MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

The Future of Data Science: A Convergence of Academia, Industry, and Government
Speakers

October 6, 2015

Michigan Institute for Data Science (MIDAS) Overview

Al Hero and Brian Athey

MIDAS Co-directors

midas.umich.edu

Why Data Science? Why now?

- **Information Explosion:** Big Data methods of knowledge discovery are transforming all academic disciplines
- **Educational Transformation:** Digital data and information are transforming teaching, learning, and knowledge creation
- **Societal Demands:** New technical, social, and political solutions are required to address emerging privacy and security issues
- **Industry and Social Sectors Desire:** A new class of knowledge worker – the data science trained domain specialist
- **Cloud and Mobile Solutions:** Data Science Services and infrastructures that are modern, adaptable, and cost-effective

Data Comes in Many Forms at U of M

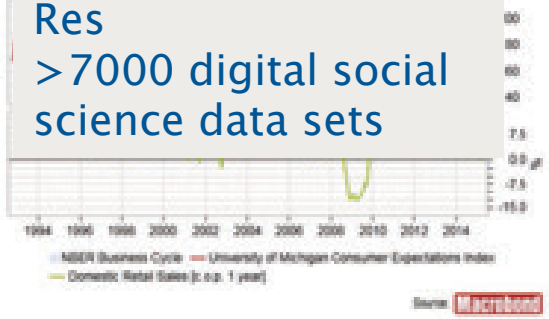
UM Health System:
15 years and >4.91M unique patient records

Electronic Health Records (UMHS)

Trans. Res. Institute
>1 petabyte cts data from >9000 vehicles

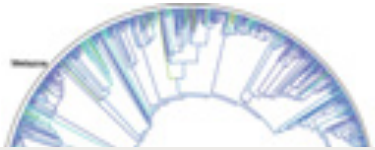
Electronic Sensor Data (UMTRI)

Institute for Social Res
>7000 digital social science data sets



Economic/Financial Data (UMISR)

Lab of Stephen A. Smith | http://blackrim.org



S. Smith – Ecol&EvolBio
>2-3 million species over 3.5 Billion years

Open Tree of Life (UMLSA)

http://socialmedia.umich.edu/

School of Information
>1 petabyte cts
Twitter feed data

Social Media Feeds (UMSI)

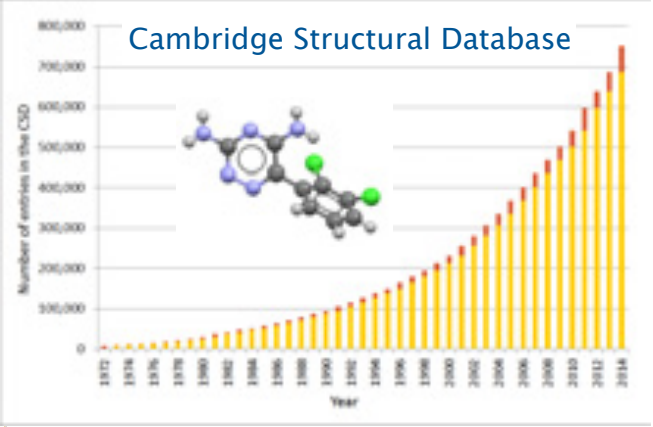
CVPR A2D Actor Action Dataset

Jason Corso –EECS
4000 videos with a total 2.5M frames

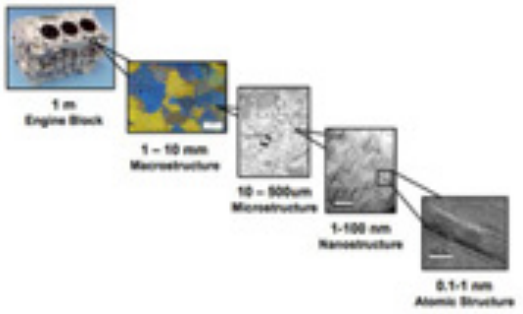
Annotated Images/Videos (UMENG)

