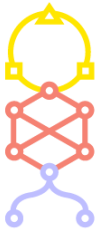
Data-driven Intelligent Systems

Introduction

Dr. Cornelius Weber, Prof. Stefan Wermter,
Kyra Ahrens, Dr. Burhan Hafez, Kerim Erekmen



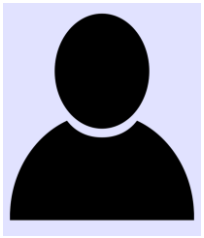
KNOWLEDGE
TECHNOLOGY

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

A bit about us…

- Research Group Knowledge Technology (WTM)
  - Research interest in Hybrid Neural/Symbolic Systems
- Head: Prof. Stefan Wermter
  - Prior to Hamburg at Univ. of Sunderland, Berkeley, Dortmund, Massachusetts

- Team for DAIS:
  - Dr. Cornelius Weber
  - Kyra Ahrens
  - Dr. Burhan Hafez
  - Kerim Erekmen



What is Data? What is Knowledge?

π = 3,14159 26535 89793 23846

IF winter
THEN cold

marietta.jpg

http://map-of-world.com

information and skills acquired by *education* and *experience*

information and processing methods acquired by *programming* and *machine learning*

Topics / Lecture Overview

week

2. Data and visualization methods
3. Pre-processing methods
4. Decision trees, decision rules
5. Classification and supervised neural networks
6. Theory and Evalutation
7. Deep and recurrent neural networks for classification
8. Ensemble learning
9. Intelligent agents: reinforcement learning and planning
10. Association Rules, Clustering and self-organisation
11. Mining structure from graphs
12. Text mining
13. Revision

supervised learning

unsupervised learning

Organisational Issues

Module Data-driven Intelligent Systems (DAIS)

- 4 SWS Lecture
  - Evaluated by a **written** exam
    (one planned near the beginning, another near the end of the semester break – see STiNE for the dates)
- 2 SWS Tutorials (Lab)
  - practical exercises related to the lecture
  - must be successfully completed

Lecture in 1st Post-Corona Semester

- Module can be done entirely from home / on-line
  - Slides are uploaded as PDF to the Moodle
  - Lecture recordings are available from Lecture2go
    - Upload will be before scheduled start of corresponding lecture
    - Some lecture recordings may be taken from year 2021
- On Mondays 14:15 on Zoom there will be a Q&A session with the lecturer
  - Students ask questions about current and previous lectures
- Additional offer in presence: On Wednesdays 10:15-12:45 in D-125 another Q&A session
  - Maybe we will quickly browse through the week's lectures

Benefits of Attending the Lectures

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

DAIS Lab / Tutorial (I/II)

- Practical part of this module to ...
  - train some methods with exercises
  - test mini-system in an own implementation
- Regular participation is mandatory
  - Prepare at home before every tutorial
  - Solve a programming practical exercise
  - Defend your solution in the end of every meeting
- Additionally, a weekly quiz needs to be answered

DAIS Lab / Tutorial (II/II)

- Groups (Tutors: Kyra Ahrens, Burhan Hafez, Kerim Erekmen)
  - Thursday 14-18
  - Friday 10-14
  - Friday 10-14
  - Friday 14-18
  - Friday 14-18
- Either Zoom or in presence (D-114 / D-118)
- Tutorial: every 14 days, starting 7./8.04.
  - But once, a 3-week delay over the lecture-free week

DAIS - Timeline

as on STiNE

week	Lecture (Mondays+Wednesdays)	Tutorials (Thursdays or Fridays)
1	**4.4. + 6.4.**	**7.4. or 8.4.**
2	**11.4. + 13.4.**	
3	**20.4.** (not: Easter Monday)	**21.4. or 22.4.**
4	**25.4. + 27.4.**	
5	**2.5. + 4.5.**	**5.5. or 6.5.**
6	**9.5. + 11.5.**	
7	**16.5. + 18.5.**	**19.5. or 20.5.**
8	("shifted" Whitsun/Pentecost)	
9	**30.5. + 1.6.**	
10	**8.6.** (not: Whit Monday)	**9.6. or 10.6.**
11	**13.6. + 15.6.**	
12	**20.6. + 22.6.**	**23.6. or 24.6.**
13	**27.6. + 29.6.**	
14	**4.7. + 6.7.**	**7.7. or 8.7.**
15	**11.7. + 13.7.**	
	First written exam planned for 18.7. Second exam near end of Semester break.	

DAIS – Time Investment

- Estimated time investment in hours (source: Modulhandbuch)

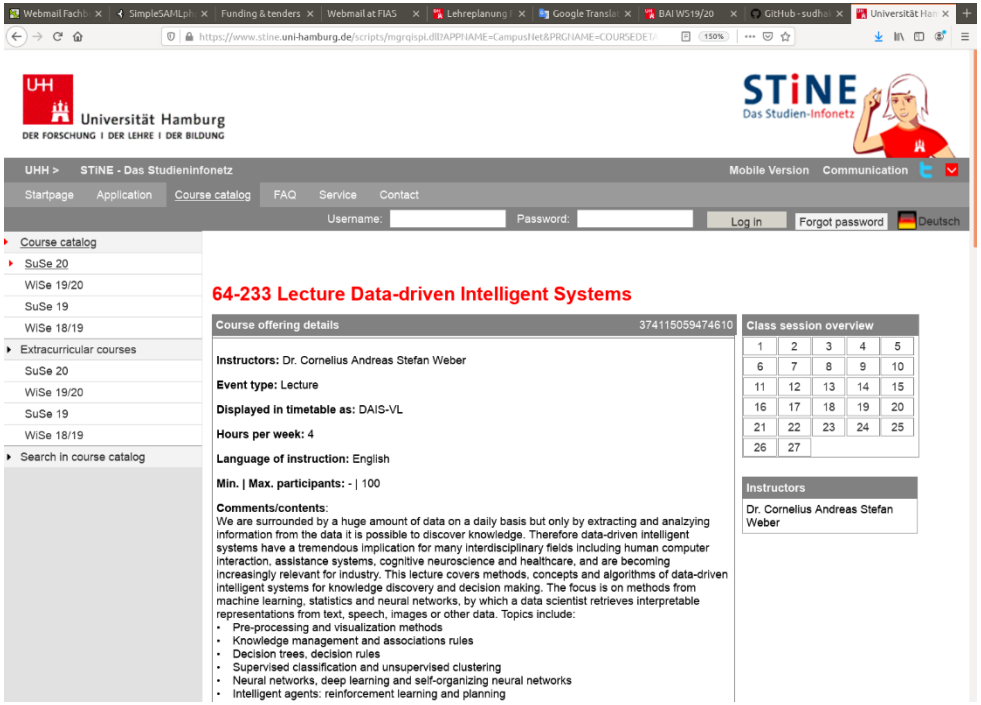
	Presence	Self-study	Exam preparation
Lecture	**56**	**84**	**40**
Tutorials	**28**	**42**	**20**

