Machine Learning Landscape

Daro VAN

Brought to you by Dynamics and Control Laboratory

Department of Industrial and Mechanical Engineering
Intelligence Mechatronics and Information Technology Research Unit
Institute of Technology of Cambodia

What is Machine Learning?

What exactly does it mean for ML to **learn** something?

If you downloaded a copy of Wikipedia, has your computer really learnt something? Is it suddenly smarter?

We will start by clarifying **what Machine Learning** is and **why** you may want to use it.

What is Machine Learning?

More general definition:

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

-Arthur Samuel, 1959

And more engineering-oriented one:

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

-Tom Mitchell, 1997

What is Machine Learning?

Example

For example, your spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (non-spam, also called "ham") emails. The examples that the system uses to learn are called the ***training set***.

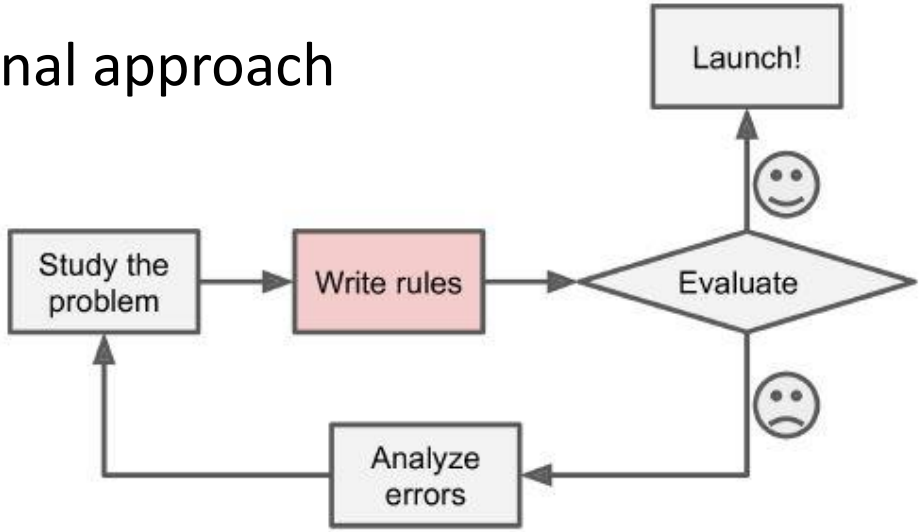
In this case, the task T is to flag spam for new emails, the experience E is the *training data*, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails. This particular performance measure is called ***accuracy*** and it is often used in classification tasks.

What is Machine Learning?
Example

If you just download a copy of Wikipedia, your computer has a lot more data, but it is not suddenly better at any task. Thus, it is not Machine Learning

Why Use Machine Learning?

The Traditional approach



1, Look at what spam typically looks like

2, Write a detection algorithm for each pattern that you noticed(probably long list of complicated rules) .

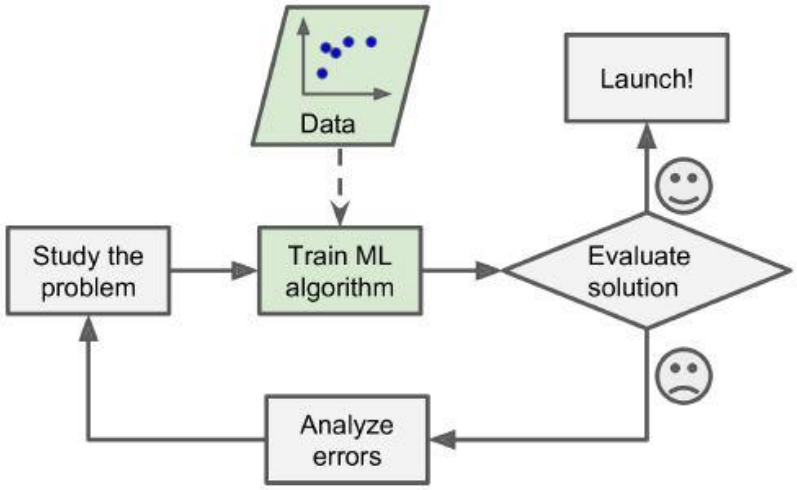
3, Test your program and repeat step 1 and 2 until it is good enough.

Why Use Machine Learning?

Example, if spammers notice that all their emails containing “4U” are blocked, they might start writing “For U” instead. A spam filter using traditional programming techniques would need to be updated to flag “For U” emails. If spammers keep working around your spam filter, you will need to keep writing new rules forever.

Why Use Machine Learning?

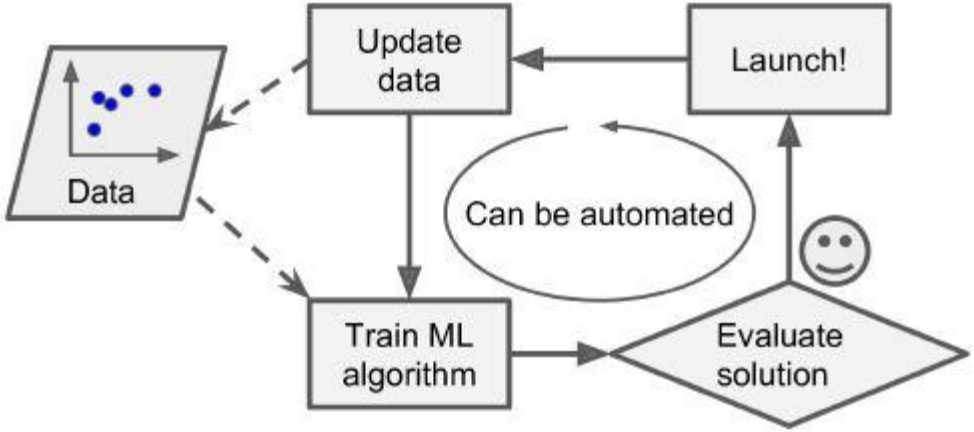
In contrast, a spam filter based on Machine Learning techniques **automatically learns** which words and phrases are good predictors of spam by **detecting unusually frequent patterns** of words in the spam examples compared to the ham examples.



