

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

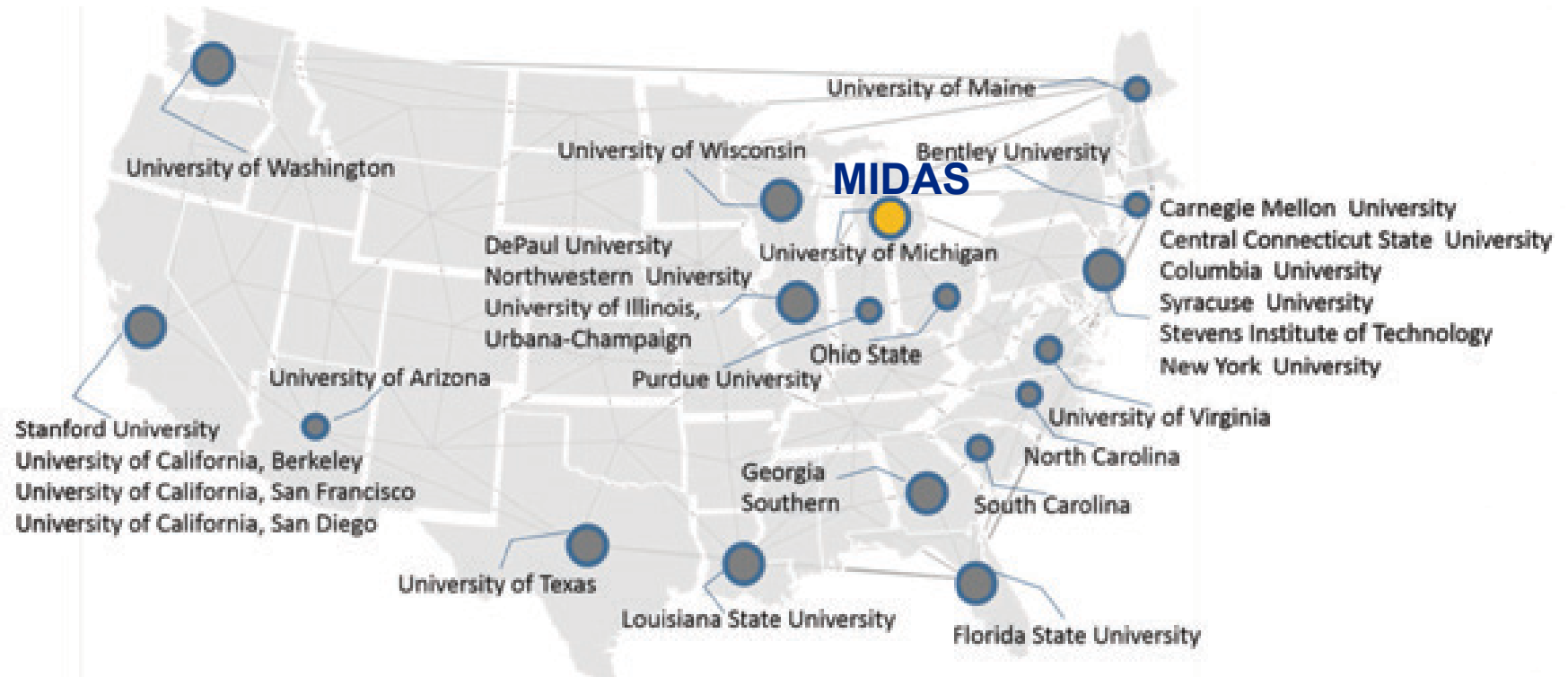
Associate Education Director, Michigan Institute for Data Science (MIDAS)

