MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

UNIVERSITY OF MICHIGAN

Data Science and Engineering

Roundtable on Data Science Education
Keck Center, National Academies
Washington, DC
Dec 14, 2016

Alfred Hero

Co-director, Michigan Institute for Data Science

Dept. of EECS, Dept. of BME, Dept. of Statistics

University of Michigan

midas.umich.edu

Outline

1. Changing landscape of data science
2. An engineering view of data science
3. Data science education
4. Closing thoughts

Outline

1. Changing landscape of data science
2. An engineering view of data science
3. Data science education
4. Closing thoughts

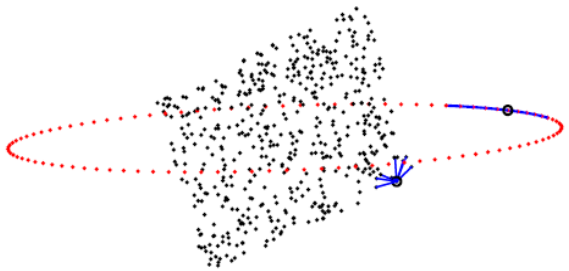
Academic disciplines engaging in data science

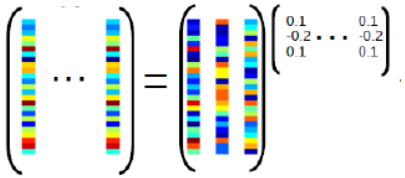
- Almost all disciplines are riding the wave of data using tools developed by data scientists:
  - Engineering, Natural Sciences, Social Science, Humanities, Music and Art, Urban Planning, Medicine, Nursing, Law, Business…
- There are several disciplines developing foundational data science principles:
  - Math, Computer Science, Statistics, Information Science, Physics, Engineering.

Multidisciplinary Landscape of Data Science

Mathematics
Data as a topological object

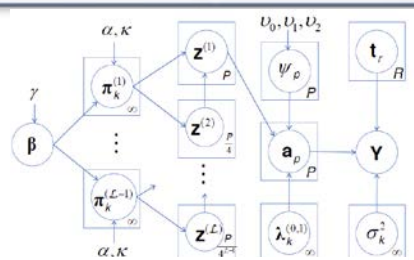
Applied topology
Harmonic analysis
Convex optimization
Num. linear algebra
Applied probability
Random matrix theory

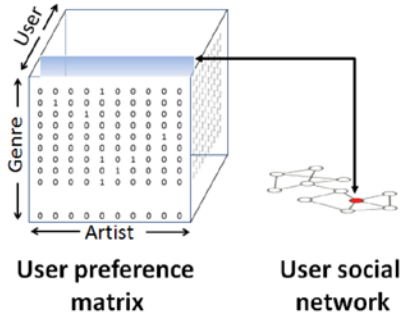




Computer Science
Data as a list/graph

Natural language proc.
Graphs and Networks
Algorithms
Database indexing
Machine learning
Privacy and security



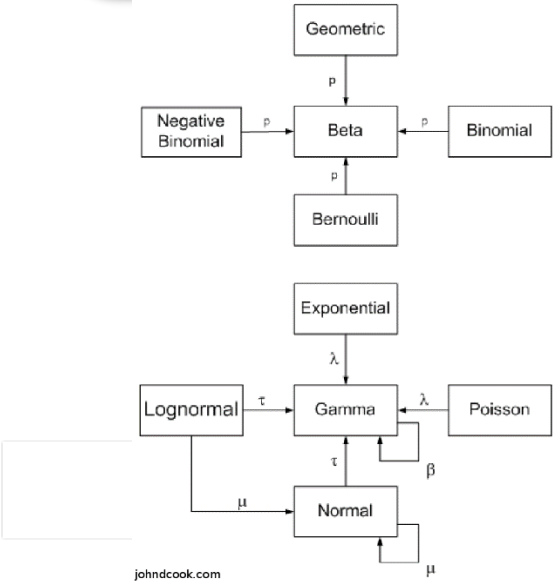


User preference matrix

User social network

Statistics
Data as a random sample

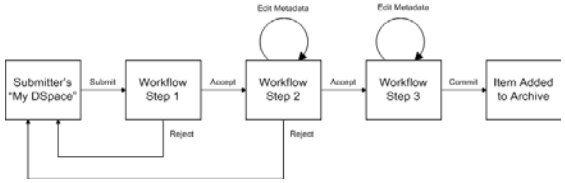
Sampling theory
Strength of evidence
Missing/anomalous data
Experimental design
Multivariate analysis
Graphical models



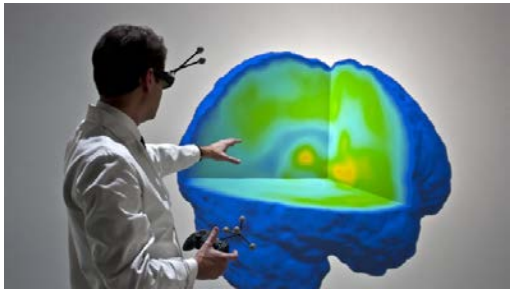
Multidisciplinary Landscape of Data Science

