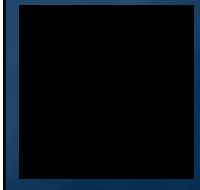
Big Brain Data & Predictive Analytics

Ivo D. Dinov

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

University of Michigan

www.SOCR.umich.edu



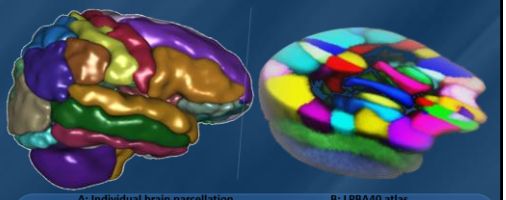
SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Outline

- AMIA Biomedical Imaging Working Group Driving Challenge:
  Does Integrative Data Analytics on Biomedical Imaging Bring Us Closer to Precision Medicine?
- Common characteristics of Big Brain Data
- Data science & predictive analytics
- Case-studies
- Applications to Neurodegenerative Disease
- Data Dashboarding
- Compressive Big Data Analytics (CBDA)

Integrative Data Analytics ↔ Precision Medicine

- **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.



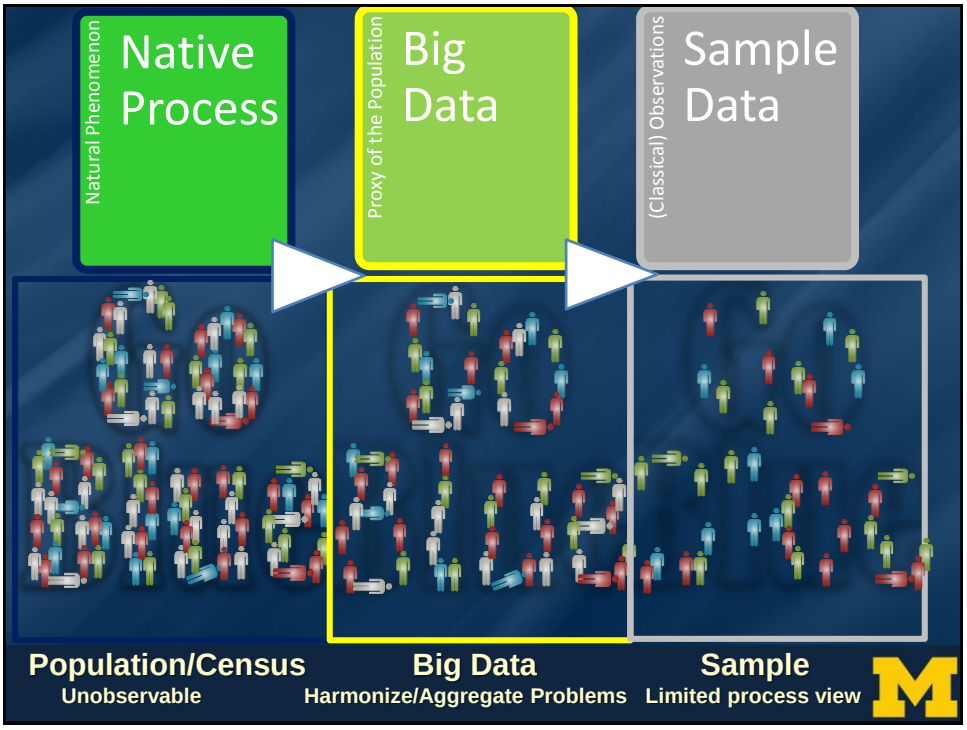
A: Individual brain parcellation B: LPBA40 atlas

