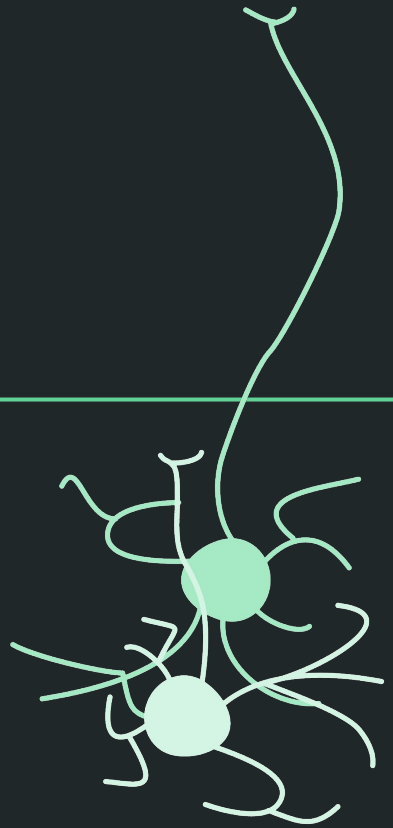
Welcome to BIPN162/BGGN240

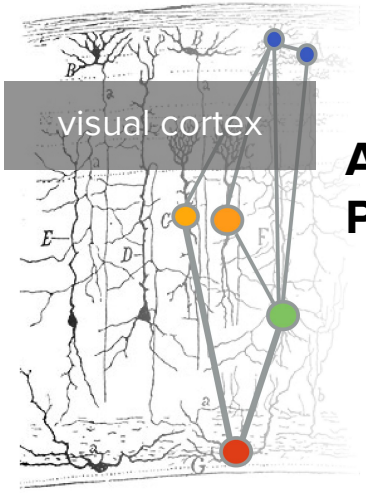
Dr. Juavinett
Jah-vah-nett (or, Dr. J)
she/they

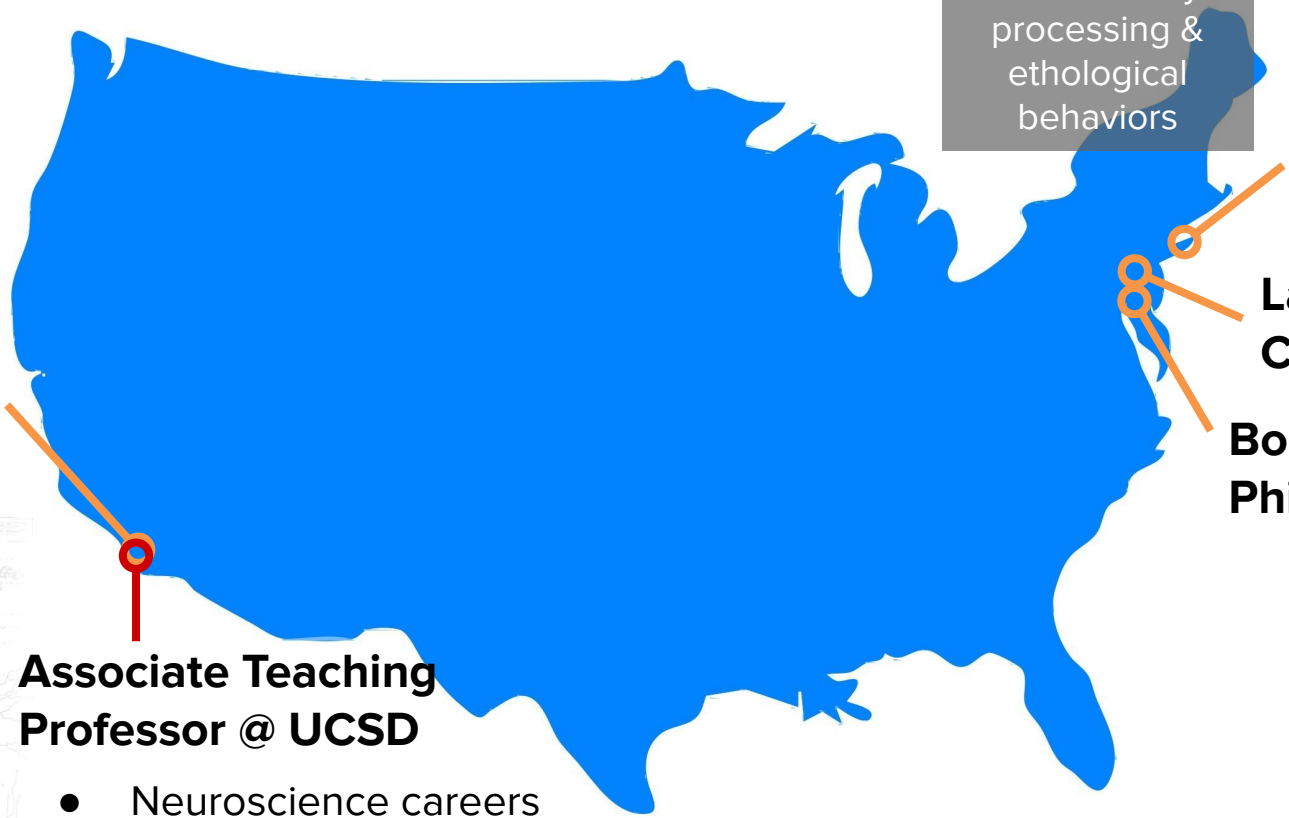


Objectives for this morning

- **Introduce the teaching staff, students, and class**
- Motivate “neural data science” as a field of study
- Discuss course logistics, expectations, & tools

