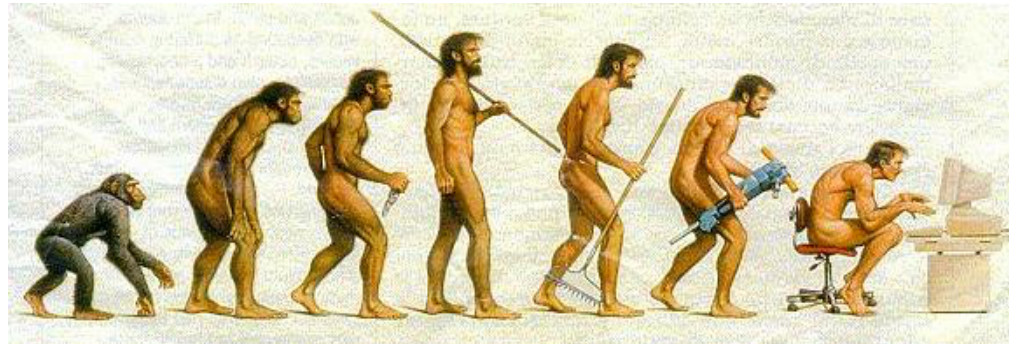


Scientific Data Asset Management — The Missing Link in Data Driven Discovery



Carl Kesselman

University of Southern California



Acknowledgements

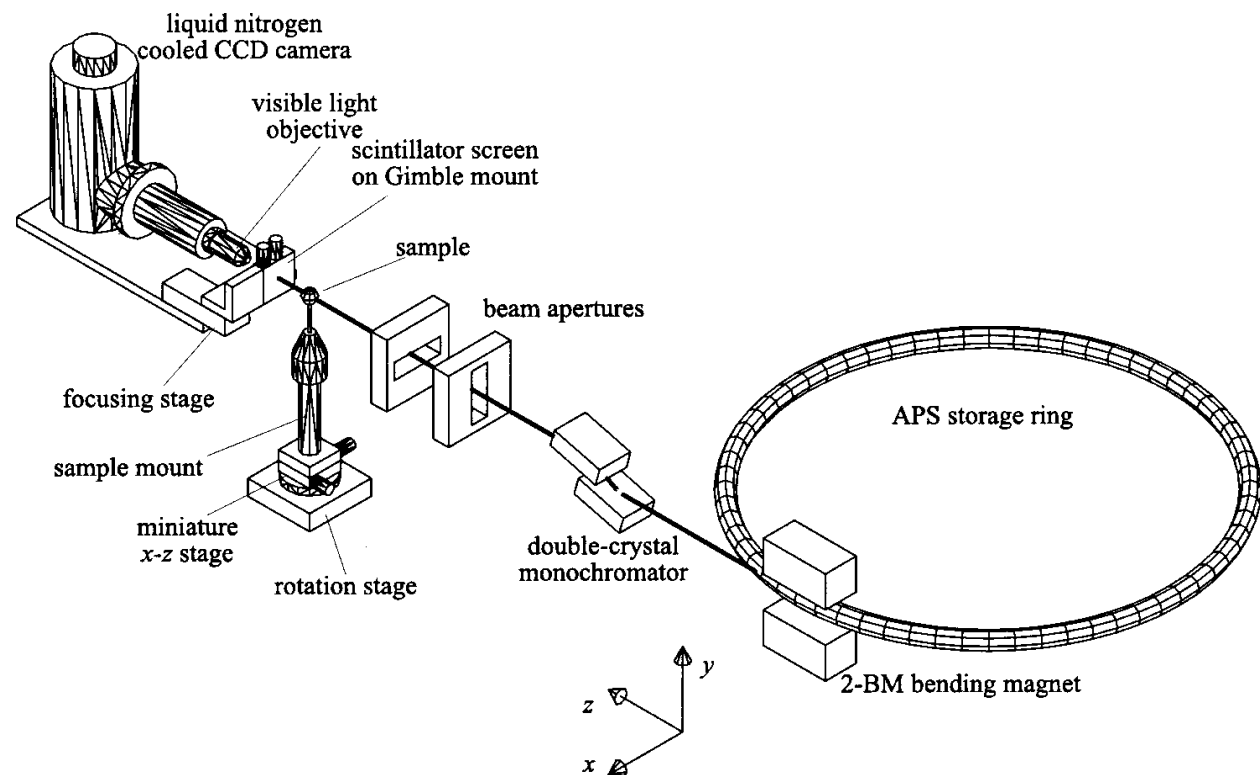


- ▶ Karl Czajkowski, Mike D'Arcy, Hongsuda Tangmunarunkit, Robert Schuler, Anoop Kumar, Alejendro Bugacov
- ▶ Kristi Clark, Lu Zhau
- ▶ Ian Foster, Kyle Chard, Ravi Madduri,
- ▶ Mike Hanson, Jeff Su, Ray Stevens

- ▶ Funded in part by NIH Big Data for Discovery Science Center of Excellence.

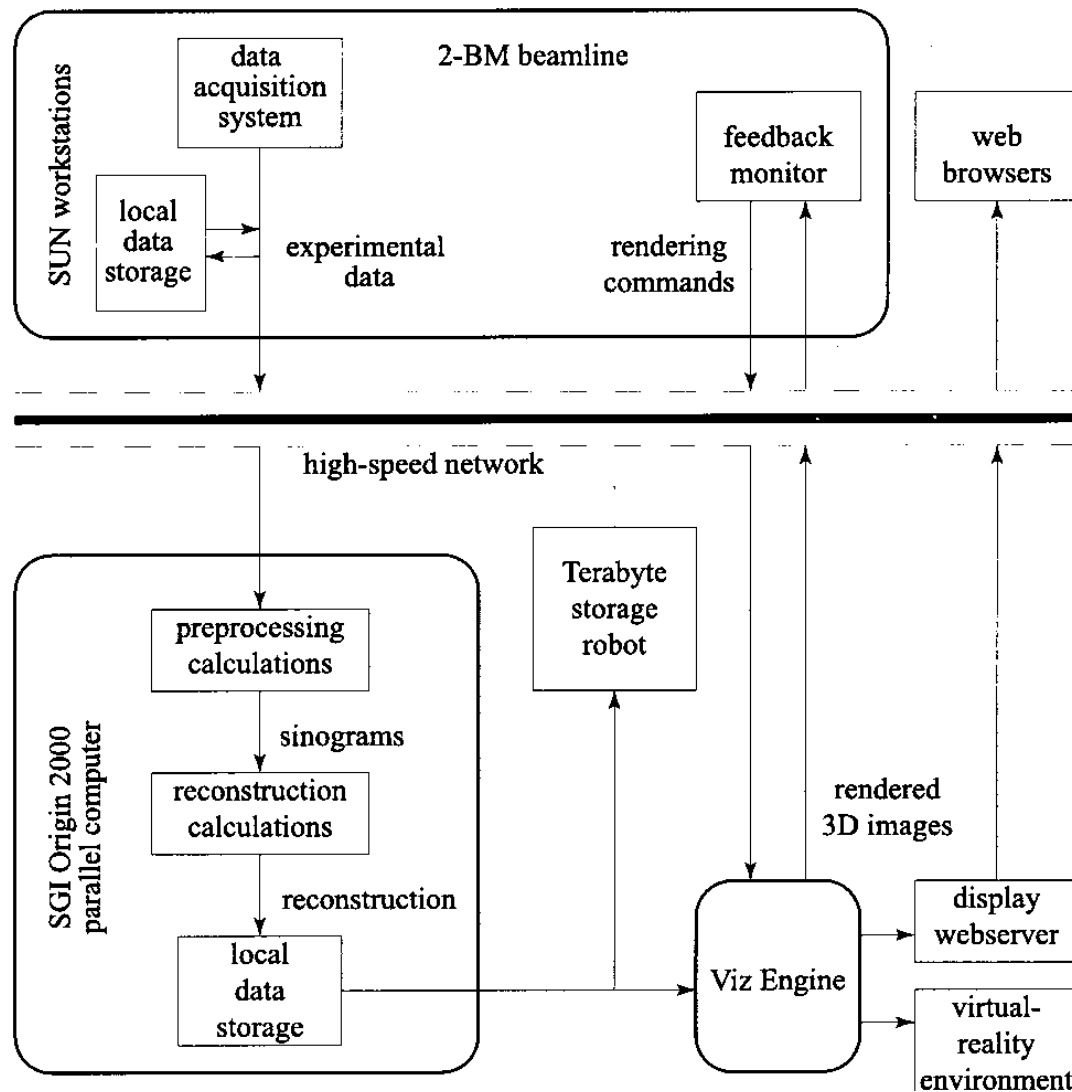
Set the way back machine....

