BIG DATA FOR DISCOVERY SCIENCE

Predictive Big Data Analytics

A Study of Parkinson's Disease using Large, Complex, Heterogeneous, Incongruent, Multi-source & Incomplete Observations

Ivo D. Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard , Mike Darcy, Ravi Madduri , Judy Pa, Cathie Spino, Carl Kesselman, Ian Foster,
Eric W. Deutsch, Nathan D. Price, John D. Van Horn, Joseph Ames, Kristi Clark, Leroy Hood, Benjamin M. Hampstead, William Dauer, and Arthur W. Toga
University of Michigan, Ann Arbor, Institute for Systems Biology, Seattle, University of Southern California, Los Angeles, University of Chicago, Chicago

USC Stevens Neuroimaging and Informatics Institute

USC Laboratory of Neuro Imaging

National Institutes of Health

Computation Institute

USC Viterbi School of Engineering Information Sciences Institute

Institute for Systems Biology

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

Goals

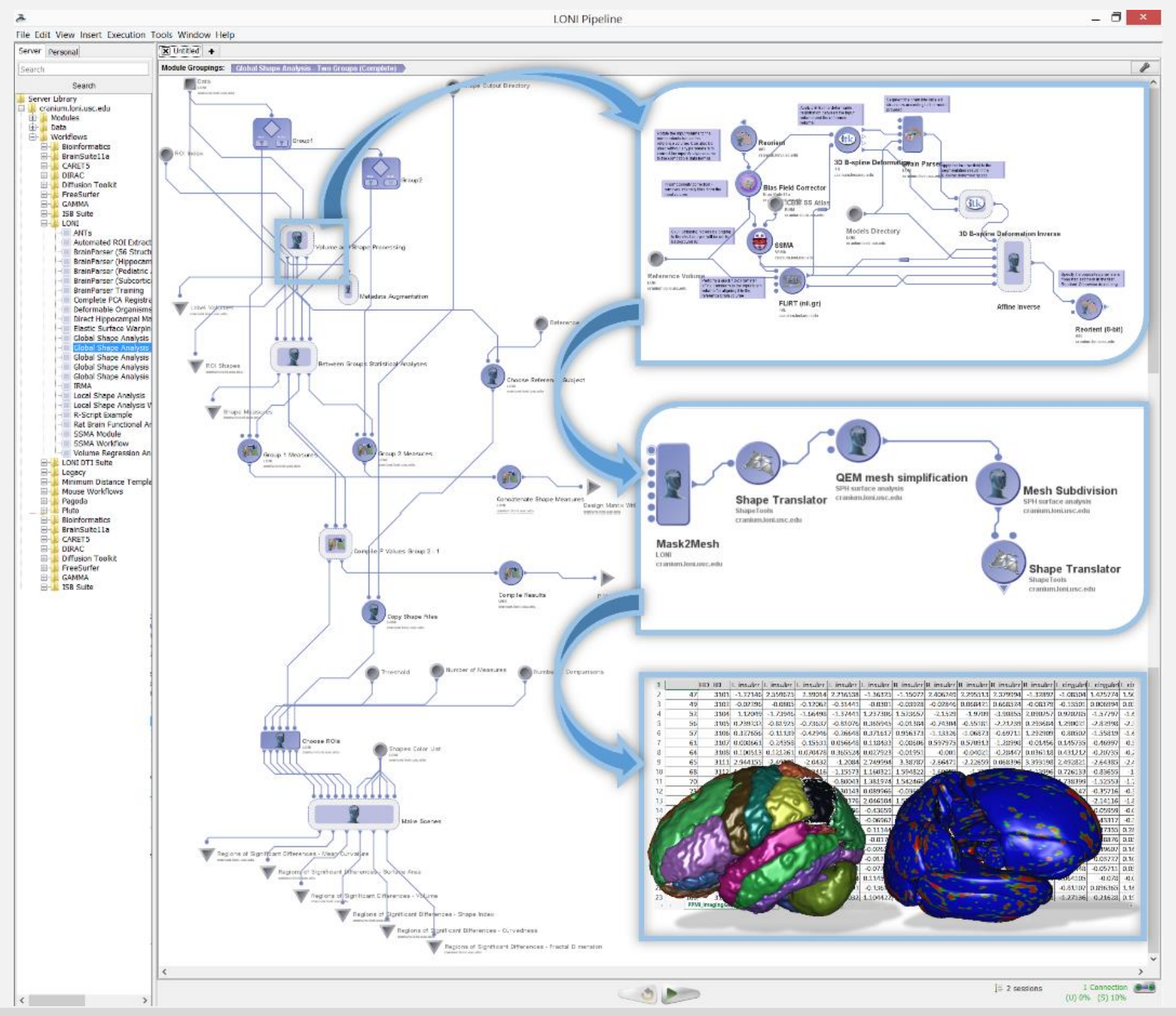
We propose, implement, test, and validate complementary model-based and model-free approaches for Parkinson's Disease (PD) classification and prediction. To explore PD risk using Big Data methodology, we jointly processed complex PPMI imaging, genetics, clinical and demographic data.

