

ECAC

Extracção de Conhecimento e Aprendizagem Computacional
Knowledge Extraction and Machine Learning

José Luís Borges (DEIG – L215) - jlborges@fe.up.pt

João Mendes Moreira (DEI – I023) - jmoreira@fe.up.pt

1

Knowledge Extraction

Objectives

To provide the students with knowledge so that they can use

analysis and

knowledge extraction techniques

on large data quantities' patterns.

2

Classes and Evaluation

- Lecture + lab session (1.5h+1.5h)
- **Integrated Masters**
 - Distributed evaluation with final exam
 - Course work: 50% (teams of two students)
 - Final exam: 50%
- **Doctoral Programmes**
 - As above or
 - only by an extended individual course-work

3

Classes' tentative plan

week	week					Part I (Lecture)	Part II (Lab Session)
year	term	day	month			17:00 - 18:30	18:40 - 20:00
37	1	14	Set	JMB		Introduction to Data Mining	
38	2	21	Set	JMB		Descriptive Statistics	lab session
39	3	28	Set	JMB		Data preparation	lab session
40	4	5	Out			Feriado - implantação da República Portuguesa em 1910	
41	5	12	Out	JMB		Association Rules	lab session
42	6	19	Out	JMB		Clustering	lab session
43	7	26	Out			Semana da FEUP	
44	8	2	Nov	JMB	Progress report	Clustering	lab session
45	9	9	Nov	JMM		Prediction: Classification and regression	lab session
46	10	16	Nov	JMM		Decision trees	lab session
47	11	23	Nov	JMM		Decision trees pruning, rules and evaluation	lab session
48	12	30	Nov	JMM		Neural networks and support vector machines	lab session
49	13	7	Dez	JMM	Final report	Ensemble learning	lab session
50	14	14	Dez		Presentations	Students' presentations of the course works.	lab session

4

companion website

- <http://paginas.fe.up.pt/~ec/>
 - slides
 - links to useful resources
 - information about the course work

IMPORTANT DATES

2nd November – Progress report

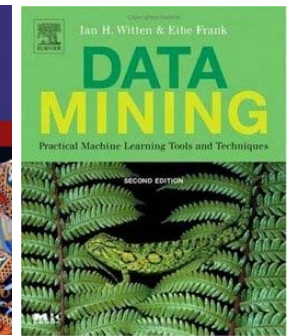
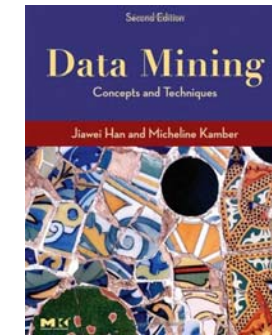
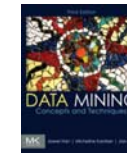
7th December – Final report

14th December - Presentations

5

Bibliography

- *Data Mining: Concepts and Techniques*. Jiawei Han and Micheline Kamber. Morgan Kaufmann Publishers
- *Data Mining: Practical Machine learning tools with JAVA implementations*. Ian H. Witten and Eibe Frank. Morgan Kaufmann Publishers



Introduction to Data Mining

7

Motivation: "Necessity is the Mother of Invention"

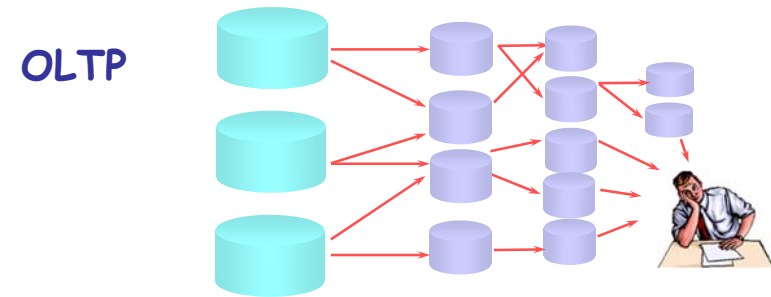
- Data explosion problem
 - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- There is a tremendous **increase in the amount of data** recorded and stored on digital media
 - We are producing over two exabites (10^{18}) of data per year
 - Storage capacity, for a fixed price, appears to be doubling approximately every 9 months

8

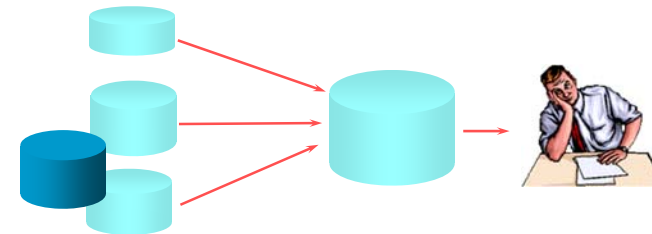
Motivation: "Necessity is the Mother of Invention"

- We are drowning in data, but starving for knowledge!
 - "The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as none at all." (W.H. Auden)
- Solution: Data warehousing and data mining
 - Data warehousing and On-Line Analytical Processing (OLAP)
 - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

9



Data Warehouse DSS (OLAP)



10

Big Data Examples

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces 1 **Gigabit/second** of astronomical data over a 25-day observation session
 - storage and analysis a big problem
- AT&T handles billions of calls per day
 - so much data, it cannot be all stored -- analysis has to be done "on the fly", on streaming data
- Web
 - Alexa internet archive: 7 years of data, 500 TB
 - Google searches 4+ Billion pages, many hundreds TB
 - IBM WebFountain, 160 TB (2003)
 - Internet Archive (www.archive.org), ~ 300 TB

11

Data Growth Rate Estimates

- Data stored in world's databases doubles every 20 months
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human
- Knowledge Discovery is **NEEDED** to make sense and use of data.

