MACHINE LEARNING

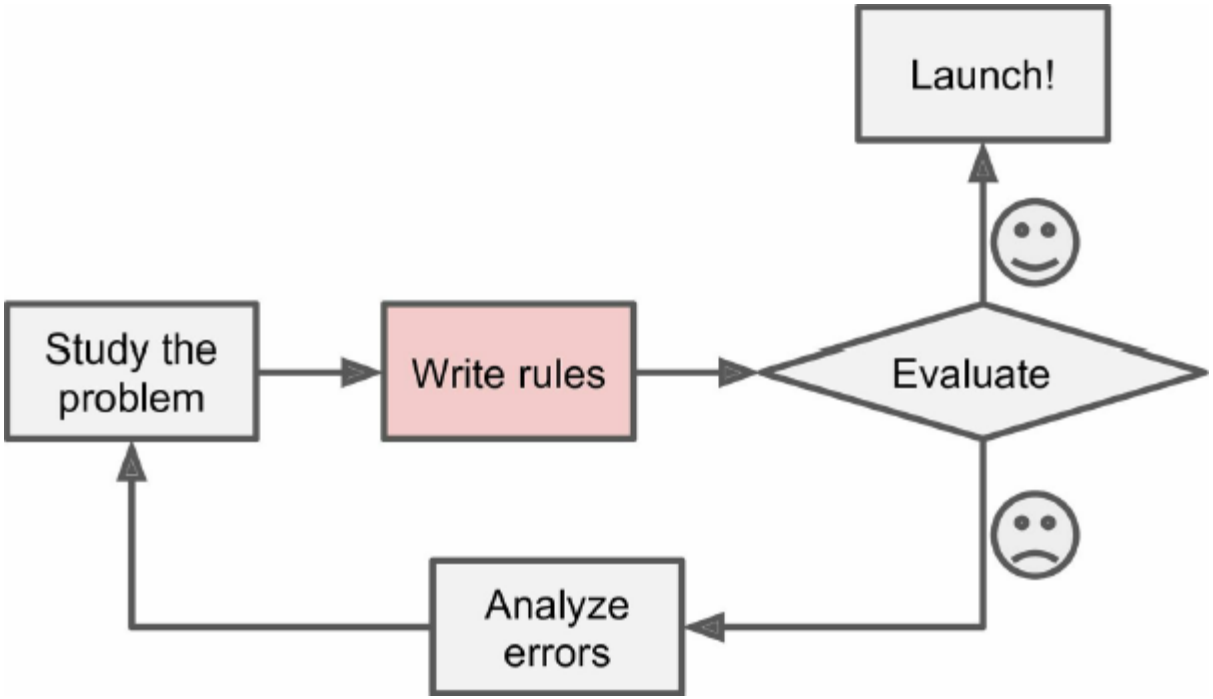
Jahan Ghofraniha, Ph.D.

What is Machine Learning?

- Machine Learning is the science (and art) of programming computers so they can learn from data.
- [Machine Learning is 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, 1997

Why ML?

- Traditional approach:

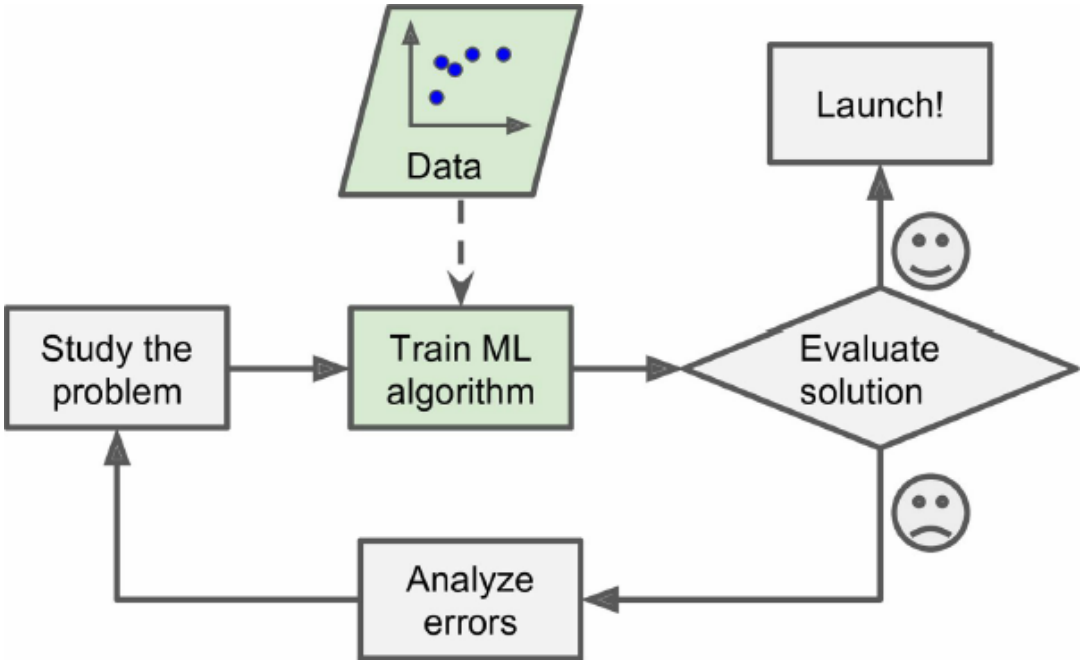


Why ML?

- In contrast, a spam filter based on Machine Learning techniques automatically learns which words and phrases are good predictors of spam.
- By detecting unusually frequent patterns of words in the spam examples compared to the ham examples.

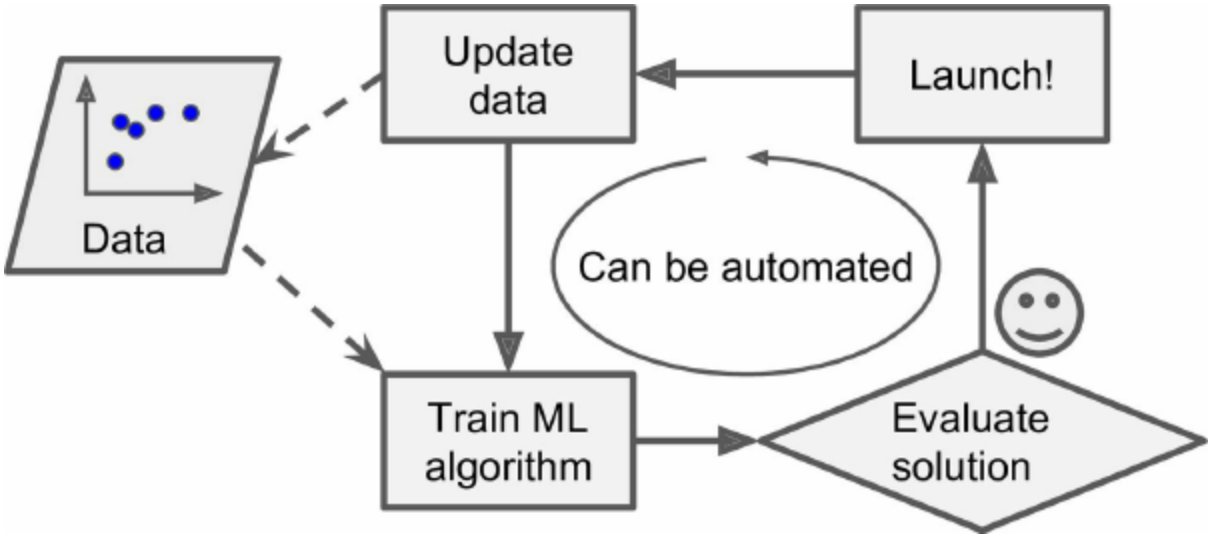
Why ML?

- The program is much shorter, easier to maintain, and most likely more accurate even if the hackers change their strategy.



Why ML?

- A spam filter based on Machine Learning techniques automatically notices that “For U” has become unusually frequent in spam flagged by users, and it starts flagging them without your intervention



Why ML?

- Another area where Machine Learning shines is for problems that either are too complex for traditional approaches or have no known algorithm.
- For example, consider speech recognition: say you want to start simple and write a program capable of distinguishing the words “one” and “two.” You might notice that the word “two” starts with a high-pitch sound (“T”), so you could hardcode an algorithm that measures high-pitch sound intensity and use that to distinguish ones and twos.

Why ML?

- Obviously this technique will not scale to thousands of words spoken by millions of very different people in noisy environments and in dozens of languages. The best solution (at least today) is to write an algorithm that learns by itself, given many example recordings for each word.

