Predictive Big Data Analytics:

Using Large, Complex, Heterogeneous, Incongruent, Multi-source & Incomplete Observations to Study Neurodegenerative Disorders

Ivo D. Dinov

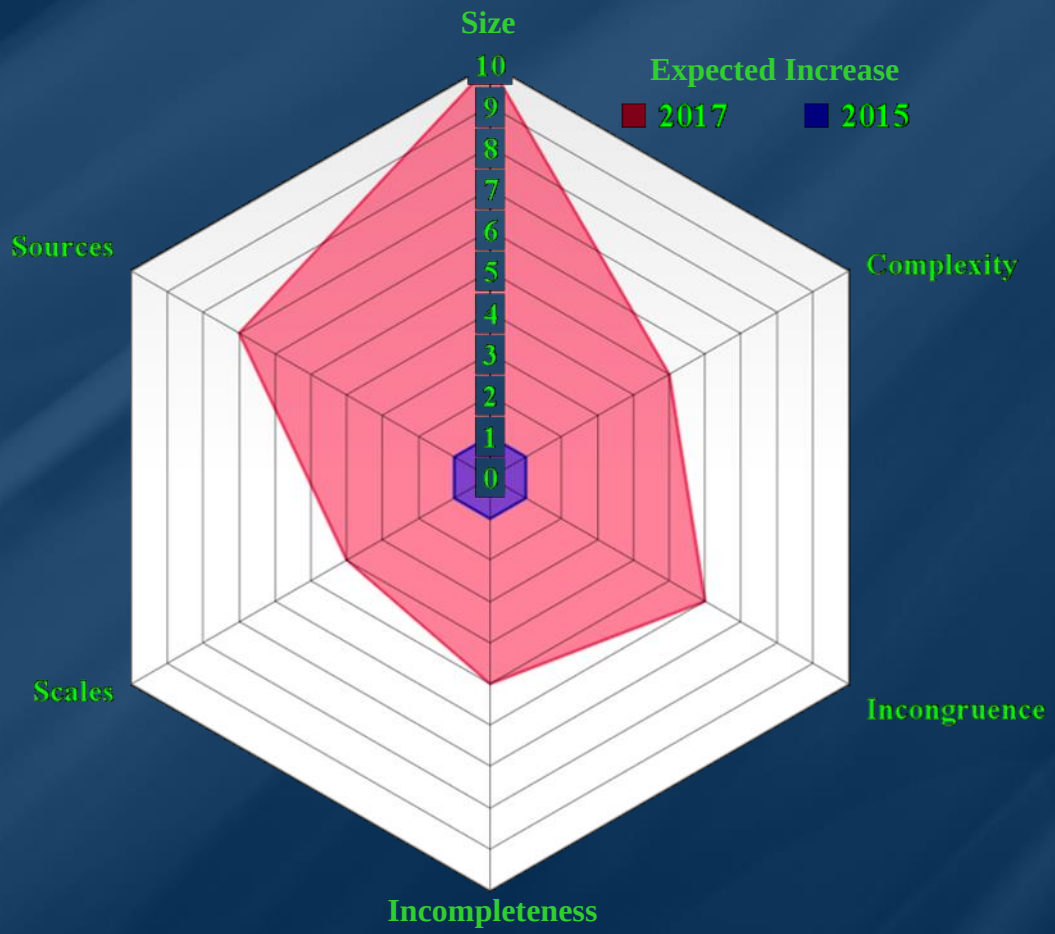
Statistics Online Computational Resource
Michigan Institute for Data Science
School of Nursing
University of Michigan

www.SOCR.umich.edu

Outline

- Big, Deep & Dark Data
- Big Data Analytics
- Applications to Neurodegenerative Disease
- Data Dashboarding
- Open Problem: Representation and Joint BD Analytics – Compressive Big Data Analytics (CBDA)

Characteristics of Big Biomed 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.

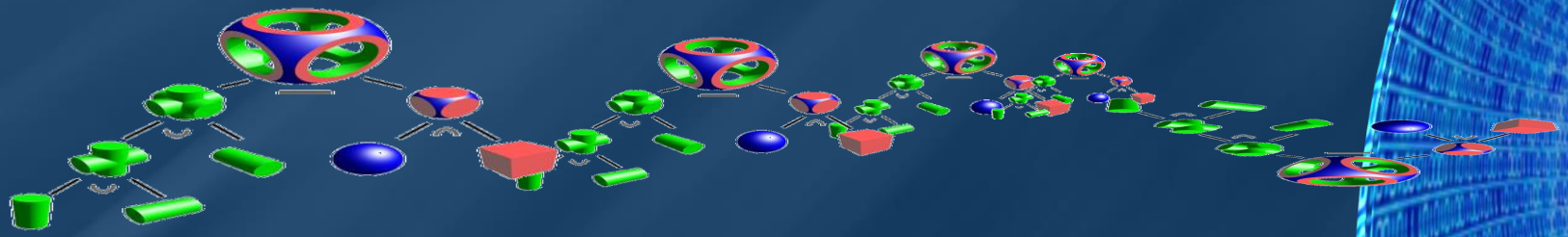
Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

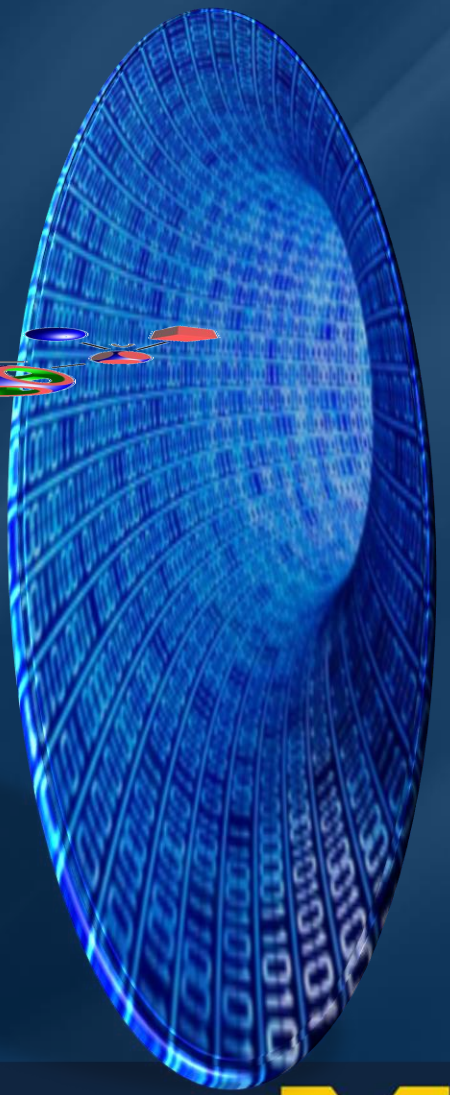
Mixture of quantitative & qualitative estimates

Dinov, et al., 2014

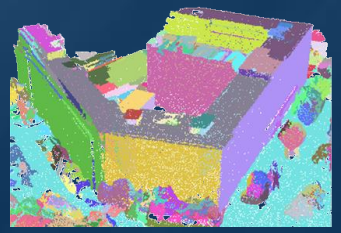
Big, Deep & Dark Data

- **Big**: Size + Complexity + Variability + Heterogeneous

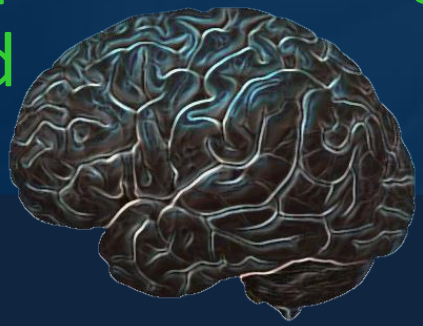




- **Deep**: Covering + Multi-source + Multi-dimensional



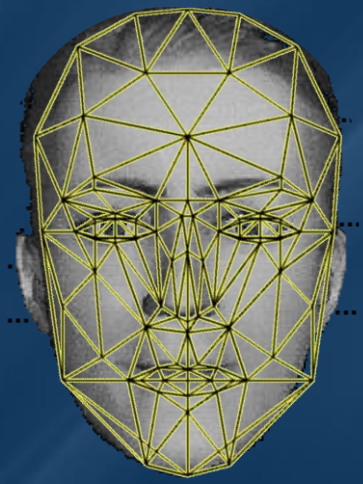
- **Dark**: Incomplete + Incongruent + Unconcealed

