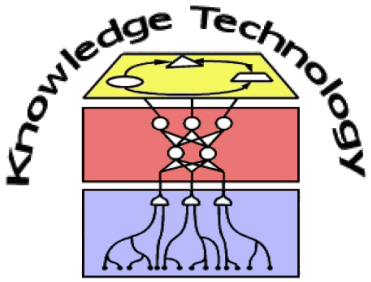
Data Mining

Introduction

Prof. Dr. Stefan Wermter



http://www.informatik.uni-hamburg.de/WTM/

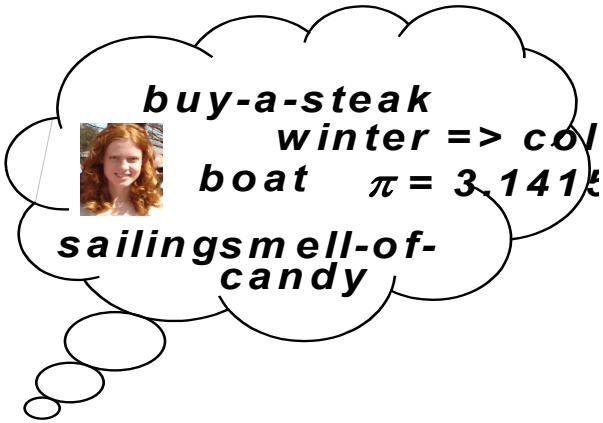
Why do YOU take this Module?…

A bit about us…

- Research Group Knowledge Technology (WTM)
  - Main research interest in Hybrid Neural/Symbolic Knowledge Technology
- Head: Prof. Stefan Wermter
  - Prior to Hamburg taught in Sunderland, Berkeley, Dortmund, Massachusetts
- Team for Data Mining:
  - Dr. Cornelius Weber
  - Doreen Jirak
  - Stefan Heinrich




What is Knowledge?



information and skills acquired by education and experience

π = 3,14159 26535 89793 23846

IF winter
THEN cold

marietta.jpg

http://best-steakhouse.com

information and processing methods acquired by programming and machine learning

Topics

- Theoretical and practical methods for data mining, knowledge management and assistive systems

- Pre-processing and visualization methods
- Knowledge management and associations rules
- Decision trees, decision rules
- Supervised classification and neural networks
- Unsupervised clustering and self-organizing neural networks
- Genetic algorithms and learning
- Fuzzy reasoning and neuro-fuzzy architectures
- Hybrid systems and ensemble learning

Organisational Issues

- Module Data Mining
- 4 SWS Lecture Data Mining
  - Wednesday 10-14, F-132, with a (lunch) break in between?
- 2 SWS Tutorials in Data Mining (Lab)
- Examinations: **written** (Klausur)!, 15. July and 29. September 2014
- The tutorials will contain exercises and practical assignments related to this lecture and must be attended

http://www.informatik.uni-hamburg.de/WTM/teaching/SoSe14_DataMining_V.shtml

Benefits of attending the lectures and labs

- Regular and effective learning of main concepts
- Discussions about provided methods and approaches as well as about emerging questions
- Access to video demonstrations and live demos in our lab
- Links to staff members and related research in our group
- Focus for examinations

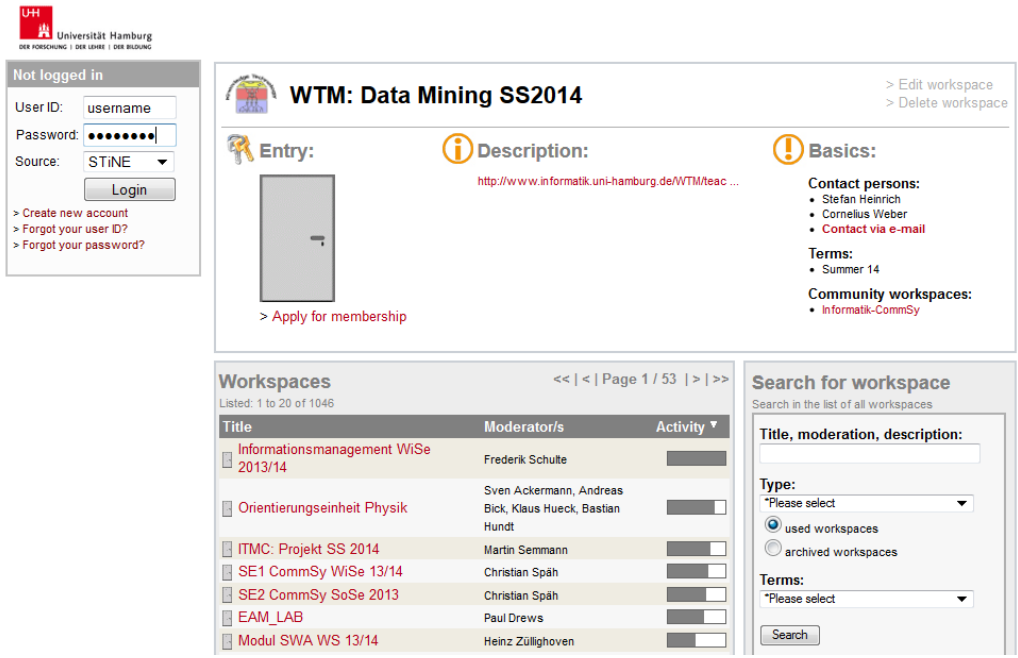
Lab / Tutorial – Data Mining

- Practical part of this module to ...
  - train some methods with exercises
  - test some KBS in an own implementation
- Participation is mandatory
  - Attend the meetings of your group of choice
  - Solve a programming practical exercise
  - Defend your solution in the end of every meeting
- Groups
  - Thursday 10-14 (Doreen Jirak, D-114)
  - Thursday 14-18 (Stefan Heinrich, D-114)
  - Thursday 14-18 (Doreen Jirak, D-118)
  - Friday 10-14 (Cornelius Weber, D-114)

every other week, starting 10./11.04.!

Communication: MIN-CommSy

- Our platform for:
  - Latest news
  - Teaching material
- Participation:
  - Visit the page
  - Apply for membership

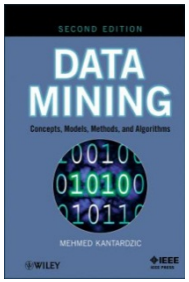


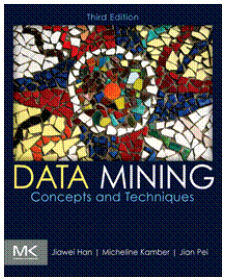
Website: https://www.mincommsy.uni-hamburg.de/commsy.php?cid=5930621

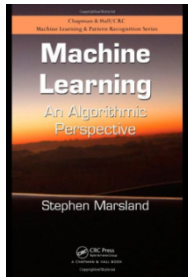
Or follow the link on our Website: http://www.informatik.uni-hamburg.de/WTM/teaching/

Literature and Acknowledgements

- Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: Concepts and techniques*. Morgan Kaufmann, 2011.
- Mehmed Kantardzic. *Data Mining*. Wiley, 2011.
  main text book
- Stephen Marsland. *Machine Learning - An Algorithmic Perspective.* CRC Press, 2009.







