INTRODUCTION TO DATA MINING

Hari Sundaram

hs1@illinois.edu

http://sundaram.cs.illinois.edu

adapted from slides by Jiawei Han and Kevin Chang

DAIS@UIUC:

DATA MINING, DATABASE SYSTEMS, TEXT INFORMATION SYSTEMS, NETWORKS

Different classes in Database and Information Systems

Zhai

Sundaram

Parameswaran

Chang

Han

DATA MINING

Intro. to data mining (CS412: Han, Chang, Sundaram, Spring and Fall)

Data mining: Principles and algorithms (CS512: Han, Chang, Spring and Fall)

Seminar: Advanced Topics in Data mining (CS591Han: Fall and Spring, 1 credit)

DATABASE SYSTEMS

Introduction to database systems (CS411: Chang, Parameswaran, Sinha — Spring and Fall)

Advanced database systems (CS511: Chang, Parameswaran — Fall or Spring)

Seminar: Human in the Loop Data Management (CS598: Parameswaran, Fall)

TEXT INFORMATION SYSTEMS

Text information system (CS410 Zhai: Spring)

Advanced text information systems (CS510) Zhai: Fall)

NETWORKS + ADVERTISING

Advanced topics in Social & Information Networks (CS598, Sundaram, Spring, every two years; next class, Spring 2021)

Social & Information Networks (CS498, Sundaram, Spring, every two years; next class: Spring 2019)

Computational Advertising (CS498, Fall every year, starting Fall 2018)

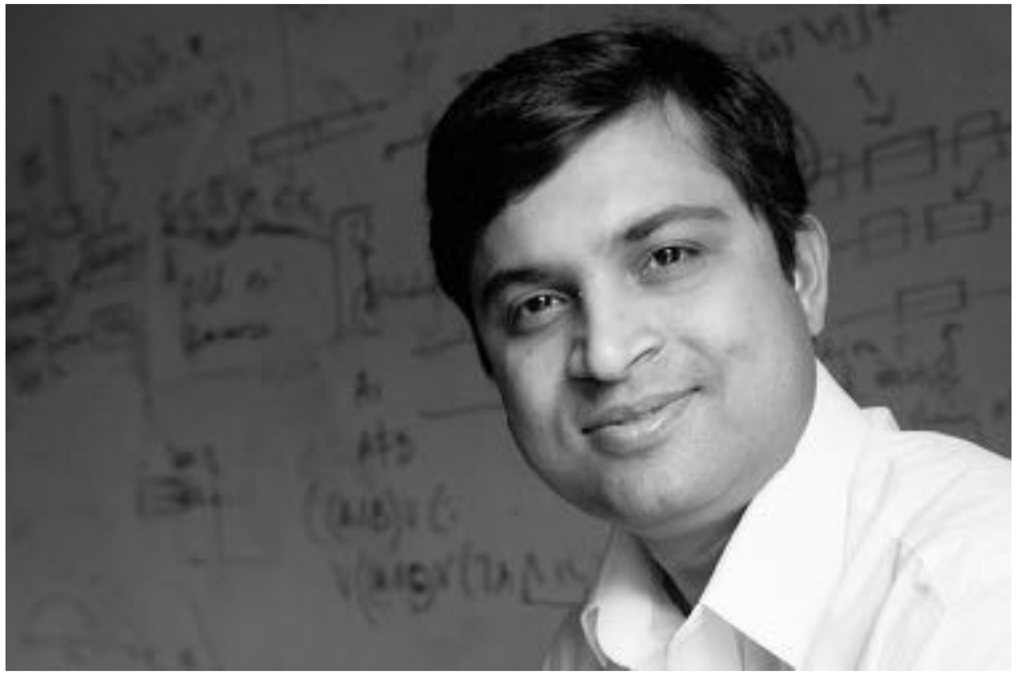
YAHOO DAIS SEMINAR

Keep in Mind

BIOINFORMATICS

Peng

Warnow



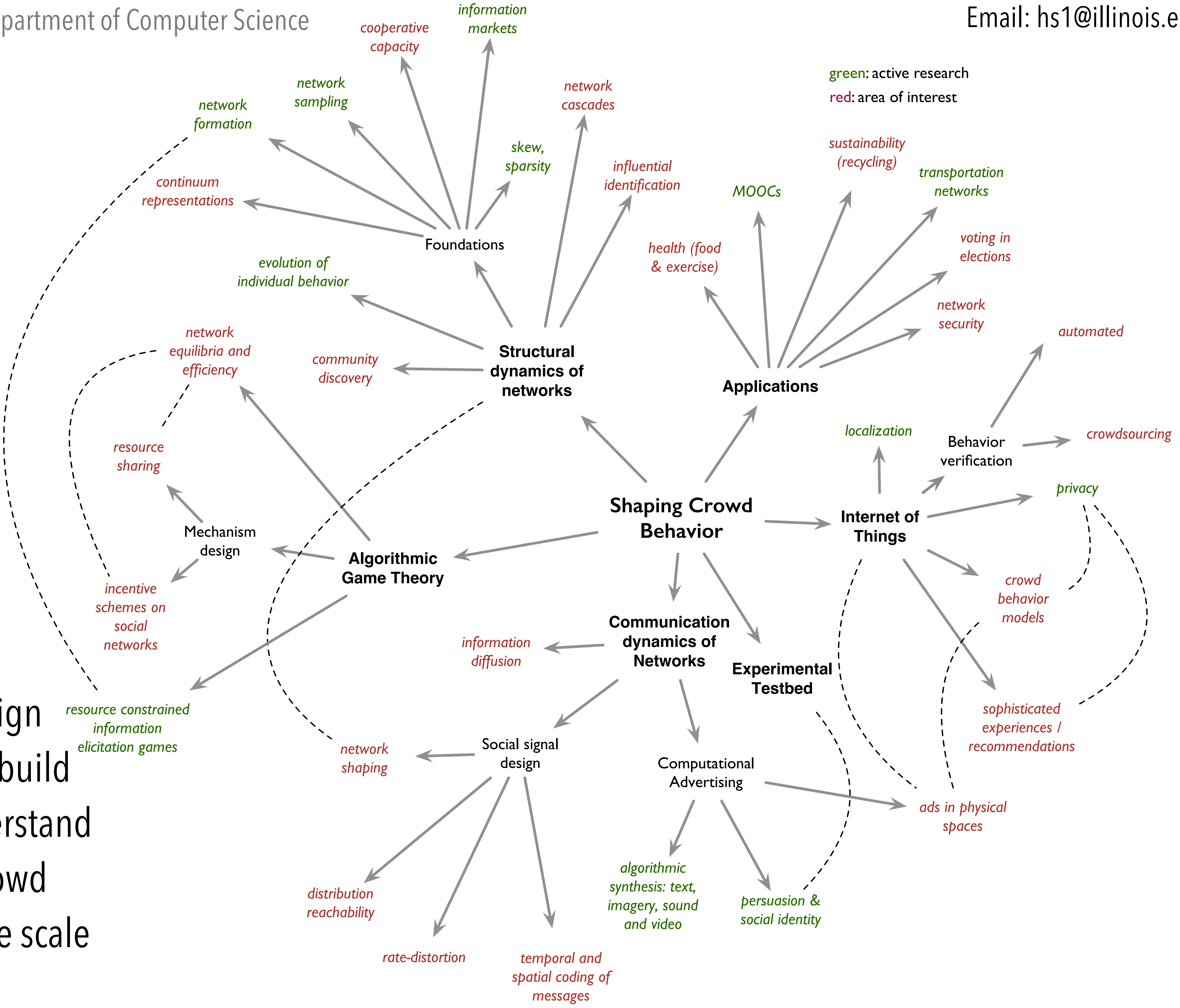
Sinha

Hari Sundaram

Associate Professor, Department of Computer Science

Web: http://sundaram.cs.illinois.edu/

Email: hs1@illinois.edu



green: active research

red: area of interest

Shaping Crowd Behavior

Structural dynamics of networks

Foundations

network formation

network sampling

cooperative capacity

information markets

skew, sparsity

continuum representations

network cascades

influential identification

evolution of individual behavior

community discovery

Applications

health (food & exercise)

MOOCs

sustainability (recycling)

transportation networks

voting in elections

network security

Internet of Things

localization

Behavior verification

automated

crowdsourcing

privacy

crowd behavior models

sophisticated experiences / recommendations

Algorithmic Game Theory

network equilibria and efficiency

Mechanism design

resource sharing

incentive schemes on social networks

resource constrained information elicitation games

Communication dynamics of Networks

information diffusion

Social signal design

network shaping

distribution reachability

rate-distortion

temporal and spatial coding of messages

Computational Advertising

algorithmic synthesis: text, imagery, sound and video

persuasion & social identity

ads in physical spaces

Experimental Testbed

What I do: design algorithms and build systems to understand and to shape crowd behavior in large scale social networks

Motahhare

Himel (**online TA**)

Suhansanu

MEET THE TA'S

Kanika

Subham

CS412 CLASS MECHANICS

Everything you wanted to know

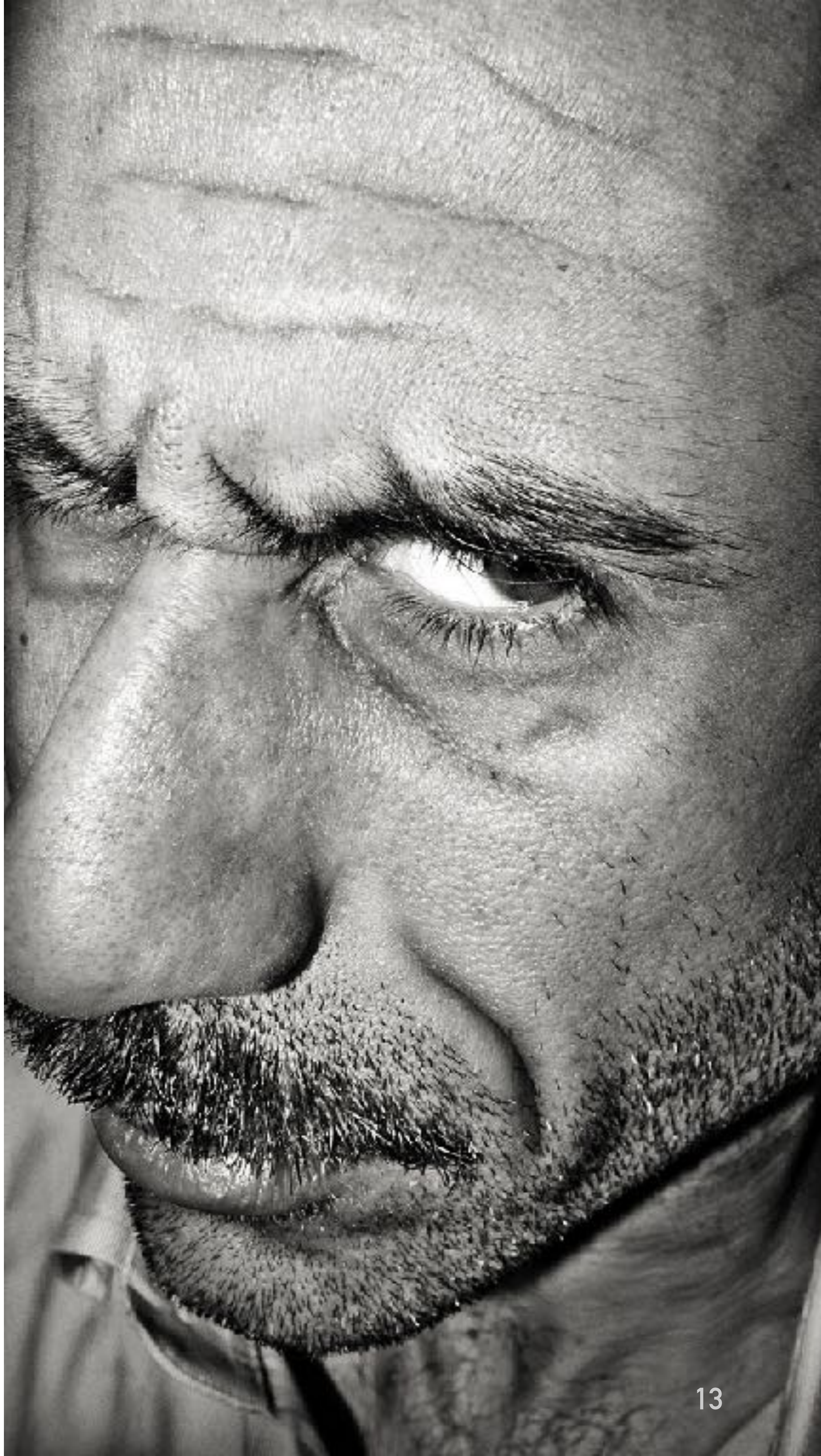
class website:

https://wiki.illinois.edu/wiki/display/cs412sp18/Syllabus

lectures are online

https://echo360.org/home

BUT..



