Machine Learning Overview

Michael Wang

What is machine learning?

The science of programming computers so they can learn from data.

- Géron (2019)

The field of study that gives computers the ability to learn without being explicitly programmed.

- Arthur Samuel, 1959

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, 1977

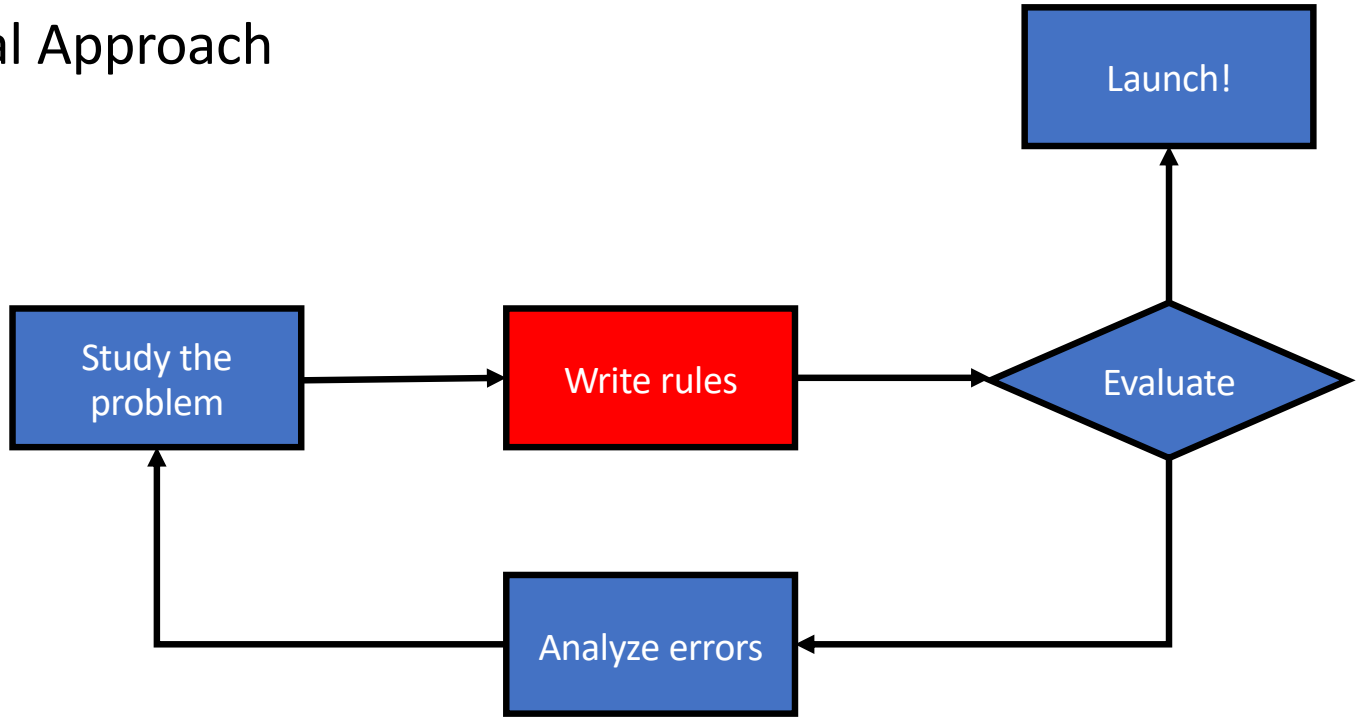
Data… and its “unreasonable effectiveness”

- Mathematics can provide elegant descriptions of some parts of our world
  - $f = m \cdot a$
  - $e = m \cdot c^2$
- But it is difficult, sometimes even impossible, to use math to provide descriptions of other aspects of our world
  - E.g., psychology, economics, etc.
- Data comes to the rescue!

Halevy, Norvig, Pereira, 2009

Why use machine learning?

