Michigan Institute for Data Science (MIDAS)

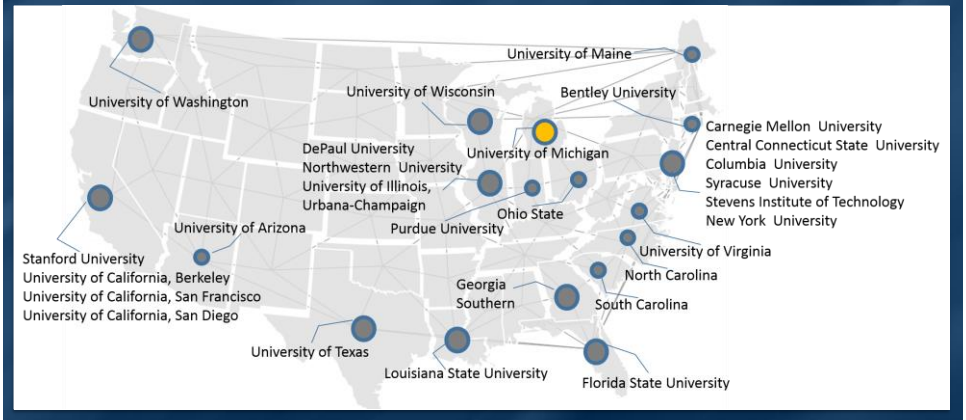
Computational Challenges & Research Opportunities