WHY YOU SHOULD COME

A lot of research shows that students who come to class tend to score better in exams

We'll be solving problems in class that will help with understanding of the material

sign up on piazza!

https://piazza.com/class/jc0rew6qwc14lf

assignments

Written + programming assignments (5)

8x5=**40** points

quizzes

goal: regular review

quizzes (5) 12x5=**60** points

Final exam

37 points

rest of the grade scaled to 75%

project

kaggle.com

Four credit (**mandatory**)	25 points	25%
extra credit, three credit (PS3)	10 points	10%

quizzes and **assignments** are **interleaved**

expect something every week

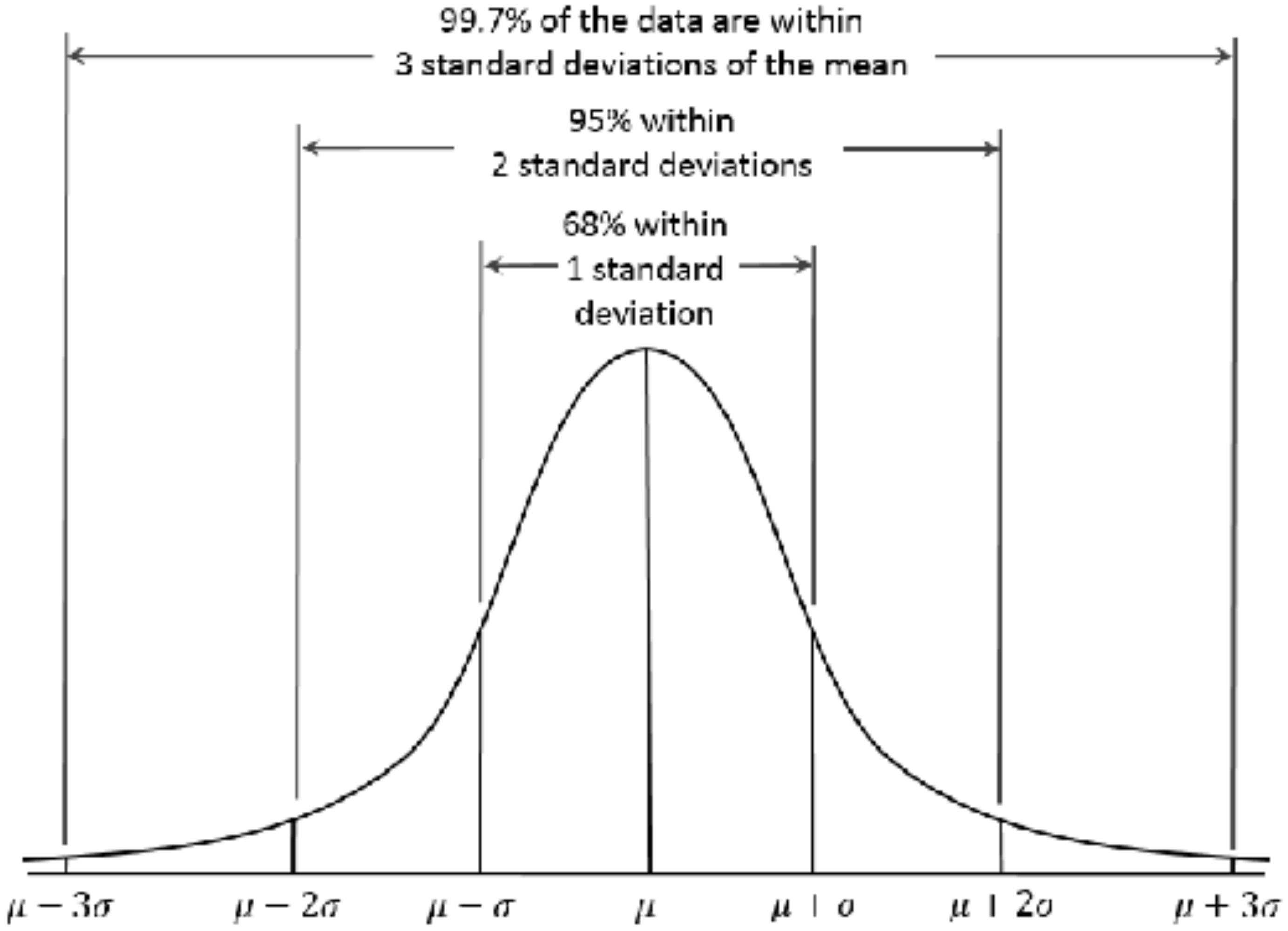
first HW in **two weeks**

first quiz in **three weeks**

Class participation on Piazza: **up to 5% extra credit**

mapping from effort to points will be finalized this week

GRADING



Will grade on a curve

Will grade undergrads and grads on the same curve—there is no difference in performance.

Grad students taking the 4 credit class will need to do an extra project worth 25% of the grade

Note: In Spring 2017, the median grade for CS 412 was **'A-'**

academic integrity

zero tolerance policy!

You are encouraged to form a study group to discuss the homework and the programming assignments but are expected to compete the homework and programming assignments **completely on your own without recourse to notes from the group discussions**.

CS Honor code http://www.cs.uiuc.edu/academics/honor-code

Plagiarism: **It is an academic violation to copy, to include text from other sources, including online sources, without proper citation.**

CS Honor code http://www.cs.uiuc.edu/academics/honor-code

Any student found to be violating this code will be subject to **disciplinary** action.

CS Honor code http://www.cs.uiuc.edu/academics/honor-code

wade

swim

dive

what do you want to do?

Why are you taking this class?

why are you excited about this class?

what are you concerned about?

WHY DATA MINING?

There is an explosive growth of data: from terabytes to petabytes

MAJOR SOURCES OF DATA

Business:

Web, e-commerce, transactions, stocks, …

Science:

Remote sensing, astronomy, bioinformatics, scientific simulation, …

Society and everyone:

news, digital cameras, YouTube

Cisco expects 70% of **all** internet data to be video

“

We are drowning in data, but starving for knowledge

-John Naisbitt, 1982.

WHAT IS DATA MINING?

Extraction of interesting (**non-trivial, implicit, previously unknown and potentially useful**) patterns or knowledge from huge amount of data

Is Data Mining a misnomer?

We don’t mine for data!

Also known as

Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

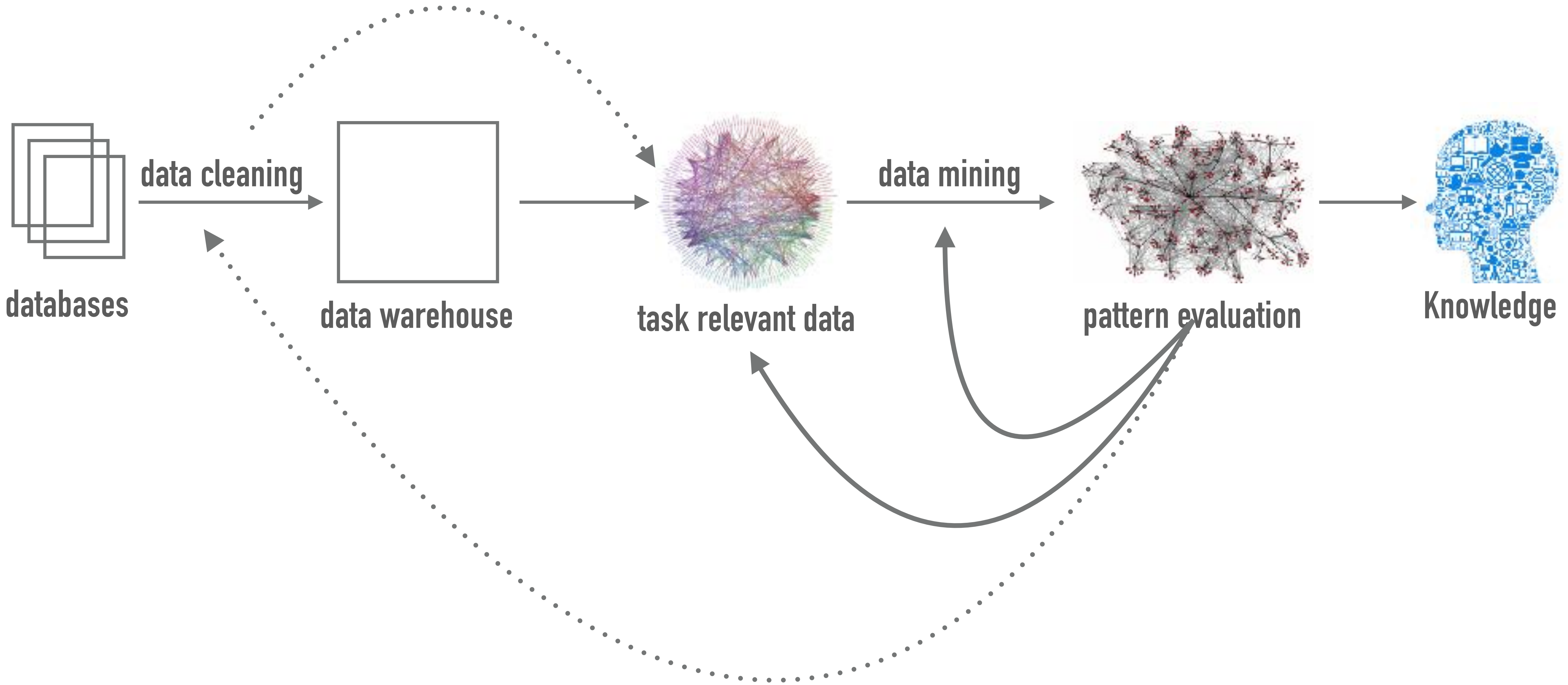
Is everything data mining?

Simple search and query processing

(Deductive) expert systems

The Knowledge discovery process

database, data warehousing community view



EXAMPLE: RETAIL MINING

Data cleaning

Data integration from multiple sources

Warehousing the data

Data cube construction

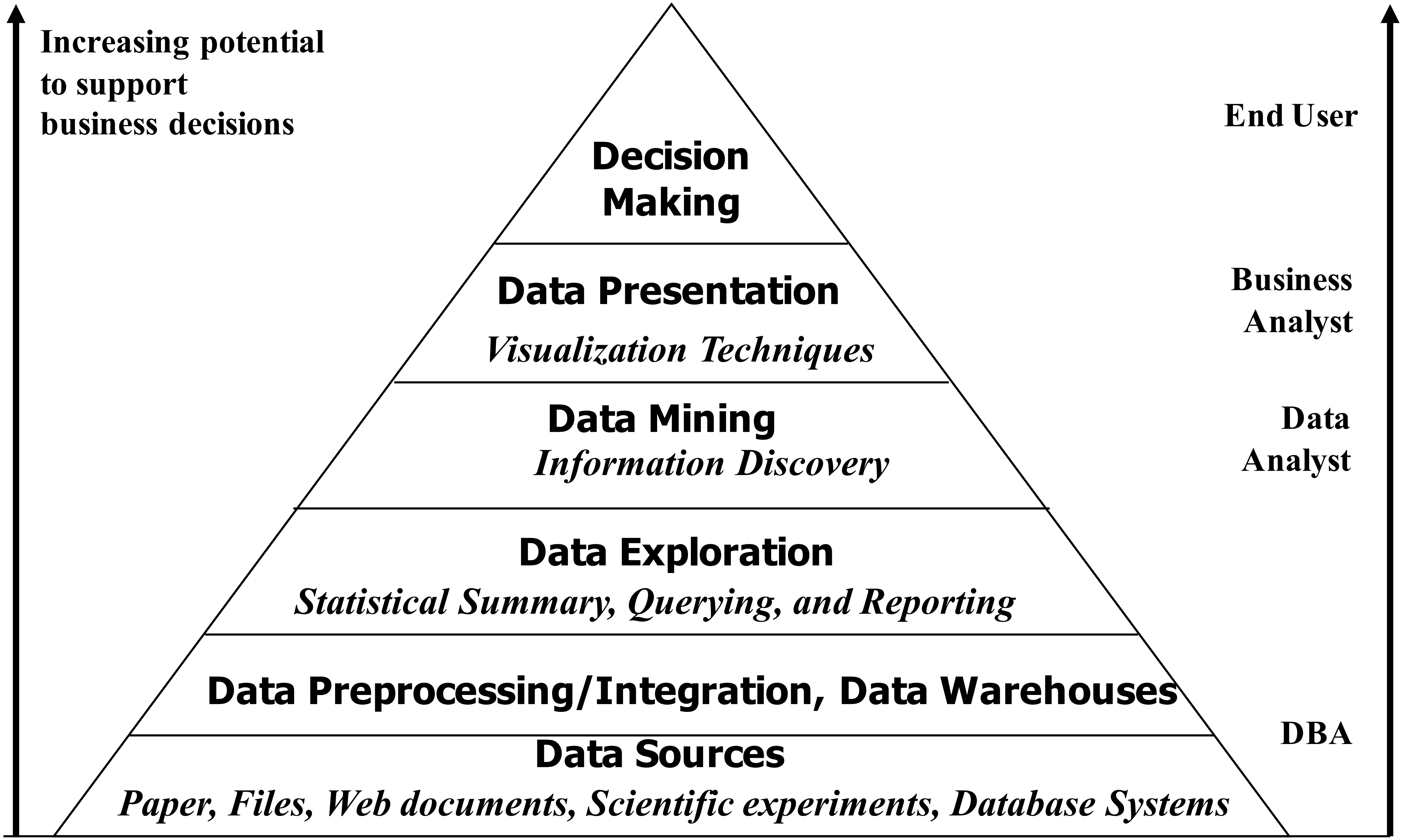
Data selection for data mining

Data mining

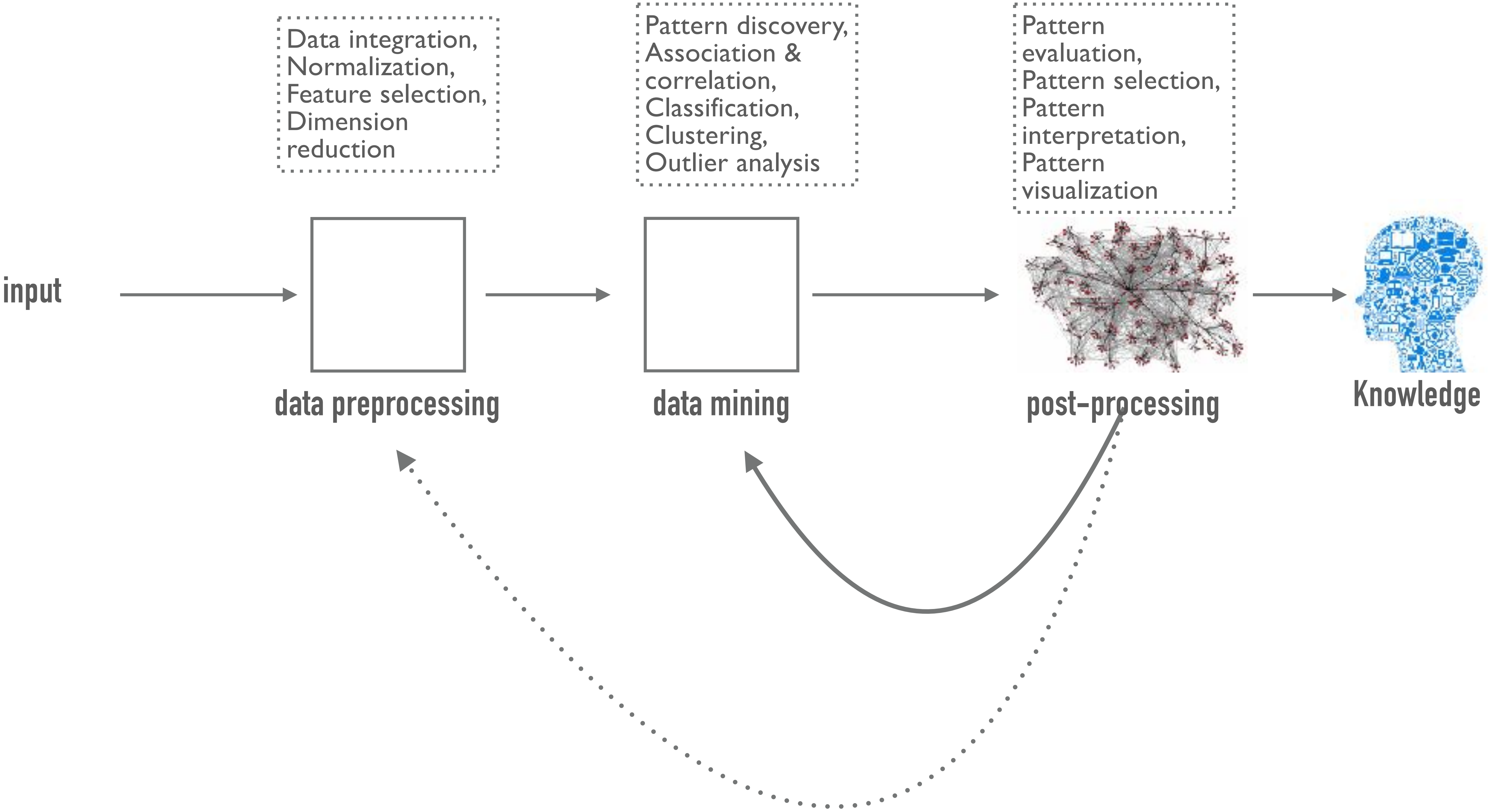
Presentation of the mining results

Patterns and knowledge to be used or stored into knowledge-base

Data mining in business intelligence



The Machine Learning / Statistics View



input

Data integration, Normalization, Feature selection, Dimension reduction

data preprocessing

Pattern discovery, Association & correlation, Classification, Clustering, Outlier analysis

