Towards a science of complex data

Alfred Hero

Co-director, Michigan Institute for Data Science

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

University of Michigan

Oct 17, 2016

midas.umich.edu

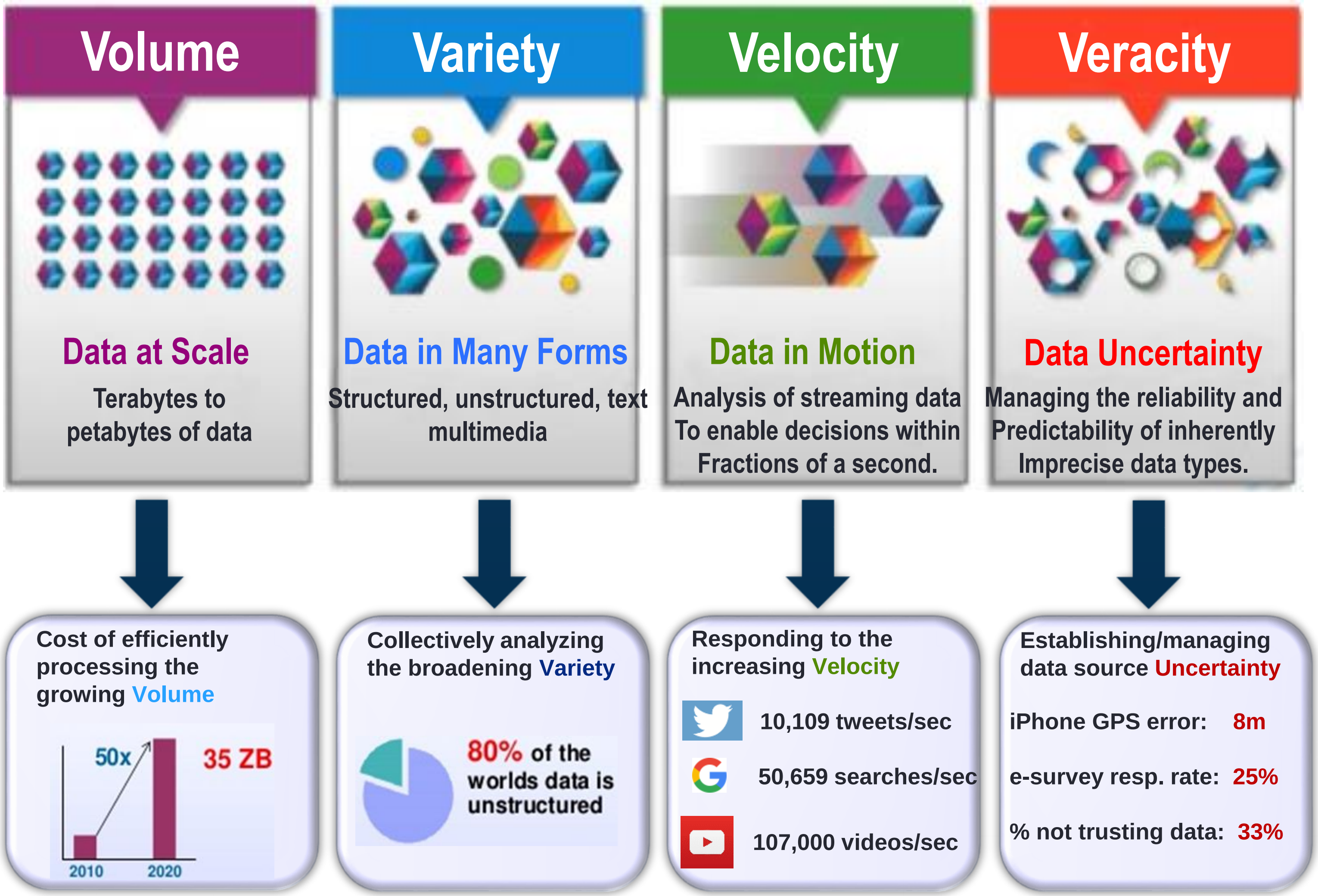
Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

Complex data: the four V's



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

Information explosion

http://blogs.cdc.gov/

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

Electronic Health Records (UMHS)

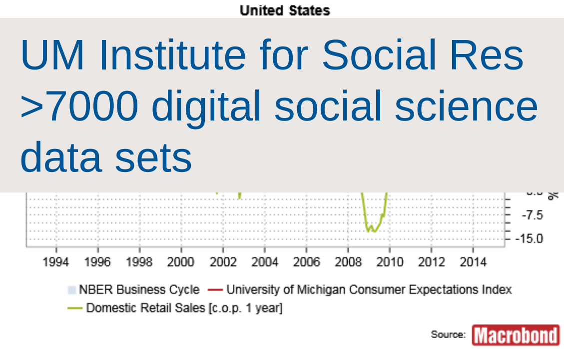
http://www.umtri.umich.edu/content/rr_45_2.pdf

