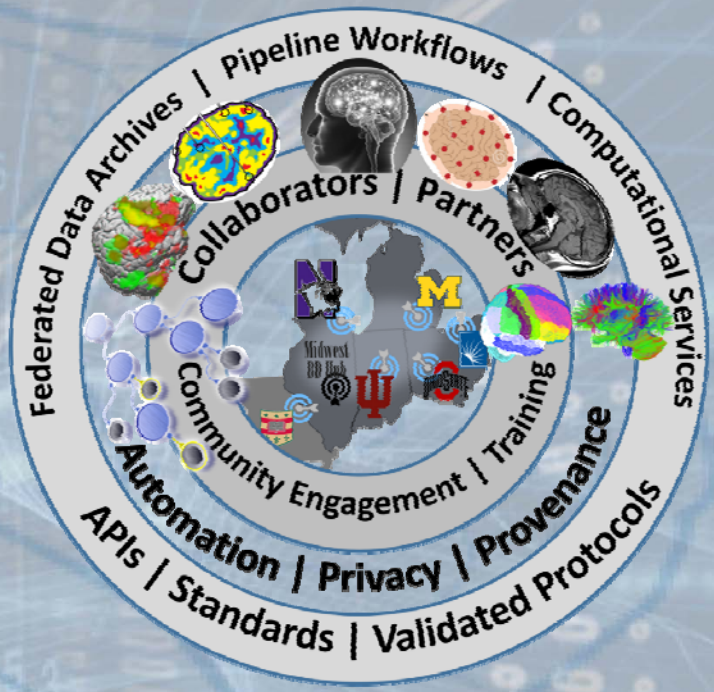
Midwest Workshop on Big Neuroscience Data, Tools, Protocols & Services



Computational Neuroscience Network (ACNN)

http://www.NeuroscienceNetwork.org/ACNN_Workshop_2016.html

Predictive Big Data Analytics

Ivo D Dinov

Statistics Online Computational Resource (SOCR)

Michigan Institute for Data Science (MIDAS)

University of Michigan

http://www.umich.edu/~dinov

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

MIDAS

Characteristics of Big Biomed Data

IBM Big Data 4V's: Volume, Variety, Velocity & Veracity

BD 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

Dinov, et al. (2014)

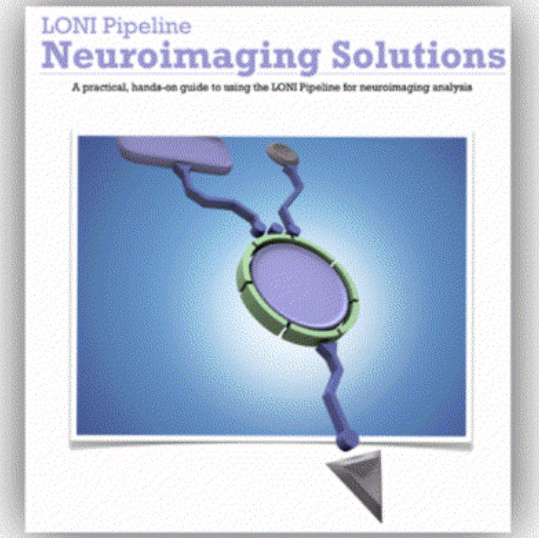
Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements.

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

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

End-to-end Pipeline Workflow Solutions



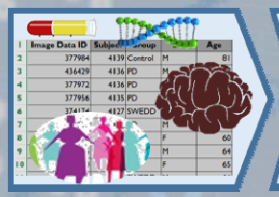
Dinov, *et al*., 2014, Front. Neuroinform.;

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

Predictive Big Data Analytics in Parkinson's Disease

- A **unique archive of Big Data**: Parkinson's Progression Markers Initiative (PPMI). Defining data characteristics – large size, incongruency, incompleteness, complexity, multiplicity of scales, and heterogeneity of information-generating sources (imaging, genetics, clinical, demographic)
- **Approach**
  - introduce methods for rebalancing imbalanced cohorts,
  - utilize a wide spectrum of classification methods to generate phenotypic predictions,
  - reproducible machine-learning based classification
- **Results** of machine-learning based classification show significant power to predict Parkinson's disease in the PPMI subjects (consistent accuracy, sensitivity, and specificity exceeding 96%, confirmed using internal statistical 5-fold cross-validation). Clinical (e.g., Unified Parkinson's Disease Rating Scale (UPDRS) scores), demographic (e.g., age), genetics (e.g., rs34637584, chr12), and derived neuroimaging biomarker (e.g., cerebellum shape index) data all contributed to the predictive analytics and diagnostic forecasting.
- **Model-free** Big Data machine learning-based classification methods (e.g., 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. Even without longitudinal UPDRS data, however, the accuracy of model-free machine learning based classification is over 80%. The methods, software and protocols developed here are openly shared and can be employed to study other neurodegenerative disorders (e.g., Alzheimer's, Huntington's). Dinov, et al., *PLoS*, 2016