Information Science
Data at the interface

Human Computer Interaction (HCI)
Data sharing and reuse
Process and workflow
Data curation
Visualization



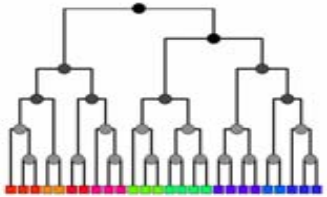
http://dspace.org/sites/dspace.org/

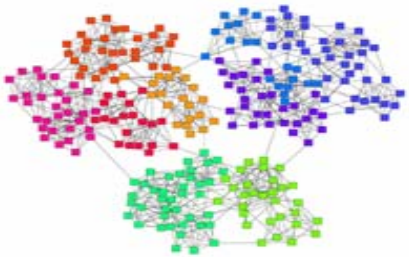


http://um3d.dc.umich.edu/visualization/

Physics
Data as natural phenomena

Network science
Complex systems
Statistical physics
Physico-mimetics
Phase transitions
Scaling&power laws

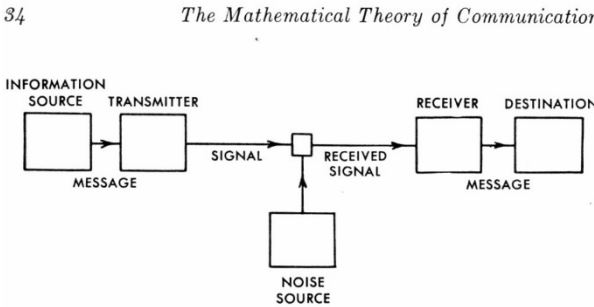


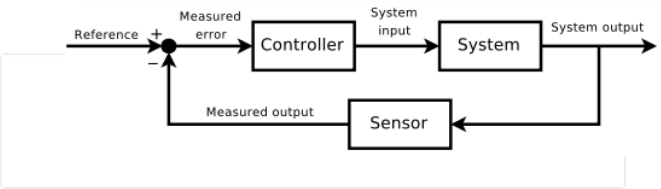


Mark Newman, *Networks* 2010

Engineering
Data-to-Decision

Comm. & info. theory
Signal processing
Sensing and control
Software engineering
Real-time HP computing
Cyberphysical systems

