Michigan Institute for Data Science (MIDAS)

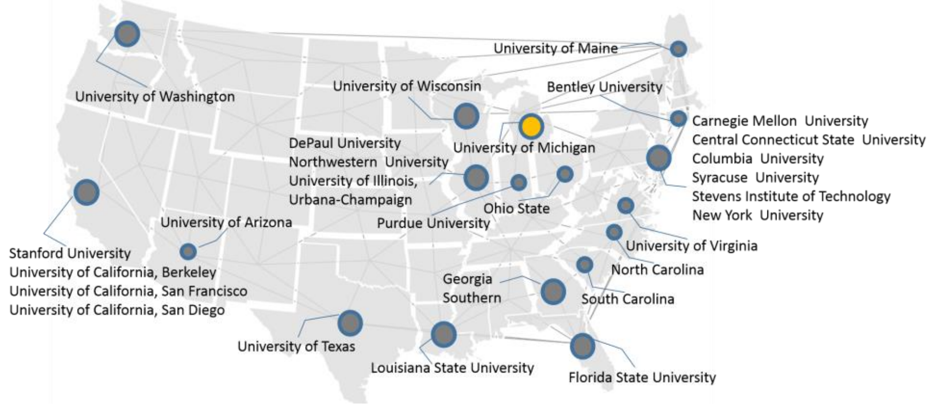
Foundations, Challenges & Opportunities

Ivo D Dinov

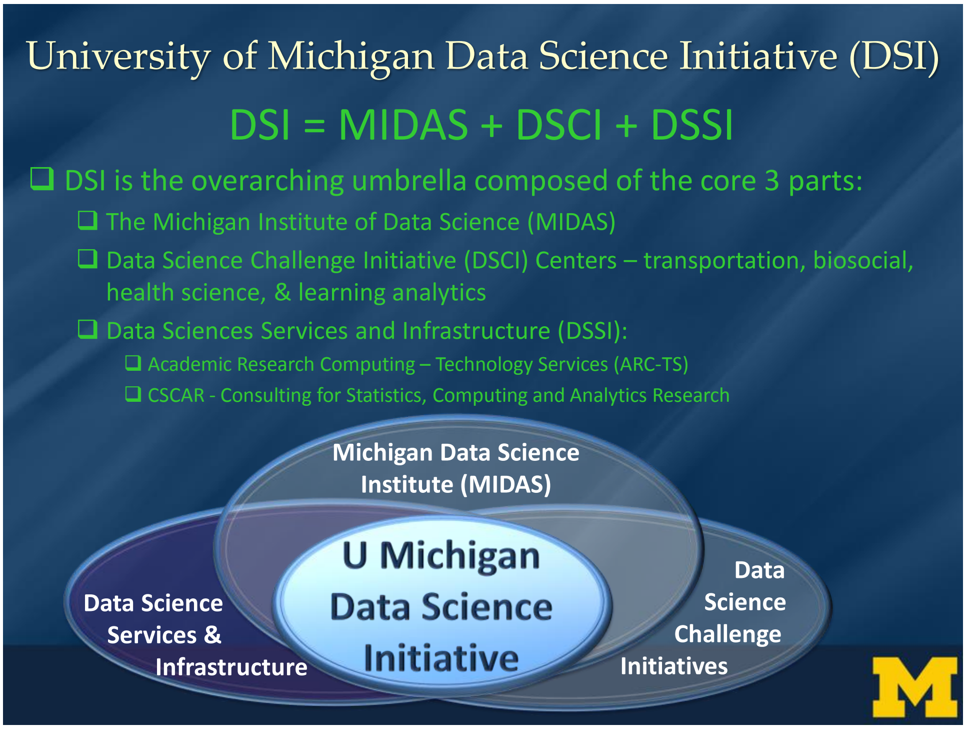
www.MIDAS.umich.edu

Michigan Institute for Data Science (MIDAS)
University of Michigan

National Big Data Science Curricula Constellation



Recently established Data Science instituters and curricular programs



University of Michigan Data Science Initiative (DSI)

DSI = MIDAS + DSCI + DSSI

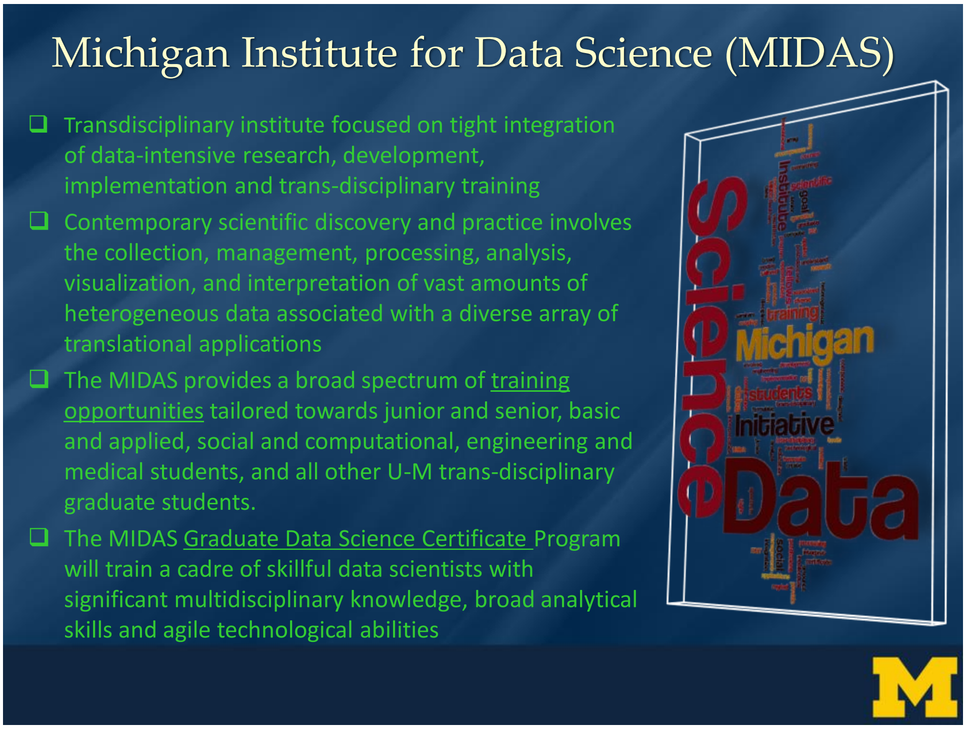
- ❑ DSI is the overarching umbrella composed of the core 3 parts:
  - ❑ The Michigan Institute of Data Science (MIDAS)
  - ❑ Data Science Challenge Initiative (DSCI) Centers – transportation, biosocial, health science, & learning analytics
  - ❑ Data Sciences Services and Infrastructure (DSSI):
    - ❑ Academic Research Computing – Technology Services (ARC-TS)
    - ❑ CSCAR - Consulting for Statistics, Computing and Analytics Research

