Big Brain Data Science and Predictive Health Analytics

Ivo D. Dinov

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

University of Michigan

www.SOCR.umich.edu

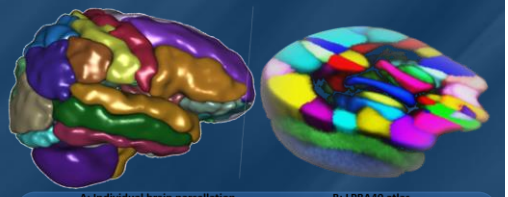
SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Outline

- Driving biomedical & health challenges
- Common characteristics of Big Brain Data
- Data science & predictive analytics
- Case-studies
  - Applications to Neurodegenerative Disease
  - Data Dashboarding
- Compressive Big Data Analytics (CBDA)

Driving Biomedical/Health Challenges

- **Neurodegeneration**: Structural Neuroimaging in Alzheimer's Disease illustrates the Big Data challenges in modeling complex neuroscientific data. 808 ADNI subjects, 3 groups: 200 subjects with Alzheimer's disease (AD), 383 subjects with mild cognitive impairment (MCI), and 225 asymptomatic normal controls (NC). The 80 neuroimaging biomarkers and 80 highly-associated SNPs.



A: Individual brain parcellation

Index	Volume Intensity	ROI Name
1	21	L superior frontal gyrus
2	24	R middle frontal gyrus
3	50	R precuneus
4	181	cerebellum
5	47	L angular gyrus
6	122	R cingulate gyrus
7	83	L middle temporal gyrus
8	90	R lingual gyrus
9	81	L superior temporal gyrus
10	91	L fusiform gyrus
11	44	R superior parietal gyrus
12	66	R inferior occipital gyrus
13	87	L parahippocampal gyrus
14	162	R caudate
15	85	L inferior temporal gyrus
16	182	brainstem
17	43	L superior parietal gyrus
18	28	R precentral gyrus
19	23	L middle frontal gyrus
20	89	L lingual gyrus
21	41	L postcentral gyrus

B: LPBA40 atlas

Index	Volume Intensity	ROI Name
29	65	L inferior occipital gyrus
30	164	R putamen
31	61	L superior occipital gyrus
32	30	R middle orbitofrontal gyrus
33	42	R postcentral gyrus
34	27	L precentral gyrus
35	32	R lateral orbitofrontal gyrus
36	121	L cingulate gyrus
37	31	L lateral orbitofrontal gyrus
38	92	R fusiform gyrus
39	45	L supramarginal gyrus
40	88	R parahippocampal gyrus
41	22	R superior frontal gyrus
42	29	L middle orbitofrontal gyrus
43	48	R cuneus
44	62	R superior occipital gyrus
45	33	L gyrus rectus
46	48	R angular gyrus
47	64	R middle occipital gyrus
48	84	R middle temporal gyrus
49	49	L precuneus

http://DSPA.predictive.space

Moon, Dinov, *et al.* (2015)

Driving biomedical/health challenges

- Phenotype-Genotype-Environmental

Chance ↑

Phenotype:

Outward manifestations of the form, shape, or characteristics of specific individuals, cohorts, or morbid conditions

