MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

UNIVERSITY OF MICHIGAN

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

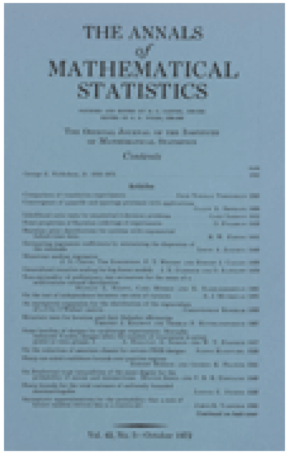
Data Science

- Origins in statistics: “50 years of Data Science,” David Donoho, Oct. 2015 Data Science Central THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS

Philosophical Magazine

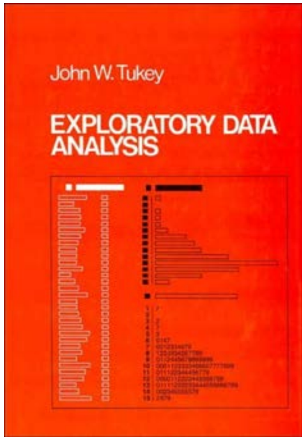
Series 6 (1901-1925)

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



THE ANNALS of MATHEMATICAL STATISTICS

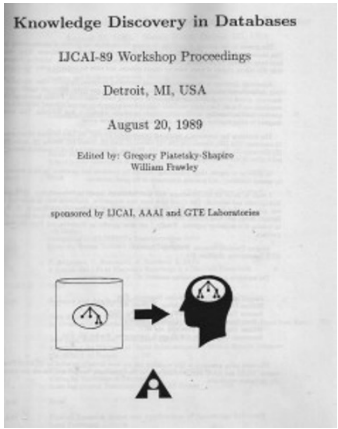
John Tukey (1962)
“Future of data analysis”



John W. Tukey

EXPLORATORY DATA ANALYSIS

John Tukey (1977)
EDA



Knowledge Discovery in Databases

IJCAI-89 Workshop Proceedings

Detroit, MI, USA

August 20, 1989

Edited by: Gregory Piatetsky-Shapiro
William Frawley

sponsored by IJCAI, AAAI and GTE Laboratories

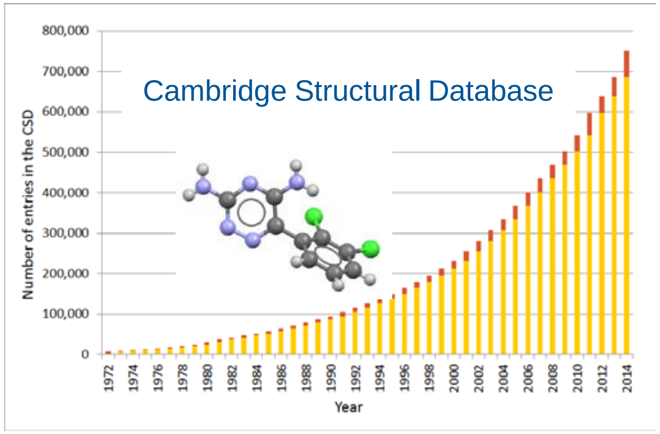
IAAI (1987)
KDD (Detroit)

Data Science

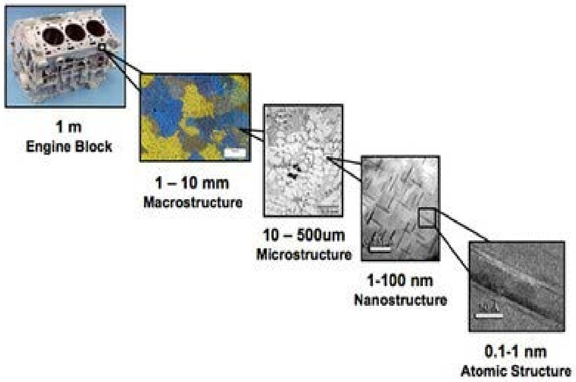
- Origins in statistics: “50 years of Data Science,” David Donoho, Oct. 2015 Data Science Central THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS
- Developing into widely embraced multi-disciplinary field
- Elements driving evolution of data science
  - Datasets are getting larger, faster with more complex structure
  - Data frequently is poorly annotated: provenance unknown
  - Privacy concerns: anonymization, fair use, reuse, ethics

Explosion in volume, velocity, variety of data

Materials Genome



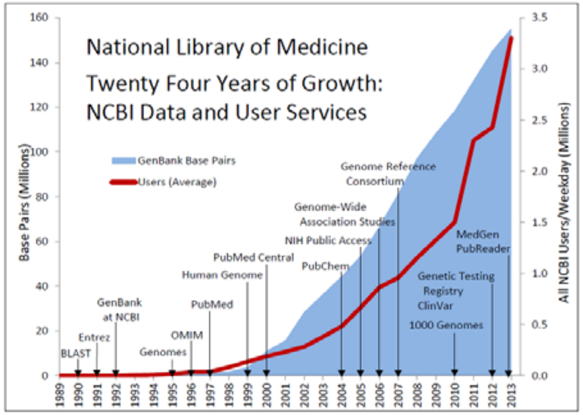
John Allison, Mat. Sci and Eng



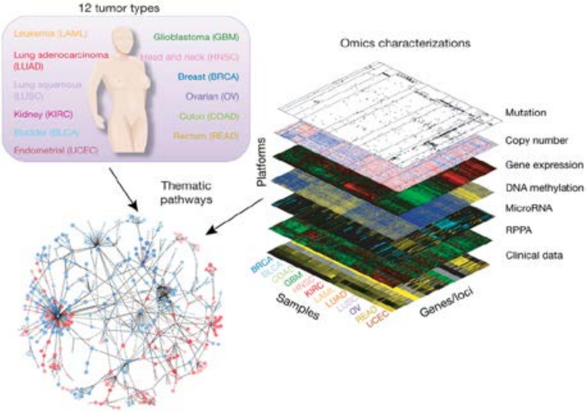
160,000 Engineering materials
Multiscale Multiphysics

