HCP Informatics: ConnectomeDB and Cloud-based Processing

Daniel Marcus, PhD

Associate Professor & Director, Neuroinformatics Research Group @ Washington University School of Medicine

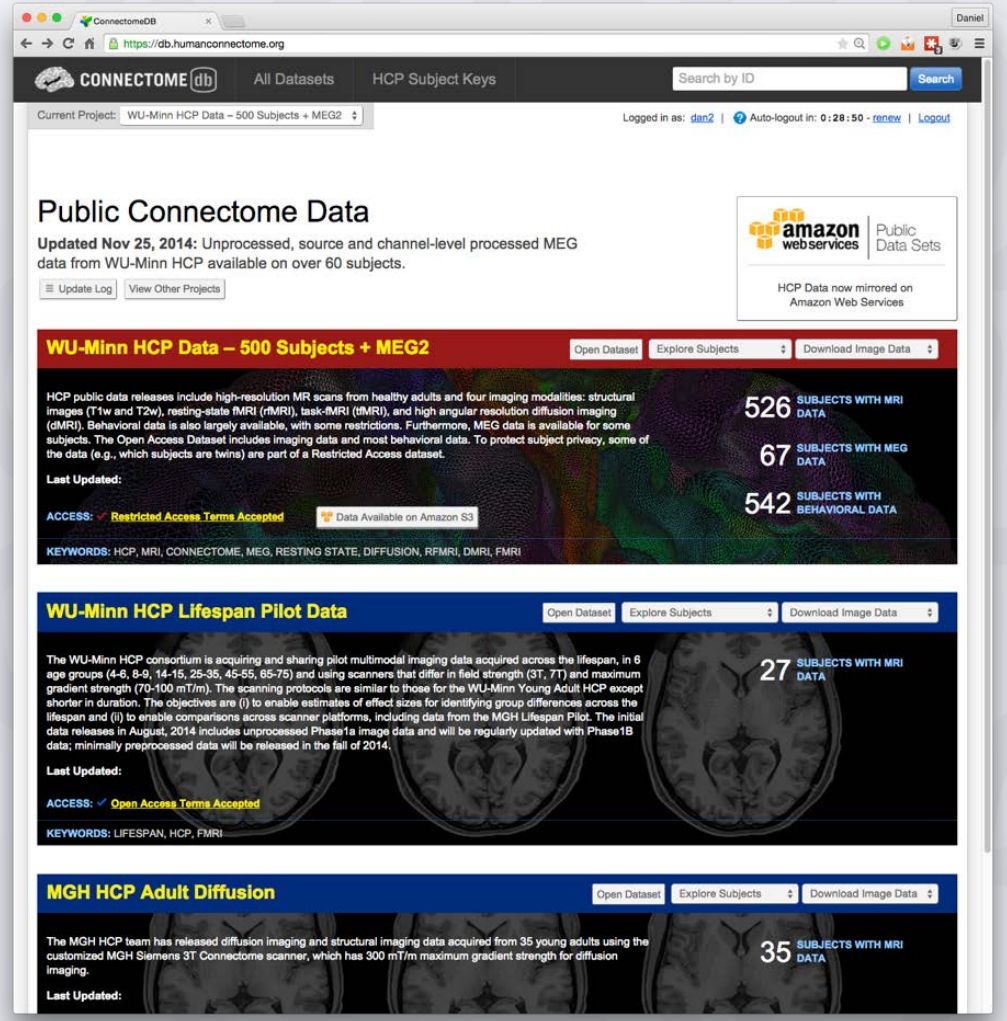
NRG Neuroinformatics Research Group

Outline

- ConnectomeDB
  - Activities from the browser
  - Activities from command line
- Cloud computing
  - HCP data on the cloud
  - HCP pipelines and analysis on the cloud
- Running your own HCP Informatics System

