

Data Science & Predictive Neuro-Analytics

Ivo D. Dinov

Statistics Online Computational Resource

Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science

University of Michigan

<http://SOCR.umich.edu>

<http://Predictive.Space>



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

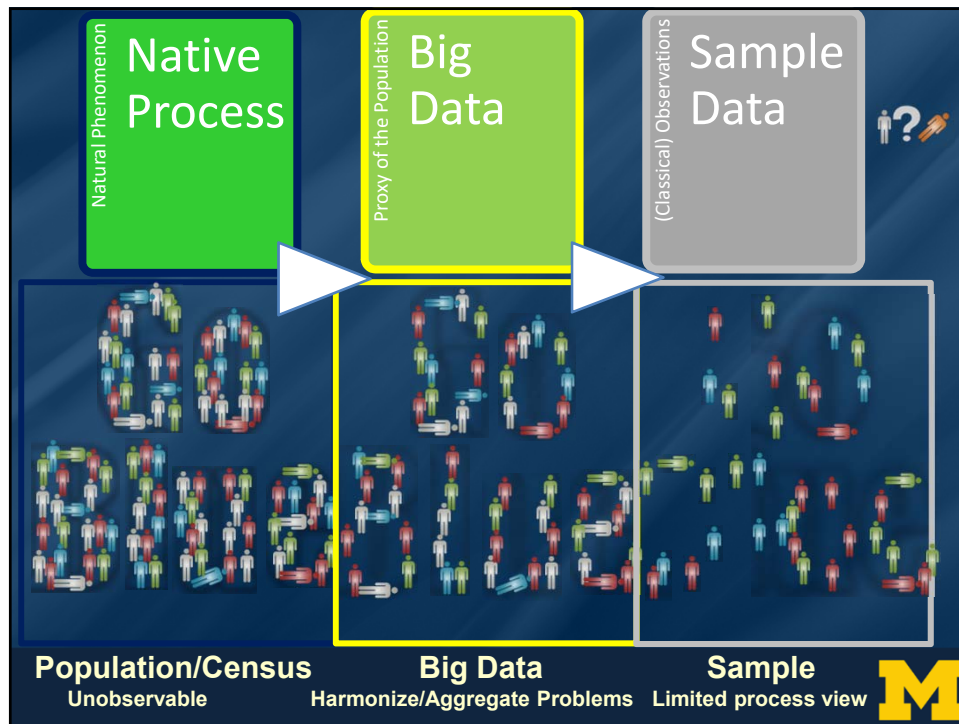
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)

Slides Online:
"SOCR News"

Outline

- ☐ Driving biomedical & health challenges
- ☐ Common characteristics of Big Neuroscience Data
- ☐ Data science & predictive neuro-analytics
- ☐ Compressive Big Data Analytics (CBDA)
- ☐ Case-studies
 - ☐ Applications to Neurodegenerative Disease
 - ☐ Population Census-like Neuroscience





Characteristics of Big Biomed Data

IBM Big Data 4V's: Volume, Variety, Velocity & Veracity

Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale observations
Time	Techniques accounting for longitudinal patterns in the data
Incomplete	Reliable management of missing data

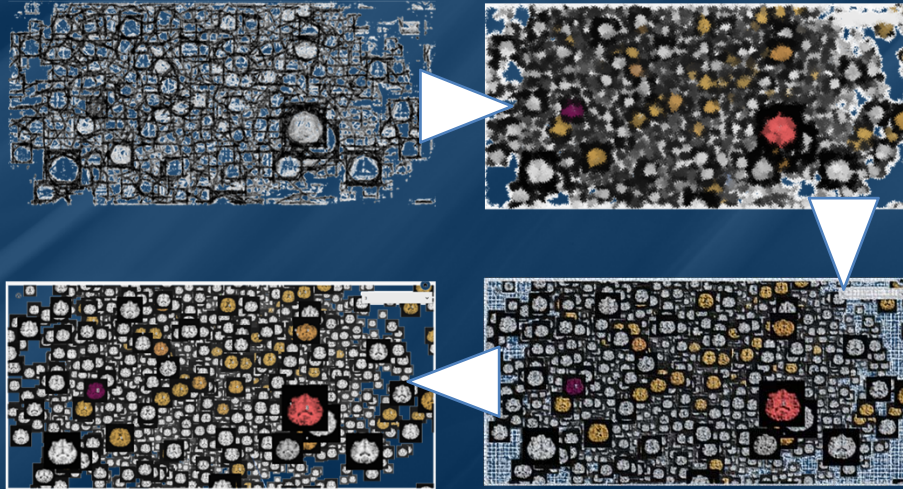
Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Dinov, et al. (2016) PMID:26918190