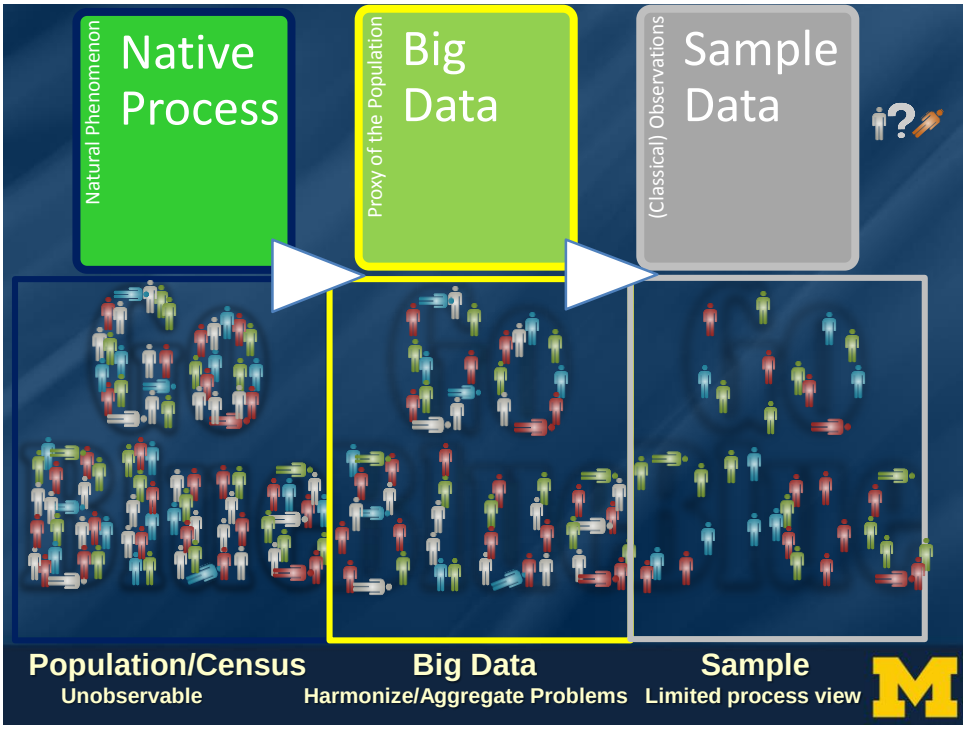
Nature/Genome:

Read the genome →
protein synthesis →
2 basic functions →
structural proteins determine *physical traits* or *functional traits* of protein enzymes catalyzing chemical reactions

Nurture/Environment:

Ambient environment directly effects Pheno + Geno →
Provides raw materials needed for the synthetic processes controlled by Nature/Genes →
Organisms either synthesize of obtain amino acids with their diet

↓ $SNR = \frac{Signal}{Noise}$



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
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

Data science & predictive analytics

- **Data science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and biosocial 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 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)

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	Forecasts, Predictions
Semantic-Mapping	Derived Biomarkers	Linkages, Associations	Healthcare Outcomes

Dinov, et al. (2016) PMID:26918190

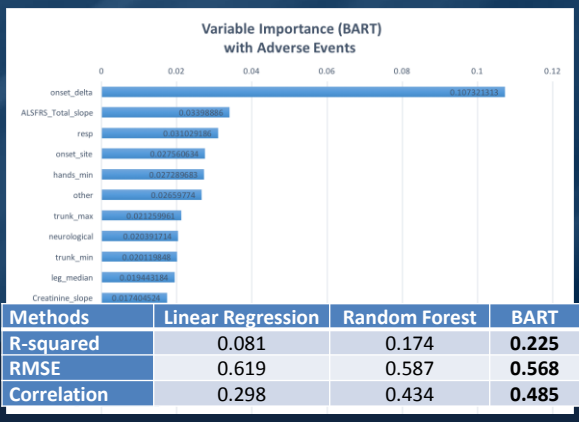
Case-Studies – ALS

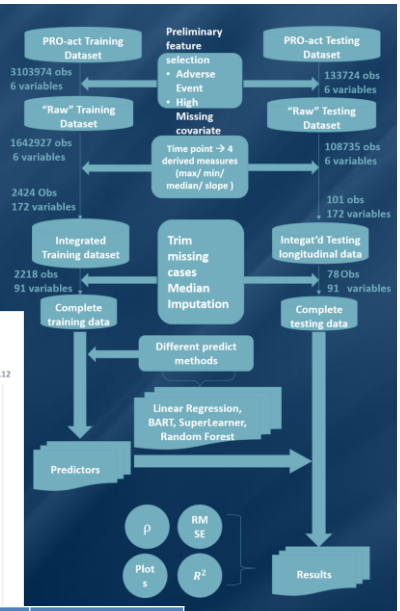
- Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: Amyotrophic Lateral Sclerosis 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)

Case-Studies – ALS

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





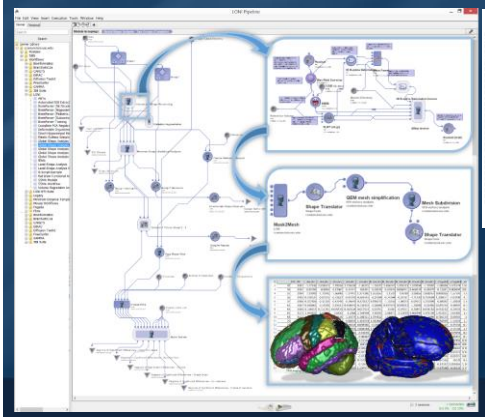
Methods	Linear Regression	Random Forest	BART	SuperLearner
R-squared	0.081	0.174	**0.225**	0.178
RMSE	0.619	0.587	**0.568**	0.585
Correlation	0.298	0.434	**0.485**	0.447

