Knowledge Discovery and Data Mining 1 (VO) (707.003)

Denis Helic

KTI, TU Graz

Oct 1, 2015

Lecturer

Name:	Denis Helic
Office:	IWT, Inffeldgasse 13, 5th Floor, Room 070
Office hours:	Tuesday from 12 til 13
Phone:	+43-316/873-30610
email:	dhelic@tugraz.at

Lecturer

Name:	Roman Kern
Office:	IWT, Inffeldgasse 13, 6th Floor, Room 072
Office hours:	By appointment
Phone:	+43-316/873-30860
email:	rkern@know-center.at

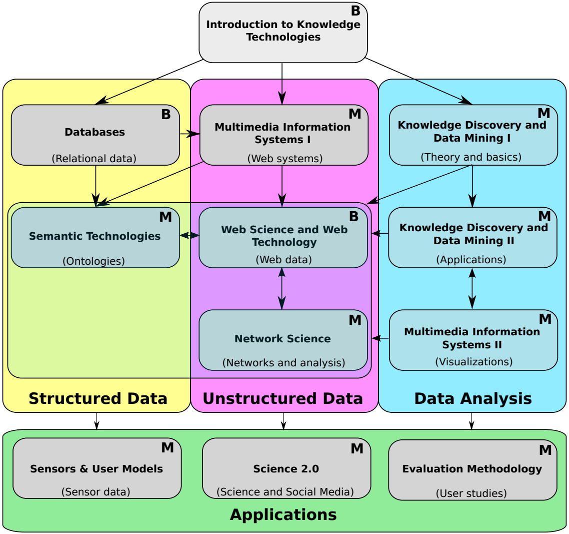
Language

- Lectures in English
- Communication in German/English
- If in German: please informally (Du)!
- Examination: German/English

Outline

1. Welcome and Introduction
2. Course Organization
3. Motivation
4. Course Overview
5. Course Highlights

Teaching @ KTI



+ Projects, Bachelor Thesis, Master Projects, Master Thesis, PhD Thesis

Course context

- Knowledge Discovery and Data Mining 1 (VO) (707.003)
- Obligatory course Master Software Development and Business (1st Semester)
- Elective course in subject catalogue “Knowledge Technologies”
- Computer Science, Telematics

Course context

- Knowledge Discovery and Data Mining 1 (KU) (707.004)
- Free course
- An add-on for the theoretical part
- Highly suggested

Goals of the course

- The overall goal of KDDM and related courses is to learn how to discover patterns and models in data. Discovered patterns need to be:
  (i) Valid: hold for new data with high probability
  (ii) Useful: we can base further actions on them
  (iii) Unexpected: non-obvious
  (iv) Understandable: humans can intepret them

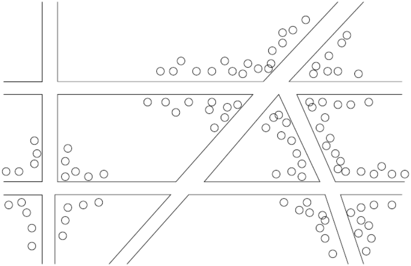
Goals of the course: patterns example

1854 Broad Street cholera outbreak

Extracting clusters of cholera outbreak in the city of London in 1854. The cases clusterd around some intersections of roads in London. These had contaminated water wells.

- http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Goals of the course: patterns example



Goals of the course

- Specific goals of this course are to learn about **two** (three) cruicial elements of that discovery:
  (i) **Tools: mathematical tools such as probability theory, linear algebra, information theory, and statistical inference**
  (ii) Infrastructure: models of computation for large data (KDDM2 has also a practical part)
  (iii) **Process: steps that are needed to discover patterns**
- I assume here that you already know how to program and develop software

Goals of the course

- Student goals: to pass the examination
- Bonus goal for all: to have fun!

Course Calendar

- 01.10.2015: Course organization, Introduction and Motivation (Denis)
- 08.10.2015: Preprocessing (Roman)
- 15.10.2015: Feature Extraction (Roman)
- 22.10.2015: **Partial Exam 1** / Project presentations (KU)
- 29.10.2015: Feature Engineering (Roman)

Course Calendar

- 05.11.2015: Data Matrices (Denis)
- 12.11.2015: Principal Component Analysis (Denis)
- 19.11.2015: SVD and Latent Semantic Analysis (Denis)
- 26.11.2015: Recommender Systems: Matrix Factorization (Denis)
- 03.12.2015: **Partial Exam 2** / Project presentations (KU)

Course Calendar

- 10.12.2015: Classification (Denis)
- 17.12.2015: Clustering (Roman)
- 14.01.2016: Pattern Mining (Roman)
- 21.01.2016: **Partial Exam 3** / Project presentations (KU)
- 28.01.2016: Examination

Course Logistics

- Course website: http://kti.tugraz.at/staff/denis/courses/kddm1
- Slides will be made available on the course website
- Additional readings, references, links, etc. also on the website
- We expect that you have basic knowledge in **probability theory** and **linear algebra**
- To freshen the knowledge in the first midterm there will be questions from these topics!
- Please check the homepage for details
- As a side note: we also expect that you know how to **program** (relevant for the practical part)

Grading

- Partial examinations within the class
- Written examination at the end of the class
- Two additional examinations in winter semester
- Three examinations in summer semester
- Examination material: contents of slides
- In class we will discuss certain types of exam questions

Partial Examinations

- 3 written examinations
- In the beginning of a lecture: 30 minutes
- Each partial examination 2 questions
- Difficulty adjusted to solve both problems in approx. 25 minutes
- Max 15 points for each question
- Total points: 90

Examination

- Written examination 90 minutes
- 4 questions
- Difficulty adjusted to solve all four in approx. 70-80 minutes
- Max 20 points for each question
- Total points: 80

Grading

- 0-40 points: 5
- 41-50 points: 4
- 51-60 points: 3
- 61-70 points: 2
- 71-80 points: 1

KU Organization

- KU organization on 22.10.2015

Questions?

- Raise them now (+1 +1)
- Ask after the lecture (+1)
- Visit me in the office hours (+1)
- Send me an e-mail (±1)
- As a side note: you should(!) interrupt me immediately (+1 +1 +1) and ask any question you might have during the lecture

How much information is being produced?

- Study at Berkley: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
- The World Wide Web contains about 170 terabytes of information on its surface
- This is seventeen times the size of the Library of Congress print collections
- Instant messaging generates five billion messages a day
- This is 274 terabytes a year

How much information is being produced?

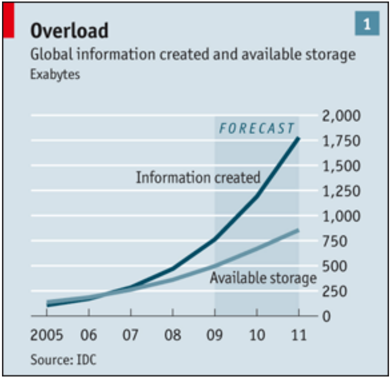
- Email generates about 400,000 terabytes of new information each year worldwide
- P2P file exchange on the Internet is growing rapidly
- That was 2003, what do we have today?
- In 1993: 100.000G transferred over the Internet per year

How much information is being produced?

- Email generates about 400,000 terabytes of new information each year worldwide
- P2P file exchange on the Internet is growing rapidly
- That was 2003, what do we have today?
- In 1993: 100.000G transferred over the Internet per year
- In 2008: 100.000G transferred over the Internet in a *second*

How much information is being produced?

- Data, data everywhere: http://www.economist.com/node/15557443



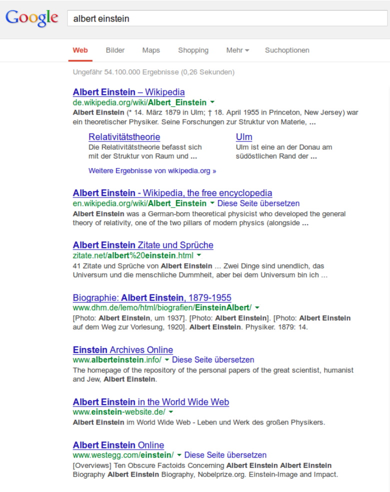
How much information is being produced?

- We are producing more data than we are able to store
- We should extract and describe useful data
- We can store that data
- We should also try to predict future data

Knowledge discovery

- Discover patterns and models in data. These patterns need to be:
  1. (i) Valid: hold for new data with high probability
  2. (ii) Useful: we can base further actions on them
  3. (iii) Unexpected: non-obvious
  4. (iv) Understandable: humans can interpret them

Examples



albert einstein

Ungefähr 54.100.000 Ergebnisse (0,26 Sekunden)

Albert Einstein – Wikipedia
de.wikipedia.org/wiki/**Albert_Einstein**
Albert Einstein (* 14. März 1879 in Ulm; † 18. April 1955 in Princeton, New Jersey) war ein theoretischer Physiker. Seine Forschungen zur Struktur von Materie, ...

Relativitätstheorie
Die Relativitätstheorie befasst sich mit der Struktur von Raum und ...

Ulm
Ulm ist eine an der Donau am südöstlichen Rand der ...

Weitere Ergebnisse von wikipedia.org »

Albert Einstein - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Albert_Einstein** Diese Seite übersetzen
Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics (alongside ...

Albert Einstein Zitate und Sprüche
zitate.net/**albert**%20**einstein**.html
41 Zitate und Sprüche von **Albert Einstein** ... Zwei Dinge sind unendlich, das Universum und die menschliche Dummheit, aber bei dem Universum bin ich ...

Biographie: **Albert Einstein**, 1879-1955
www.dhm.de/lemo/html/biografien/**EinsteinAlbert**/
[Photo: **Albert Einstein**, um 1937]. [Photo: **Albert Einstein**]. [Photo: **Albert Einstein** auf dem Weg zur Vorlesung, 1920]. **Albert Einstein**. Physiker. 1879: 14.

Einstein Archives Online
www.**alberteinstein**.info/ Diese Seite übersetzen
The homepage of the repository of the personal papers of the great scientist, humanist and Jew, **Albert Einstein**.

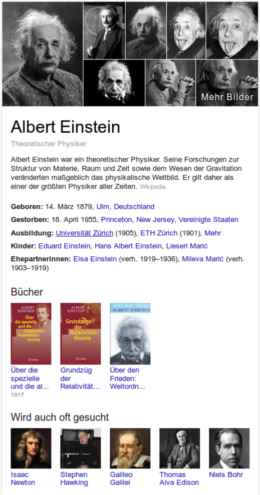
Albert Einstein in the World Wide Web
www.**einstein**-website.de/
Albert Einstein im World Wide Web - Leben und Werk des großen Physikers.

Albert Einstein Online
www.westegg.com/**einstein**/ Diese Seite übersetzen
[Overviews] Ten Obscure Factoids Concerning **Albert Einstein Albert Einstein** Biography **Albert Einstein** Biography, Nobelprize.org. Einstein-Image and Impact.

Examples

- PageRank
  - (i) Valid: holds for all new data
  - (ii) Useful: we base rankings of search results on PageRank
  - (iii) Unexpected: non-obvious and non-trivial to calculate
  - (iv) Understandable: popularity, importance of Web pages, etc.

Examples



Albert Einstein

Theoretischer Physiker

Albert Einstein war ein theoretischer Physiker. Seine Forschungen zur Struktur von Materie, Raum und Zeit sowie dem Wesen der Gravitation veränderten maßgeblich das physikalische Weltbild. Er gilt daher als einer der größten Physiker aller Zeiten. Wikipedia

Geboren: 14. März 1879, Ulm, Deutschland

Gestorben: 18. April 1955, Princeton, New Jersey, Vereinigte Staaten

Ausbildung: Universität Zürich (1905), ETH Zürich (1901), Mehr

Kinder: Eduard Einstein, Hans Albert Einstein, Lieserl Marić

Ehepartnerinnen: Elsa Einstein (verh. 1919–1936), Mileva Marić (verh. 1903–1919)

Bücher

Über die spezielle und die al... 1917

Grundzüg der Relativität...

Über den Frieden: Weltordn...

Wird auch oft gesucht

Isaac Newton

Stephen Hawking

Galileo Galilei

Thomas Alva Edison

Niels Bohr

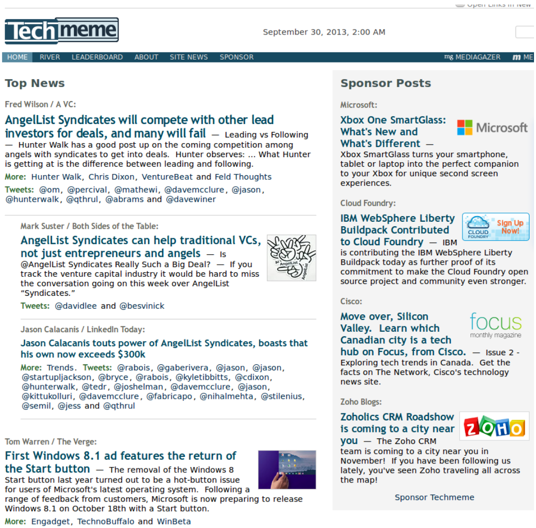
Examples

- Knowledge Graph
  - (i) Valid: holds for new data with high probability
  - (ii) Useful: users can explore connections between concepts
  - (iii) Unexpected: non-obvious and non-trivial
  - (iv) Understandable: related, similar, etc

Examples

- https://www.google.at/#q=albert+einstein
- http://www.youtube.com/watch?v=mmQl6VGvX-c

Examples



September 30, 2013, 2:00 AM

Top News

Fred Wilson / A VC:

AngelList Syndicates will compete with other lead investors for deals, and many will fail — Leading vs Following — Hunter Walk has a good post up on the coming competition among angels with syndicates to get into deals. Hunter observes: ... What Hunter is getting at is the difference between leading and following.

More: Hunter Walk, Chris Dixon, VentureBeat and Feld Thoughts

Tweets: @om, @percival, @mathewi, @davemcclure, @jason, @hunterwalk, @qthrul, @abrams and @davewiner

Mark Suster / Both Sides of the Table:

AngelList Syndicates can help traditional VCs, not just entrepreneurs and angels — Is @AngelList Syndicates Really Such a Big Deal? — If you track the venture capital industry it would be hard to miss the conversation going on this week over AngelList "Syndicates."

Tweets: @davidlee and @besvinick

Jason Calacanis / LinkedIn Today:

Jason Calacanis touts power of AngelList Syndicates, boasts that his own now exceeds $300k

More: Trends. Tweets: @rabois, @gaberivera, @jason, @jason, @startupljackson, @bryce, @rabois, @kyletibbitts, @cdixon, @hunterwalk, @tedr, @joshelman, @davemcclure, @jason, @kittukolluri, @davemcclure, @fabricapo, @nihalmehta, @stilenius, @semil, @jess and @qthrul

Tom Warren / The Verge:

First Windows 8.1 ad features the return of the Start button — The removal of the Windows 8 Start button last year turned out to be a hot-button issue for users of Microsoft's latest operating system. Following a range of feedback from customers, Microsoft is now preparing to release Windows 8.1 on October 18th with a Start button.

More: Engadget, TechnoBuffalo and WinBeta

Sponsor Posts

Microsoft:

Xbox One SmartGlass: What's New and What's Different — Xbox SmartGlass turns your smartphone, tablet or laptop into the perfect companion to your Xbox for unique second screen experiences.

Cloud Foundry:

IBM WebSphere Liberty Buildpack Contributed to Cloud Foundry — IBM is contributing the IBM WebSphere Liberty Buildpack today as further proof of its commitment to make the Cloud Foundry open source project and community even stronger.

Cisco:

Move over, Silicon Valley. Learn which Canadian city is a tech hub on Focus, from Cisco. — Issue 2 - Exploring tech trends in Canada. Get the facts on The Network, Cisco's technology news site.

Zoho Blogs:

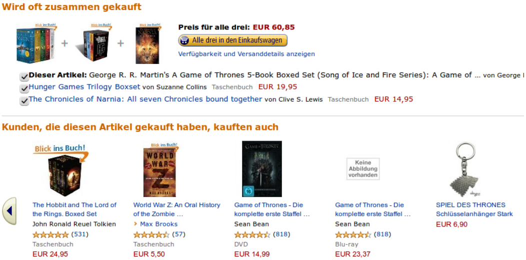
Zoholics CRM Roadshow is coming to a city near you — The Zoho CRM team is coming to a city near you in November! If you have been following us lately, you've seen Zoho traveling all across the map!

Sponsor Techmeme

Examples

- TechMeme
  (i) Valid: holds for new data with high probability
  (ii) Useful: users can find and explore the news about technology
  (iii) Unexpected: non-obvious and non-trivial
  (iv) Understandable: summaries of tech news, etc.

Examples



Wird oft zusammen gekauft

Preis für alle drei: EUR 60,85

Verfügbarkeit und Versanddetails anzeigen

☑ **Dieser Artikel:** George R. R. Martin's A Game of Thrones 5-Book Boxed Set (Song of Ice and Fire Series): A Game of ... von George I

☑ Hunger Games Trilogy Boxset von Suzanne Collins Taschenbuch EUR 19,95

☑ The Chronicles of Narnia: All seven Chronicles bound together von Clive S. Lewis Taschenbuch EUR 14,95

Kunden, die diesen Artikel gekauft haben, kauften auch

The Hobbit and The Lord of the Rings. Boxed Set
John Ronald Reuel Tolkien
(531)
Taschenbuch
EUR 24,95

World War Z: An Oral History of the Zombie ...
› Max Brooks
(57)
Taschenbuch
EUR 5,50

Game of Thrones - Die komplette erste Staffel ...
Sean Bean
(818)
DVD
EUR 14,99

Game of Thrones - Die komplette erste Staffel ...
Sean Bean
(818)
Blu-ray
EUR 23,37

SPIEL DES THRONES Schlüsselanhänger Stark
EUR 6,90

Examples

- Amazon product recommendations
  (i) Valid: holds for new data with high probability
  (ii) Useful: users can find and explore new products
  (iii) Unexpected: non-obvious and non-trivial
  (iv) Understandable: related articles, etc.

Examples

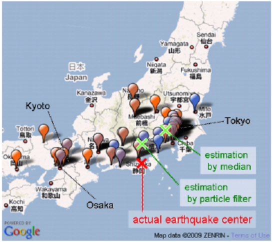


Figure 9: Earthquake location estimation based on tweets. Balloons show the tweets on the earthquake. The cross shows the earthquake center. Red represents early tweets; blue represents later tweets.

Examples

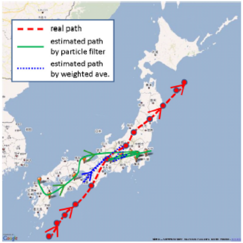
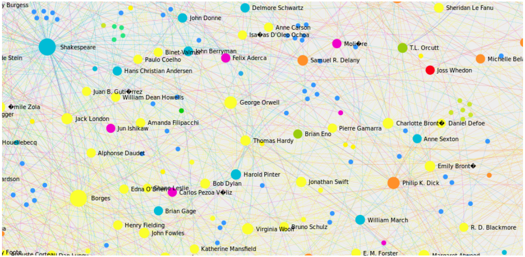

real path
estimated path by particle filter
estimated path by weighted ave.

Figure 10: Typhoon trajectory estimation based on tweets.

Examples

- Twitter earthquake and typhoon prediction
  1. (i) Valid: holds for new data with high probability
  2. (ii) Useful: can save lives
  3. (iii) Unexpected: non-obvious and non-trivial
  4. (iv) Understandable: trajectories of typhoons, positions of earthquakes

Examples


y Burgess
Delmore Schwartz
Sheridan Le Fanu
John Donne
Anne Carson
Isa�as D'Oleo Ochoa
Shakespeare
Moli�re
T.L. Orcutt
Binet-Valmer
John Berryman
le Stein
Paulo Coelho
Felix Aderca
Samuel R. Delany
Michelle Bela
Hans Christian Andersen
Joss Whedon
Juan B. Guti�rrez
William Dean Howells
George Orwell
�mile Zola
gger
Jack London
Amanda Filipacchi
Jun Ishikaw
Charlotte Bront�
Daniel Defoe
Pierre Gamarra
Brian Eno
Thomas Hardy
Anne Sexton
Houellebecq
Alphonse Daudet
Emily Bront�
Harold Pinter
ardson
Jonathan Swift
Bob Dylan
Philip K. Dick
Edna O'Brien
Shane Leslie
Carlos Pezoa V�liz
Borges
Brian Gage
William March
Henry Fielding
a
Virginia Woolf
Bruno Schulz
R. D. Blackmore
John Fowles
Katherine Mansfield
y Foote
Auguste Corteau
Dan Lungu
E. M. Forster
Margaret Atwood

Examples

- Graph of ideas http://zoom.it/l3dq
  - (i) Valid: holds for new data with high probability
  - (ii) Useful: development of ideas
  - (iii) Unexpected: non-obvious and non-trivial
  - (iv) Understandable: who influenced whom, etc

Knowledge discovery vs. data mining

- Knowledge discovery refers to the entire process, of which knowledge is the end-product
- It is iterative and interactive
- Data mining refers to a specific step in this process
- It is the step consisting of applying data analysis and discovery algorithms that produce a particular enumeration of patterns over data
- Additional steps are necessary to ensure that the process produces useful knowledge

Steps in the knowledge discovery process

1. Developing an **understanding of the application domain** and the relevant prior knowledge and identifying the goal of the KDD process from the customers viewpoint
2. Creating a target data set: **selecting a data set** or focusing on a subset of variables or data samples on which discovery is to be performed
3. **Data cleaning and preprocessing**: basic operations such as the removal of noise. If appropriate collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes

Steps in the knowledge discovery process

4. Data reduction and projection: **finding useful features to represent the data** depending on the goal of the task. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data
5. Matching the goals of the KDD process step to a particular **data mining method** e.g. summarization, classification, regression, clustering, etc
6. Choosing the **data mining algorithms**: selecting methods to be used for searching for patterns in the data. This includes deciding which models and parameters may be appropriate e.g. models for categorical data are different than models on vectors over the reals. Matching a particular data mining method with the overall criteria of the KDD process e.g. the enduser may be more interested in understanding the model than its predictive capabilities

Steps in the knowledge discovery process

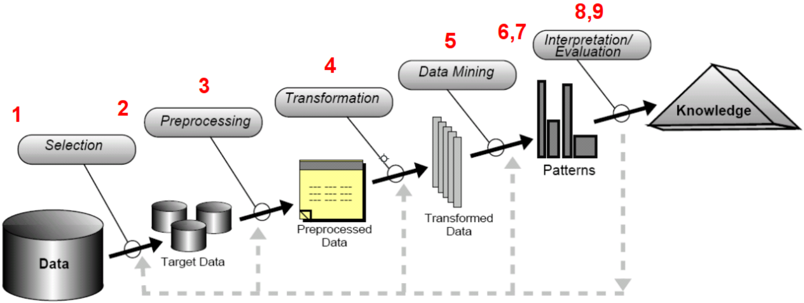
7. **Data mining searching for patterns** of interest in a particular representational form or a set of such representations, classification rules or trees, regression, clustering and so forth. The user can significantly aid the data mining method by correctly performing the preceding steps
8. **Interpreting mined patterns**: possibly return to any of the steps for further iteration. This step can also involve visualization of the extracted patterns, models or visualization of the data given the extracted models
9. **Consolidating discovered knowledge**: incorporating this knowledge into another system for further action or simply documenting it and reporting it to interested parties. This also includes checking for and resolving potential conflicts with previously believed or extracted knowledge

Steps in the knowledge discovery process

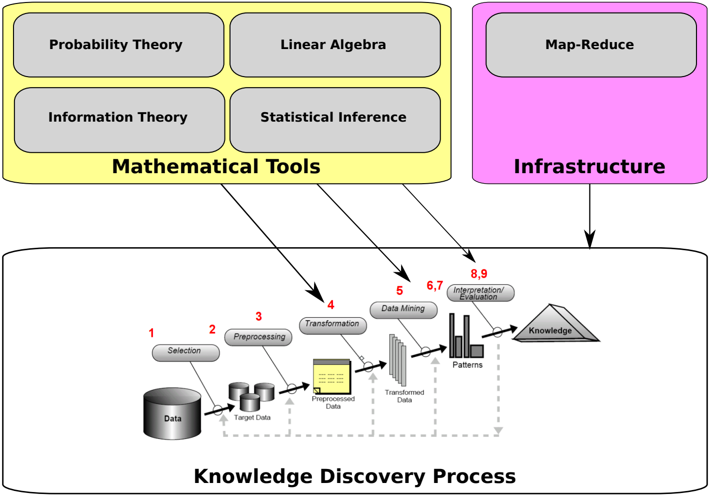
Reading!

Knowledge Discovery and Data Mining: Towards a Unifying Framework (1996) Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth

Steps in the knowledge discovery process



Big picture: KDDM



Probability Theory

Linear Algebra

Information Theory

Statistical Inference

Mathematical Tools

Map-Reduce

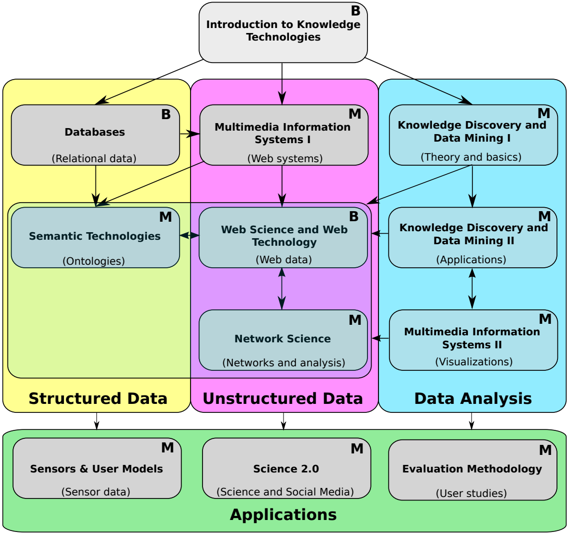
Infrastructure

Knowledge Discovery Process

Teaching @ KTI and KDDM1

- KDDM1: Basics, theory and KDD process until step 8 (no visualization, no interpretation). We will only shortly mention graph-based data mining
- KDDM2: Implementation and practice of the theory from KDDM1
- MMIS2: KDD process steps 8 and 9 (interpretation and visualization)
- Network Science: graph and networks mining

Teaching @ KTI

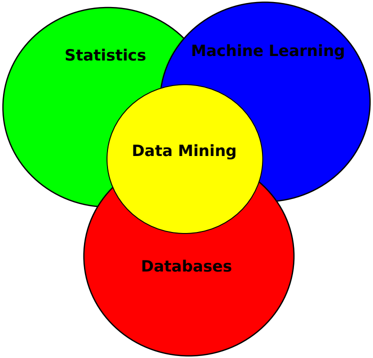


+ Projects, Bachelor Thesis, Master Projects, Master Thesis, PhD Thesis

Data mining and other fields

- Data mining overlaps with
  1. (i) Databases: Large-scale data, simple queries
  2. (ii) Machine learning: Small data, complex models, model parameters
  3. (iii) Statistics: Theory, predictive models, no algorithms
- Data mining: **Algorithms**, simple and predictive models, large-scale data

Data mining and other fields



KDDM1: some thoughts

- Big data, data science, . . .
- Many buzzwords
- But in the end:
  1. You need to know how to program and how to develop software systems
  2. You need to understand the math behind it: linear algebra, probability, statistics
- If you know both of these you are in the top 10% of developers in the field ;)

KDDM1: some highlights

- Bayesian inference: update your prior knowledge with new evidence
- Recommender systems: decompose the matrix and find out what your users really like :)
- PCA: reduce the dimensionality of the data
- LSA: analyze relationships between set of documents with linear algebra