M

Multiscale/Multimodal NI Data



http://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB



Data Science & Predictive Analytics

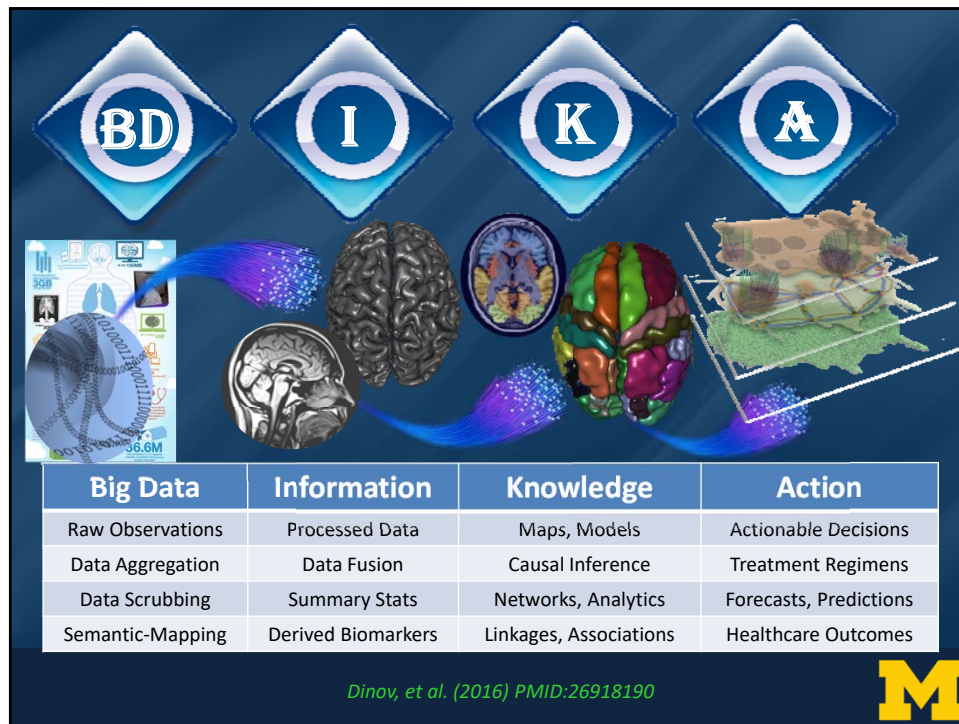
- ❑ **Data Science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and applied areas. Deals with enormous amounts of complex, incongruent and dynamic data from multiple sources. Aims to develop algorithms, methods, tools, and services capable of ingesting such datasets and supplying semi-automated decision support systems
- ❑ **Predictive Analytics**: process utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools, and distributed web-services to represent, interrogate, and interpret complex data. Aims to forecast trends, cluster patterns in the data, or prognosticate the process behavior either within the range or outside the range of the observed data (e.g., in the future, or at locations where data may not be available)



<http://DSPA.predictive.space>

Dinov, Springer (2018)





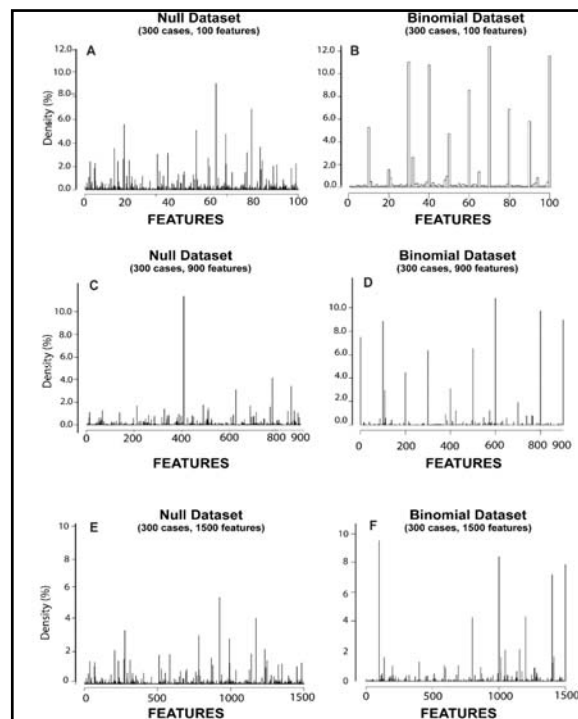
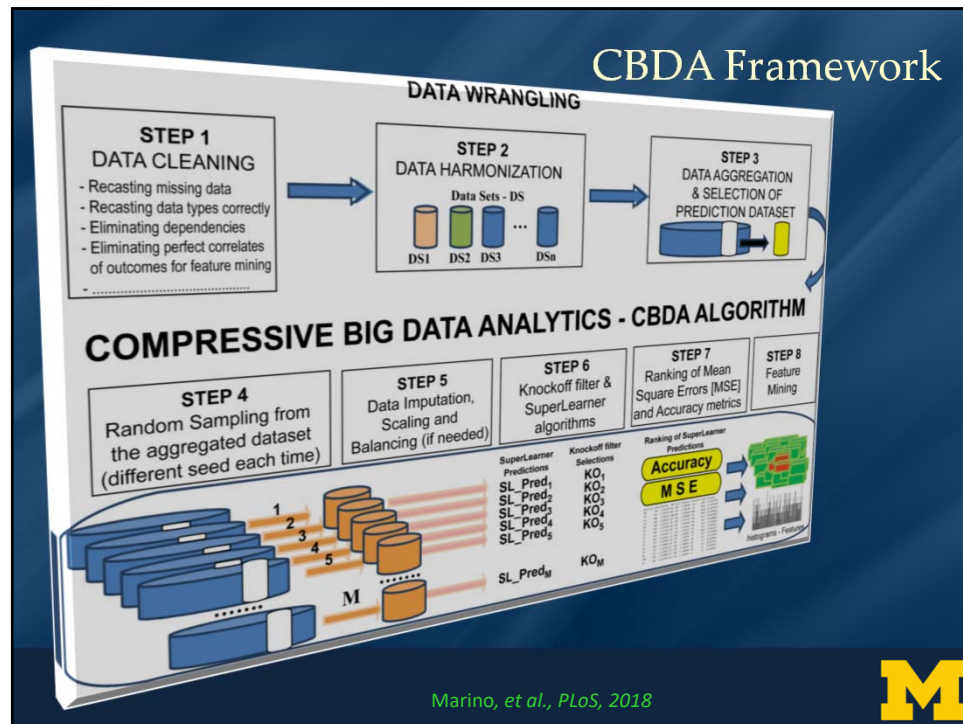
Compressive Big Data Analytics (CBDA)