Predictive Big Data Analytics: Applications to Parkinson's Disease



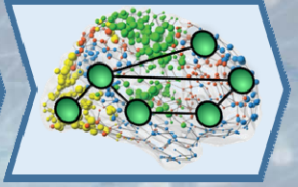
Data Representation Fusion Harmonization Aggregation

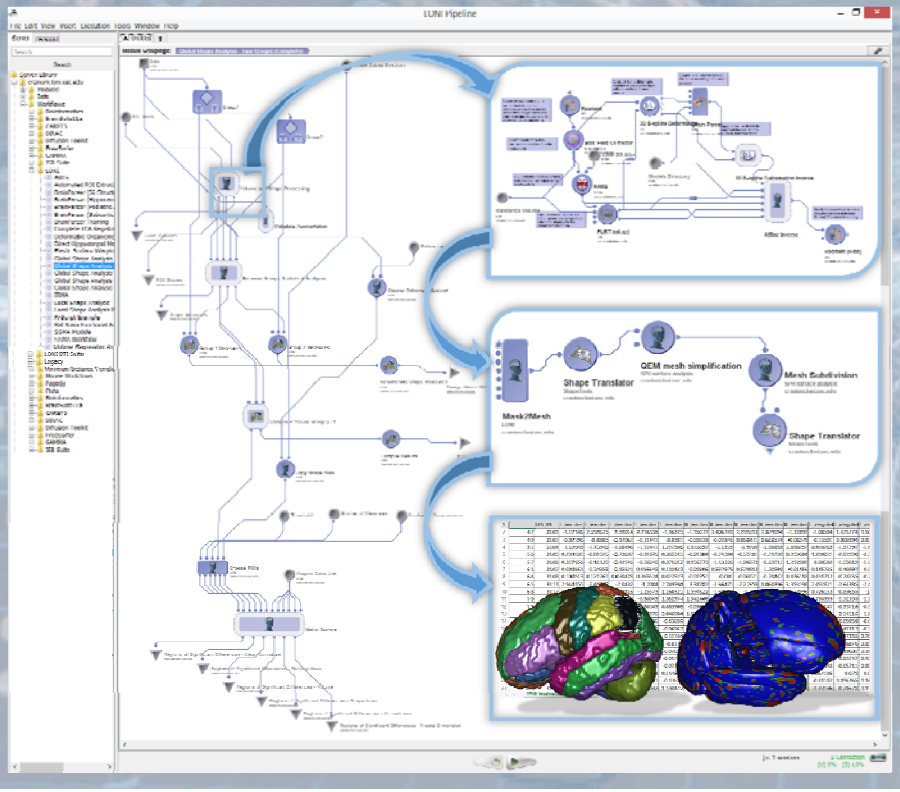
Cleaning Imputation Wrangling Synthesis

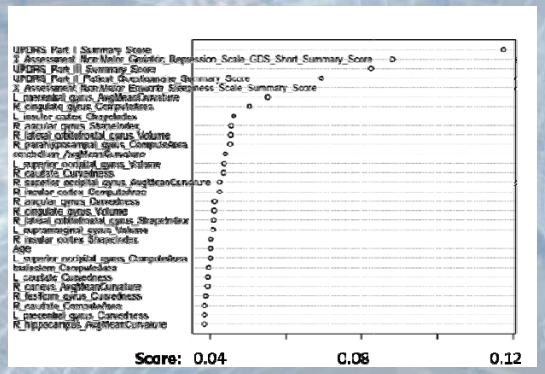
Rebalancing, Feature Selection

Model-based, Model-free, Classification

Prediction, Statistical Validation, Inference







Varplot illustrating:

- the critical predictive data elements (Y-axis)
- and their impact scores (X-axis)

