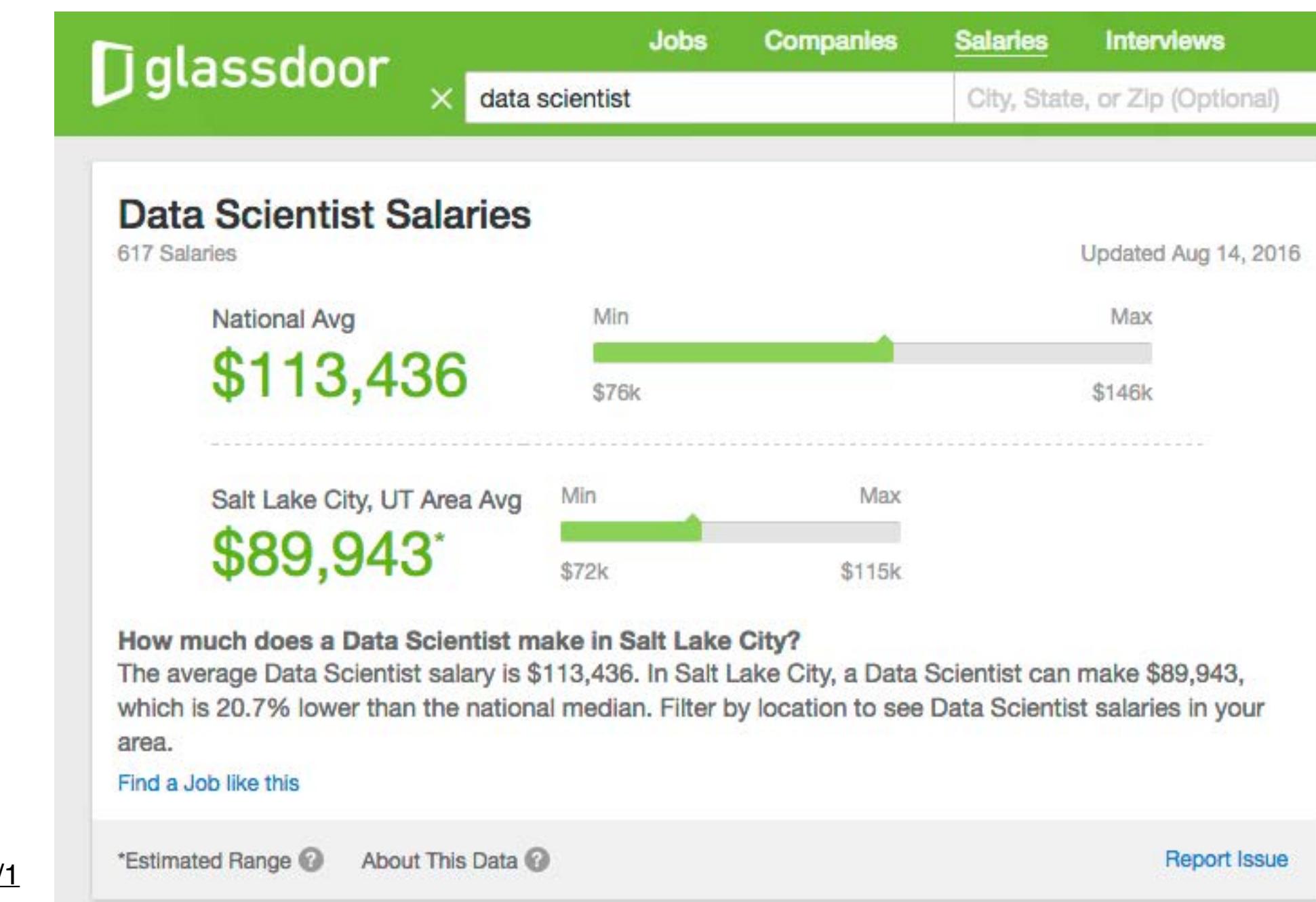
Camera Policy for Zoom Participants

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 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.

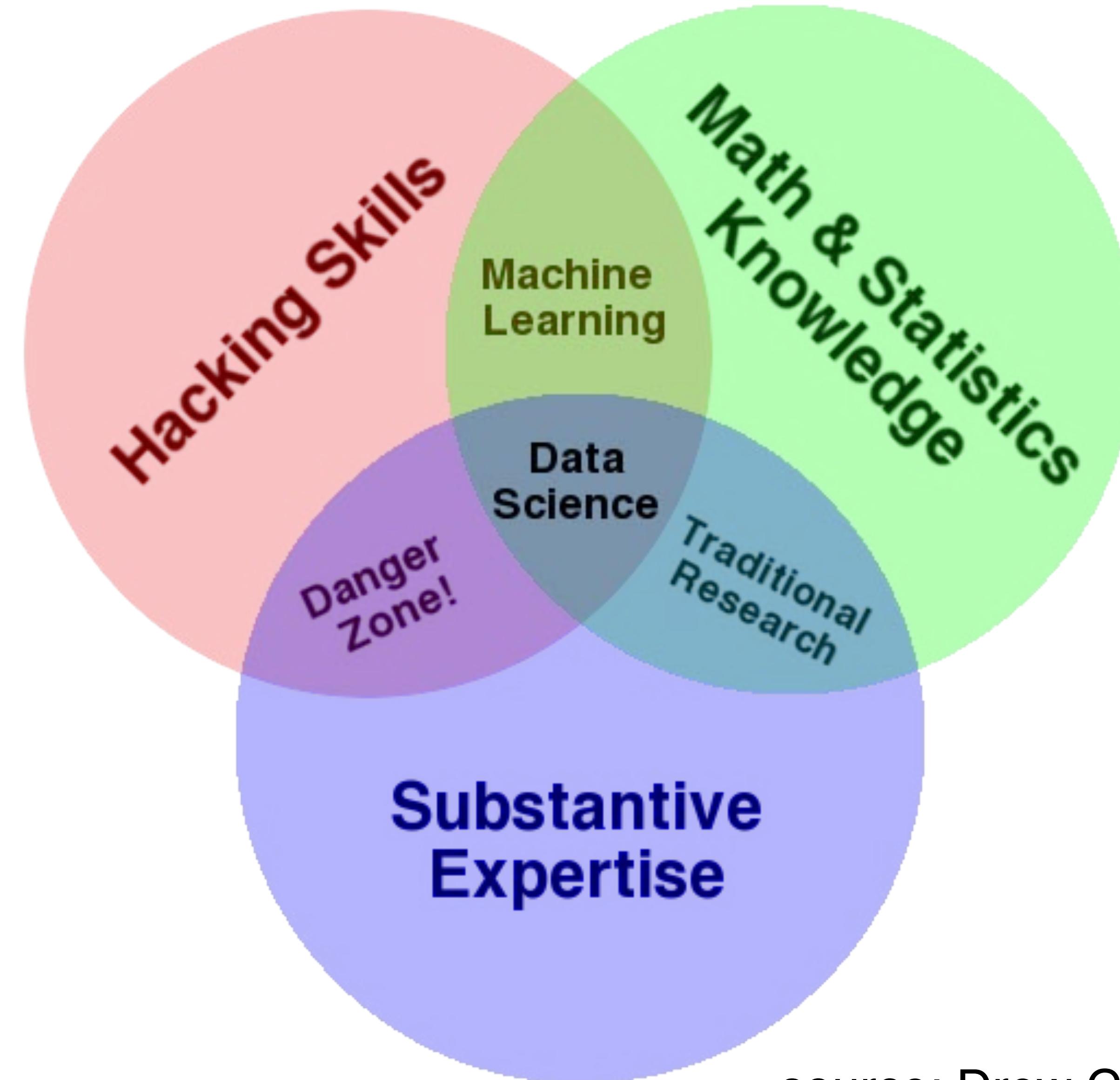


What is Data Science?

evidence collect approach
warehouse related derived system describes much
scale small new system whether system lives
term enough insights storage anything
continues often things velocity scope
refers diverse really machine think world like
means technological actually datasets complex
information society standard advances capturing
changes way used analytical every
change decisions organization behavior people fit
organization tool gather better lot essential
gather behavior better

ethical software commerce tools excel
otherwise disparate requires questions amounts
use many aspects later
something aspect later
possibilities get solve main computing
main meaning collecting relative deal
help deal
business products
processes methods issues
internet memory also
mean issues

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.

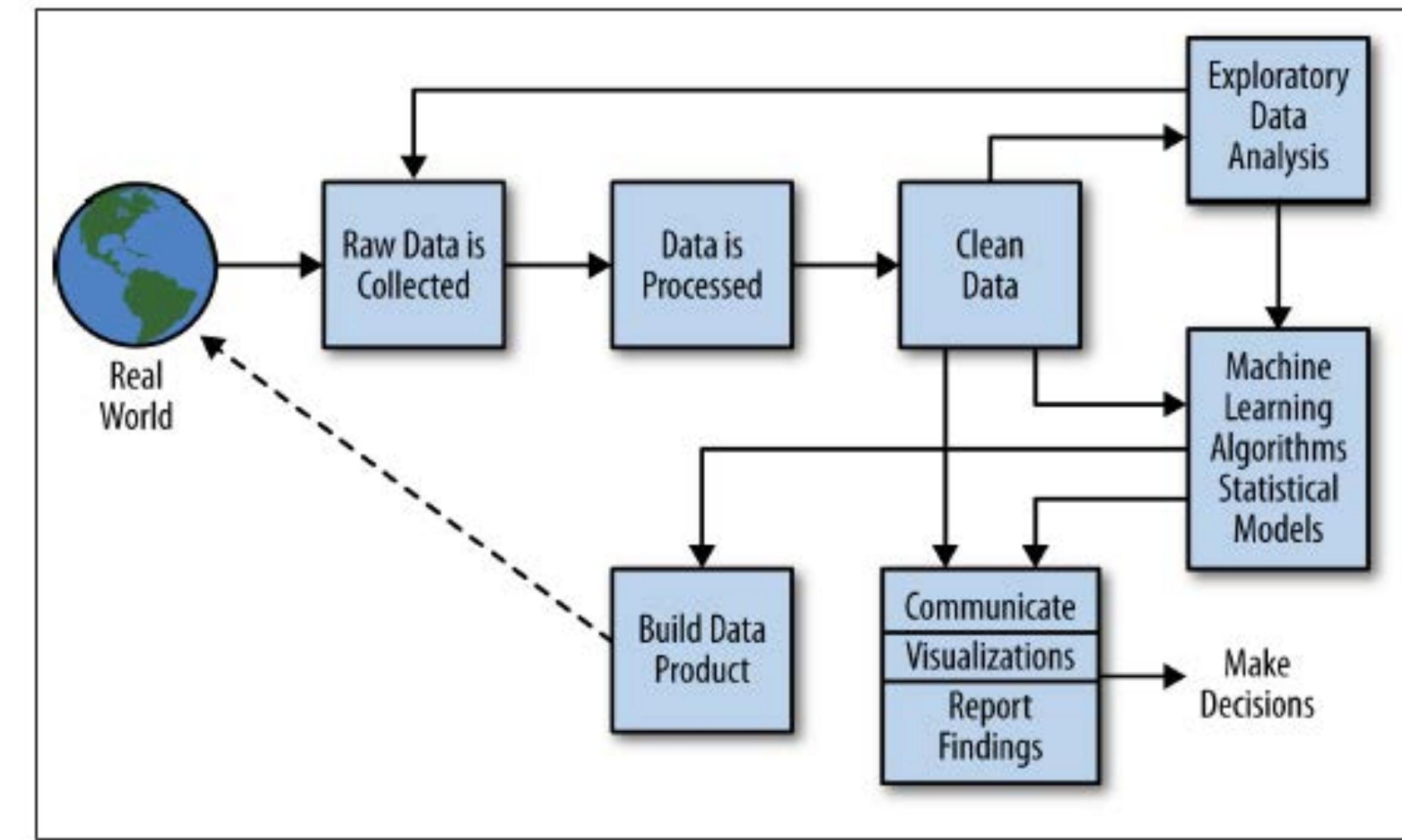


Figure 2-2. The data science process

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 main search results section shows a link titled "TOP 10 BEST CAT VIDEOS OF ALL TIME! - YouTube" with a thumbnail image of a cat and the text "www.youtube.com/watch?v... YouTube Sep 6, 2012 - Uploaded by WatchTheDaily We've scoured the internet and found the cutest and funniest cat videos of all time. Any we missed? Let us ...".

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.

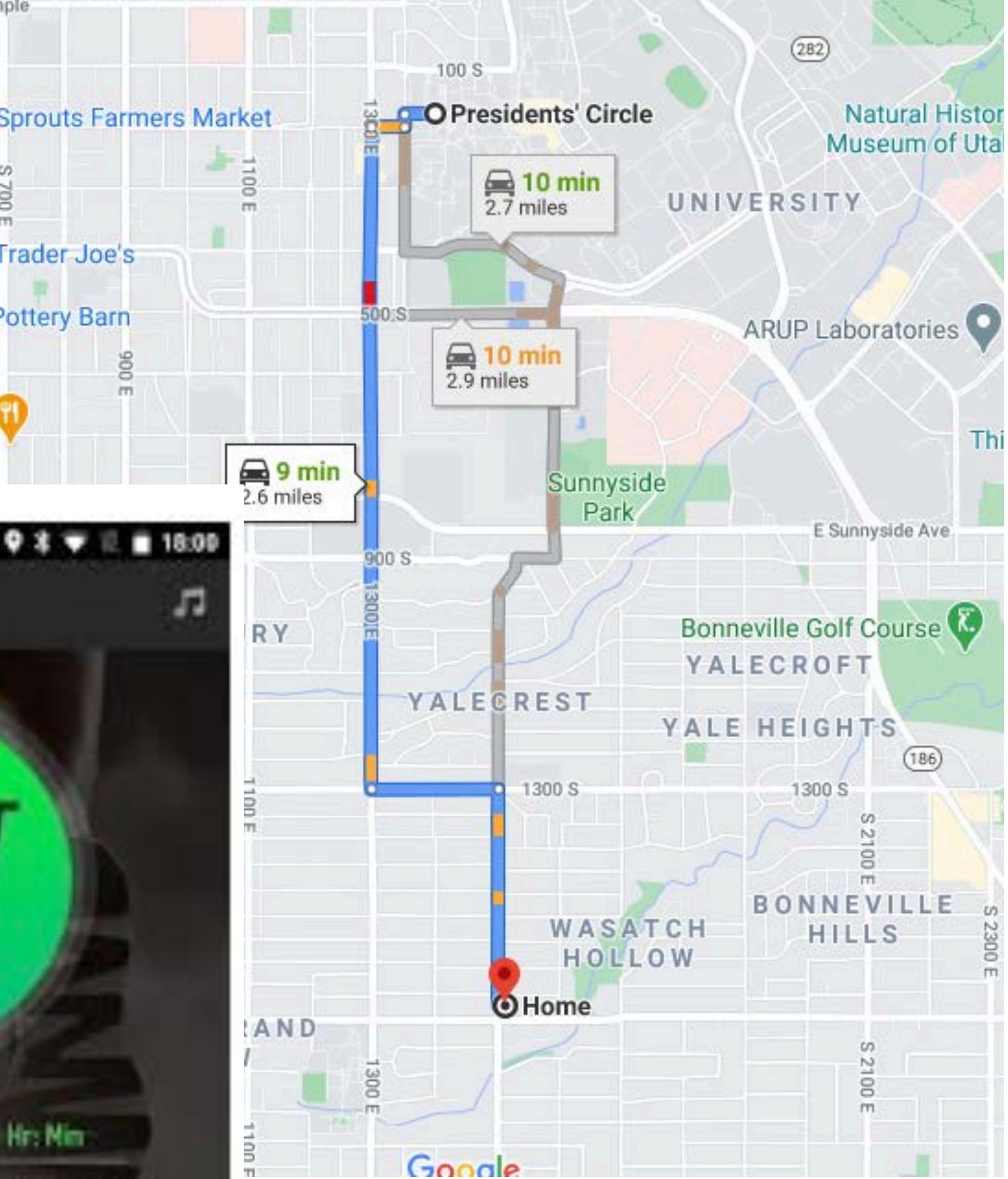
... and many more applications

Example: Personal Data

The Zillow search interface shows a map of Salt Lake City, UT with numerous red dots representing homes for sale. The map includes major roads like I-15, I-80, and I-215. A blue boundary is drawn around the central business district. Buttons for "Remove Boundary" and "Re-center" are visible. The search bar shows "Salt Lake City, UT". Other filters include "For Sale", "Price", "Beds & Baths", "Home type", and "More". The results page displays "193 Agent listings" and "48 Other listings" sorted by "Homes for You". Two house listings are shown:

12 hours ago
\$322,000
3 bds, 2 ba, 1,155 sqft - House for sale
1333 N Capistrano Dr, Salt Lake City, UT 84116
Utah Key Real Estate

3 days on Zillow
Utah Real Estate.co



Big Data in Science and Engineering

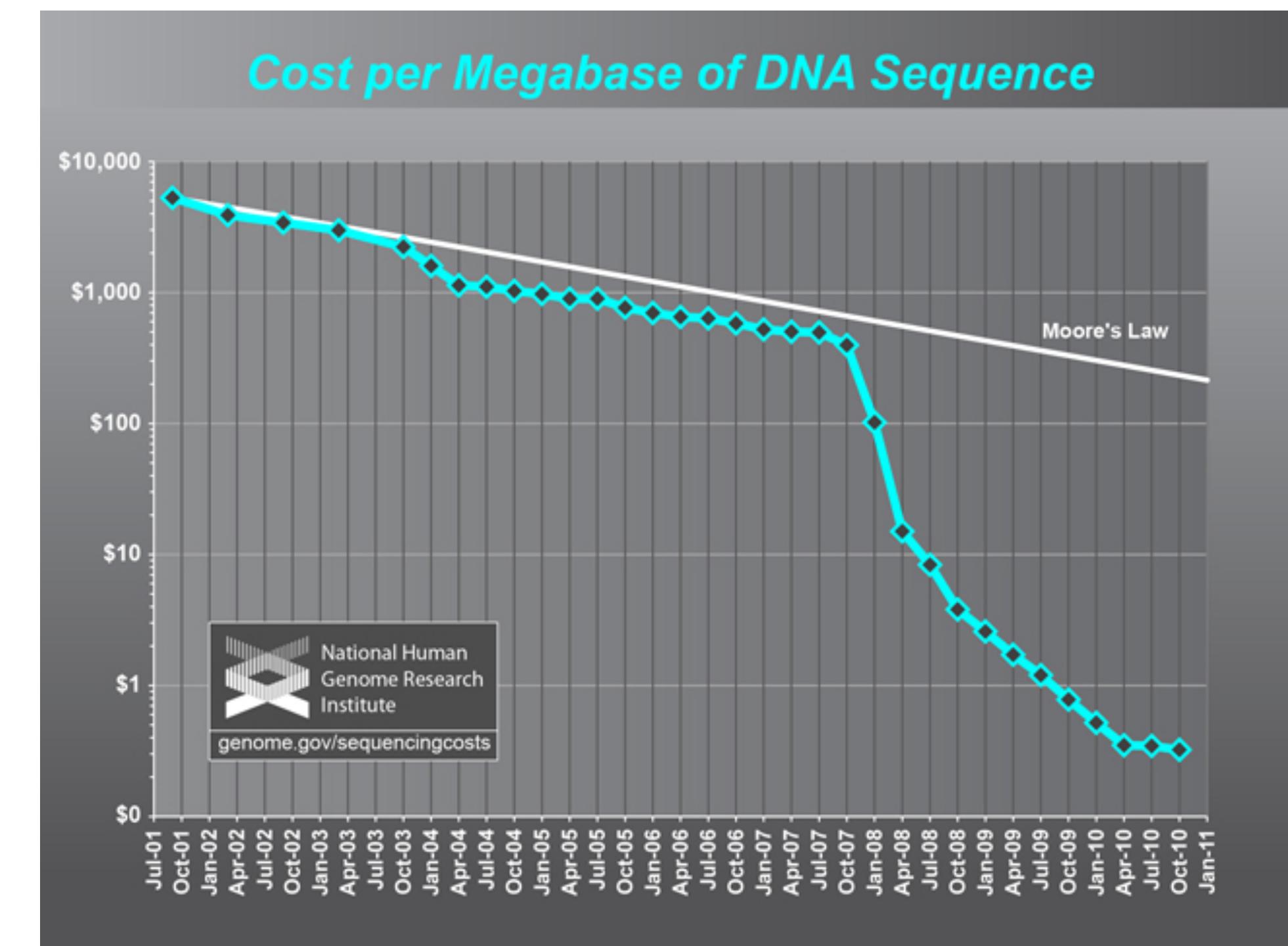
Big Data transformed science and engineering.

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 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

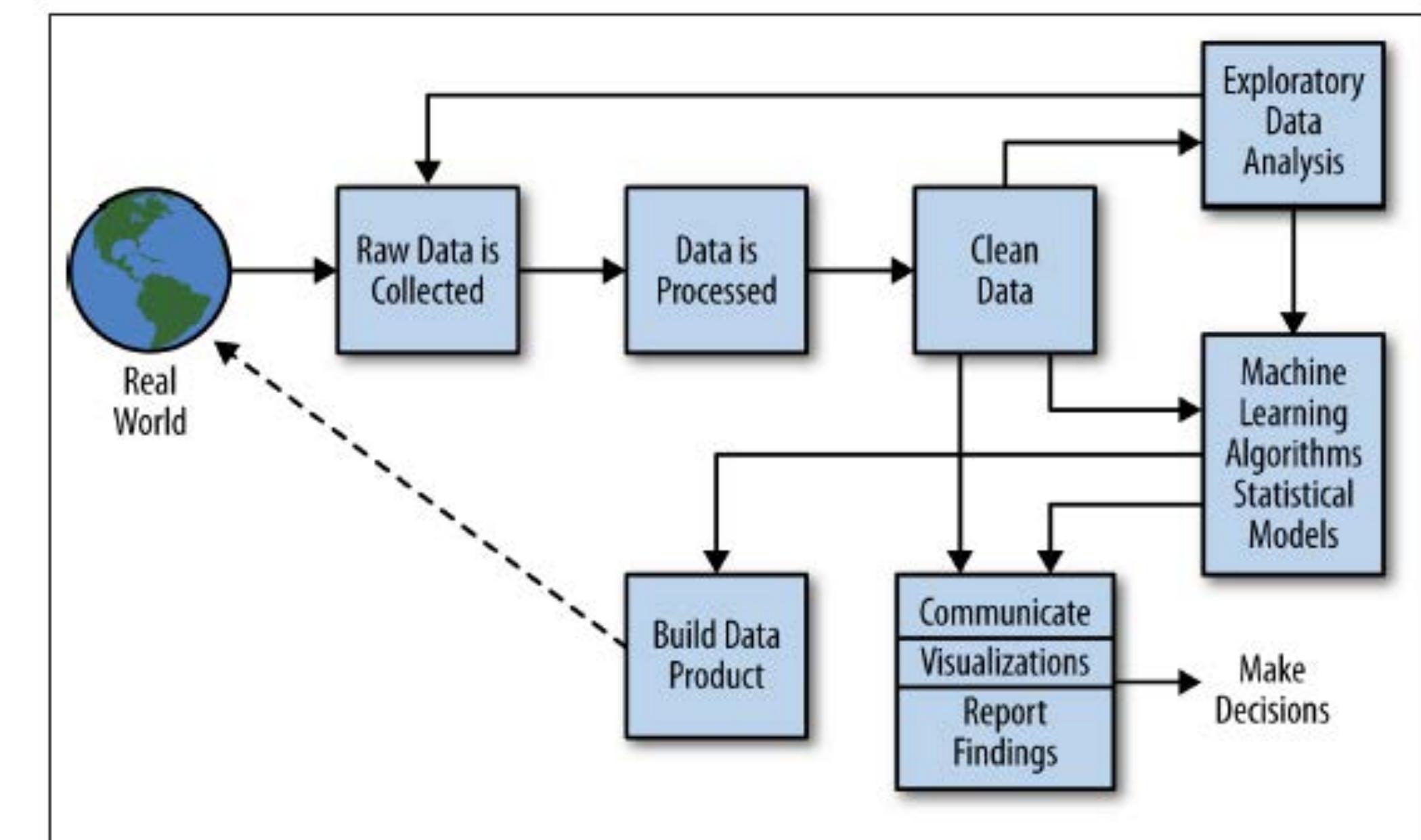


Figure 2-2. The data science process

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**

Teaching Assistants



CS PhD

Mingzhe Li



Math PhD

Haoyu Chen



CS MS

Prasanth Yalamanchili

Anna Little

Assistant Professor, Mathematics

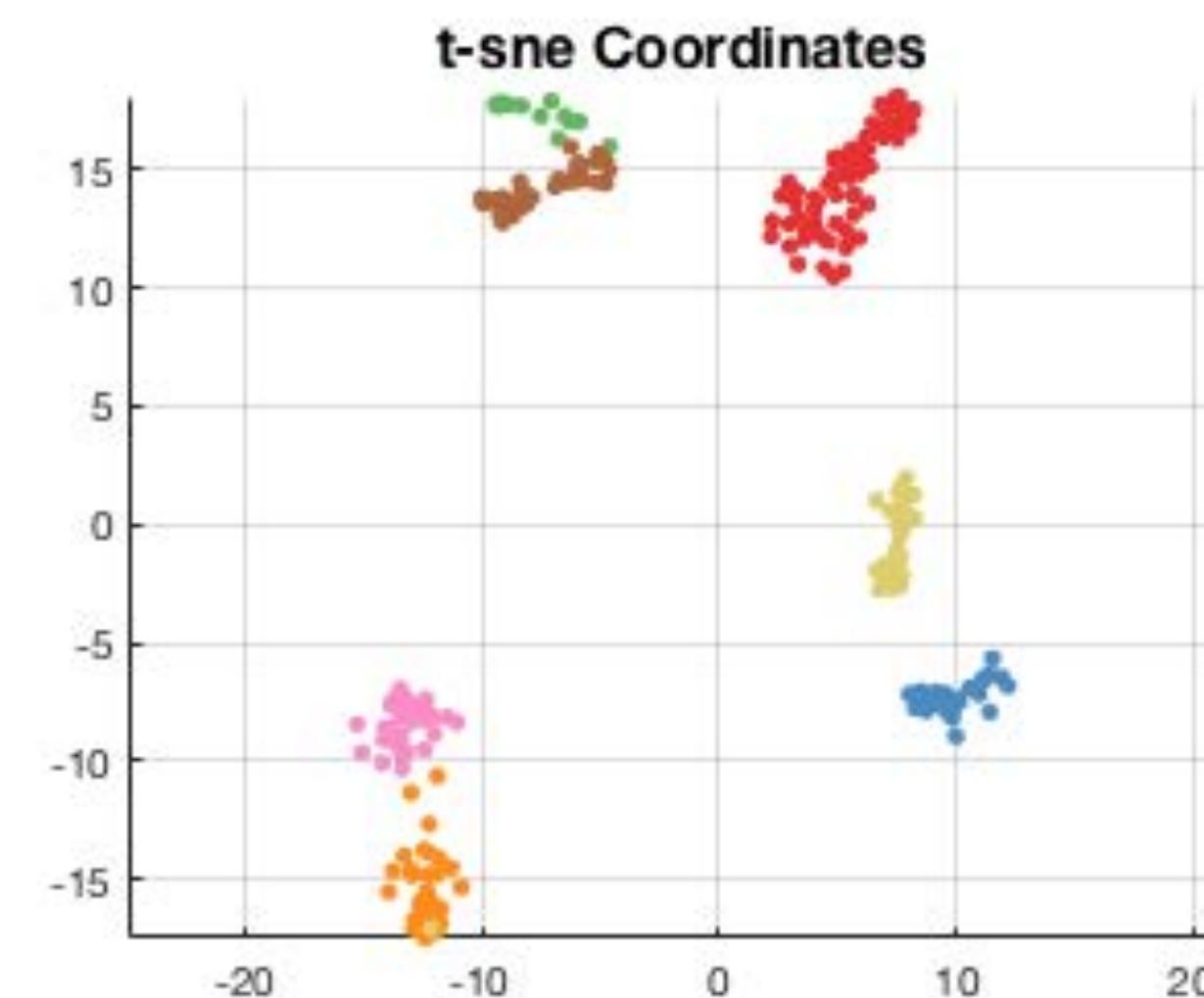
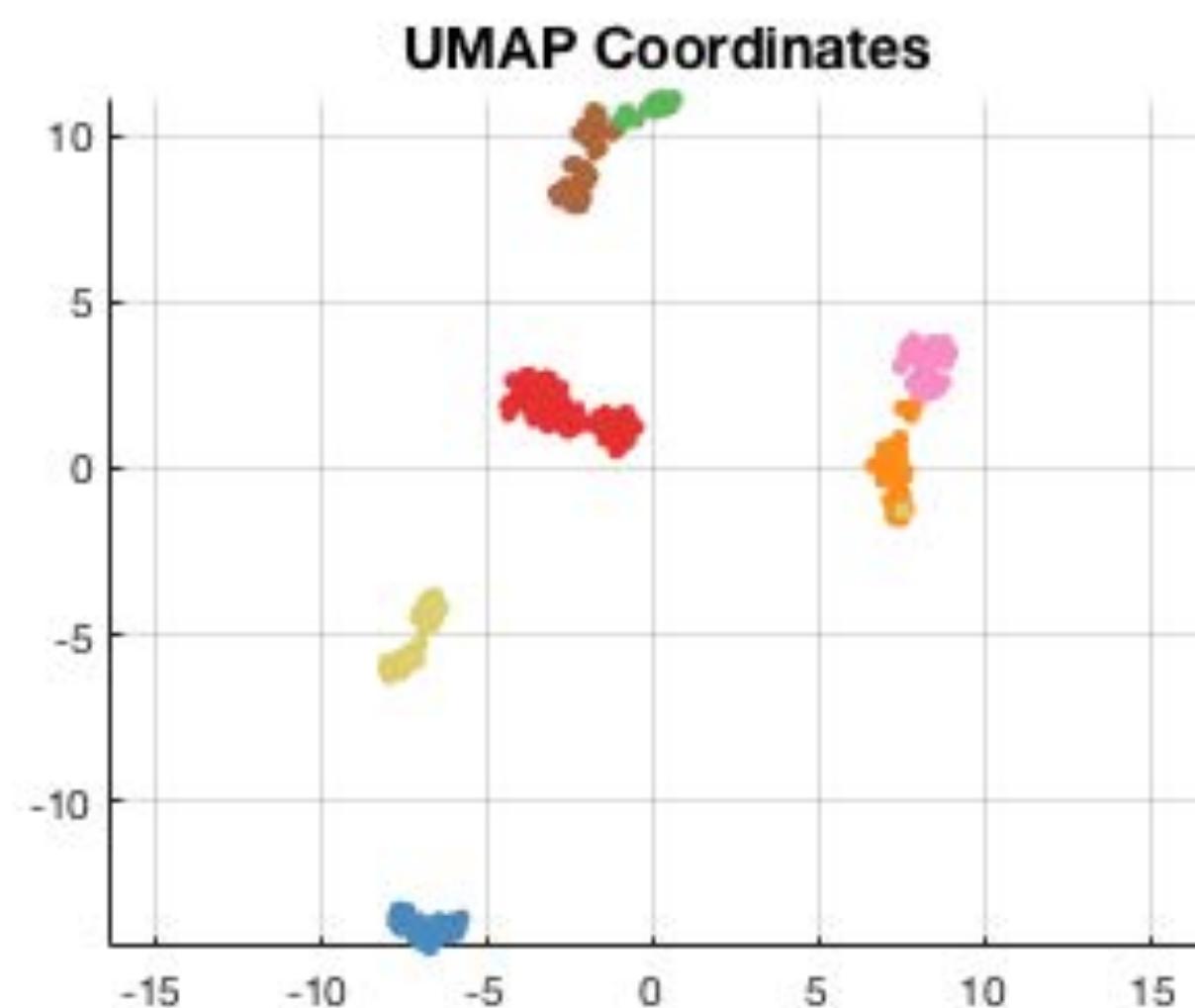
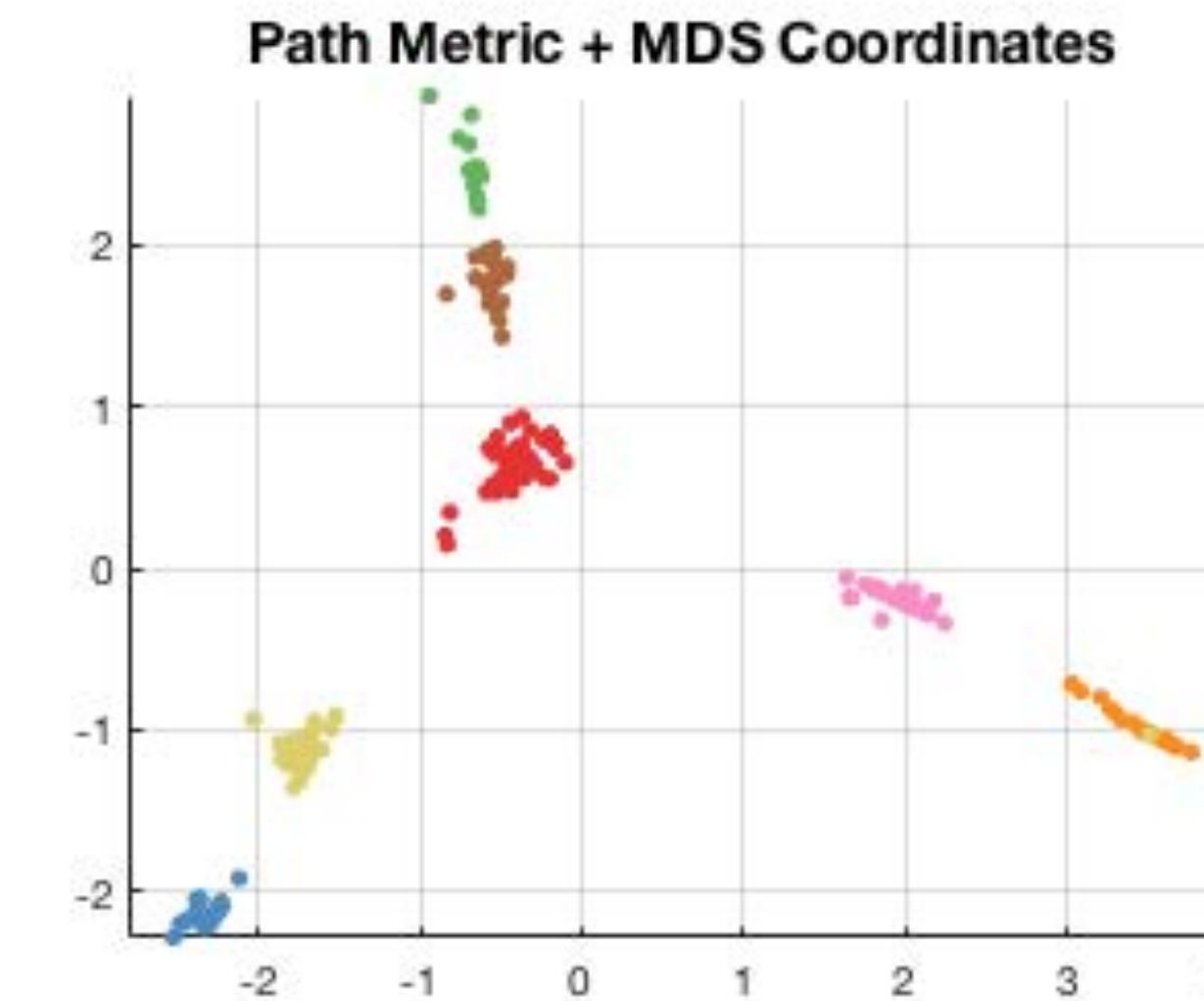
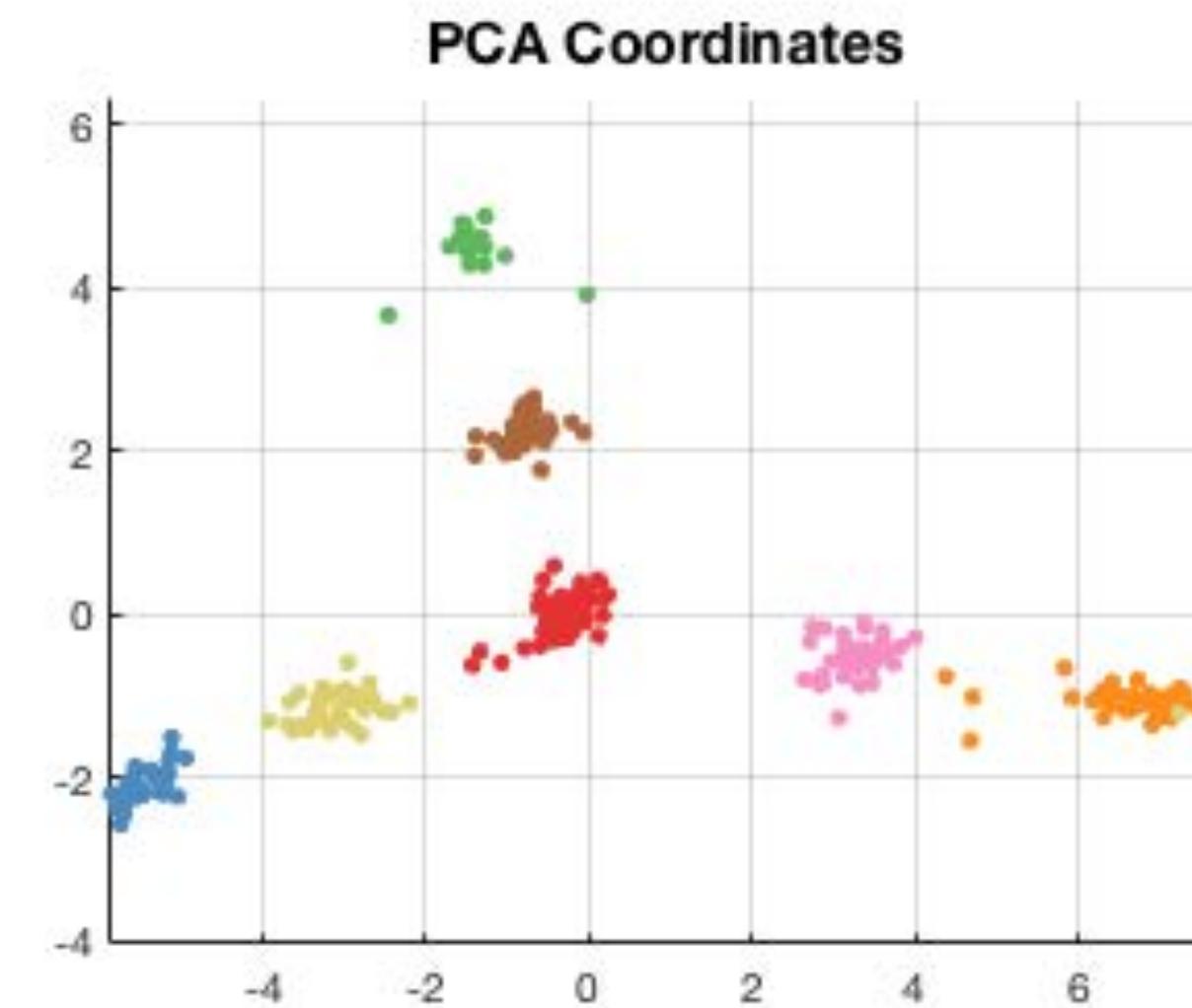
Before that: Postdoc, Michigan State University

