COMP 345
Data Mining

Introduction &
Course Overview

Rhodes College

What Is Data Mining?

https://medium.com/@SunTecIndia/effective-data-mining-strategies-to-boost-your-business-db23a0594ecd

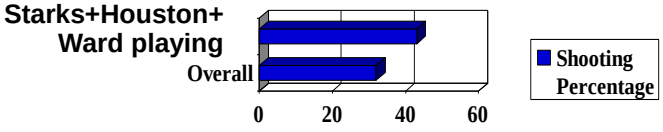
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

3

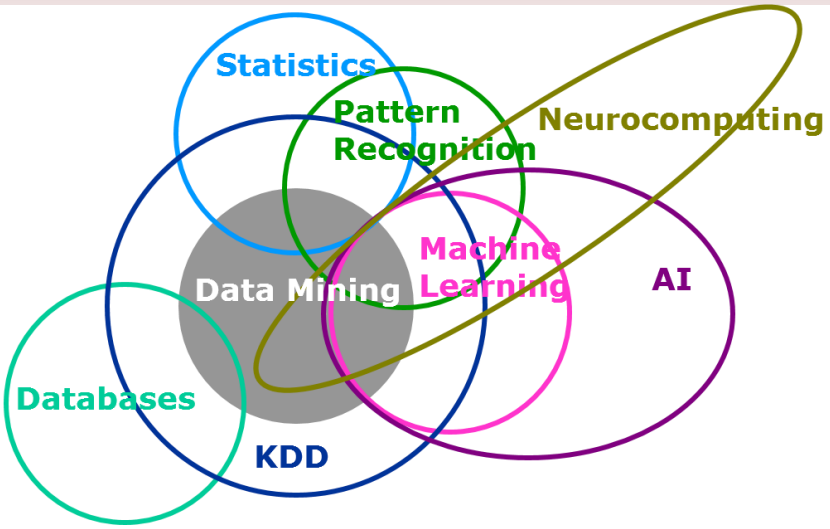
What is Data Mining?
Real Example from the NBA

- Play-by-play information recorded by teams
  - Who is on the court
  - Who shoots
  - Results
- Coaches want to know what works best
  - Plays that work well against a given team
  - Good/bad player matchups
- Advanced Scout (from IBM Research) is a data mining tool to answer these questions



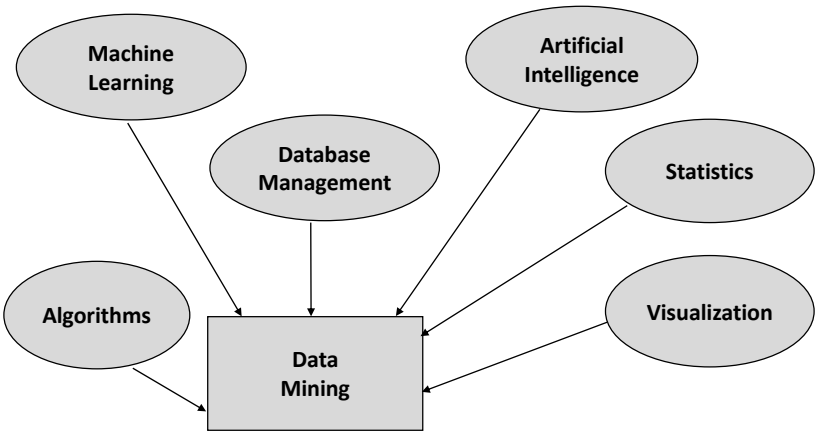
4

What is Data Mining?



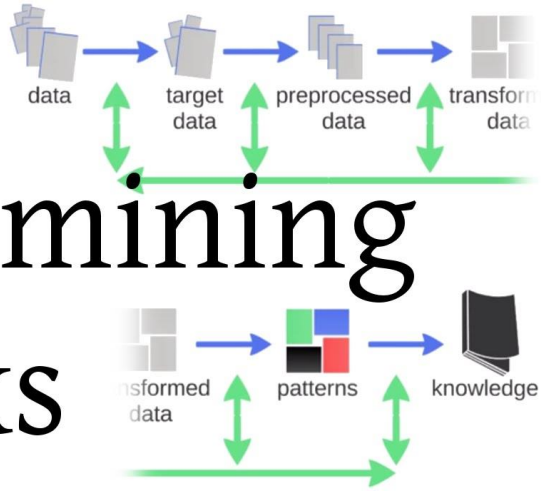
5

Integration of Multiple Technologies



6

How data mining works



https://www.youtube.com/watch?v=W44q6qszdqY

7

Data Mining Phases / Steps

1

1 **Define the Problem**
Identify business goals
Identify data mining goals

2 **Identify Required Data**
Assess needed data
Collect and understand data

3 **Prepare and Pre-process**
Select required data
Cleanse/format data as necessary

