

למידה חישובית

כריית מידע

250200

הרצאה 1 - מבוא

ד"ר אריאלה ריכרדסון – קמפוס טל

(ד"ר יוסף סולביץ' – קמפוס לב)

yosef.solewicz@gmail.com

קצת על הקורס

• מודל / DRIVE

• סילבוס

מקורות מידע מומלצים

- Books

- Hands-on Machine Learning with Scikit-Learn and TensorFlow
 - יש קישור במודל

- Online courses

- Introduction to Machine Learning from Intel
 - (<https://www.intel.com/content/www/us/en/developer/learn/course-machine-learning.htm>).
- Machine Learning Crash Course at Google
 - (<https://developers.google.com/machine-learning/crash-course>).

- Web sources

- Stack overflow
- GitHub

שימוש במקורות בקורס (ובכלל)

- לא להעתיק באופן "עיוור"

- תקראו

- תלמדו ותבינו

- תשלבו מספר מקורות

- תבצעו התאמות

- תרחיבו

- חובה לציין את המקור/ות ולתת קרדיט כנדרש

זכויות יוצרים של מצגות הקורס

- השקפים גזורים מתוך מגוון מקורות, ובעיקר:

- קורס Introduction to Machine Learning של Intel

- <https://www.intel.com/content/www/us/en/developer/learn/course-machine-learning.html>

- ובמודל, ודרייב לפי הצורך עם דגשים

- שקפים של מרצי קורסים במרכז האקדמי לב משנים קודמות

- ד"ר יוסף סולביץ', ד"ר אלישבע בנש"ק-דוקוב, ד"ר מאיר גולדנברג, ד"ר אריאלה ריכרדסון ועוד

- מקורות נוספים על פי הצורך

דרישות (הרחבה על מה שכתוב בסילבוס)

- תרגילים/פרויקט 50%
- 10% תרגילים שבועיים (של תחילת הקורס)
- 40% פרויקט (פרטים בהמשך)
- כולל רכיב של השתתפות פעילה בקורס (לשיקול דעת המרצה)
- עבודה מותרת בזוגות (ולא יותר), הציון אישי, שני בני הזוג מחויבים לבקיאות
- מבחן 50%
- (לא בזוגות...)
- ראש גדול ובוגר
- מטלות 'פתוחות'
- נוכחות אינה חובה, אבל מומלצת