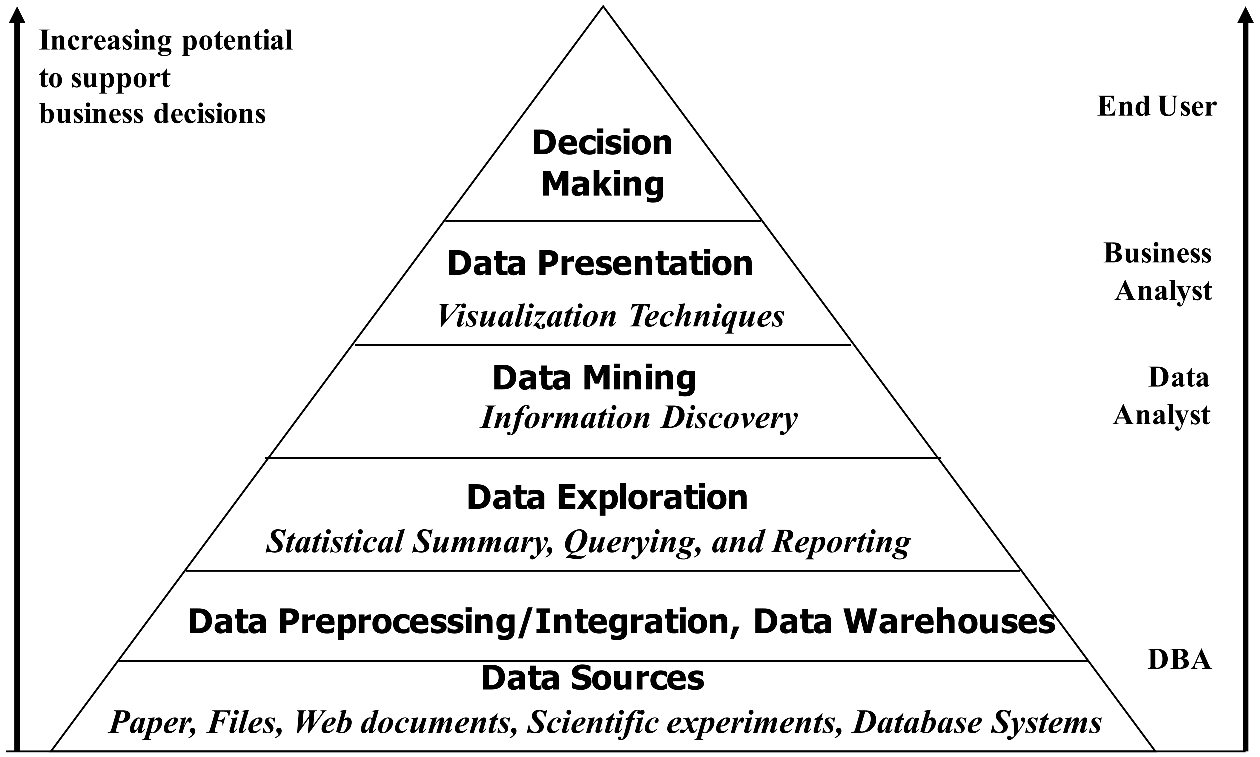
data mining

Pattern evaluation, Pattern selection, Pattern interpretation, Pattern visualization

post-processing

Knowledge

WHICH VIEW DO YOU PREFER?



Increasing potential to support business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

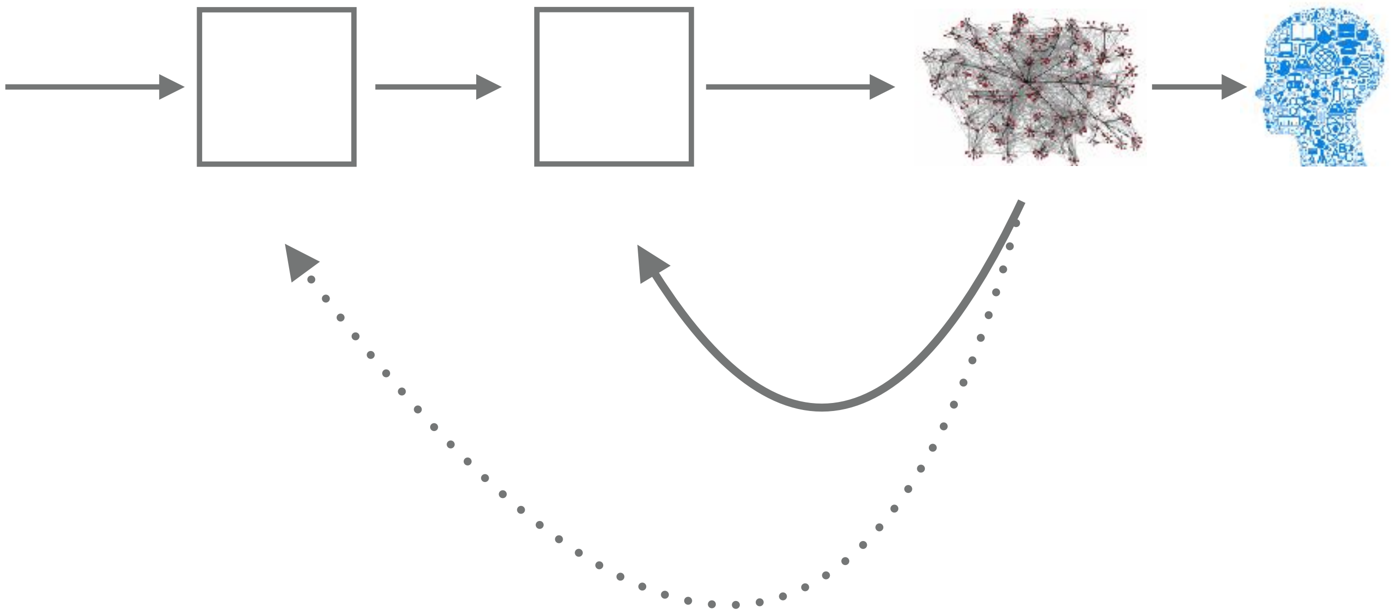
Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

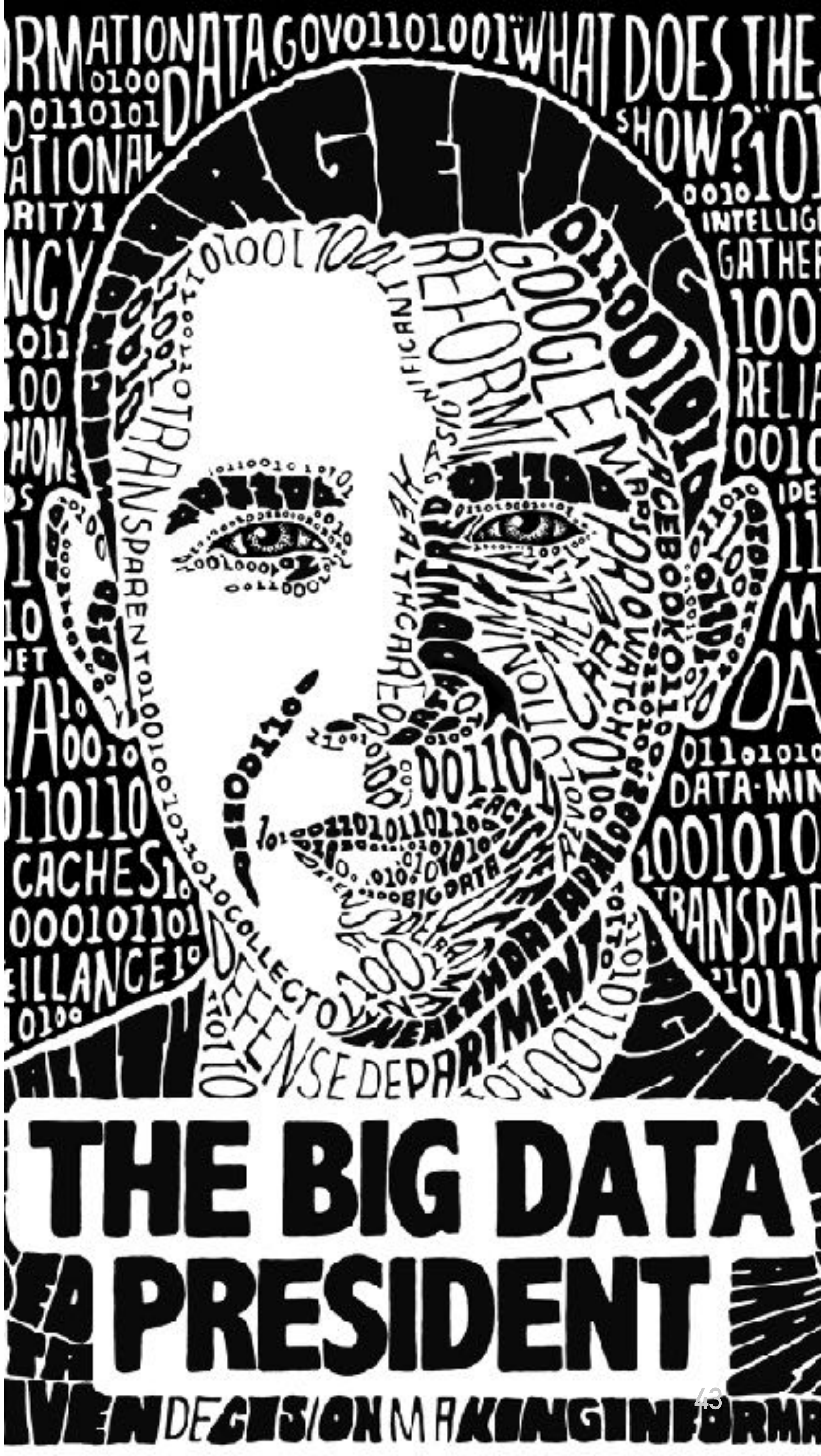
Data Analyst

DBA

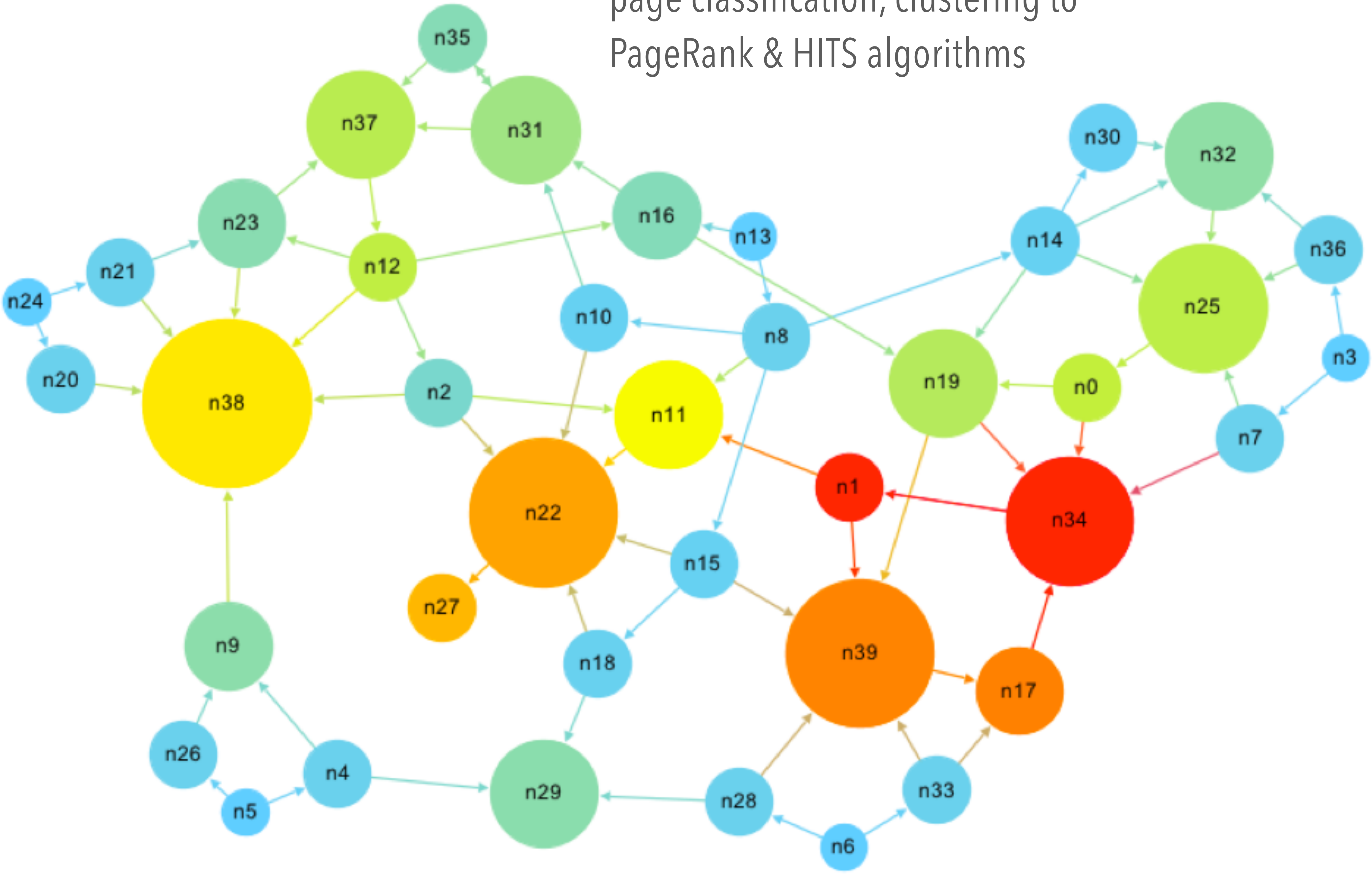


- Which view do you prefer?
  - KDD vs. ML/Stat. vs. Business Intelligence
  - Depending on the data, applications, and your focus
- Data Mining vs. Data Exploration
  - Business intelligence view
    - Warehouse, data cube, reporting but not much mining
  - Business objectives vs. data mining tools
  - Supply chain example: mining vs. OLAP vs. presentation tools
  - Data presentation vs. data exploration

