Lecture 01

What is Machine Learning?
An Overview.

STAT 451: Intro to Machine Learning, Fall 2020

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat451-fs2020/

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

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

About this Course

For details -> http://stat.wisc.edu/~sraschka/teaching/stat451-fs2020/

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

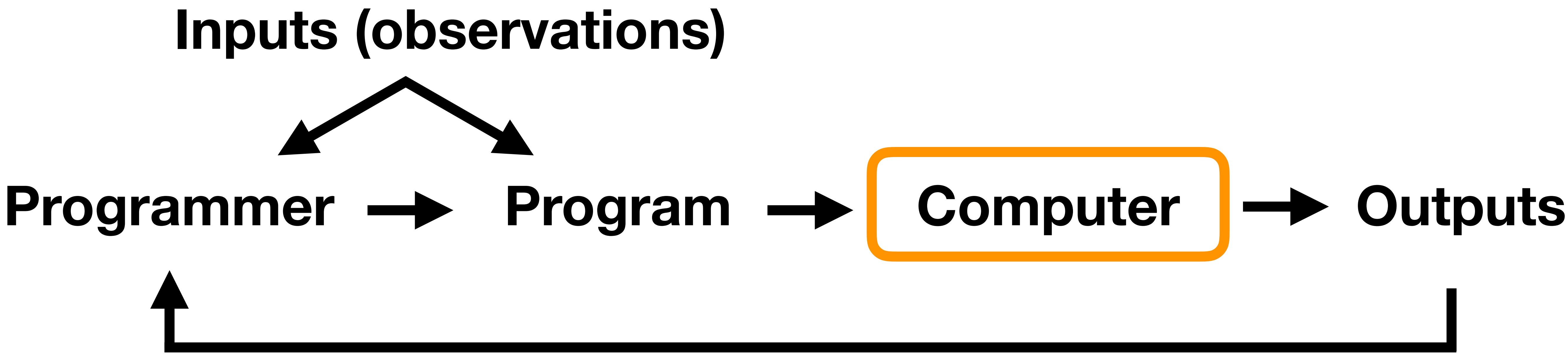
"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

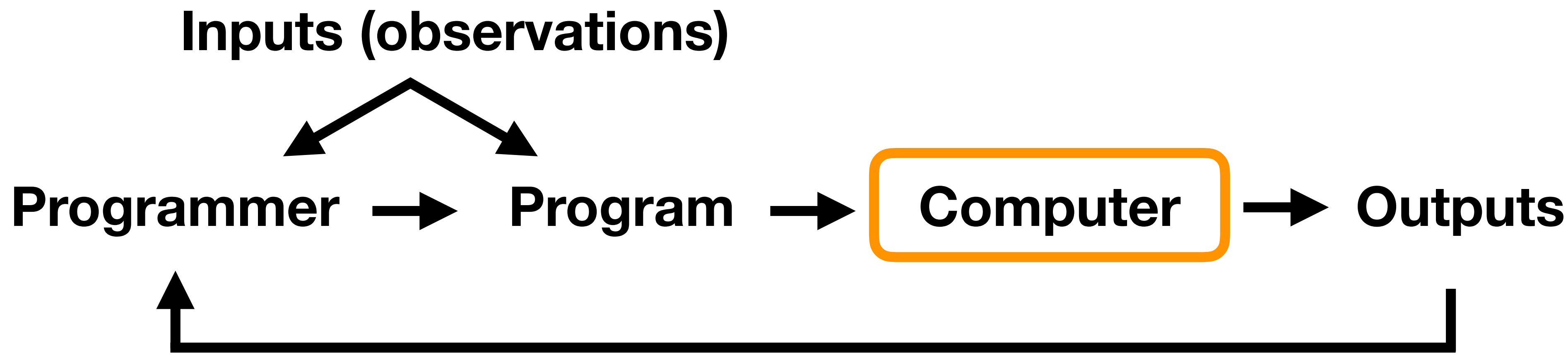
— Arthur L. Samuel, AI pioneer, 1959

(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

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

Tom M Mitchell et al. "Machine learning. 1997". In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

Handwriting Recognition Example:

- Task T: ?
- Performance measure P : ?
- Training experience E: ?

Some Applications of Machine Learning:

-
-
-
-
-

Lecture 1 Overview

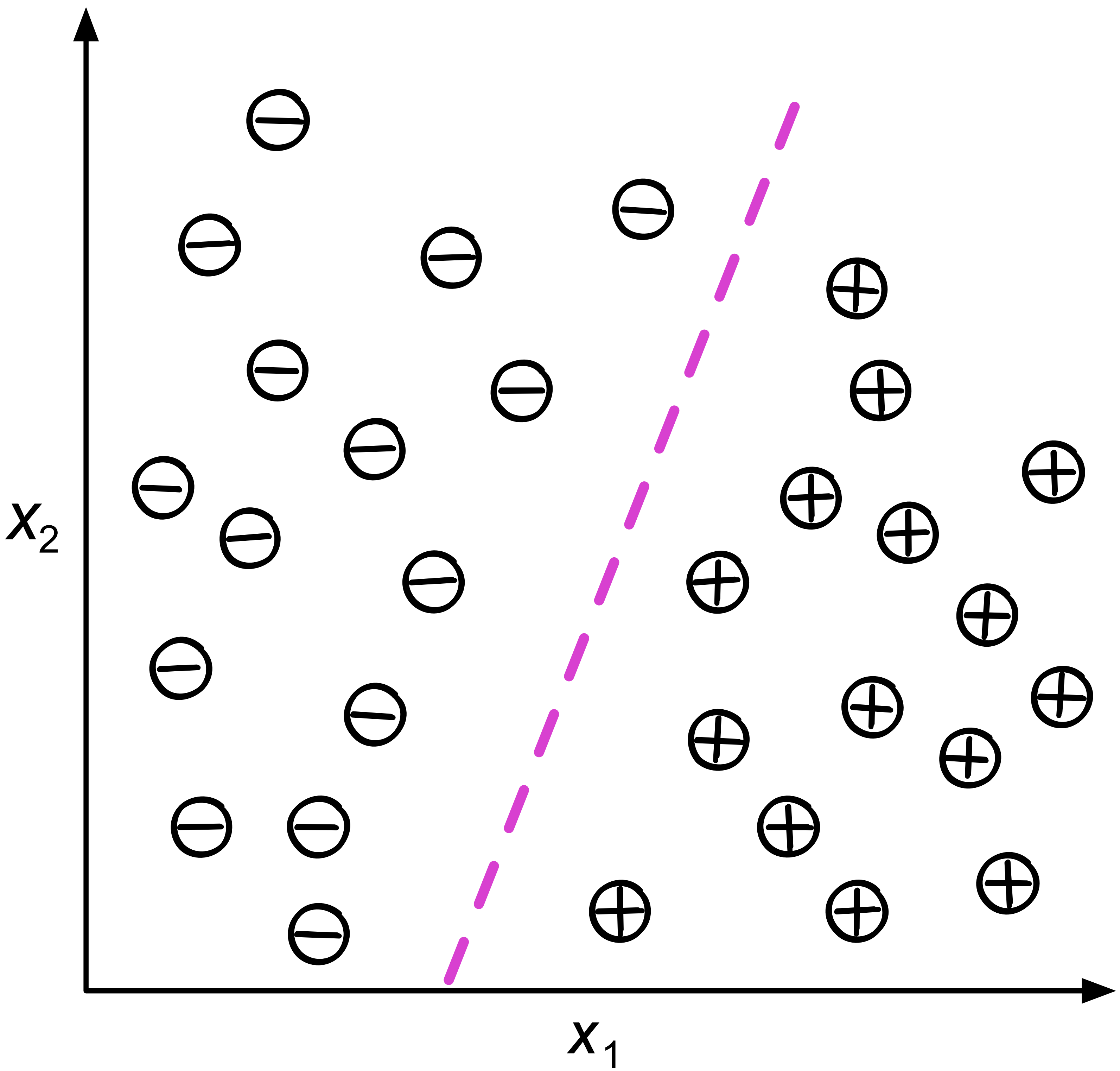
1. About this course
2. What is machine learning
3. **Categories of machine learning**
4. Notation
5. Approaching a machine learning application
6. Different machine learning approaches and motivations