PhD in Mathematics, Duke University

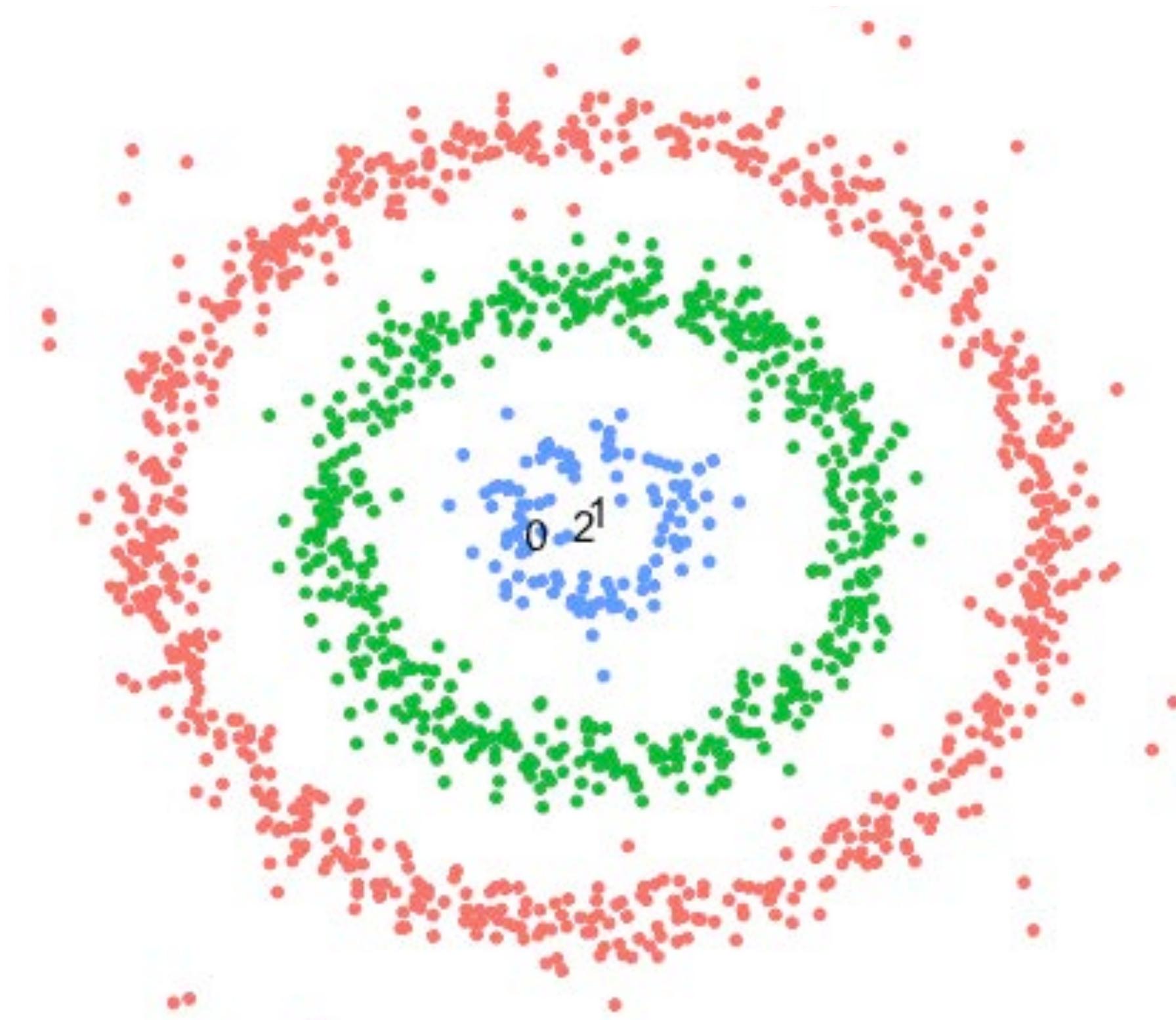
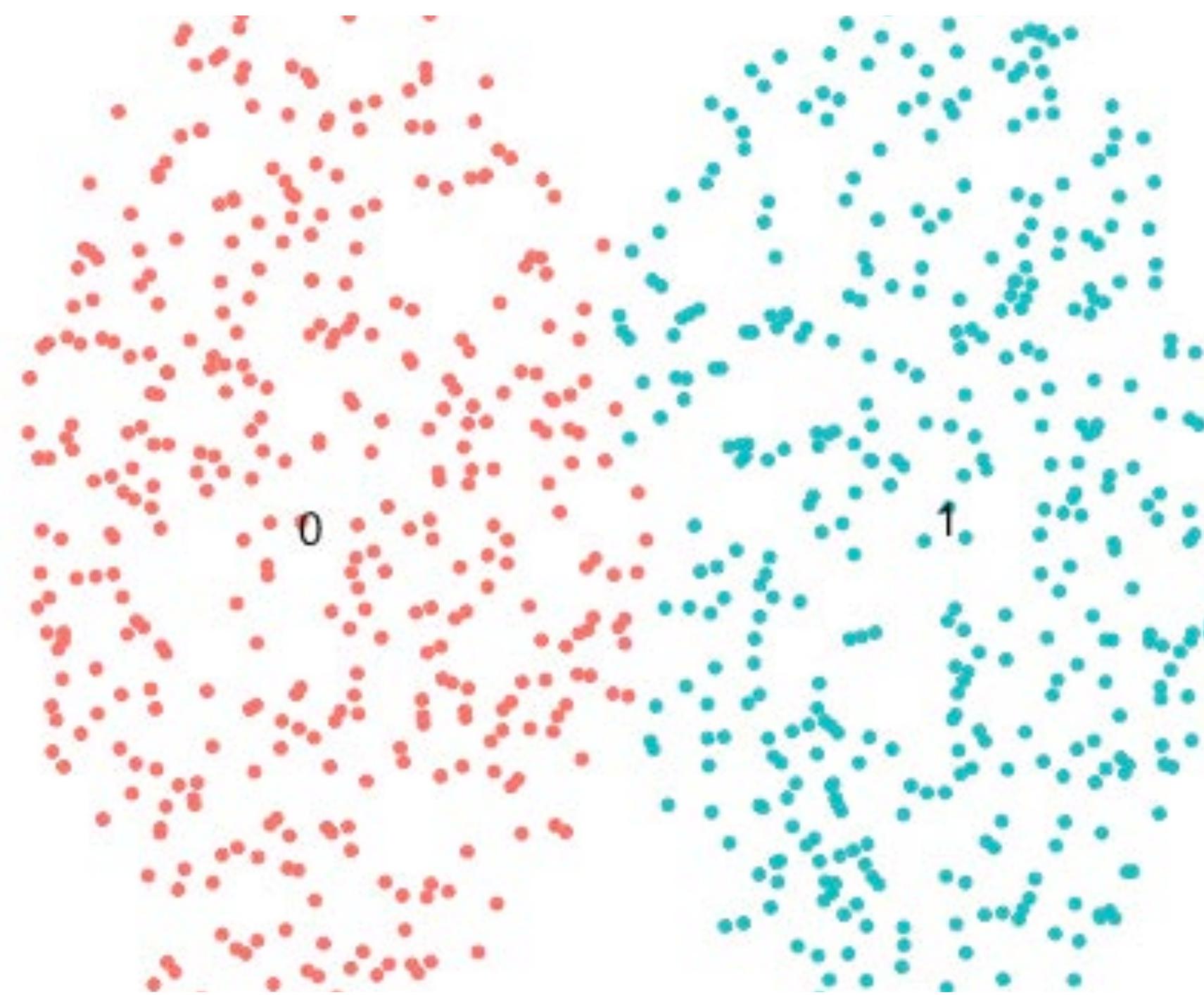