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

Electronic Sensor Data (UMTRI)

http://www.sca.isr.umich.edu/

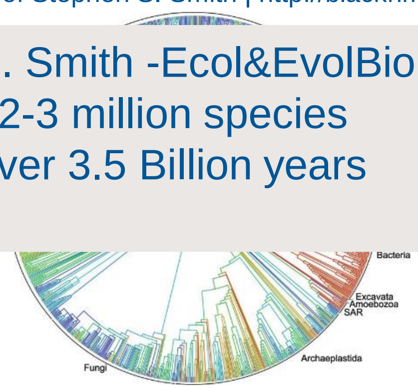
UM Institute for Social Res
>7000 digital social science
data sets



Economic/Financial Data (UMISR)

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

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



Open Tree of Life (UMLSA)

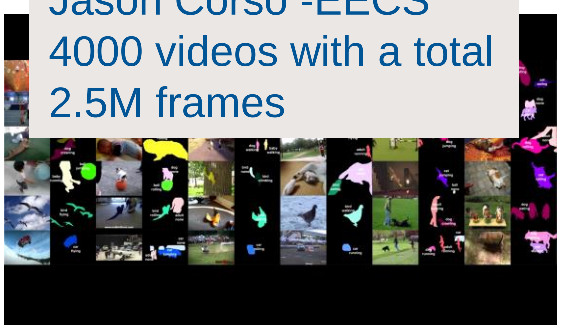
http://socialmedia.umich.edu/

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



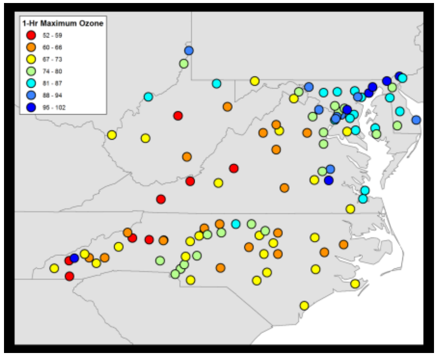
Annotated Images/Videos (UMENG)

Information explosion

Ward,MIDAS 2016

Physiological signal analysis (UM MED)

Brown, Xiao, Graham, Balzano, AGU 2015

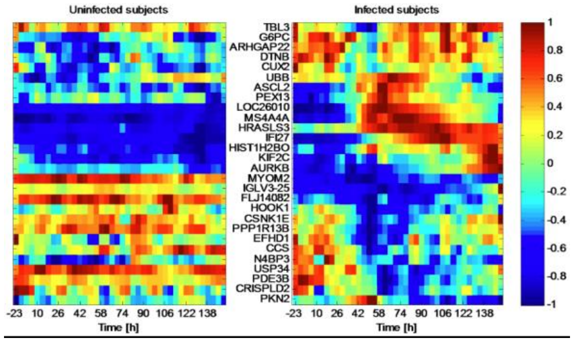


Air pollution monitoring (UM CoE)

Lipor, Balzano. Kerkez. Scavia. Allerton 2015

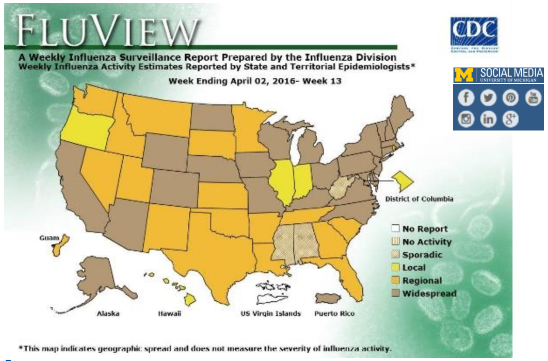
Water quality monitoring (UM SNRE)

Hero, BMC Bioinformatiocs 2015



Integrated Personal Omics profile (iPOP) (CoE)

http://www.cdc.gov/flu/weekly



Epidemic monitoring (UM SPH)

