Applied Machine Learning
Introduction to the Course

UNIVERSITY OF
GOTHENBURG

CHALMERS

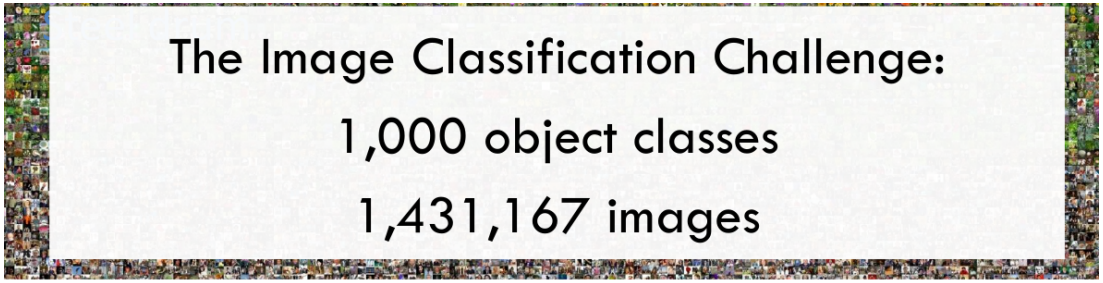
Richard Johansson

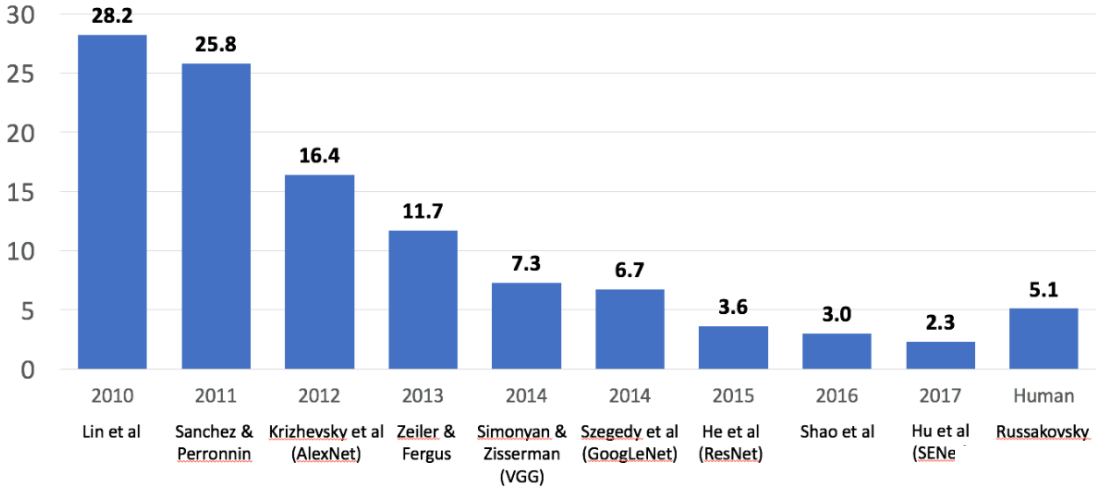
richajo@chalmers.se

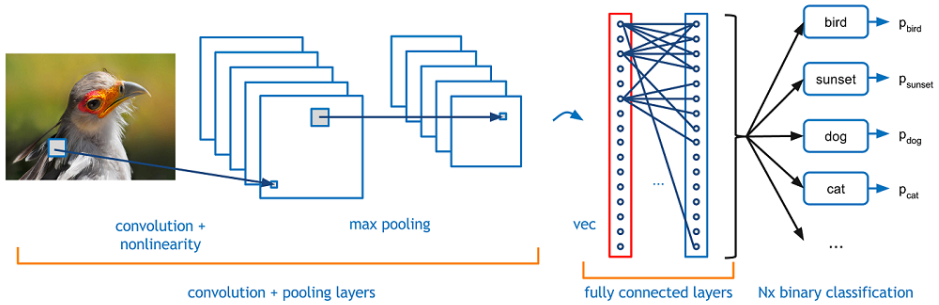
Welcome to the course!

- Machine learning is increasingly popular among students
  - our courses take increasing volumes
  - many thesis projects develop or apply ML models
- ...and in industry, public sector
  - many companies come to us looking for students
  - joint research projects
- Why the fuss and why now?