Michigan Data Science Institute (MIDAS)

U Michigan Data Science Initiative

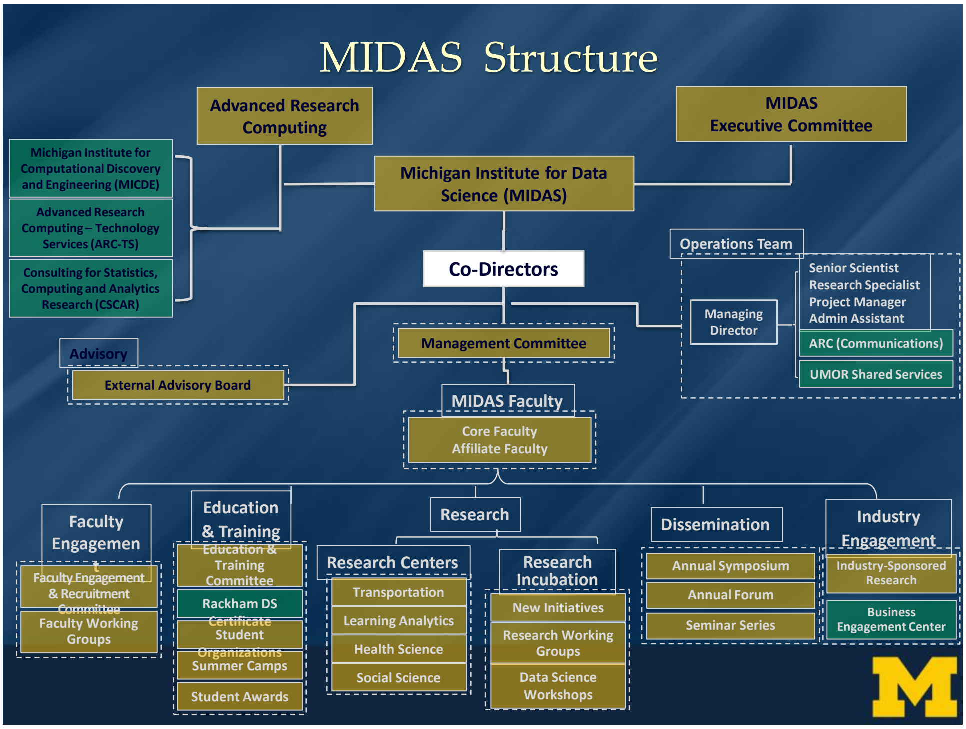
Data Science Services & Infrastructure

Data Science Challenge Initiatives



Michigan Institute for Data Science (MIDAS)

- ❑ Transdisciplinary institute focused on tight integration of data-intensive research, development, implementation and trans-disciplinary training
- ❑ Contemporary scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of translational applications
- ❑ The MIDAS provides a broad spectrum of training opportunities tailored towards junior and senior, basic and applied, social and computational, engineering and medical students, and all other U-M trans-disciplinary graduate students.
- ❑ The MIDAS Graduate Data Science Certificate Program will train a cadre of skillful data scientists with significant multidisciplinary knowledge, broad analytical skills and agile technological abilities



MIDAS Structure

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)

Operations Team

Co-Directors

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 Certificate
- Student Organizations
- Summer Camps
- Student Awards

Research

Research Centers

- Transportation
- Learning Analytics
- Health Science
- Social Science

Research Incubation

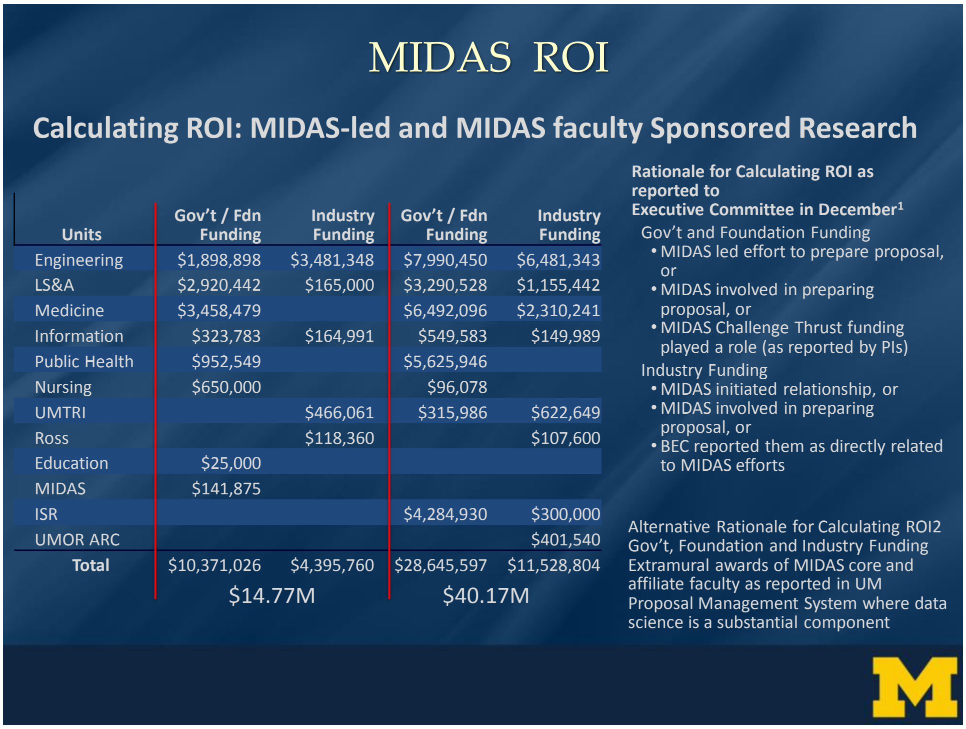
- New Initiatives
- Research Working Groups
- Data Science Workshops