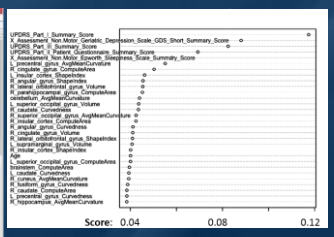
Case-Studies – Parkinson's Disease

- Predict the clinical diagnosis of patients using all available data (with and without the UPDRS clinical assessment, which is the basis of the clinical diagnosis by a physician)
- Compute derived neuroimaging and genetics biomarkers that can be used to model the disease progression and provide automated clinical decisions support
- Generate decision trees for numeric and categorical responses (representing clinically relevant outcome variables) that can be used to suggest an appropriate course of treatment for specific clinical phenotypes

Data Source	Sample Size/Data Type	Summary
PPMI Archive	Demographics: age, medical history, sex. Clinical data: physical, verbal learning and language, neurological and olfactory, UPSIT, UPDRS scores, ADL, GDS-15, … Imaging data: structural MRI. Genetics data: APOE genotypes e2/e3 Cohorts: Group 1 = {PD Subjects}, N_1 = 263; Group 2 = {PD Subjects with Scans without Evidence of a Dopaminergic Deficit (SWEDD)}, N_2 = 40; Group 3 = {Control Subjects}, N_3 = 127.	The longitudinal PPMI dataset including clinical, biological and imaging data (screening, baseline, 12, 24, and 48 month follow-ups) may be used conduct model-based predictions as well as model-free classification and forecasting analyses

Predictive Big Data Analytics: Applications to Parkinson's Disease





Varplot

- Critical predictive data elements (Y-axis)
- Their impact scores (X-axis)

AdaBoost classifier for Controls vs. Patients prediction

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

Dinov, et al., (2016) PMID:27494614

Open-Science & Collaborative Validation

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

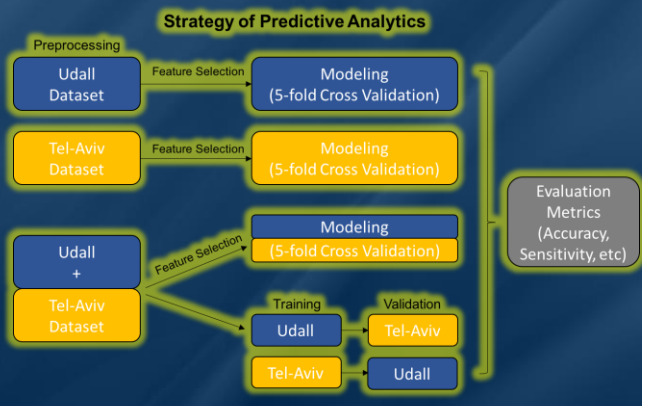
- Methods for rebalancing of imbalanced cohorts
- ML classification methods generating consistent and powerful phenotypic predictions
- Reproducible protocols for extraction of derived neuroimaging and genomics biomarkers for diagnostic forecasting

https://github.com/SOCR/PBDA

PD Patient Falls – Predictive BD Analytics

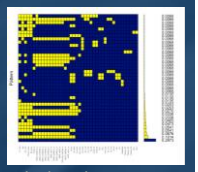
- Goals
  1) Harmonize and aggregate complex, multisource, and multi-site PD data
  2) Identify highly predictive features associated with specific clinical traits, e.g., falls
  3) Forecast falls using ML techniques and validate using statistical methods