- **http://www.informatik.uni-hamburg.de/bib/vib/index.shtml.de**

- Thanks to slides by J. Han and M.Kamber, S. Marsland, and M. Kantardzic.
  - Slides follow mainly textbook of Kantardzic
  - Additional Slides from Shane Warren and Brittney Ballard

Other Optional Literature and Links

- "Economic Modeling Using Artificial Intelligence Methods / by Tshilidzi Marwala " 2013 Springer E-Book-Paket *Computer Science* via Campus-Katalog : https://kataloge.uni-hamburg.de/DB=1/XMLPRS=N/PPN?PPN=744996422

A few word about English…

A service for your future professional life

Questions About Data Mining

- What is data mining?
- Why data mining? Motivation and benefits?
- What kind of data to mine?
- When to mine the data?
- How to organize the mining process?
- What are the challenges in data mining?

Definitions in Dictionaries

- Data Mining is the practice of searching through large amounts of computerized data to find useful patterns or trends (Merriam Webster Dictionary)

- Data Mining is the process of identifying ***valid***, ***novel***, potentially ***useful***, and ultimately ***comprehensible*** knowledge from databases that is used to make crucial business decisions (G. Piatetsky-Shapiro)

Why Data Mining?

Trends Leading to Data Flood:

- Bank, telecom, other business transactions ...
- Scientific data: astronomy, biology, etc.
- Web, text, and e-commerce

Petabytes of Data?

- MEDLINE text database
  - 12 million published articles
- Google
  - More than 1 billion search requests per day
- CALTRANS loop sensor data
  - Every 30 sec., thousands of sensors, 2 GB per second
- NASA MODIS satellite
  - Coverage: 250m resolution, 37 bands, whole earth, every day
- Walmart transaction data
  - Order of 100 million transactions per day

Zetabytes of Data?

COMPUTERWOCHE

Meet the IBM EXPERTS Diskutieren Sie mit!

Technologie | Management | Karriere | Mittelstand | Whitepaper | Events &

Big Data

Hintergrund | Ratgeber | Bilder | Video | News

ZETTABYTE-BARRIERE GEKNACKT

Big Data - die Datenflut steigt

17.08.2012 | von Martin Bayer

Die explodierenden Datenmengen werden für Unternehmen zu einem ernsthaften Problem. Wer die Kontrolle behalten und möglichst viel Nutzen aus den Informationen ziehen will, muss die gesamte IT-Infrastruktur hinterfragen.

XING

Share

Die Datenflut steigt.
Foto: fotolia.com/ktsdesign

Die Information ist das Öl des 21. Jahrhunderts, und Analytics der Verbrennungsmotor, der damit läuft" - Peter Sondergaard, Senior Vice President von Gartner, bemühte eine Metapher, um die Herausforderung deutlich zu machen. Den Rohstoff Information aus gewaltigen Datenmengen zu extrahieren und zu verarbeiten sei eine der künftigen Kernaufgaben für Unternehmen.

Why Data Mining Now?

- Data Explosion causes Data Wasting:
  - Only a small portion (5% - 10%) of the collected data is ever analyzed.
  - Data that may be never analyzed continues to be collected at great expenses.

WE ARE DROWNING IN DATA,
BUT STARVING FOR KNOWLEDGE!

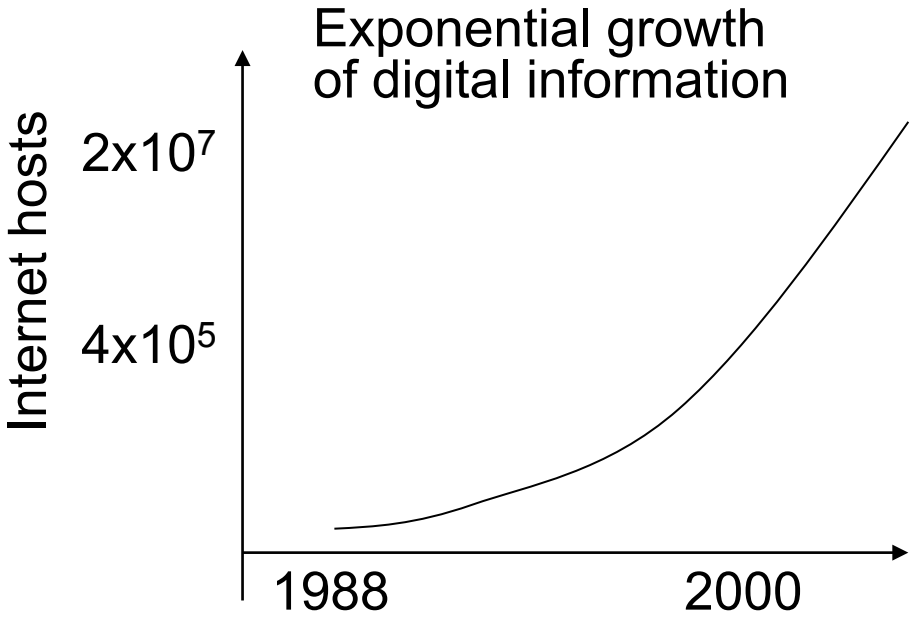
Where is the knowledge we have lost in information?

—T. S. Eliot, *The Rock*

Why Data Mining Now? (cont.)

- Sources of data overload:
  - Distributed data sources
  - Remote sensing
  - Internet
  - Multimedia data

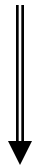
….



- ***A Gap*** between
  - data collection and organization capabilities, and
  - analysis and extraction of useful information for decision processes.

Stories: Managers believe …

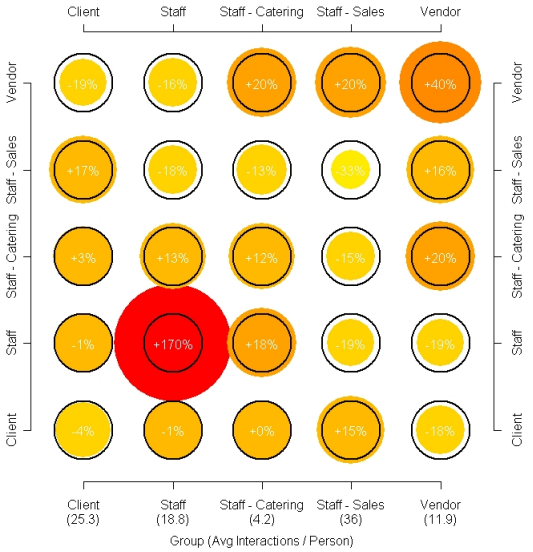
- 61% believe that information overload is present in their workplace.
- 80% believe the situation will get worse.
- 50% ignore large data sets in current decision process.
- 84% store the data for future with current use or analysis.



