Predictive Data Analytics

Ivo D. Dinov

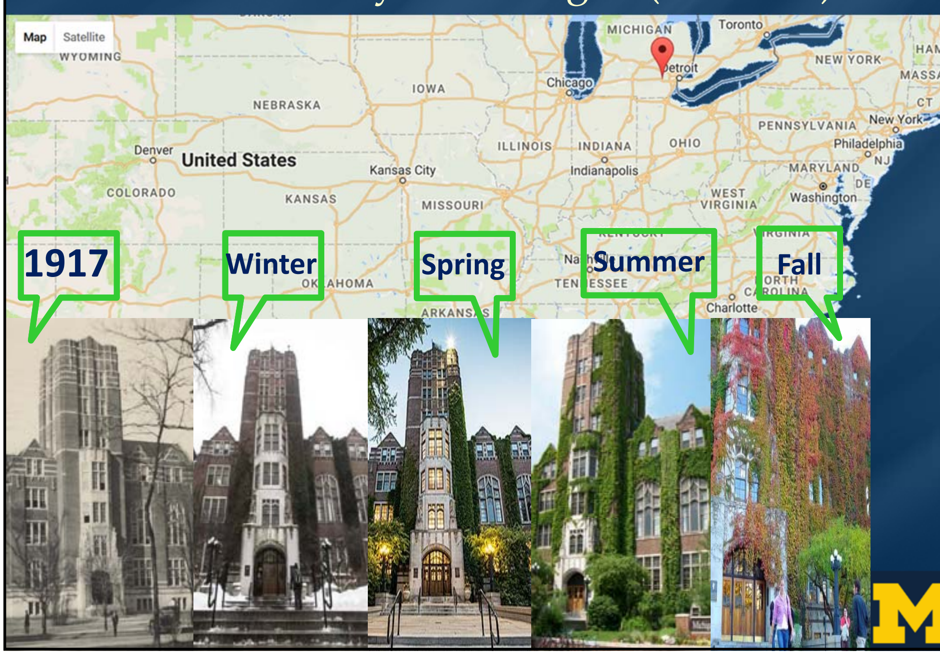
Statistics Online Computational Resource
Health Behavior and Biological Sciences
Michigan Institute for Data Science

University of Michigan

www.SOCR.umich.edu

STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)

The University of Michigan (est. 1817)



Outline

- Driving biomedical & health challenges
- Common characteristics of Big Data
- Data science & predictive analytics
- Case-studies
- Applications to Neurodegenerative Disease
- Data Dashboarding
- Compressive Big Data Analytics (CBDA)
- Tomorrow's Healthcare: The Age of Disruptions
- Demo(s)

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.

Index	Volume Intensity	ROI Name	Index	Volume Intensity	ROI Name
1	21	L superior frontal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	30	164	R putamen
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	32	30	R middle orbitofrontal gyrus
5	47	L angular gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	77	L precentral gyrus
7	63	L middle temporal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	L superior temporal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	44	R superior parietal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	40	88	R parahippocampal gyrus
[illegible]	[illegible]	L parahippocampal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	44	62	R superior occipital gyrus
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	23	L middle frontal gyrus	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	41	L postcentral gyrus	[illegible]	[illegible]	[illegible]

http://DSPA.predictive.space

Driving biomedical/health challenges

- Phenotype-Genotype-Environmental

Phenotype:

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

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

Chance

$SNR = \frac{Signal}{Noise}$

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 generating semi-automated decision support systems
- **Predictive analytics**: utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools and services to represent, interrogate and interpret complex data. Aims to forecast trends, predict 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

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 highly 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 will be used in out processing, modeling and analysis	The time points for all longitudinally varying data elements will be aggregated into signature vectors. This will facilitate the modeling and prediction of ALSFRS slope changes over the first three months (baseline to month 3)

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}, N1 = 263; Group 2 = {PD Subjects with Scans without Evidence of a Dopaminergic Deficit (SWEDD)}, N2 = 40; Group 3 = {Control Subjects}, N3 = 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

