Northwestern Medicine®

SchizConnect and DataBridge for Big Data Neuroscience

Lei Wang

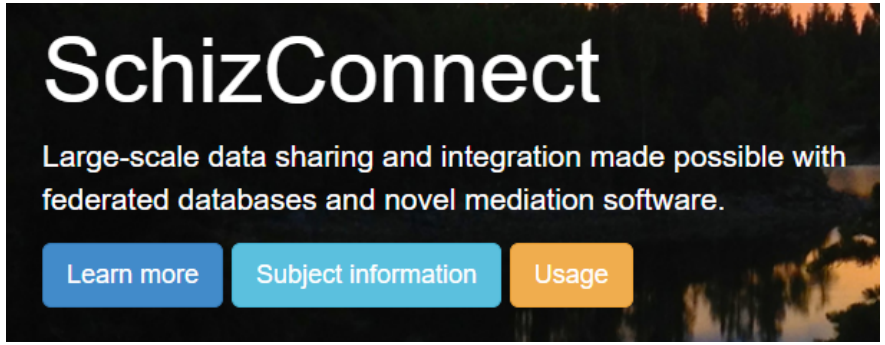
Big Data Neuroscience Workshop 2018
September 6, 2018
Case Western Reserve University

Big data, dark data, the long tail of data

- Dark Data is a problem
  - Legacy data and data in silos
  - Not well published
  - Not well described (non-standard metadata)
- Big Data is a problem
  - Volume - Increasing sizes of studies
  - Volume - Increasing sizes of data (higher resolution)
  - Variety - Differing data formats and instrumentation
  - Variety - Multiple species including simulated brain
- Exploding search space is a problem
  - Many specialized repositories
  - Metadata Standards
  - Publications are increasing

Big data, dark data, the long tail of data

- Dark Data is a problem
  - Legacy data and data in silos
  - Not well published
  - Not well described (non-standard metadata)
- Big Data is a problem
  - Volume - Increasing sizes of studies
  - Volume - Increasing sizes of data (higher resolution)
  - Variety - Differing data formats and instrumentation
  - Variety - Multiple species including simulated brain
- Exploding search space is a problem
  - Many specialized repositories
  - Metadata Standards
  - Publications are increasing

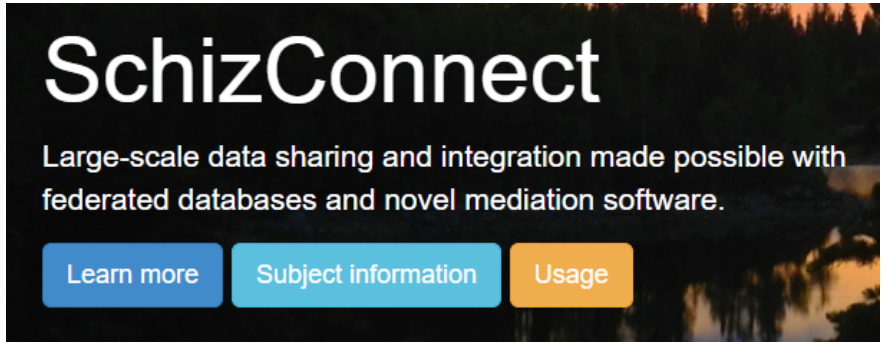


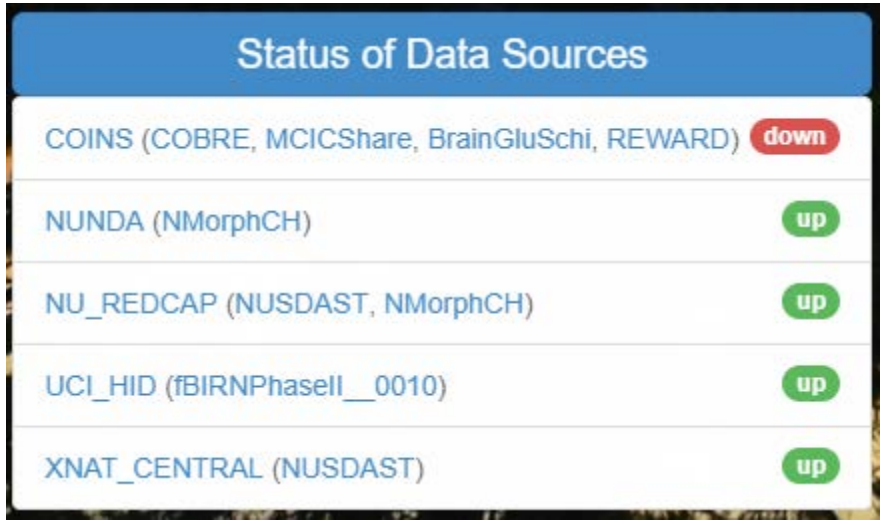
DATABRIDGE

SchizConnect

Data

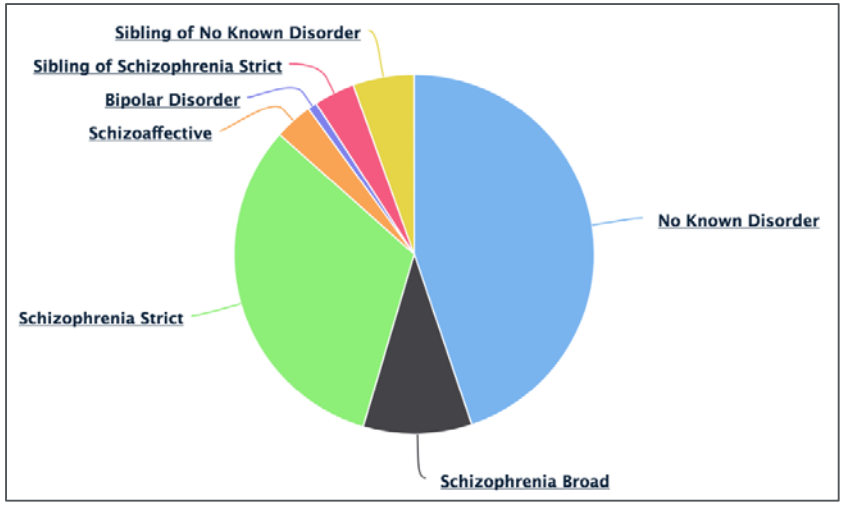
- Virtual database for schizophrenia and related disorders
  - Schizophrenia, schizoaffect, bipolar, siblings, healthy
  - 8 sources, 5 databases
- Neuroimaging data
  - Structural MRI (sMRI)
  - Resting-state functional MRI (fMRI)
  - Task-paradigm fMRI
  - Diffusion MRI (dMRI)
- Non-imaging data
  - Demographics
  - Clinical evaluation
  - Neuropsychological assessments





SchizConnect

Data



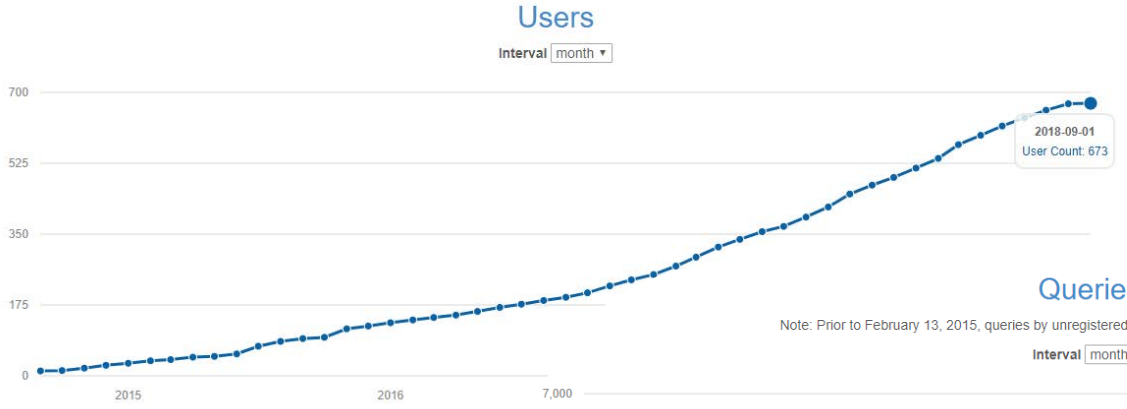
- Virtual database for schizophrenia and related disorders
  - Schizophrenia, schizoaffect, bipolar, siblings, healthy
- 8 sources, 5 databases
- Neuroimaging data
  - Structural MRI (sMRI)
  - Resting-state
  - Task-paradig
  - Diffusion MR
- Non-imaging da
  - Demographi
  - Clinical evalu
  - Neuropsychological assessments
- 1,392 (additional 680 pending, 2,092 total)