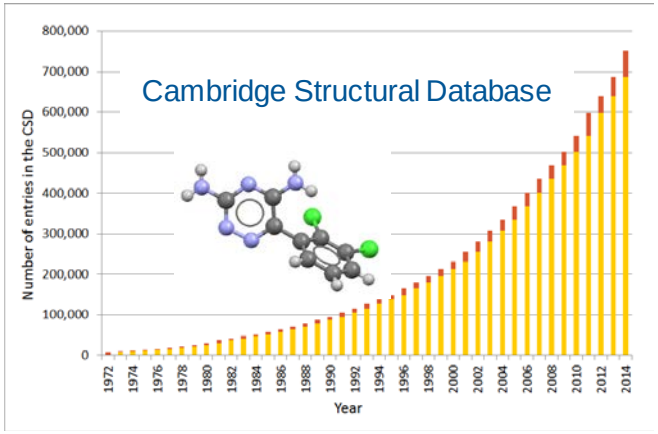


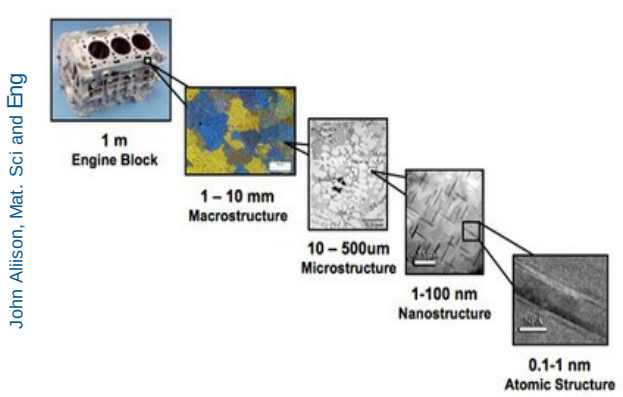


http://en.wikipedia.org/wiki/Control_theory

Examples of data-enabled engineering

Materials Genome



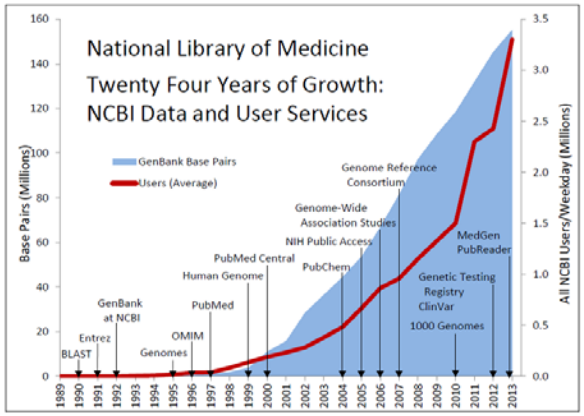


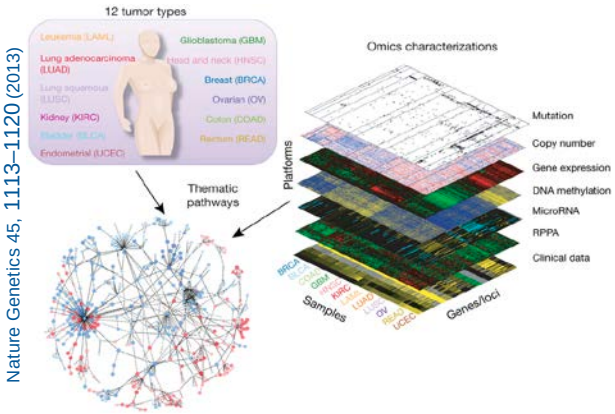
John Allison, Mat. Sci and Eng

160,000 Engineering materials
Multiscale Multiphysics

CSE, ChemE, ECE, ME, MSE

Precision Medicine



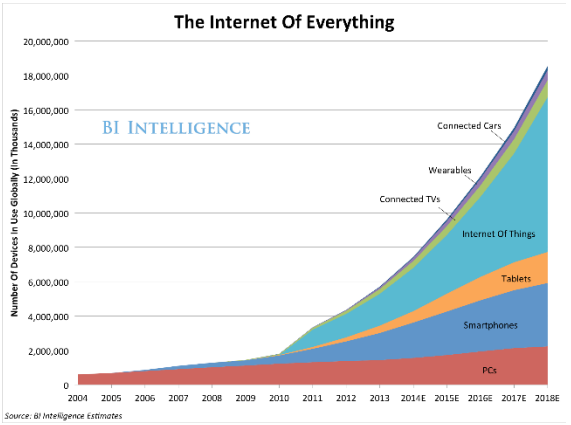


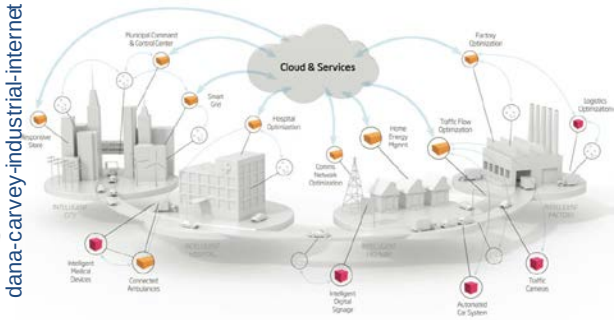
Nature Genetics 45, 1113–1120 (2013)

The Cancer Genome Atlas (TCGA)