Ivo D Dinov

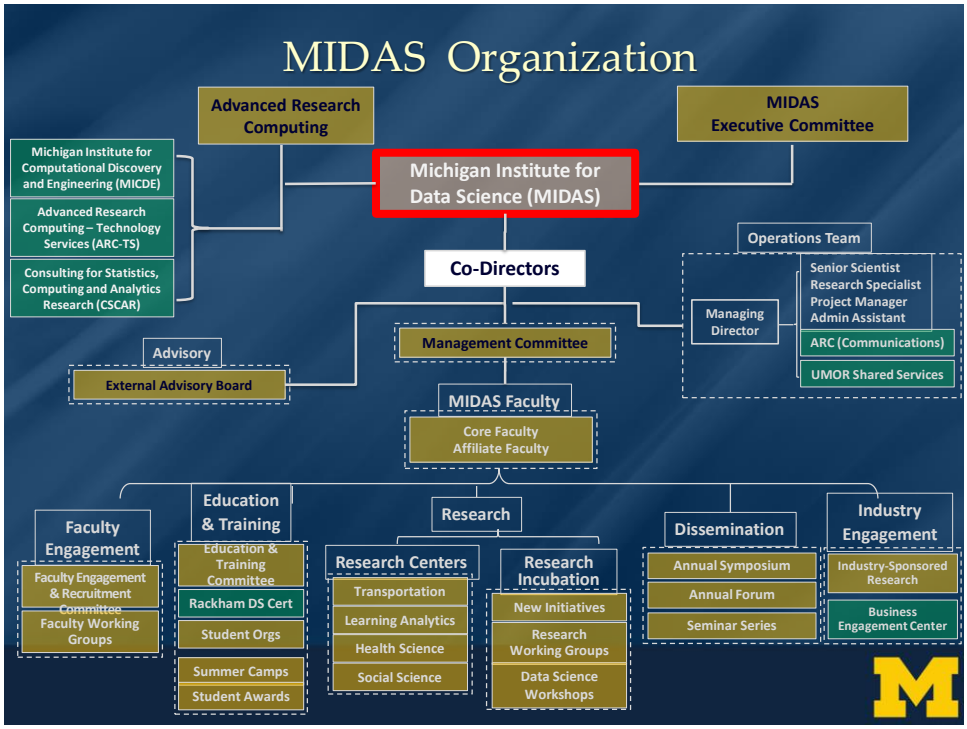
www.MIDAS.umich.edu

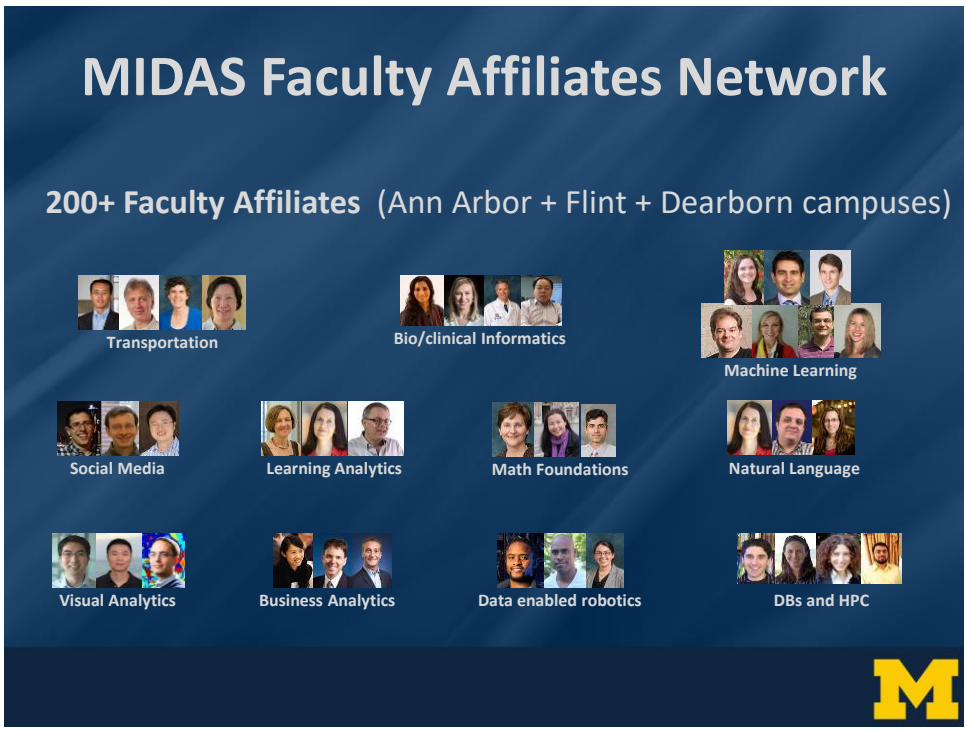
Michigan Institute for Data Science (MIDAS)
University of Michigan

National Big Data Science Curricula Constellation



Examples of recently established Data Science instituters and curricular programs in the US


MIDAS Organization
Advanced Research Computing
MIDAS Executive Committee
Michigan Institute for Computational Discovery and Engineering (MICDE)
Advanced Research Computing – Technology Services (ARC-TS)
Consulting for Statistics, Computing and Analytics Research (CSCAR)
Michigan Institute for Data Science (MIDAS)
Co-Directors
Operations Team
Managing Director
Senior Scientist
Research Specialist
Project Manager
Admin Assistant
ARC (Communications)
UMOR Shared Services
Advisory
External Advisory Board
Management Committee
MIDAS Faculty
Core Faculty
Affiliate Faculty
Faculty Engagement
Faculty Engagement & Recruitment Committee
Faculty Working Groups
Education & Training
Education & Training Committee
Rackham DS Cert
Student Orgs
Summer Camps
Student Awards
Research
Research Centers
Transportation
Learning Analytics
Health Science
Social Science
Research Incubation
New Initiatives
Research Working Groups
Data Science Workshops
Dissemination
Annual Symposium
Annual Forum
Seminar Series
Industry Engagement
Industry-Sponsored Research
Business Engagement Center


MIDAS Faculty Affiliates Network
200+ Faculty Affiliates (Ann Arbor + Flint + Dearborn campuses)
Transportation
Bio/clinical Informatics
Machine Learning
Social Media
Learning Analytics
Math Foundations
Natural Language
Visual Analytics
Business Analytics
Data enabled robotics
DBs and HPC

MIDAS Student Organizations

Computational Social Science Rackham Interdisciplinary Workshop

- 160+ members from 17 depts / institutes
- Panels, skill building workshops, reading groups, roundtable discussions

sites.lsa.umich.edu/css

Michigan Data Science Team (MDST)

"teaches practical data science skills by solving impactful problems"

- 50+ active members from CoE, LS&A, Ross and other units
- Competitions, Tutorials, Projects

midas.umich.edu/mdst

Statistics in the Community (STATCOM)

"promotes student-driven programs that provide statistical consulting as a community service"