AdaBoost classifier for Controls vs. Patients prediction

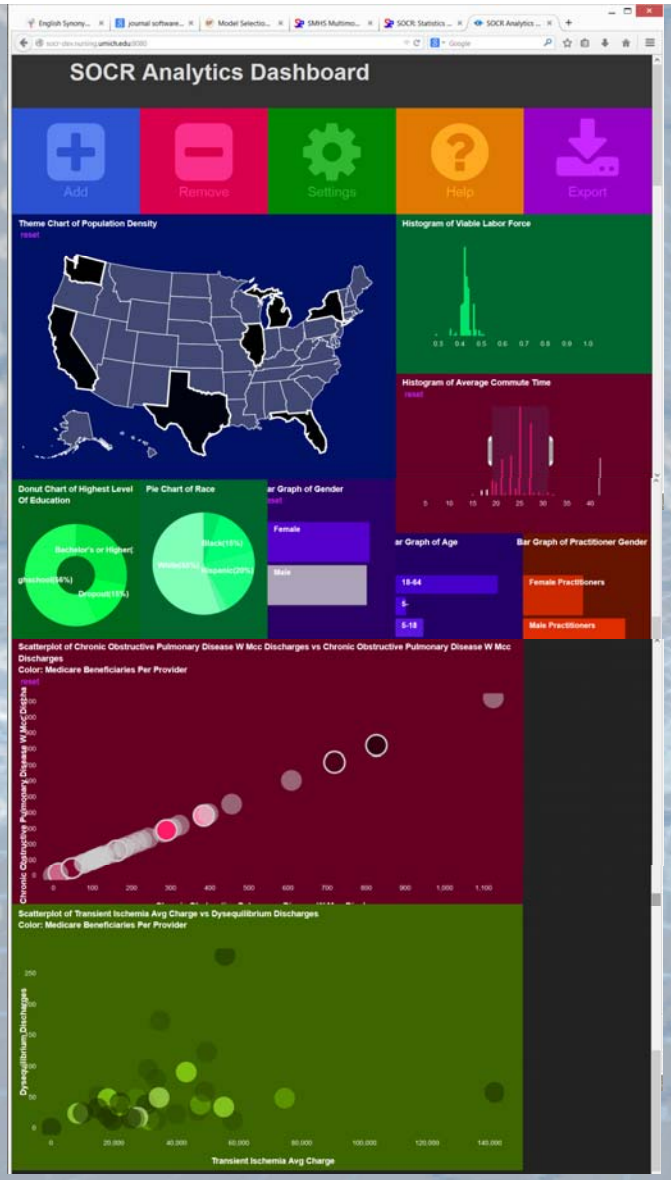
ML classifier	accuracy	sensitivity	specificity	positive predictive value	negative predictive value	log odds ratio (LOR)
AdaBoost	0.996324	0.994141	0.998264	0.9980392	0.9948097	11.4882058
SVM	0.985294	0.994140	0.977431	0.9750958	0.9946996	8.902166

Dinov, et al., *PLoS*, 2016

Predictive Big Data Analytics: Applications to Parkinson's Disease

Best Machine Learning Based Classification Results (according to average measures of 5-fold cross-validation)											
Data Strata	Classifier	FP	TP	TN	FN	Accuracy	Sensitivity	Specificity	positive predictive value	negative predictive value	log odds ratio (LOR)
balanced control/PD	ada	0.2	102	115	0.6	0.99632	0.99414	0.99826	0.998039	0.9948097	11.4882
balanced control/PD	svm	2.6	102	113	0.6	0.98529	0.99414	0.97743	0.975096	0.9946996	8.902166
balanced control/PD&SWEDD	svm	3.4	101	112	1	0.97978	0.99023	0.97049	0.967557	0.9911348	8.1120092
balanced control/PD&SWEDD	ada	1.4	101	114	1.8	0.98529	0.98242	0.98785	0.986275	0.9844291	8.4214
unbalanced control/PD	ada	0.2	25	54	0.8	0.98753	0.96875	0.99634	0.992	0.9855072	9.039789
unbalanced control/PD&SWEDD	ada	1.2	24	61	1.8	0.96599	0.92969	0.98083	0.952	0.971519	6.516987

