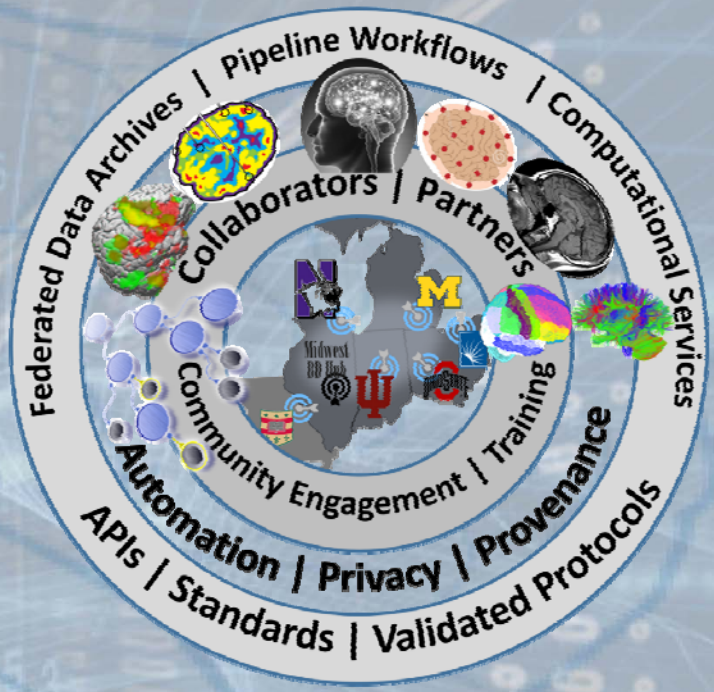
Midwest Workshop on Big Neuroscience Data, Tools, Protocols & Services



Computational Neuroscience Network (ACNN)

http://www.NeuroscienceNetwork.org/ACNN_Workshop_2016.html

Michigan Institute for Data Science

Ivo D Dinov

Statistics Online Computational Resource (SOCR)

Michigan Institute for Data Science (MIDAS)

University of Michigan

http://www.umich.edu/~dinov

SCHOOL OF NURSING
STATISTICS ONLINE
COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

