



ALLEN INSTITUTE *for*
NEURAL DYNAMICS

OPEN SCIENCE at the ALLEN INSTITUTE FOR NEURAL DYNAMICS

11/14/2022

- Mission & data
- Open science and data sharing

Allen Institute for Neural Dynamics - Mission

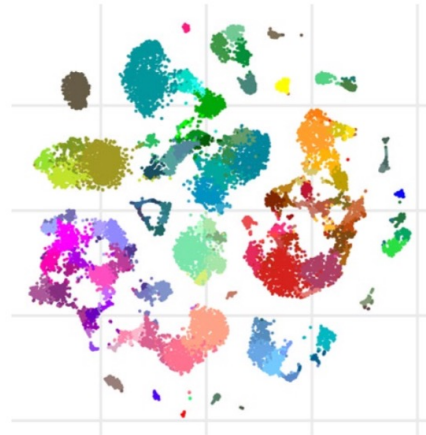
Discover how the brain's neurons produce our emotions, memories and actions.

Answers will be in terms of neural activity in defined neuron types interacting across the whole brain and body.

This requires next-generation neurotechnologies.

Knowledge, data, and tools will be widely shared, to facilitate science elsewhere and to support the development of therapies for brain disorders.

Transcriptomic clusters



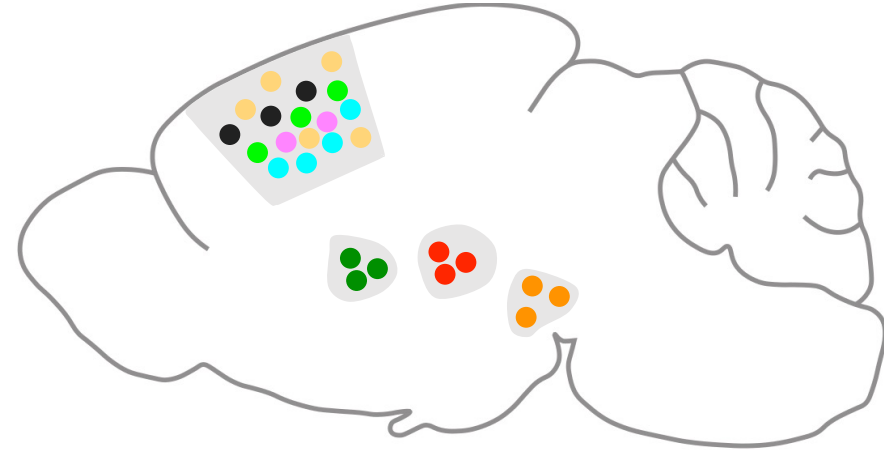
