Lecture 01

What is Machine Learning? An Overview.

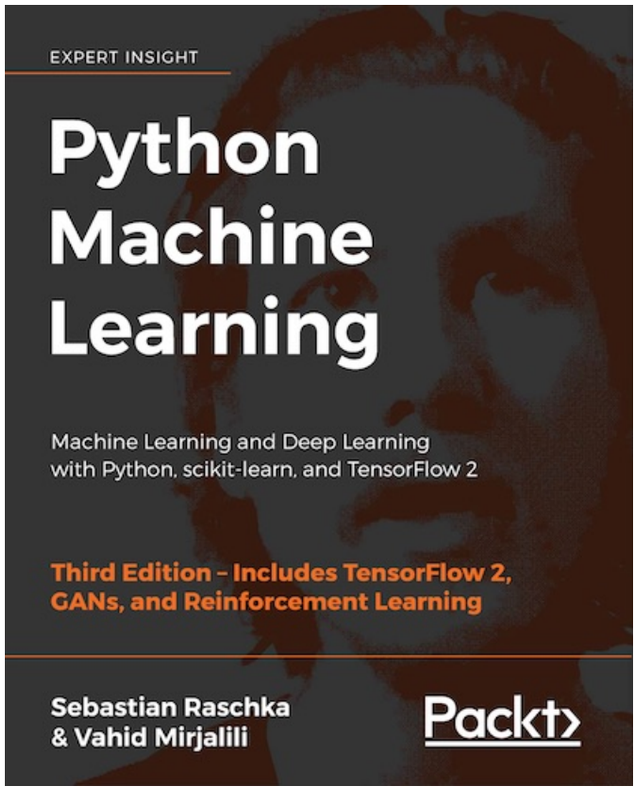
STAT 451: Intro to Machine Learning, Fall 2021

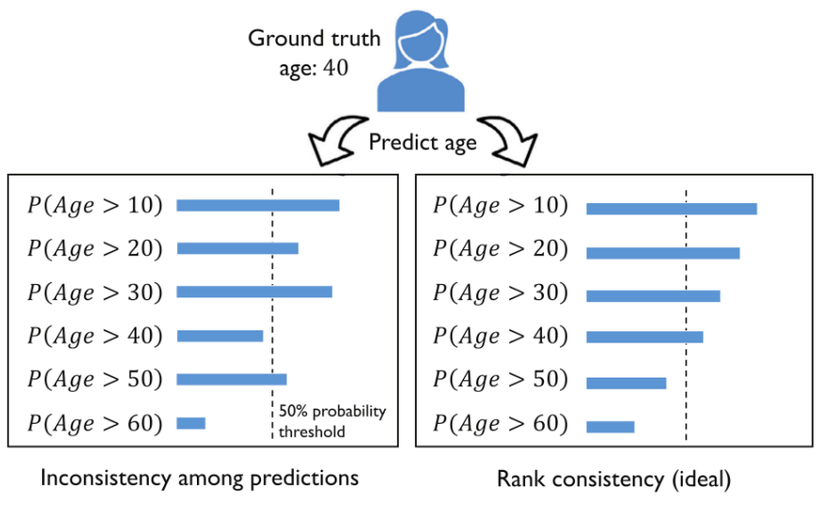
Sebastian Raschka

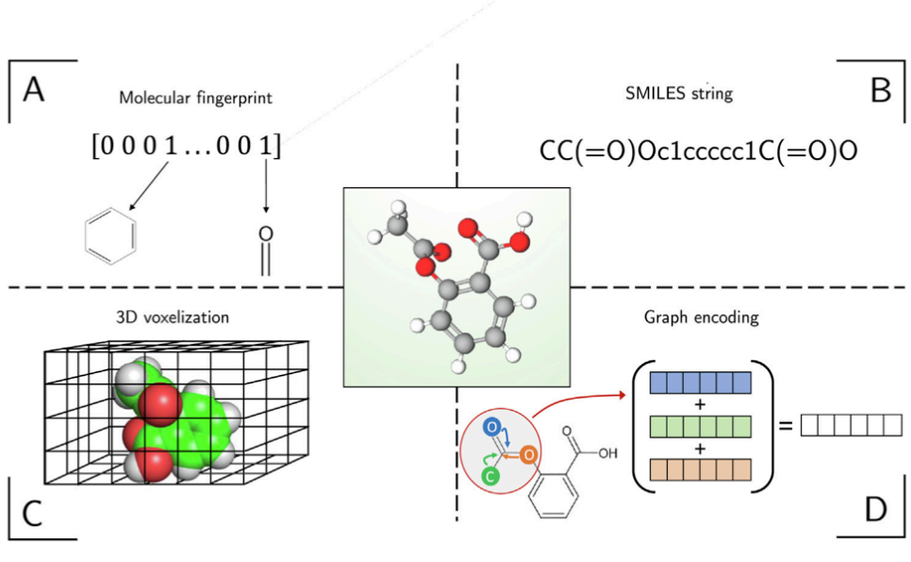
Lecture 1 Overview

1. **About this course**
2. What is machine learning
3. Categories of machine learning
4. Notation
5. Approaching a machine learning application
6. Different machine learning approaches and motivations