PhD in Neuroscience @ UCSD





multisensory processing & ethological behaviors

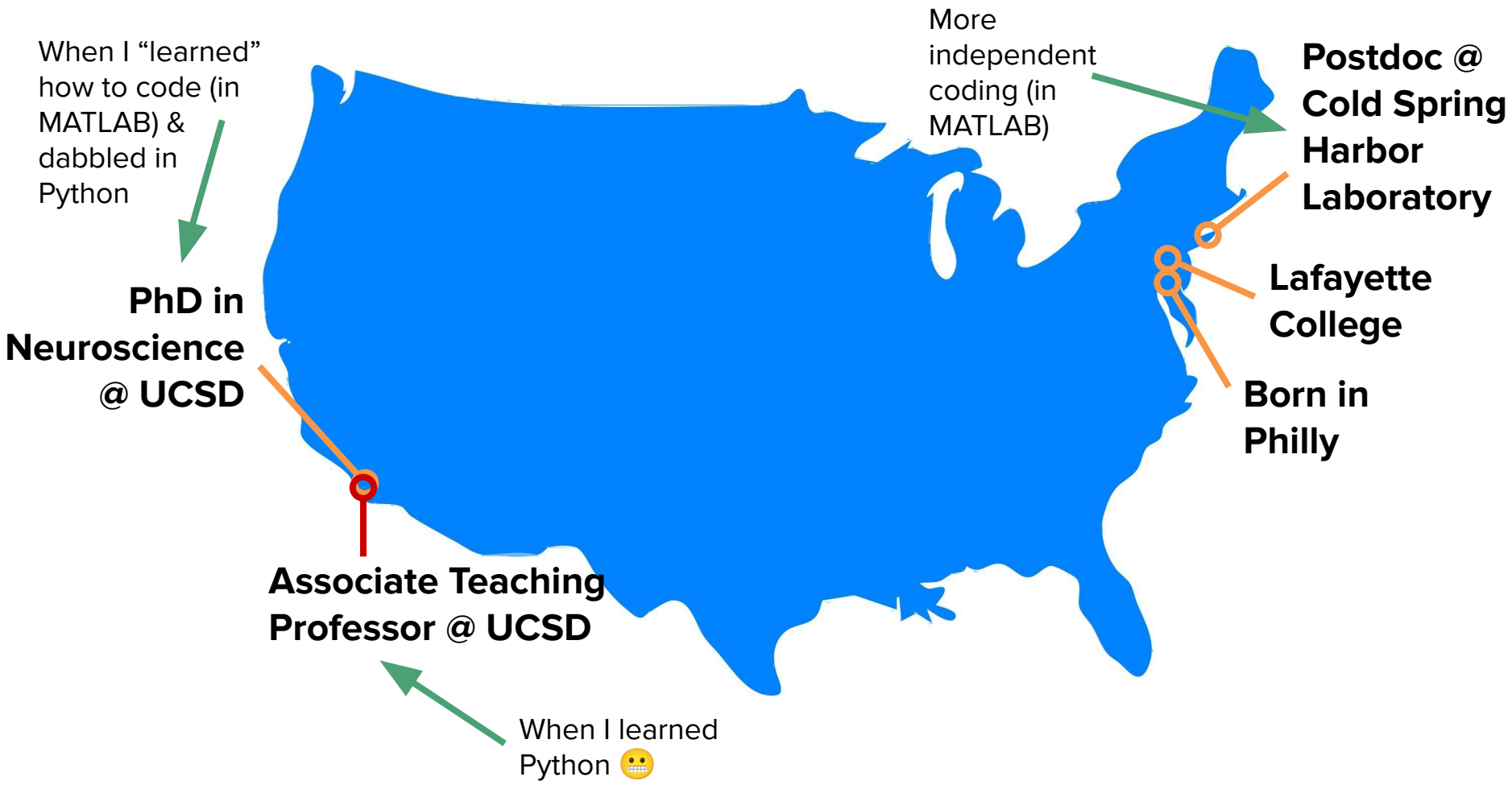
Postdoc @ Cold Spring Harbor Laboratory

Lafayette College

Born in Philly

Associate Teaching Professor @ UCSD

- Neuroscience careers & education
- Open-source data


When I “learned” how to code (in MATLAB) & dabbled in Python
PhD in Neuroscience @ UCSD
Associate Teaching Professor @ UCSD
When I learned Python 😬
More independent coding (in MATLAB)
Postdoc @ Cold Spring Harbor Laboratory
Lafayette College
Born in Philly

Introduction to our Instructional Assistant!

Jeffrey Liu

(he/him)

Bioengineering & Neurobiology

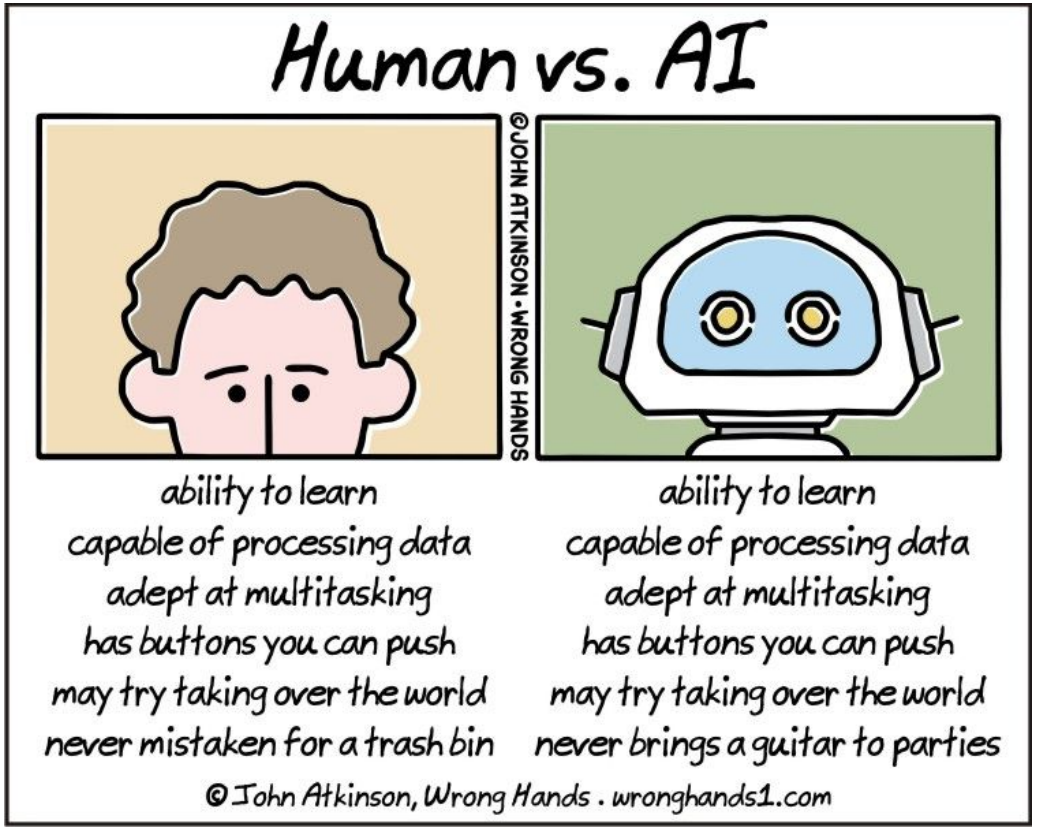
(Double Major)

jcliu@ucsd.edu

Let’s be human, for just a second.

With one or two people next to you, share:

- Your name, major, and preferred pronouns
- Why you’re taking this course

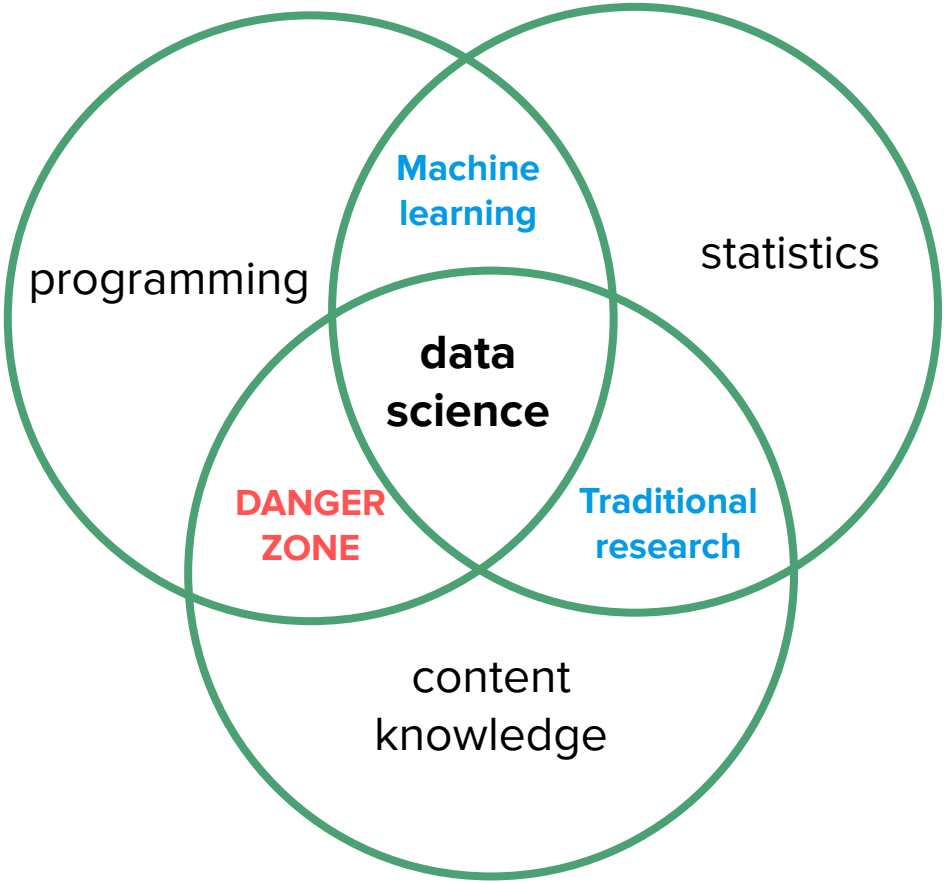


Objectives for this morning

- Introduce the teaching staff, students, and class
- **Motivate “neural data science” as a field of study**
- Discuss course logistics, expectations, & tools

What is *data science*?

What is data science?



class

Why this Book is Needed

Reason 1. Data Science is Hot. Really hot. *Bloomberg* called data scientist "the hottest job in America."[1] *Business Insider* called it "The best job in America right now."[2] *Glassdoor.com* rated it the best job in the world in 2018 for the third year in a row.[3] The *Harvard Business Review* called data scientist "The sexiest job in the 21st century."[4]

🔥

From "Data Science Using Python & R" (2019)
Quote from: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

"[W]hat data scientists do is **make discoveries while swimming in data**... At ease in the digital realm, they are able to bring structure to large quantities of formless data & make analysis possible."

- *D.J. Patil, & Jeff Hammerbacher (2013)*
  (who apparently coined "data scientist" as a term)

What does *data science* have to do with *neuroscience*?

Neuroscience has more data than it knows what to do with right now.

And we have the computing power & statistical frameworks to make some sense of it!

