MACQUARIE University

MACHINE LEARNING

COMP8220

01 – Introduction

Xuyun Zhang (Department of Computing)

Teaching Staff

MACQUARIE University

- Dr Xuyun (Sean) Zhang
  - Lecturer for 1st half
  - Rm 287, BD Building
  - xuyun.zhang@mq.edu.au
  - Ph: 02 9850 8229
  - Webpage: https://researchers.mq.edu.au/en/persons/xuyun-zhang

- Career path

- Research interests
  - Scalable and secure machine learning
  - Big data privacy and cyber security

Unit Scheduling

MACQUARIE University

Machine Learning

Part I: Conventional machine learning techniques for regression, classification, and clustering

Part II: Deep learning models

Lecture Outline

MACQUARIE University

- What is Machine Learning?
- Types of Machine Learning Systems
- Main Challenges
- Machine Learning in Python

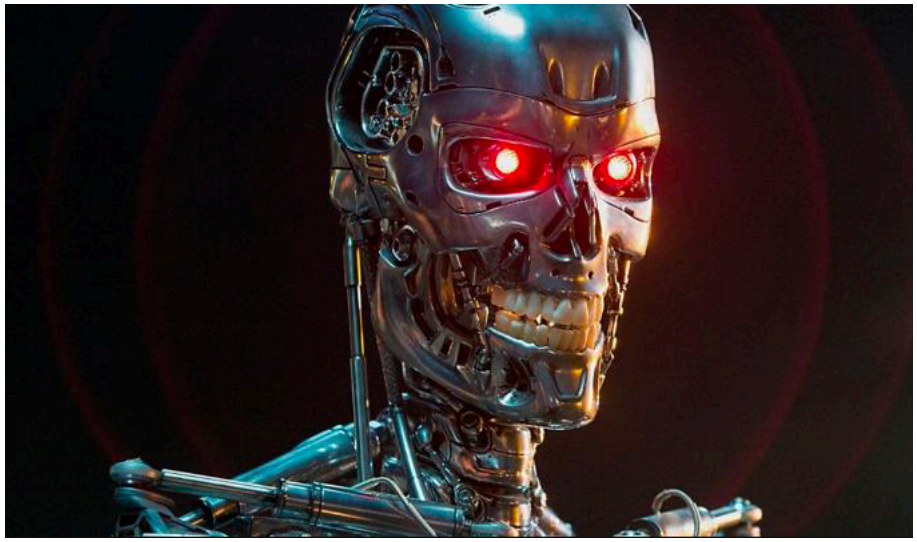
Lecture Outline

MACQUARIE University

- What is Machine Learning?
- Types of Machine Learning Systems
- Main Challenges
- Machine Learning in Python

What is Machine Learning?

MACQUARIE University



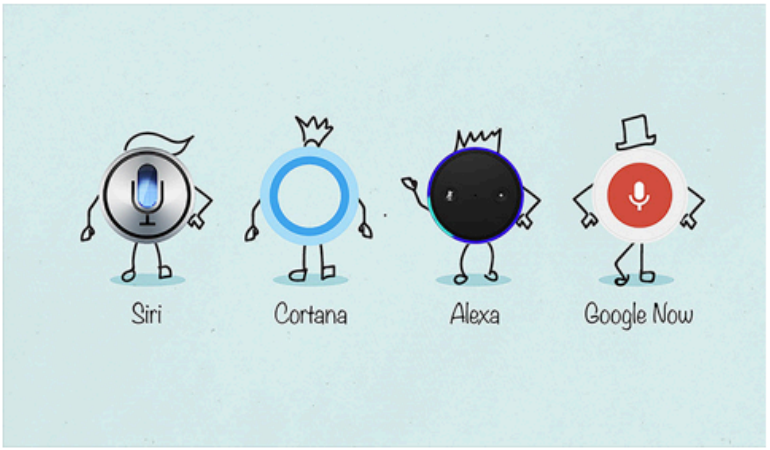
What is your first impression?

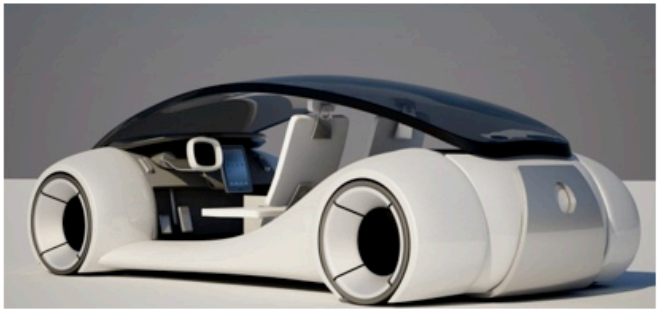
State-of-the-art Applications

MACQUARIE University

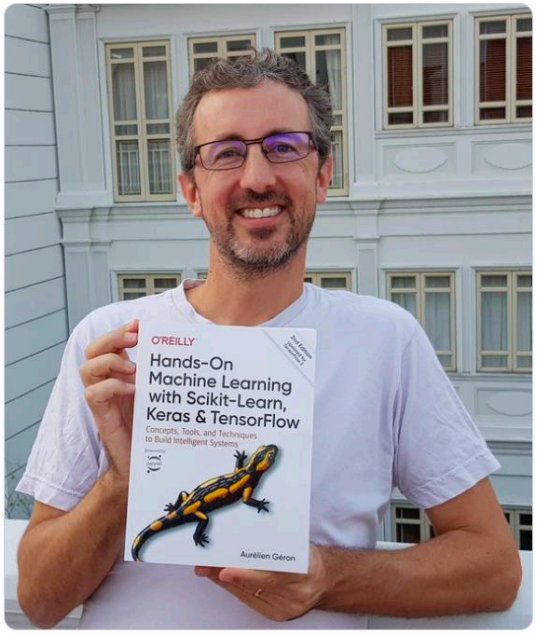
AlphaGo

Google DeepMind





Machine Learning Definition



Machine Learning is the science of programming computers so that they can learn from data.

(Aurélien Géron, 2019)

Machine Learning Definition

MACQUARIE University

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

(Arthur Samuel, 1959)

