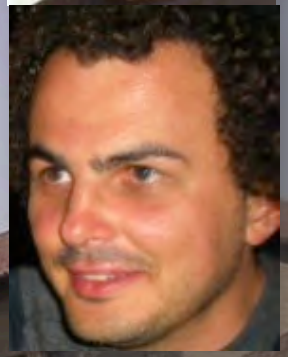
Project on Scientific Transparency

Prof. Brian Wandell*

Director
Stanford Center for Cognitive and Neurobiological Imaging (CNI)

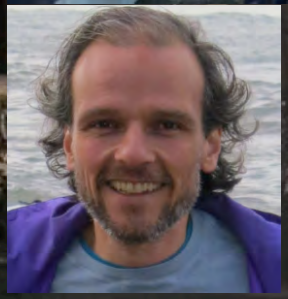
Deputy Director
Stanford Neurosciences Institute (SNI)

Michael Perry*



Renzo Frigato*

Gunnar Schaefer*



Bob Dougherty

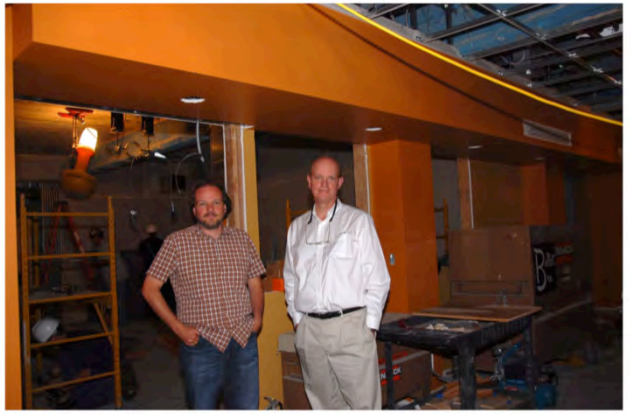
We thank the Simons Foundation and the Arnold Foundation for their support

Commercial disclosure*:
We are active participants in a commercial venture related to the work described here, Flywheel Exchange, LLC.

Planning for Stanford's Center for Neurobiological Imaging

- The story behind this presentation starts with Bob D. and me, working in our lab, and eager to improve software
- Then, as we were putting together the CNI, we had a chance to think some more but at a larger scale about infrastructure
- This talk is about what we learned from our colleagues about these issues over the last few years, and what we are trying to do



Scientific goals of the project

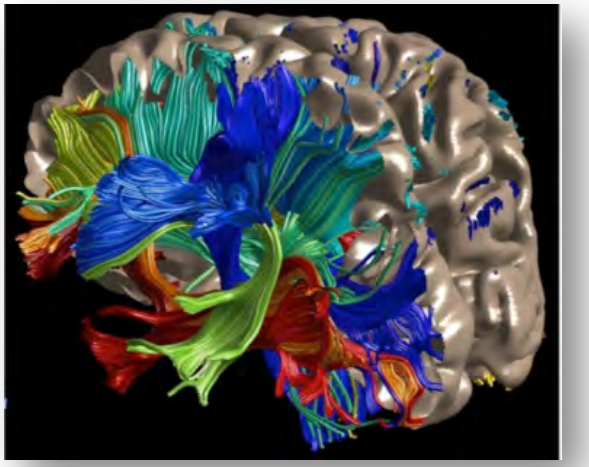
- Reproducible Research
- Personalized neuroscience
- Software design and technologies
- Where we want to go

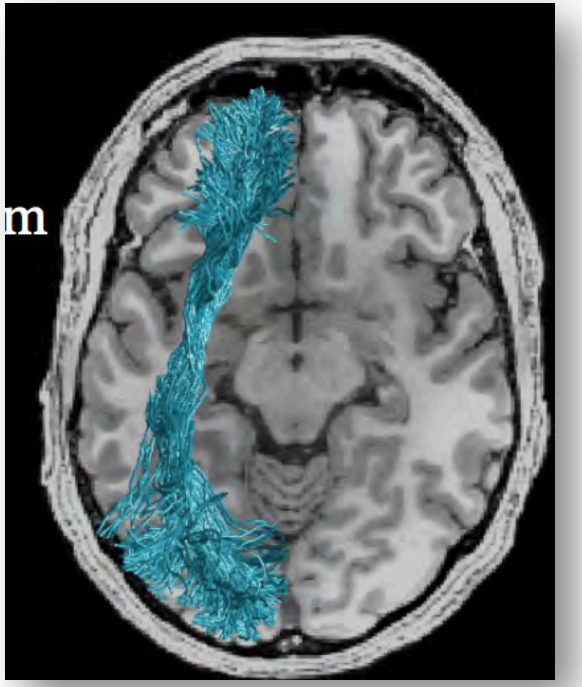
Replication is often impractical

- Neuroscience is transitioning from a cottage industry to larger labs and research organizations

- The complexity of tools and computations puts us into the field of 'big science' in which full replication is impractical

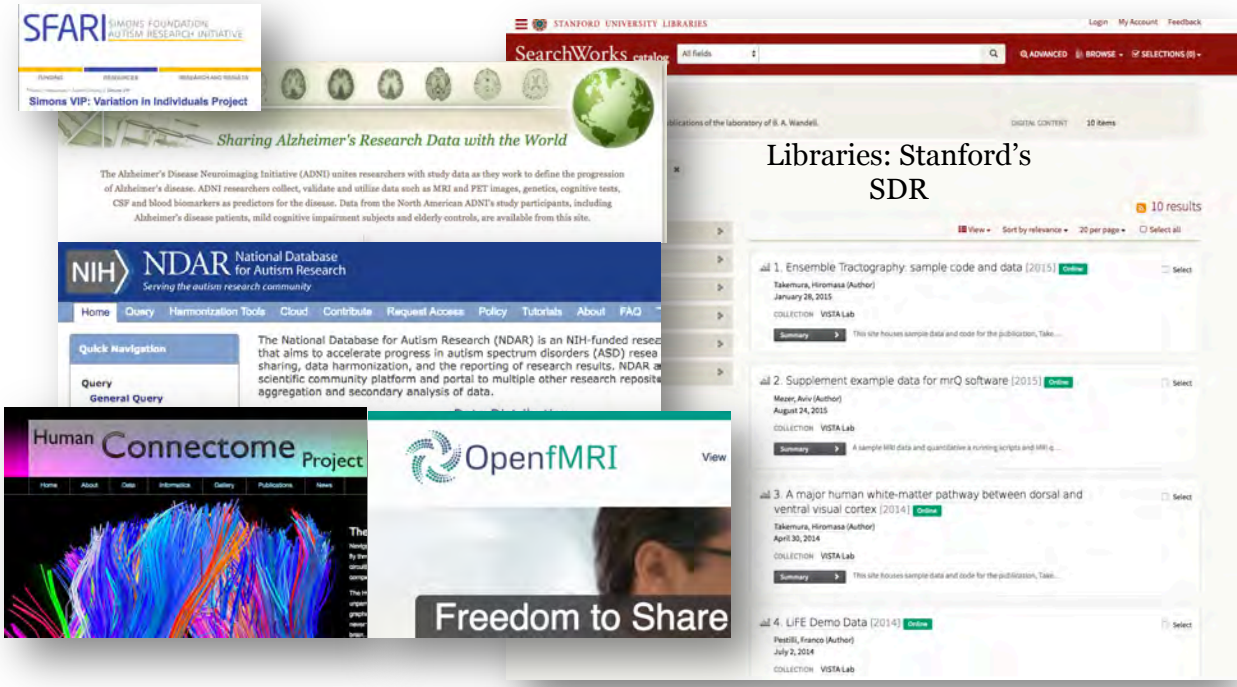
  - **Example 1:**
    Analyses of 320 subjects with autism, each measured using MRI, at 3 time points over four years
  - **Example 2:**
    Structural MRI of 50 illiterate adults flown in from the Amazon rain forest





What tools are on offer?

- Assemble your work and put it in a shared resource (archive)
- Many sites offer archival data storage; I will return to this later
- I particularly admire Amy Hodges and the work at Stanford University Libraries; we use their repository for permanent URL a lot
- A few computational sites that seem important to me. Hurlin, Perignon and Stodden, "runmycode" is very cool for archival code.



Libraries: Stanford's SDR

Reproducible research is within reach

Replication means obtaining the data again, usually by independent investigators using similar methods, equipment and protocols

Many big science experiments are too expensive and time-consuming to replicate

Reproducible means that starting with the data gathered by the scientists, we can confirm the derived results (e.g., statistics, summary curves and images, numerical relationships)

Proper tools enable scientists to achieve reproducibility much more often

Reproducible Research

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Donoho

Reproducible Research

Computational reproducibility is not an afterthought—it is something that must be designed into a project from the beginning. **One does need to develop a whole set of programming and research disciplines** with the end result in mind and stick with them.

Donoho

Scientific data management goals

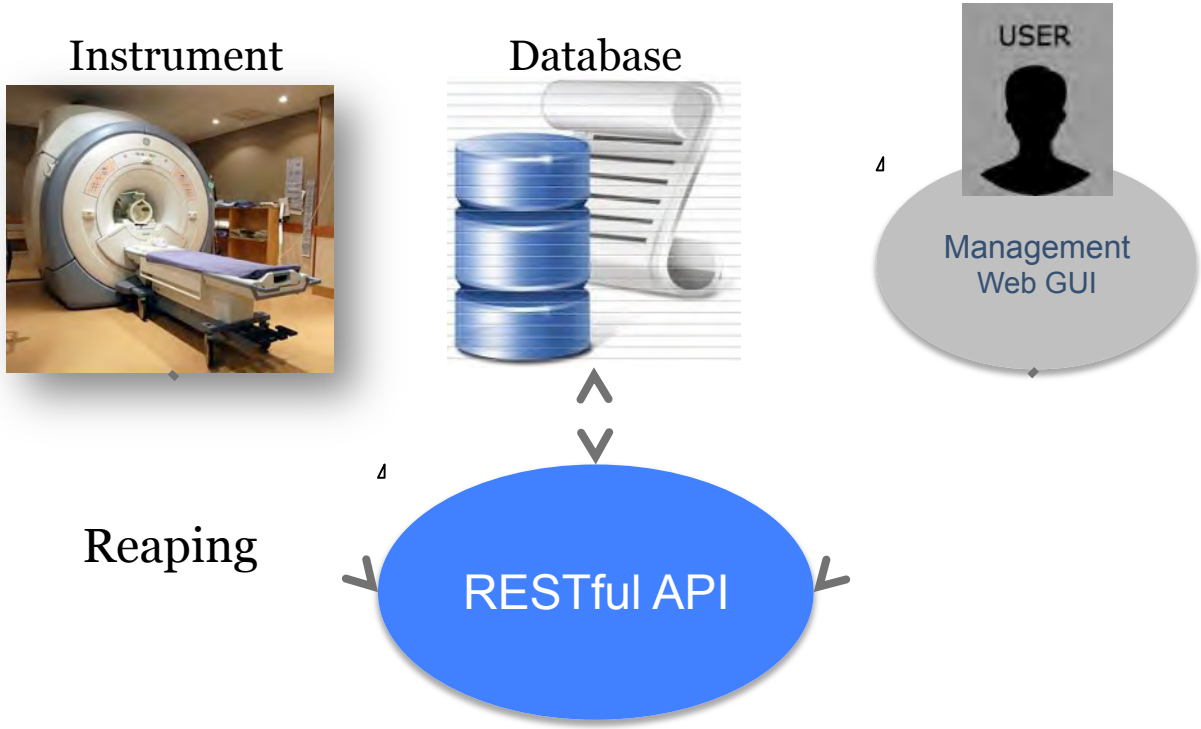
- Reproducible research requires **scientific data management** tools
- These should be expected as part of science and scientific training
- When we started the **CNI** we had these goals in mind, but no tools; Gunnar and Reno arrived; Simons and Neuroventures helped