MIDAS

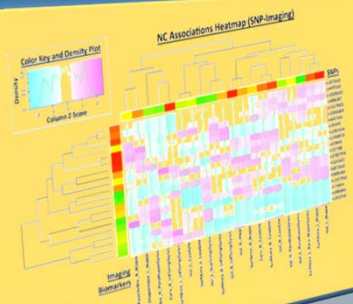
Big Neuroscience Data

Data Wrangling

Data Modeling

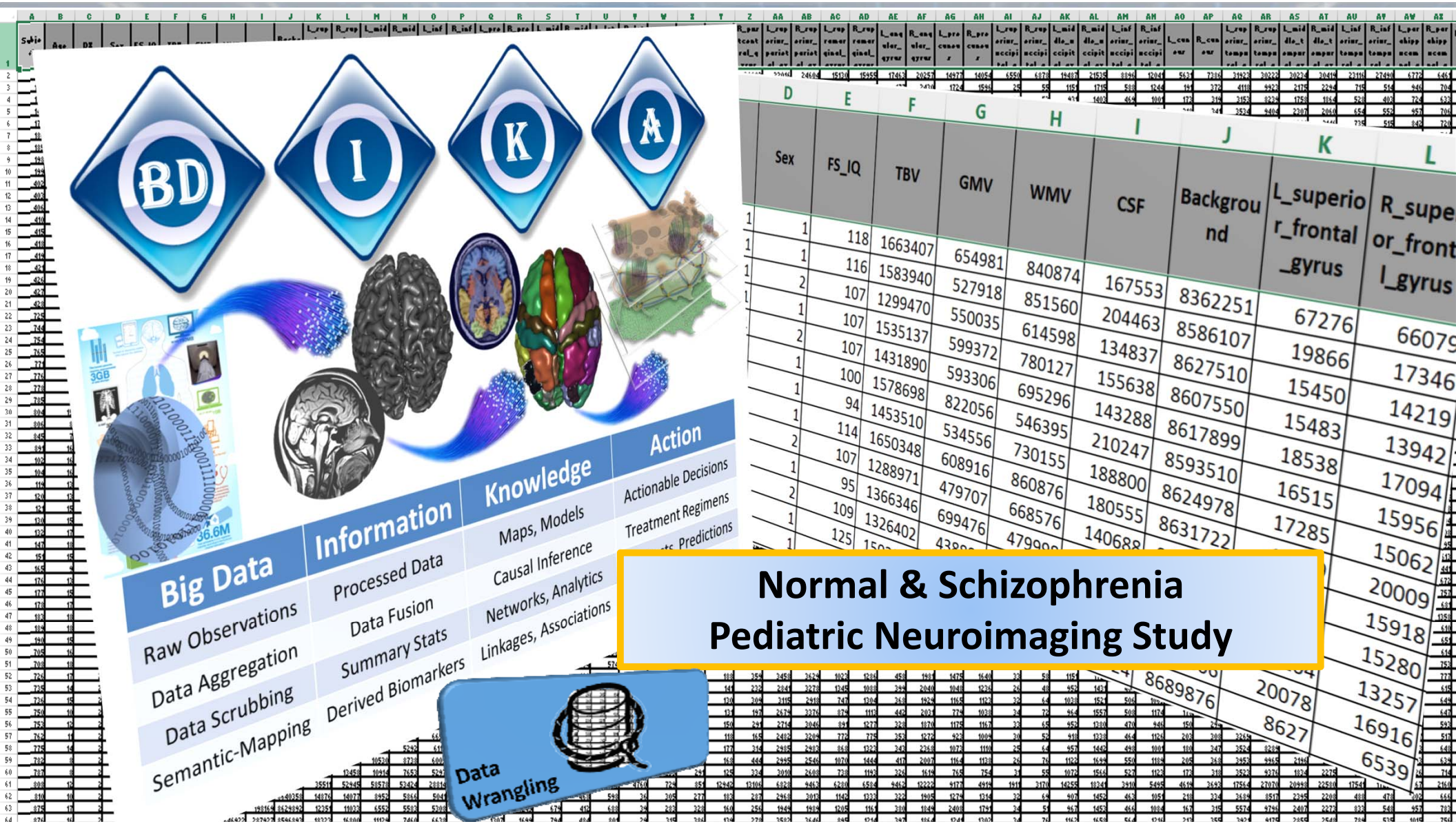
Data Analytics

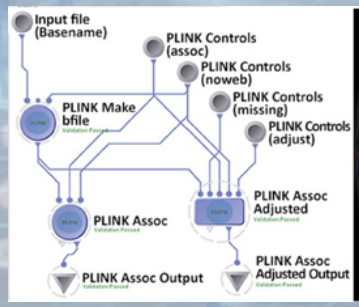
Data Visualization

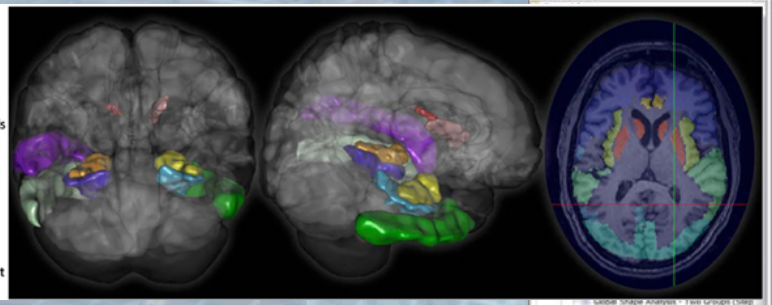


Big Neuroscience Data

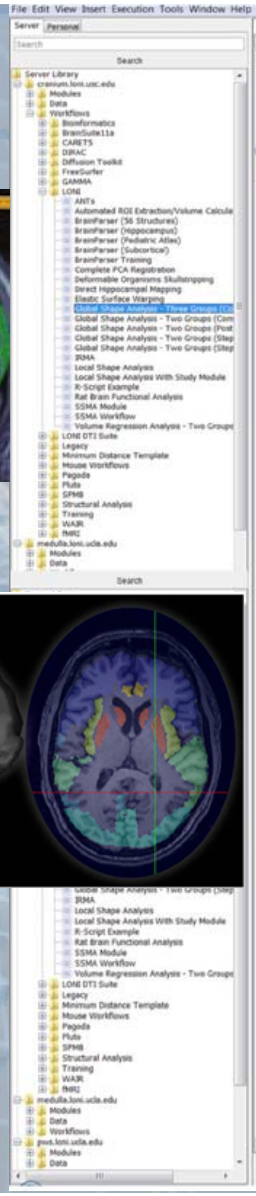
Case-Studies	Sample-size	Data Elements	Description
ALS	Longitudinal data for 8K patients (data comes from 2 independent EU registries)	>300 variables including incomplete clinical, physiological, and cognitive data	This case-study needs innovative machine learning methods for automated diagnostic classification (e.g., patients vs. controls) and unsupervised prediction of cognitive and behavioral decline in ALS patients. The multi-source data will be partitioned into training (estimation) and testing (validation) sets. We plan to fit different models and estimate model-free classifiers to either cluster the participants into groups or hierarchies, or to forecast the progression of the disease over time
Depression / SZ	Extremely large longitudinal data (ms samples), 150 subjects	Study design includes 20 (2*2*5) stimuli types, 12 electrode locations, 1,000's of ms measurements, 4 summary measures	This case-study is focused on identifying a set of P300 biomarkers (using the incomplete high frequency data) that can classify the core cohorts (Bipolar, Schizophrenia, Depression, and healthy controls). We also aim to investigate if P300 responses to emotional stimuli classify the groups better than those to non-emotional stimuli (from standard Go/NoGo tasks). Mean and peak amplitude and latency are candidates based on previously reported results
PD	550 PPMI subjects with 1-10 observation time points	Demographics, clinical tests, vital signs, MDS-UPDRS scores, ADL, MoCA, sMRI, ESS Sleepiness Scale, GDS-15, genetics	Using heterogeneous data of Parkinson's Disease (PD), the study aims to develop a comprehensive end-to-end protocol for data characterization, manipulation, processing, cleaning, analysis and validation. Specifically: (1) introduce methods for rebalancing imbalanced cohorts, (2) utilize a wide spectrum of classification methods to generate consistent and powerful phenotypic predictions, and (iii) generate reproducible machine-learning based classification that enables the reporting of model parameters and diagnostic forecasting based on new data
TBI / Trauma	Over 2,000 patients and controls, acute and multiple chronic times	Dozens of structured (imaging, phenotypic, clinical) and unstructured (injury type, notes) data elements	Three major data sets are included: Volumetric in Brain Trauma (VBT), HeadSmat, and PROTECT II. Each of these datasets include patient's brain CT and/or MRI scans at time of admission, and in some cases during ICU stay, and even during long-term follow up after hospital discharge. These datasets also include time-course data on some physiological measures and blood factors, captured throughout the course of treatment. The UMich/Massey Foundation Grand Challenge provides additional motivation and testing data. The PIs are involved in research funded by this foundation to integrate and analyze these datasets