Machine Learning Definition

MACQUARIE University

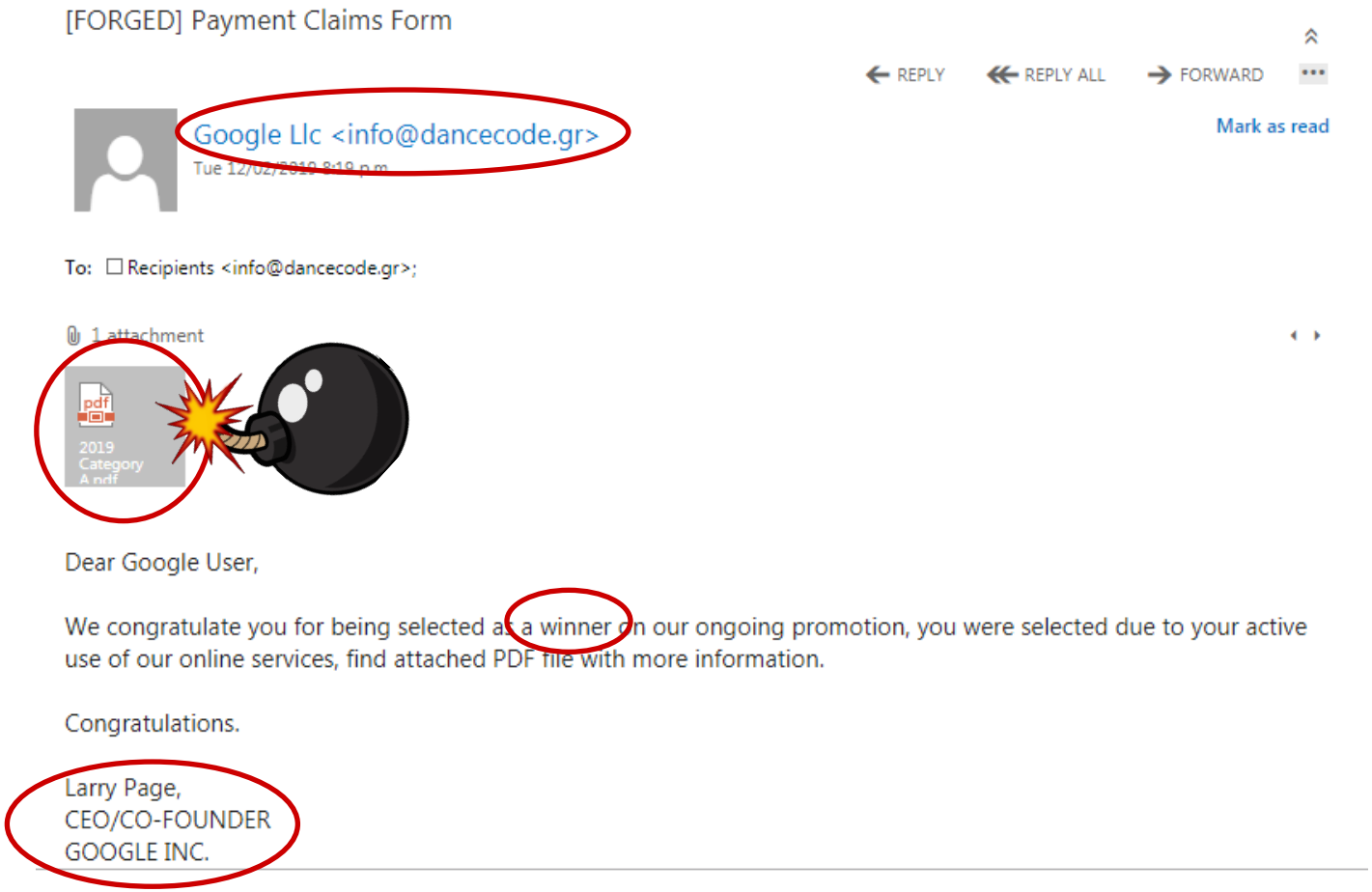
A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

(Tom Mitchell, 1997)

Example: Spam Filtering

MACQUARIE University

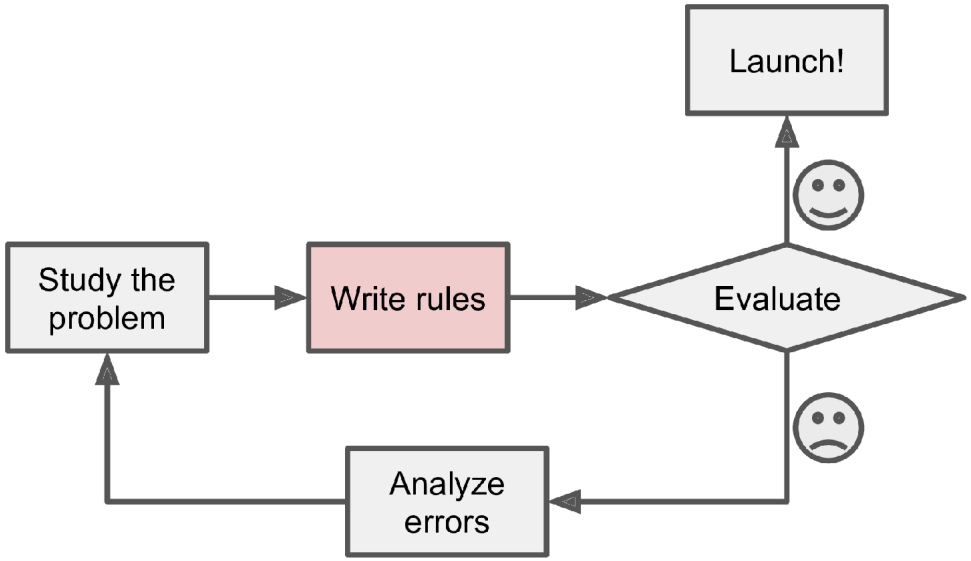
- Task: Given an email, a spam filter must classify the email as bad (spam) or as good (ham)



Example: Spam Filtering

MACQUARIE University

- Traditional way: hand-crafted rules
  - E.g., if it contains 'winner', then classified as a spam

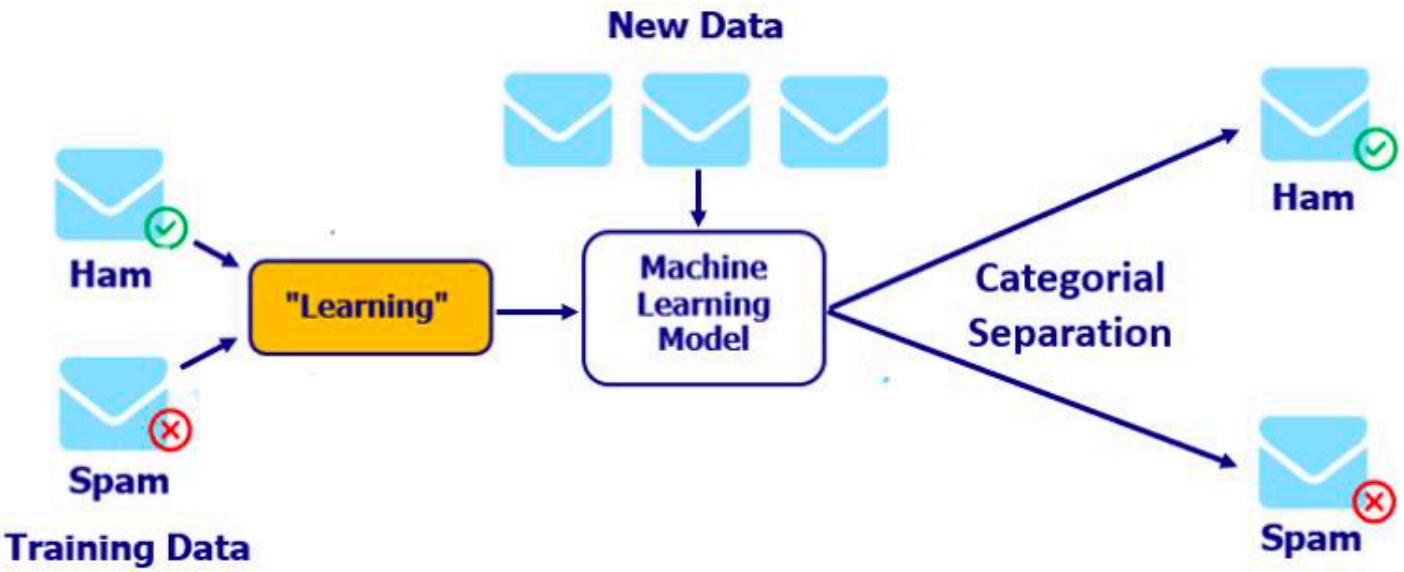


- Limitations
  - Need to memorize many complex rules to detect patterns
  - Spam creators often change their tactics over time