Dinov, et al., *PLoS*, 2016



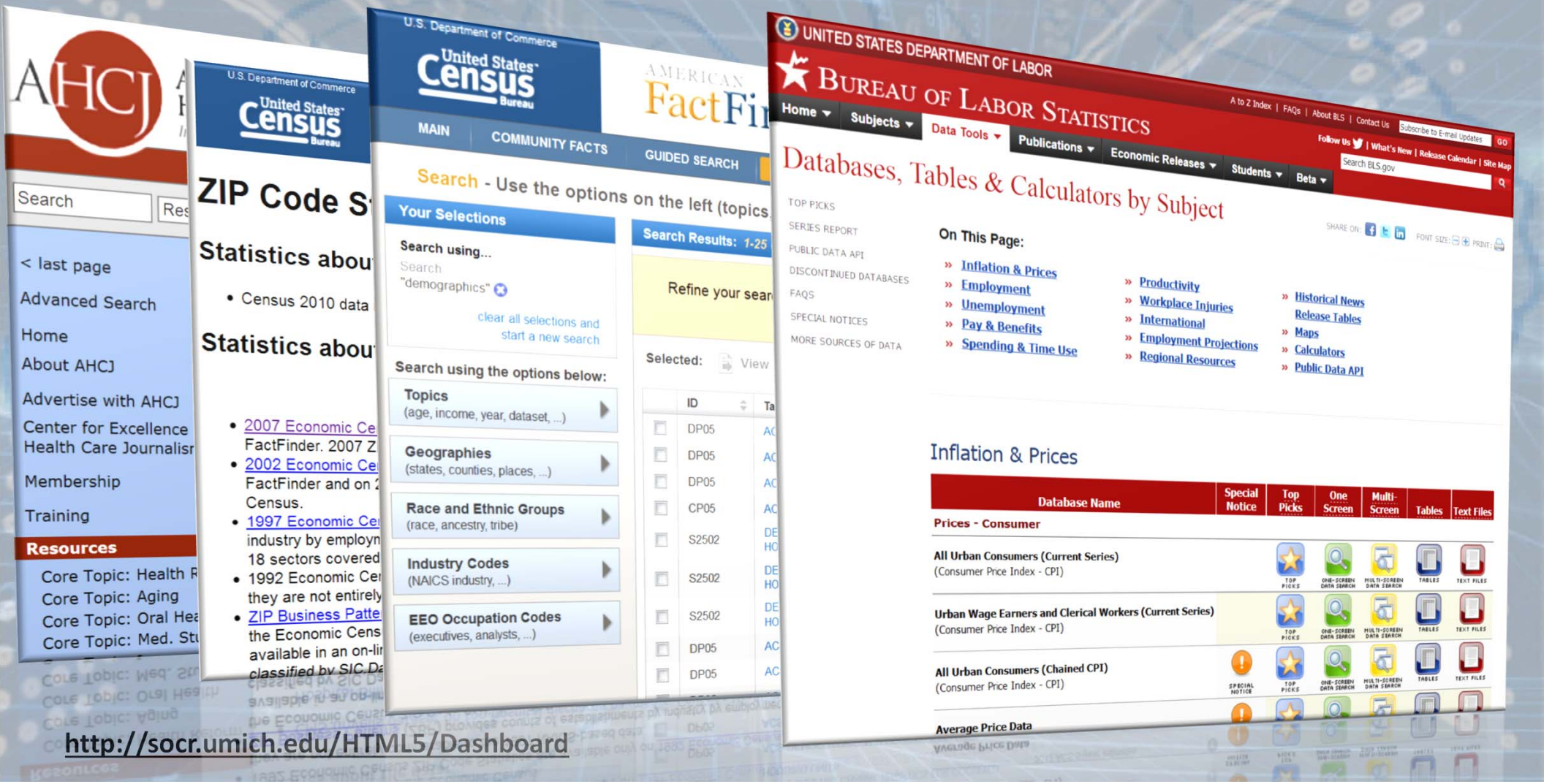
SOCR Big Data Dashboard

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

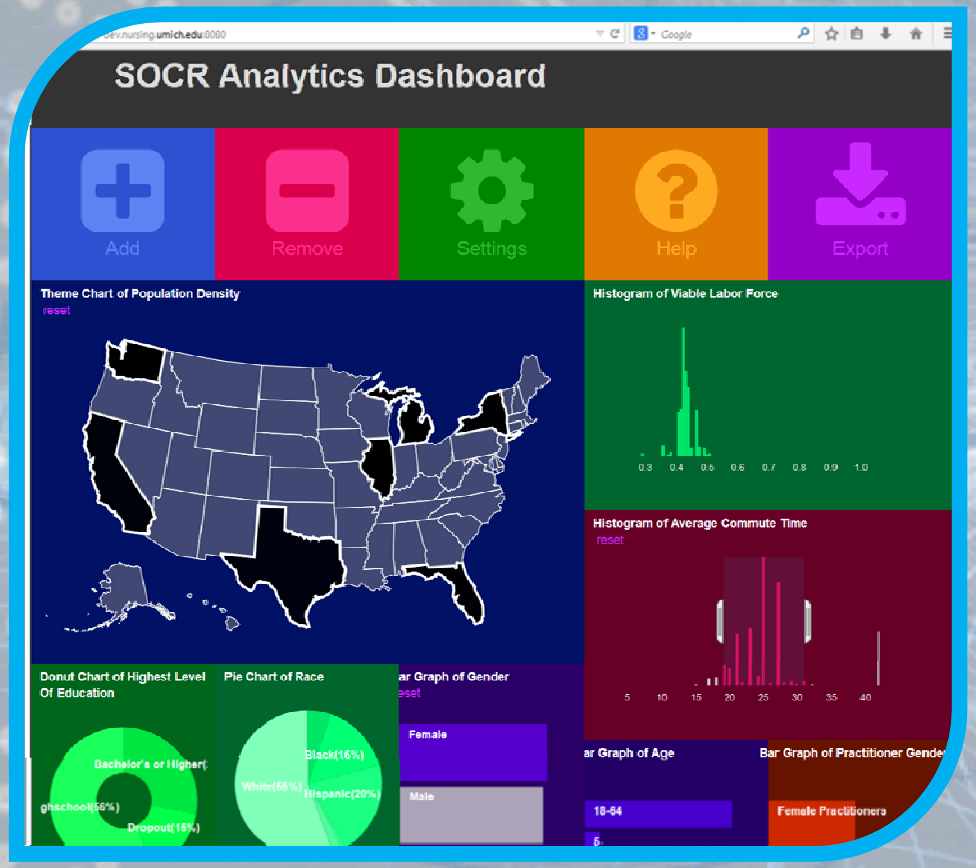
Husain, et al., 2015, J Big Data

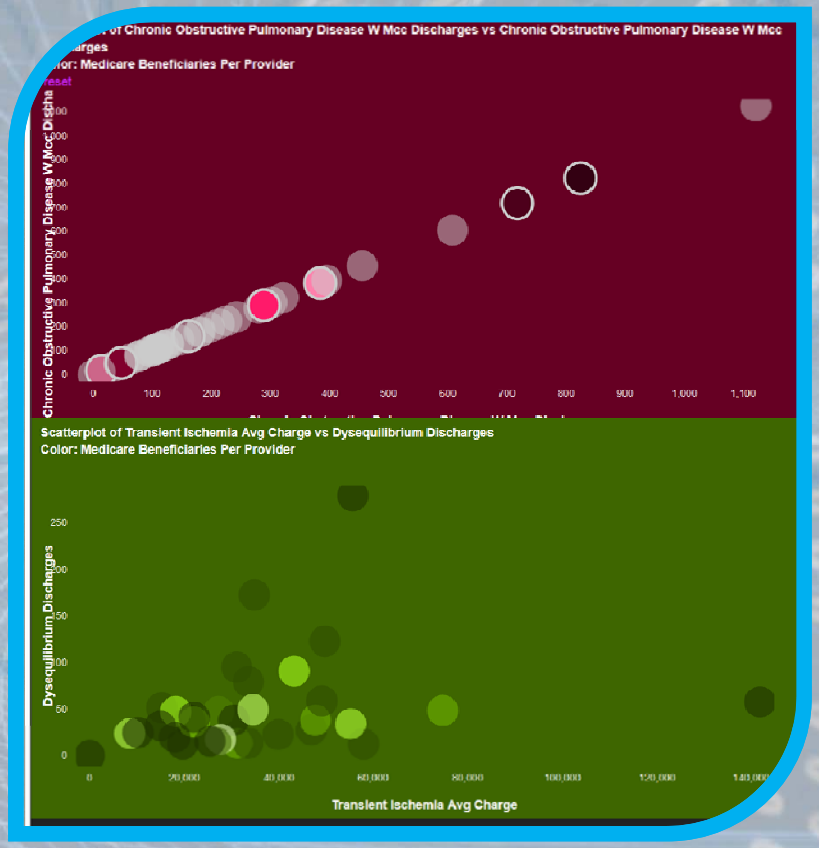
SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion



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

SOCR Dashboard (Exploratory Big Data Analytics): Associations





Compressive Big Data Analytics (CBDA)

- The foundation for Compressive Big Data Analytics (CBDA) involves
  - Iteratively generating 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.

- Repeating the (re)sampling and inference steps many times (with or without using the results of previous iterations as priors for subsequent steps).

Dinov, J Med Stat & Info, 2016

Big Data Analytics – Compressive Sensing

- Define the nested sets

$$S_k = \{x: \|x\|_o \stackrel{\text{def}}{=} |supp(x)| \leq k\},$$

where the data x, as a vector or tensor, has at most k non-trivial elements. Note that if $x, z \in S_k$, then $x + z \in S_{2k} \supseteq S_k$

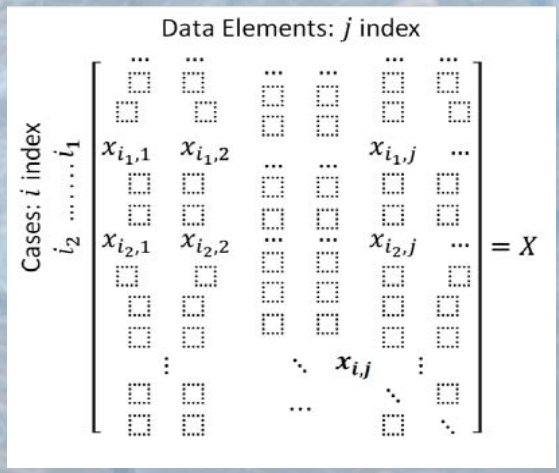