BICCN; AIBS



Cell types



Spatial transcriptomics → map of cell types



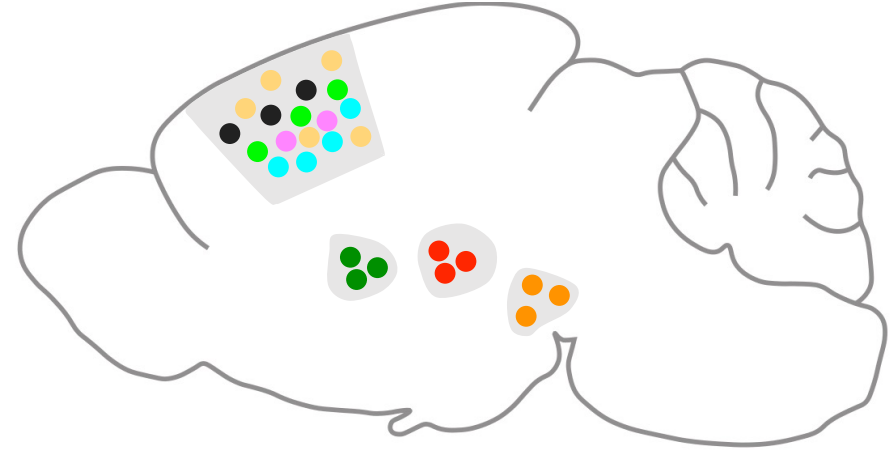
Transcriptomic clusters



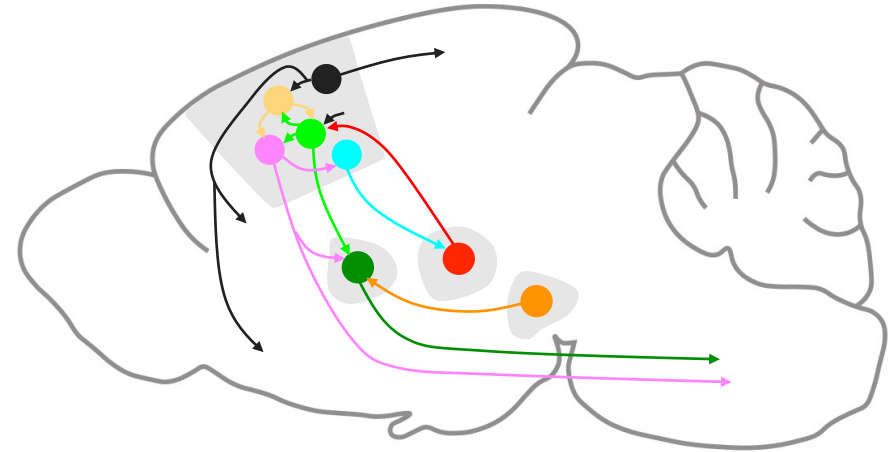
Cell types

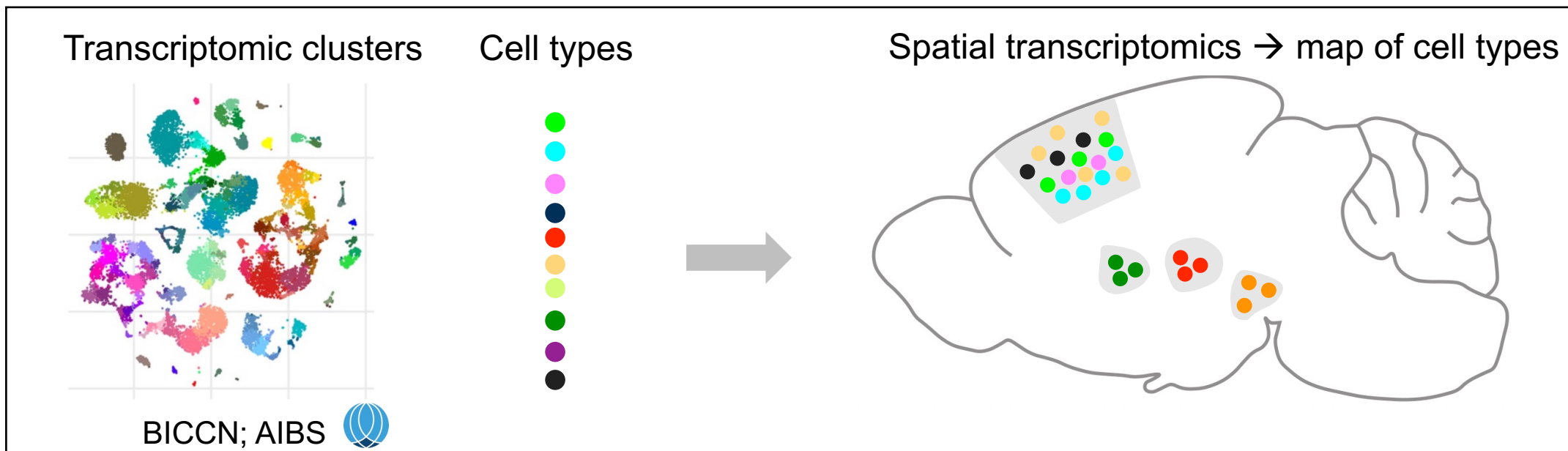


Spatial transcriptomics → map of cell types

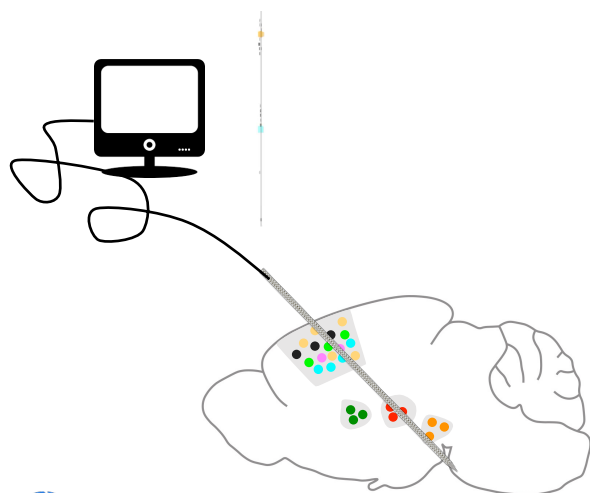


Structure and connectivity → neural circuits

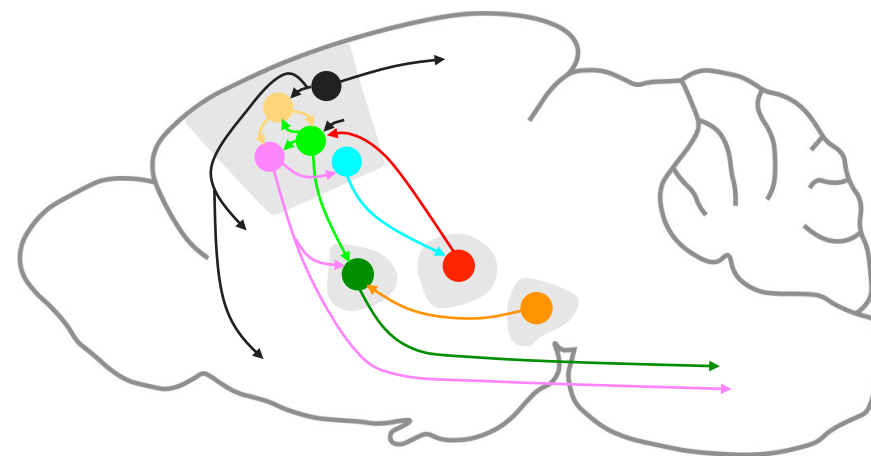




Neural signals → coding of information



Structure and connectivity → neural circuits

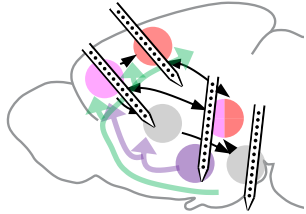


AIND organization: **Groups**

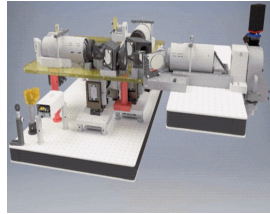
Behavior



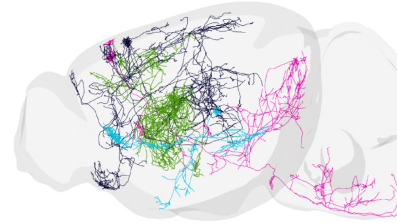
Neural dynamics
Ephys



