Machine learning landscape, build a machine learning project

What is Machine Learning?

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*

- Examples: spam filters, recommendation system， customer segmentation

Why use machine learning?

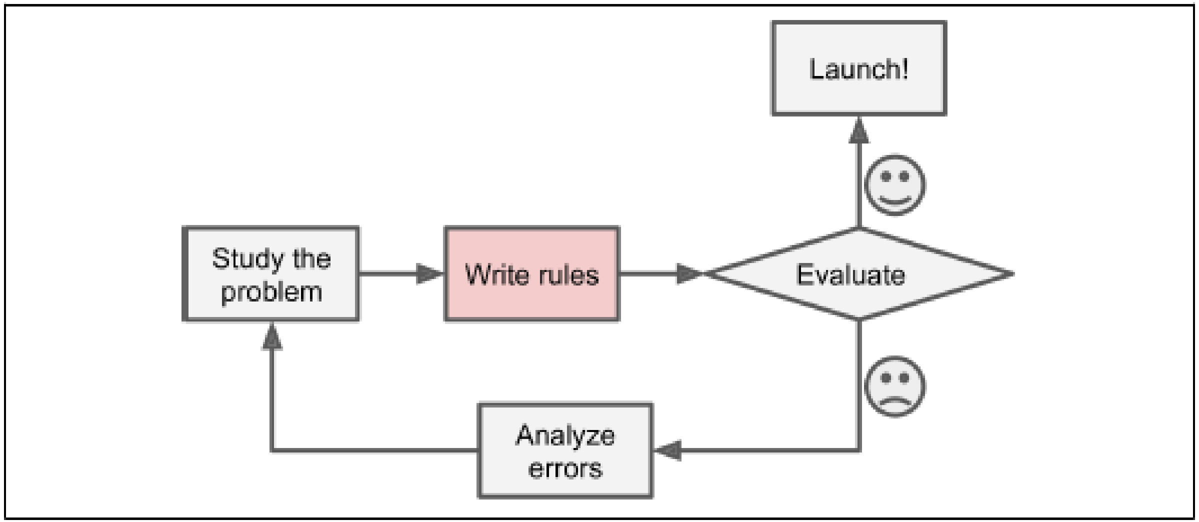

Figure 1-1. The traditional approach

- Complex rules

Why use machine learning?

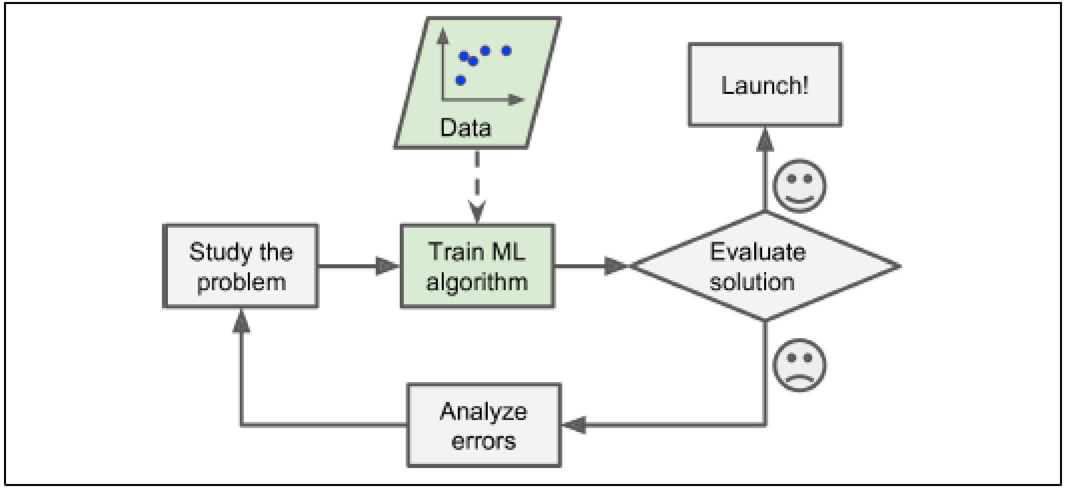

Figure 1-2. Machine Learning approach

- Machine Learning techniques automatically learns which words and phrases are good predictors of spam

Why use machine learning?