CSE, ChemE, ECE, ME, MSE

Biomedicine



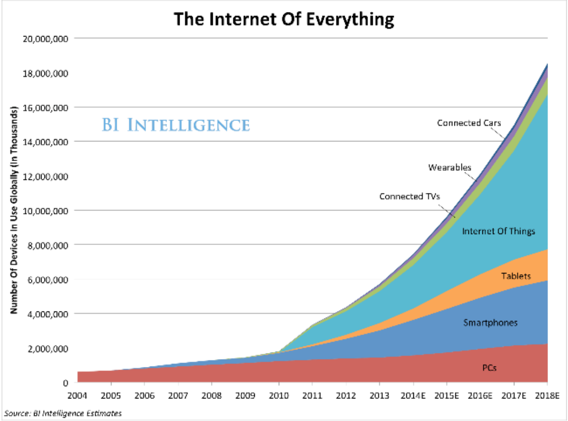
Nature Genetics 45, 1113–1120 (2013)



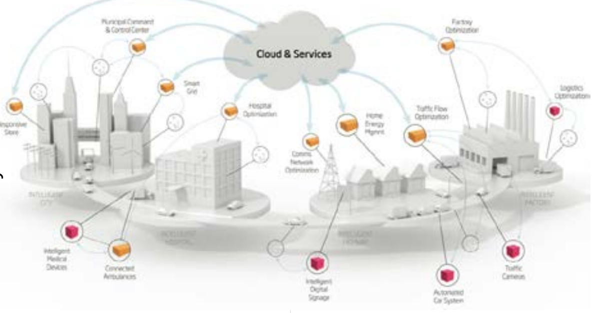
The Cancer Genome Atlas (TCGA)

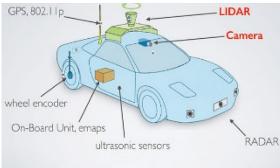
BME, CSE, ChemE, ECE, MED

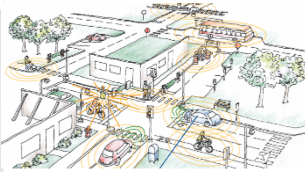
Cyberphysical Networks



dana-carvey-industrial-internet







UM Mobility Transformation Center (MTC)

AE, CSE, CivE, ECE, IOE, ME

Data science dimensions

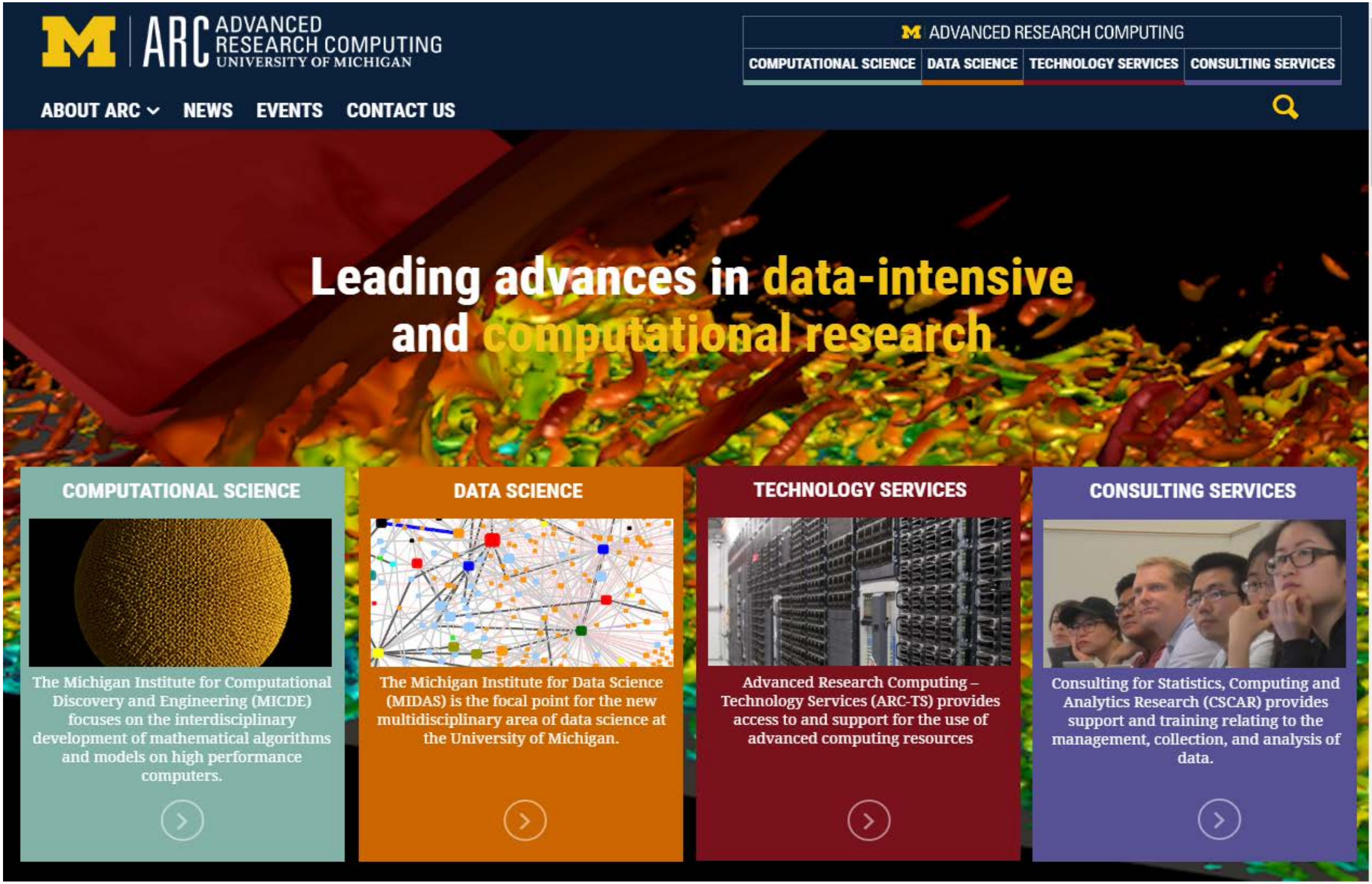
Data science is concerned with

- 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.