- In 2000 we described how real-time data acquisition could be integrated into the Grid for diffraction studies and tomography

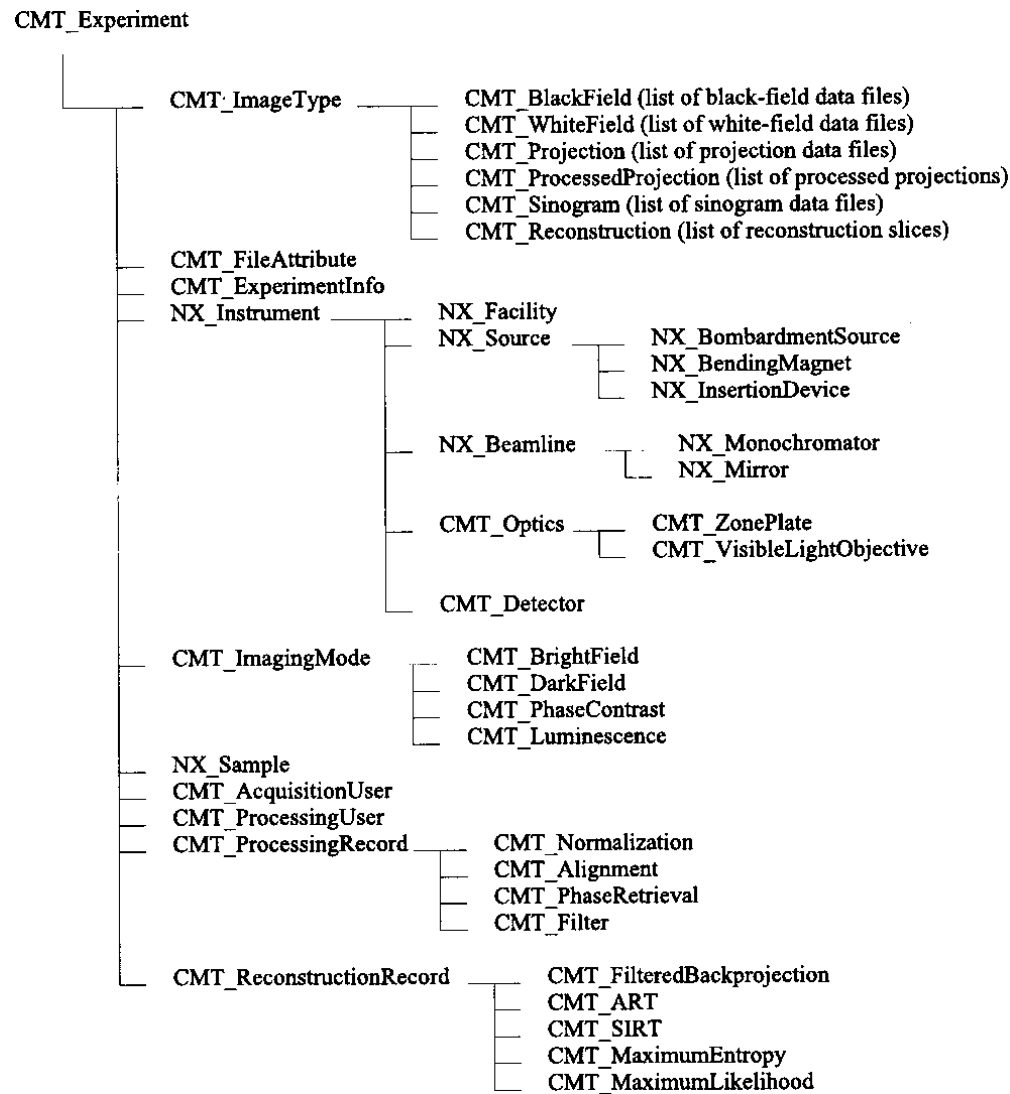


System architecture....