- Study Design

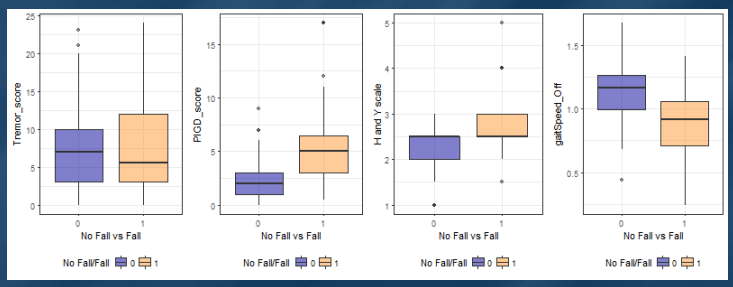


PD Patient Falls – Predictive BD Analytics

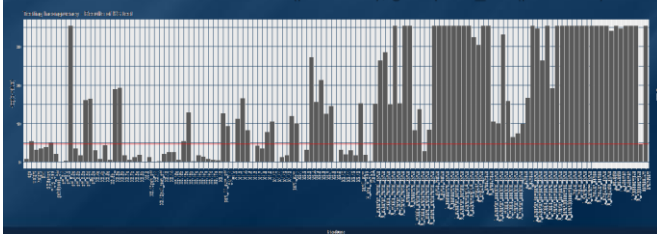
Results



Missing data pattern



Udall data differences in MDS_TREM (p = 0.5465), MDS_PIGD (p <0.001), H and Y scale (p < 0.001), gaitSpeed_Off (p < 0.001) between patients with/without falls



Udall vs. Tel-Aviv data

KS tests on 126 features comparing the distributions. Red horizontal line is the cutoff of $-log(\alpha)$, where α (desired FDR) = 0.01.

PD Patient Falls – Predictive BD Analytics

Results

Feature selection for the Udall data using RF and KO.

7 common features selected by both methods: MDS_PIGD, gaitSpeed_Off, MOT_EDL, NON_MOTOR_EDL, walk, pos_stab

Binary classification of falls/no-fall (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)

Random Forests (RF)		Knockoff (KO)	
Features	Frequency	Features	Frequency
MDS_PIGD	0.888	hx_smoke	0.764
gaitSpeed_Off	0.860	high_bp	0.751
R_middle_temporal_gyrus	0.662	walk	0.718
R_inferior_temporal_gyrus	0.618	MDS_PIGD	0.672
Caudate_DA	0.554	SLEEP_APNEA	0.602
Striatum_DA	0.534	head_inj	0.598
MOT_EDL	0.516	SLEEP_RBD	0.552
time_upgo	0.494	out_bed	0.515
L_middle_temporal_gyrus	0.436	gaitSpeed_Off	0.502
NON_MOTOR_EDL	0.418	HY	0.477
UPSIT40	0.410	NON_MOTOR_EDL	0.440
Putamen_DA	0.408	hal_psy	0.415
R_middle_orbitofrontal_gyrus	0.364	chair	0.415
	354	pos_stab	0.407
	336	Caudate_DA	0.403
	324	MOT_EDL	0.398
	322	gait	0.374
	320	gender	0.361
	318	turn	0.361
	318	depression	0.324

Method	acc	sens	spec	ppv	npv	lor
Logit	0.728	0.537	0.855	0.710	0.736	1.920
Random Forests	**0.796**	**0.683**	**0.871**	**0.778**	**0.806**	**2.677**
AdaBoost	0.689	0.610	0.742	0.610	0.742	1.502
XGBoost	0.699	0.707	0.694	0.604	0.782	1.699
SVM	0.709	0.561	0.806	0.657	0.735	1.672
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

Case-Studies – General Populations

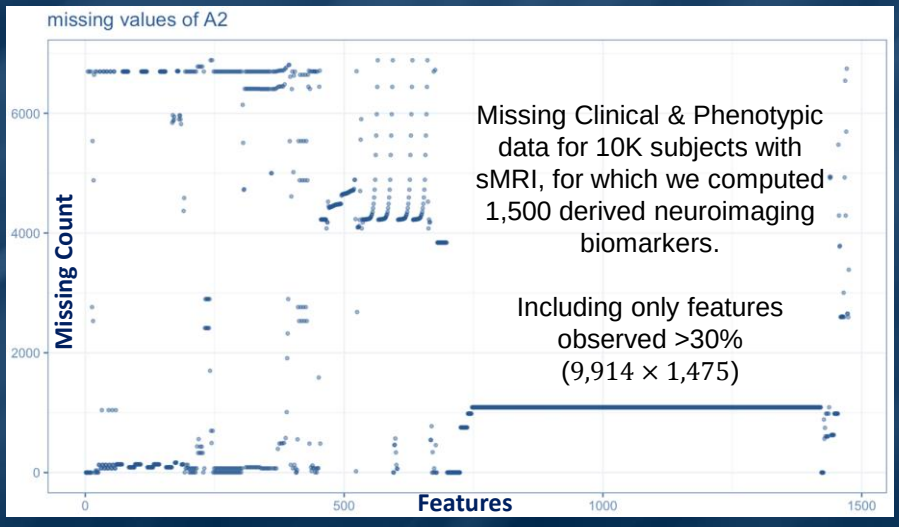
2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain 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
101	22684	Carotid ultrasound	Quality control indicator for IMT at 210 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