Dissemination

- Annual Symposium
- Annual Forum
- Seminar Series

Industry Engagement

- Industry-Sponsored Research
- Business Engagement Center



MIDAS ROI

Calculating ROI: MIDAS-led and MIDAS faculty Sponsored Research

Units	Gov't / Fdn Funding	Industry Funding	Gov't / Fdn Funding	Industry Funding
Engineering	$1,898,898	$3,481,348	$7,990,450	$6,481,343
LS&A	$2,920,442	$165,000	$3,290,528	$1,155,442
Medicine	$3,458,479		$6,492,096	$2,310,241
Information	$323,783	$164,991	$549,583	$149,989
Public Health	$952,549		$5,625,946	
Nursing	$650,000		$96,078	
UMTRI		$466,061	$315,986	$622,649
Ross		$118,360		$107,600
Education	$25,000			
MIDAS	$141,875			
ISR			$4,284,930	$300,000
UMOR ARC				$401,540
Total	$10,371,026	$4,395,760	$28,645,597	$11,528,804
	$14.77M		$40.17M	

Rationale for Calculating ROI as reported to Executive Committee in December[1]

Gov't and Foundation Funding

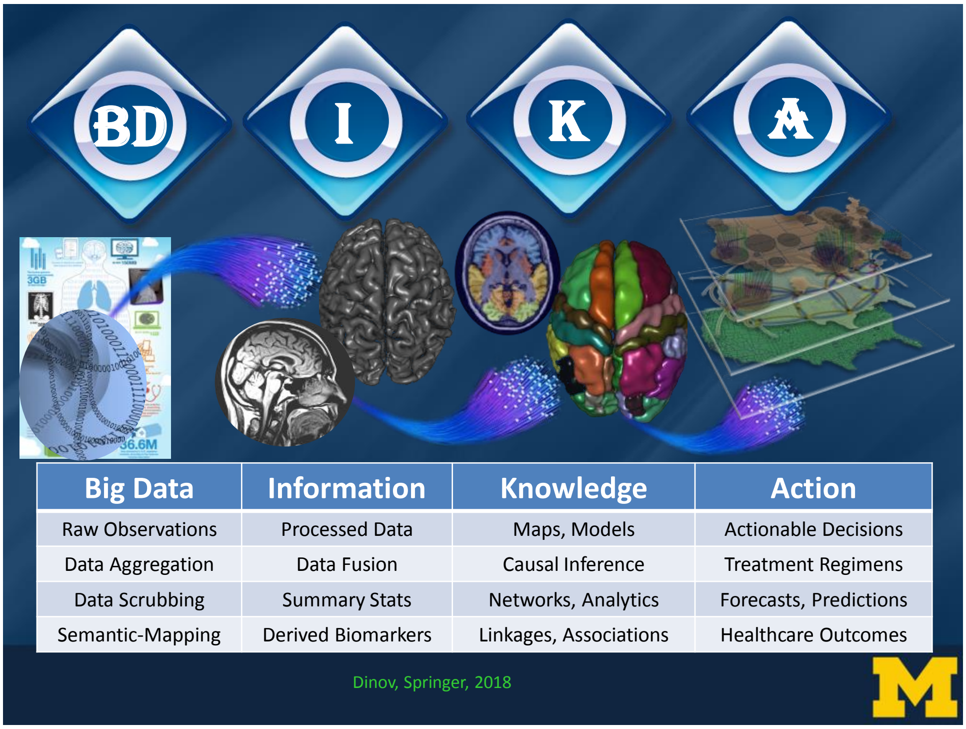
- MIDAS led effort to prepare proposal, or
- MIDAS involved in preparing proposal, or
- MIDAS Challenge Thrust funding played a role (as reported by PIs)

Industry Funding

- MIDAS initiated relationship, or
- MIDAS involved in preparing proposal, or
- BEC reported them as directly related to MIDAS efforts

Alternative Rationale for Calculating ROI2
Gov't, Foundation and Industry Funding
Extramural awards of MIDAS core and affiliate faculty as reported in UM Proposal Management System where data science is a substantial component