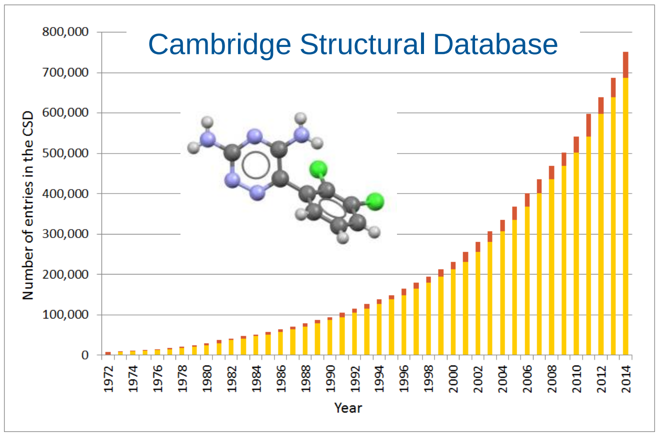
S. Murphy MIDAS 2016

MD2K
Center of Excellence for
Mobile Sensor
Data-to-Knowledge

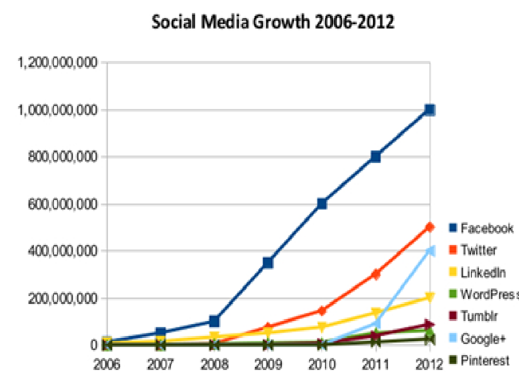
Obesity and smoking cessation (UM LSA)

Information explosion

Materials Science

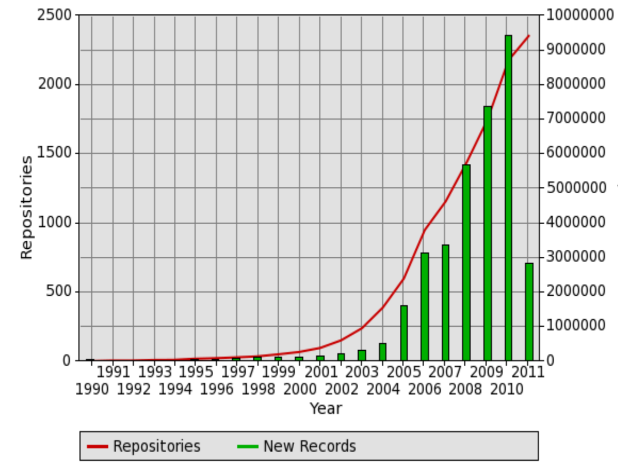


Social media users

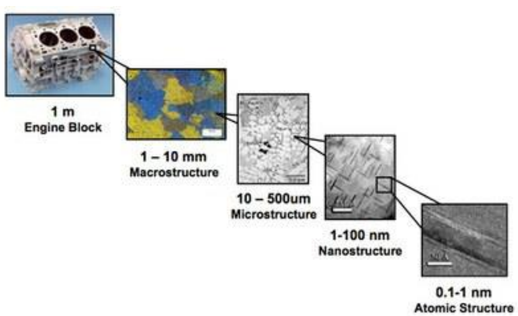


http://www.dstevenshite.com

Registered OA Repositories



John Allison, Mat. Sci and Eng



160,000 Engineering materials
Multiscale multiphysics

Materials Genome Initiative (UM-ENG)

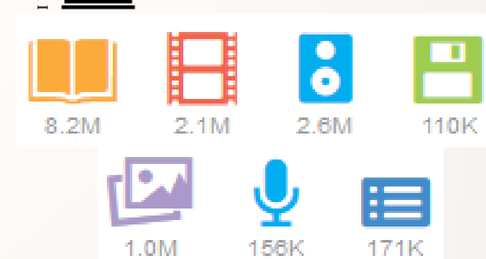
Qaozhu Mei, SI

Twitter firehose generates
10,109 Tweets/sec

Twitter Rumor Tracking (UM-SI)