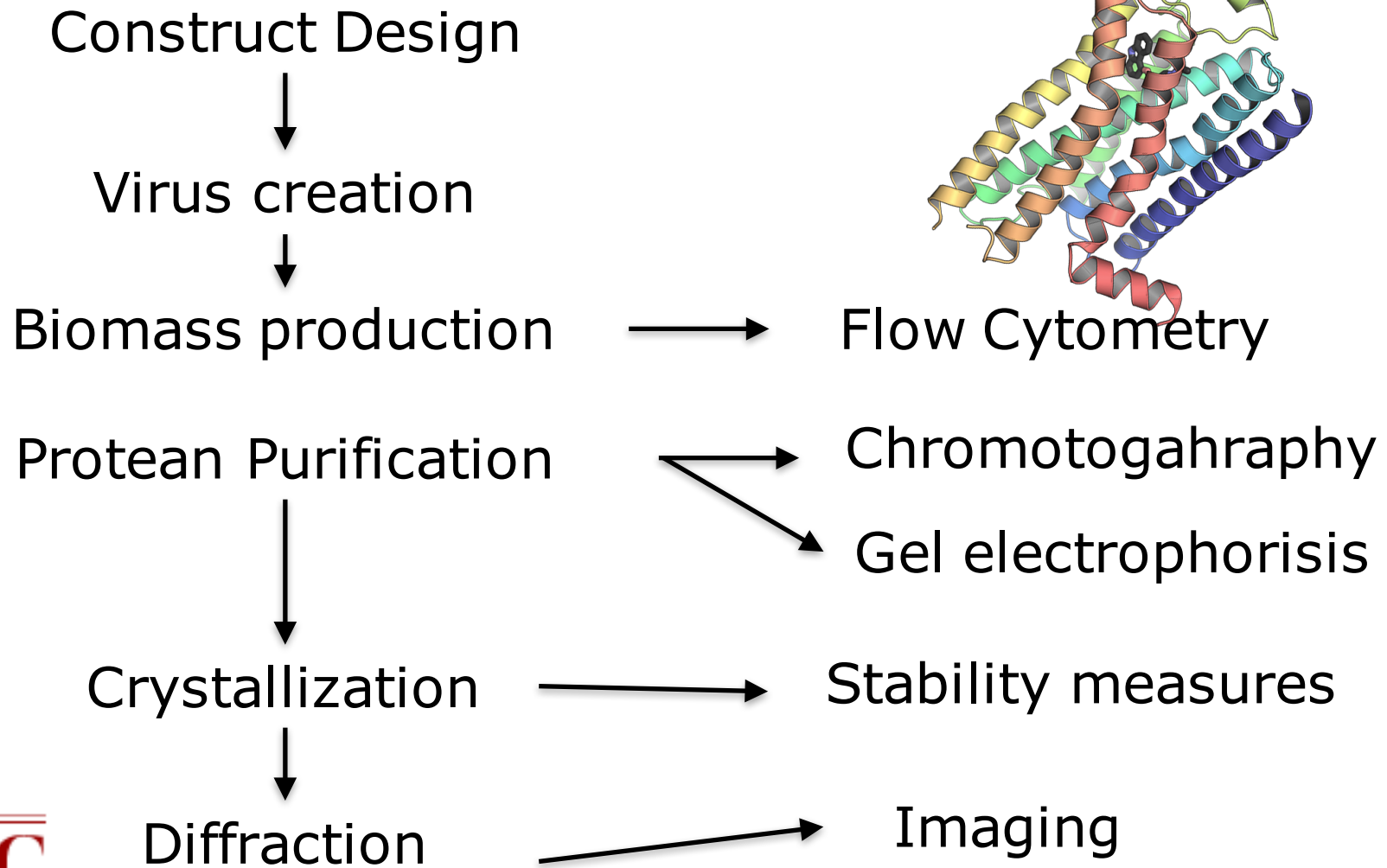
Advanced Photon Source



Fancy file systems....

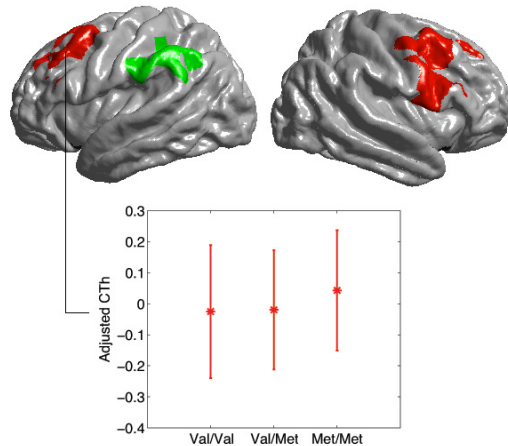


The whole story (kind of....)

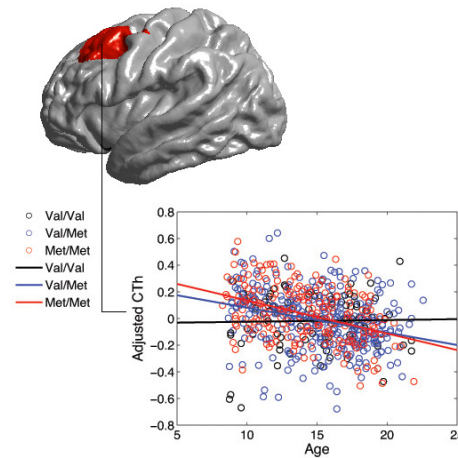


PheWAS findings

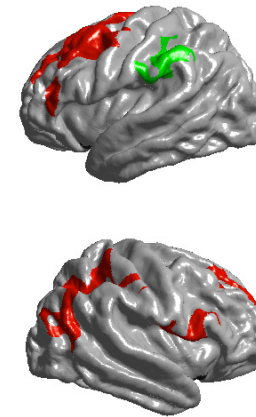
Robust COMT-CTh associations



Robust COMT effect on CTh-age associations



More prominent COMT effect in Caucasians



Shaw, Molecular Psychiatry (2009) 14, 348–355

Raznahan, Neuroimage (2011) 57, 1517-23

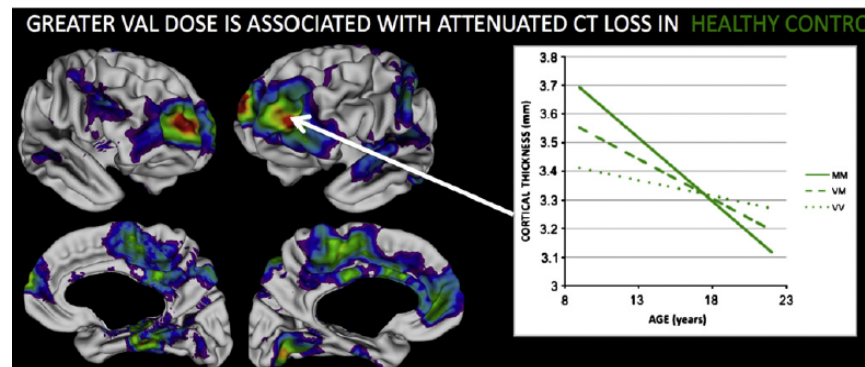
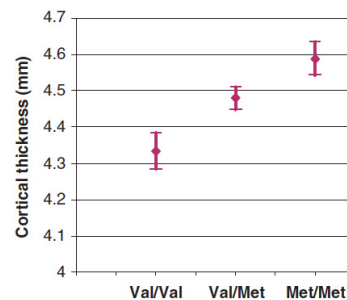
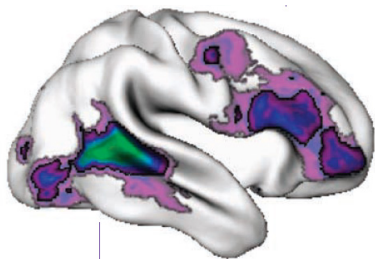
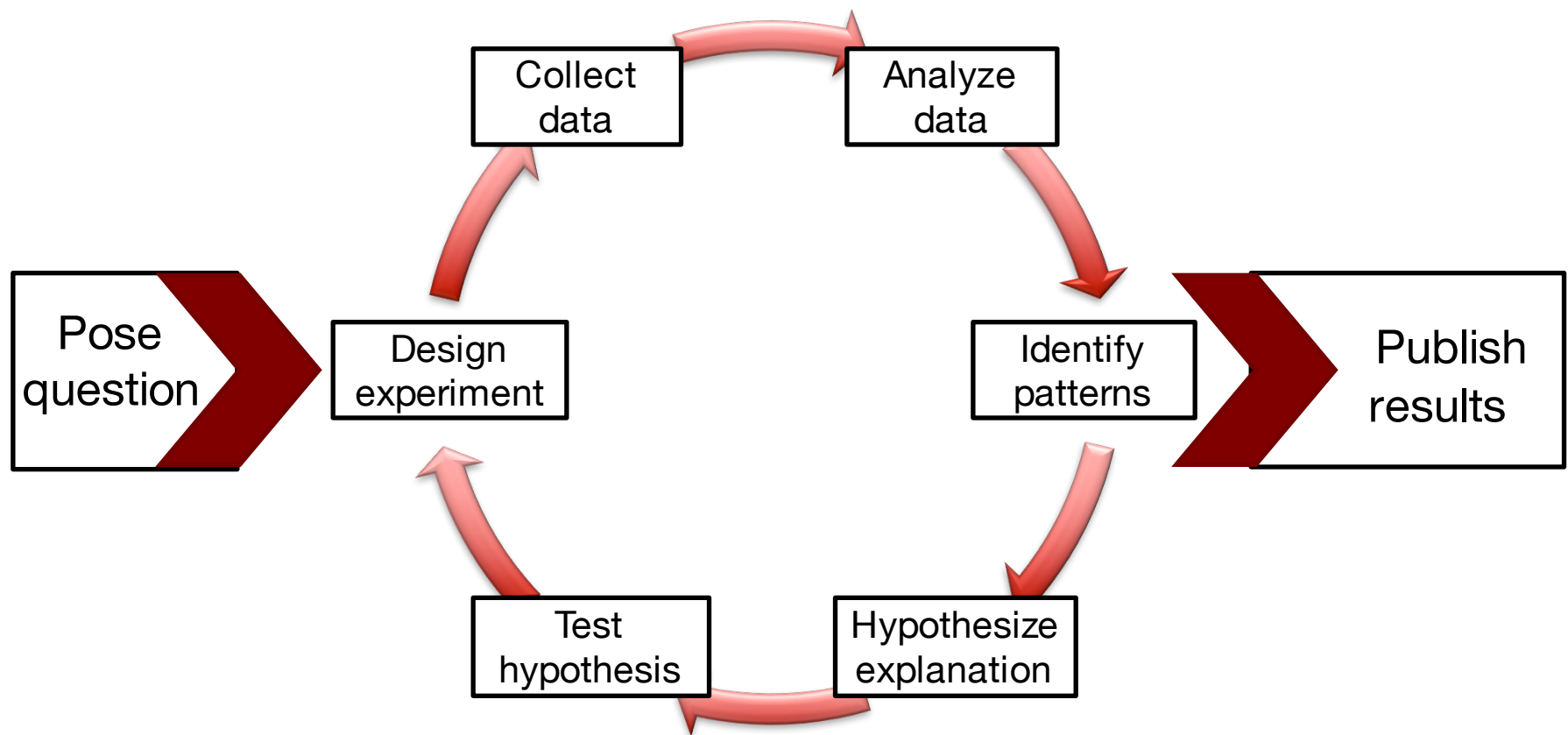