University of Michigan

<http://MIDAS.umich.edu/education>

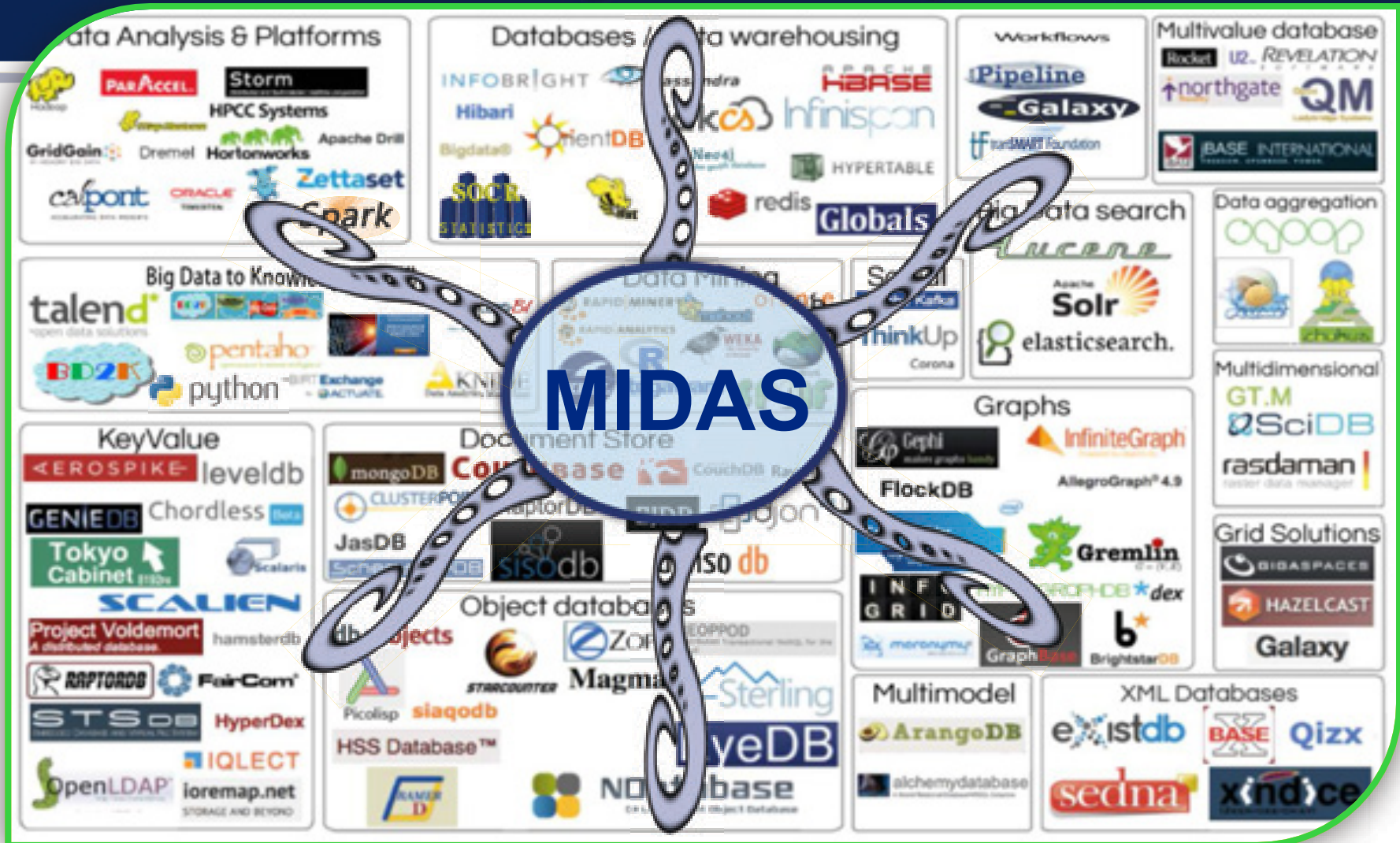


National Big Data Science Curricula Constellation



Recently established Data Science institutes and curricular programs

MIDAS Big Data Ecosystem



<http://socr.umich.edu/docs/BD2K/BigDataResourceome.html>



MIDAS

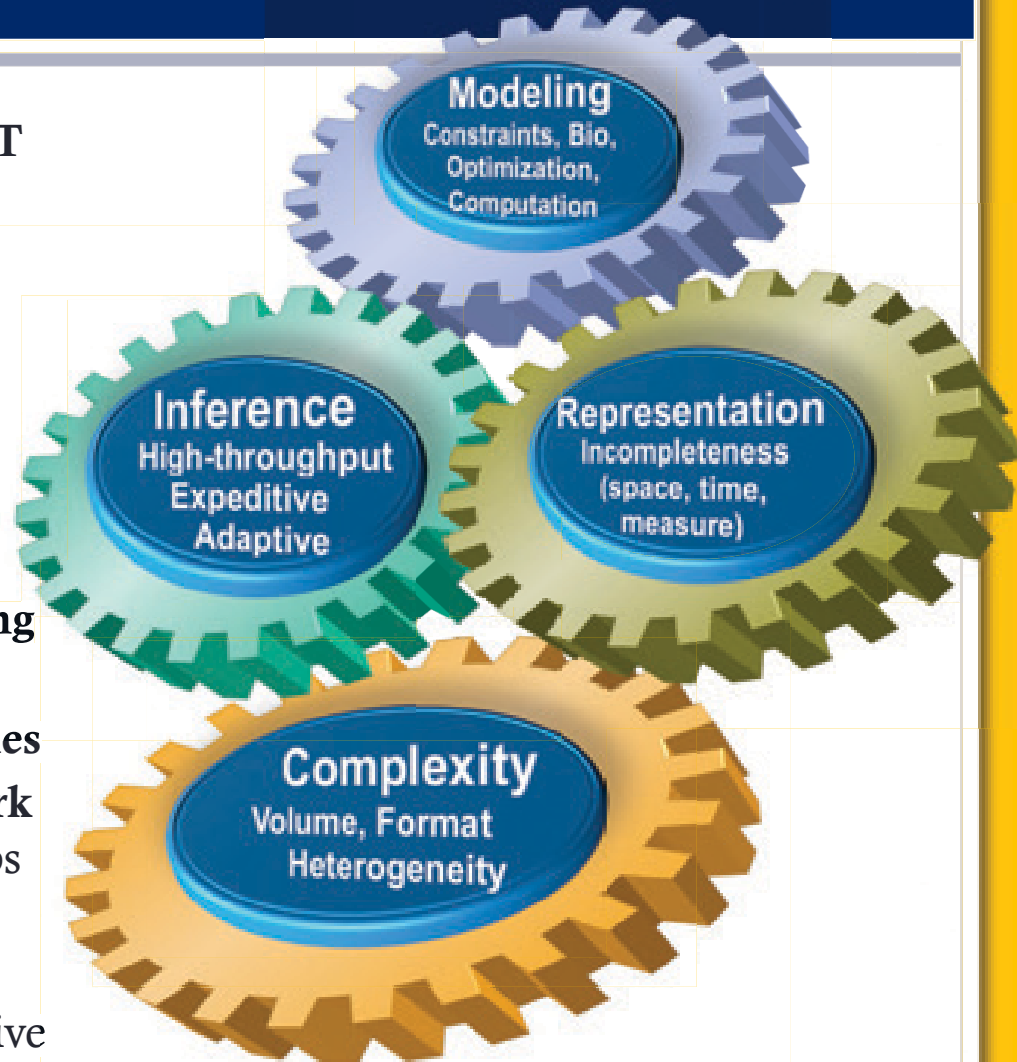
FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

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

Specific Data Science Training Challenges

- 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 & 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

Existent 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
- Graduate DS Certificate Program
 - [ibug-um15.github.io/2015-summer-camp](https://github.com/ibug-um15/2015-summer-camp)
 - Transdisciplinary (SM,SPH,LS&A,SN,CoE,SI)
 - 12 cr, Modeling+Technology+Practice
 - <http://MIDAS.umich.edu/certificate>

Undergraduate Program in Data Science

Undergraduate Programs

- Computer Engineering
- Computer Science - Eng
- Computer Science - USA

Transforming Analytical Learning in the Era of Big Data
An Undergraduate Summer Institute in Biostatistics
June 1-26, 2015, Department of Biostatistics, University of Michigan Ann Arbor

Source from Home

Main Menu

COS Big Data Summer Camp
University of Michigan

Room R0220 - Ross School of Business - 701 Tappan Street, Central Campus
June 1-4, 2015
9:00 am - 5:00 pm

General Information

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Home

Data Science Initiative

About MIDAS

Affiliated Faculty

Education & Training

Graduate Data Science (DS) Certificate Program

Challenges Initiatives

Industry Engagement

Events

Grant Opportunities

Contact Us

Graduate Data Science (DS) Certificate Program

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 Data Science Certificate Program is now open for Fall 2015 enrollment. U-M graduate students from any field are eligible to enroll. Merit-based Graduate Data Science top-off fellowships may be provided. Minority and underrepresented students are strongly encouraged to enroll and complete the Data Science training program. The 9 course credits must be outside the student's regular degree program/department. Completion of the program is expected in 3-4 semesters. The Data Science Certificate program aims to provide core experiences in:

DS Certificate Program

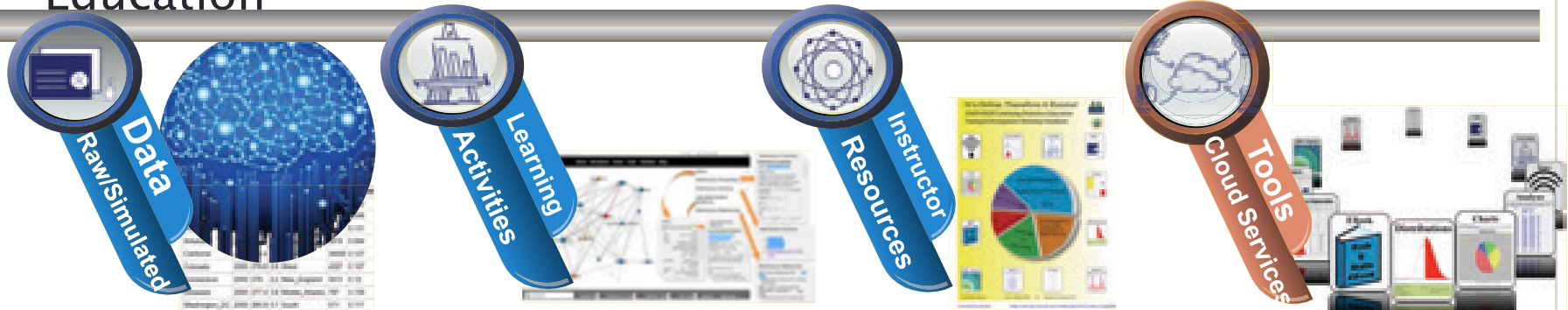
- [Academic Requirements](#)
- [Application Procedures](#)
- [Approved Courses](#)
- [Example Course Choices](#)
- [Graduation Checklist](#)

MIDAS Education & Training (Going Forward)

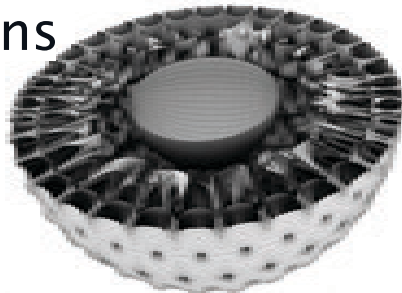
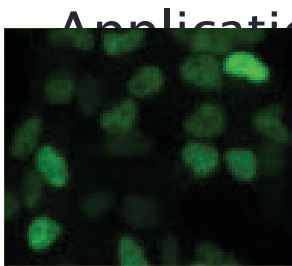
- Graduate Data Science Certificate Program
 - Enrollment of 100 (UMich) students
- Develop the Online Graduate Data Science Certificate Curriculum
- Develop a 5-yr dual undergrad-grad MS Program in Data Science
- DS Summer Institute (Public Health)
- Big Data Summer Bootcamp
- MIDAS Seminar Series (academia, 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

