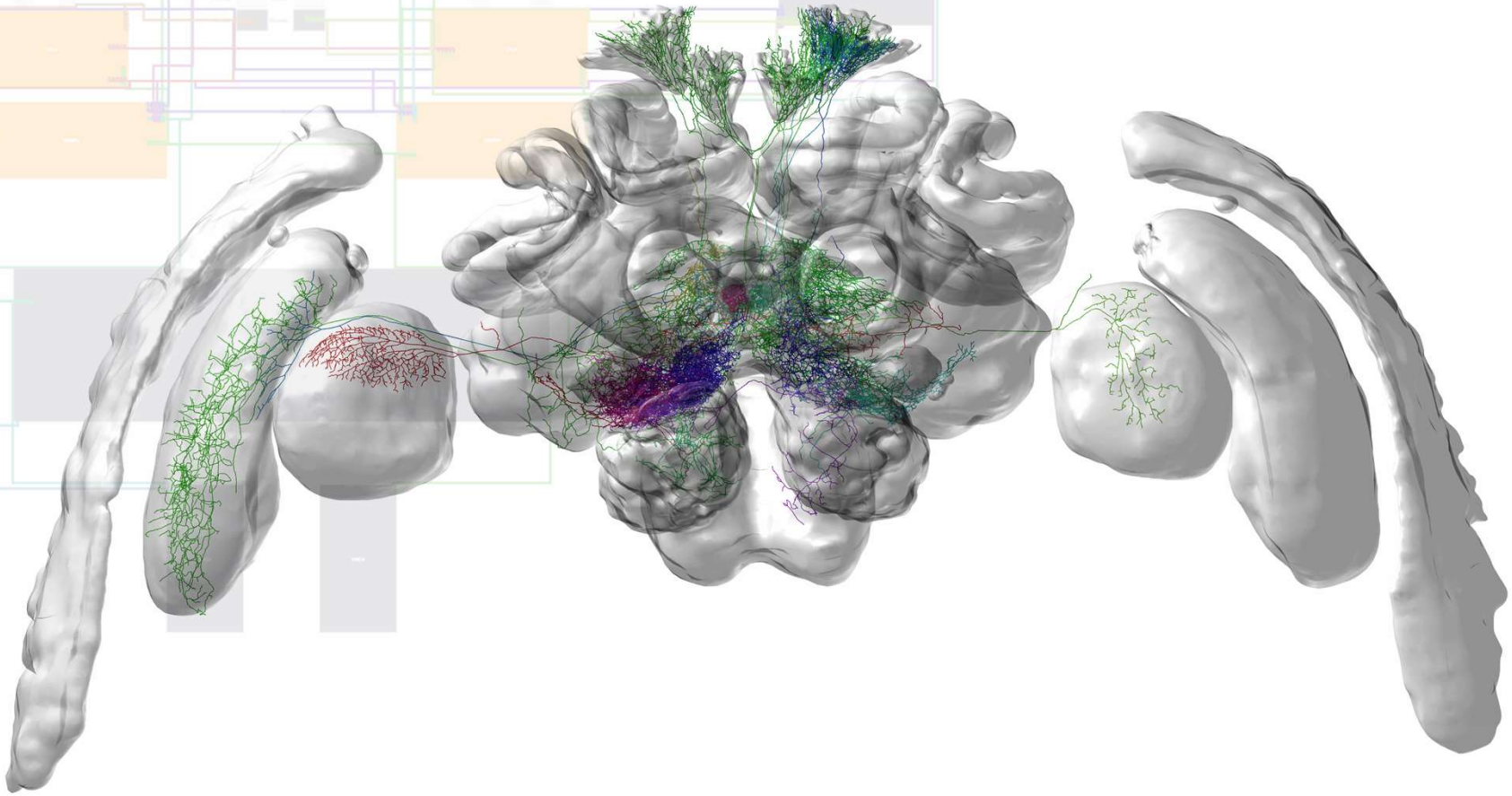
Data Management Plans

A neurobiologist's perspective

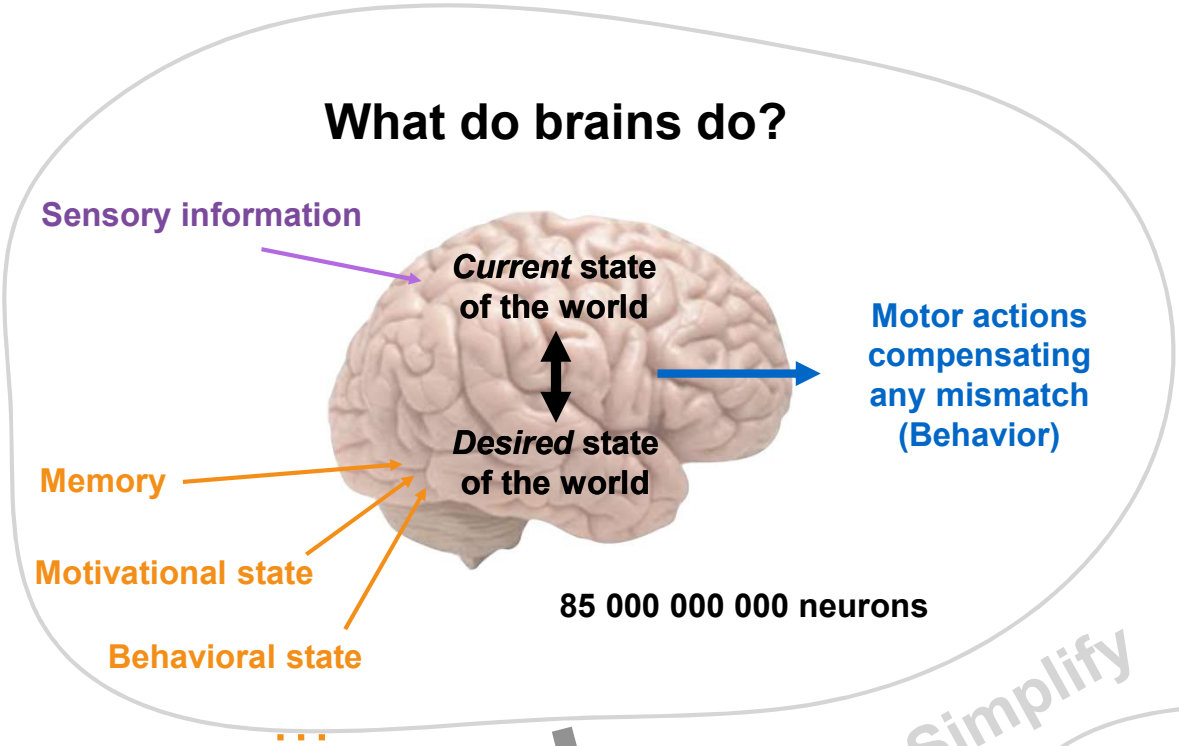
Stanley Heinze, Lund University



ERC Starting grant: The Neurobiological Basis of Navigational Decisions

European Research Council

What do brains do?



All behavioral decisions = what do I do next?

Navigational decisions = where do I go next?

Simplify

...

