MACHINE LEARNING

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

OVERVIEW

- What is machine learning (ML)?
- Why (do we need / study) ML?
- ML application areas
- Kinds of ML tasks
- Basic steps / elements of the ML process
- Bias (under-fitting) vs. Variance (over-fitting)

What is Machine Learning (ML) ?

ML refers to the capacity of a software system to:

- *generalize* based on the previous *experience*,
  - *experience* = data about entities / instances the system is learning about
- to use the created generalizations to answer questions or solve tasks related to the entities / instances it hasn't encountered before

WHAT IS ML?

A computer system is learning from the *experience* ***E***

related to the *task* ***T***, and the performance measure(s) ***P***,

if its performance on the task ***T***, measured with metrics ***P***,

improve with the experience ***E***

Tom Mitchell (1997)

WHAT IS ML?

Example: a program that labels messages as spam or not-spam

- Task (***T***): classification of email messages into spam and not-spam messages
- Experience (***E***): a set of email messages labelled either as spam or not-spam; “observation” of the user while he/she is interacting with the received messages (marking them as spam)
- Performance metric (***P***): the percentage of email messages that are correctly classified as spam/not-spam (accuracy)

WHY (DO WE NEED/STUDY) ML?

1) There are tasks that people solve easily, but are unable to explain / describe precisely (algorithmically) how they perform those tasks

Examples: image recognition, interpretation of the speech

2) There are tasks that can be solved by specifying an algorithmic procedure, but the algorithms tend to be overly complex and/or require huge knowledge bases

Example: automated translation

WHY (DO WE NEED/STUDY) ML?

3) In many application areas, computer systems are collecting ever increasing collections of data, and it is expected that some knowledge can be mined/distilled from those data; e.g.:

- in medicine: data about patients and their therapies
- in sport: data about games/matches and individual players
- in marketing: data about users/customers, their purchases, their interests, product reviews and ratings,…

Such data sets need to be analyzed using methods capable of mining regularities / patterns in the data that are neither known, nor obvious, but might be useful

APPLICATION AREAS

Numerous and diverse application areas

- Classification of documents based on their topics, or expressed sentiment, opinion and/or emotions
- Machine/automated translation of documents
- Natural language understanding
- Image recognition
- Market segmentation
- Detecting patterns in user behaviour
- Self-driving cars
- ...

Types of ML tasks

Three basic types of ML tasks:

- Supervised learning
- Unsupervised learning
- Reinforced learning

SUPERVISED LEARNING

Refers to a set of problems and methods for resolving them where the program receives:

- a set of input data (x_1, x_2, …, x_n) – attributes that describe a set of problem-specific items / observations
- a set of correct / desired output values, so that for each input data item (observation) x_i, there is the correct / desired output y_i

The program's task is to "learn" to assign the correct output value to a new, unlabeled data item (observation)

SUPERVISED LEARNING

Depending on the type of the output value, there are two kinds of supervised ML tasks:

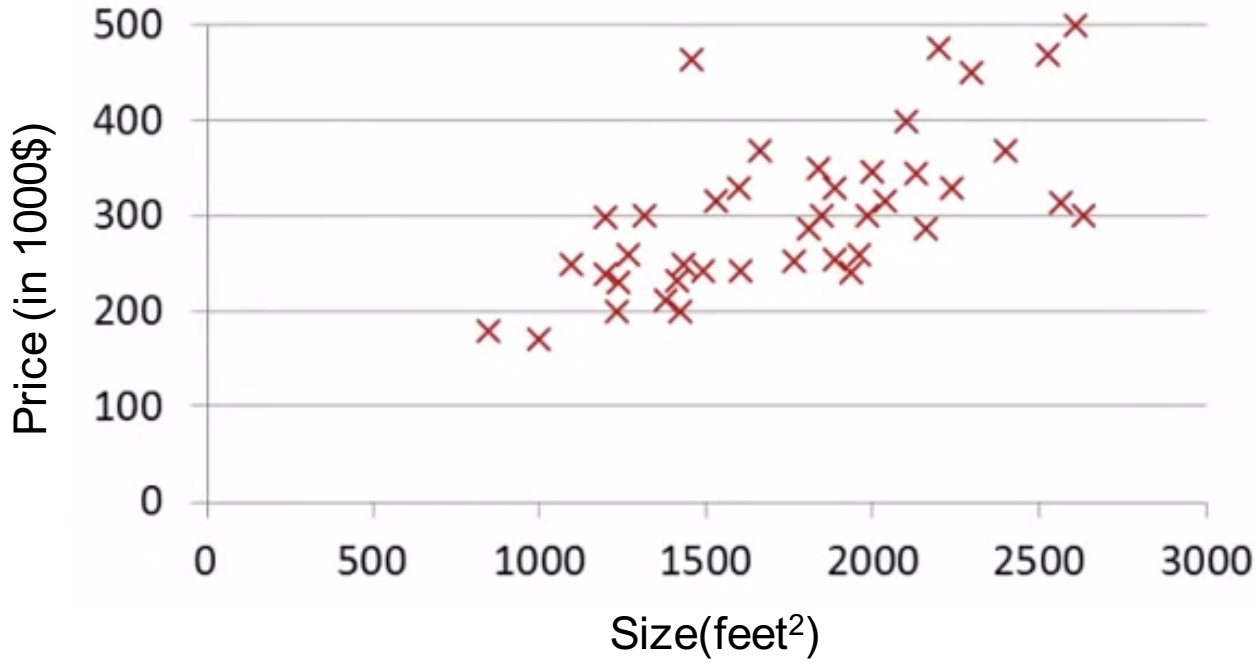
- Classification task – output values are labels (or nominal values)
- Regression task – output values are real numbers