□ Foundation for Compressive Big Data Analytics (CBDA)

- Iteratively generate random (sub)samples from the Big Data collection
- Then, using classical techniques to obtain model-based, model-free, non-parametric inference based on the sample
- Next, compute likelihood estimates (e.g., probability values quantifying effect sizes, relations, and other associations)
- Repeat – the process continues iteratively until a convergence criterion is met – the (re)sampling and inference steps many times (with or without using the results of previous iterations as priors for subsequent steps)

Dinov, J Med Stat Inform, 2016;

Marino, et al., PLoS, 2018



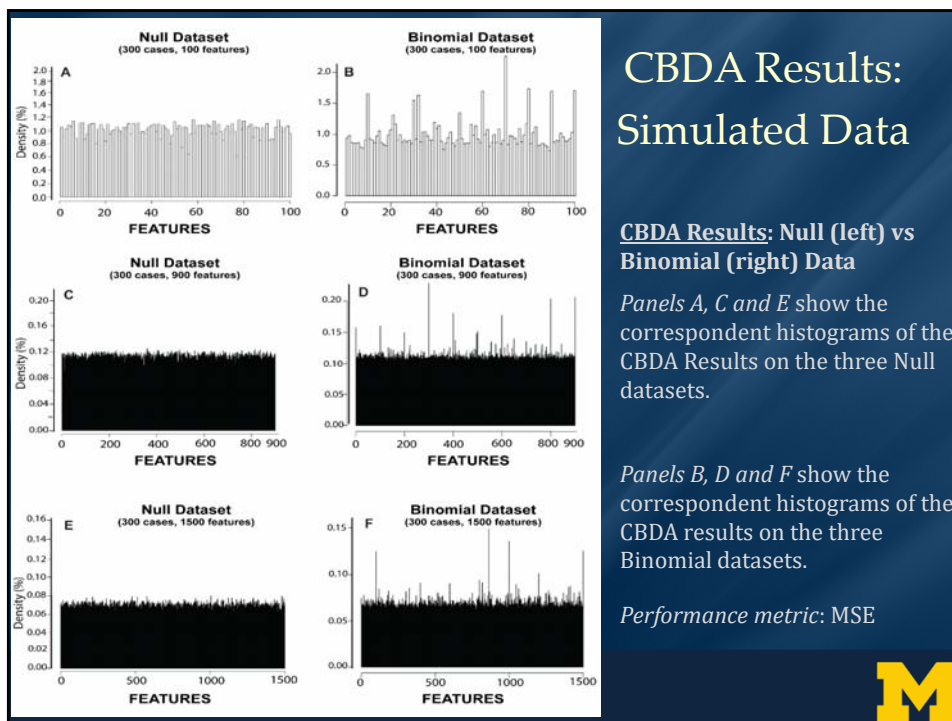
CBDA Results: Simulated Data

Knockoff of Null (left) vs Binomial (right) Data

Panels A, C and E show the correspondent histograms generated from the Knockoff Filter algorithm on the three Null datasets.

Panels B, D and F show the correspondent histograms generated from the Knockoff Filter algorithm on the three Binomial datasets.

Performance metric: MSE



CBDA Results: Biomed Data (ADNI)

	Reference		
Prediction	AD	MCI	Normal
AD	69	17	1
MCI	12	243	8
Normal	0	9	140

Overall Statistics	
Accuracy	0.9058 [95% CI = (0.8767, 0.93)]
No Information Rate	0.5391
P-Value [Acc > NIR]	<2e-16
Kappa	0.8426
McNemar's Test P-Value	0.589

Statistics by Diagnostic Class			
	AD	MCI	Normal
Sensitivity	0.8519	0.9033	0.9396
Specificity	0.9569	0.9130	0.9743
Positive Pred Value	0.7931	0.9240	0.9396
Negative Pred Value	0.9709	0.8898	0.9743
Prevalence	0.1623	0.5391	0.2986
Balanced Accuracy	0.9044	0.9082	0.9569

CBDA multinomial classification results (ADNI)