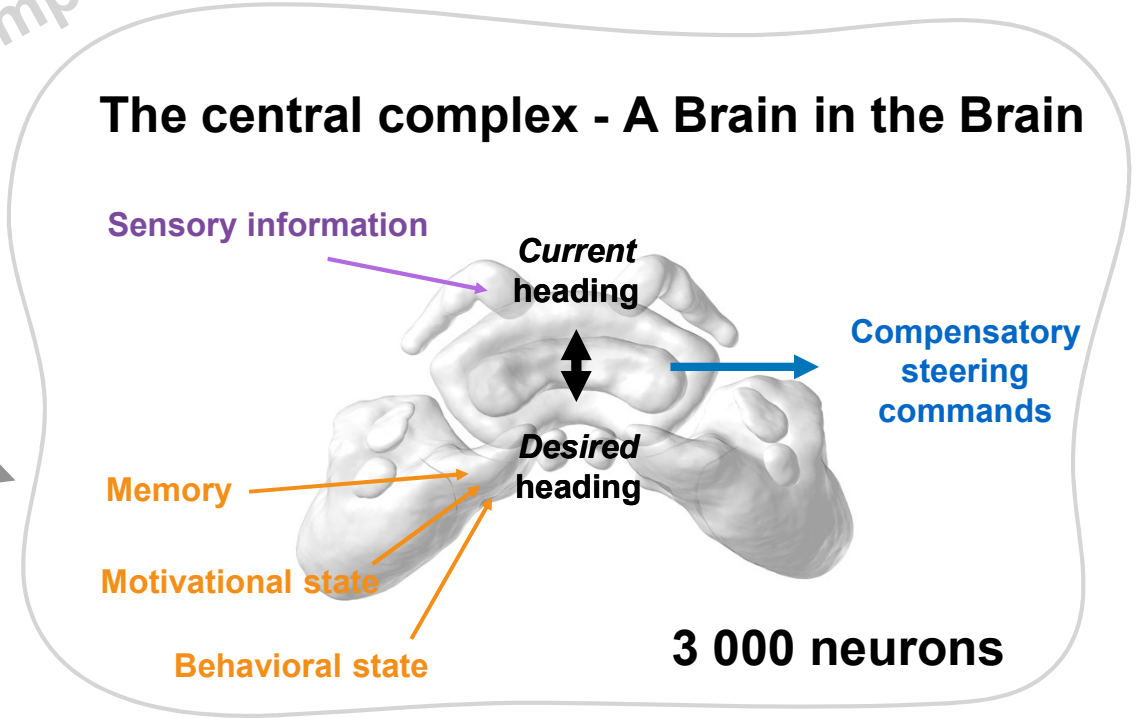
Human

Animals

Insects

One insect brain region

The central complex - A Brain in the Brain



What is a Data Management Plan (DMP)?

Before starting a series of experiments, have a concrete plan about what to do with the data resulting from the experiments.

ERC template asks how to address five points:

1. MAKING DATA FINDABLE
2. MAKING DATA OPENLY ACCESSIBLE
3. MAKING DATA INTEROPERABLE
4. INCREASE DATA RE-USE
5. ALLOCATION OF RESOURCES and DATA SECURITY

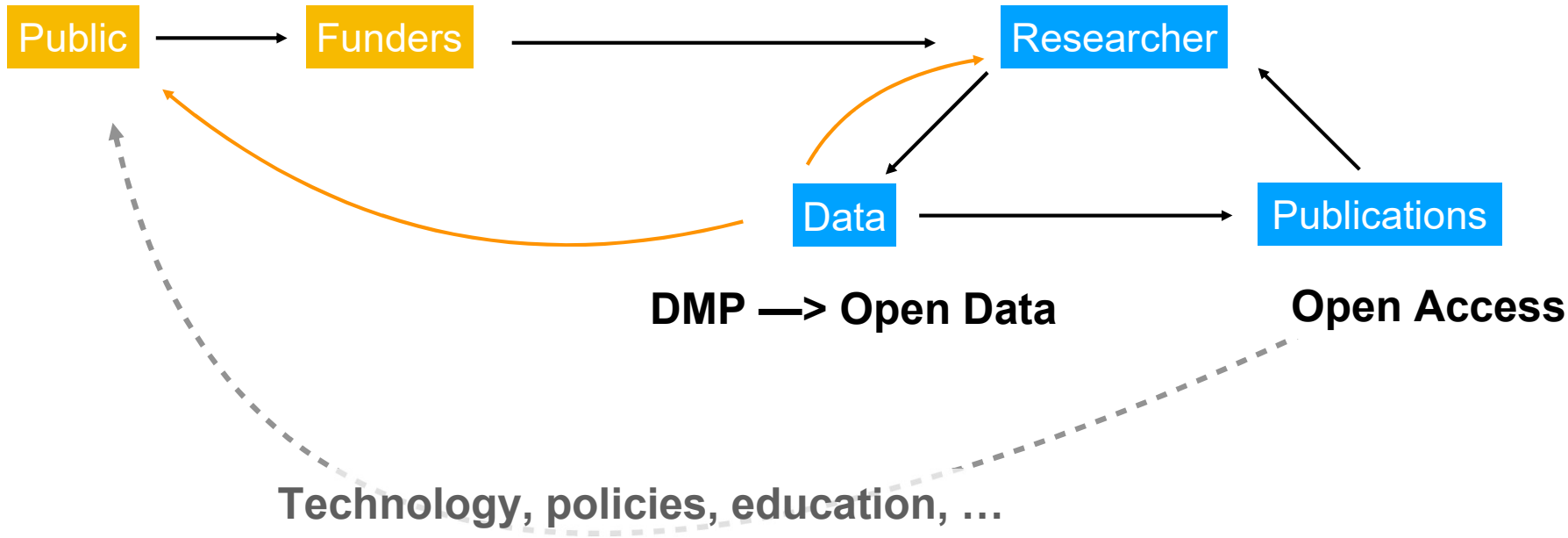
This all makes sense and sounds reasonable, but remains a bit fuzzy

Why are DMPs so low on researchers priority lists?

