Data Mining

Lecture 1

Lecture outline

- What is Data Mining?
- Data
- Methods and stages of Data Mining

WHAT IS DATA MINING?

What is Data Mining?

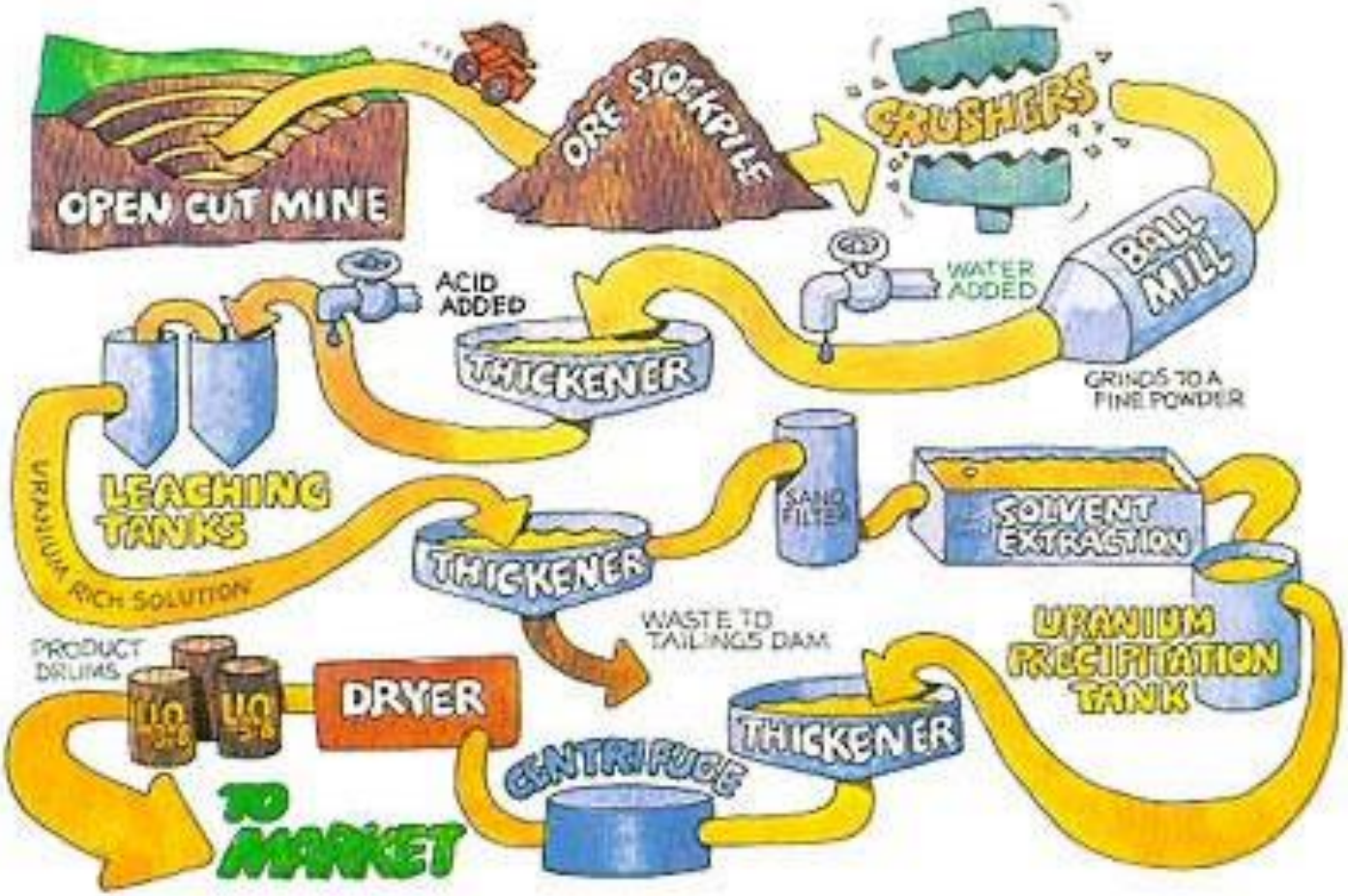
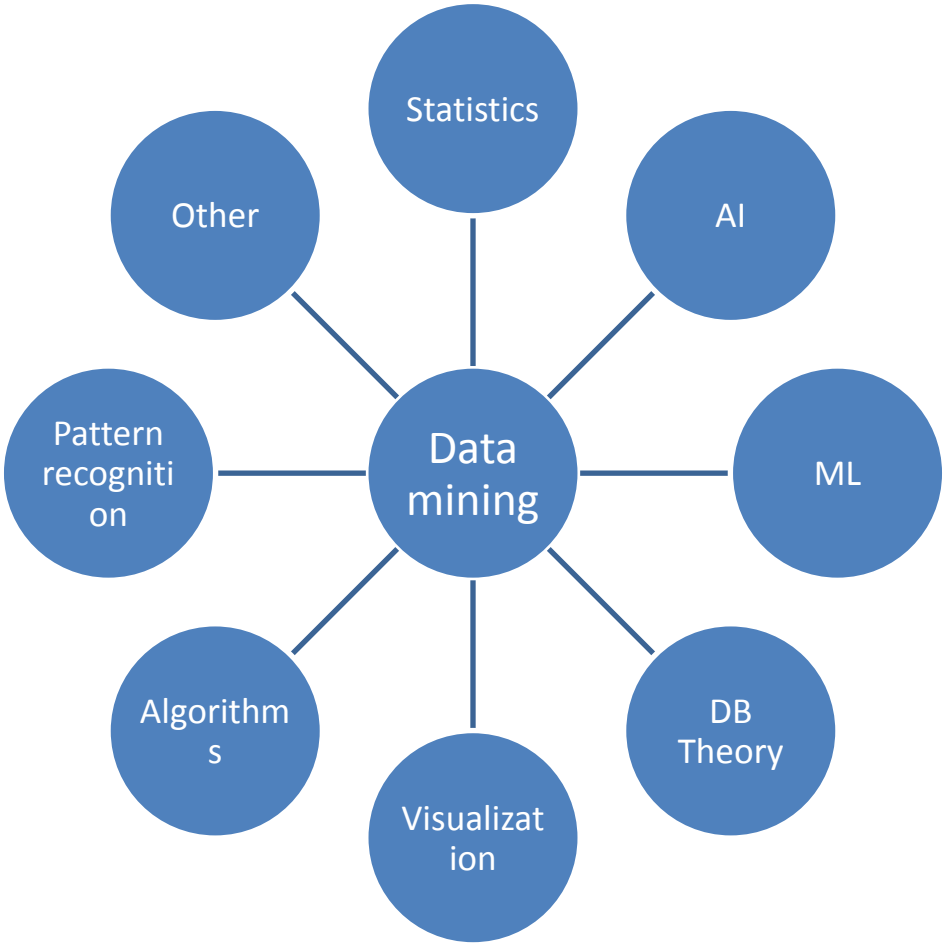