INTERNET ARCHIVE

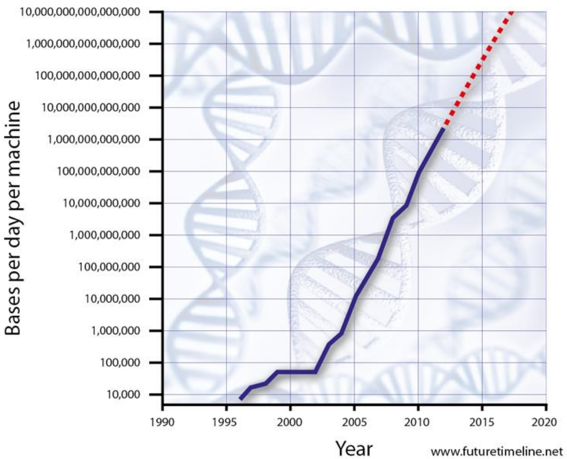
Open access library and data repository



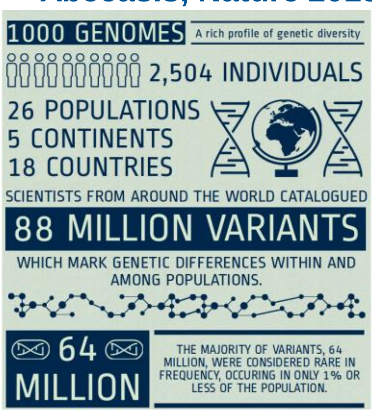
2,227,787 books from US libraries
145,227 from US Lib Congress
205,046 from UM Library

Information explosion

Gene sequencing

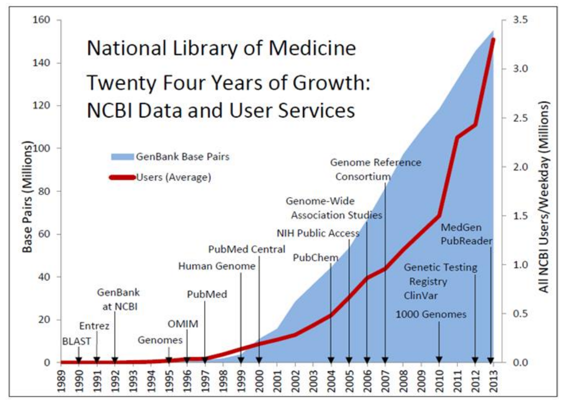


Abecasis, Nature 2015



Human genetic diversity (UM-SPH)

National Library of Medicine



RELATION TO TIME / GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	**Biological Process (GO)**
CELL AND CELLULAR COMPONENT	Cell (CL)	**Cellular Component (FMA, GO)**	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RnaO, PrO)		**Molecular Function (GO)**		Molecular Process (GO)