Neuroscience data is getting bigger

One ~500 mm^3 mouse brain…

assuming ~1,000 connections for each neuron, the resulting connection matrix contains ~10^{11} entries

≃ a few hundred gigabytes

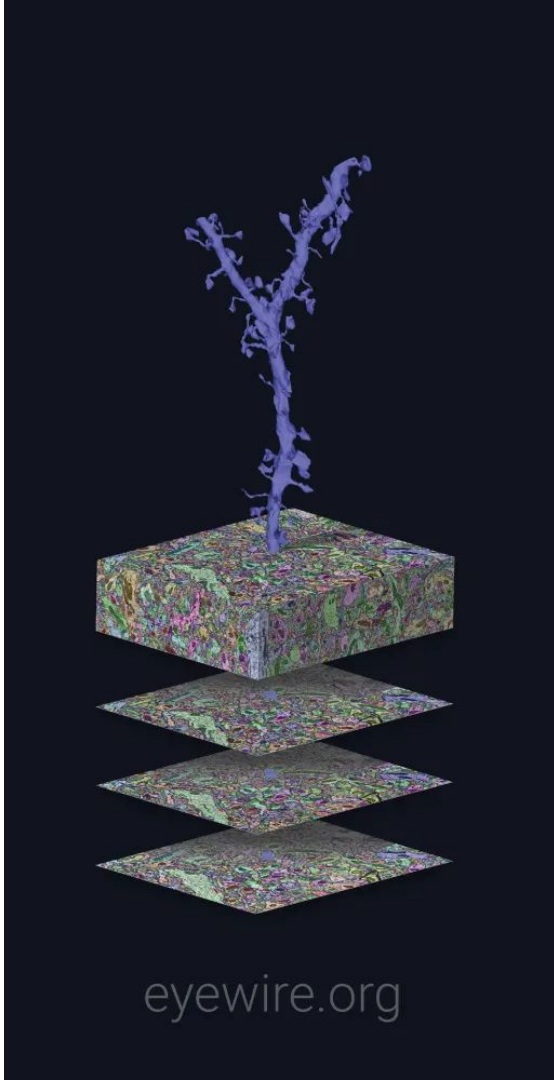
imaging at 5 nm × 5 nm × 40 nm resolution

≃ 500 petabytes

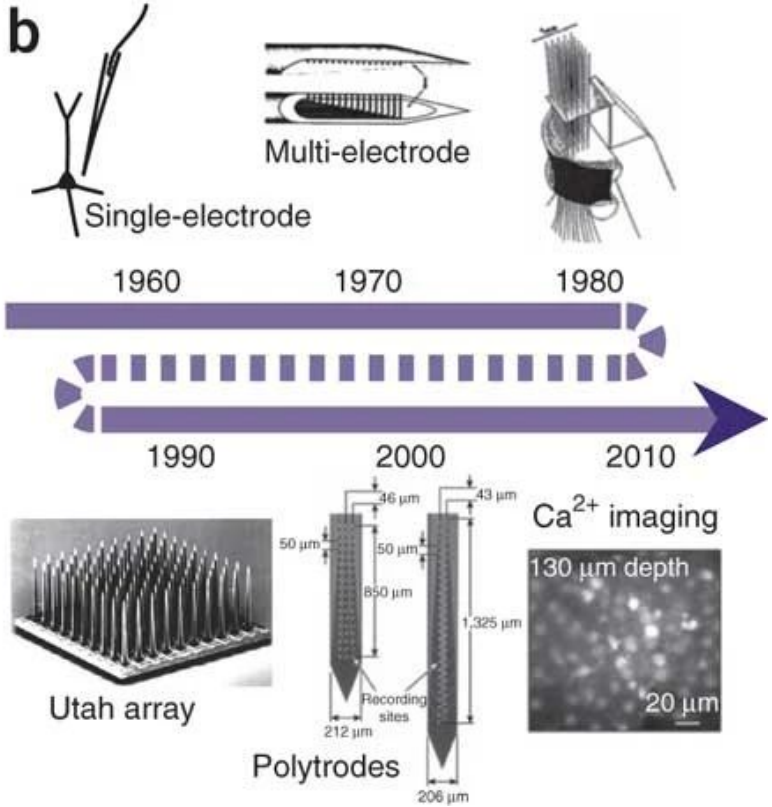
recording from each pixel for 20 min at 1000 Hz

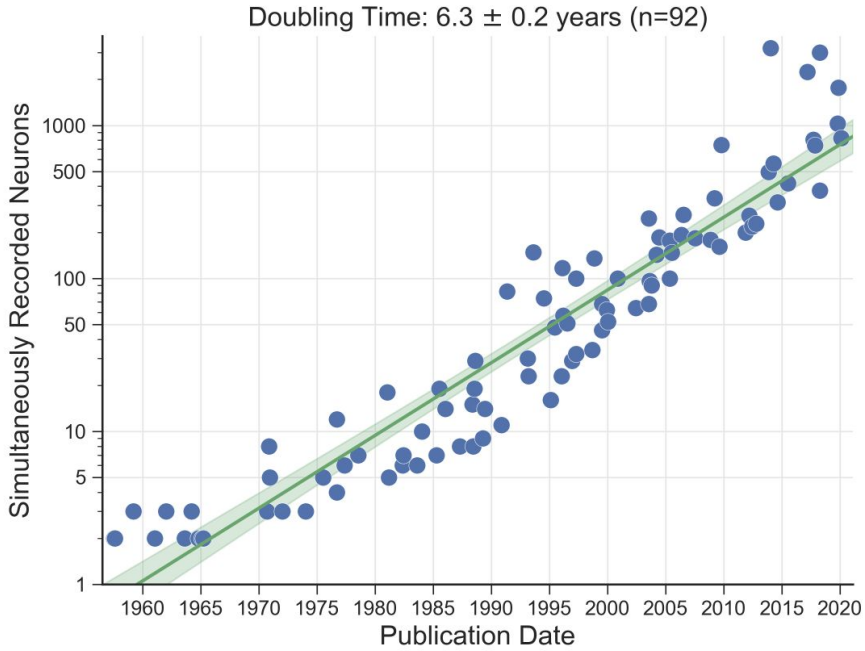
≃ 500 petabytes

Estimates: Engert, 2014; Image: EyeWire



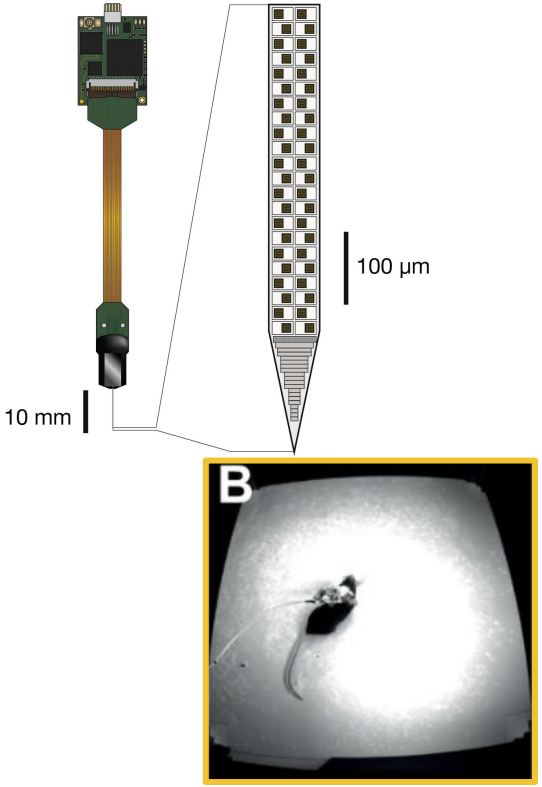
The number of neurons we can record from is growing exponentially