12

“Every time the amount of data increases by a factor of ten, we should totally **rethink the way we analyze it**”

Jerome Friedman, Data Mining and Statistics: What's the Connection (paper 1997)

13

Data Mining

- Data Mining query differs from Database query
 - Query not well formulated
 - Data in many sources
 - Discover **actionable** patterns & rules
- Traditional Analysis
 - Did sales of product X increase in Nov.?
 - Do sales of product X decrease when there is a promotion on product Y?
- Data mining is result oriented
 - What are the factors that determine sales of product X?

14

Data Mining

- Traditional analysis is incremental
 - Does billing level affect turnover?
 - Does location affect turnover?
 - Analyst builds model step by step
- Data Mining is result oriented
 - Identify the factors and predict turnover

15

“The key in business is to know something that nobody else knows.”



HULTON-DEUTSCH COLL

— Aristotle Onassis



LUCINDA DOUGLAS-MENZIES

“To understand is to perceive patterns.”


— Sir Isaiah Berlin

16

An Application Example

- A person buys a book (product) at Amazon.com
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
 - customers who bought "Advances in Knowledge Discovery and Data Mining", also bought "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations"
- Recommendation program is quite successful

17



Click to LOOK INSIDE!

Data Mining, Second Edition: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (Hardcover)
 by [Jiawei Han](#) (Author), [Micheline Kamber](#) (Author), [Jian Pei](#) (Author)
 ★★★★★ (2 customer reviews)

RRP: ~~£34.99~~
 Price: **£33.24** & this item **Delivered FREE in the UK** with Super Saver Delivery. [See details and conditions](#)
 You Save: **£1.75 (5%)**

In stock.
 Dispatched from and sold by **Amazon.co.uk**. Gift-wrap available.

Only 3 left in stock--order soon (more on the way).

Want guaranteed delivery by Friday, September 25? Order it in the next 9 hours and 33 minutes, and choose **Express** delivery at checkout. [See Details](#)

24 new from £28.99 **8 used** from £29.02

Other Editions: RRP: ~~£34.99~~ Our Price: **£33.24** Other Offers:
 Hardcover: ~~£34.99~~ **£33.24** 13 used & new from £8.90

This title is part of our **Up to 35% Off 1000s of Textbooks** offer.

Frequently Bought Together

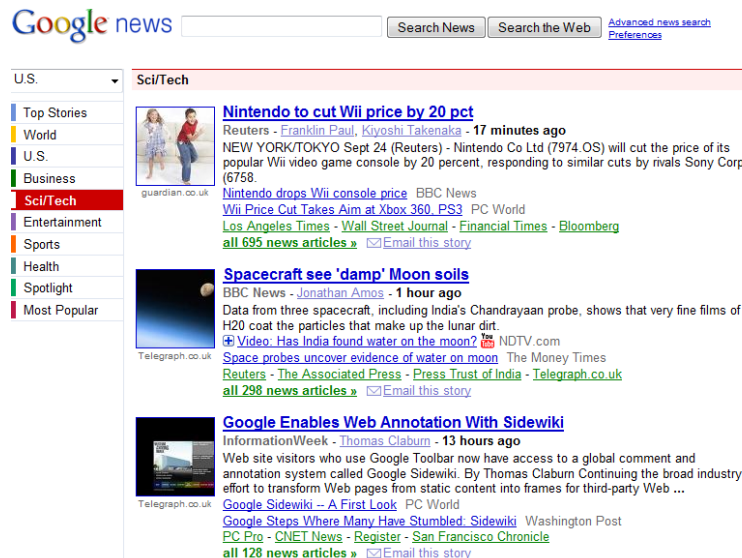
 +  +  **Price For All Three: £123.37**
[Add all three to Basket](#)
[Show availability and shipping details](#)

This Item: Data Mining, Second Edition: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han

Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten

18


Google news example





Google news [Advanced news search](#) [Preferences](#)

U.S. **Sci/Tech**

Top Stories
 World
 U.S.
 Business
Sci/Tech
 Entertainment
 Sports
 Health
 Spotlight
 Most Popular

 **Nintendo to cut Wii price by 20 pct**
 Reuters - [Franklin Paul](#), [Kiyoshi Takenaka](#) - **17 minutes ago**
 NEW YORK/TOKYO Sept 24 (Reuters) - Nintendo Co Ltd (7974.OS) will cut the price of its popular Wii video game console by 20 percent, responding to similar cuts by rivals Sony Corp (6758).
[Nintendo drops Wii console price](#) BBC News
[Wii Price Cut Takes Aim at Xbox 360, PS3](#) PC World
[Los Angeles Times](#) - [Wall Street Journal](#) - [Financial Times](#) - [Bloomberg](#)
[all 695 news articles »](#) [Email this story](#)

 **Spacecraft see 'damp' Moon soils**
 BBC News - [Jonathan Amos](#) - **1 hour ago**
 Data from three spacecraft, including India's Chandrayaan probe, shows that very fine films of H2O coat the particles that make up the lunar dirt.
[Video: Has India found water on the moon?](#) NDTV.com
[Space probes uncover evidence of water on moon](#) The Money Times
[Reuters](#) - [The Associated Press](#) - [Press Trust of India](#) - [Telegraph.co.uk](#)
[all 298 news articles »](#) [Email this story](#)

 **Google Enables Web Annotation With Sidewiki**
 InformationWeek - [Thomas Claburn](#) - **13 hours ago**
 Web site visitors who use Google Toolbar now have access to a global comment and annotation system called Google Sidewiki. By Thomas Claburn Continuing the broad industry effort to transform Web pages from static content into frames for third-party Web ...
[Google Sidewiki -- A First Look](#) PC World
[Google Steps Where Many Have Stumbled: Sidewiki](#) Washington Post
[PC Pro](#) - [CNET News](#) - [Register](#) - [San Francisco Chronicle](#)
[all 128 news articles »](#) [Email this story](#)