M

Marino, et al., PLoS, 2018

Case-Studies – ALS

- Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

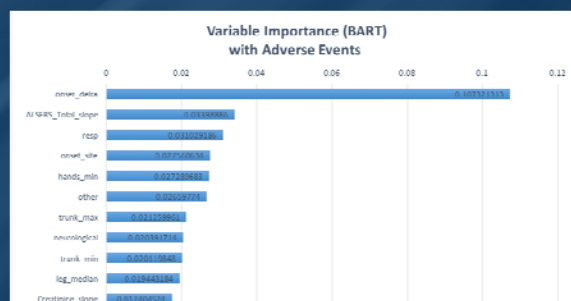
Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 variables are recorded for all subjects including: <u>Demographics</u> : age, race, medical history, sex; <u>Clinical data</u> : <u>Amyotrophic Lateral Sclerosis</u> Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole) The PRO-ACT training dataset contains clinical and lab test information of 8,635 patients. Information of 2,424 study subjects with valid gold standard ALSFRS slopes used for processing, modeling and analysis	The time points for all longitudinally varying data elements are aggregated into signature vectors. This facilitates the modeling and prediction of ALSFRS slope changes over the first three months (baseline to month 3)

Tang, et al. (2018), in review

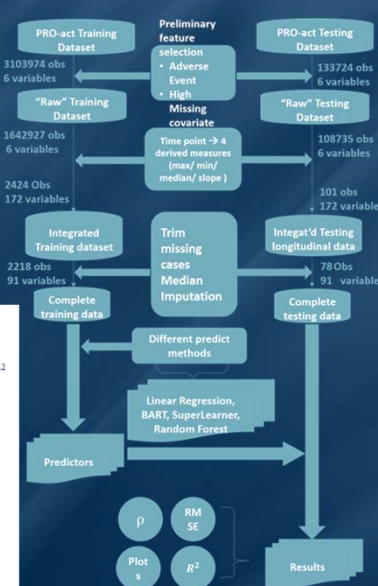


Case-Studies – ALS

- Detect, track, and prognosticate the progression of ALS
- Predict adverse events based on subject phenotype and 0-3 month clinical assessment changes



Methods	Linear Regression	Random Forest	BART	SuperLearner
R-squared	0.081	0.174	0.225	0.178
RMSE	0.619	0.587	0.568	0.585
Correlation	0.298	0.434	0.485	0.447



Case-Studies – ALS

- ❑ **Main Finding:** predicting univariate clinical outcomes may be challenging, the (information energy) signal is very weak. We can cluster ALS patients and generate evidence-based ALS hypotheses about the complex interactions of *multivariate factors*
- ❑ **Classification vs. Clustering:**
 - ❑ Classifying univariate clinical outcomes using the PRO-ACT data yields only marginal accuracy (about 70%).
 - ❑ Unsupervised clustering into sub-groups generates stable, reliable and consistent computable phenotypes whose explication requires interpretation of multivariate sets of features

Data
Representation
Fusion
Harmonization
Aggregation

Cleaning
Imputation
Wrangling
Synthesis

Model-based,
Model-free,
Classification,
Clustering,
Inference

Cluster	Consistency	Variance	Cluster-Size	Silhouette
1	1	0	565	0.58
2	0.986	0.018	427	0.63
3	0.956	0.053	699	0.5
4	0.985	0.018	733	0.5

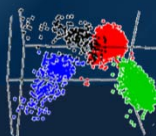
Tang, et al. (2018), in review



Case-Studies – ALS – Explicating Clustering

Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
onset_age						
onset_delta.x	1	1	1			
onset_delta.y	1	1	1			
RedBloodCells_RBC_min						
RedBloodCells_RBC_median						
RedBloodCells_RBC_slope						
Q4_Handwriting_min						
Q4_Handwriting_max						
Q4_Handwriting_median						
Q9_Climbing_Stairs_min						
Q9_Climbing_Stairs_max						
Q9_Climbing_Stairs_slope						
Q9_Walking_min						
Q9_Walking_max						
Q9_Walking_median						
trunk_min						
trunk_max						
trunk_slope						
Protein_slope						
Creatinine_max						
Creatinine_min						
Creatinine_median						
respiratory_rate_max						
hand_min						
hand_max						
hand_slope						
Q8_Dressing_and_Hygiene_max						
Q8_Dressing_and_Hygiene_min						
Q8_Dressing_and_Hygiene_median						
Q9_Turning_in_Bed_min						
Q9_Turning_in_Bed_max						
Q9_Turning_in_Bed_slope						
SLFRT_Total_max						
ALFRT_Total_min						
ALFRT_Total_max						
ALFRT_Total_min						
Hematocrit_min						
Hematocrit_max						
Hematocrit_slope						
leg_min						
leg_max						
leg_slope						
muscle_min						
AbsoluteNeutrophilCount_min						
AbsoluteNeutrophilCount_max						
AbsoluteNeutrophilCount_slope						
AbsoluteNeutrophilCount_min						
AbsoluteNeutrophilCount_max						
AbsoluteNeutrophilCount_slope						
AbsoluteLymphocyteCount_min						
AbsoluteLymphocyteCount_max						
AbsoluteLymphocyteCount_slope						

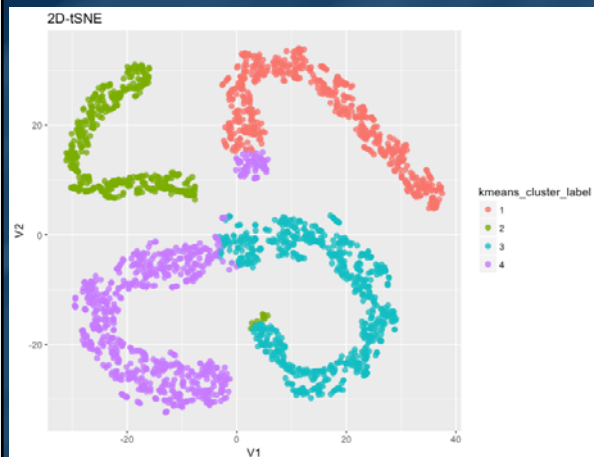
Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
...						
onset_delta.x	1	1	1	1	1	1
...						
Q9_Climbing_Stairs_slope	1			1		
...						
leg_max		1	1	1	1	
...						