Example: Spam Filtering

MACQUARIE University

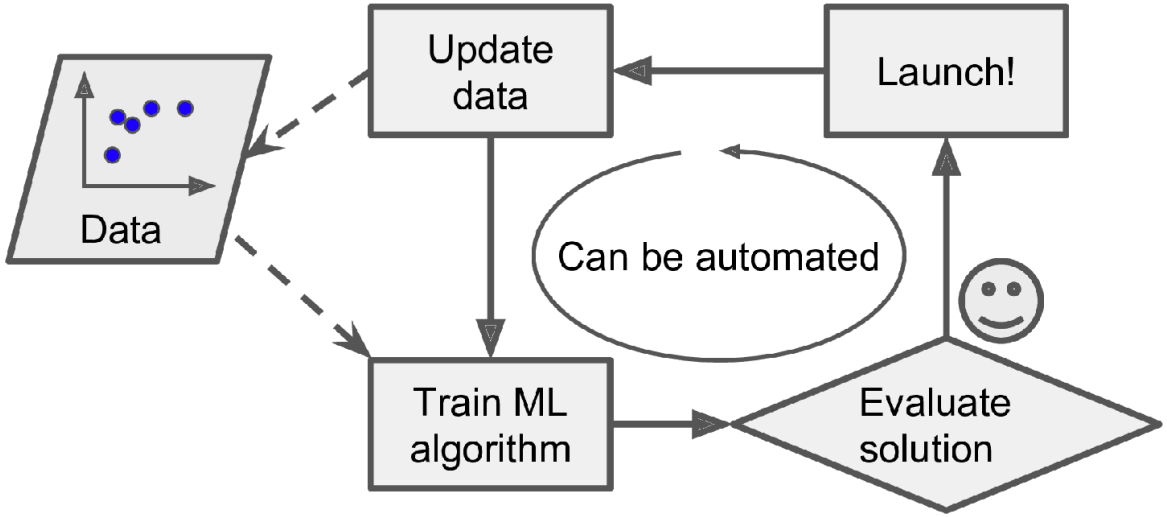
- A spam filter based on machine learning can automatically learn which words and phrases are good predictors for spam using term frequency
- Architecture



Example: Spam Filtering

MACQUARIE University

- Machine learning based detection
  - It requires: examples of regular emails and spam emails



- Three learning aspects
  - T: to separate spam from regular emails
  - E: both regular and spam emails
  - P: can be defined, e.g., ratio of correctly classified emails

What if No Learning?

MACQUARIE University

- Experience (email data):

ID	'winner'	'payment'	'attachment'	spam
1	yes	yes	no	yes
2	yes	no	yes	yes
3	no	yes	yes	no

- Task: to predict if an email is a spam or not
- Performance: accuracy
- No learning: to randomly guess for a future email
  - Accuracy: 50%
- A simple learning: predict a future email with 'yes'
  - Accuracy: 66.7% > 50%

Learning vs Memorizing

Sees:

So, what really matters in learning is the capability of generalization from experience for unseen instances

Learning vs Memorizing

MACQUARIE University

- When do we use machine learning?
  - Human expertise does not exist (navigating on Mars)
  - Humans are unable to explain their expertise (speech recognition; face recognition; driving)
  - Complex problems for which no good solution exists
  - Solution changes in time (routing on a computer network; browsing history; driving)
- When do we **not** use machine learning?
  - Calculating payroll
  - Sorting a list of words
  - Monitoring CPU usage
  - Querying a database

Machine Learning Ingredients

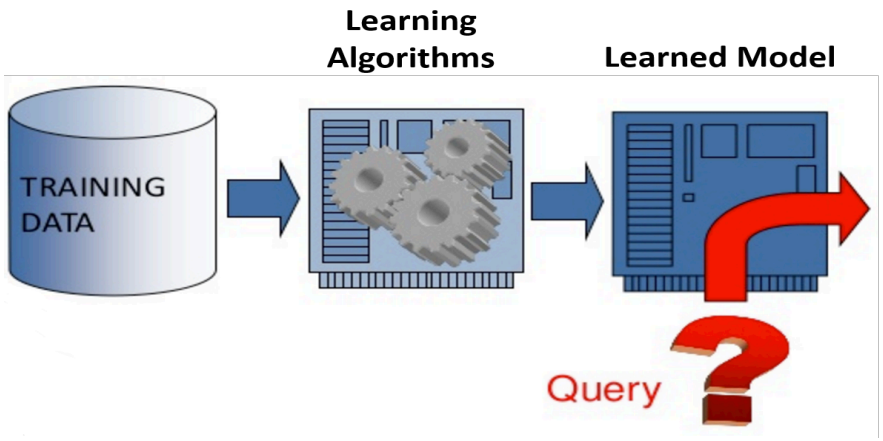
MACQUARIE University

- Data: experience (E) is embedded in the form of data
  - E.g., email text (raw data)
- Learning models: the task (T) we need to specify
  - Regression, classification, or clustering?
  - High-level abstraction learned from observations
- Learning algorithm: the computer program
  - E.g., decision trees, Naïve Bayes classifier, stochastic gradient descent, maximum likelihood estimation, etc.
- Testing and evaluation: check if the performance (P) has been improved
  - E.g., k-folder cross validation

Two Stages of Machine Learning

MACQUARIE University

- Training stage
  - Training: to generate a model from observed data
- Testing stage
  - Testing: to use the learned model to predict unseen data



E.g., does a patient with coughing, running nose and fever suffer from COVID-19 or flu?

Data

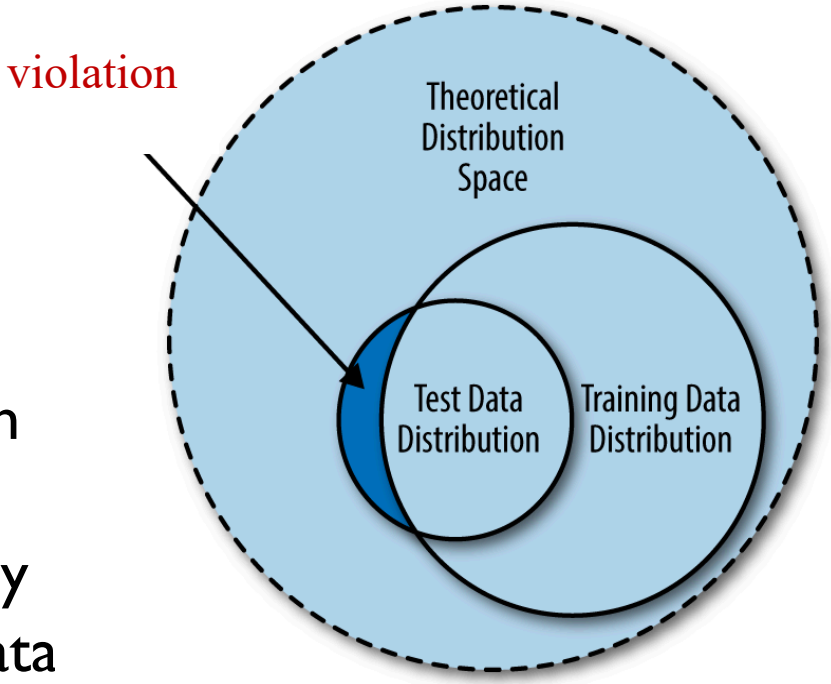
MACQUARIE University

- **Data set** D: a set of observed data instances
  - $D = \{d_1, d_2, \cdots, d_{|D|}\}$
  - Many types: numerical vectors, an image, a graph, …
  - Assumption: i.i.d. (independent & identical distributed)
    - Instances in D follow a (unknown) distribution, from which each instance is independently sampled
- **Training data**: data used in the training stage
- **Testing data**: data used in the testing stage
- **Validation data**: used to select learning models
  - Part of a training data

Data

MACQUARIE University

- Assumption: the distribution of training examples is identical to the distribution of test examples (including future unseen examples)
  - In practice, this is often violated to certain degree
  - Strong violations will clearly result in poor performance
  - To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data



Data

MACQUARIE University

- Label/target of data
  - The interesting attribute(s) for prediction
  - Further, $d_i \equiv \langle x_i, y_i \rangle$
  - **Supervised learning models**
    - **Regression**: y is continuous
    - **Classification**: y is discrete (binary or multi-class)
- No explicit label/target information
  - Then, $d_i \equiv \langle x_i, \cdot \rangle$
  - **Unsupervised learning models**
    - **Clustering**
- **Semi-supervised learning** (labeling is often costly)

Model

MACQUARIE University

- **Model**: a map from input space to output space
  - i.e., $f : \mathcal{X} \to \mathcal{Y}$
  - An infinite number of such maps
  - Input space: space spanned by feature attributes
  - Output space: space spanned by label/target attributes
- **Hypothesis space** H: space of all possible maps
  - Functional space: $\mathcal{H} \equiv \{f \mid \mathcal{X} \to \mathcal{Y}\}$
- **Ground truth**: underlying true mechanisms of generating the observed data
  - But never known in reality
  - The purpose of learning: approximate the ground truth

Parameter Space

MACQUARIE University

- What do we really learn from data for a model?
  - Hypothesis space is usually pre-specified in terms of problem domains
  - Different models are determined by parameters
  - Let θ denote the parameter vector, we have

$$\mathcal{H} \equiv \{f \mid y = f_\theta(x)\}$$

- Parameter space
  - Can be real value spaces, e.g., $\mathbb{R}^n$
  - Structure of a model (more implicit), e.g., tree or graph structures, as well as partitions of the input space

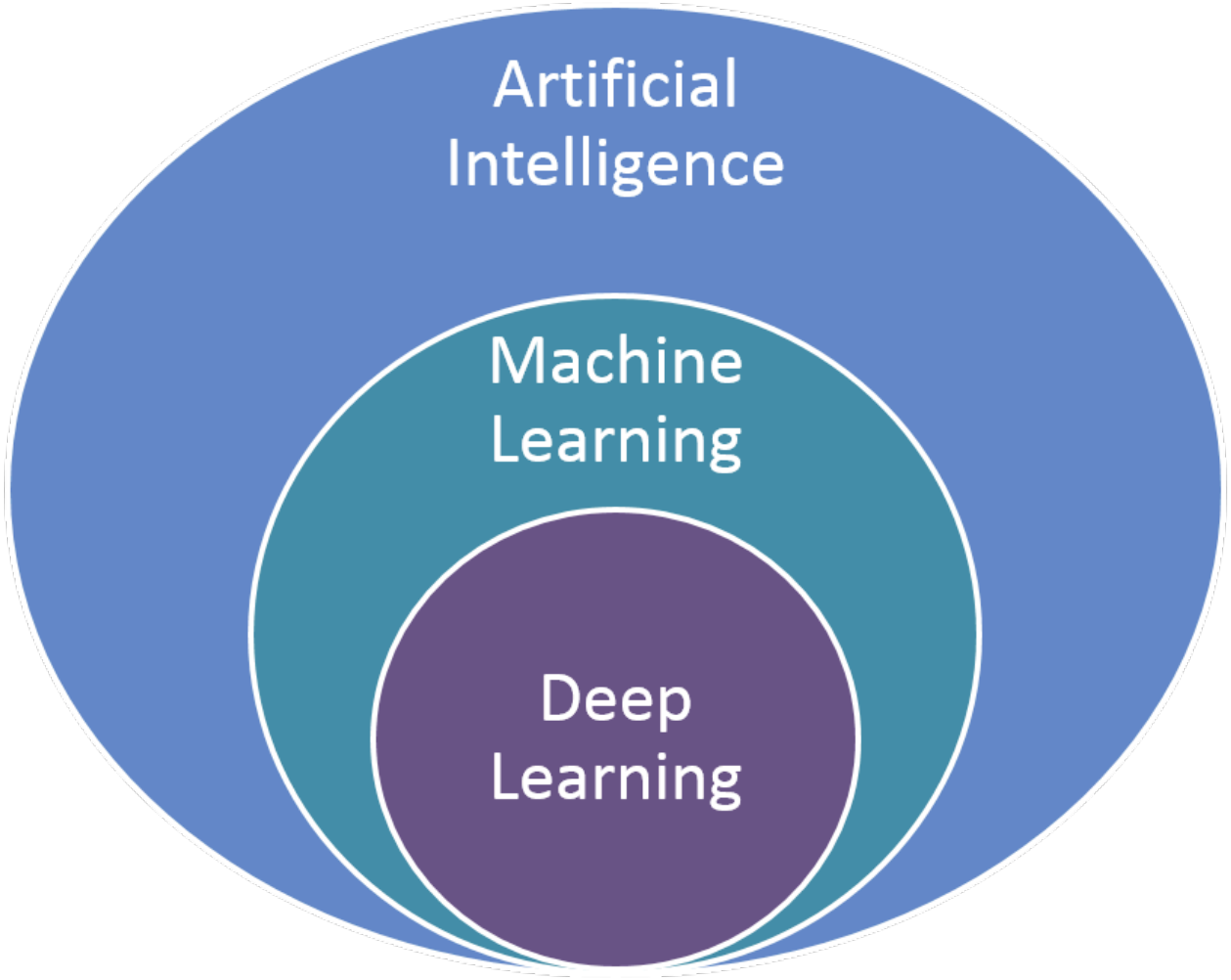
Lecture Outline

MACQUARIE University

- What is Machine Learning?
- Types of Machine Learning Systems
- Main Challenges
- Machine Learning in Python

AI, ML, and DL

MACQUARIE University



AI, ML, and DL

MACQUARIE University

- 1950s Deductive reasoning stage
  - Formal logic, e.g., proof of mathematic theorems
- 1970s Knowledge engineering stage
  - Knowledge representation, e.g., expert systems
- 1980s Machine learning stage
  - Symbolism: based on symbols, e.g., first-order logic-based learning
  - Statistical learning: applying probability theory and statistics, e.g., co-locating diaper and beer
  - Connectionism: reverse engineering brain mechanism for artificial neural nets
    - Black-box model with low interpretability

Types of ML Systems

MACQUARIE University

- One can classify ML systems along the following questions:
- Are they trained with or without human supervision?
  - => supervised versus unsupervised learning
- Can they learn incrementally or not?
  - => batch versus online learning
- Do they measure the similarity between data points or do they detect patterns and build a predictive model?
  - => instance-based versus model-based learning

Supervised vs Unsupervised

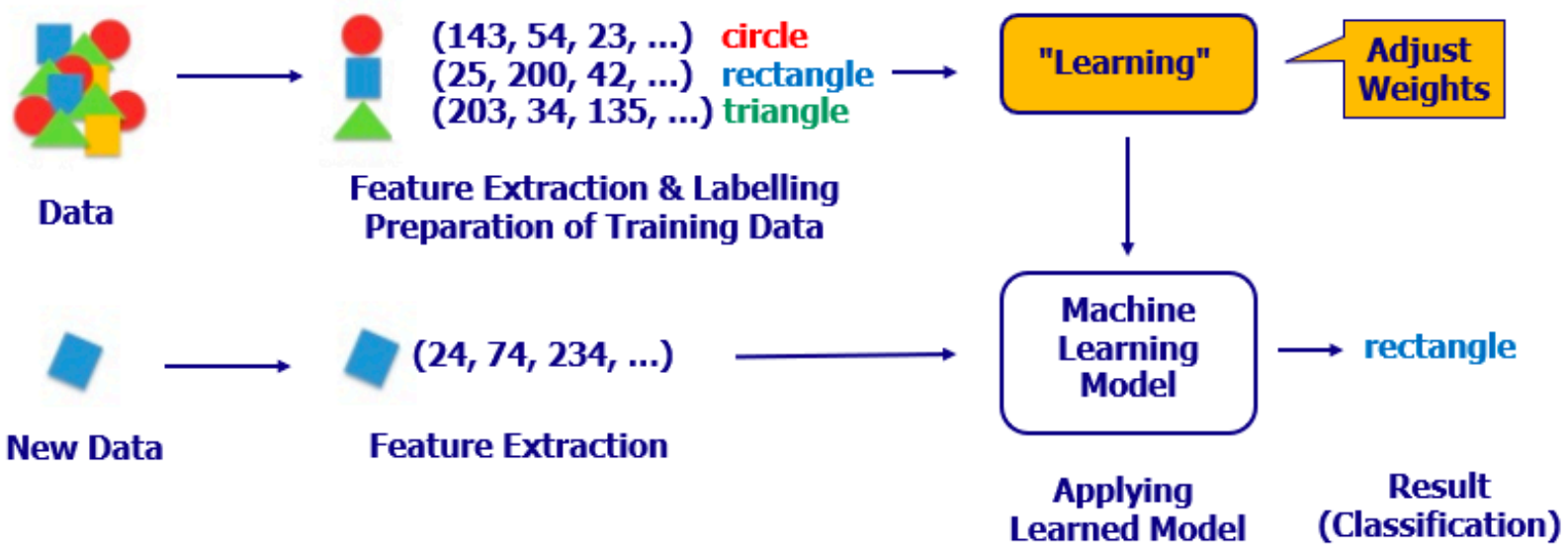
MACQUARIE University

- Supervised learning, e.g.,
  - Linear Regression, Logistic Regression
  - Support Vector Machines (SVMs)
  - Neural Networks
- Unsupervised learning, e.g.,
  - K-Means for clustering
  - PCA for dimension reduction
  - Apriori for association rule learning
- Semi-supervised learning
- Reinforcement learning

Supervised Learning

MACQUARIE University

- Training data is labelled with expected solutions

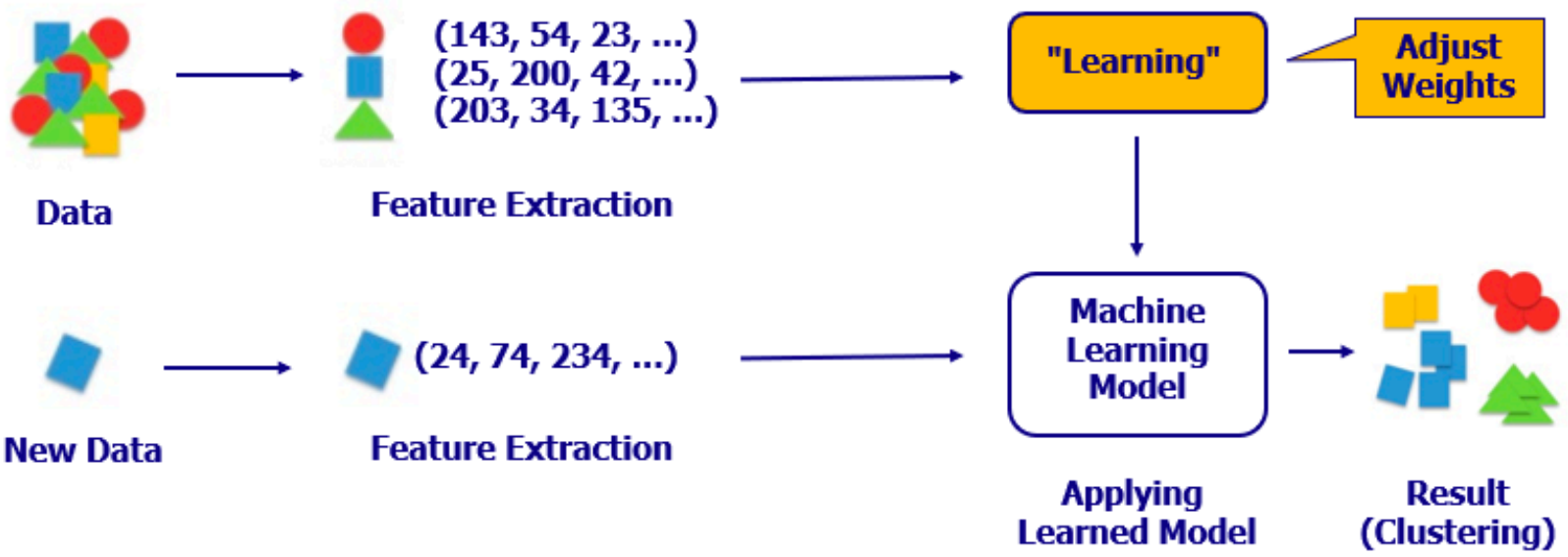


- Tasks: classification and regression

Unsupervised Learning

MACQUARIE University

- Training data is NOT labelled with expected solutions

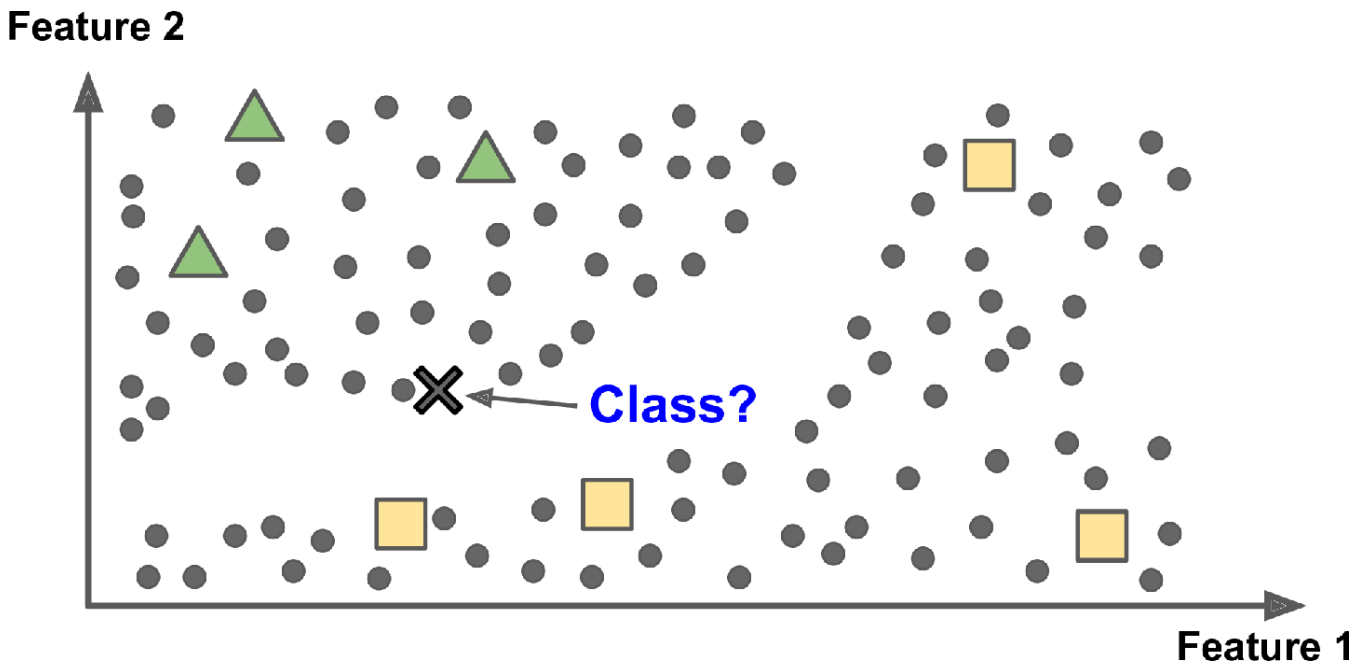


- E.g., a clustering algorithm can detect groups of similar objects

Semi-supervised Learning

MACQUARIE University

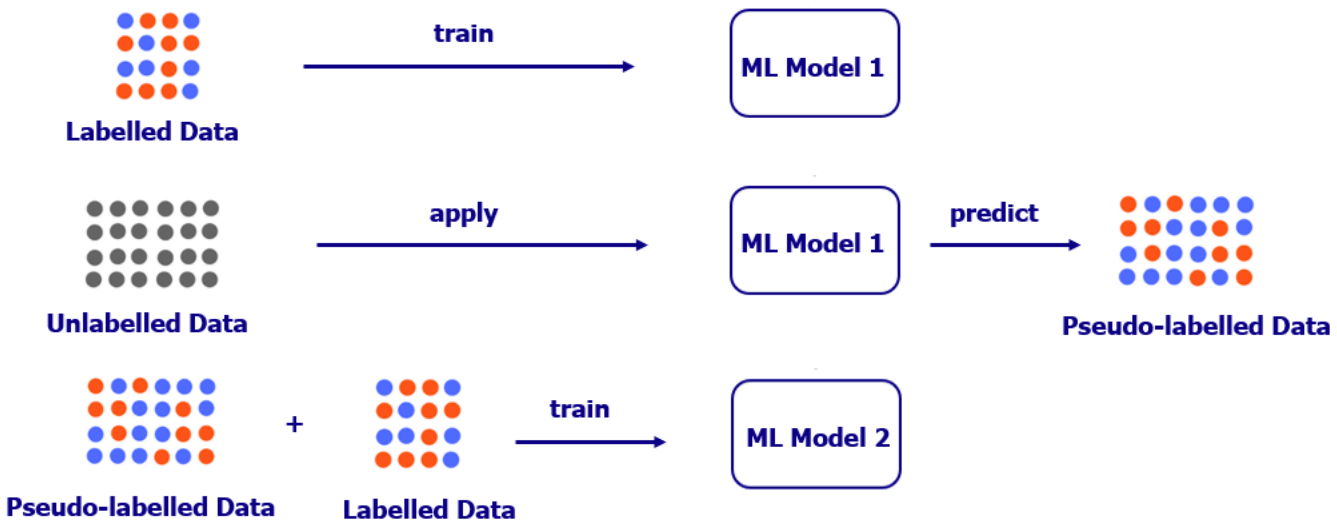
- Semi-supervised learning uses
  - A small amount of labelled data and
  - A large amount of unlabelled data



Example: Pseudo-Labelling

MACQUARIE University

- One important technique is pseudo-labelling
  - Pseudo labels are predicated using a first model trained on the labelled data
  - Pseudo labelled data and labelled data are then used to train a second model
  - Second model is then used for predictions on test data



Reinforcement Learning

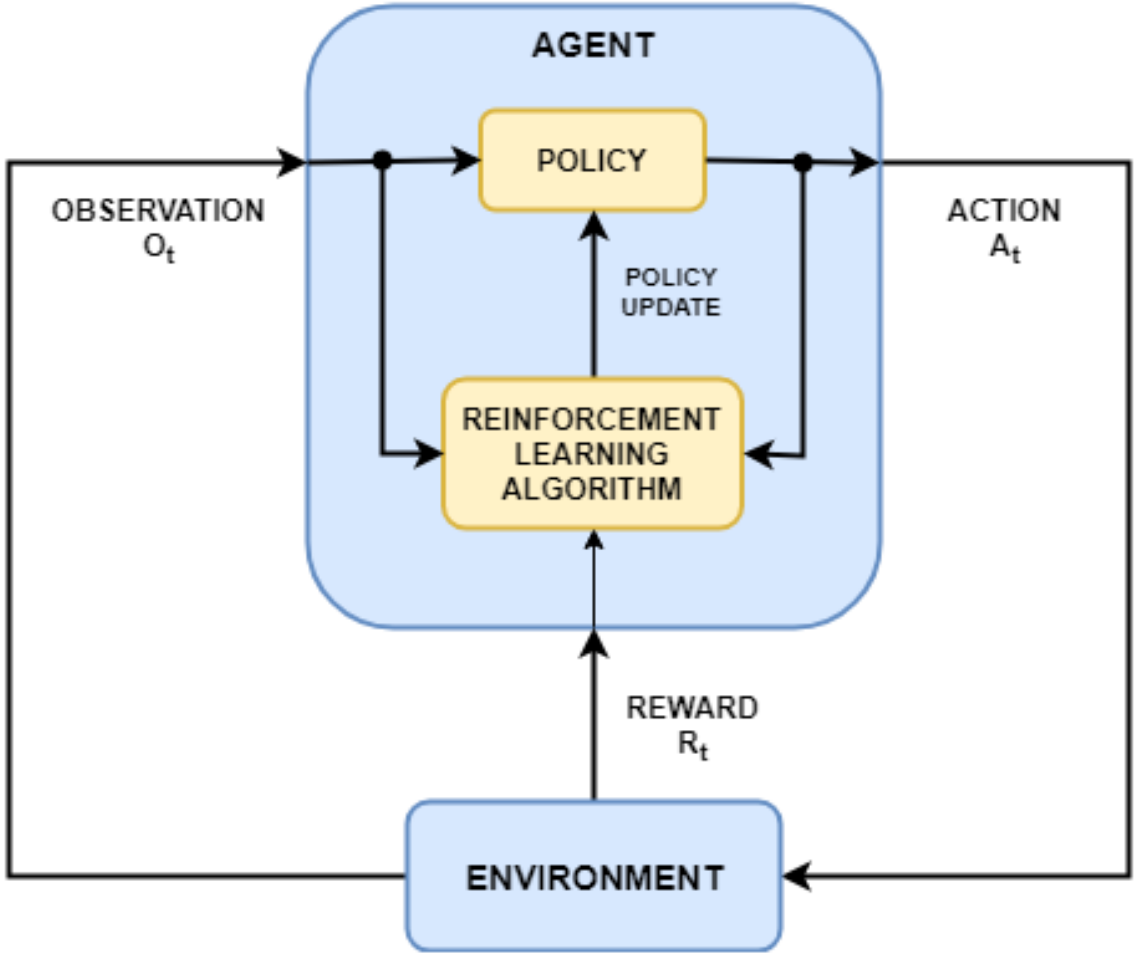
MACQUARIE University

- The aim of reinforcement learning is to train an agent via trial and error for a task in an unknown environment
  - E.g. AlphaGo

- The agent receives observations and rewards from the environment and sends actions to the environment
- The agent consists of a policy and a learning algorithm:
  - The policy defines what action the agent should choose
  - The learning algorithm continuously updates the policy parameters to maximize the cumulative reward

Reinforcement Learning

MACQUARIE University



Batch versus Online Learning

MACQUARIE University

- Can learn incrementally from a stream of data or not?
- Batch learning
  - System is not capable of learning incrementally
  - System must be trained using all available data
  - Note that re-training using the full dataset may be time consuming and may require a lot of computing resources
- Online learning
  - System is capable of learning incrementally
  - System can be trained by feeding data sequentially in small groups of data (mini batches) or as individual instances
  - Can deal with datasets that cannot fit into memory

Instance-based vs Model-based

MACQUARIE University

- One can classify ML systems by how they generalize
  - Most ML tasks are about making predictions
  - Given a number of training examples, the system needs to generalize to new instances it has never seen before
- Two main approaches to generalization
  - Instance-based learning: compares new instances with instances seen in the training data, stored in memory
    - E.g, k nearest neighbors algorithm
  - Model-based learning: uses features in the training data to predict the target variable
    - E.g., linear regression models: A linear regression model predicts the target variable as a weighted sum of the feature inputs

Lecture Outline

MACQUARIE University

- What is Machine Learning?
- Types of Machine Learning Systems
- Main Challenges
- Machine Learning in Python

Main Challenges

MACQUARIE University

- Basically two things can go wrong in a ML project:

1. Bad Data

"Garbage in, garbage out"

Your analysis is as good as your data.

2. Bad Algorithms

Bad Data

MACQUARIE University

- Insufficient quantity of training data
  - Humans can classify using few examples
  - Simple ML problems require thousands of examples
  - Complex problems like image and speech recognition require millions of examples
- Nonrepresentative training data
  - Violation of the i.i.d. assumption
  - In order to generalize well, data needs to be representative
  - Too small sample sizes lead to sampling noise
  - Flawed sampling methods lead to sampling bias

Bad Data

MACQUARIE University

- Poor quality data
  - Underlying patterns are difficult to detect if training data is full of errors, outliers and noise
  - Good idea to clean up training data (manually)
  - Remove clear outliers and fix errors
  - Many ways to deal with missing features
- Irrelevant features
  - Training data needs enough relevant features
  - Irrelevant features make too much noise
  - Some ML models are sensitive to irrelevant features
  - Feature engineering is concerned with the selection, of a set of good features

Bad Algorithm

MACQUARIE University

- Underfitting the data
  - Occurs when the model is too simple to learn the underlying structure of data
  - E.g., linear models are prone to underfit; reality is more complex than the model

- Possible solutions:
  - Select a more powerful model with more parameters
  - Feed better features to the learning algorithm
  - Reduce the constraints on the model
    - E.g. reduce the number of hyperparameters

Bad Algorithm

MACQUARIE University

- Overfitting training data
  - Overfitting means that the model performs well on the training data, but does not generalize well
  - Overfitting happens when the model is too complex to the amount and noise of training data

- Possible solutions:
  - Simplify the model; select one with fewer parameters
  - Gather more training data
  - Reduce the noise in the training data
  - Constrain the model
    - Regularization

Lecture Outline

MACQUARIE University

- What is Machine Learning?
- Types of Machine Learning Systems
- Main Challenges
- Machine Learning in Python

Machine Learning in Python

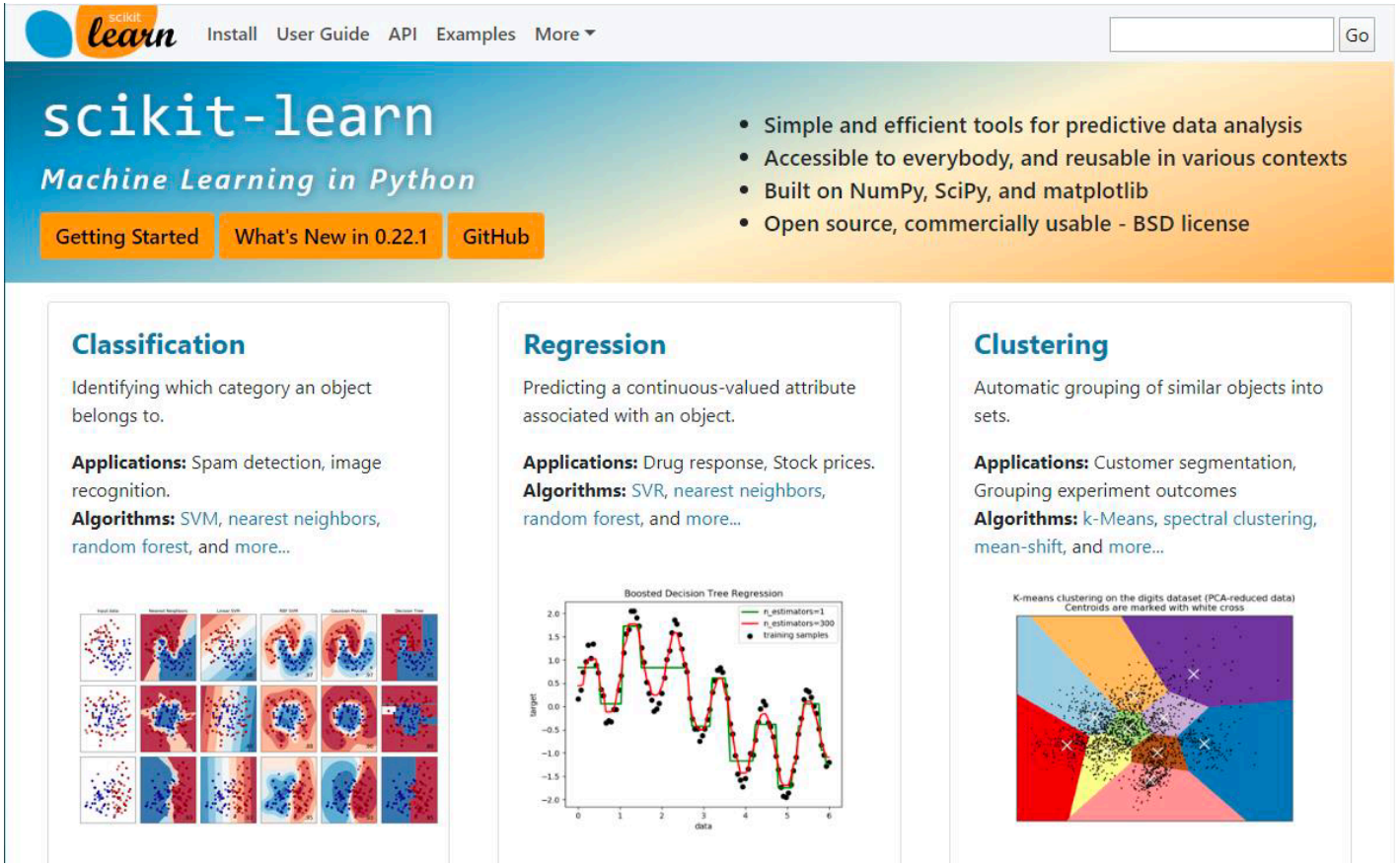
MACQUARIE University

- We use Python's Anaconda distribution: https://www.anaconda.com/distribution/
- Enables data scientists to
  - Perform data science and machine learning on various operating systems
  - Develop and train models for example with Scikit-Learn and TensorFlow
  - Analyze data for example with NumPy and Pandas
  - Visualize results for example with MatplotLib
  - Manage libraries and dependencies
  - Include Jupyter Notebook for interactive data anlysis

Scikit-Learn

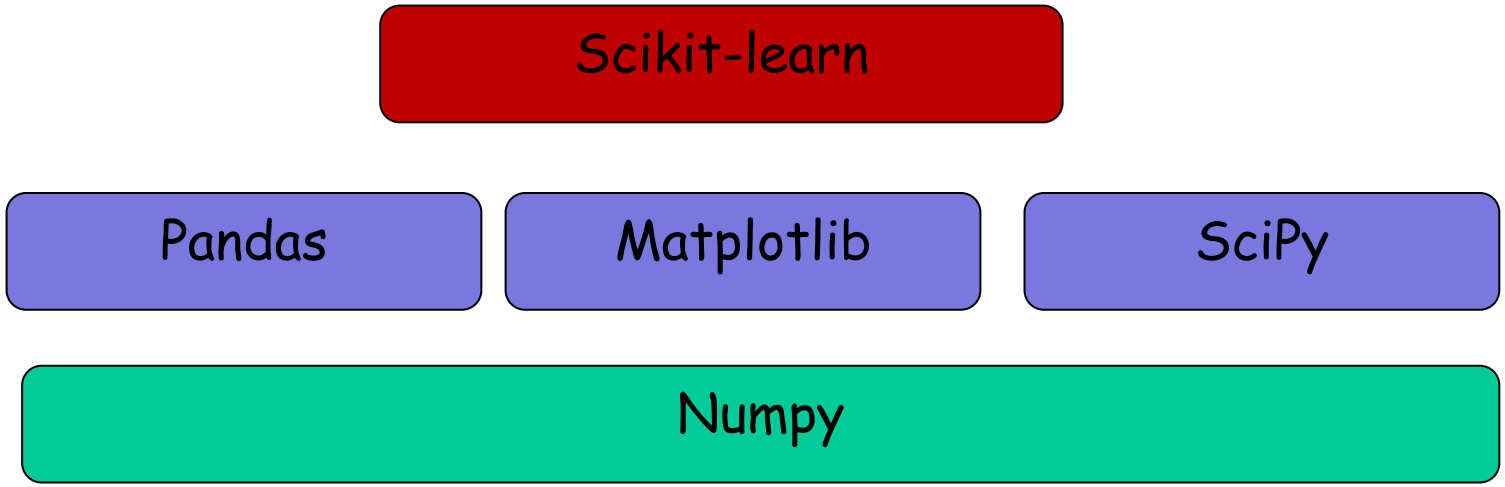
MACQUARIE University

- https://scikit-learn.org/stable/



Scikit-Learn: Libraries

MACQUARIE University



Scikit-Learn: Libraries

MACQUARIE University

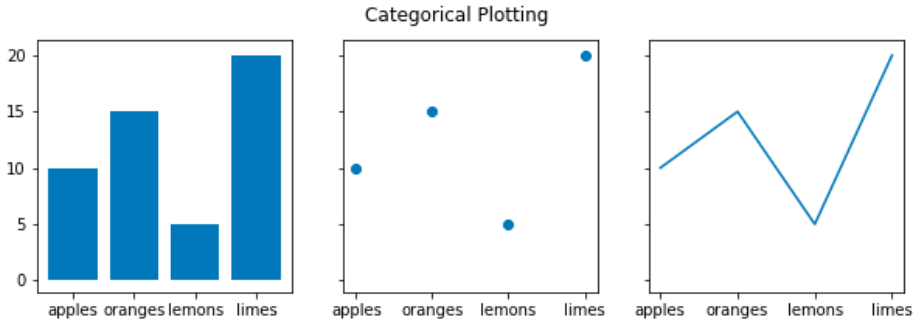
- NumPy:
  - Written in C and adds support for large, multi-dimensional arrays and matrices
  - Provides mathematical functions to operate on these arrays

- Pandas:
  - Offers data structures and operations for manipulating numerical tables and time series
  - Provides in memory 2d table object called Dataframe
  - Like a spreadsheet with column names and row labels

Scikit-Learn: Libraries

MACQUARIE University

- Matplotlib:
  - A plotting library,
  - Generate plots, histograms, bar charts, error charts, etc.



- SciPy:
  - Provides many user friendly and efficient numerical routines for integration, interpolation, optimization, linear algebra, and statistics
  - Uses NumPy arrays as the basic data structure

Take-Home Messages

MACQUARIE University

- Machine Learning is the science of programming computers so that they can learn from data.
- Supervised learning uses labelled training data; unsupervised learning does not.
- Online learning learn on the fly, batch learning not.
- Instance-based learning measures the similarity between instances, while model-based learning generates a model from training data to make predications.
- Scikit-learn is a machine learning library in Python built on NumPy, Pandas, SciPy, and Matplotlib.