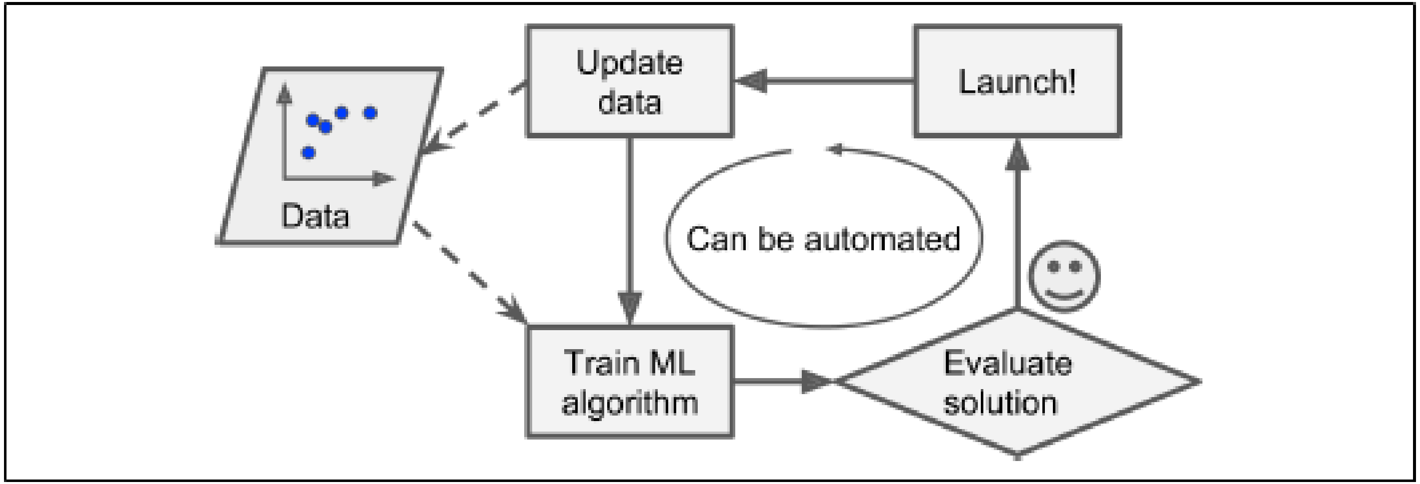

Figure 1-3. Automatically adapting to change

- 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?

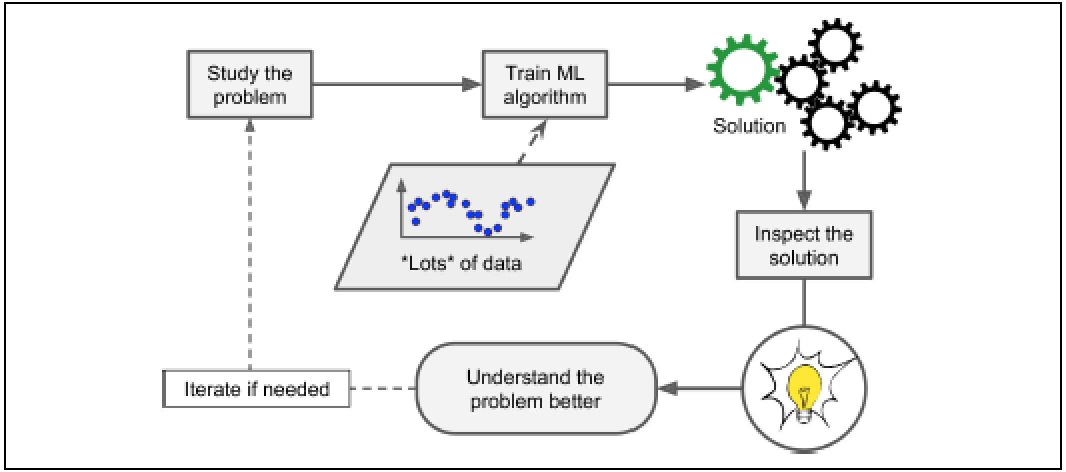

Figure 1-4. Machine Learning can help humans learn

- Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent.

Why use machine learning?

- 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

- Supervised / Unsupervised learning / Semisupervised learning / Reinforcement learning /
  - Whether or not the systems are trained with human supervision
  - Whether or not the data has labels

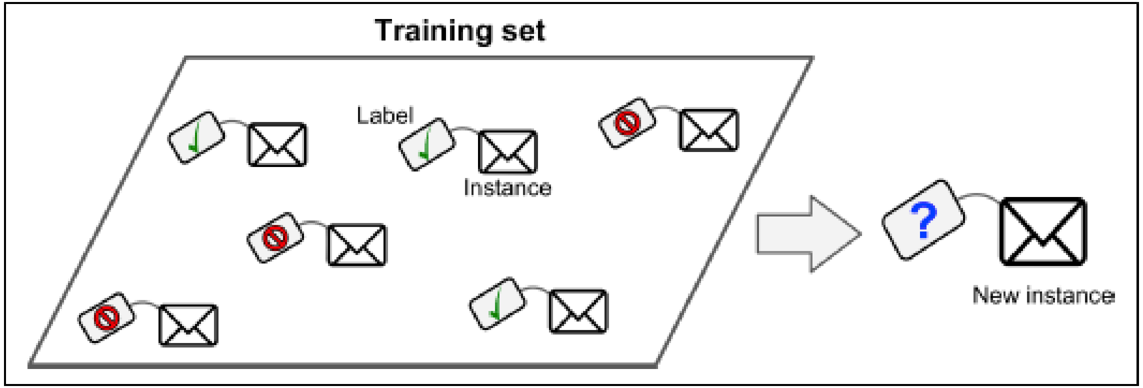

Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

Supervised learning algorithms

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

Unsupervised learning algorithms

- Data has no labels

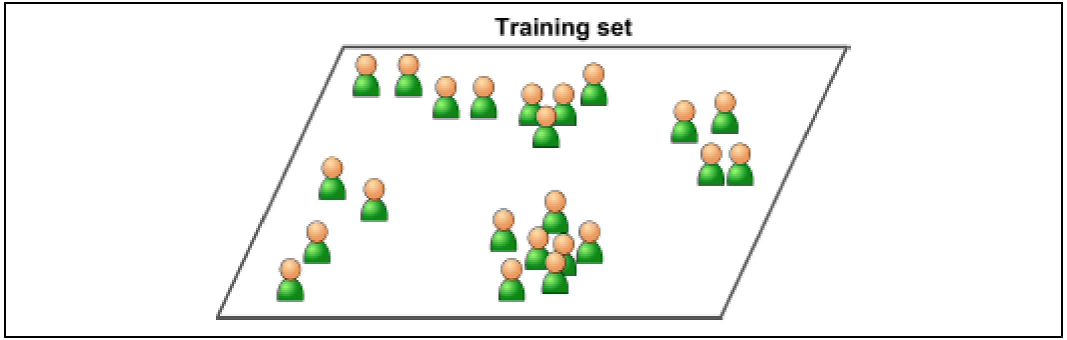

Figure 1-7. An unlabeled training set for unsupervised learning

Unsupervised learning algorithms

- **Clustering**
  - K-Means
  - DBSCAN
  - Hierarchical Cluster Analysis (HCA)
- **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

Unsupervised learning algorithms

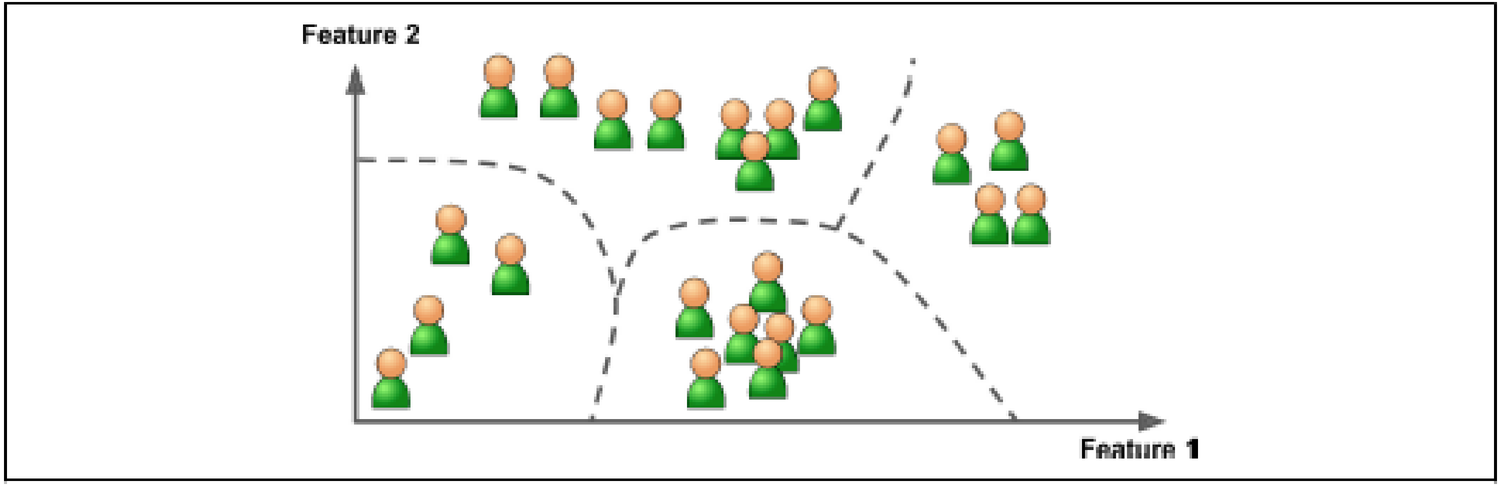

Figure 1-8. Clustering

Semisupervised learning algorithms

- Some data has labels

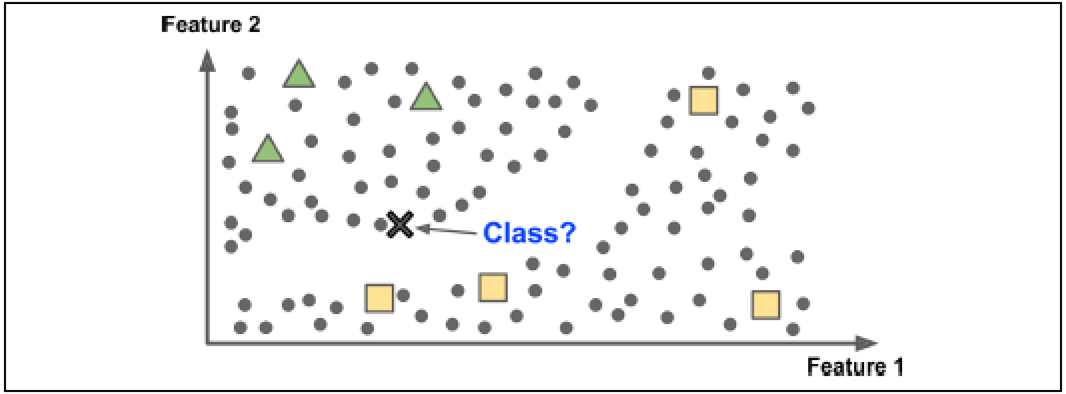

Figure 1-11. Semisupervised learning

Reinforcement learning

- Agent select and perform actions, and get *rewards* in return (or *penalties* in the form of negative rewards)

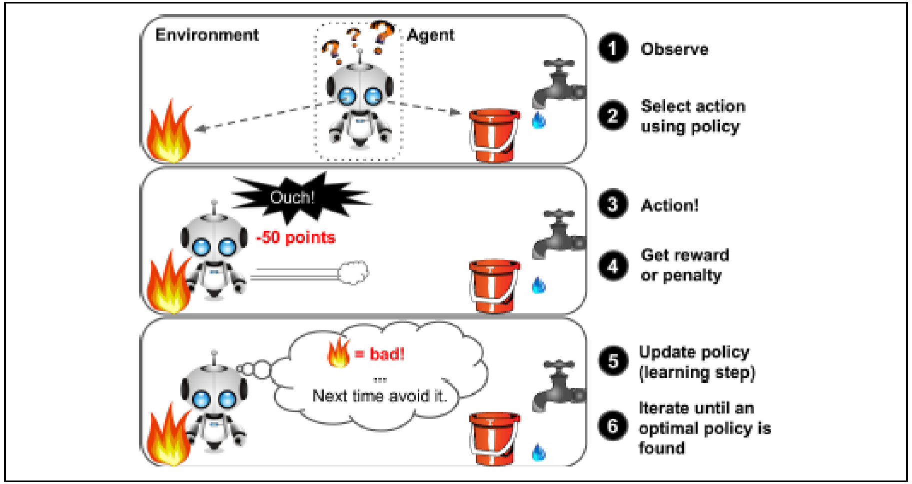

Figure 1-12. Reinforcement Learning

Batch learning vs. online learning

- Another criterion used to classify Machine Learning systems is whether or not the system can learn incrementally from a stream of incoming data
  - Batch learning
  - Online learning

Instance-based vs. model-based learning

- One more way to categorize Machine Learning systems is by how they *generalize*.
  - Instance-based learning
    - ✓ This is called *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.
  - 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.

Main challenges of Machine Learning

- Insufficient quantity of training data

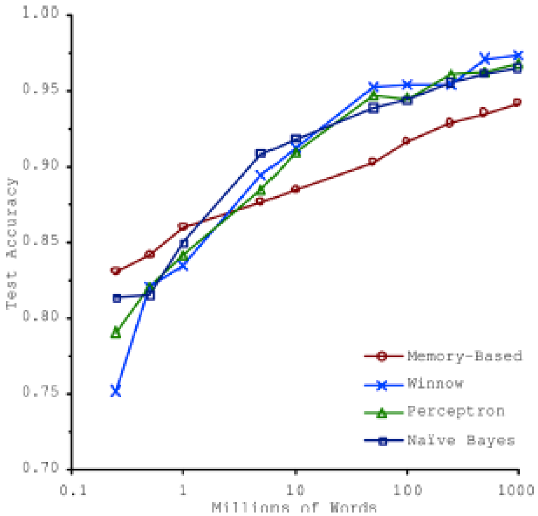

Figure 1-20. The importance of data versus algorithms[9]

Main challenges of Machine Learning

- Non-representative training data

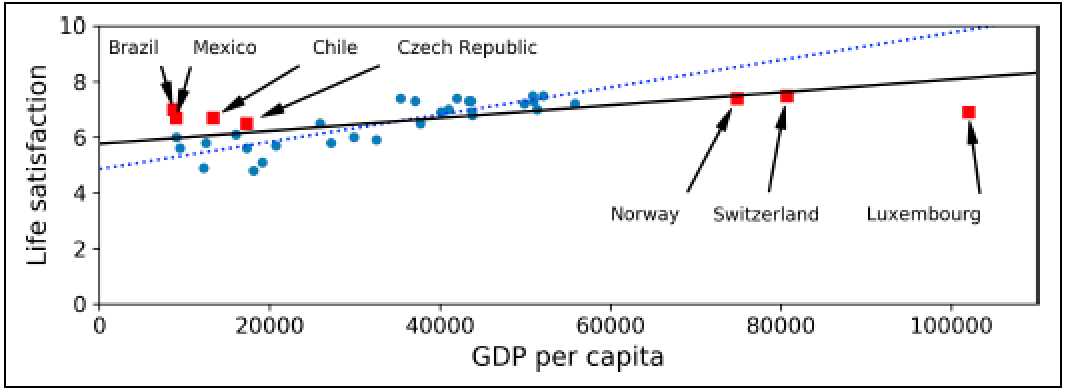

Figure 1-21. A more representative training sample

Main challenges of Machine Learning

- Insufficient quantity of training data
- Non-representative training data
- Poor-quality of data: errors, outliers, and noise
- Irrelevant features: garbage in, garbage out
- Overfitting the training data
- Underfitting the training data

Main challenges of Machine Learning

- Overfitting the training data

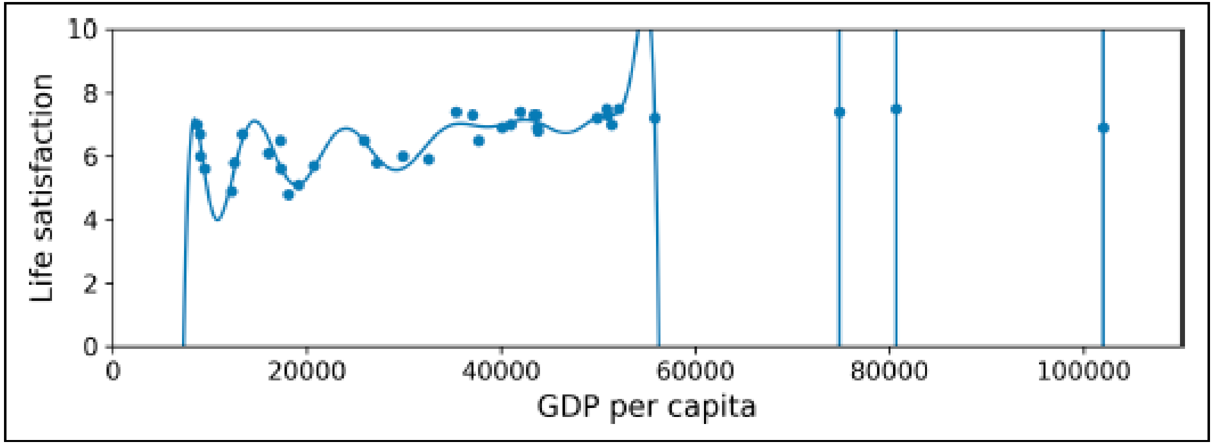

Figure 1-22. Overfitting the training data

Main challenges of Machine Learning

- Overfitting the training data
  - *Regularization reduces the risk of overfitting*
  - *Adjusting hyper-parameter*

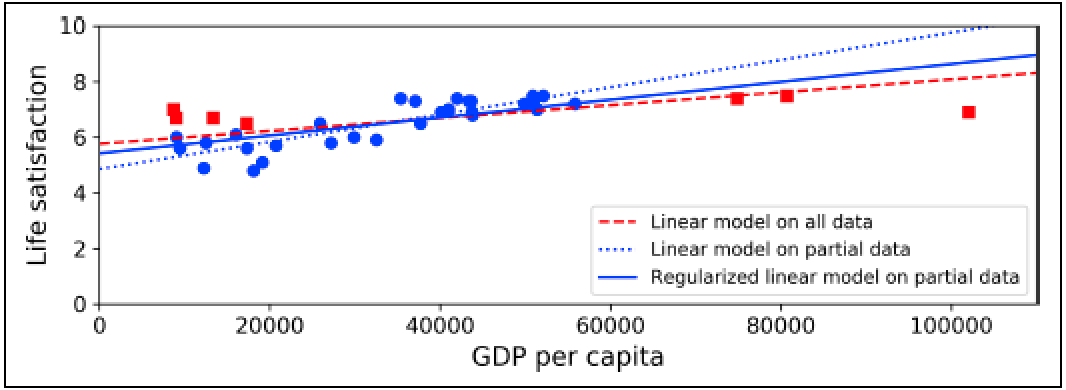

Figure 1-23. Regularization reduces the risk of overfitting

Main challenges of Machine Learning

- Underfitting the training data: it occurs when your model is to simple to learn the underlying structure of the data
  - Selecting a more powerful model, with more parameters
  - Feeding better features to the learning algorithm (feature engineering)
  - Reducing the constraints on the model (e.g., reducing the regularization hyperparameter)

Testing and validating

- Training set vs. test set (50% vs. 50%, 80% vs 20%, 70% vs 30%)
- Generalization error (or out-of-sample error) on unseen data
- Hyperparameter tuning and model selection
- No free lunch theorem
  - In a famous 1996 paper, David Wolpert demonstrated that if you make absolutely no assumption about the data, then there is no reason to prefer one model over any other. This is called the *No Free Lunch* (NFL) theorem. For some datasets the best model is a linear model, while for other datasets it is a neural network. There is no model that is *a priori* guaranteed to work better (hence the name of the theorem).