14 weeks * 4 h

14 weeks * 4 h * 1.5

Communication I/III: STiNE

- Our platform for:
  - Latest news
  - Link to Moodle



you are a member in STiNE already

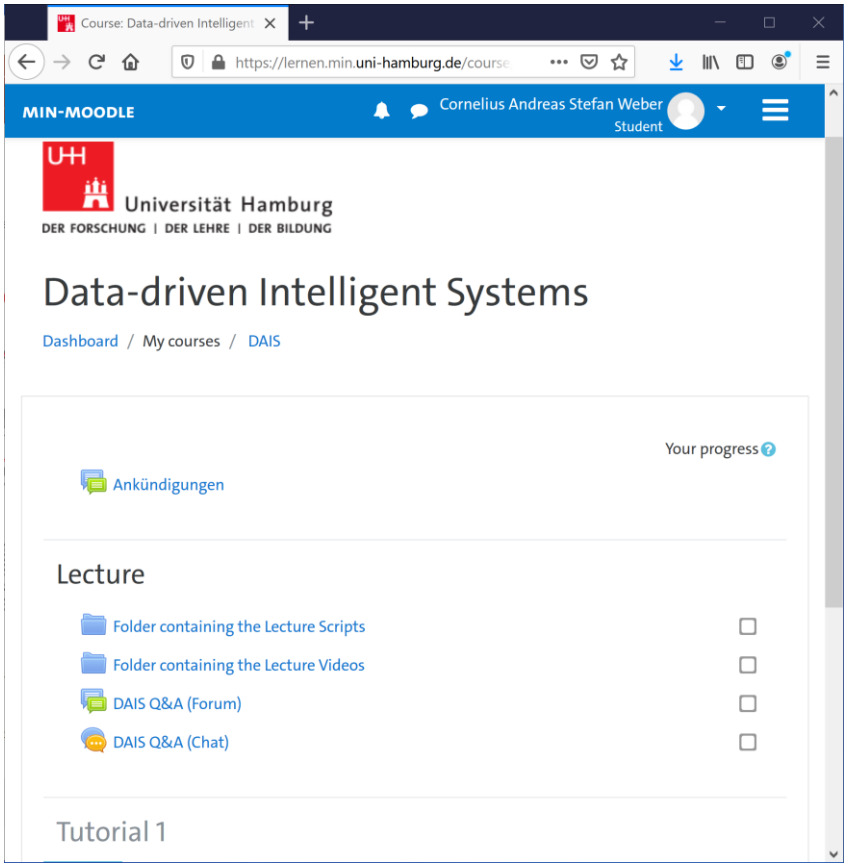
Website: https://www.stine.uni-hamburg.de

Or follow the link on our Website:

https://www.inf.uni-hamburg.de/en/inst/ab/wtm/teaching.html

Communication II/III: Moodle

- Our platform for:
  - Lecture slides and link to Lecture2Go
  - Tutorial materials
  - Latest news
- Visit the page regularly



- Enrolment key:

22~DAISy

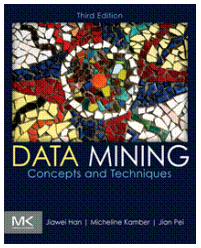
Web link:

https://lernen.min.uni-hamburg.de/course/view.php?id=2247

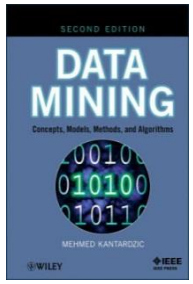
Communication III/III: Zoom

- Our platform for:
  - Tutorials
  - Lecture Q&A Sessions
- Zoom link:
  - https://uni-hamburg.zoom.us/j/66667846898?pwd=MExIcC9DZFpDYVVtcDdDMXpmaE5jdz09
  - Meeting ID: 66667846898
  - Passcode: 22~DAISy

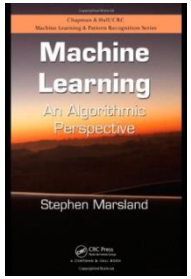
Literature and Acknowledgements



- Mehmed Kantardzic. *Data Mining*. Wiley, 2011.
  main text book
- Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: Concepts and techniques*. Morgan Kaufmann, 2011.



- Stephen Marsland. *Machine Learning - An Algorithmic Perspective.* CRC Press, 2009.

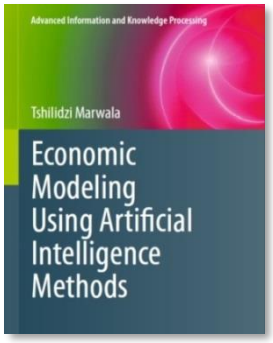


- **Virtual Bookcase**: https://www.inf.uni-hamburg.de/inst/bib/service/vib.html
- Thanks to slides by M. Kantardzic, J. Han and M.Kamber, and S. Marsland.
  - Slides follow mainly textbook of Kantardzic
  - Additional slides from Shane Warren and Brittney Ballard

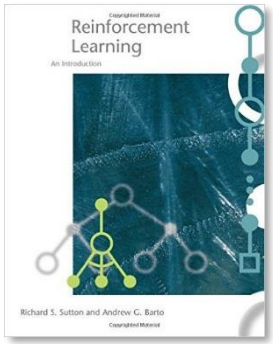
Other Optional Literature

- Tshilidzi Marwala. *Economic Modeling Using Artificial Intelligence Methods*

  Springer E-Book-Paket *Computer Science* via Campus-Katalog: https://kataloge.uni-hamburg.de/DB=1/XMLPRS=N/PPN?PPN=744996422



- Richard S. Sutton and Andrew G. Barto *Reinforcement Learning: An Introduction*
  A Bradford Book The MIT Press
  https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf



Advantages of Lectures in English Language

- Research-oriented
  - available papers, your future papers
  - most computer science resources originate in English
  - international colleagues
- Avoid German-English mixtures or bad translations
  - e.g. „Datenbergbau" (→ *„Wissensentdeckung aus Daten")*
- Train your English skills (you won't forget German anyway)
  - companies are becoming more and more international
  - may help your future professional life

Tutorials will be offered in both English and German

Questions about Data-driven Intelligent Systems

- What are data-driven intelligent systems?
- Why data-driven? Motivation and benefits?
- What kind of data to drive your intelligent systems?
- How to organize the process?
- What are the challenges?