Case-Studies – General Populations

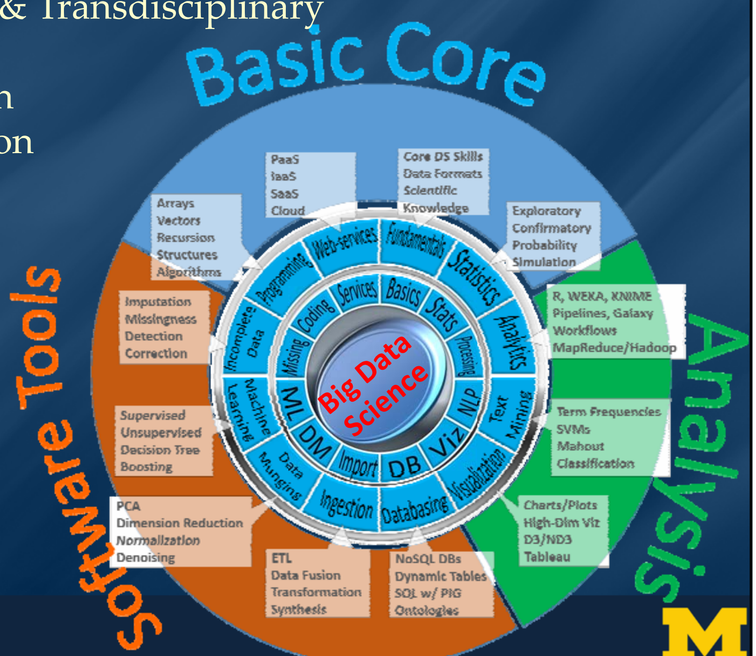
2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain 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

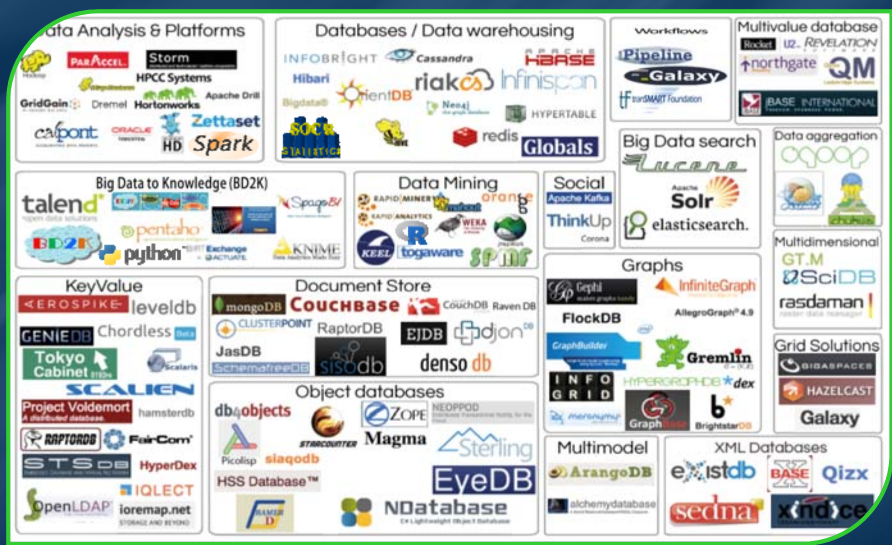
Data Source	Sample Size/Data Type	Summary
UK Biobank	**Demographics**: > 500K cases **Clinical data:** > 4K features **Imaging data**: T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI **Genetics data**	The longitudinal archive of entire UK population (NHS)

Integrated & Transdisciplinary

- Research
- Education
- Practice

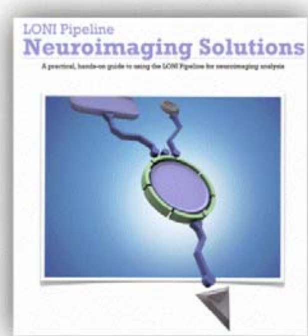


Big Data Analytics Resourceome



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

End-to-end Pipeline Workflow Solutions



Dinov, *et al.*, 2014, Front. Neuroinform.; Dinov, *et al.*, Brain Imaging & Behavior, 2013