Image PheWAS



1. Assemble Data Collections
2. Identify subjects with images and extract images
3. Compute image phenotypes
 - Use Freesurfer with different atlases and computed measures
4. Associate Freesurfer results with each subject.
5. Quality control on derived data. Rerun on bad results
6. Identify subset of subjects that have variant of interest in SNP being considered
7. Collect up all phenotype data associated with identified subset
8. Do correlation analysis of phenotypes for the SNP to look for predictive correlations.
9. Repeat until discovery

How do we accelerate discovery?



A view from 1960.....



“my choices of what to attempt and what not to attempt [are] determined to an embarrassingly great extent by considerations of clerical feasibility, not intellectual capability”

Man-Computer Symbiosis

J. C. R. Licklider

The View From 2016



Scientists report *50-80% of their time* is spent “wrangling” messy data, not analyzing it

- The problem is not the cost of computing!!

Repeatability of results from papers is shockingly low: 10%

- ▶ Lack of comprehensive tools for organizing, contextualizing, and sharing data
- ▶ Ad hoc processes and practices for managing and sharing information
- ▶ Messy Data → Reusable Data → Discovery



How to get from point A to point B?

GPCR Data Explorer

Selected by:

Clear All Filters

Site **USC**

Biomass

Current Status **completed**

Showing 1-3 of 3 results, sort by: Select an attribute

Biomass ID	Uniprot Name	Biomass Date
IMPT-17887	MTR1A	2015-03-24
IMPT-18239	OPRK	2015-04-09
IMPT-18290	ADRB2	2015-04-10

© 2014-2015 University of Southern California



- ☐
- ☐ complete
- ☐ Complete
- ☐ completed
- ☒ completed
- ☐ Completed
- ☐ Completed
- ☐ Completed
- ☐ Completed
- ☐ completed, contaminated
- ☐ inprgress
- ☐ inprogress
- ☐ inprogress
- ☐ Inprogress
- ☐ InProgress
- ☐ In Progress
- ☐ Inrpogress
- ☐ requested
- ☐ requested
- ☐ Requested

esselman@globusid.org Logout

plete
plete
pleted
pleted
pleted
pleted
pleted
pleted, contaminated
ress
gress
gress
gress
gress
gress
gress
sted
sted
sted

What if....



- ▶ Every piece of data produced in was “citable”
 - Microscope, flow cytometry, mass spec, sequence, mouse, zebrafish, material sample
- ▶ Data flowed instantly and seamlessly
 - From points of production/acquisition
 - between dynamically evolving research teams
- ▶ Data was contextualized
- ▶ You had automated support to help discover data, extract interesting features, point you to related data, assemble data sets...

It's the data, not the analysis!!

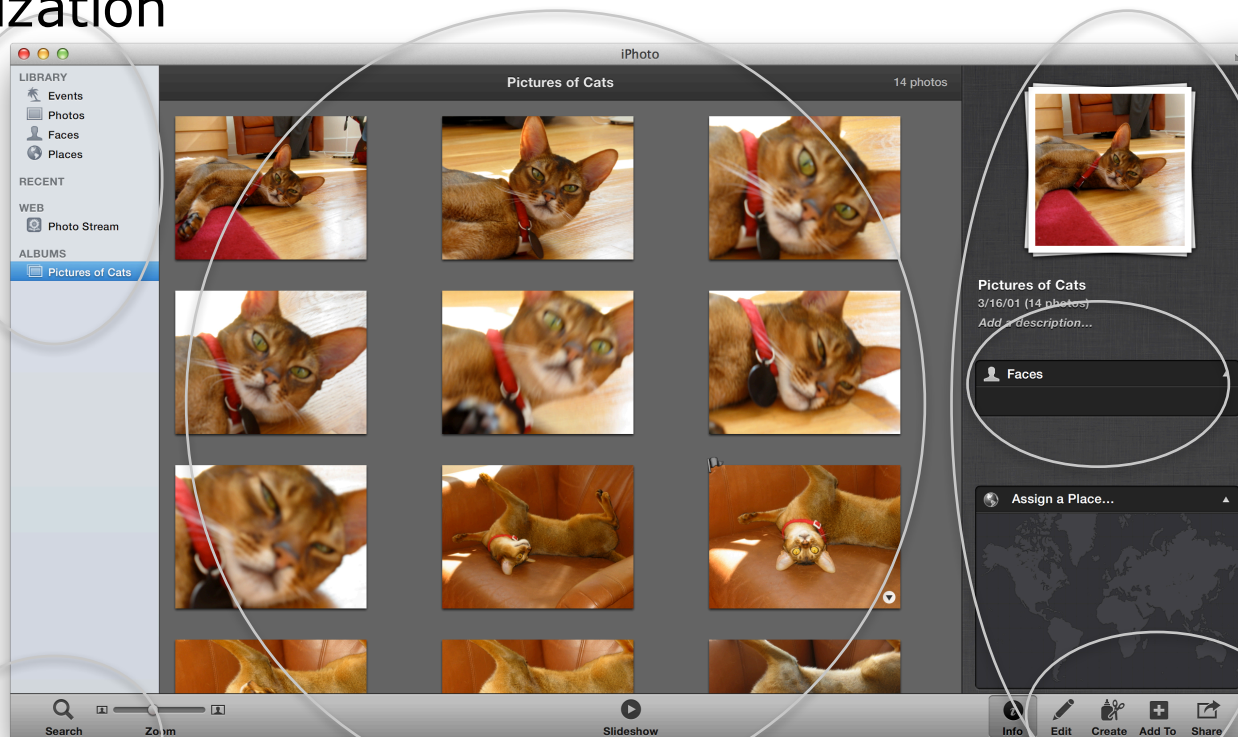


Data is a precious thing and will last longer
than the systems themselves.

