University of Pittsburgh

ECE 2195: Special Topics – Computers Machine Learning

Introduction to Machine Learning

Mai Abdelhakim, PhD
ECE Department
Swanson School of Engineering
University of Pittsburgh
maia@pitt.edu

Outline for this unit

- Syllabus, Course objectives, Logistics …
- Background needed - and review of probability theory
- Machine learning (ML) definitions – when to use ML?
- Applications
- Supervised vs. unsupervised learning
- Regression vs. classification

Course Objectives

- Explain concepts, process, and algorithms of machine learning
- Enables you to differentiate between different machine learning algorithms
- Assess the performance of learning algorithms
- Describe best practices in applying machine learning
- Apply machine learning algorithms with python

Contact

- Instructor: Mai Abdelhakim,
  - PhD from Michigan State University

- Contact me if you have any question or need to discuss anything
  - Email address: maia@pitt.edu
  - **Office Hours**: check Canvas page – Faculty and grader information

- Grader
  - **Mr. Renfan Yang,** renfan.yang@pitt.edu

Textbook & References

- **An Introduction to Statistical Learning: with Applications in R** , by James Gareth et al., 2013
  Available online: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
- **Pattern Recognition and Machine Learning**, by Christopher Bishop, 2006.
- Hands-on Machine Learning with Scikit-Learn & TensorFlow, by Aurelien Geron, O'Reilly, 2017 **(Available online through Pitt library)**
- Introduction to Machine Learning with Python, by Andreas Müller et al., 2016
- Python Machine Learning, by Sebastian Raschka, 2015
- **Deep Learning, by Ian Goodfellow, MIT Press, 2016 (https://www.deeplearningbook.org/)**
- Deep Learning with Python, Francois Chollet, Manning, 2018
- Elements of Statistical Learning, by Trevor Hastie et al.
- Pattern Classification, Richard Duda et al.
- Additional reading may be posted

Course Outline (1) – Subject to Change

- Week 1: Introduction to machine learning, ML models
- Week 2-3: Performance measures, tradeoffs, KNN classification
  Bayes classifiers, Linear regression
- Week 4-5: Linear regression single feature and multiple features, polynomial regression, regularization, Cross validation, logistic regression
- Week 6-7: Classifiers, LDA, QDA, Midterm
- Week 8: Support Vector Machines

Course Outline (2) – Subject to Change

- Week 9: Decision trees, Ensemble methods
- Week 10 -11: Neural networks, Deep learning
- Week 12: Dimensionality reduction, Unsupervised learning ,
- Week 13: No class (Thanksgiving)
- Week 14-16: Contemporary machine learning algorithms, Projects presentations
- Week 16: Final Exam (Dec. 16)

Course Requirements

- Class exercises, Assignments (40%)
- Midterm (20%)
- Term Project (20%)
- Final exam (20%)

Term project

- Groups: 2 members per team, each team selects an application of machine learning that is of interest to team members!

- It is expected that each team member will **read at least one related article/paper per week and discuss it with team members…**
    - **Description of the papers read is expected in reports**

- Graded based on effort/approach/attempts and not only final result

Project Milestones

- **Projects proposal due ~ Oct 1:** In your proposal should include
  - Description of the system/problem
  - Explain how machine learning will be used to solve your problem
  - Mention the type and source of data you will use
  - Related work covered by the team so far
  - Include the main responsibilities of each team member in the project

- **Project reports**
  - **Progress report due Nov. 1**
  - **Final report 1st week of Dec.:** Comprehensive description of the problem, related work, data set, solution, and analysis/evaluation

- **Project presentations (expected Dec 9, 14)**

Topics that are NOT accepted for project

- The following projects are not recommended unless discussed with instructor prior to project proposal:
  - Housing price prediction
  - MNIST handwritten recognition
  - Wisconsin breast cancer detection

Course material on Canvas

- Class meeting: Tu, Th – 2:30 – 3-45pm, 132 Chevron Science Center
  - Check Pitt Heath Guidelines!

- Canvas
  - Please check the Canvas regularly
    - Lectures, announcements, exercise/assignment submissions will be there
    - Go to the Syllabus tab to check the syllabus: grading policy and outline.
  - Ask questions or share related material there (through creating or participating in a discussion)

Prerequisites – Required Background

- Probability theory & statistics
- Calculus
- Linear algebra

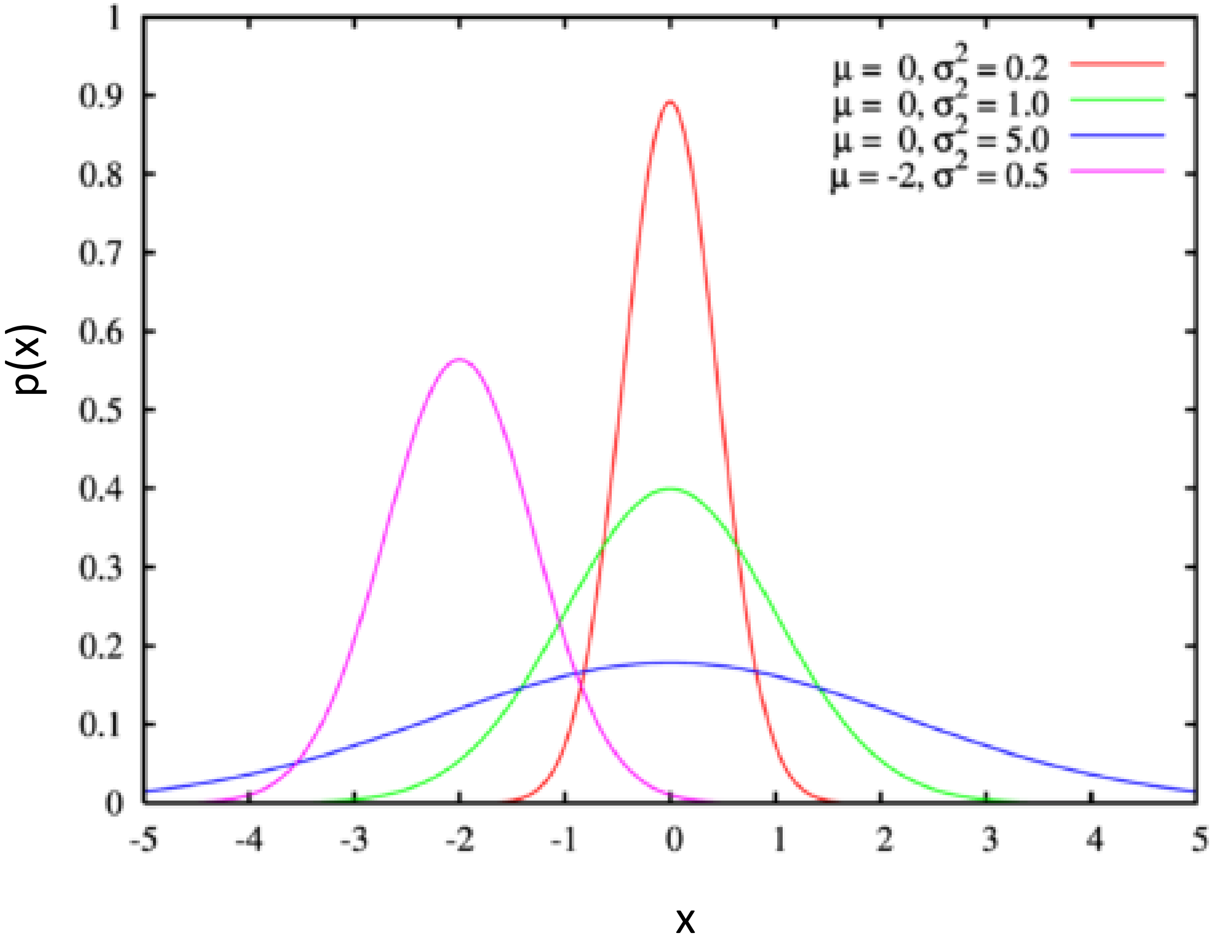
Probability Theory – Basics

You should be familiar with the concepts of probability theory such as:

1. **Random variable**
   Uncertainty about the outcome
2. **Probability**
   measures the likelihood / frequency of occurrence of a random variable
3. **Expected value/mean**
   average value of a random variable
4. **Variance**
   measures the deviation from the mean value

Probability Theory: Probability Distribution

- **Probability distribution**
  - **E.g. Gaussian/ Normal distribution** is common
    - Fully characterized by mean and variance



Normal distribution, mean μ and variance σ^2

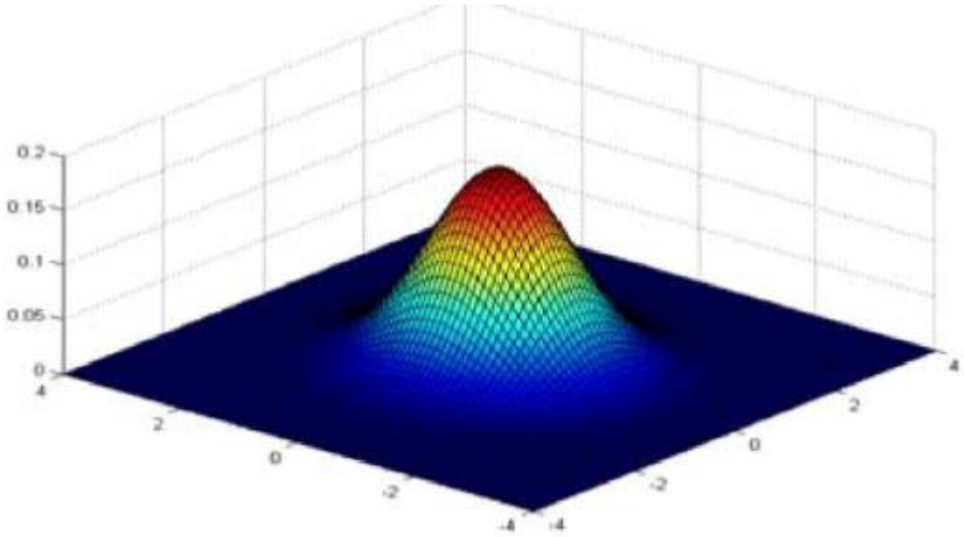
Gaussian Distribution

- 1-Dimensional Gaussian

$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- 2-Dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



- d-Dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Probability Theory: Conditional and Joint Distribution

- **Conditional probability [P(X|Y)]**: Given that an event has occurred, what is the probability that another event will occur

- **Joint probability [P(X,Y)]**:

$$P(X = x, Y = y) = P(X = x|Y = y)\,P(Y = y) = P(Y = y|X = x)\,P(X = x)$$

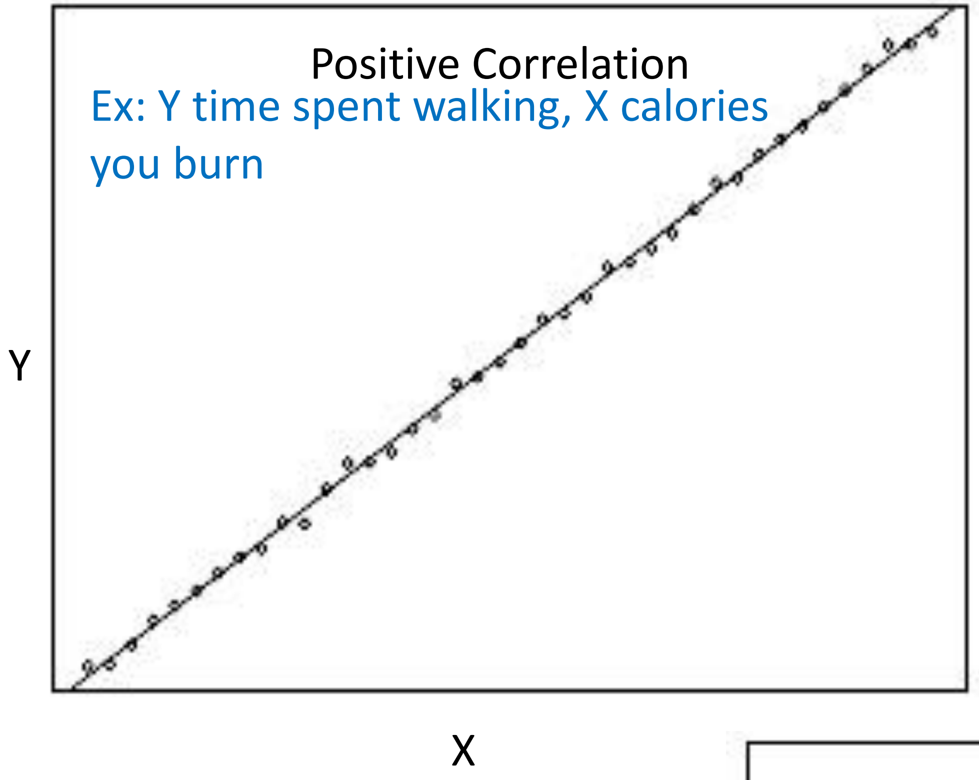
- **Bayes Rule**
  - $P(Y|X) = P(X|Y)P\,(Y)/P\,(X)$

