# Predictive Analytics of Big Neuroscience Data

Ivo D. Dinov

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

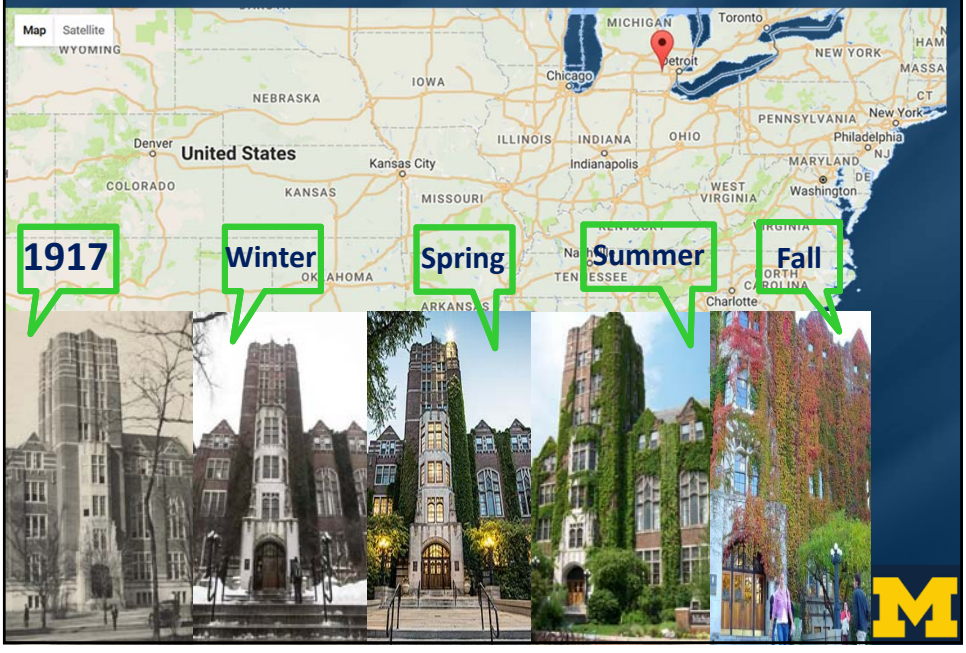
University of Michigan

http://SOCR.umich.edu

*Slides Online: "SOCR News"*

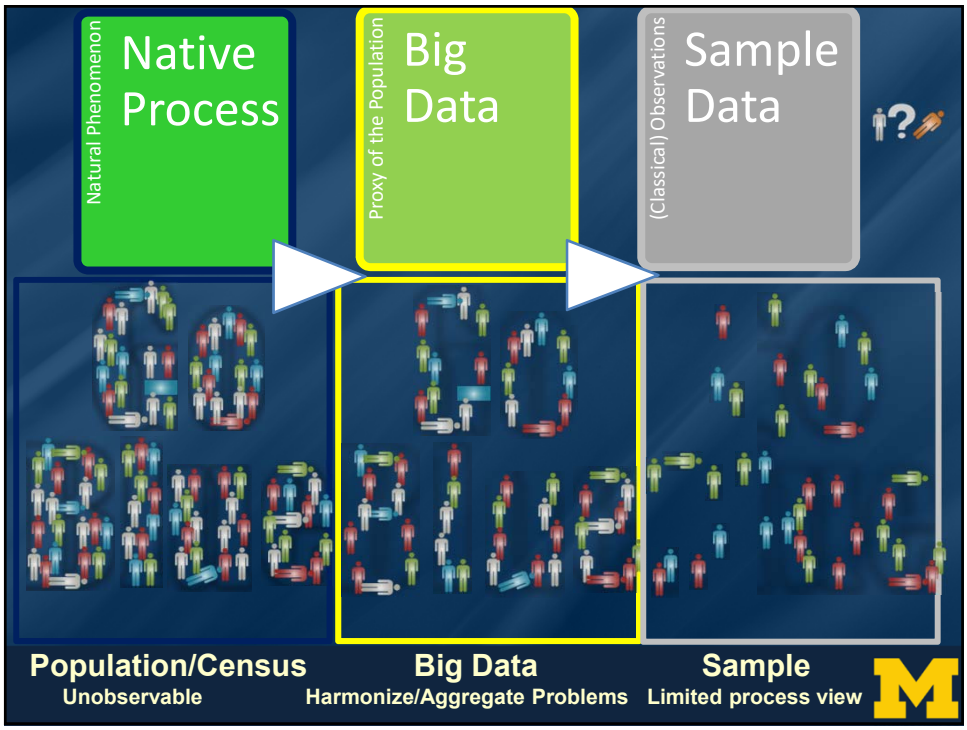
SCHOOL OF MEDICINE | STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR) | UNIVERSITY OF MICHIGAN

