

Chapter 1 Data Mining Overview

2025 Autumn

Lei Sun

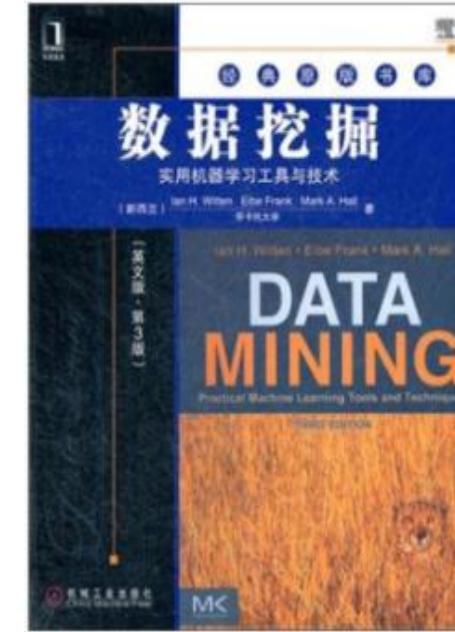


Course material

- Recommended books

刘金玲, Python数据挖掘算法与应用, 清华大学出版社, 2024
数据挖掘导论(完整版)》, (美) Pang-Ning Tan, 人民邮电出版社
薛薇, R语言数据挖掘方法及应用, 电子工业出版社, 2016.

- Python R and Rstudio



Grade

- Assignments: 40%
- Final exam: 60%

Course material



Kaggle.com

The screenshot shows the Kaggle website interface. On the left is a sidebar with navigation links: kaggle (selected), Create, Home, Competitions, Datasets (selected), Code, Discussions, Courses, and More. The main content area has a search bar, sign-in/register buttons, and a "Datasets" section. It includes a sub-section for "Explore, analyze, and share quality data.", a "New Dataset" button, and a search bar for datasets. At the bottom are category filters: Computer Science, Education, Classification, Computer Vision, NLP, Data Visualization, and Pre-Trained Model.



Github.com

基于git的代码托管平台

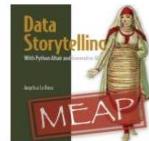
Course material

Resources---kdnuggets.com



A purple banner at the top of the page for the TDWI San Diego event. It includes the TDWI logo, the text "SAN DIEGO Aug. 6-11", "New Hands-On ML Bootcamp + Courses in MLOps, Data Viz, and more!", "Education designed for instant on-the-job value", and a yellow box with the text "Save 30% with KD30 thru June 30". Below the banner is a link: "TDWI San Diego | Aug 6-11 | Save 30% with KD30 through 6/16".

Latest Posts

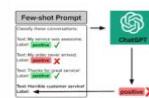


Data storytelling – the art of telling stories through data

Data Storytelling with Python Altair and Generative AI teaches you how to turn raw data into effective, insightful data stories. You'll learn exactly what goes into an effective...

By **KDNuggets** on July 18, 2023 in **Partners**

Search KDNuggets...



Ensuring Reliable Few-Shot Prompt Selection for LLMs

Data-centric techniques for better Few-Shot Prompting when applying LLMs to Noisy Real-World Data.

[Data Science, Machine Learning, AI & Analytics - KDNuggets](#)

