Open Data Science & 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

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

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

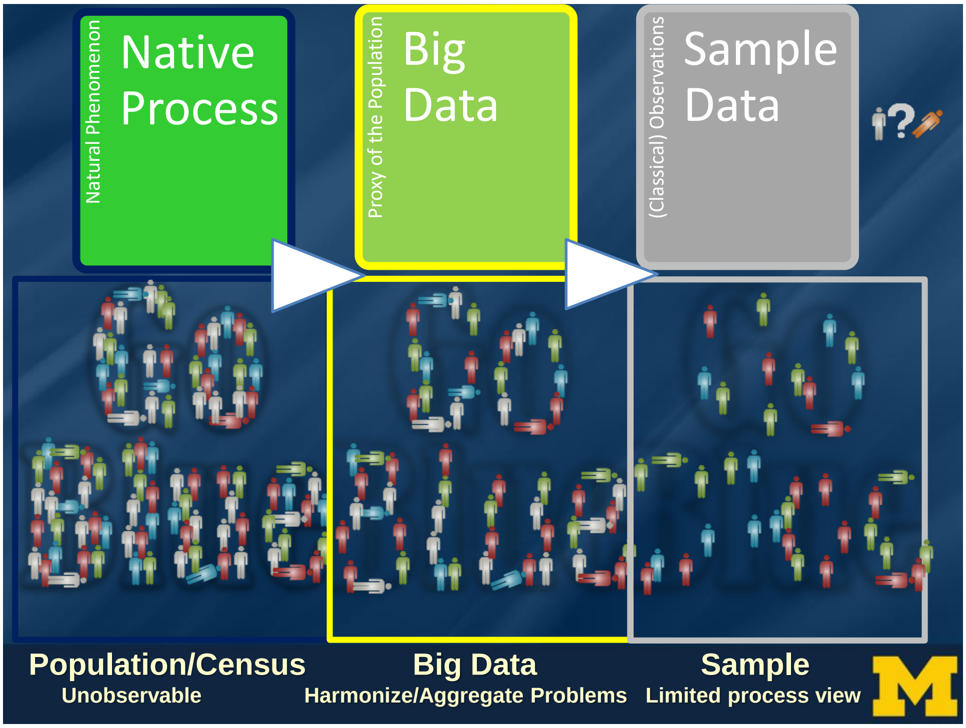
Slides Online:
"SOCR News"

Outline

- ❑ Driving biomedical & health challenges
- ❑ Common characteristics of Big Neuroscience Data

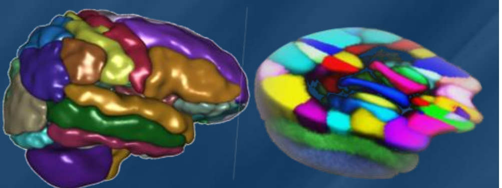
- ❑ Data science & predictive neuro-analytics methods

- ❑ Case-studies
  - ❑ Prenatal Exposure to Methamphetamine & Alcohol
  - ❑ Parkinson's disease (PD)
  - ❑ UK Biobank (Hands-on Demo)



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 | 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	61	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	66	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
Moon, Dinov, *et al.* (2015)

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

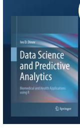
Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

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

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)

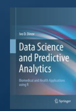


http://DSPA.predictive.space

Dinov, Springer (2018)

Data Science & Predictive Analytics

- Dimensionality Reduction
- Lazy Learning: Classification Using Nearest Neighbors
- Probabilistic Learning: Classification Using Naive Bayes
- Decision Tree Divide and Conquer Classification
- Forecasting Numeric Data Using Regression Models
- Black Box Machine-Learning Methods: Neural Nets/Support Vector Machines
- Apriori Association Rules Learning
- k-Means Clustering
- Model Performance Assessment
- Improving Model Performance
- Specialized Machine Learning Topics
- Variable/Feature Selection
- Regularized Linear Modeling and Controlled Variable Selection
- Big Longitudinal Data Analysis
- Natural Language Processing/Text Mining
- Prediction and Internal Statistical Cross Validation
- Deep Learning, Neural Networks