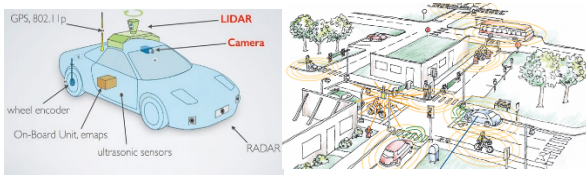
BME, CSE, ChemE, ECE, MED

Cyberphysical Networks





dana-carvey-industrial-internet

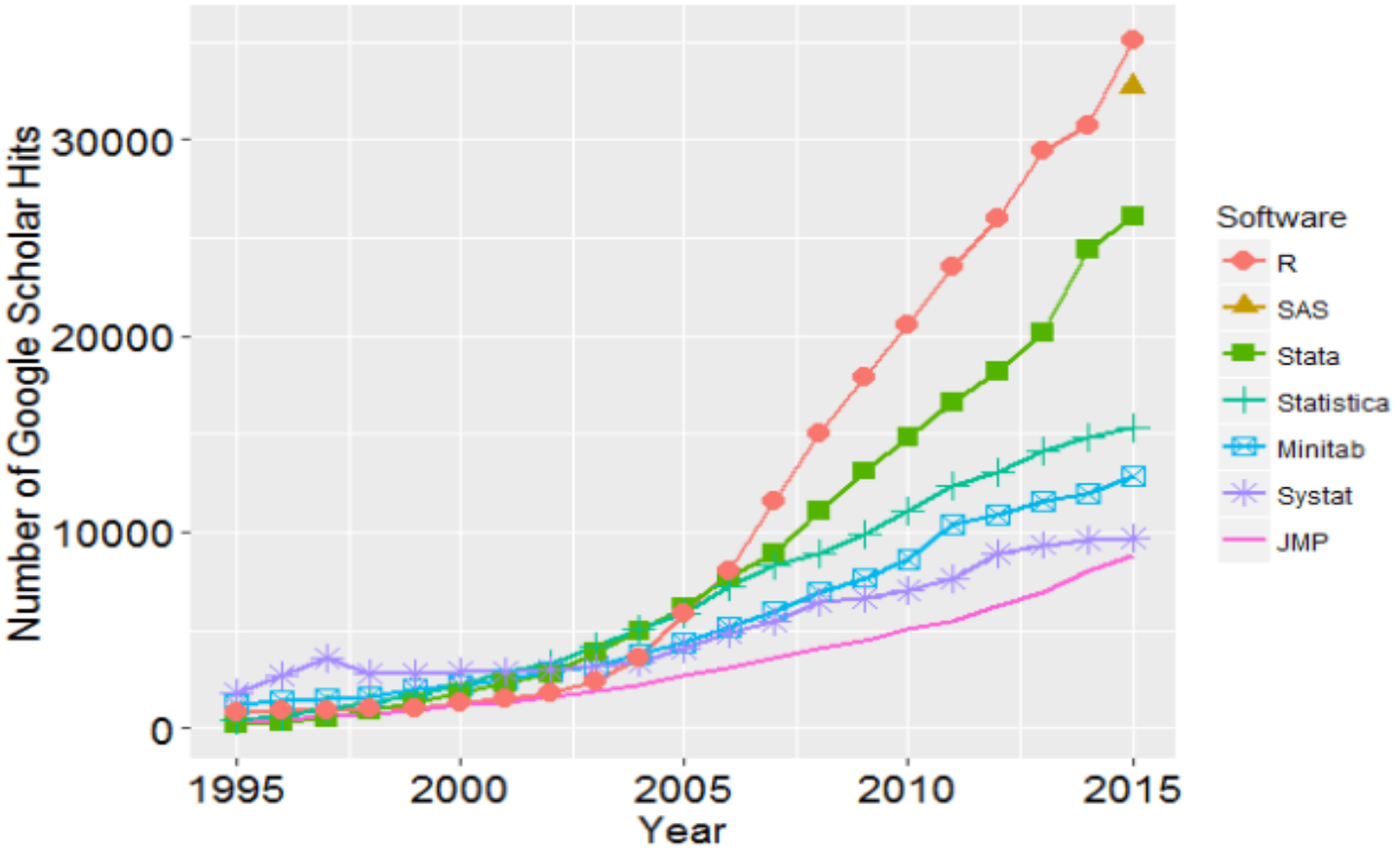


UM Mobility Transformation Center (MTC)

AE, CSE, CivE, ECE, IOE, ME

Data Analysis Software Usage

http://r4stats.com/articles/popularity/

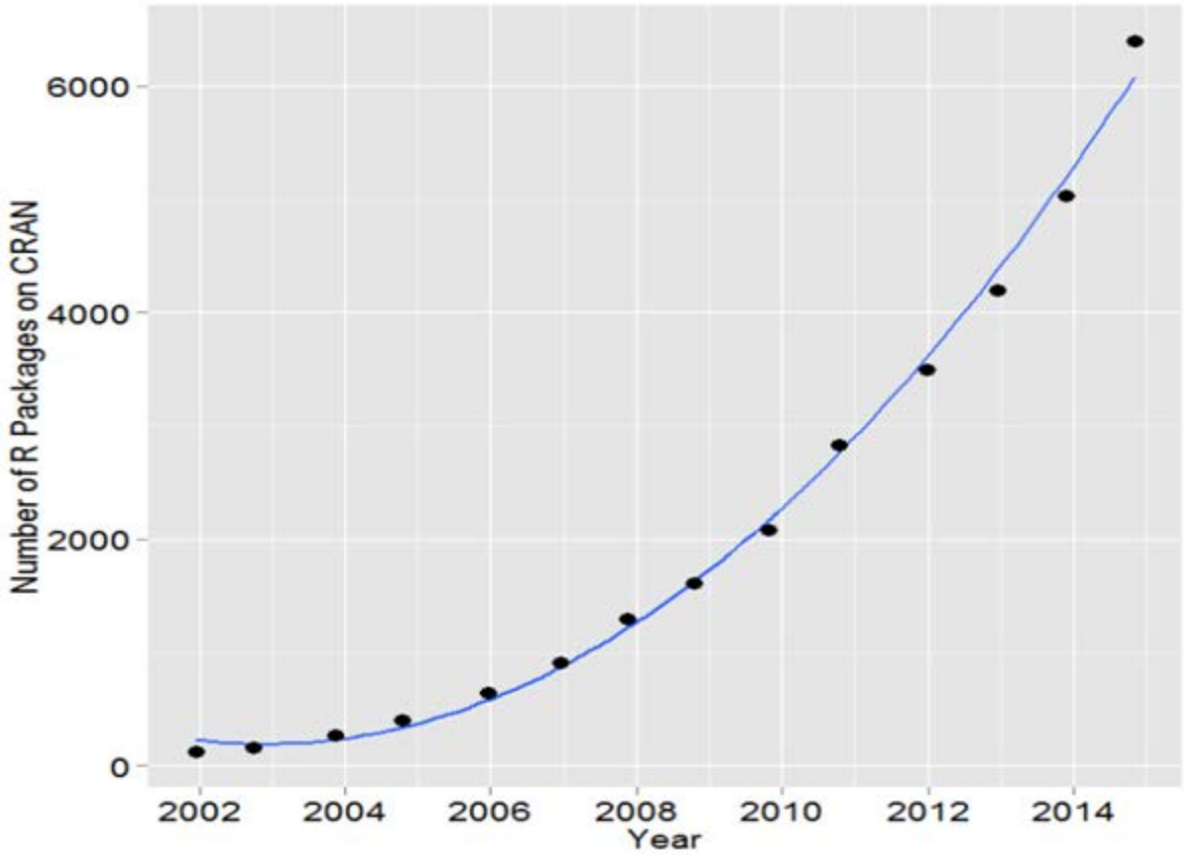


Explosion in # citations to analysis software

- Packages have better memory management and cloud support
- more data cleaning and diagonostic features
- more versatile data analysis and data visualization tools

