Advanced Statistics & Machine Learning

A focus on Particle Physics Applications

HASCO Summer School 2024

M. Sc. Stephanie Käs
JLU Gießen & RWTH Aachen

Göttingen
30.07.24

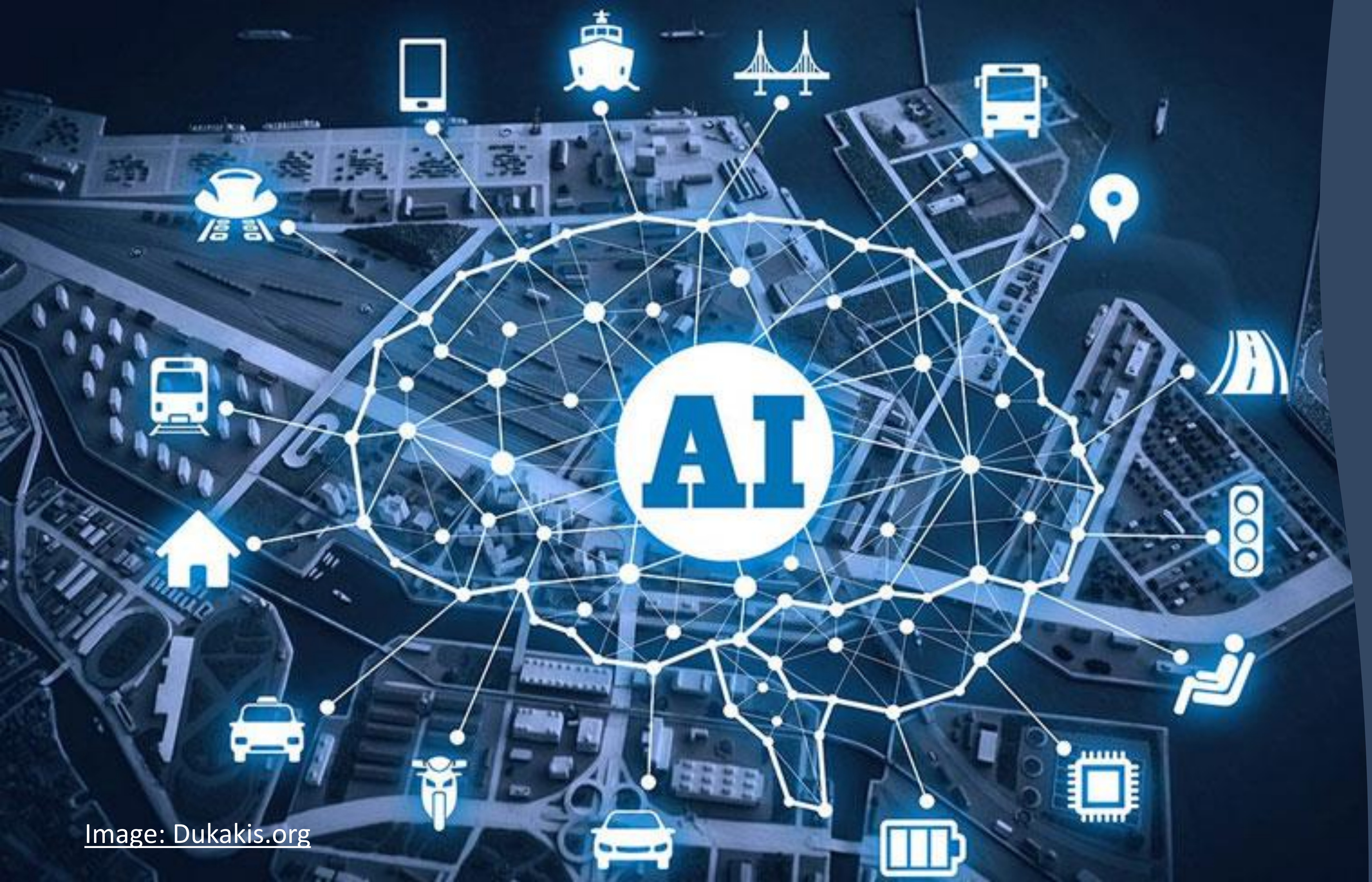
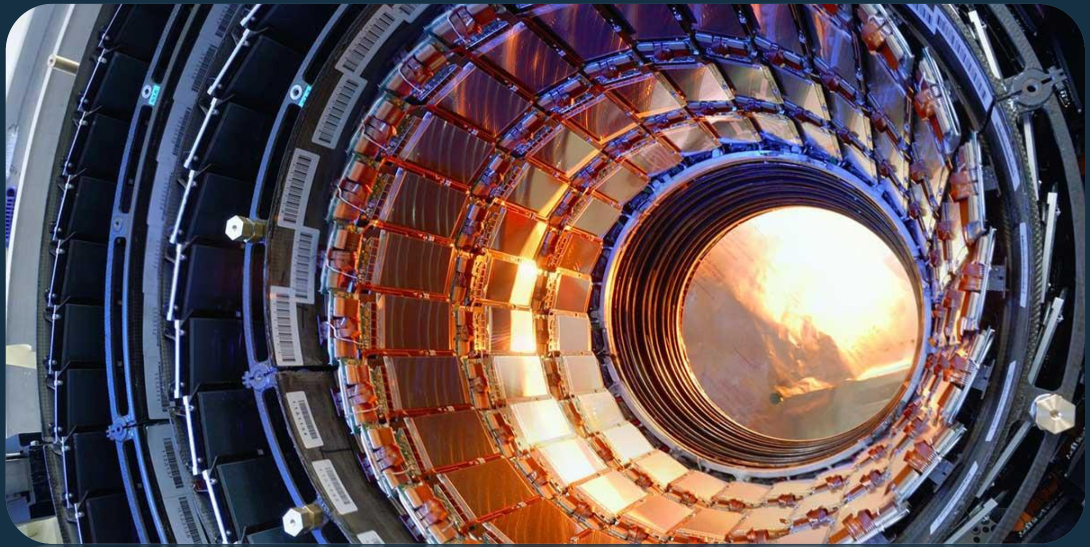


Image: Dukakis.org

Did you know?

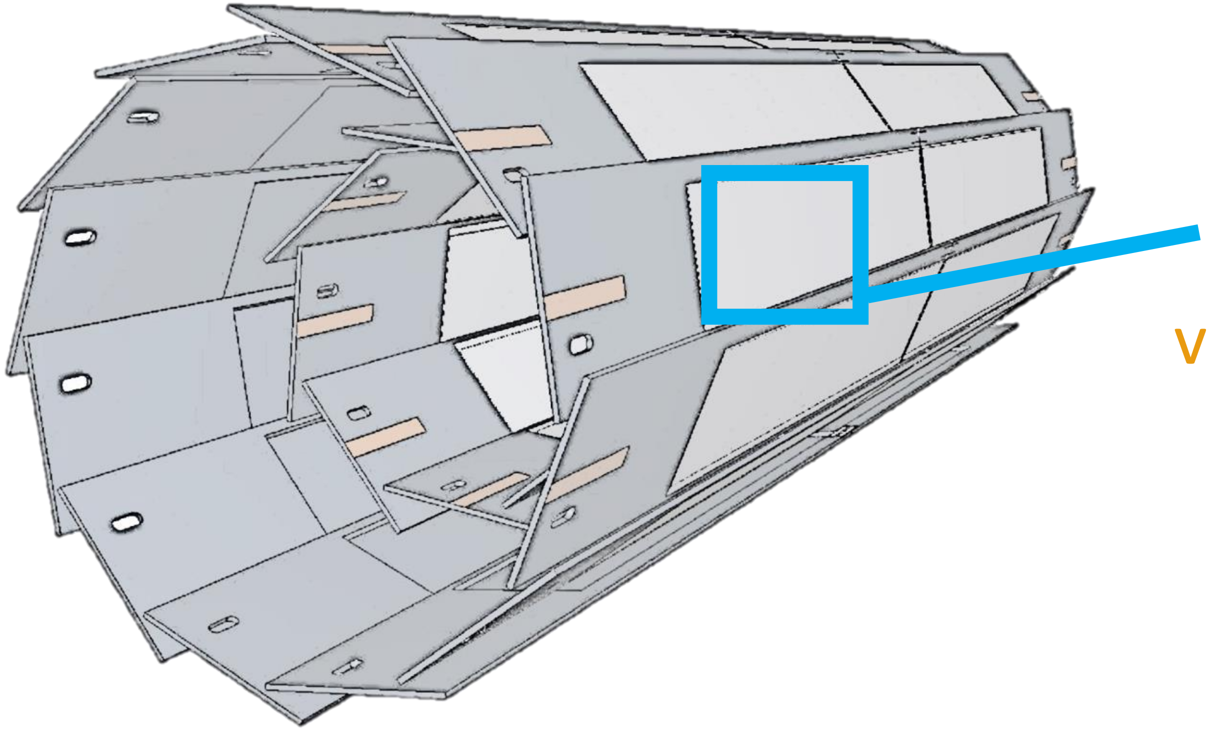


*Particles collide in the LHC detectors approximately 1 billion times per second, generating about **one petabyte of collision data per second**.*

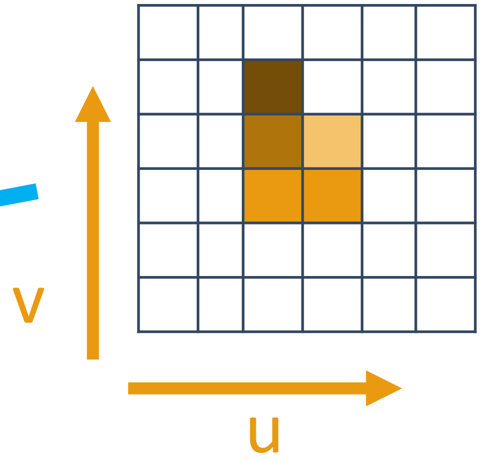
Image: uni-wuppertal.de

Example – Data Analysis in Particle Physics

Pixeldetector (JLU Gießen / Belle II Experiment Japan)



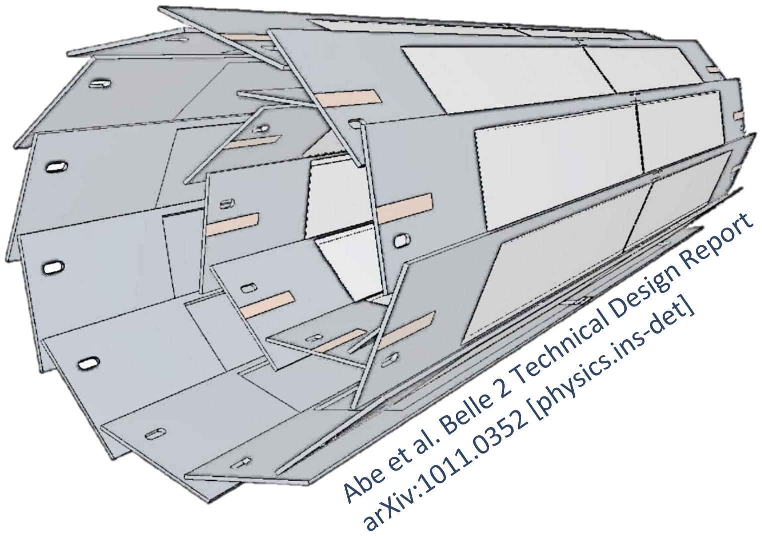
belle2.desy.de

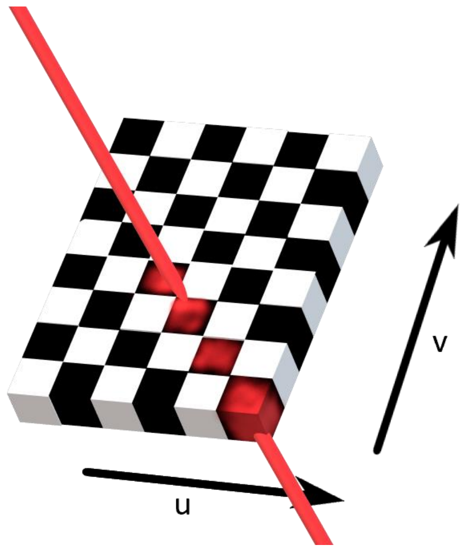


Pixel color = deposited charge

Spoiler: Here Neural Networks have not been the best choice.

Belle II Pixeldetector



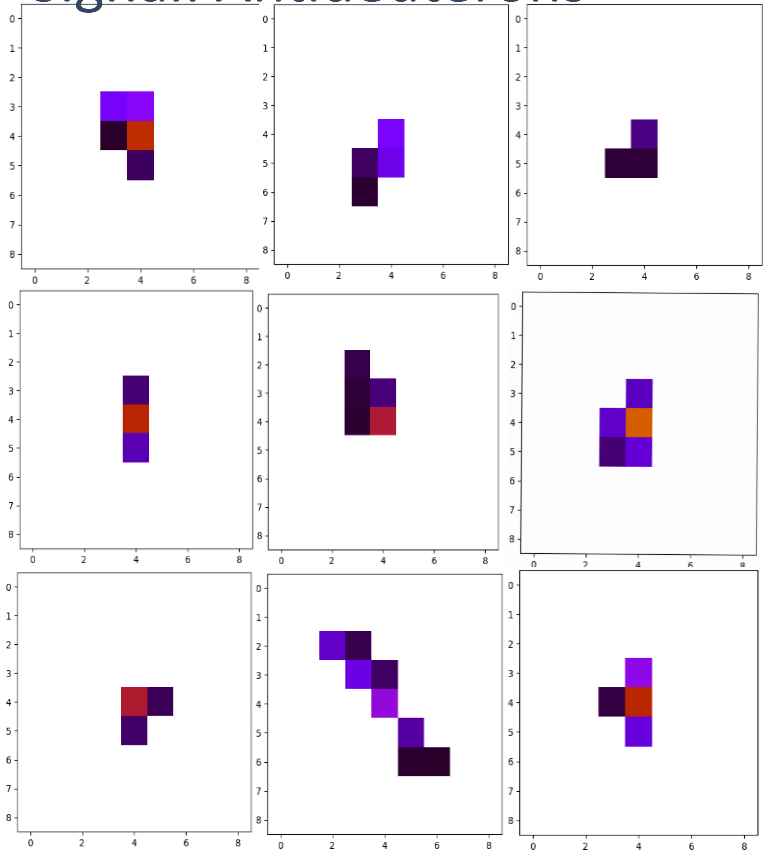


- Innermost detector
- Pixelated silicon sensors (PXD)
- 2 layers of 40 sensors each
- 8 M pixels

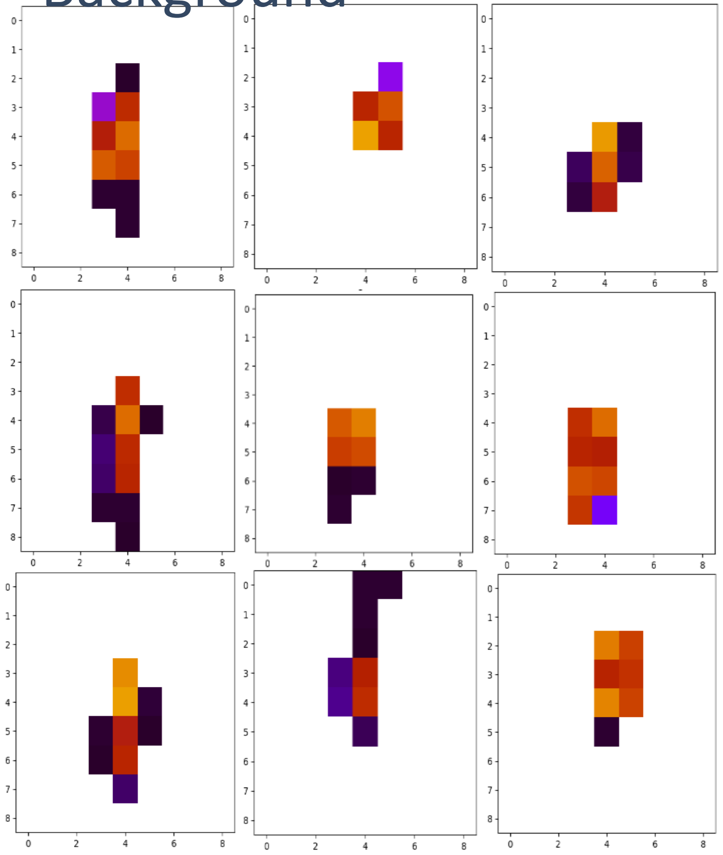
Captures highly ionizing particles.