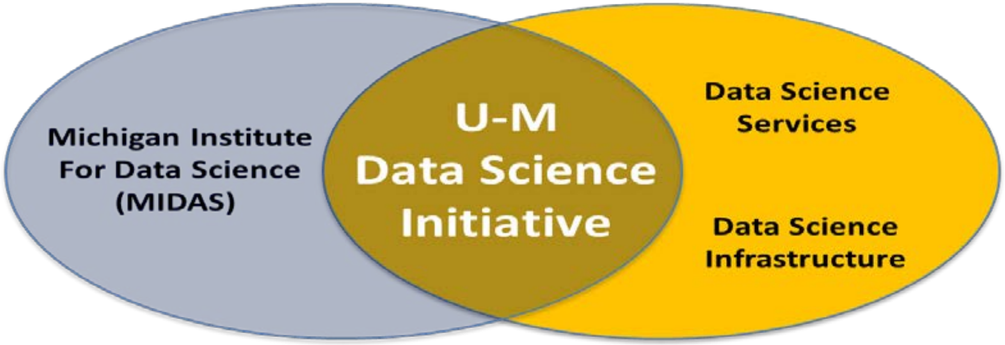
TECHNOLOGY SERVICES

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

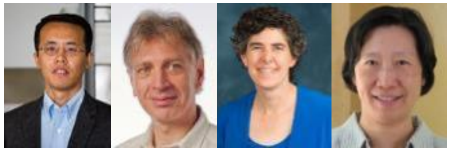
Data Science Infrastructure (ARC-TS)

- Hadoop, SPARK
- SQL, NoSQL databases
- Analytics Platforms
- Integration with HPC Flux Platform

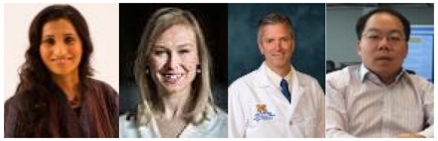
MIDAS affiliate faculty

- **MIDAS Faculty Affiliates**

200+ Faculty Affiliates (3 campuses)



Transportation



Bio/clinical Informatics

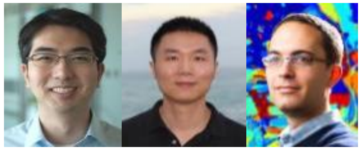
Machine Learning

Social Media

Learning Analytics

Math Foundations

Natural Language

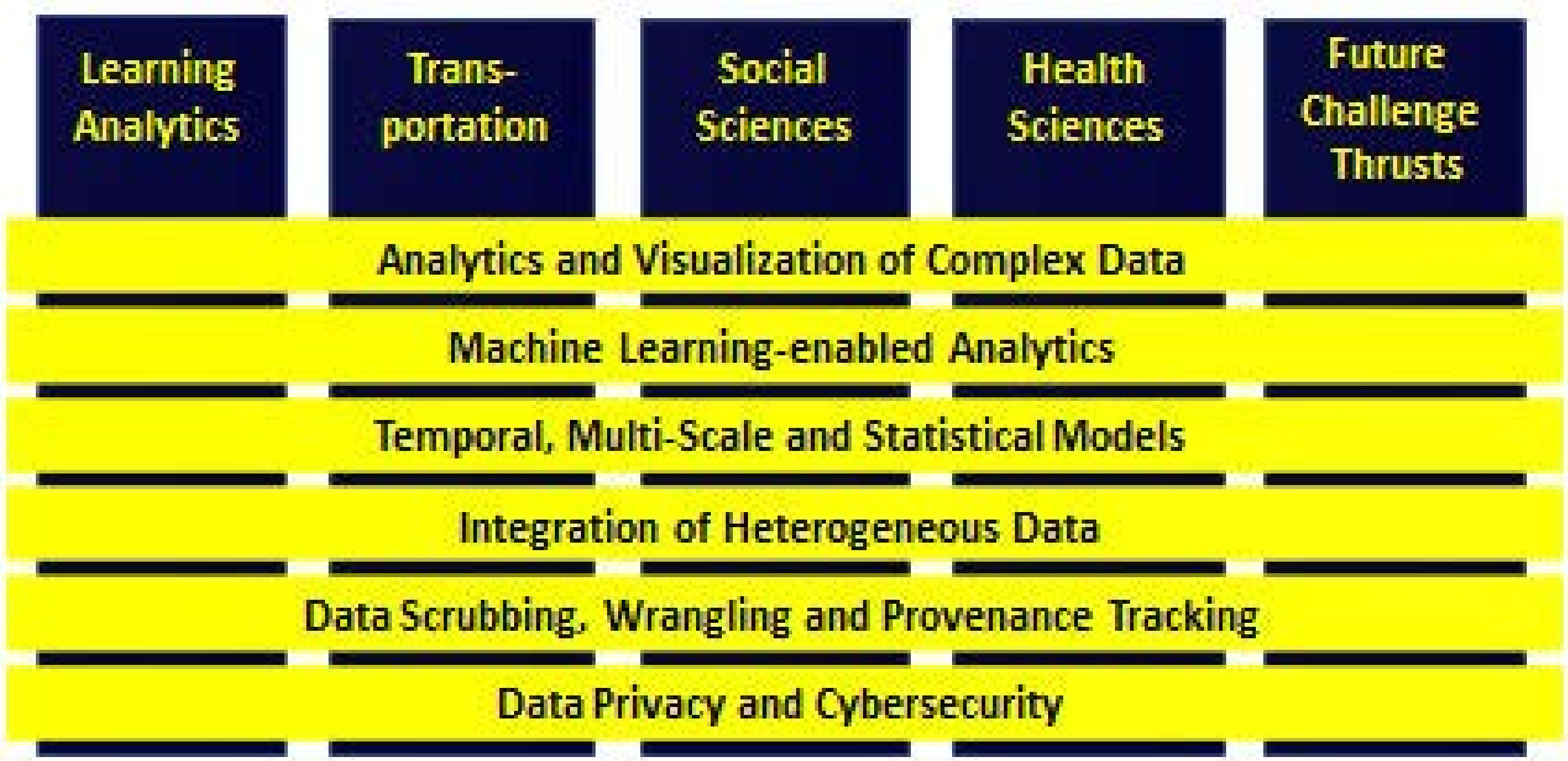


Visual Analytics

Business Analytics

Data enabled robotics

MIDAS research challenge initiative programs



Learning Analytics

Trans-portation

Social Sciences

Health Sciences

Future Challenge Thrusts

Analytics and Visualization of Complex Data

Machine Learning-enabled Analytics

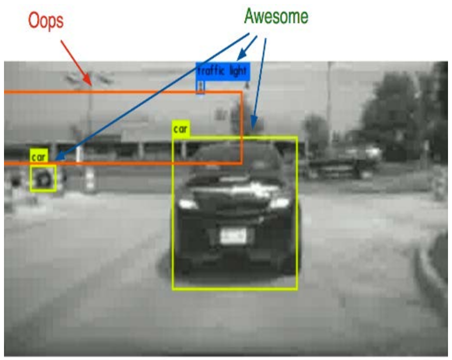
Temporal, Multi-Scale and Statistical Models

