Data Mining and Machine Learning

Class Info

Class Objective: This course covers the mathematical and programming foundations of data mining(DM) and machine learning (ML) using Python programming languages and software tools.

Prerequisites:
Python, Basic knowledge in prob & statistics

Software tools:
scikit-learn, numpy/pandas

Assignments:
5 assignments(theory + coding)

Class Info

Grading: (* subject to change)
There are 5 assignments.
Each assignment accounts for 20% of total grade.

Textbook
No official textbooks. Some chapters may be from the following book.

1) Mohammed J. Zaki and Wagner Meira, Jr, “Data Mining and Machine Learning: Fundamental Concepts and Algorithms” 2nd Edition, Cambridge University Press, 2020

Supplementary/Recommended Readings
1) Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow” O’Reily Media Press, 2019
2) Pattern Recognition and Machine Learning, by Christopher M. Bishop. Springer, 2006, ISBN-13: 978-0-3873-1073-2.
** Free ebook from author website
https://www.microsoft.com/en-us/research/people/cmbishop/prml-book

Class Info

Instructor
Chang-hwan Lee,
EE Room 521
Tel: 561-297-3496
Email: changhwanlee@fau.edu
Office hour: Tu/Th 12:20-1:50

TAs
Devi Vara Prasad Thonangi: dthonangi2023@fau.edu
Mohamad Kawssarani: mkawssarani2020@fau.edu

Course Outline

Introduction of DM/ML
Data preprocessing
Association
Linear regression
Logistic regression
Kernel methods
Feature selection
PCA
Decision tree
Neural Network (Deep Learning)
Performance evaluation

Hyperparameter tuning
Bias variance
Svm
knn
Gradient descent
Overfitting + regularization
Bagging/boosting
Random forest
k means
Agglomerative + EM
Reinforcement learning

* subject to change

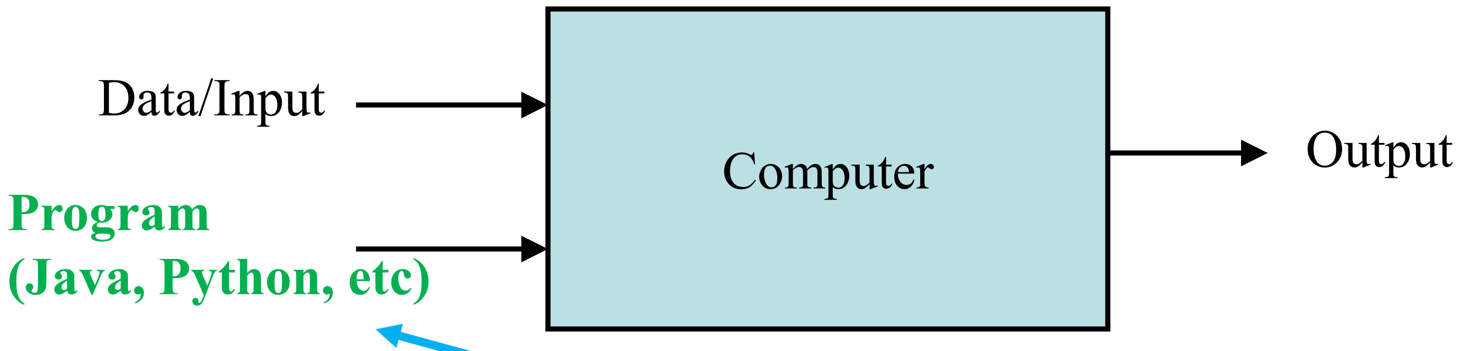
Introduction to Data Mining & Machine Learning

What Is Machine Learning?

- Program is an automation tool
- Machine learning is about automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

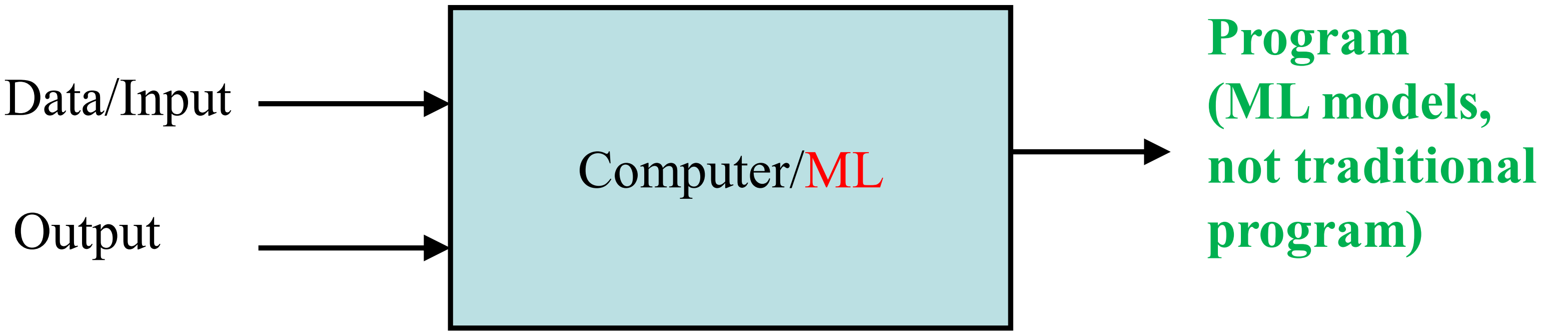
Traditional Programming

- given data, program generates output



Machine Learning

- given data, output generates program



Magic?

No, more like gardening

- **Gardening = Machine Learning**
- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs(ML model)

What is Machine Learning? (Jeoffrey Hinton)