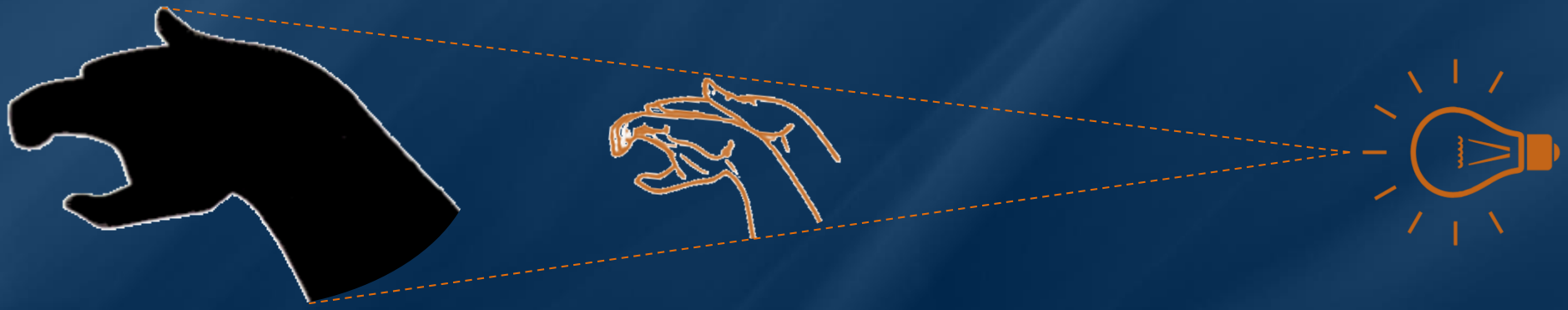


Data: Process & Model Representation

Native Process

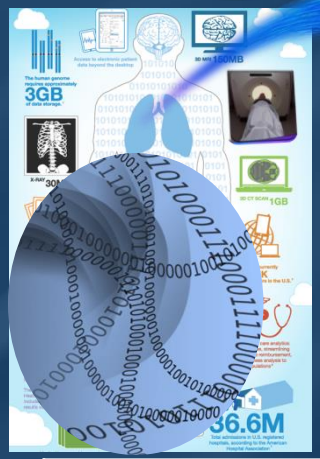
Model Representation



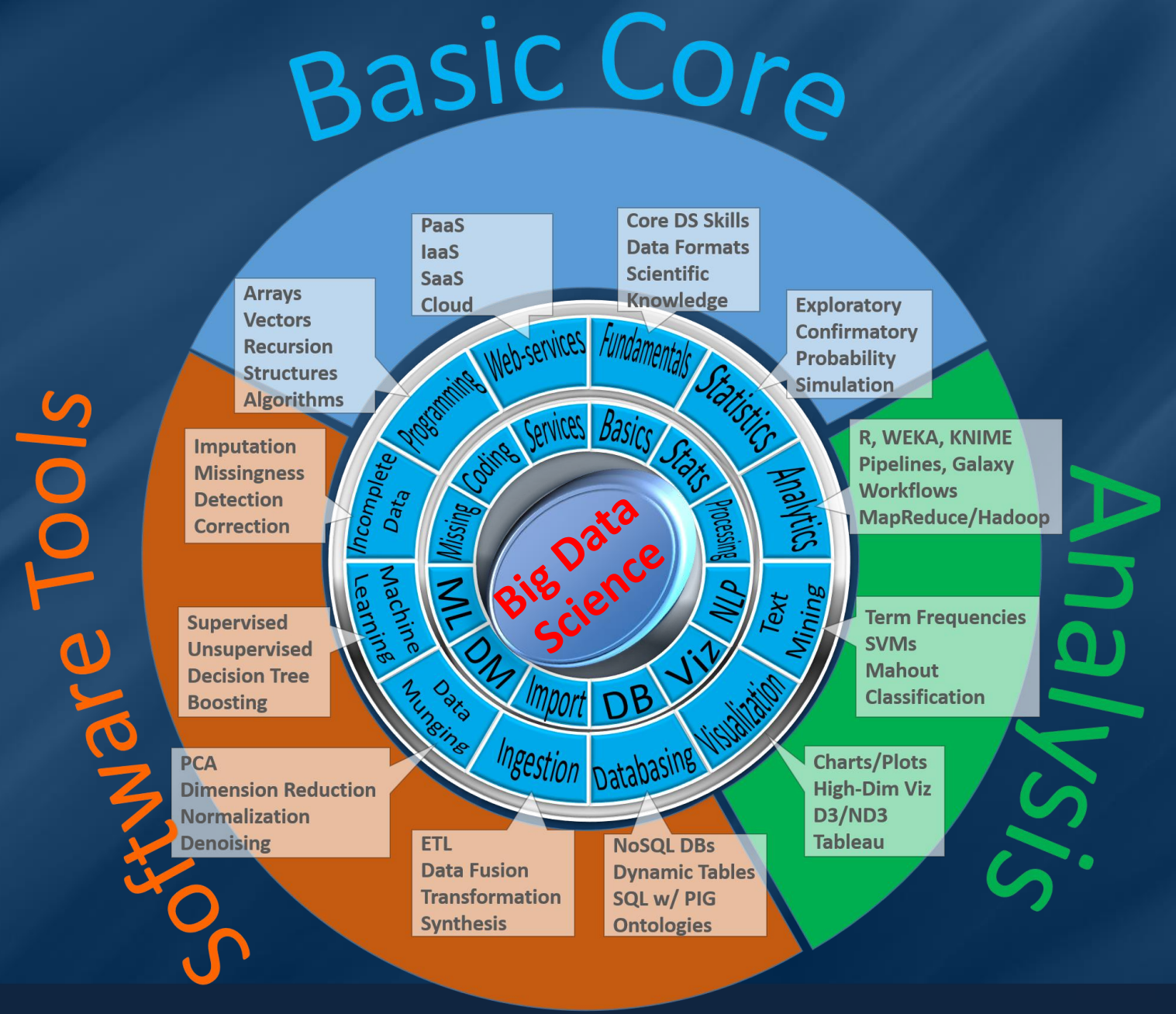


Observation

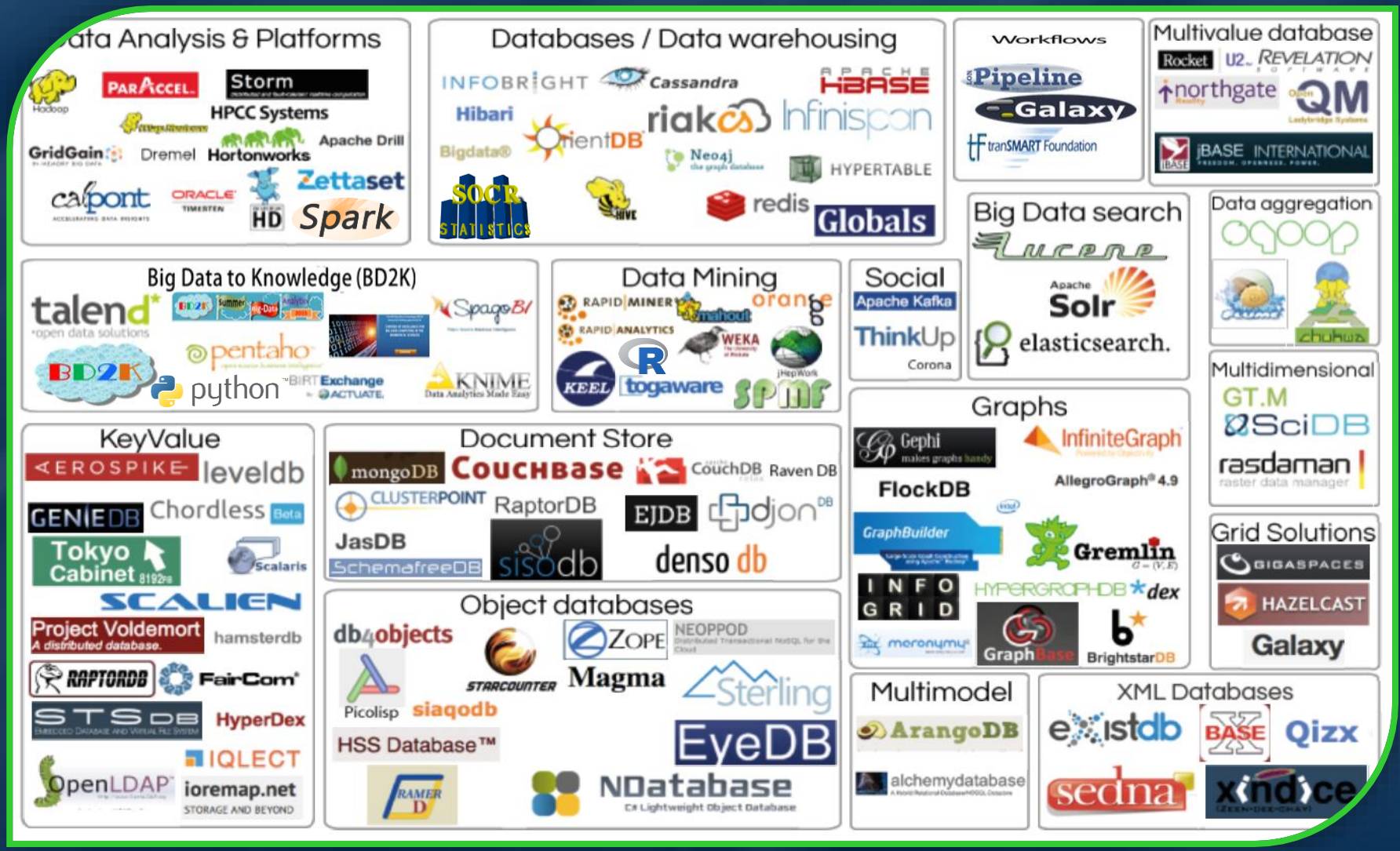
Source



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

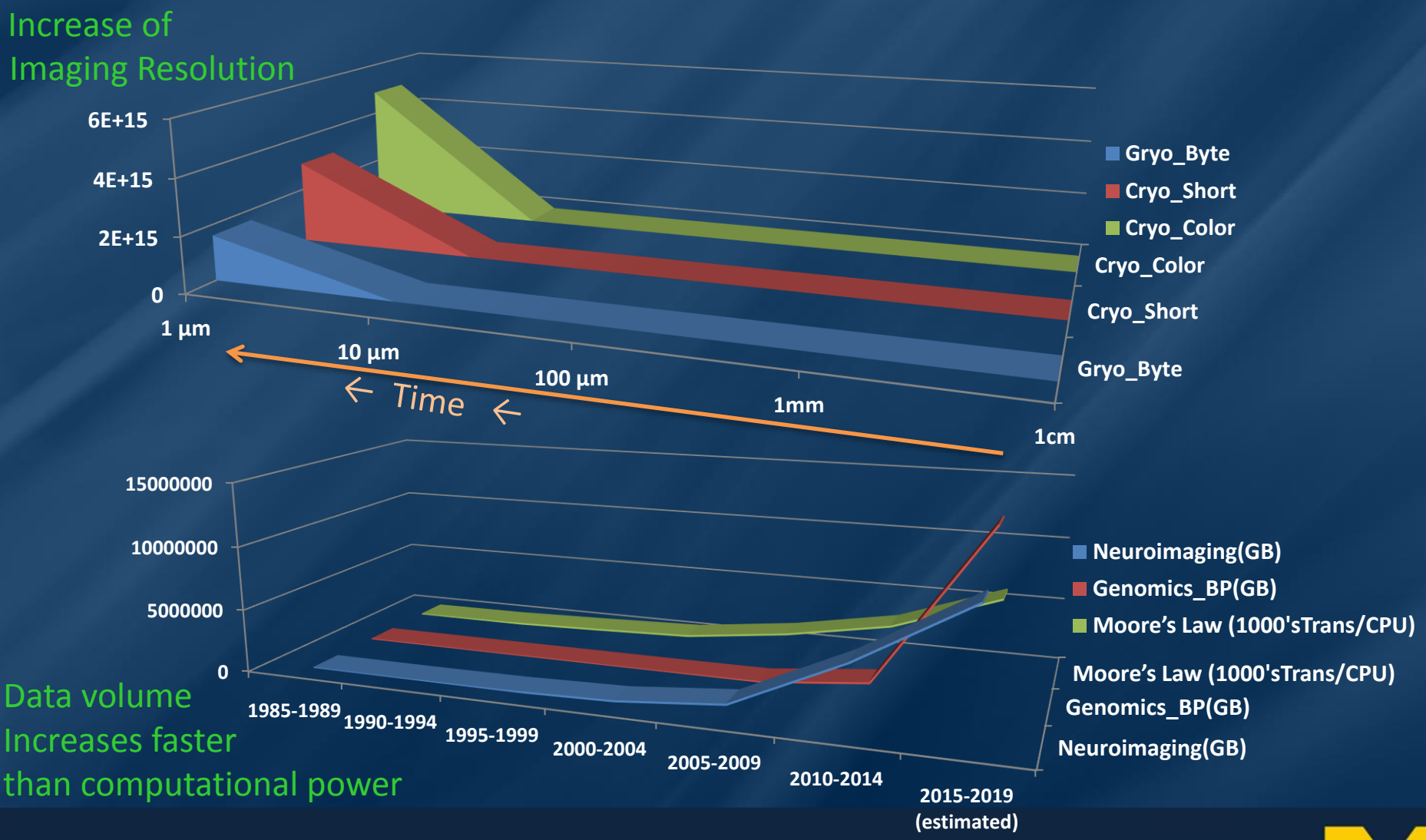

Basic Core
Software Tools
Analysis
Big Data Science
Web-services
Fundamentals
Statistics
Analytics
Text Mining
Visualization
Databasing
Ingestion
Data Munging
Machine Learning
Incomplete Data
Programming
Services
Basics
Stats
Processing
NLP
Viz
DB
Import
DM
ML
Missing
Coding
PaaS
IaaS
SaaS
Cloud
Core DS Skills
Data Formats
Scientific
Knowledge
Arrays
Vectors
Recursion
Structures
Algorithms
Exploratory
Confirmatory
Probability
Simulation
R, WEKA, KNIME
Pipelines, Galaxy
Workflows
MapReduce/Hadoop