SUPERVISED LEARNING - REGRESSION

Example: simple linear regression

Predict the price of a real estate based on its size

Data used for learning: size (x) and price (y) of real estate in a city suburb



Source: https://www.coursera.org/learn/machine-learning

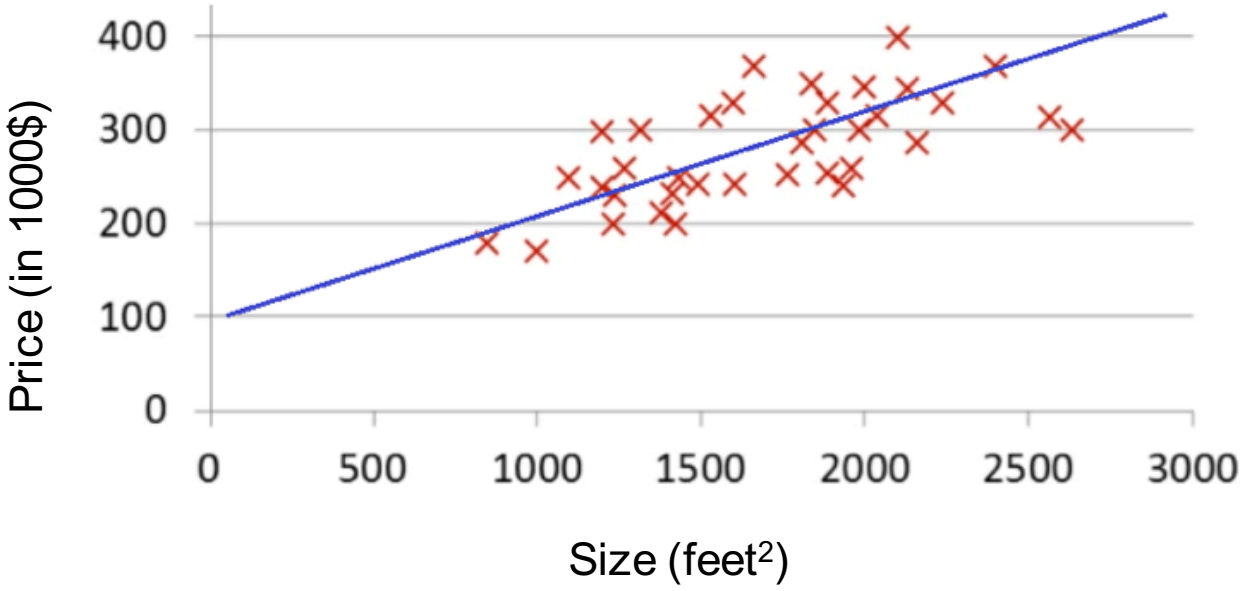
SUPERVISED LEARNING - REGRESSION

Example: simple linear regression (cont.)

Linear function to be learnt in this case has the following form:

$$h(x) = a + bx$$

- *a* and *b* are coefficients that the program should *estimate* based on the available data;
- learning in this case refers to estimating the values of these two parameters



UNSUPERVISED LEARNING

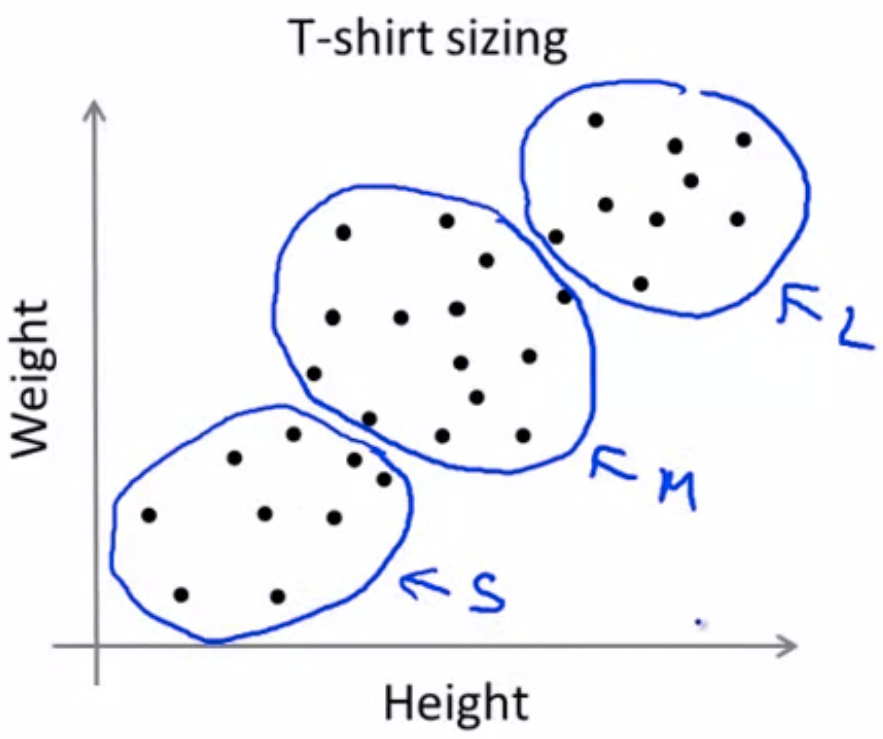
In the case of unsupervised learning task:

- we have no information about the desired/correct output values
- the program receives only a set of input data items (x_1, x_2, …, x_n) – attributes describing the observations / items relevant for the problem under study

The program's task is to discover patterns in the input data, that is, to identify hidden structure / regularities in the data