- If $\Phi_{n\times n} = (\varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_n)$ represents an orthonormal basis, the data may be expressed as $x = \Phi c$, where $c_i = \langle x, \varphi_i \rangle$, i.e., $c = \Phi^T x$, and $\|c\|_o \leq k$. Even if x is not strictly sparse, its representation c may be sparse. For each dataset, we can assess and quantify the error of approximating x by an optimal estimate $\hat{x} \in S_k$ by computing

$$\sigma_k(x)_p = \min_{\hat{x}\in S_k} \|x - \hat{x}\|_p$$

Big Data Analytics – Compressive Sensing

- In compressive sensing, if $x \in R^n$, and we have a data stream generating m linear measurements, we can represent $y = Ax$, where $A_{m \times n}$ is a dimensionality reducing matrix ($m \ll n$), i.e.,

$$A_{m \times n}: R^n \rightarrow R^m$$



- The null space of A,

  $N(A) = \{z \in R^n : Az = 0 \in R^m\}$.

  A uniquely represents all $x \in S_k \Leftrightarrow N(A)$ contains no vectors in S_{2k}.

- The spark of a matrix A represents the smallest number of columns of A that are linearly dependent. If $A_{m \times n}$ is a random matrix whose entries are independent and identically distributed, then $spark(A) = m + 1$, with probability 1.