Imputation
Missingness
Detection
Correction
Term Frequencies
SVMs
Mahout
Classification
Supervised
Unsupervised
Decision Tree
Boosting
Charts/Plots
High-Dim Viz
D3/ND3
Tableau
PCA
Dimension Reduction
Normalization
Denoising
ETL
Data Fusion
Transformation
Synthesis
NoSQL DBs
Dynamic Tables
SQL w/ PIG
Ontologies

Big Data Analytics Resourceome



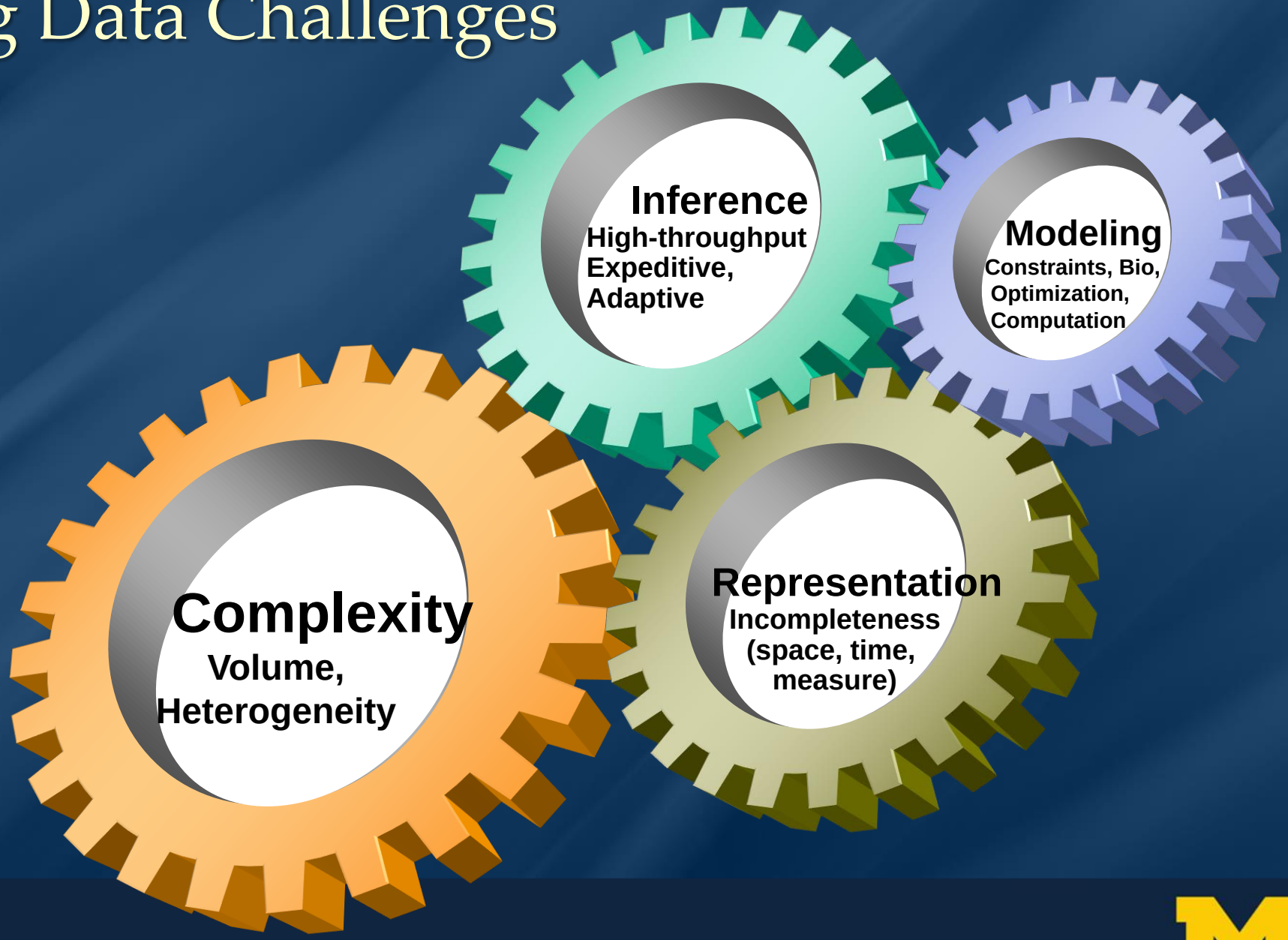
http://socr.umich.edu/docs/BD2K/BigDataResourceome.html

Kryder's law: Exponential Growth of Data



Dinov, et al., 2014

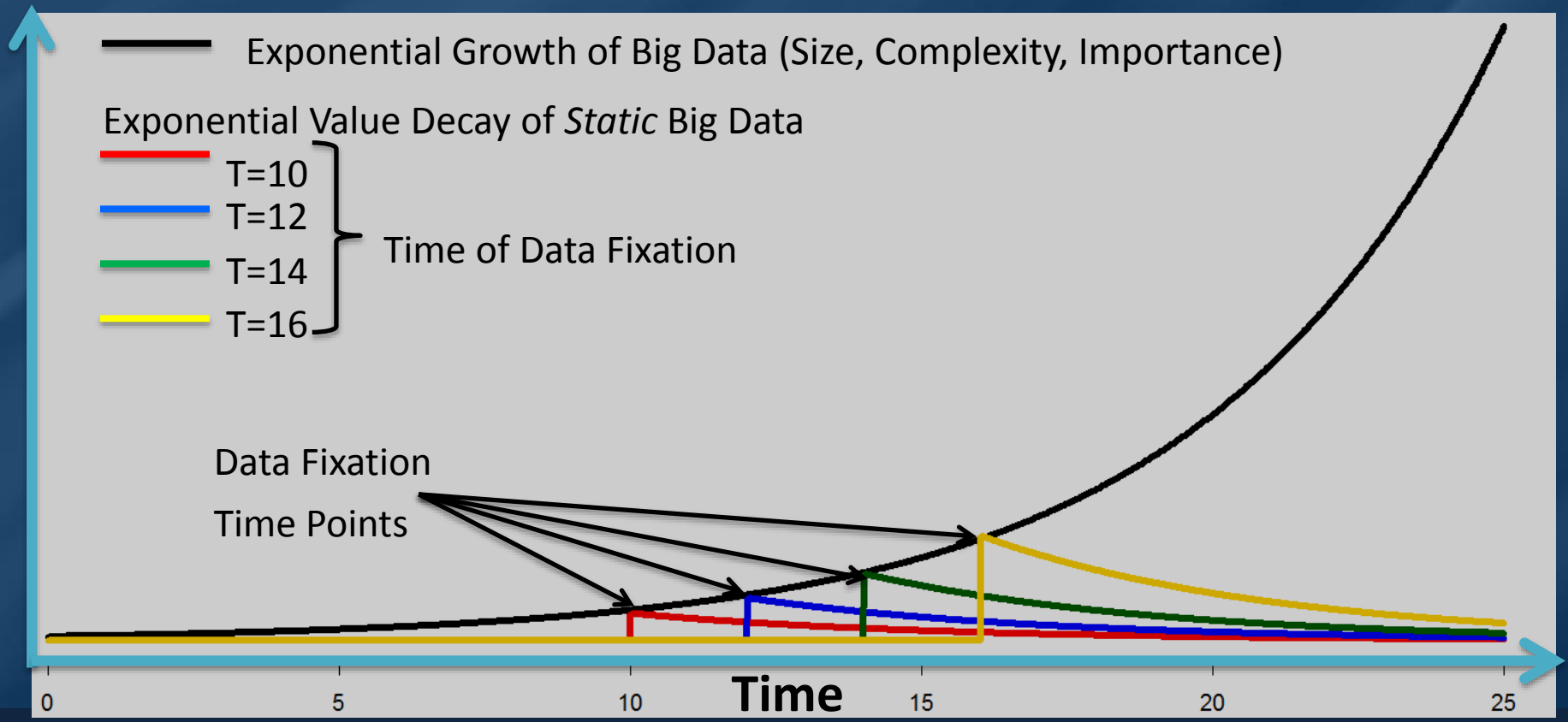
Big Data Challenges



Energy & Life-Span of Big Data

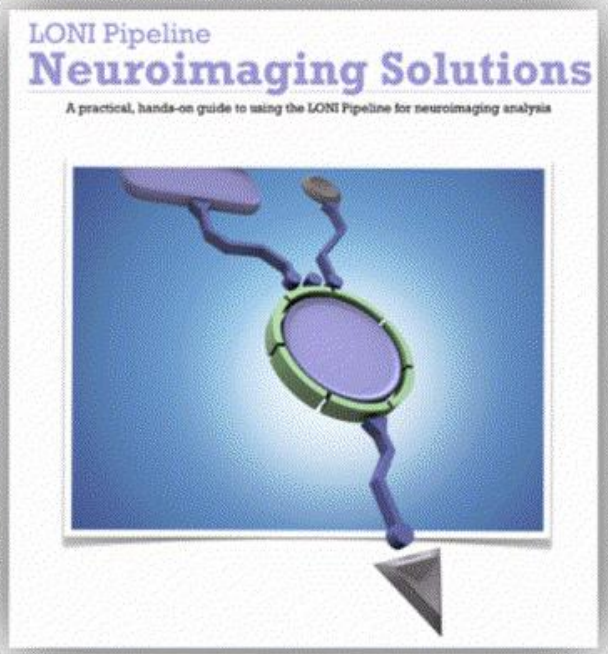
Energy encapsulates the holistic information content included in the data (exponentially increasing with time), which may often represent a significant portion of the joint distribution of the underlying healthcare process

