Data Science at UM

Alfred Hero

Co-director, Michigan Institute for Data Science
Dept. of EECS, Dept. of BME, Dept. of Statistics
University of Michigan – Ann Arbor
June 8, 2017

midas.umich.edu
Outline

1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks
Outline

1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks
Data Science


Karl Pearson (1901)  
“On lines and planes ... of closest fit to points”

John Tukey (1962)  
“Future of data analysis”

John Tukey (1977)  
EDA

IAAI (1987)  
KDD (Detroit)
Data Science


• Developing into widely embraced multi-disciplinary field

• Elements driving evolution of data science
  – Datasets are getting larger, faster with more complex structure
  – Data frequently is poorly annotated: provenance unknown
  – Privacy concerns: anonymization, fair use, reuse, ethics
Alfred Hero, Univ. Michigan

2017 ICOS Big Data Summer Camp

Explosion in volume, velocity, variety of data

Materials Genome

- Cambridge Structural Database

Biomaterials

- Materials Genome
- 160,000 Engineering materials
- Multiscale Multiphysics

Nature Genetics 45, 1113–1120 (2013)

Cyberphysical Networks

- The Cancer Genome Atlas (TCGA)
- UM Mobility Transformation Center (MTC)

AE, CSE, CivE, ECE, IOE, ME
Data science is concerned with

• Collecting data: sensing instruments and data repositories
  • Extract maximum value from data sources for end-use
  • Fuse data from diverse sources giving actionable information

• Managing data: resilient protected databases
  • Efficiently store, annotate, access and protect data
  • Develop standard formats for diverse data types

• Analyzing data: integrated computational algorithms
  • Develop automated algorithms that handle uncertainty
  • Summarize/visualize results to maximize interpretability
Outline

1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks
Michigan Data Science Initiative

Leading advances in data-intensive and computational research

- **COMPUTATIONAL SCIENCE**
  The Michigan Institute for Computational Discovery and Engineering (MICDE) focuses on the interdisciplinary development of mathematical algorithms and models on high performance computers.

- **DATA SCIENCE**
  The Michigan Institute for Data Science (MIDAS) is the focal point for the new multidisciplinary area of data science at the University of Michigan.

- **TECHNOLOGY SERVICES**
  Advanced Research Computing – Technology Services (ARC TS) provides access to and support for the use of advanced computing resources.

- **CONSULTING SERVICES**
  Consulting for Statistics, Computing and Analytics Research (CSCAR) provides support and training relating to the management, collection, and analysis of data.
Michigan Data Science Initiative

Michigan Institute for Data Science (MIDAS)
- 202 U-M Core/Affiliate Faculty
- Cross-cutting Data Science Methodologies & Analytics
- Education & Training
- Industry Engagement
- 4 Challenge Thrusts
- 30 existing U-M faculty slots
- 12 new core faculty slots

Data Science Services (CSCAR)
*Consulting for*
- Database Creation, Preparation & Ingestion
- Data Visualization
- Data Access
- Data Analytics

Data Science Infrastructure (ARC-TS)
- Hadoop, SPARK
- SQL, NoSQL databases
- Analytics Platforms
- Integration with HPC Flux Platform
MIDAS affiliate faculty

- MIDAS Faculty Affiliates

200+ Faculty Affiliates (3 campuses)

- Transportation
- Bio/clinical Informatics
- Machine Learning
- Social Media
- Learning Analytics
- Math Foundations
- Natural Language
- Visual Analytics
- Business Analytics
- Data enabled robotics

Alfred Hero, Univ. Michigan
MIDAS research challenge initiative programs

<table>
<thead>
<tr>
<th>Learning Analytics</th>
<th>Transportation</th>
<th>Social Sciences</th>
<th>Health Sciences</th>
<th>Future Challenge Thrusts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytics and Visualization of Complex Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine Learning-enabled Analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal, Multi-Scale and Statistical Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration of Heterogeneous Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scrubbing, Wrangling and Provenance Tracking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Privacy and Cybersecurity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MIDAS Funded Research: Transportation

Building a Transportation Data Ecosystem: creating a system for data on driver behavior, traffic, weather, accidents, vehicle messages, traffic signals and road characteristics, with a parallel and distributed computing platform.