Big Data Science



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, Springer, 2018

Characteristics of Big Biomed Data

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

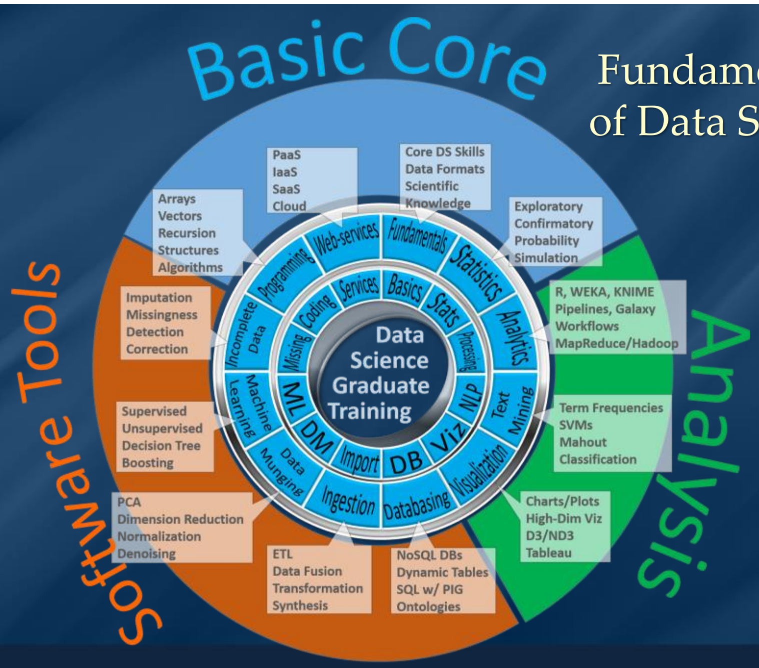
BD Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro to meso to micro scale observations
Incomplete	Reliable management of missing data

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, et al. (2014)

Fundamentals of Data Science



Basic Core

Analysis

Software Tools

Data Science Graduate Training

Web-services, Services: PaaS, IaaS, SaaS, Cloud

Fundamentals, Basics: Core DS Skills, Data Formats, Scientific Knowledge

Statistics, Stats: Exploratory, Confirmatory, Probability, Simulation

Analytics, Processing: R, WEKA, KNIME, Pipelines, Galaxy, Workflows, MapReduce/Hadoop

Text Mining, NLP: Term Frequencies, SVMs, Mahout, Classification

Visualization, Viz: Charts/Plots, High-Dim Viz, D3/ND3, Tableau

Databasing, DB: NoSQL DBs, Dynamic Tables, SQL w/ PIG, Ontologies

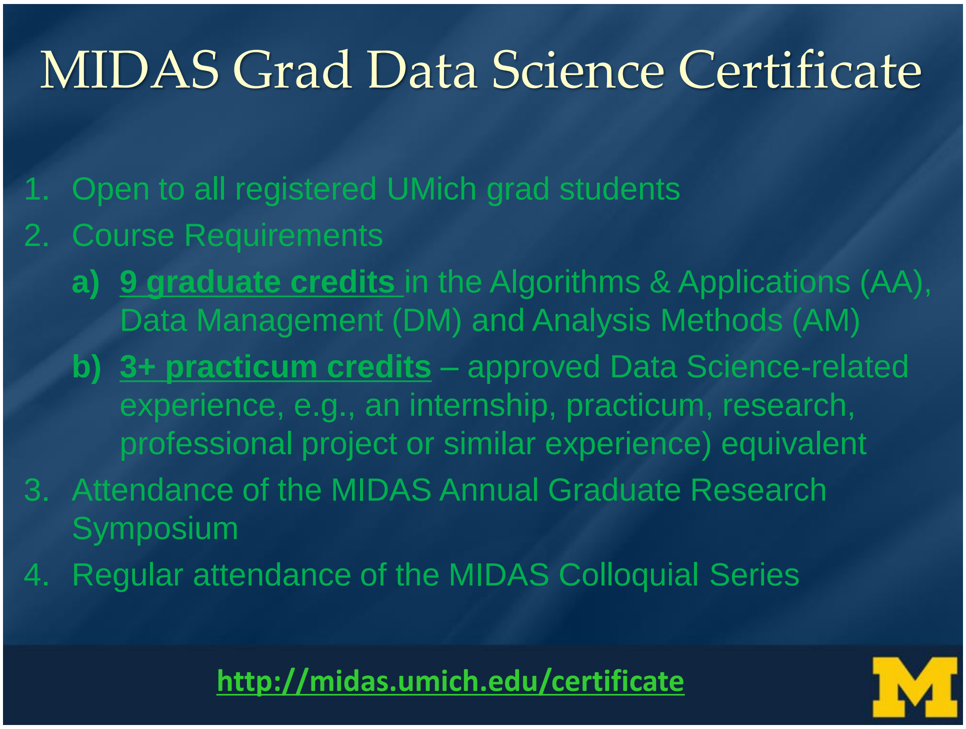
Ingestion, Import: ETL, Data Fusion, Transformation, Synthesis

Data Munging, DM: PCA, Dimension Reduction, Normalization, Denoising

Machine Learning, ML: Supervised, Unsupervised, Decision Tree, Boosting

Incomplete Data, Missing: Imputation, Missingness, Detection, Correction

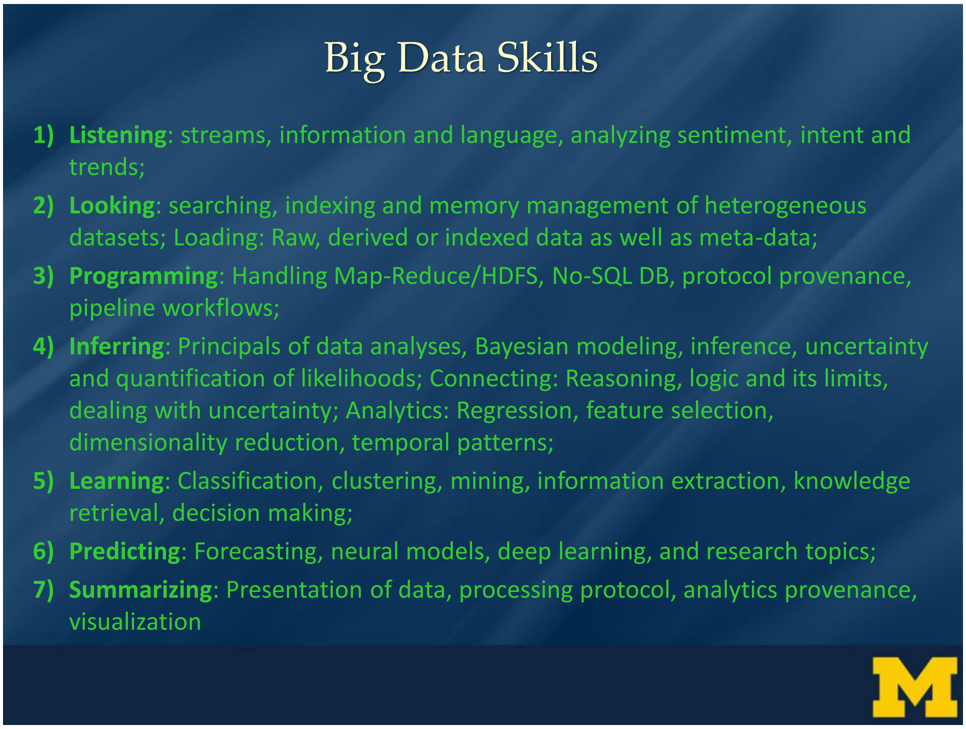
Programming, Coding: Arrays, Vectors, Recursion, Structures, Algorithms