Life-span refers to the relevance and importance of the Data (exponentially decreasing with time)



http://www.aaas.org/news/big-data-blog-part-v-interview-dr-ivo-dinov

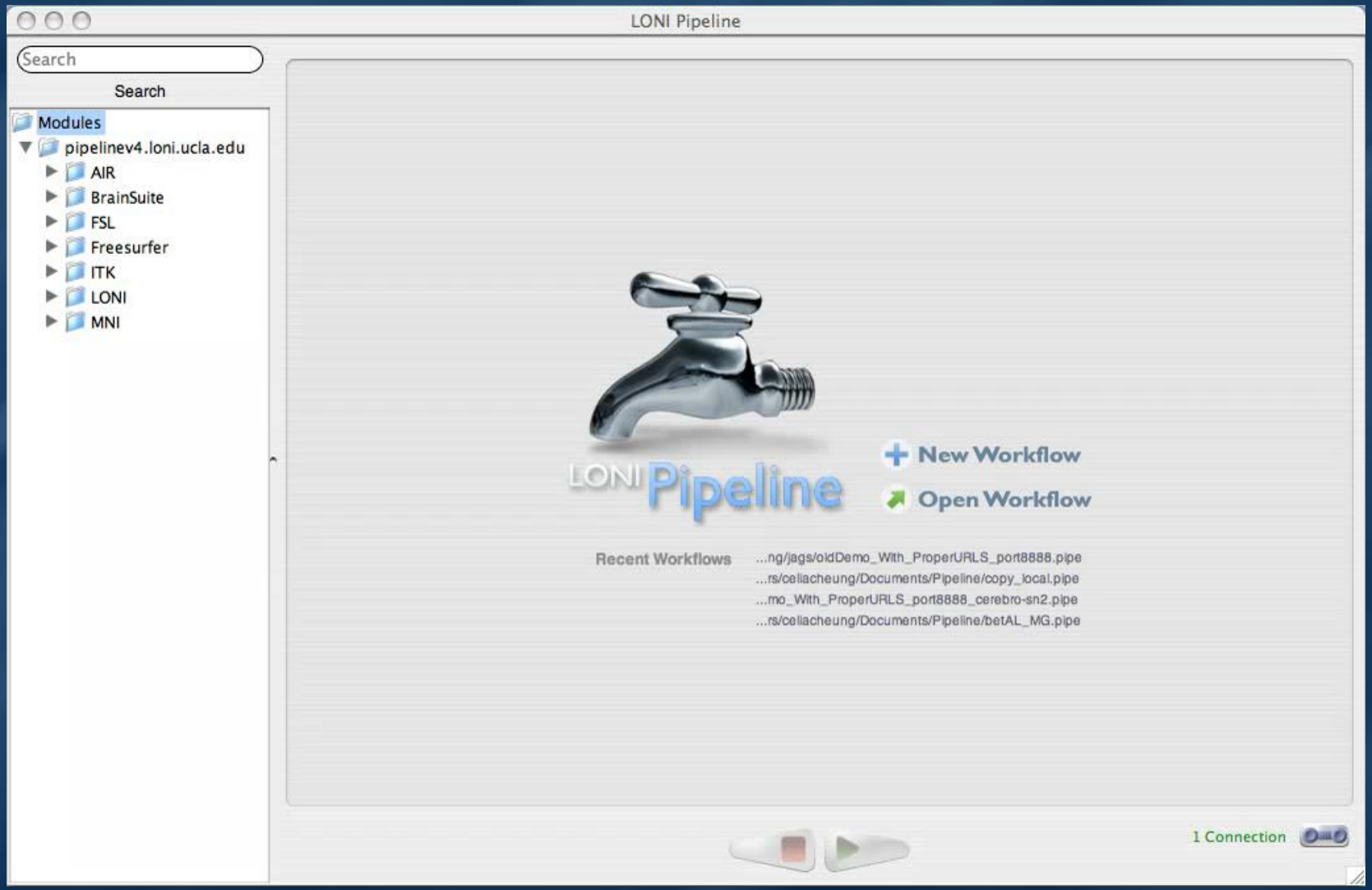
End-to-end Pipeline Workflow Solutions



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

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

End-to-end Pipeline Workflow Solutions



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

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

Neurodegenerative Disease Applications:

Imaging-Genetic Biomarker Interactions in Alzheimer's Disease

Goals

Investigate AD subjects (age 55 – 65) using Neuroimaging Initiative (ADNI) database to understand early-onset (EO) cognitive impairment using neuroimaging and genetics biomarkers

Data

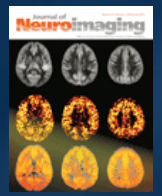
EO-AD and EO-MCI (mild cognitive impairment)

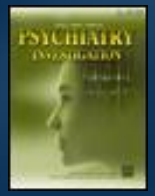
Derived 15 most impactful neuroimaging markers (out of 336 morphometry measures)

Obtained 20 most significant single nucleotide polymorphisms (SNPs) associated with specific neuroimaging biomarkers (out of 620K SNPs)

Approach

Global Shape Analysis (GSA) Pipeline workflow





Moon, Dinov et al., 2015, J Neuroimaging
Moon, Dinov et al., 2015, Psych. Investig.

Results: AD Imaging-Genetic Biomarker Interactions

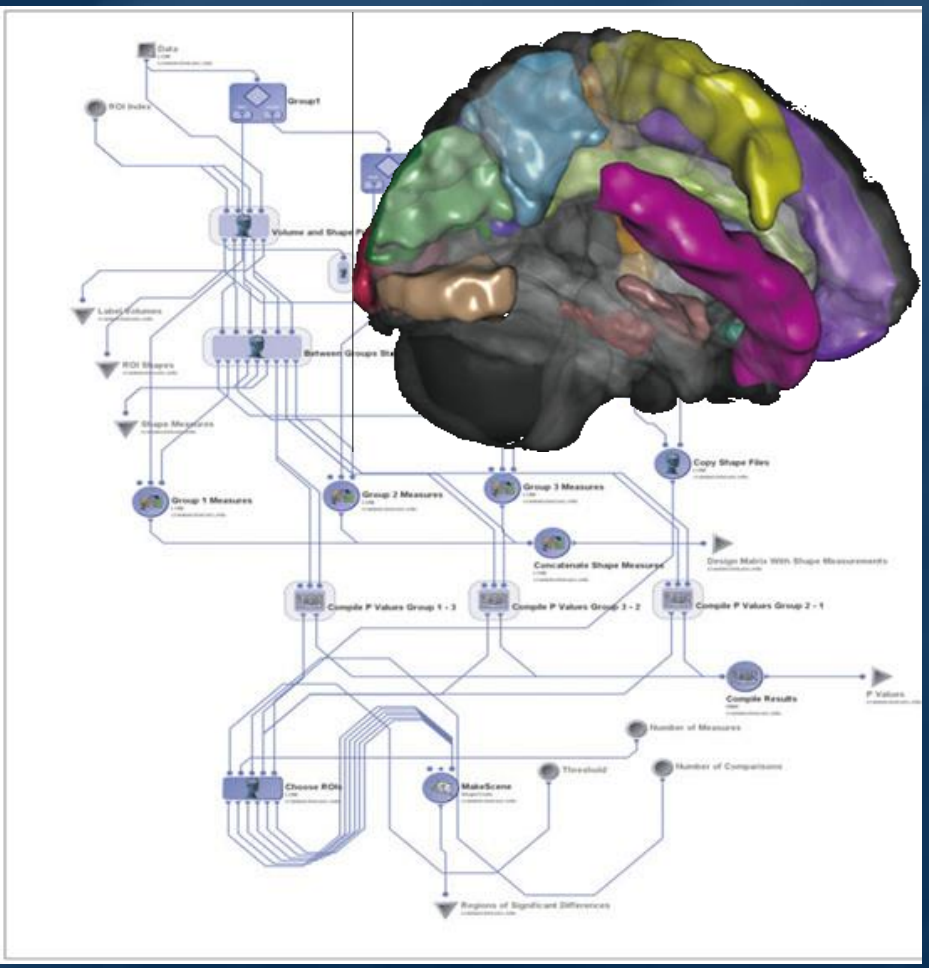
- Identified associations between neuroimaging phenotypes and genotypes for the EO cohort

- Overall most significant associations:
  - rs7718456 (Chr 15) and L_hippocampus (volume)
  - rs7718456 and R_hippocampus (volume)

- For the 27 EO-MCI's, most significant associations
  - rs6446443 (Chr 4, JAKMIP1 janus kinase and microtubule interacting protein 1 gene) and R_superior_frontal_gyrus (volume)
  - rs17029131 (Chr 2) and L_Precuneus (volume)

Moon, Dinov et al., 2015, Psych. Investig.