4 **Model the Data**
Select algorithms
Build predictive models

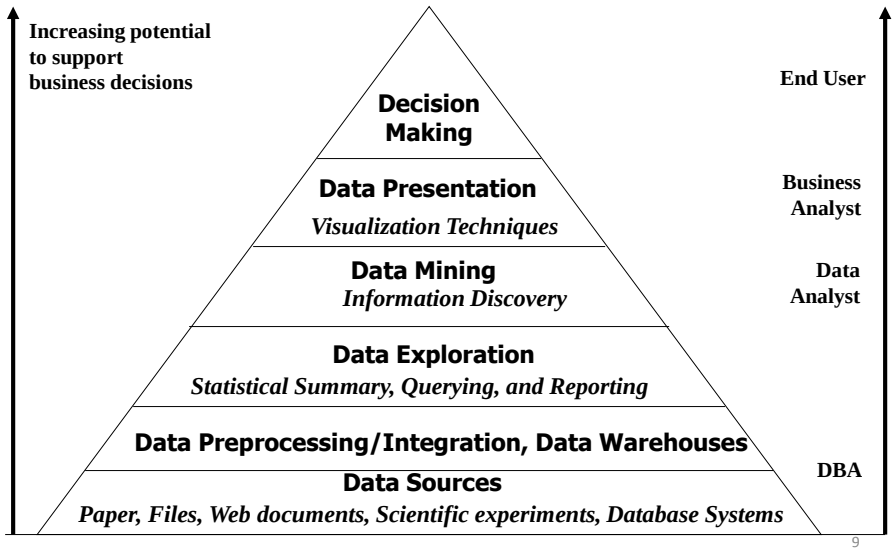
5 **Train and Test**
Train the model with sample data sets
Test and iterate

6 **Verify and Deploy**
Verify final model
Prepare visualizations and deploy

https://digitaltransformationpro.com/wp-content/uploads/2017/06/Datamining.png

8

Data Mining in Business Intelligence



9

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

10

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis

11

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

12

Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

13

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

14

What Can Data Mining Do?

- Cluster
- Classify
  - Categorical, Regression
- Summarize
  - Summary statistics, Summary rules
- Link Analysis / Model Dependencies
  - Association rules
- Sequence analysis
  - Time-series analysis, Sequential associations
- Detect Deviations

15

Data Mining Complications

- Volume of Data
  - Clever algorithms needed for reasonable performance
- Interest measures
  - How do we ensure algorithms select “interesting” results?
- “Knowledge Discovery Process” skill required
  - How to select tool, prepare data?
- Data Quality
  - How do we interpret results in light of low quality data?
- Data Source Heterogeneity
  - How do we combine data from multiple sources?

16

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

17

Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

18

Top Tools for Data Mining

https://www.kdnuggets.com/2015/05/poll-analytics-data-mining-data-science-software-used.html

19

Course Page & Class Schedule

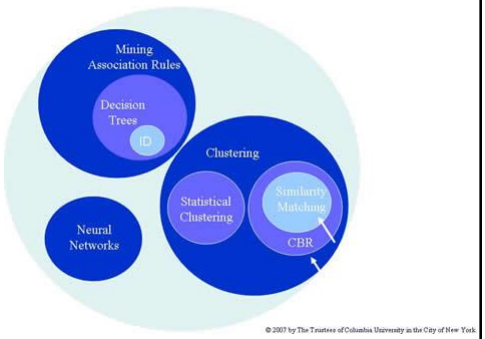
http://cs.rhodes.edu/welshc/COMP345_F18/

- What's there?
  - Course info
  - Course schedule
  - Lecture media (slides, handouts, etc)
  - Assignments (reading & things to hand in)
  - Project information (coming soon…)
  - Presentation information (coming soon…)

20

Course Topics

- Data acquisition and pre-processing
- Data mining process
- Association Rule Mining
- Introduction to WEKA
- Classification
- Clustering
- Big Data
- Recommender Systems



21

Course Grades

- 10% In-Class Exercises, Online quizzes
- 20% Assignments
- 10% Paper Presentation
- 25% Final Project
- 15% Midterm -in class
- 20% Final

22

Next Time

- Read Chapter 1 in Han Book
- Complete Assignment 0 – details on website – upload to Moodle