UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3253 - Analytic Techniques and Machine Learning

Module 1: Introduction to Machine Learning

Land Acknowledgment

I (we) wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and the Mississaugas of the Credit. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Course Plan

Module Titles
Current Focus: Module 1 – Introduction to Machine Learning
Module 2 – End to End Machine Learning Project
Module 3 – Classification
Module 4 – Clustering and Unsupervised Learning
Module 5 – Training Models and Feature Selection
Module 6 – Support Vector Machines
Module 7 – Decision Trees and Ensemble Learning
Module 8 – Dimensionality Reduction
Module 9 – Introduction to TensorFlow
Module 10 – Introduction to Deep Learning and Deep Neural Networks
Module 11 – Distributing TensorFlow, CNNs and RNNs
Module 12 – Final Assignment and Presentations (no content)

Learning Outcomes for this Module

- Define Machine Learning
- Consider when Machine Learning is applicable
- Enumerate the types of Machine Learning
- Discuss challenges of Machine Learning
- Begin to apply Machine Learning tools and techniques

Topics for this Module

- **1.1** What is machine learning?
- **1.2** Why use machine learning?
- **1.3** Types of machine learning
- **1.4** Modeling
- **1.5** Challenges of machine learning
- **1.6** Tools & Techniques
- **1.7** Resources and Wrap-up

Certificate in Data Science

- Understand the techniques and methods of predictive and Big Data analytics
- Learn how to use tools such as Python and Hadoop to tackle data analysis challenges
- Develop and use models tools to solve business problems and mine data for fresh insights

Certificate in Data Science (Cont'd)

What You'll Learn

- Explore the evolution of data science and predictive analytics
- Know statistical concepts and techniques including regression, correlation and clustering
- Apply data management systems and technologies that reflect concern for security and privacy
- Adopt techniques and technologies including data mining, neural network mapping and machine learning
- Represent big data findings visually to aid decision-makers

Certificate in Data Science (Cont'd)

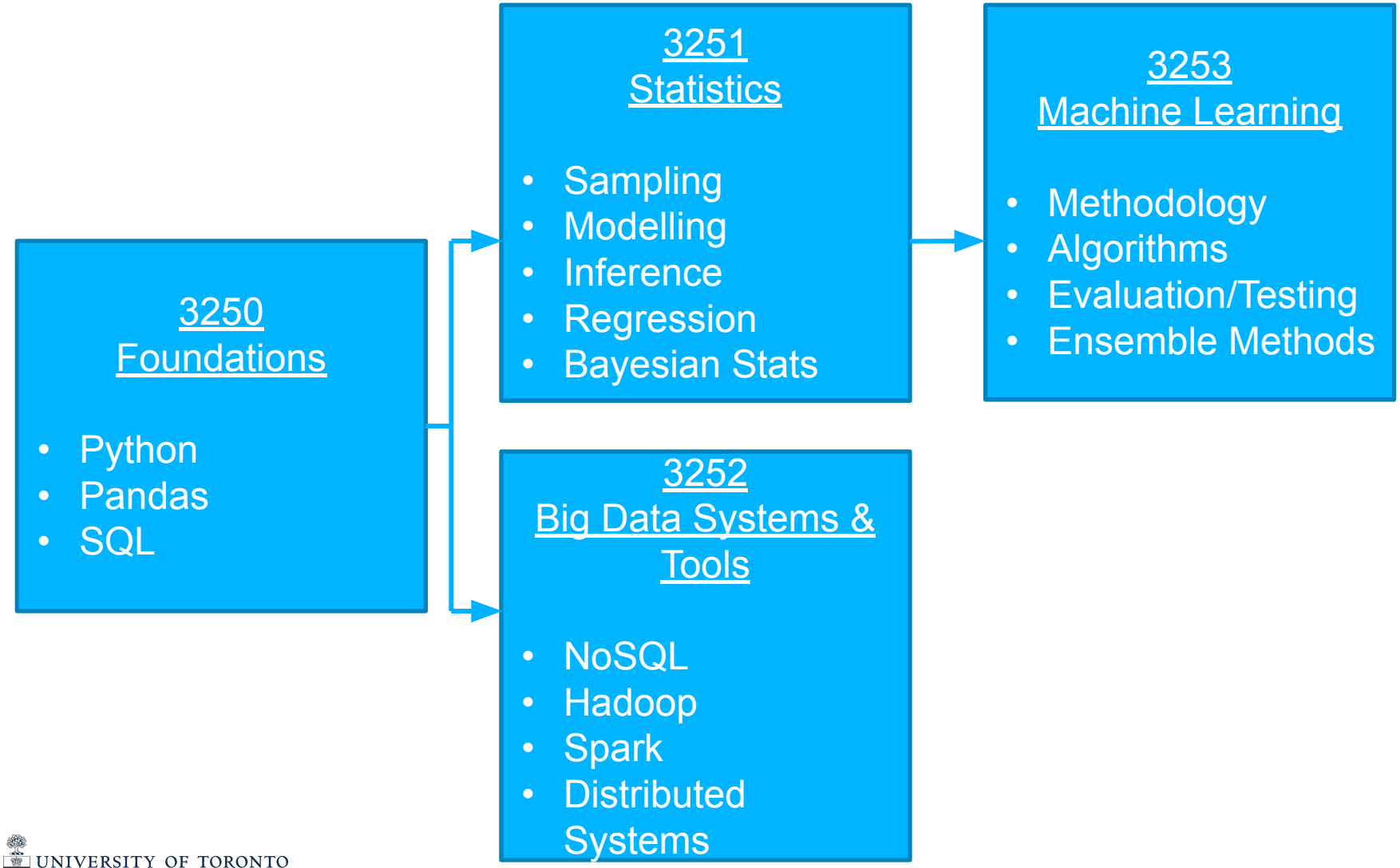
Courses

- SCS 3250 – Foundations of Data Science
- SCS 3251 – Statistics for Data Science
- SCS 3252 – Big Data Management Systems & Tools
- **SCS 3253 – Machine Learning**

Prerequisites

- This course assumes that you are comfortable programming in Python and that you are familiar with Python's main scientific libraries, in particular NumPy, Pandas, and Matplotlib
- Experience with notebook environments
- Understanding of college-level math as well (calculus, linear algebra, probabilities, and statistics)

Certificate in Data Science Fundamentals (Cont'd)



Certified Analytics Professional

- Industry Certification
- Operated by INFORMS, the world's largest professional society for those in the field of analytics, operations research (O.R.), and the management sciences
- Requires experience doing analytics and a related degree (or equivalent additional experience)
- Code of ethics

The CAP Domains

Coverage in this certificate program

	3250	3251	3252	3253
I. Business Problem (Question) Framing	✓	✓✓	✓	✓✓✓
II. Analytics Problem Framing	✓	✓✓✓	✓	✓✓✓
III. Data	✓✓	✓✓✓	✓✓	✓✓✓
IV. Methodology (Approach) Selection	✓	✓✓		✓✓✓
V. Model Building		✓✓✓	✓	✓✓✓
VI. Deployment			✓✓✓	✓
VII. Model Life Cycle Management			✓✓	✓✓✓

✔ = Introductory content

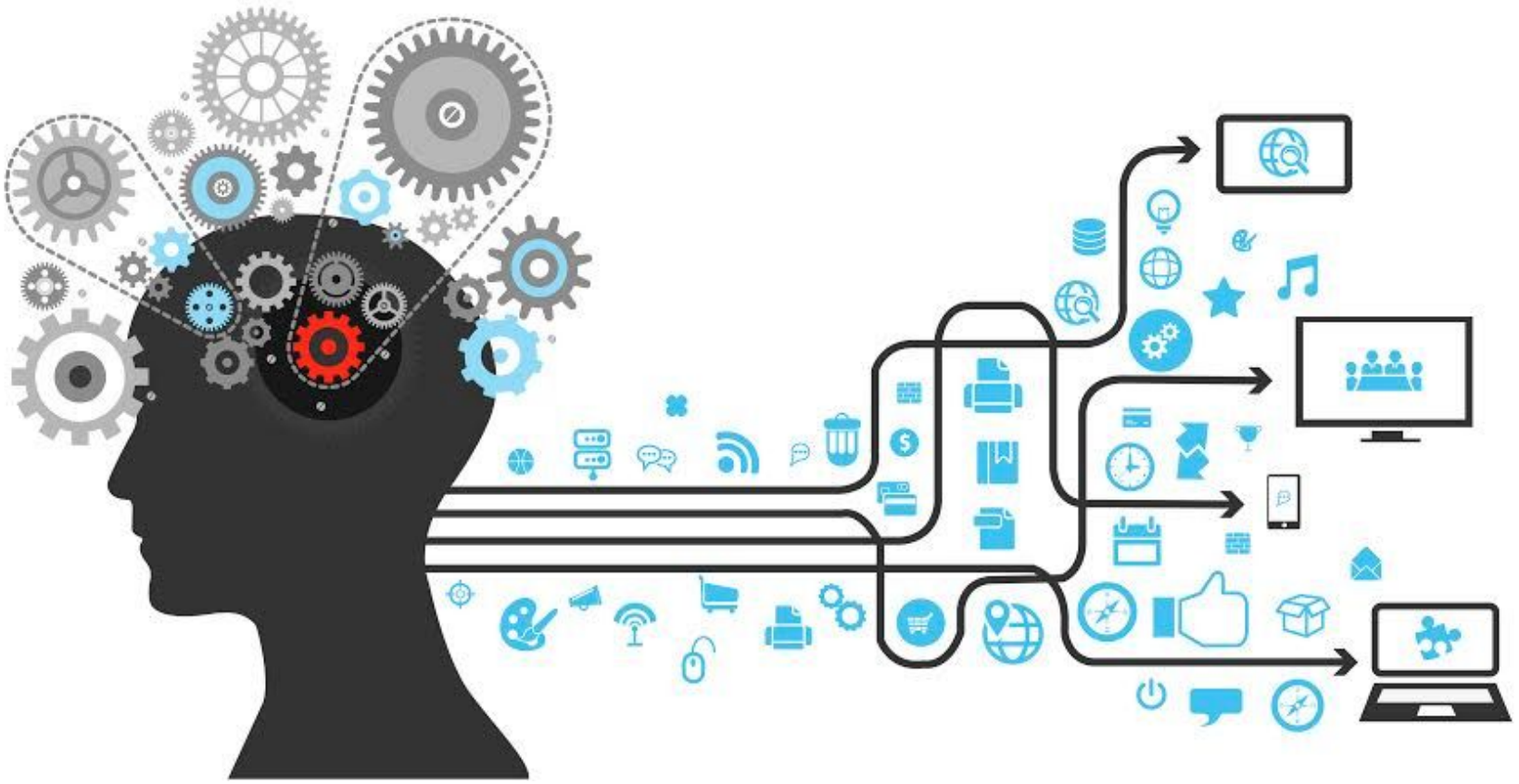
✔ ✔ = Substantial coverage

✔ ✔ ✔ = Major focus

Module 1 – Section 1

What is Machine Learning?

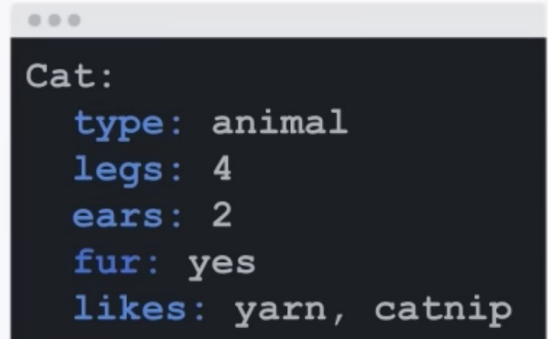
Machine Learning



Machine Learning

- [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

  Arthur Samuel, 1959



- "Modern AI is modeled after ideas about how the brain works…instead of programming the computer you show it lots of examples…and it learns to produce the right answers without you ever programming."

  Geoffrey Hinton

  (Emeritus Prof. Comp Sci, U.of.T)

Module 1 – Section 2

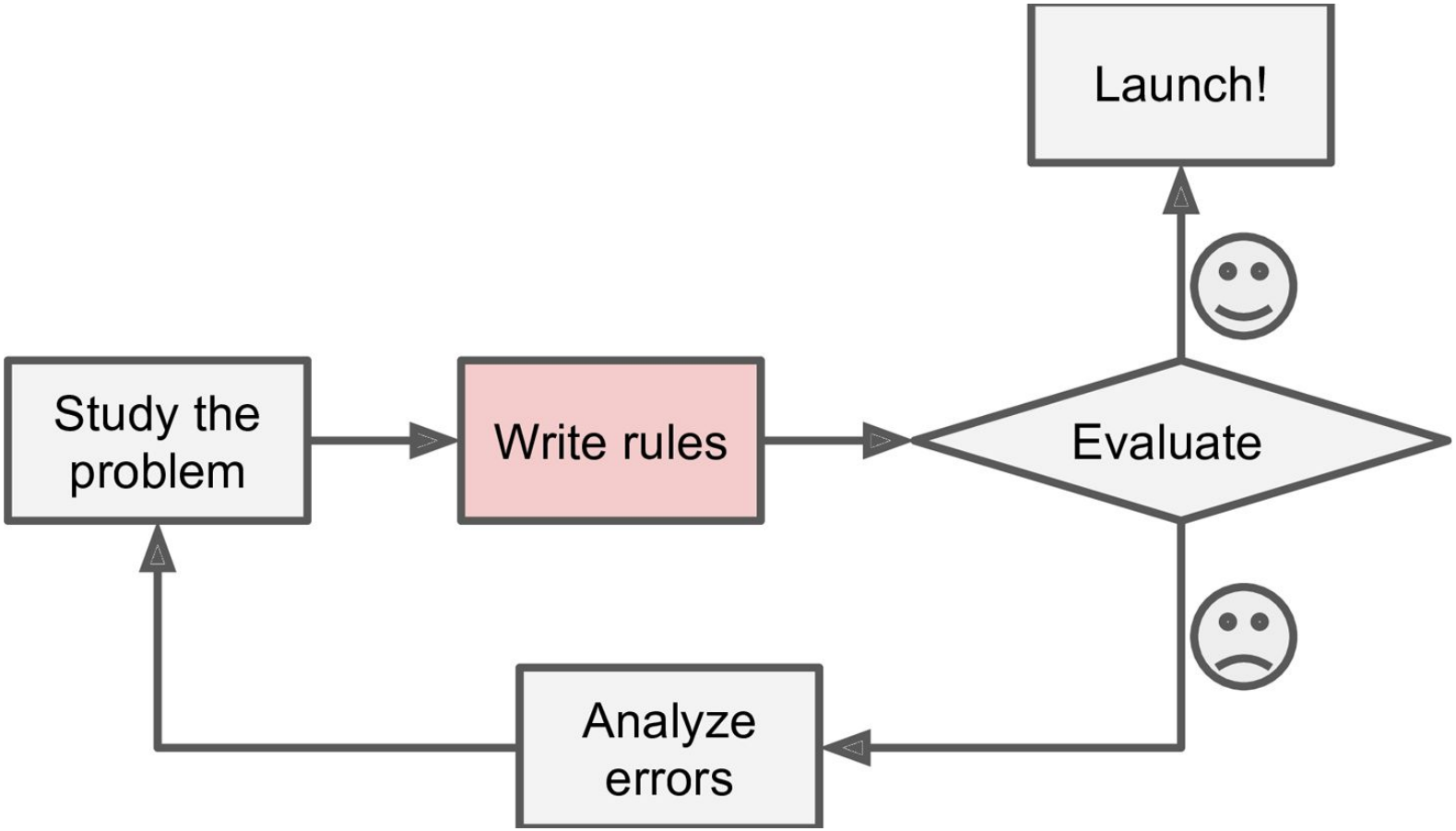
Why Machine Learning?