Image source: https://www.mystorybook.com/books/151814

What is Data Mining?

- Data Mining is…
  - Information extraction
  - Data excavation
  - Data intellectual analysis
  - Search for regularities
  - Knowledge extraction
  - Pattern analysis
  - Knowledge Discovery in Databases, KDD

What is Data Mining?



What is Data Mining?

- **Statistics** – science of data collecting, processing and analysis for detecting the regularities peculiar to the researched object.
- **Machine learning (ML)** – algorithmic learning of new knowledge by a computer program from the data.
- **Artificial Intelligence (AI)** – research area of human intellectual process modelling.

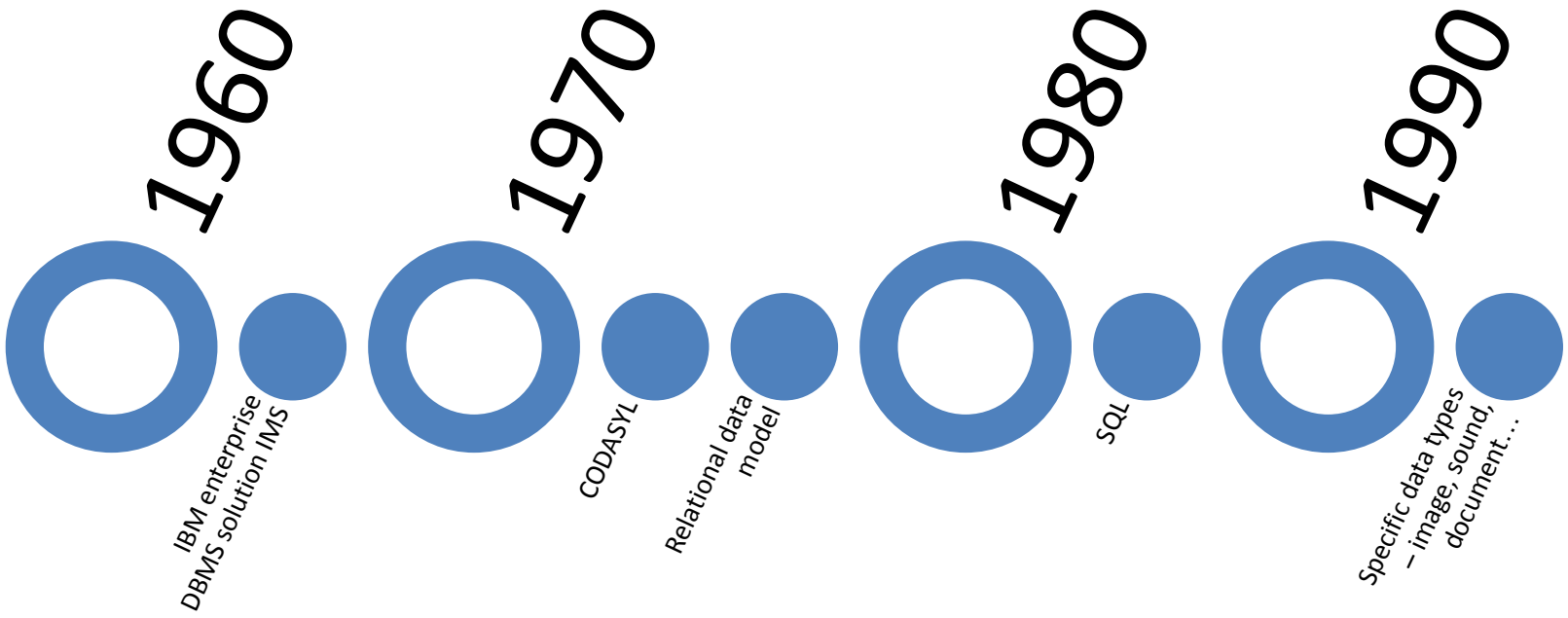
What is Data Mining?

Comparison of statistics, machine learning and Data Mining

- Statistics
  - More than Data Mining is based on theory
  - More concentrated on hypothesis checking.
- Machine learning
  - More heuristic in nature.
  - Concentrated on the enhancing of learning agents.
- Data Mining.
  - Integration of theory and heuristics
  - Concentrated on the data analysis process as a whole, including data cleaning, learning, integration and visualization of the obtained results.

What is Data Mining?