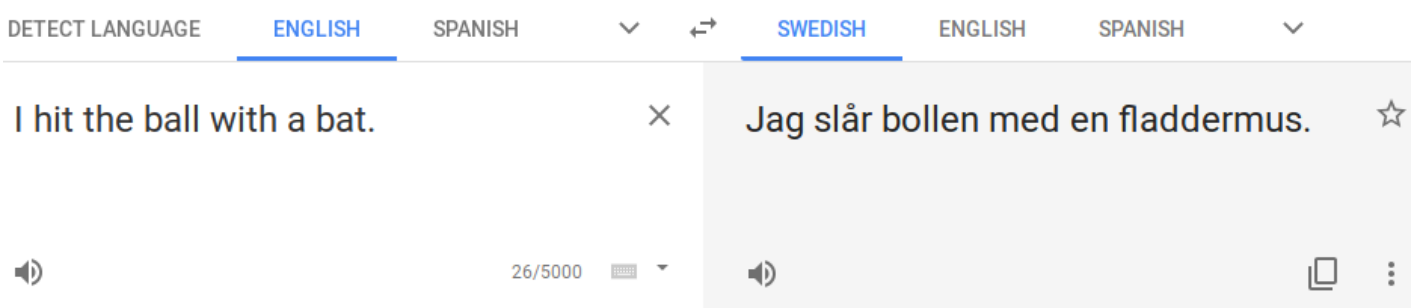
Success stories: image recognition

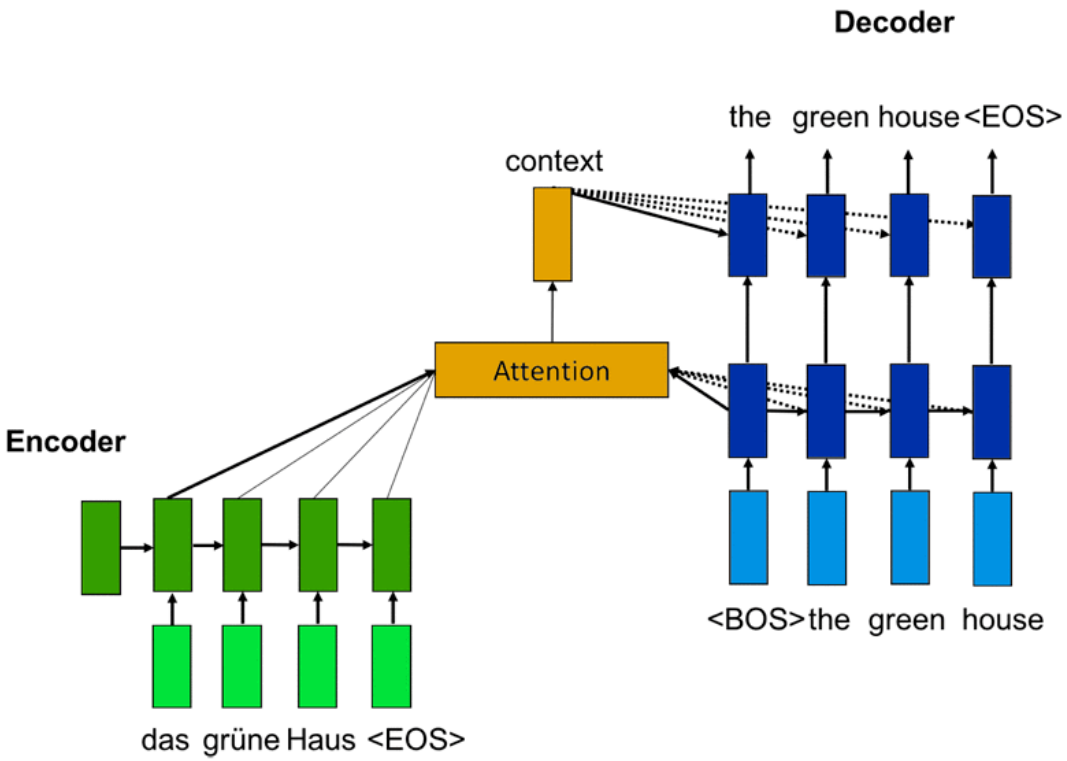




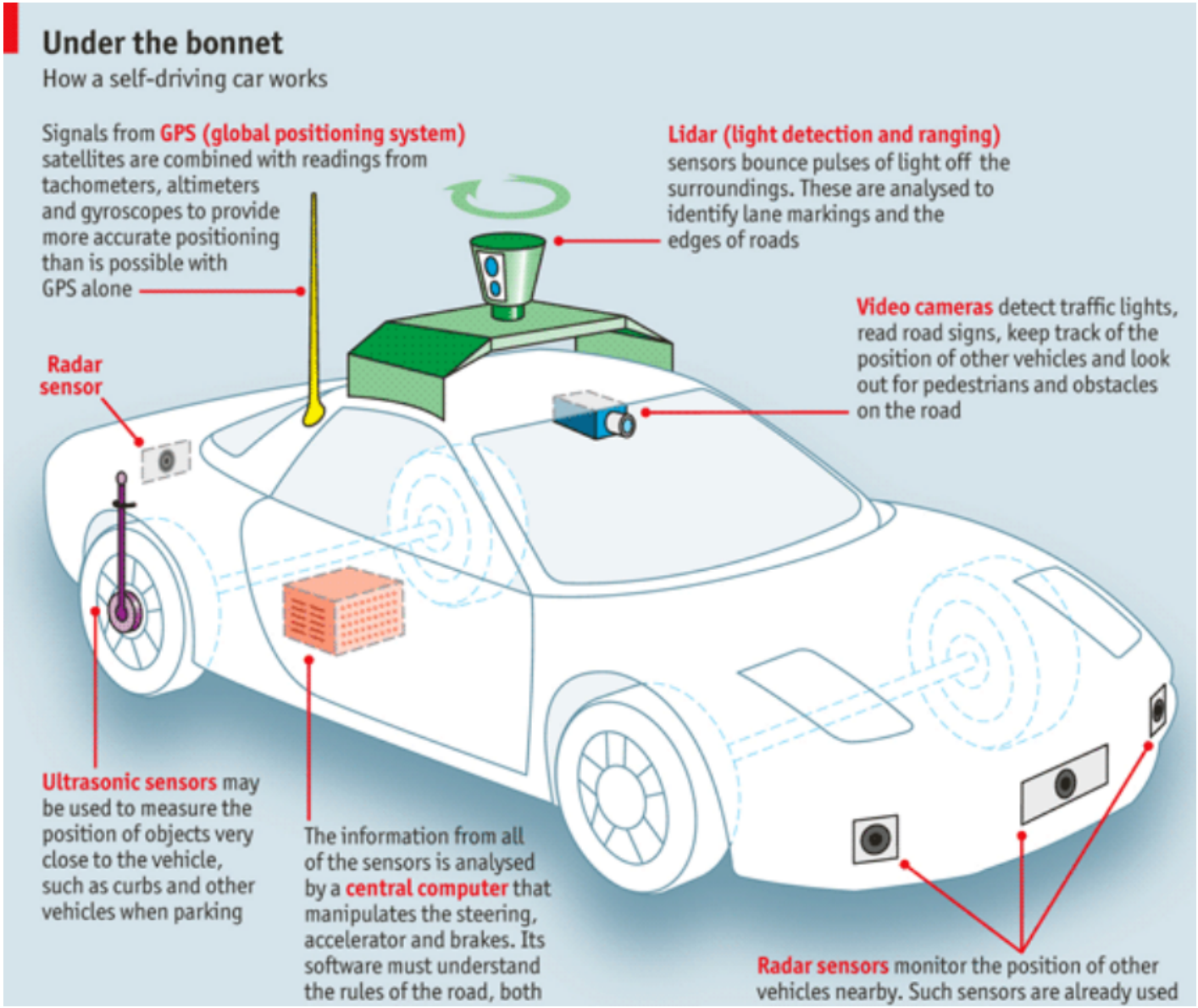


Success stories: machine translation



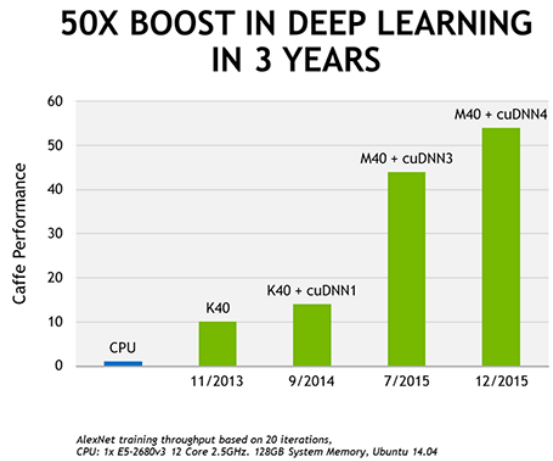


Data


Under the bonnet
How a self-driving car works
Signals from GPS (global positioning system) satellites are combined with readings from tachometers, altimeters and gyroscopes to provide more accurate positioning than is possible with GPS alone
Lidar (light detection and ranging) sensors bounce pulses of light off the surroundings. These are analysed to identify lane markings and the edges of roads
Video cameras detect traffic lights, read road signs, keep track of the position of other vehicles and look out for pedestrians and obstacles on the road
Radar sensor
Ultrasonic sensors may be used to measure the position of objects very close to the vehicle, such as curbs and other vehicles when parking
The information from all of the sensors is analysed by a central computer that manipulates the steering, accelerator and brakes. Its software must understand the rules of the road, both
Radar sensors monitor the position of other vehicles nearby. Such sensors are already used

[source]

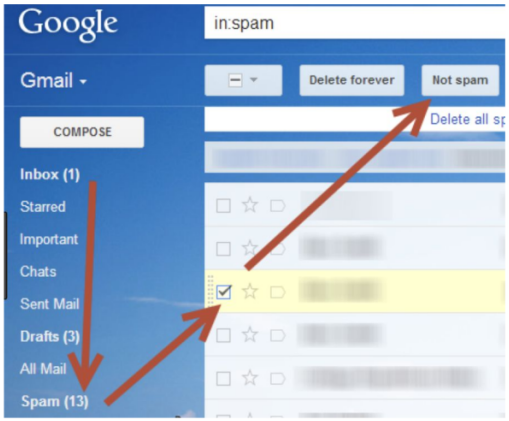

Titan X
VS
GTX 980
VS
Tesla M40
VS
Tesla K80

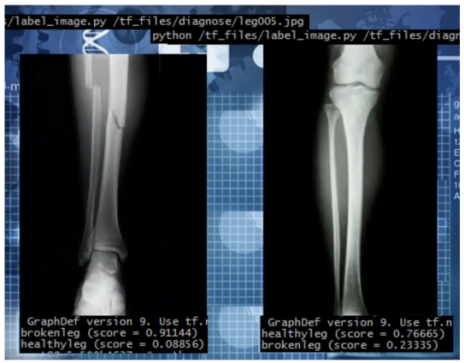

50X BOOST IN DEEP LEARNING IN 3 YEARS
Caffe Performance
60
50
40
30
20
10
0
CPU
K40
K40 + cuDNN1
M40 + cuDNN3
M40 + cuDNN4
11/2013
9/2014
7/2015
12/2015
AlexNet training throughput based on 20 iterations,
CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04

[source]


APACHE
Spark
orange
WEKA
The University
