Education & Training Program Presenters

Ivo D. Dinov
Assoc. Director
MIDAS
Ed & Training

Erin Shellman
Research Scientist
AWS

Patrick Harrington
Co-Founder
Chief Data Scientist
CompGenome.com

Nandit Soparkar
CEO, Ubiquiti
MIDAS Data Science Education & Training: Challenges & Opportunities

Ivo D. Dinov
Associate Professor, School of Nursing
Associate Education Director, Michigan Institute for Data Science (MIDAS)

University of Michigan
http://MIDAS.umich.edu/education
Recently established Data Science institutes and curricular programs
MIDAS Big Data Ecosystem

http://socr.umich.edu/docs/BD2K/BigDataResourceome.html
MIDAS Focus on Developing Big Data Skills

- **Listening**: streams, analyzing sentiment, intent and trends;
- **Looking**: searching, indexing and memory management of heterogeneous datasets; Raw, derived or indexed data as well as meta-data;
- **Programming**: Handling Map-Reduce/HDFS, No-SQL DB, protocol provenance, algorithm development/optimization, pipeline workflows;
- **Inferring**: Principles of data analyses, Bayesian modeling, inference, uncertainty and quantification of likelihoods; Reasoning & logic; Analytics: Regression, feature selection, dimensionality reduction, temporal patterns, validation; Actionable Knowledge
- **Machine Learning**: Classification, clustering, mining, information extraction, knowledge retrieval, decision making;
- **Predicting**: Forecasting, neural models, deep learning, and research topics;
- **Summarizing**: Presentation of data, processing protocol, analytics provenance, visualization, synthesis
Technological advances and students’ IT skills far outpace educators’ technological expertise and the current rate of IT adoption in college curricula.

Lack of open learning resources and interactive interoperable platforms.

Difficulty sharing data, tools, materials, and activities across different disciplines.

Limited technology-enhanced continuing instructor education opportunities.

Discipline-specific knowledge boundaries.

Skills for communication and teamwork involving cooperation in dynamic groups of dispersed researchers.

Ability to aggregate & harmonize data (e.g., fusion of qualitative and quantitative elements).
Core MIDAS Education Components

- **Drivers**: Motivations, datasets (6D of Big Data), challenges, applications
- **Methods**: foundational and trans-disciplinary scientific techniques
- **Tools**: web-services, software, code, platforms
- **Analytics**: practice of data interrogation
Existential Data Science Education & Training

- Undergraduate DS Degree Program
  - Stats + Engineering
  - Started Fall 2015
  - www.eecs.umich.edu/eecs/undergraduate/data-science

- DS Summer Institute (Public Health)
  - 20+ faculty, 40+ trainees
  - Full support/in residence (1 month)
  - BigDataSummerInst.sph.umich.edu

- Big Data Summer Bootcamp
  - 5 years running, Business-Engineering
  - Practice of Econ-Bio-Social Analytics
  - ibug-um15.github.io/2015-summer-camp

- Graduate DS Certificate Program
  - Transdisciplinary (SM,SPH,LS&A,SN,CoE,SI)
  - 12 cr, Modeling+Technology +Practice
  - http://MIDAS.umich.edu/certificate
MIDAS Education & Training (Going Forward)

- Graduate Data Science Certificate Program
  - Enrollment of 100 (UMich) students
- Develop the Online Graduate Data Science Certificate Curriculum
- Develop a 5–yr dual undergrad–grad MS Program in Data Science
- DS Summer Institute (Public Health)
- Big Data Summer Bootcamp
- MIDAS Seminar Series (academia, industry, government, partners)
  - Web–streamed and archived for broader impact
- National Events
  - BD2K, Midwest Big–Data–Hub, Math, Stats, Computer Vision, Machine Learning, Informatics, …
- Funding: NIH, NSF, Foundation: Research & Traineeship Applications
Michigan Difference in Data Science Education & Training

- Public–Private–Partnership (PPP) Model
- Integration of Research + Practice + Education
- Problems + Modeling + Technology + Applications
Probability and Statistics EBook
- Motivation, Methods, Techniques, Practice
  - wiki.socr.umich.edu/index.php/EBook

Data sets
- 100’s of Research, Observed & Simulated
  - wiki.socr.umich.edu/index.php/SOCR_Data

Hands-on Activities (Team-focused)
- Modeling, Inference, Concept Demos
  - wiki.socr.umich.edu/index.php/SOCR_EduMaterials

Applets and Webapps
- Over 500 Web tools
- Distributed services based on Java, HTML5/JavaScript

Features: Free, Cloud-service, no-barrier access, LGPL/CC-BY
Ref: Dinov et al., TS (2013), DOI: 10.1111/test.12012

SOCR has over 9.6M users worldwide since 2002
www.SOCR.umich.edu
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
MIDAS plans to:

- Develop new hands-on **MOOC DS courses** (data, learning modules, services, code snippets, applications/partners) …

- Deploy MIDAS **Training as a Service (TaaS)** MOOC platform

- Compile **Datasets** (research-derived, observational, simulated) – identify, aggregate, manage, navigate, and service exemplary Big Data sets

- **Faculty/Mentors**—instructors, collaborators, partners, students and staff to develop a core DS course-series (4 courses)

- Instructional resources and complete end-to-end **learning modules**

- **Track and Validate the DSEP Program** – annual reports (MIDAS ETC), review of evaluations, sustainability (financial, resources, space), impact
How to Engage & Partner in MIDAS Education?

