INTRODUCTION TO DATA MINING

OUTLINE

What is Data Mining ?

Specificities of Data Mining

Some examples

Typology of Methods

WHAT IS DATA MINING ?

WHAT IS DATA MINING

SS / CS

xác suất

stochastic

Datawarehouse

Data Mining is a « new » field

Crossing of

- Statistics
- Information technology
  - Databases
  - Artificial Intelligence
  - Machine Learning

Aims at discovering informations in big data sets

DEFINITIONS

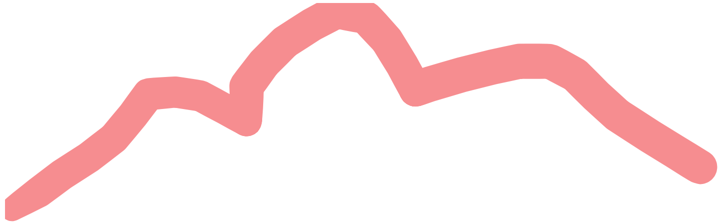
uptrend

U.M.Fayyad, G.Piatetski-Shapiro : “ Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data ”

D.J.Hand : “ I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets”

downtrend

DEFINITIONS



Data Mining metaphor :

- some « treasures » are hidden under mountains of datas and we want to discover them wwith specialized tools

Data Mining analyses datas that were collected for a different goal

- Secondary analysis of databases, mostly built for the management of personnal datas (Kardaun, T.Alanko,1998)

Data mining does not deal with collecting data efficiently (*survey, experience plans*

IS IT NEW ?

« Data Analysis is a tool to draw from the coating of datas a pure diamond of natural truth »
J.P.Benzécri 1973

« Statistics is the science of learning from data. Statistics is essential for the proper running of government, central to decision making in industry,and a core component of modern educational curricula at all levels » J.Kettenring, 1997

HISTORY : MANY DIFFERENT NAMES

Data Fishing, Data Dredging: 1960

- used by statisticians (as bad name)

Data Mining : 90's

- used in DB community, business

Knowledge Discovery in Databases : 90's KDD

- used by AI, Machine Learning Community

also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...

HT

Currently: Data Mining and Knowledge Discovery are used interchangeably

TRENDS LEADING TO DATA FLOOD

More data is generated:

- Bank, telecom, other business transactions ...
- Scientific data: astronomy, biology, etc
- Web, text, and e-commerce

BIG DATA EXAMPLES

PB = 1000000GB

eBay two data warehouses at 7.5 petaBytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising

Archive.org : in October 2016, collection topped 15 petabytes

NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations

BIG DATA EXAMPLES

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second :

- 600 millions of collisions / seconds
- Filtering → refraining 99.99995% of data, represents 25 petabytes annual rate

The Square Kilometre Array is a radio telescope built of thousands of antennas →operationnal by 2024

- expected to gather 14 exabytes and store one petabyte per day

DATA GROWTH RATE

exponential data growth toward 2020 and beyond

Size of the digital universe will double every two years

50-fold growth from 2010 to 2020

Very little data will ever be looked at by a human

Knowledge Discovery is **NEEDED** to make sense and use of data.

MACHINE LEARNING / DATA MINING APPLICATION AREAS

FIS GM

ERP SAP

Science

- astronomy, bioinformatics, drug discovery, …

Business

- CRM (Customer Relationship management), fraud detection, e-commerce, manufacturing, sports/entertainment, telecom, targeted marketing, health care, …

Web:

- search engines, advertising, web and text mining, …

Government

- surveillance (?|), crime detection, profiling tax cheaters, …

APPLICATION AREAS

What do you think are some of the most important and widespread business applications of Data Mining?

DATA MINING FOR CUSTOMER MODELING

Customer Tasks:

- attrition prediction
- targeted marketing:
  - cross-sell, customer acquisition
- credit-risk
- fraud detection

Industries

- banking, telecom, retail sales, …

CUSTOMER ATTRITION: CASE STUDY

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year!
- With this in mind, what is our task?
  - Assume we have customer information for the past N months.