Definitions in Dictionaries

- 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)

- An Intelligent System is a living or technical system capable of intelligent ***behaviour***. This includes capability of ***adaptation*** to changing environmental conditions.

  (de.wikipedia.org/wiki/Intelligentes_System)

Why Data-driven Systems?

We are flooded by data:

- Business transactions (bank, telecom, ...)
- Web (text, e-commerce, social networks, …)
- Multimedia (images, voice, video, …)
- Scientific data (astronomy, biology, …)

Collection, storage, and processing technologies improve

- ➢ We can extract interesting information from the data
- ➢ Discovered patterns can improve our decisions
- ➢ Better than just rule-based behaviour

Petabytes* of Data and More?

- MEDLINE text database
  - 22 million references to published articles in life sciences
- Google
  - Indexed over 4 billions web pages
  - Over 40,000 search requests per second
  - More than 3.5 billion search requests per day
- NASA MODIS satellite
  - Coverage: 250m resolution, 37 bands, whole earth, every day
- Walmart transaction data
  - Order of 100 million transactions per day

* 10^{15} bytes

Zettabytes of Data?

$2^{70} \approx 10^{21}$ →

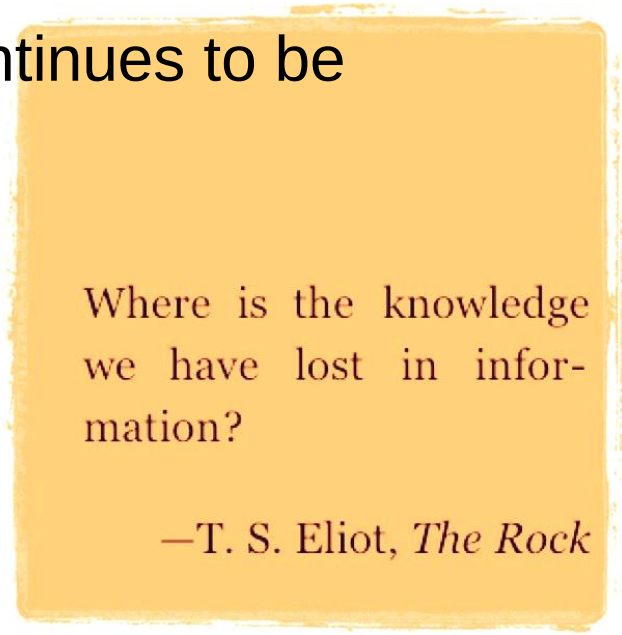
Zettabyte and its Friends

Amount	In Byte	~	English	German	
1 Bit	1/8				
1 Byte	1		One	Eins	10^0
1 Kilobyte	1024		Thousand	Tausend	10^3
1 Megabyte	1024^2		Million	Million	10^6
1 Gigabyte	1024^3		Billion	Milliarde	10^9
1 Terabyte	1024^4		Trillion	Billion	10^{12}
1 Petabyte	1024^5		Quadrillion	Billiarde	10^{15}
1 Exabyte	1024^6		Quintillion	Trillion	10^{18}
1 Zettabyte	1024^7		Sextillion	Trilliarde	10^{21}
1 Yottabyte	1024^8		Septillion	Quadrillion	10^{24}
1 Brontobyte	1024^9		Octillion	Quadrilliarde	10^{27}

Why Data-driven Technologies Now?

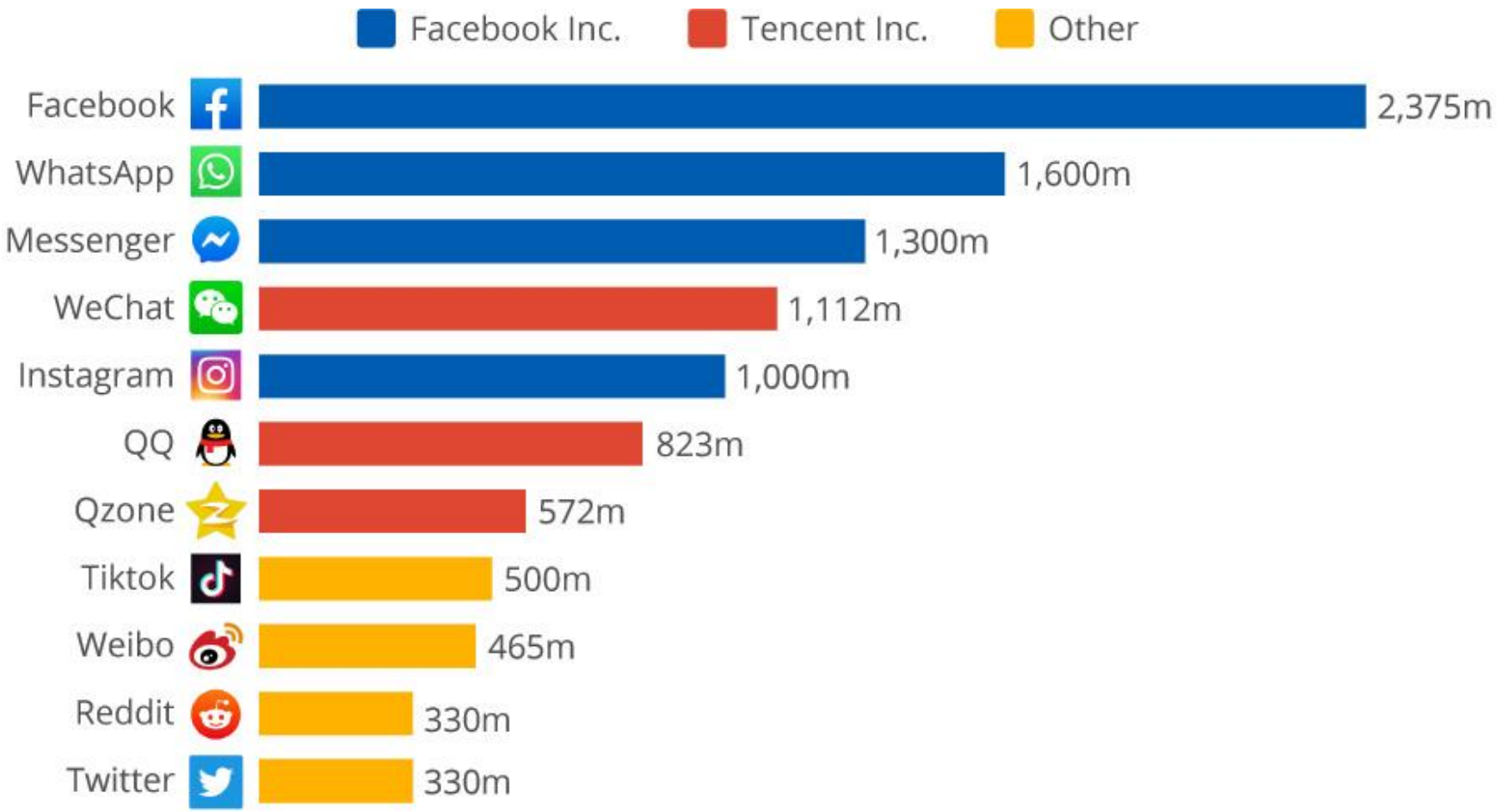
- Only a small portion (5% - 10%) of the collected data is ever analyzed.
- Data that may be never analyzed continues to be collected at great expenses.