SCITRAN: The scientific transparency foundation

- Use data management tools from the beginning of the project; archival tools at the end
- Simplify data sharing and data discovery
  - User-rights management
  - Search
- Support reproducible research methods
  - Complete software
  - Portable methods
- Be helpful and offer value
  - Data archiving; note taking
  - Automated quality assurance
  - Visualizations
  - Platform independence (browsers, docker)

An active experimental system needs more

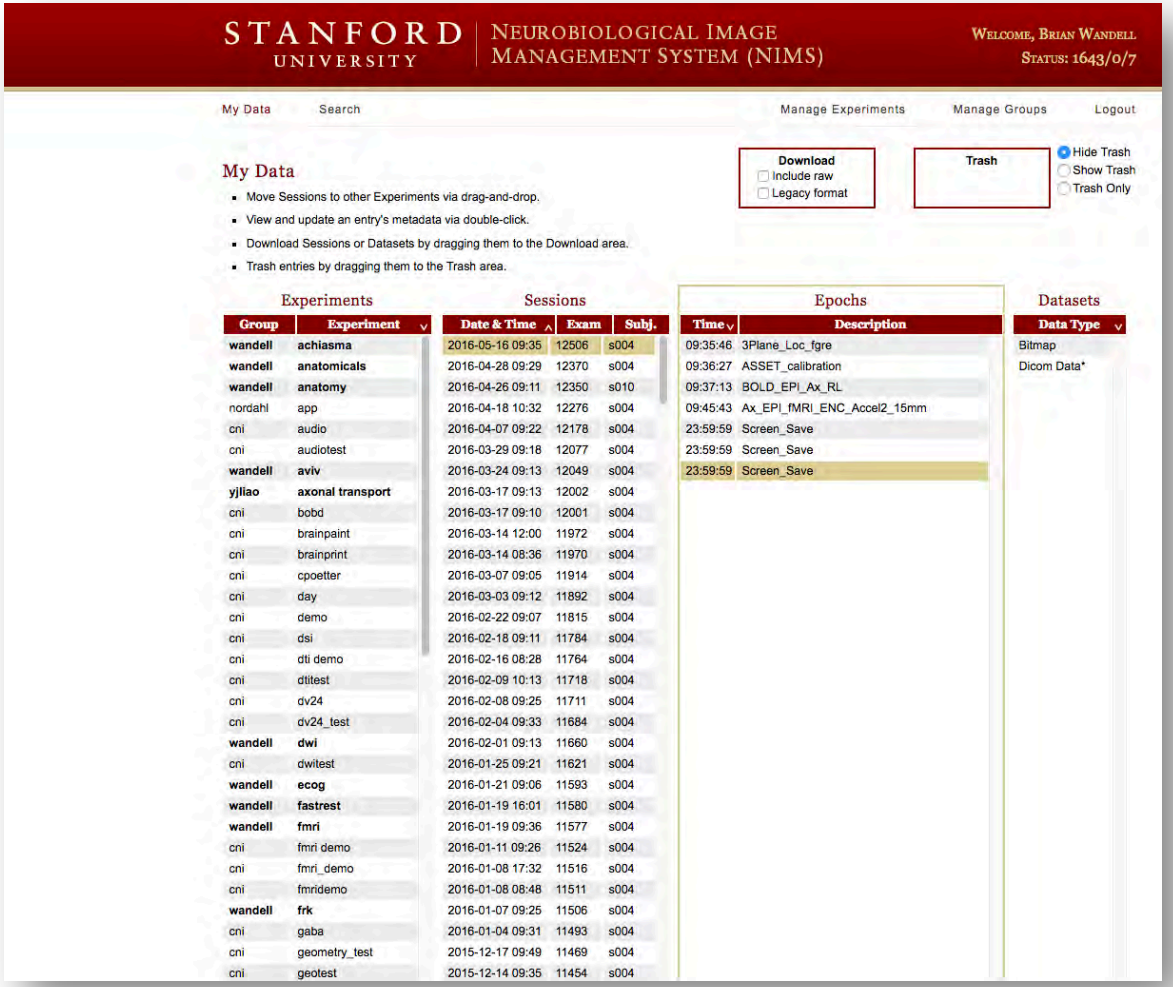
- Data from the CNI scanner are parsed and stored in a permanent database, no user action required
- The raw data are converted to common formats (e.g., NIfTI, bvec, bval, physiological)



Current status at Stanford CNI (04/12/2015)

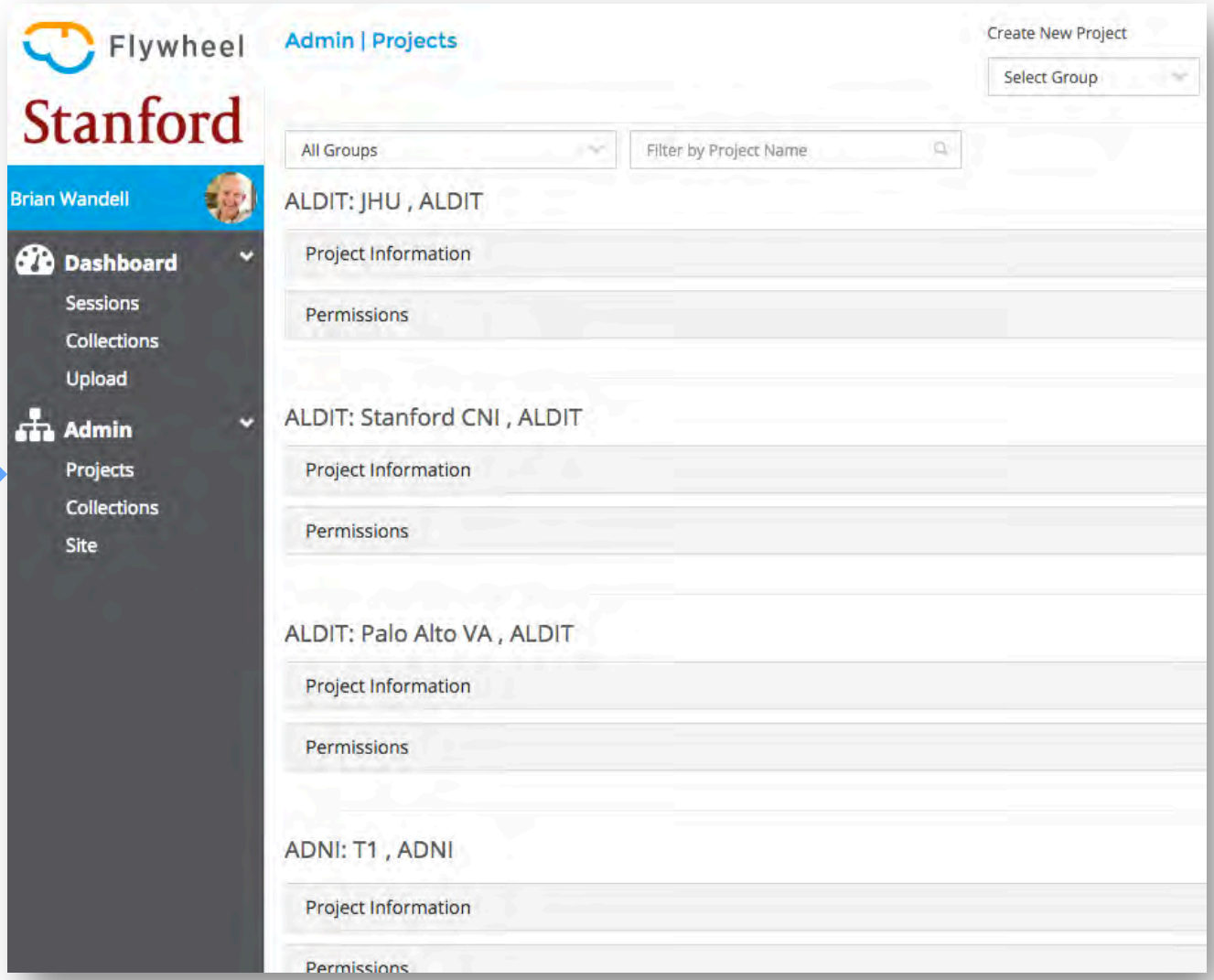
People
700 scientists
11,469 scan sessions
6232 research subjects
64 research groups