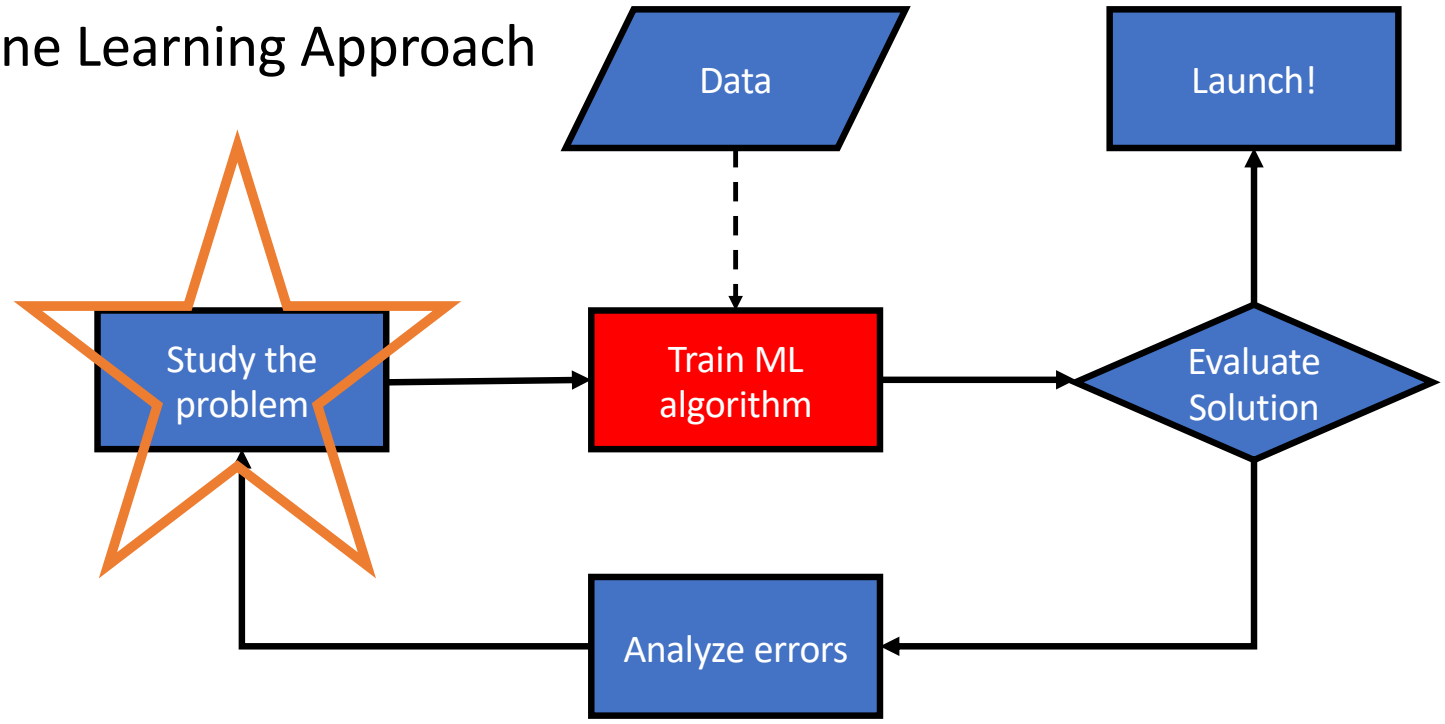
Traditional Approach



Géron, 2019

Why use machine learning?

