High-Dimensional Biomedical Data & Predictive Health Analytics

Ivo D. Dinov

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

University of Michigan

http://SOCR.umich.edu http://Predictive.Space

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

Slides Online:
"SOCR News"

Outline

- Common characteristics of Big Biomed/Health Data
- Data science & predictive health analytics
- Compressive Big Data Analytics (CBDA)
- Case-studies
  - Applications to Neurodegenerative Disease (ADNI)
  - Population Census-like Neuroscience (UKBB)

Characteristics of Big Biomed Data

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

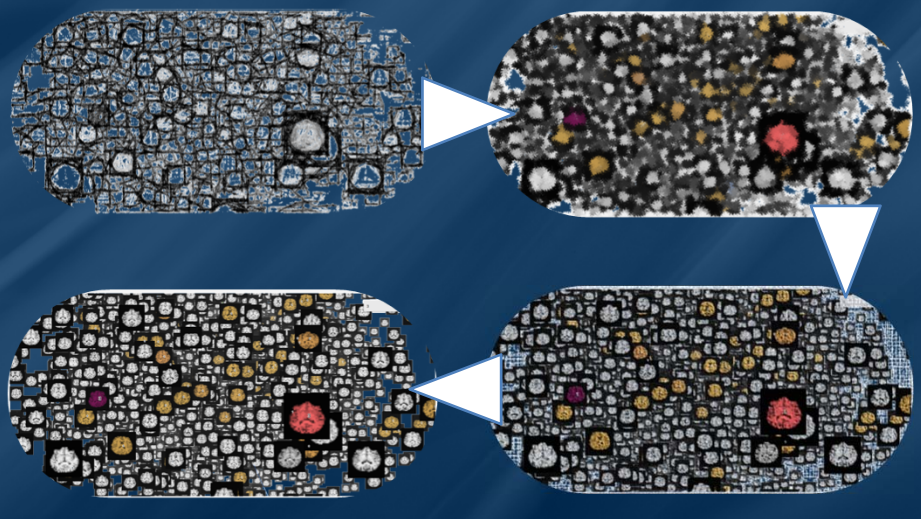
Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale 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

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)

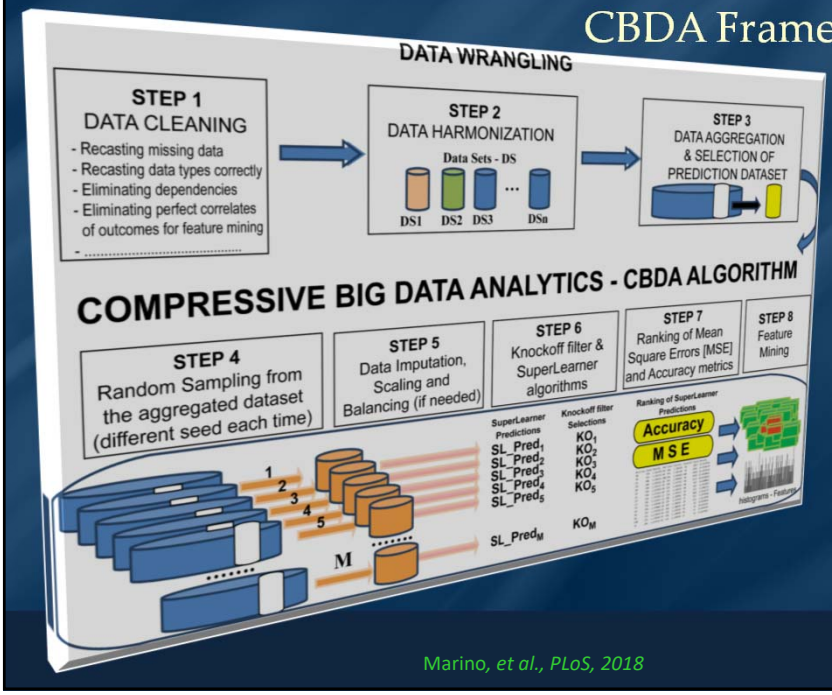
http://DSPA.predictive.space Dinov, Springer (2018)