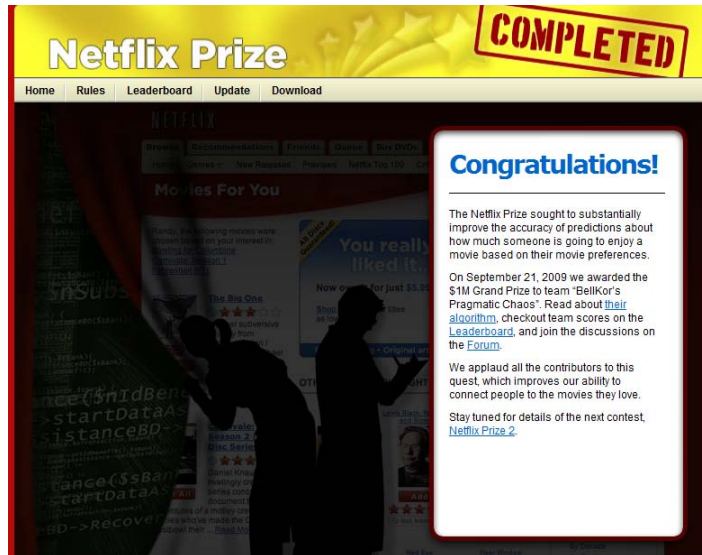
19

Another Application Example

- Netflix prize
- <http://www.netflixprize.com/>
- The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set.
- ~~You can win~~ could have won **one million dollars**

20

Netflix



21

Netflix - Some Details

- Dataset with 100 million date stamped movie ratings performed by anonymous Netflix customers (Dec 1999 and Dec 2005), about 480,189 users and 7,770 movies.
- A Hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user. The remaining data made up the training set.
- The Hold-out set was randomly split three ways, into subsets called Probe, Quiz, and Test. The labels were attached to the Probe. The Quiz and Test sets made up an evaluation set, which is known as the Qualifying set, that competitors were required to predict ratings for. Once a competitor submits predictions, the prizemaster returns the error achieved on the Quiz set on a public leaderboard.
- The winner of the prize is the one that scores best on the Test set, and those scores were never disclosed by Netflix.

22

Netflix - Lessons...

- The biggest lesson learned, according to members of the two top teams, **was the power of collaboration**. It was not a single insight, algorithm or concept that allowed both teams to surpass the goal Netflix.
- Instead, they say, the formula for success was to bring together people with complementary skills and combine different methods of problem-solving.
- When BellKor's announced that it had passed the 10 percent threshold, it set off a 30-day race, under contest rules, for other teams to try to best it. That led to another round of team-merging by BellKor's leading rivals, who assembled a global consortium of about 30 members, appropriately called the Ensemble.

23

Problems Suitable for Data-Mining

- The business problem is unstructured
- Accurate prediction is more important than the explanation
- Have accessible, sufficient, and relevant data
- The data are highly heterogeneous with a large percentage of outliers, leverage points, and missing values
- Require knowledge-based decisions
- Have a changing environment
- Have sub-optimal current methods
- Provides high payoff for the right decisions!
- Privacy considerations important if personal data is involved

24

What is Data Mining?

- **Knowledge Discovery** in Databases
 - Is the non-trivial process of identifying
 - implicit (by contrast to explicit)
 - valid (patterns should be valid on new data)
 - novel (novelty can be measured by comparing to expected values)
 - potentially useful (should lead to useful actions)
 - understandable (to humans)
 - patterns in data
- **Data Mining**
 - Is a **step** in the KDD process

25

What Is Data Mining?

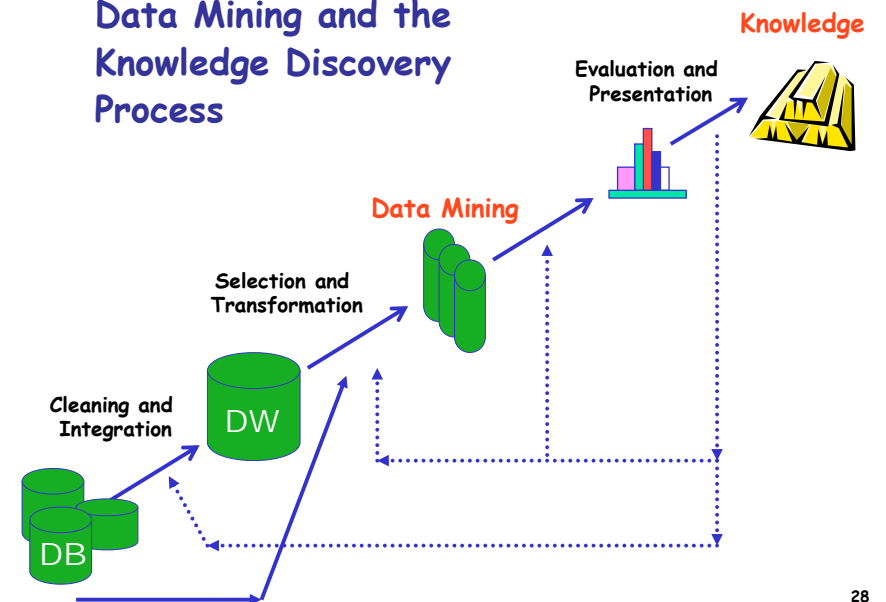
- Alternative names:
 - **Data Mining**: a misnomer?
(knowledge mining from data?)
 - Knowledge discovery (mining) in databases (KDD),
 - knowledge extraction,
 - data/pattern analysis,
 - data archeology,
 - data dredging,
 - information harvesting,
 - business intelligence, etc.