Machine Learning approach

The program is much shorter, easier to maintain, and most likely more accurate.

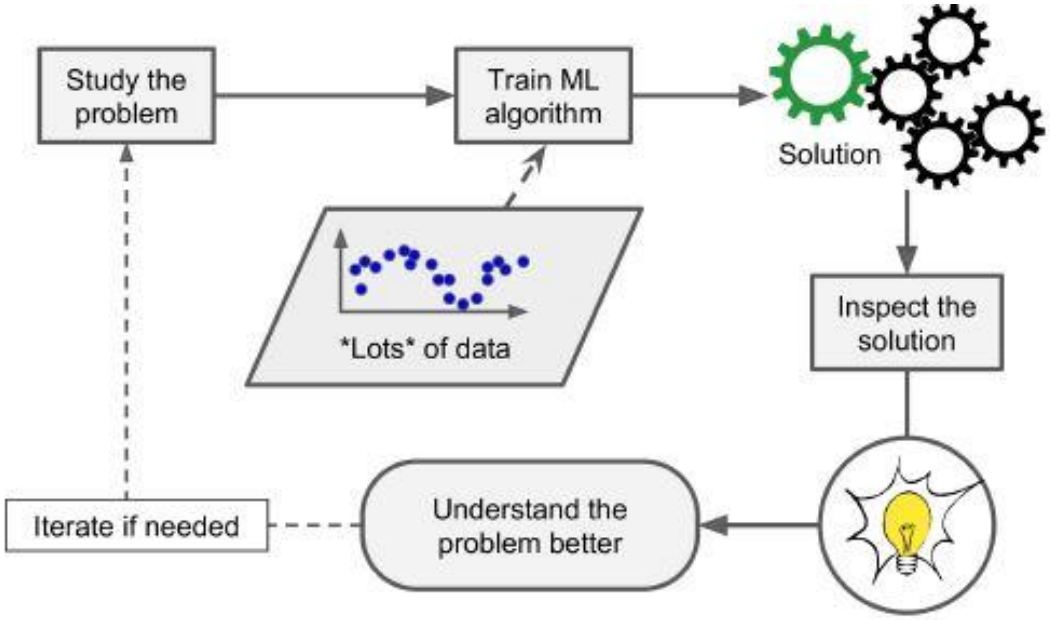
Why Use Machine Learning?



Automatically adapting to change

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 Use Machine Learning?



Machine Learning can help humans learn

What is machine Learning great for?

- 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.**

Types of Machine Learning Systems

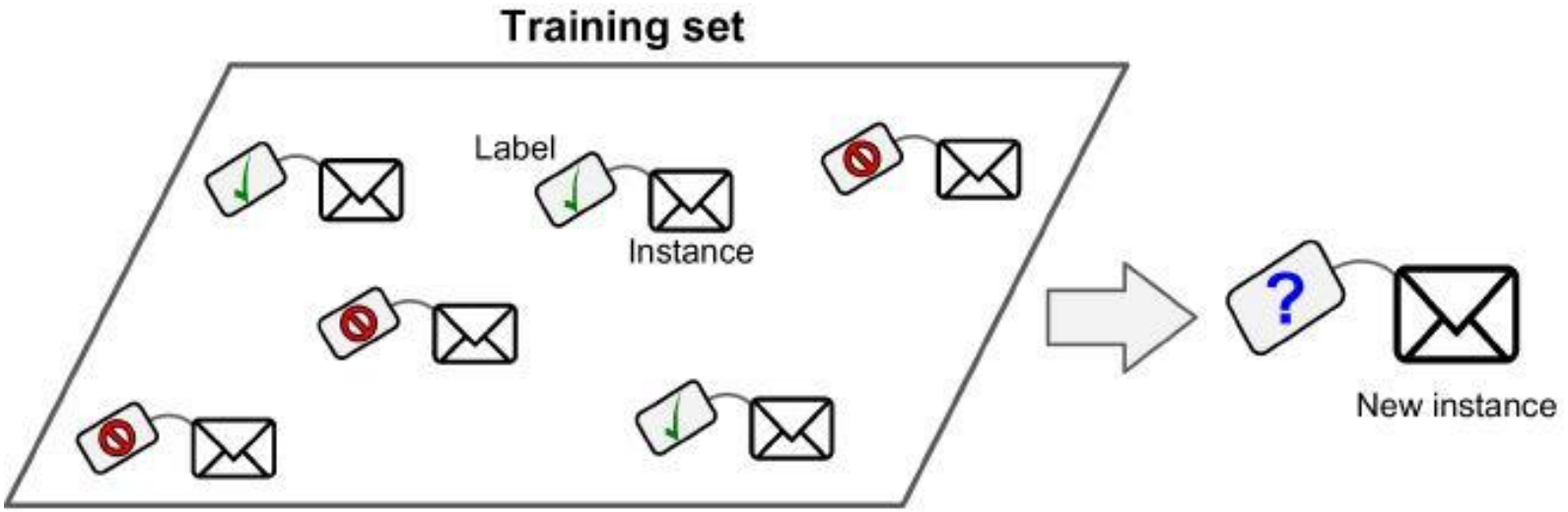
it is useful to classify them in broad categories based on:

- 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**)