Tang, et al. (2018), in review



Case-Studies – ALS – Dimensionality Reduction



2D t-SNE Manifold embedding

Learn a mapping: $f: R^n \xrightarrow{n \gg d} R^d$
 $\{x_1, x_2, \dots, x_n\} \rightarrow \{y_1, y_2, \dots, y_d\}$
preserves closely the *original distances*, $p_{i,j}$ and represents the *derived similarities*, $q_{i,j}$ between pairs of embedded points:

$$q_{i,j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}$$

$$\min_f KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

Tang, et al. (2018), in review

$$0 = \frac{\partial KL(P||Q)}{\partial y_i} = 2 \sum_j (p_{i,j} - q_{i,j}) f(\|x_i - x_j\|) u_{i,j}$$

$$f(z) = \frac{z}{1+z^2} \text{ and } u_{i,j} \text{ is a unit vector from } y_j \text{ to } y_i.$$



Case-Studies – Parkinson's Disease

- ❑ **Investigate falls in PD patients** using clinical, demographic and neuroimaging data from two independent initiatives (UMich & Tel Aviv U)
- ❑ Applied **controlled feature selection** to identify the most salient predictors of patient falls (gait speed, Hoehn and Yahr stage, postural instability and gait difficulty-related measurements)
- ❑ **Model-based** (e.g., GLM) and **model-free** (RF, SVM, Xgboost) analytical methods used to forecasts clinical outcomes (e.g., falls)
- ❑ Internal statistical cross **validation** + external out-of-bag validation
- ❑ Four specific **challenges**
 - ❑ Challenge 1, harmonize & aggregate complex, multisource, multisite PD data
 - ❑ Challenge 2, identify salient predictive features associated with specific clinical traits, e.g., patient falls
 - ❑ Challenge 3, forecast patient falls and evaluate the classification performance
 - ❑ Challenge 4, predict tremor dominance (TD) vs. posture instability and gait difficulty (PIGD).
- ❑ **Results:** model-free machine learning based techniques provide a more reliable clinical outcome forecasting, e.g., falls in Parkinson's patients, with classification accuracy of about 70-80%.

Gao, et al. SREP (2018)



Case-Studies – Parkinson's Disease



Falls in PD are extremely difficult to predict ...

PD phenotypes
Tremor-Dominant (TD)
Postural Instability & gait difficulty (PI & GD)



Case-Studies – Parkinson's Disease

Method	acc	sens	spec	ppv	npv	lor	auc
Logistic Regression	0.728	0.537	0.855	0.710	0.736	1.920	0.774
Random Forests	<u>0.796</u>	<u>0.683</u>	<u>0.871</u>	<u>0.778</u>	<u>0.806</u>	<u>2.677</u>	<u>0.821</u>
AdaBoost	0.689	0.610	0.742	0.610	0.742	1.502	0.793
XGBoost	0.699	0.707	0.694	0.604	0.782	1.699	0.787
SVM	0.709	0.561	0.806	0.657	0.735	1.672	0.822
Neural Network	0.699	0.610	0.758	0.625	0.746	1.588	
Super Learner	0.738	0.683	0.774	0.667	0.787	1.999	

Results of binary fall/no-fall classification (5-fold CV) using top 10 selected features (gaitSpeed_Off, ABC, BMI, PIGD_score, X2.11, partII_sum, Attention, DGI, FOG_Q, H_and_Y_OFF)

Gao, et al. SREP (2018)



Open-Science & Collaborative Validation

End-to-end Big Data analytic protocol jointly processing complex imaging, genetics, clinical, demo data for assessing PD risk

- Methods for rebalancing of imbalanced cohorts
- ML classification methods generating consistent and powerful phenotypic predictions
- Reproducible protocols for extraction of derived neuroimaging and genomics biomarkers for diagnostic forecasting

<https://github.com/SOCR/PBDA>



Case-Studies – General Populations

2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain MRI measuring method
100	12663	Brain MRI	Reason believed unsafe to perform brain MRI
100	12704	Brain MRI	Reason brain MRI not completed
100	12652	Brain MRI	Reason brain MRI not performed
101	12292	Carotid ultrasound	Carotid ultrasound measurement completed
101	12291	Carotid ultrasound	Carotid ultrasound measuring method
101	20235	Carotid ultrasound	Carotid ultrasound results package
101	22672	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 120 degrees
101	22675	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 150 degrees
101	22678	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 180 degrees
101	22681	Carotid ultrasound	Maximum carotid IMT (intima-medial thickness) at 210 degrees
101	22671	Carotid ultrasound	Mean carotid IMT (intima-medial thickness) at 120 degrees
101	22674	Carotid ultrasound	Mean carotid IMT (intima-medial thickness) at 150 degrees
101	22677	Carotid ultrasound	Mean carotid IMT (intima-medial thickness) at 180 degrees
101	22680	Carotid ultrasound	Mean carotid IMT (intima-medial thickness) at 210 degrees
101	22670	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 120 degrees
101	22673	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 150 degrees
101	22676	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 180 degrees
101	22679	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 240 degrees
101	22682	Carotid ultrasound	Quality control indicator for IMT at 120 degrees
101	22683	Carotid ultrasound	Quality control indicator for IMT at 150 degrees
101	22684	Carotid ultrasound	Quality control indicator for IMT at 180 degrees
101	22685	Carotid ultrasound	Quality control indicator for IMT at 210 degrees
101	22686	Carotid ultrasound	Quality control indicator for IMT at 240 degrees

