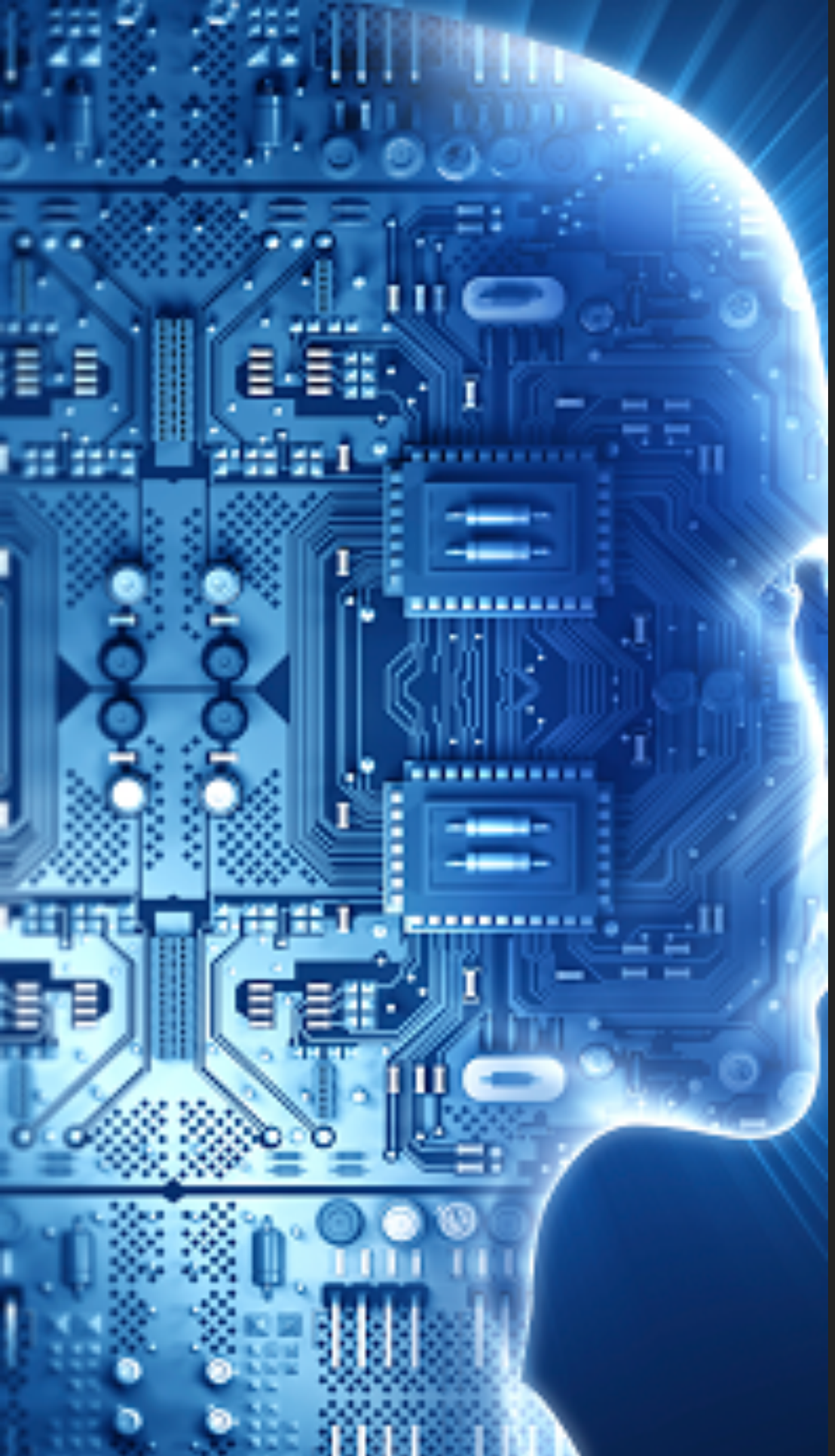
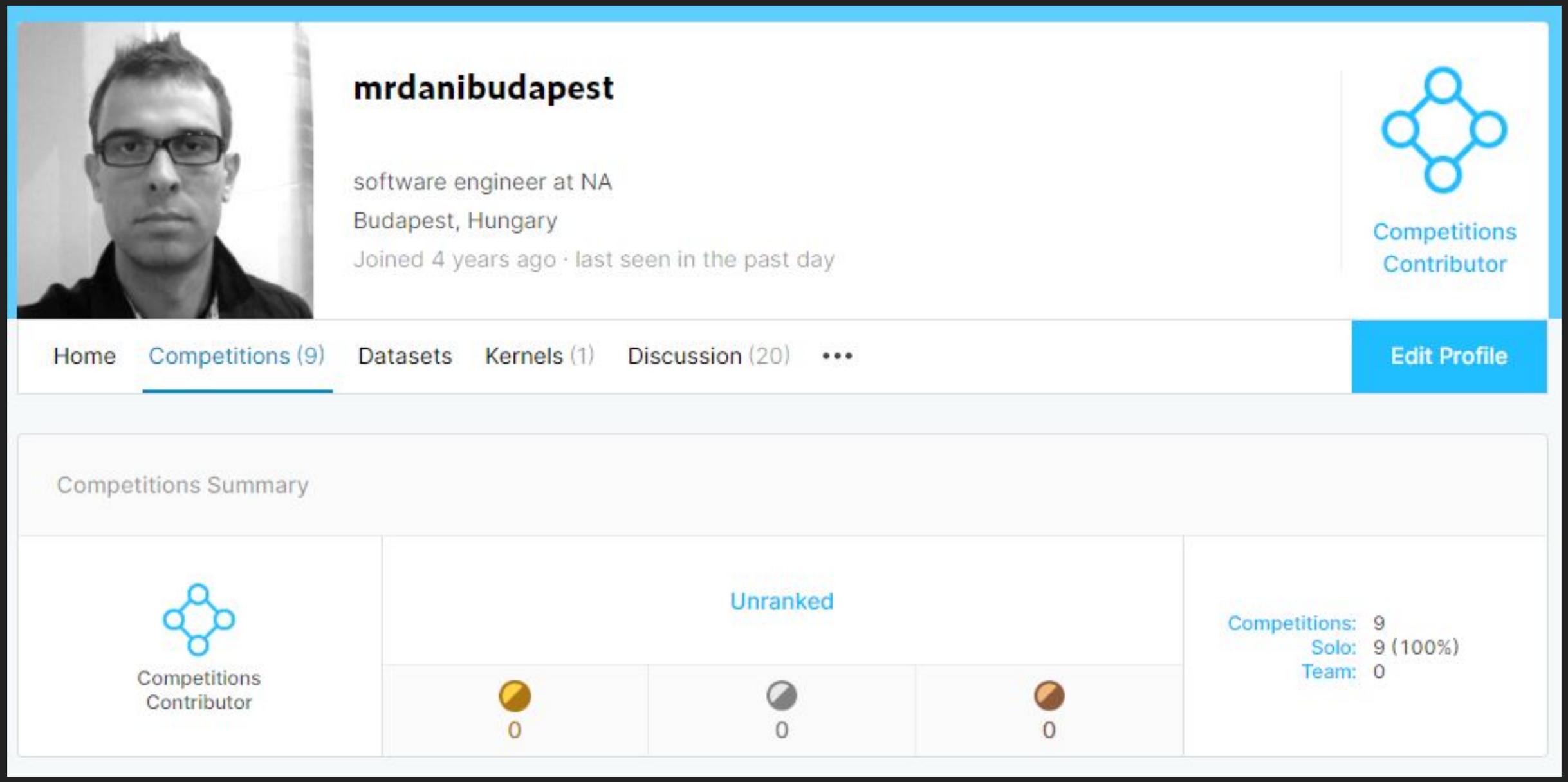