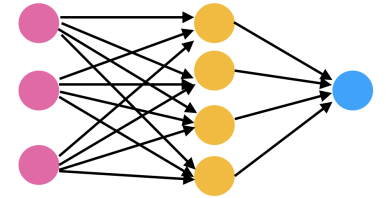
Neural dynamics
Ophys



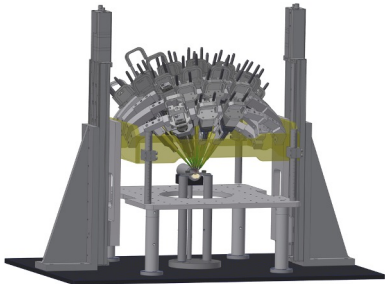
Mapping brain-wide
neural circuits



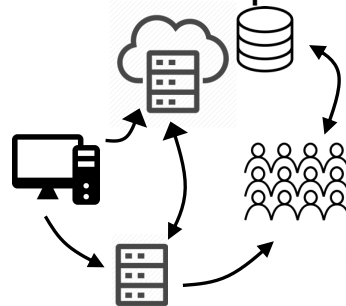
Machine-learning
and theory



Instrumentation



Scientific computing



Project mgt & administration



David Feng



Saskia de Vries



**Open science
increases scientific
collaborations and sharing
of information for the benefits
of science and society**

**OPEN
SCIENCE**

**makes multilingual scientific
knowledge openly available,
accessible and reusable for
everyone**

**opens the processes of scientific
knowledge creation, evaluation and
communication to societal actors
beyond the traditional scientific
community.**