<https://www.anna-little.com/>

Data Analysis, Dimension Reduction & Estimation

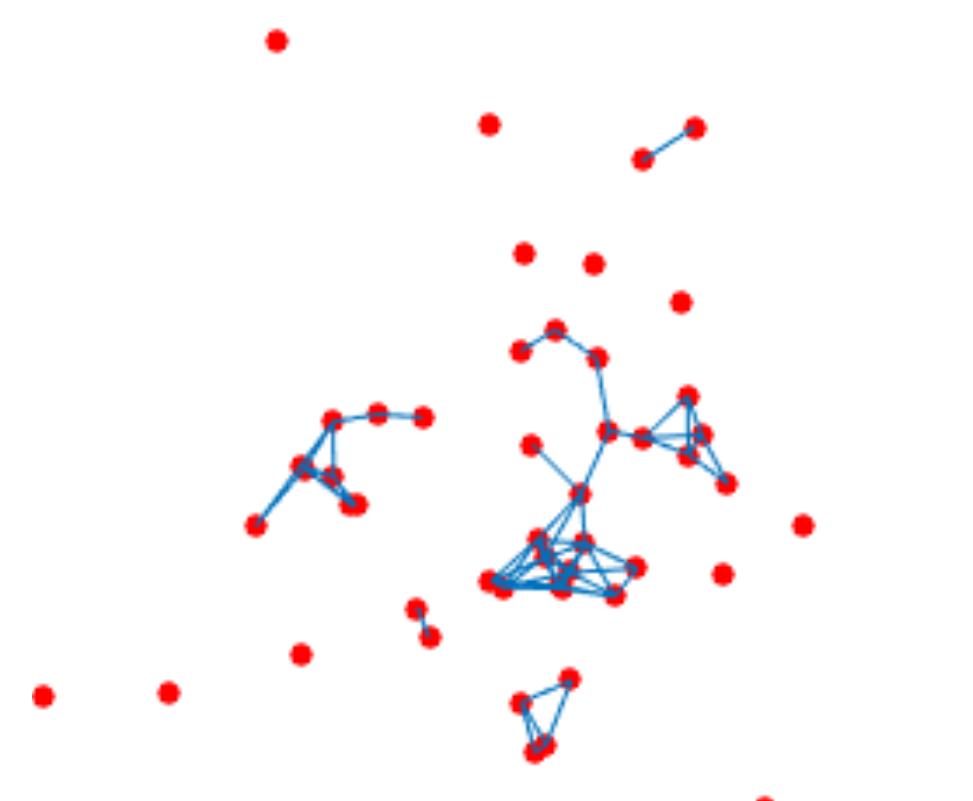


Clustering and Notions of Similarity

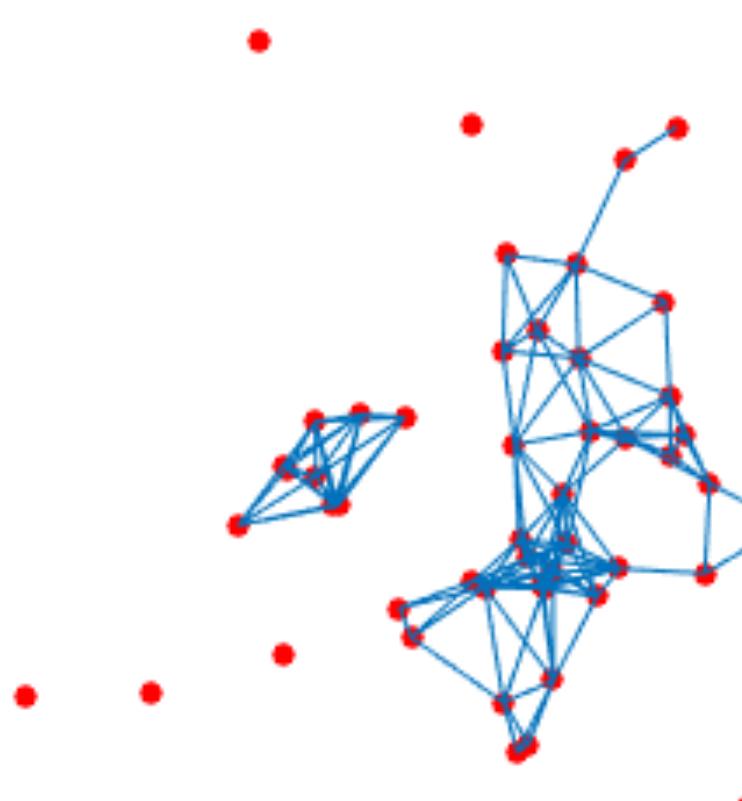


Graph-based Methods, Fast Computation

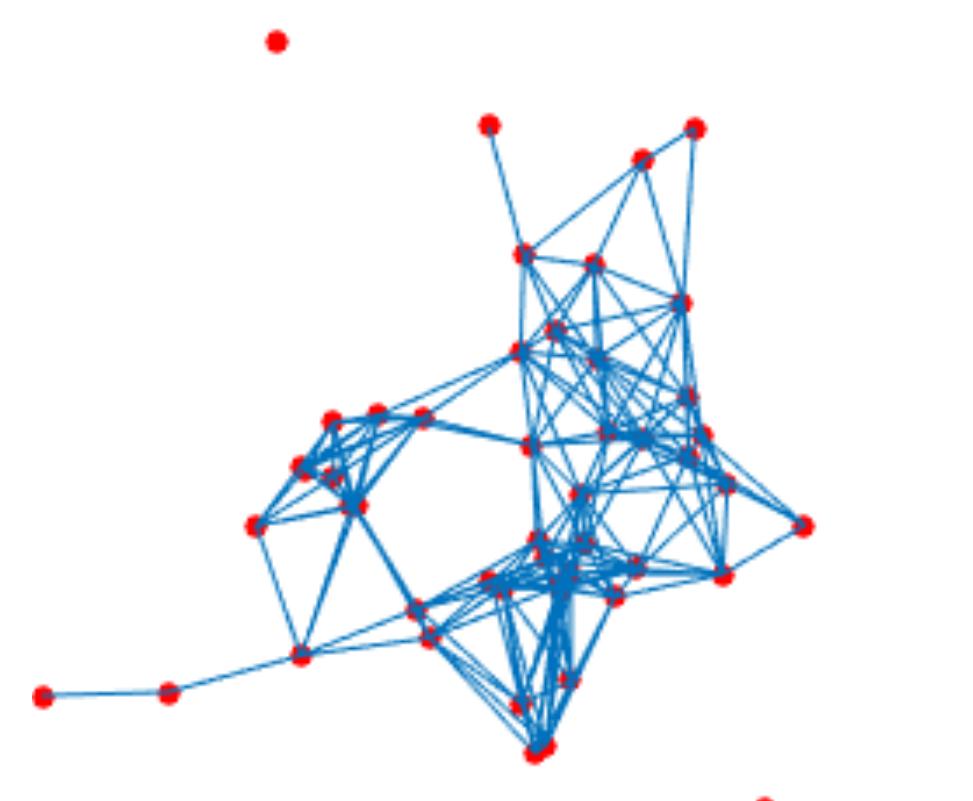
G_1 : All Edges < 0.36



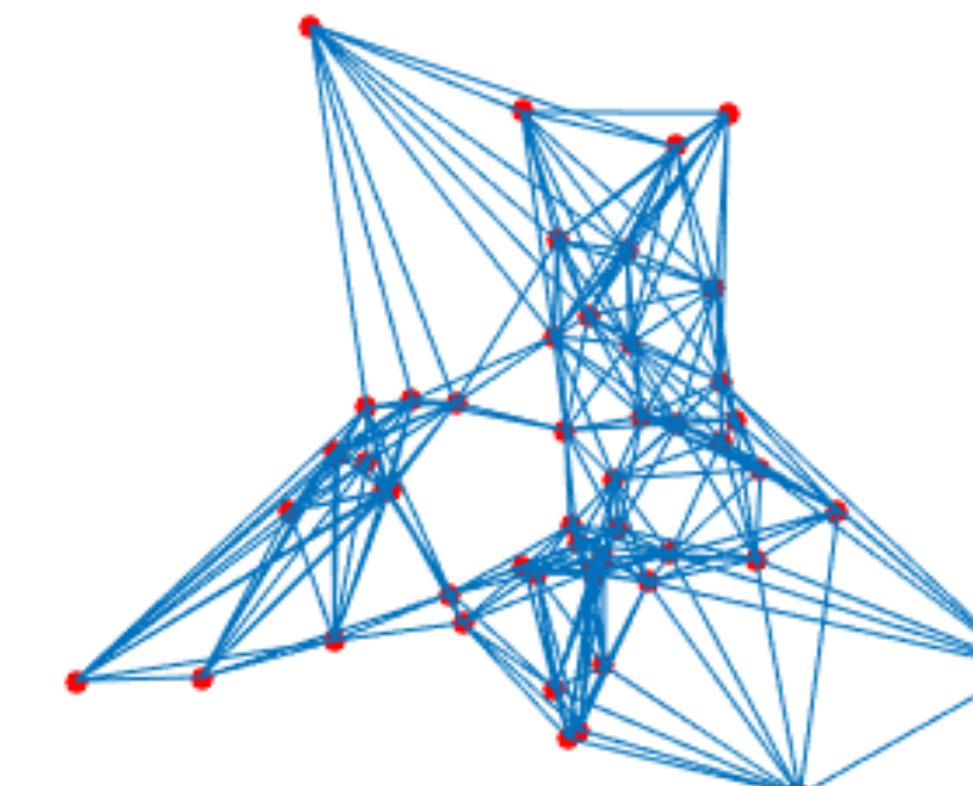
G_2 : All Edges < 0.56



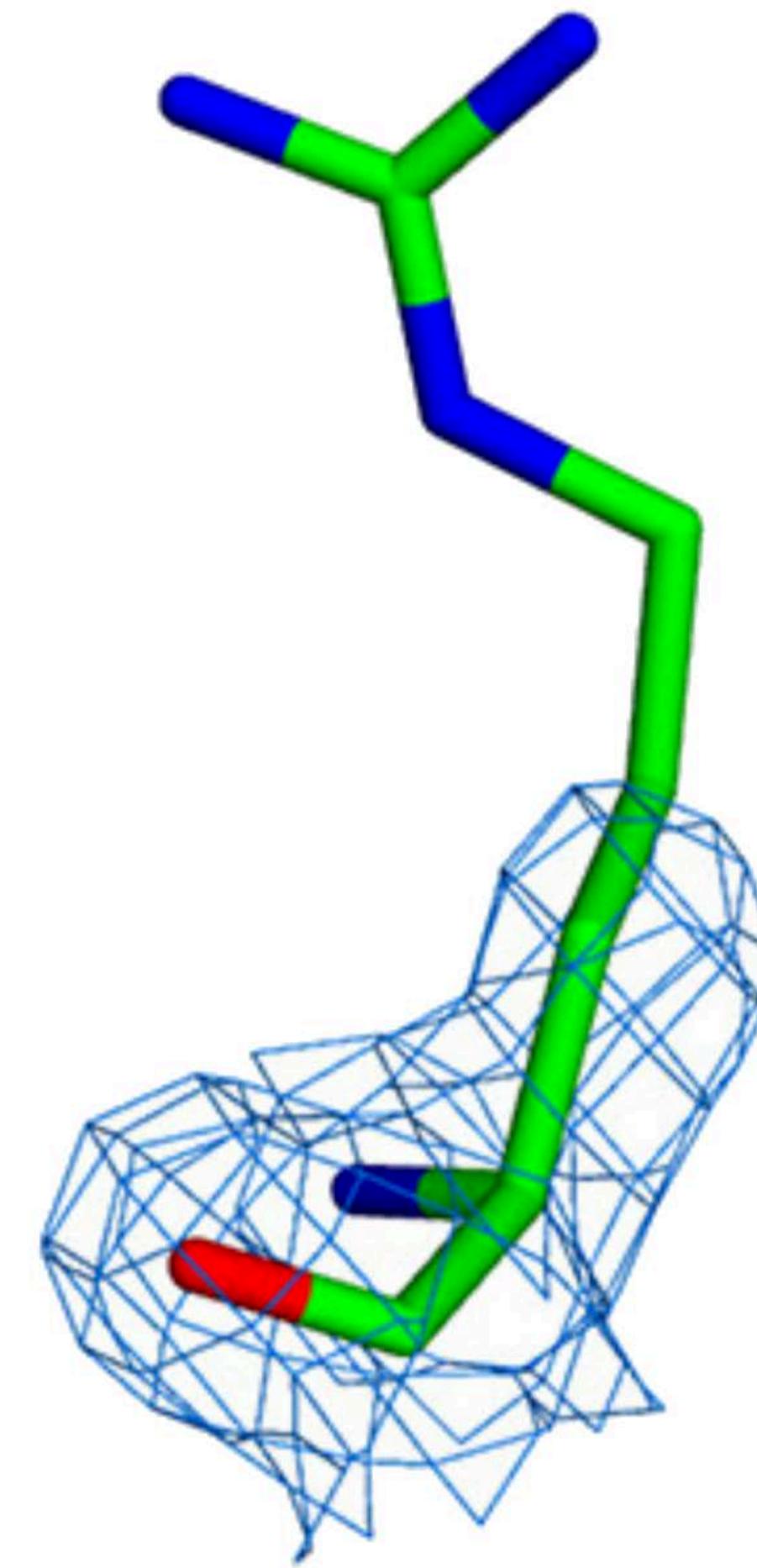
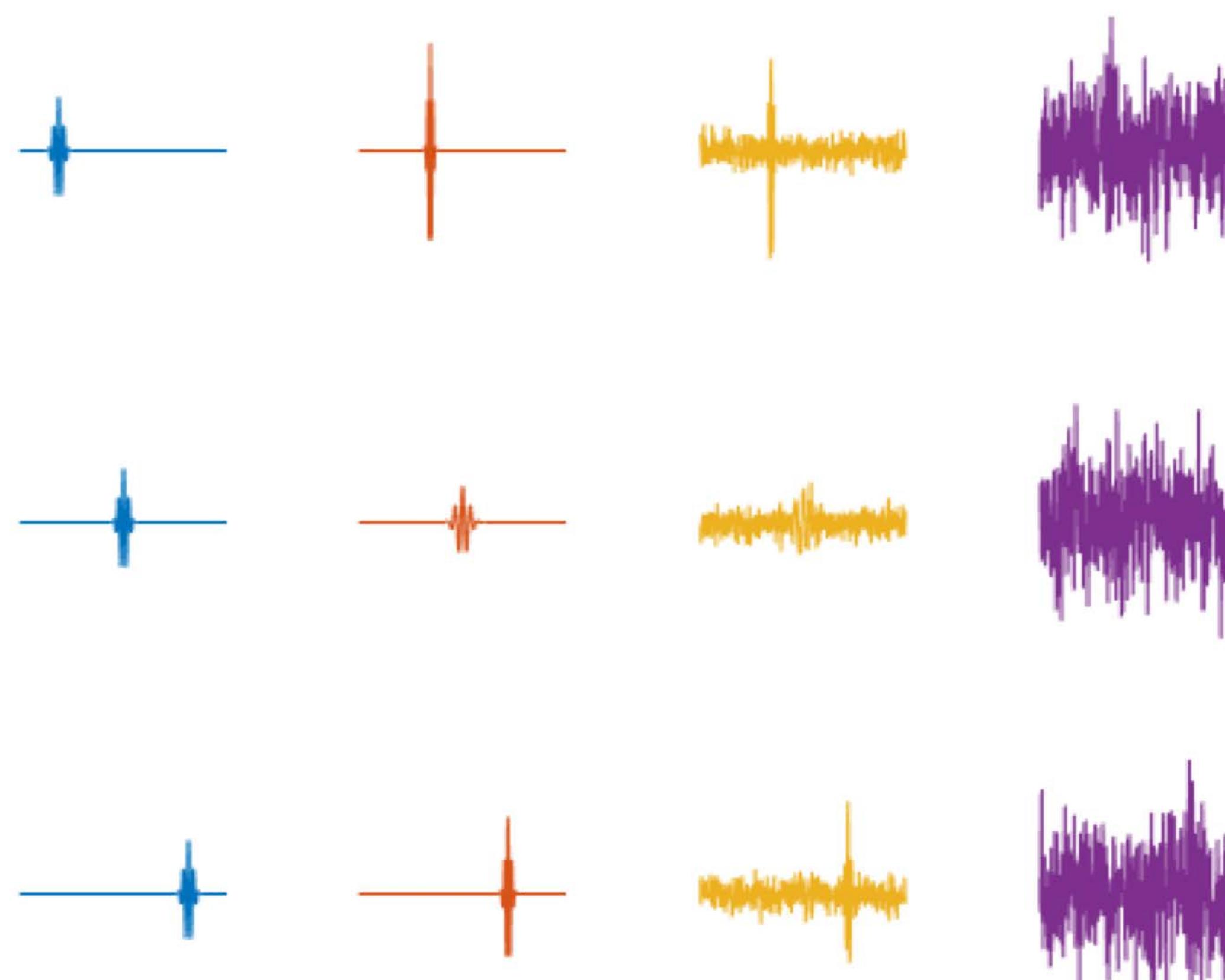
G_3 : All Edges < 0.8