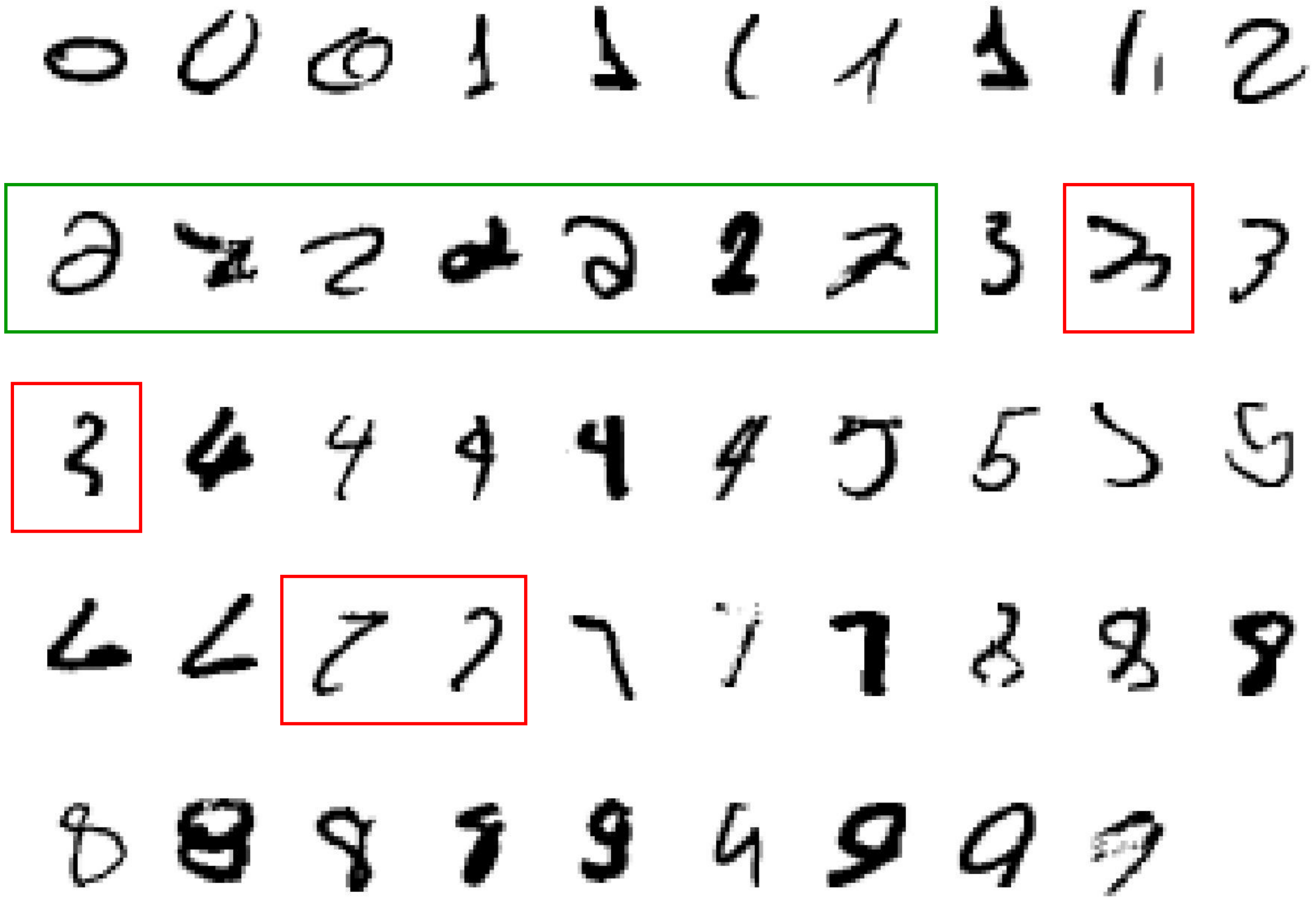
- It is very hard to write programs that solve problems like recognizing a face.
  - We don’t know what program to write because we don’t know how our brain does it.
  - Even if we had a good idea about how to do it, the program might be horrendously complicated.

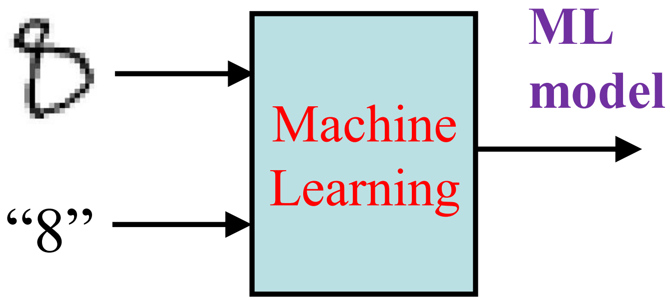
- Instead of writing a program by hand, we collect lots of examples that specify the correct output for a given input.

- A machine learning algorithm then takes these examples and produces a program that does the job.
  - The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers.
  - If we do it right, the program works for new cases as well as the ones we trained it on.

An example of a task that requires machine learning:

It is very hard to say what makes a 2





Machine Learning vs. Statistics

- Both Statistics and ML need a lot of data
  - What is the difference then?

- Statistics is known for:
  - well defined hypotheses used to learn about a specifically chosen population studied using carefully collected data providing inferences with well known properties.
  - Build a hypothesis (knowledge) first, and then verify it using data

- Machine learning isn’t that careful. It is:
  - data driven discovery of models and patterns from massive and observational data sets
  - Generate knowledge from data (no need of hypothesis)

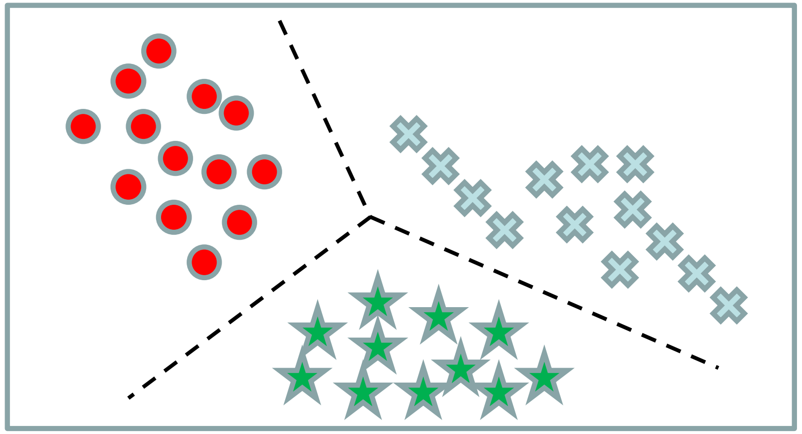
Machine Learning vs. Statistics

- Traditional statistics
  - first hypothesize, then collect data, then analyze
  - often model-oriented (strong parametric models)
  - Focused on understanding
- Machine Learning:
  - few if any a priori hypotheses
  - data is usually already collected a priori
  - analysis is typically data-driven not hypothesis-driven
  - Often algorithm-oriented rather than model-oriented
  - Focused on prediction
- But
  - statistical ideas are very useful in machine learning, e.g., in validating whether discovered knowledge is useful
  - increasing overlap at the boundary of statistics and ML
  - cultures could learn from each other

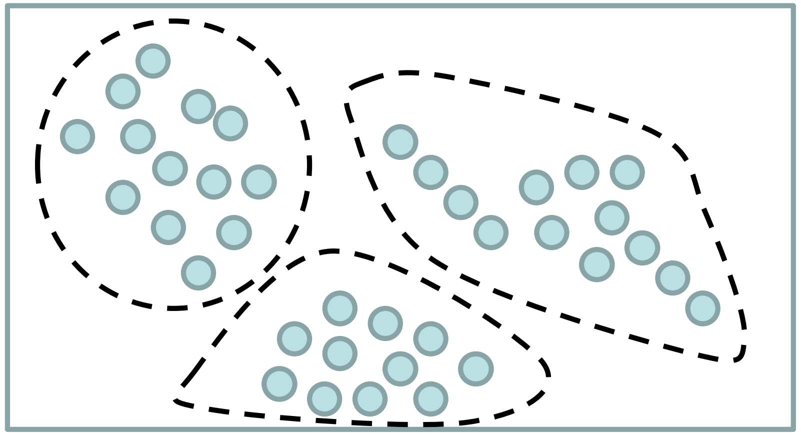
Types of Machine Learning Methods

- **Supervised learning(Classification/Regression)**
  - Each data is given class(target) value
  - Learn to predict class value when given an input data
- **Unsupervised learning(Clustering)**
  - No class values are given
  - Find structure that exists in the data
- **Semi-Supervised learning**
  - Both Labelled and Unlabelled data
- **Reinforcement learning(RL)**
  - Interacts with environment, learns by trial and error method
  - Learn actions to maximize rewards(goal)
  - Sometimes combined with Deep Learning(DRL)