APPLICATIONS OF DATA MINING

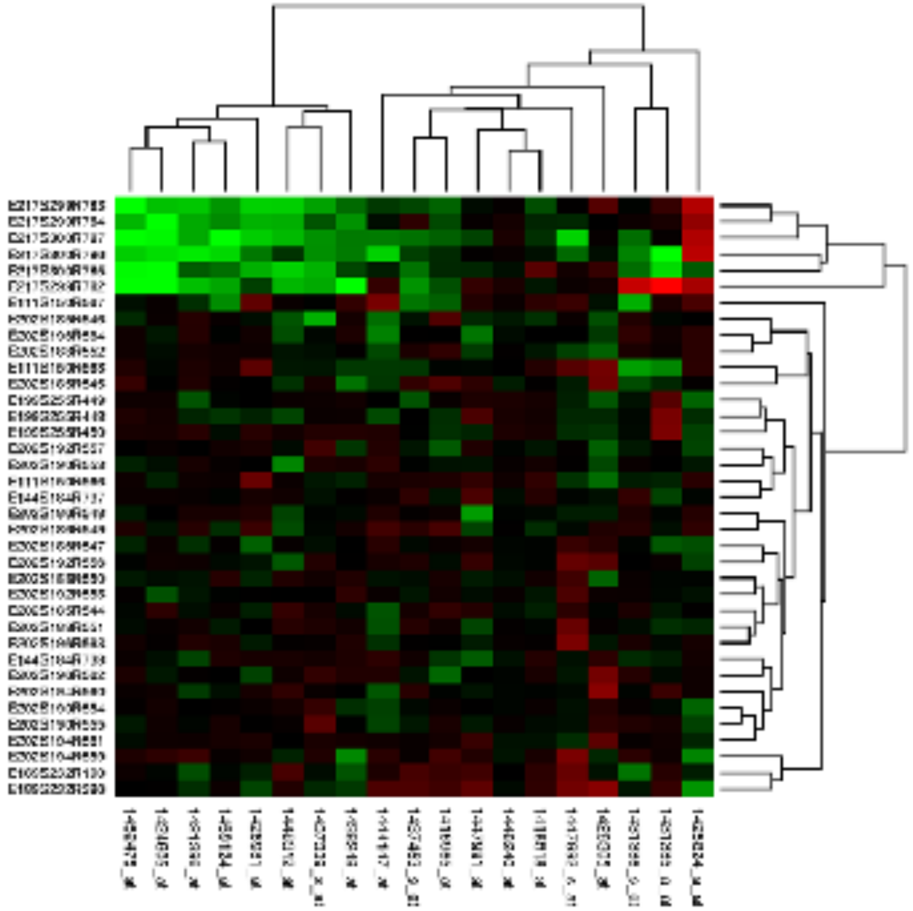


Web page analysis: from web page classification, clustering to PageRank & HITS algorithms



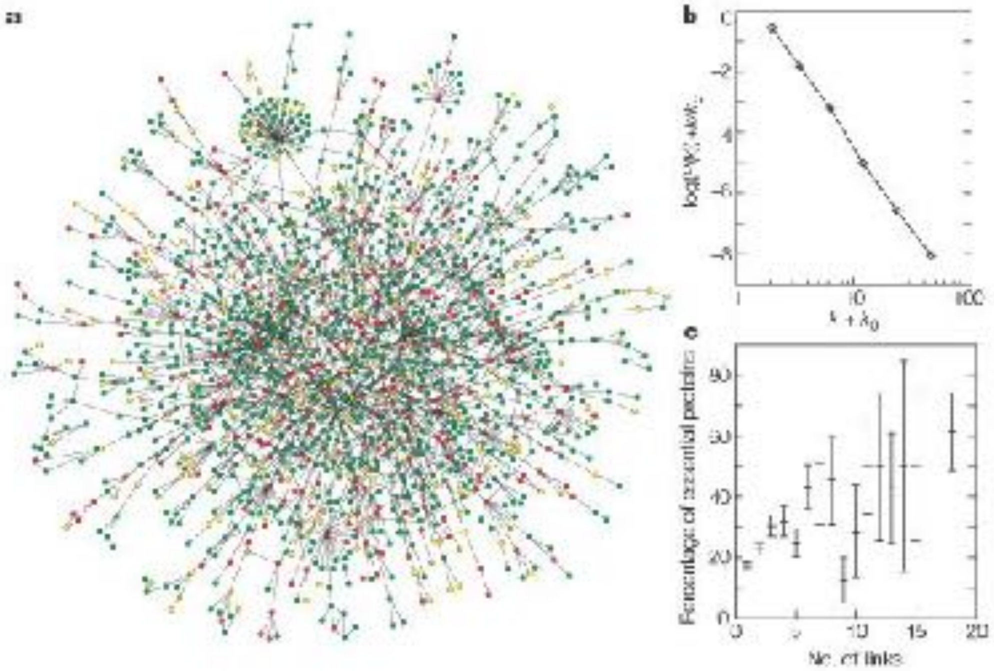
Collaborative analysis & recommender systems

Basket data analysis to targeted marketing



microarray

biological network



Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis source code

commits

fix bugs

execution traces

estimate costs

Software engineering and data mining

binaries

code optimization

code completion

SAS

dedicated data mining tools

Orcale data mining tools

MS SQL Server Tools

DATA MINING: A MULTI-DIMENSIONAL VIEW

The Data

Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

Data Mining Functions

Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.

Descriptive vs. predictive data mining

Multiple/integrated functions and mining at multiple levels

Techniques

Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

Applications

Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

WHAT KINDS OF DATA?

Database-oriented datasets and applications

Relational databases, data warehouse, transactional databases

Object-relational databases, Heterogeneous databases and legacy databases

Multimedia databases

Data streams
and sensor data

Time-series data,
temporal data, sequence
data (incl. bio-sequences)

Advanced datasets and advanced applications

Spatial data and
spatiotemporal
data

Structure data,
graphs, social
networks and
information networks

Text databases

The World-Wide Web

WHAT CAN WE DISCOVER WITH THIS DATA?

Generalization

Information integration and data warehouse construction

Data cleaning, transformation, integration, and multidimensional data model

Generalization

Scalable methods for computing (i.e., materializing) multidimensional aggregates

Data cube

OLAP (online analytical processing)

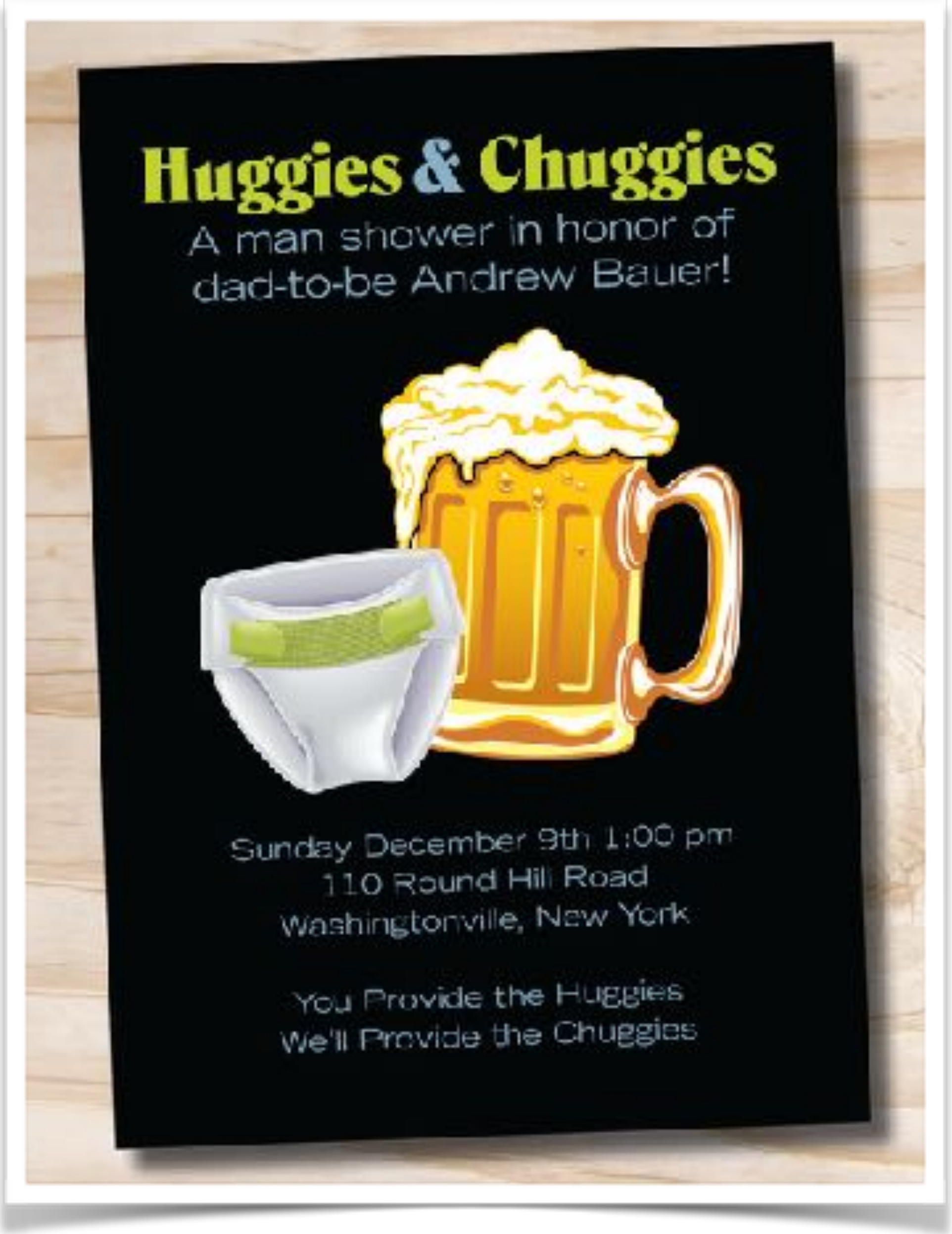
Generalization

Multidimensional concept description: characterization and discrimination

Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Association and Correlation Analysis

Frequent patterns



What items are frequently purchased together at the local grocery store?

Association and Correlation Analysis

Association, Correlation, and Causality

there are subtle differences between them!

An Association Rule

How to mine such patterns and rules efficiently in large datasets?

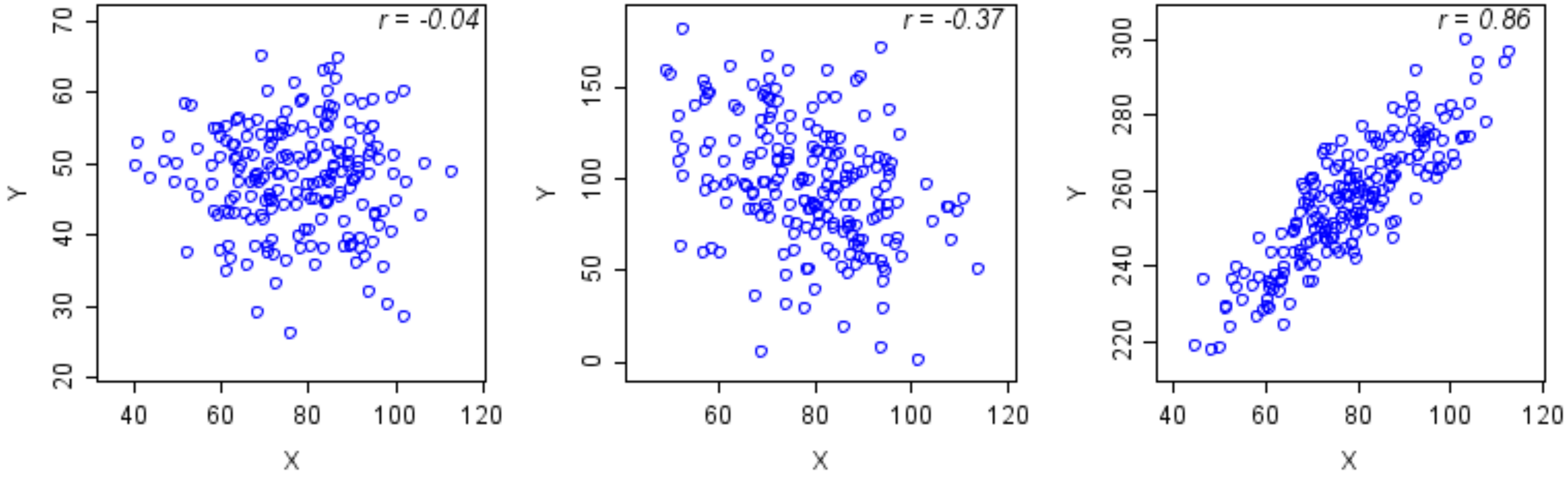
Diapers ⇒ Beer, [0.5%, 75%]

support confidence

How to use such patterns for classification, clustering, and other applications?

Association and Correlation Analysis

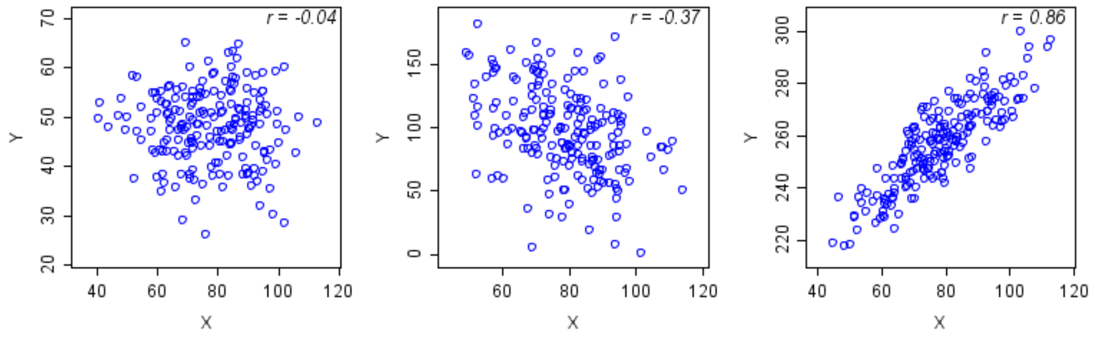
correlation measures the linear dependence between two numeric variables



Association and Correlation Analysis

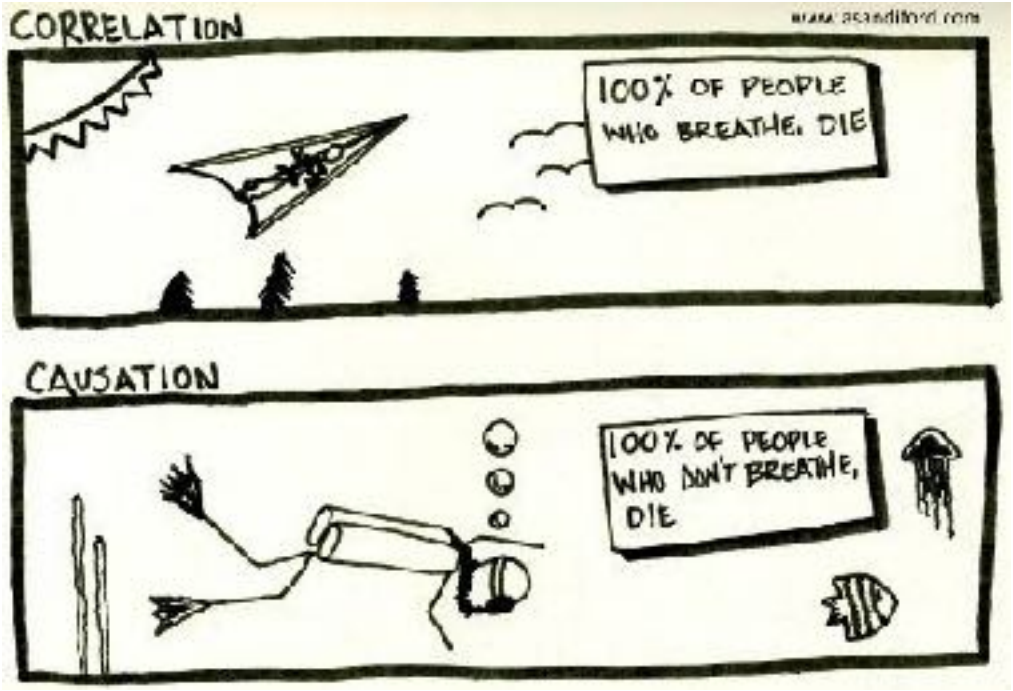
What is the difference between association and correlation?