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)

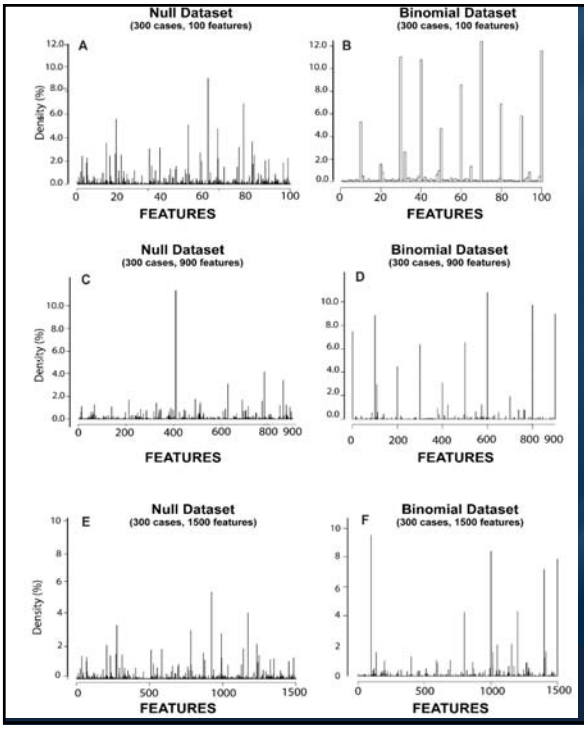
Dinov, J Med Stat Inform, *2016;* Marino, *et al., PLoS, 2018*

CBDA Framework



Marino, *et al., PLoS, 2018*

Controlled Feature Selection Results: Simulated Data

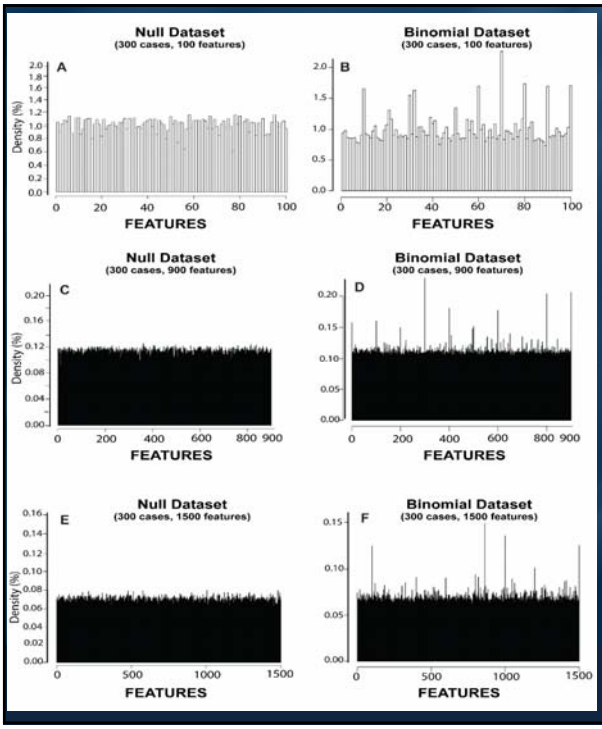


<u>Knockoff</u> of Null (left) vs Binomial (right) Data

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: Simulated Data

CBDA Results: Null (left) vs Binomial (right) Data

Panels A, C and E show the correspondent histograms of the CBDA Results on the three Null datasets.