Types of Machine Learning Systems

Supervised Learning

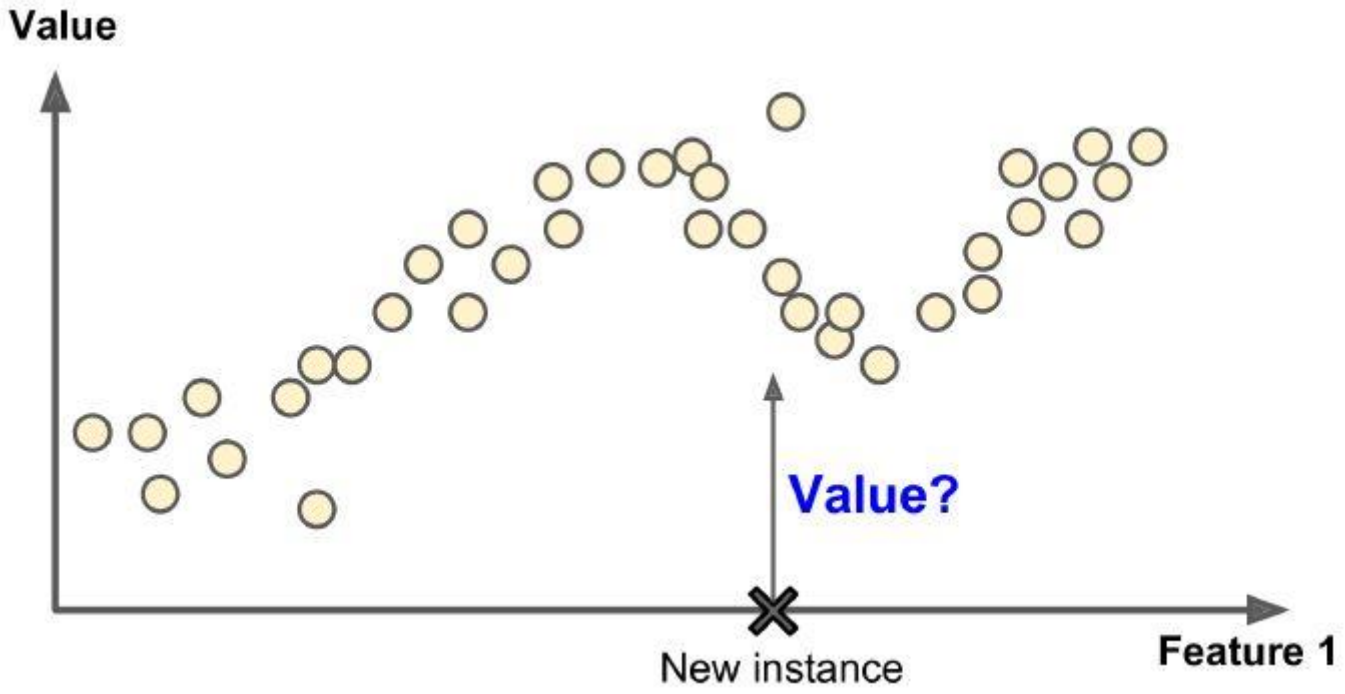
In *supervised learning*, the **training data** you feed to the algorithm includes the desired solutions.



classification

Types of Machine Learning Systems

Supervised Learning



regression

Types of Machine Learning Systems

Supervised Learning

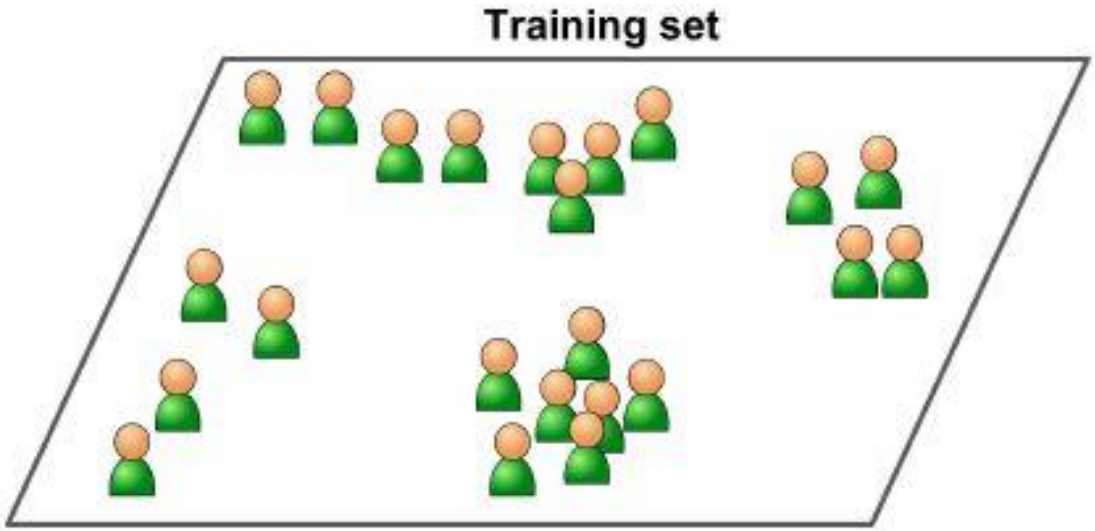
Here are some of the most important supervised learning algorithms:

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

Types of Machine Learning Systems

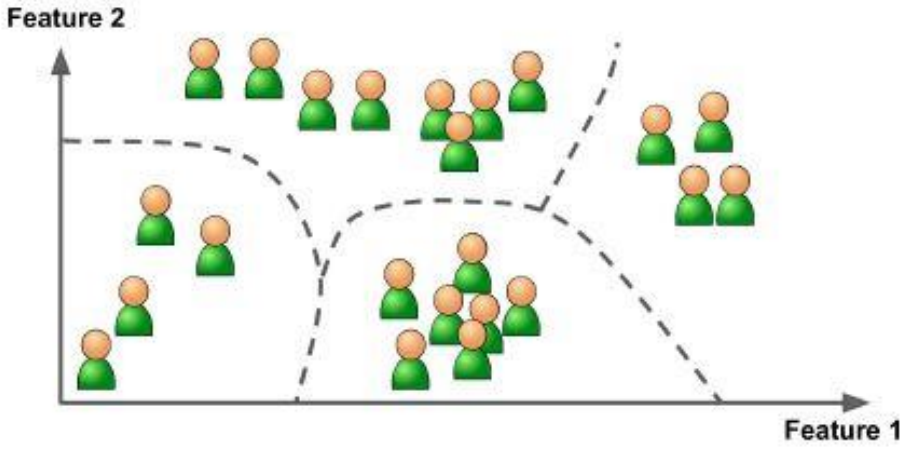
unsupervised Learning

In ***unsupervised learning***, as you might guess, the training data is unlabeled. The system tries to learn without a teacher.