Big Data Analytics – Compressive Sensing

- If the entries of A are chosen according to a sub-Gaussian distribution, then with high probability, for each k, there exists $\delta_{2k} \in (0,1)$ such that

$$(1-\delta_{2k})\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_{2k})\|x\|_2^2 \quad (1)$$

for all $x \in S_{2k}$ (RIP=Restricted isometry property)

- When we know that the original signal is sparse, to reconstruct x given the observed measurements y, we can solve the optimization problem:

$$\hat{x} = \arg \min_{z:Az=y} \|z\|_0$$

Big Data Analytics – Compressive Sensing

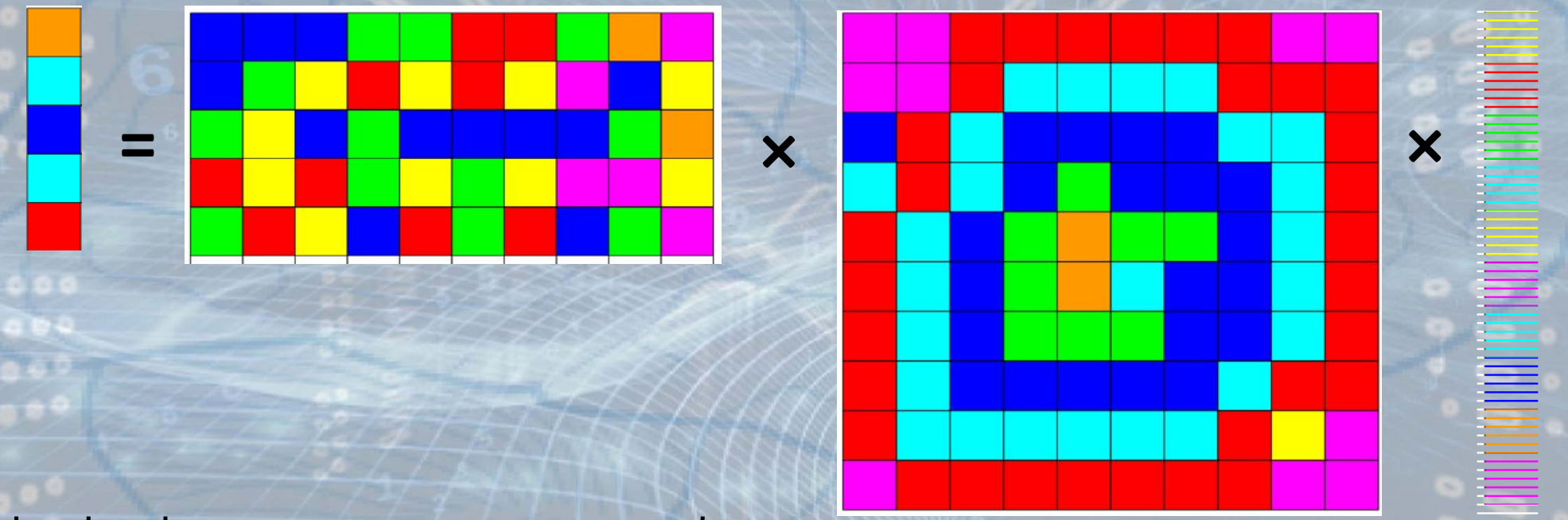
- Linear programming may be used to solve the optimization problem if we replace the zero-norm by its more tractable convex approximation, the l_1-norm, $\hat{x} = \arg \min_{z:Az=y} \|z\|_1$
- Given that $A_{m\times n}$ has the above property and $\delta_{2k} < \sqrt{2} - 1$, if we observe $y = Ax$, then the solution $\hat{x}$ satisfies

$$\|\hat{x} - x\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}}$$

- Thus, in compressive sensing applications, if $x \in S_k$ and A satisfies the RIP, condition (1), we can recover any k-sparse signal x exactly (as $\sigma_k(x)_1 = 0$) using only $O(k \log(n/k))$ observations, since $m = O\left(\frac{k \log(n/k)}{\delta_{2k}^2}\right)$
- Finally, if $A_{m\times n}$ is random (e.g., chosen according to a Gaussian distribution) and $\Phi_{n\times n}$ is an orthonormal basis, then $A_{m\times n} \times \Phi_{n\times n}$ will also have a Gaussian distribution, and if m is large, $A' = A \times \Phi$ will also satisfy the RIP condition (1) with high probability.

Big Data Analytics – Compressive Sensing

$Z_{N\times 1}$ = $D_{N\times M}$ × $S_{M\times T}$ × $X_{T\times 1}$



Obs

Proxy Prism (dim reducing matrix)

Orthonormal basis

Nat Process

Compressive Big Data Analytics (CBDA)

- The foundation for Compressive Big Data Analytics (CBDA) involves
  - Iteratively generating 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.

- Repeating the (re)sampling and inference steps many times (with or without using the results of previous iterations as priors for subsequent steps).

Dinov, J Med Stat & Info, *2016*

Compressive Big Data Analytics (CBDA)

- Bootstrapping techniques may be employed to quantify joint probabilities, estimate likelihoods, predict associations, identify trends, forecast future outcomes, or assess accuracy of findings.
- The goals of compressive sensing and compressive big data analytics are different.
  - CS aims to obtain a stochastic estimate of a complete dataset using sparsely sampled incomplete observations.
  - CBDA attempts to obtain a quantitative joint inference characterizing likelihoods, tendencies, prognoses, or relationships.
  - However, a common objective of both problem formulations is the optimality (e.g., reliability, consistency) of their corresponding estimates.