Panels B, D and F show the correspondent histograms of the CBDA results 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)

Marino, *et al., PLoS, 2018*

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- degrees
101	22681	Carotid ultrasound	Maximum carotid IMT (intima- degrees
101	22671	Carotid ultrasound	Mean carotid IMT (intima-med
101	22674	Carotid ultrasound	Mean carotid IMT (intima-med
101	22677	Carotid ultrasound	Mean carotid IMT (intima-med
101	22680	Carotid ultrasound	Mean carotid IMT (intima-med
101	22670	Carotid ultrasound	Minimum carotid IMT (intima- degrees
101	22673	Carotid ultrasound	Minimum carotid IMT (intima- degrees
101	22676	Carotid ultrasound	Minimum carotid IMT (intima-medial thickness) at 210 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

- 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

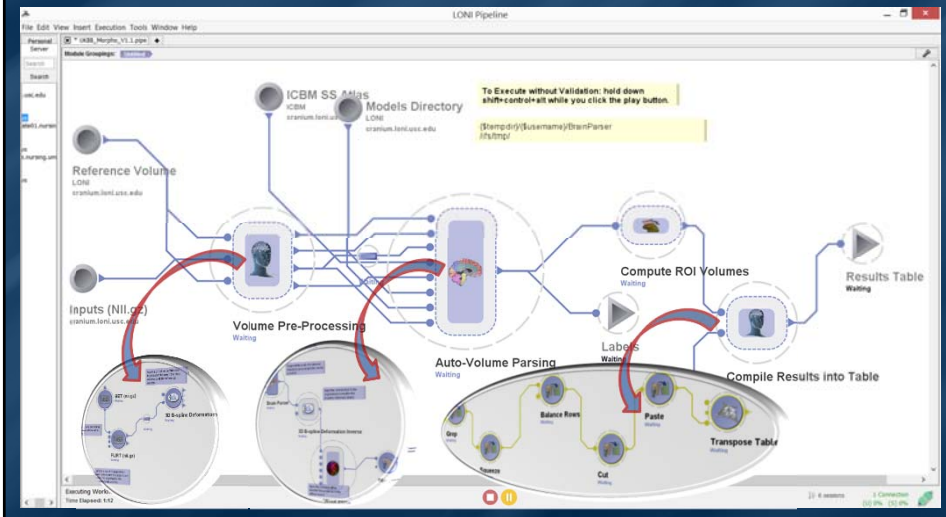
Case-Studies – UK Biobank (Complexities)

Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1,500 derived neuroimaging biomarkers.

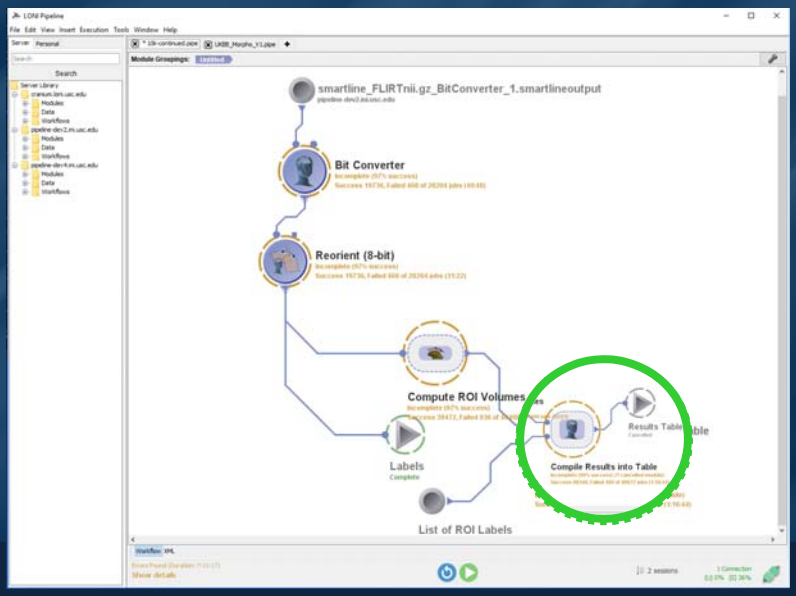
Including only features observed >30%
$(9{,}914 \times 1{,}475)$