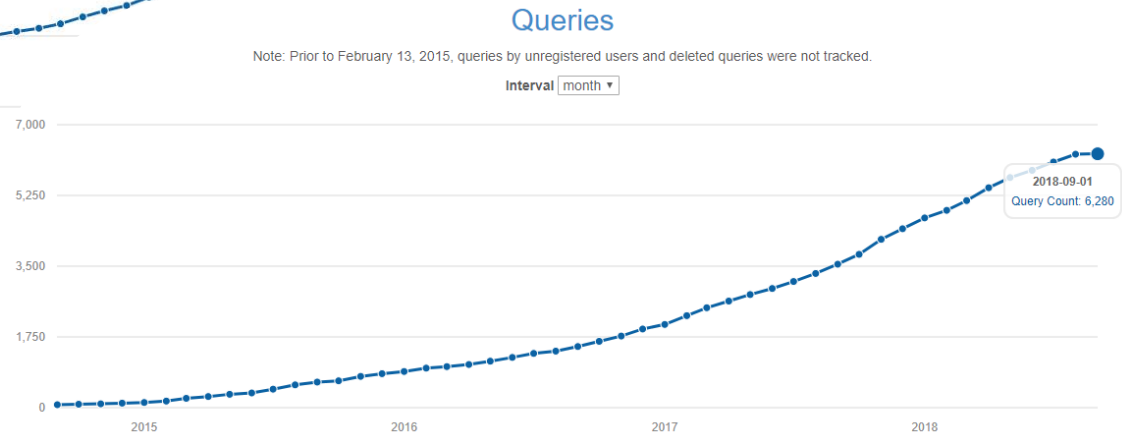
Statistic/DX	No Known Disorder	Schizophrenia Broad	Schizophrenia Strict	Schizoaffective	Bipolar Disorder	Sibling of Schizophrenia Strict	Sibling of No Known Disorder
Number of Current Subjects	632	215	384	41	10	44	66
Gender (m/f)	392/240	173/42	278/106	25/16	5/5	21/23	16/50
Age (years) (mean ± s.d.)	34.7 ± 12.6	34.9 ± 12.9	35.3 ± 12.4	39.2 ± 10.0	46.6 ± 14.4	21.6 ± 3.7	20.4 ± 3.5
Age range (years) (min - max)	13.0 - 67.0	0.0 - 65.0	17.0 - 66.0	19.0 - 59.0	21.0 - 64.0	14.0 - 28.0	14.0 - 28.0

SchizConnect

Usage

- 673 users, 6,280 queries, 1,785 downloads
- Neurodegenerative and Neurodevelopmental Subcortical Shape Diffeomorphometry (NIBIB R01, MPI: Miller, Paulsen, Mostfosky, Wang)
- ReproNim: Center for Reproducible Neuroimaging Computation (CRNC) (NIBIB P41, PI: Kennedy)
- DataBridge for Neuroscience (NSF EAGER, MPI: Arcot Rajasekar, Howard Lander)