Proliferation of software packages

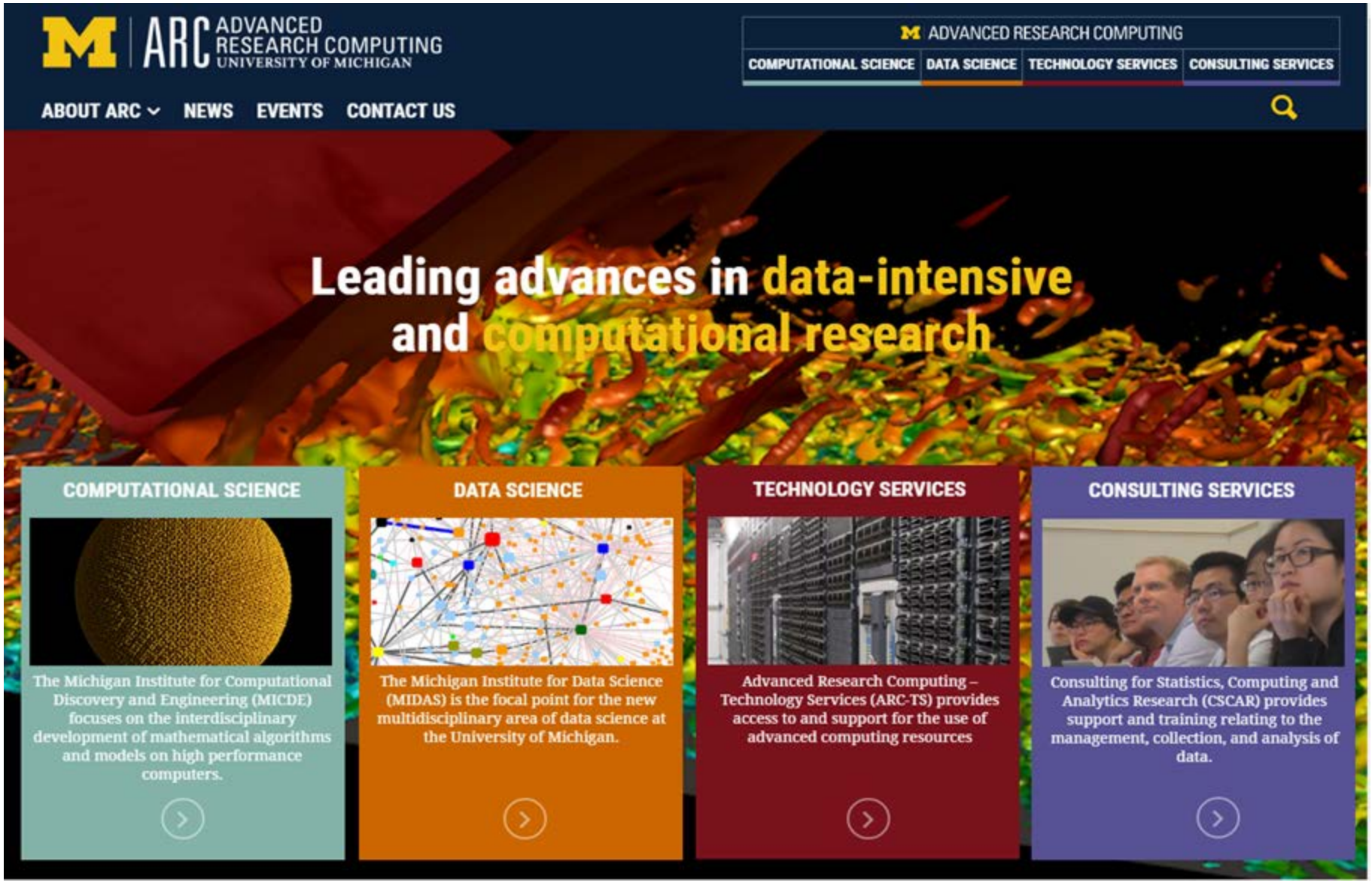
http://r4stats.com/articles/popularity/



Number of software packages is increasing

- Need for better package curation, navigation and certification
- Need for better package interoperability
- Consensus-based UL-like software standards?

