Machine Learning

Introduction

by *Marina Barsky*

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

- Machine learning teaches machines to **learn how to carry out tasks by themselves**, without giving explicit instructions

- How can a machine learn something new, if all the instructions are given by a human programmer?

- This is possible only if the instructions are of a special type: they mimic the ways that humans learn

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What do we mean by learning?

- Based on the previous experiences assign a label to a new object
- Group similar things together into a single category
- Identify patterns

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What do we mean by learning?

- Based on the previous experiences assign a label to a new object
- Group similar things together into a single category
- Identify patterns

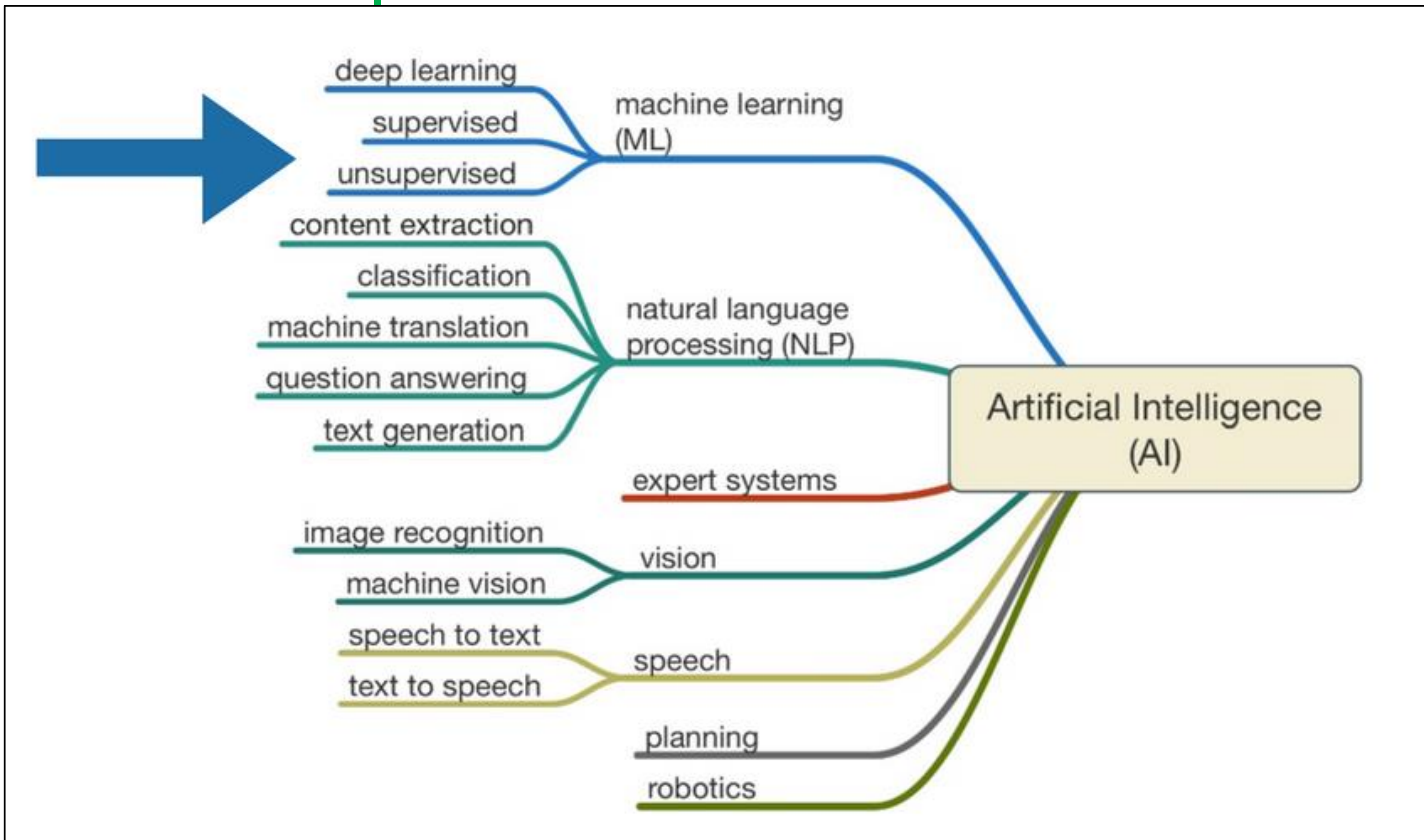
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

ML algorithms learn from previous experiences

- The previous experiences are encoded as a set of data points
- If data is non-random, it contains *patterns*
- Based on these patterns, ML algorithm discovers a *generalized model of data*
- That allows it to make *predictions* about other data that it might see in the future

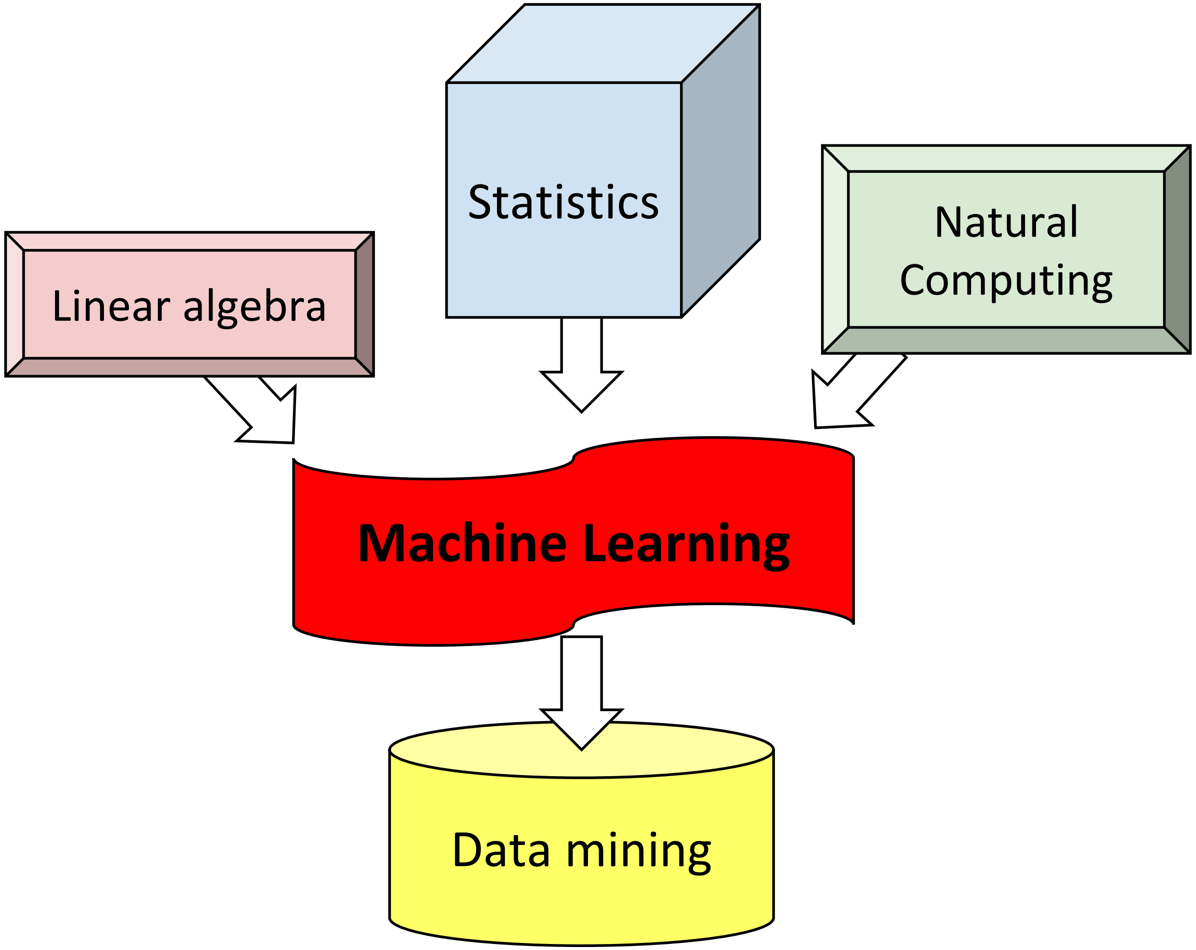
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Machine Learning is a subfield of AI



