COGS138: Neural Data Science

Lecture 5

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

- Announcements
- Assignment 1 overview
- Review - Last time
- Asking the right questions in data science
- LISC, hypothesis generation (automated)
- Gene expression studies introduction, animal models
- F.A.I.R. data, what is it and why?
- NWB data and BIDS data - definition, accessing, usage and relevance
- DANDI - putting datasets together and making it all available, reusable and documented

Announcements

- Final reminder to check on your FinAID status
- A1 - due **a week from release**, which will be tonight or tomorrow
- Reading 1 - Released on canvas and in web site password protected area tonight, lecture quiz due **a week from release**, released tonight
- **Group formation** - time to start choosing who you want to work with for your project group

Last time

Course links

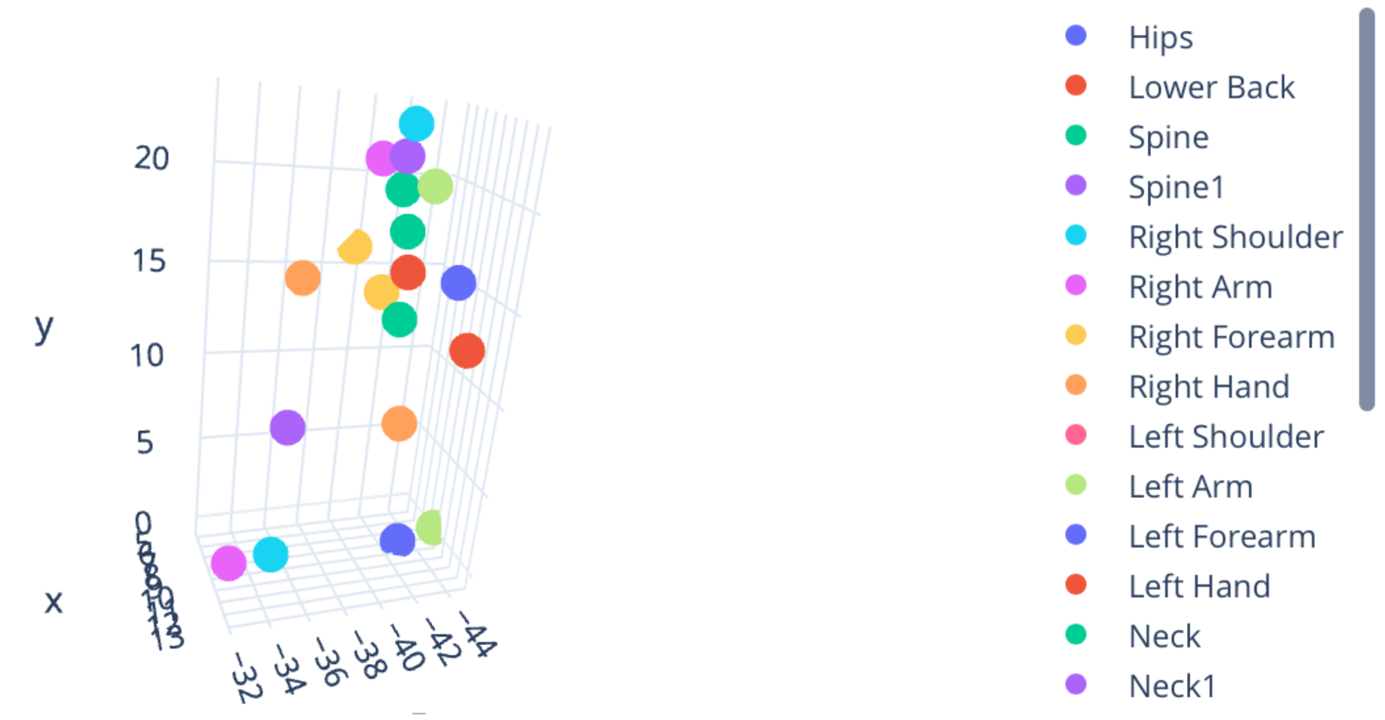
Website	http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23	Main face of the course and everything will be linked from here. Lectures, Readings, Handouts, Files, links
GitHub	https://github.com/drsimpkins-teaching	files/data, additional materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	https://piazza.com/ucsd/spring2023/cogs138_sp23_a00/home (course code on canvas home page)	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/44897	grades, lecture videos
Anonymous Feedback	Will be able to submit via google form	If I ever offend you, use an example you are uncomfortable with, or to provide general feedback. Please remain constructive and polite

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)

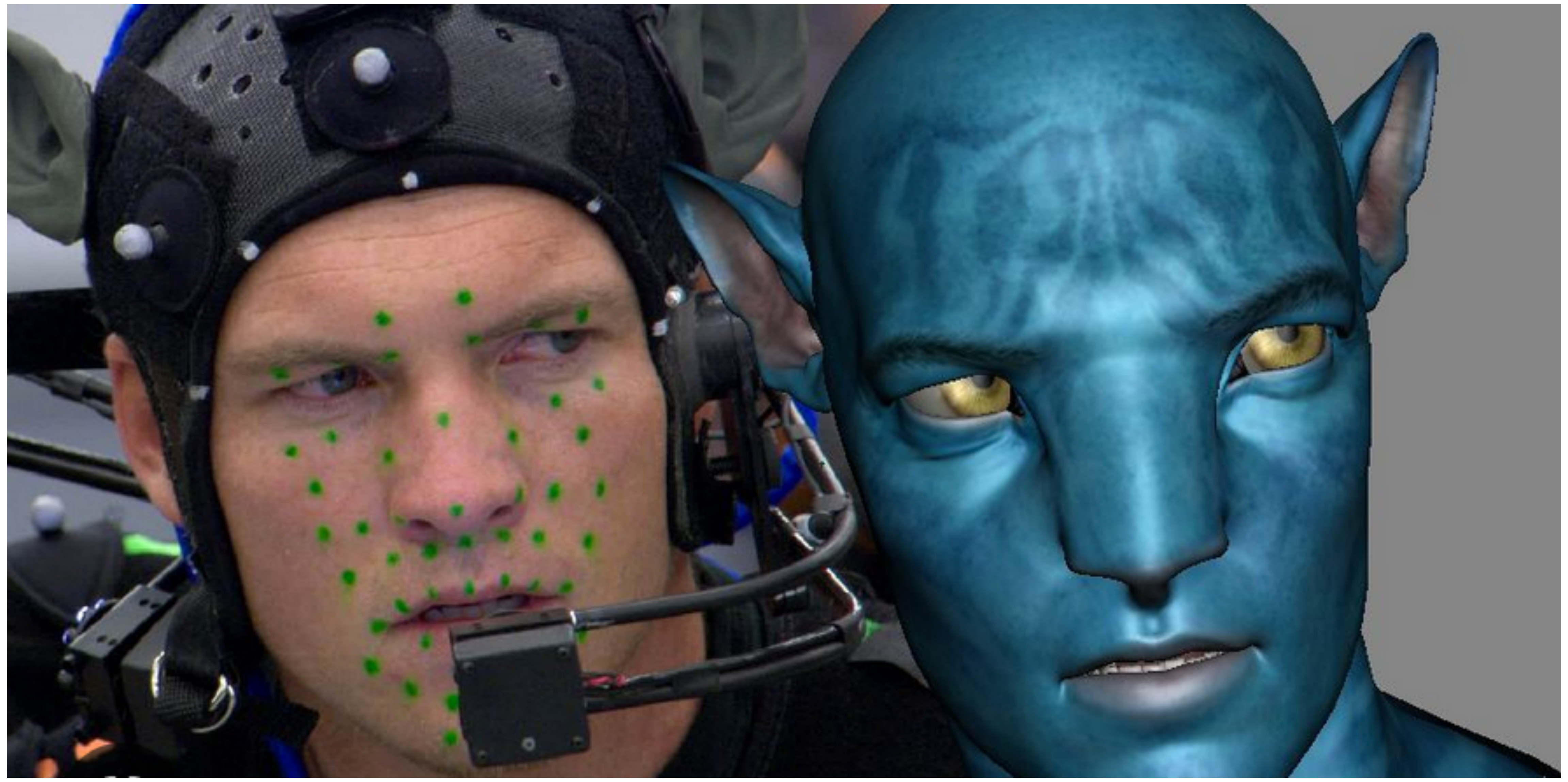


Facial motion capture

- Facial expression capture

- Using markers and a fixed perspective camera tracking with the subject

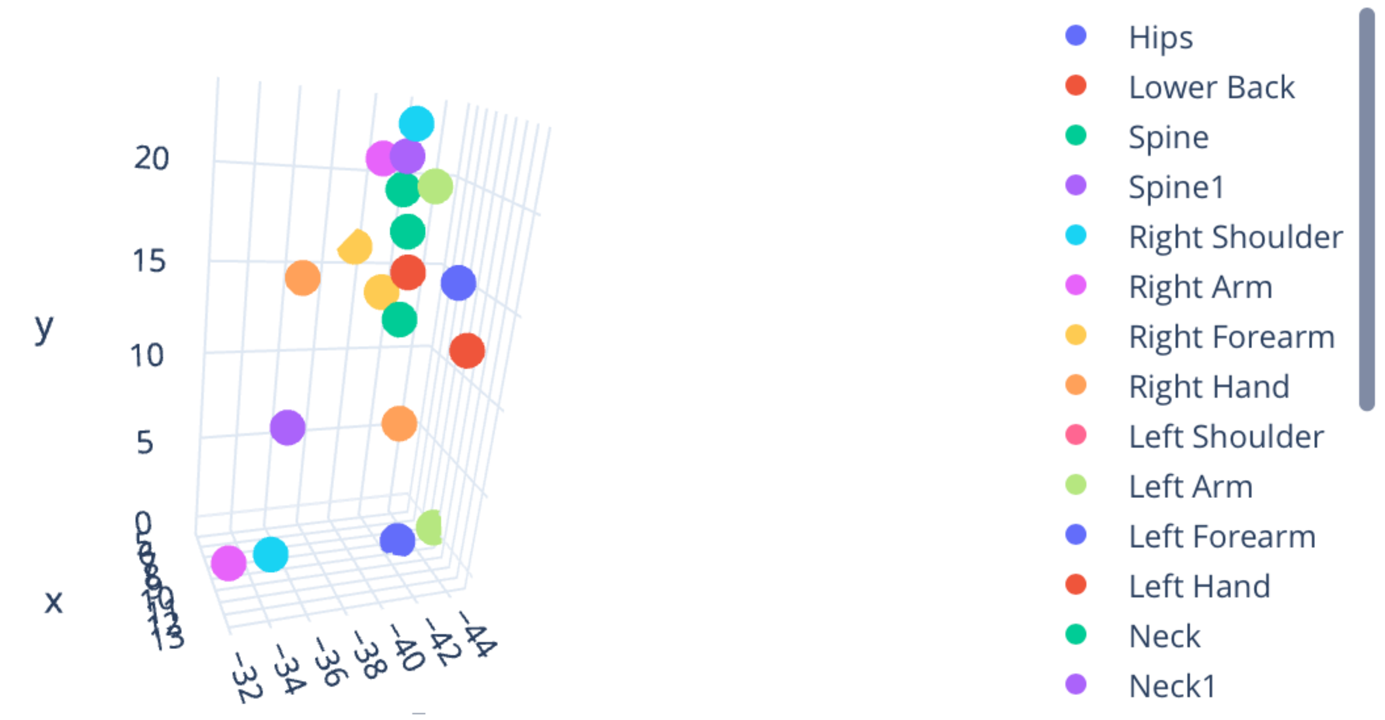
- Combined with positional markers for mapping



