# Introduction to Machine Learning

# Overview

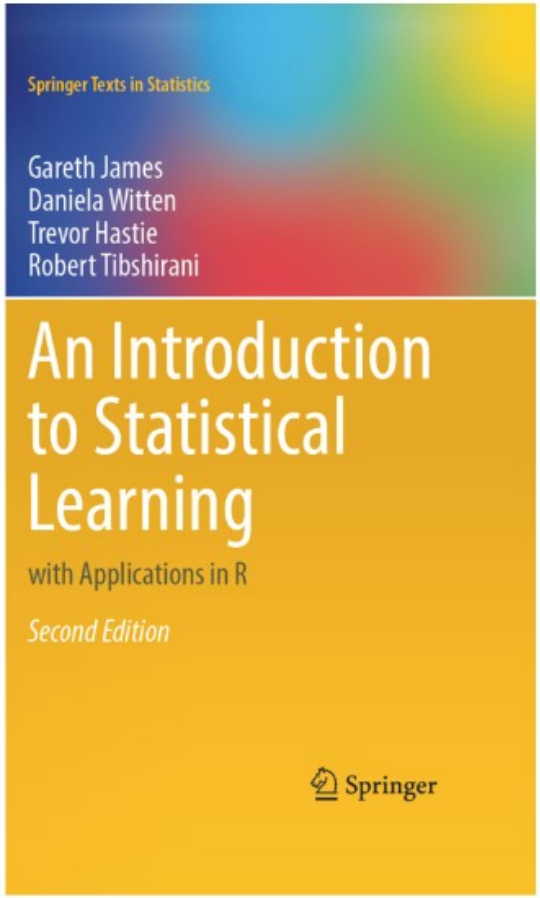
## First lecture, 05.01.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

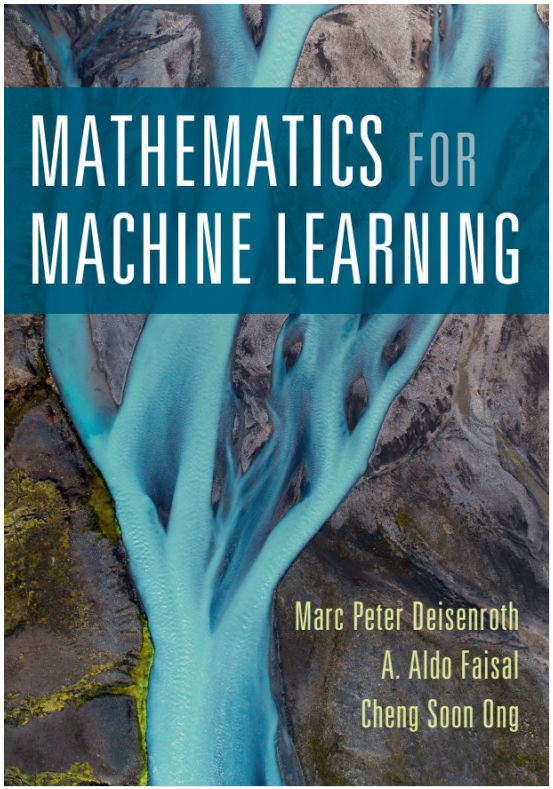
# Organization of the lecture

- Introduction to Machine Learning: 05.01.2022 to 06.04.2022 (12 weeks + 2 weeks)
- Prerequisites:
  - Basic programming skills with R and Python
  - Linear Algebra
  - Statistics
- Exercises:
  - weekly exercises, theoretical and practical work (roughly alternating),
  - practical exercises will be in R or Python,
  - 50% of the points in the exercises are needed to take part in the exams.
- Final Exam:
  - presenting the course project (13.04.2022 at 3pm)
  - script
- Grading: passed if you get at least 50% of the points.

# Books, slides, exercises and solutions to exercises








https://www.statlearning.com/
https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf
https://www.dataschool.io/15-hours-of-expert-machine-learning-videos/
https://blog.princehonest.com/stat-learning/ (solutions)

https://machinelearningcoban.com

# Roadmap of the lecture

- Introduction to machine learning
- Review of statistics
- Review of linear algebra and information theory
- Review of convex optimization
- Supervised learning
  - Bayesian decision theory
  - Linear methods for regression and classification
  - Support Vector Machine (SVM) and Kernel methods
  - Evaluation/Comparison of classifiers, Model selection
  - Boosting and decision trees, random forest, prototype methods
  - Bayseian Naive
  - Neural networks
  - Fuzzy Logic
  - Feature selection (!Feature extraction)
- Semi-supervised learning
- Unsupervised learning
  - Clustering (Kmean, hierarchical clustering, spetral clustering)
  - Dimensionality Reduction (PCA, tSNE, UMAP, MDS)