Index	Volume Intensity	ROI Name	Index	Volume Intensity	ROI Name
1	21	L superior frontal gyrus	29	65	L inferior occipital gyrus
2	24	R middle frontal gyrus	30	164	R putamen
3	50	R precuneus	31	69	L superior occipital gyrus
4	181	cerebellum	32	30	R middle orbitofrontal gyrus
5	47	L angular gyrus	33	42	R postcentral gyrus
6	122	R cingulate gyrus	34	27	L precentral gyrus
7	83	L middle temporal gyrus	35	32	R lateral orbitofrontal gyrus
8	90	R lingual gyrus	36	121	L cingulate gyrus
9	81	L superior temporal gyrus	37	31	L lateral orbitofrontal gyrus
10	91	L fusiform gyrus	38	92	R fusiform gyrus
11	44	R superior parietal gyrus	39	45	L supramarginal gyrus
12	96	R inferior occipital gyrus	40	88	R parahippocampal gyrus
13	87	L parahippocampal gyrus	41	22	R superior frontal gyrus
14	162	R caudate	42	29	L middle orbitofrontal gyrus
15	85	L inferior temporal gyrus	43	68	R cuneus
16	182	brainstem	44	62	R superior occipital gyrus
17	43	L superior parietal gyrus	45	33	L gyrus rectus
18	28	R precentral gyrus	46	48	R angular gyrus
19	23	L middle frontal gyrus	47	64	R middle occipital gyrus
20	89	L lingual gyrus	48	84	R middle temporal gyrus
21	41	L postcentral gyrus	49	49	L precuneus

http://DSPA.predictive.space

Integrative Data Analytics ↔ Precision Medicine

- **Information Complexity** – large, incongruent, time-varying data
- **Precision Medicine** – customized medical decisions, clinical practice, treatments, or healthcare products to individual patients
- **Individual vs. Population Studies** – inductive (discriminative) vs. deductive (generative) models for clinical decision support
- **Tools** – molecular diagnostics, imaging, clinical, wearables, analytics, …



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**: 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

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

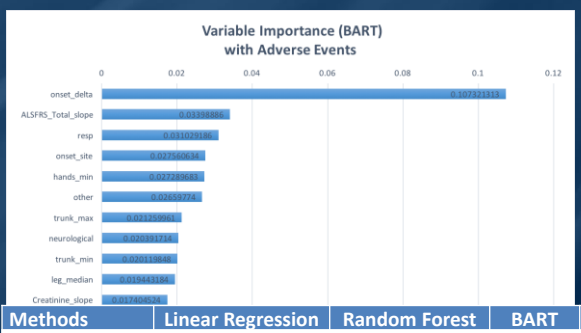
Case-Studies – ALS

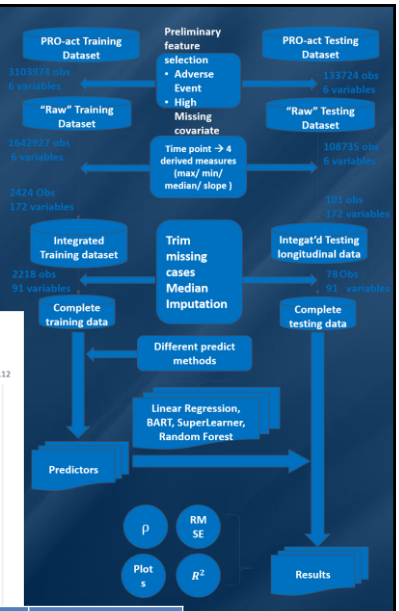
- Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

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





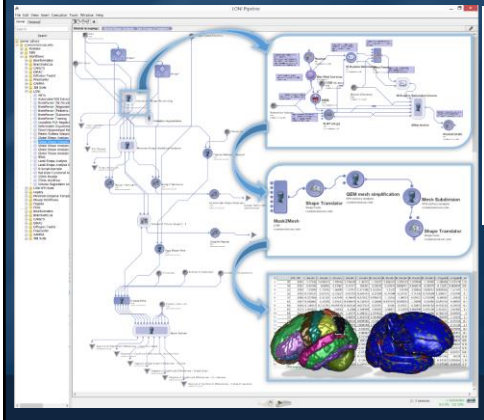
Methods	Linear Regression	Random Forest	BART	SuperLearner
R-squared	0.081	0.174	**0.225**	0.178
RMSE	0.619	0.587	**0.568**	0.585
Correlation	0.298	0.434	**0.485**	0.447