Amount of Global Data is Growing Exponentially

Materials Science



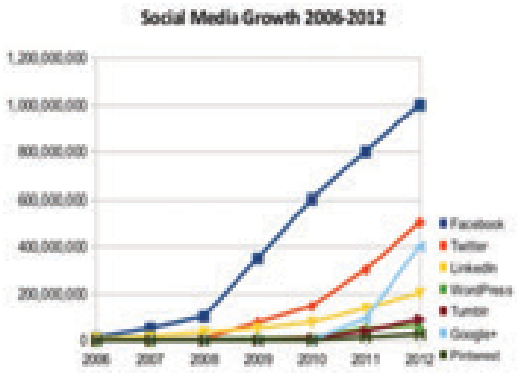
John Aliison, Mat. Sci and Eng



160,000 Engineering materials
Multiscale multiphysics

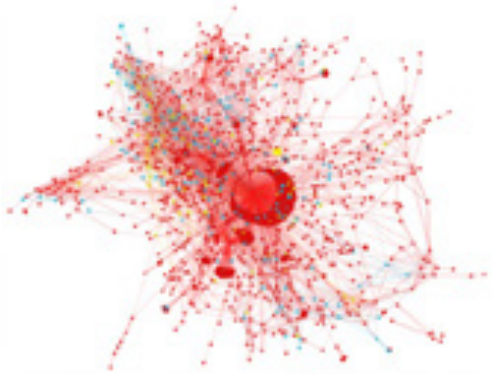
Materials Genome Initiative (UM-ENG)

Social media users



http://www.dstevenshite.com

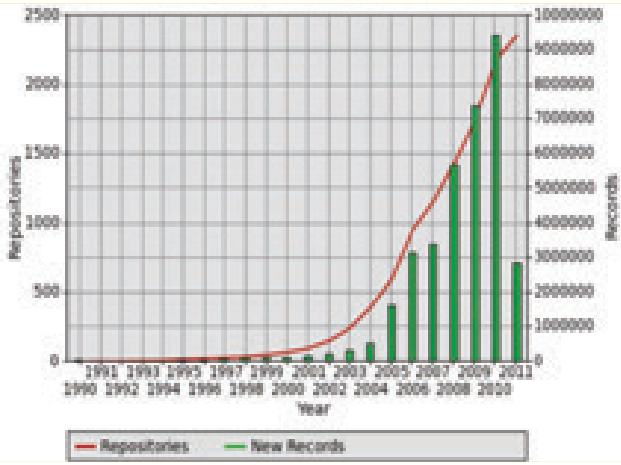
Qiaozhu Mei, SI

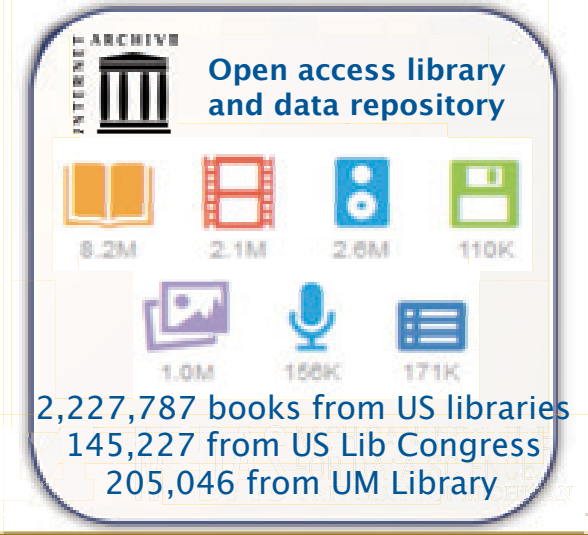


Twitter firehose generates
10,109 Tweets/sec

Twitter Rumor Tracking (UM-SI)