Michigan Data Science Initiative



Outline

1. Changing landscape of data science
2. An engineering view of data science
3. Data science education
4. Closing thoughts

An engineering view of data science

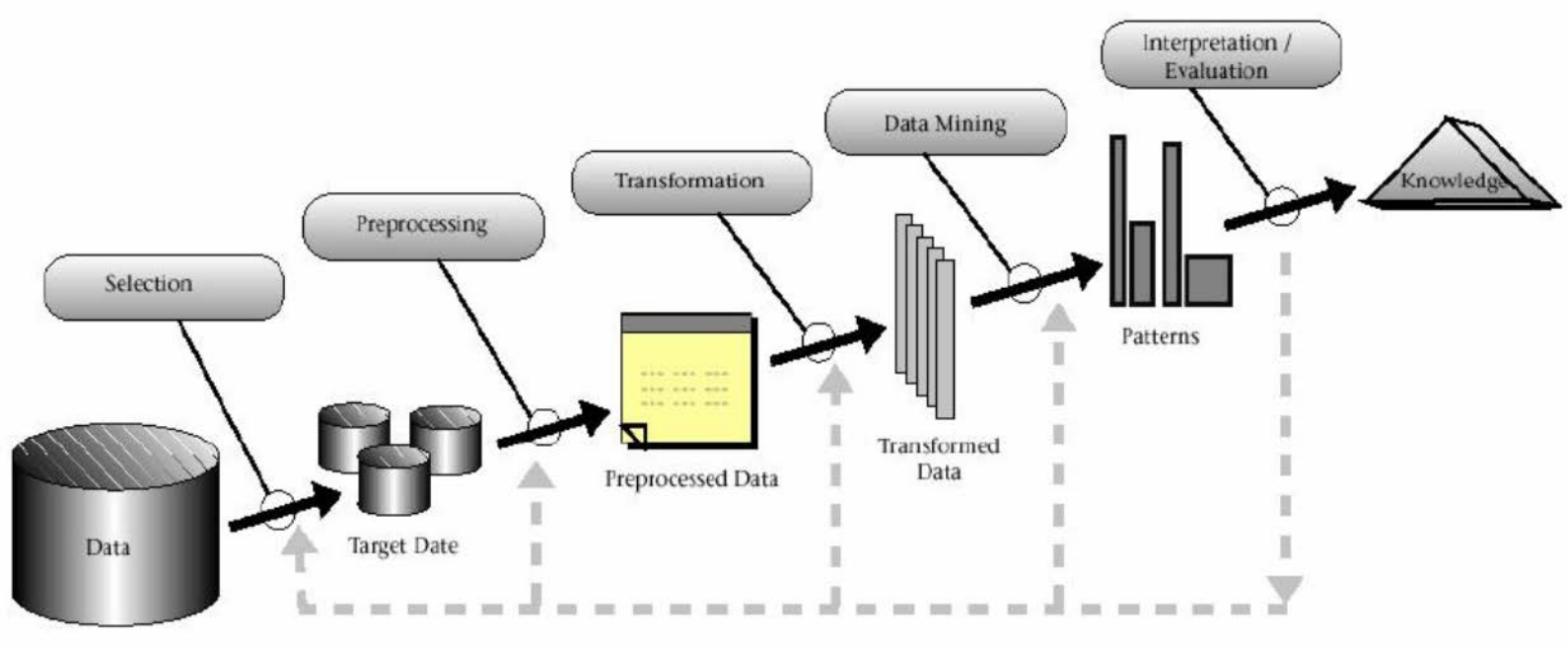
Goal develop design principles for systems that

- Collect data: sensing instruments and data repositories
  - Extract maximum value from data sources for end-use
  - Fuse data from diverse sources giving actionable information
- Manage data: resilient protected databases
  - Efficiently store, annotate, access and protect data
  - Develop standard formats for diverse data types
- Analyze data: integrated computational algorithms
  - Develop automated algorithms that handle uncertainty
  - Summarize/visualize results to maximize interpretability