of Waikato
mxnet
scikit
learn
TensorFlow
K
Keras
CNTK
PYTORCH
R

Applications...





[source]

Topics covered in the course

- The usual "zoo": a **selection of machine learning models**
  - what's the idea behind them?
  - how are they implemented? (at least on a high level)
  - what are the use cases?
  - how can we apply them practically?
- But hopefully also the "**real-world context**":
  - extended "messy" practical assignments requiring that you think of what you're doing
  - invited talks from industry and/or the healthcare sector
  - annotation of data, evaluation
  - ethical and legal issues, interpretability

Overview

Practical issues about the course

Fundamental concepts in machine learning

Machine learning libraries in Python

Course webpage

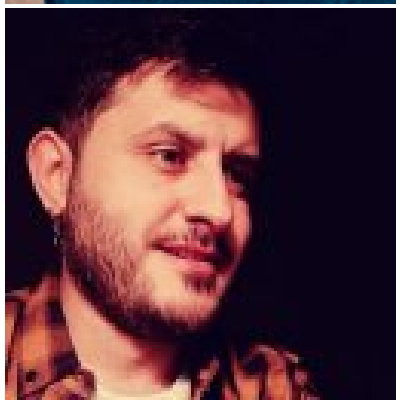
- The official course webpage is the **Canvas page**
  `https://chalmers.instructure.com/courses/27917`

People involved in the course

- **Richard**: examiner, responsible for the course
- **Anton**, **Jack**, **Newton**, **Peter**, **Yossra**: helping you with the assignments








Structure of teaching

- **Lecture discussions** Tuesdays and Fridays 13–15
  - we will use a **flipped classroom** format with pre-recorded lectures you are expected to have watched **before** the session
  - summary and discussion of the content of the recorded lectures
  - interactive coding
  - solving a few exercises when we have time
  - feel free to ask questions before the session!
- **Assistance sessions** Thursdays 13–17
  - our TAs help you work on your assignments
  - please let me know if it's too crowded
  - **in a computer lab room** (with possibly additional remote sessions)