Data
46,229 fMRI
9708 T1 anatomical
5081 diffusion
536 spectroscopy



Project and user rights management

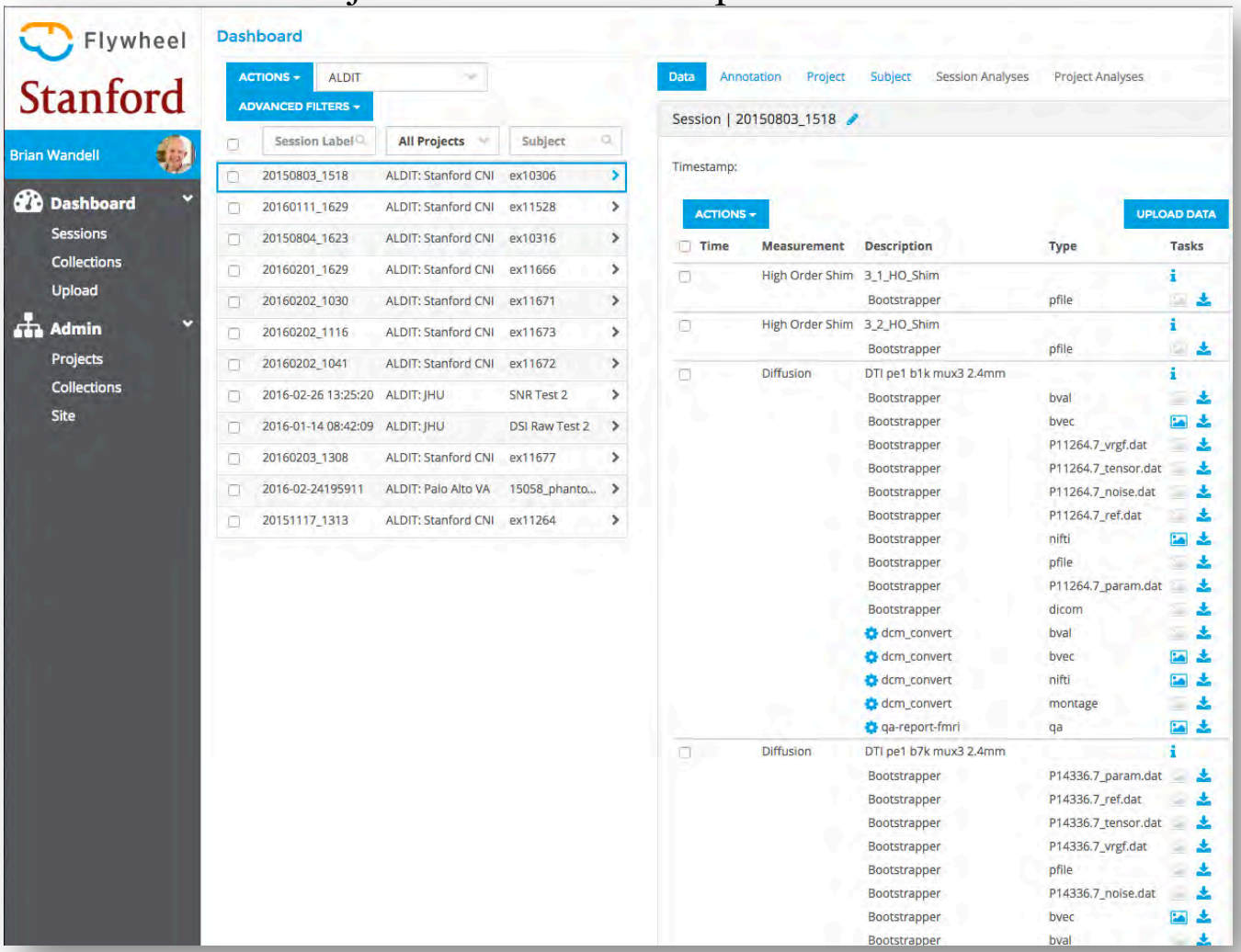
- Create projects
- Assign access to people (RA, Collaborators, Students)
- Levels of access from admin, read-only, read-write.



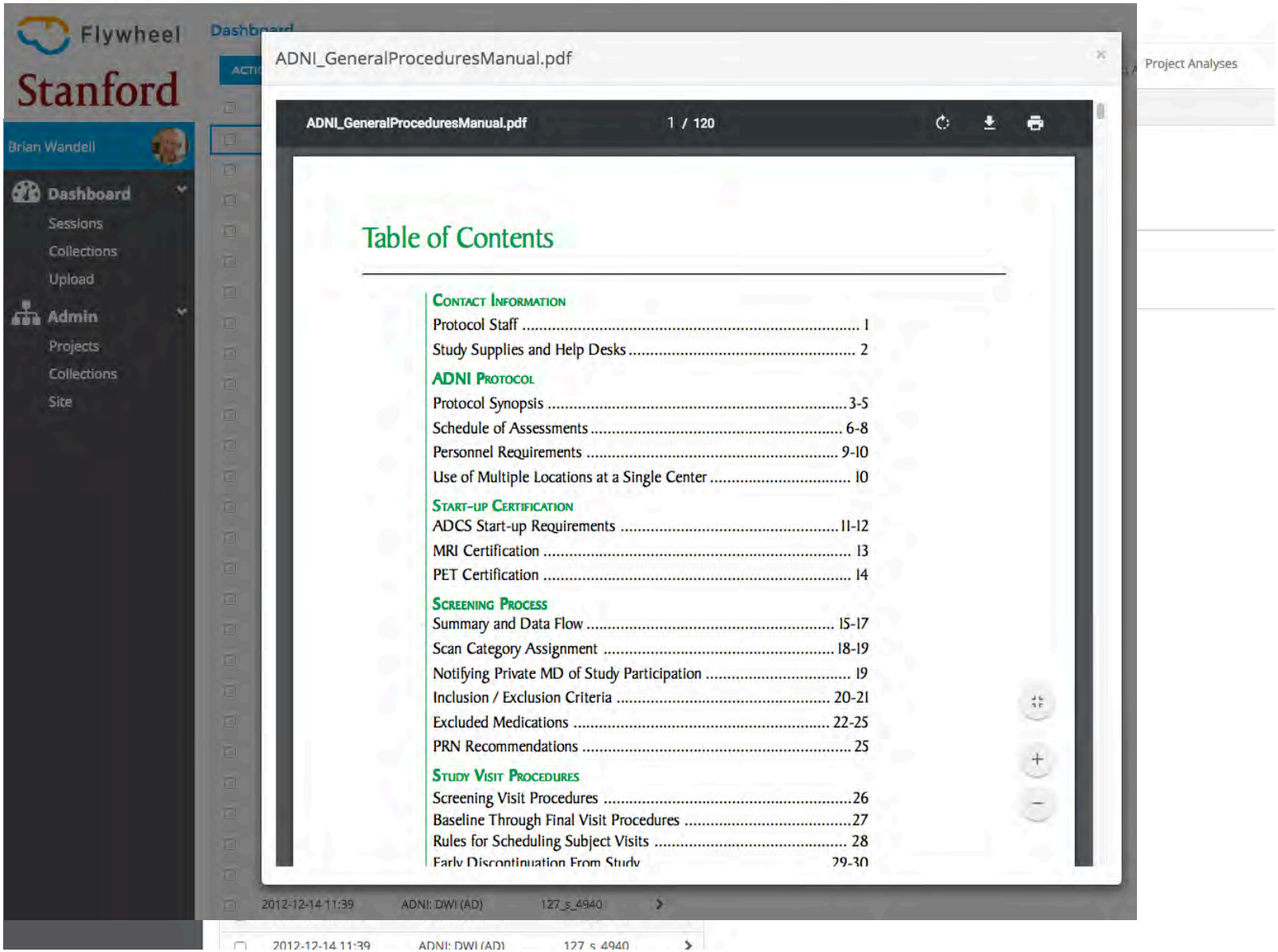
Value: Monitor acquisition and keep data organized

- We check which data have been collected and processed from any web-browser (desktop, mobile)
- We can visualize data to check quality and monitor experimental progress

It took us a long time to realize this simple organization
Project -> Session -> Acquisition -> File