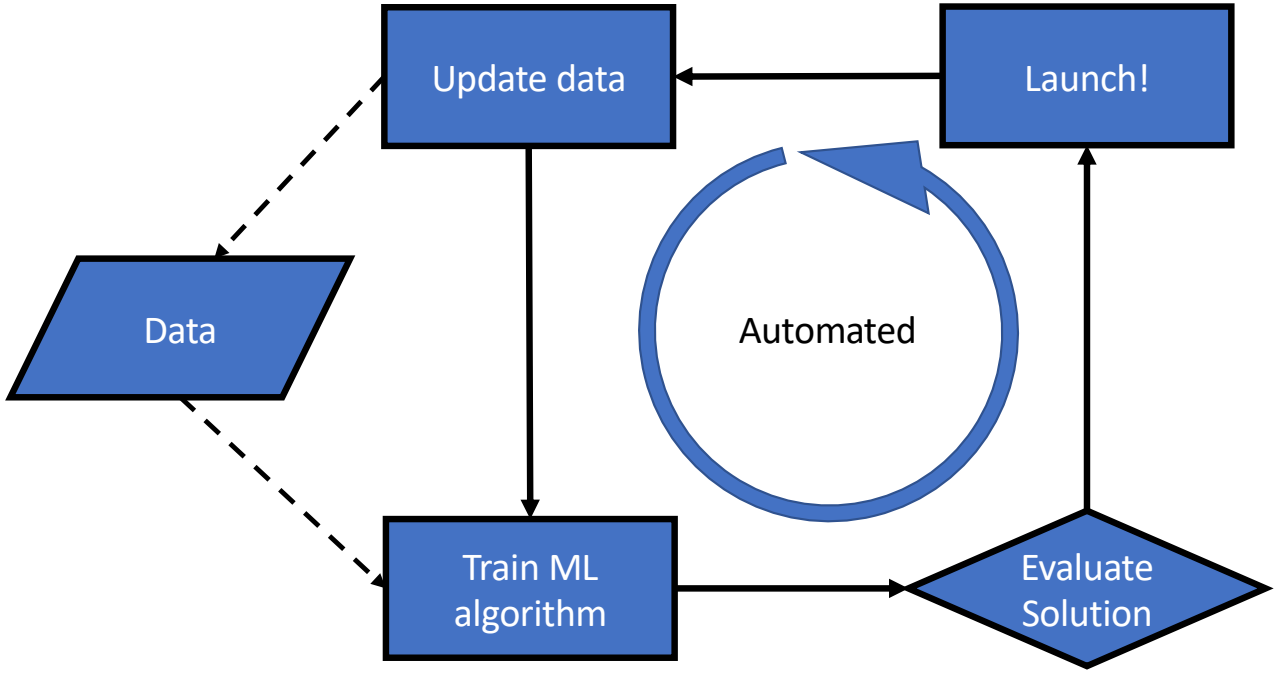
Machine Learning Approach



Géron, 2019

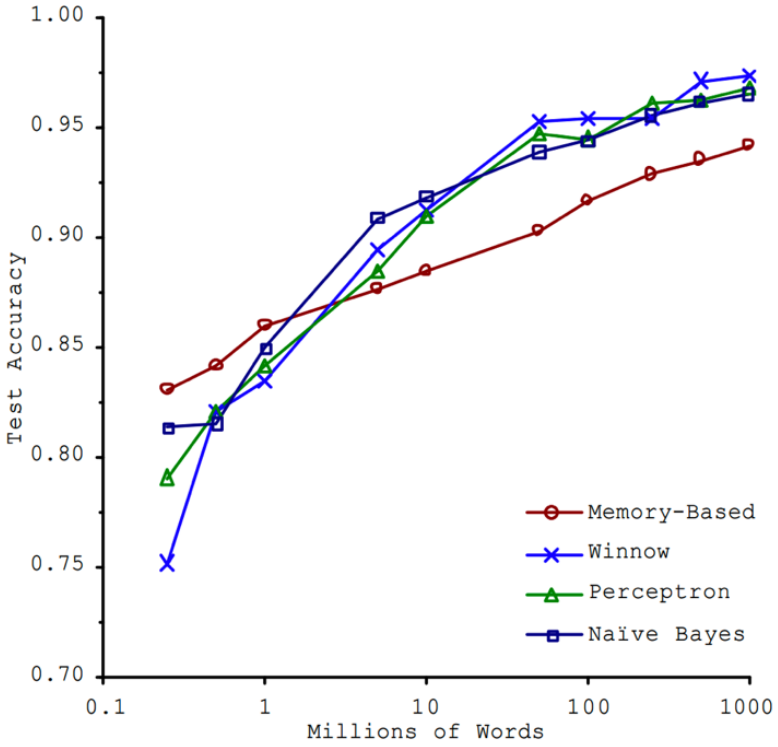
Why use machine learning?