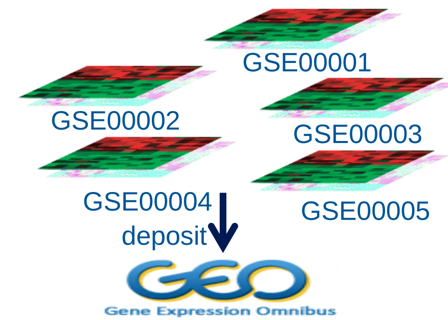
OBO Foundry Data Servers

155 ontologies, 1,768,134 terms, 100K users

Ontologic Data integration

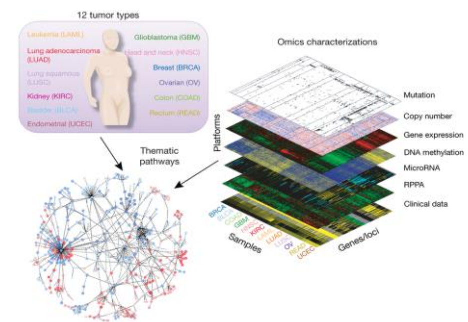
Yonqun He, Medical School

NCBI GEO Flat Data Repository



# of Samples: 1345828

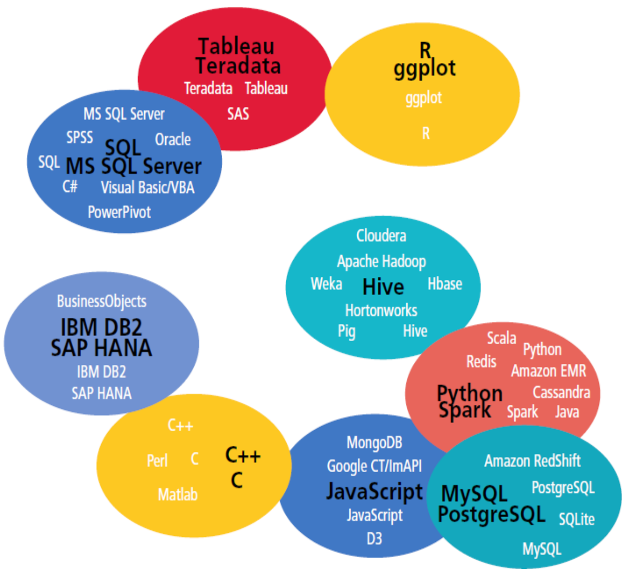
NCI Integrated Dataset

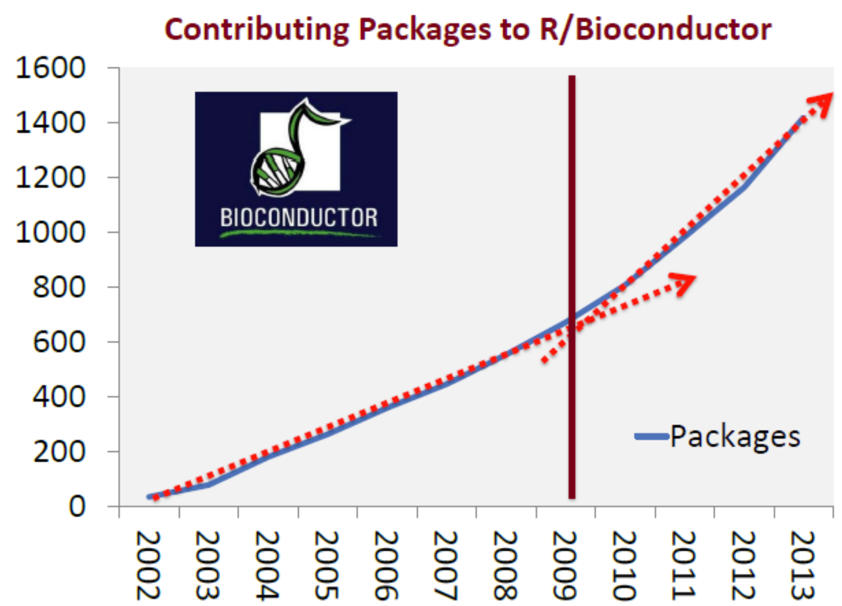


The Cancer Genome Atlas (TCGA)

Nature Genetics 45, 1113–1120 (013)

Information explosion: open source software





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

Handling data at scale

first this...

...then this.

Image from Joshua Bloom, UC Berkeley

Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

How is explosion being addressed? Data science

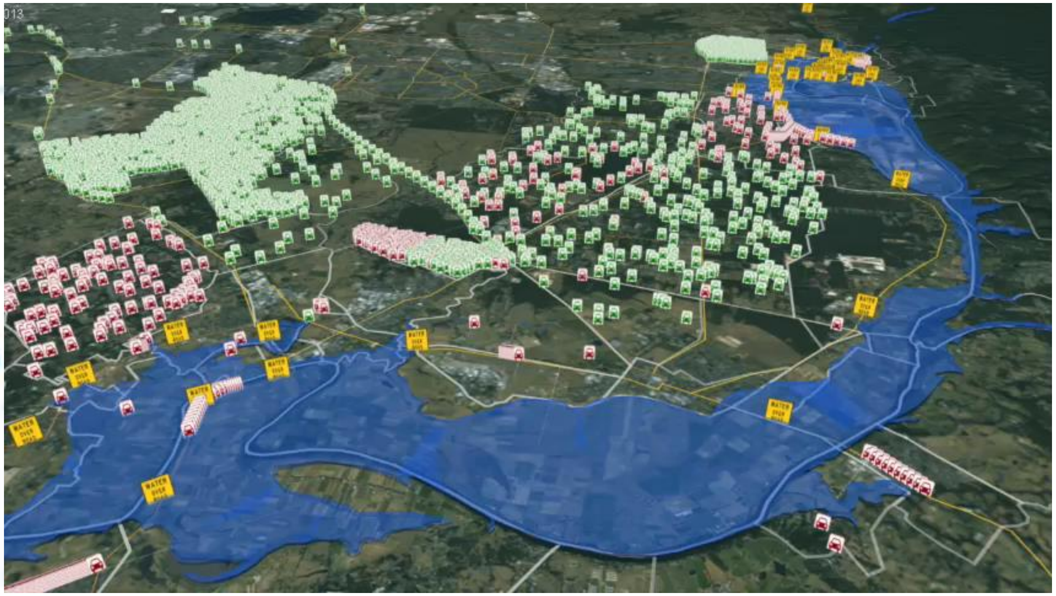
- Data collection
  - What is ultimate value of a data source to end-user?
  - How best to fuse data from diverse sources?
- Data management
  - How best to efficiently store, annotate and protect the data?
  - How best to verify provenance/veracity of data?
- Data Analysis
  - How best to process and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