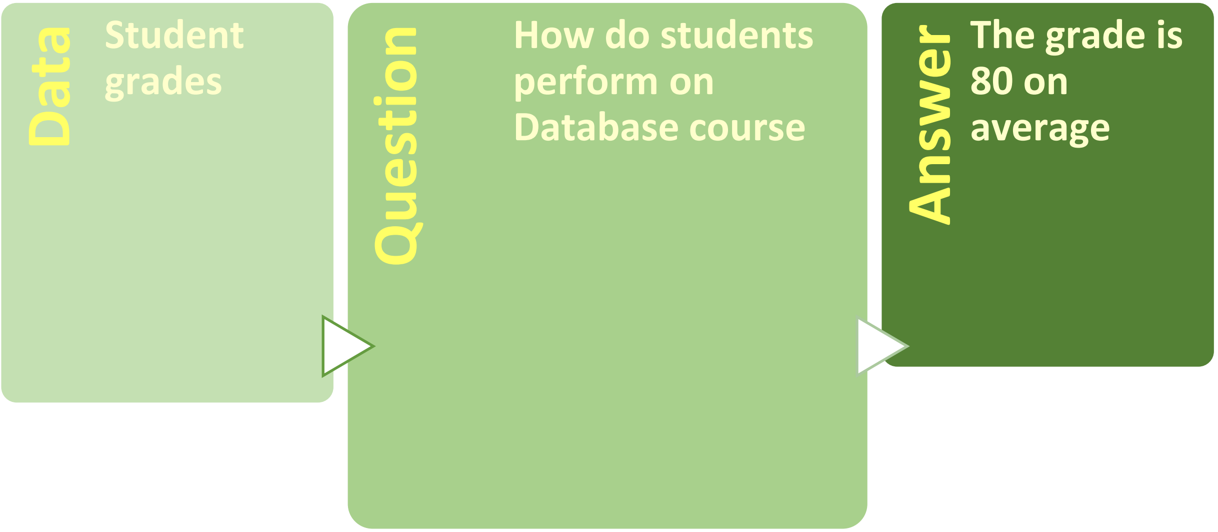
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Machine Learning overlaps



- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What is (not) Machine Learning



- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What is (not) Machine Learning

Data: Student grades

Question: How do students perform on Database course

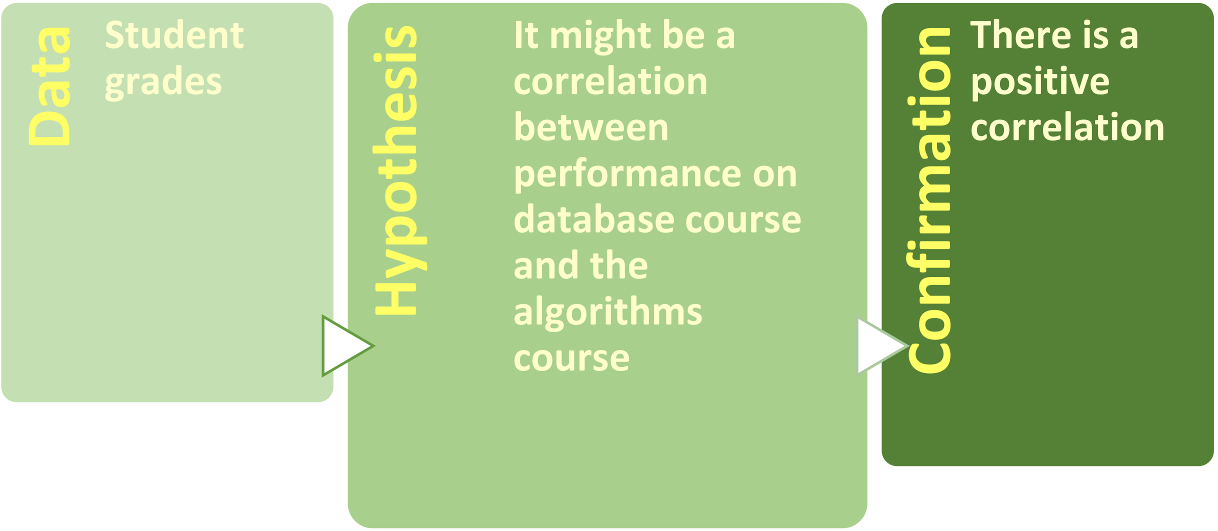
Answer: The grade is 80 on average

Not ML

- data manipulation (query)

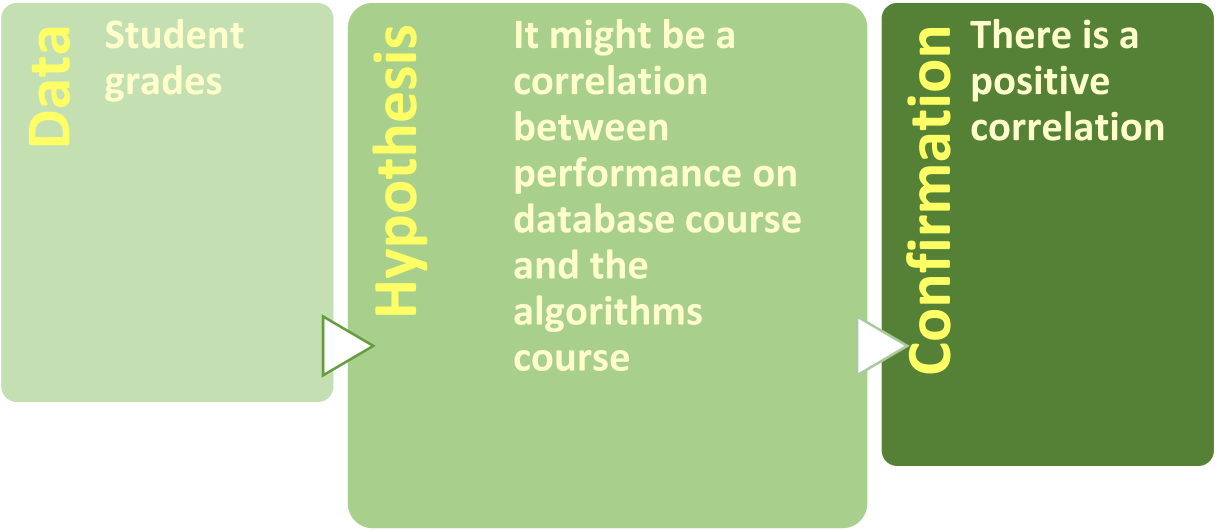
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What is (not) Machine Learning



- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What is (not) Machine Learning



Not ML problem

- pure statistics (hypothesis testing)

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

What is (not) Machine Learning

Data
- Student grades

Interestingness criteria
- Is there any correlation in performance on computer science courses?
- Given performance in some courses can we predict performance in the others?

Patterns
- If good at algorithms then good at Databases
- If good at Hardware courses, then not good at Java

Machine learning!

- What is machine learning
- **Why ML**
- Types of ML tasks
- Course requirements

Why study Machine Learning? 1/3

Get **competitive advantage** in **business**

- Google uses web links to rank pages, it gathers your every click and **learns to adapt** its search to your preferences
- Amazon and Netflix use information about the things people buy or watch to **learn which people or items are similar** to one another, and then make recommendations
- Pandora and Last.fm use your ratings of songs to create custom radio stations with music they think you will enjoy
- The predictions made by the Hollywood Stock Exchange are routinely better than those made by individual experts
- eHarmony uses information collected from participants to determine who would be a good match

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Why study Machine Learning? 2/3

In **science**:

- Classify faint galaxies
- Find similar gene expressions for different drug treatments
- Predict the structure of a chemical from magnetic resonance data

…

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Facial recognition?

http://www.pictriev.com/

Why study Machine Learning? 3/3

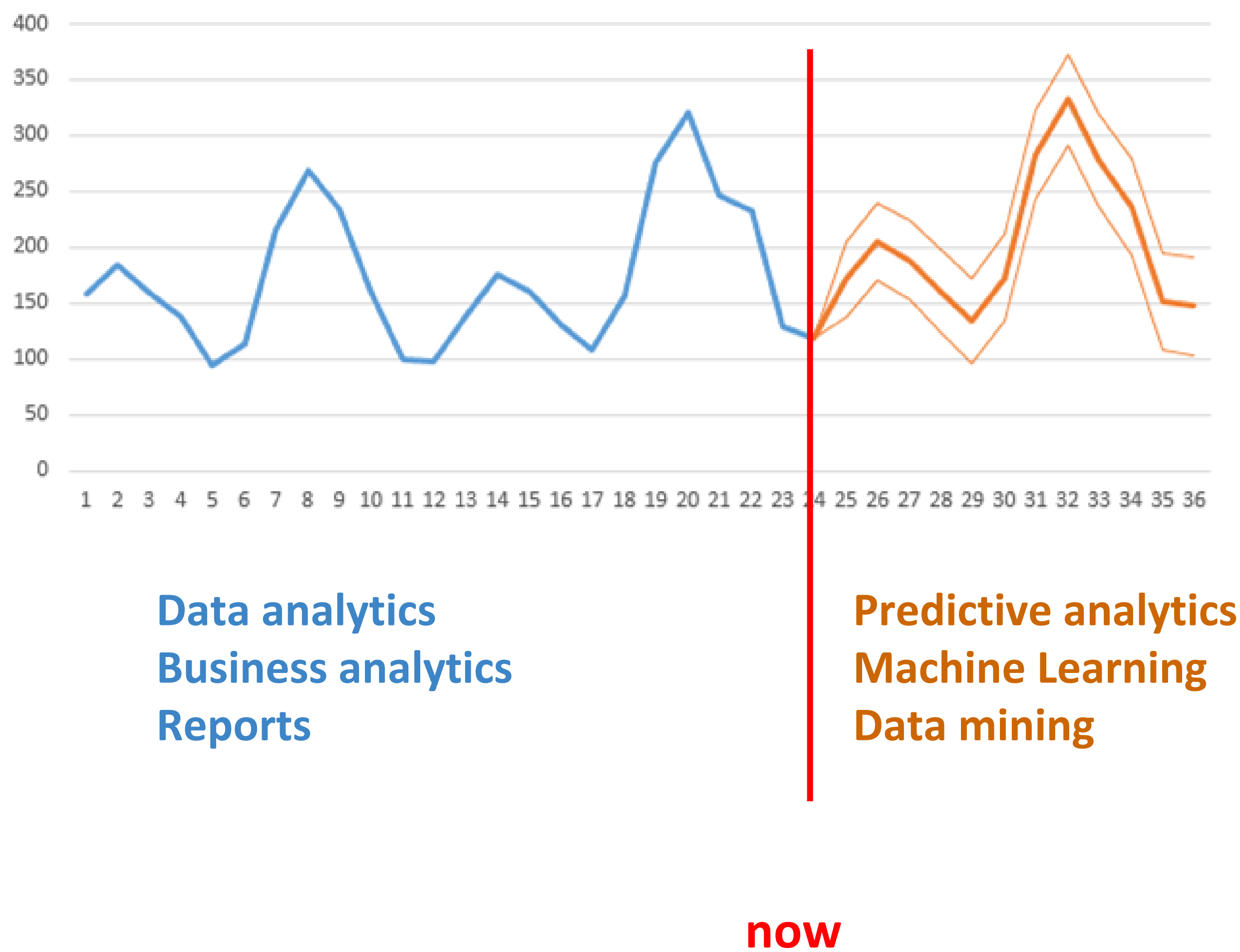
Automate everyday tasks

- Show to the algorithm which messages you consider a spam, and the task of separating spam can be carried out automatically
- Collect only positive or only negative news articles
- …

Once you learn about a few machine-learning algorithms, you'll start seeing places to apply them just about everywhere

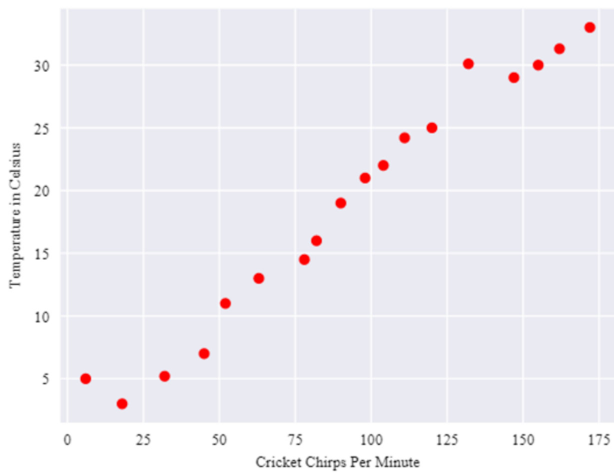
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

ML predicts the future

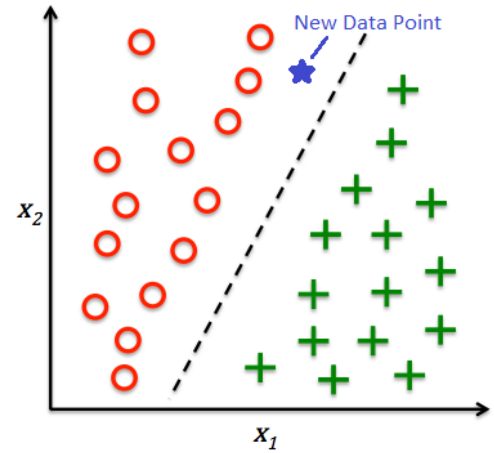


Types of learning tasks

Supervised learning

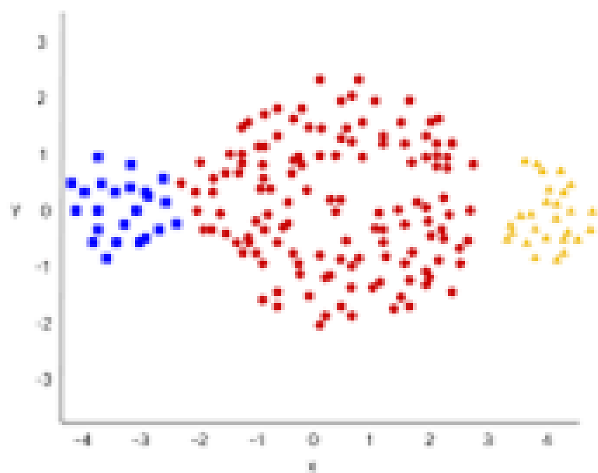


Prediction



Classification

Unsupervised learning



Clustering

TransactionId	Items
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}
5	{A,B,C,D}

Associations

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Supervised learning: Classification

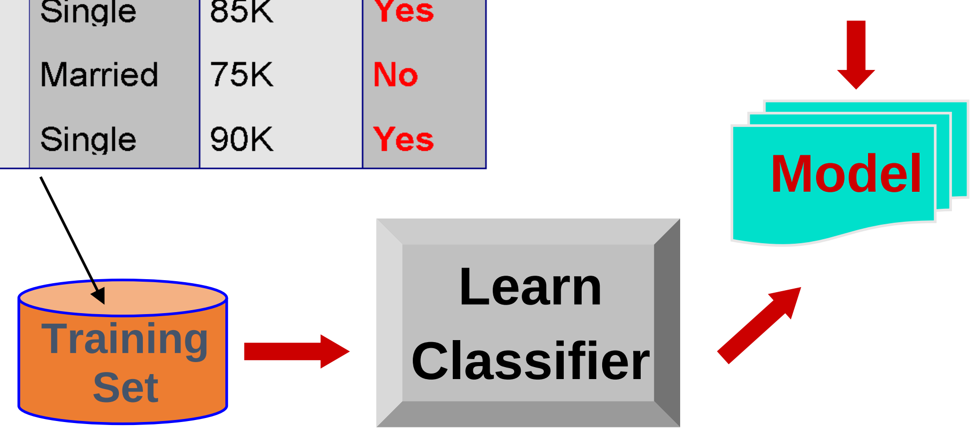
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find ("learn") a *model* for the class attribute as a function of the values of the other attributes.
- Objective: new **previously unseen** records should be assigned a class as accurately as possible.

- What is machine learning
- Why ML
- **Types of ML tasks**
- Course requirements

Classification example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Toy classification problem

My neighbour dataset

class label

Temp	**Precip**	**Day**	**Shop**	**Clothes**	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**

(Adopted from Leslie Kaelbling's example in the MIT courseware)

- What is machine learning
- Why ML
- **Types of ML tasks**
- Course requirements

Classify:

class label

Temp	Precip	Day	Shop	Clothes	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**?**

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Classify: memory

class label

Temp	Precip	Day	Shop	Clothes	
25	None	Sat	No	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**
15	Snow	Mon	Yes	Casual	**Walk**
-5	Snow	Mon	Yes	Casual	**Drive**

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Classification problem: noise

Temp	**Precip**	**Day**	**Clothes**	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**?**

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Classification: averaging

Temp	Precip	Day	Clothes	
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Drive**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**
25	None	Sat	Casual	**Walk**

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Classification: generalization

Temp	Precip	Day	Clothes	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Three main ideas used for classification:

- memory
- averaging
- generalization

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Unsupervised learning. Associations

The Market-Basket Model

- A large set of *items*, e.g., things sold in a supermarket.
- A large set of *baskets*, each of which is a small set of the items, e.g., the things one customer buys in one transaction.

Fundamental question

- What sets of items are often bought together?

Motivation

- If a large number of baskets contain both hot dogs and mustard, we can use this information. How?

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Solving association problem: market basket

	Transactions
1	{bread, milk, peanut butter}
2	{bread, milk}
3	{beer, potato chips}
4	{beer, diapers}
5	{beer, milk, diapers}
6	{bread, milk, yogurt}
7	{beer, bread, diapers}
8	{bread, milk, jelly}
9	{beer, cigarettes, diapers}
10	{bread, milk}

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Solving association problem: market basket

	Transactions
1	{**bread**, **milk**, peanut butter}
2	{**bread**, **milk**}
3	{**beer**, potato chips}
4	{**beer**, **diapers**}
5	{**beer**, **milk**, **diapers**}
6	{**bread**, **milk**, yogurt}
7	{**beer**, **bread**, **diapers**}
8	{**bread**, **milk**, jelly}
9	{**beer**, cigarettes, **diapers**}
10	{**bread**, **milk**}

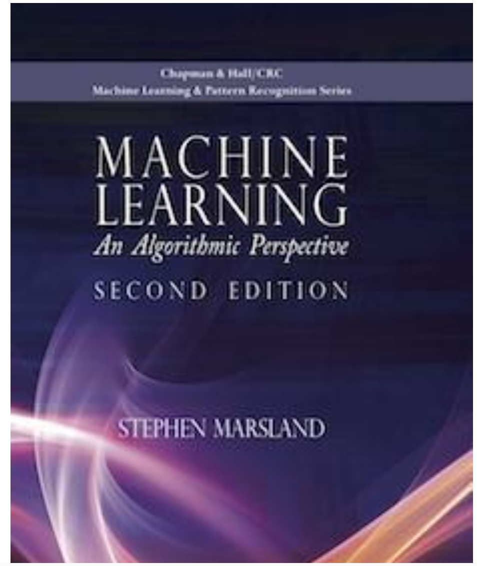
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Beer and diapers?

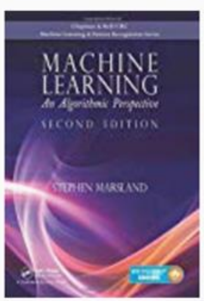
	Transactions
1	{bread, milk, peanut butter}
2	{bread, milk}
3	{**beer**, potato chips}
4	{**beer**, **diapers**}
5	{**beer**, milk, **diapers**}
6	{bread, milk, yogurt}
7	{**beer**, bread, **diapers**}
8	{bread, milk, jelly}
9	{**beer**, cigarettes, **diapers**}
10	{bread, milk}

- What is machine learning
- Why ML
- **Types of ML tasks**
- Course requirements

Amazon example

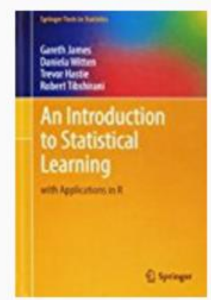


Customers who viewed **Machine Learning: An Algorithmic Perspective (Chapman & Hall/Crc Machine...** also viewed



Machine Learning: An Algorithmic Perspective, Second Edition
46
$69.29
prime
48 used and new from $59.61

Hands-On Machine Learning with Scikit-Learn and TensorFlow:
251
$29.35
prime
85 used and new from $22.86



An Introduction to Statistical Learning: with Applications in R (Springer Texts
200
$49.60
prime
20 used and new from $39.95

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

?

Customers Who Bought This Item Also Bought

Revere Polished Aluminum 8-Inch Nonstick Skillet by Revere
(16)
$14.99

Pyrex Smart Essentials 8-Piece Mixing Bowl Set by Pyrex
(66)
$26.82

Kodak Portra 400 Professional ISO 400, 35mm, 36 Exposures, Color...
(5)
$29.88

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

This course:
Computer Science part of Machine Learning

- We focus on ***algorithms***
- By the end you understand ideas behind ML algorithms
- You will experiment with code based on these algorithms and see by yourselves whether machines can or cannot learn

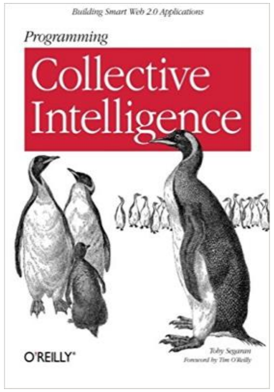
Far-reaching Course Objectives

- Develop interest in math as a tool for learning about the world
- Learn how to handle ambiguity
- Formalize mental models of learning
- Make your future programs smart by incorporating ML algorithms
- Invent new ML approaches and new algorithms

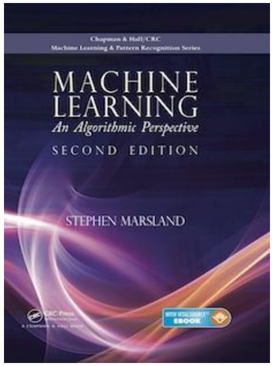
- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Books, blogs, videos

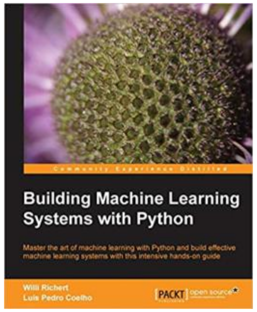
Programming Collective Intelligence
by Toby Segaran



Machine Learning: An Algorithmic Perspective
by Stephen Marsland



Building Machine Learning Systems with Python
by Willi Richert and Luis Pedro Coelho



Web sites:

- Analytics Vidhya
- Kaggle
- Datacamp
- KDNuggets

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Types of assignments

- Problem solving quizzes and short labs – every week: 35%
- Mini-projects – real coding, real problems: 35%
- Final project: large, professional, open-ended: 30%

- What is machine learning
- Why ML
- Types of ML tasks
- Course requirements

Learning by doing

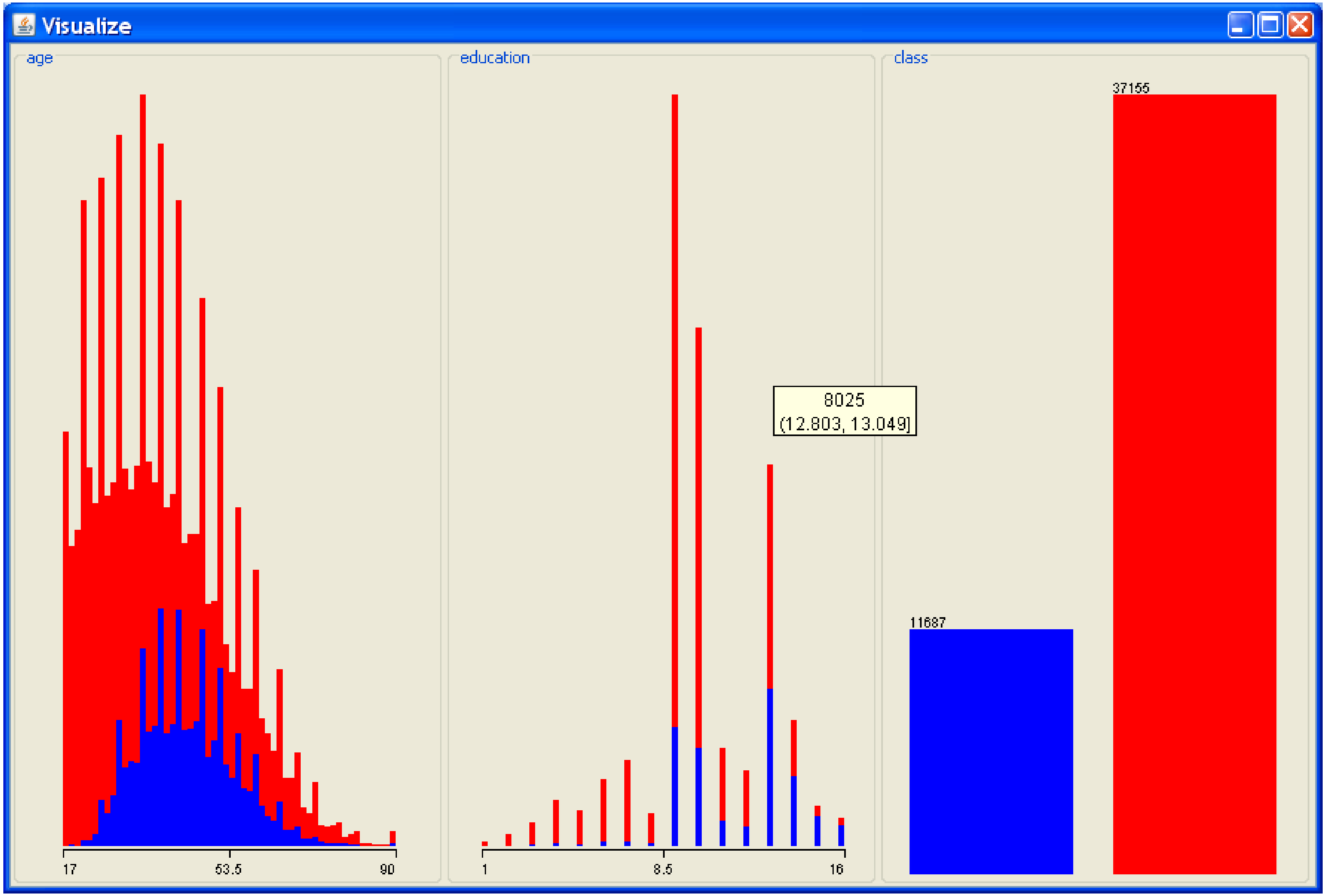
- Learning by example: on toy datasets which exhibit features of real-life datasets
- Implementation of some algorithms in Python
- Python library of ML algorithms: *sklearn*
- Analysis of real-life datasets

Mini-project example: what determines high salary

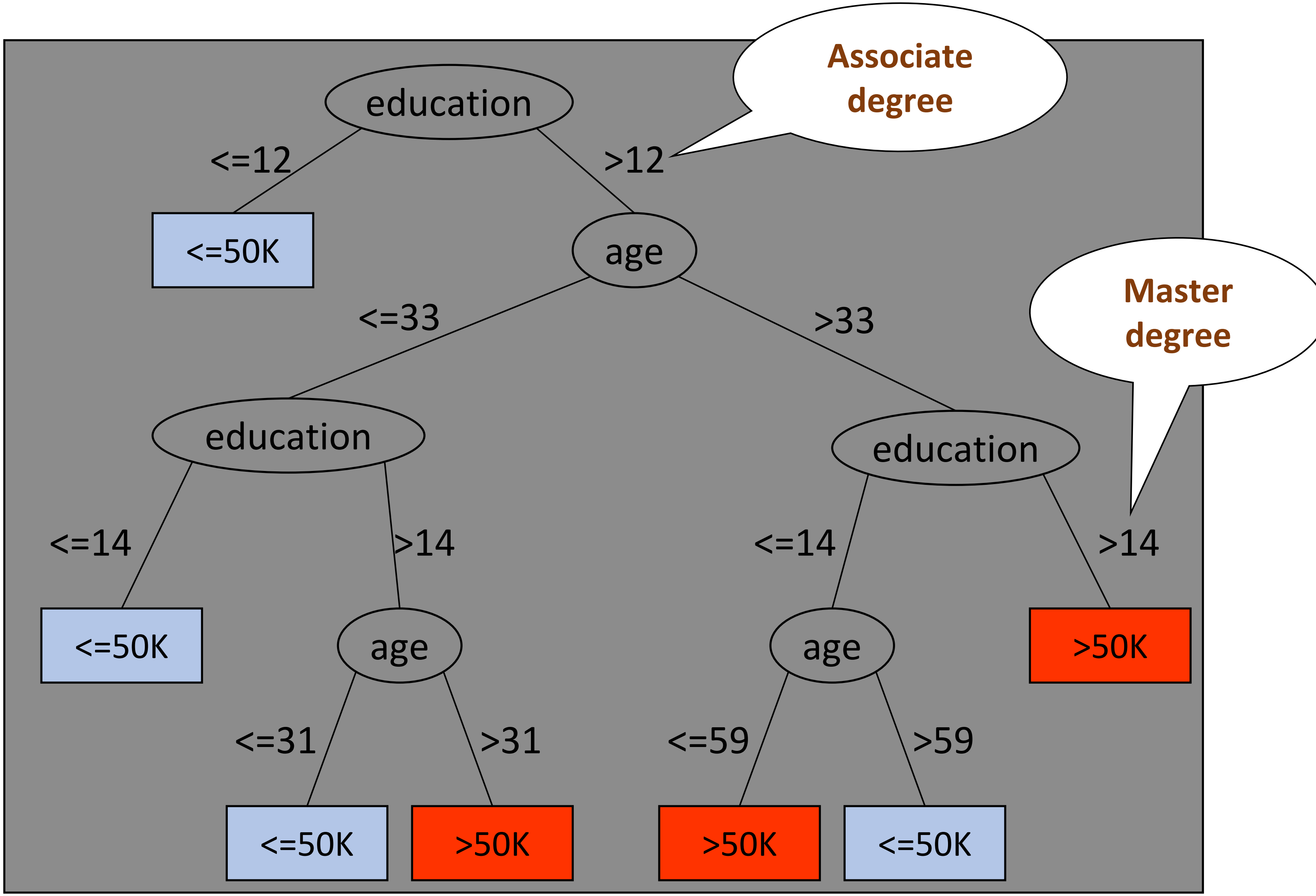
Adult income dataset (US census 1994)

Age	Education	Mar. status	Occupation	Race	Sex	Born in	Yearly income
39	Bachelors	Never-married	Adm-clerical	White	M	US	**<=50 K**
50	Bachelors	Married-civ-spouse	Exec-managerial	White	M	US	**<=50 K**
54	7th-8th	Married-civ-spouse	Machine-op-inspct	White	M	US	**>50K**
37	Bachelors	Never-married	Exec-managerial	Black	M	US	**>50K**
28	Bachelors	Married-civ-spouse	Prof-specialty	Black	F	Cuba	**<=50 K**
37	Masters	Married-civ-spouse	Exec-managerial	White	F	US	**<=50 K**

Visualization of attributes: *age* and *education*



The result of learning: decision tree on age and education attributes



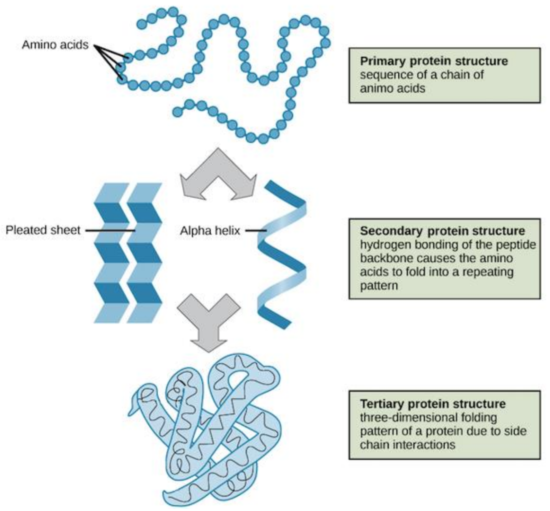
Sample Past Final Projects: 1/2

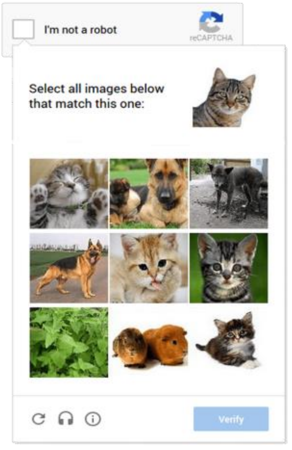
- **Predicting protein secondary structure using Recurrent Neural Networks** by *Joyee Wang*

Sample input sequence: Human Hemoglobin

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSA
QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVT
LAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

LINK





- **Solving Captcha Challenge with Convolutional Neural Networks** by *Aung Wai Yan Hein*

Can machines emulate humans and pretend that they are not robots?

LINK

- **Creating a Pokémon Battle AI with Decision Trees** by *Kai Dai*

LINK

Sample Past Final Projects: 2/2

- **Parsimonious Gymnosperms Identification using Decision Trees**

  LINK

- **Course Schedule Optimization with Genetic Algorithm**

  LINK

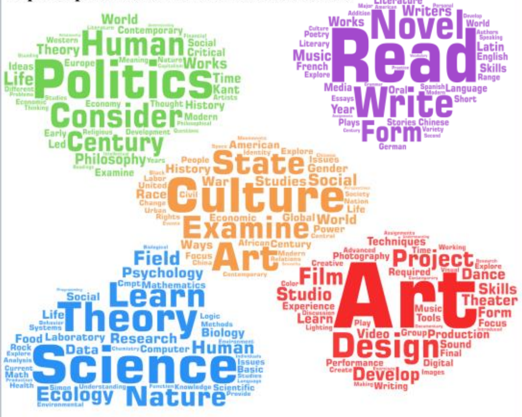
- **Study of International Pop Songs**

  LINK

- **Exploring Liberal Arts Curriculum Using Latent Dirichlet Allocation**

  LINK

Top 5 topics in Simon's Rock curriculum



Most frequent terms in Simon's Rock course descriptions

This llama word cloud represents the top 100 words in each topic combined. The most apparent topic words belong to Cultural Perspective and Literature. STEM and Social Sciences fade in the background

Course syllabus (minimalistic version)

Part I. Basic algorithms

Learning to optimize

1. Hill climbing, Simulated annealing, Genetic algorithm

Supervised learning

2. Decision trees and classification rules
3. Regression vs Logistic regression
4. Nearest neighbors

Unsupervised learning

5. Clustering
6. Associations and correlations

Part II. Advanced topics

Probabilistic classifiers

7. Naive Bayes. Bayesian Belief Networks
8. Evaluating and comparing classifiers

Artificial Neural networks

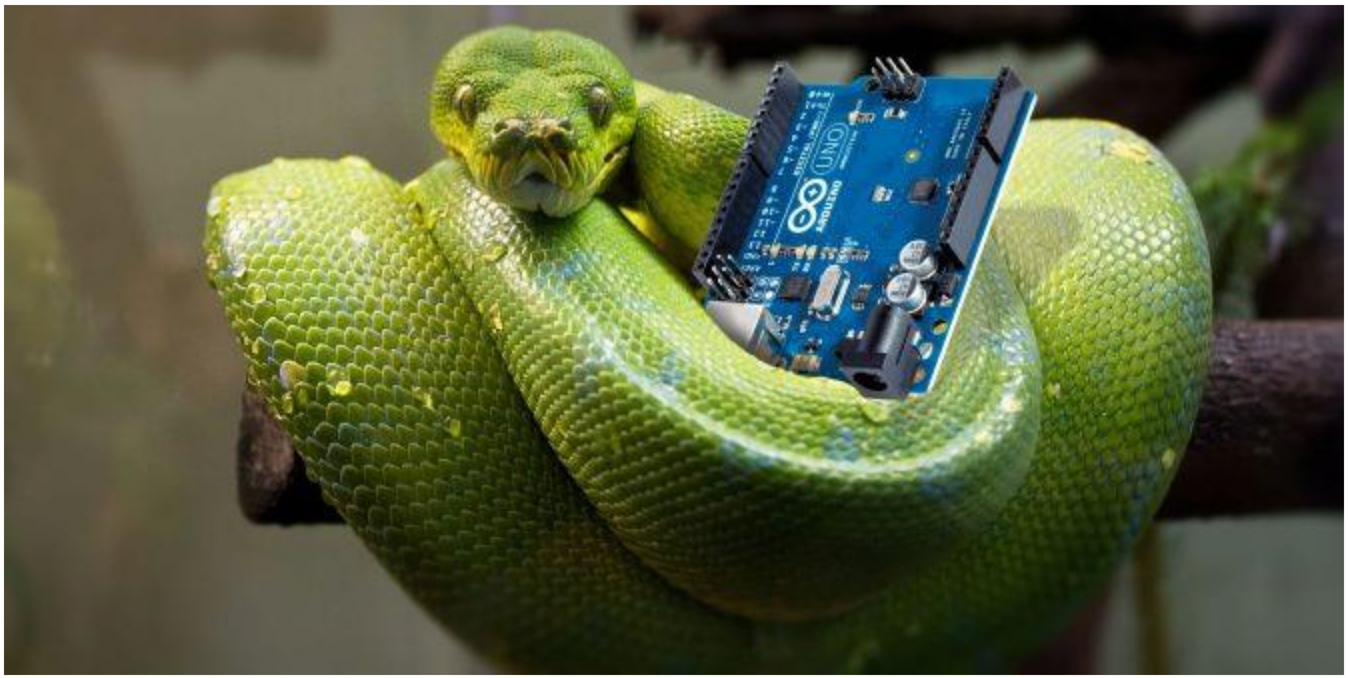
9. Classification with ANNs
10. Bias-variance trade-off. Regularization

Project ideas live web page: [LINK]

We all are familiar with Python

It's time to get a faster cleaner **Anaconda**

https://docs.anaconda.com/anaconda/install/



It comes with all the libraries plus Jupyter notebooks

Launch Anaconda Navigator →
Jupyter notebooks

Directories

Check what local directories you can access from Jupyter tree.

Create a lab folder inside one of these directories.

Create a separate folder where you are going to store all the datasets.

On github:

First, fork the repository, and then clone it into the lab directory.

Publishing notebooks

You can publish your notebooks using your google account in Google colab:

https://colab.research.google.com/notebooks/intro.ipynb

If you have a nice notebook for one of kaggle datasets, you can also publish it on kaggle:

https://www.kaggle.com/notebooks

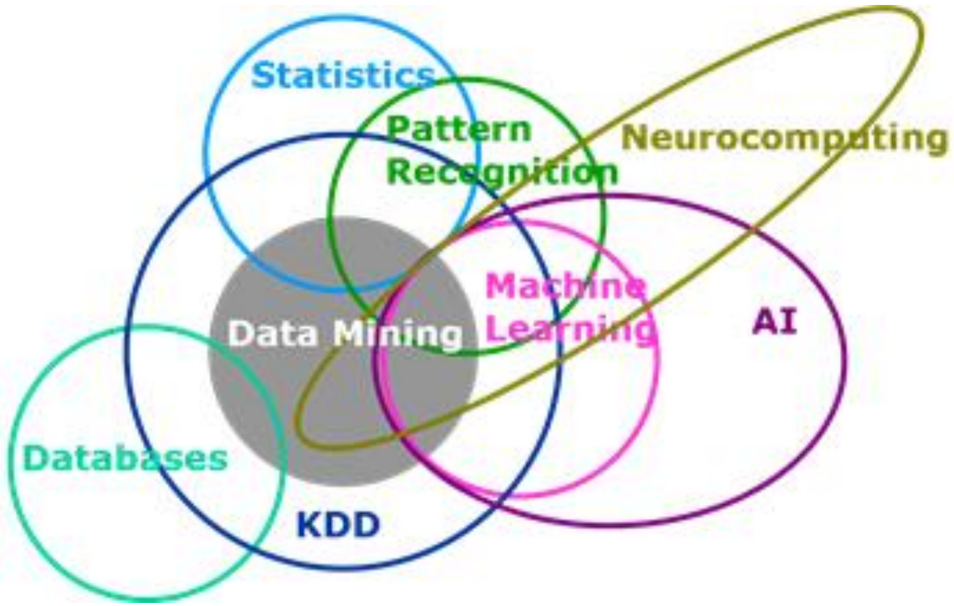
To avoid any problems with sharing and security, we are going to use local notebooks.

You can publish the notebook with your final project.

To do (a lot):

- Take Quiz 0: what is ML? We will discuss the answers in the next meeting
- Research and answer the writing prompt in Assignment 0.
- Install Anaconda
- Get familiar with Jupyter notebooks: follow this tutorial
- Create a github account

What is the difference?



Some discussions

- https://discuss.analyticsvidhya.com/t/what-is-the-difference-between-machine-learning-data-analysis-data-mining-data-science-and-ai/572
- https://bernardmarr.com/what-is-the-difference-between-data-mining-and-machine-learning/#:~:text=Data%20mining%20is%20used%20on,predictions%20about%20new%20data%20sets.
- https://www.simplilearn.com/data-mining-vs-machine-learning-article
- https://www.quora.com/Whats-the-relationship-between-machine-learning-and-data-mining
- https://www.kdnuggets.com/2021/11/3-differences-coding-data-science-machine-learning.html
- https://www.researchgate.net/post/What_is_the_difference_between_machine_learning_and_data_mining2
- https://www.kdnuggets.com/2016/11/machine-learning-vs-statistics.html

Practitioner Observations

- Statistics: applies statistically sound sampling techniques to make the input small. Many famous algorithms used in ML and DM are invented by statisticians and are a part of Statistics
- Machine Learning: encompasses all the techniques and algorithms which allow machines extract new insights from data
- Data Mining: adopts all of the Machine Learning algorithms plus their efficient implementation for very large datasets (Big Data, parallel processing)
- Data Science: a new way of learning about the world – from data

Evolution of Science

- Empirical Science – collect and systematize facts
- Theoretical Science – formulate theories and empirically test them
- Computational Science – run automatic proofs, simulations
- **e-Science** (Data Science) – collect data without clear goal - and test theories, find patterns **in the data itself**

SLOAN DIGITAL SKY SURVEY

Science is about asking questions

Traditionally: "Query the world"

Data acquisition for a specific hypotheses

Data science: "Download the world"

Data acquired en masse in support of future hypotheses

Computational challenge

The cost of data acquisition has dropped

The cost of **processing**, **integrating** and **analyzing** data is the new bottleneck

"…the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences – arguably the key scientific theme of our times"

F. Diebold

Efficient data manipulation

Poll: How much time modern scientists spend "handling data" as opposed to "doing science"?

Mode answer: 90%

"the Next Wave of InfraSress" *(*J. Mashey)