מבוא ללמידה חישובית / כריית
מידע

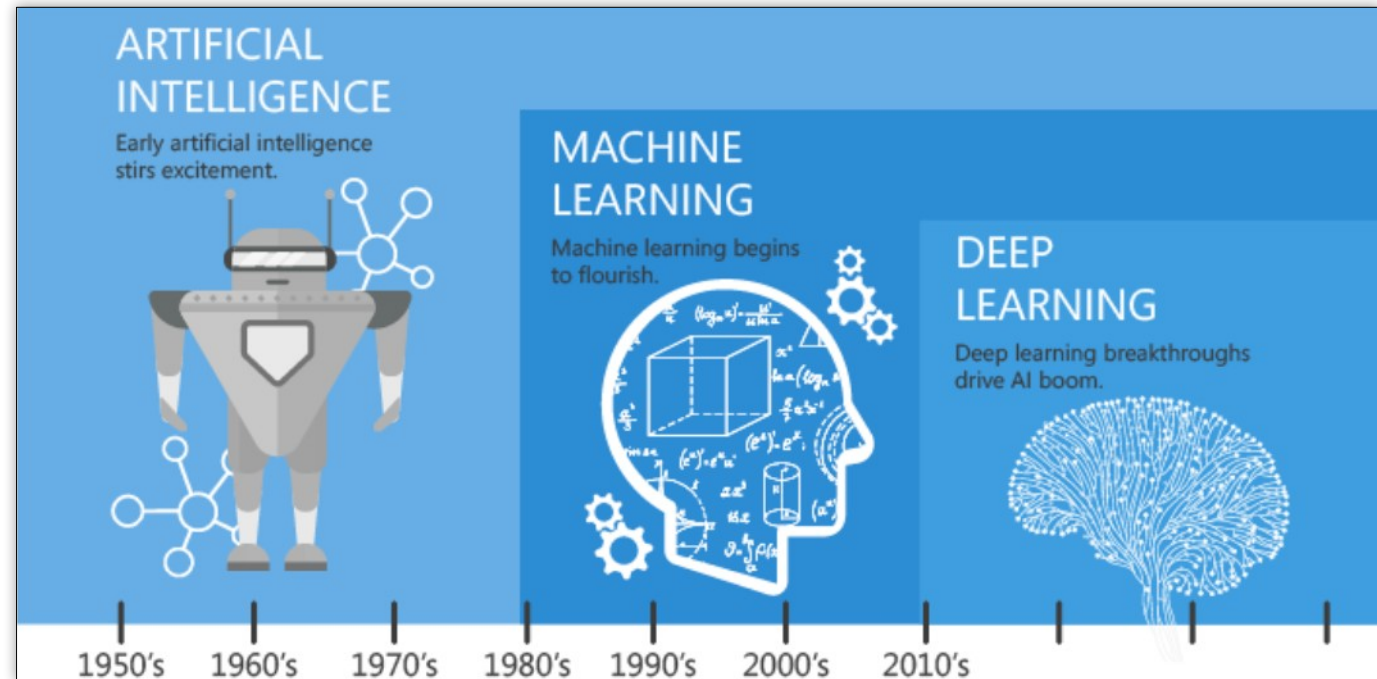
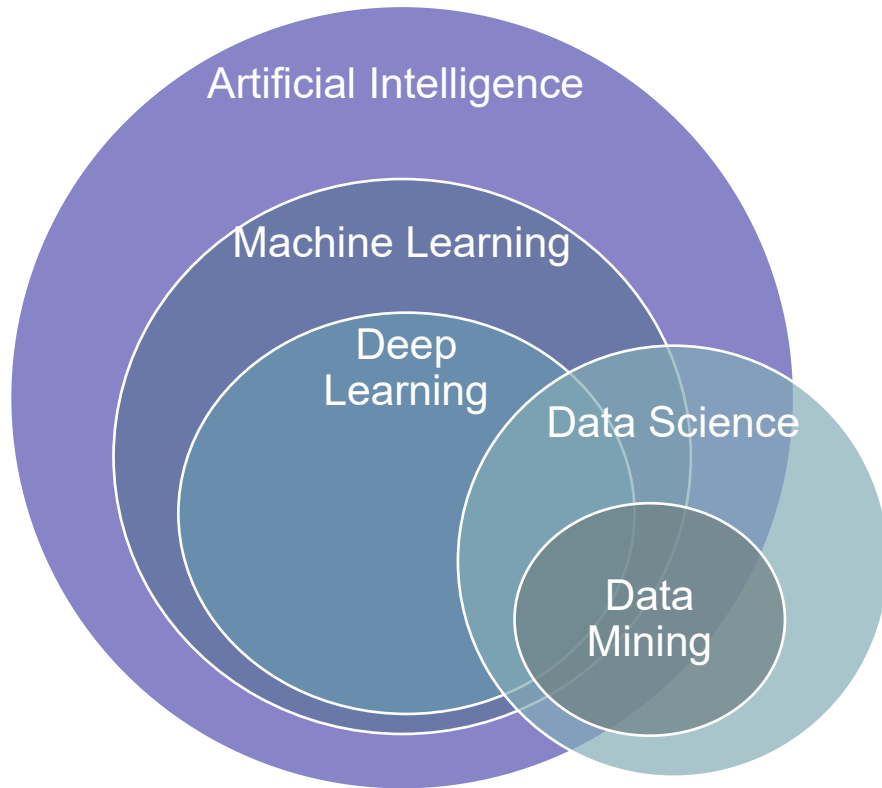
What is the course about?

- What is Machine Learning?
- What is Data Mining?
- Data