Zhou, et al. (2018), in review

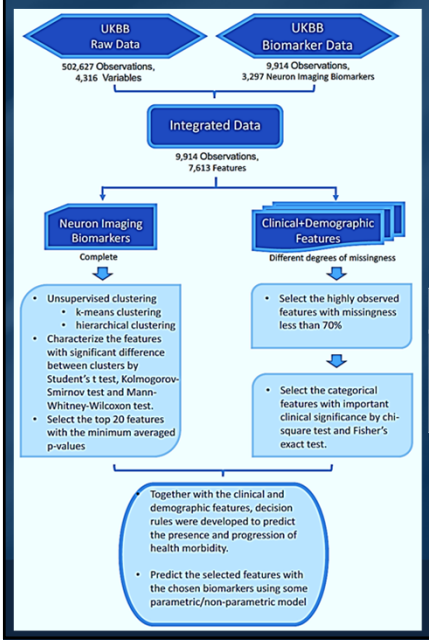
Case-Studies – UK Biobank – NI Biomarkers

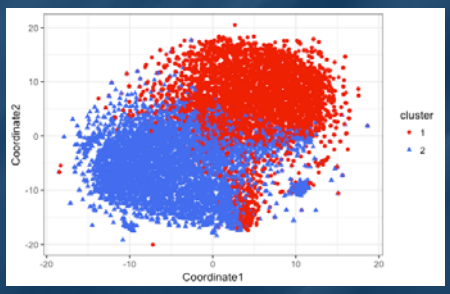


Case-Studies – UK Biobank – Successes/Failures



Case-Studies – UK Biobank – Results





t-SNE plot of the brain neuroimaging biomarkers

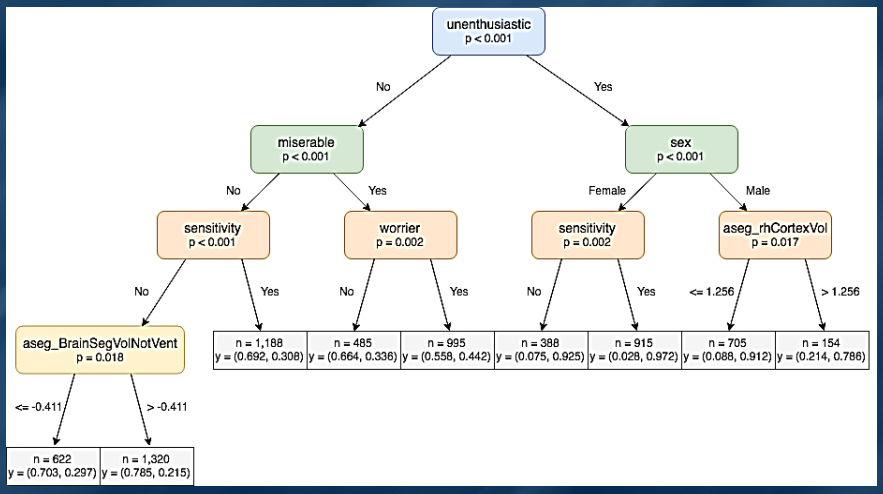
k-means clustering			
		Cluster 1	Cluster 2
Hierarchical clustering	Cluster 1	3768 (38.0%)	528 (5.3%)
	Cluster 2	827 (8.3%)	4791 (48.3%)

Cluster	Consistency	Variance	Cluster-size	Silhouette
1	0.997	0.001	5344	0.09
2	0.934	0.001	4570	0.05

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%)
. . .	. . .	
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	2,402 (53.0%)	2,489 (47.8%)
Not at all	1,770 (39.0%)	2,127 (40.9%)
Several days	187 (4.1%1)	300 (5.8%)
More than half the days	177 (3.9%)	287 (5.5%)
Nearly everyday		
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 Velev, 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 Macciardi, 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