Compressive Big Data Analytics (CBDA)

- Suppose we represent (observed) Big Data as a large matrix $Y \in R^{n \times t}$, where n= sample size (instances) and t = elements (e.g., time, space, measurements, etc.)
- To formulate the problem in an *analytical framework*, let's assume $L \in R^{n \times t}$ is a low rank matrix representing the mean or background data features, $D \in R^{n \times m}$ is a (known or unknown) design or dictionary matrix, $S \in R^{m \times t}$ is a sparce parameter matrix with small support ($supp(S) \ll m \times t$), $E \in R^{n \times t}$ denote the model error term, and $\Lambda_{\Omega}(.)$ be a sampling operator generating incomplete data over the indexing pairs of instances and data elements $\Omega \subseteq \{1, 2, \dots, n\} \times \{1,2, \dots, t\}$
- In this generalized model setting, the problem formulation involves estimation of L, S (and D, if it is unknown), according to this model representation: $\Lambda_{\Omega}(Y) = \Lambda_{\Omega}(L + DS + E)$ (2)

Compressive Big Data Analytics (CBDA)

- Having quick, reliable and efficient estimates of L, S and D would allow us to make inference, compute likelihoods (e.g., p-values), predict trends, forecast outcomes, and adapt the model to obtain revised inference using new data

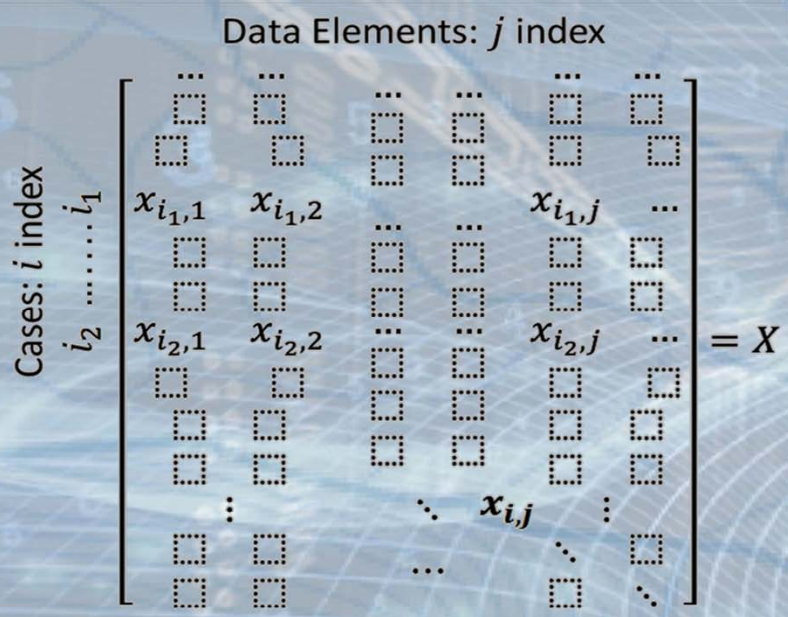
- When D is known, the model in equation (2) is jointly convex for L and S, and there exist iterative solvers based on sub-gradient recursion (e.g., alternating direction method of multipliers)

- However, in practice, the size of Big Datasets presents significant computational problems, related to slow algorithm convergence, for estimating these components that are critical for the final study inference

Compressive Big Data Analytics (CBDA)

- One strategy for tackling this optimization problem is to use a random Gaussian sub-sampling matrix $A_{m\times n}$ (much like in the compressive sensing protocol) to reduce the rank of the observed data ($Y_{m\times l}$, where $(m, l) \in \Omega$) and then solve the minimization using least squares
- This *partitioning* of the difficult general problem into smaller chunks has several advantages. It reduces the hardware and computational burden, enables algorithmic parallelization of the global solution, and ensures feasibility of the analytical results
- Because of the stochastic nature of the index sampling, this approach may have desirable analytical properties like predictable asymptotic behavior, limited error bounds, estimates' optimality and consistency characteristics

Compressive Big Data Analytics (CBDA)



Data Structure (Representation)

Sample Data (Instance)

Compressive Big Data Analytics (CBDA)

- One can design an algorithm that searches and keeps only the most informative data elements by requiring that the derived estimates represent optimal approximations to y within a specific sampling index subspace $\{(m,l)\} \subseteq \Omega$

- We want to investigate if CBDA inference estimates can be shown to obey error bounds similar to the upper bound results of point imbedding's in high-dimensions (e.g., Johnson-Lindenstrauss lemma) or the restricted isometry property

Compressive Big Data Analytics (CBDA)

- The Johnson-Lindenstrauss lemma guarantees that for any $0 < \varepsilon < 1$, a set of points $\{P_k\}_1^K \in R^n$ can be linearly embedded $\left(\Psi: R^n \rightarrow R^{n'}\right)$ into $\{\Psi(P_k) = P_k'\}_1^K \in R^{n'}$, for $\forall\, n' \geq 4\left(\frac{ln(K)}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}}\right)$, almost preserving their pairwise distances, i.e., $(1-\epsilon)\left\|P_i - P_j\right\|_2^2 \leq \left\|P_i' - P_j'\right\|_2^2 \leq (1+\epsilon)\left\|P_i - P_j\right\|_2^2$
- The restricted isometry property ensures that if $\delta_{2k} < \sqrt{2} - 1$ and the estimate $\hat{x} = \arg\min_{z:Az=y} \|z\|_1$, where $A_{m \times n}$ satisfies property (1), then the data reconstruction is reasonable, i.e., $\|\hat{x} - x\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}}$
- Can we develop iterative space-partitioning CBDA algorithms that either converge to a fix point or generate estimates that are *close* to their corresponding inferential parameters?