DB technology evolution

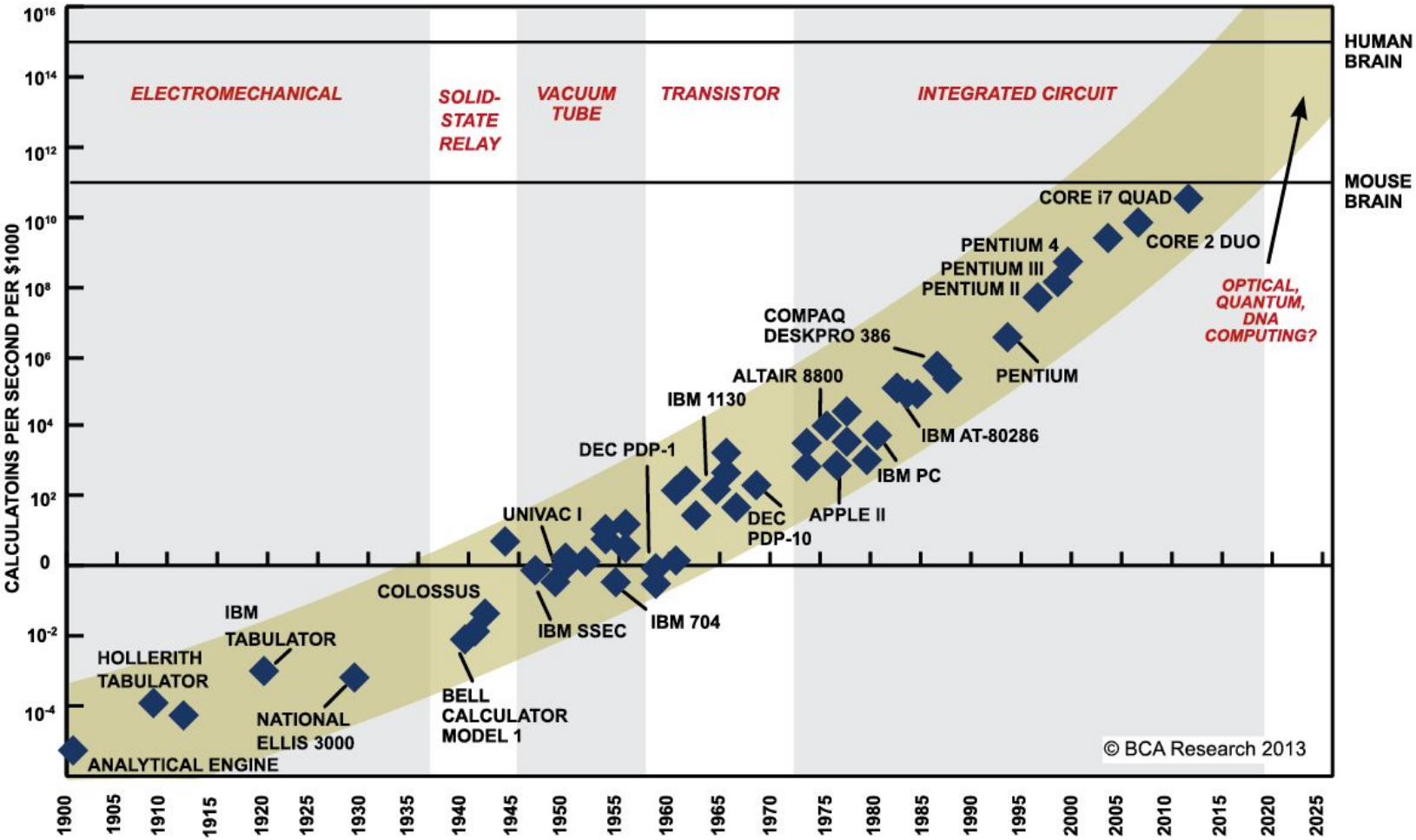


What is Data Mining?

Basic factors for emerging and development of Data Mining:

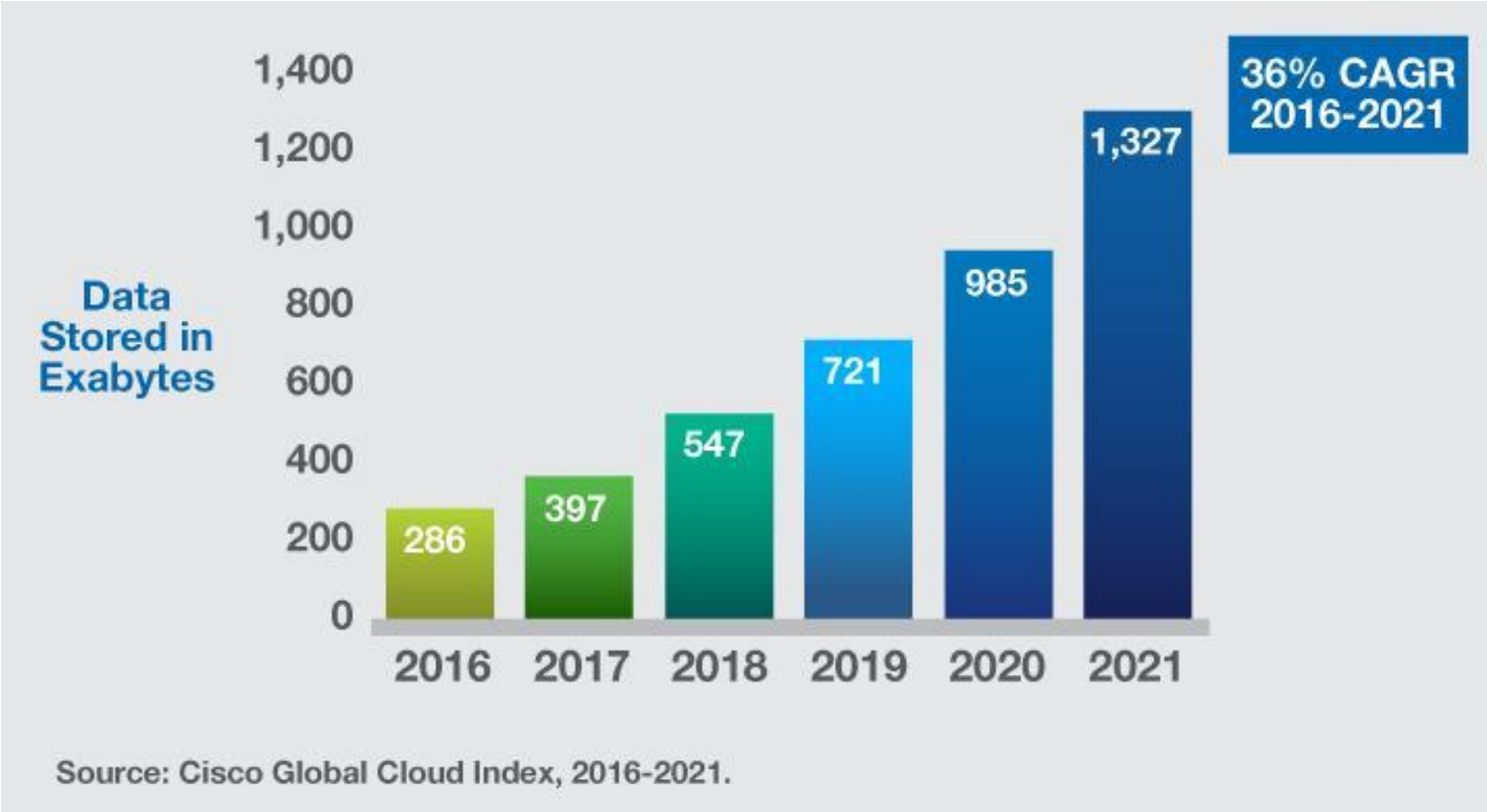
- Hardware and software technological improvement
- Improvement of data record and storage technologies
- Accumulation of large volume of retrospective data
- Improvement of data processing algorithms

