Machine Learning and Deep Learning
Lecture-01

By: Somnath Mazumdar
Assistant Professor
sma.digi@cbs.dk

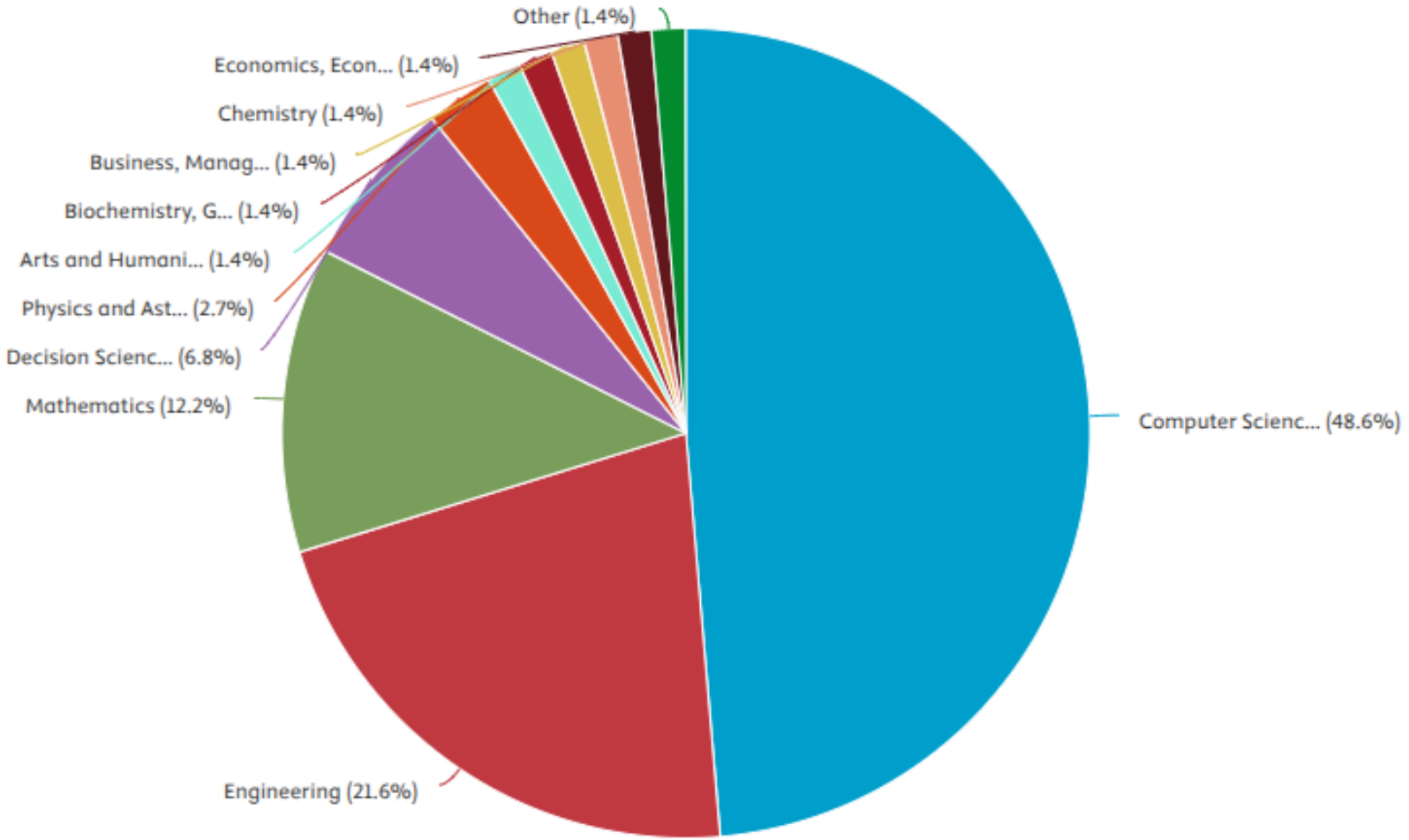

WELCOME

This session's primary AIM is to set **expectations**,
rules and other practicalities.

Overview

- About Me & You
- About Course
  - Learning Objectives
  - N9
  - Prerequisites
- Lecture Plan
- Exam & Guidelines
- About Ucloud

My Research


Other (1.4%)
Economics, Econ... (1.4%)
Chemistry (1.4%)
Business, Manag... (1.4%)
Biochemistry, G... (1.4%)
Arts and Humani... (1.4%)
Physics and Ast... (2.7%)
Decision Scienc... (6.8%)
Mathematics (12.2%)
Engineering (21.6%)
Computer Scienc... (48.6%)

About You

About THIS Course

Team

- Total: 157 Students..
- Teaching Assistants:
  - Alexandru Plecan --> alp.digi@cbs.dk
  - Yue Wu --> yw.digi@cbs.dk
- Teacher: Somnath Mazumdar (sma.digi@cbs.dk) (SM)
  - Student hour: Every Tuesday 16:00 hrs to 18:00 hrs (Email me before)

Four Basic Goals

- What is ML/DL?
- What are its components/models/complexities/issues?
- How can we pick a fitting model for an application?
- How can we bring ML advantages to the application?

Learning Objectives

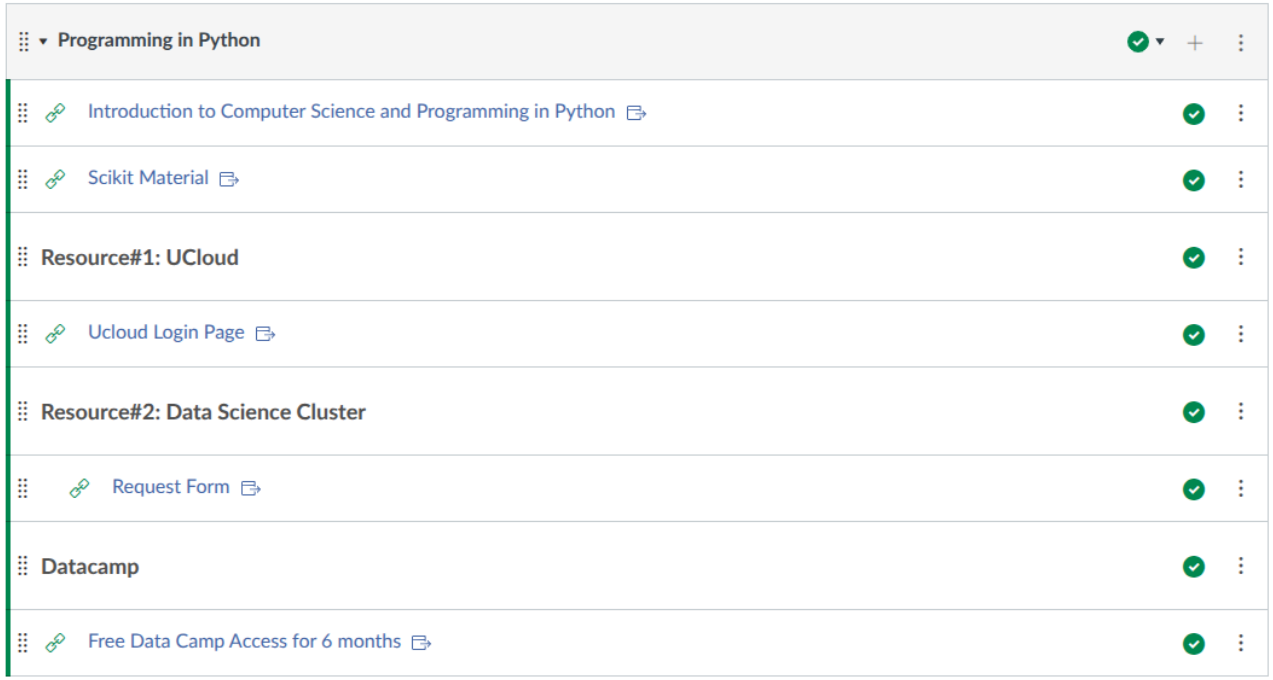
1. Understand fundamental challenges of machine learning (ML) models (selection, complexity).
2. Detect strengths and weaknesses of ML models.
3. Design, implement, ML models and deep learning techniques for realistic applications.
4. Summarize application areas, trends, and challenges in ML.
5. Exhibit deeper knowledge and understanding of covered topics.
6. Reflect on critical awareness of methodological choices with written skills to accepted academic standards.