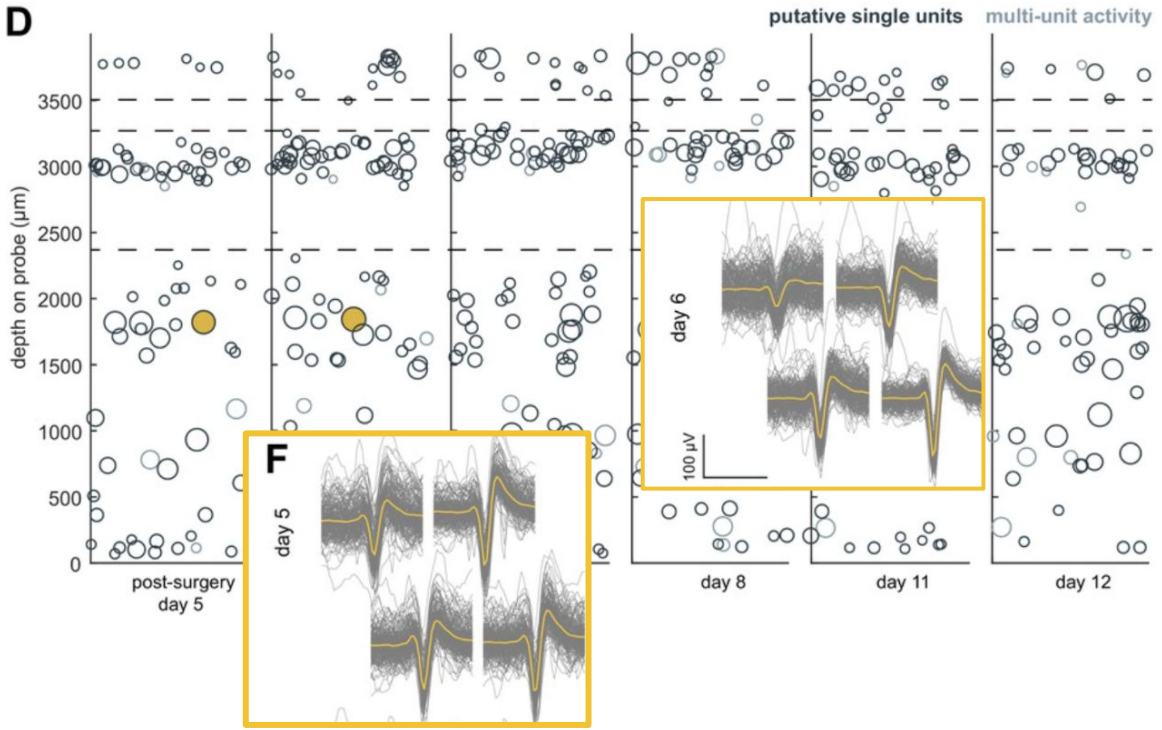




Stevenson & Kording, 2011;
https://stevenson.lab.uconn.edu/scaling/

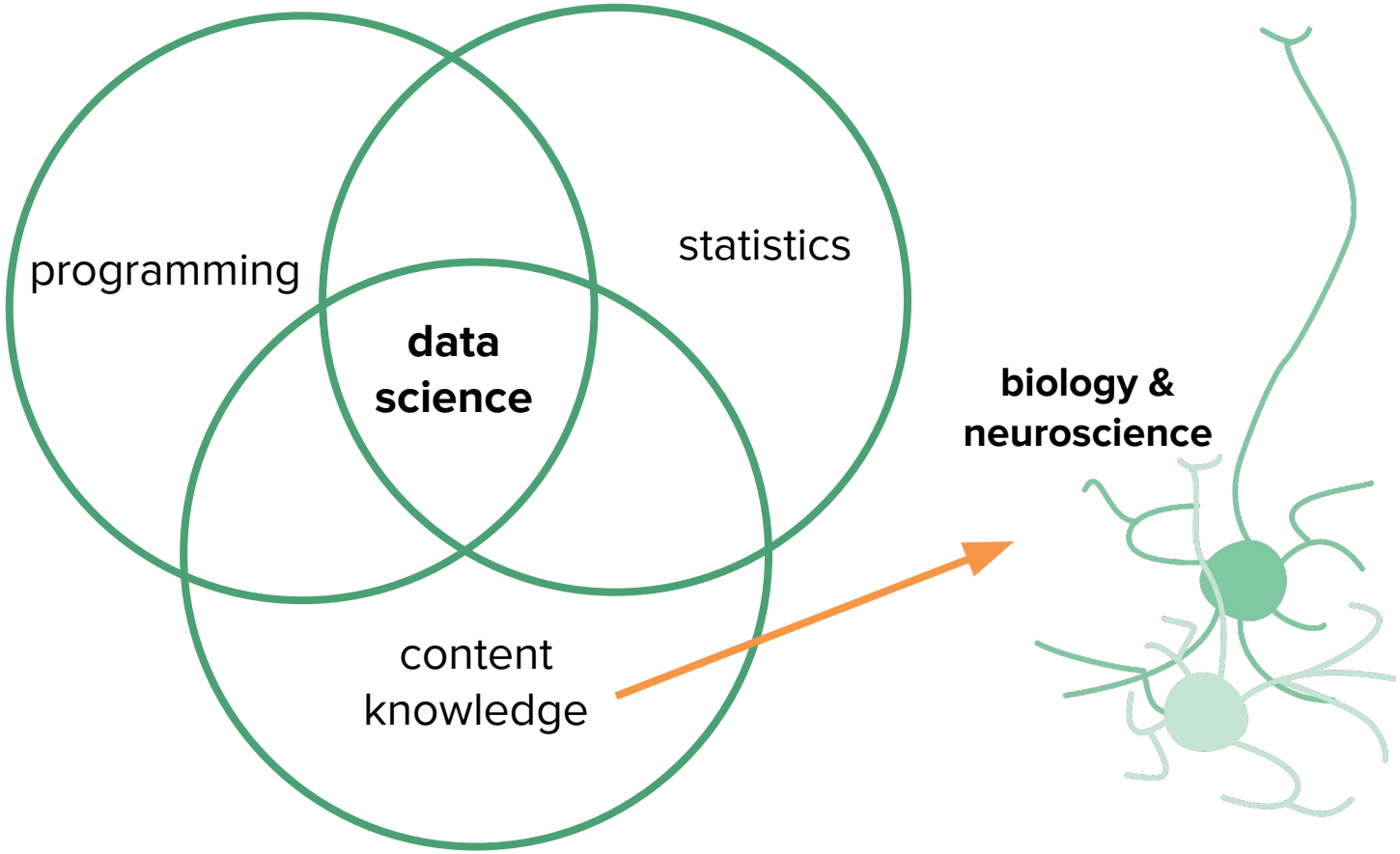
Recording from hundreds of neurons simultaneously in a freely moving animal





Juavinett, Bekheet, Churchland, *eLife*, 2019

What is *neural* data science?



Many of the big questions in neuroscience are **statistical questions in disguise**. *What information is encoded in neural populations? How do we define cell types? How do we decide which experiments to run next? What is one region of the brain telling another?* We need more people who can speak the language of neural science and statistics simultaneously to translate neuroscience questions, theories and prior knowledge into statistical models. There will be a huge impact from people who can bridge these two worlds and make progress on both sides.

Liam Paninski (Professor at Columbia) — in this interview

https://www.simonsfoundation.org/2018/11/19/why-neuroscience-needs-data-scientists/

Neural Data Science

"[I]nherently *interdisciplinary*, bringing principles from data science to bear on neural data to answer questions that are neuroscientifically relevant."

- Pascal Wallisch (2017)

What does the data tell us?

Computational Neuroscience

Using mathematical approaches to:

- build biophysically plausible models of neural processes
- identify the computations corresponding to physiological processes

Can we model the brain using computational approaches?

Case study: What are the mechanisms of human brain development?

How would you address this question?

Case study: What are the mechanisms of human brain development?

Whitaker et al. (2016) used MRI to measure cortical thickness (CT) and myelination (MT) in 14-24 year olds

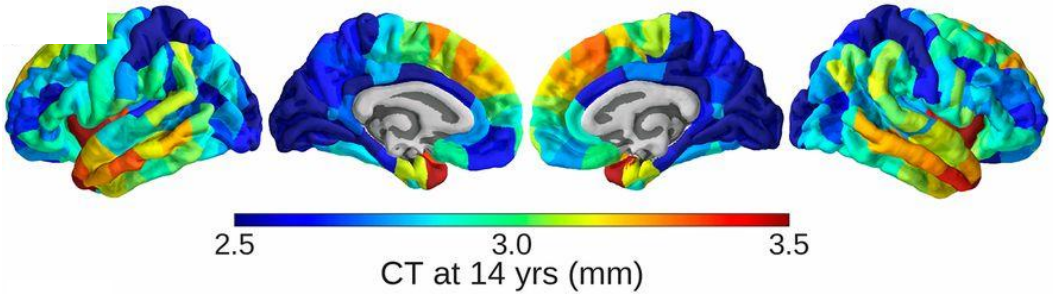
They used a data science approach to answer the question of what drives cortical shrinkage in adolescence.

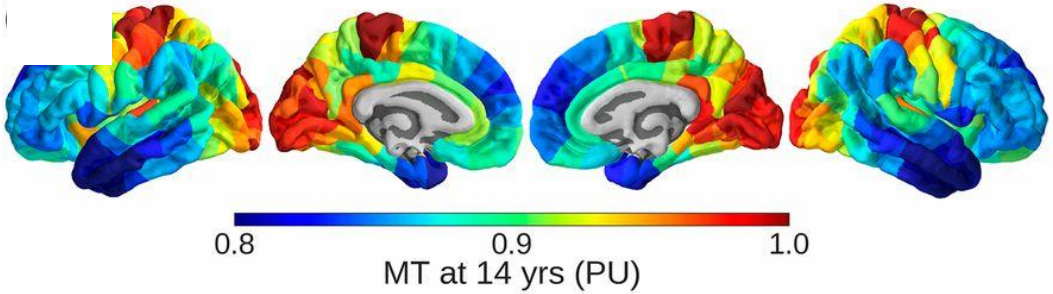
(The cortex shrinks from 3.5 mm at age 13 to 2.2 mm at 75!)