What is Data Mining?



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

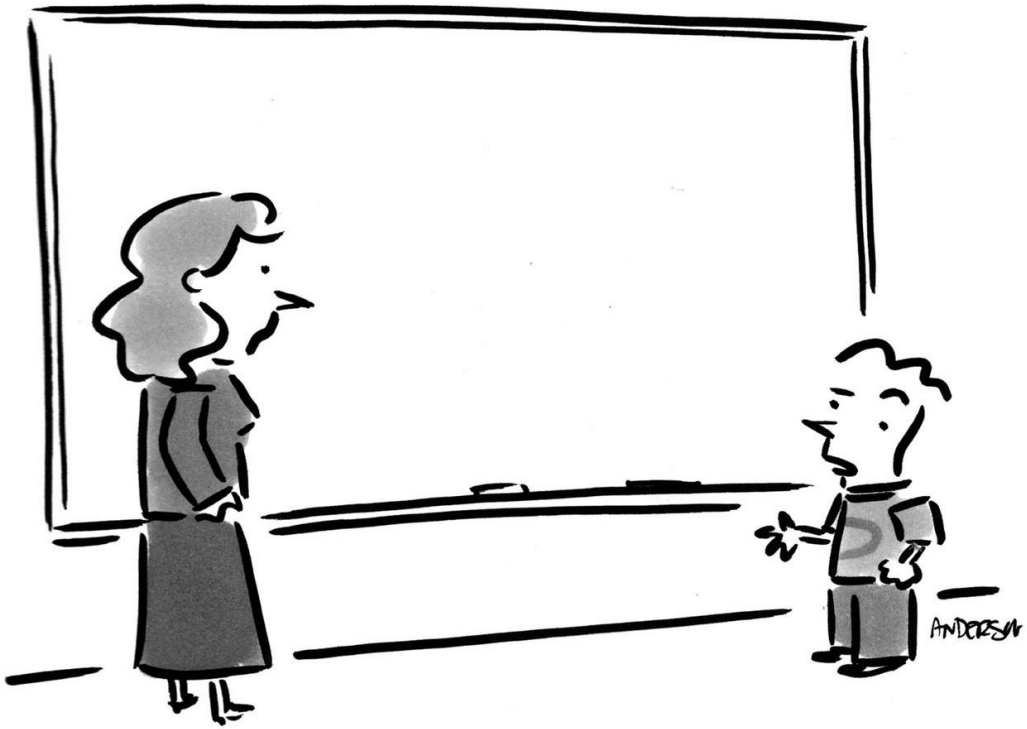
What is Data Mining?



What is Data Mining?