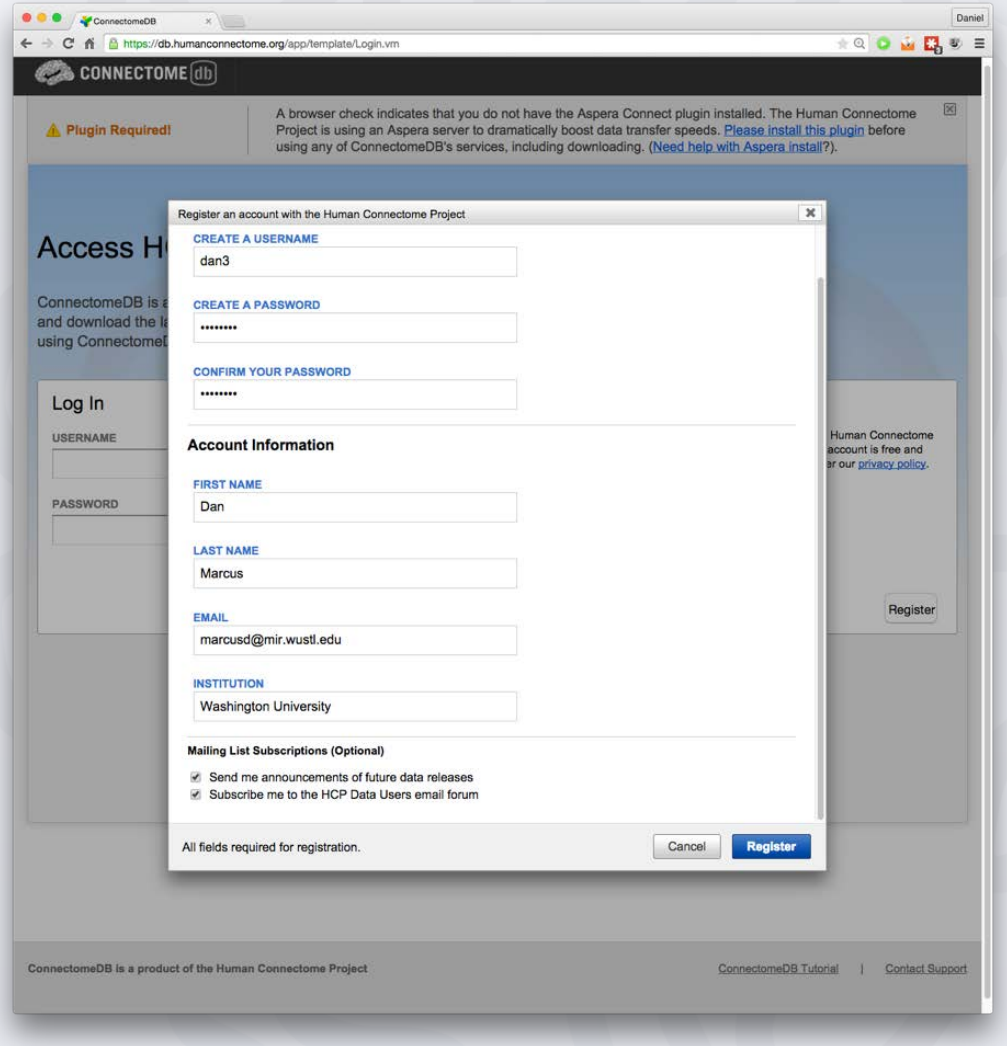
What is ConnectomeDB?

- Public resource for accessing HCP data
- Includes WU-Minn data along with MGH and other data sets
- Future home of Connectome Coordination Facility (CCF) data



Accessing ConnectomeDB

- Registration required
- Data use terms: open and restricted.
- Aspera plugin required for bulk downloads (i.e. packages)
- Browser for easy navigation
- Command line for scripted access



Alternative sources of HCP data

- Connectome in a Box
- Amazon Public Data Sets
- All options require first signing the data use terms in ConnectomeDB

Data organization

- The 4 standard imaging sessions have been merged into 1 “mega” session.
- The scans in the mega session are ordered in a standard sequence (not necessarily reflecting the temporal order).
- Key data files have been assembled into “packages” to streamline distribution.
- Some interesting processed data sits outside the package structure.

Packages

- Specific packages for structural, resting, diffusion, and each task.
- File organization is well documented here: http://www.humanconnectome.org/documentation/S500/HCP_S500+MEG2_Release_Appendix_III.pdf
- Same file organization on ConnectomeDB, CinaBox, and Amazon.
- All files can be retrieved individually from ConnectomeDB.