## The University of Michigan (est. 1817)



# Outline

- Driving biomedical & health challenges
- Common characteristics of Big Neuroscience Data
- $\varepsilon$-Differential Privacy & Homomorphic Encryption
- *DataSifter: Statistical obfuscation*
- Case-studies
  - Applications to Neurodegenerative Disease (Udall/MADC)
  - Autism Brain Imaging Data Exchange (ABIDE)
  - Population Census-like Neuroscience



# 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 (2016) GigaScience Dinov (2018) Springer

# $\varepsilon$-Differential Privacy ($\varepsilon$DP) vs. fully Homomorphic Encryption (fHE)

| Category | $\varepsilon$DP | fHE |
|---|---|---|
| **Goal** | Mine information in a DB without compromising privacy; no access to inspect individual DB entries | Provide a secure encryption allowing program execution on encrypted data; encrypt results, interpretation requires ability to decrypt the data |
| **Pros** | Theoretical limits on the balance between utility and risk of sharing data | Elegant and powerful math framework for bijective (encode/decode) encryption. Fast |
| **Cons** | Difficult for unstructured, skewed, and categorical data | There are limitations on deriving $f'$ – commutative analytic evaluators |

# ε-Differential privacy (εDP)

- **Data-features**: $\{C_1, C_2, \dots, C_k\}$, categorical or numerical.
- **DB** = list of cases $\{x_1, x_2, \dots, x_n\}$, $x_i \in C_1 \times C_2 \times \cdots, \times C_k$, $1 \le i \le n$.

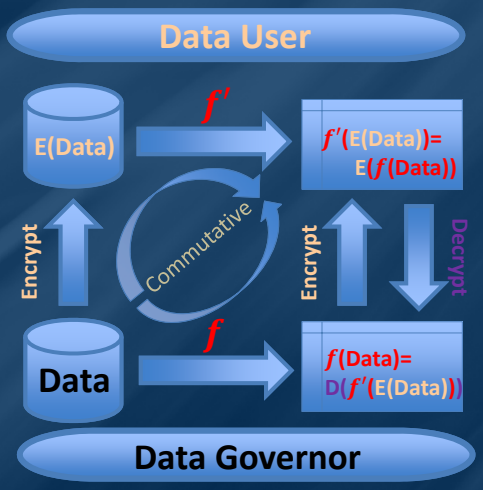
- $\varepsilon$-Differential privacy relies on adding noise to data to protect the identities of individual records. An **algorithm $f$** is $\varepsilon$-differentially private if for all possible inputs (datasets/DBs) $D_1, D_2$ that differ on a single record, and all possible $f$ outputs, $y$, the probability of correctly guessing $D_1$ knowing $y$ is not significantly different from that of $D_2$:

$$\frac{P(f(D_1) = y)}{P(f(D_2) = y)} \le e^{\varepsilon}, \qquad \forall y \in Range(f).$$

- The global sensitivity of $f$ is the smallest number $S(f)$, such that $\forall D_1, D_2$ that differ on at most one element $\|f(D_1) - f(D_2)\|_1 \le S(f)$
- There are many differentially private algorithms, e.g., random forests, decision trees, k-means clustering, etc.
- E.g., $f: D = DB \to R^m$, the algorithm outputting $f(D) + (y_1, y_2, \dots, y_m)$, with $y_i \in Laplace\left(\mu = 0, \sigma = \sqrt{2}\frac{S(f)}{\varepsilon}\right), \forall i$ is $\varepsilon$-differentially private

Dwork, LNCS, 2008

# Fully Homomorphic Encryption (fHE)




Legend

Data – dataset/DB
$f$ – data analytic process
$f'$ – commutative analytic evaluator
E – encryption protocol
D – decryption protocol