Summary of ML Applications

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

Machine Learning Types

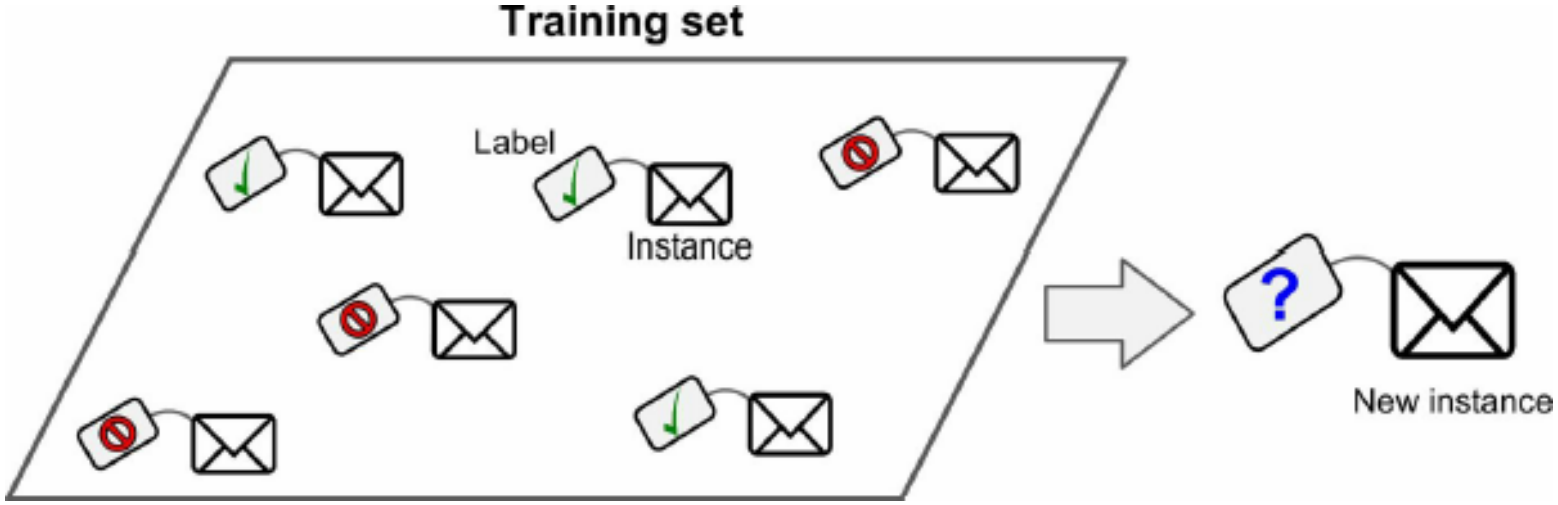
- There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:
- Whether or not they are trained with human supervision (supervised, unsupervised, semi-supervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)
- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

ML Types

- These criteria are not exclusive; you can combine them in any way you like.
- For example, a state-of-the-art spam filter may learn on the fly using a deep neural network model trained using examples of spam and ham; this makes it an online, model-based, supervised learning system.