MIDAS Grad Data Science Certificate

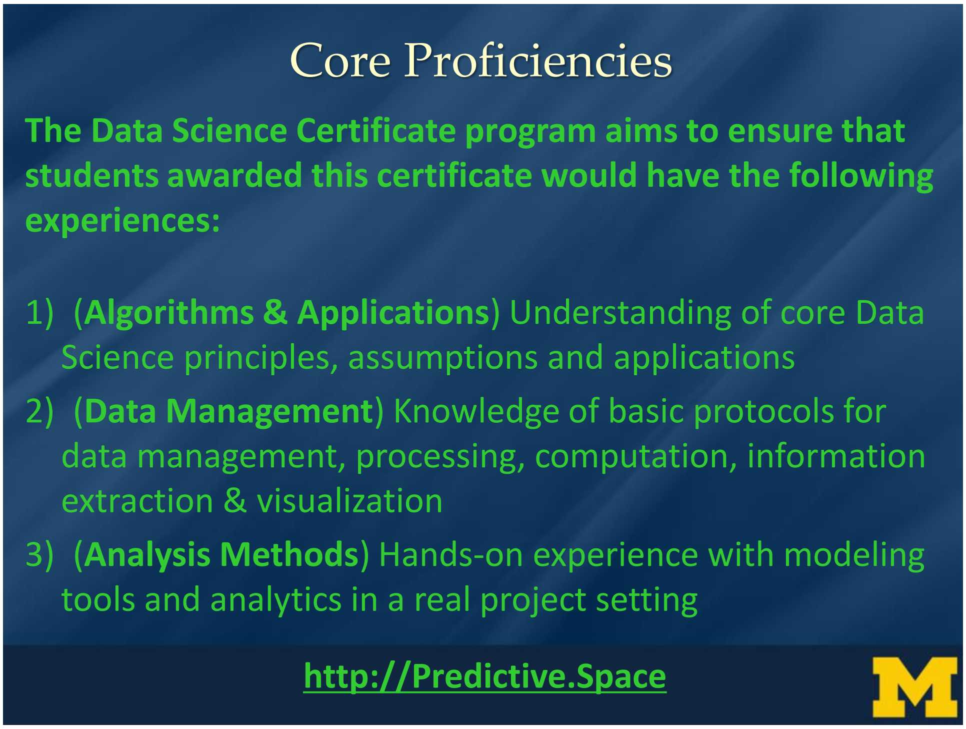
1. Open to all registered UMich grad students
2. Course Requirements
   a) **9 graduate credits** in the Algorithms & Applications (AA), Data Management (DM) and Analysis Methods (AM)
   b) **3+ practicum credits** – approved Data Science-related experience, e.g., an internship, practicum, research, professional project or similar experience) equivalent
3. Attendance of the MIDAS Annual Graduate Research Symposium
4. Regular attendance of the MIDAS Colloquial Series

http://midas.umich.edu/certificate



Big Data Skills

1) **Listening**: streams, information and language, analyzing sentiment, intent and trends;
2) **Looking**: searching, indexing and memory management of heterogeneous datasets; Loading: Raw, derived or indexed data as well as meta-data;
3) **Programming**: Handling Map-Reduce/HDFS, No-SQL DB, protocol provenance, pipeline workflows;
4) **Inferring**: Principals of data analyses, Bayesian modeling, inference, uncertainty and quantification of likelihoods; Connecting: Reasoning, logic and its limits, dealing with uncertainty; Analytics: Regression, feature selection, dimensionality reduction, temporal patterns;
5) **Learning**: Classification, clustering, mining, information extraction, knowledge retrieval, decision making;
6) **Predicting**: Forecasting, neural models, deep learning, and research topics;
7) **Summarizing**: Presentation of data, processing protocol, analytics provenance, visualization