Results: AD Imaging-Genetic Biomarker Interactions



Shape Index

$$SI = \frac{2}{\pi} \arctan\left(\frac{\kappa_2 + \kappa_1}{\kappa_2 - \kappa_1}\right)$$

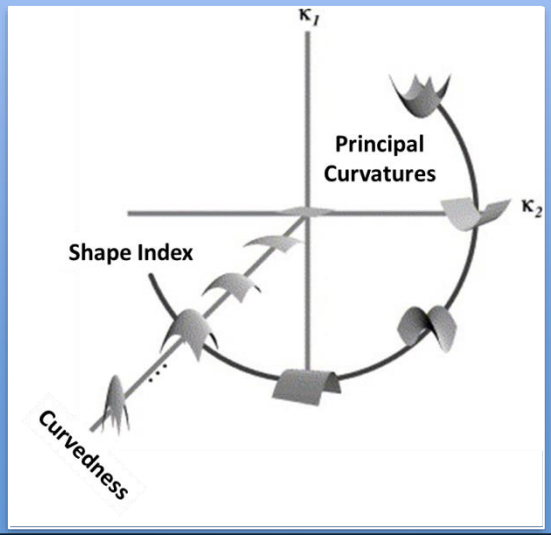
$$-1 \leq SI \leq 1$$

Cup	Rut	Saddle	Ridge	Cap
-1	-0.5	0	0.5	+1

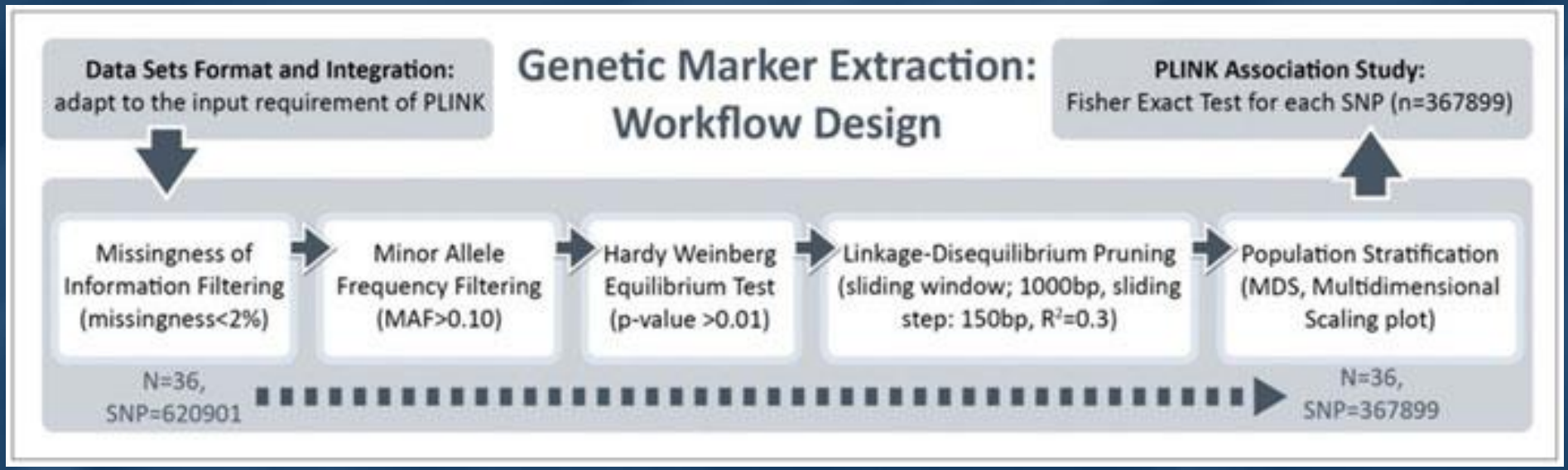
Curvedness

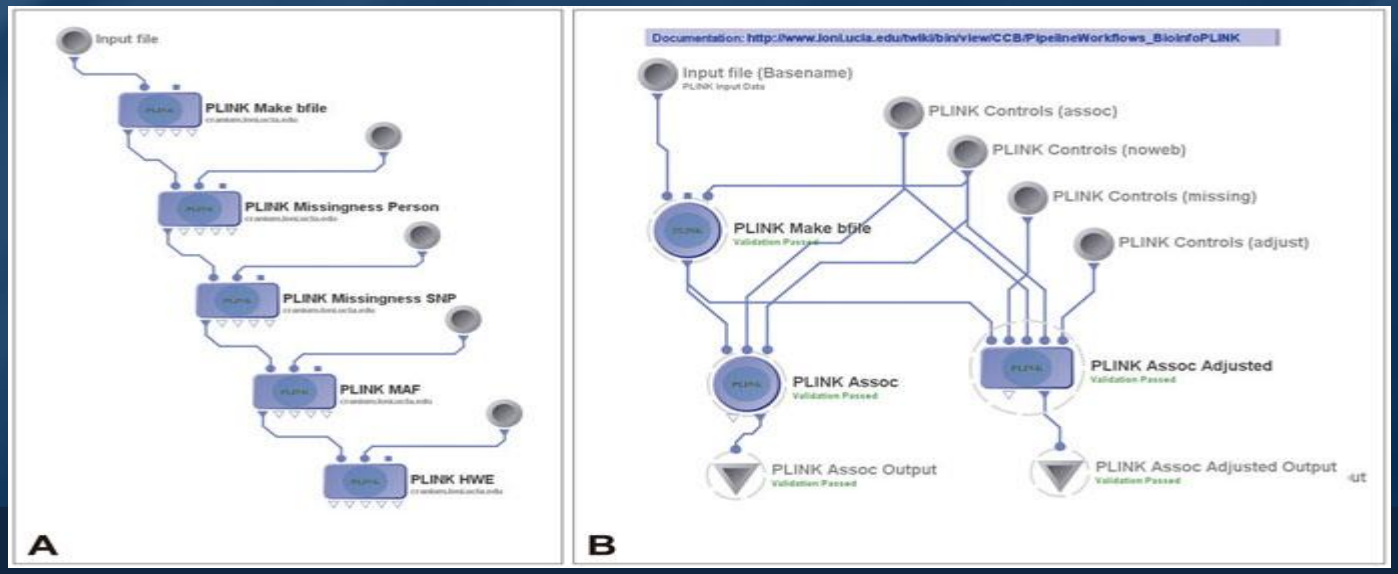
$$CV = \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}}$$

$$0 \leq CV$$



Results: AD Imaging-Genetic Biomarker Interactions



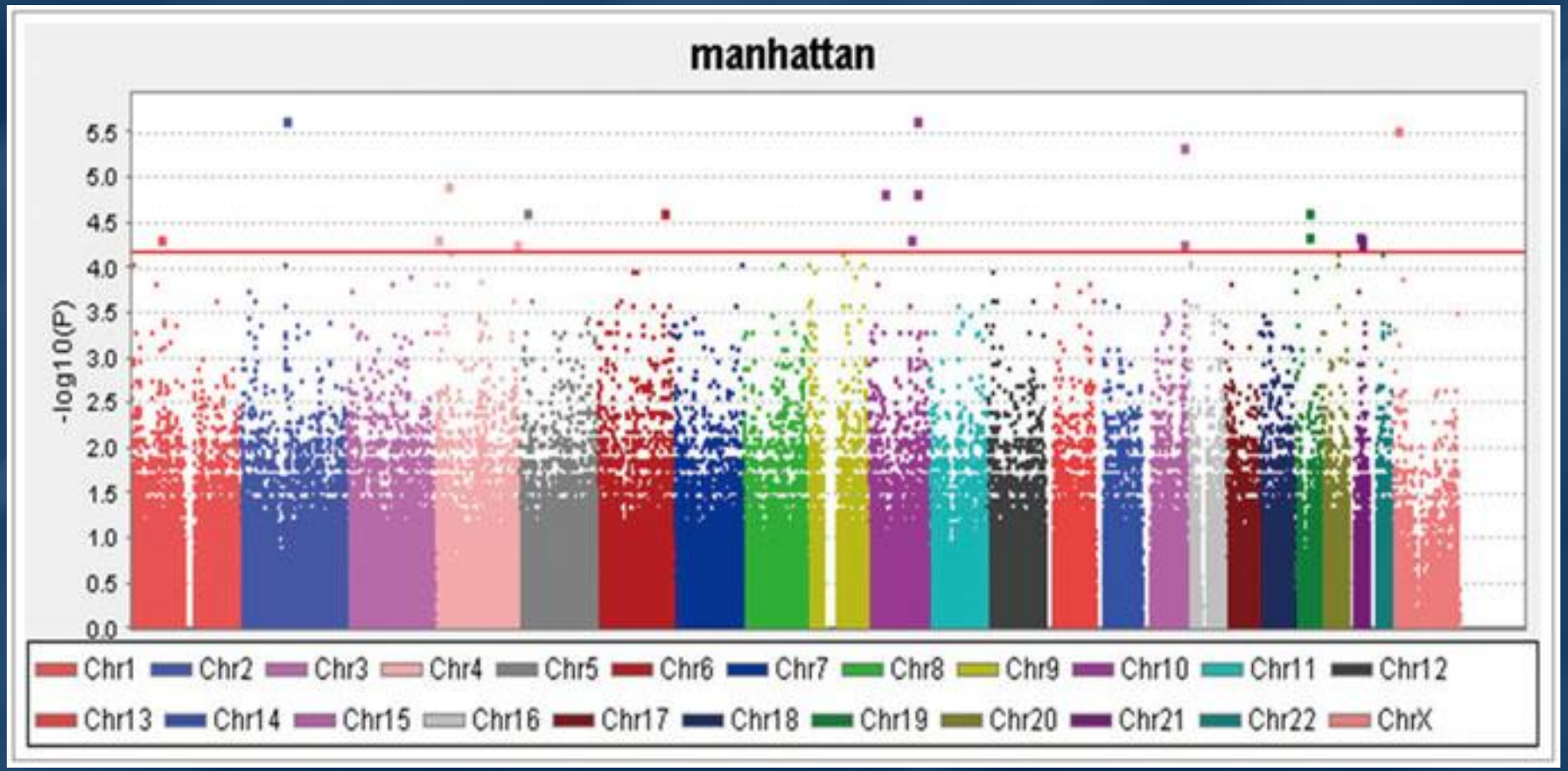


QC Protocol

A. QC Plink workflow
B. Genetic association study

http://Pipeline.loni.usc.edu

Results: AD Imaging-Genetic Biomarker Interactions



Manhattan plot for all the single nucleotide polymorphisms (SNPs)

Results: AD Imaging-Genetic Biomarker Interactions

Neuroimaging phenotypes	p-value	Index	SNPs	Chromosome	p-value	Gene
L_cingulate_gyrus (Average mean curvature)	0.0335	1	rs17029131	2	3.52E-06	
L_gyrus_rectus (Surface area)	0.01728	2	rs1822144	2	2.28E-06	
R_cuneus (Surface area)	0.04203	3	rs6446443	4	6.68E-05	JAKMIP1 (janus kinase & microtubule interacting protein 1)
R_superior_frontal_gyrus (Volume)	0.03706	4	rs12506164	4	1.75E-05	
L_precentral_gyrus (Volume)	0.04125	5	rs7718456	5	3.36E-05	
L_precuneus (Volume)	0.0508	6	rs9377090	6	3.36E-05	
L_middle_occipital_gyrus (Volume)	0.01805	7	rs2776932	10	2.20E-05	NRP1 (neuropilin1)
R_superior_temporal_gyrus (Volume)	0.03353	8	rs4933672	10	6.48E-05	
L_hippocampus (Volume)	0.00067	9	rs11193270	10	3.52E-06	
R_hippocampus (Volume)	0.00539	10	rs11193272	10	3.52E-06	
R_precentral_gyrus (Shape index)	0.03411	11	rs11193274	10	3.52E-06	
R_precuneus (Shape index)	0.03186	12	rs12218153	10	3.52E-06	
L_cuneus (Shape index)	0.04952	13	rs1338956	10	2.20E-05	
R_inferior_occipital_gyrus (Curviness)	0.05037	14	rs1338025	10	2.20E-05	
R_putamen (Curviness)	0.03504	15	rs12101936	15	7.08E-06	
		16	rs16964473	19	3.53E-05	Intergenic
		17	rs12972537	19	6.14E-05	
		18	rs2212356	21	6.23E-05	
		19	rs2831165	21	6.68E-05	
		20	rs1266320	23	4.46E-06	

