CS 4821 – Data Mining

CS 5831 – Advanced Data Mining

Laura E. Brown

Jan. 8, 2024

Some slides from G. Piatetsky-Shapiro; Han, Kamber, & Pei;P. Smyth; C. Volinsky; Tan, Steinbach, & Kumar; J. Taylor; G. Dong; M. Hahsler

About Me

Laura Brown

- Rekhi 307
- lebrown@mtu.edu
- Office hours: TBA

Research Interests

- Machine Learning
- Artificial Intelligence
- Data Science

Courses I teach:

- CS 2311
- CS 4811 / CS 5811
- CS 4821 / CS 5831
- UN 5550

Agenda

- Course Logistics
- What is Data Mining?
- Why Data Mining?
- Examples of Data Mining

COVID-19 and other Illness

- If you are not feeling well?
  - **Wear a mask, Do not come to class, let me know**

- If you test positive for covid-19?
  - **Do not come to class, let me know**

All class lectures will be recorded and posted on Canvas

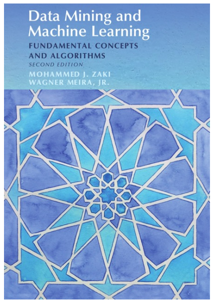
Course Logistics

Website on Canvas holds all relevant information

- Syllabus
- Schedule
- Modules
  - Lecture materials (slides, videos)
  - Assignments
  - Additional references, Materials, and Links
- Discussion – hosted on EdStem
- Assignments
  - Submission on Canvas and Gradescope
- Grades

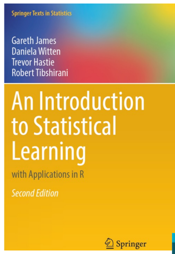
Textbooks

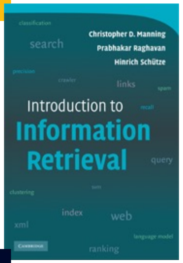
Required Textbook

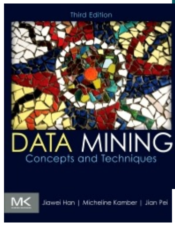

Data Mining and Machine Learning
FUNDAMENTAL CONCEPTS AND ALGORITHMS
SECOND EDITION
MOHAMMED J. ZAKI
WAGNER MEIRA, JR.

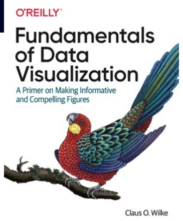
- Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Ed. Mohammed J. Zaki, Wagner Meira, Jr. Available online: https://dataminingbook.info/

Supplemental Reading





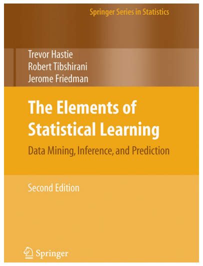



O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

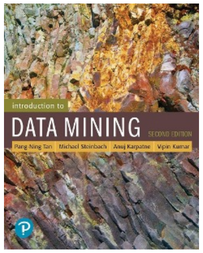
- Introduction to Statistical Learning, 2nd Ed.
  by James, Witten, Hastie, Tibshirani
  https://statlearning.com/
- Introduction to Information Retrieval
  by Manning, Raghavan, Schutze
  https://nlp.stanford.edu/IR-book/
- Data Mining: Concepts and Techniques, 3rd Ed.
  by Han, Kamber, and Pei
  Can be found online
- Fundamentals of Data Visualization, 1st Ed.
  By Claus Wilke
  https://clauswilke.com/dataviz/

Textbooks

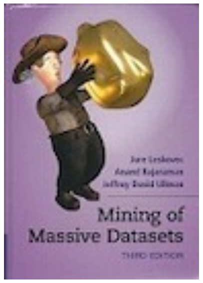
Recommended References


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

- Elements of Statistical Learning, 2nd Ed.
  by Hastie, Tibshirani and Friedman
  https://hastie.su.domains/ElemStatLearn/



- Introduction to Data Mining
  by Tan, Steinbach, Karpatne, and Kumar
  https://www-users.cs.umn.edu/~kumar001/dmbook/index.php



- Mining of Massive Datasets, 3rd Ed.
  by Leskovec, Rajaraman, and Ullman
  http://i.stanford.edu/~ullman/mmdsn.html

Course Components

Course Component	CS 4821	CS 5831
Assignments	65%	60%
Exam	20%	20%
Presentations	15%	
Final Project		20%
Total	100%	100%

Grading Scheme

- Your grade will be determined by your performance on assignments, project/presentations, and exams according to the weighted groups on the prior slide.
- Your score will determine your letter grade as shown below

Score	Letter Grade
> 93%	A
[88 – 93)	AB
[83 – 88)	B
[78 – 83)	BC
[73 – 78)	C
[68 – 73)	CD
[60 – 68)	D
< 60%	F

Grading

Assignments (~5-7), 65% / 60%

- You are expected to turn in individual work for the first two assignments, then group work for the remaining
- Many assignments divided into two parts: written problems and programming exercises
- Most assignments due through Gradescope
- Late assignments will receive a grade of zero with the following exception:
  - Each student has **8 late days**, which allow an assignment to be turned in 2 days late without penalty
  - A max of 2 late days can be used for a single assignment

Grading (2)

Programming exercises will support the following languages:

- R
  - Example code available for **all** topics
  - Many packages available and advised to be used with base functionality provided
- Python 3.9
  - Example code available for **all** topics
  - Many packages available and advised to be used, with base functionality provided
- Matlab
  - Example code available for **most** topics
  - Toolbox support for graphics, statistics and machine learning tasks, other methods not as well supported

Grading (3)

Exam, 20%

- Exam will be in Week 10-12
- The exam will be announced at least 7 days ahead of time
- Make-up exams must be arranged ahead of time

Grading (4)

- CS 4821 – Presentations
  - Read and present a current paper discussing data mining methods

- CS 5831 – Final Projects
  - Frame a data mining problem, collect data, select methods to solve problem, apply the methods, and evaluate the performance.
  - Several stages to the final project: initial idea, references, report and presentation

Collaboration and Cheating

- Students are encouraged to discuss course materials to understand the topics of the course, but not to produce homework/programming solutions
- Examples of Acceptable Collaboration:
    - Clarifying ambiguities or vague points in class handouts, textbooks, or lectures.
    - Discussing or explaining the general class material.
    - Providing assistance with general Python/R/Matlab questions, in using the system facilities, or with editing, debugging, and Python/R/Matlab tools.
    - Discussing the code that we give out on the assignment.
    - Getting help concerning programming issues which are clearly more general than the specific assignment (e.g., what does a particular error message mean?)

Collaboration and Cheating (2)

- Your submissions should be your own work
- Examples of Unacceptable Collaboration:
  - Copying files from another person or source, including retyping their files, changing variable names, etc.
  - Using a solution you find on the web and copying it into your solutions is cheating.
  - Allowing someone else to copy your code or written assignment, either in draft or final form.
  - Looking at someone else's files containing draft solutions, even if the file permissions are incorrectly set to allow it.
  - Receiving help from students or using course materials from previous years.
  - Use of any AI tools or software to generate your submissions.

Assignments will be checked using software to detect cheating. Any violations will be reported to the Office of Student Affairs

Collaboration and Cheating (3)

If you have questions on the course, consider the following options:

- Post a clear question on Ed.
  - If general, post publicly
  - If specific to your solution, post privately to instructors/TA
- Stop by office hours
- Send an email

Build a great community

- Help out your peers on Ed!
- Be mindful of the tone you use – be respectful and supportive, help everyone feel at home.

My Expectations

- Come to every class / watch every class
- Complete your work yourself / with your group
- Be respectful of your instructor and fellow classmates
- Ask questions
- Check-in to Canvas for announcements
- Adhere to MTU academic integrity policy

What is Data Mining?

What is Data Mining

One of many definitions:

"Data mining is the science ***of extracting useful knowledge*** *from huge data repositories."*

ACM SIGKDD, Data Mining Curriculum: A Proposal

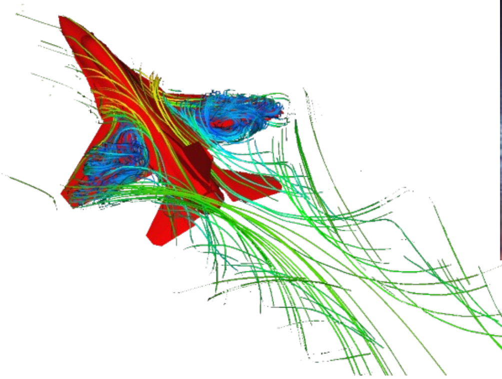
http://www.kdd.org/curriculum

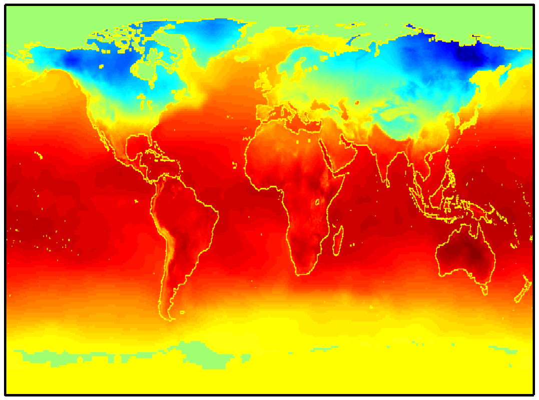
Why Data Mining? Commercial Viewpoint

- Businesses collect and warehouse lots of data.
  - Purchases at department/grocery stores
  - Bank/credit card transactions
  - Web and social media data
  - Mobile and IOT
- Computers are cheaper and more powerful.
- Competition to provide better services.
  - Mass customization and recommendation systems
  - Targeted advertising
  - Improved logistics

Why Mine Data?
Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
    - remote sensors on a satellite
    - telescopes scanning the skies
    - microarrays generating gene expression data
    - scientific simulations generating terabytes of data
- Data mining may help scientists
    - identify patterns and relationships
    - to classify and segment data
    - formulate hypotheses





Data Everywhere

- Drowning in data, but starving for knowledge
- Data may contain hidden information and patterns
- Human analysis may takes days/weeks/months/never find useful information
- Much of the data in never analyzed at all

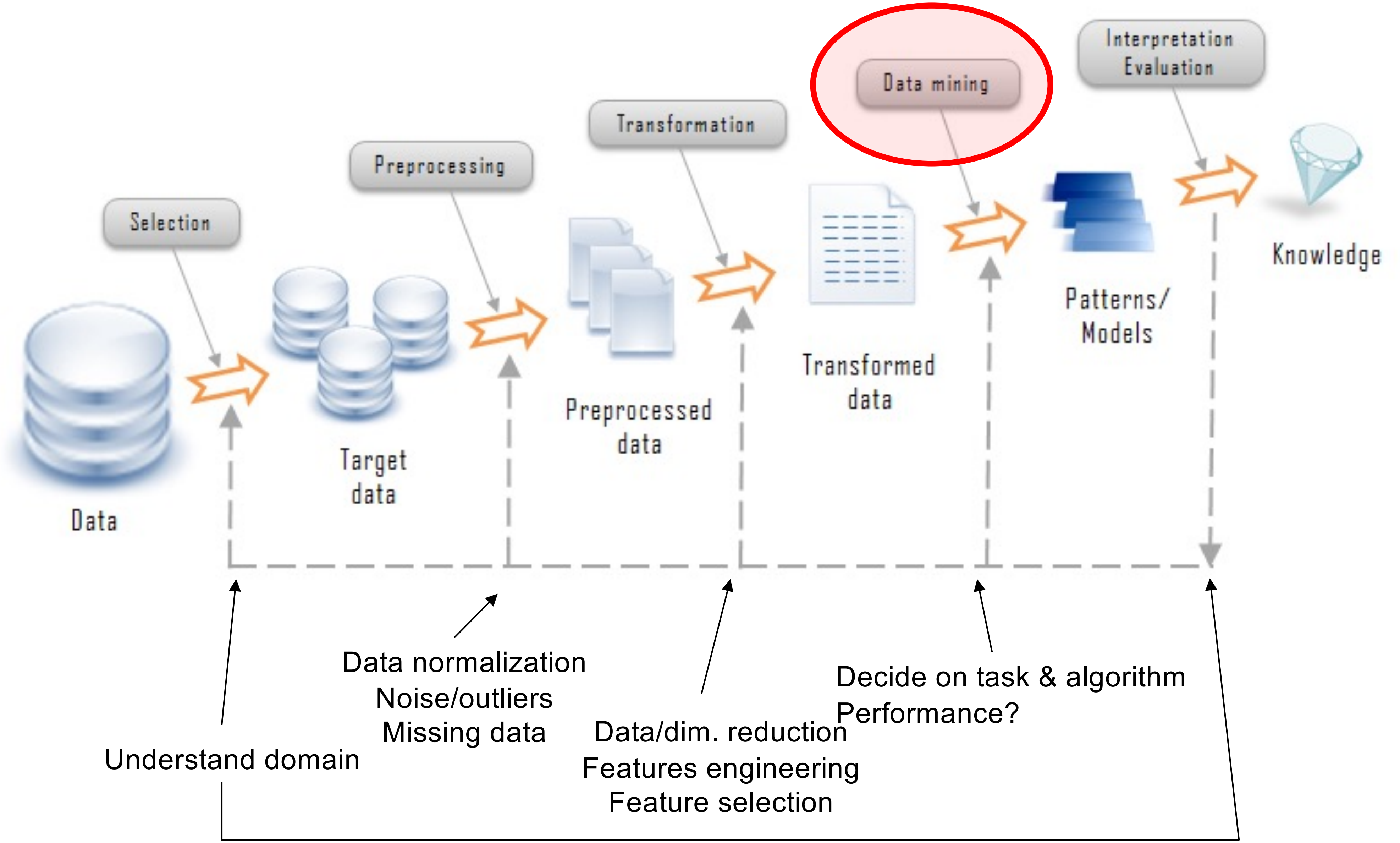
Alternative and Related Names

- Knowledge discovery in databases (KDD)
- Knowledge extraction
- Data / pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence
- Predictive analytics
- …

Closely Related Fields

- Statistics
  - Often more theory-based (parametric models)
  - Descriptive statistics, statistical inference, design of experiments
    - Hypothesize, collect data, analyze
- Machine Learning
  - Study of Algorithms to extract information automatically
  - Many sub-fields not part of data mining (robotics, RL)
- Data Mining
  - Integrates theory and heuristics
  - Focus on entire process: data collection, cleaning, learning, integration and visualization
  - Applications- and algorithm-oriented

Knowledge Discovery in Databases (KDD) Process



Understand domain

Data normalization
Noise/outliers
Missing data

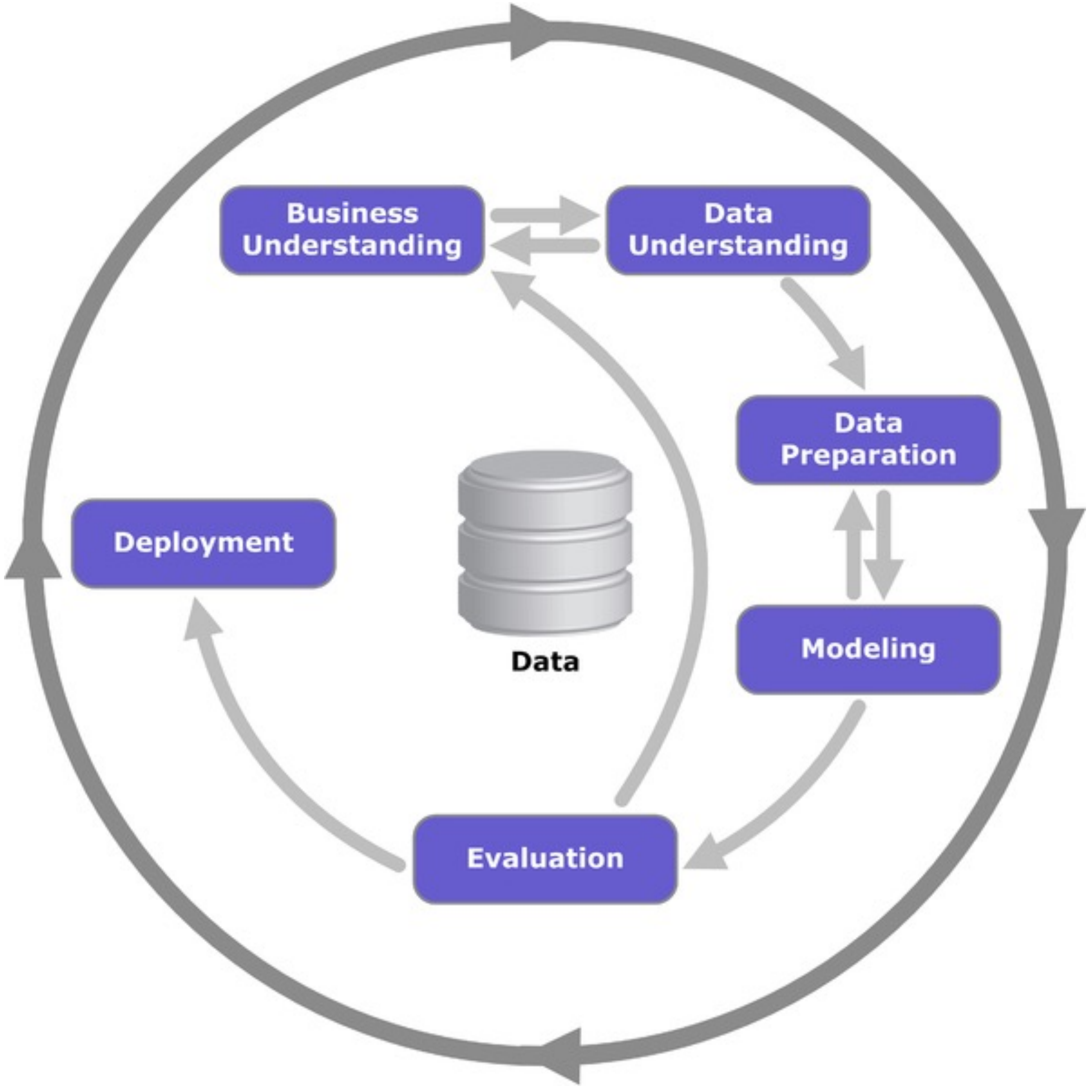
Data/dim. reduction
Features engineering
Feature selection

Decide on task & algorithm
Performance?

Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery: an overview.

CRISP-DM Reference Model

- Cross Industry Standard Process for Data Mining
- Open standard process model
- Industry, tool and application neutral
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own similar process.



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Examples of Data Mining

Data Mining Tasks

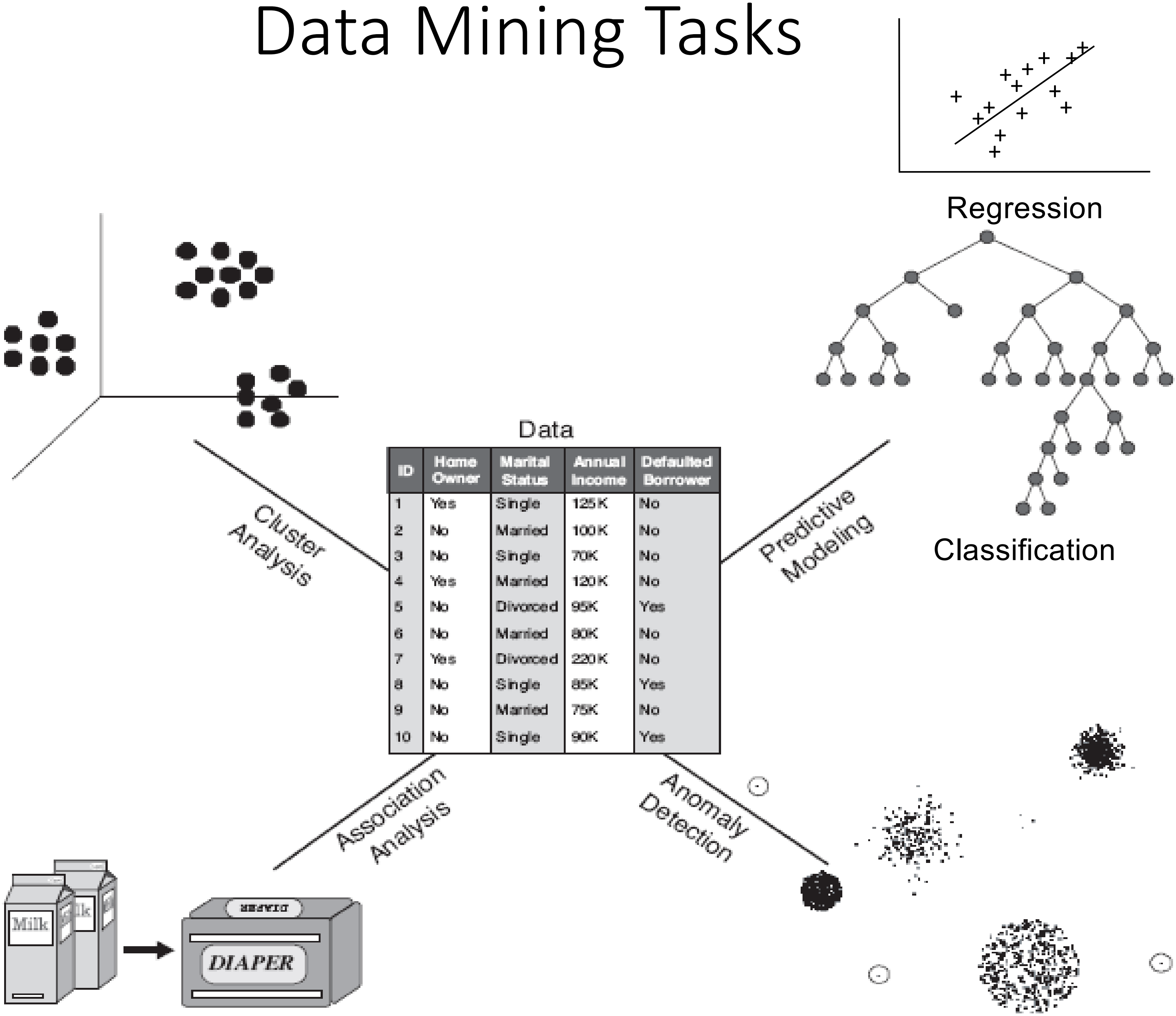
Descriptive Methods

Find human-interpretable patterns that describe the data.

Predictive Methods

Use some features (variables) to predict unknown or future value of other variable.

Data Mining Tasks

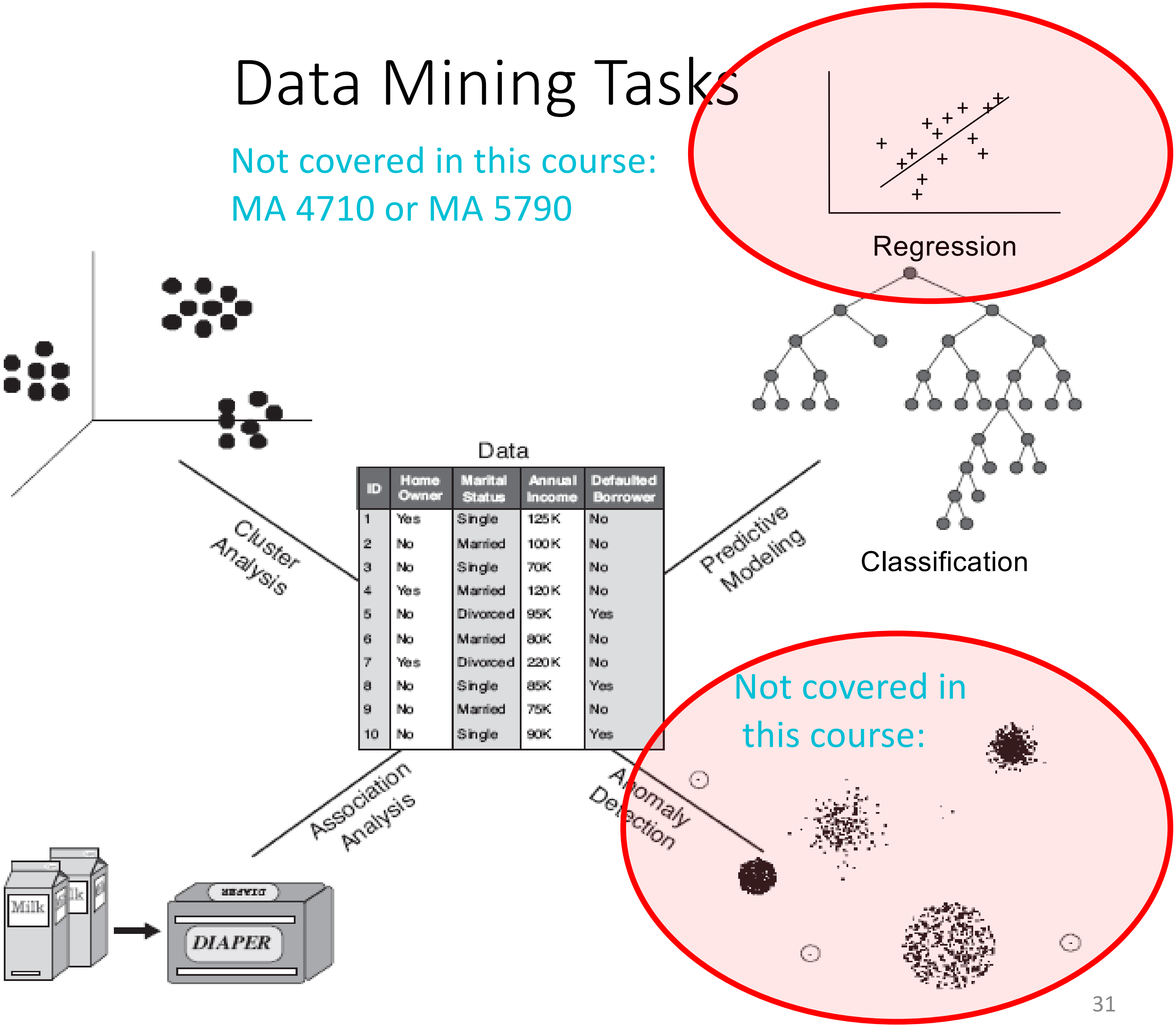


ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70K	No
4	Yes	Married	120 K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220 K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley, 2006

Data Mining Tasks

Not covered in this course:
MA 4710 or MA 5790

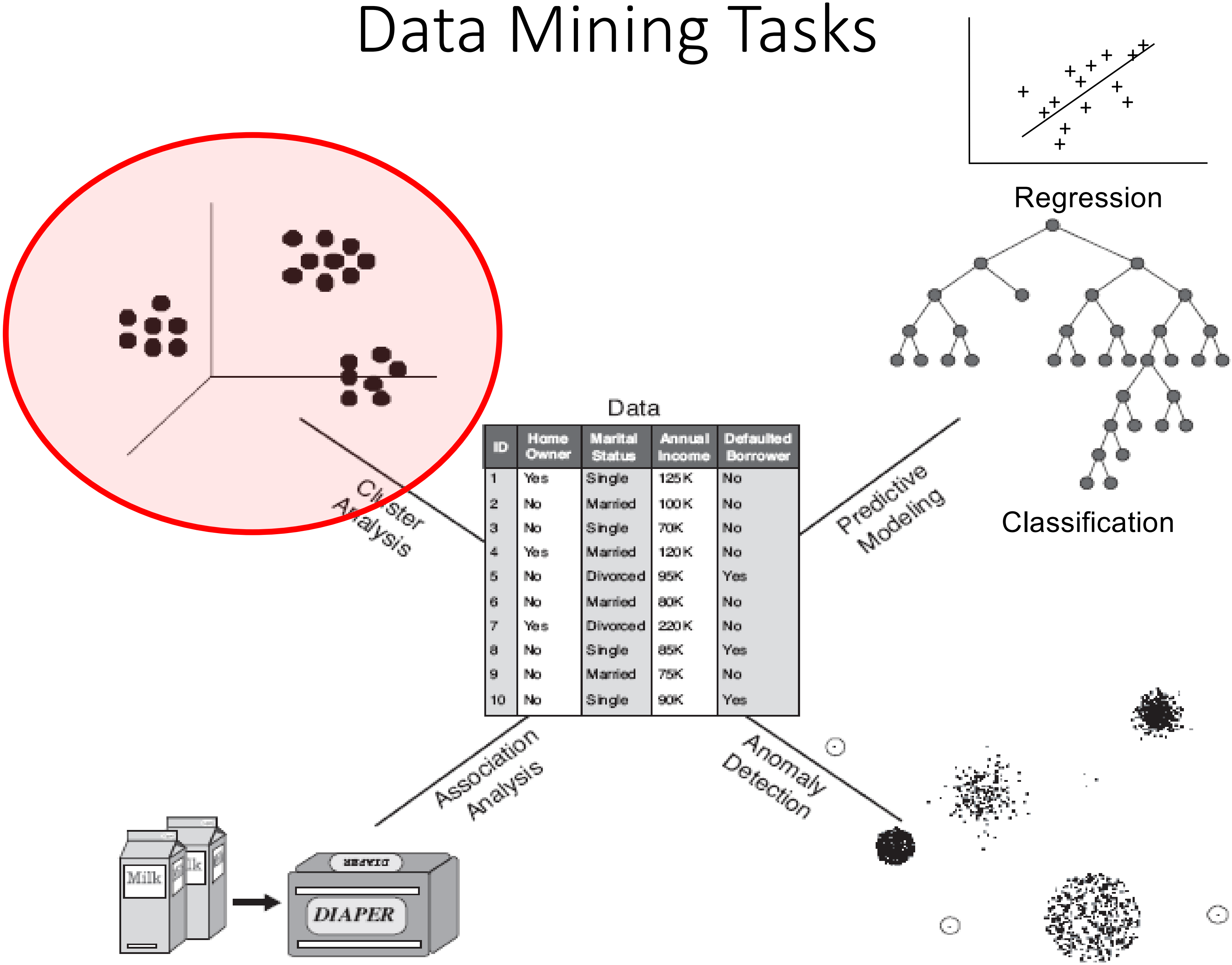


Data

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	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Not covered in this course:

Data Mining Tasks



Data

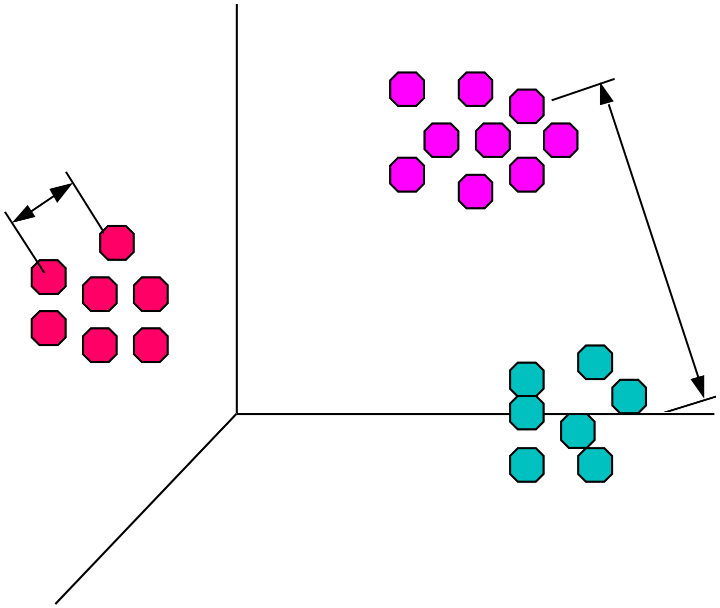
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Clustering

Group points such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar each other.

Ideal grouping is not known → Unsupervised Learning



Intracluster distances are minimized

Intercluster distances are maximized

Euclidean distance based clustering in 3-D space.

Clustering: Market Segmentation

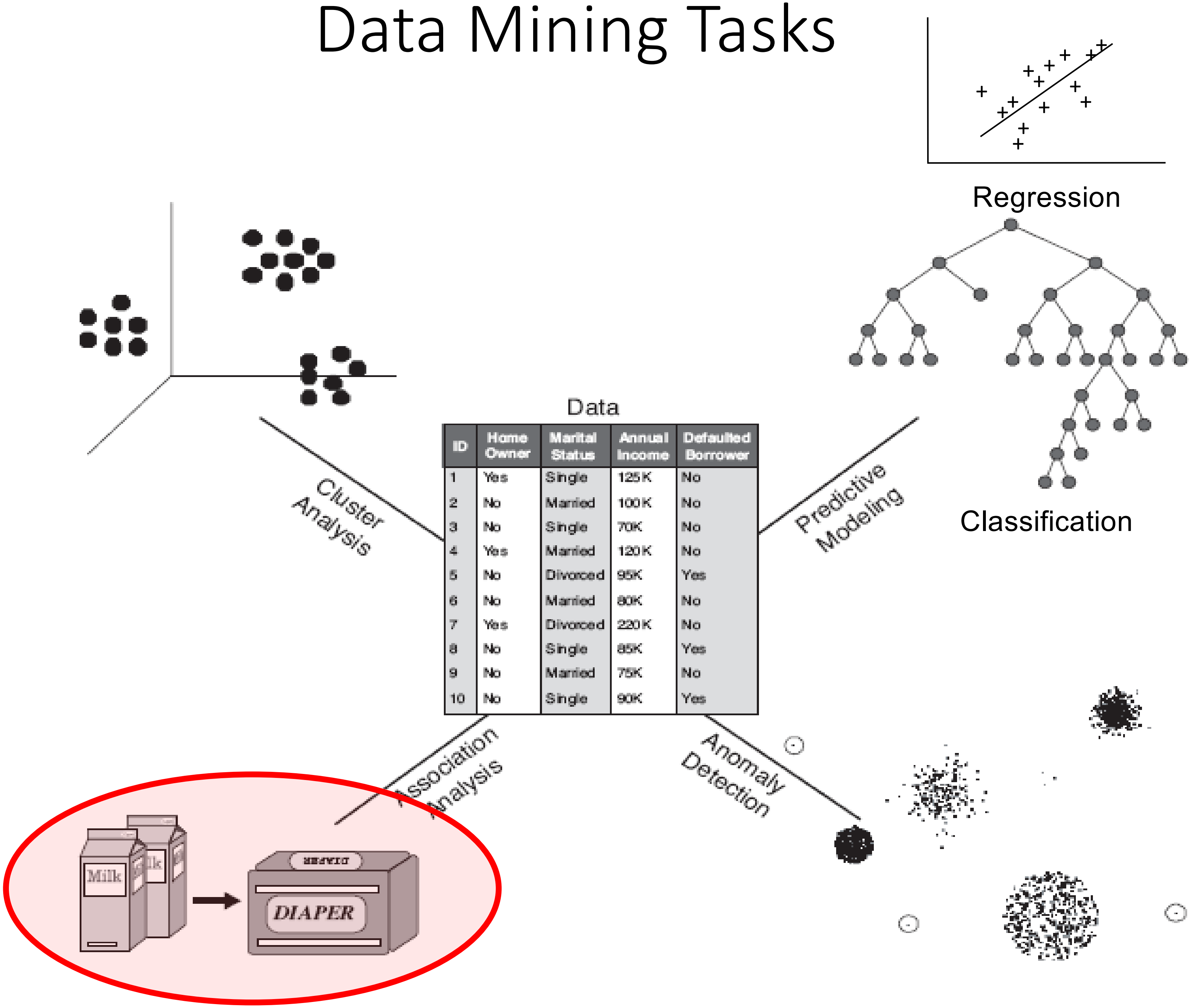
Goal: subdivide a market into distinct subsets of customers. Use a different marketing mix for each segment.

Approach:

1. Collect different attributes of customers based on their geographical and lifestyle related information and observed buying patterns.

2. Find clusters of similar customers.

Data Mining Tasks



Data

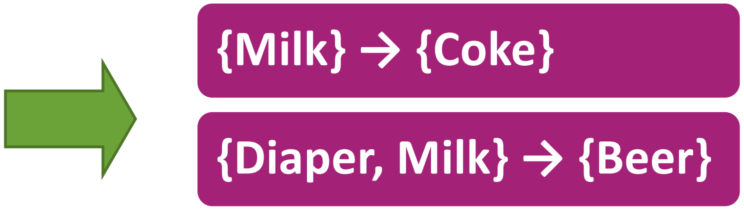
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Association Rule Discovery

- Data is a set of transactions. Each contains a number of items.
- Produce dependency rules of the form
  LHS → RHS
  which indicate that if the set of items in the LHS are in a transaction, then the transaction likely will also contain the RHS item.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction data

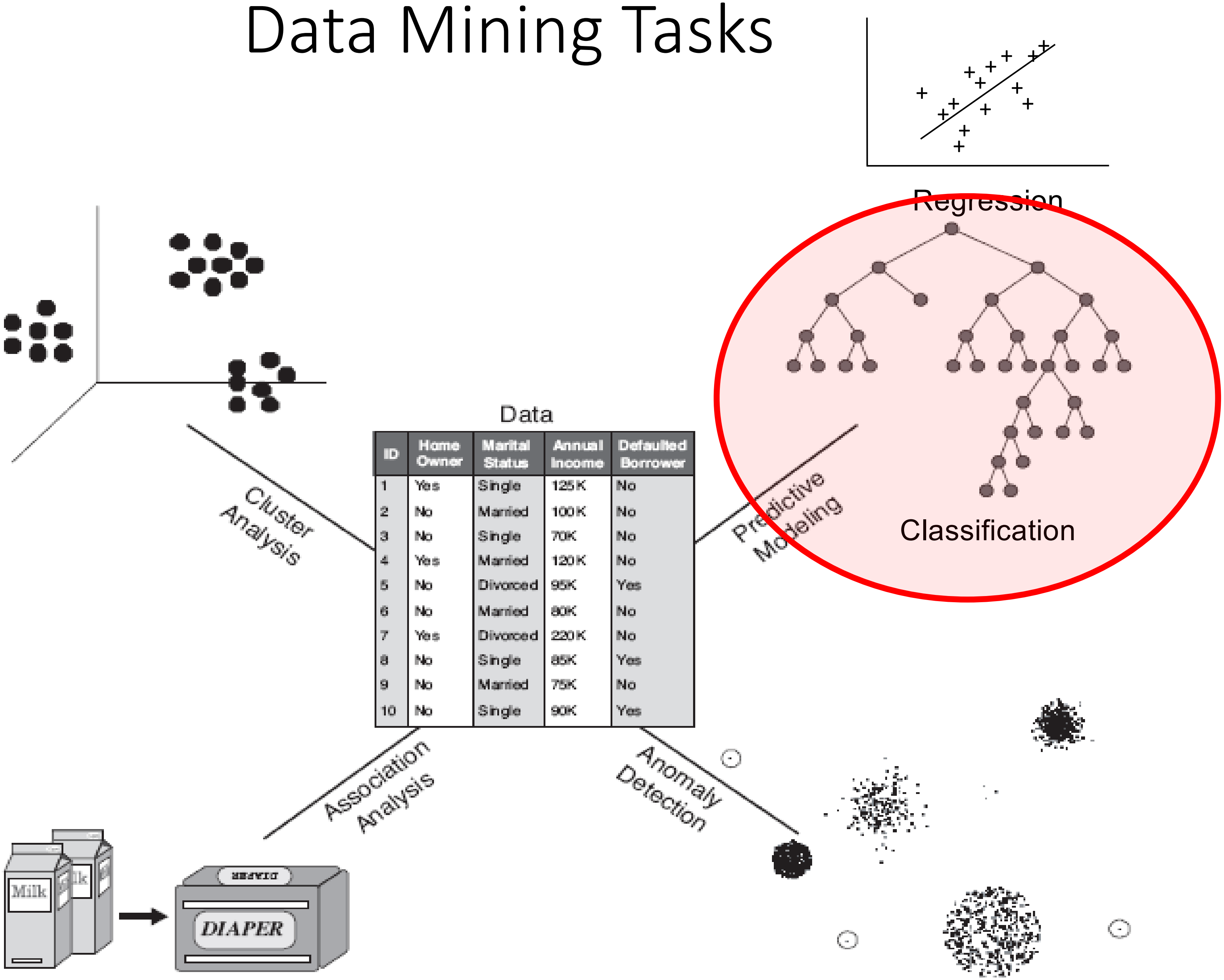


{Milk} → {Coke}

{Diaper, Milk} → {Beer}

Discovered Rules

Data Mining Tasks



Data

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	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Classification: Customer Attrition/Churn

Goal: To predict whether a customer is likely to be lost to a competitor.

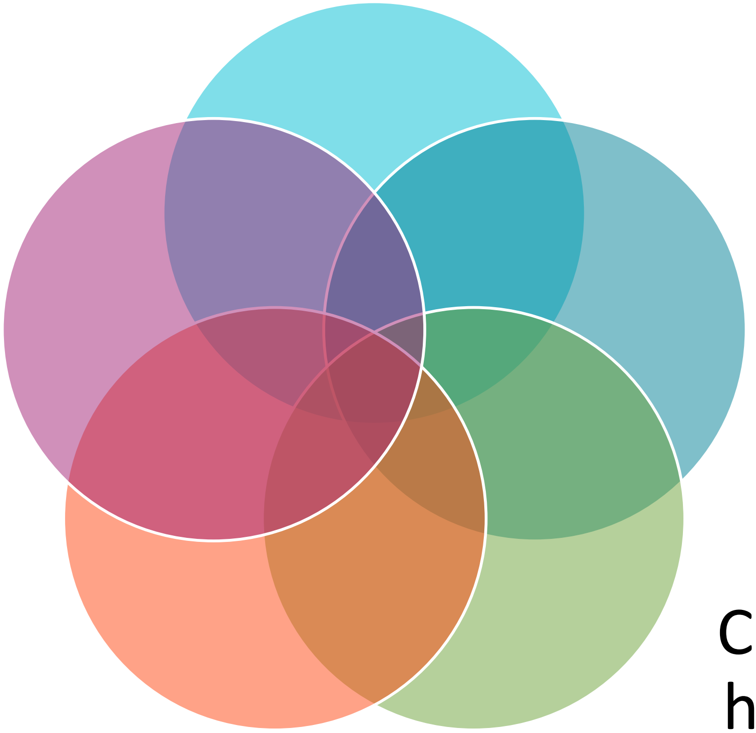
Approach:

- Use detailed record of transactions with each of the past and present customers, to find attributes (frequency, recency, complaints, demographics, etc.).
- Label the customers as loyal or disloyal.
- Find a model for disloyalty.
- Rank each customer on a loyal/disloyal scale (e.g., churn probability).

Other Data Mining Tasks

- Text mining – document clustering, topic models
- Graph mining – social networks
- Data stream mining/real time data mining
- Mining spatiotemporal data (e.g., moving objects)
- Visual data mining
- Distributed data mining

Challenges of Data Mining



Scalability

Data ownership and privacy

Dimensionality

Data quality

Complexity and heterogeneous data