G_4 : All Edges



Multi-reference Alignment, Signal Processing



Bei Wang Phillips

Assistant Professor, Computer Science

Adjunct Assistant Professor, Mathematics

Faculty, Scientific Computing and
Imaging (SCI) Institute

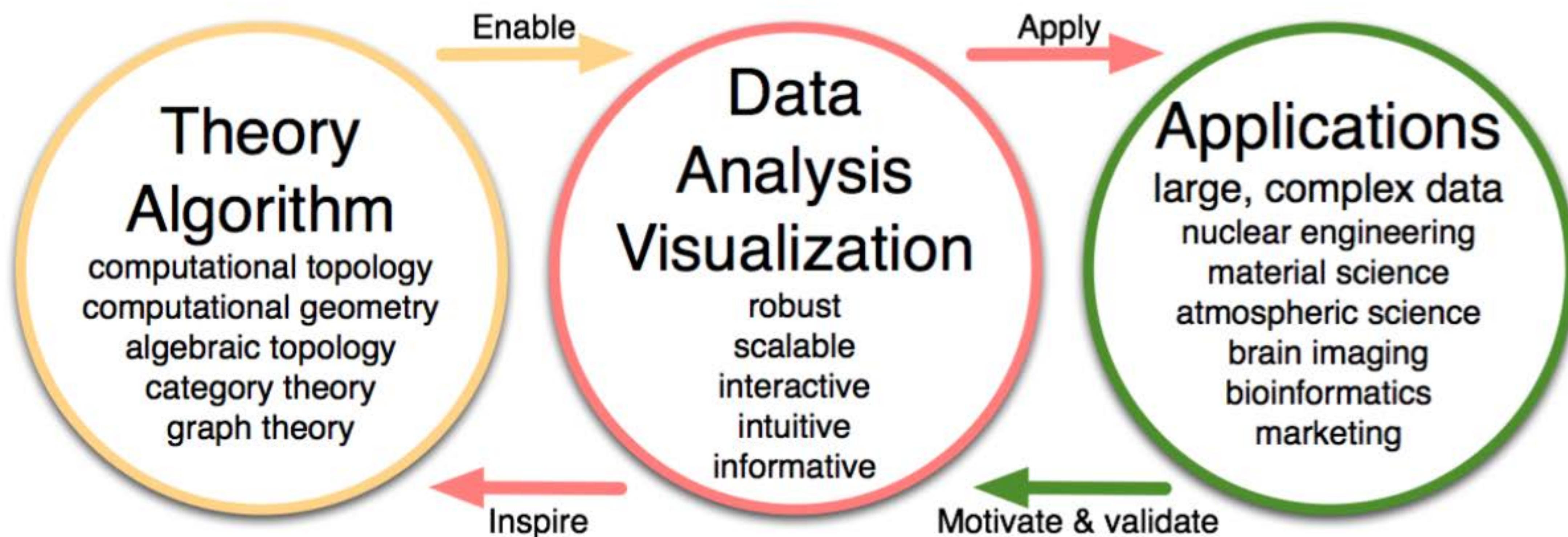
PhD in CS, Duke University



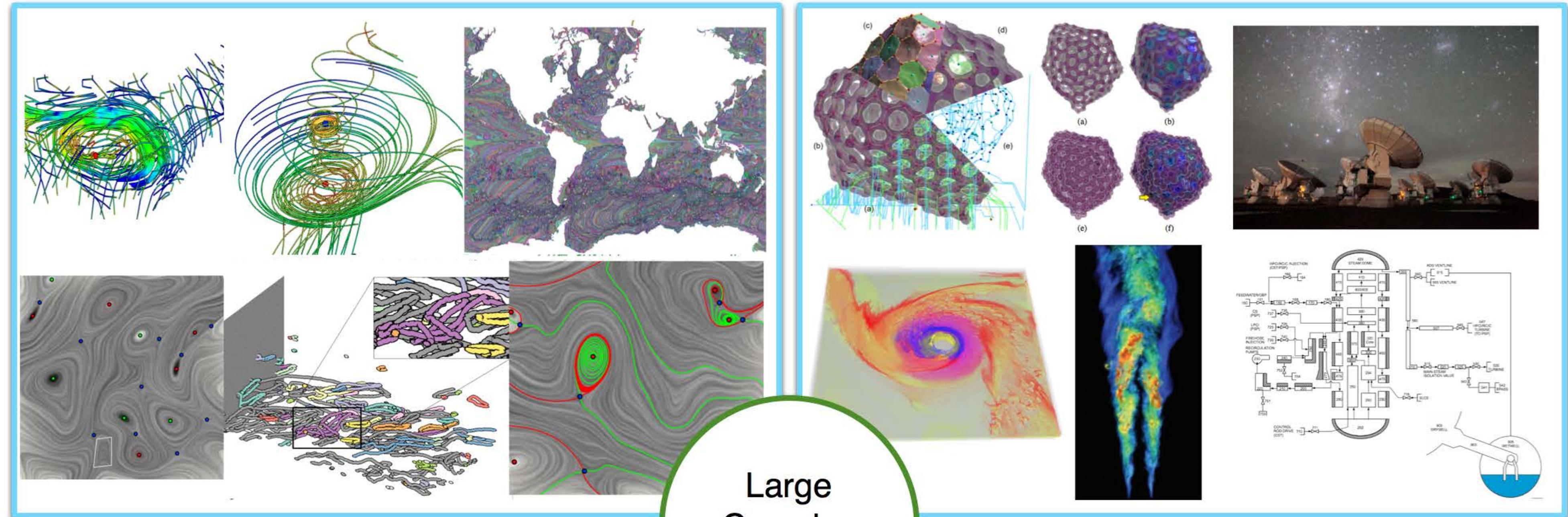
<http://www.sci.utah.edu/~beiwang/>

Topological data analysis, data visualization, applied and computational topology/geometry, machine learning, data mining; previously, bioinformatics, robotics

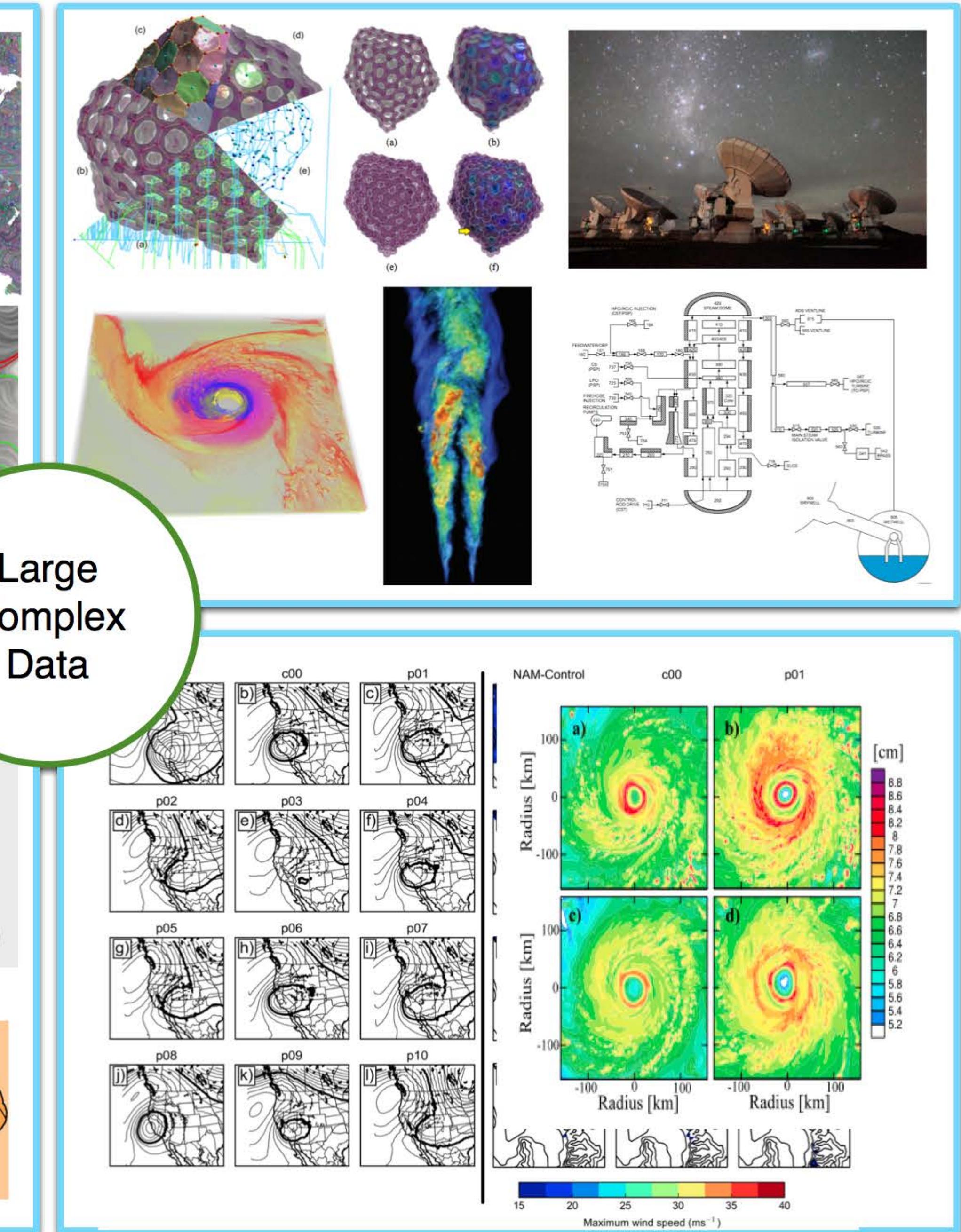
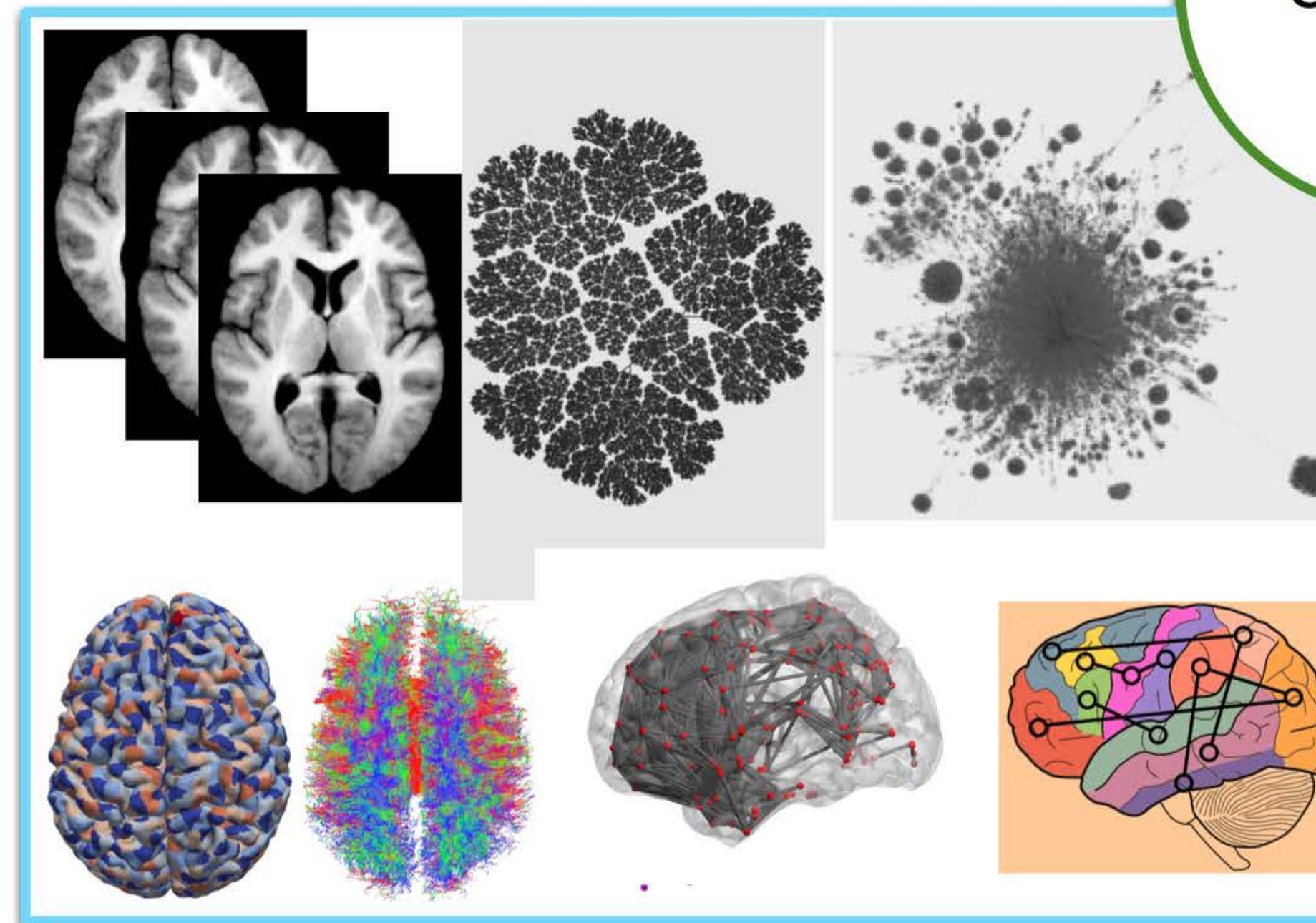
Innovation and progress in data visualization require an end-to-end view

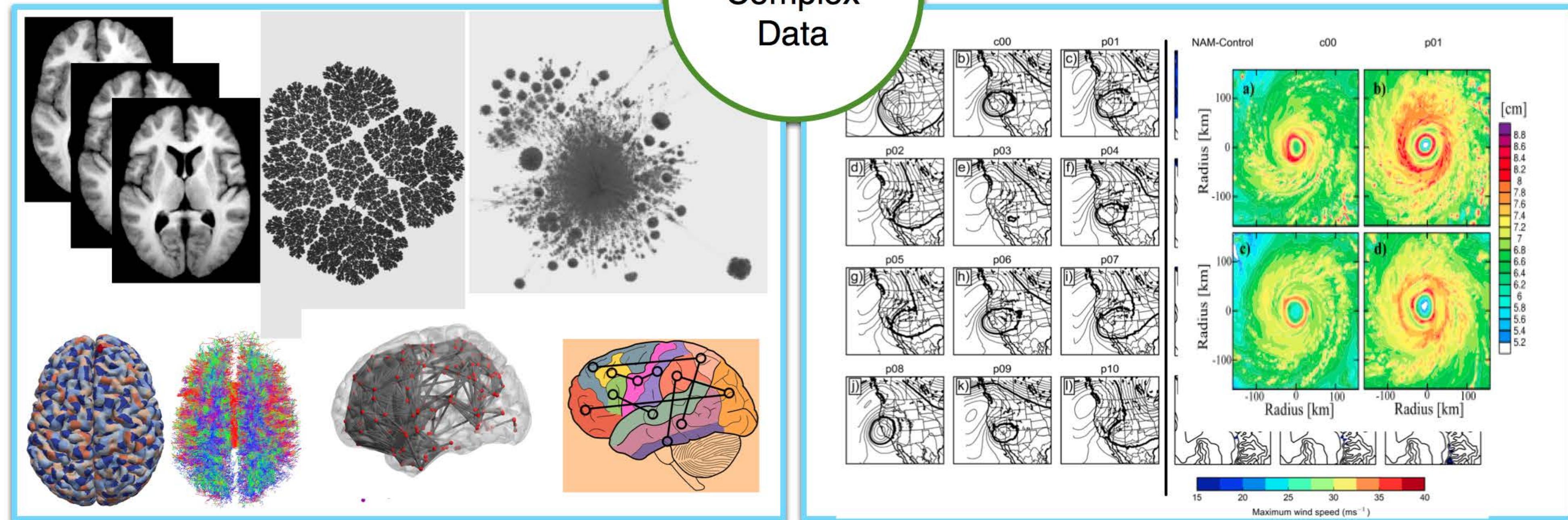
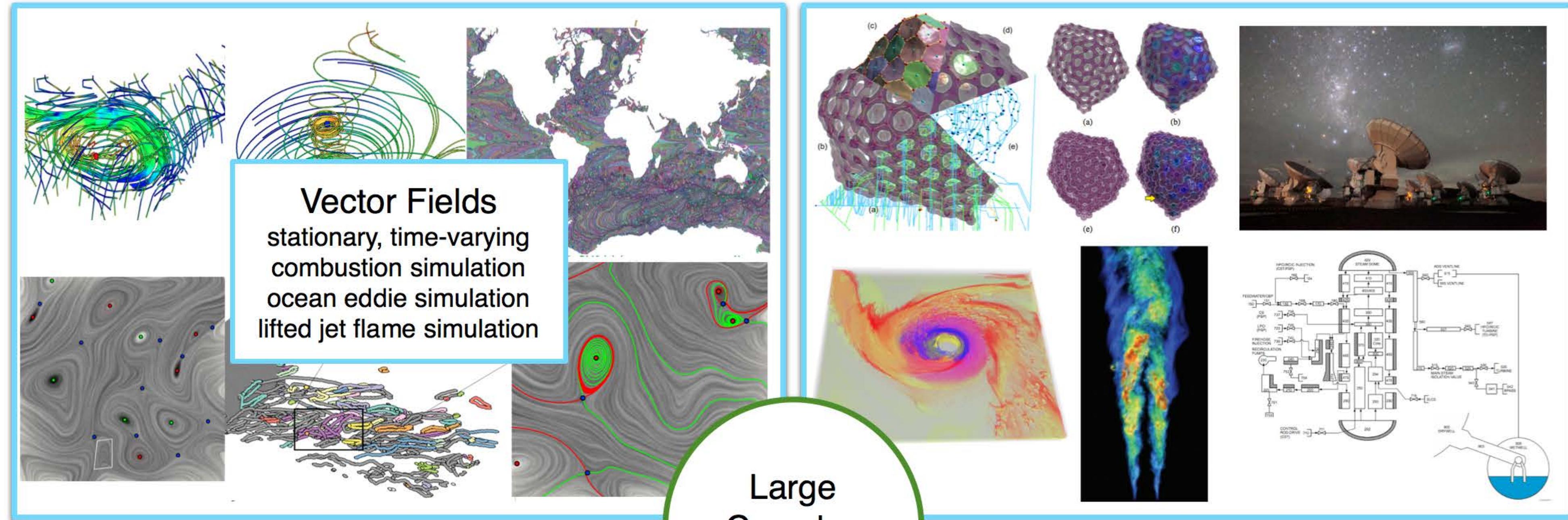