Diapers ⇒ Beer, [0.5%, 75%]



Association and Correlation Analysis

correlation
≠
causation


CORRELATION
www.asandiford.com
100% OF PEOPLE WHO BREATHE, DIE
CAUSATION
100% OF PEOPLE WHO DON'T BREATHE, DIE

Classification

Classification and label prediction

Construct models (functions) based on some training examples

Predict some unknown class labels

Describe and distinguish classes or concepts for future prediction

climate, gas mileage

Classification

Typical methods

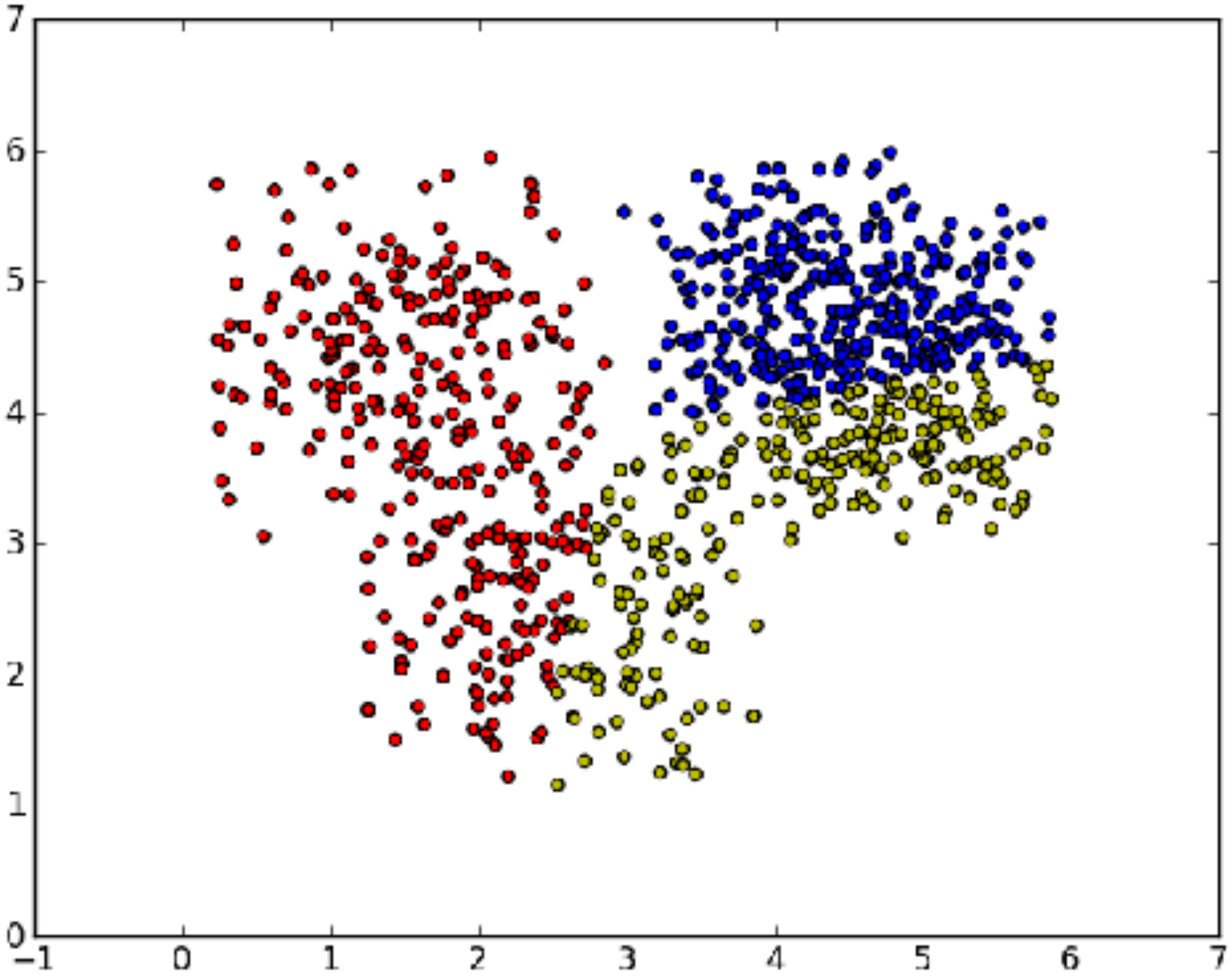
Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …

Classification

Typical applications

Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

CLUSTER ANALYSIS



Unsupervised learning (i.e., Class label is unknown)

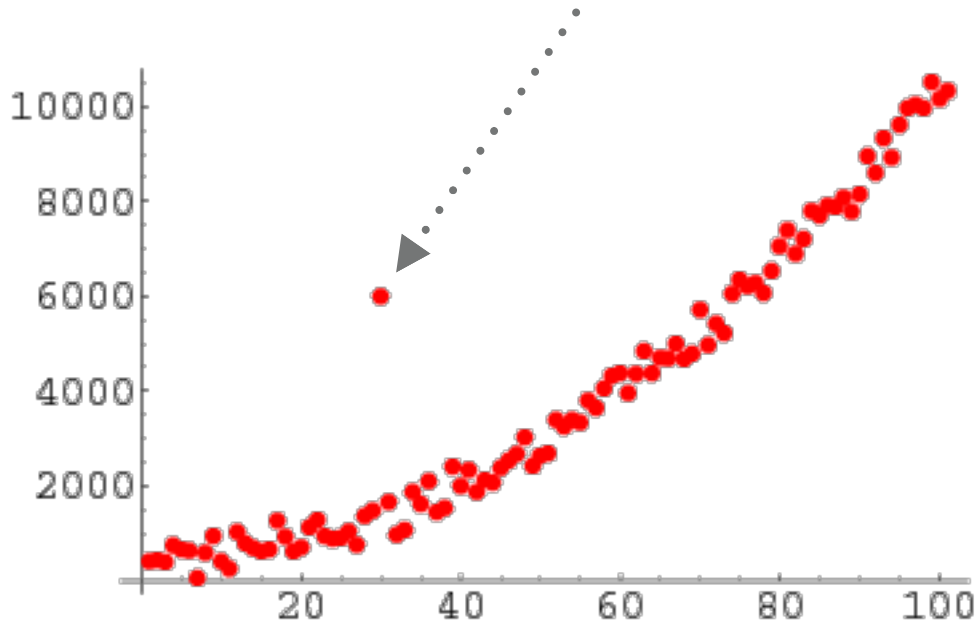
Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns

Principle: Maximizing intra-class similarity & minimizing interclass similarity

Many methods and applications

OUTLIER ANALYSIS

A data object that does not comply with the general behavior of the data

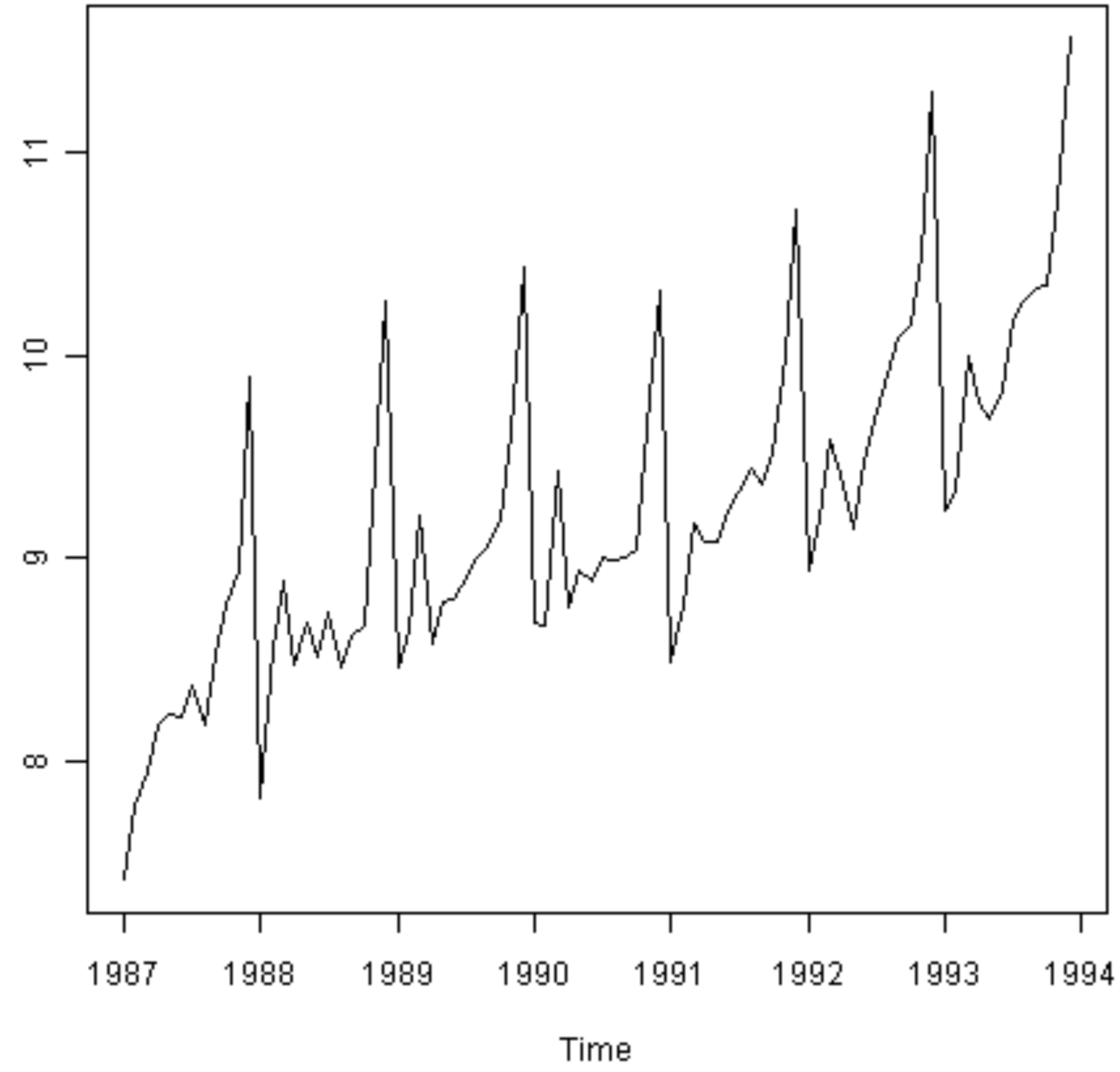


Noise or exception? — the value of the outlier is application dependent

Methods: byproduct of clustering or regression analysis, …

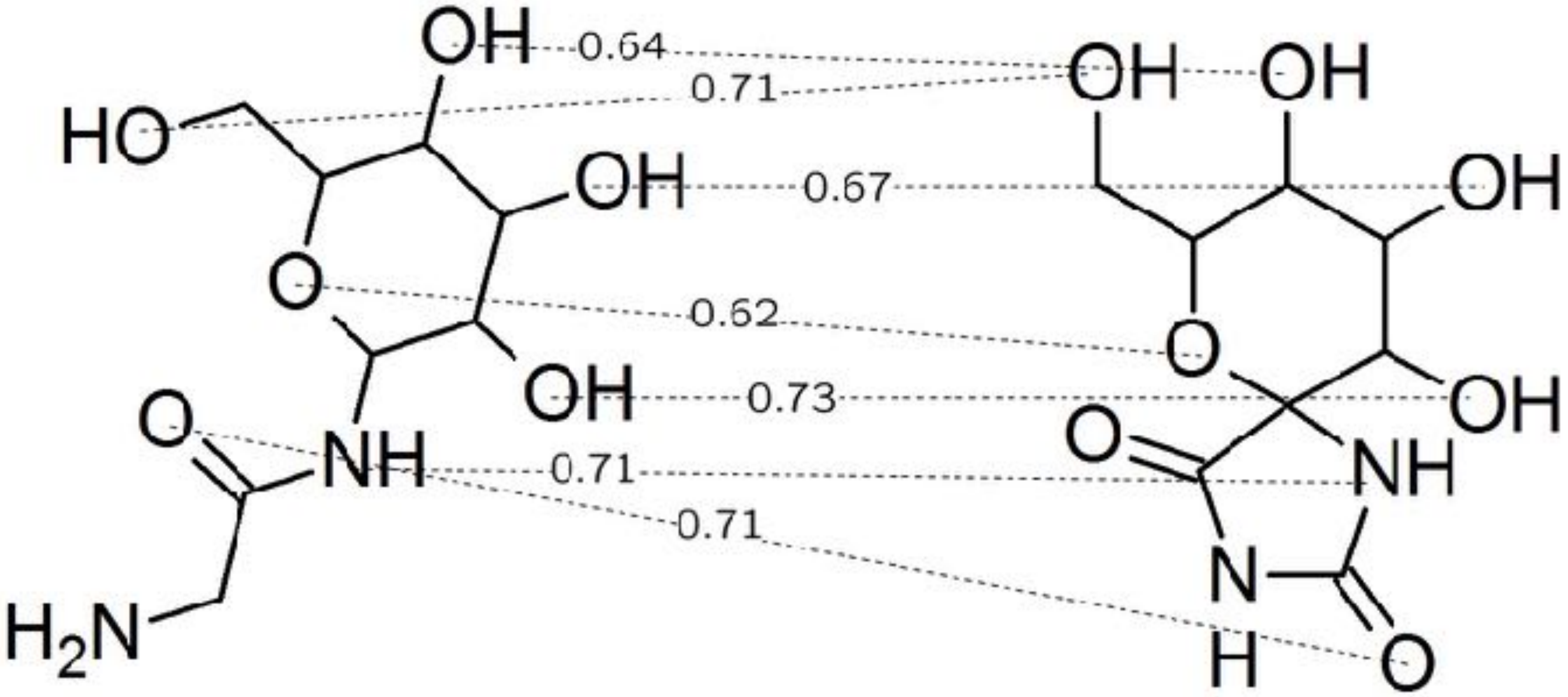
Useful in fraud detection, rare events analysis