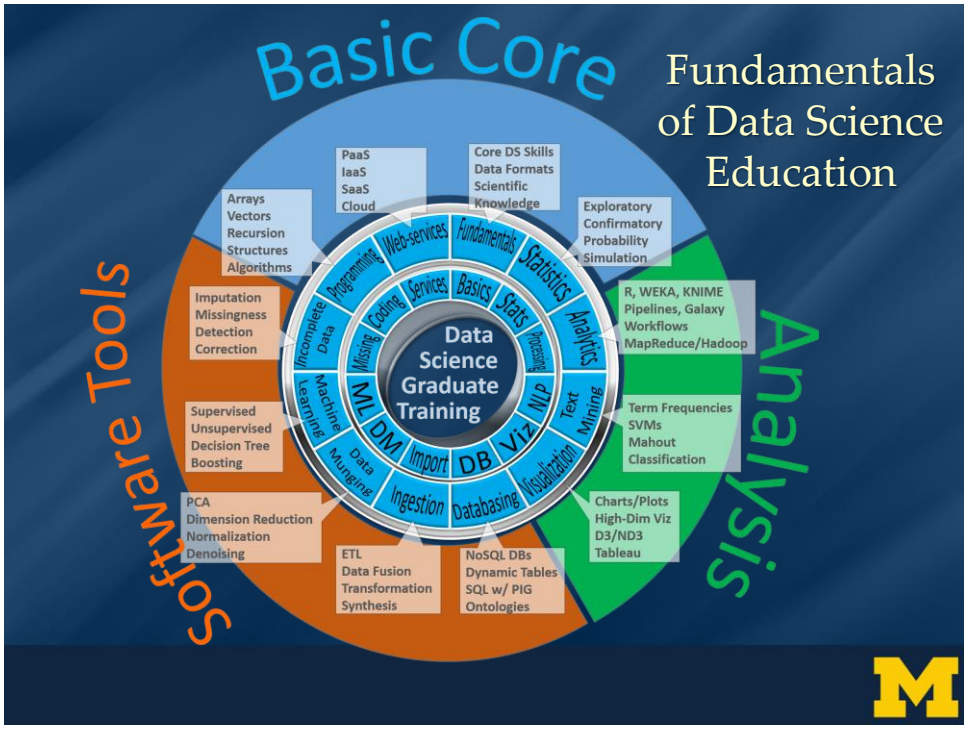
- 75+ members from Biostatistics, Statistics and Survey Methodology
- Offers services to local governmental and nonprofit community groups

sph.umich.edu/biostat/statcom

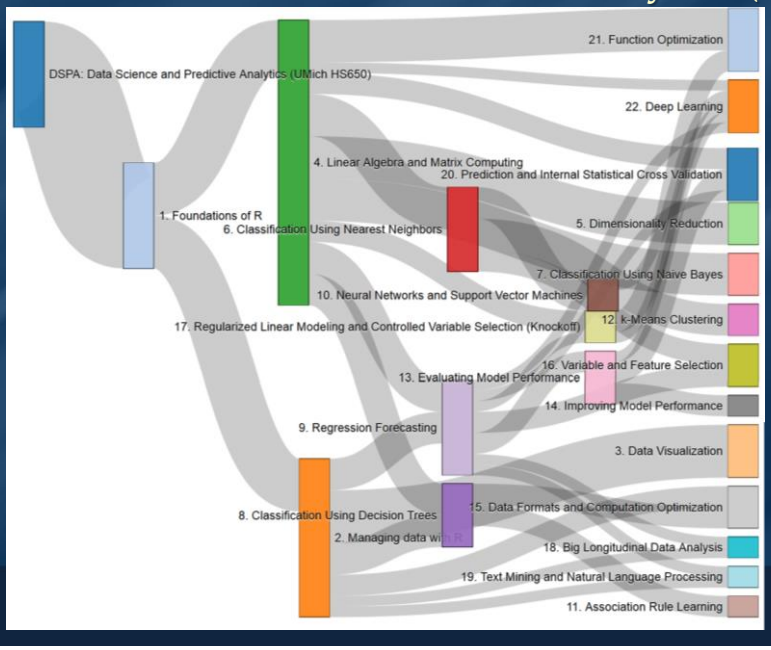
Michigan Student Artificial Intelligence Lab (MSAIL)