Why Machine Learning

*IDC Digital Universe report, 2014 http://www.emc.com/infographics/digital-universe-2014.htm

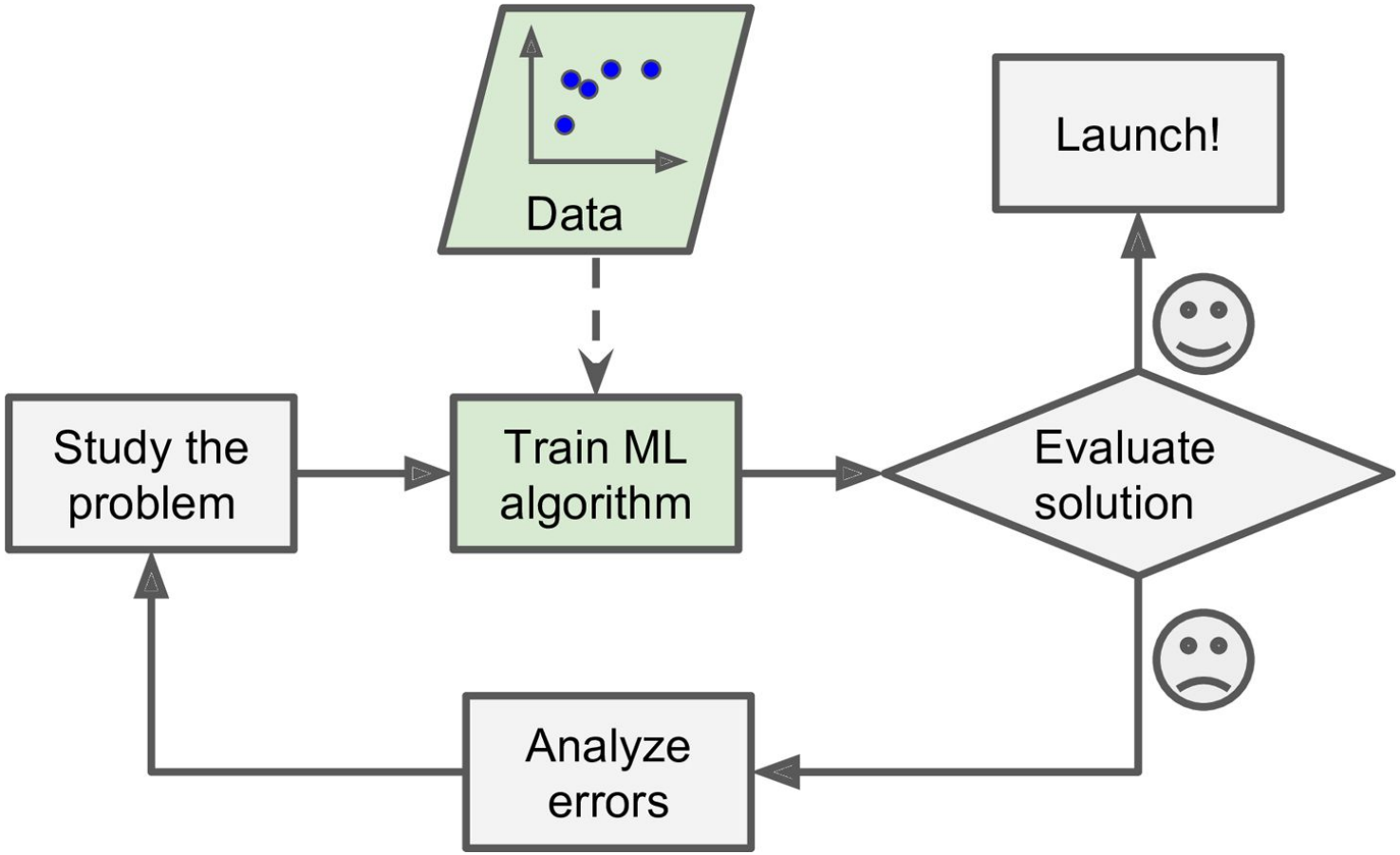
**Data Scientist: The Sexiest Job of the 21st Century, Oct 2012 https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Traditional Approach



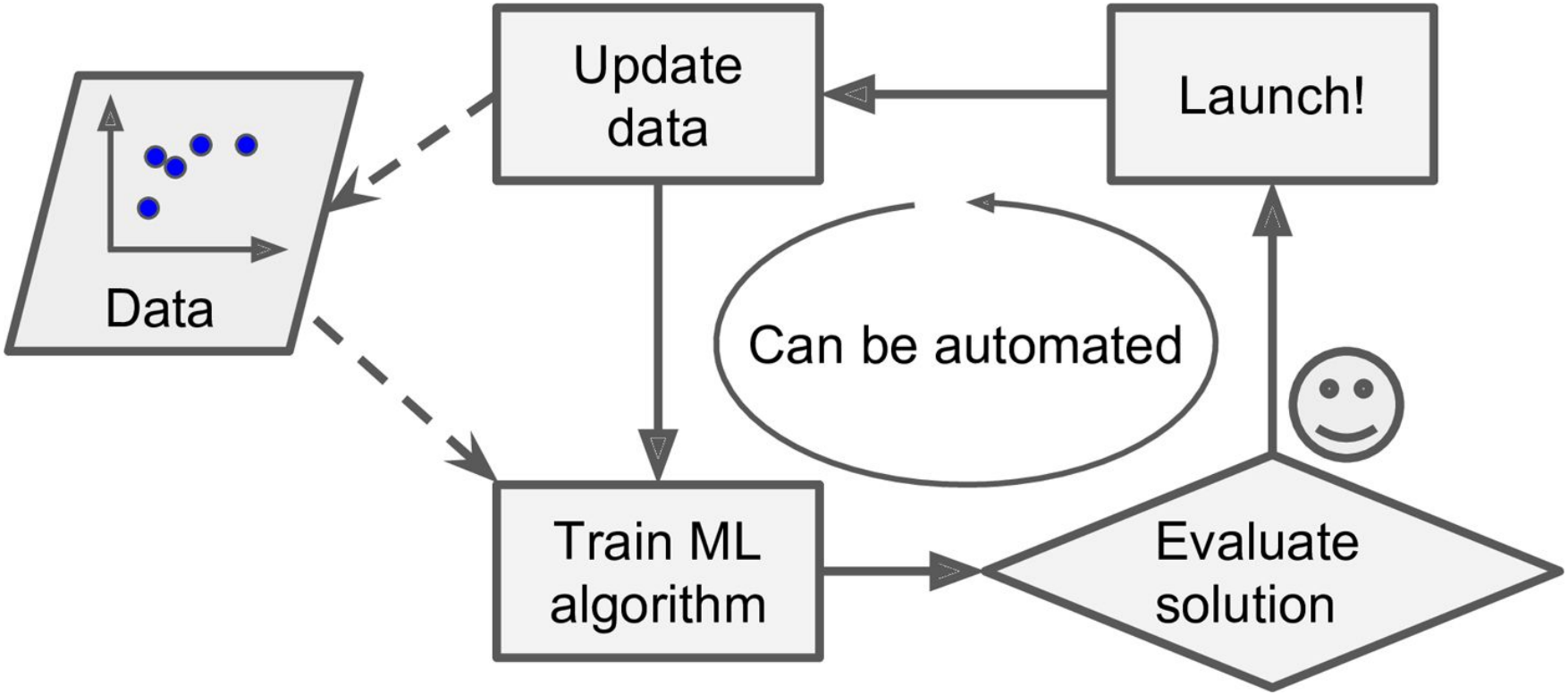
Long list of rules that can break!

Machine Learning Approach



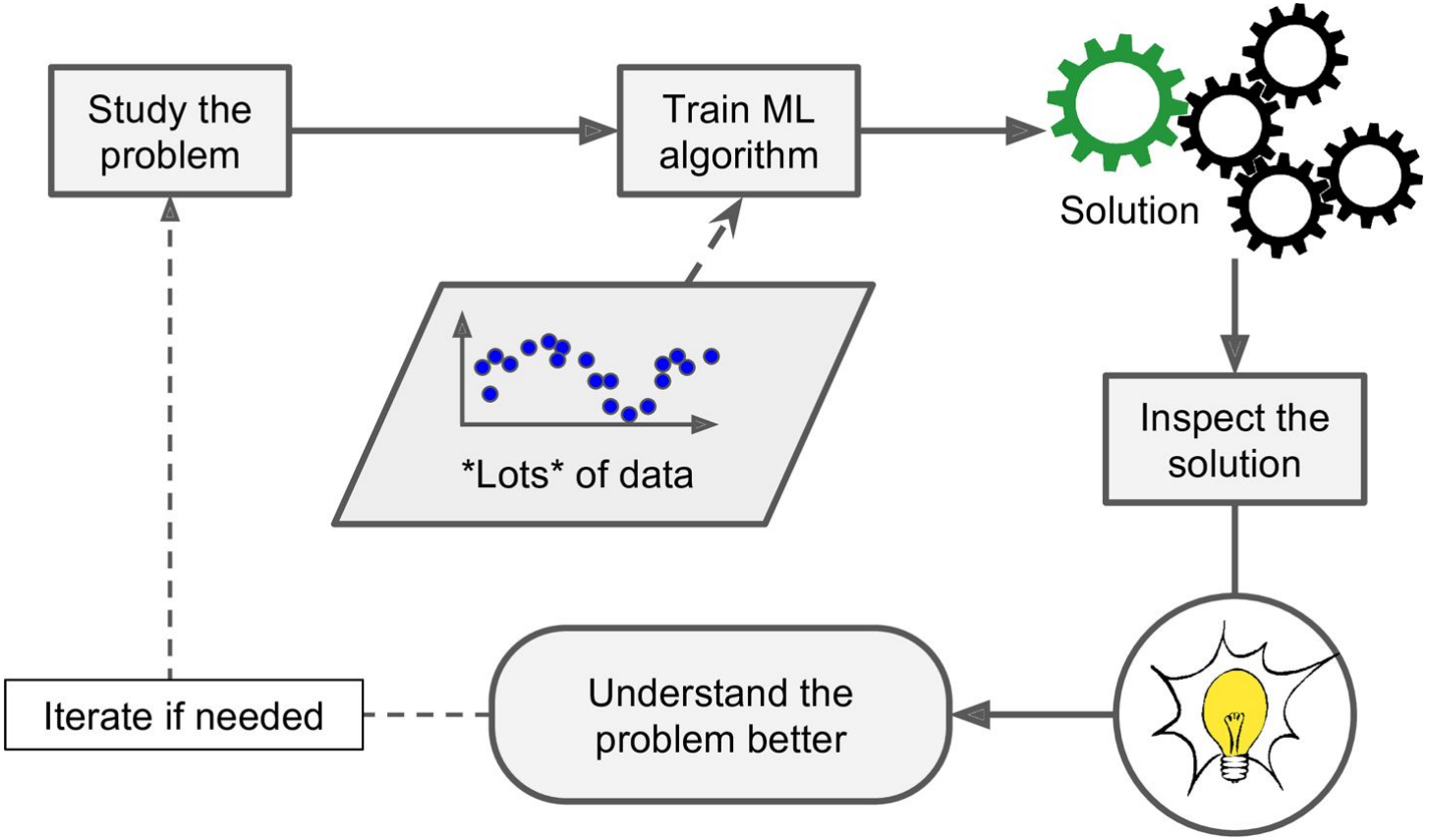
Learn from data.
Generalize!

Machine Learning Automation



Update model as new data arrives

Machine Learning Automation (cont'd)



Find non-obvious patterns.

Module 1 – Section 3

Types of Machine Learning

Machine Learning

Supervised Learning

- Labeled target variable
- Model is “trained” using labeled target variables

Machine Learning

Supervised Learning

- Labeled target variable
- Model is “trained” using labeled target variables

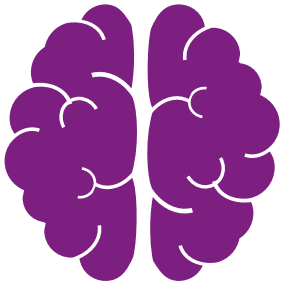
Machine Learning

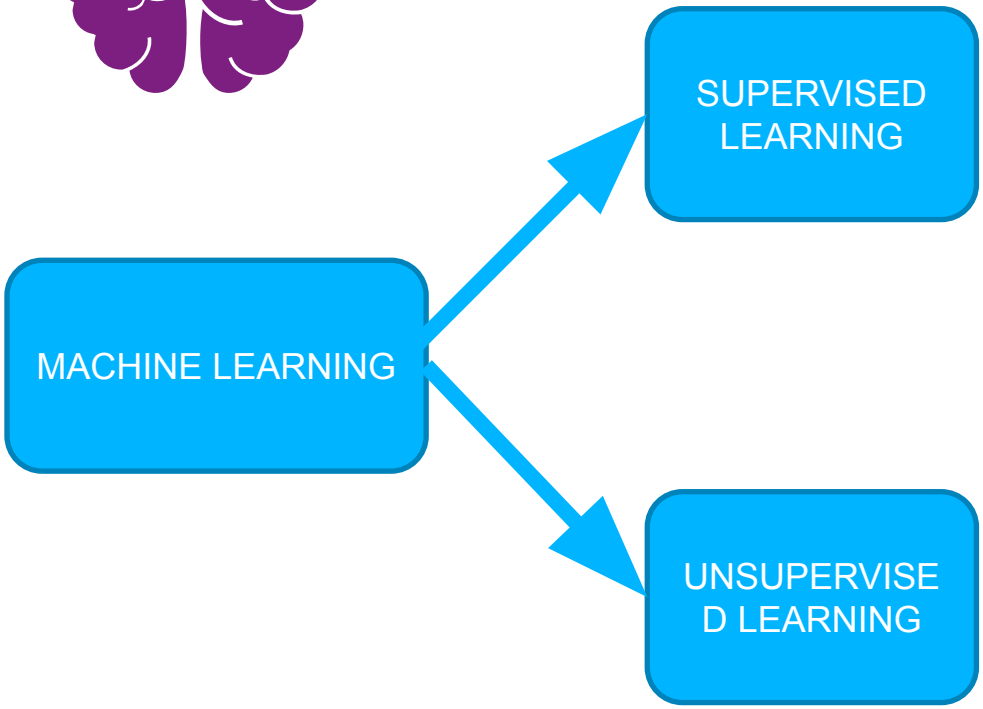
Supervised Learning

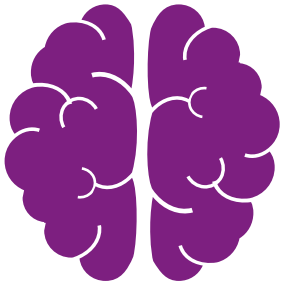
- Labeled target variable
- Model is “trained” using labeled target variables

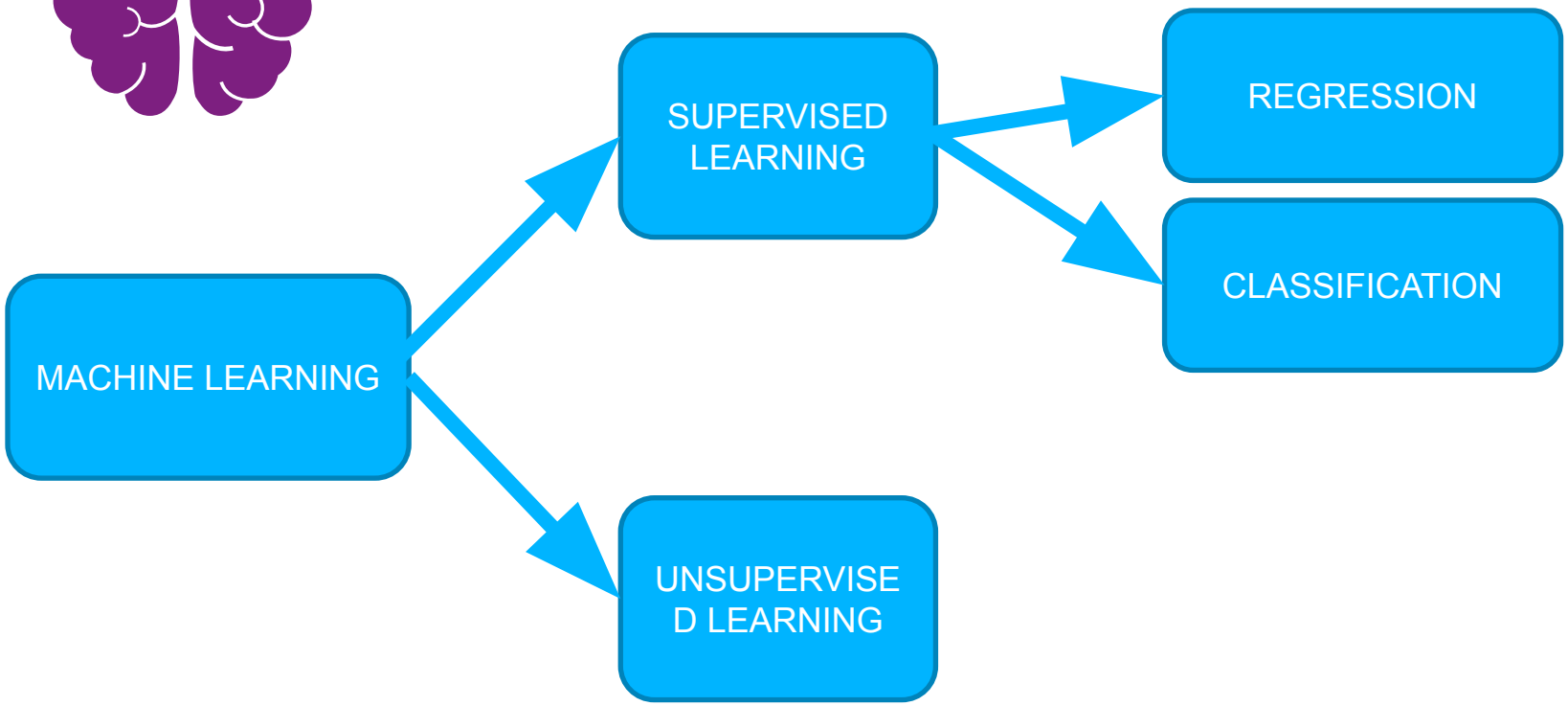
Unsupervised learning

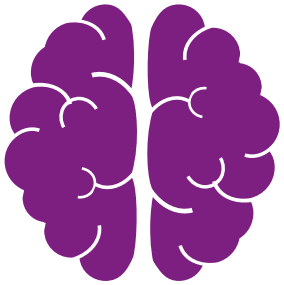
- No target variable
- No model to train/fit
- Data are clustered/grouped

