JADS Jheronimus Academy of Data Science

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Wine Quality Classification

Hands-on

Introduction to Data Science Elective

On behalf of the PDEng Data Science

5 November 2019

Agenda

Yesterday:

Exploratory analysis of the Wine Quality dataset.

Today:

Machine learning: wine quality classification.

Learning from data

Inputs and Outputs

- **Inputs**
  - Words, sentences
  - Images, videos
  - Sensor observations, time series
  - Voice

$\mathbf{x}$

- **Outputs**
  - Class label

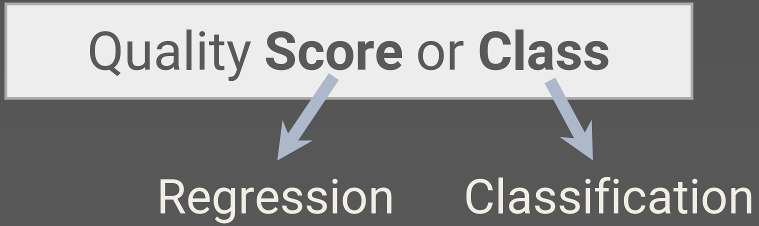
$\mathbf{y}$

Inputs and Outputs in the Wine Data

Inputs

Fixed acidity	Total sulfur dioxide
Volatile acidity	Density
Citric acid	pH
Residual sugar	Sulphates
Chlorides	Alcohol
Free sulfur dioxide	

Outputs



Quality **Score** or **Class**

Regression

Classification

What is learning?

- **Learning** or **training** refers to estimate the parameters of a model.

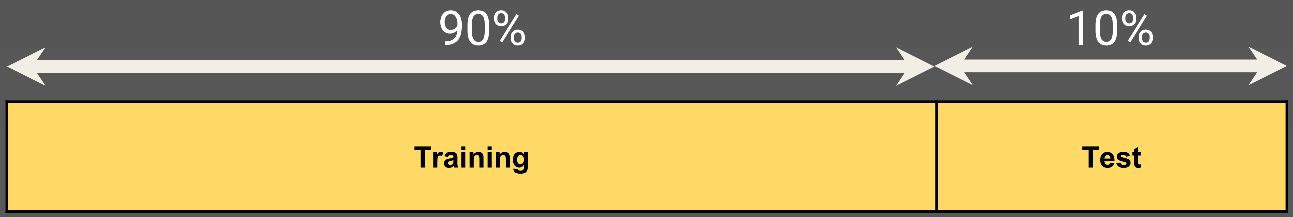
Dataset

Machine learning is about generalizing to unseen data.

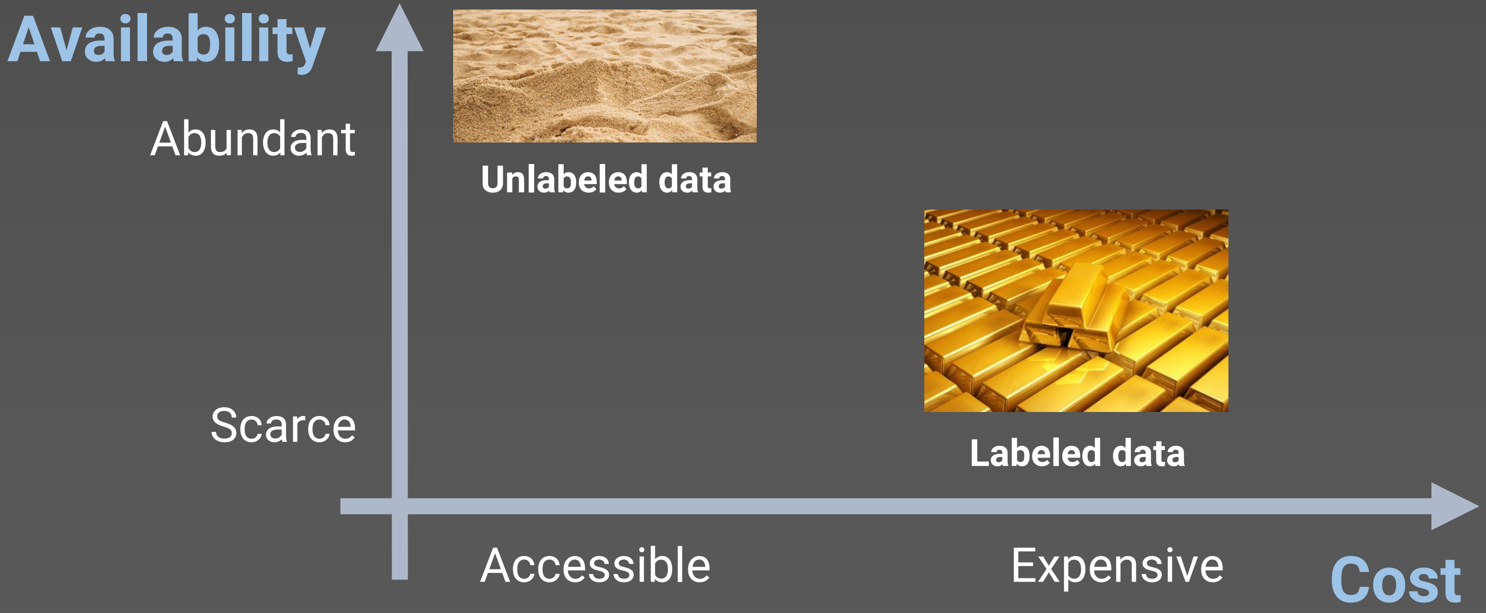
Split Data

- Split dataset into a training and test set:

```
x_train, x_test, y_train, y_test = train_test_split(data_features,
data['grade'], test_size=0.1)
```



What data is out there?



Availability

Abundant

Scarce

Unlabeled data

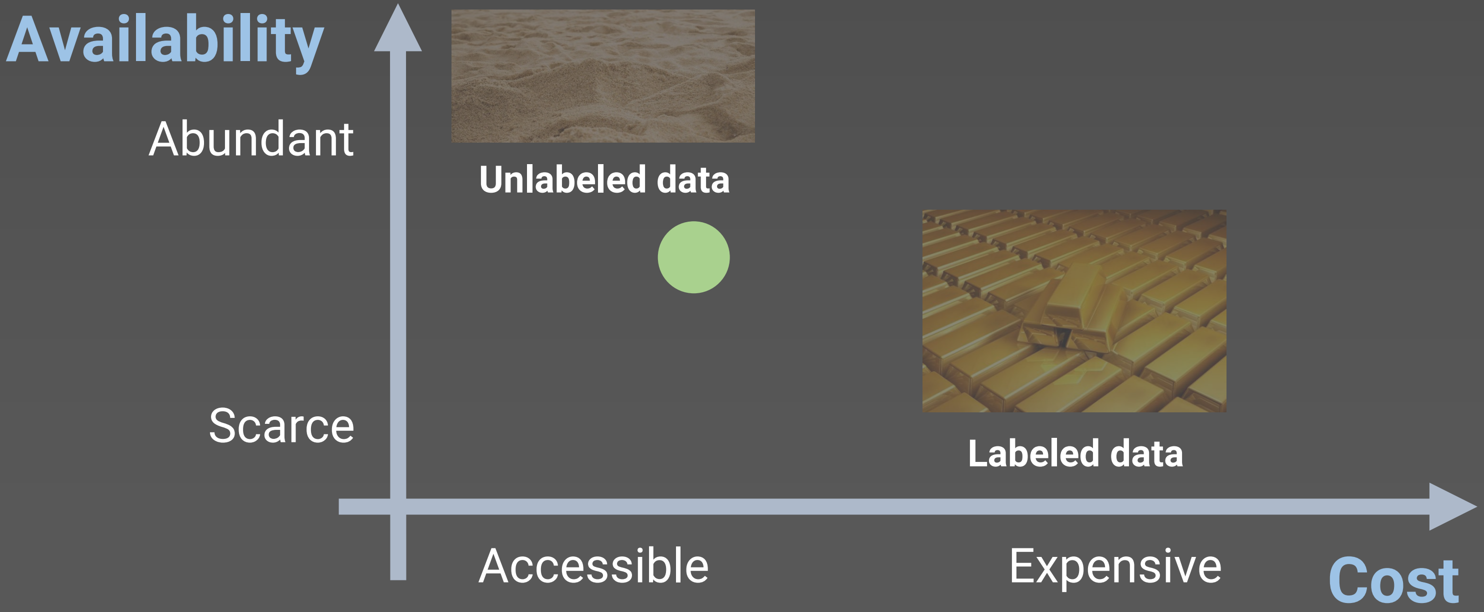
Labeled data

Accessible

Expensive

Cost

Data in the PDEng Data Science



Availability

Abundant

Scarce

Unlabeled data

Labeled data

Accessible

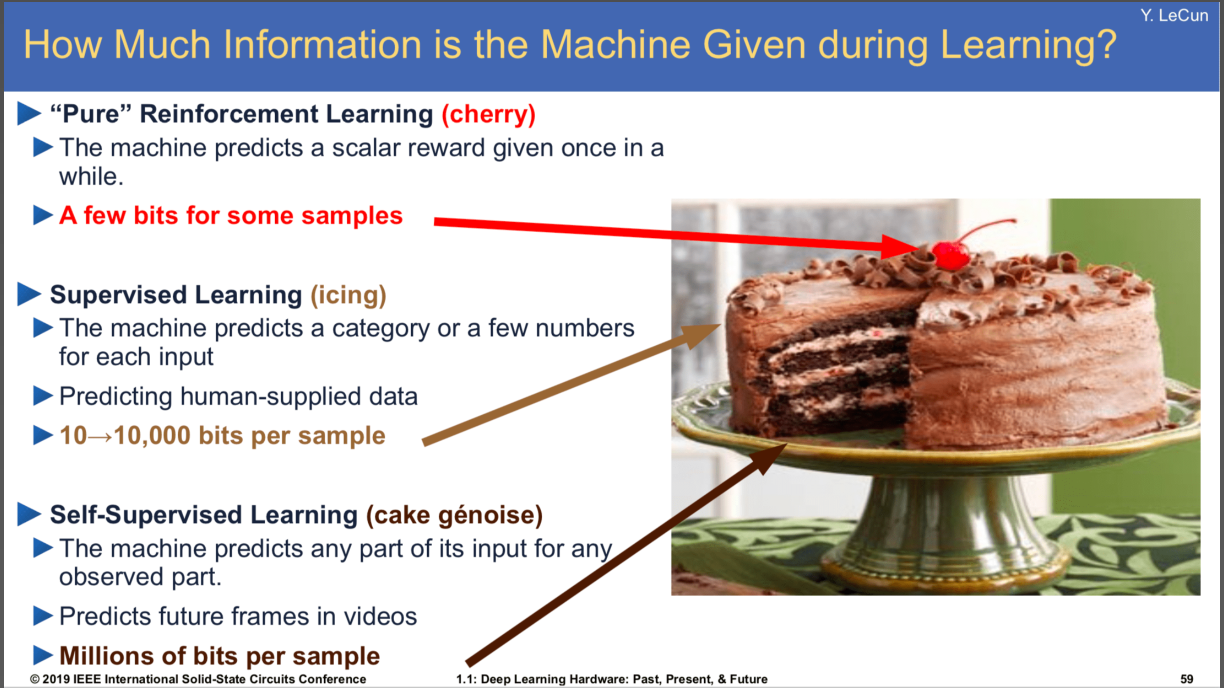
Expensive

Cost

Types of learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

LeCun’s Cake Analogy



Y. LeCun

How Much Information is the Machine Given during Learning?

- “Pure” Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - A few bits for some samples
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - 10→10,000 bits per sample
- Self-Supervised Learning (cake génoise)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - Millions of bits per sample

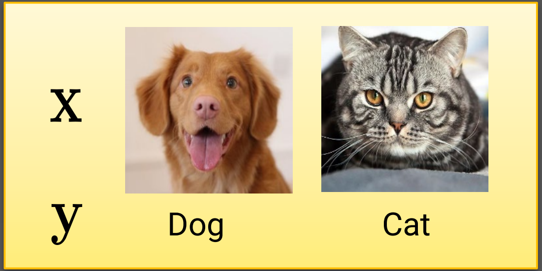
© 2019 IEEE International Solid-State Circuits Conference

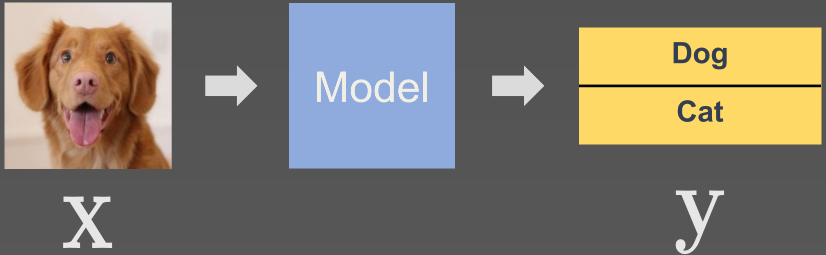
1.1: Deep Learning Hardware: Past, Present, & Future

59

Supervised Learning

- Given a dataset D of inputs x and labeled targets y, ***learn*** **to predict** y from x.

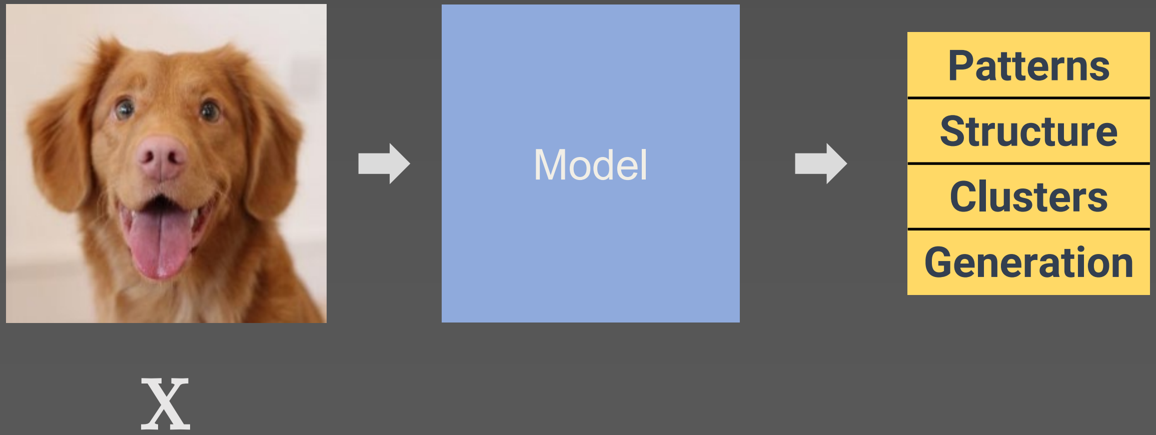




- Most successful paradigm in machine learning.

Unsupervised Learning

- Given only the inputs x, models $p(\mathbf{x})$ and find



$\mathbf{x}$

Model

Patterns

Structure

Clusters

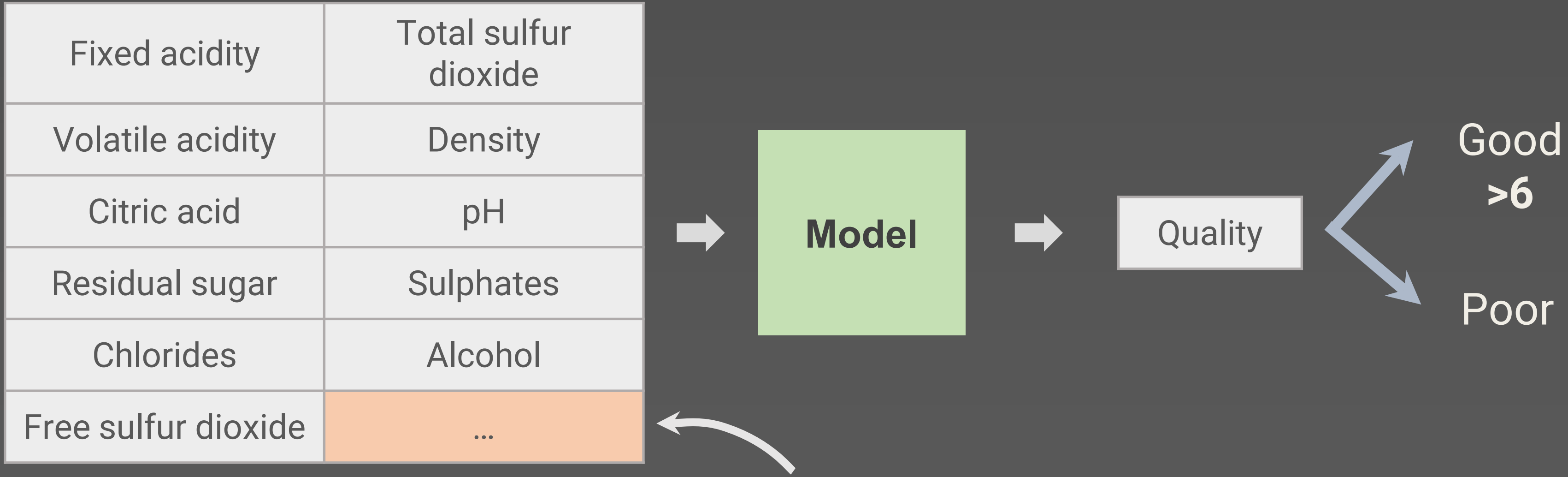
Generation

Wine Quality Dataset

Supervised

Inputs

Outputs



Fixed acidity	Total sulfur dioxide
Volatile acidity	Density
Citric acid	pH
Residual sugar	Sulphates
Chlorides	Alcohol
Free sulfur dioxide	…

Handcrafted features

Example: Decision Tree

Model

```
decisiontree = DecisionTreeClassifier(max_depth=4)

decisiontree.fit(x_train_values, y_train)

decisiontree_pred = decisiontree.predict(x_test_values)
```

Learn decision tree on the **training data**

Predict wine quality on the **test data**

How to evaluate our model?

Use a performance metric (e.g., accuracy)

Number of correct predictions made divided by the total.

True value

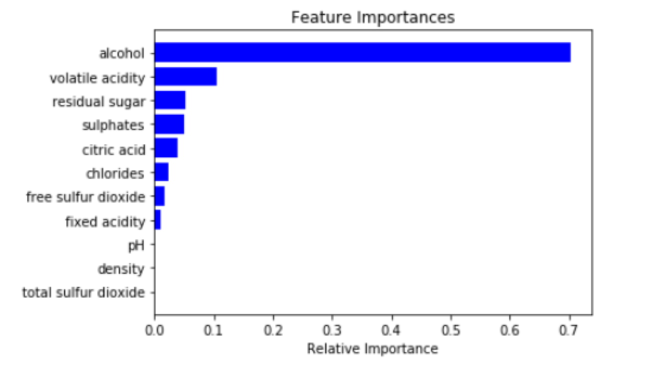
```
accuracy_score(y_test, decisiontree_pred)*100
```

Predicted value

How to interpret our model?

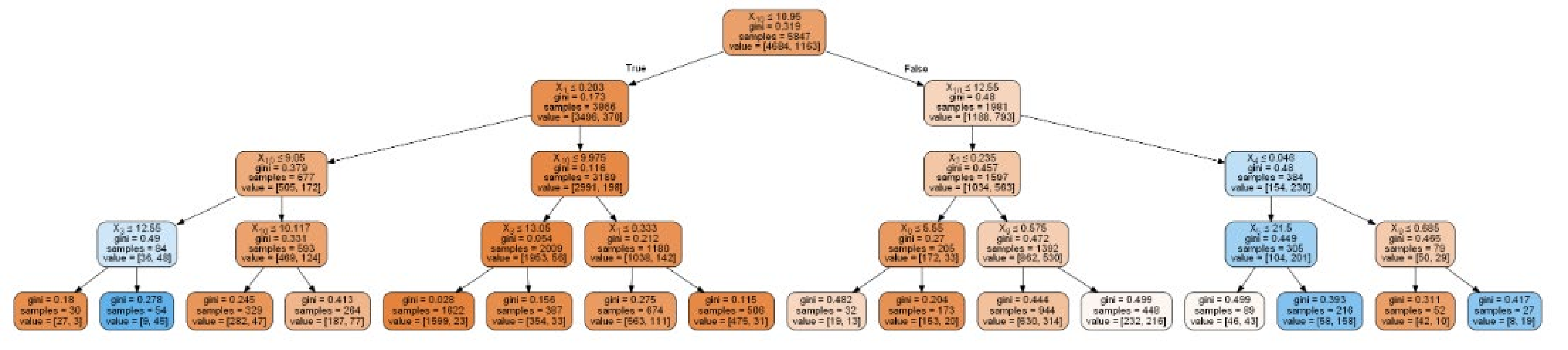
In a decision tree, compute **feature importance**.

```
importances = decisiontree.feature_importances_
indices = np.argsort(importances)
```



How to visualize our model?

Visualize the decision tree!


$X_{10} \leq 10.95$ gini = 0.319 samples = 5847 value = [4684, 1163]
True
False
$X_{1} \leq 0.203$ gini = 0.173 samples = 3866 value = [3496, 370]
$X_{10} \leq 12.55$ gini = 0.48 samples = 1981 value = [1188, 793]
$X_{10} \leq 9.05$ gini = 0.379 samples = 677 value = [505, 172]
$X_{10} \leq 9.975$ gini = 0.116 samples = 3189 value = [2991, 198]
$X_{9} \leq 0.235$ gini = 0.457 samples = 1597 value = [1034, 563]
$X_{4} \leq 0.046$ gini = 0.48 samples = 384 value = [154, 230]
$X_{3} \leq 12.55$ gini = 0.49 samples = 84 value = [36, 48]
$X_{10} \leq 10.117$ gini = 0.331 samples = 593 value = [469, 124]
$X_{3} \leq 13.05$ gini = 0.054 samples = 2009 value = [1953, 56]
$X_{1} \leq 0.333$ gini = 0.212 samples = 1180 value = [1038, 142]
$X_{0} \leq 5.55$ gini = 0.27 samples = 205 value = [172, 33]
$X_{9} \leq 0.575$ gini = 0.472 samples = 1392 value = [862, 530]
$X_{5} \leq 21.5$ gini = 0.449 samples = 305 value = [104, 201]
$X_{0} \leq 0.685$ gini = 0.465 samples = 79 value = [50, 29]
gini = 0.18 samples = 30 value = [27, 3]
gini = 0.278 samples = 54 value = [9, 45]
gini = 0.245 samples = 329 value = [282, 47]
gini = 0.413 samples = 264 value = [187, 77]
gini = 0.028 samples = 1622 value = [1599, 23]
gini = 0.156 samples = 387 value = [354, 33]
gini = 0.275 samples = 674 value = [563, 111]
gini = 0.115 samples = 506 value = [475, 31]
gini = 0.482 samples = 32 value = [19, 13]
gini = 0.204 samples = 173 value = [153, 20]
gini = 0.444 samples = 944 value = [630, 314]
gini = 0.499 samples = 448 value = [232, 216]
gini = 0.499 samples = 89 value = [46, 43]
gini = 0.393 samples = 216 value = [58, 158]
gini = 0.311 samples = 52 value = [42, 10]
gini = 0.417 samples = 27 value = [8, 19]

Support Material

Jupyter Notebook:

Wine Quality classification using decision tree.

Available in the shared folder @ https://bit.ly/34r6YUs

References

- **A Few Useful Things to Know About Machine Learning**, Domingos, 2012 (Link)

- Dataset source (Link)
- Modelling wine preferences by data mining from physicochemical properties (Link)
- Predicting wine quality using data analytics (Link)

References

- Predicting quality of wine based on chemical attributes (Link)
- Data analysis on the wine dataset (Link)
- Wine Quality Classification (Link)
- Vinho Verde webpage (Link)

Thank you for your attention!