- 50+ members from CoE, LS&A, Ross and other units
- Machine learning reading group, research projects, tutorials

http://msail.org



Data Science and Predictive Analytics (HS650)



http://Predictive.Space

Data Science and Predictive Analytics (HS650)

Areas	Competency	Expectation
Algorithms and Applications	Tools	Working knowledge of basic software tools (command-line, GUI based, or web-services)
	Algorithms	Knowledge of core principles of scientific computing, applications programming, API's, algorithm complexity, and data structures
	Application Domain	Data analysis experience from at least one application area, either through coursework, internship, research project, etc.
Data Manage-ment	Data validation & visualization	Curation, Exploratory Data Analysis (EDA) and visualization
	Data wrangling	Skills for data normalization, data cleaning, data aggregation, and data harmonization/registration
	Data infrastructure	Handling databases, web-services, Hadoop, multi-source data
Analysis Methods	Statistical inference	Basic understanding of bias and variance, principles of (non)parametric statistical inference, and (linear) modeling
	Study design and diagnostics	Design of experiments, power calculations and sample sizing, strength of evidence, p-values, False Discovery Rates
	Machine Learning	Dimensionality reduction, k-nearest neighbors, random forests, AdaBoost, kernelization, SVM, ensemble methods, CNN

Desired Competencies

Big Data Science Challenges & Opportunities

MIDAS Challenge Initiatives

Data-Intensive Transportation Research Hub
- Reinventing Public Urban Transportation and Mobility
- Building a Transportation Data Ecosystem

Data-Intensive Learning Analytics Hub
- LEAP: Analytics for LEarners As People
- HOME: Holistic Modeling of Education

Data-Intensive Social Science Research Hub
- Computational Approaches for the Construction of Novel Macroeconomic Data
- A Social Science Collaboration for Research on Communication and Learning based upon Big Data

Data-Intensive Health Science Research Hub
- Michigan Center for Single-Cell Genomic Data Analytics
- Michigan Integrated Center for Health Analytics & Medical Prediction (MiCHAMP)
- Identifying Real-Time Data Predictors of Stress and Depression Using Mobile Technology

Data Science for Music Hub
- Understanding how the brain processes music through the Bach trio sonatas
- Mining patterns of audience engagement / crowdsourced music performances
- The sound of text
- A computational study of patterned melodic structures across musical cultures

http://midas.umich.edu/research

Characteristics of Big 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 effects
Incomplete	Reliable management of missing data

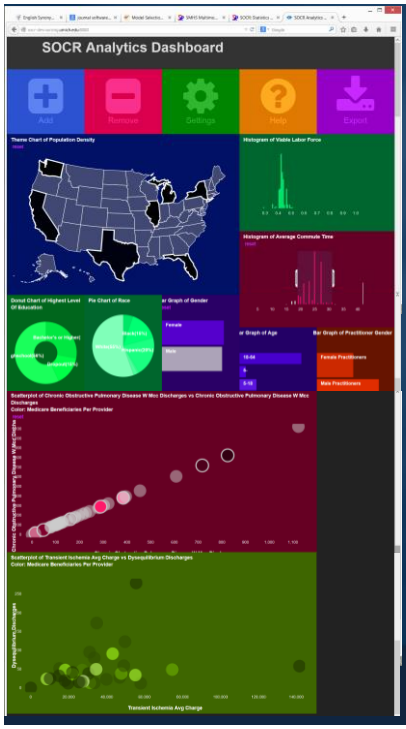
Biomed Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

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, *GigaScience* (2016) PMID:26918190

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 (2016) PMID:26918190



Big Data Dashboard

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

- Web-service combining and integrating multi-source socioeconomic and medical datasets
- Big Data Visual Analytics
- 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.*, J Big Data, 2015