Registered OA Repositories

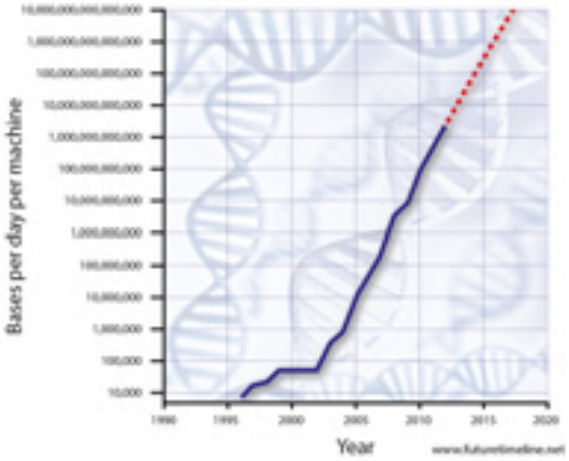




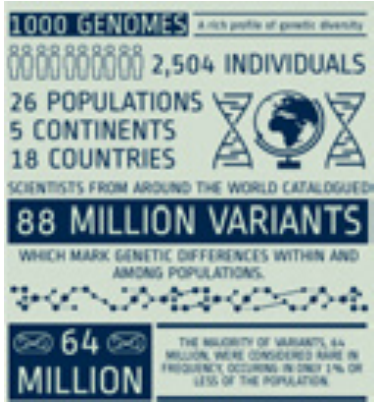
2,227,787 books from US libraries
145,227 from US Lib Congress
205,046 from UM Library

Amount of Health and Life Science Data is also Growing Exponentially

Gene sequencing

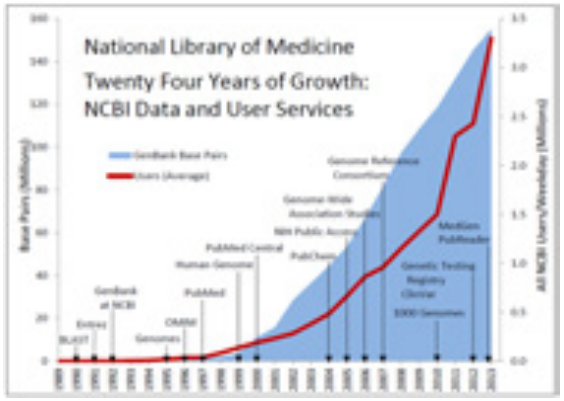


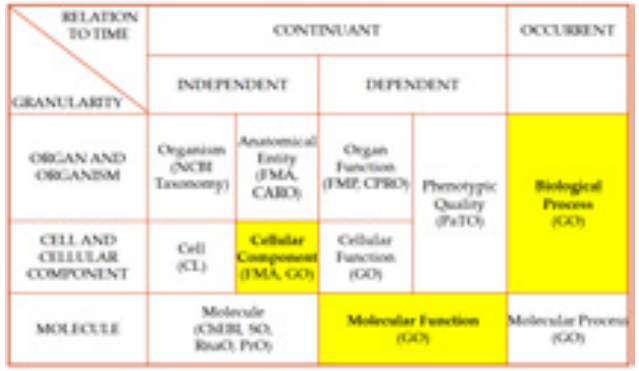
Abecasis, Nature 2015



Human genetic diversity (UM-SPH)

National Library of Medicine





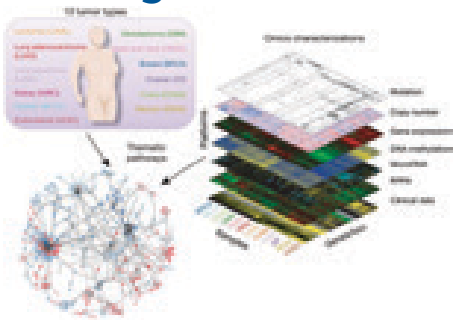
OBO Foundry Data Servers
155 ontologies, 1,768,134 terms, 100K users

Ontologic Data integration
Yonqun He, Medical School

NCBI GEO Flat Data Repository

# of Samples: 1345828

NCI Integrated Dataset



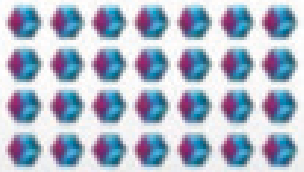
The Cancer Genome Atlas (TCGA)

Nature Genetics 45, 1113-1120 (2013)

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Dimensions of Data – the 4 V's of "Big Data"

Volume



Data at Scale

Terabytes to petabytes of data

Variety

Data in Many Forms

Structured, unstructured, text multimedia

Velocity

Data in Motion

Analysis of streaming data To enable decisions within Fractions of a second.