Categories of Machine Learning

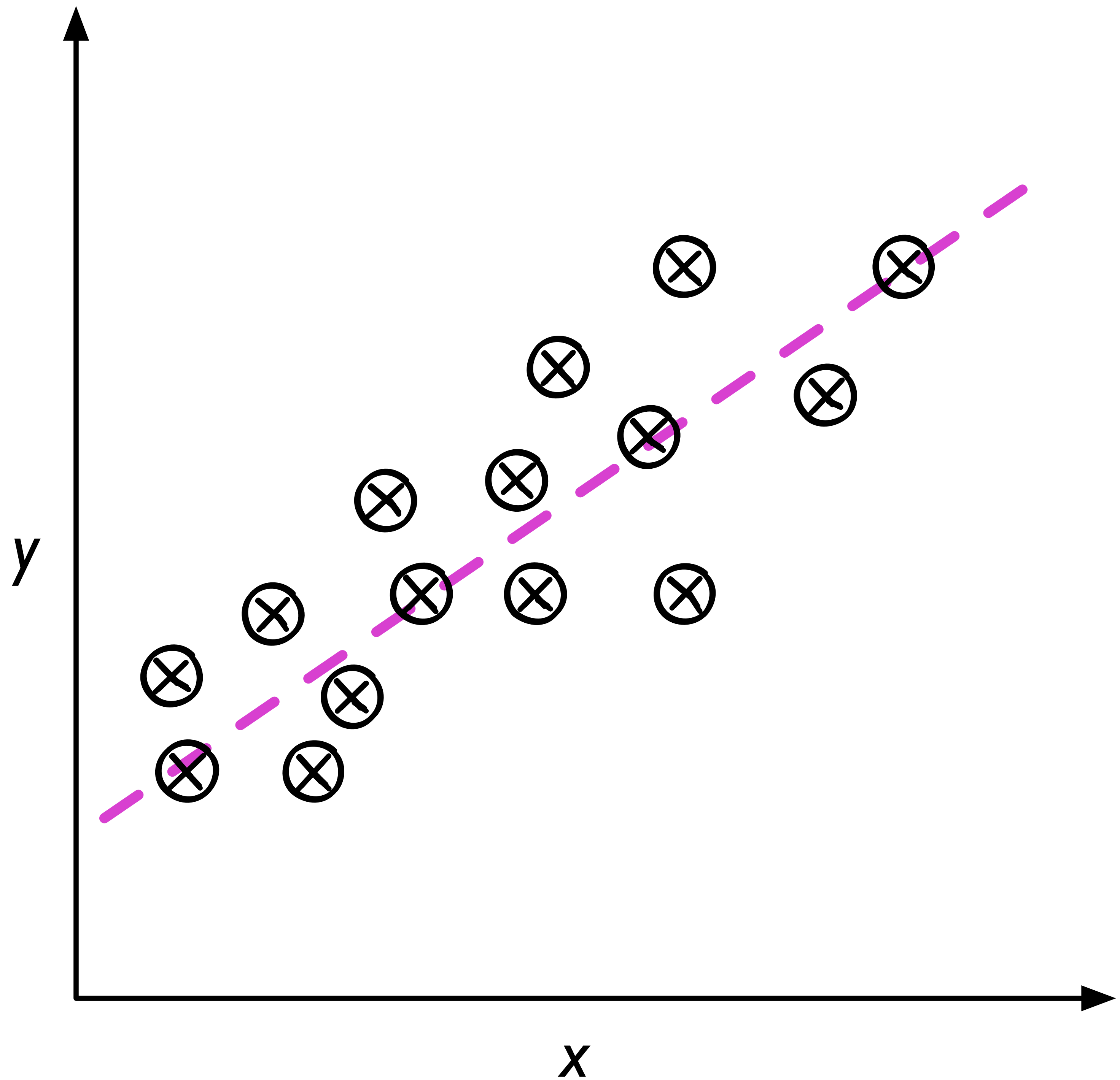
Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

Supervised Learning: Classification



Supervised Learning: Regression



Categories of Machine Learning

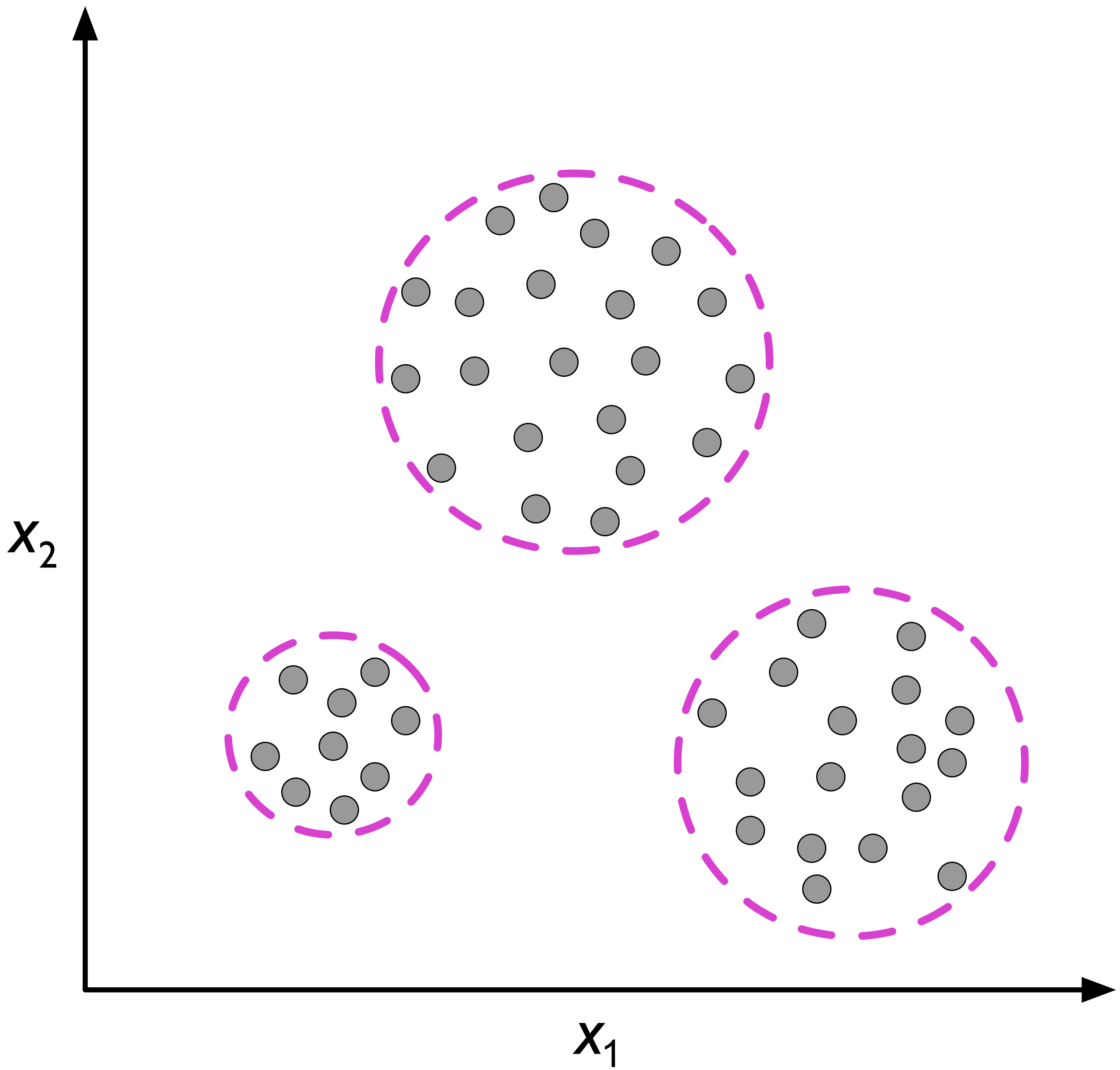
Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

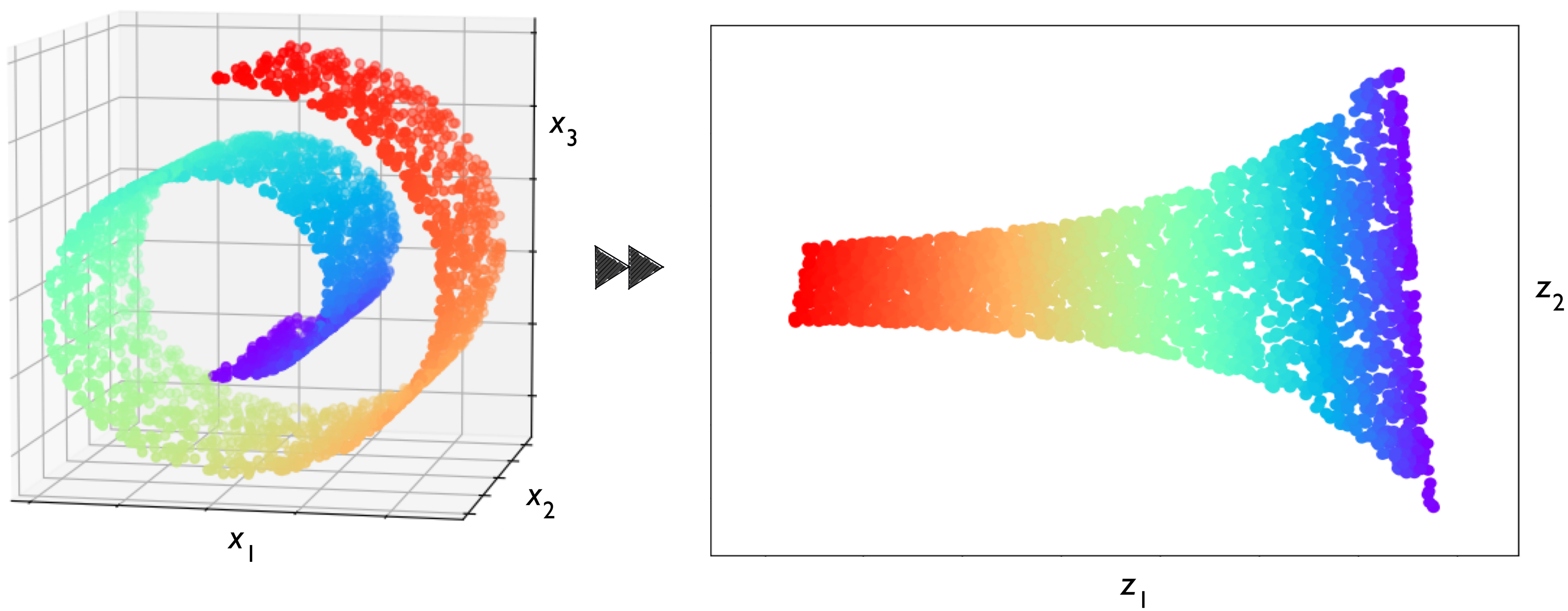
Unsupervised Learning

- 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

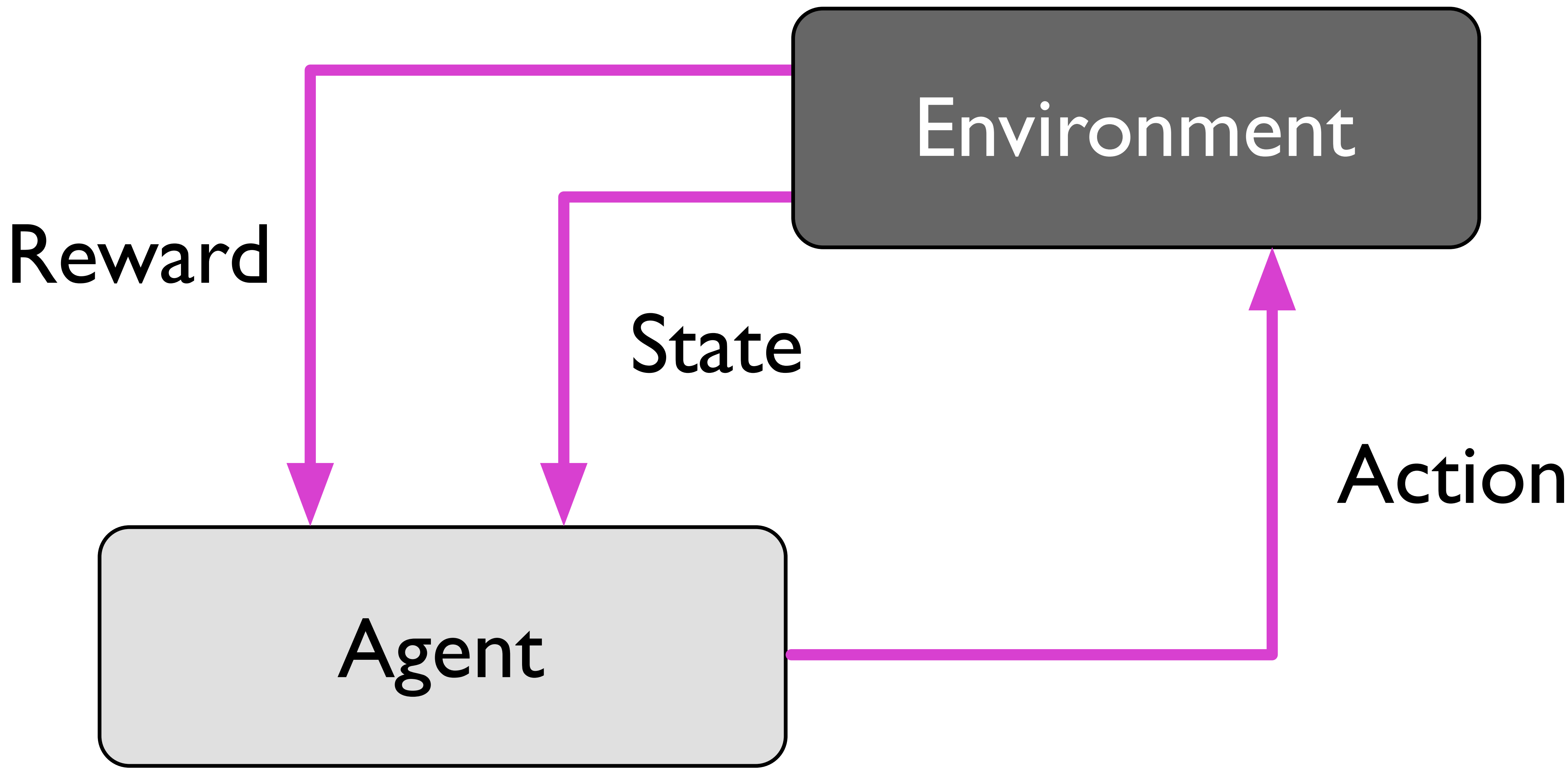
Unsupervised Learning

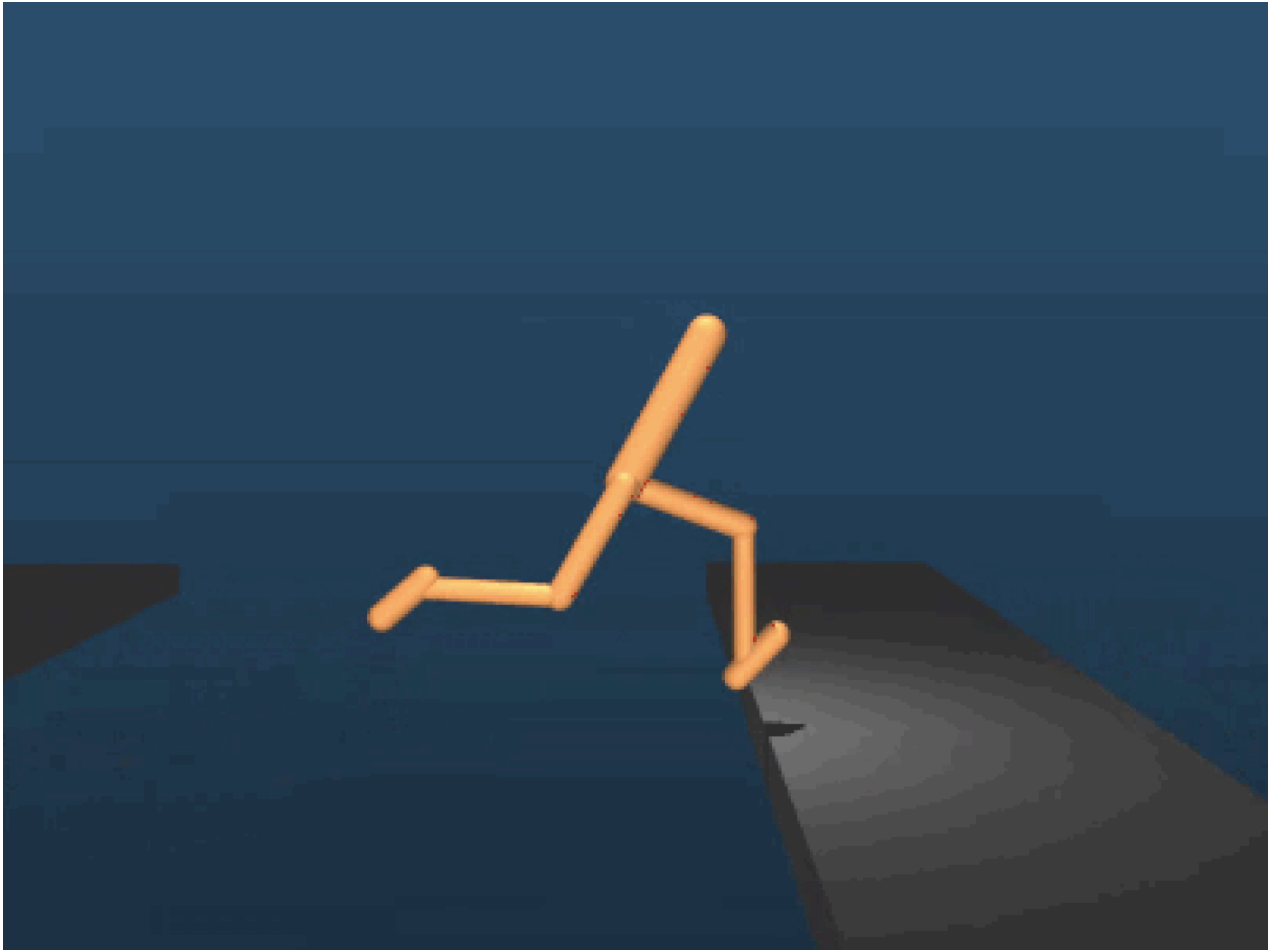
- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning

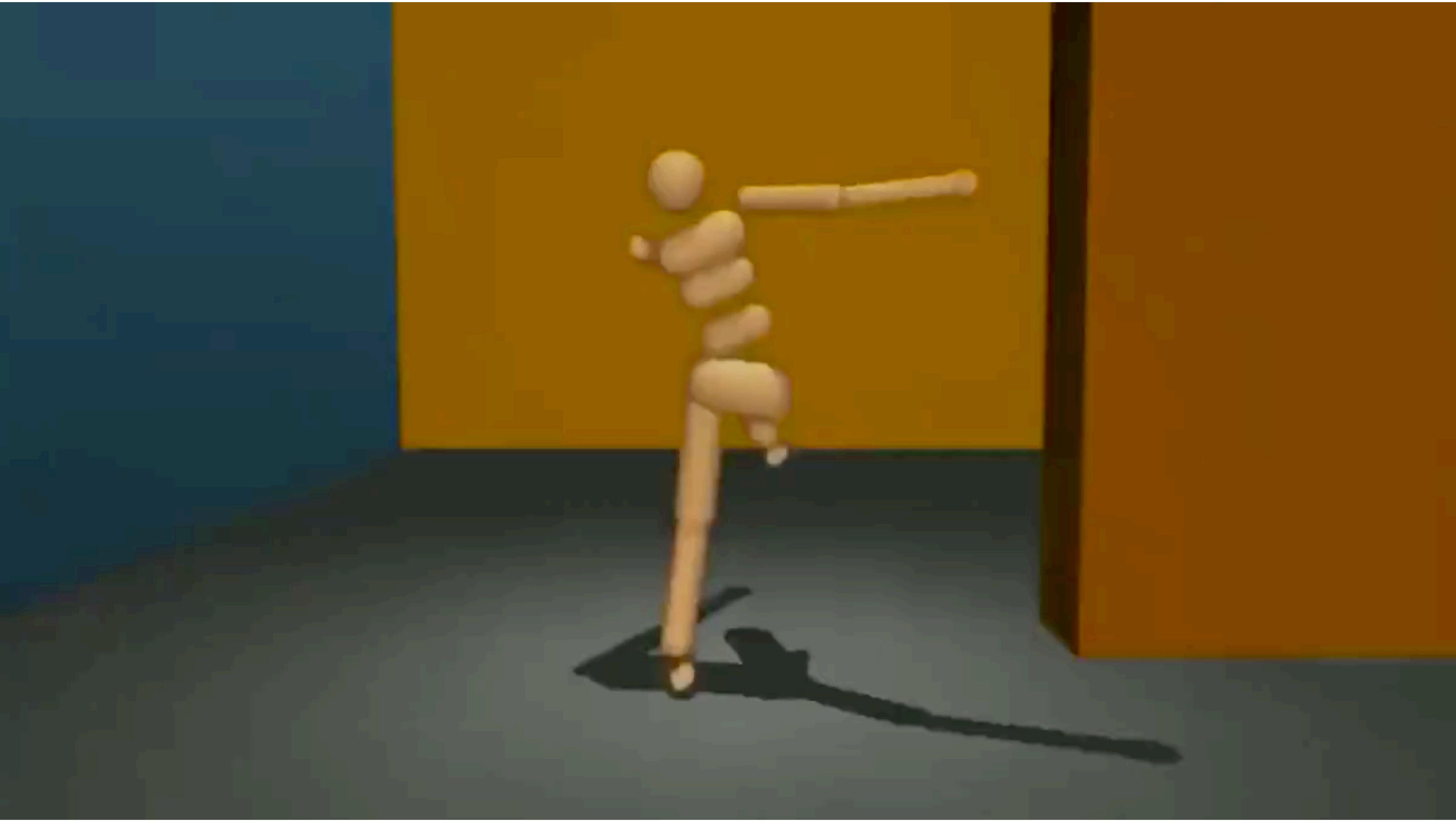
- Decision process
- Reward system
- Learn series of actions

Reinforcement Learning





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

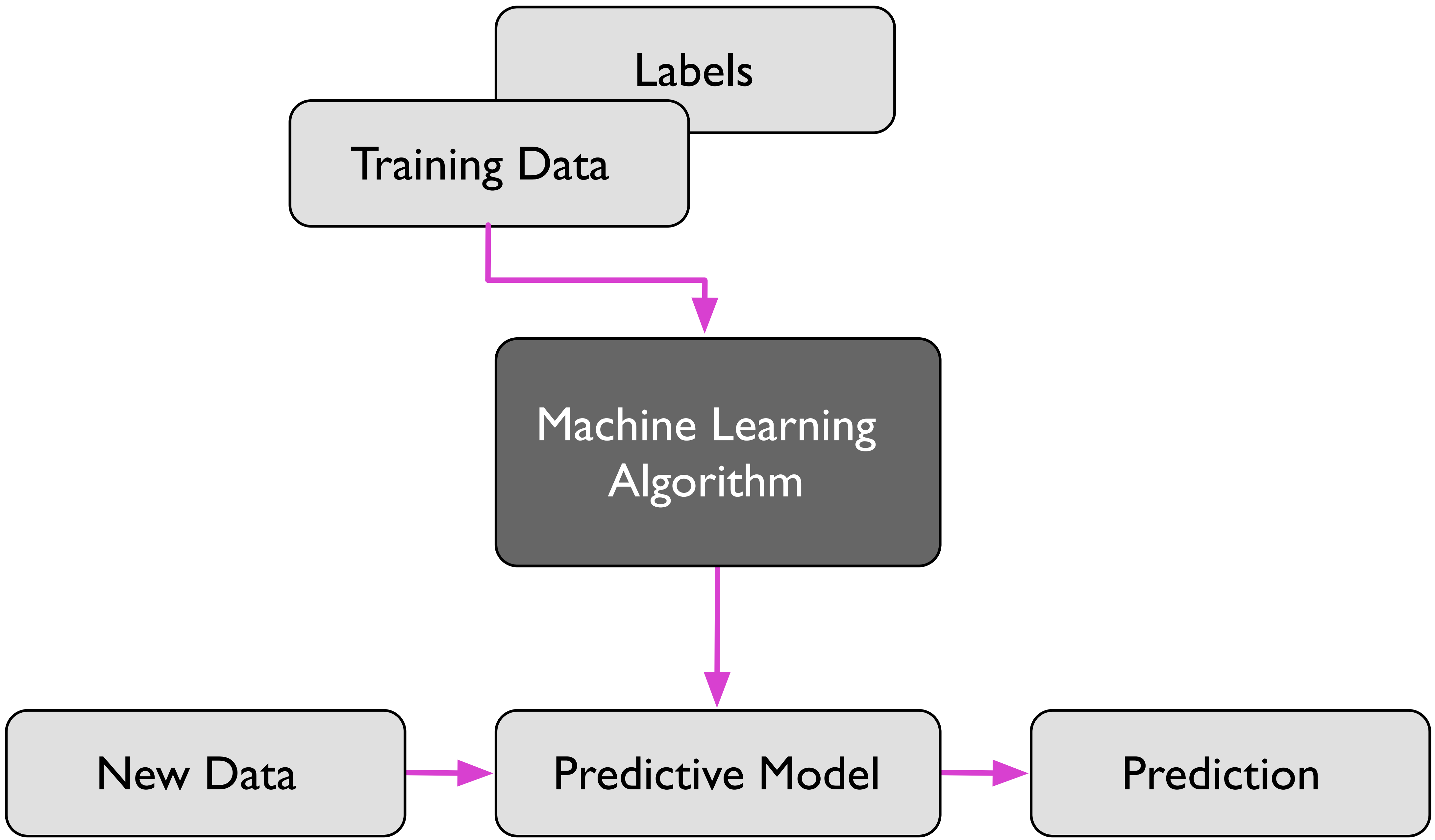


https://video.twimg.com/ext_tw_video/1111683489890332672/pu/vid/1200x674/WqUJEhUETw0M0gCl.mp4?tag=8

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

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

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

Classification | **Regression**

$h : \mathbb{R}^m \rightarrow$ __ | $h : \mathbb{R}^m \rightarrow$ __

Data Representation

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

Feature vector

Data Representation

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

Feature vector

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

D___n m_________

Data Representation

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

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

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

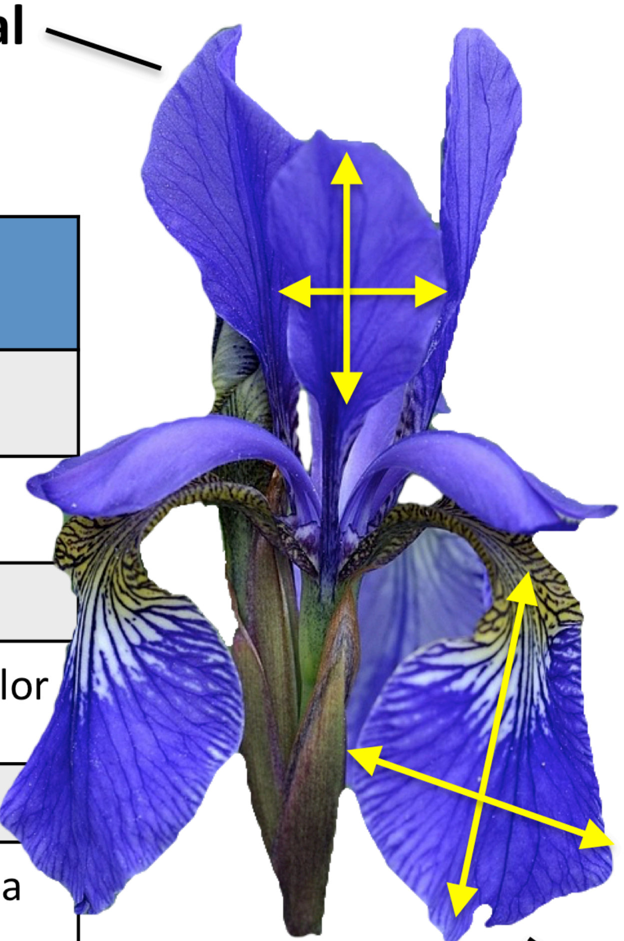
Feature vector

Data Representation

$m =$ ____

$n =$ ____

	Sepal length	Sepal width	Petal length	Petal width	
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



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)

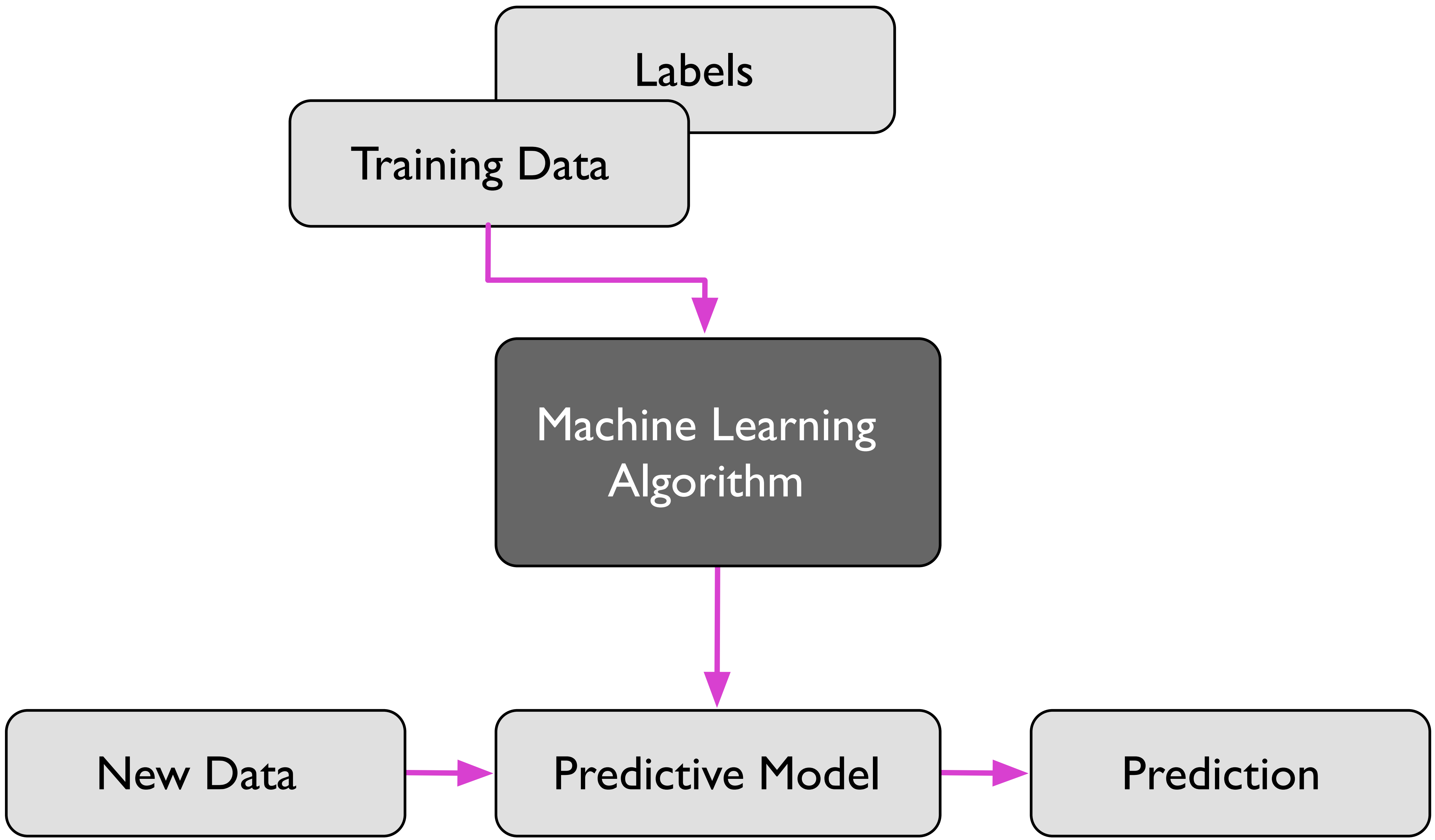
Classes of Machine Learning Algorithms

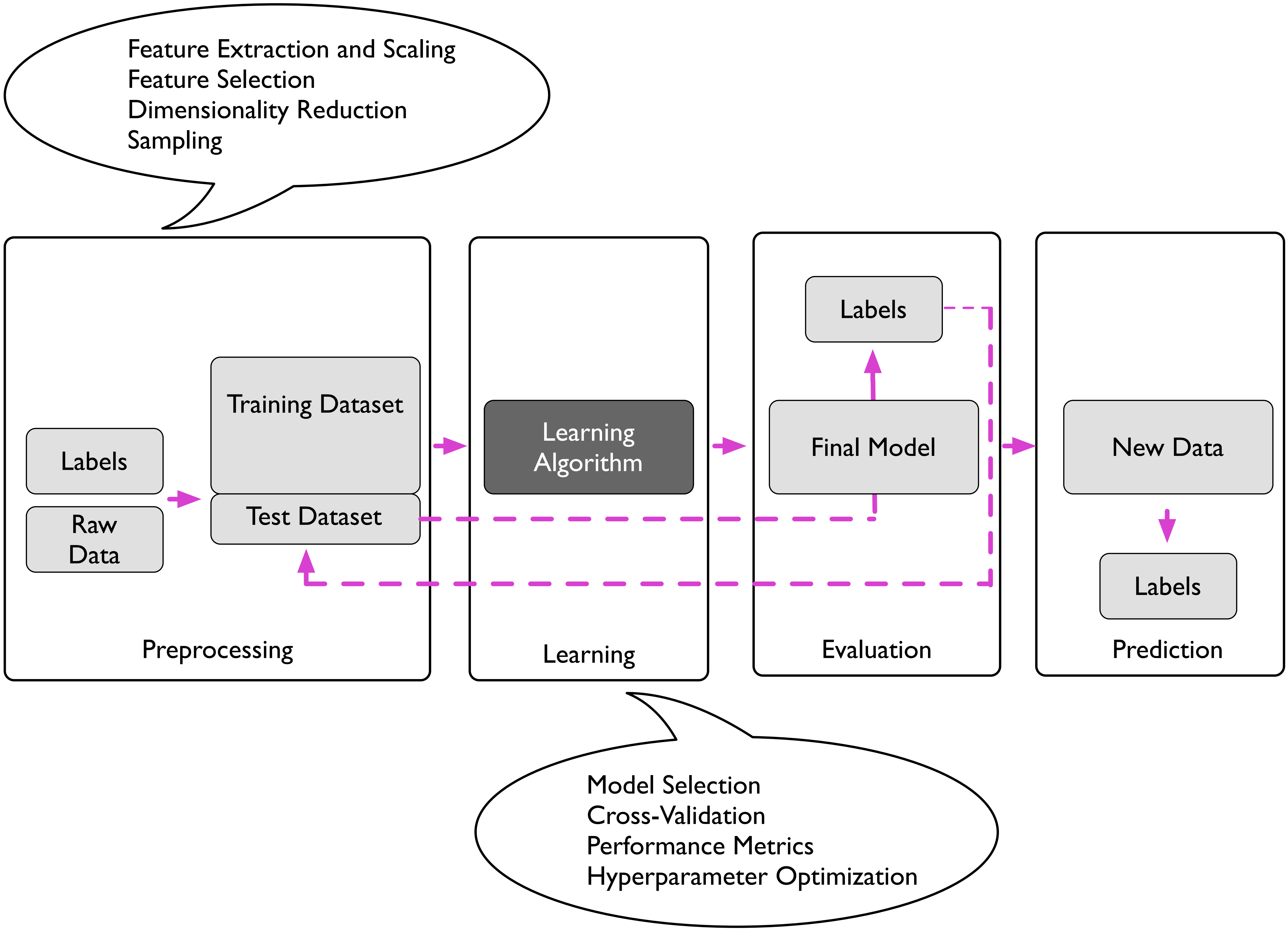
- Generalized linear models (e.g.,
- Support vector machines (e.g.,
- Artificial neural networks (e.g.,
- Tree- or rule-based models (e.g.,
- Graphical models (e.g.,
- Ensembles (e.g.,
- Instance-based learners (e.g.,

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

Supervised Learning Workflow -- Overview




Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Preprocessing
Learning Algorithm
Learning
Labels
Final Model
Evaluation
New Data
Labels
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

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.

Objective Functions

- Maximize the posterior probabilities (e.g., naive Bayes)
- Maximize a fitness function (genetic programming)
- Maximize the total reward/value function (reinforcement learning)
- Maximize information gain/minimize child node impurities (CART decision tree classification)
- Minimize a mean squared error cost (or loss) function (CART, decision tree regression, linear regression, adaptive linear neurons, ...)
- Maximize log-likelihood or minimize cross-entropy loss (or cost) function
- Minimize hinge loss (support vector machine)

Optimization Methods for Different Learning Algorithms

- Combinatorial search, greedy search (e.g., decision trees)
- Unconstrained convex optimization (e.g.,
- Constrained convex optimization (e.g.,
- Nonconvex optimization, here: using backpropagation, chain rule, reverse autodiff. (e.g.,
- Constrained nonconvex optimization (e.g.,

Evaluation -- Misclassification Error

$$L(\hat{y}, y) = \begin{cases} 0 & \textbf{if } \hat{y} = y \\ 1 & \textbf{if } \hat{y} \neq y \end{cases}$$

$$ERR_{\mathscr{D}^{\textbf{test}}} = \frac{1}{n} \sum_{i=1}^{n} L\Big(\hat{y}^{[i]}, y^{[i]}\Big)$$

ML Terminology (Part 2)

- **Loss function:** Often used synonymously with cost function; sometimes also called error function. In some contexts the loss for a single data point, whereas the cost function refers to the overall (average or summed) loss over the entire dataset. Sometimes also called empirical risk.

Other Metrics in Future Lectures

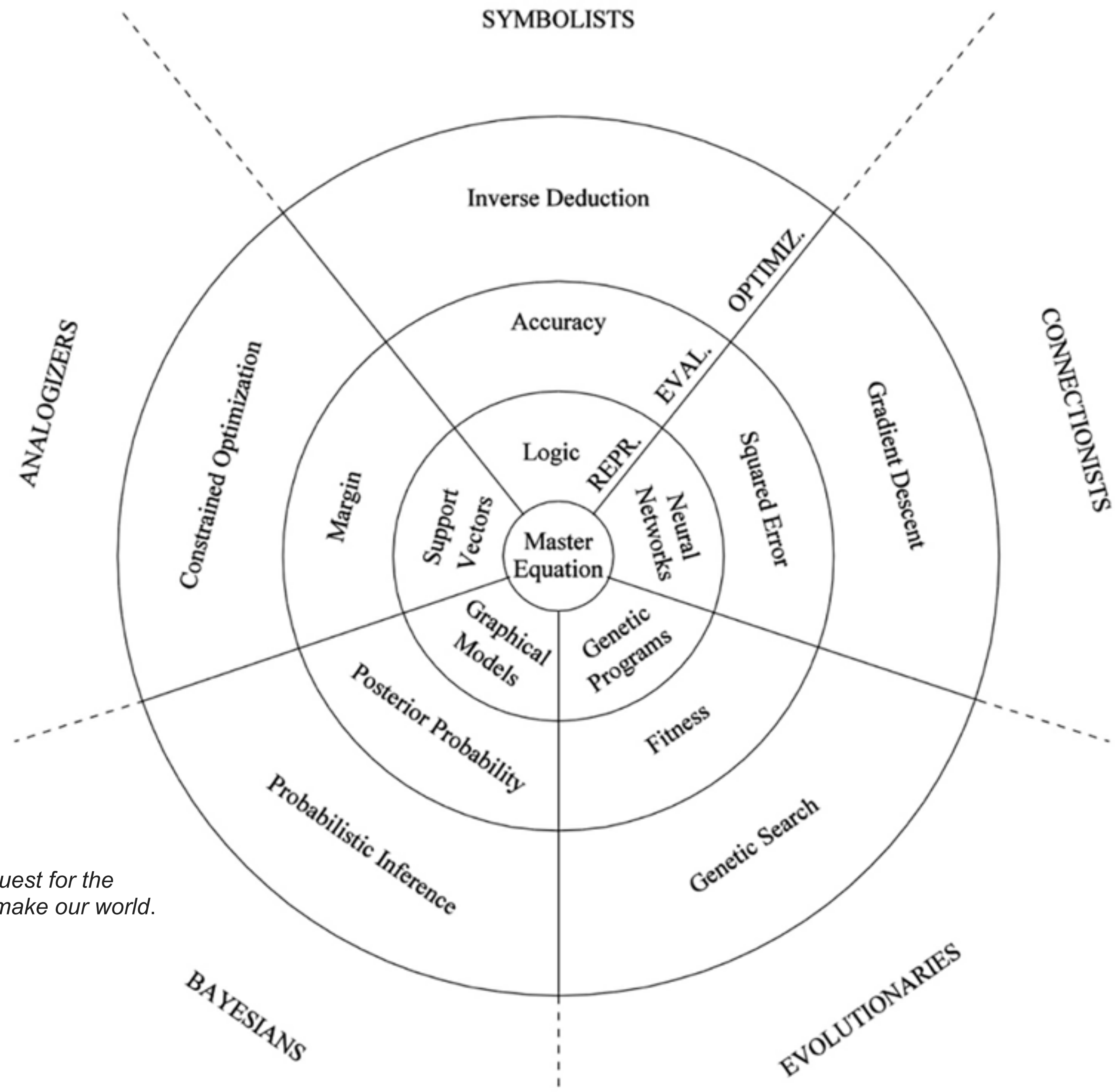
- Accuracy (1-Error)
- ROC AUC
- Precision
- Recall
- (Cross) Entropy
- Likelihood
- Squared Error/MSE
- L-norms
- Utility
- Fitness
- ...

But more on other metrics in future lectures.

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**

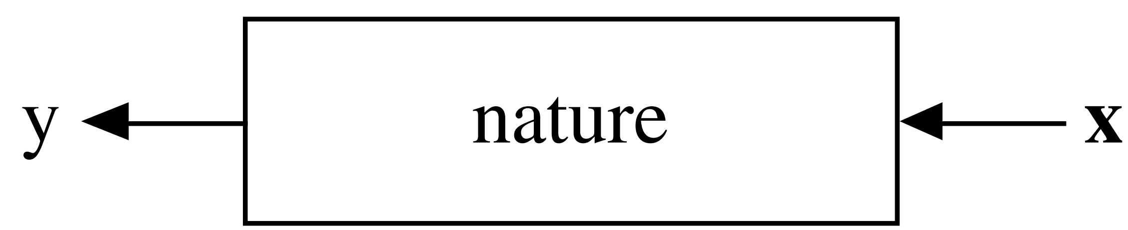
Pedro Domingos's 5 Tribes of Machine Learning



Source: Domingos, Pedro.
The master algorithm: How the quest for the ultimate learning machine will remake our world.
Basic Books, 2015.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author).
" *Statistical science* 16.3 (2001): 199-231.

A



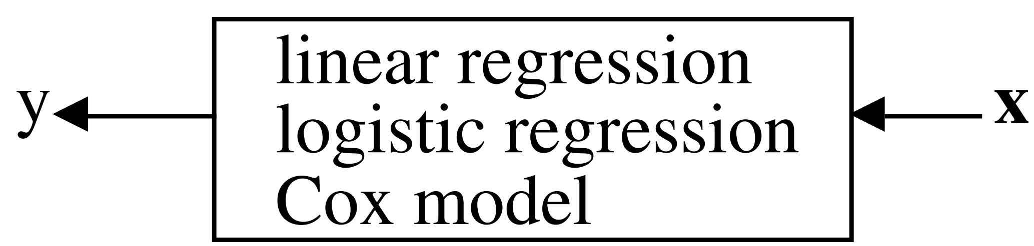
There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author). " *Statistical science* 16.3 (2001): 199-231.

B

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

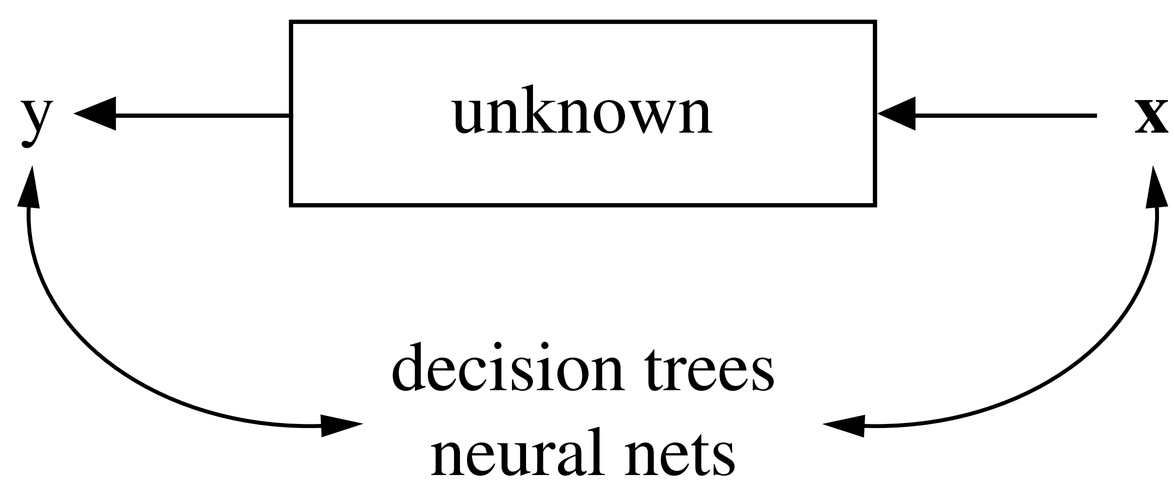


Model validation. Yes–no using goodness-of-fit tests and residual examination.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author). " *Statistical science* 16.3 (2001): 199-231.

C

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.

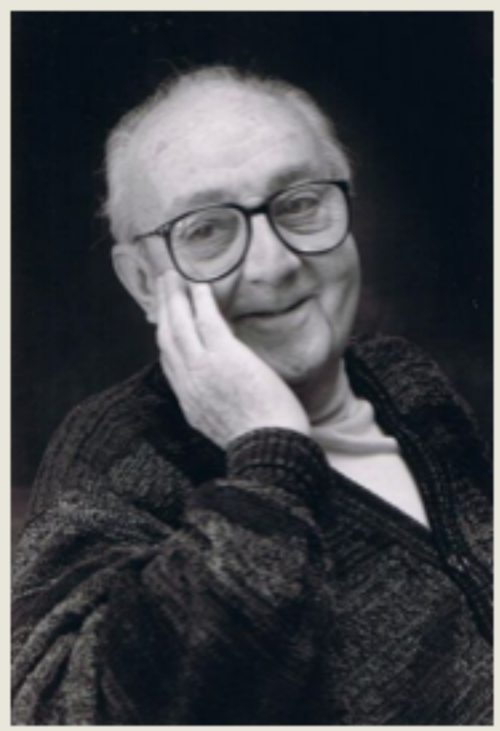

0 1 2 cm

Evolved antenna (Source: https://en.wikipedia.org/wiki/Evolved_antenna) via evolutionary algorithms; used on a 2006 NASA spacecraft.

Black Boxes vs Interpretability

Black Boxes vs Interpretability

GEORGE BOX, 1919 –2013



"All models are wrong but some are useful."

George Box, professor emeritus of Statistics and of Industrial & Systems Engineering, died on Thursday, March 28, 2013, at the age of 93. Founder of the Department of Statistics...

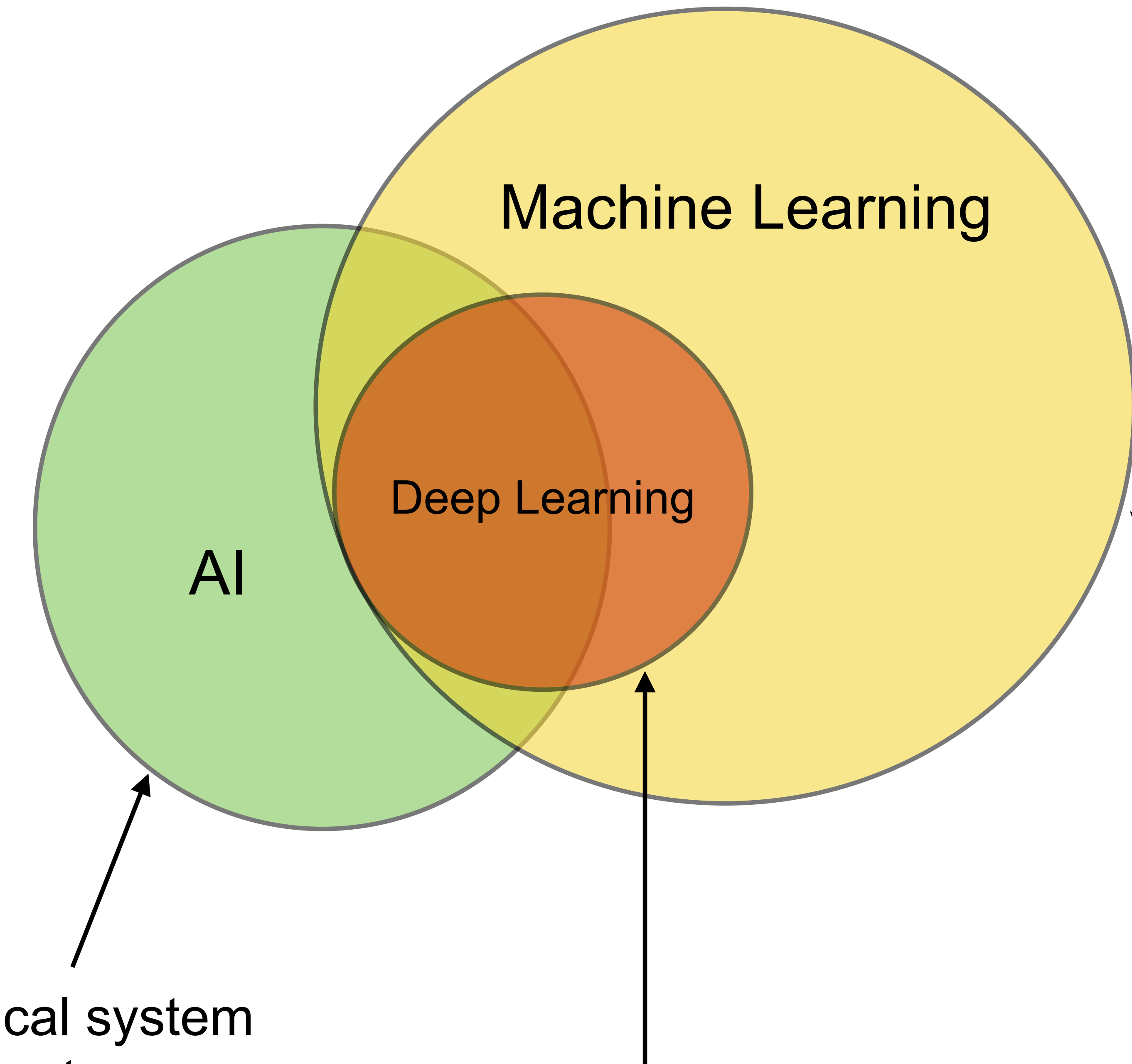
Different Motivations for Studying Machine Learning

- Engineers:

- Mathematicians, computer scientists, and statisticians:

- Neuroscientists:

Machine Learning, AI, and Deep Learning



Algorithms that learn models/representations/ rules automatically from data/examples

A non-biological system that is intelligent through rules

Algorithms that parameterize multilayer neural networks that then learn representations of data with multiple layers of abstraction

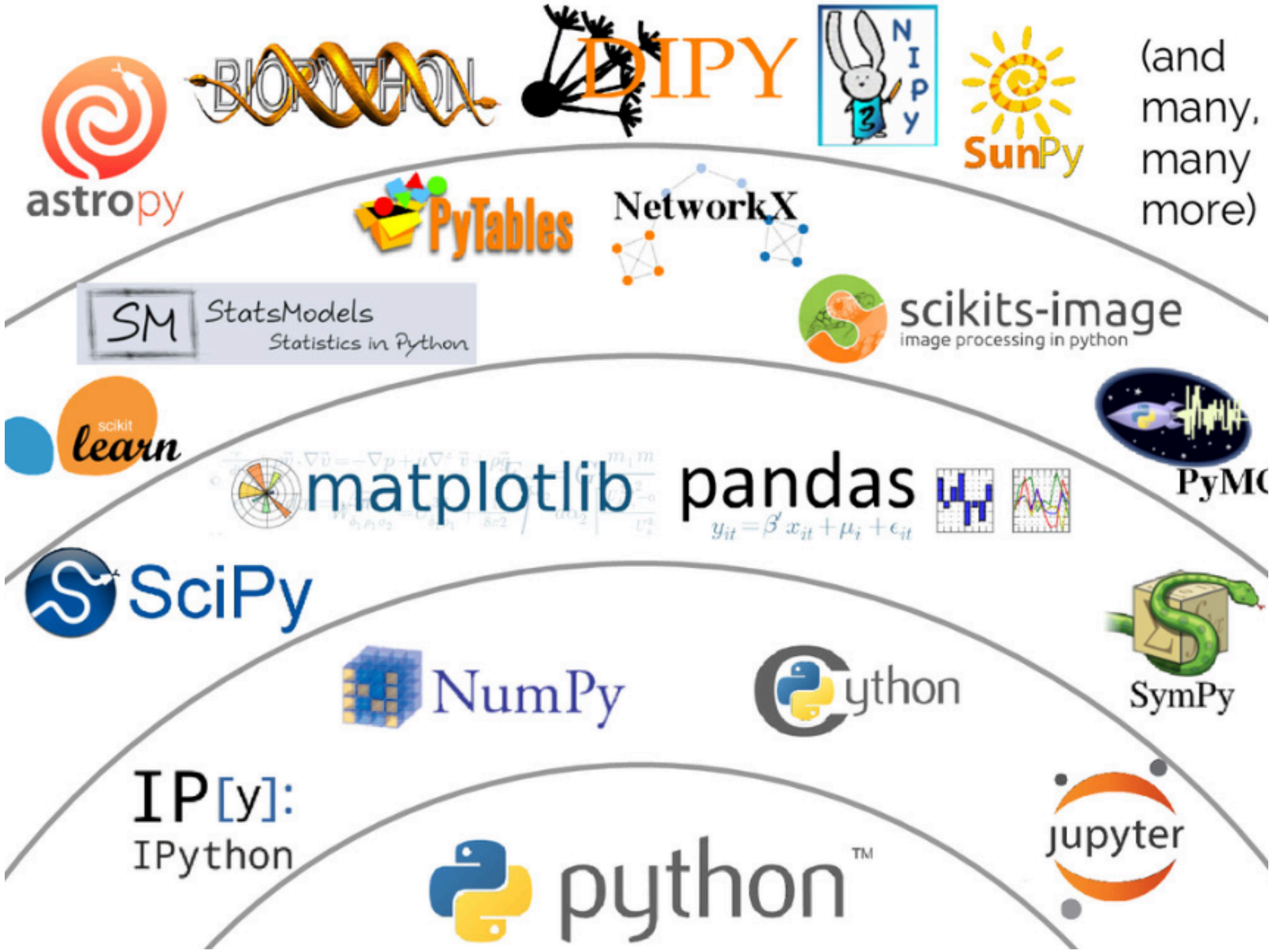


Image by Jake VanderPlas; Source:
https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8)

Spam

https://en.wikipedia.org/wiki/Spam_(food)

"It has become the subject of a number of appearances in pop culture, notably a Monty Python sketch which repeated the name many times, leading to its name being borrowed for unsolicited electronic messages, especially email."

Spam

https://en.wikipedia.org/wiki/Spam_(food)

Monty Python

The Pythons in 1969:
Back row: Chapman, Idle, Gilliam
Front row: Jones, Cleese, Palin

Medium	Television · film · theatre · audio recordings · literature
Nationality	British[1]
Years active	1969–1983, 1989, 1998–1999, 2002, 2013–2014

https://en.wikipedia.org/wiki/Monty_Python

Spam

https://en.wikipedia.org/wiki/Spam_(food)

https://en.wikipedia.org/wiki/Monty_Python

"Python's name is derived from the British comedy group Monty Python, whom Python creator Guido van Rossum enjoyed while developing the language. "

https://en.wikipedia.org/wiki/Python_(programming_language)

ML Terminology (Part 3)

- **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

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 Assignments

- Raschka and Mirjalili: Python Machine Learning, 3rd ed., Ch 1
- Elements of Statistical Learning, Ch 01 (https://web.stanford.edu/~hastie/ElemStatLearn/)
- Optional: Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". *Statistical science* 16.3 (2001): 199-231. https://projecteuclid.org/euclid.ss/1009213726