Adaptive to change



Géron, 2019

Data… and its “unreasonable effectiveness”



“Bass”:
I went fishing for some bass.
The bass of this song is too strong.
She plays bass in a jazz band.

Scaling to Very Very Large Corpora for Natural Language Disambiguation.
Banko & Brill, 2001

Common Machine Learning Tasks

- Binary classification
  - Twitter comment sentiment, “positive” or “negative”
  - Disease diagnosis
  - Email spam
- Multiclass classification – Feb 6
  - Dog breed
  - Categorize hotel reviews (location, price, cleanliness, etc.)
- Regression – Feb 4
  - Predicting housing price based on house attributes
  - Predicting stock prices based on historical data and current market trends

https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/tasks

Common Machine Learning Tasks

- Clustering
  - Hotel guest segments based on habits and characteristics of hotel choices
  - Targeted advertising campaigns based on customer segments and demographics
- Anomaly detection
  - Fraudulent transaction detection
  - Identifying network intrusion
- ……

https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/tasks

Machine Learning Designs

- Batch learning
  - Trained with all available data.
    - If received new data, need to train the system from scratch.
  - Takes a lot of time and computing resources.
  - Typically done offline.
- Online learning
  - Trained incrementally by feeding data instances sequentially.
  - Suitable for a system that receives data as a continuous flow and needs to adapt to change rapidly or autonomously.
  - Does not require a lot of computing resources.

Géron, 2019

Machine Learning Approaches

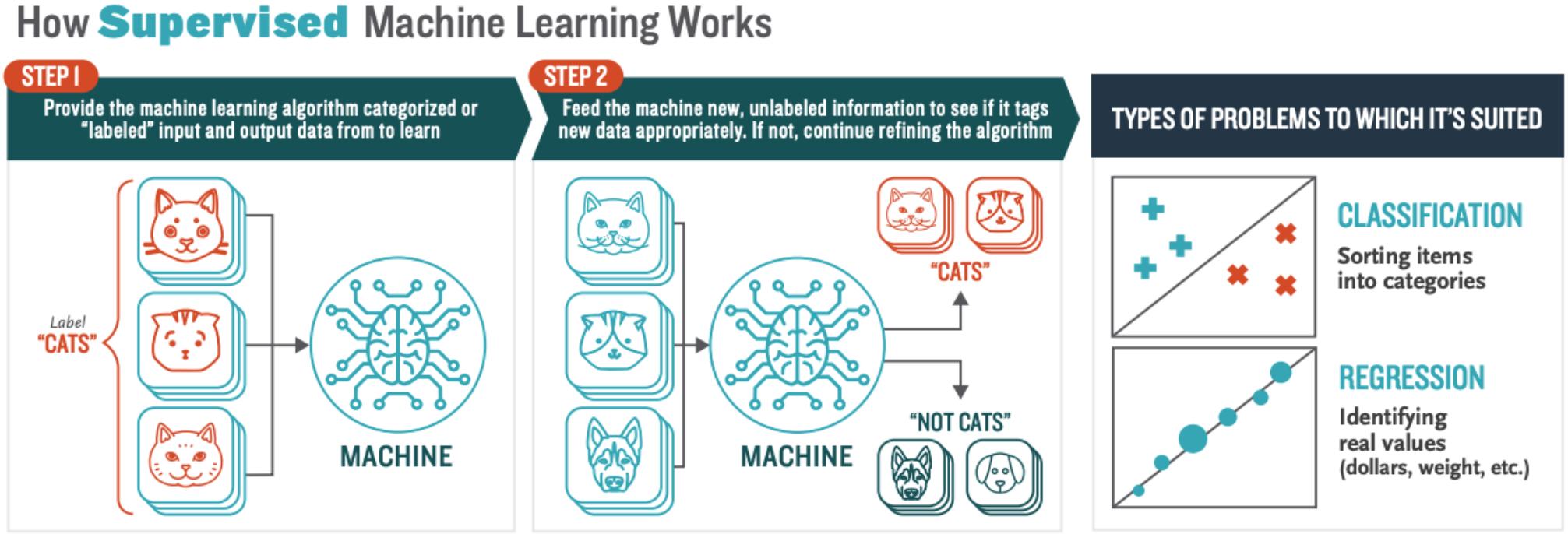
- Supervised learning
  - The algorithm receives both data and desired solutions (labels) during training.
- Unsupervised learning
  - The algorithm receives unlabeled data during training.
- Semi-supervised learning
  - The combination of supervised and unsupervised learning.
- Reinforcement learning
  - Completely different from the previous ones.
  - Uses an agent inside an environment. The agent can perform certain actions and get rewards or penalties in return. The agent needs to identify the best strategy (a policy) to maximize reward overtime.

