INTRODUCTION

CSE801B

Contact Information

- Instructor: Zijun Cui
  - Email: cuizijun@msu.edu
  - Office location: Engineering building 2212
  - Office hours: By appointment via email

- TA: Francisco Santos
  - Email: santosf3@msu.edu
  - Office location: 3211 Anthony Hall
  - Office hours: Tuesday 3-4pm; Wednesday 3-4pm

  **For anyone who cannot attend either of these two slots, talk to me after class*

Course Webpage

- https://zijunjkl.github.io/teaching/

- Syllabus
- Course breakdown and slides

- Note: subject to minor changes as class progresses

Textbooks

- **Introduction to Data Mining (Second edition)**
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar
- **Data Mining: Practical Machine Learning Tools and Techniques (Third edition) ***
  - Ian Witten, Eibe Frank, Mark Hall
- **Data Mining For Business Analytics (Third edition) ***
  - Galit Shmueli, Peter C. Bruce, Nitin R. Patel
- **Python Data Science Handbook**
  - Jake VanderPlas

** Online through magic.msu.edu (MSU libraries online catalog)*

Assessment

- Homework: 35%
  - 5; every two weeks
- Exam 1: 15%
- Exam 2: 15%
- Project: 35%
  - Project proposal
  - Final report & Presentation
  - Group of 3-5

Class Policies

- Assignments must be submitted by the given deadline or special permission must be requested from instructor before the due date.
- Use D2L for assignment/submission/grade posting
- Late homework with no extension permission:
  - Within 2 days after the deadline: 50% penalty
  - More than 2 days after the deadline: Not Accepted
- Extensions will not be given beyond the next assignment except under extreme circumstances.

Class Policies

- [Plagiarism] May discuss ideas but submitted write up must be your own work. Any work that is found cheating will receive zero and will be reported to the University.

- [Attendance] Attendance at all regularly scheduled class meetings is a requirement of this course.

AI Guidelines

- You are welcome to use generative AI tools (e.g. ChatGPT, etc.) in this class as doing so aligns with the course learning goal.
- You are responsible for the information you submit based on an AI query. Remember, AI is not likely to generate a response that would be seen as quality work.
- No AI tools are allowed during exams and final presentation.

Important Dates

- Tentative Exam Dates:
  - Exam 1: Tuesday Oct 8th
  - Exam 2: Thursday November 21th
- Project Presentation:
  - 12/3 and 12/5
- Subject to minor change as the class progresses
  - Exam 1 will for sure be completed before the last day to drop the class (Oct 14th)

Programming Assignments

- Python

- HW Zero: get yourself familiar with Python.
  - Successfully install Python into your laptop/computer
  - Successfully import functions

  - Resources:
    - Supplemental slides
    - First two weeks of TA hours will provide help in this regard

Why Data Mining?

- Large amounts of data collected daily
  - **Business**: sales transactions, customer feedback, stock trading record, product descriptions: Walmart customers per week ~ 100 million
  - **Telecommunication**: Networks carry terabytes of data everyday
  - **Medical field**: generates huge amount of medical record, patient monitoring
  - **Engineering**: scientific experiments, environment monitoring, process measuring
- Data is of different types, and may not follow a particular distribution
- Difficult to analyze manually, important decisions made based on intuition not on data

Data Mining

- Powerful tools needed to automatically uncover valuable information
- Gap between data and information calls for development of data mining tool
- Natural evolution of information technology
  - Data collection
  - Database creation and management
  - Advanced data analysis
  - Data mining

Applications - Business

- Collect all information about customers purchases and interests
  - Point of sale data collection
  - Web logs from e-commerce
- Make informed business decisions
  - Customer profiling
  - Targeted marketing
  - Workflow management
  - Store layout
  - Fraud detection

Questions - Banking

- What potential factors will draw investors to the bank?
- What are the main factors that leave customers unsatisfied?
- What are the potential types of loans that might bring profit?
- What methods are commonly used to commit fraud?

Questions - Supermarket

- What items in the store are popular among teenagers?
- How likely is it that a vegetarian customer will buy non-vegetarian products?
- If an item is purchased by a customer, what other items are likely to be purchased at the same time?
- What kind of items should be stocked during the holiday seasons?

Applications - Healthcare

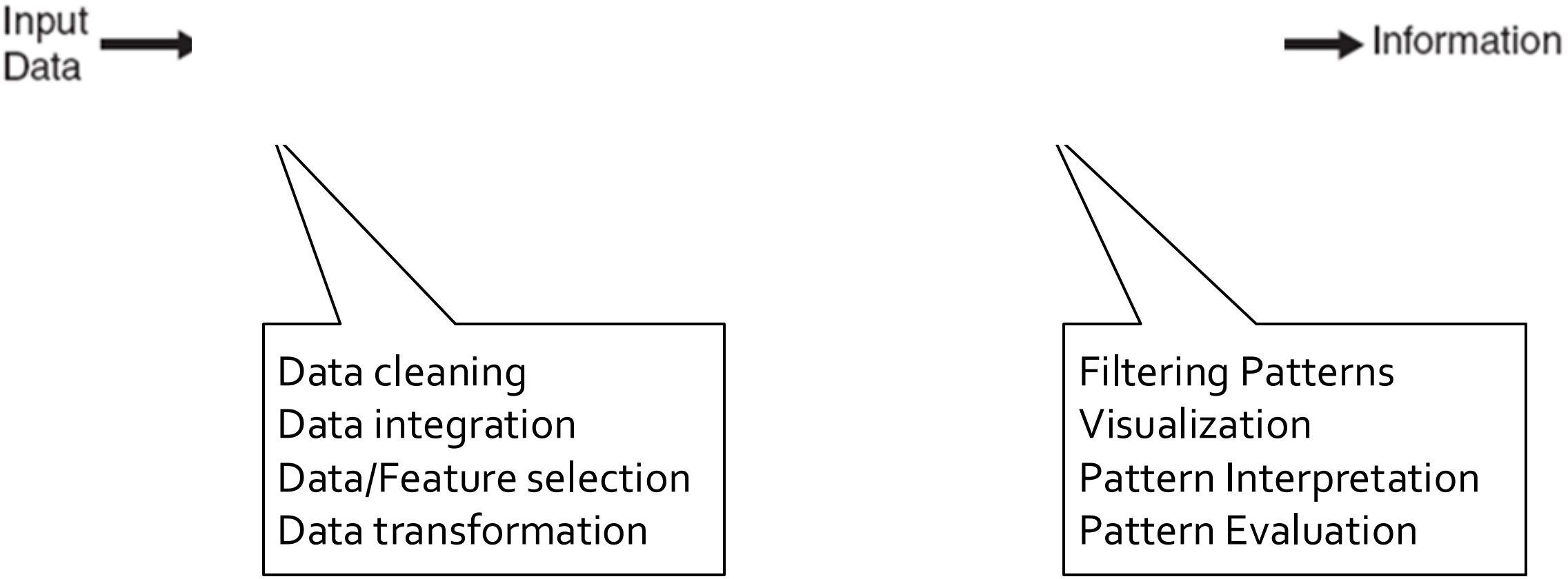
- Prediction patient outcomes
- Infection control
- Clinical research
- Treatment effectiveness

- Sample questions
  - How likely is it that an adult whose age is more than 70 and who has had a stroke will have a heart attack?
  - What are the characteristics of patients with a history of at least one occurrence of stroke?
  - What hospitals provide patients the best recovery rate?

What is Data Mining?

- The process of discovering interesting patterns and knowledge from large amounts of data
  - Involves data analysis methods with sophisticated algorithms
  - Part of Knowledge Discovery in Databases (KDD) process: converting raw data into useful information

What is Data Mining?



What kind of Data?

- Any data as long as it is meaningful for the target application
  - Tabular data
  - Sequence data
  - Graph data
  - Spatial data
  - Text data

Technologies

- Build upon methodology from existing fields:
  - Statistics: Sampling, estimation, modeling techniques, hypothesis testing
  - Machine learning and Pattern recognition: search algorithms, modeling techniques and learning theory
  - Information Retrieval
  - Database systems
  - Parallel and distributed computing

Challenges

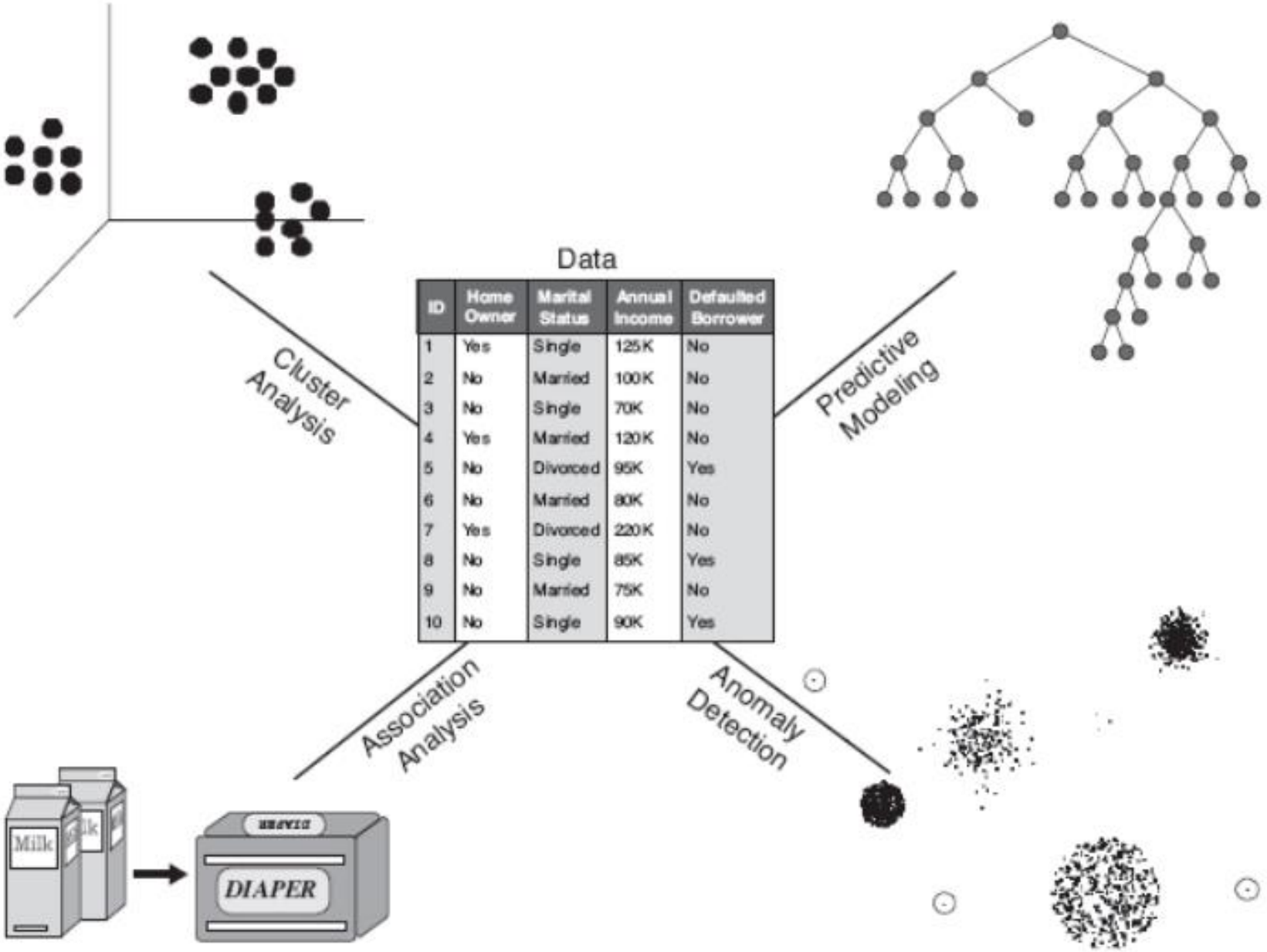
- Scalability: terabytes of data
  - need for efficient algorithms
- High dimensionality:
  - data with hundreds or thousands of attributes
- Heterogeneous and complex data:
  - web pages, DNA data, data with temporal and special correlation

Challenges

- Data ownership and distribution: data at different physical locations
  - Reduce communication
  - Consolidate results from multiple sources
  - Address data security issues
- Data analysis: hypothesis generation and tests
  - Thousands of hypotheses

Data Mining Tasks

- Two major categories:
  - Predictive tasks: predict the value of a particular attribute (**target variable**) based on the values of other attributes (**explanatory variables**)
  - Descriptive tasks: derive patterns that summarize relationships in the data
    - Correlations, trends, clusters, anomalies



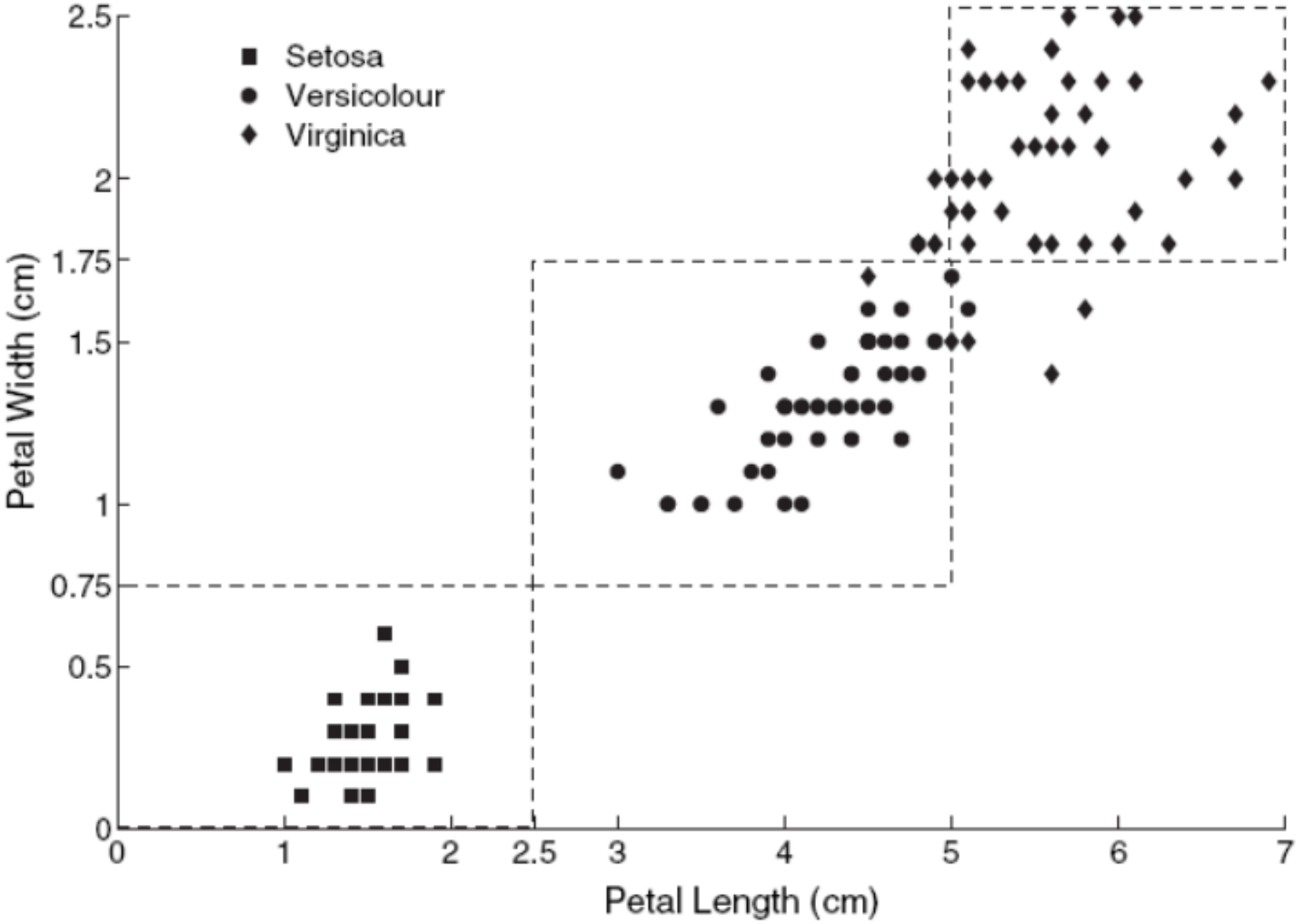
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Predictive Modeling

- Build a model for the target variable as a function of the explanatory variables.
- **Classification**: discrete target variables
  - Example: Predict whether a customer will renew contract (yes/no)
- **Regression**: continuous target variables
  - Example: Predict the future price of a stock

Classification Example

- Goal: classify an Iris flower to one of three Iris species
- Data: Iris data set
- *(Sepal width, sepal length, petal width, petal length, class)*


Setosa
Versicolour
Virginica
Petal Width (cm)
Petal Length (cm)
2.5
2
1.75
1.5
1
0.75
0.5
0
0
1
2
2.5
3
4
5
6
7

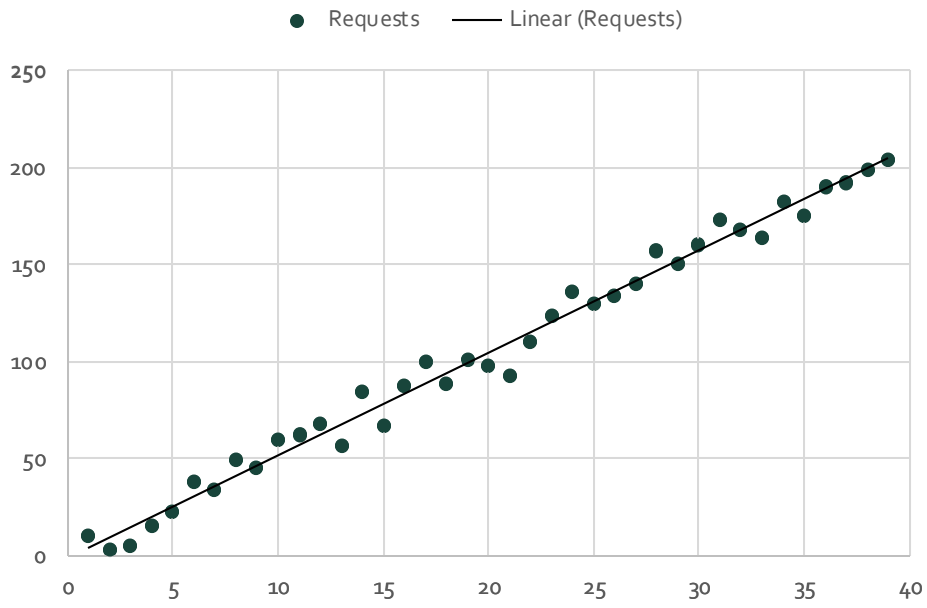
Classification Example

- Divide widths attributes into classes (low, medium, high) to simplify
- Rules:
  - *Petal width low and petal length low => Setosa*
  - *Petal width medium and petal length medium => Versicolour*
  - *Petal width high and petal length high => Virginica*

- Good classification but not perfect

Regression Example

- Goal: predict the number of help desk requests in the upcoming weeks
- Dataset: help desk logs
- Good prediction but some error



Association Analysis

- Used to discover patterns that describe strongly associated features in the data
- Discovered patterns represented as implication rules
- Search space is exponential
- Goal is to extract the most interesting patterns

Association Example

- Goal: find items that are frequently bought together

- Rules:
- *{Diapers} -> {Milk}*
- *{Bread} -> {Butter, Milk}*

Trans. ID	Items
1	{bread, butter, diapers, milk}
2	{coffee, sugar, cookies, salmon}
3	{bread, butter, tea, eggs, milk}
4	{butter, diapers, milk, eggs, cookies}
…	…

Clustering

- Finds groups of closely related observations such that observations that belong to the same group are more similar to each others than to those belonging to other clusters
- Applications:
  - Astronomy: aggregation of stars, galaxies, …
  - Biology: Plants and animal ecology
  - Medical imaging
  - Market research