Data science epistemology

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

Complex Data

analysis



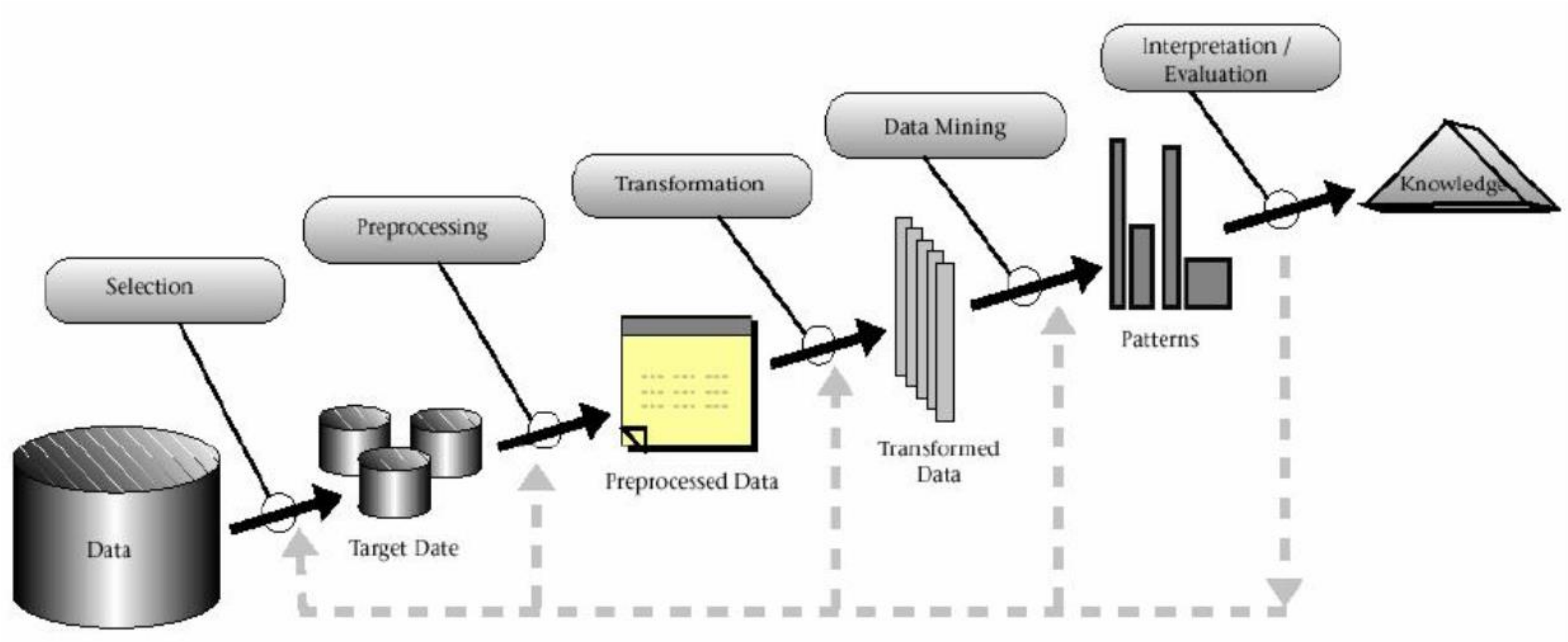
Pascal Van Hentenryk, Dept IOE Univ of Michigan