closed

transparent

access

***Research
question***

Design &
build

Collect data

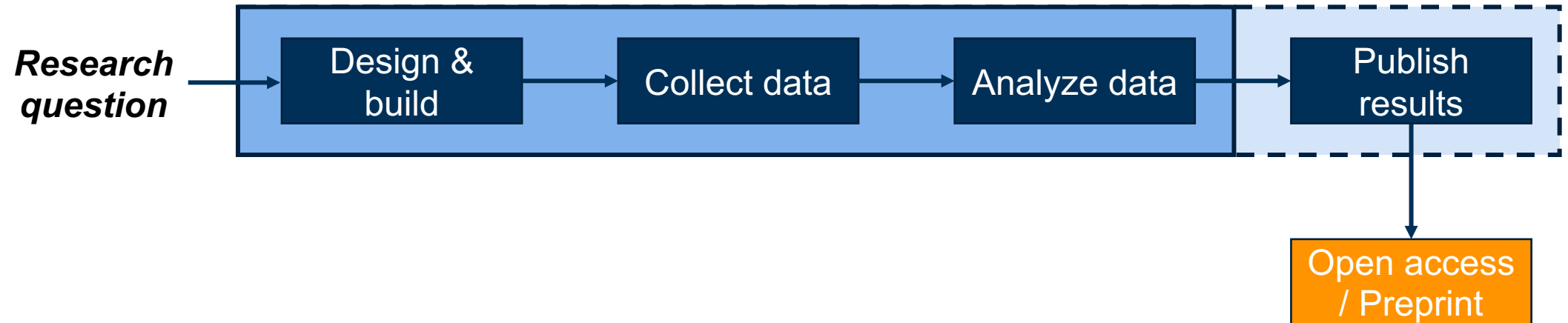
Analyze data

Publish
results

closed

transparent

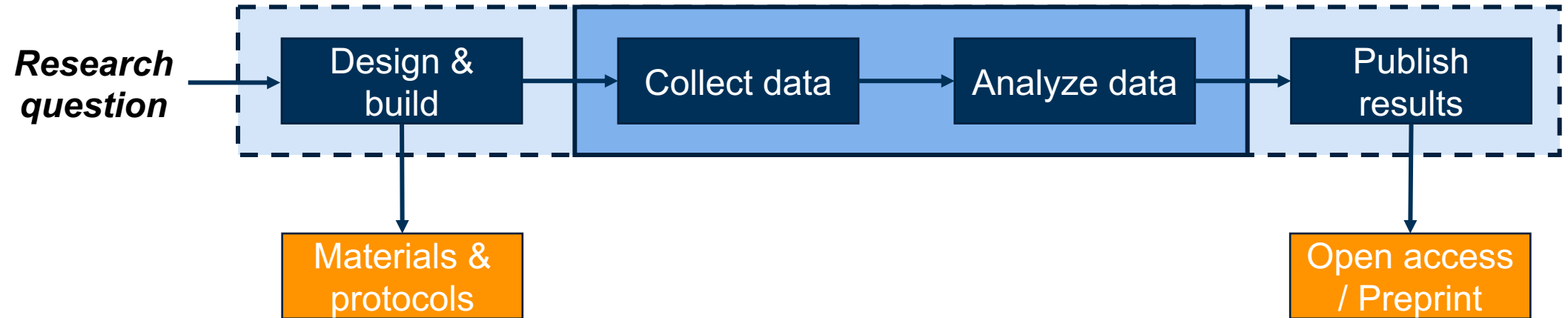
access



closed

transparent

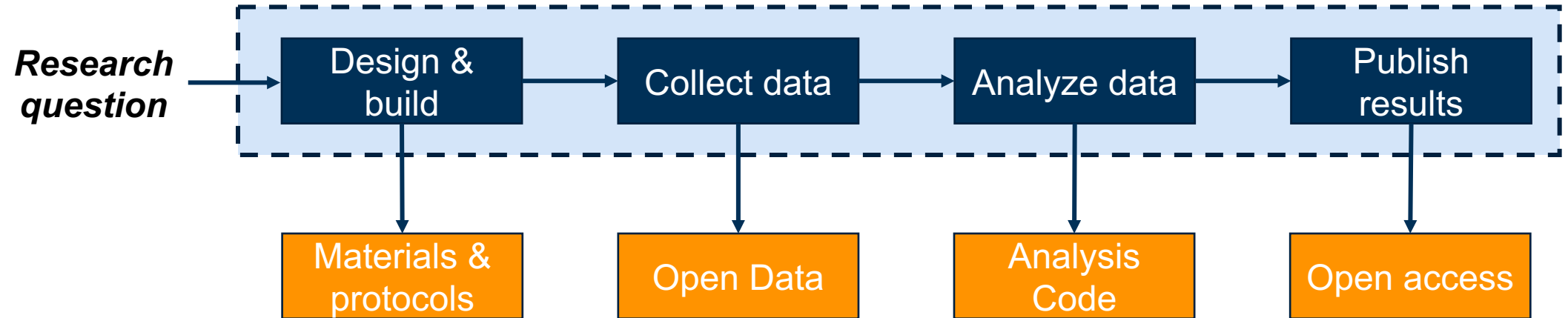
access



closed

transparent

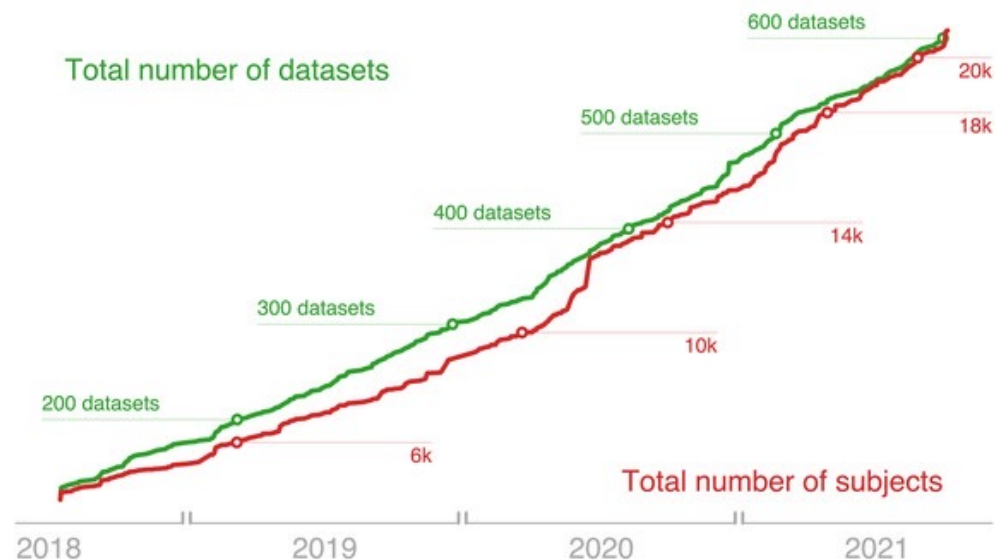
access



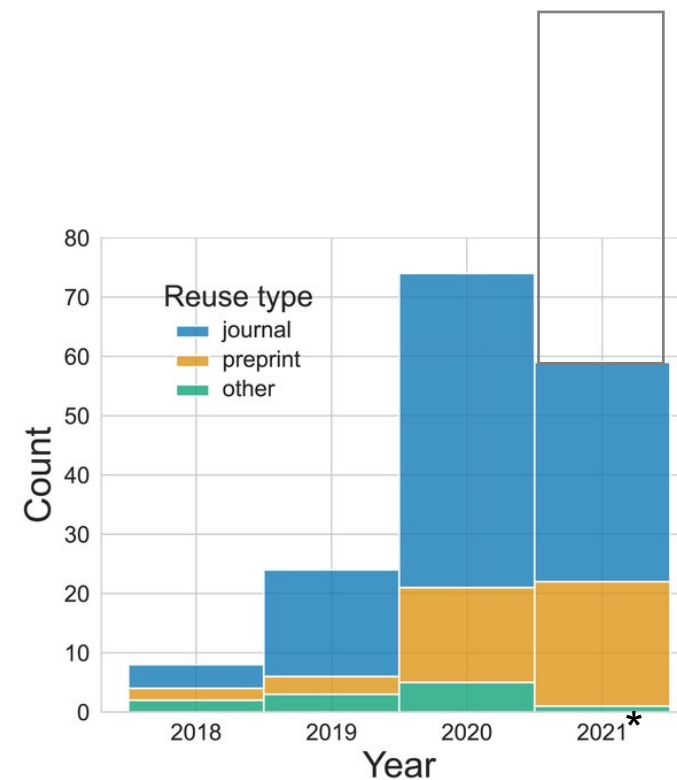
Sharing:
Tools, data, code, research findings

Data re-use in neuroscience: MRI

OpenNeuro datasets



Reuses of OpenNeuro datasets



Markiewicz *et al.*, 2021

Data re-use: cellular neurophysiology

CRCNS.org

CRCNS – Collaborative Research in Computational Neuroscience – Data sharing

Home News Data Sets Download Marketplace Forum About Publications Other Resources Hosted Projects NWB project Course Contribute

You are here: Home → Data Sets → Methods → cai-1 → About cai-1