Rivest & Adleman, Academic Press, 1978

# DataSifter

- DataSifter is an iterative statistical computing approach that provides the data-governors controlled manipulation of the trade-off between sensitive information obfuscation and preservation of the joint distribution.
- The DataSifter is designed to satisfy data requests from pilot study investigators focused on specific target populations.
- Iteratively, the DataSifter stochastically identifies candidate entries, cases as well as features, and subsequently selects, nullifies, and imputes the chosen elements. This statistical-obfuscation process relies heavily on nonparametric multivariate imputation to preserve the information content of the complex data.

http://DataSifter.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

# DataSifter

- A detailed description and *dataSifter()* R method implementation are available on our GitHub repository (https://github.com/SOCR/DataSifter).
- Data-sifting different data archives requires customized parameter management. Five specific parameters mediate the balance between protection of sensitive information and signal energy preservation.

| Obfuscation level | $0 \leq \eta = \eta(k_0 + k_1 + k_2 + k_3 + k_4) \leq 1$ | | | | |
|---|---|---|---|---|---|
| | $k_0$ | $k_1$ | $k_2$ | $k_3$ | $k_4$ |
| None | 0 | 0 | 0 | 0 | **0** |
| Small | 0 | 0.05 | 1 | 0.1 | **0.01** |
| Medium | 1 | 0.25 | 2 | 0.6 | **0.05** |
| Large | 1 | 0.4 | 5 | 0.8 | **0.2** |
| Indep | Output synthetic data with independent features | | | | |

$k_0$: A Boolean; obfuscate the unstructured features?

$k_1$: proportion of artificial missing data values that should be introduced

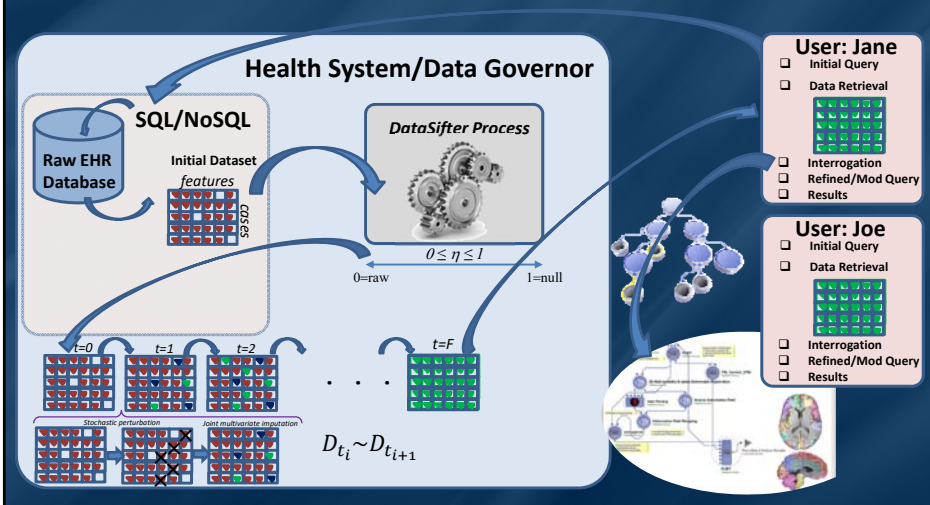
$k_2$: The number of times to iterate

$k_3$: The fraction of structured features to be obfuscated in all the cases

$k_4$: The fraction of closest subjects to be considered as neighbours of a given subject

http://DataSifter.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

# *DataSifter*



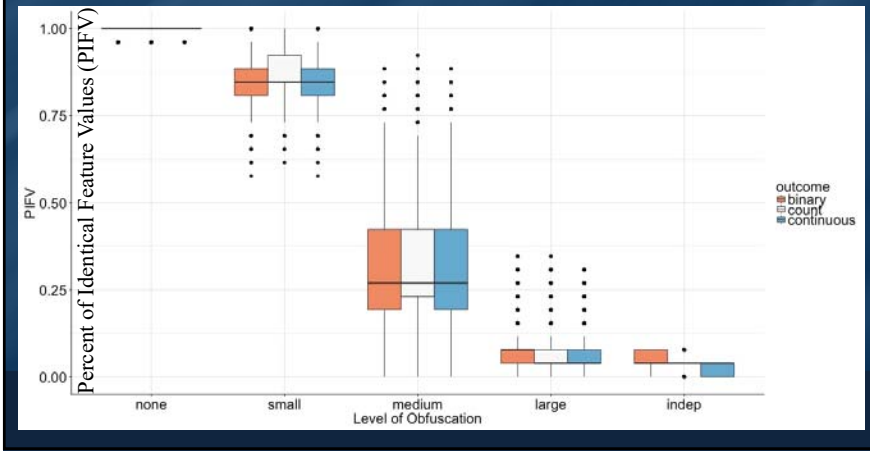

http://*DataSifter*.org US patent #16/051,881 Marino, *et al.*, *JSCS* (2019)

# *DataSifter* Validation

I. Protection of sensitive information (privacy)

PIFV under Different Privacy Levels. Binary outcome refers to the first experiment; Count refers to the second experiment; Continuous refers to the third experiment.