Predictive Big Data Analytics in Parkinson's Disease

- Big Data: Parkinson's Progression Markers Initiative (PPMI). Defining data characteristics – large size, incongruency, incompleteness, complexity, multiplicity of scales, and heterogeneity of sources (imaging, genetics, clinical, demographic)
- Approach – machine-learning based classification
  - introduce methods for rebalancing imbalanced cohorts,
  - utilize a wide spectrum of classification methods to generate phenotypic predictions,
  - reproducible machine-learning based classification

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

- Results
  - Predicted Parkinson's disease in the PPMI subjects (consistent accuracy, sensitivity, and specificity exceeding 96%
  - Confirmed using internal statistical 5-fold cross-validation
  - Clinical features: Unified Parkinson's Disease Rating Scale (UPDRS) scores demographic (e.g., age), genetics (e.g., rs34637584, chr12)
  - Neuroimaging biomarkers (e.g., cerebellum shape index)
- Model-free Big Data machine learning-based classification methods (Adaptive boosting, support vector machines) outperform model-based techniques (GEE, GLM, MEM) in terms of predictive precision and reliability (e.g., forecasting patient diagnosis).
- UPDRS scores play a critical role in predicting diagnosis, which is expected based on the clinical definition of Parkinson's disease.
- Excluding longitudinal UPDRS data, the accuracy of model-free machine learning based classification is over 80%. The methods, software and protocols are openly shared and can be employed to study other neurodegenerative disorders

Probabilistic Bayesian Inference – Chance Encounters

- Suppose a patient visits a primary care clinic and is seen by a male provider not wearing a badge/insignia
- To address the clinician appropriately, the patient is trying to figure out if he is more likely to be a doctor (D) or a nurse (N).
- Notation $F = Female$, $M = Male$, $D = Doctor$, and $N = Nurse$.
- Traditional stereotypes may suggest that a male provider is more likely to be a doctor than a nurse.

Is the odds likelihood ratio, $\frac{P(N|M)}{P(D|M)} < 1$?

Probabilistic Bayesian Inference – Chance Encounters

- Actually, the odds are that the *Male* healthcare provider is a *Nurse*!

Data: Kaiser Family Foundation

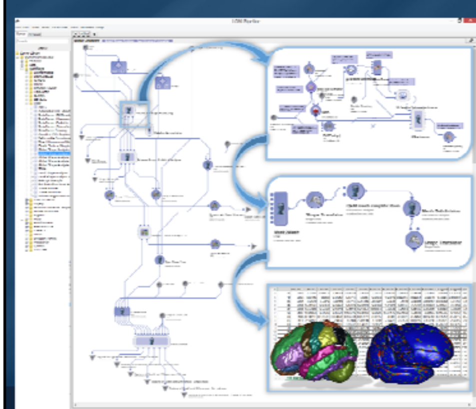
$$\underbrace{\frac{P(N|M)}{P(D|M)}}_{\text{odds likelihood ratio}} = \frac{\frac{P(M|N)\times P(N)}{P(M)}}{\frac{P(M|D)\times P(D)}{P(M)}} = \underbrace{\frac{P(M|N)}{P(M|D)}}_{\text{likelihood ratio}} \times \underbrace{\frac{P(N)}{P(D)}}_{\text{base rate}} =$$

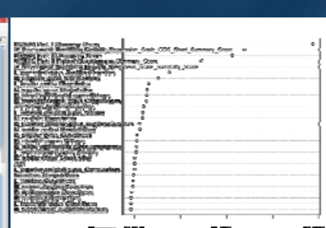
$$= \frac{\frac{1}{13}}{\frac{2}{3}} \times \frac{\frac{4{,}500{,}000}{US}}{\frac{435{,}000}{US}} = \frac{3}{26} \times 10.3 = 1.2.$$

- An odds likelihood ratio >1 dispels an initial stereotypic vision of females and males as predominantly nurses and physicians, respectively.
- This is a simple example of data-driven/evidence-based Inference.