Navigation

- Visual cortex
- Auditory cortex
- Inner ear/cochlea
- Frontal cortex
- Prefrontal cortex (PFC)
- Parietal cortex
- Motor cortex

About cai-1
Information about the data including data types, experiments, format of the data.

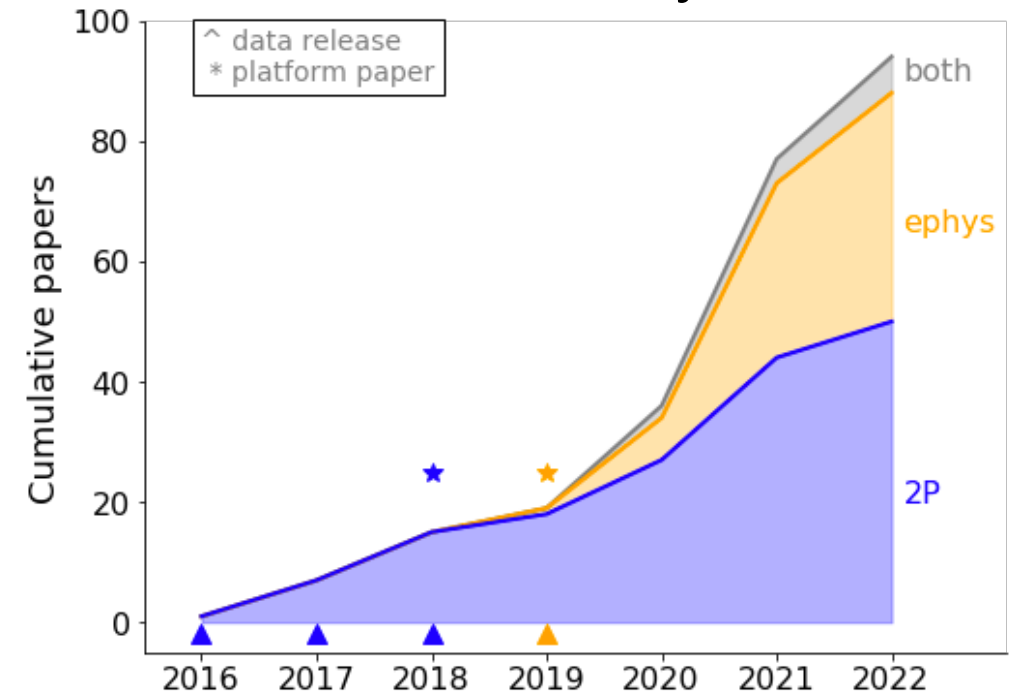
Summary of the data
This data set contains hard won calibration data for in vivo imaging data using a variety of genetically encoded calcium indicators. Spikes were recorded independently for imaged neurons. Specifically, the data set contains simultaneous imaging with loose-seal cell-attached recording in GCaMP expressing neurons. The data are described in the following two publications:

Akerboom, et al., (2012). Optimization of a GCaMP calcium indicator for neural activity imaging. *The Journal of Neuroscience* 32(40), 13819–13840. doi:10.1523/JNEUROSCI.2601-12.2012