WE ARE DROWNING IN DATA,
BUT STARVING FOR KNOWLEDGE!



- There is ***a gap*** between
  - data collection and organization capabilities
  - extraction of useful information for decision processes.

Global Monthly Active Social Platform Users



* July 2019 or latest available

Source: Company data via DataReportal Q3 Global Digital Statshot

Stories: Managers Believe …

- 61% believe that information overload is present in their workplace,
- 50% ignore data in current decision process because of overload,
- 80% believe the situation will get worse,
- 84% store the data for future – without any current analysis,
- 60% believe that the cost of gathering information outweighs its value!

→ not a promising conclusion!

Data Mining: What is it not?

- Potential point of confusion:
  - The “mining of ore from rock” metaphor ***does not*** really apply to data mining
  - If it did, then database queries (SQL, …) would define data mining
- Data mining is furthermore ***not***:
  - Simply crunching data, transforming it into different formats
  - Only visualization, presenting data in different ways
  - “Blind” application of algorithms, finding relationships where none exist

Data Mining: What is it?

- Classical modeling and analyses is based on ***first principle***.
- Data mining attempts to ***develop 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***.

From Data to Knowledge

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

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

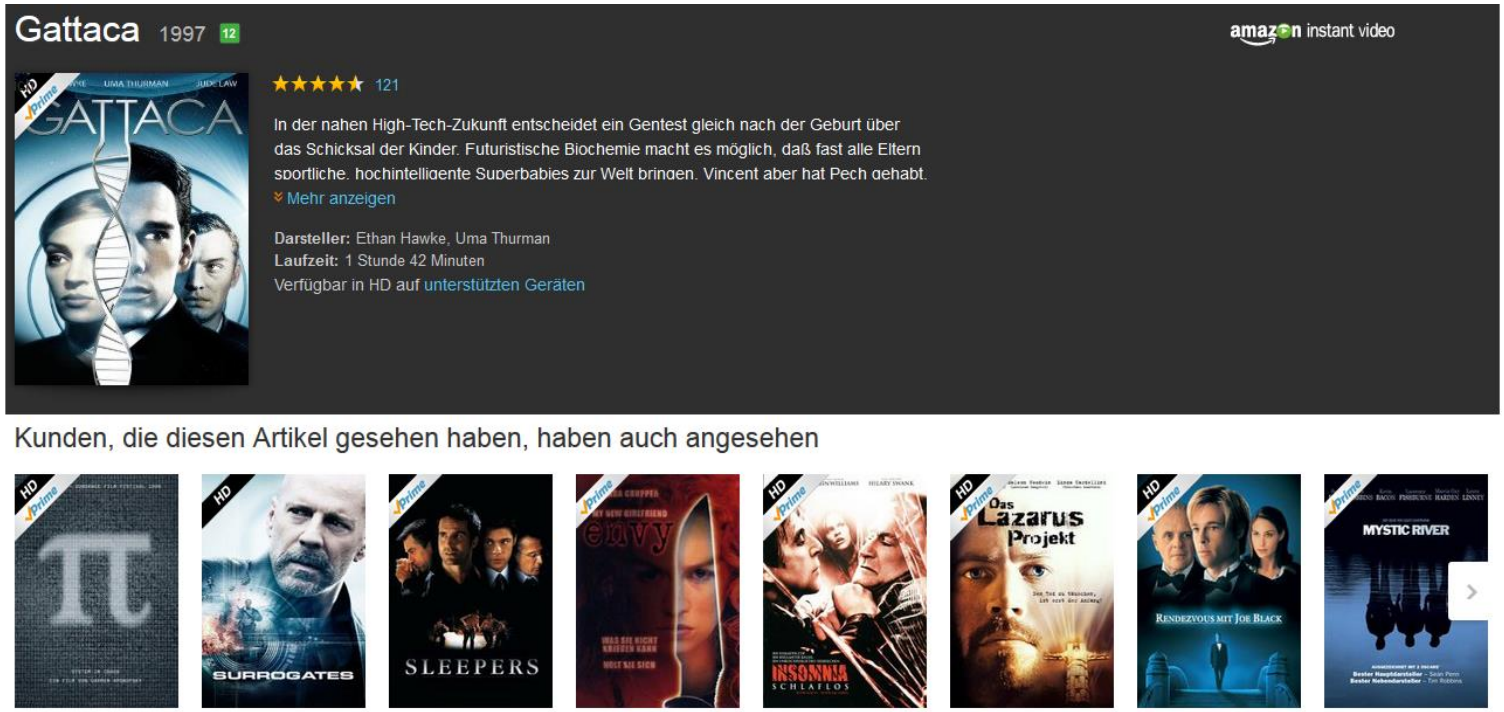
- Can I develop a general characterisation/profile of different investor types? (***clustering or classification***)
- What characteristics distinguish between Online and Broker investors? (***discrimination***)
- Can I develop a model that predicts the average trades/month for a new investor? (***prediction***)