CONNECTOMEDB ACTIVITIES FROM THE BROWSER

ConnectomeDB Browser Demo -- Main themes

- Basic navigation
- Some special fields (metadata)
- Filtering data
- Saving subject groups
- Digging deeper into individual subjects

CONNECTOMEDB ACTIVITIES FROM COMMAND LINE

Why use the command line?

- ConnectomeDB has a rich programming interface (API)
  - Use the XNAT REST API
  - https://wiki.xnat.org/display/XNAT16/Using+the+XNAT+REST+API
- Build ConnectomeDB access directly into scripts.
- Batch download specific files.
- WARNING: The API does NOT use Aspera.

Some handy command line tools

- CURL
  - Access specific URLs
- XNAT Data Client
  - Access data using XNAT tags
- Python, bash, or your favorite scripting language
  - Access data programmatically

ConnectomeDB Command Line Demo -- Main themes

- Basic REST API patterns
- Accessing subject groups
- Accessing specific files follows standard package structure
- Querying on subject metadata
- Batch retrieval of files

HCP ON THE CLOUD

MEMES & FUNNY PICS FRABZ.COM

What is the “cloud”?

<u>Amazon Web Services (AWS)</u>

- “’Cloud Computing’, by definition, refers to the on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing.”
- Simple Storage Service (S3): File storage
- Elastic Compute Cloud (EC2): Computers
- Many other services (databases, block storage, workflow, etc.) not used by HCP

Why use the cloud for HCP projects?

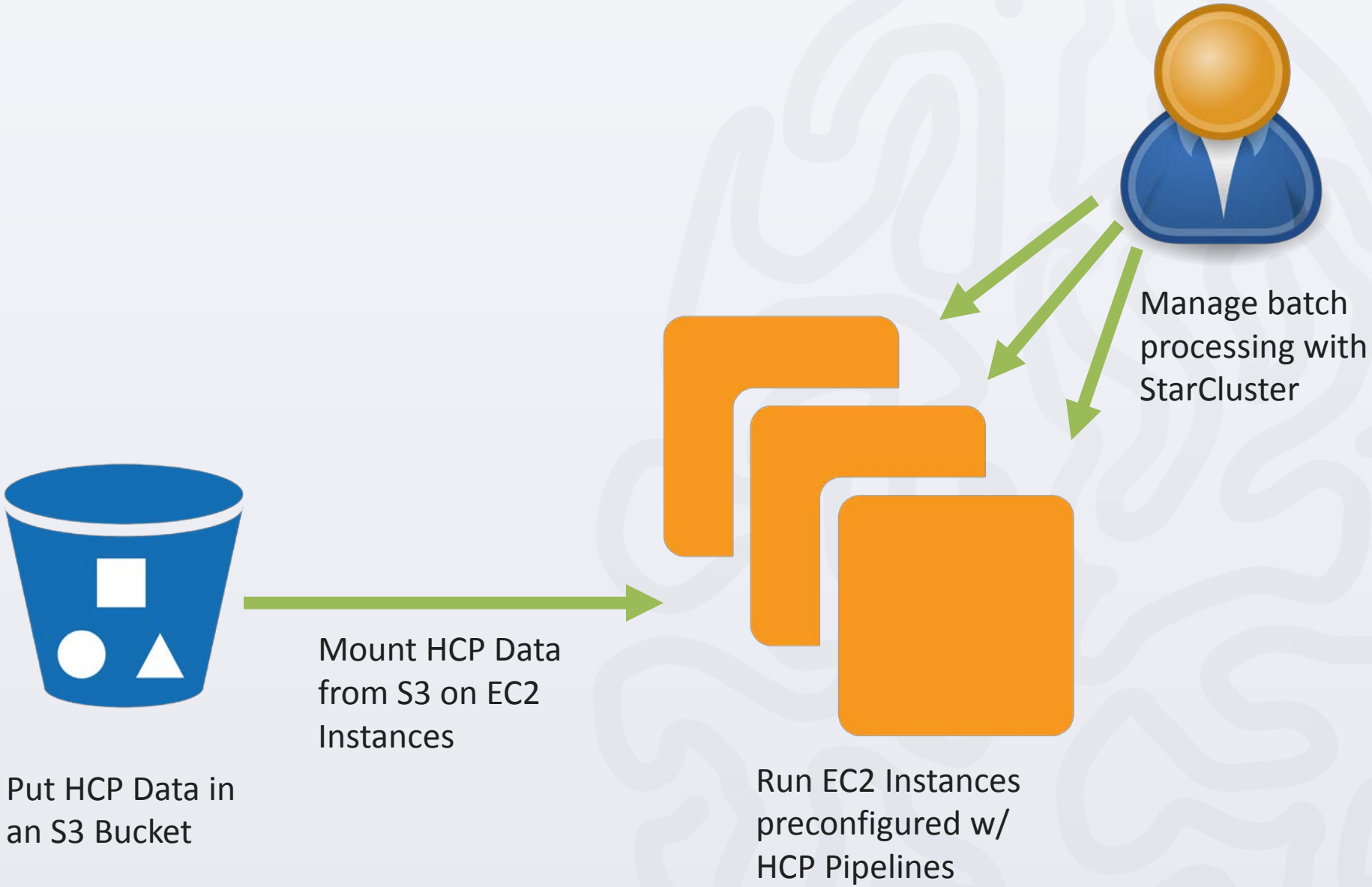
- HCP data is hosted on S3
  - You don't need to download and store it locally!
  - You don't need to buy Connectome in a Box!
- HCP pipelines are preconfigured on EC2
  - You don't need to install and configure the pipelines locally.
  - You don't need to have local computing resources to process large quantities of data.
- The cloud ain't free (but it can be pretty cheap).