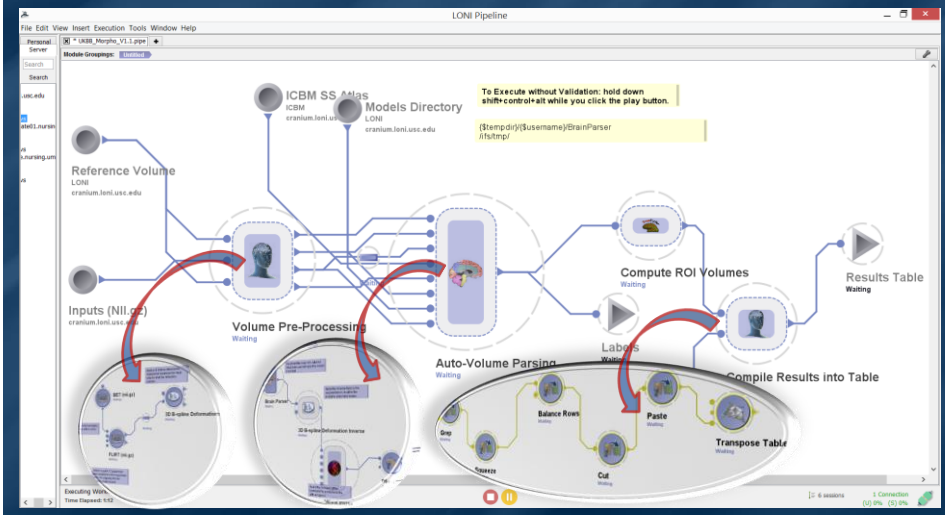
Case-Studies – UK Biobank (Complexities)



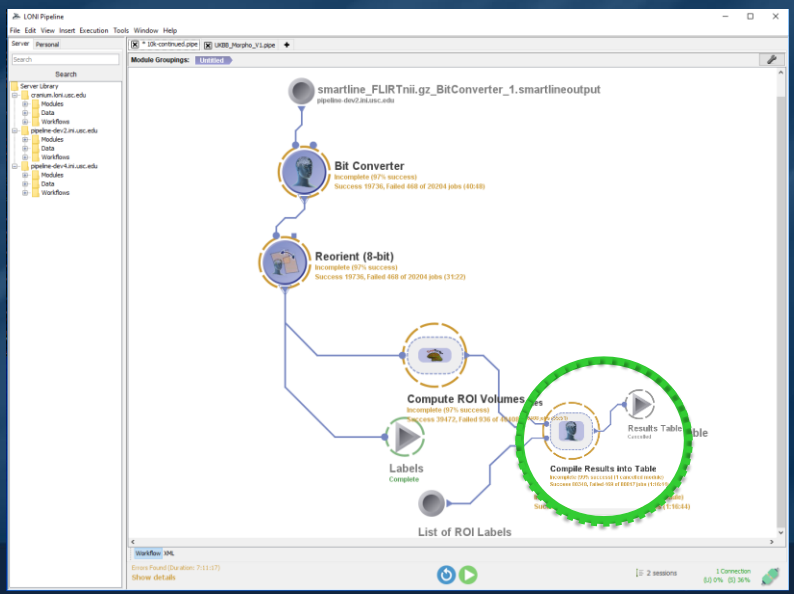
Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1,500 derived neuroimaging biomarkers.