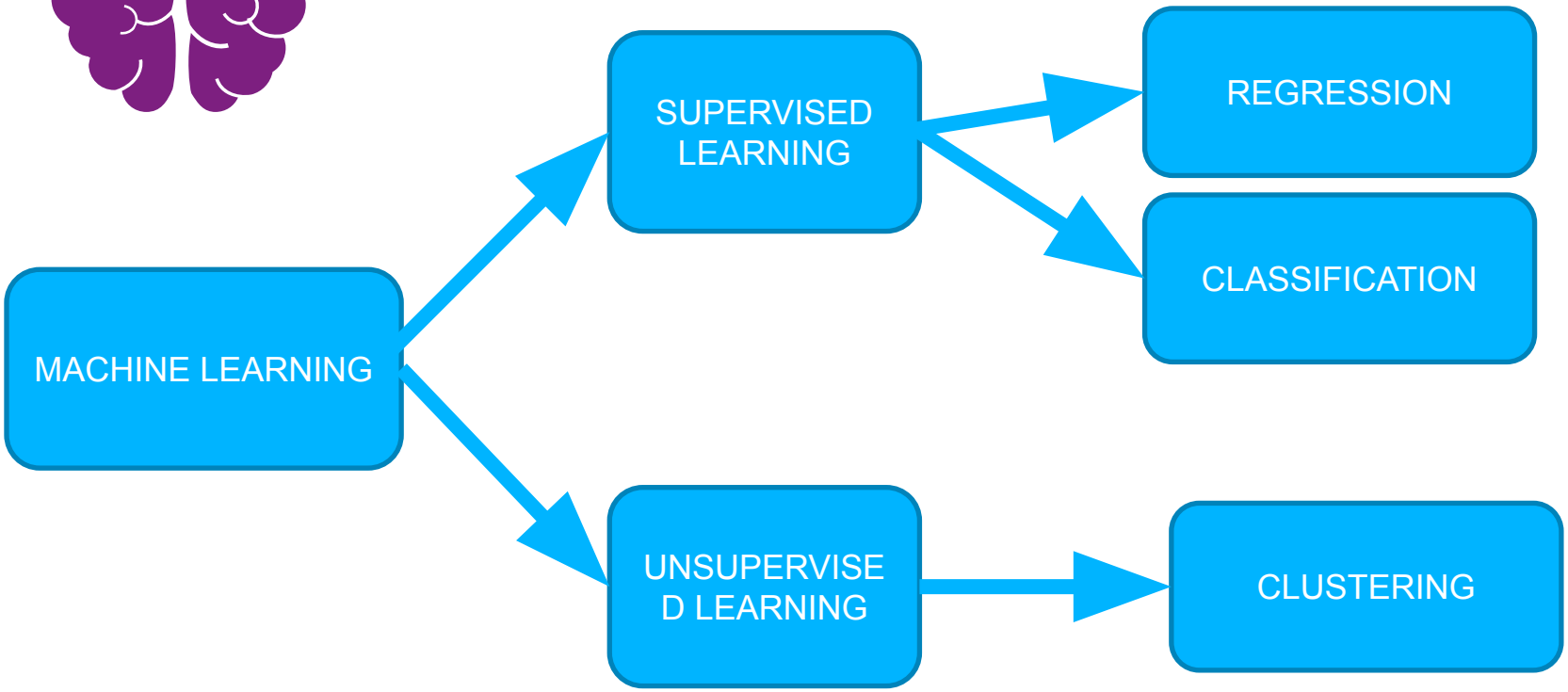



SUPERVISED LEARNING
MACHINE LEARNING
UNSUPERVISE D LEARNING




MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISE D LEARNING
REGRESSION
CLASSIFICATION

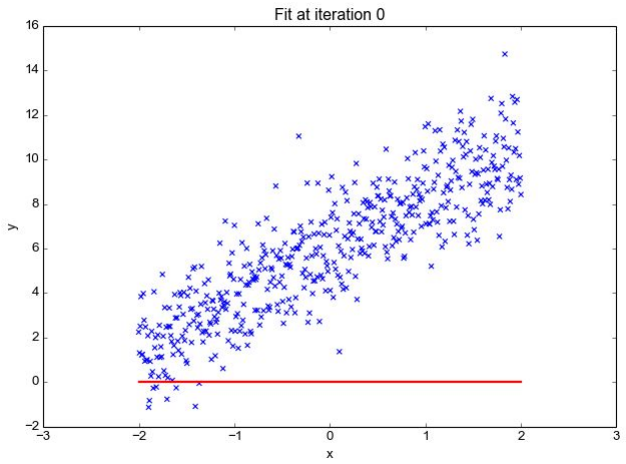


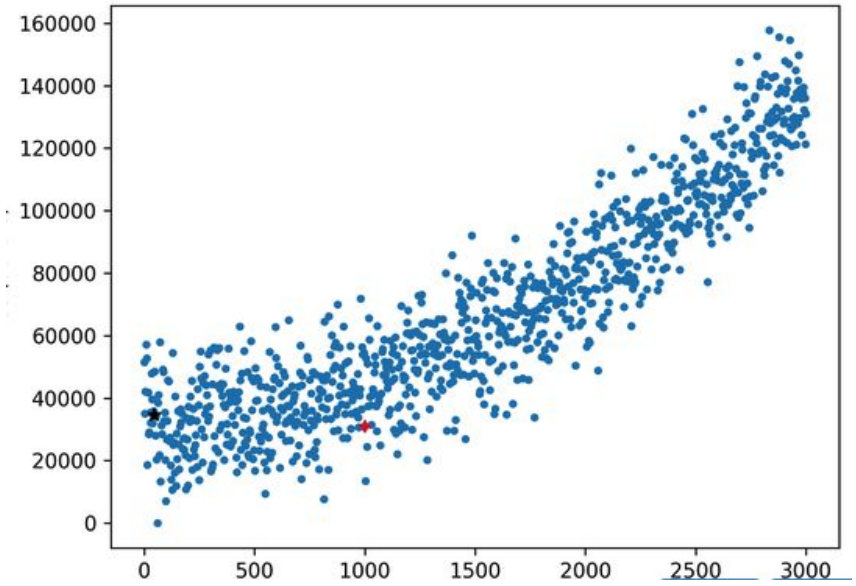

MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISE D LEARNING
REGRESSION
CLASSIFICATION
CLUSTERING

Regression (Supervised Learning)

Predicts value of a continuous variable

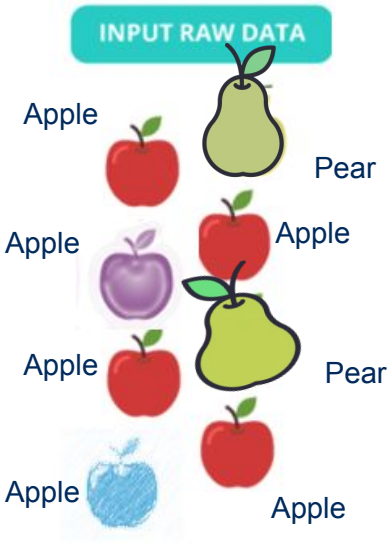
- **Age**
- **Income**
- **Price of a Car**





Level of Pain in a patient

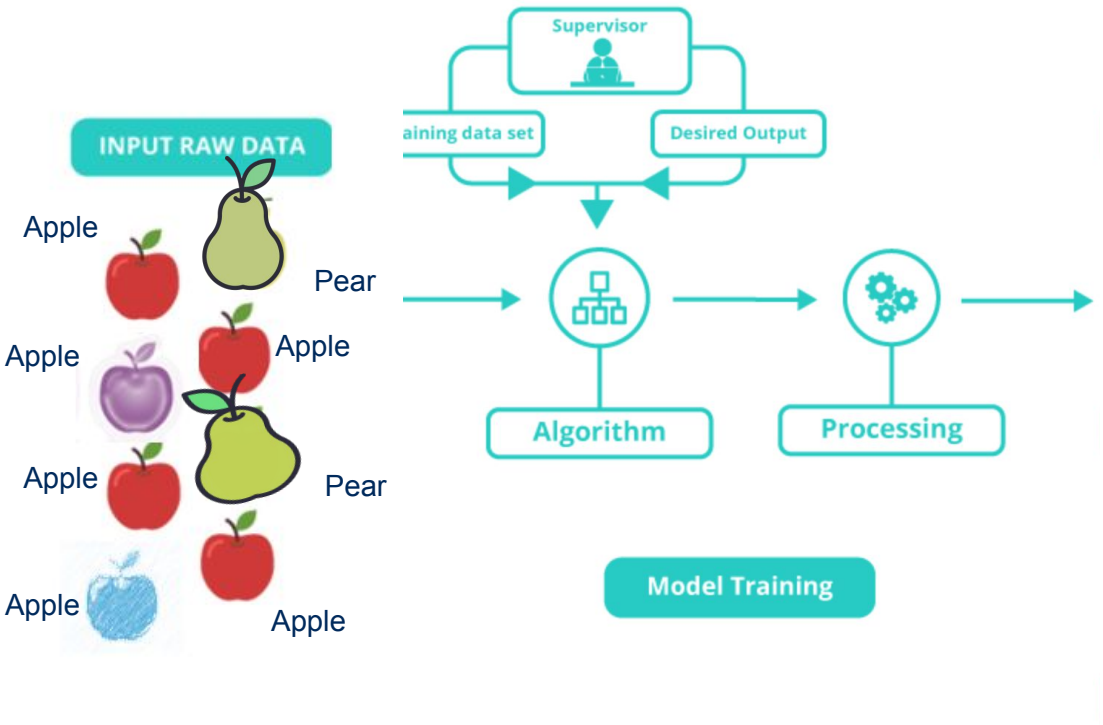
Classification (Supervised Learning)



For each apple, we know their propertise

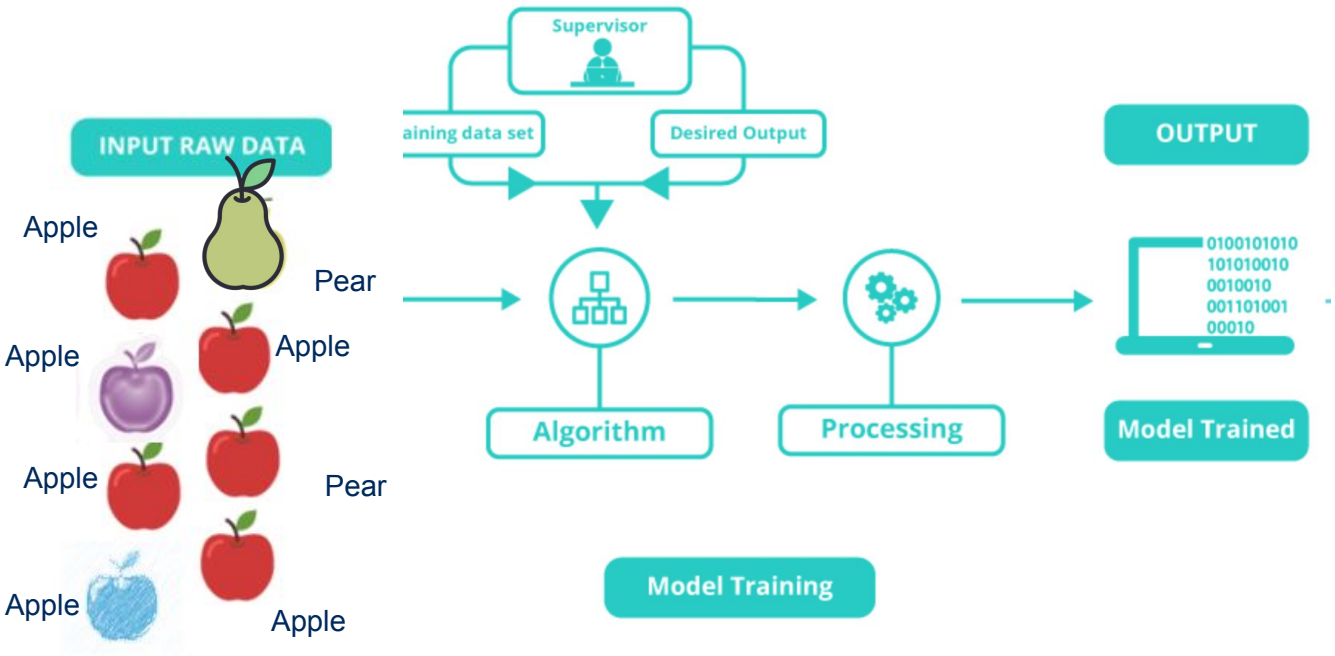
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Classification (Supervised Learning)



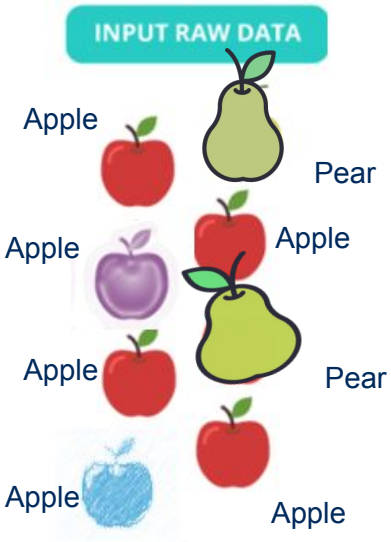
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

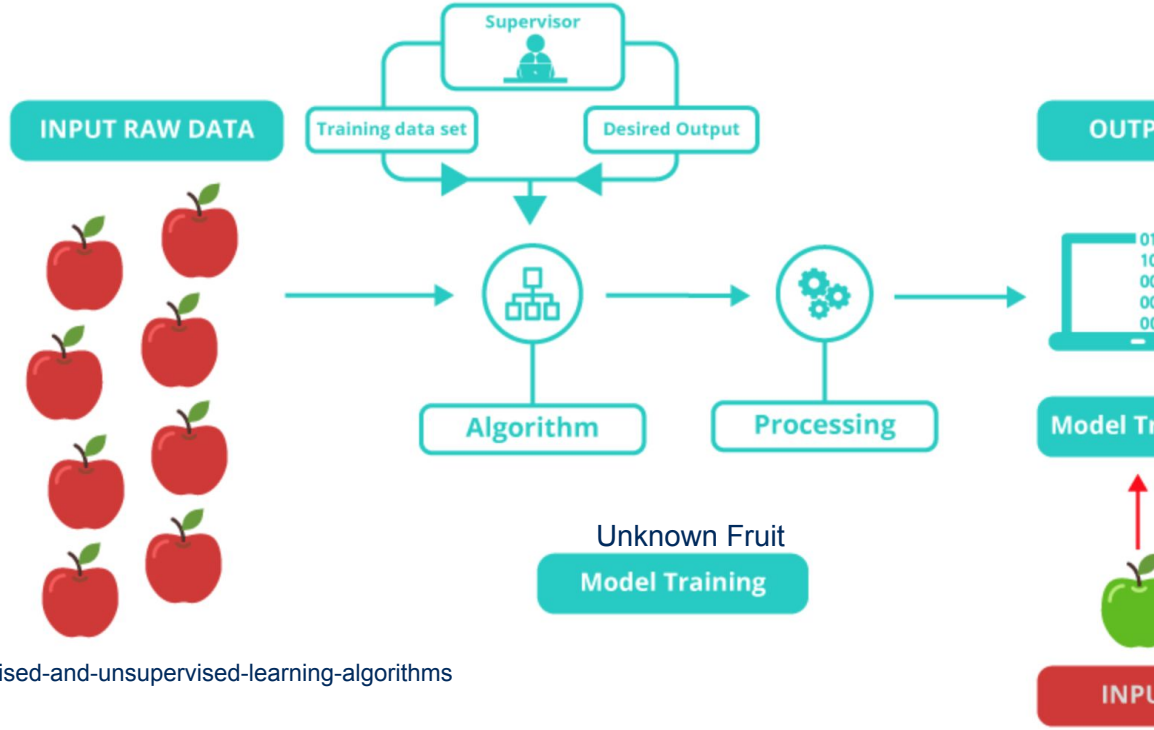
Classification (Supervised Learning)



Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

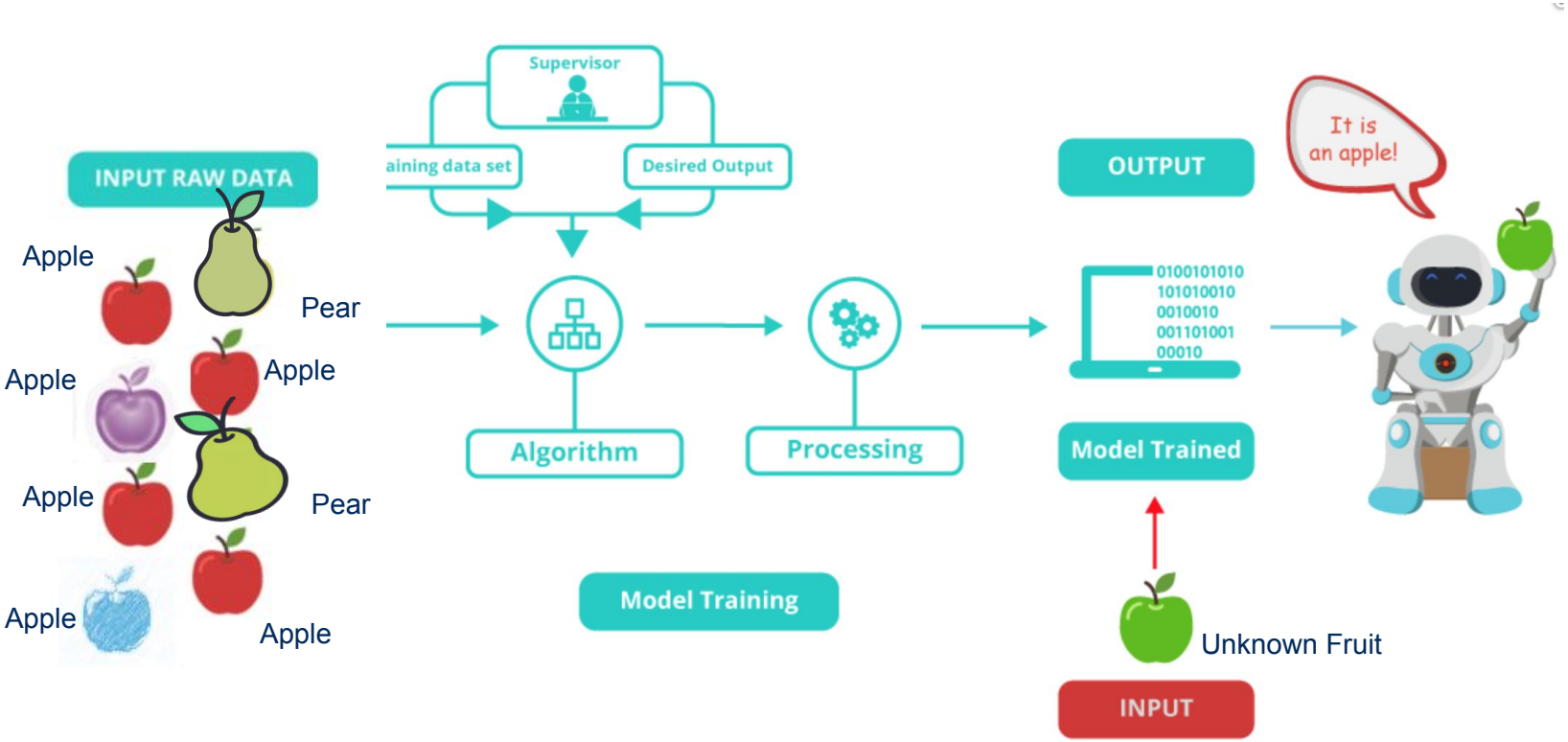
Classification (Supervised Learning)