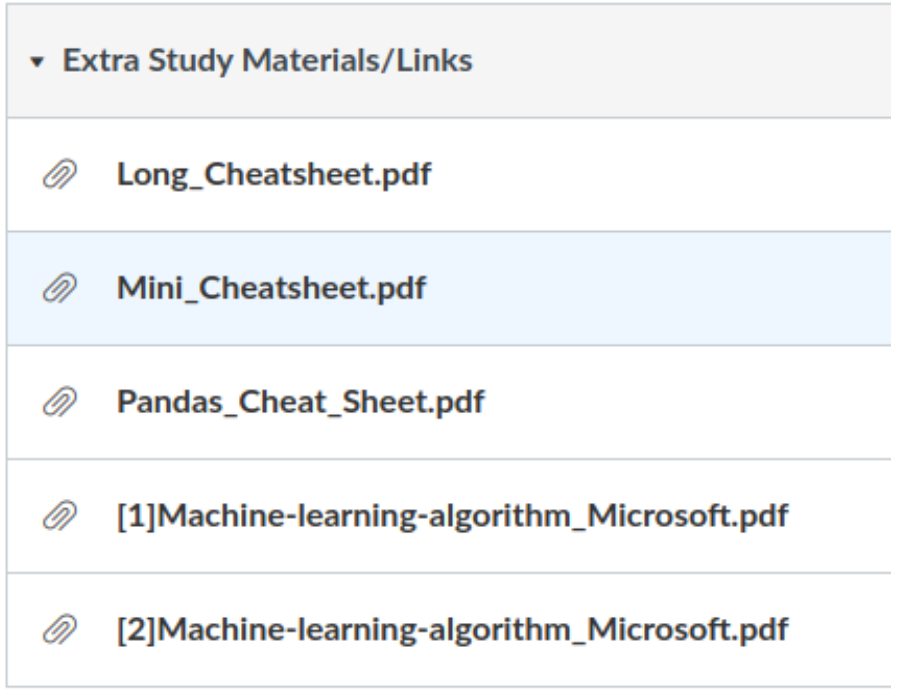
NORDIC NINE

Copenhagen Business School develops disciplinary skills and transformational capabilities. Together we pursue knowledge that builds values, and values that prepare for action.
Graduating from CBS means that

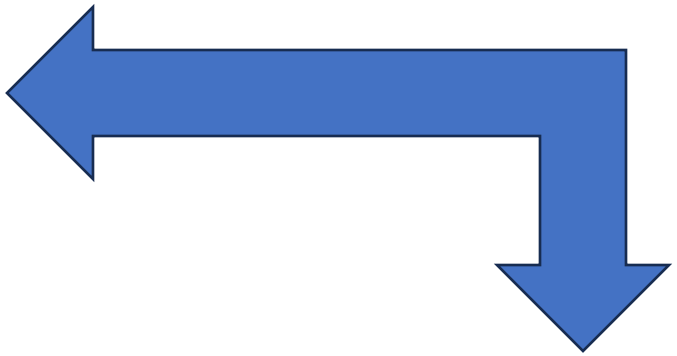
KNOWLEDGE	VALUES	ACTION
you have deep business knowledge placed in a broad context	you are competitive in business and compassionate in society	you produce prosperity and protect the prosperity of next generations
you are analytical with data and curious about ambiguity	you understand ethical dilemmas and have the leadership values to overcome them	you grow by relearning and by teaching others to do the same
you recognise humanity's challenges and have the entrepreneurial knowledge to help resolve them	you are critical when thinking and constructive when collaborating	you create value from global connections for local communities

Helping Materials





Lecture Plan

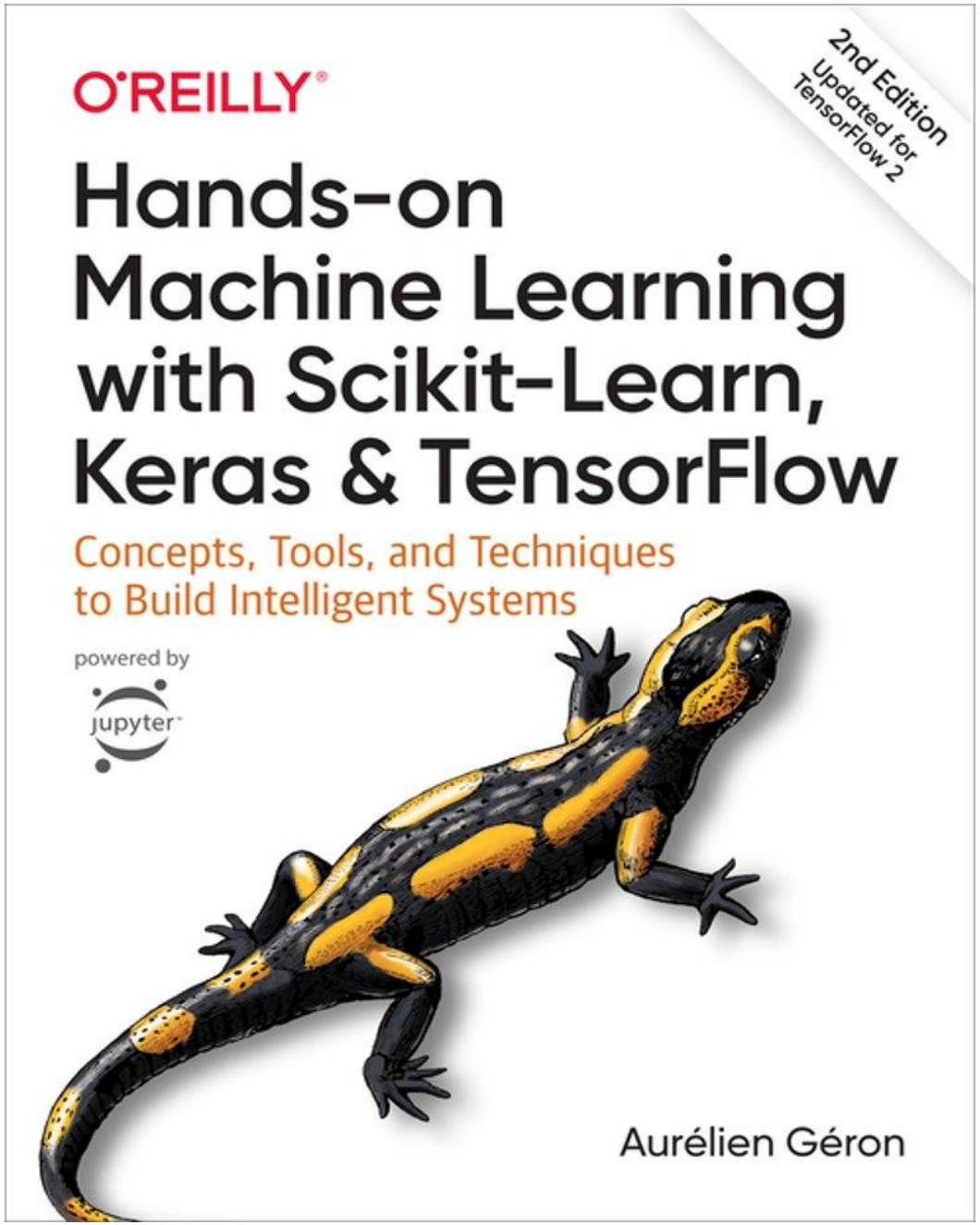


Lecture	Teacher	Topic(s)	Readings
Lecture-01 Week #06	SM	Course practicalities & Introduction to course	[HML] Ch 1
Lecture-02 Week #06	SM	Introduction to machine learning; Data pre-processing and exploratory data analysis	[HML] Ch 2
Lecture-03 Week #08	SM	Principles of unsupervised machine learning; K-Means, DBSCAN, Hierarchical clustering	[HML] Ch 9
Lecture-04 Week #08	SM	Principles of supervised machine learning; KNN, Linear and Logistic regression	[HML] Ch 3 and 4
Lecture-05 Week #09	SM	Dimensionality Reduction: Principal Component Analysis, Decision Trees, Ensembles: Random Forests	[HML] Ch 8
Lecture-06 Week #09	SM	Boosting: Gradient Boosting, Support Vector Machines (SVM); Performance Metrices of ML	SVM: [HML] ch, 5; Decision trees: [HML] ch. 6; Ensembles and random forest: [HML] ch. 7
Lecture-07 Week #10	SM	Outlier Detection: Isolation Forest, Recommender Systems, Class imbalance: SMOTE and ADASYN	SMOTE: [J01] section 4.2 and ADASYN: [J02]
Lecture-08 Week #10	SM	Gradient Descent problem, Batch and mini-batch gradient descent, Regularization (L0, L1, early stopping and dropout regularization), Batch normalization.	[HML] Ch 4
Lecture-09 Week #11	SM	Neural Networks: General Introduction, Threshold Logic Units (TLU), Multi-Layer Perceptron with Many Layers, Feedforward network	[HML]: Ch 10
Lecture-10 Week #11	SM	Introduction Tensor flow, RNN, Distributed DL	[https://www.tensorflow.org/tutorials/] & [HML]: Ch 10
Lecture-11 Week #12	SM	Long short-term memory (LSTM), Gated Recurrent Unit.	[HML]: Ch 15
Lecture-12 Week #12	SM	Autoencoder, Hyper Parameter Optimizations	
Lecture-13 Week #13	SM	CNN, Adversarial attacks	[HML]: Ch 14
Lecture-14 Week #13	SM	Philosophy of AI, Ethics of ML and DL, AI alignment	
Lecture-15 Week #15	SM	Federated learning, Explainable artificial intelligence, Reinforcement learning	
Lecture-16 Week #15	SM	Recap and conclusion; Project Consultations; Groups meetings	

Course is too vast!!

It covers many models.

Book


O'REILLY
2nd Edition
Updated for
TensorFlow 2
Hands-on
Machine Learning
with Scikit-Learn,
Keras & TensorFlow
Concepts, Tools, and Techniques
to Build Intelligent Systems
powered by
jupyter
Aurélien Géron

Exam & Guidelines

Compulsory Assignments

- Activity Type: Mandatory Activities
- Number of mandatory activities: 3
- Activities to be approved to qualify for final exam: 2

Compulsory Assignment 01 (Opens: 14-02-2025 & Closes: 28-02-2025)

Compulsory Assignment 02 (Opens: 28-02-2025 & Closes: 14-03-2025)

Compulsory Assignment 03 (Opens: 18-03-2025 & Closes: 01-04-2025)

Submission Process

- Teacher adds the question paper by creating "CANVAS Assignment".
- Students must add their student numbers on the mandatory home assignment (on Canvas) to make it easier for the teacher to find them on Digital Exam.
- Students hand in their mandatory assignment on CANVAS.
- Teacher checks the assignments on CANVAS.
- Teacher grades them on Digital Exam.