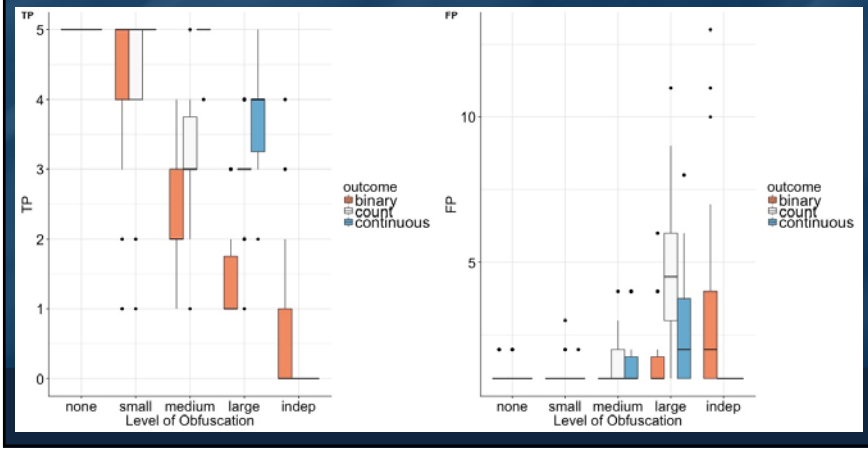
Each box represents 30 different "sifted" data or 30,000 "sifted" cases.



# *DataSifter* Validation

II. Preserving utility information of the original dataset

Logistic Model with Elastic Net Signal Capturing Ability. TP is the number of true signals (total true predictors = 5) captured by the model. FP is the number of null signals that the model has falsely selected (total null signals=20).




# *DataSifter* Validation

III. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data

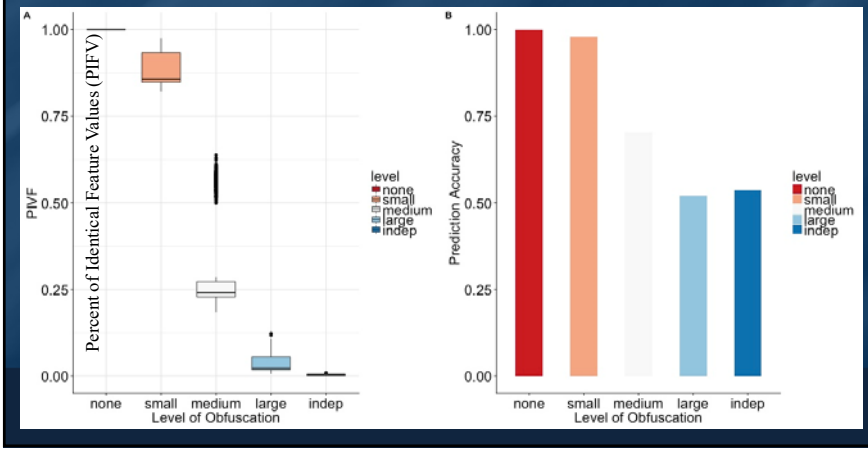
Comparing the Original and "Sifted" Data for the 22nd ABIDE Subject