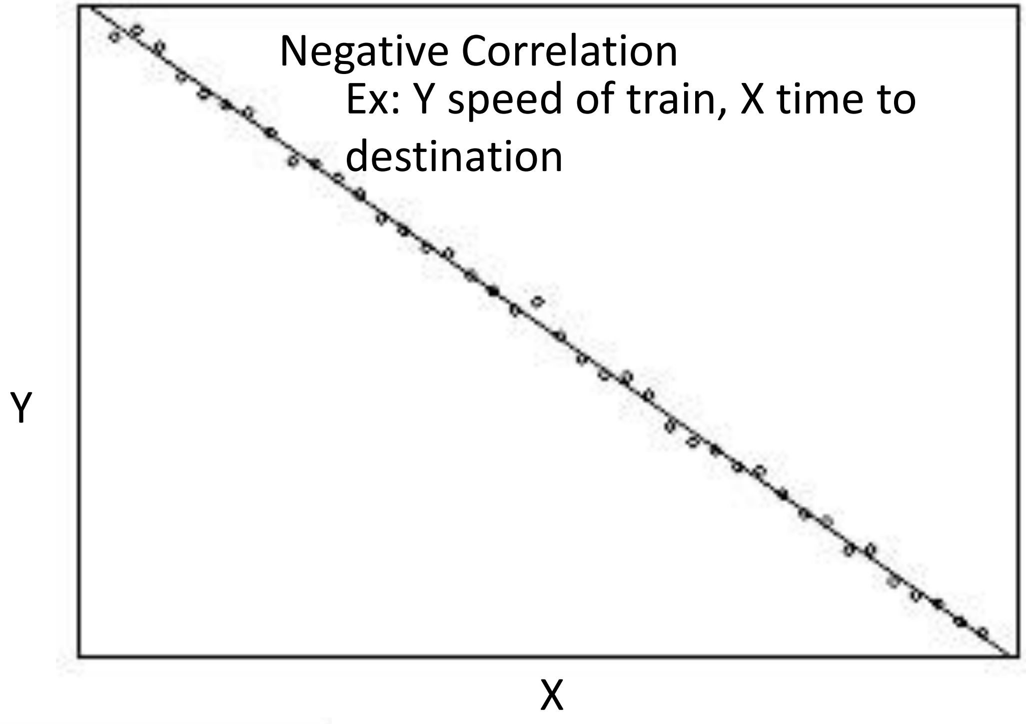
Probability Theory - Correlation

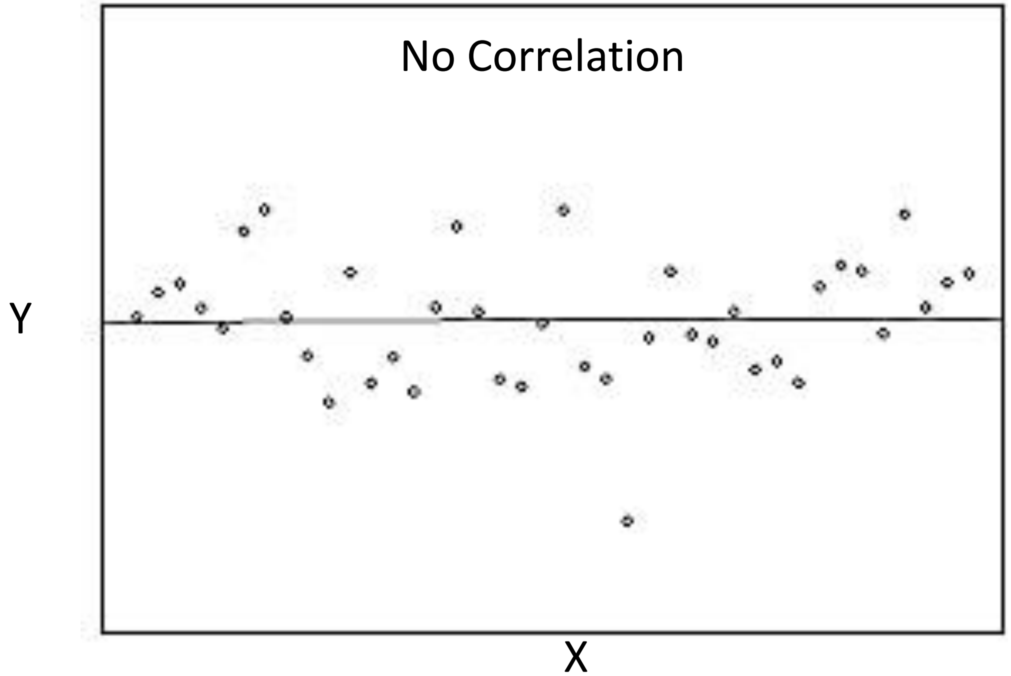
- **Correlation** between two variables describes how strong they are associated with each other
- Measured by covariance matrix or correlation coefficient
  - Covariance between two variables X, Y: COV(X,Y)
  - Correlation coefficient between X, Y ($\rho_{x,y}$) Is a value in the range of [-1,1]

$$\rho_{x,y} = \frac{COV(X,Y)}{\sqrt{VARIANCE(X)}\sqrt{VARIANCE(Y)}}$$

  - If $\rho_{x,y}$=0, then there is no correlation
  - If $\rho_{x,y}$=1, then there is a positive correlation
  - If $\rho_{x,y}$=-1, then there is a negative correlation


Positive Correlation
Ex: Y time spent walking, X calories you burn
Y
X


Negative Correlation
Ex: Y speed of train, X time to destination
Y
X


No Correlation
Y
X

Correlation does not imply Causation!

More research is needed to determine causation

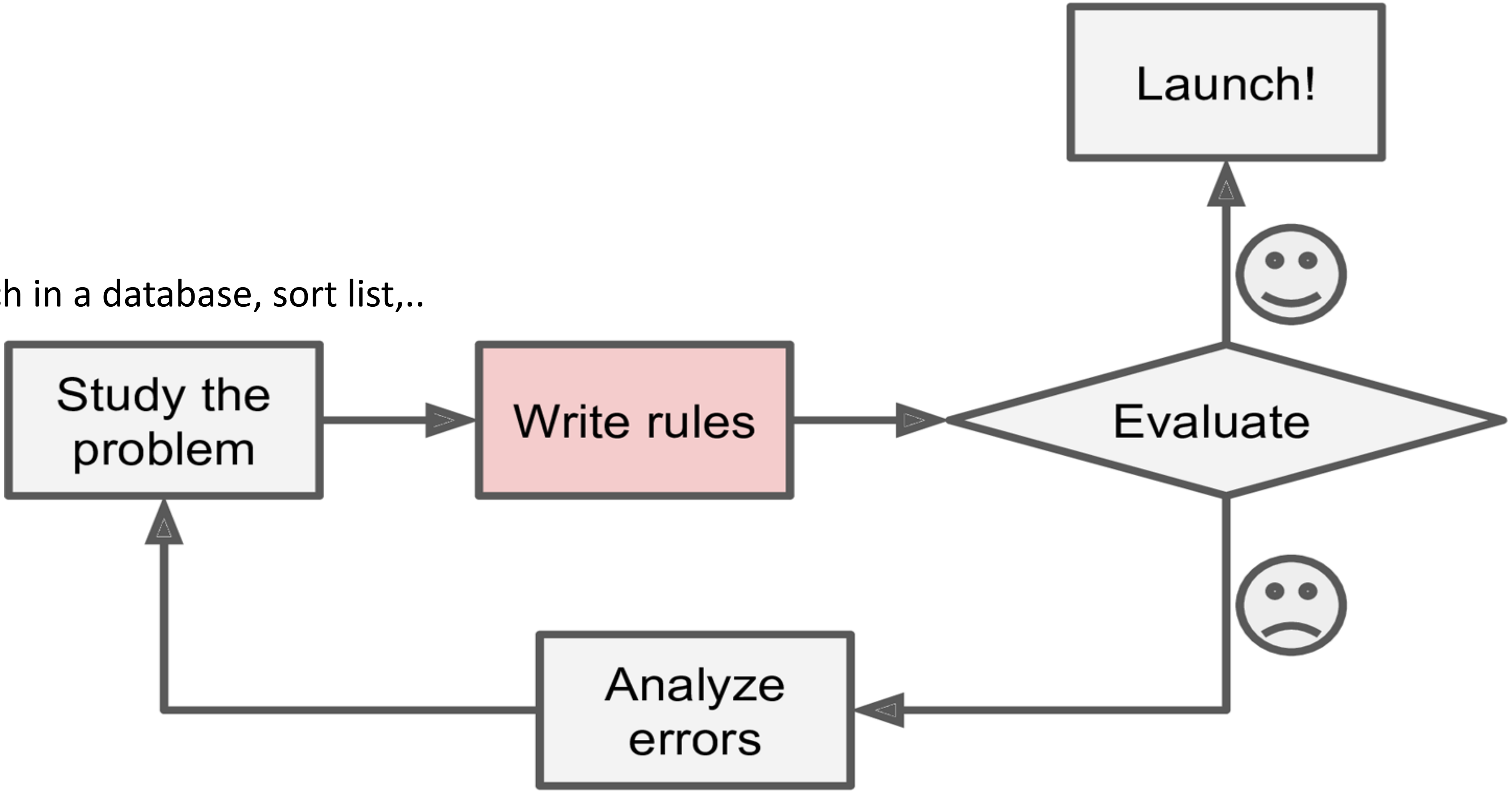
Machine Learns without Explicit Programming

- Machine learning is a subfield of artificial intelligence

- Definition: Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).
  - Arthur Samuel, Stanford University, Pioneer in artificial intelligence & computer gaming

Explicit Programming

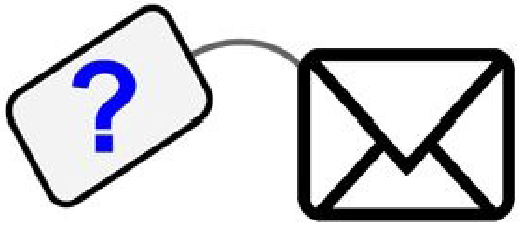
E.g. search in a database, sort list,..