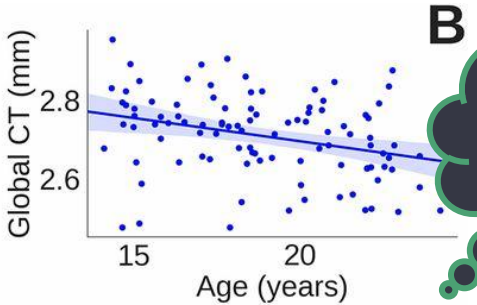
Kirstie Whitaker, lead director of *The Turing Way*
https://the-turing-way.netlify.app/index.html

Case study: What are the mechanisms of human brain development?

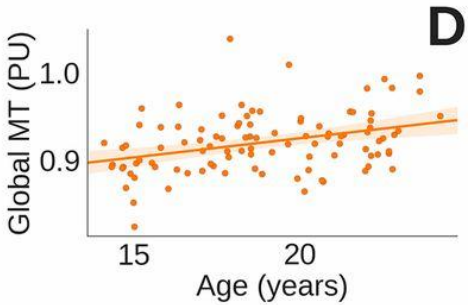
Whitaker et al. (2016) used MRI to measure cortical thickness (CT) and myelination (via MT) in 14-24 year olds







Linear regression!

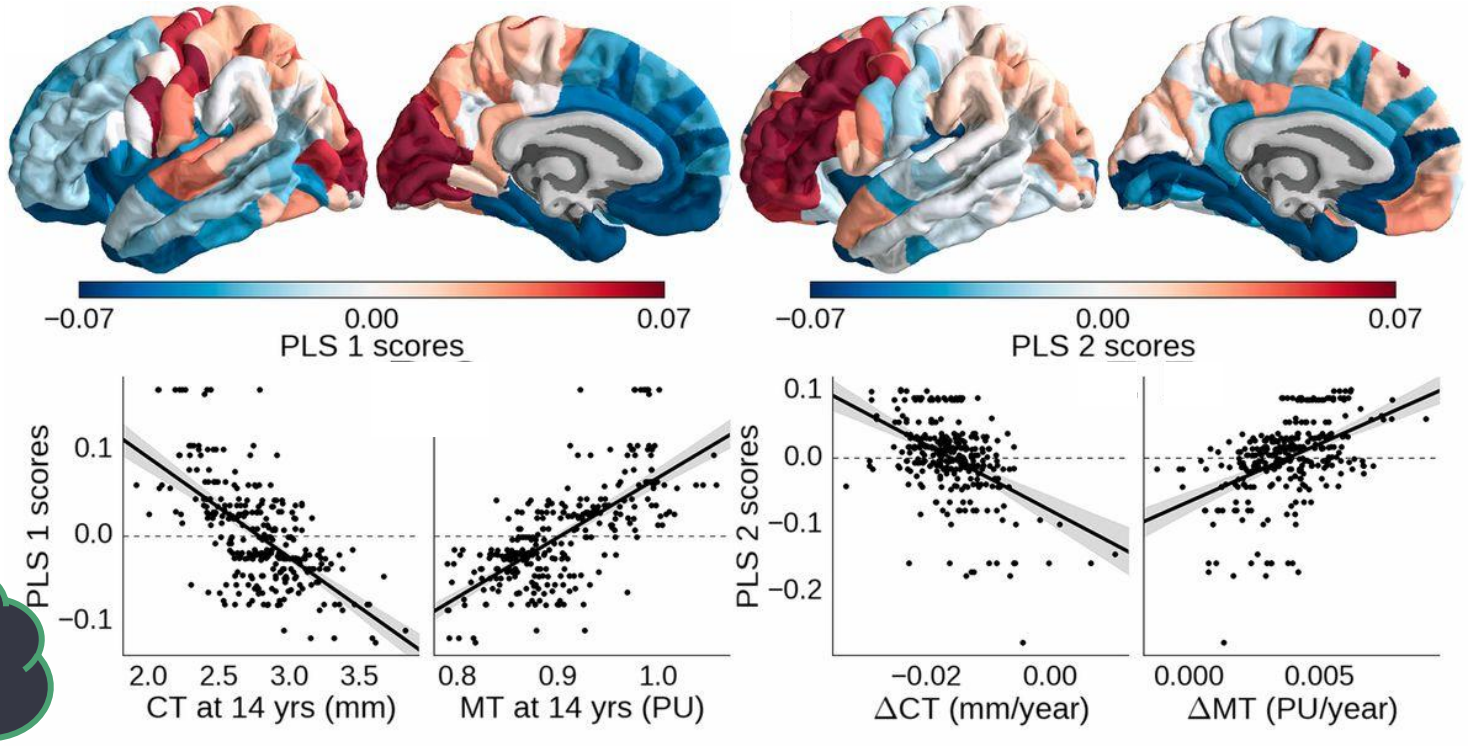


Which genes correlate with thickness & myelination patterns?

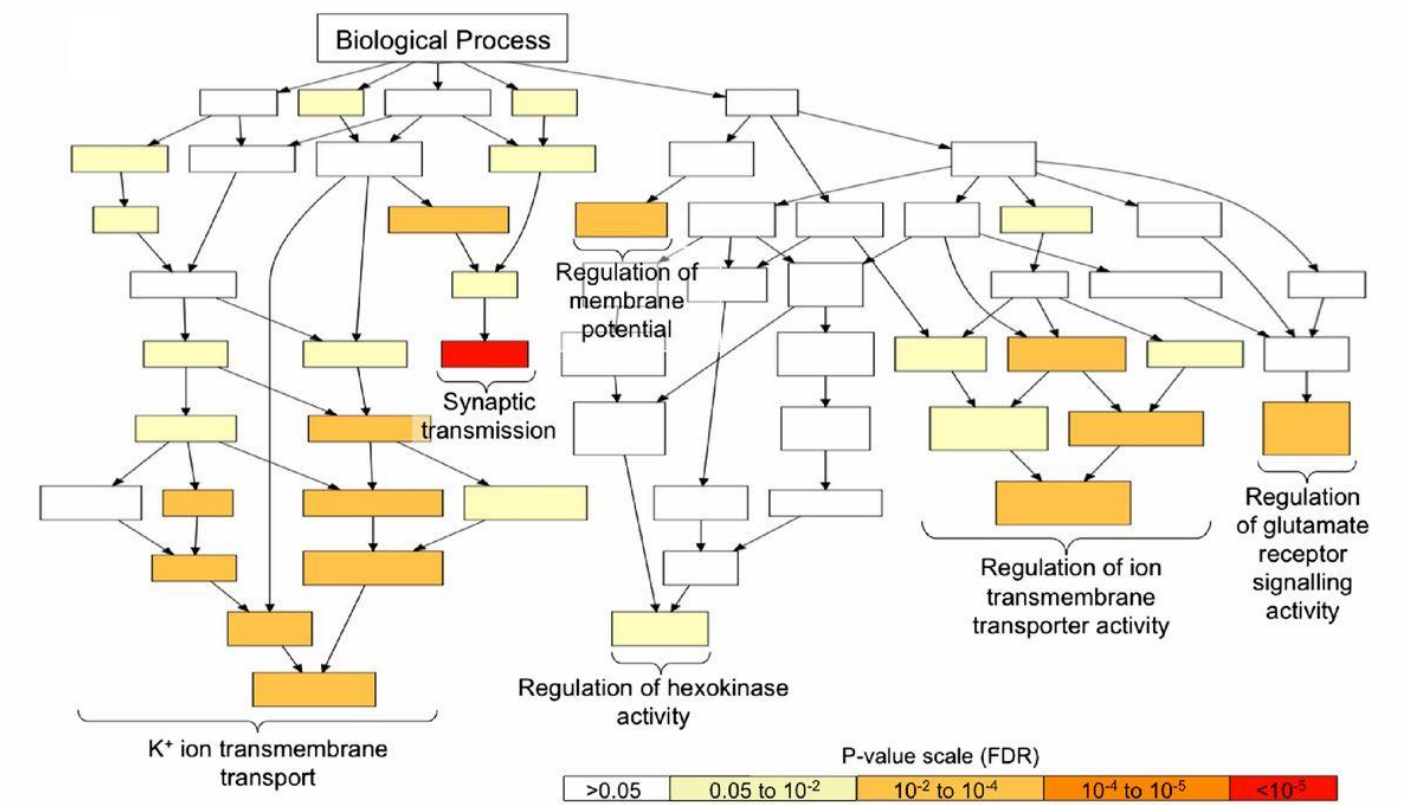
Aligned MRI data with Allen Institute gene expression data

PLS = partial least squares component; represents many genes

Dimensionality reduction!