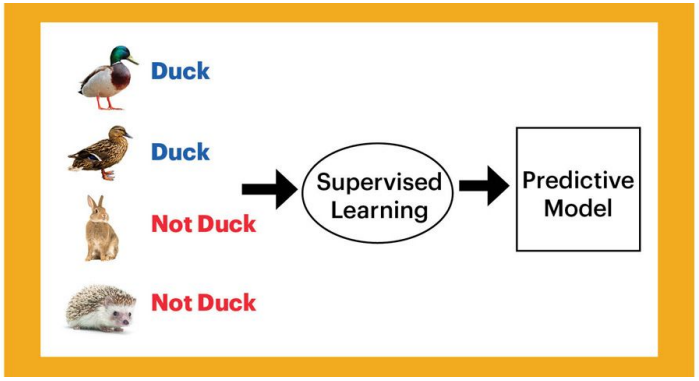
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Classification (Supervised Learning)



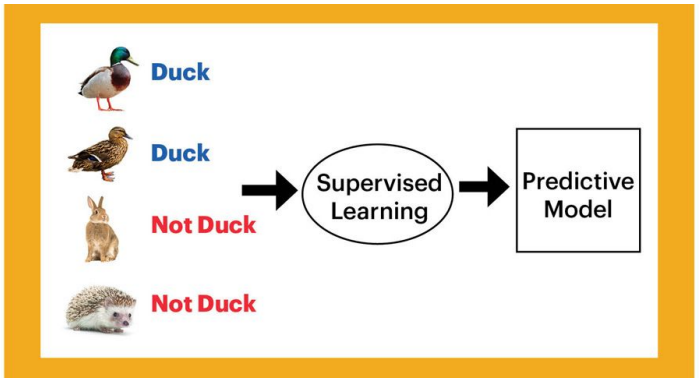
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

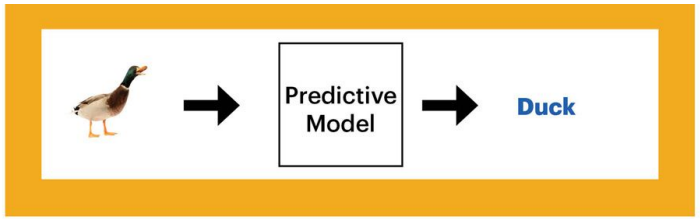
Classification (Supervised Learning)



Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

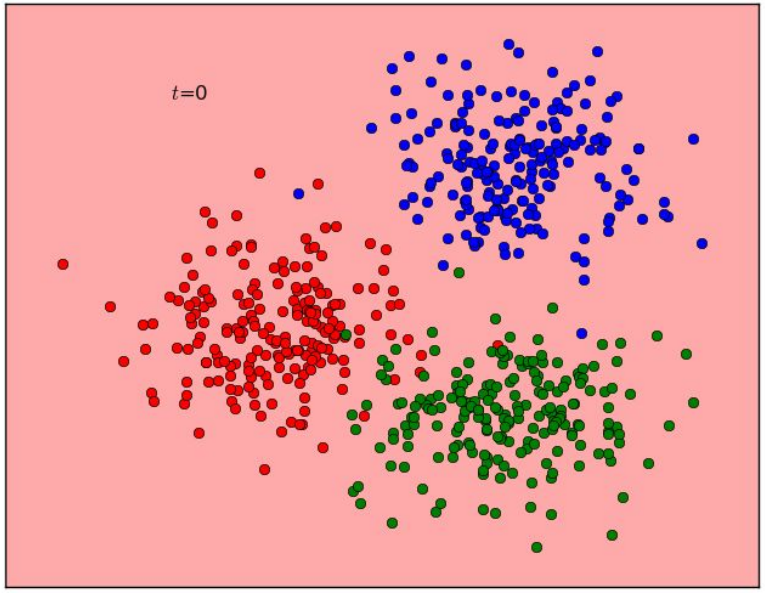
Classification (Supervised Learning)





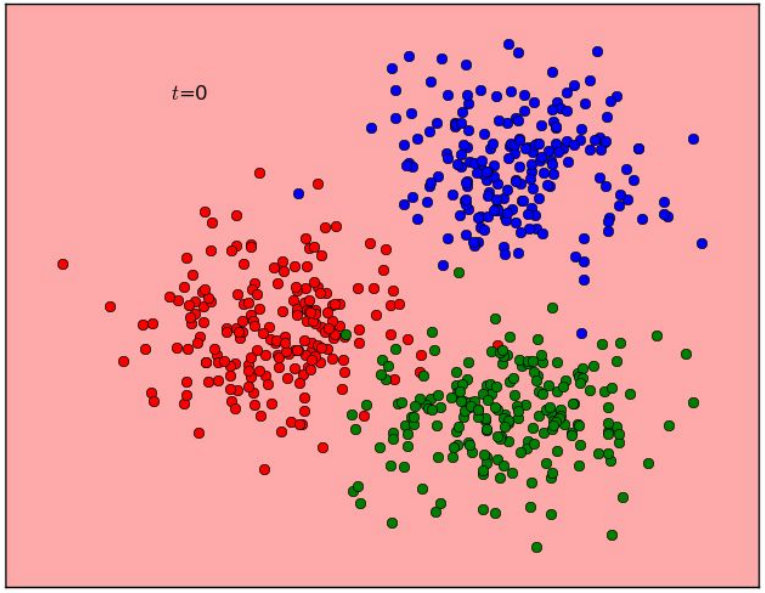
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

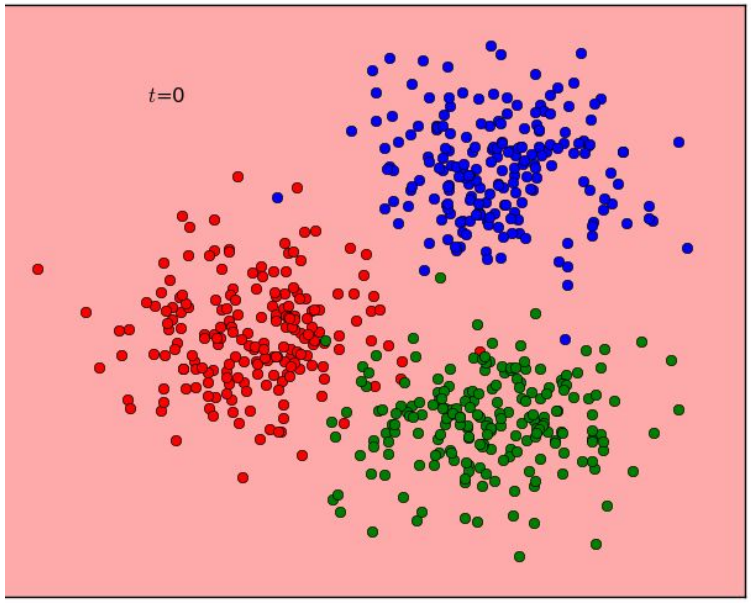
Classification (Supervised Learning)



Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

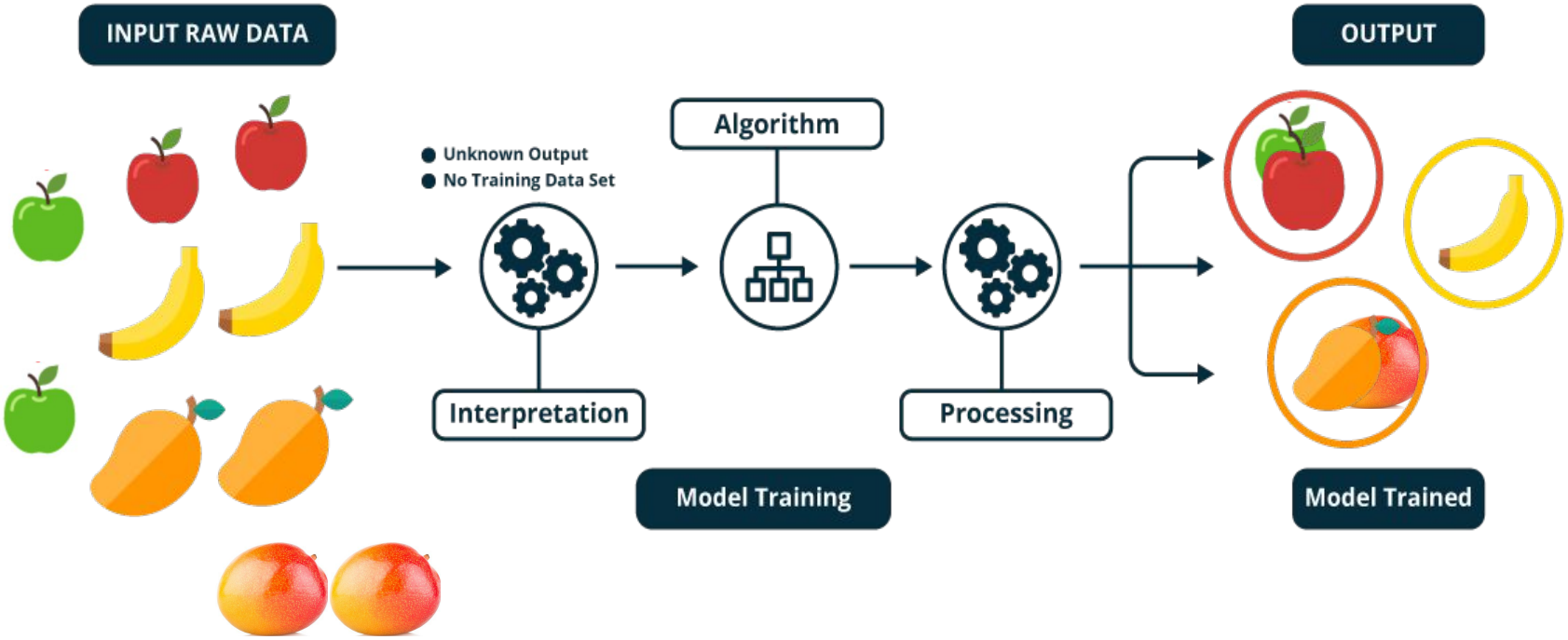
Classification (Supervised Learning)





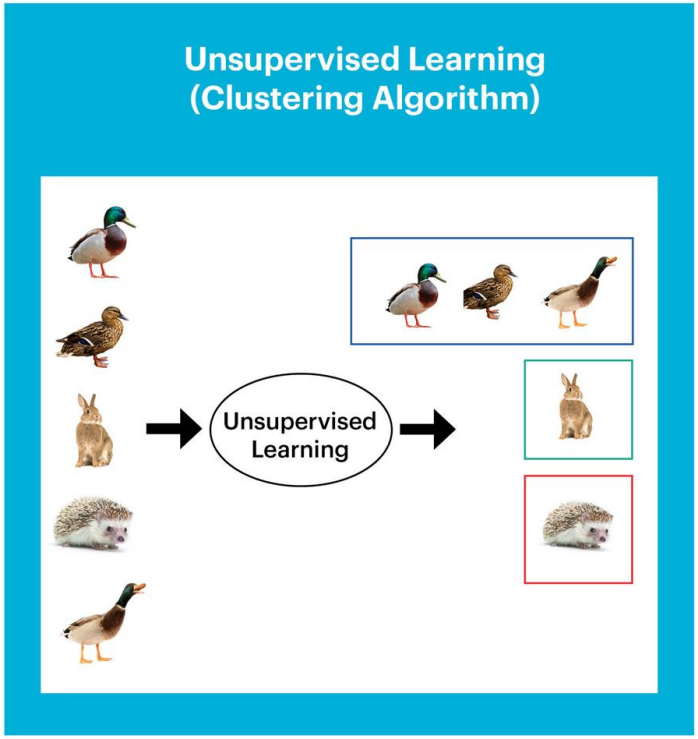
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Clustering (Unsupervised Learning)



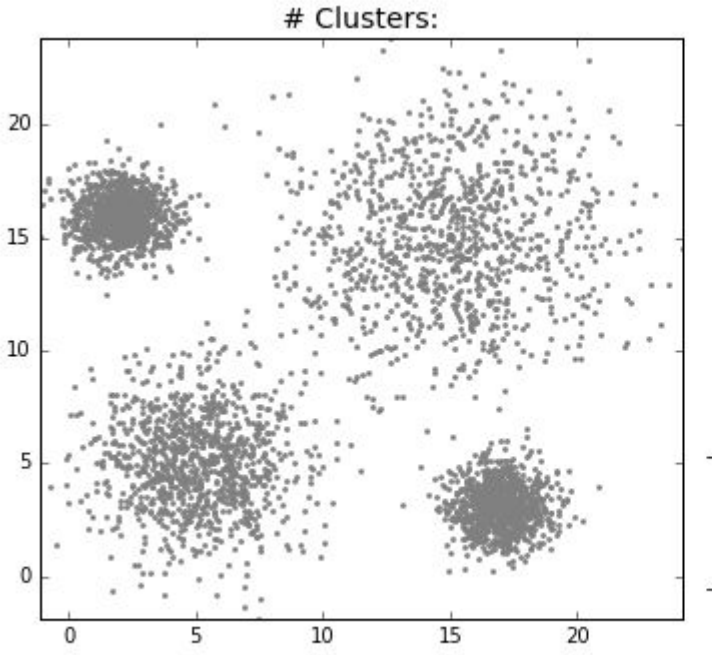
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Clustering (Unsupervised Learning)



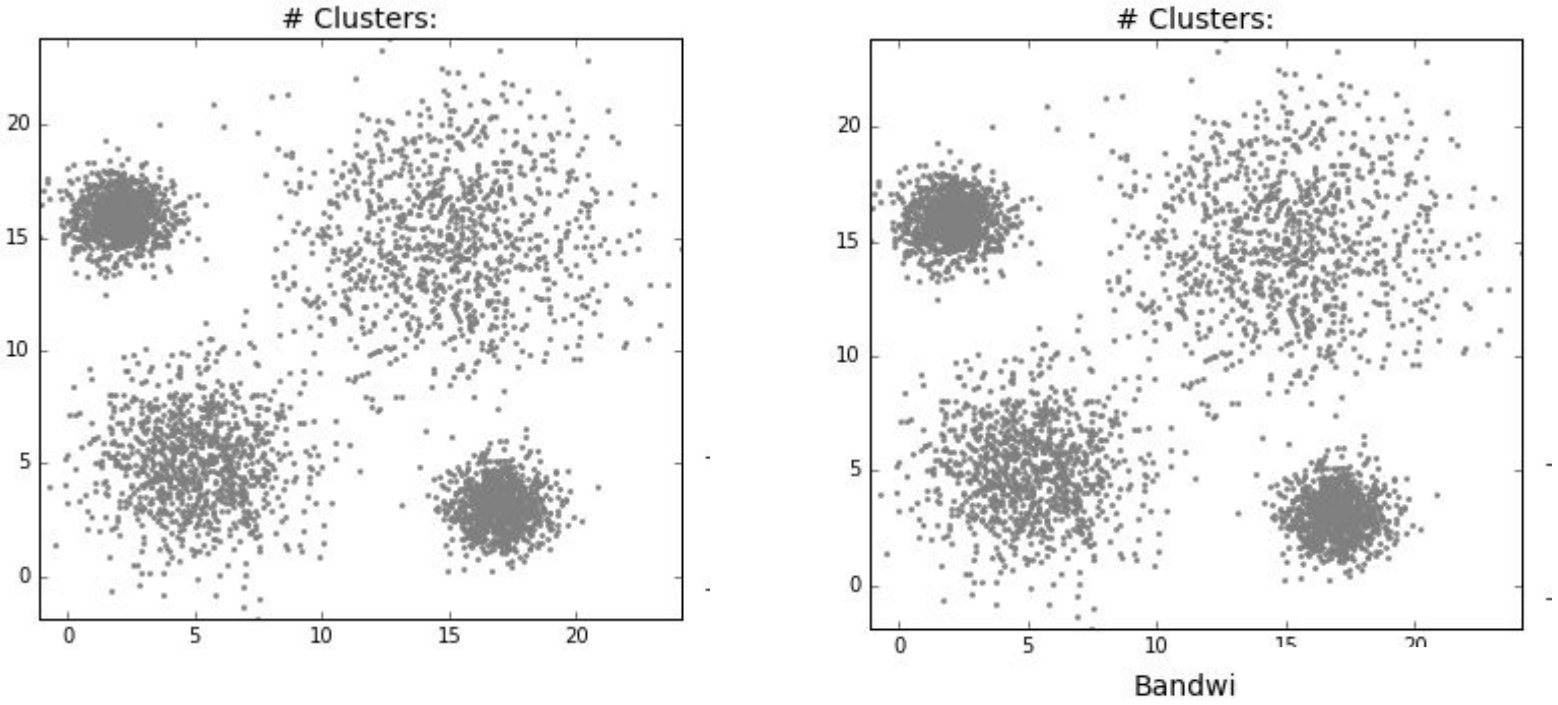
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Clustering (Unsupervised Learning)