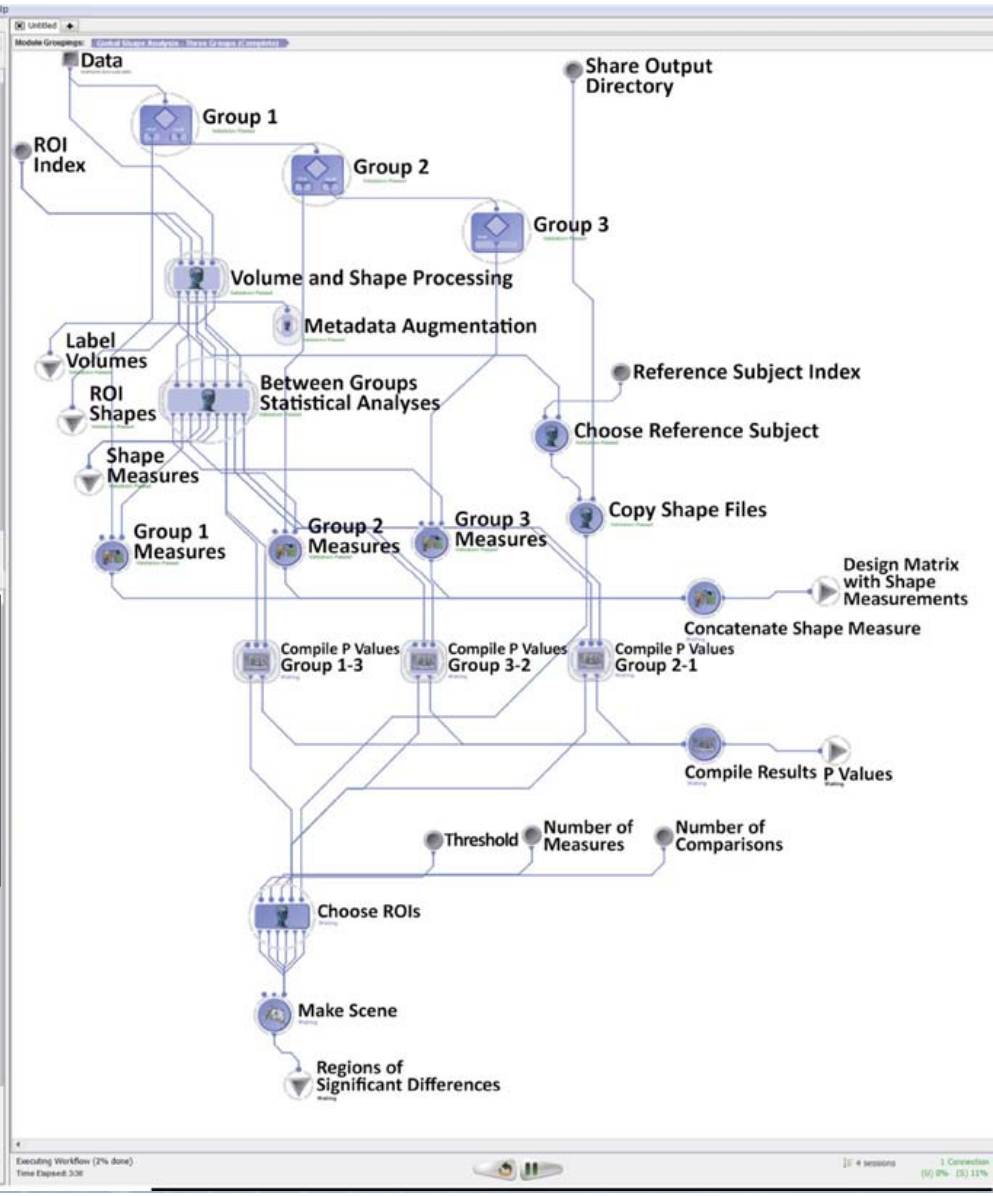

BD
I
K
A
Big Data
Information
Knowledge
Action
Raw Observations
Processed Data
Maps, Models
Actionable Decisions
Data Aggregation
Data Fusion
Causal Inference
Treatment Regimens
Data Scrubbing
Summary Stats
Networks, Analytics
Semantic-Mapping
Derived Biomarkers
Linkages, Associations
3GB
36.6M
Data Wrangling
Sex
FS_IQ
TBV
GMV
WMV
CSF
Background
L_superior_frontal_gyrus
R_superior_frontal_gyrus
Normal & Schizophrenia
Pediatric Neuroimaging Study

