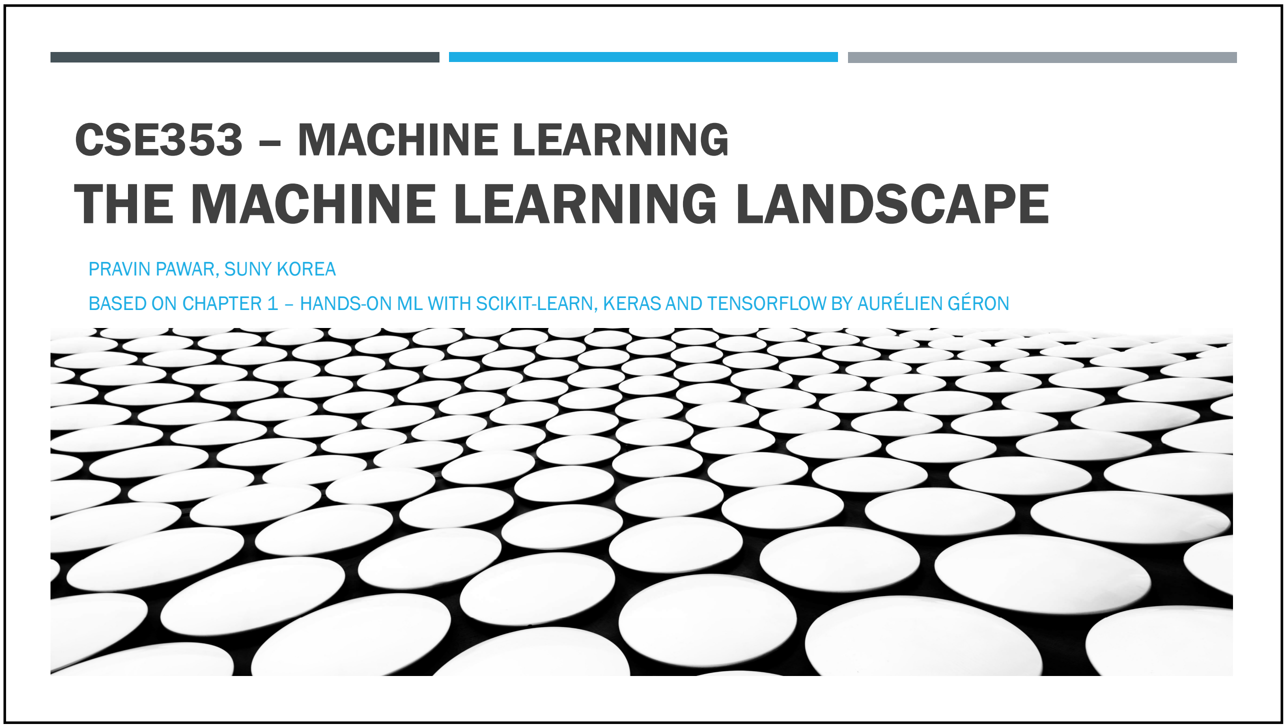
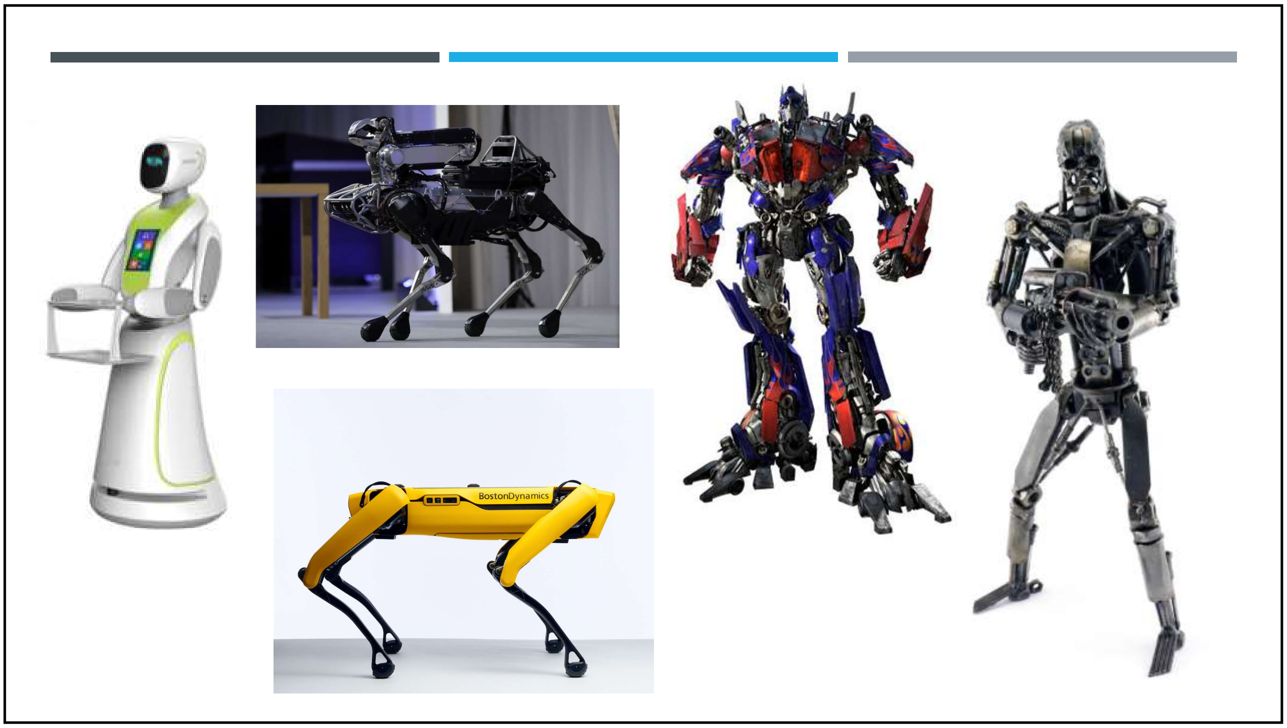


