

COGS138: Neural Data Science

Lecture 20: Final lecture

C. Alex Simpkins, PhD

UCSD Dept. of Cognitive Science, Spring 2023

RDPRobotics, LLC

http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23

rdprobotics@gmail.com | csimpkinsjr@ucsd.edu

Plan for today

- Thank you to Siddhant
- Announcements
- In class discussion about EDA checkpoint
- Concluding the class

Let's please thank Siddhant for his time and effort

Announcements

- **Deadlines upcoming this week:**
- **Tuesday:**
- **Wednesday:**
- **Saturday (6/10):**
 - EDA checkpoint 11:59pm
- **Sunday:**
 - Reading Quiz 4 6/11 @11:59pm
- **Next Saturday 6/17** for A5 and EC assignment, missing quizzes, etc, final lecture quizzes, etc

Announcements

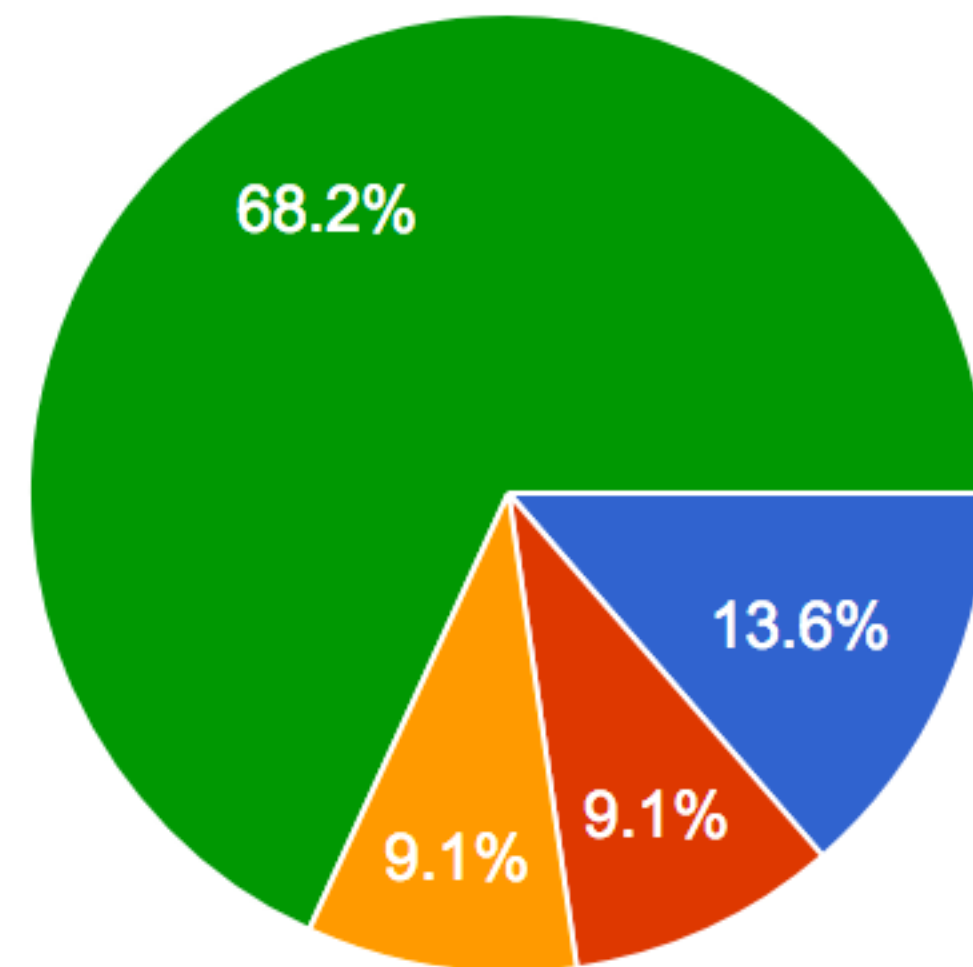
- Extra office hours for Dr. Simpkins this week (Friday at 12-1pm)
 - Will be available all day either over zoom via quick appointment or piazza/email as always
- Project feedback for the Data checkpoint will be released by the end of the evening if not there yet

Preference for your **favorite** method of doing the final presentations and feedback for each other



(choose the top that you prefer, and you can vote on a secondary option that works for you, as well as comment)

22 responses

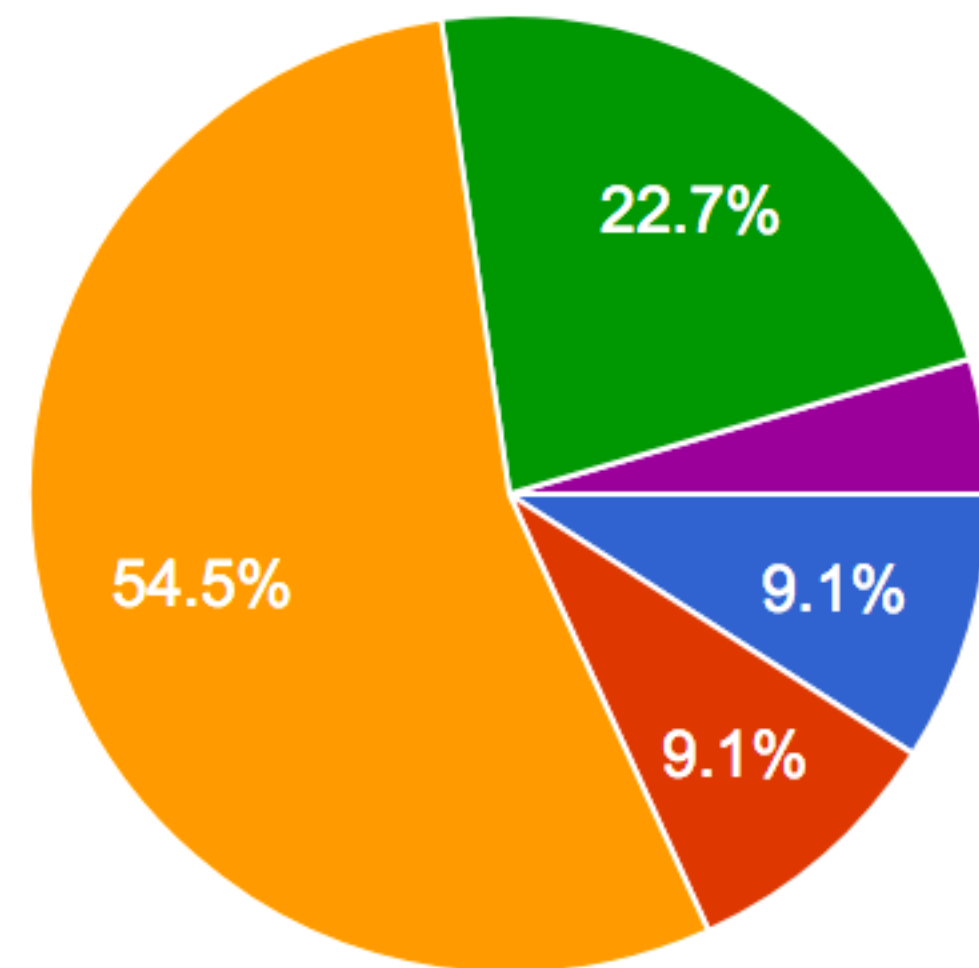


- Live presentations during the scheduled final in class, with remote people who cannot physically go attending over zo...
- Live presentations during another scheduled time later in finals week (we could locate a room that has availabilit...
- Live Zoom presentations all remote during some scheduled time we all vot...
- Offline pre-recorded presentations (5-10 min each), and everyone has to watch...

Preference for your **second favorite** method of doing the final presentations and feedback for each other

(choose the second favorite that you prefer, and feel free to add detail in the comments area)

22 responses

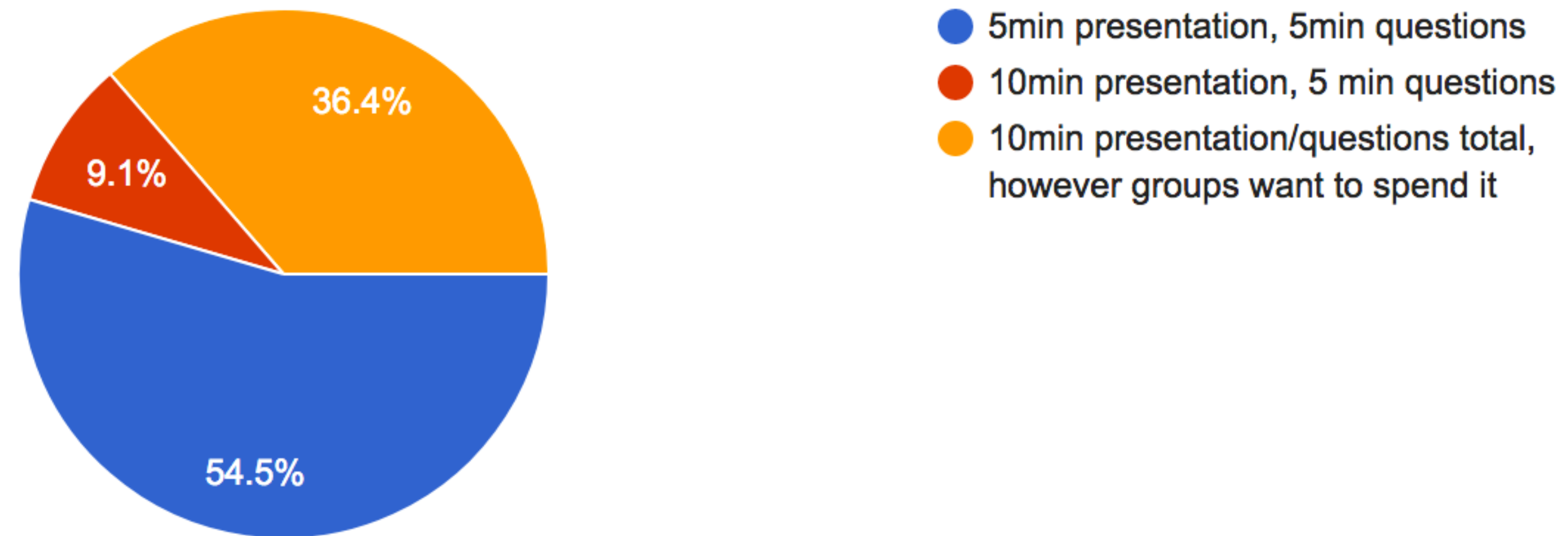


- Live presentations during the scheduled final in class, with remote people who...
- Live presentations during another scheduled time later in finals week (w...
- Live Zoom presentations all remote during some scheduled time we all vot...
- Offline pre-recorded presentations (5-10 min each), and everyone has to watch...
- Pre-recorded presentations with a paper write-up alternative

What length do you prefer? 5min presentation with potentially 5min questions is short to get it all across, but takes less overall time, 10min is longer especially if we do 5min questions, or we could set 10min total however the groups want to spend it.



22 responses



Announcements II

- Final presentations
 - Vote was mainly for offline, about 1/3 online
 - Backup was for zoom - technical issues
- **We will do offline - simpler**
 - Every group will record a 5-10min presentation, share a link (you can do however it works but we must all be able to access it)
 - Each individual will review 3 at minimum, fill out google form, then each additional one you review will give you an additional 0.5% on the group review project grade portion
 - We will encourage a piazza discussion for each project
 - End of quarter optional additional hour discussion

Project schedule

Task due	Date due	Description
Previous project review	5/23/2023 at 11:59pm (Tuesday)	Select 2 of the 3 available, review as individuals and then come together as a group to submit your responses to the questions after a discussion. This will orient you to the class project
Project proposal	5/26/2023 at 11:59pm (Friday wk8)	Generate your question, hypothesis, initial data sets you'll be working with, etc., describe your plan, schedule, who is doing what, potential issues, suggested analysis and how it will answer your question
Data checkpoint	6/2/2023 at 11:59pm (Friday wk9)	Builds on the proposal by taking the feedback from PP above and actually getting, loading, describing your data,
<i>EDA checkpoint</i>	<i>6/10/2023 at 11:59pm (Saturday wk10)</i>	<i>Builds on the previous checkpoint, essentially most of your analysis should be done by this point</i>
Final report	6/15/2023 at 11:59pm (Thursday Fin wk)	Due Thursday of finals week so we can grade before the Tuesday deadline, otherwise your grade may be delayed
Group evaluations	6/15/2023 at 11:59pm (Thursday Fin wk)	You will evaluate each other based on participation and performance, this will contribute to your overall final project grade 5%)

EDA checkpoint

- Link to EDA checkpoint:
 - https://github.com/drsimpkins-teaching/cogs138/blob/main/main_project/EDACheckpoint_groupXXX.ipynb
- One additional question to add - what do you think given the exploration you have done that your biggest challenges are and how will you address them?
- Link to outline of what to include:
 - https://github.com/drsimpkins-teaching/cogs138/tree/main/main_project

Group issues?

- Communicate with us for assistance working things out
- Group strategies, clear communication to avoid misunderstandings, regular updates, be open with ideas (no shooting down approach)

Remaining assignments schedule

- A5 wk10, A4 extra credit
- Lecture quizzes - will be released and you complete by the end of finals week
- Final course survey
- Otherwise just project

A4 - getting pysurfer working is a task...

- Working with UCSD IT on the dependencies since the user install version does not appear to work well - complex paths to update
- This will be an optional assignment or for your edification

A5 - Mouse v. Human cells

- Less dependent on complex dependencies
- <https://allensdk.readthedocs.io/en/latest/install.html>
- installation
- path
- you don't need to have any data included in your repo, it's all downloaded by the allensdk by command

Last time...

Course review

Pulling it all together

What is Neural Data Science?

What is neural data science?

- A new field of neuroscience emerging, still evolving, in its infancy, incomplete
- Data science methodology applied to neuroscience to ask and answer scientific questions
- When addressing the brain we are faced with unique challenges, and must draw upon not only neuroscience or basic data science, but heterogeneous data and insights we have in order to gain new insight, and that has *many* complications as well as requires perspective
- The brain functions in complex ways that at times defy normal approaches to expose underlying dynamics
 - e.g. Sparse encoding within a population - significant properties such as cognitive or behavioral function is often encoded in such a way that averaging over many measures wipes out or creates spurious signals

There is no ground truth!

- This is a huge challenge - there is no correct answer, no training labels for neural data, only behavior
 - This differentiates the field
- We have no idea how brains encode signals
- All the behavioral labels we create are guesses, and almost always only “clues”
 - Neuroanatomy and neurophysiology - everything we think we know is based on assumptions guided by measurement and calculations
- All the data we are recording is incomplete, it's like trying to understand a sports game by watching one individual running around, not seeing anything else at all and only seeing a 2 second snapshot at different times

So how do we address this?

- We must ask the right questions, be open to new interpretations of data that discard established assumptions at times
- We must consider not just a single but a wide variety of data as part of a study
- This is often a necessity with modeling the brain - we have to consider
 - Neural signals, connectivity, spiking (which might be measured from EEG, fMRI, MEG, single unit recording, high density recording, eCOG, LFP, other),
 - Behavior (motion capture, eye tracking, others),
 - Gene expression
 - Literature (digested with LISC type automation)
 - Stimuli/environment
 - Cognitive context/other state-dependent effects

Why consider a variety?

- More information can draw links that may not be clear otherwise - these are all pieces of a picture
- Limited data source sets may not contain the necessary data for the question we want to ask
 - **Sparsity** - improved results with ***sparse*** datasets (sparse patterns)
 - **Modality** - one set might have patterns, but lack the content explaining patterns, the meaning underlying
 - **Reliability** - one dataset showing statistical significance vs. many confirming from various perspectives
 - **Validity** - are we measuring what we think we are measuring?

Why is it a challenge to integrate them?

- Sampling rate mismatch
- Time/frequency/spatial domains - what is the best form of representation?
 - <https://www.sciencedirect.com/science/article/pii/S1053811919300497>
- Sample rate variability (why does this matter?)
- Sample time mismatch
- Format, software
- Missing data, data mixture/non-tabular etc
- Memory usage
- (Not an exhaustive list)

Now what fundamental types of questions can we ask?

1. Is there structure in the spikes, fields, patterns and other recordings?
2. What might be causing that neural activity (spike, field, pattern, etc)?
3. What information is encoded in the spikes, fields, patterns and other recordings?
4. What connection do the patterns have to the outside world or the body?

What were the course objectives?

Learn how to:

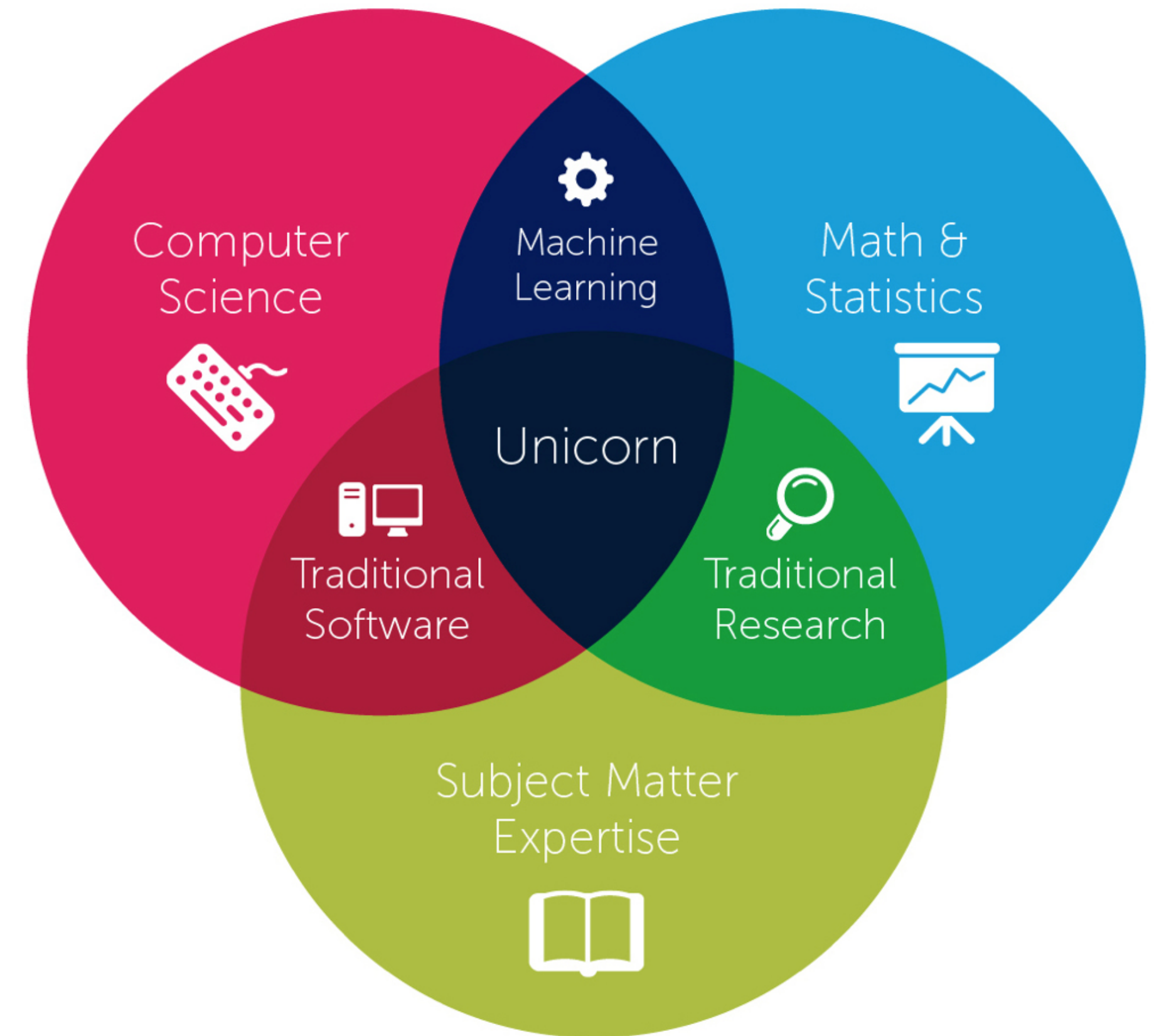
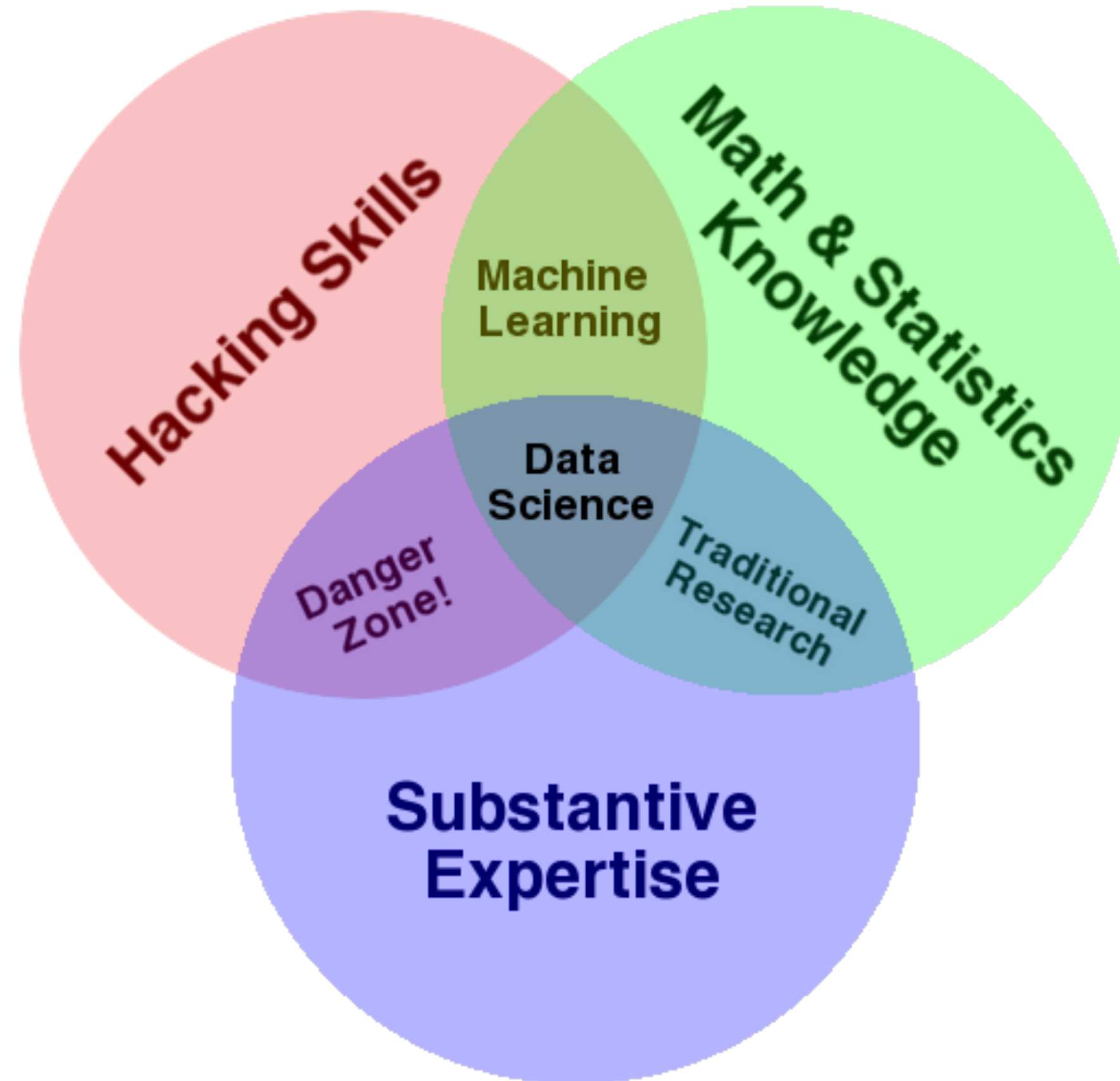
- think from a “data first” perspective: what data *would* you need to answer your scientific questions of interest?
- develop hypotheses specific to big data environments in neuroscience.
- work with many different neuroscience data types that might include data on behavior, brain structure and connectivity, single-unit spiking, field potential, gene expression, and even text-mining of the peer-reviewed neuroscientific literature.
- read and analyze data stored in standard formats (e.g., Neurodata Without Borders and Brain Imaging Data Structure).