01 Why data mining

02 What is data mining

03 Applications

04 Data mining activities



01 Why data mining



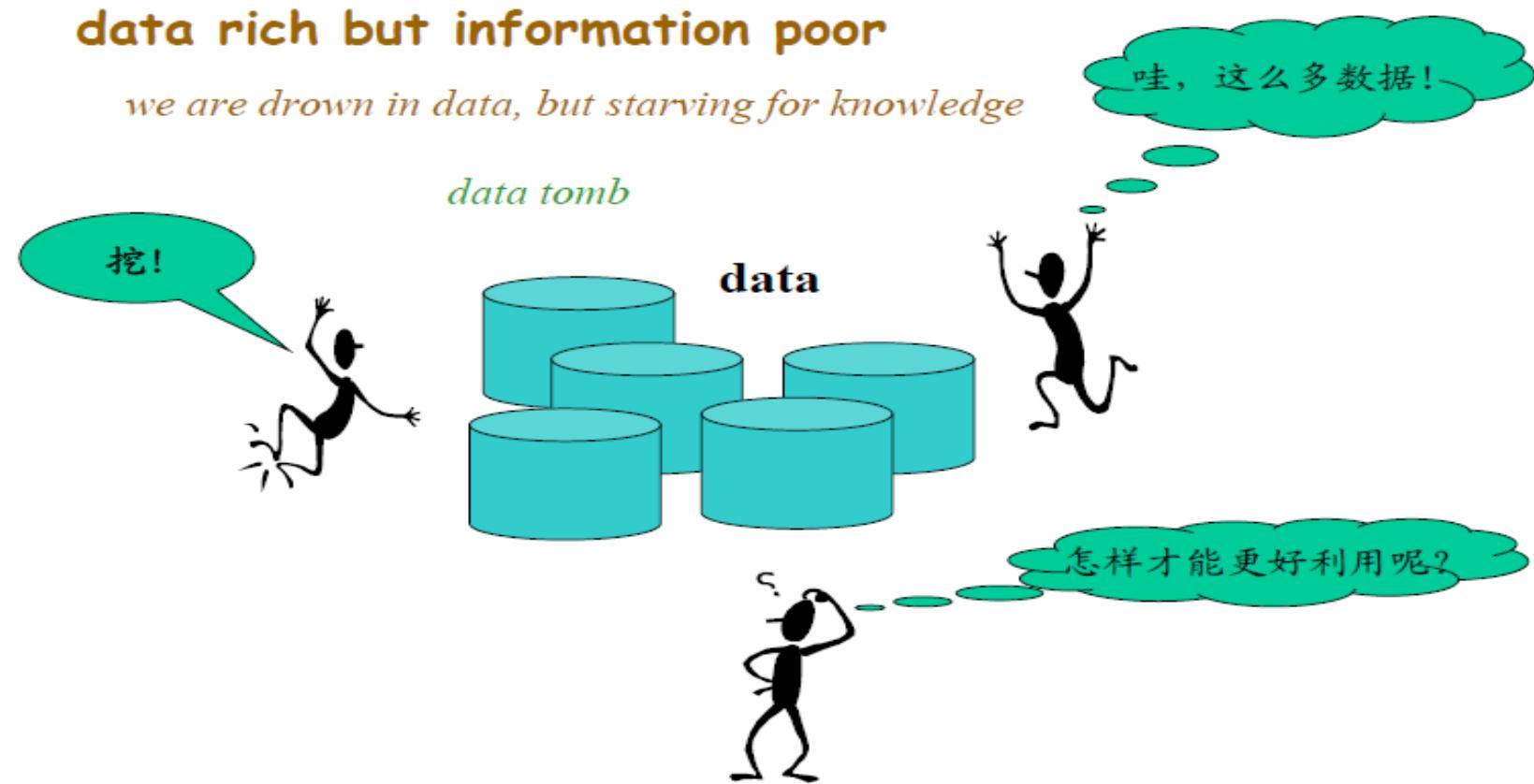
01

Why data mining

Why data mining?

data rich but information poor

we are drown in data, but starving for knowledge



01 Why data mining



Problems

How to place supermarket shelves?
What kinds of promotion for certain person?

Customer churn and package
strategy development

Bundles sell, personalized recommendations
and ad placement and value

01 Why data mining

Data



Economic
Government
eMail
Video
Text
.....
....