Relevant information

Insights Decisions Validation

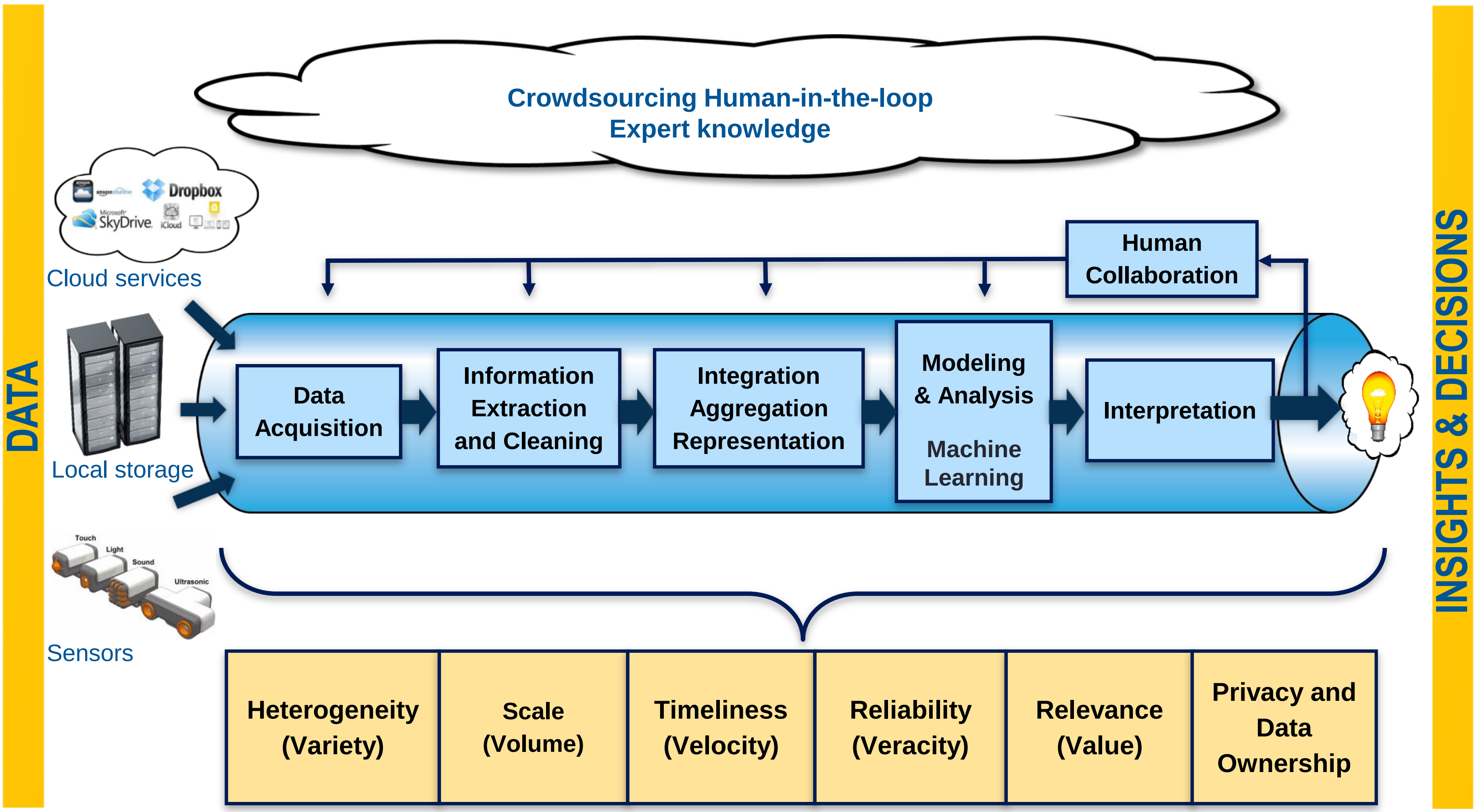
The Data Mining Pipeline in 1995

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

The Data Mining Pipeline in 2016 and Beyond



Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

New methods and algorithms for complex data

Visualization of Complex Data —networked single-user and collaborative visualization of massive multi-modality datasets.

Machine Learning-enabled Analytical Tools —Machine learning methods such as anomaly detection, dictionary learning, reinforcement learning, similarity learning, and transfer learning must be scalable to massive data scales.

Temporal, Multi-Scale and Statistical Models — Mathematical, computational and statistical models are needed to integrate multimodal data collected at many different time and length scales.

Integration of Heterogeneous Data —Integration of numerical data, symbolic data, structured data, and streaming data at various stages of the analysis pipeline.

Data Preparation, Annotation and Provenance Tracking – Automation of data preparation steps such as normalization, calibration, outlier treatment, and annotation.

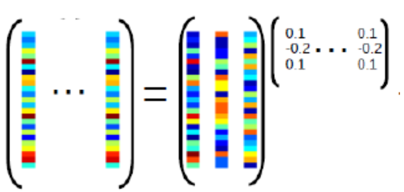
Data Privacy and Cybersecurity — The tradeoffs between data privacy/security and data utility must be understood in the context of the specific application, e.g., medicine, transportation, or business analytics, throughout the data storage, management, and analysis pipeline.

Alfred Hero, Univ. Michigan

Requires collaboration across disciplines

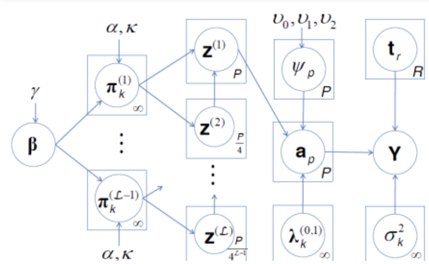
Mathematics:
Data as a matrix

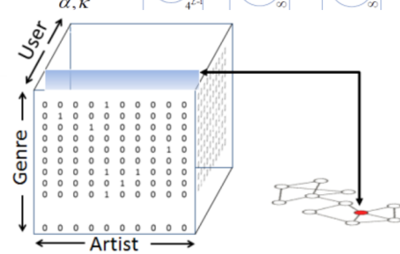
Applied topology
Harmonic analysis
Convex optimization
Num. linear algebra
Applied probability
Random matrix theory