Genes involved in synaptic transmission, regulation of glutamatergic signaling, and potassium ion channels are related to cortical shrinkage & increased myelination



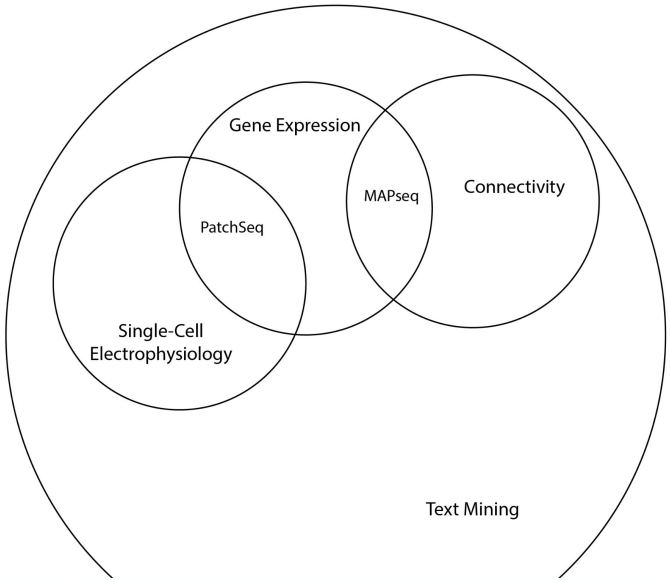
This kind of “big science” approach, combining different kinds of data, led to several interesting conclusions

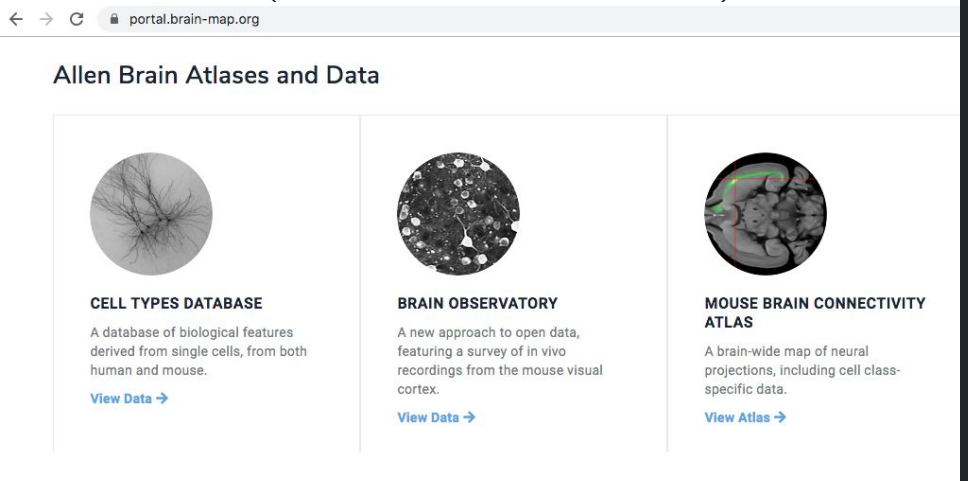
Gene expression profile associated with adolescent change was enriched for genes related to oligodendroglial function, as well as neuronal genes (especially for remodeling of synapses)

... perhaps deviation from the typical developmental trajectory of the cortex is an intermediate phenotype underlying the high incidence of schizophrenia in young people.

“The set of gene transcription markers most strongly associated with this late maturational process was enriched for genes known to confer risk for schizophrenia”

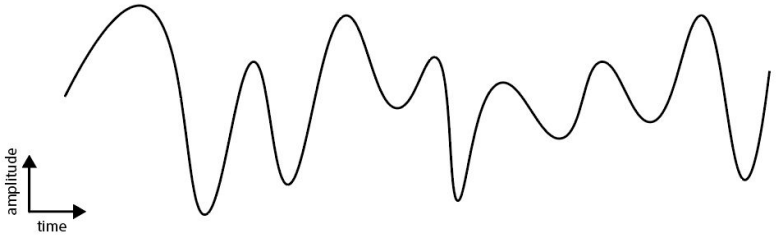
Whitaker et al. (2016)





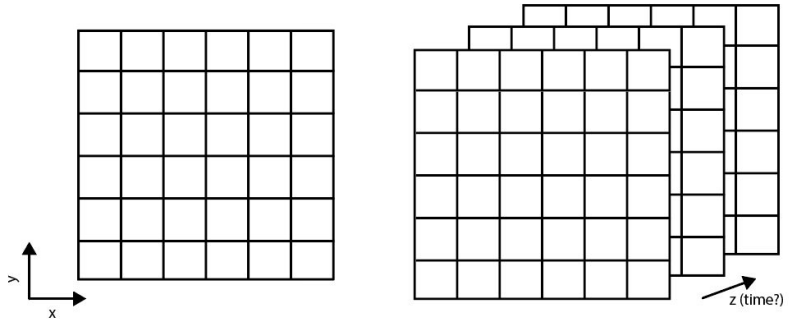
Types of neural data we'll encounter in this course

- Gene expression data
- Electrophysiology
- Imaging data
- Connectivity
- Text & metadata



Time series: something recorded over time. In neuroscience research, this could be voltage, fluorescence intensity, or an animal's movement. The trace above is an example of a continuous time series.

Time series may also be binary, with two discrete states. Such time series may be used to signal when a neuron spikes or not, or a binary regressor in a model.



Images: two- or three-dimensional array of values. The third dimension could be depth in space, different color channels of the image (e.g. RGB), or time (as in fMRI or two-photon calcium imaging).

Types of data structures we'll encounter in this course

- **Metadata:** information *about* the experiment
- **Categorical data**: data that is lumped into categories (e.g., eye color)
- **Time series**: continuous, ordered data, usually physiological or behavioral
- **Images**

How do we ask relevant, interesting questions of big data sets?

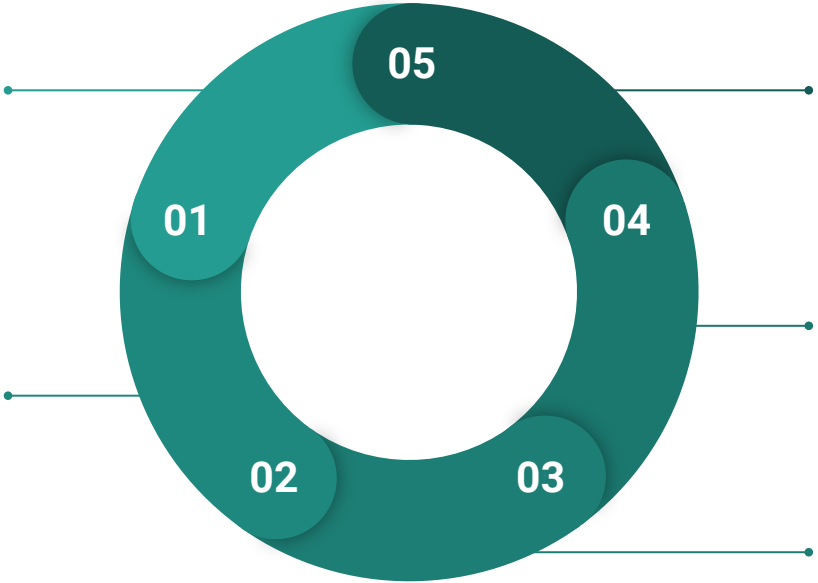
The process of (neural) data science

1. Devise a question
2. Find a dataset that addresses your question
3. Clean & understand the dataset
4. Pull the dataset into Python/your programming tools
5. Perform analyses
    a. Integrate with your existing/other datasets (maybe)
    b. Exploratory analyses
    c. Modeling of the data

The life cycle of neural data

Experimental Design
(Often out of the hands of neural data scientists!)

