Michigan Institute for Data Science (MIDAS)

Foundations, Challenges & Opportunities

Ivo D Dinov

www.MIDAS.umich.edu

National Big Data Science Curricula Constellation

Recently established Data Science institutes 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 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**
- **Michigan Institute for Data Science (MIDAS)**
- **MIDAS Executive Committee**
- **Co-Directors**
- **Management Committee**
- **MIDAS Faculty**
  - **Core Faculty**
  - **Affiliate Faculty**

**Advisory**
- **External Advisory Board**

**Research**
- **Research Centers**
  - Transportation
  - Learning Analytics
  - Health Science
  - Social Science
- **Research Incubation**
  - New Initiatives
- **Research Working Groups**
  - Data Science
- **Data Science Workshops**

**Education & Training**
- **Education & Training Committee**

**Industry Engagement**
- **Industry-Sponsored Research**
- **Business Engagement Center**

**Faculty Engagement**
- **Faculty Engagement & Recruitment Committee**
- **Faculty Working Groups**
- **Faculty Awards**

**Research Centers**
- **Rackham**
- **Student Programs**
  - Summer Camps
  - Student Awards

**Dissemination**
- **Annual Symposium**
- **Annual Forum**
- **Seminar Series**

**MIDAS ROI**

**Calculating ROI: MIDAS-led and MIDAS faculty Sponsored Research**

<table>
<thead>
<tr>
<th>Units</th>
<th>Gov’t / Fdn Funding</th>
<th>Industry Funding</th>
<th>Gov’t / Fdn Funding</th>
<th>Industry Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>$1,898,898</td>
<td>$3,481,348</td>
<td>$7,990,450</td>
<td>$6,483,343</td>
</tr>
<tr>
<td>LS&amp;A</td>
<td>$2,920,442</td>
<td>$165,000</td>
<td>$3,290,528</td>
<td>$11,155,442</td>
</tr>
<tr>
<td>Medicine</td>
<td>$3,458,479</td>
<td>$5,492,096</td>
<td>$6,492,096</td>
<td>$2,310,241</td>
</tr>
<tr>
<td>Information</td>
<td>$233,783</td>
<td>$164,991</td>
<td>$5,625,946</td>
<td>$149,989</td>
</tr>
<tr>
<td>Public Health</td>
<td>$952,549</td>
<td>$315,986</td>
<td>$622,649</td>
<td>$107,600</td>
</tr>
<tr>
<td>Nursing</td>
<td>$650,000</td>
<td>$96,078</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UMTRI</td>
<td>$466,061</td>
<td>$315,986</td>
<td>$622,649</td>
<td>$107,600</td>
</tr>
<tr>
<td>Ross</td>
<td>$118,360</td>
<td>$315,986</td>
<td>$622,649</td>
<td>$107,600</td>
</tr>
<tr>
<td>ISR</td>
<td>$141,875</td>
<td>$4,294,930</td>
<td></td>
<td>$300,000</td>
</tr>
<tr>
<td>MIDAS</td>
<td>$25,000</td>
<td>$4,284,930</td>
<td></td>
<td>$300,000</td>
</tr>
<tr>
<td>ISR ARC</td>
<td>$401,540</td>
<td>$4,284,930</td>
<td></td>
<td>$300,000</td>
</tr>
</tbody>
</table>

**Total**
- $10,371,026
- $4,395,760
- $28,645,597
- $11,528,804

**Rationale for Calculating ROI as reported to Executive Committee in December**

- **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 ROI**

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

<table>
<thead>
<tr>
<th>Big Data</th>
<th>Information</th>
<th>Knowledge</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Observations</td>
<td>Processed Data</td>
<td>Maps, Models</td>
<td>Actionable Decisions</td>
</tr>
<tr>
<td>Data Aggregation</td>
<td>Data Fusion</td>
<td>Causal Inference</td>
<td>Treatment Regimens</td>
</tr>
<tr>
<td>Data Scrubbing</td>
<td>Summary Stats</td>
<td>Networks, Analytics</td>
<td>Forecasts, Predictions</td>
</tr>
<tr>
<td>Semantic-Mapping</td>
<td>Derived Biomarkers</td>
<td>Linkages, Associations</td>
<td>Healthcare Outcomes</td>
</tr>
</tbody>
</table>

Dinov, Springer, 2018
Characteristics of Big Biomed Data

Dinov, et al. (2014)

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

<table>
<thead>
<tr>
<th>BD 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>Incomplete</td>
<td>Reliable management of missing data</td>
</tr>
</tbody>
</table>

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.
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

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

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

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

Graduate Data Science Certificate Program

http://midas.umich.edu/certificate
Examples of Core Grad DS Courses

- **Computational Data Science (EECS 598)**
  

- **Data Science and Predictive Analytics (HS650)**
  
  [http://Predictive.Space](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.

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

---

**Prerequisites**

- 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.

---

**Data Science and Predictive Analytics (HS650)**

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

---

**http://Predictive.Space**
## 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</td>
<td>Tools</td>
<td>Working knowledge of basic software tools (command-line, GUI based, or web-services)</td>
</tr>
<tr>
<td>Applications</td>
<td></td>
<td>Knowledge of core principles of scientific computing, applications programming, API's, algorithm complexity, and data structures</td>
</tr>
<tr>
<td>Data Management</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></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>

**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
- George Alter: Institute for Social Research, History, LS&A
- Brian Astley: Computational Medicine and Bioinformatics, SoM
- Mike Cafarella: Computer Science and Engineering, CoE
- Ivo Dinov, Chair, Leadership and Effectiveness Science, Bioinformatics, SoN/SoM
- Karthik Durasamy: 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