Aim: to engineer a reliable data-to-decision pipeline

The Data Mining Pipeline in 1995

http://www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf



Data sources are centralized

All data Is stored locally

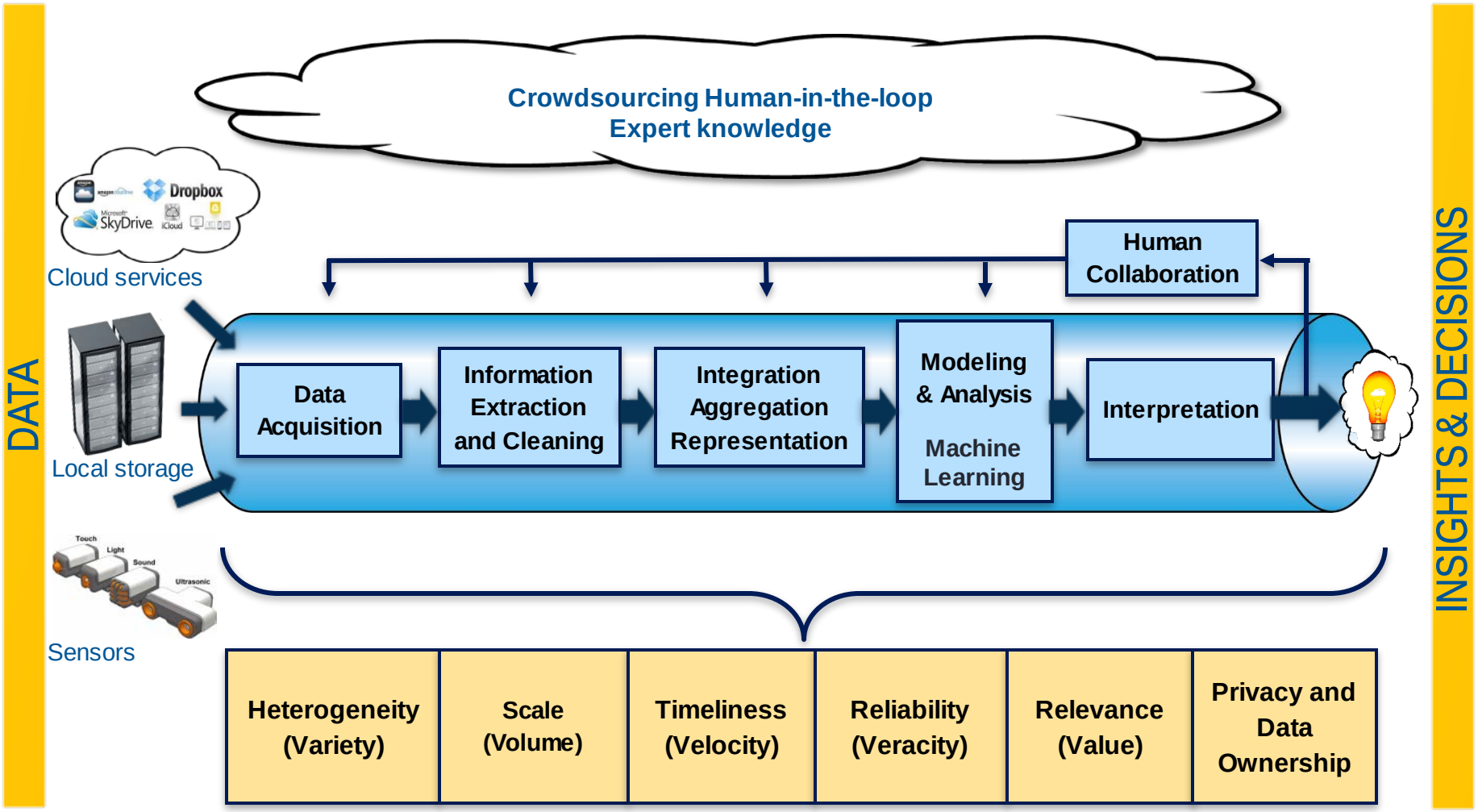
Data is homogeneous and small

Data is structured as simple list

Algorithms are primitive by today's standard

Processing not designed for Decision making

Designing the Data Mining Pipeline of Tomorrow



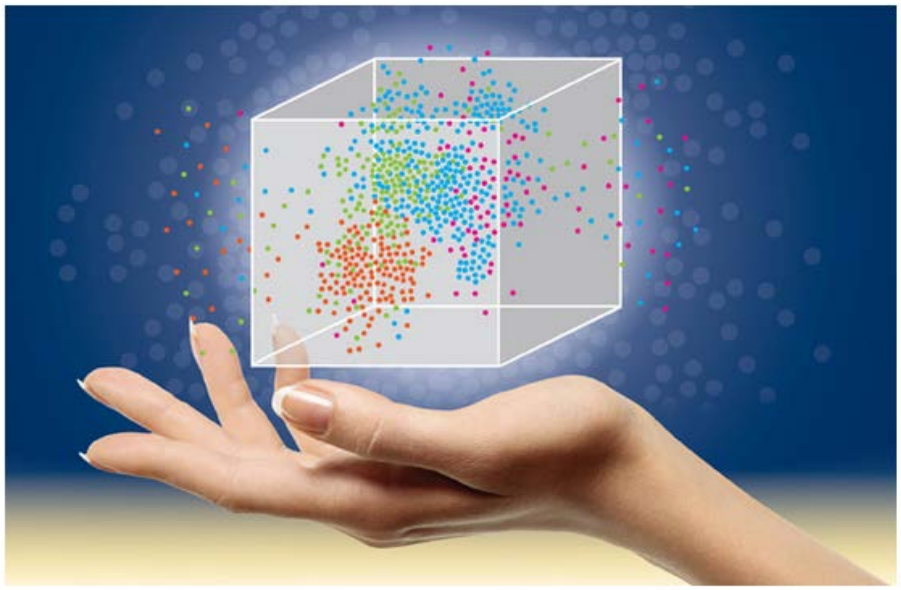
Outline

1. Changing landscape of data science
2. An engineering view of data science
3. Data science education
4. Closing thoughts