Collecting measurements & preprocessing
Data wrangling, source separation, image filtering, spike sorting, dimensionality reduction



Hypothesis testing & deriving scientific conclusions
Bootstrap, permutation, multiple comparisons, interpretation

Model building, optimization, and parameter estimation
Dimensionality reduction, neural coding, decoding

Exploratory data analysis
Dimensionality reduction, data management, visualization

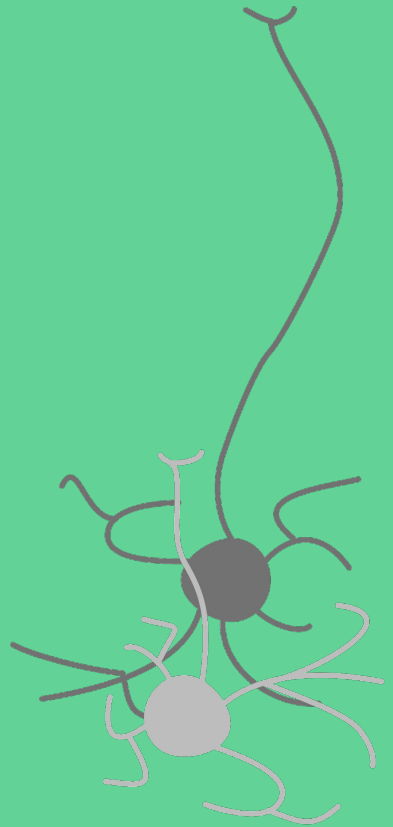
What *can't* we do with open data sets, that we'd normally do in biology experiments?

Let's discuss in small groups

Objectives for this morning

- Introduce the teaching staff, students, and class
- Motivate “neural data science” as a field of study
- **Discuss course logistics, expectations, & tools**

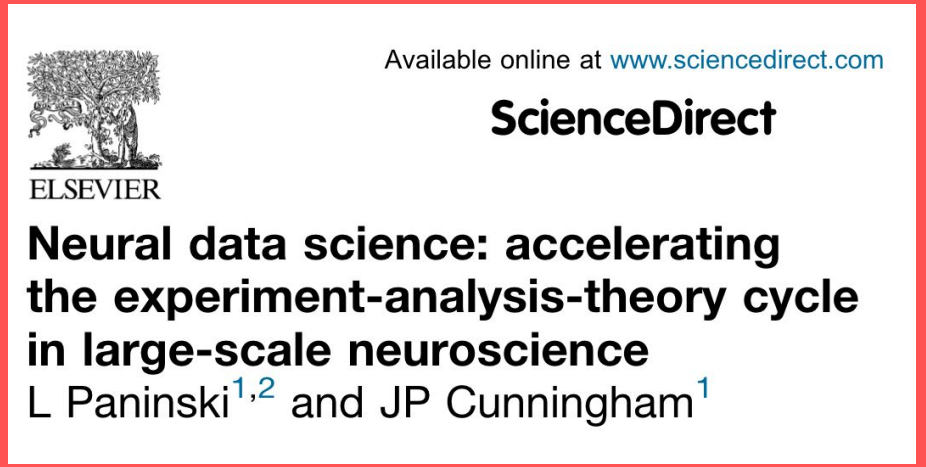
Course logistics



We will not be using our scheduled discussion section time.

Instead, I will assign you things to read or watch.

This week's readings:



Course Objectives

- Develop hypotheses specific to big data environments in neuroscience
- Design a neural data science experiment and excavate data from open sources
- Integrate data from multiple datasets to answer a biological question
- Describe the fundamentals of statistical machine learning tools used in neuroscience
- Dissect data analysis sections of computational / data-heavy neuroscience papers
- Interpret results from common methods in neural data science

In this course, we'll be developing conceptual & technical skills in parallel:

How different neural data sets are collected, preprocessed, and analyzed

Programming, math, and statistics skills necessary for data science

Grading breakdown

Quizzes (25%) -- three, dates on syllabus

In-class work & participation (10%) -- make up as needed

Assignments (25%)

- Due ***most* Wednesdays at 5 pm**
- Worth 2.5-10% each
- Completed individually
- Programmatically graded (via Datahub/NBGrader)

If you're in BGGN, you'll have an extra assignment to write a summary of a neural data science paper. Details are on Canvas.

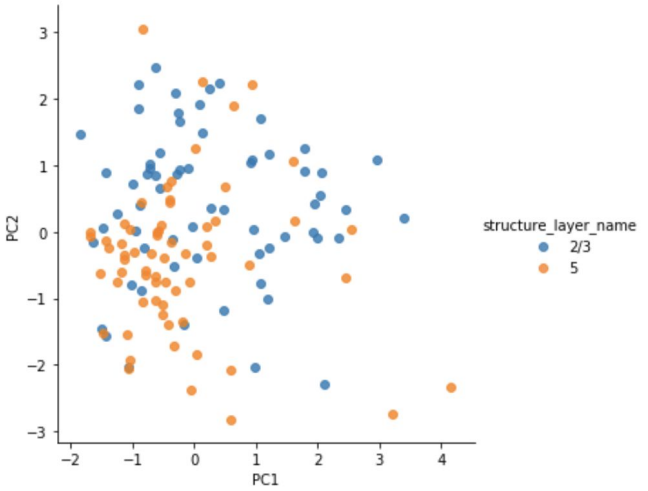
Projects (40%) ...

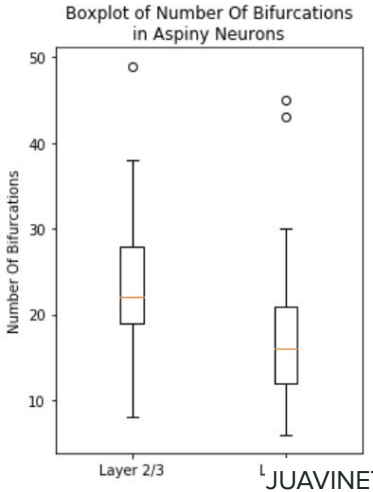
Note: Assignments & project components lose 10% each day they are late.

Projects

- Cell Types Project (15%, groups of 2-3): The first project will ask you to investigate specific **cell types** in the brain, combining information across at least **two** different data sets of your choosing.

- Final Project (25%, groups of 2-3): There are two options for this project.
  Option 1: Integrate **three** different datasets to address a question about brain function.
  Option 2: Choose a paper that uses a new data analysis technique that has been published no more than 10 years ago (most should have code and data publicly available). Implement this method using a new data set (either your own or from publicly available sources).





Office hours

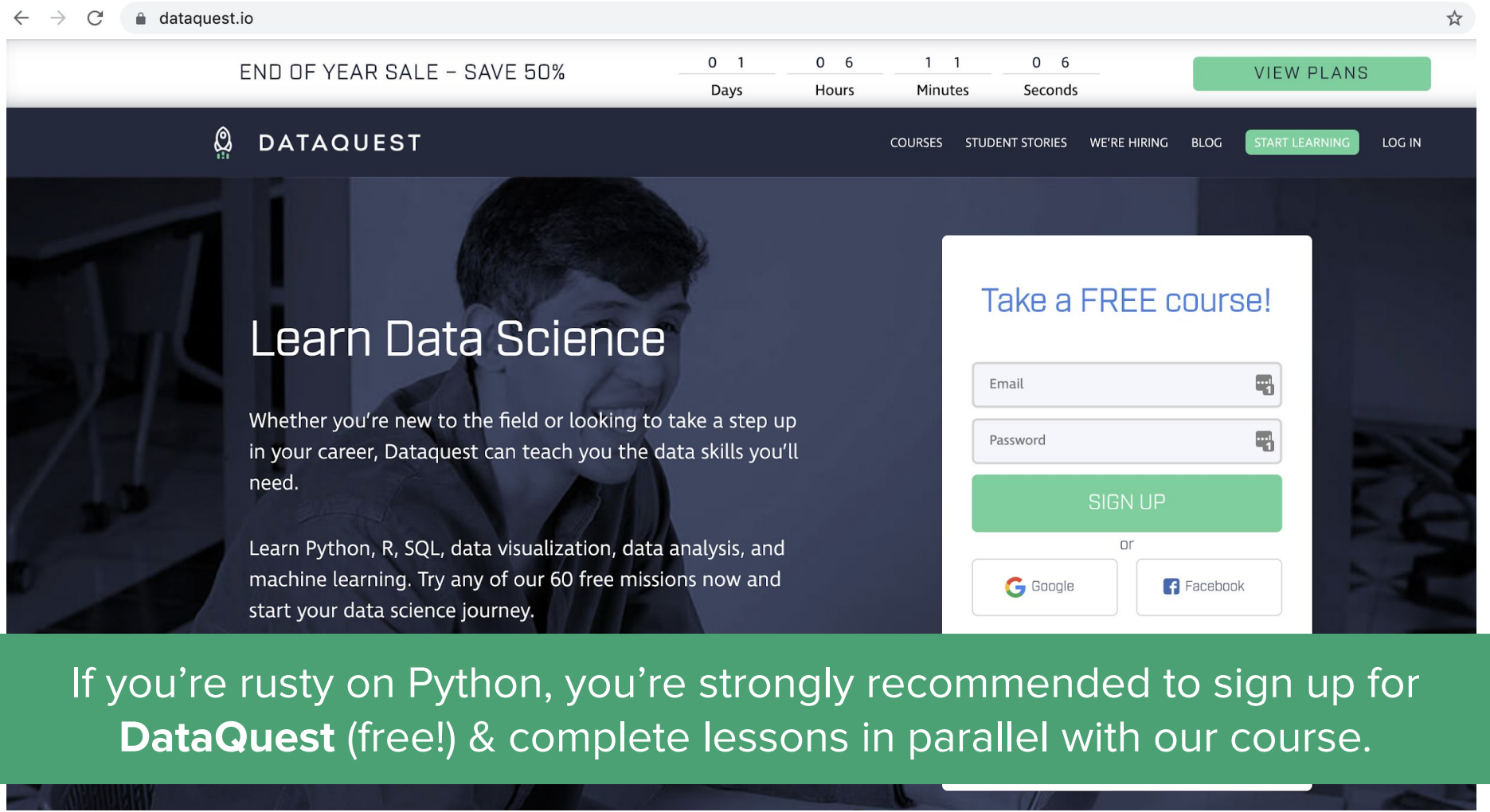
For Dr J, Mondays 3-4 pm

Online or hybrid: https://ucsd.zoom.us/my/ajuavinett, waiting room enabled)

Why should you come to office hours?