SEQUENTIAL PATTERN, TREND AND EVOLUTION ANALYSIS

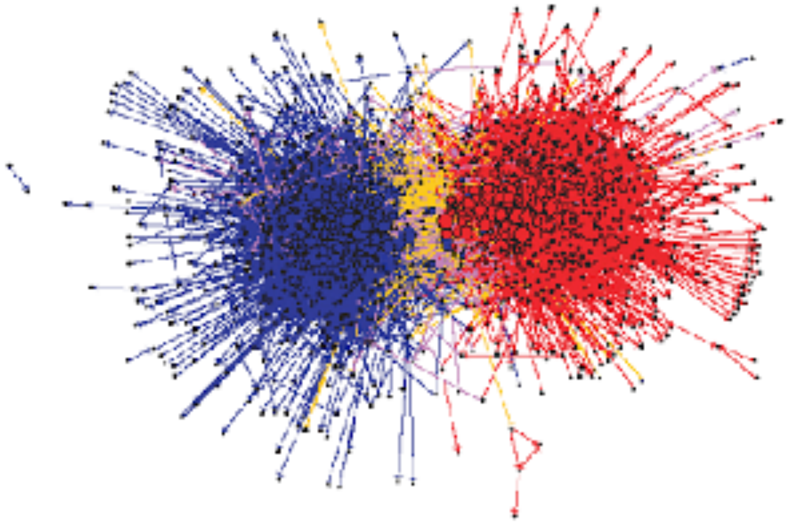


- Trend, time-series, and deviation analysis: e.g., regression and value prediction
- Sequential pattern mining
  - e.g., first buy a digital camera, then buy large memory cards
- Periodicity analysis
- Motifs and biological sequence analysis
  - Approximate and consecutive motifs
- Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

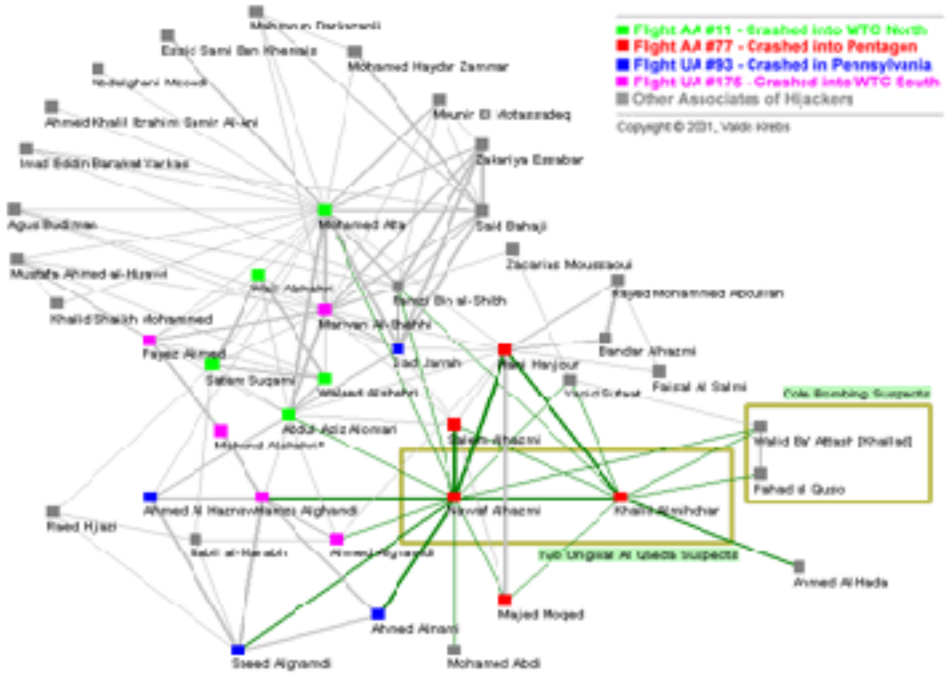
MINING GRAPHS



Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)



political blogs



terrorist networks

NETWORK ANALYSIS

Social networks: actors (objects, nodes) and relationships (edges)

e.g., author networks in CS, terrorist networks

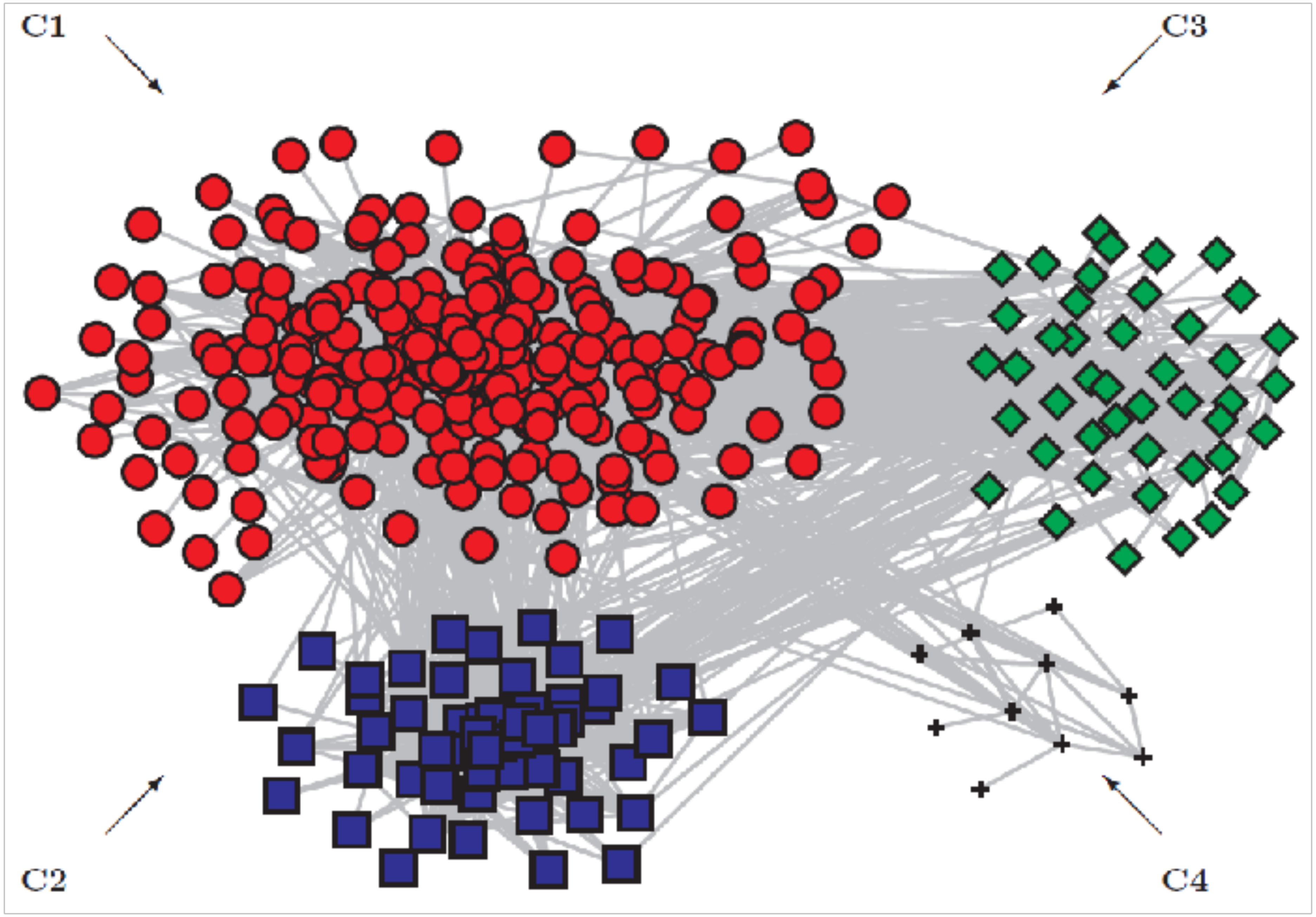
Multiple heterogeneous networks

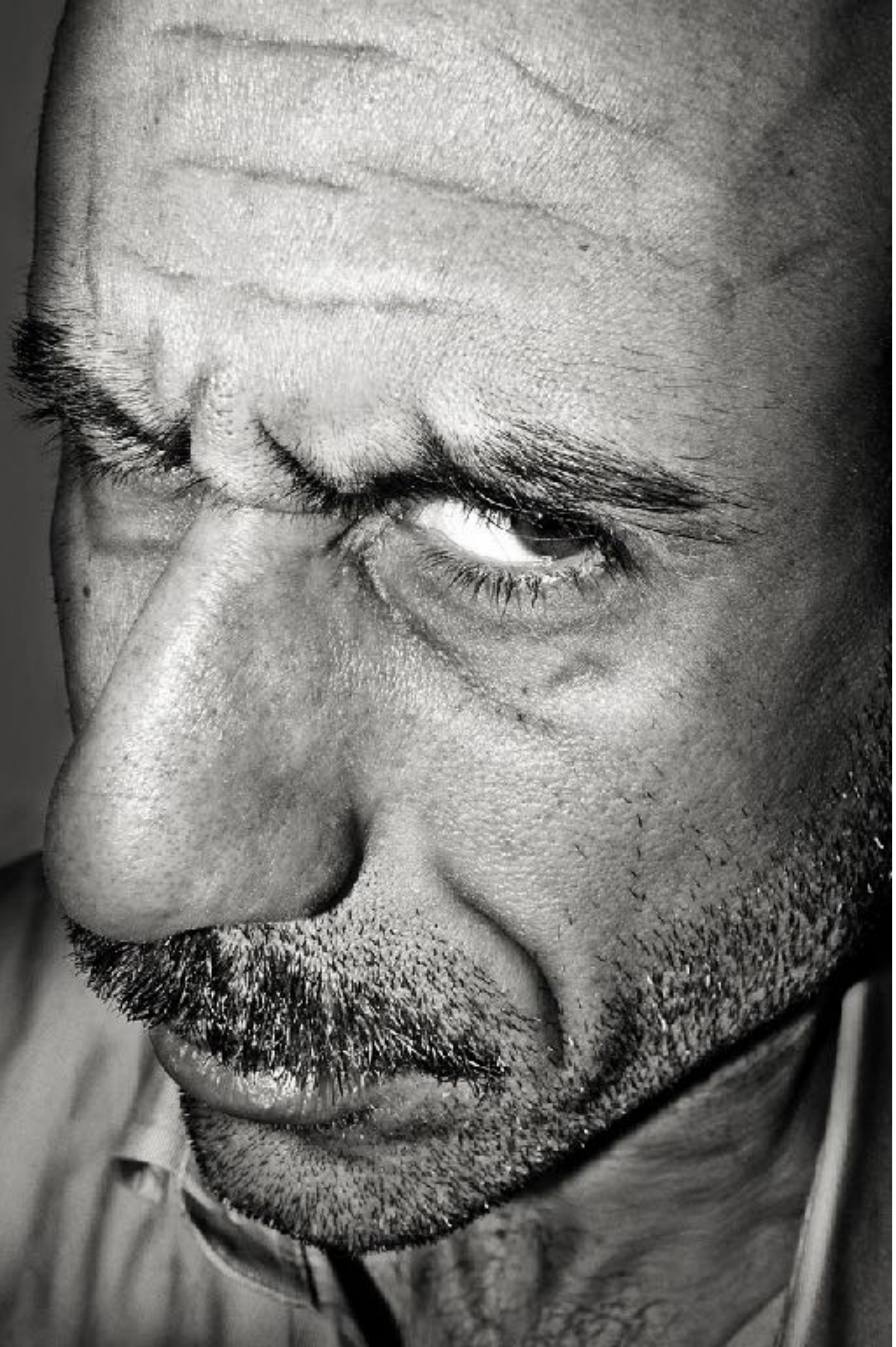
A person could be multiple information networks: friends, family, classmates, …

Links carry a lot of semantic information: Link mining

MINING THE WEB

Web community discovery, opinion mining, usage mining, …





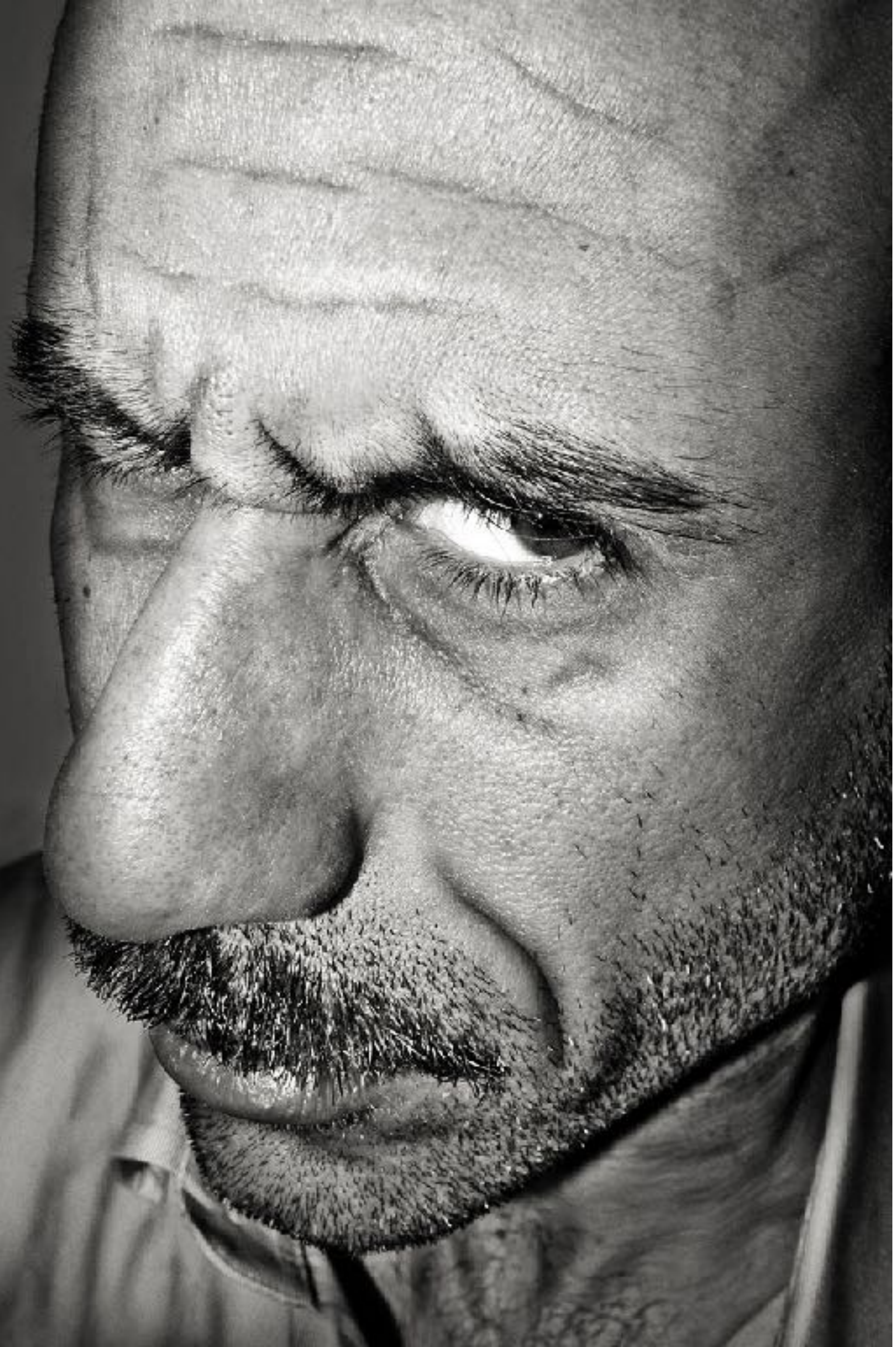
EVALUATION

Is all discovered knowledge interesting?

