Hasso Plattner Institut
Digital Engineering · Universität Potsdam

# Data Management for Digital Health
# Revision of Exercise IV

Borchert, Rasheed, Dr. Bayat, Dr. Schapranow

Data Management for Digital Health
Winter 2022

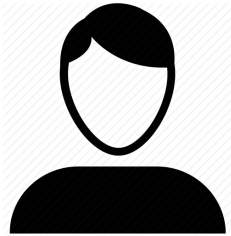
# Exercise III
# Topics

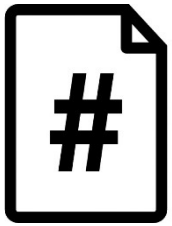
- Medical Use Case Infectious Diseases
- Medical Imaging
- Deep Learning
- Unsupervised Learning

# Exercise III
# Key Stats

Hasso Plattner Institut

| 25 Questions<br>50 Points | 40<br>Submissions<br>40 Passed | Average score<br>43.55 / 87.1% | Average time<br>57.6 min | << 3h |
|---|---|---|---|---|





# Q3: What are effective measures to generally prevent spread of infectious diseases especially if you have no treatment at hand?

Hasso Plattner Institut

- ✗ Wearing a mask
- ✗ Use of high-dose antimycotics and antibiotics combinations to prevent additional infections
- ✓ **Systematic contact tracing of infected persons**
- ✓ **Monitoring of persons suspected to be in close contact with infected persons for symptoms over the incubation time.**

Frequently missed

Frequent incorrect answer

# What To Take Home?

- Infections may
  - Affect different body locations
  - Be triggered by numerous agents
  - Result in life-threatening events
  - Require intensive care

- If no therapy available: containment is the **only option** to fight a pandemic spread.

# Infectious Diseases: Definitions

- Infection := Invasion + multiplication of disease-causing agents + host reaction
- Disease-causing agents := Pathogenic microorganisms, e.g. bacteria, viruses, parasites or fungi
- Infectious diseases can be spread directly or indirectly from one person to another
- Zoonotic diseases := Infectious diseases transmitted from animals to humans or vice versa.

# Q4: Please select all correct options for vaccinations as discussed in class.

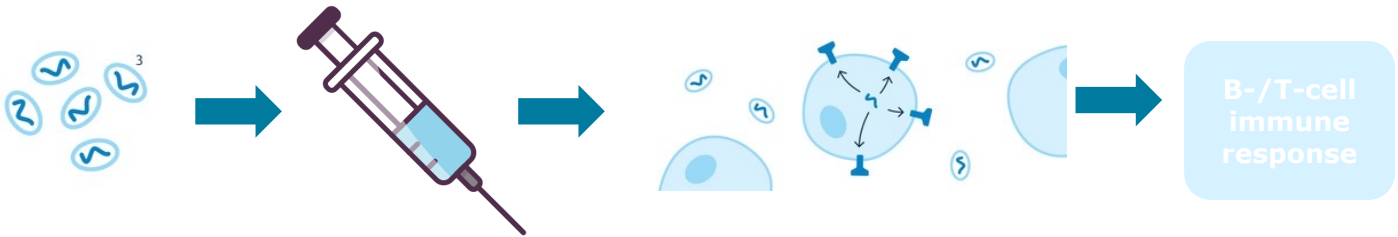
- ✗ Inactivated vaccines are more efficient than mRNA-based vaccines.
- ✓ Vector-based vaccines make use of a non-pathogenic/disabled vector to invade body cells and release the disease-specific RNA there.
- ✗ Due to the high resistance of RNA molecules, mRNA-based vaccines can be stored at room temperature over a long period of time (months and year) before its use (long durable vaccines)
- ✗ A single vaccination dose is sufficient to obtain life-long immunity for most infectious diseases.

Frequently missed

Frequent incorrect answer

# Types of Vaccination: mRNA Vaccine

- Requires synthesis and handling of very fragile RNA




- mRNA for spike protein covered by lipid hull

- Lipid hull allows to mRNA to enter body cells
- Ribosomes read RNA and assemble spike protein
- Proteins are released by cell

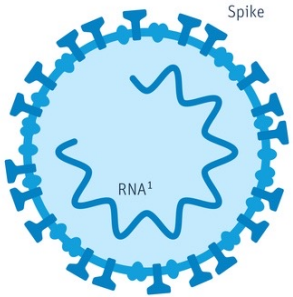



Image Credits: vfa, https://icon-icons.com/

# Q13: Which statements are true about unsupervised learning? Select all that apply

Hasso Plattner Institut

- ✗ Expert annotators are required to define baseline parameters.
- ✓ Clustering is a type of unsupervised learning.
- ✗ Explicit labels are required for unsupervised learning.
- ✗ Unsupervised learning always needs the number of clusters as input.



Frequently missed

Frequent incorrect answer

# Most common problem settings in Machine Learning

Hasso Plattner Institut

## Supervised Learning (Labels available for training)

### Classification

Categorical output

e.g. x ∈ Fruits, y ∈ {"apple", "orange"}

f( )= "apple"

f( )= "orange"

### Regression

Continuous output

e.g.: x ∈ Fruits, y ∈ $\mathbb{R}_+$ ≙ t until ripe)

f ( ) = 12 days

### Structured Prediction

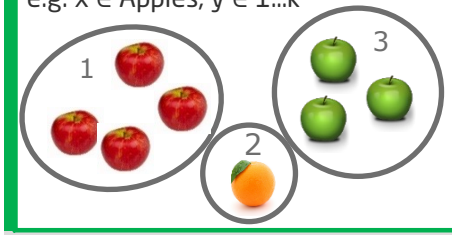
e.g. x ∈ $\mathbb{R}^{w \times h \times d}$, y ∈ $\mathbb{R}^{w \times h}$ ≙ pixels

f( ) =

## Unsupervised Learning (No labels during training)

### Clustering

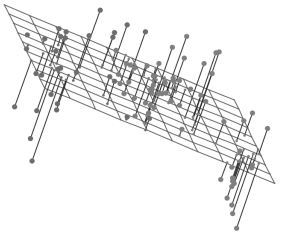
e.g. x ∈ Apples, y ∈ 1...k




### Dimensionality reduction

$x \in \mathbb{R}^d$, $x' \in \mathbb{R}^p$, $p < d$

e.g., projecting all features of a fruit to 2 dimensions for visualization

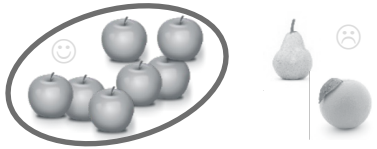


## Semi-Supervised Learning (Some labels for training)

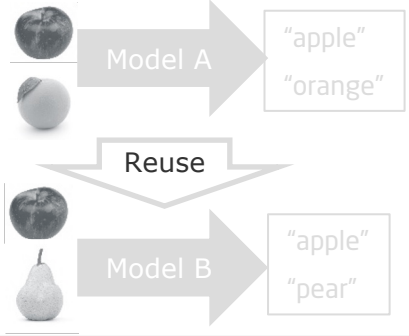
### Anomaly / novelty detection

trained only on "normal" samples

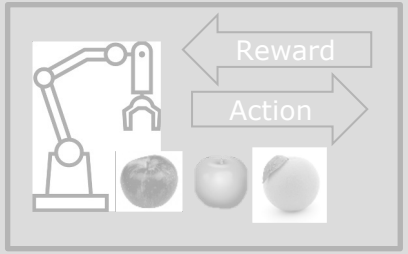
e.g. x ∈ Apples, y ∈ { ☺ , ☹ }



## Transfer Learning




## Reinforcement Learning




https://en.wikipedia.org/wiki/Apple
https://cdn4.vectorstock.com/i/1000x1000/16/58/robot-arm-line-icon-sign-on-vector-17841658.jpg

**Evaluation Exercise IV**

Data Management for Digital Health, Winter 2022

# DBSCAN: Pros and cons

- Pros:
  - Do not need number of cluster
  - Could identify noise in data
- Cons
  - Only assign one data to one cluster point. No hierarchy
  - Slow to run on large amounts of data
- Other density-based clustering algorithms
  - HDBscan: Improved version of Dbscan uses hierarchy of clusters
  - Optics: Uses K-nearest neighbour search to improve algorithm

# Q15: What are evaluation metrics for clustering algorithms? Select all that apply:

Hasso Plattner Institut

- ✗ Euclidean distance
- ✗ Pearson's correlation coefficient
- ✓ Rand index
- ✗ Manhattan distance

Frequently missed

Frequent incorrect answer

# Entrinsic Evaluation

- Evaluated based on the data that was clustered itself
- Ground truth not needed
- Assigns the best score to clusters with high similarity within a cluster and low similarity between clusters.
- Silhouette Coefficient
- Dunn Index
- Davies-Bouldin index

# Extrinsic Evaluation

- Uses ground truth to perform the cluster evluation
- Evaluates the ability of clustering algorithms ability to separate class compared to ground truth
- Rand Index: Measures the similarity of the two assignments.
- Mutual information based: Uses ideas from Shannon's entropy
  - Mutual Information Gain
  - Homogeneity, completeness, and V-measure

$$H(U) = -\sum_{i=1}^{|U|} P(i)\log(P(i))$$

# Q21: Select all correct answers about Convolutional Neural Networks (CNN)

Hasso Plattner Institut

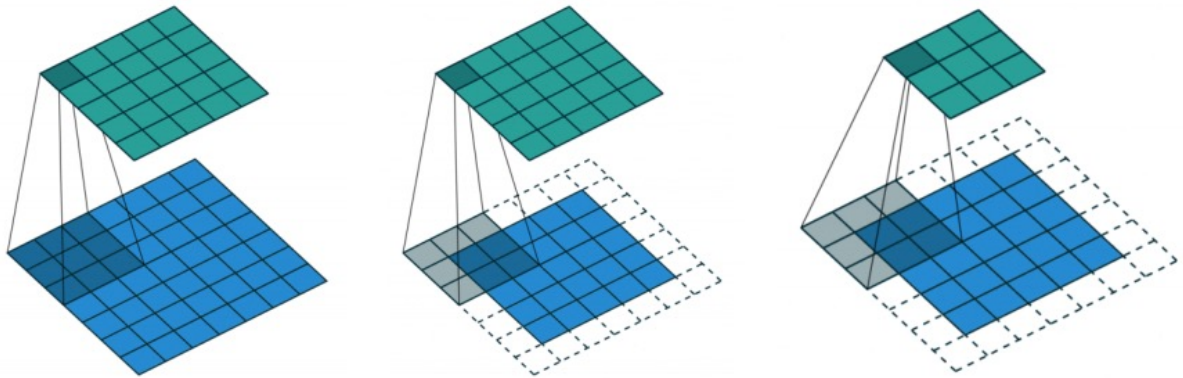
- ✗ The values of filters are fixed during training of a CNN, as it is not a learnable parameter.
- ✓ Pooling is a type of down sampling method used in CNN.
- ✓ When defining a CNN architecture, one should specify the number of filters along with the filter size and stride of each filter.
- ✗ Padding in a convolution layer averages all the value in a window.

Frequently missed

Frequent incorrect answer

# CNN
# Convolution Layer

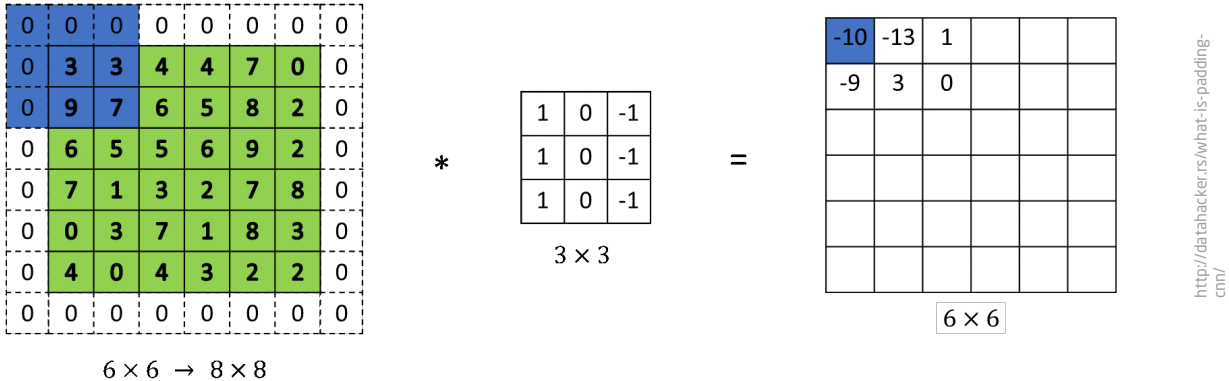
- In practice, a CNN learns the values of these filters on its own during the training process
- Although we still need to specify parameters such as number of filters, filter size, padding, and stride before the training process



https://lipyeow.github.io/ics491f17/morea/deepnn/ImageClassification-CNN.pdf

# Convolution Layer
# Padding

- Adding extra pixels of filler around the boundary of input image → Increasing the effective size of the image
- Typically, set values of the extra pixels to zero
- **Valid convolutions** ↔ no padding
- **Same convolution** ↔ pad so that output size is the same as the input size




http://datahacker.rs/what-is-padding-cnn/