- UK Biobank – discriminate between HC, single and multiple comorbid conditions
- Predict likelihoods of various developmental or aging disorders
- Forecast cancer

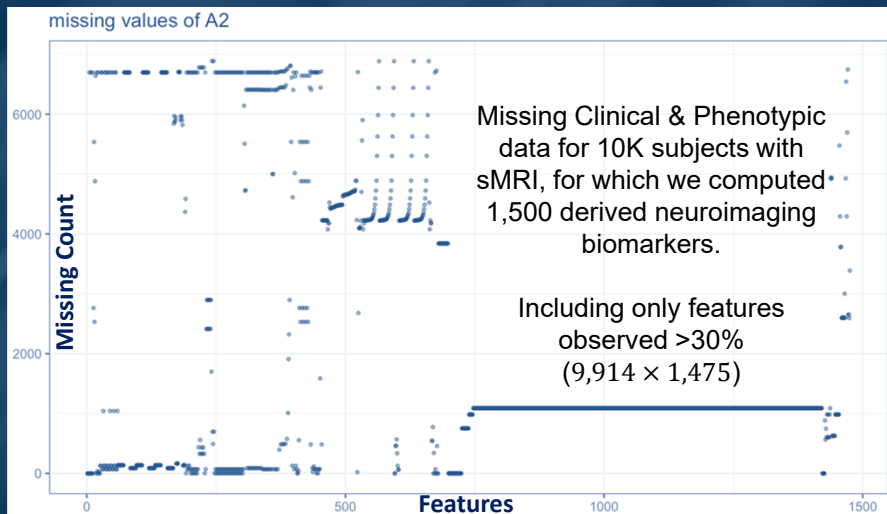
Data Source	Sample Size/Data Type	Summary
UK Biobank	Demographics: > 500K cases Clinical data: > 4K features Imaging data: T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI Genetics data	The longitudinal archive of the UK population (NHS)

<http://www.ukbiobank.ac.uk>

<http://bd2k.org>



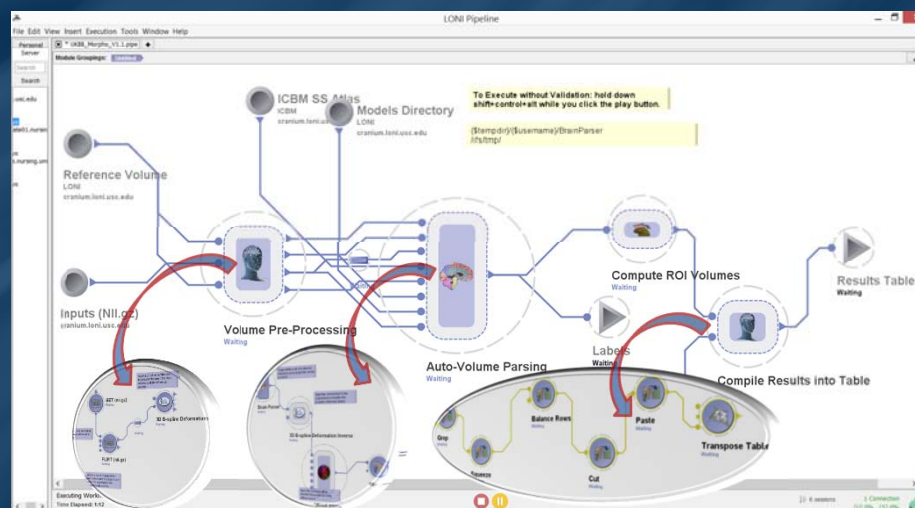
Case-Studies – UK Biobank (Complexities)