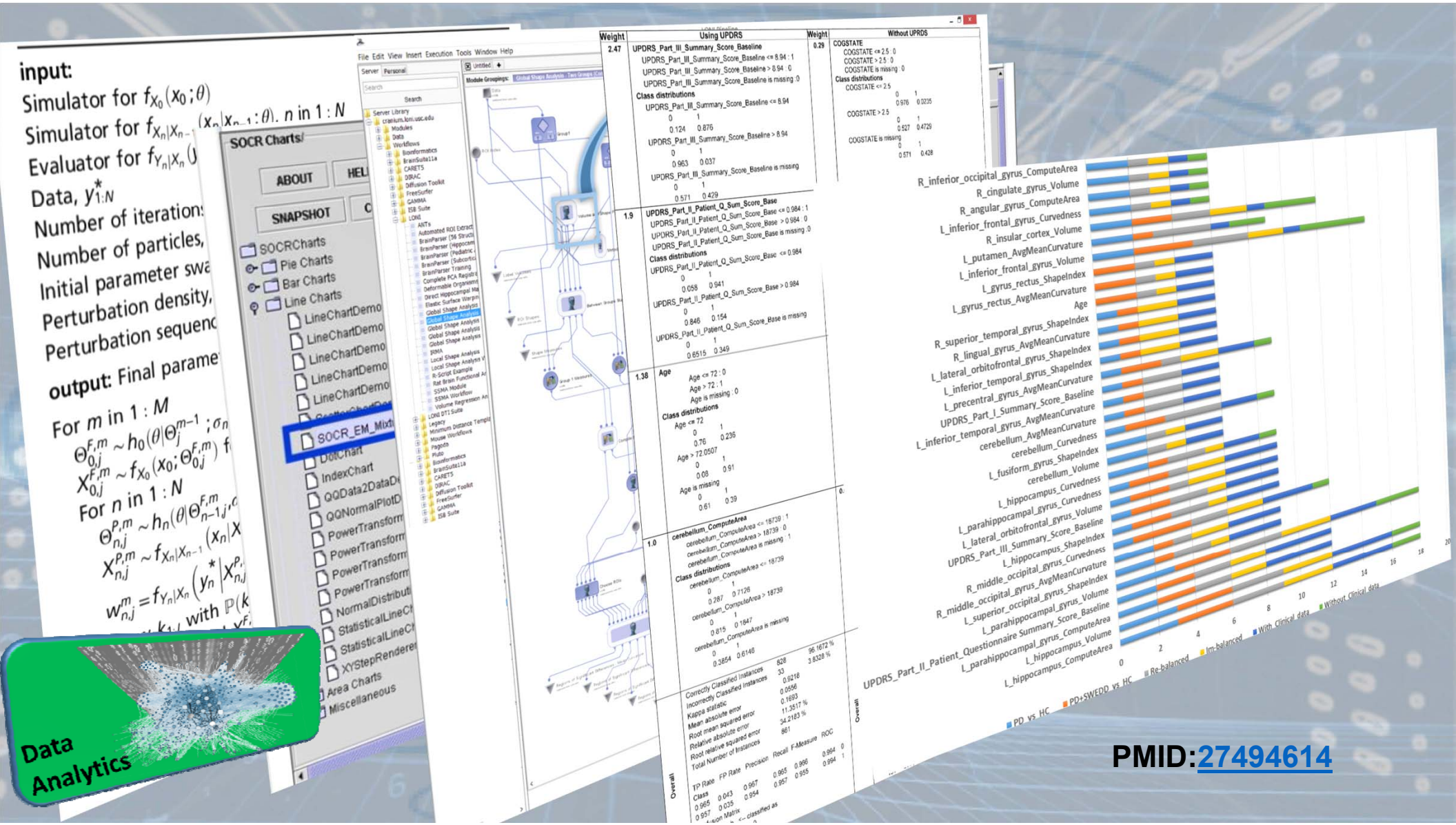

Input file (Basename)
PLINK Controls (assoc)
PLINK Controls (noweb)
PLINK Controls (missing)
PLINK Controls (adjust)
PLINK Make bfile
PLINK Assoc
PLINK Assoc Adjusted
PLINK Assoc Output
PLINK Assoc Adjusted Output



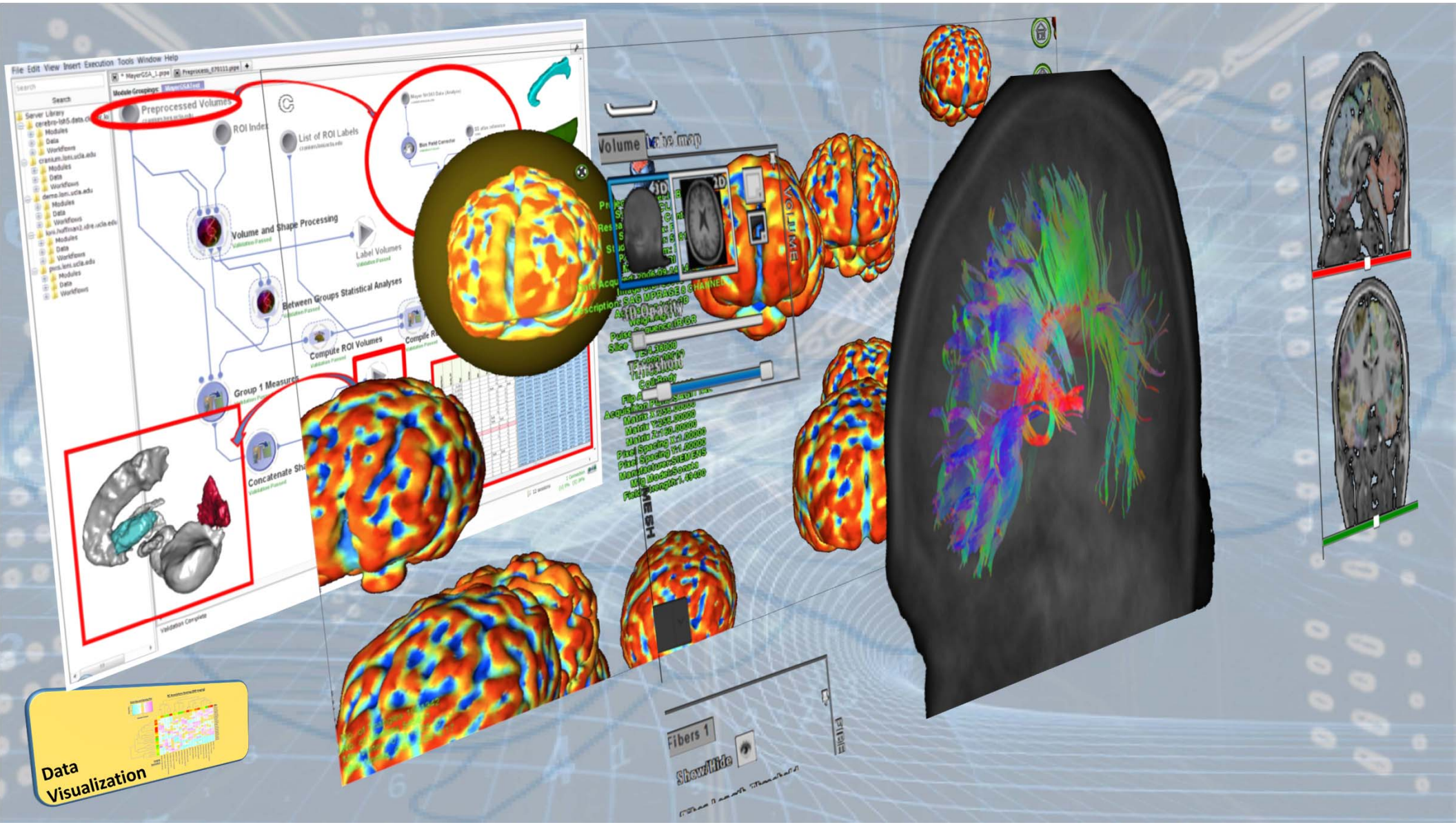

Data Modeling




Data
Share Output Directory
Group 1
ROI Index
Group 2
Group 3
Volume and Shape Processing
Metadata Augmentation
Label Volumes
ROI Shapes
Between Groups Statistical Analyses
Reference Subject Index
Choose Reference Subject
Shape Measures
Copy Shape Files
Group 1 Measures
Group 2 Measures
Group 3 Measures
Design Matrix with Shape Measurements
Concatenate Shape Measure
Compile P Values Group 1-3
Compile P Values Group 3-2
Compile P Values Group 2-1
Compile Results P Values
Threshold
Number of Measures
Number of Comparisons
Choose ROIs
Make Scene
Regions of Significant Differences