Supervised & Unsupervised Learning

- Supervised vs. Unsupervised
  - Classification & regression

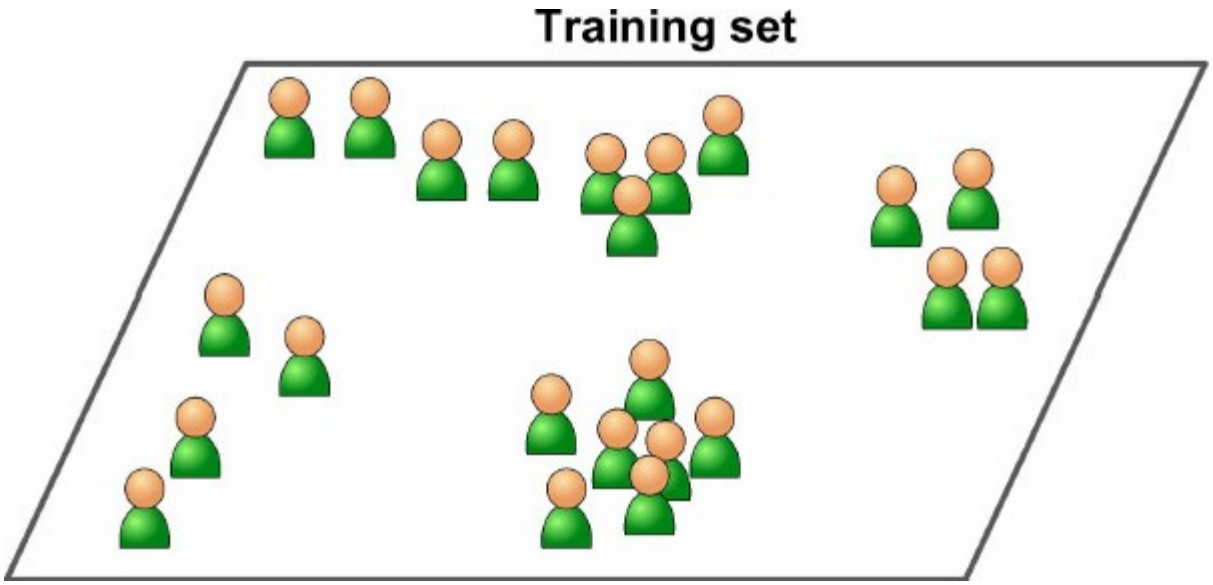


Supervised Algorithms

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

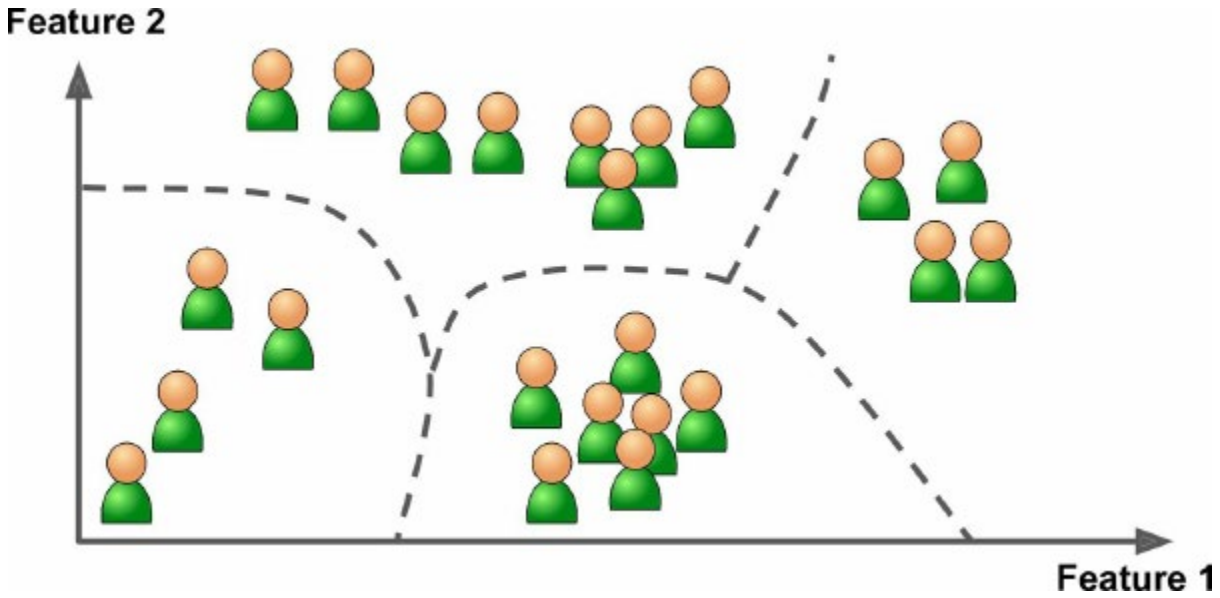
Unsupervised

- In unsupervised learning the training data is unlabeled.
- The system tries to learn without a teacher.