60% believe that the cost of gathering information outweights its value!

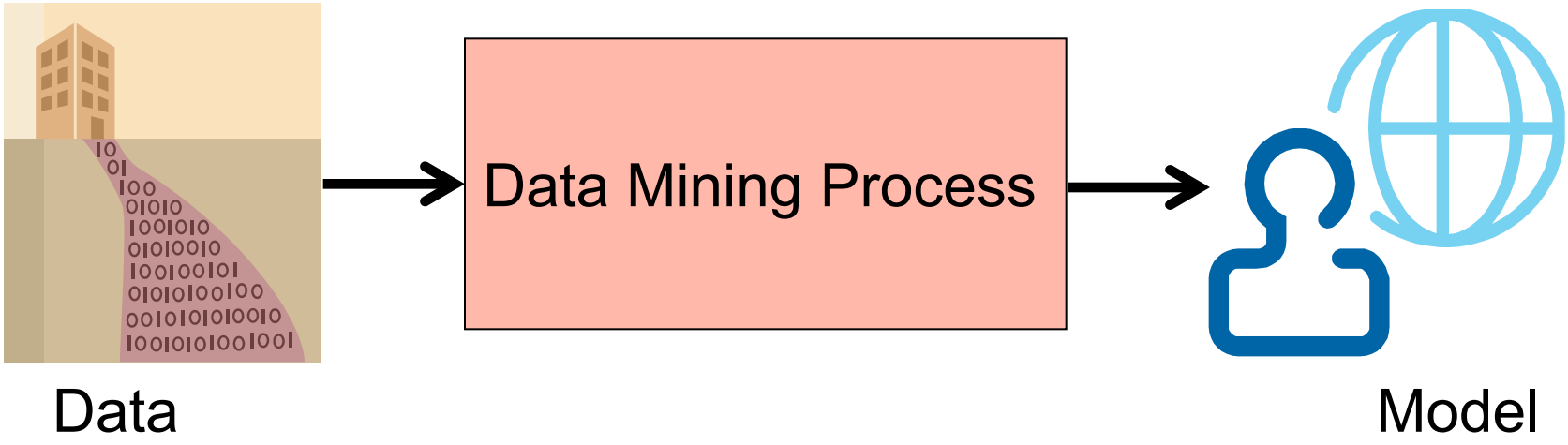
Data Mining is NOT…

- Brute-force crunching of bulk data
- “Blind” application of algorithms
- Going to find relationships where none exists
- Presenting data in different ways (visualization)
- A database intensive task (SQL)



What is Data Mining?

- In many domains there is a shift from classical modeling and analyses based on ***first principle*** to developing models and corresponding analyses ***directly from data***.



Data Mining is ***a process*** for the ***automatic extraction*** of non-obvious, hidden ***knowledge*** from ***large volumes of data***.

What Is Data Mining? (cont.)

- Potential point of confusion:
  - The “extracting of ore from” rock metaphor ***does not*** really apply to the practice of data mining
  - If it did, then standard database queries would fit under the rubric of data mining

- In practice, DM refers to:
  - ***Finding patterns/models*** across large databases
  - ***Discovering*** unknown ***information and knowledge***

From Data to Knowledge

Medical Data by Dr. X, Tokyo Med. & Dent. Univ., 38:

10, M, 0, 10, 10, 0, 0, 0, SUBACUTE, 37, 2, 1, 0,15,-,-, 6000, 2, 0, abnormal, abnormal,-,
2852, 2148, 712, 97, 49, F,-,multiple,,2137, negative, n, n, ABSCESS,***VIRUS***

12, M, 0, 5, 5, 0, 0, 0, ACUTE, 38.5, 2, 1, 0,15, -,-, 10700,4,0,normal, abnormal, +, 1080,
680, 400, 71, 59, F,-,ABPC+CZX,, 70, negative, n, n, n, BACTERIA***, BACTERIA***

15, M, 0, 3, 2, 3, 0, 0, ACUTE, 39.3, 3, 1, 0,15, -, -, 6000, 0,0, normal, abnormal, +, 1124,
622, 502, 47, 63, F, -,FMOX+AMK, , 48, negative, n, n, n, BACTE(E), ***BACTERIA***

16, M, 0, 32, 32, 0, 0, 0, SUBACUTE, 38, 2, 0, 0, 15, -, +, 12600, 4, 0,abnormal,
abnormal, +, 41, 39, 2, 44, 57, F, -, ABPC+CZX, ?, ? ,negative, ?, n, n, ABSCESS,***VIRUS***

Numerical attribute Categorical attribute Missing values Class labels

IF cell_poly <= 220 AND Risk = n AND Loc_dat = + AND Nausea > 15
THEN Prediction = VIRUS [87,5%]

Predictive accuracy

Possible Business Discoveries

Cus-tomer-ID	Account Type	Margin Account	Transac-tion Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30-39	Tennis	40-59k
1013	Custodial	No	Broker	0.5	F	50-59	Skiing	80-99k
1245	Joint	No	Online	3.6	M	20-29	Golf	20-39k
2110	Individual	Yes	Broker	22.3	M	30-39	Fishing	40-59k
1001	Individual	Yes	Online	5.0	M	40-49	Golf	60-79k