Progress: The team has set up a baseline computing system for computer vision algorithms on integrated driving and sensor data. The team is improving algorithms, developing analyses to produce nationally representative results, and developing comprehensive statistical approach to identifying theme-based epochs in the data.

Flannagan (PI), UMITRI; Elliott, ISR; Hampshire, UMITRI; Jagadish, CoE
Jin, CoE; Mars, CoE; Murphey, UM-Dearborn; Nair, LS&A and CoE
Rupp, UMITRI; Shedden, LS&A; Tang, CoE; Witkowski, ISR

Reinventing Public Urban Transportation and Mobility: using predictive models for travel demand, accessibility, driver behavior, and transportation networks to design an on-demand public transportation system for urban areas.

Van Hentenryck (PI), CoE; Budak, SI; Cohn, CoE; Cunningham, Med.Sch and SPH
Dillahun, SI; Hampshire, UMITRI; Levine, Taubman College
Merlin, Taubman Coll.; Ortiz, UM-Dearborn; Wellman, COE

Progress: The RITMO project has developed and simulated an on-demand, multimodal transit system for Ann Arbor and is ready to deploy it. It improves convenience, cost, and accessibility.
MIDAS Funded Research: Learning analytics

**LEAP: analytics for LEarners As People:** creating learning analytics tools to directly link academic success and mental health with personal attributes such as values, beliefs, interests, behaviors, background, and emotional state.

Mihalcea (PI), CoE; Baveja, CoE; Eisenberg, SPH & ISR; Karabenick, SE & EMUI; Provost, CoE; Samson, SI; Collins-Thompson, SI; McKay, LS&A; Samson, SL; Shedden, LS&A

**Progress:** collecting data from 100 students and will start piloting a data collection with StudentLife in the fall. Methods developed to: (1) infer values, behavior, and sentiment from social media; (2) make cross-group comparisons using textual datasets; (3) extract linguistic features from classroom forums for predicting academic performance.

**Holistic Modelling of Education (HOME):** developing a holistic learning model, using cutting-edge data science methods, to examine the relationship of learner behavior, learning strategies, learner interaction with the learning environment, and academic achievements measured in multiple ways.

Teasley (PI), SI; Brooks, SI; Collins-Thompson, SI; Evrard, LS&A; Gere. LS&A; McKay, LS&A; Samson, SI

**Progress:** Data virtualization infrastructure for merging datasets across disparate sources. A funded NSF project utilizes what HOME is teaching us about how to form a more holistic model of the student.
MIDAS Funded Research: Social Science

Computational Approaches for the Construction of Novel Macroeconomic Data: creating a versatile and user-friendly system that processes and analyzes massive social media data for research in macroeconomics.

Shapiro (PI), LS&A and ISR; Cafarella, CoE; Deng, CoE; Levenstein, ISR
MIDAS Funded Research: Social Science

A Social Science Collaboration for Research on Communication and Learning based upon Big Data: developing methods to integrate geospatial, social media and survey data and examine patterns of communication that influence political choices and health awareness.

Traugott (PI), ISR; Budak, SI; Mneimneh, ISR; Singh, Georgetown; Ragunathan, SPH & ISR; Davis-Keane, LS&A and ISR; Pasek, LS&A; Bode, Georgetown; Ladd, Georgetown; Ryan, Georgetown; Soroka, LS&A
MIDAS Funded Research: Health Science

**Michigan Center for Single-Cell Genomic Data Analytics**: developing state-of-the-art approaches to analyze single-cell sequencing data.

- Li (co-PI), Med. Sch.
- Gagnon-Bartsch, LS&A
- Scott, CoE
- Gilbert (co-PI), LS&A
- Balzano, CoE
- Guan, Med. Sch.
- Vershynin, LS&A
- Hammoud, Med. Sch.
- Omenn, Med. Sch.
- Colacino, SPH
- Wicha, Med. Sch.

3D density map of 13,000 germ cells, as distinguished in their gene expression PC1-PC2 space.
MIDAS Funded Research: Health Science

Michigan Center for Health Analytics and Medical Prediction (M-CHAMP): developing innovative data science methods to extract features and patterns in complex time varying patient data

Nallamothu (PI), Med.Sch.; Harris, SON; McCullough, SPH; Shedden, LS&A; Vydiswaran, Med. Sch. & SI; Waljee, Med. Sch.
Iwashyna, Med. Sch.; Prescott, Med. Sch.; Sjoding, Med. Sch.; Wiens, CoE;
Kellenberg, Med. Sch.; Ryan, SPH; Sussman, Med. Sch.; Zhu, LS&A
MIDAS Funded Research: Health Science

Identifying Real-Time Data Predictors of Stress and Depression Using Mobile Technology: predicting the onset of depression by using mobile and wearable data from medical interns about their physiology, behavior and environment.

MIDAS Seminar series

- Seminars are on Fridays at 4pm
- A mix of internal and external seminars
- Open to the public and required of all Graduate Certificate students

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Attendees</th>
<th>In Person/ (Webcast)</th>
</tr>
</thead>
<tbody>
<tr>
<td>09/09/2016</td>
<td>Geoff Ginsburg (Duke)</td>
<td>24</td>
<td>(na)</td>
</tr>
<tr>
<td>09/23/2016</td>
<td>Rebecca Willett (Wisconsin)</td>
<td>37</td>
<td>(na)</td>
</tr>
<tr>
<td>09/30/2016</td>
<td>Jake Abernethy (UM)</td>
<td>24</td>
<td>(na)</td>
</tr>
<tr>
<td>10/07/2016</td>
<td>Gary King (Harvard)</td>
<td>129</td>
<td>(74)</td>
</tr>
<tr>
<td>11/04/2016</td>
<td>Yuejie Chi (OSU)</td>
<td>26</td>
<td>(na)</td>
</tr>
<tr>
<td>11/09/2016</td>
<td>Tamara Kolda (Sandia Labs)</td>
<td>61</td>
<td>(na)</td>
</tr>
<tr>
<td>12/16/2016</td>
<td>Bing Liu (UI-Chicago)</td>
<td>59</td>
<td>(na)</td>
</tr>
<tr>
<td>01/06/2017</td>
<td>Lav Varshney (UIUC)</td>
<td>40</td>
<td>(na)</td>
</tr>
<tr>
<td>01/13/2017</td>
<td>Dimitris Papanikolaou (Harvard)</td>
<td>36</td>
<td>(19)</td>
</tr>
<tr>
<td>01/27/2017</td>
<td>Emily Mower Provost (UM)</td>
<td>26</td>
<td>(17)</td>
</tr>
<tr>
<td>02/03/2017</td>
<td>Yao Xie (GA Tech)</td>
<td>39</td>
<td>(na)</td>
</tr>
<tr>
<td>02/17/2017</td>
<td>Carol Flannagan (UM)</td>
<td>17</td>
<td>(17)</td>
</tr>
<tr>
<td>02/24/2017</td>
<td>Jose Perea (MSU)</td>
<td>32</td>
<td>(16)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker</th>
<th>Attendees</th>
<th>In Person/ (Webcast)</th>
</tr>
</thead>
<tbody>
<tr>
<td>03/09/2017</td>
<td>David Blei (Columbia)</td>
<td>55</td>
<td>(19)</td>
</tr>
<tr>
<td>03/10/2017</td>
<td>Robert Nowak (Wisconsin)</td>
<td>39</td>
<td>(29)</td>
</tr>
<tr>
<td>03/17/2017</td>
<td>Laura Balzano (UM)</td>
<td>30</td>
<td>(19)</td>
</tr>
<tr>
<td>03/24/2017</td>
<td>Tianxie Cai (Harvard)</td>
<td>35</td>
<td>(18)</td>
</tr>
<tr>
<td>04/07/2017</td>
<td>Michael Cavaretta (Ford)</td>
<td>28</td>
<td>(14)</td>
</tr>
<tr>
<td>04/21/2017</td>
<td>Dania Koutra (UM)</td>
<td>32</td>
<td>(na)</td>
</tr>
</tbody>
</table>

Fall 2017 seminar schedule is under construction
Artificial Intelligence Goes All-In: Computers Playing Poker
Thurs, July 15
3:30 pm to 5 pm, Rm 1690 BBB

Prof. Michael Bowling
Department of Computing Science
University of Alberta

Artificial intelligence has seen several breakthroughs in recent years, with games such as checkers, chess, and go often serving as milestones of progress. Poker is another game entirely, with players having their own asymmetric information about what’s happening in the game. In this talk, I’ll describe a decade long research program to build AI that can cope with the hallmarks of poker -- deception, bluffing, and manipulating what other players know. This research has culminated in two landmark results: Cepheus playing a perfect game of limit poker, and most recently DeepStack beating poker pros at the game of no-limit poker. These two computer programs take very different approaches, and shed light on what is required to play a game at an expert-level and what is required to play it perfectly.
Outline

1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks
Data science educational programs

• Undergraduate programs:
  – Data Science BS degree program (2015 - MIDAS advisory)
  – Data Science BA degree program (2015 - MIDAS advisory)

• Graduate programs
  – Data Science Certificate program (2015 - MIDAS run)
  – Data Science Masters degree (2018 - MIDAS advisory)

• Post-graduate activities
  – MIDAS massive online open courseware (Moocs)
  – MIDAS distinguished seminar series (Northrop Grumman)
  – MIDAS workshops (Shannon, MBDH)

• Educational outreach activities
  – Big Data Summer School (MIDAS supported)
  – Michigan Data Science Team (MIDAS run)
  – Data science high school summer camp (MIDAS run)
Data science education at UM

- Two Data Science programs at University of Michigan

Undergraduate Program in Data Science

UG program is joint between EECS and Statistics and provides
- Rigorous foundation in CS, Stats, and Math
- Practical use of DS methods&algorithms
Capstone course is required for DS-Eng

A 9 credit G program certifying training in
- (Modeling) Understanding of core Data Science principles, assumptions & applications;
- (Technology) management, computation, information extraction & analytics;
- (Practice) Hands-on experience with modeling tools and technology using real data
Open to all graduate students on campus

NB: An MS/MA in DS is in planning stages
Graduate Certificate Requirements

1. Course Requirements
   a) 9 **graduate credits** in 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

2. Attendance at the MIDAS Annual Graduate Research Symposium

3. Regular attendance at the MIDAS Seminar Series (1 year)

http://midas.umich.edu/certificate
# Grad Certificate Areas and Competencies

<table>
<thead>
<tr>
<th>Areas &amp; Applications</th>
<th>Competency</th>
<th>Expectation</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithms &amp; Applications</strong></td>
<td><strong>Tools</strong></td>
<td>Working knowledge of basic software tools (command-line, GUI based, or web-services)</td>
<td>Familiarity with statistical programming languages, e.g., R or SciKit/Python, and database querying languages, e.g., SQL or NoSQL</td>
</tr>
<tr>
<td></td>
<td><strong>Algorithms</strong></td>
<td>Knowledge of core principles of scientific computing, applications programming, API’s, algorithm complexity, and data structures</td>
<td>Best practices for scientific and application programming, efficient implementation of matrix linear algebra and graphics, elementary notions of computational complexity, user-friendly interfaces, string matching</td>
</tr>
<tr>
<td></td>
<td><strong>Application Domain</strong></td>
<td>Data analysis experience from at least one application area, either through coursework, internship, research project, etc.</td>
<td>Applied domain examples include: computational social sciences, health sciences, business and marketing, learning sciences, transportation sciences, engineering and physical sciences</td>
</tr>
<tr>
<td><strong>Data Management</strong></td>
<td><strong>Data validation &amp; Visualization</strong></td>
<td>Curation, Exploratory Data Analysis (EDA) and visualization</td>
<td>Data provenance, validation, visualization - histograms, QQ plots, scatterplots (ggplot, Dashboard, D3.js)</td>
</tr>
<tr>
<td></td>
<td><strong>Data Wrangling</strong></td>
<td>Skills for data normalization, data cleaning, data aggregation, and data harmonization/registration</td>
<td>Data imperfections include missing values, inconsistent string formatting ('2016-01-01' vs. '01/01/2016', PC/Mac/Linux time vs. timestamps, structured vs. unstructured data)</td>
</tr>
<tr>
<td></td>
<td><strong>Data Infrastructure</strong></td>
<td>Handling databases, web-services, Hadoop, multi-source data</td>
<td>Data structures, SOAP protocols, ontologies, XML, JSON, streaming</td>
</tr>
<tr>
<td><strong>Analysis Methods</strong></td>
<td><strong>Statistical Inference</strong></td>
<td>Basic understanding of bias and variance, principles of (non)parametric statistical inference, and (linear) modeling</td>
<td>Biological variability vs. technological noise, parametric (likelihood) vs non-parametric (rank order statistics) procedures, point vs. interval estimation, hypothesis testing, regression</td>
</tr>
<tr>
<td></td>
<td><strong>Study design &amp; diagnostics</strong></td>
<td>Design of experiments, power calculations and sample sizing, strength of evidence, p-values, FDR</td>
<td>Multistage testing, variance normalizing transforms, histogram equalization, goodness-of-fit tests, model overfitting, model reduction</td>
</tr>
<tr>
<td></td>
<td><strong>Machine Learning</strong></td>
<td>Dimensionality reduction, k-nearest neighbors, random forests, AdaBoost, kernelization, SVM, ensemble methods, CNN</td>
<td>Empirical risk minimization. Supervised, semi-supervised, and unsupervised learning. Transfer learning, active learning, reinforcement learning, multiview learning, instance learning</td>
</tr>
</tbody>
</table>
## Illustrative 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(Ross)</td>
<td>Personal health and political views</td>
<td>Biostat 521</td>
<td>Psych 614</td>
<td>HS 853</td>
</tr>
</tbody>
</table>
## Sampling of courses

<table>
<thead>
<tr>
<th>Course Number Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EECS 584: Advanced Database Management Systems</td>
<td>Masters/Ph.D. level course for students in Computer Science, Electrical Engineering, and Information School</td>
</tr>
<tr>
<td>EECS EECS453: Applied Data Analysis</td>
<td>Applied matrix algorithms for signal processing, data analysis and machine learning</td>
</tr>
<tr>
<td>EECS 545: Machine Learning</td>
<td>Foundations of machine learning, mathematical derivation and implementation of the algorithms, and their applications</td>
</tr>
<tr>
<td>Math 571, Numerical Linear Algebra</td>
<td>Numerical methods for solving linear algebra problems (linear systems and eigenvalue problems), matrix decompositions, and convex optimization</td>
</tr>
<tr>
<td>Stats 415: Data Mining and Statistical Learning</td>
<td>This course covers the principles of data mining, exploratory analysis and visualization of complex data sets, and predictive modeling. The presentation balances statistical concepts (such as over-fitting data, and interpreting results) and computational issues. Students are exposed to algorithms, computations, and hands-on data analysis in the weekly discussion sessions.</td>
</tr>
</tbody>
</table>
Graduate Certificate Program Advisors

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, Leadership and Effectiveness Science, Bioinformatics, SoN&M
Karthik Duraisamy: Atmospheric, Oceanic, and Space Sciences CoE
August Evrard: Physics; Astronomy, LS&A
Anna Gilbert: Mathematics, LS&A
Richard Gonzales, Psychology, LS&A
Alfred Hero: EECS; Biomedical Engineering, Statistics, CoE
H. V. Jagadish: Electrical Engineering and Computer Science, CoE
Judy Jin: Industrial & Operations Engineering, CoE
Carl Lagoze: School of Information, SI
Honglak Lee, Electrical Engineering and Computer Science, CoE
Qiaozhu Mei: School of Information, SI
Christopher Miller: Astronomy, LS&A
Stephen Smith: Ecology and Evolutionary Biology, LS&A
Jeremy Taylor, Biostatistics, SPH
Ambuj Tewari: Statistics; Computer Science and Engineering, LS&A
Massive Open Online Courses (MOOC)

The following UM MOOCS are available on Coursera or EdX

– Foundations
  • Python for Everybody Series
  • Survey Data Collection and Analytics Series
  • Discrete Optimization
  • Probabilistic Graphical Models

– Core Data Science
  • Practical Learning Analytics
  • Data Science Ethics
  • Introduction to Natural Language Processing

– Advanced Data Science and Predictive Analytics
  • Data Science and Predictive Analytics

– U-M launches 2 specializations for new generation of data scientists
  • Applied Data Science with Python
  • Data Collection and Analysis

[Links to MOOCS and related articles provided]
PyData provides a forum for the international community of users and developers of data analysis tools to share ideas and learn from each other. The global PyData network promotes discussion of best practices, new approaches, and emerging technologies for data management, processing, analytics, and visualization. PyData communities approach data science using many languages, including (but not limited to) Python, Julia, and R.