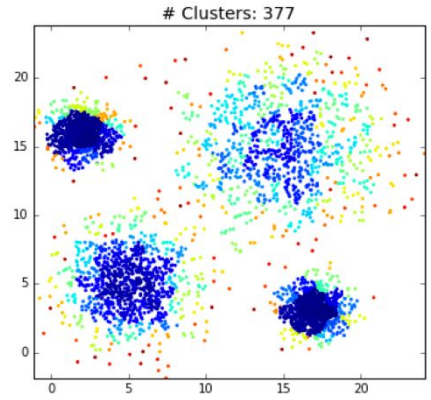
Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

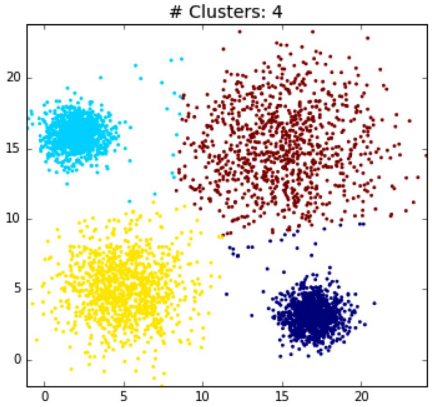
Clustering (Unsupervised Learning)

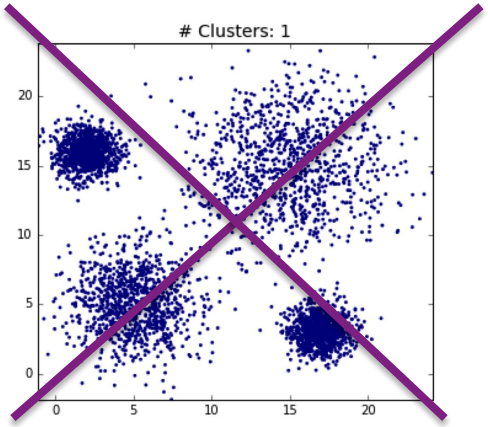


Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Clustering (Unsupervised Learning)

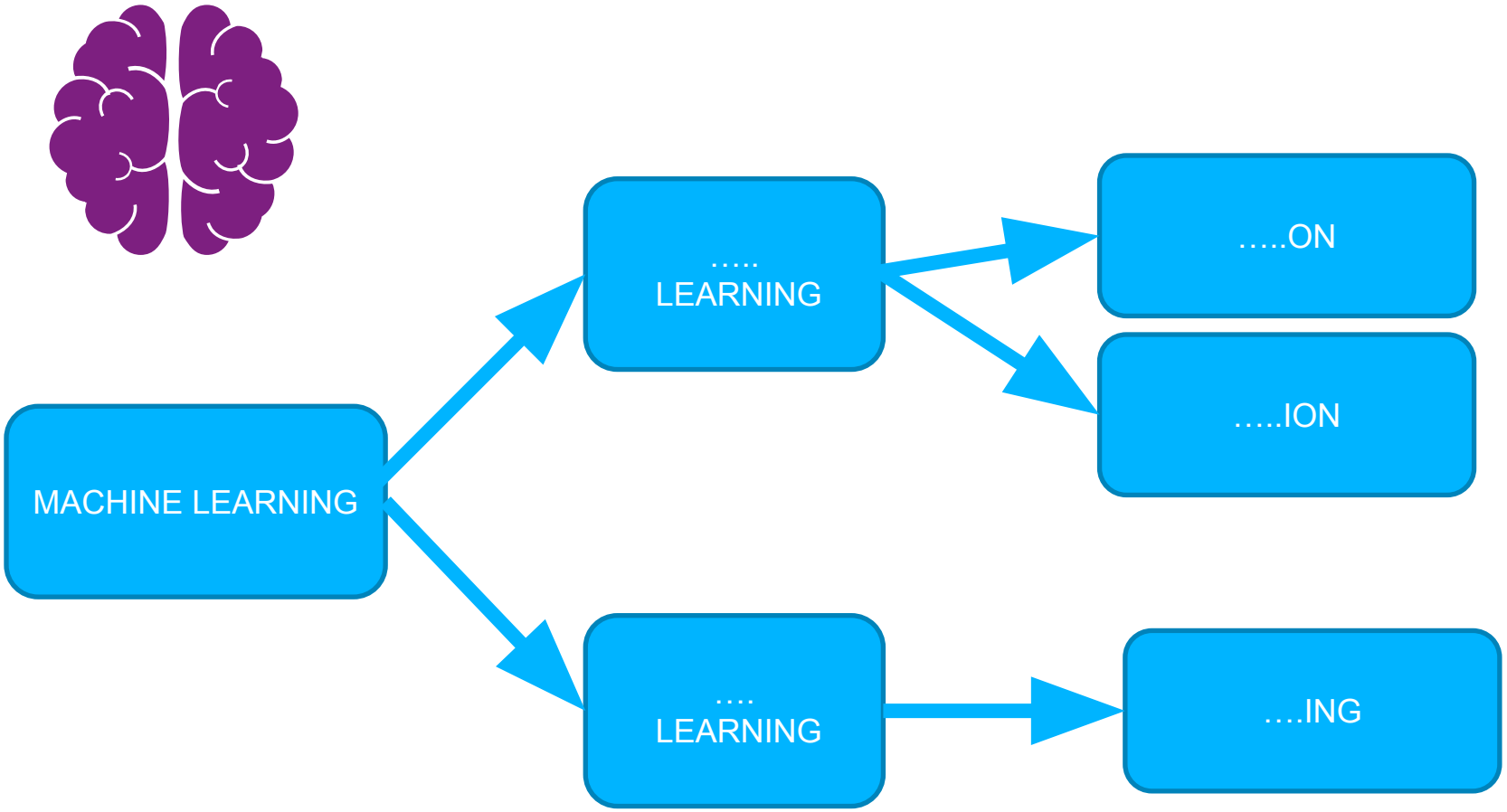






Source:https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

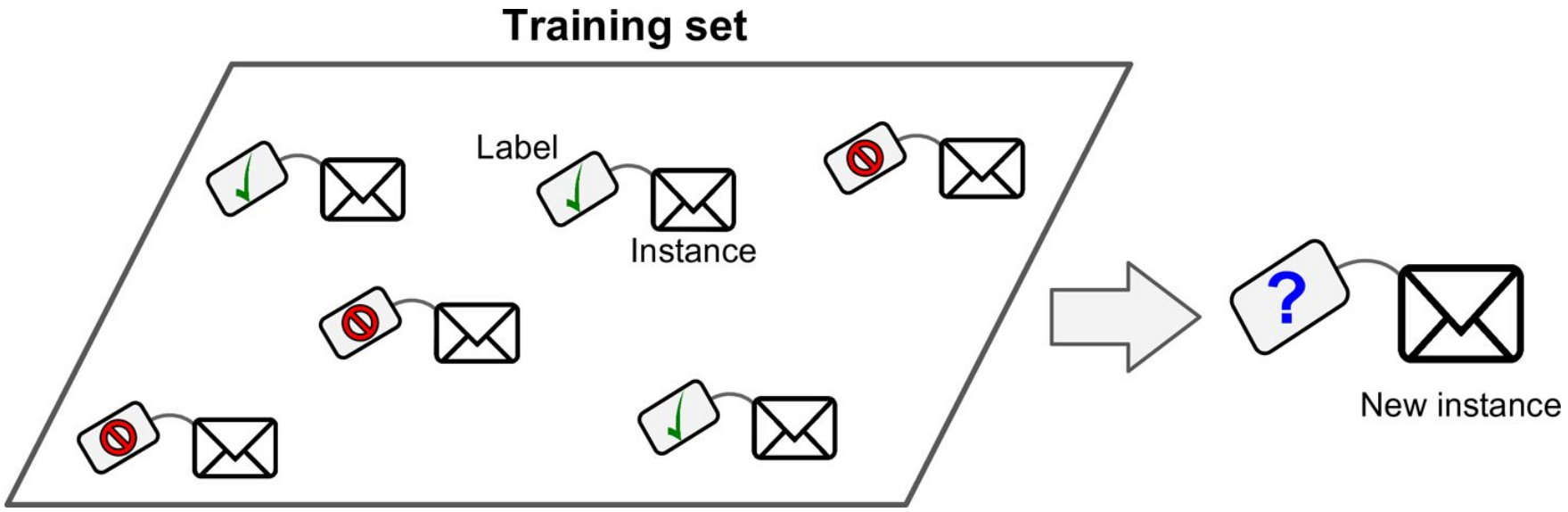
Look Again & Fill In the Blank!



Types of Machine Learning

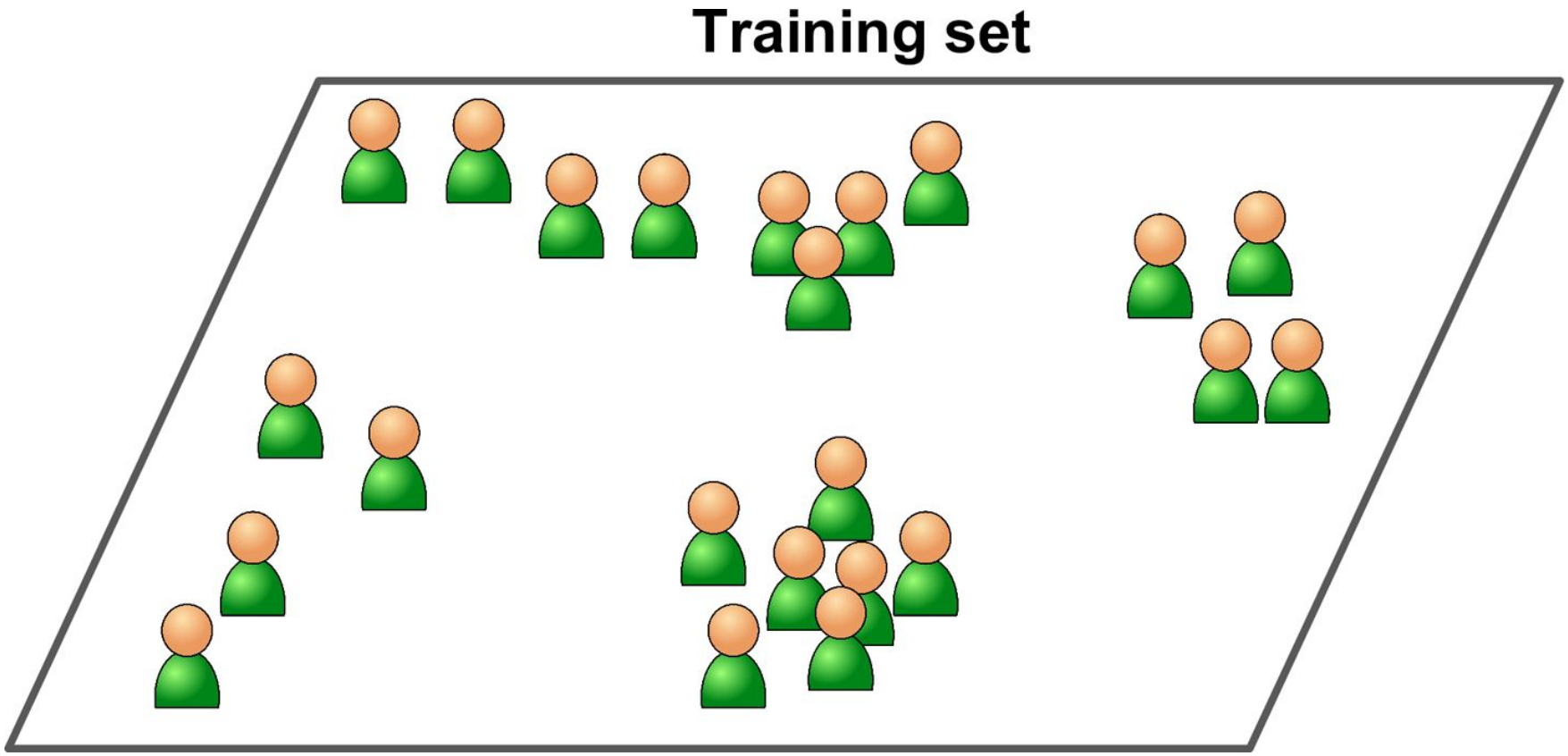
- Whether or not they are trained with human supervision (Supervised, Unsupervised, Semi-supervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)
- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model (instance-based versus model-based learning)

Quiz: Supervised or not?



The training data you feed to the algorithm includes the desired solutions, called labels

Quiz 2: Supervised or not?

