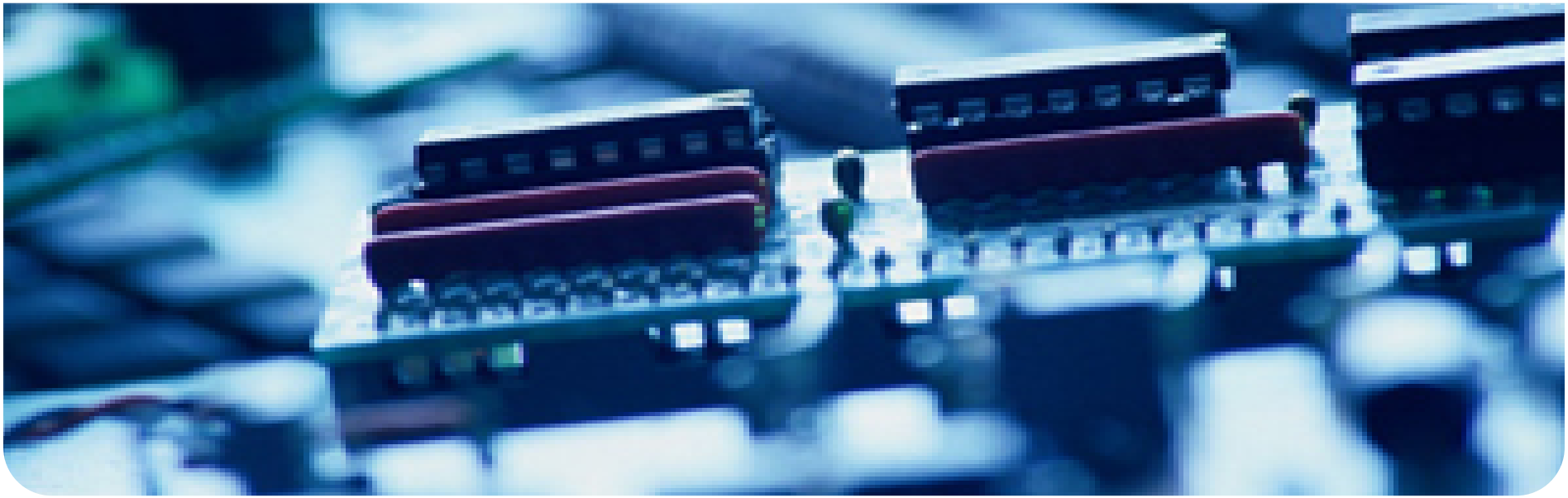
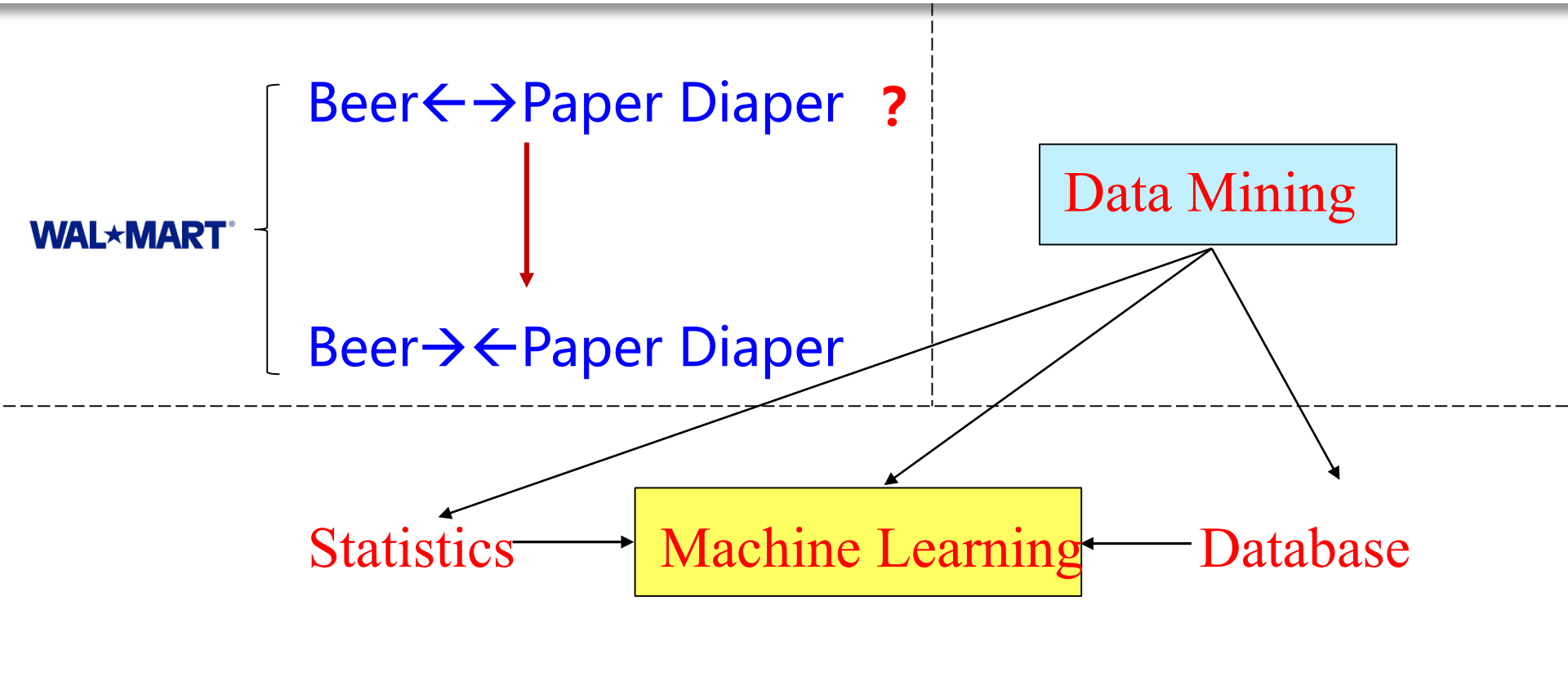
数据挖掘和大数据分析