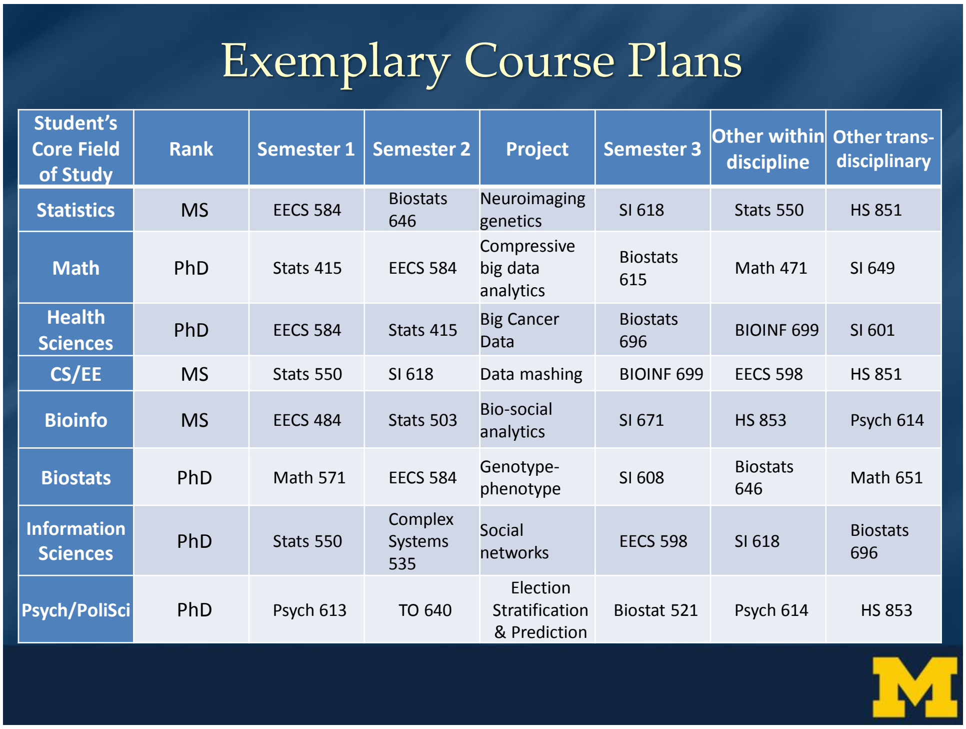


Core Proficiencies

The Data Science Certificate program aims to ensure that students awarded this certificate would have the following experiences:

1) (**Algorithms & Applications**) Understanding of core Data Science principles, assumptions and applications
2) (**Data Management**) Knowledge of basic protocols for data management, processing, computation, information extraction & visualization
3) (**Analysis Methods**) Hands-on experience with modeling tools and analytics in a real project setting

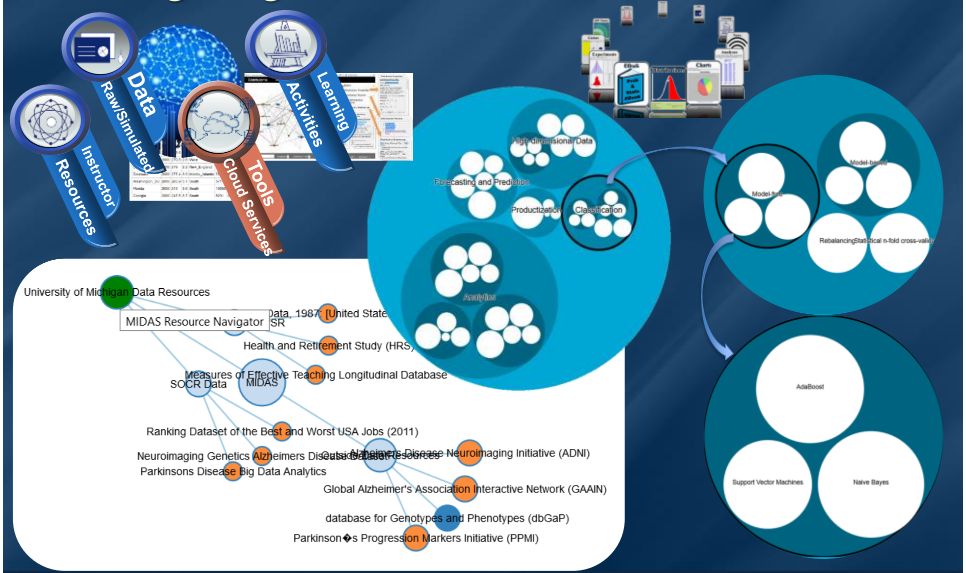
http://Predictive.Space



Exemplary Course Plans

Student's Core Field of Study	Rank	Semester 1	Semester 2	Project	Semester 3	Other within discipline	Other trans-disciplinary
Statistics	MS	EECS 584	Biostats 646	Neuroimaging genetics	SI 618	Stats 550	HS 851
Math	PhD	Stats 415	EECS 584	Compressive big data analytics	Biostats 615	Math 471	SI 649
Health Sciences	PhD	EECS 584	Stats 415	Big Cancer Data	Biostats 696	BIOINF 699	SI 601
CS/EE	MS	Stats 550	SI 618	Data mashing	BIOINF 699	EECS 598	HS 851
Bioinfo	MS	EECS 484	Stats 503	Bio-social analytics	SI 671	HS 853	Psych 614
Biostats	PhD	Math 571	EECS 584	Genotype-phenotype	SI 608	Biostats 646	Math 651
Information Sciences	PhD	Stats 550	Complex Systems 535	Social networks	EECS 598	SI 618	Biostats 696
Psych/PoliSci	PhD	Psych 613	TO 640	Election Stratification & Prediction	Biostat 521	Psych 614	HS 853

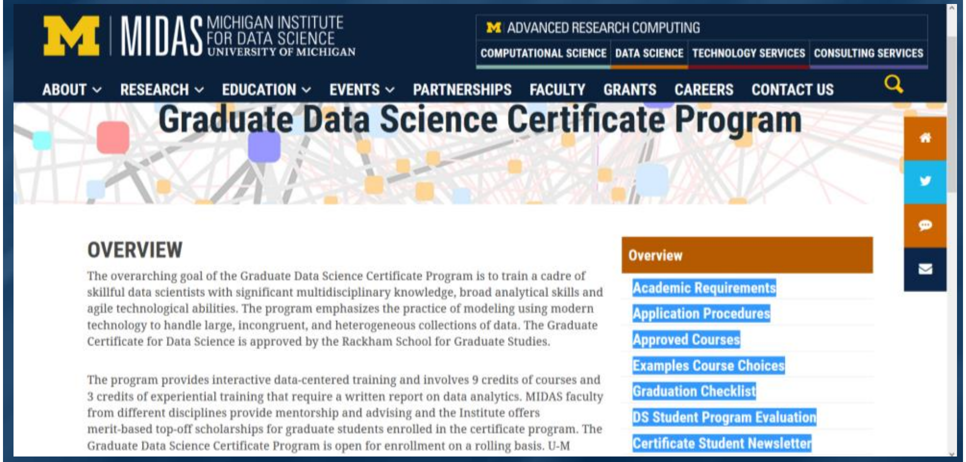
Navigating the MIDAS Curricular Materials