Acme Investors Incorporated

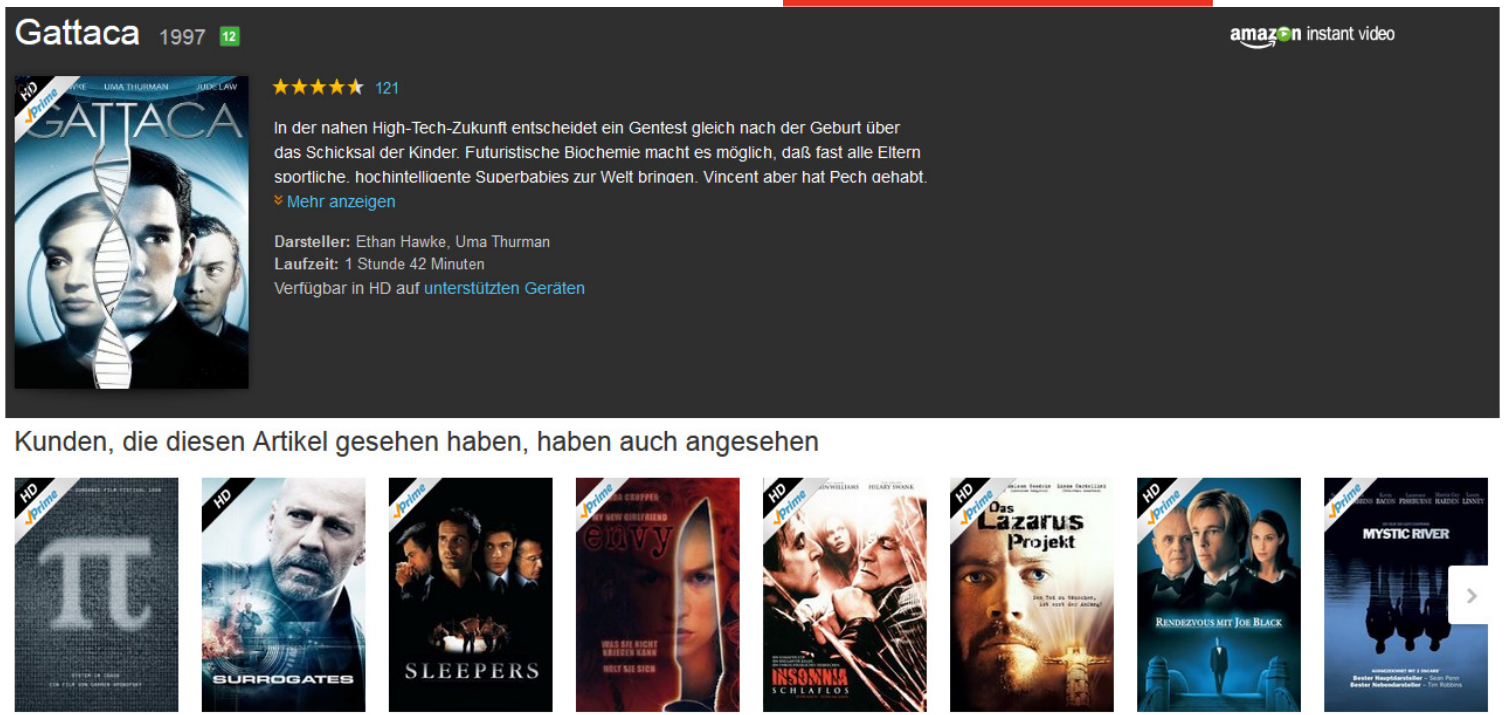
- Can I develop a general characterisation/profile of different investor types? (***classification***)
- What characteristics distinguish between Online and Broker investors? (***discrimination***)
- Can I develop a model which will predict the average trades/month for a new investor? (***prediction***)

Data Mining: An application

- Movies recommendations in online video libraries

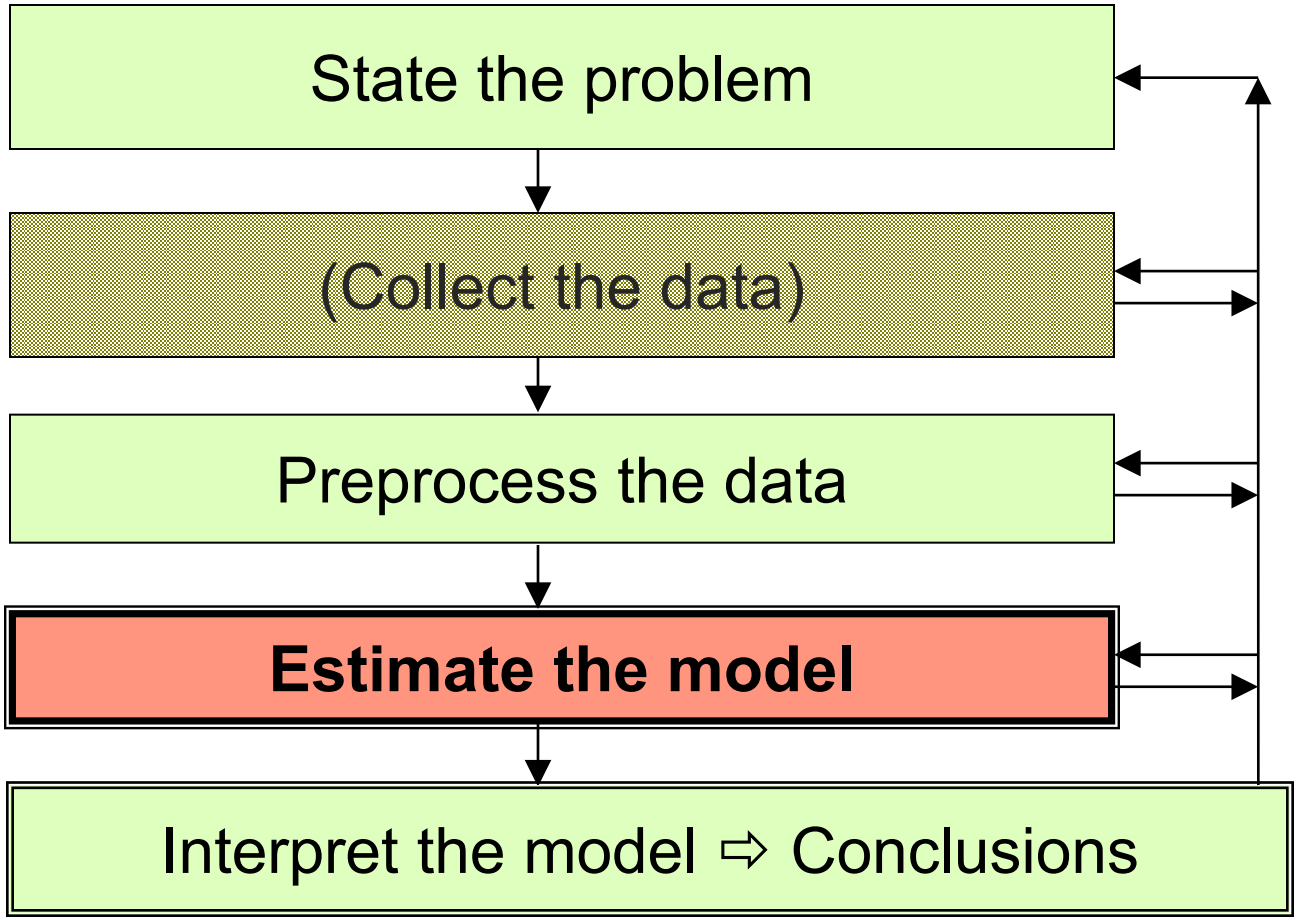
amazon.de Prime

NETFLIX



Data Mining – What‘s in a Name?

Data Mining as a simplified Process



Up and Down …

- Data mining is an ***iterative*** and ***interactive*** process:

  Be prepared to generate *"potentially right or wrong hypotheses"* before you arrive at actionable, ***meaningful*** and useful knowledge

State the Data Mining Problem

- There must be a well-defined problem
- The data must be available
- The data must be relevant, adequate and clean
- The problem should not be solvable by means of ordinary query, OLAP or other tools only

The results must be actionable.

Characteristics of Raw Data

- Missing data
- Misrecorded data
- Data may be from another population (heterogeneous)
- Different structures & formats
- With or without compression
- Redundant data
- With implicit temporal & spatial components

...

Why Data Preprocessing?