UNSUPERVISED LEARNING

Example: identifying T-shirt sizes based on the people's height and weight



Source: https://www.coursera.org/learn/machine-learning

REINFORCED LEARNING

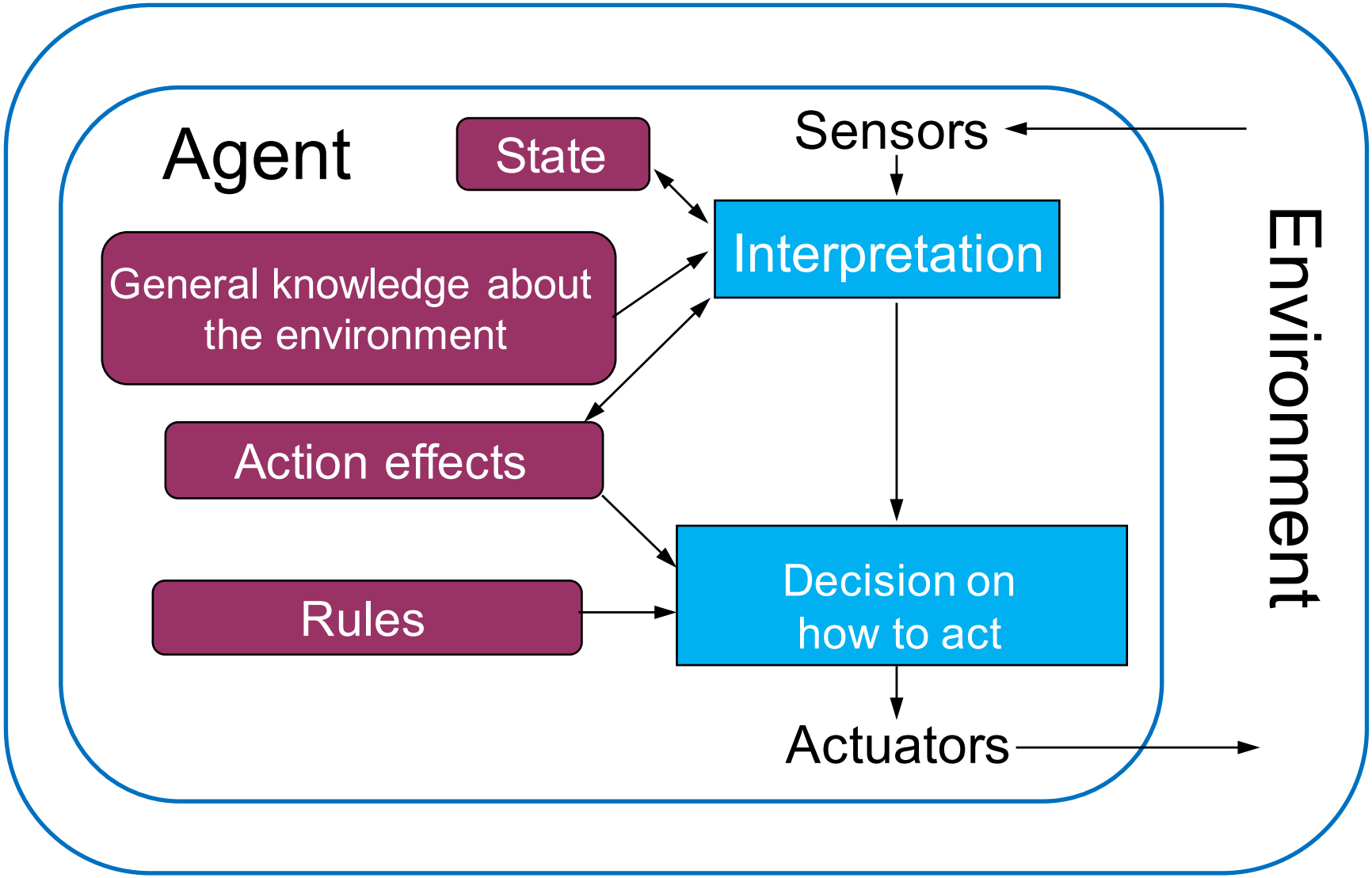
In this type of ML, a program (agent) interacts with its environment by producing actions a1, a2,…

These actions affect the state of the environment, which in turn reacts by providing the program with a feedback in the form of some scalar rewards (or punishments) r1, r2,…

The goal of the program is to learn to act in a way that maximizes the future rewards it receives (or minimizes the punishments) over its lifetime