- Task:
- Predict who is likely to attrite next month.
- Estimate customer value and what is the cost-effective offer to be made to this customer.

CUSTOMER ATTRITION RESULTS

Verizon Wireless built a customer data warehouse

Identified potential attriters

Developed multiple, regional models

Targeted customers with high propensity to accept the offer

Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

ASSESSING CREDIT RISK: CASE STUDY

Situation: Person applies for a loan

Task: Should a bank approve the loan?

Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle

CREDIT RISK - RESULTS

Banks develop credit models using variety of machine learning methods.

Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan

Widely deployed in many countries

E-COMMERCE

A person buys a book (product) at Amazon.com

What is the task?

SUCCESSFUL E-COMMERCE – CASE STUDY

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

UNSUCCESSFUL E-COMMERCE CASE STUDY (KDD-CUP 2000)

Data: clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer

Q: Characterize visitors who spend more than $12 on an average order at the site

Dataset of 3,465 purchases, 1,831 customers

Very interesting analysis by Cup participants

- thousands of hours - $X,000,000 (Millions) of consulting

Total sales -- $Y,000

Obituary: Gazelle.com out of business, Aug 2000

GENOMIC MICROARRAYS – CASE STUDY

Given microarray data for a number of samples (patients), can we

Accurately diagnose the disease?

Predict outcome for given treatment?

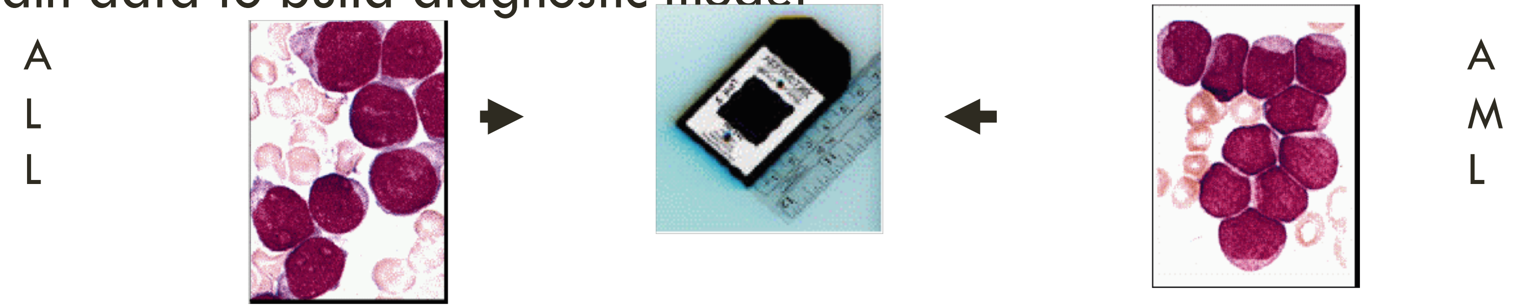
Recommend best treatment?

EXAMPLE: ALL/AML DATA

38 training cases, 34 test, ~ 7,000 genes

2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)

Use train data to build diagnostic model



Results on test data:
33/34 correct, 1 error may be mislabeled

SECURITY AND FRAUD DETECTION - CASE STUDY

Credit Card Fraud Detection

Detection of Money laundering

- FAIS (US Treasury)