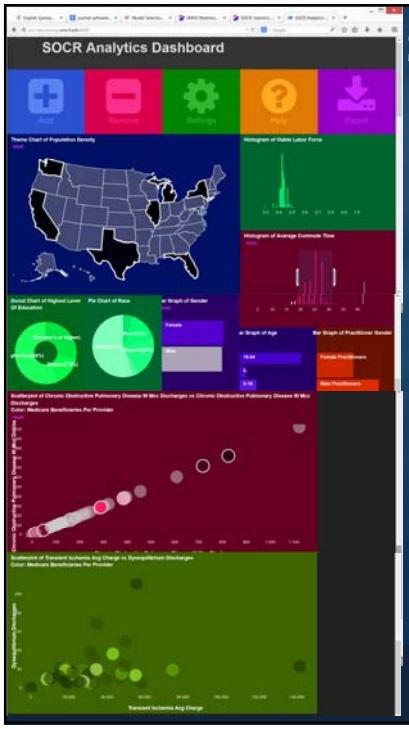
| *η* | Output | Sex | Age | Acquisition Plane | IQ | thick_std_ctx .lh.cuneus | curv_ind_ctx_lh_G_front_inf.Triangul | gaus_curv_ctx.lh. medialorbitofrontal | curv_ind_ctx_lh_S_interm_prim.Jensen |
|---|---|---|---|---|---|---|---|---|---|
| original | Autism | M | 31.7 | Sagittal | 131 | 0.475 | 2.1 | 0.315 | NA |
| none | Autism | M | 31.7 | Sagittal | 131 | 0.475 | 2.1 | 0.315 | 0.51 |
| small | Autism | M | 31.7 | Sagittal | 131 | 0.475 | 2.1 | 0.315 | 0.4589 |
| medium | Autism | M | 31.7 | Sagittal | 111 | 0.548 | 2.85 | 0.315 | 0.463 |
| large | Control | M | 18.2 | Sagittal | 104 | 0.5347 | 3.198 | 0.1625 | 0.4524 |
| indep | Control | M | 15.4 | Coronal | 104 | 0.4842 | 3.383 | 0.1079 | 1.002 |

Autism Brain Imaging Data Exchange (ABIDE) case-study

# *DataSifter* Validation

IV. Clinical Data Application: Using DataSifter to Obfuscate the ABIDE Data
PIFVs for ABIDE under different levels of DataSifter obfuscations.
Each box represents 1098 subjects among the ABIDE sub-cohort
Random forest prediction of binary clinical outcome - autism spectrum disorder – (ASD) status (ASD vs. control)




# 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*

# 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 (2018) Springer

# Case-Studies – General Populations

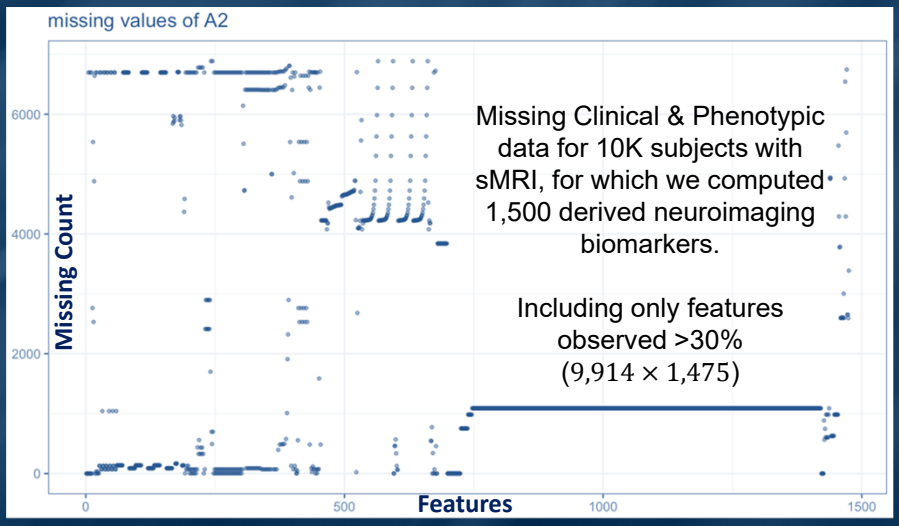
| | | | |
|---|---|---|---|
| 2 | 20005 | Ongoing characteristics | Email access |
| 2 | 110007 | Ongoing characteristics | Newsletter communications, date sent |
| 100 | 25780 | Brain MRI | Acquisition protocol phase. |
| 100 | 12139 | Brain MRI | Believed safe to perform brain MRI scan |
| 100 | 12188 | 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<br>**Clinical data:** > 4K features<br>**Imaging data**: T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI<br>**Genetics data** | The longitudinal archive of the UK population (NHS) |