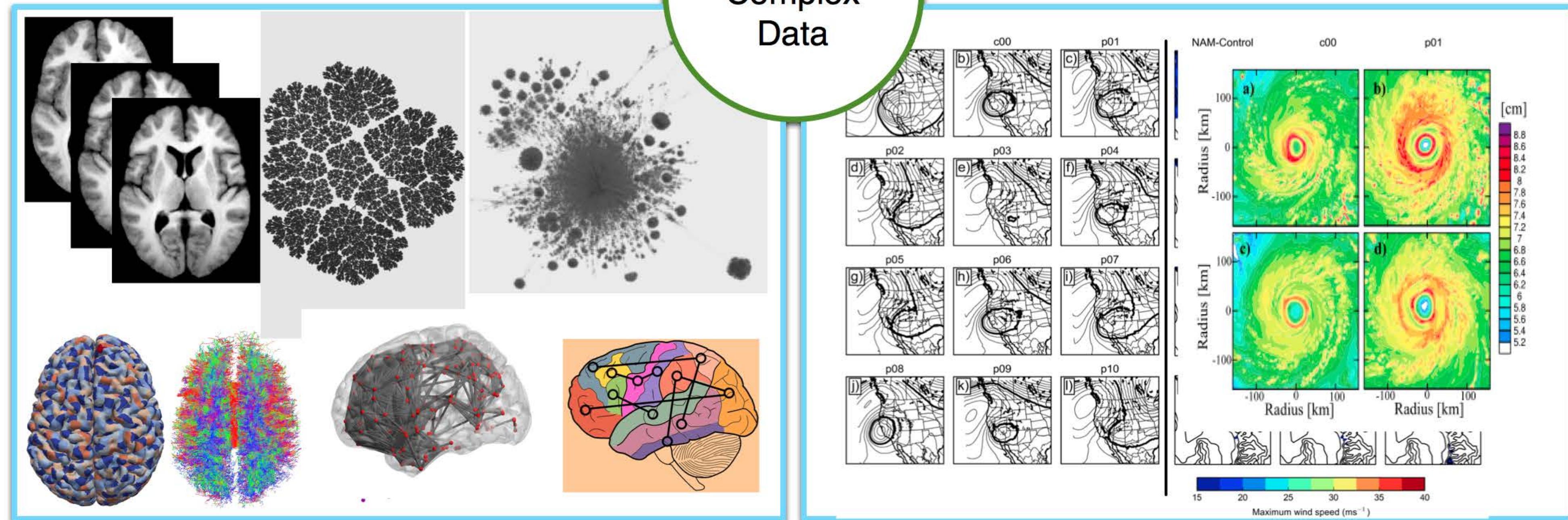
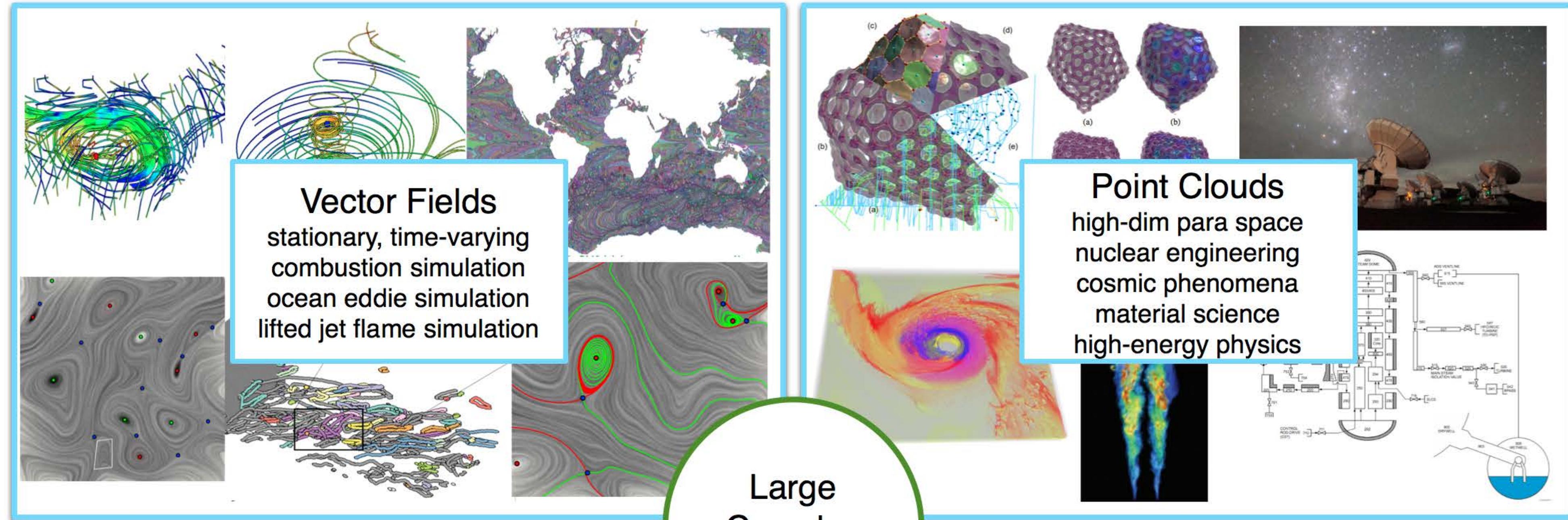
Combine topological, geometric, data analysis and visualization techniques
to **study** large and complex data that require rich structural descriptions

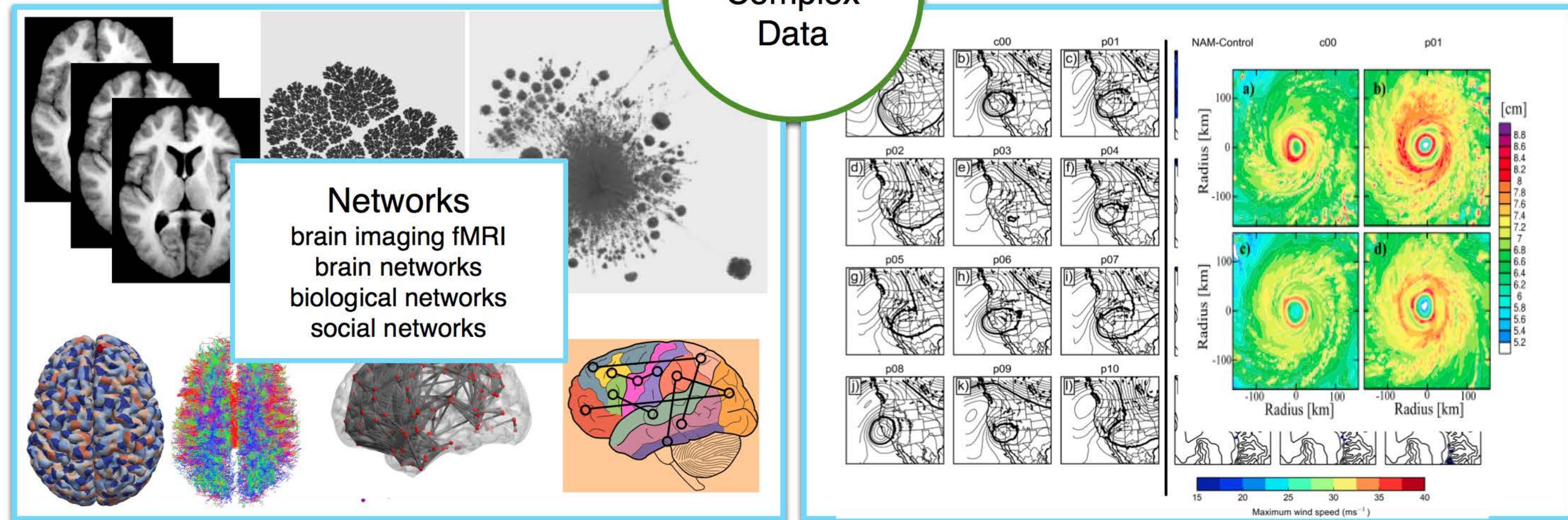
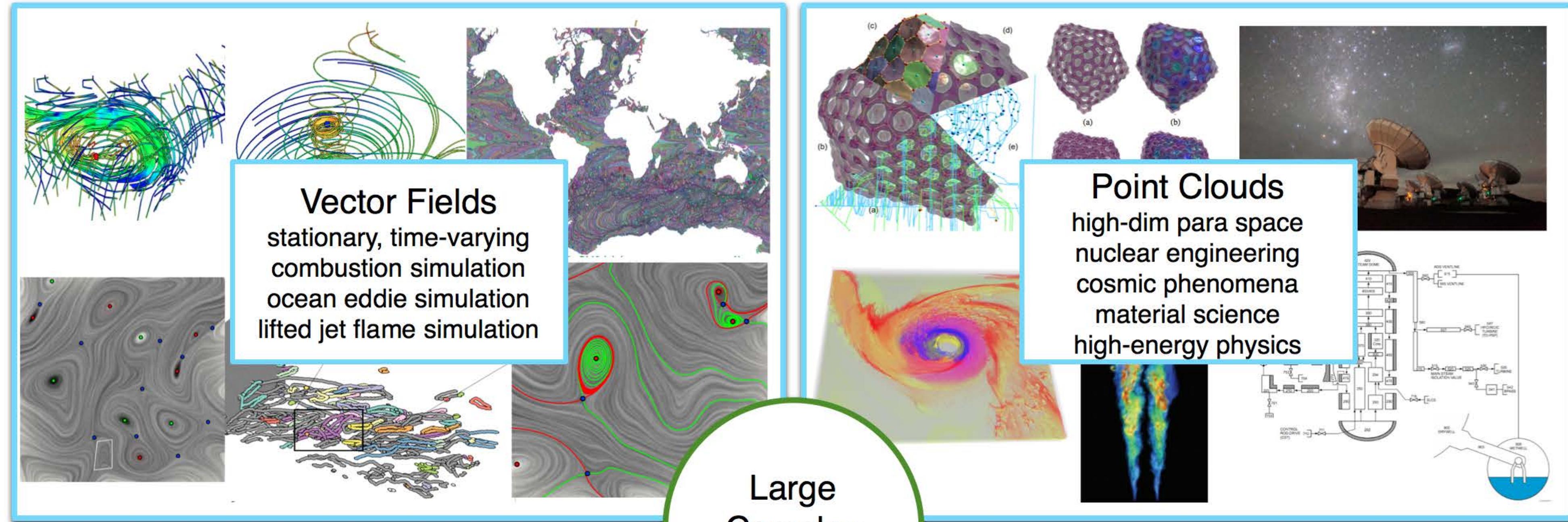


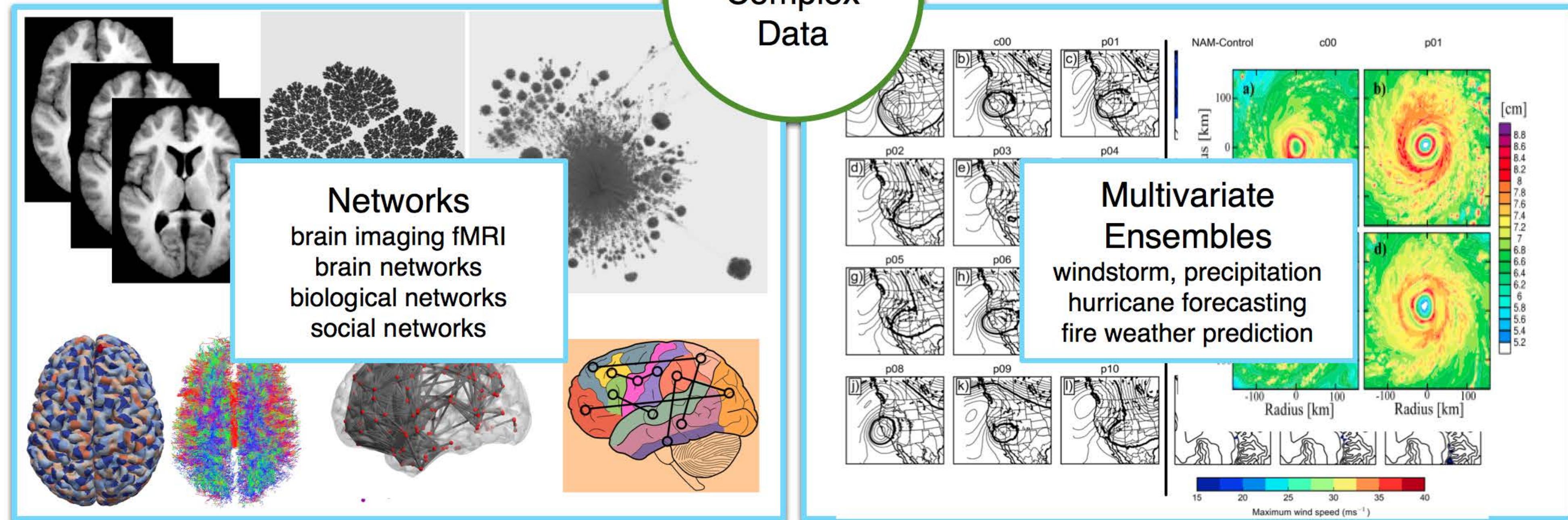
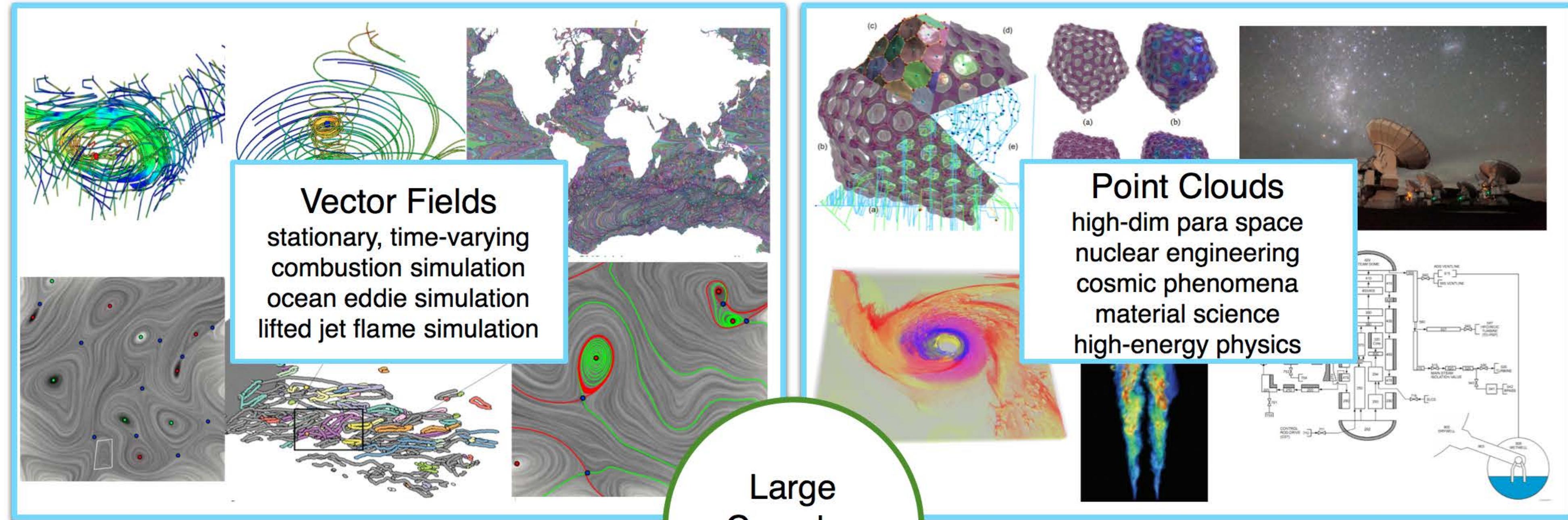
Large
Complex
Data