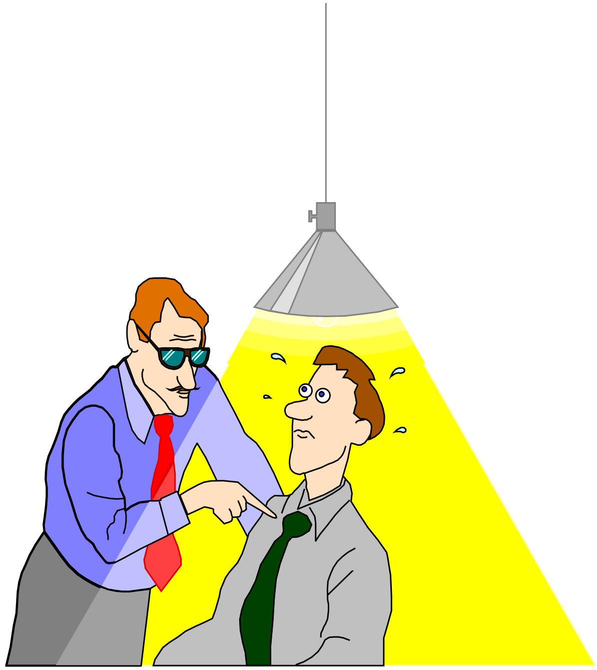
Securities Fraud

- NASDAQ KDD system

Phone fraud

- AT&T, Bell Atlantic, British Telecom/MCI

Bio-terrorism detection at Salt Lake Olympics 2002



DATA MINING AND PRIVACY

in 2006, NSA (National Security Agency) was reported to be mining years of call info, to identify terrorism networks

Social network analysis has a potential to find networks

Invasion of privacy – do you mind if your call information is in a gov database?

What if NSA program finds one real suspect for 1,000 false leads ? 1,000,000 false leads?

PROBLEMS SUITABLE FOR DATA-MINING

require knowledge-based decisions

have a changing environment

have sub-optimal current methods

have accessible, sufficient, and relevant data

provides high payoff for the right decisions!

Privacy considerations important if personal data is involved

KNOWLEDGE DISCOVERY DEFINITION

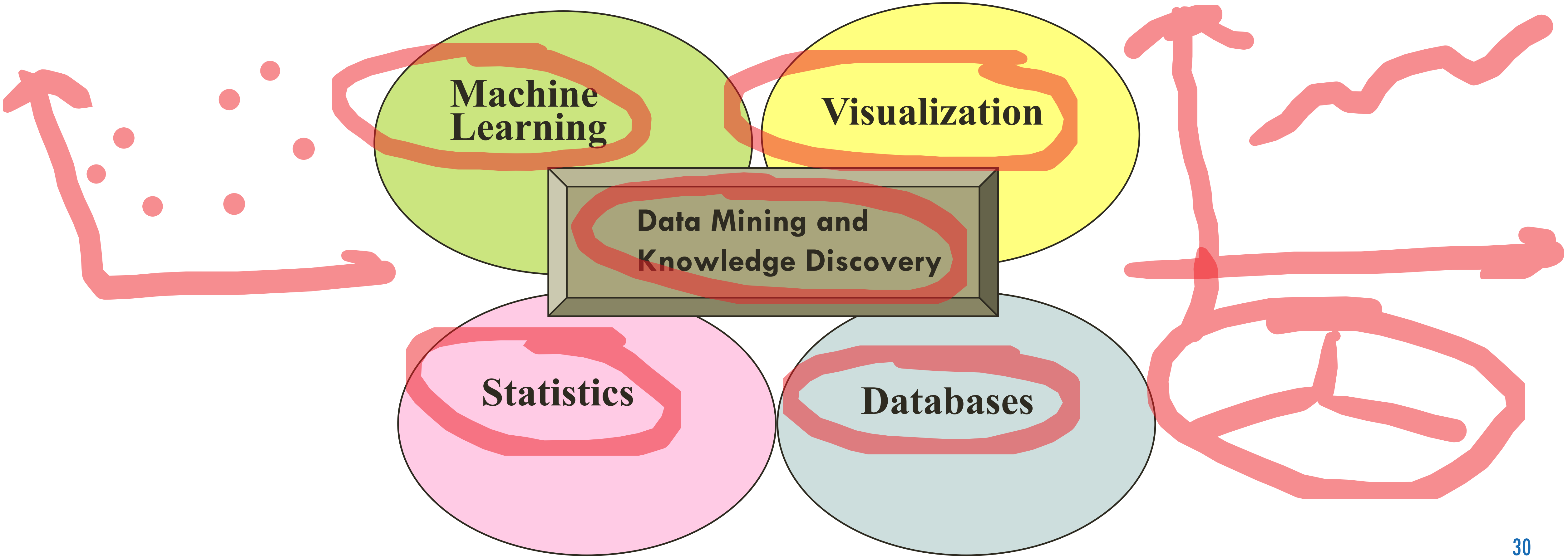
Knowledge Discovery in Data is the

non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining*, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