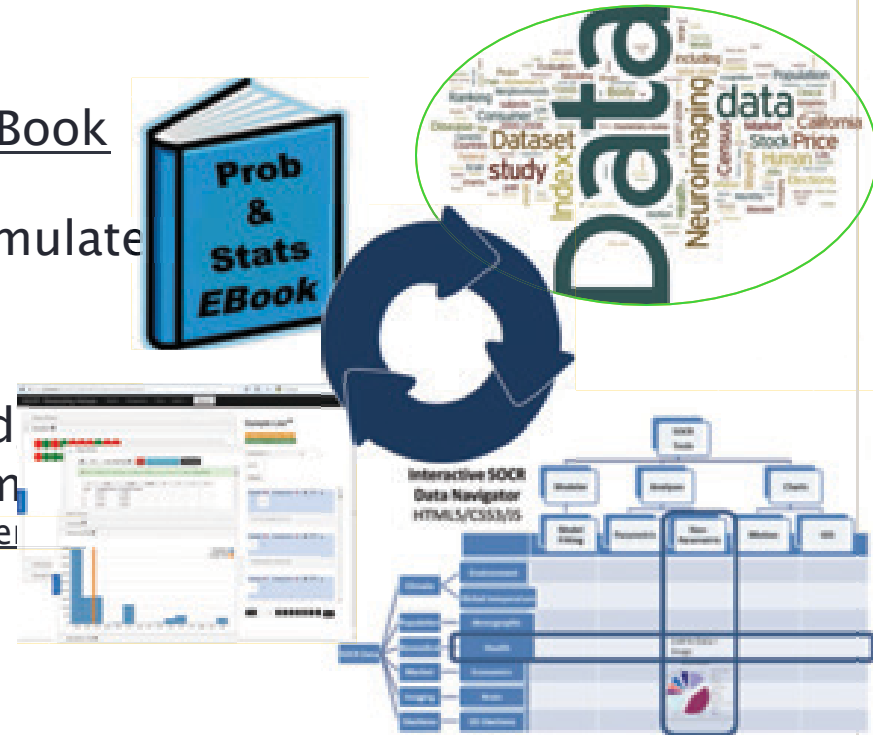


Geometric Measure	Mathematical Formula
Volume	$V = \int \int \int \text{d}x \text{d}y \text{d}z$
Surface Area	$A_s = \int \int \sqrt{1 + f_x^2 + f_y^2} \text{d}x \text{d}y$
Mean Curvature	$H = \frac{1}{2} (\kappa_1 + \kappa_2)$
Shape Index	$S = \frac{1}{2} \ln \left(\frac{\kappa_1 + \kappa_2}{\kappa_1 \kappa_2} \right)$
Curvature	$\kappa = \frac{1}{R}$
Principal Direction	$\theta = \tan^{-1} \left(\frac{f_y}{f_x} \right)$



Statistics Online Computational Resource (SOCR) Integration of Education, Research & Practice

- 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_EduMater
- 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

MIDAS-SOCR Dashboard: Big Data Fusion Example

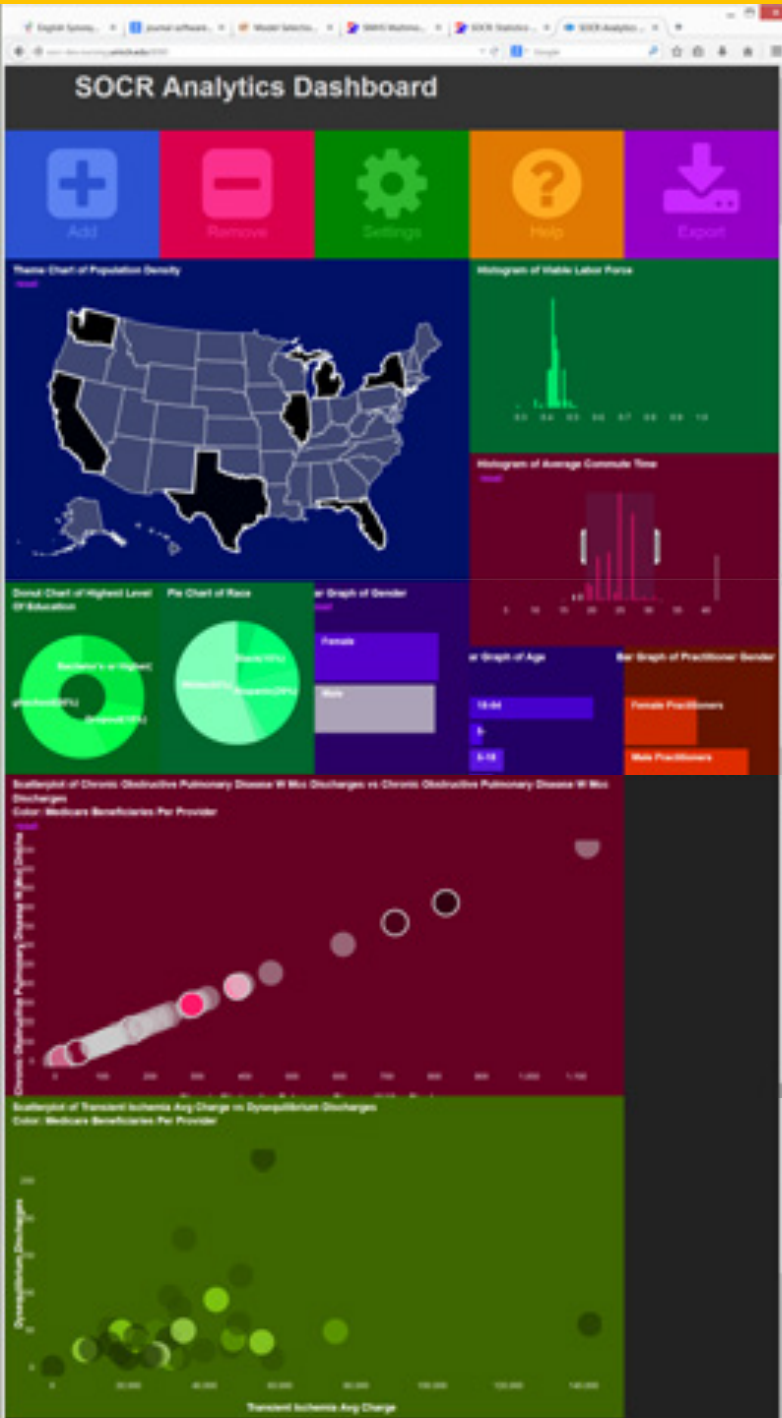
The screenshot displays the MIDAS-SOCR Dashboard, a Big Data Fusion Example. It features three overlapping web interfaces:

- Association of Health Care Journalists (AHCJ):** Shows a search bar, navigation links (Home, About AHCJ, Advertise with AHCJ, Center for Excellence in Health Care Journalism, Membership, Training), and a list of core topics (Health Re, Aging, Oral Heal, Med. Stu).
- United States Census Bureau:** Displays the "ZIP Code" search interface with options for "People" and "Business". It includes a search bar, a list of "Your Selections" (e.g., "Demographics"), and a list of "Search using the options below" (Topics, Geographies, Race and Ethnic Groups, Industry Codes, EEO Occupation Codes).
- Bureau of Labor Statistics:** Shows the "Databases, Tables & Calculators by Subject" page. It includes a search bar, a list of "On This Page" (Inflation & Prices, Employment, Unemployment, Pay & Benefits, Spending & Time Use, Productivity, Workplace Injuries, International, Employment Projections, Regional Resources, Historical News Release Tables, Maps, Calculators, Public Data API), and a table of "Inflation & Prices" data.

The "Inflation & Prices" table lists the following data series:

Database Name	Special Notice	Top Picks	One Screen	Multi-Screen	Tables	Text Files
Prices - Consumer						
All Urban Consumers (Current Series) (Consumer Price Index - CPI)		★	🔍	📄	📊	📄
Urban Wage Earners and Clerical Workers (Current Series) (Consumer Price Index - CPI)		★	🔍	📄	📊	📄
All Urban Consumers (Chained CPI) (Consumer Price Index - CPI)	⚠️	★	🔍	📄	📊	📄
Average Price Data	⚠️	★	🔍	📄	📊	📄

The URL <http://socr.umich.edu/HTML5/Dashboard> is displayed at the bottom left. The MIDAS logo and the text "MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN" are visible at the bottom right.



Big Data Analytics

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

- 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
MICHIGAN INSTITUTE
FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

MIDAS Online Data Science Education Program (DSEP)

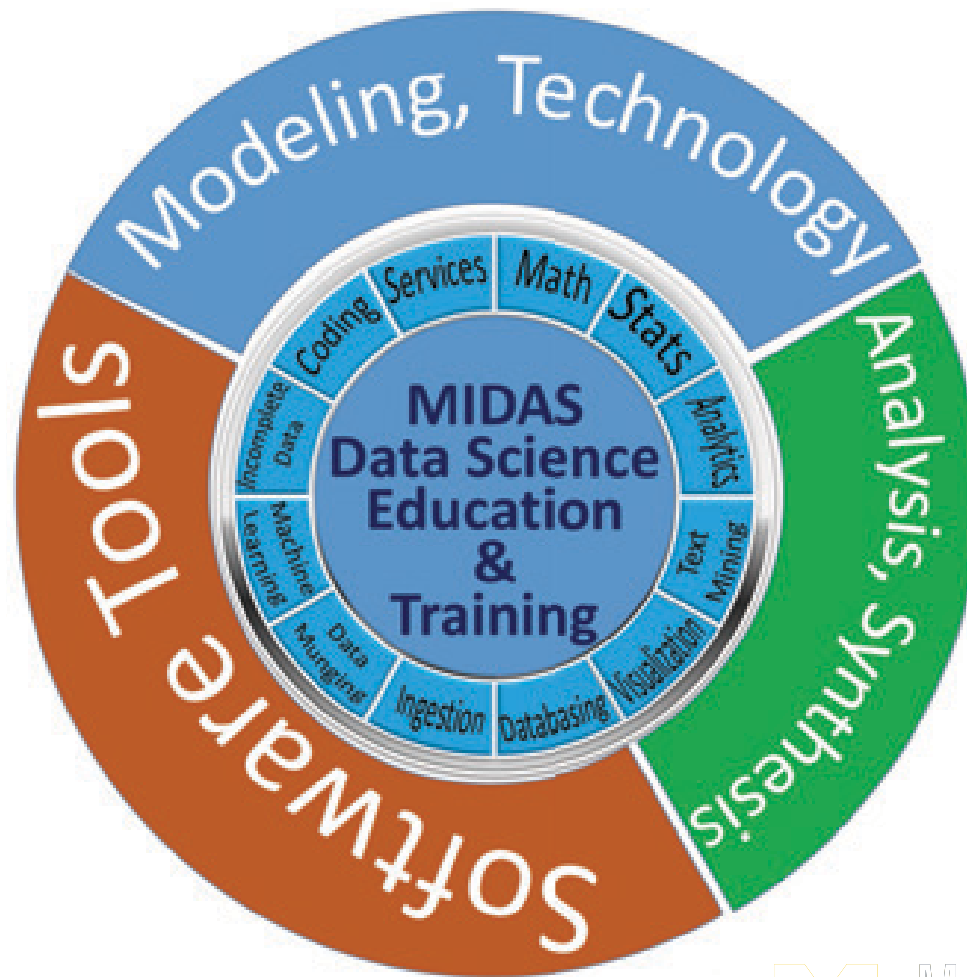
MIDAS plans to:

- 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 Team

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](mailto:midas-contact@umich.edu)

Dinov@umich.edu

M | **MIDAS** MICHIGAN INSTITUTE
FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

Education & Training Program Presenters

Ivo D. Dinov
Assoc. Director MIDAS
Ed & Training



Erin Shellman
Research Scientist, AWS



Nandit Soparkar
CEO, Ubiquiti

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



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



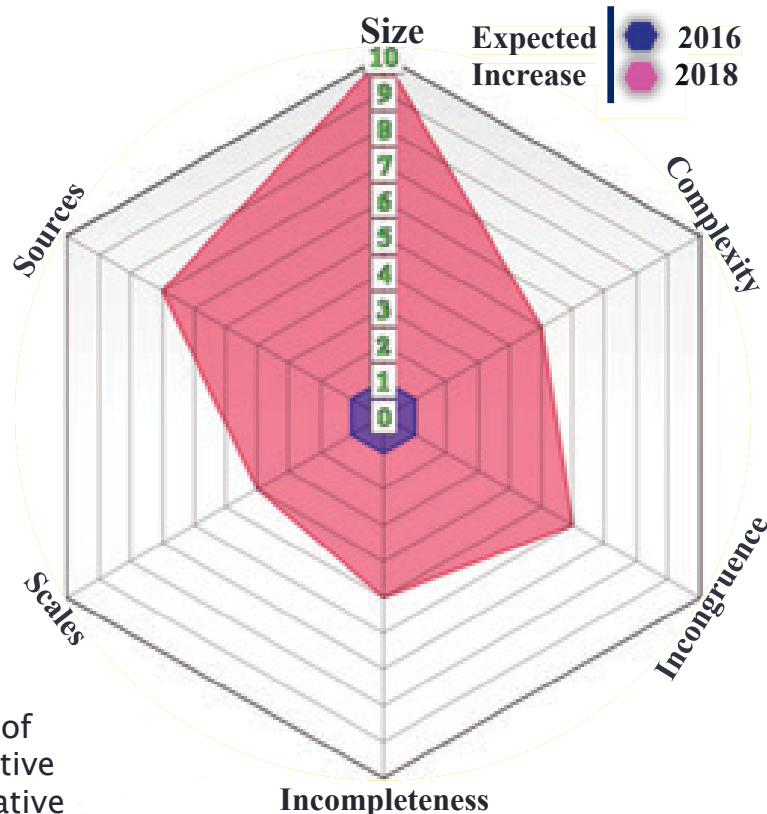
Nandit Soparkar
CEO, Ubiquiti



Big Data Science

From a descriptive to a constructive definition of Big Data

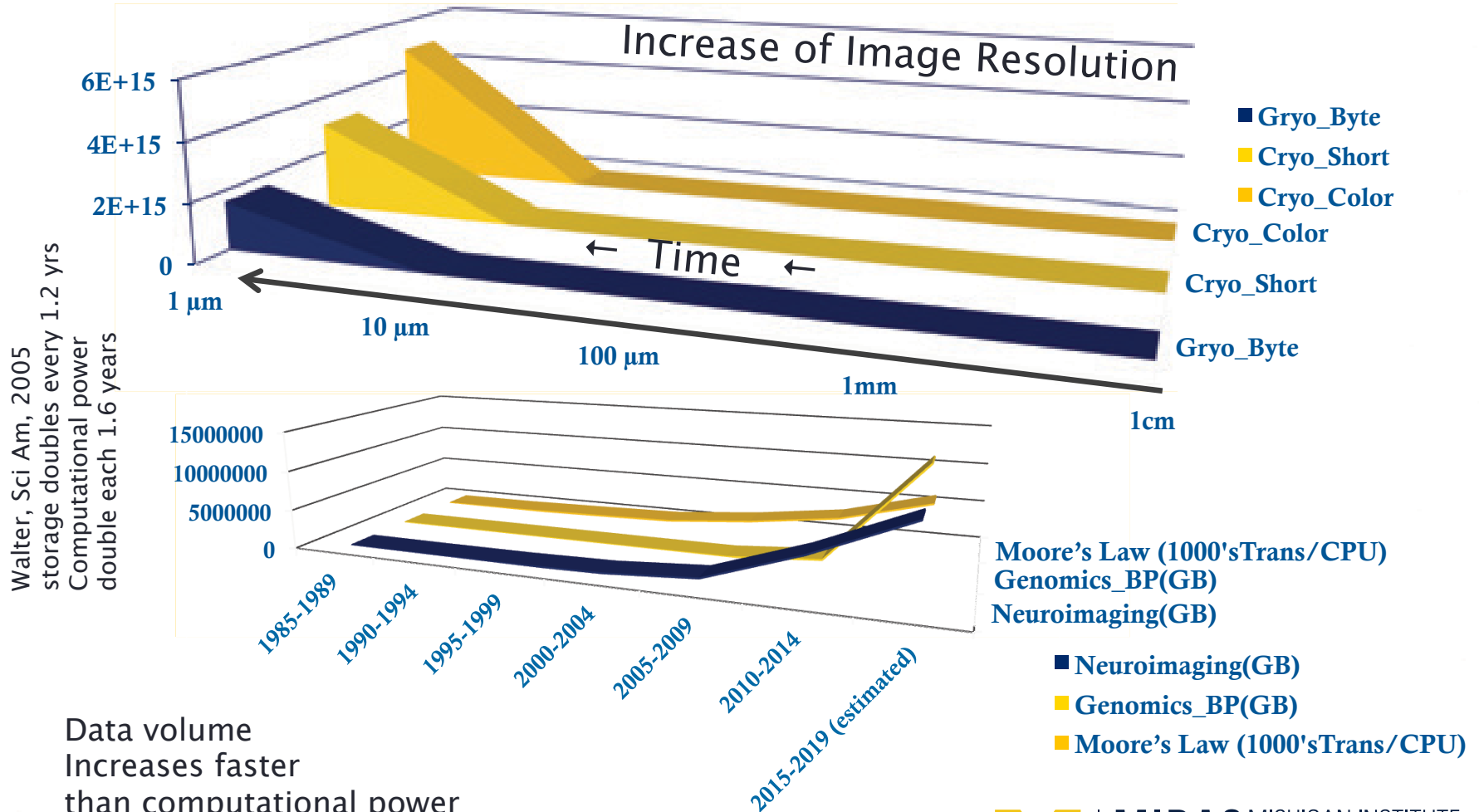
- IBM's 4V's – volume, variety, velocity (speed) and veracity (reliability)
- MIDAS Big Data Characterization – identifies gaps, challenges & needs



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

Kryder's law: Exponential Growth of Data



Data Science Training Program: Core Curriculum

Themes	Training	Examples
Data Scrubbing	Big Data Management	Big Data technology, Searching, indexing, memory management, Information extraction, feature selection, Supervised and unsupervised-learning, stream mining
	Big Data Representation	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)
Data Mining	Mining Social-Network Graphs	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
	Modeling	Statistical Modeling, Machine Learning, Computational Modeling, Feature Extraction, Power Laws, Map-Reduce/Hadoop
	Link Analysis	PageRank, Spider Traps, Taxation and loops in Big Data traversal, Random Walks
Data Inference	Clustering	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
	Classification	Random Forest, SVM, Neural Networks, Latent Class Models, Finite Mixture Models
	Dimensionality Reduction	Eigenvalues and Eigenvectors, Principal-Component Analysis, Singular-Value Decomposition, Wrapper-Based vs Filter-Based Feature Selection, Independent Components Analysis, Multidimensional Scaling, CUR Decomposition



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