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

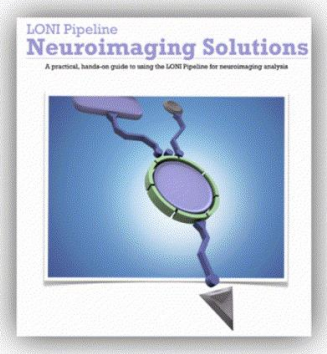
Case-Studies – UK Biobank – NI Biomarkers



Case-Studies – UK Biobank – Successes/Failures

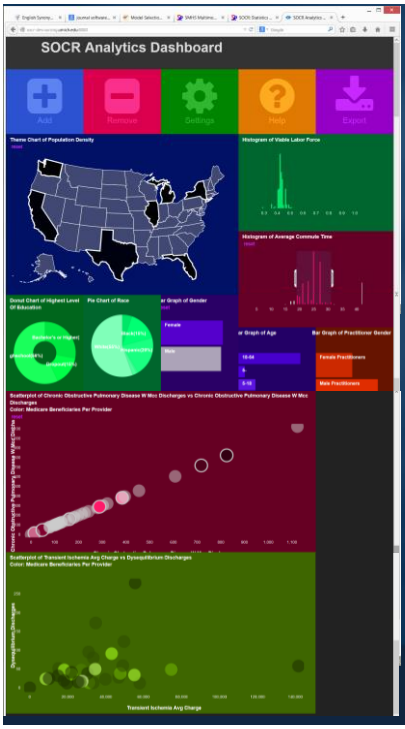


End-to-end Pipeline Workflow Solutions



Dinov, *et al*., 2014, Front. Neuroinform.;

Dinov, *et al*., Brain Imaging & Behavior, 2013



SOCR Big Data Dashboard

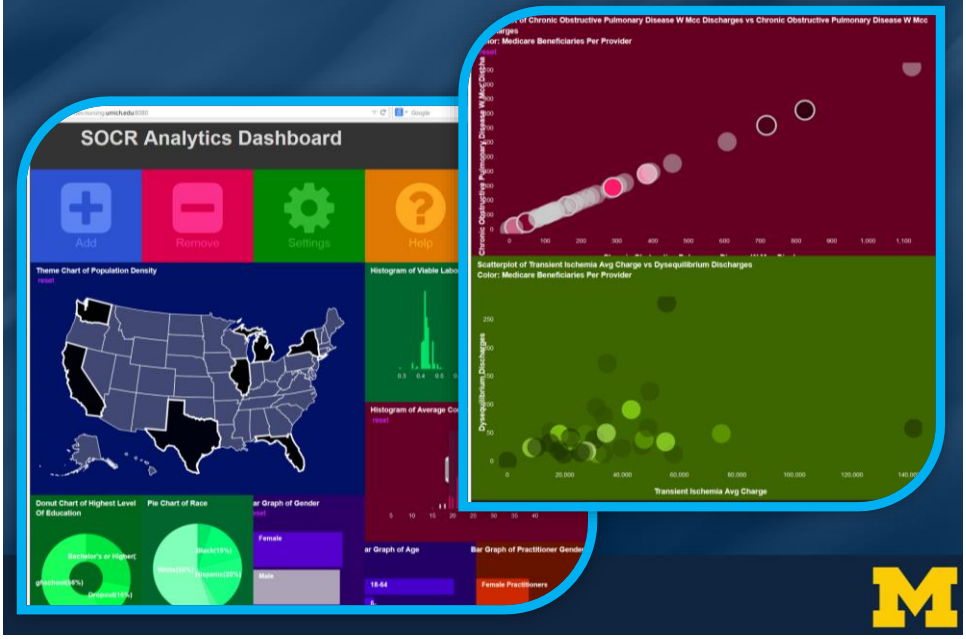
http://socr.umich.edu/HTML5/Dashboard

- Web-service combining and integrating multi-source socioeconomic and medical datasets
- Big data analytic processing
- Interface for exploratory navigation, manipulation and visualization
- Adding/removing of visual queries and interactive exploration of multivariate associations
- Powerful HTML5 technology enabling mobile on-demand computing