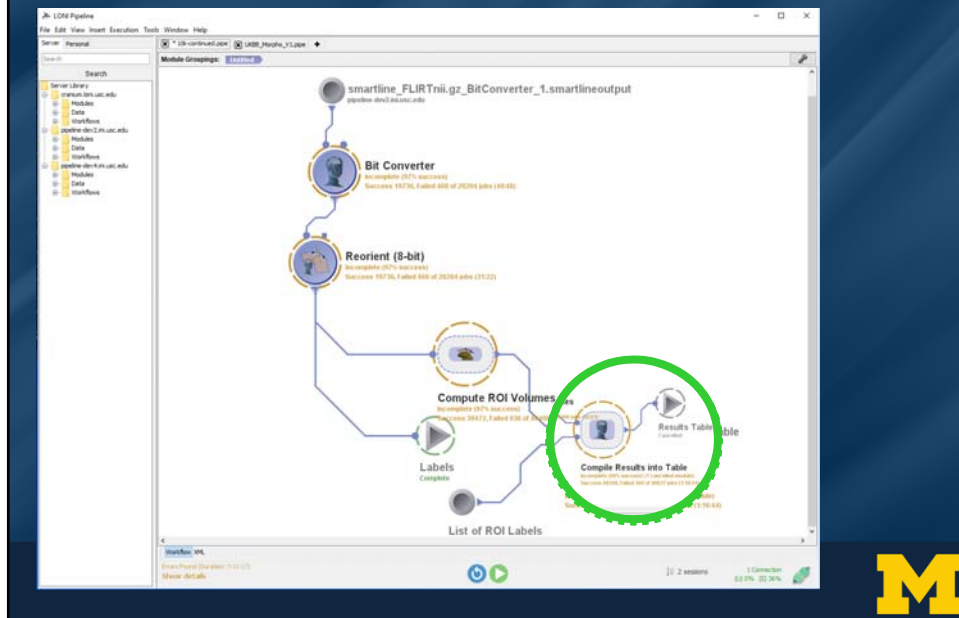
Zhou, et al. (2018), in review



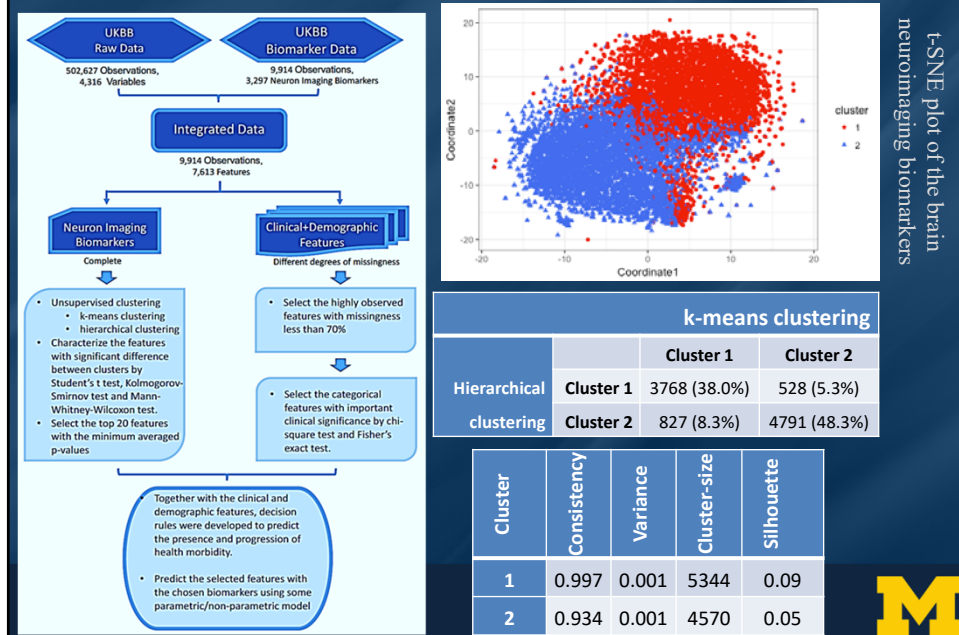
Case-Studies – UK Biobank – NI Biomarkers