GWAS Imaging-Genetics Approach

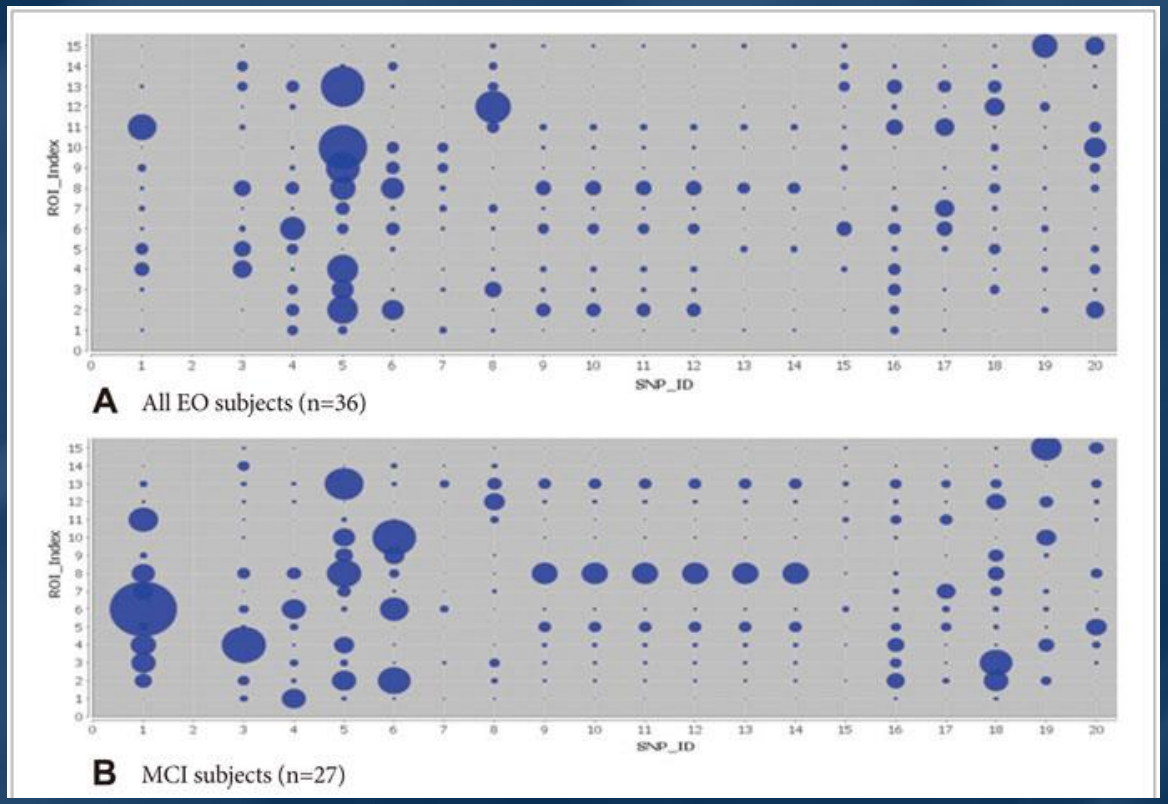
- **SNPs**
  - ***E.g., C/T polymorphism***
- Model
  - ***Phenotype:*** Y_i be the imaging-biomarker for i^{th} *subject*
  - ***Genotype:*** X_i be the genotype i^{th} *subject* at a particular:

$$\text{SNP } X_i = \begin{cases} 0, & BB \\ 1, & BA \\ 2, & AA \end{cases}$$

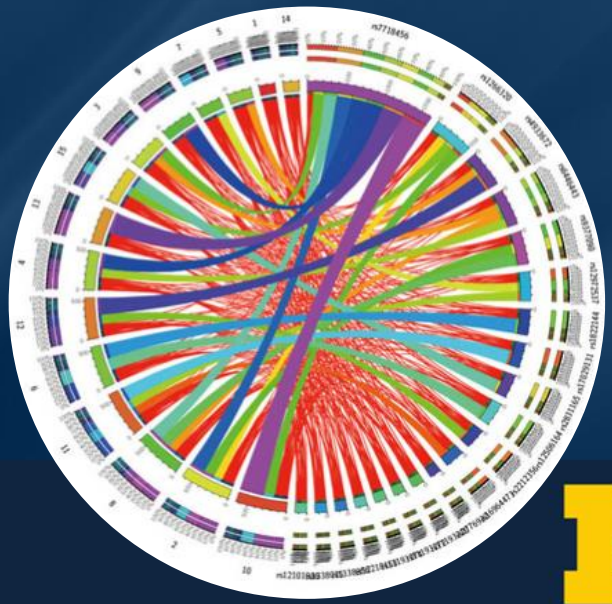
- SOCR Multivariate Regression Models
  - $Y_i = \beta_0 + \beta_1 X_i$
  - $In\ general, Y_i = \sum_{k=0}^{K} \beta_k X_i^{(k)} + \varepsilon$
  - Stat analysis: $\boldsymbol{\beta_k \neq 0}$

Genotype-Phenotype Relation			
Parents		♂	
		B	A
♀	B	BB	BA
	A	BA	AA

SNP-Neuroimaging interactions in Alzheimer's Disease



- Overall Associations
  - rs7718456 and L_hippocampus (volume)
  - rs7718456 and R_hippocampus (volume)
- EO-MCI associations
  - rs6446443 R_superior_frontal_gyrus (volume)
  - rs17029131 and L_Precuneus (volume)

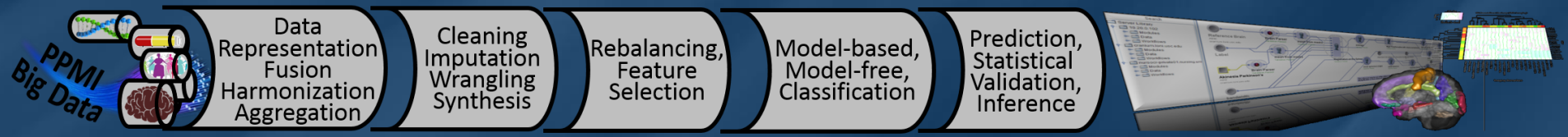


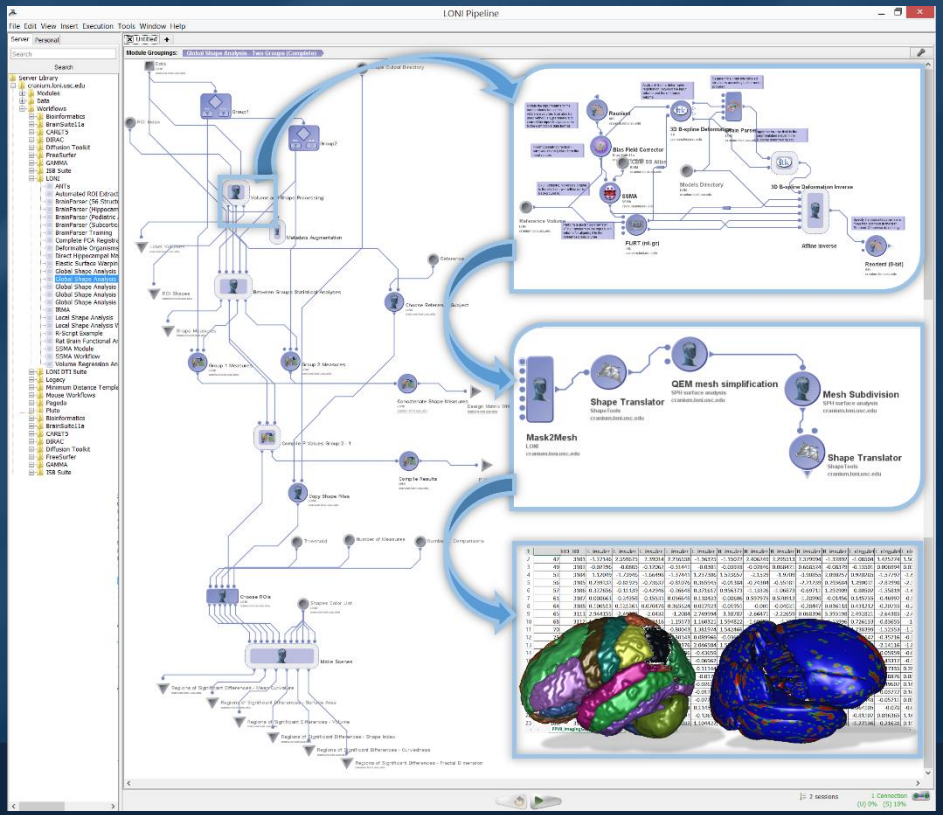
Predictive Big Data Analytics in Parkinson's Disease

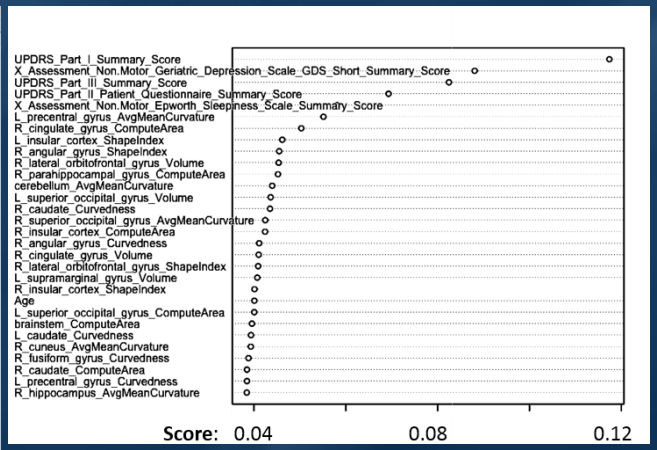
- A **unique archive of Big Data**: Parkinson's Progression Markers Initiative (PPMI). Data characteristics – large size, incongruency, incompleteness, complexity, multiplicity of scales, and heterogeneity of information-generating sources
- **Approach**
  - introduce methods for rebalancing imbalanced cohorts,
  - utilize a wide spectrum of classification methods to generate consistent and powerful phenotypic predictions,
  - generate reproducible machine-learning based classification that enables the reporting of model parameters and diagnostic forecasting based on new data.
- **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 statistical n-fold cross-validation).
- **Model-free** machine learning-based classification methods (e.g., adaptive boosting, support vector machines) outperform **model-based techniques** in terms of predictive precision and reliability (e.g., forecasting patient diagnosis).