What were the course objectives?

Learn how to:

- integrate multiple heterogeneous datasets in scientifically meaningful ways.
- choose statistical model(s) informed by the underlying data.
- design a big data experiment and integrate data from multiple open data sources.
- consider alternative hypotheses and assess for spurious correlations and results.

We asked: What is data science?



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

Defining Data Science

a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.^[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. -Wikipedia

"This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary actions." -David Donoho ("50 years of Data Science)

"an emerging discipline that draws upon knowledge in statistical methodology and computer science to create impactful predictions and insights for a wide range of traditional scholarly fields" - from a panel Rafael Irizarry moderated, shared on SimplyStatistics ("The role of academia in data science education")

"an umbrella term used by organizations to describe the processes used to extract value from data" -Rafael Irizarry's personal definition in "The role of academia in data science education"

"The study of how the quantification of observable phenomena can lead to human understanding of the processes giving rise to those phenomena—or even the ability to predict future outcomes absent human understanding—and why certain phenomena require more or less data to lead to human understanding and/or prediction accuracy". -Brad Voytek's definition

"The scientific process of extracting value from data"

Data scientists ask
interesting questions
& answer them with
data

We then defined neural data
science

Introduction to DataHub

- datahub.ucsd.edu
- Logging in
- Navigating
 - Intro to file structures and how they relate to your computer
- Upload
- Download
- Rename files
- Make folders
- Delete
- Submitting assignments, fetching assignments
- Validating!

The screenshot shows the homepage of the UC San Diego Jupyterhub (Data Science) Platform. At the top, it says "DATA SCIENCE / MACHINE LEARNING PLATFORM" and "UC San Diego". Below that, it says "Information Technology Services - Academic Technology Services" and "Help" and "FAQ". The main header features the "jupyterhub" logo and a large image of a bear sculpture. A yellow "Log In" button is visible, with the text "Registered Users" and "username@ucsd.edu" below it. The main content area has the title "UC San Diego Jupyterhub (Data Science) Platform" and a note: "If you are unable to log in: Please try opening a private/incognito window in your browser | FAQ". Below this, there are two columns of resources: "Student Resources" and "Instructor Resources".

DATA SCIENCE / MACHINE LEARNING PLATFORM

UC San Diego

Information Technology Services - Academic Technology Services

Help ▾ FAQ

jupyterhub

Log In

Registered Users
"username@ucsd.edu"

UC San Diego Jupyterhub (Data Science) Platform

If you are unable to log in: Please try opening a private/incognito window in your browser | [FAQ](#)

Student Resources

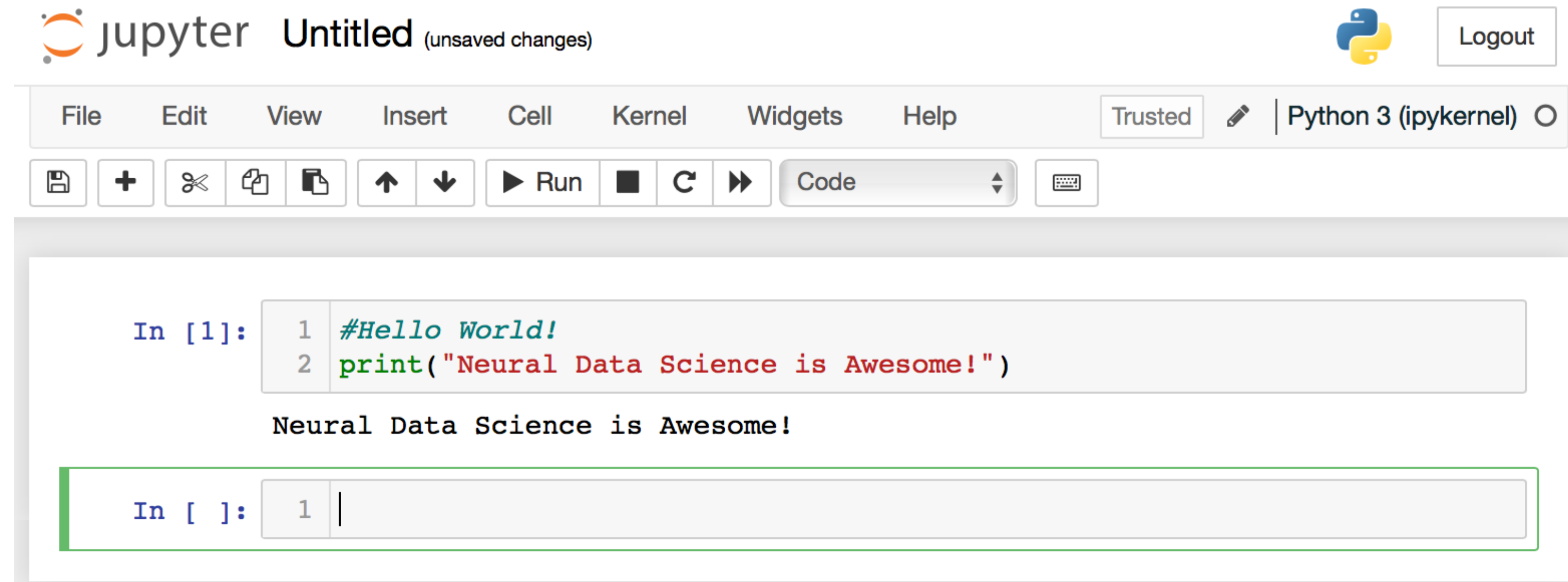
- [Datahub/DSMLP Cluster Status](#)
- [Independent Study Access Request](#)
- [Data Science Resources](#)
- [Datahub/DSMLP Knowledge Base](#)
 - [Launching Containers from the Command Line](#)
 - [Configuring Your Container Launch](#)

Instructor Resources

- [Request Datahub/DSMLP - Instructional Technology Request \(CINFO\)](#)
- [Instructor Guidance for Datahub/DSMLP](#)
- [Educational Technology Services Instructional Github](#)
- [Blink Documentation](#)
- [Datahub Grading Tools](#)

Jupyter notebooks review

- Installing anaconda
- <https://github.com/COGS108/Tutorials>
- <https://github.com/NeuralDataScience/Tutorials>
- Correcting common issues
- Up to students to correct and resubmit so grading can be timely



The screenshot shows the Jupyter Notebook interface. At the top, it says "jupyter Untitled (unsaved changes)" with the Python logo and a "Logout" button. Below this is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". To the right of the menu bar, it says "Trusted" and "Python 3 (ipykernel)". Below the menu bar is a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, and running code. The main area contains two code cells. The first cell has the following code:

```
In [1]: 1 #Hello World!
        2 print("Neural Data Science is Awesome!")
```

The output of this cell is "Neural Data Science is Awesome!". The second cell is empty and has the following code:

```
In [ ]: 1 |
```

We asked: What is a program?

- Generally a **program** is a **set of instructions** the programmer defines for a device or entity (usually a computer but not always) to follow
- Regarding computers-> programmer writes a set of instructions (“program”) that tells the computer to perform a set of operations
- We also discussed python, jupyter, when to use and not use jupyter and python

How do you write a program in Jupyter notebooks and python?

- datahub.ucsd.edu
- or your machine with anaconda
- The notebooks we will review are listed below and available in the lectures directory of the github and linked from the website and will be on canvas as well
 - 00-Introduction.ipynb
 - 01-Python.ipynb
 - 02-JupyterNotebooks.ipynb
 - 01_01_python-checkpoint.ipynb

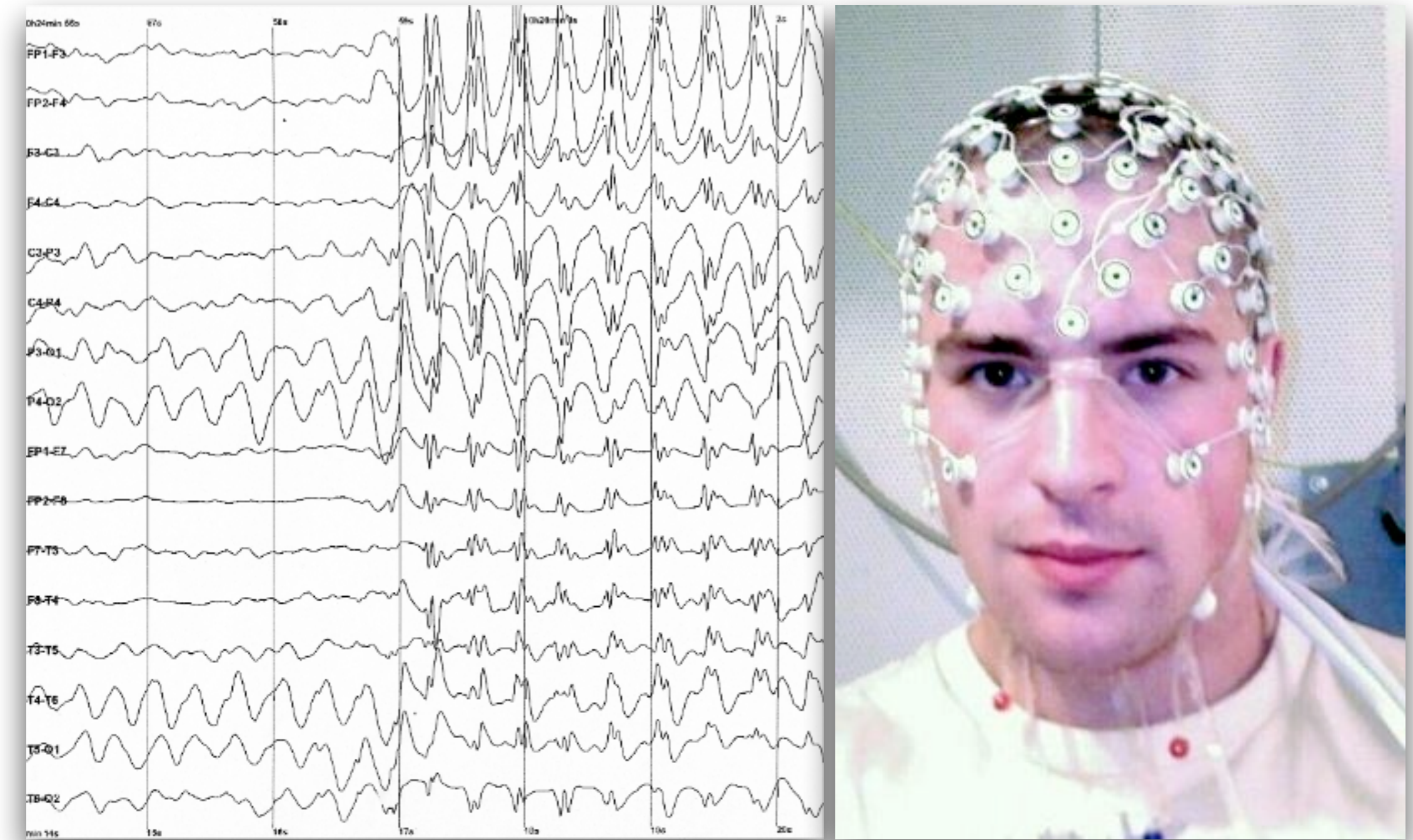
Neural data science toolsets

- There are a variety of toolsets employed in neural data science
- Assist in processing data types, e.g.
 - **EEG and MEG analysis**
 - MNE - <https://mne.tools/stable/index.html>
 - **Linguistics**
 - NLTK- <https://www.nltk.org>
 - **Motion capture data** - kinematic/inverse kinematic and dynamic analysis

Why EEG and MEG analysis?

- **EEG - Electroencephalography**

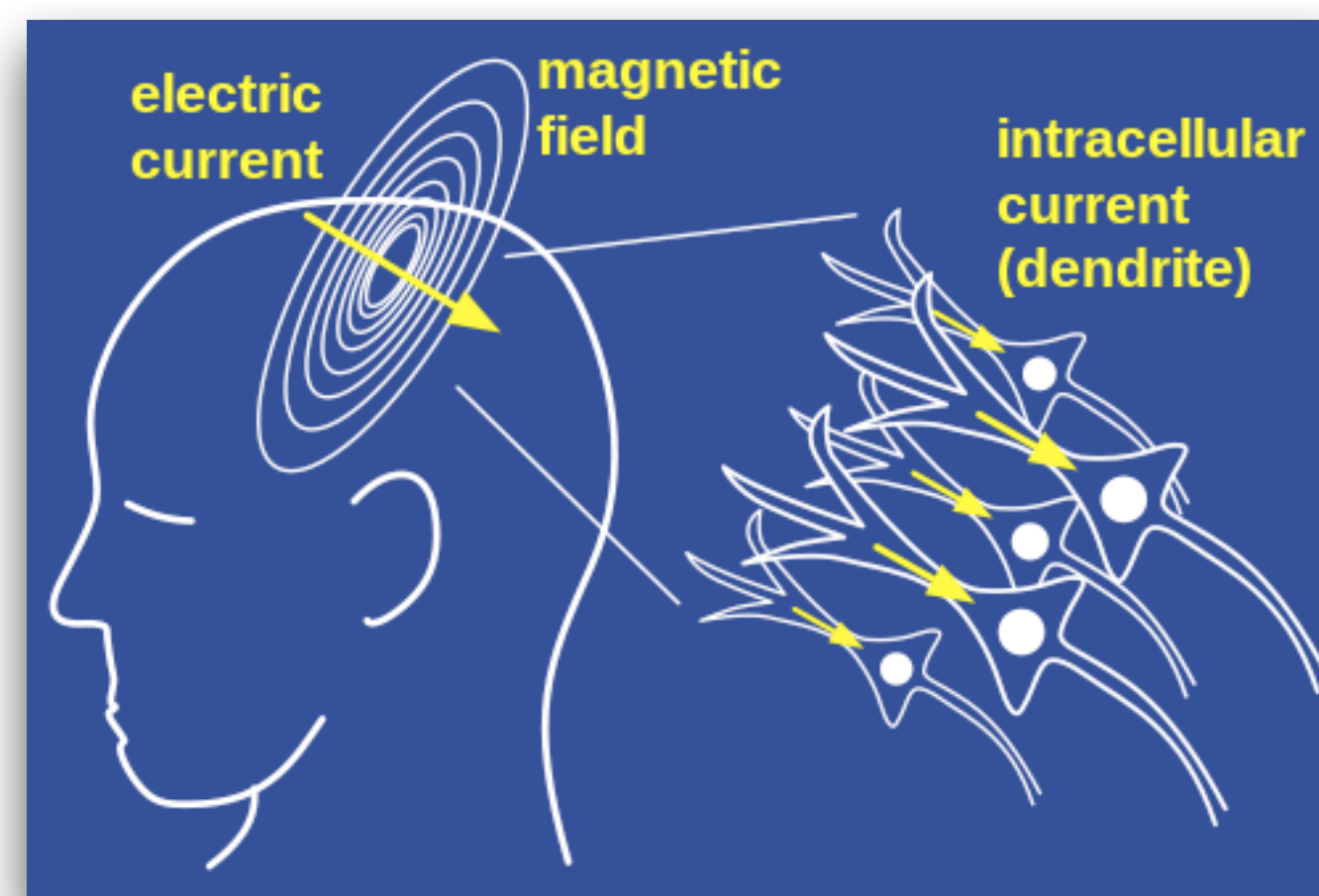
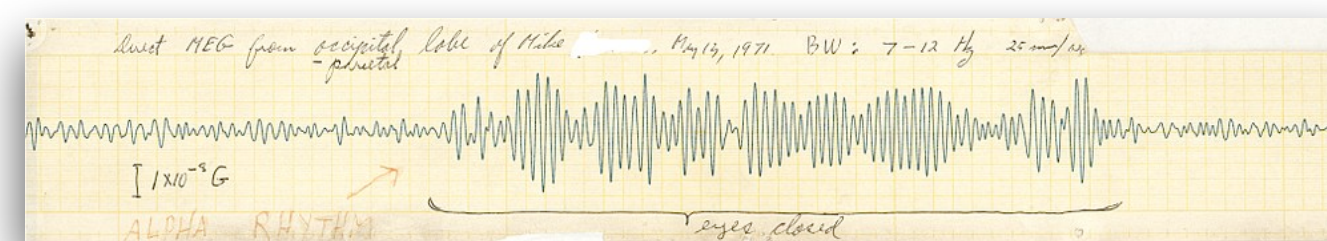
- Standard location patterns of sensors for recording (20-10, 10-10 systems)
- EEG records electrical activity generated in your brain at the scalp
- Global types of signals such as decision processes, spelling, gross body movement, etc
- Why useful?



(Source: <https://en.wikipedia.org/wiki/Electroencephalography>)

MEG - Magnetoencephalography

- **MEG** - measurement of the magnetic field generated by electrical activity of neurons
- Mapped onto structural image from MRI
- **Advantages**
 - Provides a higher spatial (mm)/temporal (msec) resolution, no distortion through head
 - Decay relative to dist. is more pronounced than electrical fields thus useful for measuring superficial cortical activity
 - Shows absolute neuronal activity vs. fMRI shows relative activity (fMRI must always be compared to some reference neural activity)
 - Can be recorded for sleeping subjects, unconscious subjects other
 - Safe, no exposure to radiation/emf, noninvasive, easy to use



(Source: <https://en.wikipedia.org/wiki/Magnetoencephalography>)

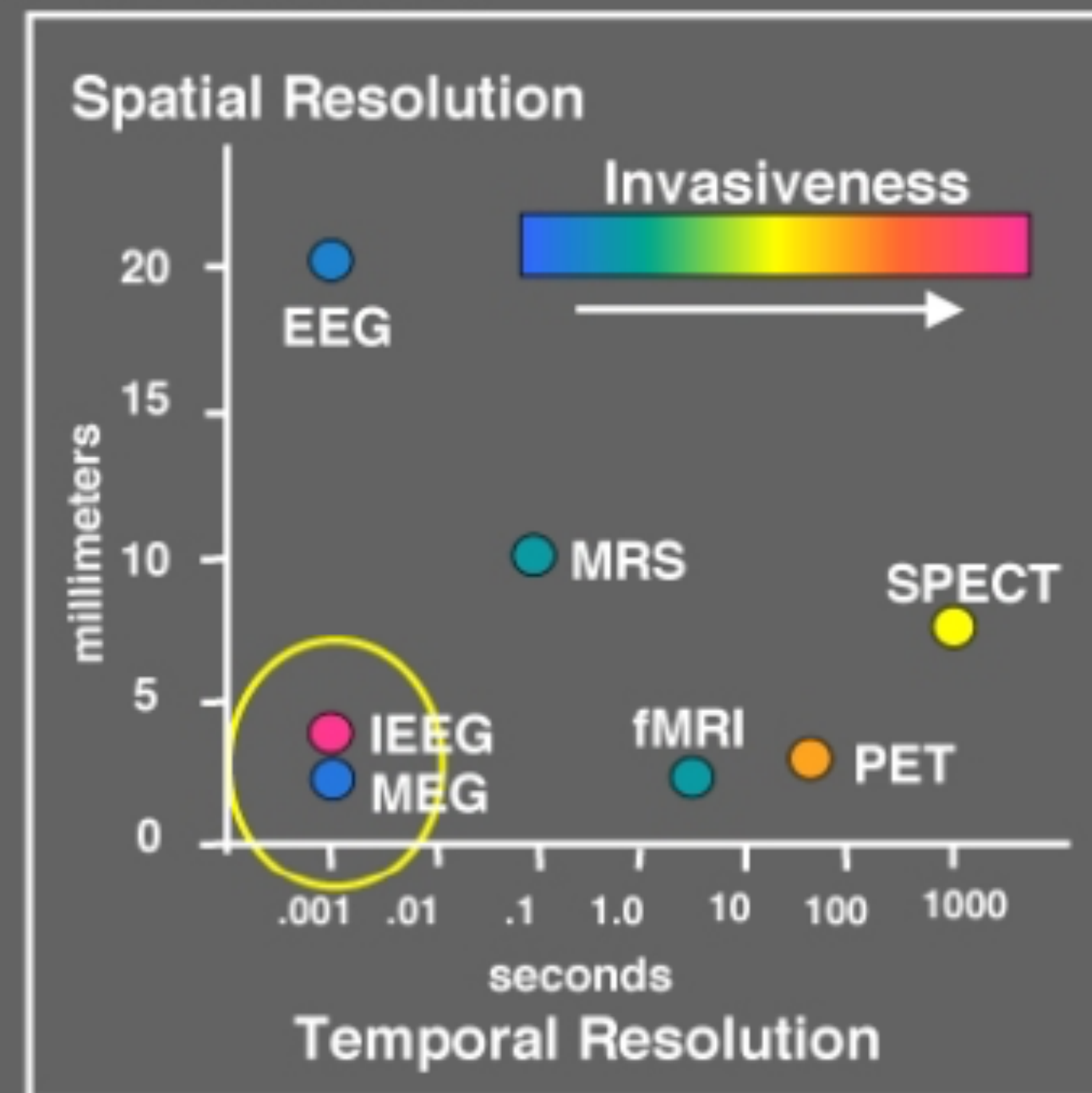
Properties and challenges

Problem of biomagnetism:

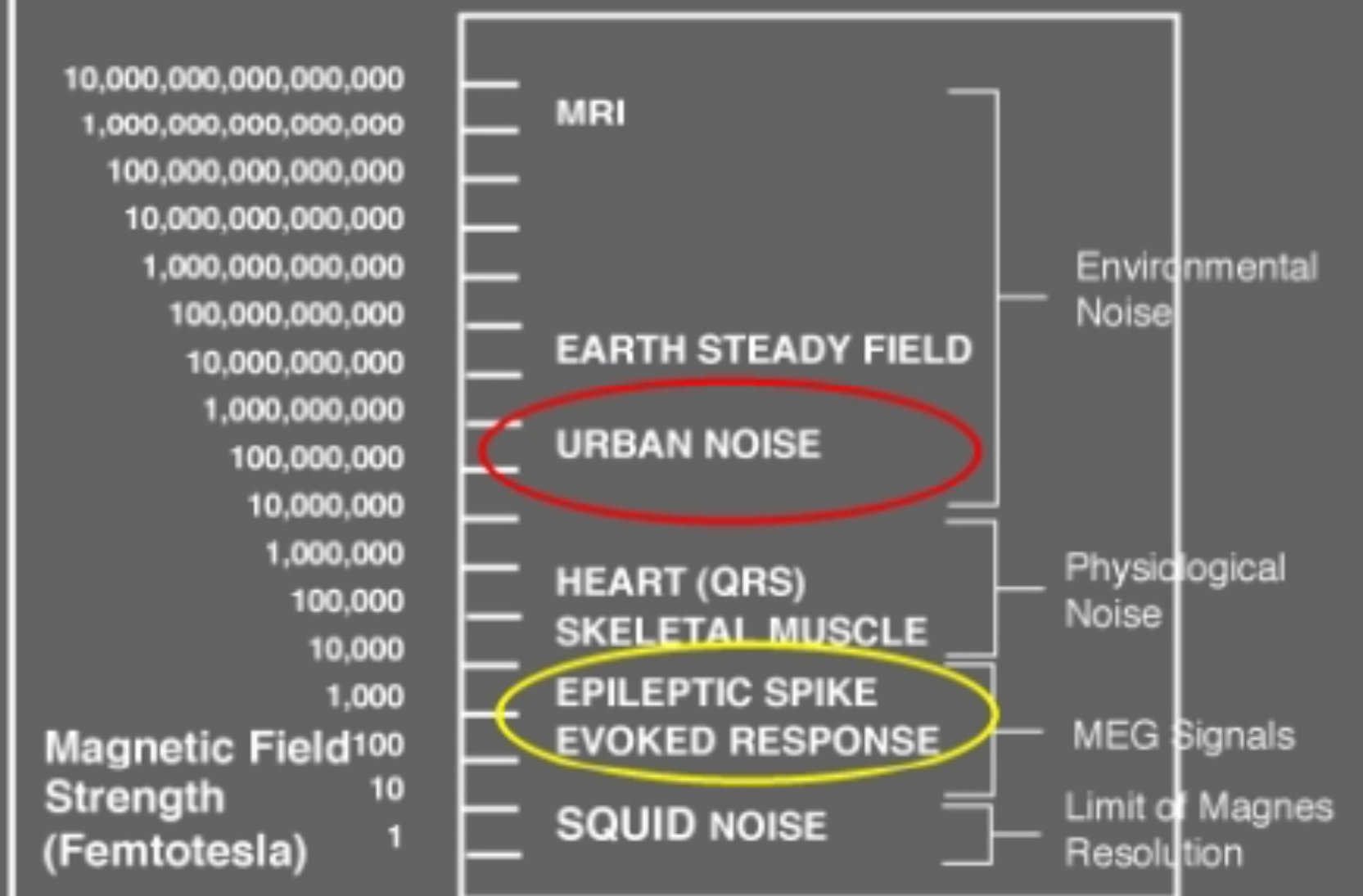
- The brain's magnetic field, measuring at 10 femtotesla (fT) for cortical activity and 103 fT for the human alpha rhythm
- Ambient magnetic noise in an urban environment, which is on the order of 108 fT or 0.1 μ T
- 50k Neurons for measurement
- Signals must be aligned \rightarrow pyramidal cells (perp. to cortical surface)

Properties of MEG

MEG Provides High Spatial and High Temporal Resolution



Strengths of Biological and Environmental Magnetic Fields



Introduction to MNE

- <https://mne.tools/stable/index.html>
- https://mne.tools/stable/auto_tutorials/index.html

Natural Language Processing

- **NLTK** - natural language toolkit (python)
 - <https://www.nltk.org/>
- Easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet
- Libraries for easily performing - classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries
- Documentation and discussion forums

TF-IDF

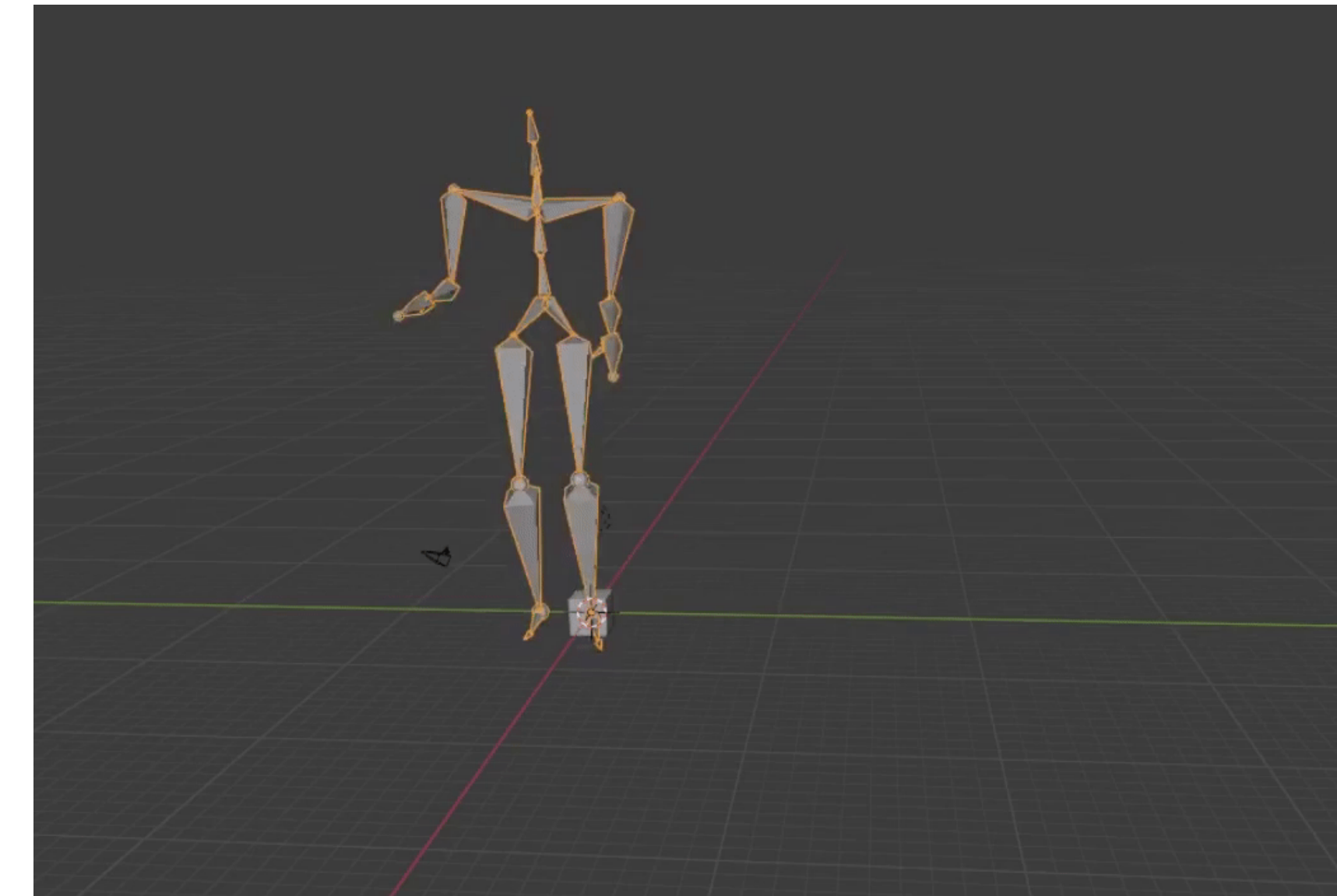
Term Frequency - Inverse Document Frequency

- Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents
- Calculating TF-IDF attempts to find the words that are important (i.e., common) in a text, but not *too* common

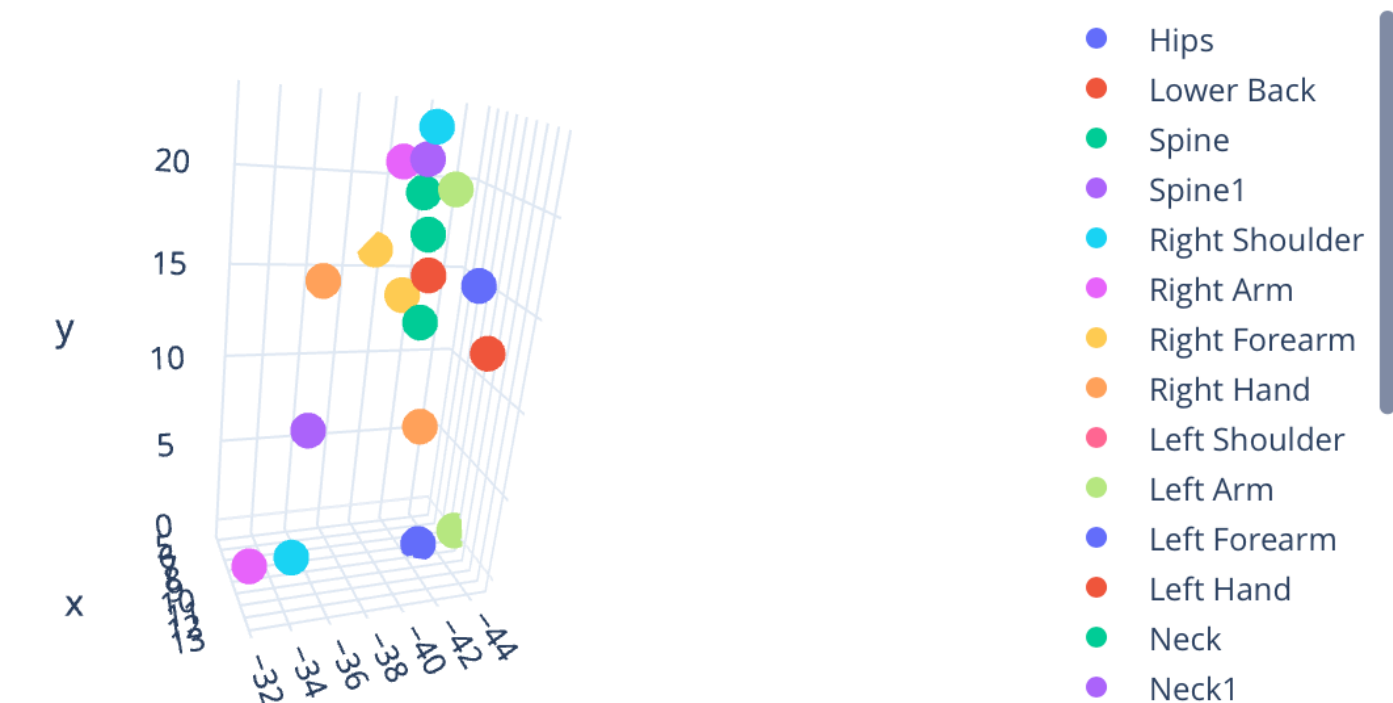
Motion capture and Eye Tracking

Motion capture data

- Recorded via
 - MoCap cameras - excellent, multiple types
 - Video - ok, issues
 - IMUs - ok, some disadvantages

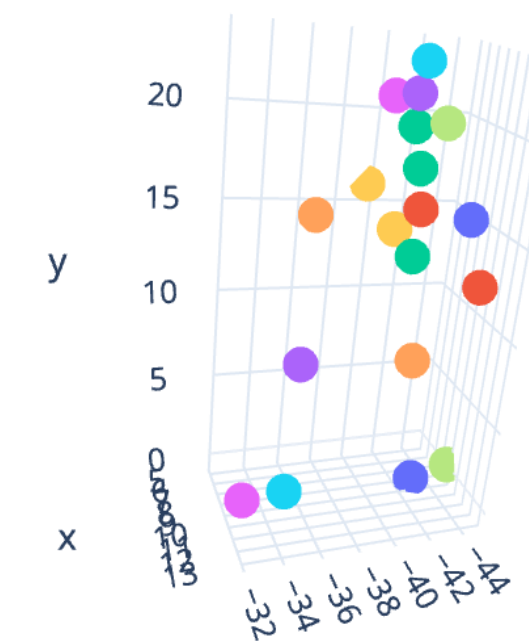


(Source: <https://medium.com/swlh/movement-classification-b98614084ec6>)



Motion capture data - challenges

- Hand manipulation involves many occlusions
 - Estimation
 - High camera density
 - Active markers
- Predictive estimation
- Marker occlusions generally, jumps and discontinuities, open/closed chain complexity
- Active systems require power, wires, may be delicate
- <https://www.engadget.com/2018-05-25-motion-capture-history-video-vicon-siren.html>



- Hips
- Lower Back
- Spine
- Spine1
- Right Shoulder
- Right Arm
- Right Forearm
- Right Hand
- Left Shoulder
- Left Arm
- Left Forearm
- Left Hand
- Neck
- Neck1



Motion capture systems

- Two main branches of tech:

- **Inertial** - IMUs track p/v/a (estimating p typically but can measure angle via gravity)

- Lower cost

- **Optical** - typically track markers, active or passive in IR to highlight marker positions relative to other data

- Higher cost

- Two main optical approaches

- **Active** systems

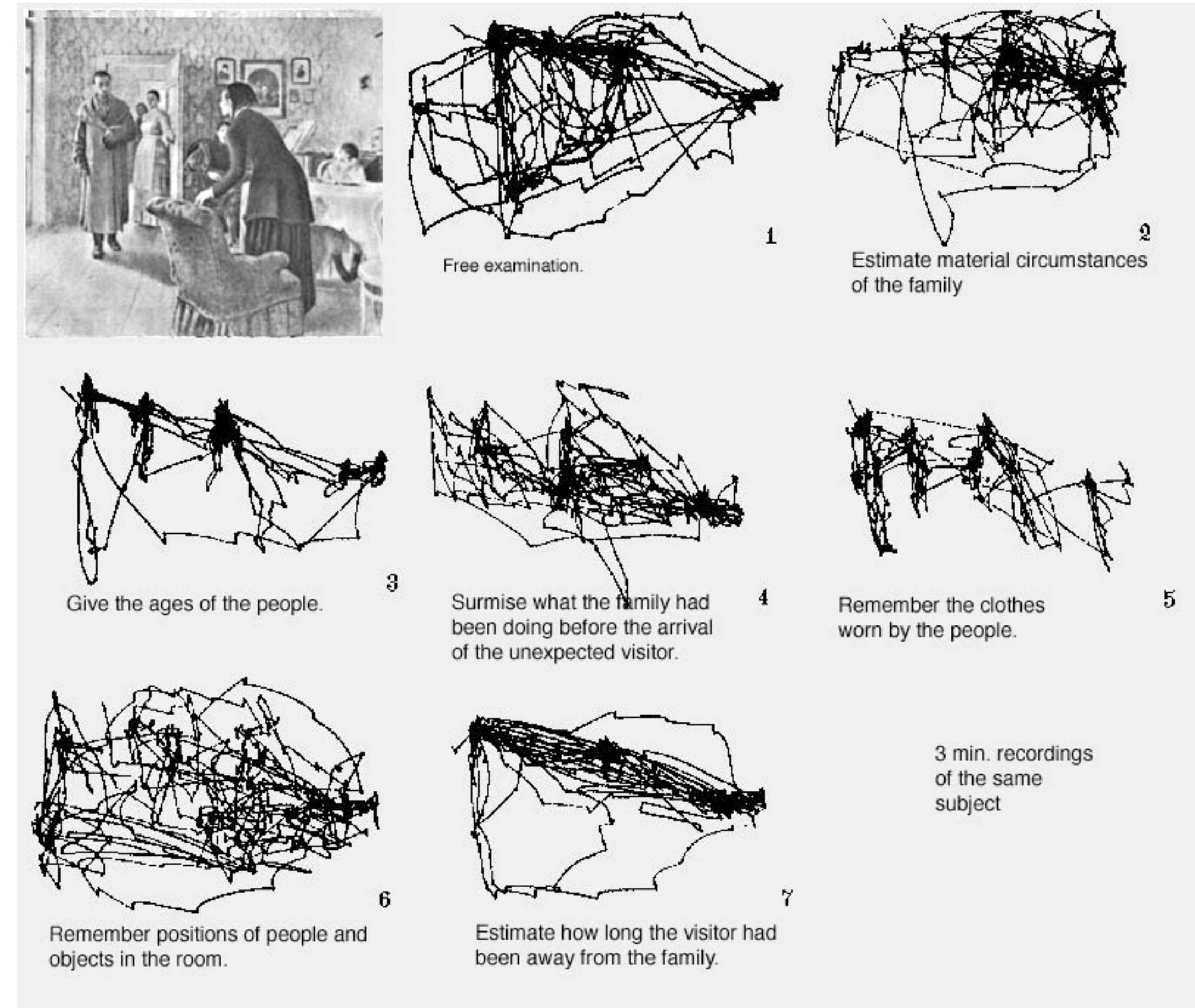
- **Passive** systems

- Combinations are possible



Eye tracking

- Human eye movements are complex and indicate many things about cognitive and neurological states as well as dynamics
- Yarbis (1967) - task given to a person affects eye movement
- “Unknown water balloon release time”
- Pencil stuck in the ceiling tile going to fall but when?
- Eye position, pupil dilation indications, focal point



Eye tracking - applications

- Cognitive loading
- Neurological diagnosis
- HCI
- Language reading
- Human factors/ergonomics
- Marketing research
- Operating interfaces without other means
- Safety, game theory, aviation, other assistive applications, augmented systems, engineering, automotive, etc



Eye tracking technology

- Eye trackers use one of the following to track retinal position and other bio-optic parameters
 - Cameras
 - Electrodes
 - Eye-attached technology (special contacts etc)
- Low speed vs. high speed
- Historically way back to 1800s by observation
 - Saccades

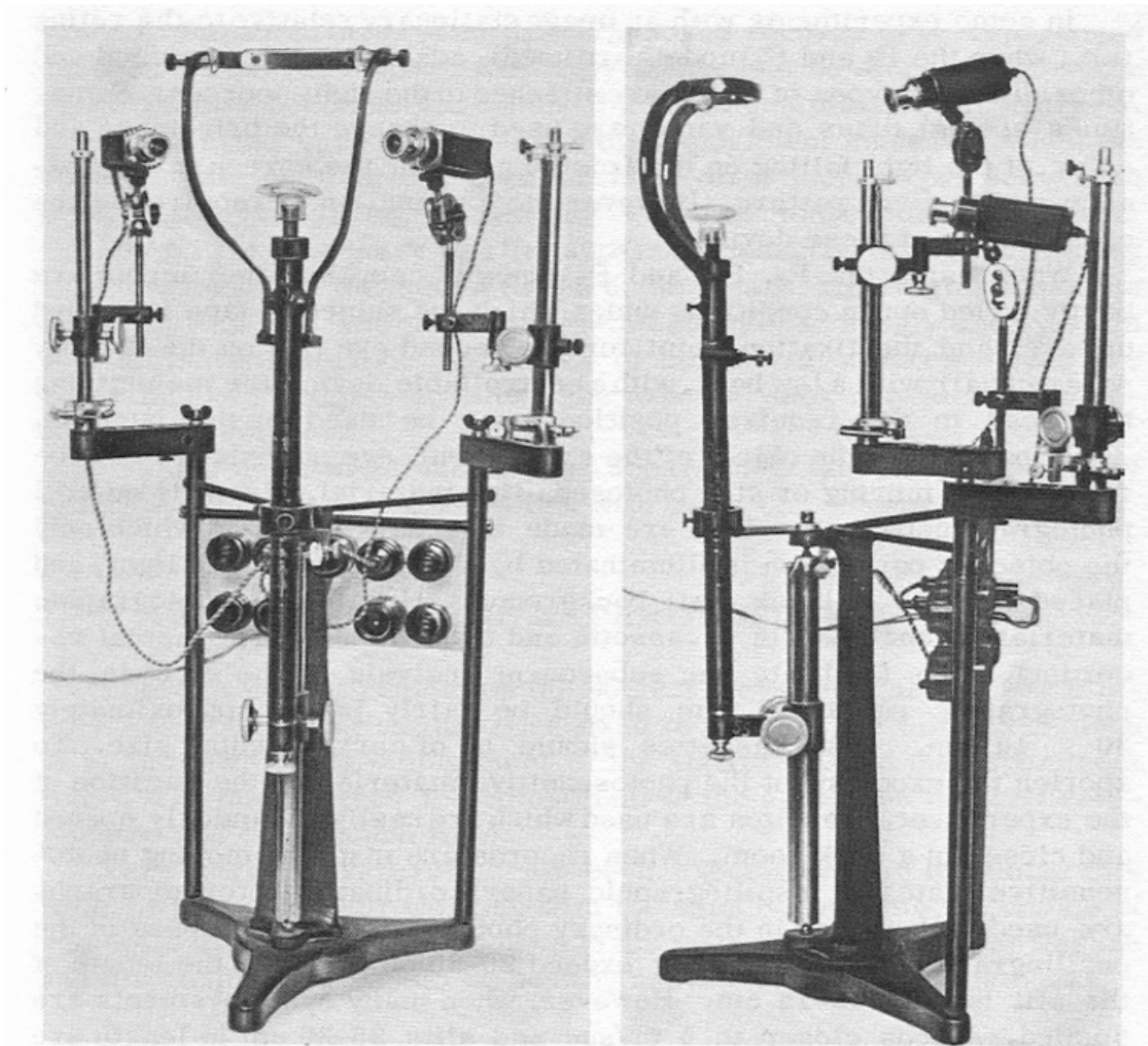


Fig. 21. The apparatus used in recording eye movements.