Dinov, et al., (2016) PMID: 26998309

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

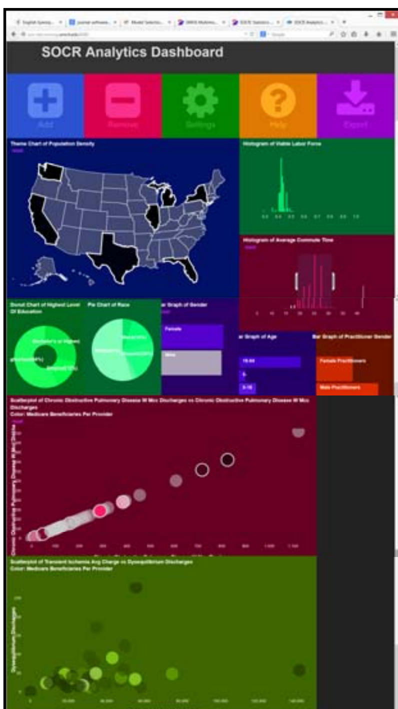
Dinov, et al., (2016) PMID:27494614

Tools Developed, Validated & Shared

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

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

https://github.com/SOCR/PBDA



SOCR Big Data Dashboard

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

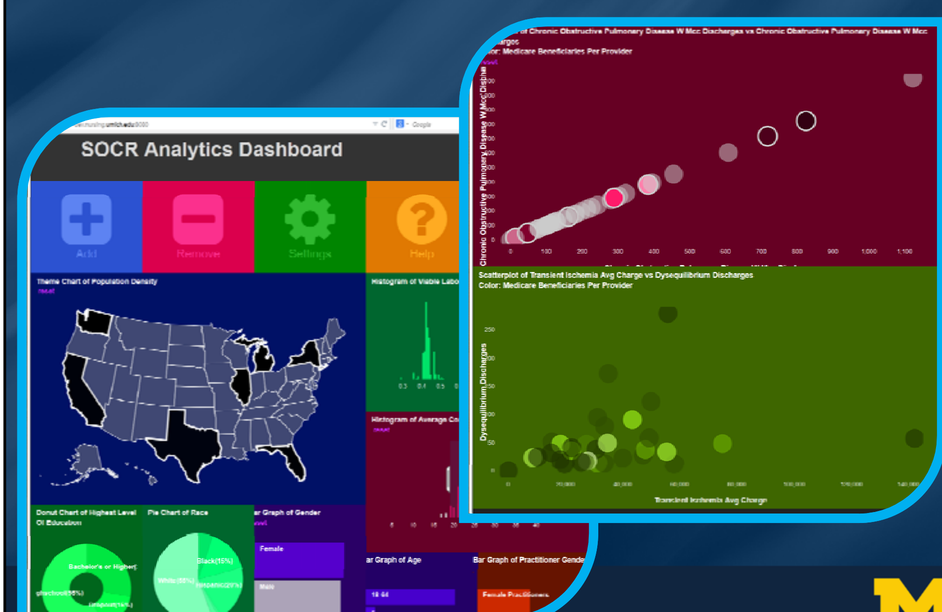
- ❑ Web-service combining and integrating multi-source socioeconomic and medical datasets
- ❑ Big data analytic processing
- ❑ Interface for exploratory navigation, manipulation and visualization
- ❑ Adding/removing of visual queries and interactive exploration of multivariate associations
- ❑ Powerful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