Predictive Big Data Analytics using Large, Complex, Incongruent, Heterogeneous, Multi-source & Incomplete Observations

- **A Big Data Study of Parkinson's Disease**







Varplot illustrating:

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

AdaBoost classifier for Controls vs. Patients prediction

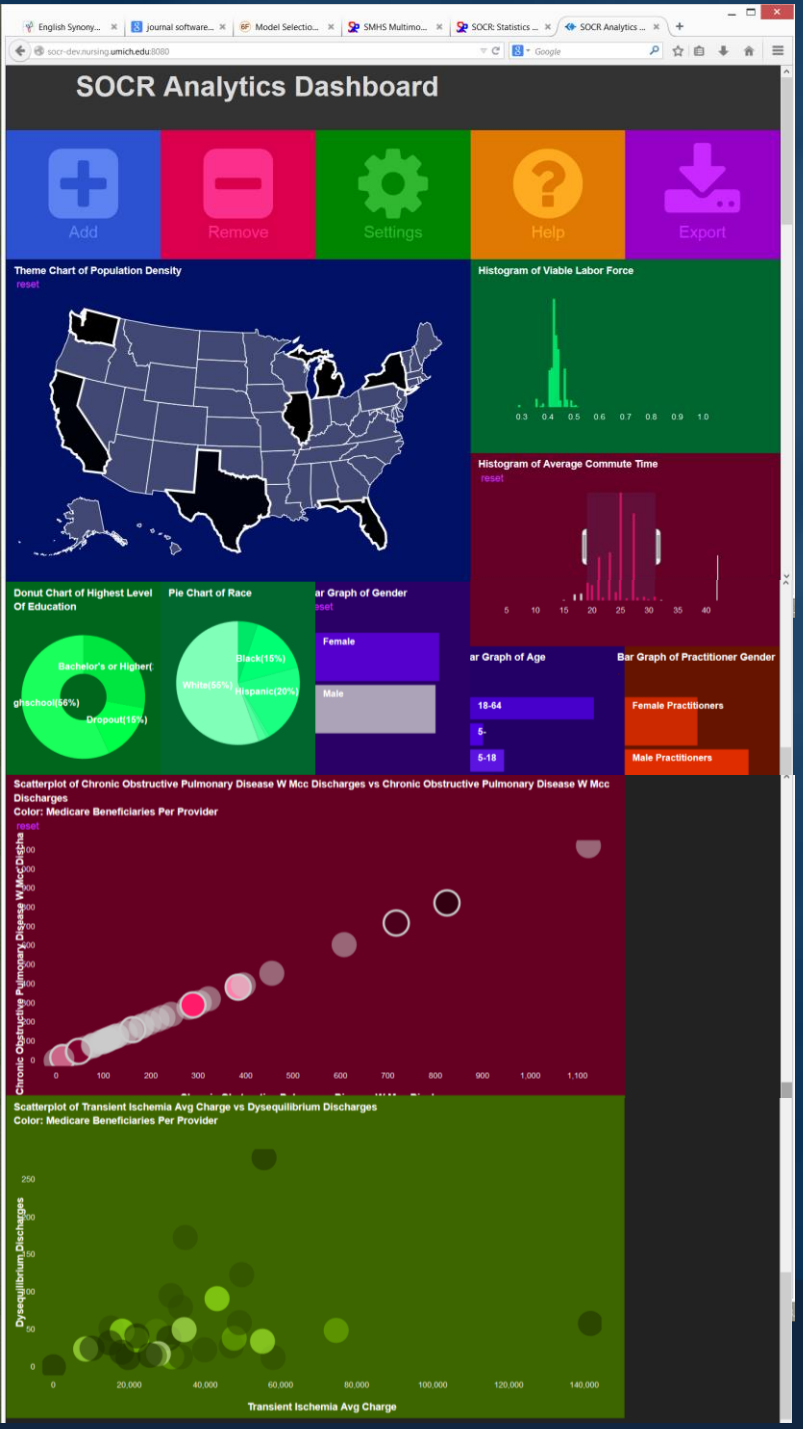
ML classifier	accuracy	sensitivity	specificity	positive predictive value	negative predictive value	log odds ratio (LOR)
AdaBoost	0.996324	0.994141	0.998264	0.9980392	0.9948097	11.4882058
SVM	0.985294	0.994140	0.977431	0.9750958	0.9946996	8.902166