DATA ANALYTICS:

DATA MINING AND BIG DATA

What is Data Mining?



What is Data Mining?

- **Discovery of useful, possibly unexpected, patterns in data.**
- **What is Pattern?**
  - Statistic Patterns
  - Machine Learning

投票　最多可选1项

设置

How do you understand DM and AI?

A　DM = =AI

B　DM ! = AI

提交

Data Mining

【Materials】

Data Dataset Database （Diamond）

【Tools】

traditional method/Modern （ML）

【Objective】

Knowledge Rules …… （Diamond）

Data Mining Contents

① Statistics

concentrate on **models**.

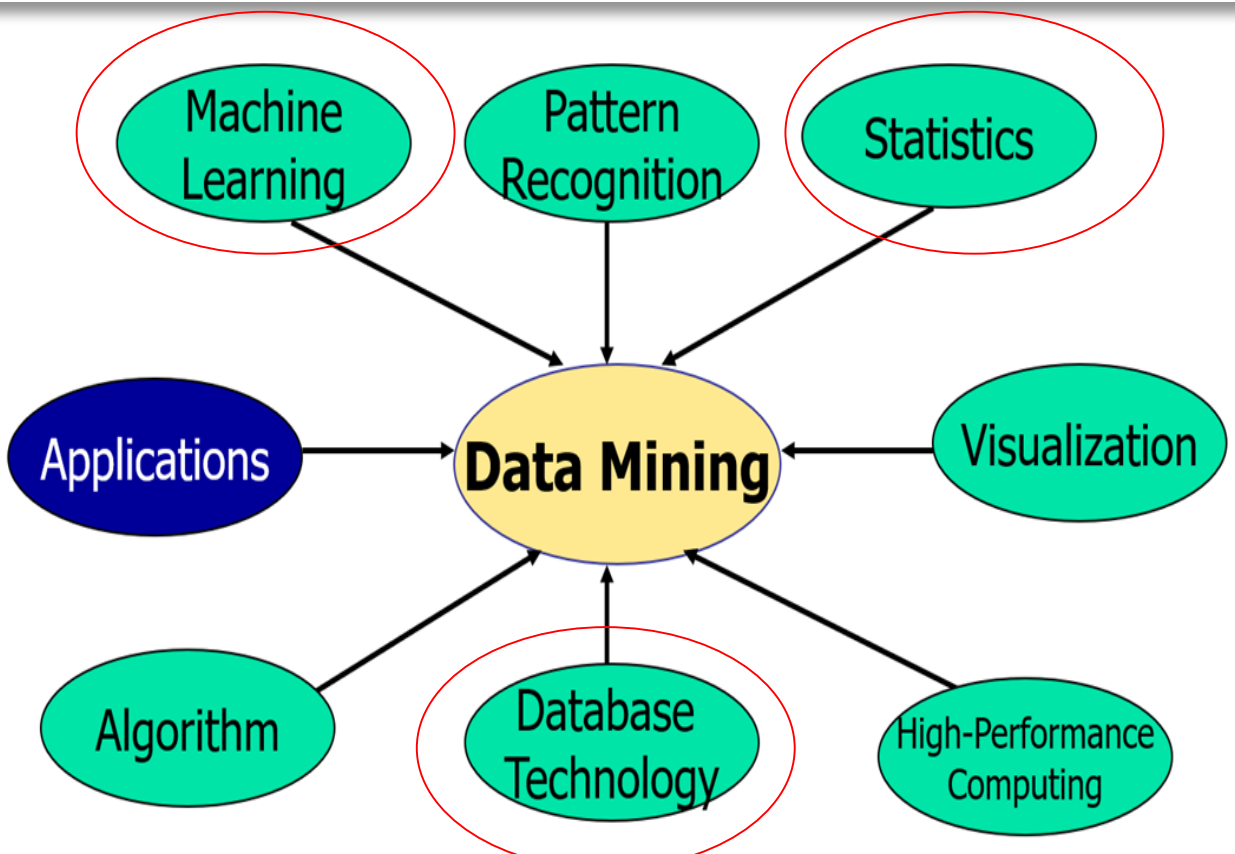
② AI (Machine-Learning)

concentrate on **complex methods**, **small** data.

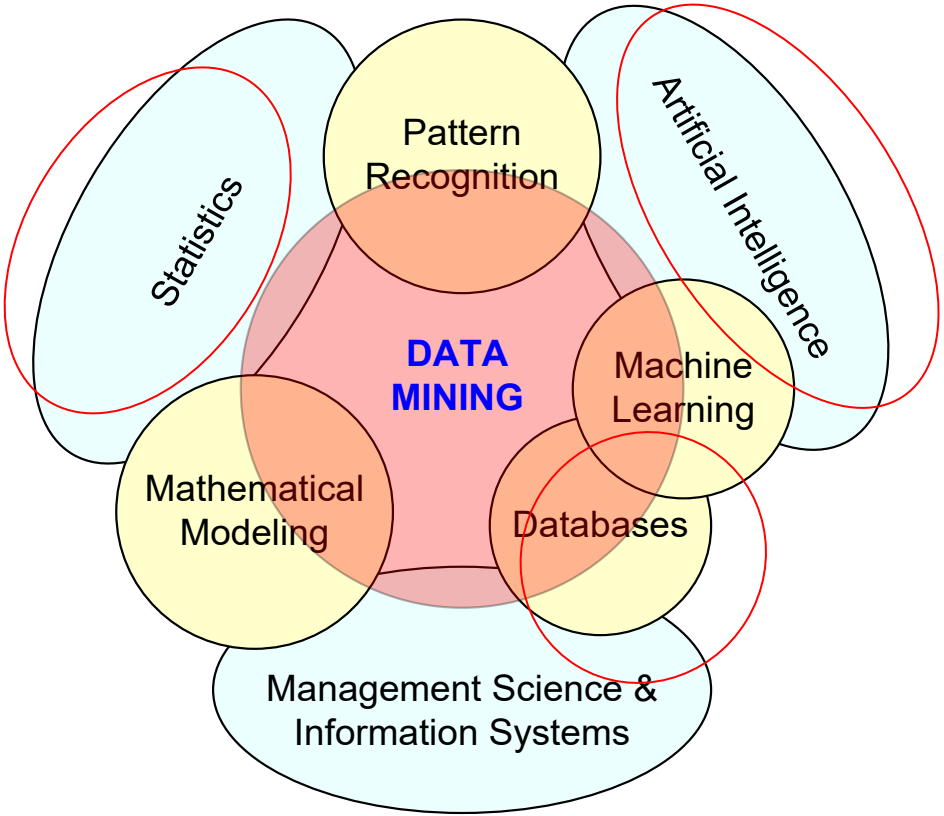
③ Databases

concentrate on **large-scale** (non-main-memory) data.

Data Mining Contents



Data Mining Contents



Objectives and Schedule

O1: Master Machine Learning classic algorithm

O2: Master classic Statics Methods

O3: Master Basic Data Mining Experiments

O4: Solve some realitic examples

Tools

Application

W1: Statics

W2: Statics Quiz1

W3: Machine Learning

W4: Machine Learning Quiz2

W5: Machine Learning Quiz3

W6: Map-Reduce /Spark / Hadoop Quiz4

W7: Project Defence

W8: Project Defence Final Exam

60% = 15% *4

30%

Team Project: 10%

投票　最多可选1项

设置

Do you understand Objectives and Schedule of the course?

A　Yes

B　No

提交

Data Mining Process

① **Data Cleaning** 数据清理（消除噪声或不一致数据）

② Data Integration**数据集成**（多种数据源可以组合在一起）

③ Data Selection 数据选择（从数据库中检索与分析任务相关的数据）

④ Data transformation数据变换（数据变换或统一成适合挖掘的形式）

⑤ Data Mining Method 挖掘方法（使用各种方法提取数据模式）

⑥ Pattern Assessment 模式评估（使用某种度量，识别真正有价值的模式）

⑦ Knowledge Representation 知识表示（使用可视化和知识表示技术，向用户提供挖掘的知识）

投票　最多可选4项

设置

Do you think which section is more important?

A Data Cleaning

B Data Integration

C Data Selection

D Data transformation

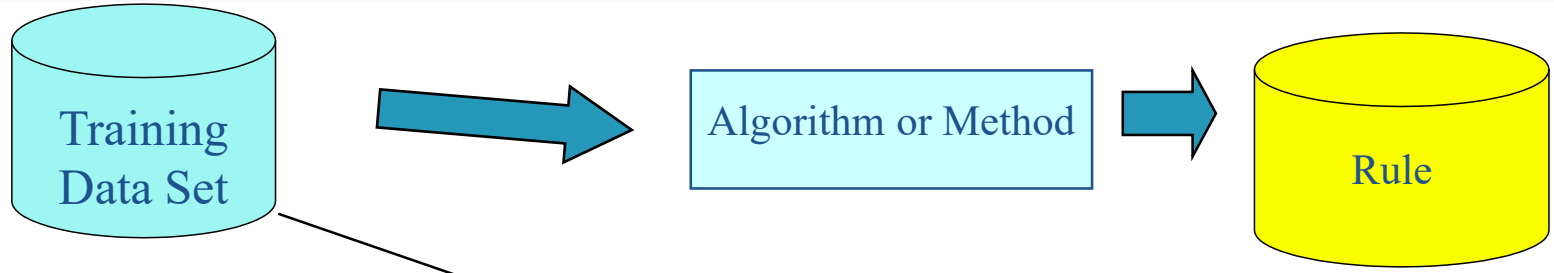
E Data Mining Method

F Pattern Assessment

G Knowledge Representation

提交

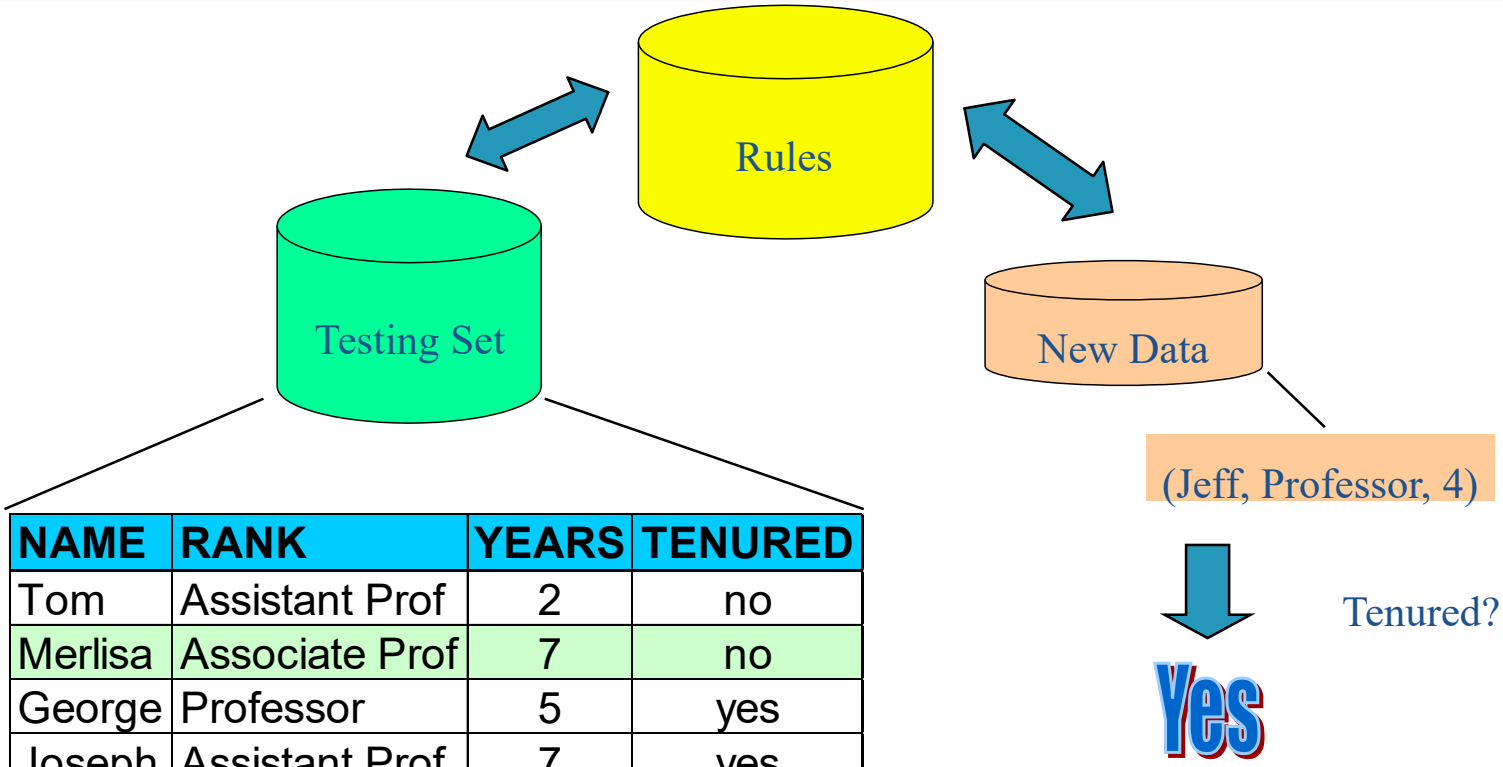
DM Lab 1



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’ OR years > 6
THEN
tenured = ‘yes’

DM Lab 1



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

DM Tools -Orange

orange

Features Screenshots Workflows Download Blog Docs

Data Mining
Fruitful and Fun

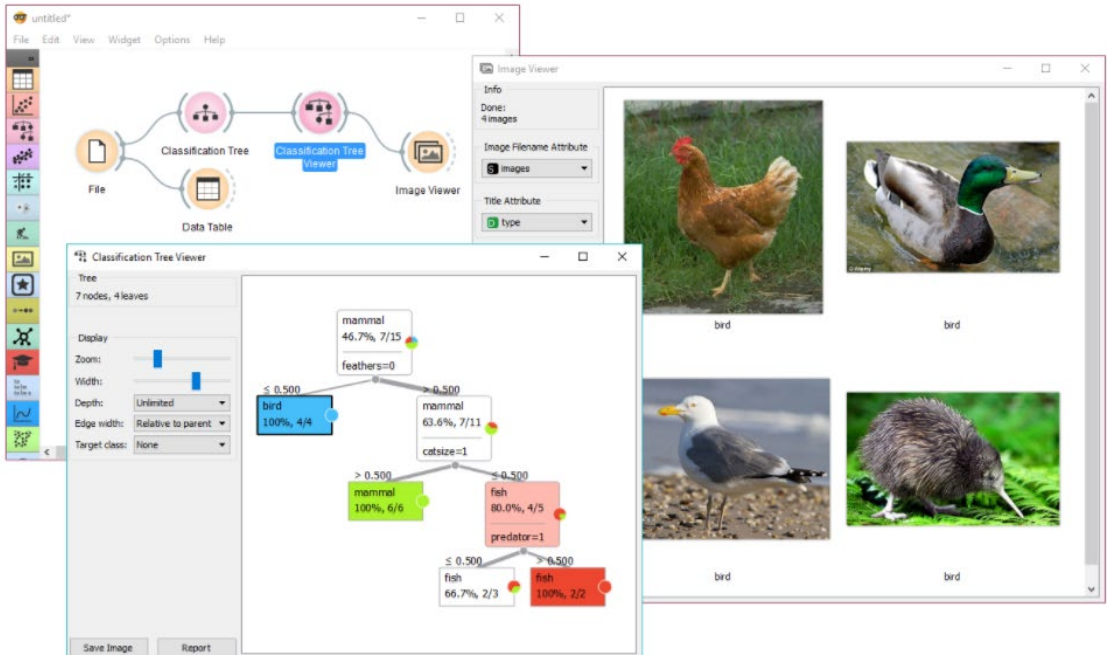
Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.

Download Orange

DM Tools –Orange

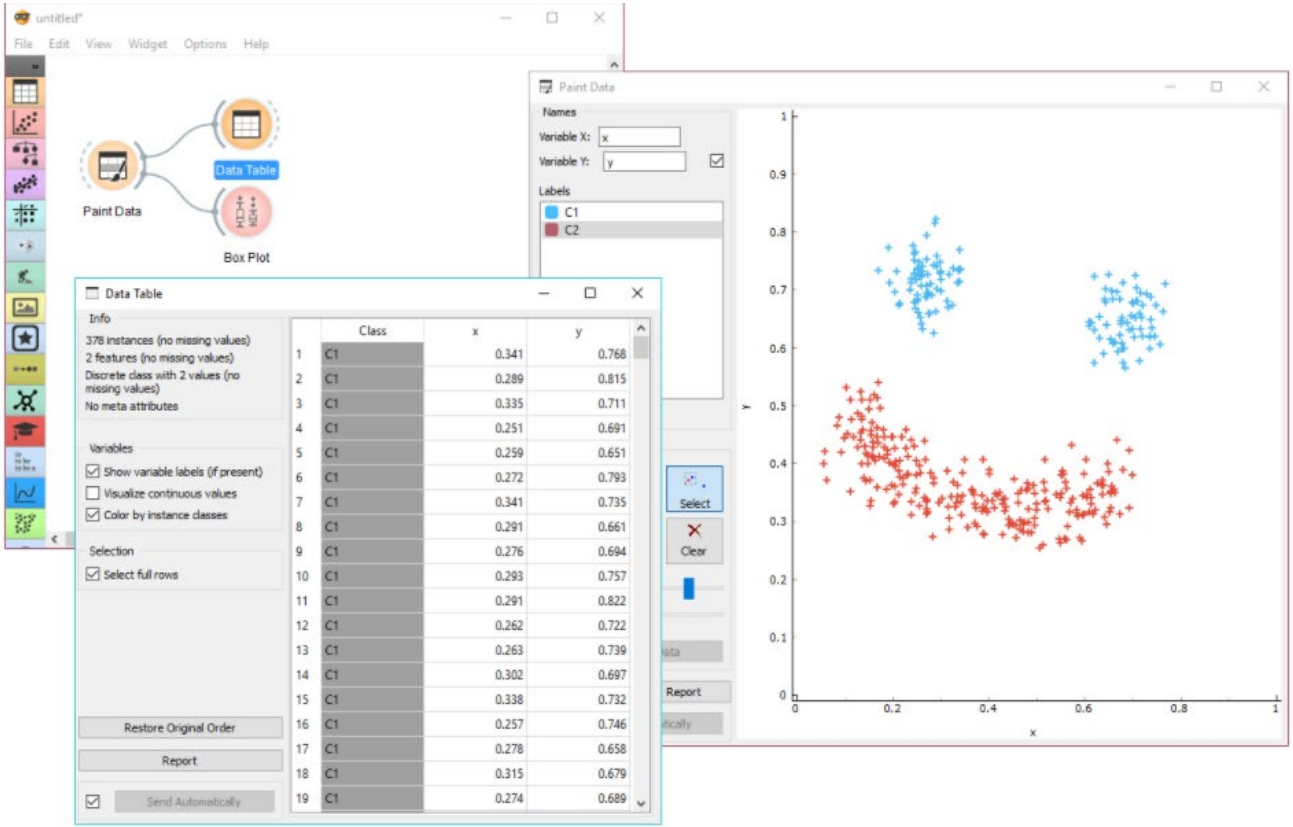
Orange (http://orange.biolab.si/)

Example 1



DM Tools -Orange

Example 2

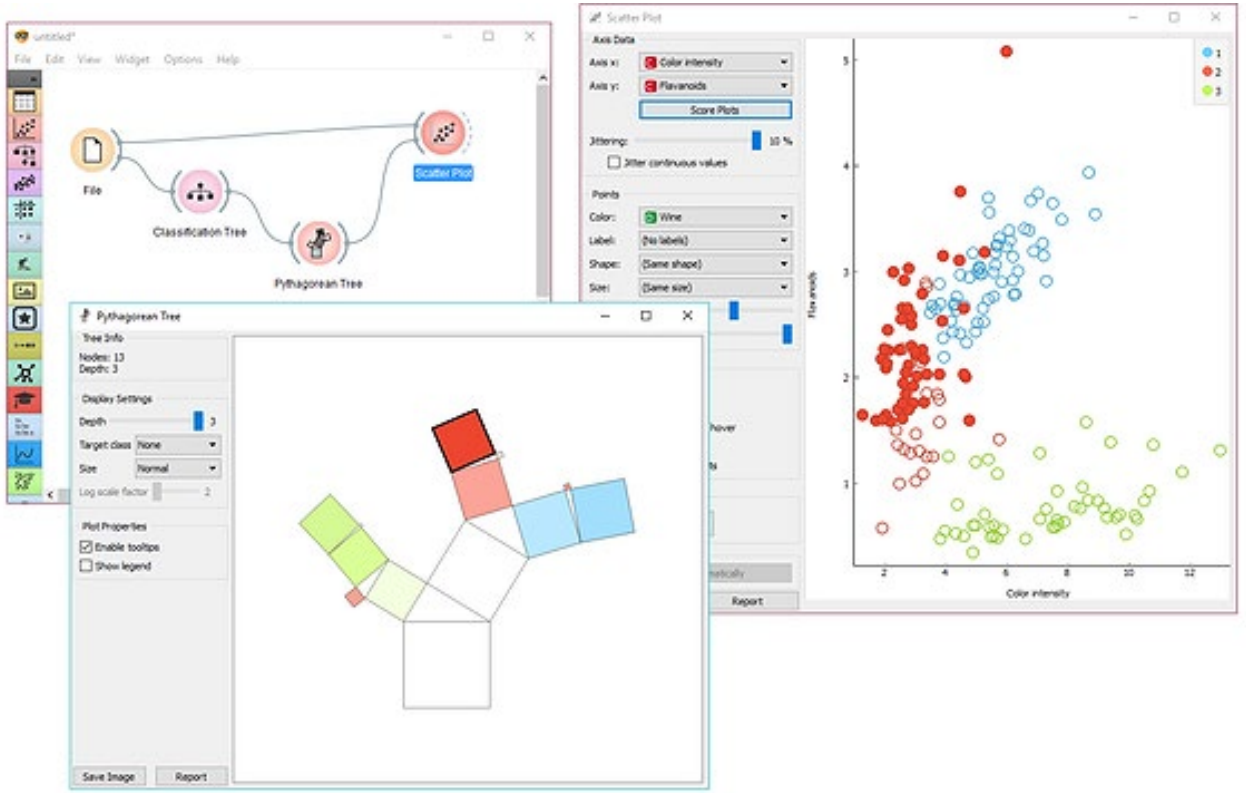


DM Tools -Orange

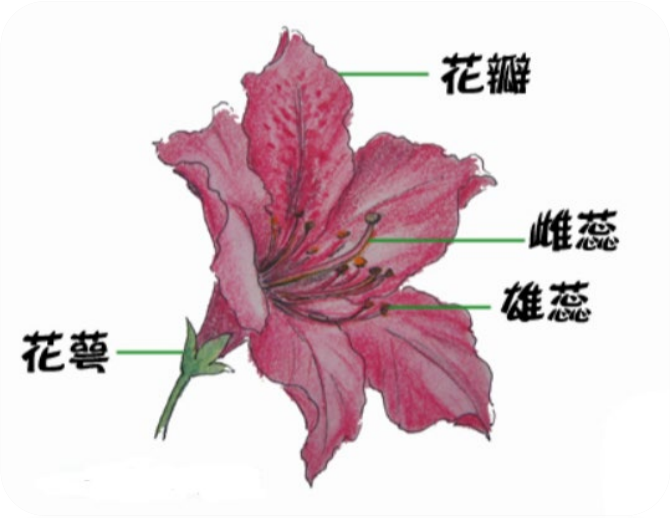
Example 3

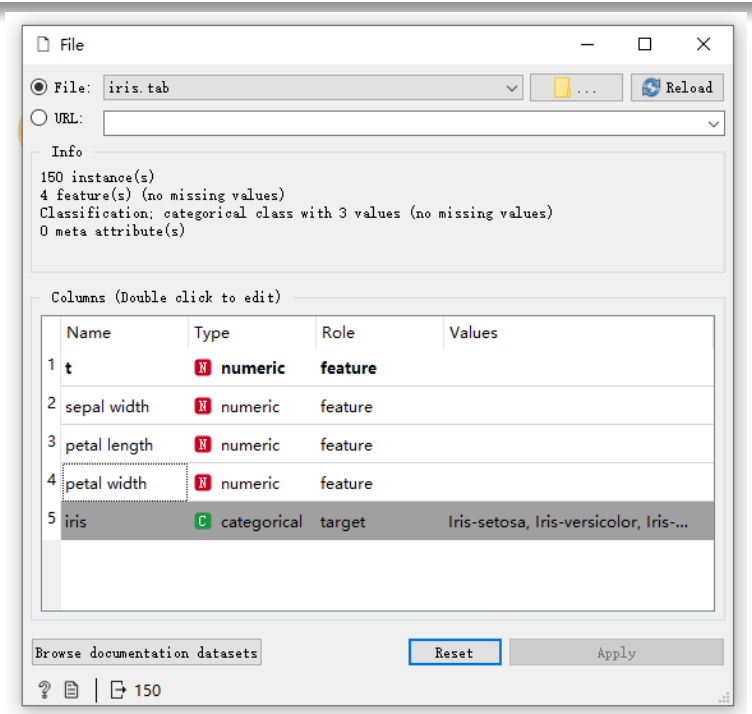
DM Tools -Orange

Example 4



DM Tools －Orange Dataset





https://docs.biolab.si//3/data-mining-library/reference/data.io.html

DM Tools －Orange & Python Load

```
>>> import Orange
>>> data = Orange.data.Table("iris")
```

相对路径

https://blog.csdn.net/qq_42571592/article/details/90734149

DM Tools -Orange Chinese

DM Lab 1 - Show Data

```
import Orange
data = Orange.data.Table("lenses")
print("Attributes:", ", ".join(x.name for x in data.domain.attributes))
print("Class:", data.domain.class_var.name)
print("Data instances", len(data))

target = "soft"
print("Data instances with %s prescriptions:" % target)
atts = data.domain.attributes
for d in data:
    if d.get_class() == target:
        print(" ".join(["%14s" % str(d[a]) for a in atts]))
```

Summary

- Content: Introduce to Data Mining and Data Analytics
- Hope: I will change it into "Data Analytics under Business View"

Gregory Piatetsky-Shapiro

贵在坚持！