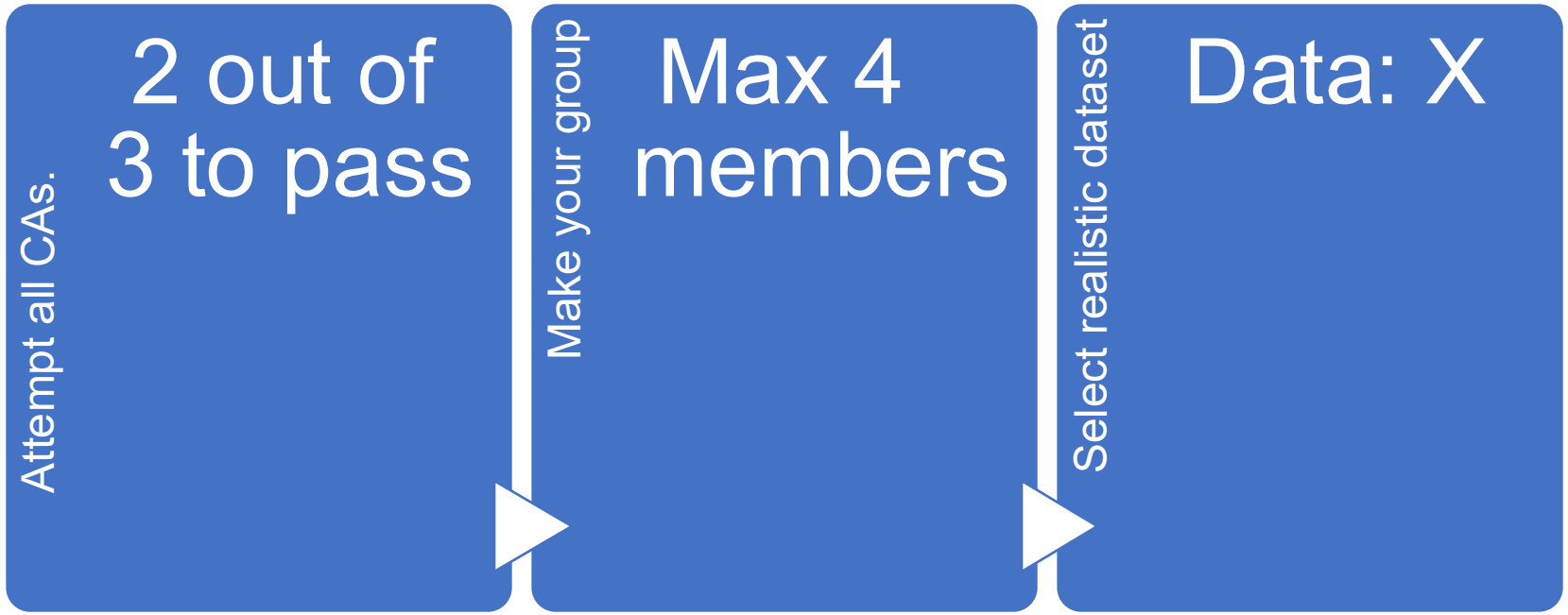
Final Exam

- Assignment type: Project
- Form: Individual oral exam based on written group product.
  - Model Implementation is important!!
- Number of people in group: 2-4.
- Size of written product: Max. 15 pages.
- Examiner(s): Examiner and internal examiner.

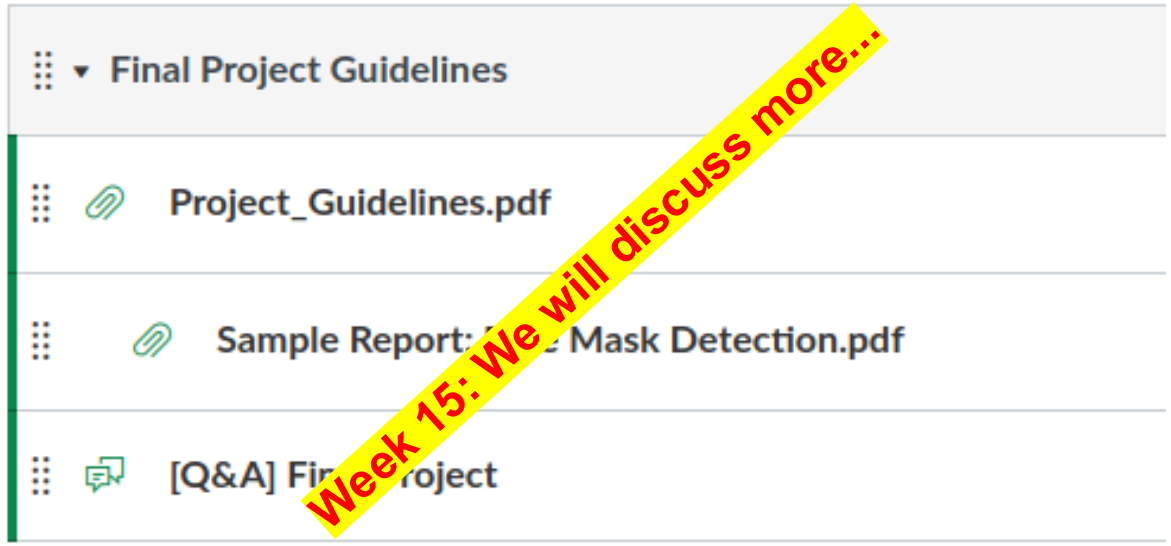
Final Exam Guidelines

- Choose a realistic data set.
- Analyse data set applying covered models.
- Discuss the rationale behind the usage selected model(s).
- Analyse the result for lay man use.

Before Week 15



Report Structure [Week 15]



Remember!!

Rule: Compulsory Assignments

- If a student did not get his/her compulsory assignments approved, s/he cannot participate in the final exam.
- Student should make a decent try in ALL compulsory assignments.
- We assess students' effort.

[For More]: Check CBS program regulations

Rule: Final Exam & Compulsory Assignments

- For doing group work alone:
  - Go to administration and *never through the teacher/course coordinator*.
  - Applies to both final exam and compulsory assignments.
    - A student *should talk to the teacher*/course coordinator before applying the Study Board for exemption to do group work alone.
    - When applying for exemption the student is asked to document that the student has *made an effort* to find a group.

Rule: Final Exam & Compulsory Assignments

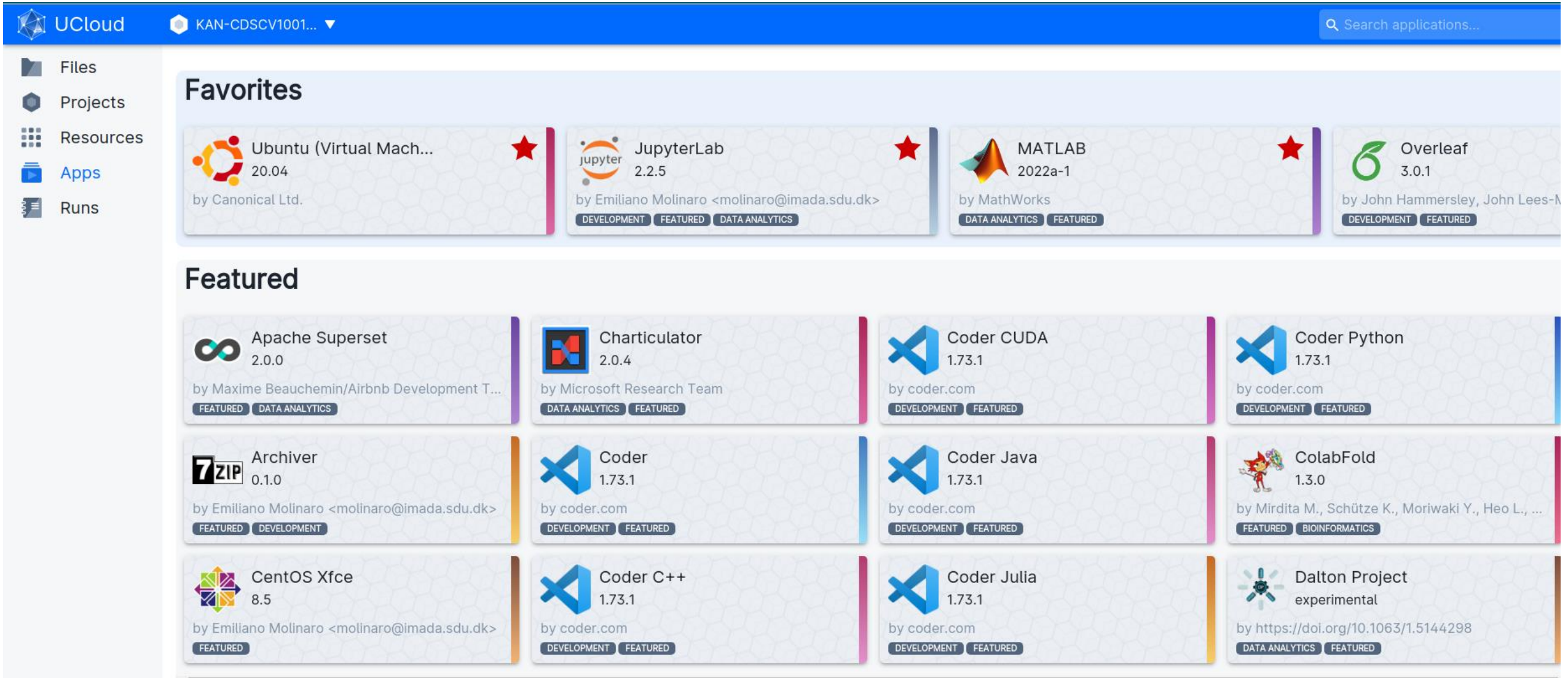
- Guideline also applies where students wishes *for a group size that differs* from what is described in the course catalogue.