PyData Ann Arbor is a group for amateurs, academics, and professionals currently exploring various data ecosystems. Specifically, we seek to engage with others around analysis, visualization, and management. We are primarily focused on how Python data tools can be used in innovative ways but also maintain a healthy interest in leveraging tools based in other languages such as R, Java/Scala, Rust, and Julia.

Upcoming Events

- **June 8**, [Intro to Azure Machine Learning: Predict Who Survives the Titanic](www.meetup.com/PyData-Ann-Arbor)
- **July 13**, [Designing an Algorithmic Trading Strategy with Python](www.meetup.com/PyData-Ann-Arbor)
Computational Social Science Workshop

We are excited to announce two more python skills workshops in partnership with CSCAR! In order to attend, participants should register for them as soon as they become available on the CSCAR website. Registration is free to UM affiliated people.

Data Science with Social Science data: an introduction to Python’s Pandas
Thursday, March 30th, 2-4 pm, MLB 2001A
Register
This workshop introduces participants to Python’s NumPy, Pandas DataFrames, Matplotlib and StatsModels using an advertising dataset. Participants will use these tools to model (OLS) associations between advertising expenditures and product sales in example data. We will start with an introductory explanation of Anaconda and the Jupyter notebook environment (although not required for the participant, the instructor will be using these tools). We will proceed with topics including: reading data files; creation, indexing and slicing of Pandas DataFrames; creation and handling of Matplotlib objects; and creation and interpretation of models using Python’s StatsModels. Although not required, we recommend that participants have a basic knowledge of Python.

Data Science with Social Science data: building predictive model Python’s Scikit-learn
Thursday, April 6th, 2-4 pm, MLB 2001A
We will use Python’s Pandas DataFrames, Matplotlib and Scikit-learn data. Participants will use Scikit-learn tools to predict whether income particular dollar amount based on the census data. This workshop covers steps to building a predictive model in Python. We will start with an intro...

ORGANIZERS
Faculty Sponsor: Elizabeth Bruch
Graduate Student Coordinators
• Teddy Dewitt
• Jeff Lockhart
• Dylan Nelson

sites.lsa.umich.edu/css/
MIDAS Michigan Data Science Team

A student run organization with faculty oversight

Grassroots activity w/o academic credit.
Student-led tutorials + data hackathon project

Started in 2015 to facilitate student teaming for Kaggle prediction challenges
Transitioned to public service projects (2016)
- Flint Water Crisis
- Drunk Driving Forecasting
- Data-driven marketing

Sponsored by Nvidia and Google (2016)

Eric Schwartz (Mrkting) and Jake Abernethy (CSE)
MDST and Flint in the news

Google, U-M to build digital tools for Flint water crisis

How big data and algorithms are slashing the cost of fixing Flint’s water crisis

Google donates $250,000 to help bring information to Flint residents

Google, U-M to build digital tools to help manage Flint water crisis
A weeklong HS Summer Camp

A commuter camp open to all 9-12 graders.

Theme: Data science foundations in art and sports analytics

2016 camp held at UM in Ann Arbor
- 15 campers, 1 faculty, 2 graduate students, 2 staff

2017 camp to be at UM Detroit center
- Over 95 applications from 4 countries.
  28(15) accepted (Art/Sports)
Outline

1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks
Concluding remarks

• Data Science Discipline
  – Data science exists in an ecosystem of different disciplines
  – New data science applications are constantly being uncovered
  – Foundational principles for handling big data sets are evolving
  – Institutes and Centers cohere activities and build community

• Data Science Education:
  – Students cannot be expected to become universal experts
  – Statistical inference, computation, algorithms, and data management must be basic foundations of DS skills
  – Experience with empirical hands-on applications is a must
  – Interdisciplinary communication skills are especially important