RELATED FIELDS



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

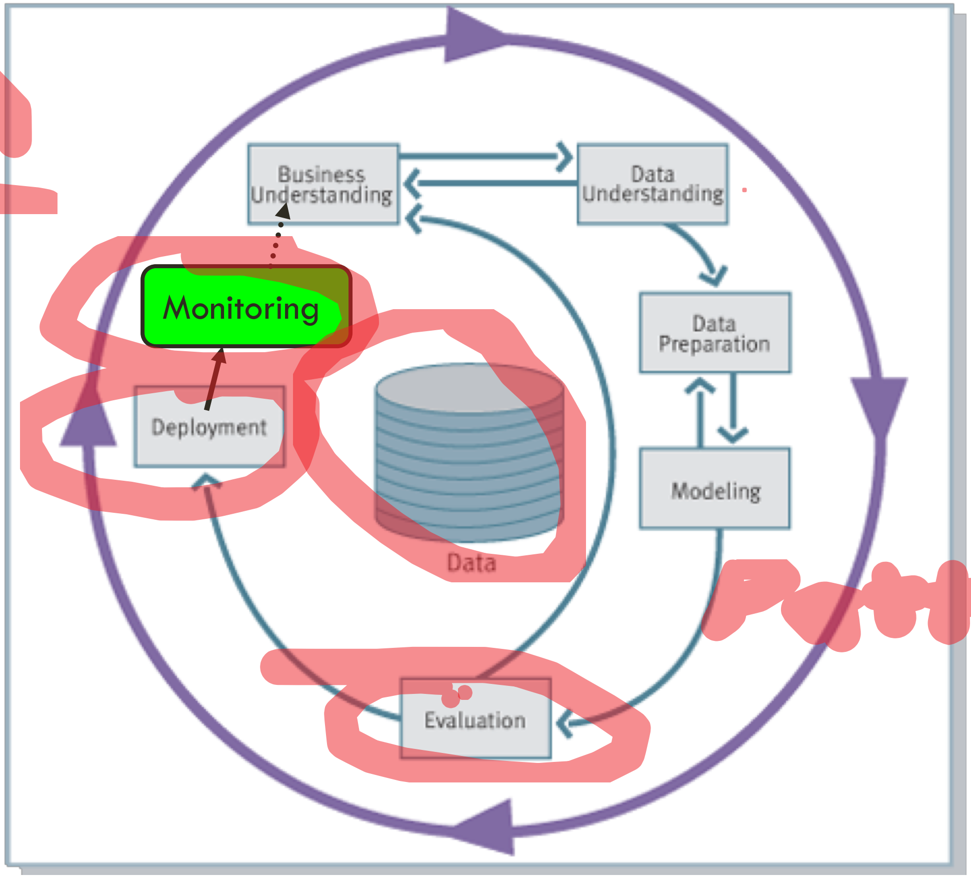
alternative

[illegible]

Distinctions are fuzzy

KNOWLEDGE DISCOVERY PROCESS FLOW, ACCORDING TO CRISP-DM

PDCA



see
www.crisp-dm.org
for more
information

P

Pattern

DATA MINING TASKS

SOME DEFINITIONS

Instance (also Item or Record): 25/12

- an example, described by a number of attributes,
- e.g. a day can be described by temperature, humidity and cloud status

Attribute or Field

- measuring aspects of the Instance, e.g. temperature

Day

Class (Label)