A solo member-based group project is FULLY discouraged!!

Teacher *will not ask* any group to include someone!!

Students must *co-operate*!!

UCloud Platform

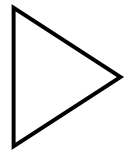


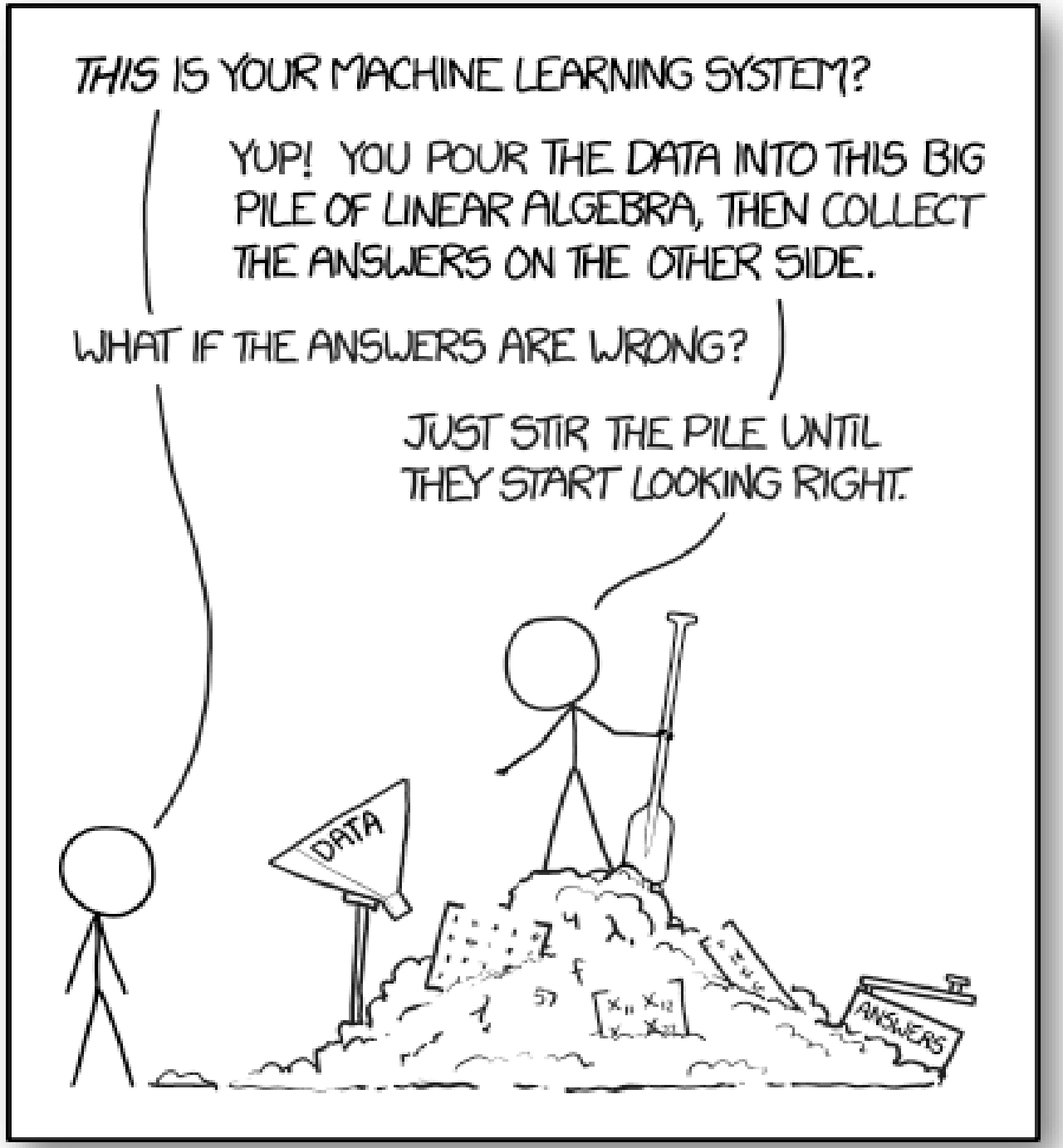
Exam/Course related Remarks/Questions/Confusions

Good-to-Know Survey

vote at **Slido.com** with
Code #838856

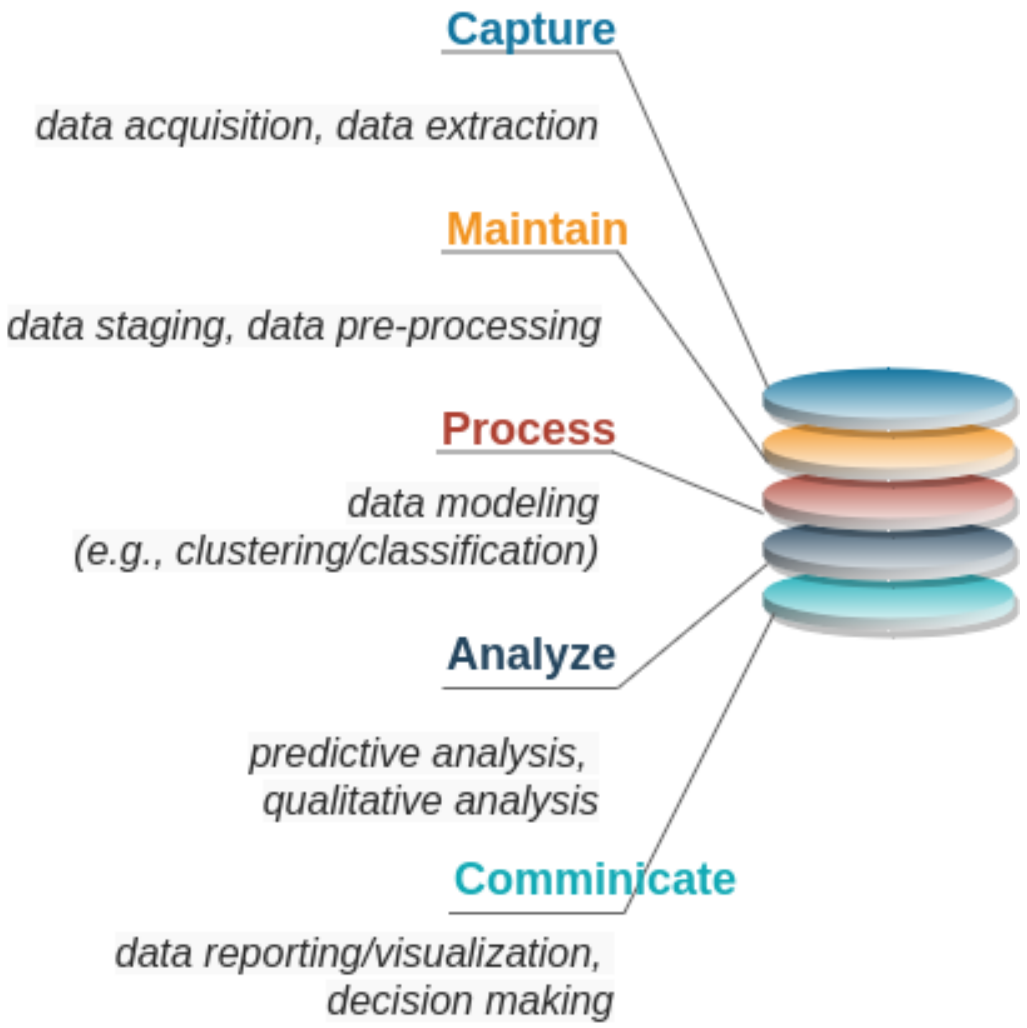
Let's Get Started





What is Data Science?