- additional effort
- no credits

Mandatory DMPs are now common for most grants.

—> why is this good?



Accountability, scientific integrity

Data reuse within and across research fields

More efficient use of resources

Organizing data has immediate benefits to the researcher

- The researcher knows where all data are
- There are no conflicting copies of files generated by several storage places
- Data areless likely lost
- A system for file names etc. is created before the data is generated (minimizing confusion)

- Group members know where data are, enabling better collaborations within the group
- As data formats are known in advance, data analysis pipelines can be developed early
- Automatic analysis is facilitated

- Organized data can be moved to public repositories without much additional effort

—> Fewer mistakes
—> More accountability
—> Higher productivity

What?

Step 1: What will my data be?

Electrophysiology —> Large files containing voltage recordings (specific formats, e.g. smr, matlab or text format)

Anatomy —> Large confocal image series of each neuron injection; image series from reference brains; 3D surfaces of brain areas, label fields, 3D skeletons of single neurons

Connectomics —> Gigantic image series of serial section electron microscopy work (terabytes of data) with associated skeletons of thousands of neurons.

Behavioral experiments —> High speed videos of bee behavior and analyzed flight tracks.

Analysis and computational modeling —> software code

File formats? —> mat, tiff, smr, swc, txt, am

Type of data? —> Images, text, videos, sound

Size? —> MB, GB, TB?

Where?

Step 2: Where will I store my data?

Upon acquisition:
- local computer
- notebooks
- camera (SD card)

Analysis stage
- local computer
- external hard drives
- NAS server (with data redundancy)
- cloud

Post-publication
- repository
- local backup (NAS, cloud, hard drives)

Accessibility
(who can access from where?)

Safety
(can data be lost?)

Security
(can data leak out?)

How?

Step 3: How do I store my data? Make them FAIR

Findable → Unique identifiers in repository, persistence, visibility
Accessible → Openly accessible (not behind paywall)
Interoperable → Standard data formats, proper annotation
Reusable → Open licenses, quality assurance, maintenance

Find a suitable repository

General repositories:
- *Zenodo* (https://zenodo.org):
- *Dryad* (https://datadryad.org):
- *Figshare* (https://figshare.com):
- *The Dataverse Project* (https://dataverse.org):
- *Open Science Framework* (https://osf.io

Field specific repositories:

e.g. Flybase, WormBase, MonarchBase, Genome databases, Image data repositories

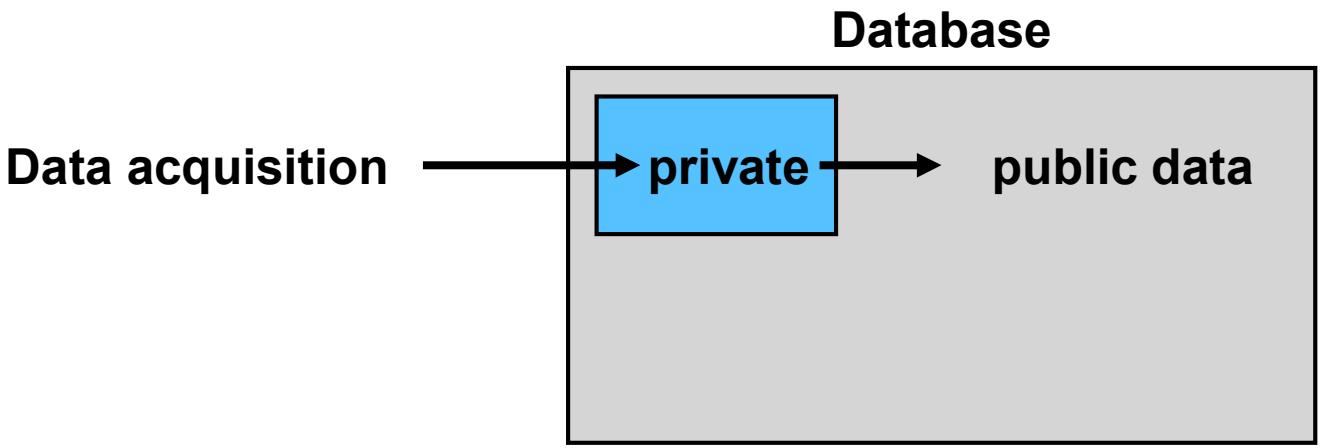
We created our own repository

www.insectbraindb.org

Dual data repository and management tool

Idea:
Data formats and structure during analysis stage should ideally be identical formats and data structure in final repository