Data
Raw/Simulated
Instructor
Resources
Tools
Cloud Services
Learning
Activities
University of Michigan Data Resources
MIDAS Resource Navigator
Data, 1987: [United State
SR
Health and Retirement Study (HRS)
Measures of Effective Teaching Longitudinal Database
SOCR Data
MIDAS
Ranking Dataset of the Best and Worst USA Jobs (2011)
Neuroimaging Genetics
Alzheimers Disease Data
Outside Dataset Resources
Alzheimers Disease Neuroimaging Initiative (ADNI)
Parkinsons Disease Big Data Analytics
Global Alzheimer's Association Interactive Network (GAAIN)
database for Genotypes and Phenotypes (dbGaP)
Parkinson�s Progression Markers Initiative (PPMI)
High-dimensional Data
Forecasting and Prediction
Productization
Classification
Analytics
Model-free
Model-based
Rebalancing
Statistical n-fold cross-valid
AdaBoost
Support Vector Machines
Naive Bayes
Experiments
Ebook
Prob & Stats EBook
Distributions
Charts
Analyses
Games

http://socr.umich.edu/tests/2015/MIDAS/DataNavigator

http://socr.umich.edu/tests/2015/MIDAS/MIDAS_LearningModuleResourceNavigator

Graduate Data Science Certificate Program



Graduate Data Science Certificate Program

OVERVIEW

The overarching goal of the Graduate Data Science Certificate Program is to train a cadre of skillful data scientists with significant multidisciplinary knowledge, broad analytical skills and agile technological abilities. The program emphasizes the practice of modeling using modern technology to handle large, incongruent, and heterogeneous collections of data. The Graduate Certificate for Data Science is approved by the Rackham School for Graduate Studies.

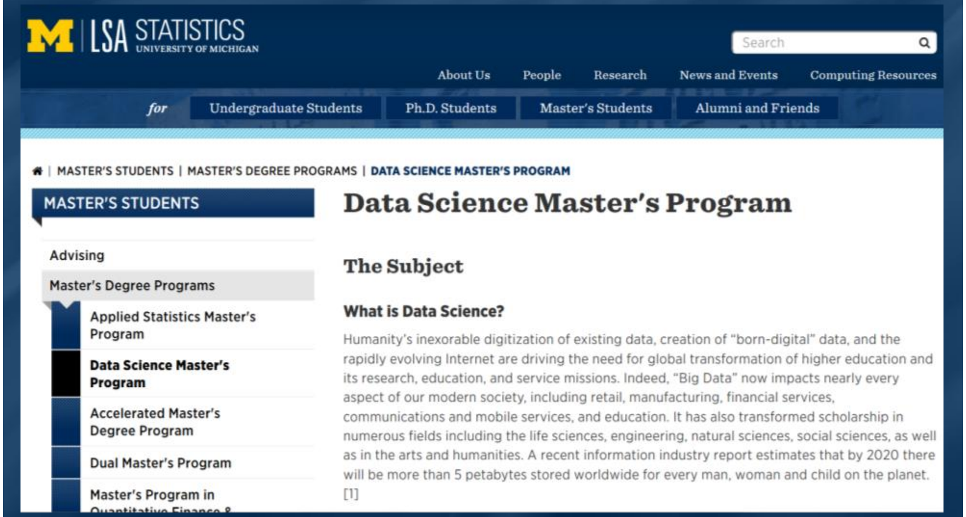
The program provides interactive data-centered training and involves 9 credits of courses and 3 credits of experiential training that require a written report on data analytics. MIDAS faculty from different disciplines provide mentorship and advising and the Institute offers merit-based top-off scholarships for graduate students enrolled in the certificate program. The Graduate Data Science Certificate Program is open for enrollment on a rolling basis. U-M

Overview

- Academic Requirements
- Application Procedures
- Approved Courses
- Examples Course Choices
- Graduation Checklist
- DS Student Program Evaluation
- Certificate Student Newsletter

http://midas.umich.edu/certificate

MS in Data Science Degree Program (Fall'18)



MASTER'S STUDENTS | MASTER'S DEGREE PROGRAMS | DATA SCIENCE MASTER'S PROGRAM

MASTER'S STUDENTS

- Advising
- Master's Degree Programs
  - Applied Statistics Master's Program
  - Data Science Master's Program
  - Accelerated Master's Degree Program
  - Dual Master's Program
  - Master's Program in Quantitative Finance &

Data Science Master's Program

The Subject

What is Data Science?

Humanity's inexorable digitization of existing data, creation of "born-digital" data, and the rapidly evolving Internet are driving the need for global transformation of higher education and its research, education, and service missions. Indeed, "Big Data" now impacts nearly every aspect of our modern society, including retail, manufacturing, financial services, communications and mobile services, and education. It has also transformed scholarship in numerous fields including the life sciences, engineering, natural sciences, social sciences, as well as in the arts and humanities. A recent information industry report estimates that by 2020 there will be more than 5 petabytes stored worldwide for every man, woman and child on the planet. [1]

https://lsa.umich.edu/stats/masters_students/mastersprograms/data-science-masters-program.html