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

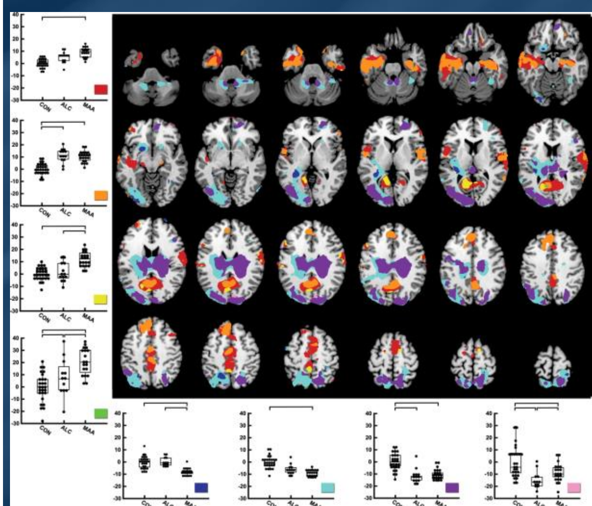
Dinov, GigaScience (2016) PMID:26918190

Case-Studies – Prenatal Exposure to Methamphetamine & Alcohol

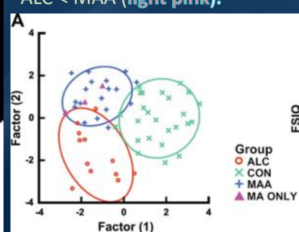
- Goals: Examine the local brain effects of prenatal exposure to methamphetamine (MA)
- Data: structural magnetic resonance imaging (sMRI). Compared local brain volumes differed among 61 children (ages 5–15 years),
  - 21 with prenatal MA exposure,
  - 18 with concomitant prenatal alcohol exposure (the MAA group),
  - 13 with heavy prenatal alcohol but not MA exposure (ALC group), and
  - 27 unexposed controls.
- Methods: Brain morphometry (sMRI processing) & Discriminant analysis (prediction)
- Results:
  - Bilateral volume reductions in both exposure groups relative to controls in striatal and thalamic regions, right prefrontal and left occipitoparietal cortices.
  - MAA group had negative correlation between full-scale intelligence quotient (FSIQ) scores and caudate volume.
  - LDA prediction of group membership correctly classified 72% of participants.
- Conclusions: Striatal and limbic structures, known to be sites of neurotoxicity in adult MA abusers, may be more vulnerable to prenatal MA exposure than alcohol exposure; Severe striatal damage is associated with more severe cognitive deficit

Sowell, et al. JNeurosci (2010)

Case-Studies – Prenatal Exposure to Methamphetamine & Alcohol



p maps representing the following contrast categories between groups: (1) CON < MAA (**red**), (2) CON < MAA and CON < ALC (**orange**), (3) CON < MAA and ALC < MAA (**yellow**), (4) CON < ALC, CON < MAA, and ALC < MAA (**green**), (5) CON > MAA and ALC > MAA (**dark blue**), (6) CON > MAA (**light blue**), (7) CON > ALC and CON > MAA (**purple**), and (8) CON > ALC, CON > MAA, and ALC < MAA (**light pink**).



Sowell, et al. JNeurosci (2010)

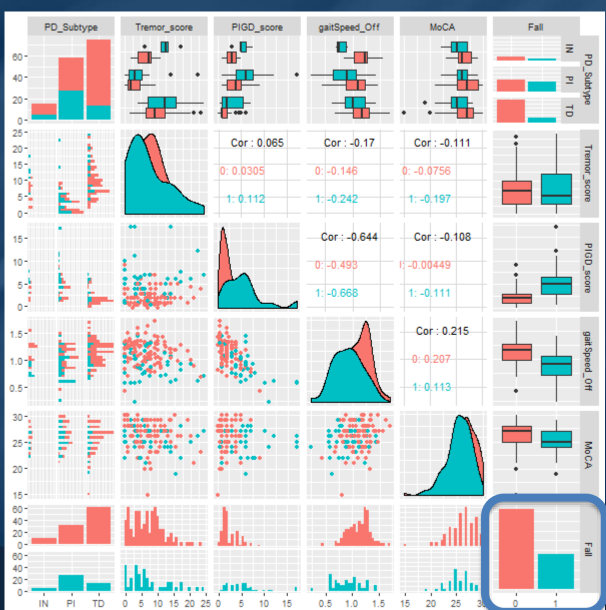
Factor analysis using Jacobian values for 14 ROIs

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	**0.796**	**0.683**	**0.871**	**0.778**	**0.806**	**2.677**	**0.821**
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

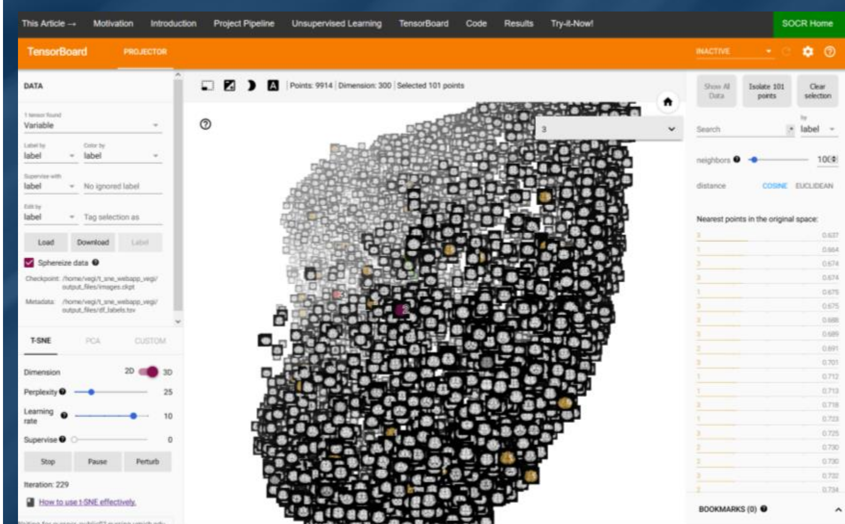
2 20005 Ongoing characteristics Email access
2 110007 Ongoing characteristics Newsletter communications, date sent
100 25780 Brain MRI Acquisition protocol phase.
100 12139 Brain MRI Believed safe to perform brain MRI scan
100 12188 Brain MRI Brain MRI measurement completed
100 12187 Brain MRI Brain 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- 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
http://myumi.ch/6wQgv

Case-Studies – UK Biobank



Try-It-Now
With your own data!

http://myumi.ch/6wQgv
http://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB/

Acknowledgments

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**: 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, Jerome Choi, Yuming Sun, Yi Zhao, Nina Zhou, 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