Computer Science:
Data as a list/graph

Natural language proc.
Graph theory
Algorithms
Database indexing
Machine learning
Privacy and security



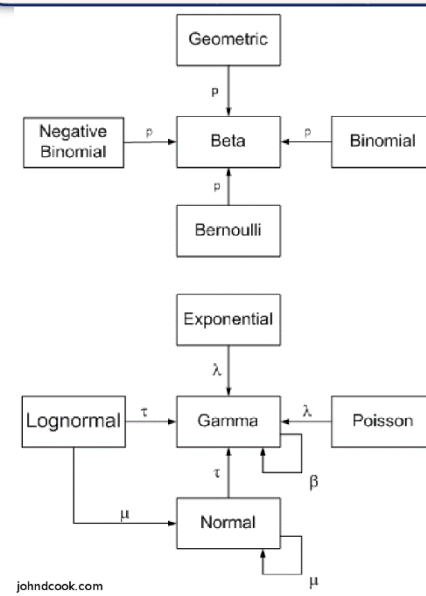


User preference matrix

User social network

Statistics:
Data as a random sample

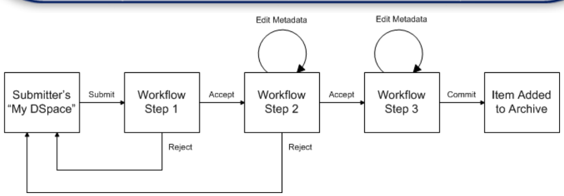
Sampling theory
Handling missing data
Robust procedures
Experimental design
Multivariate analysis
Graphical models



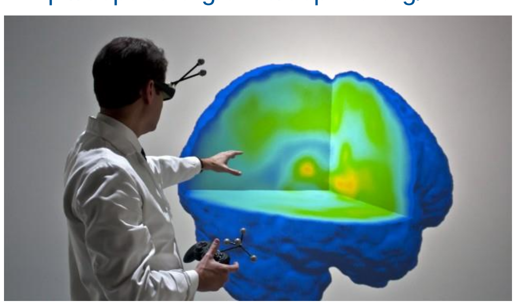
Requires collaboration across disciplines

Information Science
Data as an interface

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



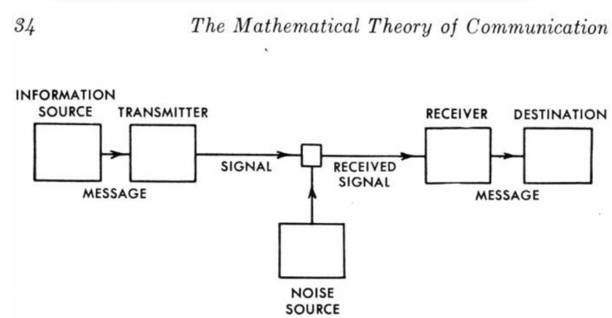
http://dspace.org/sites/dspace.org/

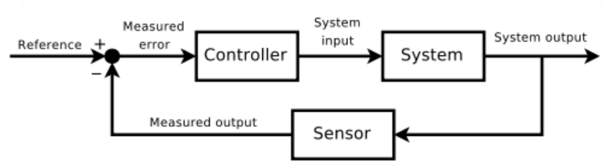


http://um3d.dc.umich.edu/visualization/

Engineering
Data2Decision

Comm. & info. theory
Signal processing
Sensors and control
Operations engineering
Real-time computing
Cyberphysical systems

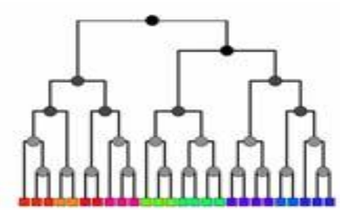


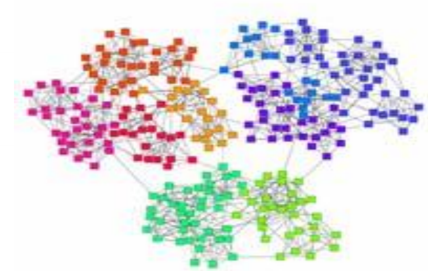


http://en.wikipedia.org/wiki/Control_theory

Physics
Data as natural phenomenon

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





Mark Newman, UM Physics

Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

Outline

1. The increasing complexity of data
2. Data science: the new science of complex data
3. Foundational principles and algorithms
4. Focus area: graph mining
5. Conclusions

Conclusions

- **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
- **Needs of Industry and Social Sectors:** 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