- **Data mining** - is the process of discovering previously unknown, nontrivial, practically useful and interpretable knowledge from the raw data and for use in decision making processes in a wide range of human activities.

Gregory Piatetsky-Shapiro



"Before I write my name on the board, I'll need to know how you're planning to use that data."

DATA

Data

What is Data?

- Data are the facts:
  - Numbers
  - Texts
  - Images
  - Sounds
  - Video records
- Data sources:
  - Measurements
  - Experiments
  - Arithmetic and logical operations
  - Records

Data

Attributes/Features

Objects

ID	Age	Marital status	Income	Gender
1	28	Single	100	male
2	22	Married	50	female
3	45	Divorced	67	female
4	30	Single	80	male
5	18	Single	20	female
6	26	Divorced	50	male
7	60	Widowed	50	female
8	34	Married	120	male
9	25	Married	80	male

Data

- Variable/Attribute/Feature/Charachteristic
  - Value
    - Discrete/Continuous
    - Numeric/Categorial
  - Dependent/Independent
- Studied objects
  - Population - parameters
  - Sample - statistics

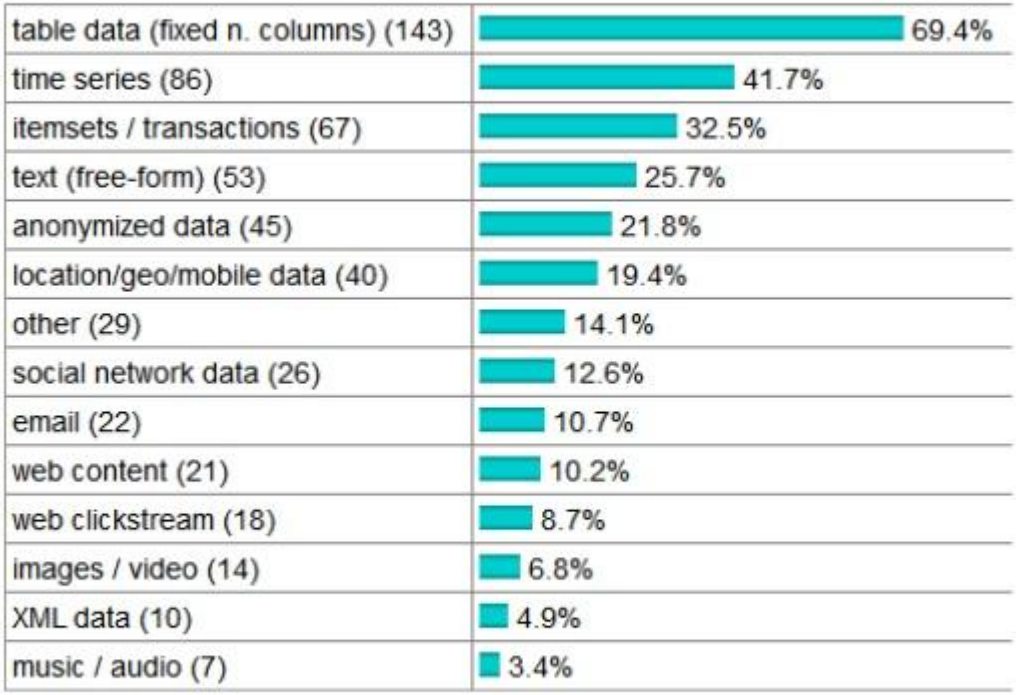
Data

Types of datasets:

- Table data
- Transactional data
- Graphical data
  - Graphs
  - Molecular structures
  - Maps

Data

Data Types Analyzed/Mined



www.KDnuggets.com/polls/2011/data-types-analyzed-mined.html

Data

- Data base – is electronic data organized and stored in a specific way.
- Data scheme – description of the data logic structure
- DBMS – shell for organizing interrelated tables with data into a data base.