 - Examples, Features, Labels...
- Types of Learning
 - Supervised / Unsupervised / Reinforcement
- Evaluation Measures

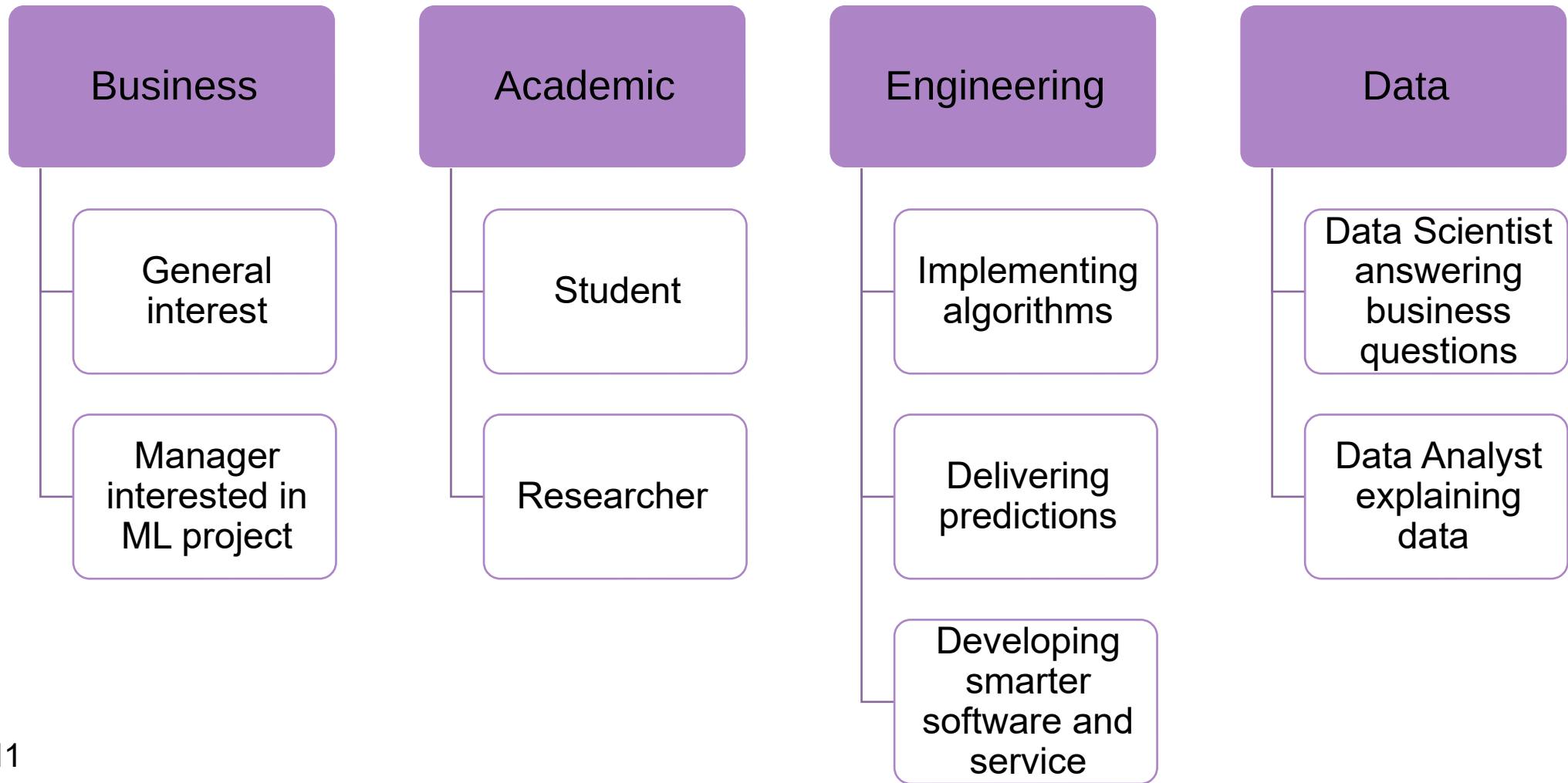
Machine Learning / Data mining (Wikipedia)