- Data in the real world is ***messy***:
  - Incomplete/missing: lacking attribute values
    - e.g., `occupation=""`
  - Noisy: containing erroneous outliers
    - e.g., `Salary="-10"`
  - Inconsistent: containing discrepancies in codes or names
    - e.g., `Age="42" Birthday="03/07/1997"`
    - e.g., Was rating `"1,2,3"`, now rating `"A, B, C"`
    - e.g., discrepancy between duplicate records

Data Mining Techniques

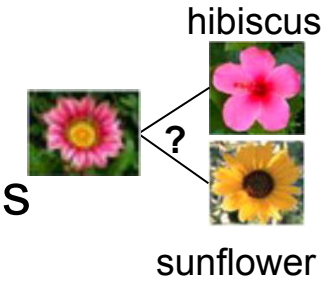
Algorithms for ***preprocessing***:

Raw Data = Messy Data

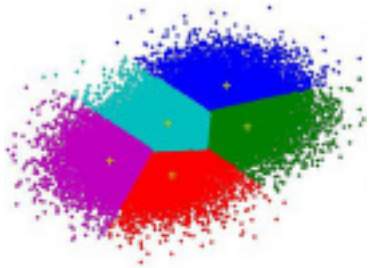
- Scaling & Normalization
- Encoding
- Outlier Detection & Removal
- Feature Selection & Composition
- Data Cleansing & Scrubbing
- Data Smoothing
- Missing Data Elimination
- Sampling

Primary Tasks of Data Mining I

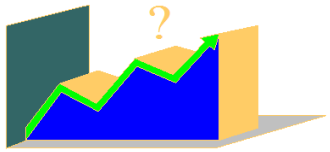
- ***Classification***:
  - Find the description of several predefined classes
  - Classify a data item into one



- ***Clustering***:
  - Identify a finite set of categories
  - … or clusters to describe the data

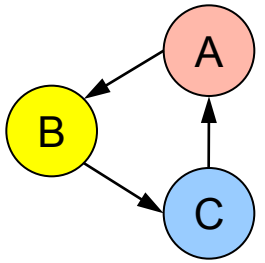


- ***Regression***:
  - Maps a data item to a real-valued prediction variable



Primary Tasks of Data Mining II

- ***Dependency modeling***:
  - Find a model that describes significant dependencies between variables



- ***Deviation*** and ***change detection***:
  - Discover the most significant changes in the data

- ***Summarization***:
  - Find a compact description for a subset of data

Challenges for Data Mining

- Technical
  - Tera-bytes and Peta-bytes and Zeta-bytes…
  - Complex, multi-media, unstructured data
  - Integration with domain knowledge
- Business
  - Finding good application areas
- Societal
  - Privacy & security issues

Data, Information, Knowledge? (Websters)

- ***Data***: Factual information (e.g. measurements or statistics) used as a basis for calculation, discussion or reasoning.
- ***Information***:
  - Communication or reception of knowledge
  - Obtained from investigation, study or instruction
- ***Knowledge***:
  - Understanding gained by actual experience
  - Awareness of information
  - Perception of truth
  - Something learnt and kept in mind

  - As we see the terms are sometimes defined “overlapping”

Example: Kyoto Traffic Simulator

- Modeling various driving behaviors and by varying the mixture ratio of the driver models

http://domino.research.ibm.com/library/cyberdig.nsf/0/d69a62935563ec2d852575450003ada3?OpenDocument

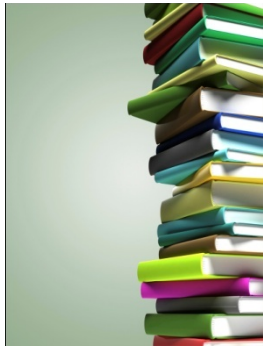
A semiotic view of Data, Information, Knowledge

- ***Data***:
  - Syntactic phenomena, e.g. numbers, bitcodes
- ***Information***:
  - Contains syntax and semantics (form and content).
  - E.g. HH-AB 694
- ***Knowledge***:
  - Contains a pragmatic version in addition to syntax and semantics.
  - Linked to usage or a **purpose**.
  - Functional relationships and associations between information or data

Knowledge in Humans

- ***Tacit implicit knowledge***:
  - difficult to communicate
  - stored in the brain
  - embodied knowledge
  - difficult to formalize

- ***Explicit knowledge***:
  - can be communicated
  - can be formalised at different levels of abstraction
  - can be stored in different media
  - often disembodied knowledge



Knowledge in Organizations

- **Knowledge and know-how** of employees are vital for economical success of an organization.
- Methods for **preserving, enhancing and communicating** knowledge are in high demand.
- **Formalizing human knowledge** is the main topic of "Knowledge Management in Organizations“

- This leads to Knowledge discovery based on data mining…

Knowledge-based Systems

Systems which exploit knowledge (in analogy to human knowledge) for problem solving

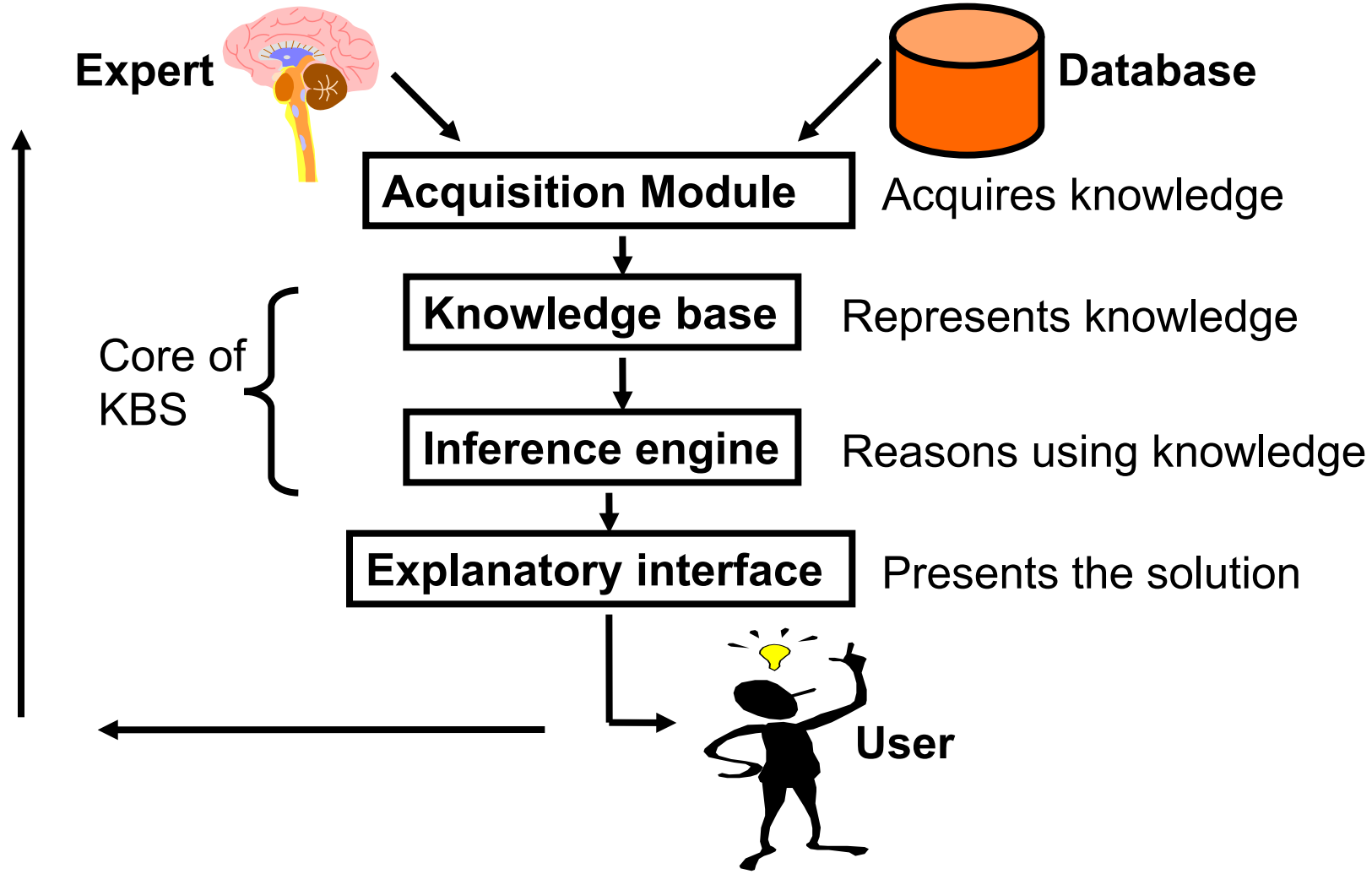
Examples:

- Public traffic information systems
  - knowledge of timetable
  - search of best connection
- Knowledge discovery system
  - facts and rules
  - association of facts

Content and organisation of system knowledge may be different from human knowledge

System knowledge processing methods may be different from human knowledge processing methods

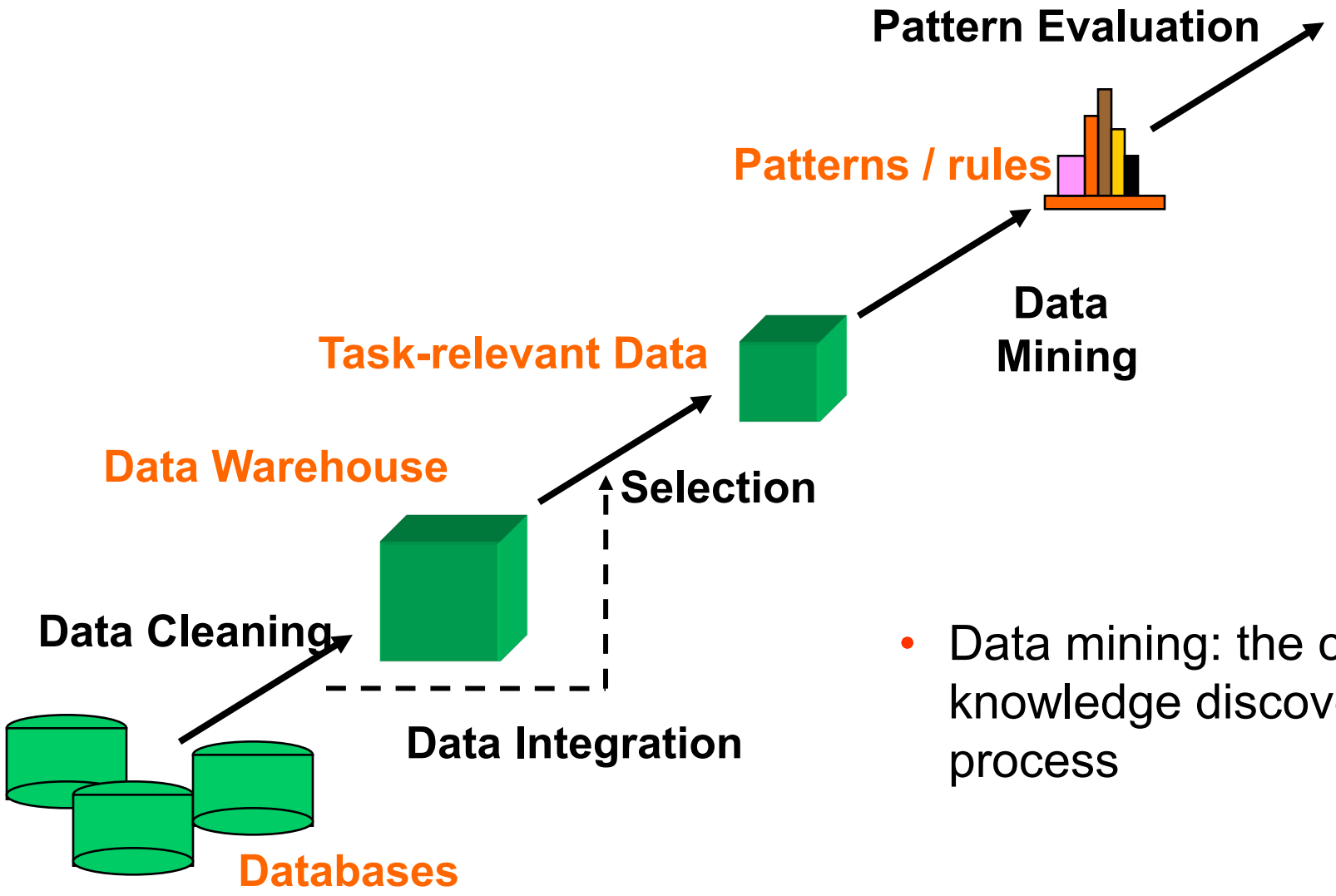
Knowledge-based System Structure



Examples of earliest up to recent Knowledge-based Systems

Year	System	Author(s)	Task
1956		Newell, Simon & Shaw	Proved logic theorems
1961		Minsky & Slagle	Solved mathematical calculus problems
1973	DENDRAL	Feigenbaum	Derived chemical structures from mass spectrograph
1976	MYCIN	Shortliffe	Medical diagnosis of blood disorders
1978	PROSPECTOR	Duda	Prospecting for mineral ores
	...		
1997 -2004	STATUTE	Softlaw Corp.	Human resources ES
1999	OSHA	Stern	Hazard awareness advisor
1999 - ...	Gensym G2	Gensym Corp.	G2 real-time expert system