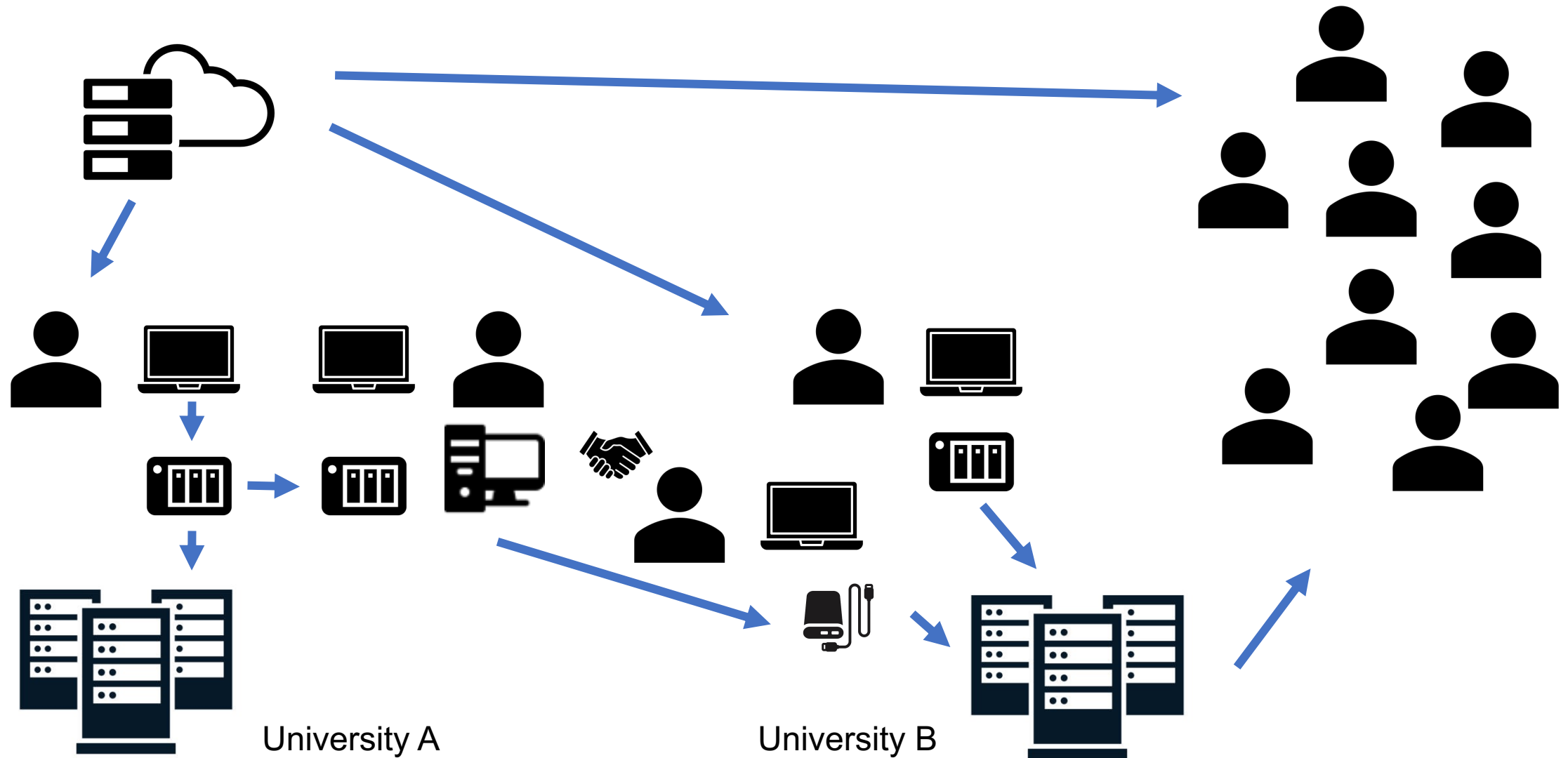
Chen, et al., (2013). Ultra-sensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458), 295–300. doi:10.1038/nature12354

Data reused in dozens of studies focused on the interpretation of calcium imaging data