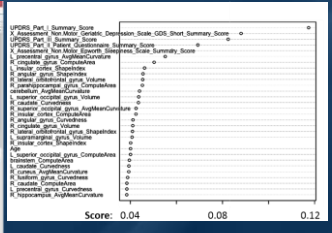
Case-Studies – Parkinson's Disease

- 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

Case-Studies – General Populations

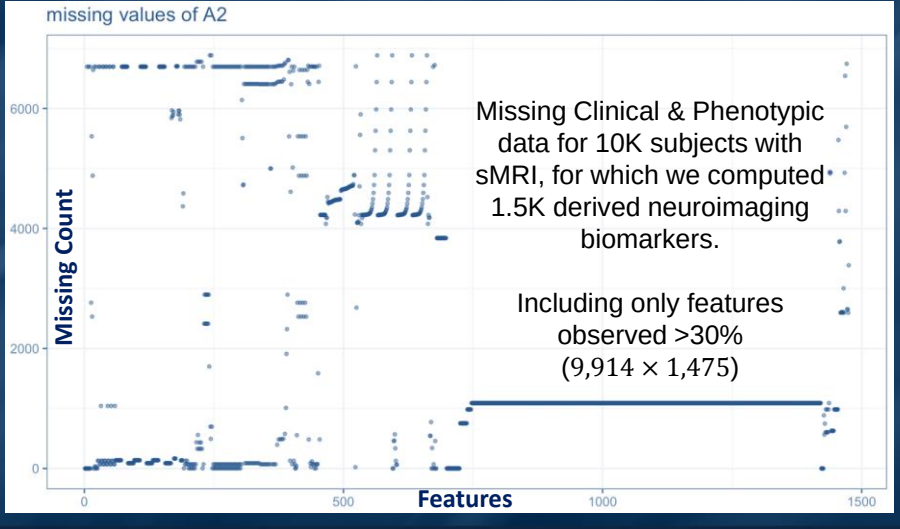
2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain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

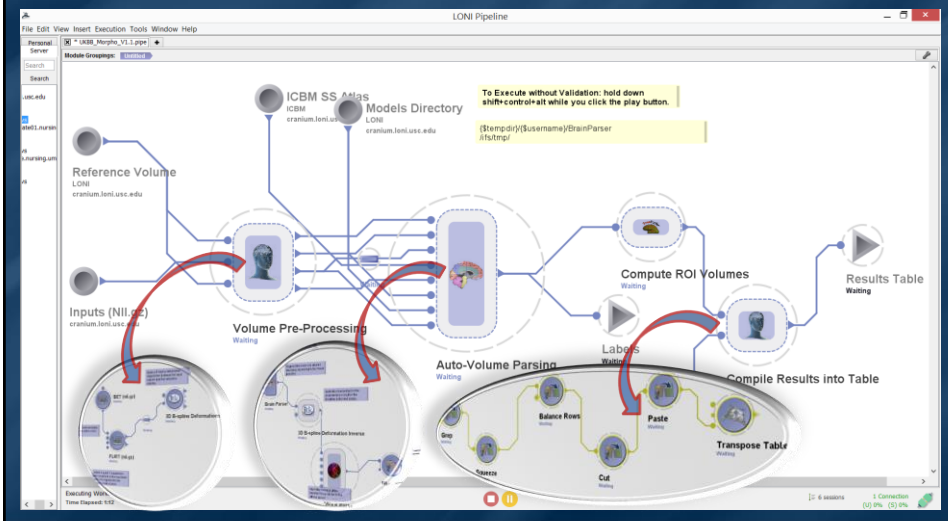
Case-Studies – UK Biobank - Complexities