About Me

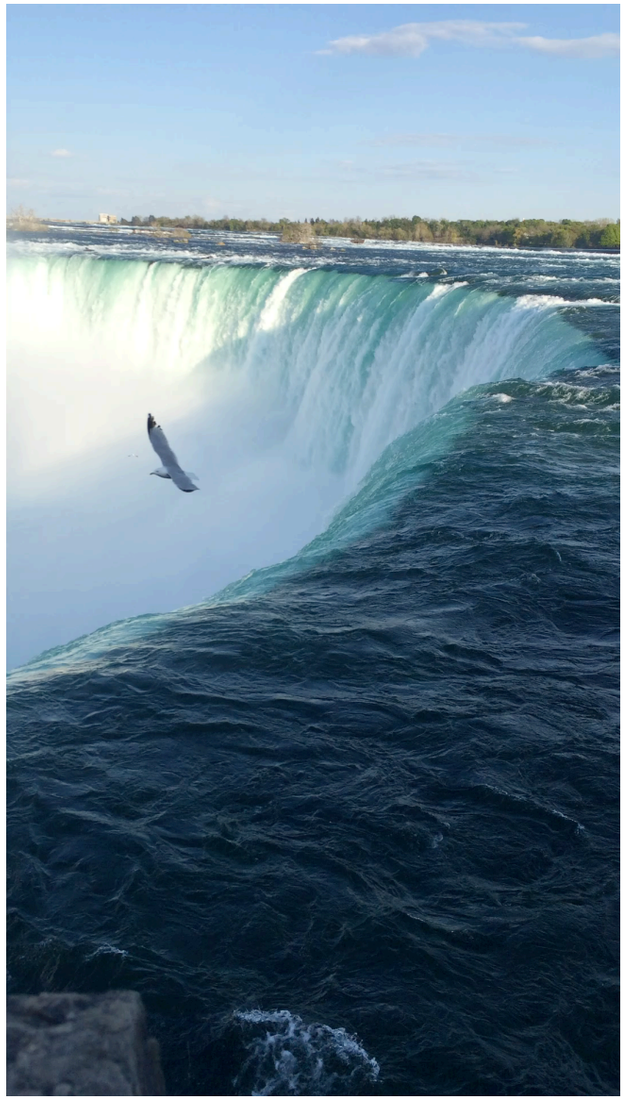


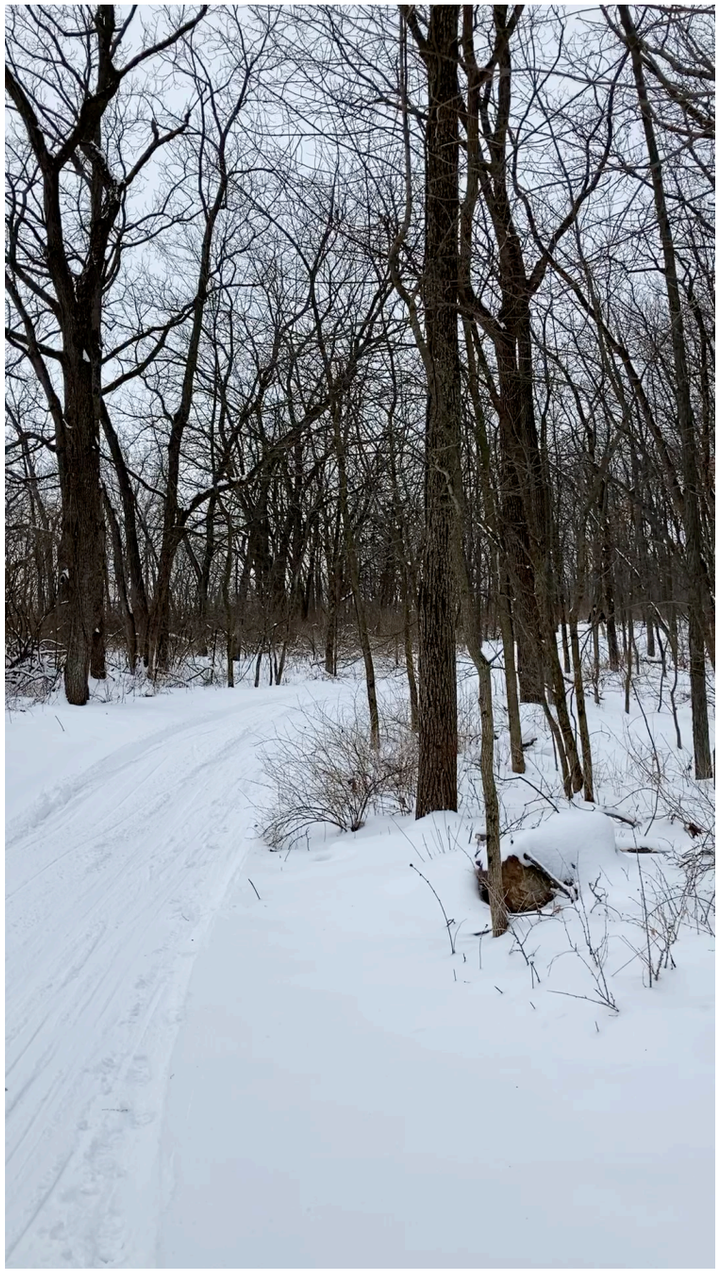




https://sebastianraschka.com/publications/

About Me





Course Topics

Part 1: Introduction

Part 2: Computational foundations

Part 3: Tree-based methods

Part 4: Model evaluation

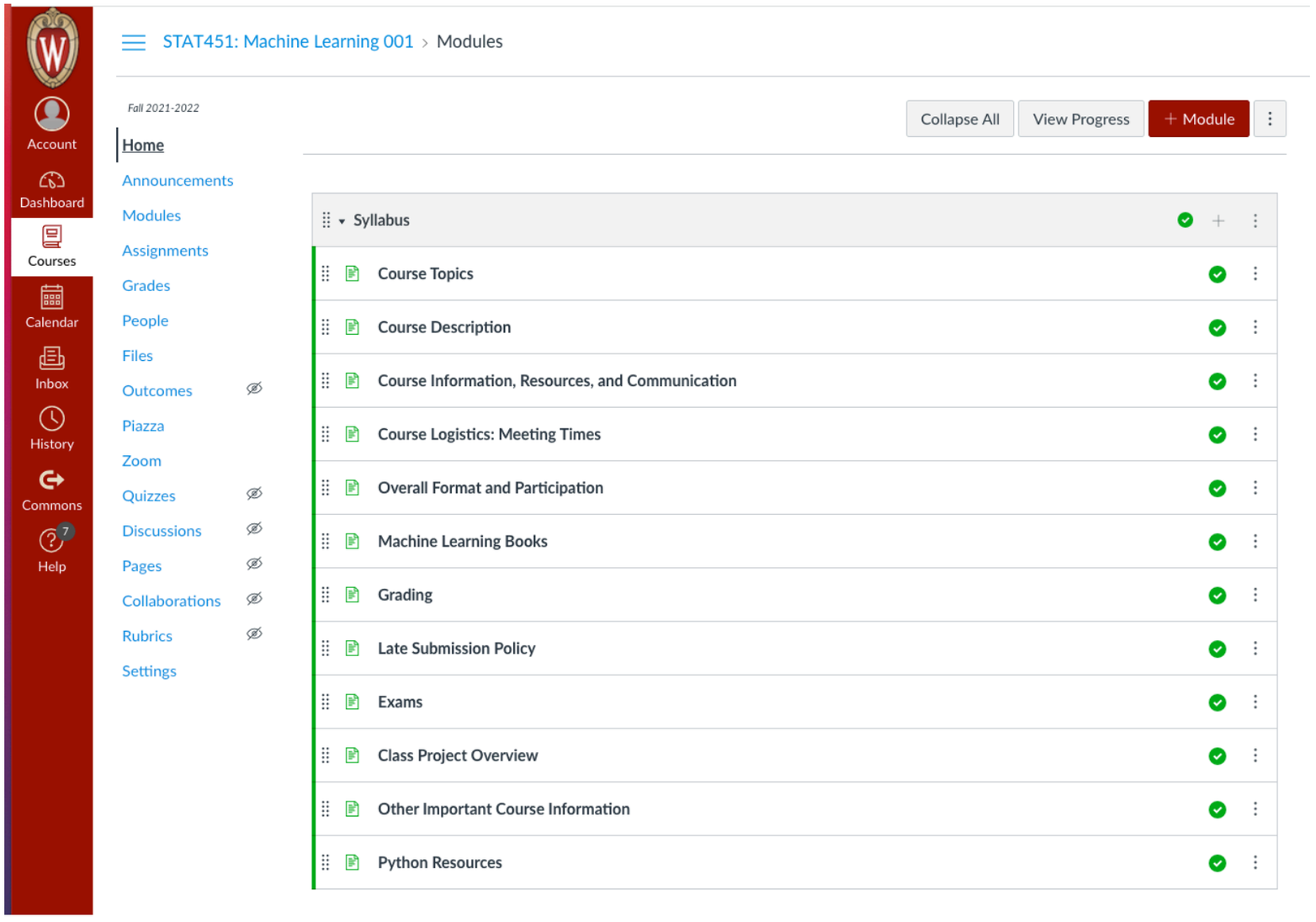
Part 5: Dimensionality reduction and unsupervised learning

Part 6: Bayesian methods

Part 7: Class project presentations

About this Course

Canvas:



For those who don't have access to Canvas, yet:

https://sebastianraschka.com/teaching/stat451-fs2021/

When and Where

Section 1 – Lec 001:

- TuTh 9:30AM - 10:45AM in VILAS 4028

Section 2 – Lec 002

- TuTh 2:30PM - 3:45PM in SMI 331

Office Hours

- **Dr. Sebastian Raschka (Instructor)**
  Time: Tuesday 1:00 - 2:00 pm
  Location: Medical Sciences Center room 1171 (If you enter the building through the main entrance head straight to the elevators. Then, turn left and walk down the hallway. My office should be the 3rd or 4th door on the left.)

- **Jitian Zhao** (Teaching Assistant for Lec 001)
  Time: Thursday 1:00 - 2:00 pm
  Location: virtual via Zoom

- **Yanbo Shen** (Teaching Assistant for Lec 002)
  Time: Thursday 3:50 - 4:50 pm
  Location: virtual via Zoom

Overall Format and Participation

- There are two lectures each week to deliver the main course content.

- A short self-assessment quiz will be posted at the end of each lecture week (Fridays) asking conceptual questions about the lecture's contents. It quiz will be due on the Friday of the following week.

- There will be ~3 hands-on homework assignments involving coding, which will be posted approximately every 3 weeks.

- Starting after the first few weeks of the semester, students will form teams of three to work and collaborate on an individual class project throughout the semester. Students should meet on a regular and weekly basis to make progress towards their project goals.

Grading

The final grade will be computed using the following weighted grading scheme:

- 30% Problem Sets (Homeworks and quizzes)
- 20% Midterm Exam
- 50% Class Project:
  - 5% Project proposal
  - 20% Project presentation
  - 25% Project report

The final letter grade will be based on the percent of the total points accumulated in the course. The proposed grade cut-offs are as follows:

- A: >= 93%
- AB: >= 90%
- B: >= 85%
- BC: >= 80%
- C: >= 70%
- D: >= 50%
- F: < 50%

However, please note that the grades are subject to curving and the cut-offs may be adjusted such that a grade distribution similar to previous semester can be achieved.

Lecture 1 Overview

1. About this course

2. **What is machine learning**

3. Categories of machine learning

4. Notation

5. Approaching a machine learning application

6. Different machine learning approaches and motivations

What is Machine Learning?

"Machine learning is the hot new thing."

-- John L. Hennessy, President of Stanford (2000-2016)

Image Source: https://www.innovateli.com/hennessy-grad-keeps-gifting/

"A breakthrough in machine learning would be worth ten Microsofts"

-- Bill Gates, Microsoft Co-founder

Image source: https://www.gatesnotes.com/Books

[...] machine learning is a subcategory within the field of computer science, which allows you to implement artificial intelligence. So it's kind of a mechanism to get you to artificial intelligence.

-- Rana el Kaliouby, CEO at Affectiva

Image Source: https://fortune.com/2019/03/08/rana-el-kaliouby-ceo-affectiva/

Image Source: https://history-computer.com/ModernComputer/thinkers/images/Arthur-Samuel1.jpg

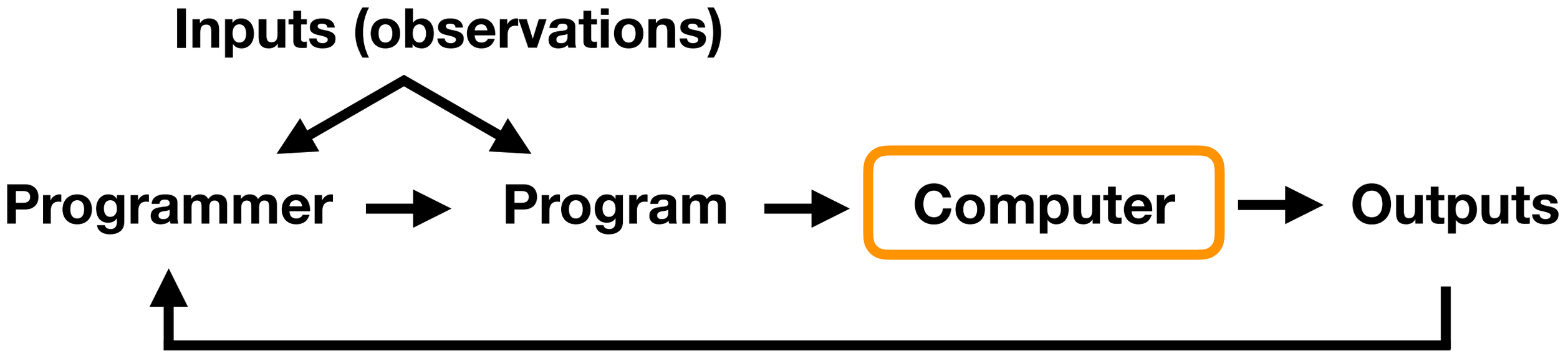
"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

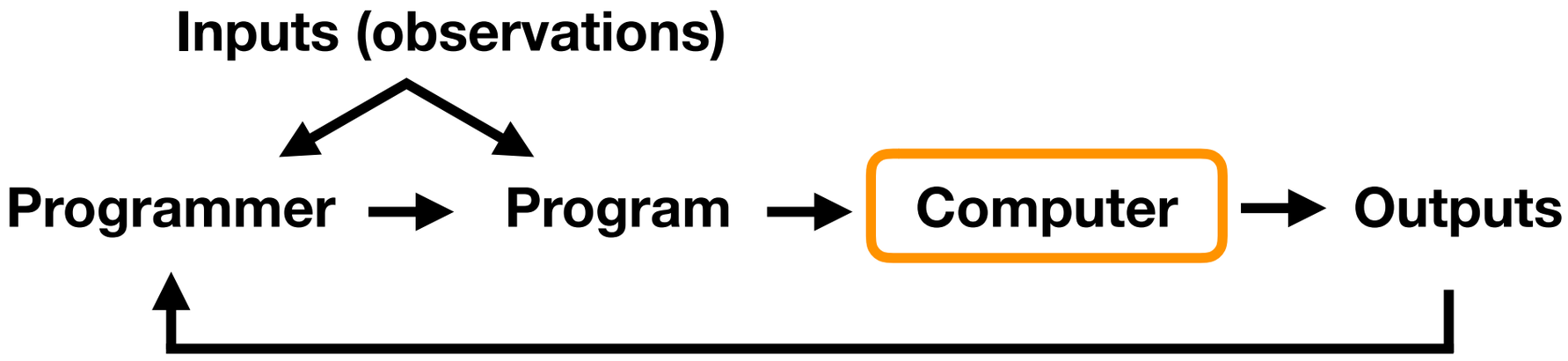
— Arthur L. Samuel, AI pioneer, 1959

(This is likely not an original quote but a paraphrased version of Samuel's sentence "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.")

Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

The Traditional Programming Paradigm





Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed

— *Arthur Samuel (1959)*

We will not only use the machines for their intelligence, we will also collaborate with them in ways that we cannot even imagine.

