Examples of recently established Data Science institutes and curricular programs in the US
MIDAS Organization

Michigan Institute for Data Science (MIDAS)

Co-Directors

Management Committee

Advisory External Advisory Board

MIDAS Faculty

Core Faculty

Affiliate Faculty

Research Centers

Transportation

Learning Analytics

Health Science

Social Science

Research Incubation

New Initiatives

Research Working Groups

Data Science Workshops

Research Education & Training

Education & Training Committee

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

Faculty Engagement

Faculty Engagement & Recruitment

Faculty Working Groups

Education & Training

Education & Training Committee

Transportation

Learning Analytics

Health Science

Social Science

Research Incubation

New Initiatives

Research Working Groups

Data Science Workshops

Research Education & Training

Education & Training Committee

Transportation

Learning Analytics

Health Science

Social Science

Research Incubation

New Initiatives

Research Working Groups

Data Science Workshops

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

Michigan Institute for Computational Discovery and Engineering (MICDE)

Advanced Research Computing - Technology Services (ARC-TS)

Consulting for Statistics, Computing and Analytics Research (CSCAR)
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

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

Basic Core
Fundamentals of Data Science Education

Software Tools
- PCA, Dimension Reduction
- Visualization, Dimension Reduction
- ETL, Data Fusion

Analysis
- Core DS Skills
- Data Formats
- Scientific Knowledge

Exploratory
- Confirmatory
- Probability
- Simulation

 Term Frequencies
- SVMs
- Mahout
- Classification

Charts/Plots
- High-Dim Vis
- D3/3D
- Tableau
## Desired Competencies

### Data Science and Predictive Analytics (HS650)

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

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

<table>
<thead>
<tr>
<th>Big Bio Data Dimensions</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Harvesting and management of vast amounts of data</td>
</tr>
<tr>
<td>Complexity</td>
<td>Wranglers for dealing with heterogeneous data</td>
</tr>
<tr>
<td>Incongruency</td>
<td>Tools for data harmonization and aggregation</td>
</tr>
<tr>
<td>Multi-source</td>
<td>Transfer and joint modeling of disparate elements</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>Macro to meso to micro scale observations</td>
</tr>
<tr>
<td>Time</td>
<td>Techniques accounting for longitudinal effects</td>
</tr>
<tr>
<td>Incomplete</td>
<td>Reliable management of missing data</td>
</tr>
</tbody>
</table>

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 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%.

Falls in PD are extremely difficult to predict …

PD phenotypes
Tremor-Dominant (TD)
Postural Instability &
gait difficulty (PI & GD)

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)

<table>
<thead>
<tr>
<th>Method</th>
<th>acc</th>
<th>sens</th>
<th>spec</th>
<th>ppv</th>
<th>npv</th>
<th>lor</th>
<th>auc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.728</td>
<td>0.537</td>
<td>0.855</td>
<td>0.710</td>
<td>0.736</td>
<td>1.920</td>
<td>0.774</td>
</tr>
<tr>
<td>Random Forests</td>
<td><strong>0.796</strong></td>
<td><strong>0.683</strong></td>
<td><strong>0.871</strong></td>
<td><strong>0.778</strong></td>
<td><strong>0.806</strong></td>
<td><strong>2.677</strong></td>
<td><strong>0.821</strong></td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.689</td>
<td>0.610</td>
<td>0.742</td>
<td>0.610</td>
<td>0.742</td>
<td>1.502</td>
<td>0.793</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.699</td>
<td>0.707</td>
<td>0.694</td>
<td>0.604</td>
<td>0.782</td>
<td>1.699</td>
<td>0.787</td>
</tr>
<tr>
<td>SVM</td>
<td>0.709</td>
<td>0.561</td>
<td>0.806</td>
<td>0.657</td>
<td>0.735</td>
<td>1.672</td>
<td>0.822</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.699</td>
<td>0.610</td>
<td>0.758</td>
<td>0.625</td>
<td>0.746</td>
<td>1.588</td>
<td></td>
</tr>
<tr>
<td>Super Learner</td>
<td>0.738</td>
<td>0.683</td>
<td>0.774</td>
<td>0.667</td>
<td>0.787</td>
<td>1.999</td>
<td></td>
</tr>
</tbody>
</table>