- **Machine learning (ML)** is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of **artificial intelligence**
- **Data mining** is the process of extracting and discovering patterns in large **data sets** involving methods at the intersection of **machine learning**, **statistics**, and **database systems**.

Or...



What can I do?



What do I need?

- ML basics, linear algebra, derivatives, statistics
- Python, TensorFlow
- Spark (will learn in BigData course)
- Pandas, Scikit-learn



- How much of each? Depends on what you do...

Programming

- Language - Python
 - <https://www.learnpython.org/>, <https://docs.python.org/3/tutorial/>
 - מהשקפים בתואר ראשון - במודל PYTHON
- Python Packages:
 - Numpy, Pandas, Matplotlib, Scikit-Learn
- Platform - Jupyter Notebook
 - Local: WinPython, Anaconda,...
 - Cloud:
 - Google's Colab (nice overview - <https://www.youtube.com/watch?v=inN8seMm7UI>)
 - Kaggle (nice overview - https://youtu.be/TNzDMOg_zsw)

Examples of problems

Spam Filtering

Web Search

Postal Mail Routing

Fraud Detection

Movie
Recommendations

Vehicle Driver
Assistance

Web Advertisements

Social Networks

Speech Recognition

Data (example, labeled)

- CSV file (Excel)

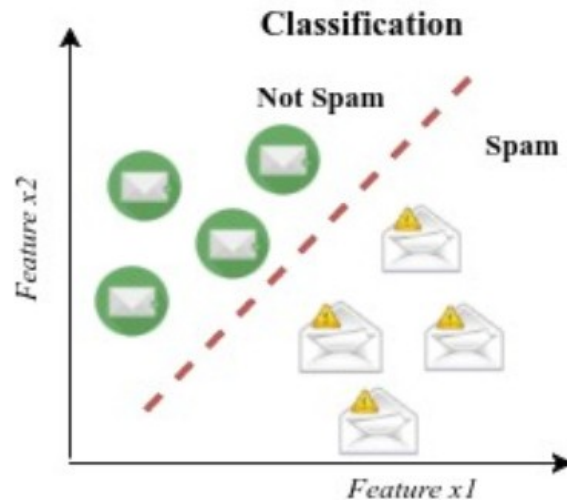
| sepal_length | sepal_width | petal_length | petal_width | species |
|--------------|-------------|--------------|-------------|-----------------|
| 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 5.1 | 2.5 | 3 | 1.1 | Iris-versicolor |
| 5.7 | 2.8 | 4.1 | 1.3 | Iris-versicolor |
| 6.3 | 3.3 | 6 | 2.5 | Iris-virginica |
| 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |

- Examples- rows
- Attributes (Features)- columns
- Labels- first/last column



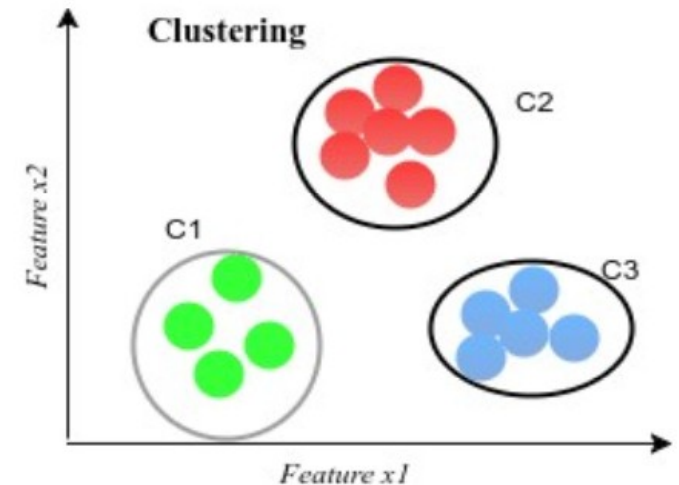
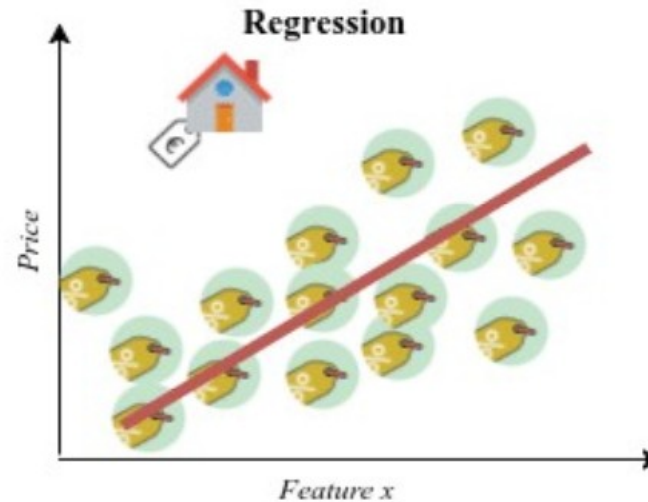
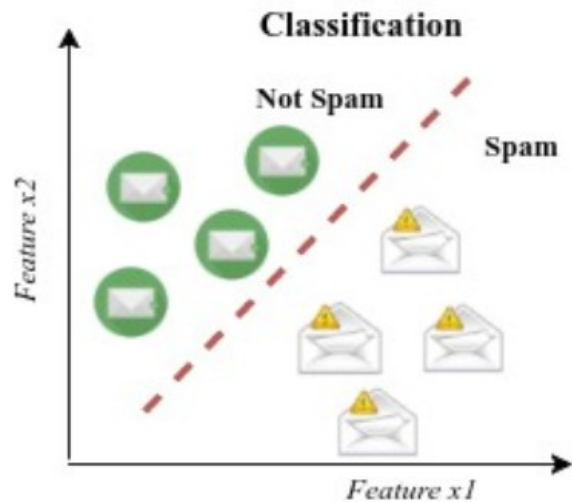
Iris Data – type of problem

- Supervised – labeled data
 - Classification (type of Iris)



Other Types of ML Problems

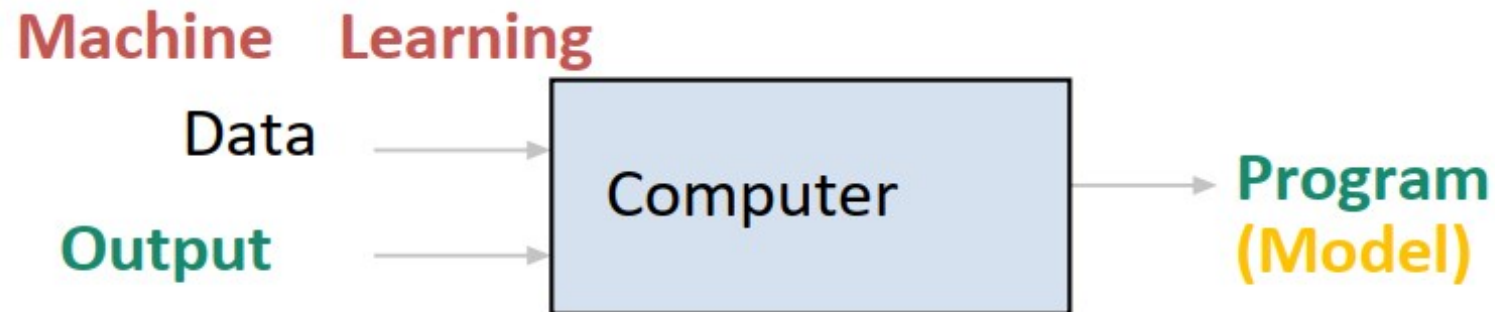
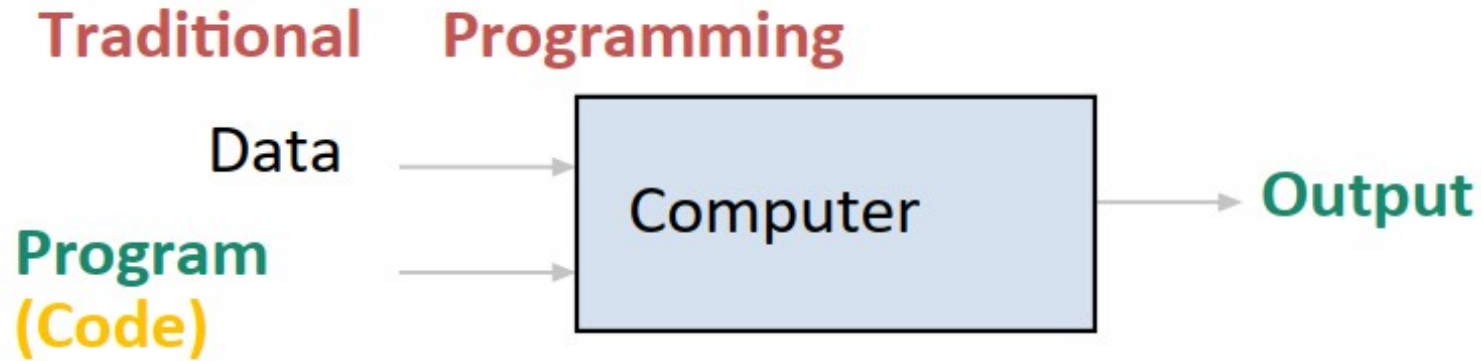
- Supervised – labeled data
 - Classification, Regression
- Unsupervised – unlabeled data
 - Clustering