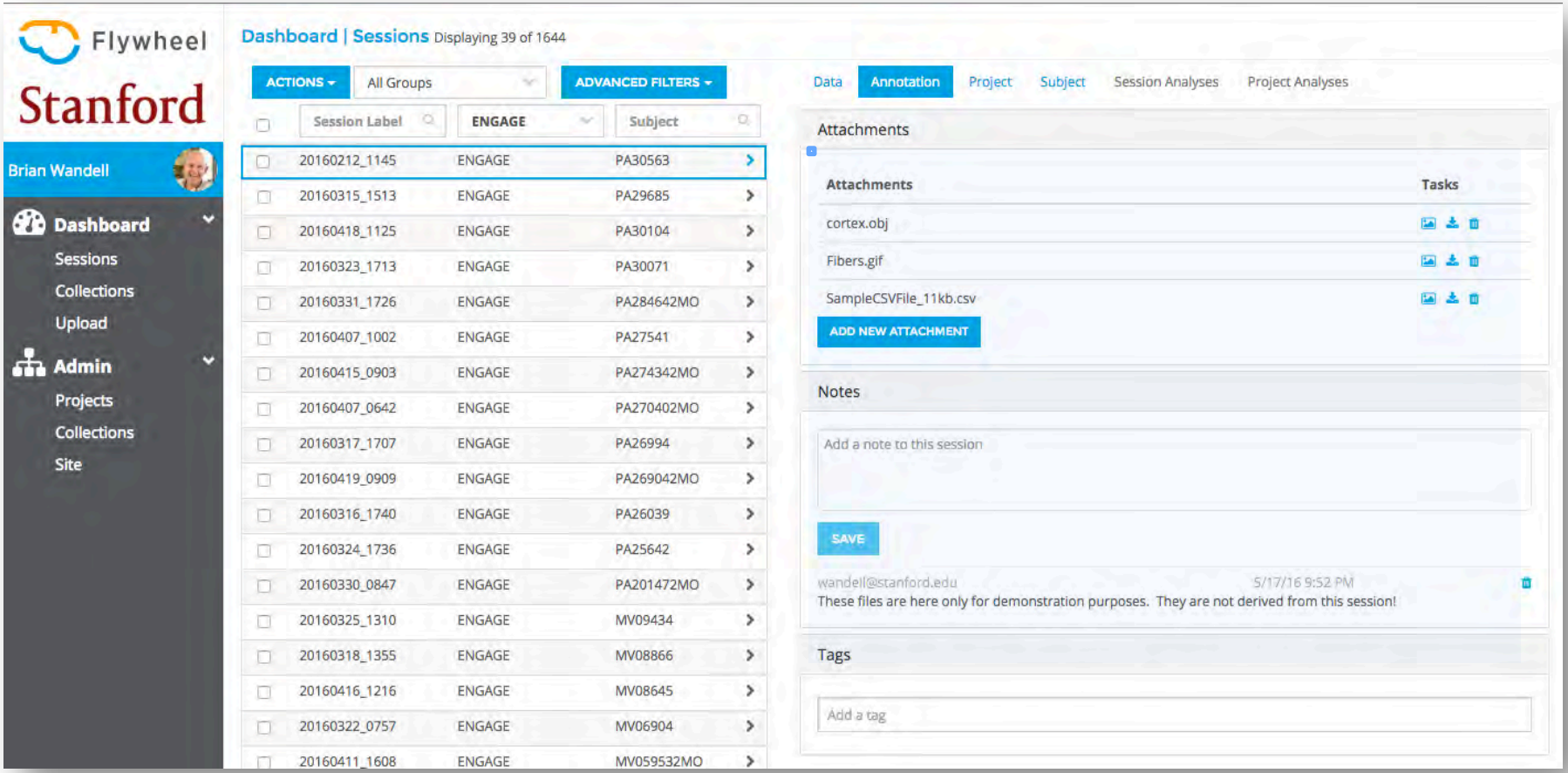


Value: Provide information about the project



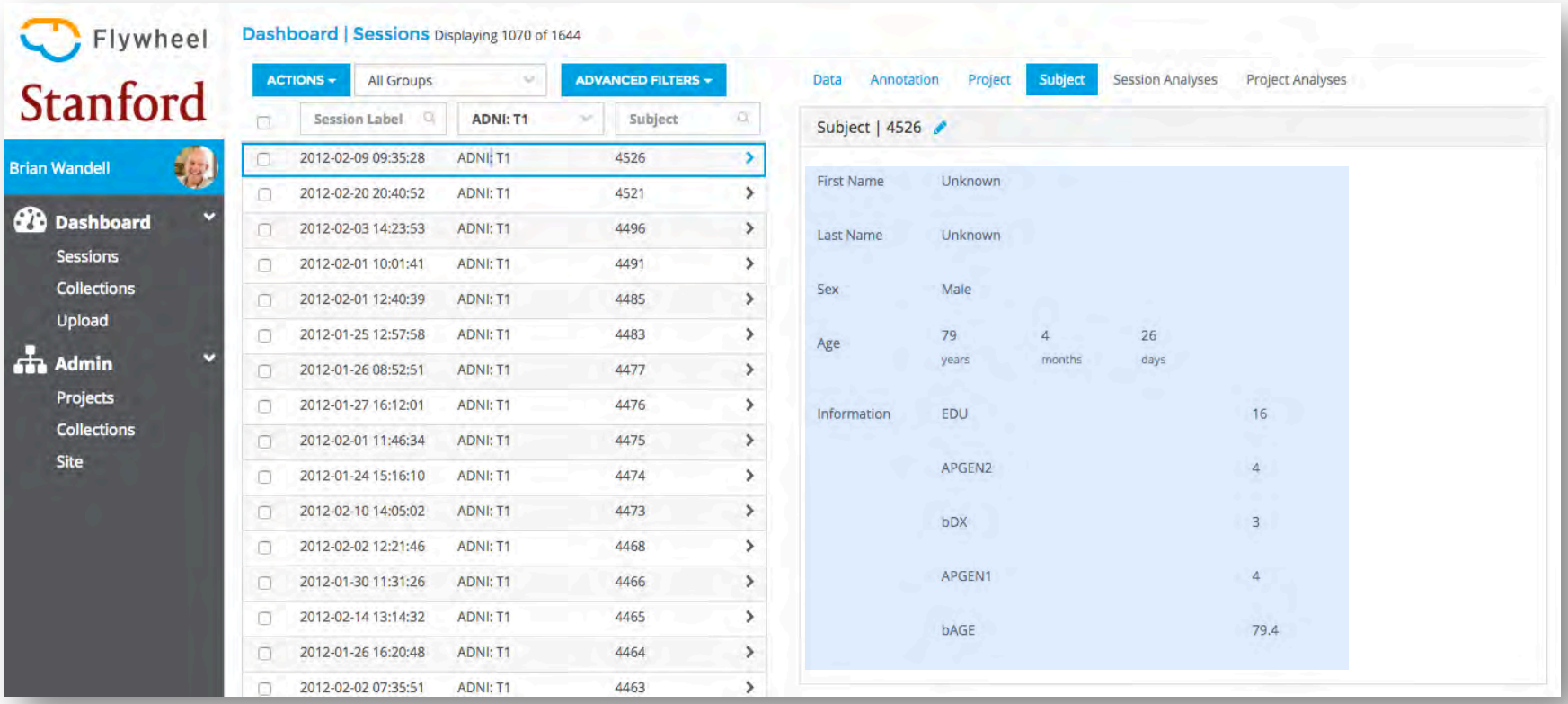
Value: Notes, attachments, viewers

- Create notes about a project, or a session, or a file
- Add attachments describing the project, or session (e.g., something special happened during the session)
- Viewers for attachments



Value: Edit and search for subject information

- Subject information
- Basic fields included (age, sex)
- Additional fields can be created and values stored

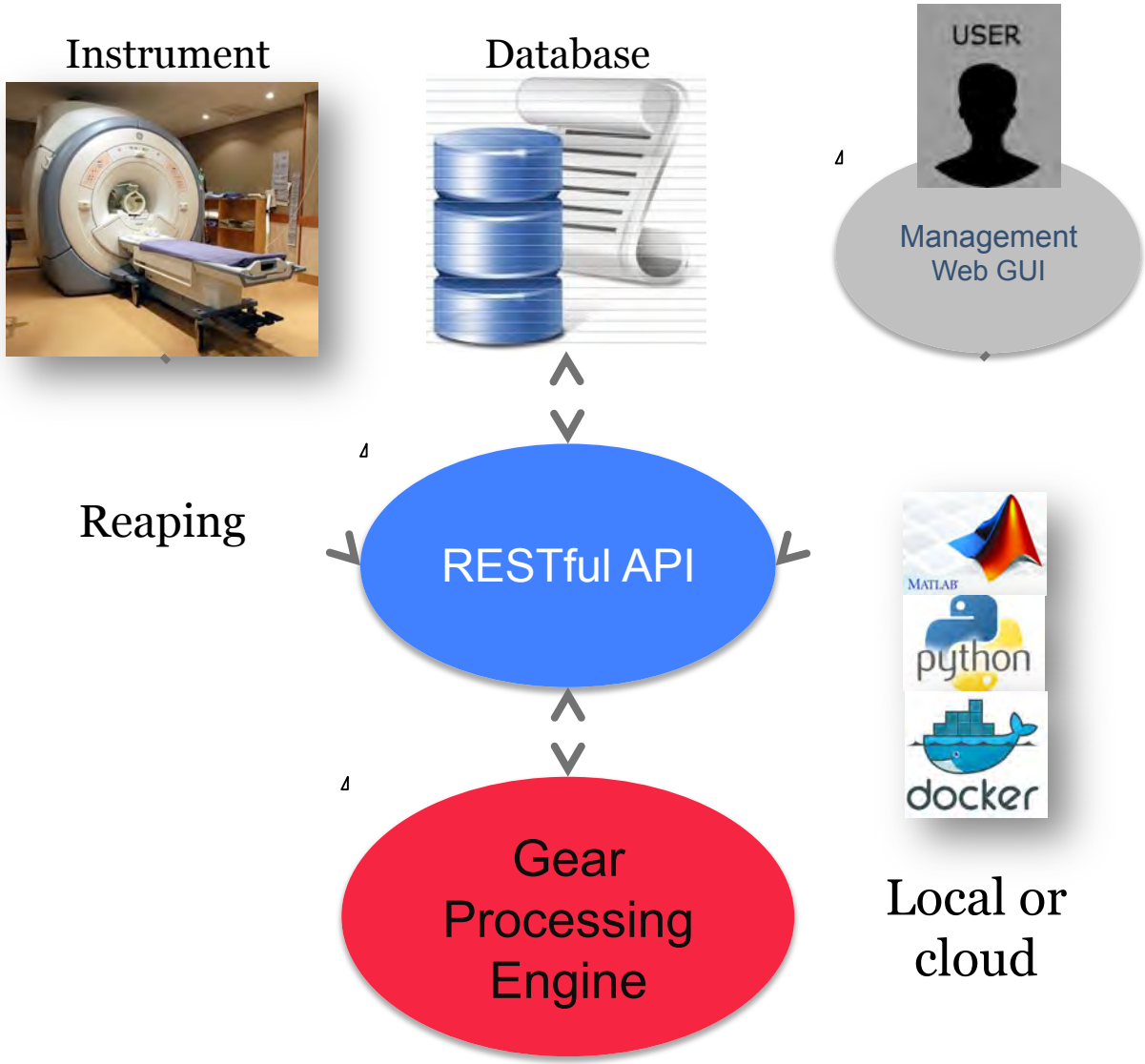


Value: Time machine for data organization

- Data from colleagues, legacy, or public repositories can be added (openFMRI, HCP, ADNI)
- BIDS, organizers, multisite integration

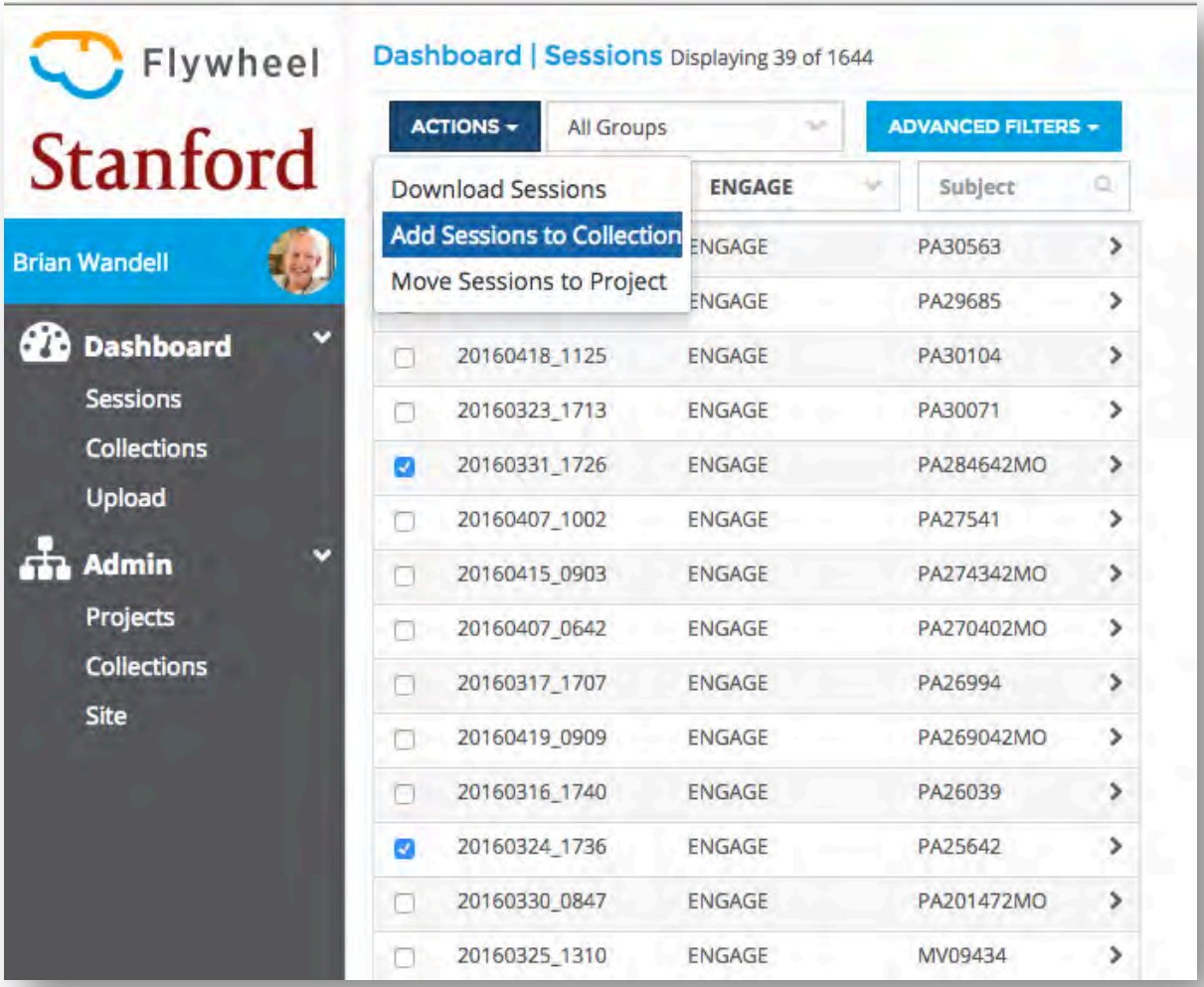
Analysis methods

- Given permission, data can be collected into a virtual experiment for further analysis
- There is a developing format for creating a defined container (**Gear**) to analyze data
- Processing will be managed by an engine
- The Gear analyses – including results and software – are described in the database for reproducibility
- Gears are based on docker containers for completeness and platform independence – Matlab and Python interface are being developed