PMID:27494614


Preprocessed Volumes
ROI Index
List of ROI Labels
Volume and Shape Processing
Label Volumes
Between Groups Statistical Analyses
Compute ROI Volumes
Group 1 Measures
Concatenate Sha
Validation Complete
Volume
Fibers 1
Show/Hide
Data
Visualization

Gaps, Barriers & Opportunities

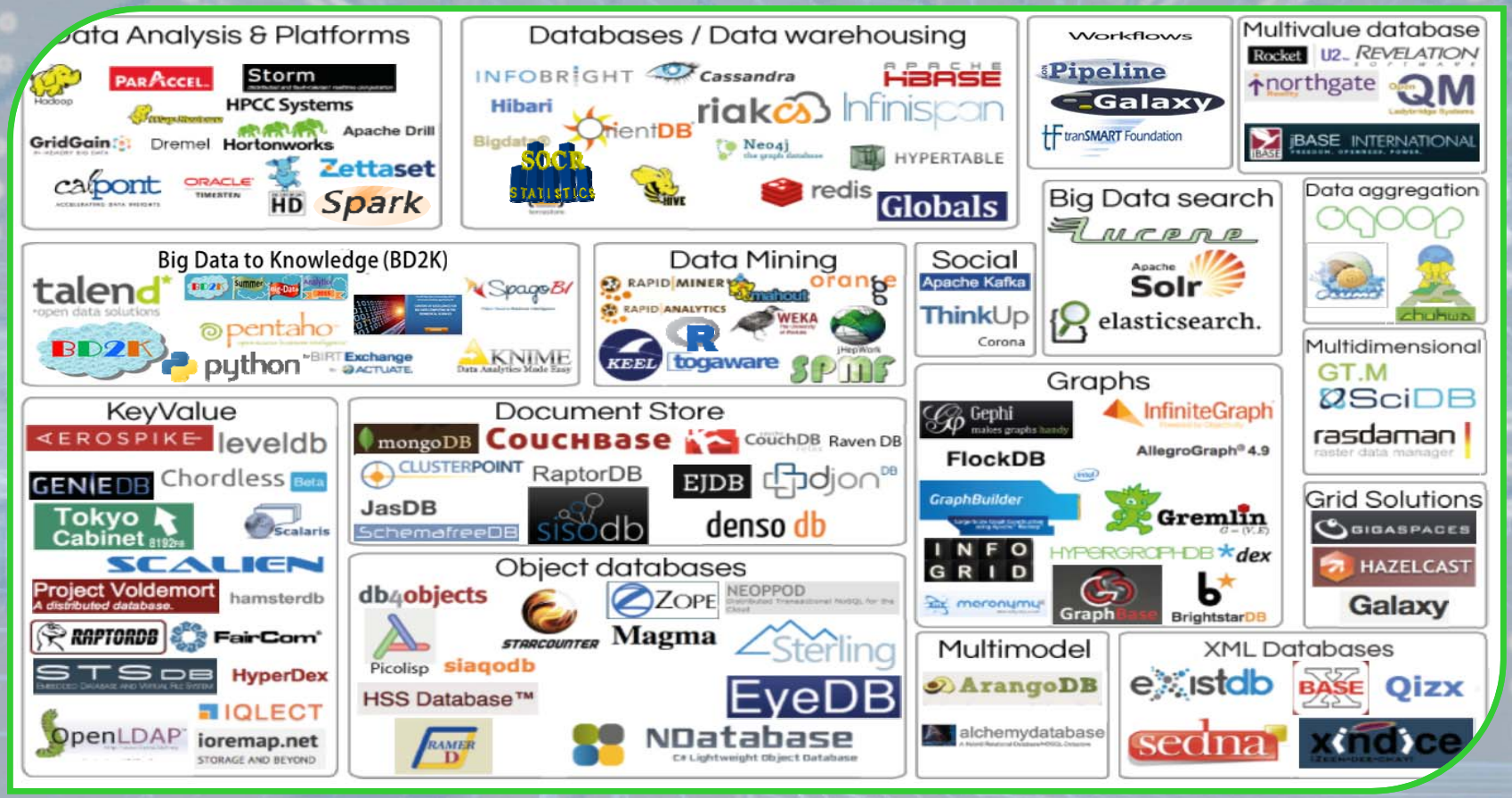
- There is no analytical foundation for systematic representation of Big Data that facilitates the handling of data complexities and at the same time enables joint modeling, information extraction, high-throughput and adaptive scientific inference (cf. CBDA, PMCID: PMC479548)

- Kryder's Law ≫ Moore's Law (more data than we can possibly manage with projected increase of computational power) PMCID: PMC3933453

- Enormous opportunities for algorithm development, trans-disciplinary data-science training, collaborative research using Big Neuroscience Data

- Advance "Open-Science"

Big Data Analytics Resourceome



http://socr.umich.edu/docs/BD2K/BigDataResourceome.html

Examples of Available Resources

- Source Code: https://github.com/SOCR
- End-to-End Pipeline Workflows:
  - Docs: https://wiki.loni.usc.edu
  - Protocols: http://pipeline.loni.usc.edu/explore/library-navigator
- Pubs: http://www.socr.umich.edu/people/dinov/publications.html
- Training/Learning Resources: http://wiki.socr.umich.edu/index.php/SMHS
- Data:
  - Classical: http://wiki.socr.umich.edu/index.php/SOCR_Data
  - Case-Studies: https://umich.instructure.com/courses/38100/files/folder/Case_Studies

Distributed Services

Processing

- Pipeline Try-It-Now Graphical Workflow Environment (Guest access): http://pipeline.loni.usc.edu/products-services/pws/
- socr-pipeline.nursing.umich.edu
- AWS/Galaxy: http://GalaxyProject.org
- tranSMART: https://github.com/transmart