INTRODUCTION TO AI AND MACHINE LEARNING

SESSION #1

WHY ME???


mrdanibudapest
software engineer at NA
Budapest, Hungary
Joined 4 years ago · last seen in the past day
Competitions Contributor
Home
Competitions (9)
Datasets
Kernels (1)
Discussion (20)
•••
Edit Profile
Competitions Summary
Competitions Contributor
Unranked
0
0
0
Competitions: 9
Solo: 9 (100%)
Team: 0

WHY ME???

9 Completed Competitions

Elo Merchant Category Recommendation Help understand customer loyalty Featured · a year ago · banking, tabular data, regression	952/4127	Top 24%
TalkingData Mobile User Demographics Get to know millions of mobile device users Featured · 3 years ago · demographics, mobile web, tabular data, multiclass classification	1524/1688	Top 91%
Avito Demand Prediction Challenge Predict demand for an online classified ad Featured · 2 years ago · tabular data, image data, text data	1677/1871	Top 90%
Toxic Comment Classification Challenge Identify and classify toxic online comments Featured · 2 years ago · arguments, text data	3207/4550	Top 71%
TalkingData AdTracking Fraud Detection Challenge Can you detect fraudulent click traffic for mobile app ads? Featured · 2 years ago	3484/3946	Top 89%
Home Credit Default Risk Can you predict how capable each applicant is of repaying a loan? Featured · a year ago · home, banking, tabular data	3711/7190	Top 52%
IEEE-CIS Fraud Detection Can you detect fraud from customer transactions? Research · 4 months ago · tabular data, binary classification	4267/6381	Top 67%
Nagyházi feladat Adatelemzési platformok és Customer Analytics 2019 InClass · 7 months ago	49/58	Top 85%
DonorsChoose.org Application Screening Predict whether teachers' project proposals are accepted Playground · 2 years ago · crowdfunding, binary classification	223/580	Top 39%

COURSE AGENDA

AFTER THIS COMPLETING THIS COURSE YOU WILL:

- know the basic theory behind machine learning
- know the essential machine learning techniques and libraries
- get some hands-on machine learning programming practice in Python
- be able to decide on machine learning applicability to a given problem.

COURSE AGENDA

AFTER THIS COMPLETING THIS COURSE YOU WILL NOT BE:

- a machine learning expert
- a Python programmer
- offered a job as a machine learning engineer at the Firm.

COURSE AGENDA

Session #1: Introduction to machine learning, concepts, basics, capabilities. Classification basics.

Session #2: Feature engineering, data wrangling. Regression basics.