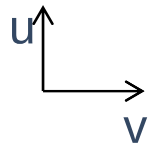
PXD Clusters

Signal: Antideuterons

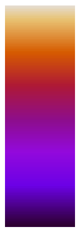


Background





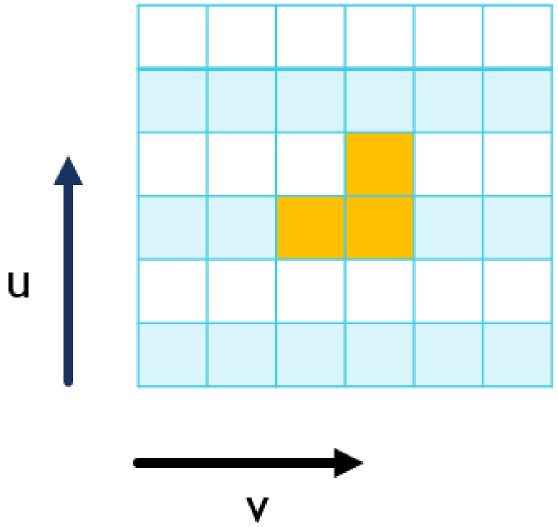
9x9 matrix
ADC values



M. Peter, Unpublished

Belle II Pixeldetector

Cluster properties



Total charge	Minimum charge	Maximum charge
Total size	Size in u	Size in v

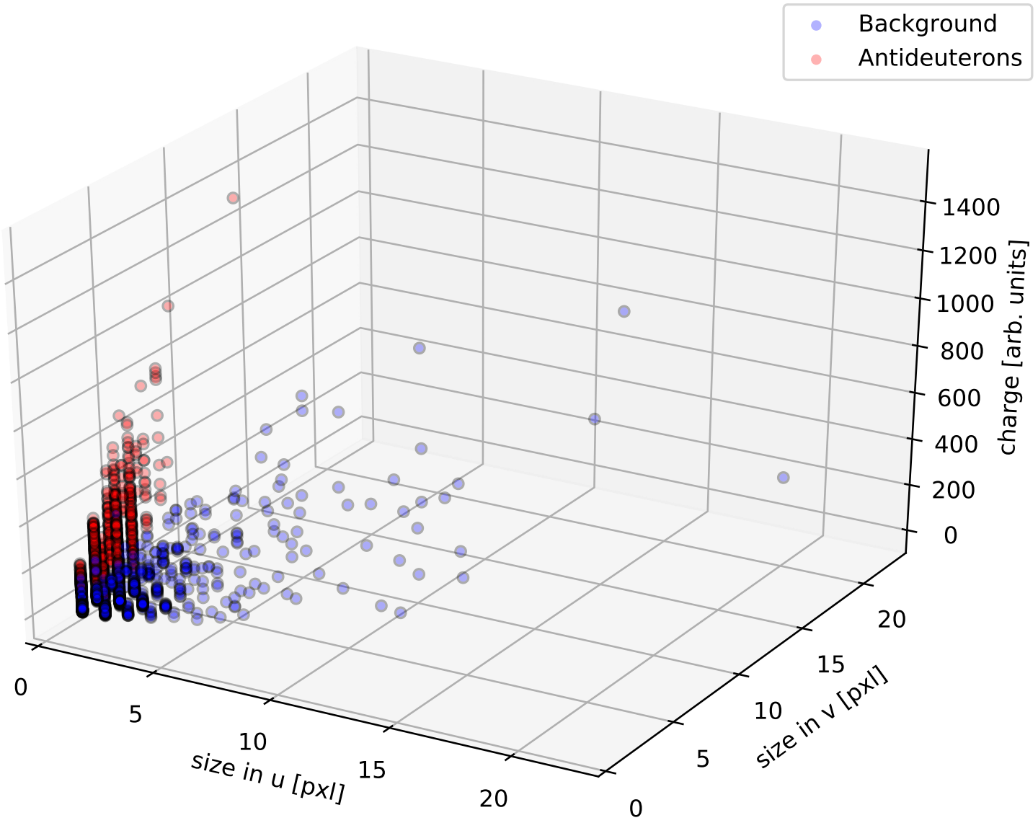
Antideuteron Dataset

Goal

Differenciate between….

$\bar{d}$

Back-ground

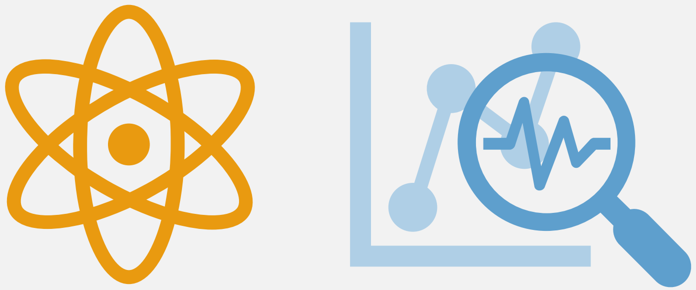


S. Käs, Multiparameter Analysis of the Belle II Pixeldetector's Data.

Problem Areas in Particle Physics

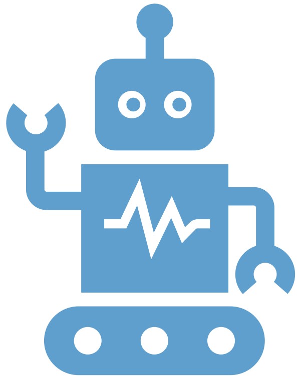
which can be tackled using Stats & ML

Main Challenges in Particle Physics



Complexity of modern particle physics

Enormous data volumes



Does AI help?

Main Challenges in Particle Physics

Complexity of modern particle physics

Simulations

Experiments

Simulations and AI in Particle Physics

Simulations

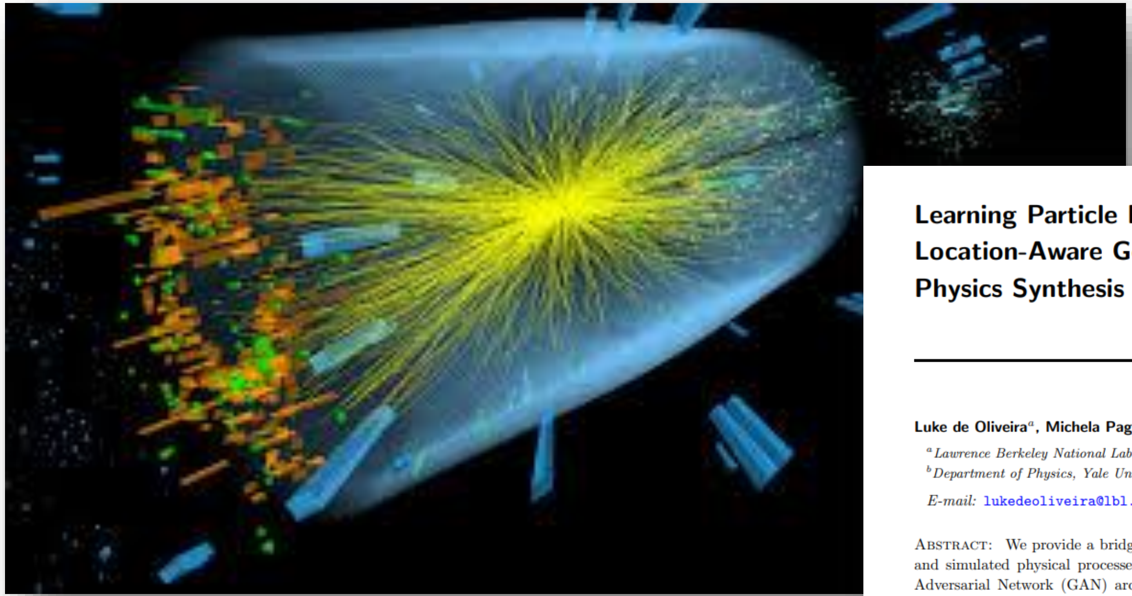
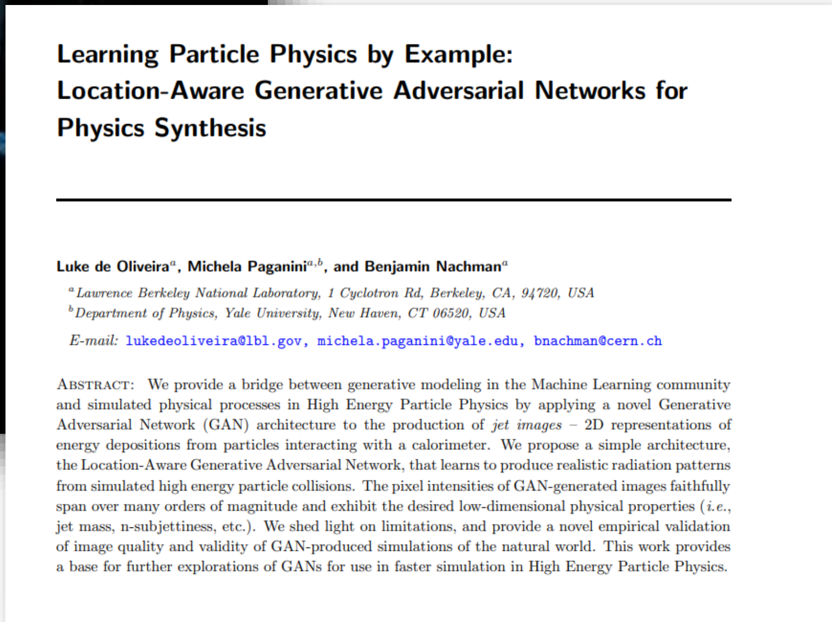


Image: Encrypted-tbn0.gstatic.com



Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis

Luke de Oliveira[a], **Michela Paganini**[a,b], **and Benjamin Nachman**[a]

[a] *Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA, 94720, USA*
[b] *Department of Physics, Yale University, New Haven, CT 06520, USA*

E-mail: lukedeoliveira@lbl.gov, michela.paganini@yale.edu, bnachman@cern.ch

Abstract: We provide a bridge between generative modeling in the Machine Learning community and simulated physical processes in High Energy Particle Physics by applying a novel Generative Adversarial Network (GAN) architecture to the production of *jet images* – 2D representations of energy depositions from particles interacting with a calorimeter. We propose a simple architecture, the Location-Aware Generative Adversarial Network, that learns to produce realistic radiation patterns from simulated high energy particle collisions. The pixel intensities of GAN-generated images faithfully span over many orders of magnitude and exhibit the desired low-dimensional physical properties (*i.e.*, jet mass, n-subjettiness, etc.). We shed light on limitations, and provide a novel empirical validation of image quality and validity of GAN-produced simulations of the natural world. This work provides a base for further explorations of GANs for use in faster simulation in High Energy Particle Physics.

https://arxiv.org/pdf/1701.05927

Main Challenges in Particle Physics

Handling
big data

Find patterns humans
cannot observe

New physics!

Main Challenges in Particle Physics

a) Simulations

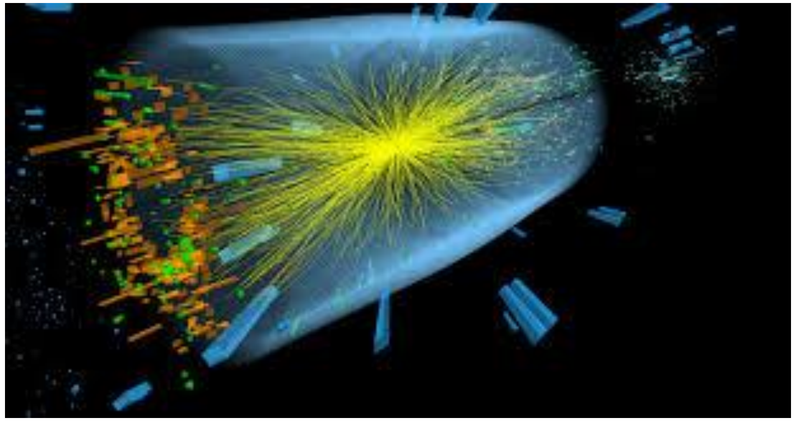
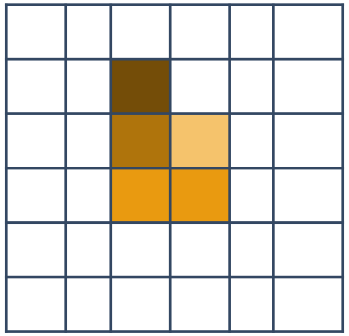


Image: Encrypted-tbn0.gstatic.com



Classification of Particles

Image: encrypted-tbn0.gstatic.com

b) Cluster Analysis

c) Pattern Recognition

d) Event Classification

Model Building

How do we choose a (good) model for our specific problem?

What is a model?

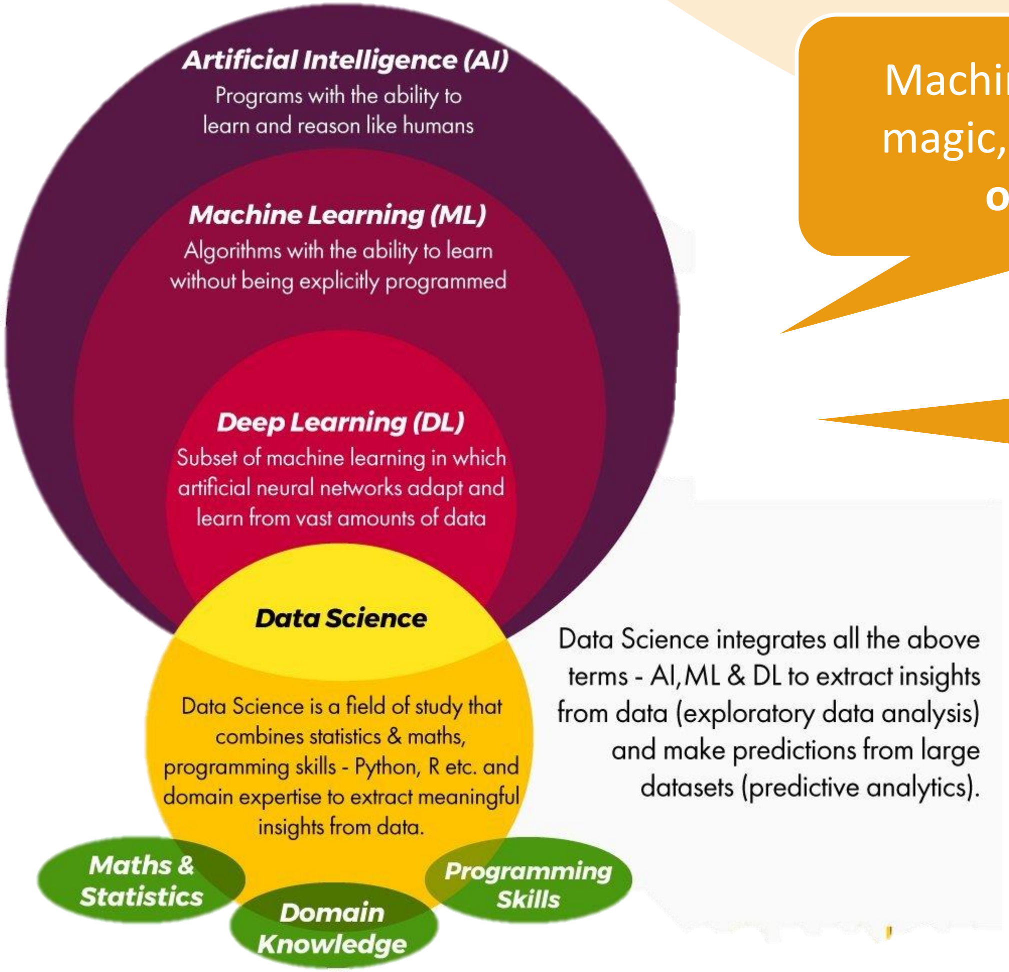
A model is **a simplified representation of a syst**em or phenomenon that helps to **understand, analyze, predict, or control** its behavior.

Key Components:

- **Variables**: Represent the quantities of interest
- **Parameters**: Constants that define system behavior
- **Equations/Rules**: Mathematical expressions/rules governing the relationships between variables.

$$s(t) = v * t + s_0$$

Models can be „traditional“ or „intelligent“



Machine Learning is not magic, but **statistics** and **optimization**.

Altough Deep learning is super popular, it might **not always be the best choice**.

[Original image: corpnce.com]

How do you create an AI?

Define Task

Which problem to you wish to solve?

Collect Data

Which data includes relevant information?

Pick Model

Which algorithm is well-suited?

Prepare Data

Train AI Model

AI learns from data.

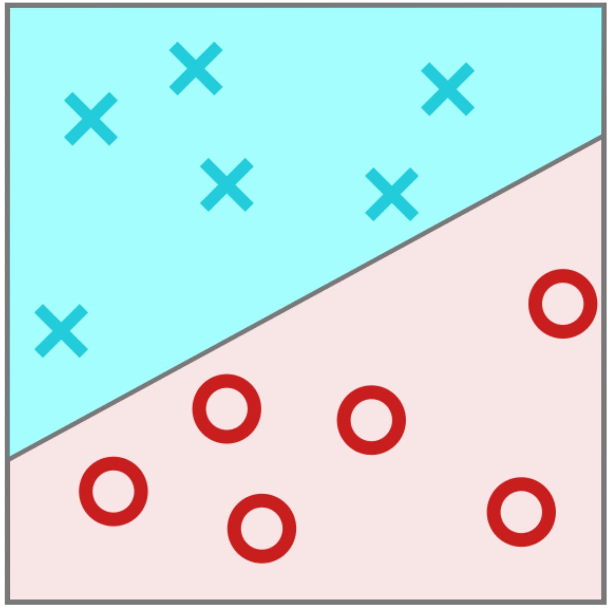
Test Model

Is the model able to perform well on unseen data?

Apply AI Model

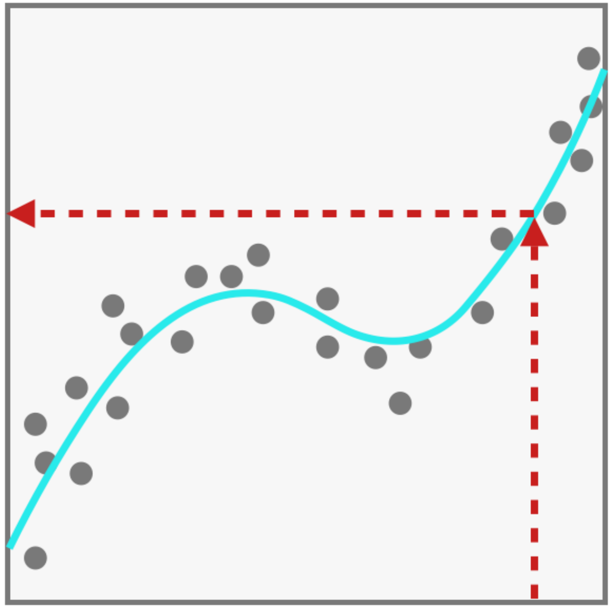
Classification vs. Regression

Classification Groups observations into "classes"



Here, the line classifies the observations into X's and O's

Regression predicts a numeric value



Here, the fitted line provides a predicted output, if we give it an input

Make sure to define exactly what you want to do!

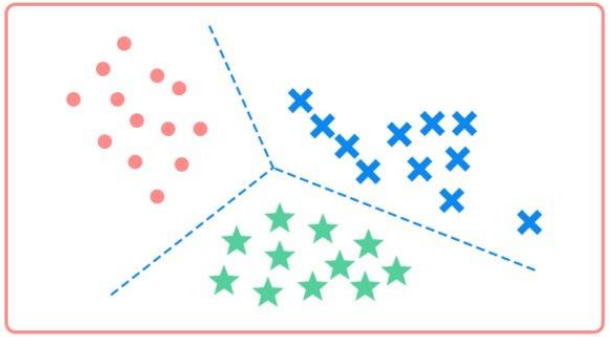
Image: r-craft.org

How does an AI learn?

Thee General Learning Methods

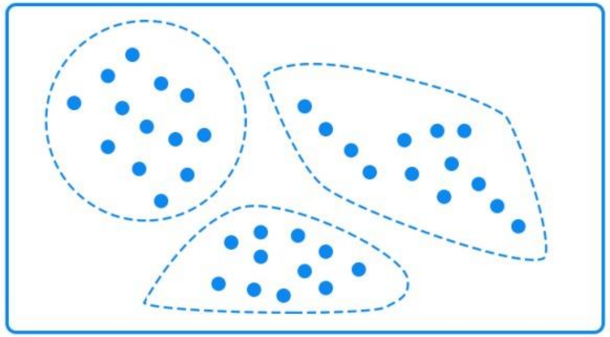
Supervised Learning	Unsupervised Learning	Reinforcement Learning

Classification

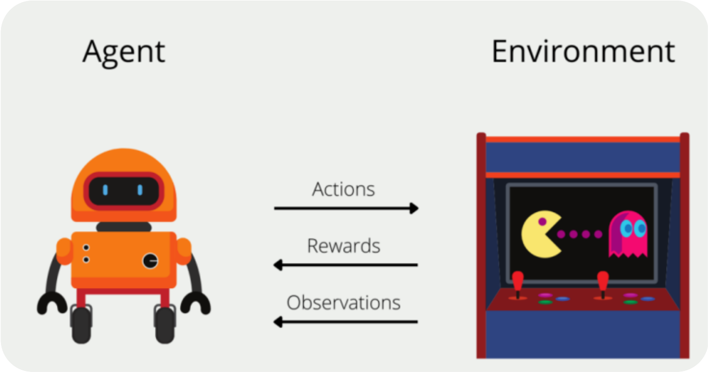


Supervised learning

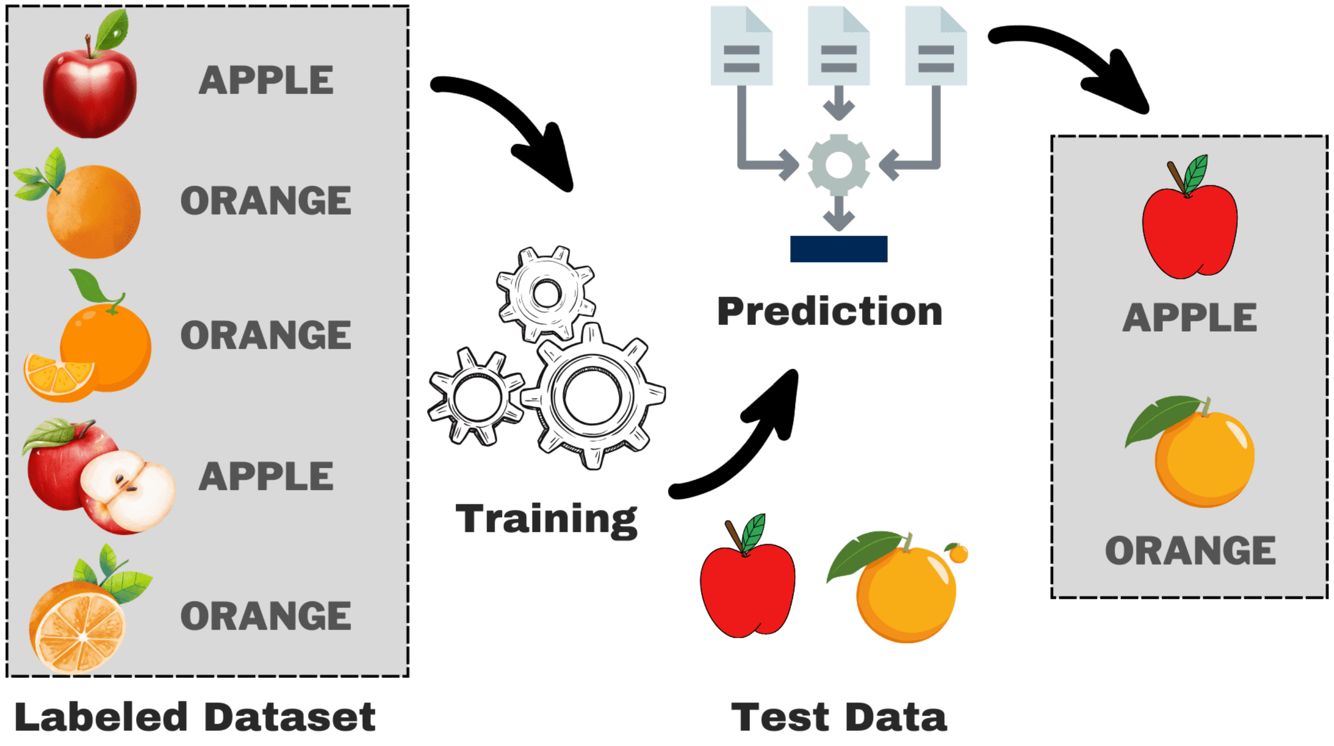
Clustering



Unsupervised learning

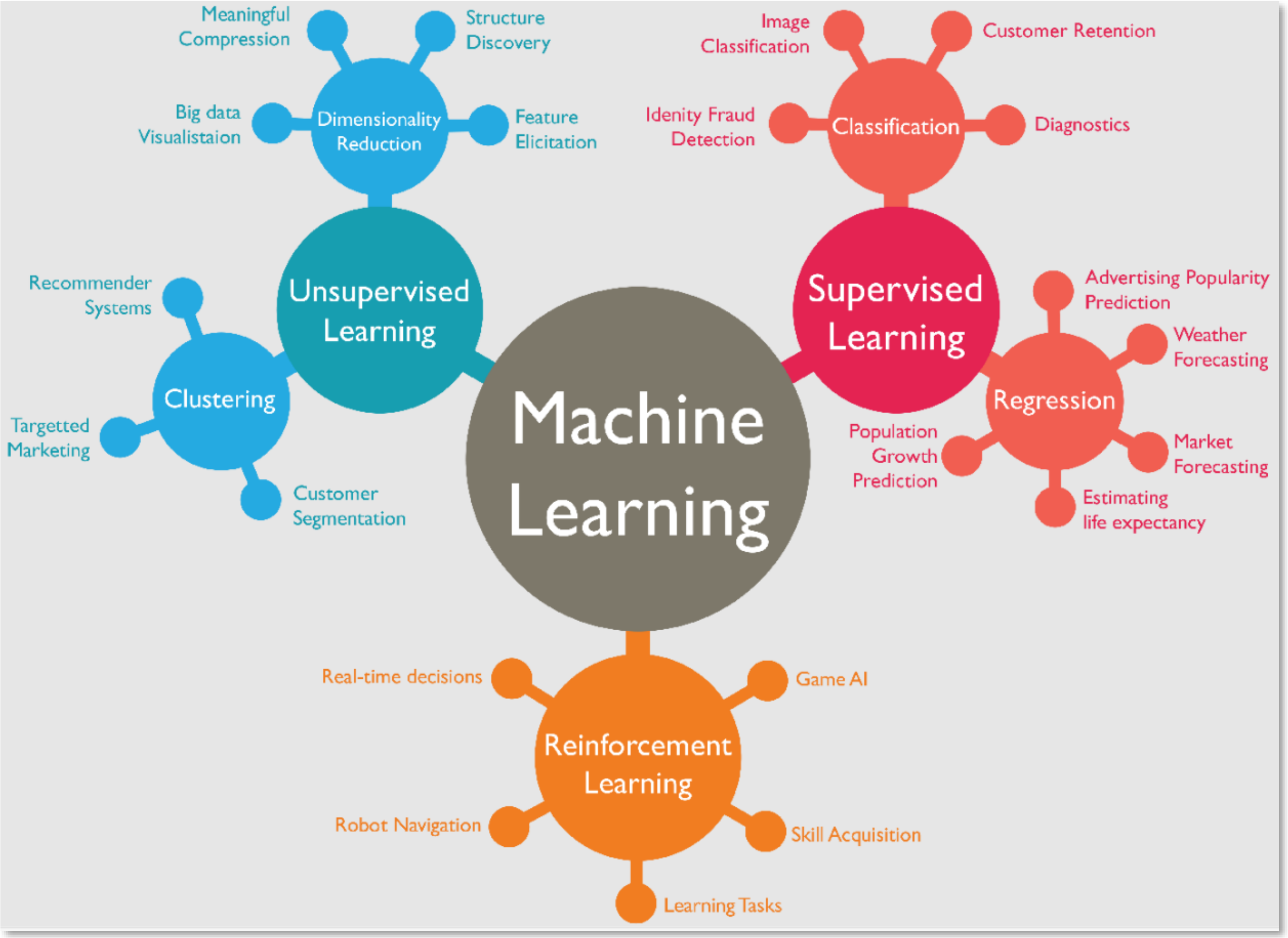


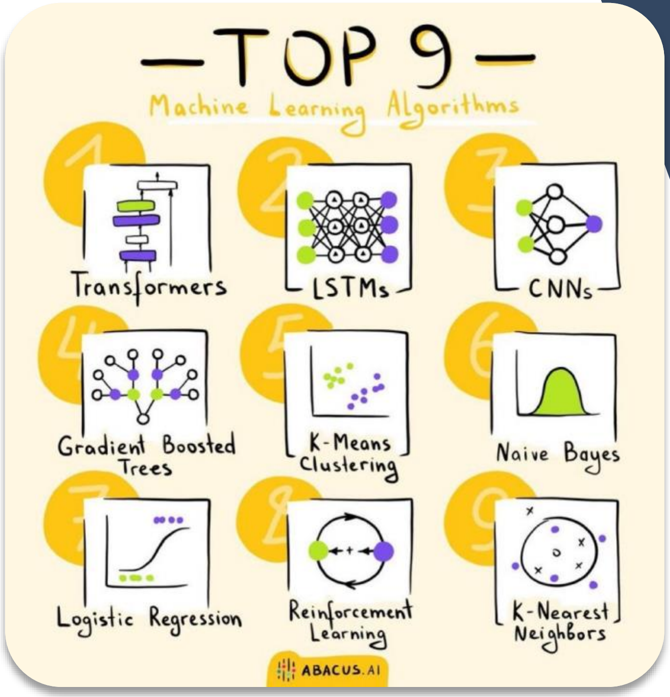
How does an AI learn?



Beispiel für überwachtes Lernen

Machine Learning is a Broad Field

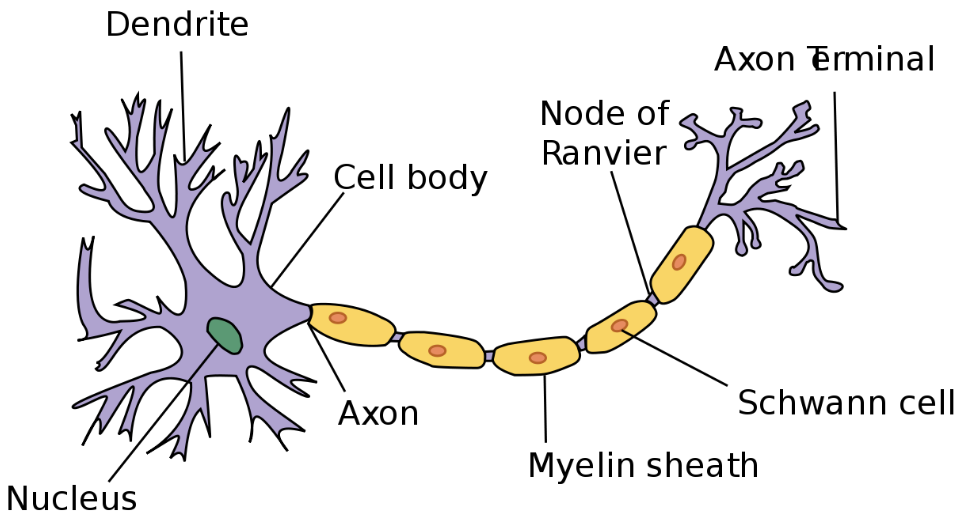




Neural Network Basics

Neural Networks – Main Idea

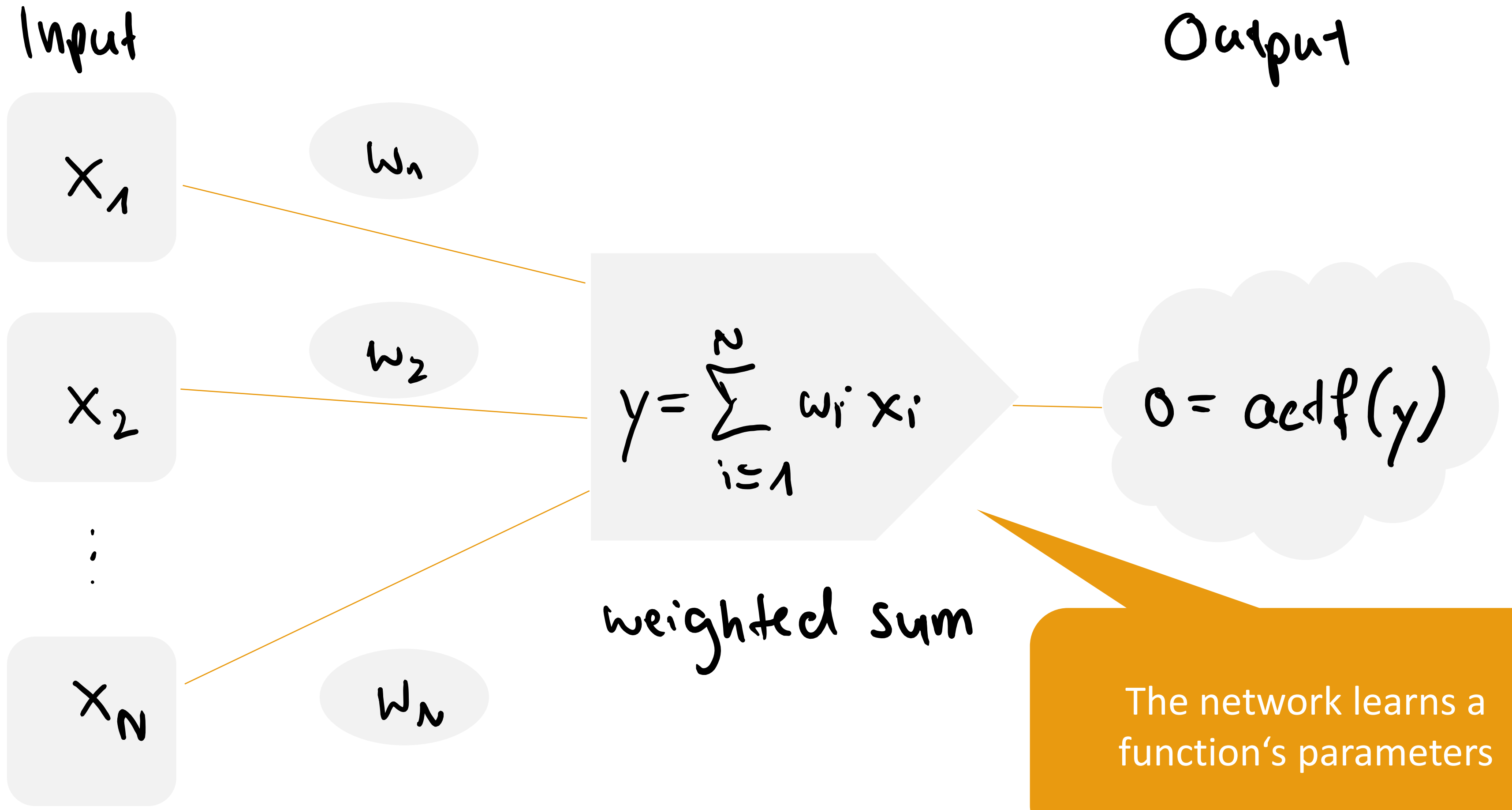
Artificial human brain -> *Neurons* which are connected



Wikimedia.org

Purdue.edu

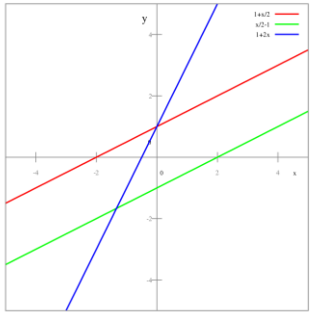
Neural Networks – Single Neuron (Perceptron)



The network learns a function's parameters

Neural Networks & Mathematical Models

A mathematical model has **parameters** which can be adapted:



Wikipedia.org

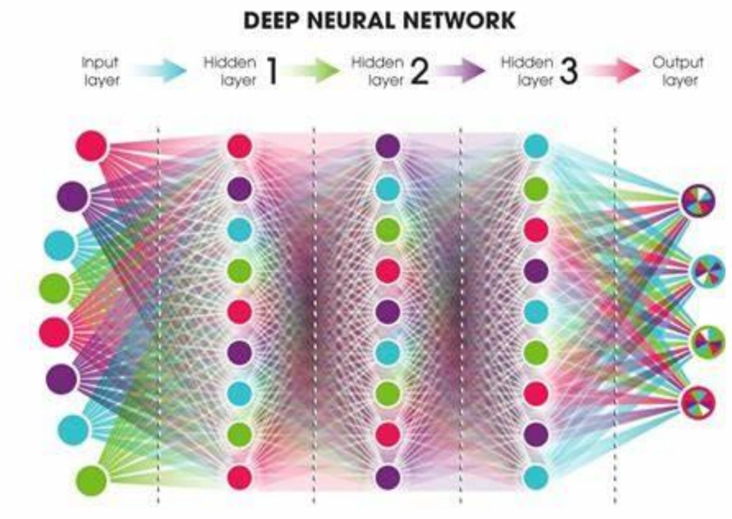
Linear model: $y = mx + b$	Parameters: m, b

In NNs **weights** are parameters...

...but there are also **hyperparameters**
which have to be chosen correctly to receive good results:

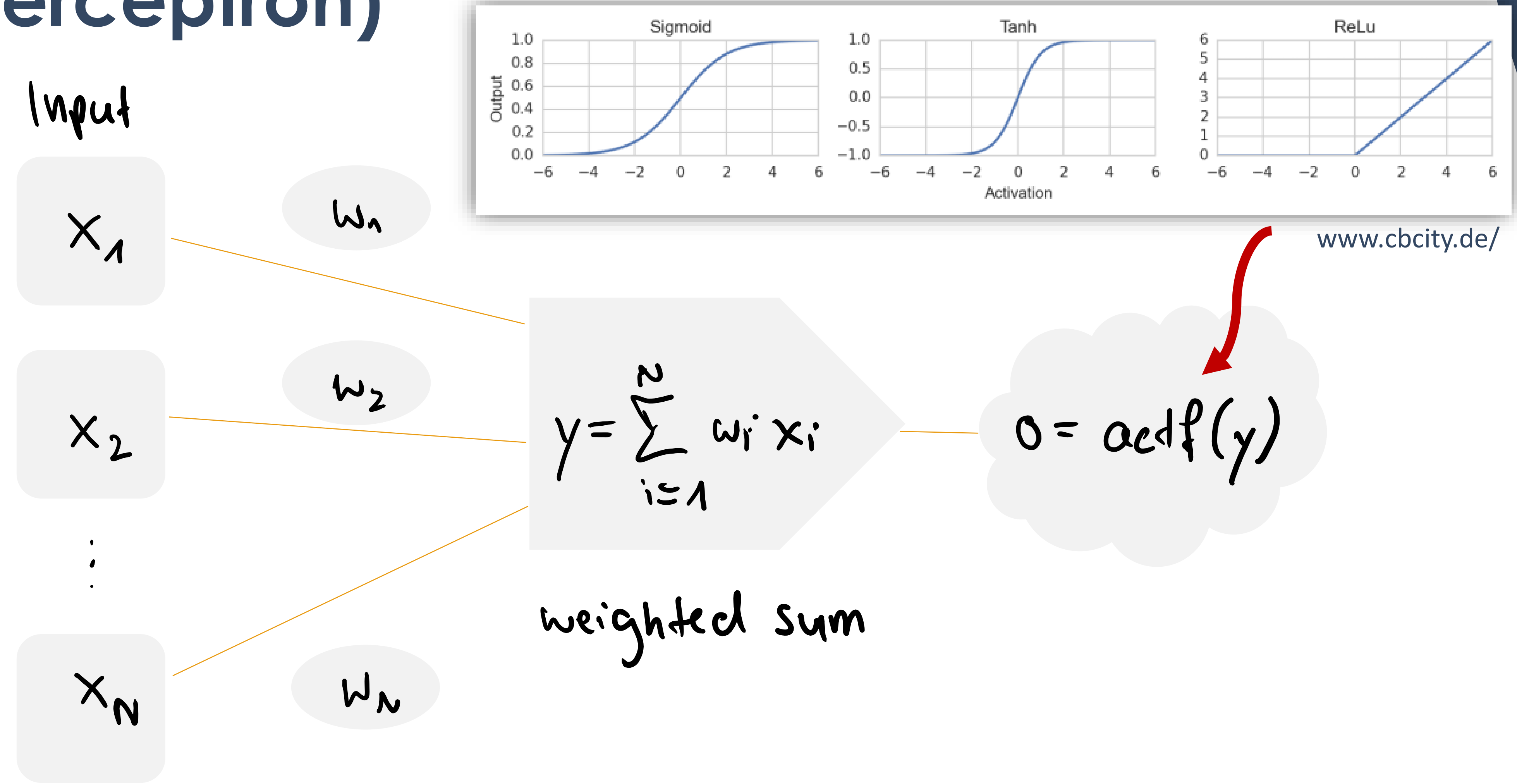
- Number of layers
- Number of neurons per layer
- Activation function
- Optimization algorithm
- ...

Millions of parameters!



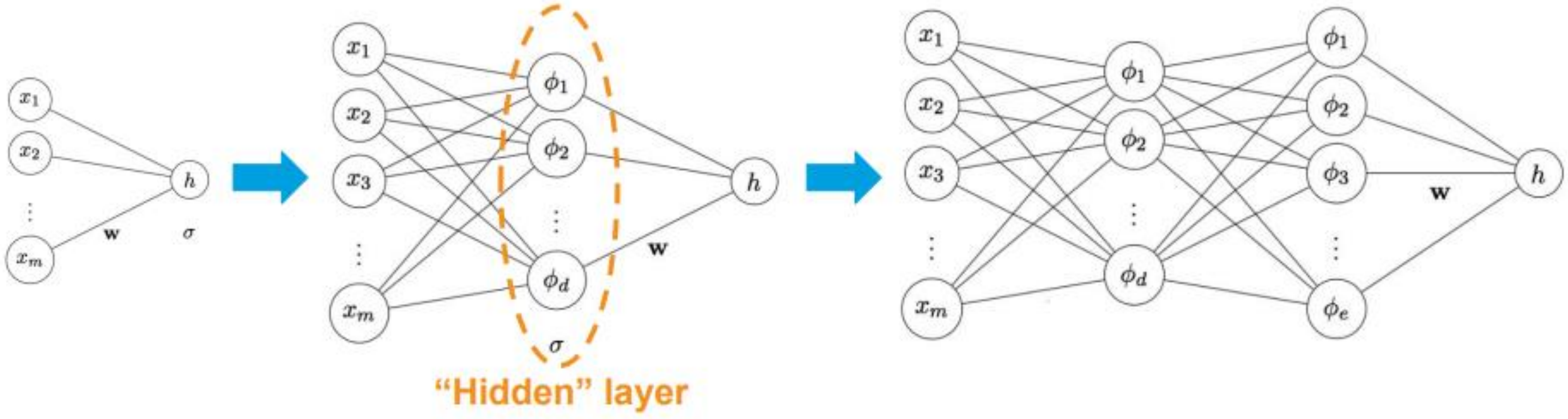
neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016

Neural Networks – Single Neuron (Perceptron)



www.cbcity.de/

Neural Networks – Multi-Layer-Perceptron (MLP)

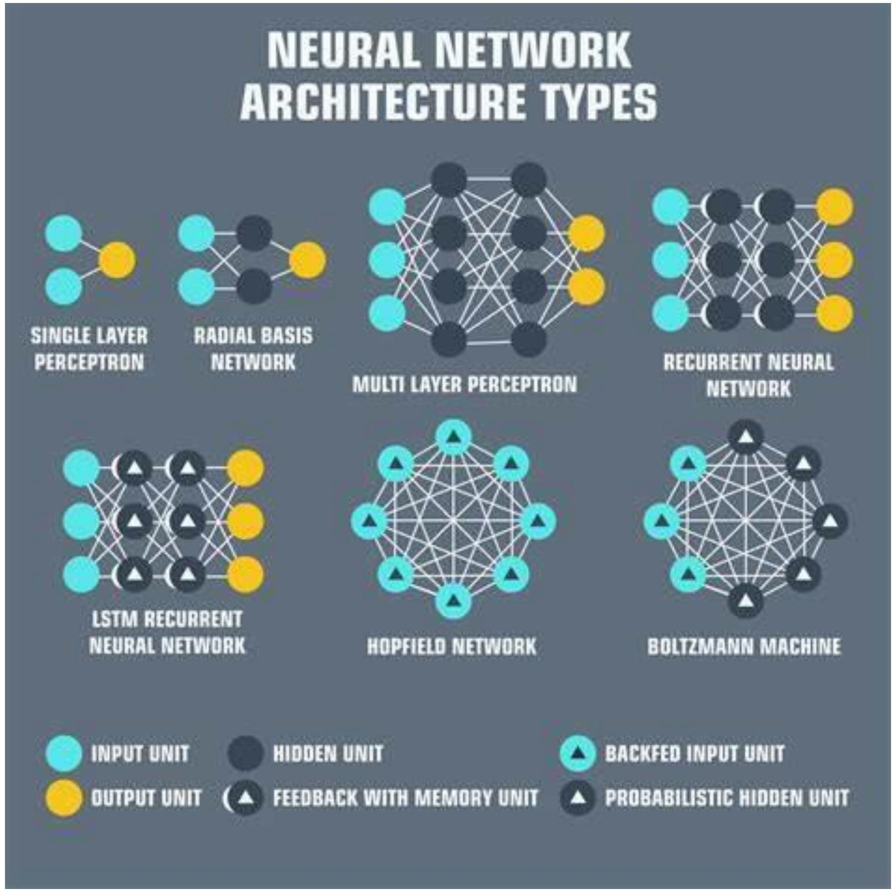


- Stack several nerons -> enlarge model‘s complexity
- Each layer adapts basis functions based on previous layer
- Allows non-linear decision boundaries

Image: Federico Meloni, HASCO Summer School 2023

Interesting Networks for Particle Physics

Data Generation:	GAN
Compressing Data & Anomaly Detection:	Autoencoder
Image Processing:	CNN
Classification:	FFN, SOM
Handling graph-like structures:	GraphNN



mygreatlearning.com

Model Evaluation

How can I check if my model did a good job?

Model Building Process: How AI Learns

Choosing appropriate Metrics

Classification

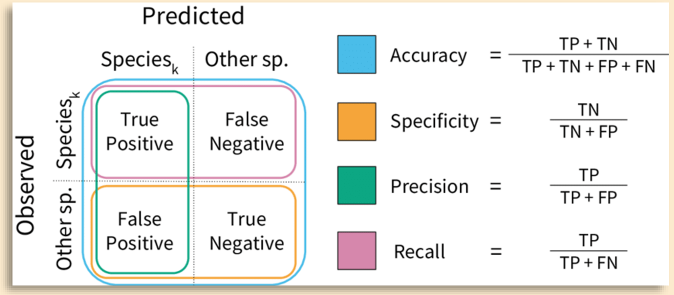


Image: Almabetter.com

Regression

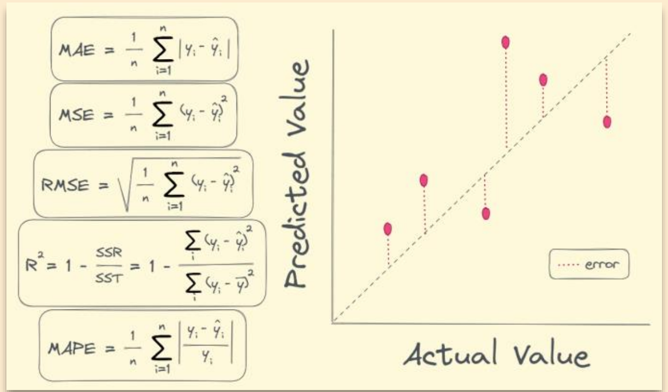


Image: Towardsdatascience

Others

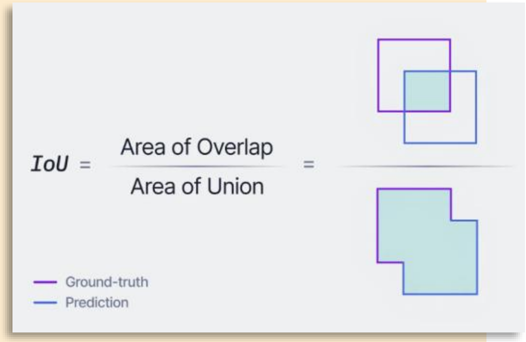
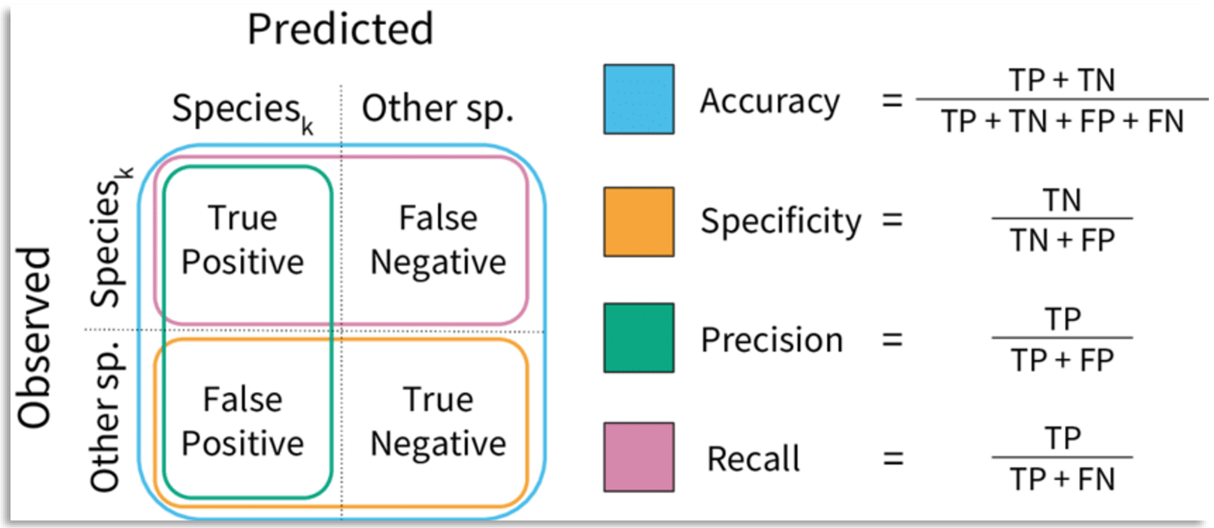


Image: v7labs.com

(Particle) Classification Metrics

Confusion Matrix & ROC-AUC Curve



Check your usecase!

F1 Score:

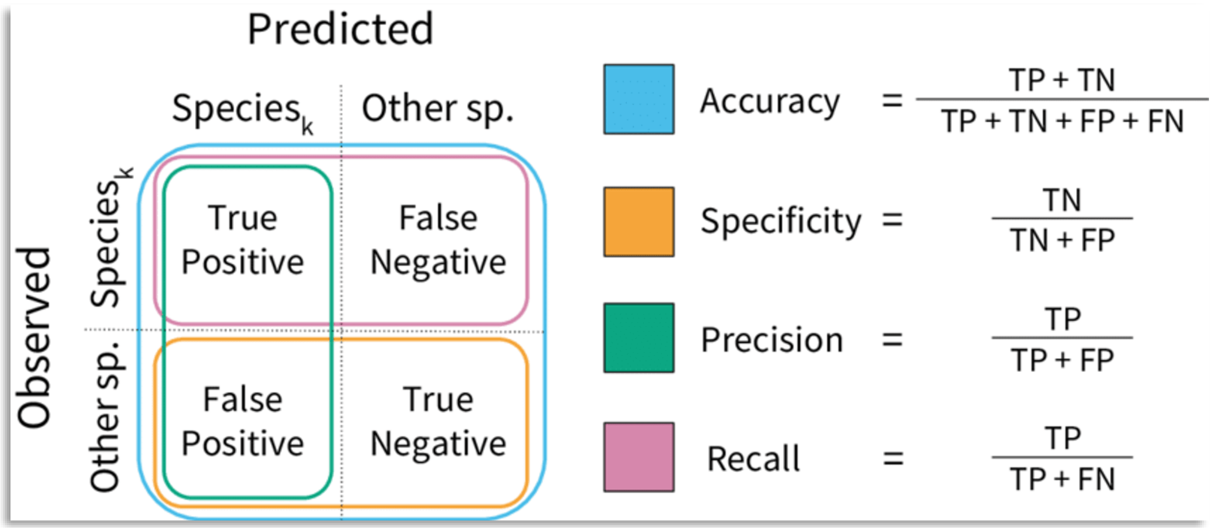
- The harmonic mean of Precision and Recall, balancing both concerns.
- Formula: $\mathrm{F1\ Score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$
- Useful when you need a balance between Precision and Recall.

Sometimes you want to **detect all potential candidates**…
(maximize *Recall*)

Image: almabetter

(Particle) Classification Metrics

Confusion Matrix & ROC-AUC Curve



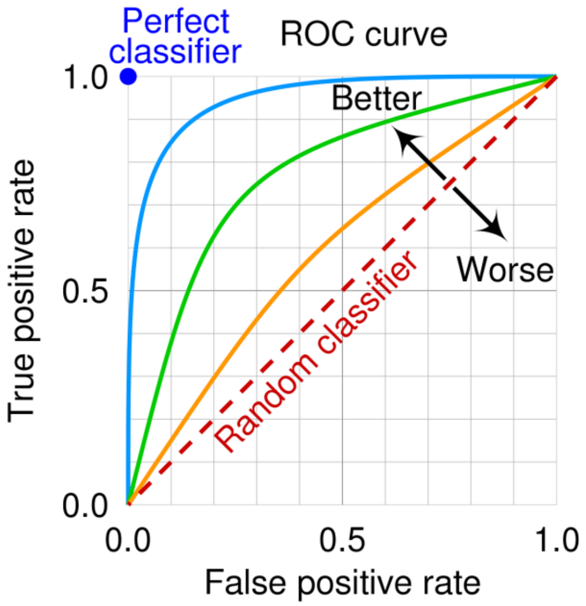


Image: almabetter

Image: Medium

Model Training & Evaluation

Train, Test, and Validation Dataset Split

Ensure **robust evaluation** of a machine learning model's performance without **over- or underfitting**.

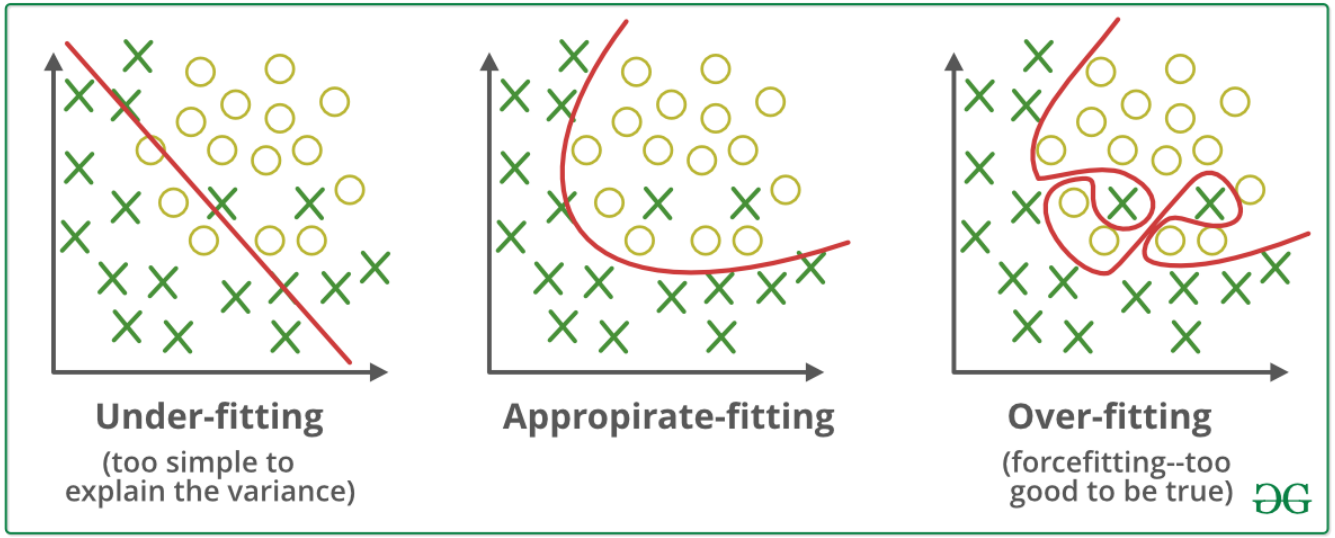


Image: Kaggle

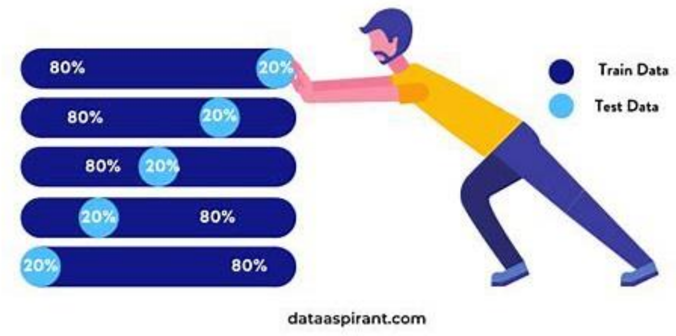
Model Training & Evaluation

Train, Test, and Validation Dataset Split

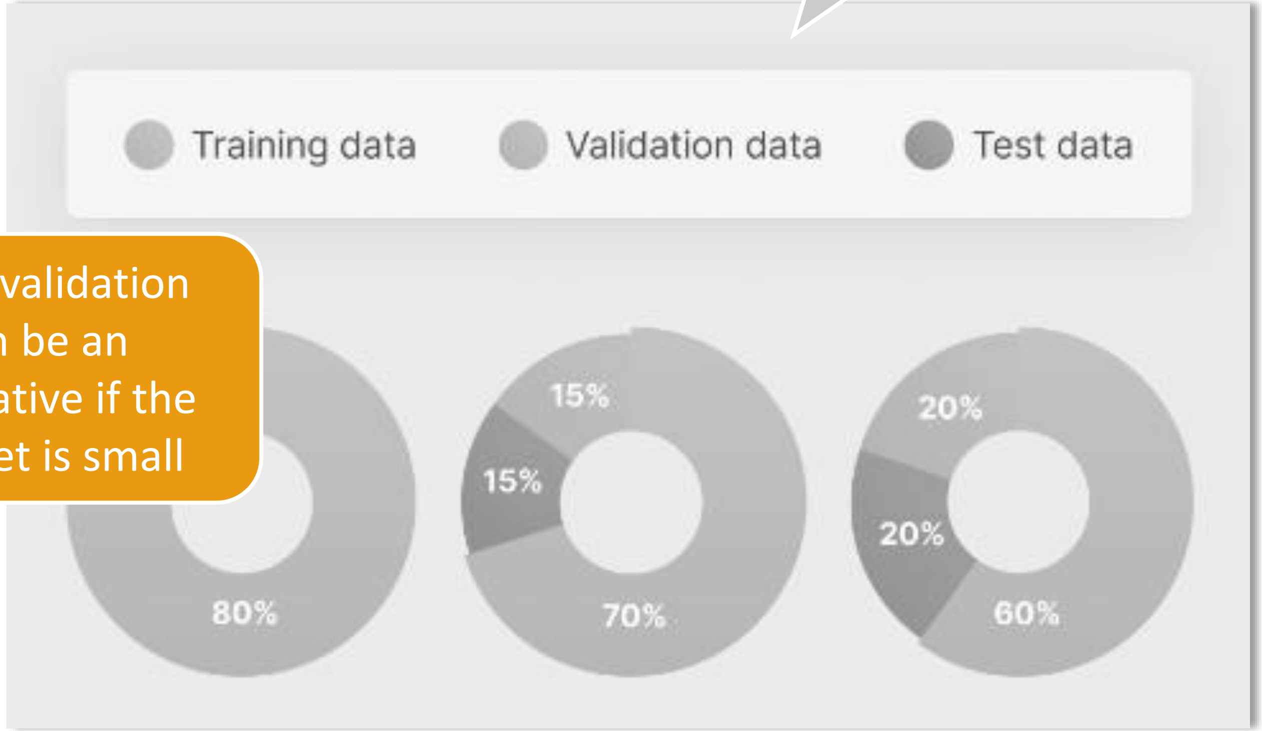
Traditionally, we use a **split**

Ensure **robust evaluation** of a machine learning model's performance without over- or underfitting.

Cross-validation can be an alternative if the dataset is small



[Image: dataaspirant.com](https://dataaspirant.com)



[Image: v7labs.com](https://v7labs.com)

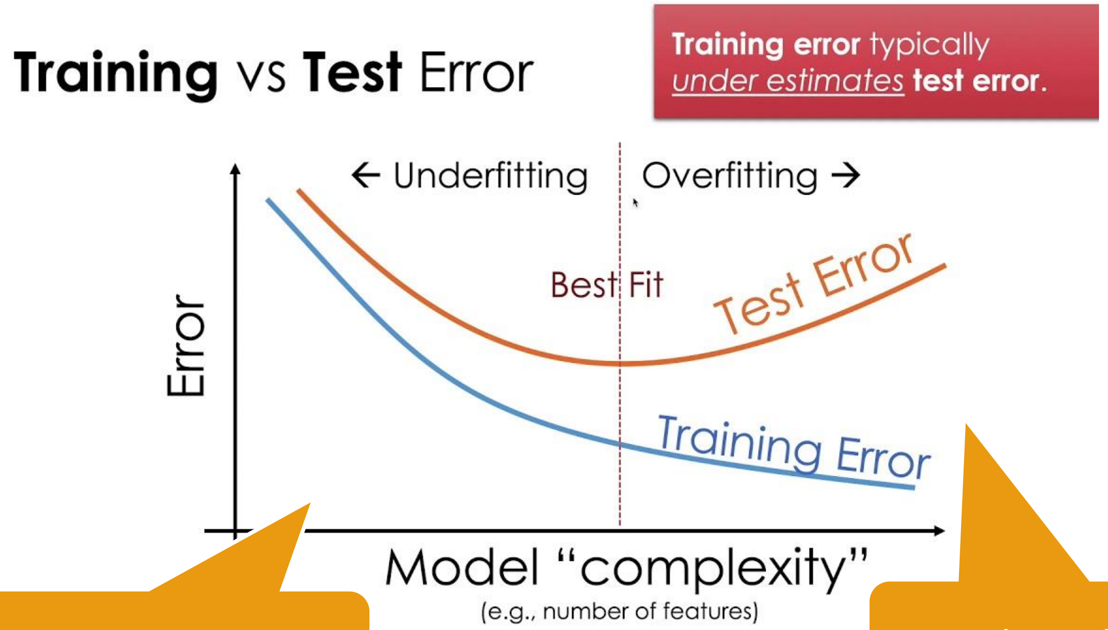
Model Training & Evaluation

Train, Test, and Validation Dataset Split

Bias-Variance Tradeoff

Increased model complexity

- more parameters to fit to
- requires mode data



High bias, low variance

Low bias, high variance

Image: YouTube | J. E. Gonsalez

Back to Particle Physics

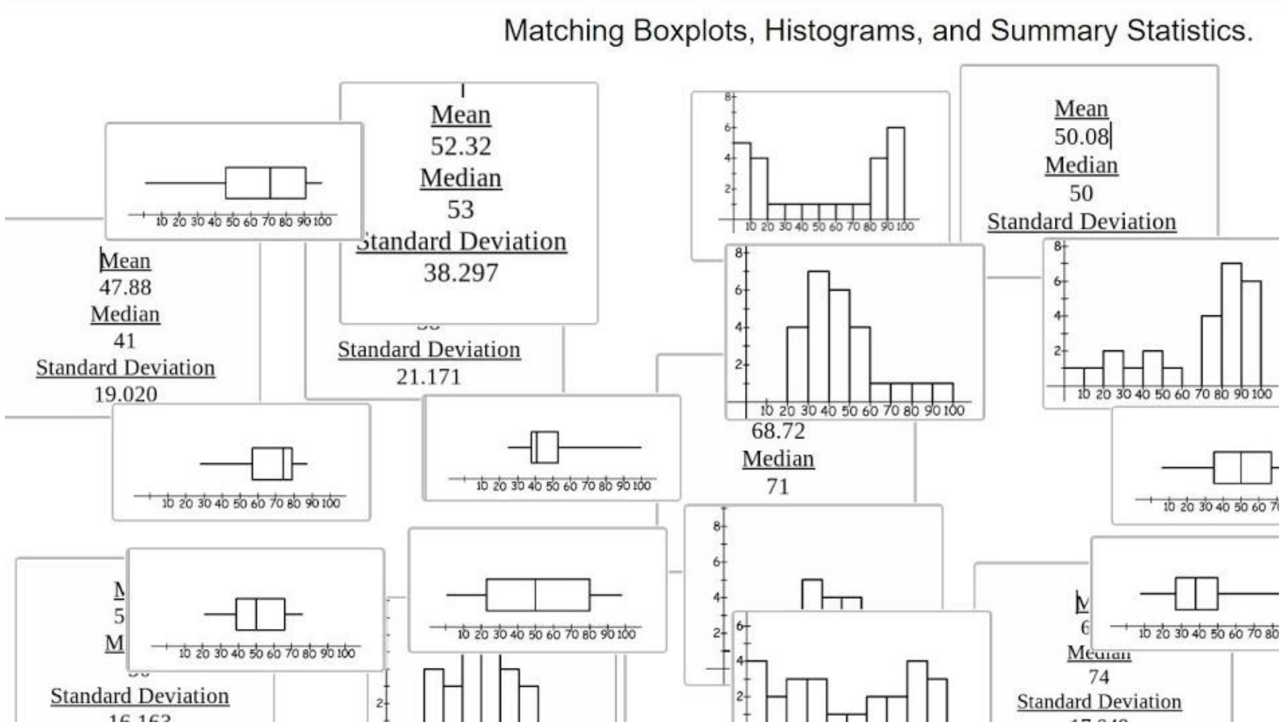
Data Preprocessing

Data Preprocessing

Transforming **raw data** into a clean and usable format for machine learning models.

Enhances data quality, improves model accuracy, and reduces computational costs.

Basic Statistics First + Actually *Look* at Your Data



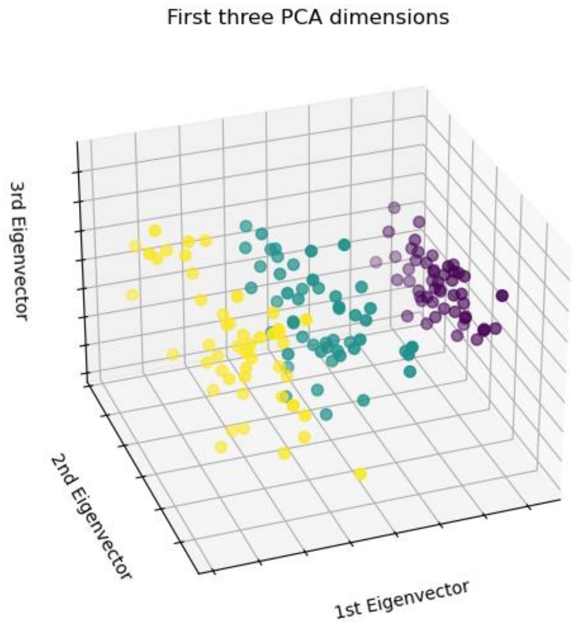
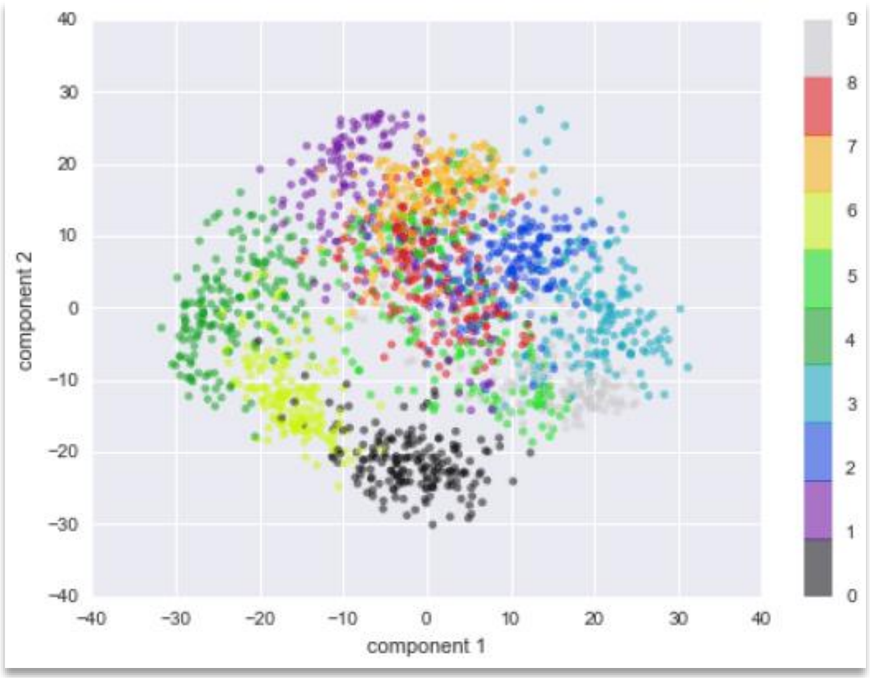


Image: YouTube | D. Reeves

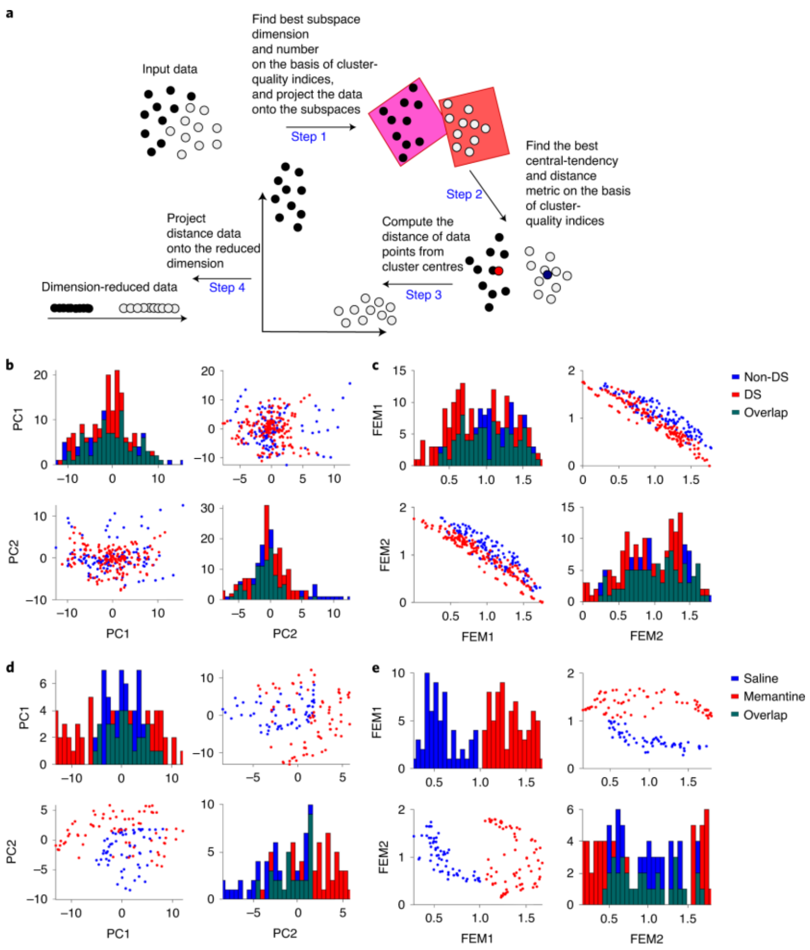
https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

Dimensionality Reduction



[Image: Neptuune.ai]

Check out this [link] for more information about **data visualization.**



[Image: Nature.com]

Dimensionality Reduction

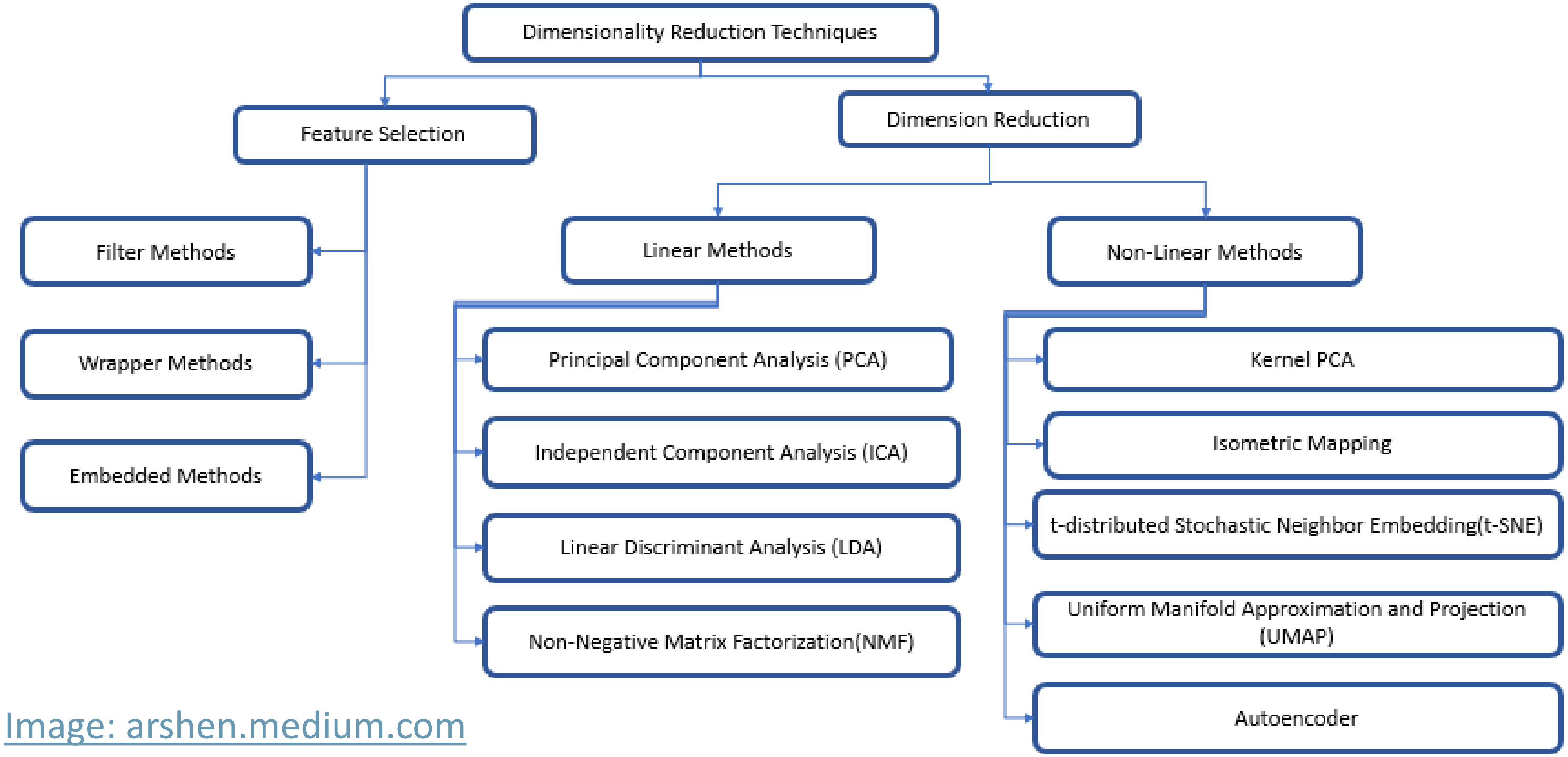
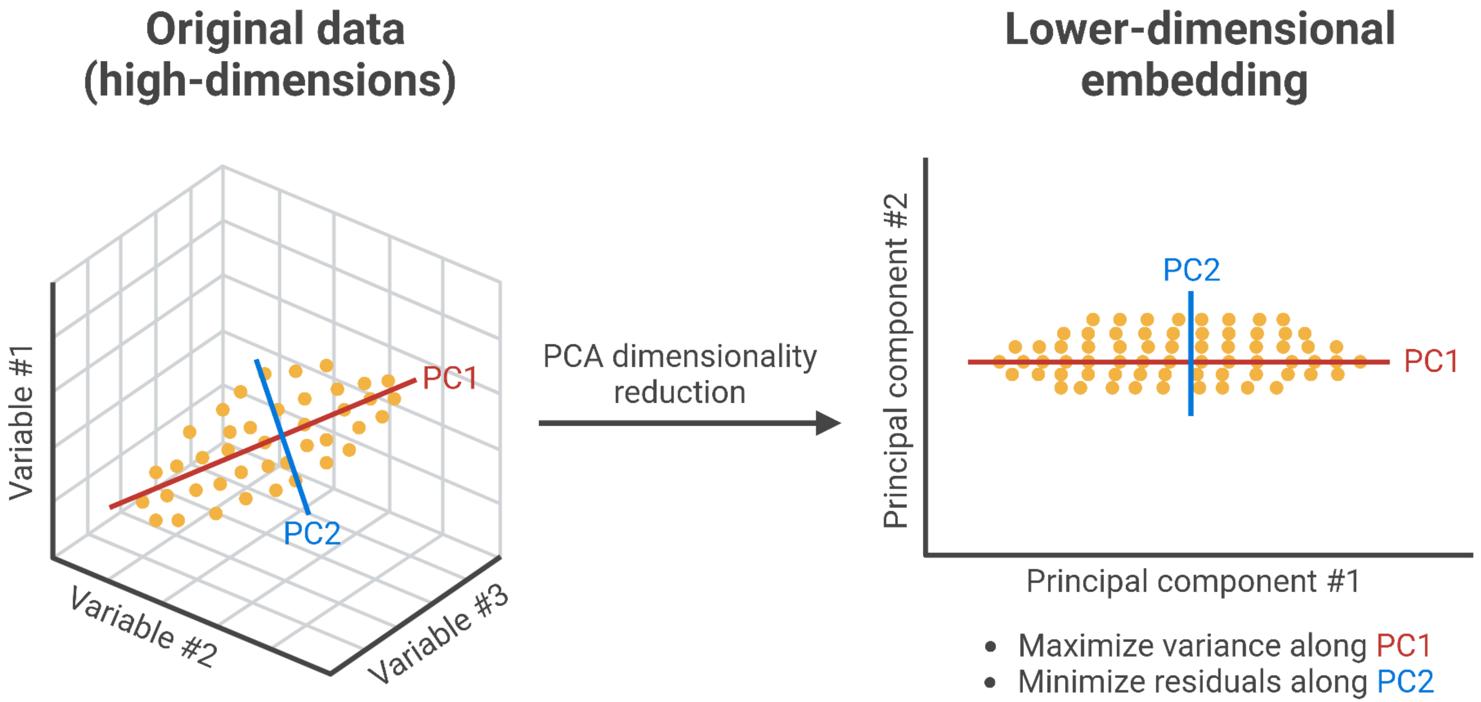


Image: arshen.medium.com

Principal Component Analysis



- Maximize variance along PC1
- Minimize residuals along PC2

Image: Biorender.com

Handling Images

Convolutional Neural Networks

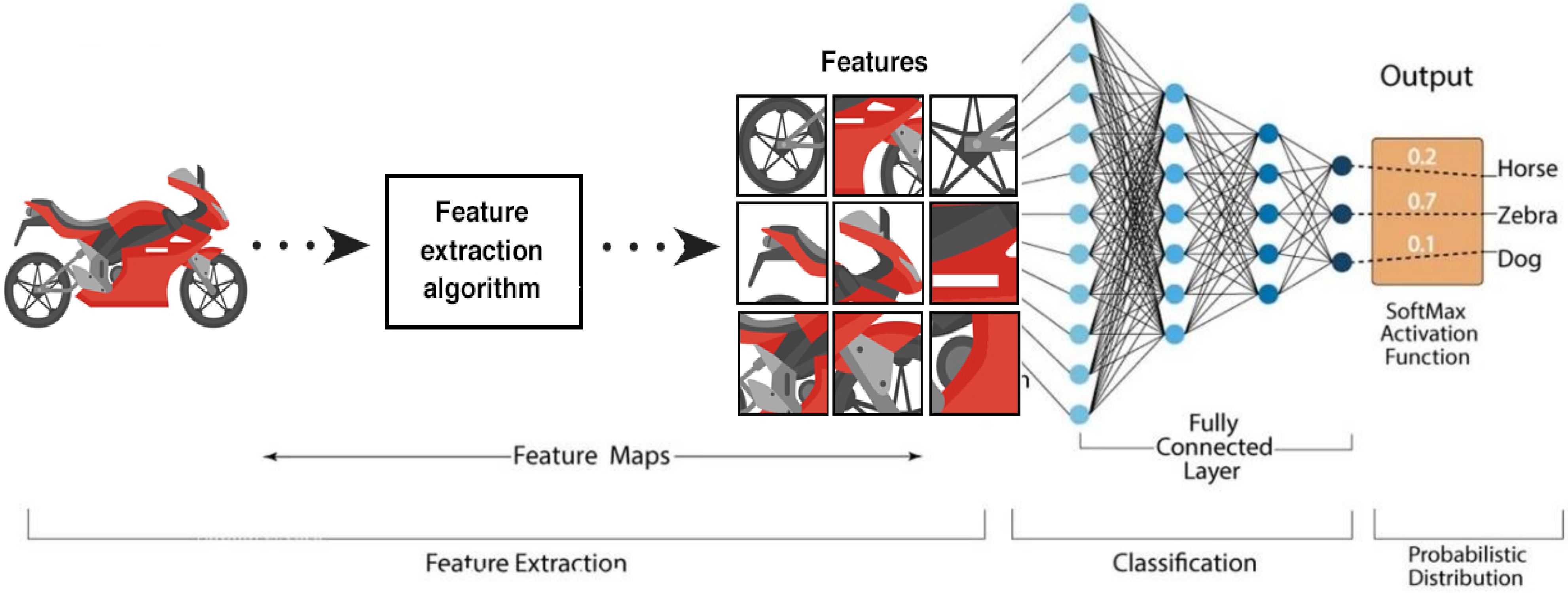
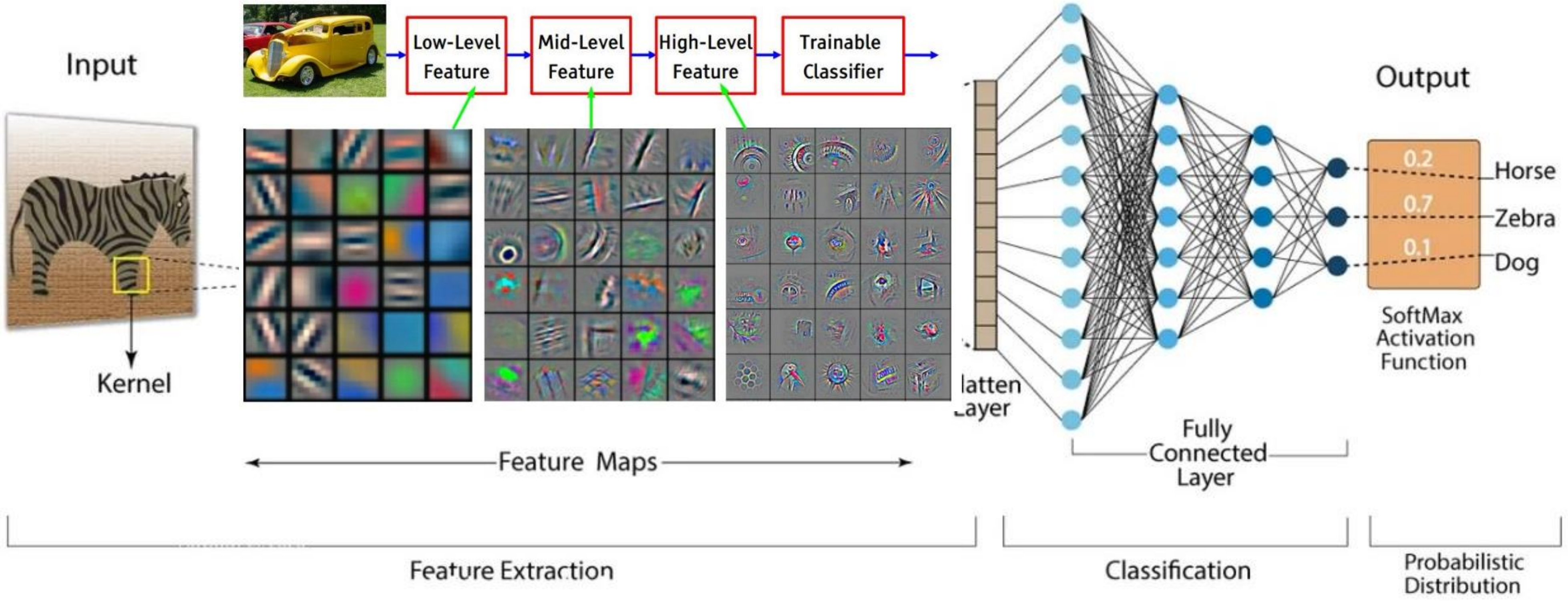


Image: towardsdatascience.com

Convolutional Neural Networks



[Image: towardsdatascience.com]()

Convolutional Neural Networks

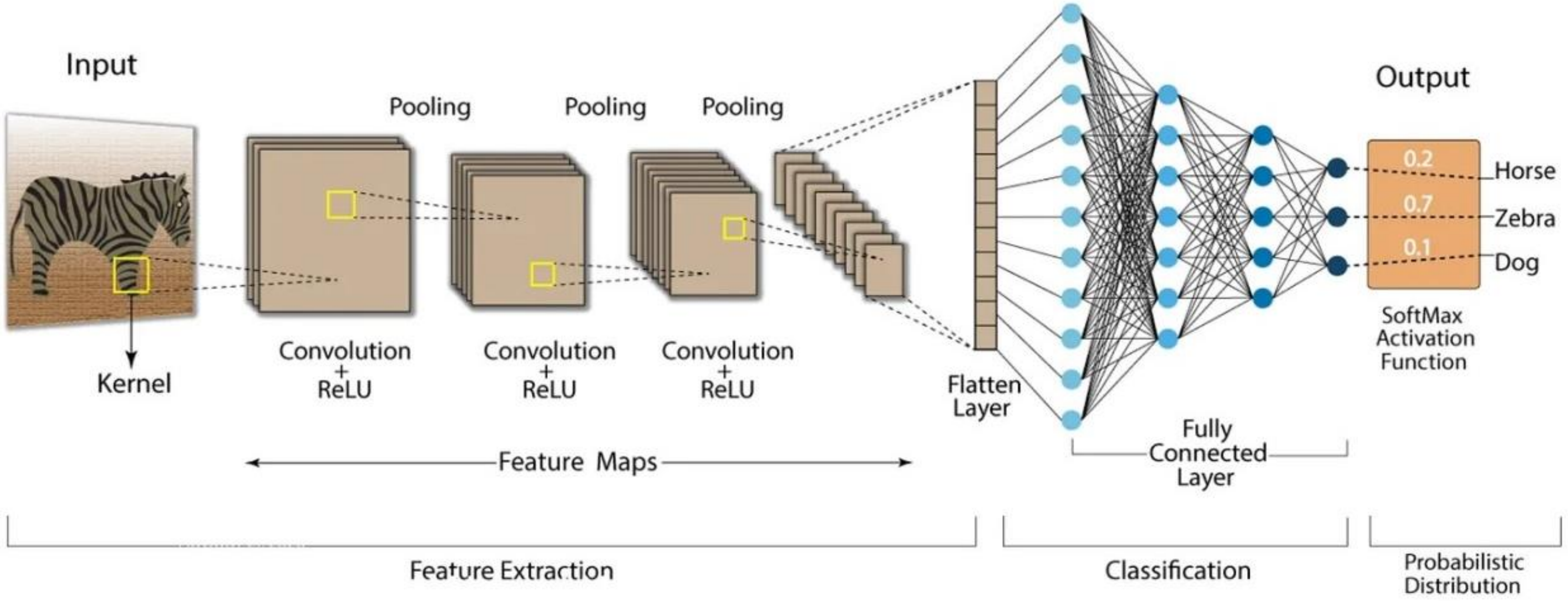


Image: towardsdatascience.com

Convolutional Neural Networks

What we see:

How its encoded

Image: Erum Data Hub Deep Learning School 2024

Colour images are encoded in RGB.

Convolutional Neural Networks

What we see:

How its encoded

5	10	2	4	2	0
3	6	2	0	1	3
0	7	10	4	8	0
11	10	8	0	3	7
10	4	8	6	0	4
5	12	11	0	2	0

5	6	2	3	2	0
5	6	2	0	1	3
0	5	12	2	5	6
2	2	8	0	3	7
0	4	5	10	0	4
4	1	4	0	2	9

6	10	2	4	2	2
3	6	12	0	1	6
0	5	2	4	5	4
11	13	8	1	3	7
9	4	4	6	0	9
5	12	5	8	2	9

Image: Erum Data Hub Deep Learning School 2024

Images are matrices.

Convolutional Neural Networks

Filter operations detect patterns

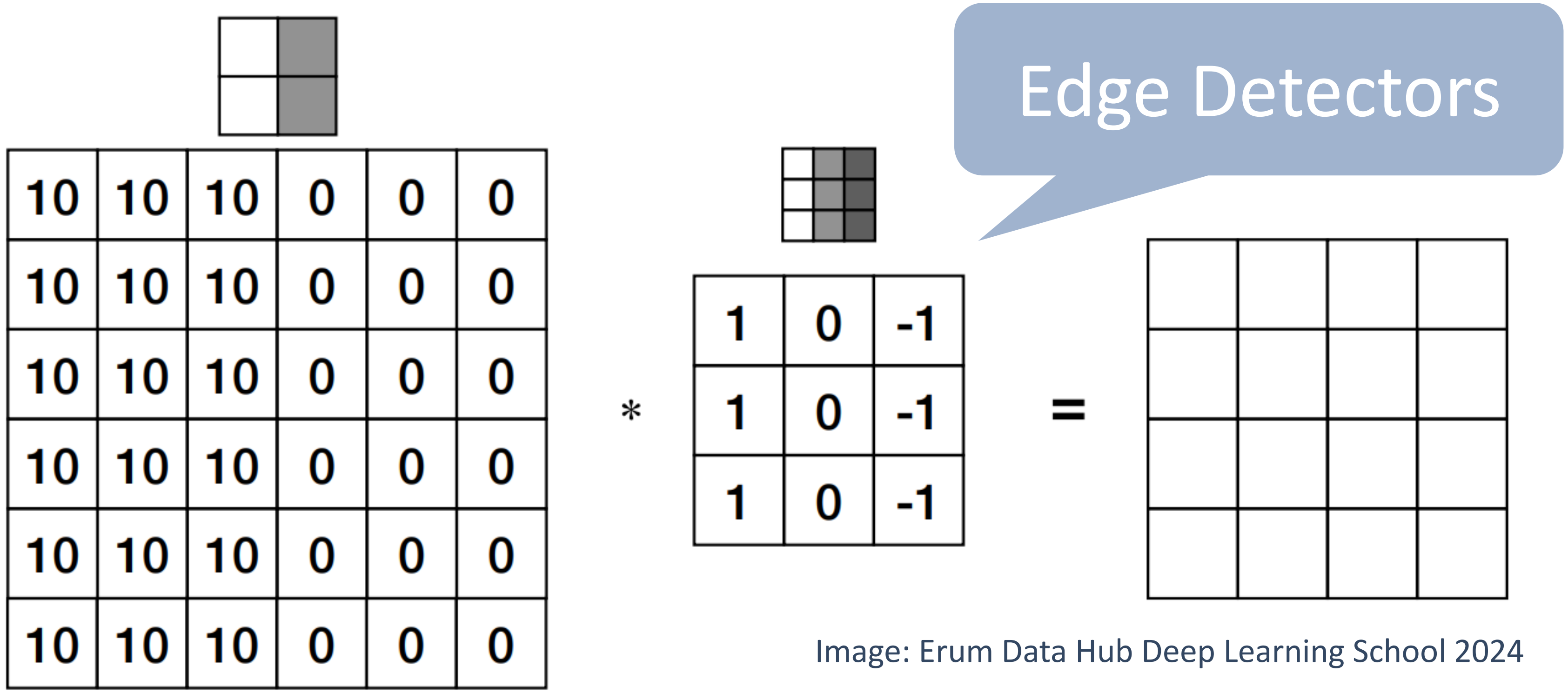


Image: Erum Data Hub Deep Learning School 2024

Convolutional Neural Networks

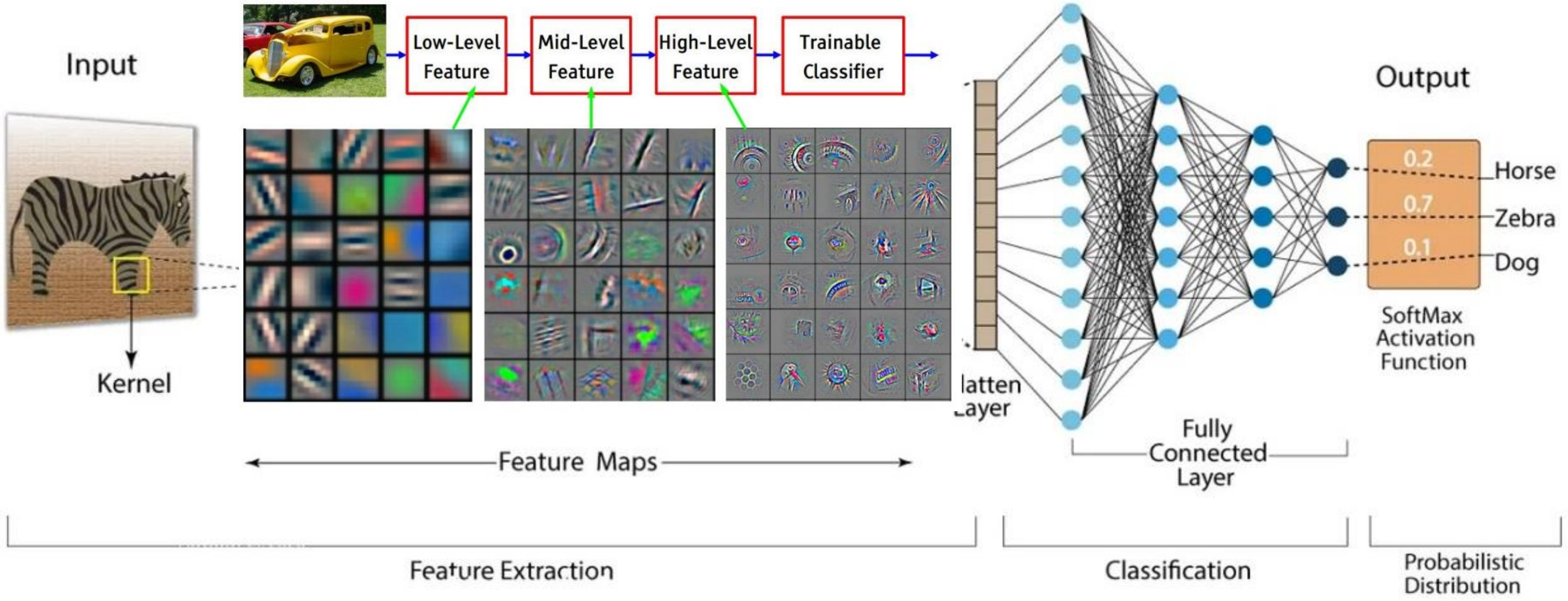


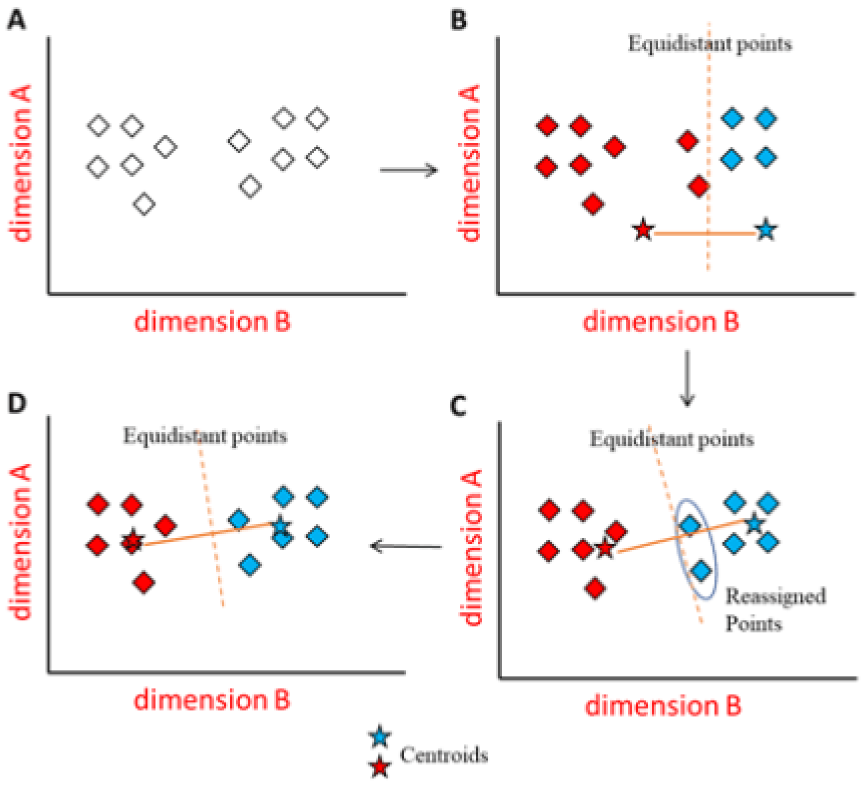
Image: Erum Data Hub Deep Learning School 2024

Cluster Analysis Methods

Which data does belong together?

Finding Clusters of Data

Identify natural groupings in data without predefined labels.



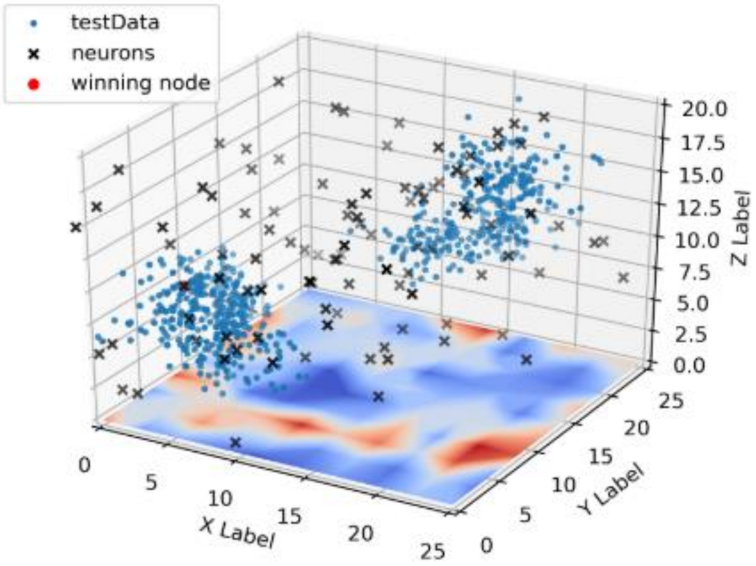
k-Means Clustering

1. Choose the number of clusters k.
2. Initialize k cluster centroids randomly.
3. Assign each data point to the nearest centroid.
4. Recompute centroids as the mean of assigned points.

Others: Hierarchical Clustering, Self-organizing Maps, …

Self-Organizing Maps

- Unsupervised learning
- Self-organizing



Low distance between nodes

High distance between nodes

J. Bilk & J. Budak, Detecting Clusters in Highdimensional Data

Figure 6: The first step of a self-organizing map. One can see the data clouds in blue, the vectors of each neuron as black x's and on the floor the U-matrix.

Classification Methods

*How can we see **anything** in a huge pile of data?*

Classification Algorithms

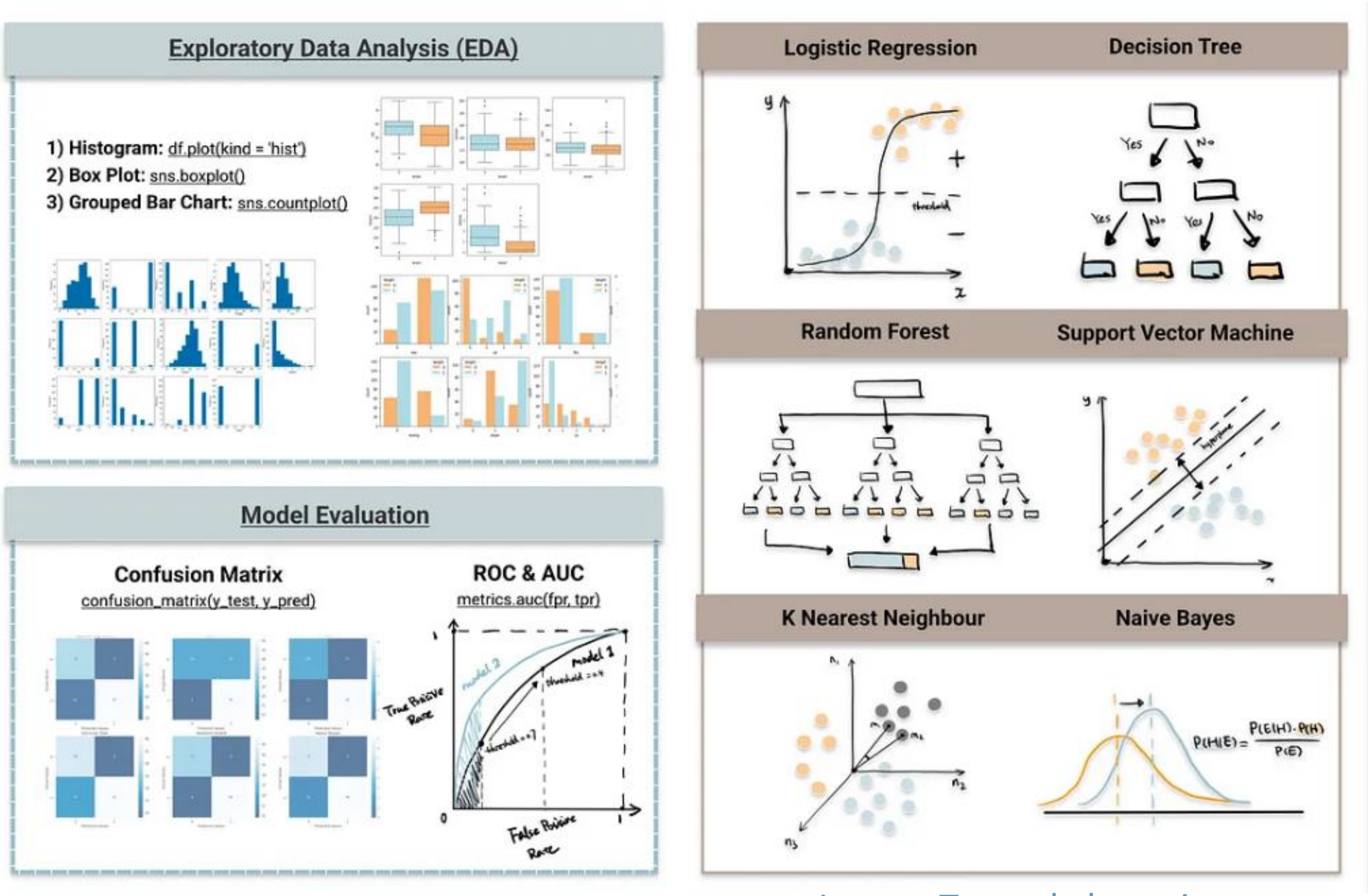


Image: Towardsdatascience.com

Anomaly Detection

How can we detect something unusual?

Autoencoder

Neural network for data **compression** and **reconstruction**.

Example: Using autoencoders for

- **noise reduction** in detector data and
- **detecting anomalies** that might indicate new physical phenomena.

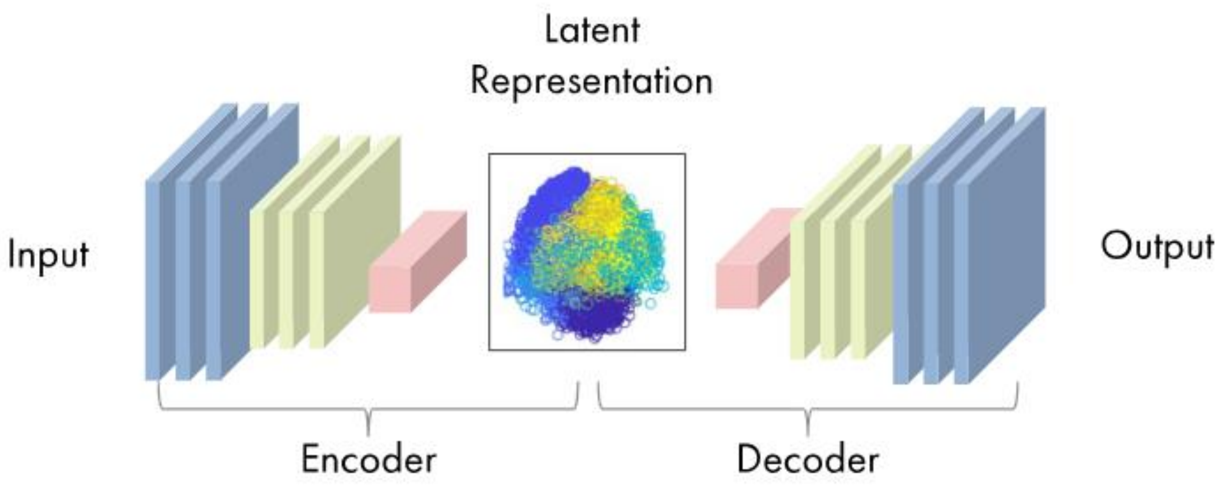


Image: Mathworks.com

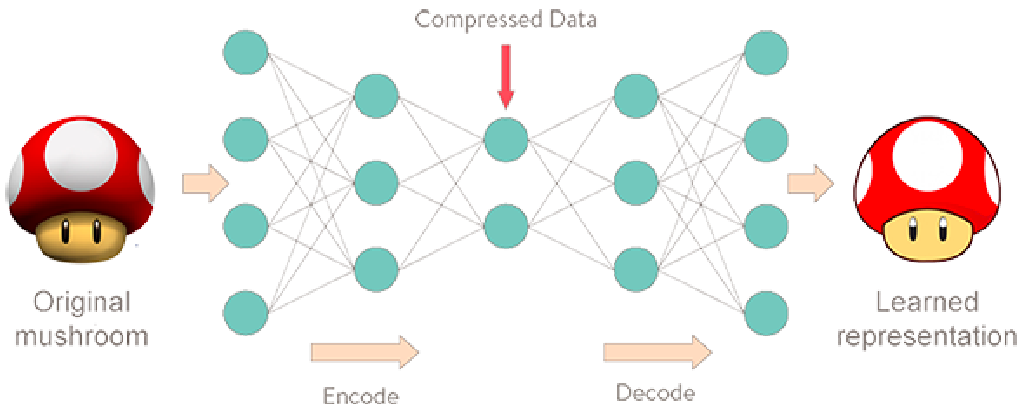


Image: bpesquet.fr

Autoencoder

Reconstruction error can be used to detect anomalies.

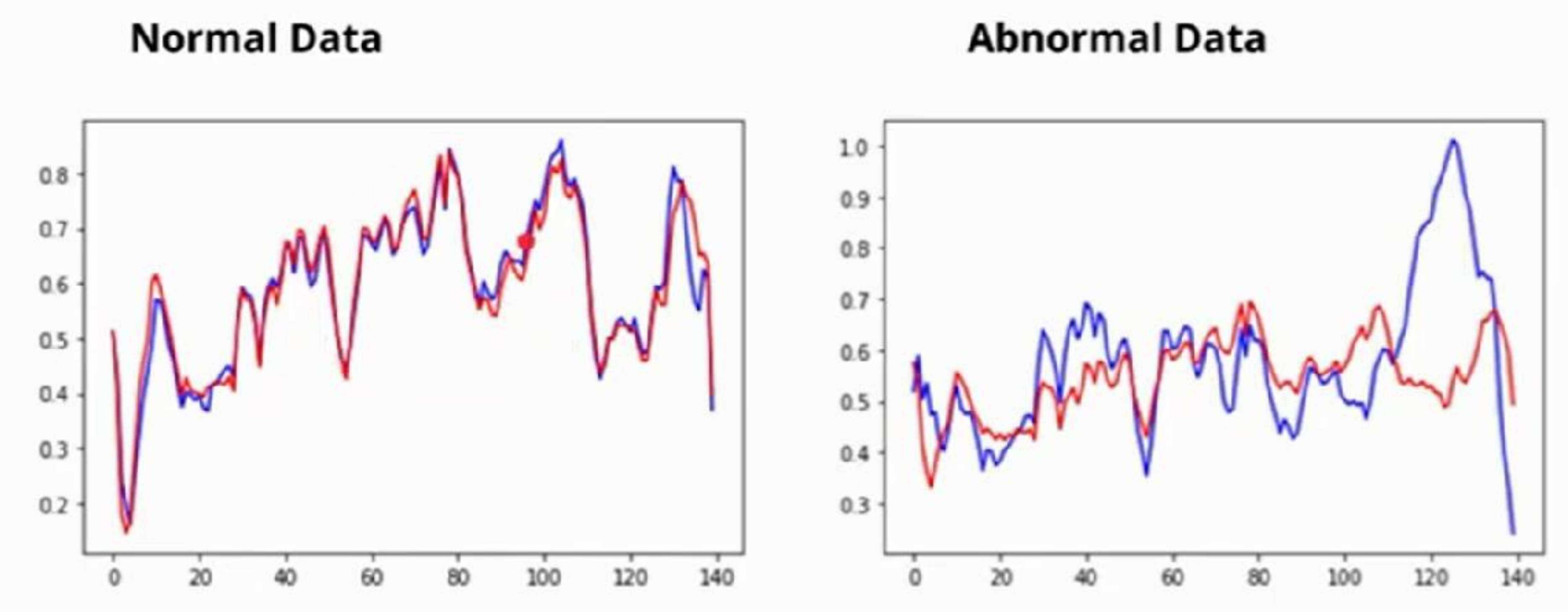


Image: analyticsvidhya.com

Recap

What should you take from today‘s lesson?

Summary

- AI can be used for **Simulations & Data Analysis** in HEP

- Challenges in HEP include
  - Data Preprocessing
  - Clustering
  - Pattern Recognition
  - Classification
  - Anomaly Detection

- **AI is not all** – we need to be **good Data Scientists** i.e. understand Statistics, Classic Machine Learning & Deep Learning