Géron, 2019

Supervised Learning

- The machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- It infers a function from labeled training data consisting of a set of training examples.

Supervised Learning



https://www.boozallen.com/s/insight/blog/how-do-machines-learn.html

Supervised Learning

Given a set of N training examples of form

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$$

The supervised learning algorithm seeks to find a function $h: X \to Y$ such that

$$f^* = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{N} \mathcal{L}(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

Objective (loss) function

Hypothesis space

Training data

Supervised Learning

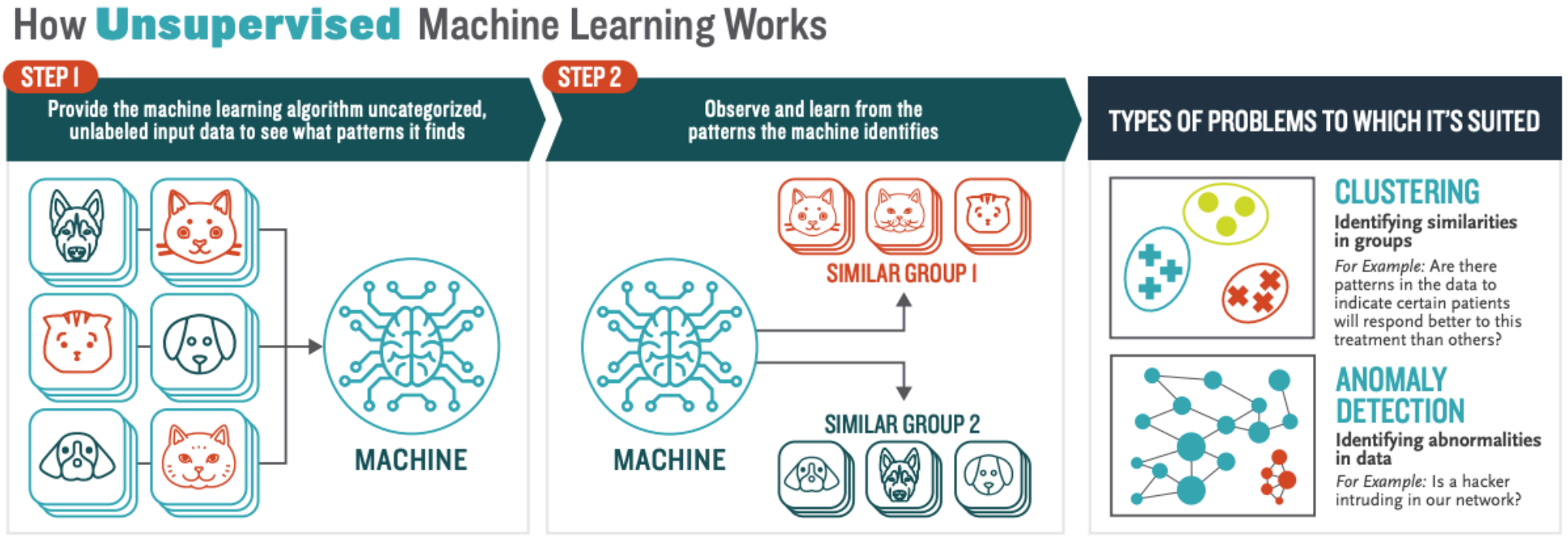
- Common supervised learning algorithms
  - k-Nearest Neighbors
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVMs)
  - Decision Trees and Random Forests
  - Neural networks*