An Application

- Movies recommendations in online video libraries

amazon.de Prime

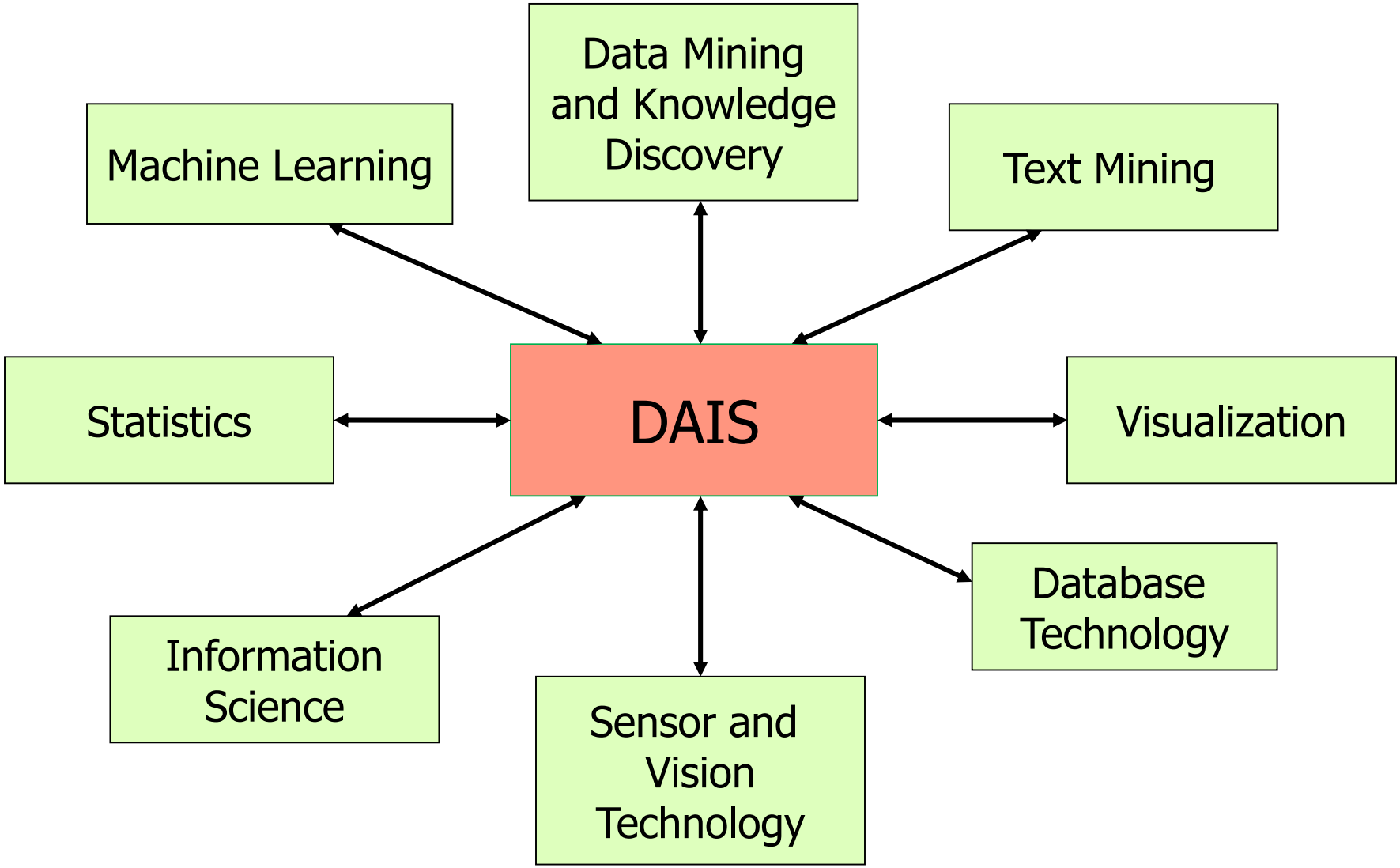
NETFLIX



Challenges for Data-driven Systems

- Technical
  - From Tera-bytes and Peta-bytes and Zetta-bytes…
  - Complex, multi-media, unstructured data
  - Integration with domain knowledge and human expertise
- Business
  - Finding good application areas
  - Finding suitable techniques for knowledge extraction
- Societal
  - Legal and ethical issues: privacy, security and accountability

Data-driven Intelligent Systems in Context



DAIS – What‘s in a Name?

Figure created with wortwolken.com using these words: **Data-driven Intelligent Systems**, Intelligent Agents, **Machine Learning**, Unsupervised Learning, Supervised Learning, Reinforcement Learning, **Data Mining**, Intelligent Data Analysis, Data Fishing, Data Dredging, Data Archeology, Information Harvesting, **Data Science**, Knowledge Extraction, **Knowledge Discovery**, Databases

Data, Information, Knowledge? (Webster's)

- ***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 defined "overlapping"

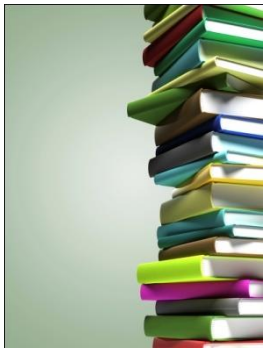
A Semiotic View of Data, Information, Knowledge