Veracity

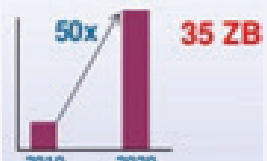
Data Uncertainty

Managing the reliability and Predictability of inherently Imprecise data types.

http://www.slideshare.net/ibmsverige/building-confidence-in-big-data

---

Cost of efficiently processing the growing Volume



Collectively analyzing the broadening Variety

80% of the worlds data is unstructured

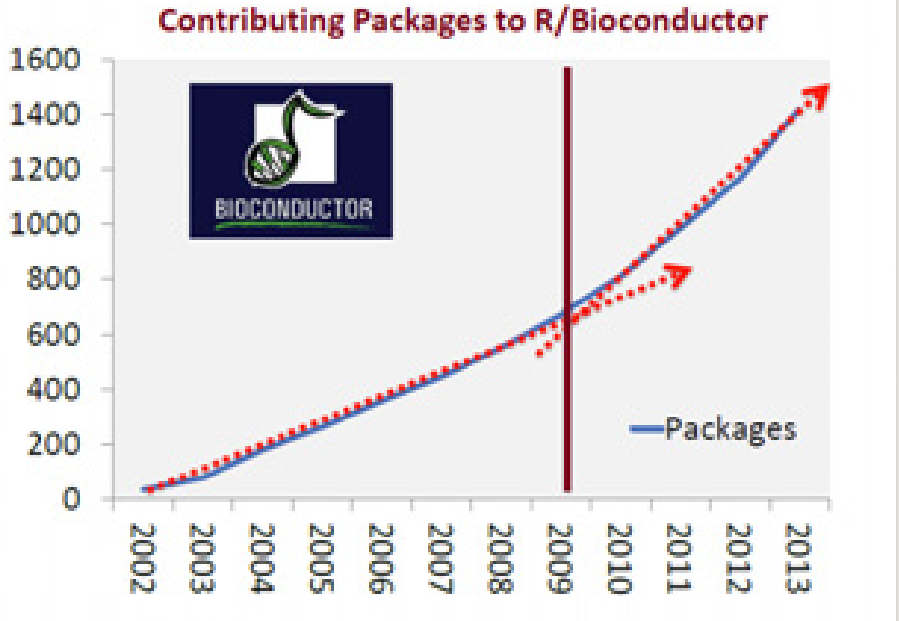
Responding to the increasing Velocity

- 10,109 tweets/ sec
- 50,659 searches/ sec
- 107,000 videos/ sec

Establishing/managing data source Uncertainty

- iPhone GPS error: 8m
- e-survey resp. rate: 25%
- % not trusting data: 33%

Open Source Software is Diverse and Growing



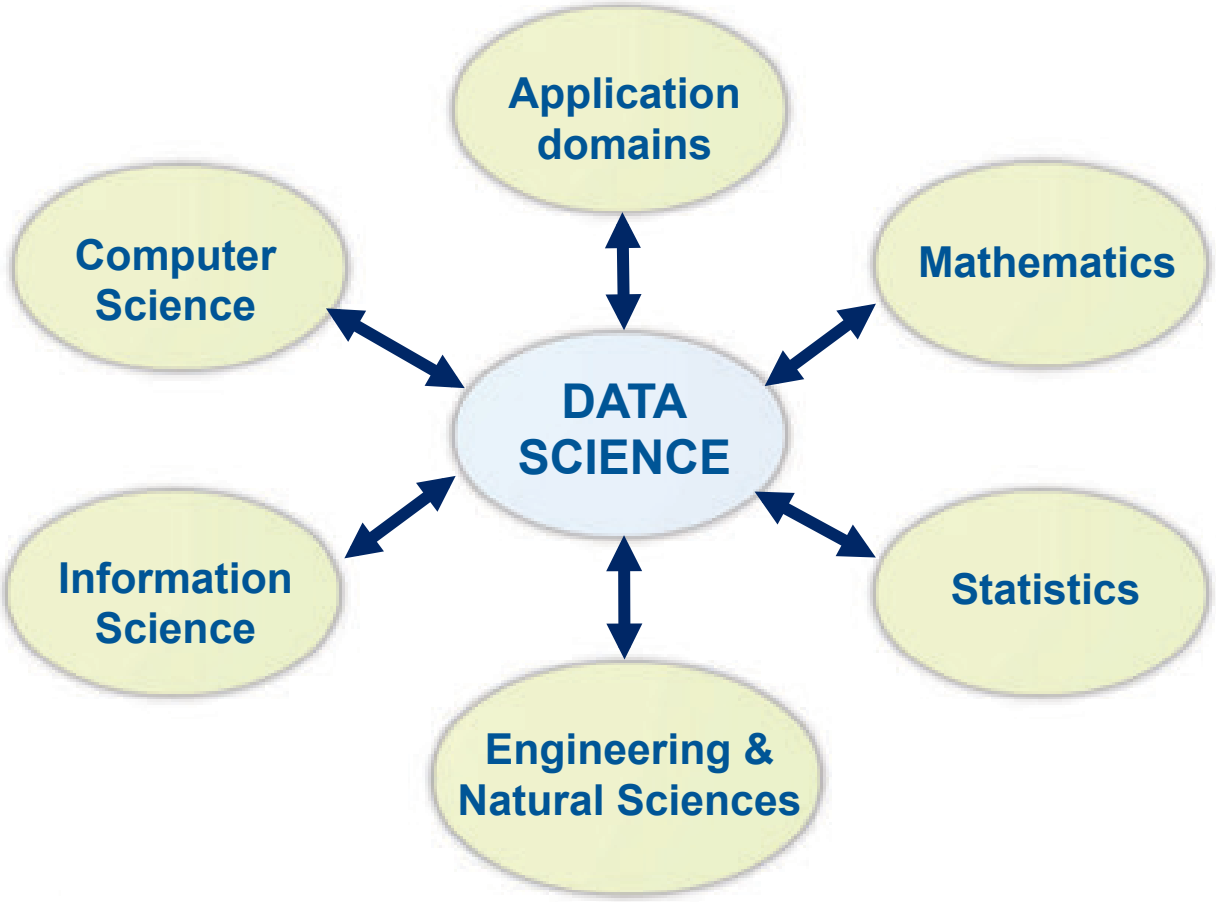
https://www.oreilly.com/ideas/2015-data-science-salary-survey

Software packages are improving year-to-year

- Faster computation and better memory management
- Better package curation and interoperability
- More data diagnostics and data cleaning features
- More reliable data analysis and data visualization

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Data Science Lies at the Multidisciplinary Interface



MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Some Questions Addressed by Data Science

- Data collection
  - What is ultimate value of a data source to end-user?
  - How best to fuse data from diverse sources?
- Data management
  - How best to efficiently store, annotate and protect the data?
  - How best to verify provenance/veracity of data?
- Data Analysis
  - How best to process and analyze complex data?
  - How best to summarize and visualize complex data?
- How to automate data-to-decision pipeline?

Data Science Paradigm

- **Principles** for turning complex data into insights and decisions
- **Methods** for data collection, mining, management, and analysis