- grouping of instances, e.g. days good for playing

MAJOR DATA MINING TASKS

Domain

Classification: predicting an item class

Clustering: finding clusters in data

Associations: e.g. A & B & C occur frequently

Visualization: to facilitate human discovery

Summarization: describing a group

Deviation Detection: finding changes

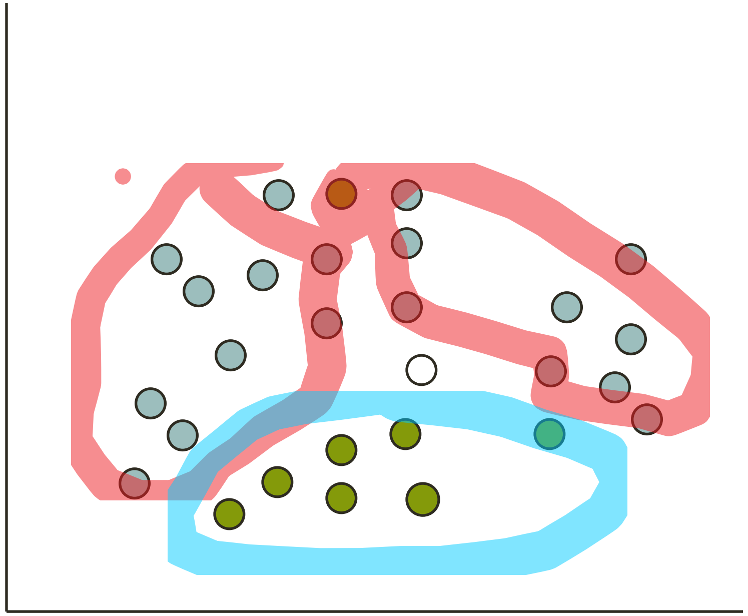
Estimation: predicting a continuous value

Link Analysis: finding relationships

…

CLASSIFICATION

Learn a method for predicting the instance class from pre-labeled (classified) instances