Case-Studies – UK Biobank – Successes/Failures



Case-Studies – UK Biobank – Results



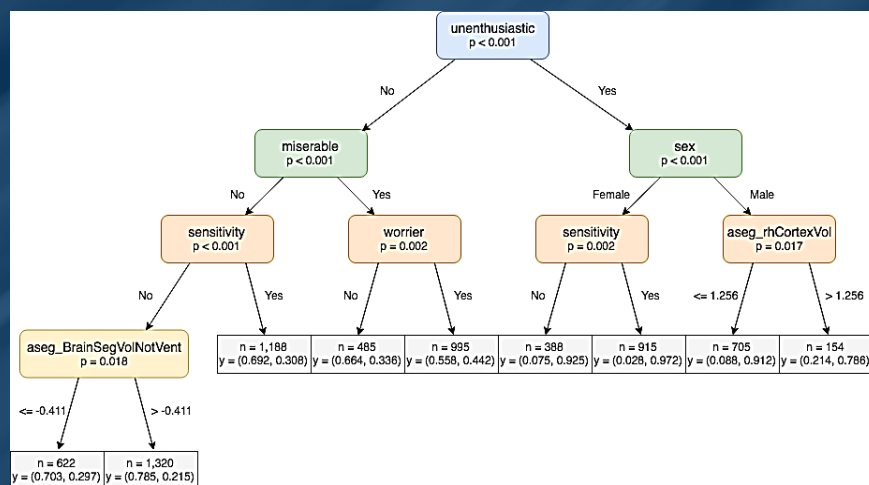
Case-Studies – UK Biobank – Results

Variable	Cluster 1	Cluster 2
Sex		
Female	1,134 (24.7%)	4,062 (76.4%)
Male	3,461 (75.3%)	1,257 (23.6%)
Sensitivity/hurt feelings		
Yes	2,142 (47.9%)	3,023 (58.4%)
No	2,312 (52.1%)	2,151 (41.6%)
Worrier/anxious feelings		
Yes	2,173 (48.2%)	2,995 (57.4%)
No	2,317 (51.8%)	2,208 (42.6%)
Risk taking		
Yes	1,378 (31.0%)	1,154 (22.2%)
No	3,064 (69.0%)	3,933 (77.8%)
Guilt feelings		
Yes	1,100 (24.4%)	1,697 (32.4%)
No	3,417 (75.6%)	3,536 (67.6%)
Seen doctor for nerves, anxiety, tension or depression		
Yes	1,341 (29.3%)	1,985 (37.9%)
No	2,227 (70.7%)	2,210 (62.1%)
Alcohol usually taken with meals		
Yes	1,854 (66.7%)	2,519 (76.4%)
No	924 (33.3%)	771 (23.6%)
Snoring		
Yes	1,796 (41.1%)	1,652 (33.4%)
No	2,577 (58.9%)	3,306 (66.6%)
Worry too long after embarrassment		
Yes	1,978 (44.3%)	2,675 (52.3%)
No	2,491 (55.7%)	2,462 (47.7%)
Miserableness		
Yes	1,715 (37.7%)	2,365 (45.3%)
No	2,829 (62.3%)	2,882 (54.7%)
Ever highly irritable/argumentative for 2 days		
Yes	485 (10.7%)	749 (14.5%)
No	4,006 (89.3%)	4,448 (85.5%)
Nervous feelings		
Yes	751 (16.6%)	1,071 (20.8%)
No	3,763 (83.4%)	4,076 (79.2%)
Ever depressed for a whole week		
Yes	2,176 (48.1%)	2,739 (52.8%)
No	2,347 (51.9%)	2,438 (47.2%)
Ever unenthusiastic/disinterested for a whole week		
Yes	1,346 (30.3%)	1,743 (34.3%)
No	3,089 (69.7%)	3,344 (65.7%)
Sleeplessness		
Never/rarely	1,367 (29.8%)	1,181 (22.8%)
Sometimes	2,202 (47.9%)	2,571 (49.4%)
Usually	1,024 (22.3%)	1,549 (29.8%)
Getting up in morning		
Not at all easy	139 (3.1%)	249 (4.7%)
Not very easy	538 (11.9%)	830 (15.8%)
Fairly easy	2,327 (51.4%)	2,663 (50.9%)
Very easy	1,526 (33.7%)	1,505 (28.6%)
Nap during day		
Never/rarely	2,497 (54.5%)	2,238 (62.3%)
Sometimes	1,774 (38.8%)	1,798 (34.7%)
Usually	307 (6.7%)	228 (4.3%)
Frequency of tiredness/lethargy in last 2 weeks		
Not at all	2,402 (53.0%)	2,489 (47.8%)
Several days	1,770 (39.0%)	2,127 (40.9%)
More than half the days	187 (4.1%)	300 (5.8%)
Nearly everyday	177 (3.9%)	287 (5.5%)
Alcohol drinker status		
Never	81 (1.8%)	179 (3.4%)
Previous	83 (1.8%)	146 (2.7%)
Current	4,429 (96.4%)	4,992 (93.9%)

Variable	Cluster 1	Cluster 2
Sex		
Female	1,134 (24.7%)	4,062 (76.4%)
Male	3,461 (75.3%)	1,257 (23.6%)
...
Nervous feelings		
Yes	751 (16.6%)	1,071 (20.8%)
No	3,763 (83.4%)	4,076 (79.2%)
...
Frequency of tiredness/lethargy in last 2 weeks		
Not at all	2,402 (53.0%)	2,489 (47.8%)
Several days	1,770 (39.0%)	2,127 (40.9%)
More than half the days	187 (4.1%)	300 (5.8%)
Nearly everyday	177 (3.9%)	287 (5.5%)
Alcohol drinker status		
Never	81 (1.8%)	179 (3.4%)
Previous	83 (1.8%)	146 (2.7%)
Current	4,429 (96.4%)	4,992 (93.9%)



Case-Studies – UK Biobank – Results



Decision tree illustrating a simple clinical decision support system providing machine guidance for identifying depression feelings based on categorical variables and neuroimaging biomarkers. In each terminal node, the y vector includes the percentage of subjects being labeled as “no” and “yes”, in this case, answering the question “Ever depressed for a whole week.” The p-values listed at branching nodes indicate the significance of the corresponding splitting criterion.



Case-Studies – UK Biobank – Results

	Accuracy	95% CI (Accuracy)	Sensitivity	Specificity
Sensitivity/hurt feelings	0.700	(0.676, 0.724)	0.657	0.740
Ever depressed for a whole week	0.782	(0.760, 0.803)	0.938	0.618
Worrier/anxious feelings	0.730	(0.706, 0.753)	0.721	0.739
Miserableness	0.739	(0.715, 0.762)	0.863	0.548

Cross-validated (random forest) prediction results for four types of mental disorders

Zhou, et al. (2018), in review



Acknowledgments

Slides Online:
"SOCR News"

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760, UL1TR002240
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

<http://SOCR.umich.edu>

Collaborators

- **SOCR:** Milen Velez, Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang, Simeone Marino
- **LONI/INI:** Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Young Sung, Carl Kesselman, Fabio Maciardi, Federica Torri
- **UMich MIDAS/MNORC/AD/PD Centers:** Cathie Spino, Chuck Burant, Ben Hampstead, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, Brian Athey