"The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades." - Hal Varian, 2009*



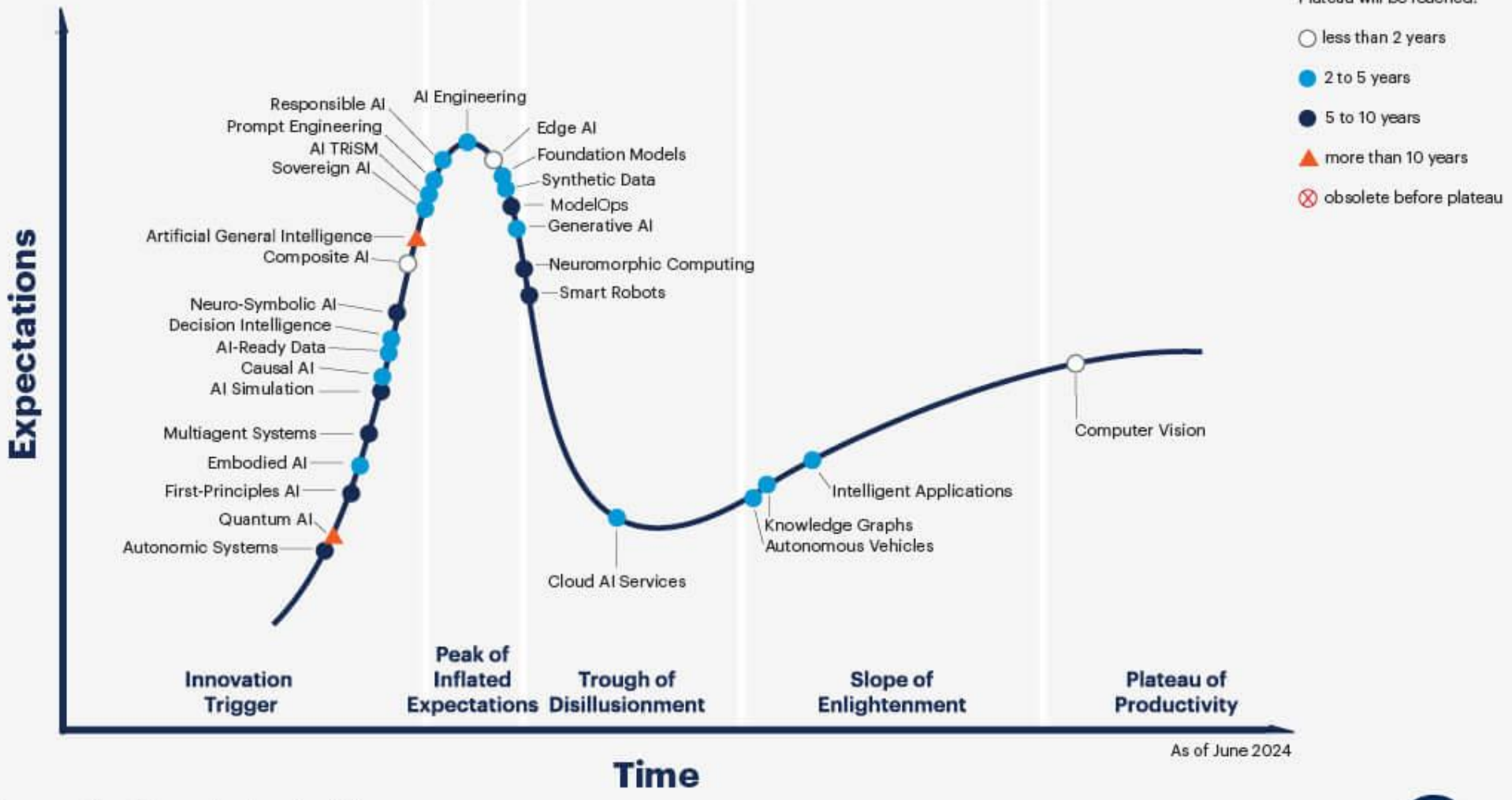
*https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers

What is Data Science?

- Data science is primarily a collection of statistical and machine learning models.
  - It supports and guides the information and knowledge extraction from data.
  - Offers insights, establishes causality, predictions.
- Application Domains: Banking, Manufacturing, supply chain management, Transportation, Healthcare and many more.

*https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers

Hype Cycle for Artificial Intelligence, 2024



Source: Gartner

Gartner®

Jun 2018, USA:
Tesla Model S running on autopilot..

Feb 2020, USA:
Tesla Model X running "semi-autonomously" of Autopilot.

https://www.wired.com/story/tesla-autopilot-why-crash-radar/
https://www.bbc.com/news/technology-51645566

MONDAY, AUGUST 5, 2024

8/5/2024 12:00:00 PM Share This Episode

Hidden Autopilot Data Shows Patterns in Tesla Crashes

A WSJ investigation reveals previously unknowable patterns in crashes involving Tesla's driver-assistant system, Autopilot. Frank Matt, a WSJ senior video journalist, joins host Zoe Thomas to explain the comprehensive analysis of crash data and the longstanding concerns about Tesla's Autopilot. Plus, why Amazon is expanding its ultrafast delivery to rural U.S. communities.

Sign up for the WSJ's free Technology newsletter.

00:27 / 12:31 1x

FULL TRANSCRIPT

This transcript was prepared by a transcription service. This version may not be in its final form and may be updated.

Zoe Thomas: Welcome to Tech News Briefing. It's Monday, August 5th. I'm Zoe Thomas for the Wall Street Journal. Amazon's last frontier of ultra-fast delivery in the US is reaching into the remote corners of America. We'll tell you why and what it could mean for the US Postal Service. And then, Tesla's semi-autonomous driving system relies mostly on cameras, which differs from the rest of the industry. A WSJ analysis has revealed previously unknowable patterns in crashes involving this system called Autopilot, and found Autopilot sometimes struggles to recognize obstacles or stay on the road. We'll bring you details of that investigation. But

https://www.youtube.com/watch?v=mPUGh0qAqWA

What is machine learning?

- Goal of AI is to create computer models that exhibit “intelligent behaviors” like humans.
- Machine Learning (ML) is one way to use AI.

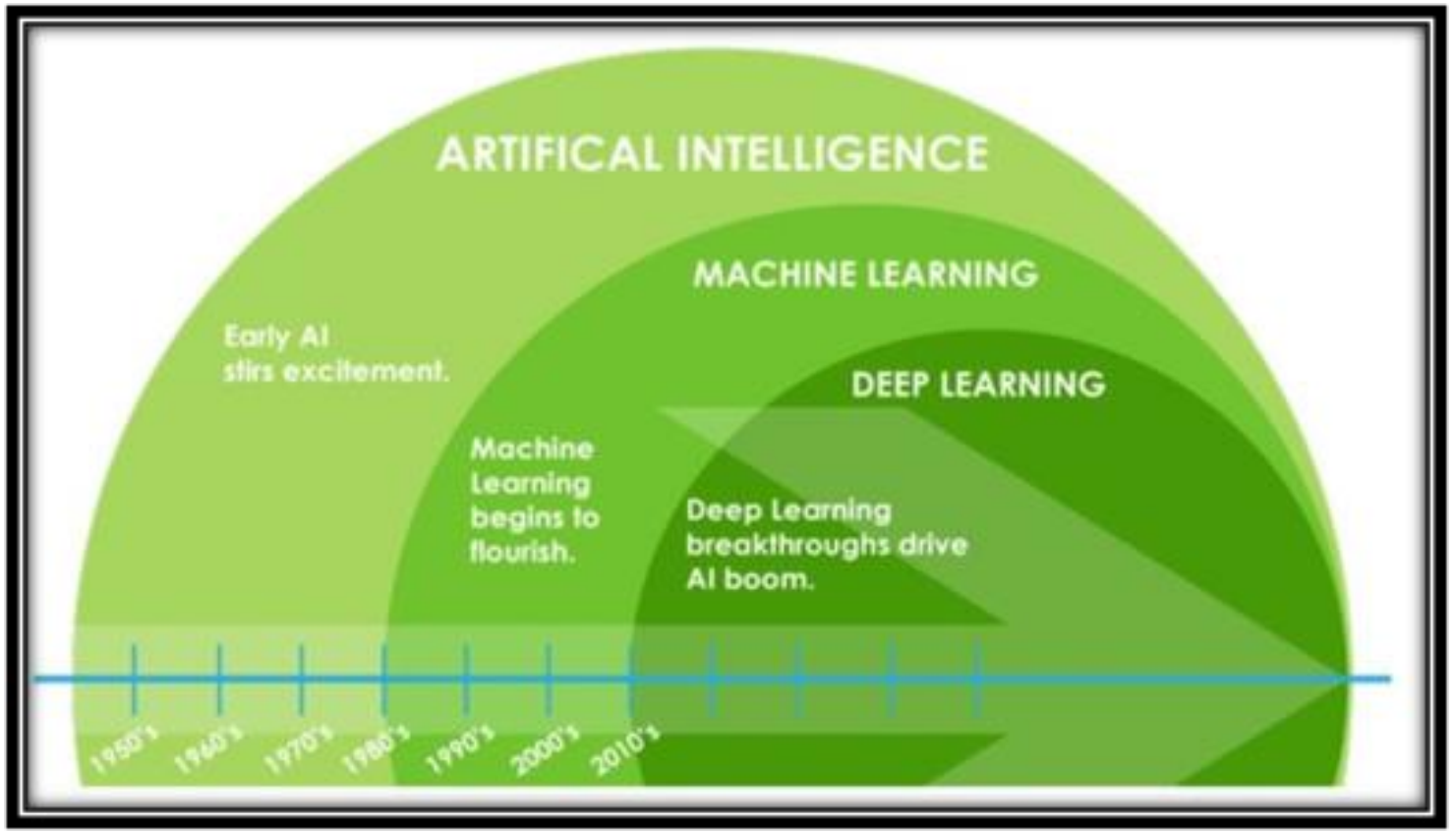


Image Src: Introduction to Deep Learning and Tensorflow By Cineca, 2018 Italy.