Motion capture data

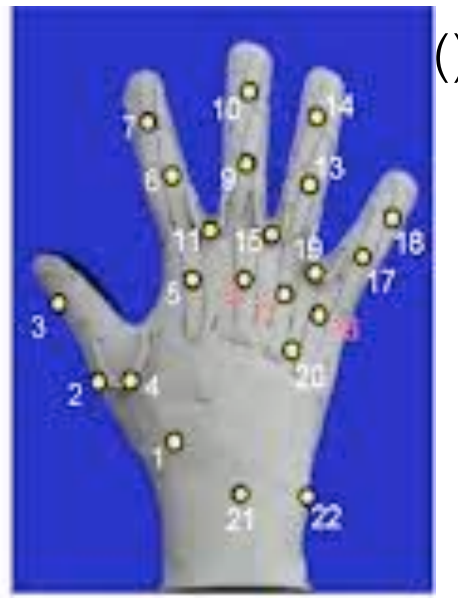
- PyMO - https://omid.al/projects/pymo/
- pypi - https://pypi.org/project/mocaplib/
- Others
- Not well standardized yet

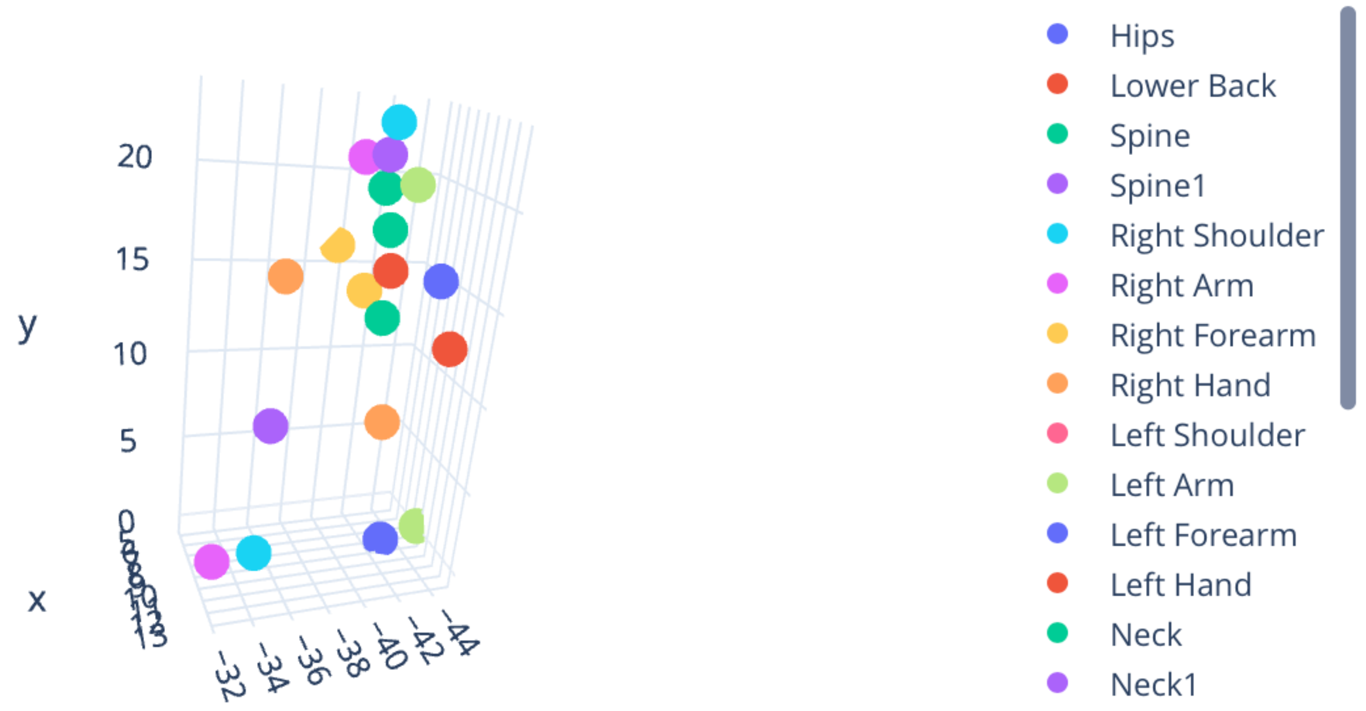
(Source: https://medium.com/swlh/movement-classification-b98614084ec6)



Motion capture data - challenges

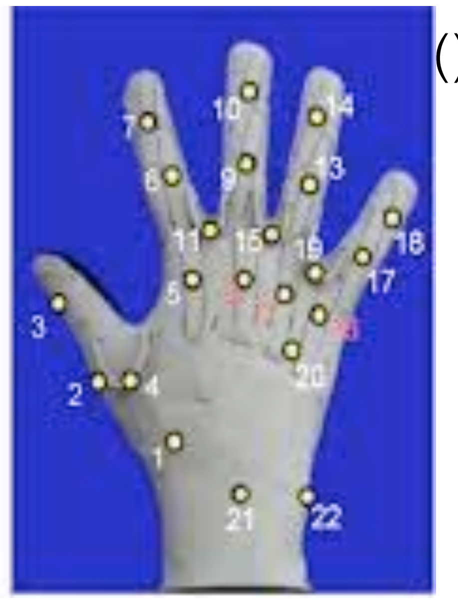
- Hand manipulation involves many occlusions
    - Estimation
    - High camera density
    - Active markers
- Predictive estimation
- Marker occlusions generally, jumps and discontinuities, open/closed chain complexity
- Active systems require power, wires, may be delicate
- https://www.engadget.com/2018-05-25-motion-capture-history-video-vicon-siren.html





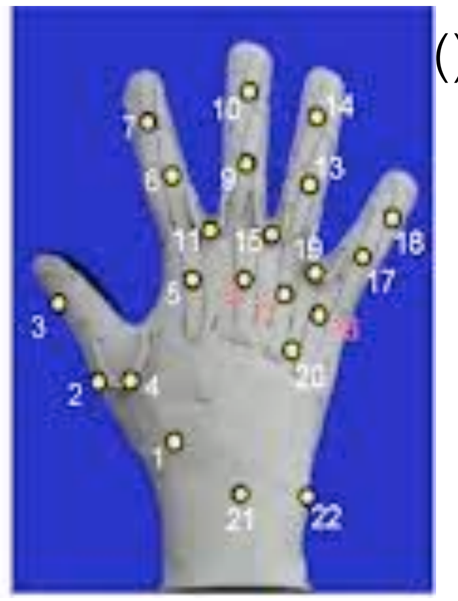
Motion capture systems

- 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



()

Motion capture systems



- **VICON**: https://www.youtube.com/watch?v=HBD6vA0Xi6Y

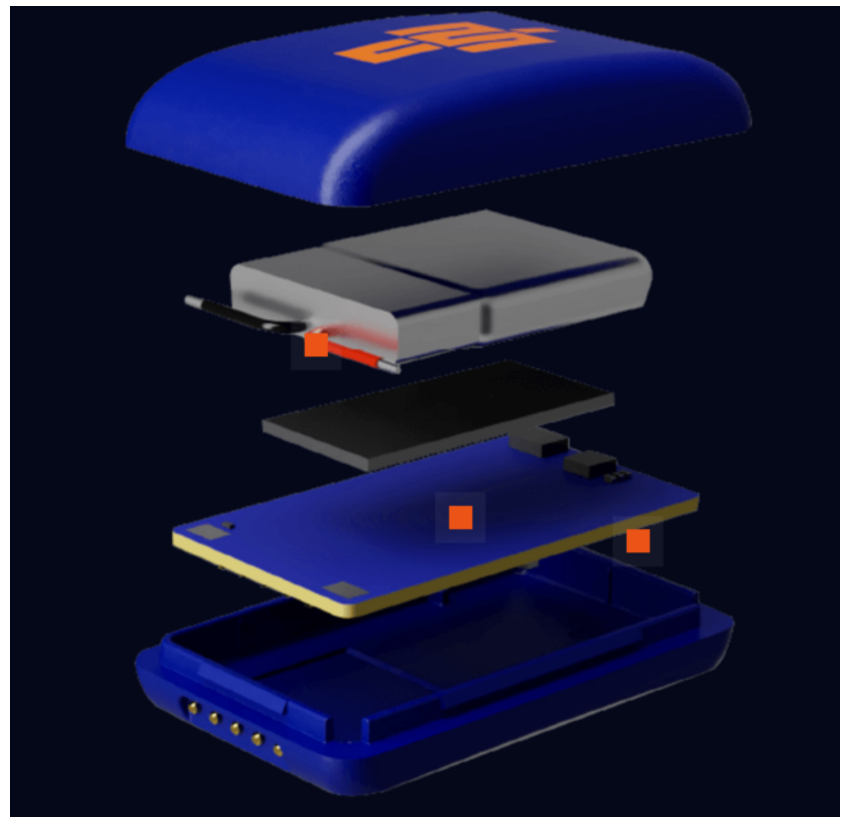
- **PhaseSpace**:
  - https://www.youtube.com/watch?v=A1BrYmC1Vpo
  - https://www.youtube.com/watch?v=iklXUxpq-T4