Data

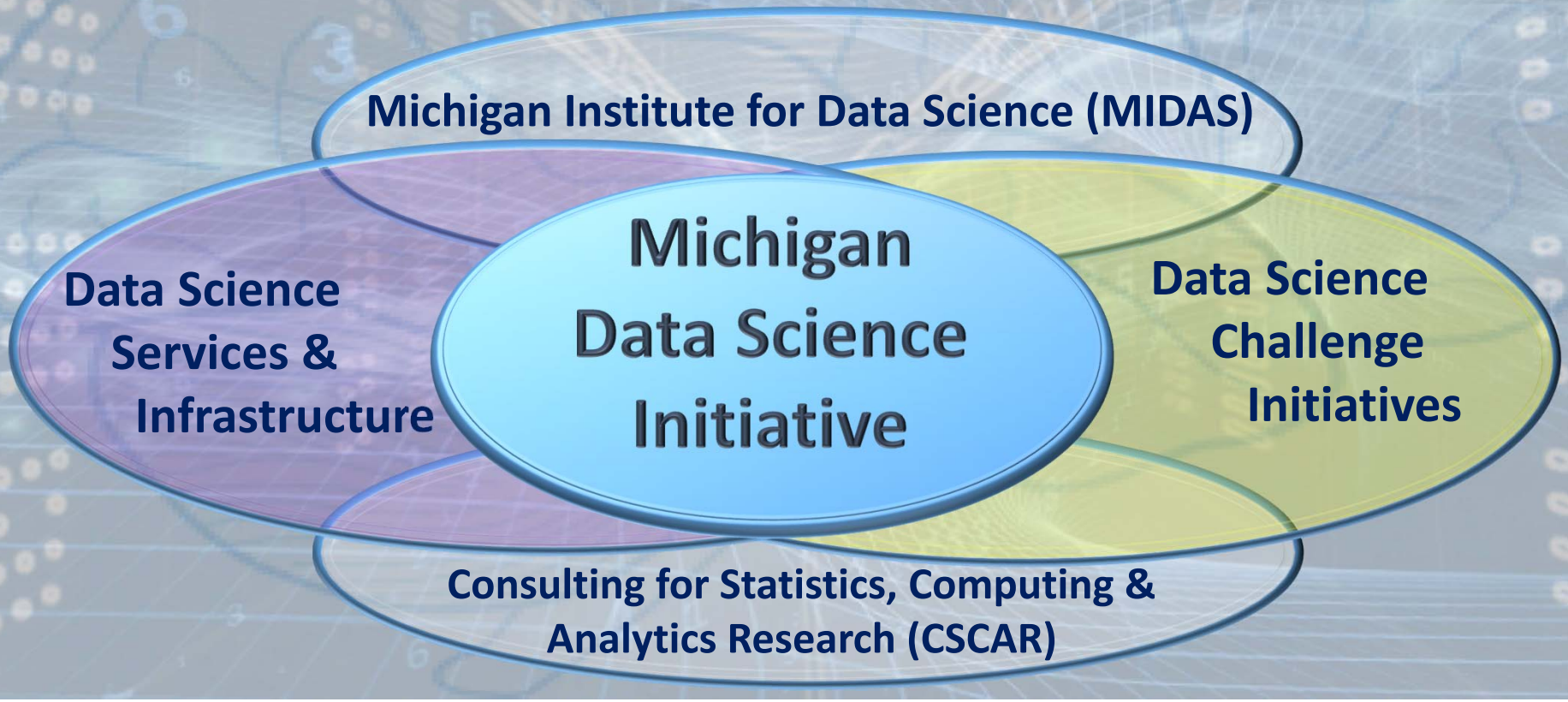
- dbGaP http://dbgap.ncbi.nlm.nih.gov
- Neuroimaging http://IDA.loni.usc.edu
- XNAT: https://central.xnat.org

Transfer

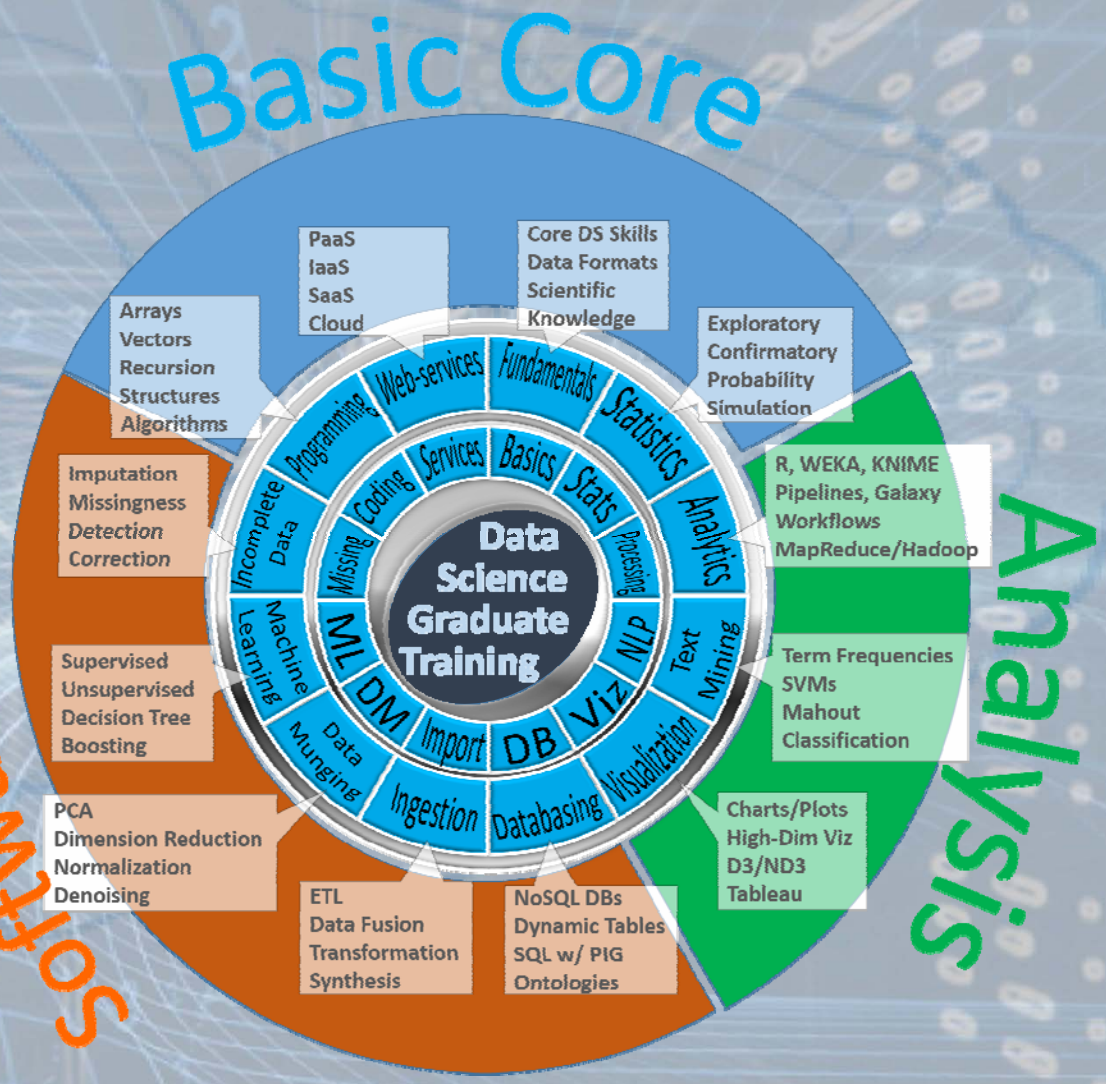
- Globus: http://www.globusonline.org
- GridFTP: http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/