Knowledge



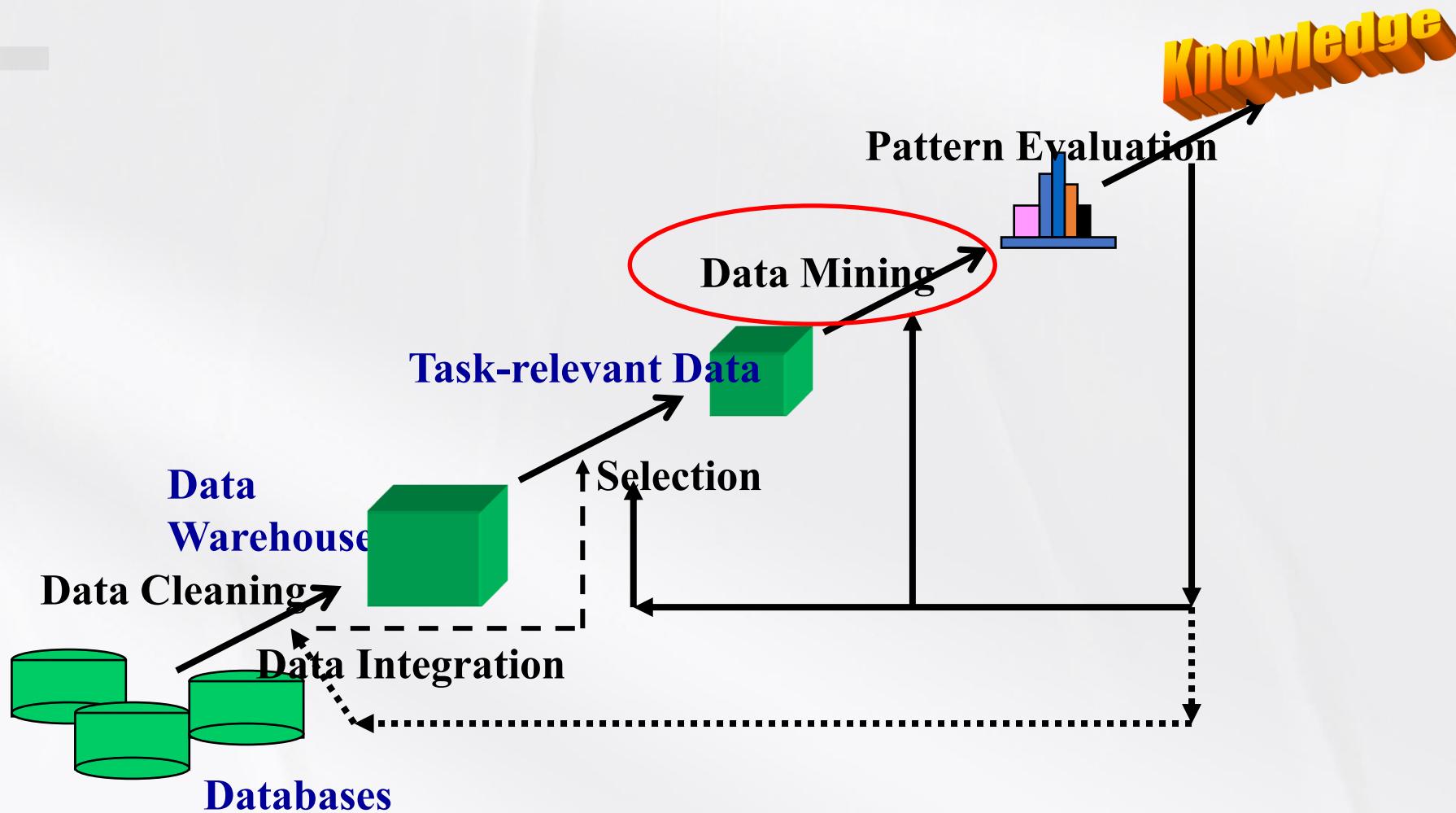
Model
Relationship
Patent
Rules
.....

Decision



Marketing
Relationship
Optimize ads
.....

02 What is data mining



02

What is data mining

Problems:

- ✓ Why select these data?

Fayyad model ignores business problems specific issue that will be analyzed.

- ✓ How to apply the model?

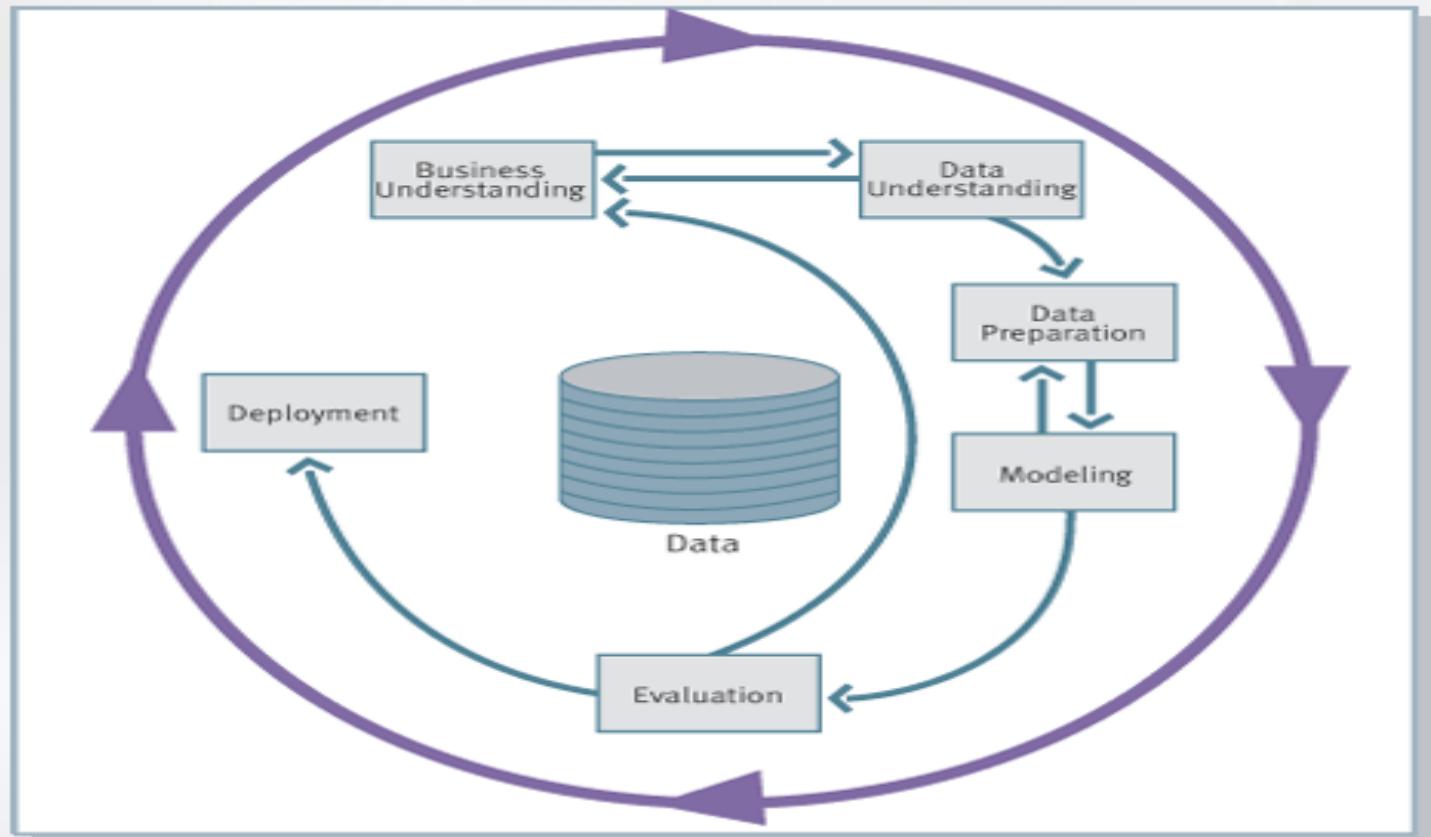
Round information flow

feedback

02 What is data mining

CRISP-DM is an iterative, adaptive process.

Cross-Industry Standard Process for Data Mining



<http://www.crisp-dm.org>

02 What is data mining

Exploration and analysis of large quantities of data to discover meaningful patterns.

Data mining is the **non-trivial process** of identifying **valid, novel, potentially useful, and ultimately understandable patterns** from huge volume of data.

derived from U.Fayyad,et al.'s definition of KDD at KDD96



数据挖掘是从海量数据中获取正确的、新颖的、潜在有用的、最终可理解的模式的非凡过程。

02 What is data mining

Interdisciplinary Studies

Data mining draws from a wide range of disciplines, such as **statistics, machine learning, artificial intelligence, and database management.**

This multidisciplinary approach enables data miners to employ various techniques and tools to address different data analysis challenges effectively.





Medical field

- Infections Disease Prediction
- Customer Assisted Therapy Program



Finance

Identify defaulting customers



E-commerce

Sales performance analysis

Products Recommended

Media



- Box Office Forecast

website



- Click-through rate analysis
- Advertising analysis



Telecommunication

CRM

Telecom Package Decision

service industry



- Associated Food
- Recommendations
- Airbnb room classification



Government

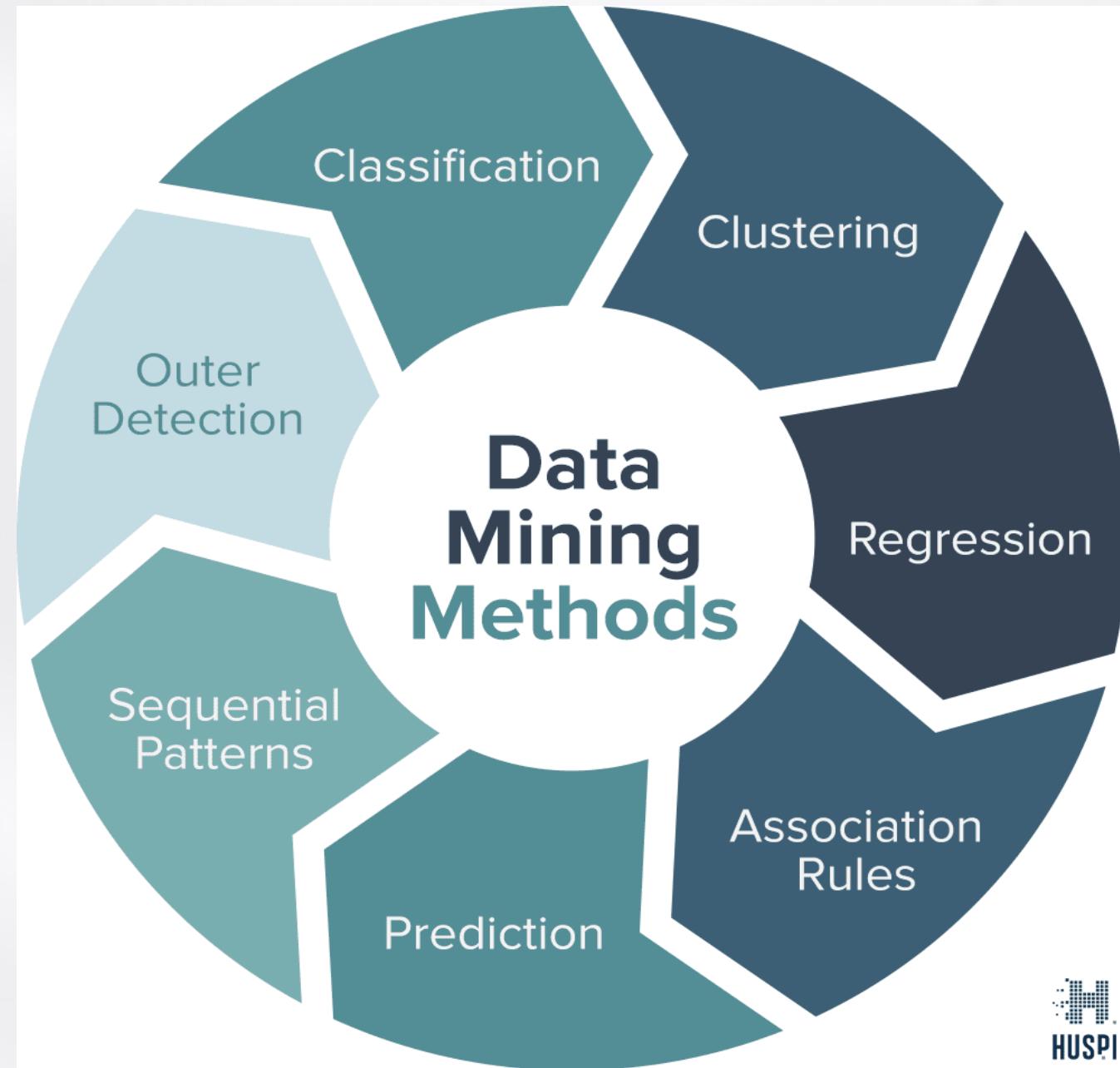
- GDP trend analysis
- City Classification by pollution

04 Data mining activities

Classification: build a model that accurately predicts the class labels of new instances based on their features. Each instance in a dataset is assigned a **class label**.

Clustering: divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups.

Without class label



04 Data mining activities

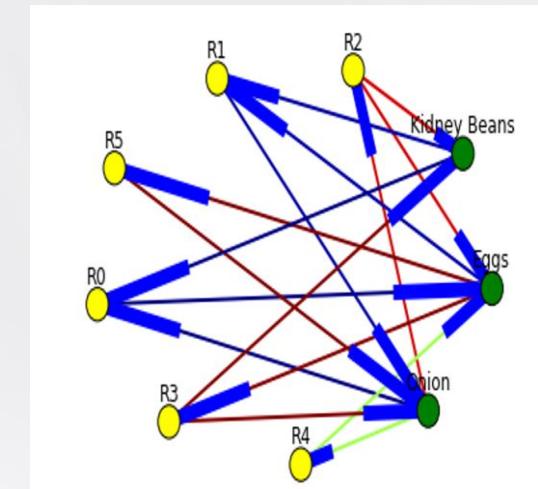
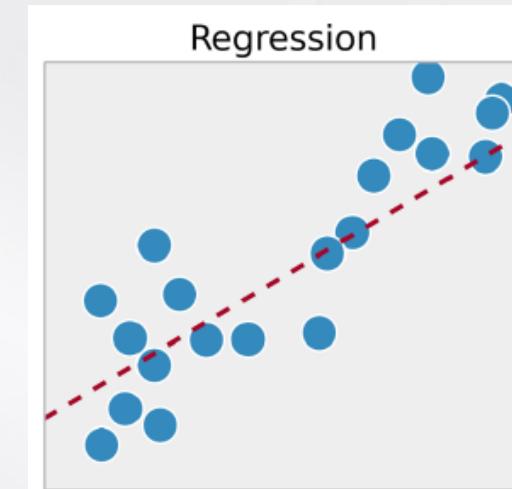
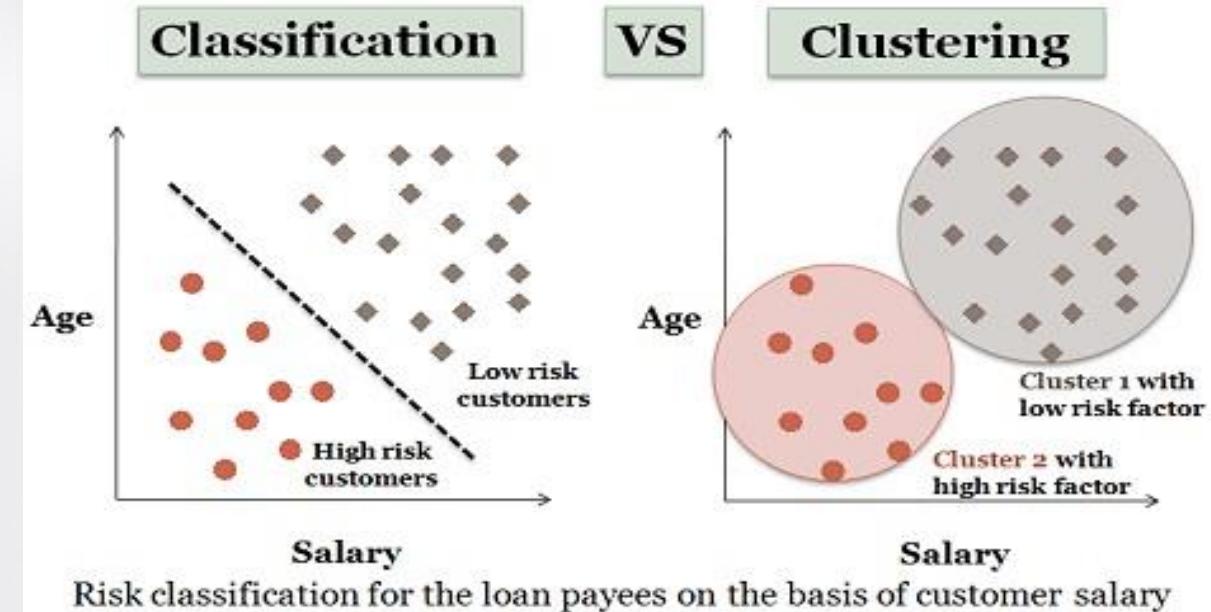
Data Mining Algorithms