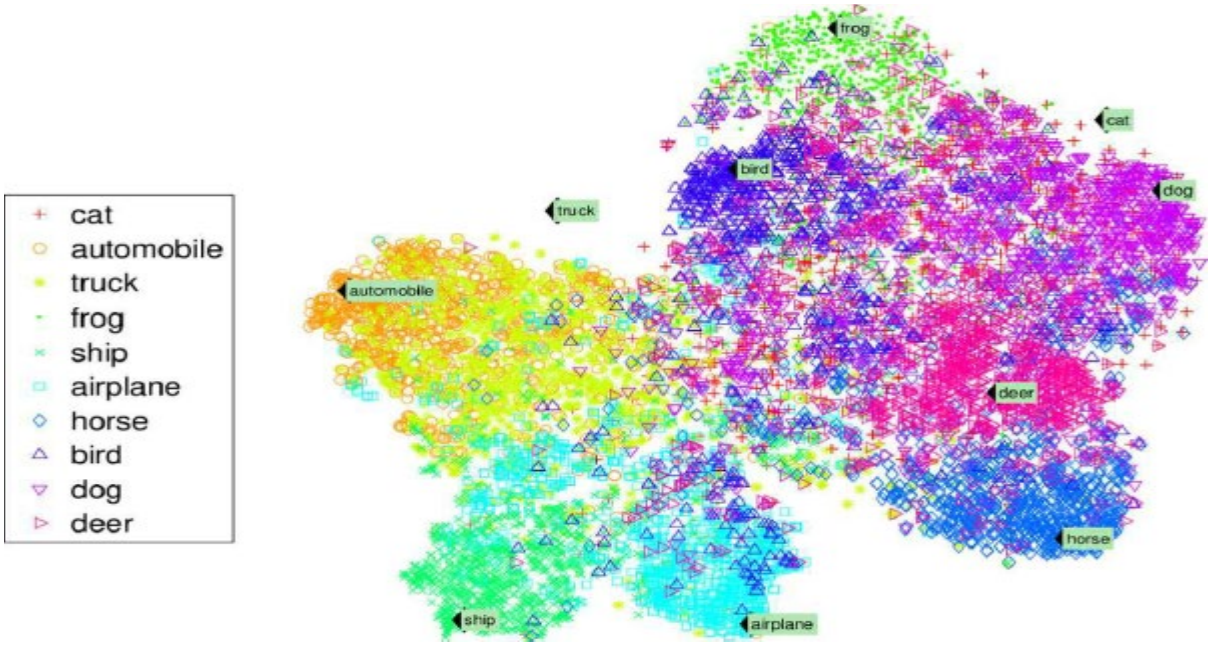
Unsupervised

- Clustering
  - k-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization



Unsupervised

- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE), t-distributed Stochastic Neighbor Embedding (t-SNE)



Unsupervised

- Association Rule Learning:
  - Apriori
  - Eclat

Unsupervised

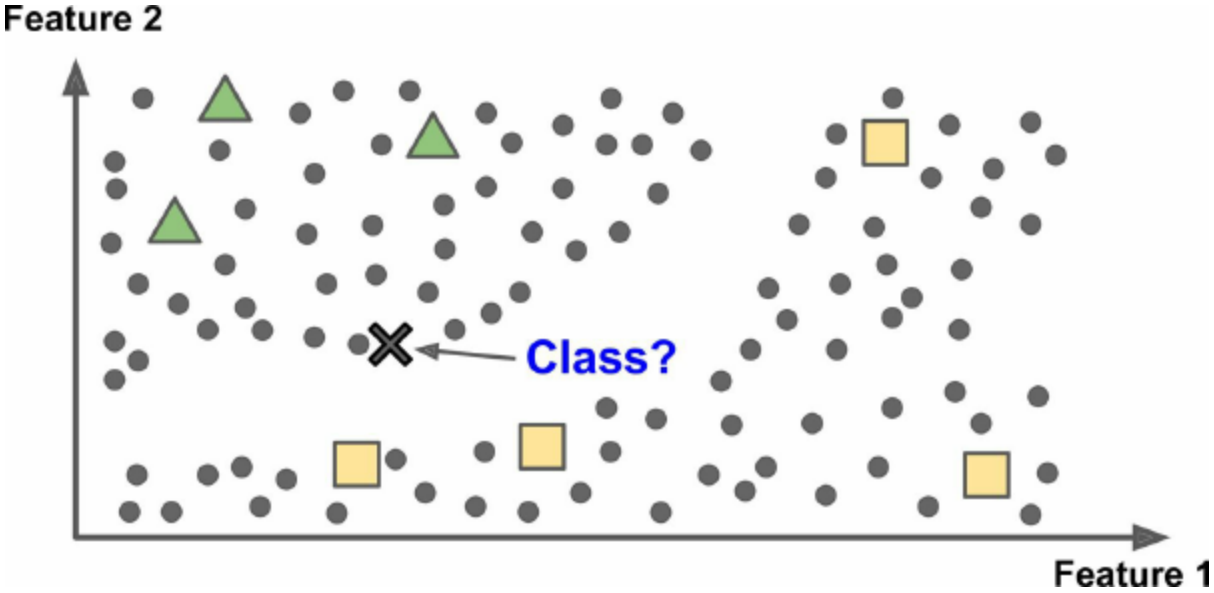
- Anomaly detection

Semi-supervised

- Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. This is called semi-supervised learning
- A good example of this is Google Photos
- Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11

Semi-supervised

- Most semi-supervised are combination of supervised and unsupervised.