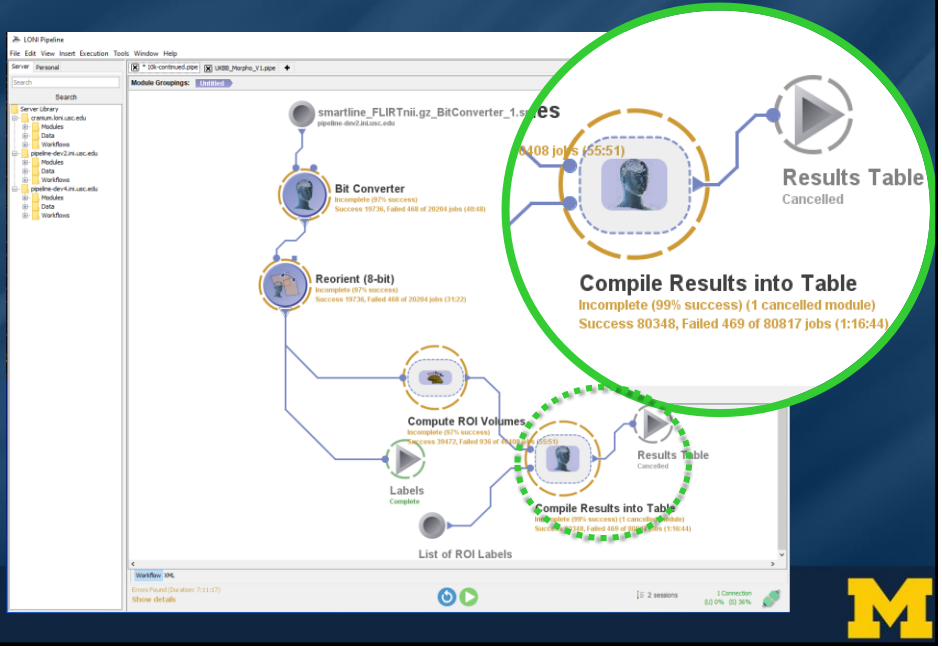
Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1.5K derived neuroimaging biomarkers.

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

Case-Studies – UK Biobank – NI Biomarkers



Case-Studies – UK Biobank – Successes/Failures



Case-Studies – UK Biobank – Derived NI Biomarkers

- Subjects=9,915
- Clinical Features=4,100
- Derived NI Biomarkers=3,297

lh_aparc.a2009s_volume__lh_S_oc.temp_med.Lingual_volume
lh_aparc.a2009s_volume__lh_S_orbital_lateral_volume
lh_aparc.a2009s_volume__lh_S_orbital_med.olfact_volume
lh_aparc.a2009s_volume__lh_S_orbital.H_Shaped_volume
lh_aparc.a2009s_volume__lh_S_parieto_occipital_volume