Data

Data base requirements:

- High speed performance
- Data updating simplicity
- Data independence
- Multiuser usage
- Data safety
- Standardization of building and exploitation of the DB
- Data adequacy
- User-friendly interface

Data

Data type classification:

- Relational data
- Multidimensional data
- Permanency
  - Variable
  - Constant
  - Conditionally constant
- Function
  - Operational
  - Archive
  - Reference
- Time
  - Periodic
  - Point

Data

Metadata – is the data about the data

- Catalogues
- References
- Registries

METHODS AND STAGES OF DATA MINING

Methods and Stages of Data Mining

- Data Mining employs a wide variety of tools ranging from classical statistics to the latest information technology achievements.
- Data Mining methods:
  - Artificial neural networks
  - Decision trees
  - Symbolic rules
  - K-nearest neghbors
  - SVM
  - Bayes networks
  - Linear regression
  - Correlation-regression analysis
  - Clustering (hierarchical, k-means and etc.)
  - Association rules (Apriori algorithm)
  - Genetic algorithm
  - Visualization methods

Methods and Stages of Data Mining

- Most of Data Mining methods are well known mathematical algorithms and methods.
- The novelty of Data Mining is in its application to solve specific science or business problems, which became possible because of tech advances.
- Algorithm – exact step by step description of inputs and actions required to achieve desired output.

Methods and Stages of Data Mining

- Abu Adallah Muhammad ibn Musa Al-Horezmi – medieval scientist and mathematician
- The book: Al-kitāb al-mukhtaṣar fī <u>hisāb</u> **al-ǧabr** wa'l-muqābala
  - Decimal system
  - Solving of quadratic equation algorithm
  - Latin translation – Algebra, was the starting point of European math
  - Contained compilation of Indian mathematicians' achievements

Methods and Stages of Data Mining

Stage 1

- Discovery
- Regularity detection
- Laws and rules

Stage 2

- Using regularities to foretell unknowns.
- Forecasting

Stage 3

- Exception analysis
- Anomaly detection in regularities

Methods and Stages of Data Mining

- **Stage 1 – Discovery**
  - Conditional logic
  - Associations and affinities
  - Trends and variations
  - Rules validation on the test dataset
- **Example**: Using HH database (induction)
  - Using queries analyst could detect mean desired salary of specialists in the age range 25-35 years is $1200
  - Using Data Mining methods, after defining the target variable:
    - If age<20 and desired salary>$700 then position searched is **programmer** (target)
    - If age>35 and desired salary>$1200 than **managing position** is searched
    - If managing position is searched and years of experience>15 then **age is 35** in 65% of cases

Methods and Stages of Data Mining

- **Stage 2 – Forecasting**
  - Use rules detected on Stage 1 to predict the unknowns
  - Classification and regression
- **Example**: Using the rules derived from HH database analysis (deduction)
  - If <u>age<20</u> and desired <u>salary>$700</u> then position searched is **programmer** (target)
  - If <u>age>35</u> and desired <u>salary>$1200</u> than **managing position** is searched
  - If <u>managing position is searched</u> and years of <u>experience>15</u> then **age is 35** in 65% of cases

Methods and Stages of Data Mining

- Stage 3 – Exception analysis
  - Detect anomalies, deviations and exceptions
- Example:
  - If age >35 and desired salary>$1200 then 90% of cases managing position is searched. What is the other 10% of cases?
    - Second rule
    - Error (use in data cleaning)

Methods and Stages of Data Mining

- Technological method classification
  - Data preservation
    - Data is stored in the detailed state and used directly
    - Problems with large amounts of data
    - Methods – clustering, analogy
  - Data distillation
    - Feature engineering
    - Dimensionality reduction
    - Methods:
      - Logical methods: induction, fuzzy logic queries, symbolic rules, decision trees, genetic algorithms
      - Cross-tabulation methods: agents, Bayesian networks, cross-table visualization
      - Equation-based methods: statistical methods (correlations, regressions), neural networks

Methods and Stages of Data Mining

- Learning method classification
  - Statistical methods based on retrospective data
    - Descriptive analysis (homogeneity, stationarity hypothesis testing, distribution analysis)
    - Relation analysis (correlation, regression analysis)
    - Multidimensional statistical analysis (linear and non-linear discriminant analysis, clustering, component analysis, factor analysis)
    - Time series analysis
  - Cybernetic methods
    - Neural networks
    - Evolutionary algorithms
    - Genetic algorithms
    - Association rules
    - Fuzzy logic
    - Decision trees
- Both types rely on statistics

Summary

- What is Data Mining?
  - Information extraction
  - Data excavation
  - Data intellectual analysis
  - Search for regularities
  - Knowledge extraction
  - Pattern analysis
  - Knowledge Discovery in Databases, KDD
  - Statistics and ML
- Data
  - Facts
  - Sources
  - Metadata
- Methods and stages of Data Mining
  - Discovery
  - Forecasting
  - Exception analysis