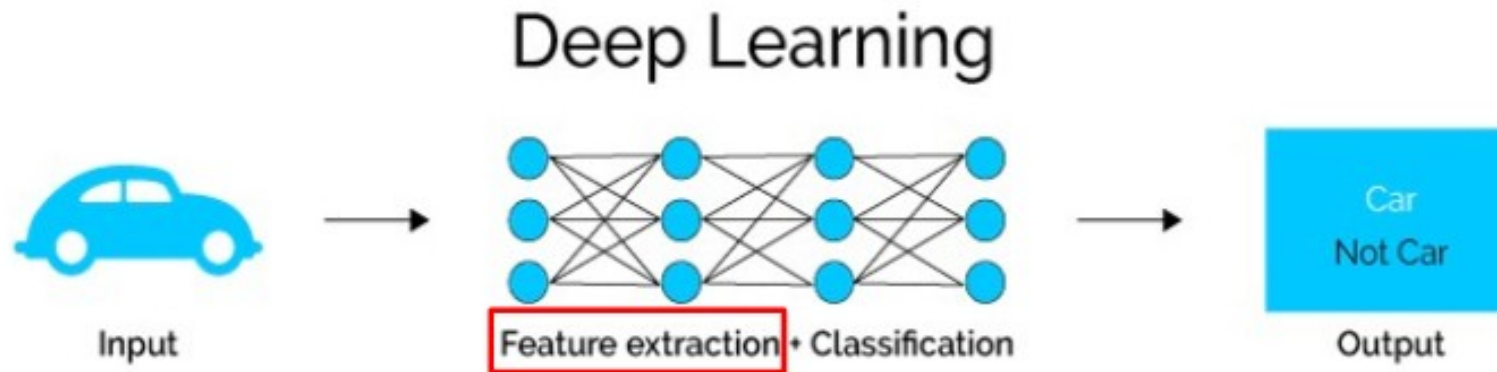
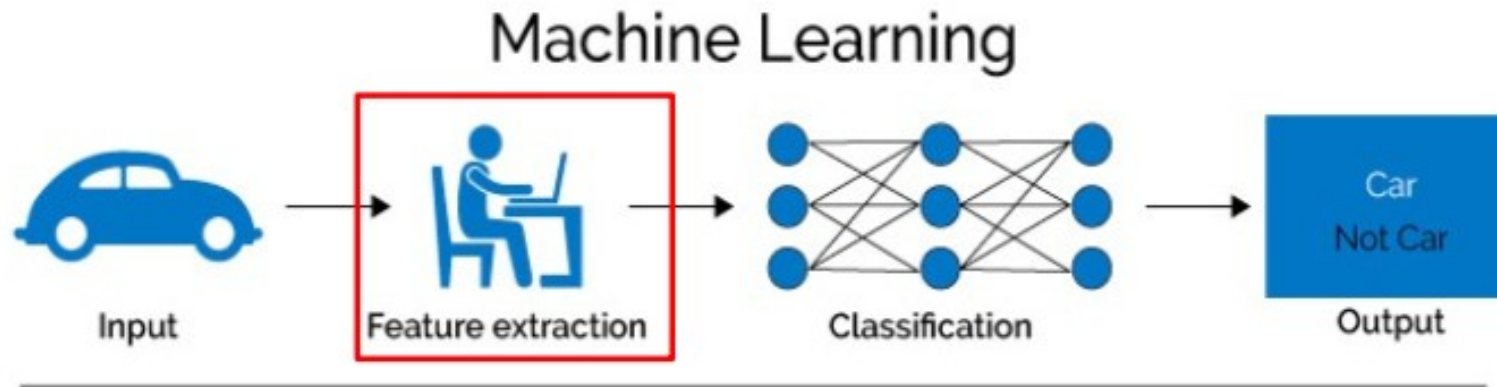
Evaluation measures (later in the course)

- How good is our model?
 - We need to define measures
 - Precision, Recall, Accuracy...
- Training Data vs. Testing Data
 - What is the difference?
 - Why do we need them?
- What type of mistakes does our model make?
 - False positive, False negative...