26

KDD Process

Data Mining and the Knowledge Discovery Process



27

28

Steps of a KDD Process

- **Data cleaning:** missing values, noisy data, and inconsistent data
- **Data integration:** merging data from multiple data stores
- **Data selection:** select the data relevant to the analysis
- **Data transformation:** aggregation (daily sales to weekly or monthly sales) or generalisation (street to city; age to young, middle age and senior)
- **Data mining:** apply intelligent methods to extract patterns
- **Pattern evaluation:** interesting patterns should contradict the user's belief or confirm a hypothesis the user wished to validate
- **Knowledge presentation:** visualisation and representation techniques to present the mined knowledge to the users

29

More on the KDD Process

60 to 80% of the KDD effort is about **preparing the data** and the remaining 20% is about mining

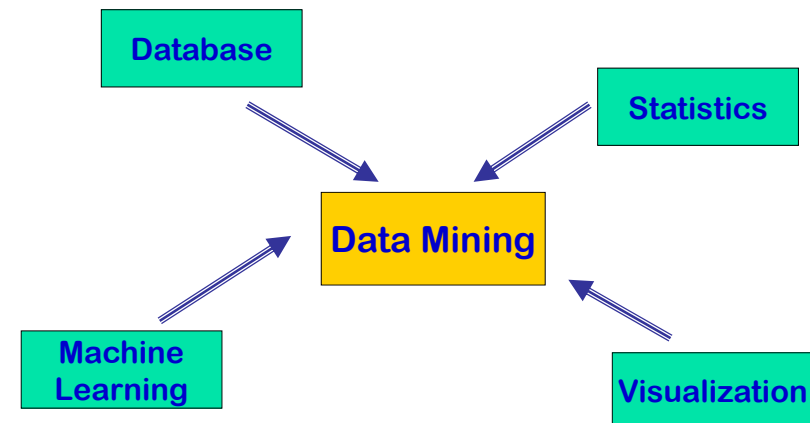
30

More on the KDD Process

- A data mining project should always start with an analysis of the data with **traditional query tools**
 - 80% of the interesting information can be **extracted using SQL**
 - how many transactions per month include item number 15?
 - show me all the items purchased by Sandy Smith.
 - 20% of hidden information requires more **advanced techniques**
 - which items are frequently purchased together by my customers?
 - how should I classify my customers in order to decide whether future loan applicants will be given a loan or not?

31

Data Mining: Related Fields



32

Statistics, Machine Learning and Data Mining

- Statistics
 - more theory-based
 - more focused on testing hypotheses
- Machine learning
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics - areas not part of data mining
- Data Mining and Knowledge Discovery
 - integrates theory and heuristics
 - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

33

More on Data Mining

- Data mining is sometimes also referred to as **secondary** data analysis
- Very large datasets have problems associated with them beyond what is traditionally considered by statisticians
 - Many statistical methods require some type of **exhaustive search**
- Many of the techniques & algorithms used are shared by both statisticians and data miners
- While data mining aims at pattern detection statistics aims at assessing the reality of a pattern
 - (example: finding a cluster of people suffering a particular disease which the doctor will assess if it is random or not)

34

DM and Non-DM examples

Data Mining:

-Certain names are more prevalent in certain US locations

(O'Brien, O'Rourke, O'Reilly... in Boston area)

-Group together similar documents returned by search engine according to their context

(e.g. Amazon rainforest, Amazon.com, etc.)

• NOT Data Mining:

-Look up phone number in phone directory

-Query a Web search engine for information about "Amazon"

35

Rhine Paradox

- A great example of how not to conduct scientific research.
- David Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception (ESP).
- He devised an experiment where subjects were asked to guess 10 hidden cards --- **red** or **blue**.
- He discovered that almost 1 in 1000 had ESP --- they were able to get all 10 right!

36

Rhine Paradox

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?

You shouldn't tell people that they have ESP: it causes them to lose it

37

Rhine Paradox

- What has really happened:

There are 1024 combinations of red and blue combinations of red and blue of length 10.

Thus with probability 0.98 at least one person (in 1000) will guess the sequence of red blue correctly

38

Data Mining Applications

39

Data Mining - Applications

- Market analysis and management
 - Target marketing, customer relation management, market basket analysis, cross selling, market segmentation
 - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
 - Determine customer purchasing patterns over time
- Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis, credit scoring

40

Data Mining - Applications

- Fraud detection and management
 - Use **historical data to build models** of fraudulent behavior and use data mining to help identify similar instances
- Examples
 - auto insurance: detect a group of people who stage accidents to collect on insurance
 - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
 - medical insurance: detect professional patients and ring of doctors and ring of references (ex. doc. prescribes expensive drug to a Medicare patient. Patient gets prescription filled, gets drug and sells drug unopened, which is sold back to pharmacy)

41

Fraud Detection and Management

- Detecting inappropriate medical treatment
 - Charging for unnecessary services, e.g. performing \$400,000 worth of heart & lung tests on people suffering from no more than a common cold. These tests are done either by the doctor himself or by associates who are part of the scheme. A more common variant involves administering more expensive blanket screening tests, rather than tests for specific symptoms

42

Fraud Detection and Management

- Detecting telephone fraud
 - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
 - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
 - ex. an inmate in prison has a friend on the outside set up an account at a local abandoned house. Calls are forwarded to inmate's girlfriend three states away. Free calling until phone company shuts down account 90 days later.