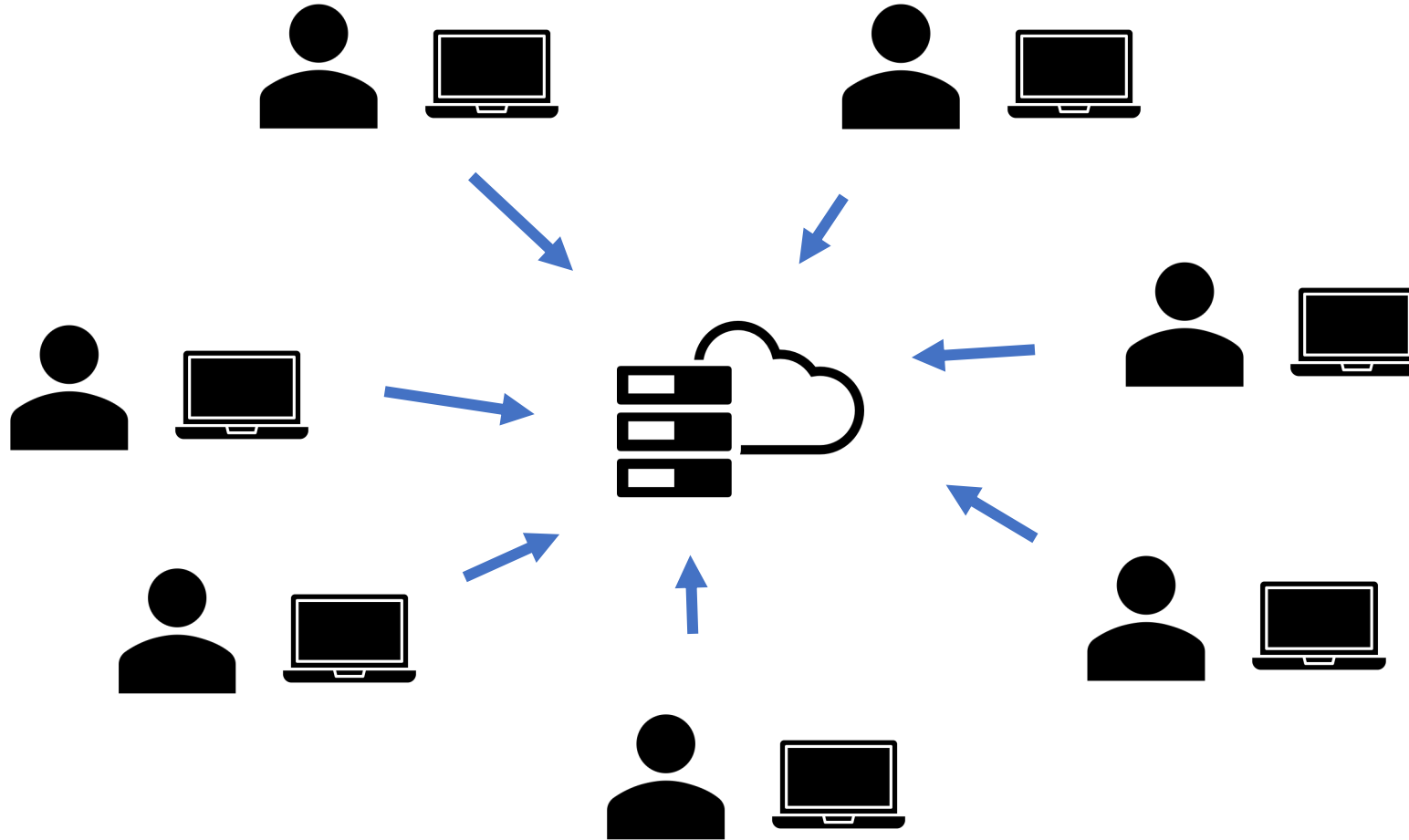
Allen Brain Observatory / DANDI



Moving data is inefficient



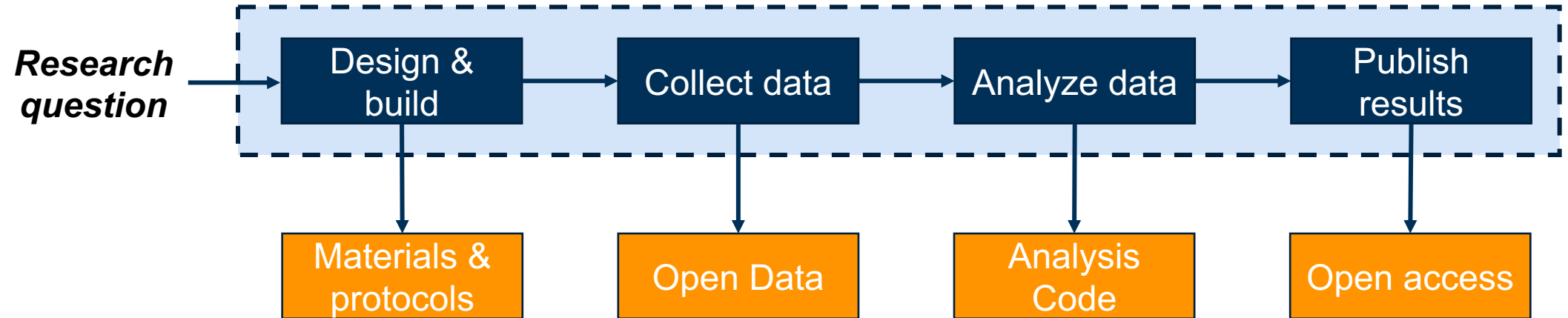
Bringing the community to data in the public cloud is more efficient and powerful