Husain, et al., 2015, PMID:26236573

SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

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

SOCR Dashboard (Exploratory Big Data Analytics): Associations

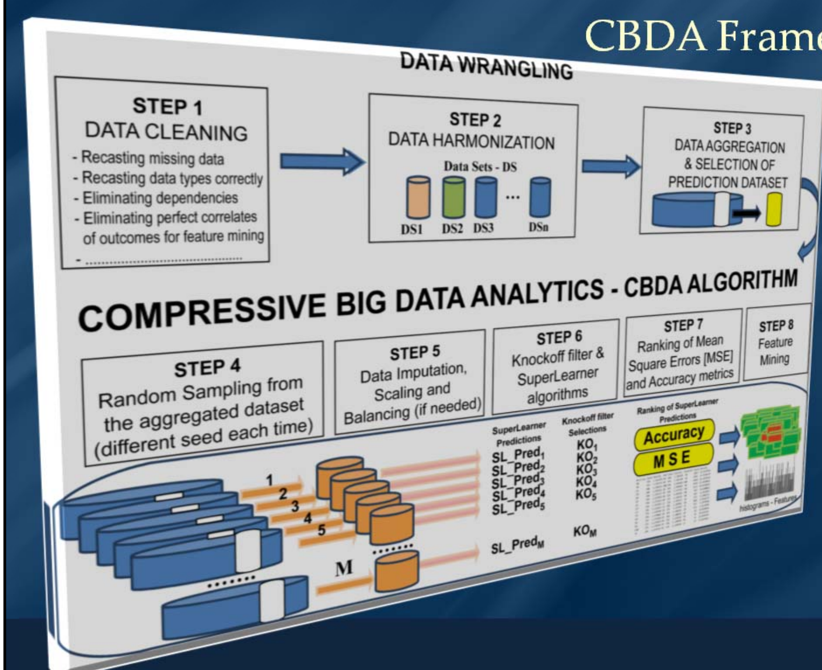


Compressive Big Data Analytics (CBDA)

- 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)

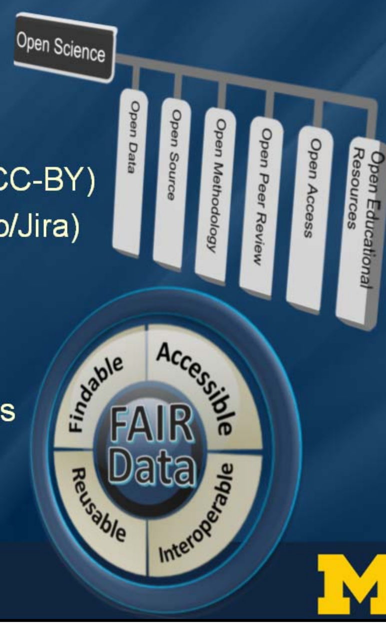
Dinov, *2016, PMID: 26998309*

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

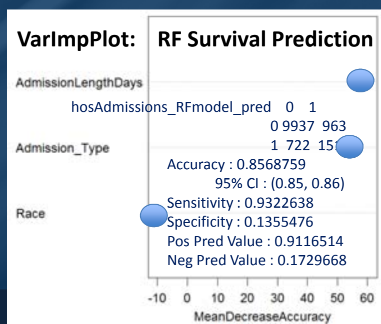


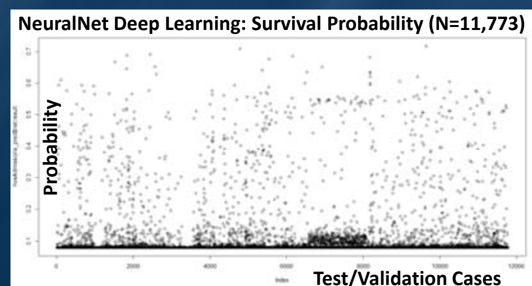
Tomorrow’s Healthcare: The Age of Disruptions