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

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

BS in DS-ENG program requirements

1. Program core (19 credits):

- EECS 203 Discrete Mathematics.
- EECS 280 Programming and Elementary Data Structures.
- EECS 281 Data Structures and Algorithms.
- STATS 412 Introduction to Probability and Statistics.
- STATS 413 Applied Regression Analysis

2. Advanced Technical Electives (at least 8 credits from list):

- Machine learning and data mining: at least 1 course
- Data management and databases: at least 1 course
- Data science applications: at least 1 course

3. Flexible Technical Electives (at least 11 credits from list)

4. Capstone Experience (4 credits from list)

5. Technical Communication and Professionalism (9 credits from list)

MIDAS Michigan Data Science Team

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

MDST

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

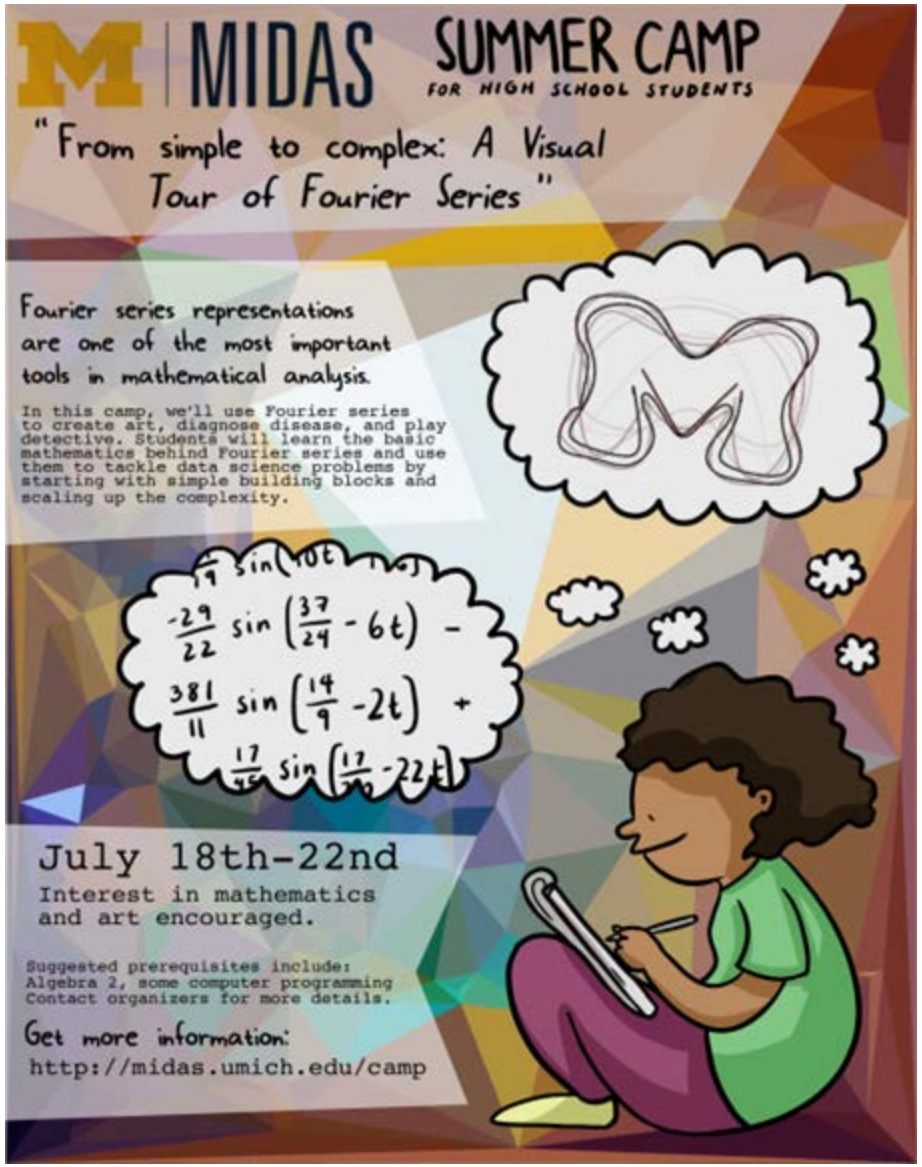
Started in 2015 to facilitate student teaming for Kaggle prediction challenges kaggle

Transitioned to public service projects (2016)

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

Sponsored by Nvidia and Google (2016)

MIDAS High School Summer Camp



A weeklong HS Summer Camp

A commuter camp open to all 9-12 graders.

2016 camp held at UM in Ann Arbor
2016 theme: Data science through Fourier series

2017 camp at UM Detroit center
2017 theme: Data science through sports data

Outline

1. Changing landscape of data science
2. An engineering view of data science
3. Data science education
4. Closing thoughts

Closing Thoughts

- Data science exists in an ecosystem of different disciplines
- Students cannot be expected to become universal experts
- Any BS/MS/PhD DS program must distill to their special brand

"A BA/BS degree in DS with a concentration in XYZ"

- Statistical inference, computation, algorithms, and data management are basic foundations of curriculum
- Experience with empirical hands-on applications is a must
- Communication skills are especially important