closed

transparent

access

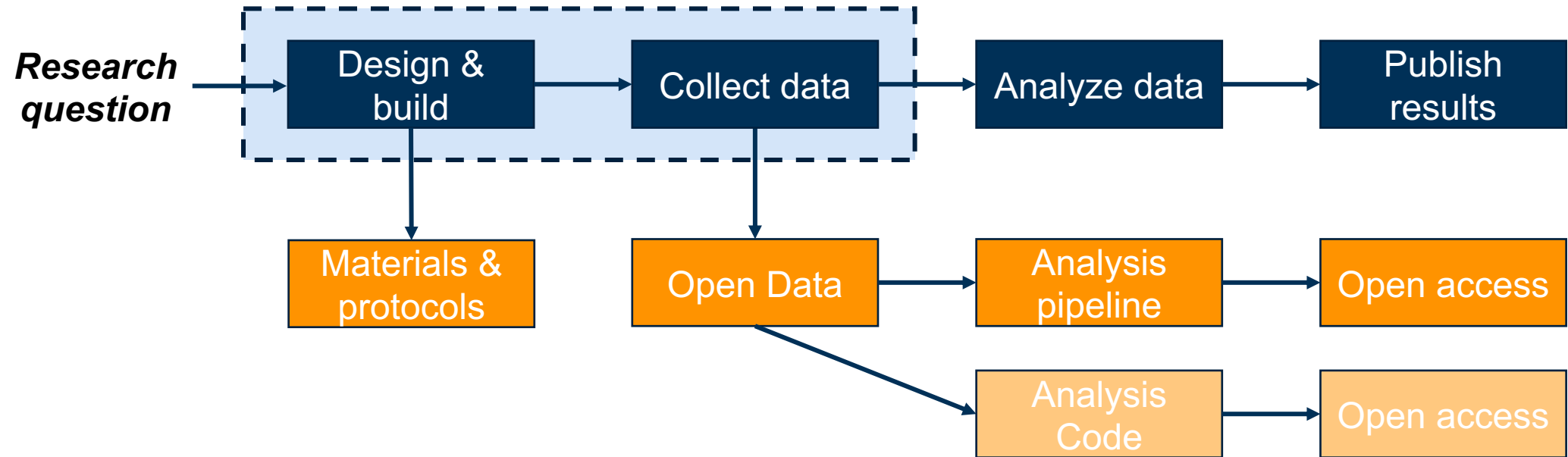


Sharing:
Tools, data, code, research findings

closed

transparent

access



Public cloud data lake architecture

Goals:

- Reproducibility
- Flexibility
- Collaboration
- Dissemination
- Inclusion

Challenges

- Data rate, size (**> 500 TB/w, 2 GB/s**)
- Metadata complexity
- Rate of change (science, tools, standards)
- Cutting edge hardware, software

Approach:

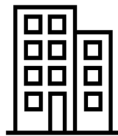
- Minimize on-prem footprint
- Bring scientists to cloud data pipelines
- Accelerate, simplify data ingestion
- All data shareable the day of acquisition
- Use scientist-friendly cloud tools
- If possible buy/use, don't build

Acquisition



- Ephys
- Imaging
- Behavior
- Ophys
- ...more

Manage the labs
electronic lab notebooks
workflow management
separate schemas per lab



On-premises

Data Lake

cloud bucket + CODE OCEAN

Data Transfer



compress + upload

Ecephys_625463_2022-10-06_10-14-25_sorted-Ks2.5	
Ecephys_625463_2022-09-28_16-34-22_sorted-Ks2.5	
Ecephys_625464_2022-10-06_11-22-45_sorted-Ks2.5	
Ecephys_625464_2022-09-29_16-51-32_sorted-Ks2.5	
Ecephys_625463_2022-09-29_15-26-36_sorted-Ks2.5	
Ecephys_639379_2022-10-25_15-10-27	
SmartSPIM_597305_2022-09-27_00-07-58	
SmartSPIM_623711_2022-10-27_16-48-54	

Self-describing file sets
one per acquisition or analysis
JSON metadata validation
searchable via metadata



Cloud

Ad-hoc Analysis



CODE OCEAN

Scalable, reproducible
no-code docker, git
elastic, on-demand hardware

Analysis Pipelines



Automated, reproducible
results go back into lake
launched on demand

Analysis Views



Data Summaries
easy to build (avoid SQL)
updated automatically

Progress

- **Compression → save \$**
- **Upload throughput**
 - Multi-node SLURM jobs
 - 1-2GBps
- **Metadata schema**
 - Inspiration from NWB, DANDI, HCA, BIDS
 - Need more!
 - Database agnostic, versioned

Pain Points

- Poor support for cloud-friendly formats
 - e.g. OME-Zarr, NWB-Zarr
- High-value desktop applications not supported
 - e.g. imagej, napari, phy
- Metadata standardization still very limited
- Cloud archives prefer polished data
- Expectation of free download/compute

Is Open Science Free?

Libre ("free speech") != Gratis ("free beer")

Academia expects both:

1. Download any data locally
2. Run any analysis (maybe on HPC)



Large, hidden, subsidized costs
(network, storage, compute, admin)

Cloud pay-as-you-go makes costs very clear

Science must be Libre, but it is never Gratis.

The cloud is a better model, but who pays?



Questions?