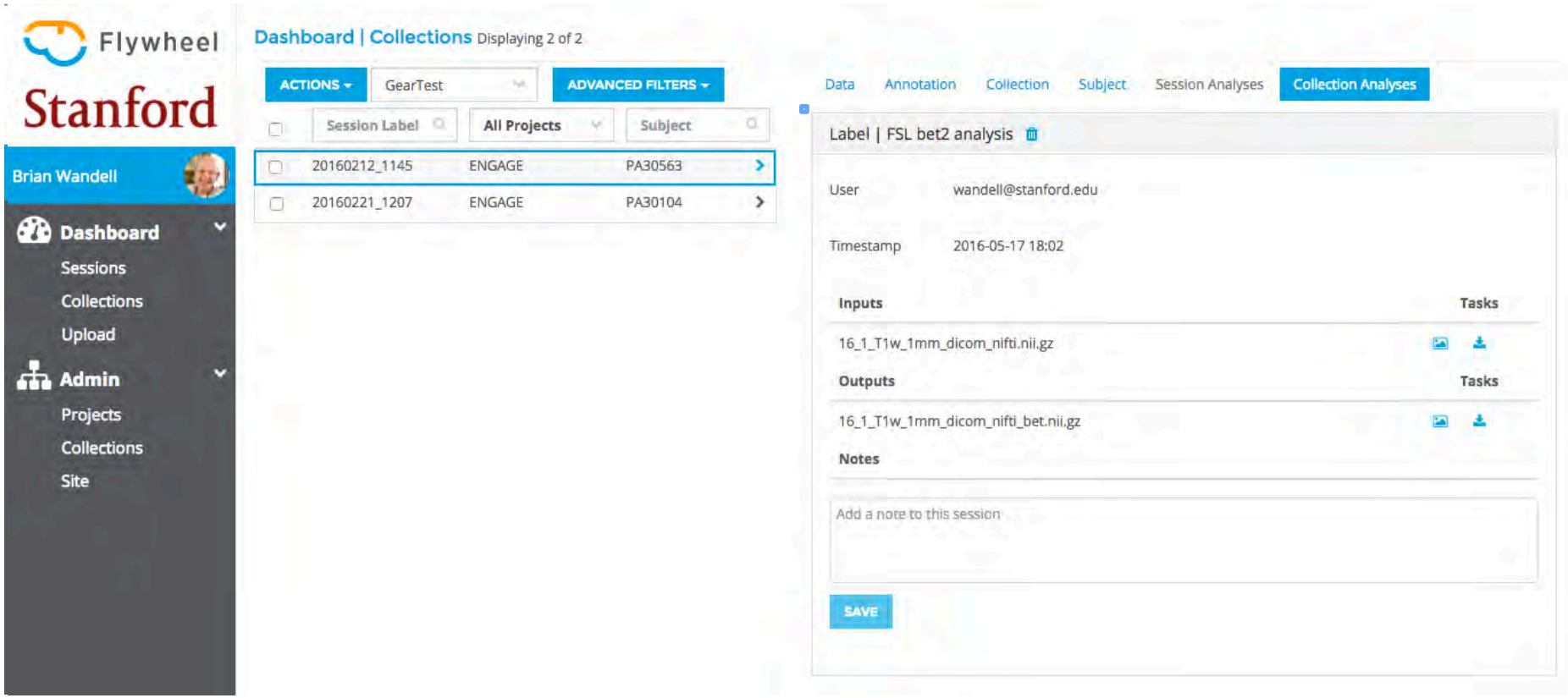
Analysis example: Create a collection from the database

- Select sessions or acquisitions to group into a collection
- We expect people will do their analyses on collections
- Ask me about file duplication afterward – the answer is cool

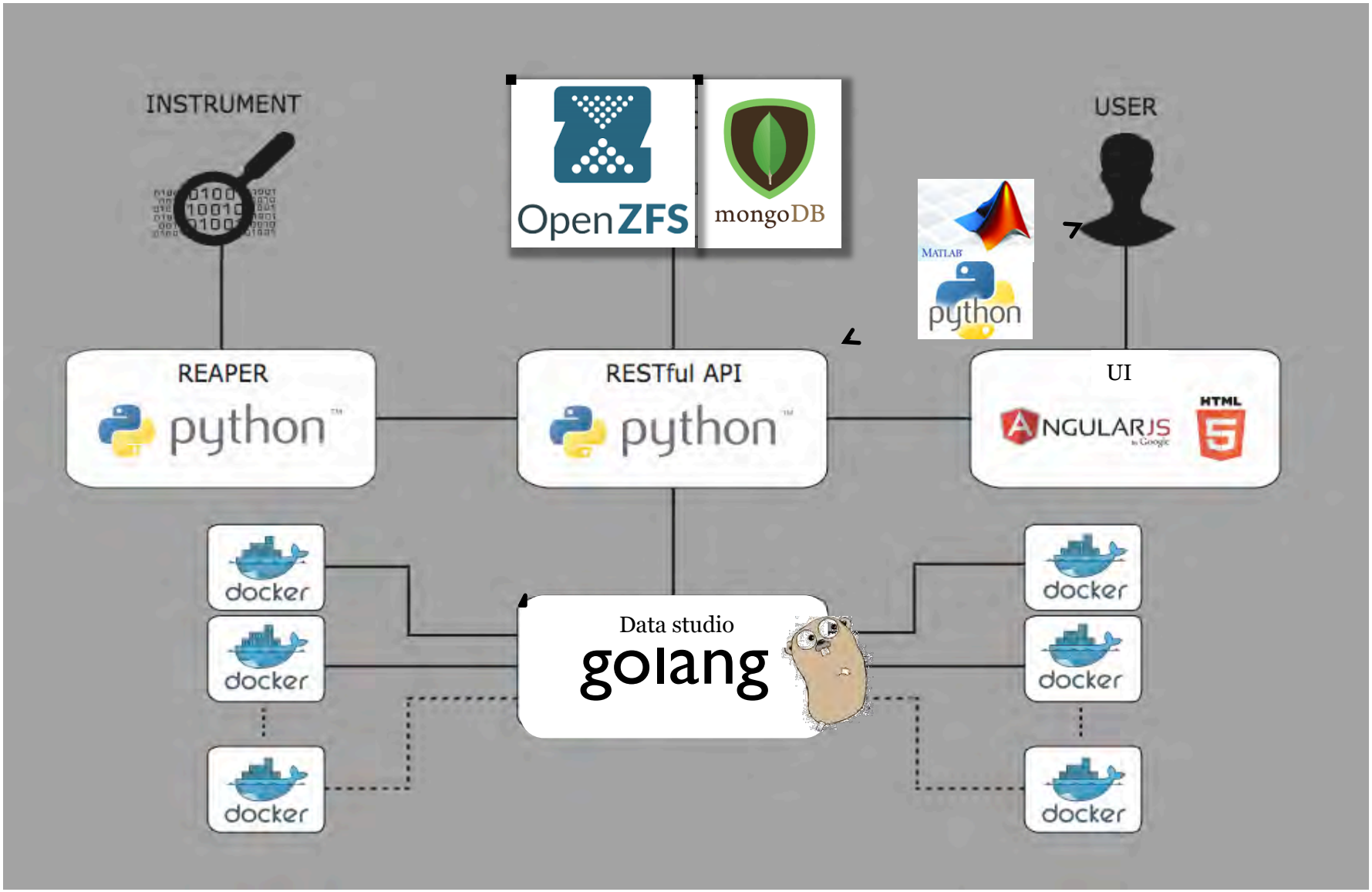


Analyze collection data using Gears

Gears can be executed on a file from the UI or
Using an API with both Matlab and Python interfaces
The analyses posted to scitran