-- Fei Fei Li, Director of Stanford's artificial intelligence lab

Image Source: https://en.wikipedia.org/wiki/Fei-Fei_Li#/media/File:Fei-Fei_Li_at_AI_for_Good_2017.jpg

Some Applications of Machine Learning:

-
-
-
-

Lecture 1 Overview

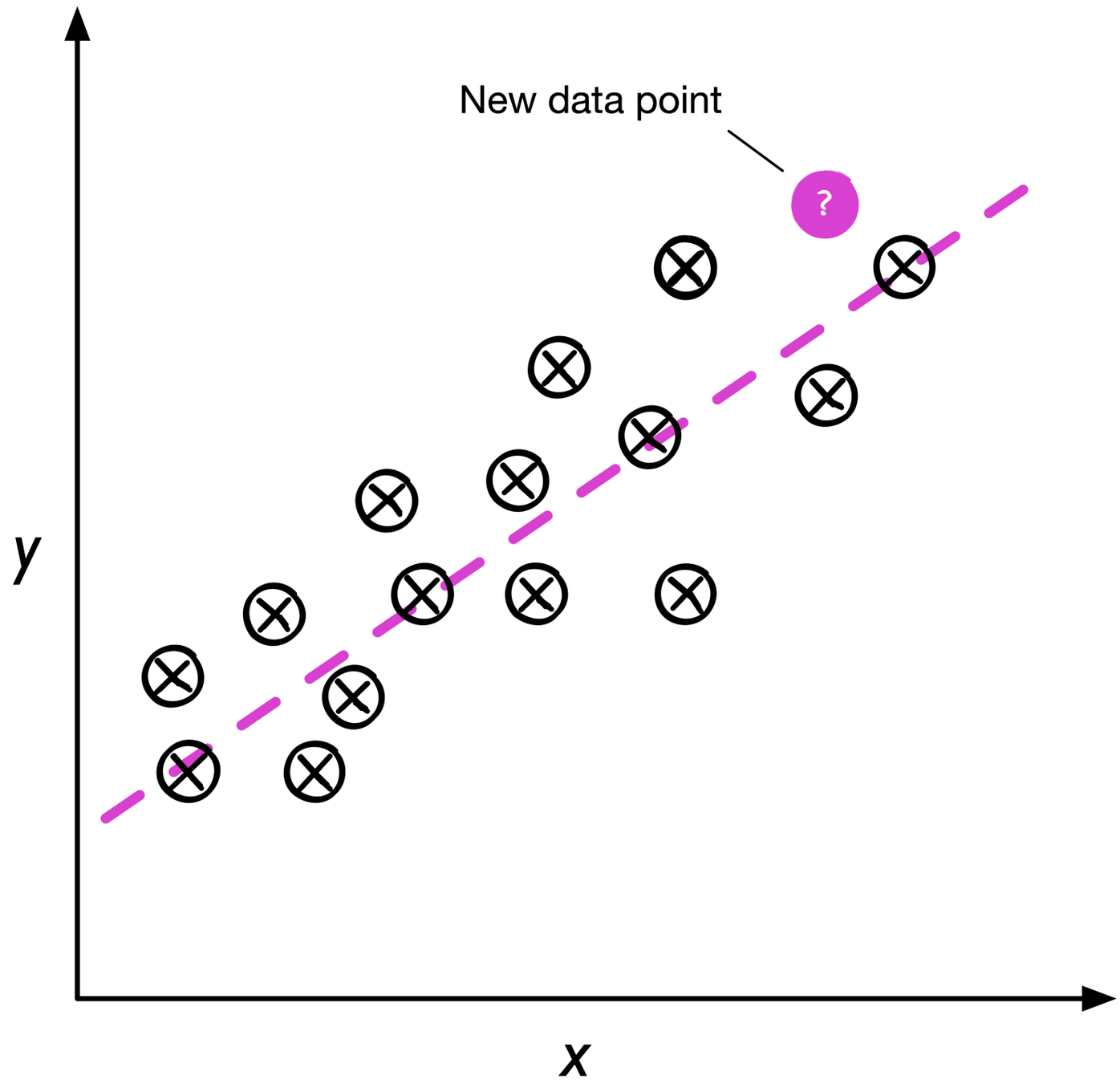
1. About this course
2. What is machine learning
3. **Categories of machine learning**
4. Notation
5. Approaching a machine learning application
6. Different machine learning approaches and motivations

Categories of Machine Learning

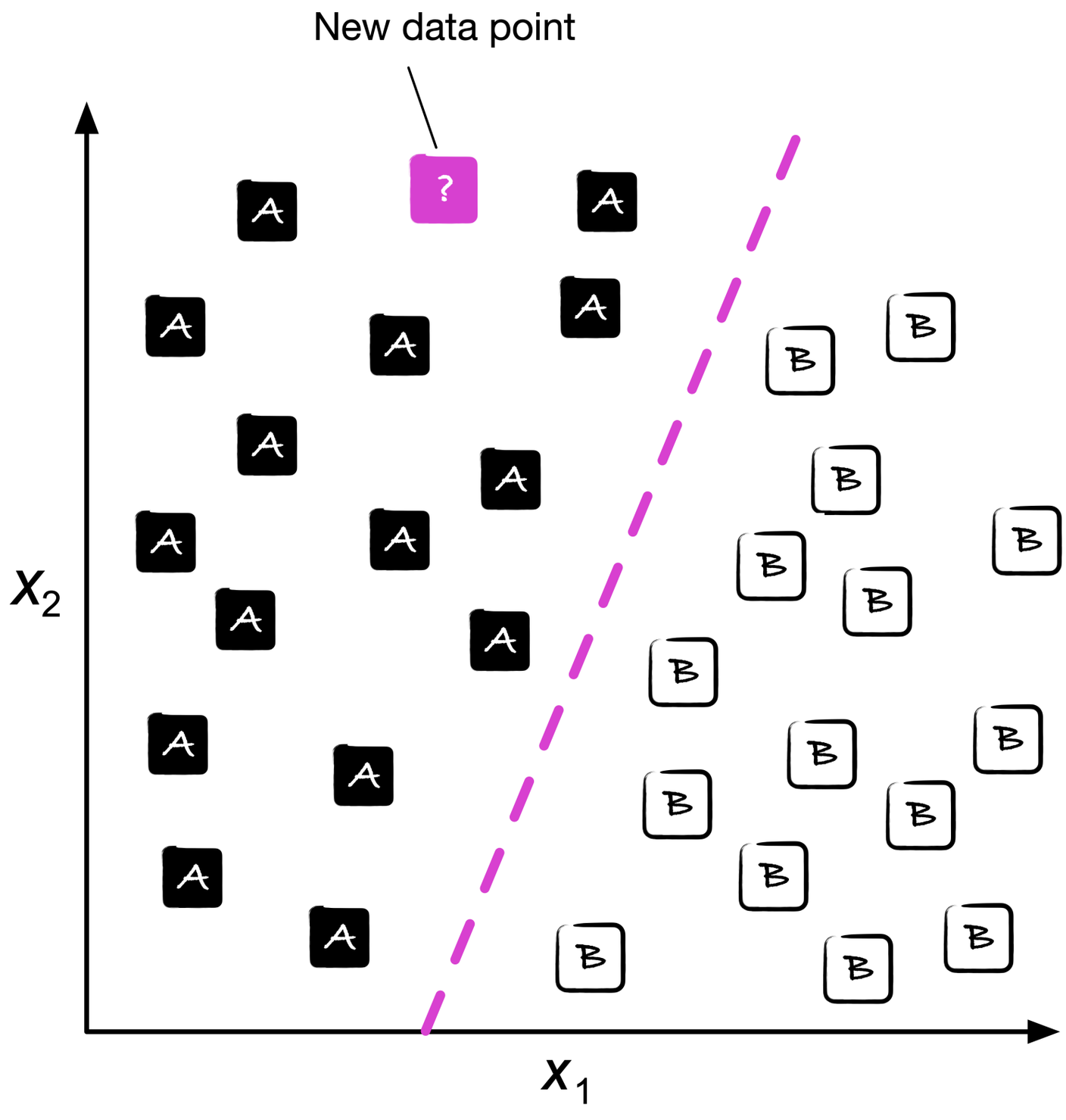
Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

Supervised Learning: Regression



Supervised Learning: Classification



Categories of Machine Learning

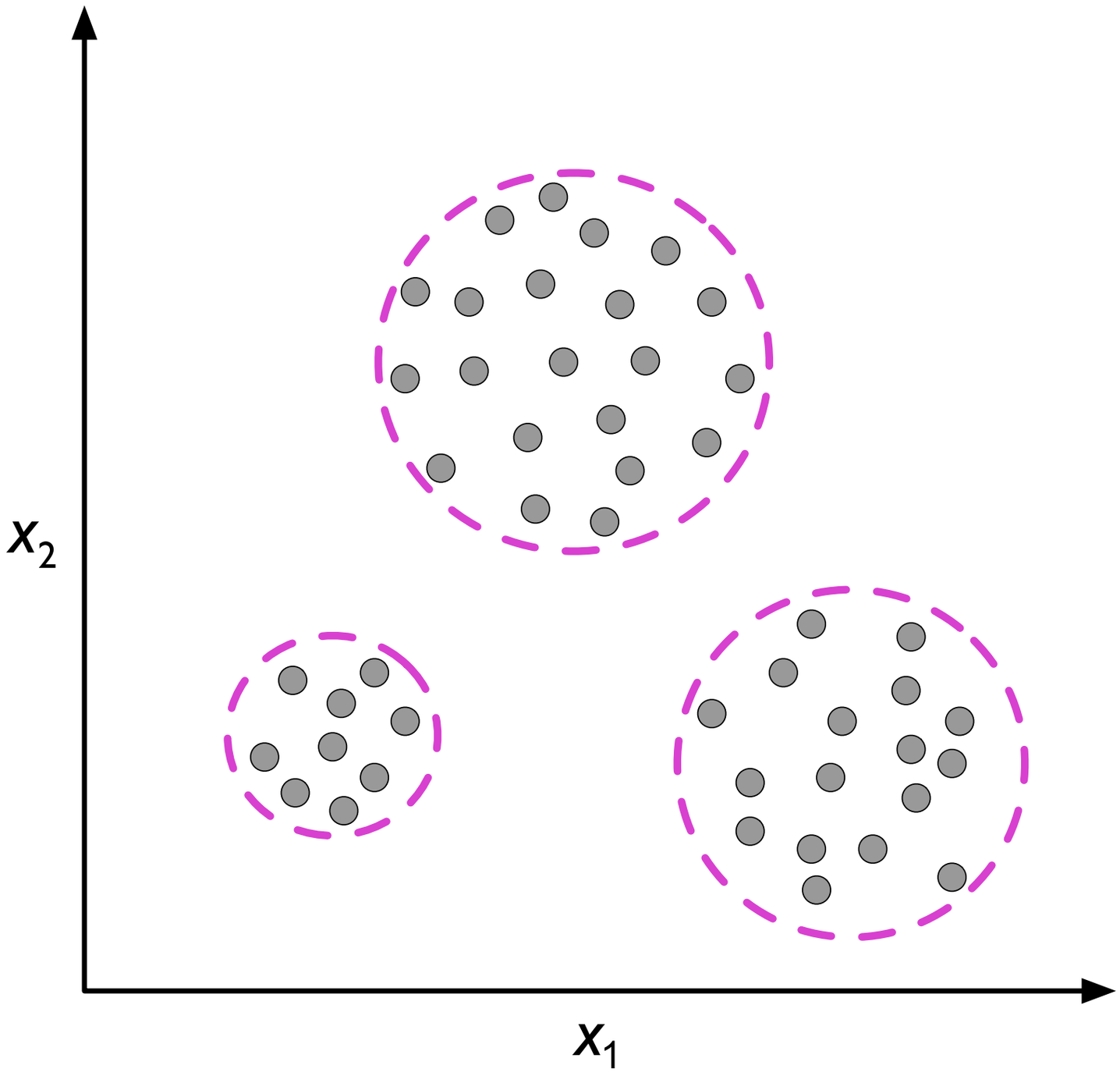
Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