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

Knowledge in Humans

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

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

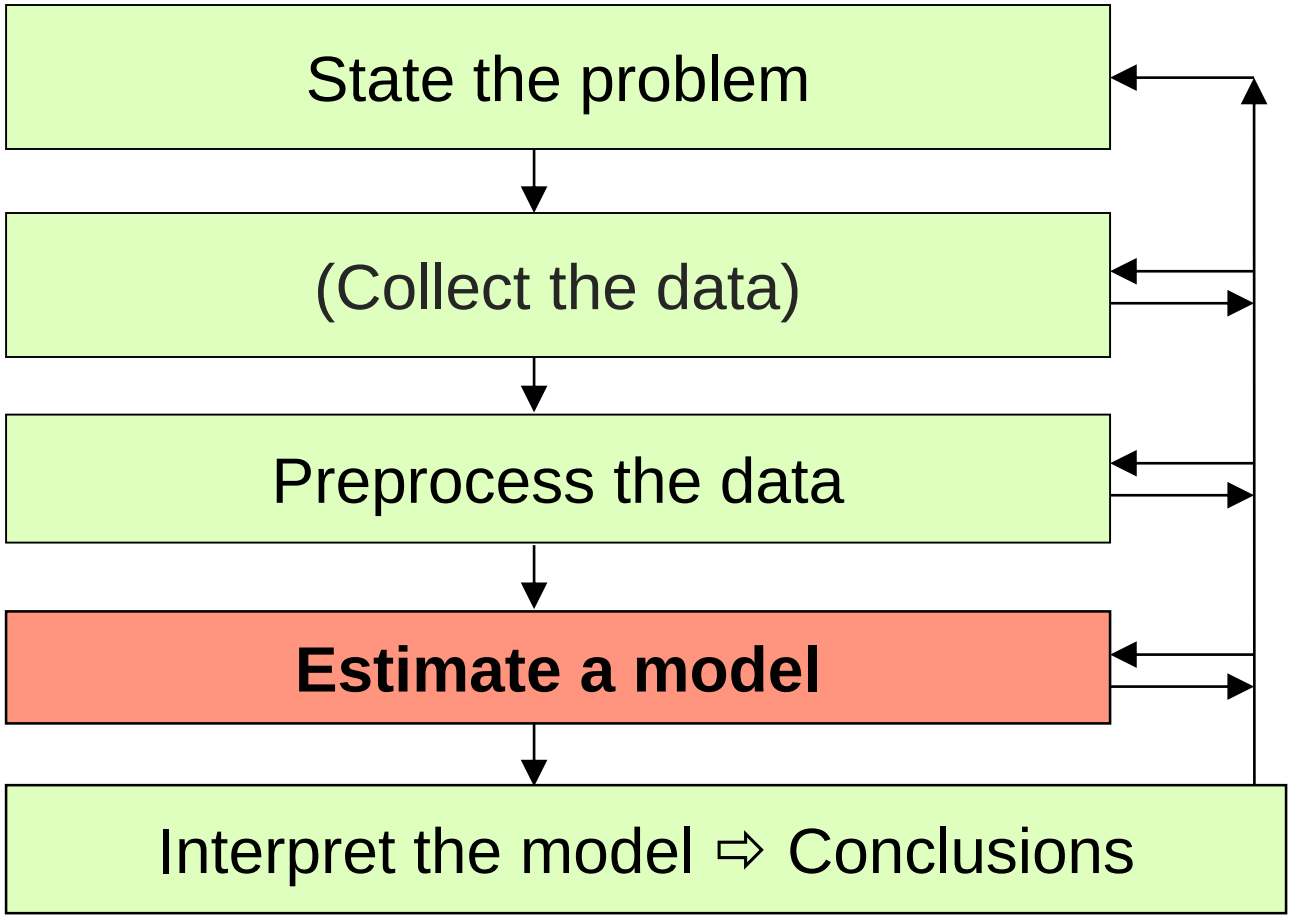


Knowledge in Organizations

- Knowledge and know-how of employees is vital for economical success of an organization
- Methods to **preserve, communicate and enhance** 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…