CSE353 – MACHINE LEARNING
THE MACHINE LEARNING LANDSCAPE

PRAVIN PAWAR, SUNY KOREA

BASED ON CHAPTER 1 – HANDS-ON ML WITH SCIKIT-LEARN, KERAS AND TENSORFLOW BY AURÉLIEN GÉRON

1



2

LIST SOME MACHINE LEARNING APPLICATIONS YOU KNOW

3

MACHINE LEARNING - DEFINITION

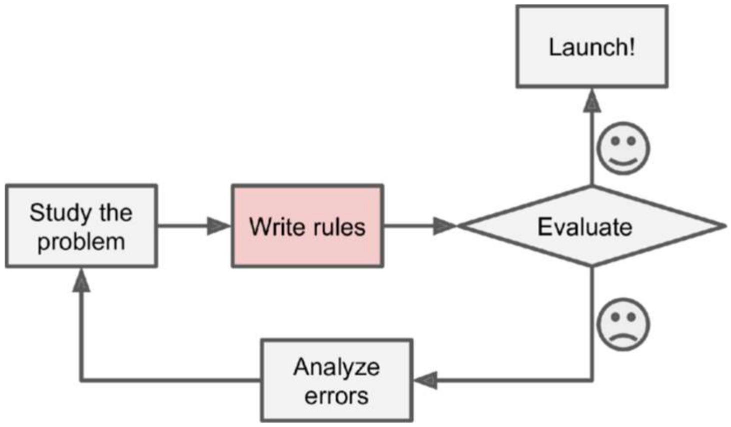
- Machine Learning is the science (and art) of programming computers so they can learn from data.
- Here is a slightly more general definition:
  - [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.
  - —Arthur Samuel, 1959
- And a more engineering-oriented one:
  - A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
  - —Tom Mitchell, 1997

4

ELABORATE ON HOW ML APPLICATIONS FOLLOW THESE DEFINITIONS

5

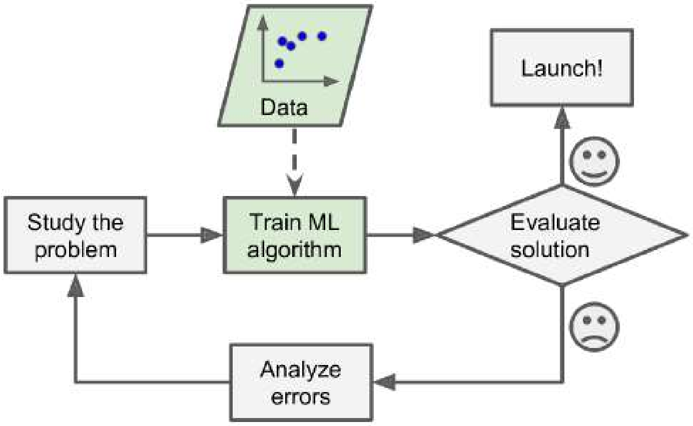
WHY USE MACHINE LEARNING?
SPAM FILTER EXAMPLE – TRADITIONAL APPROACH



- Detect a pattern
- Create a rule
- Evaluate
- Launch
- Program becomes a long list of complex rules hard to maintain

6

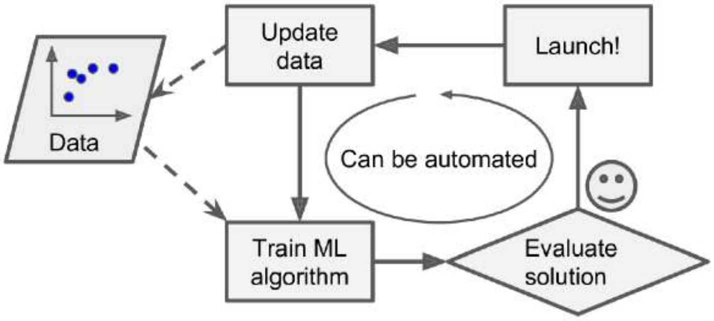
WHY USE MACHINE LEARNING?
SPAM FILTER EXAMPLE – MACHINE LEARNING APPROACH



- Flag email as spam
- ML algorithm compare spam vs ham examples
- Automatically learn words and phrases which are good predictors of spam

7

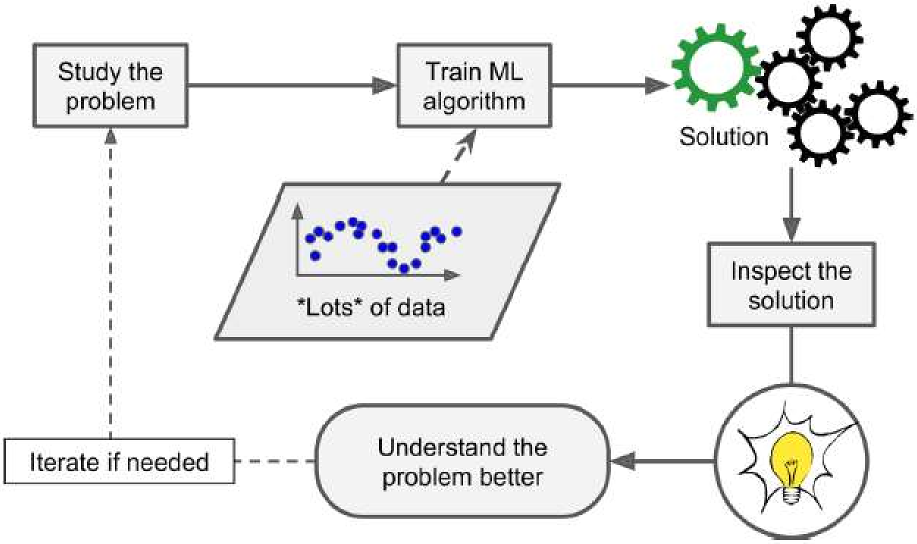
WHY USE MACHINE LEARNING?
SPAM FILTER EXAMPLE – ADAPTIVE MACHINE LEARNING APPROACH



- “4U” vs “For U”
- Adaptive ML algorithm learns on the fly

8

DATA MINING – APPLYING ML TO DIG INTO LARGE AMOUNTS OF DATA



TYPES OF MACHINE LEARNING SYSTEMS

- Machine Learning systems can be classified into broad categories, based on the following criteria:
  - 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 by detecting patterns in the training data and building a predictive model, much like scientists do (instance-based versus model-based learning)
- These criteria are not exclusive and can be combined in any way

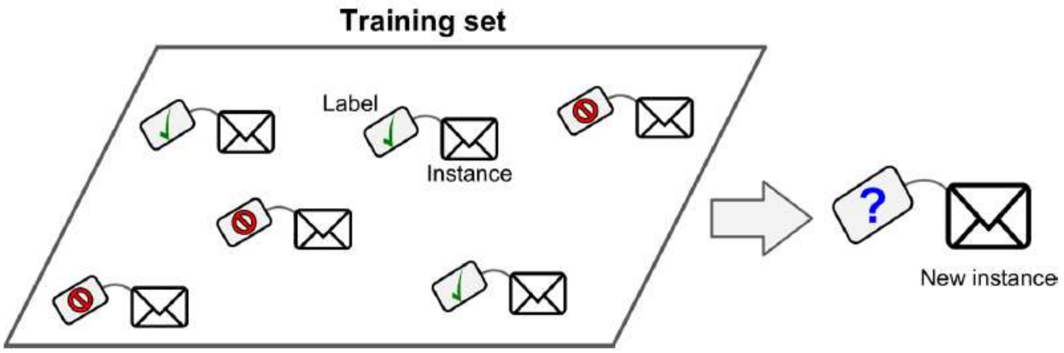
SUPERVISED LEARNING TASKS

- Classification: Learn how to classify examples given labeled data
  - K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, Neural Networks
- Regression: predict a target numeric value given a set of features (predictors)
  - Linear Regression, Neural Networks
- Regression for classification: Output a value that corresponds to the probability of belonging to a given class
  - Logistic Regression

11

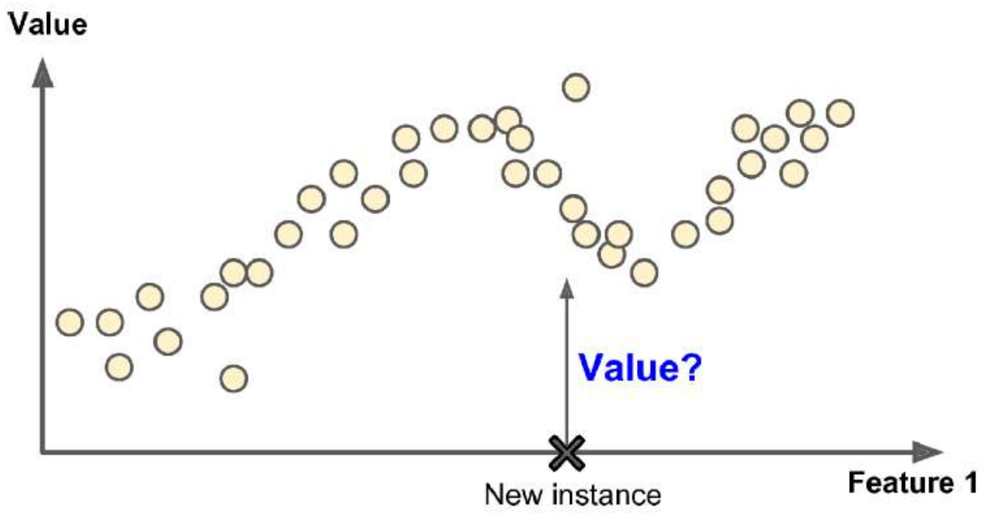
SUPERVISED LEARNING

- A training set fed to the algorithm includes desired solutions – *labels*



12

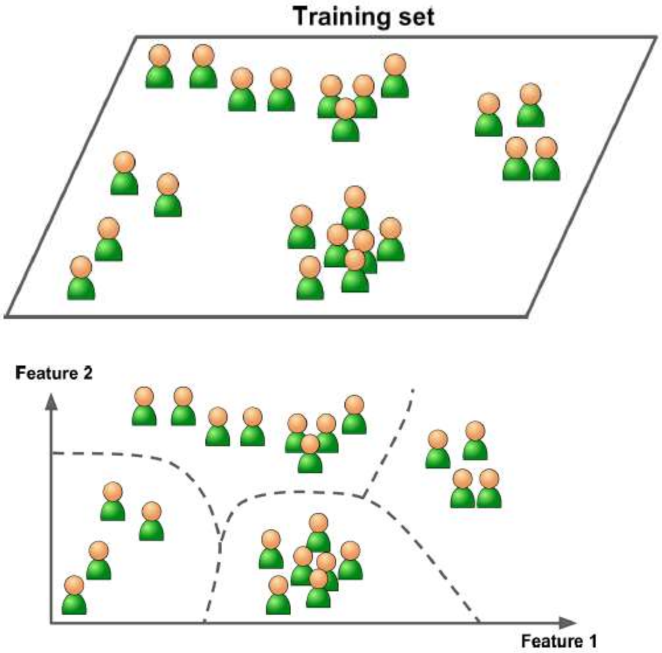
REGRESSION PROBLEM



13

UNSUPERVISED LEARNING

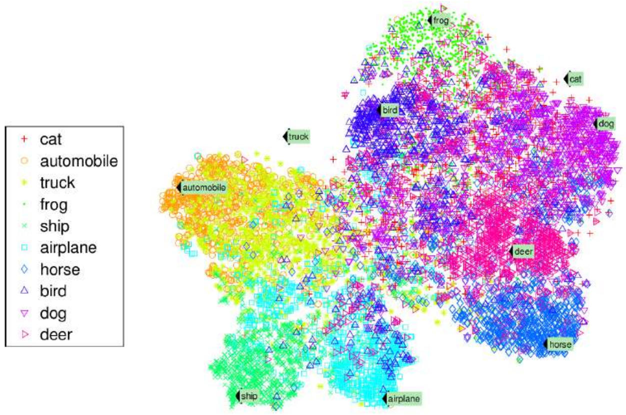
- Training data is unlabeled
- System tries to learn without a teacher
- Clustering
  - K-Means, DBSCASN, Hierarchical Cluster Analysis
- Anomaly detection
  - One-class SVM
- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
- Association rule mining
  - Apriori



14

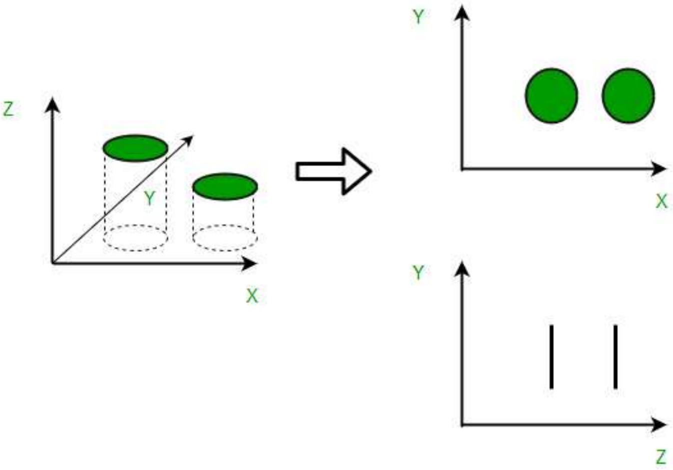
VISUALIZATION ALGORITHMS (E.G. T-SNE – T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

- Visualization algorithms such as t-SNE output a 2D or 3D representation of data given a lot of complex and unlabeled data
- Richard Socher et al., “Zero-Shot Learning Through Cross- Modal Transfer,” Proceedings of the 26th International Conference on Neural Information Processing Systems 1(2013): 935–943



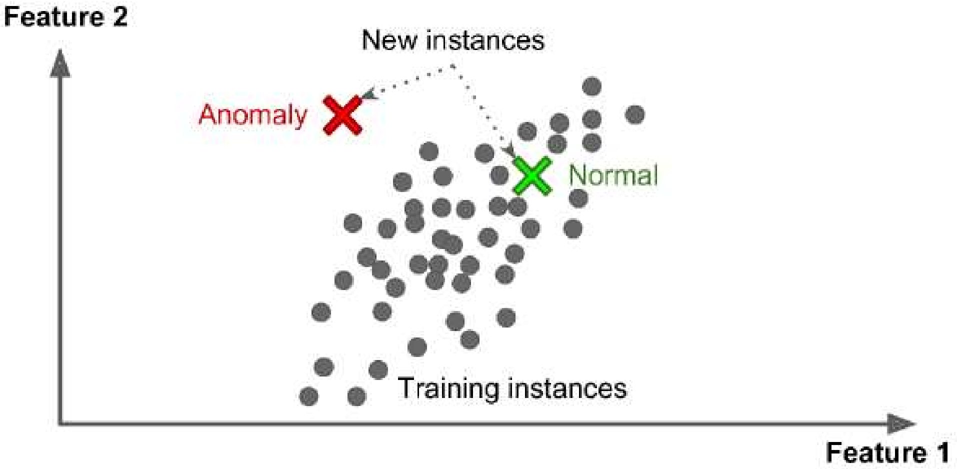
15

DIMENSIONALITY REDUCTION



16

ANOMALY DETECTION



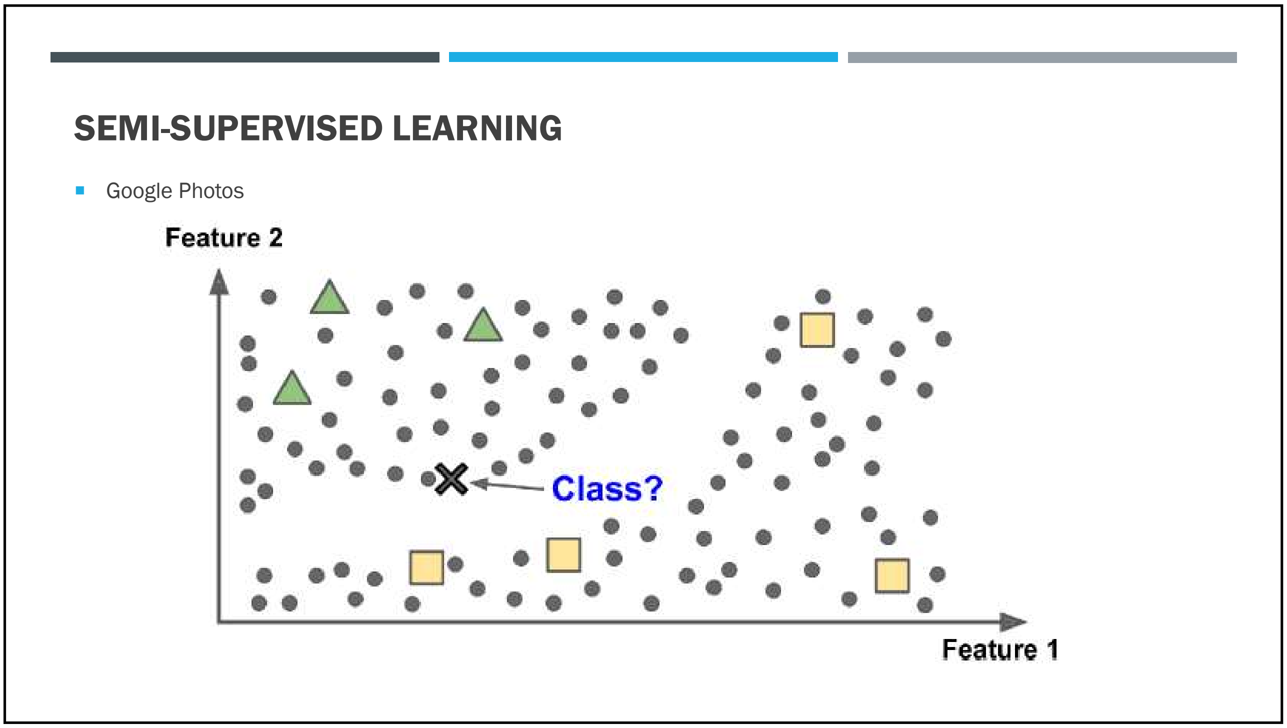
ASSOCIATION RULE LEARNING

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

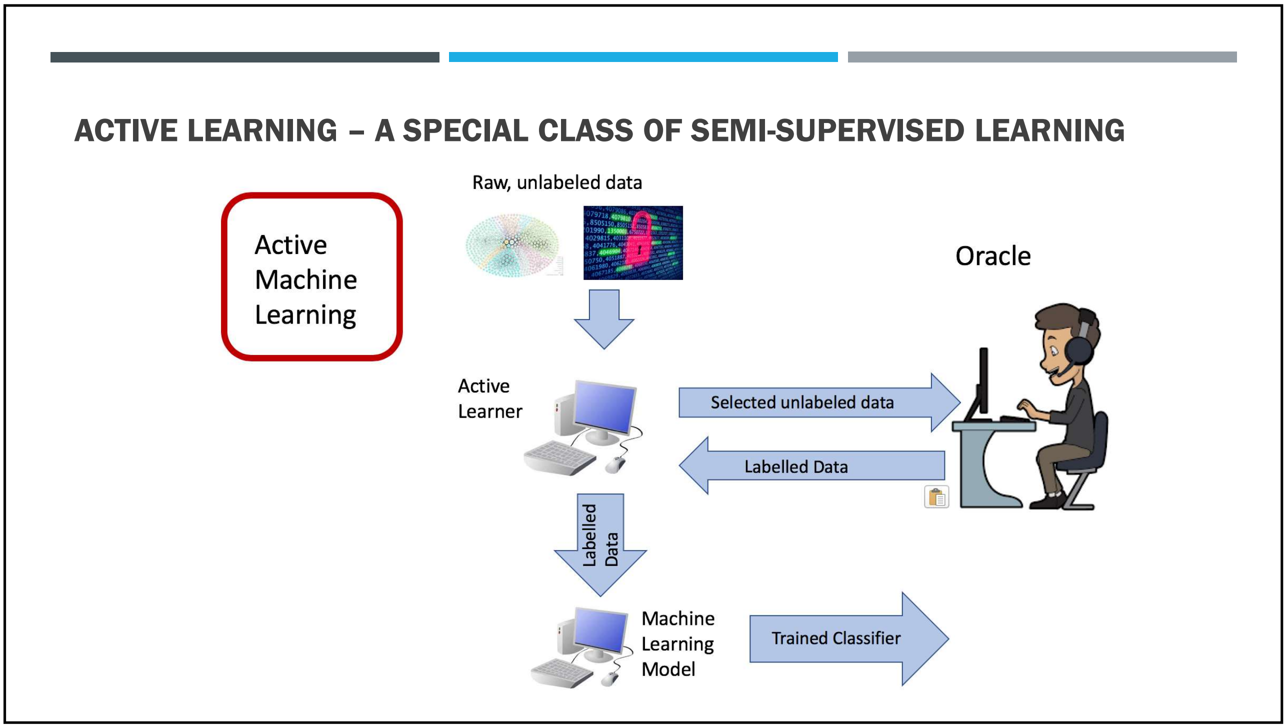
{Diapers} → {Beer} Example of an association rule



SEMI-SUPERVISED LEARNING

- Google Photos

19



ACTIVE LEARNING – A SPECIAL CLASS OF SEMI-SUPERVISED LEARNING

Active Machine Learning

Raw, unlabeled data

Oracle

Active Learner

Selected unlabeled data

Labelled Data

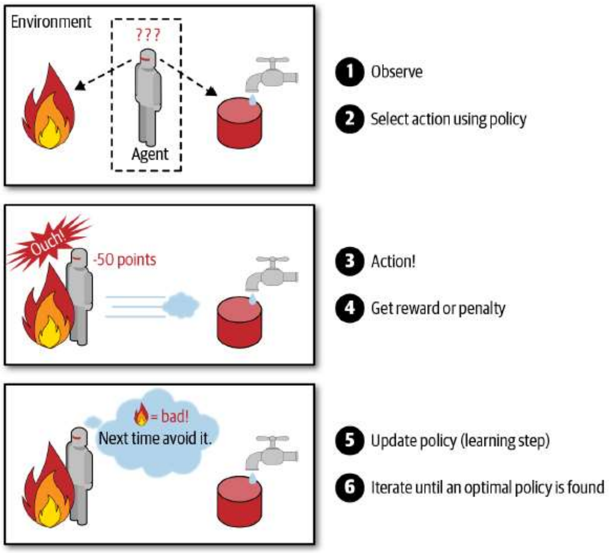
Labelled Data

Machine Learning Model

Trained Classifier

20

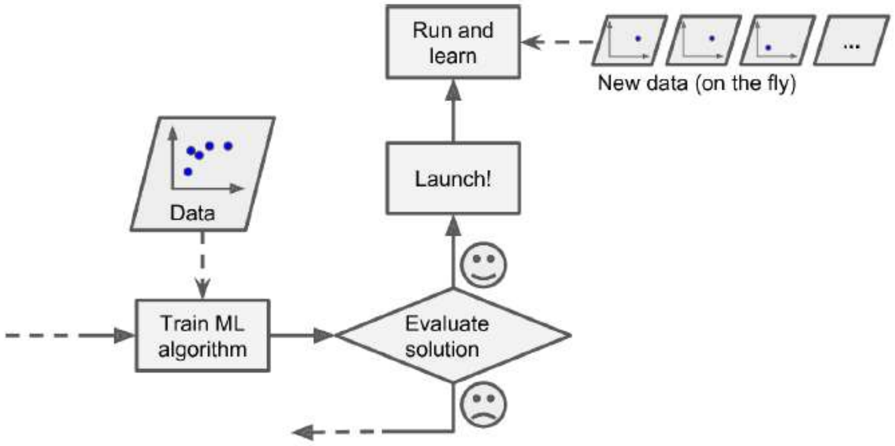
REINFORCEMENT LEARNING


Environment
???
Agent
1 Observe
2 Select action using policy
Ouch!
-50 points
3 Action!
4 Get reward or penalty
= bad!
Next time avoid it.
5 Update policy (learning step)
6 Iterate until an optimal policy is found

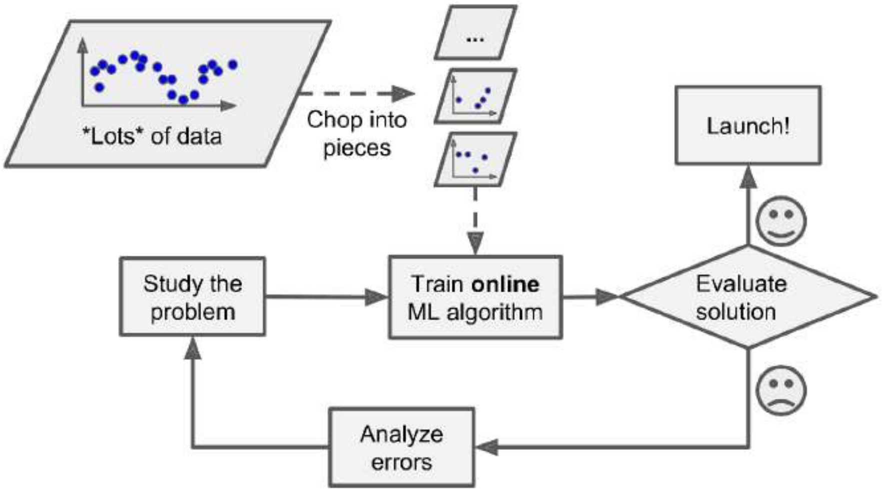
- Agent (Learning system) observes the environment, select and perform actions and gets rewards in return
- Creates a best strategy – called policy to get most reward over time
- Robotics
- DeepMind's AlphaGo program

BATCH LEARNING VS ONLINE LEARNING
GREAT FOR SYSTEMS THAT RECEIVE CONT. DATA FLOW- E.G. STOCK PRICES

- In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data.
- In online learning, a model is trained and launched into production. It keeps learning as new data comes in.

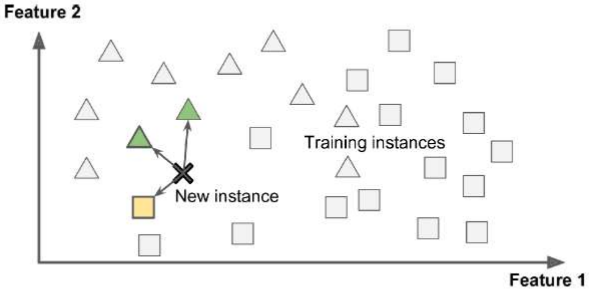


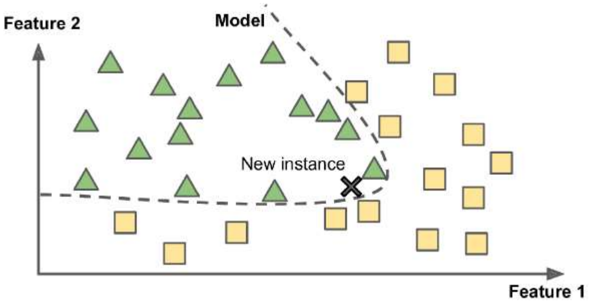
ONLINE LEARNING - HANDLING HUGE DATASETS



23

INSTANCE BASED LEARNING VS MODEL BASED LEARNING

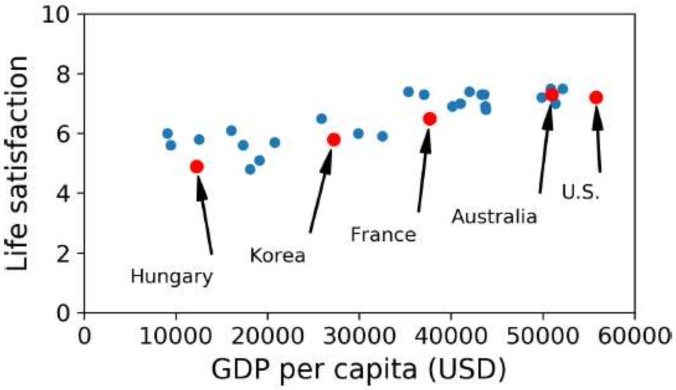




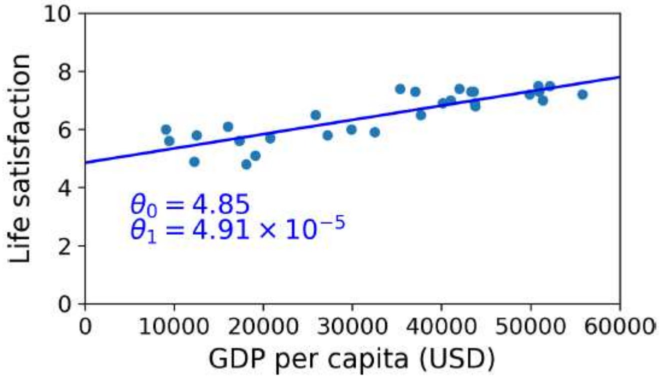
24

EXAMPLE OF A MODEL – DOES MONEY MAKE PEOPLE HAPPIER?

Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2

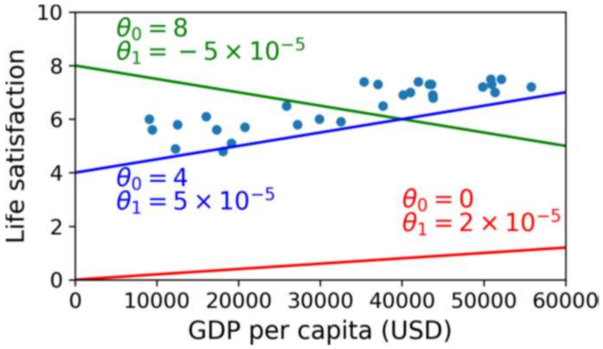


EXAMPLE OF A MODEL – DOES MONEY MAKE PEOPLE HAPPIER?



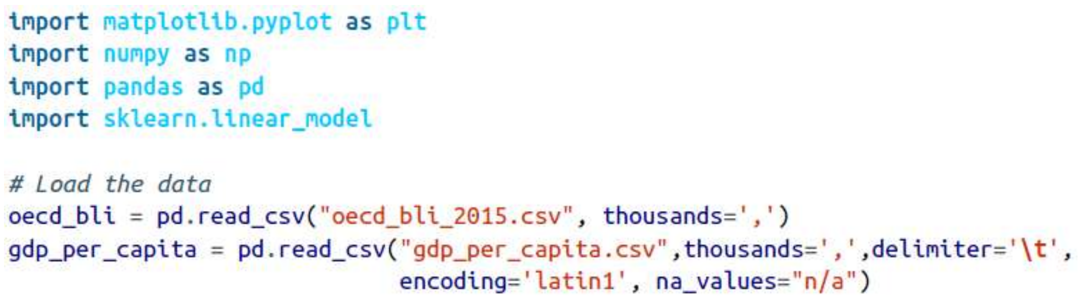
- Simple linear model

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



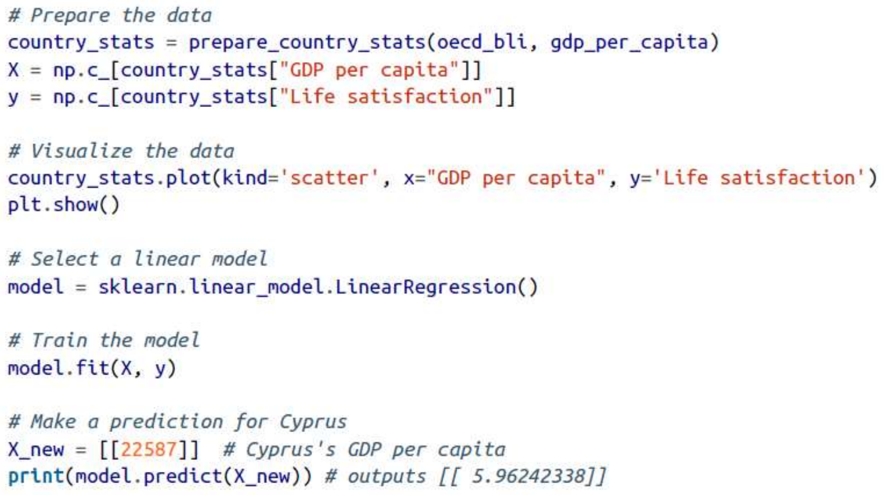
- Many models are possible
- Which one is the best?
- Define a cost function which measures how good/bad the model is
- For linear model, use cost function that measures the distance between the linear model's predictions and the training examples
- The objective is to minimize this distance.

USING SCIKIT-LEARN FOR LINEAR REGRESSION MODEL



```
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")
```

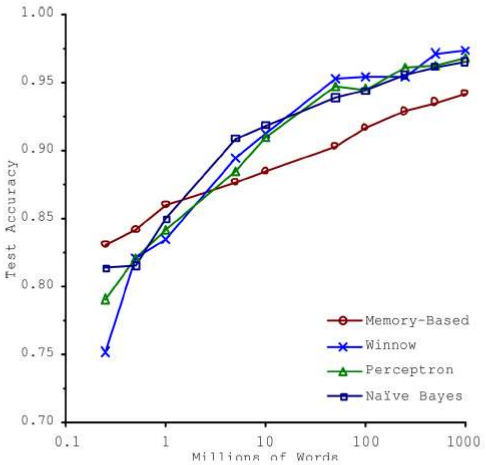
USING SCIKIT-LEARN FOR LINEAR REGRESSION MODEL

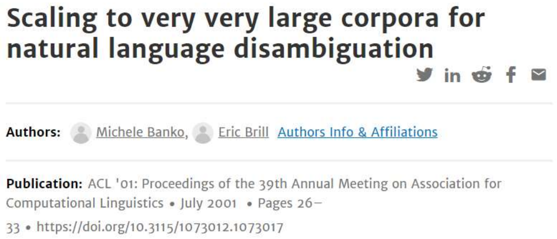


```
# Prepare the data
country_stats = prepare_country_stats(oecd_bli, gdp_per_capita)
X = np.c_[country_stats["GDP per capita"]]
y = np.c_[country_stats["Life satisfaction"]]

# 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's GDP per capita
print(model.predict(X_new)) # outputs [[ 5.96242338]]
```

29

ML CHALLENGES – INSUFFICIENT QUANTITY OF TRAINING DATA





Scaling to very very large corpora for natural language disambiguation

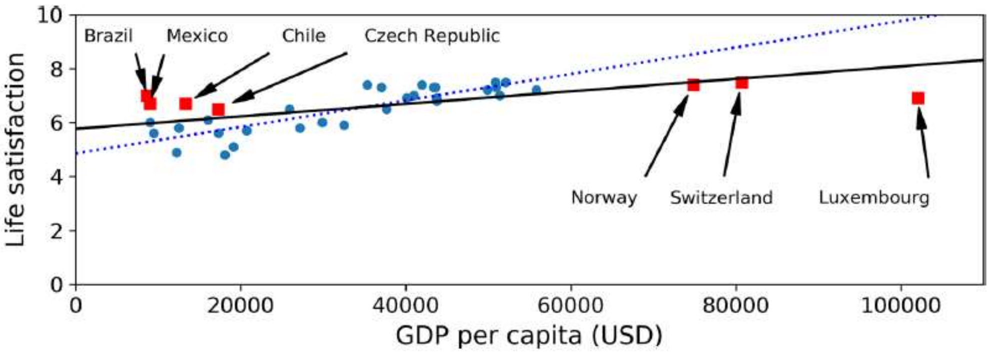
Authors: Michele Banko, Eric Brill Authors Info & Affiliations

Publication: ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics • July 2001 • Pages 26–33 • https://doi.org/10.3115/1073012.1073017

- Reconsider the trade-off between spending the time and money on algorithm development vs. corpus development
- However, small- and medium sized datasets are still very common, and it is not always easy or cheap to get extra training data

30

ML CHALLENGES – NONREPRESENTATIVE TRAINING DATA



- Use a training set that is representative of the cases you want to generalize to
- If the sample is too small, you will have sampling noise
- Very large samples can be nonrepresentative if the sampling method is flawed - known as sampling bias

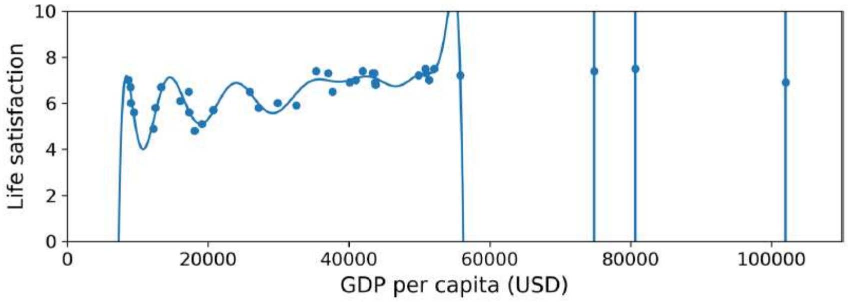
31

ML CHALLENGES

- Poor quality data
  - If the training data is full of errors, outliers, and noise the system will not be able to detect the underlying patterns
  - If some instances are clearly outliers, it may help to simply discard them or try to fix the errors manually
  - If some instances are missing a few features, you must decide whether to ignore this attribute altogether, ignore these instances, fill in the missing values, or train one model with the feature and one model without it
- Irrelevant features
  - Feature engineering is a process to come up with a good set of features to train on
  - Feature selection (selecting the most useful features)
  - Feature extraction (combining existing features)
  - Creating new features by gathering new data

32

OVERFITTING THE TRAINING DATA

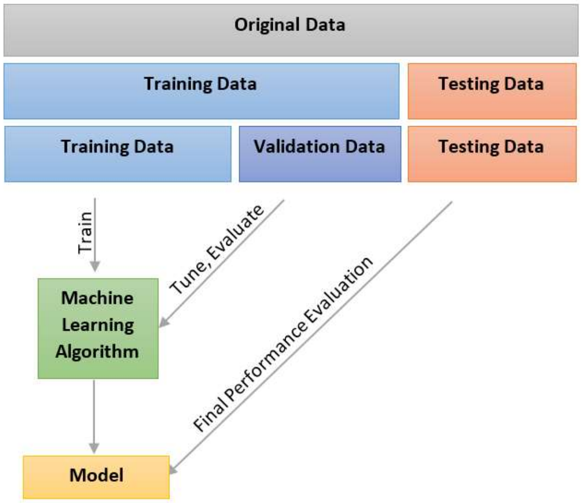


- Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. Possible solutions:
  - Simplify the model by selecting one with fewer parameters, by reducing the number of attributes in the training data
  - Constrain the model to make it simpler – also called regularizartion (we will see in detail in later lectures)
  - Gather more training data
  - Reduce the noise in the training data

UNDERFITTING THE TRAINING DATA

- *Underfitting* is the opposite of overfitting
- It occurs when your model is too simple to learn the underlying structure of the data.
- For example, a linear model of life satisfaction is prone to underfit; reality is just more complex than the model.
- Here are the main options for fixing this problem:
- Select a more powerful model, with more parameters.
- Feed better features to the learning algorithm (feature engineering)
- Reduce the constraints on the model

TESTING AND VALIDATING



Original Data

Training Data

Testing Data

Training Data

Validation Data

Testing Data

Train

Tune, Evaluate

Final Performance Evaluation

Machine Learning Algorithm

Model

NO FREE LUNCH THEOREM

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

⇩

Simplification is based on assumptions (model bias)

⇩

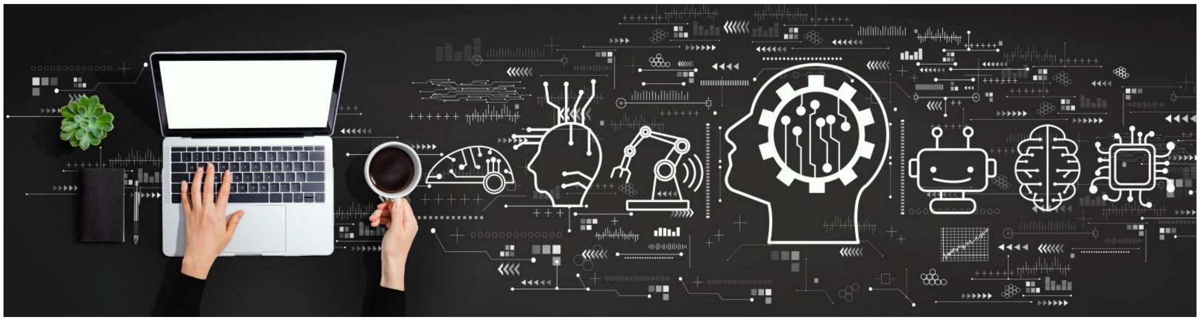
Assumptions fail in certain situations

Roughly speaking:

"No one model works best for all possible situations."

- There is no model that is a priori guaranteed to work better
- The only way to know for sure which model is best is to evaluate them all (not possible)
- In practice make some reasonable assumptions about the data and evaluate only a few reasonable models
  - E.g. for simple tasks you may evaluate linear models with various levels of regularization
  - For a complex problem use various neural networks

QA



37