Simple Storage Service (S3)

- Files are stored on S3 in “buckets”, which are used to partition and manage the data.
- Each file is accessible at a specific URL.
  - Example: http://bucket1.s3.amazonaws.com/file1
- But working with URLs is complicated – file systems are easier.
- S3fs can be used to “mount” an S3 bucket as a file system on a computer – they look just like regular files.

Elastic Compute Cloud (EC2)

- Computers on EC2 are called “instances”.
- “Amazon Machine Images” (AMIs) are preconfigured computer templates with specific applications preinstalled.
- Each instance runs a selected AMI.
- Each instance runs on selected “hardware”.
- Run lots of instances to build your own high performance compute cluster.
- Use Starcluster to manage jobs on your cluster


Manage batch processing with StarCluster
Mount HCP Data from S3 on EC2 Instances
Put HCP Data in an S3 Bucket
Run EC2 Instances preconfigured w/ HCP Pipelines

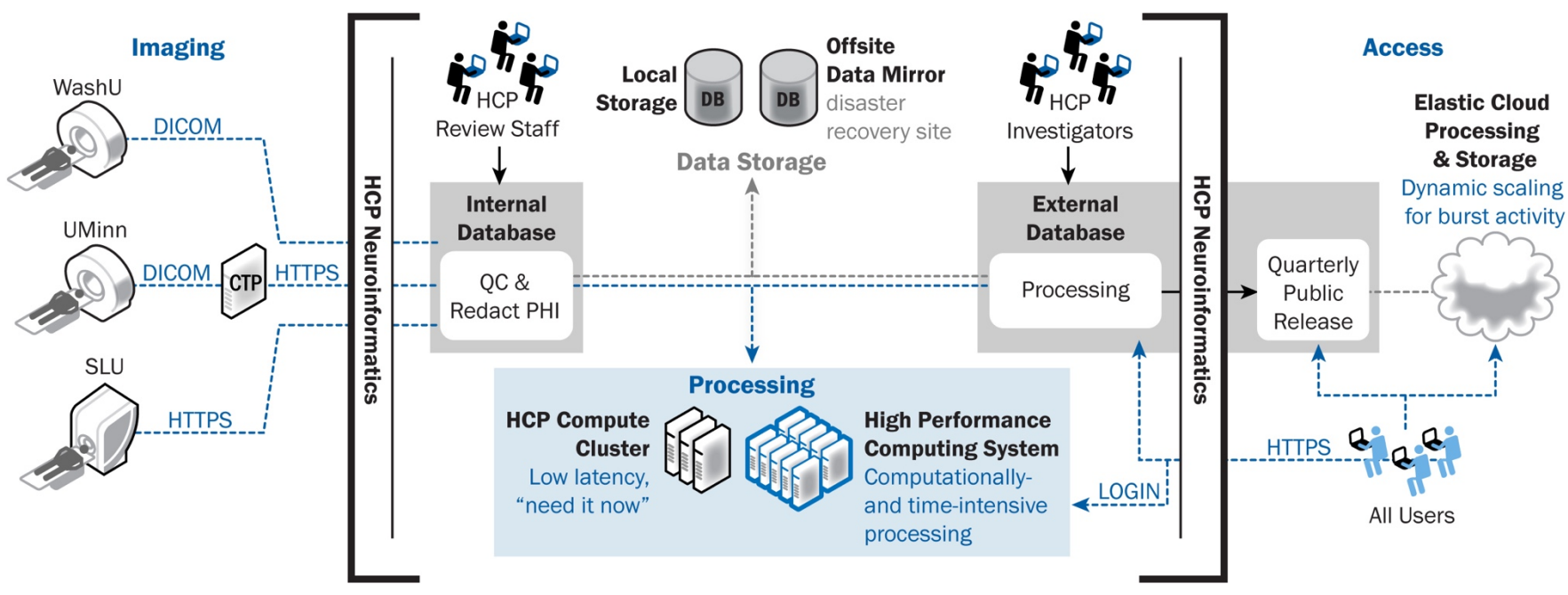
Strategies for using AWS

- Use Elastic Block Storage (EBS) for your own data.
- Use EC2 Spot Instances.
- Fork the NITRC HCP AMI to customize pipelines.