We aim to aggregate and harmonize all the data, jointly model the entire data, test model-based and model-free predictive analytics, and statistically validate the results using n-fold cross validation.

Data

The **defining characteristics of Big Data** include: large size, incongruency, incompleteness, complexity, multiplicity of scales, and heterogeneity of information-generating sources. These pose challenges to the classical techniques for data management, processing, visualization and interpretation. A unique archive of Big Data on Parkinson's Disease is collected, managed and disseminated by the Parkinson's Progression Markers Initiative (PPMI), N=600.

Data Elements demographics, clinical tests (physical, verbal learning and language, neurological and olfactory (U Penn Smell Identification Test, UPSIT) tests), vital signs, MDS-UPDRS scores (Unified Parkinson's Disease Rating Scale), ADL (activities of daily living), Montreal Cognitive Assessment (MoCA), Epworth Sleepiness Scale, REM sleep behavior questionnaire, Geriatric Depression Scale (GDS-15), and State-Trait Anxiety Inventory for Adults, and 3D sMRI (http://www.ppmi-info.org/access-data-specimens).



Materials & Methods

We tested generalized linear models (with fixed and random effects) as well as multiple classification methods to discriminate between Parkinson's disease (PD) patients and asymptomatic healthy controls (HC).

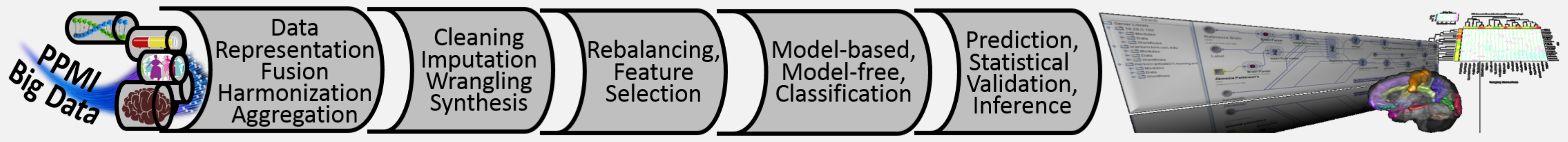
Previous studies have reported results of integrating multiple types of data to diagnose, track and predict Parkinson's disease using imaging and genetics, genome-wide association studies, animal phenotypic models, molecular imaging, pharmacogenetics, phenomics and genomics. However, **few studies have reported strategies to efficiently and effectively handle all available multi-source data to produce high-fidelity predictive models** of neurodegenerative disorders.

The main **contributions** of this study include:

- an approach for rebalancing initially imbalanced cohorts,
- applying a wide spectrum of automated classification methods that generate consistent and powerful phenotypic predictions (e.g., diagnosis),
- developing a reproducible machine-learning based protocol for classification that enables the reporting of model parameters and outcome forecasting, and
- using generalized estimating equations models to assess population-wide differences based on incomplete longitudinal Big Data.

Predictive Analytics

- **Model-based** approaches included generalized linear models (GLM), mixed effect modeling with repeated measurements (MMRM), change-based models, and generalized estimating equations (GEE),
- **Model-free** predictive analytics involved forecasting, classification, and data mining. Specific examples of such model-free methods include AdaBoost, support vector machine (SVM), Naïve Bayes, Decision Tree, KNN, and K-Means classifiers. Both types of approaches (model-based or model-free) facilitate classification, prediction, and outcome forecasting (e.g., disease state) using new or testing data containing the same clinical, demographic, imaging and phenotypic data elements.


PPMI Big Data
Data Representation Fusion Harmonization Aggregation
Cleaning Imputation Wrangling Synthesis
Rebalancing, Feature Selection
Model-based, Model-free, Classification
Prediction, Statistical Validation, Inference

Resource Availability: All raw PPMI data is available from the PPMI consortium (www.ppmi-info.org/data). The computational protocol, source-code, scripts, and derived data we generated as part of this study, along with the complete GSA pipeline workflow are available in the BDDS GitHub repository (https://github.com/BD2K/BDDS).

