Uniwersytet Ekonomiczny we Wrocławiu

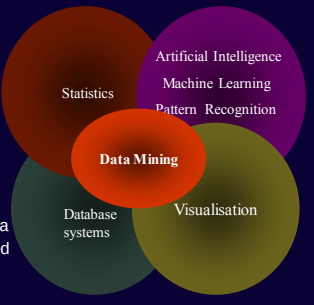
IT in Finance

Mining of Financial Databases Introduction

Jerzy KORCZAK
email: jerzy.korczak@ue.wroc.pl
http://www.korczak-leliwa.pl

Origins of Data Mining

- Draws ideas from AI/machine learning, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining

- Data mining (knowledge discovery in databases, KDD)
  - Extraction of interesting, non-trivial, implicit, previously unknown and potentially useful information (knowledge) or patterns from data in large databases or other information repositories
- Scientific point of view: data abstraction and KDD
- Commercial point of view: competitive pressure
- Necessity is the mother of invention
  - *Data is everywhere — data mining should be everywhere, too!*
  - Understand and use data — an imminent task!

Statistics vs Data Mining

- Statistics: a discipline dedicated to data analysis
- What are the differences?
  - Huge amount of data—in Giga to Tera bytes
  - Fast computing—quick response, interactive analysis
  - Multi-dimensional, powerful, thorough analysis
  - High-level, "declarative"—user's ease and control
  - Automated or semi-automated—mining functions hidden or built-in in many systems

Types of Decision-Support Systems (DSS)

Model-driven DSS:

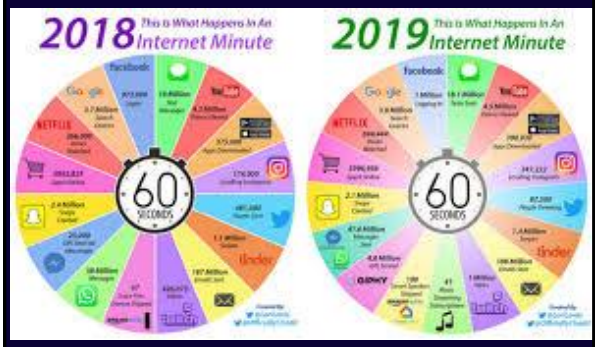
- Primarily stand-alone systems
- Use a strong theory or model to perform "*what-if*" analyses

Data-driven DSS:

- Integrated with large pools of data in major enterprise systems and Web sites
- Support decision making by enabling user to extract useful information
- Data mining: can obtain types of information such as associations, sequences, classifications, clusters, and forecasts

Data Sets, Database, Images

- Relational database — A commodity of every enterprise
- Huge data warehouses are under construction
- POS (Point of Sales): Transactional DBs in terabytes
- Object-relational databases, distributed, heterogeneous, and legacy databases
- Spatial databases (GIS), remote sensing database (EOS), and scientific/engineering databases
- Time-series data (e.g., stock trading) and temporal data
- Text (documents, emails) and multimedia databases
- WWW: A huge, hyper-linked, dynamic, global information system

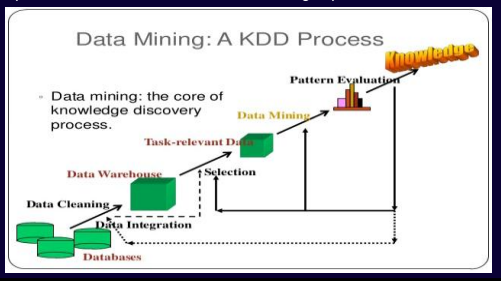


7

What is Data Mining?

Many definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

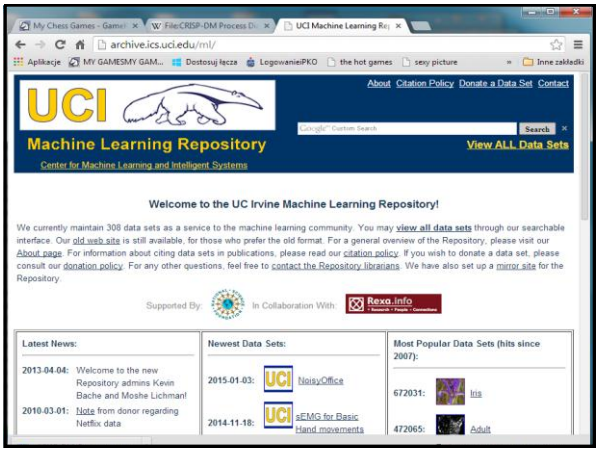


8

A Brief History of Data Mining

- Most scientific discoveries involve "data mining"
  - Kepler's Law, Newton's Laws, periodic table of chemical elements, …, from "big bang" to DNA
- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995- now International Conferences on Knowledge Discovery in Databases and Data Mining (KDD)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2005 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, DaWaK, SPIE-DM, etc.