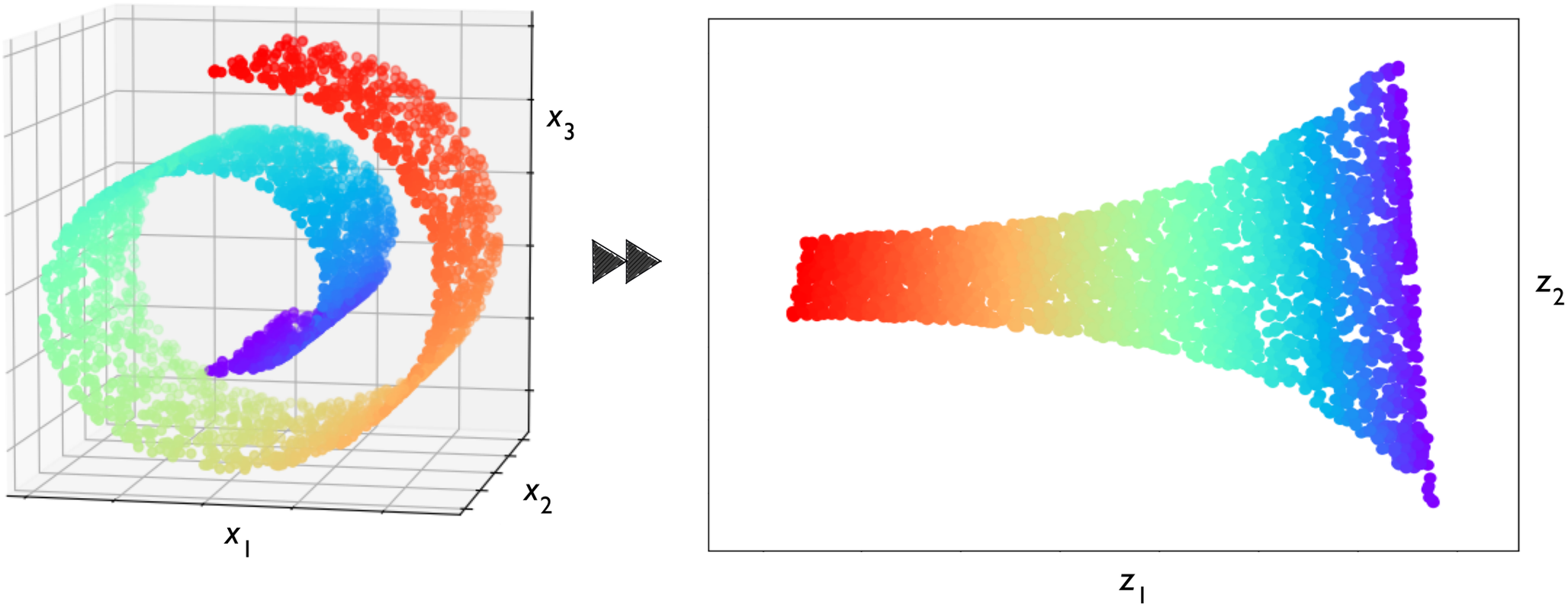
Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

Unsupervised Learning -- Clustering



Unsupervised Learning
-- Dimensionality Reduction



Categories of Machine Learning

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

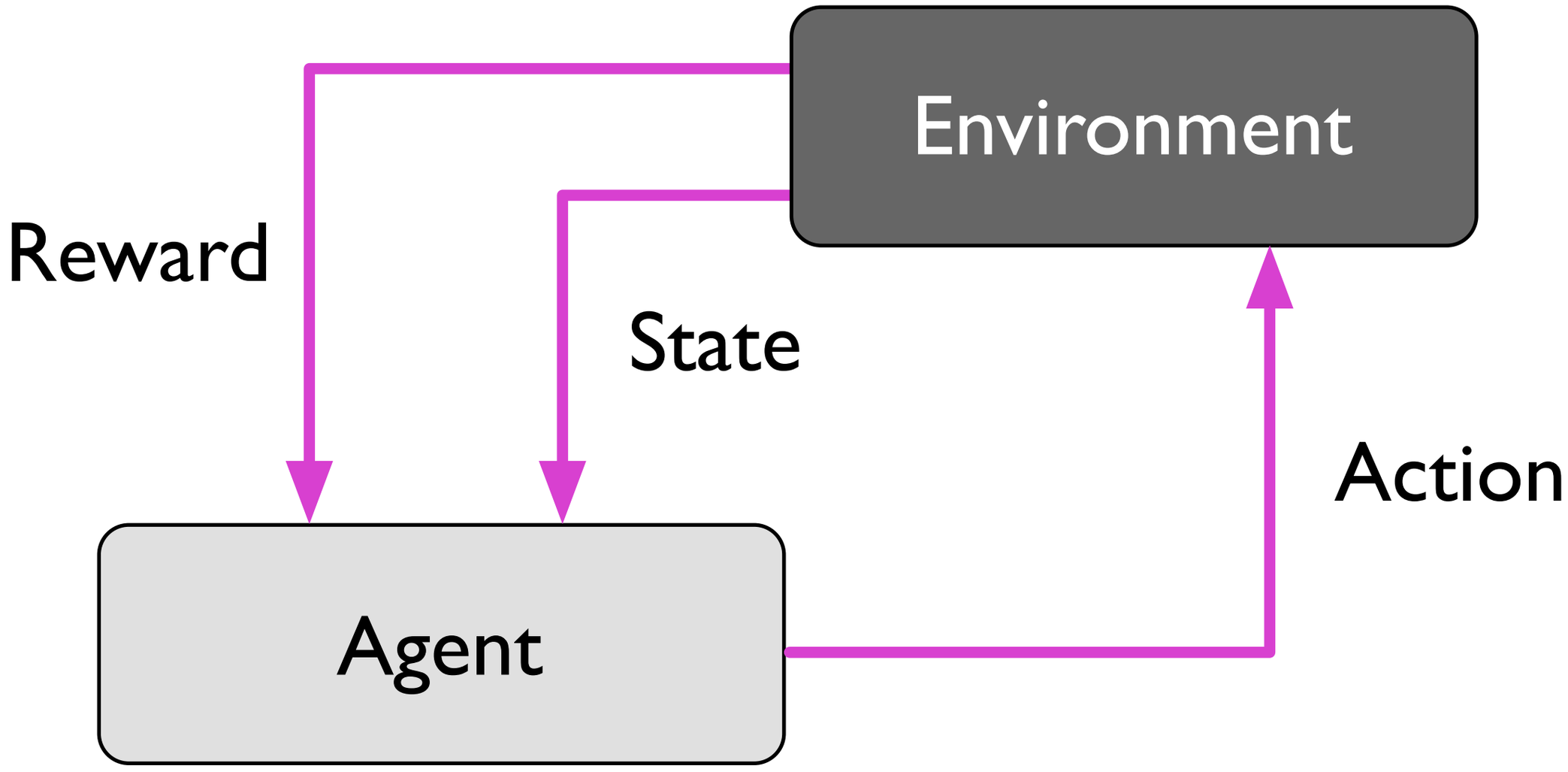
Unsupervised Learning

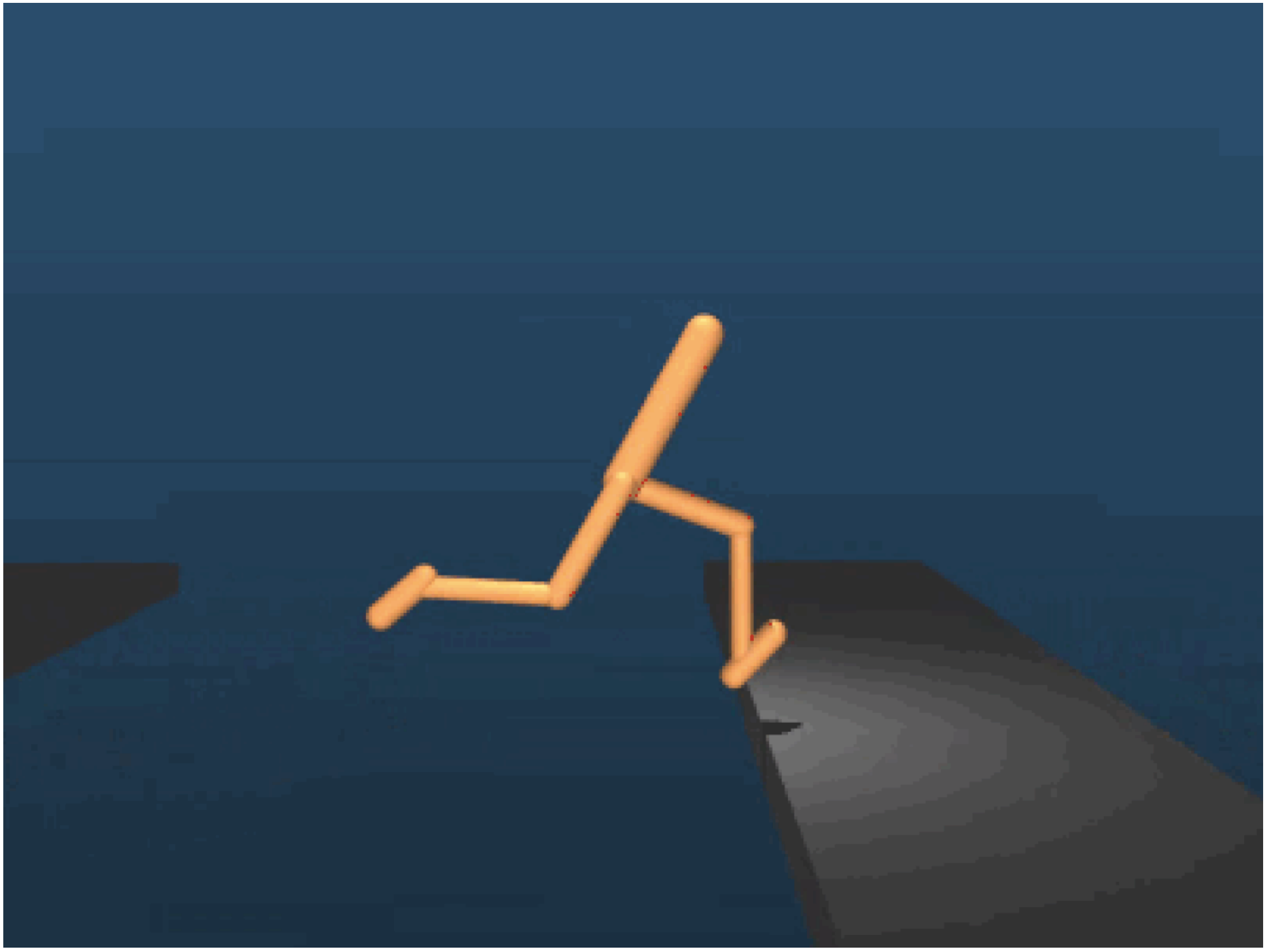
- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning

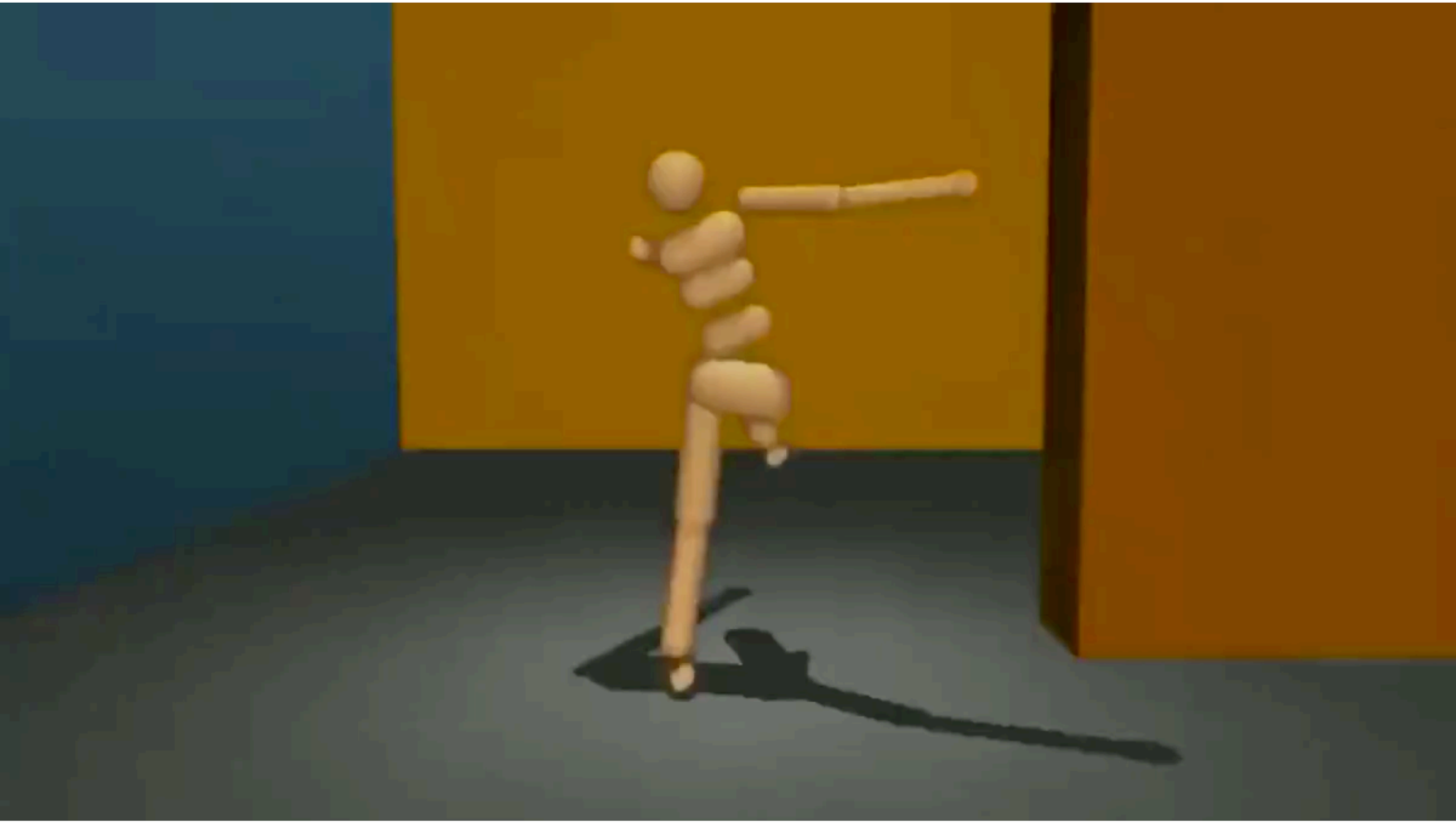
- Decision process
- Reward system
- Learn series of actions

Reinforcement Learning





https://www.theverge.com/tldr/2017/7/10/15946542/deepmind-parkour-agent-reinforcement-learning

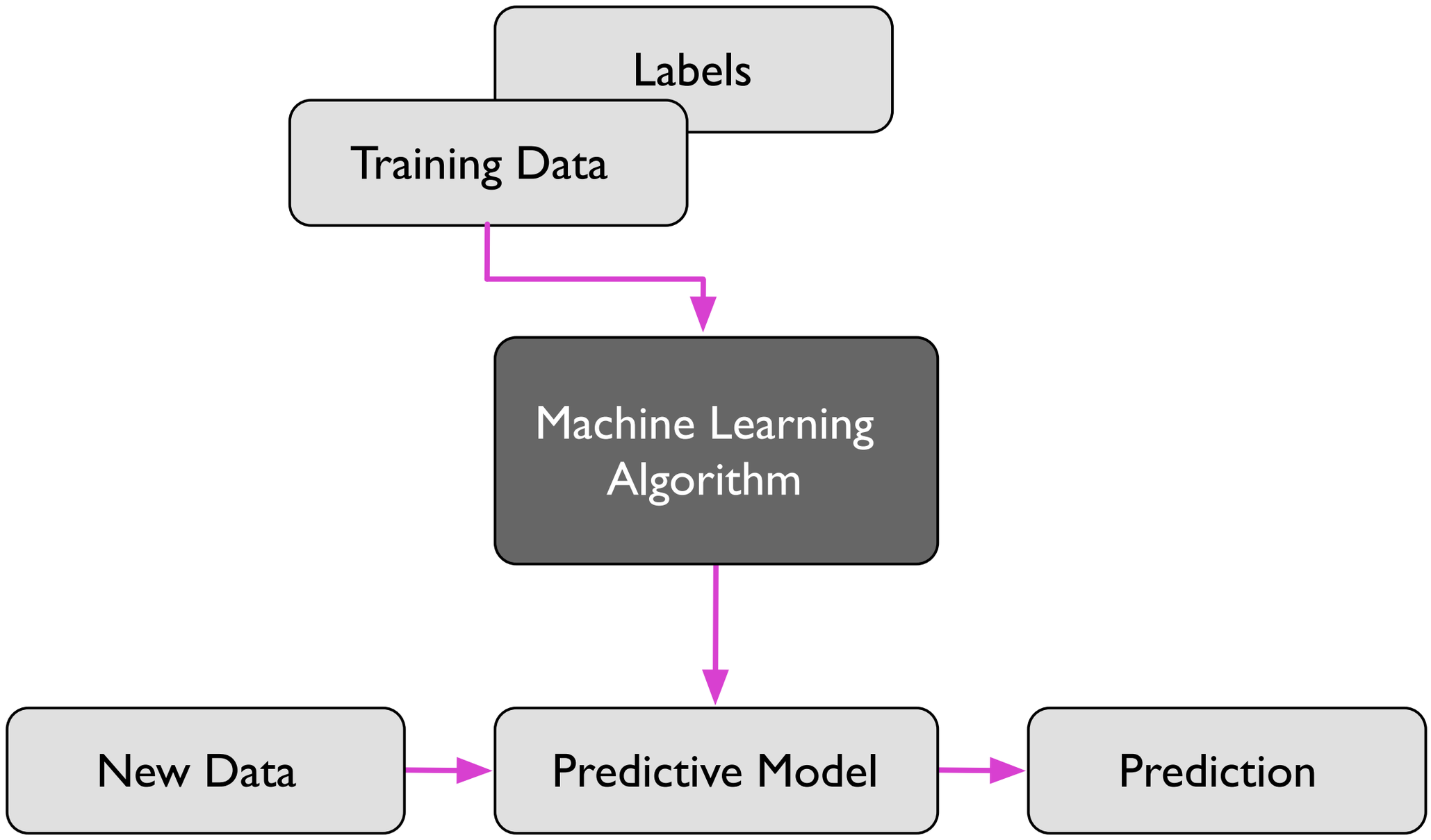


https://video.twimg.com/ext_tw_video/1111683489890332672/pu/vid/1200x674/WqUJEhUETw0M0gCl.mp4?tag=8

Lecture 1 Overview

1. About this course
2. What is machine learning
3. Categories of machine learning
4. **Notation**
5. Approaching a machine learning application
6. Different machine learning approaches and motivations

Supervised Learning Workflow -- Overview



Supervised Learning Notation

Training set: $\mathscr{D} = \{\langle \mathbf{x}^{[i]}, y^{[i]} \rangle, i = 1, \ldots, n\},$

Unknown function: $f(\mathbf{x}) = y$

Hypothesis: $h(\mathbf{x}) = \hat{y}$

Classification

$h : \mathbb{R}^m \rightarrow \_\_$

Regression

$h : \mathbb{R}^m \rightarrow \_\_$

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature vector

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Feature vector

D___n m_____

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

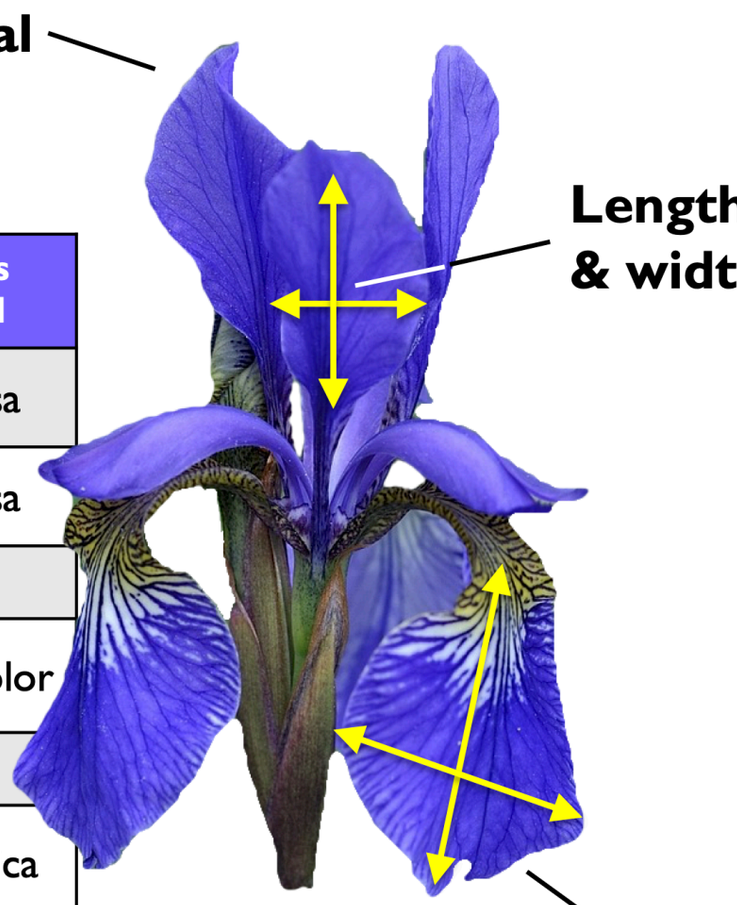
$$\mathbf{X} = \begin{bmatrix} x_1^{[1]} & x_2^{[1]} & \cdots & x_m^{[1]} \\ x_1^{[2]} & x_2^{[2]} & \cdots & x_m^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{[n]} & x_2^{[n]} & \cdots & x_m^{[n]} \end{bmatrix}$$

Feature vector

Data Representation

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(inputs, attributes, measurements, dimensions)

Class labels
(targets, outcomes)

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{[1]} \\ y^{[2]} \\ \vdots \\ y^{[n]} \end{bmatrix}$$

Input features

ML Terminology (Part 1)

- **Training example**: A row in the table representing the dataset. Synonymous to an observation, training record, training instance, training sample (in some contexts, sample refers to a collection of training examples)