Integration of Heterogeneous Data

Data Scrubbing, Wrangling and Provenance Tracking

Data Privacy and Cybersecurity

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, UMTRI; Jagadish, CoE
Jin, CoE; Mars, CoE; Murphey, UM-Dearborn; Nair, LS&A and CoE
Rupp, UMTRI; Shedden, LS&A; Tang, CoE; Witkowski, ISR

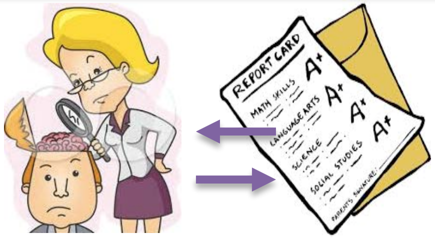
Watson
Downer
Lyneham
Dickson
O'Connor
Braddon
Ainslie
Turner

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, UMTRI; Lynch, CoE; Levine, Taubman College
Merlin, Taubman Coll.; Ortiz, UM-Dearborn; Sayer, UMTRI; 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



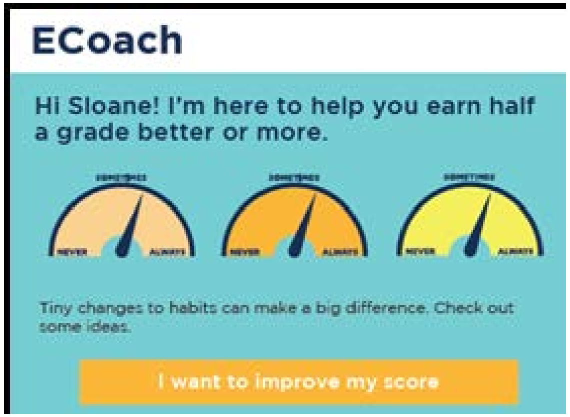
- Personality, Values
- Behaviors
- Interests
- Sentiment

- Grades
- Courses
- Major

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; Collins-Thompson, SI;
Eisenberg, SPH & ISR Karabenick, SE & EMUI; McKay, LS&A;
Provost, CoE; Samson, SI; 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.



ECoach

Hi Sloane! I'm here to help you earn half a grade better or more.

Tiny changes to habits can make a big difference. Check out some ideas.

I want to improve my score

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

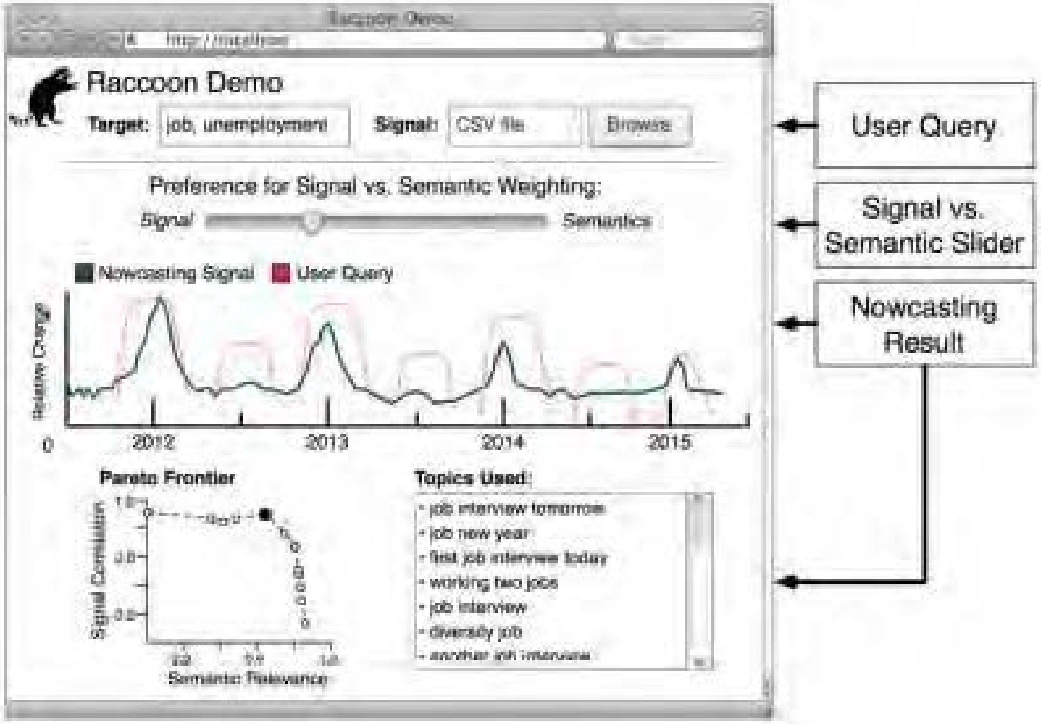
Progress: Data virtualization infrastructure for merging datasets across disparate sources. A funded NSF project utilizes what HOME is teaching us about how to form a more holistic model of the student.

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;
Deng, CoE;

Cafarella, CoE;
Levenstein, ISR



Raccoon Demo

Target: job, unemployment

Signal: CSV file

Preference for Signal vs. Semantic Weighting:

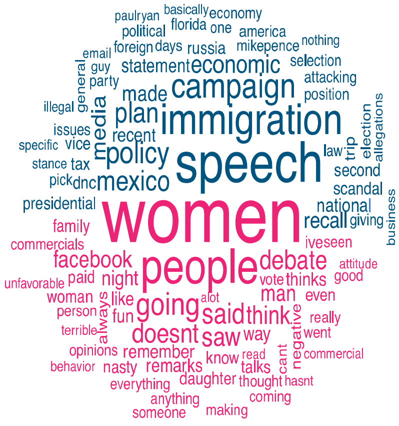
Topics Used:

- job interview tomorrow
- job new year
- first job interview today
- working two jobs
- job interview
- diversify job
- another job interview

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.