	subject_id	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_	lh_aparc_
2	1000262	981	442	1768	1398	380	2881	3794	2576	947	4526	2350	2489	1646	2620	692
3	1000489	1138	434	2090	1675	504	3448	5261	3446	1158	5640	2646	3423	1960	3001	664
4	1000649	989	561	1764	1389	391	2798	4366	2856	1163	4881	2431	2639	1586	2997	593
5	1001782	916	664	2144	1464	551	2830	4344	3237	811	4783	2124	2766	1893	2910	589
6	1001993	937	550	2109	1371	490	2910	4673	3476	1116	5213	2851	2980	1960	3170	522
7	1002402	1143	444	2045	1459	390	2869	4381	3499	1007	5969	2741	3007	1916	3380	594
8	1002595	1138	549	2165	1822	381	3166	5218	3407	1122	5275	2907	3536	2093	3519	592
9	1003472	1101	666	2049	1249	434	2943	4518	3904	912	4666	2611	2700	1907	3156	671
10	1003481	1004	688	2003	1709	477	2977	4808	3405	1073	5765	2626	3572	1795	3141	716
11	1003536	1128	561	1871	1674	463	3050	4293	2853	814	5143	2602	2479	1752	2859	567
12	1005842	1027	504	2503	1787	550	3572	3756	3802	957	5151	2548	3690	2023	3295	669
13	1005912	968	579	2007	1690	475	3159	3752	3779	907	5104	2536	3342	1840	2730	576
14	1006011	1059	728	2160	1518	464	2986	4786	3624	1151	4963	2673	3211	2018	3383	762
15	1006164	892	490	1990	1932	468	3143	3142	3186	1152	5663	2653	3584	1762	2925	603
16	1006577	932	560	2006	1541	395	2716	3988	3454	876	4488	2467	2876	1704	3209	691
17	1006835	1133	789	2544	1317	457	3386	4308	3151	830	4746	2345	2293	2026	2874	674
18	1007354	826	449	2448	1389	458	2814	3968	3080	827	4132	2399	2814	1840	2765	658
19	1007820	883	650	2500	1636	448	3068	4460	3537	1159	4788	3044	3170	1510	3193	580
20	1008359	833	480	1773	1382	438	2964	4202	3245	1204	4919	2637	2823	2185	3014	598
21	1008610	699	618	1957	1445	308	2327	3125	2121	909	4803	1953	2667	1482	2401	524

lh_aparc.DKTatlas_area__lh_caudalmiddlefrontal_area
lh_aparc.DKTatlas_area__lh_cuneus_area
lh_aparc.DKTatlas_area__lh_entorhinal_area
lh_aparc.DKTatlas_area__lh_fusiform_area
lh_aparc.DKTatlas_area__lh_inferiorparietal_area
lh_aparc.DKTatlas_area__lh_inferiortemporal_area

Case-Studies – UK Biobank – Unstructured Data

- n=11K cases with clinical diagnosis (notes)
- Computed Corpus Fingerprint – signature tensor
- LASSO regularize modeling (outcome death in 12 months)
  - max AUC = 0.5657 (poor)
- External Validation using 8 patients from the SOCR Hospital Admissions Data

```
predTestPatients <- predict(fit_test, s = fit_test$lambda.1se,
        newx = dtm_testPatients, type="response");
predTestPatients  # Probabilities of Death
## 1 0.9698585
## 2 0.9698585
## 3 0.9698585
## 4 0.9698585
## 5 0.9698585
## 6 0.9698585
## 7 0.9691101
## 8 0.9698585
```

- …

Initial

```
<<DocumentTermMatrix (documents: 11134, terms: 2566)
>> ## Non-/sparse entries: 607,58/28,509,086
## Sparsity : 100%
## Maximal term length: 25
## Weighting : term frequency (tf)
```

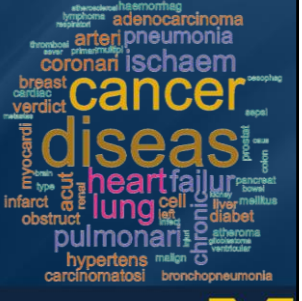
Post-processing

```
<<DocumentTermMatrix (documents: 384, terms: 64)
>> ## Non-/sparse entries: 1,421/23,155
## Sparsity : 94%
## Maximal term length: 16
## Weighting : term frequency (tf)
```

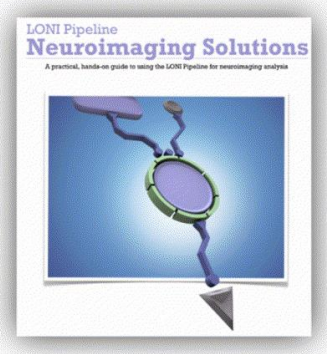
Training + Cross-Validation Testing on Hospital Admissions Data

table(TestY, predTestPatients)