Complex Data

analysis

Pascal Van Hentenryk, Dept IOE Univ of Michigan

Relevant information

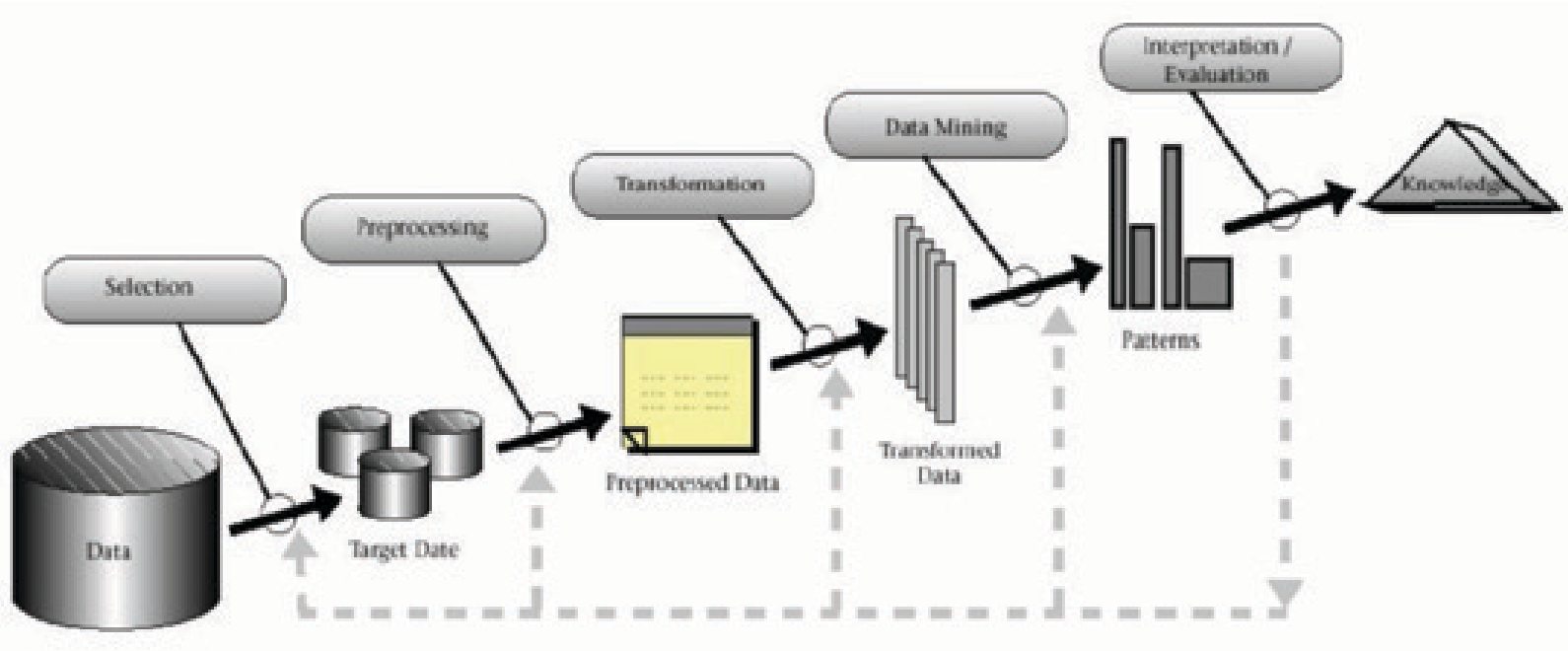
Insights

Decisions

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Data-to-Knowledge Pipeline (1996)