Partners
- Open-Source Community
- Academia
- Trainees
- Industry
- Non-profit
- Government
- Philanthropy
- Community Orgs

Activities
- Drivers
  - Challenges
  - Datasets
  - Applications
- Methods
- Technologies
- Tools/Services
- Training Opps
  - Mentorship
  - DS Projects
  - Internships
MIDAS Education & Training Committee
Ivo Dinov, Margaret Hedstrom, Honglak Lee, Sebastian Zöllner, Richard Gonzalez, Kerby Shedden

Other Contributing MIDAS Faculty

Engineering
Alfred Hero: Electrical Engineering and Computer Science; Biomedical Engineering, Stats
H. V. Jagadish: Electrical Engineering and Computer Science
Mike Cafarella: Computer Science and Engineering
Karthik Duraisamy: Atmospheric, Oceanic, and Space Sciences
Judy Jin: Industrial & Operations Engineering
Dragomir Radev: School of Information; Computer Science and Engineering; Linguistics

Information & Health Sciences
Brian Athey: Computational Medicine and Bioinformatics, Medicine
Carl Lagoze: School of Information
Qiaozhu Mei: School of Information
Jeremy Taylor, Biostatistics, Public Health

Basic Sciences
Vijay Nair: Statistics & Engineering
George Alter: Institute for Social Research, History
Christopher Miller: Physics & Astronomy
August Evrard: Physics & Astronomy
Anna Gilbert: Mathematics, Engineering
Stephen Smith: Ecology and Evolutionary Biology
Ambuj Tewari: Statistics; Computer Science and Engineering

Contact
http://MIDAS.umich.edu
midas-contact@umich.edu
Dinov@umich.edu
Education & Training Program Presenters

Ivo D. Dinov
Assoc. Director MIDAS
Ed & Training

Erin Shellman
Research Scientist, AWS

Patrick Harrington
Co-Founder/Chief Data Scientist
CompGenome.com

Nandit Soparkar
CEO, Ubiquiti
From a descriptive to a constructive definition of Big Data

- **IBM’s 4V’s** – volume, variety, velocity (speed) and veracity (reliability)
- **MIDAS Big Data Characterization** – identifies gaps, challenges & needs

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)
Kryder’s law: Exponential Growth of Data

Data volume Increases faster than computational power

Dinov, et al., 2014
# Data Science Training Program: Core Curriculum

<table>
<thead>
<tr>
<th>Themes</th>
<th>Training</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scrubbing</td>
<td>Big Data Management</td>
<td>Big Data technology, Searching, indexing, memory management, Information extraction, feature selection, Supervised and unsupervised-learning, stream mining</td>
</tr>
<tr>
<td></td>
<td>Big Data Representation</td>
<td>Matrix Representation of Sets, Minhashing, Jaccard Similarity, Distance Measures, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance, Networks, Graph similarity, Sets as Strings, Prefix Indexing, Data Streams and Processing, Representative Samples, Bloom Filter, Flajolet-Martin Algorithm, TF/IDF, MoM, MLE, Alon-Matias-Szegedy Algorithm for Second Moments, Datar-Gionis-Indyk-Motwani Algorithm (DGIM)</td>
</tr>
<tr>
<td>Data Mining</td>
<td>Mining Social-Network Graphs</td>
<td>Bio-Social Network Graphs, Root-Mean-Square Error, UV- Decomposition, The NetFlix Challenge, Tweeter Challenge, Distances and Clustering of Social-Network Graphs, Network Cluster Betweenness, Complete Bipartite Subgraphs, Graph Partition, Matrices and Graphs, Eigen-values/ vectors of the Laplacian Matrix, Graph Neighborhood Properties, Diameter of a Graph, Transitive Closure and Reachability, Crowdsourcing Algorithms</td>
</tr>
<tr>
<td></td>
<td>Modeling</td>
<td>Statistical Modeling, Machine Learning, Computational Modeling, Feature Extraction, Power Laws, Map-Reduce/Hadoop</td>
</tr>
<tr>
<td></td>
<td>Link Analysis</td>
<td>PageRank, Spider Traps, Taxation and loops in Big Data traversal, Random Walks</td>
</tr>
<tr>
<td>Data Inference</td>
<td>Clustering</td>
<td>Curse of Dimensionality, Classification and Regression Trees/Random Forests, Hierarchical Clustering, K-means Algorithms, Bradley, Fayyad, and Reina (BFR) Algorithm, CURE Algorithm, Clusters in the GRGPF Algorithm</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Random Forest, SVM, Neural Networks, Latent Class Models, Finite Mixture Models</td>
</tr>
<tr>
<td></td>
<td>Dimensionality Reduction</td>
<td>Eigenvalues and Eigenvectors, Principal-Component Analysis, Singular-Value Decomposition, Wrapper-Based vs Filter-Based Feature Selection, Independent Components Analysis, Multidimensional Scaling, CUR Decomposition</td>
</tr>
<tr>
<td>Big Data</td>
<td>Information</td>
<td>Knowledge</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Raw Observations</td>
<td>Processed Data</td>
<td>Maps, Models</td>
</tr>
<tr>
<td>Data Aggregation</td>
<td>Data Fusion</td>
<td>Causal Inference</td>
</tr>
<tr>
<td>Data Scrubbing</td>
<td>Summary Stats</td>
<td>Networks, Analytics</td>
</tr>
<tr>
<td>Semantic-Mapping</td>
<td>Derived Biomarkers</td>
<td>Linkages, Associations</td>
</tr>
</tbody>
</table>