Data Science and Machine Learning

- Components of data science: data, (pre-) processing, statistical model and *machine learning*.
- One of the key feature of machine learning is prediction.
  - In machine learning, we train models and evaluate models.
  - Deep neural network has many hidden layers.
    - Ex- Image classification, Natural language processing
- Multiple machine learning problems overlap with statistics.

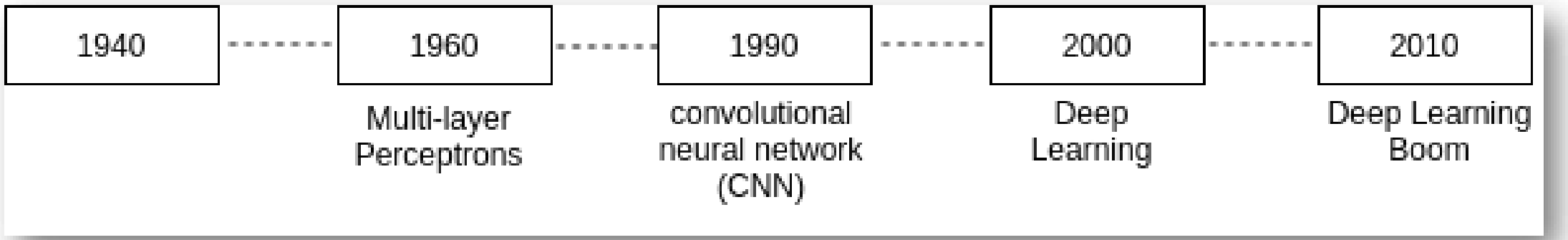
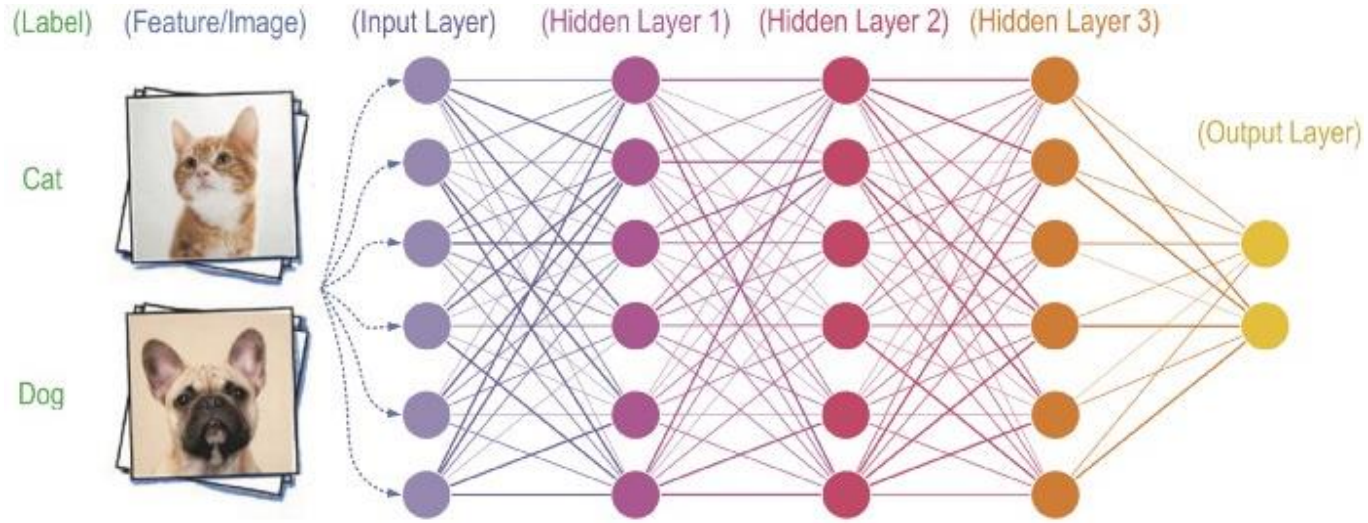


Image Src: Introduction to Deep Learning and Tensorflow By Cineca, 2018 Italy.

ML to Deep Learning

- ML was defined in 1950s as “the field of study that gives computers the ability to learn without explicitly being programmed.”
- Deep learning (DL) is a subfield of ML which uses neural networks with many layers.
    - Layered network can process extensive amounts of data and determine the “weight” of each link in the network.



Src: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Terminology

- **Algorithm:** is a set of procedures that creates a model when trained. E.g., linear regression.
- **Model:** is a fitted algorithm that has been trained. E.g. a linear regression model that has been trained to predict prices of 'X'.

- **Parameters** are the internal variables within a model that are adjusted automatically as you train a ML model
- **Hyperparameters** are set by the user to control the algorithm and they define how it learns from the data.

Example

sklearn.naive_bayes.GaussianNB

class sklearn.naive_bayes.**GaussianNB**(*, *priors=None*, *var_smoothing=1e-09*) [source]

Gaussian Naive Bayes (GaussianNB).

Can perform online updates to model parameters via `partial_fit`. For details on algorithm used to update feature means and variance online, see Stanford CS tech report STAN-CS-79-773 by Chan, Golub, and LeVeque:

http://i.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf

Read more in the User Guide.

Parameters:

priors : *array-like of shape (n_classes,), default=None*

Prior probabilities of the classes. If specified, the priors are not adjusted according to the data.

var_smoothing : *float, default=1e-9*

Portion of the largest variance of all features that is added to variances for calculation stability.

New in version 0.20.

https://scikit-learn.org/stable/tutorial/index.html

Functioning of ML

- Functions:
  - **Descriptive** meaning that system uses data to explain what happened.
  - **Predictive** meaning that system uses data to predict what will happen.
  - **Prescriptive** meaning that system will use data to make suggestions about what action to take.

Src: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Categories of ML

- Subcategories:
  - **Supervised** ML models are trained with labeled data sets.
  - **Unsupervised** ML models looks for patterns in unlabeled data.
  - **Reinforcement** ML trains machines through trial and error to take the best action by establishing a reward system.
    - Difficult to precisely specify the task.
    - Learning process of task can be dangerous.

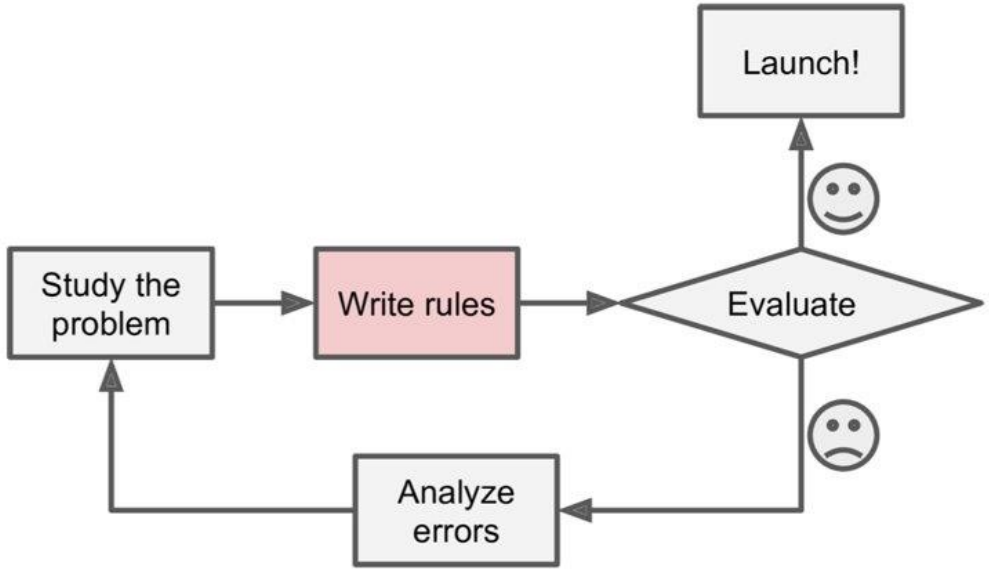
Src: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Subfields of ML

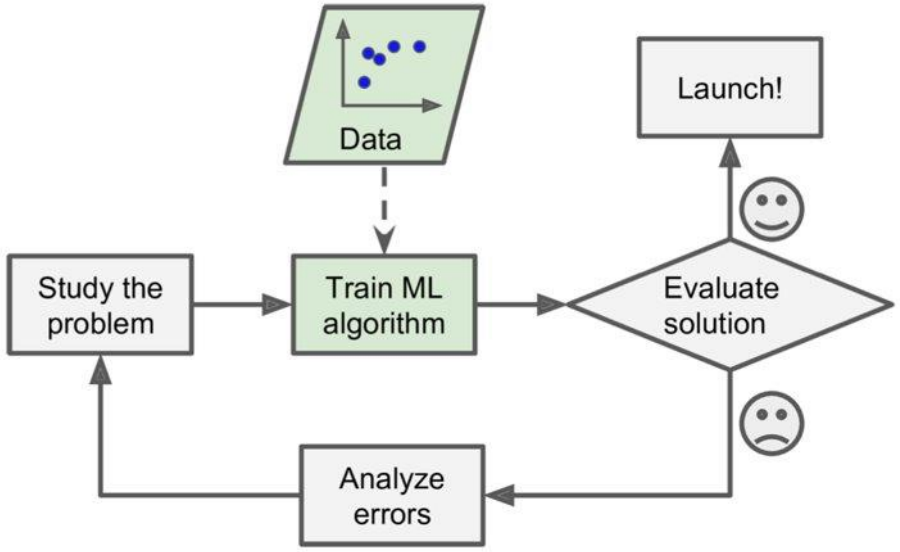
- Subfields:
    - Natural language processing is a field of ML in which machines learn to understand natural language as spoken and written by humans.
    - Neural networks are modeled on the human brain.
    - Deep learning networks are neural networks with many layers.