Géron, 2019

Unsupervised Learning

- Forces the machine to build a compact internal representation of its world via feature spaces.
  - Feature: an individual measurable property or characteristic of a phenomenon being observed.
- The machine learning task of learning a new feature space that captures the characteristics of the original space by maximizing some objective function or minimizing some loss function.

Unsupervised Learning



https://www.boozallen.com/s/insight/blog/how-do-machines-learn.html

Unsupervised Learning

Given a set of N training inputs of form

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$$

The unsupervised learning algorithm seeks to find a function $\mathrm{h}: X \to X$ such that

$$f^* = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{N} \mathcal{L}(h_\theta(\boldsymbol{x}_i), \boldsymbol{x}_i)$$

https://www.boozallen.com/s/insight/blog/how-do-machines-learn.html

Dimensionality Reduction - PCA

Original Data 537 (frames) × 52 (joints) × 3 (xyz)

PCA Components: number of PCs × 52 (joints) × 3 (xyz)

Original

PC1

PC2

PC3

PC6

Unsupervised Learning

- Common unsupervised learning algorithms
  - Clustering
    - k-Means
    - DBSCAN
    - Hierarchical Cluster Analysis
  - Anomaly detection and novelty detection
    - One-class SVM
    - Isolation Forest
  - Visualization and dimensionality reduction
    - Principal Component Analysis (PCA)
    - Kernel PCA
    - Locally-Linear Embedding (LLE)
    - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Association rule learning
    - Apriori
    - Eclat

Géron, 2019

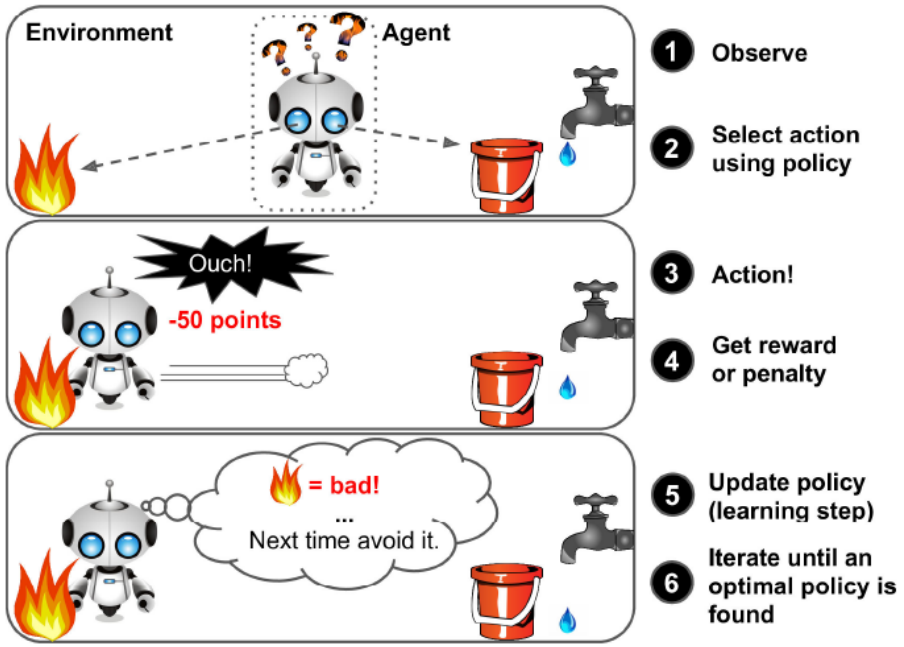
Semi-supervised Learning

- Photo apps that can identify the same person and let you tag his/her name.
- Mostly are combinations of unsupervised and supervised algorithms.
- Deep belief networks (DBNs)
  - Consisted of layers of unsupervised components called restricted Boltzmann Machines (RBMs).
  - Trained sequentially and unsupervised.
  - The trained system then act as feature detectors and be further trained with supervision to perform classification.

Géron, 2019

Reinforcement Learning

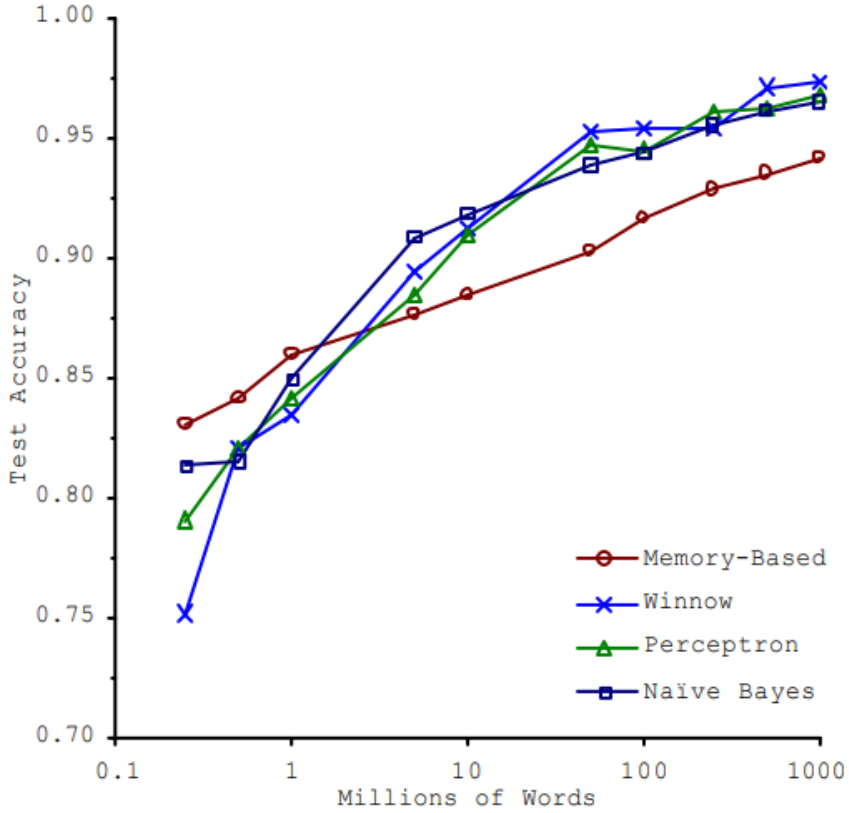
- Agent-Environment-Rewards(Penalties)-Policy
- E.g. DeepMind's AlphaGo