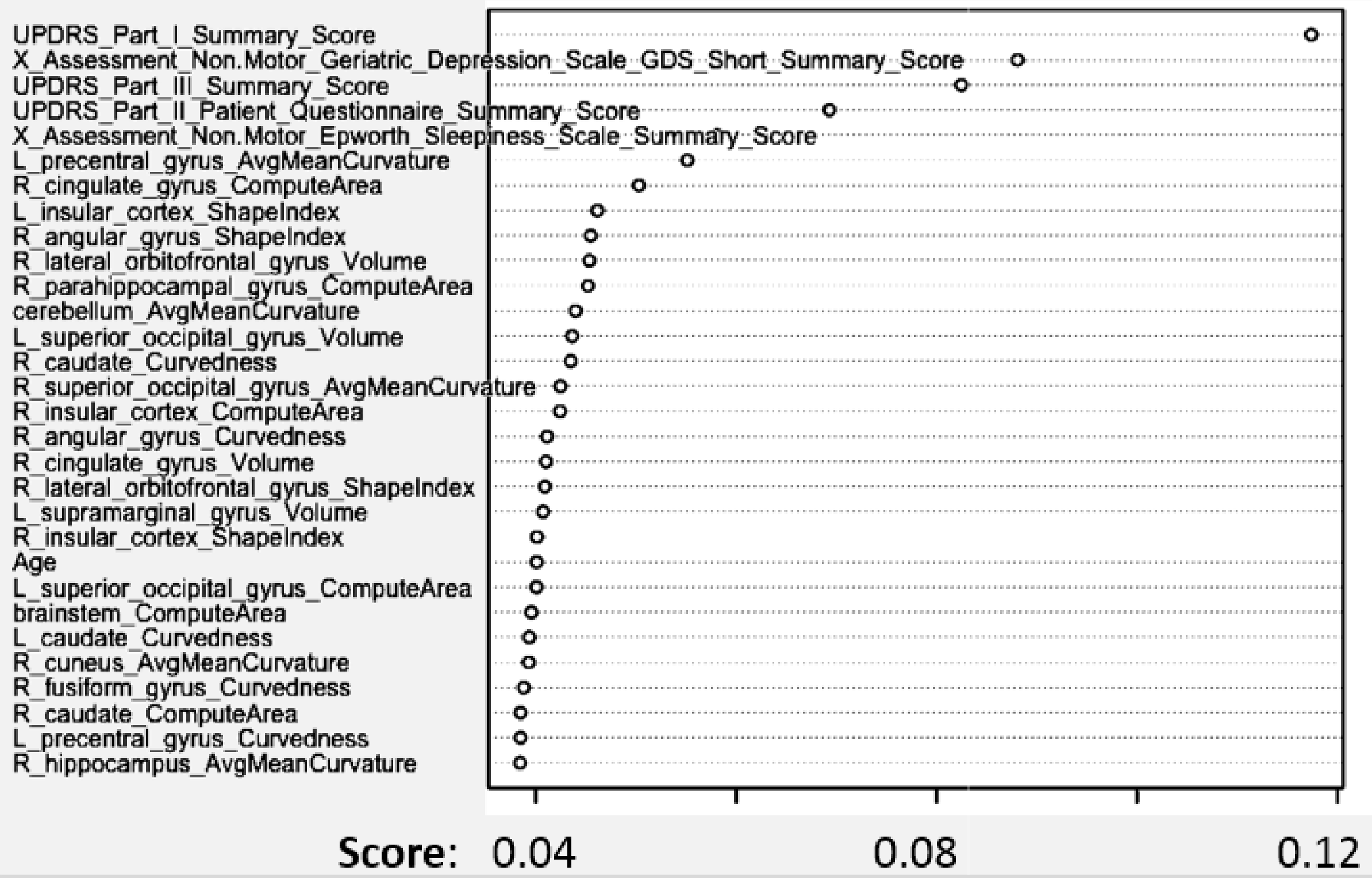
Results

Developed **reproducible protocols** for end-to-end data analytics to provide a scalable solution to discovery-based Big Data science, facilitate active trans-disciplinary collaborations, and entice independent community validation of algorithmic modules, atomic tools, and complete end-to-end workflows.

High power to predict Parkinson's disease (consistent accuracy, sensitivity, and specificity exceeding 96%, confirmed using statistical n-fold cross-validation). Clinical (e.g., Unified Parkinson's Disease Rating Scale (UPDRS) scores), demographic (e.g., age), genetics (e.g., rs34637584, chr12), and derived neuroimaging biomarker (e.g., cerebellum shape index) data all contributed to the predictive analytics and diagnostic forecasting.

ML classifier	accuracy	sensitivity	specificity	positive predictive value	negative predictive value	log odds ratio (LOR)
AdaBoost	0.996324	0.994141	0.998264	0.9980392	0.9948097	11.4882058
SVM	0.985294	0.994140	0.977431	0.9750958	0.9946996	8.902166

PD-predictive Data Elements



Varplot illustrates

- critical predictive data elements (Y-axis)
- and their impact scores (X-axis)

AdaBoost classifier for Controls vs. Patients prediction

Acknowledgments

Big Data Discovery Science
www.BD2K.org, INI, USC

NSF 1023115 & NIH U54 EB020406

Statistics Online Computational Resource
www.SOCR.umich.edu, HBBS, U Michigan