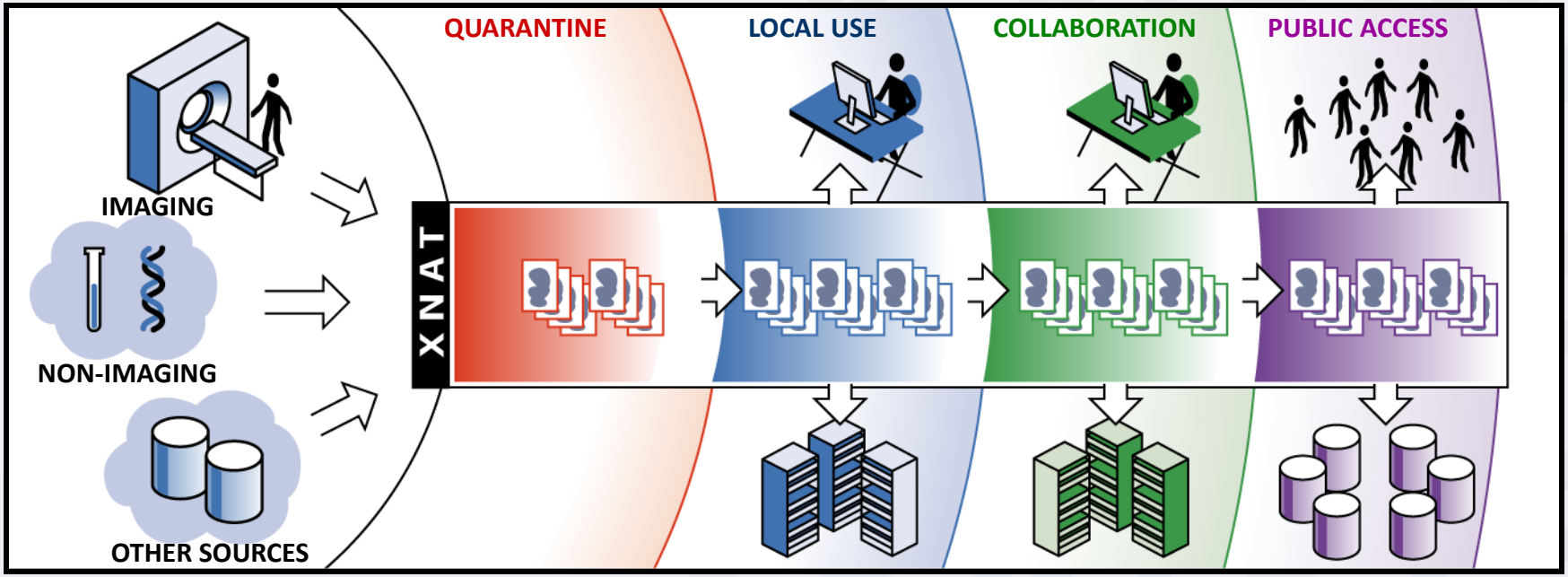
RUNNING YOUR OWN HCP INFORMATICS SYSTEM

Why run your own HCP Informatics System?

- You are running an HCP-like study.
- You enjoy pain.



The XNAT workflow



- Data organization