Database software
- helps with online upload of data
- facilitates annotation by providing automatic meta-data
- automatic cross links to other public data
- provides elaborate visualization tools
- *allows to keep data private until explicitly published in the database*
- *private data can be shared between users*



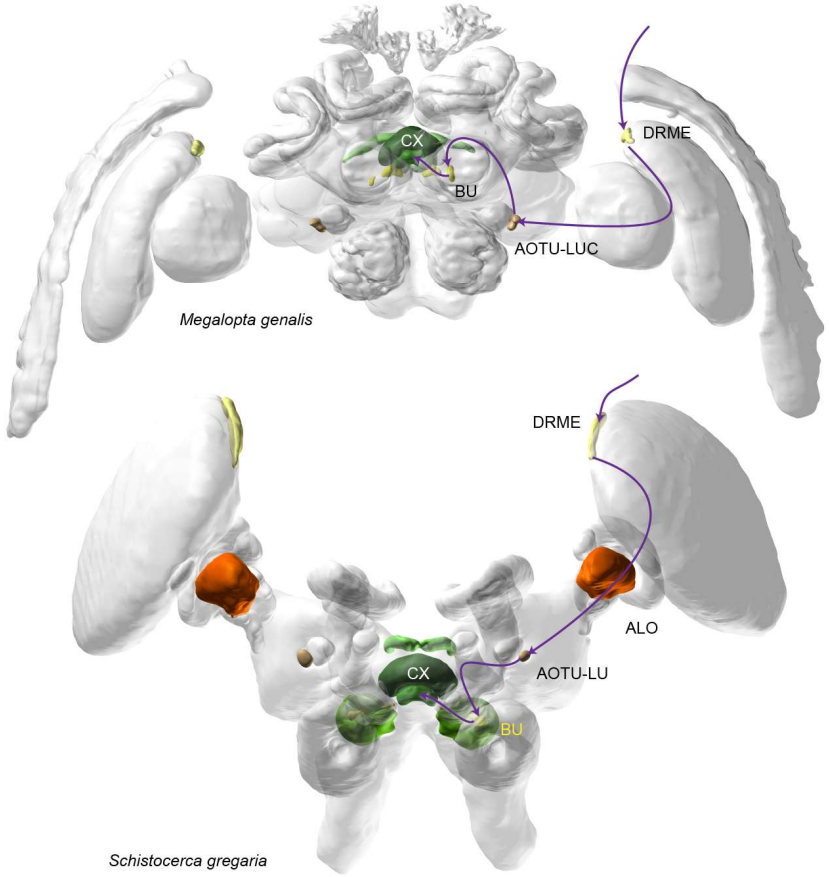
Pre-publication and post-publication data are in the same space, using same structure and formats