Data Mining for Knowledge Discovery

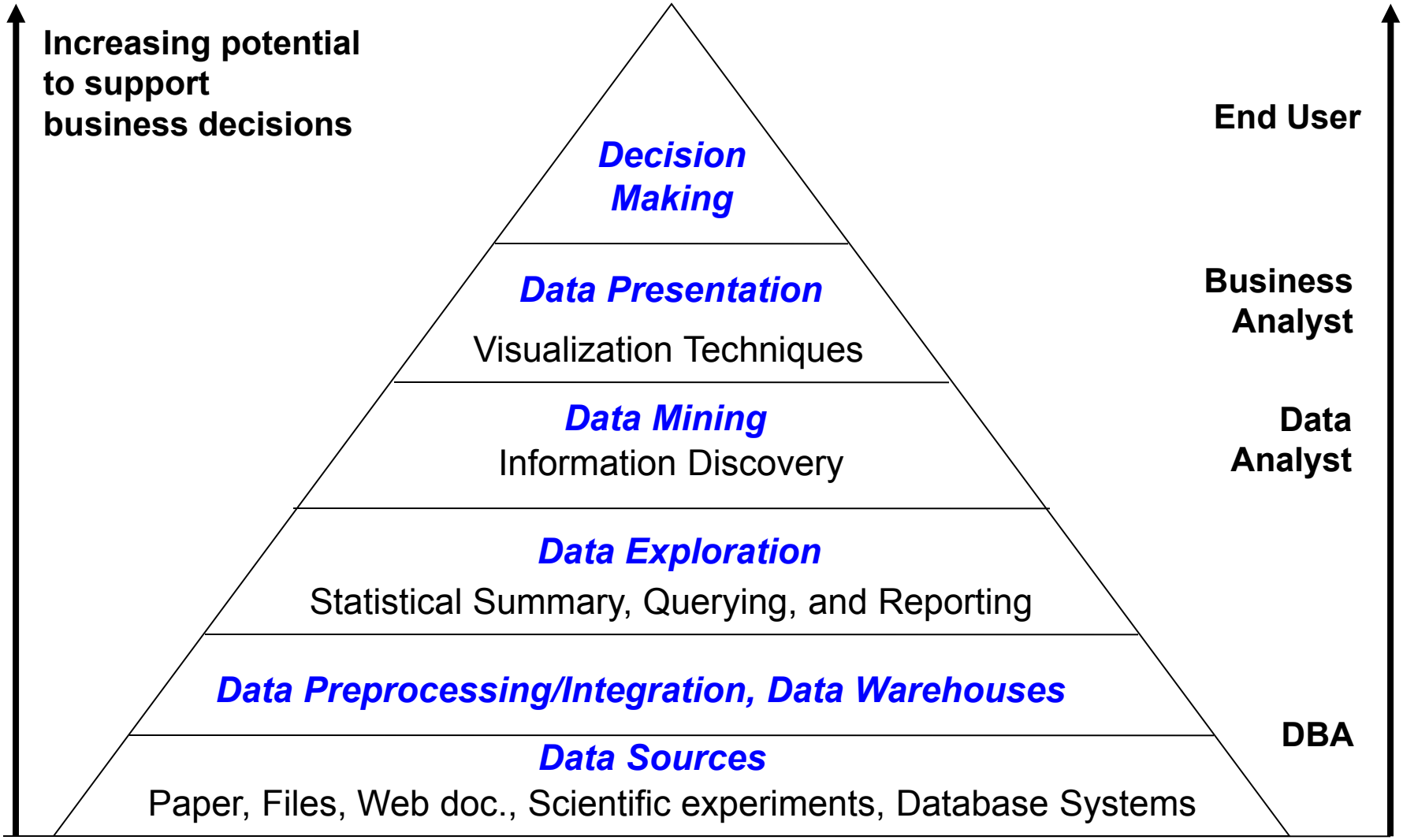


- Data mining: the core of knowledge discovery process

Steps of a Knowledge Discovery Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Identifying or creating a target data set: data selection
- Data cleaning and pre-processing: substantial effort!
- Data reduction and transformation:
  - find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- ***Data mining***: search for interesting knowledge patterns
- Pattern evaluation and knowledge representation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Knowledge Discovery in Business Intelligence



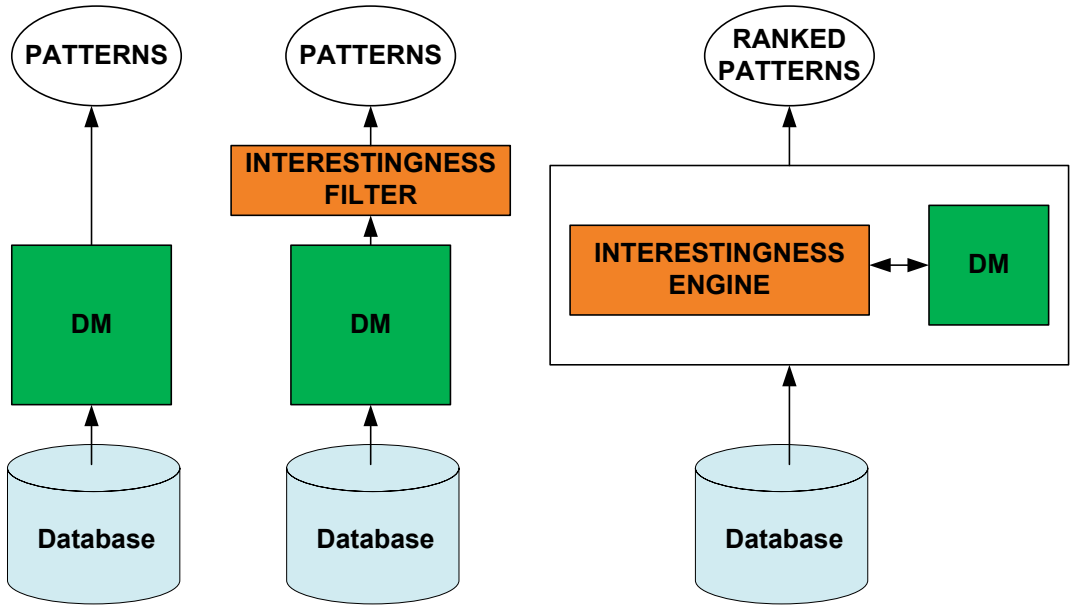
Example: Business Intelligence in Industry

Knowledge Management based on Data Warehouse

- A decision support database that is maintained ***separately*** from the organization's operational database
- Support ***information processing*** by providing a solid platform of consolidated, historical data for analysis
- A ***subject-oriented, integrated, time-variant*** collection of data in support of management's decision-making process
- Process of constructing and using data warehouses

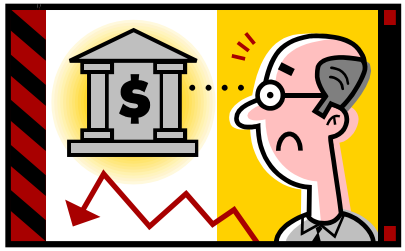
Knowledge Discovery

- Discovery of:
  - new relationships or patterns
- Methods
  - dredging (A)
  - selective (B)
  - interactive (C)
- Legal and ethical issues
  - privacy
  - accountability