Tim Berners-Lee

An Ecosystem for Data

Why don't we have tools for managing data sets of cancer and kidneys that are as good as the tools we have for managing data sets of cats and kids?
Flexible data organization



Editable
attributes
and
Automatic
metadata
analysis

Edit and
share



Full text
search

Data browsing

Apple iPhoto

Applied to other types of work?



- ▶ Can we create a reusable platform that enables us to address data centric integration of
 - devices,
 - computation,
 - human interactions
 - ...

Digital Asset Management



- ▶ “management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets”
- ▶ streamline free-form “creative” processes rather than enforce predefined business processes.
- ▶ Many commercial DAM offerings, but not well suited to biomedical data
 - Complex and diverse data types
 - Specialized data ingest requirements
 - Data size (big data)

Scientific Digital Asset Management



- ▶ Discovery Environment for Relational Information and Versioned Assets (DERIVA).
- ▶ Model discovery as process of creating and updating contextualized digital assets.
- ▶ Web services platform
 - “Data Oriented Architecture”
- ▶ Adaptive and extensible

Platform Elements



ERMRest

- ▶ **Object/Relational data store**
 - Pub/Search/Retrieve structured data

HATRAC

- ▶ **Object store**
 - Pub/Retrieve immutable objects

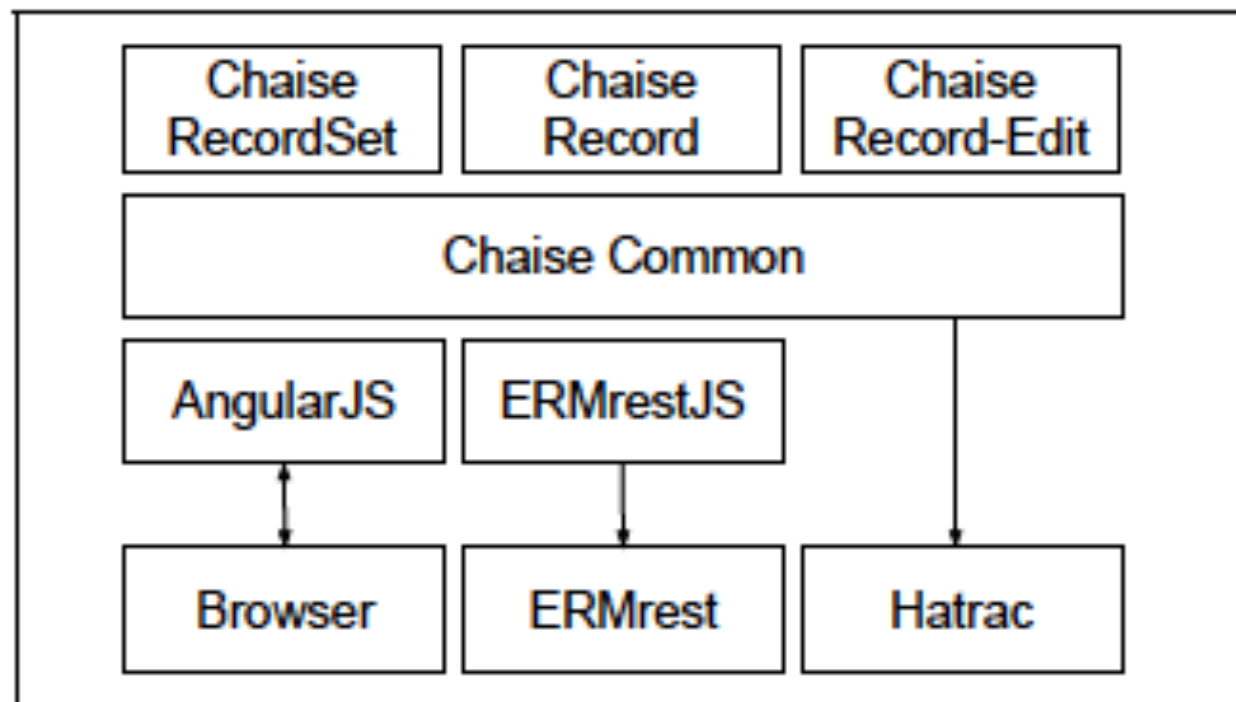
IObox

- ▶ **Batch publish/retrieval tool**
 - Watch file system and publish data bundle

Chaise

- ▶ **Model-driven UI**
 - Introspect and adapt to data model

Software ecosystem





- ▶ Configurable tools for enabling arbitrary endpoint
 - Files, databases, microscopes, etc.
 - IoT like
- ▶ Contextualize data based on time and location
 - Ruleset per location
 - Metadata extraction, publication to catalog, management of asset
 - Simple recovery mechanisms based on retry/notification
- ▶ Triggers per asset ingest pipeline in “cloud”

ERMRest



- ▶ Relational data storage service for web-based, data-oriented collaboration.
 - general entity-relationship modeling of data resources manipulated by RESTful access methods.
- ▶ RESTful interface → data views as named resource
- ▶ Focus on introspection and evolution
 - Data model can change over time to reflect evolving understanding of problem space

Chaise – Adaptive User Interface



► How little can we assume?

- discovery, analysis, visualization, editing, sharing and collaboration over tabular data (ERMRest).

► Makes almost no assumptions about data model

- Introspect the data model from [ERMrest](#).
- Use heuristics, for instance, how to flatten a hierarchical structure into a simplified presentation for searching and viewing.
- Schema annotations are used to modify or override its rendering heuristics, for instance, to hide a column of a table or to use a specific display name.
- Apply user preferences to override, for instance, to present a nested table of data in a transposed layout

One platform, many use cases



- ▶ High-resolution 2D and 3D microscopy
- ▶ GPCR protein conformation studies
- ▶ Kidney reconstruction using stem-cells
- ▶ Mapping dynamic synaptome in vivo
- ▶ Gene expression studies for craniofacial dysmorphia
- ▶ Digital cell line for cancer
- ▶ Developmental biology

Neuroimaging PheWAS



► What is PheWAS?

- One SNP -> a wide variety of neuroimaging phenotypes (inverse of GWAS)

► Why PheWAS?

- explores system-level genetic associations.

► Challenges

- Complexity, heterogeneity, and volume of the data
- Complex and sophisticated brain image processing
- Multiple-comparison correction
- Result visualization

Philadelphia Neurodevelopmental Consortium



- ▶ 8719 subjects in study
 - Baseline clinical elements
- ▶ 6 different SNP array chipsets resulting in a combined set of 1,873,486 distinct SNPs (out of a possible 85 million in the human genome).
 - The total combinatorial space of the genomic data is 5,435,533,460 (SNP, subject, allele) tuples across the 8719 subjects
- ▶ 997 of the subjects have MRI imaging data

Managing data collections

PheWas PNC Data Explorer

https://bdds-dev.isrd.isi.edu/phewas/pnc/search/#3/pncsubject?facets=(subject:gender::eq:M/metrics_v:tissue::eq:gray/metrics_v:prima

Search

BDDS PheWas PNC Data Explorer

root Logout

Search within: **Subject (452)**

[Tour](#) [Permalink](#)

Selected by:

Clear All Filters

Gender **M**

Tissue **gray**

Primary Lobe **occipital**

SNP ID **rs6265 , rs133885**

Showing 1-25 of 452 results, sort by:

Select an attribute

Switch view:

Sample ID	Subject ID	Birth Year	Age (Years)	Age (Months)	Race/Ethnicity	Gender
600031697545	PNC0004_M20	1990	20	242	AA	M
600039015048	PNC0006_M11	1999	11	139	EA, AA, HI	M
600039665619	PNC0007_M09	2001	9	113	EA	M
600054124128	PNC0011_M18	1991	18	223	AA	M
600062084650	PNC0014_M12	1998	12	145	AA	M
600084088680	PNC0015_M10	2000	10	125	EA	M
600109657100	PNC0018_M10	2001	10	123	EA	M
600110501017	PNC0019_M14	1995	14	176	EA	M
600114922498	PNC0021_M14	1996	14	173	AA	M
600116672720	PNC0022_M10	2000	10	127	EA	M
600137870077	PNC0025_M08	2002	8	104	EA	M
600173623767	PNC0027_M16	1995	16	196	EA	M
600185621034	PNC0028_M17	1993	17	215	AA	M
600209790043	PNC0029_M09	2001	9	109	AA	M
600209995267	PNC0030_M12	1997	12	152	AA	M
600210683444	PNC0032_M15	1995	15	187	EA	M
600263649795	PNC0035_M13	1996	13	165	AA	M
600282088524	PNC0036_M18	1992	18	219	AA	M

Search within attributes...

CHOOSE ATTRIBUTES:

Subject ID 39556

Sample ID 39556

Gender 83072

Birth Year 38764

Age (Years) 39556

Age (Months) 39556

Race/Ethnicity 39204

Atlas 39556

Region 39556

Hemi 39556

Tissue 46748

Structure 39556

Division 39556

Primary Lobe 258910

Secondary Lobe 39556

Volume 39556

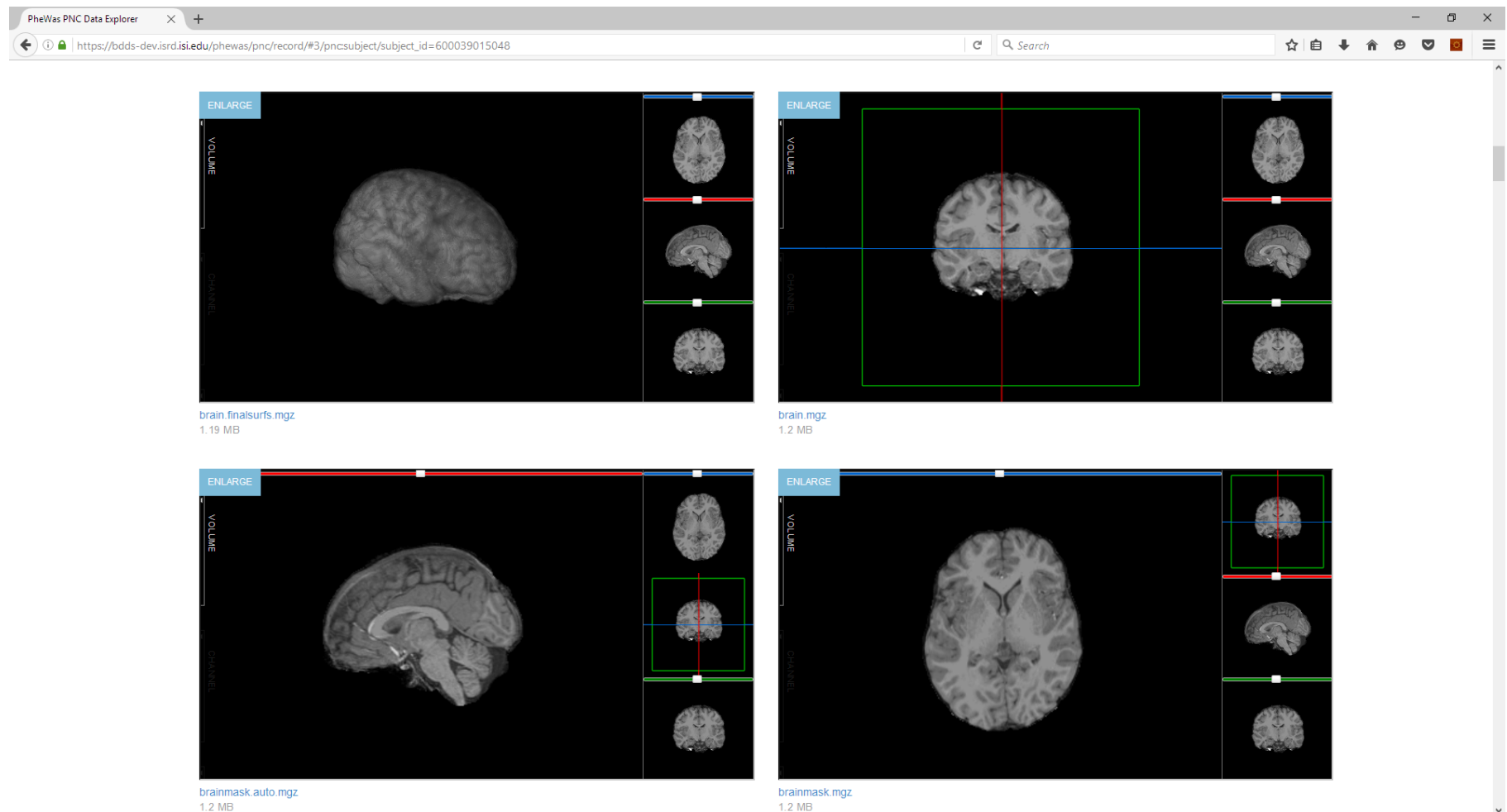
SNP ID 181852

Genotype 39556

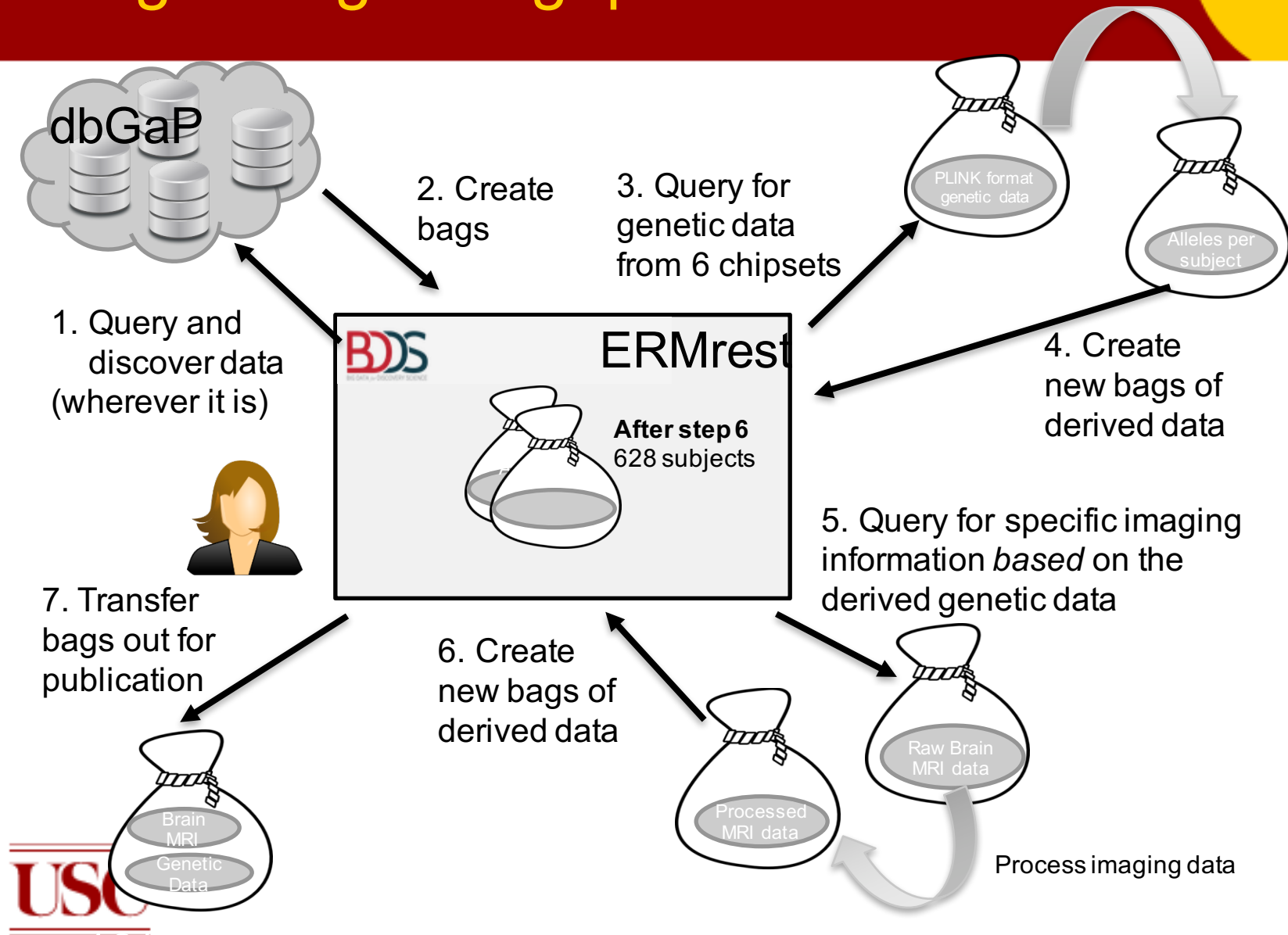
Array Chipset 39556

View all attributes (31)

Heterogeneous source data




Bags bridge the gap between tools



Details on one data element

PheWas PNC Data Explorer

https://bdds-dev.isrd.isi.edu/phewas/pnc/record/#3/pnc:subject/subject_id=600039015048

 PheWas PNC Data Explorer

SUBJECT

Subject ID	PNC0006_M11
Sample ID	600039015048
Gender	M
Birth Year	1999
Age (Years)	11
Age (Months)	139
Race/Ethnicity	EA, AA, HI

+ SUBJECT IMAGING (72)

+ SUBJECT GENETIC VARIANTS (9)

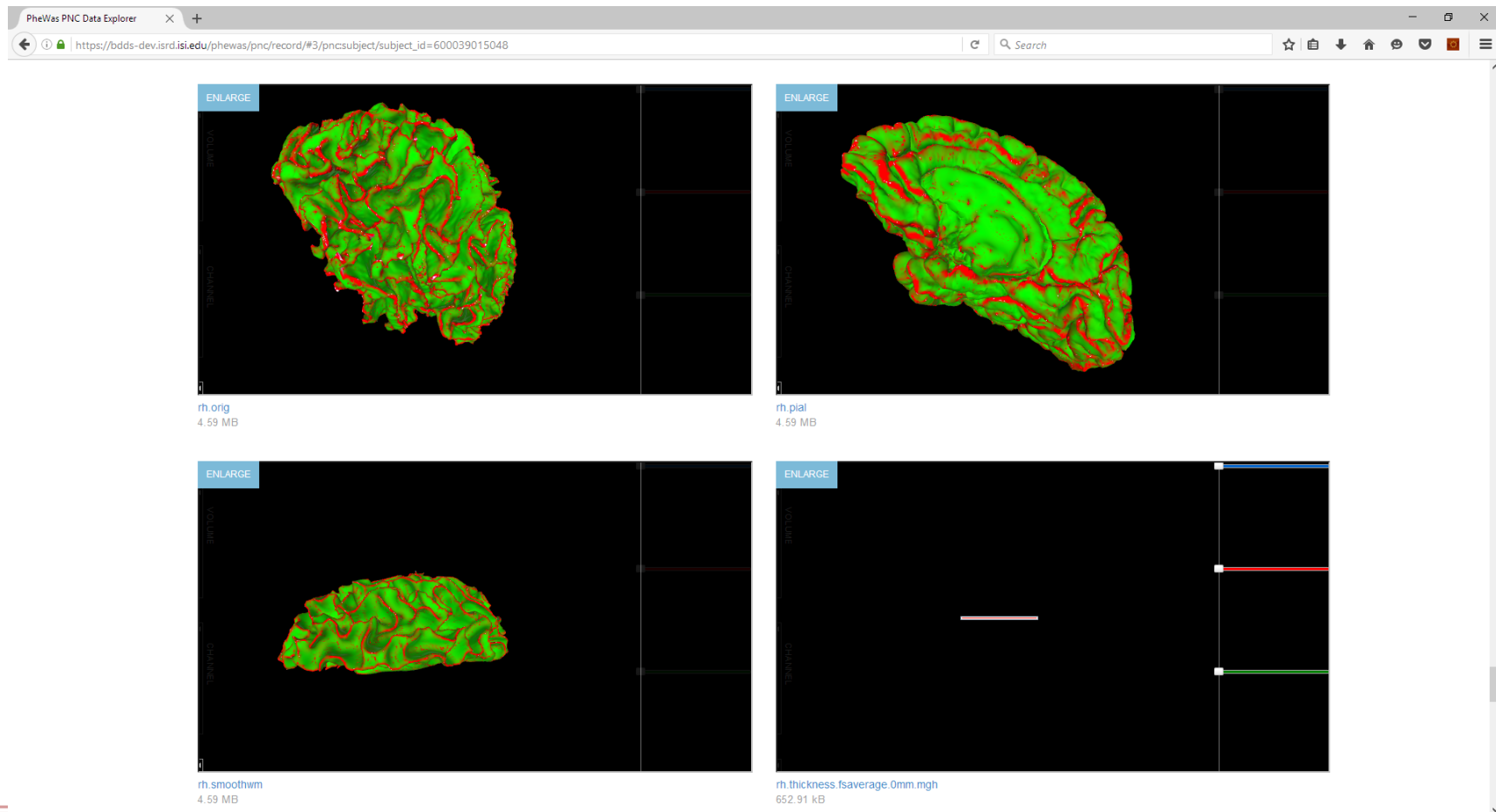
+ SUBJECT IMAGING METADATA (8)

+ SUBJECT PARCELLATION METRICS (381)

+ SUBJECT PARCELLATION METRICS METADATA (10)

+ SUBJECT PHENOTYPES (1)

QC on derived data



Complex data relationships...

PheWas PNC Data Explorer

https://bdds-dev.isrd.isi.edu/phewas/pnc/record/#3/pncsubject/subject_id=600039015048

Search

- SUBJECT GENETIC VARIANTS (9)

VIEW Default | Transpose

SNP ID	Genotype	Array Chipset
rs10868235	0/1	Human610_Quadv1_B
rs1147198	0/0	Human610_Quadv1_B
rs133885	0/1	Human610_Quadv1_B
rs1867283	0/1	Human610_Quadv1_B
rs3739722	0/0	Human610_Quadv1_B
rs4680	1/1	Human610_Quadv1_B
rs4767492	0/0	Human610_Quadv1_B
rs6265	0/0	Human610_Quadv1_B
rs786992	0/1	Human610_Quadv1_B

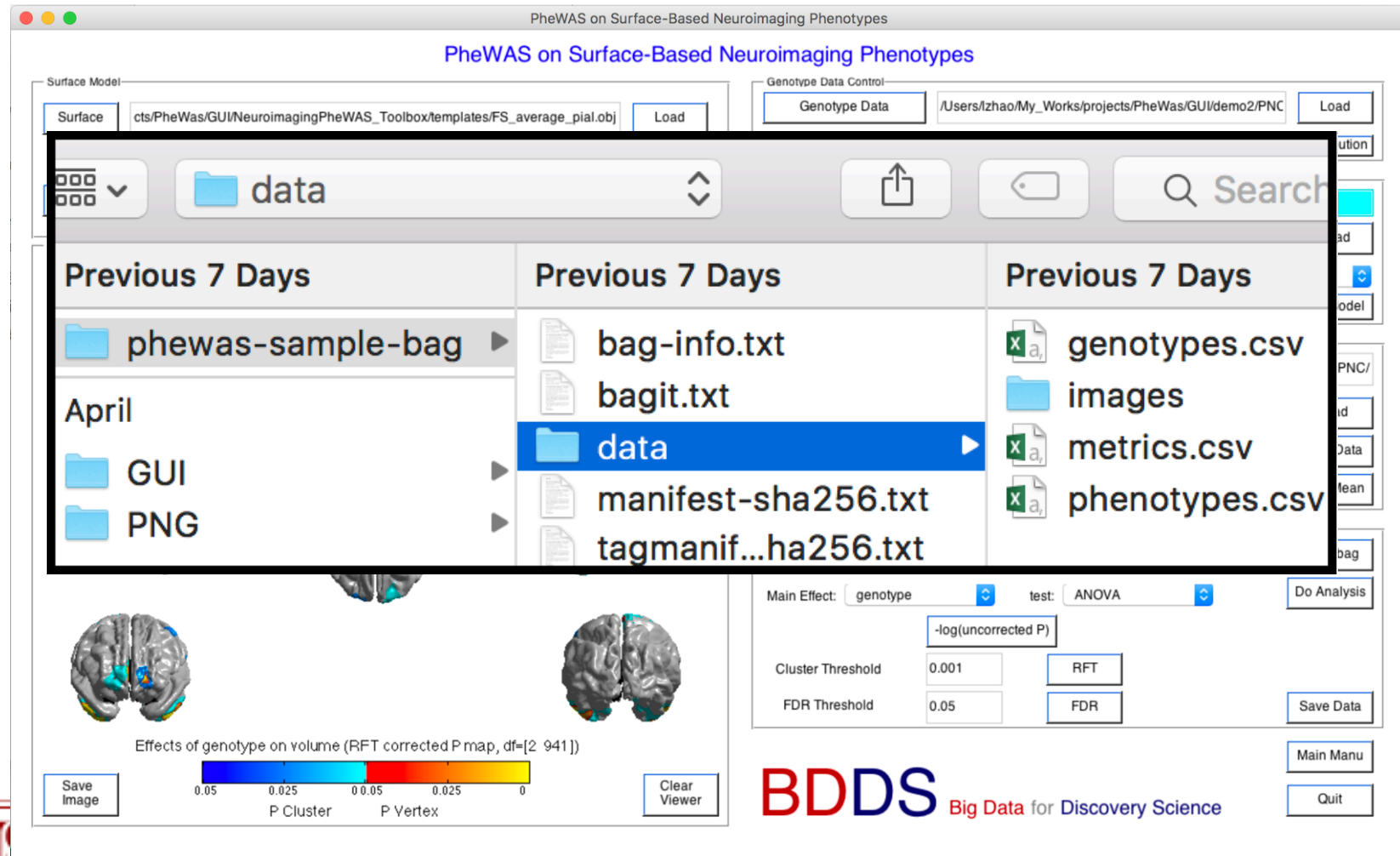
+ SUBJECT IMAGING METADATA (8)

- SUBJECT PARCELLATION METRICS (381)

VIEW Default | Transpose

File Id	Atlas	Label Name	Hemi	Tissue	Structure	Division	Primary Lobe	Secondary Lobe	Sup Inf	Med Lat	Ant Post	Num Vert	Surf Area	Th Av
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_paracentral	L	gray	cortex	telencephalon	frontal	parietal	none	med	none	1350	903	2.37
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_subcentral	L	gray	cortex	telencephalon	frontal	parietal	none	lat	none	1597	1015	2.80
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_and_S_transv_frontopol	L	gray	cortex	telencephalon	frontal	none	none	none	ant	913	625	2.63
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cingul-Post-dorsal	L	gray	cortex	telencephalon	limbic	none	none	med	post	516	354	3.21
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cingul-Post-ventral	L	gray	cortex	telencephalon	limbic	none	none	med	post	252	159	2.39
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_cuneus	L	gray	cortex	telencephalon	occipital	none	none	med	none	1905	1156	1.94
lh.aparc.a2009s.stats	FS_aparc_2009	ctx_lh_G_front_inf-Opercular	L	gray	cortex	telencephalon	frontal	none	none	lat	none	1280	870	2.93

NeuroimagingPheWAS Toolbox



Summary



- ▶ Exponential increases in computing/storage imposes additional complexity on the end user.... What to do?
- ▶ Scientific Digital Asset Management is the missing link
 - Make science data as good as consumer data
- ▶ We have demonstrated that generally applicable software ecosystem for DAM is feasible