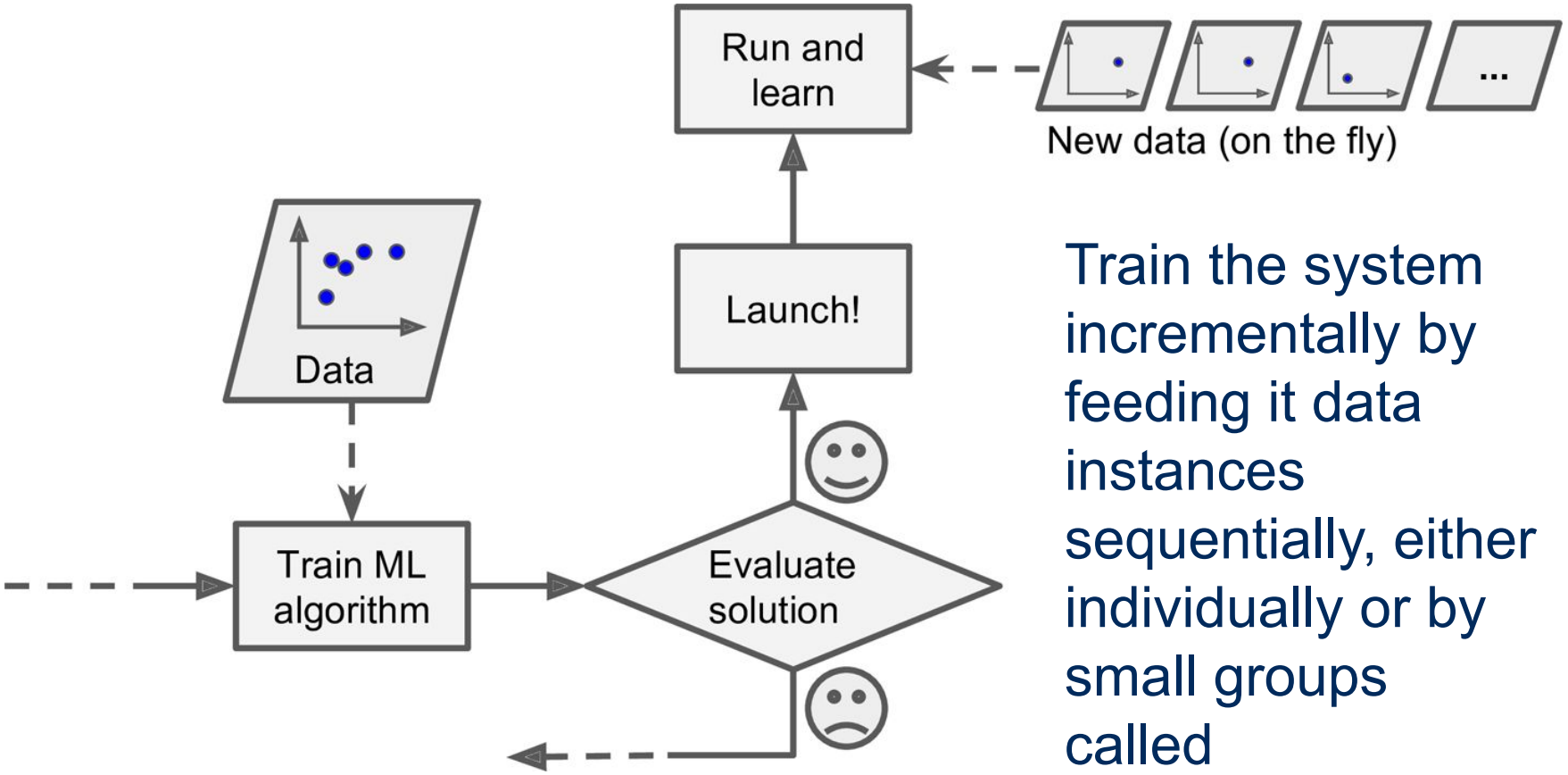


Data is unlabeled

Batch Learning

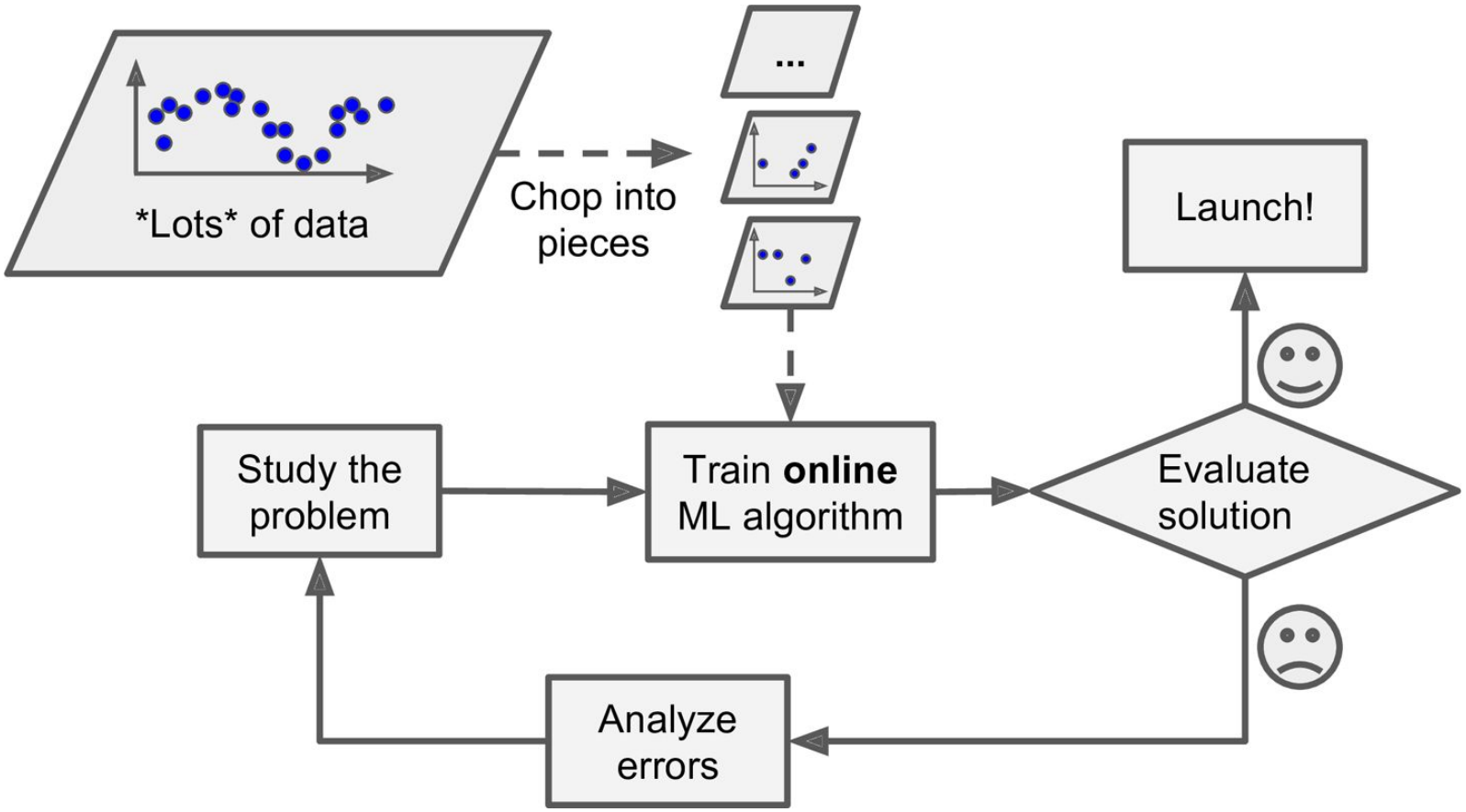
- Train model using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline
- Train algorithm from scratch using new + old data
- Can be automated

Online Learning

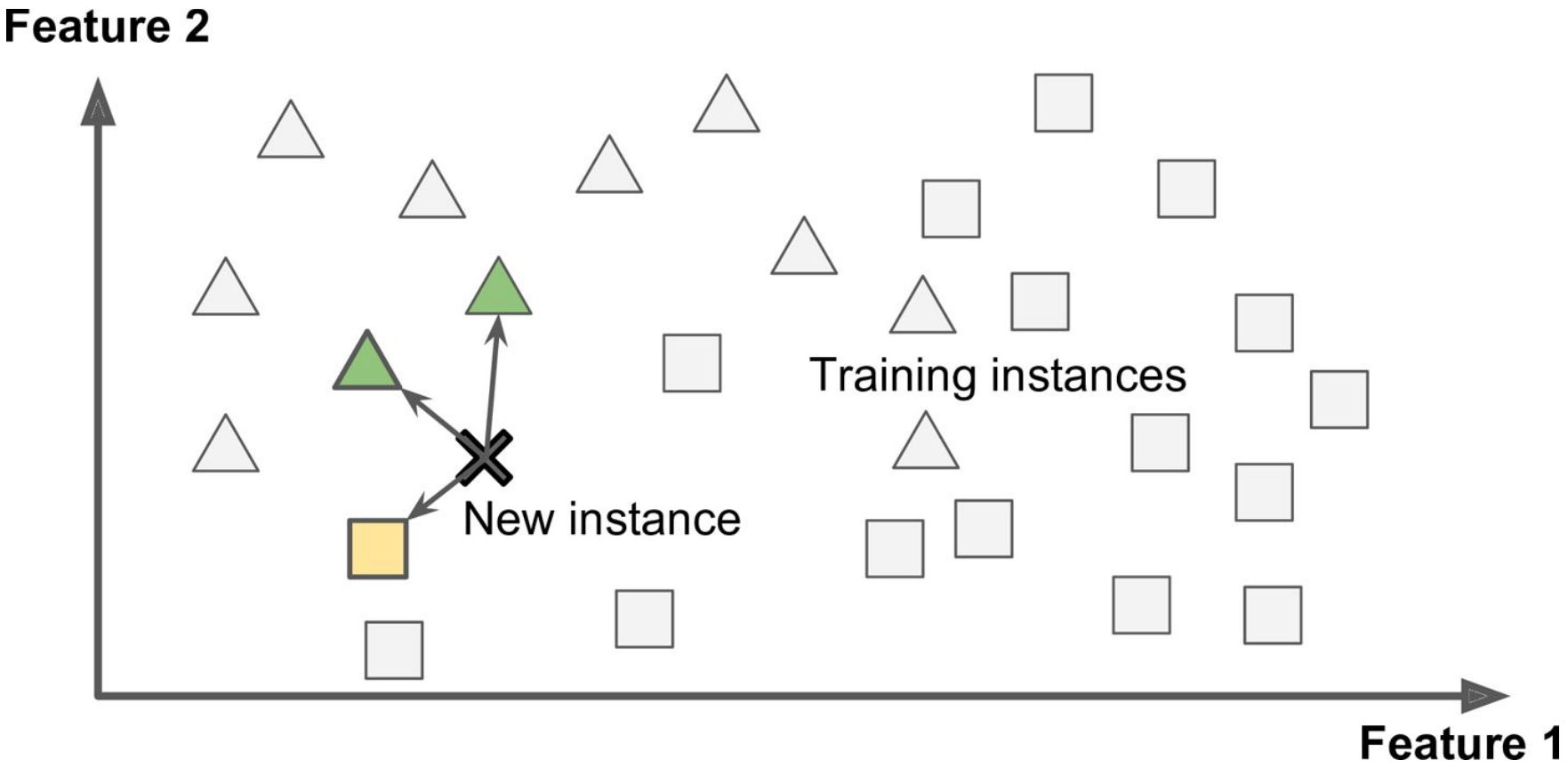


Train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.

Online Learning for Large Datasets



Instance Based Learning

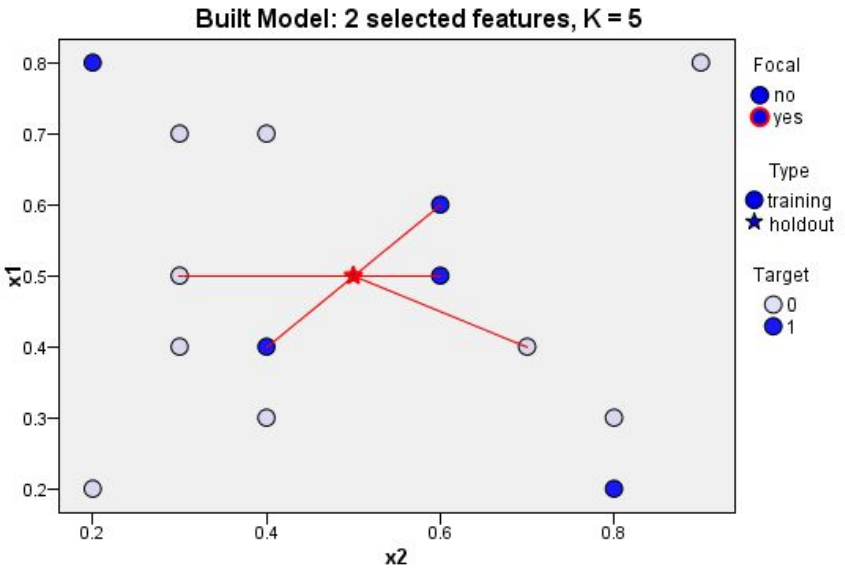


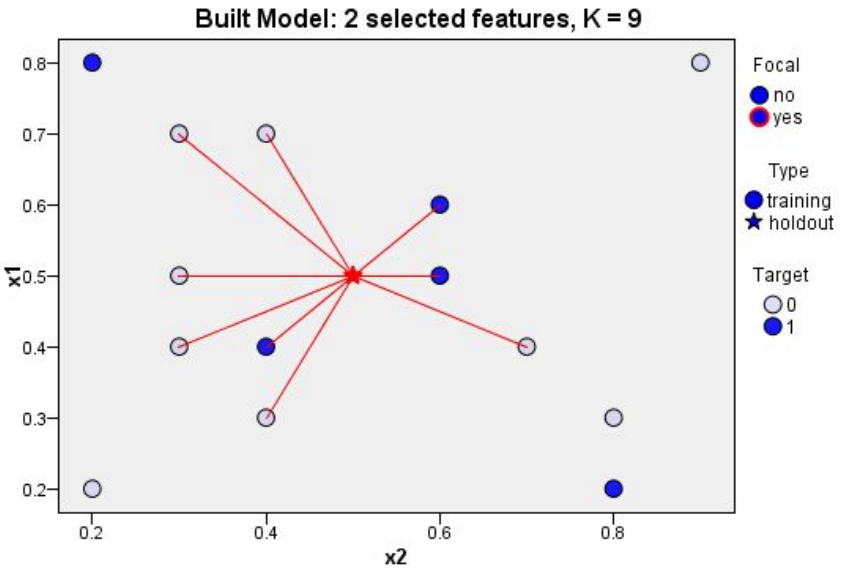
This is called instance-based learning: the system learns the examples by heart, then generalizes to new cases using a similarity measure

Instance-based Methods

- Methods for approximating discrete-valued or real-valued target functions (classification or regression)
- A new instance gets a classification equal to the classification of the nearest instance
- Learning becomes tied to data storage
- Assumptions:
  - Output varies smoothly with input
  - Non-prior model assumption

How K-Nearest Neighbor Works?





Nearest Neighbors

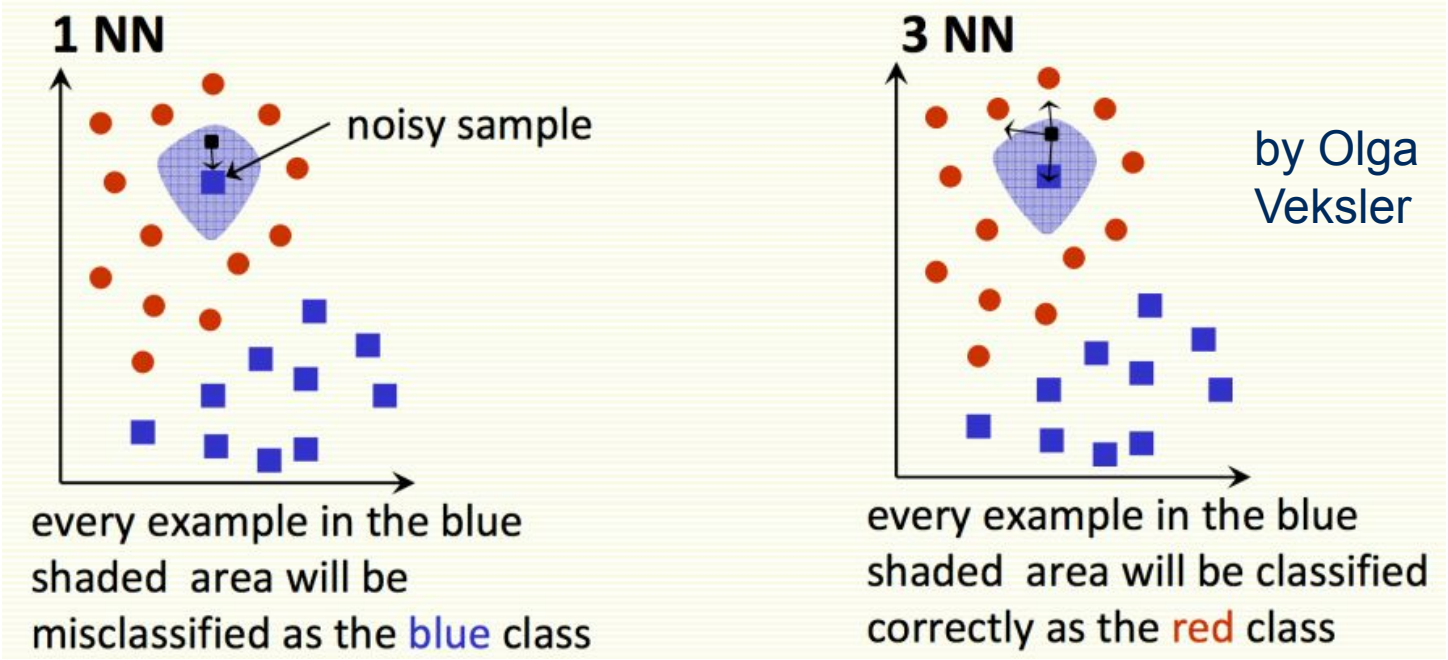
- Training examples correspond to points in d-dim space
- The value of the target function for a new query is estimated from the known value(s) of the nearest training example(s)
- Euclidean distance:

$$||\mathbf{x}^{(a)} - \mathbf{x}^{(b)}||_2 = \sqrt{\sum_{j=1}^{d} (x_j^{(a)} - x_j^{(b)})^2}$$

Choosing k

- Larger k may lead to better performance
- But if we set k too large we may end up looking at samples that are not neighbors
- We can use cross-validation to find k
- Rule of thumb is k < sqrt(n), where n is the number of training examples

Mislabeled Data

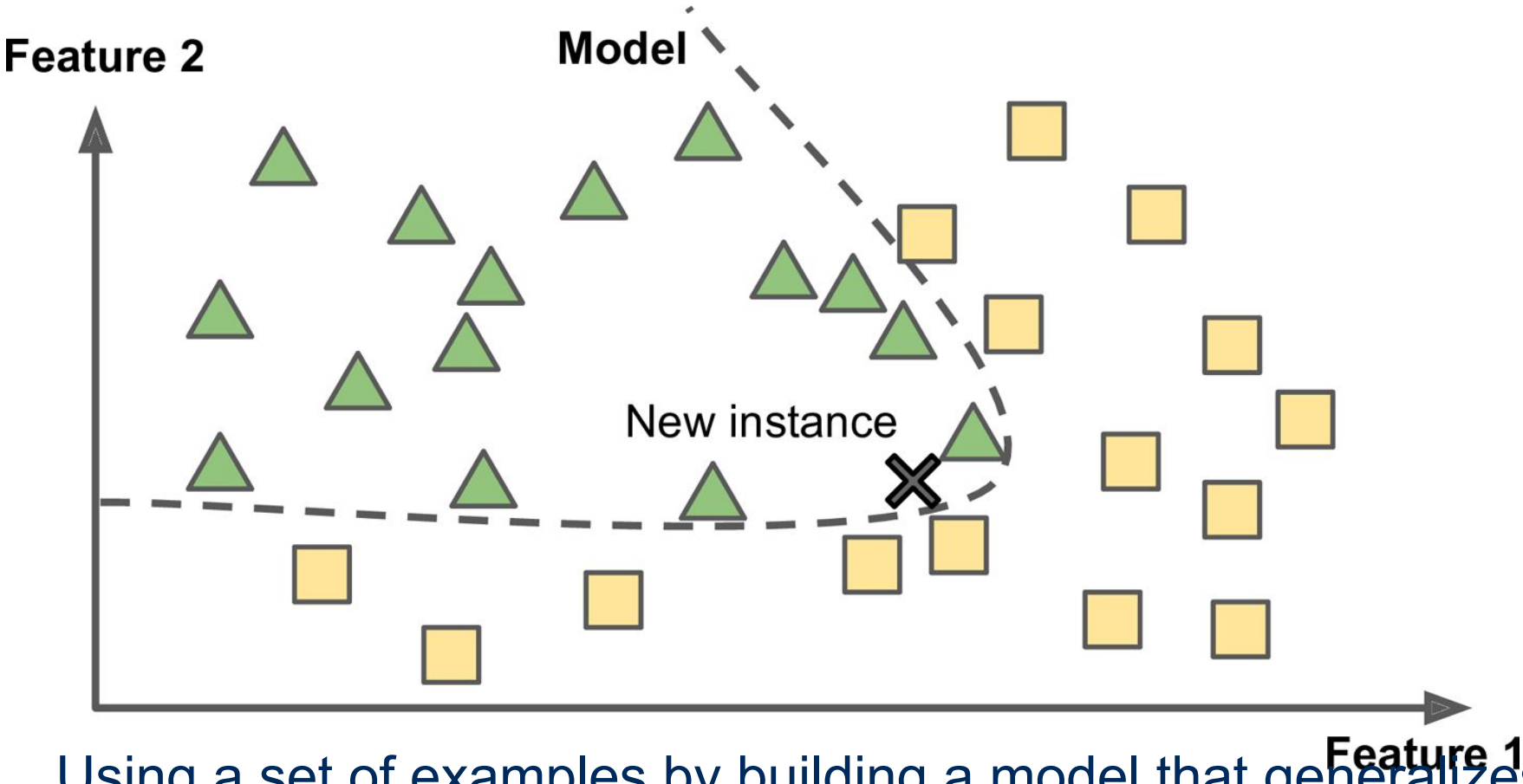


- Nearest neighbors sensitive to mislabeled data ("class noise"). Solution?
  - Smooth by having k nearest neighbors vote

KNN Issues and Solutions

- Features have different dimensions
  - Rescale data
- Irrelevant correlated features
  - Eliminate features, add weights to distance
- Non-metric attributes
  - Use alternative metric: Edit distance (Hamming)
- Expensive at test time
  - Subset of dimension, kd-trees (sort), approximate distance, remove redundant data
- Storage requirements
  - Condense data

Model Based Learning



Using a set of examples by building a model that generalizes these examples, then use that model to make predictions.