## TestY	0	1
## 0	9127	45
## 1	897	16

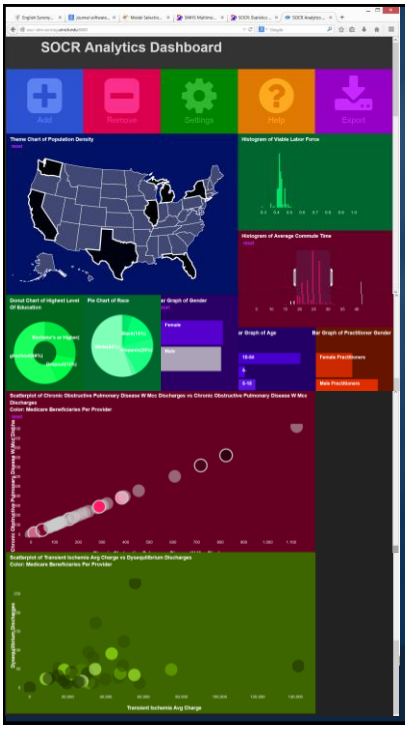


End-to-end Pipeline Workflow Solutions



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

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



SOCR Big Data Dashboard

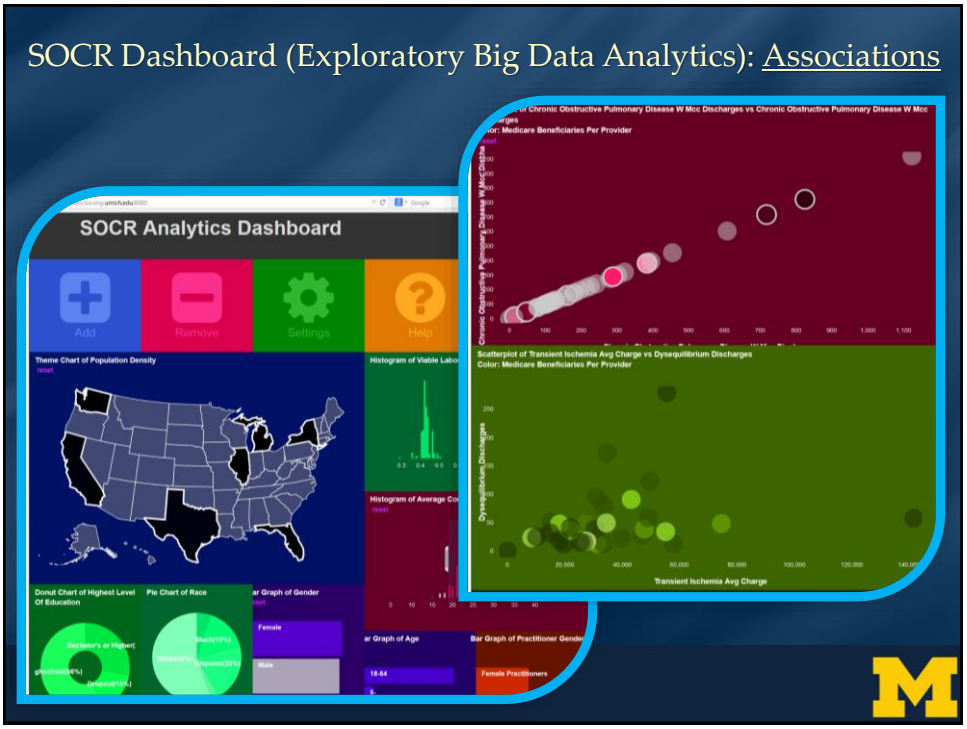
http://socr.umich.edu/HTML5/Dashboard

- 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

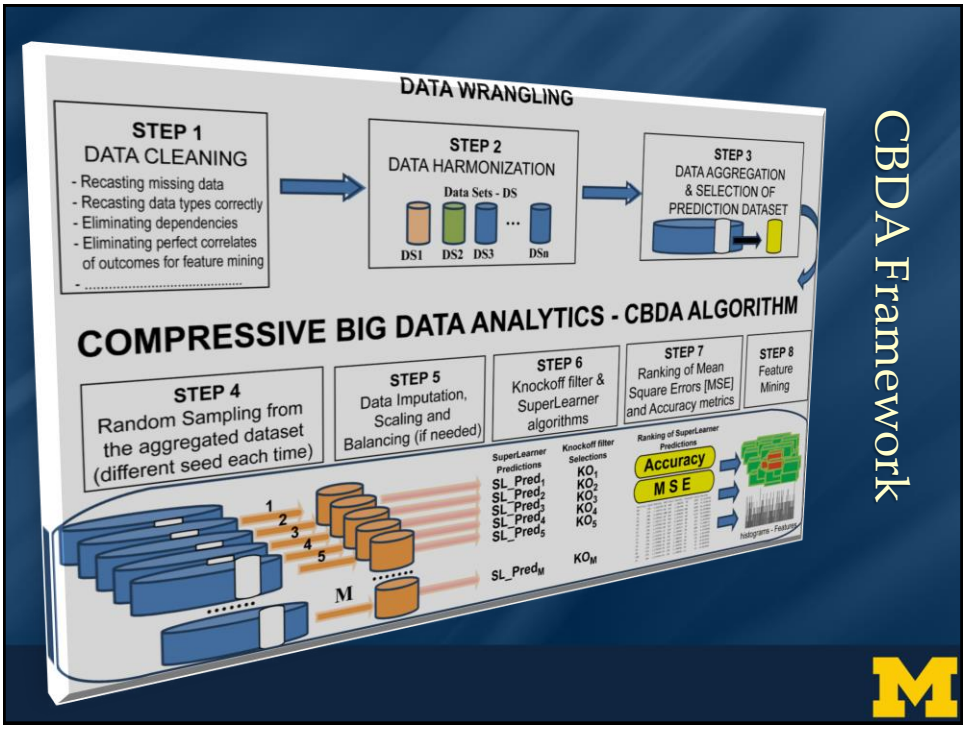
SOCR Dashboard (Exploratory Big Data Analytics): Associations



Compressive Big Data Analytics (CBDA)