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



Data sources are centralized

All data Is stored locally

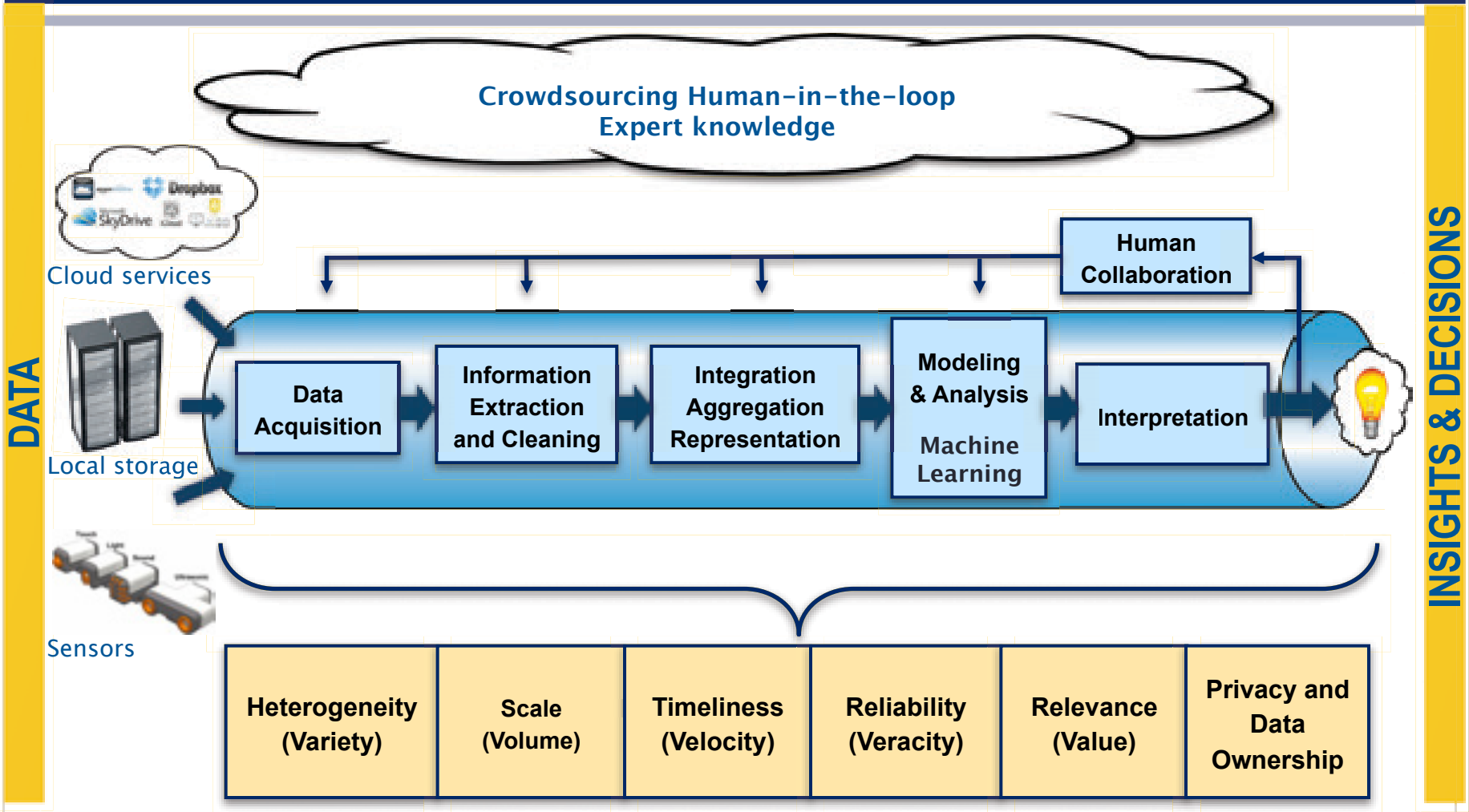
Data is homogeneous and small

Data is structured as simple list

Algorithms are primitive by today's standard

Processing not designed for Decision making

Data-to-Decision pipeline (2015 and Beyond)

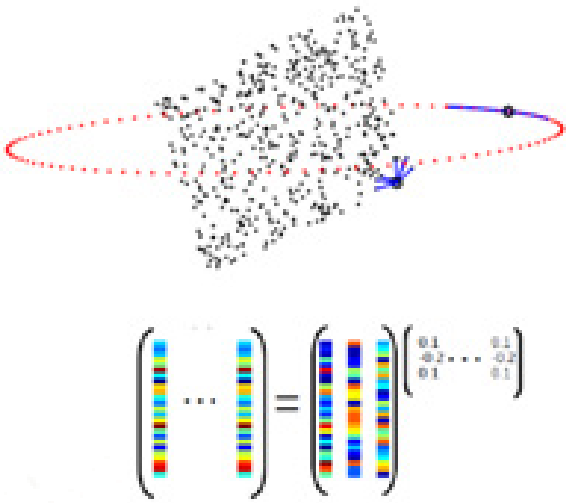


MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Data Science Methodologies

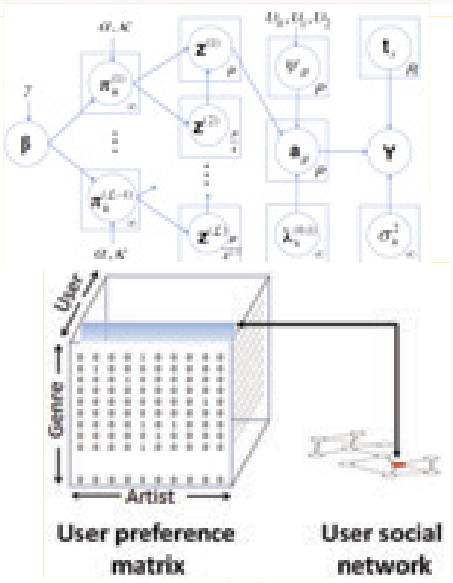
Mathematics

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



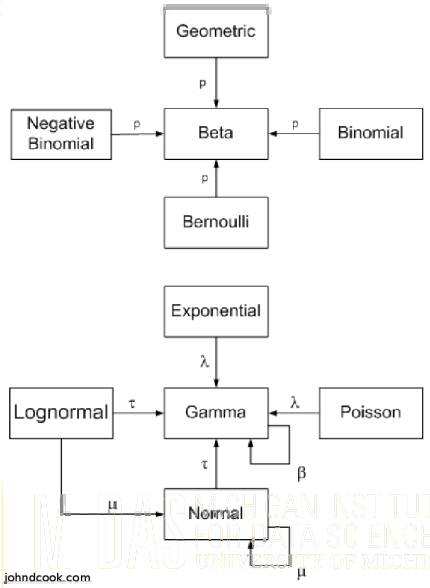
Computer Science

Natural language proc.
Graph theory
Algorithms
Database indexing
Machine learning