Session #3: Working with textual data, text classification, NLP basics

Session #4: Introduction to neural networks, deep learning, image recognition

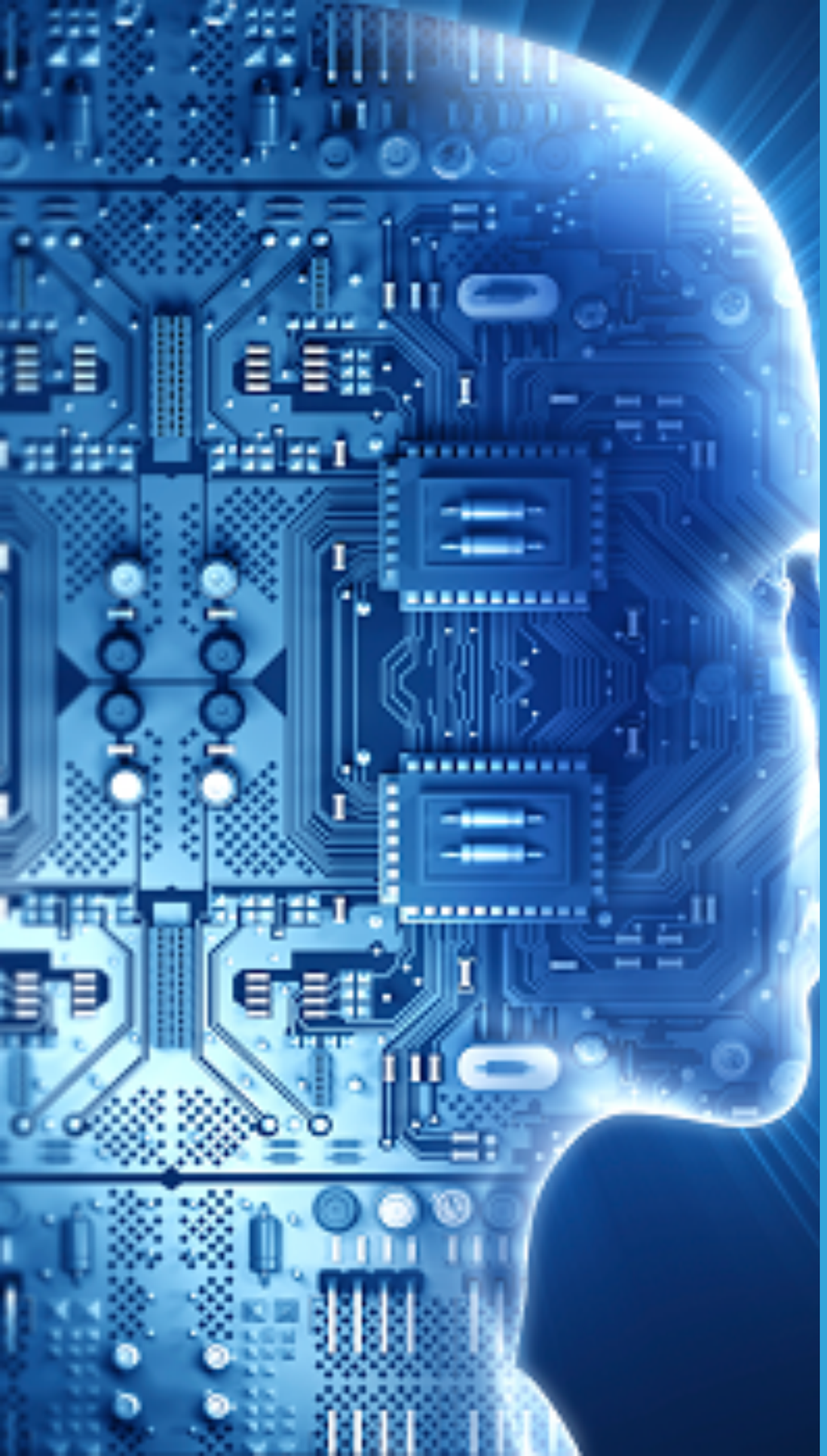
SESSION #1 AGENDA

SECTION 1

- What is machine learning?
- Essential machine learning problems & application areas
- Machine learning techniques & algorithms overview

SECTION 2

- Setting up a Python ML development environment
- Case Study: The survivals of the Titanic