in review

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

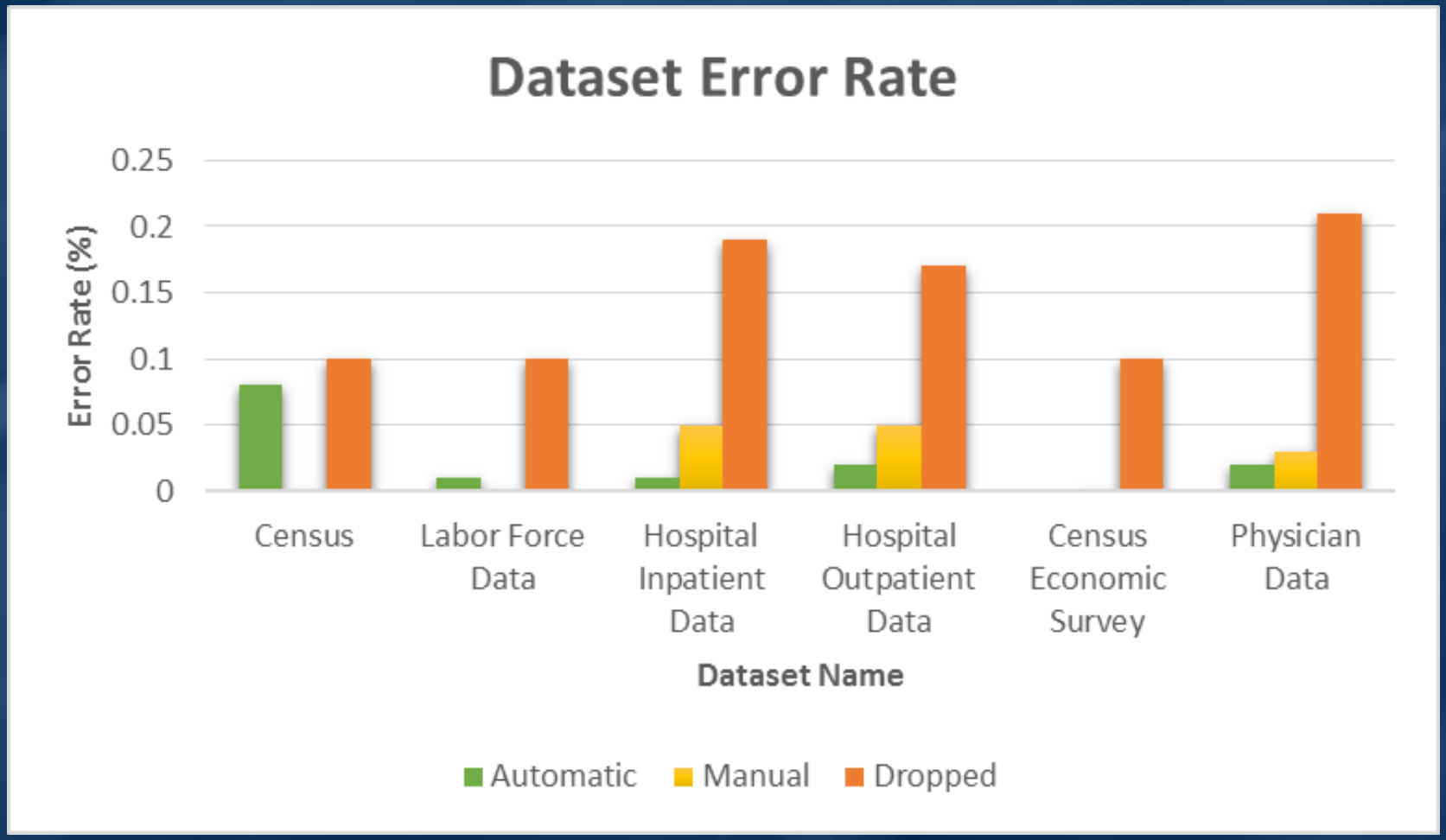


Husain, et al., 2015, J Big Data

SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

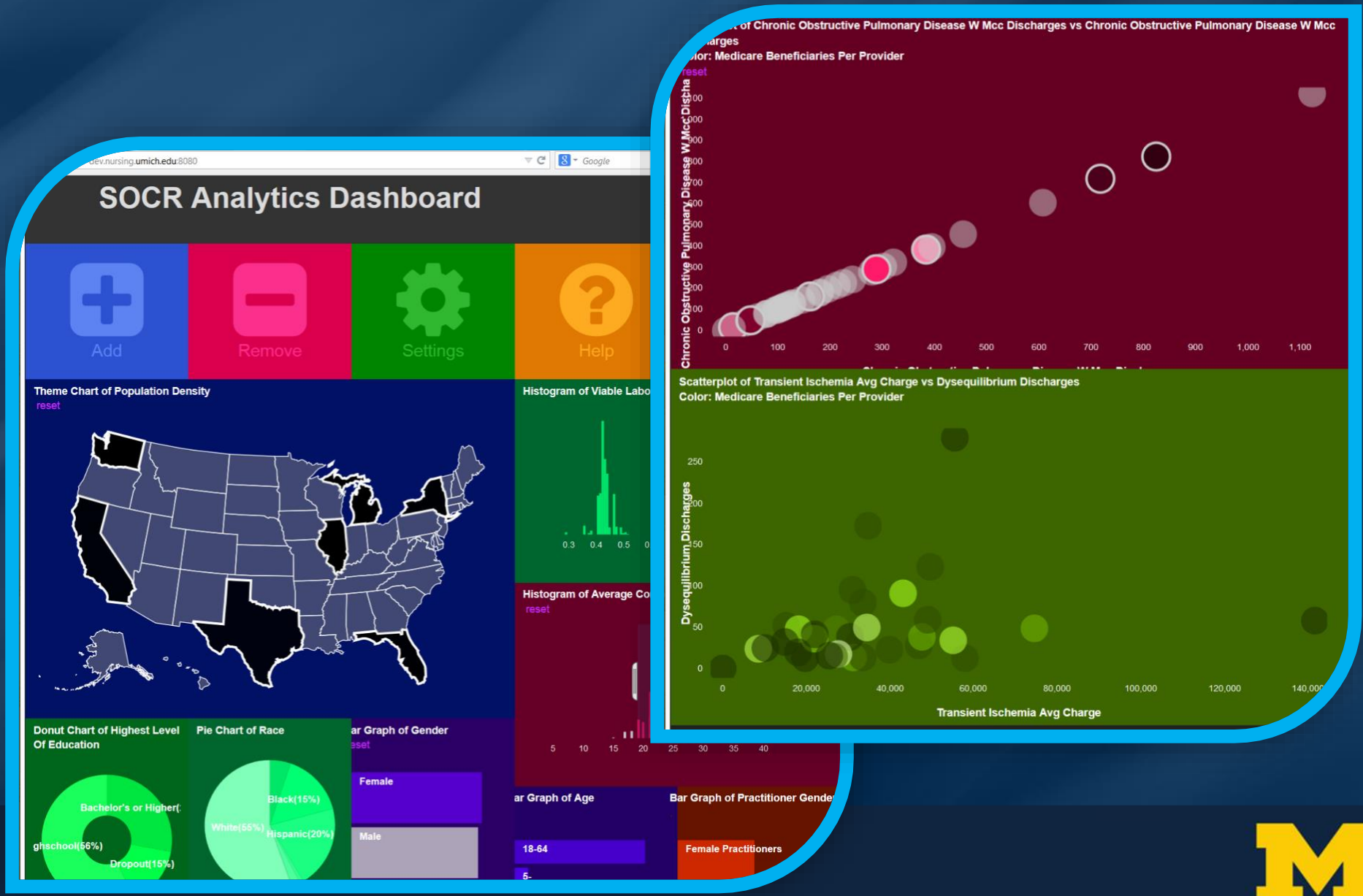
http://socr.umich.edu/HTML5/Dashboard

SOCR Dashboard (Exploratory Big Data Analytics): Data QC



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

SOCR Dashboard (Exploratory Big Data Analytics): Associations



Open Problem: Representation and Joint BD Analytics

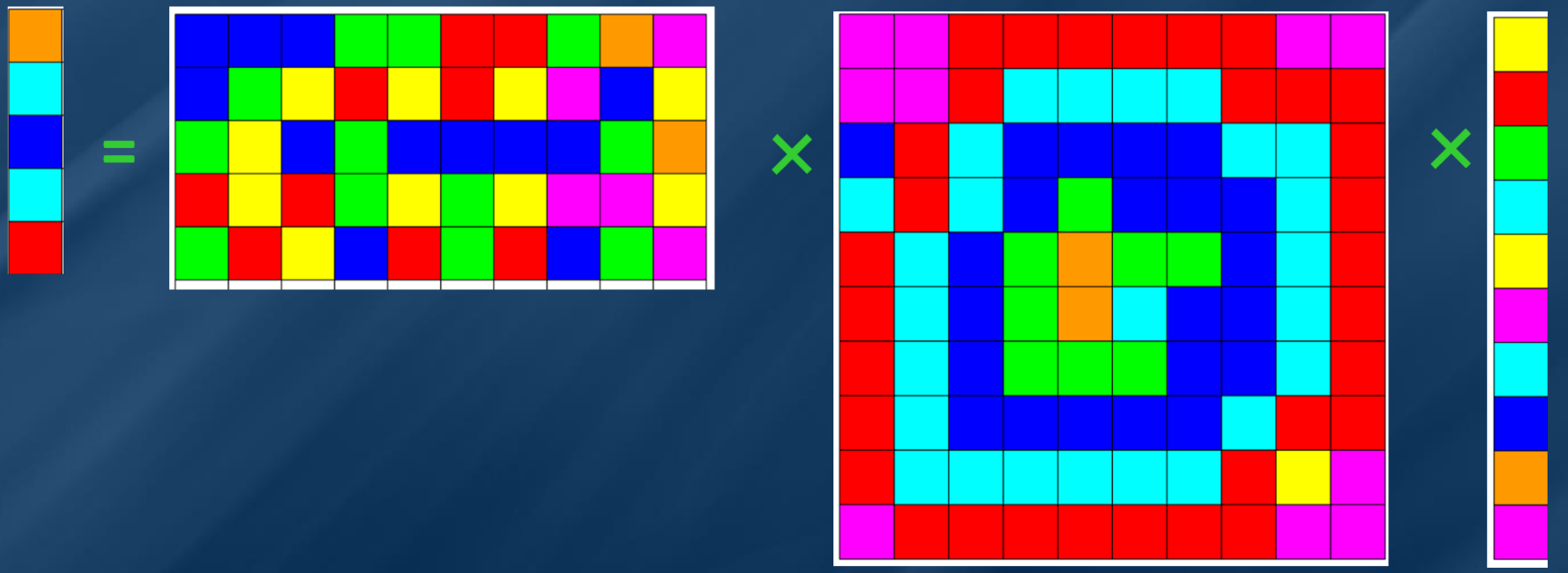
One Approach:
Compressive Big Data Analytics (CBDA)

Motivation! Big Data Analytics – Compressive Sensing

- There is currently no established analytical foundation for systematic representation of Big Data that facilitates the handling of data complexities and at the same time enables joint modeling, information extraction, high-throughput and adaptive scientific inference
- One idea is Compressive Big Data Analytics (CBDA), which borrows some of the compelling ideas for representation, reconstruction, recovery and data denoising recently developed for compressive sensing
- In compressive sensing, a sparse (incomplete) data is observed and one looks for a high-fidelity estimation of the complete dataset. Sparse data (or signals) can be described as observations with a small support, i.e., small magnitude according to the zero-norm

Big Data Analytics – Compressive Sensing

$y_{m\times 1}$ $A_{m\times n}$ $\Phi_{n\times n}$ $x_{n\times 1}$



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).

Compressive Big Data Analytics (CBDA)

- Finally, 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

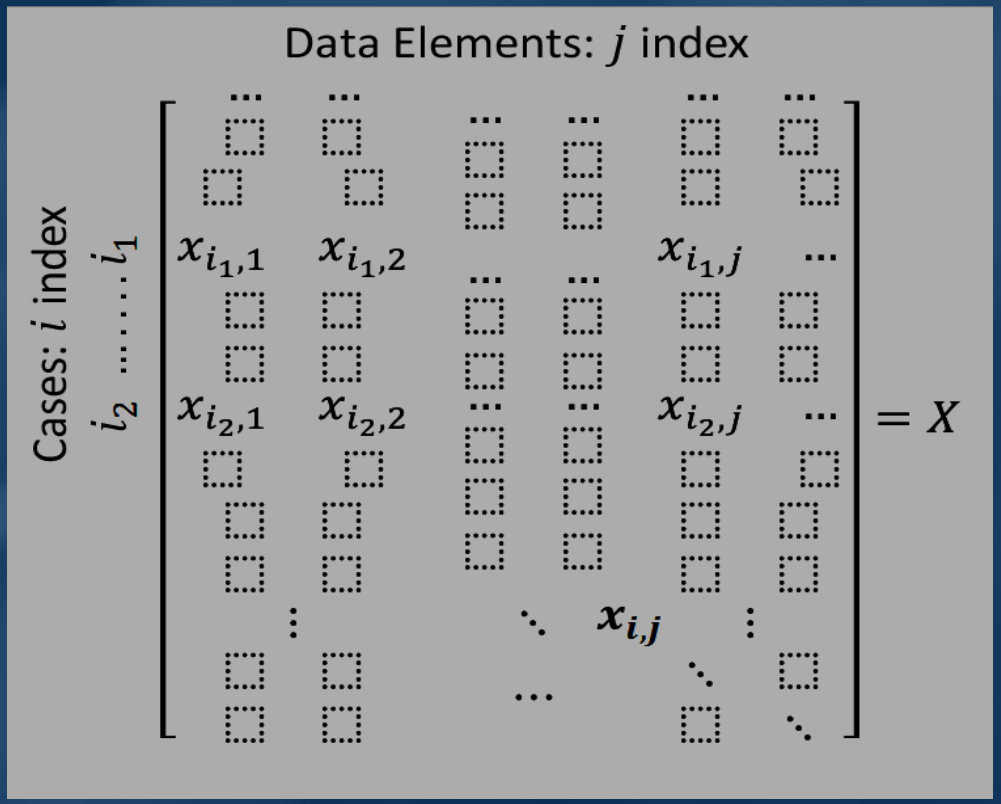
- When D is known, the model in equation (2) is jointly convex for L and S, and there exist iterative solvers based on sub-gradient recursion (e.g., alternating direction method of multipliers)

- 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

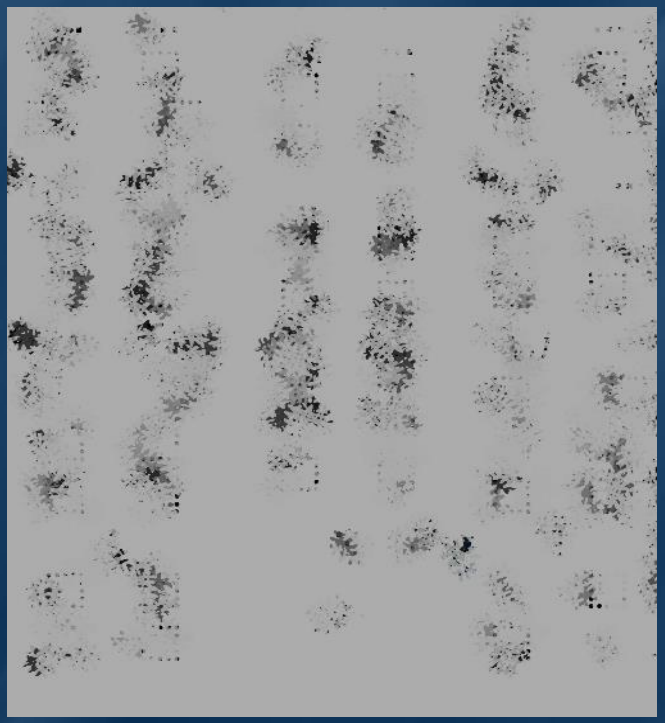
Compressive Big Data Analytics (CBDA)

- 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

Funding

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

NSF: 1416953, 0716055 1023115

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
- **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, Ben Hampstead, Bill Dauer