Traditional Programming vs ML



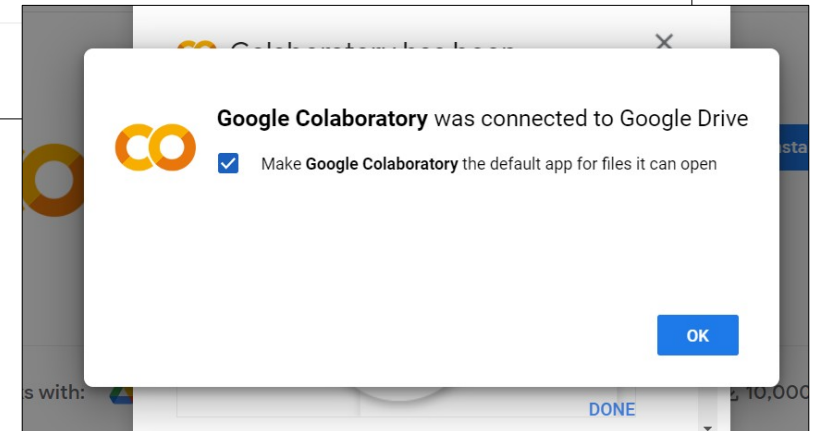
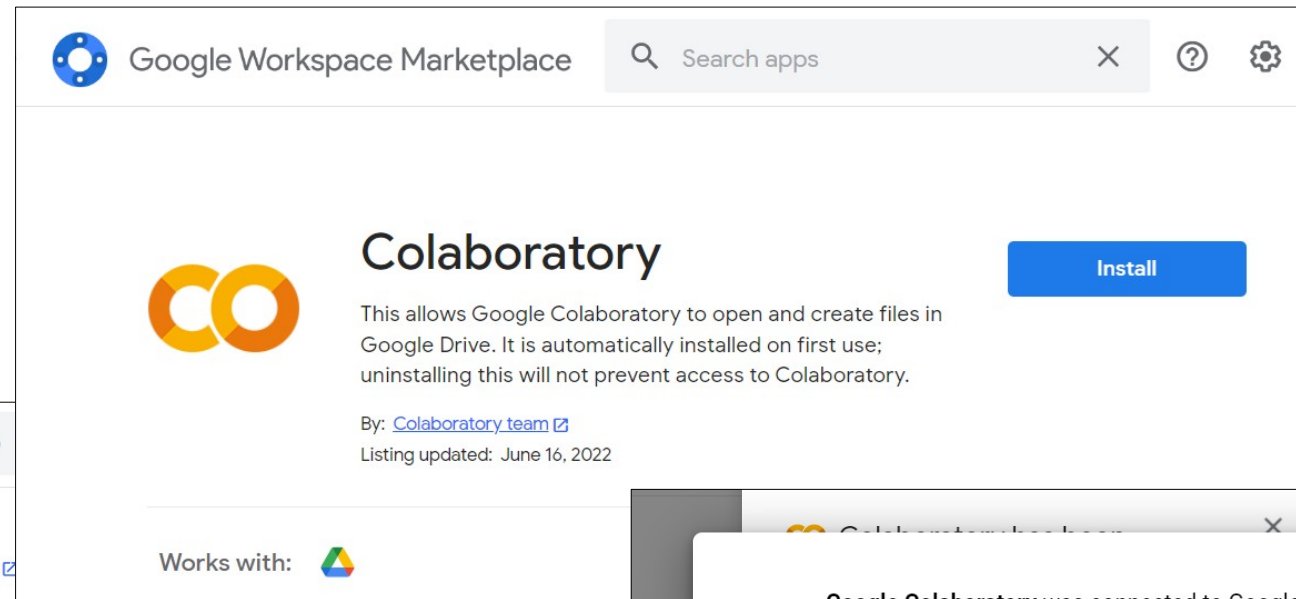
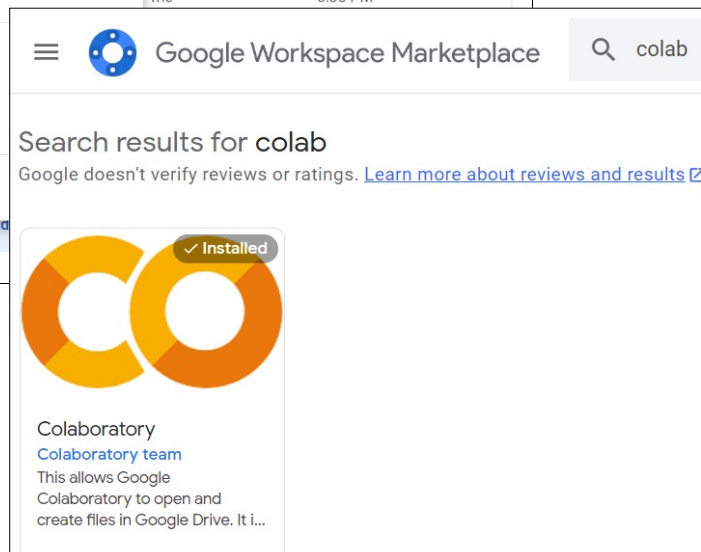
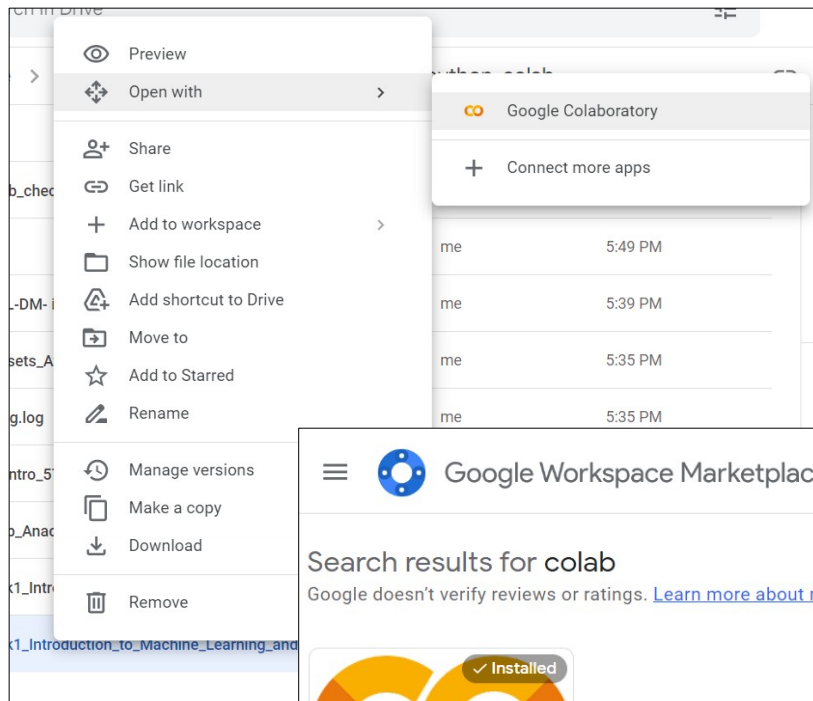
Machine Learning vs Deep Learning