Supervised

- Classification
 - Decision Tree
 - KNN
 - NB
 - SVM
 - ANN
 -
- Regression
 - linear
 - Polynomial

Unsupervised

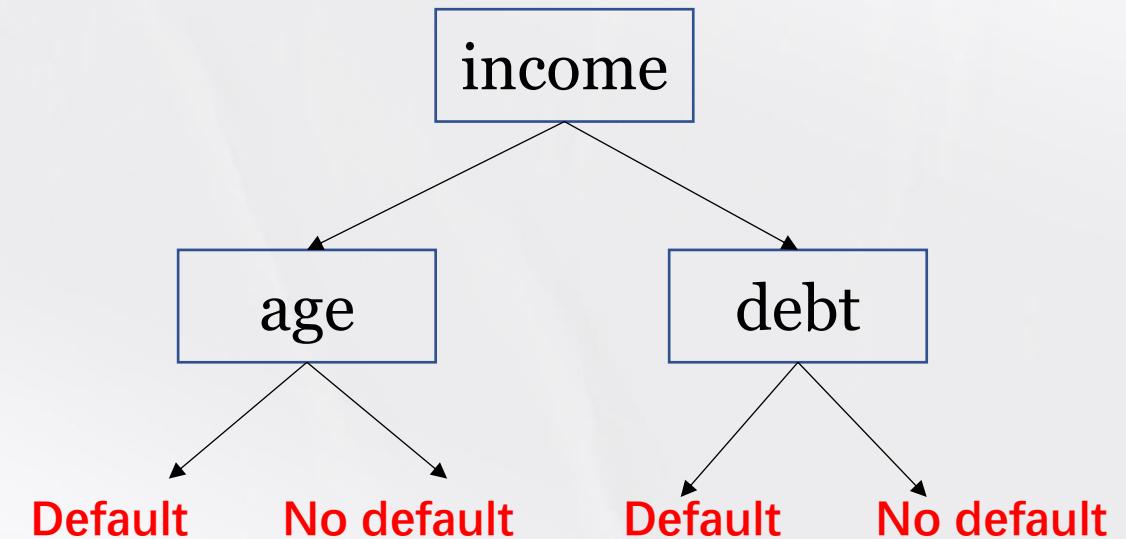
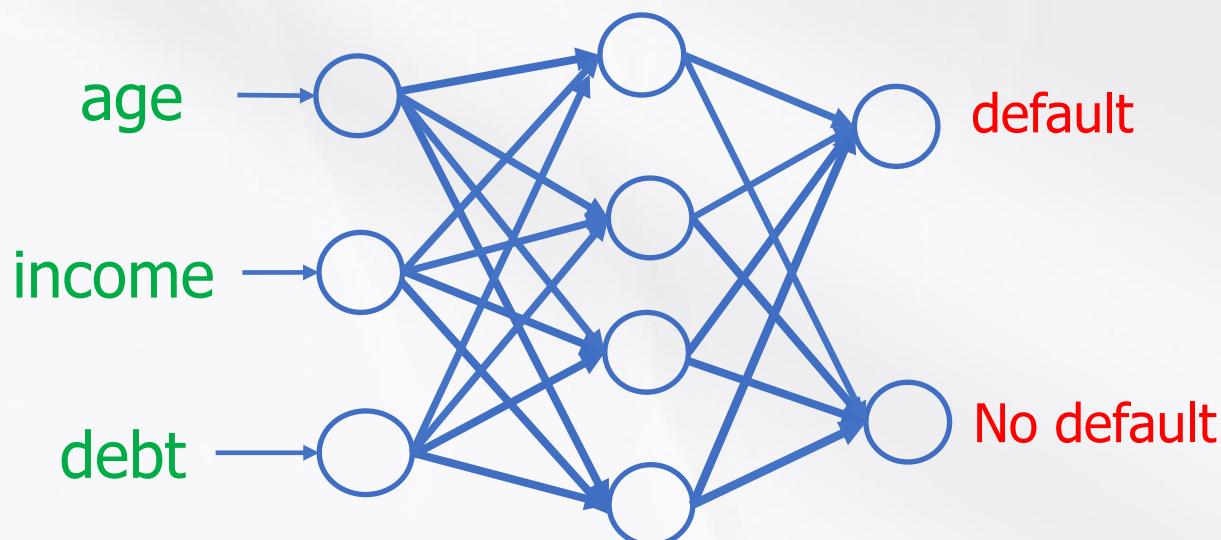
- Clustering
 - K-means
 - Density-Based Method
 - hierarchical
 - Korhonen
 -
- Association Analysis
 - Apriori
 - FP-Growth
 - Eclat Algorithm
 -