43

Other Applications

- Sports
 - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Space Science
 - SKICAT automated the analysis of over 3 Terabytes of image data for a sky survey with 94% accuracy
- Internet Web Surf-Aid
 - Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

44

Other Applications

- Social Web and Networks
- There are a growing number of highly-popular user-centric applications such as blogs, folksonomies, wikis and Web communities that generate a lot of structured and semi-structured information.
 - Ranking of social bookmark search results. Aggregating bookmarks.
 - Models to explain and predict the evolution of social networks
 - Personalized search for social interaction
 - User behaviour prediction
 - Discovering social structures and communities
 - Topic detection and topic trend analysis

45

On going projects I am involved in

- Wine tasting panel data analysis and Studying the impact of weather changes on wine quality (ADVID)
- Operating room capacity planning and scheduling optimization (CHP / KAIZEN)
- Analysis of in store customer experiences (sensory, and path analysis)
- Visualization of temporal urban mobility data

46

Data Mining Tasks

Association (correlation and causality)

- Multi-dimensional vs. single-dimensional association
- $\text{age}(X, "20..29") \wedge \text{income}(X, "20..29K") \rightarrow \text{buys}(X, "PC")$
[support = 2%, confidence = 60%]
- $\text{buys}(T, "computer") \rightarrow \text{buys}(x, "software")$ [1%, 75%]

47

Data Mining Tasks

Classification and Prediction

- Finding models (functions) that describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on climate, or classify cars based on gas mileage
- Presentation: decision-tree, classification rule, neural network
- Prediction: Predict some unknown or missing numerical values

48

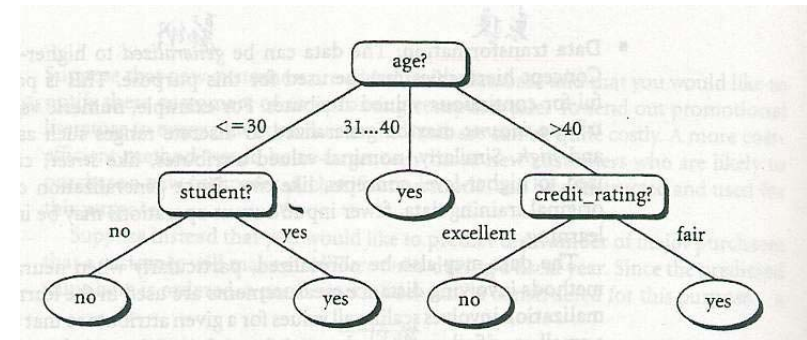
Training Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

This follows an example from Quinlan's ID3

49

Classification: A Decision Tree for "buys_computer"



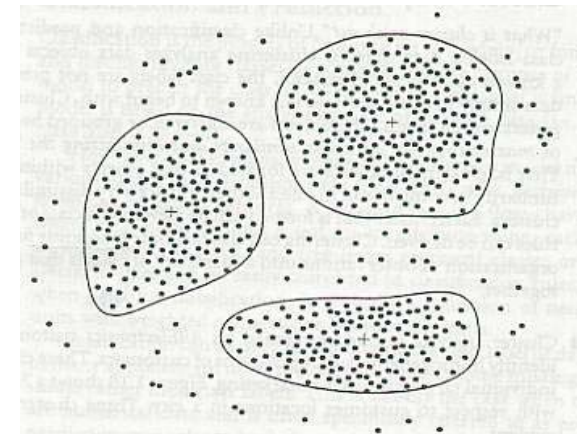
50

Data Mining Tasks

- Cluster analysis
 - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
 - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

51

Cluster Analysis



52

Data Mining Tasks

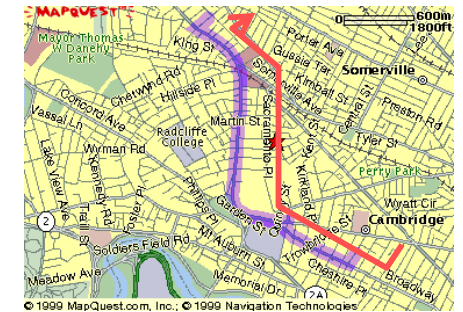
- **Outlier analysis**
 - Outlier: a data object that does not comply with the general behavior of the data
 - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- **Trend and evolution analysis**
 - Trend and deviation: regression analysis
 - Sequential pattern mining, periodicity analysis
 - Similarity-based analysis

53

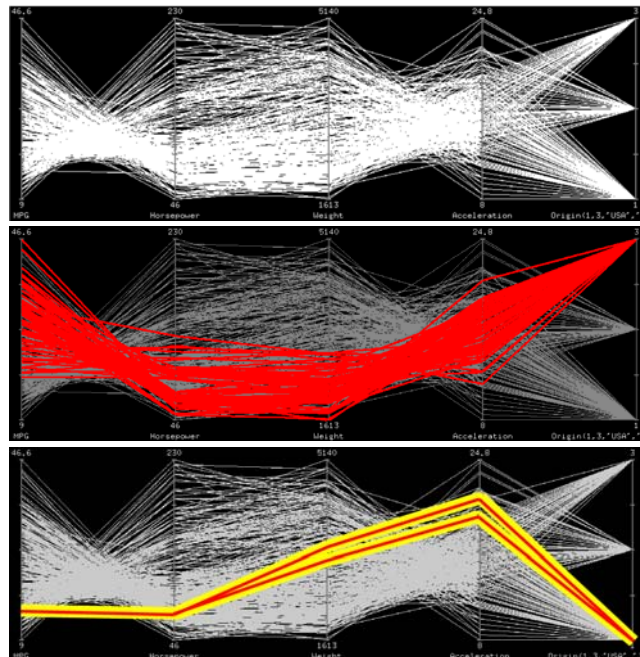
Data Mining Tasks

The Power of Visualization

1. Start out going Southwest on ELLSWORTH AVE Towards BROADWAY by turning right.
2. Turn RIGHT onto BROADWAY.
3. Turn RIGHT onto QUINCY ST.
4. Turn LEFT onto CAMBRIDGE ST.
5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
6. Turn RIGHT onto RUSSELL ST.