- Quality control
- Data access
- Security
- Visualization
- Automation
- Integration
- Data sharing

XNAT is…

Feature rich

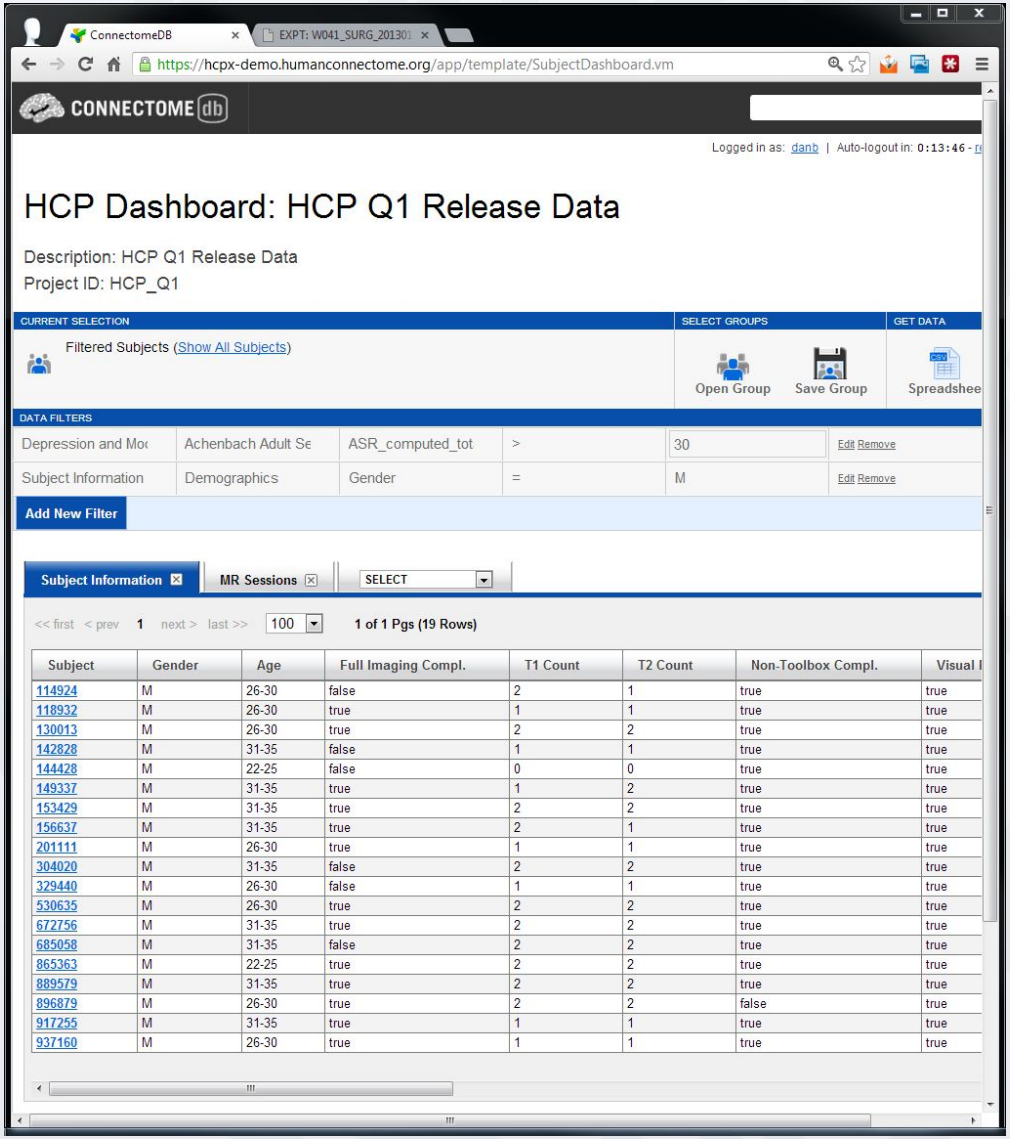
Archive, manage, process, view, and share imaging and related data.