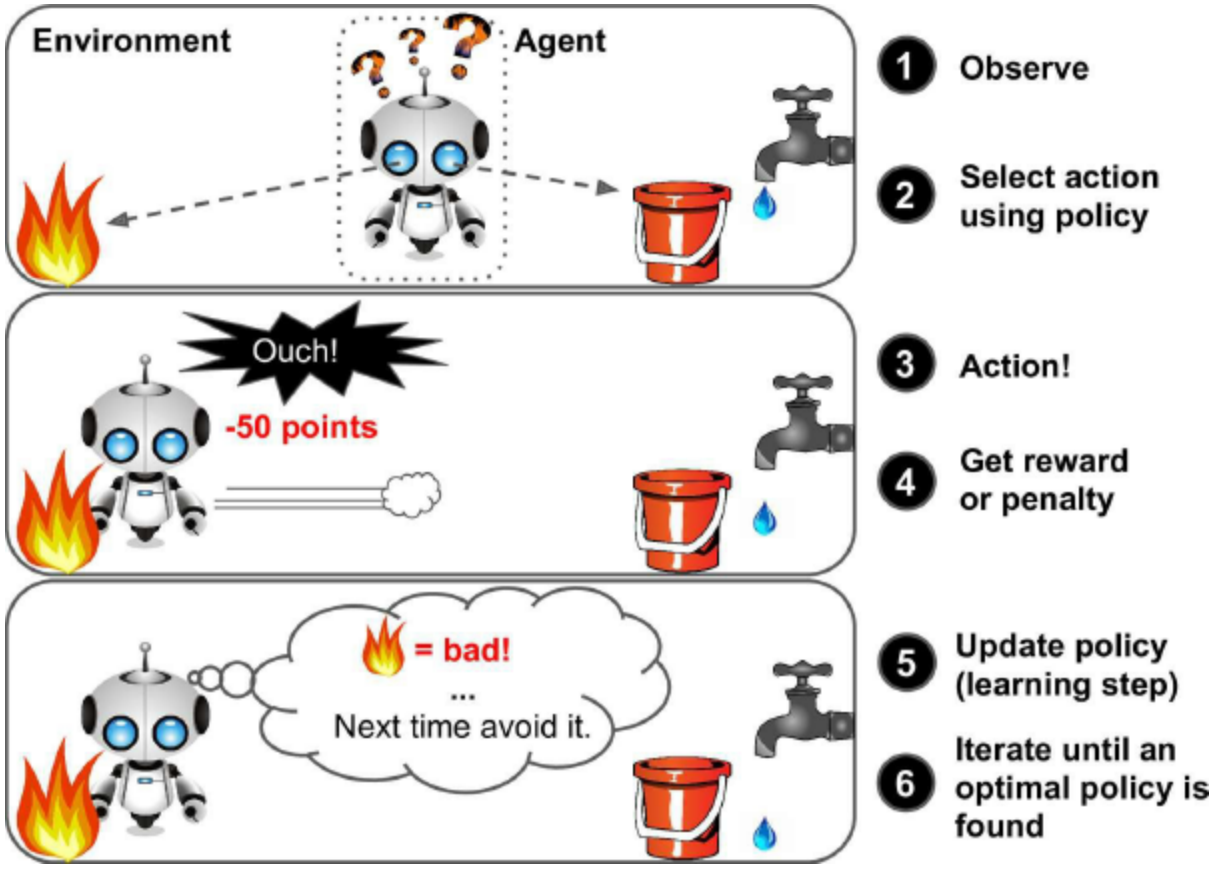


Reinforcement Learning

- Reinforcement Learning is very different in nature. The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards). It must then learn by itself what is the best strategy, called a policy, to get the most reward over time.

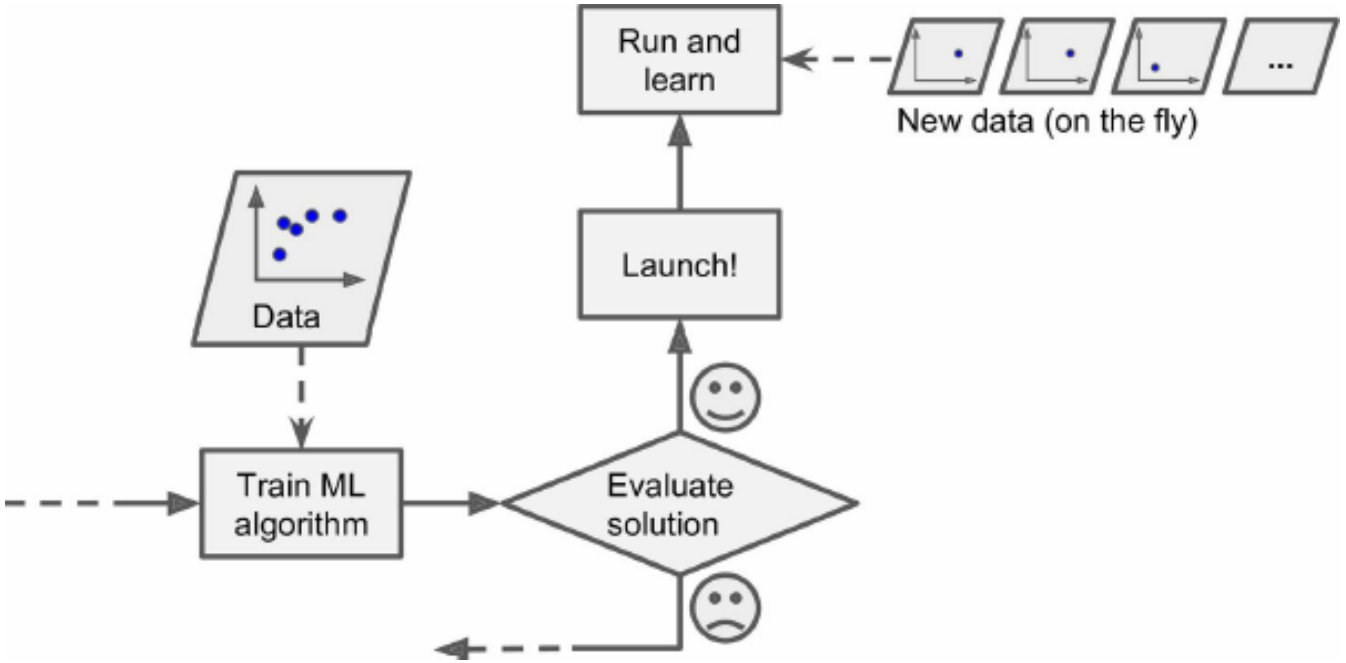
Reinforcement Learning

- A policy defines what action the agent should choose when it is in a given situation.



Online Learning

- Online learning is great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously. It is also a good option if you have limited computing resources

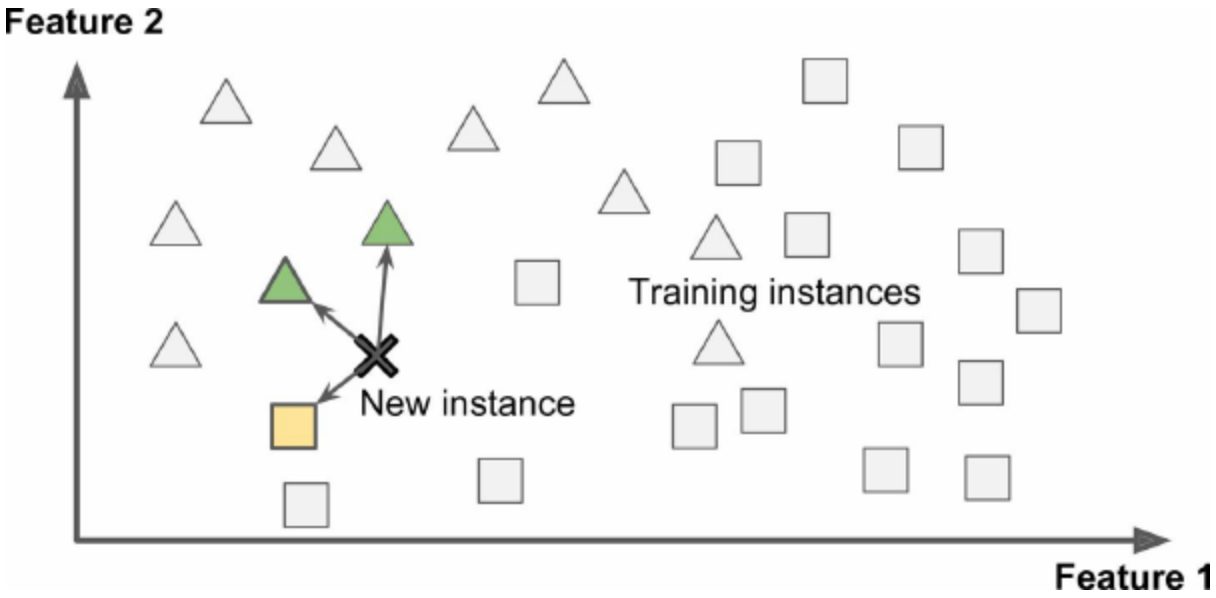


Batch Learning

- In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline. First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called offline learning.