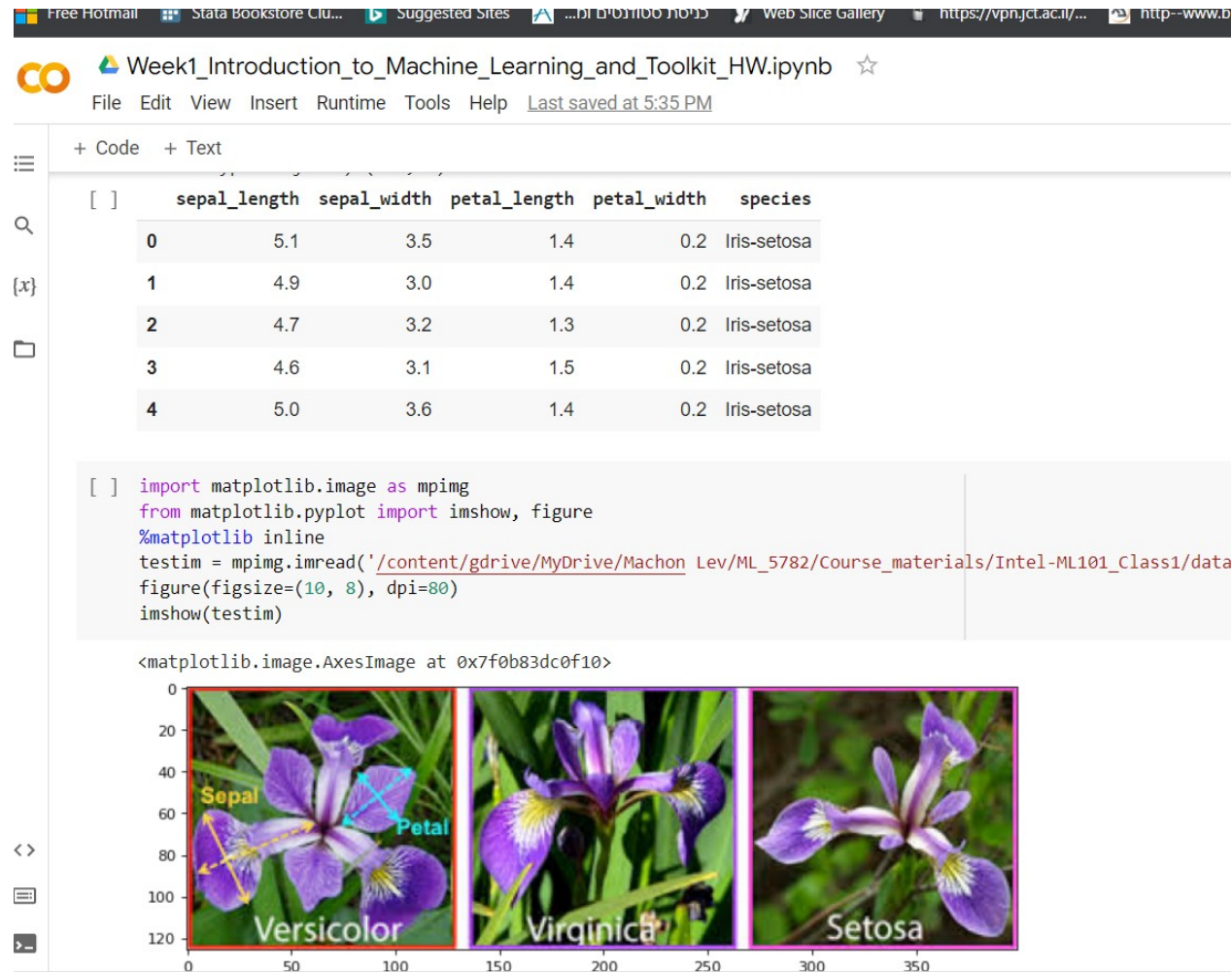
COLAB

<https://www.youtube.com/watch?v=inN8seMm7UI>

פתיחת קבצי Jupyter ב COLAB וחיבור ל DRIVE



עבודה ב COLAB , עם NOTEBOOK



המשך חיבור ל DRIVE

- מריצים את התא, ומקבלים בפעם הראשונה בקשת גישה
- בהמשך קוד לאבטחה

▼ Introduction to Machine Learning and Toolkit Exercises

```
from google.colab import drive  
drive.mount('/content/gdrive')
```

Introduction

We will be using the iris data set for this exercise. The data we will use are in a file called Iris

▼ Question 1

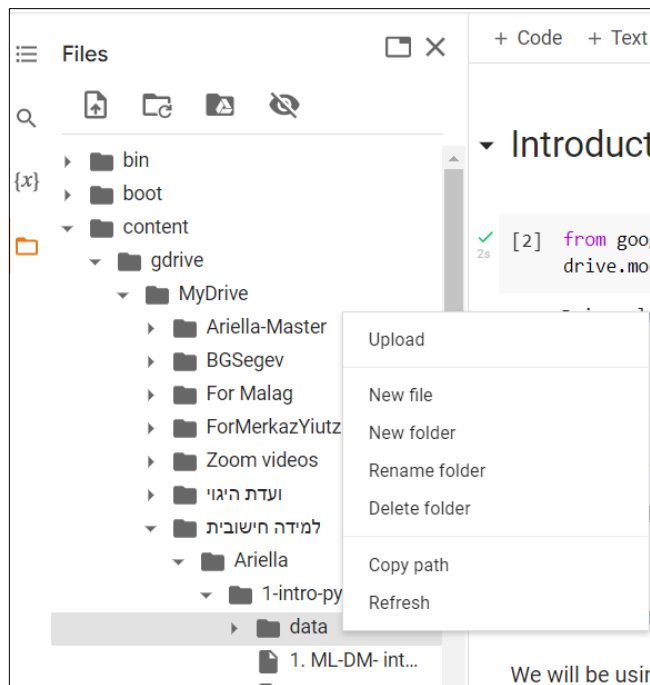
Permit this notebook to access your Google Drive files?

This notebook is requesting access to your Google Drive files. Granting access to Google Drive will permit code executed in the notebook to modify files in your Google Drive. Make sure to review notebook code prior to allowing this access.

No thanks

[Connect to Google Drive](#)

חיבור לתיקיית DATA ב COLAB



```
+ Code + Text
```

```
import numpy as np
import pandas as pd

#filepath = os.sep.join(data_path + ['Iris_Data.csv'])
filepath = '/content/gdrive/MyDrive/למידה חישובית/Ariella/1-intro-python-colab/data/Iris_Data.csv'
data = pd.read_csv(filepath)
print(data.columns,data.shape)
data.head()
```

Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
 'species'],
 dtype='object') (150, 5)

| | sepal_length | sepal_width | petal_length | petal_width | species |
|---|--------------|-------------|--------------|-------------|-------------|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |

Pandas

כמה מילים על PANDAS

- לא צריך המון פרטים, נלמד על הדרך
- מומלץ לעיין כאן
 - https://pandas.pydata.org/docs/getting_started/intro_tutorials/01_table_oriented.html#min-tut-01-tableoriented
- מומלץ גם לעיין בשקפים של INTEL
- ובכל מקור אחר שיעזור

הכרות עם סביבת העבודה

והכנה לתרגיל הבית

ועוד

- יש עוד הרבה מה לגלות על עבודה עם Colab, Jupyter Notebook וכו'
- כאן נכנסת האחריות שלכם דרך תרגילי הבית
- חצי שני של השיעור יהיה להראות את ה Notebook
- ש"ב במודל
- בהצלחה!