Assignments

- Five compulsory **assignments**:

PA 1 intro to the ML workflow, decision trees
PA 2 random forests
PA 3 text classification
PA 4 neural network software
PA 5 medical image classification

- We will use the **Python** programming language
- Please refer to the course PM for details about grading
- Assignments are done in **groups**

Programming assignment 1

- Warmup lab exercise: quick tour of the scikit-learn library
- Introduction to decision trees
- For a high grade: implement decision tree regression
- Assistance sessions this Thursday
- Submission deadline: **January 22**

Literature

- We won't follow a book closely, but we'll give pointers to reading material in this book:
  - *Machine Learning: A course for engineers and scientists* by Lindholm et al: `http://smlbook.org/`
- And additional papers to read for some topics
- Some notes to complement the lectures
- Example code will be posted on the course page

Additional material along the way

- Exercise sheets, old exams
- Online quizzes

Exam, mid-March

- This is a **take-home** exam: a written assignment
- Will be available during the whole exam period
- **Two-part structure**:
  1. a first compulsory part about basic concepts: you need to answer most of these questions correctly to pass
  2. a second optional part that requires more insight: answer these questions for a higher grade

Student representatives

- If you're interested in being a student representative, please send me an email!
- The workload is light and there will be a small reward…

Overview

Practical issues about the course

Fundamental concepts in machine learning

Machine learning libraries in Python

Predictive models

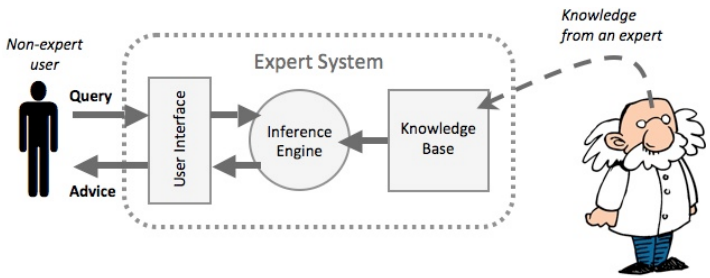
- Given some object, make a prediction
  - is this patient diabetic?
  - what animal does this image show?
  - what is the market value of this apartment?
  - what are the phonemes contained in this speech signal?

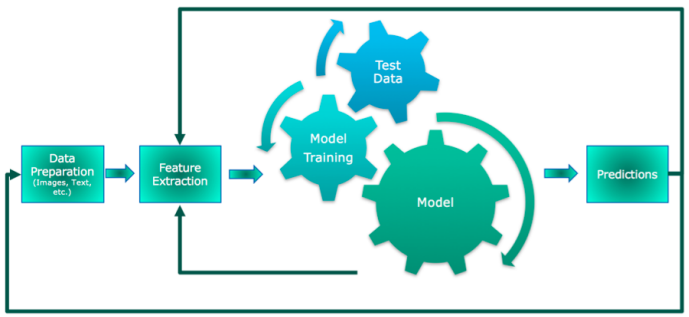
Predictive models

- Given some object, make a prediction
  - is this patient diabetic?
  - what animal does this image show?
  - what is the market value of this apartment?
  - what are the phonemes contained in this speech signal?
- The goal of machine learning is to build the predictive models by observing data

Predictive models

- Given some object, make a **prediction**
  - is this patient diabetic?
  - what animal does this image show?
  - what is the market value of this apartment?
  - what are the phonemes contained in this speech signal?
- The goal of machine learning is to build the predictive models by **observing data**
- Contrast: **expert-defined** or **data-driven**





[source]

Why machine learning?

Why would we want to "learn" the function from data instead of just implementing it?

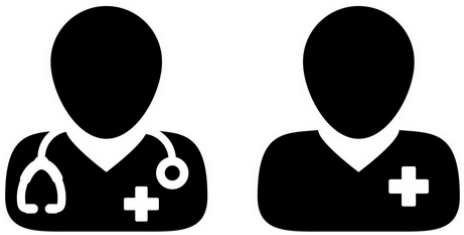
- Usually **because we don't really know** how to write down the function by hand
  - speech recognition
  - image classification
  - machine translation
  - ...
- Might not be necessary for **limited** tasks where we **know**
- What is more expensive in your case? knowledge or data?

Don’t forget your domain expertise!

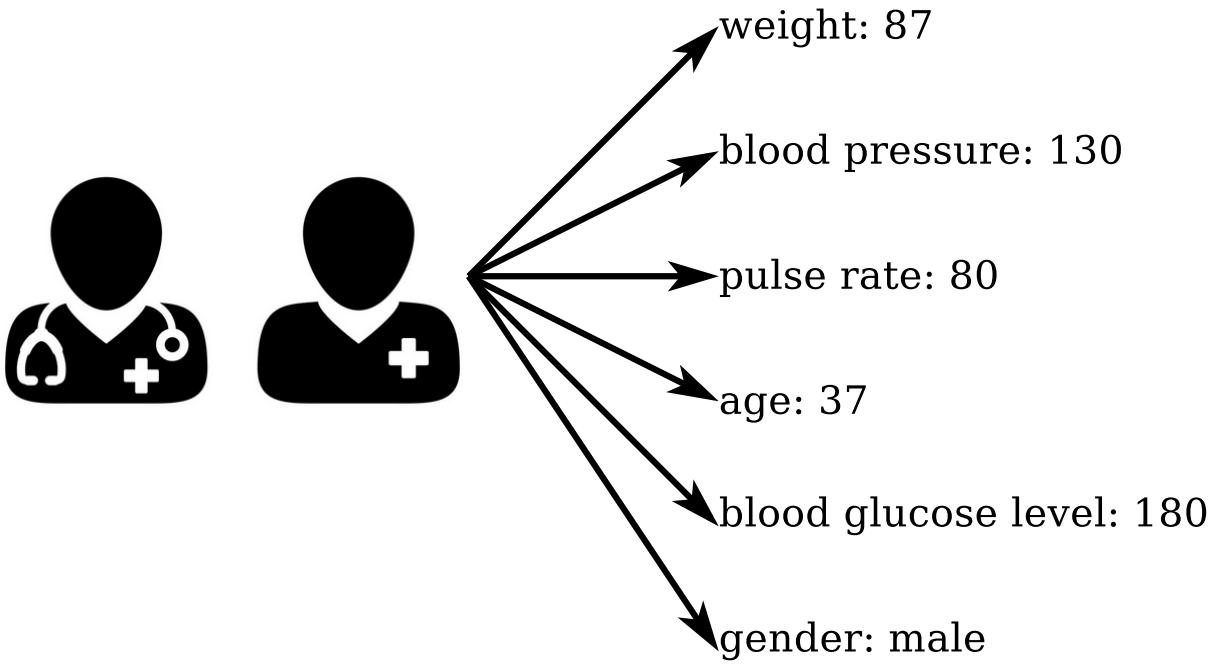
ML makes some tasks automatic, but we still need our brains:

- **defining** the tasks, terminology, evaluation metrics
- **annotating** (hand-labeling) training and testing data
- designing **features**
- **error analysis**

Example: is the patient diabetic?



Example: is the patient diabetic?



- In order to predict, we make some measurements of properties we believe will be useful: these are called the **features**

More terminology: what is the output?

- **Classification**: learning to output a category label
  - spam/non-spam; positive/negative; ...
- **Regression**: learning to guess a number
  - value of a share; number of stars in a review; ...

How is the training signal provided?

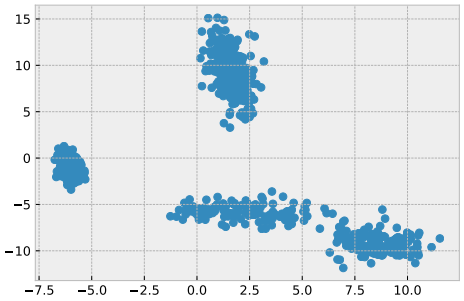
- In **supervised** learning, the training set consists of **input–output** pairs
- our goal is to learn to produce the outputs

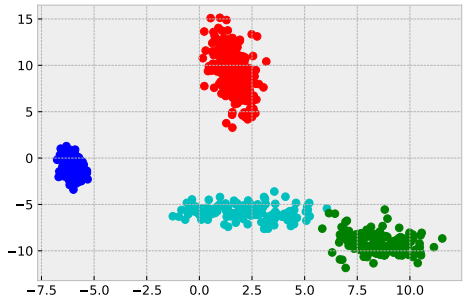
age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	target
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

	F1	F2	F3	F4	F5	F6	F7	F8	F9	RMSD
0	13558.30	4305.35	0.31754	162.1730	1.872791e+06	215.3590	4287.87	102	27.0302	17.284
1	6191.96	1623.16	0.26213	53.3894	8.034467e+05	87.2024	3328.91	39	38.5468	6.021
2	7725.98	1726.28	0.22343	67.2887	1.075648e+06	81.7913	2981.04	29	38.8119	9.275
3	8424.58	2368.25	0.28111	67.8325	1.210472e+06	109.4390	3248.22	70	39.0651	15.851
4	7460.84	1736.94	0.23280	52.4123	1.021020e+06	94.5234	2814.42	41	39.9147	7.962

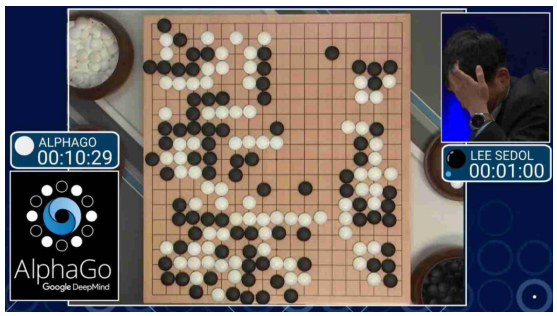
Types of supervision: alternatives

- **Unsupervised learning**: we are given "unorganized" data
  - our goal is to discover some structure





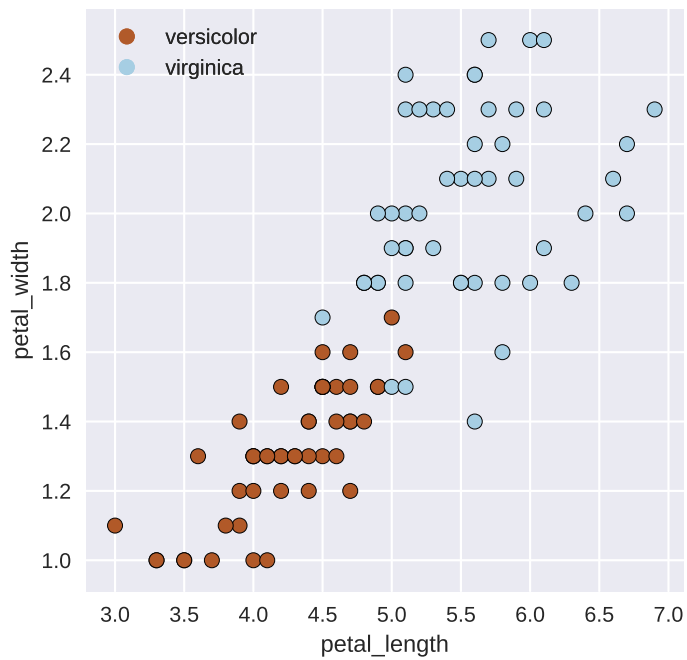
- **Reinforcement learning**: our problem is formalized as a game
  - an agent carries out actions and receives rewards