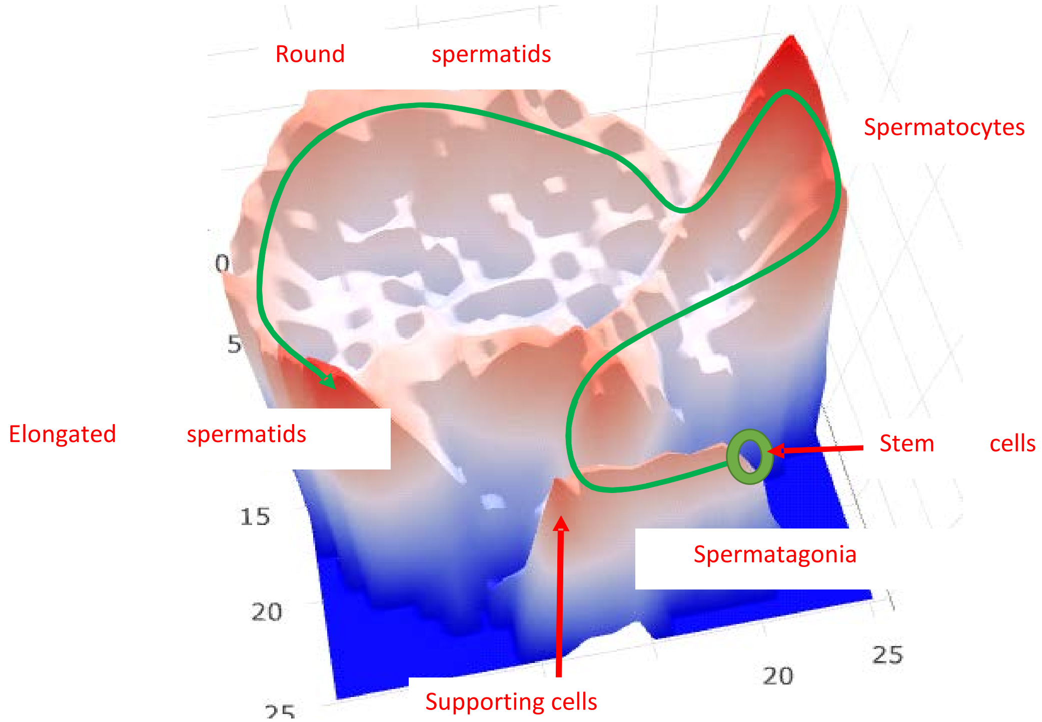
Traugott (PI), ISR;
Budak, SI;
Mneimneh, ISR;
Singh, Georgetown;
Ragunathan, SPH & ISR;
Davis-Keane, LS&A and ISR;
Pasek, LS&A;
Soroka, LS&A
Bode, Georgetown
Ladd, Georgetown;
Ryan, Georgetown;



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

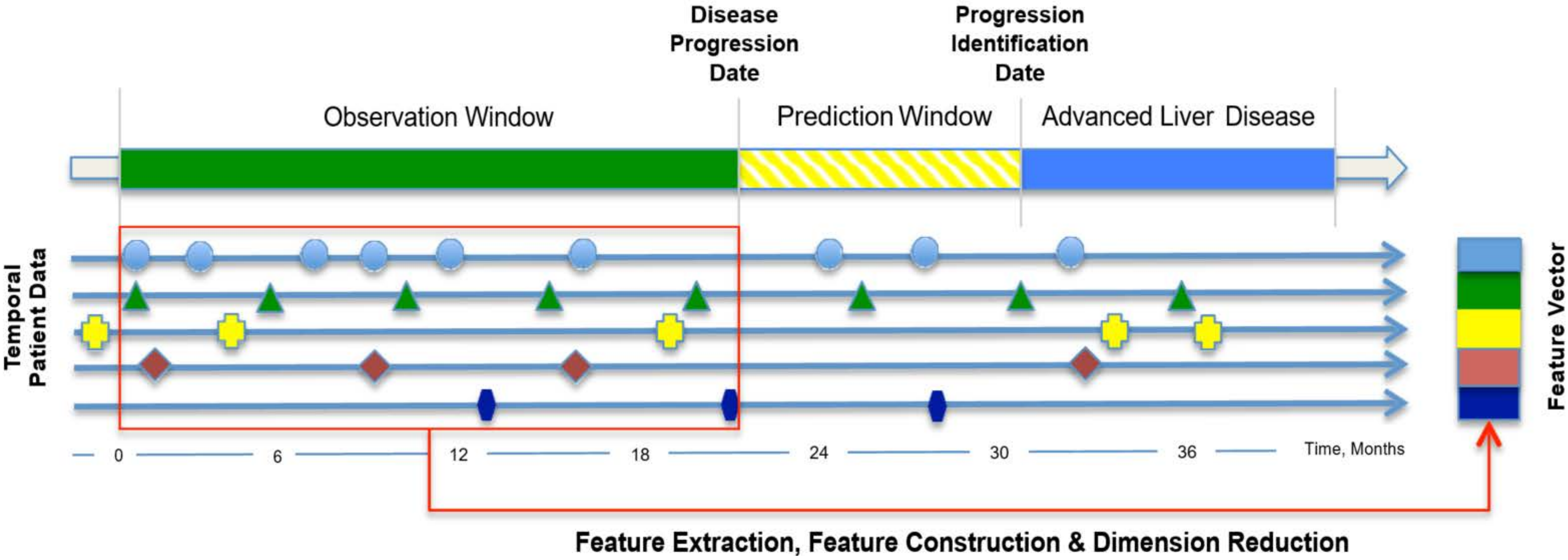


3D density map of 13,000 germ cells, as districted in their gene expression PC1-PC2 space.

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; Iwashyna, Med. Sch.; Kellenberg, Med. Sch.;
McCullough, SPH; Najarian, Med. Sch.; Prescott, Med. Sch.; Ryan, SPH;
Shedden, LS&A; Singh, Med. Sch.; Sjoding, Med. Sch.; Sussman, Med. Sch.;
Vydiswaran, Med. Sch. & SI; Waljee, Med. Sch. Wiens, CoE; 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.

Sen (PI), Med.Sch.; Burmeister, Med. Sch.; Cochran, LS&A.;
Forger, LS&A, and Med. Sch.; Murphy, LS&A, Med. Sch. And ISR; Wu, SPH

Welcome to the Intern Health Study

Thank you for choosing to participate in this study!

Participants should enter their access code to begin.