Learning from Data as a Simplified Process



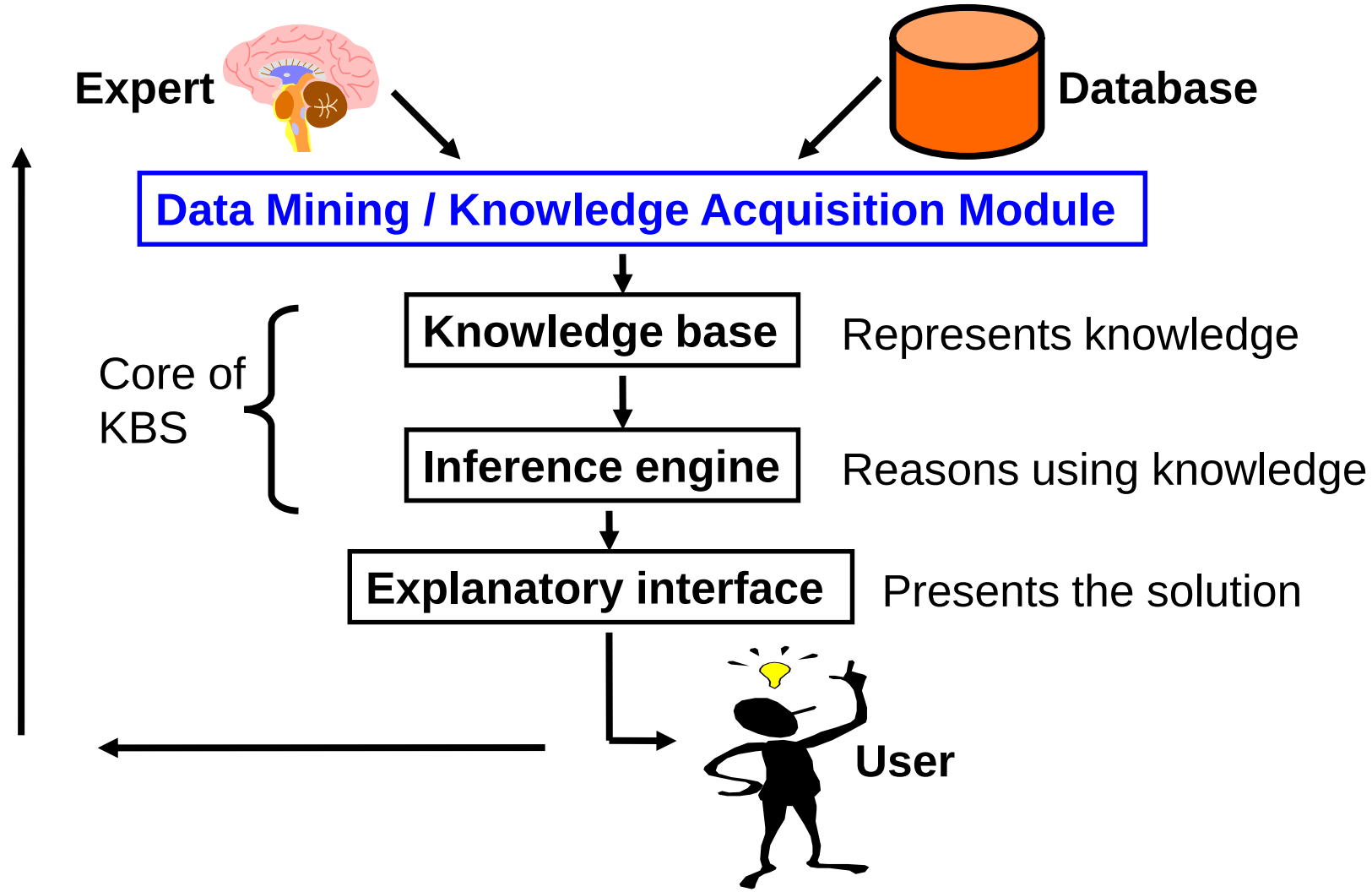
Bottom-Up and Top-Down …

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

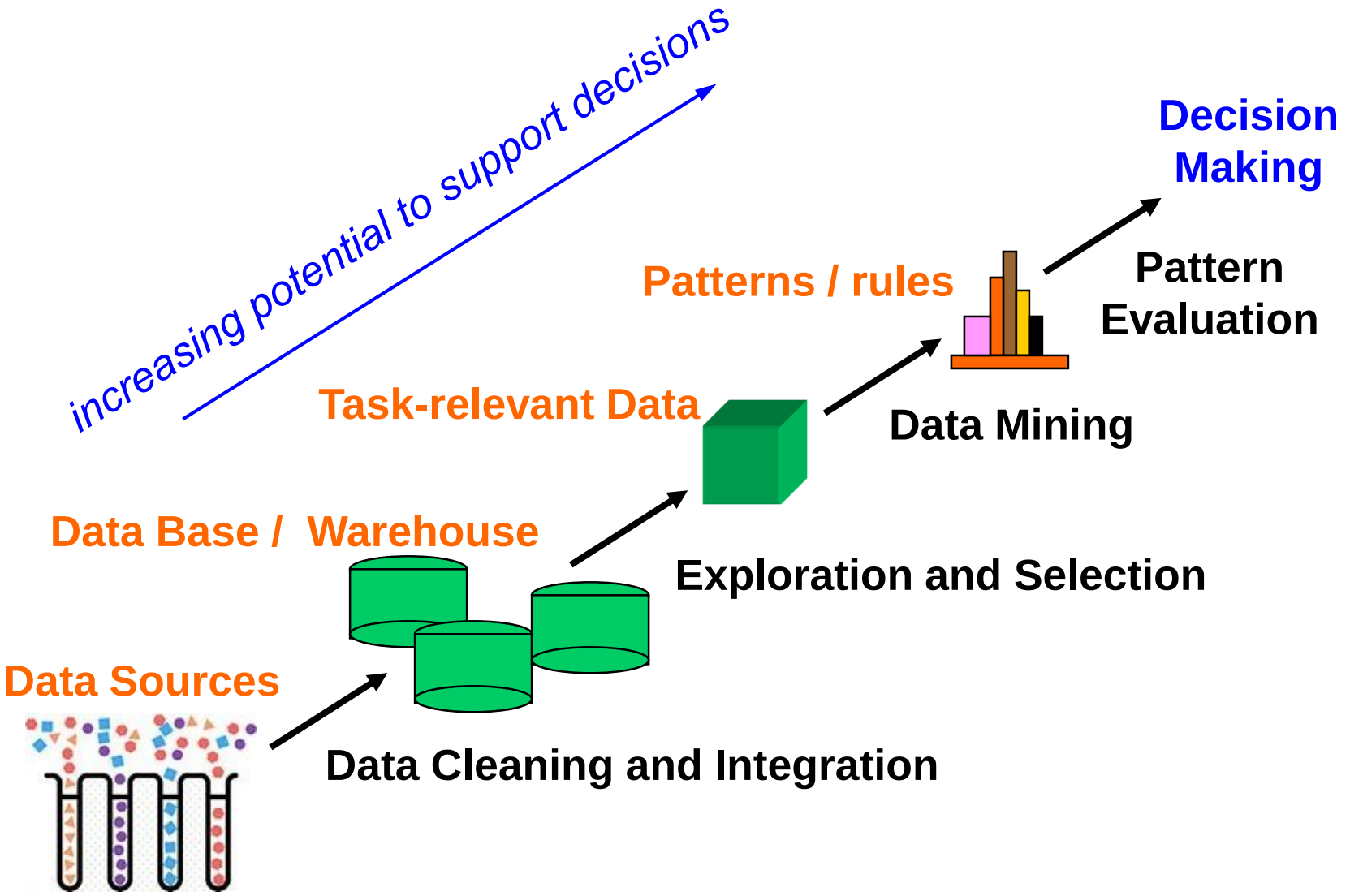
You may generate *"potentially right or wrong hypotheses"*, before you get actionable, ***meaningful*** and useful knowledge

The results must be actionable

Data Mining in Knowledge-based System



Knowledge Discovery from Data

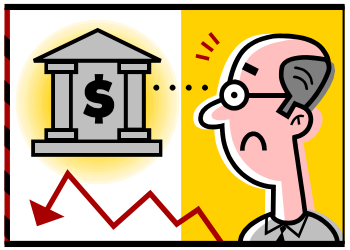


Steps of a Knowledge Discovery Process

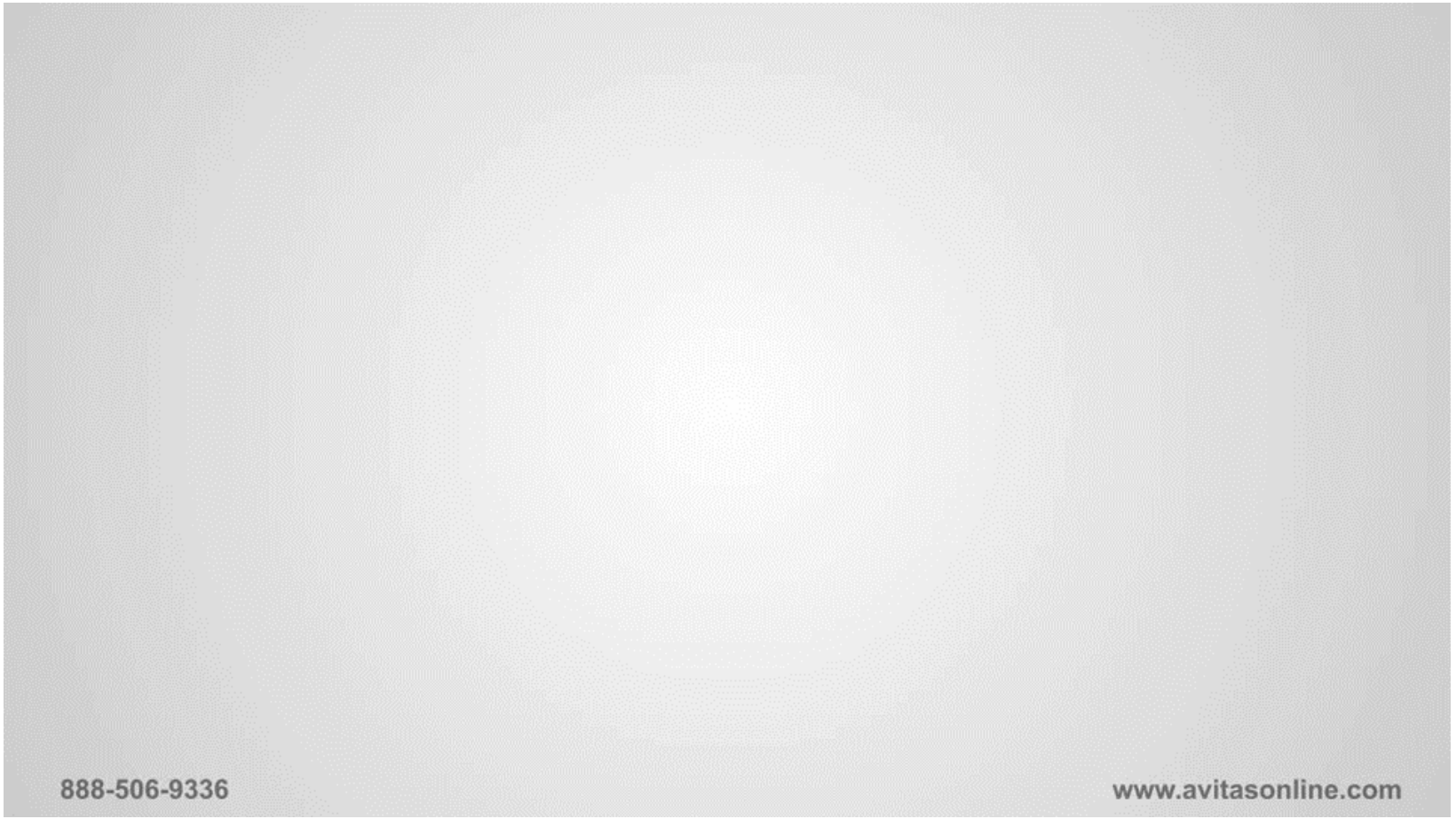
- Learning the application domain
  - relevant prior knowledge and goals of application
- Identifying or creating a target data set: data selection
- Data cleaning and pre-processing (substantial effort!)
- Data reduction and transformation
  - find useful features, dimensionality/variable reduction,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

Are all "discovered" Patterns Interesting Knowledge?