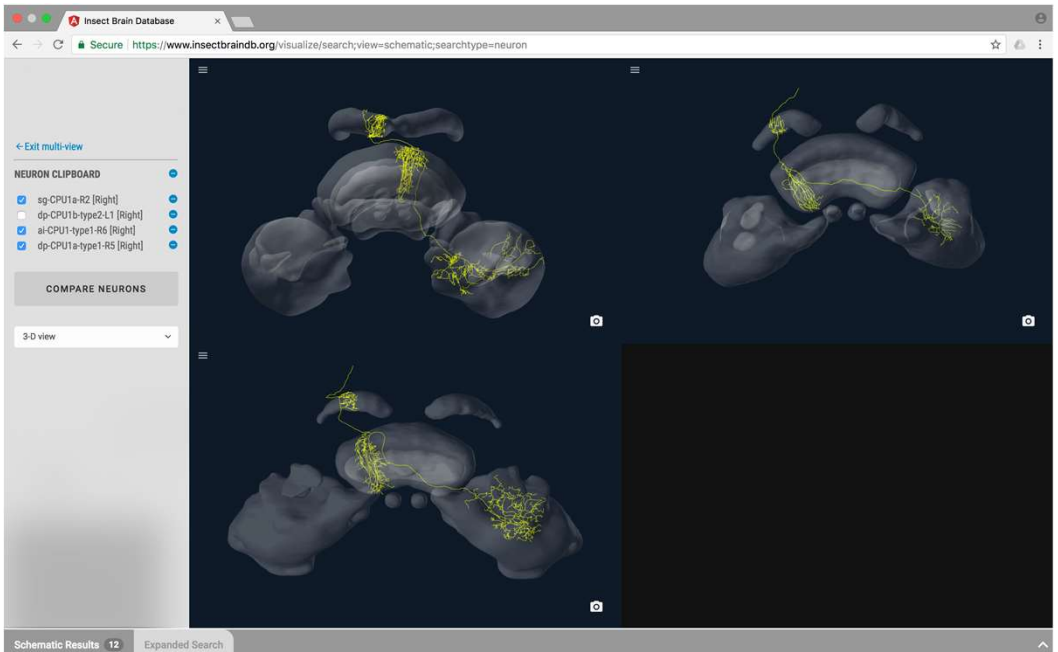
—> can be made public by one click

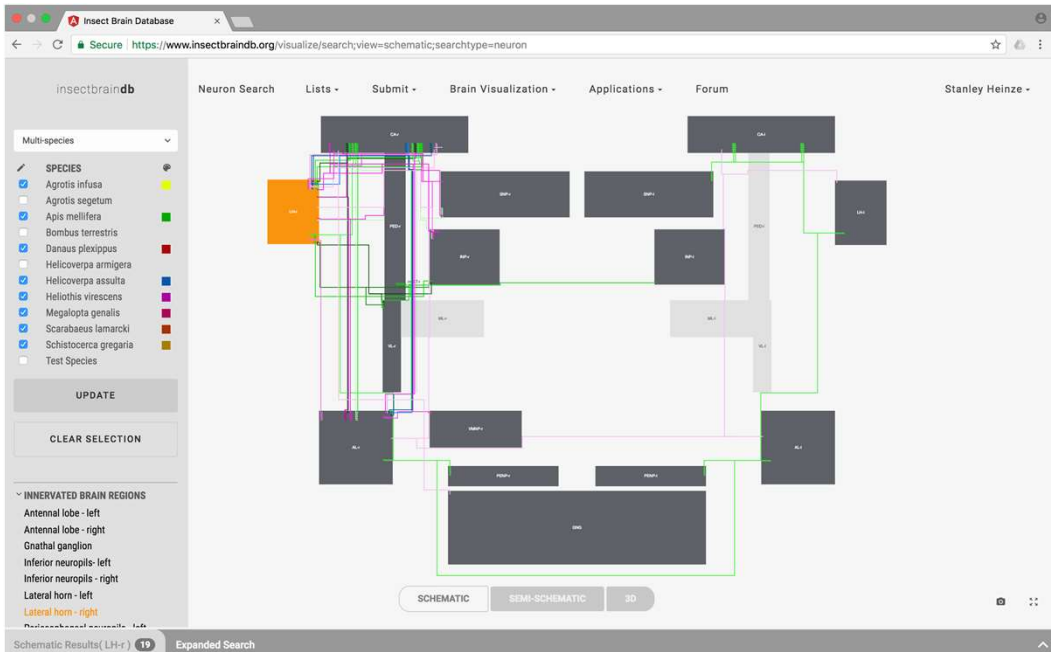
Usefulness for every-day life

(why should anyone deposit data in our database?)

- Create publication-ready snapshots without additional software
- Compare neurons within and between species side by side on site
- Locate own data and data of others faster and easier than any other way
- Easily find publications linked to data
- *Private mode:* Use as lab-internal data-repository







My data management plan evolved into a software tool:

- Provide direct advantages for researchers who deposit research data pre- and post-publication
- Eliminates the extra burden of making data available after publication
- Facilitates organization of research data in a way that is compatible with the field
- Provides a persistent identifier for each dataset
- Provides a cost-free way to manage, archive and share even complex data

Thinking about the DMP helped me realize:

- what my data is
- what the standards in the field are for certain data
- what meta-data are key to make data useful to others
- how to make data available to the field
- what the current standards for repositories are
- what the future mandates will be and how to anticipate them