Statistics

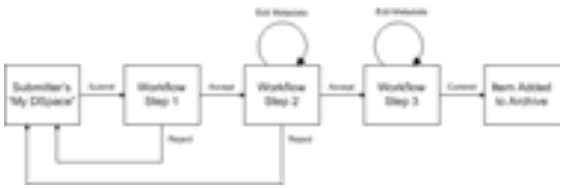
Sampling theory
Handling missing data
Experimental design
Multivariate analysis
Graphical models



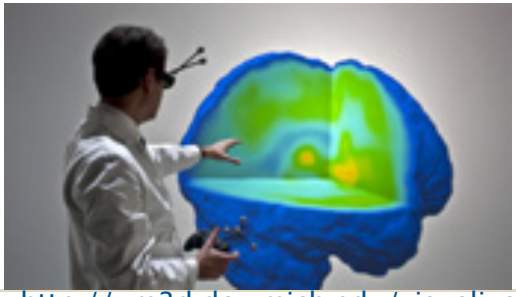
Data Science Methodologies

Information Science

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



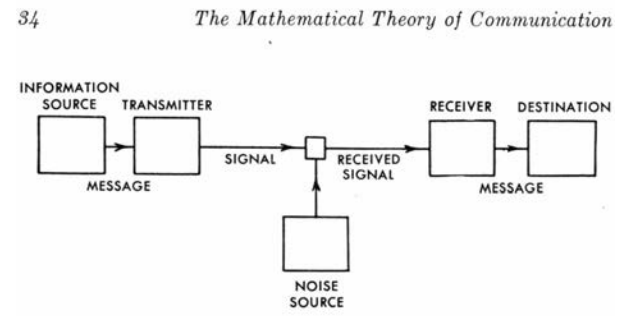
http://dspace.org/sites/dspace.org/

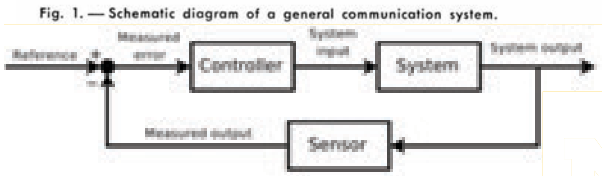


http://um3d.dc.umich.edu/visualization/

Engineering

Comm. & info. theory
Operations research
Sensors and control
Real-time computing
Cloud computing

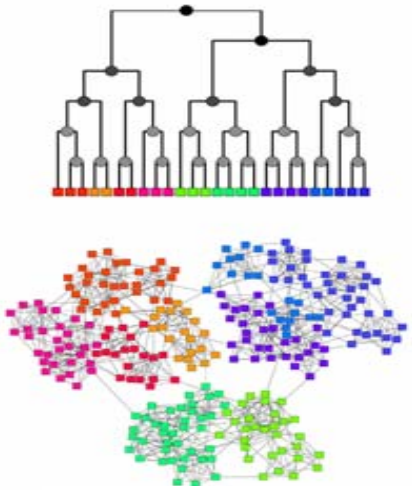




http://en.wikipedia.org/wiki/Control_theory

Physics

Network science
Complex systems
Statistical physics
Physico-mimetic models for data



Mark Newman, UM Physics

U-M is the Ideal Ecosystem for Data Science

- UM has some of the largest corpora of data and data analytics
  - Transportation and energy data (UMTRI, EI)
  - University student records
  - Electronic health records and patient testing data (UMHS)
  - Social media feeds (SI) and economic data (ISR, ICPSR, Ross)
  - Astronomy, earth sciences, materials, and phylogeny data (LSA, SNRE)
- Top ranked educational programs in data science disciplines
  - Electrical Engineering and Computer Science, Mathematics, Statistics…
- High level of engagement with business and industry
- Location in a region with a vibrant rapidly & diversifying economy

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN