Mathematical techniques in data science

Lecture 1: General information and Introductions

General information

- Classes:

  MWF 12:20pm-1:10pm, Purnell Hall 236
- Office hours:
  - MW 3pm-4pm, Ewing Hall 312 (starting from the 2^{nd} week)
  - By appointments
- Instructor: Vu Dinh
- Website:

  `http://vucdinh.github.io/m637s22`

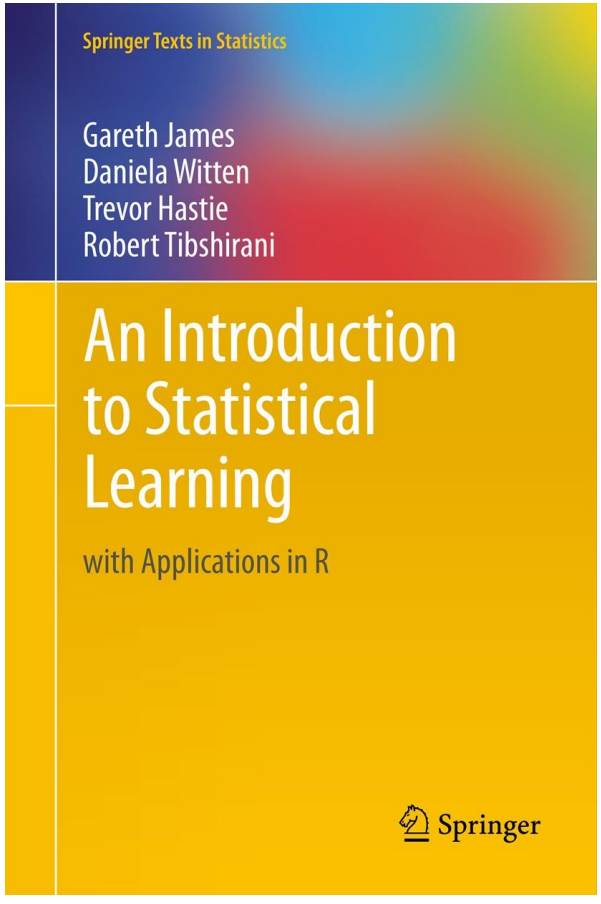
Data science

- is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data, both structured and unstructured
- is a concept to unify statistics, AI, data analytics, machine learning and their related methods in order to understand and analyze data
- employs techniques and theories drawn from many fields: mathematics, statistics, information science, and computer science

Goals of the course

- Become familiar with the basic methods used to analyze modern datasets
- Be able to analyze datasets using Python
- Understand how to select a good model for data
- Understand the mathematical theory and the standard models used in data science

Textbook



An Introduction to Statistical Learning. James, Witten, Hastie, and Tibshirani.

The pdf of the book is available at `http://www-bcf.usc.edu/~gareth/ISL/`

Topics

The materials of the course can be organized

- By problems:
  - Classification
  - Regression
  - Clustering
  - Manifold learning
- By methods:
  - Regression-based methods
  - Tree-based methods
  - Network-based methods
- By meta-level techniques:
  - Regularization
  - Kernel trick
  - Boosting
  - Bootstrapping

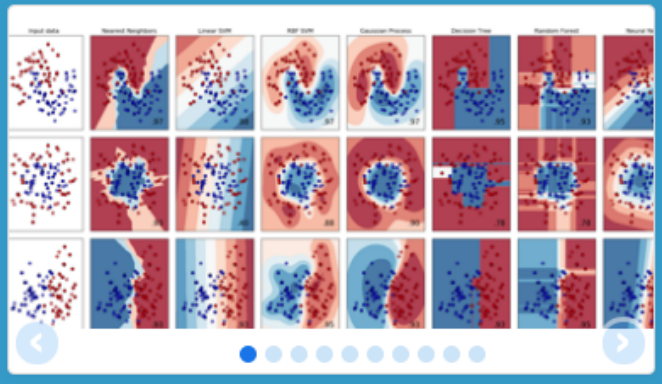
scikit-learn

scikit learn

Home | Installation | Documentation | Examples

Google Custom Search

Fork me on GitHub



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

Prerequisites

- Probability theory and basic statistics (e.g. MATH 350 and MATH 450)
- Knowledge of algorithmic concepts. Comfortable programming in a high-level language
- Multivariable calculus (e.g. MATH 243)
- Linear algebra (e.g. MATH 349)

Homework 0

- on the course webpage
- serves as a self-assessment whether you have enough back-ground knowledge for the course
- will not be graded
- an attempt of all questions (regardless of the correctness) will earn you **+2% bonus** toward the final grade
- due 02/21

Evaluation

- Homework (theoretical + programming problems): 60%
- Final project: 40% (10% presentation, 30% final report) with a possible **+5% bonus**
- Grading system:

$\geq 94\%$ At least A
$\geq 90\%$ At least A-
$\geq 80\%$ At least B-
$\geq 70\%$ At least C-
$\geq 60\%$ At least D-
$< 60\%$ F

Plattforms

- We will use Python during the course (there will be sessions to review the language). Specifically, we will use Google Colab for coding and programming assignments:

  `https://colab.research.google.com`

- We will use LaTeX to write the final report. The easiest way to use it collaboratively is to register an Overleaf account:

  `https://www.overleaf.com`

Homework policy

- Copying solutions in whole or in part from other students or **any other source** without acknowledgement constitutes cheating.
- Any student found cheating risks automatically failing the class and will be referred to the Office of Student Conduct
- You can discuss with other students, but must write up your own solutions/codes
- Please note your collaborators on your submissions

Final project

- Group project: 4 people (sign up on Canvas)
- The groups should be formed by the end of Week 4
- Data-oriented projects
  - Pick a practical learning problem with a dataset
  - Analyze the dataset
  - Write a report (in the form of a 4-page IEEE conference paper)
  - Present the project (last week of the semester)

Final project: some datasets

1. Fraud detection
2. Predict survival on the Titanic
3. Predict air pollution
4. Predict corporate credit rating
5. Predict next-day rain in Australia
6. Sign language MNIST
7. Job change prediction
8. House price prediction
9. Healthcare analytics
10. Predict water quality

Final project: scope

- your project should focus on only **one task**
- you will be graded based on:
  - how you apply the knowledge in the course to approach the problem
  - whether your experiment setups are reasonable to evaluate your methods
  - whether your conclusions are supported by your experiments
  - the clarity of your report.
- you are **not** graded based on
  - your model's accuracy
  - whether you can successfully solve the task

Final project: bonus

You can score a maximum 5% bonus points toward the final course grade by:

- using models/materials that are not covered in class, or
- adding novelty to your project (such as proposing a new method)

Final project: report

- in IEEE conference format
- maximum length: 4 pages + 1 additional page for the references
- should include:
  - An abstract (short paragraph summarizing your work)
  - An introduction (giving an overview of your work)
  - Related work (discussing briefly previous work on the problem)
  - Data and Methods (discussing the problem, the dataset, and your approaches to the problem, etc.)
  - Experiments (detailing your experiment setups to evaluate your methods, presenting and discussing the results of your experiments; include any tables or figures to show your results)
  - Discussions and conclusions (any discussions and conclusions that you can draw from your work)

Tentative schedule

- Introduction to (supervised) machine learning (4 weeks)
- Mathematical techniques in data science (8 weeks)
- Final project presentations (1 week)

Introduction to supervised learning (4 weeks)

- Week 1: Intros and reviews. Working with Python and sklearn.
- Week 2-3: Basic methods (Nearest neighbors, Logistic regression, LDA, SVM, Decision tree, Feed-forward neural nets). Formulations and demos.
- Week 4: Deep learning

Mathematical techniques in data science (8 weeks)

Meta-level techniques

- Week 5: Intro to statistical learning theory
- Week 6: SVM and the Kernel trick
- Week 7 and 9: Model selection and regularization
- Week 10: Boosting, bagging, bootstrapping

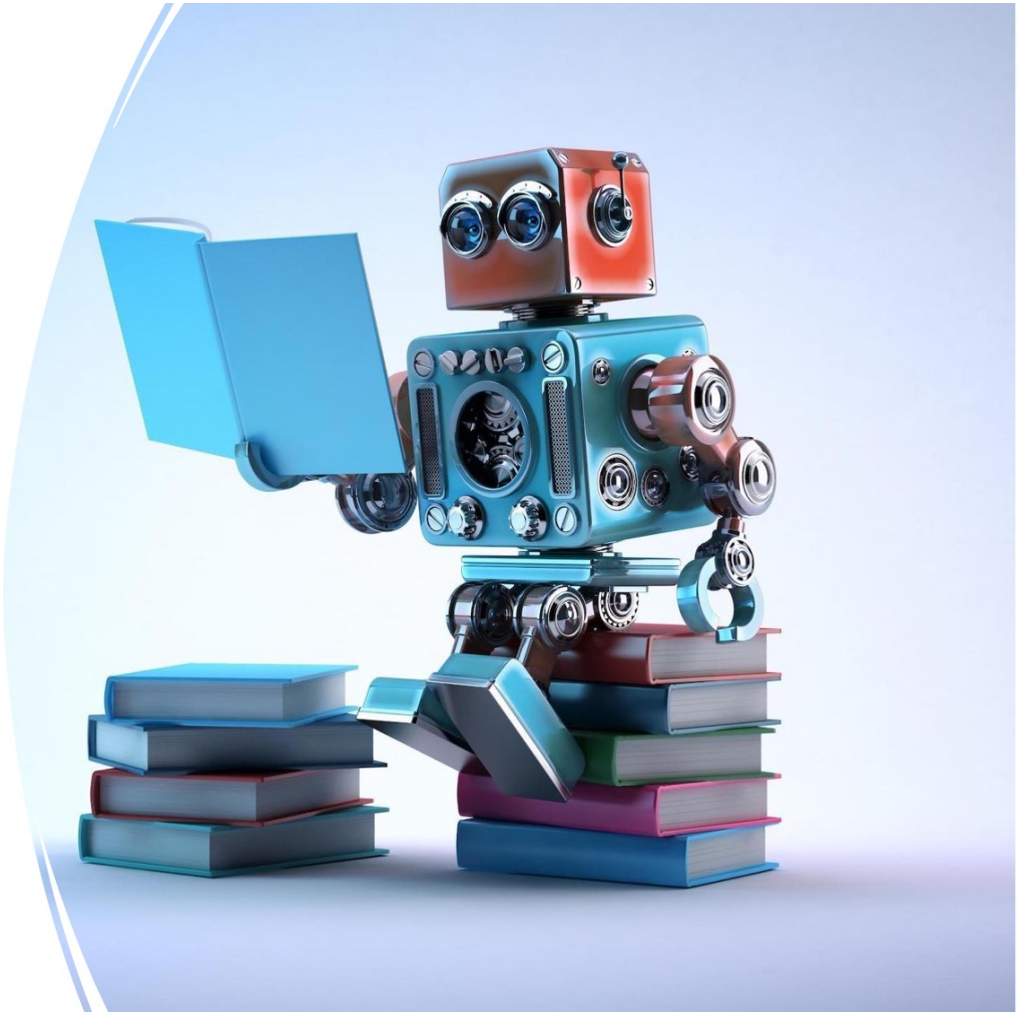
Other learning contexts

- Week 11: PCA and Manifold learning
- Week 12: Clustering
- Week 13: Selected topics

Questions?

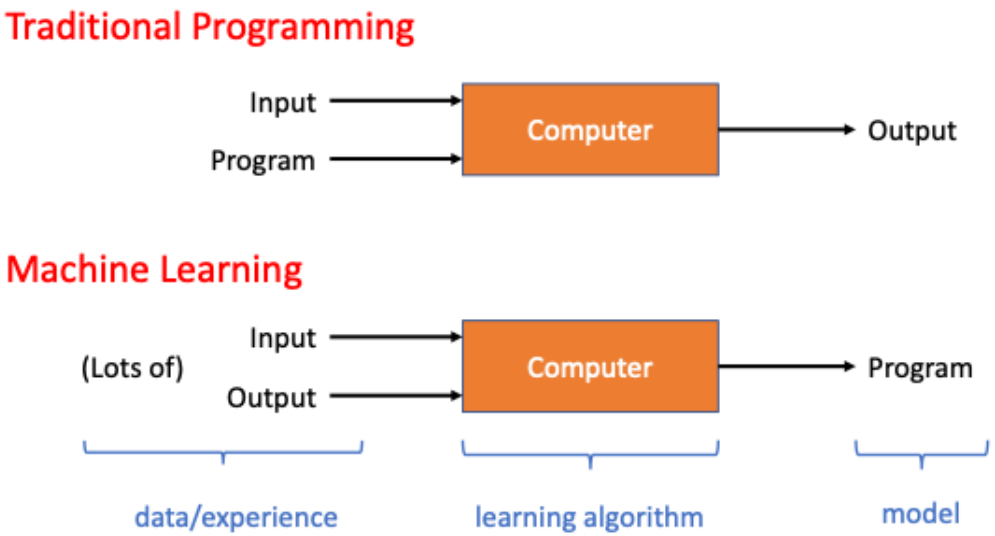
An introduction to machine learning

What is Machine Learning?

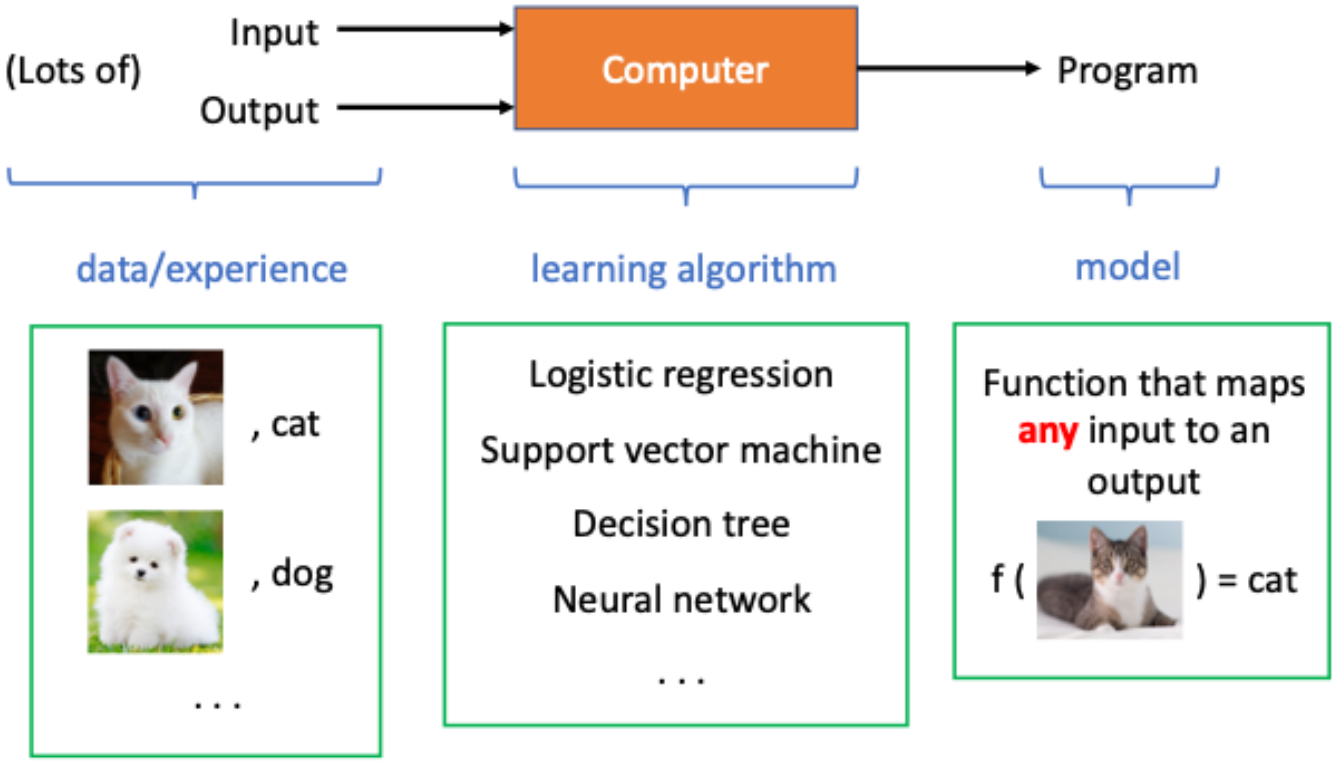


What is Machine Learning?

- A field that studies "algorithms that allow computer programs to automatically improve through experience." Tom Mitchell (1997)



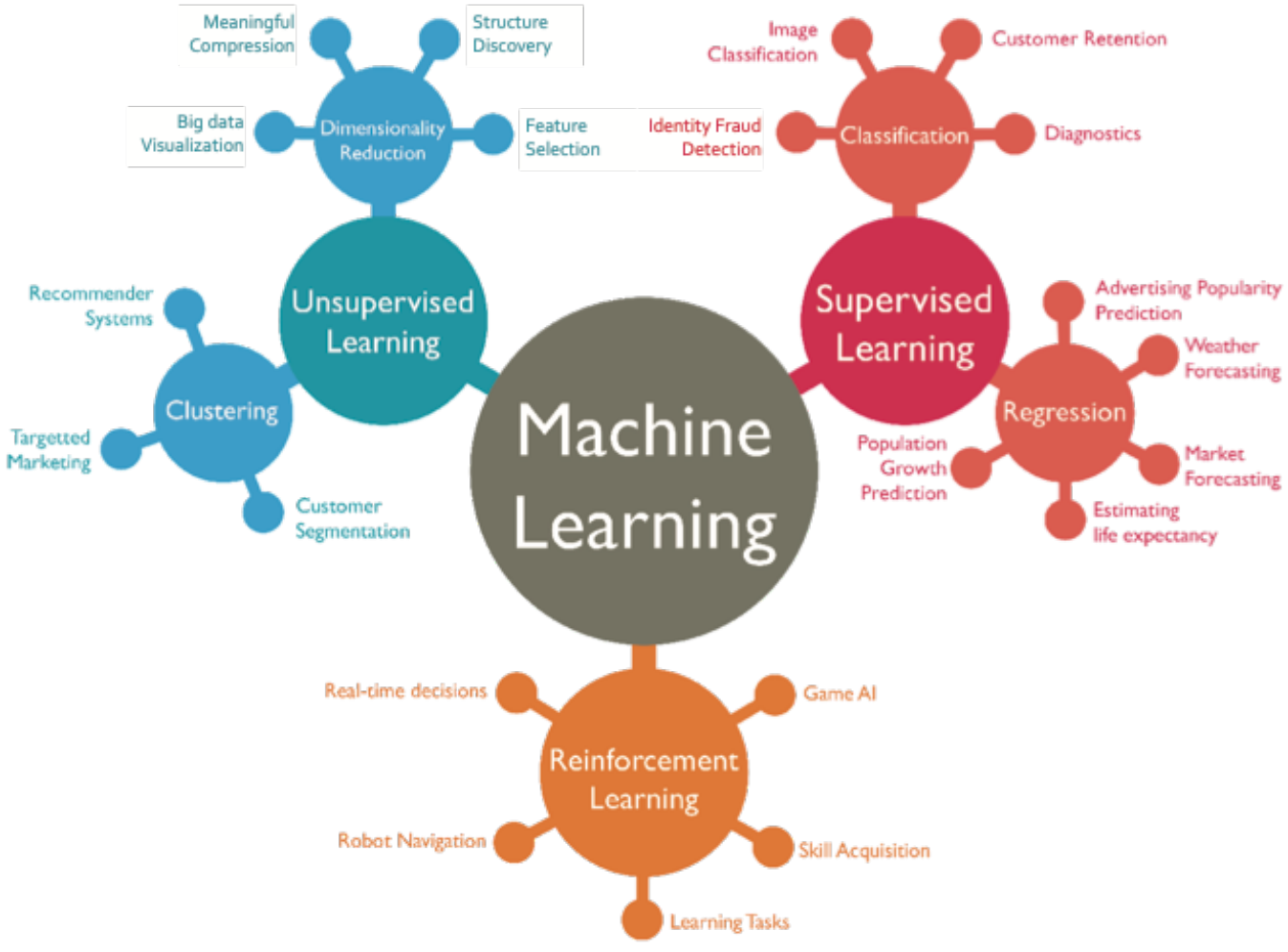
Machine learning components



ML Paradigms

- Supervised learning
  Learn a function that maps an input to an output (input, output) pairs are given as examples.
- Unsupervised learning
  Learn patterns from inputs only (no outputs).
- Reinforcement learning
  Learn to take actions to maximize some reward

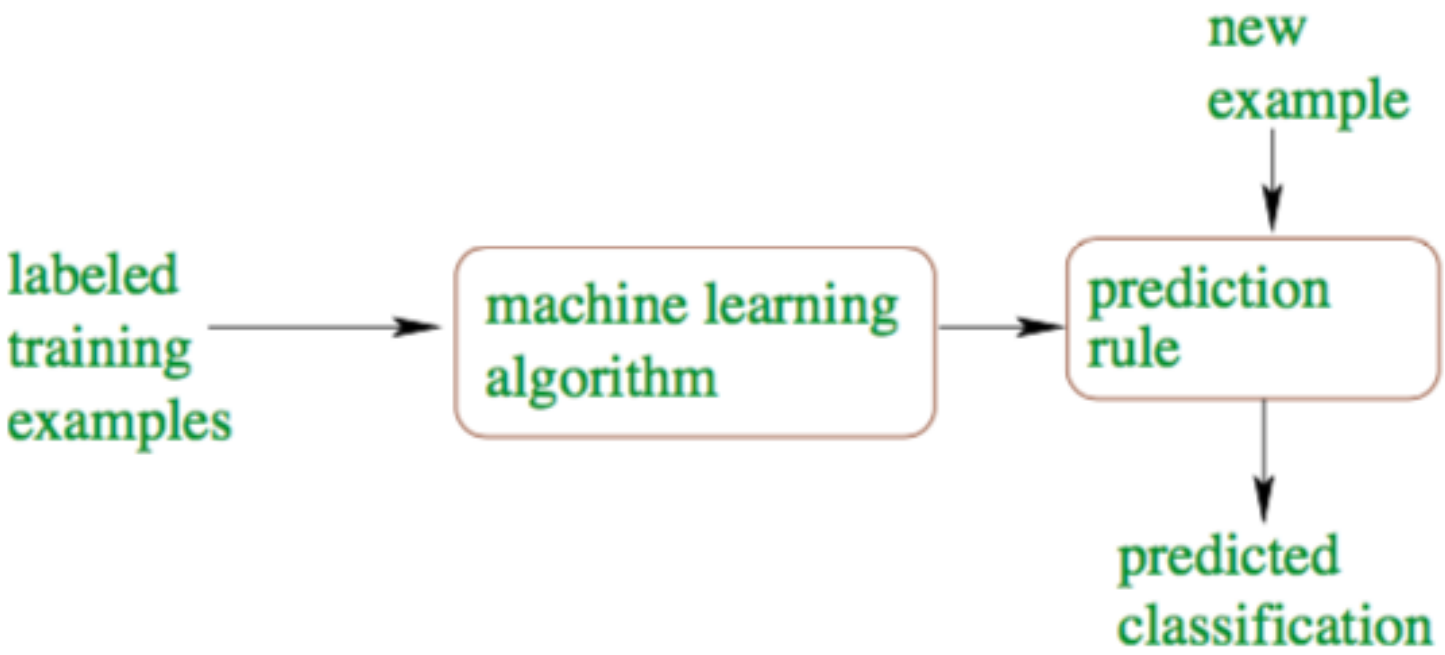
ML Paradigms



(Source: Abdul Rahid)

Supervised learning

In the first 4 weeks, we will focus on supervised learning



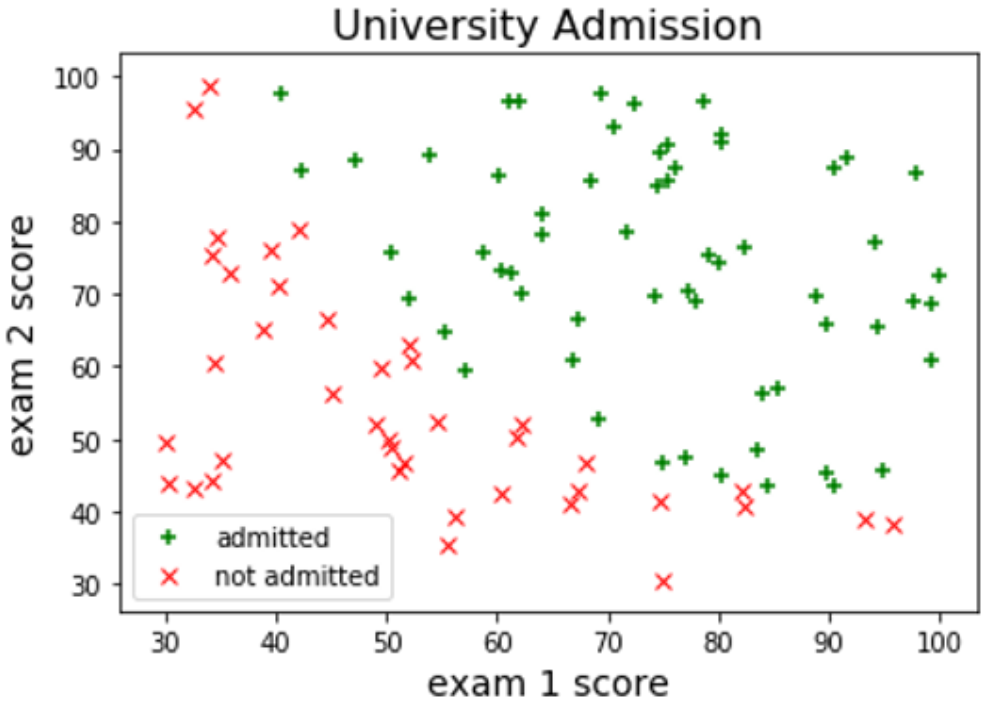
Supervised learning: learning a function that maps an input to an output based on example input-output pairs

Supervised learning

- One example contains both input (X) and output (Y)
- Two most common tasks:
  - Classification: discrete output Y
  - Regression: continuous output Y

Classification or regression?

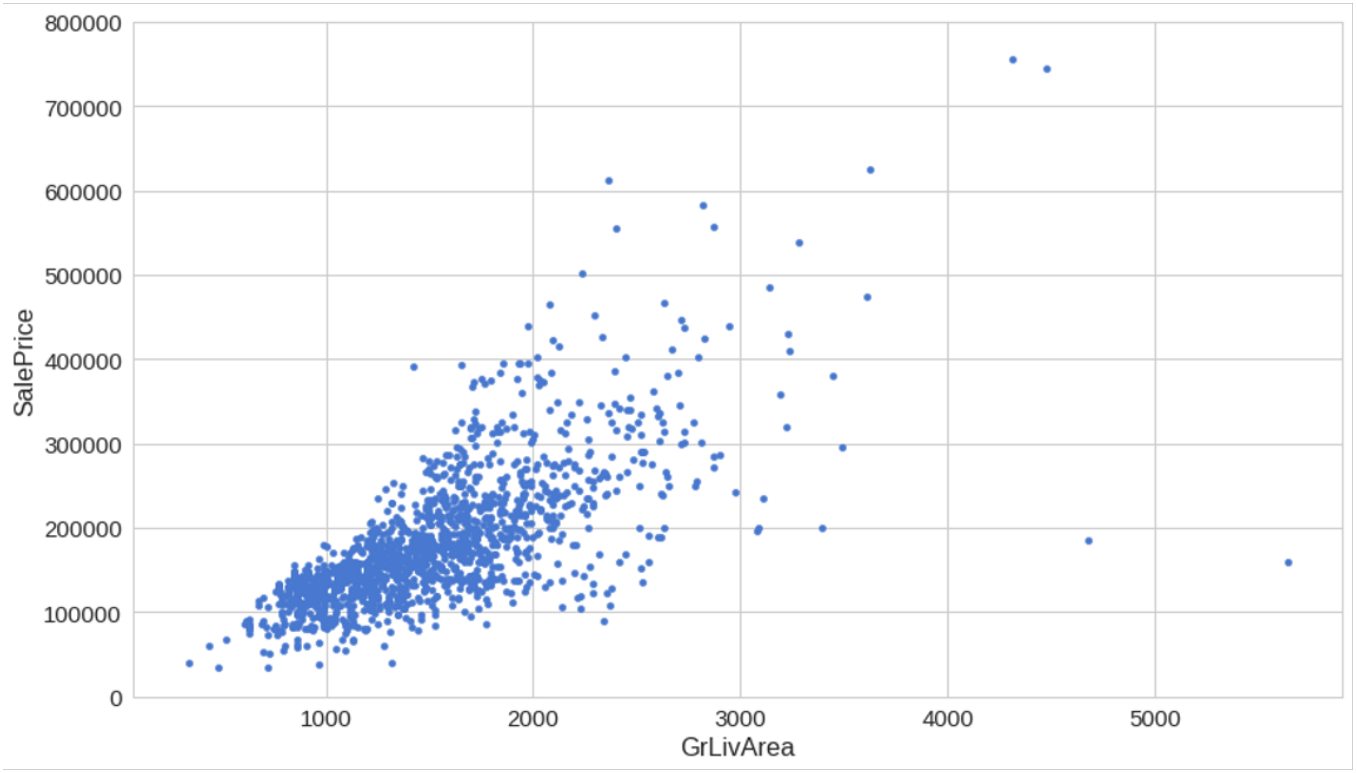
Example: Predict university admission based on exam scores



$\rightarrow$ Two classes: admitted/not admitted $\rightarrow$ Binary classification

Classification or regression?

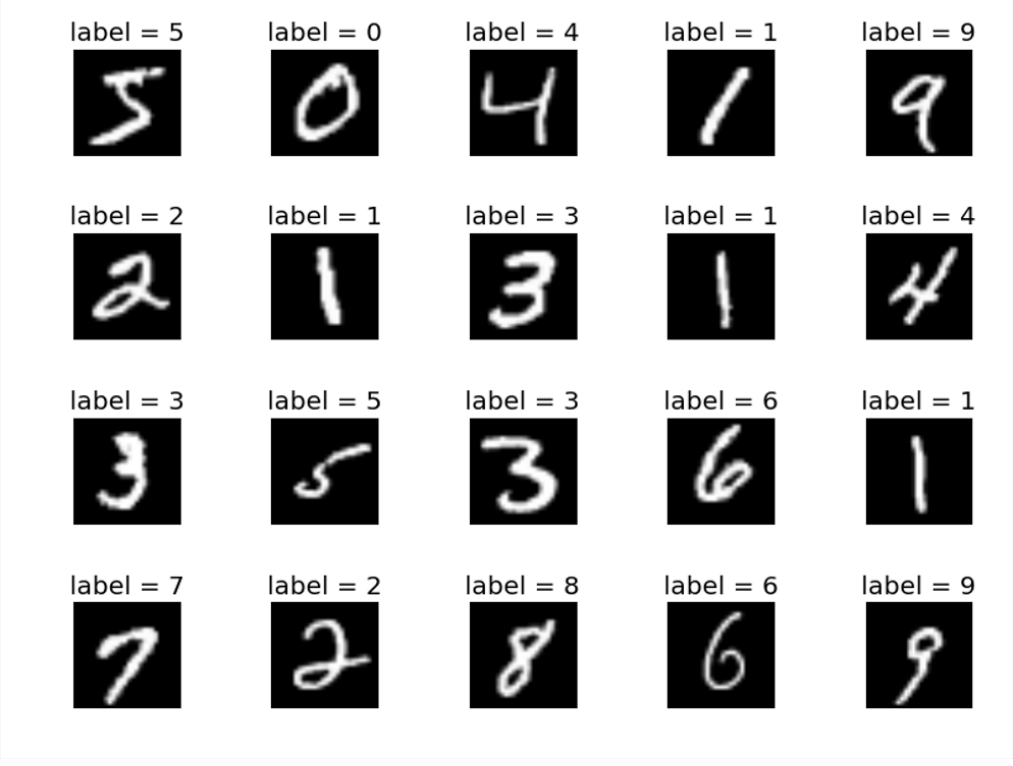
Example: Predict house price by living area



$\rightarrow$ Regression

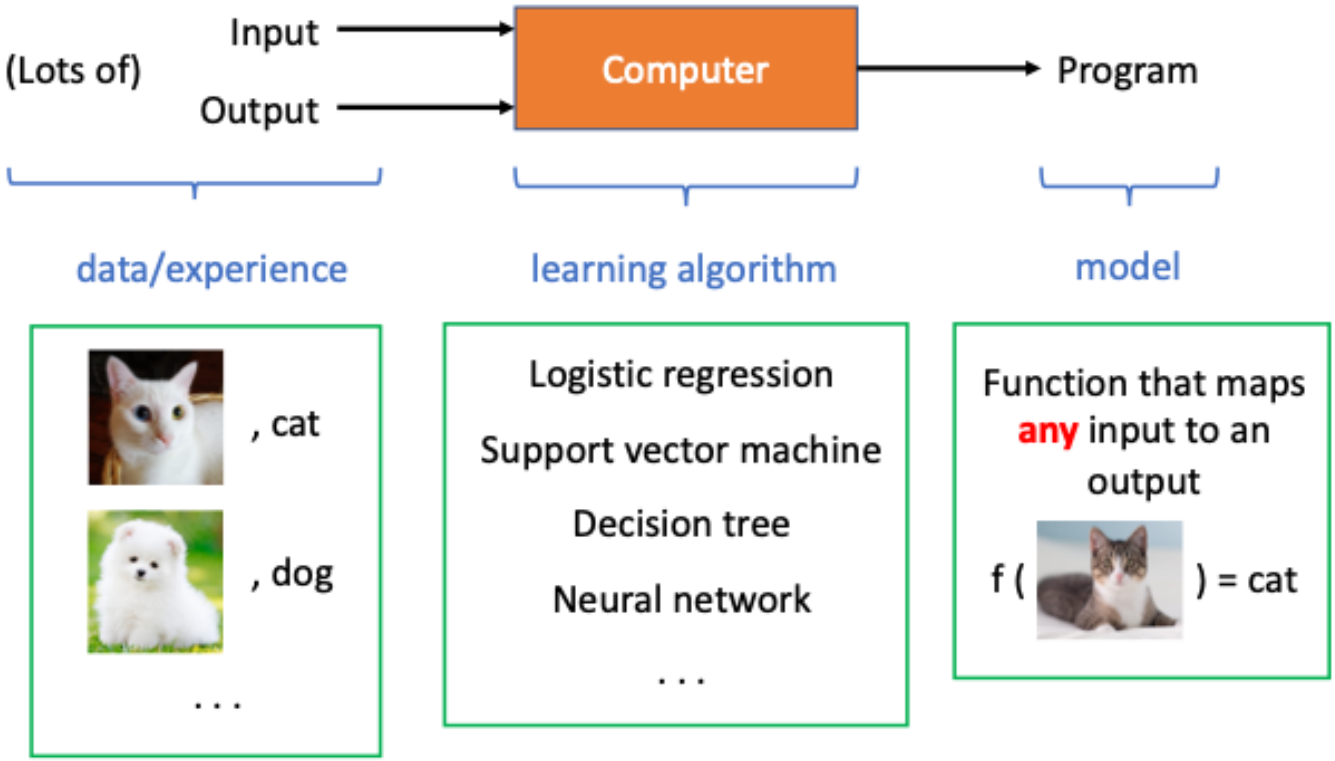
Classification or regression?

Example: Handwritten digit recognition



$\rightarrow$ Multiple classes (labels) $\rightarrow$ Multi-class classification

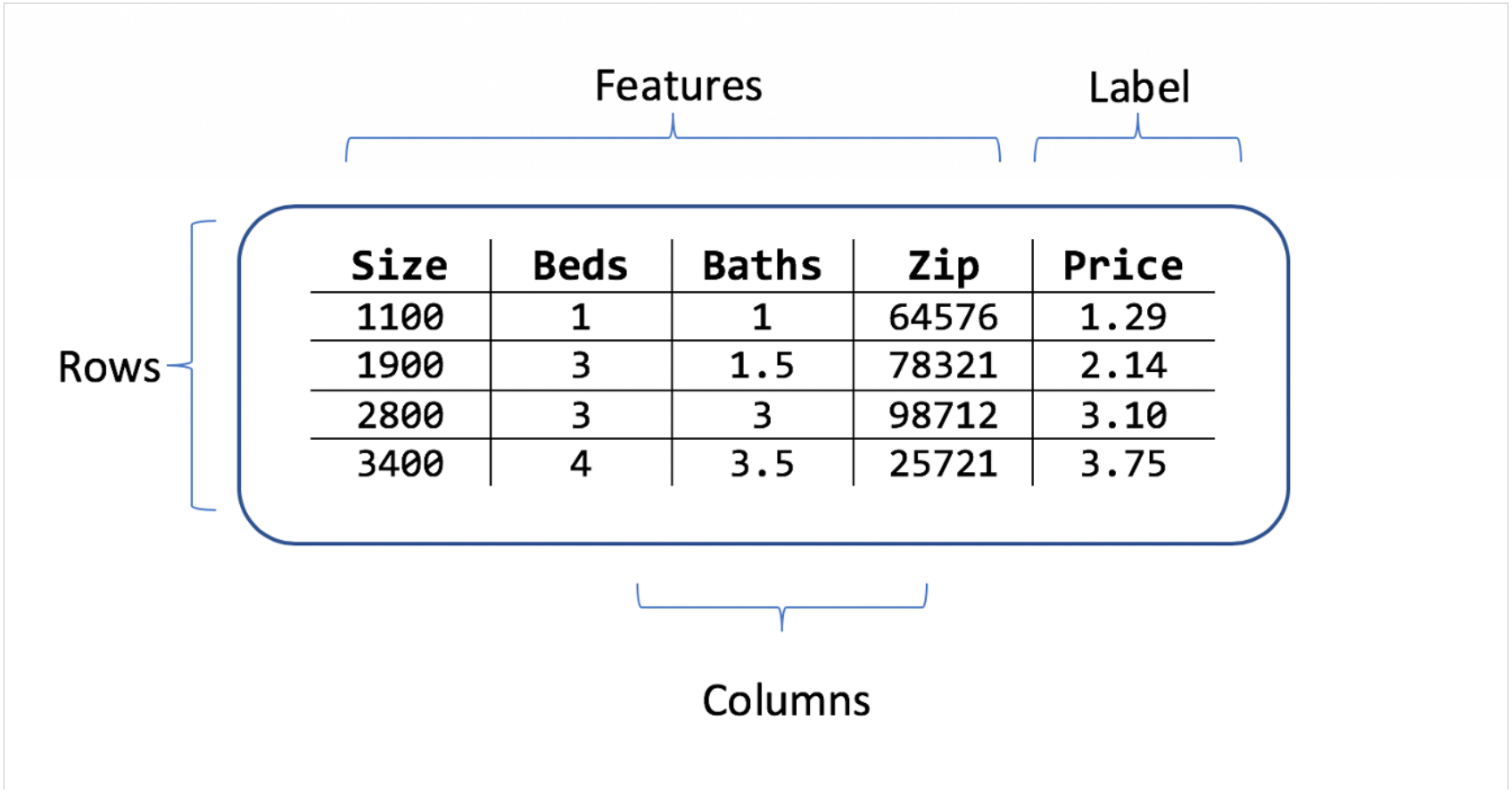
Machine learning components



Supervised learning: data

- In principle, data can be in any form
- Raw and complex data are hard to use $\rightarrow$ may need to manually (by humans) extract features before usage
- You may also need to pre-process certain features
  - Handle missing values
  - Normalize data

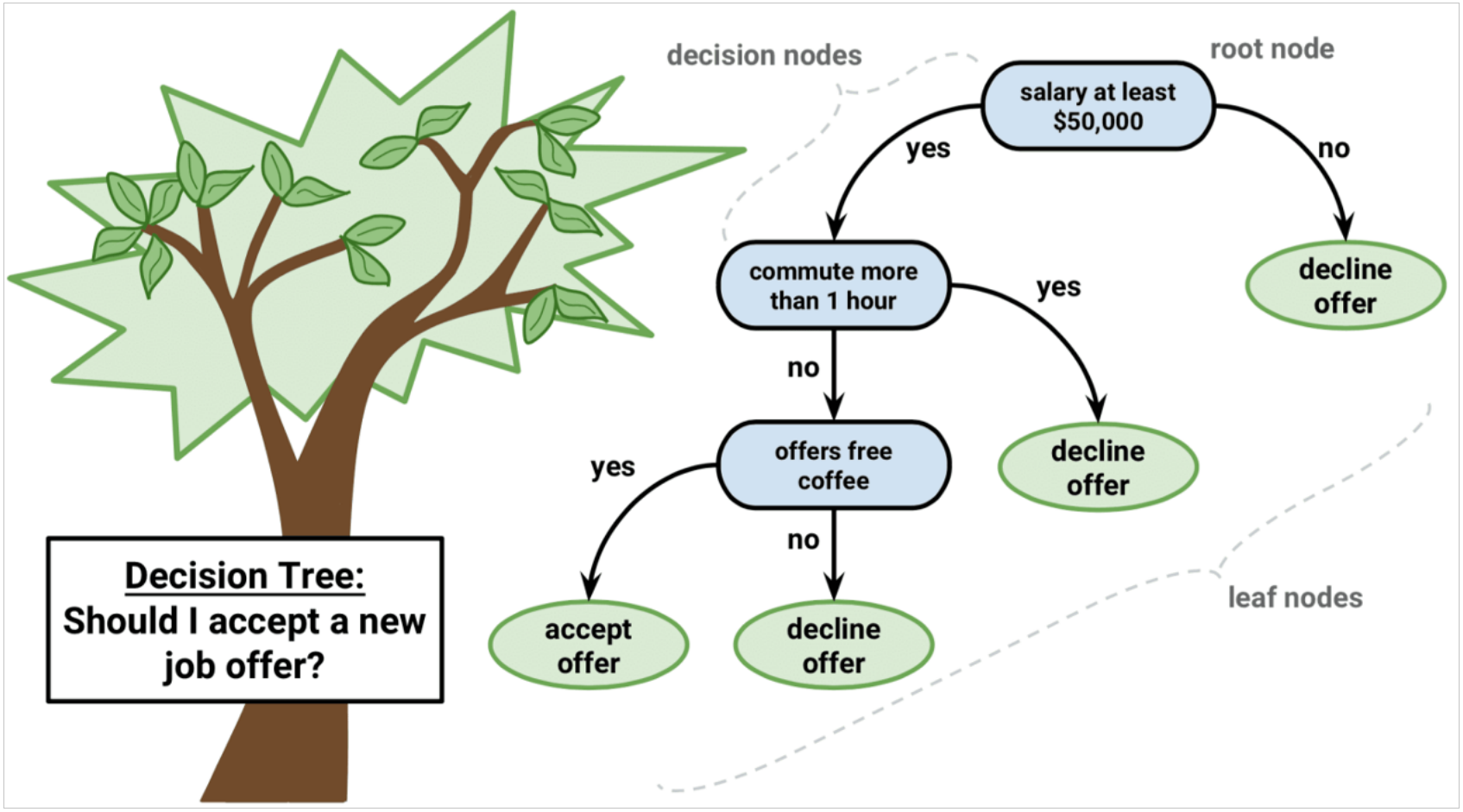
Sample data



(Source: Microsoft)

Model

Function that takes a pre-processed feature vector and predicts its label



(Source: packtpub.com)

Learning algorithm

- An algorithm that returns the **parameters** or **configurations** of the model from data
- Note: learning algorithm = training algorithm
- May need to optimize an objective or loss function

Evaluate a learned model

- How effective the model makes predictions on new (unseen) data
- Classification: accuracy or error rate
- Regression: average (squared) distance between predicted and true values (mean squared error)

Data splitting practices