Example: Fisher's iris data


```
iris.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



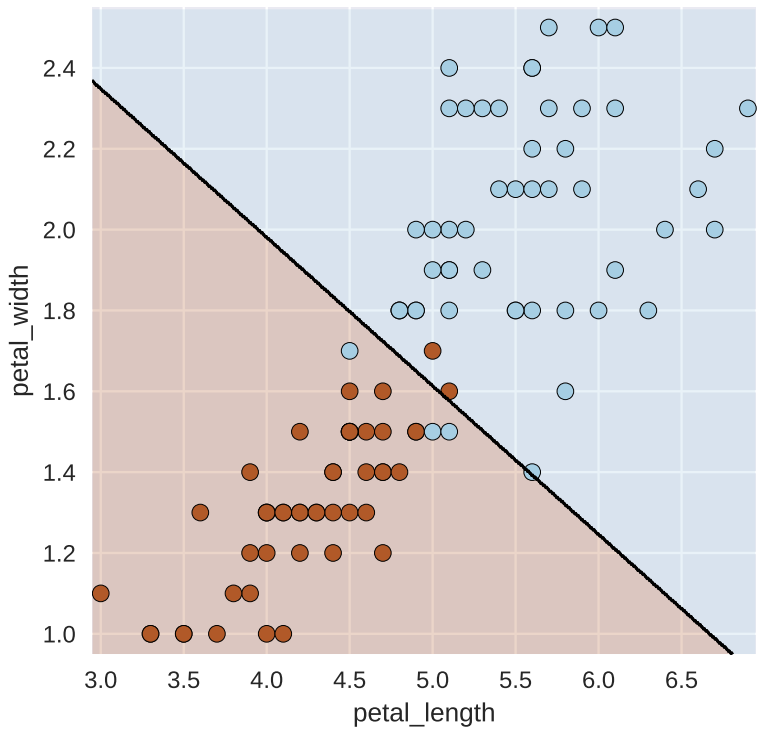
Approach 1: linear separator

if $0.85 \cdot$ petal_length $+ 2.42 \cdot$ petal_width ≥ 8.34:

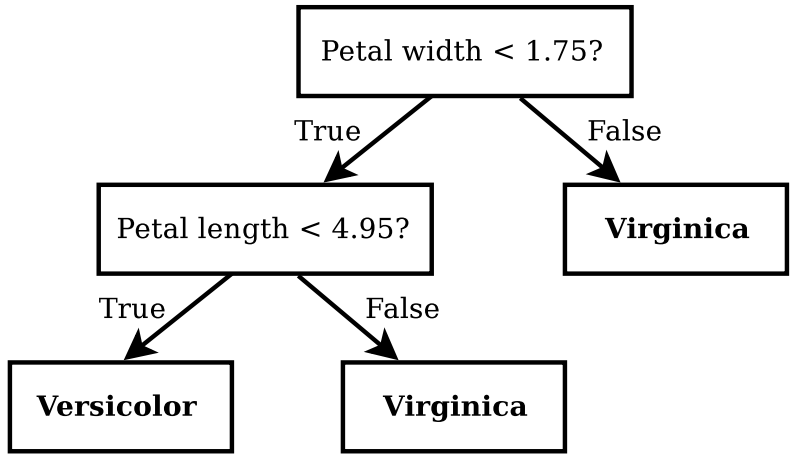
    return virginica

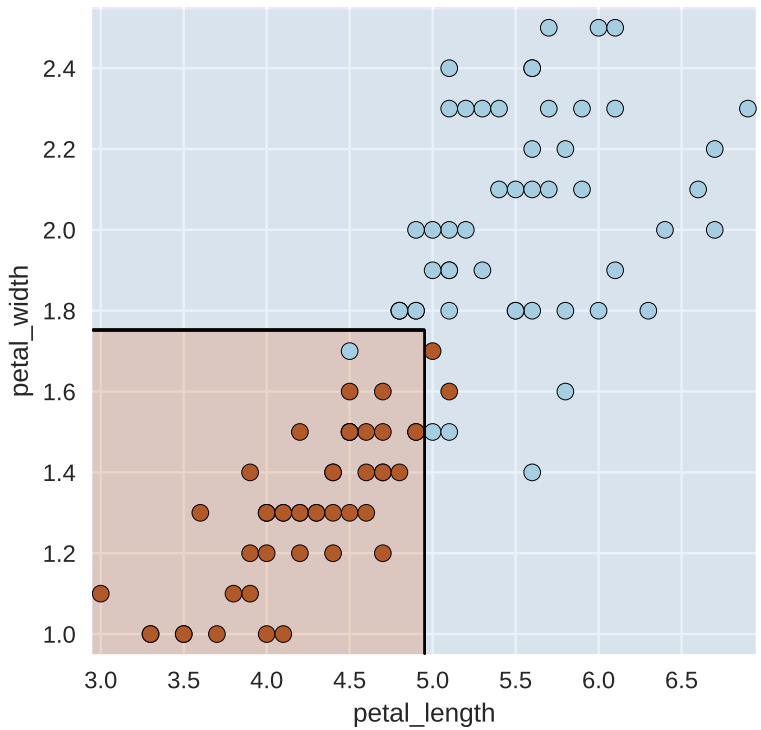
else

    return versicolor

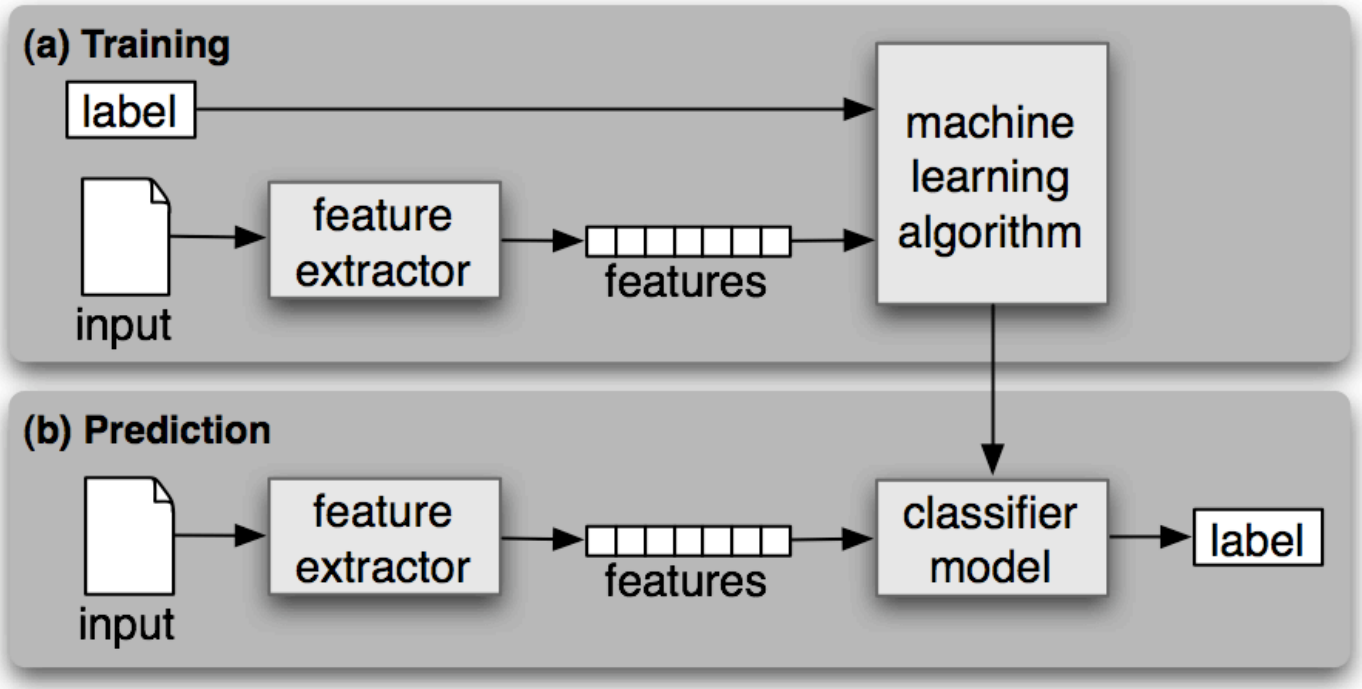


Approach 2: if/then/else tree





Basic supervised machine learning workflow

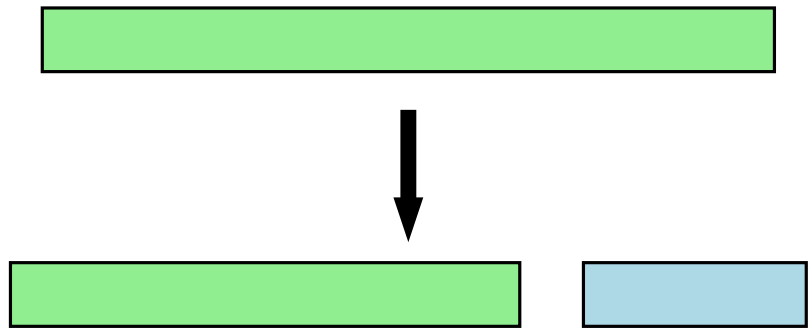


Basic ML methodology: evaluation

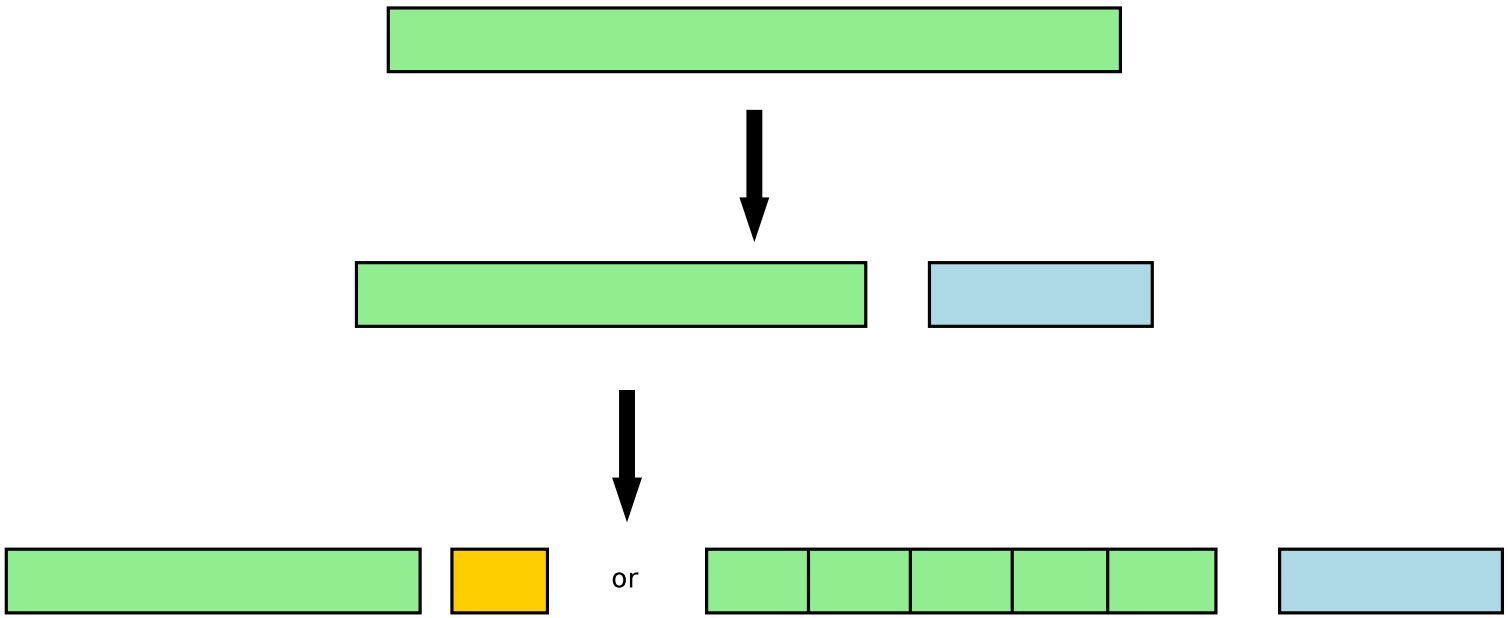
- Select an **evaluation procedure** (a "metric") such as
  - **classification accuracy**: proportion correct classifications?
  - **mean squared error** often used in regression
  - or some domain-specific metric
- Compare to one or more **baselines**
  - trivial solution
  - rule-based solution
  - existing solution