ML Application

- ML works well where:
  - (Large) data is available.
  - Problem is dynamic or fluctuating.
  - Problem requires predictions or discovering patterns.



Traditional programming



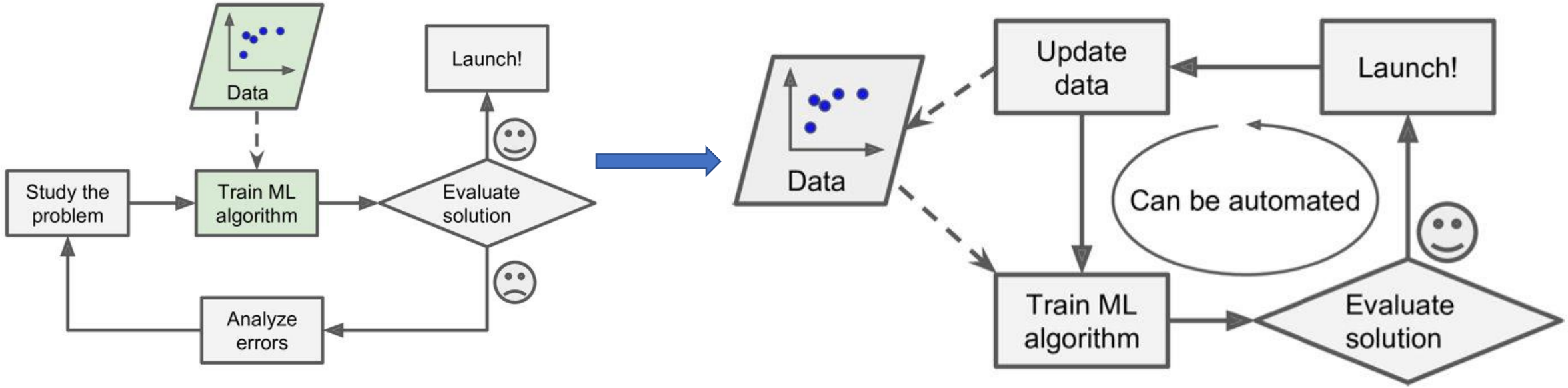
ML-based programming

Src: Image from Géron, A. (2019)

ML Application

ML works well where:

- (Large) data is available.
- Problem is dynamic or fluctuating.
- Problem requires predictions or discovering patterns.

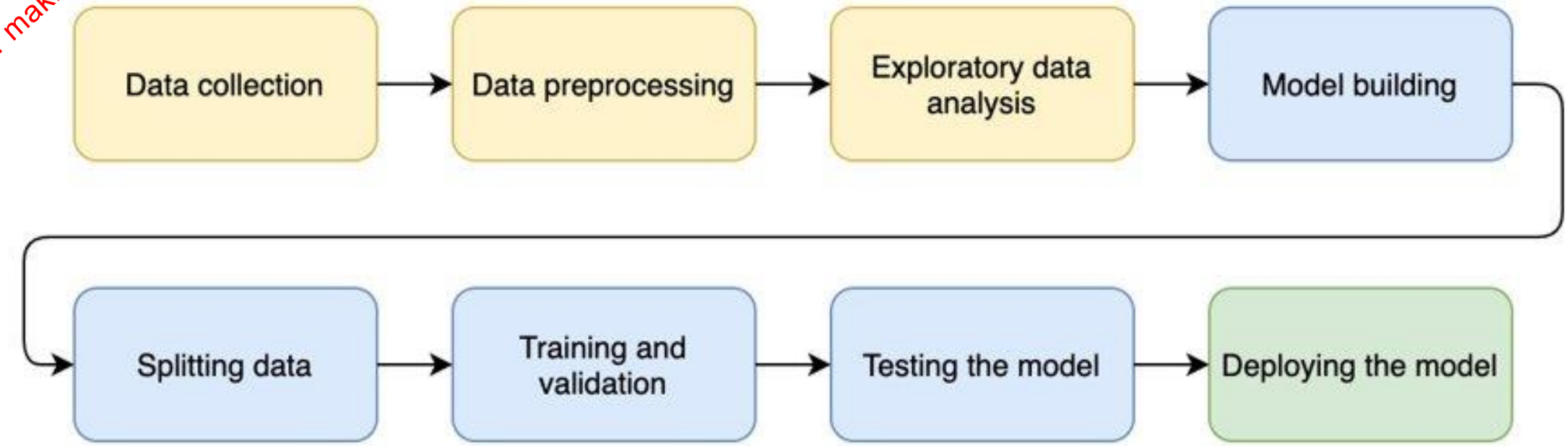


ML-based programming

Src: Image from Géron, A. (2019)

Consider making your scripts reusable!!

ML Flow



Yellow boxes represent preparation
Blue boxes represent ML model development.

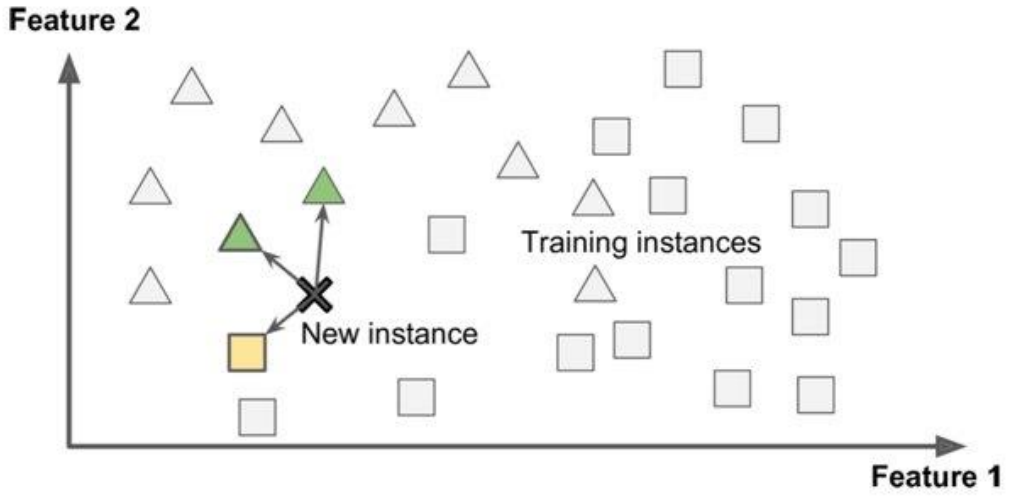
Challenges

- Explainability:
  - What ML models are doing?
  - How they make decisions?
- Bias and unintended outcomes
  - Not enough training data always.
  - Non-representative/biased training data.
  - Less feature rich data.
  - Irrelevant features.
  - Overfitting (overgeneralizing) or underfitting.

Generalization of ML

- How ML generalize?
    - Approaches: Instance-based, Model-based learning.

- Instance-based learning: Learns all previous data and compares new data to it and generalizes based on similarity.