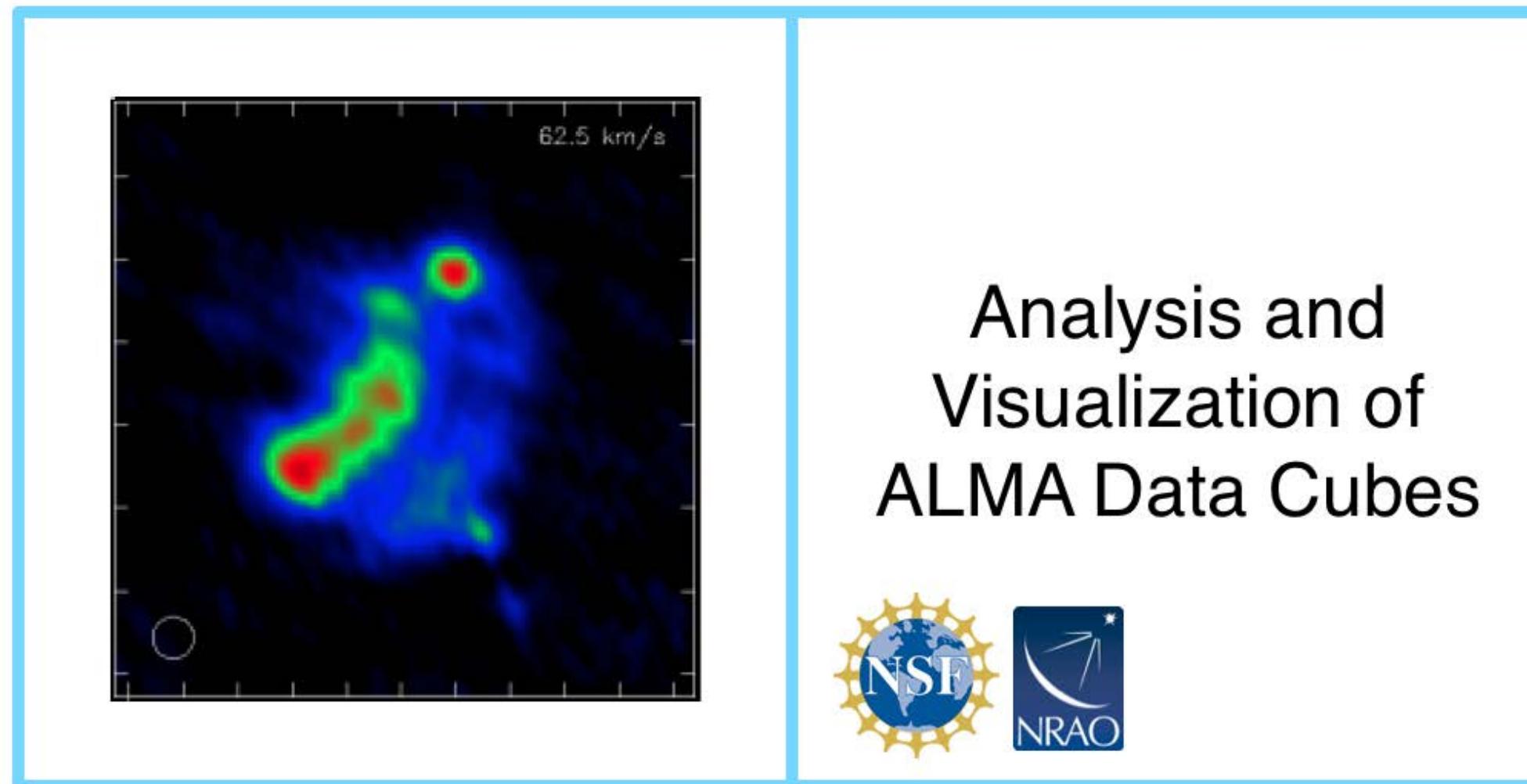




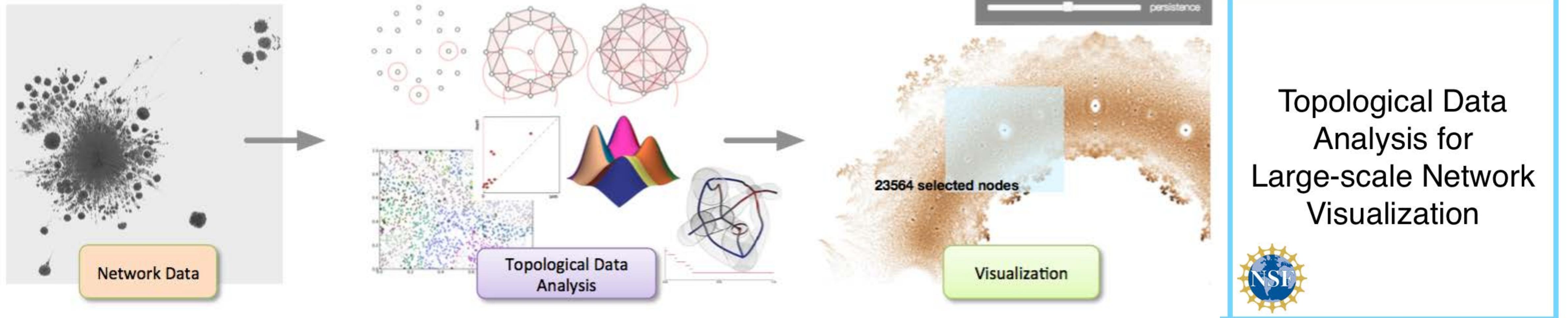




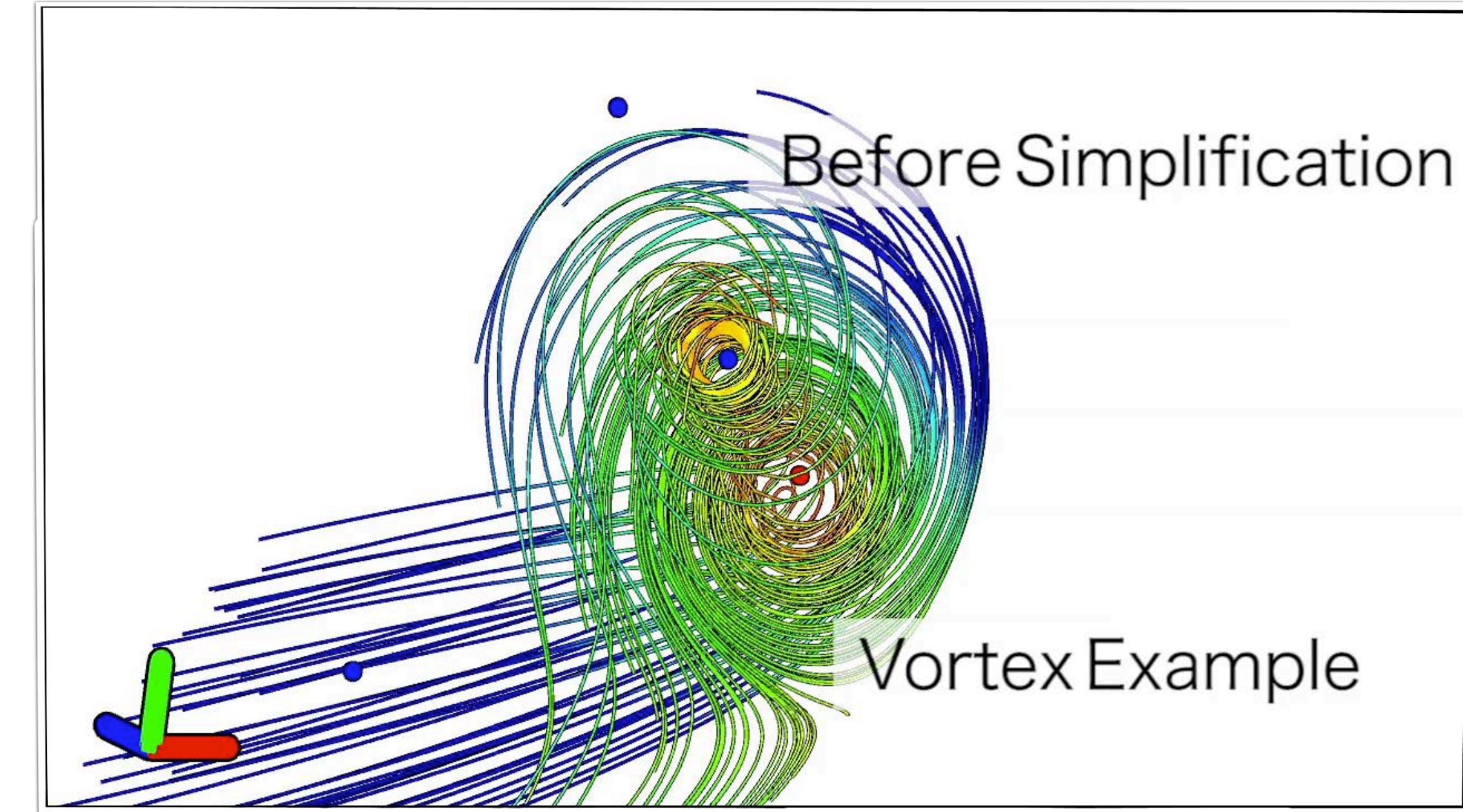
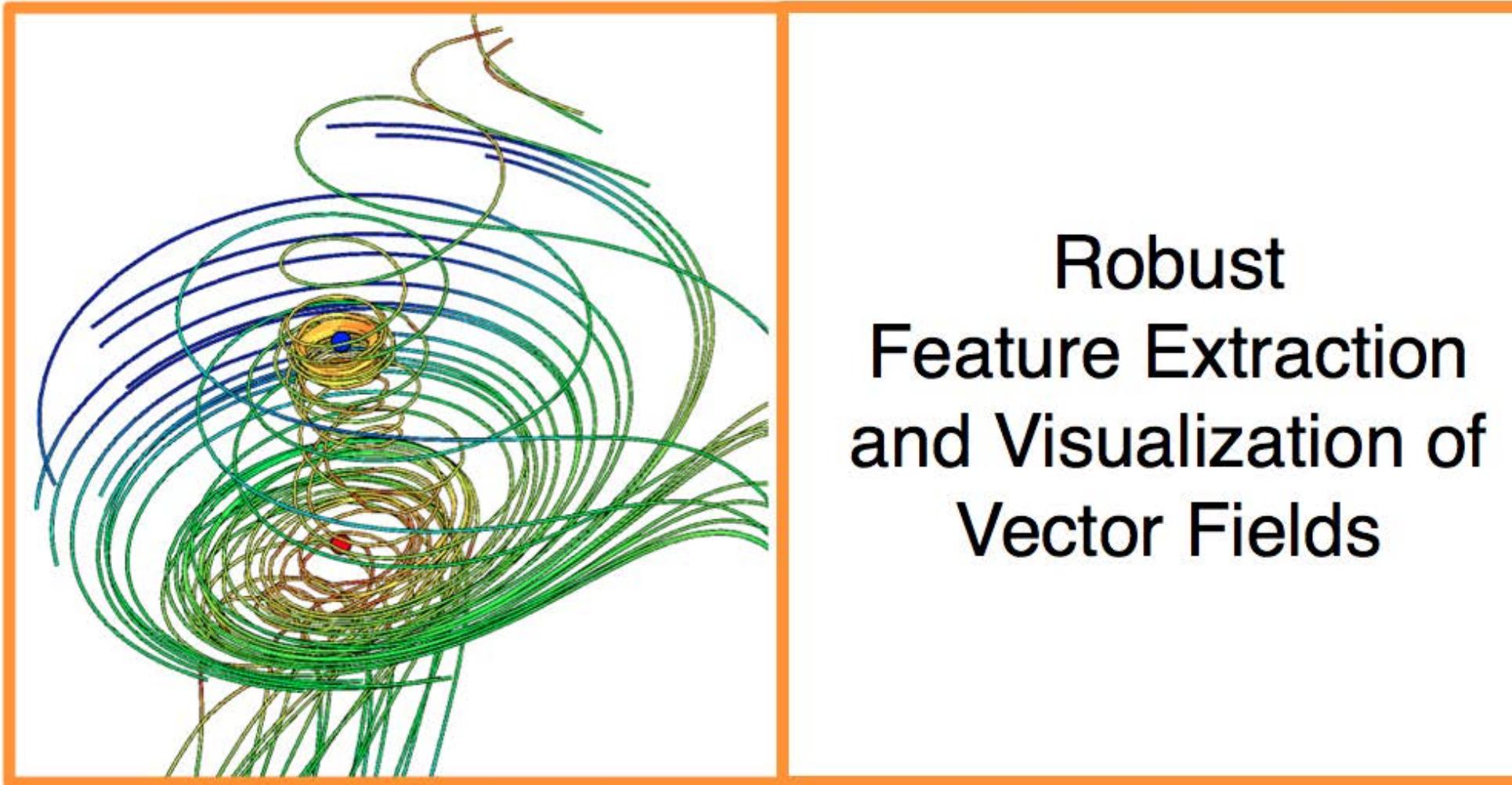
Astronomy