MD2K

Center of Excellence for Mobile Sensor Data-to-Knowledge

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

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

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

Fall 2017 seminar schedule is under construction

Upcoming Joint MIDAS and Toyota-AI Seminar

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

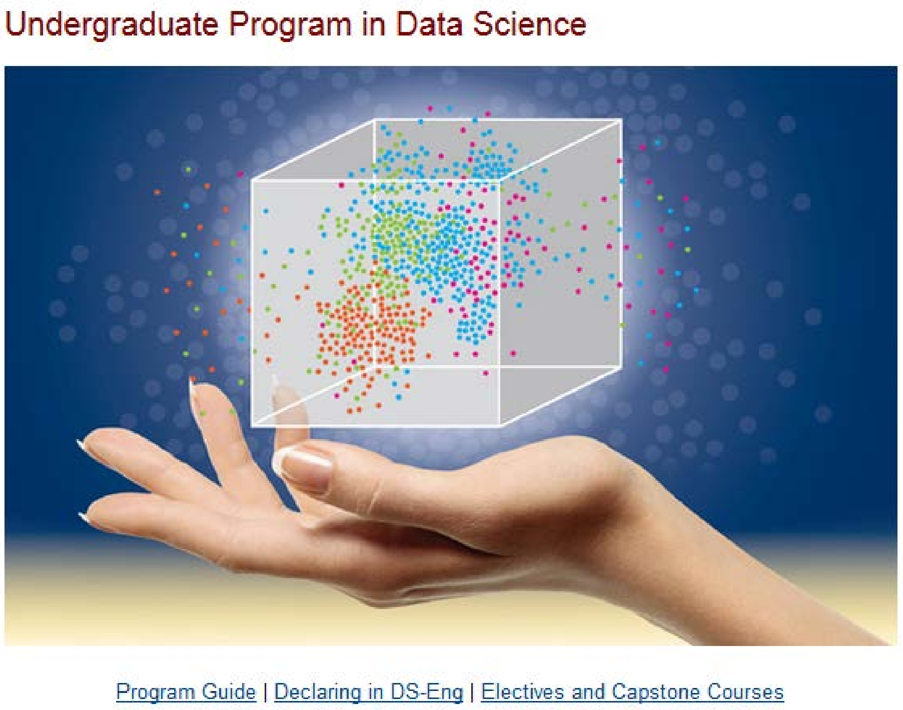
1. Emergence of data science
2. Michigan Institute for Data Science
3. Data science education and training at UM
4. Concluding remarks

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

Program Guide | Declaring in DS-Eng | Electives and Capstone Courses

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

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN
MENU

Graduate Data Science Certificate Program

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

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

Illustrative 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(Ross)	Personal health and political views	Biostat 521	Psych 614	HS 853

Sampling of courses

Course Number Title	Description
EECS 584: Advanced Database Management Systems	Masters/Ph.D. level course for students in Computer Science, Electrical Engineering, and Information School
EECS EECS453: Applied Data Analysis	Applied matrix algorithms for signal processing, data analysis and machine learning
EECS 545: Machine Learning	Foundations of machine learning, mathematical derivation and implementation of the algorithms, and their applications
Math 571, Numerical Linear Algebra	Numerical methods for solving linear algebra problems (linear systems and eigenvalue problems), matrix decompositions, and convex optimization
Stats 415: Data Mining and Statistical Learning	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.
Stats 503: Applied Multivariate Analysis	Applied multivariate analysis including Hotelling's T-squared, multivariate ANOVA, discriminant functions, factor analysis, principal components, canonical correlations, and cluster analysis. Selected topics from: Maximum likelihood and Bayesian methods, robust estimation, and survey sampling.

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

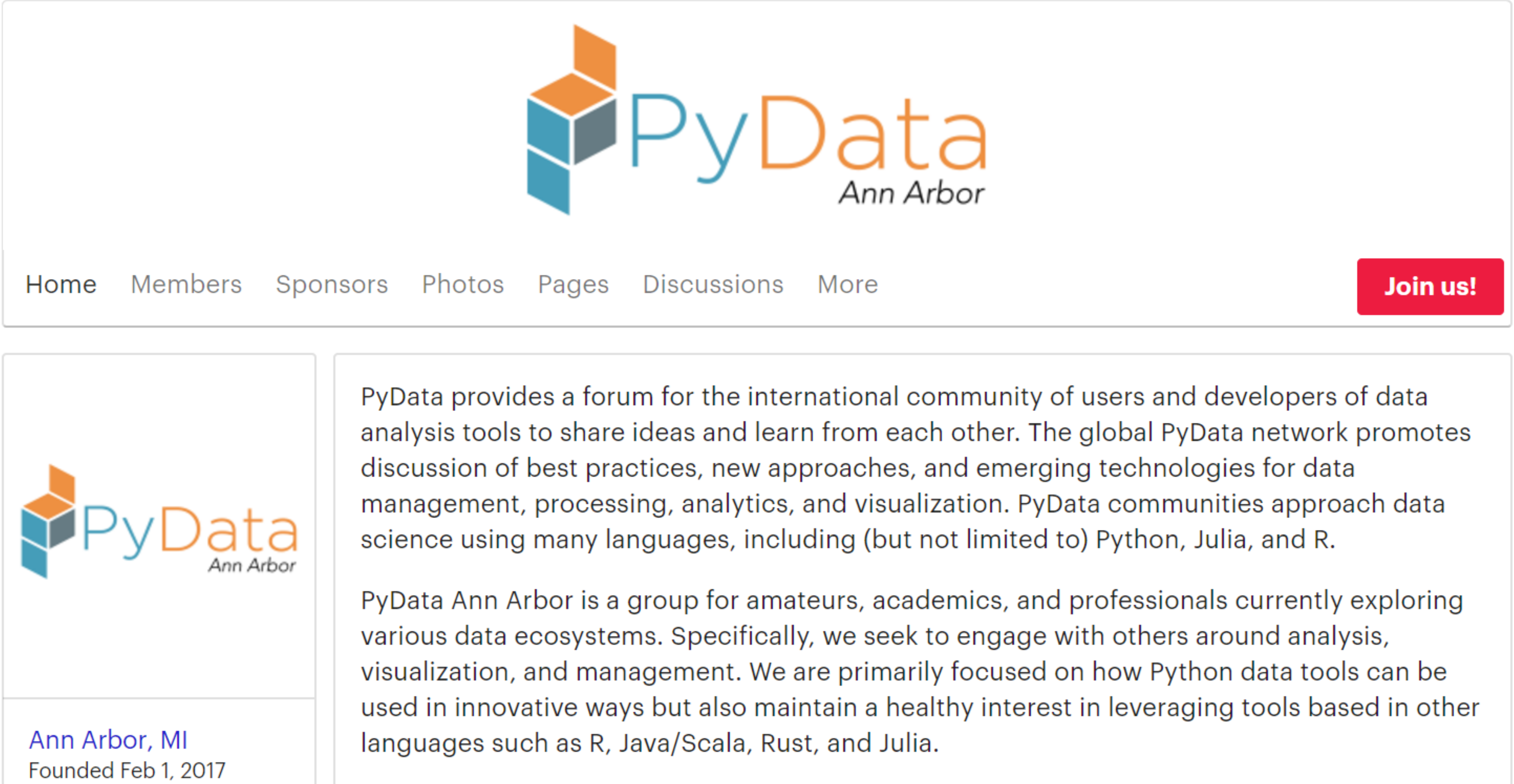
Massive Open Online Courses (MOOC)

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

midas.umich.edu/education/ds_moocs/

record.umich.edu/articles/u-m-launches-two-specializations-new-generation-data-scientists



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.