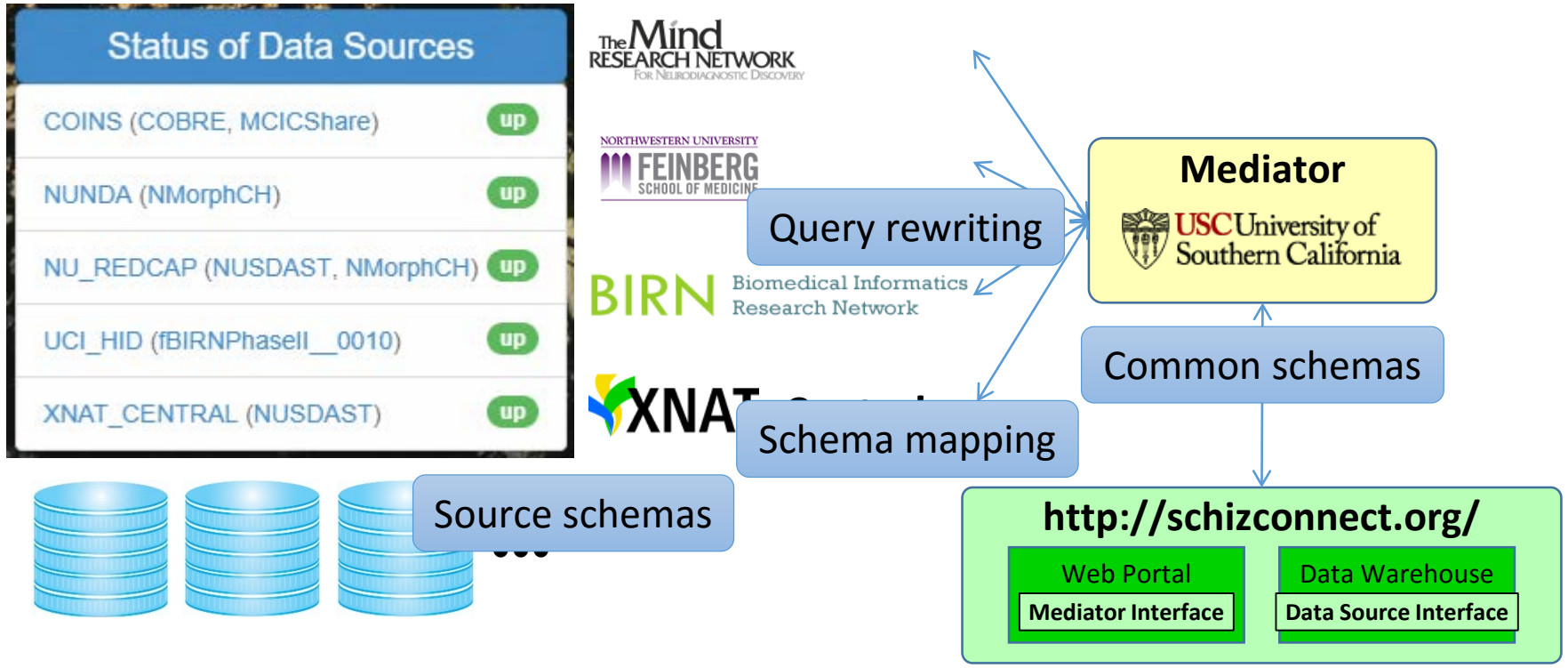


Northwestern Medicine®
Northwestern University
Feinberg School of Medicine

SchizConnect

Architecture

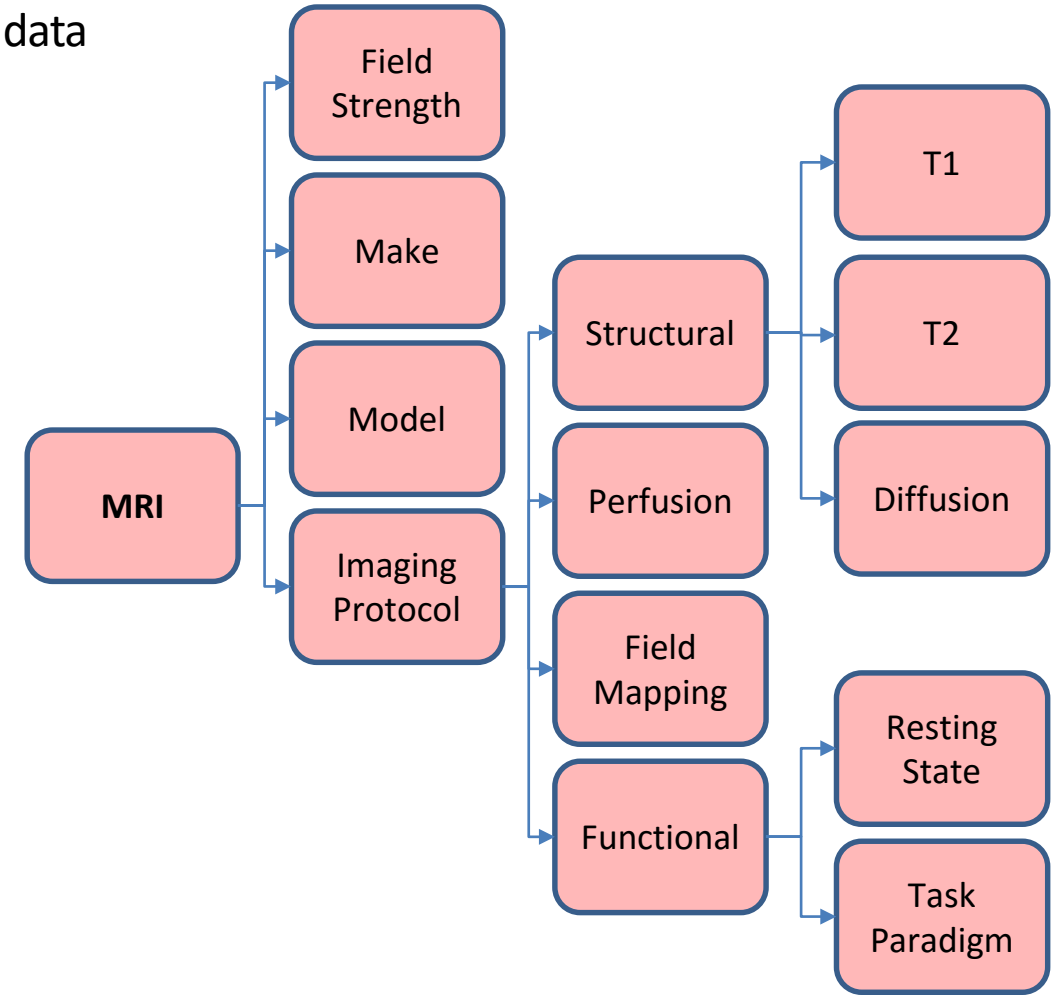
- Data mediation with schema mapping



SchizConnect

Schemas

- Common schema for imaging data



SchizConnect

Schemas

- Common schema for imaging data – Structural

Source		Protocol
HID	T1	t1;"t1"
HID	T1	t1_deface;"t1_deface"
HID	T2	t2;"t2"
HID	T2	T2 Inplane Scan;"T2 Inplane Scan"
NUSDAST	T1	FLASH1 type="T1"
NUSDAST	T1	MPR1 type="T1"
NUSDAST	T1	MPR2 type="T1"
NUSDAST	T1	MPR3 type="T1"
NUSDAST	T1	MPR4 type="T1"
NUSDAST	T1	MPR5 type="T1"
NUSDAST	T1	MPR6 type="T1"
NUSDAST	T1	FLASH3D
NUSDAST	T1	MPRAGE

SchizConnect

Schemas

- Source schema

HIDPSQLResource_nc_experiment(uniqueid:NUMBER:f, tableid:NUMBER:f, owner:NUMBER:f, modtime:DATE:f, moduser:NUMBER:f, name:STRING:f, description:STRING:f, contactperson:STRING:f, ….)

HIDPSQLResource_nc_subjexperiment(uniqueid:NUMBER:f, … subjectid:STRING:f, ….)

HIDPSQLResource_nc_expsegment(segmentid, componentid, …, subjectid, … time_stamp:DATE:f, description:STRING:f, protocolversion, protocolid:STRING:f, …)

HIDPSQLResource_nc_collectionequipment(uniqueid, tableid, owner, modtime:DATE:f, …, make, model)

…

XnatSubjectResource_xnat__subjectData(PROJECT, SUBJECT_ID)

XnatMRSessionResource_xnat__mrSessionData(PROJECT, SCANNER, MARKER)

…

COINSMySQLResource_subjects_v(ANONYMIZATION_ID, GENDER, YOB, SUBJECT_TYPE, STUDY_ID, AGE)

COINSMySQLResource_series_v(SERIES_ID, ANONYMIZATION_ID, AGE_AT_SCAN, SCAN_DATE_YEAR, SCAN_DATE, COLLECTION_TECHNIQUE, DEFINITION, SCANNER_LABEL, SCANNER_MODEL, …)

…

MappingsMySQLResource_protocol_mappings(szc_protocol_hier, source, protocolid)

MappingsMySQLResource_scanner_mappings(maker, model, field_strength, source, source_make, source_model)

…

SchizConnect

Schemas

- Schema mapping

```
imaging_protocol("HID", SUBJECTID, SZC_PROTOCOL_HIER, DATE, NOTES, DATAURI,
                 MAKER, MODEL, FIELD_STRENGTH) <-
  HIDPSQLResource_nc_scannersbyscan_mview( SUBJECTID, componentid, segmentid,
      SOURCE_PROTOCOL, DATE, nc_colequipment_uniqueid, SOURCE_MAKE, SOURCE_MODEL,
      DATAURI, NOTES) ^
  MappingsMySQLResource_protocol_mappings( SZC_PROTOCOL_HIER, "HID",
                                           SOURCE_PROTOCOL, ID1) ^
  MappingsMySQLResource_scanner_mappings( MAKER, MODEL, FIELD_STRENGTH, "HID",
                                          SOURCE_MAKE, SOURCE_MODEL, ID2)

imaging_protocol("NUSDAST", SUBJECTID, SZC_PROTOCOL_HIER, DATE, SCAN_ID, DATA_URI,
                 "SIEMENS", "VISION 1.5T", 1.5) <-
  XnatMRSessionResource_xnat__mrSessionData( SUBJECTID, IMAGE_ID, SESSION_ID, DATE,
                                             SCANNER, SCAN_ID, SCAN_TYPE, quarantine_status) ^
  MappingsMySQLResource_protocol_mappings( SZC_PROTOCOL_HIER, "NUSDAST", SCAN_TYPE, ID1) ^
  Concat(IMAGE_ID, "/scans/", SCAN_ID, DATA_URI)

...
```

SchizConnect

Schemas

- Query rewriting

Using the schema mappings the SchizConnect mediator rewrites queries over the domain schema to queries over the source schemas

- Domain Query (over domain schema):

  SELECT * FROM imaging_protocol WHERE szc_protocol_hier like '%Auditory Oddball%'

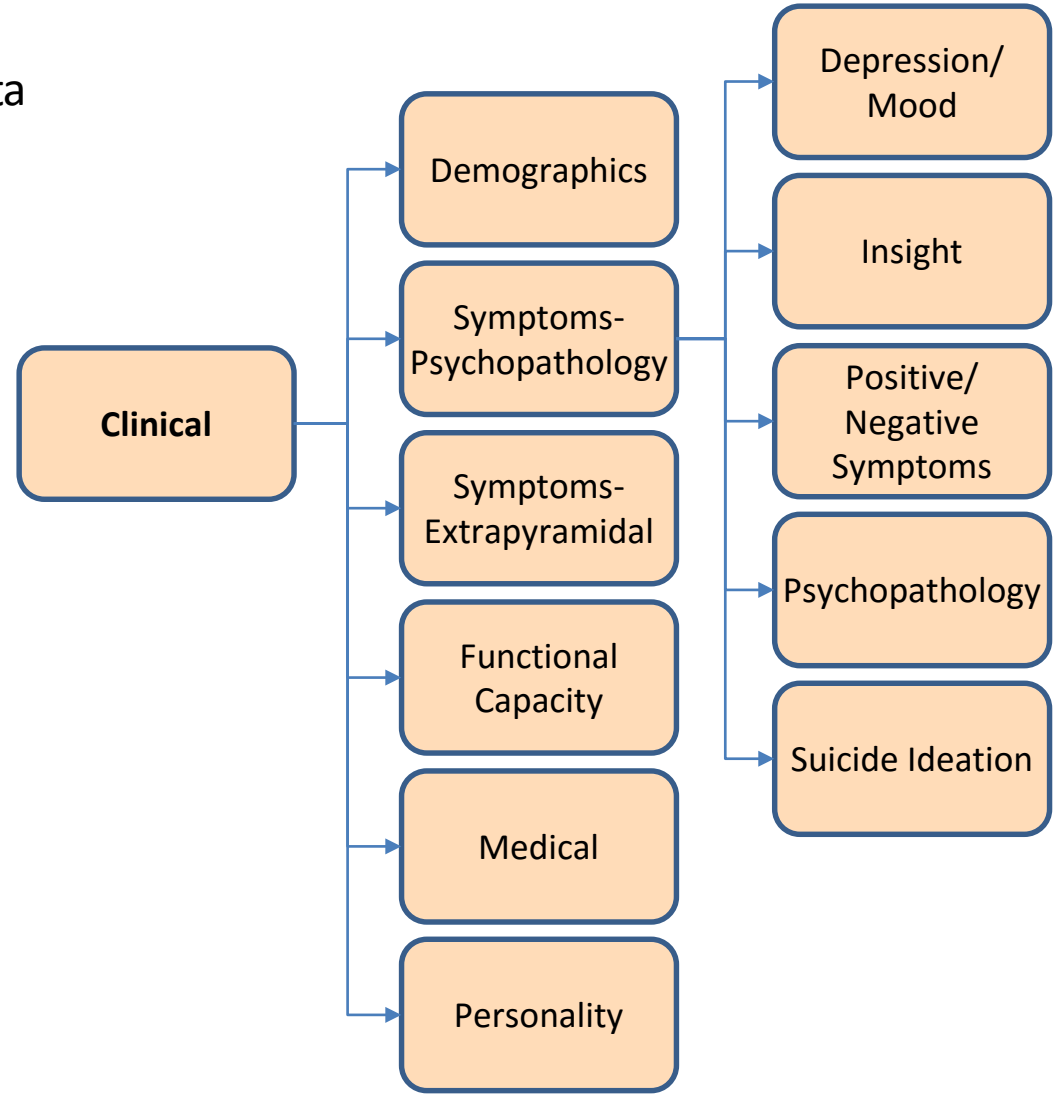
- Rewritten Query (over source schemas):

```
(SELECT  'HID' as provenance, T6.subjectid as subjectid, T4.szc_protocol_hier as szc_protocol_hier, T6.date as img_date,
          T6.description as notes, T6.datauri as datauri, T2.maker as maker, T2.model as model,  T2.field_strength as field_strength
 FROM MappingsMySQLResource_scanner_mappings T2, MappingsMySQLResource_protocol_mappings T4,
        HIDPSQLResource_nc_scannersbyscan_mview T6
 WHERE T2.source_make=T6.source_make  and T2.source_model=T6.source_model and T2.source = 'HID' and
        T4.source_protocol=T6.source_protocol and T4.source = 'HID' and T4.szc_protocol_hier LIKE '%Auditory Oddball%')
UNION
(SELECT 'NUSDAST' as provenance, T10.SUBJECT_ID as subjectid, T8.szc_protocol_hier as szc_protocol_hier, T10.SCAN_DATE as img_date,
          T10.SCAN_ID as notes, Concat(T10.IMAGE_ID,'/scans/',T10.SCAN_ID) as datauri, 'SIEMENS' as maker, 'VISION 1.5T' as model,
          1.5 as field_strength
 FROM MappingsMySQLResource_protocol_mappings T8, XnatMRSessionResource_xnat__mrSessionData T10
 WHERE T8.source_protocol=T10.SCAN_TYPE  and T8.source = 'NUSDAST' and T8.szc_protocol_hier LIKE '%Auditory Oddball%')
UNION
(SELECT 'COINS' as provenance, T12.ANONYMIZATION_ID as subjectid, T12.szc_protocol_hier as szc_protocol_hier,
          T12.SCAN_DATE as img_date, 'notes' as notes, T12.SERIES_ID as datauri, T12.SCANNER_MANUFACTURER as maker,
          T12.SCANNER_LABEL as model, T12.FIELD_STRENGTH as field_strength
FROM COINSMySQLResource_series_v T12
WHERE T12.szc_protocol_hier LIKE '%Auditory Oddball%')
```