One can discover a very large number of "patterns"

Some may fit only certain dimension space (time, location, …)

May not be representative, be transient, …



EVALUATION

Why not discover <u>only</u> interesting knowledge?

Descriptive vs. predictive

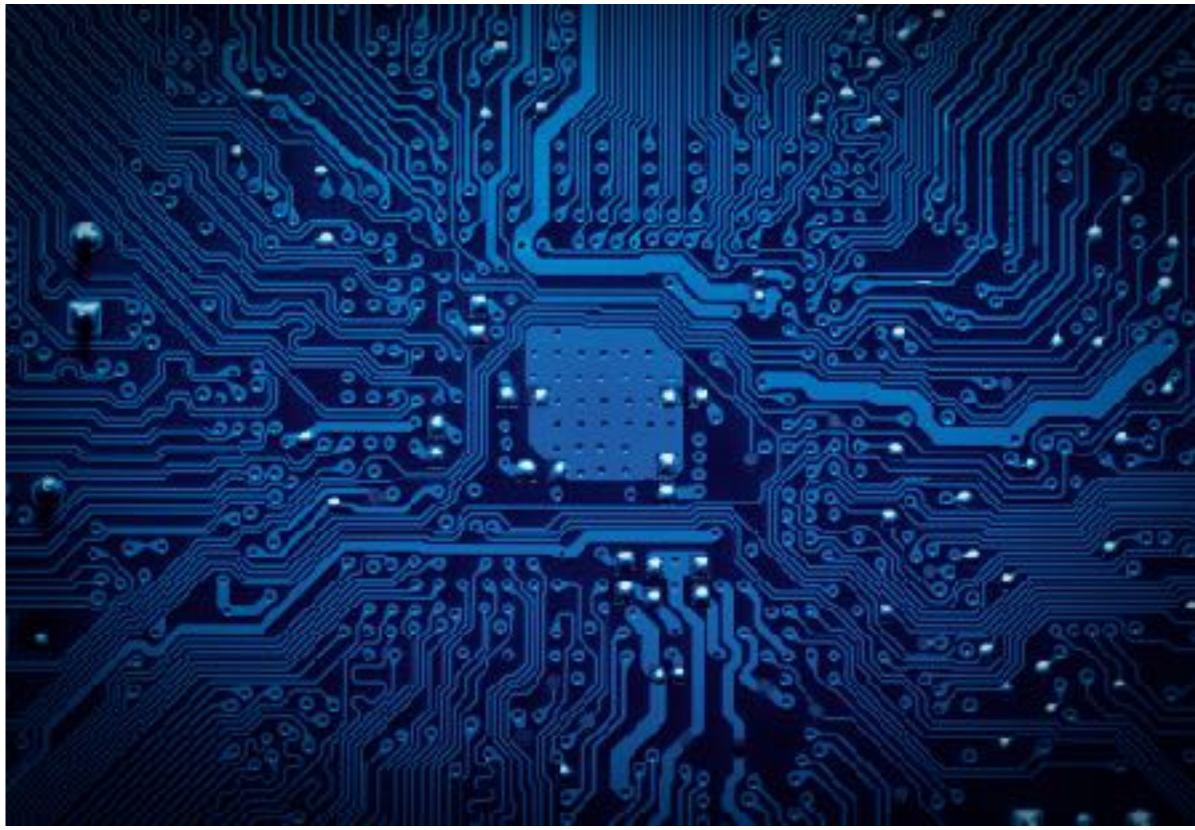
Coverage

Typicality vs. novelty

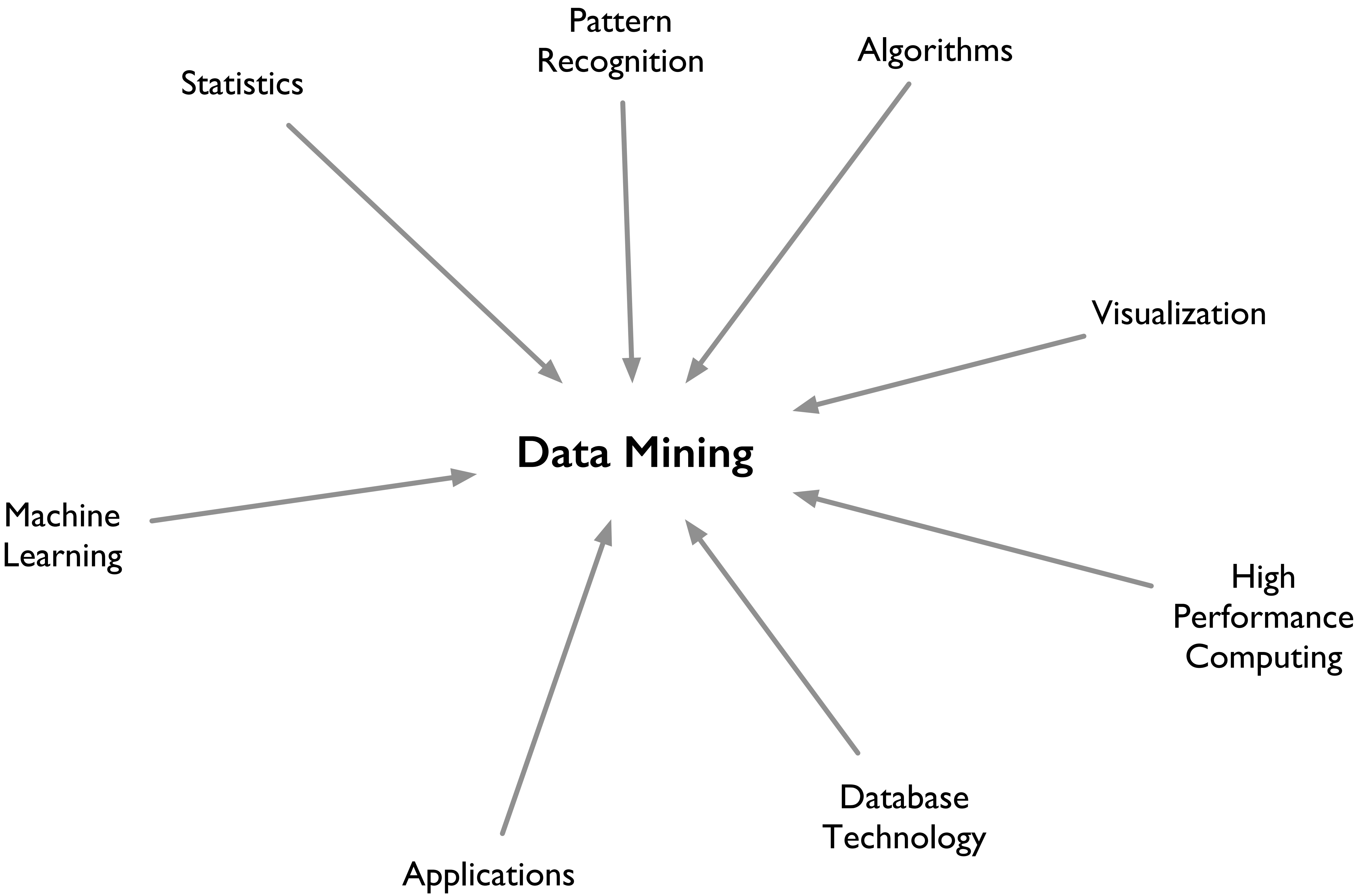
Accuracy

Timeliness

WHAT KINDS OF TECHNOLOGIES ARE USED?



A Confluence of Technologies



Statistics

Pattern Recognition

Algorithms

Visualization

Data Mining

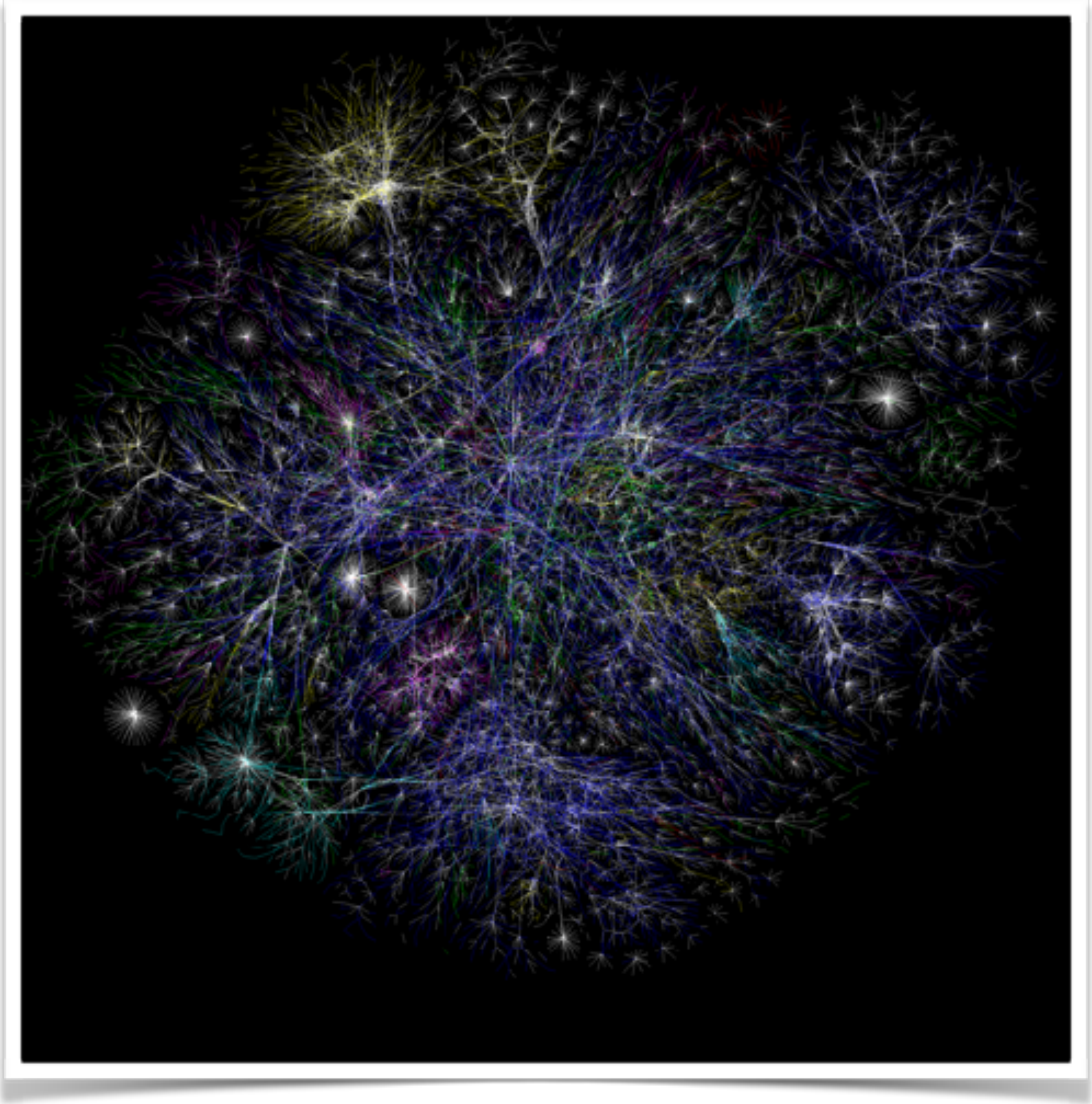
Machine Learning

High Performance Computing

Applications

Database Technology

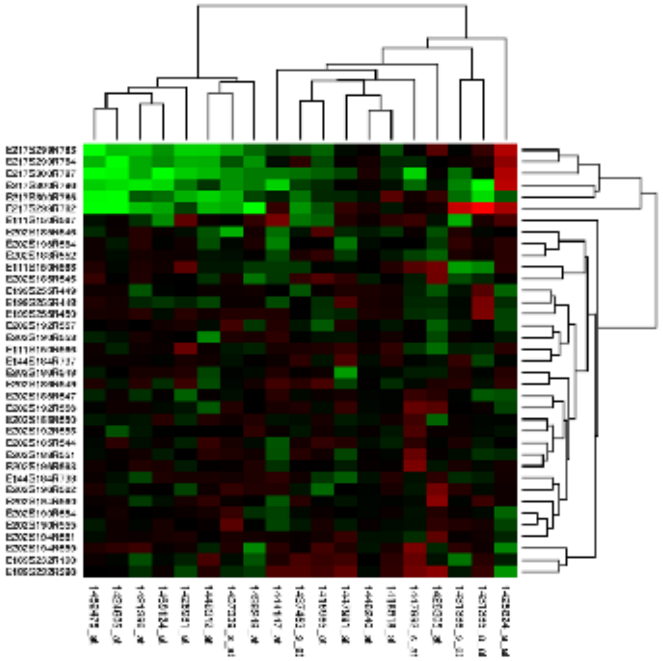
Why a confluence?