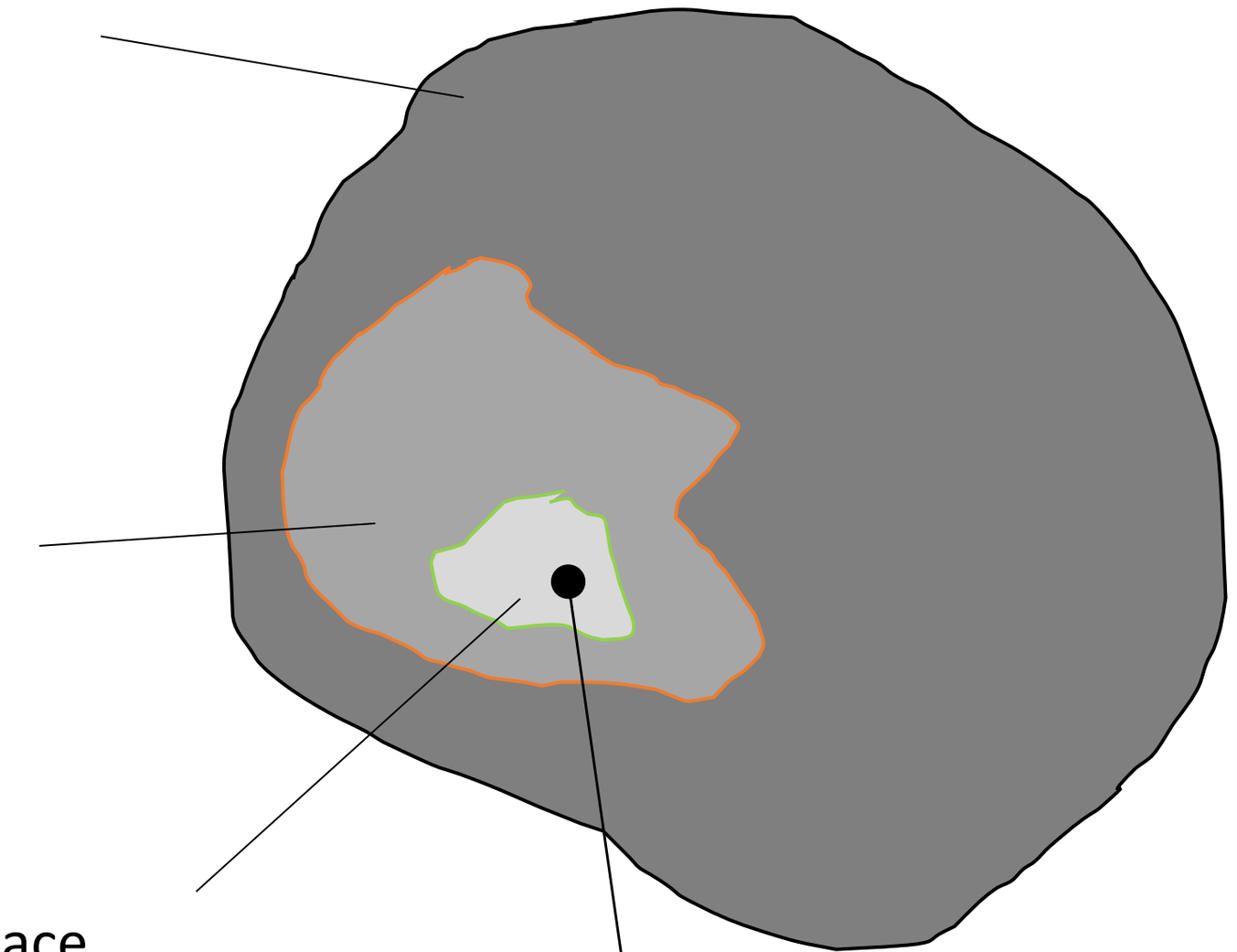
- **Feature**: a column in the table representing the dataset. Synonymous to predictor, variable, input, attribute, covariate.

- **Targets**: What we want to predict. Synonymous to outcome, output, ground truth, response variable, dependent variable, (class) label (in classification).

- **Output / prediction**: use this to distinguish from targets; here, means output from the model.

Hypothesis Space

Entire hypothesis space



Hypothesis space
a particular learning
algorithm category
has access to

Hypothesis space
a particular learning
algorithm can sample

Particular hypothesis
(i.e., a model/classifier)

Classes of Machine Learning Algorithms

- Generalized linear models
- Support vector machines
- Artificial neural networks
- Tree- or rule-based models
- Graphical models
- Ensembles
- Instance-based learners

Lecture Overview

1. About this course
2. What is machine learning
3. Categories of machine learning
4. Notation
5. **Approaching a machine learning application**
6. Different machine learning approaches and motivations

5 Steps for Approaching a Machine Learning Application

1. Define the problem to be solved.

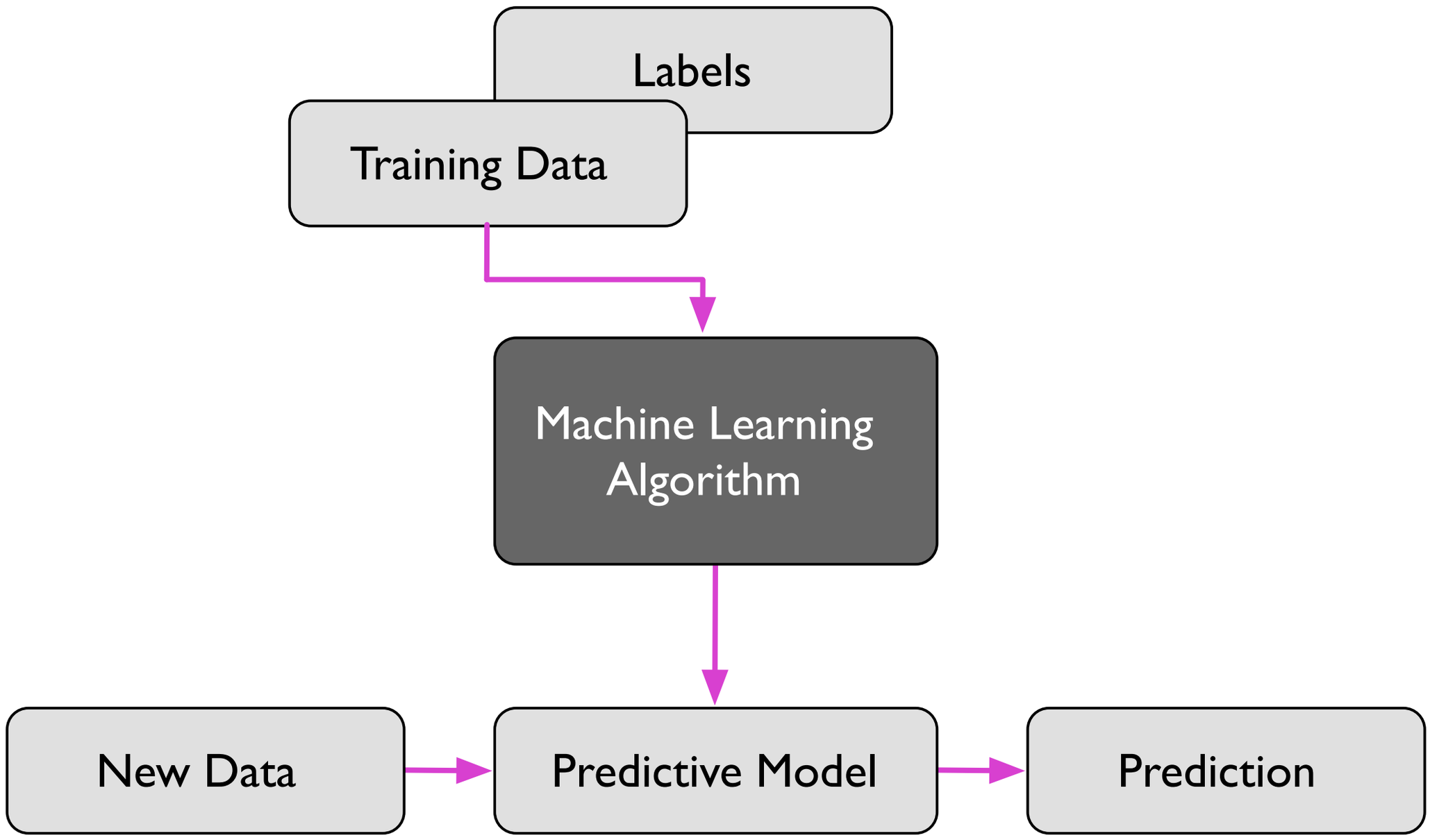
2. Collect (labeled) data.

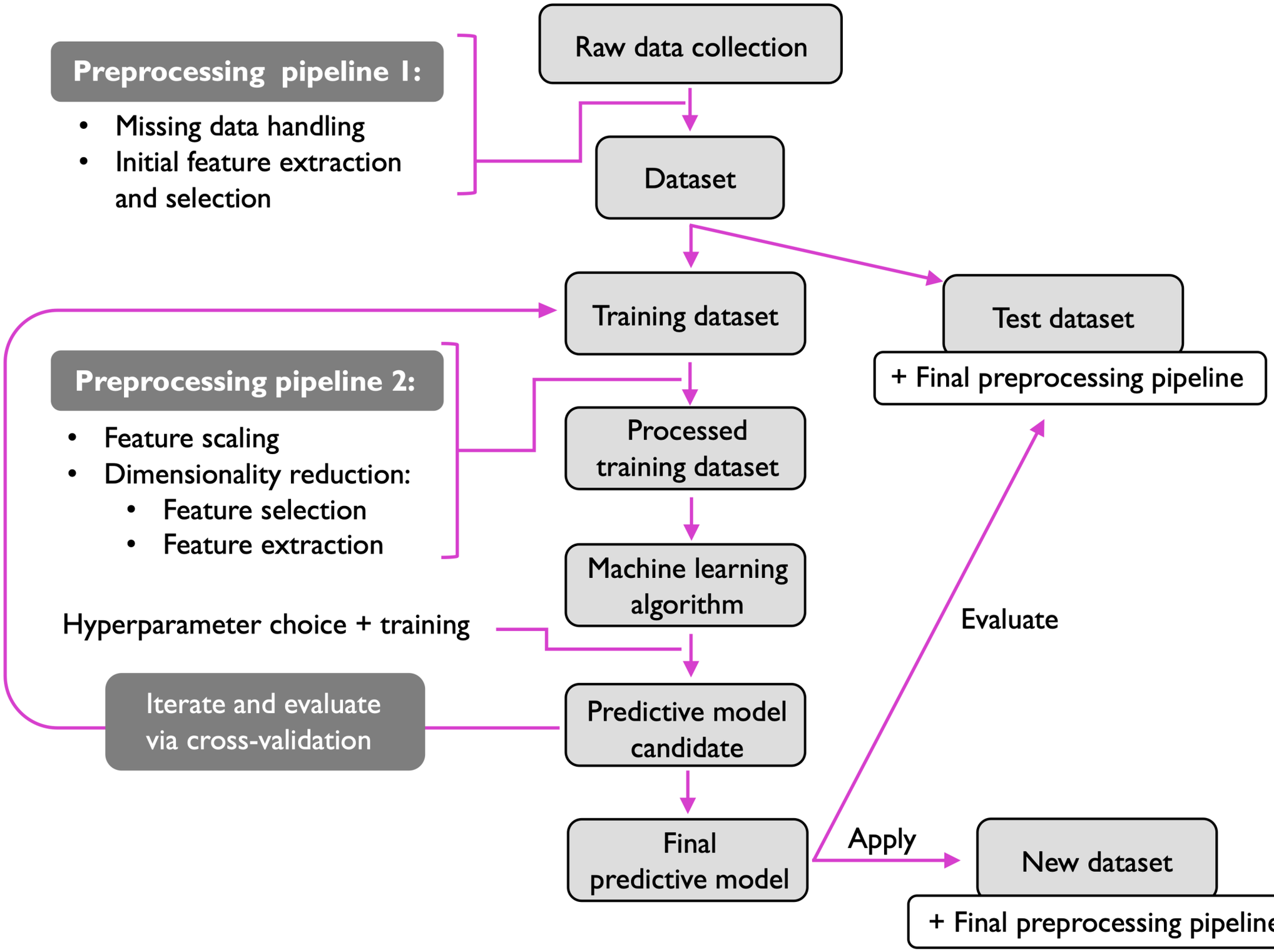
3. Choose an algorithm class.

4. Choose an optimization metric or measure for learning the model.

5. Choose a metric or measure for evaluating the model.

Supervised Learning Workflow -- Overview




Preprocessing pipeline 1:
• Missing data handling
• Initial feature extraction and selection
Raw data collection
Dataset
Training dataset
Test dataset
+ Final preprocessing pipeline
Preprocessing pipeline 2:
• Feature scaling
• Dimensionality reduction:
• Feature selection
• Feature extraction
Processed training dataset
Machine learning algorithm
Hyperparameter choice + training
Iterate and evaluate via cross-validation
Predictive model candidate
Final predictive model
Evaluate
Apply
New dataset
+ Final preprocessing pipeline

ML Terminology (Part 2)

- **Hypothesis:** A hypothesis is a certain function that we believe (or hope) is similar to the true function, the target function that we want to model.

- **Model:** In the machine learning field, the terms hypothesis and model are often used interchangeably. In other sciences, they can have different meanings.

- **Learning algorithm:** Again, our goal is to find or approximate the target function, and the learning algorithm is a set of instructions that tries to model the target function using our training dataset. A learning algorithm comes with a hypothesis space, the set of possible hypotheses it explores to model the unknown target function by formulating the final hypothesis.

- **Classifier:** A classifier is a special case of a hypothesis (nowadays, often learned by a machine learning algorithm). A classifier is a hypothesis or discrete-valued function that is used to assign (categorical) class labels to particular data points

Lecture 1 Overview

1. About this course
2. What is machine learning
3. Categories of machine learning
4. Notation
5. Approaching a machine learning application
6. **Different machine learning approaches and motivations**

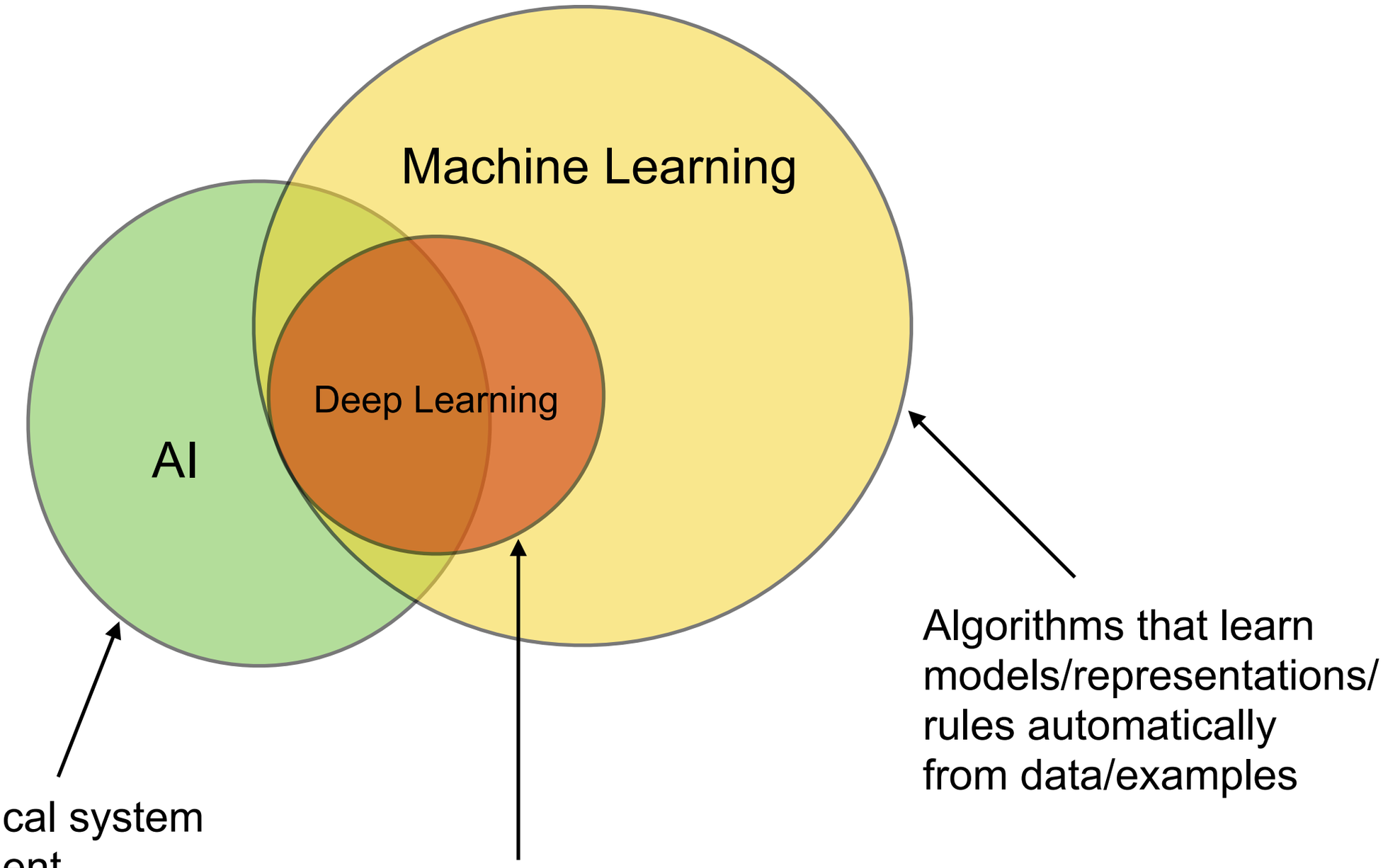
Different Motivations for Studying Machine Learning

- Engineers:

- Mathematicians, computer scientists, and statisticians:

- Neuroscientists:

Machine Learning, AI, and Deep Learning



Cornell University

arXiv.org > cs > arXiv:2106.03253

Computer Science > Machine Learning

[Submitted on 6 Jun 2021]

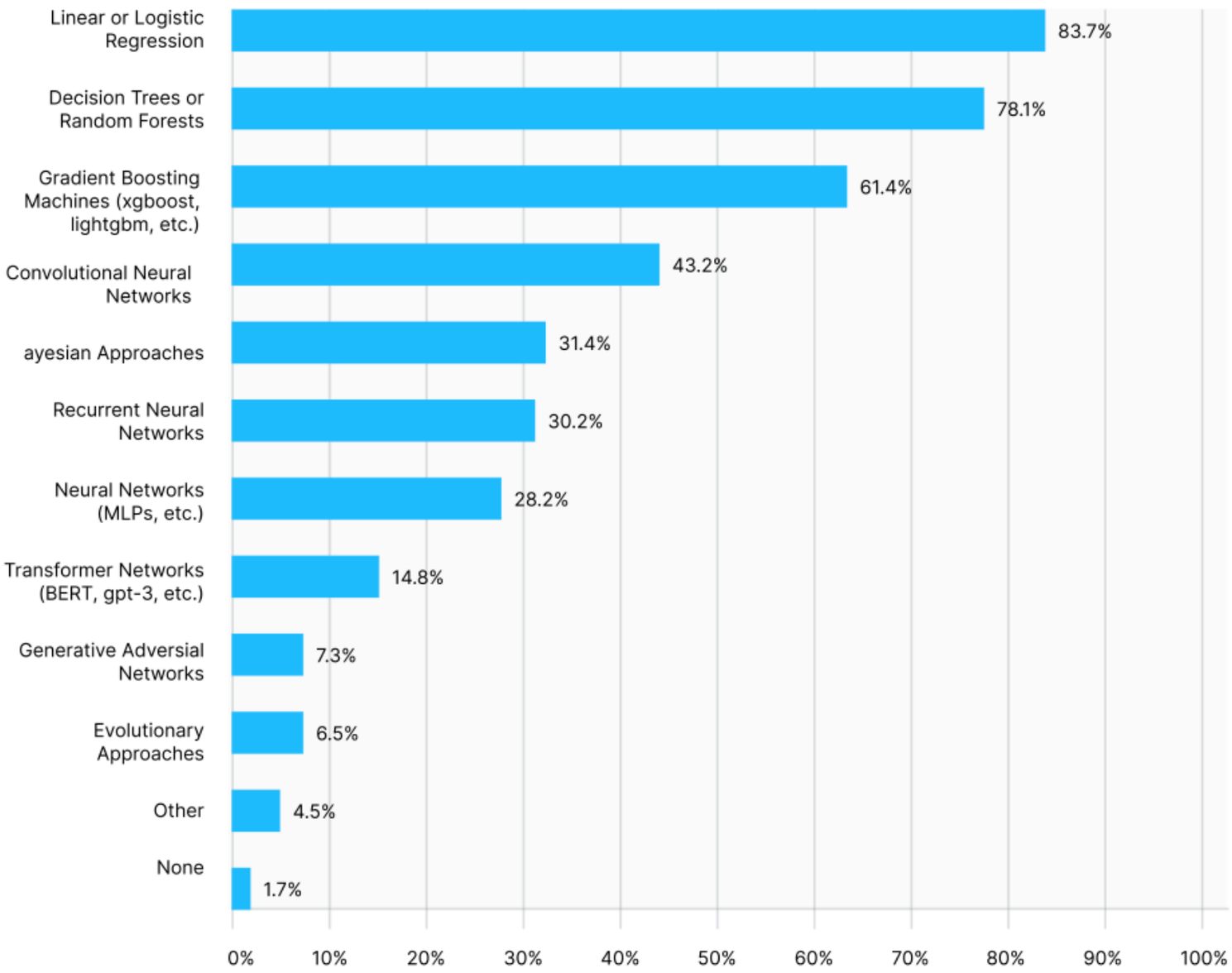
Tabular Data: Deep Learning is Not All You Need

Ravid Shwartz-Ziv, Amitai Armon

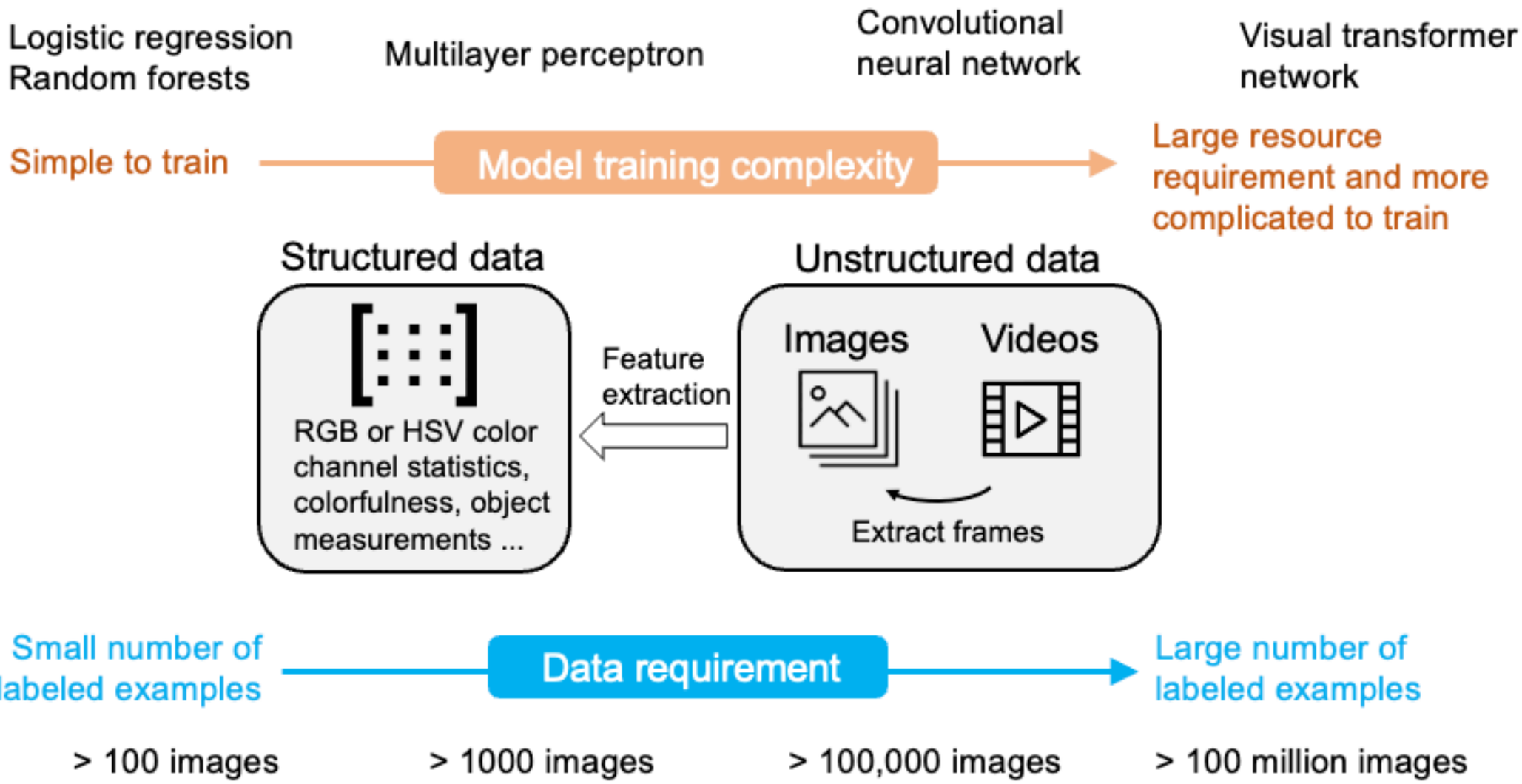
https://arxiv.org/abs/2106.03253

Methods & Algorithms

The most commonly used algorithms were linear and logistic regression, followed closely by decision trees and random forests. Of more complex methods, gradient boosting machines and convolutional neural networks were the most popular approaches.



https://www.kaggle.com/kaggle-survey-2020



https://arxiv.org/abs/2102.01163

Visual Framing of Science Conspiracy Videos: Integrating Machine Learning with Communication Theories to Study the Use of Color and Brightness
Kaiping Chen, Sang Jung Kim, Sebastian Raschka, Qiantong Gao

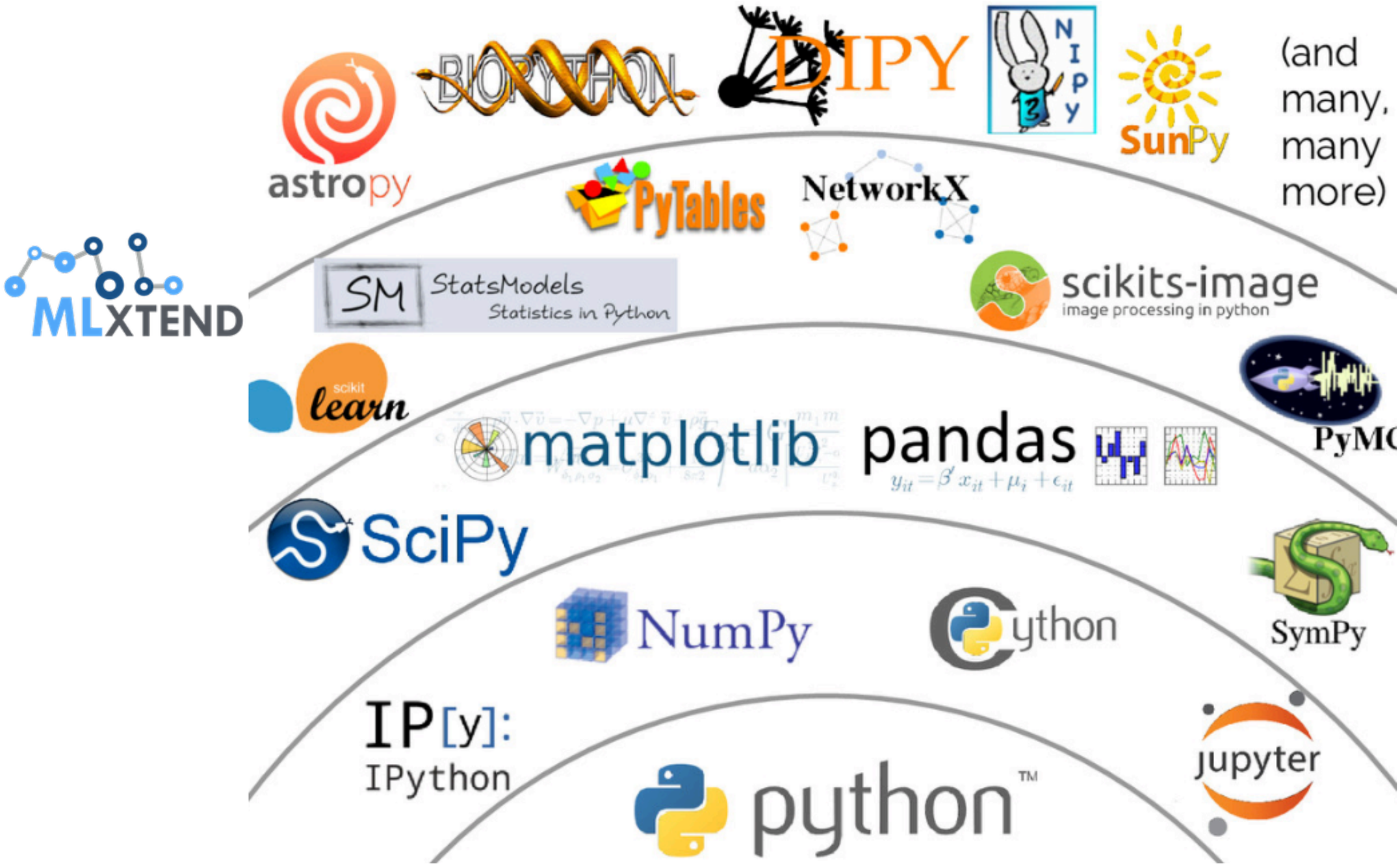


Image by Jake VanderPlas; Source: https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8)

ANALYSIS | COMPUTING

Top Programming Languages 2021 › Python dominates as the de facto platform for new technologies

BY STEPHEN CASS | 24 AUG 2021 | 3 MIN READ |

https://spectrum.ieee.org/top-programming-languages-2021

Rank	Language	Type	Score
1	Python		100.0
2	Java		95.4
3	C		94.7
4	C++		92.4
5	JavaScript		88.1
6	C#		82.4
7	R		81.7
8	Go		77.7
9	HTML		75.4
10	Swift		70.4

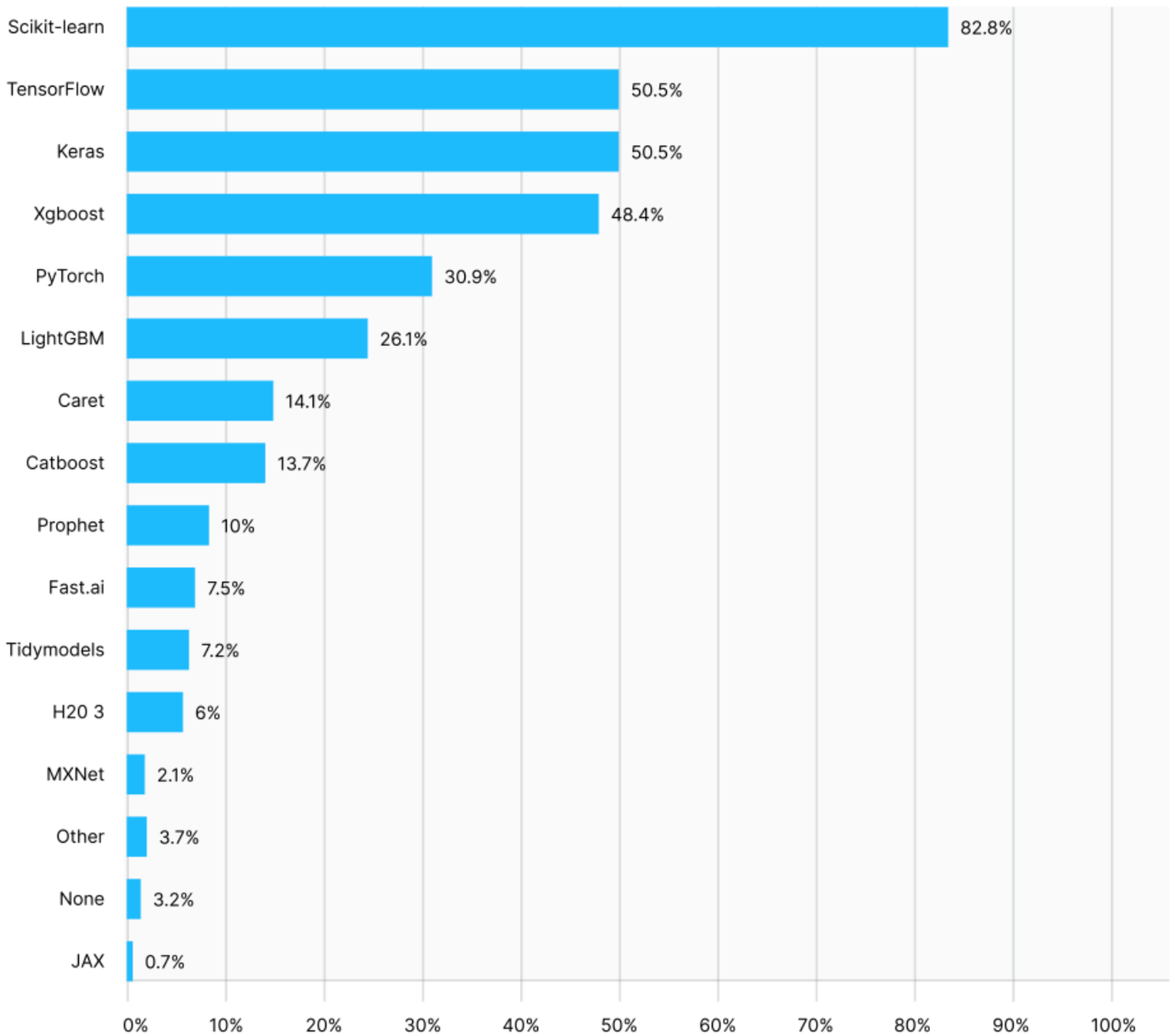
IEEE SPECTRUM

Python-based tools continue to dominate the machine learning frameworks. Scikit-learn, a swiss army knife applicable to most projects, is the top with four in five data scientists using it. TensorFlow and Keras, notably used in combination for deep learning, were each selected on about half of the data scientist surveys. Gradient boosting library xgboost is fourth, with about the same usage as 2019.

The fifth place tool, PyTorch, climbed above 30%, up from about 26% in 2019.

The most popular of the tools added to the survey this year is R-based Tidymodels, reaching over 7 percent.

MACHINE LEARNING FRAMEWORK USAGE



https://www.kaggle.com/kaggle-survey-2020

Course Topics

Part 1: Introduction

Part 2: Computational foundations

Part 3: Tree-based methods

Part 4: Model evaluation

Part 5: Dimensionality reduction and unsupervised learning

Part 6: Bayesian learning

Part 7: Class project presentations

Part 1: Introduction

- Week 01: L01 - Course overview, introduction to machine learning

- Week 02: L02 - Introduction to Supervised Learning and k-Nearest Neighbors Classifiers

Part 2: Computational foundations

- Week 03: L03 - Using Python

- Week 03: L04 - Introduction to Python's scientific computing stack

- Week 04: L05 - Data preprocessing and machine learning with scikit-learn

Reading Recommendations

- Raschka and Mirjalili: Python Machine Learning, 3rd ed., Ch 1

- Chapter 1: Introduction to Machine Learning and Deep Learning, https://sebastianraschka.com/blog/2020/intro-to-dl-ch01.html

- Lecture Notes for Lecture 01 (see Canvas)

- How to avoid machine learning pitfalls: a guide for academic researchers, https://arxiv.org/abs/2108.02497