Michigan Institute for Data Science (MIDAS)

MIDAS catalyzes data science at the University of Michigan through support for faculty, research, education and training, and industry engagement



Fundamentals of Data Science Education



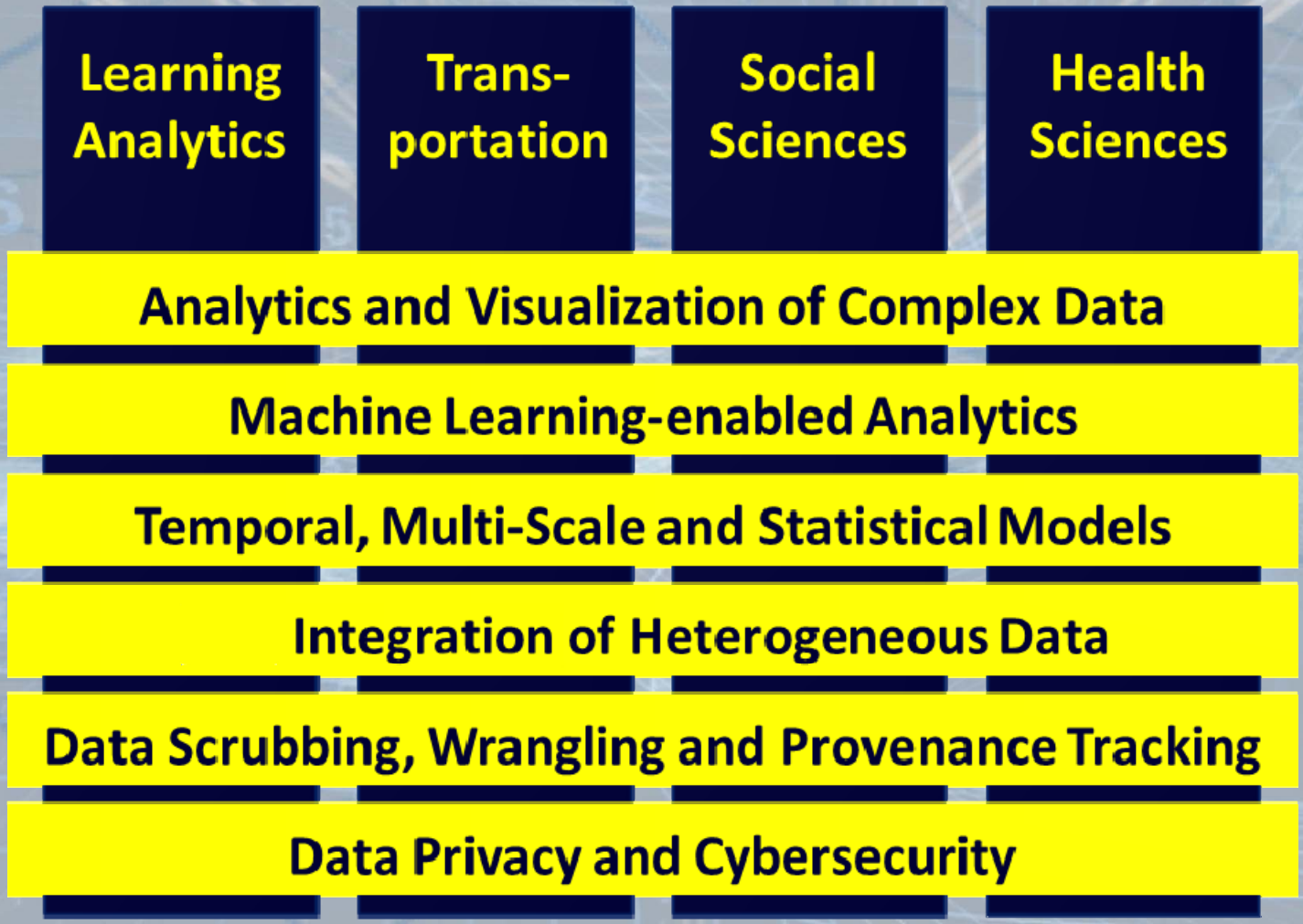
Grad DS Curriculum: Prereqs & Competencies