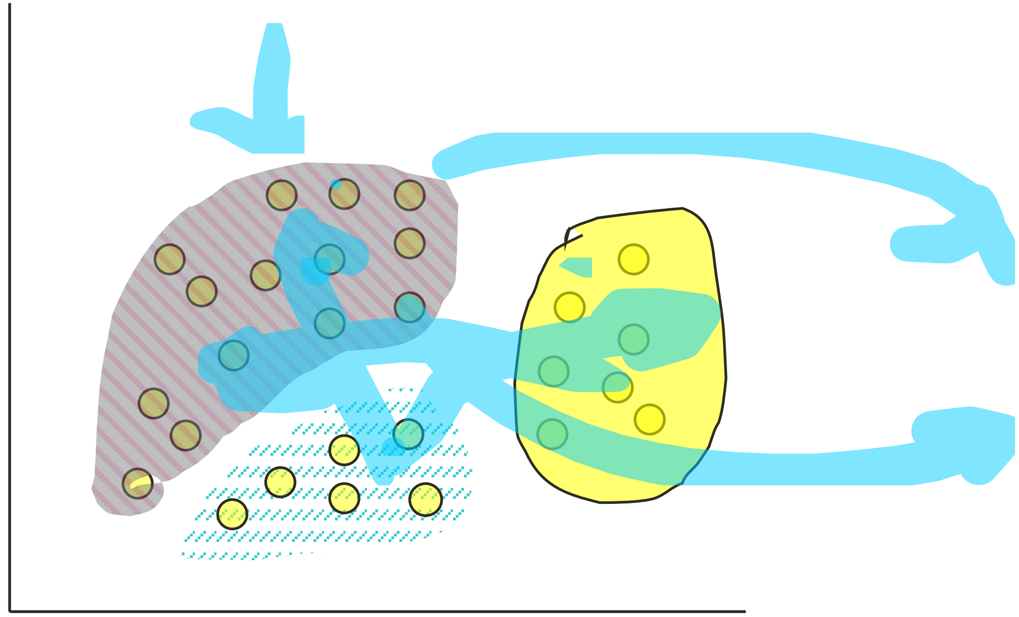


Many approaches: Statistics, Decision Trees, Neural Networks,

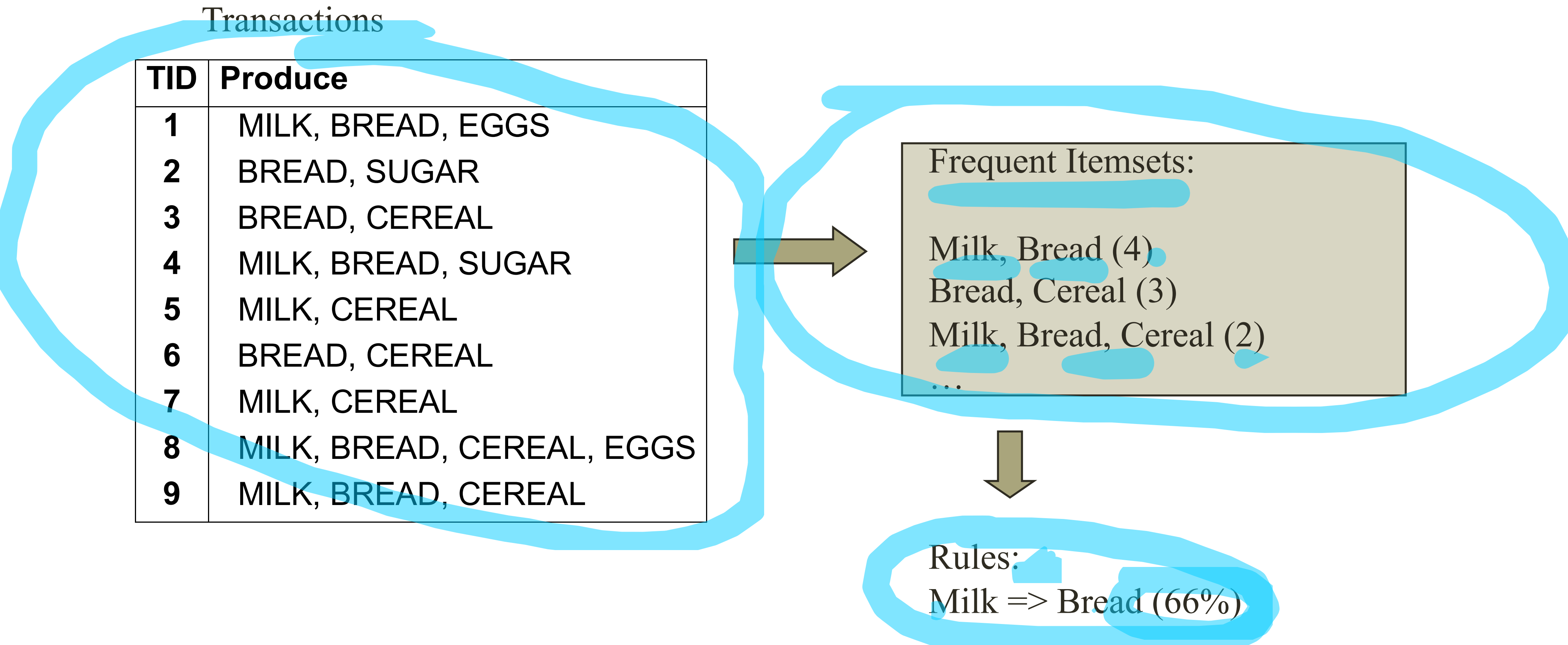
...