- A data mining system/query may generate thousands of patterns …
- 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
  - validates some hypothesis that a user seeks to confirm
- ***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.



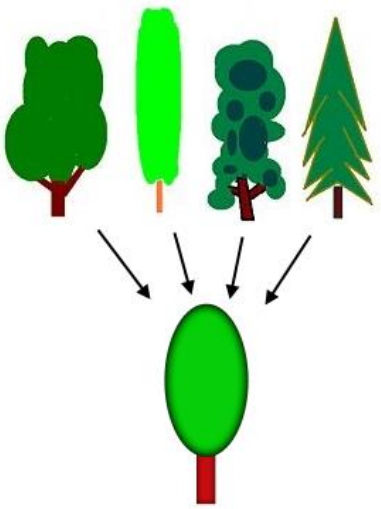
Example: from Business Intelligence to Business Analytics in Industry



Functions of 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

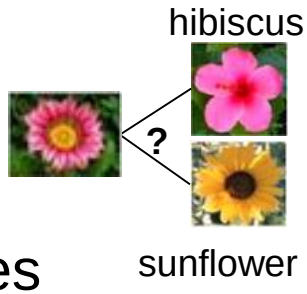


Functions of Knowledge Discovery: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your store?
- Association, correlation vs. causality
  - A typical association rule
    - Beer→ Chips [20%, 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?

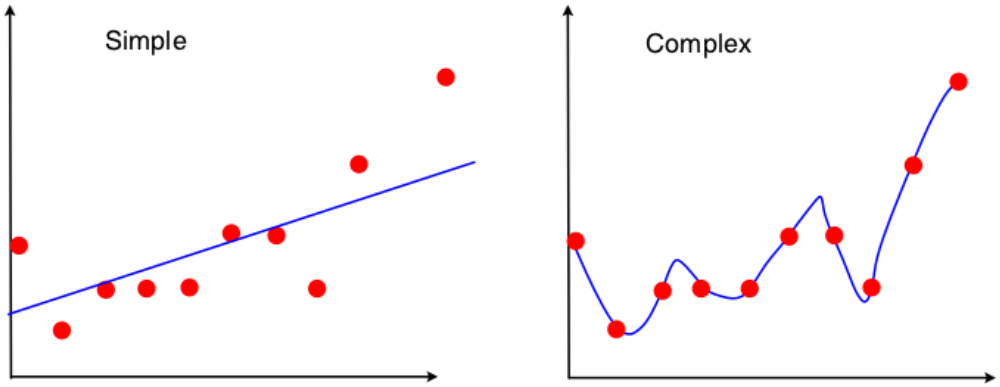
Functions of Knowledge Discovery: (3) Classification



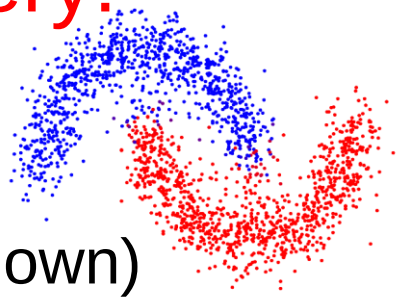
- Classification and label prediction
  - Construct models based on some training examples
  - Describe classes or concepts for future prediction
    - E.g., classify countries based on “climate”, or classify cars based on “gas mileage”
  - Predict some unknown class labels
- Typical methods
  - Decision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

Functions of Knowledge Discovery: (4) Regression

- Prediction of (a) real valued variable(s)
  - Map a data item to a real-valued prediction variable
  - Similar to classification
    - Supervised learning
    - Same set of methods

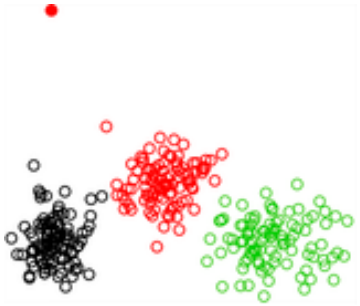


Functions of Knowledge Discovery: (5) Cluster Analysis



- Unsupervised learning (i.e., class label is unknown)
- Group data to form new categories (i.e., clusters)
- Principle: Maximizing intra-class similarity & minimizing inter-class similarity
- Typical methods
  - K-means, spectral clustering
- Applications
  - Various: identification of groups
  - Helper for classification, outlier detection, visualisation

Functions of Knowledge Discovery: (6) Outlier Analysis

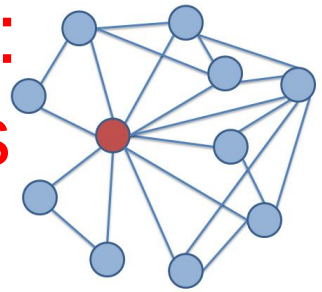


- ***Outlier***: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure? → *Think about an example!*
- Methods: by-product of clustering or regression analysis, …
- Useful in fraud detection, rare events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

Functions of Knowledge Discovery: (7) Sequential Pattern and Trend Analysis

- Finding ordered temporal patterns
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy SD memory cards
  - Periodicity analysis
  - Similarity-based analysis

Functions of Knowledge Discovery: (8) Structure and Network Analysis



- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: persons (nodes) and relationships (edges)
    - e.g., author networks in science, terrorist networks
  - Multiple heterogeneous networks
    - Persons are in multiple 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
    - Web community discovery, opinion mining, usage mining, …

Functions of Knowledge Discovery: (9) Summarization

- Find a compact description for a subset of data

- Microsoft‘s Xiaoice learns about and to interprete weather readings and then delivers reports

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.
- Intelligent systems may act, based on the learnt knowledge

Research Group (Video)

Please see our Knowledge Technology group overview video:

https://fiona.uni-hamburg.de/0eeef64f/knowledgetechnologygroup.mp4

More videos from the group can be seen here:

https://www.inf.uni-hamburg.de/en/inst/ab/wtm/media/videos.html