- Foundations 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*

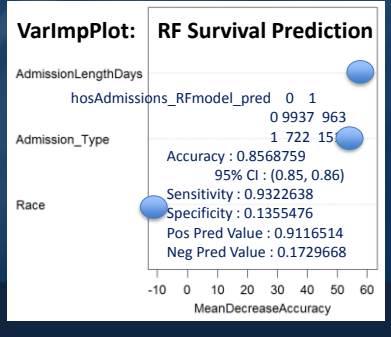


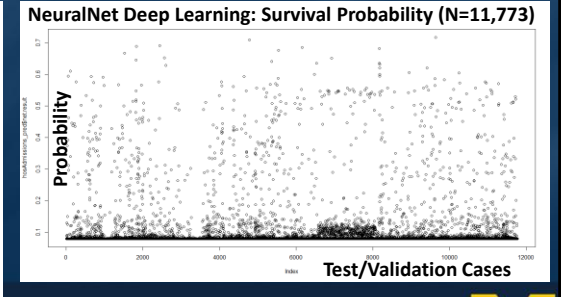
FAIR Data & Open-Science Principles

- 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

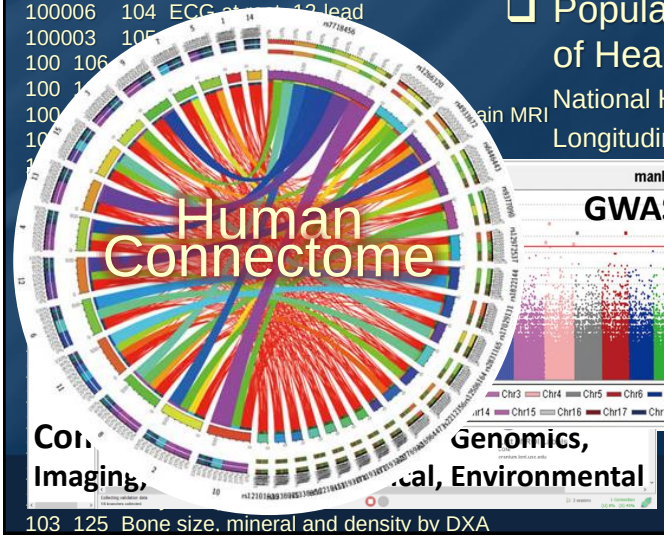


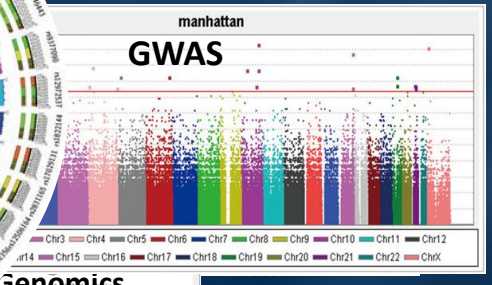


Clinical Decision Support

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
100
10
1

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



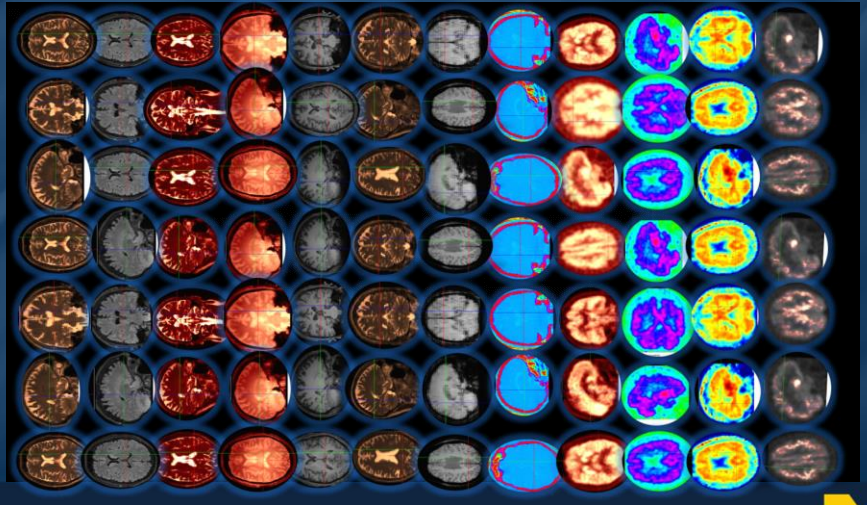


Con Genomics,
Imaging, cal, Environmental

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

Slides available online:
Google "SOCR News"

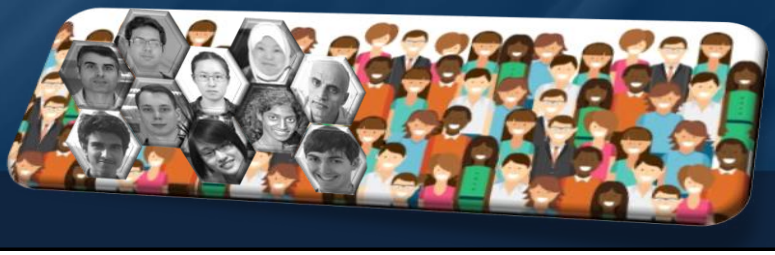
Funding

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

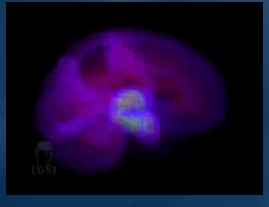
Collaborators

- **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