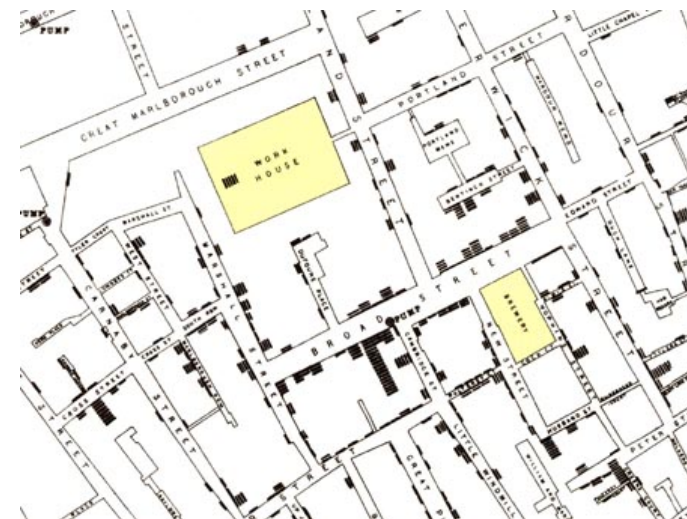
54



http://vis.computer.org/vis2006/Vis2006/Papers/outlier_preserving_focus_context.ppt

55

Visualization for Problem Solving



Cholera Map, 1855

From Visual Explanations by Edward Tufte, Graphics Press, 1997

56

Asia at night



61

Data Mining Methodology

- CRISP - Data Mining Process
- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort to develop framework for data mining tasks
- **C**Ross **I**ndustry - enables Leverage.
- **S**tandard **P**rocess - enables Competition.

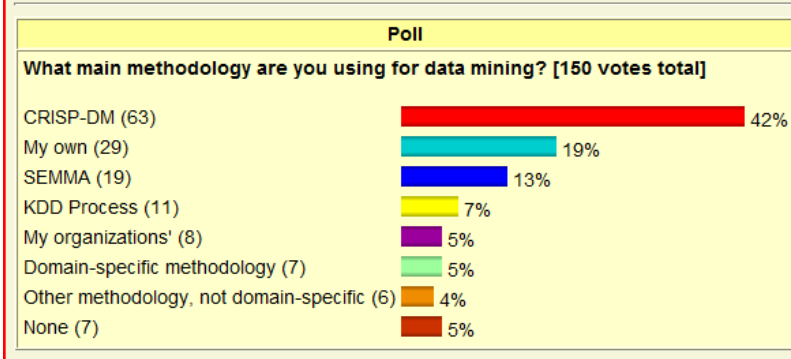
62

CRISP-DM goals

- General Objectives
 - Defining a cross industry data mining process and providing tool support, allowing for cheaper, faster, and more reliable data mining.
 - Widespread adoption of the CRISP-DM process model.
- Detailed Objectives
 - Ensure quality of Data Mining projects results.
 - Reduce skills required for Data Mining.
 - Capture experience for reuse.
 - General purpose (i.e., widely stable across varying applications, for example).
 - and robust (i.e., insensitive to changes in the environment).
 - Tool and technique independent.
 - Tool supportable.

64

KDnuggets : Polls : Data Mining Methodology (Aug 2007)



http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

63

Why Should There be a Standard Process?

- Framework for recording experience
 - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
 - Demonstrates maturity of Data Mining
 - Reduces dependency on "stars"
- Encourage best practices and help to obtain better results

65

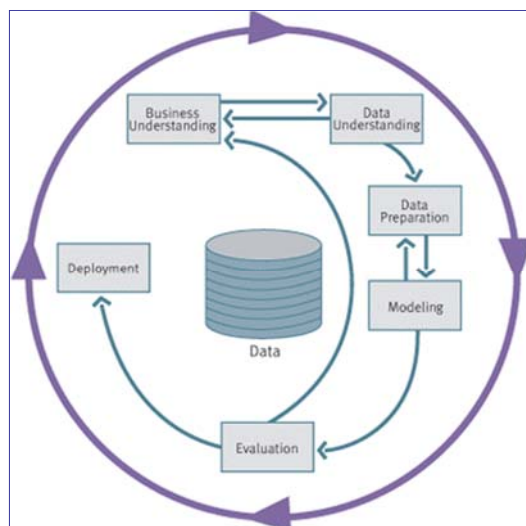
CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
 - As well as technical analysis
- Framework for guidance
- Experience base
 - Templates for Analysis



66

CRISP-DM: Overview



CRISP-DM is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project.

CRISP-DM breaks down the life cycle of a data mining project into six phases.

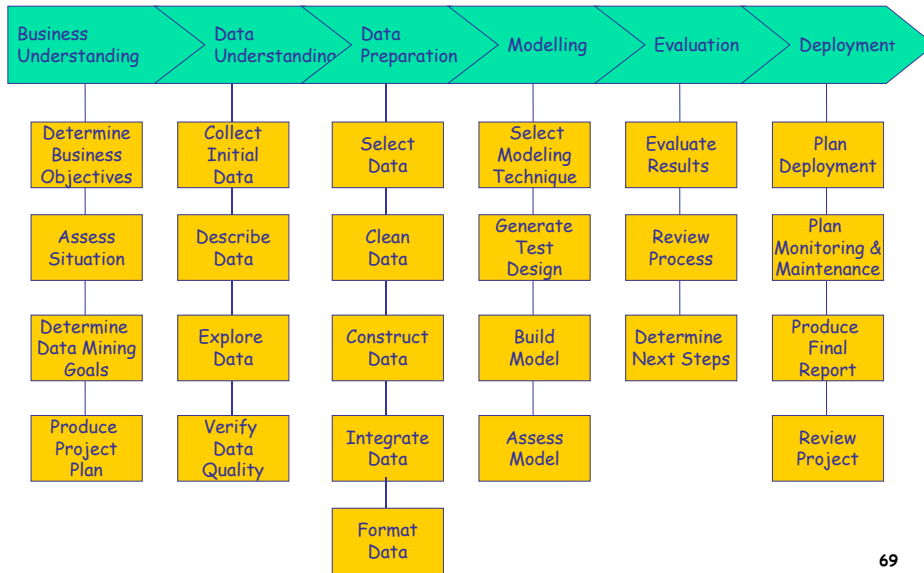
67

CRISP-DM: Phases

- **Business Understanding**
 - Understanding project objectives and requirements; Data mining problem definition.
- **Data Understanding.**
 - Initial data collection and familiarization; Identify data quality issues; Initial, obvious results.
- **Data Preparation**
 - Record and attribute selection; Data cleansing.
- **Modelling**
 - Run the data mining tools.
- **Evaluation**
 - Determine if results meet business objectives; Identify business issues that should have been addressed earlier.
- **Deployment**
 - Put the resulting models into practice; Set up for continuous mining of the data.

68

Phases and Tasks



69

True Legends of KDD

Stories – Beer and Diapers



◆ Diapers and Beer. Most famous example of market basket analysis for the last few years. If you buy diapers, you tend to buy beer.

- T. Blischok headed Terradata's Industry Consulting group.
- K. Heath ran self joins in SQL (1990), trying to find two itemsets that have baby items, which are particularly profitable.
- Found this pattern in their data of 50 stores/90 day period.
- Unlikely to be significant, but it's a nice example that explains associations well.

Ronny Kohavi ICML 1998

70

True Legends of KDD

Stories – Non-actionable Segment

- ◆ A bank discovered a cluster of customers that have left the bank:
 - Older than the average customer.
 - Less likely to have a mortgage.
 - Less likely to have a credit card.

They were also...



(*) From Berry and Linoff's Data Mining techniques book.

Ronny Kohavi ICML 1998

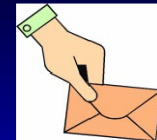
71

True Legends of KDD

Stories – Insurance for Californians

- ◆ A health insurance mailing campaign had 100% response rate from California.

Reason: the mailing never went to California in the first place!



- People who received the offers would pass them to their family members in other states.
- Anyone from California that was in the dataset was there because s/he accepted the insurance.

Ronny Kohavi ICML 1998

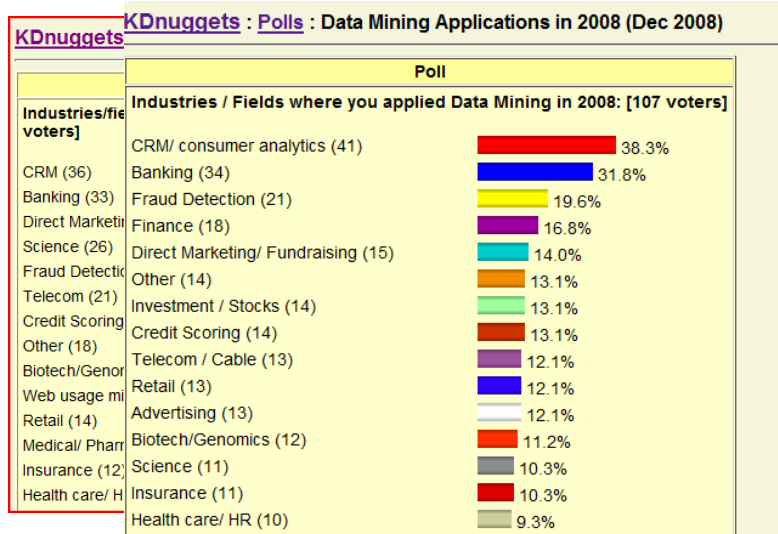
72

KDnuggets

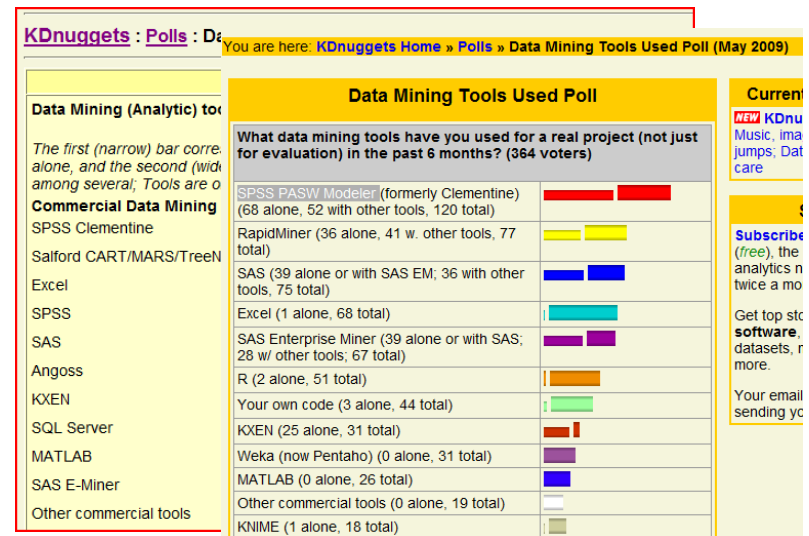
- <http://www.kdnuggets.com/>
 - Is the leading source of information on Data Mining, Web Mining, Knowledge Discovery, and Decision Support Topics, including News, Software, Solutions, Companies, Jobs, Courses, Meetings, Publications, and more.
- KDnuggets News
 - Has been recognized as the #1 e-newsletter for the Data Mining and Knowledge Discovery community

- **Companies**
Consulting, Products
- **Gregory Piatetsky-Shapiro**
Data Mining Consulting
- **Domain-specific Solutions**
CRM, Twitter, Web
- **Datasets**
Competitions, KDD Cup
- **Useful Data Mining / Analytics sites**
Blogs, Twitters, Social
- **KDnuggets Polls**
NEW Data types mined
- **FAQ**
DM Tool Comparison
- ACM SIGKDD**: The Knowledge Discovery and Data Mining Society
- **Webcasts**: live, on-demand
▶ Sep 9: Targeted Business Analytics
- **Courses**
▶ Sep 14-15, Tools for Discovering Patterns in Data
- **Meetings, Conferences**
Sep 7-11, ECML/PKDD-09, Slovenia
- **Publications**
UPDATE Books, Professional books
- **Education**:
on-line, USA, Europe
- **Data Mining Course**
lectures and teaching materials
- **Data Mining Forums**
Open Issues, Beginners, Experts
- Data Mining Crossword**
Can you solve it?

Results of a KDnuggets Poll

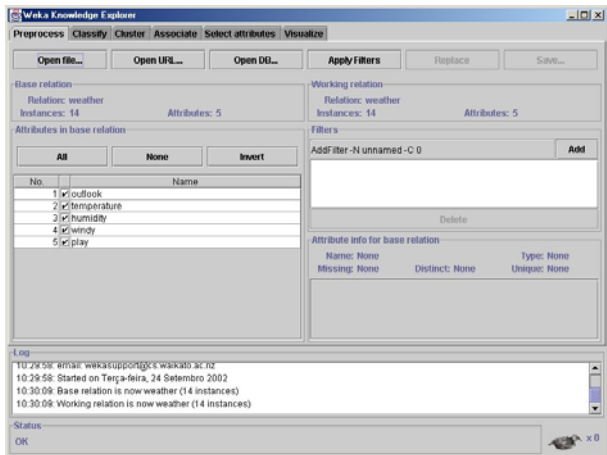


Results of a KDnuggets Poll



Weka 3 - Machine Learning Software in Java

<http://www.cs.waikato.ac.nz/~ml/weka/>

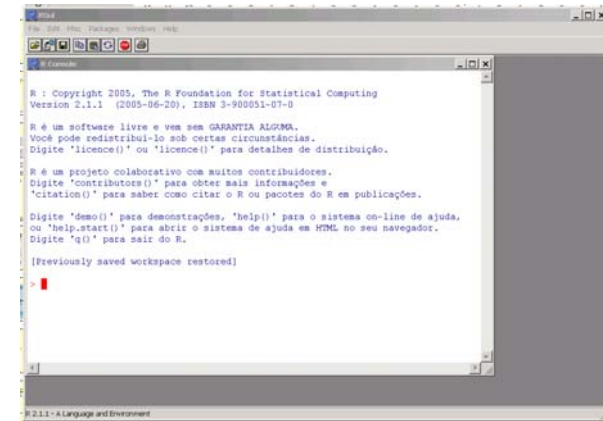


77

R - Project for Statistical Computing

<http://www.r-project.org/>

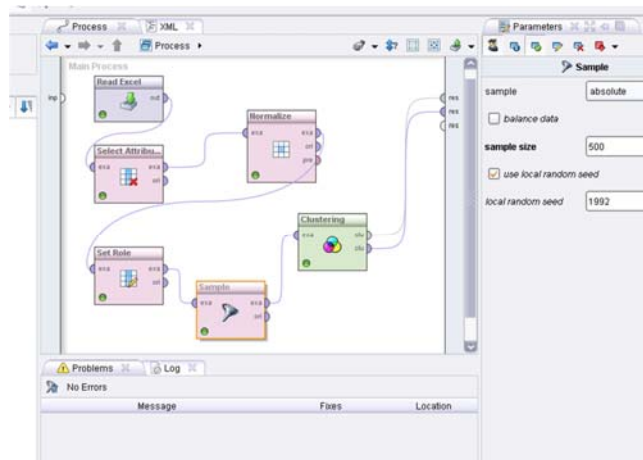
Open source and lots of libraries available.



78

Rapidminer

<http://rapid-i.com/content/view/181/190/>



79

Golden Rules for Data Mining

KDnuggets FAQ - Gregory Piatetsky-Shapiro

- Focus on what is actionable.
- Prepare and clean the data carefully.
- Verify data analysis steps.
- Use multiple data mining and machine learning methods.
- Beware of "false predictors" (also called "information leakers") fields that appear to predict the outcome too well and are actually recording events that happened after the outcome happened. Find and eliminate them.
- If the results are too good to be true, you probably have found false predictors.
- Examine the results carefully and repeat and refine the knowledge discovery process until you are confident.
- Did I emphasize that you should be beware of "false predictors"?

80

A Brief History of Data Mining Society

- [1989 IJCAI Workshop on Knowledge Discovery in Databases \(Piatetsky-Shapiro\)](#)
 - 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)
- [1998 ACM SIGKDD, SIGKDD'1999-2009 conferences, and SIGKDD Explorations](#)
- [More conferences on data mining](#)
 - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

81

Where to Find References?

- [Data mining and KDD \(SIGKDD member CDROM\):](#)
 - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery
- [Database field \(SIGMOD member CD ROM\):](#)
 - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
 - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- [AI and Machine Learning:](#)
 - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
 - Journals: Machine Learning, Artificial Intelligence, etc.
- [Statistics:](#)
 - Conference proceedings: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- [Visualization:](#)
 - Conference proceedings: CHI, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

82

Books on Data Mining

- **Data Mining: Concepts and Techniques**, Jiawei Han, Micheline Kamber, Morgan Kaufmann - Third Edition ,2011
- **Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations**, Ian H. Witten, Eibe Frank - Morgan Kaufmann, Third Edition, 2011
- **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**, Michael Berry and Gordon Linoff - John Wiley & Sons Inc - Third Edition, 2011
- **Handbook of Statistical Analysis and Data Mining Applications**, R. Nisbet, J. Elder and G. Miner - Academic Press, 2009.

83



Thank you !!!

84