Python Machine Learning In Biology:

Machine Learning Overview

Nichole Bennett

Course Information

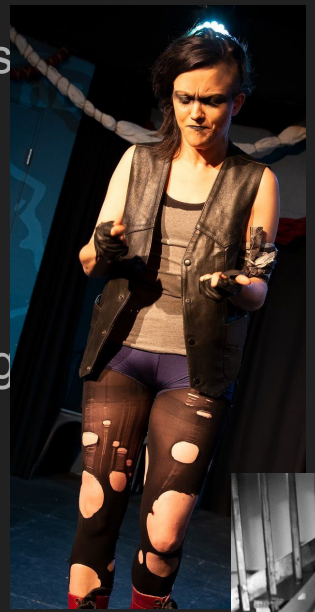
About Me

Used to research climate change impacts on butterflies (ecology, evolution, behavior, some genomics)

Then, I taught adults data science with General Assembly and taught kids coding through various programs.

Now I research science communication.

In my spare time, I do lots of theater.



About You

Find a partner, find out the following about them:

- Name
- Where they are from (hometown and current university)
- What they research
- Why they are here
- One fun fact

Then, we'll share.

Course Overview (subject to change)

Day 1: Intro to Machine Learning, Python Foundations, Pandas

Day 2: Data Viz, Logistic Regression, KNN, Bias-Variance Tradeoff

Day 3: Morning Break, Data Preprocessing, SVM

Day 4: Hyperparameter Tuning, Evaluation Metrics, Decision Trees, Ensemble Learners, Dimensionality Reduction, Feature Selection

Day 5: Clustering, Neural Networks

Anything missing on here that you were hoping for?

Course expectations

Wide and Shallow.

You'll get an overview of the tools available and be able to pick the right ones for the job at hand. (You won't be expert on any of them).

My goals are that you'll be able to match a task to a problem, know the pros and cons of several algorithms, be able to run them in Python, and be able to evaluate and interpret their output.

You'll also get a lot of "best practices" advice. We'll talk about how to preprocess your data, how to tune your models, and how to evaluate them.

How this course will go

Some theory/lecture

We'll focus on getting an intuition for how the algorithms work and being able to apply machine learning.

Code-alongs

- You can do these as you please (watch, code-along, execute-along)

Independent work (means independent from top-down instruction)

The datasets we are using and why.

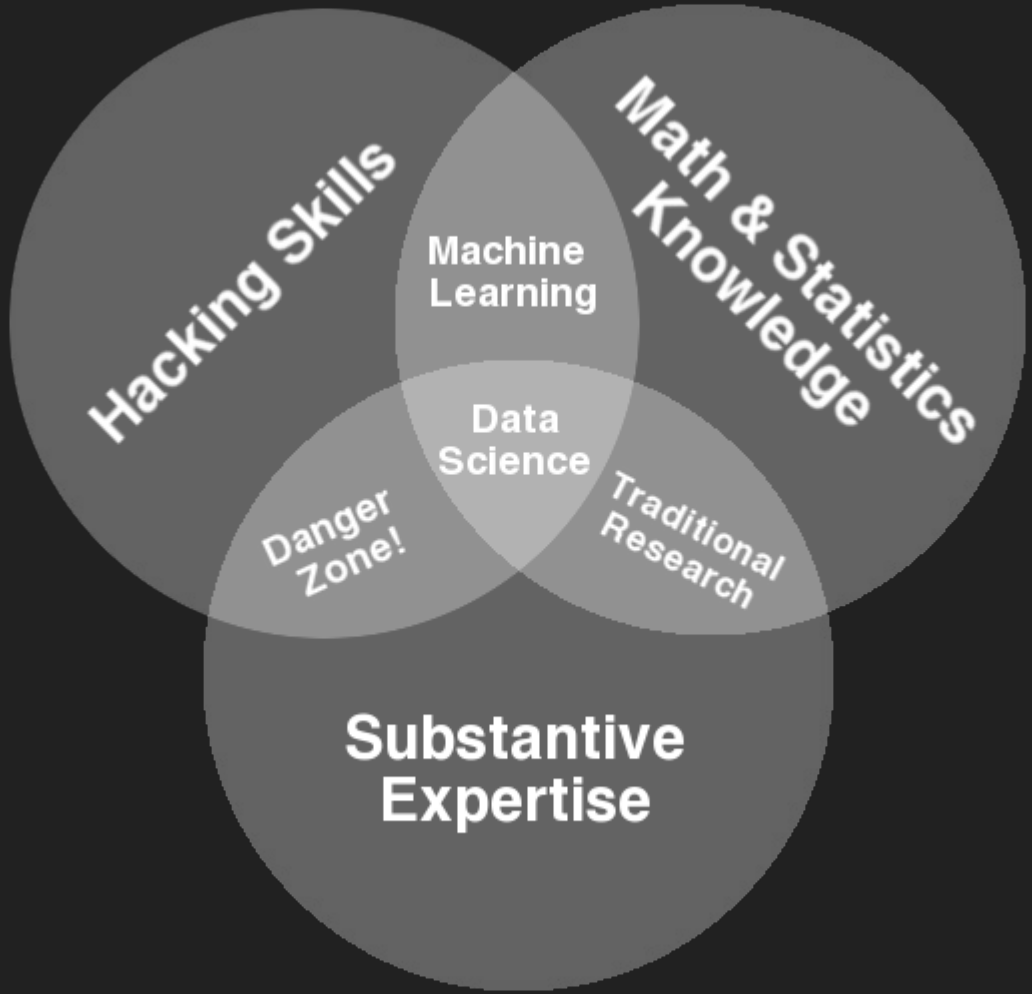
My thoughts on passive learning. Rant, rant, rant.

Prerequisites

This should be pretty beginner-friendly.

Feel free to speak out if you don’t understand something--I’ll adjust.

I’ll assume a little knowledge of statistics and Python, but we can review those as needed.



Install Help Time!

We won't use these until later, so if you don't have them, let's get those downloads started.

Mac/Linux Users: Go to your Terminal and type `conda list`

PC Users: Go to your Anaconda Prompt and type `conda list`

(optional) PC Users: Download Git Bash https://git-for-windows.github.io/

Questions About Course In General?

Any Requests?

Introduction to Machine Learning

What is machine learning?
Think-Pair-Share

What is machine learning?

The semi-automated extraction of knowledge from data.

What is machine learning?

The semi-automated extraction of knowledge from data.

Knowledge from data: Starts with a question that might be answerable using data

Automated extraction: A computer provides the insight

Semi-automated: Requires many smart decisions by a human

What is Machine Learning?

Machine Learning is the science of getting computers to learn and act like humans do and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

Instead of humans having to manually derive rules and build models for large data sets, the computer gradually improves the performance of predictive models.

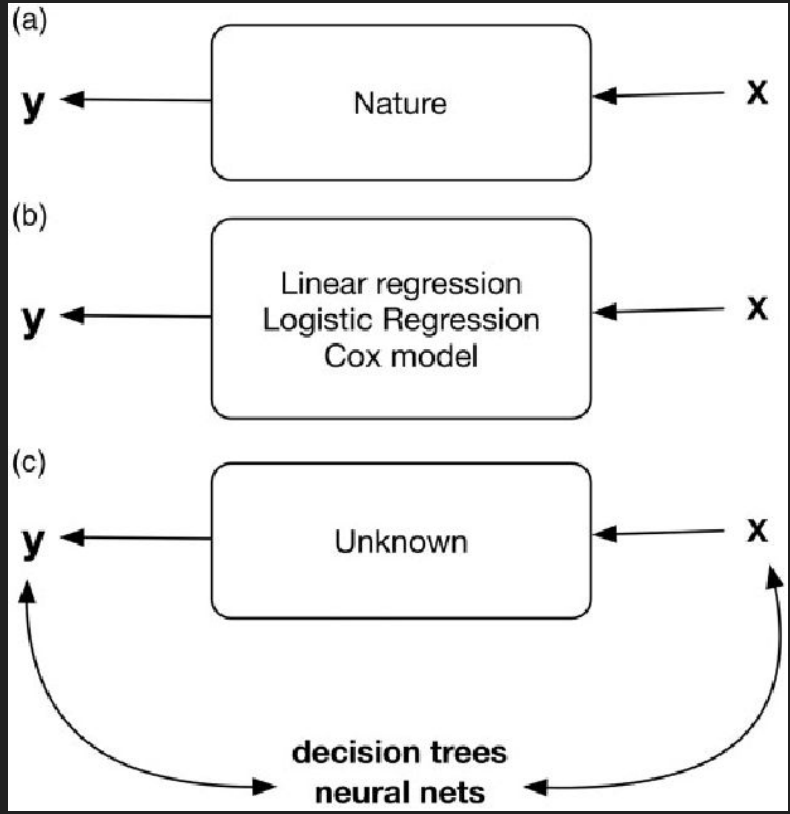
This is great for some problems, but it’s not a cure-all.

The “Two Cultures” of Statistical Modeling

a) who knows? Nature doing Nature things.

b) assume a stochastic data model, want explanations

c) assume complex and unknown, want accurate predictions

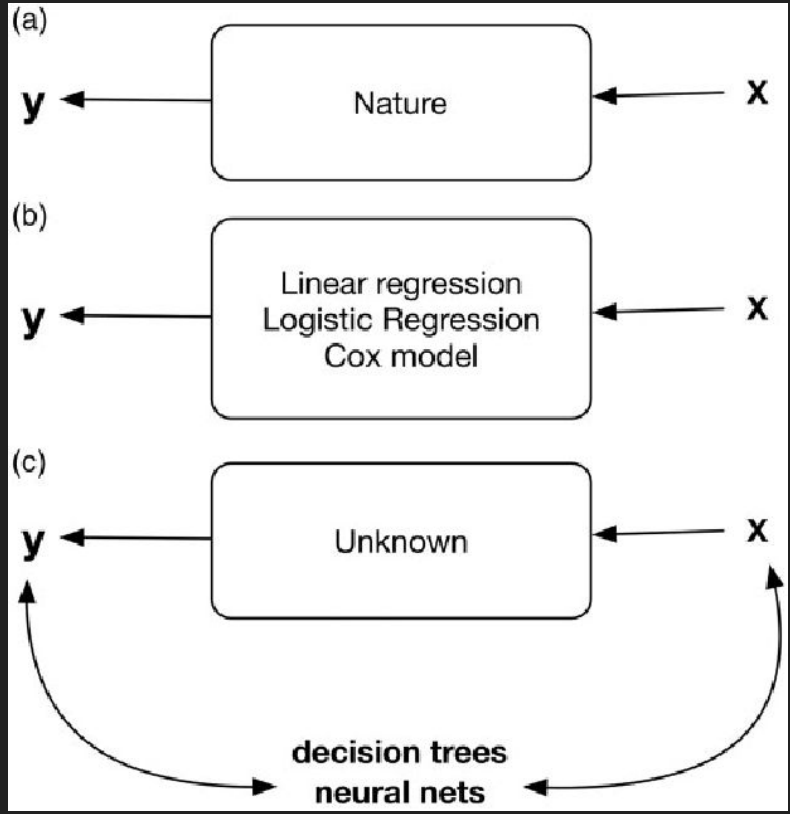


The “Two Cultures” of Statistical Modeling

Two goals for analyzing data:

Extract some information about
how nature is associating the response
variables to the input variables. (a)

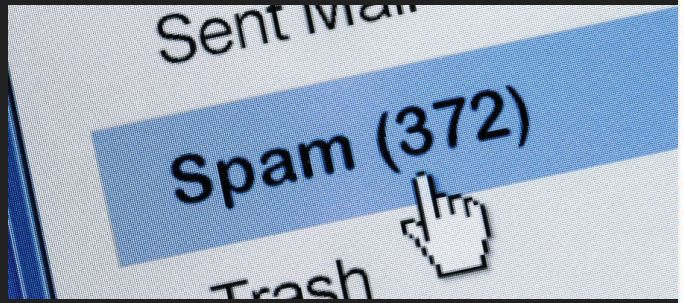
Predict what the responses are going to be to
future input variables. (b)

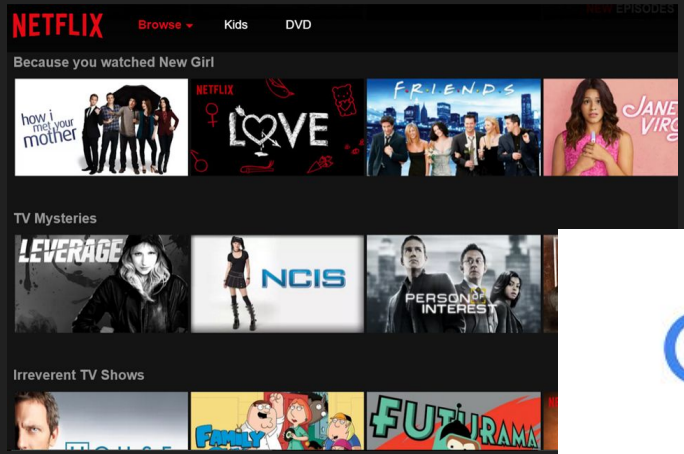


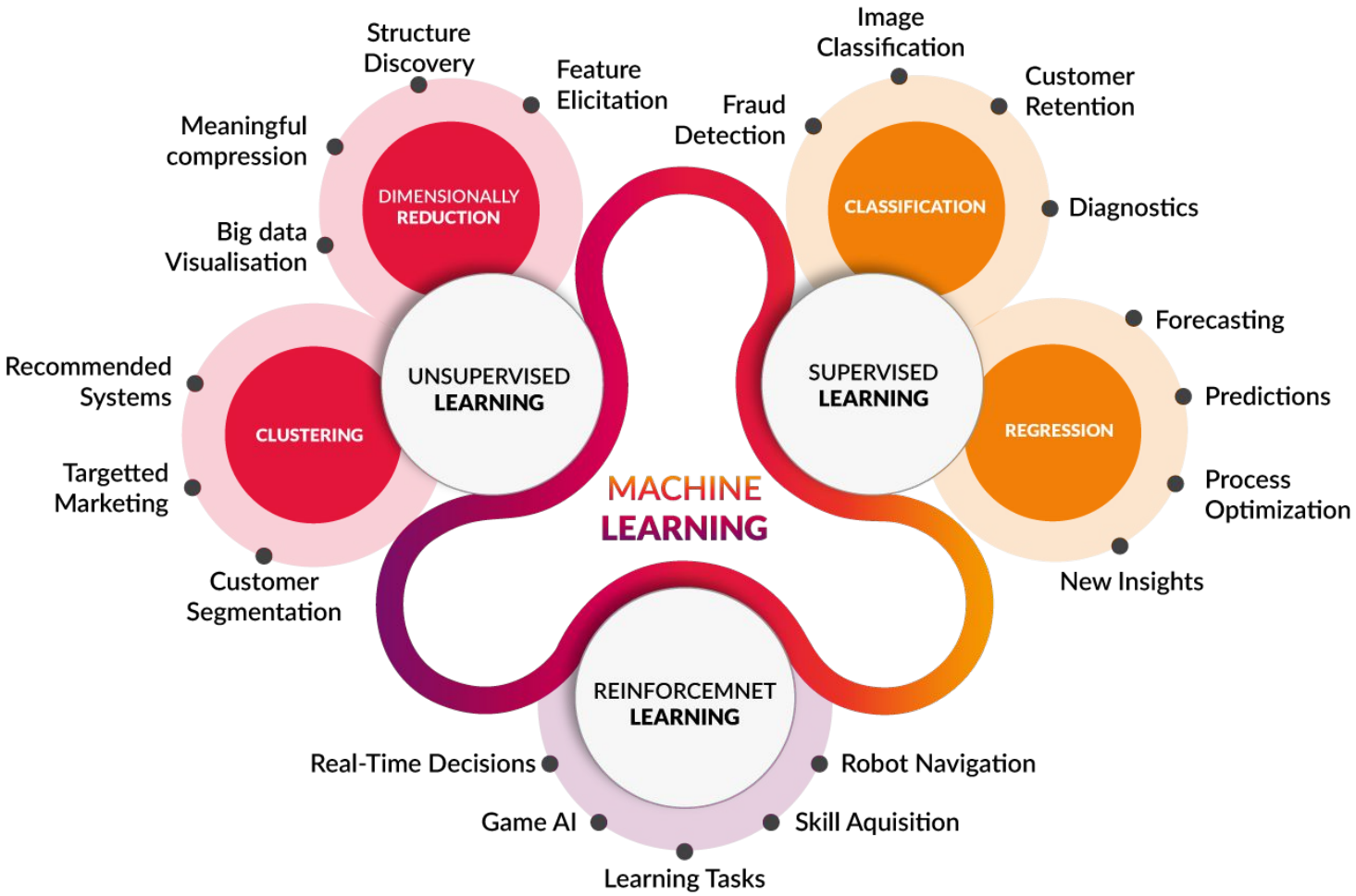
Machine Learning is useful for accurate predictions for unknown data

...not as much for explaining relationships between variables.

Applications of Machine Learning

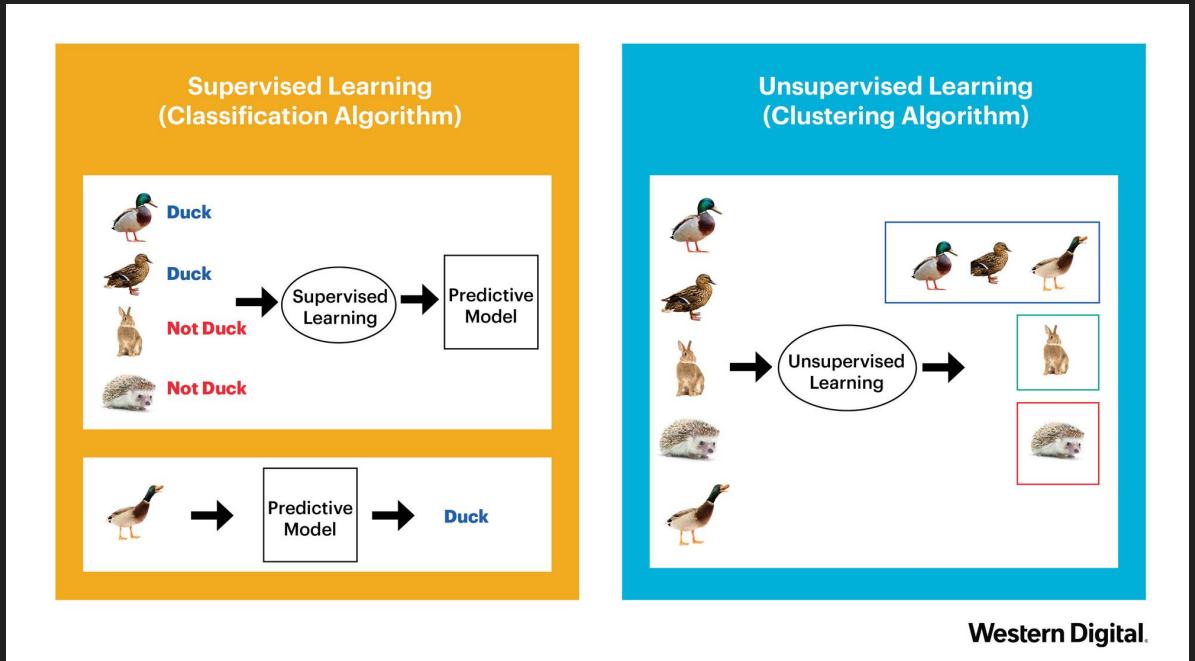





MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Supervised Learning

Learn a model from labeled training data that allows us to make predictions about unseen or future data.



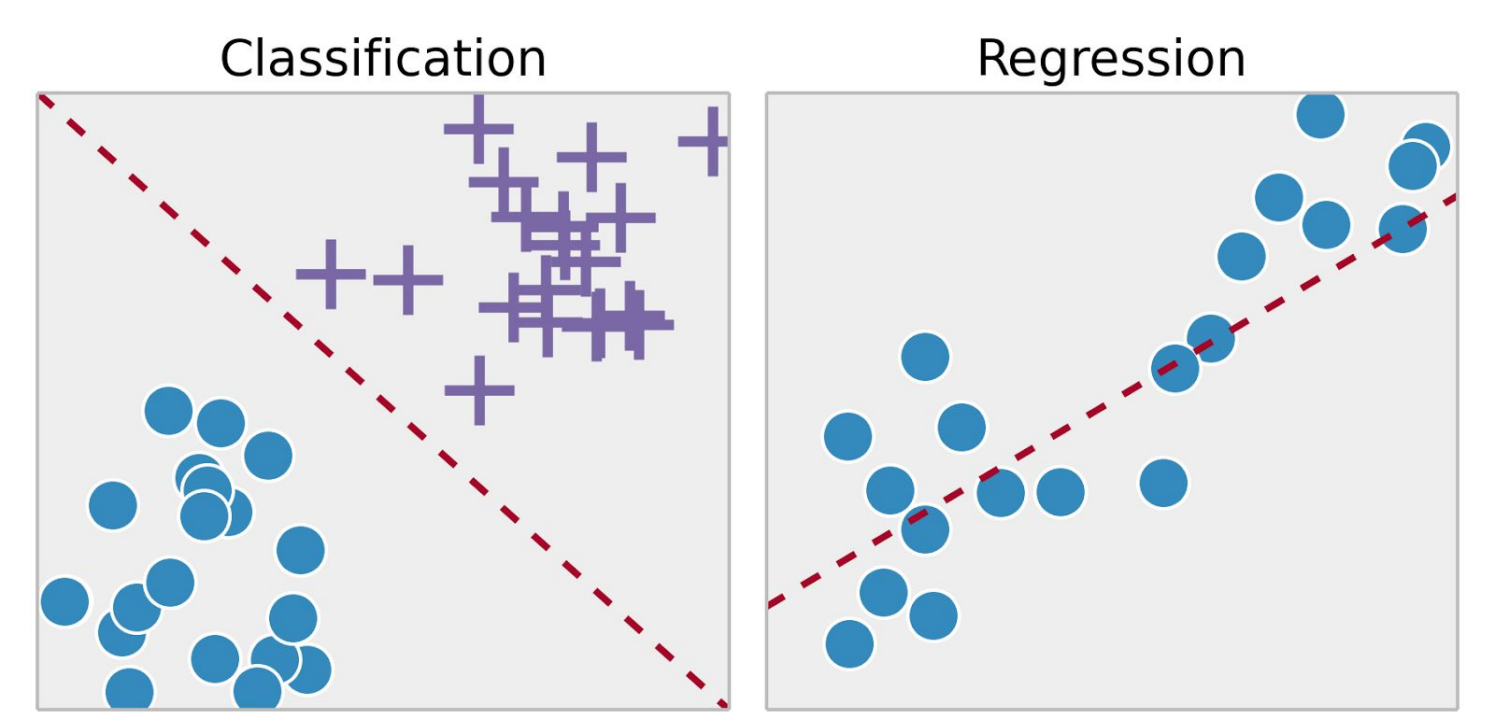
Supervised Learning

Making predictions using data

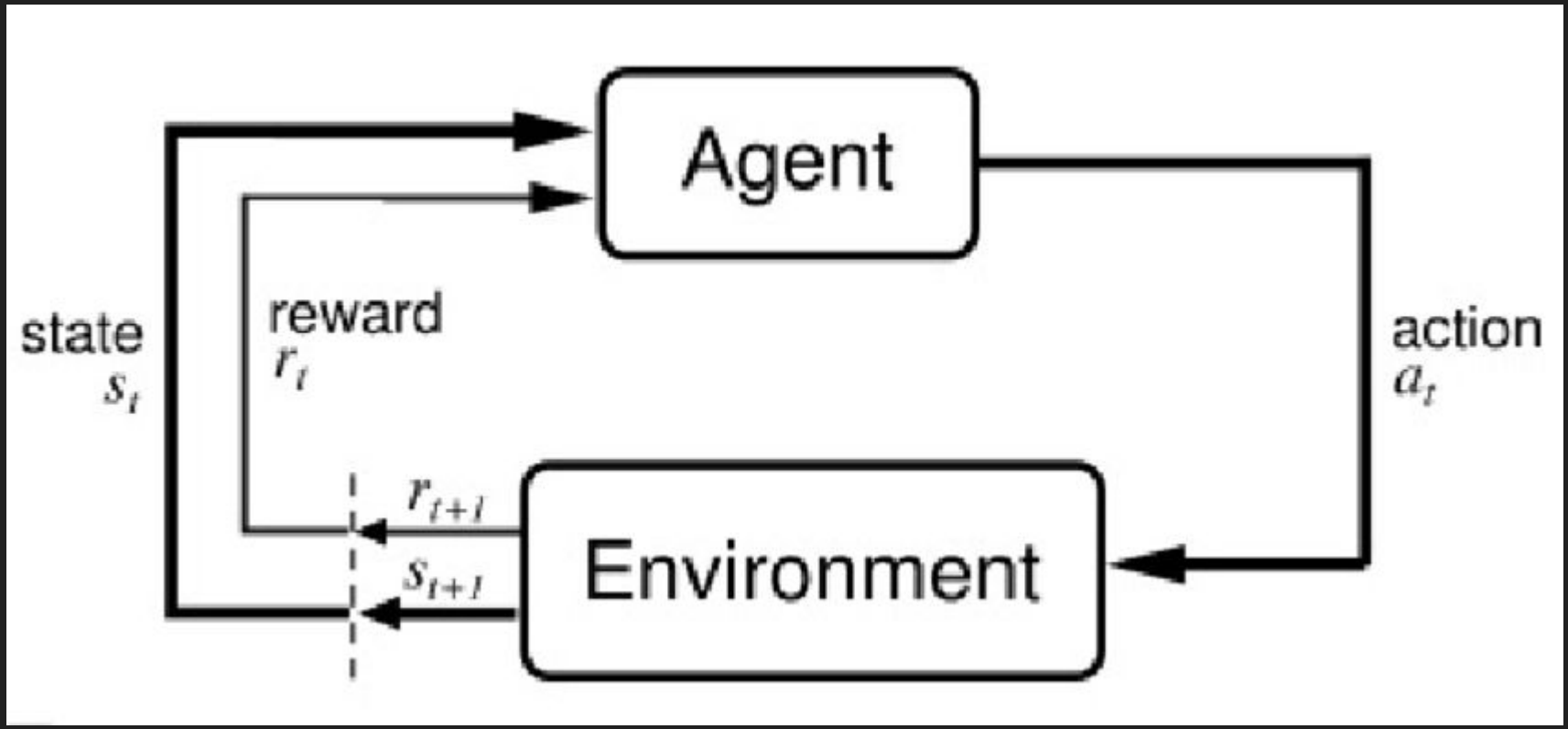
Example: Is a given email "spam" or "ham"?

There is an outcome we are trying to predict

Supervised Learning:
Classification vs. Regression

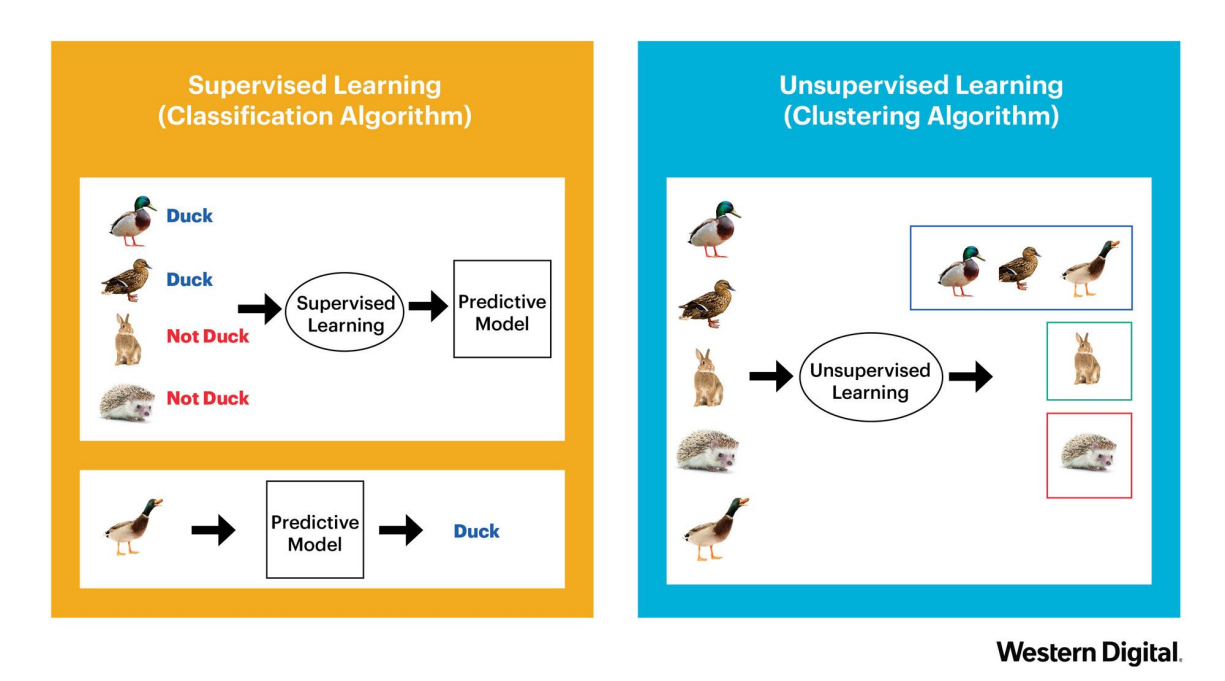


Reinforcement Learning



Unsupervised Learning

Discover hidden structure in unlabeled data.

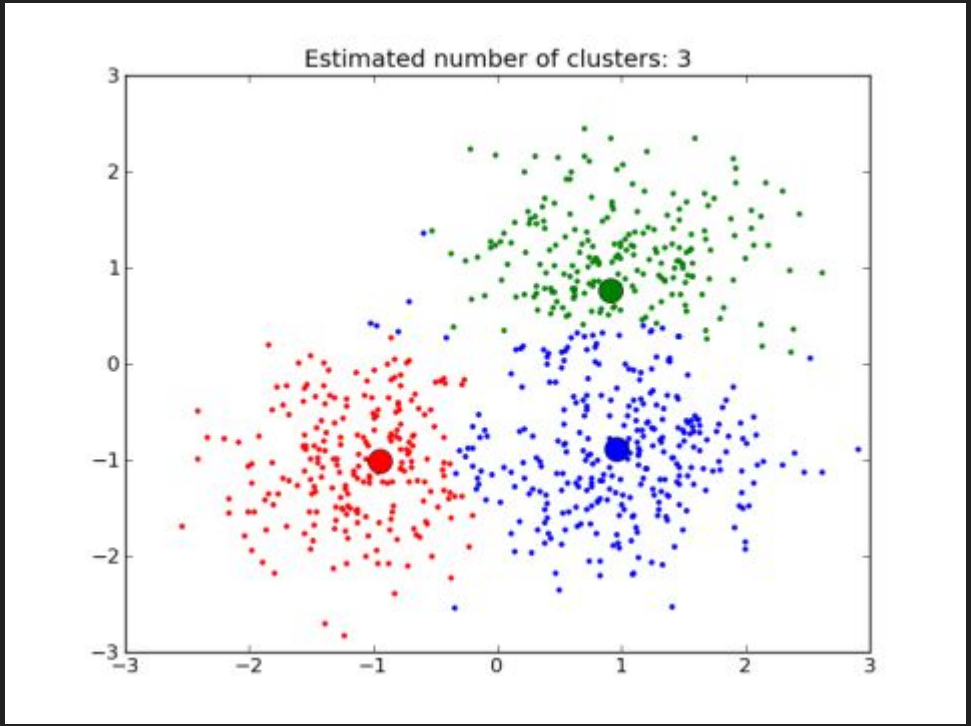


Unsupervised Learning

Extracting structure from data

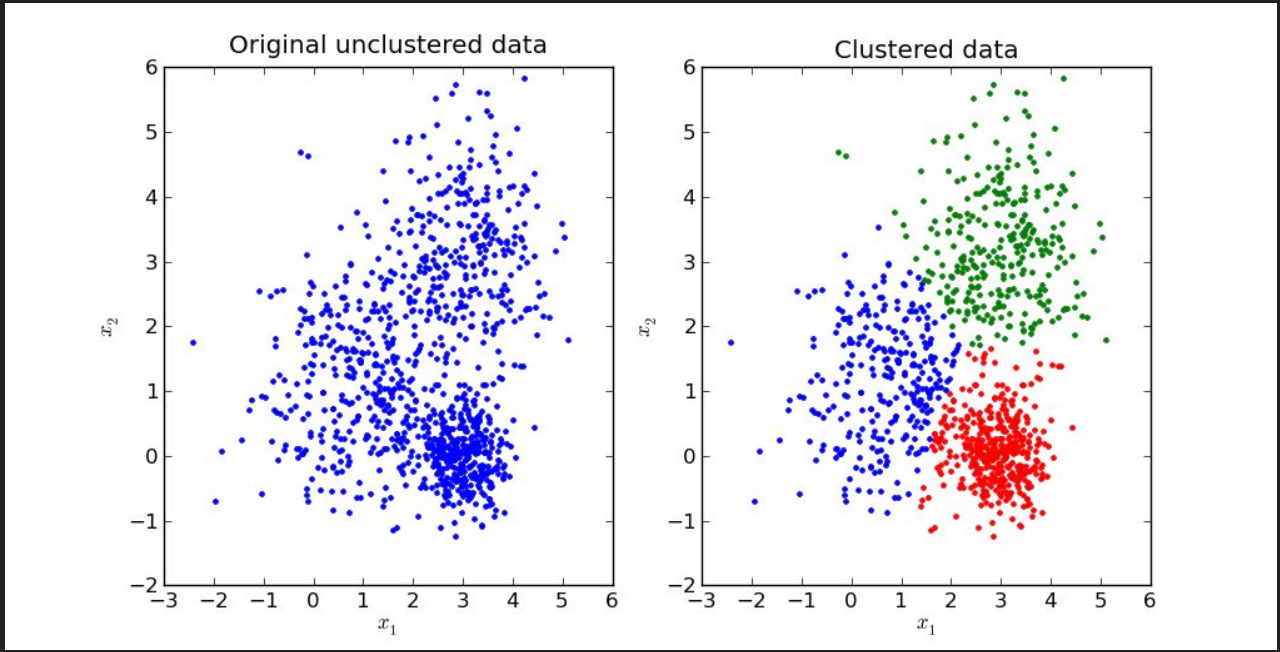
Example: Segment grocery store shoppers into clusters that exhibit similar behaviors

There is no "right answer"



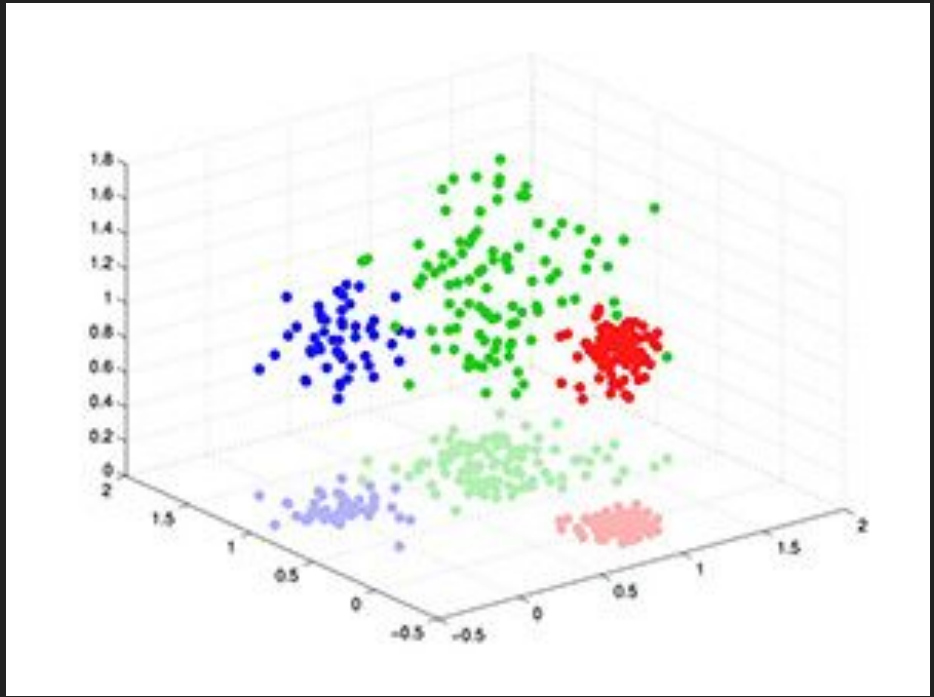
Unsupervised Learning: Clustering

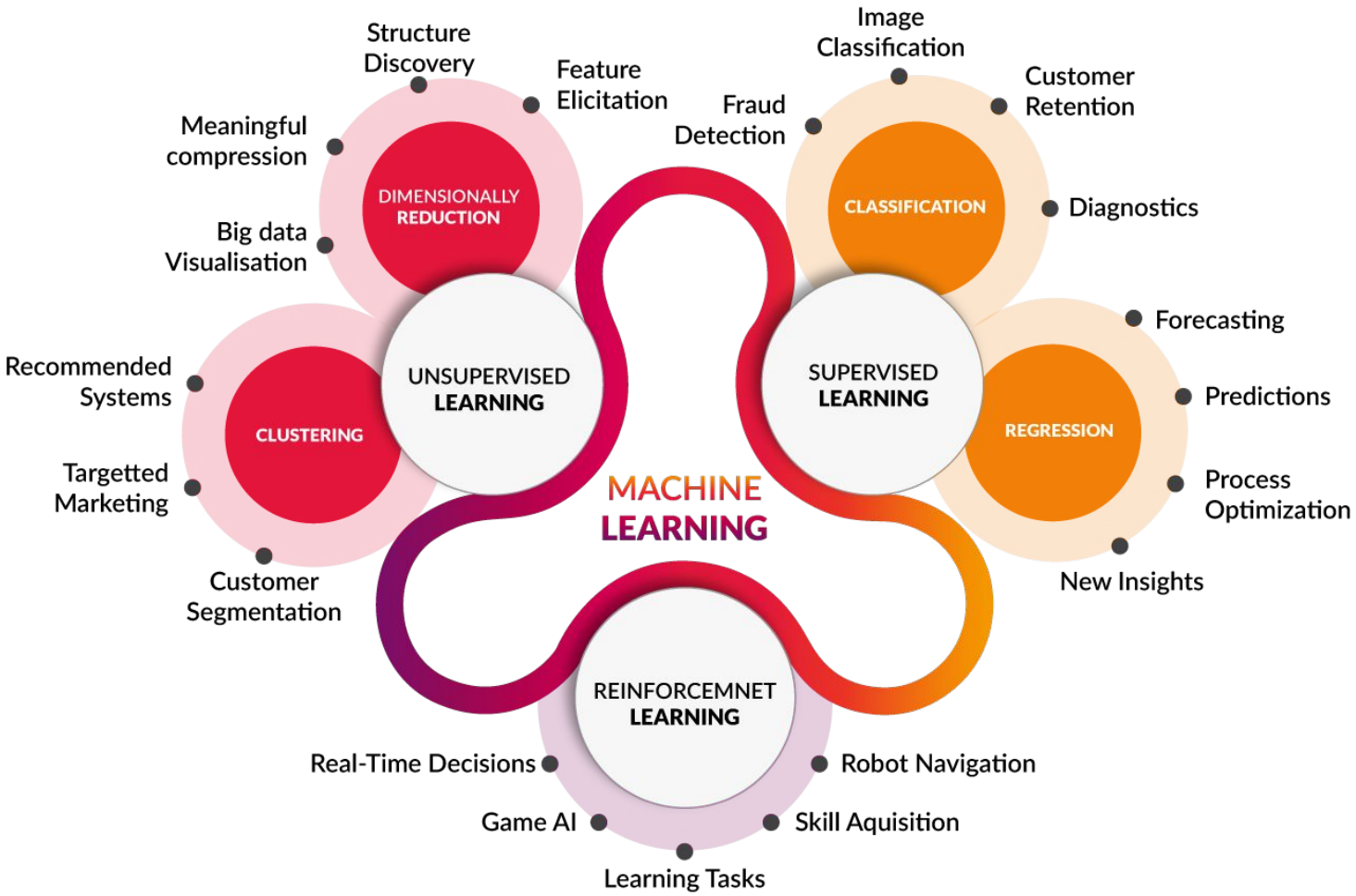
Organize data into meaningful subgroups (clusters) without any prior knowledge of their group membership

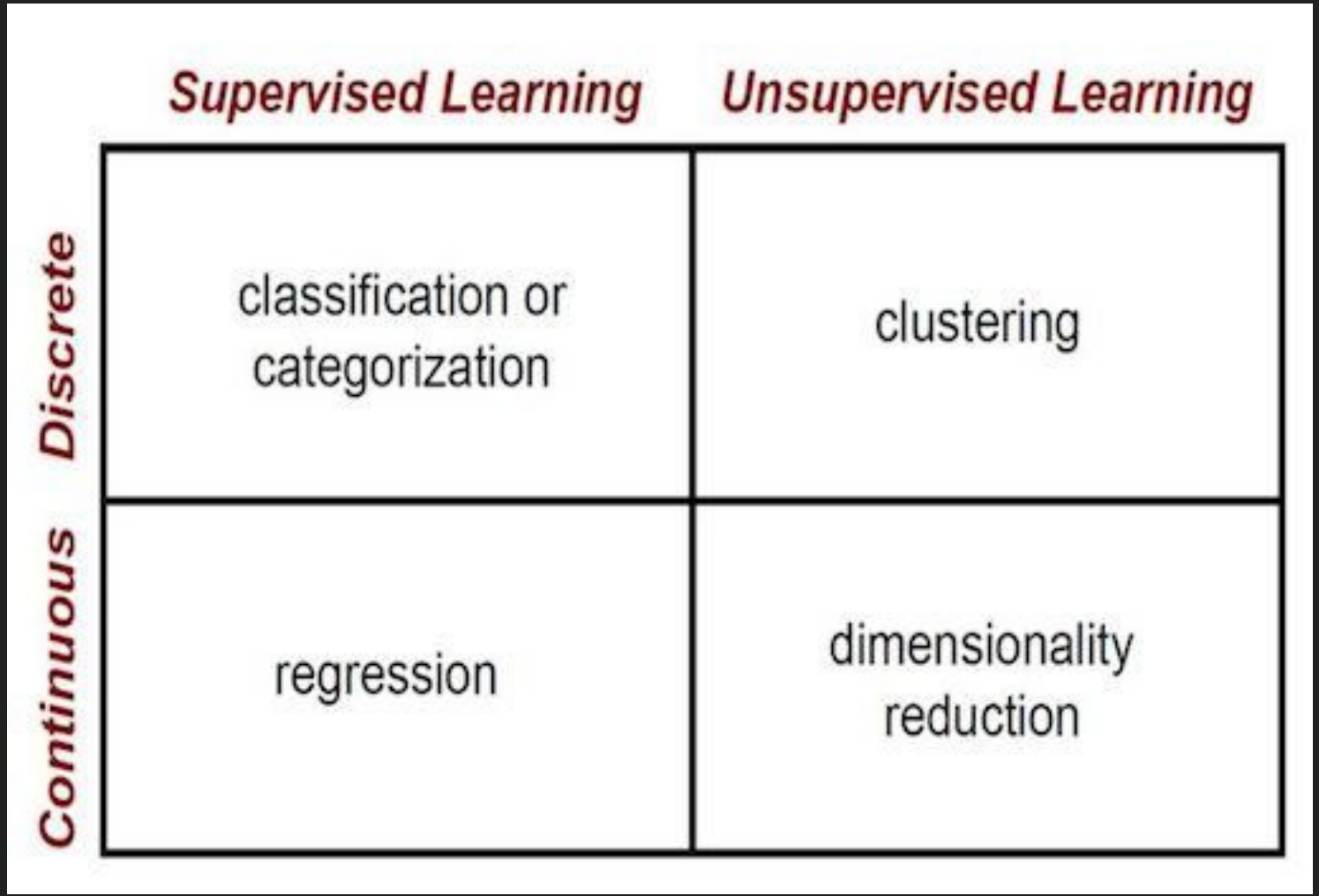


Unsupervised Learning: Dimensionality Reduction

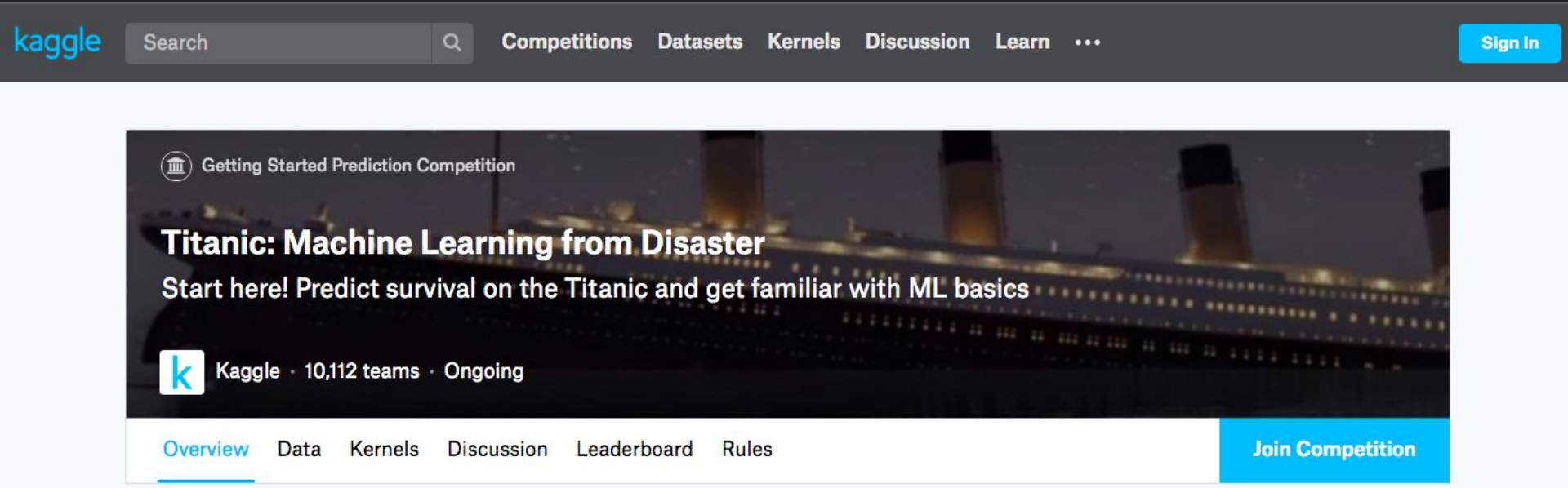
Feature preprocessing to reduce dimensionality of data (remove noise)




MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks


Supervised Learning
Unsupervised Learning
Discrete
Continuous
classification or categorization
clustering
regression
dimensionality reduction

Supervised or Unsupervised Learning?

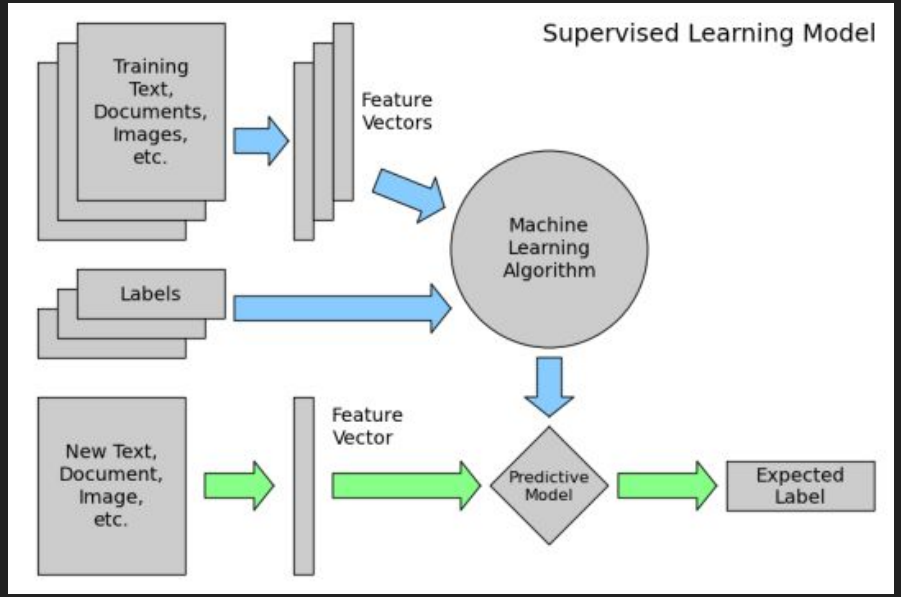


How does machine learning work?

High-level steps of supervised learning:

1) First, train a machine learning model using labeled data

    - "Labeled data" has been labeled with the outcome
    - "Machine learning model" learns the relationship between the attributes of the data and its outcome



2) Then, make predictions on new data for which the label is unknown

The primary goal of supervised learning is to build a model that "generalizes": It accurately predicts the future rather than the past!

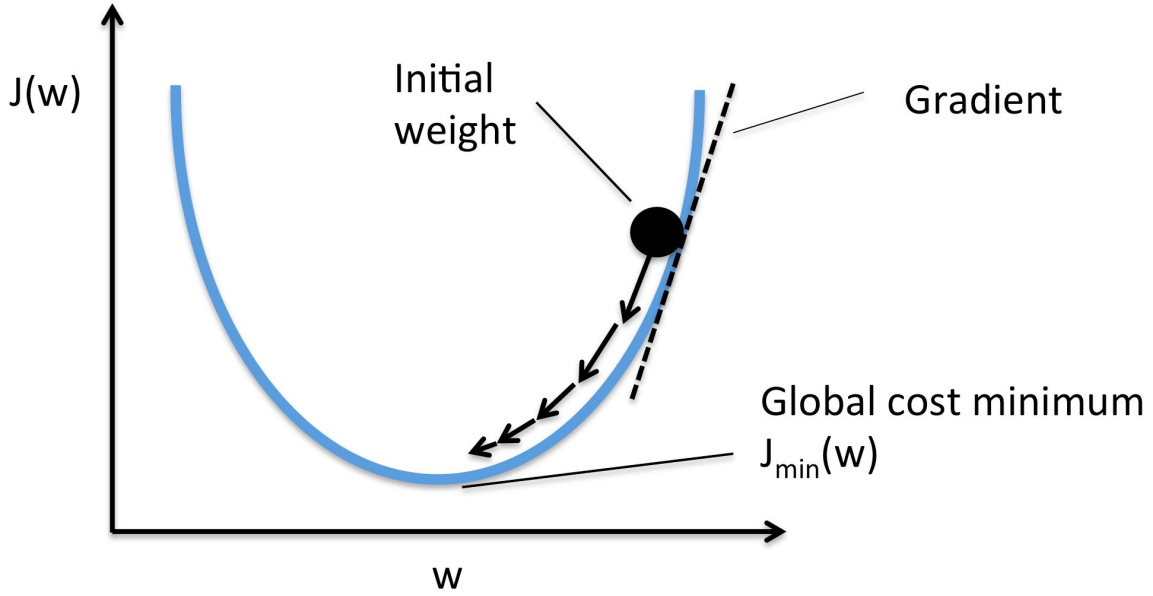
Gradient Descent

Training a robot to go to the pen

Pen: Minimize Distance
Mountain: Minimize Height

Any Machine Learning Problem: Minimize the Error

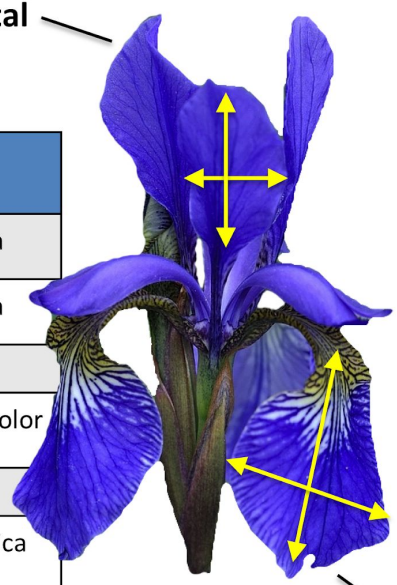
Gradient Descent is minimizing the error.



Basic Terminology

Samples
(instances, observations)

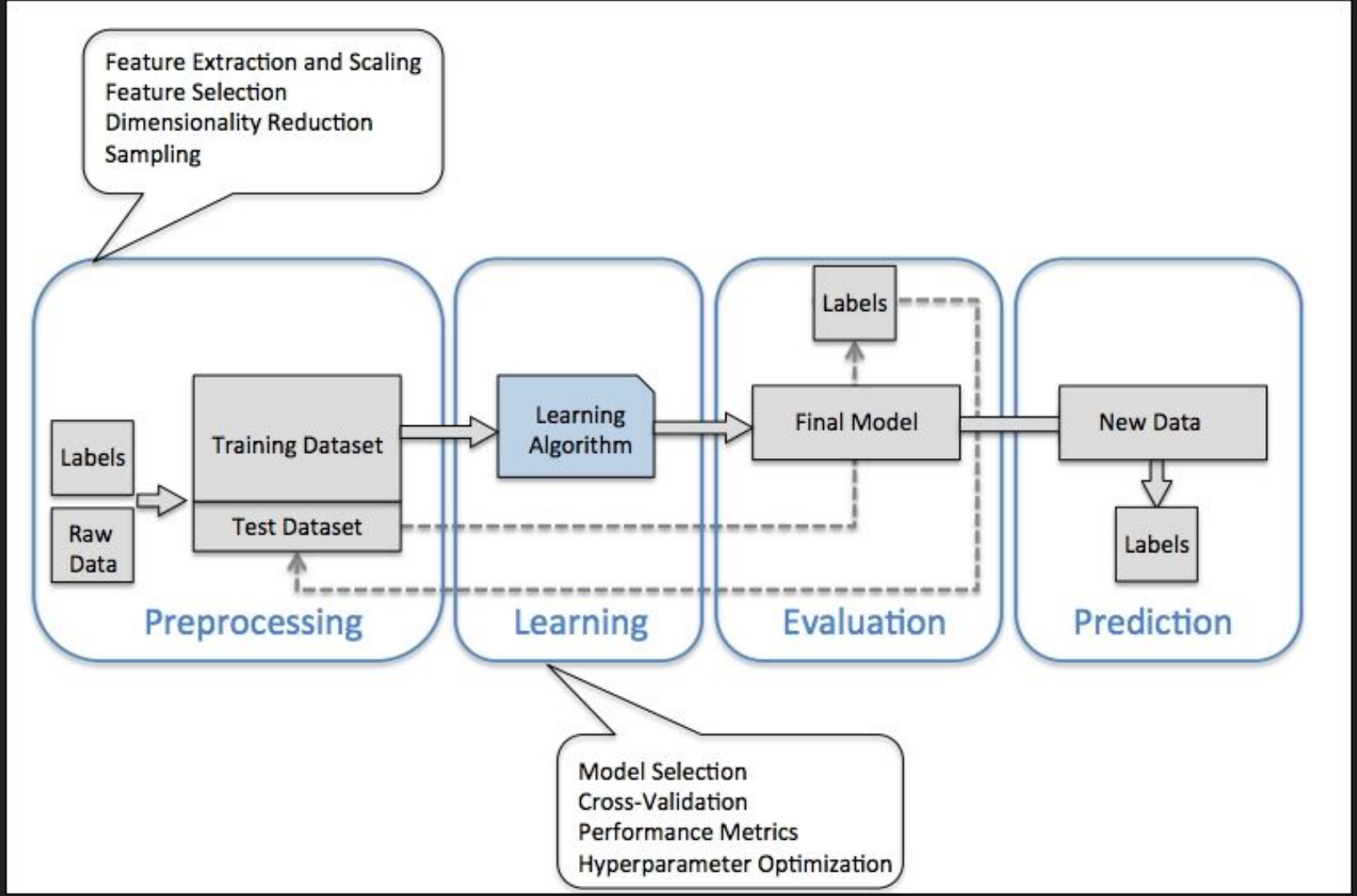
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Let’s take a look at one of the .csv files in our folder and try to use some of our new vocabulary terms...


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

Preprocessing

From raw data to measurement data.

Scaling data

Dimensionality reduction

Splitting into training and testing sets

Training and Selecting a Model

No “silver bullet” model.

Compare performance.

Cross-validation.

Hyperparameter optimization.

Evaluating and Predicting

How the model performs on unseen data (how well does it generalize?)

Use model to predict new, future data

Get to know your coursemates!

You can work in groups or on your own.

Pick one of these questions or make up your own:

- What is your current favorite tool for working with data?
- What are you most excited about learning?
- What can you help your classmates with when it comes to data analysis?

Collect data on your classmates.

Summarize your findings in a narrative, visualization, etc. (feel free to be creative)