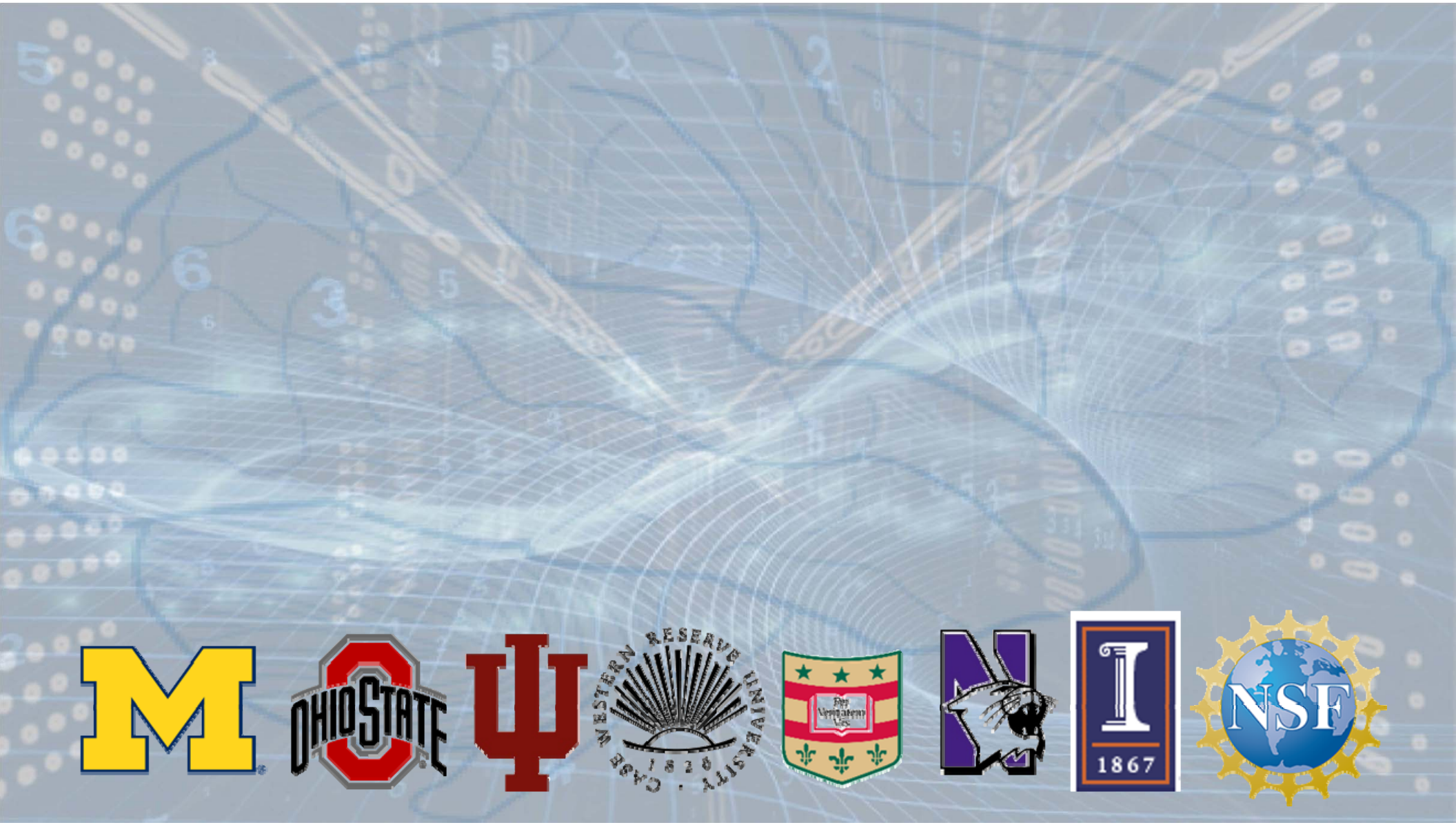
Prerequisites	Skills	Rationale
BS degree or equivalent	Quantitative training and coding skills as described below	The DS certificate is a graduate program requiring a minimum level of quantitative skill
Quantitative training	Undergraduate calculus, linear algebra and intro to probability and statistics	These are the entry level skills required for most upper-level undergrad and grad courses in program
Coding experience	Exposure to software development or programming on the job or in the classroom	Most DS practitioners need substantial experience with Java, C/C++, HTML5, Python, PHP, SQL/DB
Motivation	Significant interest and motivation to pursue quantitative data analytic applications	Dedication for prolonged & sustained immersion into hands-on and methodological research

Grad DS Curriculum: Prereqs & Competencies

Areas	Competency	Expectation	Notes
Algorithms & Applications	Tools	Working knowledge of basic software tools (command-line, GUI based, or web-services)	Familiarity with statistical programming languages, e.g., R or SciKit/Python, and database querying languages, e.g., SQL or NoSQL
	Algorithms	Knowledge of core principles of scientific computing, applications programming, API's, algorithm complexity, and data structures	Best practices for scientific and application programming, efficient implementation of matrix linear algebra and graphics, elementary notions of computational complexity, user-friendly interfaces, strings
	Application Domain	Data analysis experience from at least one application area, either through coursework, internship, research project, etc.	Applied domain examples include: computational social sciences, health sciences, business and marketing, learning sciences, transportation sciences, engineering and physical sciences
Data Management	Data validation & visualization	Curation, Exploratory Data Analysis (EDA) and visualization	Data provenance, validation, visualization via histograms, Q-Q plots, scatterplots (ggplot, Dashboard, D3.js)
	Data wrangling	Skills for data normalization, data cleaning, data aggregation, and data harmonization/registration	Data imperfections include missing values, inconsistent string formatting ('2016-01-01' vs. '01/01/2016', PC/Mac/Lynux time vs. timestamps, structured vs. unstructured data
	Data infrastructure	Handling databases, web-services, Hadoop, multi-source data	Data structures, SOAP protocols, ontologies, XML, JSON, streaming
Analysis Methods	Statistical inference	Basic understanding of bias and variance, principles of (non)parametric statistical inference, and (linear) modeling	Biological variability vs. technological noise, parametric (likelihood) vs non-parametric (rank order statistics) procedures, point vs. interval estimation, hypothesis testing, regression
	Study design and diagnostics	Design of experiments, power calculations and sample sizing, strength of evidence, p-values, False Discovery Rates	Multistage testing, variance normalizing transforms, histogram equalization, goodness-of-fit tests, model overfitting, model reduction
	Machine Learning	Dimensionality reduction, k-nearest neighbors, random forests, AdaBoost, kernelization, SVM, ensemble methods, CNN	Empirical risk minimization. Supervised, semi-supervised, and unsupervised learning. Transfer learning, active learning, reinforcement learning, multiview learning, instance learning

Vertical Integration of MIDAS Challenges and Analytical Methods




OHIO STATE
WESTERN RESERVE UNIVERSITY
CASE
1826
1867
NSF