VICON

- Passive, IR camera type
- Slower refresh, non-unique markers, dependent on larger numbers of cameras for occlusion detection and reliability for complicated kinematic systems
- Works fairly well for less complex kinematics and no occlusions
- Fairly accurate at 0.017mm max
- Provides an actual video image (low res grayscale) as well potentially
- Integrated software and calibration ‘wand’
- Integrated with other hardware and items like IMU
- Various software options from Vicon
- https://www.vicon.com/applications/engineering/
- https://docs.vicon.com/display/Shogun18/Getting+started+with+Vicon+Shogun
- https://docs.vicon.com/display/Shogun18/PDF+downloads+for+Vicon+Shogun?preview=/174784515/174785494/Python%20scripting%20with%20Vicon%20Shogun.pdf

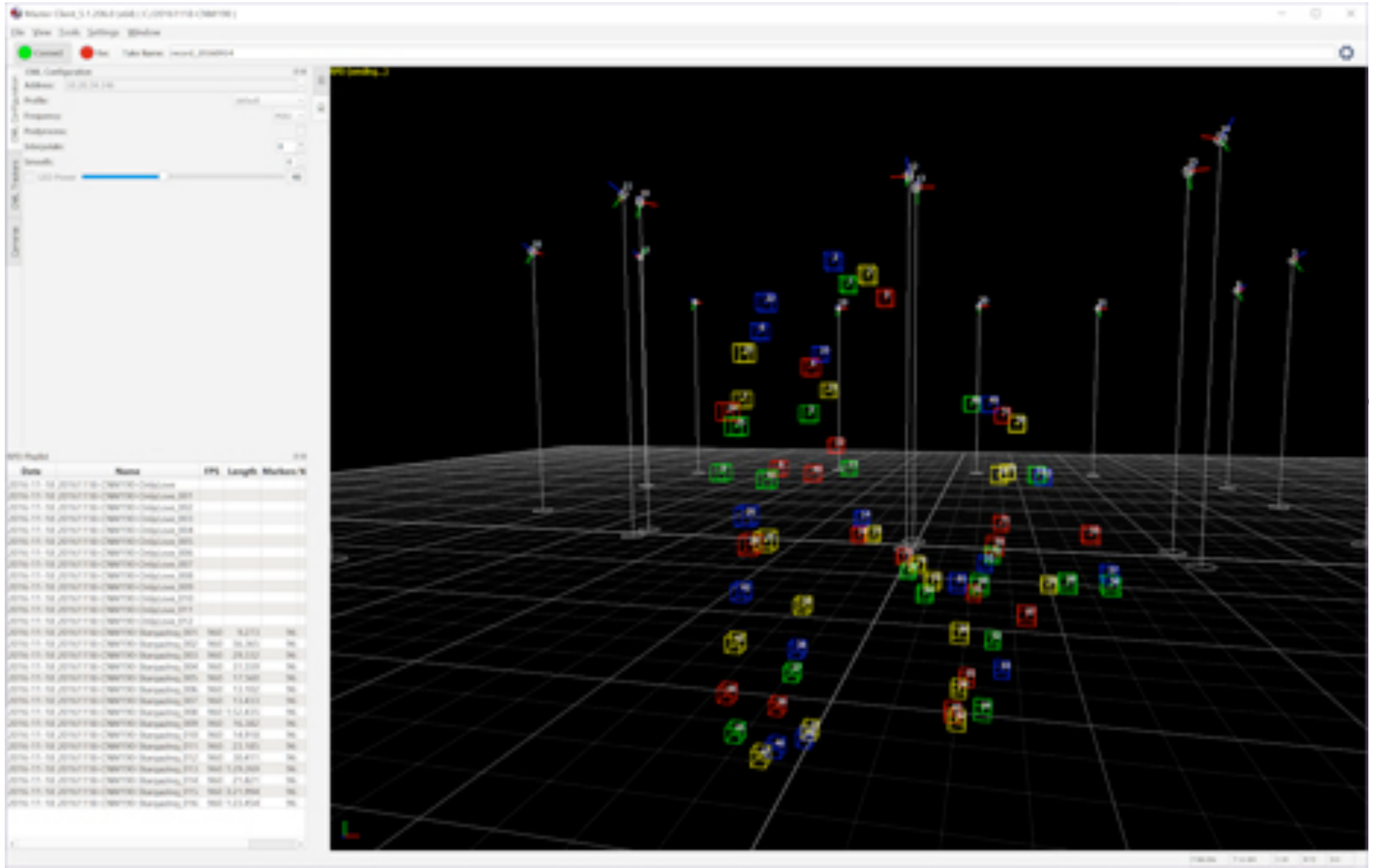
Cameras:

IMU:



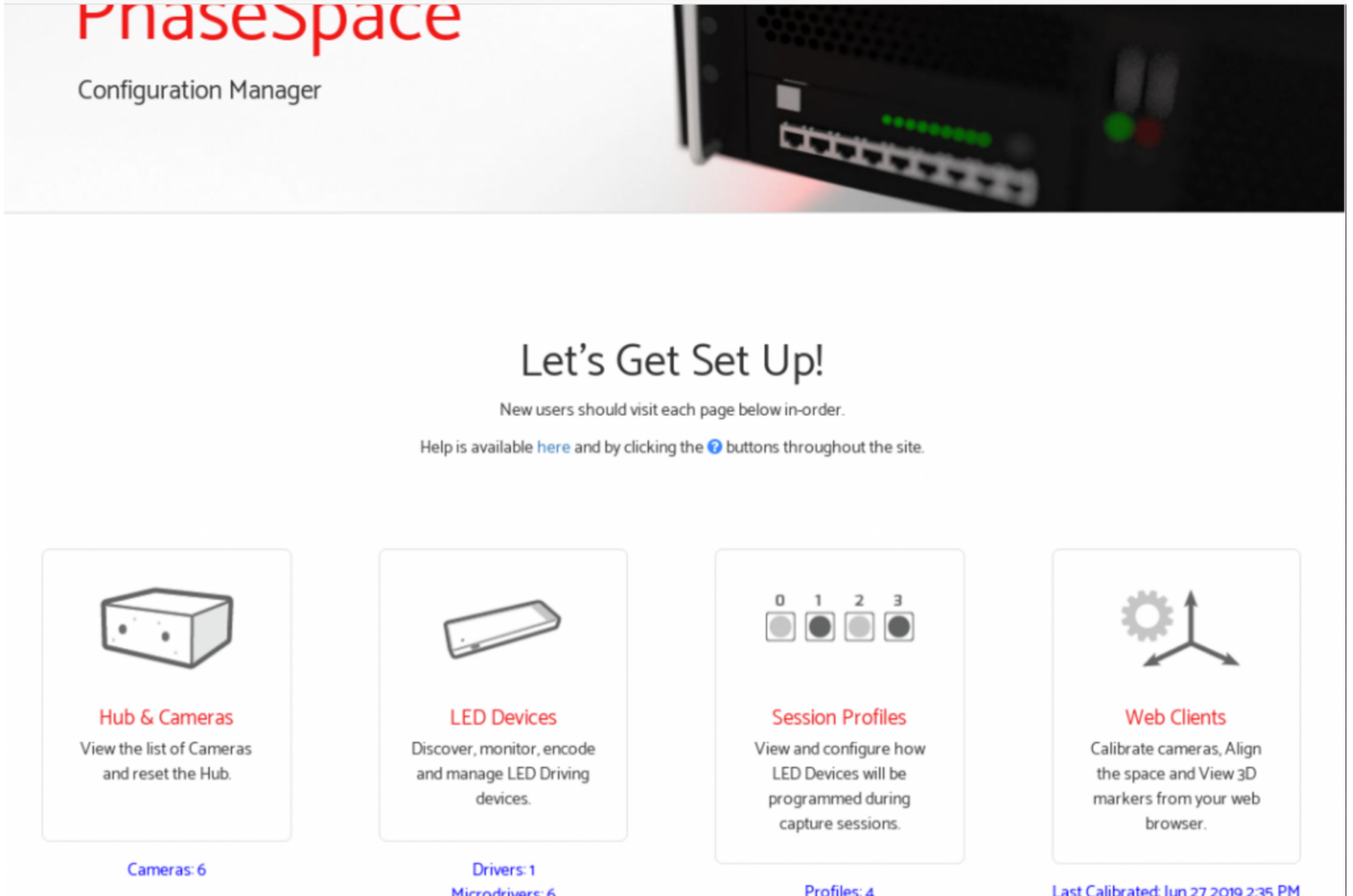
PhaseSpace

Hub:



- Much faster (960Hz)
- High precision sub-millimeter precision (~20μ) at full sub-pixel resolution of 36000 x 36000
- No confusion between markers as each is unique, active pattern
- No video image, less cameras for reliability required
- Integrated software and calibration ‘wand’
- http://www.phasespace.com/software.html
- https://www.phasespace.com/applications/robotics/
- https://www.phasespace.com/applications/sports-medical/

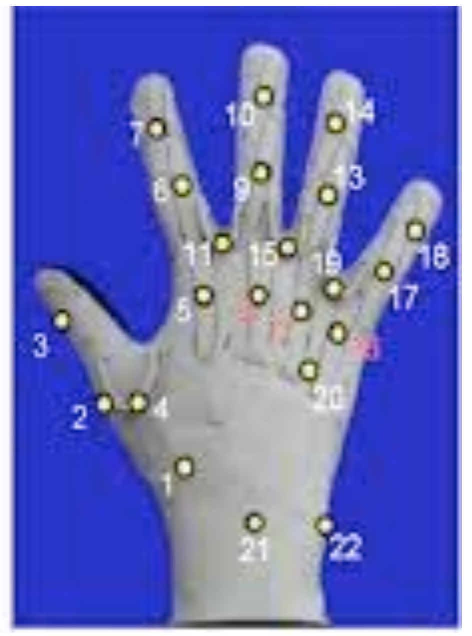
Config. Manager:

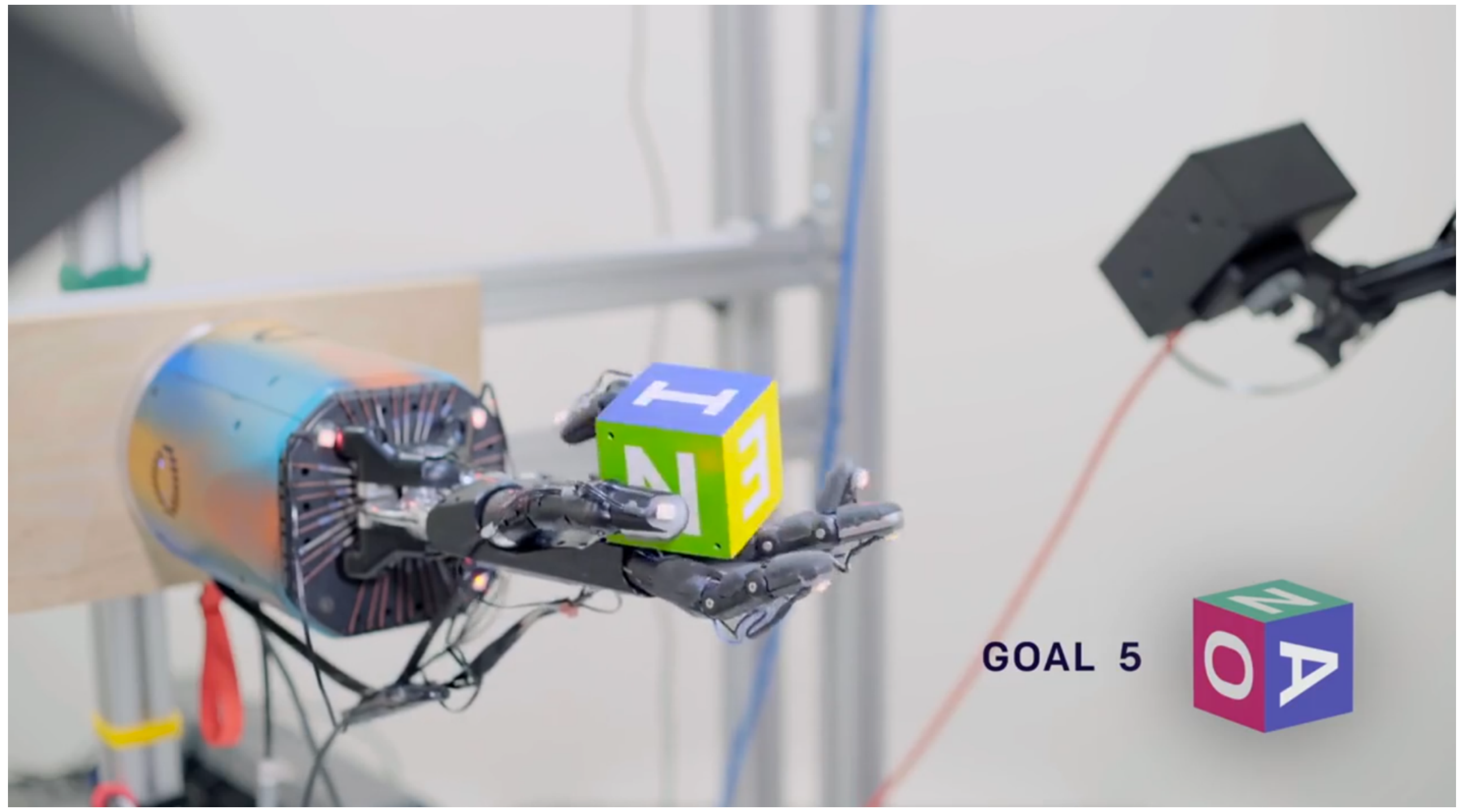


Motion capture systems

- Dextrous manipulation
  - Neither system is perfect
  - Better to have some glove and instrumented objects
  - Many of these are not perfect
    - Relative joint angles, glove covering or interfering with movement and interaction
- Not typically suitable for MRI type studies but EEG yes
- https://hub.packtpub.com/openai-reinforcement-learning-giving-robots-human-like-dexterity/

()





Motion capture data

()

- File type examples
  - https://en.wikipedia.org/wiki/List_of_motion_and_gesture_file_formats
  - VICON: https://docs.vicon.com/display/Shogun17/Post+-+File+types
  - PhaseSpace: export .C3D or .BVH files

Eye tracking

- Human eye movements are complex and indicate many things about cognitive and neurological states as well as dynamics

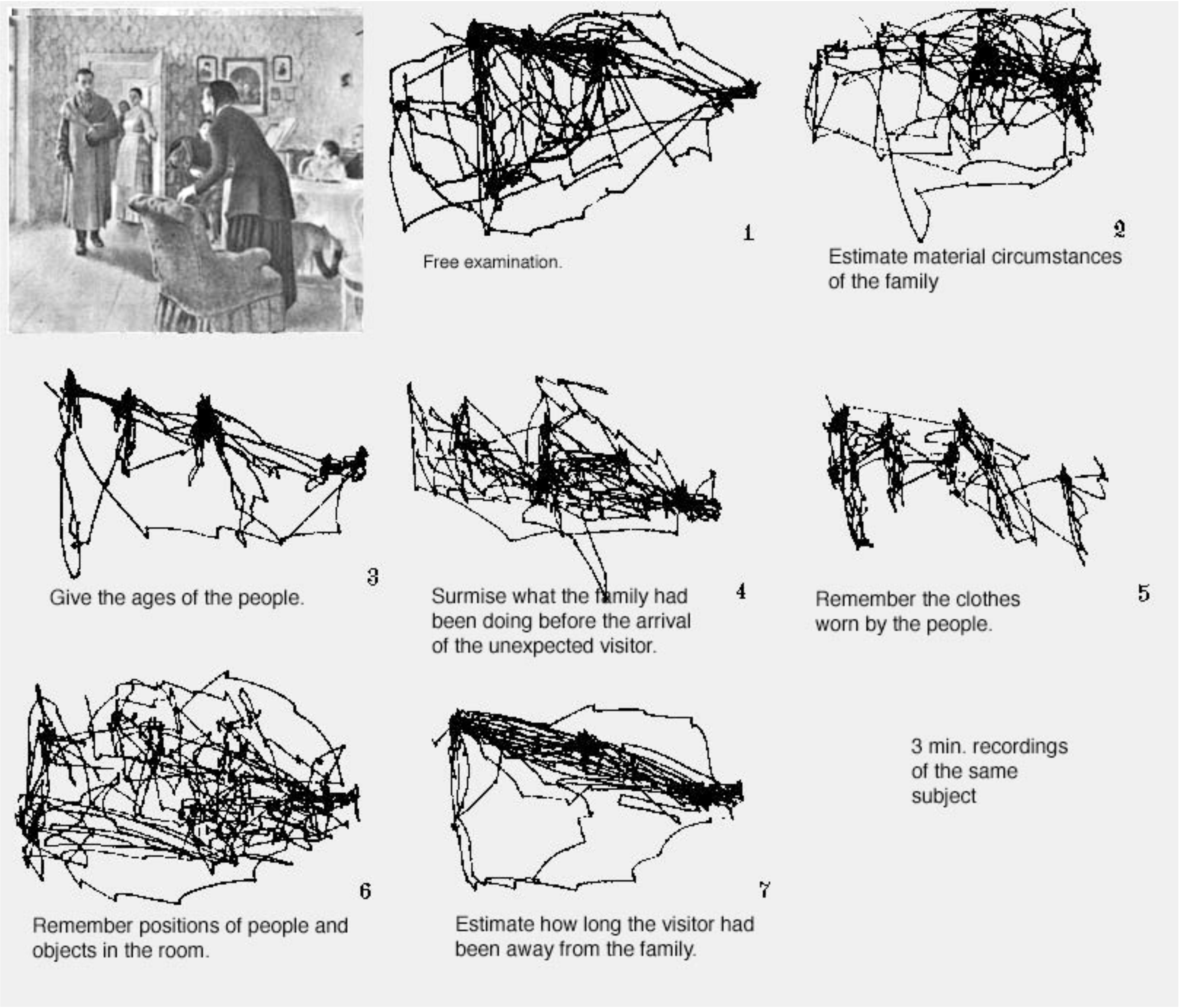
- Yarbis (1967) - task given to a person affects eye movement

- “Unknown water balloon release time”

- Pencil stuck in the ceiling tile going to fall but when?

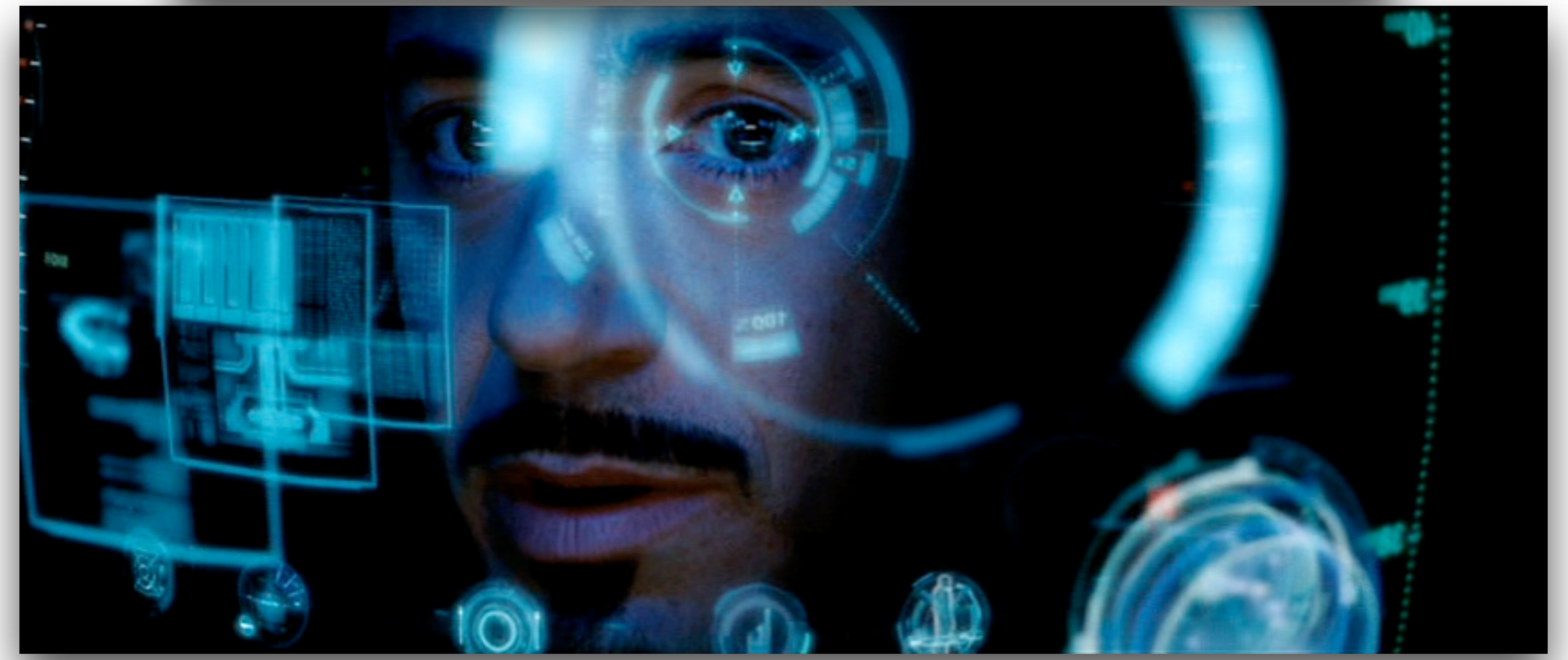
- Eye position, pupil dilation indications, focal point