- Apply your model to a held-out **test set** and evaluate
  - the test set must be different from the training set
  - also: don't optimize on the test set; use a development set or cross-validation!

Managing your data

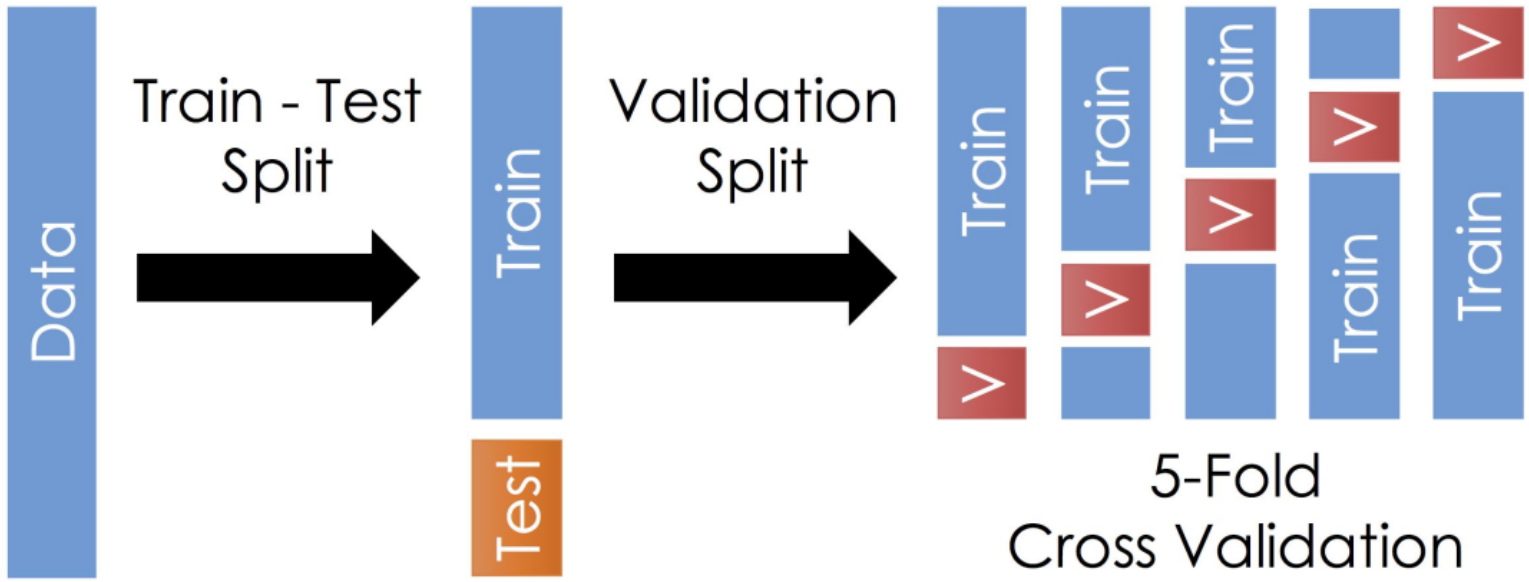
Managing your data



Managing your data



Managing your data for evaluation and cross-validation



[source]

Overview

Practical issues about the course

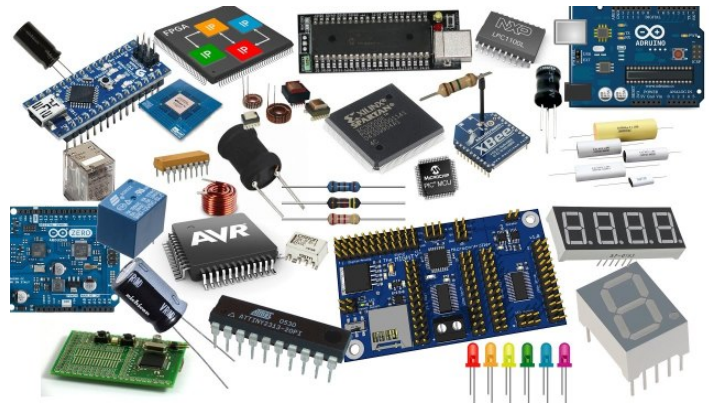
Fundamental concepts in machine learning

Machine learning libraries in Python

Use cases for machine learning

- **Standard use cases**: standard solutions are available

- **Special cases**: we may need to tailor our own solutions



The Python machine learning ecosystem (selection)

abstraction

low-level control

scikit learn

dmlc XGBoost

Keras

PYTORCH

TensorFlow

NumPy

SciPy

Intel Math Kernel Library Intel MKL Library

NVIDIA CUDA

Machine learning software: a small sample

- General-purpose software, large collections of algorithms:
  - scikit-learn: `http://scikit-learn.org`
    - Python library – will be used in this course
  - Weka: `http://www.cs.waikato.ac.nz/ml/weka`
    - Java library with nice user interface
- Special-purpose software, small collections of algorithms:
  - Keras, PyTorch, TensorFlow, JAX for neural networks
  - LibSVM/LibLinear for support vector machines
  - XGboost, lightgbm for tree ensembles
  - ...
- large-scale learning in distributed architectures:
  - Spark MLLib
  - H2O

Scikit-learn toy example

See also

https://scikit-learn.org/stable/getting_started.html

Up next

- Thursday: lab sessions for programming assignment 1
- Topic of Friday's discussion:
  - decision trees
  - ensembles and random forests
  - generalization, under/overfitting
- Please prepare for assignment 1 by reading my code and the extra reading on decision trees