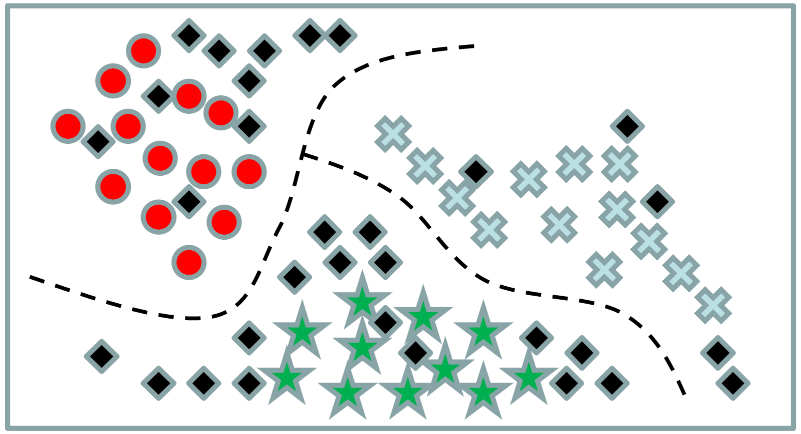
Types of Machine Learning Methods



Supervised learning



Unsupervised learning



Semi-supervised learning

Types of Machine Learning Methods

Labeled data

Input data | Class value/Label

"Cat"

"Dog"

Unlabeled data

Input data

~~"Cat"~~

~~"Dog"~~

Input data | Class value/Label

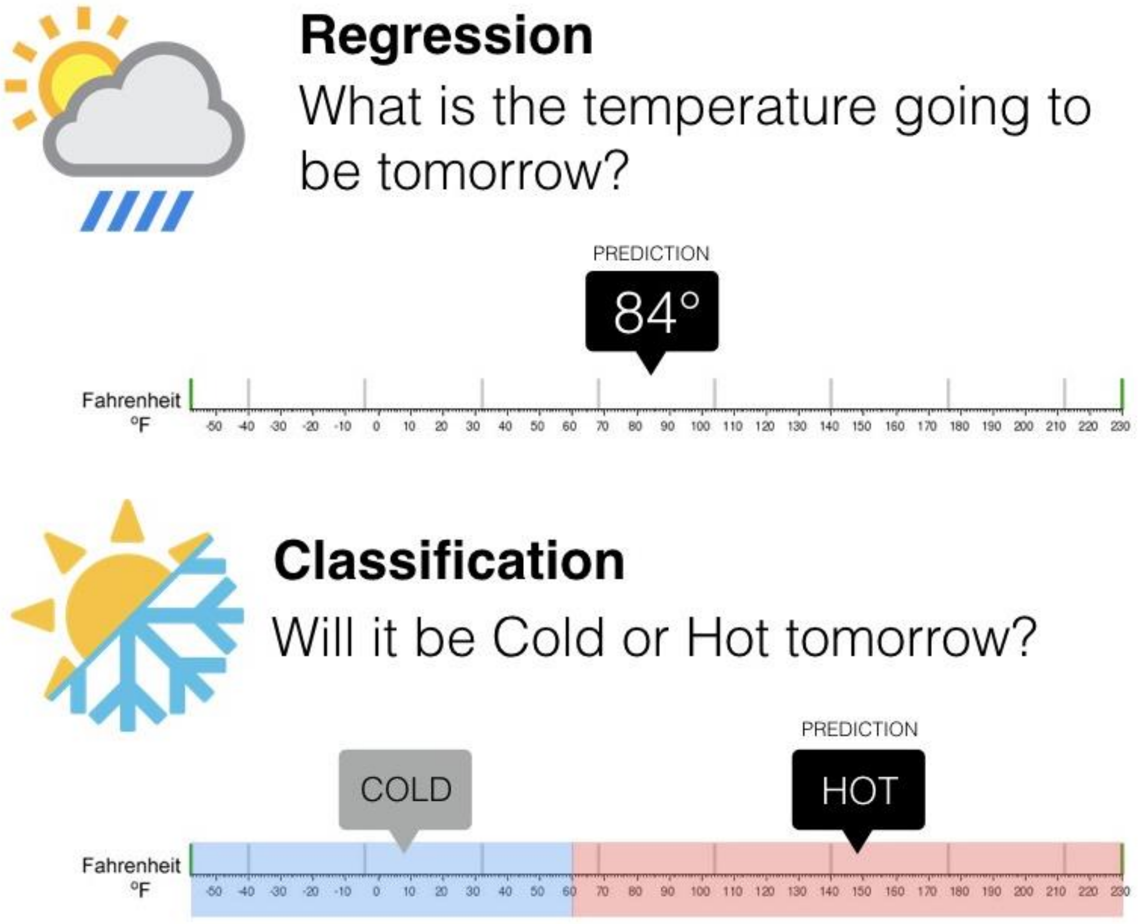
Supervised learning
(labeled data)

Unsupervised learning
(unlabeled data)

Semi-Supervised learning
(labeled + unlabeled data)

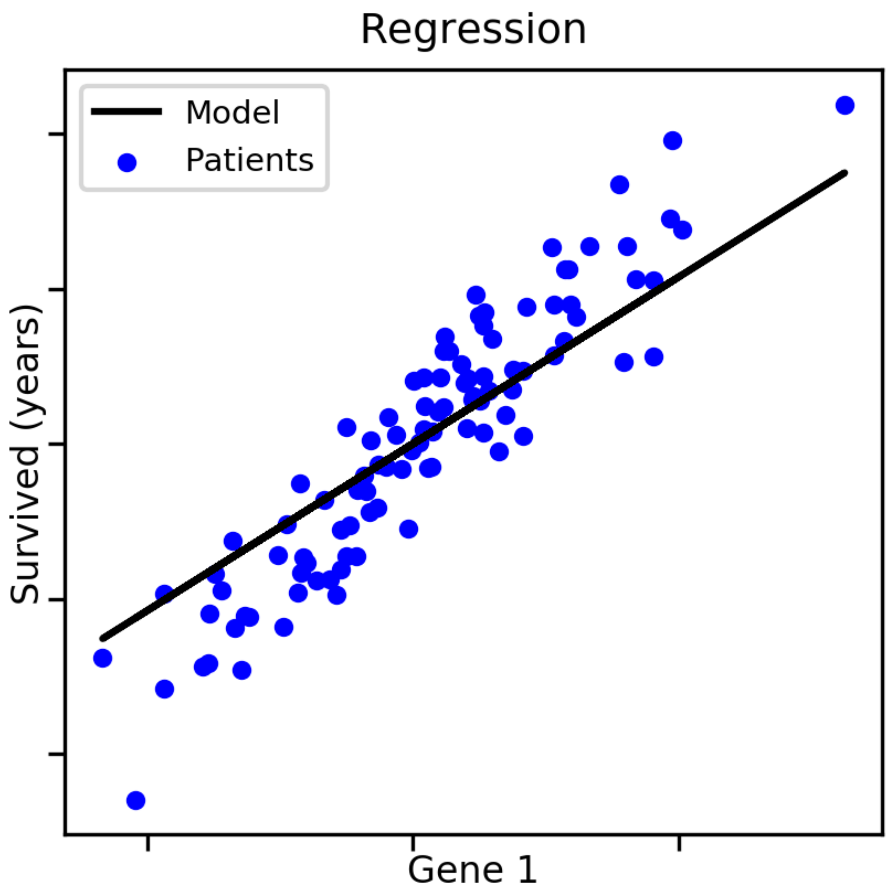
Difference between regression and classification