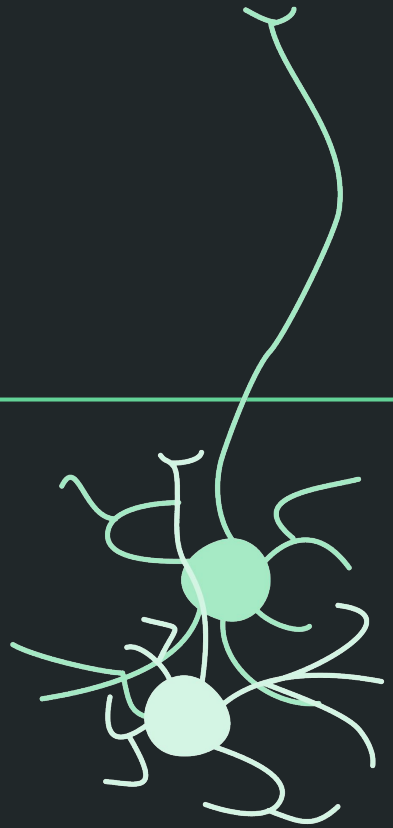
- You have clarifying questions about the course or its content
- You have concerns about the course and your progress
- You'd like to talk about career paths in biology or neuroscience

Jeffrey (UGIA's) office hours are Wednesdays at 2 pm, location TBD.



If you're rusty on Python, you're strongly recommended to sign up for **DataQuest** (free!) & complete lessons in parallel with our course.

Tools for this class



Course Tools

canvas

Submitting non-coding assignments & managing grades

“Tutoring” and project help; more on this later.

Coding exercises & assignments

Sharing public course materials
https://github.com/BIPN162

You can find all of our course materials on either Canvas or the course GitHub: https://github.com/BIPN162/BIPN162_SP24

Lectures

In other words, PDF slides shown during class.

Hosted on GitHub in the Lectures folder

Materials

Jupyter Notebooks that we'll manipulate in class

You can pull these locally or to DataHub, or look at them online via GitHub or Colab/ Binder

Assignments

Jupyter Notebooks, submitted through **Assignments** tab

Answers posted on GitHub after deadline

If I use both a PDF and a Jupyter Notebook during lecture, these numbers will match

💫 THE MAGIC LINK FOR THIS COURSE:

https://datahub.ucsd.edu/hub/user-redirect/git-sync?repo=https://github.com/BIPN162/BIPN162_SP24

💫 THE MAGIC LINK FOR THIS COURSE:

Sync with your datahub:

https://datahub.ucsd.edu/hub/user-redirect/git-sync?repo=https://github.com/BIPN162/BIPN162_SP24

Where our course content lives

To clone Materials to DataHub:

1. Click on the magic link.
2. Log in to DataHub as prompted.
3. You'll be in our course folder now!
4. Save your own copy by adding your initials to the end of the file name. **DO NOT DO THIS FOR ASSIGNMENTS!**

5. Next time you click the link, you'll have a fresh copy, plus your copy.

To interact with Jupyter Notebooks on your computer

OPTIONAL

1. Install Anaconda with Python 3.7 for your operating system.
2. If you're using Windows, download git.
3. In Terminal (Mac) or the Anaconda Prompt (Windows), clone the repository by running the following command:

   `git clone https://github.com/BIPN162/BIPN162_SP24`
4. Open Jupyter Notebook. There are two ways to open:
   - In Terminal (Mac) or the Anaconda Prompt (Windows), type **`jupyter notebook`**
   - Open Anaconda Navigator and launch jupyter notebook
5. On the Jupyter landing page, navigate to the notebook and open it.
   - It will open in a browser but is *not* using an internet connection.

There are multiple ways to interact with the Python interpreter

- Command line
- Jupyter Notebook
- Integrated Development Environments
  - A few good options are:
    - Visual Code (https://code.visualstudio.com/download)
      - Integrated with Copilot (more on this later)
    - Spyder (Included with Anaconda, the recommended install)

Integrated Development Environments (IDEs)

- Help you write, debug, and compile code
  - **Compiling** is the process of translating your **source code** into **machine code**
- Useful because they have features like **line numbers** and **syntax highlighting**, which colors your code based on the syntax.
- Often have auto-completion, memory for commands, and provide information about functions

Anaconda is an open-source distribution of Python, focused on scientific computing in Python.

Includes:

- “Conda,” a package management tool
- Useful code packages
- A couple applications for editing & running code:
  - Spyder (Python IDE)
  - Jupyter Notebooks

A few notes

Macs have a native installation of Python.

- It may be older & will not include the extra packages that you will need for this class, and is best left untouched.
- Downloading Anaconda will install a separate, independent install of Python, leaving your native install untouched.

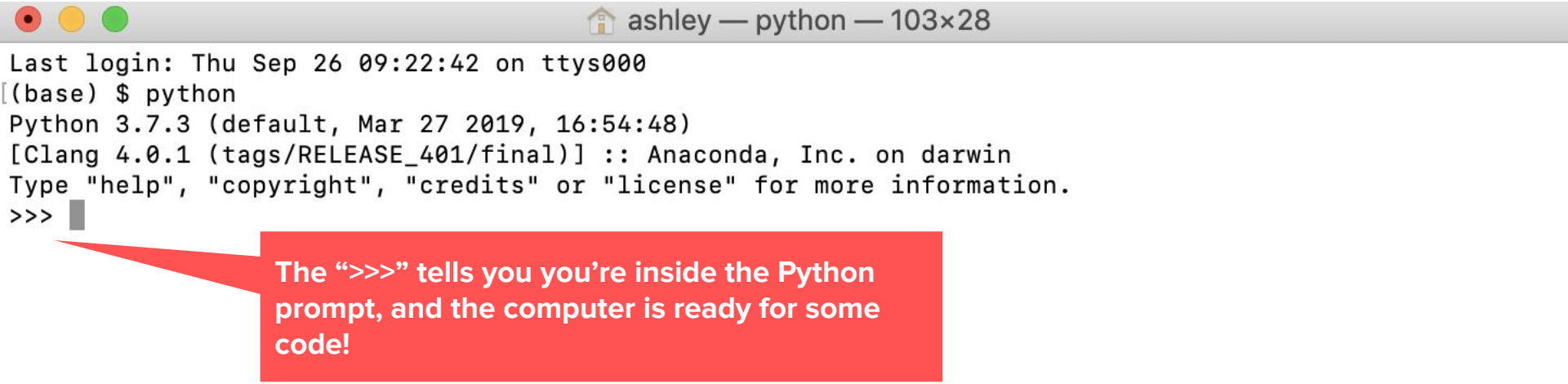
Windows does not require Python natively and so it is not typically pre-installed.

If you're not sure which Python your computer is using, ask it (in Python):

```
>>> which python
```

If you have a Mac

- Macs ship with Python already installed. You can check which version by opening **Terminal** & typing "**`python --version`**"
- This will show you the version of Python that you have installed.
- **For this course, we'll be using Python 3.6** (or above)**.**



Before next class

- Access Canvas (canvas.ucsd.edu) & DataHub (datahub.ucsd.edu)
- **Read Humphries “Neural Data Science” and Paninski & Cunningham (2018) -- links on Canvas & syllabus**
- (Optional) complete relevant DataQuest tutorials (see syllabus)
- (Optional) Install Python & an IDE (e.g., VS Code)

You only *really* need access to the DataHub, but having the ability to use Python & Jupyter Notebooks on your local computer may be useful!