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

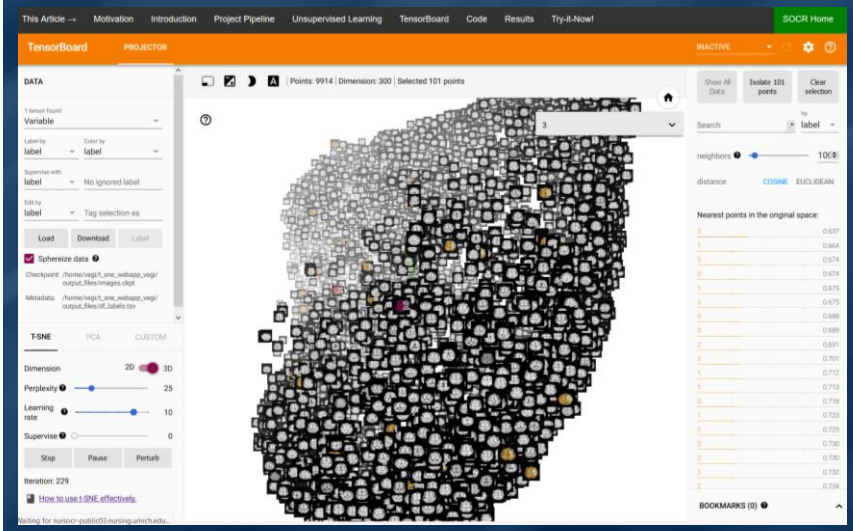
2	20005	Ongoing characteristics	Email access
2	110007	Ongoing characteristics	Newsletter communications, date sent
100	25780	Brain MRI	Acquisition protocol phase.
100	12139	Brain MRI	Believed safe to perform brain MRI scan
100	12188	Brain MRI	Brain MRI measurement completed
100	12187	Brain MRI	Brain 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-mec
101	22674	Carotid ultrasound	Mean carotid IMT (intima-mec
101	22677	Carotid ultrasound	Mean carotid IMT (intima-mec
101	22680	Carotid ultrasound	Mean carotid IMT (intima-mec
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/

Case-Studies – ALS

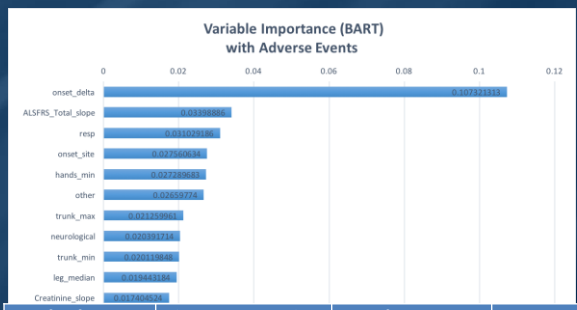
- Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

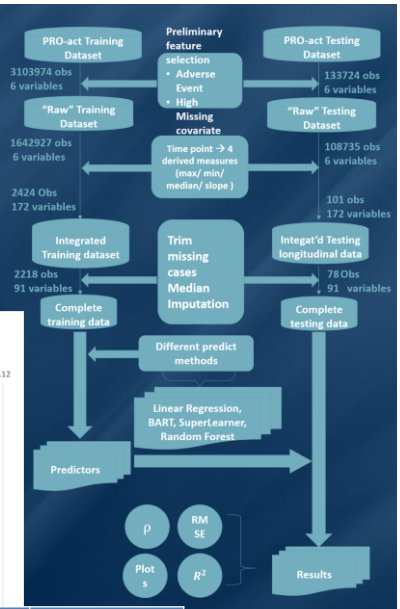
Data Source	Sample Size/Data Type	Summary
ProAct Archive	Over 100 variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: **Amyotrophic Lateral Sclerosis** Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole) 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)

Tang, et al. (2018), in review

Case-Studies – ALS

- Detect, track, and prognosticate the progression of ALS
- Predict 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 – ALS

- **Main Finding**: predicting univariate clinical outcomes may be challenging, the (information energy) signal is very weak. We can cluster ALS patients and generate evidence-based ALS hypotheses about the complex interactions of *multivariate factors*
- **Classification vs. Clustering**:
  - Classifying univariate clinical outcomes using the PRO-ACT data yields only marginal accuracy (about 70%).
  - Unsupervised clustering into sub-groups generates stable, reliable and consistent computable phenotypes whose explication requires interpretation of multivariate sets of features

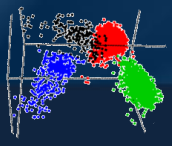
Data Representation Fusion Harmonization Aggregation → Cleaning Imputation Wrangling Synthesis → Model-based, Model-free, Classification, Clustering, Inference

Cluster	Consistency	Variance	Cluster-Size	Silhouette
1	1	0	565	0.58
2	0.986	0.018	427	0.63
3	0.956	0.053	699	0.5
4	0.985	0.018	733	0.5