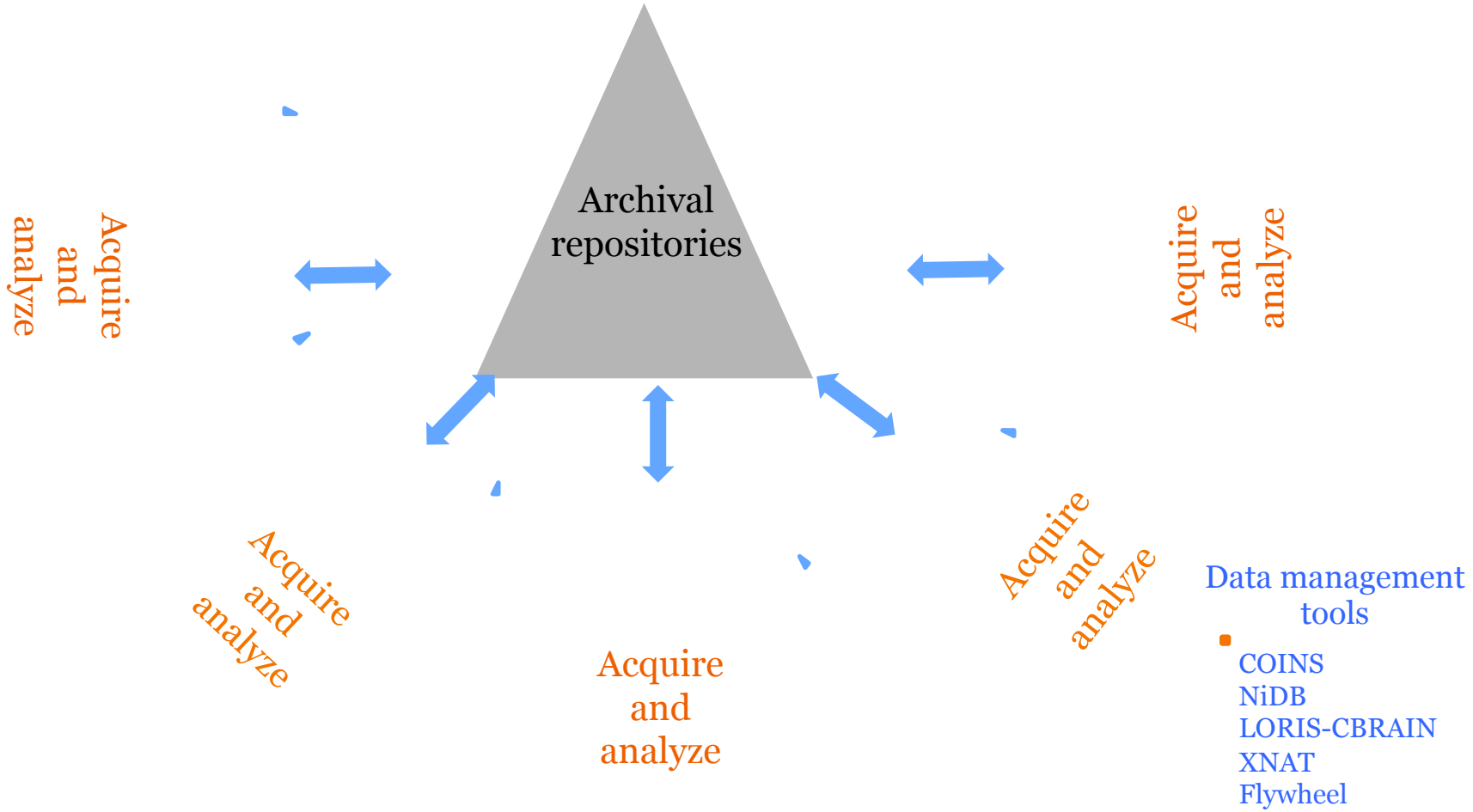


Web technologies



Archives and data management tools

LONI IDA UMCD iEEG.org MR Innovations Anima OMEGA	CBFBIRN Open Science CBS PAIN PDBP (Parkinsons) GAAIN Age-ility (nitrc)	Cimbi PedsDTI ConnectomeDB (xnat) SchizNet (xnat) NDAR	FITBIR openFMRI (scitran) Neurovault FBIRN ADNI NUNDA (xnat)	NITRC PING PNC PLORAS VUIIS (xnat) XNAT Central (xnat)	Simons VIP (scitran)



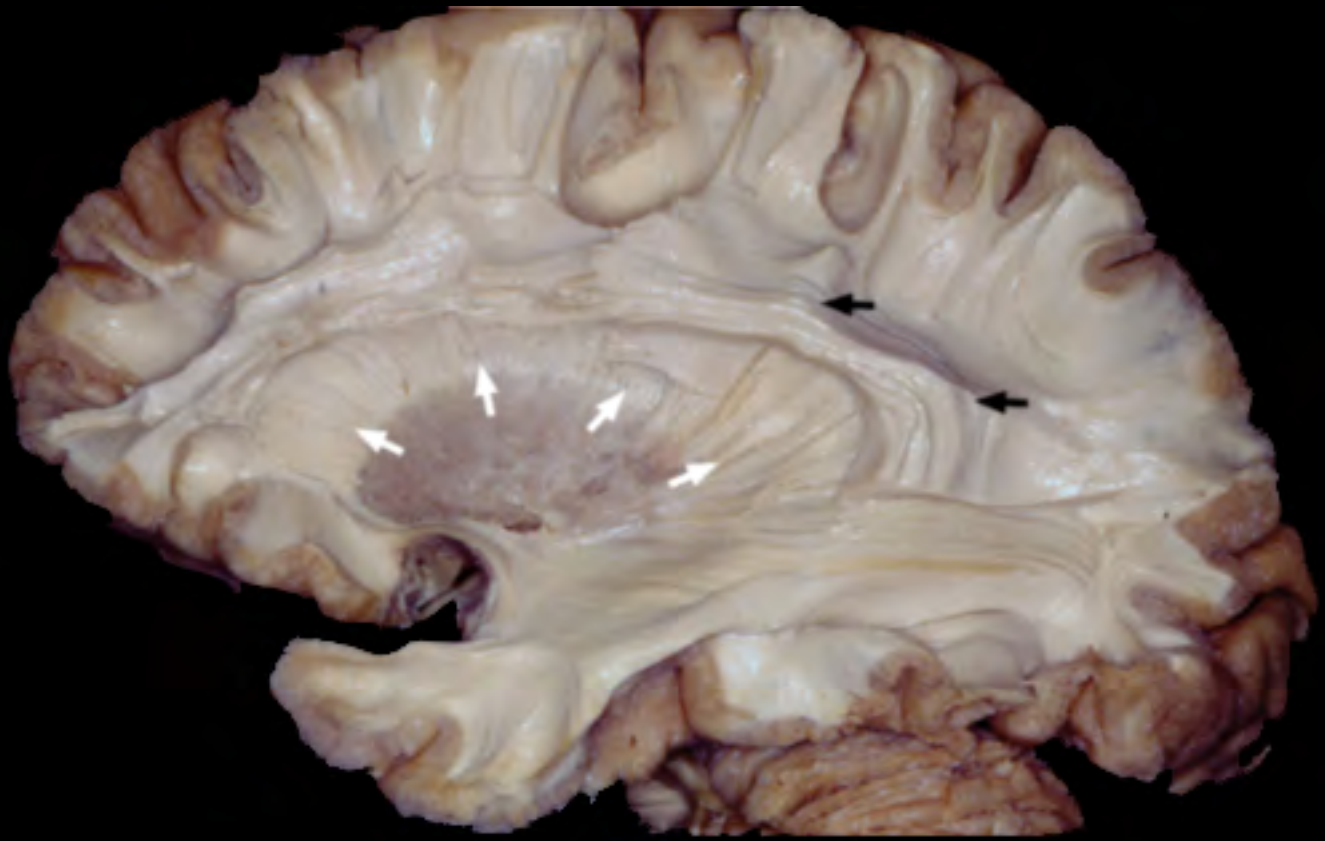
Scientific goals of the project

Personalized neuroscience

An opportunity that derives from a combination of big science and the quest for reproducible research

Human fascicles (tracts)

- There are many long-range connections
- These connections are not passive – they change their properties in response to use
- A system with active wires



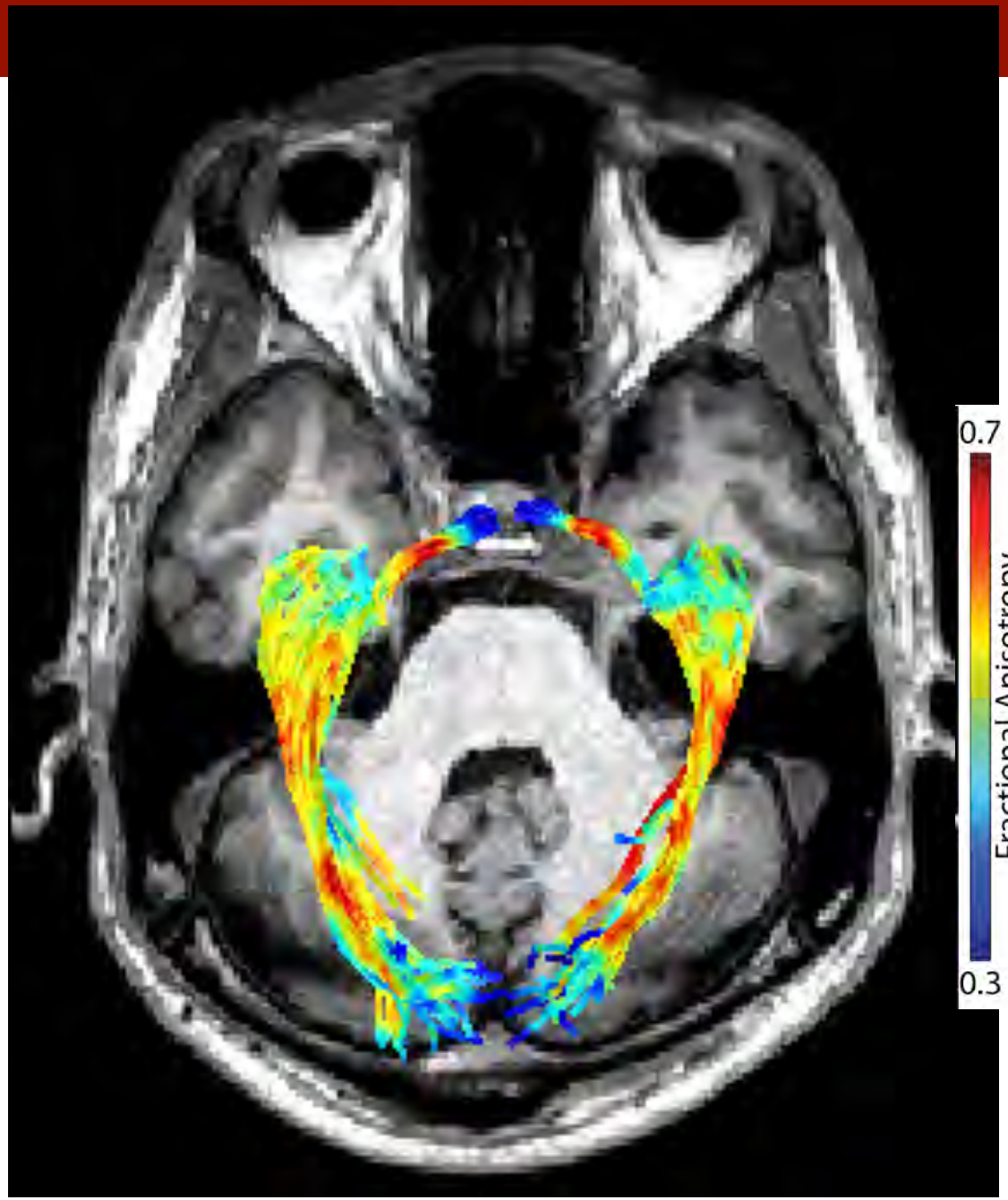
Courtesy Professor Ugur Ture

A motivating example

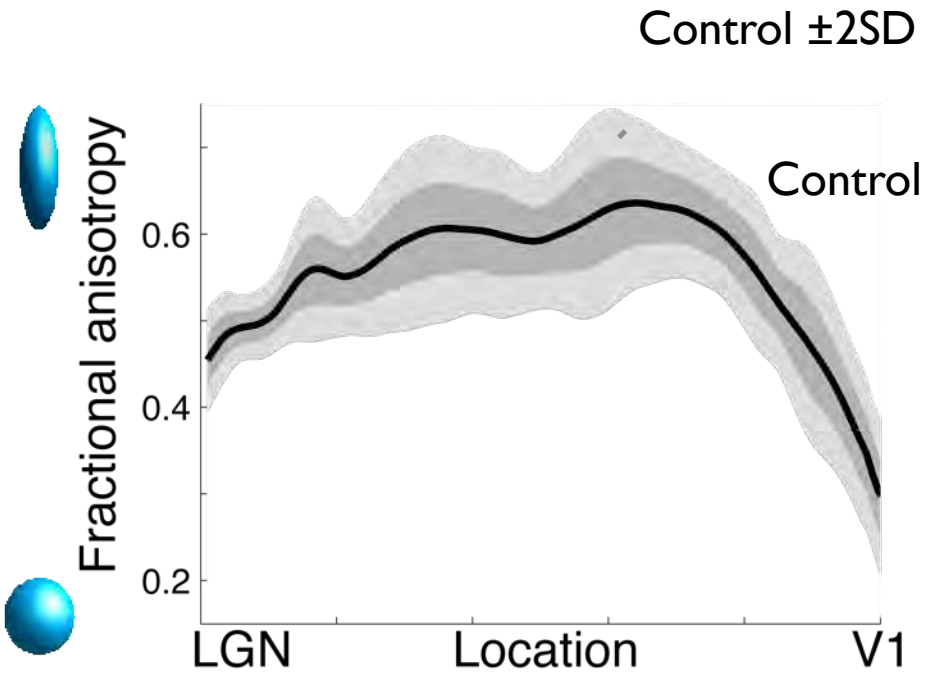
- A subject or patient with a retinal eye disease comes to the lab
- We want to know the consequences of retinal degeneration on cortical structures