Husain, et al., 2015, PMID:26236573

SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

SOCR Dashboard (Exploratory Big Data Analytics): Associations

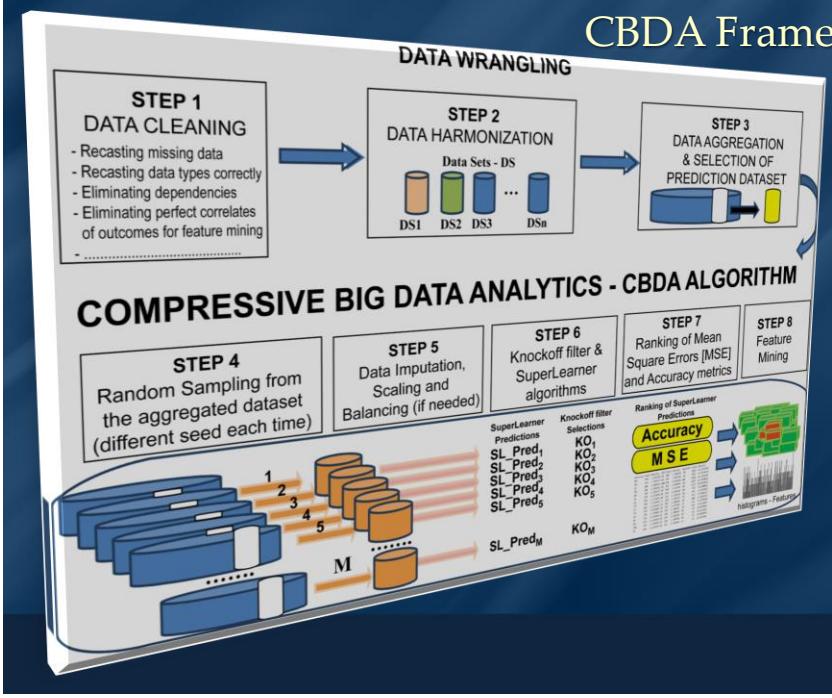


Compressive Big Data Analytics (CBDA)

- 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 or non-parametric inference based on the sample
  - Next, compute likelihood estimates (e.g., probability values quantifying effects, relations, sizes)
  - Repeat – the process continues iteratively until a 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, *2016, PMID: 26998309*

CBDA Framework

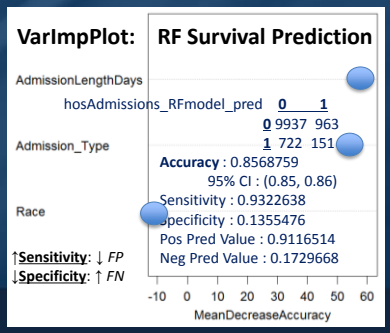


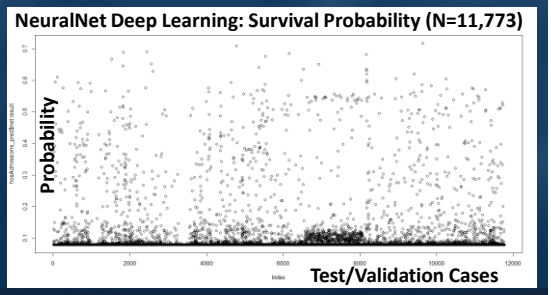
FAIR Data & Open-Science Principles

- Share resources
- Collaborate
- Permissive licenses (e.g., LGPL/CC-BY)
- Project management (e.g., GitHub/Jira)
- Open-access pubs
- Public-private partnerships
- Co-mentoring of trainees
- Effective transdisciplinary methods
- Resource Interoperability
- Result Reproducibility

Clinical Decision Support

- Hospital Admissions
  - Survival Inference and Clinical outcome forecasting using hospital admissions data (N~60K and k=9):
    - Admission_Length: Duration of hospital stay (in days)
    - Death: Indicator of Death (1) or survival (0)
    - …
    - Demographics & (Human-labeled) Diagnoses