Examples of Core Grad DS Courses

- ❑ Computational Data Science (EECS 598)

  http://midas.umich.edu/computational-data-science-eecs-598-bioinf-598

- ❑ Data Science and Predictive Analytics (HS650)

  http://Predictive.Space

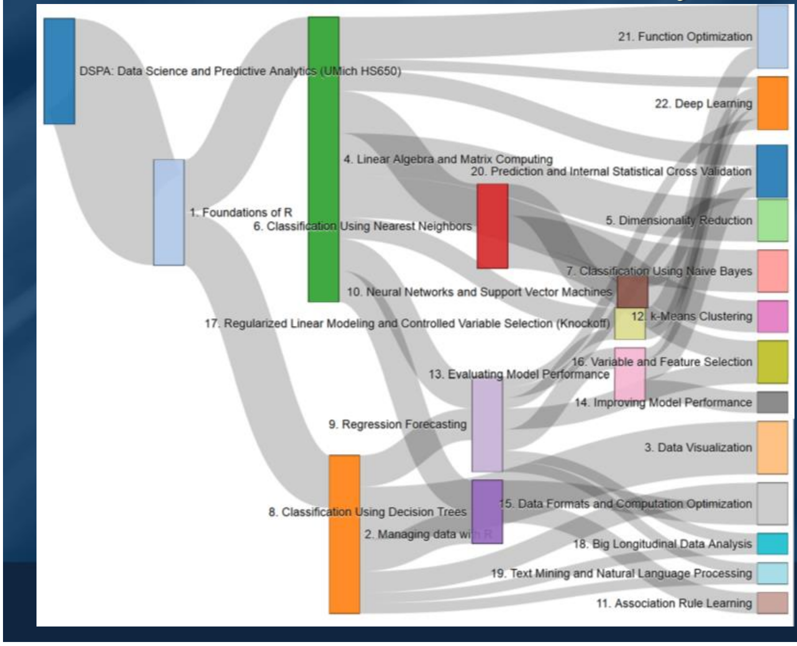
Data Science and Predictive Analytics (HS650)

There are expected variations in student backgrounds, interests, motivations, expectations, and learning styles. These prerequisites serve as a guideline of the foundational knowledge and experience for the successful completion of the Program

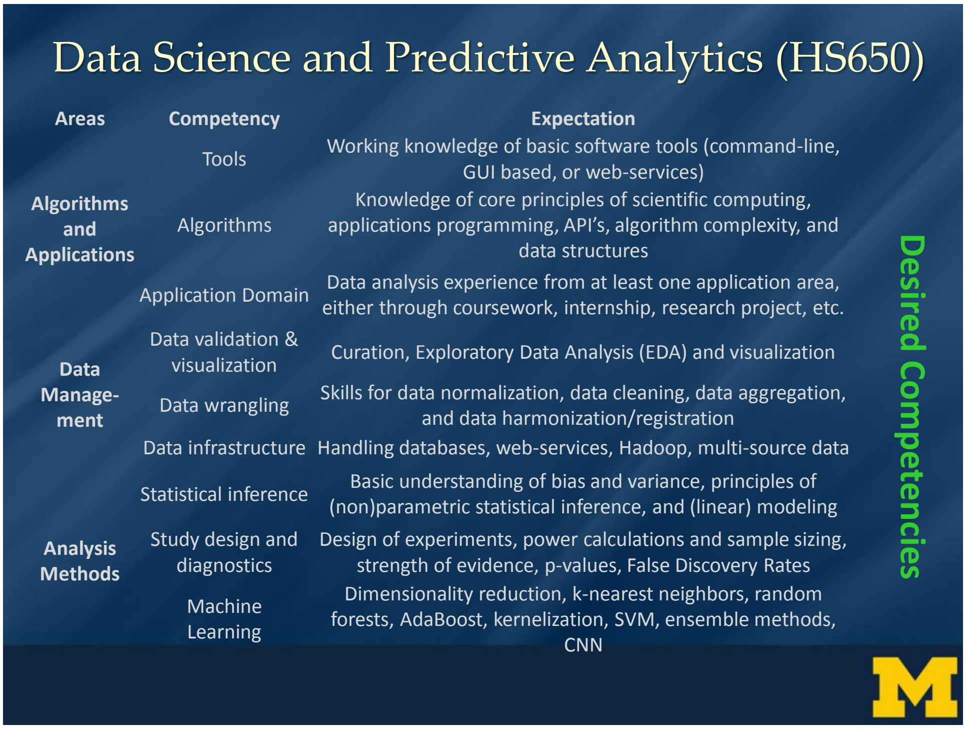
Prerequisites	Skills	Rationale
BS Degree or Equivalent	Quantitative methods/analytics training and coding skills	The DSPA graduate-level course requires a minimum level of quantitative skills
Quantitative Training	Undergraduate calculus, linear algebra and introduction to probability and statistics	These represent entry level skills required for the DSP course
Coding Experience	Exposure to software development or programming on the job or in the classroom	Most DS practitioners need substantial experience with Java, C/C++, HTML5, Python, PHP, SQL/DB
Motivation	Significant interest and motivation to pursue quantitative data analytic applications	Dedication for prolonged and sustained immersion into hands-on and methodological research

Prerequisites

Data Science and Predictive Analytics (HS650)


DSPA: Data Science and Predictive Analytics (UMich HS650)
1. Foundations of R
4. Linear Algebra and Matrix Computing
6. Classification Using Nearest Neighbors
10. Neural Networks and Support Vector Machines
17. Regularized Linear Modeling and Controlled Variable Selection (Knockoff)
13. Evaluating Model Performance
9. Regression Forecasting
8. Classification Using Decision Trees
2. Managing data with R
21. Function Optimization
22. Deep Learning
20. Prediction and Internal Statistical Cross Validation
5. Dimensionality Reduction
7. Classification Using Naive Bayes
12. k-Means Clustering
16. Variable and Feature Selection
14. Improving Model Performance
3. Data Visualization
15. Data Formats and Computation Optimization
18. Big Longitudinal Data Analysis
19. Text Mining and Natural Language Processing
11. Association Rule Learning

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

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

Acknowledgments

MIDAS Education & Training Committee

Ivo Dinov HBBS/Bioinfo, Honglak Lee, CoE/EECS, Sebastian Zöllner, SPH, Richard Gonzalez, ISR/PSY/LS&A, Kerby Shedden, Stats/LS&A

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

Michigan Institute for Data Science (MIDAS)
University of Michigan