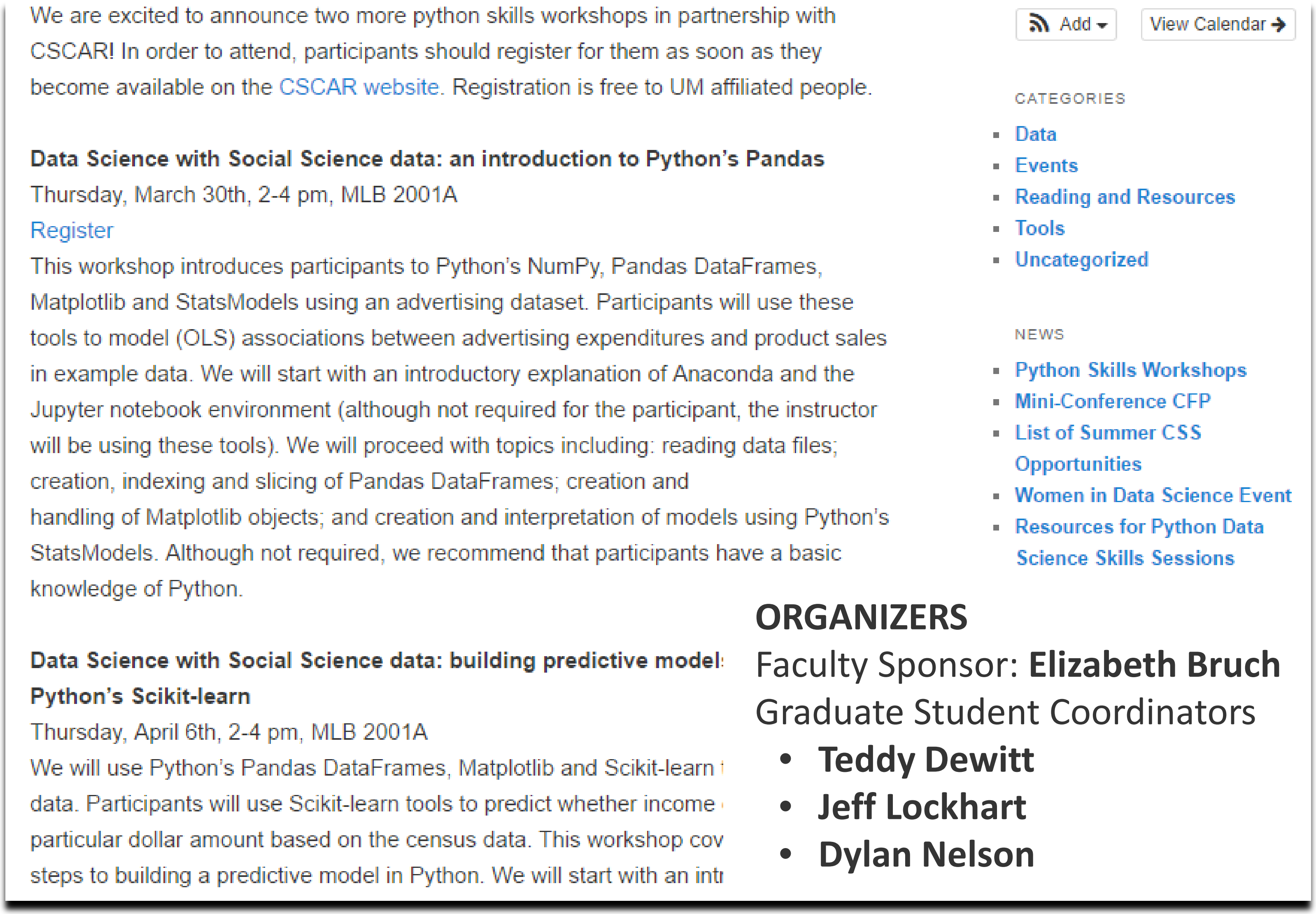
Ann Arbor, MI
Founded Feb 1, 2017

Upcoming Events

June 8, Intro to Azure Machine Learning: Predict Who Survives the Titanic
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 t data. Participants will use Scikit-learn tools to predict whether income particular dollar amount based on the census data. This workshop cov steps to building a predictive model in Python. We will start with an intr

CATEGORIES

- Data
- Events
- Reading and Resources
- Tools
- Uncategorized

NEWS

- Python Skills Workshops
- Mini-Conference CFP
- List of Summer CSS Opportunities
- Women in Data Science Event
- Resources for Python Data Science Skills Sessions

ORGANIZERS

Faculty Sponsor: **Elizabeth Bruch**

Graduate Student Coordinators

- **Teddy Dewitt**
- **Jeff Lockhart**
- **Dylan Nelson**

sites.lsa.umich.edu/css/

MIDAS Michigan Data Science Team

M | MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

A student run organization with faculty oversight

Eric Schwartz (Mrkting) and Jake Abernethy (CSE)

Grassroots activity w/o academic credit.

Student-led tutorials + data hackathon project

Started in 2015 to facilitate student teaming for Kaggle prediction challenges kaggle

Transitioned to public service projects (2016)

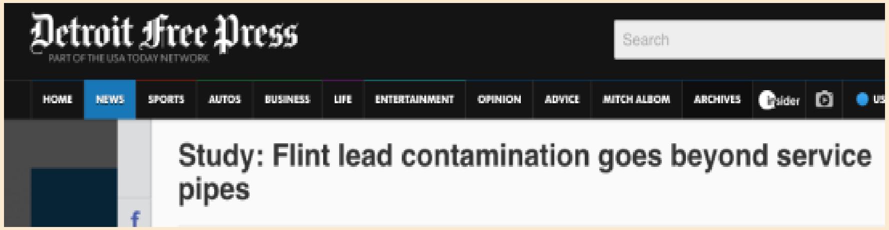
- Flint Water Crisis
- Drunk Driving Forecasting
- Data-driven marketing

Sponsored by Nvidia and Google (2016)

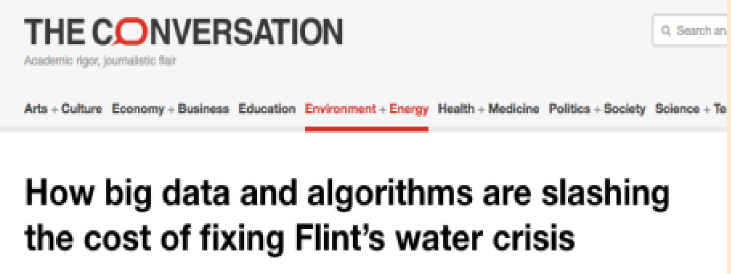
MIDAS Michigan Data Science Team

MDST and Flint in the news

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



Study: Flint lead contamination goes beyond service pipes

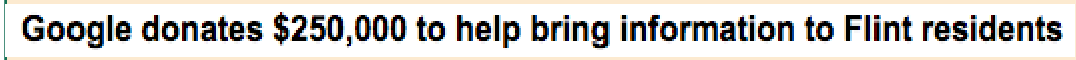


How big data and algorithms are slashing the cost of fixing Flint's water crisis

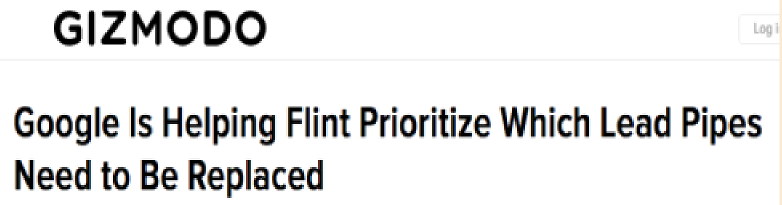
MDST Partners with UM-Flint & Google.org to Aid Locals in Flint Water Crisis

TALKING MACHINES

Machine Learning in the Flint Water Crisis

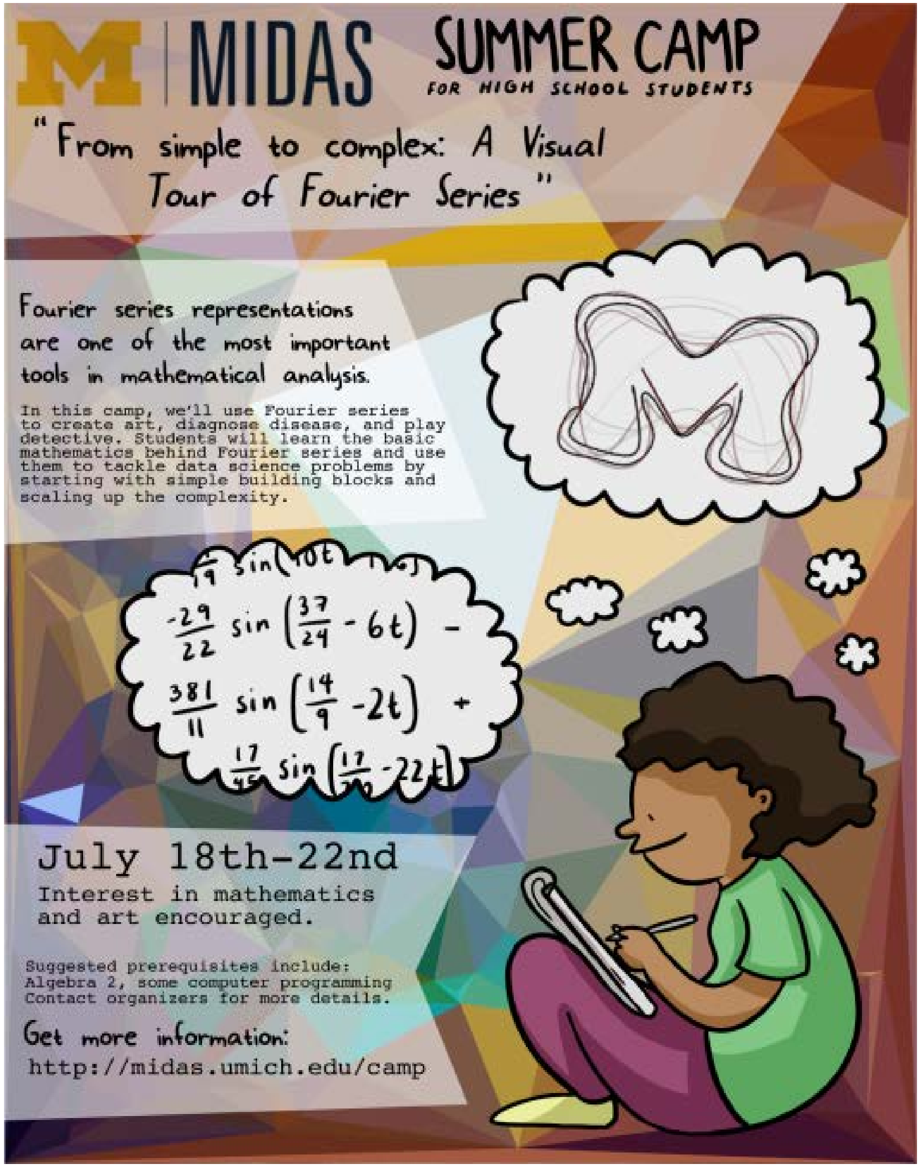

Google donates $250,000 to help bring information to Flint residents

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



Google Is Helping Flint Prioritize Which Lead Pipes Need to Be Replaced

MIDAS High School Summer Camp



M | MIDAS SUMMER CAMP

FOR HIGH SCHOOL STUDENTS

"From simple to complex: A Visual Tour of Fourier Series"

Fourier series representations are one of the most important tools in mathematical analysis.

In this camp, we'll use Fourier series to create art, diagnose disease, and play detective. Students will learn the basic mathematics behind Fourier series and use them to tackle data science problems by starting with simple building blocks and scaling up the complexity.

$\frac{-29}{22}\sin\left(\frac{37}{24}-6t\right)$ -
$\frac{381}{11}\sin\left(\frac{14}{9}-2t\right)$ +
$\frac{17}{45}\sin(\ldots-22t)$

July 18th-22nd

Interest in mathematics and art encouraged.

Suggested prerequisites include: Algebra 2, some computer programming Contact organizers for more details.

Get more information: http://midas.umich.edu/camp

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