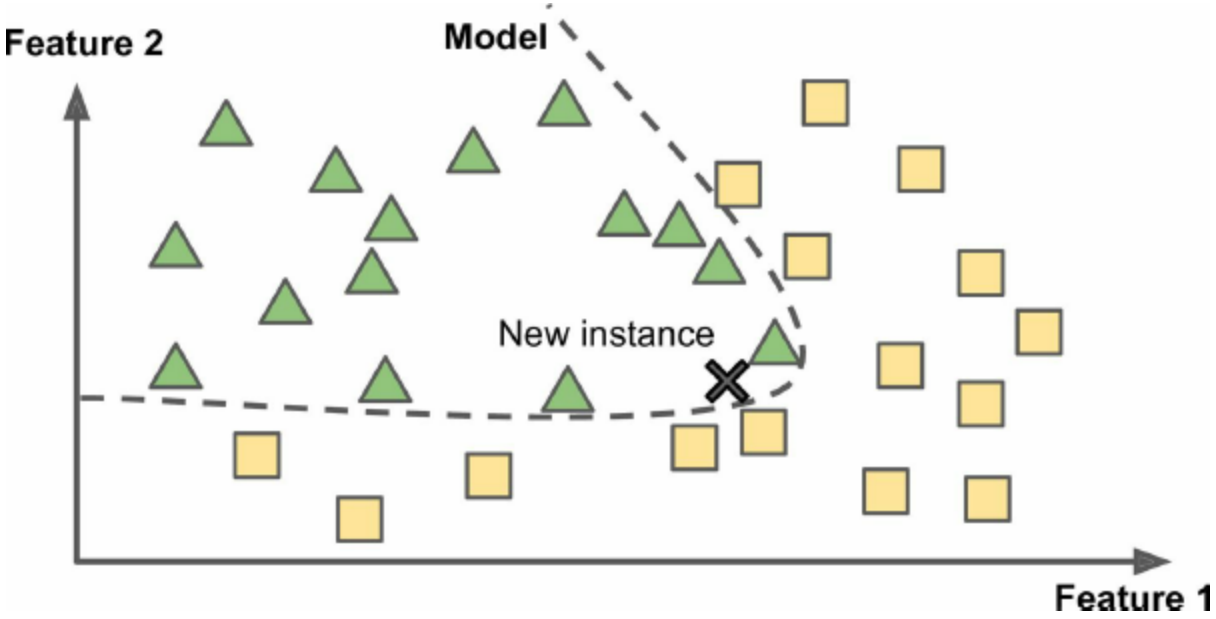
Instance-based Learning

- This is called instance-based learning: the system learns the examples by heart, then generalizes to new cases using a similarity measure



Model-Based Learning

- Another way to generalize from a set of examples is to build a model of these examples, then use that model to make predictions. This is called model-based learning



Tools

- Python as a learning tool.
  - Download Anaconda from: https://www.anaconda.com/download/
  - Instructions for using the tools will be provided and demonstrated separately.

Main Challenges of ML

- Two things can go wrong in a ML task:
  - Bad algorithm
  - Bad data
- Examples of Bad Data
  - Insufficient data
    - In a famous paper in 2001, Microsoft researchers showed that given enough data, almost all ML algorithms, even the simple ones perform well.
    - For complex problems, data matters more than algorithms (Peter Norvig, 2009)
    - However, small and medium size datasets are fairly common and we cannot abandon algorithms yet.
  - Non-representative data
    - Sampling bias (last presidential poll results, Clinton vs. Trump or 1936 campaign of London vs. Roosevelt)

Main Challenges of ML

- Poor-Quality Data
  - Outliers
  - Too many missing instances
- Irrelevant Features
  - Data should contain good and relevant features- feature engineering
  - Feature selection
  - Feature extraction (pca, …)
- Over-fitting the training data
  - Loss of prediction due to excess complexity of the model
- Under-fitting the training data
  - Not enough complexity in the model
  - Limited feature selection
- Remedy to valid modeling approach is “testing and validation”