Module 1 – Section 4

Modeling

What is a Model?

- Representation of a reality
- Reality can be car, house or any phenomena
- Representation can be **abstract** or **concrete**
- Representation can be math/stat or non-math/stat
- We focus on math/stat representation

Can you give an example of a model you know since school days ?
grade 7 or 8

Mathematical or Statistical Model

- Calculate area of a circle using its radios?

Calculate area of a circle using its radios:

$A = \pi r^2 = 3.14 * r^2$

Example:

Calculate the area of a circle with radios = 6

$A = 3.14 * 6^2 = 3.14 * 36 = 113.04$

Mathematical or Statistical Model

- Often describe relationship between variables

- Types

    1. **Deterministic** models (With No uncertainty, No error in prediction/estimation)
    2. **Probabilistic** models (With uncertainty, always has some error)

Deterministic Models

- Models exact relationships
- Suitable when prediction error is negligible
- Example: Body mass index (BMI) is measure of body fat based
- Example

$$BMI = \frac{Weight\ in\ Kilograms}{(Height\ in\ Meters)^2}$$

Probabilistic Models

1. Hypothesize 2 Components
   - Deterministic
   - Random Error

2. Example: Rent in a condo unit in downtown Toronto is a function of **size in sft, number of bedroom, and bathroom** plus some other things (level, view, amenities - hard to quantify)

$$\textit{Rent} = 1.5 * \text{size} + 400 * \text{bed} + 300 * \text{bath} + \varepsilon$$

- Random error ε explains variation that is not explained by used features

Question

- Predict the rent for a 1 bedroom, 1 bath apartment, 999 sft in downtown Toronto?

$$\textit{Rent} = 1.5 * \text{size} + 400 * \text{bed} + 300 * \text{bath} + \varepsilon$$

Question

- Predict the rent for a 1 bedroom, 1 bath apartment, 99 m^2 in downtown Toronto?

$$\textit{Rent} = 1.5 * \text{size} + 400 * \text{bed} + 300 * \text{bath} + \varepsilon$$

A modeling question

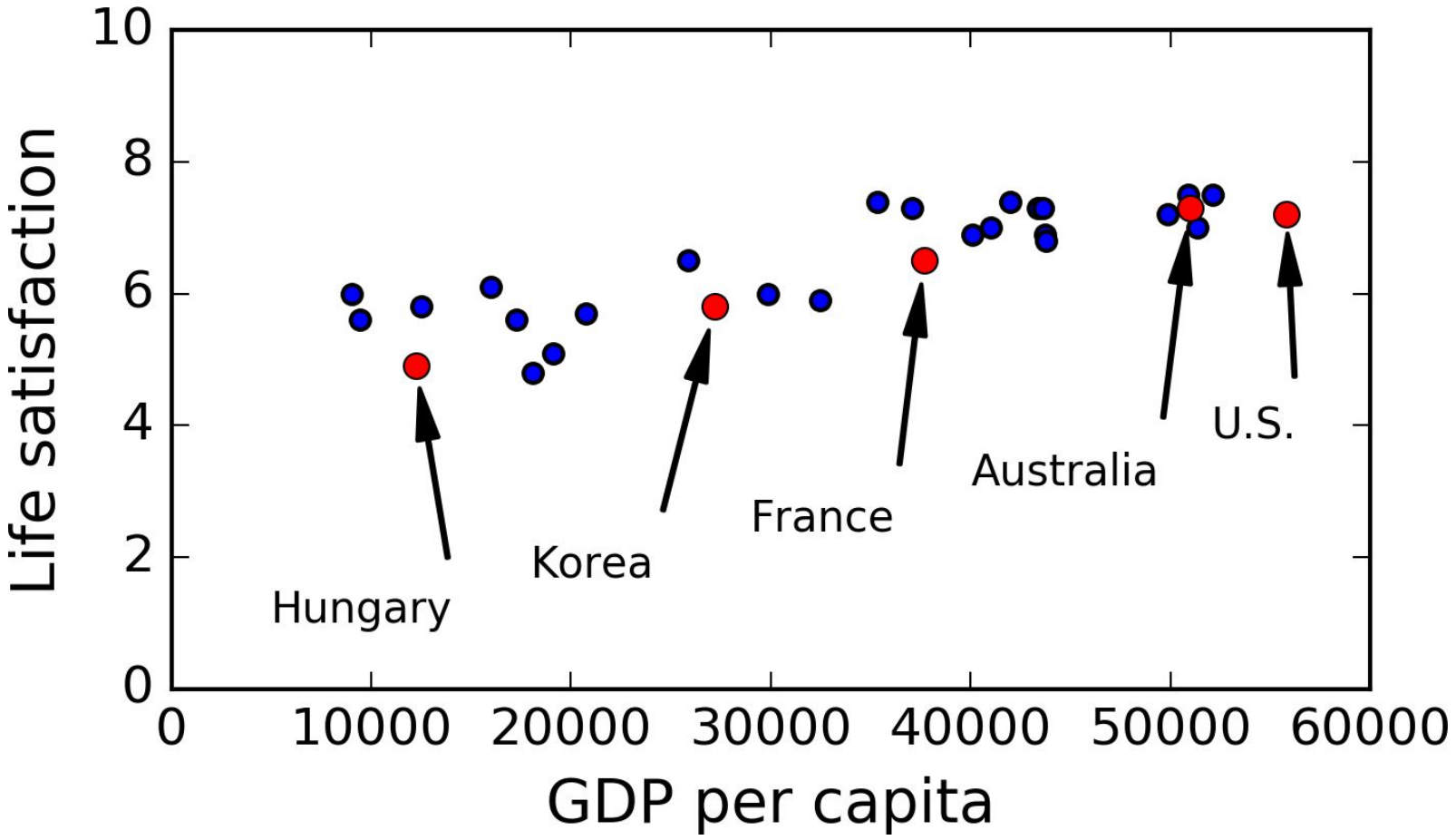
Is it possible to predict the life satisfaction of people across the globe?

A modeling question

Is it possible to predict the life satisfaction of people across the globe?

Yes! if we can find a feature that is able to explain the relationship with happiness

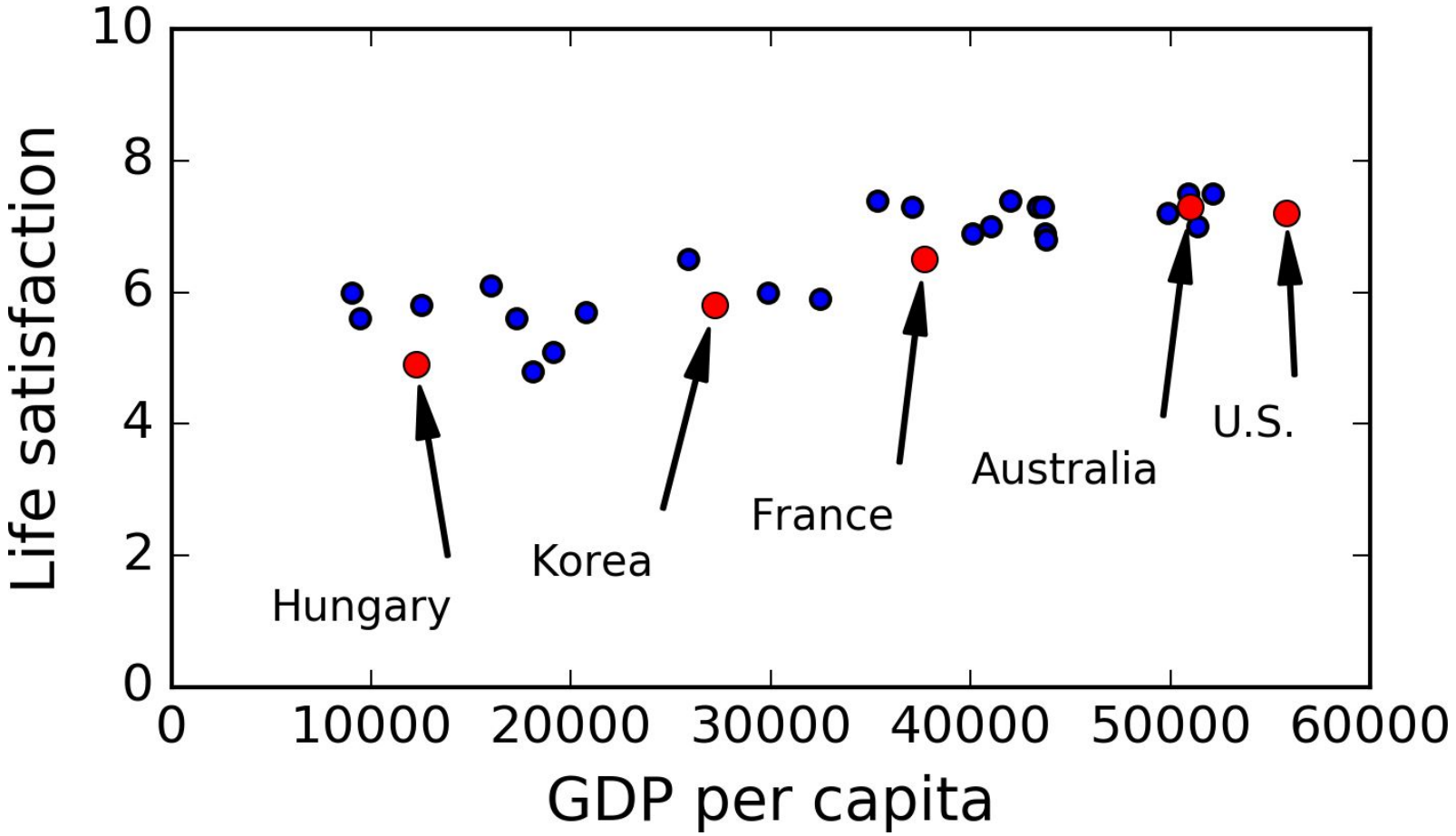
Is There a Trend?



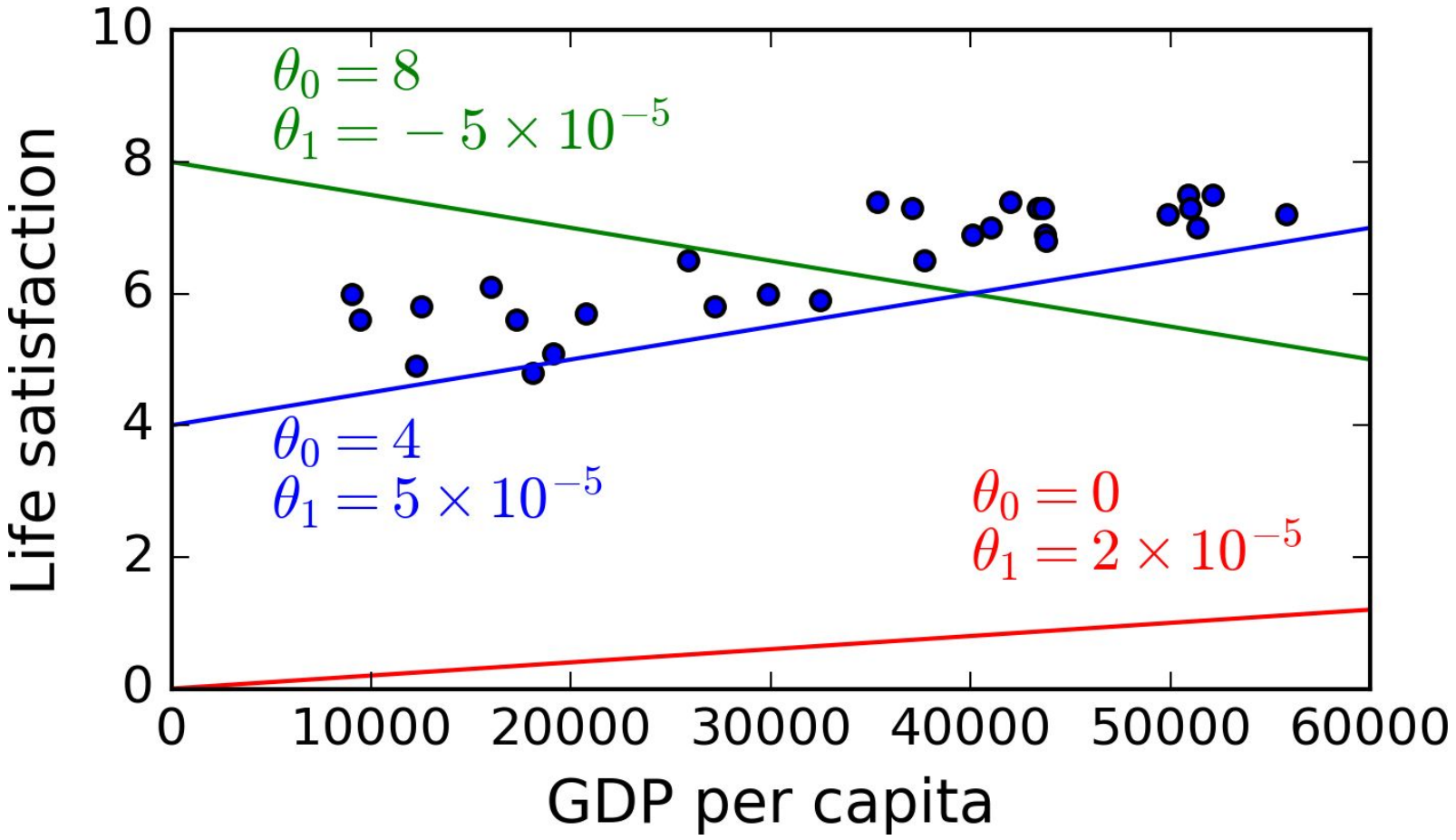
What do you observe?

Build Model

$$life\_satisfaction = \theta_0 + \theta_1 \times GDP\_per\_capita$$



Possible Models

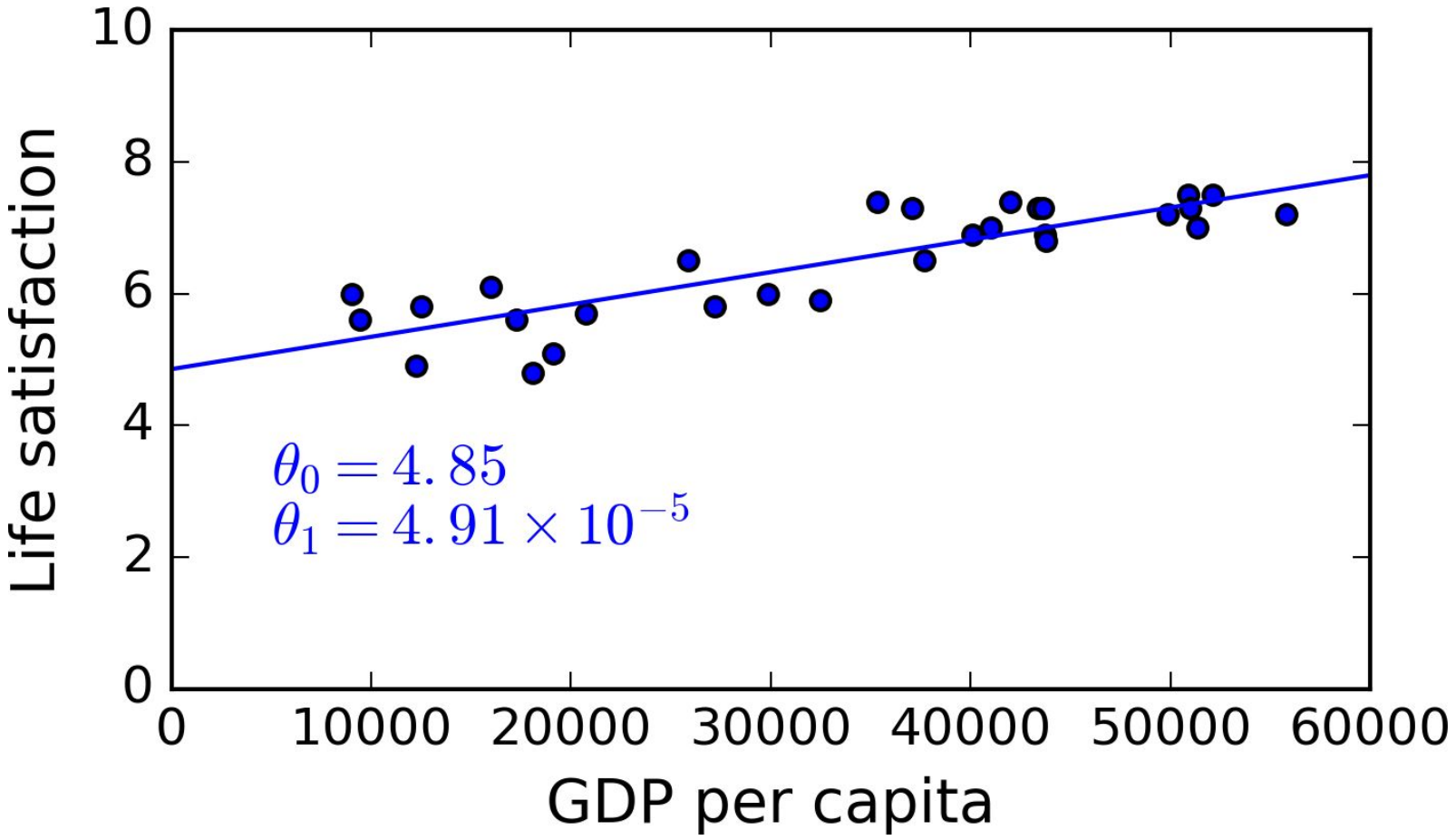


Build Model (cont'd)

$$life\_satisfaction = \theta_0 + \theta_1 \times GDP\_per\_capita$$

- How to define parameters?
- Specify a performance measure or a cost function.
  - i.e. Measure distance between examples and model's prediction
- Find optimal parameters that minimize the cost function

Best Model


Life satisfaction
GDP per capita
$\theta_0 = 4.85$
$\theta_1 = 4.91 \times 10^{-5}$
0
2
4
6
8
10
0
10000
20000
30000
40000
50000
60000

Coding

```
import matplotlib
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import sklearn

# Load the data
oecd_bli = pd.read_csv("oecd_bli_2015.csv", thousands=',')
gdp_per_capita = pd.read_csv("gdp_per_capita.csv",thousands=',',del
                             encoding='latin1', na_values="n/a")

# Prepare the data
country_stats = prepare_country_stats(oecd_bli, gdp_per_capita)
X = np.c_[country_stats["GDP per capita"]]
y = np.c_[country_stats["Life satisfaction"]]

# Visualize the data
country_stats.plot(kind='scatter', x="GDP per capita", y='Life sati
plt.show()

# Select a linear model
lin_reg_model = sklearn.linear_model.LinearRegression()

# Train the model
lin_reg_model.fit(X, y)

# Make a prediction for Cyprus
X_new = [[22587]]  # Cyprus' GDP per capita
print(lin_reg_model.predict(X_new)) # outputs [[ 5.96242338]]
```

ML Process Summary

- You studied the data.
- You selected a model.
- You trained it on the training data (i.e., the learning algorithm searched for the model parameter values that minimize a cost function).
- Finally, you applied the model to make predictions on new cases (this is called inference), hoping that this model will generalize well.