# Projects

| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 1 | (Simple and Multiple) Linear Regression | Thanh Giang | https://web.stanford.edu/~hastie/MOOC-Slides/linear_regression.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBSSCPANhTgrw82ws7w_or9 |
| 2.1 | Classification: Logistic regression | Thien | https://web.stanford.edu/~hastie/MOOC-Slides/classification.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IC4vaDba5ClatUmFppXLAhE |
| 2.2 | Classification: Linear Discriminant Analysis (LDA) | Thien | https://web.stanford.edu/~hastie/MOOC-Slides/classification.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IC4vaDba5ClatUmFppXLAhE |
| 2.3 | Naive Bayes Classifier | Thien | https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ https://www.youtube.com/watch?v=vz_xuxYS2PM https://machinelearningcoban.com/2017/08/08/nbc/ https://www.youtube.com/watch?v=xYqiIjaqydU&list=PLBv09BD7ez_7-4V3IJIzCHWQj9nd4rVWB | https://www.youtube.com/watch?v=XzSlEA4ck2I https://www.youtube.com/watch?v=os-NaA0ldGs&list=PLBv09BD7ez_6CxkuiFTbL3jsn2Qd1IU7B&index=1 |
| 2.4 | ROC curve, AUC, RMSE ... | Giang | https://medium.datadriveninvestor.com/evaluation-metrics-101-7c8b4c3421c2 https://medium.com/@xaviergeerinck/artificial-intelligence-how-to-measure-performance-accuracy-precision-recall-f1-roc-rmse-611d10e4caac | https://onionesquereality.wordpress.com/2010/11/07/auc-versus-rmse/ |
| 3.1 | Resampling Methods: Cross validation | Hoang, Cuong | https://web.stanford.edu/~hastie/MOOC-Slides/cv_boot.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IA6E6ZNXu7dp89_uv8yocmf |
| 3.2 | Resampling Methods: Bootstrap | Hoang, Cuong | https://web.stanford.edu/~hastie/MOOC-Slides/cv_boot.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IA6E6ZNXu7dp89_uv8yocmf |

# Projects

| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 4.1 | Linear Model Selection and Best Subset Selection, Forward Stepwise Selection, Backward Stepwise Selection, Estimating Test Error Using Mallow's Cp, AIC, BIC, Adjusted R-squared, Estimating Test Error Using Cross-Validation | Tam | https://web.stanford.edu/~hastie/MOOC-Slides/model_selection.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB-Xdpj_uXJpLGiRfv9UVXI |
| 4.2 | Shrinkage Methods, Ridge Regression and The Lasso | Tam | https://web.stanford.edu/~hastie/MOOC-Slides/model_selection.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB-Xdpj_uXJpLGiRfv9UVXI |
| 4.3 | Tuning Parameter Selection for Ridge Regression and Lasso | Tam | https://web.stanford.edu/~hastie/MOOC-Slides/model_selection.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB-Xdpj_uXJpLGiRfv9UVXI |
| 4.4 | Principal Components Regression and Partial Least Squares | Thong | https://web.stanford.edu/~hastie/MOOC-Slides/model_selection.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB-Xdpj_uXJpLGiRfv9UVXI |
| 5.1 | Polynomial Regression and Step Functions | Thinh | https://web.stanford.edu/~hastie/MOOC-Slides/nonlinear.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBn84fvhh-u2MU80jvo8OoR |
| 5.2 | Piecewise Polynomials and Splines | Thinh | https://web.stanford.edu/~hastie/MOOC-Slides/nonlinear.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBn84fvhh-u2MU80jvo8OoR |
| 5.3 | Smoothing Splines | Thinh | https://web.stanford.edu/~hastie/MOOC-Slides/nonlinear.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBn84fvhh-u2MU80jvo8OoR |

# Projects

| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 6.1 | Decision Trees | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |
| 6.2 | Pruning a Decision Tree | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |
| 6.3 | Classification Trees and Comparison with Linear Models | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |
| 6.4 | Bootstrap Aggregation (Bagging) and Random Forests | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |
| 6.5 | Boosting | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |
| 6.6 | Variable Importance | Truong, Phu | https://web.stanford.edu/~hastie/MOOC-Slides/trees.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh |

# Projects

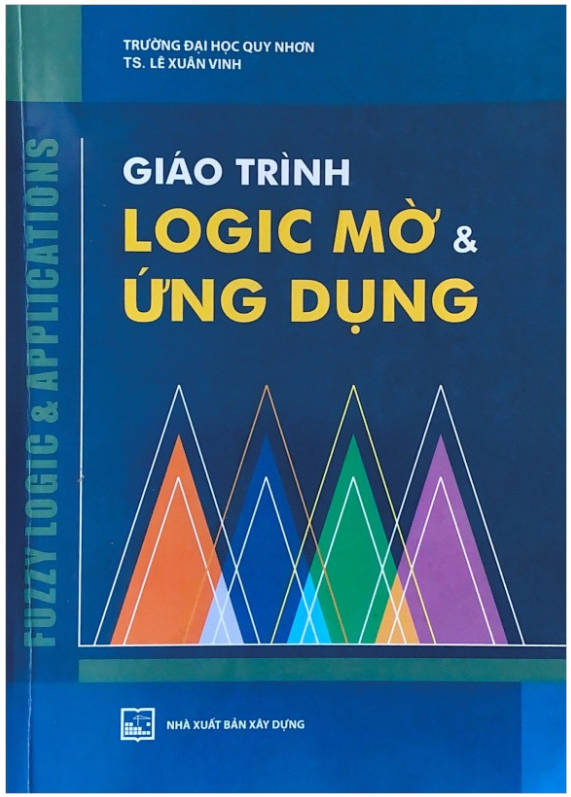
| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 7.1 | Maximal Margin Classifier | Bac, Giang | https://web.stanford.edu/~hastie/MOOC-Slides/svm.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IDl6MkmovVdZwyYOhpCxo5o |
| 7.2 | Support Vector Classifier | Bac, Giang | https://web.stanford.edu/~hastie/MOOC-Slides/svm.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IDl6MkmovVdZwyYOhpCxo5o |
| 7.3 | Support Vector Machines (*/MATH) | Bac, Giang | https://web.stanford.edu/~hastie/MOOC-Slides/svm.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IDl6MkmovVdZwyYOhpCxo5o |
| 7.4 | Kernels (*/HEIN) | Bac, Giang | https://web.stanford.edu/~hastie/MOOC-Slides/svm.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IDl6MkmovVdZwyYOhpCxo5o |
| 7.5 | Example and Comparison with Logistic Regression | Bac, Giang, Thien | https://web.stanford.edu/~hastie/MOOC-Slides/svm.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IDl6MkmovVdZwyYOhpCxo5o |

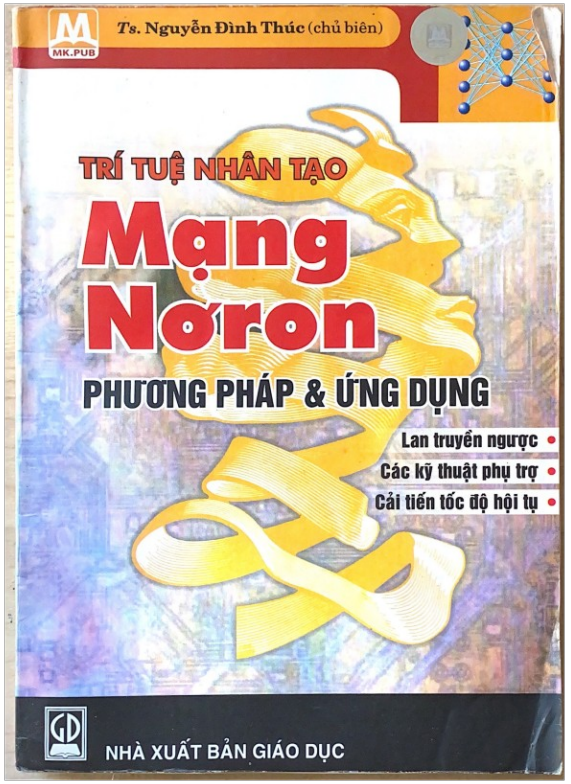
# Projects

| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 8.1 | Unsupervised Learning and Principal Components Analysis (*/Maths) | Hoang, Cuong | https://web.stanford.edu/~hastie/MOOC-Slides/unsupervised.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBC-MneTc9oBZz0C6kNJ-f2 https://www.youtube.com/watch?v=IbE0tbjy6JQ&list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM |
| 8.2 | Exploring Principal Components Analysis and Proportion of Variance Explained | Hoang, Cuong | https://web.stanford.edu/~hastie/MOOC-Slides/unsupervised.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBC-MneTc9oBZz0C6kNJ-f2 |
| 8.3 | K-means Clustering | Minh, Xuan | https://web.stanford.edu/~hastie/MOOC-Slides/unsupervised.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBC-MneTc9oBZz0C6kNJ-f2 https://www.youtube.com/watch?v=mHl5P-qlnCQ&list=PLBv09BD7ez_6cgkSUAqBXENXEhCkb_2wl |
| 8.4 | Hierarchical Clustering | Minh, Xuan | https://web.stanford.edu/~hastie/MOOC-Slides/unsupervised.pdf | https://www.youtube.com/playlist?list=PL5-da3qGB5IBC-MneTc9oBZz0C6kNJ-f2 https://www.youtube.com/watch?v=GVz6Y8r5AkY&list=PLBv09BD7ez_7qIbBhyQDr-LAKWUeycZtx |
| 8.5 | Spectral Clustering (*/Maths) | Minh, Xuan | | https://www.youtube.com/watch?v=zLuVrqlYKyg&list=PLHXZdjnADb1Q6ol2UEuPMnXpQY-KzCmBy |
| 8.6 | Expectation Maximization Algorithms (EM) | Minh, Xuan | | https://www.youtube.com/watch?v=REypj2sy_5U&list=PLBv09BD7ez_4e9LtmK626Evn1ion6ynrt |
| 8.7 | Hidden Markov Model Algorithms (HMM) | Minh | https://www.youtube.com/watch?v=i3AkTO9HLXo https://www.youtube.com/watch?v=c6OrSKsH_gg https://www.youtube.com/watch?v=3FlVNdck3NU | https://www.youtube.com/watch?v=kqSzLo9fenk&t=93s https://www.youtube.com/watch?v=WT6jI8UgROI |

# Projects

| No | Project | Students | Slide/Book | Video |
|---|---|---|---|---|
| 9 | Fuzzy Logic | Giang | | https://www.youtube.com/watch?v=a2i-lHS-c_I&list=PLIY8eNdw5tW9ZqgI9nfXxr6r-FHnLS90k |
| 10 | Neural Network (a bit introduction to DEEP LEARNING) | Thong | | https://www.youtube.com/watch?v=IbE0tbjy6JQ&list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM |



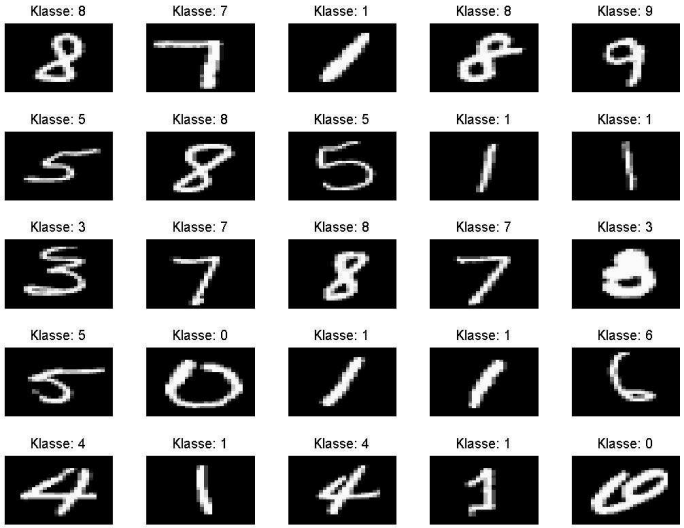



# Roadmap for today

- What is machine learning ?
- Inductive (qui nạp) Inference and machine learning
- Types of learning
- Statistical learning
- Discriminative versus generative learning
- Challenges:
  - Curse of dimensionality and over- versus underfitting
- Is there a best learning algorithm ?