()



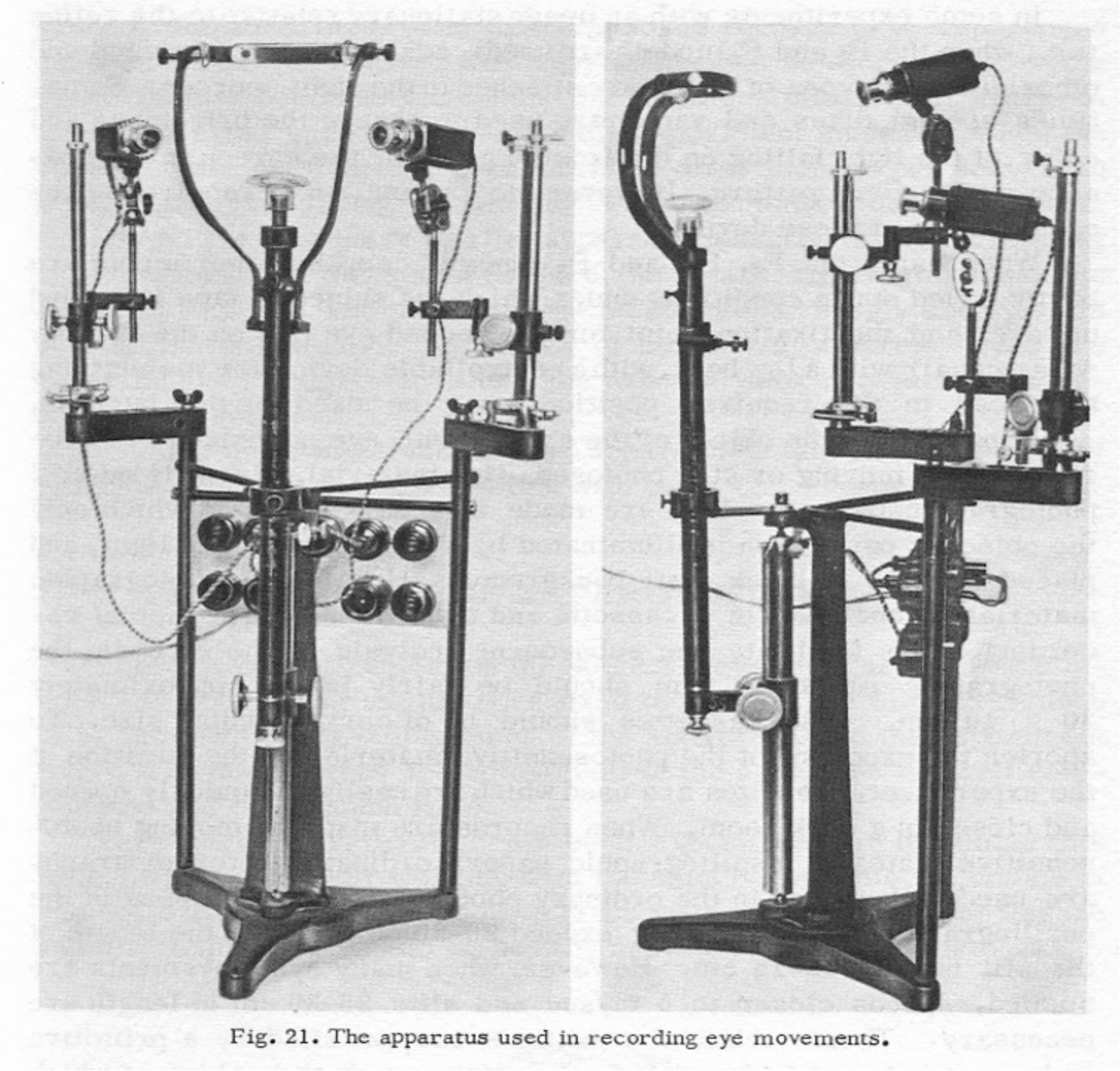
Eye tracking - applications

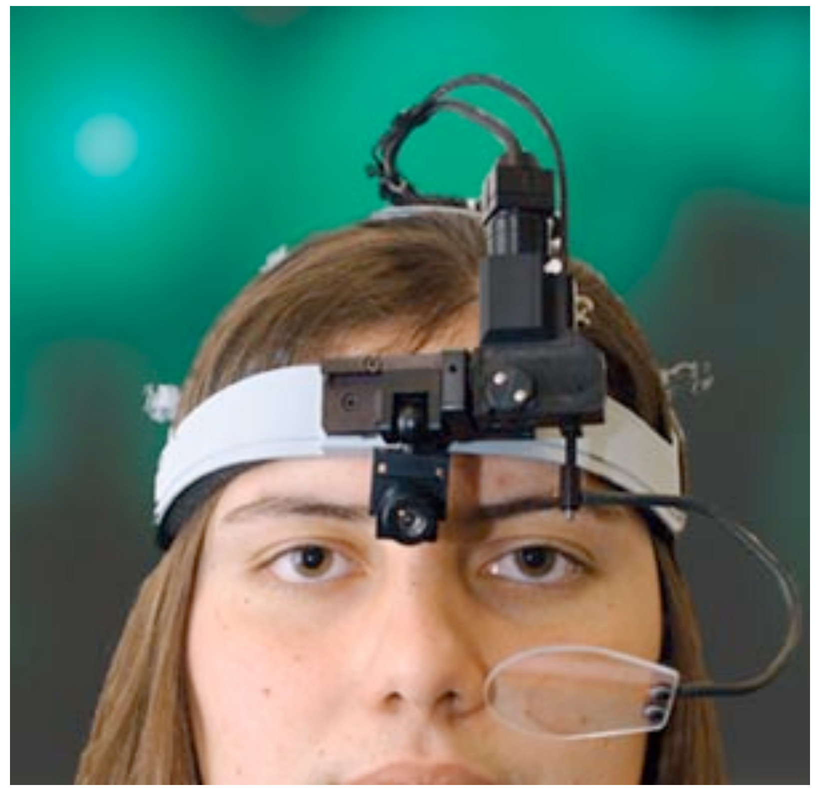
- Cognitive loading
- Neurological diagnosis
- HCI
- Language reading
- Human factors/ergonomics
- Marketing research
- Operating interfaces without other means
- Safety, game theory, aviation, other assistive applications, augmented systems, engineering, automotive, etc



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

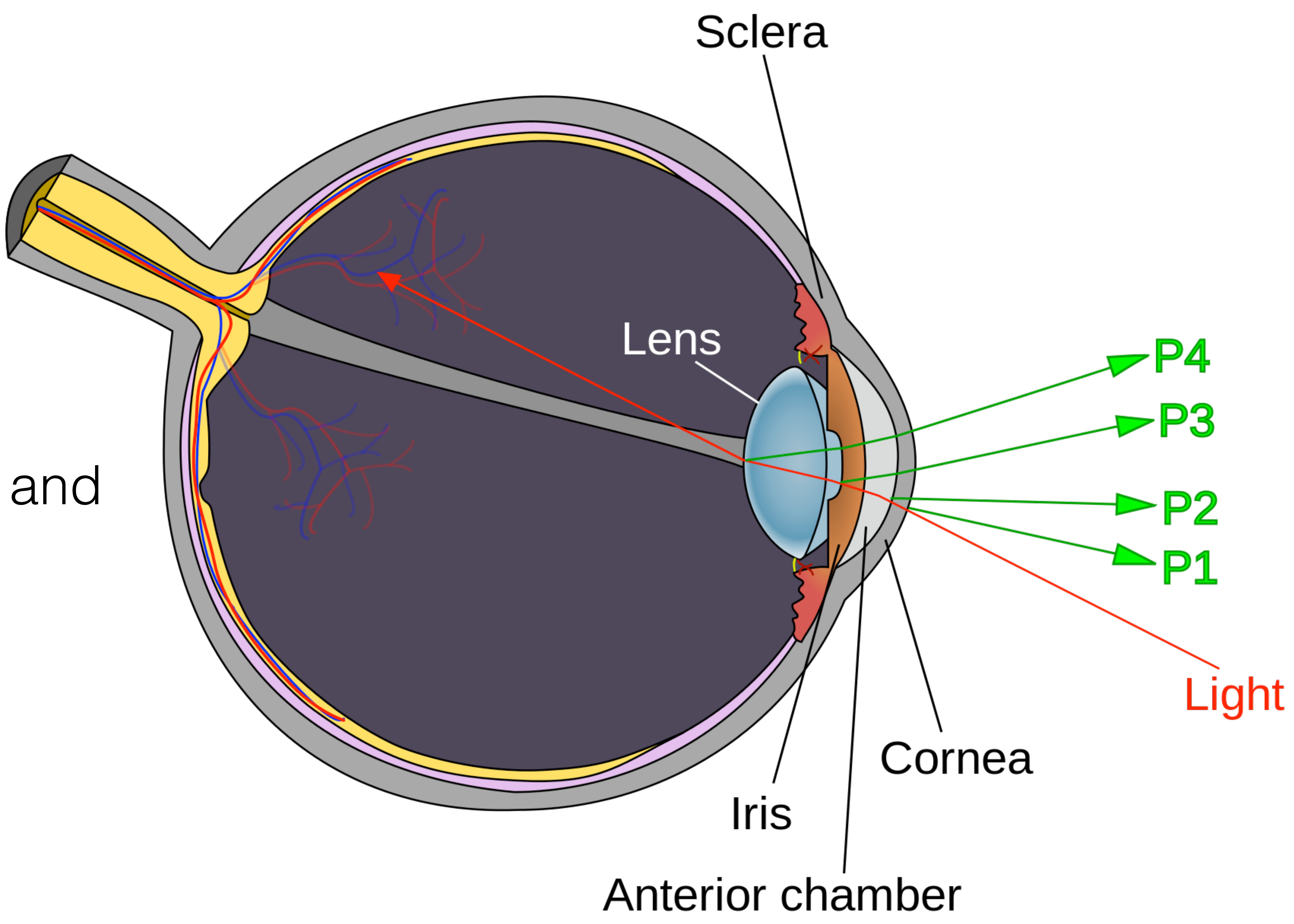




(Source: https://en.wikipedia.org/wiki/Eye_tracking)