Géron, 2019

Main Challenges

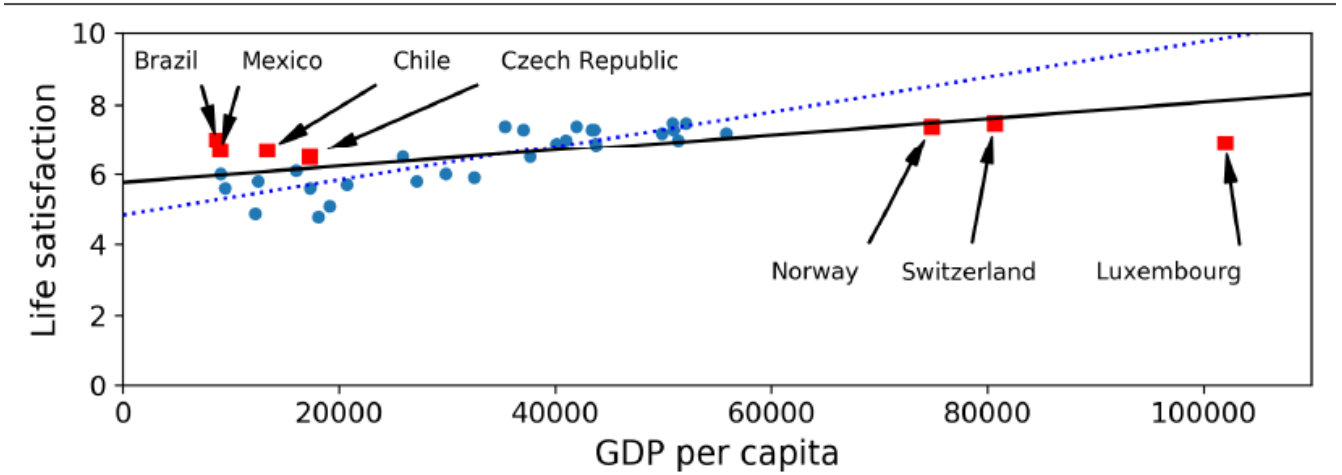
- Insufficient training data



Banko & Brill, 2001

Main Challenges

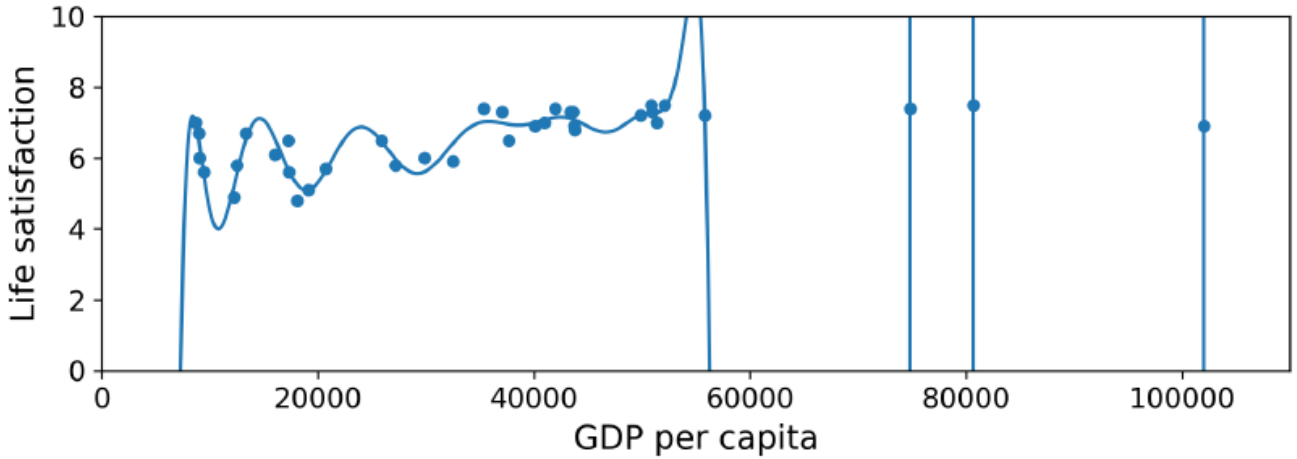
- Insufficient training data.
- Nonrepresentative training data.



Géron, 2019

Main Challenges

- Insufficient training data.
- Nonrepresentative training data.
- Overfitting.



Géron, 2019

Main Challenges

- Insufficient training data.
- Nonrepresentative training data.
- Overfitting.
- Interpretability and generalizability.
  - The “black box” metaphor.
- ……

Géron, 2019

A Final Word

- Machine learning cannot magically help you to solve all the problems.
- It still requires you to:
  - Identify and study the problem.
  - Use appropriate tools.
  - Keep identifying potential issues and continue to improve the algorithm.