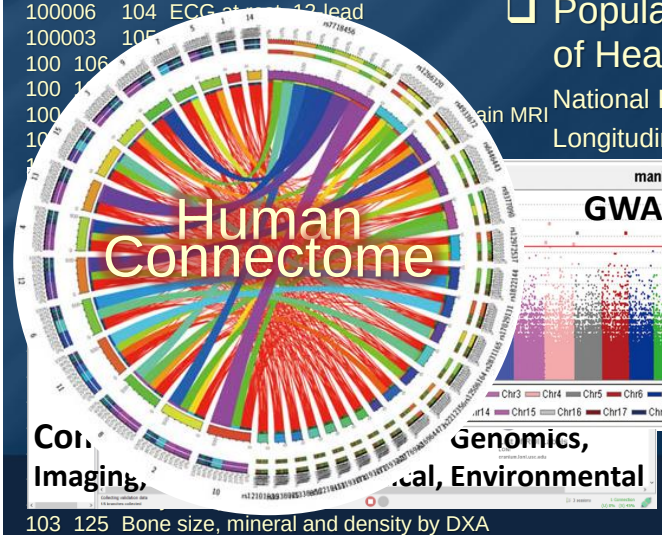


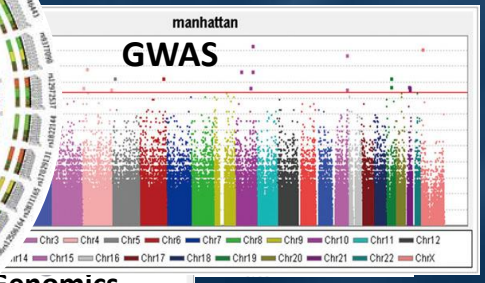


Clinical Decision Support

Top 1 Population characteristics
1 2 Ongoing characteristics
100003 100 Brain MRI
100006 101 Carotid ultrasound
100003 102 Heart MRI
100003 103 DXA assessment
100006 104 ECG at rest 12-lead
100003 105
100 106
100 1
100
1
1

- Population-wide Study of Health and Disease
  - National Health Service (NHS)
  - Longitudinal Data (N>1M, k>4,300)



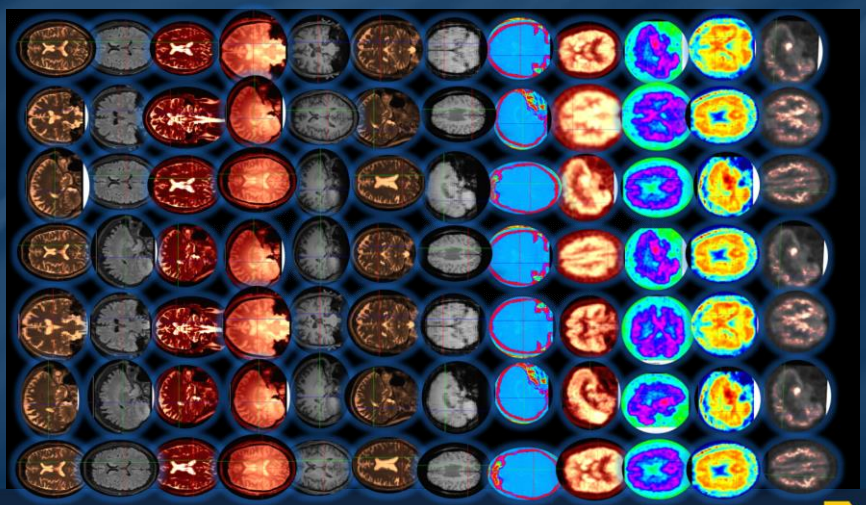


Genomics,
Imaging, cal, Environmental

103 125 Bone size, mineral and density by DXA

Clinical Decision Support

- Personalized medicine – Traumatic Brain Injury (TBI)



LONI/USC, BIRC/UCLA, SOCR/UMich

Acknowledgments

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760

NSF: 1734853, 1636840, 1416953, 0716055, 1023115

The Elsie Andresen Fiske Research Fund

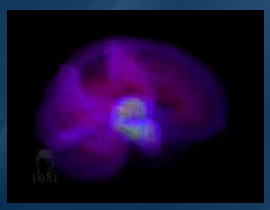
http://SOCR.umich.edu

Collaborators

- **SOCR**: 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
- **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

Demo(s)?

- Complex DB Search, retrieval (IDA)
- Multidimensional data visualization (MotionCharts, BrainViewer, R)
- Distributed high-throughput pipeline workflow computing
- SOCRAT Framework
- Data Dashboard
- Education and Training Resources
  - Probability and Statistics Ebook (EBook)
  - Scientific Methods for Health Sciences (SMHS)
  - Data Science and Predictive Analytics (DSPA) MOOC
  - SOCR Tools (distribution calculators, charts, modeler, analyses, experiments)
- Compressive Big Data Analytics (CBDA)



Demo