A motivating example

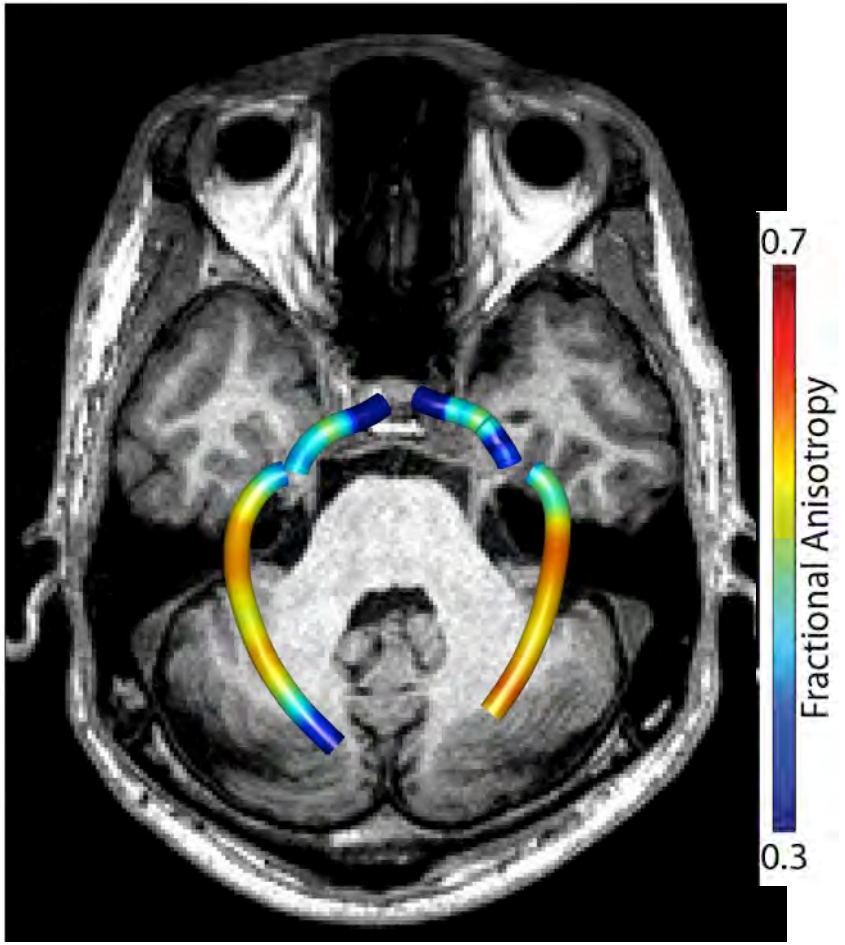
- Measure the subject's visual white matter and secure the data!
- Use validated computational tools for quality assurance
- Use open-source software for tract identification, tissue estimation and comparison with other populations



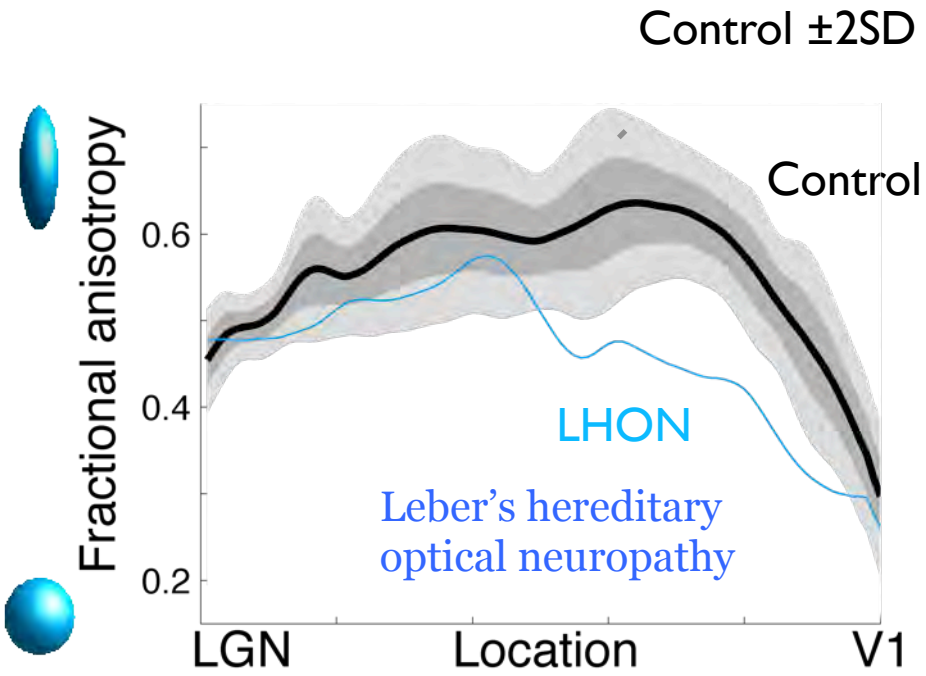
Use databases to find control data



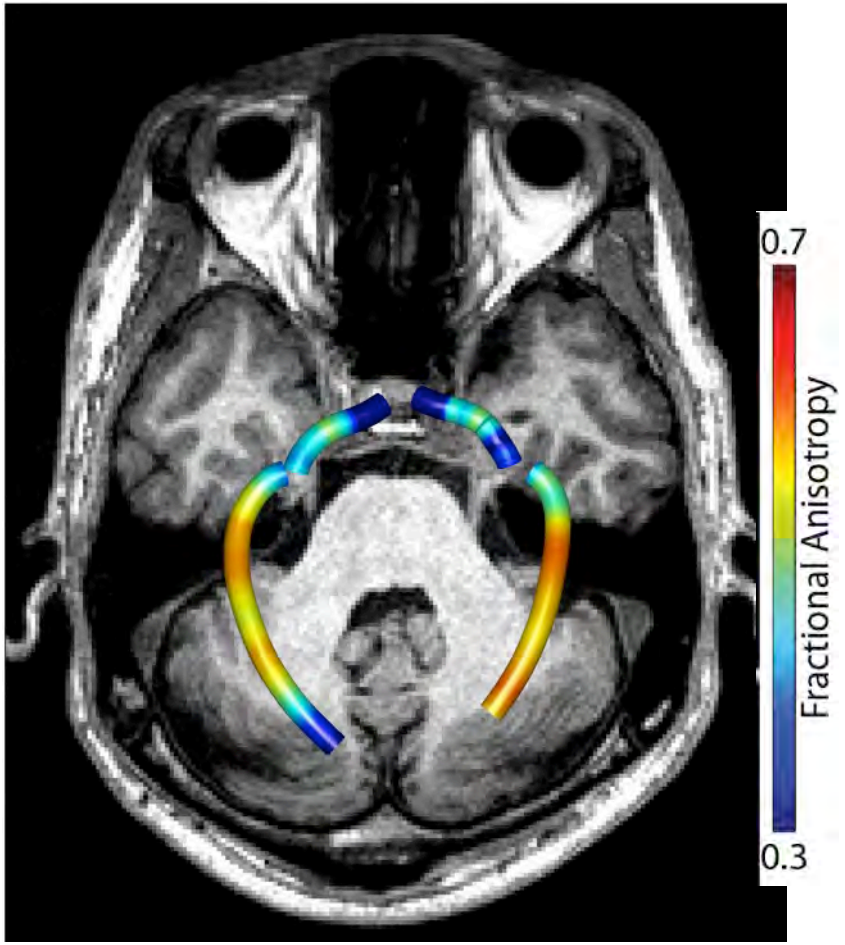
The expectation based on data acquired and stored in the SDM



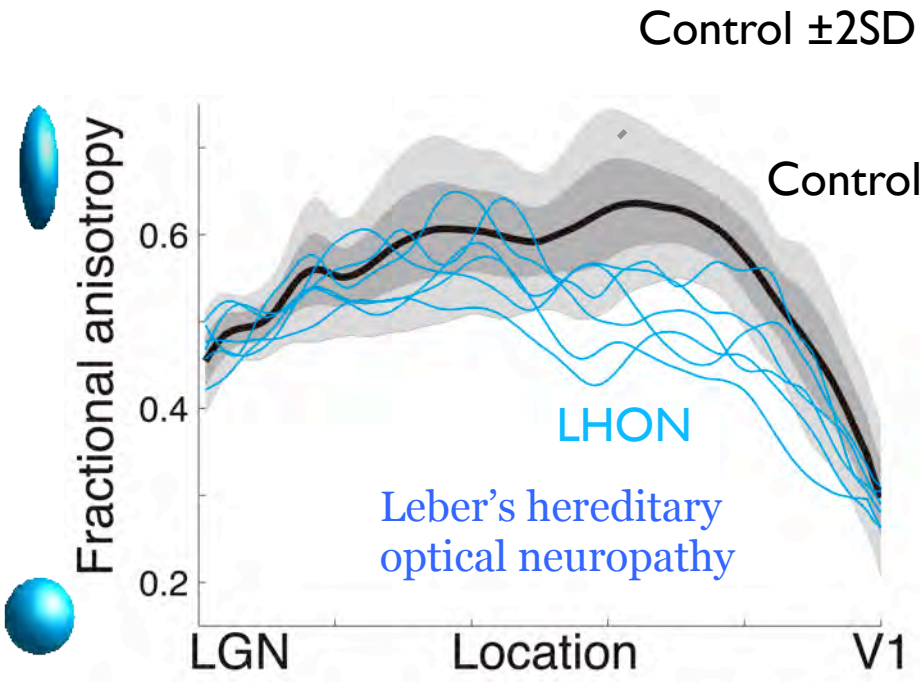
Compare your subject with the distribution and think