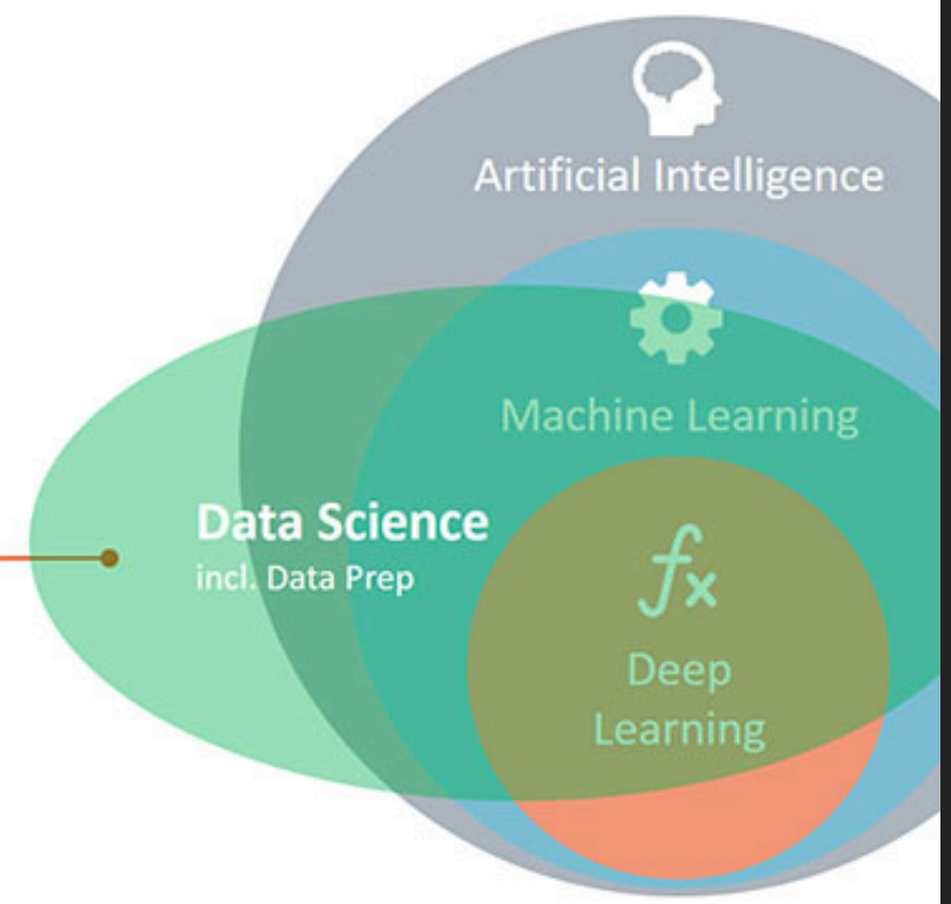
MACHINE LEARNING IS A FIELD OF STUDY THAT GIVES COMPUTERS THE ABILITY TO LEARN WITHOUT BEING EXPLICITLY PROGRAMMED

Arthur Samuel, 1959

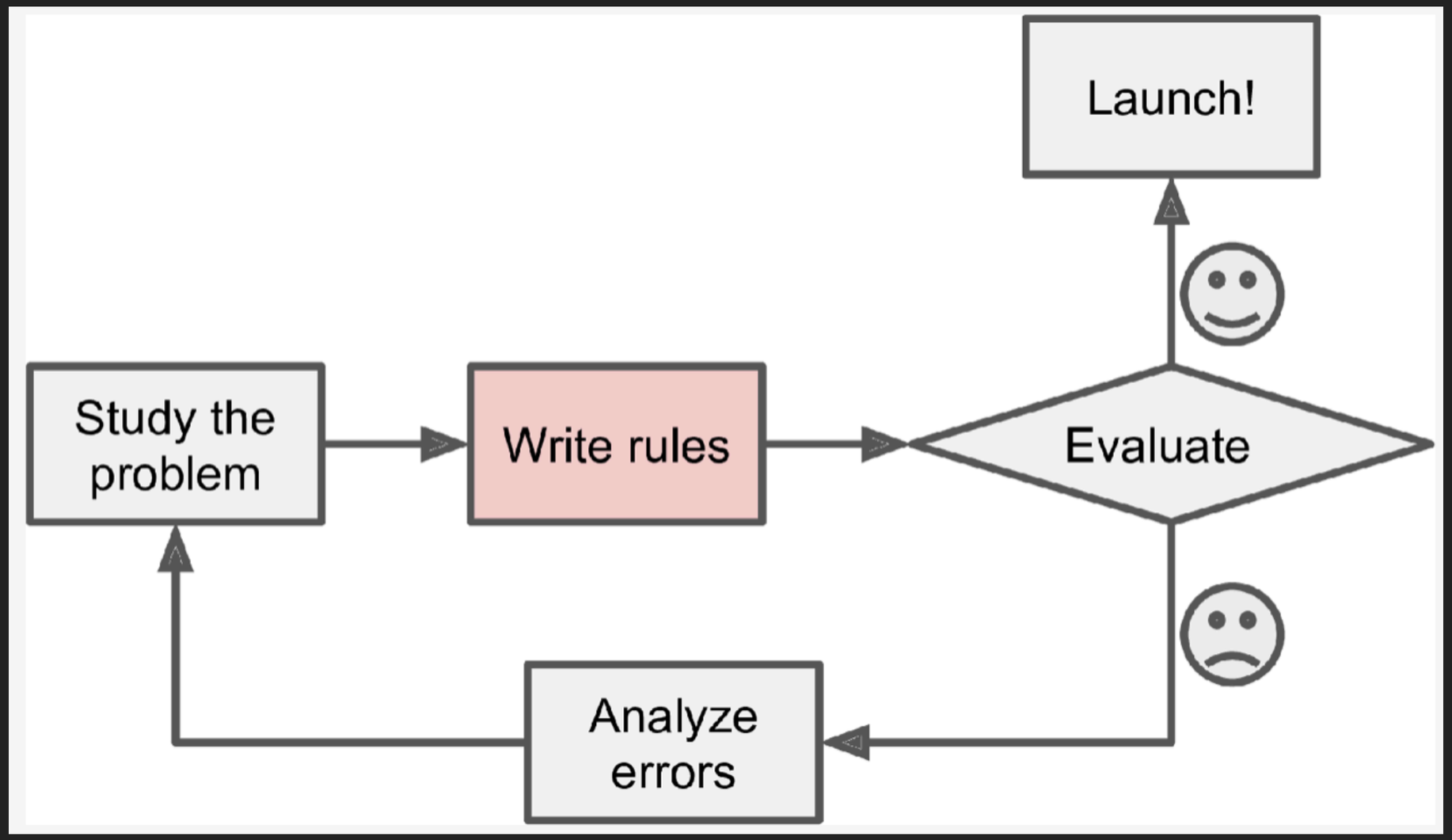
MACHINE LEARNING VS. AI. VS. DATA SCIENCE

Data Science

Covers the practical application of advanced analytics, statistics, machine learning, and the necessary data preparation in a business context.

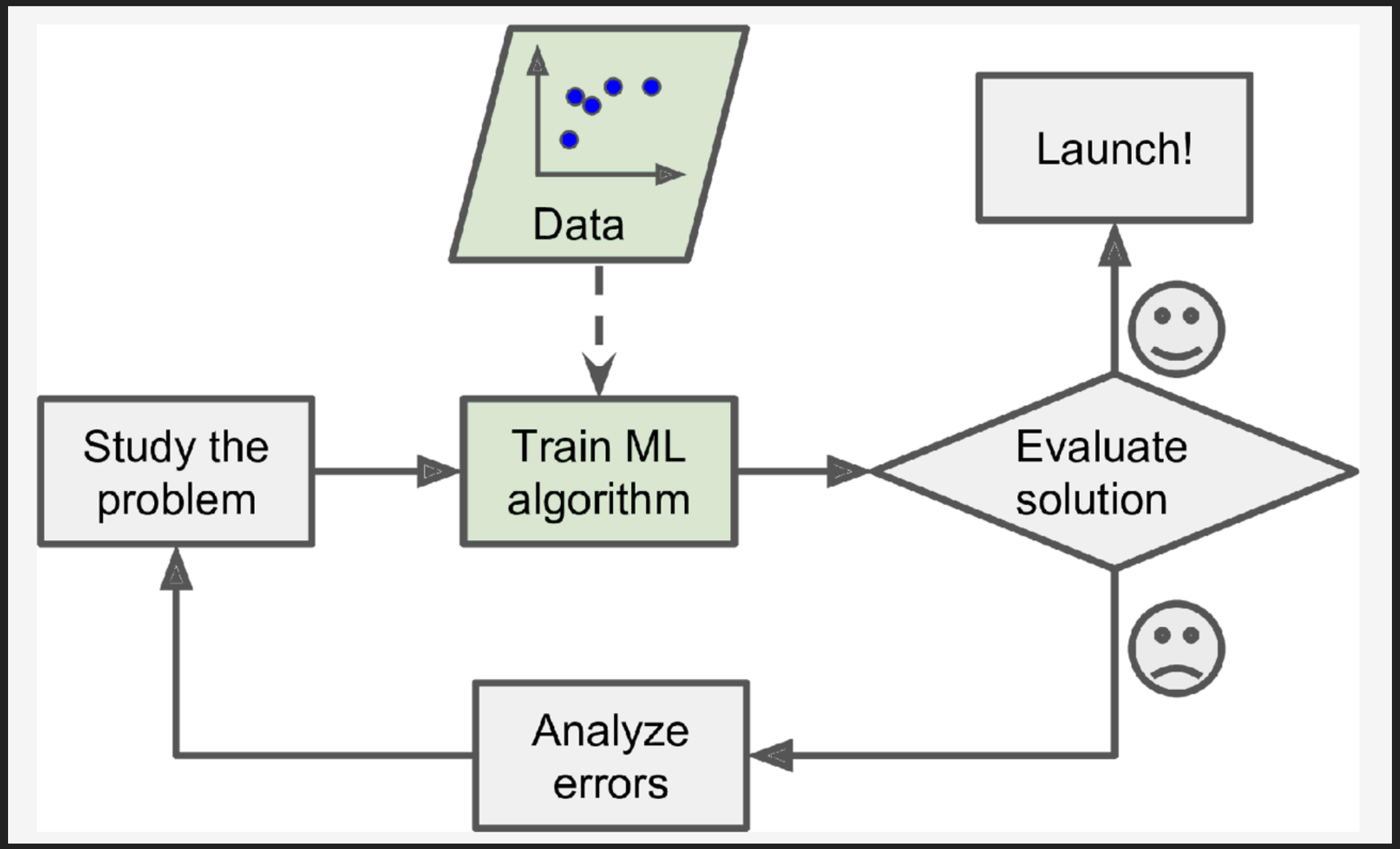


MACHINE LEARNING VS. TRADITIONAL PROGRAMMING



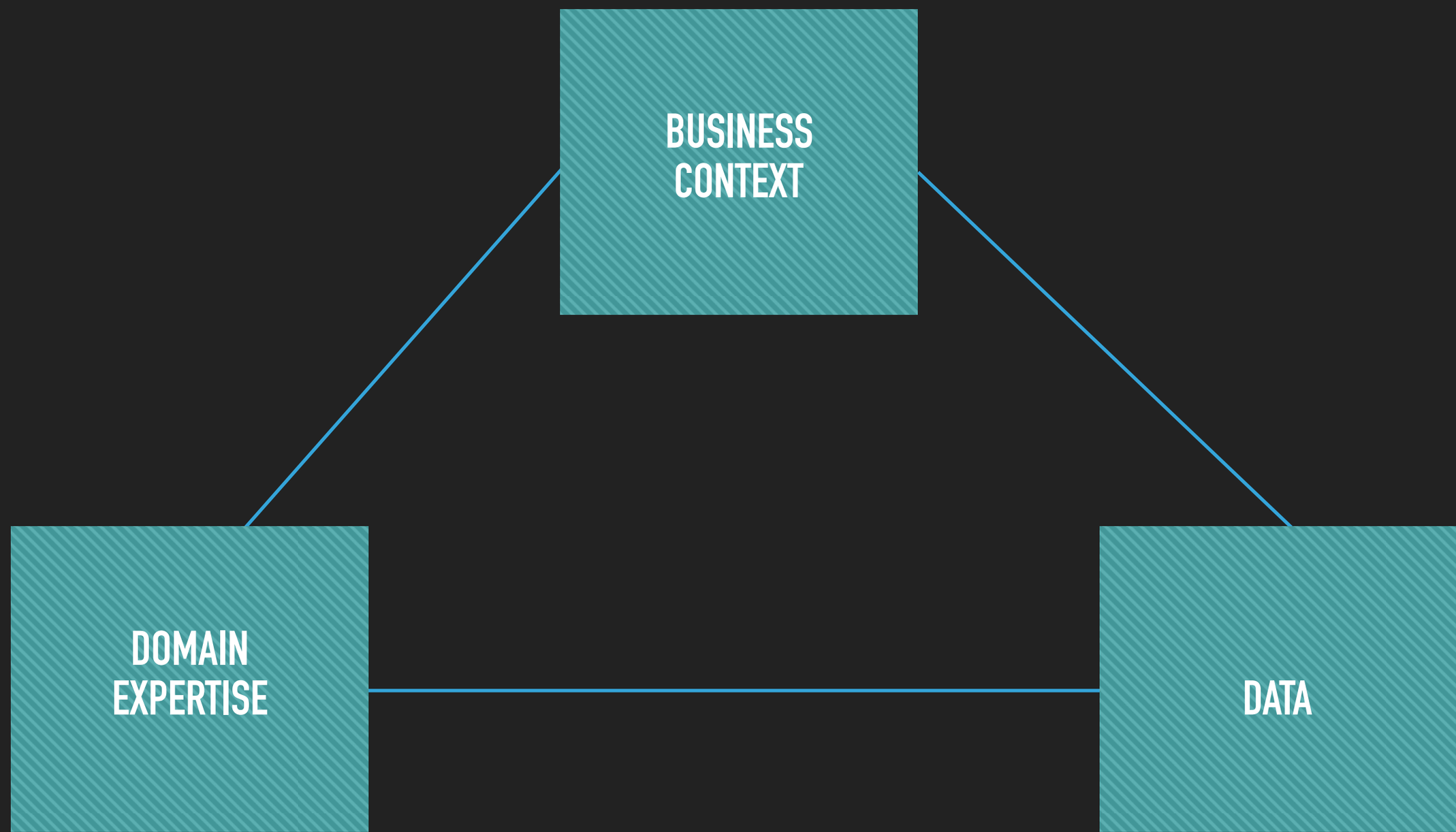
Source: Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow (Géron)

MACHINE LEARNING VS. TRADITIONAL PROGRAMMING



Source: Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow (Géron)

ENTERPRISE MACHINE LEARNING PROJECT ESSENTIAL BUILDING BLOCKS



ENTERPRISE MACHINE LEARNING PROJECT ESSENTIAL BUILDING BLOCKS

A clothes shop

Data: pictures of people entering the shop

Business context: optimise stock based on customer gender ratio

Domain expertise: how to actually optimise stock???

ENTERPRISE MACHINE LEARNING PROJECT ESSENTIAL BUILDING BLOCKS

A bank

Data: user signatures on documents

Domain expertise: ability to determine whether the client is left or right-handed

Business context: how can you make money out of it ???

ENTERPRISE MACHINE LEARNING PROJECT ESSENTIAL BUILDING BLOCKS

A printing company

Domain expertise: predict machine failures

Business context: save money by predictive maintenance

Data: ????

ESSENTIAL MACHINE LEARNING PROBLEMS

- categorisation
- numeric estimation
- forecasting
- data transformation.

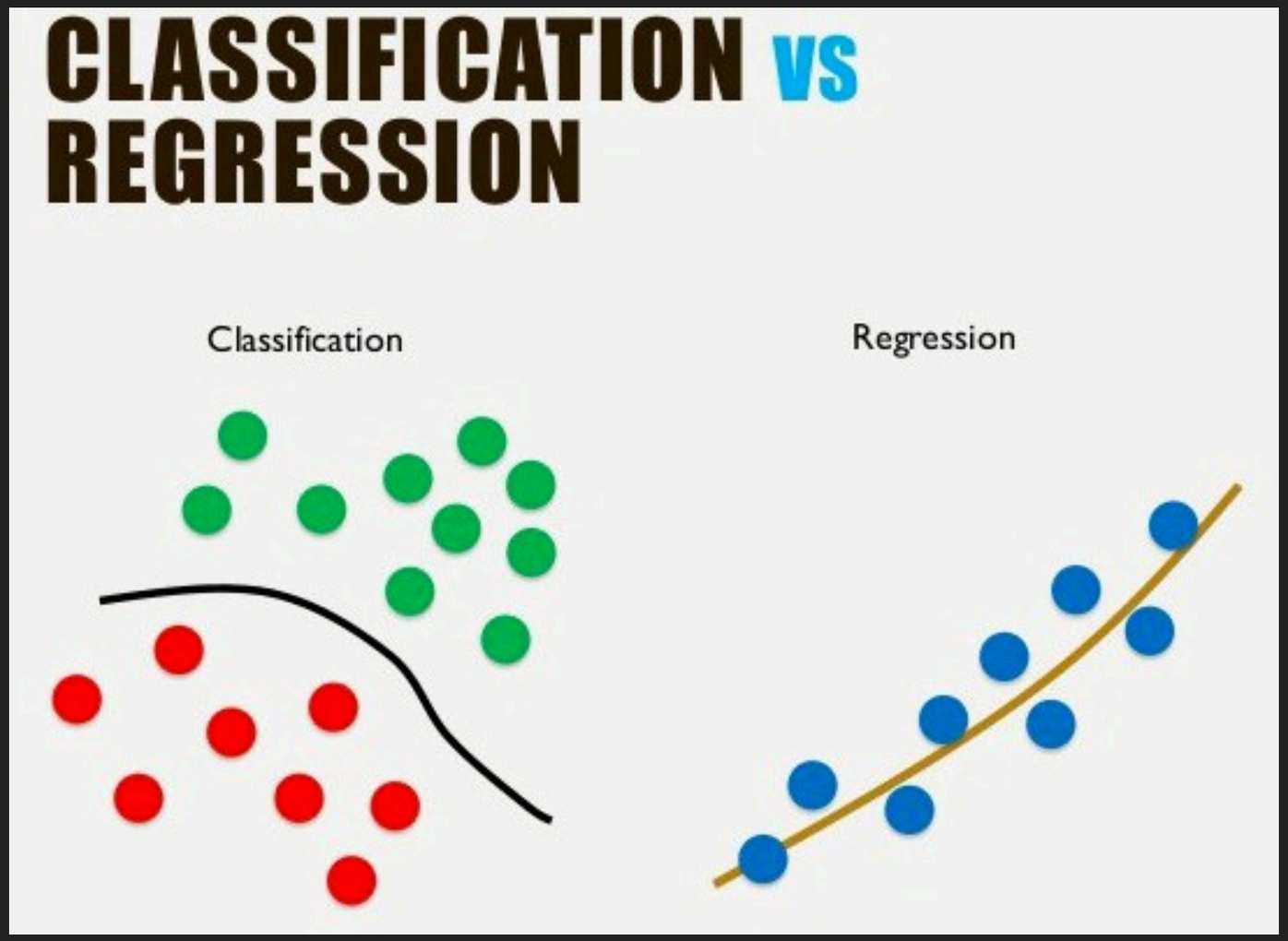
MACHINE LEARNING APPLICATION AREAS

- natural language processing
- image recognition
- signal recognition (e.g. voice, music)
- recommender systems
- anomaly detection.

HOW MACHINES LEARN?