SchizConnect

Schemas

- Schema for clinical data



SchizConnect

Schemas

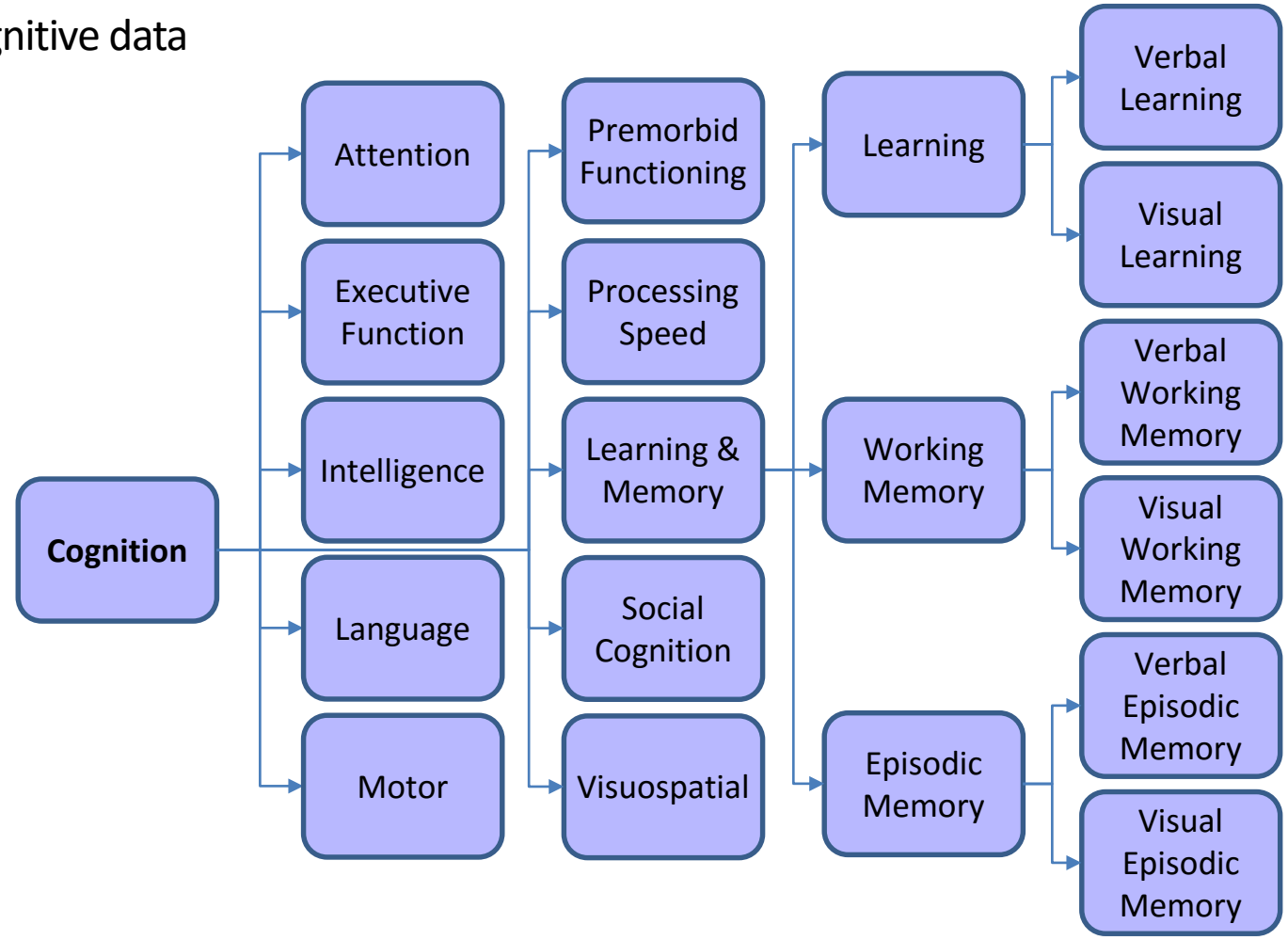
- Schema for clinical data – Psychopathology

Source	Test	
NUSDAST	SIPS	Structured Interview for Prodromal Syndromes Summary
NMorphCH	SIPS	Structured Interview for Prodromal Syndromes Summary
ConteTT	SIPS	Structured Interview for Prodromal Syndromes Summary
COBRE	PANSS	Positive and Negative Symptom Scale
BrainGluSchi	PANSS	Positive and Negative Symptom Scale
fBIRN PhaseII	Modified Positive and Negative Symptom Scale	Positive and Negative Symptom Scale
fBIRN PhaseIII	PANSS	Positive and Negative Symptom Scale
MCIC	SAPS	Scale for the Assessment of Positive Symptoms
fBIRN PhaseII	SAPS	Scale for the Assessment of Positive Symptoms
fBIRN PhaseIII	SAPS_PhaseIII	Scale for the Assessment of Positive Symptoms
NUSDAST	SAPS SANS	Scale for the Assessment of Positive Symptoms
NMorphCH	SAPS SANS	Scale for the Assessment of Positive Symptoms
ConteTT	SAPS SANS	Scale for the Assessment of Positive Symptoms
MCIC	SANS	Scale for the Assessment of Negative Symptoms
fBIRN PhaseII	SANS	Scale for the Assessment of Negative Symptoms
fBIRfBIRN PhaseIII	SANS_PhaseIII	Scale for the Assessment of Negative Symptoms
NUSDAST	SAPS SANS	Scale for the Assessment of Negative Symptoms
NMorphCH	SAPS SANS	Scale for the Assessment of Negative Symptoms
ConteTT	SAPS SANS	Scale for the Assessment of Negative Symptoms
fBIRN PhaseIII	NSA-4	Negative Symptom Assessment
fBIRN PhaseII	Deficit Syndrom ScoreSheet	Deficit Syndrome Score Sheet
fBIRN PhaseIII	Schedule of Deficit Syndrome Scale	Schedule of Deficit Syndrome Scale
fBIRN PhaseII	Hallucination	Hallucination
fBIRN PhaseII	Calgary Depression Scale	Calgary Depression Scale