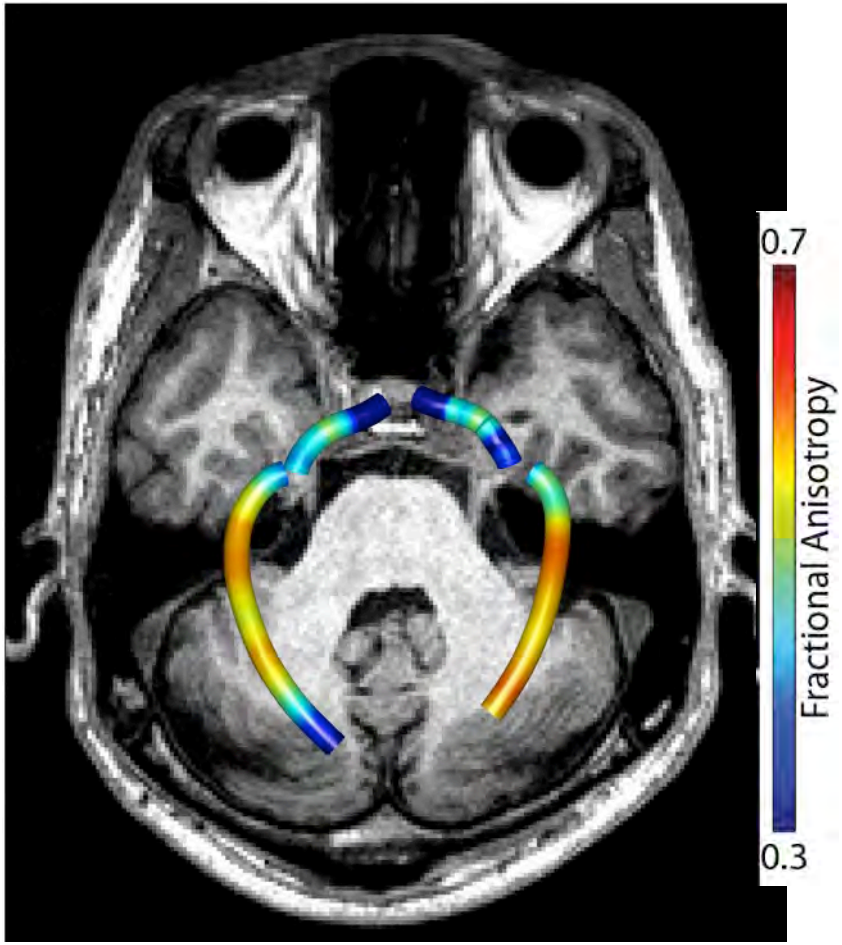


This subject compared to the expectation



Data science and statistics issues





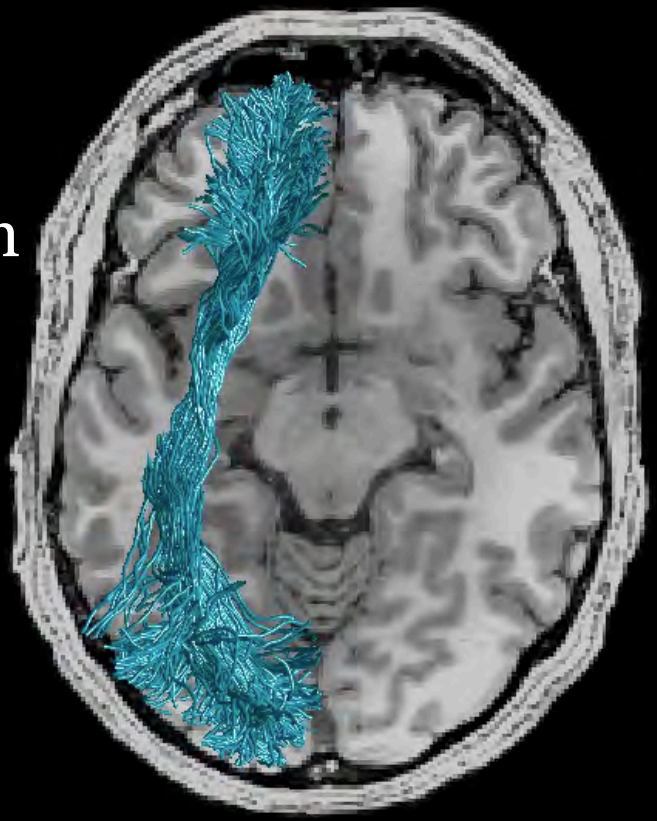
Each subject with the disease has some variation and we would like to know, and track each one over time

An example application

(image from F. Pestilli)

Inferior fronto-occipital fasciculum

Left IFOF



150 Directions, 2 mm^3, B=2000 projected on a 1 mm^3 T1 anatomical image

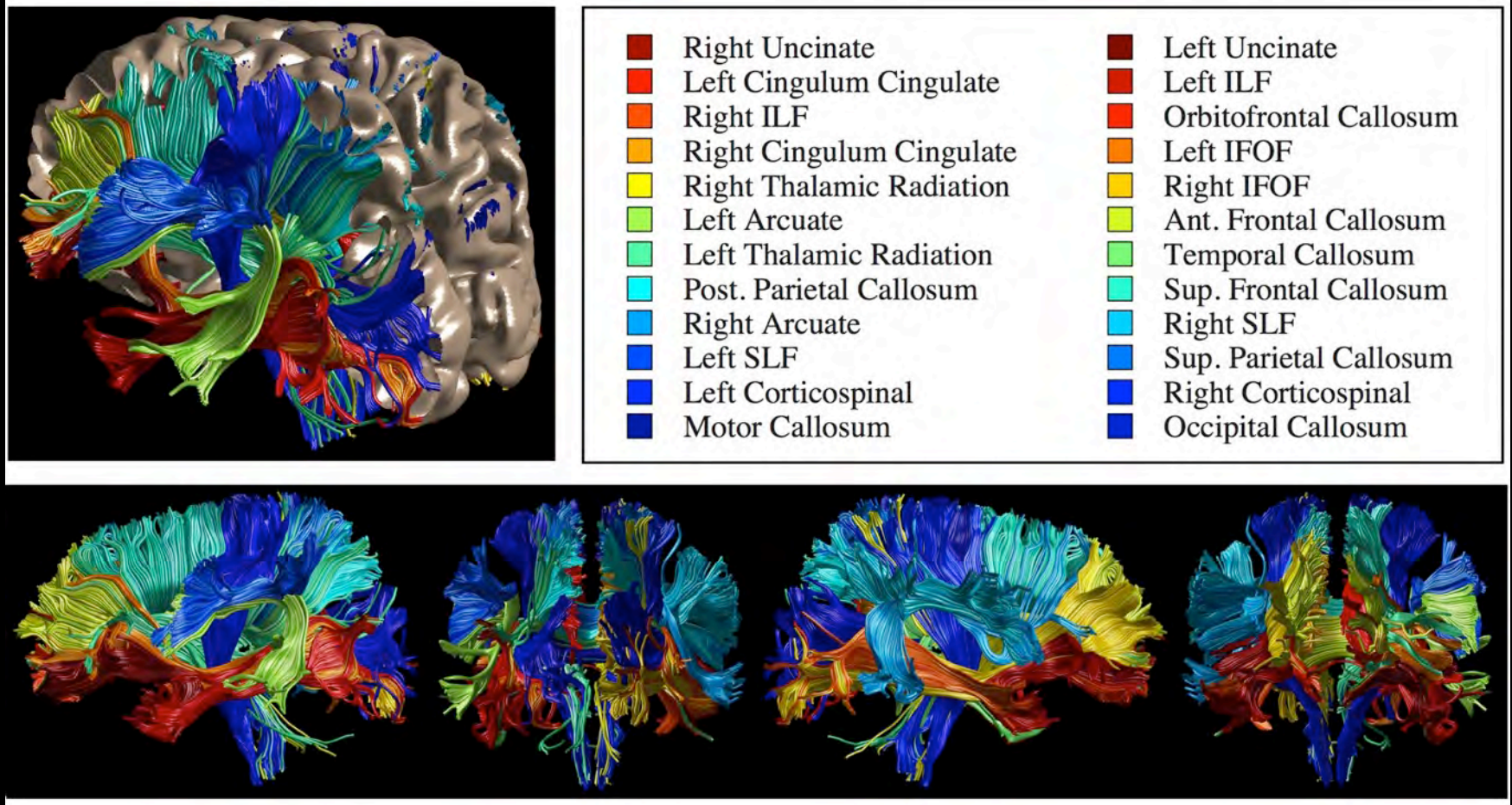
Building databases of healthy controls

Tissue development from 7-85 years

(Yeatman, Wandell, Mezer, 2014, *Nature Communications*)

- Diffusion-weighted and quantitative imaging identify the major tracts and tissue properties in individual subjects

Color indicates age of maturity. See paper for details. (N=120)



Scitran extension to methods

- "Who did skull-stripping on a T1 data set with the same MRI parameters as mine? What parameters did they choose that worked well?"
- "How did Noah find the positions of V1/V2/V3 from an anatomical MRI in an adult? Will it work on a child's brain?"
- "I would like to run a machine-learning algorithm 1000 times on a variety of data sets and store the learned parameters and keep a searchable record"
- "What analyses did I run on these data last year? Where did I store the results?"
- "How did Joyce find the arcuate fasciculus from her diffusion data? Can I use the same methods on my diffusion data?"
- "What process was used to create Figure 3 in the paper by Jones? If I change this parameter, will the results be similar?"

Scitran extension for Stanford community data sharing

- Interactions with archival databases (e.g., NDAR, ADNI)
  - We downloaded much of ADNI by hand
  - We have Simons VIP
  - We might download NDAR
  - We have part of the HCP
- Integration with publishers and libraries (e.g., SDR)
- Many SoM colleagues have been working with us to incorporate their special data sets.

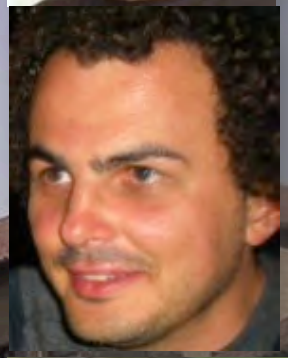
Thanks from the Project on Scientific Transparency

- Our fears
  - Sustainability
  - Hiring and keeping an excellent team
- Our plan
  - Open-source
  - Commercial venture

Summary

- Scientific data management tools are needed to support reproducible research in the era of complex science
- We have made a start for neuroimaging with the open-source scitran and corporate partner, Flywheel Exchange
- Technology exists and needs development; social barriers and risks remain significant

Michael Perry*



Renzo Frigato*

Gunnar Schaefer*

Bob Dougherty

We thank the Simons Foundation and the Arnold Foundation for their support

Commercial disclosure*:
We are active participants in a commercial venture related to the work described here, Flywheel Exchange, LLC.