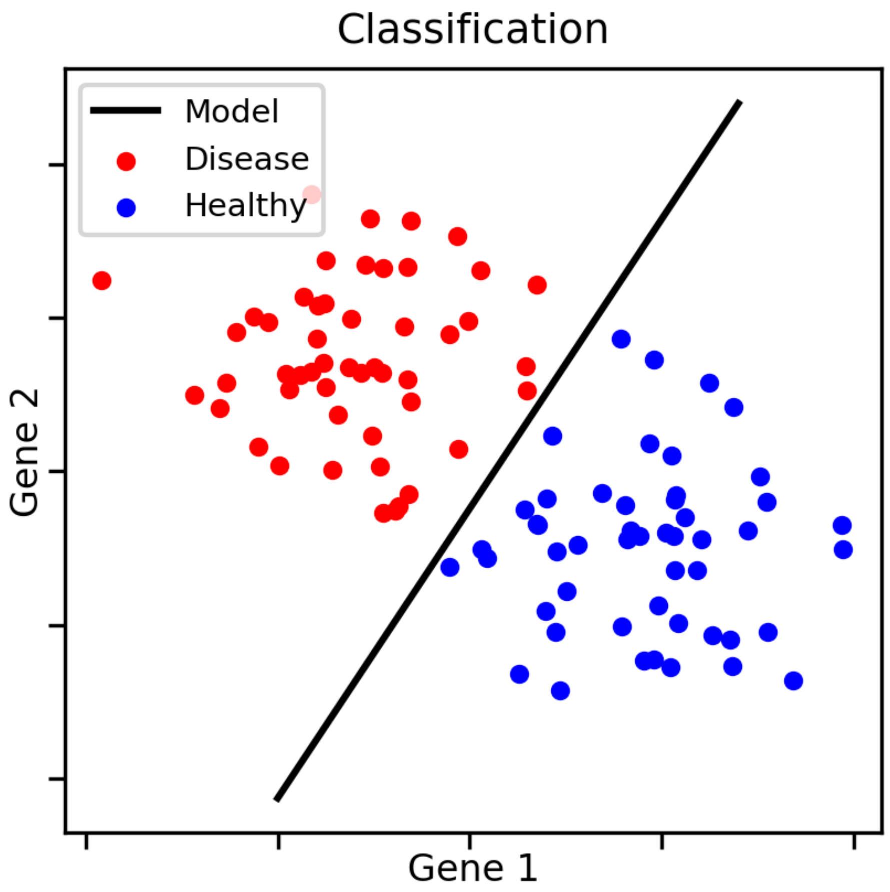
- **Regression:** Response Y is **quantitative** (**numerical**), and so predications are numbers.
- **Classification:** Response Y is **qualitative** (**categorical**), and so predictions are classes (which could be represented as numbers).



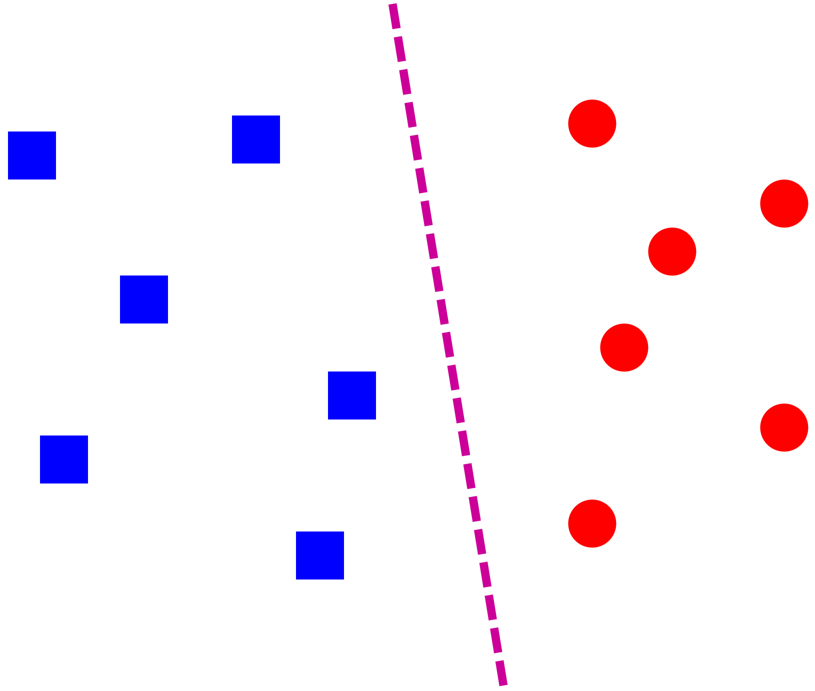
Visualizing the difference with regression

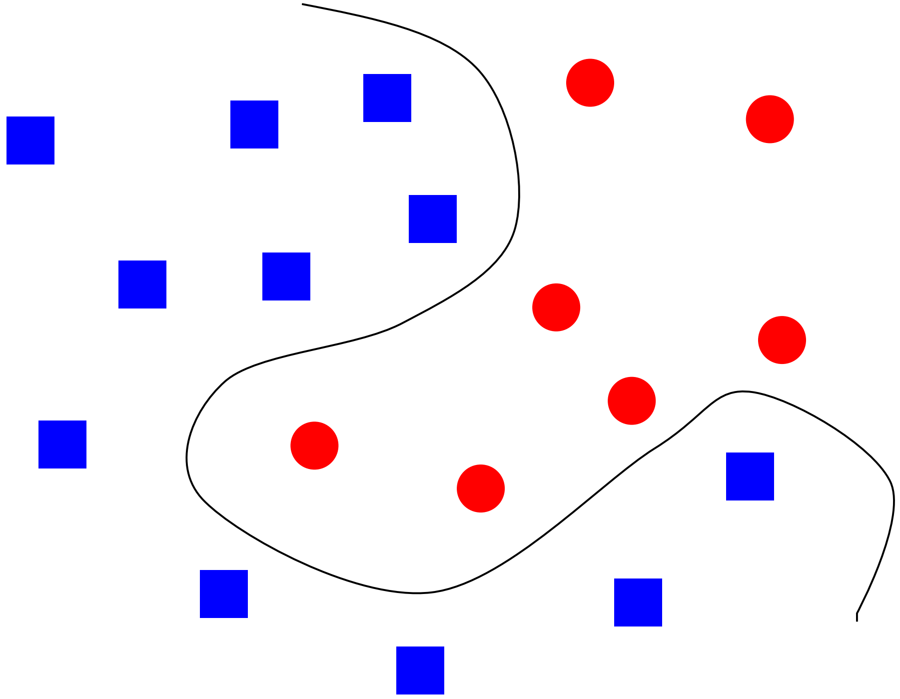
- **Regression:** Predict a **quantitative** response by **fitting** the data.
  - predict numeric/continuous values
- **Classification:** Predict a **qualitative** response by **splitting** the data.
  - predict categorical/discrete values





Classifiers: Linear vs Non-linear





- Find a *linear function* to separate the classes:

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$$

- Find a *non-linear function* to separate the classes:

Supervised Learning (Classification)

$$y = f(\mathbf{x})$$

output classification/(prediction) function input

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the classification(prediction) function f by minimizing the prediction error on the training set
- **Testing:** apply f to a new *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

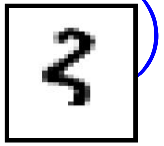
Supervised Learning (Classification)

- Apply a prediction function to a feature representation of the image to get the desired output:

f() = apple

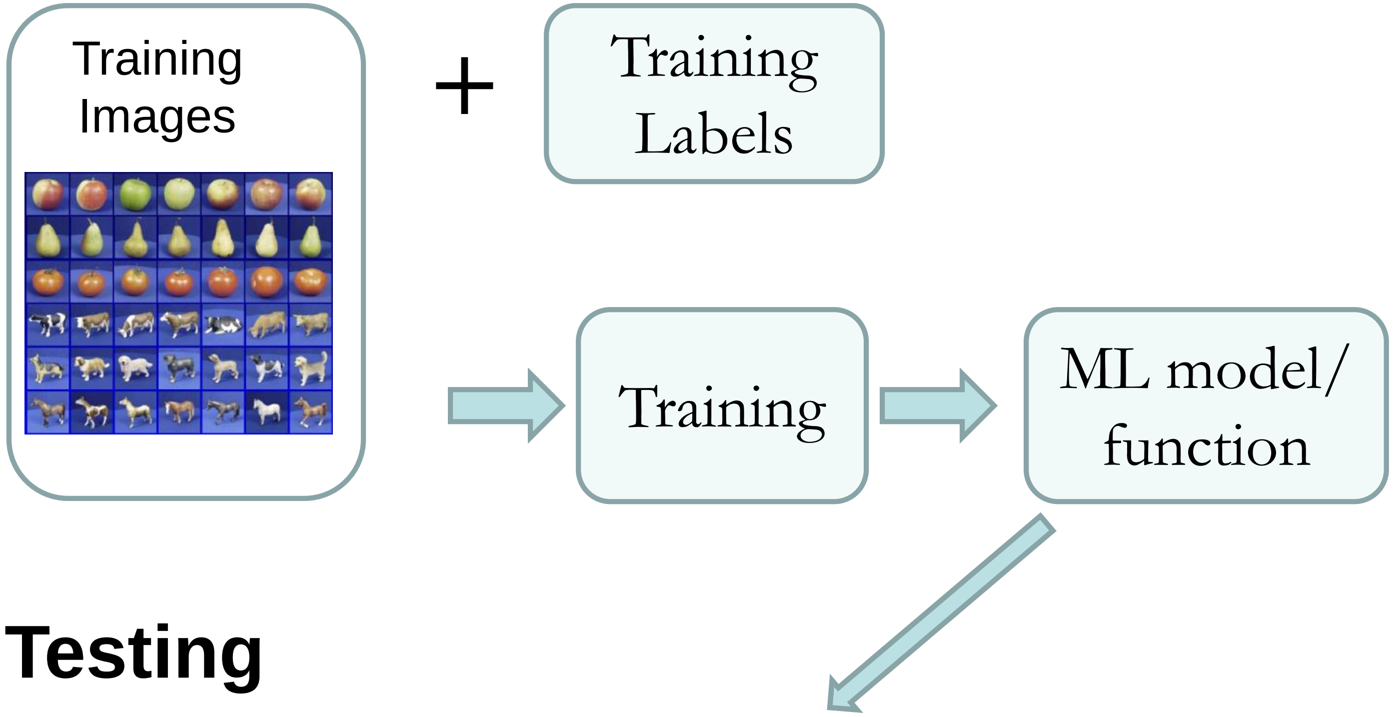
f() = cow

f() = a7(move)

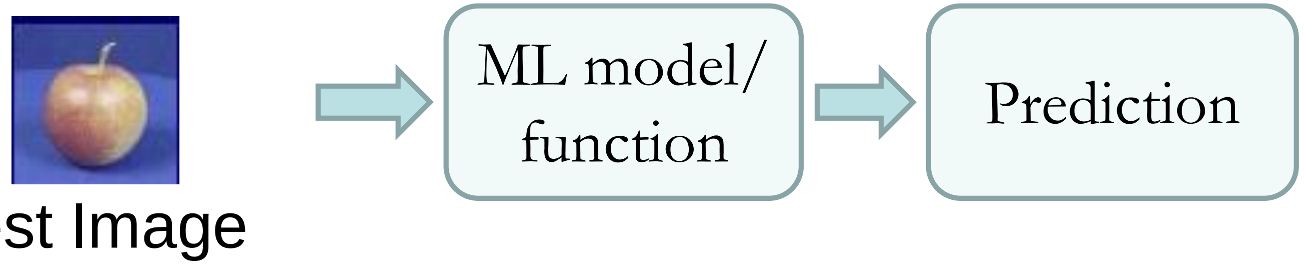
f() = 3

Supervised Learning (Classification)

Training



Testing



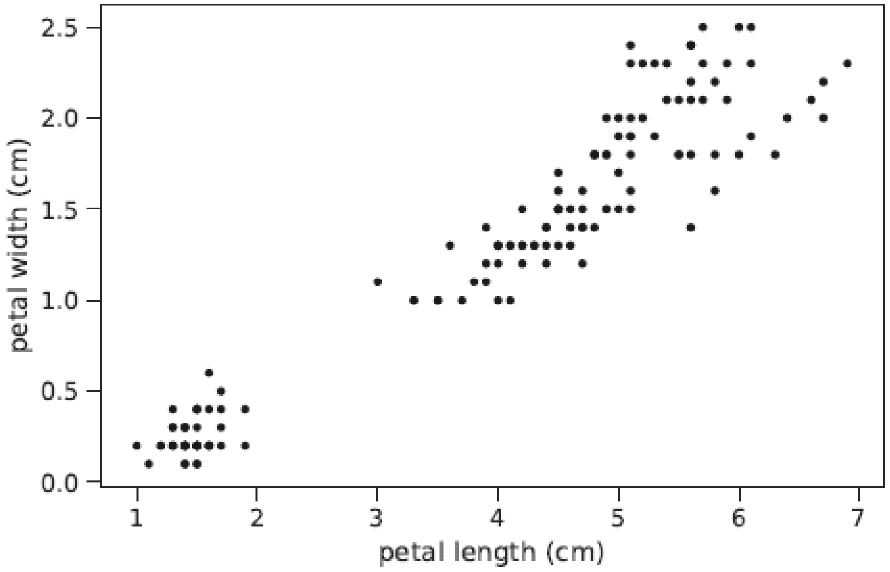
Many classifiers to choose from

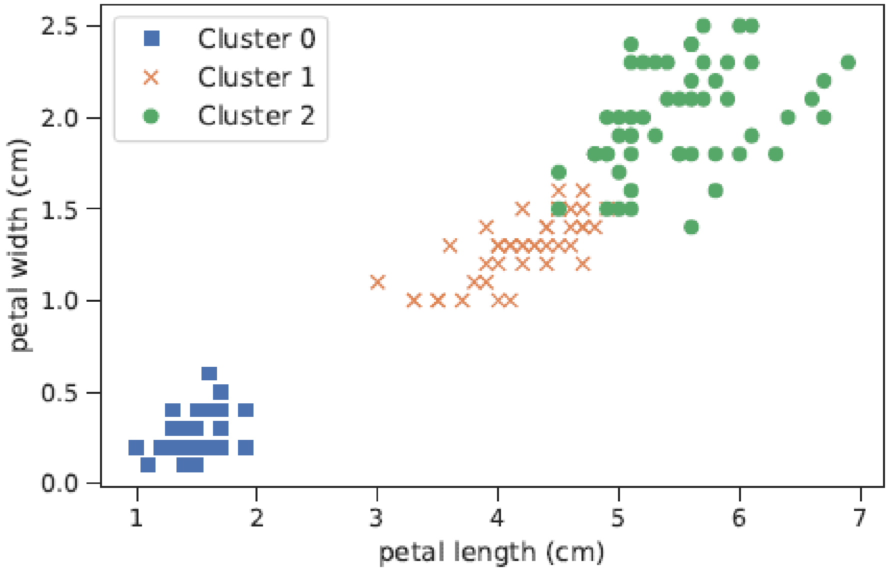
- Linear/Logistic Regression
- Decision Trees
- Neural networks
- SVM
- Random Forest
- AdaBoost
- Xgboost
- K-nearest neighbor(IBL)
- Deep Learning(CNN/RNN, etc)
- Naïve Bayes/Bayesian network
- Etc.

Unsupervised Learning (Clustering)

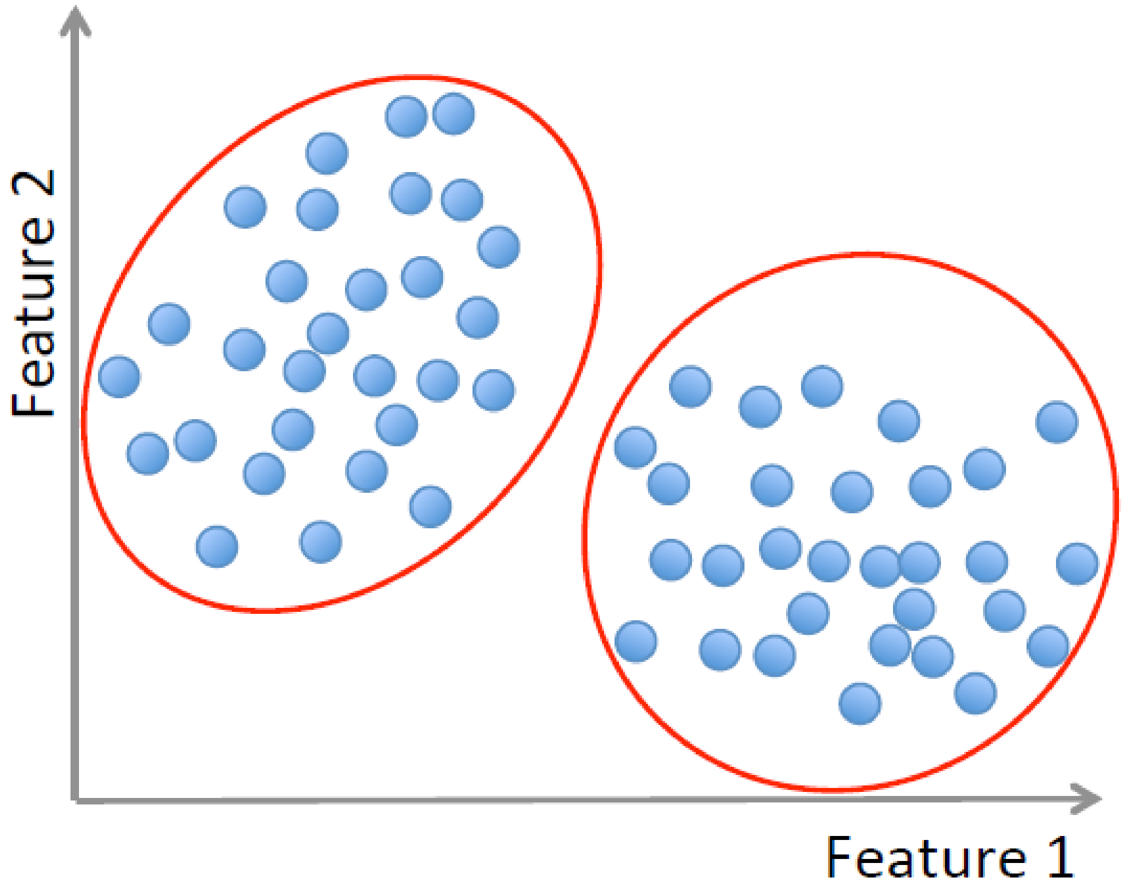
Goal:

Partition the input into regions that contain "similar" points.





Unsupervised Learning (Clustering)



Unsupervised Learning(Clustering) Algorithms

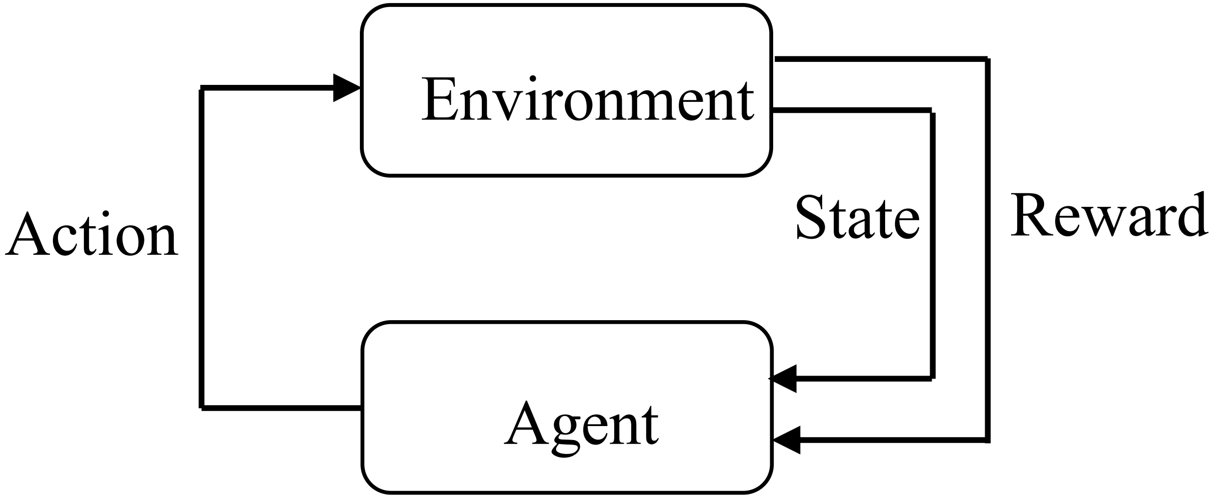
- K-means
  - Iteratively re-assign points to the nearest cluster center
- Agglomerative clustering
  - Start with each point as its own cluster and iteratively merge the closest clusters
- EM(Expectation Maximization)
  - Mixture of Gaussian Model
- etc

Reinforcement Learning

- Reinforcement learning(RL) : An area of machine learning concerned with how intelligent agents find *optimal actions* in an environment in order to achieve its goals.
  - e.g.: mobile robot, optimize operations in factories, learning to play board games

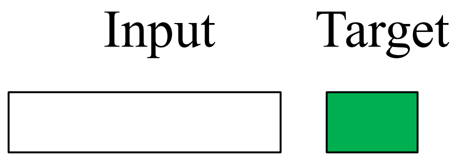
- Each time the agent performs an action, its environment may provide a reward/penalty to indicate the desirability of the resulting state

- Learn successful action policies by experimenting in their environment
  - Learn from trial and error



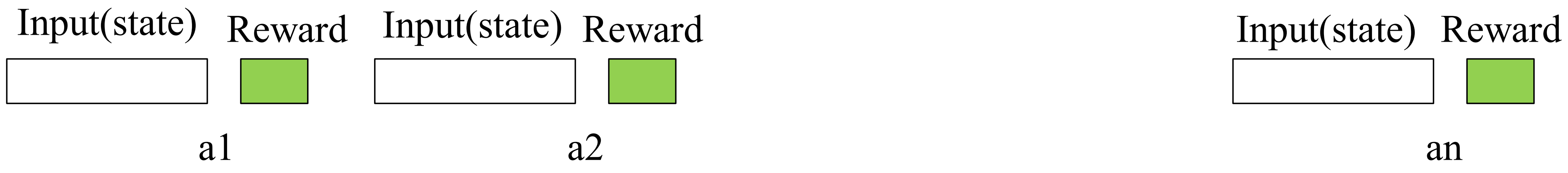
Reinforcement Learning

- Supervised Learning



Goal: predicts target values correctly

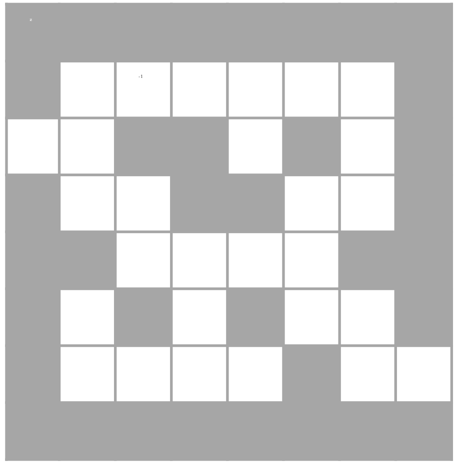
- Reinforcement Learning



Goal: find (optimal) sequence of actions that maximizes the summation of rewards

Reinforcement Learning

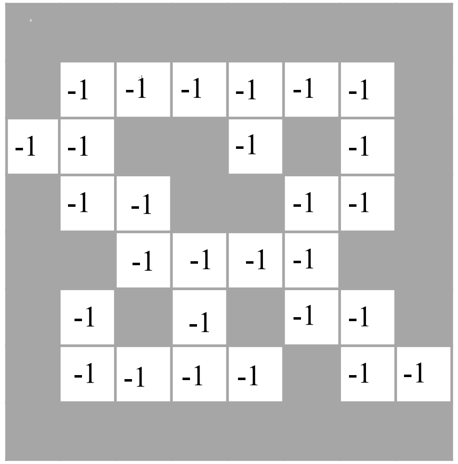
start



end

- Actions: N, E, S, W
- States: Agent's location

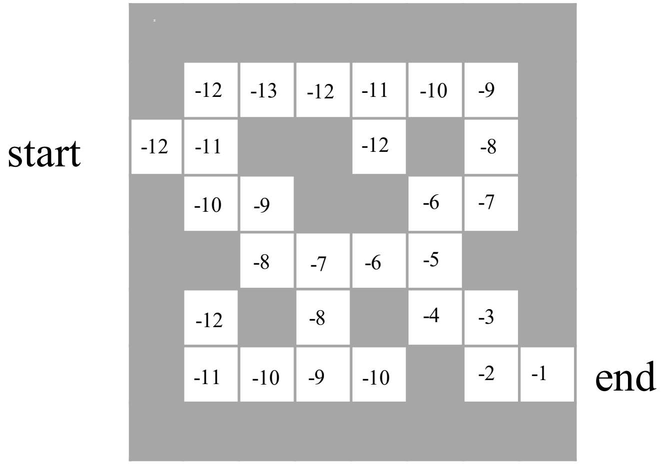
start



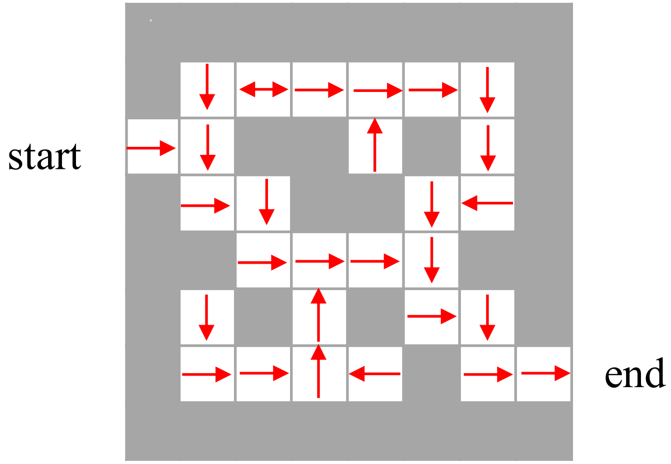
end

- Model
  - Grid layout represents transition model
  - Rewards: how much reward from each state (-1 per time-step)
  - Numbers represent immediate reward from each state s (same for all a)

Reinforcement Learning



- Value function: Numbers represent maximum rewards from each state s



- Policy: Arrows represent policy $\pi(s)$ for each state s

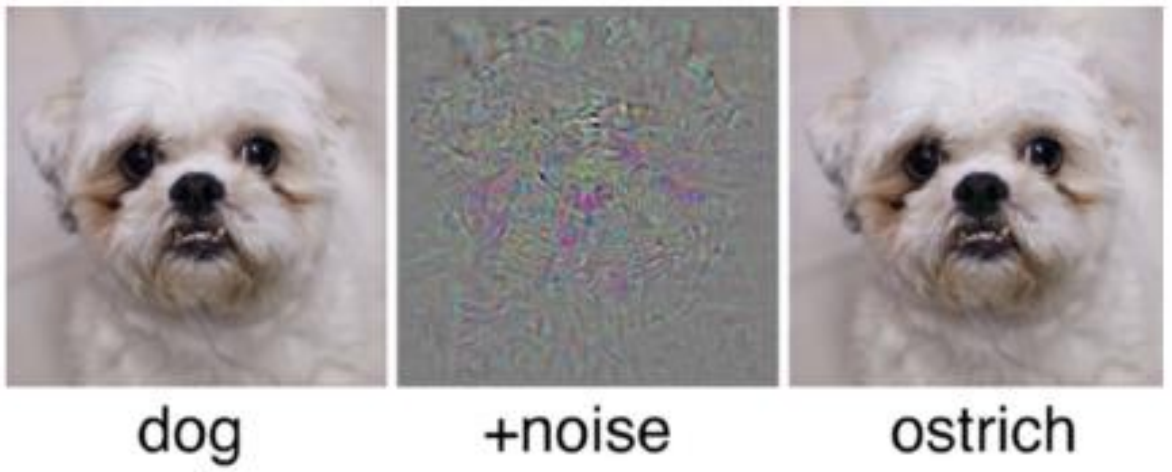
Process of Data Mining (Machine Learning) Project

- Identifying the problem: The first step is to determine what you want to achieve through data mining. This could be anything from improving sales performance to identifying potential fraud.
- Gathering data: Once the problem is identified, data from different sources is collected and combined to create a single, comprehensive dataset.
- **Preprocessing**: The most time-consuming phase. The data must be prepared for mining. This includes cleaning up missing or irrelevant values, handling noisy data, and normalizing the data for consistency.
- **Applying algorithms**: With clean data in hand, various statistical and mathematical algorithms are applied to identify patterns and relationships within the dataset.
- **Evaluating results**: After running the algorithms, the results need to be analyzed and interpreted to understand their significance in solving the identified problem.
- Utilizing insights: The final step is using these insights to inform decision-making and drive business growth or improvement.

Caveat

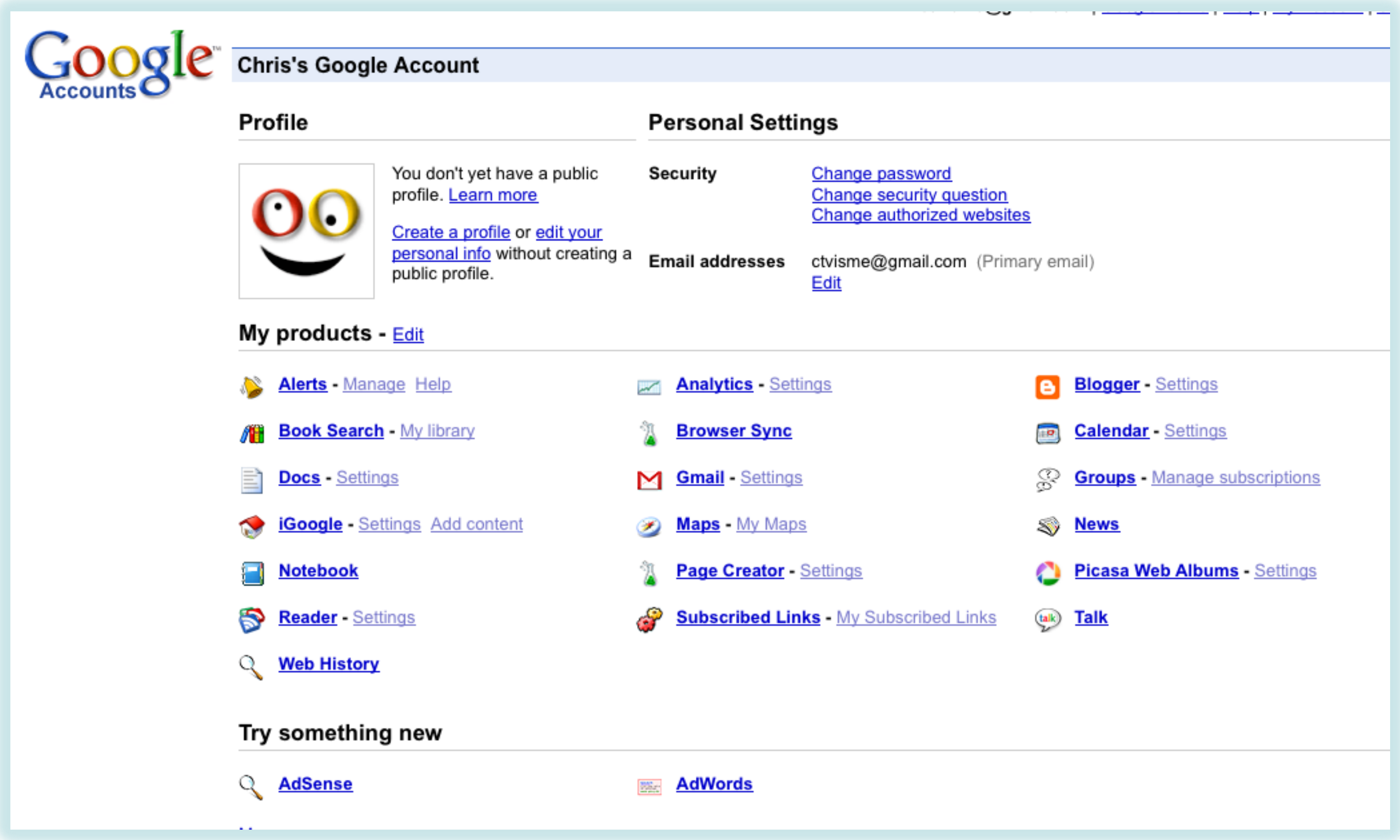
Anomaly Detection

- State-of-the-art classifiers can be fooled by adding imperceptible noise



Machine Learning vs. Privacy

- There is often tension between machine learning and personal privacy
- More data about more people in fewer places



No Free Lunch Theorem

- (simplified) For any classifier H1 and H2, if H1 ≥ H2 in some domain/data D1, there always exists other domain D2 where H1 < H2.

- If you compare H1, H2 for EVERY possible domain, no classifier is inherently better than any other
- Then why do we prefer an algorithm to others ?

- We can’t have EVERY possible domain
- Our world is full of biases(physical/chemical rules, law, science, etc)
- Thus, data generated from our world, have biases
- The goal of machine learning is to learn these ”biases in data” correctly and efficiently.

- Learning bias is the most important key in human/machine learning.
  - Human Learning is about learning *Bias in nature*
  - Machine Learning is about learning *Bias in data*