Tang, et al. (2018), in review

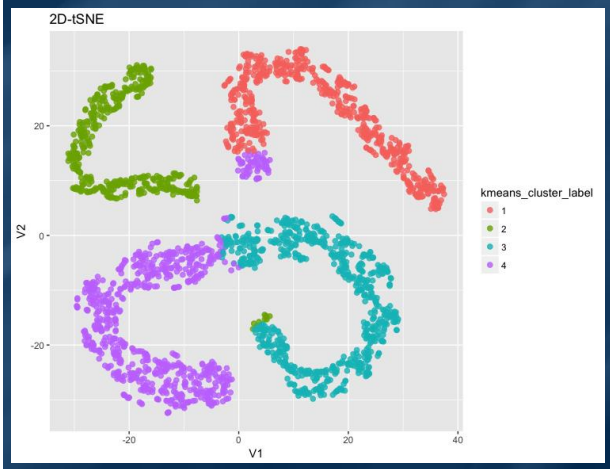
Case-Studies – ALS – Explicating Clustering

Feature Name	Between Cluster Significant Differences					
	1-2	1-3	1-4	2-3	2-4	3-4
...			...			
onset_delta.x	1	1	1	1	1	1
...			...			
Q9_Climbing_Stairs_slope	1			1		
...			...			
leg_max		1	1	1	1	
...			...			



Tang, et al. (2018), in review

Case-Studies – ALS – Dimensionality Reduction



2D t-SNE Manifold embedding

Learn a mapping: $f: R^n \overset{n \gg d}{\longrightarrow} R^d$
$\{x_1, x_2, \ldots, x_n\} \longrightarrow \{y_1, y_2, \ldots, y_d\}$
preserves closely the *original distances*, $p_{i,j}$ and represents the *derived similarities*, $q_{i,j}$ between pairs of embedded points:

$$q_{i,j} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||y_i - y_k||^2)^{-1}}$$

$$\min_f KL(P||Q) = \sum_{i \neq j} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

$$0 = \frac{\partial KL(P||Q)}{\partial y_i} = 2 \sum_j (p_{i,j} - q_{i,j}) f(|x_i - x_j|) u_{i,j}$$

$f(z) = \frac{z}{1+z^2}$ and $u_{i,j}$ is a unit vector from y_j to y_i.

Tang, et al. (2018), in review

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, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, 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

Acknowledgments

MIDAS Co-Directors

Brian Athey and Al Hero

MIDAS Education & Training Committee

Ivo Dinov HBBS/Bioinfo, Richard Gonzalez, ISR/PSY/LS&A,
Eric Schwartz Ross & Kerby Shedden, Stats/LS&A

Inaugural Program Committee Members

H. V. Jagadish: Electrical Engineering and Computer Science, CoE
Vijay Nair: Statistics & Industrial & Operations Engineering, LS&A/CoE
George Alter: Institute for Social Research; History, LS&A
Brian Athey: Computational Medicine and Bioinformatics, SoM
Mike Cafarella: Computer Science and Engineering, CoE
Ivo Dinov, Chair, Leadership and Effectiveness Science, Bioinformatics, SoN/SoM
Karthik Duraisamy: Atmospheric, Oceanic, and Space Sciences
August (Gus) Evrard: Physics; Astronomy, LS&A
Anna Gilbert: Mathematics, LS&A
Alfred Hero: Electrical Engineering and Computer Science; Biomedical Engineering, CoE
Judy Jin: Industrial & Operations Engineering, CoE
Carl Lagoze: School of Information
Qiaozhu Mei: School of Information
Christopher Miller: Astronomy, LS&A
Dragomir Radev: School of Information; Computer Science and Engineering; Linguistics, CoE
Stephen Smith: Ecology and Evolutionary Biology, LS&A
Ambuj Tewari: Statistics; Computer Science and Engineering, LS&A
Honglak Lee, Electrical Engineering and Computer Science, CoE
Jeremy Taylor, Biostatistics, SPH

Michigan Institute for Data Science
University of Michigan

www.MIDAS.umich.edu

Ivo Dinov
dinov@umich.edu

Open-ended discussion of educational challenges, research opportunities and infrastructure demands in data science