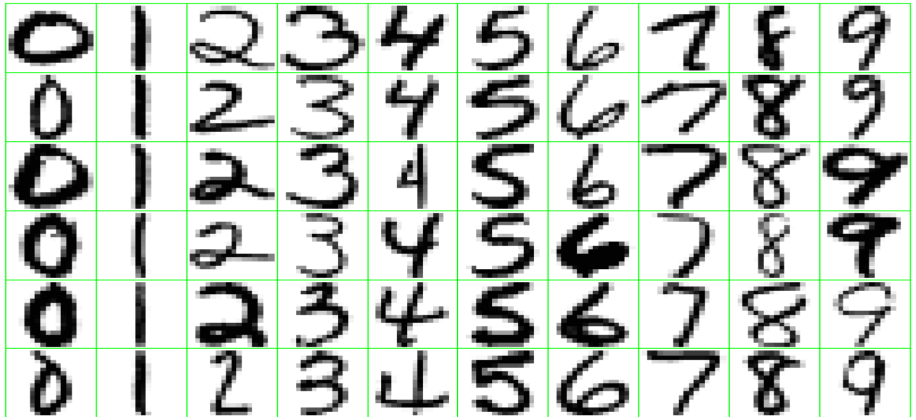
Aurelien Geron, O'Reilly, 2017

Hard to Explicitly Program

- Spam email detection



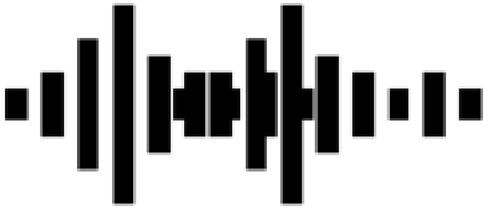
- Handwriting recognition



MNIST database: large database of handwritten digits

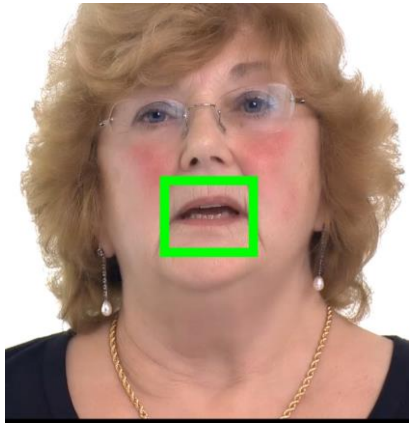
Hard to Explicitly Program

- Speech recognition system
  - Audio signal to output text



Where is the Cathedral of Learning?

- Lip-reading system
  - Video to text



Studies show that …

https://doi.org/10.1145/3192714.3192824

Machine Learning Approach

Instead of trying to write complex rules, let the machine learn from data

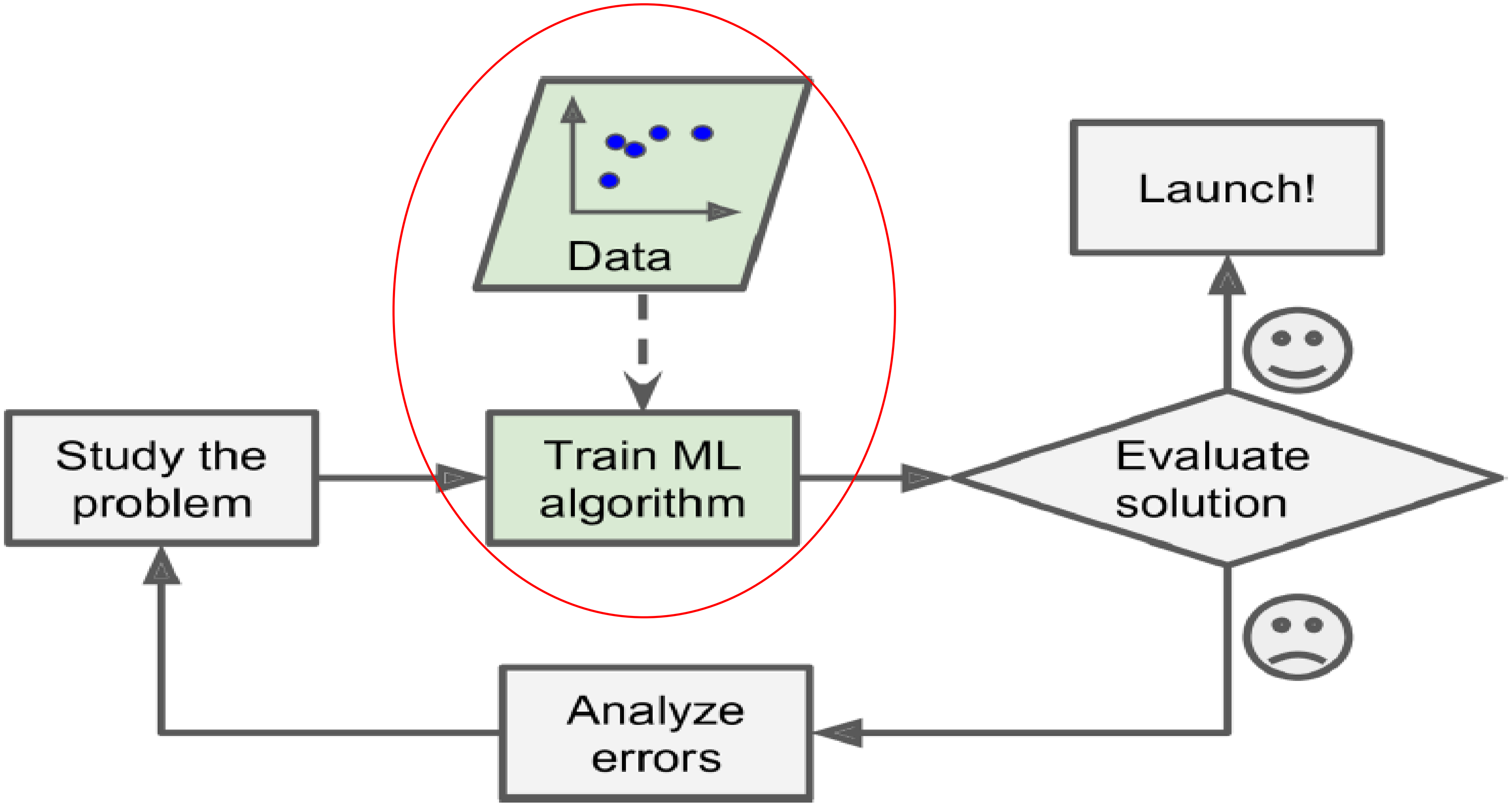


Fig. ref: modified from Aurelien Geron

What is Machine Learning?

- Machine learns with respect to a particular **task T, performance metric P** and **experience E**, if the **performance** P on task T **improves with experience E**.

- How can we build computer system that learn and improve with experience?
  - Statistics make conclusions from data, and estimate reliability of conclusions
  - Optimization and computing power to solve problems

Tom Mitchell, The Discipline of Machine Learning

Can Update to Change

If new data becomes available, model can be updated

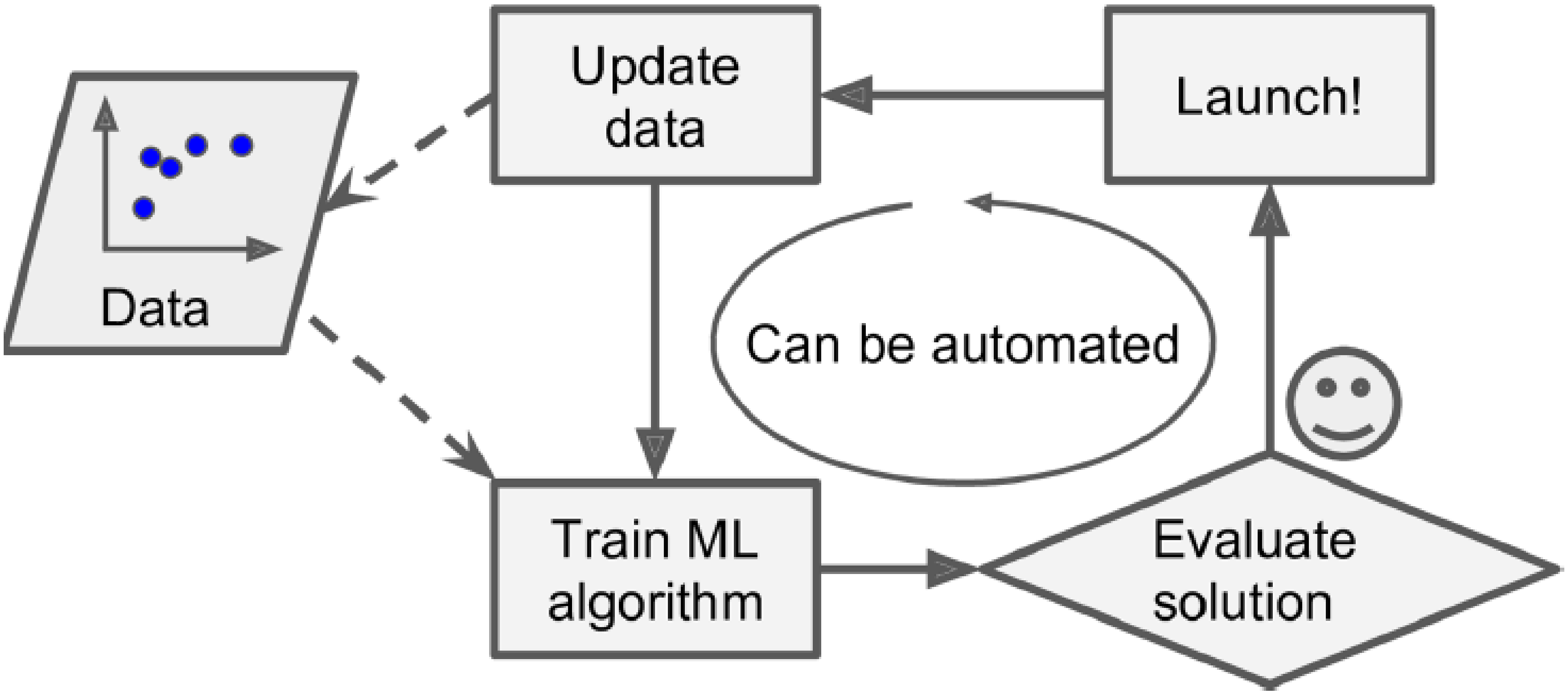


Fig. ref: modified from Aurelien Geron

Machine Learning Helps Human Learn

By inspecting the solution, we can better understand the problem
Example: spam filter reveals list of attributes (e.g. words) in a spam email