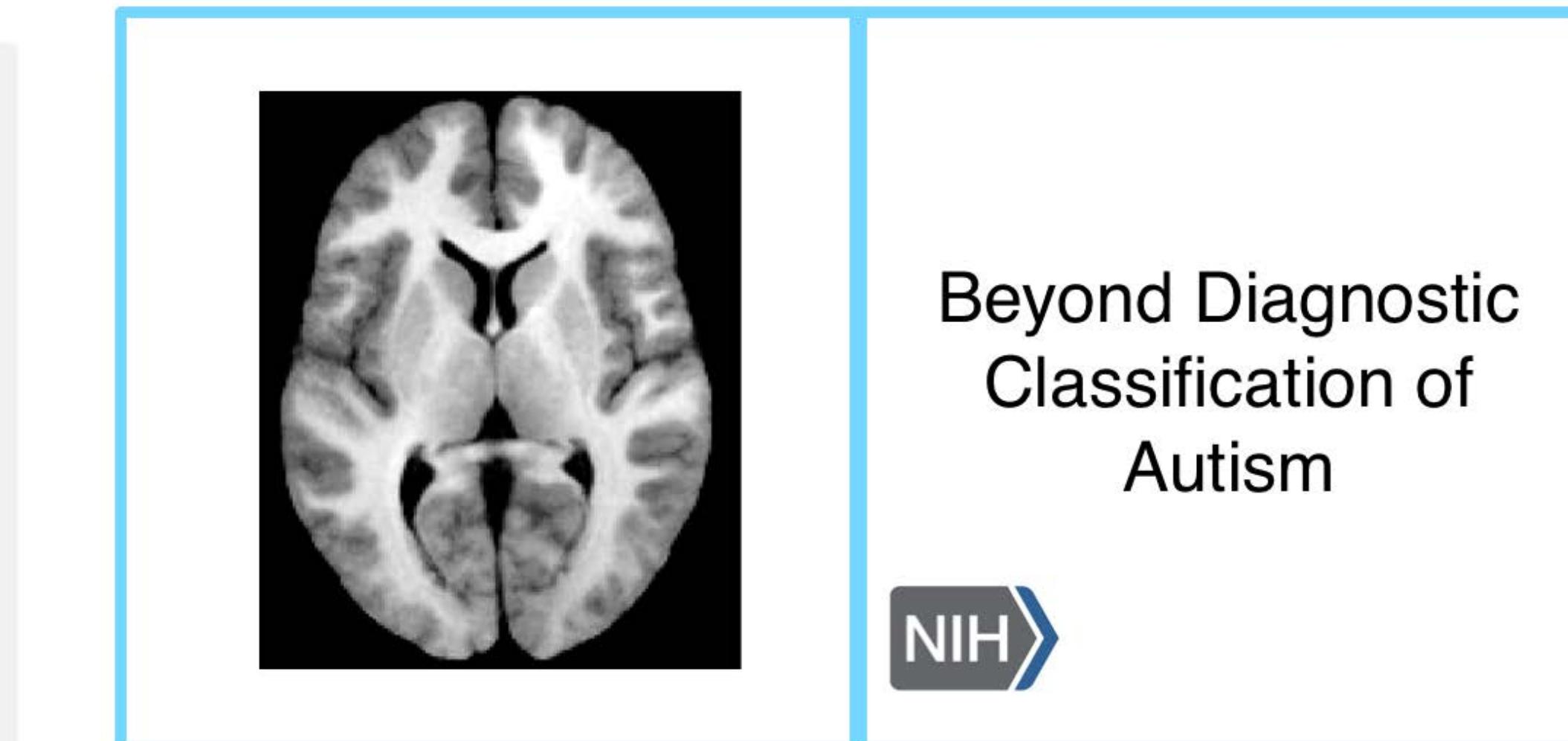
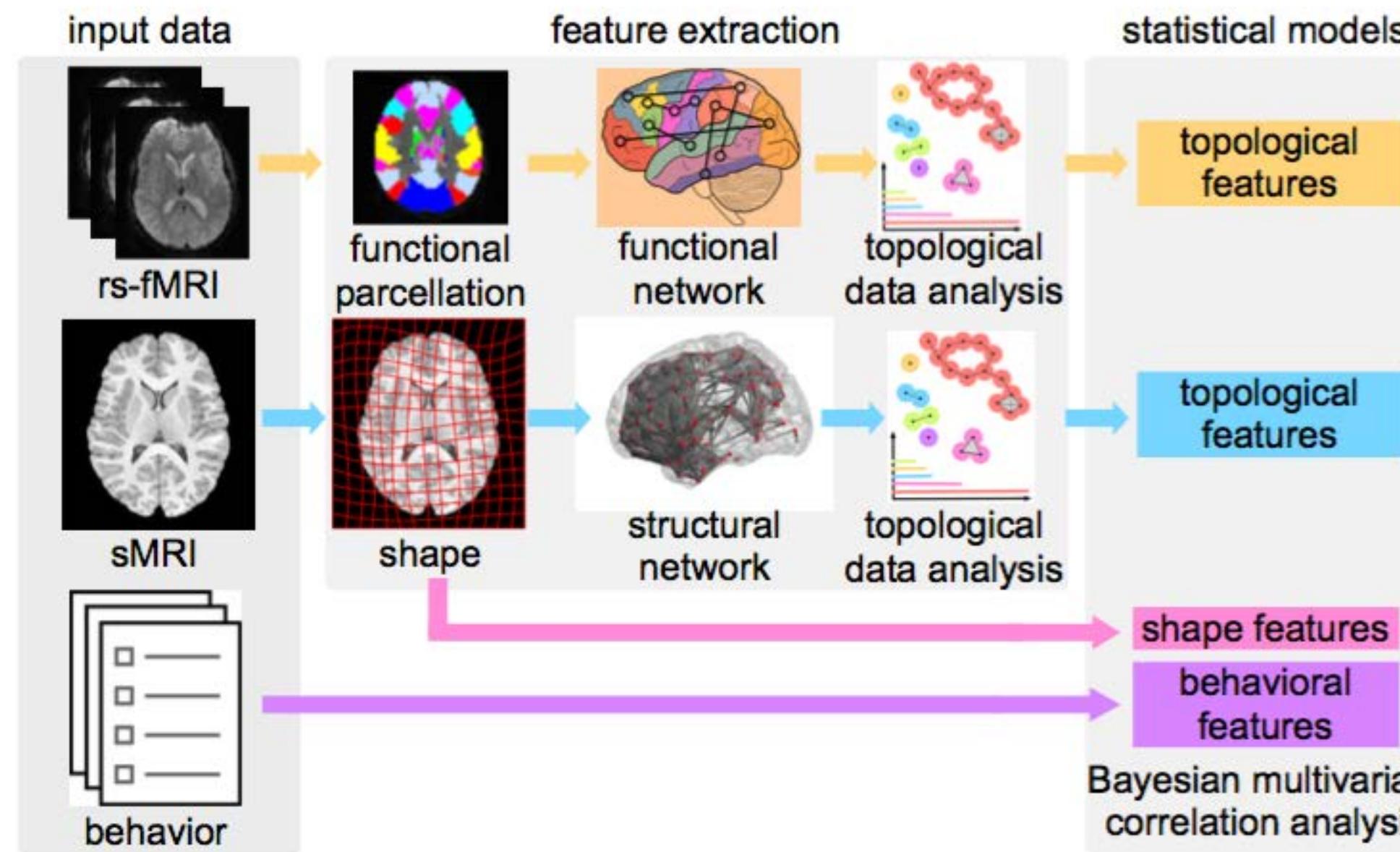
Large Graph Visualization



Flow from Scientific Simulation



Brain Initiative



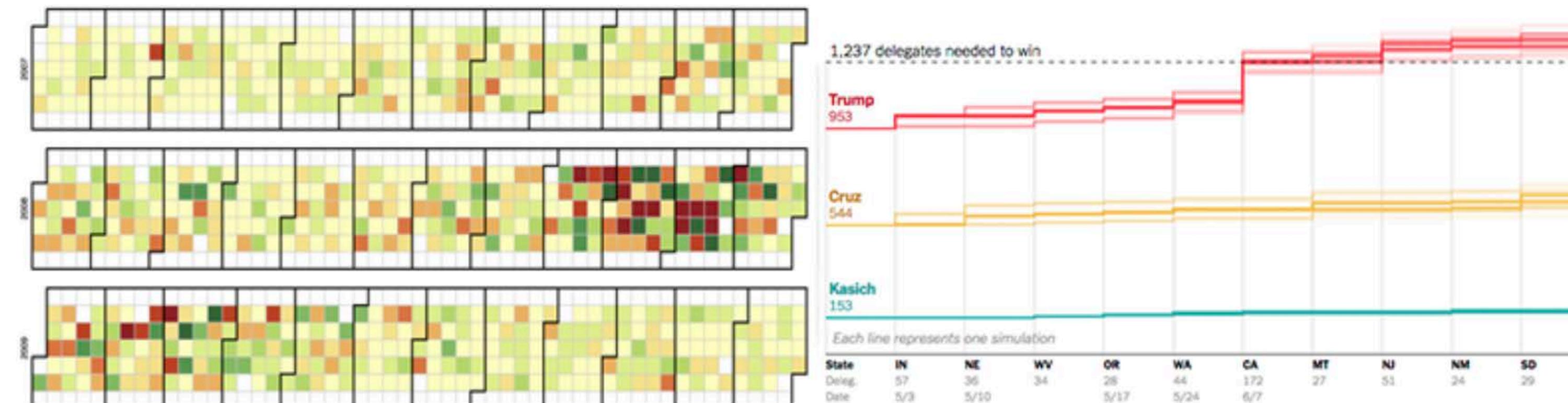
Course Structure

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.

Zoom

Streaming lectures, archived on YouTube

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

Online Lectures & 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%

Varying 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:

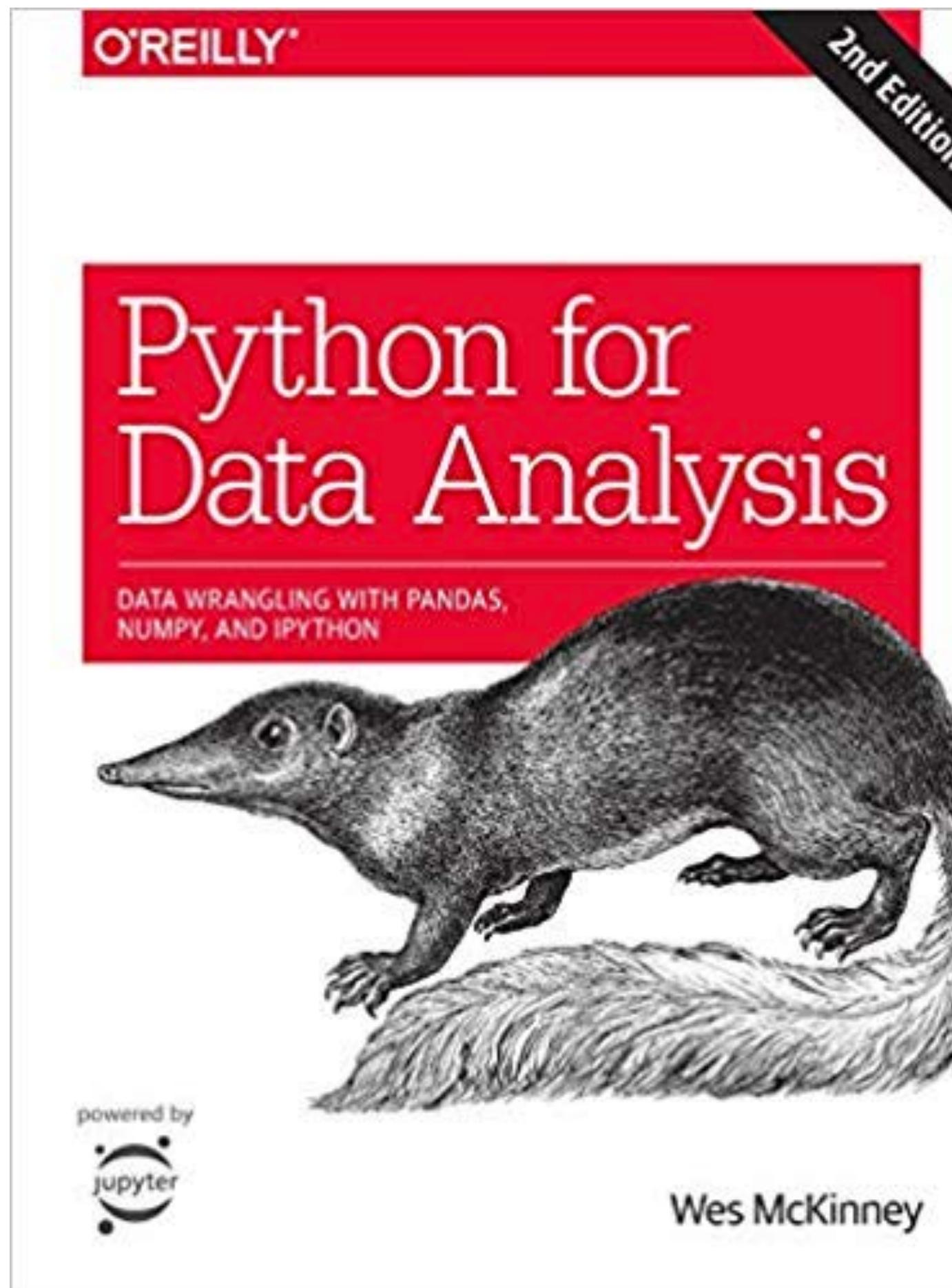
Tue / Th 3:40 - 5:00 PM MDT Hybrid (in person in L101 and on Zoom).

Lectures frequently involve computer activities.

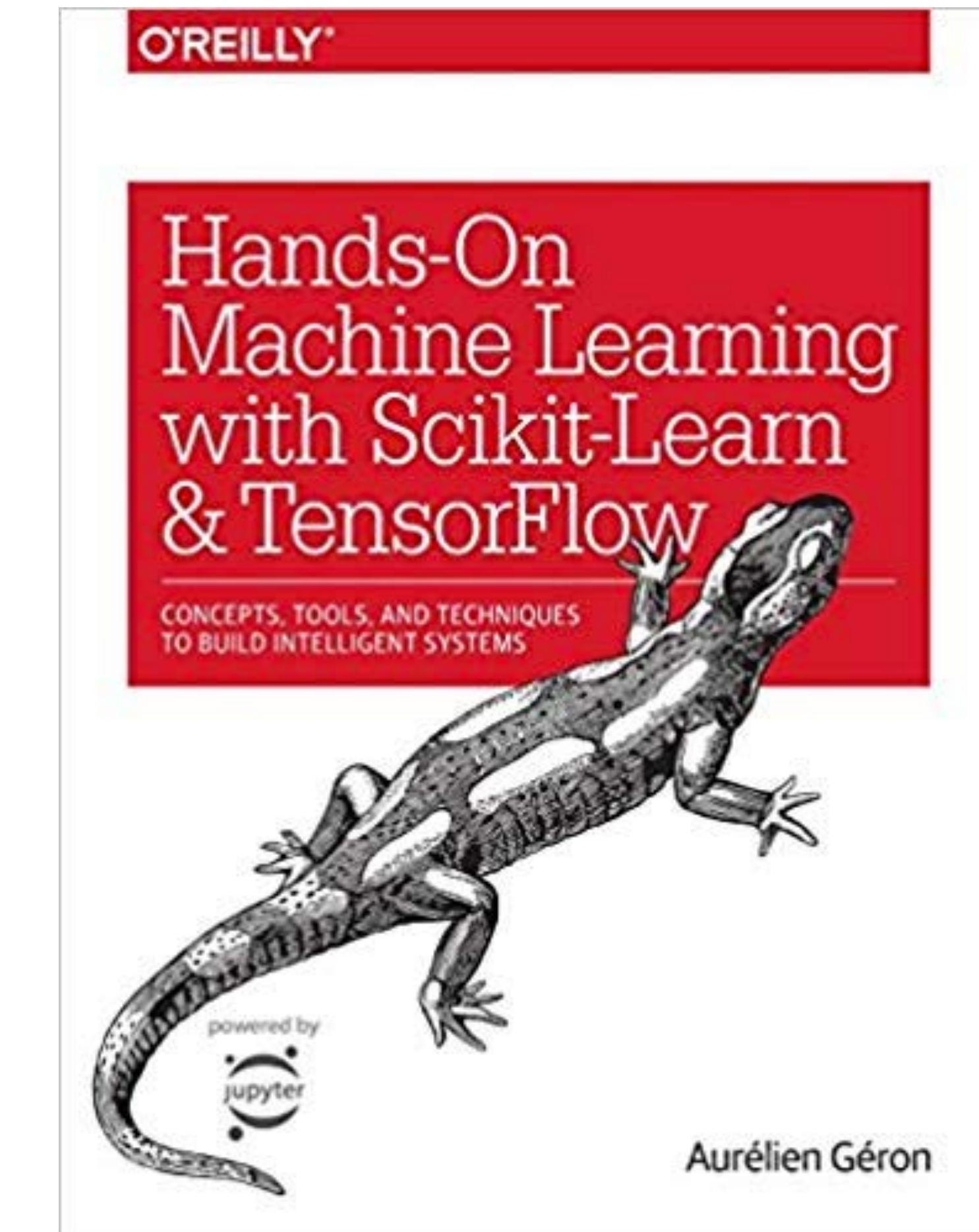
Bring your own computer!

Have Python, etc installed
(see HW0)

Books

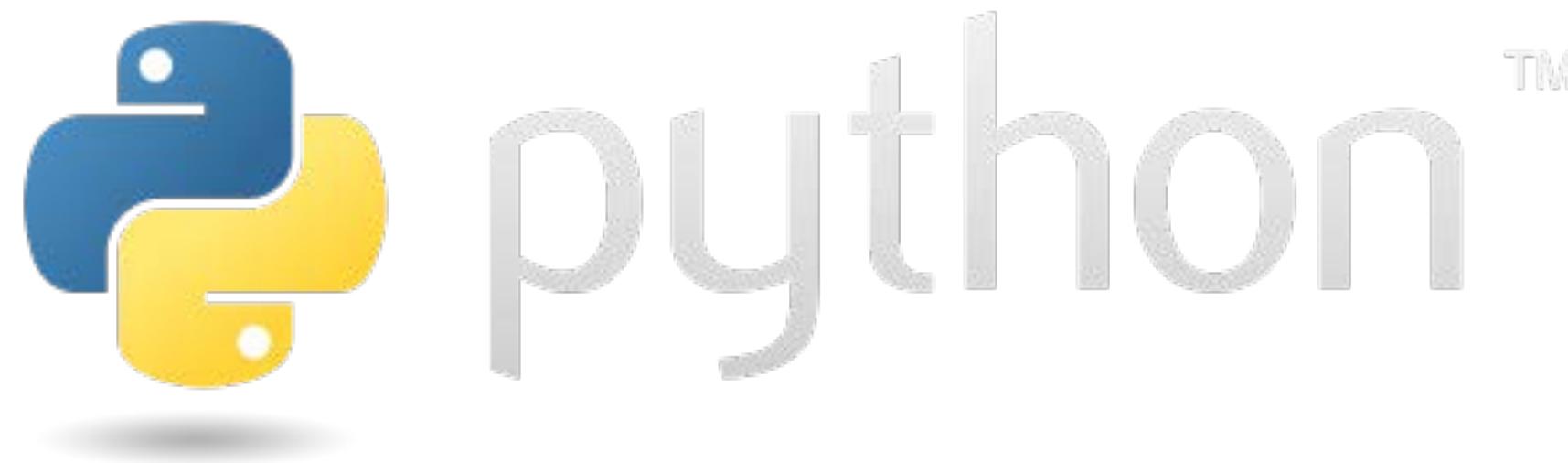


Primary Text for Readings
Available for free on Campus: [link](#)



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 in writing.

We will **manually and 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/2022-datascience-homework/tree/main/HW0>

Homework 0

Introduction to Data Science - MATH 4100 / COMP 5360.

This homework is due before class on Thursday, January 13th.

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:

Github

<https://github.com/datascience-course/2022-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/2022-datascience-homework
```

This will create a folder `2022-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: 31

COMP 5360: 44

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.

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?