# What is machine learning ?

Learning the terminology of machine learning with an example:




**input** pixel representation of the digits (image is in $\mathrm{R}^{28\times 28}$)

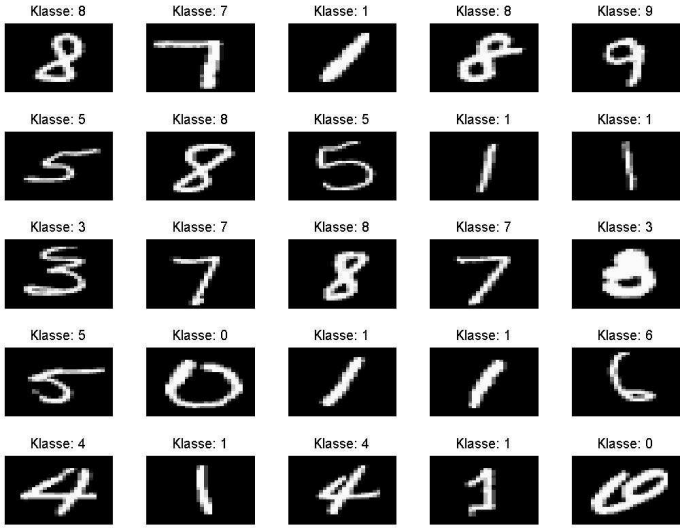
**feature** one specific property of the input (single dimension of the input) class

**output** label (ten digits) in $\{0, 1, \ldots, 9\}$ $\Longrightarrow$ multi-class problem

**classifier** a function from input to output, that is $f : \mathrm{R}^{784} \rightarrow \{0, 1, \ldots, 9\}$.

# What is machine learning ?

Learning the terminology of machine learning with an example:




**training testing** construction of the classifier (usually optimization problem) count

**generalization** errors on unseen cases

**model** classifier predicts well on unseen cases

parameterized function class in which the classifier is chosen

# Inference I

In the natural sciences, we are doing inference.

We differentiate between two types:

- Inductive (quy nap) inference: Learning general principles from observations.
- Deductive (dien dich) inference: Deriving specific assertions from general principles.

# Inference II

## Inductive inference

Inductive inference is at the heart of all natural sciences.

## General steps

1. Collecting observations/data.
2. Construction of a model.
3. Prediction.

## Falsification

Inductive results can only be falsified but not verified !

Machine Learning tries to automate the process of inductive inference.

# Applications of machine learning

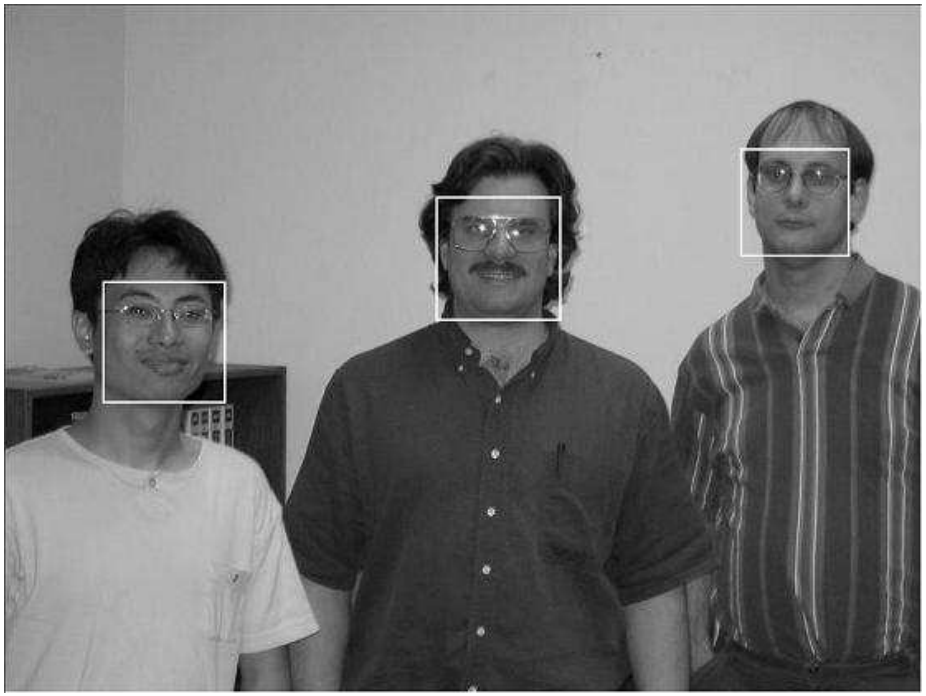
Most important application areas:

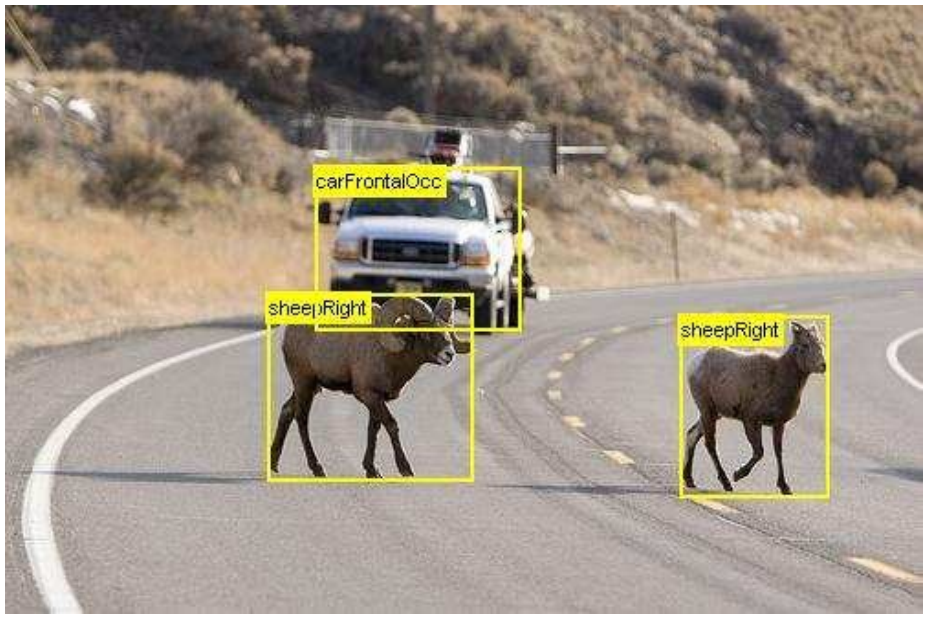
- bioinformatics,
- computer vision/image processing/computer graphics/robotics,
- information retrieval/collaborative filtering,
- natural language processing
- economics,
- other: spam filter/intrusion detection,
- new: machine learning in computer games and software engineering.

More and more data is collected in different areas. No human can analyze it.

$\Longrightarrow$ Increasing demand for machine learning.

# Applications I - Computer Vision






Object categorization in computer vision:

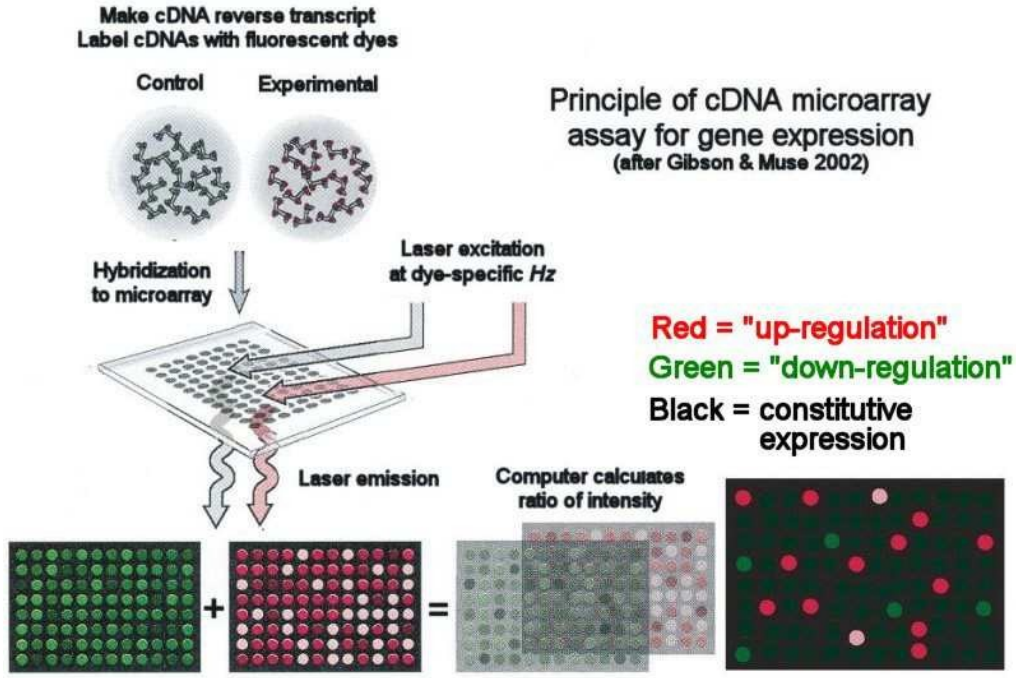
- Face detection (now in digital cameras - works well as long as you look straight into the camera)
- General Object Categorization (Competitions today with more than 20 classes).

# Applications II - Robotics

Machine learning in robotics and autonomous driving:

- Stanford won the DARPA Grand Challenge 2005
- The Grand Urban challenge 2007 has been won by CMU

# Applications III - Bioinformatics




## Bioinformatics:

- prediction of diseases etc. using microarray data.

# Image segmentation

User-guided image segmentation - Example of Semisupervised Learning:

Left: Input Image with user labels
Right: Image segmentation

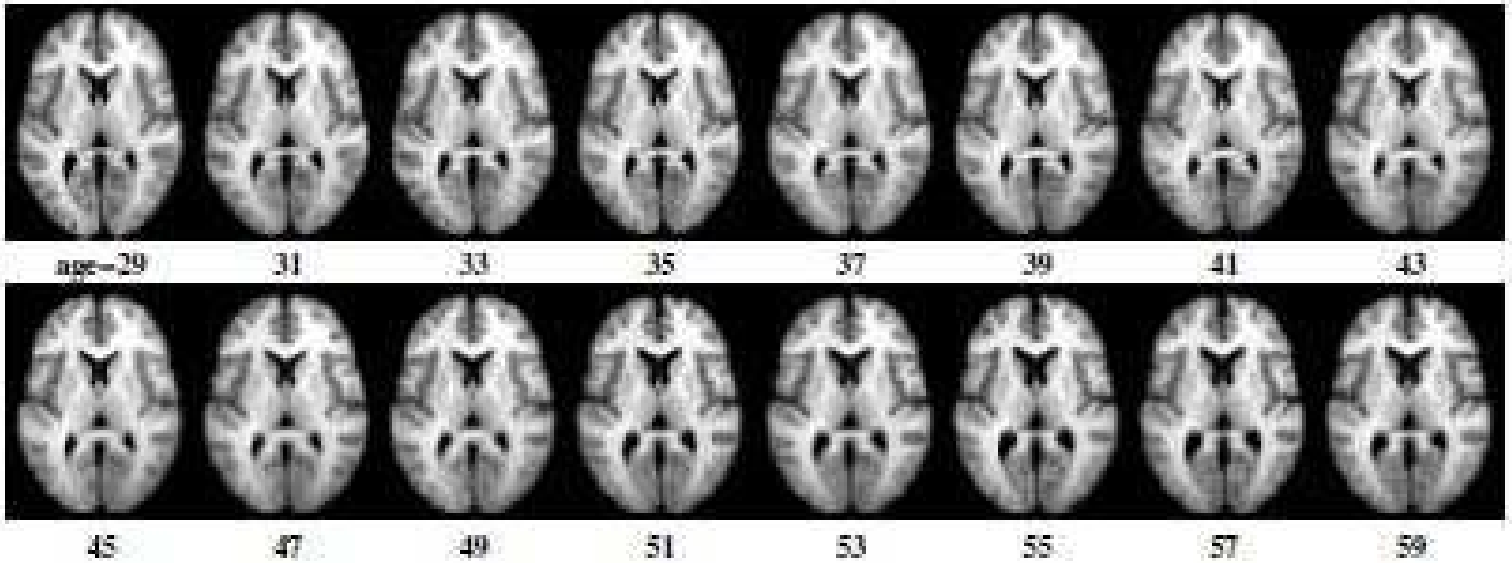
# Regression

## Regression:

Learning of a general function $f : M \rightarrow N$.

## Examples

- Prediction of temperature, wind direction in weather forecast,
- Prediction of whole voxel images,




How does the brain change when one gets older ?

# Types of learning

We distinguish between three main types of learning:

- supervised learning,
- semi-supervised learning,
- unsupervised learning.

In the following:

X: is the input space, $X_i$ are the training inputs,
Y: is the output space, $Y_i$ are the training outputs.

# Supervised Learning

**Supervised Learning:**

Given $n$ observations $T = (X_i, Y_i)^n$ construct function $f_n : \mathsf{X} \to \mathsf{Y}$.

- output space $\mathsf{Y}$ discrete $\Rightarrow$ classification
- output space $\mathsf{Y} = \mathbb{R}$ or $\mathsf{Y} = \mathbb{R}^d \Rightarrow$ (multivariate) Regression
- general output space $\mathsf{Y} \Rightarrow$ Learning with structured output

# Statistical learning I

Assumption: Data is generated by sampling from a probability measure P on $\mathsf{X} \times \mathsf{Y}$.

What does that mean ?

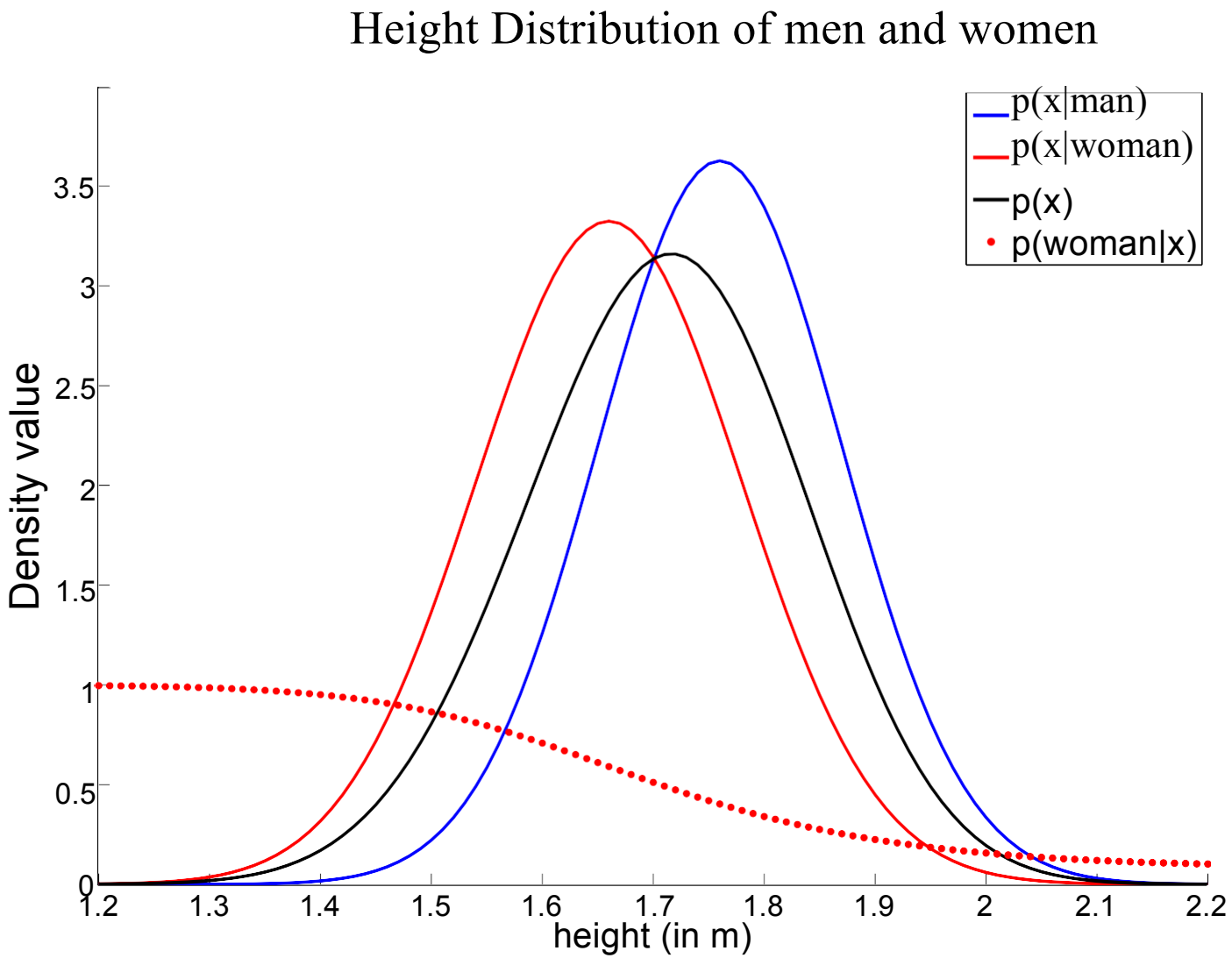
1. Training data is a random sample from P,
2. The labels $y \in \mathsf{Y}$ are non-deterministic, that means there exists not necessarily a function $y = g(x)$. Instead for a given input $x$, there exists a distribution over the possible values in $\mathsf{Y}$.
3. Since the training data underlies statistical fluctuations, the classifier should be relatively stable under small changes of the training data.

## Binary classification ($\mathsf{Y} = \{-1, 1\}$)

The joint density $p(x, y)$ of the probability measure $P$ on $\mathsf{X} \times \mathsf{Y}$ can be decomposed as follows

- The class-conditional density $p(x|y)$. It models the occurrence of the inputs $x$ of class $y$.
- The conditional probability $p(y|x)$. The probability that we observe $y$ given that the input is $x$. The most probable class $y$ for the features $x$ is then used for prediction.
- The marginal distribution $p(x)$. It models the cumulated occurrence of features $x$ over all classes.
- The class probabilities $p(y)$. The total probability of a class $y$.

Learning problem: Predict the sex of a person, $Y = \{\text{male}, \text{female}\}$, using the measured height of the person as a feature (input space is $X = R$).



Marginal distribution

$p(x) = p(x|\text{male})p(\text{male}) + p(x|\text{female})p(\text{female})$.

Using Bayes law we get the conditional distribution $p(y|x)$,

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

Classification rule: classify $x$ as female if $p(\text{female}|x) \geq 1/2$ and otherwise as male.

$\Rightarrow$ From the plot, female if $x < 1.71$ and otherwise male.

# But !


Mattia Dessi and actress wife Brigitte Nielsen


Formula One tycoon Bernie Ecclestone and wife Slavica

# Discriminative versus generative learning

Two main types of approaches to solve a (semi)-supervised statistical learning problem:

- Generative Learning: Estimation of the whole joint distribution

  $p(x, y)$ in particular the class conditional probabilities $p(x|y)$

  $\Longrightarrow$ Using Bayes rule compute the conditional probability $p(y|x)$.

  Advantage: One can create synthetically new data points.

  Disadvantage: One has to solve a harder problem

  $\Longrightarrow$ predictions often worse than in discriminative learning.

- Discriminative Learning: Just model the conditional distribution

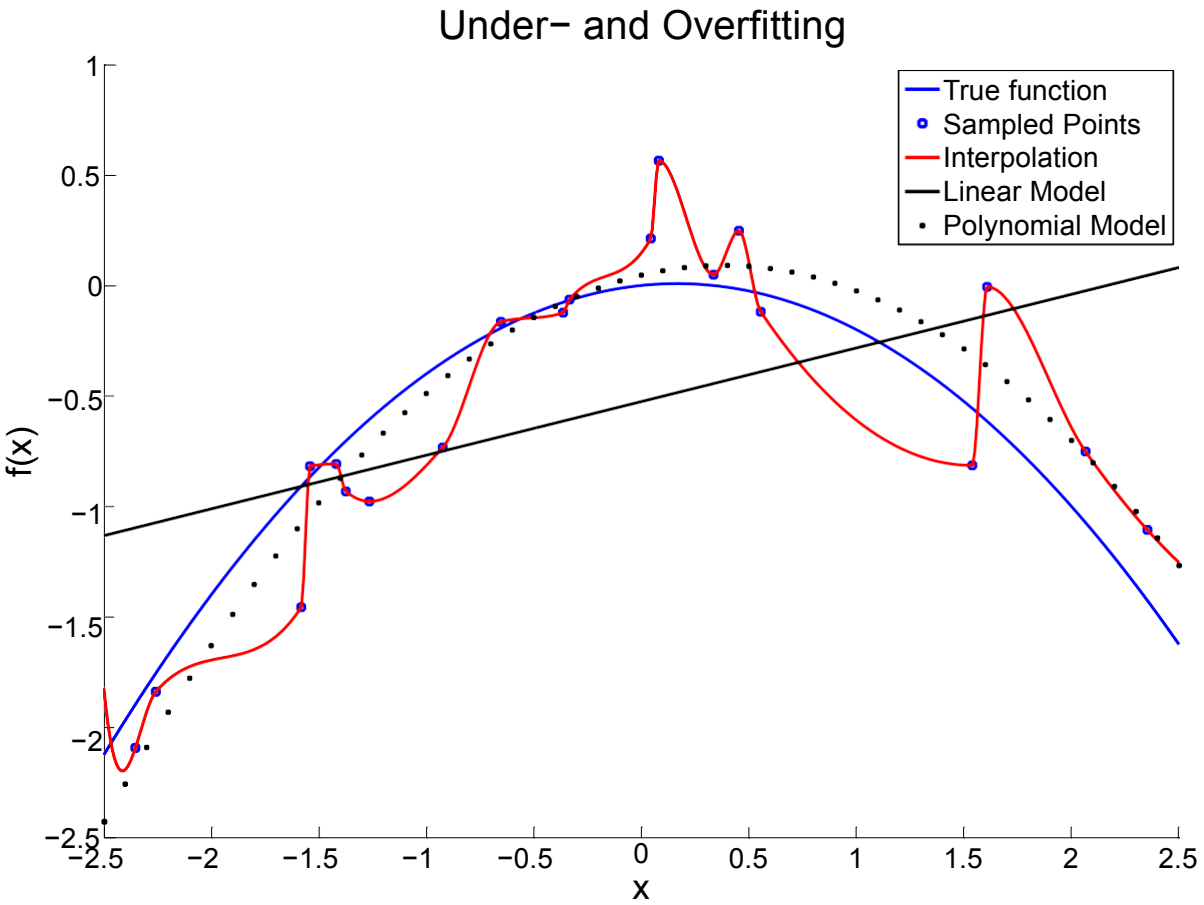
# Challenges in machine learning

Challenges in machine learning:

- choice of features
- integration of prior knowledge
- computational complexity
- curse of dimensionality
- over- and underfitting

# Overfitting and underfitting I

Regression: input $X = \mathbb{R}$, output $Y = \mathbb{R}$, training data $(X_i, Y_i)^n$




blue curve: true function, blue circles:20 noisy samples of the true function, red curve: interpolation of the training points black solid line: fitted linear model, dotted black line: polynomial model.

# Overfitting and underfitting II

- using interpolation techniques there always exists a function $f : X \rightarrow Y$ which fits the data perfectly ! (given that there are no contradictions, that is if $X_i = X_j$ , $i \neq j$, then $Y_i = Y_j$ .)
  ⇒ overfitting of the data.
  ⇒ no generalization !
- using a very simple regression model e.g. a linear one will often lead to underfitting, which means that the learned function can hardly represent the functional relationship given by the data.
  ⇒ generalization but poor performance !

# Overfitting and underfitting II

Overfitting leads to poor generalization since:

1. The training data is usually noisy. A perfect fit implies that one has fitted the noise.
2. A regression model can represent functions from a function class $\mathsf{F}$ (e.g. polynomials of degree $k$).

(a) the risk $R(f_n)$ of the function chosen based on the training data,

*(b)* $R_{\mathsf{F}} = \min_{f \in \mathsf{F}} R(f)$ is the minimal risk over the class $\mathsf{F}$,

*(c)* $R^*$ is the minimal risk over all functions, the Bayes risk,

Decomposition into the deterministic approximation error and the random estimation error,

$$R(f_n) - R^* = \underbrace{R(f_n) - R_{\mathsf{F}}}_{\text{Estimation error}} + \underbrace{R_{\mathsf{F}} - R^*}_{\text{Approximation error}}$$

# Overfitting and underfitting III

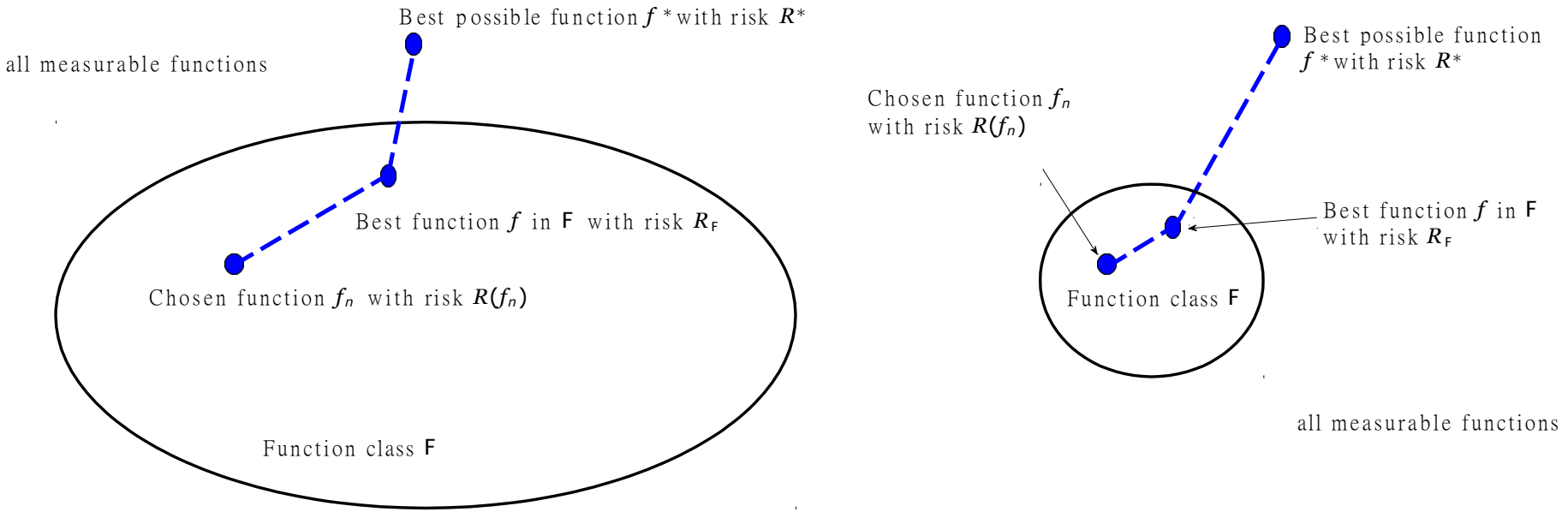



Figure 1: Left: Regression model with large capacity (small approximation error, high estimation error), Right: Regression model with small capacity (high approximation error, small estimation error)

# Purpose of different learning methods

## One learning method for all purposes ?

- Key result of statistical learning theory: there exists no universally best learning method. On 'average' (this has to be defined carefully !) they perform all the same.
- Nature is 'nice' to us - most problems are not of pathological nature.

## Purpose of different learning methods ?

- Each learning method implicitly or explicitly models different prior assumptions about the input-output relationship.The art of machine learning is to choose the method which best fits the data generating process.
- Data can never replace prior knowledge.