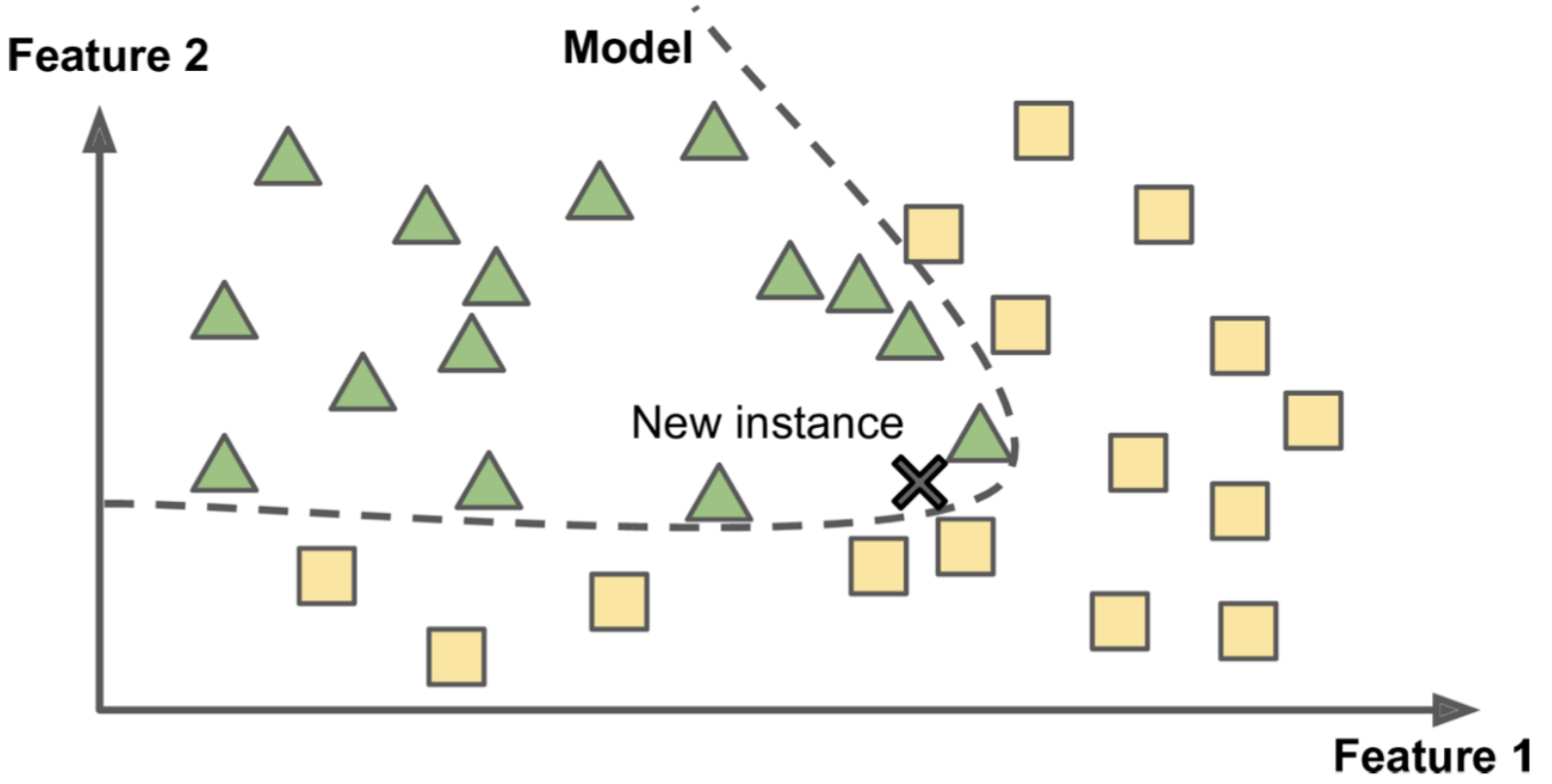


New instance would be classified as a triangle because majority of most similar instances belong to that class.

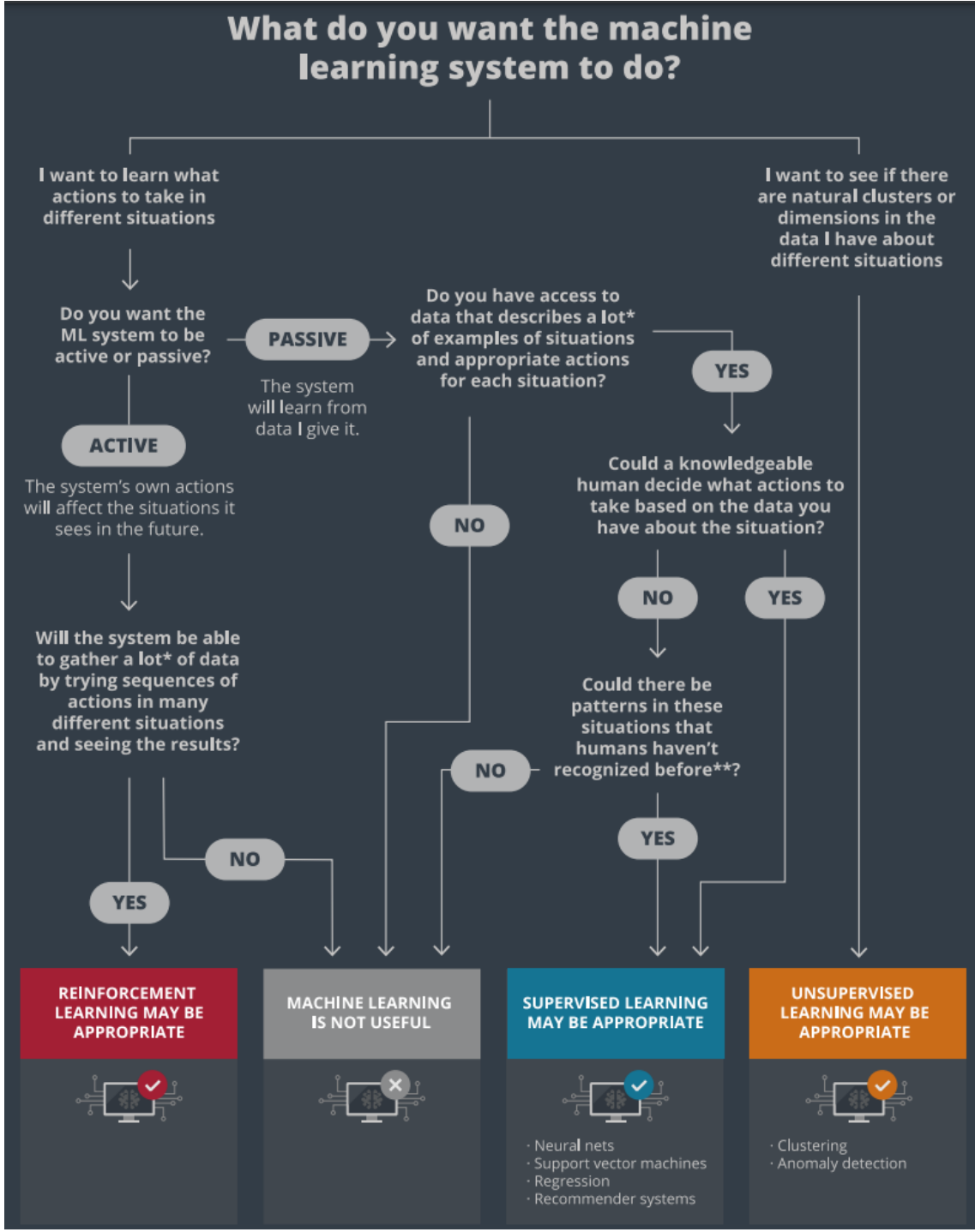
Src: Image from Géron, A. (2019)

Generalization of ML

- Model-based learning: Builds a model based on training data and predicts labels according to model.



Src: Image from Géron, A. (2019)



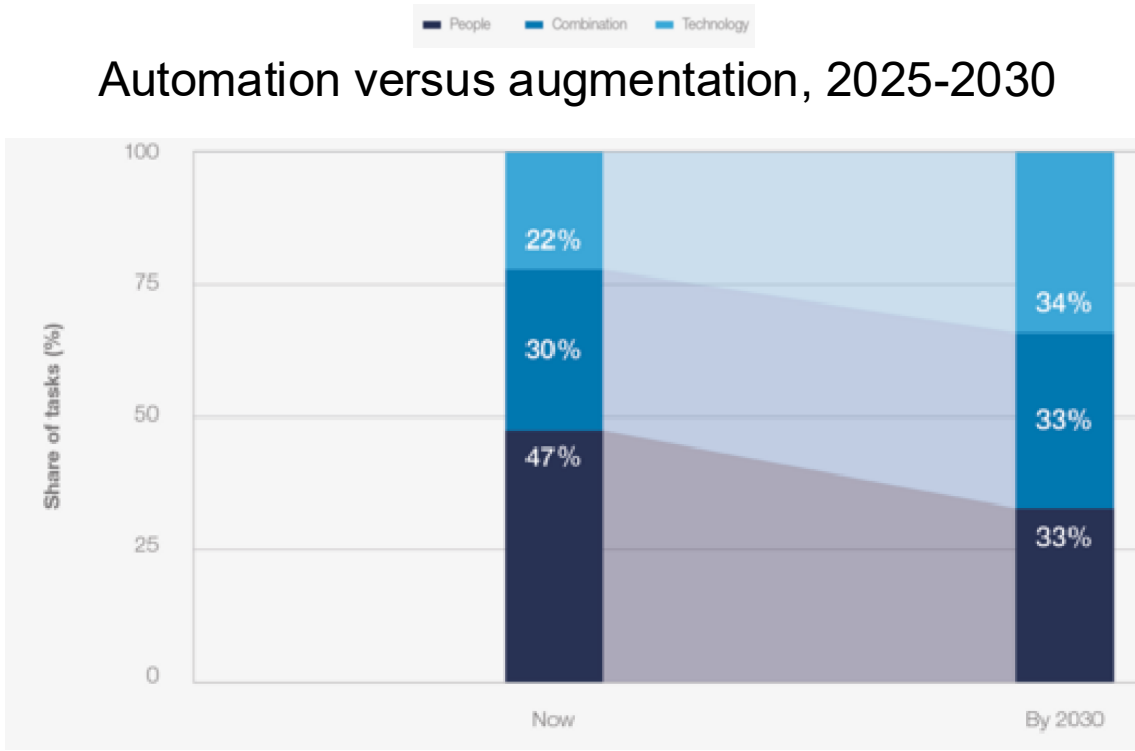
* It's very hard to say how much data will be needed until you start to analyze your problem in detail, but often you need hundreds, thousands, or millions of examples.

** For example: Have humans not focused on this problem before? Was the data about this problem not available before? Can the patterns here, if they exist, only be detected by looking at more data than humans can really process?

Src: Thomas Malone | MIT Sloan. See: https://bit.ly/3gvRho2, Figure 2.

Industry Perspective

- Who is a Data Analyst, Data Engineer and Data Scientist?
  - Data Analyst: analyzes data and infer better decisions. Good understanding of tools.
  - Data Engineer pre-processes data. They develop, tests and maintain complete model.
  - Data Scientist analyses and interpret complex data. Require strong (optimization) knowledge of statistical as well as machine learning model.



https://www.weforum.org/publications/the-future-of-jobs-report-2025/in-full/2-jobs-outlook/