Acknowledgments

http://SOCR.umich.edu

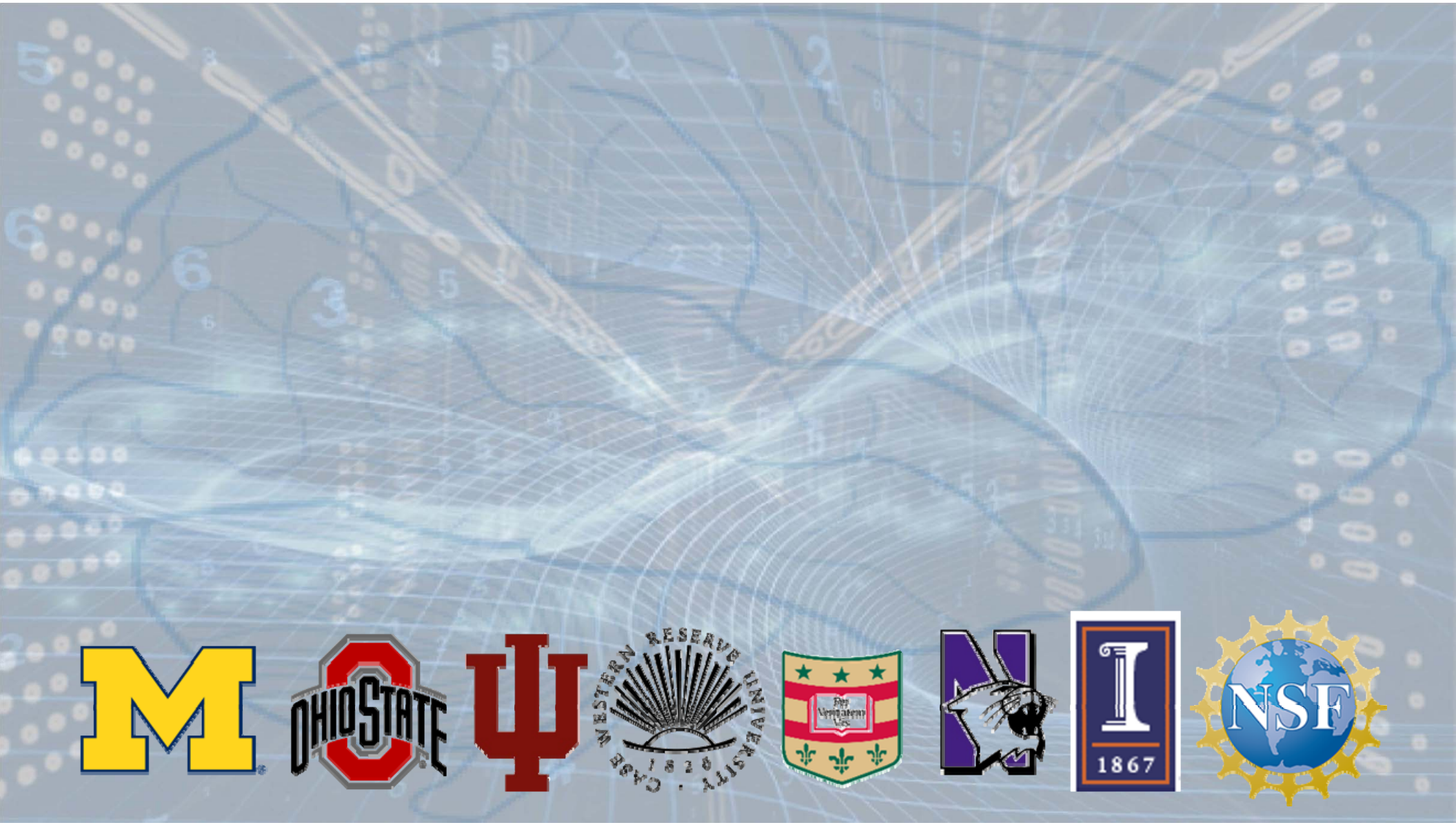
Funding

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

NSF: 1636840, 1416953, 0716055, 1023115

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
- **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, Fabio Macciardi, Federica Torri, Carl Kesselman,
- **UMich AD/PD Centers**: Cathie Spino, Brian, Athey, Hank Paulson, Ben Hampstead, Bill Dauer


OHIO STATE
WESTERN RESERVE UNIVERSITY
CASE
1826
1867
NSF