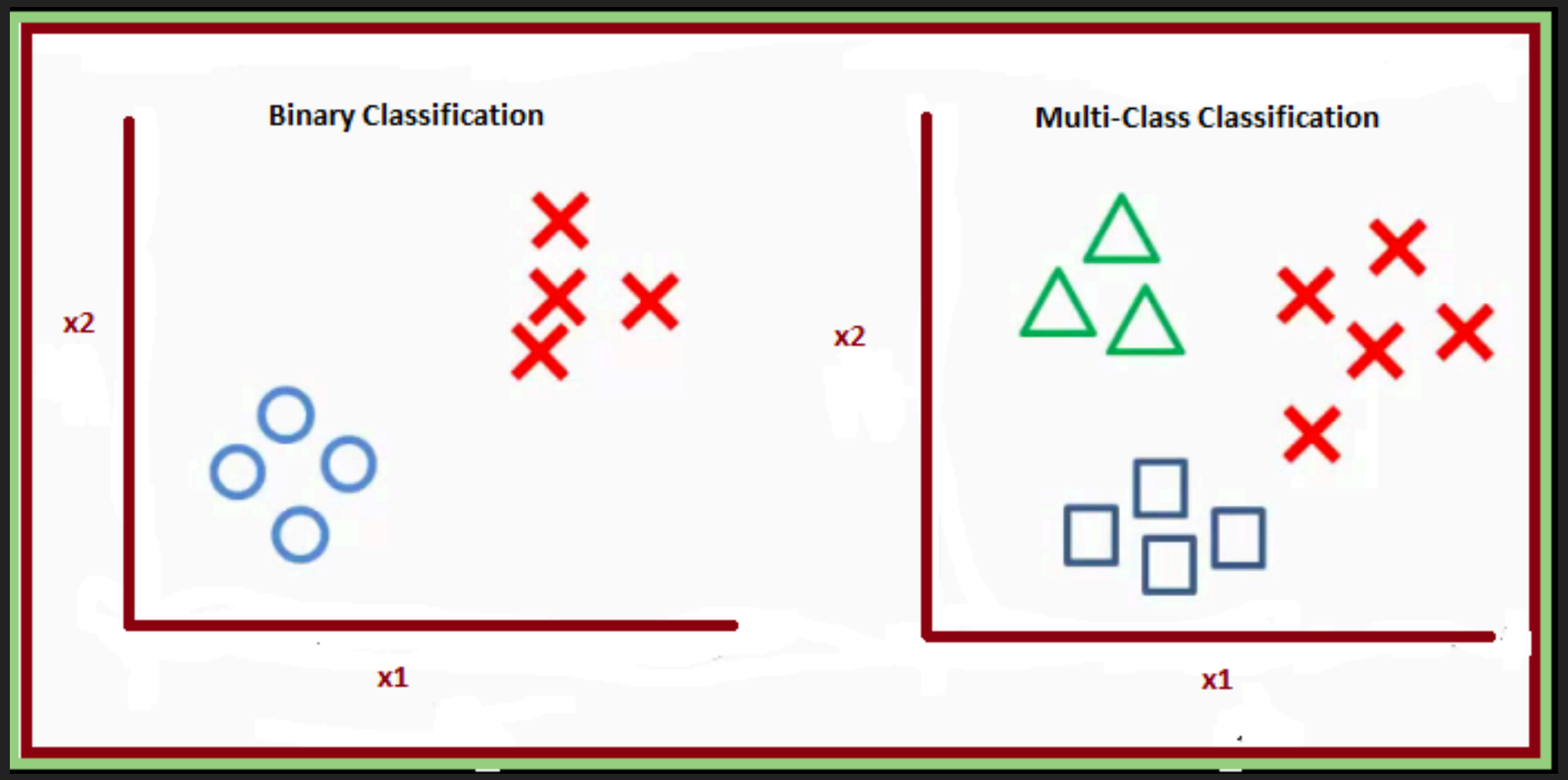
- Supervised learning
  - human labelled data, eg. spam filter
- Unsupervised learning
  - no labelled data, eg. segment customers
- Semi-supervised learning
  - combination of the two, eg. Google Photos

SUPERVISED LEARNING



Source: https://www.codeingschool.com/2019/06/regression-classification-supervised-machine-learning.html

CLASSIFICATION



Source: https://www.geeksforgeeks.org/getting-started-with-classification/binary-vs-multi-class-classification/

CLASSIFICATION

Result of a classification can be:

- true positive
- true negative
- false positive
- false negative

CLASSIFICATION

Result of a classification can be:

- true positive
- true negative
- false positive
- false negative

According to our model the patient has cancer

CLASSIFICATION

Model Evaluation

- **Accuracy**: Percentage of correct predictions made by the model.
- **Precision**: tp / (tp + fp) a.k.a positive predictive value
- **Recall**: tp / (tp + fn) a.k.a sensitivity
- **F1 score**: 2 * (precision * recall) / (precision + recall)

Ideal model: high precision, high recall

High recall, low precision: few fn, lot of tp, lot of fp

Low recall, high precision: few fp, few tp, lot of fn

CLASSIFICATION

Confusion matrices

	Actual Cancer = Yes	Actual Cancer = No
Predicted Cancer = Yes	True Positive 57	False Positive 14
Predicted Cancer = No	False Negative 23	True Negative 171

	Actual Dog	Actual Cat	Actual Rabbit
Classified Dog	23	12	7
Classified Cat	11	29	13
Classified Rabbit	4	10	24

Source: https://machinelearningknowledge.ai/confusion-matrix-and-performance-metrics-machine-learning/

BASIC STATISTICS FOR MACHINE LEARNING

You have the following data set: **7, 11, 11, 15, 20, 20, 37**

Find the following properties for the data set:

- mean
- mode
- median
- variance
- standard deviation

BASIC STATISTICS FOR MACHINE LEARNING

You have the following data set: **7, 11, 11, 15, 20, 20, 37**

Find the following properties for the data set:

- mean = (7 + 11 + 11 + 15 + 20 + 20 + 37) / 7 = 121 / 7 = 17.28
- mode = highest frequency element: 11 and 20
- median = the middle element in numerical order: 15
- variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$, standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$
- both shows how far the data is from the mean

CODE DEMO

RECAP

Today we learnt:

- what are the essential machine learning problems and their application in a business context
- supervised learning -> classification -> binary classification
- the evaluation of classification models
- the basic steps in Python to build a basic machine learning model

HOMEWORK

THE PIMA INDIAN DIABETES DATASET

Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?