A platform

Clinical/translational research
Institutional repositories
Multi-center studies
Data sharing

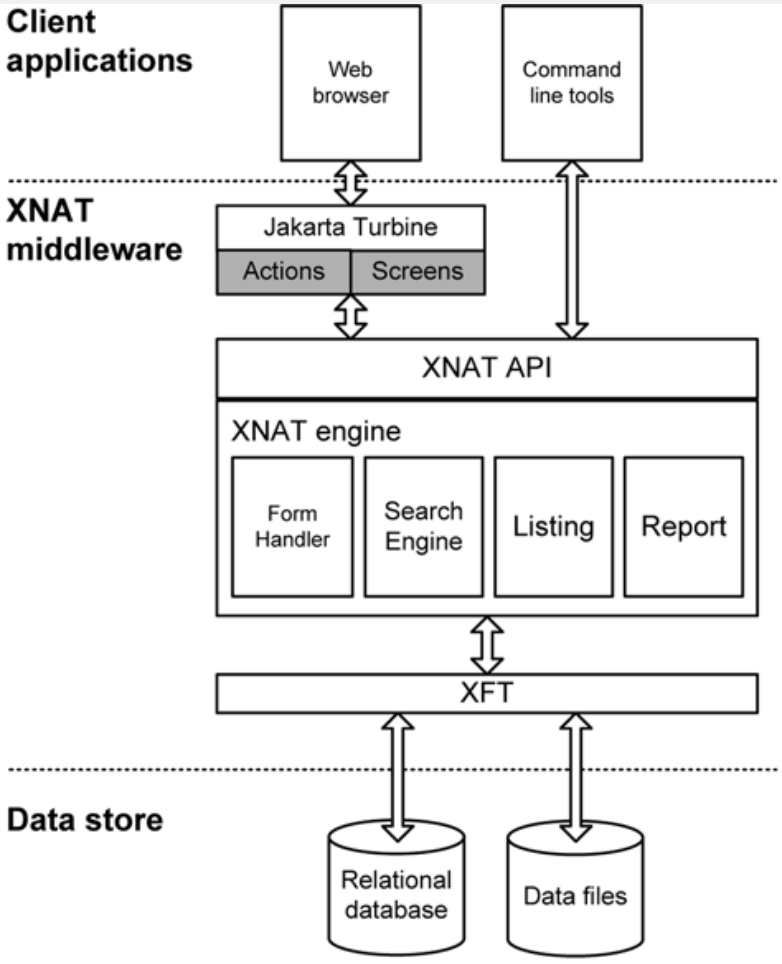
Open

Open source
Open API
Free (though commercial support is available)
Used by organizations around the world.



Subject	Gender	Age	Full Imaging Compl.	T1 Count	T2 Count	Non-Toolbox Compl.	Visual I
114924	M	26-30	false	2	1	true	true
118932	M	26-30	true	1	1	true	true
130013	M	26-30	true	2	2	true	true
142828	M	31-35	false	1	1	true	true
144428	M	22-25	false	0	0	true	true
149337	M	31-35	true	1	2	true	true
153429	M	31-35	true	2	2	true	true
156637	M	31-35	true	2	1	true	true
201111	M	26-30	true	1	1	true	true
304020	M	31-35	false	2	2	true	true
329440	M	26-30	false	1	1	true	true
530635	M	26-30	true	2	2	true	true
672756	M	31-35	true	2	2	true	true
685058	M	31-35	false	2	2	true	true
865363	M	22-25	true	2	2	true	true
889579	M	31-35	true	2	2	true	true
896879	M	26-30	true	2	2	false	true
917255	M	31-35	true	1	1	true	true
937160	M	26-30	true	1	1	true	true

XNAT Architecture



In a Nutshell, XNAT...

...has had 2,489 commits made by 53 contributors representing 643,792 lines of code

...is mostly written in Java with an average number of source code comments

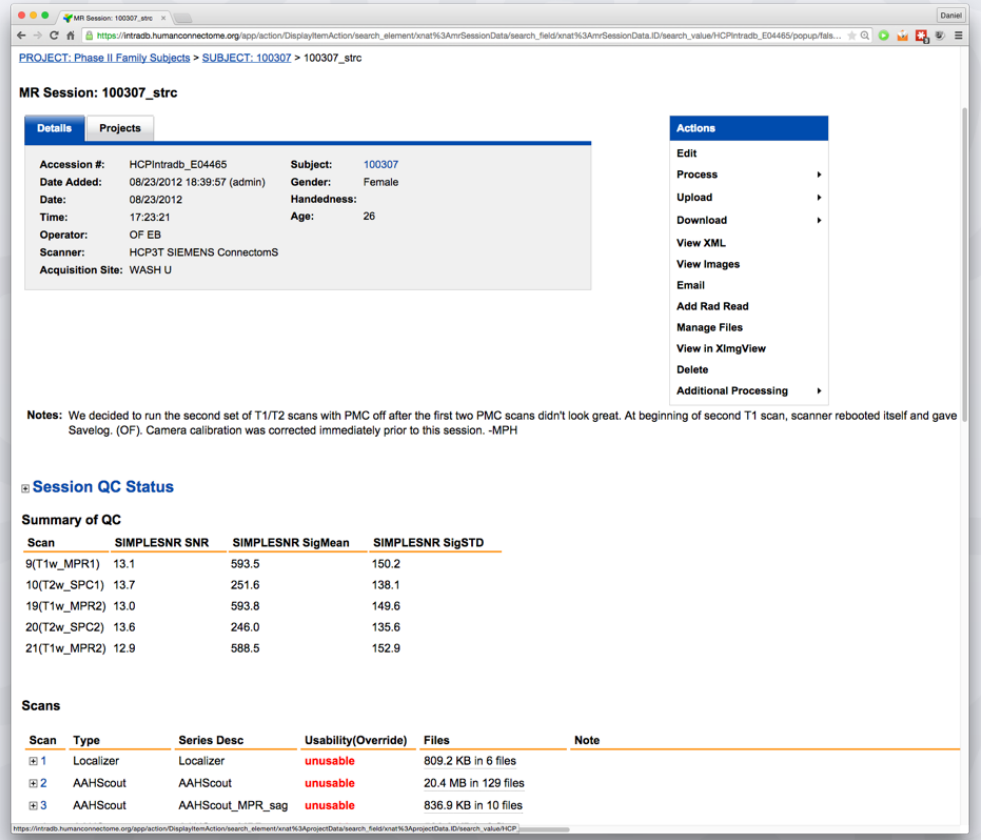
...has a codebase with a long source history maintained by a very large development team with increasing Y-O-Y commits

...took an estimated 176 years of effort (COCOMO model) starting with its first commit in April, 2010 ending with its most recent commit 1 day ago

Source: https://www.ohloh.net/p/XNAT

IntraDB Demo – Main Themes

- Receives data directly from scanners.
- Includes extensive QC pipelines and annotations.
- Accessible only to HCP personnel.
- Data that passes QC is exported to ConnectomeDB



Summary

- Get your data on ConnectomeDB, Connectome in a Box, or Amazon Public Data Sets.
- Don’t be afraid to get to know the command line.
- Get ready for some cloud computing!

Thanks to…

- Amazon Web Services
  - Hosting of HCP data thru Public Data Sets project
  - $100 credits to all course attendees
- Neuroimaging Informatics & Tools Clearinghouse (NITRC)
  - NITRC Computing Environment w/ HCP pipelines