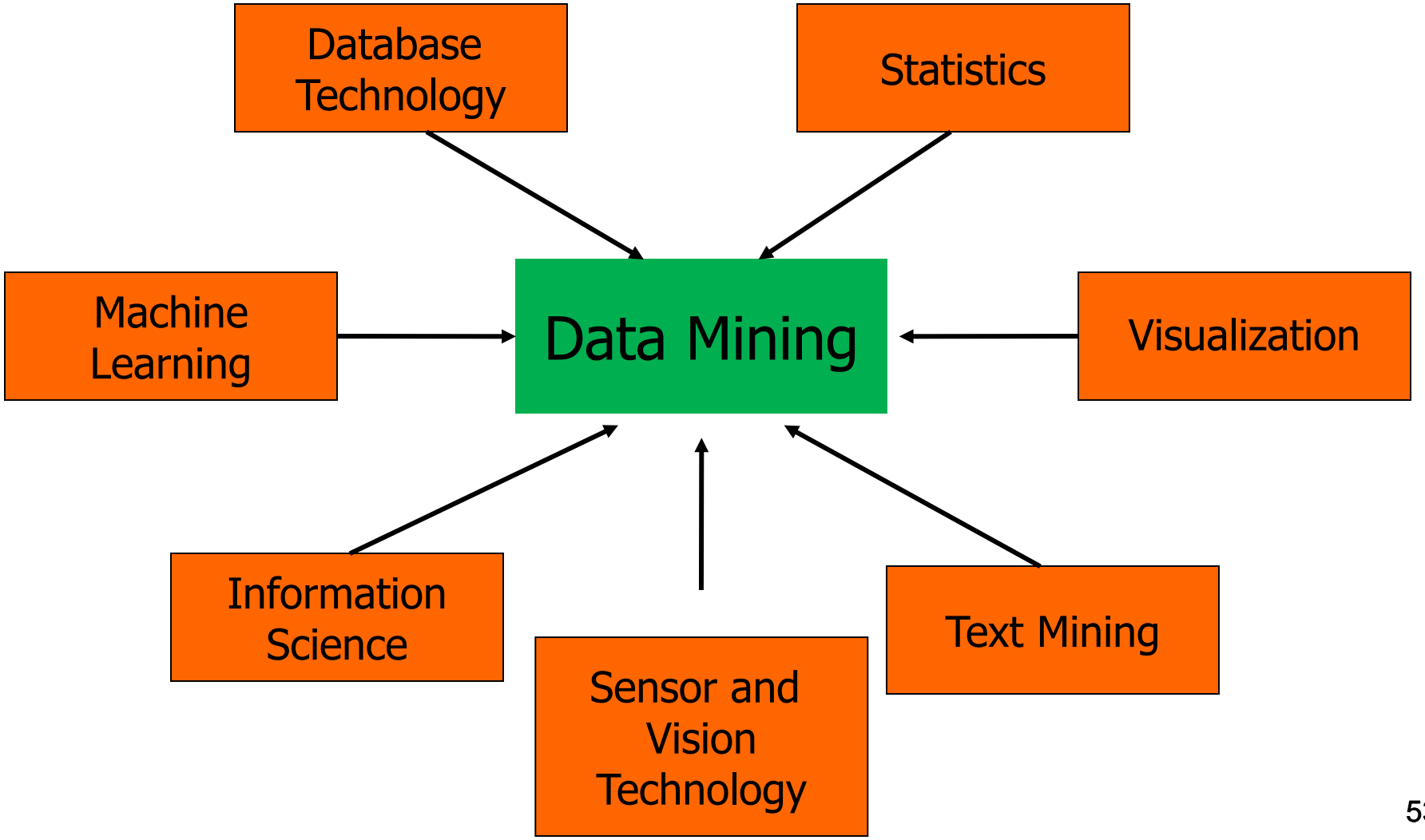


Are all “discovered” Patterns interesting Knowledge?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
- ***Interestingness measures***: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- ***Objective vs. subjective interestingness measures***:
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, actionability, etc.



Knowledge Discovery: Multiple Disciplines



Knowledge Discovery: (1) Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Knowledge Discovery: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Store?
- Association, correlation vs. causality
  - A typical association rule
    - Beer→ Chips [0.5%, 75%] (support, confidence)
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Knowledge Discovery: (3) Classification

- Classification and label prediction
  - Construct models based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Knowledge Discovery: (4) 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 inter-class similarity
- Many methods and applications

Knowledge Discovery: (5) Outlier Analysis

- ***Outlier***: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- Methods: by product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information 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
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

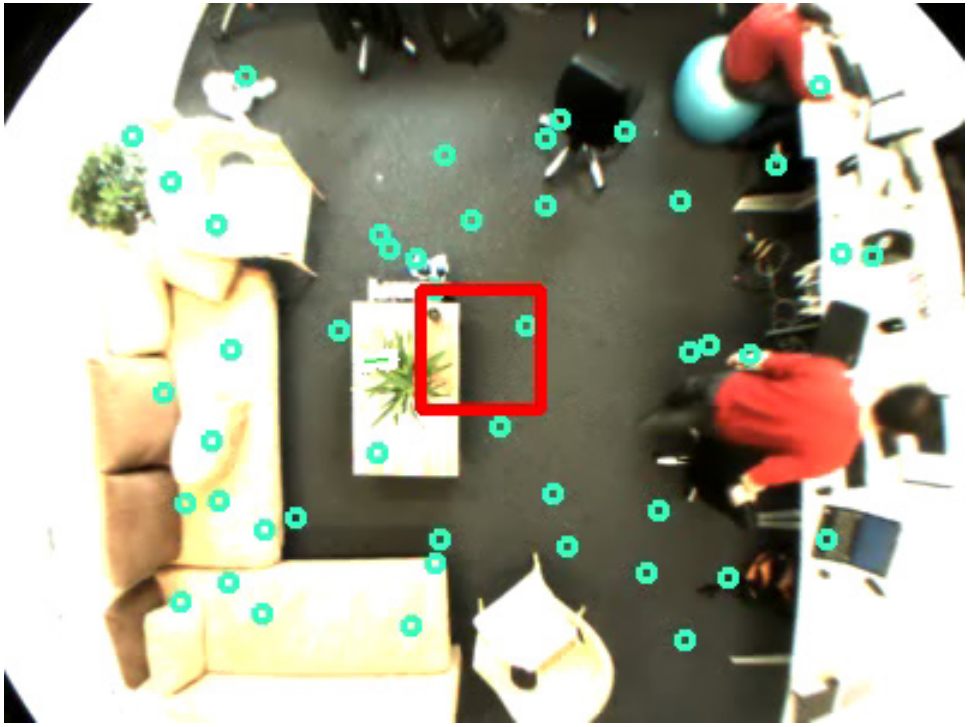
Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of “patterns” and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

Summary

- Knowledge discovery: discovering interesting patterns from large amounts of data
- Knowledge discovery process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge representation
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

Example: Which Patterns are Interesting in the Knowledge Technology Lab?



Research Group Knowledge Technology (WTM)