Example: computer games, autonomous vehicles

ILLUSTRATION OF REINFORCED LEARNING

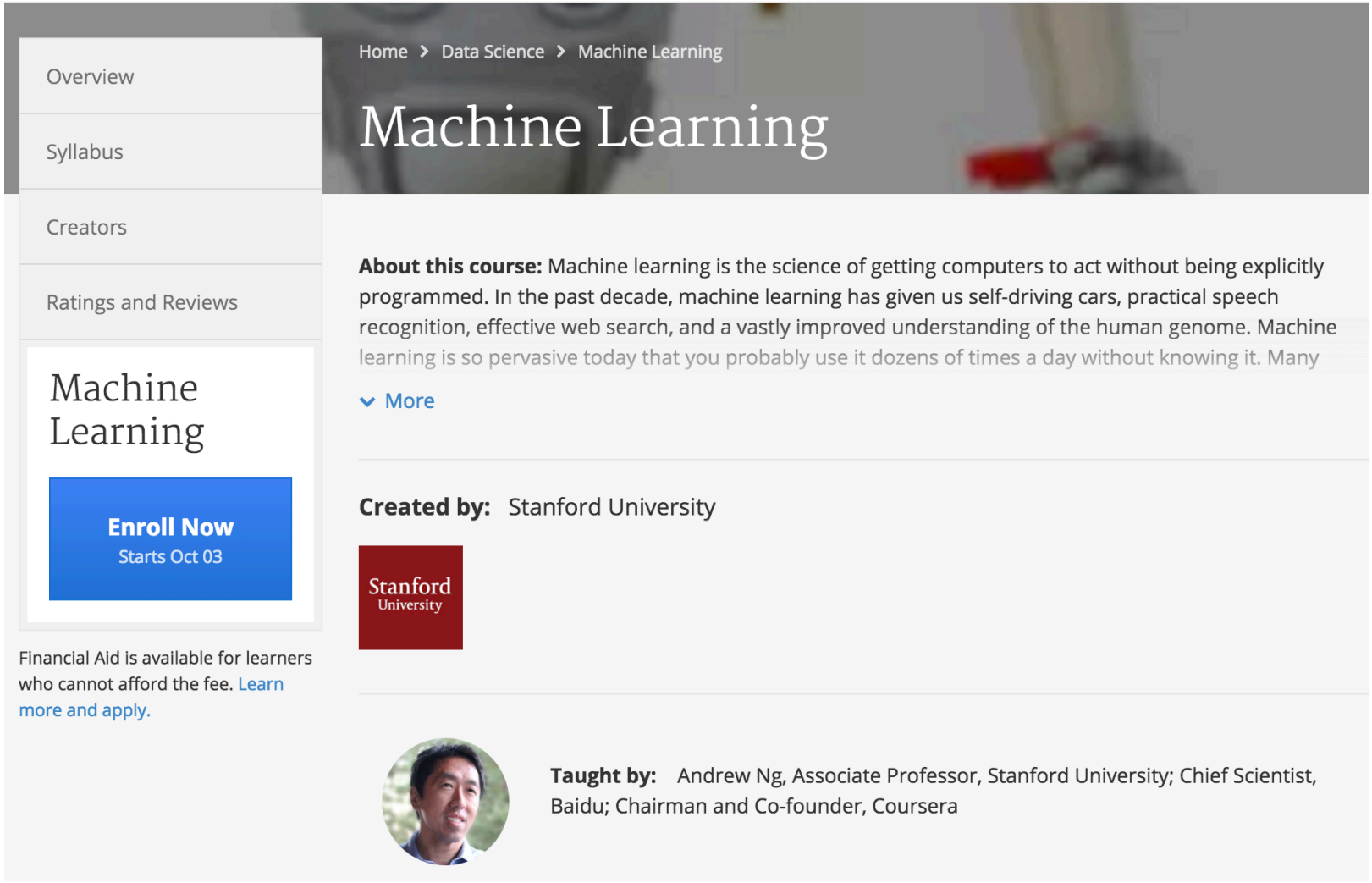


Basic steps and elements of the ML process

BASIC STEPS OF THE ML PROCESS

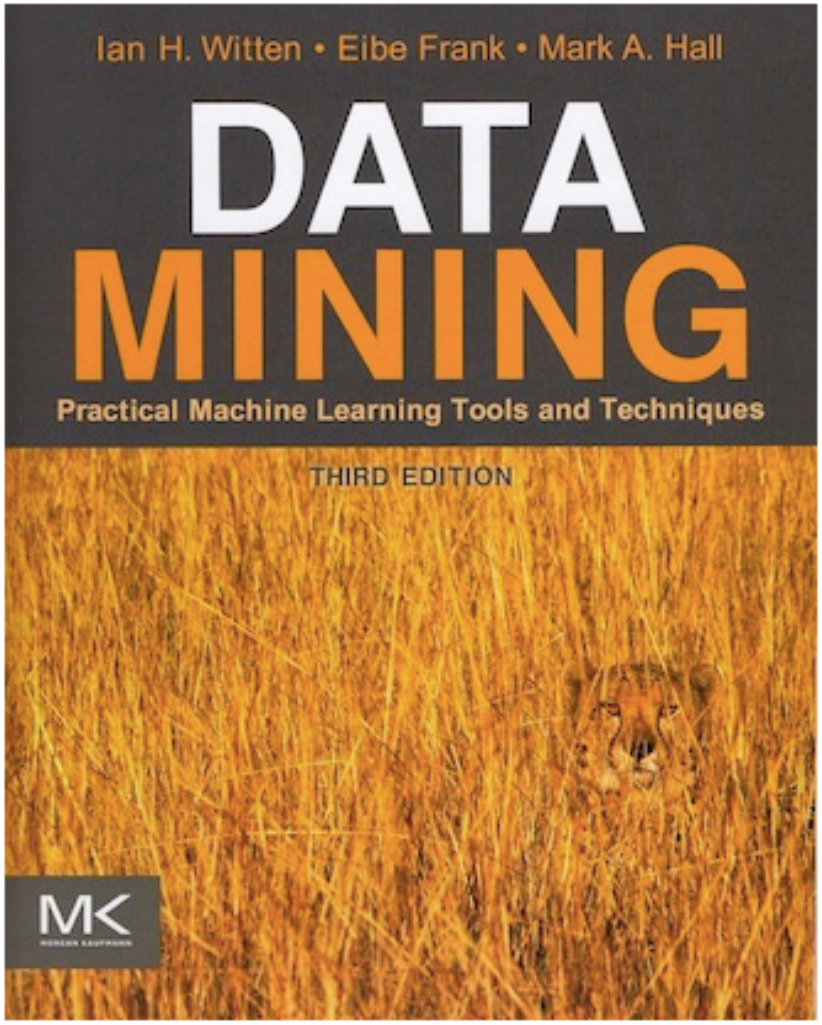
1) *Gathering the data* required for training, (validation) and testing of ML models
2) *Data preparation*, which typically includes cleaning and transformation of the gathered data
3) Analysis of the resulting datasets, and their further enhancement through *attribute selection and/or transformation*
4) *Selection of* 1 or more *ML methods / algorithms*
5) *Training, configuration and evaluation* of the built ML models
6) *Selection of the model* to be used (based on the results of step 5) *and its testing*

