Practical Introduction to Data Mining and Machine Learning

Duhai Alshukaili

Ibri College of Technology
Information Technology Department

Oct 23, 2018

الكلية التقنية بعبري
Ibri College of Technology

Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

Learning Outcomes

By the end of this tutorial you should be able to:

- Define what is meant by data mining.
- Define what is meant by machine learning.
- Understand the relation between data mining and machine learning.
- List different data mining tasks.
- Identify possible applications of data mining and machine learning.

Data Mining and **Machine Learning**

Warning!

- Basics only.
- No mathematics.

Logistics

- All relevant material are available in GitHub.
  `github.com/duhai-alshukaili/introduction-to-dm`
- Please provide your feedback form at the end of today's tutorial.
  - `http://bit.ly/2yugLM1`

Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

What is Data Mining?

Many definitions:

- *"The process if automatically discovering useful patterns in large data repositories"* [Tan et al., 2005]
- *"Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases"* [Ester and Sander, 2000]

Many names:

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archaeology, data dredging, and information harvesting [Han et al., 2011]..

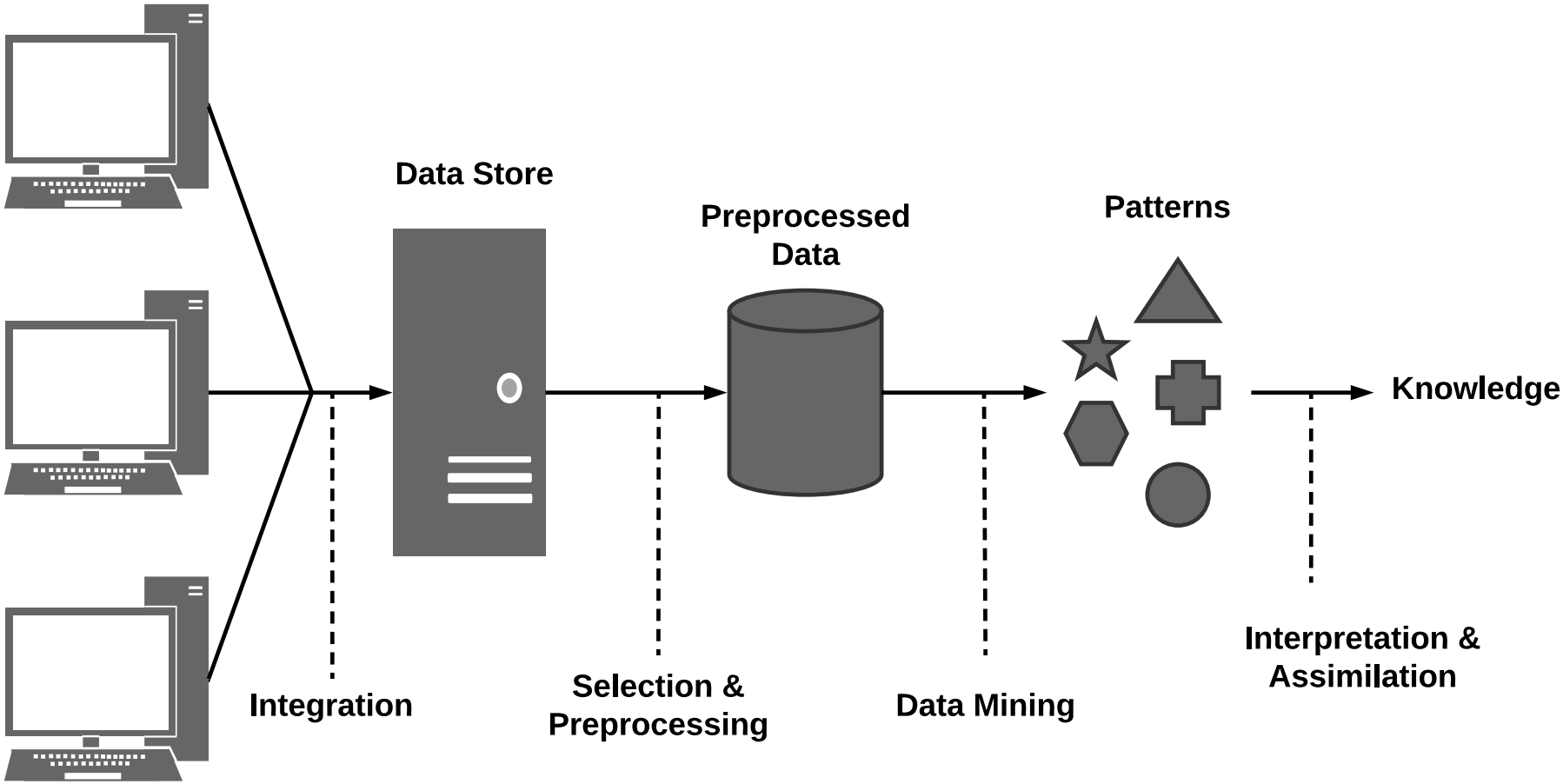
What is *Not* Data Mining?

Not every data task is a data mining task:

- Who are the students who cleared *AI* and *Compiler Construction* in the first try and have a GPA of 3.8 or more. (**querying**).
- What are the pages containing *"barack obama"* and was created after the year 2016. (**search**).
- Here are some facts about some students and a rule for exceptional students. Use the rule for listing exceptional students. (**deductive databases, e.g. Prolog**).

Knowledge Discovery in Databases (KDD)

Data mining is a core step in the KDD process.

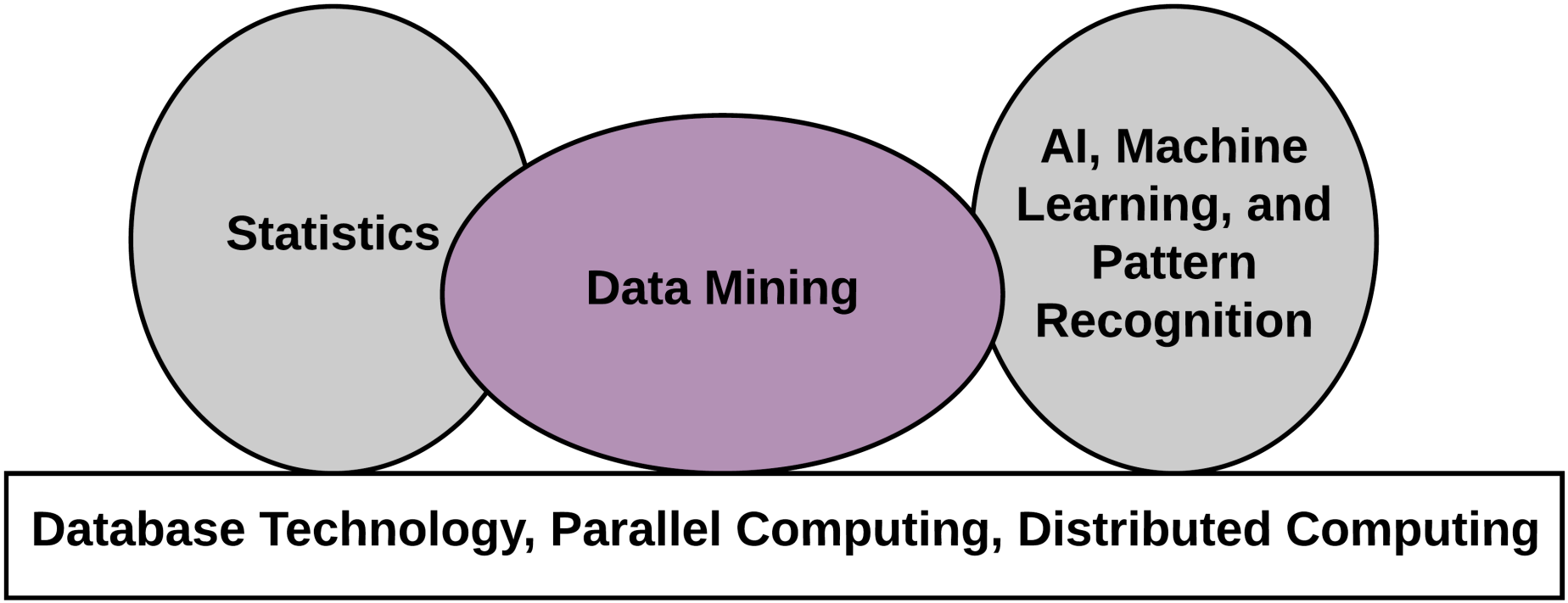


Knowledge Discovery in Databases (KDD)

- **Selection:** selecting a data set, or sources of data, on which discovery is to be performed.
- **Data Integration:** This is about providing *" ... the illusion that a single database is being accessed, when in fact data may be stored in a range of different locations."* [Hedeler et al., 2010]
- **Cleaning and Preprocessing:** removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and deciding on how to handle certain types of data such as spatial and temporal data.
- **Data Mining:** Applying (machine learning) algorithms to automatically build *models* of data.
- **Interpretation and Assimilation:** Explaining and evaluating the models and deciding on how to best deploy them in your environment.

Data Mining and Other Fields Adopted from [Tan et al., 2005]

The field of data mining borrows idea and techniques from other fields.



Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

Machine Learning

- A machine learning system is a magic box that can be used to
  - Automate some process, e.g., a process for driving a car.
  - Automate decision making, e.g., what is currently trending in twitter.
  - Extract knowledge from data.
  - Predict future event.
  - ..
- How to build such systems?

Write code to explicitly do the above tasks	Write code to make the computer **learn** how to do the above tasks

Machine Learning

- A machine learning system is a magic box that can be used to
  - Automate some process, e.g., a process for driving a car.
  - Automate decision making, e.g., what is currently trending in twitter.
  - Extract knowledge from data.
  - Predict future event.
  - ..
- How to build such systems?

Write code to explicitly do the above tasks

Write code to make the computer **learn** how to do the above tasks

Machine Learning

- A machine learning system is a magic box that can be used to
  - Automate some process, e.g., a process for driving a car.
  - Automate decision making, e.g., what is currently trending in twitter.
  - Extract knowledge from data.
  - Predict future event.
  - ..
- How to build such systems?

Write code to explicitly do the above tasks

Write code to make the computer **learn** how to do the above tasks

Machine Learning is Everywhere!

Web Search

Voice Assistants

Recommender Systems

Character Recognition

Spam Filtering

Self Driving Cars

What is Machine Learning?

Early definition of machine learning:

- “Field of study that gives computers the ability to learn without being explicitly programed.” [Samuel, 1959]
  - Arthur Samuel developed the first machine learning program that learned how to play checkers from previous moves.
  - He also developed the alpha-beta pruning which is used in tree search.

What is Machine Learning? (contd.)

A more recent definition:

- A well-posed learning problem: a computer is said to learn from the experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P,improves with experience E. [Mitchell, 1997]
  - The task is often what we want the system to do (e.g., predict some thing.)
  - The performance measure is a score on how well the system is doing the task.
  - The experience represented as data.
- *See **Lecture 1.1 What Machine Learning** (by Anderw Ng) for more on this definition.*

Exercise

Given [Mitchell, 1997] definition of machine learning, suppose your email client watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

A) Classifying emails as spam or not spam.

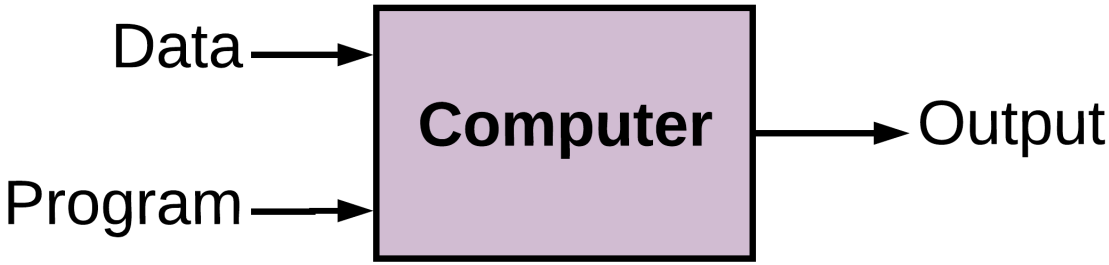
B) Watching you label as spam or not spam.

C) The number (or fraction) of emails correctly classified as spam/not spam.

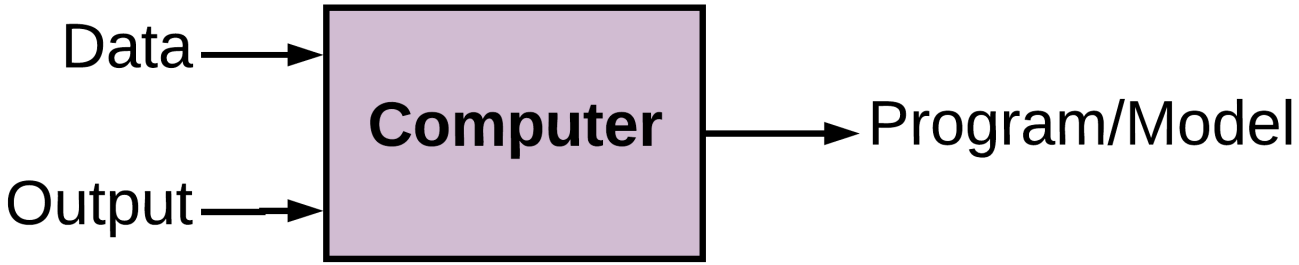
D) None of the above - this is not a machine learning problem.

Machine Learning vs. Traditional Programming

Traditional Programming

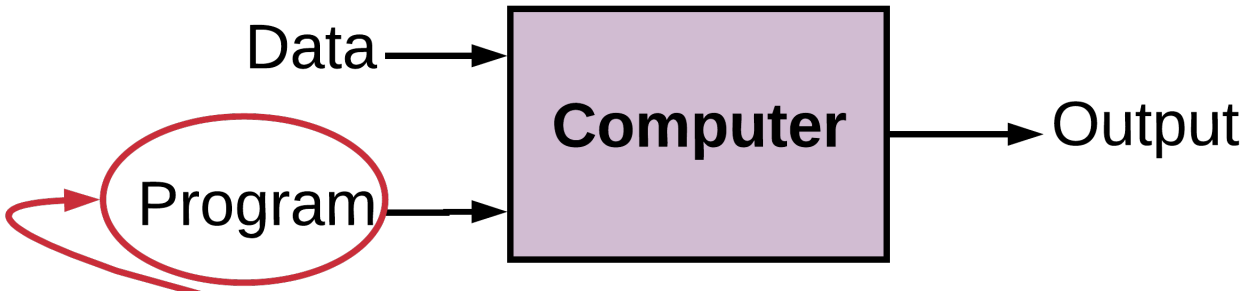


Machine Learning

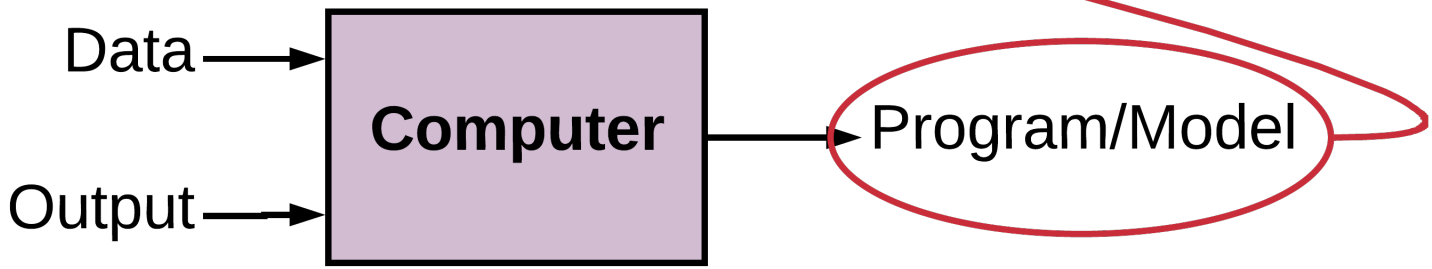


Machine Learning vs. Traditional Programming

Traditional Programming



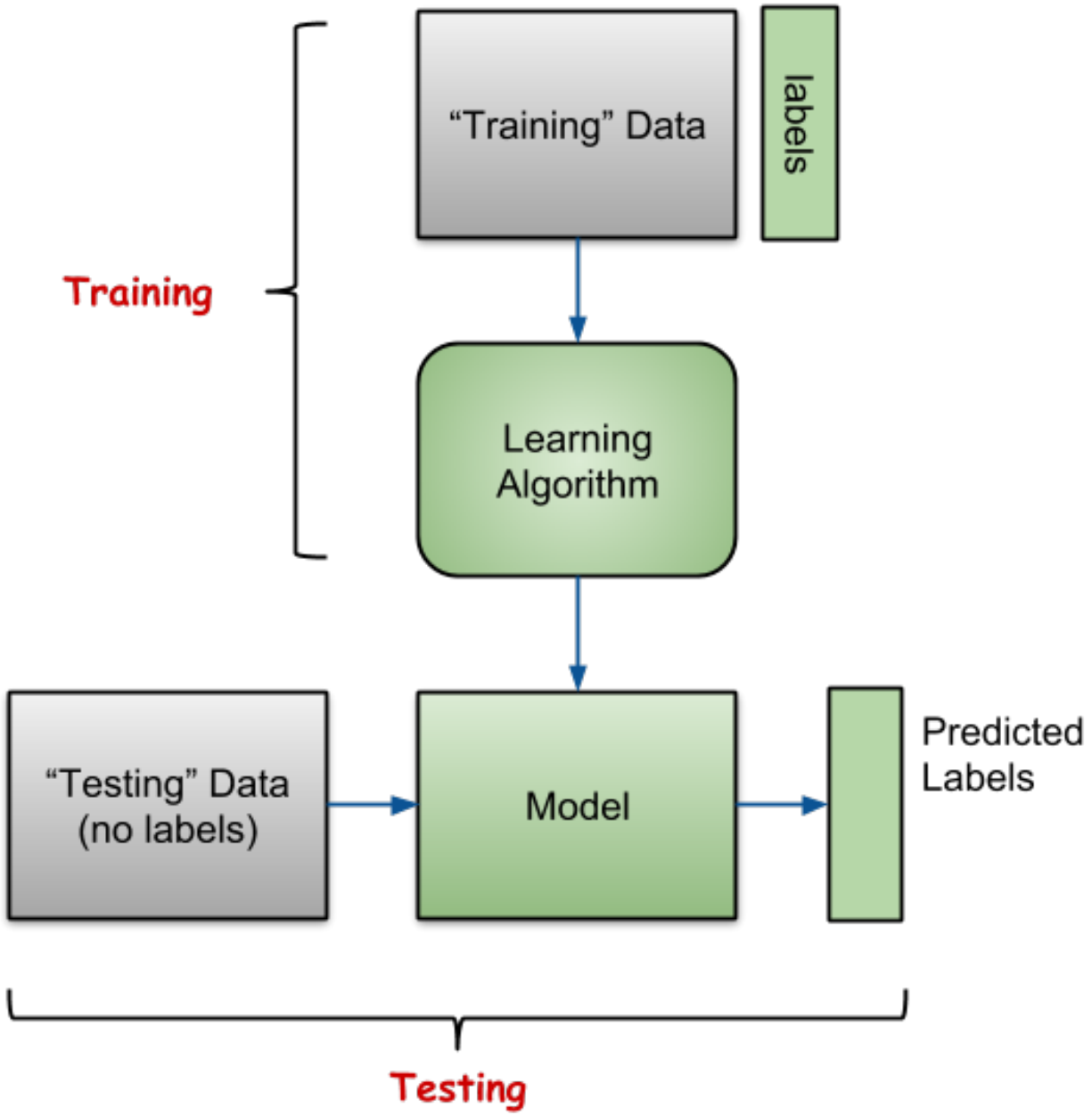
Machine Learning



Basic Paradigm

- Collect/observe a set of examples: training data.
- Infer (or learn) something about the process (or the model) that may have generated this data.
- Use the learned model to make some predictions about unseen data: test data.
- Variations on the paradigm:
  - Supervised learning: given a set of feature/label pairs, learn some rules that predicts the labels of unseen data.
  - Unsupervised learning: given a set of features vectors (without labels), group them into “natural clusters”

Basic Paradigm (contd.)



Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

Data

- All machine learning and data mining algorithms require one thing above all: data.
- Without data there is nothing to learn or extract patterns from.
- Luckily, the modern world is drowning of data:
  - Amazon process over $250,000$ book sales per day.
  - PC users uploaded over 300 billion videos in August 2014 alone, with an average of 202 videos and 952 minutes per viewer.
  - Approximately 2.5 million new scientific papers are published each year.
  - ...
- *Exactly, what do we mean by data?*

Data (contd.)

- Data for a data mining task is often recorded as table.
- Rows in such tables correspond to individuals, instances, objects, examples, or observations.
- Columns correspond to characteristics or attributes of such object. In data mining and machine learning, we call these features.
- A feature or a set features will be designated as a label for an object.
- It is often that the value of the label that we want to predict from future unseen data.

Data

- All machine learning and data mining algorithms require one thing above all: data.
- Without data there is nothing to learn or extract patterns from.
- Luckily, the modern world is drowning of data:
  - Amazon process over $250,000$ book sales per day.
  - PC users uploaded over 300 billion videos in August 2014 alone, with an average of 202 videos and 952 minutes per viewer.
  - Approximately 2.5 million new scientific papers are published each year.
  - ...
- *Exactly, what do we mean by data?*

Data: Example

The following table show a set of examples that could be used in a machine learning or data mining task for predicting the results of student taking some course:

$Quiz_1$	$Assignment$	Lab_1	$Midterm$	$Result$
10	16	15	16.5	$pass$
6	18.5	15	10.5	$pass$
7.5	18	14.25	10.25	$pass$
5	15	9.25	9.5	$fail$
8.5	15	7.25	11.25	$fail$

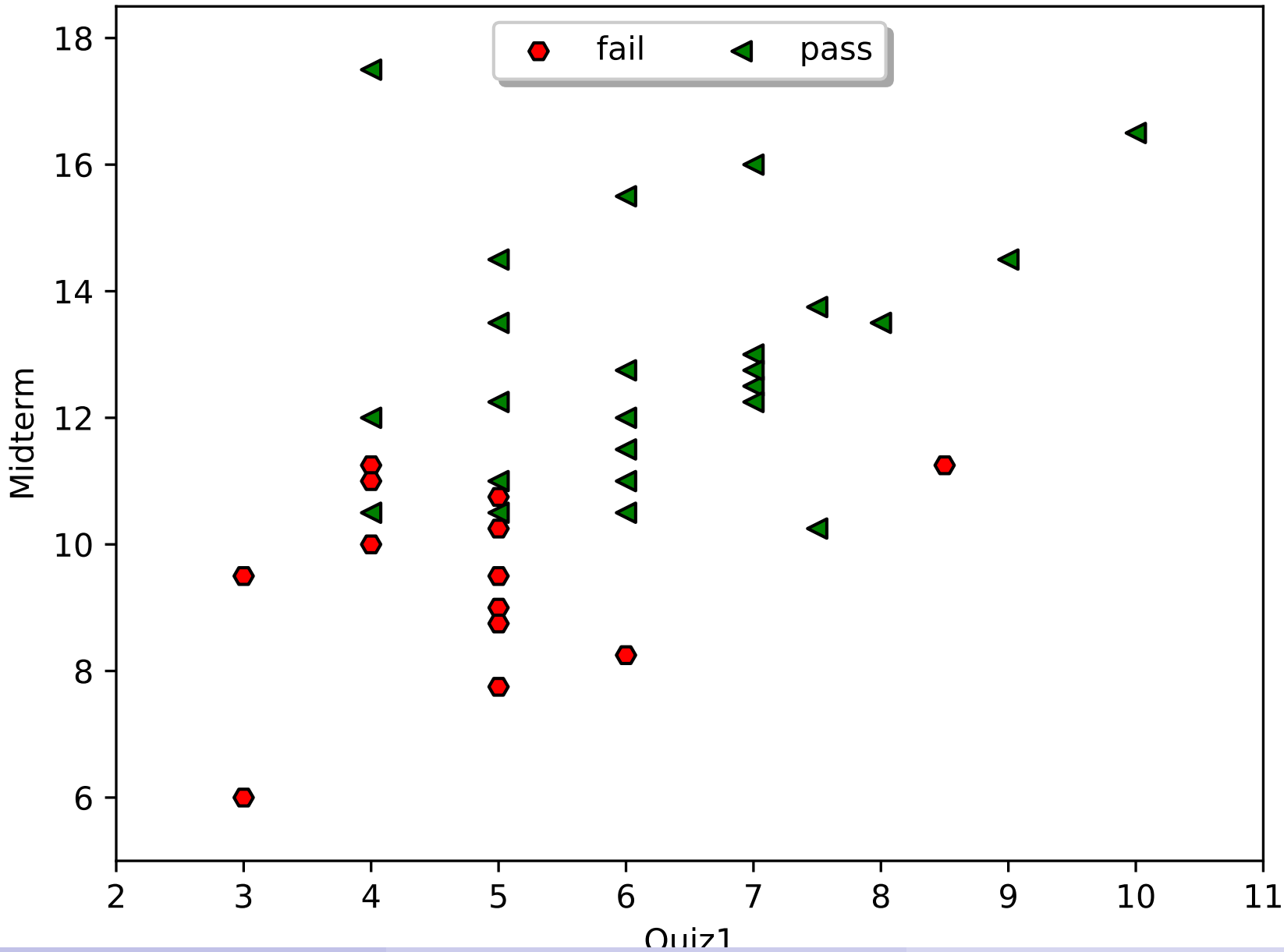
- How many examples?
- How many features (excluding the label)?
- What is the label?

Data: Example (contd.)

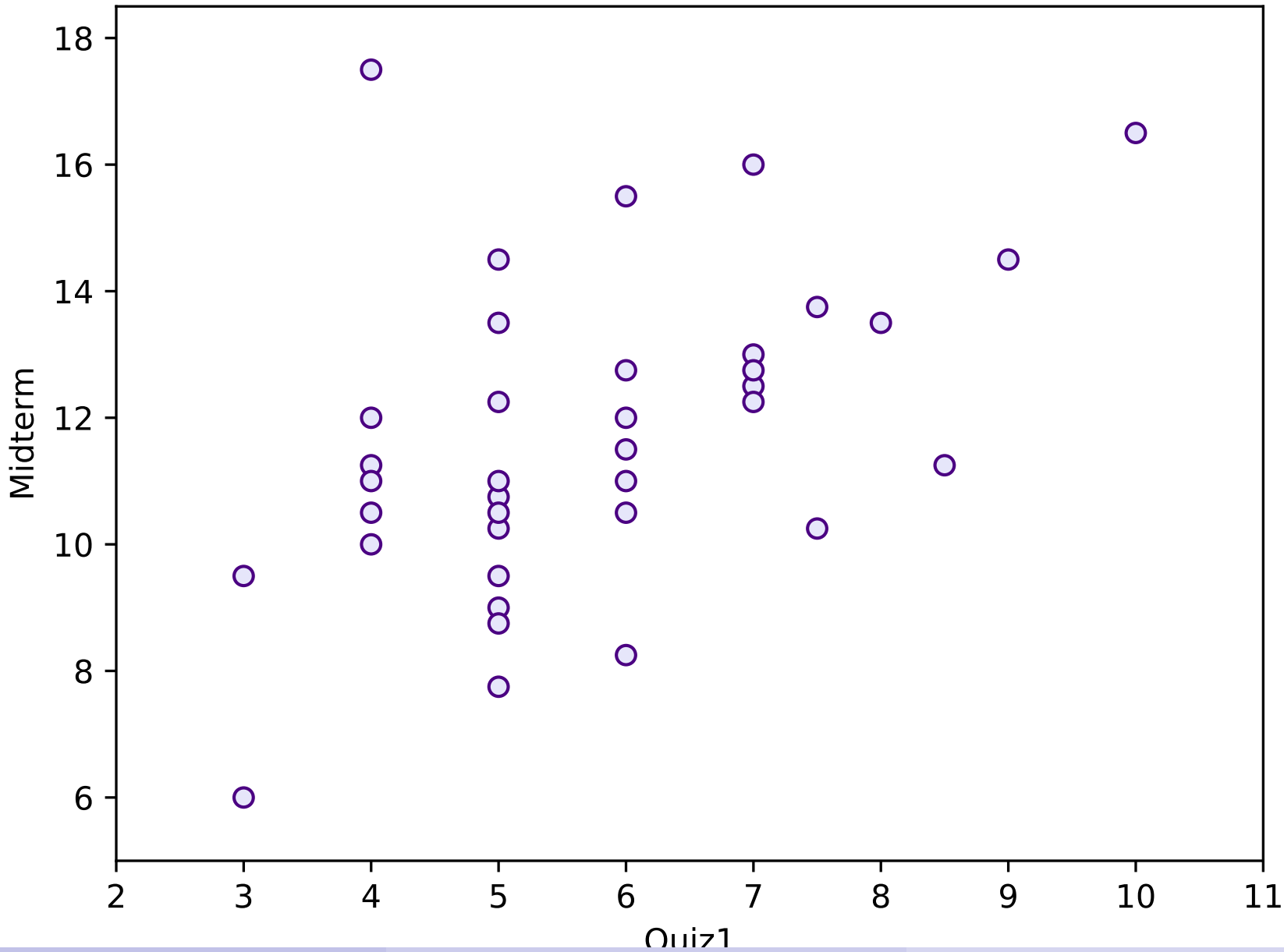
In machine learning column names are not important as the algorithm will ignore column name. so we give them shorthand names. However, In data mining, columns names might be important as it can help the data scientist in interpreting the results.

x_1	x_2	x_3	x_4	y
10	16	15	16.5	1
6	18.5	15	10.5	1
7.5	18	14.25	10.25	1
5	15	9.25	9.5	0
8.5	15	7.25	11.25	0

Labelled Data Visualization

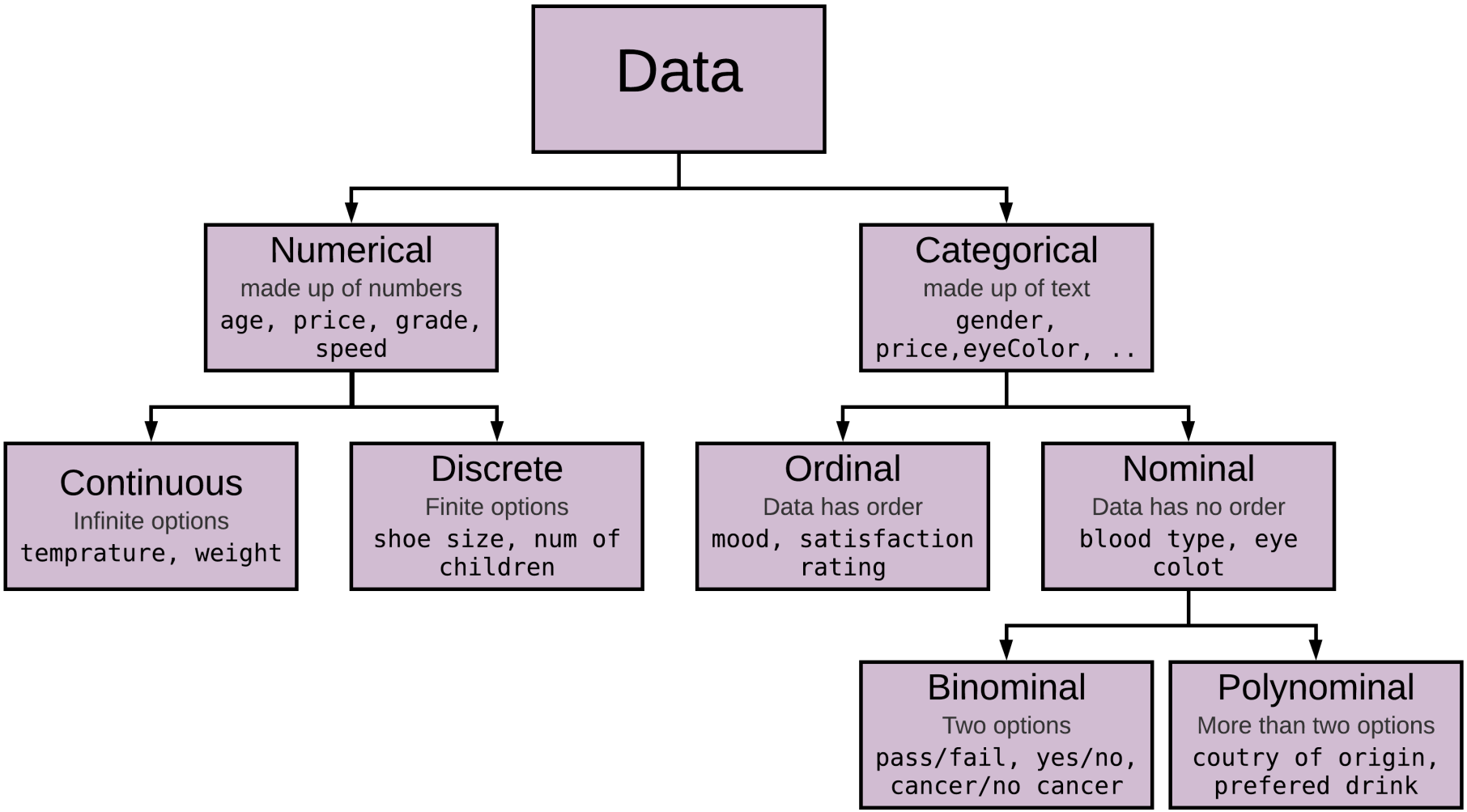


Unlabelled Data Visualization



Data Types

Features and labels can be one of different (basic) data types as shown in the following tree.



Exercise

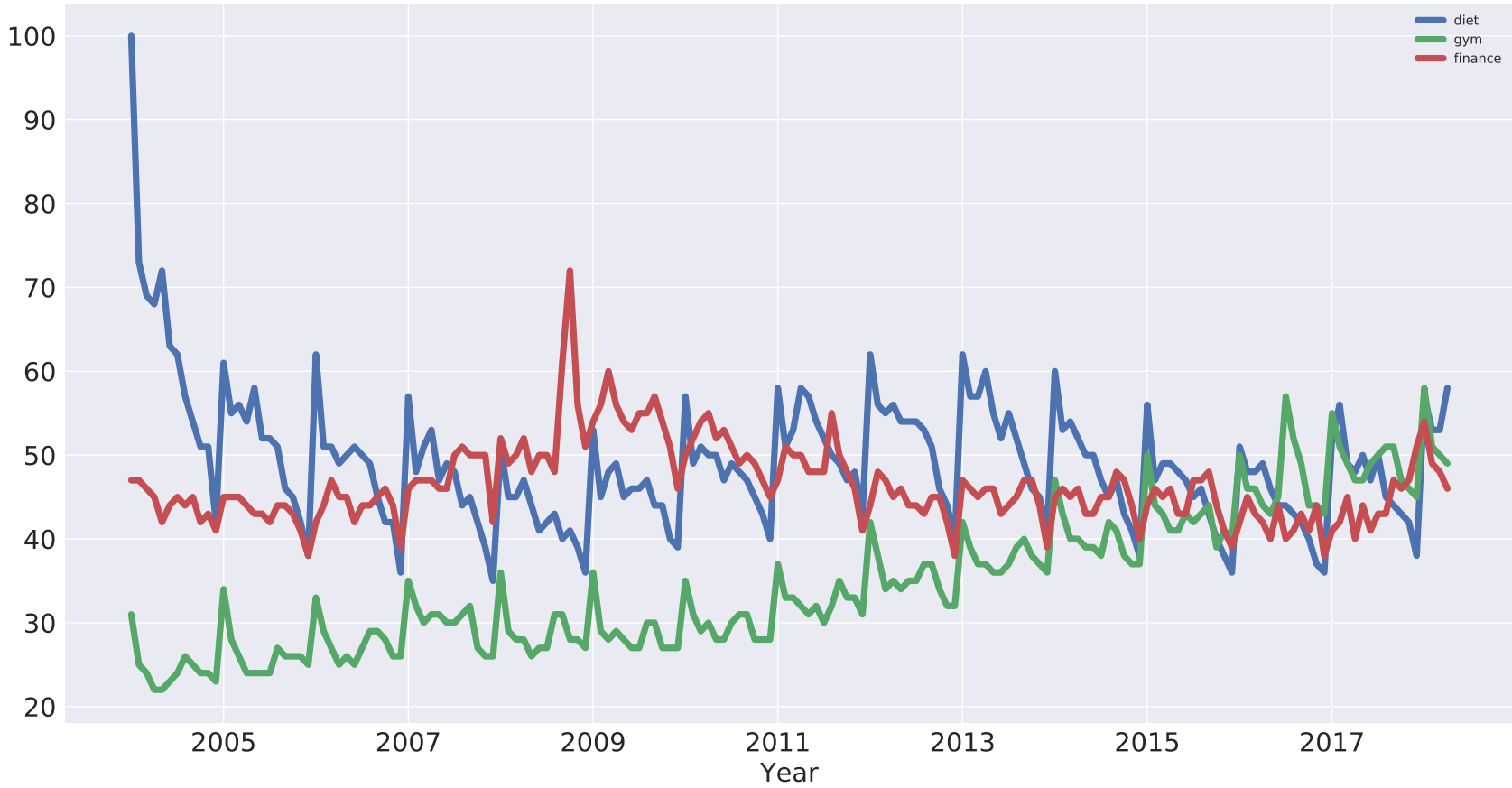
What is the most suitable type for the following attributes?

- Flight duration (in minutes).
- Quiz grade.
- Student letter grade (i.e., A, $A-$, B, etc.).
- Request status (i.e. $pending$, $rejected$, $accepted$, ...).
- Diabetes diagnosis (i.e., $diabetes$ or $no\ diabetes$).

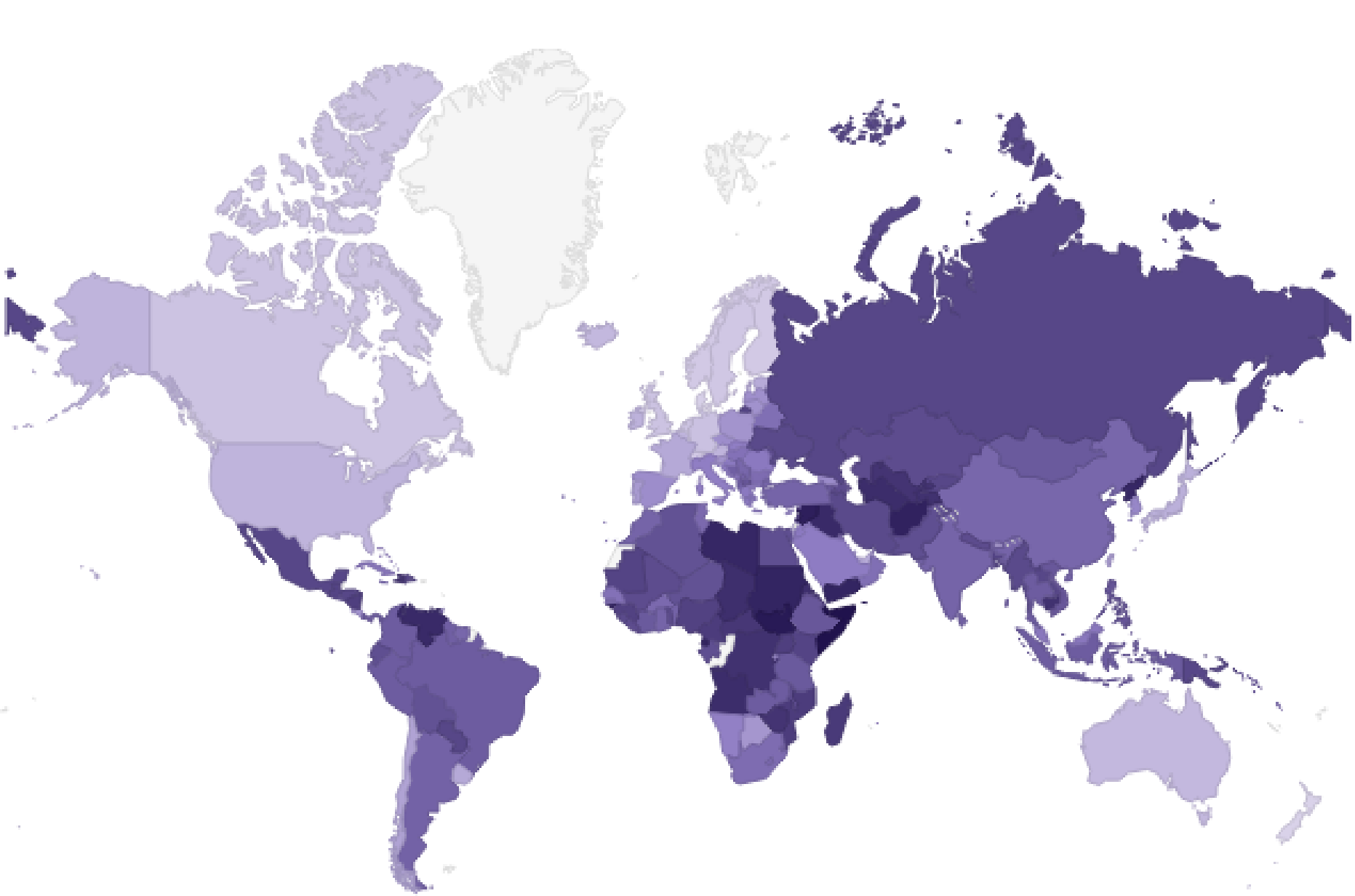
Other Types of Data

- Time-series data contain values that are typically generated by continuous measurements over time (i.e., temporal attributes).
- In spatial data many non-spatial attributes (e.g., temperature, pressure, image pixel colour intensity) are measured at spatial geographic locations.
- Spatio-temporal data combine both both spatial and temporal attributes.
- In network and graph data, the data values may correspond to nodes in the network, whereas the attributes and relationships among the data values may correspond to the edges in the network. Examples:
  - Web graph
  - Social

Time-Series Data



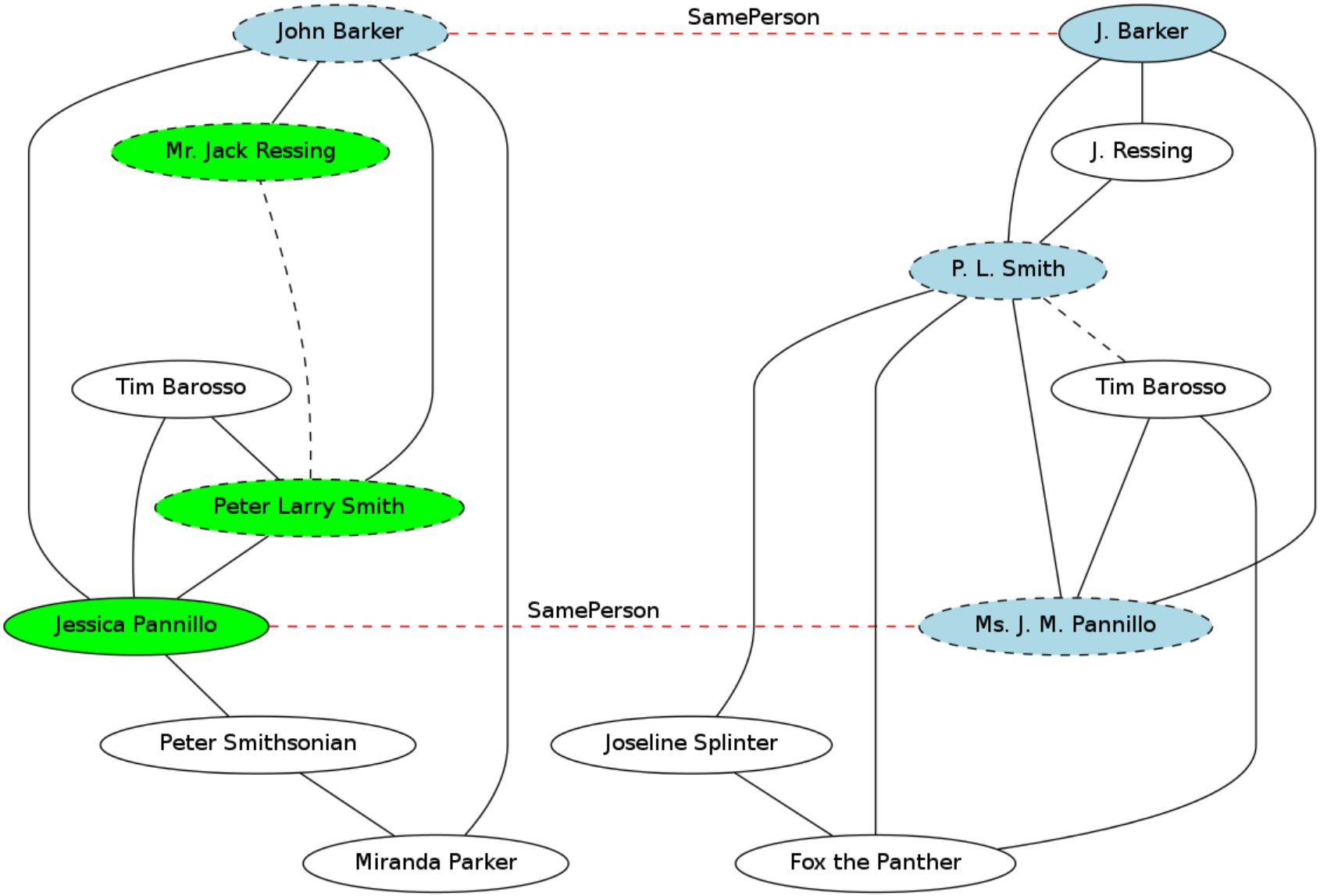
Spatial data



Spatio-temporal Data

See Gapminder tools website `http://bit.ly/2HFsUDl`

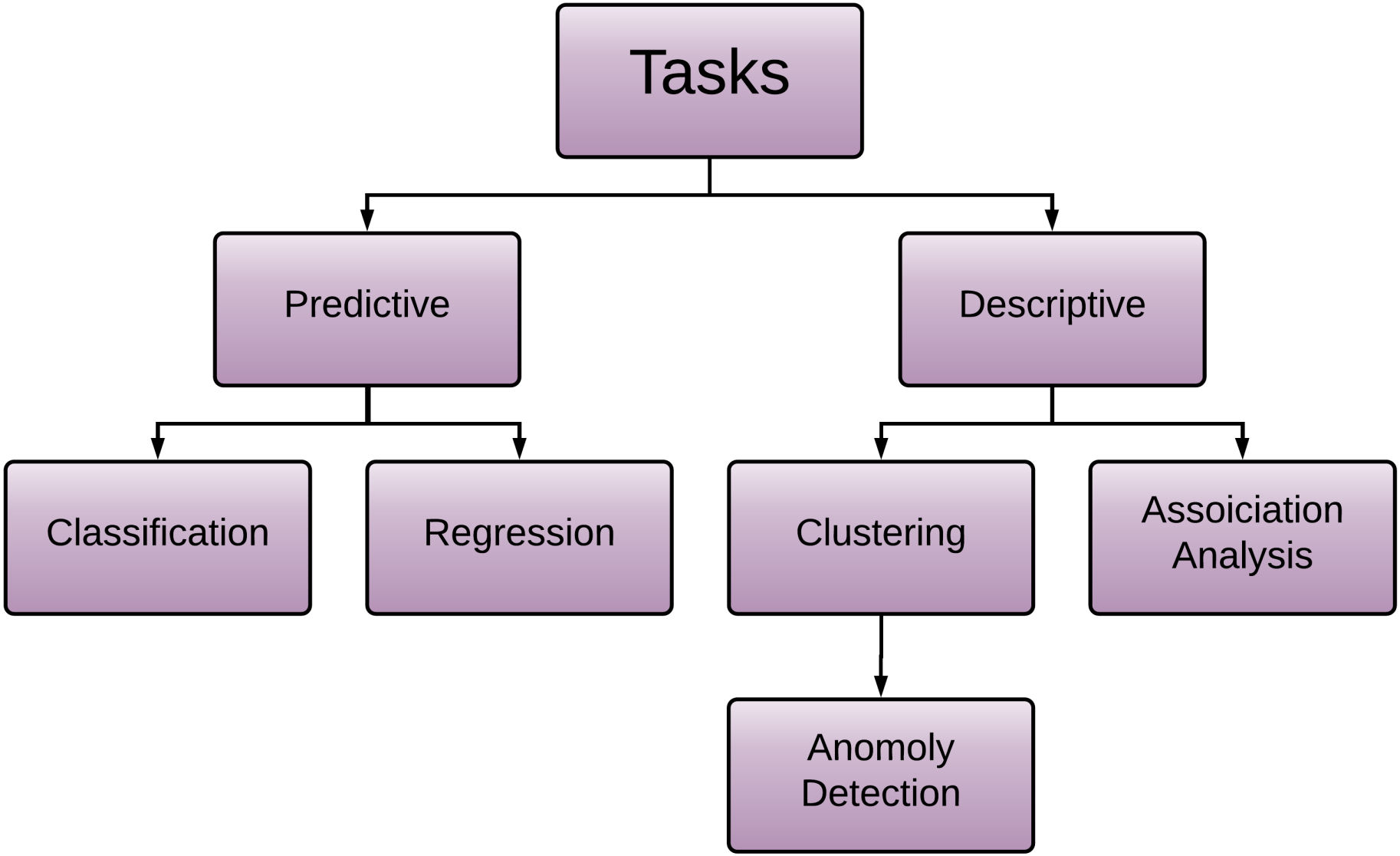
Graph Data


John Barker
SamePerson
J. Barker
Mr. Jack Ressing
J. Ressing
P. L. Smith
Tim Barosso
Tim Barosso
Peter Larry Smith
Jessica Pannillo
SamePerson
Ms. J. M. Pannillo
Peter Smithsonian
Joseline Splinter
Miranda Parker
Fox the Panther

Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

Data Mining Tasks



Predictive Tasks

- In machine leanring, predictive tasks are known as supervised learning tasks.
- In predictive tasks the learning algirhtm is given the target label for each example in the dataset. This guides the computer to build a predictive relationship between the objects and the target label. In other words, we know the ground truth, or true label for each example in the data.
- Typical predictive modelling tasks include classification and regression.
- In classification, the target label is categorical attribute.
- In regression, the target label is numeric attribute.

Classification Examples

In classification the label is a categorical attribute.

- Medical diagnosis: x=patient data, y=positive/negative of some pathology
- Optical character recognition: x=pixel values of images, y='A', 'B', 'C', ...
- Image analysis: x=image pixel features/values, y=scene/objects contained in image
- Weather: x=current & previous conditions per location, y=tomorrow's weather (i.e cloudy, sunny, windy, ..)

. . . . . this list can never end.

Regression Examples

In regression the target label is a numerical value.

- Finance: x=current market conditions and other possible side information, y=tomorrow's stock market price
- Social Media: x=videos the viewer is watching on YouTube, y=viewer's age.
- Robotics: x=control signals sent to motors, y=the 3D location of a robot arm end effector.
- Medical Health: x=a number of clinical measurements, y=the amount of prostate specific antigen (PSA) in the body.
- Environment: x=weather data, time, door sensors, etc., y=the temperature at any location inside a building.

. . . . . this list can never end, applications of classification are regression are vast and extremely active!

Supervised Learning

Classification Example

Supervised Learning

Regression Example

Algorithms for Predictive Tasks

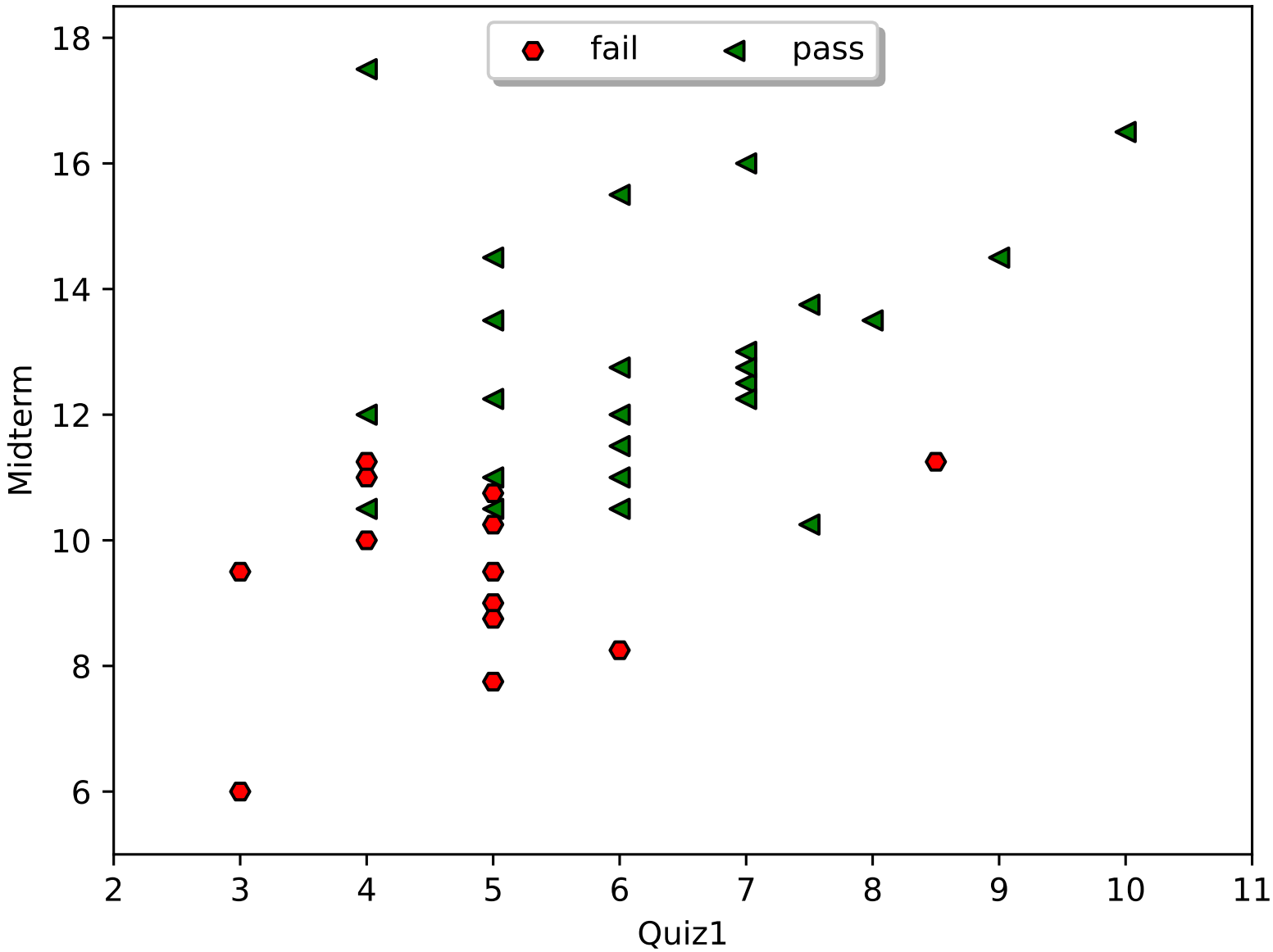
- Classification
  - k-NN
  - Naive Bayes
  - Logistic Regression
  - *Support Vector Machine (SVM)*
  - Linear Discriminant Analysis
  - ID3 (Tree Induction)
  - *Decision Tree*
  - $dots$
- Regression
  - Generalized Linear Model
  - Linear Regression
  - Polynomial Regression
  - Relevance Vector Machine
  - $dots$

Note that the names of the algorithms might vary from one tool to another, so you need to be aware of the details of the tool you use.

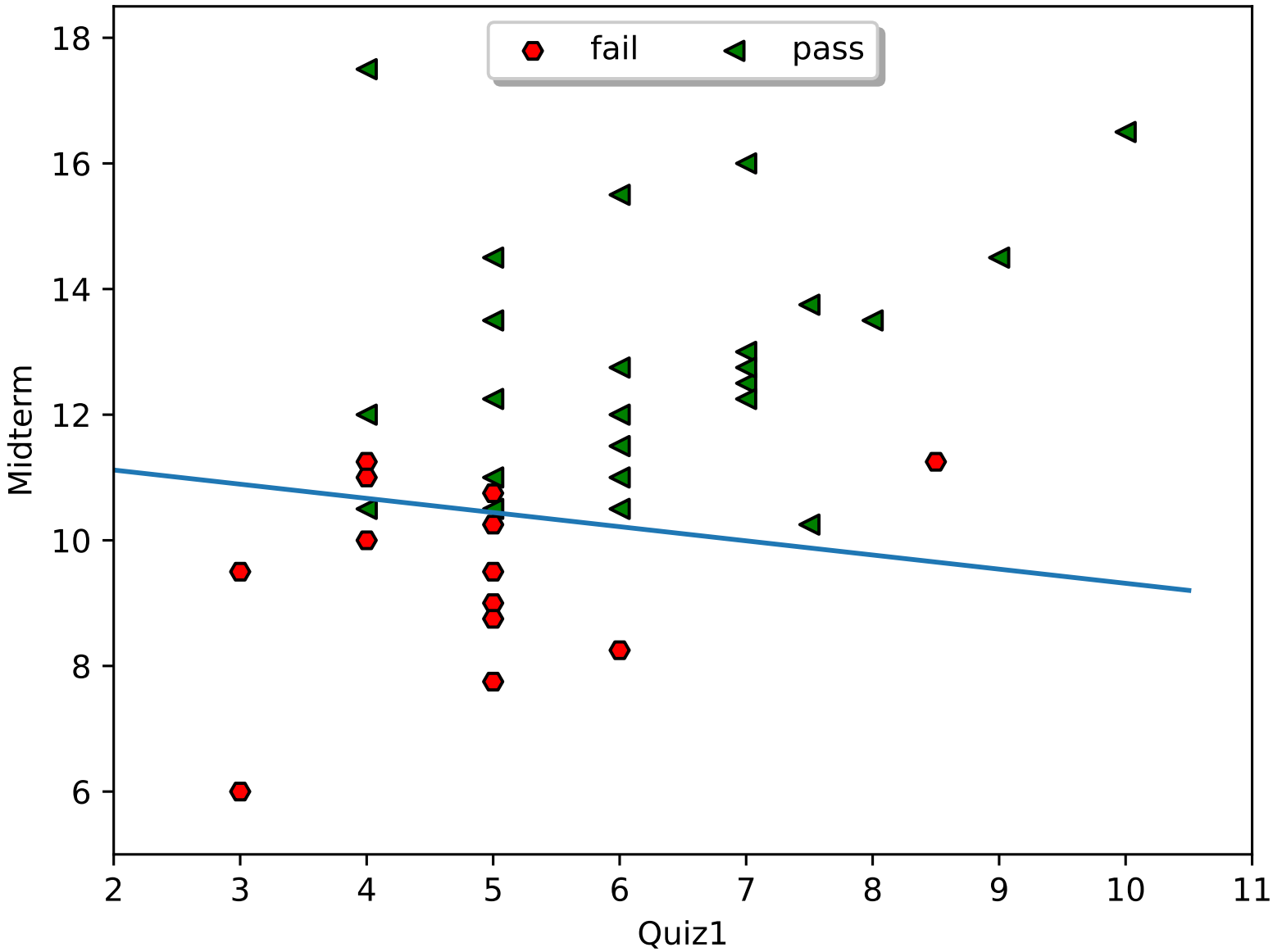
Multidimensional Data

- In reality, data will have more than one dimension. The following is an example where we use two dimensional feature vectors (i.e, *Quiz1* and *Midterm*).
- Real machine learning applications tend to be highly dimensional (order of 1000's of features).

Example: Multidimensional Data



Example: Multidimensional Data (cont.)



Exercise

Your are running a company, and you want to develop data mining systems to address each of the following two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You would like to examine individual customer accounts, and for each account you want to decide if it has been hacked/compromised.

Should you treat these as classification of regressions problems?

A) Treat both as classification problems.
B) Treat problem 1 as classification problem, problem 2 as regression.
C) Treat problem 1 as regression problem, problem 2 as classification.
D) Treat both as regression problem.

Descriptive Tasks

- Frequently, we run into situations where we only know the features, i.e., we don't know the true labels in the data.
- Descriptive tasks allows us to derive a structure from data where we don't necessarily know the effect of the variables on the different categories of the data.
- Descriptive tasks are also called unsupervised learning tasks.
- For example, we may have records of students and their interests/hobbies, and we want to group students into group based on how similar their hobbies.
- Clustering is a typical descriptive data mining task. Other tasks include association mining and anomaly detection.
- The goal of clustering is to identify those latent (hidden) groups of observations, which if done well, allows us to predict the class of new observations.

Clustering

- The goal of clustering is to identify those latent (hidden) groups of observations, which if done well, allows us to predict the class of new observations.
- Real world applications of clustering:
  - Grouping individuals based on DNA similarity.
  - Market segmentations.
  - Computer clusters analysis.
  - Identifying galaxies in space.
  - . . .

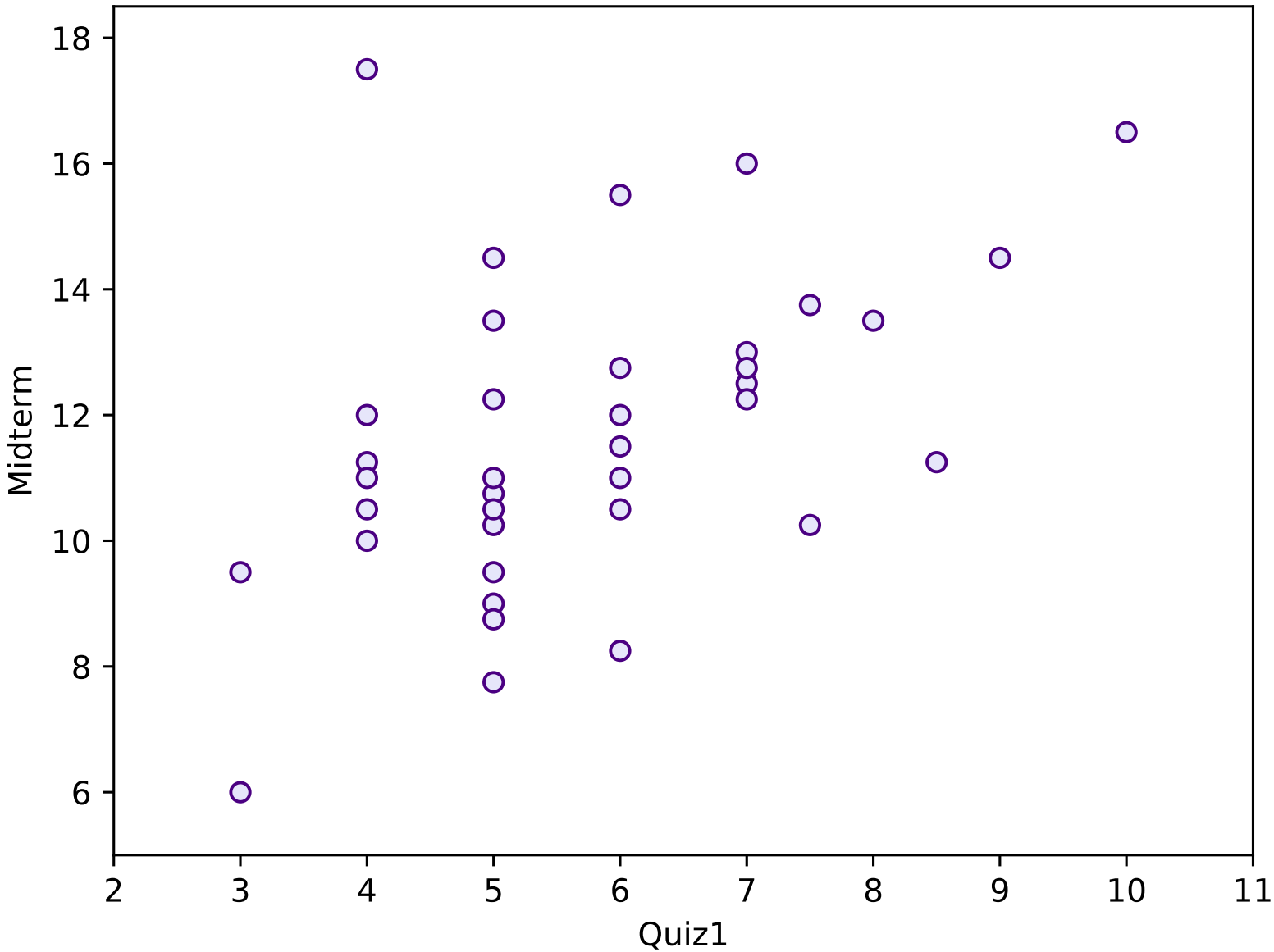
See the unsupervised learning video by Andrew Ng. for clustering examples. (Available on Moodle!) http://bit.ly/2pDFokD.

Clustering Algorithms

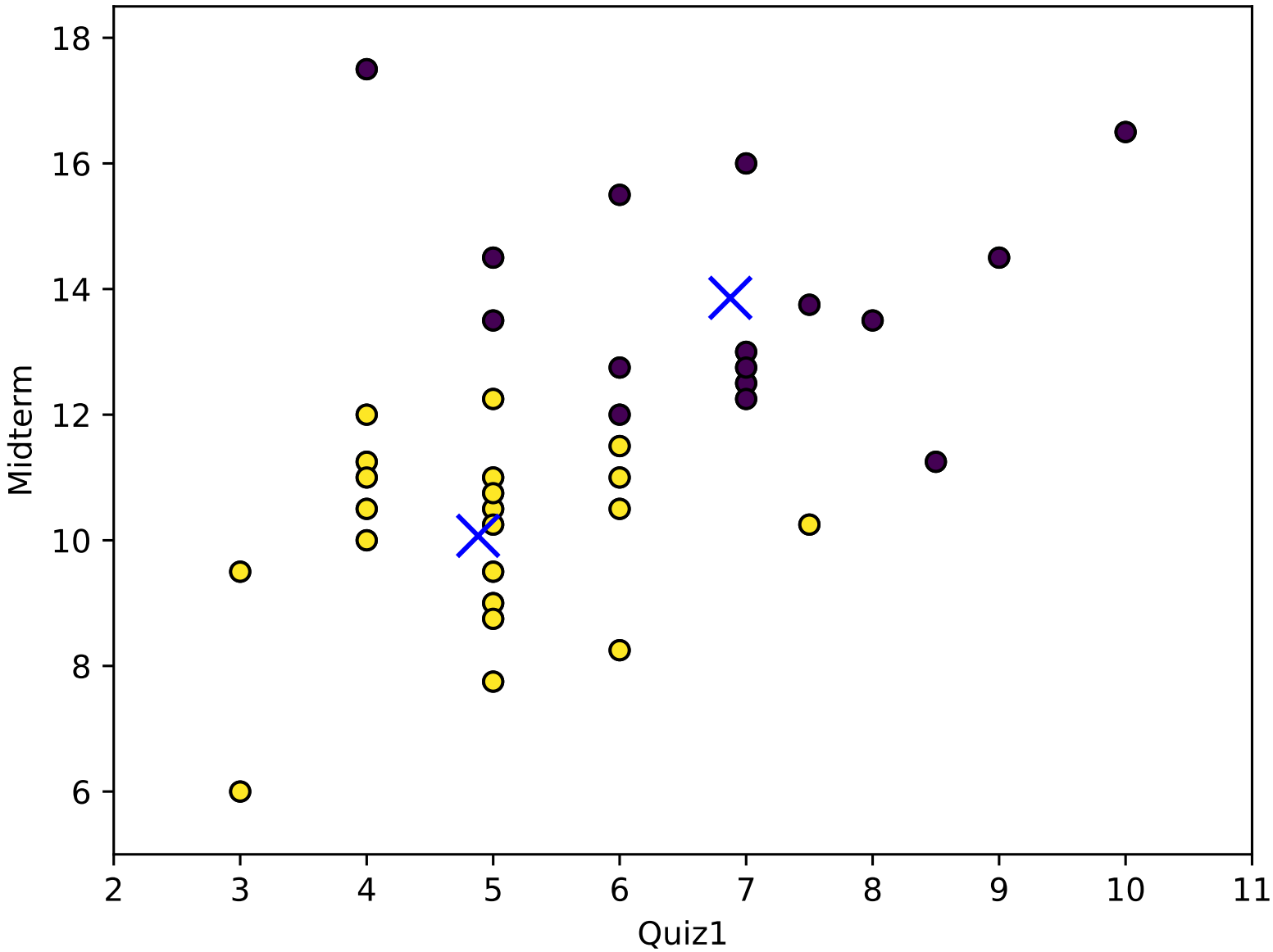
- K-Means
- DBSCAN
- Agglomerative Clustering
- Affinity Propagation
- Spectral clustering
- Mean Shift
- . . .

Note in RapidMiner clustering algorithms are referred to as segmentation algorithms.

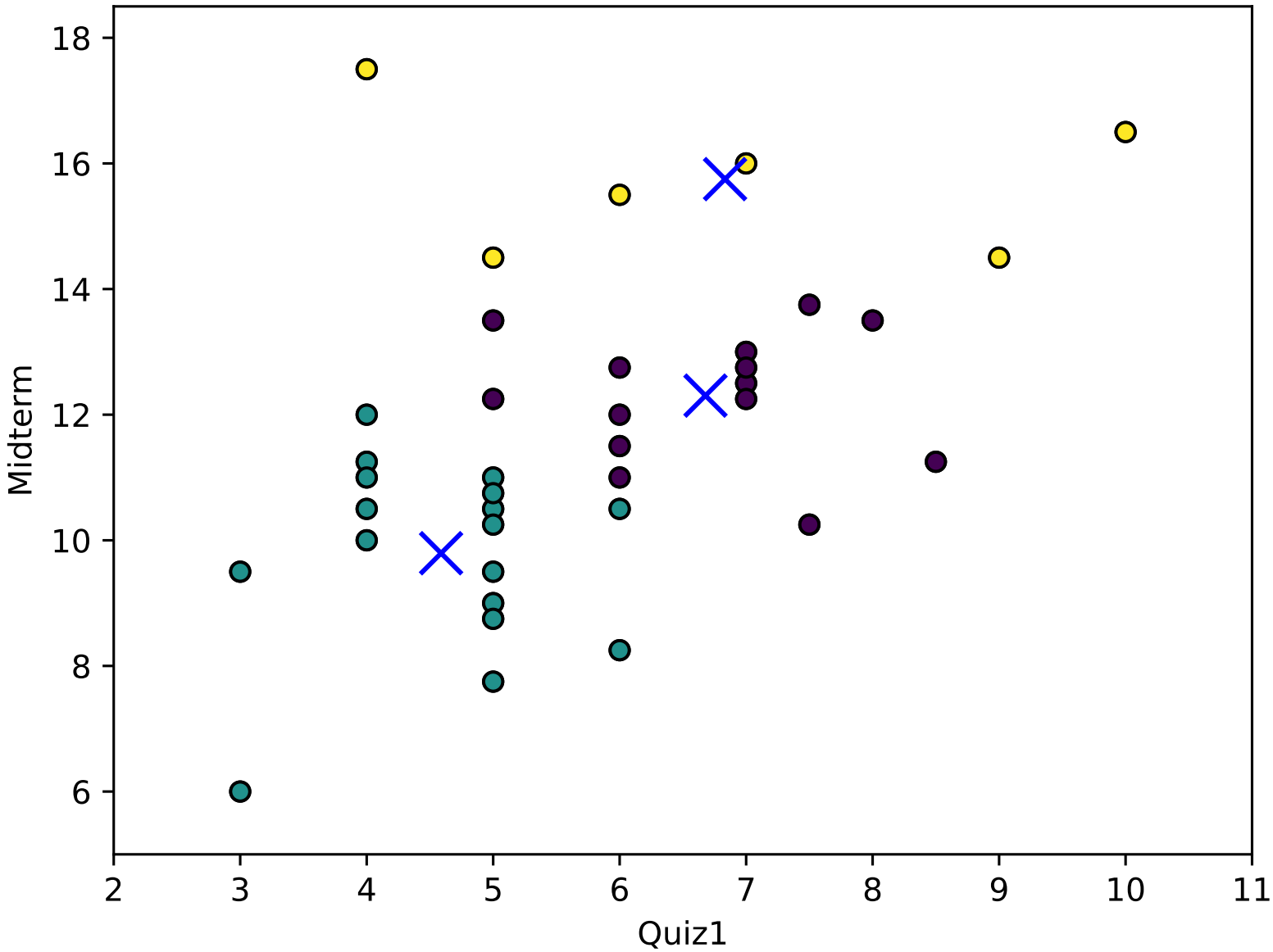
Unlabelled Data



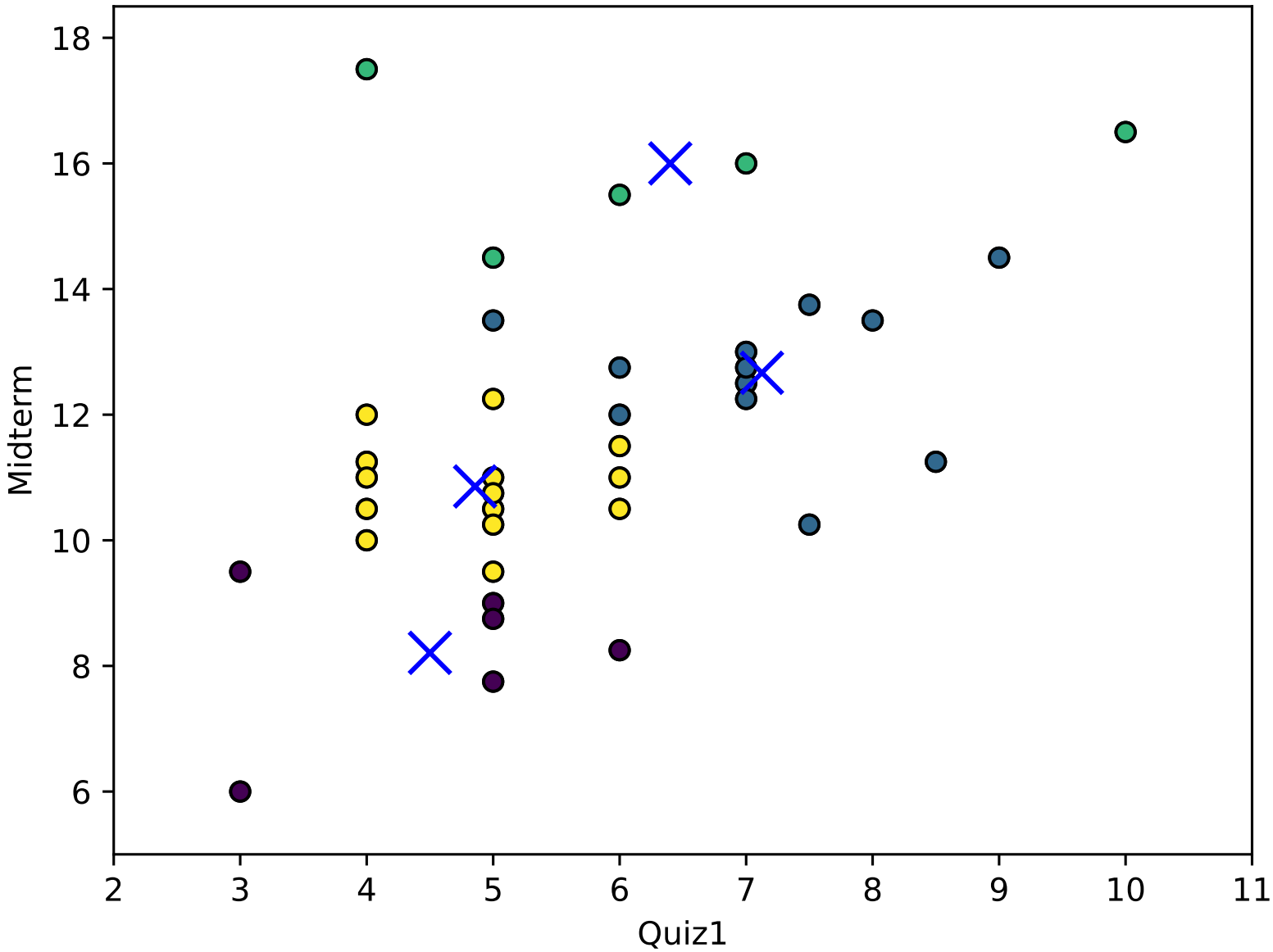
Clustered Data ($k = 2$)



Clustered Data ($k = 3$)



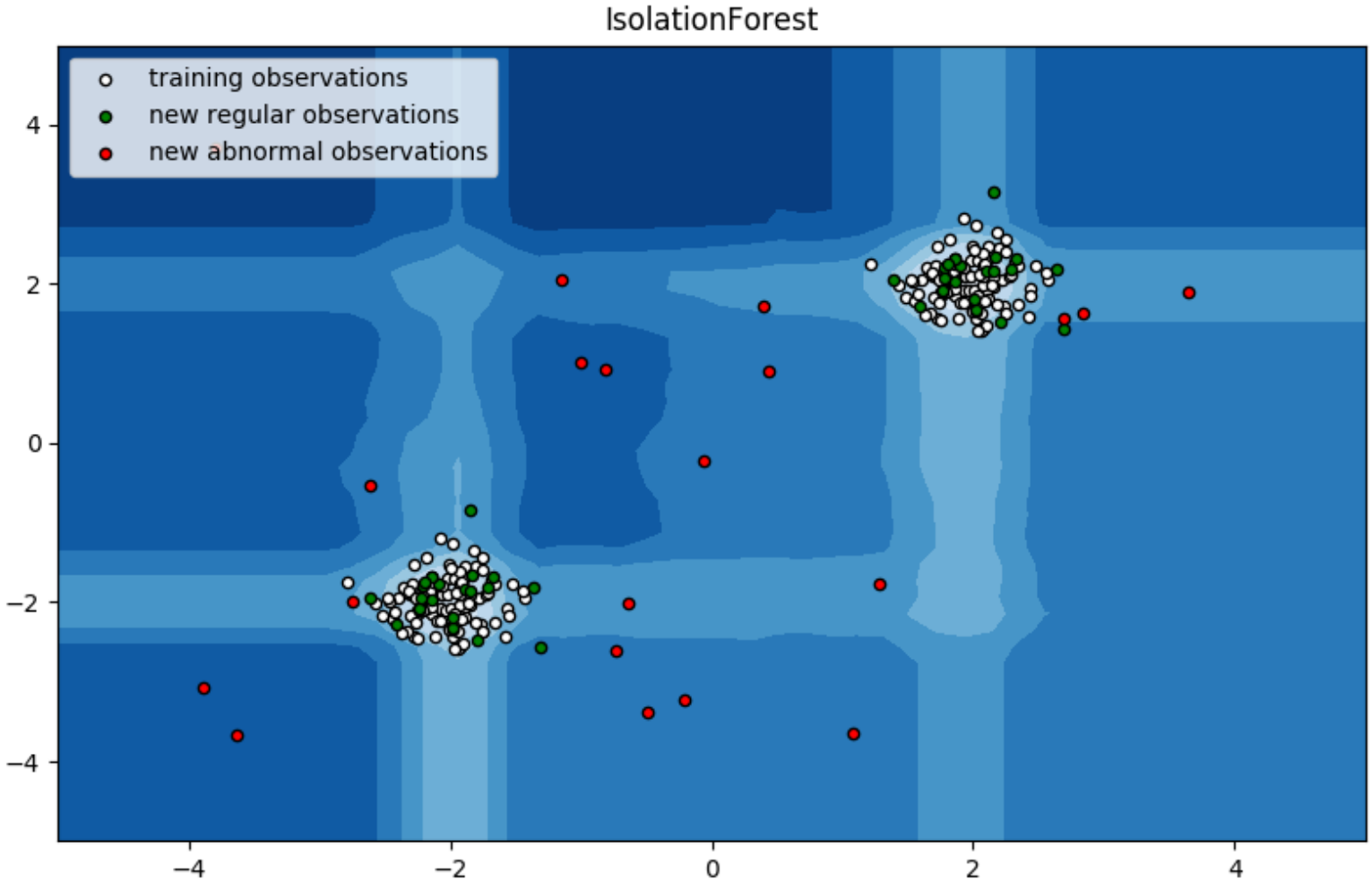
Clustered Data ($k = 4$)



Anomaly Detection

- Anomaly detection is the task of identifying observations whose features are significantly different from the rest of data.
- Also known as outlier detection.
- Example applications:
  - Credit card fraud detection.
  - Defective sensors in a Wireless Sensor Network (WSN).
  - Students copying assignment solutions form others.
  - Intrusion detection (strange patterns in a network system).
  - . . .

Anomaly Detection: Illustration



Anomaly Detection Algorithms

- One Class SVM.
- Covariance Estimators [Schölkopf et al., 2001].
- Isolation Forest [Liu et al., 2012].
- Clustering based algorithms.
- . . .

Association Rule Analysis

- Association rule analysis is an unsupervised task that allows for identifying associations between different attributes in dataset.
- The following is an example of rules generated from this task:

```
IF the customer age is 18 AND
the customer buys a paper
THEN the customer buys a binder
```

- Example applications:
  - Market basket analysis.
  - Retail transactions.
  - Word associations/co-occurrence in textual data.
  - Subjects frequently taken together by student in a university.
  - ...

Market Basket Analysis Example

Adopted from [Tan et al., 2005]

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Do you notice any association between items?

Market Basket Analysis Example

Adopted from [Tan et al., 2005]

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Market Basket Analysis Example

Adopted from [Tan et al., 2005]

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

How to put this data in tabular form?

Association Rule Algorithms

Commonly used algorithms include, but not limited to

- Apriori algorithm.
- Frequent Pattern growth (FP-growth) algorithm.
- Equivalence Class Transformation (ECLAT) algorithm.

Outline

1. Learning Outcomes and Logistics
2. What is Data Mining?
3. What is Machine Learning?
4. Data
5. Data Mining Tasks
6. Summary

Summary

- There is a strong link between data mining and machine learning.
- Clustering and classification are common tasks in data mining.
- The key to any data mining tasks is data.

Arigatō gozaimasu

Acknowledgements

The material presented in the tutorial mixes some original material by the author as well as material adopted from

- [Brown, 2017]
- [Grimson, 2016]
- [Tan et al., 2005]

The author gratefully acknowledges the work of the authors cited while assuming complete responsibility for any mistake introduced in the adaptation.

References


Brown, G. (2017).
The COMP61011 Not-Very-Scary Guide to Machine Learning.


Ester, M. and Sander, J. (2000).
Knowledge Discovery in Databases - Techniken und Anwendungen.
Springer.


Grimson, E. (2016).
Introduction to machine learning.
Accessed 22-Mar-2018.


Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques, 3rd edition.
Morgan Kaufmann.


Hedeler, C., Belhajjame, K., Paton, N. W., Campi, A., Fernandes, A. A. A., and Embury, S. M. (2010).
Dataspaces.
In *Search Computing*, pages 114–134. Springer.


Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2012).
Isolation-based anomaly detection.
ACM Trans. Knowl. Discov. Data, 6(1):3:1–3:39.

References (contd.)


Mitchell, T. M. (1997).
Machine learning.
McGraw Hill series in computer science. McGraw-Hill.


Samuel, A. L. (1959).
Some studies in machine learning using the game of checkers.
IBM Journal of research and development, 3(3):210–229.


Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. (2001).
Estimating the support of a high-dimensional distribution.
Neural Comput., 13(7):1443–1471.


Tan, P., Steinbach, M., and Kumar, V. (2005).
Introduction to Data Mining.
Addison-Wesley.