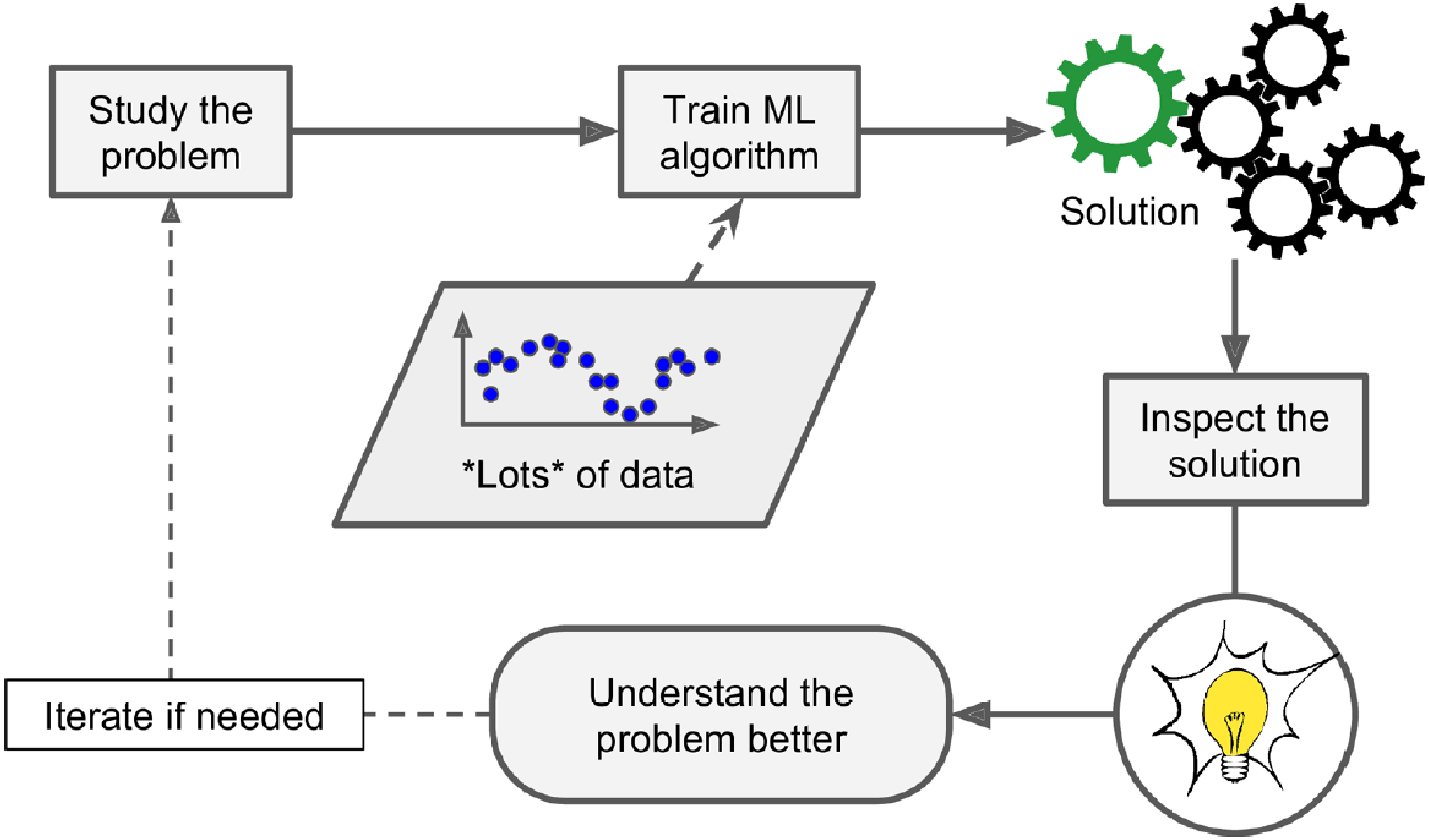


Fig. ref: modified from Aurelien Geron

Machine Learning Provide solution to complex problems that cannot be easily programmed

- Learns from data
- Can adapt to new data
- Helps us to understand complicated phenomena
- Can outperform human performance
  - DeepMind’s AlphaGo defeated a human champion (Ke Jie) in the board game Go, 2017

Applications

- Computer vision: e.g., pattern recognition,
  - US post office: automatically sort letters containing handwritten addresses
- Speech recognition
- Lip reading systems

Applications – Cont.

- Robot systems: e.g. Autonomous driving – use real-time image recognition and video processing
- Learning capabilities make them more capable, flexible, and safer

Other Applications

- Health applications:
  - Drug design and discovery, find tumors in medical images that are hard to detect
  - **Bio–surveillance - detect and track disease** (track emergency room admission reports, purchases over-the-counter medicines)
- eCommerce:
  - Product recommendations (Netflix, Amazon)
- Used across industries to improve efficiency, productivity, flexibility, safety and create new business models

Machine Learning in Industries

Healthcare
Diagnose disease, Predict personalized health outcomes

Automotive
Autonomous Driving,

Navigation

Finance
Identify fraudulent transactions, approve loans

E-commerce
Personalized advertising

Manufacturing
Automation, predictive maintenance

Agriculture
Personalized crops to individual conditions

Network security
Detect and identify attacks

Types of Machine Learning

- **Supervised Learning**
- **Unsupervised learning**
- **Semisupervised**
- **Reinforcement learning**

Variables – labels and features

- **Label Y**, correct answer/required output of the machine learning algorithm
  - Also called **target value, response, dependent variable**
    - Example, digit (0-9), spam/not spam, wages, price
- Input to the machine learning algorithm is set of **features** X
  - Also called **predictors, inputs, independent variables**
    - Example: pixel values, number of words, time, location, area, age,
  - Selection of features has a huge impact on the performance

Example

- Email spam detection: One of the earliest applications
  - Input features (X): relative frequency of most commonly occurring words, punctuation marks

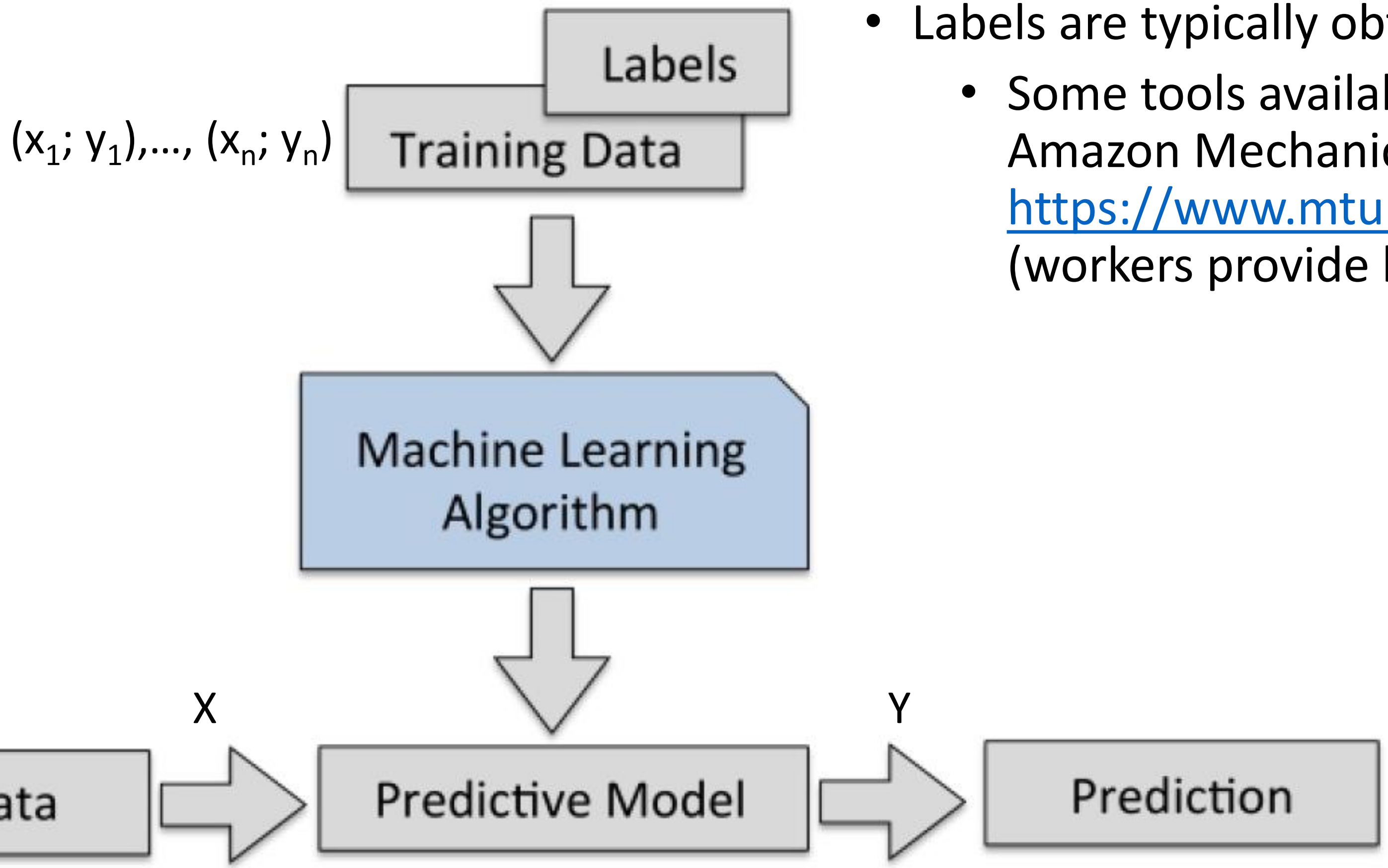
Label (y)	free	!	edu
Spam	0.52	0.51	0.01
Not spam	0.07	0.11	0.29

Types of Machine Learning

- **Supervised Learning**
- Unsupervised learning
- Semisupervised
- Reinforcement learning

Supervised Learning

- Learn using **labeled data (correct answers are given in learning phase)**
- Then, make predictions of previously unseen data



- Labels are typically obtained manually
  - Some tools available, e.g. Amazon Mechanical Turk: https://www.mturk.com/ (workers provide labels)

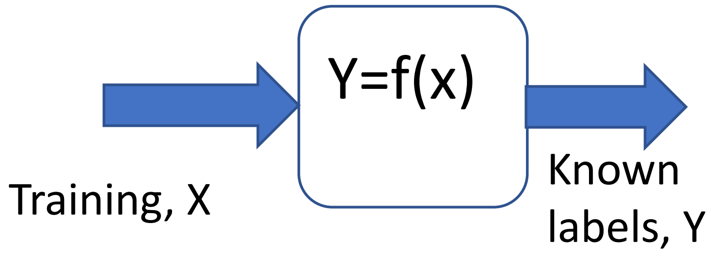
Learning phase

Prediction phase

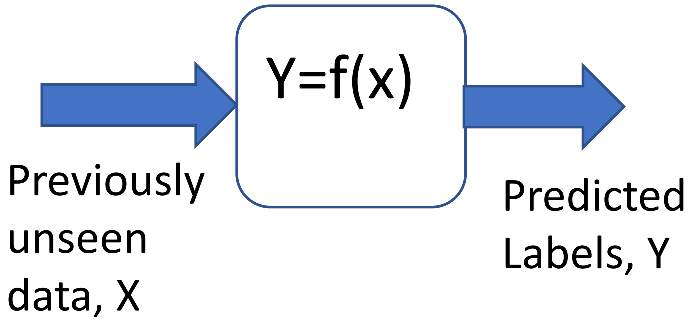
Sabastian Raschka, Python Machine Learning

Supervised Learning: Learning and Prediction

Training Phase: using labelled examples (training), the model learns, i.e., obtain function f, where Y=f(x)



Prediction phase: trained model is used to predict labels for previously unseen data.
Estimate Y for new X



Supervised Learning: Regression and Classification

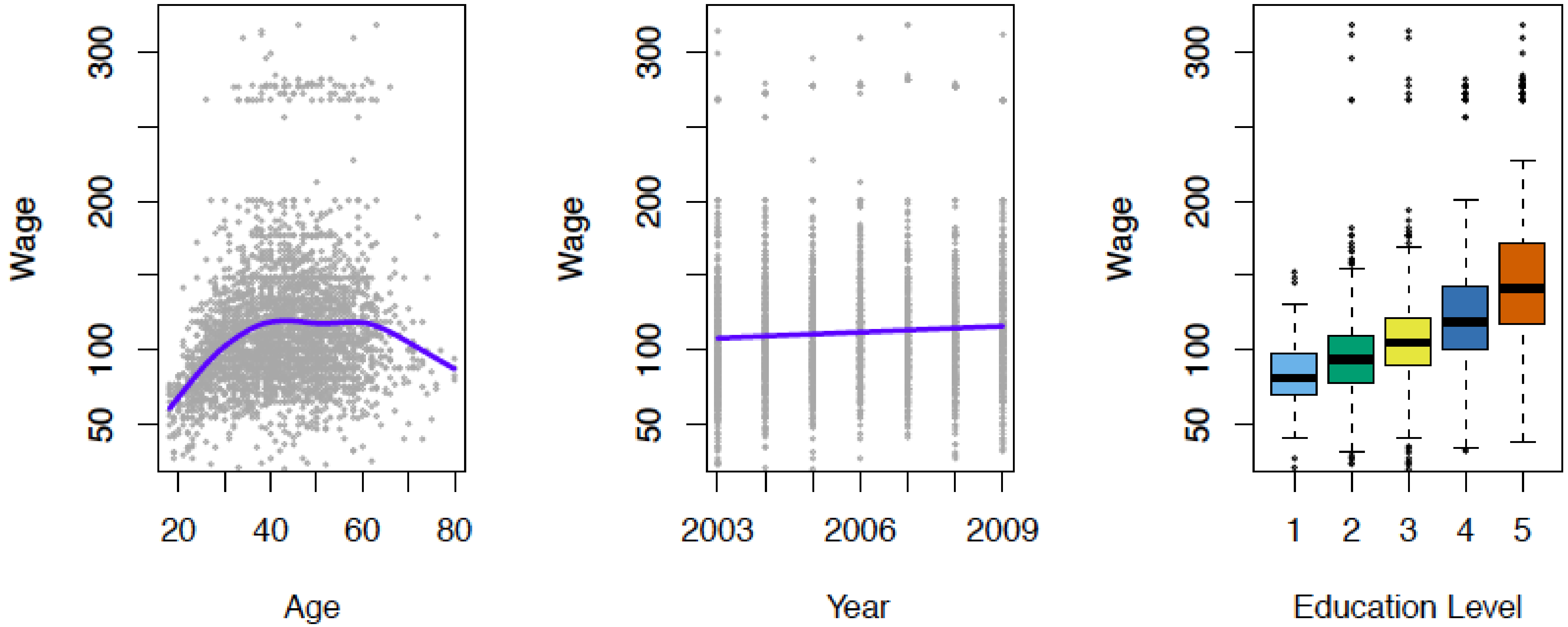
- Learn to predict target values from labeled data (Y available)

- Two types of problems
  - **Regression**: Target values (Y) are continuous/quantitative
    - E.g. price, wage, blood pressure

  - **Classification**: Target values (Y) are discrete/finite/qualitative
    - E.g. gender, digits 0-9, cancer type

Supervised Learning – Regression Example

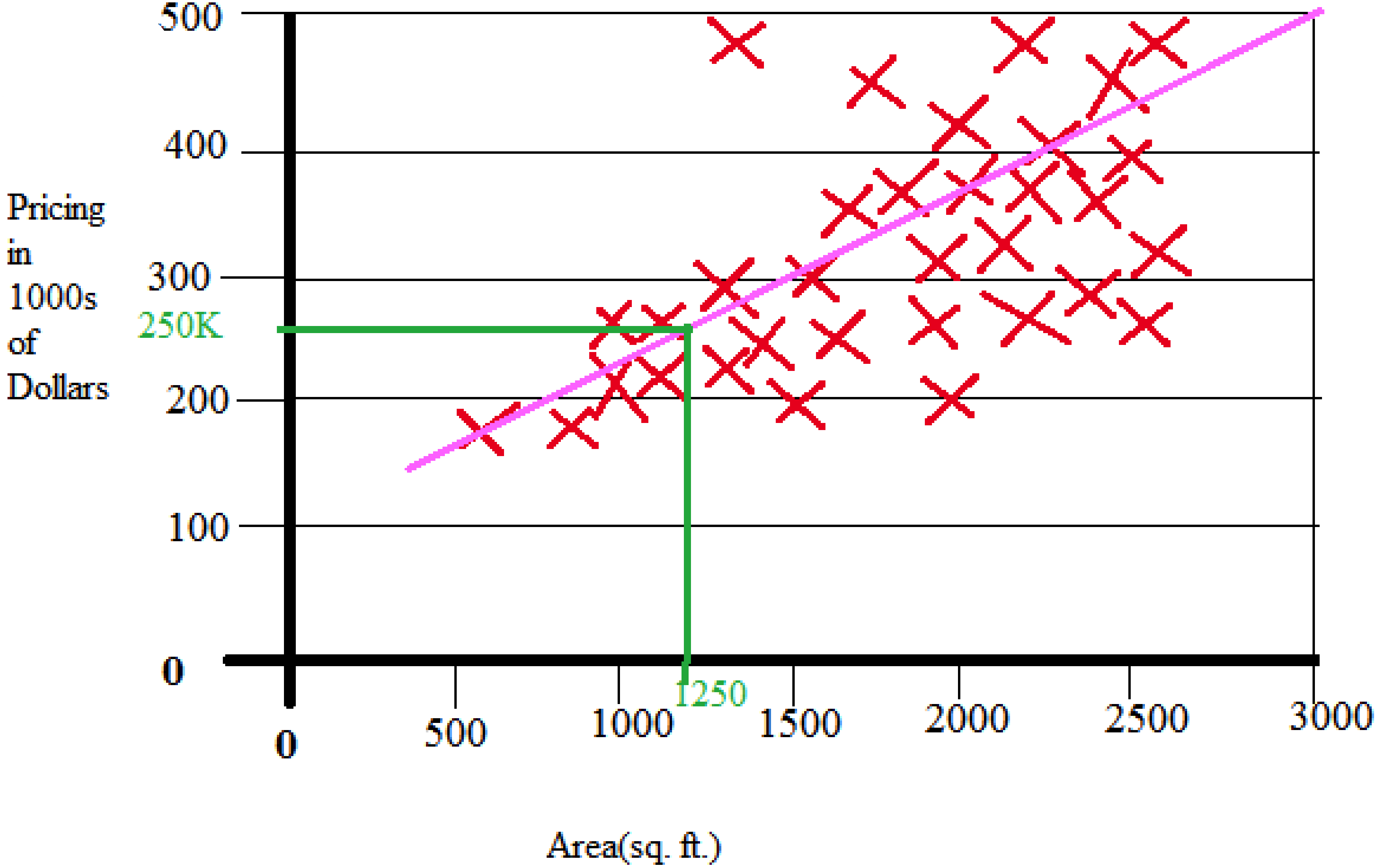
- Income survey data from the central Atlantic region-USA.
  - Label (Y): wage   Features (X): Age, Year & Education level



Trevor Hastie and Robert Tibshirani

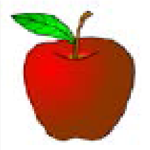
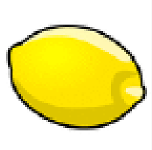
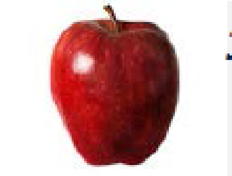
Supervised Learning – Regression Example

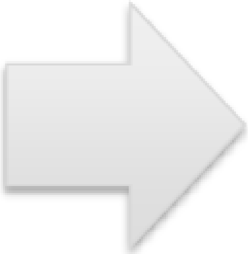
Predicting house price

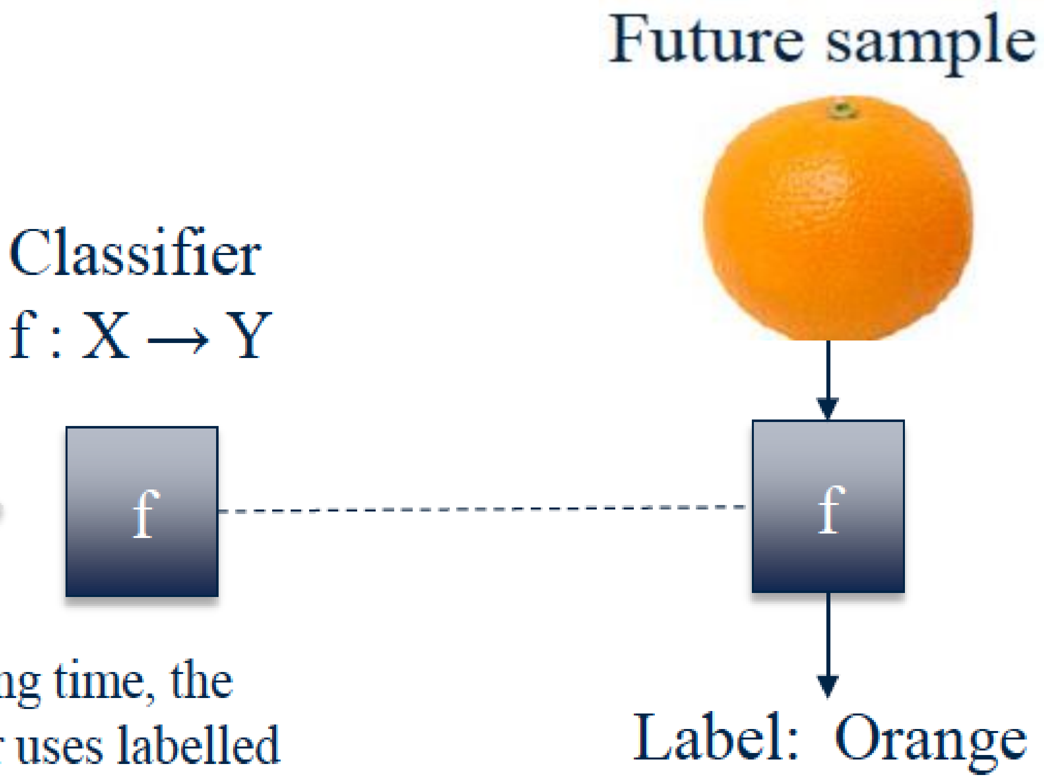


Supervised Learning – Classification Example

Fruit dataset: Apples, lemon, oranges

X Sample	Y Target Value (Label)
 x_1	Apple y_1
 x_2	Lemon y_2
 x_3	Apple y_3
 x_4	Orange y_4



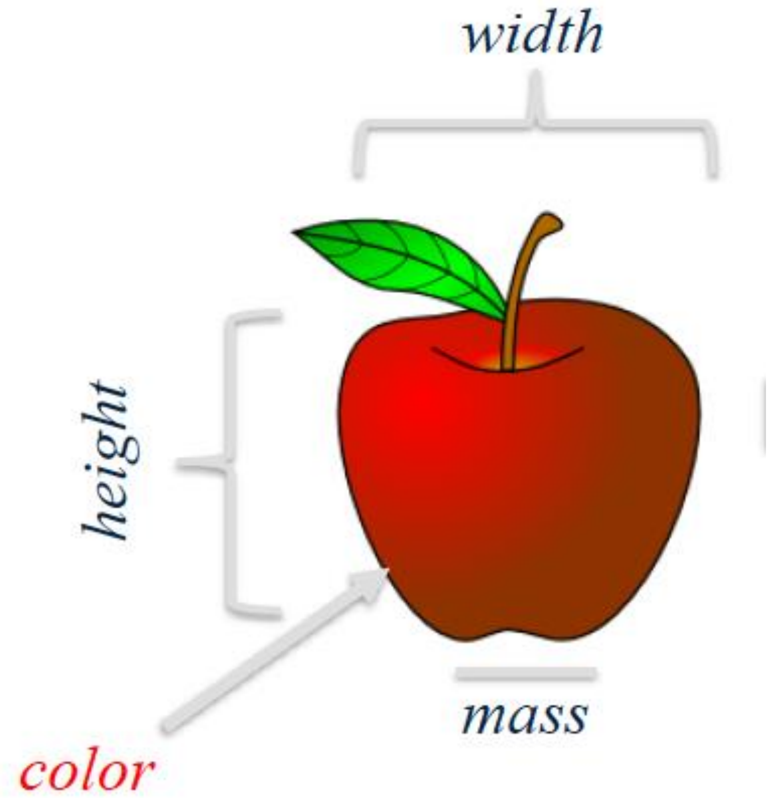


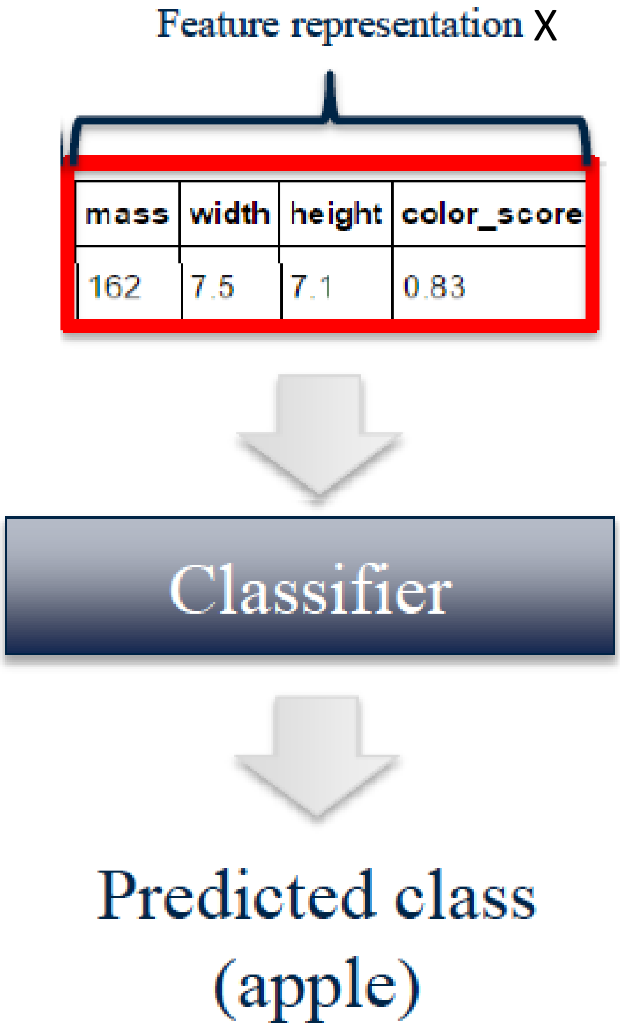
At training time, the classifier uses labelled examples to learn rules for recognizing each fruit type.

After training, at prediction time, the trained model is used to predict the fruit type for new instances using the learned rules.

Supervised Learning – Classification Example
Feature Representation

How to represent an observation?



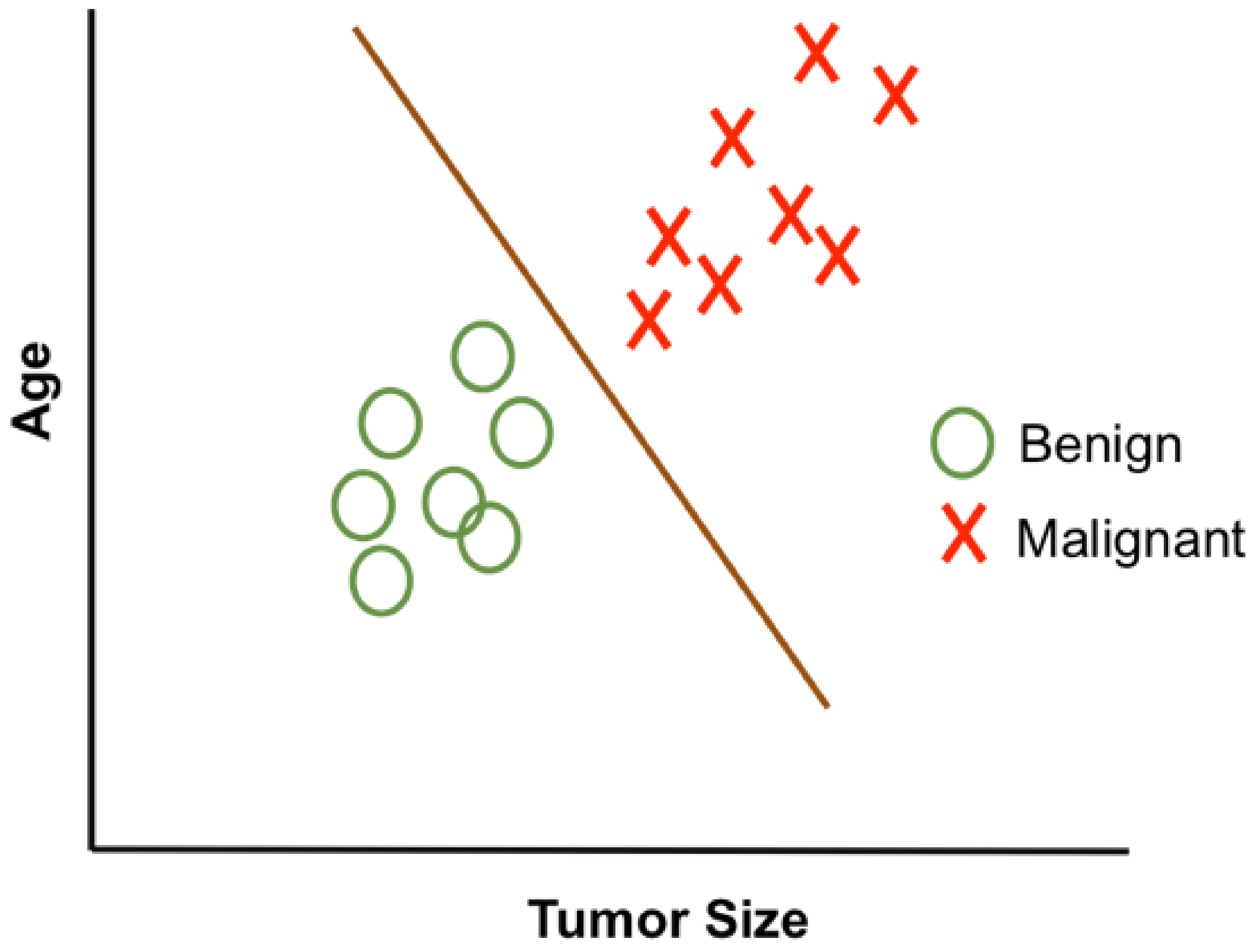


mass	width	height	color_score
162	7.5	7.1	0.83

Supervised Learning – Classification Example

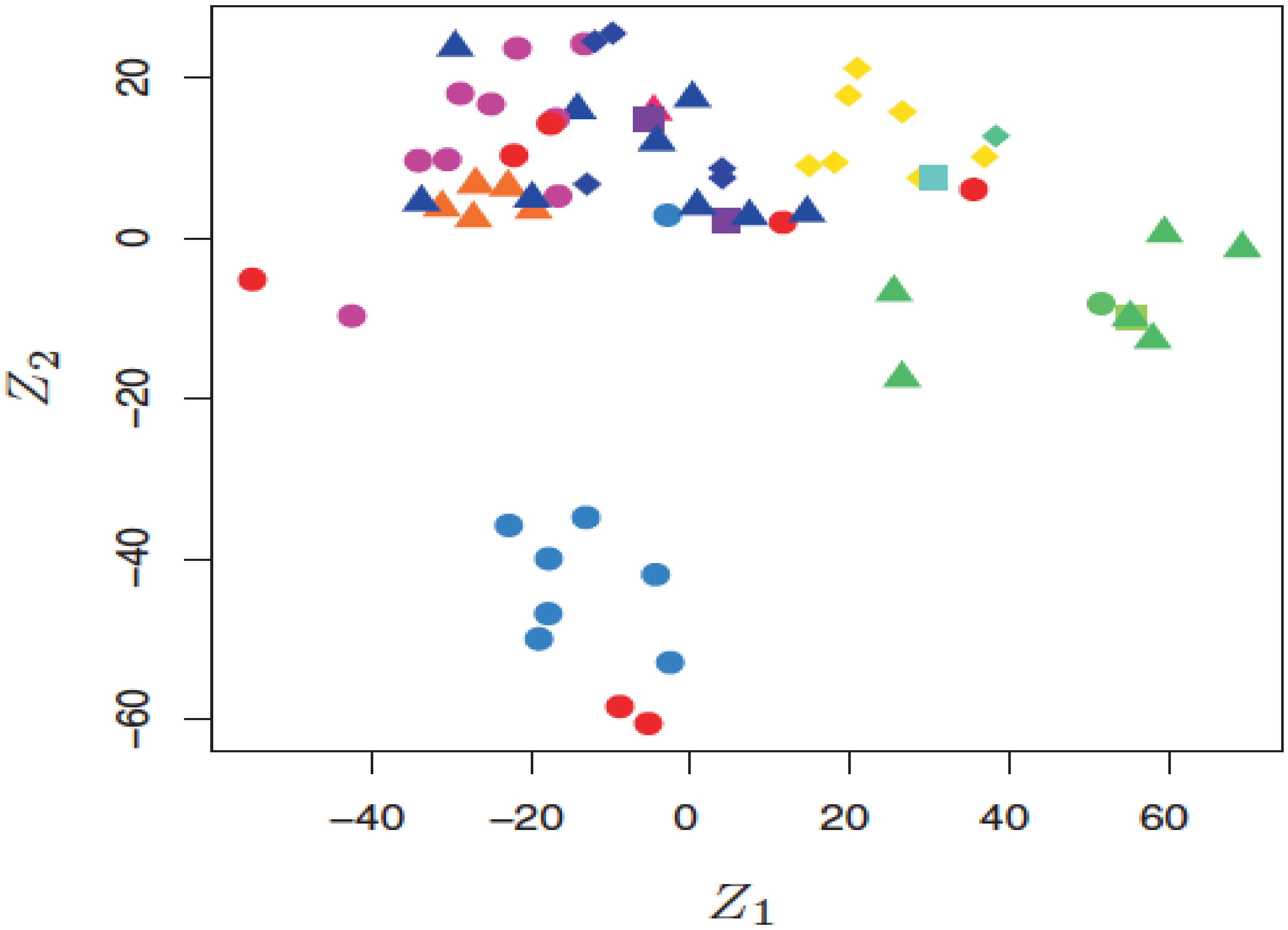
- Cancer classification example.
- Binary classification
    - Benign or Malignant cancer
    - Features: tumor size, age

Figure shows feature space with classification label



Supervised Learning – Classification Example

- Gene expression measurement for different cancer cells **classify cancer class (13 classes)**
  - From NCI60 dataset - National Cancer Institute



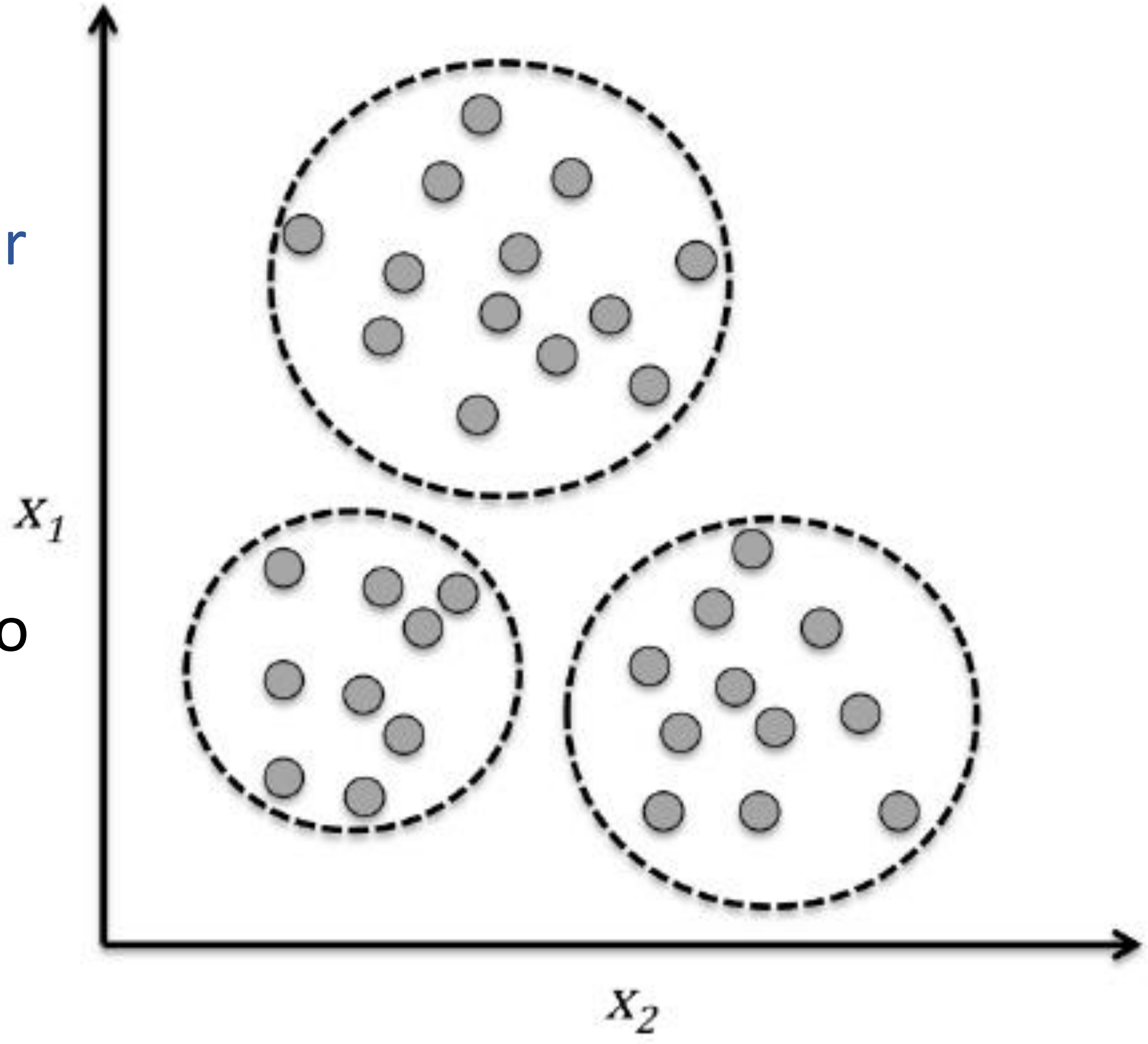
Gareth James et al.

Types of Machine Learning

- Supervised Learning
- **Unsupervised learning**
- Semisupervised
- Reinforcement learning

Unsupervised Learning

- No labels provided
- Arrange data into clusters (similar groups)
- Difficult to evaluate
- Example: Google news – look into thousands of news and group similar news together into a cohesive news stories.

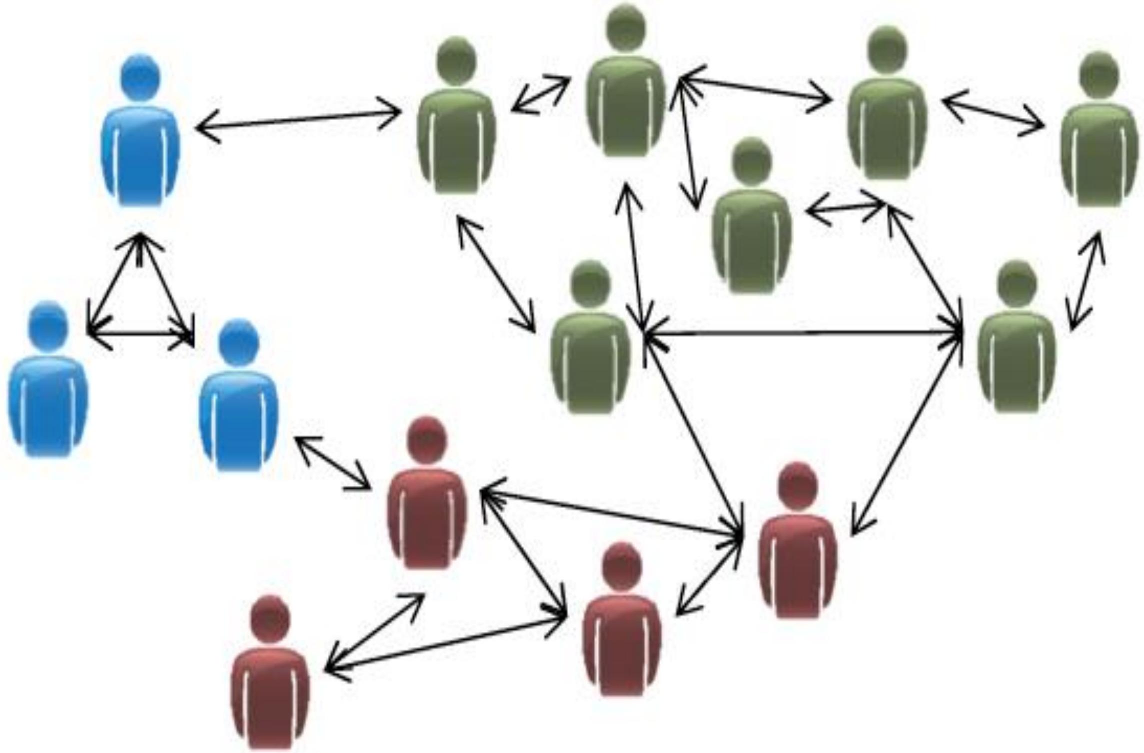


Unsupervised Learning

- Training samples are unlabeled
- Objective: find similarities/groups
  - Ex. Dogs group vs cats group

Unsupervised Learning

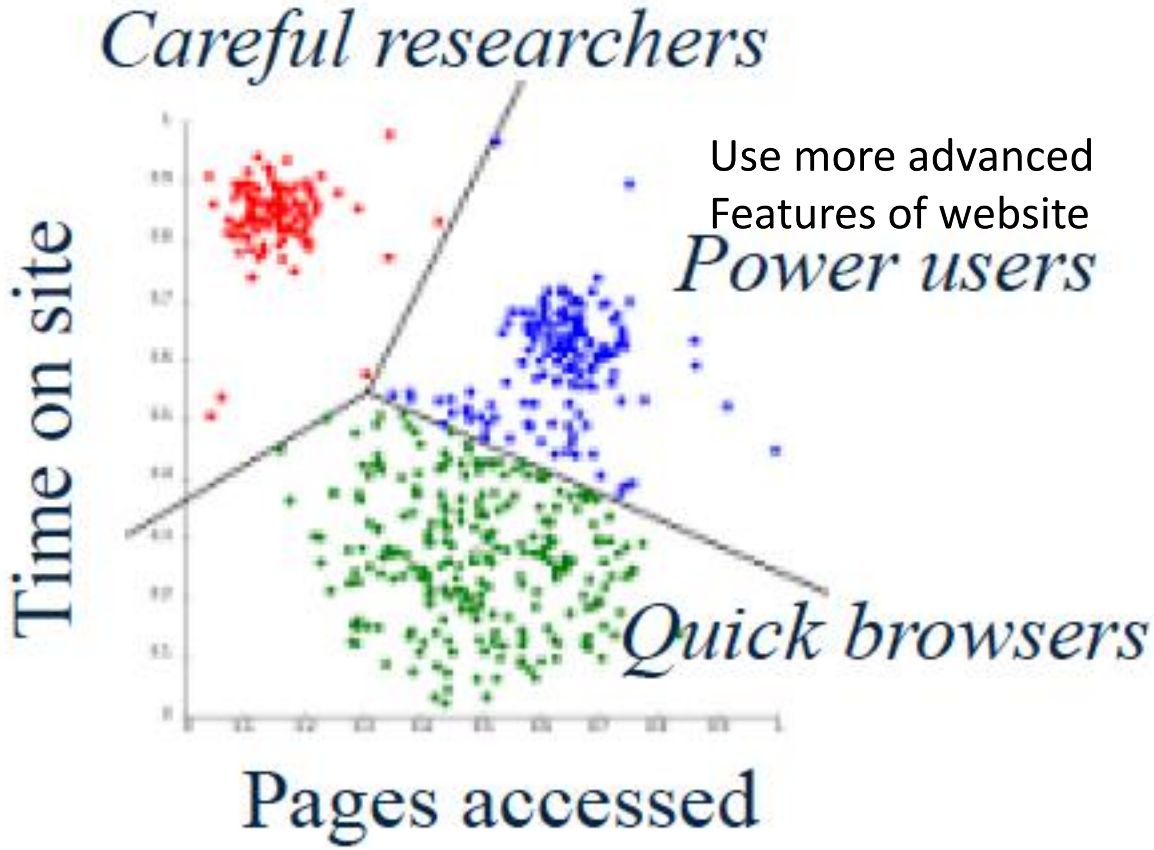
- Clustering analysis
- Finding groups of similar users



Social network analysis

Spend a lot of time on website

Use more advanced
Features of website

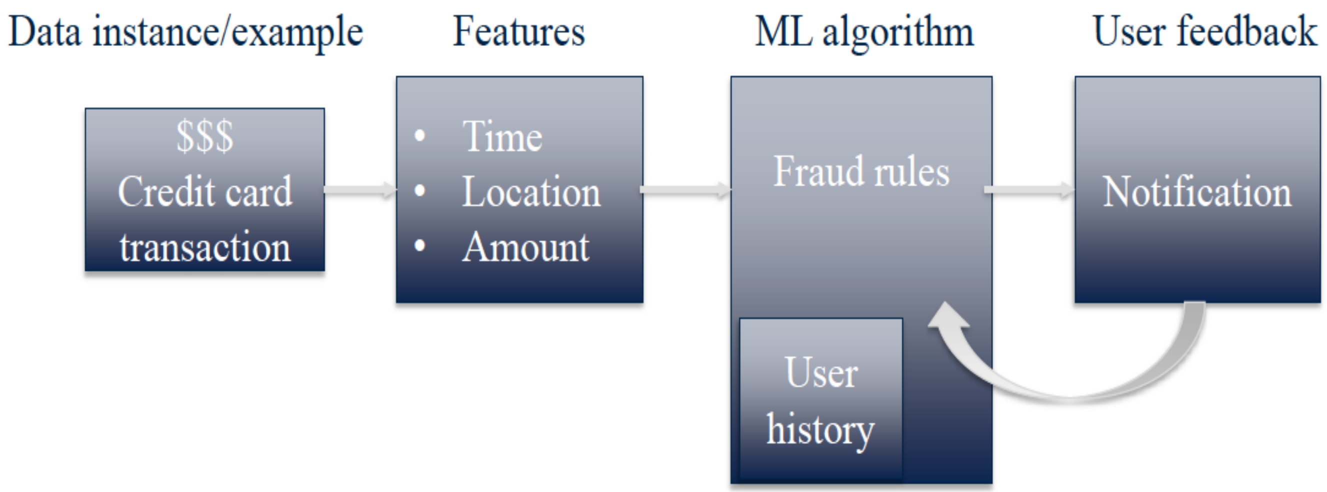


e-commerce example: Tailor website for each group

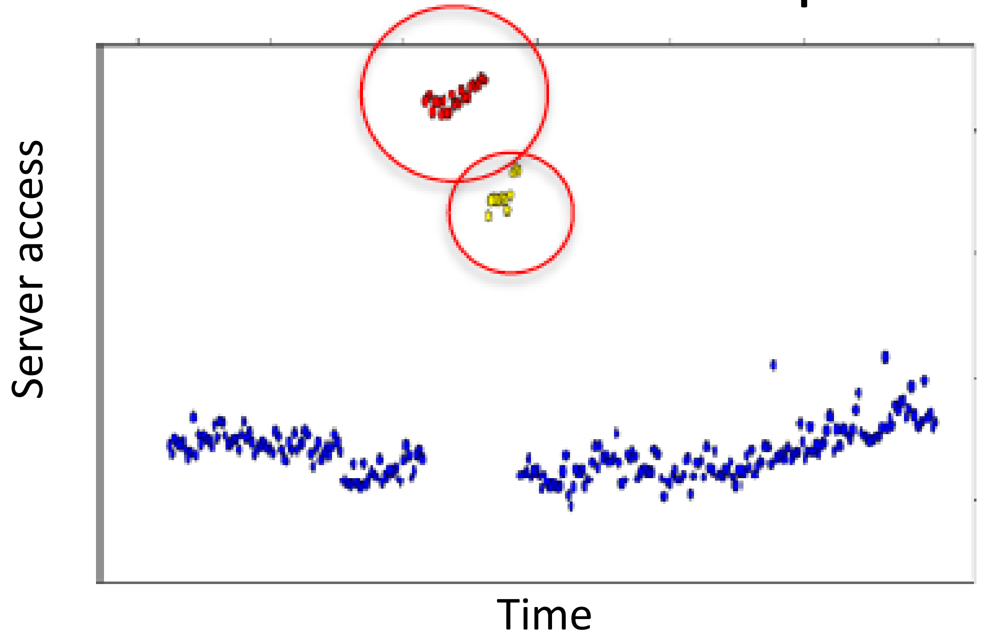
Unsupervised Learning

- Detecting abnormal patterns

Credit card fraud detection



Abnormal server access pattern

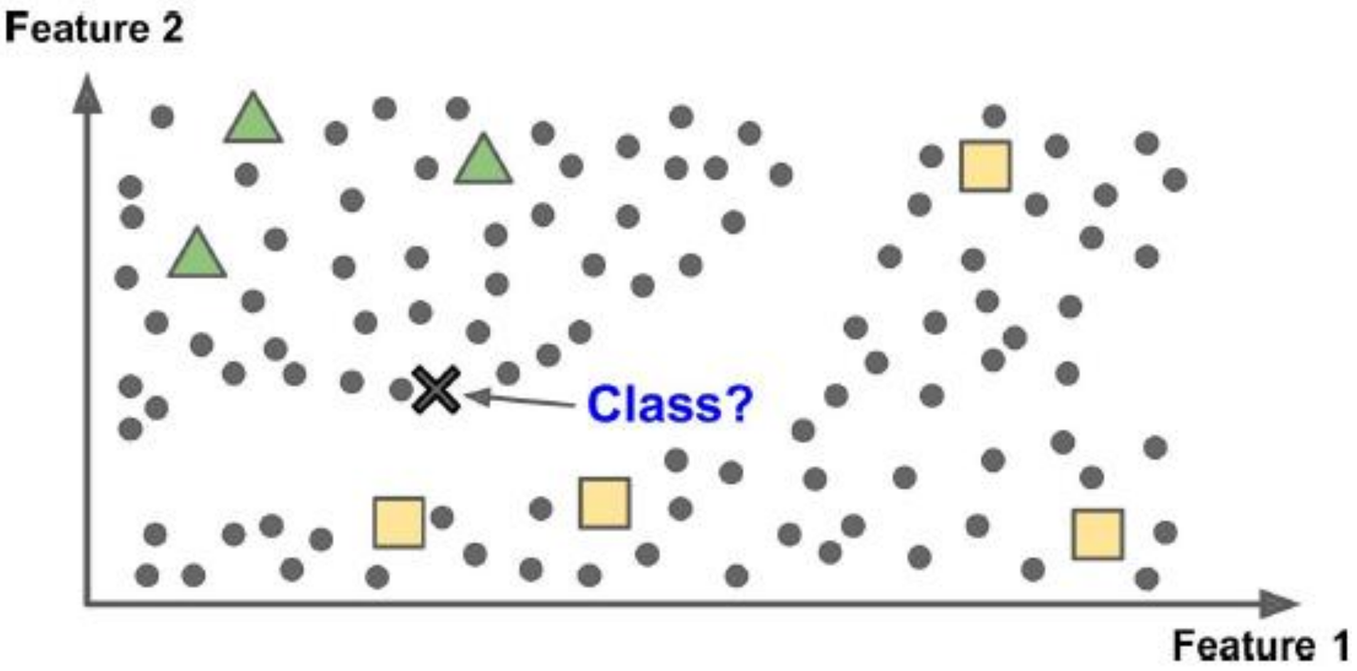


Kevyn Collins-Thompson, Applied Machine Learning

Types of Learning

- Supervised Learning
- Unsupervised learning
- **Semi supervised:**

some labelled data
– e.g. tagging people in photos

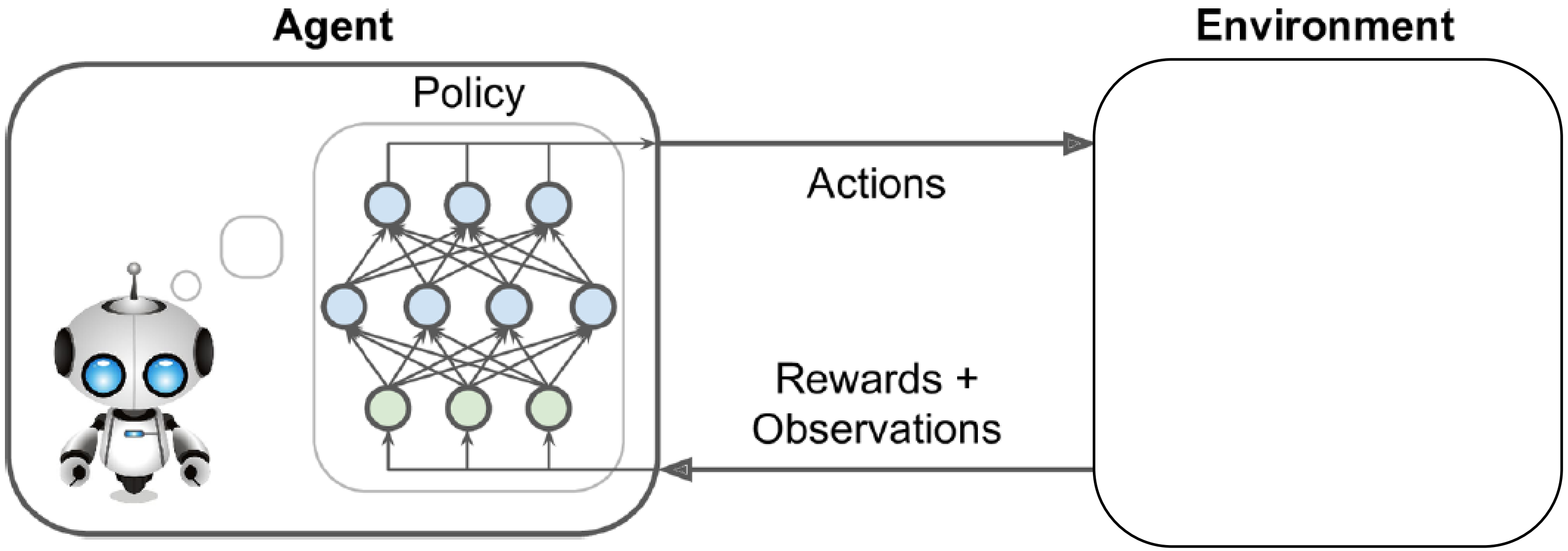


- Reinforcement learning

Aurelien Geron, O'Reilly, 2017

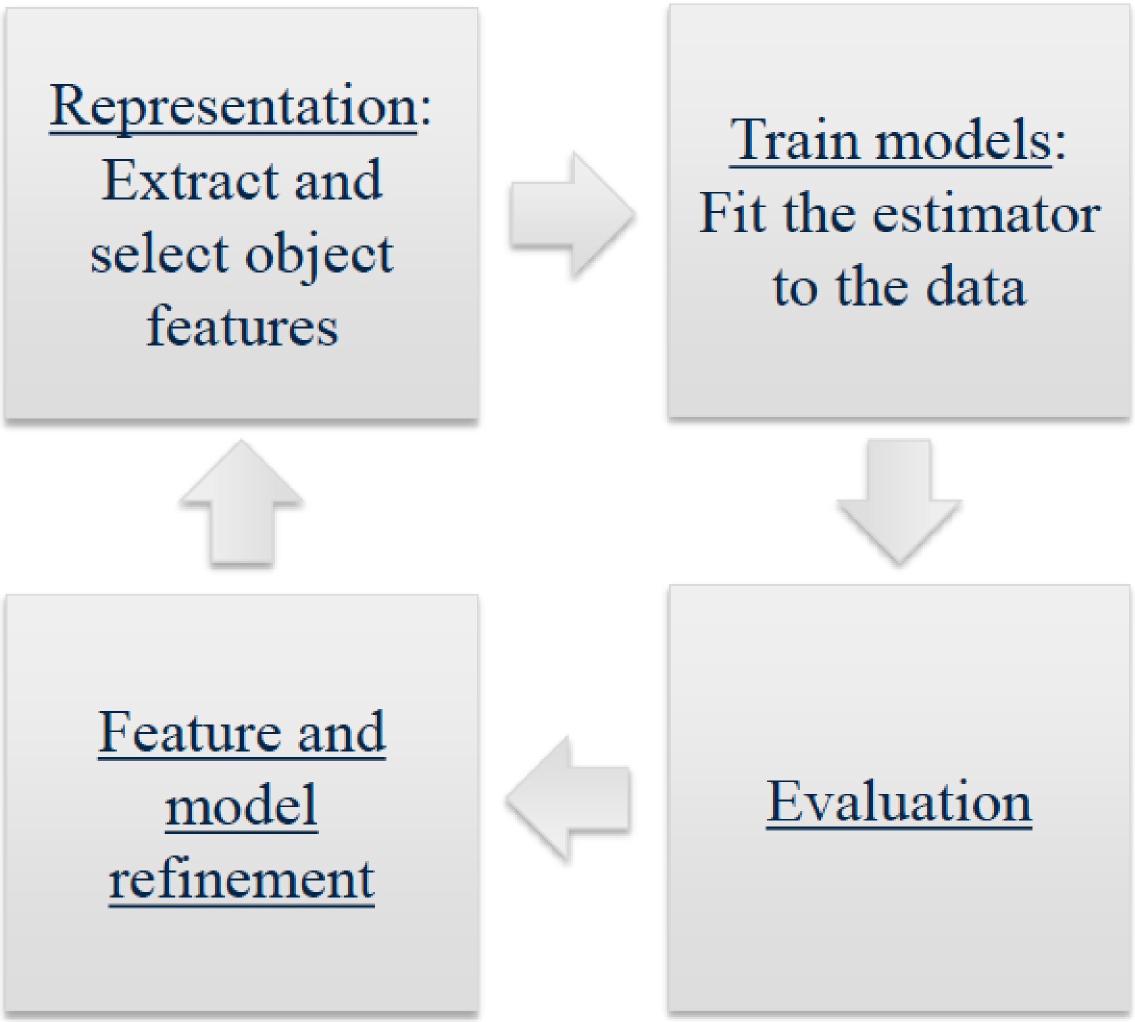
Types of Learning

- Supervised Learning
- Unsupervised learning
- Semi supervised
- **Reinforcement learning**



Aurelien Geron, O'Reilly, 2017

Machine Learning Process



Search for model and features that results in high accuracy

Performance is function of:
What features?
What learning model & parameters?
…

Data

- Massive amounts of data are available and can be used to train machine learning models
  - online click streams, ratings,
  - voice and video
  - sensors readings

- Machine learning performance heavily depends on the data sets used to train the algorithms

- We will use synthetic data as well as real data (e.g. Iris data set, MNIST data set of handwritten digits, breast Cancer Wisconsin, house prices dataset)

Example of data sets that can be used in final project

- Many public sources
- Public Data Sets from Amazon
  http://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1
- HealthData.gov
  https://www.healthdata.gov/search/type/dataset
- Stanford Large Network Dataset Collection
  http://snap.stanford.edu/data/
- Government's open data: https://www.data.gov/
- Western Pennsylvania Regional Data Center: https://data.wprdc.org/dataset
- Machine learning competitions: https://www.kaggle.com/competitions
- More (check discussion board/CourseWeb)

Inspect Data

- Inspect your data
- Missing information
- Wrong readings
    - Correct or discard

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	192
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	apple	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn		7.4	7.0	0.89
10	1	apple	braeburn		6.9	7.3	0.93
11	1	apple	braeburn		7.1	7.6	0.92
12	1	apple	braeburn		7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69

Software

- Python: (https://www.python.org/doc/)
  - **Python basics: A Whirlwind Tour of Python**, by Jake VanderPlas (available online)

- Installation: Anaconda (Recommended) **https** **https://www.anaconda.com/download/**

  Choose **Python 3**

- Scikit-learn (http://scikit-learn.org/)
- Keras (**https://keras.io/**)

Summary

- Machine learning vs. Explicit programing

- Data is a key!

- Types of machine learning