Clustering Example

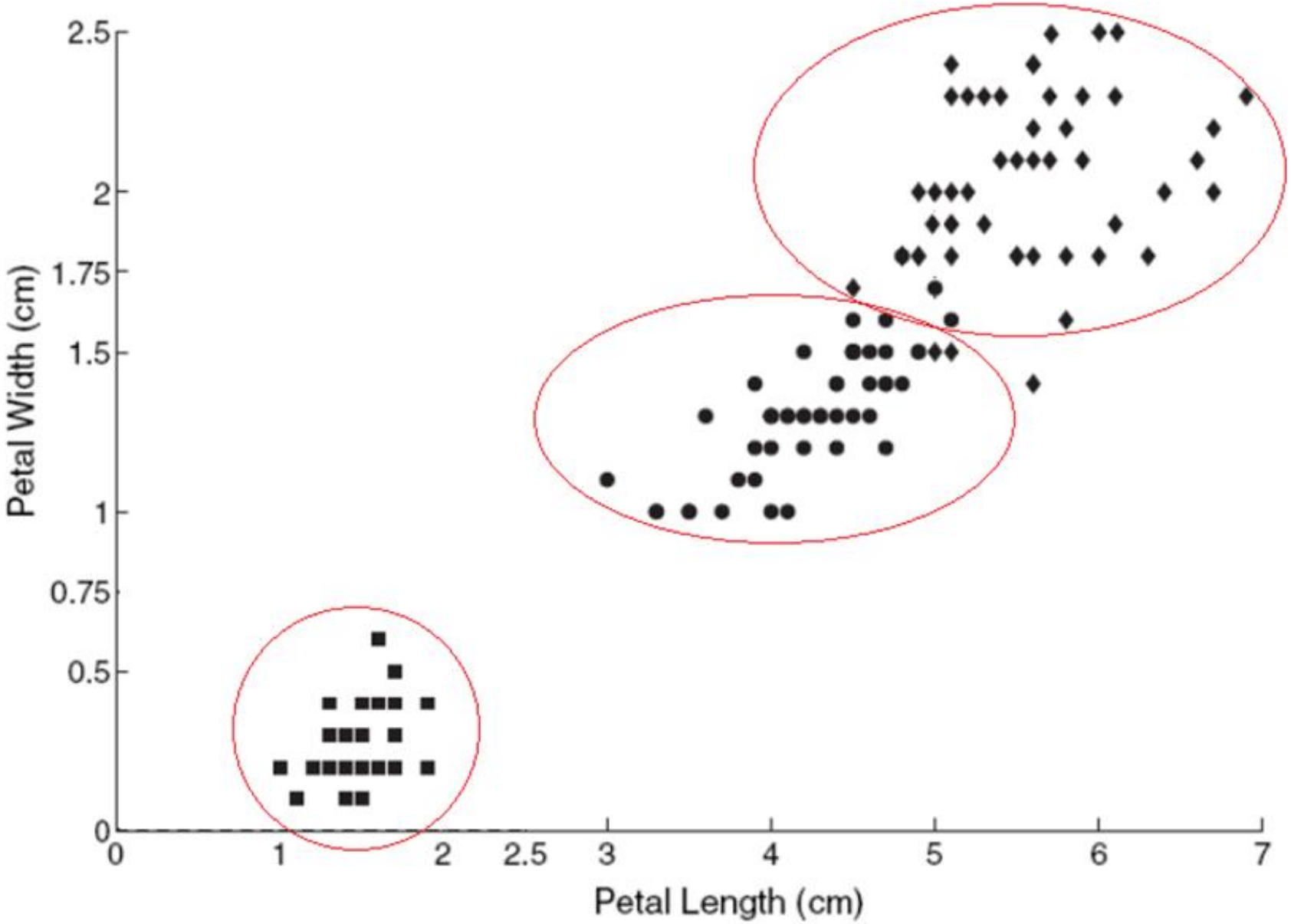
- Goal: group related document together
- Each document represented by list of pairs (w, c) denoting each word and number of occurrences

1: (dollar, 1), (industry, 4), (country, 2), (labor, 2), (death, 1)

2: (machinery, 2), (labor, 3), (market, 4), (country, 1)

3: (death, 2), (cancer, 1), (health, 3)

….


Petal Width (cm)
2.5
2
1.75
1.5
1
0.75
0.5
0
0
1
2
2.5
3
4
5
6
7
Petal Length (cm)

Anomaly Detection

- Identifies observations whose characteristics are significantly different from the rest of the data => **Anomalies or Outliers**
- Applications:
  - Fraud detection
  - Network intrusions
  - Unusual patterns of disease
  - Ecosystem disturbances

Summary

- Why Data Mining
- What is Data Mining
- Steps/Technologies involved
- Challenges

Course Outline

- Preprocessing techniques
- Classification
- Association
- Regression
- Clustering
- Anomaly detection
- Text mining
- Time series mining
- Project Presentations

Case studies
Applications in Python