Genes and text, LISC

- Leveraging NLTK for research like gene expression studies
- Creating gene dictionaries
- Looking through literature to collect information about topics of interest, data and results using python (LISC)

Formulating Data Science Questions

When you and your group sit down to figure out what you're going to do for your final project in this class, you'll have to formulate a strong question - one that is specific, can be answered with data, and makes clear what exactly is being measured.

The Data Science Process

Ask an interesting question.

What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

Get the data.

How were the data sampled?
Which data are relevant?
Are there privacy issues?

Explore the data.

Plot the data.
Are there anomalies?
Are there patterns?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://www.cs109.org/>.

Hypothesis testing

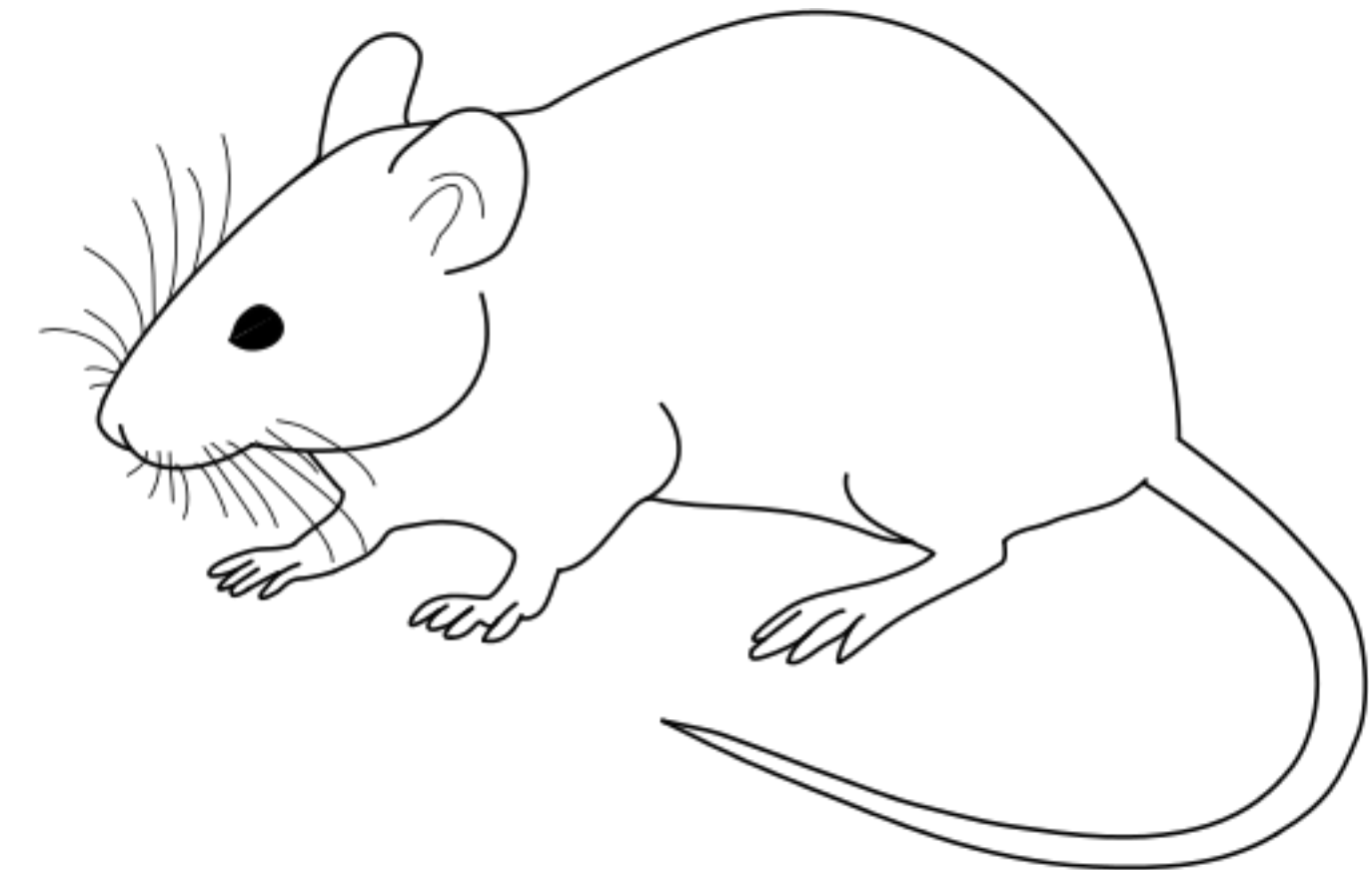
-Cannot prove hypothesis

-Can only reject or fail to reject null hypothesis

-Why?

Why Animal Models?

- We use **animal models** for gene expression because, unless a human is undergoing brain surgery where tissue can be sampled, **we cannot currently measure** gene expression in the brain otherwise
 - So to avoid harming a human (ethics are complicated!)
- Animals are found that have certain genomic similarities and assumptions are made about mapping behaviors, diseases and gene patterns into insights about humans
- Often an animal is bred for the study with specific genes or “knockouts” are created with certain genes removed in order to understand effects



(Source: https://en.wikipedia.org/wiki/Laboratory_mouse)

Why **Not** Animal Models?

- Ethical considerations
- Differences between animals and humans
- Time
- Cost
- Space, resources, pollution, energy use

Alternatives to animal models

- Simulation/computational modeling
- Artificial hardware systems/embodied systems
- Organoids
- Others?

F.A.I.R.

Findable **A**ccessible **I**nteroperable **R**eusable

Data

Science and reproducibility

- Understanding the brain requires broad, diverse and complex sets of data taken from many species of creatures, simulation, models and worldwide contributors
- The data must be findable, accessible, interoperable and reusable (FAIR)

Neurodata **W**ithout **B**orders (N.W.B.)

Introduction, tools, definitions and relevance

Use **NWB** for

- Use this for cellular neurophysiology, such as electrophysiology and optical physiology

NWB Introduction

- <https://www.nwb.org/>
- <https://nwb-overview.readthedocs.io/en/latest/>
- So essentially
 - A data format for sharing/archiving
 - Standardized (set of rules and best practices)
 - Packages Data and Metadata together so human- and machine-readable

Brain Imaging Data Structure (B.I.D.S.)

Introduction, tools, definitions and relevance

Use **BIDS** for

- Use for neuroimaging data such as MRI

Brain Imaging Data Structure

- <https://bids.neuroimaging.io/>
- A second data standard

Distributed Archives for
Neurophysiology Data Integration
(D.A.N.D.I.)

What is DANDI?

- The BRAIN Initiative archive for publishing and sharing neurophysiology data including
 - Electrophysiology, Optophysiology, Behavioral time-series, Images from immunostaining experiments.
- A persistent, versioned, and growing collection of standardized datasets
- A place to house data to collaborate across research sites
- Supported by the BRAIN Initiative and the AWS Public dataset programs

a) Web application

The screenshot shows the DANDI Archive website. At the top is a navigation bar with links for 'DANDI', 'WELCOME', 'PUBLIC DANDISETS', 'MY DANDISETS', 'ABOUT', 'DOCUMENTATION', and 'HELP'. A 'NEW DANDISET' button and a user count '11' are also visible. The main content area features a large brain graphic and the text 'The DANDI Archive' followed by a description: 'The BRAIN Initiative archive for publishing and sharing neurophysiology data including electrophysiology, optophysiology, and behavioral time-series, and images from immunostaining experiments.' Below this is a search bar with the placeholder text 'Search Datasets by name, description, identifier, or contributor name'. At the bottom, three statistics are displayed: '138 datasets', '311 users', and '157 TB total data size'.

b) Supported standards

This section displays logos for various standards and initiatives. At the top right is the AWS logo. Below it is the NIH The BRAIN Initiative logo. The central focus is the 'NEURODATA WITHOUT BORDERS' logo, which includes a network diagram of nodes and connections. Below this are the 'BIDS' (Brain Imaging Data Structure) logo, featuring a brain with vertical bars, and the 'NIDM' (Neuroinformatics Data Model) logo, featuring a brain with a network diagram.

c) Analysis platform



d) Python clients

Collaborator(s)

Lab Member(s)

(Meta)-Data Flow

JSON/JSON-LD, NWB, NIFTI, TIFF

User Interactions

Web Browser, Shell, API

GIT, GitHub and Version Control

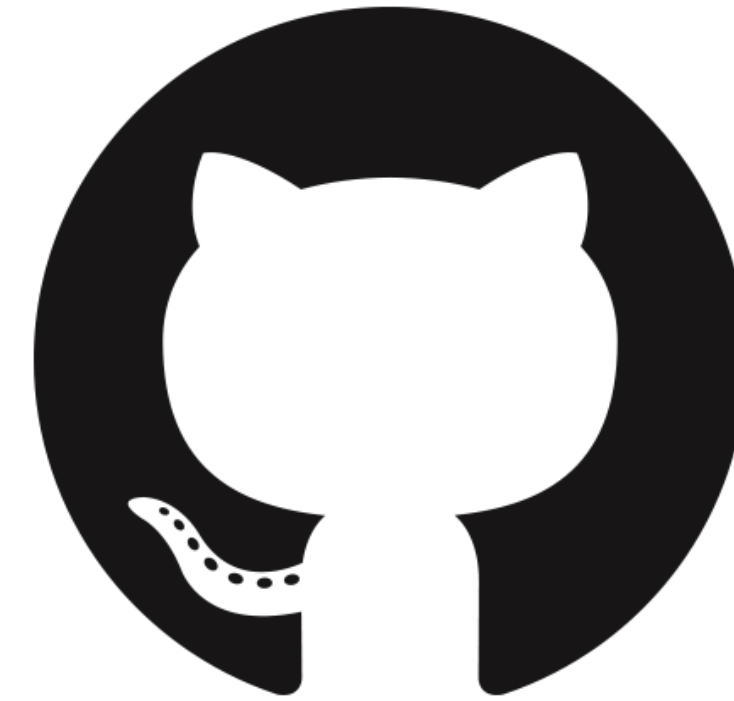
- Enables multiple people to simultaneously work on a single project.
- Each person edits their own copy of the files and chooses when to share those changes with the rest of the team.
- Thus, temporary or partial edits by one person do not interfere with another person's work

git & GitHub

git

the version control system

~ Track Changes
from Microsoft
Word....on
steroids



GitHub (or Bitbucket or
GitLab) is the home **where**
your git-based projects live
on the Internet.

~ Dropbox....but
way better

Data structures (Types, Tidy Data,
Data Intuition), Data Cleaning

Neural data and structures

- Neural data science generates and processes large amounts of data
- Data must be stored in some organized way for analysis - “Structure”
- There are three classes of data storage we will discuss - *structured, semi-structured, unstructured*

Data Structures Review

Structured data

- Can be stored in database SQL
- Tables with rows and columns
- Requires a relational key
- 5-10% of all data

Semi-structured data

- Doesn't reside in a relational database
- Has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

- Non-tabular data
- 80% of the world's data
- Images, text, audio, videos

Data Intuition



Variability in cleaning

- There is no one process to clean data
- Varies from set to set, project to project, software to software
- But can establish a 'template' procedure/process of 'check-offs' to make sure you've done your best to address it

Visualization of neural data

- https://mne.tools/stable/auto_tutorials/evoked/20_visualize_evoked.html
- https://mne.tools/stable/auto_tutorials/inverse/70_eeg_mri_coords.html#sphx-glr-auto-tutorials-inverse-70-eeg-mri-coords-py

Visualization

• **Tools:**

- seaborn - generating plots
- pandas - wrangling data
- matplotlib - fine-tuning plots

• **Plotting**

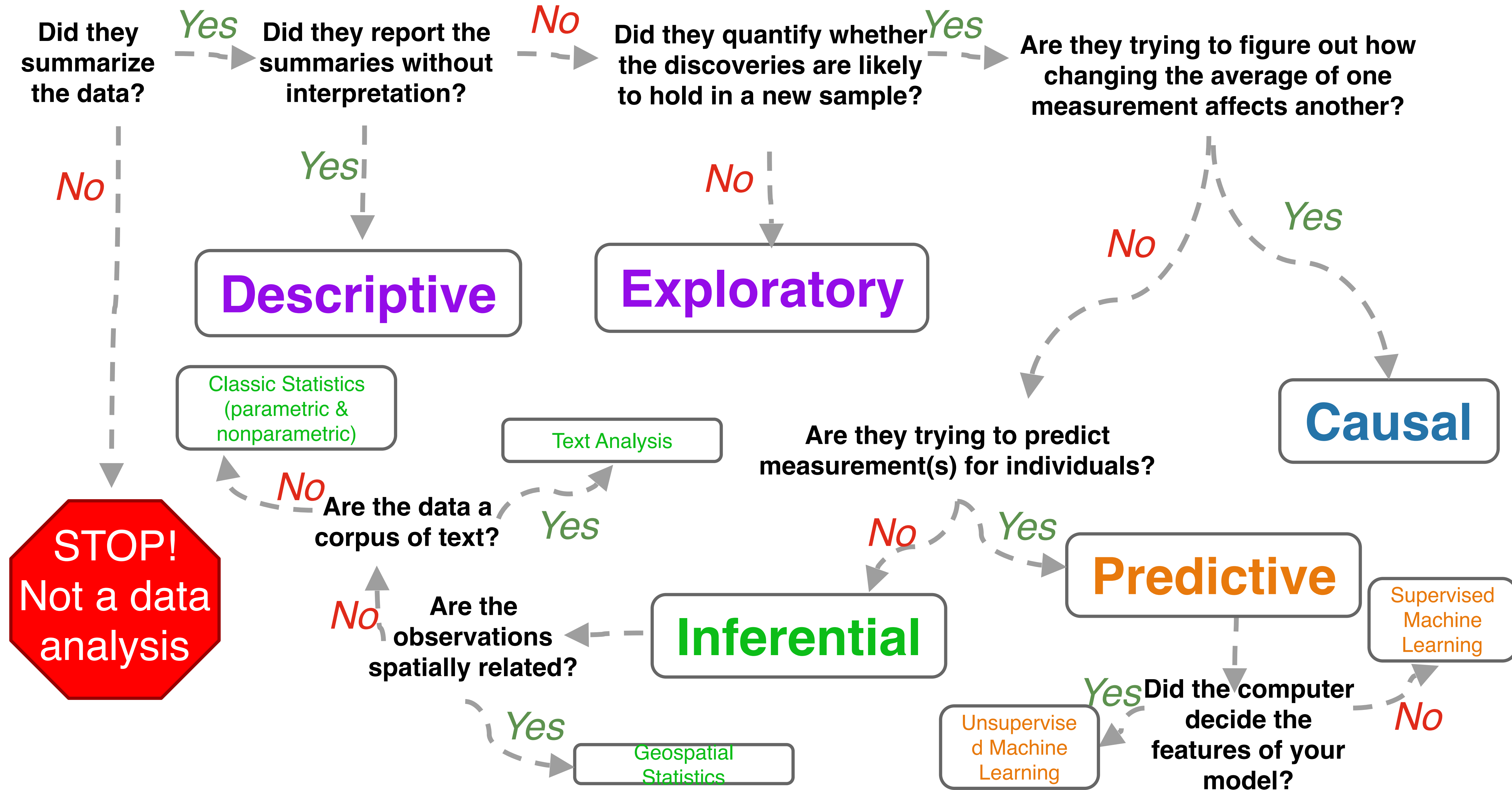
- quantitative data
- categorical data

• **Customizing visualizations**

How we went about addressing these questions

- Covered

- Intro to neural data science, questions, jupyter, python, pip
- Data sources, issues, advantages, caveats and modalities - EEG/MEG, Eye tracking, Behavior, Motion Capture (Phasespace, VICON, IMU, kinematics/dynamics), Text/Speech, gene expression, sequencing, animal models, animal model limitations, alternatives
- Tools for data science NLTK, MNE, Sentiment analysis, PyMo, mocaplib
- FAIR data, NWB, BIDS, DANDI



Statistical Data Analysis

- There are various definitions
- “Statistics” - the science of gathering data and discovering patterns
- “the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**” [[dictionary.com](https://www.dictionary.com)]

What are the 2 types of statistics?

- **Descriptive** - Summarizing the characteristics of data
- **Inferential** - Modeling, making 'inferences' from data

Descriptive statistics

- **Summarizing** the **characteristics** of data
 - Central tendency - (“center”) mean, median, mode
 - Variability - (“dispersion”) variance, standard deviation
 - Frequency distribution - (“occurrence within data”) counts
- Charts, plots, probability distribution shapes

Inferential statistics

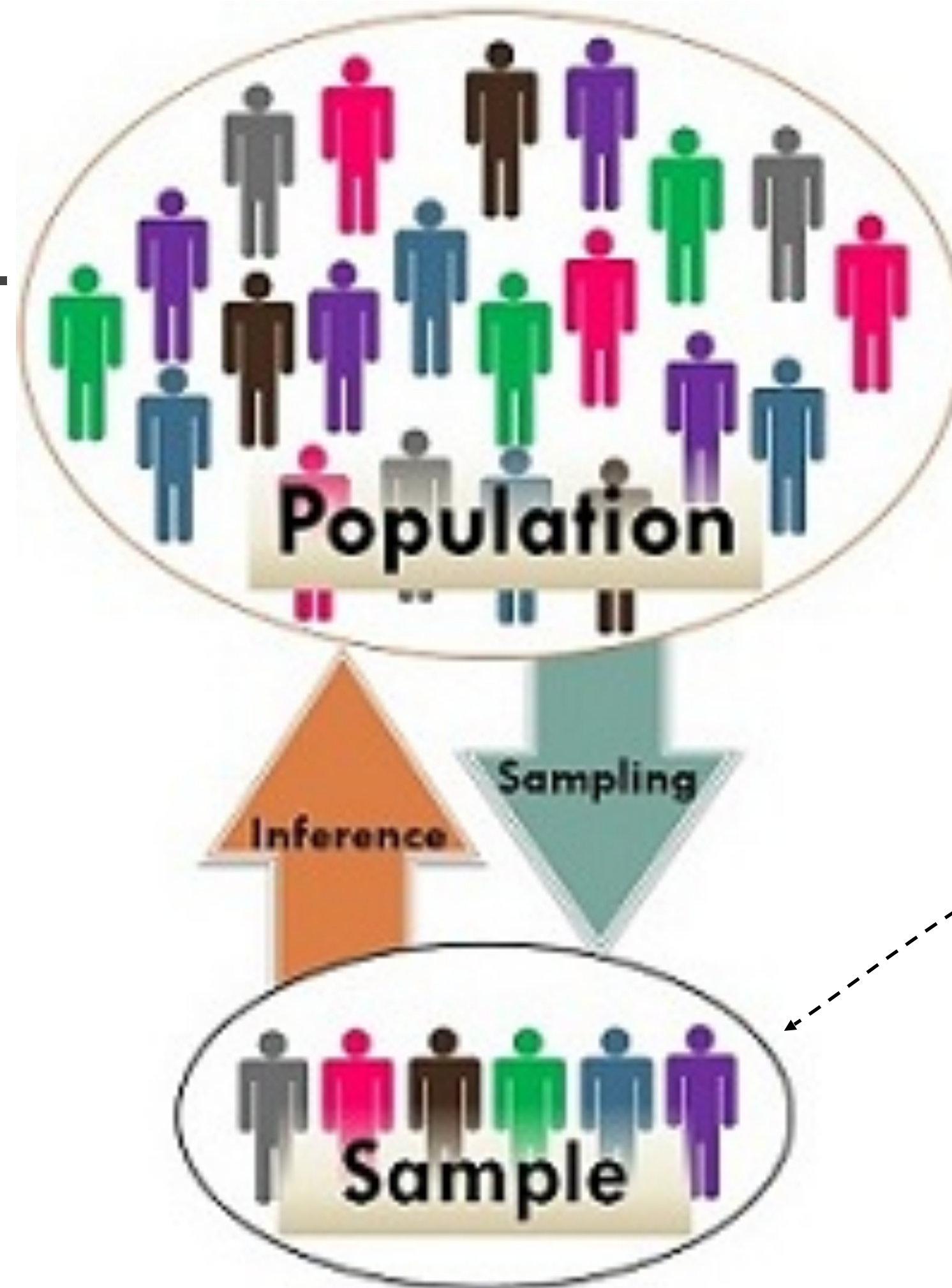
- “Modeling” or making ‘inferences’ from the data
- Taking data from samples and making predictions about populations
- 2 types
 - *Estimating parameters*
 - *Hypothesis tests*

Hypothesis testing

- Non-parametric data (no parameters)

Populations & Samples

We want to learn something about this..



Our population: *all* Neurons in the motor cortex

Our sample: LFP ~ 1-10k neurons

....but we can only *actually* collect data from this

The general principles of exploratory analysis:

- Look for missing values
- Look for outlier values
- Calculate numerical summaries
- Generate plots to explore relationships
- Use tables to explore relationships
- If necessary, transform variables

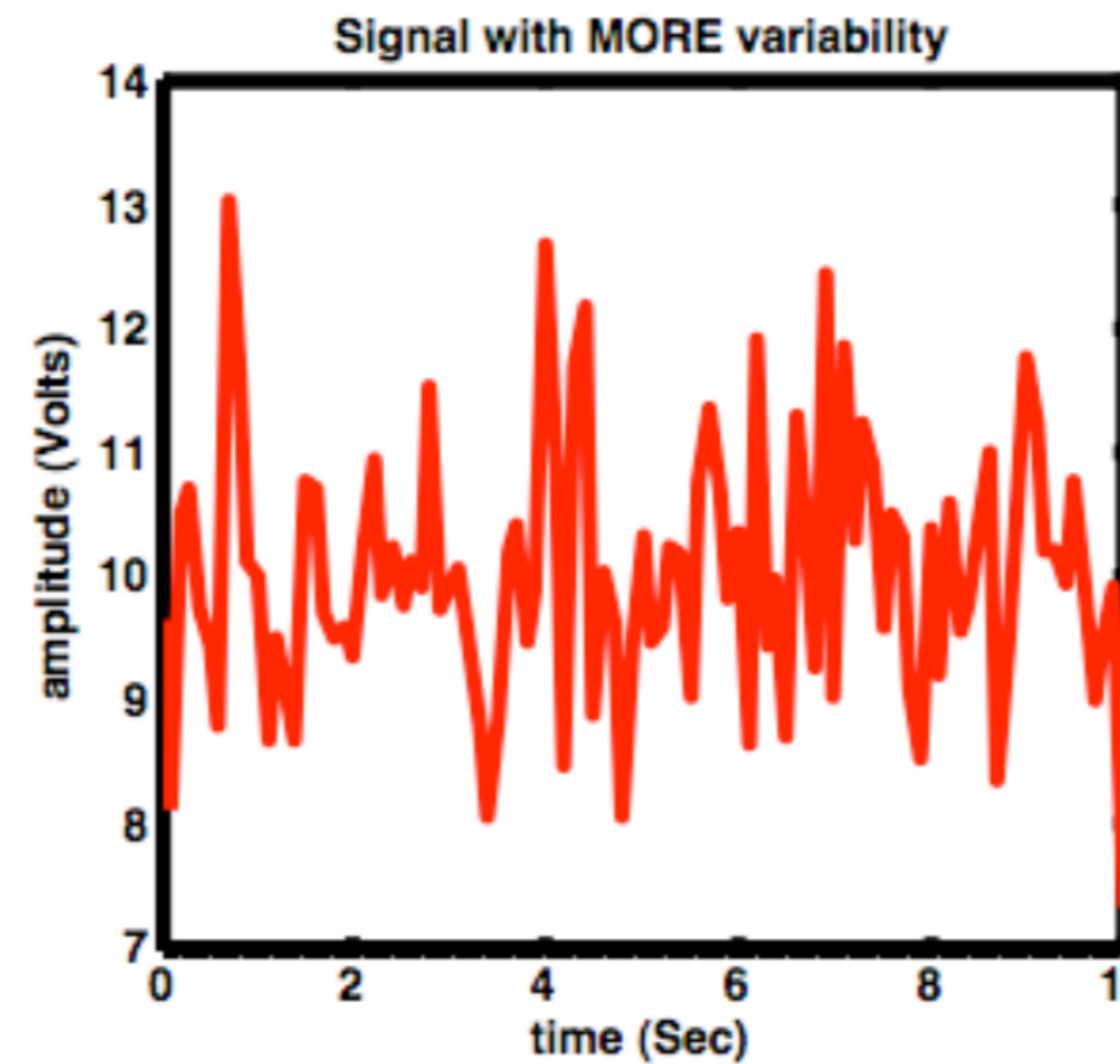
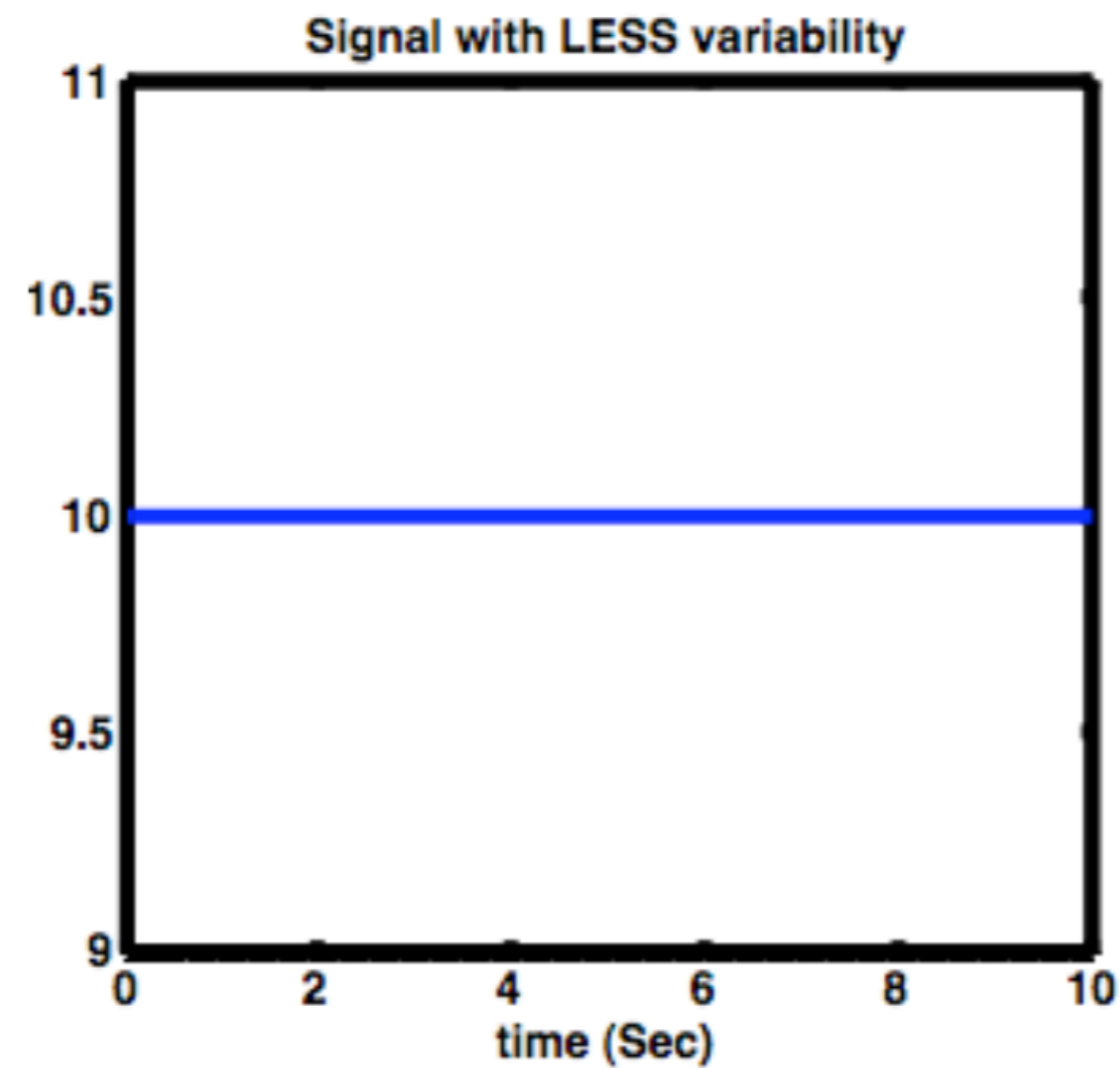
Correlations analysis, covariance and Time series analysis

PIP package manager

- [pip \(package manager\) - Wikipedia](#)
- Written in python
- Used to install, remove, manage software packages
- Connects to online package repository of public software (Python Package Index)
- Most python packages come with PIP installed
- Home page: [pip documentation v23.1.2 \(pypa.io\)](#)

Why we need a measure of variability

Same means, different variability of the signal

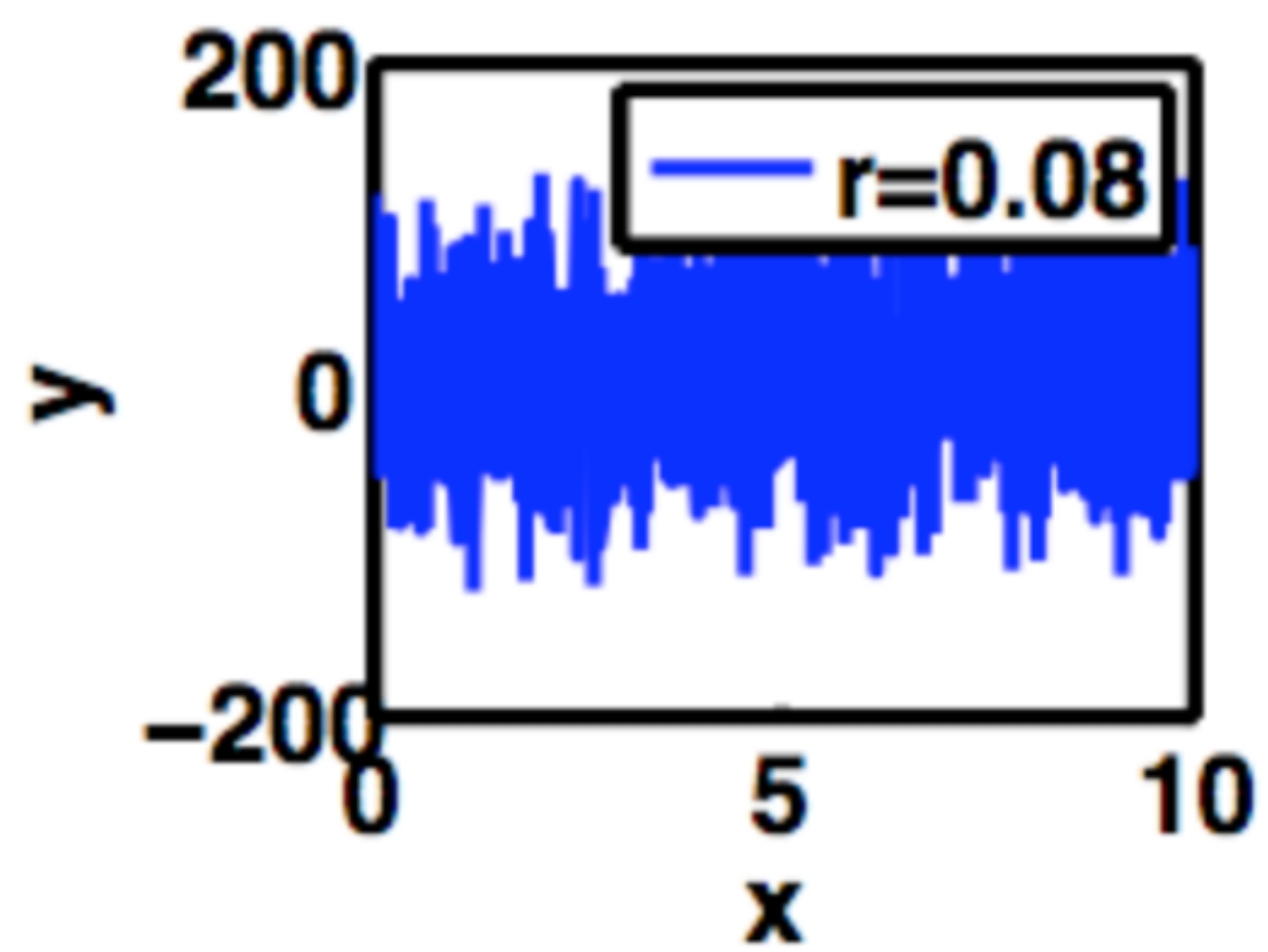
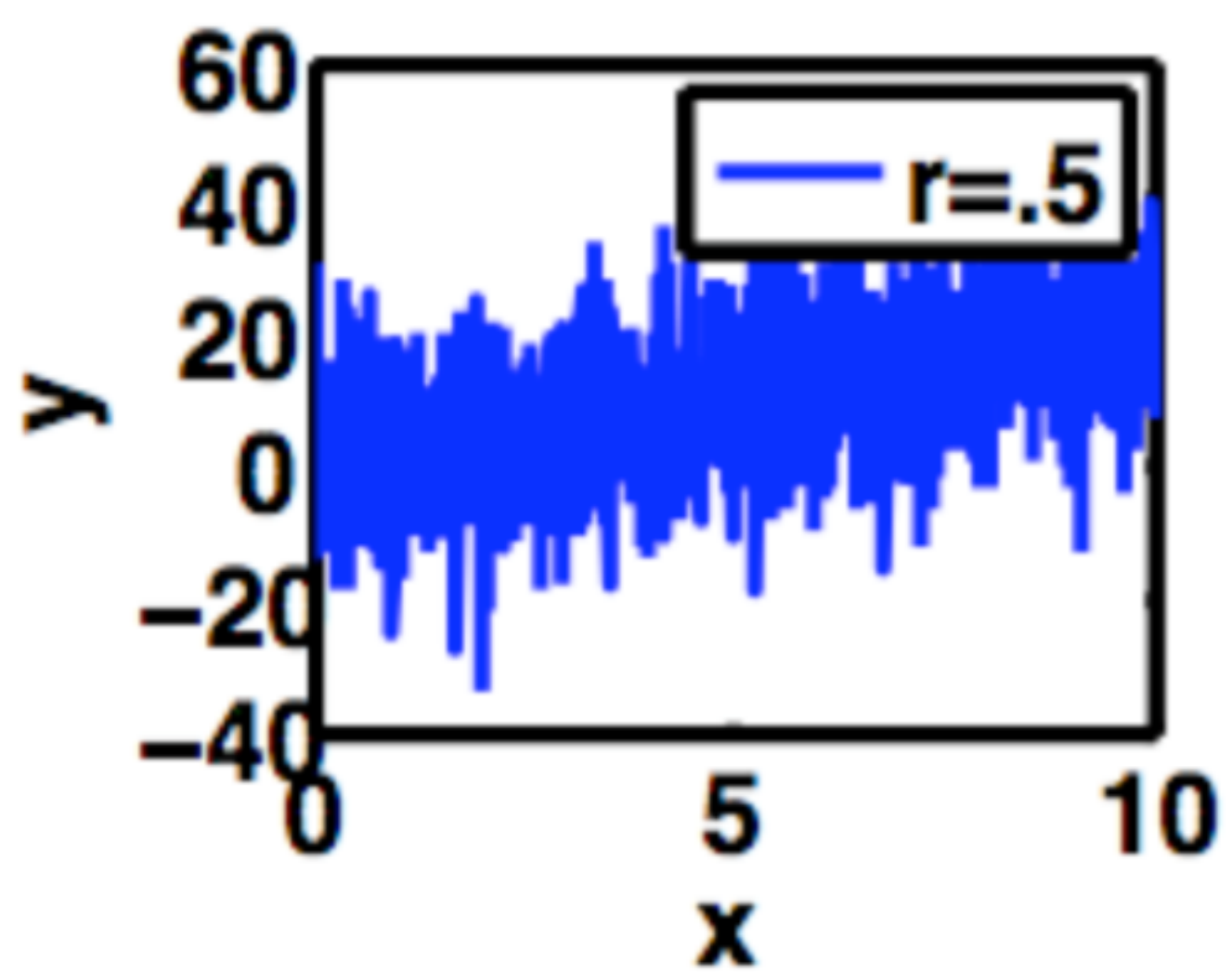
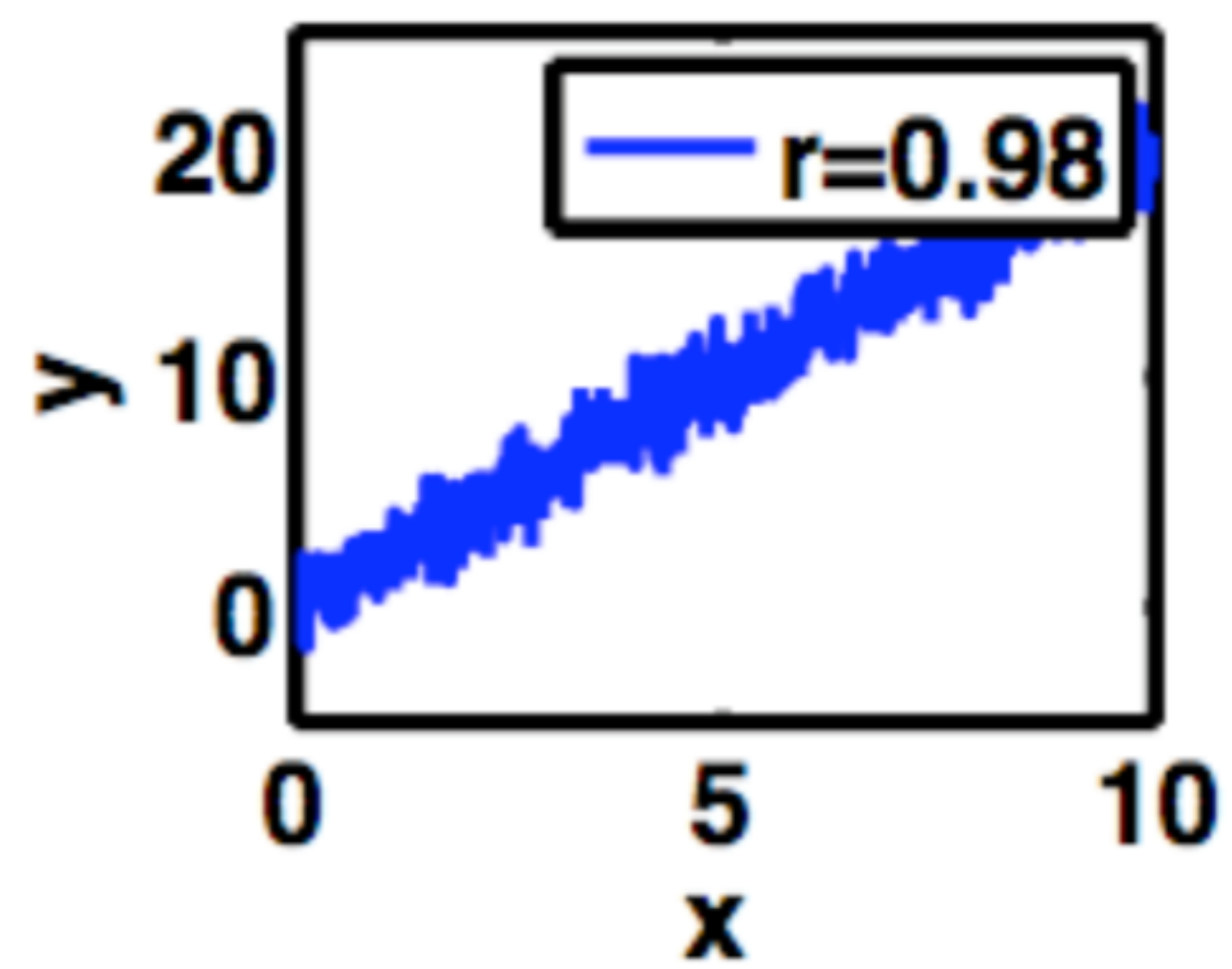
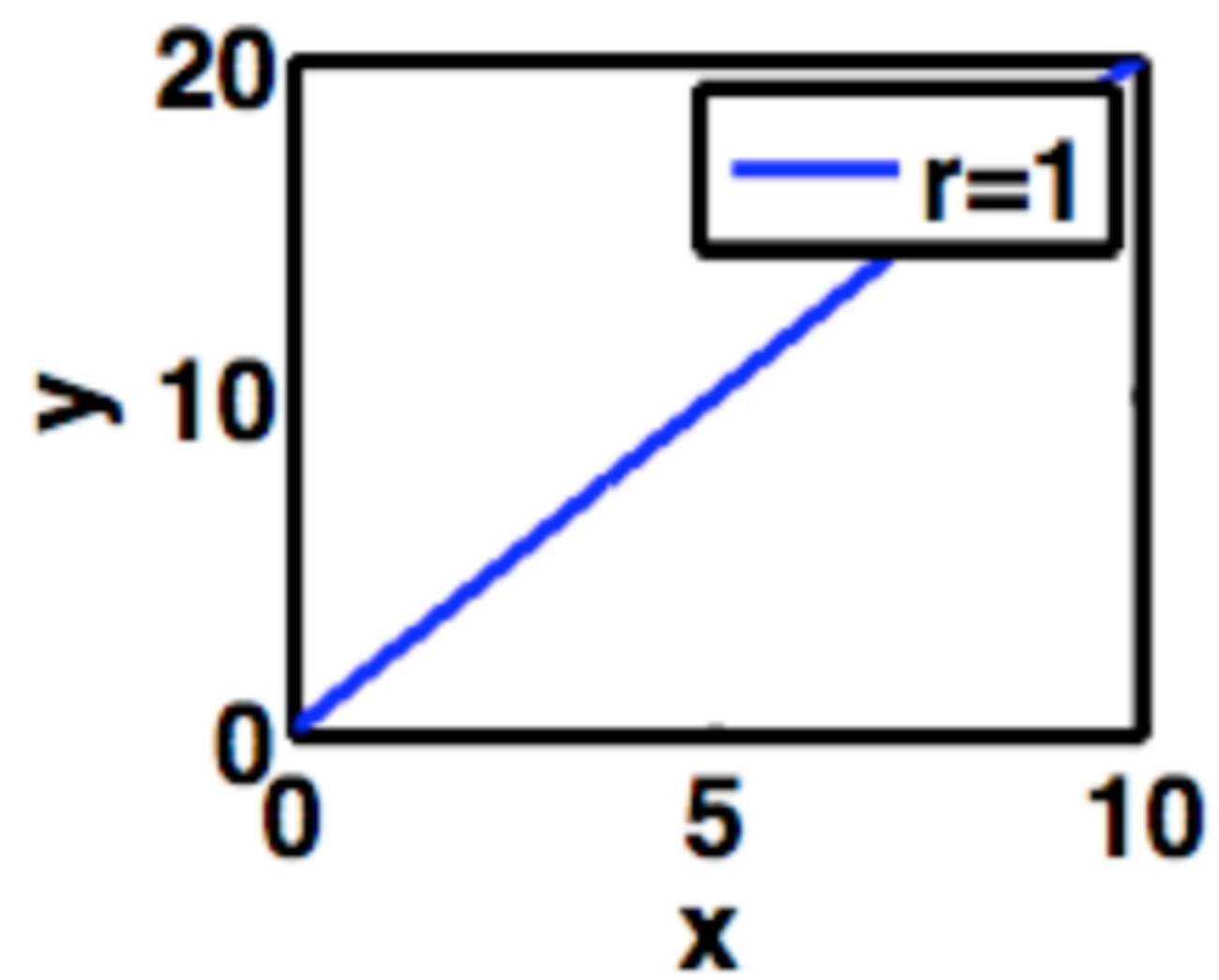


Correlation coefficient

$$\rho(j, k) = \frac{\sum_{i=1}^N Z_{ij} Z_{ik}}{N}$$

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$\rho(X, Y) = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y}$$



Math and symbol review

- http://casimpkinsjr.radiantdolphinspress.com/pages/cogs138_sp23/handouts/greek_letters_review.pdf
- http://casimpkinsjr.radiantdolphinspress.com/pages/cogs138_sp23/handouts/math_review.pdf
- Handouts page on website:
 - http://casimpkinsjr.radiantdolphinspress.com/pages/cogs138_sp23/handouts.html

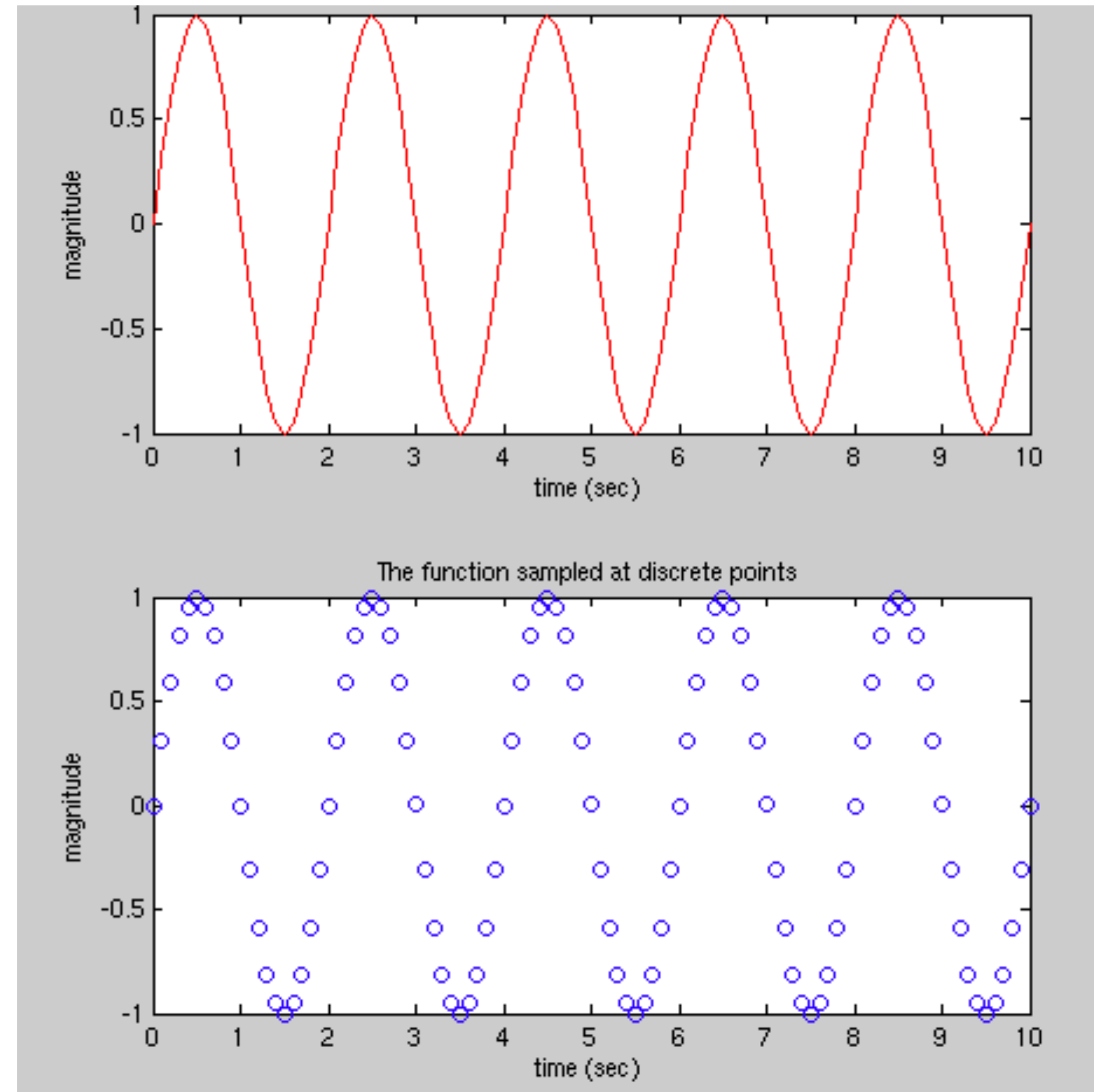
Correlations and pitfalls in neural data science

More on sampling, discretization, filtering

- Review of continuous vs. discrete quantities
- Analog vs. Digital
- Discretization, sampling, aliasing
- Filter theory, frequency response, filter types
- Linearity

Continuous vs. Discrete quantities

- Information storage
 - **Continuous** signals have information at every point in time
 - **Discrete** signals have info only at specified intervals (fixed or variable)



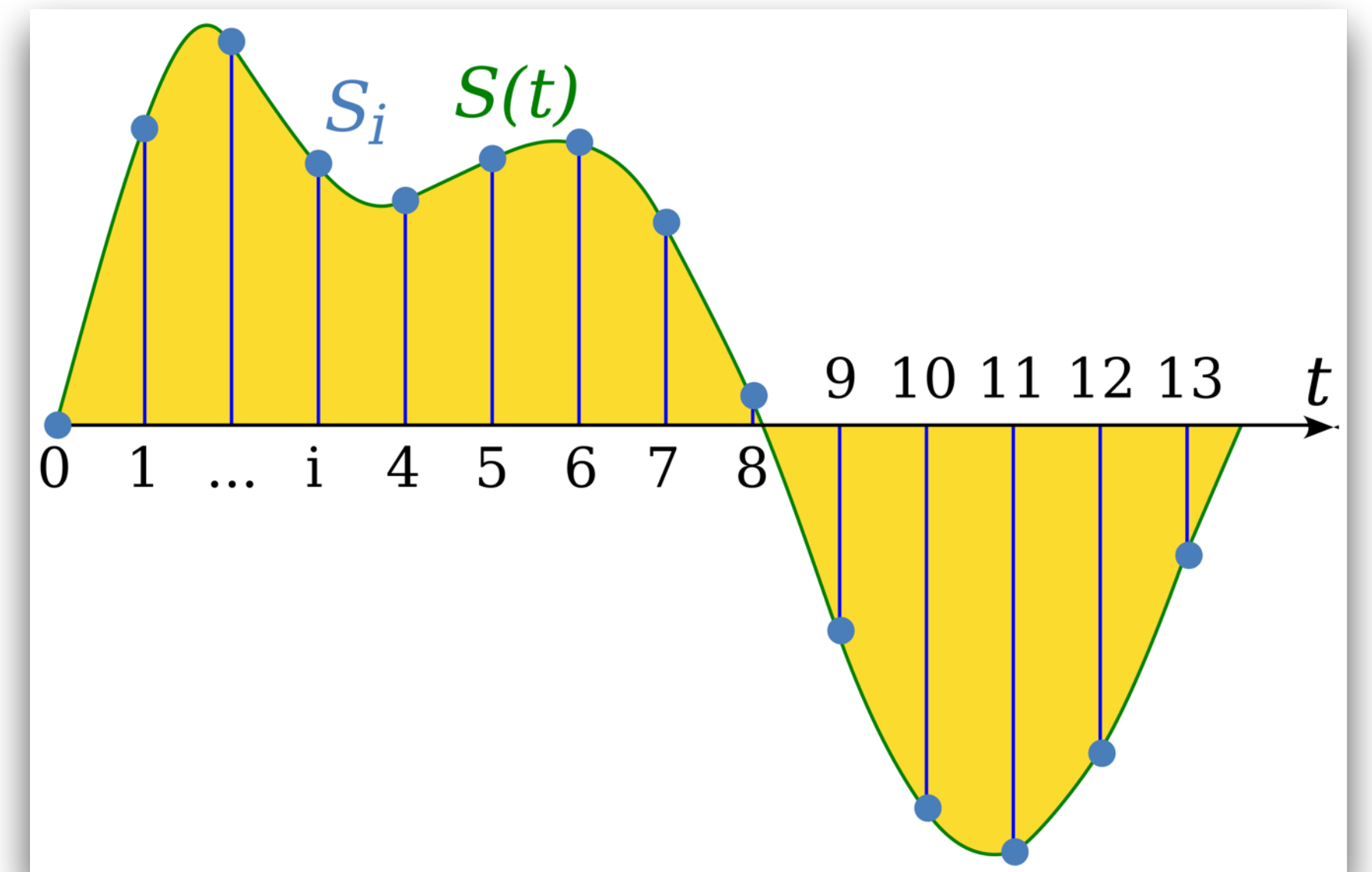
Discretization

- Measuring a continuous (analog) signal means capturing information at specified (fixed or variable) intervals
 - **Sampling frequency** - the frequency at which data is recorded from a signal (Typically in Hz, ie 5kHz)
- When capturing data, or when manipulating data which has been discretized, there are several issues to consider
 - Aliasing (not the TV show:)
 - Sampling rates
 - Post-processing – filtering data to remove unwanted information while retaining desired information

Sampling

- **Sample** - We record data at specific points in time
- **Period** - The time between samples, **T [sec]**
- **Sample frequency** - The frequency of sampling, f [Hz]

$$f = \frac{1}{\Delta T}$$



Computational filtering

- *Noisy auditory data can be filtered to remove undesired signals*
- *EEG signals can be filtered to remove 60Hz noise from AC lines nearby*
- *Other sensor signals can be filtered to improve results*

Frequency Response

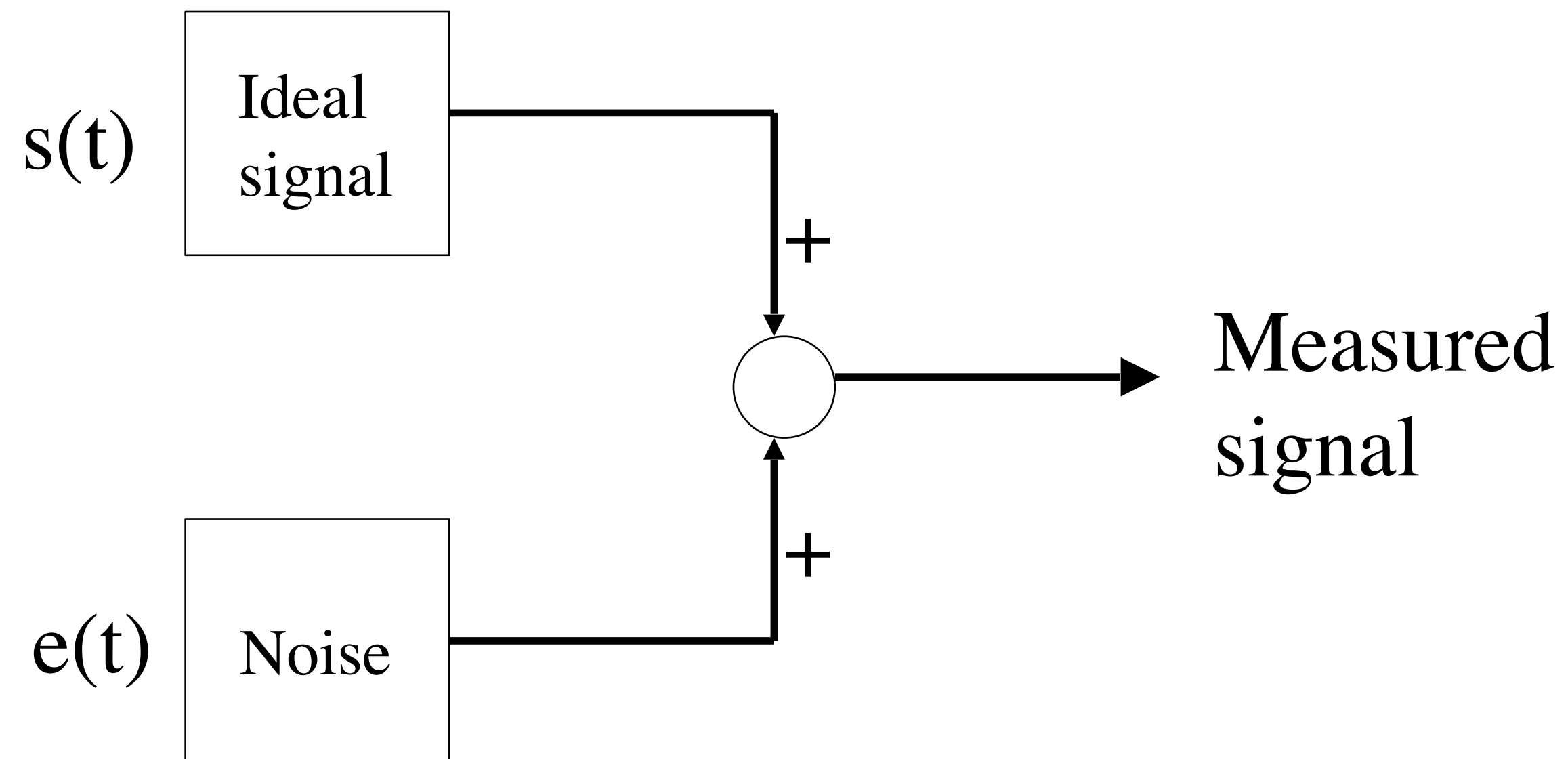
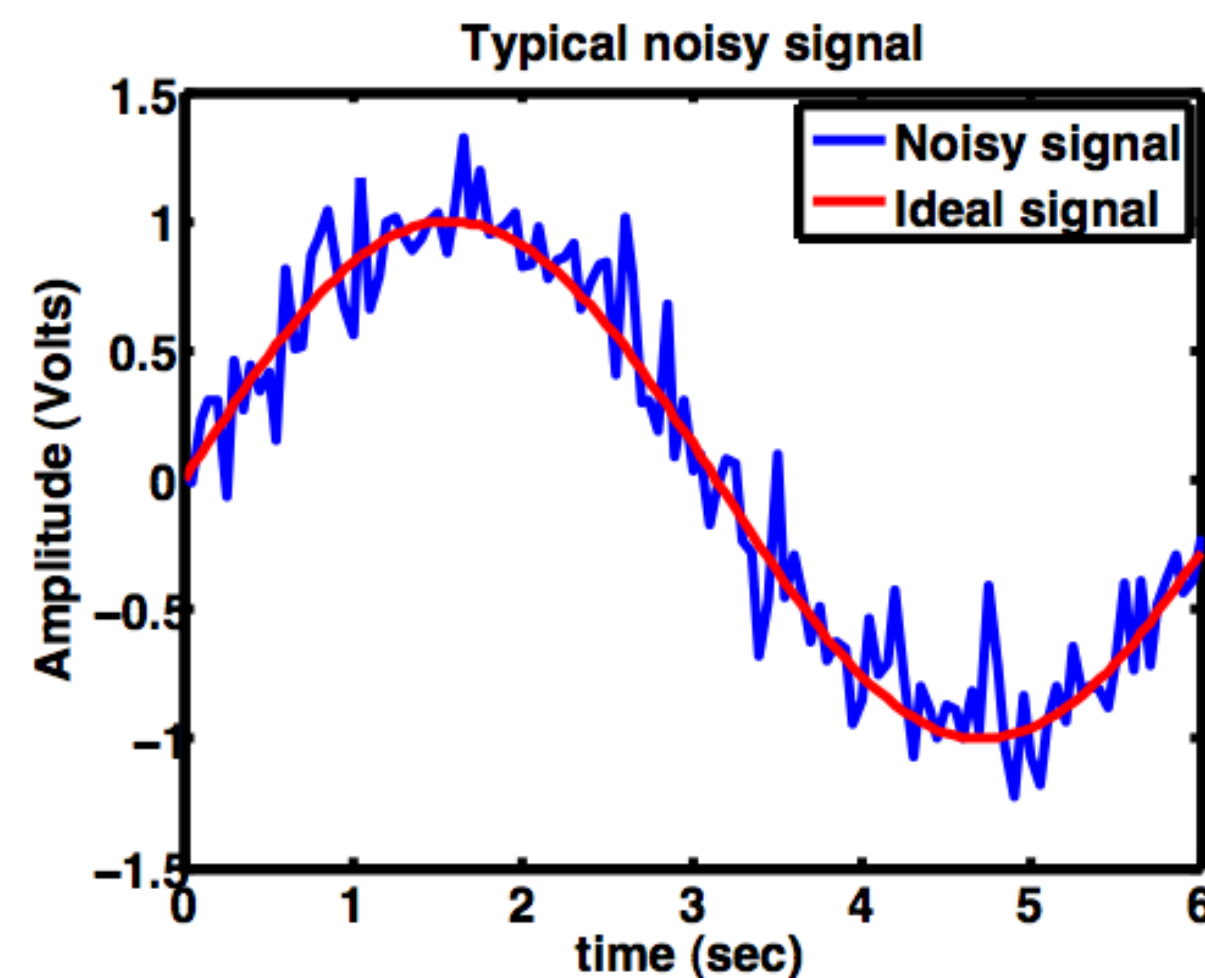
- Linearity of systems vs. nonlinearity
- The response of a linear system to a sinusoidal input is a sinusoidal output with the amplitude and phase shifted in some way
- This is useful for characterizing the behavior of some signal over a range of possible input frequencies
- Example with the chalk

Common filter types in signal processing

- **Low-pass filter** - (ideal) attenuates high frequency data, while allowing low frequency data to pass unchanged
- **High-pass filter** - (ideal) attenuates low frequency data, while allowing high frequency data to pass unchanged
- **Band-pass filter** - (ideal) attenuates all frequencies except a particular frequency band (or bands)
- **Band-stop filter** - (ideal) attenuates one or a selection of frequency ranges of data, allowing all the rest to pass unchanged
- Actual filters are not exactly ideal...which we will discuss

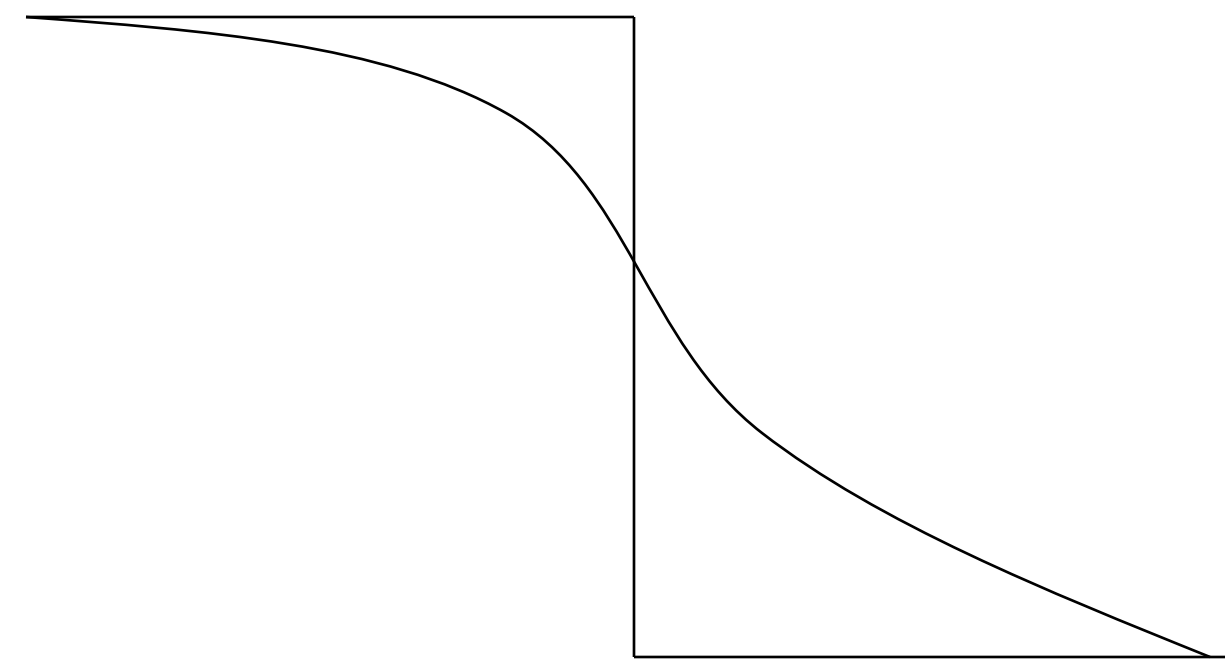
Signals and noise...

- By making assumptions about the properties of the unwanted 'noise' $e(t)$, we can reconstruct an appropriate *estimate* of the original signal $s(t)$
 - **Noise** - any unwanted portion of a signal, lumped together. It may come from multiple sources but tends toward some statistically predictable properties



Disadvantages...

- Need to have all data in memory already, so it isn't an 'online' filter
- Causality
 - **If we care about an exact event timing, this is a poor filter to use:**



Signal anticipates
changes!

Physicalist perspective

- Neuroscience perspective
- Other perspectives
- Which is 'true?'
- Does it matter?
- Animal model assumption of mapping

Underlying dynamics need to be exposed

- Tacoma narrows bridge disaster
 - 1st order vs. higher order
 - [https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_\(1940\)](https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_(1940))



[https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_\(1940\)#/media/File:Opening_day_of_the_Tacoma_Narrows_Bridge,_Tacoma,_Washington.jpg](https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_(1940)#/media/File:Opening_day_of_the_Tacoma_Narrows_Bridge,_Tacoma,_Washington.jpg)

How do we then learn about unknown dynamics?

- Learn by **experience, experimentation, hypothesis generation, data science!**
- “The Tacoma Narrows bridge failure has given us invaluable information ... It has shown [that] every new structure [that] projects into new fields of magnitude involves new problems for the solution of which neither theory nor practical experience furnish an adequate guide. It is then that we must rely largely on judgment and if, as a result, errors, or failures occur, we must accept them as a price for human progress.” [Othmar Ammann]
- Following the incident, engineers took extra caution to incorporate aerodynamics into their designs, and wind tunnel testing of designs was eventually made mandatory.

Motivation and warnings for the use of neural data science to answer big questions

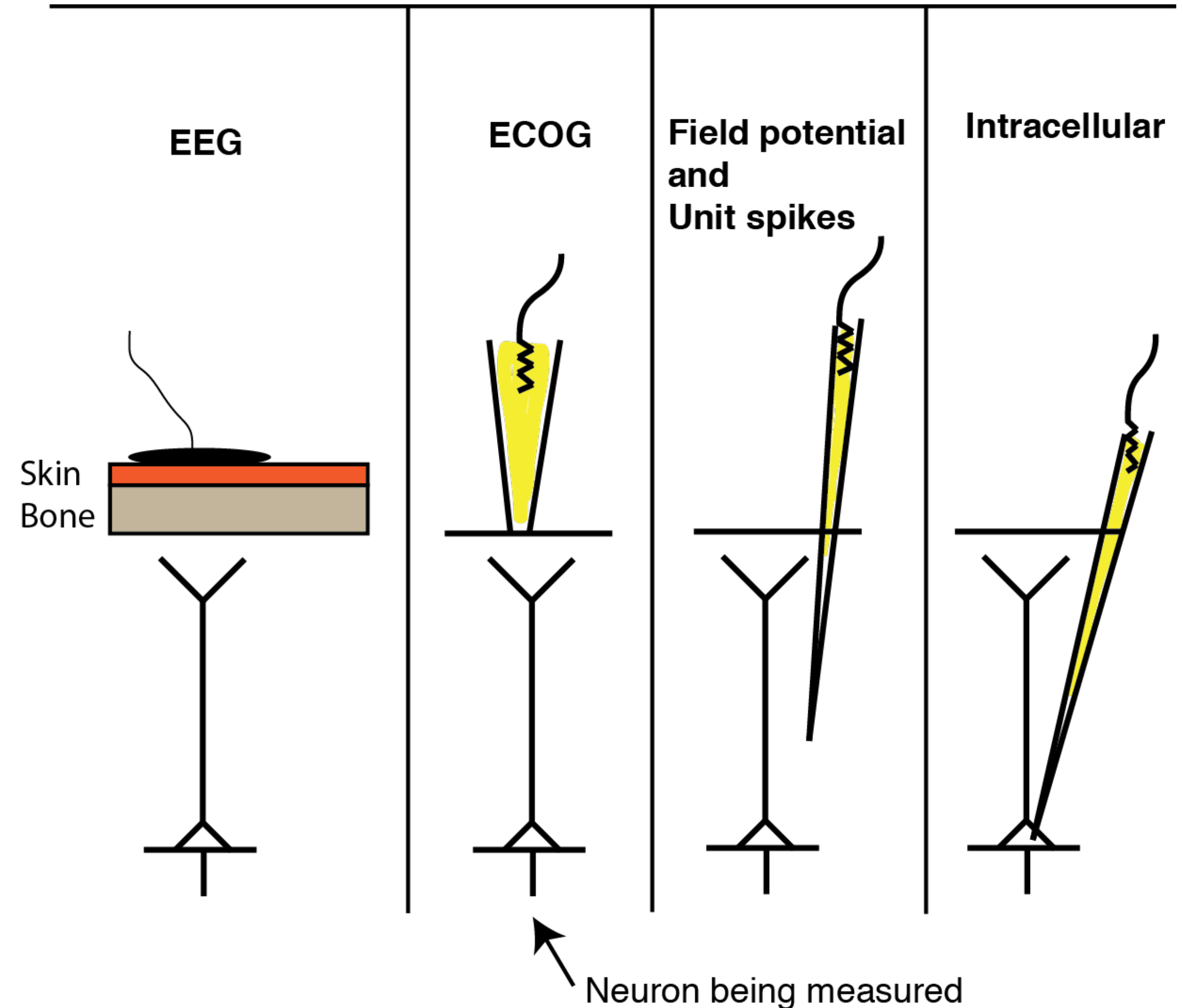
- **Power**- It's powerful
- **Scope is bigger** - We are making models and theories that are useful, of increasing complexity and scope - drawing connections from broad sources
- **Models are finite, world is infinite** - The models are only models, and thus finite, and do not capture the infinite dynamics of the system
- **All models have assumptions** - All models and fields are based upon assumptions
 - Recognize that and seek to make more and more useful studies, models, data science tools for neuroscience and beyond
 - Treat all who say they are displaying the underlying mechanisms with skepticism (take the best of their theory and apply it but recognize for what it is)

From statistics to recordings...

- We have discussed parametric and nonparametric models
- NWB, DANDI, BIDS
- EEG/MEG, MOCAP, LISC
- Data science and neuroscience perspective on it
- Modeling concepts
- Now let's consider maps between neurons and systems

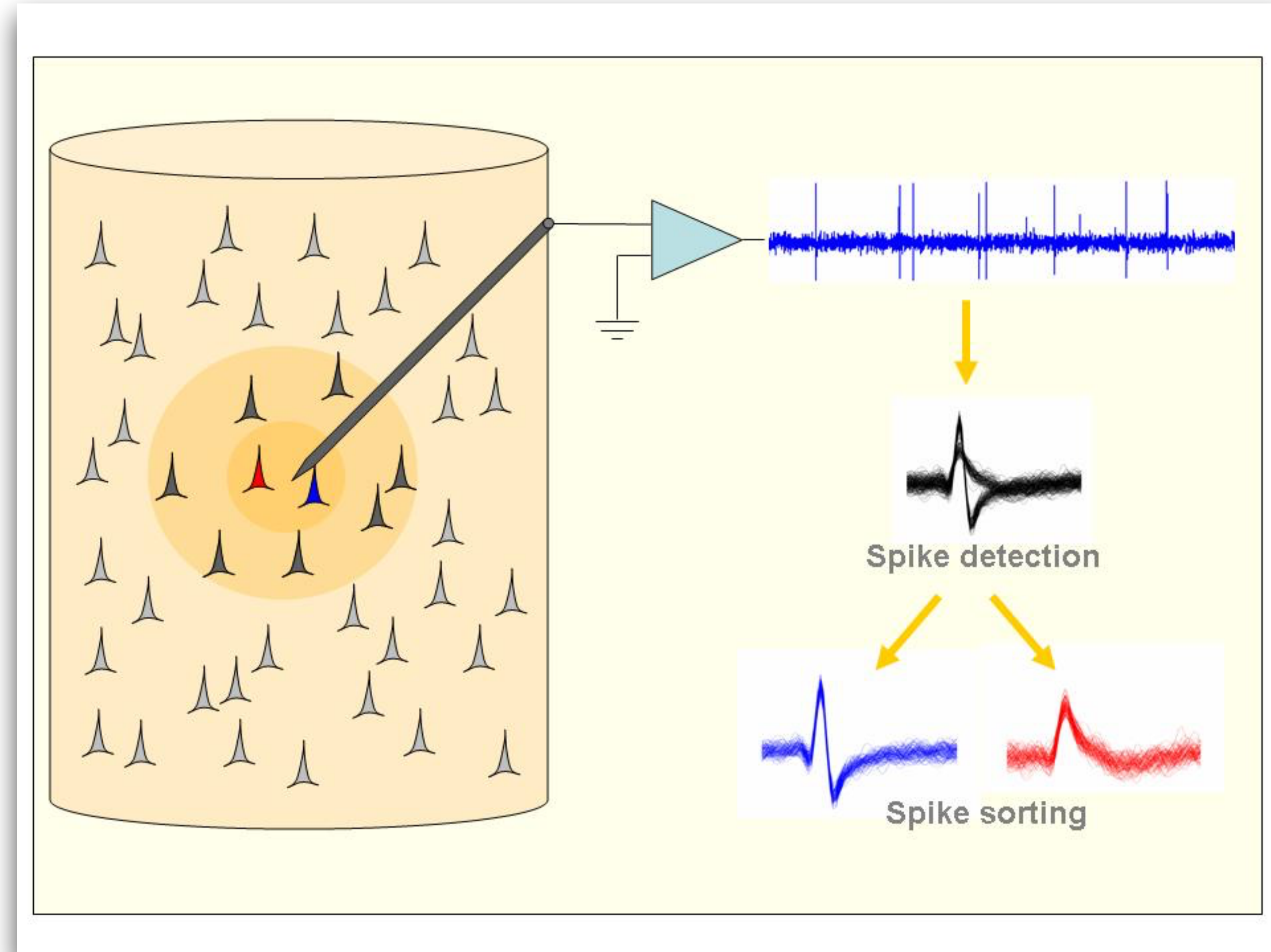
Neuronal recording approaches

- Electroencephalography (EEG)
- Electrooculogram (ECOGG)
- Local Field Potential (LFP)
- Unit Spikes (US)
- Intracellular (IC)
- Number of units recorded by size:
 - EEG > ECOG > FP >> US > IC



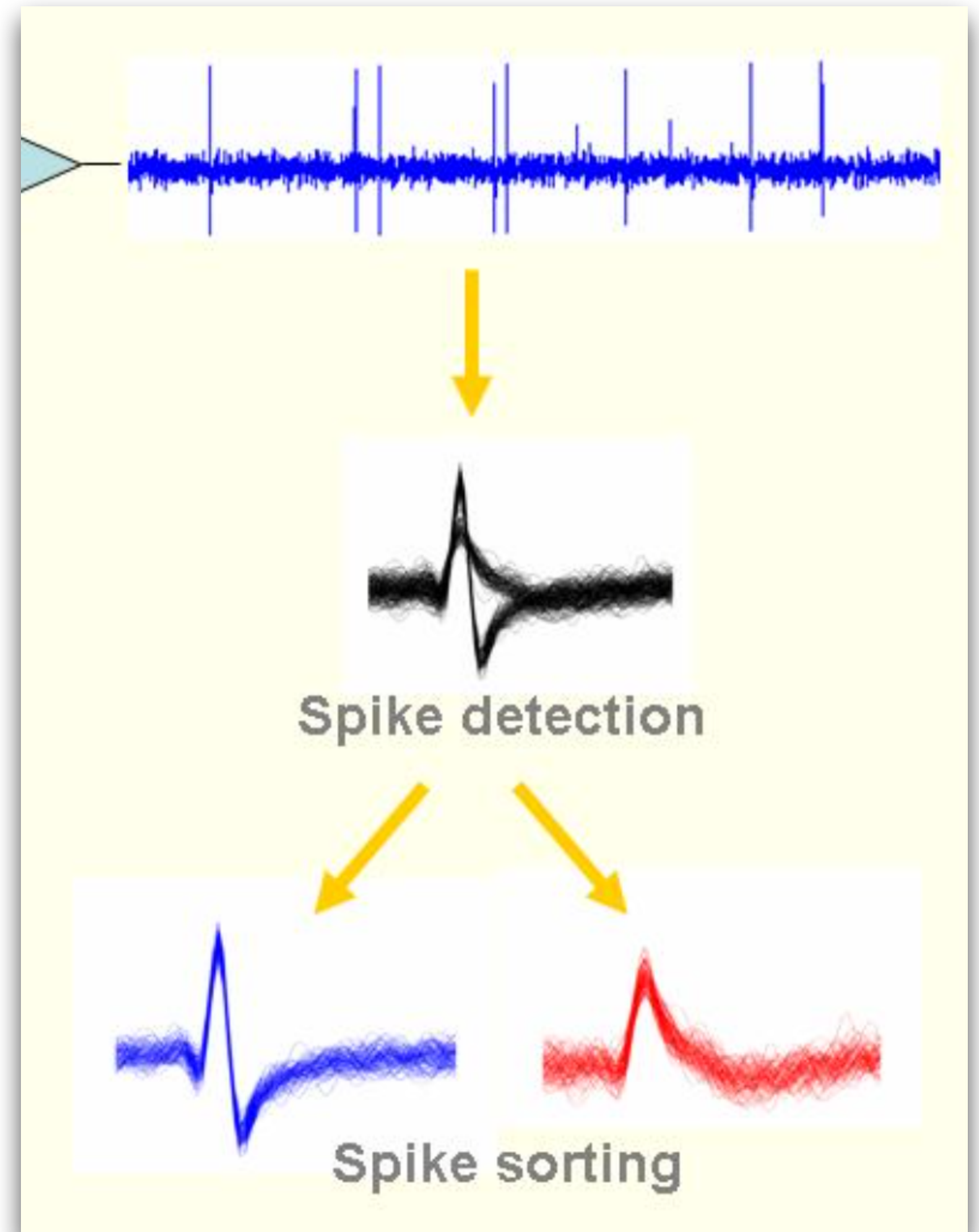
What is “Single unit recording?”

- Methods of measuring the electrophysiological response from single neurons with a micro-electrode
- An action potential generated by a neuron firing travels as a current down the excitable membrane regions through the soma and axon
 - The rate of change in voltage w.r.t. time is recorded
- Recorded extracellularly (several possible approaches)
- Micro-electrodes - High impedance, fine-tipped and conductive



Spike sorting

- Spike sorting is a process of processing neuronal data and grouping neuronal spikes into clusters based on their firing characteristics/shape
- ‘Which spike corresponds to which neuron’ from a cluster that is recorded



Spike sorting methods

- Many exist, from simple to sophisticated (a sample here)
 - **Amplitude discriminator**
 - **Window discriminator**
 - **Characteristic shape/template matching**
 - **Supervised learning**

Issues and challenges

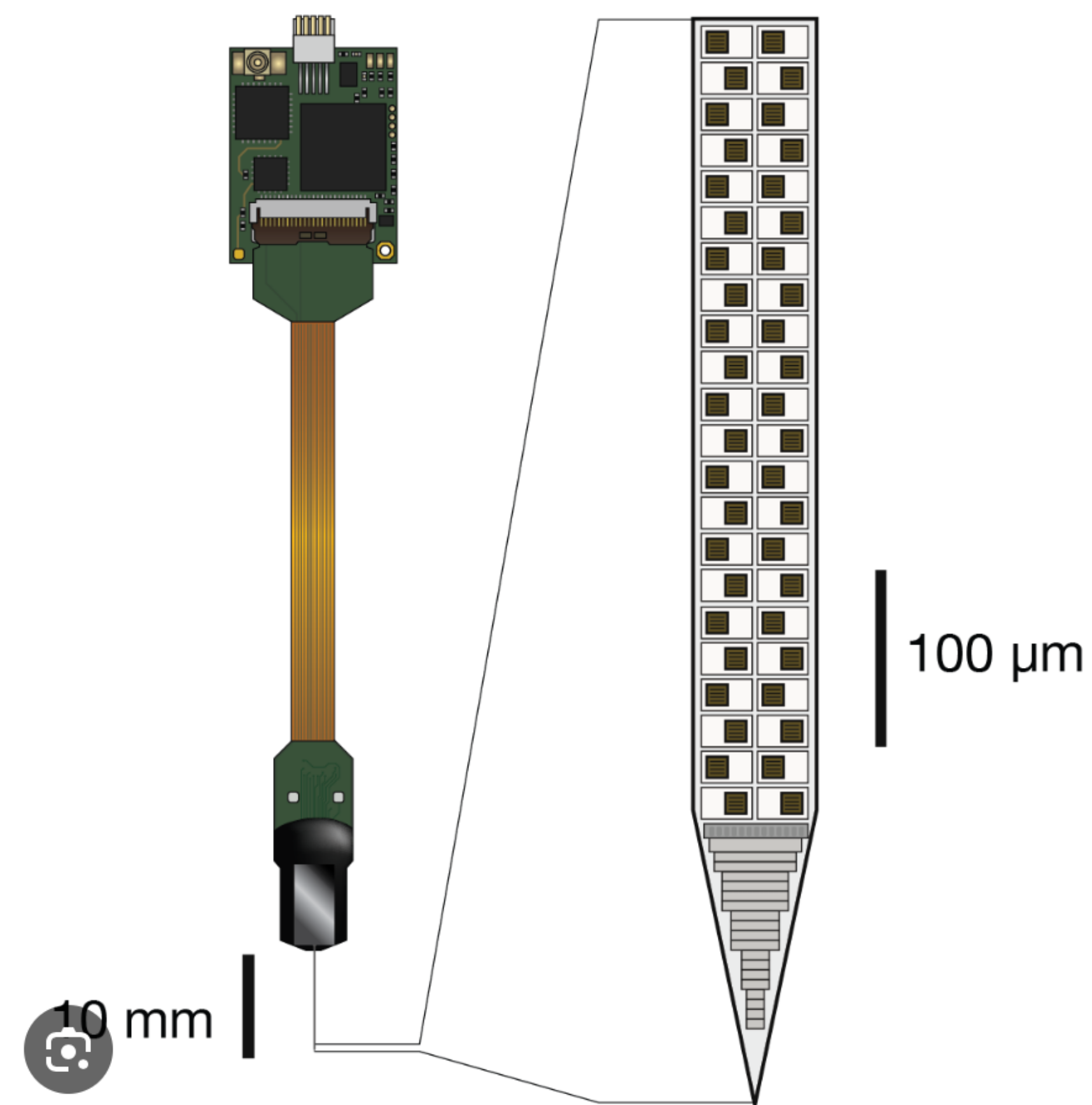
- Tetrodes
- Overlapping spikes
- Bursting cells
- Non-gaussian clusters

Introduction to python modules associated

- <https://pypi.org/project/spikeinterface/>
- <https://elifesciences.org/articles/61834>
- <https://github.com/topics/spike-sorting>
- <https://core.ac.uk/download/pdf/52193212.pdf>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7704107/>

High density single cell recording

- Multiple organizations are developing micro electrodes with multiple recording sites per electrode
- Record thousands of neurons simultaneously instead of only a few
- Better picture of entire brain areas in realtime
- Animal models and recently human



High density recording defined

- Chung, J. E., Sellers, K. K., Leonard, M. K., Gwilliams, L., Xu, D., Dougherty, M. E., Kharazia, V., Metzger, S. L., Welkenhuysen, M., Dutta, B., & Chang, E. F. (2022). High-density single-unit human cortical recordings using the Neuropixels probe. *Neuron*, *110*(15), 2409–2421.e3. <https://doi.org/10.1016/j.neuron.2022.05.007>

Parameterizing heterogeneous datasets

- Definition, review
 - What do we mean by **parameterization**?
 - Reminder of what data is and stepping back to the big picture - ***representation***
 - What are **heterogeneous** datasets?
 - What are the **challenges** and solutions?
- **Tools and practice** in neural data science
 - https://nwb-overview.readthedocs.io/en/latest/tools/tools_home.html
- Examples

Parameterization vs. Hyperparameterization

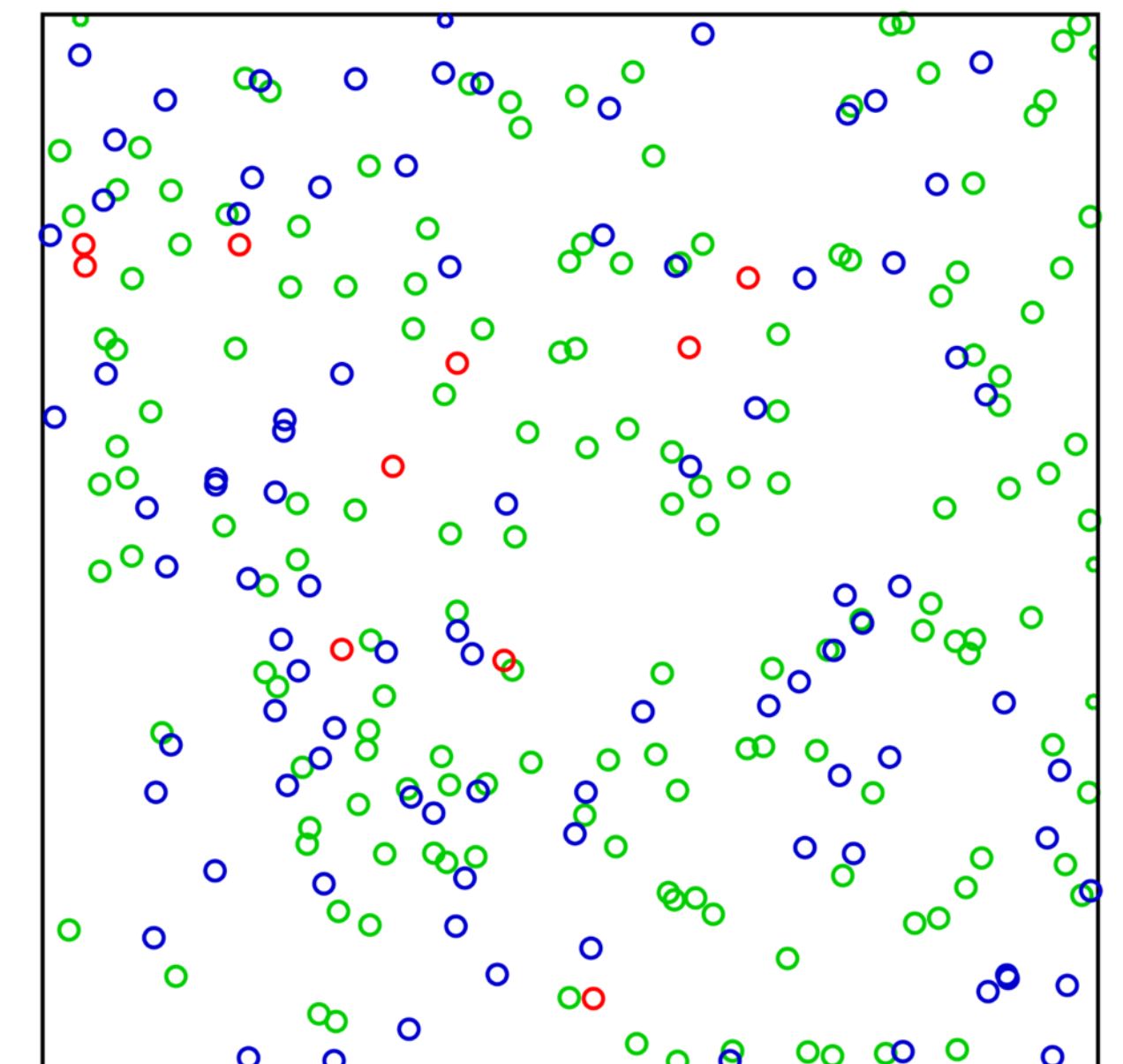
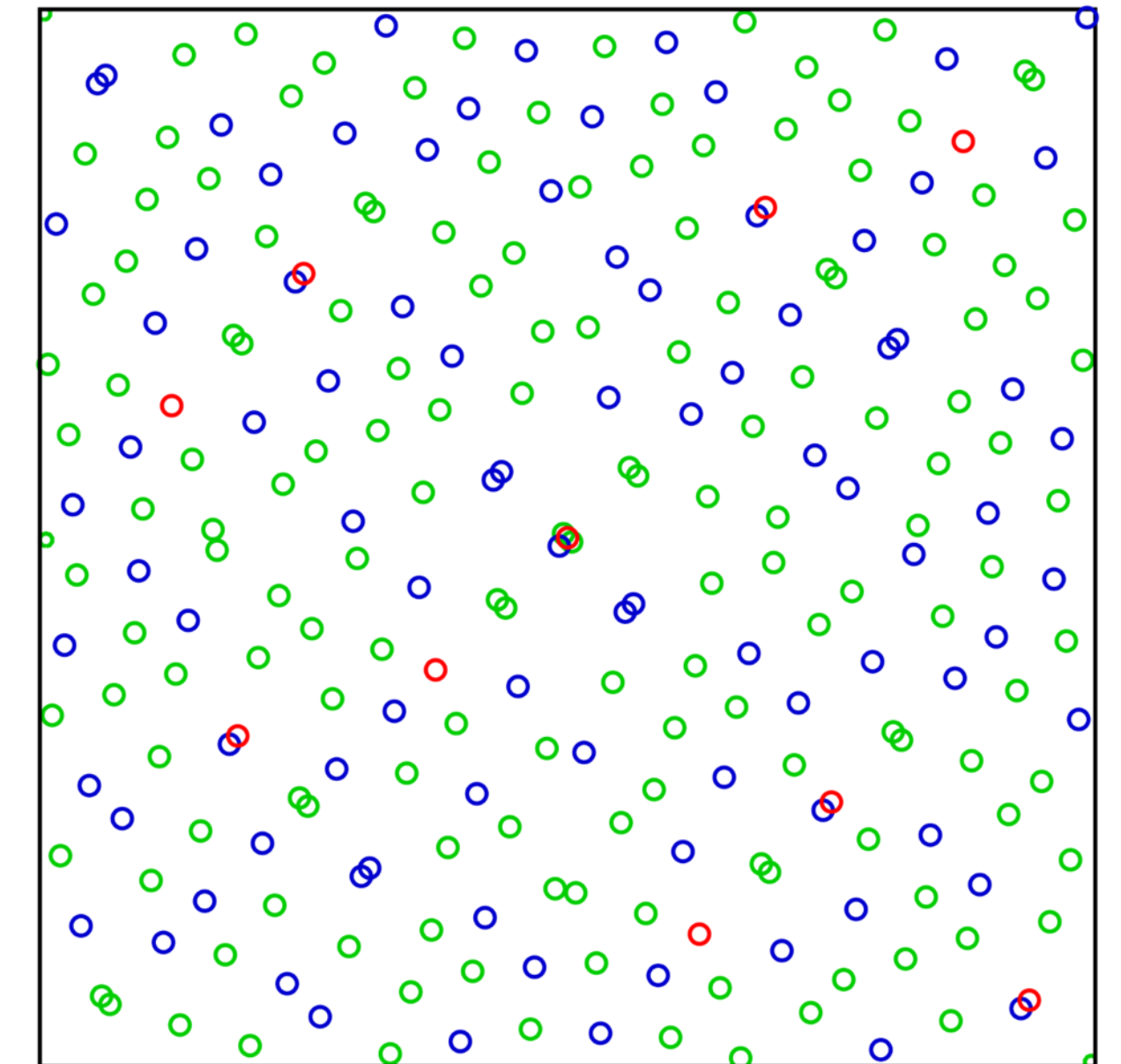
- **Parameterization** - the set of parameters that define the model unknowns to be fit, typically from data
 - For example, for $y = ax + b$, what are the parameters?
 - ANN - network weights
 - Calculated/learned from data
- **Hyperparameterization** - the set of parameters for machine learning in particular that define and control the learning process and are external to the model
 - Bisection algorithm for optimization - bisection parameter
 - ANN - parameters of the learning algorithm itself
 - Heuristic, can be set by practitioner, tunable for a given problem

Why is it a challenge to integrate them?

- Sampling rate mismatch
- Time/frequency/spatial domains
- Sample rate variability (why does this matter?)
- Sample time mismatch
- Format, software
- Missing data, data mixture/non-tabular etc
- Memory usage
- (Not an exhaustive list)

Sobol sequences

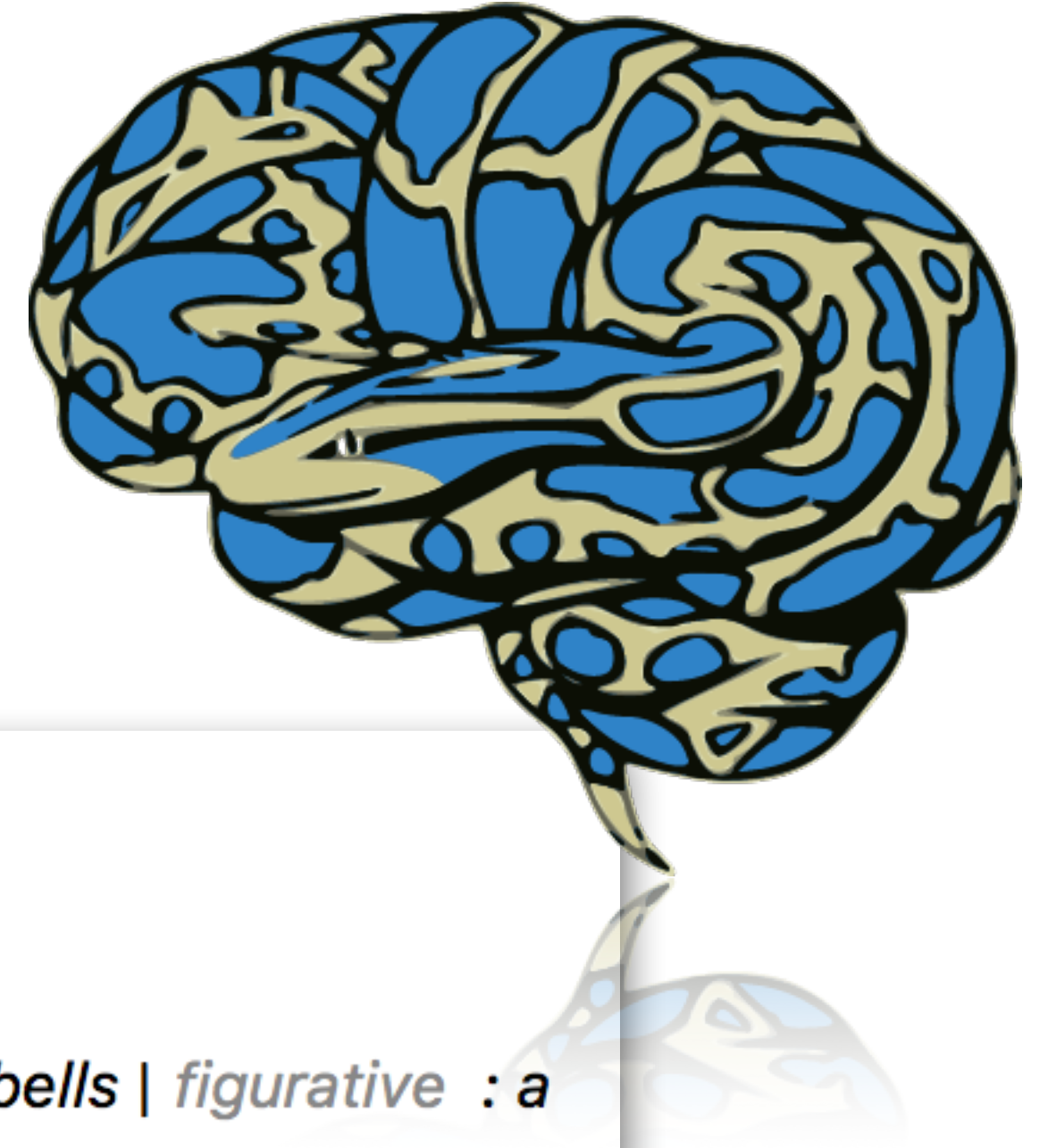
- Quasi-random low-discrepancy sequences
- https://en.wikipedia.org/wiki/Sobol_sequence
- Which one covers the space more evenly, just by eye?
 - Sobol or pseudorandom
- **Sobol sensitivity analysis** to analyze influence of parameters in computational neuroscience models
 - <https://hal.science/hal-03464025/file/root.pdf>
 - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8184610/>
 - Model reproducibility





- PySurfer is a Python library for **visualizing cortical surface representations of neuroimaging data**.
- The package is primarily intended for use with [Freesurfer](#), but it can plot data that are drawn from a variety of sources.
- PySurfer extends [Mayavi's](#) powerful rendering engine with a high-level interface for working with MRI and MEG data.

NiBabel - definition



- “Access a cacophony of neuro-imaging file formats”

- Cacophony?

cacophony | kə'käfənē |

noun (pl. **cacophonies**)

a harsh, discordant mixture of sounds: *a cacophony of deafening alarm bells* | *figurative* : *a cacophony of architectural styles* | *songs of unrelieved cacophony.*

- Read and write access to common neuroimaging file formats,
 - including: [ANALYZE](#) (plain, SPM99, SPM2 and later), [GIFTI](#), [NifTI1](#), [NifTI2](#), [CIFTI-2](#), [MINC1](#), [MINC2](#), [AFNI BRIK/HEAD](#), [ECAT](#) and Philips PAR/REC.
 - In addition, NiBabel also supports [FreeSurfer](#)'s [MGH](#), geometry, annotation and morphometry files,
 - provides some limited support for [DICOM](#)

In class report development (~30m)

- Define this course's intent
- Draw comparisons between this course and requirements
- How does this course build upon what came before?
- How can you use your starting point in this course to expand your understanding?

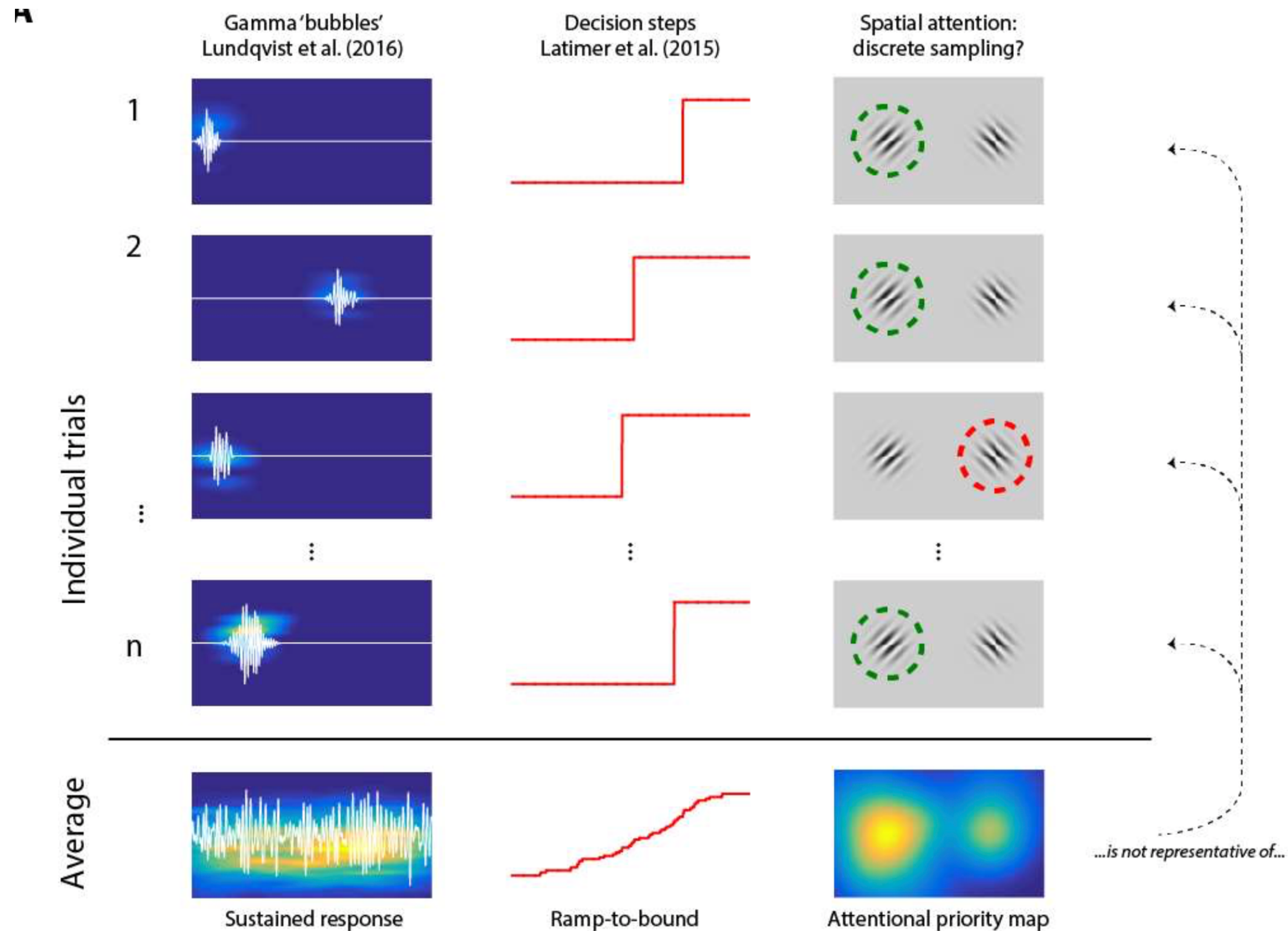
Motivation for single trial analysis

- Traditionally neuroimaging techniques are used to compute differences between means over many trials/subjects/studies
 - e.g. in classical cognitive neuroscience, theories of working memory assumed that task-relevant info. is maintained by persistent neural activity
 - Representations kept online by persistent activity patterns, evidence based on averaging massive numbers of subject trials
 - Assumption is if true distribution is contained in noisy measures, measure many times, average, you recover the noise-free representative pattern

Single Trial Analysis definition and classes

- All methods that consider ***variance within subjects***
 - 2 classes of methods
 1. Univariate methods
 2. Multivariate methods
- Applications - *Useful for both behavioral and neuroimaging experiments*

Combining by computing mean doesn't necessarily create a good representation



Combining by computing mean doesn't necessarily create a good representation

- Traditionally **statistical power** = more observations (i.e., trials) to average data

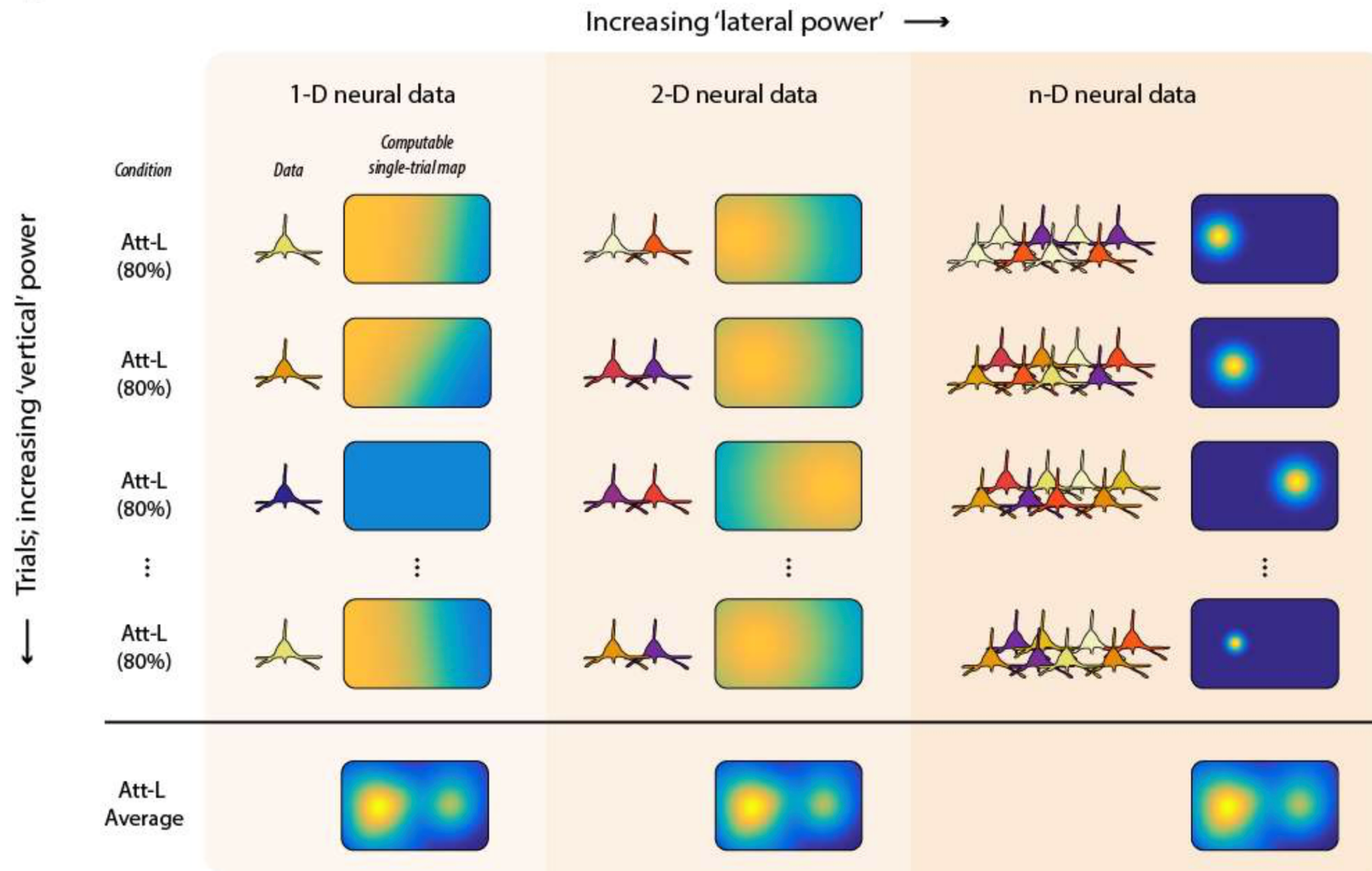
- “Vertical power”

- **Lateral power:** adding more measurement density (spatial dimension).

Larger lateral power ->

Necessary for characterizing neural dynamics in single trial (we'll come back to this)

B



Working toward a strong data science question

Vague: How does the brain change when you have a brain injury?

Better: What neurological changes are there after a stroke?

Even better: What neurological and behavioral changes can be measured with EEG and motion capture between an average normal subject and a stroke patient who had a recent stroke that impaired motor function?

Best?

Practical challenges in neural data science

What are the biggest challenges of making neural data science work?

- Many different libraries, many different dependencies
 - Have to get good at or build a team with the skill set to make complicated practical analyses, measurements possible
 - New attitude
- Many different modalities means you need to be aware of the issues associated with each in order to avoid spurious conclusions
 - Need skills or to know how to develop the skills for each - where do you look, how to ask questions (technical and content)

What are the biggest challenges of making neural data science work?

- A lot of data requires strategies to deal with it - you can't just load terabytes into RAM (at this point)
- Changing landscape of what is available
- Increased heterogeneity of teams, and interdisciplinary has challenges as well

What are the biggest challenges of making neural data science work?

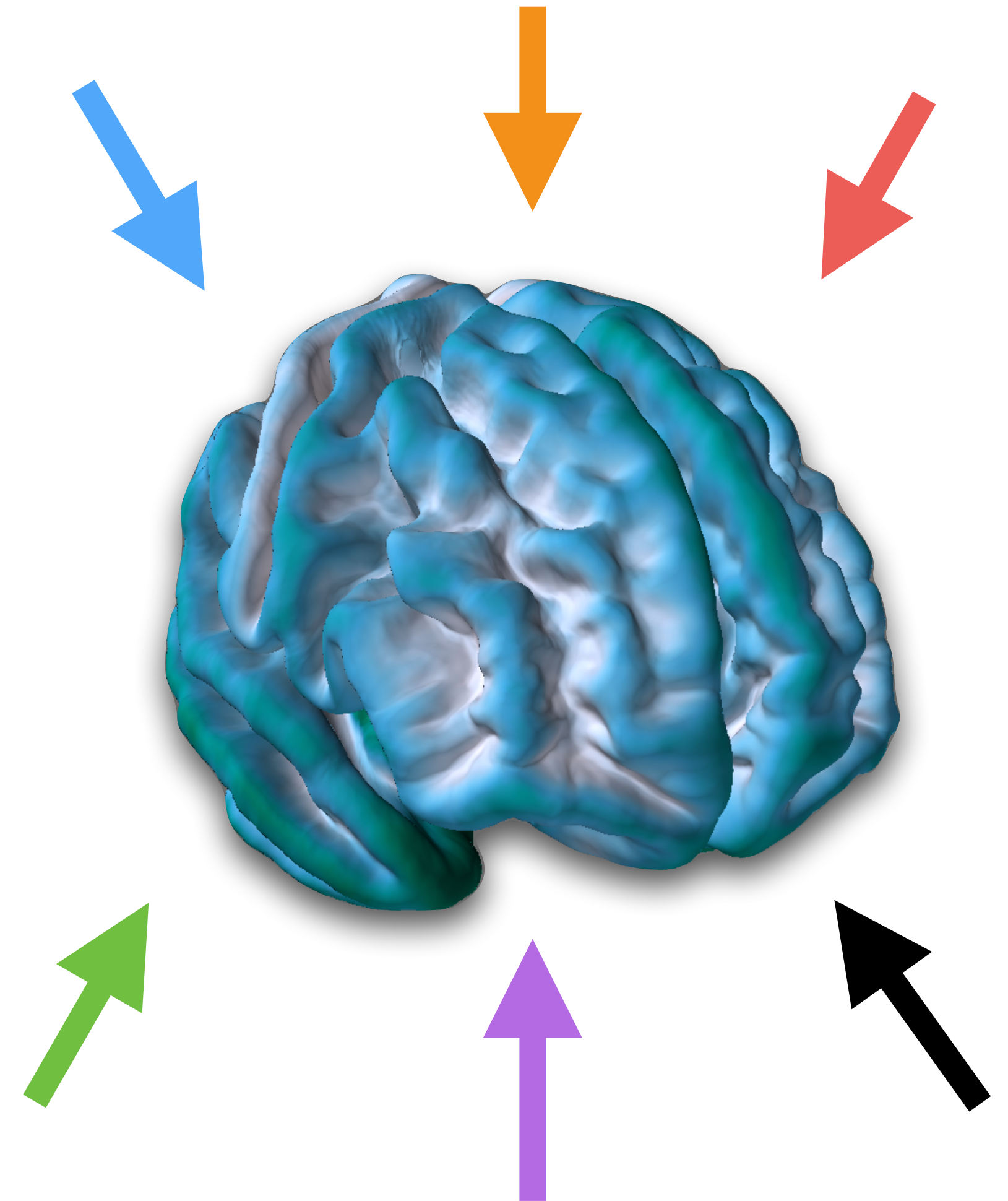
- How do you learn about, for example motion capture? Or how do you learn about eye tracking?
- How do you review all the literature?
 - Reading unfamiliar literature?
- Where do you get the data?
- How do you know if the data is good?

What are the biggest challenges of making neural data science work?

- Free and open does not mean good, easy to use, understand, or relevant
 - Need to use judgment, sometimes ask questions, consider peer reviewed, should be well documented
- Integration of different datasets that may not have been meant to be combined
 - We discussed integration of heterogeneous sets
- Reducing dimensionality
 - Care not to cut out important information early
 - Sometimes reducing data can lead to focusing on richer content
 - Other times the rich content is not at all obvious and we must operate on it all to determine that

Neural Data Science

- New way of putting together disparate methods
- Integrate many perspectives to build a better picture of brain, behavior, cognition
- We have explored many different approaches, discussed their integration, discussed how to think in all the ways you need to in order to implement techniques in a single study/groups of studies



Thank you