9



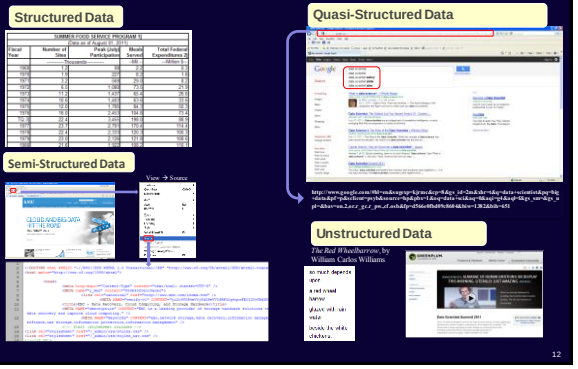
Mining of Big Data

- Concepts of Big Data:
  - **"Big Data" is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value**
    - Requires new data architectures, analytic sandboxes
    - New tools
    - New analytical methods
    - Integrating multiple skills into new role of data scientist
  - Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities

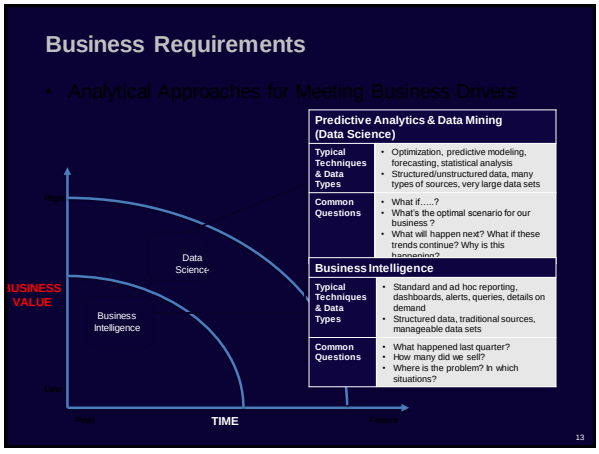
Source: McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

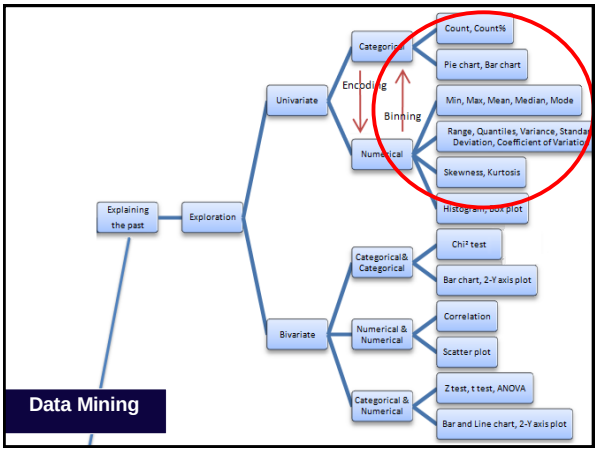
11

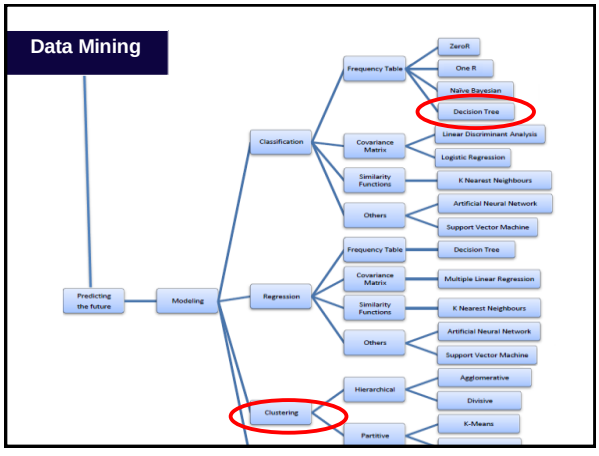
Types of Data Structures:



12


Business Requirements
• Analytical Approaches for Meeting Business Drivers
Predictive Analytics & Data Mining (Data Science)
Typical Techniques & Data Types
• Optimization, predictive modeling, forecasting, statistical analysis
• Structured/unstructured data, many types of sources, very large data sets
Common Questions
• What if.....?
• What's the optimal scenario for our business ?
• What will happen next? What if these trends continue? Why is this
Business Intelligence
Typical Techniques & Data Types
• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand
• Structured data, traditional sources, manageable data sets
Common Questions
• What happened last quarter?
• How many did we sell?
• Where is the problem? In which situations?
BUSINESS VALUE
Data Science
Business Intelligence
TIME
13


Data Mining
Explaining the past
Exploration
Univariate
Categorical
Encoding
Binning
Numerical
Count, Count%
Pie chart, Bar chart
Min, Max, Mean, Median, Mode
Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
Skewness, Kurtosis
Histogram, Box plot
Bivariate
Categorical& Categorical
Chi² test
Bar chart, 2-Y axis plot
Numerical & Numerical
Correlation
Scatter plot
Categorical & Numerical
Z test, t test, ANOVA
Bar and Line chart, 2-Y axis plot


Data Mining
Predicting the future
Modeling
Classification
Regression
Clustering
Frequency Table
Covariance Matrix
Similarity Functions
Others
ZeroR
One R
Naïve Bayesian
Decision Tree
Linear Discriminant Analysis
Logistic Regression
K Nearest Neighbours
Artificial Neural Network
Support Vector Machine
Frequency Table
Covariance Matrix
Similarity Functions
Others
Decision Tree
Multiple Linear Regression
K Nearest Neighbours
Artificial Neural Network
Support Vector Machine
Hierarchical
Partitive
Agglomerative
Divisive
K-Means


Research Progress
• Multi-dimensional data analysis: Data Warehouse and OLAP
• Association, correlation, and causality analysis
• Classification: scalability and new approaches
• Clustering and outlier analysis
• Sequential patterns and time-series analysis
• Similarity analysis: curves, trends, images, texts, etc.
• Text mining, Web mining and Weblog analysis
• Social networks, link analysis
• Spatial, multimedia, scientific data analysis
• Smart sensors: IoT
• Image classification and interpretation
• Data preprocessing and database compression
• Data visualization and visual data mining
• Many others, e.g., collaborative filtering
16