Types of Machine Learning Systems
unsupervised Learning

Here are some of the most important unsupervised learning algorithms:



Clustering

—K-Means

—DBSCAN

—Hierarchical Cluster Analysis (HCA)

Types of Machine Learning Systems

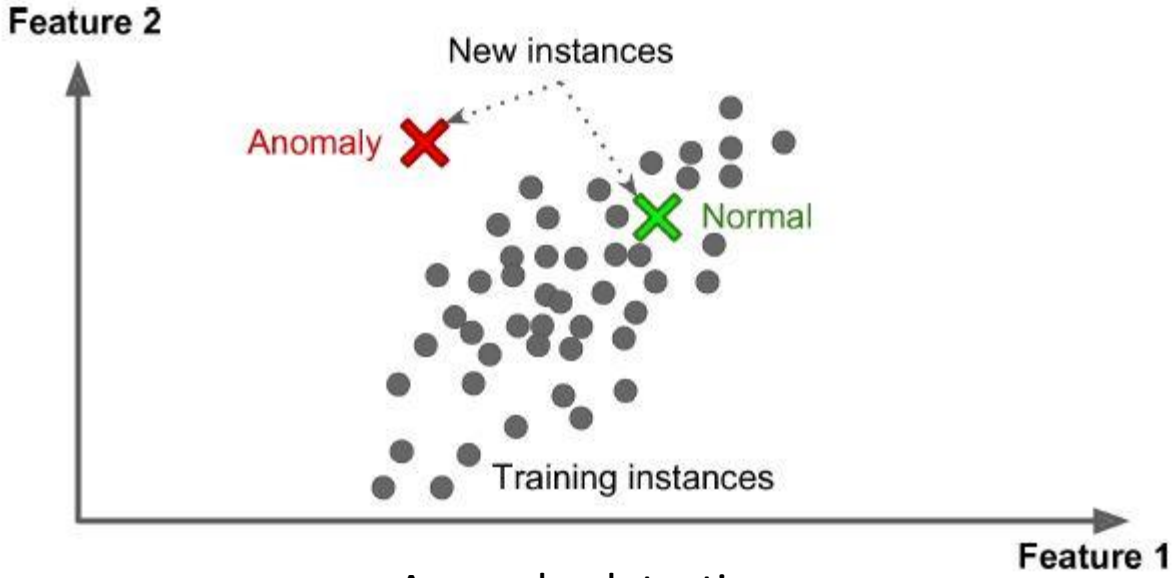
unsupervised Learning

For example, say you have a lot of data about your **blog's visitors**. You may want to run a *clustering* algorithm **to try to detect groups of similar visitors**. At no point do you tell the algorithm which group a visitor belongs to: **it finds those connections without your help.**

It might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on.

Types of Machine Learning Systems
unsupervised Learning

- **Anomaly detection and novelty detection**
  - —One-class SVM
  - —Isolation Forest



Anomaly detection

Types of Machine Learning Systems

unsupervised Learning

For example, detecting unusual credit card transactions to prevent fraud, catching manufacturing defects, or automatically removing outliers from a dataset before feeding it to another learning algorithm.

The system is shown mostly normal instances during training, so it learns to recognize them and when it sees a new instance it can tell whether it looks like a normal one or whether it is likely an anomaly.

Types of Machine Learning Systems

unsupervised Learning

very similar task is ***novelty detection***: the difference is that novelty detection algorithms expect to see only normal data during training.

Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training

Note: for novelty detection, the training data is not polluted by outliers and we are interested in detecting whether a **new** observation is an outlier. In this case, **novelty** means **unusual**.

Types of Machine Learning Systems
unsupervised Learning

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

Types of Machine Learning Systems

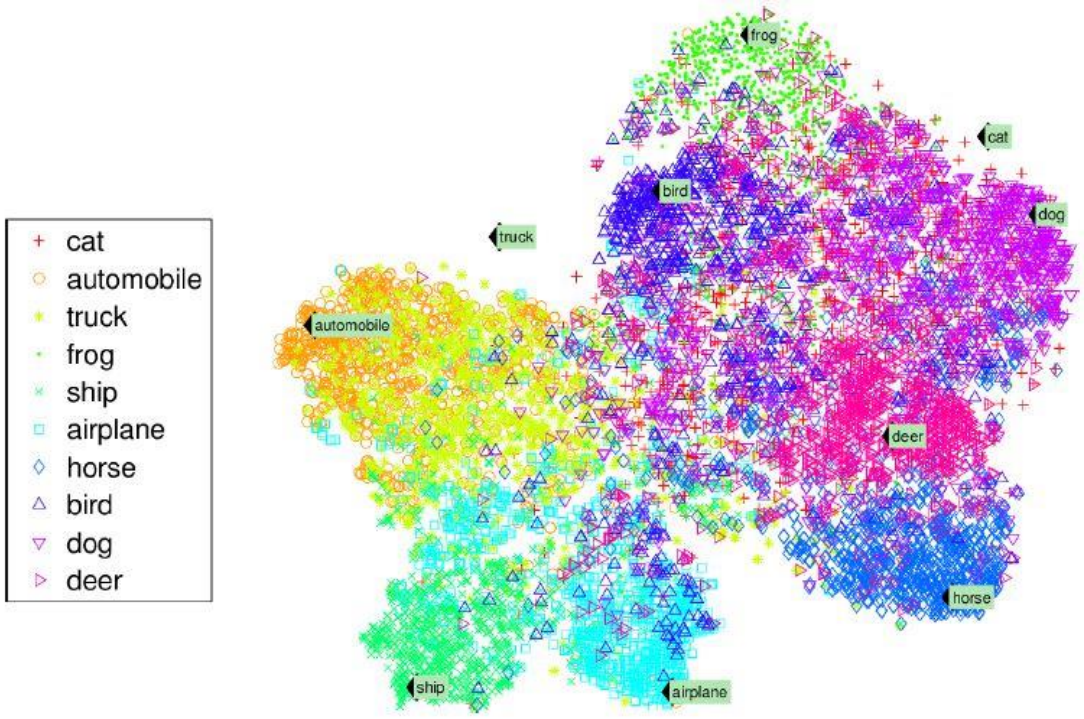
unsupervised Learning

Visualization algorithms are also good examples of unsupervised learning algorithms: **you feed them a lot of complex and unlabeled data**, and they **output a 2D or 3D representation of your data** that can easily be plotted.

These algorithms try to preserve as much structure as they can so you can understand how the data is organized and perhaps identify unsuspected patterns.

Types of Machine Learning Systems

unsupervised Learning



Notice how animals are rather well separated from vehicles, how horses are close to deer but far from birds, and so on.

Types of Machine Learning Systems

unsupervised Learning

It is often a good idea to try to reduce the dimension of your training data using a dimensionality reduction algorithm before you feed it to another Machine Learning algorithm (such as a supervised learning algorithm). It will run much faster, the data will take up less disk and memory space, and in some cases it may also perform better.

Types of Machine Learning Systems

unsupervised Learning

- **Association rule learning**
  - —Apriori
  - —Eclat

Finally, another common unsupervised task is *association rule learning*, in which **the goal is to dig into large amounts of data and discover interesting relations between attributes.**

For example, suppose you own a supermarket. Running an association rule on your sales logs may reveal that people who purchase barbecue sauce and potato chips also tend to buy steak. Thus, you may want to place these items close to each other.

Types of Machine Learning Systems

semi-supervised Learning

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*

Types of Machine Learning Systems

semi-supervised Learning

Some photo-hosting services, such as Google Photos, are good examples of this. 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, while another person B shows up in photos 2, 5, and 7. This is the unsupervised part of the algorithm (clustering).

Now all the system needs is for you to tell it who these people are. Just one label per person, 4 and it is able to name everyone in every photo, which is useful for searching photos.

Types of Machine Learning Systems

Reinforcement Learning

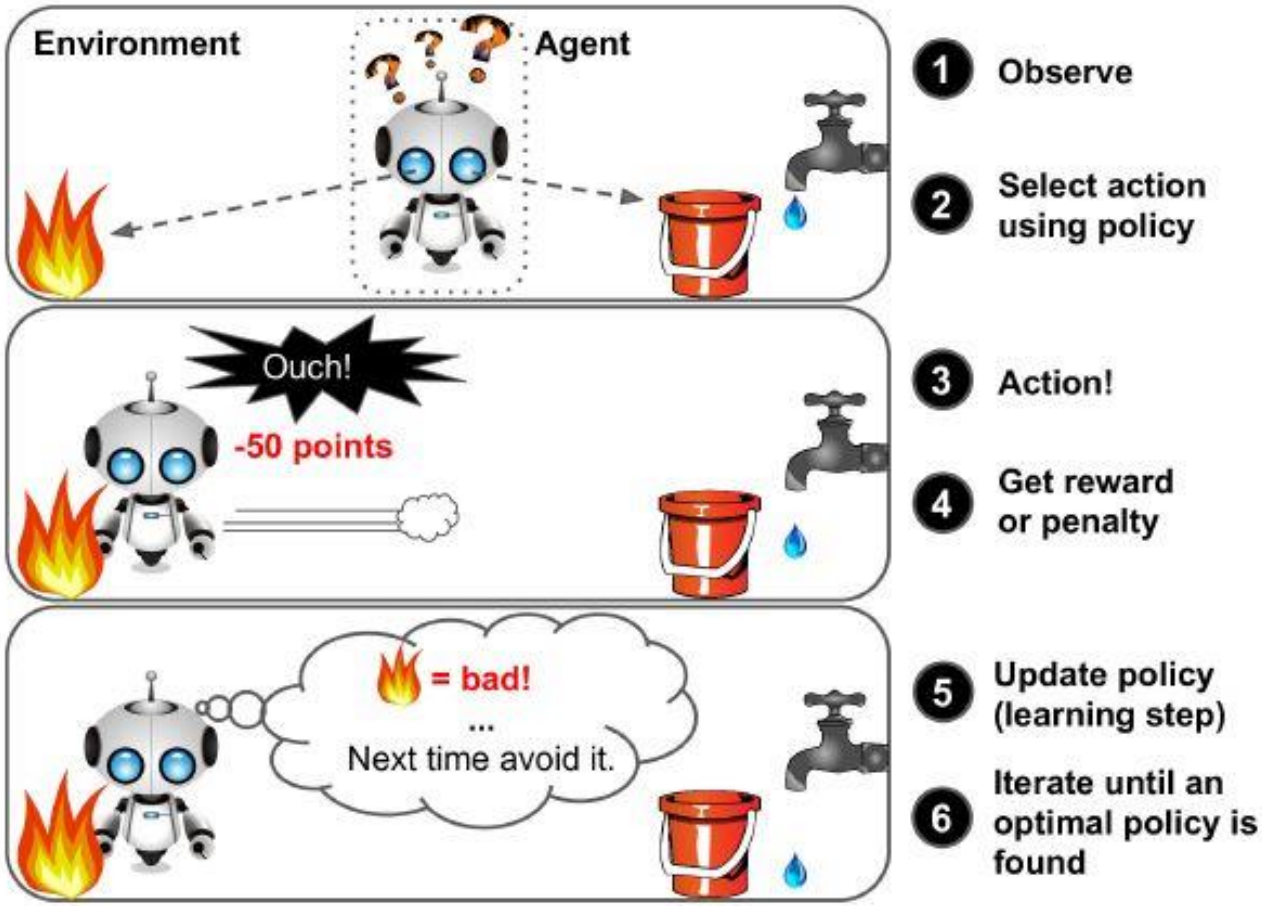
***Reinforcement Learning* is a very different beast!**

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.

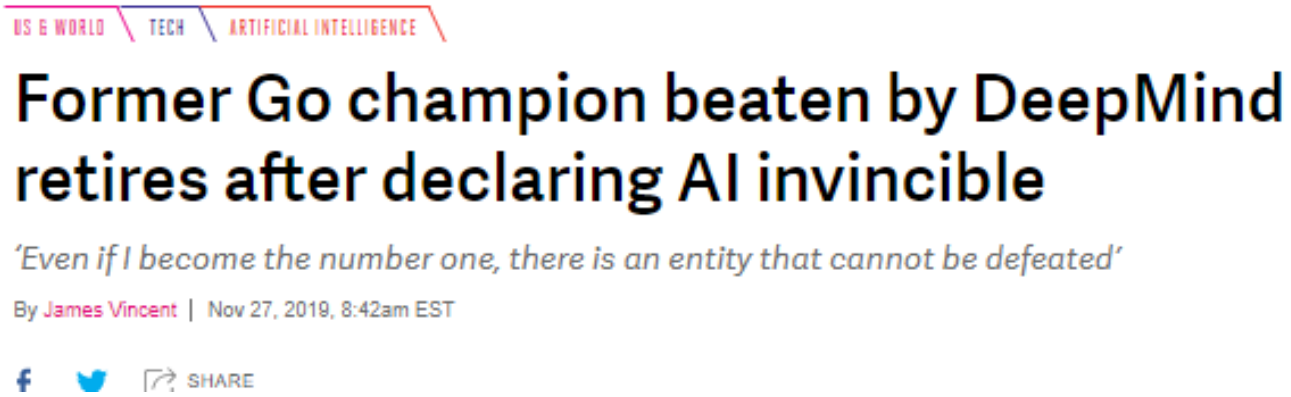
It must then learn by itself what is the best strategy, called a ***policy***, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.

Types of Machine Learning Systems

Reinforcement Learning



Types of Machine Learning Systems
Reinforcement Learning



US & WORLD \ TECH \ ARTIFICIAL INTELLIGENCE

Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is an entity that cannot be defeated'

By James Vincent | Nov 27, 2019, 8:42am EST

SHARE

https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat

Types of Machine Learning Systems

Batch and Online 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***.

Types of Machine Learning Systems

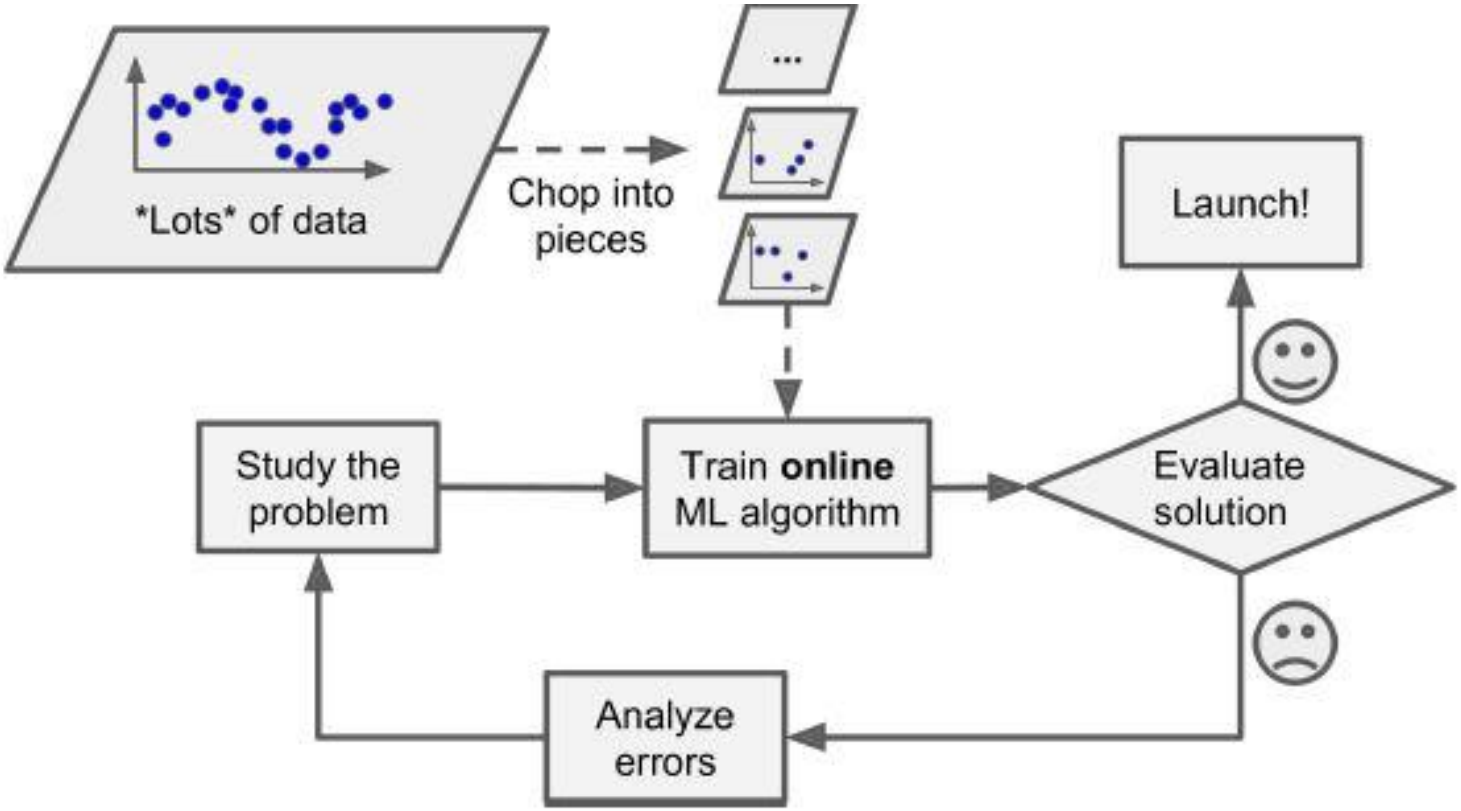
Batch and Online Learning

In ***online learning***, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called *mini-batches*. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.

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

Types of Machine Learning Systems

Batch and Online Learning



Using online learning to handle huge datasets

Types of Machine Learning Systems

Instance-Based and Model based Learning

One more way to categorize Machine Learning systems is by how they ***generalize***.

Most Machine Learning tasks are about **making predictions**. This means that **given a number of training examples**, the system needs **to be able to generalize to examples it has never seen before.**

Types of Machine Learning Systems

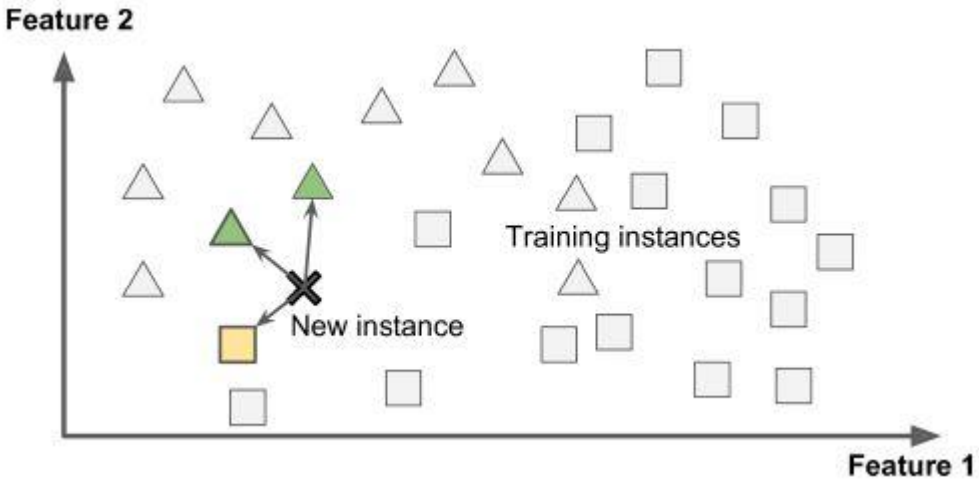
Instance-Based and Model based Learning

Possibly the most trivial form of learning is simply to **learn by heart**. If you were to create a spam filter this way, it would just flag all emails that are identical to emails that have already been flagged by users—not the worst solution, but certainly not the best.

Instance-based learning: The system learns the examples by heart, then generalizes to new cases by comparing them to the learned examples (or a subset of them), using a similarity measure.

Types of Machine Learning Systems

Batch and Online Learning

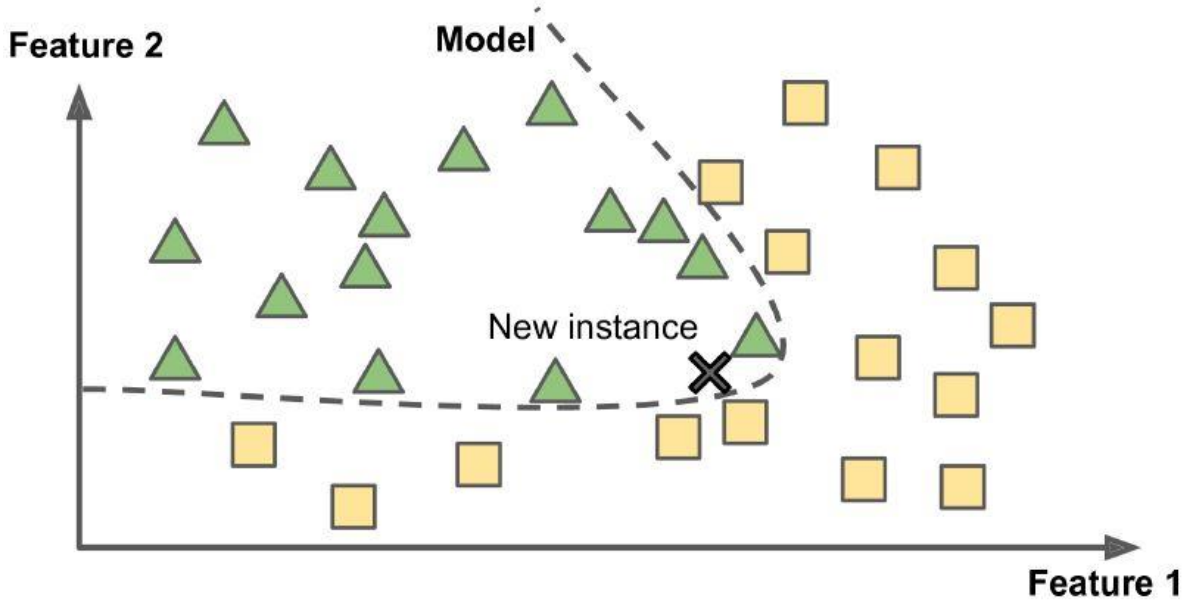


the new instance would be classified as a triangle because the majority of the most similar instances belong to that class.

Types of Machine Learning Systems

Batch and Online 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.*



Main Challenges

In short, since your main task is to select a learning algorithm and train it on some data, the two things that can go wrong are **“bad algorithm”** and **“bad data.”**

Main Challenges

- Insufficient quantity of training data
- Nonrepresentative training data
- Poor-Quality data
- Irrelevant Features
- Overfitting the training data
- Underfitting the training data
- ……

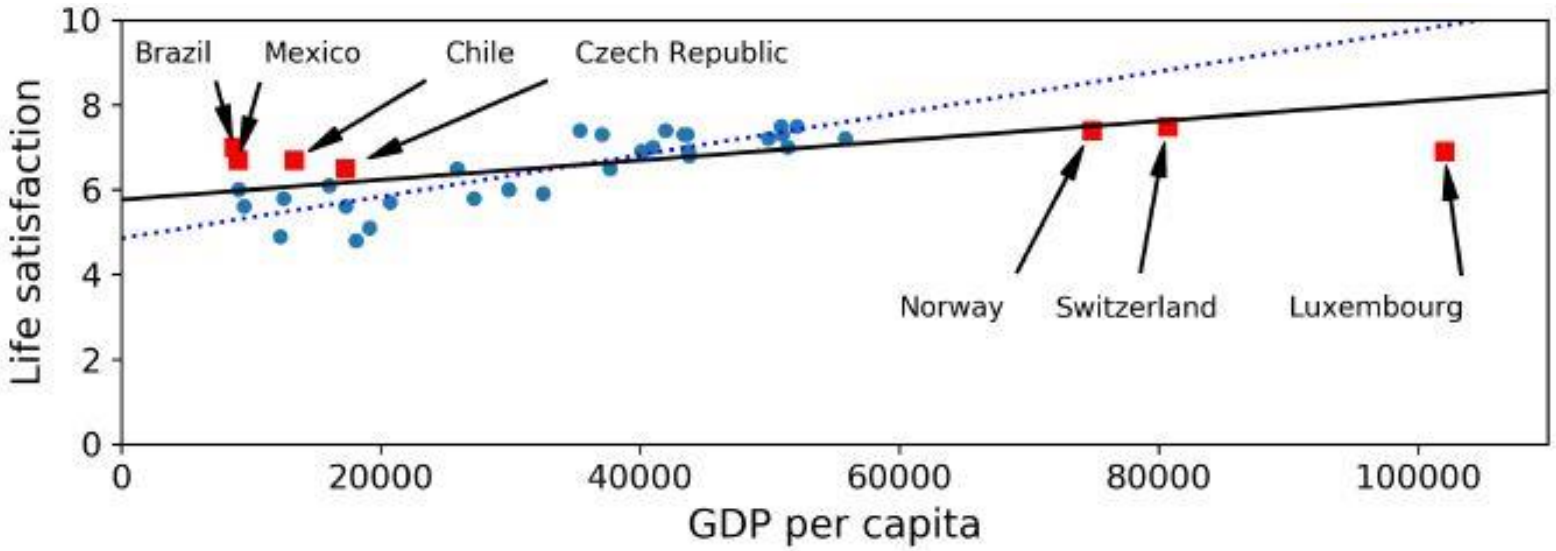
Main Challenges

Insufficient Quantity of Training Data

For a toddler to learn what an apple is, all it takes is for you to point to an apple and say "apple" (possibly repeating this procedure a few times). Now the child is able to recognize apples in all sorts of colors and shapes. **Genius**!

Machine Learning is **not quite there yet**; it takes **a lot of data** for most Machine Learning algorithms to work properly. Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples.

Main Challenges
Nonrepresentative Training Data



If you train a linear model on this data, you get the solid line, while the old model is represented by the dotted line.

Main Challenges

Poor-Quality Data

Obviously, if your training data is **full of errors**, **outliers**, and **noise**, it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.

- If some instances are clearly outliers, it may help to simply discard them or try to fix the errors manually.
- If some instances are missing a few features (e.g., 5% of your customers did not specify their age), you must decide whether you want to ignore this attribute altogether, ignore these instances, fill in the missing values , or train one model with the feature and one model without it, and so on.

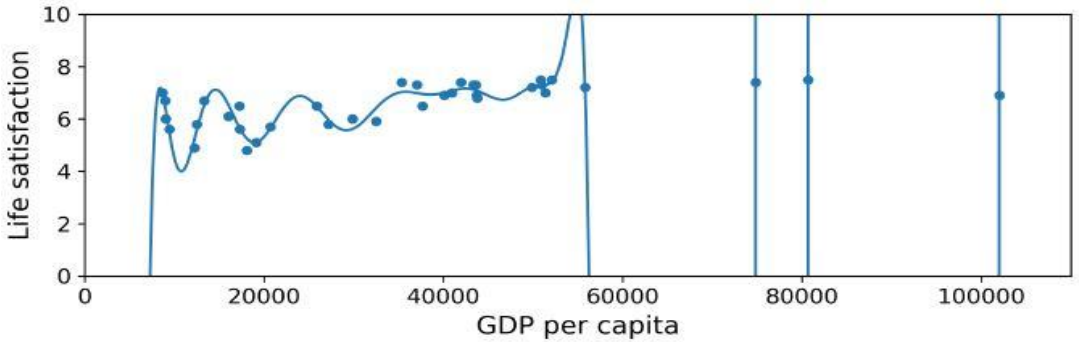
Main Challenges

Irrelevant Features

Your system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones.

- ***Feature selection***: selecting the most useful features to train on among existing features.
- ***Feature extraction***: combining existing features to produce a more useful one.
- **Creating new features** by gathering new data.

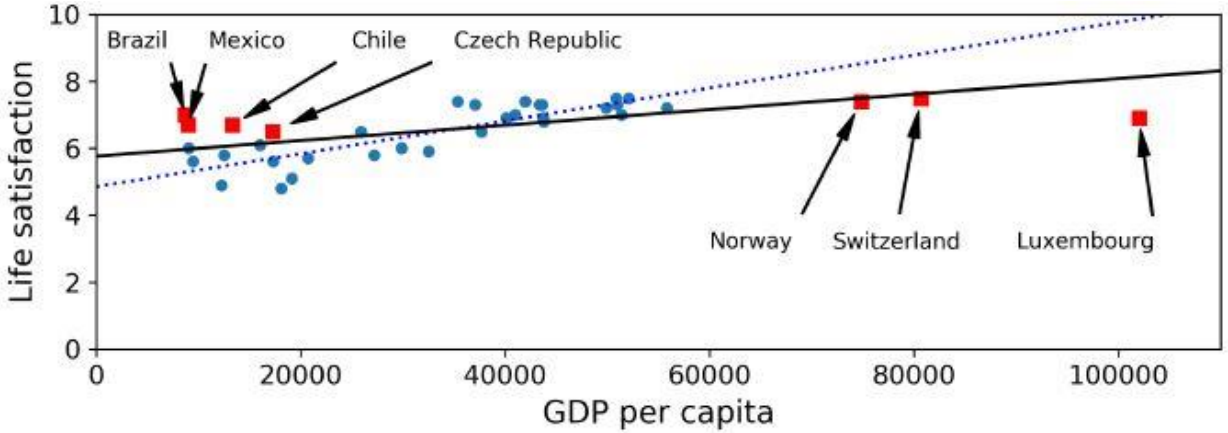
Main Challenge
Overfitting the Training Data



- **Simplify the model** by selecting one with fewer parameters
- **Gather more training data**
- **Reduce the noise in the training data** (e.g., fix data errors and remove outliers)

Main challenges

Underfitting the Training Data



The main options to fix this problem are:

- **Selecting a more powerful model**, with more parameters
- **Feeding better features** to the learning algorithm (feature engineering)

References

- **Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow by Aurelien Geron, OREILLY,2019**