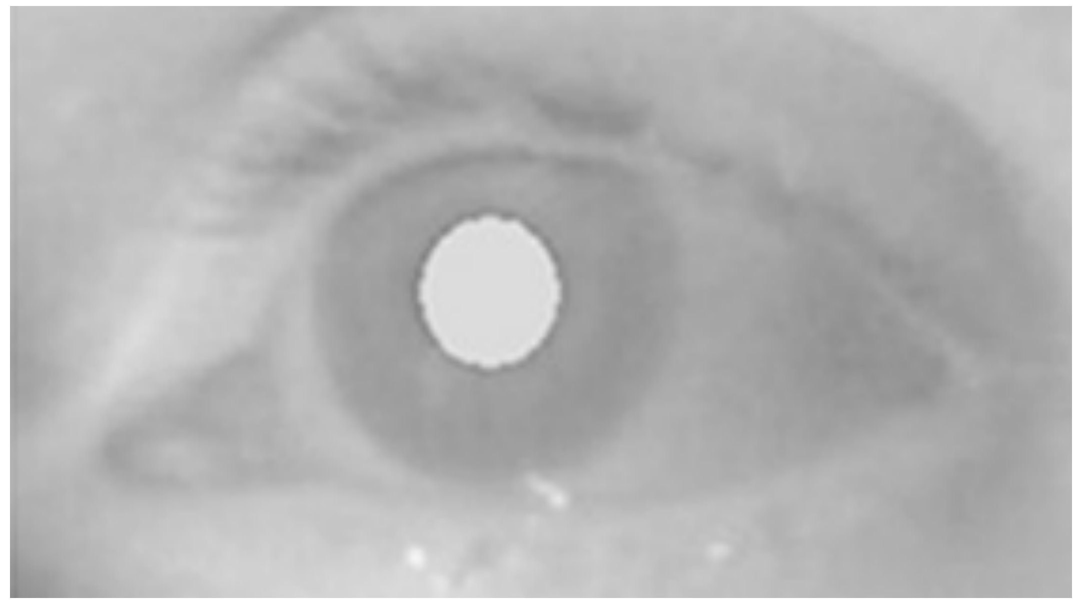
Eye tracking technology

- Most often video-based
  - Simpler, quicker to connect patient, less complications, direct measures
- Measures often infrared light reflected from eye and detected by a special camera
  - Data inferred by changes in reflections
  - i.e. Purkinje image (P1 and P4 typical)
  - or optic features like retinal blood vessel patterns

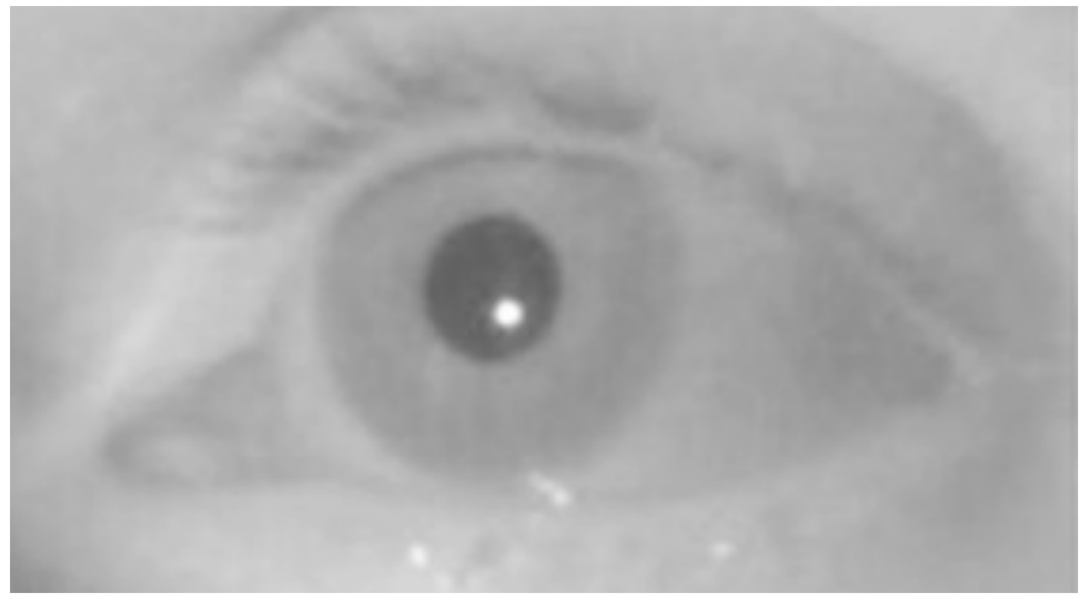


Eye tracking technology

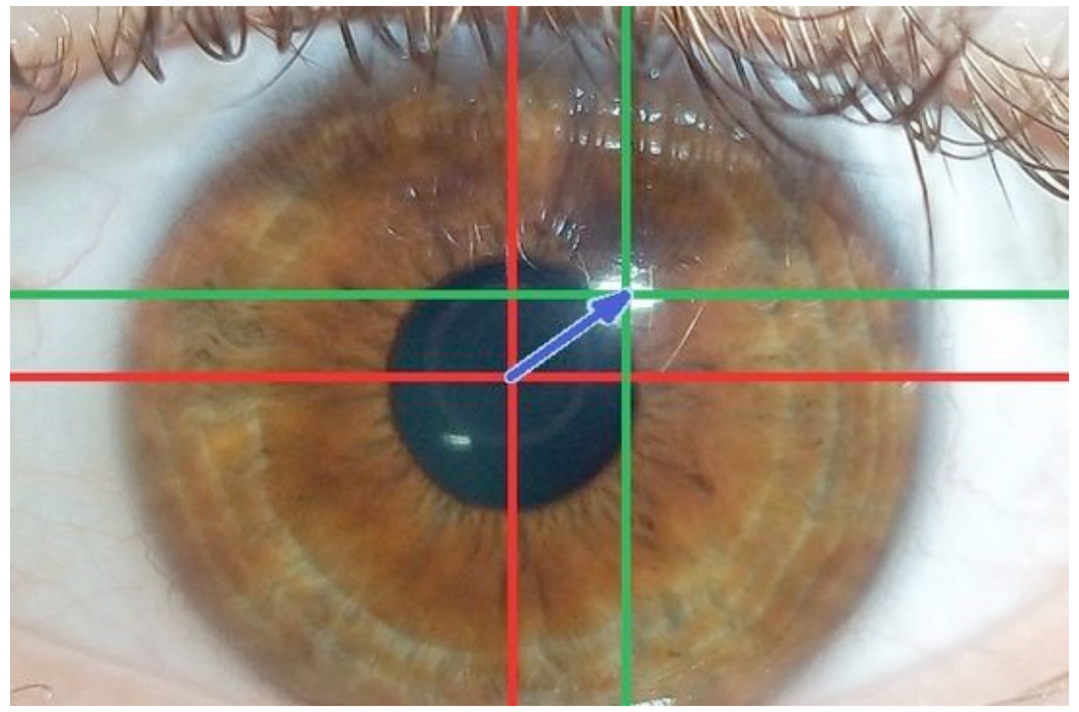
- IR/near-IR: Bright pupil,



- IR/near-IR: Dark pupil & corneal reflection



- Visible light: center of iris (red), corneal reflection (green), and output vector (blue)



Cheaper - 30Hz
But commonly > 1.3kHz

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

Eye tracking data representation

- Animated representations of a point on the interface
- Static representations of the saccade path
- Heat maps
- Blind zones maps, or focus maps
- Saliency maps

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

Eye tracking data sets and one example of raw data

- https://www.eyetracking-eeg.org/testdata.html
- https://github.com/dvlastos/eye-tracking-data
- https://englelab.gatech.edu/dataprep/eye-tracking-data.html

On to today…

So far we have discussed

- Neural Data science
- Programming
- Tools for data exploration, modeling, visualization (Python, Jupyter, Matlab, others)
- NLP
- EEG, MEG, associated analysis and tools (at a high level), other imaging
- MOCAP
- Eye tracking
- Other behavioral observations

That's a lot of data!