04 Data mining activities

✓ Classification

Assign object to one of a number of predefined classes.



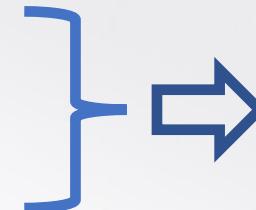
04 Data mining activities

✓ Clustering

Segmenting a population into some non-predefined classes
– Customer classes

call duration during working hours
call duration at other times
local call duration

– Market segments



business users
common users
users who make
fewer calls

04 Data mining activities

✓ Affinity grouping (关联分组) /association Rules

Determine what things go together

- Heart of “market-basket analysis”
- Examples:
 - People who liked the book of “Jane Eyre ” often liked “Pride and Prejudice ”
 - The Advanced Scoutsystem analyzed the logs of NBA games to uncover interesting pieces of information which might otherwise go unnoticed by coaches (e.g., “**IF player X is on the floor, then player Y's shot accuracy decreases from 75% to 30% (sup.=80%, conf.=70%) .**”)

POPULAR DATA MINING TOOLS

IBM SPSS

IBM'S STATISTICAL PACKAGES FOR THE SOCIAL SCIENCES IS A HIT FOR LARGE-SCALE PROJECTS



PYTHON

THE EASE OF USE HAS MADE PYTHON ONE OF THE MOST EFFECTIVE DATA MINING TOOLS



KNIME

KNIME'S USER-FRIENDLY FRAMEWORK FOCUSES ON DATA PIPE-LINING AND INTERACTIVE TABLES



H2O

WITH CUTTING-EDGE TECHNOLOGY, H2O HAS AN ENTHUSIASTIC USER COMMUNITY



R

FREE AND EASY TO PICK UP FOR NON-PROGRAMMING BACKGROUNDS



RAPIDMINER

AN OPEN-SOURCE PREDICTIVE ANALYTICS SOFTWARE FOR DATA MINING PROJECTS



SAS

THE DATA MINING TASKS IN SAS ARE GREAT FOR ENTERPRISE-LEVEL WORK



ORANGE

FREE AND IDEAL FOR BEGINNERS, IT HAS PRE-INSTALLED DATA MINING WORKFLOWS

SPARK

SPARK ALLOWS YOU TO FLOAT THROUGH OCEANS OF COLLECTED DATA WITH EASE