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

UK Biobank – discriminate between HC, single and multiple comorbid conditions
- Predict likelihoods of various developmental or aging disorders
- Forecast cancer

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Biobank</td>
<td>Demographics: &gt; 500K cases</td>
<td>The longitudinal archive of the UK population (NHS)</td>
</tr>
<tr>
<td></td>
<td>Clinical data: &gt; 4K features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Imaging data: T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genetics data</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UK Biobank</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum carotid IMT (intima-media-thickness) at 150 degrees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum carotid IMT (intima-media-thickness) at 210 degrees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quality control indicator for IMT at 120 degrees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quality control indicator for IMT at 150 degrees</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quality control indicator for IMT at 210 degrees</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProAct Archive</td>
<td>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</td>
<td>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)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Linear Regression</th>
<th>Random Forest</th>
<th>BART</th>
<th>SuperLearner</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.081</td>
<td>0.174</td>
<td>0.225</td>
<td>0.178</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.619</td>
<td>0.587</td>
<td>0.568</td>
<td>0.585</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.298</td>
<td>0.434</td>
<td>0.485</td>
<td>0.447</td>
</tr>
</tbody>
</table>

- **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

Cluster Consistency Variance Cluster-Site Silhouette

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Consistency</th>
<th>Variance</th>
<th>Cluster-Site</th>
<th>Silhouette</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>565</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>0.986</td>
<td>0.018</td>
<td>427</td>
<td>0.63</td>
</tr>
<tr>
<td>3</td>
<td>0.956</td>
<td>0.053</td>
<td>699</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.985</td>
<td>0.018</td>
<td>733</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Tang, et al. (2018), in review
### Case-Studies – ALS – Explicating Clustering

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Between Cluster Significant Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-2</td>
</tr>
<tr>
<td>_____________________________</td>
<td>____</td>
</tr>
<tr>
<td>onset_delta.x</td>
<td>1</td>
</tr>
<tr>
<td>_____________________________</td>
<td>____</td>
</tr>
<tr>
<td>Q9_Climbing_Stairs_slope</td>
<td>1</td>
</tr>
<tr>
<td>_____________________________</td>
<td>____</td>
</tr>
<tr>
<td>leg_max</td>
<td>1</td>
</tr>
<tr>
<td>_____________________________</td>
<td>____</td>
</tr>
</tbody>
</table>

Tang, et al. (2018), in review

### Case-Studies – ALS – Dimensionality Reduction

2D t-SNE Manifold embedding

Learn a mapping: \( f : \mathbb{R}^n \rightarrow \mathbb{R}^d \)

\[ \{x_1, x_2, \ldots, x_n\} \rightarrow \{y_1, y_2, \ldots, y_d\} \]

preserves closely the original distances, \( p_{i,j} \) and represents the derived similarities, \( q_{i,j} \) between pairs of embedded points:

\[
q_{i,j} = \left( 1 + ||y_i - y_j||^2 \right)^{-1}
\]

\[
\min_KL(P || Q) = \sum_{i \neq j} p_{i,j} log \frac{p_{i,j}}{q_{i,j}}
\]

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

\[
f(z) = \frac{z}{1 + z^2}
\]

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, Sehram Palanikumar, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkevat, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Stengist, Rob Gould, Jingshu Xu, Nellie Ponorari, Ming Tang, Asiyah Lin, Nicolas Christou, Haribo Sun, Tuo Wang, Simeone Marino

- **LONI**: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yingping Shi, David Shattuck, Elizabeth Sowell, Katherine Nair, Abigail Jones, Shantianjai, Paul Thompson, Luminita Vese, Stan Ogier, Stefano Soatto, Sreek Moen, Junling Li, Young Sung, Carl Kesselman, Fabio Maciardi, Federica Toni

- **UMich MIDAS/MNORC/AD/PD Centers**: Caihong Spina, Chuck Burant, Ben Hampstead, Stephen Goulman, Stephen Strabbing, Hiroko Dodge, Hank Paulson, Bill Dauer, Brian Athey

MIDAS Co-Directors
Brian Athey and AI Hero

MIDAS Education & Training Committee

Inaugural Program Committee Members

H. V. Jagadish: Electrical Engineering and Computer Science, 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