SchizConnect

Schemas

- Schema for cognitive data



SchizConnect

Schemas

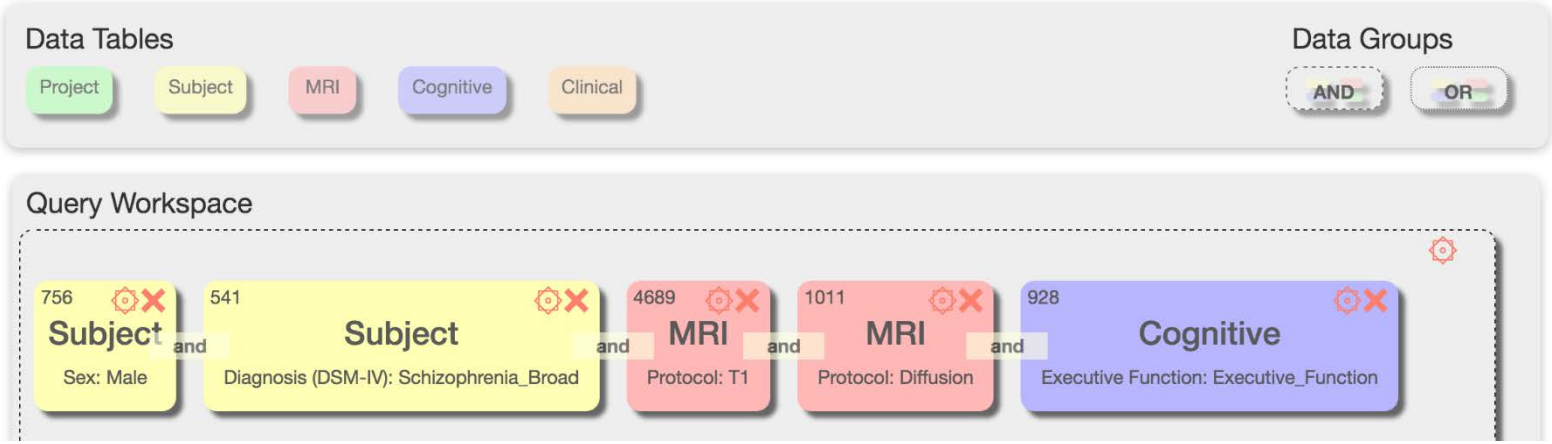
- Schema for cognitive data – Attention

Source	Test	
MCICShare	CalCap	California Computerized Assessment Package
NUSDAST	CPT-AX	A-X Continuous Performance Test - context version
NMorphCH	CPT-AX	A-X Continuous Performance Test - context version
ConteTT	CPT-AX	A-X Continuous Performance Test - context version
COBRE	CPT-II	Conners' Continuous Performance Test-II
COBRE	CPT-IP	Continuous Performance Test - Identical Pairs version
fBIRN PhaseIII	CPT CMINDS	Continuous Performance Test - Identical Pairs version
NMorphCH	CPT-IP	Continuous Performance Test - Identical Pairs version
BrainGluSchi	CPT-IP	Continuous Performance Test - Identical Pairs version
ConteTT	CPT-IP	Continuous Performance Test - Identical Pairs version
BrainGluSchi	MATRICS Attention_Vigilance	MATRICS Consensus Cognitive Battery (MCCB) Attention Vigilance
COBRE	MATRICS Attention_Vigilance	MATRICS Consensus Cognitive Battery (MCCB) Attention Vigilance
fBIRN PhaseIII	Stroop Test CMINDS	Stroop Test

SchizConnect

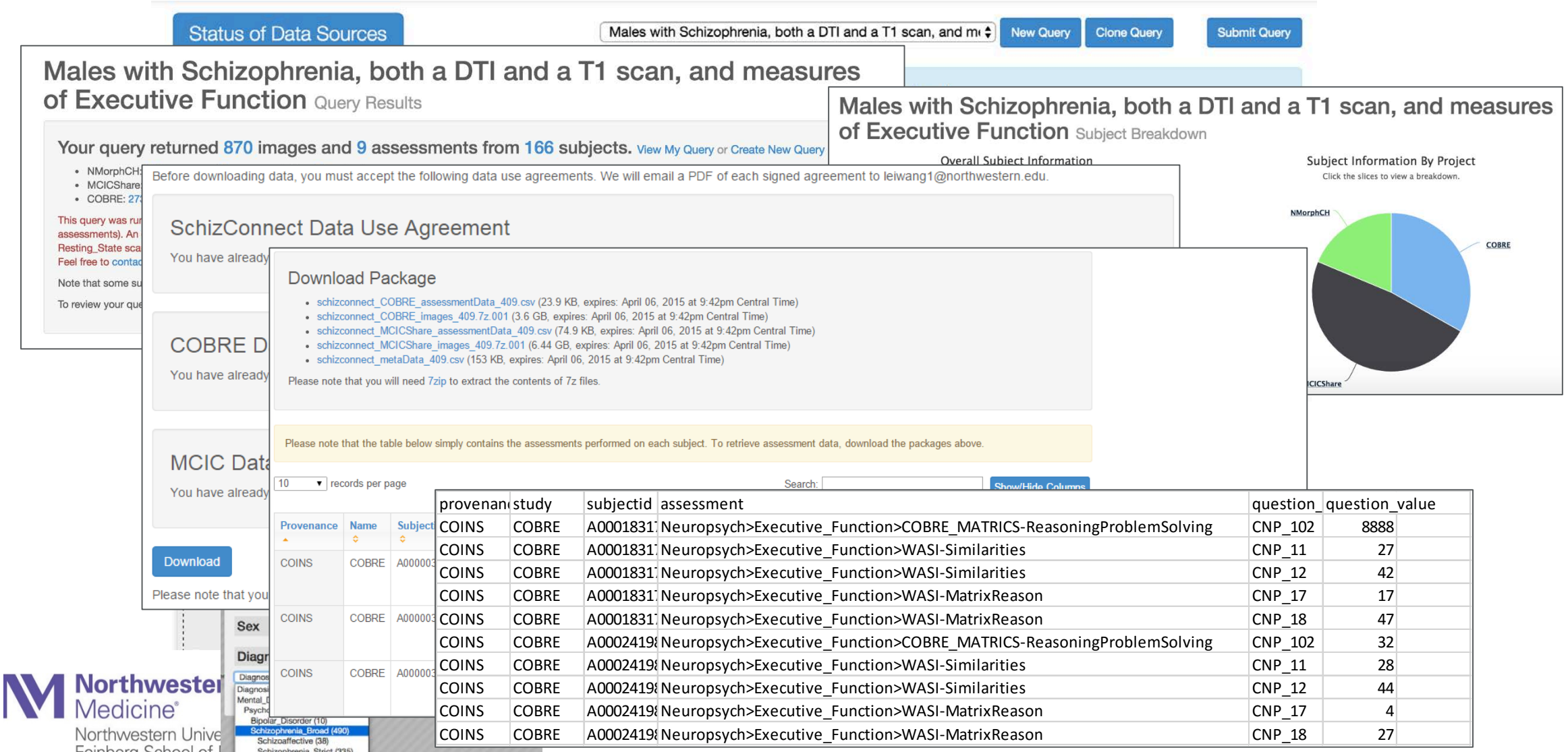
Web portal query

- Males with Schizophrenia, both a DTI and a T1 scan, and measures of Executive Function



SchizConnect

- Males with Schizophrenia, both a DTI and a T1 scan, and measures of Executive Function



provenan	study	subjectid	assessment	question_	question_value	
COINS	COBRE	A0001831	Neuropsych>Executive_Function>COBRE_MATRICS-ReasoningProblemSolving	CNP_102	8888	
COINS	COBRE	A0001831	Neuropsych>Executive_Function>WASI-Similarities	CNP_11	27	
COINS	COBRE	A0001831	Neuropsych>Executive_Function>WASI-Similarities	CNP_12	42	
COINS	COBRE	A0001831	Neuropsych>Executive_Function>WASI-MatrixReason	CNP_17	17	
COINS	COBRE	A0001831	Neuropsych>Executive_Function>WASI-MatrixReason	CNP_18	47	
COINS	COBRE	A0002419	Neuropsych>Executive_Function>COBRE_MATRICS-ReasoningProblemSolving	CNP_102	32	
COINS	COBRE	A0002419	Neuropsych>Executive_Function>WASI-Similarities	CNP_11	28	
COINS	COBRE	A0002419	Neuropsych>Executive_Function>WASI-Similarities	CNP_12	44	
COINS	COBRE	A0002419	Neuropsych>Executive_Function>WASI-MatrixReason	CNP_17	4	
COINS	COBRE	A0002419	Neuropsych>Executive_Function>WASI-MatrixReason	CNP_18	27	

Northwestern Medicine®
Northwestern Univ
Feinberg School of

SchizConnect

Download in BIDS format

- Before BIDS …

  COBRE/human/dicom/triotim/PI/cobre_ID/SUBID/SESID/TYPE/*.dcm

  fBIRNPhaseII__0010/Data/SUBID/VISITID/EXAMTYPE/TYPE/Native/Original/NIFTI/*.img

  MCICShare/SITEID/dicom/triotim/PI/mcicshare_ID/SUBID/SESID/TYPE/*.dcm

  NMorphCH/NUNDA_ID/SESLABEL/scans/SCANID_TYPE/resources/DICOM/files/*.dcm

  NUSDAST/CENTRAL_ID/SESLABEL/scans/SCANID/resouces/ANALYZE/files/*.img

- Different file structure for each source
- Required review of source-specific specification to understand
- Onus on data source manager to keep documentation current
- Pain for processing data

SchizConnect

Download in BIDS format

- With BIDS
  - Standardized file struc

PROJECT/

sub-SUBJID/

ses-SESDATE

D

```
|-- MCICShare
|   |-- dataset_description.json
|   |-- participants.tsv
|   |-- sub-A00036216
|   |   `-- ses-20040101
|   |       |-- anat
|   |       |   |-- sub-A00036216_ses-20040101_acq-mprage_run-01_T1w.json
|   |       |   |-- sub-A00036216_ses-20040101_acq-mprage_run-01_T1w.nii.gz
|   |       |   |-- etc.
|   |       |-- func
|   |       |   |-- sub-A00036216_ses-20040101_task-sternbergitemrecognition_run-01_bold.json
|   |       |   |-- sub-A00036216_ses-20040101_task-sternbergitemrecognition_run-01_bold.nii.gz
|   |       |   |-- etc.
|   |       |-- dwi
|   |       |   |-- sub-A00036216_ses-20040101_run-01_dwi.json
|   |       |   |-- sub-A00036216_ses-20040101_run-01_dwi.nii.gz
|   |       |   |-- etc.
|   |       `-- sub-A00036216_ses-20040101_scans.tsv
|   `-- etc.
|-- NMorphCH
|   |-- dataset_description.json
|   |-- participants.tsv
|   |-- sub-CH7131b
|   |   `-- ses-20110930
|   |       |-- anat
|   |       |   |-- sub-CH7131b_ses-20110930_acq-mprage_run-01_T1w.json
|   |       |   |-- sub-CH7131b_ses-20110930_acq-mprage_run-01_T1w.nii.gz
|   |       |   |-- etc.
|   |       |-- func
|   |       |   |-- sub-CH7131b_ses-20110930_task-2back_bold.json
|   |       |   |-- sub-CH7131b_ses-20110930_task-2back_bold.nii.gz
|   |       |   |-- sub-CH7131b_ses-20110930_task-2back_bold_events.tsv
|   |       |-- dwi
|   |       |   |-- sub-CH7131b_ses-20110930_run-01_dwi.json
|   |       |   |-- sub-CH7131b_ses-20110930_run-01_dwi.nii.gz
|   |       |   |-- etc.
|   |       `-- sub-CH7131b_ses-20110930_scans.tsv
|   `-- etc.
```

SchizConnect

Download in BIDS format

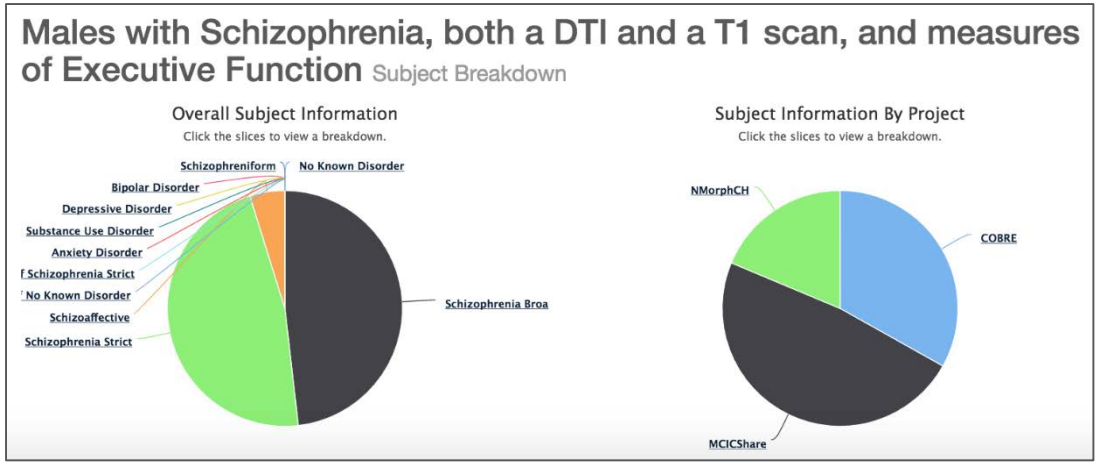
- BIDS apps

Table 1. List of currently available BIDS Apps.

App name	Description
example	Example App that also serves as a template for new apps. Calculates intracranial volume.
Freesurfer	Surface extraction, longitudinal pipeline and study specific template calculation using FreeSurfer.
ndmg	One-click reliable and reproducible pipeline for T1w + DWI weighted MRI connectome estimation.
BROCCOLI	Fast fMRI analysis on many-core CPUs and GPUs.
FibreDensityAndCrosssection	Fixel-Based Analysis (FBA) of Fibre Density and Fibre Cross
SPM	Statistical Parametric Mapping.
MRIQC	Quality Assessment of structural and functional MRI.
FMRIPREP	A generic fMRI preprocessing pipeline providing results robu data quality as well as informative reports.
Quality Assessment Protocol	Quality Assessment of structural and functional MRI.
Configurable Pipeline for the Analysis of Connectomes	Pipeline for high throughput processing and analysis of struct functional MRI data.
Hyperalignment	Computes hyperalignment transformations for functional alig
mindboggle	Pipeline to improve the accuracy, precision, and consistency of automated labeling and shape analysis of human brain image data.
MRtrix3 connectome	Robust generation and statistical analysis of structural connectomes estimated from diffusion tractography.
nilearn	Extraction of time-series and connectomes for population analysis.
nipypelines	Preprocessing of functional time series for resting or task analysis
automatic analysis (aa)	Neuroimaging pipeline system written in Matlab.
Niak Preprocessing	Noise reduction, segmentation, coregistration, motion estimation, resampling.
HPCPipelines	Anatomical and functional preprocessing pipelines used in the Human Connectome Project.
BrainIAK-SRM	Functional alignment using Shared Response Model implementation from the Brain Imaging Analysis Kit.
OPPNI	Optimization of Preprocessing Pipelines for NeuroImaging, for analysis of fMRI data.
MAGeTbrain	Multiple Automatically Generated Templates brain segmentation algorithm
tracula	Automatic reconstruction of a set of major white-matter pathways from diffusion-weighted MR images

Big data, dark data, the long tail of data

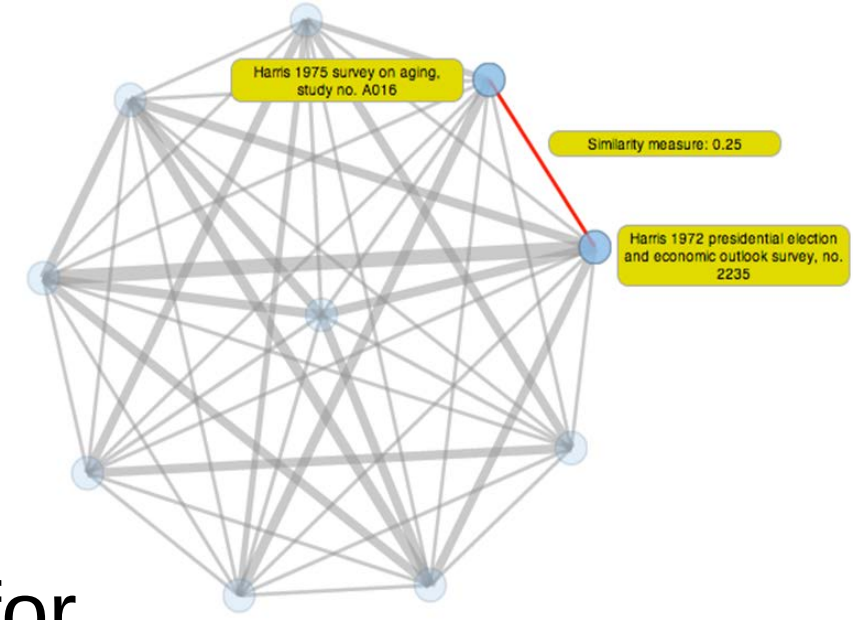
- Are there other datasets out there that are “similar?”
  - NITRC, NDA, LONI, XNAT Central, openfMRI.org, 1000_FNC/INDI?
  - Frontiers, Plos ONE, Nature?
- For meta analysis
  - ENIGMA?
- For reproducibility research
  - OpenNeuro?
  - ReproNim?



- DataBridge
  - Explore novel techniques for data discovery in neuroscience
  - Arcot Rajasekar, Howard Lander @ UNC

The DataBridge Vision

- Assist scientists in discovering “interesting” data sets by automatically forming communities of data
- Domain scientists can create their own algorithms defining “interesting”
- Build an extensible, adaptable platform for building communities of data
- Search for relevant data sets through community defined linkages

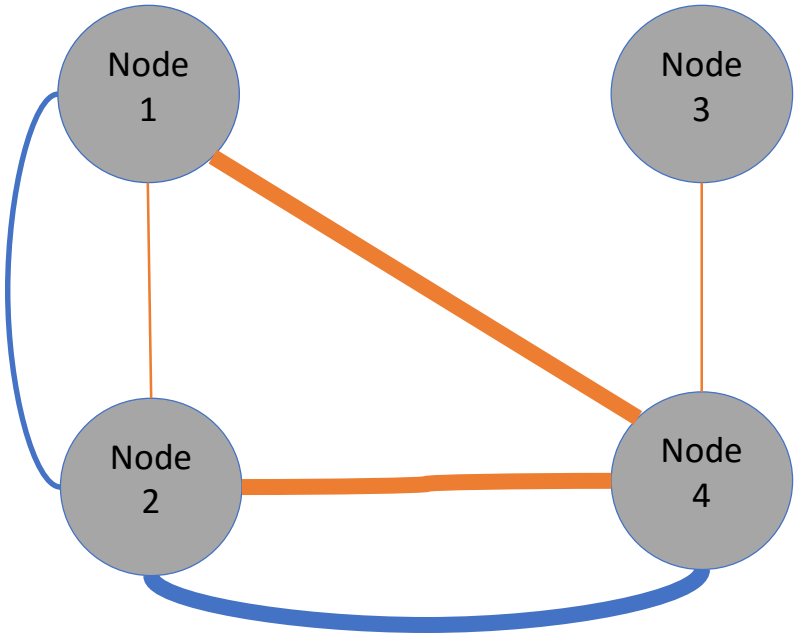


How the DataBridge Works

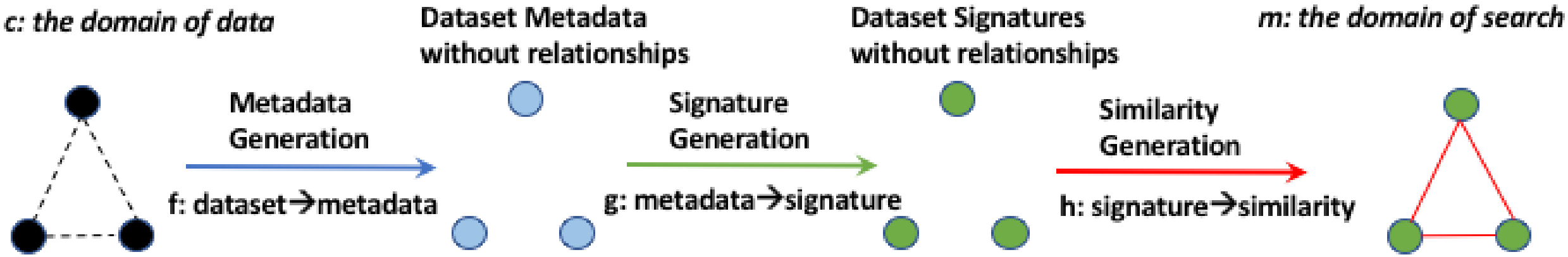
Data communities are searchable multi-dimensional networks. Four steps to building them:

- Extract or create metadata for collections of data sets
- Derive metadata-based signatures for the data sets
- Evaluate the similarity of data sets
- Detect network of communities using the resulting set of similarities

Algorithms for each step are pluggable



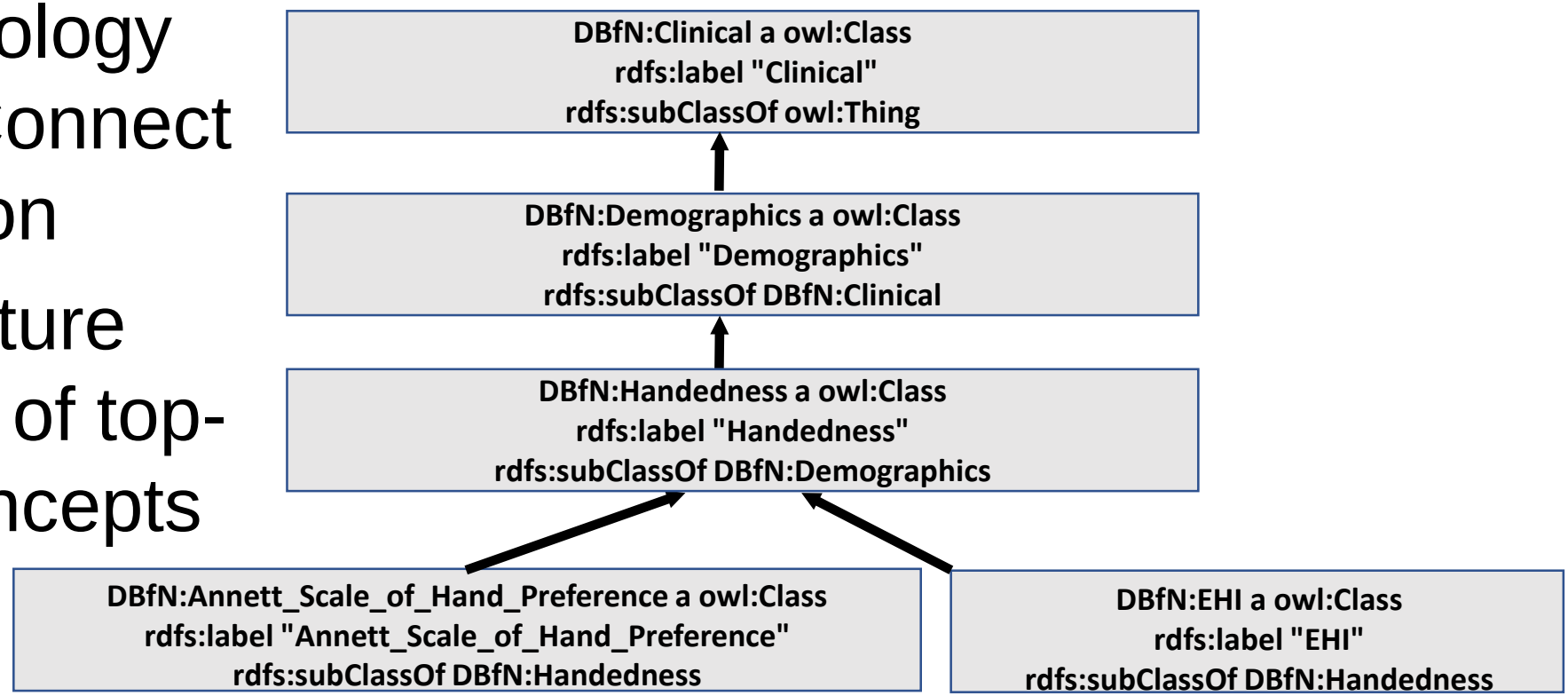
There is a (Category) theory behind this!



- Datasets have some relationship to each other. But these "true" relationships are in principle unknowable
- Through various processes, we derive estimates of these relationships that are searchable
- But these are still estimates! Metric is usefulness, not "truth"

DataBridge Signature Using Ontology Based Harmonization

- Build an RDF ontology version of SchizConnect data harmonization
- DataBridge signature vector composed of top-level ontology concepts
- Map instruments from each study to corresponding element of top-level ontology. Not all studies have a match for each concept



Complete signature vectors for the SchizConnect studies

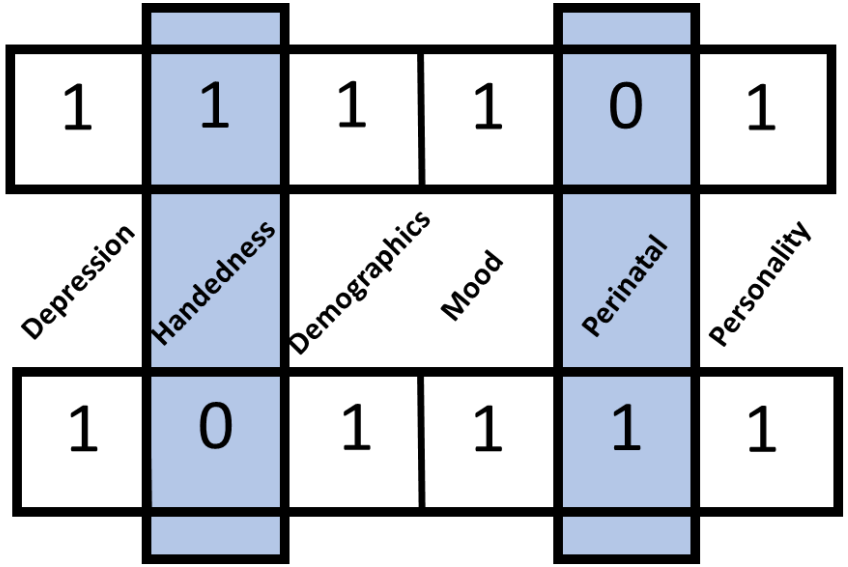
Studies

Concepts

	ConteTT	NMorphCH	NUSDAST	MCIC	BrainGluSchi	COBRE	fBIRNII	fBIRNIII	Total
Demographics	1	1	1	1	1	1	1	1	8
Depression	1	1	1	0	1	1	1	1	7
Extrapyramidal_Symptoms	1	1	1	1	1	1	1	1	8
Functional_Capacity	1	1	1	0	0	1	0	0	4
Genetic_Risk	1	1	1	0	0	0	0	0	3
Hallucination_Type	0	0	0	0	0	0	1	0	1
Insight	0	0	1	0	0	0	0	0	1
Medical_History	0	0	0	0	1	1	0	0	2
Medication_Log	0	0	0	0	1	1	0	1	3
Mental_Health_Diagnosis	1	1	1	0	1	1	1	1	7
Mood	0	0	0	0	0	0	1	0	1
Neuroleptic_Naive	0	0	0	1	0	0	0	0	1
Neurological_Exam	0	0	0	0	0	1	0	1	2
Nicotine_Dependence	0	0	0	0	1	1	1	1	4
Perinatal	1	0	1	0	0	0	0	0	2
Personality	1	0	1	0	0	0	0	0	2
Positive_Negative_Symptoms	1	1	1	1	1	1	1	1	8
Prodromal_Symptoms	1	1	1	0	0	0	0	0	3
Progress_Treatment_Response	0	0	0	0	1	1	0	0	2
Psychopathology	1	1	1	0	1	0	0	0	4
Quality_of_Life	0	0	0	0	1	1	0	0	2
Severity_Manic_Episodes	0	0	0	0	1	1	0	0	2
Social_Responsiveness	1	0	1	0	0	0	0	0	2
Suicidal_Ideation	0	0	0	0	0	0	1	0	1
Suicidal_Ideation	0	0	0	0	0	0	1	0	1
Total	12	9	13	4	12	13	9	8	80

SchizConnect and DataBridge: Similarity Calculation

- Method 1: Hamming Distance



Hamming distance calculation

- Method 2: Cosine Similarity

$$\text{cosine similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Differences in the similarity methods

- Similarity methods produce different results. Collaboration with domain scientists of algorithm selection is critical
- In this case:
  - Hammings Similarity doesn't distinguish between 2 zeros and 2 ones. So missing concepts are as powerful as matching ones
  - Cosine similarity is only based on the matching concepts: If either of the values in the denominator is 0, that concept does not contribute to similarity
- Which is correct? Depends on what you are looking for!
  - Interested in common concepts? Pick Cosine
  - Interested in total semantic distance? Pick Hammings

Hammings and Cosine Results

Hammings Similarity Matrix

	ConteTT	NMorphCH	NUSDAST	MCIC	BrainGluSchi	COBRE	fBIRNPhaseII	fBIRNPhaseIII
ConteTT	1.00	0.87	0.96	0.43	0.50	0.48	0.48	0.51
NMorphCH	0.87	1.00	0.83	0.50	0.58	0.55	0.56	0.59
NUSDAST	0.96	0.83	1.00	0.42	0.48	0.46	0.46	0.49
MCIC	0.43	0.50	0.42	1.00	0.43	0.42	0.50	0.53
BrainGluSchi	0.50	0.58	0.48	0.43	1.00	0.88	0.58	0.71
COBRE	0.48	0.55	0.46	0.42	0.88	1.00	0.55	0.78
fBIRNPhaseII	0.48	0.56	0.46	0.50	0.58	0.55	1.00	0.71
fBIRNPhaseIII	0.51	0.59	0.49	0.53	0.71	0.78	0.71	1.00

Cosine Similarity Matrix

	ConteTT	NMorphCH	NUSDAST	MCIC	BrainGluSchi	COBRE	fBIRNPhaseII	fBIRNPhaseIII
ConteTT	1.00	0.88	0.96	0.58	0.50	0.46	0.54	0.58
NMorphCH	0.88	1.00	0.83	0.71	0.63	0.58	0.67	0.71
NUSDAST	0.96	0.83	1.00	0.54	0.46	0.42	0.50	0.54
MCIC	0.58	0.71	0.54	1.00	0.58	0.54	0.71	0.75
BrainGluSchi	0.50	0.63	0.46	0.58	1.00	0.88	0.63	0.75
COBRE	0.46	0.58	0.42	0.54	0.88	1.00	0.58	0.79
fBIRNPhaseII	0.54	0.67	0.50	0.71	0.63	0.58	1.00	0.79
fBIRNPhaseIII	0.58	0.71	0.54	0.75	0.75	0.79	0.79	1.00

Qualitative Analysis

- Required by lack of ground truth: possible because of small corpus of datasets in the study
- We found results encouraging:
  - ConteTT, NUSDAST and NMorphCH studies have high similarity. Created over a 15-year span at two different institutions, but are from the same research team
  - BrainGluSchi, COBRE, and fBIRNPhaseIII studies have high similarity measures, which is justified by similar study designs
  - FBIRN phase II and phase III were conducted with the same consortium of institutions, but the study design was different across the two phases, so low result is in line with expectations

Big data, dark data, the long tail of data

- The FAIR Guiding Principles, as described in Scientific Data by Wilkinson et al:
- To be **Findable**:
  F1. (meta)data are assigned a globally unique and persistent identifier
  F2. data are described with rich metadata (defined by R1 below)
  F3. metadata clearly and explicitly include the identifier of the data it describes
  F4. (meta)data are registered or indexed in a searchable resource
- To be **Accessible**:
  A1. (meta)data are retrievable by their identifier using a standardized communications protocol
    A1.1 the protocol is open, free, and universally implementable
    A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  A2. metadata are accessible, even when the data are no longer available
- To be **Interoperable**:
  I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (meta)data use vocabularies that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data
- To be **Reusable**:
  R1. meta(data) are richly described with a plurality of accurate and relevant attributes

Big data, dark data, the long tail of data

- The FAIR Guiding Principles, as described in Scientific Data by Wilkinson et al:
- To be **Findable**:
  F1. (meta)data are assigned a globally unique and persistent identifier
  F2. data are described with rich metadata (defined by R1 below)
  F3. metadata clearly and explicitly include the identifier of the data it describes
  F4. (meta)data are registered or indexed in a searchable resource
- To be **Accessible**:
  A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  A2. metadata are accessible, even when the data are no longer available
- To be **Interoperable**:
  I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (meta)data use vocabularies that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data
- To be **Reusable**:
  R1. meta(data) are richly described with a plurality of accurate and relevant attributes

Big data, dark data, the long tail of data

Acknowledgements

- NIH 1 U01 MH097435, 1 R01 EB020062
- NSF SP0037646, BCS 1734853
- SchizConnect
  - Core technical team: Jose Luis Ambite, Jessica Turner, Kathryn Alpert, Joel Matthew, Alex Kogan
  - Data sources: Steven Potkin, Vince Calhoun, Deanna Barch, Juan Bustillo, David Keator, Margaret King, Brittny Miller
- DataBridge for Neuroscience
  - Howard Lander, Arcot Rajasekar
  - Jessica Turner, Matthew Turner

BIRN Biomedical Informatics Research Network

The Mind RESEARCH NETWORK For Neurodiagnostic Discovery

Northwestern Medicine®
Northwestern University
Feinberg School of Medicine

renci RESEARCH ENGAGEMENT INNOVATION

DATABRIDGE