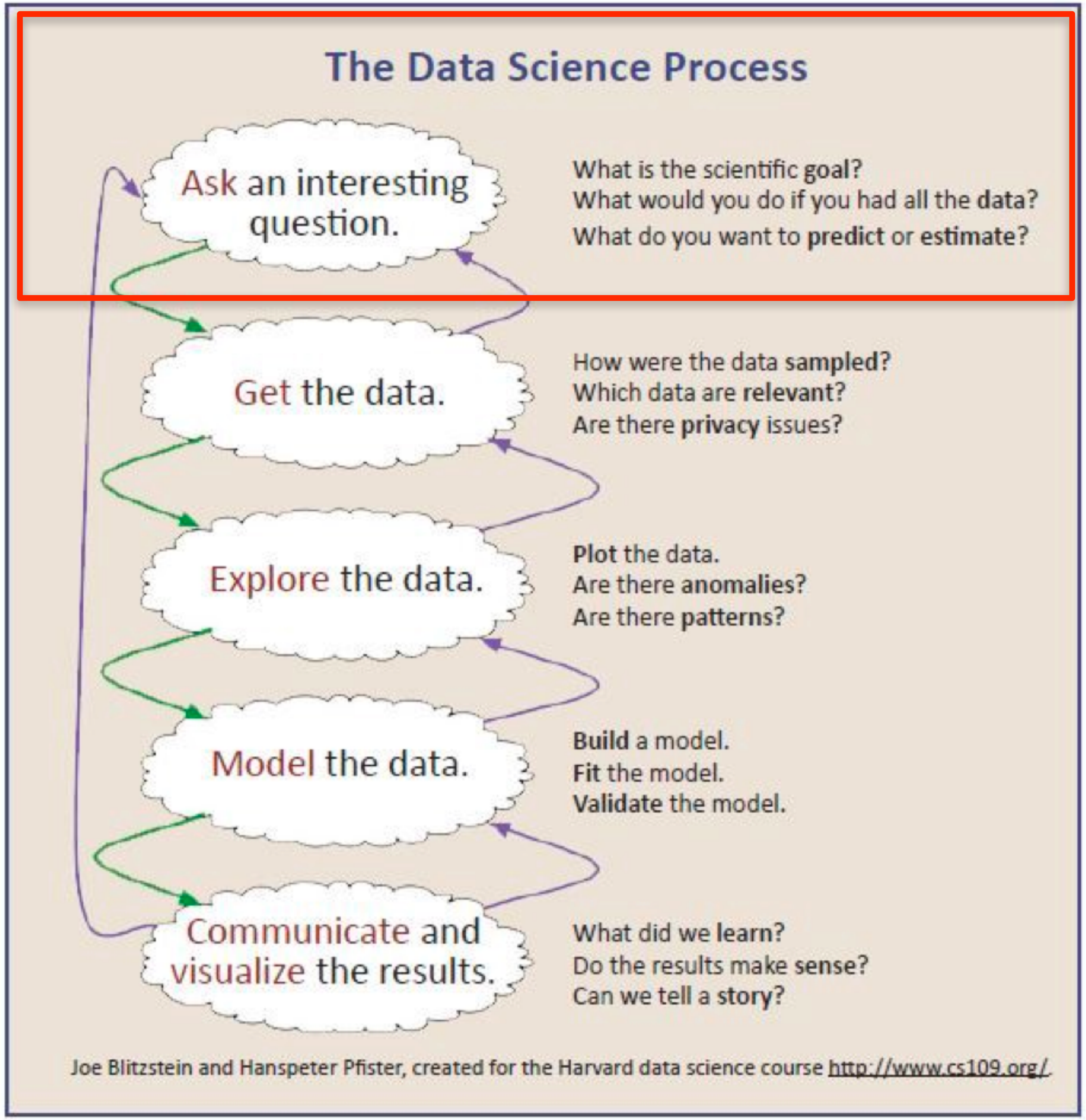
- How do you deal with it all, standardize, organize, communicate it?
- How can you talk across disciplines?
- How do you collaborate and work in teams with this?
- How can you ask questions with all that data and the results generated?

Data science questions, hypothesis generation (automated), Genes/gene expression, animal models, FAIR, Neurodata Without Borders (NWB), Brain Imaging Data Structure (BIDS), DANDI

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. It should be:

1. **Specific**,
2. Can be answered with **data**,
3. 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?

Data Science questions should...

- Be specific
- Be answerable with data
- Specify what's being measured

What makes a question a good question?

Specifying what you're going to measure is important

Examples of poor questions that leave wiggle room for useless answers:

- What can my data tell me about the brain?
- What should I do about the brain?
- How can I increase my neuroscience?

Examples of good questions where the answer is impossible to avoid:

- Does a subject's reaching trajectory change when put under a static force field? Is this change static or dynamic?
- What is the average/maximum grip strength required to manipulate a pen during writing tasks (pen and paper)?
- What is the minimum light intensity perceptible by the average subject of age range 18-24yrs in pitch black darkness for a point light at a distance of 2m?

Working toward a strong data science question

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?

Practicing asking questions…

- Could reflex be measured with brain activity

Previous questions asked during this class's projects…

Genes and text, LISC

()

- Leveraging LISC and 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)

LISC project

()

- Open source python module “Literature Scanner”
  - https://github.com/lisc-tools/lisc
- Donoghue, Thomas. (2019). LISC: A Python Package for Scientific Literature Collection and Analysis. Journal of Open Source Software. 4. 1674. 10.21105/joss.01674.
- https://www.researchgate.net/publication/336082537_LISC_A_Python_Package_for_Scientific_Literature_Collection_and_Analysis
- LISC is based on BRAIN-SCANR by Voytek (2012)

LISC- Automated methods for digesting vast information

()

- Scientific literature is vast, expanding and beyond a single researcher's ability to digest completely
- By the time an article is read, more are published
- >30M published articles as of 2019 in biomedical sciences alone!
- Automated methods for curation and digestion of literature has been explored to enhance a researcher's abilities to absorb information
- "Knowledge discovery, literature-based discovery, hypothesis generation"

LISC- Automated methods for digesting vast information

()

- Easily accessible
- Connects to several external resources through APIs
- e.g. PubMed, OpenCitations database
- Supports utilities to analyze collected data

LISC- types of data collection

()

- **Counts**: tools to collect and analyze data on the co-occurrence of specified search terms

- **Words**: tools to collect and analyze text and meta-data from scientific articles

- **Citations**: tools to collect and analyze citation and reference data

LISC- includes for supporting use cases

()

- URL management and requesting for interacting with integrated APIs
- Custom data objects for managing collected data
- A database structure, as well as save and load utilities for storing collected data
- Functions and utilities to analyze collected data
- Data visualization for plotting collected data and analysis outputs

LISC vs. Moliere

()

- LISC takes a lightweight, fast and efficient approach to hypothesis generation
- A complement for other tools like Moliere or Meta (www.meta.org)
- More customizable (LISC), tools included for efficient analysis on the results
- Connective interface to Natural Language Processing (NLP) tools such as NLTK
- Moliere/Meta better for more complex analyses

Caveats

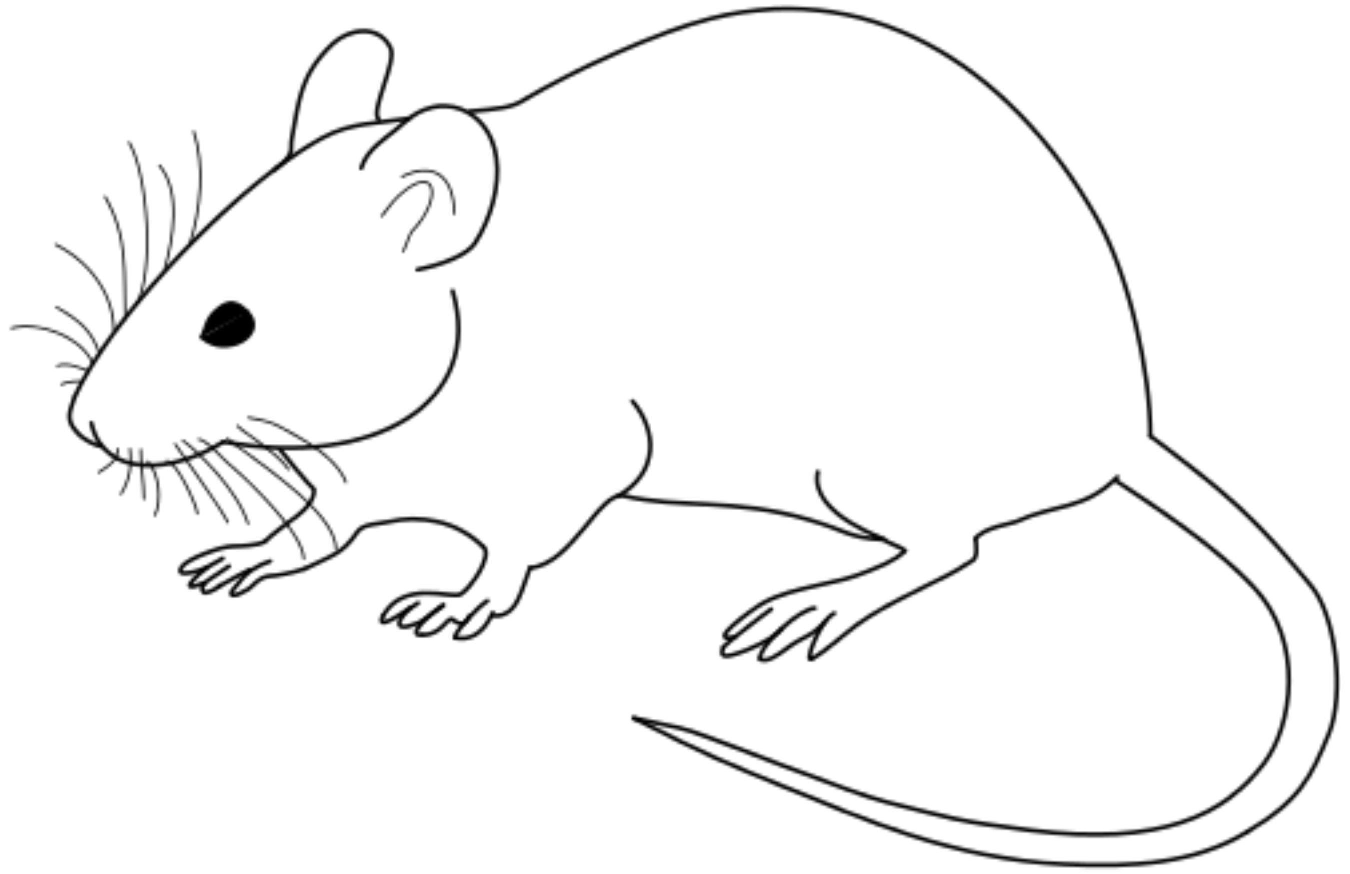
- Take care using automated systems since they don't "understand" the literature as a human does ()
- Programming biases are inevitable
  - Chatbot knowledge biases
  - Programmer biases
- Statistics can be biased
- Use with a grain of salt - it's a tool
  - ***"The hammer does not make the building" [Simpkins 2023]***