http://www.ukbiobank.ac.uk
http://bd2k.org

# Case-Studies – UK Biobank (Complexities)

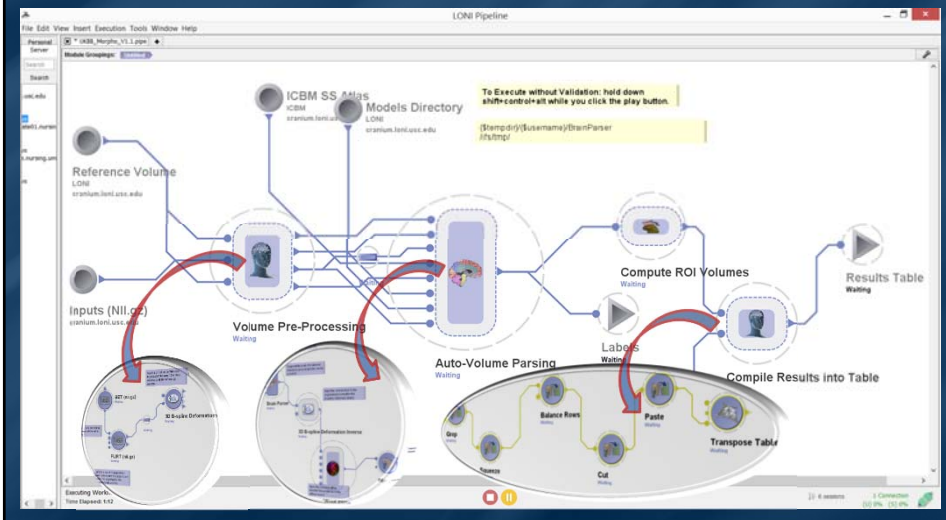



Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1,500 derived neuroimaging biomarkers.

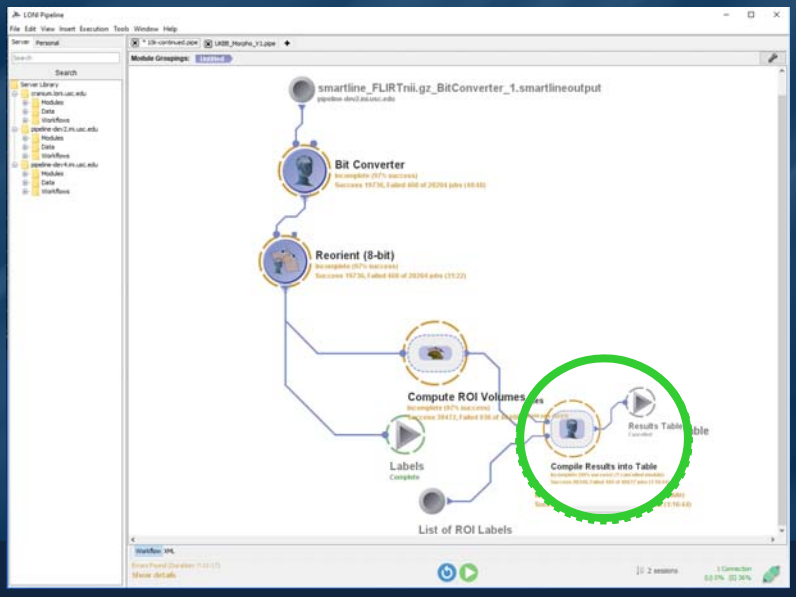
Including only features observed >30% $(9{,}914 \times 1{,}475)$

*Zhou, et al. (2019), in press, SREP* | https://github.com/SOCR/UKBB_Analytics

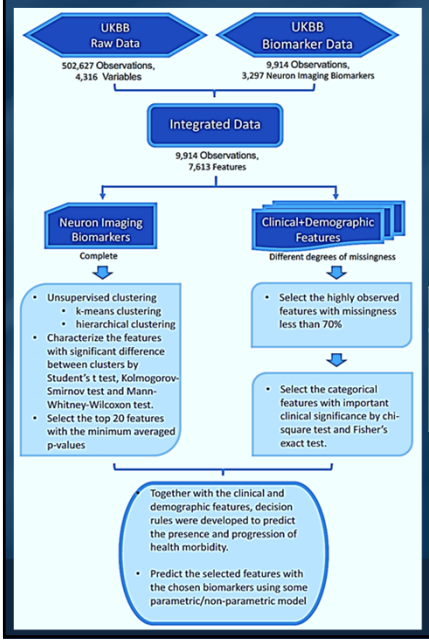
# Case-Studies – UK Biobank – NI Biomarkers

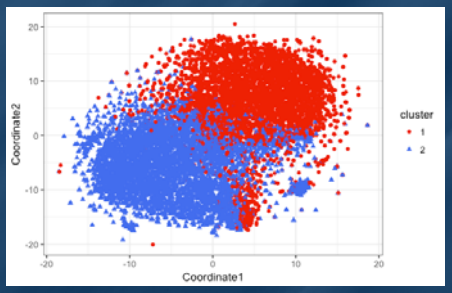


# Case-Studies – UK Biobank – Successes/Failures




# Case-Studies – UK Biobank – Results



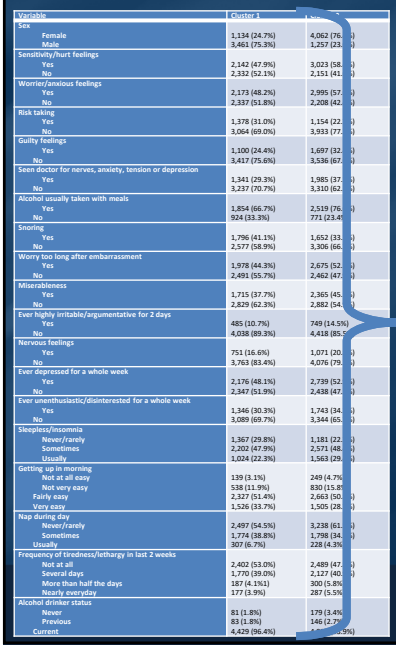




t-SNE plot of the brain neuroimaging biomarkers

| 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%) |

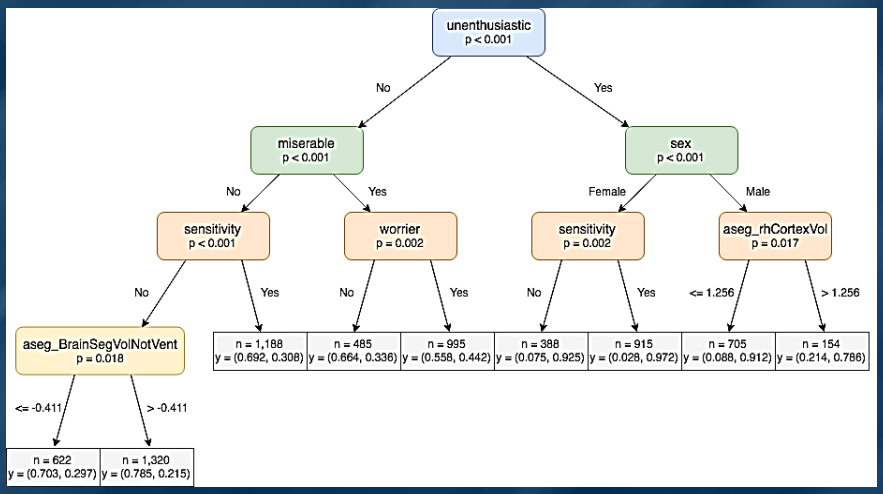
| Cluster | Consistency | Variance | Cluster-size | Silhouette |
|---|---|---|---|---|
| 1 | 0.997 | 0.001 | 5344 | 0.09 |
| 2 | 0.934 | 0.001 | 4570 | 0.05 |

# 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** | | |
| Not at all | 2,402 (53.0%) | 2,489 (47.8%) |
| Several days | 1,770 (39.0%) | 2,127 (40.9%) |
| More than half the days | 187 (4.1%1) | 300 (5.8%) |
| Nearly everyday | 177 (3.9%) | 287 (5.5%) |
| **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. (2019), in press SREP*

## What's Next?

- Lots of "open problems" in data-science, e.g., fundamentals of data representation & analytics
- The SOCR team is developing:
  - Compressive Big Data Analytics (CBDA) technique – an ensemble learning meta-algorithm
  - DS Time-Complexity and Inferential-Uncertainty
- Need lots of community, institutional, state, federal, and philanthropic support to advance data science methods, enhance the computing infrastructure, train/support students/fellows, and tackle the $Kryder\ Law \gg Moore\ Law$ trend

# Acknowledgments

Slides Online:
"SOCR News"

US patent #16/051,881


## 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, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang. Simeone Marino, Nina Zhou, Yi Zhao, Lu Wang, Qiucheng Wu
- **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