Massive

Algorithms must be scalable to handle big data

High dimensional



Micro-array may have tens of thousands of dimensions

Spatial, spatiotemporal, multimedia, text and Web data

New and sophisticated applications

Software programs, scientific simulations

Time-series data, temporal data, sequence data

Complex, Diverse

Structure data, graphs, social and information networks

Data streams and sensor data

MAJOR ISSUES IN DATA MINING

Mining various and new kinds of knowledge

Handling noise, uncertainty, and incompleteness of data

Mining knowledge in multi-dimensional space

Mining Methodology

Pattern evaluation and pattern- or constraint-guided mining

An interdisciplinary effort

Boosting the power of discovery in a networked environment

Interactive mining

Presentation and visualization of data mining results

User Interaction

Incorporation of background knowledge

Parallel, distributed, stream, and incremental mining methods

Efficiency and Scalability

Space and Time complexity of data mining algorithms

Mining dynamic,
networked, and global
data repositories

Data Type diversity

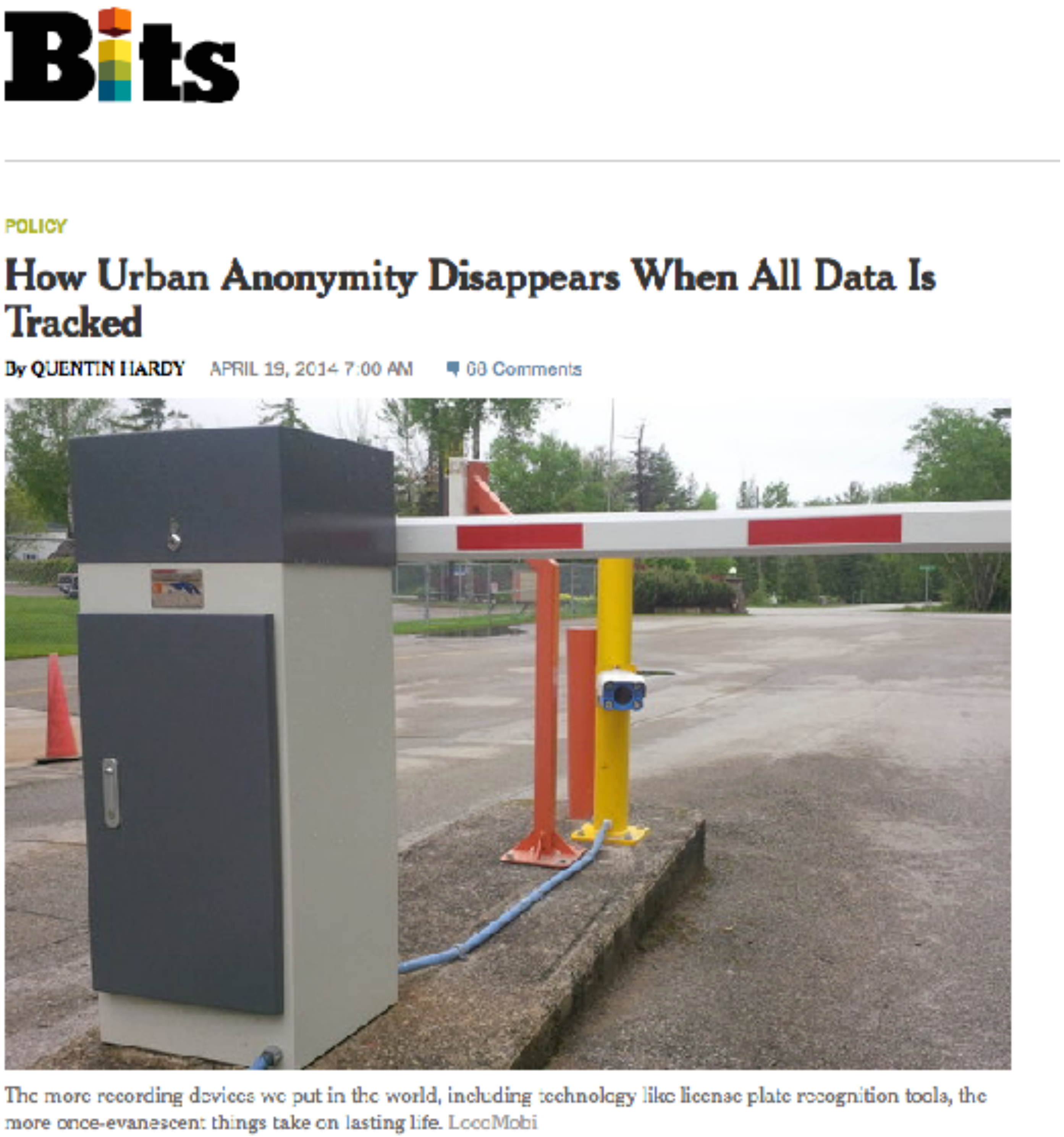
Handling complex data types

what is the social impact
of data mining?

Data mining and society

Invisible

Privacy-preserving



Bits

POLICY

How Urban Anonymity Disappears When All Data Is Tracked

By QUENTIN HARDY APRIL 19, 2014 7:00 AM 68 Comments

The more recording devices we put in the world, including technology like license plate recognition tools, the more once-evanescent things take on lasting life. LocoMobi

THE VERGE

Amazons Black Friday sale begins early with cheap Kindles, 4K TVs and game consoles

Why your favorite mobile games look a lot different in China

Bill Gates on The Verge: Can online classrooms help the developing world catch...

Will the internet of things finally kill privacy?

Why the FTC's new report doesn't go far enough

By [illegible]

HEADLINES

Facebook's new fundraiser tool makes it easier for nonprofits to raise money

The official Apple Watch magnetic charging dock is now on sale

Hasbro's new robotic cats will make you feel a little less alone, and a

SundayReview | OPINION

Facebook Is Using You

By LORI ANDREWS FEB. 4, 2012

Email

Share

Tweet

Save

More

LAST week, Facebook filed documents with the government that will allow it to sell shares of stock to the public. It is estimated to be worth at least $75 billion. But unlike other big-ticket corporations, it doesn't have an inventory of widgets or gadgets, cars or phones. Facebook's inventory consists of personal data — yours and mine.

Facebook makes money by selling ad space to companies that want to reach us. Advertisers choose key words or details — like relationship statu location, activities, favorite books and employment — and then Facebook runs the ads for the targeted subset of its 845 million users. If you indicate that you like cupcakes, live in a certain neighborhood and have invited friends over, expect an ad from a nearby bakery to appear on your page. The magnitude of online information Facebook has available about each o us for targeted marketing is stunning. In Europe, laws give people the righ to know what data companies have about them, but that is not the case in the United States.

Privacy is an important issue

Summary

Data mining: Discovering interesting patterns and knowledge from massive amount of data

A natural evolution of science and information technology, in great demand, with wide applications

Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.

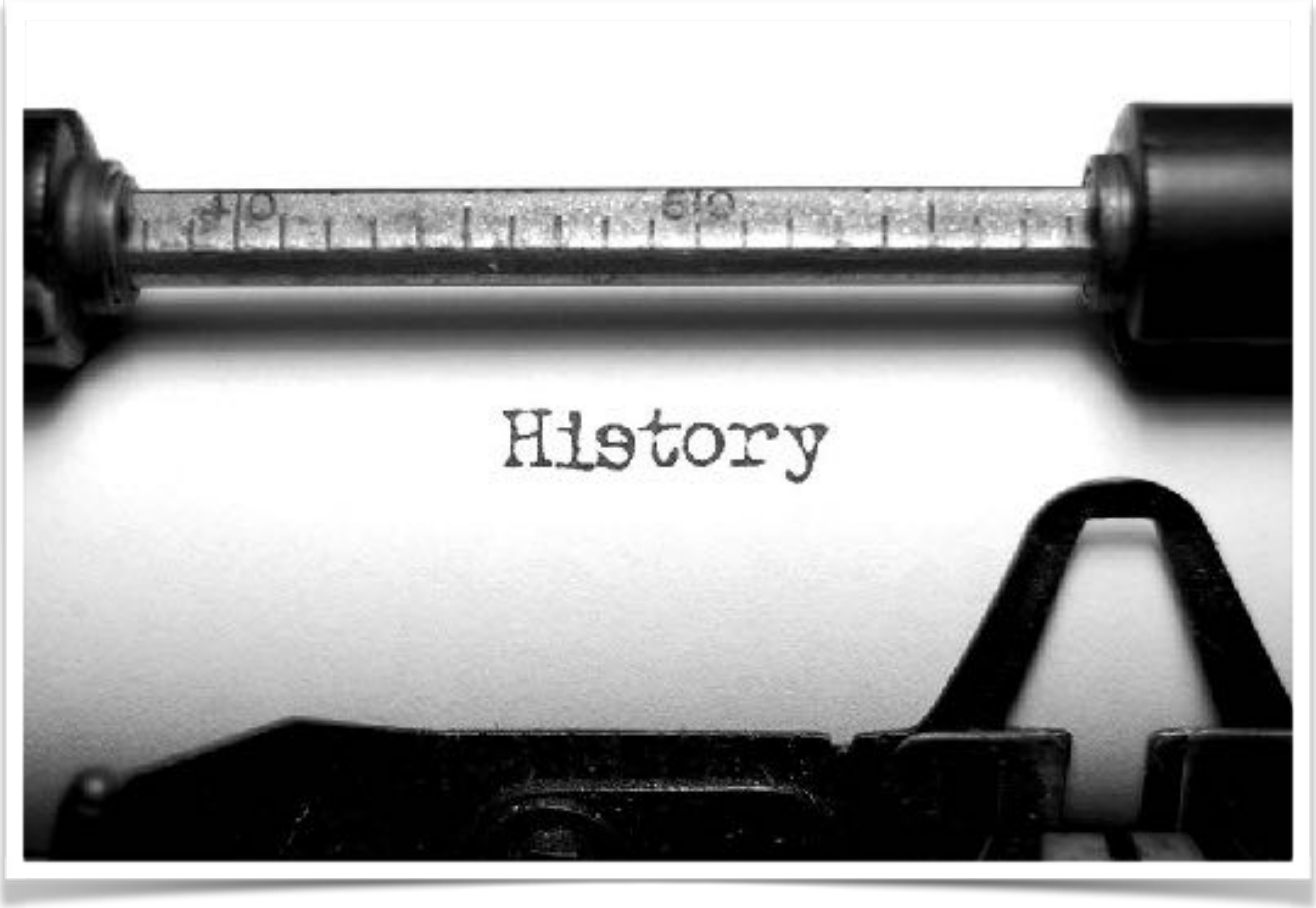
Major issues

A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

Mining can be performed on a variety of data

Data mining technologies and applications

A BRIEF HISTORY



- 1989 IJCAI Workshop on Knowledge Discovery in Databases
    - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
    - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
    - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
    - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

KDD CONFERENCES

KDD2014
This year's special theme
Data Science for Social Good
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 24-27, 2014
New York City

ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)

SIAM Data Mining Conf. (SDM)

(IEEE) Int. Conf. on Data Mining (ICDM)

European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)

Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

Int. Conf. on Web Search and Data Mining (WSDM)

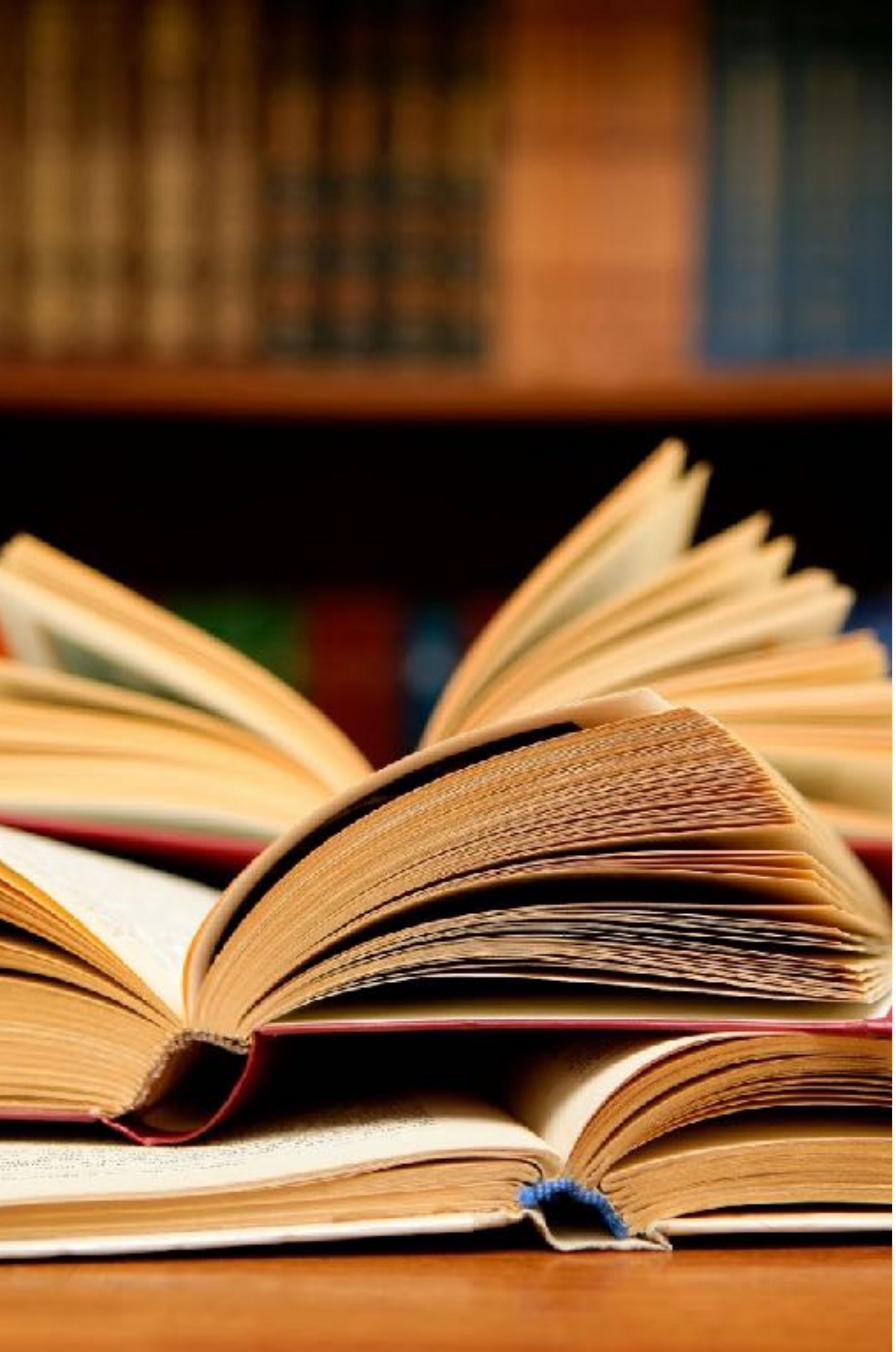
RELATED JOURNALS AND CONFERENCES

Journals:

- Data Mining and Knowledge Discovery (DAMI or DMKD)
- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- KDD Explorations
- ACM Trans. on KDD

Conferences:

- DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
- Web and IR conferences: WWW, SIGIR, WSDM
- ML conferences: ICML, NIPS
- PR conferences: CVPR, ICCV



REFERENCE BOOKS

E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011

J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011

T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009

B. Liu, Web Data Mining, Springer 2006

Y. Sun and J. Han, Mining Heterogeneous Information Networks, Morgan & Claypool, 2012

P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005

I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

T. M. Mitchell, Machine Learning, McGraw Hill, 1997

S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002

R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000

T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001

S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998