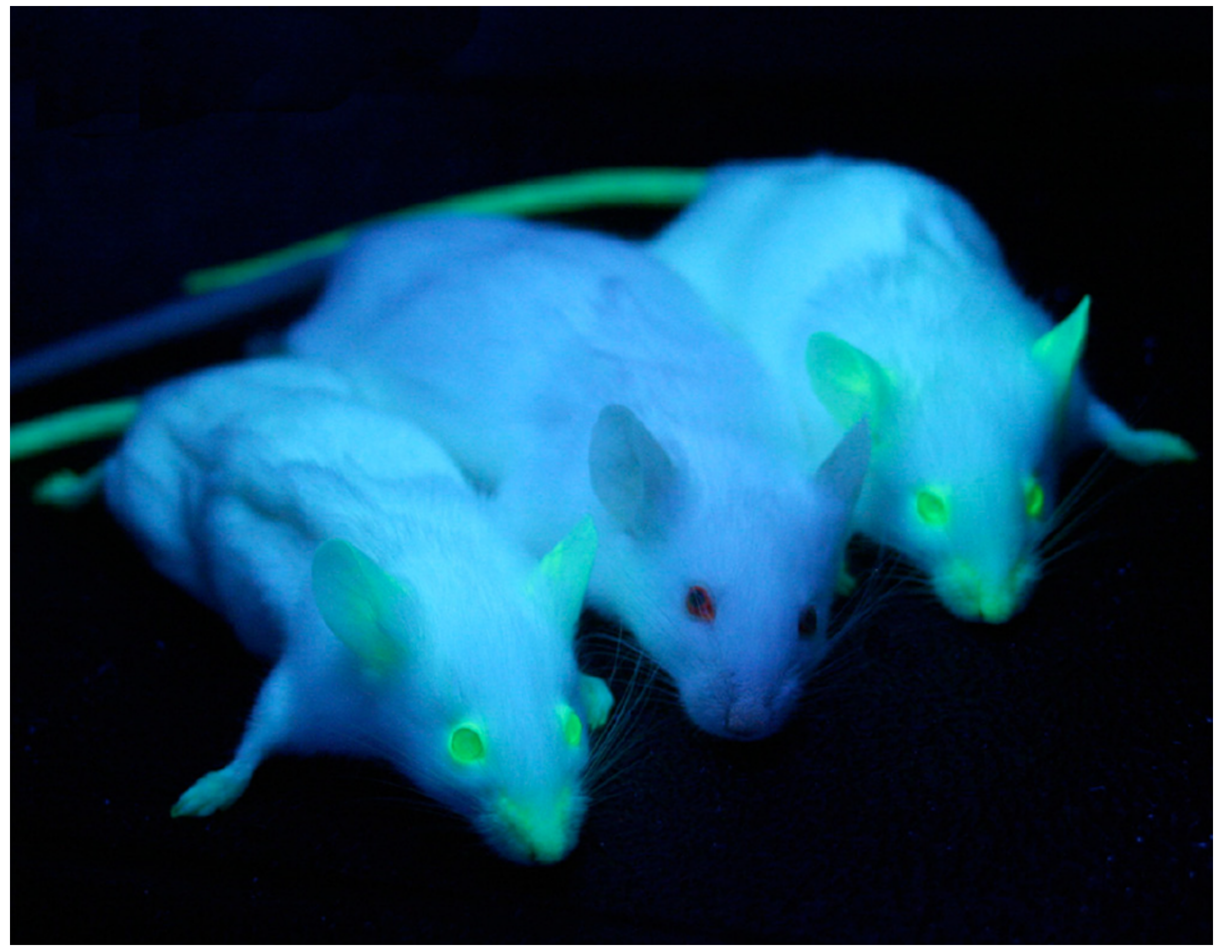
Gene expression studies

- **Gene expression definition** - *the process by which the information encoded in a gene is turned into a function. This mostly occurs via the transcription of RNA molecules that code for proteins or non-coding RNA molecules that serve other functions.*

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)

The FAIR Data Principles

- https://force11.org/info/the-fair-data-principles/
- “One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Here, we describe FAIR – a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable. The term FAIR was launched at a Lorentz workshop in 2014, the resulting FAIR principles were published in 2016.”

To be Findable

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

To be Accessible

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

To be Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

To be Re-usable

- R1. (meta)data have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

FAIR Principles Working Detailed Document

- https://force11.org/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/

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 Definition

- https://www.nwb.org/
- “**Neurodata Without Borders (NWB)** is a ***data standard*** for neurophysiology, providing neuroscientists with a common standard to share, archive, use, and build analysis tools for neurophysiology data. NWB is designed to store a variety of neurophysiology data, including data from intracellular and extracellular electrophysiology experiments, data from optical physiology experiments, and tracking and stimulus data.” [www.nwb.org]

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

NWB Introduction

- Take advantage of established techniques for processing, analysis, visualization tools
- Makes data easier to reuse - additional scientific insights
- Essential step to getting data into the DANDI archive (https://dandiarchive.org/)

Brain **I**maging **D**ata **S**tructure
(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
DANDI
WELCOME
PUBLIC DANDISETS
MY DANDISETS
ABOUT
DOCUMENTATION
HELP
NEW DANDISET
The DANDI Archive
The BRAIN Initiative archive for publishing and sharing neurophysiology data including electrophysiology, optophysiology, and behavioral time-series, and images from immunostaining experiments.
Search Dandisets by name, description, identifier, or contributor name
138
311
157 TB
b) Supported standards
AWS
NIH
The BRAIN Initiative
NEURODATA
WITHOUT BORDERS
BIDS
BRAIN IMAGING DATA STRUCTURE
NIDM
c) Analysis platform
jupyterhub
d) Python clients
Collaborator(s)
Lab Member(s)
(Meta)-Data Flow
JSON/JSON-LD, NWB, NIfTI, TIFF
User Interactions
Web Browser, Shell, API

Benefits of DANDI

- A FAIR (Findable, Accessible, Interoperable, Reusable) data archive to house standardized neurophysiology and associated data
- Rich metadata to support search across data
- Consistent and transparent data standards to simplify data reuse and software development.
  - Uses NWB, BIDS, Neuroimaging Data Model (NIDM), and other BRAIN Initiative standards to organize and search the data.
  - The data can be accessed programmatically allowing for software to work directly with data in the cloud
- The infrastructure is built on a software stack of open source products, thus enriching the ecosystem

DANDI compatibility

- Uses NWB for core data language
- “Dandisets” - DANDI datasets - collection of NWB files recorded over multiple sessions, organized together
- Viewable from a web browser
- Can interact through Jupyterhub interface for exploring, visualizing and analyzing the data stored in the archive

DANDI python client

- Organize data locally into the required structure
- Download/upload data from/to the DANDI archive

a) Web application
DANDI
WELCOME
PUBLIC DANDISETS
MY DANDISETS
ABOUT
DOCUMENTATION
HELP
NEW DANDISET
The DANDI Archive
The BRAIN Initiative archive for publishing and sharing neurophysiology data including electrophysiology, optophysiology, and behavioral time-series, and images from immunostaining experiments.
Search Dandisets by name, description, identifier, or contributor name
138
311
users
157 TB
AWS
b) Supported standards
NIH
The BRAIN Initiative
NEURODATA
WITHOUT BORDERS
BIDS
BRAIN IMAGING DATA STRUCTURE
NIDM
c) Analysis platform
jupyterhub
d) Python clients
Collaborator(s)
Lab Member(s)
(Meta)-Data Flow
JSON/JSON-LD, NWB, NIfTI, TIFF
User Interactions
Web Browser, Shell, API

DANDI archive

- **Public DANDI sets:** https://dandiarchive.org/dandiset
- **Documentation:** https://www.dandiarchive.org/handbook/10_using_dandi/

DANDI Properties

- **Data identifiers:** The archive provides persistent identifiers for versioned datasets and assets, thus improving reproducibility of neurophysiology research

- **Data storage:** Cloud-based platform on AWS. Data are available from a public S3 bucket. Data from embargoed datasets are available from a private bucket to owners only

- **Type of data:** The archive accepts cellular neurophysiology data including electrophysiology, optophysiology, and behavioral time-series, and images from immunostaining experiments and other associated data (e.g. participant information, MRI or other modalities)

- **Accepted Standards** and Data File Formats: NWB (HDF5), BIDS (NIfTI, JSON, PNG, TIF, OME.TIF, OME.BTF, OME.ZARR) (see Data Standards for more details)

Neurophysiology Informatics Challenges and DANDI Solutions

Challenges	Solutions
Most raw data stays in laboratories.	DANDI provides a public archive for dissemination of raw and derived data.
Non-standardized datasets lead to significant resource needs to understand and adapt code to these datasets.	DANDI standardizes all data using NWB and BIDS standards.
The multitude of different hardware platforms and custom binary formats requires significant effort to consolidate into reusable datasets.	The DANDI ecosystem provides tools for converting data from different instruments into NWB and BIDS.
There are many domain general places to house data (e.g. Open Science Framework, G-Node, Dropbox, Google drive), but it is difficult to find relevant scientific metadata.	DANDI is focused on neurophysiology data and related metadata.
Datasets are growing larger, requiring compute services to be closer to data.	DANDI provides Dandihub, a JupyterHub instance close to the data.
Neurotechnology is evolving and requires changes to metadata and data storage.	DANDI works with community members to improve data standards and formats.
Consolidating and creating robust algorithms (e.g. spike sorting) requires varied data sources.	DANDI provides access to many different datasets.

DANDI archive

- https://elifesciences.org/articles/78362
- **Oliver Rübel, Andrew Tritt, Ryan Ly, Benjamin K Dichter, Satrajit Ghosh, Lawrence Niu, Pamela Baker, Ivan Soltesz, Lydia Ng, Karel Svoboda, Loren Frank, Kristofer E Bouchard (2022) The Neurodata Without Borders ecosystem for neurophysiological data science eLife 11:e78362**
- https://doi.org/10.7554/eLife.78362