- ❑ Address Some Challenging Open Problems
  - ❑ Powerful data wrangling strategies
  - ❑ Techniques for data harmonization, appending, aggregation
  - ❑ Mathematical framework for Big Data representation (cf. 6D, CBDA)
  - ❑ Reliable and secure Biomed/Health data communication/sharing
  - ❑ Advanced machine-learning decision support systems
- ❑ Future Healthcare Innovation & Delivery
  - ❑ On-demand, service-oriented, geo-location-agnostic health delivery
  - ❑ Rapid deployment, continuous development/innovation/refinement
  - ❑ (Evidence-based) Data Science and Predictive Health Analytics
  - ❑ Personalized Medicine (from diagnosis, to treatment and prognosis)
  - ❑ End-to-end Doctronic (Human-Machine) services – Clinical Decision Support Systems (improve overall population health, reduce costs, better prognostication, enhanced reliability, rapid response)
- ❑ Some examples …

Clinical Decision Support Doctronics

- Hospital Admissions
  - Survival Inference and Clinical outcome forecasting using a hospital admissions (N=58,863 and k=9):
    - Admission_Length: Duration of hospital stay (in days)
    - Death: Indicator of Death (1) or survival (0)
    - …
    - Demographics & (Human-labeled) Diagnoses





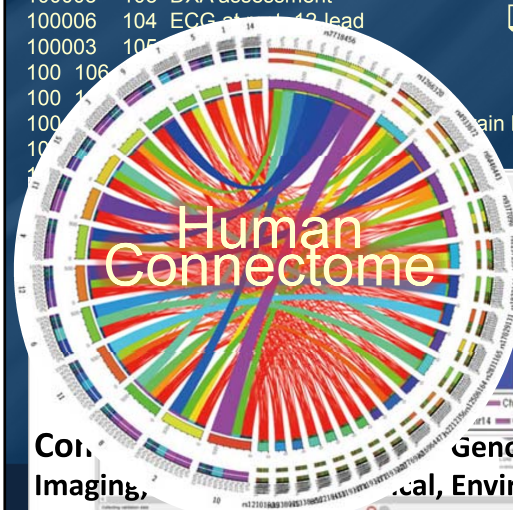
Clinical Decision Support Doctronics

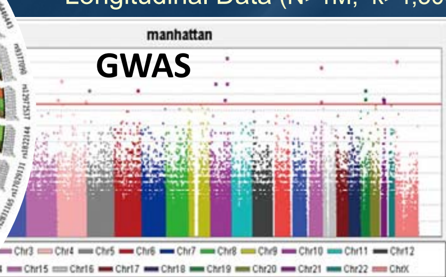
- Population-wide Study of Health and Disease

National Health System

Longitudinal Data (N>1M, k>4,300)

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 ... 12-lead
100003 105
100 106
100 1
100 ...ain MRI
10



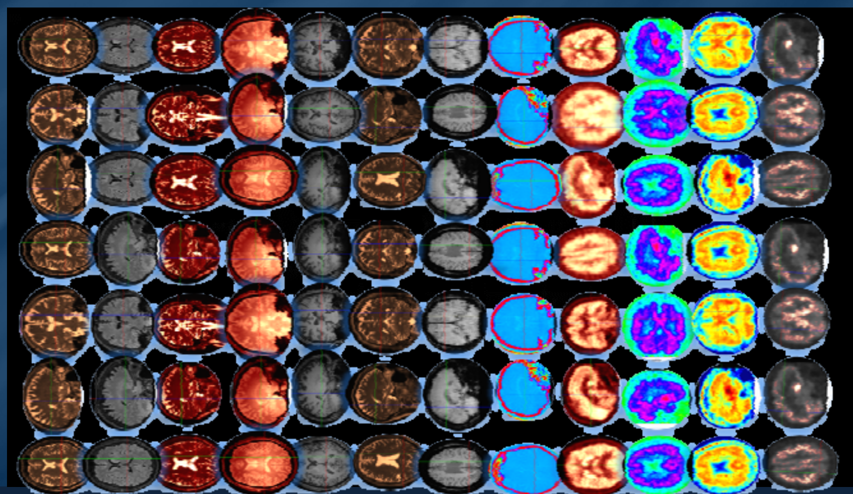


Co... Genomics,
Imaging, ...cal, Environmental

103 125 Bone size, mineral and density by DXA

Clinical Decision Support Doctronics

- ❑ 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: 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