ACKNOWLEDGEMENTS AND RECOMMENDATIONS

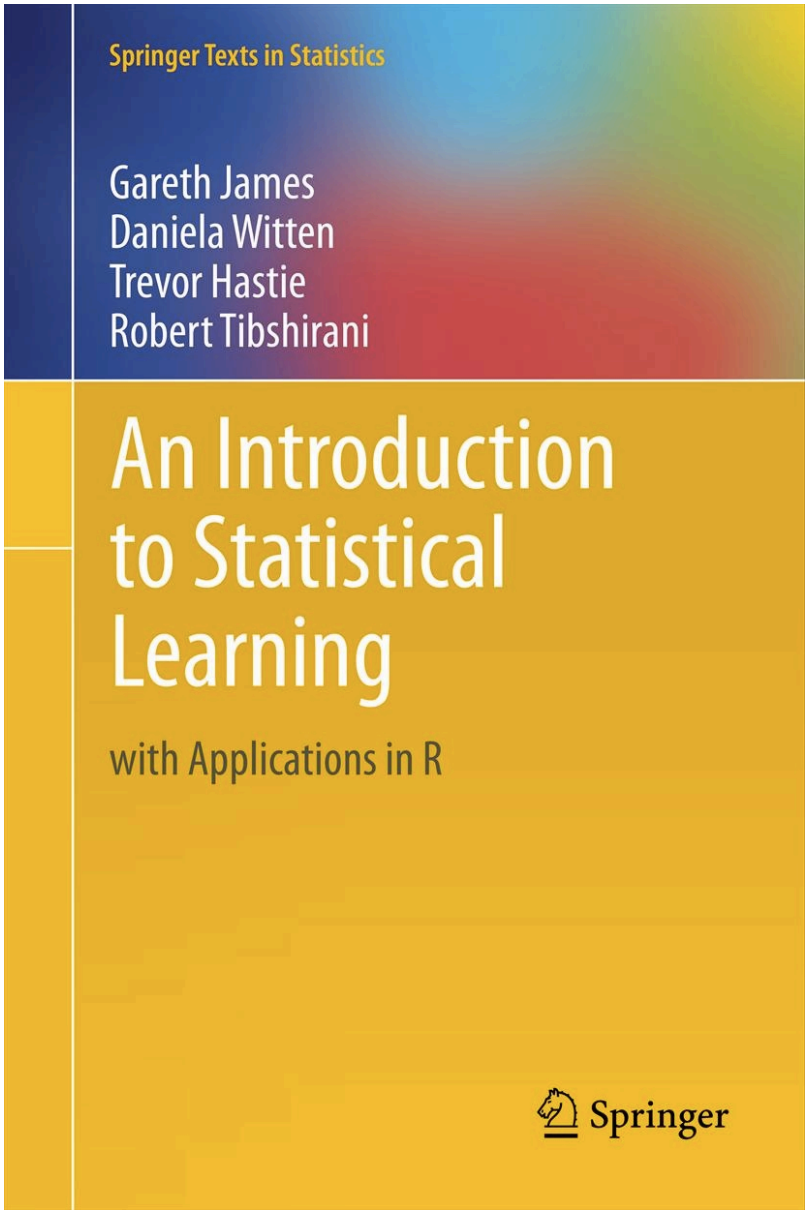


Coursera:
https://www.coursera.org/learn/machine-learning

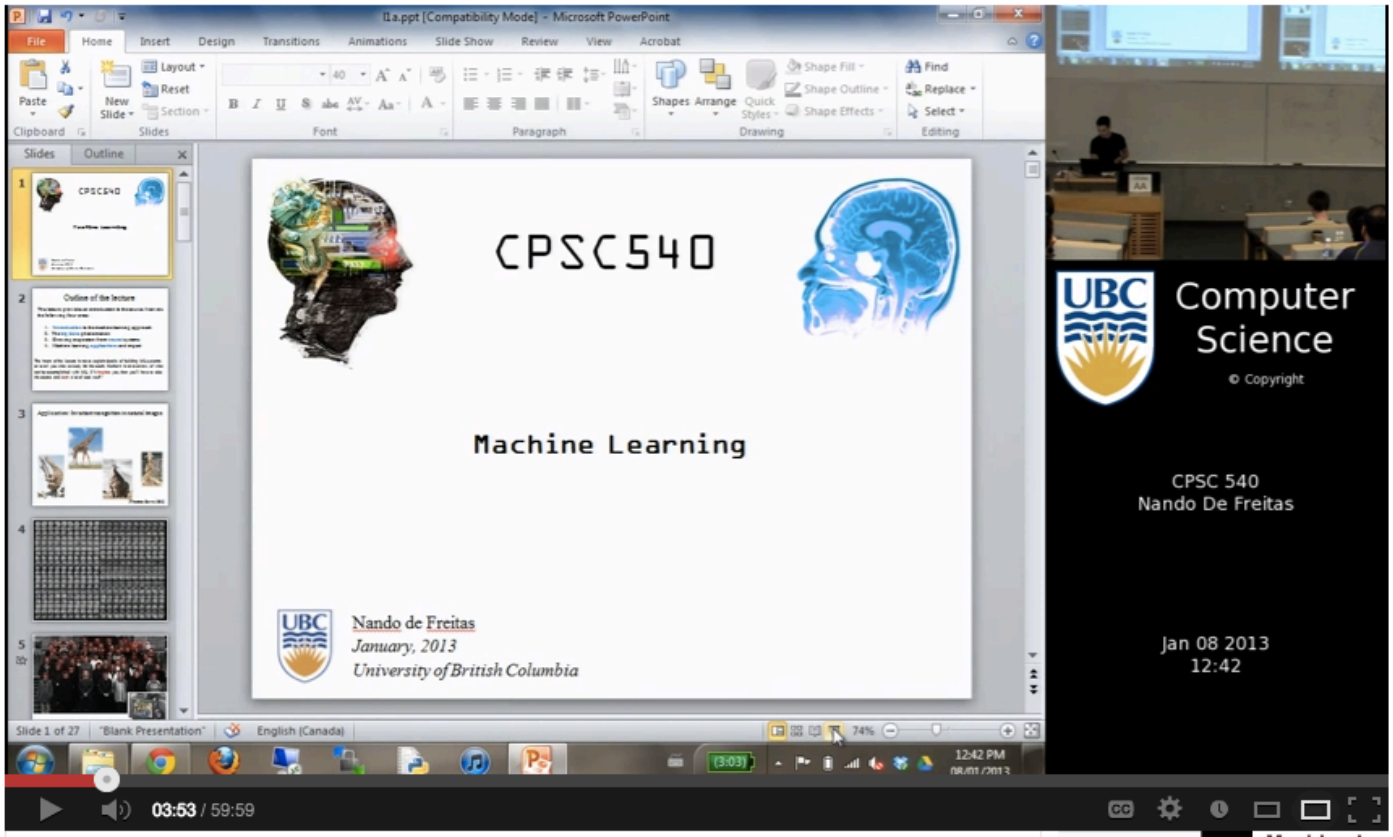
Stanford YouTube channel:
http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599



http://www.cs.waikato.ac.nz/ml/weka/book.html



http://www-bcf.usc.edu/~gareth/ISL/index.html



Lectures by Nando de Freitas

at the University of British Columbia

http://www.youtube.com/watch?v=w2OtwL5T1ow

RECOMMENDED READINGS

- [blog post] Designing ML models (link)
  - simple and well explained example of the creation and evaluation of a ML model
- [blog post] Preparing data for analysis (link)
  - concise explanation of why data preparation (cleansing and transformation) is (so much) important
- [article] A Tour of Machine Learning Algorithms (link)
  - nice, succinct overview of a large number of popular ML algorithms
- [post] Machine Learning Algorithm Cheat Sheet (link)
  - pros & cons of several popular algorithms (DT, Lin. Regression, NN, SVM, kNN); very practical and useful
- [article] Essentials of Machine Learning Algorithms (with Python and R Codes) (link)

Some (potentially) interesting ML application examples

- NELL - Never Ending Language Learner (website) (NYT article) (video lecture)
- Relationship mining on Twitter
- What Facebook Knows (data analysis at Facebook)
- Using Location Data to Predict the Events You Will Want to Attend
- Smart Autofill - Harnessing the Predictive Power of Machine Learning in Google Sheets
- Deep Learning (what it is about)
- Learning Analytics and MOOCs (TED talk) (research paper on the use of ML to predict dropouts in MOOCs)

(Anonymous) survey asking for your opinion, critique, praise, comments about this course:

http://goo.gl/cqdp3l