CLUSTERING

Find "natural" grouping of instances given un-labeled data



ASSOCIATION RULES & FREQUENT ITEMSETS



Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Frequent Itemsets:

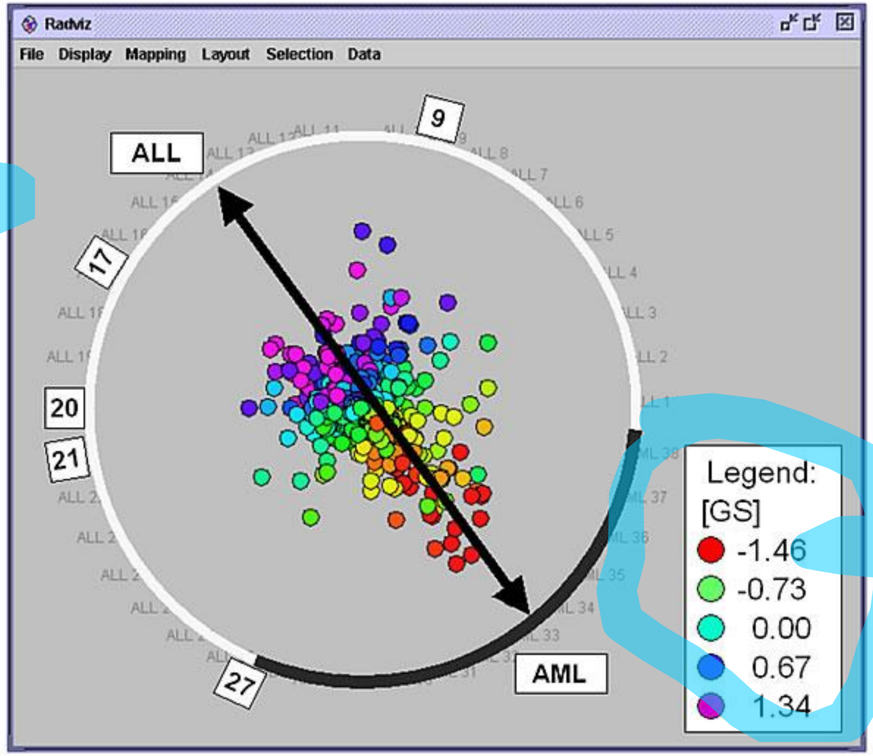
Milk, Bread (4)
Bread, Cereal (3)
Milk, Bread, Cereal (2)
…

Rules:
Milk => Bread (66%)

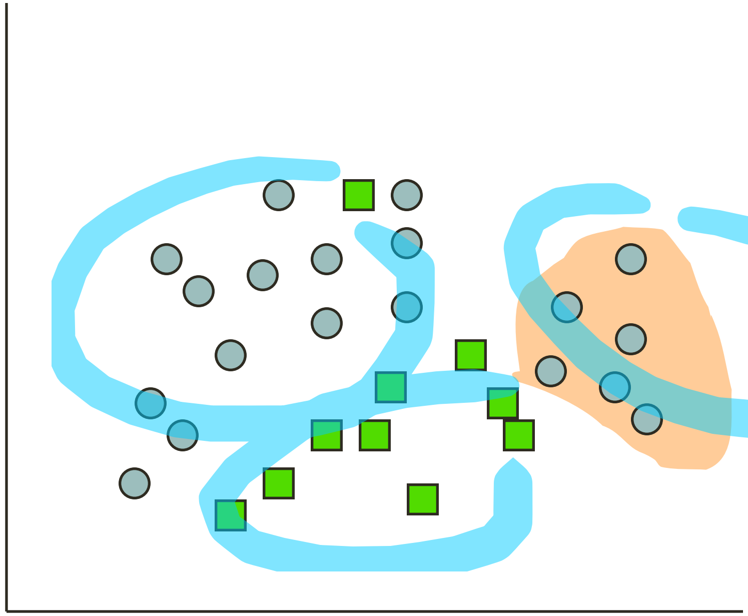
VISUALIZATION & DATA MINING

Visualizing the data to facilitate human discovery

Presenting the discovered results in a visually "nice" way



SUMMARIZATION



- Describe features of the selected group
- Use natural language and graphics
- Usually in Combination with Deviation detection or other methods

Average length of stay in this study area rose 45.7 percent, from 4.3 days to 6.2 days, because ...