Machine Learning, Spring 2020

Course Overview

Instructor:

Yi Fang (yfang@nyu.edu)

Teaching Assistants:

Tadesse Mulugeta <tsm8@nyu.edu>

Jing Zhu <jingzhu@nyu.edu>

Shuaihang Yuan <sy2366@nyu.edu>

Note: course materials are prepared based 1) NYU Prof. Daniel Neill's Machine Learning 2) Textbook of this course

Why machine learning?

Critical importance of addressing urban challenges: disease, crime, terrorism, poverty, environment…

Increasing size and complexity of available data, thanks to the rapid growth of new and transformative technologies.

Much more computing power, and scalable data analysis methods, enable us to extract actionable information from all of this data.

Machine learning techniques have become increasingly essential for data analysis, and for the development of new, practical information technologies that can be directly applied to address critical challenges in many areas (Engineering, Science, Social Science, Economics and many more)

Some motivating examples

Early detection of emerging disease outbreaks

Discovering new "best practices" for patient care

Substance abuse and overdose surveillance

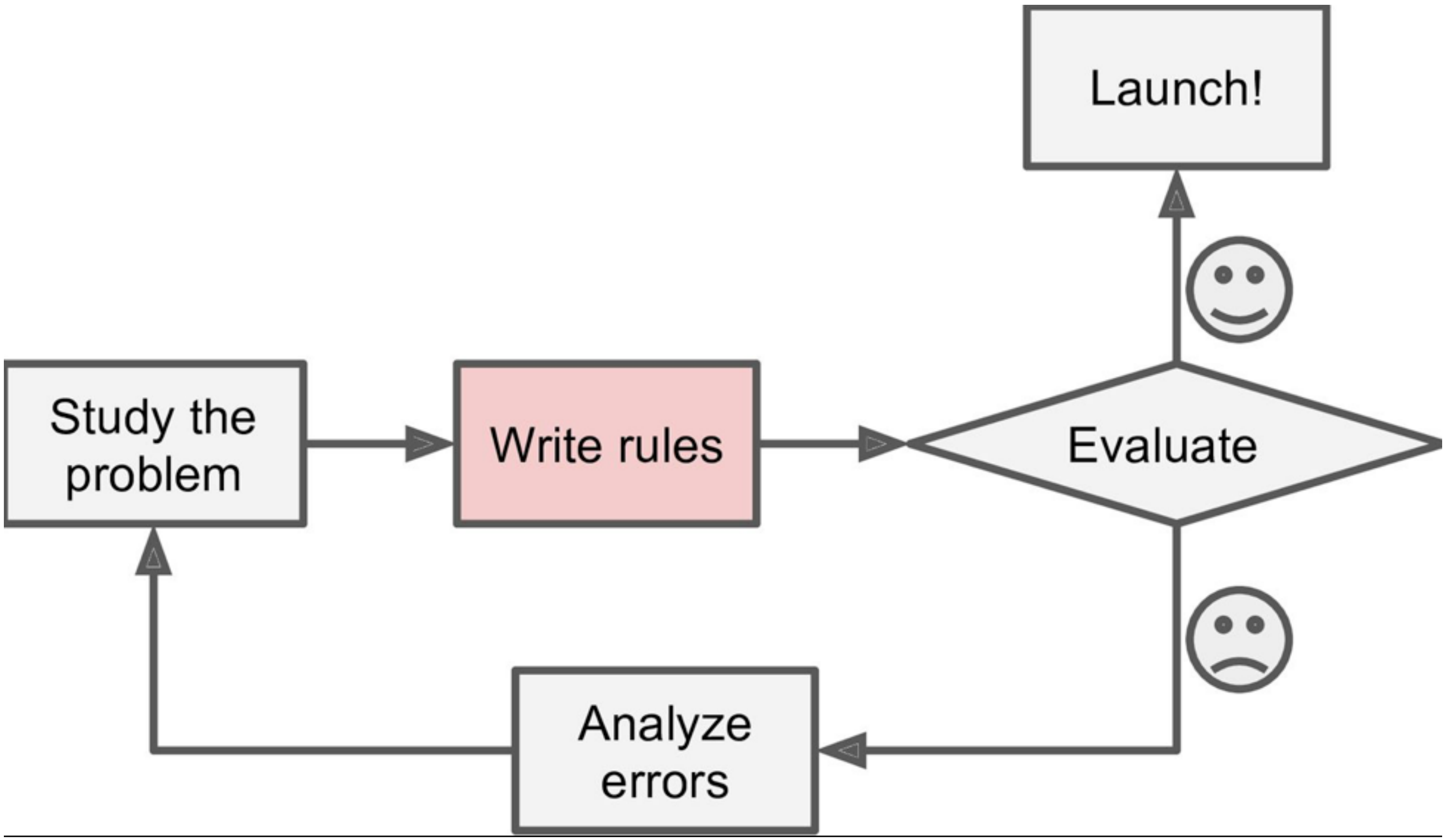
Preventing rat infestations (using "311" service calls)

Predicting civil unrest (using Twitter data)

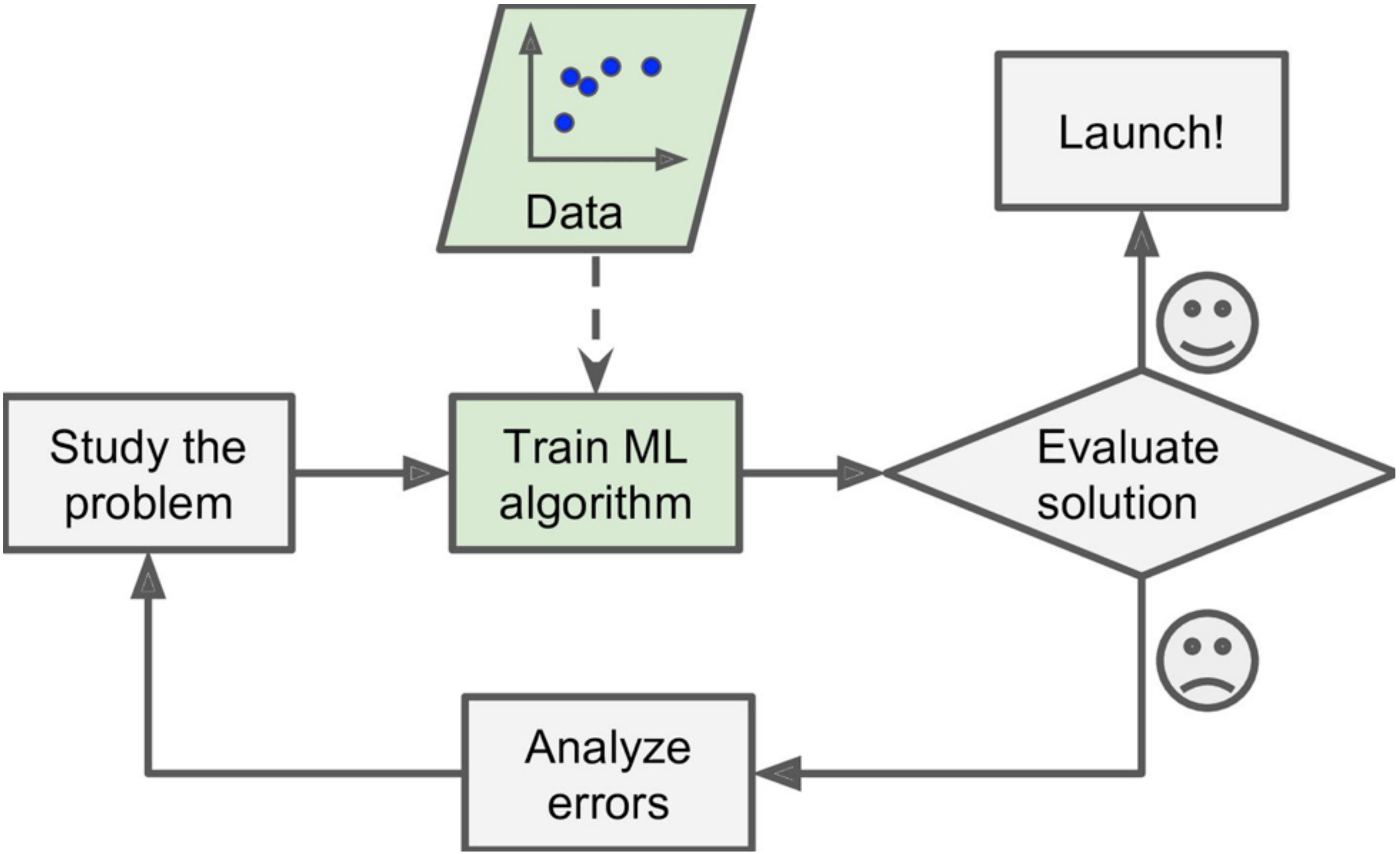
Preventing violent crime (in Chicago & Pittsburgh)

Machine Learning Methods V.S. Traditional Ones

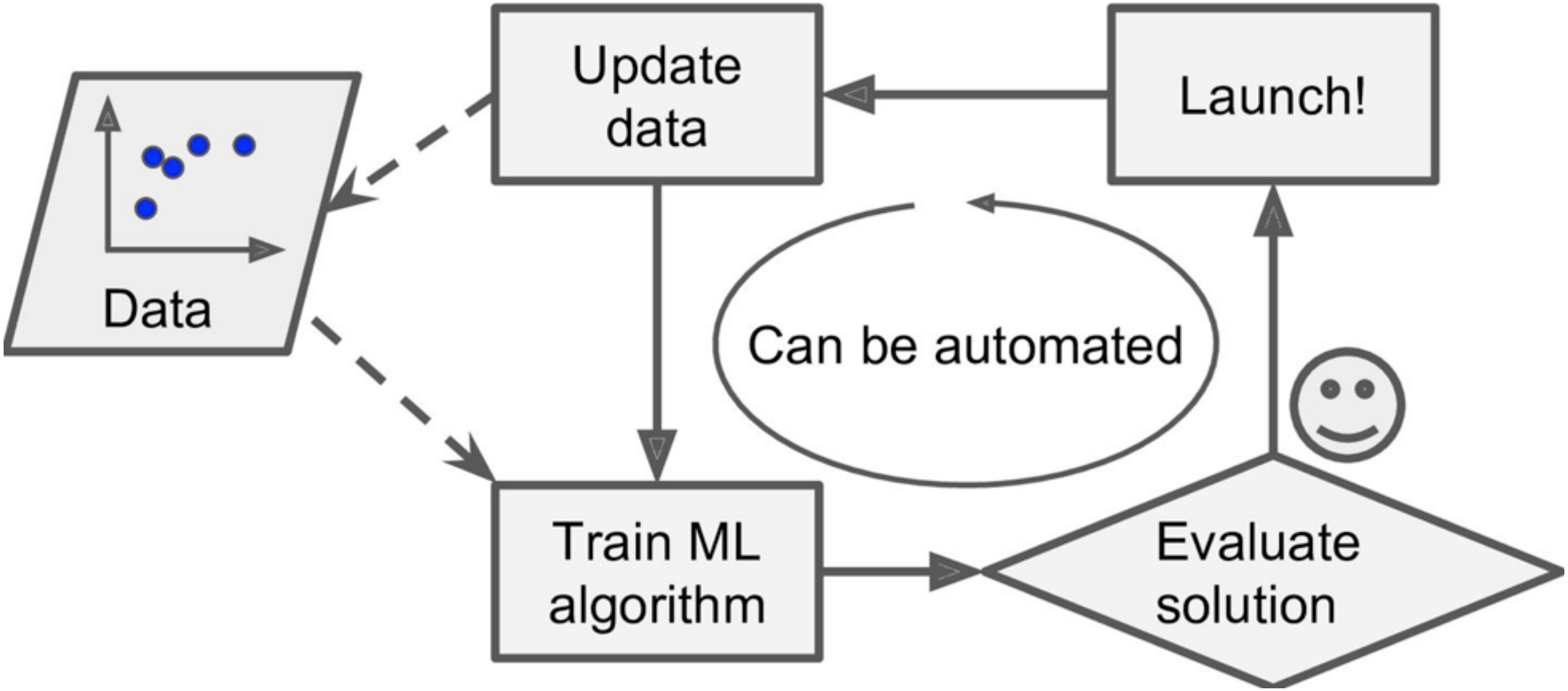
Traditional Method



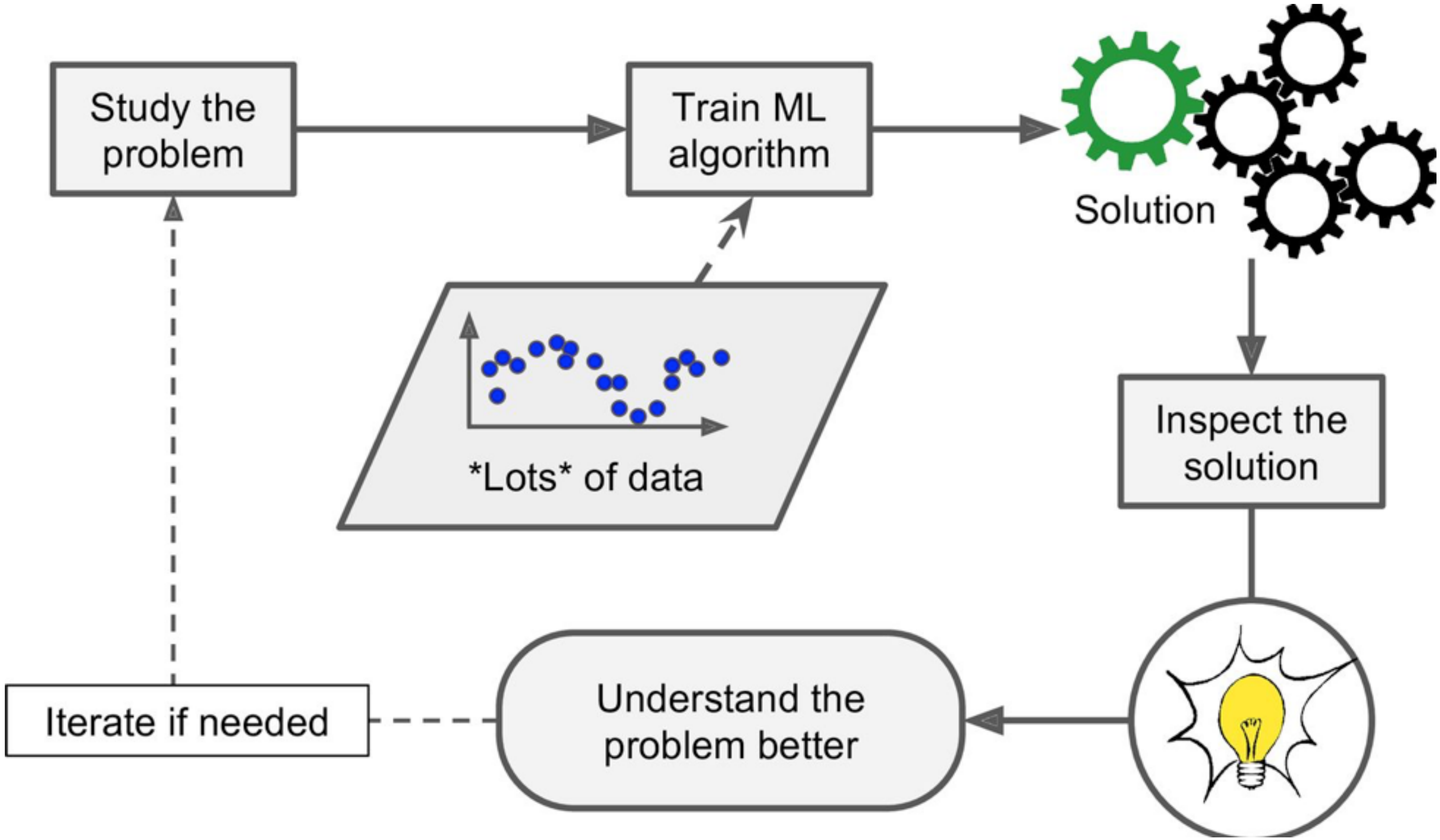
Machine Learning Method



New data coming in …



Help human learns



What is machine learning?

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

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

"Learning denotes changes in the system that are adaptive in the sense that they enable it to do a task, or tasks drawn from the same population, more efficiently and effectively next time." (H. Simon)

Learning as **generalization**: the ability to perform a task in a situation which has never been encountered before!

ML vs. computer programming

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

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

"Learning denotes changes in the system that are adaptive in the sense that they enable it to do a task, or tasks drawn from the same population, more efficiently and effectively next time." (H. Simon)

Learning as **generalization**: the ability to perform a task in a situation which has never been encountered before!

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths." (A. Lovelace, 1842)

ML vs. computer programming

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

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

"Learning denotes changes in the system that are adaptive in the sense that they enable it to do a task, or tasks drawn from the same population, more efficiently and effectively next time." (H. Simon)

Learning as **generalization**: the ability to perform a task in a situation which has never been encountered before!

"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort." (A. Samuel, 1959)

Samuel coined the term "machine learning" and was best known for his self-learning checkers program.

ML vs. measly humans

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

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

"Learning denotes changes in the system that are adaptive in the sense that they enable it to do a task, or tasks drawn from the same population, more efficiently and effectively next time." (H. Simon)

Learning as **generalization**: the ability to perform a task in a situation which has never been encountered before!

Computers now consistently beat the top players in the world at checkers (1995), chess (1997), and Go (2016), as well as the game show Jeopardy (2011).

ML vs. Humans

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

“A computer program is said to learn from experience E wrt. some class of tasks T and performance measure P, if its performance at tasks in T as measured by P improves with experience.” (T. Mitchell)

“Learning denotes changes in the system that are adaptive in the sense that they enable it to do a task, or tasks drawn from the same population, more efficiently and effectively next time.” (H. Simon)

Learning as **generalization**: the ability to perform a task in a situation which has never been encountered before!

ML methods have had tremendous success on tasks including control (self-driving cars), image recognition, speech recognition, recommender systems, machine translation, etc.

CMU’s “Boss”, winner of the DARPA Urban Challenge

Examples of performance metrics

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

“A computer program is said to learn from experience E wrt. some class of tasks T and performance measure P, if its performance at tasks in T as measured by P improves with experience.” (T. Mitchell)

Task	Performance metric	Experience
Play checkers	Percentage of wins vs. given opponent	Games previously played w/ outcomes
Recognize handwritten digits	Percentage of correct recognitions	Set of digit writing w/ labels
Control a self-driving car	Average speed in given conditions provided that safety standards are met	Previous driving record w/ evaluation
Predict stock prices	Average prediction accuracy	History of stock prices

Table credit: Stanislav Sobolevsky

ML as Optimization

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

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

1. Select performance metric and dataset to evaluate it

3. Look for the set of model parameters that optimize the given performance metric

Figure credit: Stanislav Sobolevsky

ML draws from many disciplines

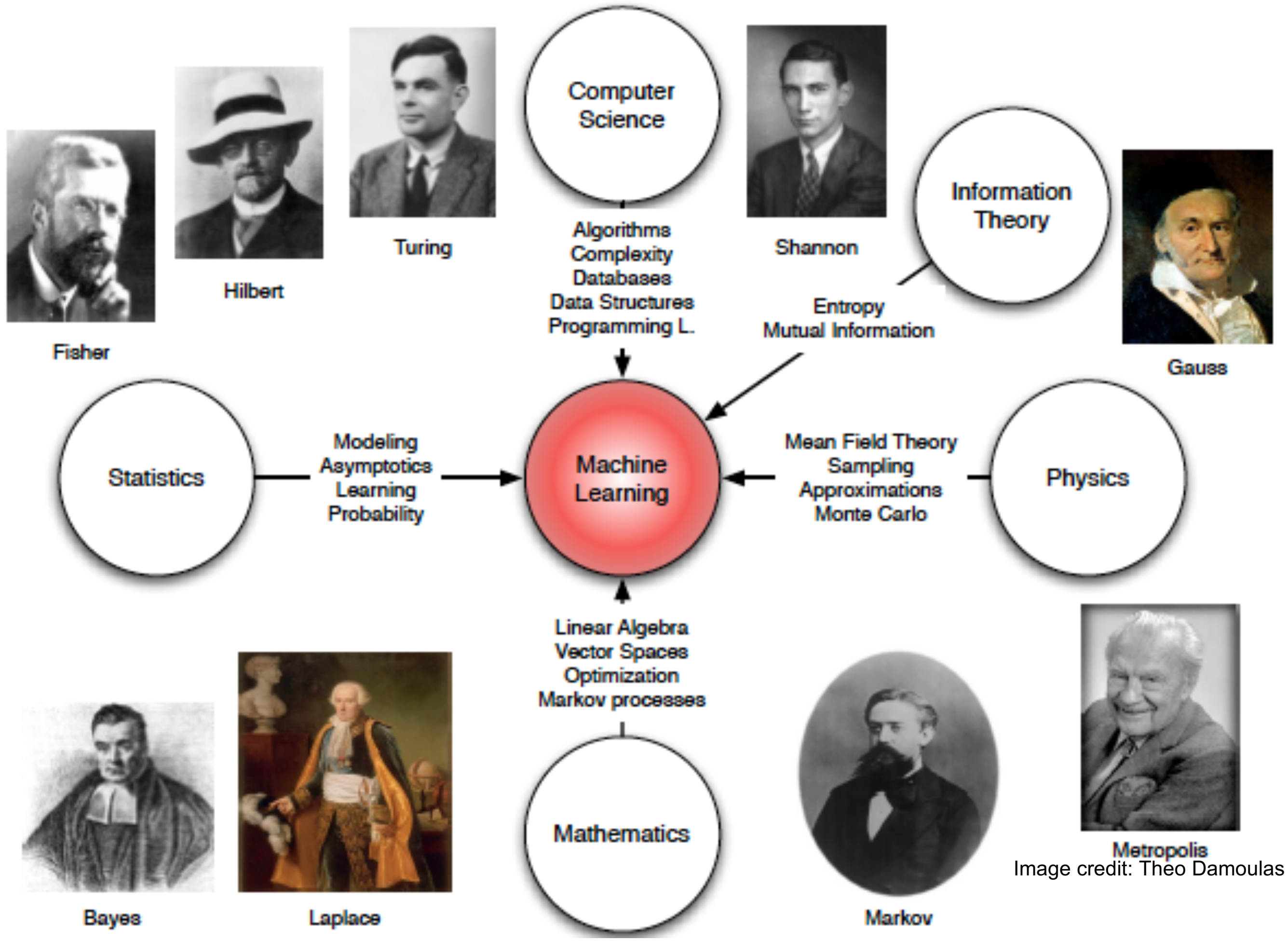


Image credit: Theo Damoulas

Also cognitive psychology, evolution, economics, neuroscience, and many more!

ML and related fields

Machine Learning (ML) is the study of systems that improve their performance with experience (typically by **learning** from data).

Artificial Intelligence (AI) is the science of automating complex behaviors that normally require human intelligence: vision, language understanding, learning, problem solving, decision making, etc.

Data Mining (DM) is the process of extracting useful information from massive quantities of complex data.

I would argue that these are not three distinct fields of study! While each has a slightly different emphasis, there is a tremendous amount of overlap in the problems they are trying to solve and the techniques used to solve them.

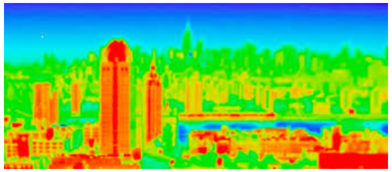
Many of the techniques we will learn are **statistical** in nature, but are very different from classical statistics.

ML/AI/DM systems and methods:

Scale up to large, complex data
Learn and **improve** from experience
Perceive and **change** the environment
Interact with humans or other agents
Explain inferences and decisions
Discover new and useful patterns

Urban Applications of ML

- Inferring urban dynamics from heterogeneous data
- Computer vision: pedestrian/traffic counts, security/law enforcement (face recognition), traffic accident detection, remote sensing (air content, IR, etc.)
- Street noise (decomposition, localization, classification)
- Economic patterns detection and prediction
- Health pattern detection and prediction
- Energy usage prediction
- Traffic modeling and prediction
- Land use classification
- 3-D landscape recognition
- Event detection from urban activity
- Detecting trends from social media



Infrared data from CUSP's Urban Observatory

Course objectives

<u>Goals</u>: familiarize students with modern machine learning methods and show how they can be effectively applied to visual data.

More precisely, this course is intended to enable students to:

1) Understand the **motivation** behind different machine learning methods and their **applicability** in a given practical context.
2) Be able to **implement** methods adapted to the problem at hand using existing software libraries.*
3) Know how to **interpret** the results appropriately.

* This course includes a significant programming component, with **Python** as the primary programming language.

Please bring your laptop (Mac or Ubuntu) and follow along.

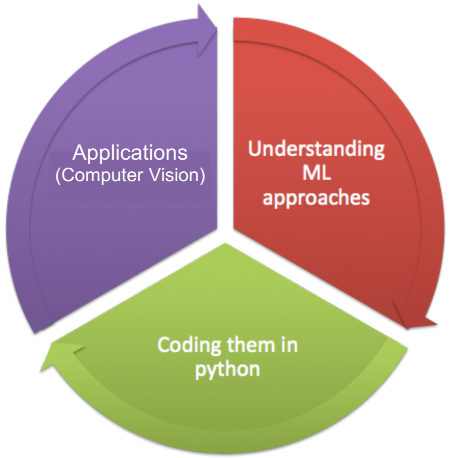


Image credit: Stanislav Sobolevsky

Structure of the course

- Lectures
- Combination of core ML methods and ML topics most relevant to visual data analysis
  - Motivating examples and applications.
- Five Mini projects
- One Large project (semester-long)
- Midterm and final presentation
- First part of course: mainly standard machine learning techniques.
- Second half of course: deep learning and its applications in computer vision.

Structure of the course

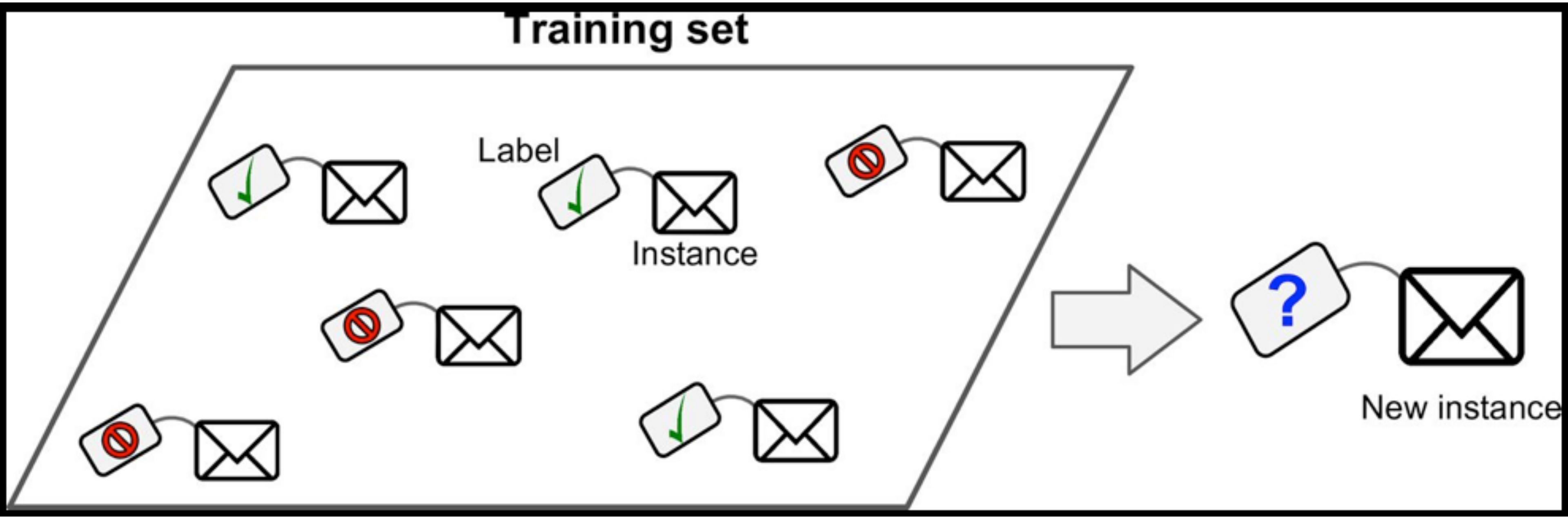
- Two Tracks of Assignments:
  - Track One
    - Six mini projects (10% * 6 = 60%)
    - One large project (40% *1 = 40%)
  - Track Two
    - Six mini projects (10% * 6 = 60%)
    - Two Medium project (20% *2 = 40%)
- See syllabus for :
  - Detailed course schedule (subject to change!)

Machine Learning, Spring 2020

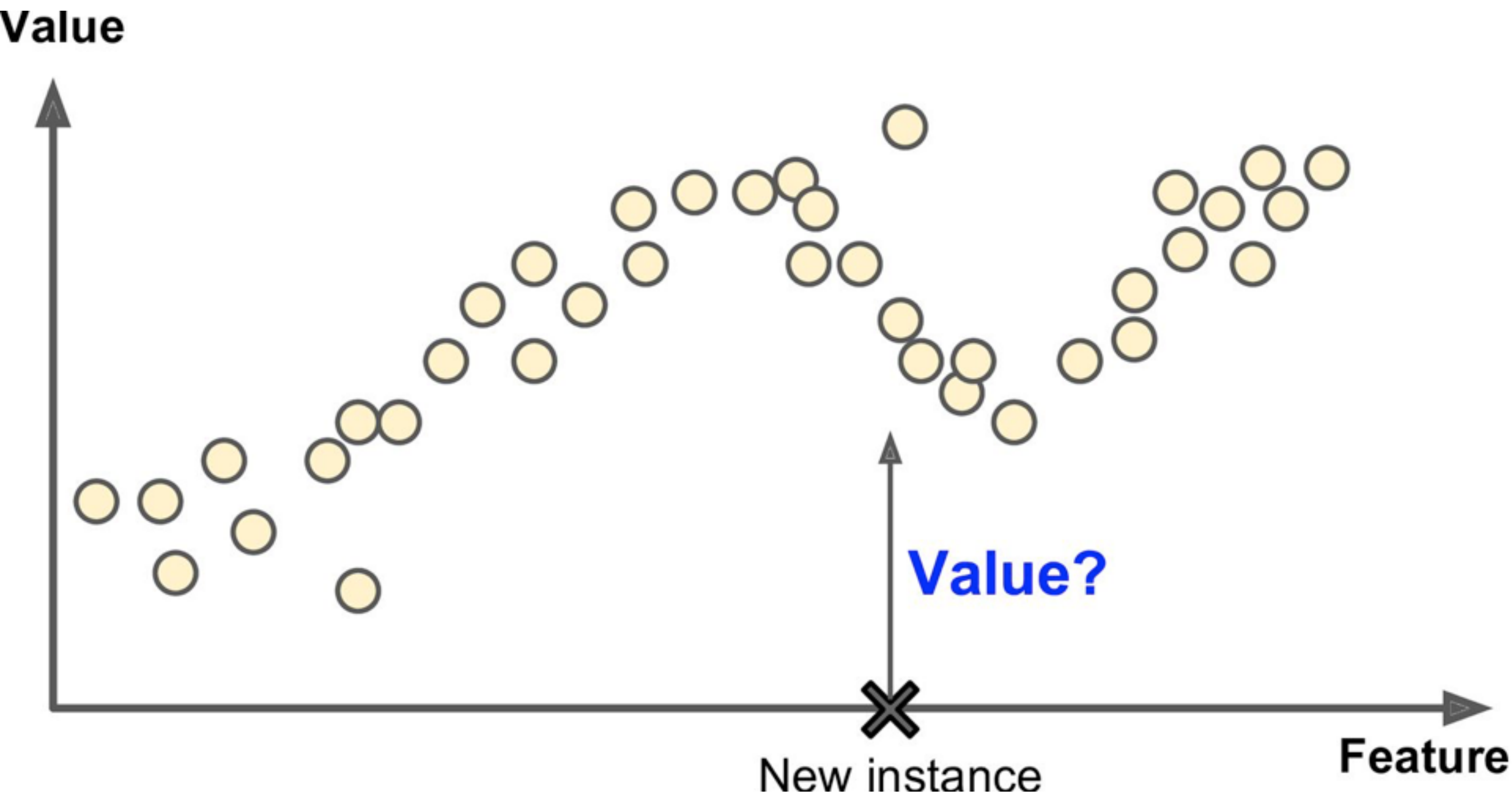
Overview of ML Methods

Supervised Learning

Labeled training set for supervised learning (classification)

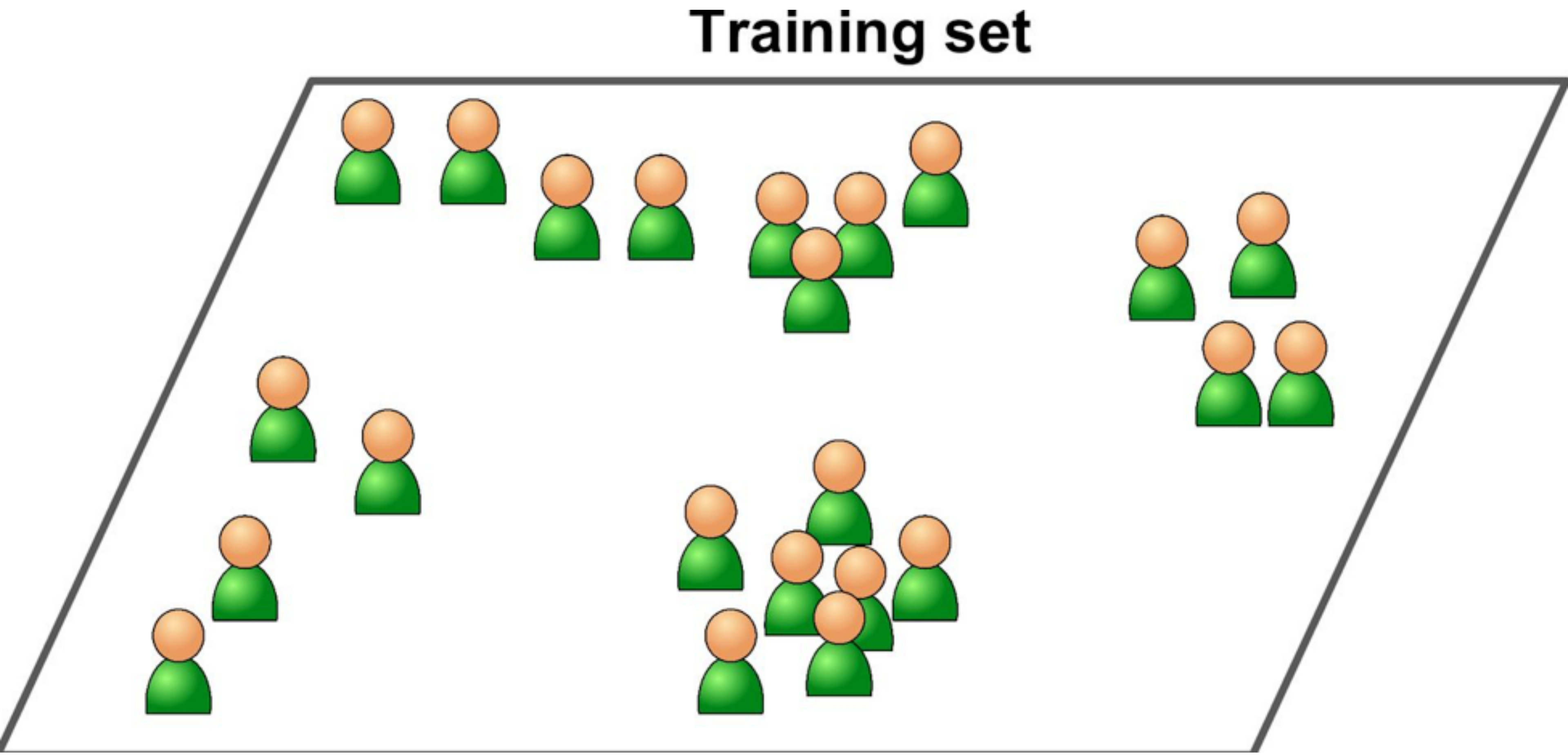


Labeled training set for supervised learning (regression)

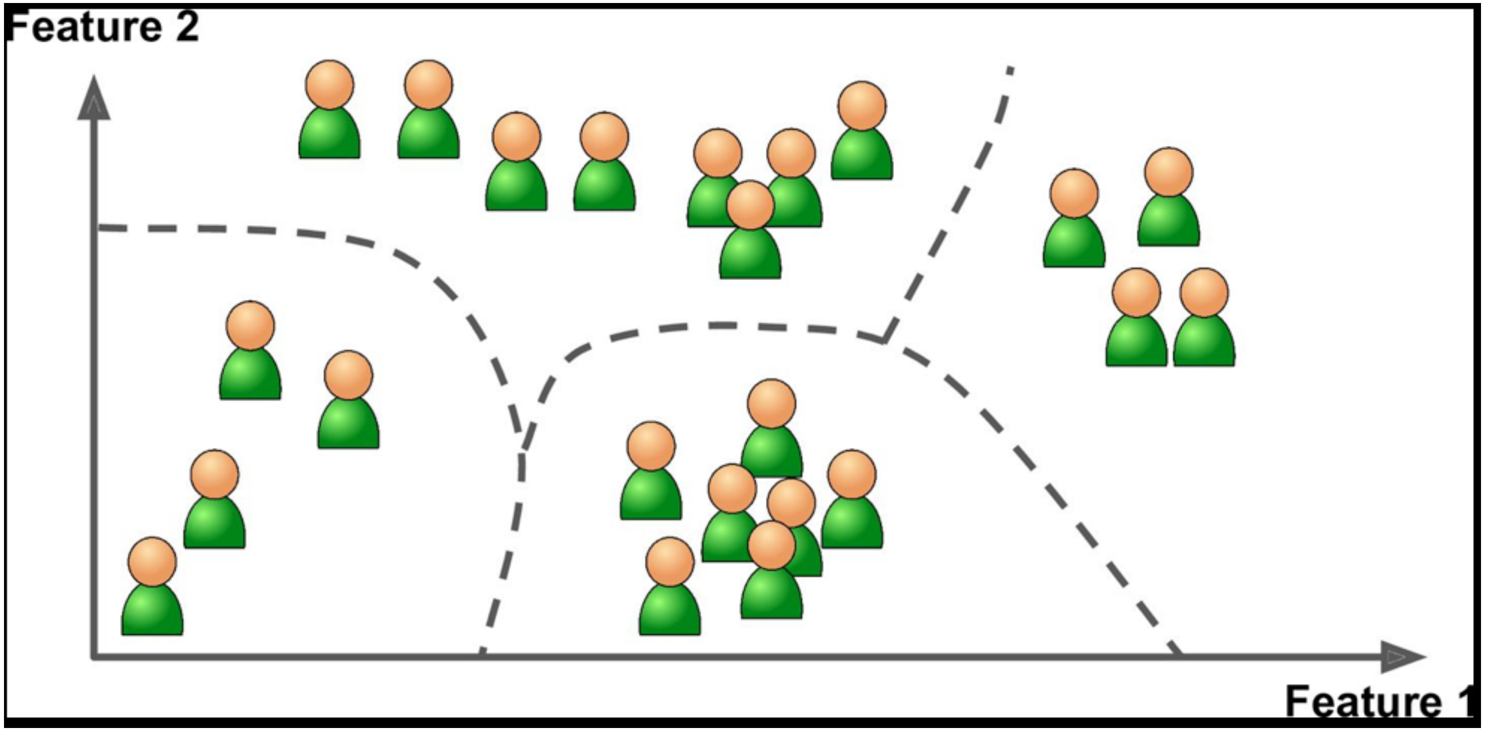


Unsupervised Learning

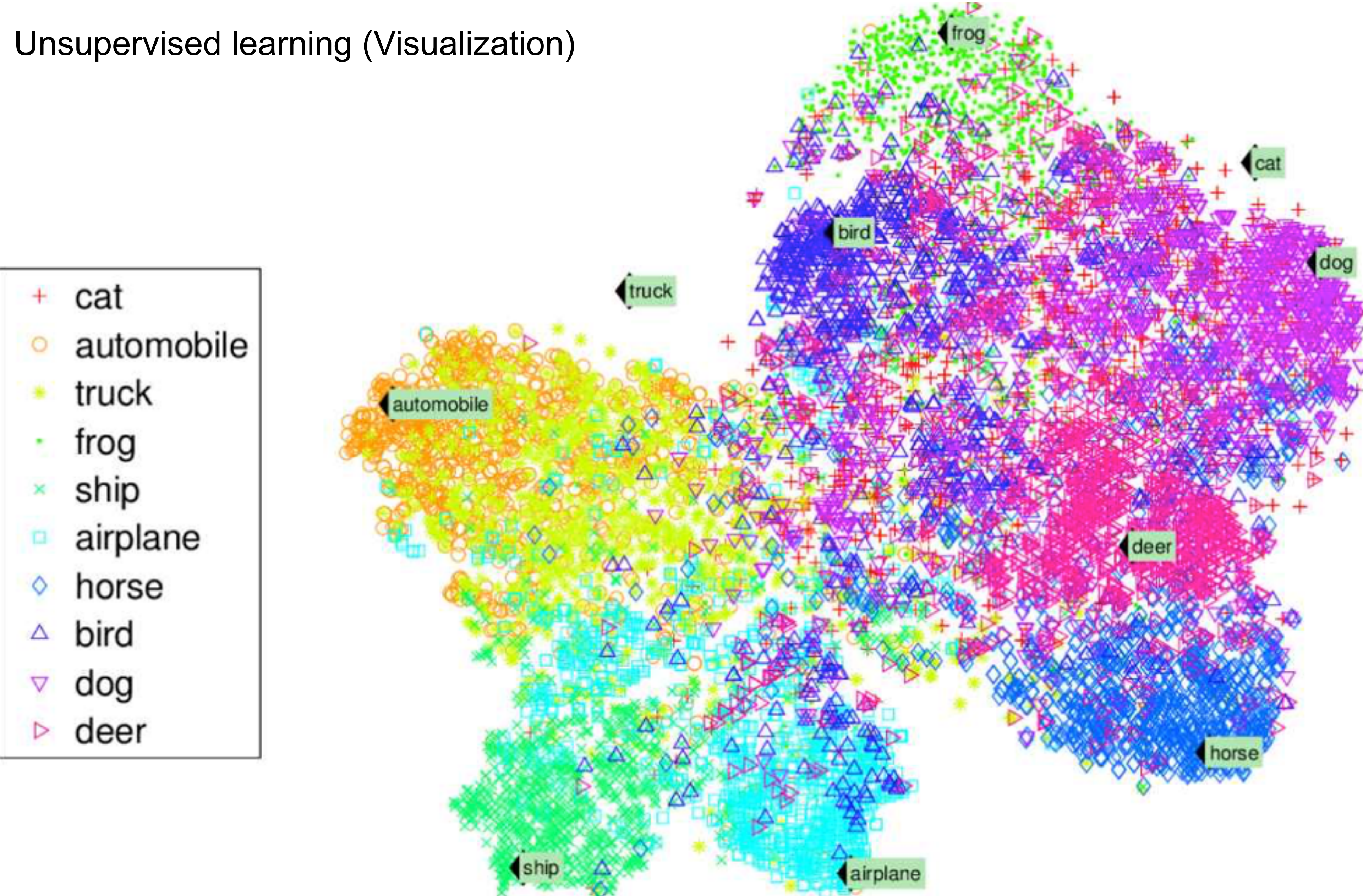
Unlabeled training set for unsupervised learning



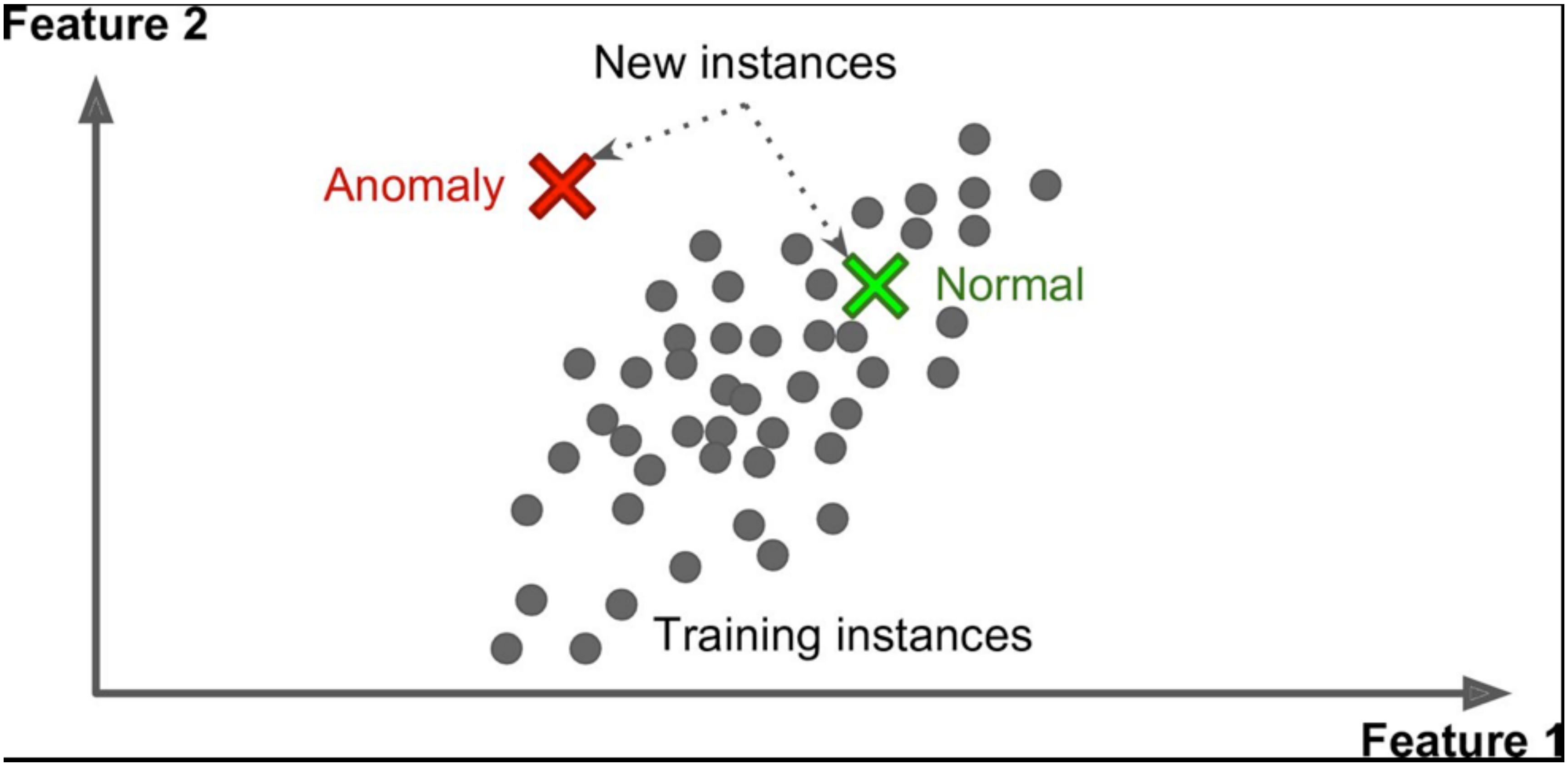
Unlabeled training set for unsupervised learning (clustering)



Unsupervised learning (Visualization)

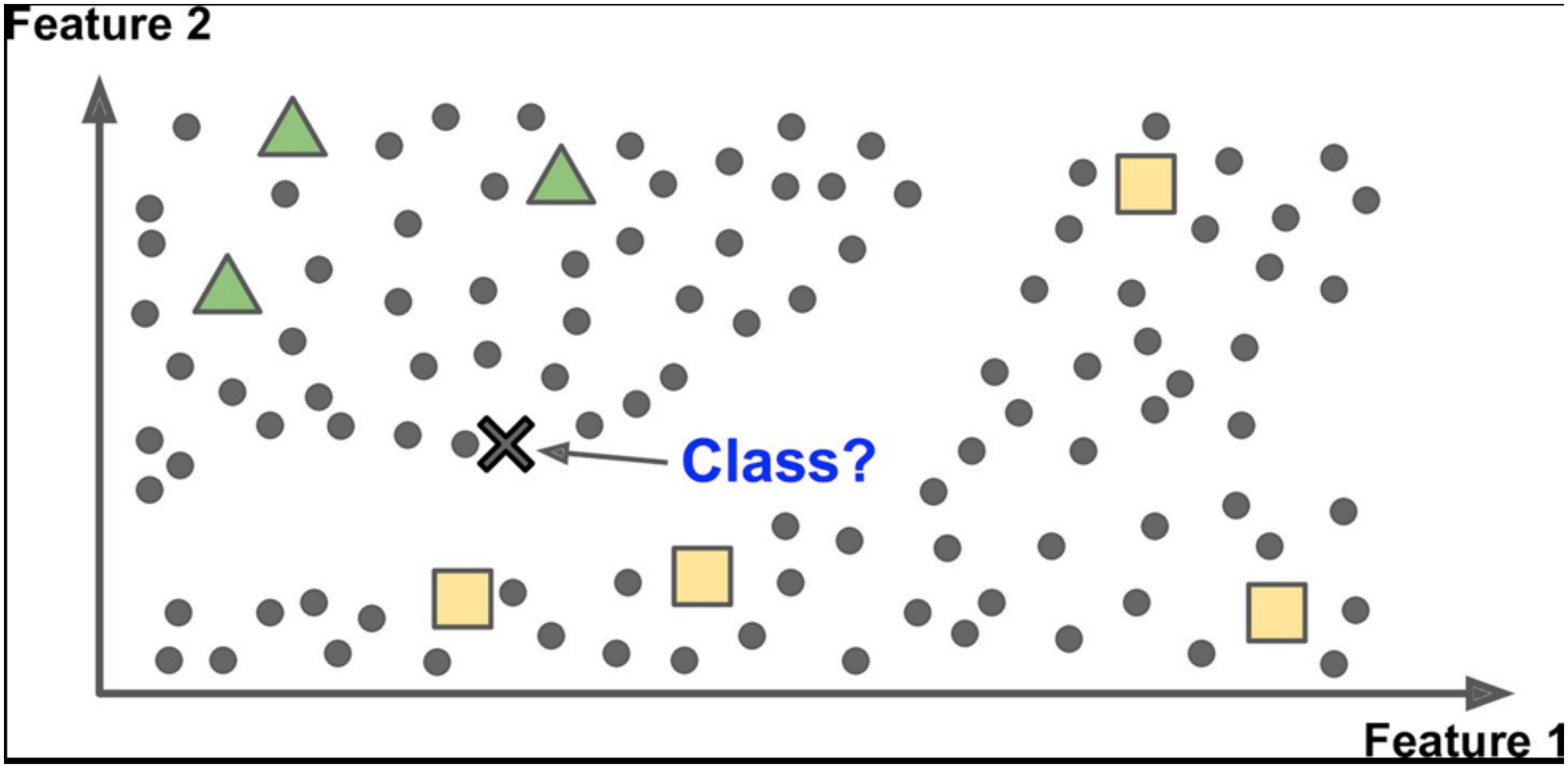


Unsupervised learning (Anomaly detection)



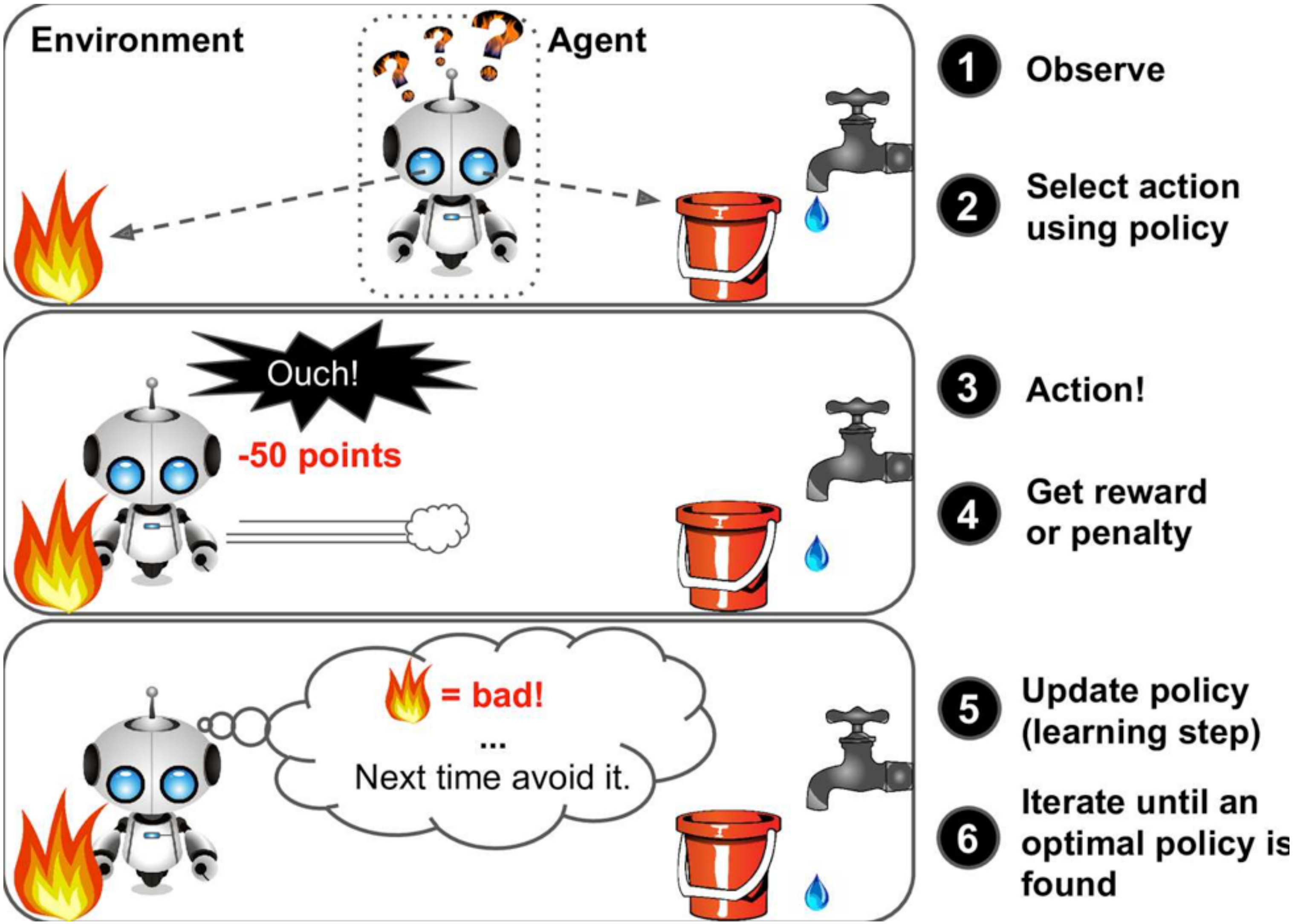
Semi-supervised Learning

Semi-supervised learning



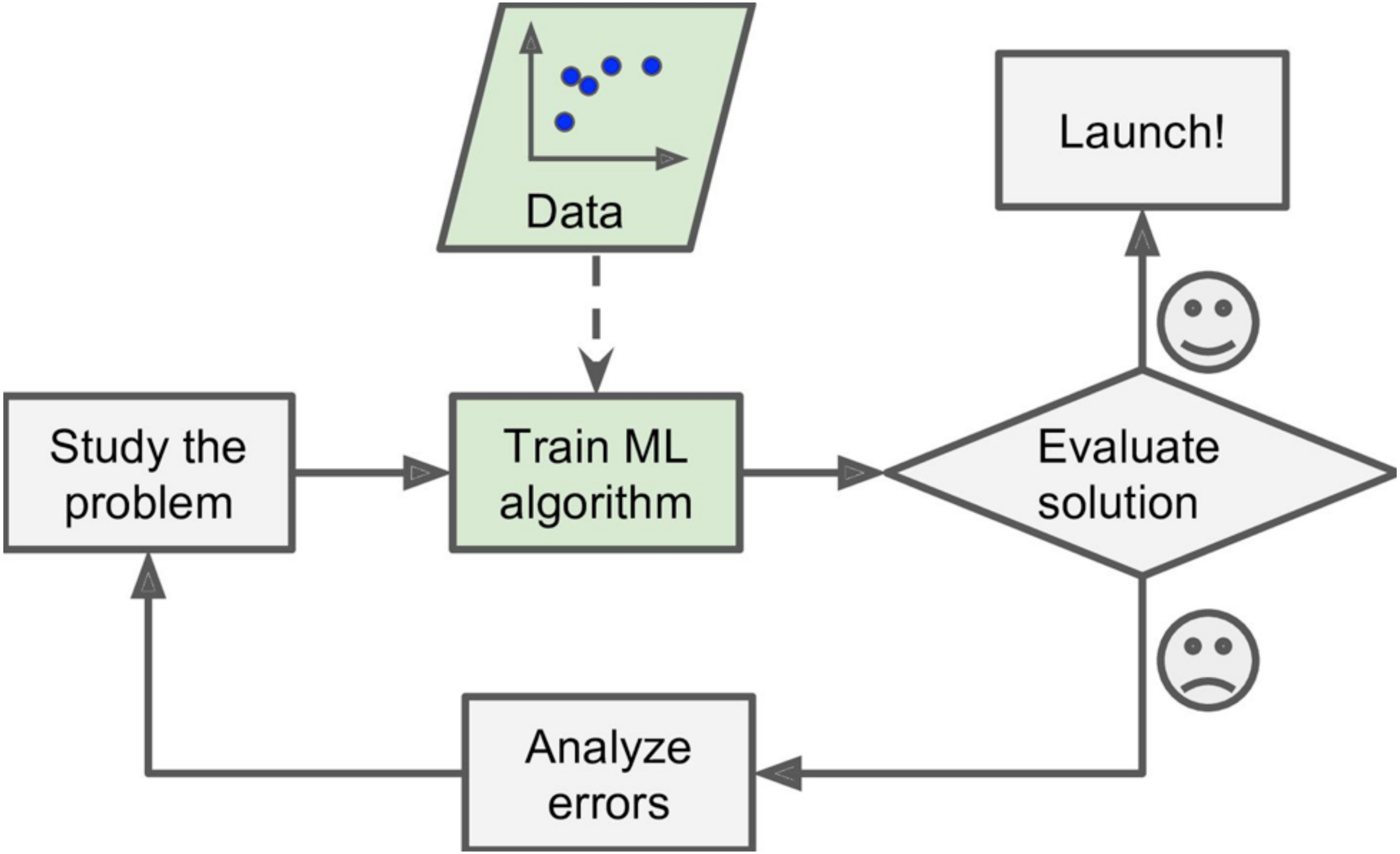
Reinforcement Learning

Reinforcement Learning



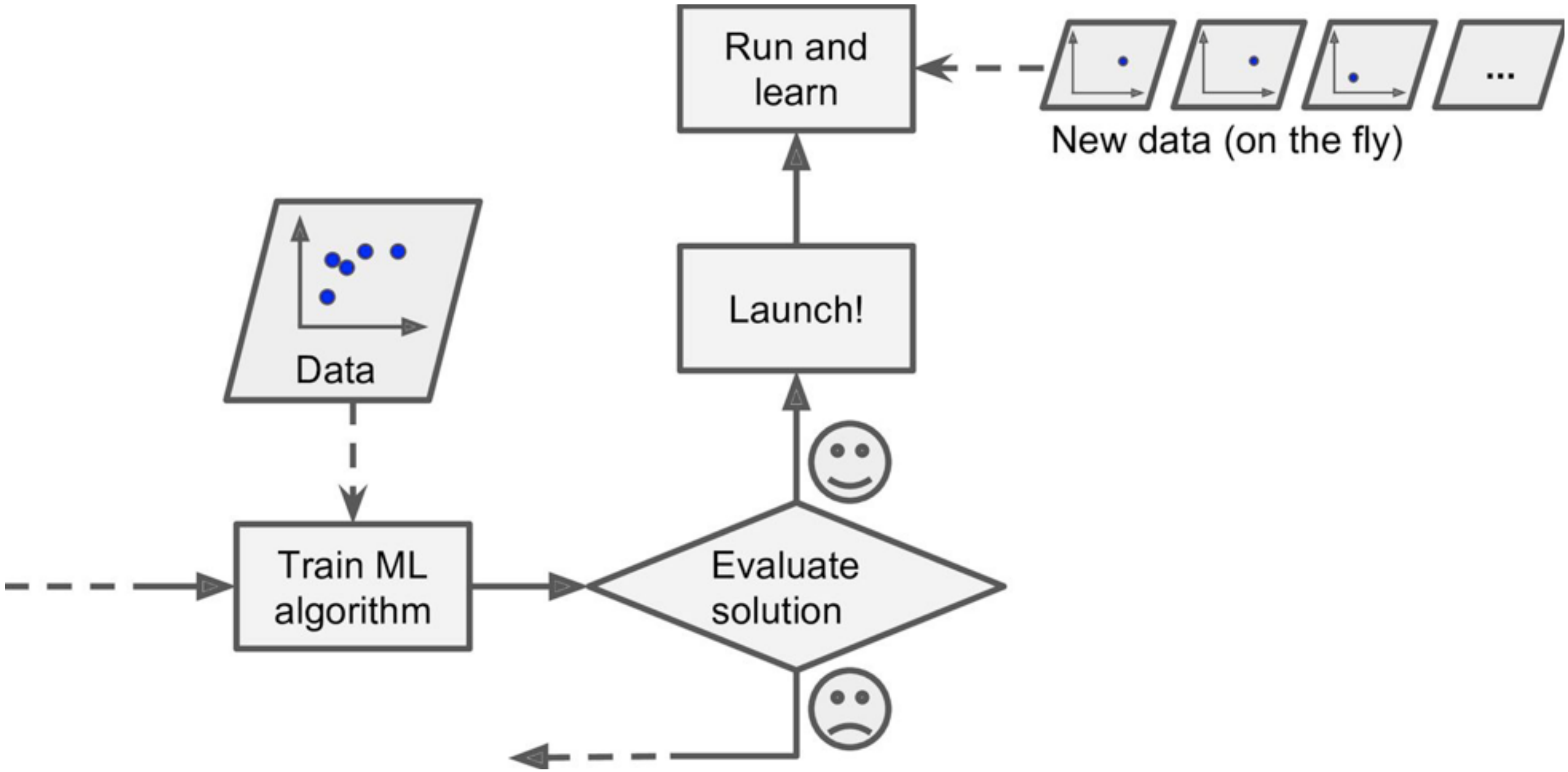
Batch Learning
VS
Online Learning

Batch Learning (offline)

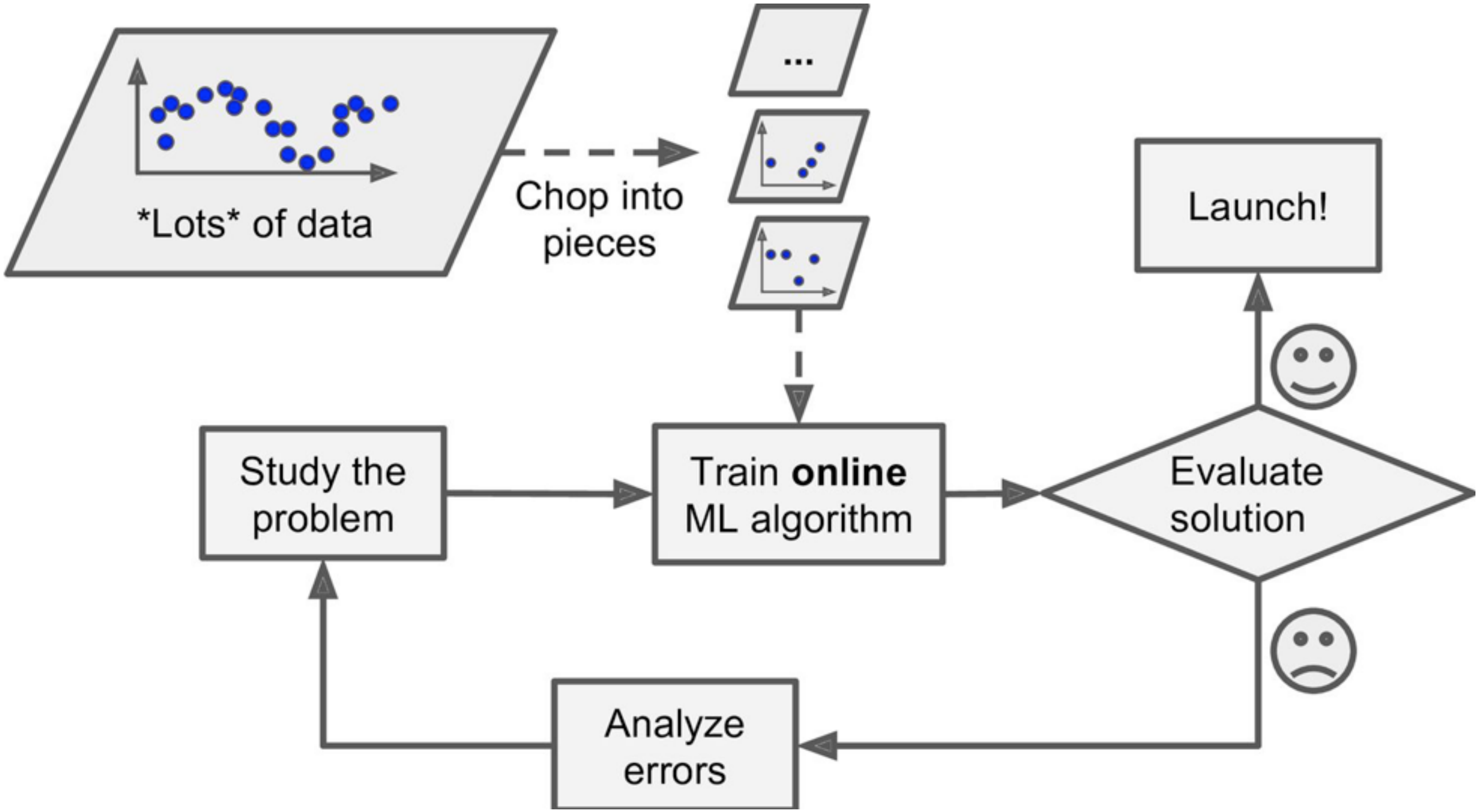


Online Learning

Online Learning (incremental learning)

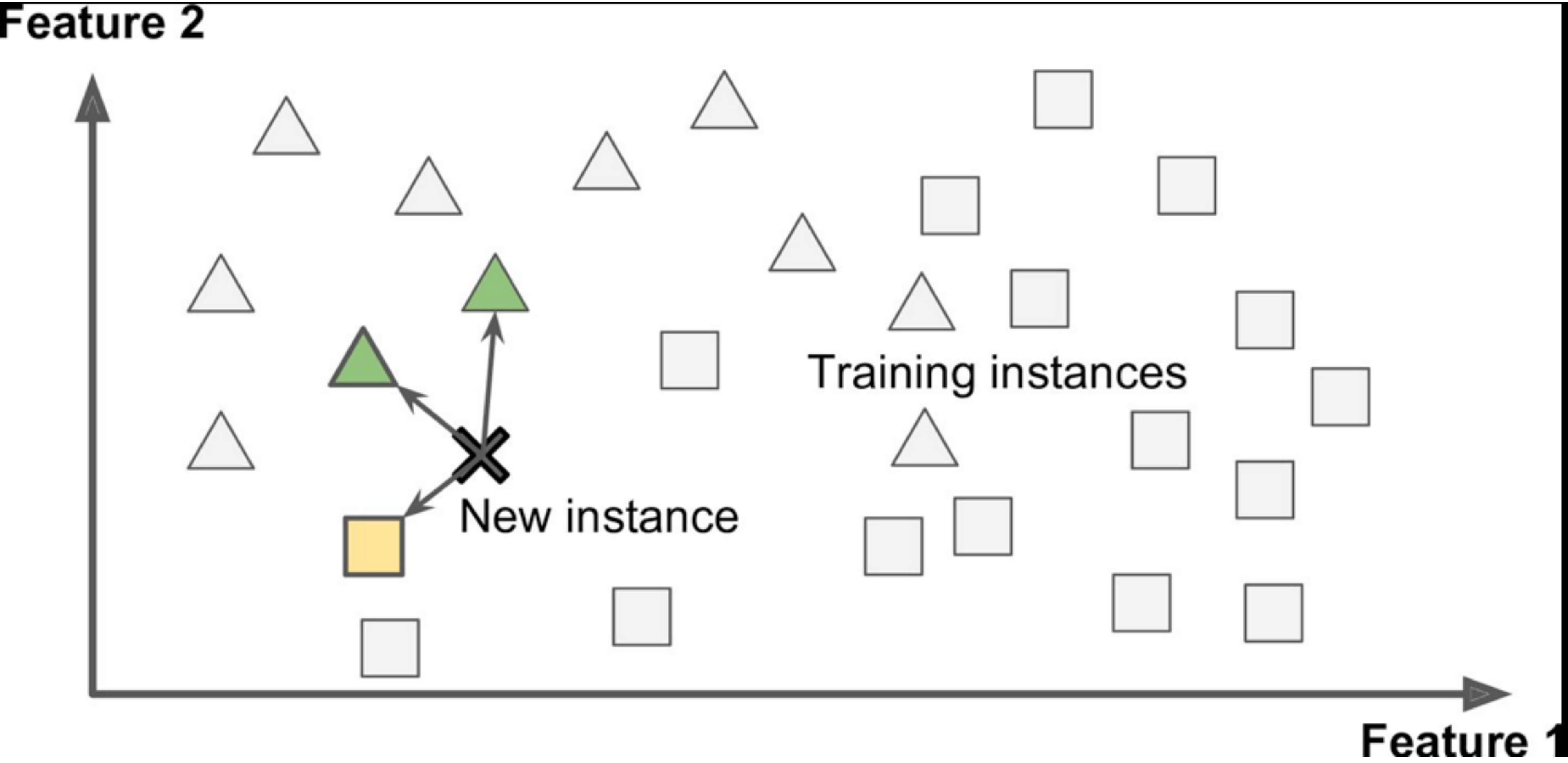


Online Learning (handle huge datasets)



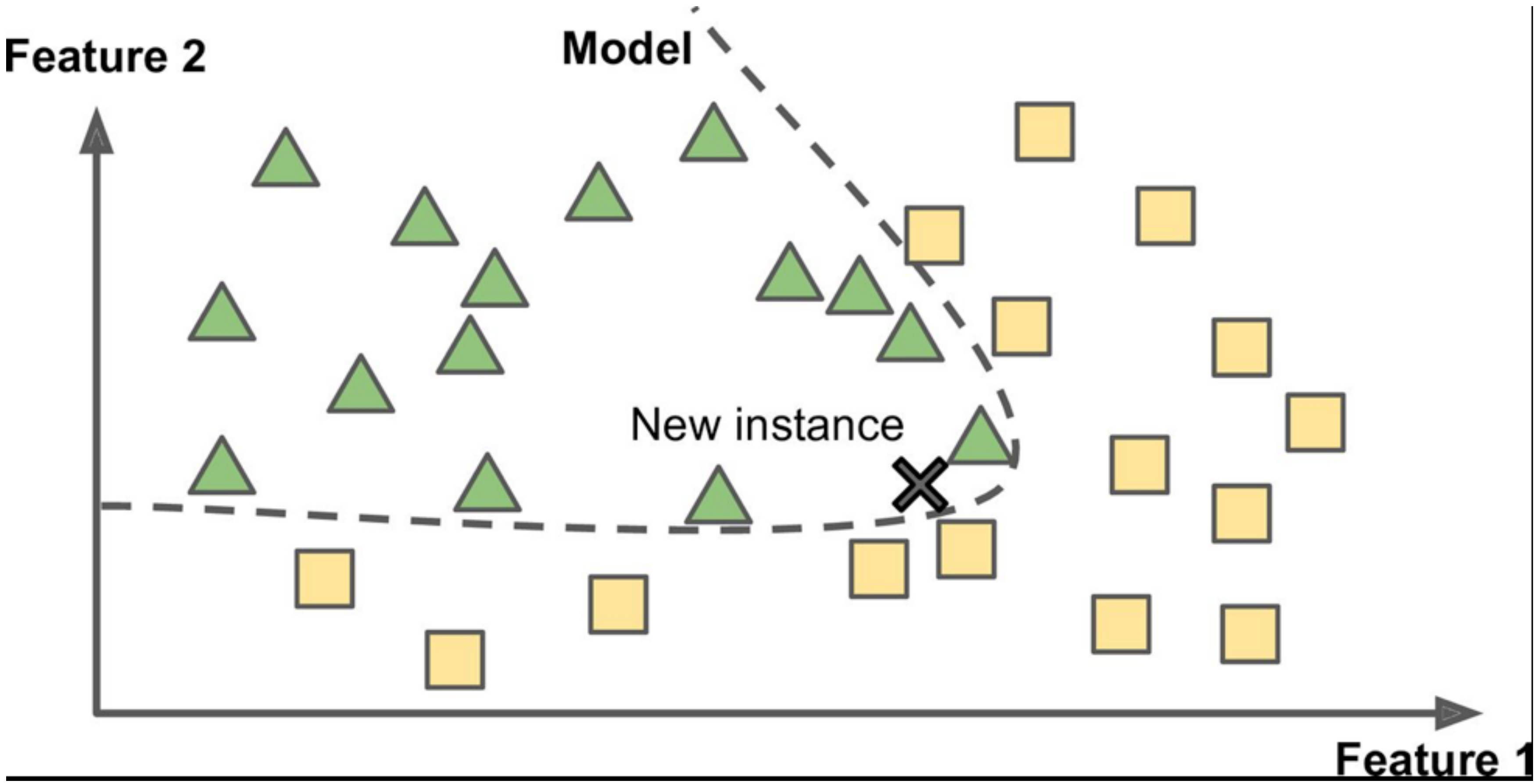
Instance-based Learning

Instance-based learning

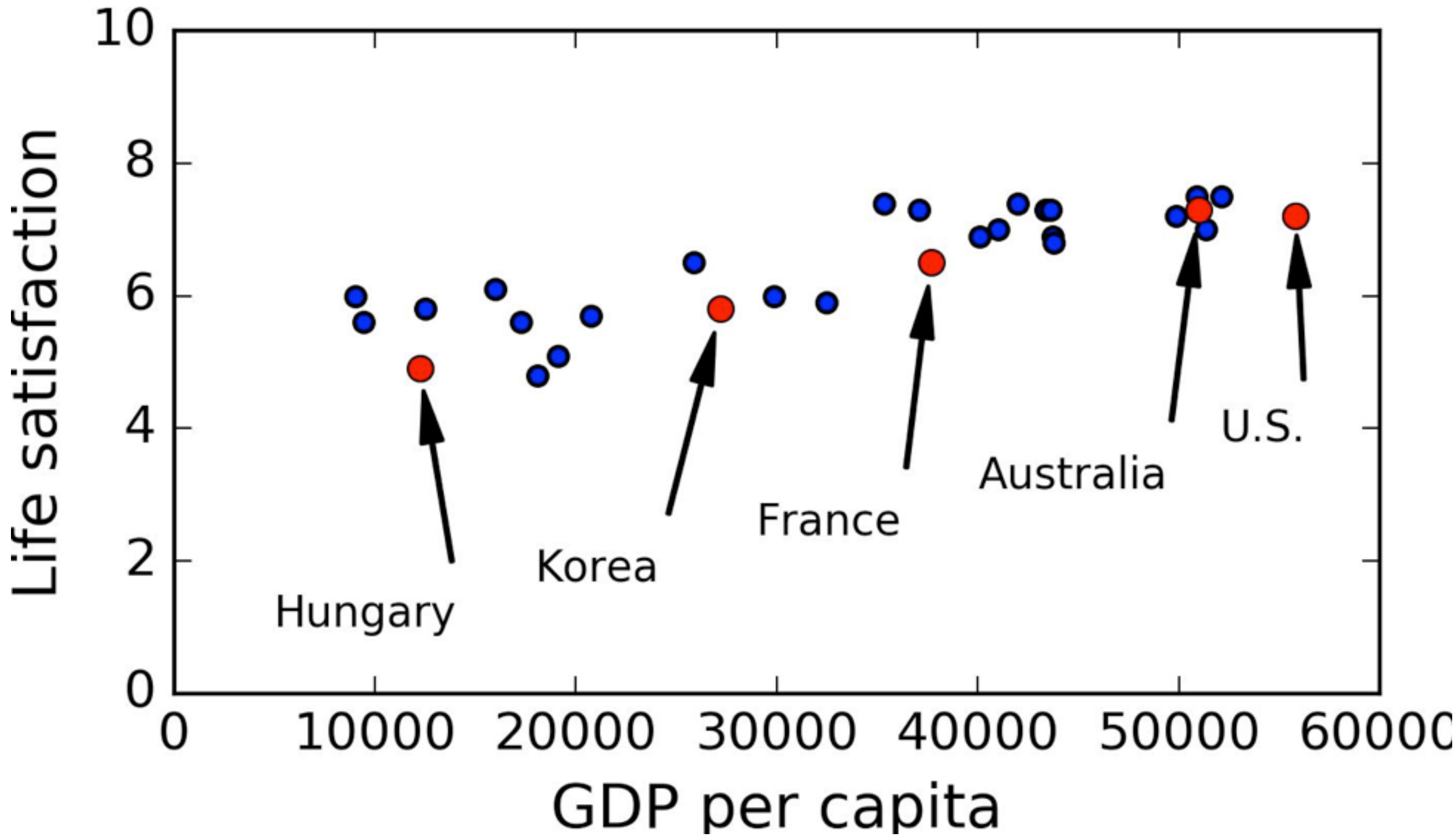


Model-based Learning

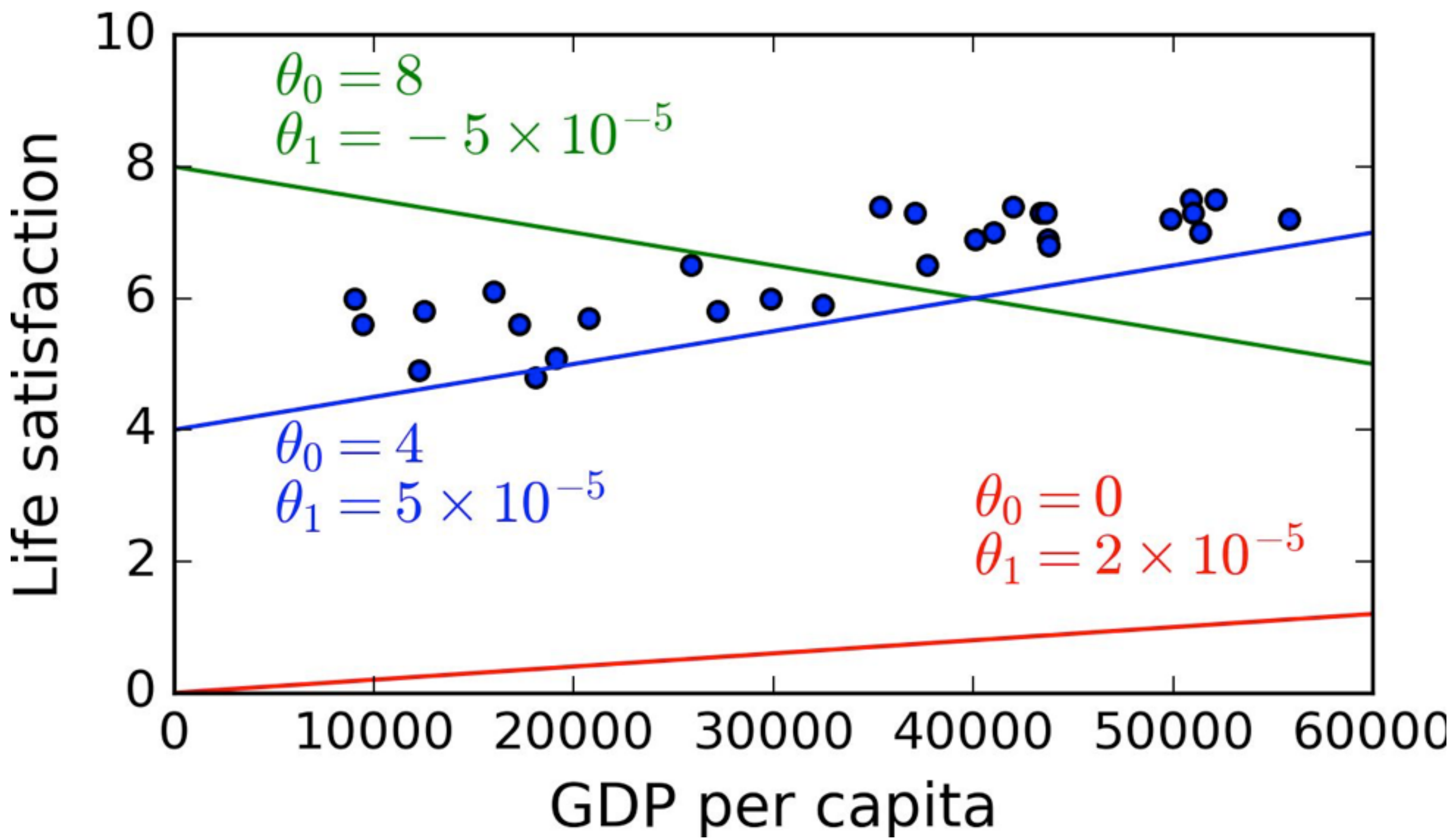
Model-based learning



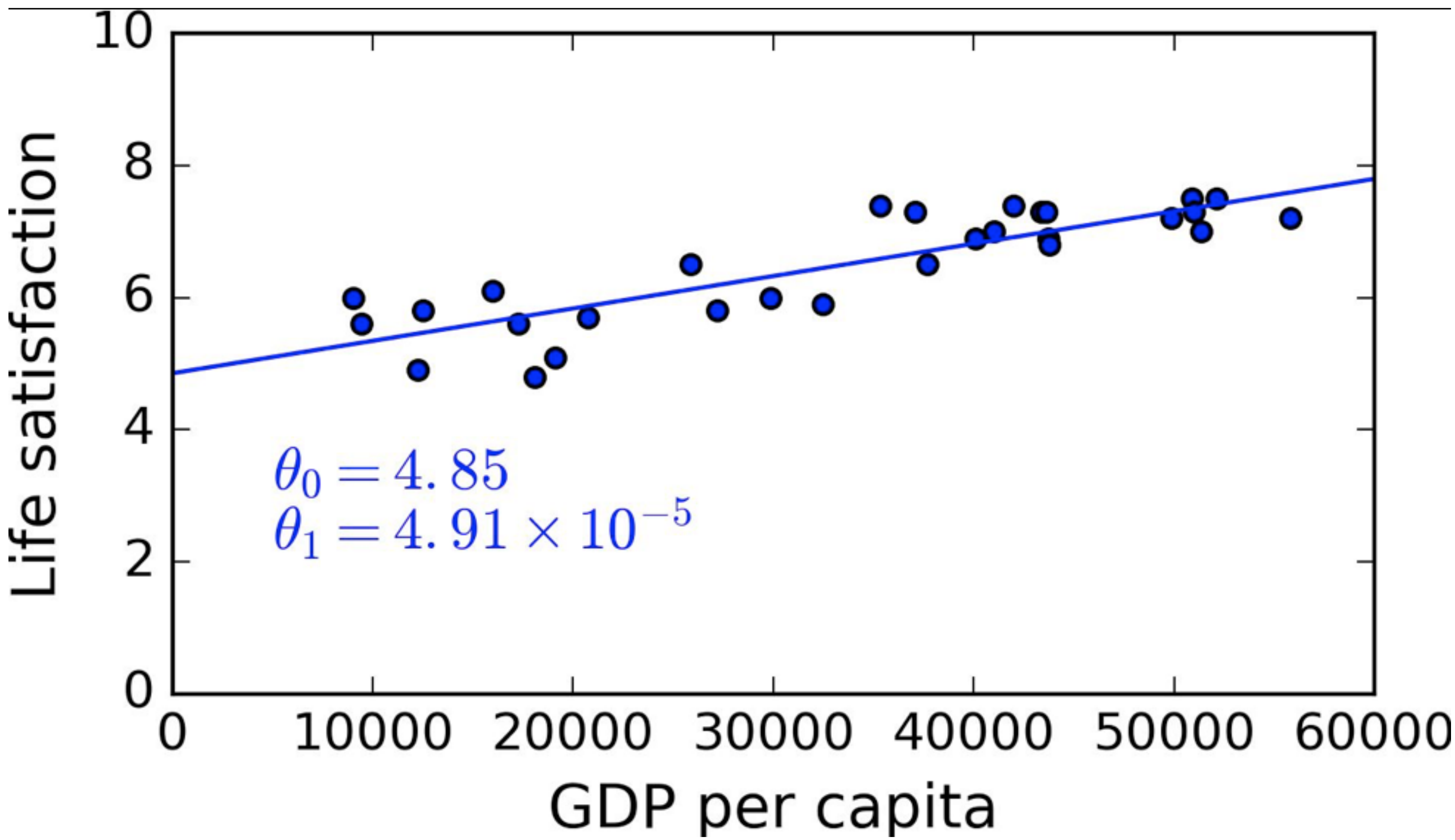
Trend?


Life satisfaction
10
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000
GDP per capita
Hungary
Korea
France
Australia
U.S.

Possible linear models



Optimal linear model



Example

Example 1-1. Training and running a linear model using Scikit-Learn

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

# Load the data
oecd_bli = pd.read_csv("oecd_bli_2015.csv", thousands=',')
gdp_per_capita = pd.read_csv("gdp_per_capita.csv",thousands=',',delimiter='\t',
                             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 satisfaction')
plt.show()

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

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

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

NOTE

If you had used an instance-based learning algorithm instead, you would have found that Slovenia has the closest GDP per capita to that of Cyprus ($20,732), and since the OECD data tells us that Slovenians' life satisfaction is 5.7, you would have predicted a life satisfaction of 5.7 for Cyprus. If you zoom out a bit and look at the two next closest countries, you will find Portugal and Spain with life satisfactions of 5.1 and 6.5, respectively. Averaging these three values, you get 5.77, which is pretty close to your model-based prediction. This simple algorithm is called *k-Nearest Neighbors* regression (in this example, k = 3).

Replacing the Linear Regression model with k-Nearest Neighbors regression in the previous code is as simple as replacing this line:

```
model = sklearn.linear_model.LinearRegression()
```

with this one:

```
model = sklearn.neighbors.KNeighborsRegressor(n_neighbors=3)
```

Machine Learning, Spring 2020

Main Challenges of Machine Learning

Reading Assignment: Chapter 1 & 2

Python tutorial: http://learnpython.org/

TensorFlow tutorial: https://www.tensorflow.org/tutorials/

PyTorch tutorial: https://pytorch.org/tutorials/