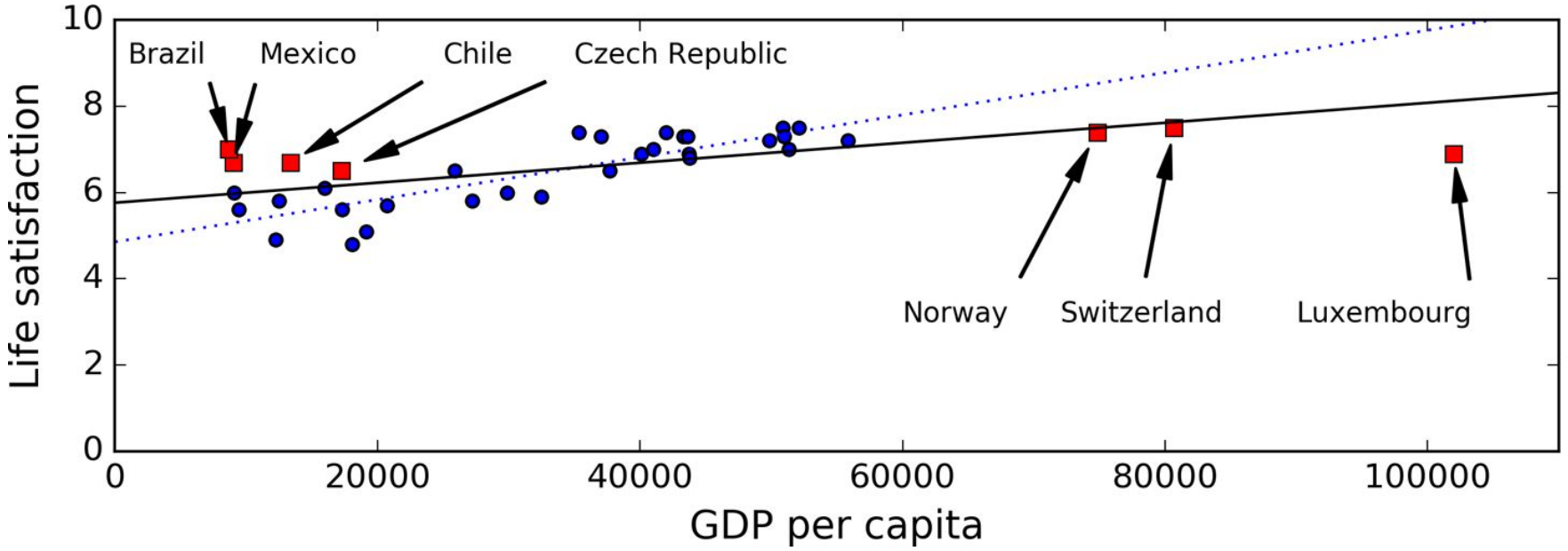
Module 1 – Section 5

Challenges of Machine Learning

Common ML Challenges

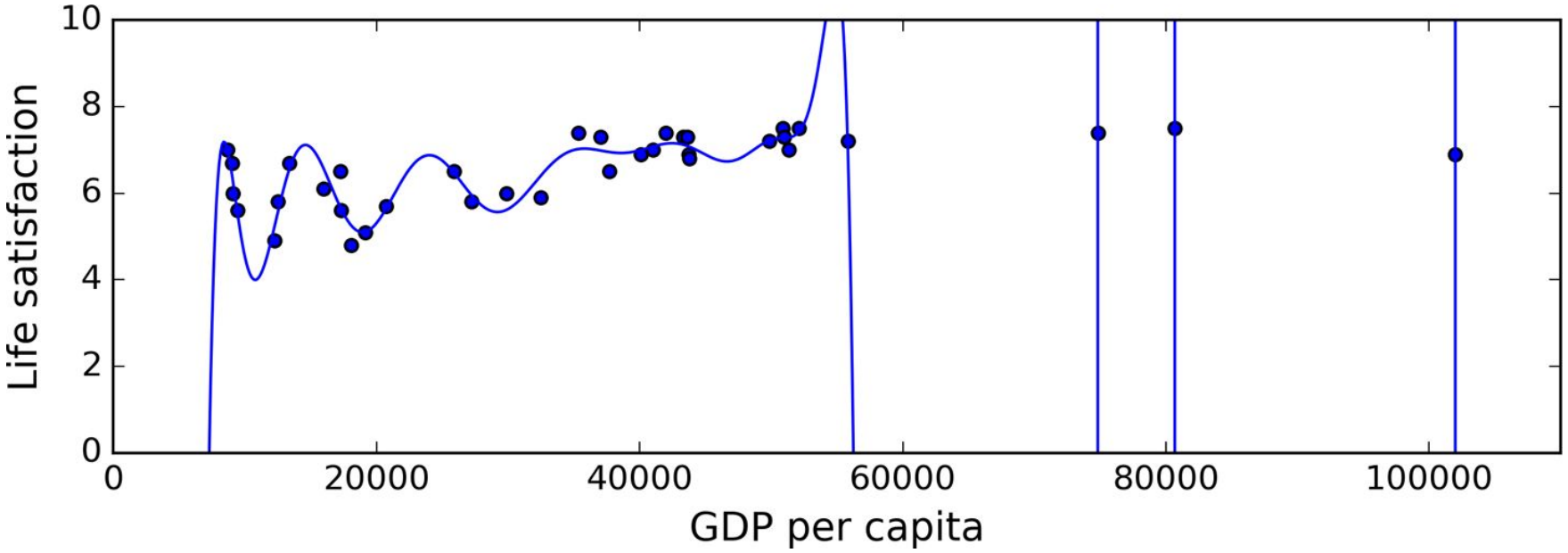
- Insufficient Quantity of Data
- Non representative Training Data
  - New data has different structure
- Poor-Quality Data
  - Missing data, outliers
- Irrelevant Features
  - Features selection, extraction
- Overfitting or Underfitting
- Heavy Hidden Technical Debt

Representative Data



In order to generalize well, it is crucial that your training data be representative of the new cases you want to generalize to. This is true whether you use instance-based learning or model-based learning.

Data Overfitting

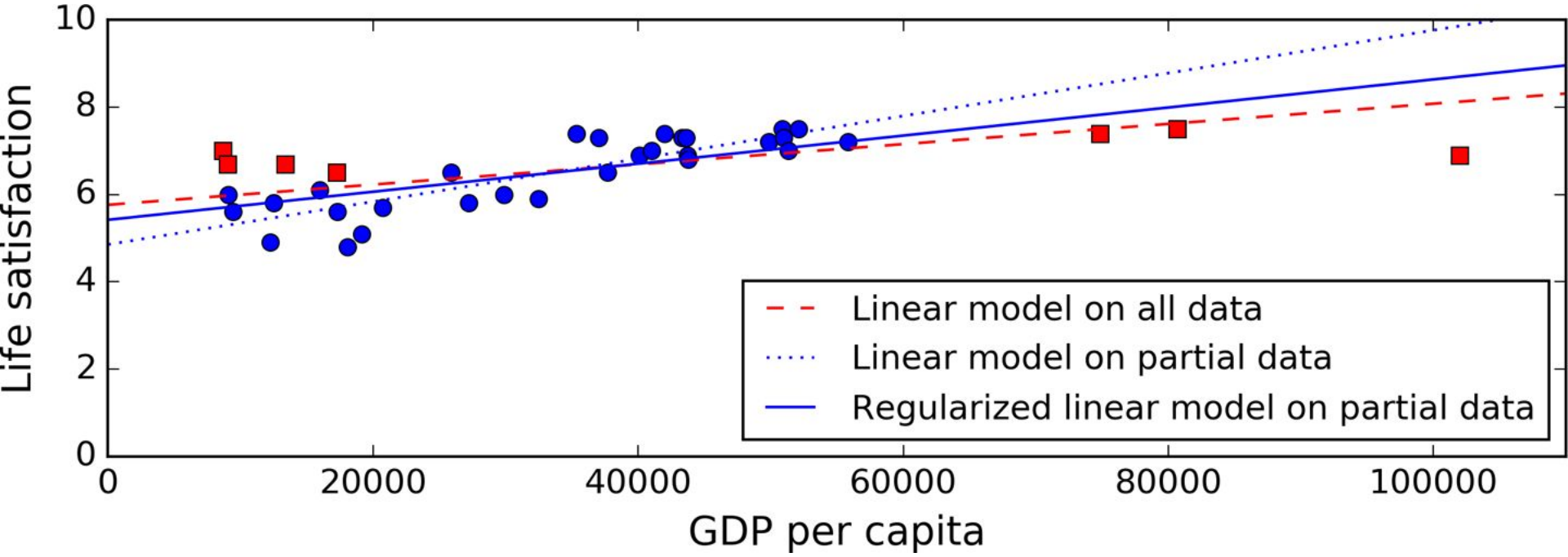


Overfitting means that the model performs well on the training data, but it does not generalize well.

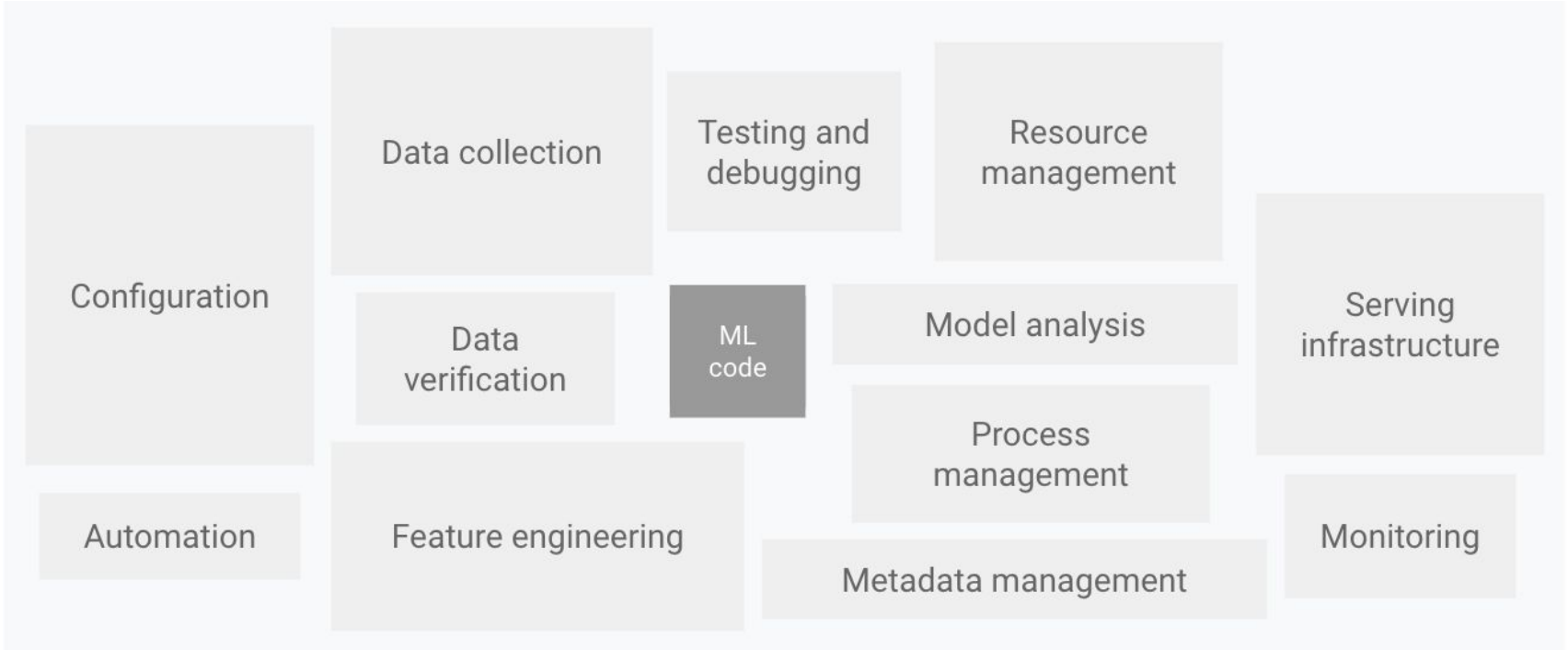
Solutions:

- Simplify model
- Gather more data
- Reduce noise (fix data)

Regularization Data



Hidden Technical Debt in Machine Learning Systems



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Module 1 – Section 6

Recap

Recap

- Machine Learning is about making machines get better at some task by learning from data, instead of having to explicitly code rules.
- There are many different types of ML systems: supervised or not, batch or online, instance-based or model-based, and so on.
- In a ML project you gather data in a training set, and you feed the training set to a learning algorithm.

So Far … (cont'd)

- If the algorithm is model-based it tunes some parameters to fit the model to the training set (i.e., to make good predictions on the training set itself), and then hopefully it will be able to make good predictions on new cases as well.
- If the algorithm is instance-based, it just learns the examples by heart and uses a similarity measure to generalize to new instances.
- The system will not perform well if your training set is too small, or if the data is not representative, noisy, or polluted with irrelevant features (garbage in, garbage out).
- Lastly, your model needs to be neither too simple (in which case it will underfit) nor too complex (in which case it will overfit).

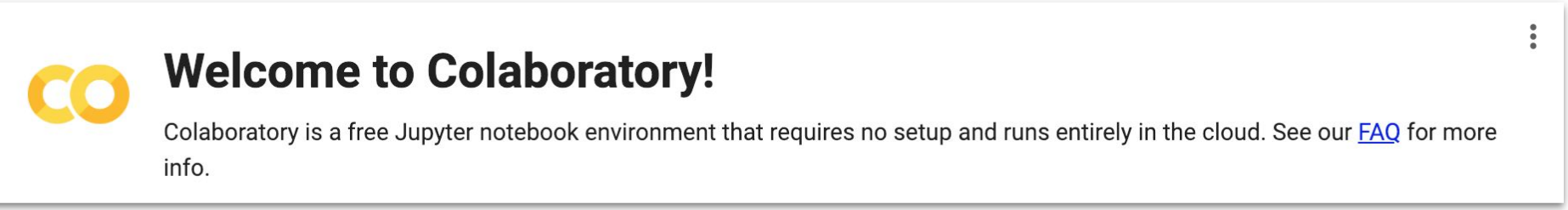
UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 1 – Section 7

Tools & Techniques

Google Colaboratory!

- Google Colaboratory Notebook
- Jupyter Notebook which is easier to share
- Online
- Free Cloud Machine
- Local Execution



sklearn

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD licence
- http://scikit-learn.org/stable/

NumPy

- NumPy is the fundamental package for scientific computing with Python. It contains among other things:
  - a powerful N-dimensional array object
  - sophisticated (broadcasting) functions
  - tools for integrating C/C++ and Fortran code
  - useful linear algebra, Fourier transform, and random number capabilities
- NumPy Quickstart tutorial

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 1 – Section 8

Resources and Wrap-up

Homework

- Complete assignment 1
- 5 multiple choice questions for 10 mark.
- Make sure you have reviewed the contents before attempting
- There is only one chance
- You will get score immediately after submission
- You may access the test from here

Next Class

- End to End Machine Learning Project
- Reading: Chapter 2 textbook

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies