UFPR

UNIVERSIDADE FEDERAL DO PARANÁ

Federated Learning and applications of data visualization techniques in machine learning

Bruno Henrique Meyer
Aurora Pozo
Wagner M. Nunan Zola

2023

Introduction

- Bruno Henrique Meyer

- Currently PhD student in Computer Science (Artificial Intelligence)
  - CBio (Bio-inspired Computation g
  - Advisors: Aurora Pozo (UFPR), Wagner M. Nunan Zola (UFPR), Michele Nogueira (UFMG)

- Usage of Artificial Intelligence for cybersecurity issues

Summary

- Introduction to Artificial Intelligence, Machine Learning and Deep Learning
- Data visualization
- Federated Learning
- Hypothesis and experiments
- Opportunities and conclusion remarks

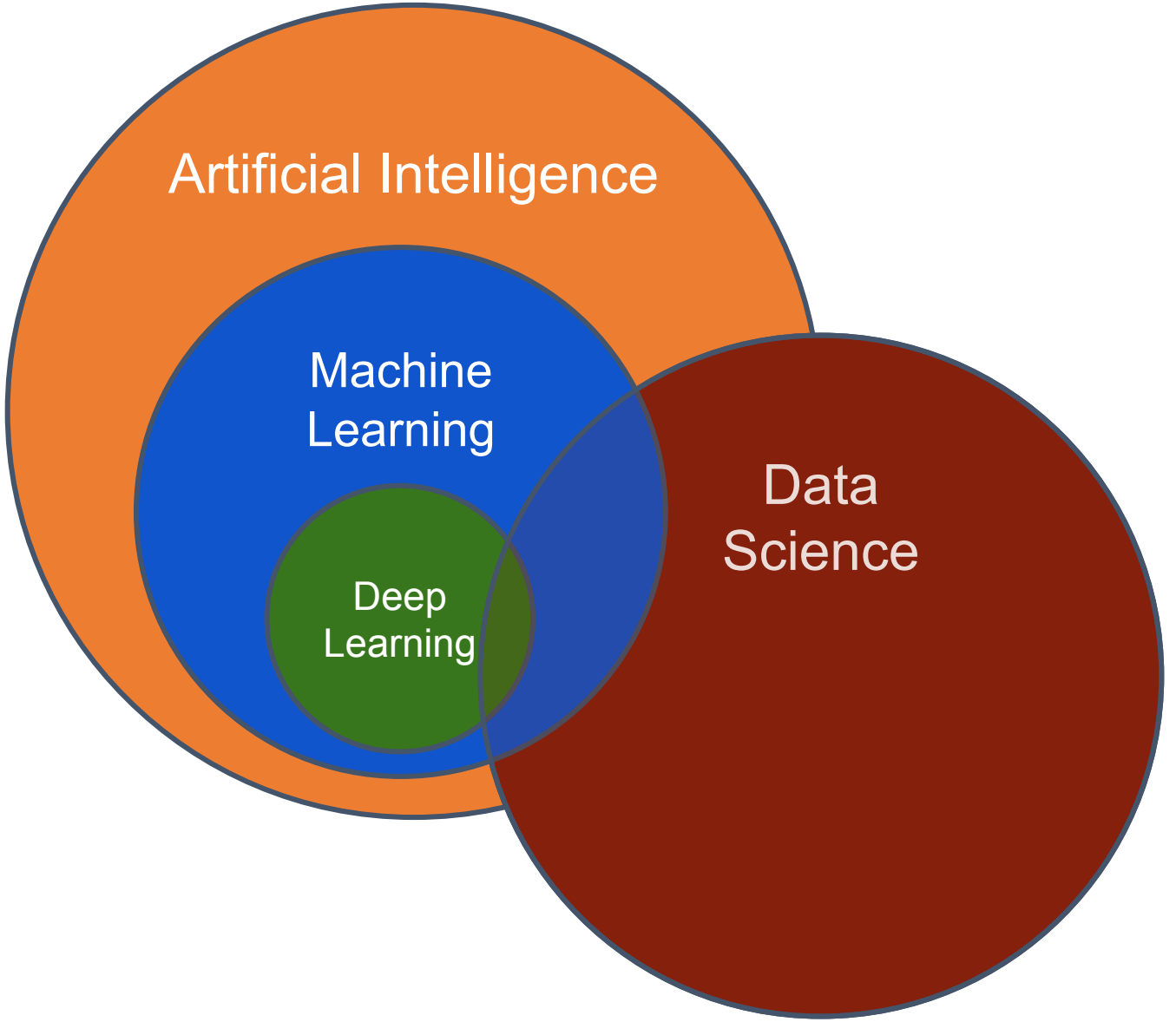
Context

- **Data** is present in large amounts in many projects
  - Hard to **interpret**
  - Hard to **manage**
  - Can be used by **Artificial Intelligence** solutions
  - Different formats: text, image, networks, etc.

- Complex problems
  - Expert knowledge (biology, physics, etc.)
  - Experiments and **data interpretation**
  - Technologies to accelerate processes and identify important phenomena

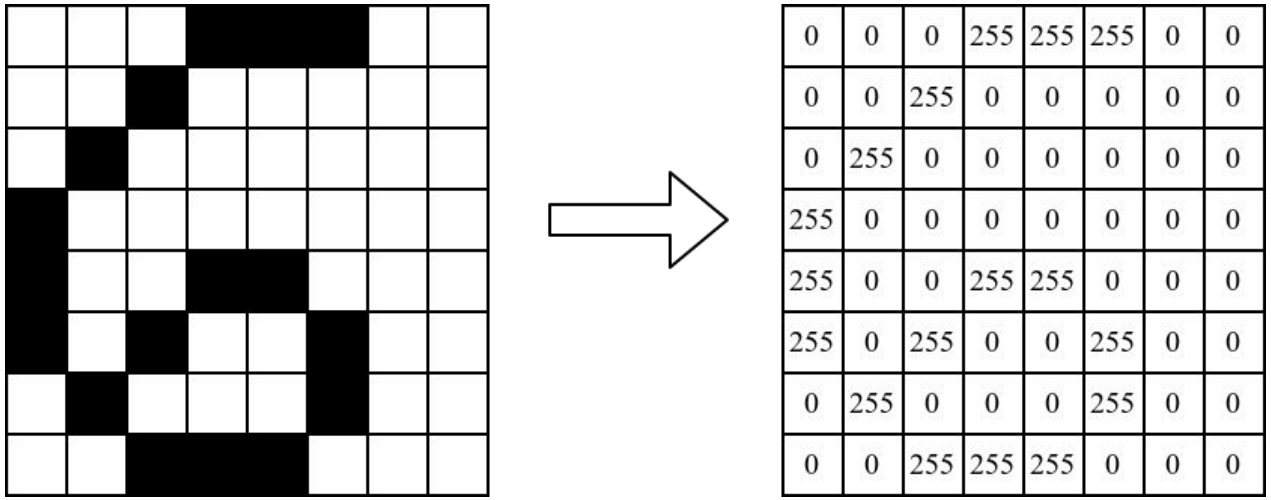
Introduction

- **Artificial Intelligence**
  - Many applications
  - Many problems
- **Machine Learning**
  - Data representation
  - Labeled and unlabeled data
- **Deep Learning**
  - Neural networks
  - Automatically learn data representation
  - “Black Box”



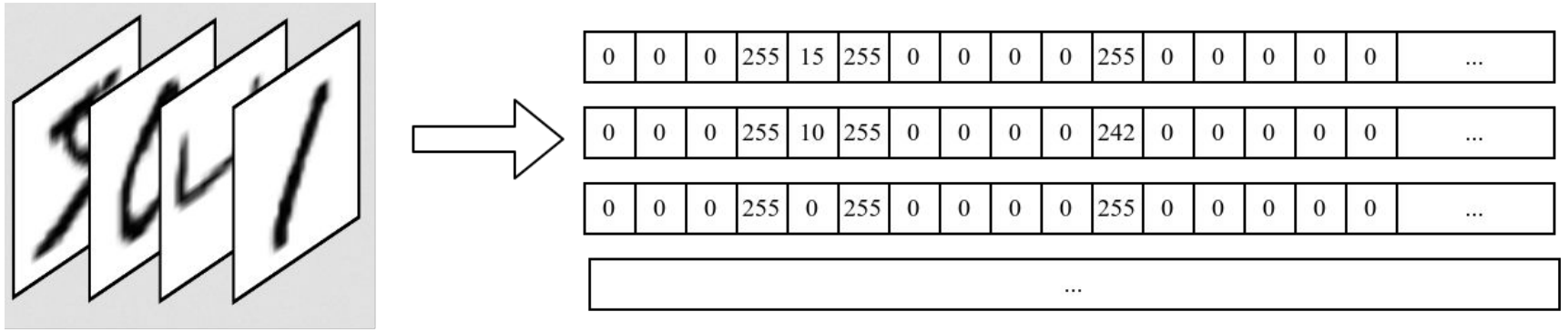
Data in Machine Learning

- Raw data in machine learning
- Structured format
- Data requires an representation
- Quantitative values



0	0	0	255	255	255	0	0
0	0	255	0	0	0	0	0
0	255	0	0	0	0	0	0
255	0	0	0	0	0	0	0
255	0	0	255	255	0	0	0
255	0	255	0	0	255	0	0
0	255	0	0	0	255	0	0
0	0	255	255	255	0	0	0

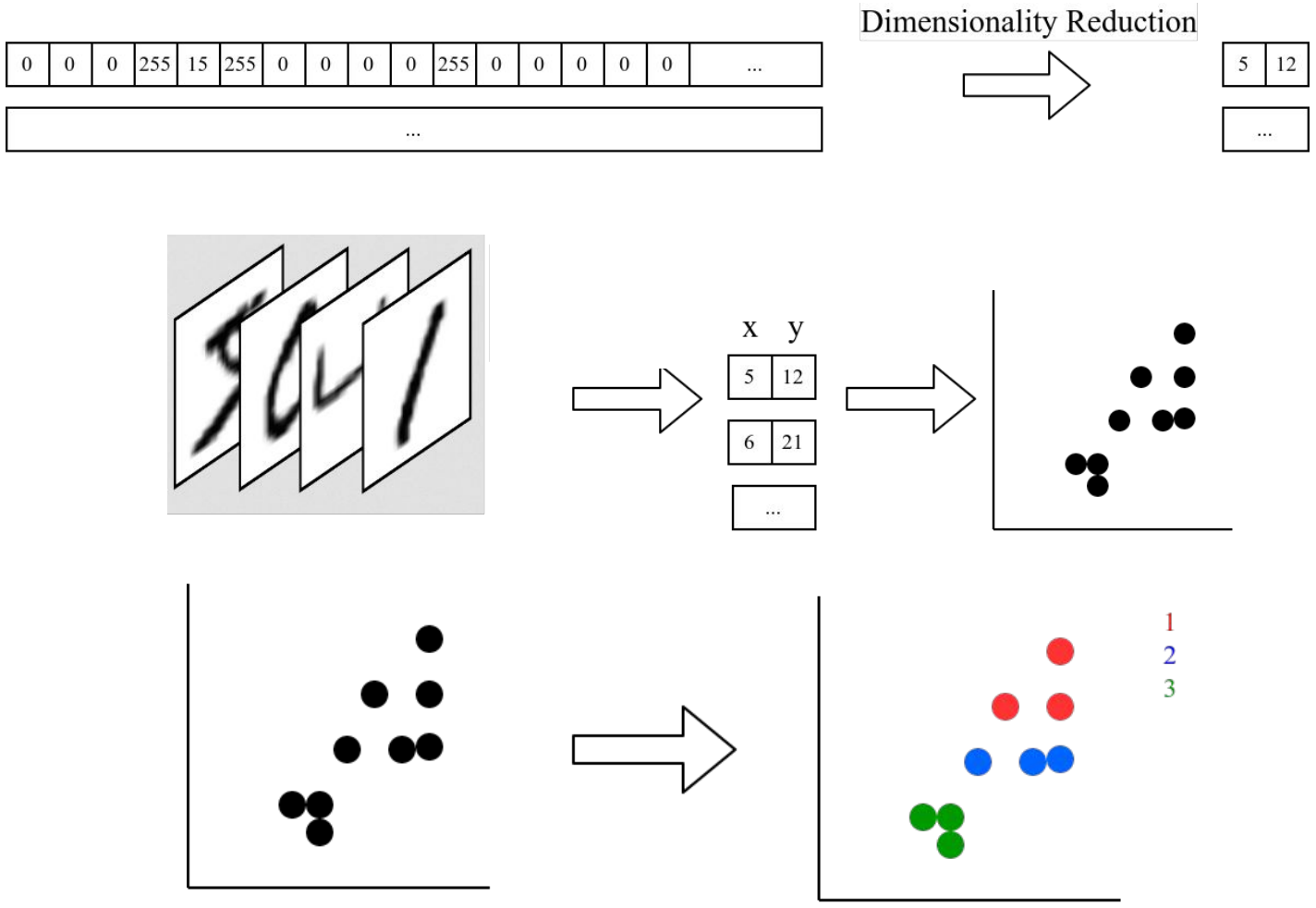
Data representation



- Most common approach: Several **high-dimensional vectors**
- In the past: Human engineering to draw features for each application
- Today: Automatically learn representations from raw-data with deep learning

Visualizing data with machine learning

- Dimensionality reduction
- Interpret machine learning representations of raw data
- Famous algorithm used for dimensionality reduction: **t-sne**



Visualizing data with machine learning

Journal of Bioinformatics and Computational Biology | Vol. 15, No. 04, 1750017 (2017) | Research Papers

Full Access

Application of t-SNE to human genetic data

Wentian Li, Jane E. Cerise, Yaning Yang, and Henry Han

https://doi.org/10.1142/S0219720017500172 | **Cited by:** 98

< Previous

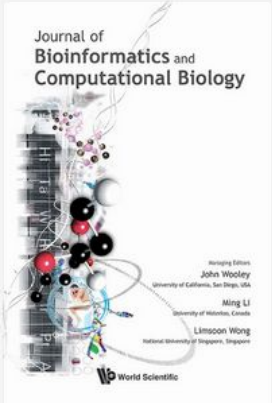
PDF/EPUB

Tools Share Recommend To Library

Abstract

The t-distributed stochastic neighbor embedding t-SNE is a new dimension reduction and visualization technique for high-dimensional data. t-SNE is rarely applied to human genetic data, even though it is commonly used in other data-intensive biological fields, such as single-cell genomics. We explore the applicability of t-SNE to human genetic data and make these observations: (i) similar to previously used dimension reduction techniques such as principal component analysis (PCA), t-SNE is able to separate samples from different continents; (ii) unlike PCA, t-SNE is more robust with respect to the presence of outliers; (iii) t-SNE is able to display both continental and sub-continental patterns in a single plot. We conclude that the ability for t-SNE to reveal population stratification at different scales could be useful for human genetic association studies.

Figures References Related Details



Vol. 15, No. 04

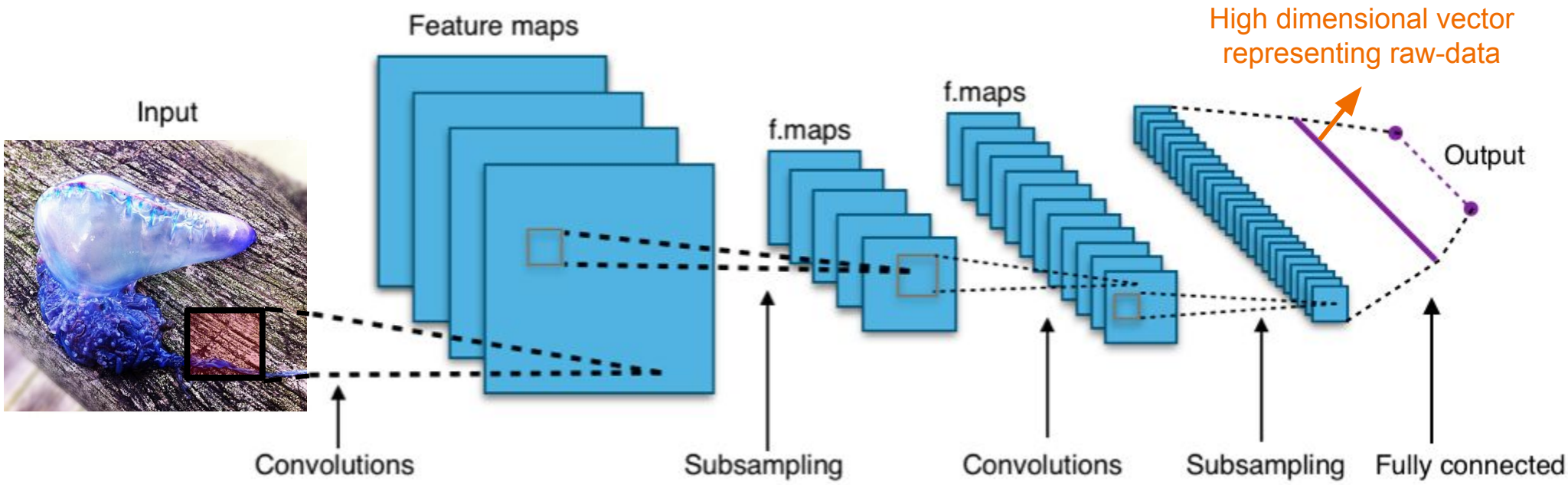
Metrics

Downloaded 317 times

Citations 120

History

Convolutional Neural Networks (CNN)



Accepted as man-of-war

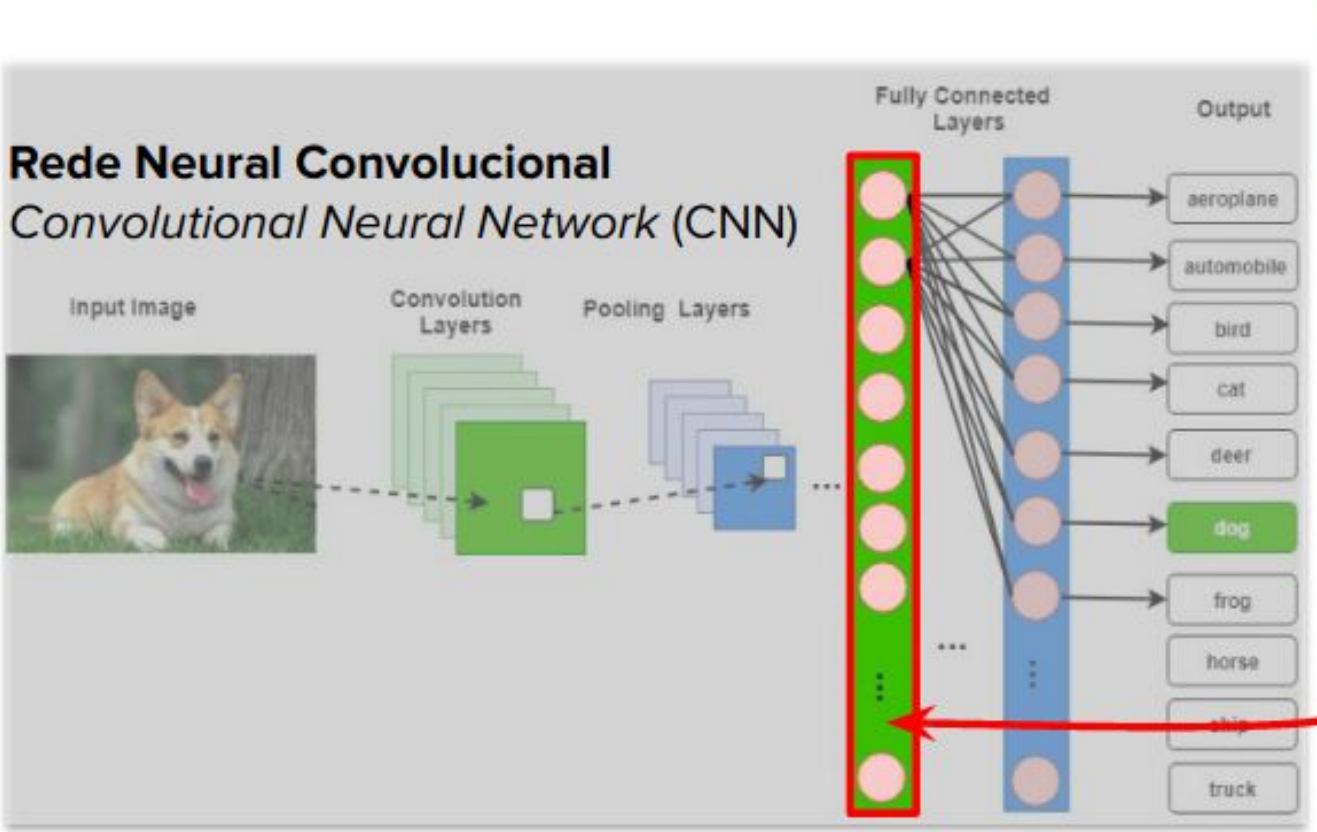
or

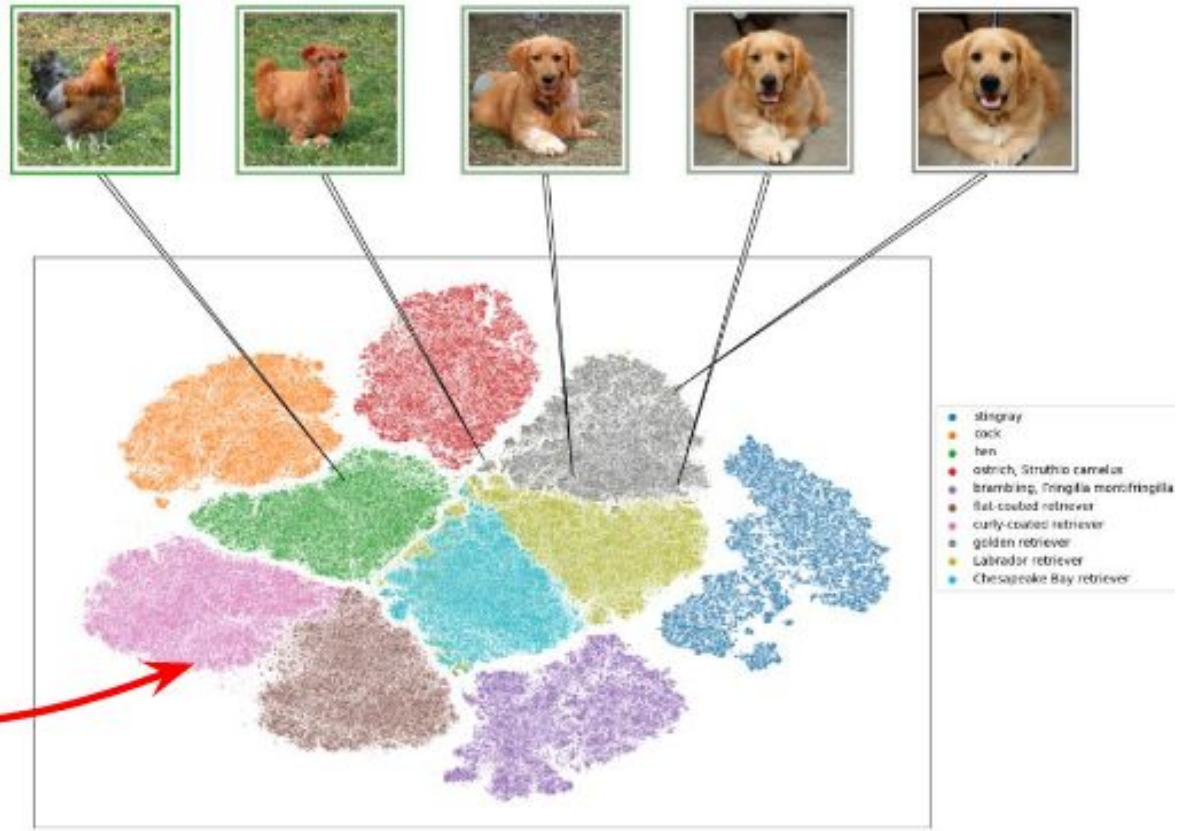
Rejected as man-of-war

- Popular technique used in Deep Learning
- Advantage: High classification **accuracy** (sometimes better than experts)
- Disadvantage: **Black Box** model

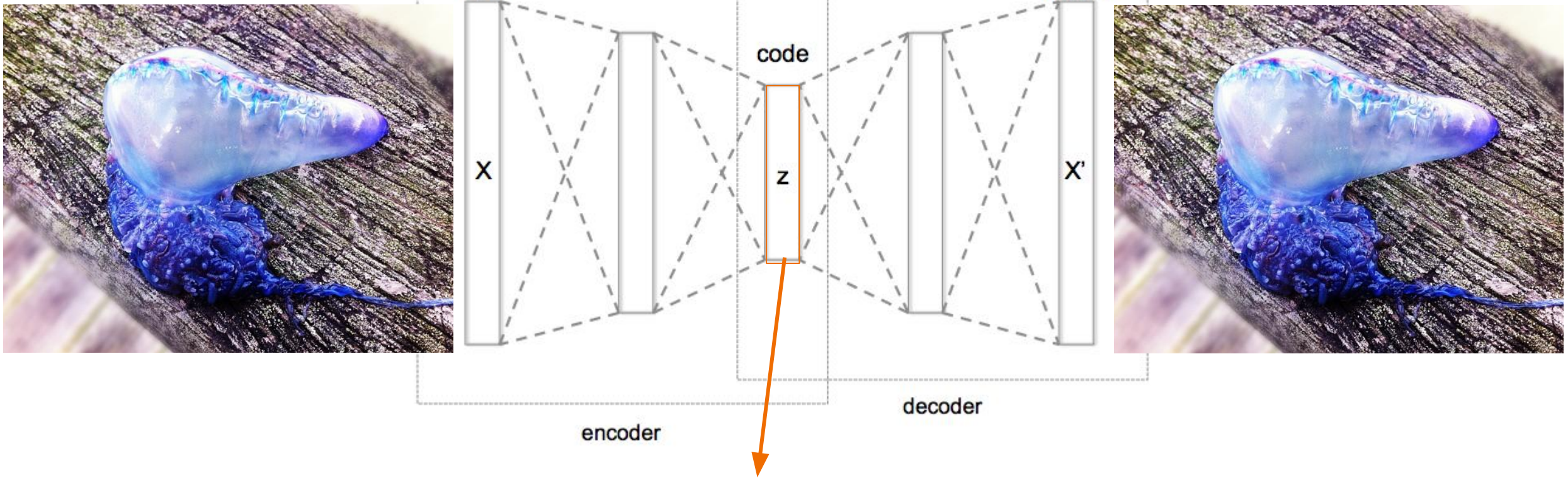
Visualizing CNN representation

With Visualization techniques: **Black box → "Gray box"**



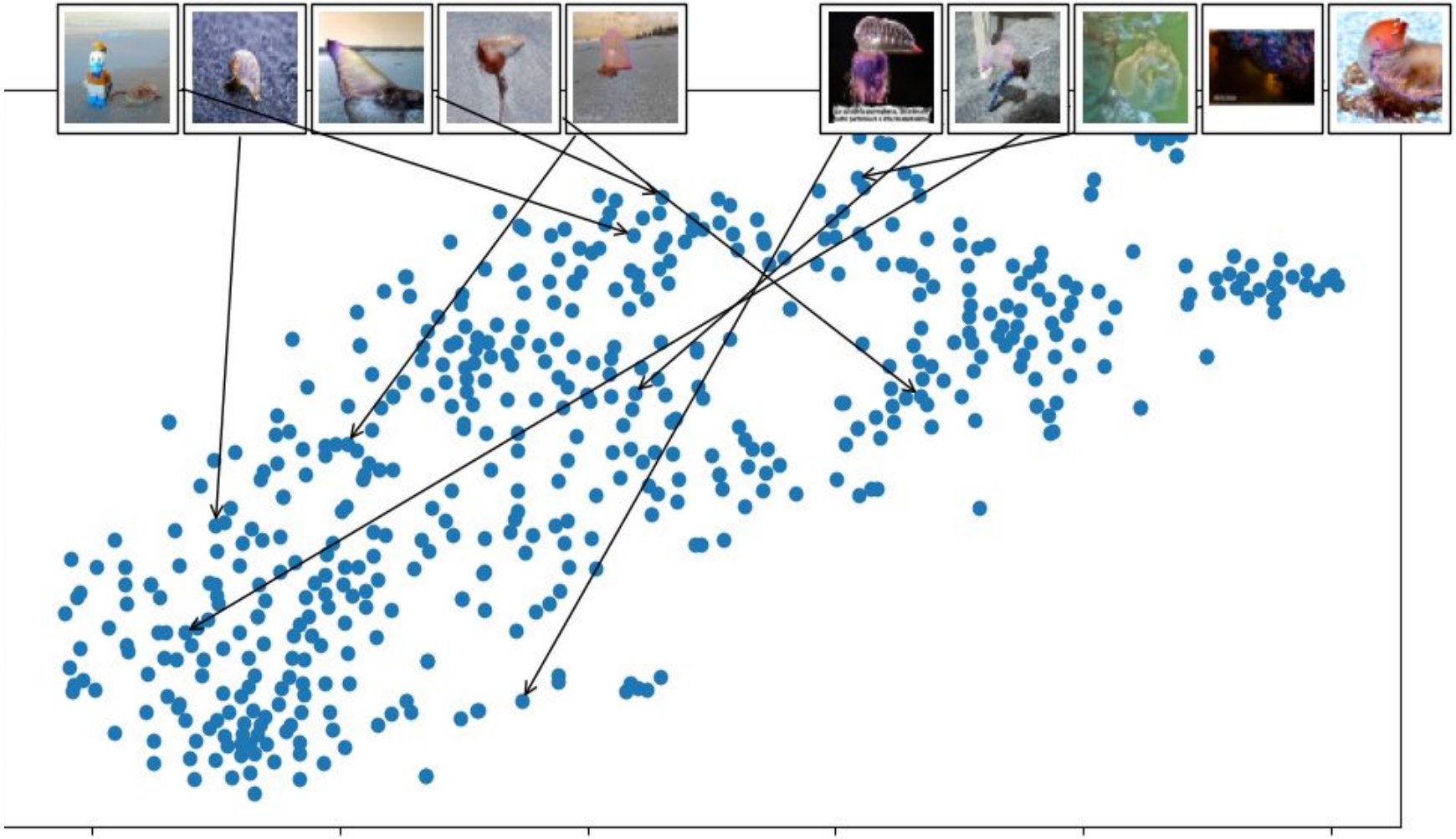


Autoencoder

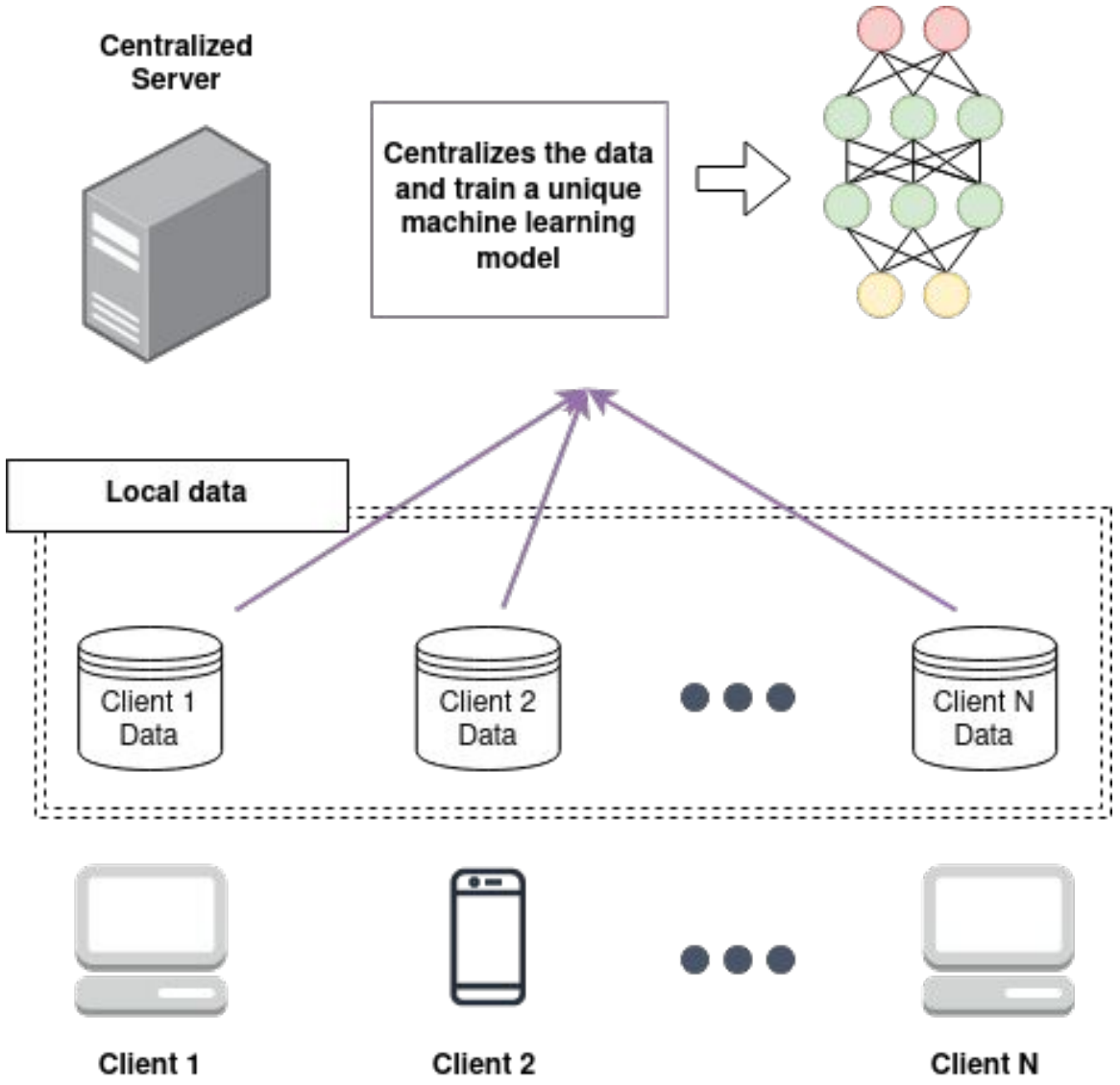


- High dimensional vector representing raw-data
- Unsupervised learning: Learn **without labels**

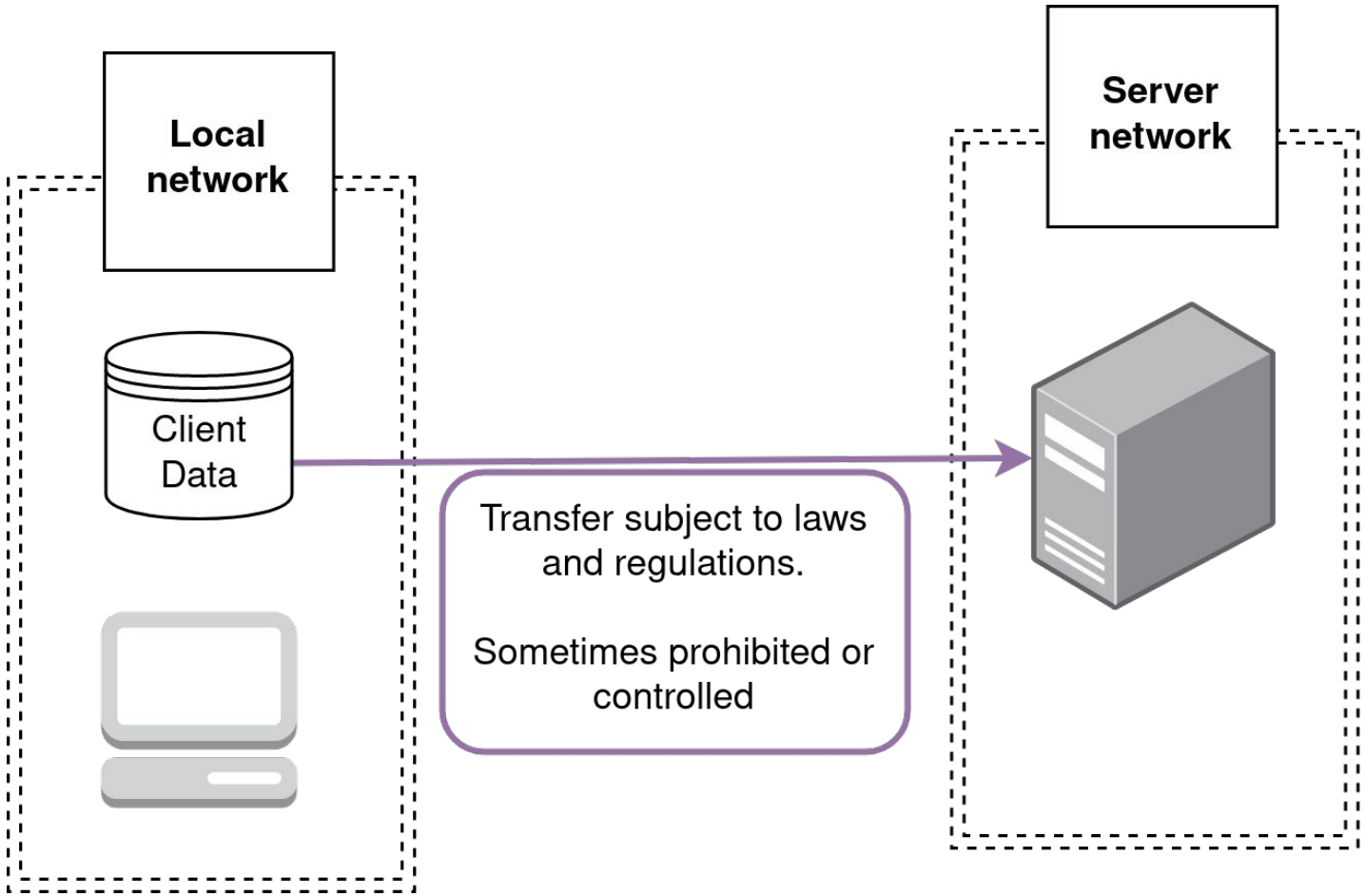
Autoencoder



Basic workflow for applications

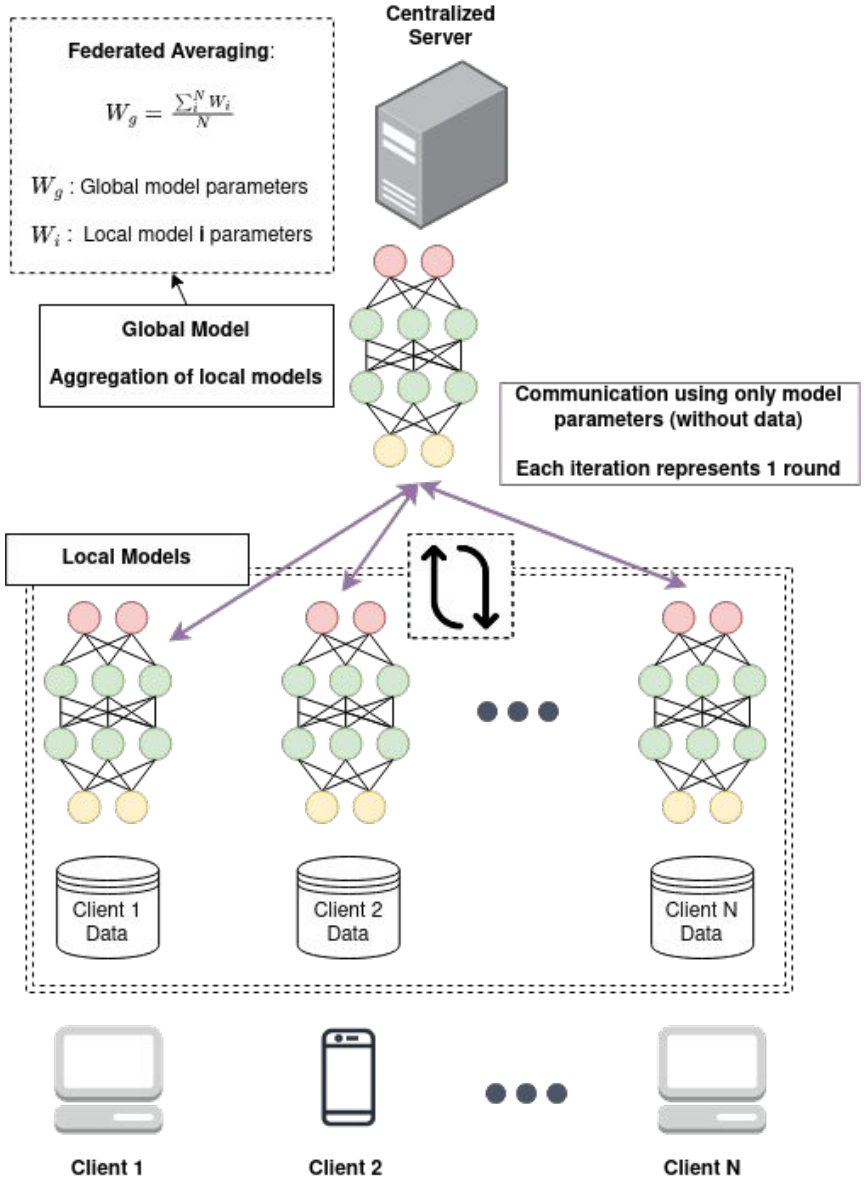


Data privacy



- GDPR (General Data Protection Regulation)
- LGPD (General Data Protection Law in Brazil)

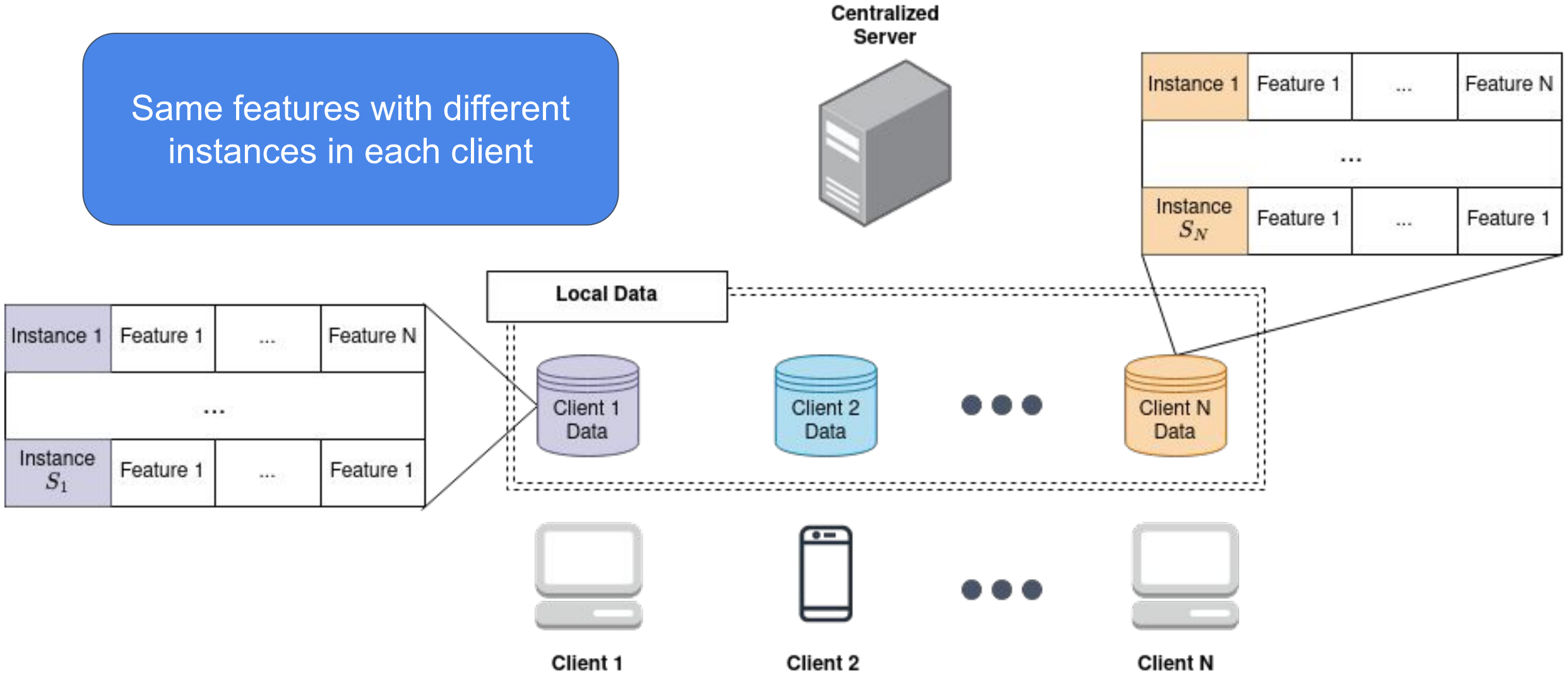
Federated Learning



- Learn machine learning models **without centralizing** data
- Most popular algorithm: FedAvg
- Can learn almost any type of deep learning model
- Challenges to create good models considering different types of data
- Distributed and parallel computing

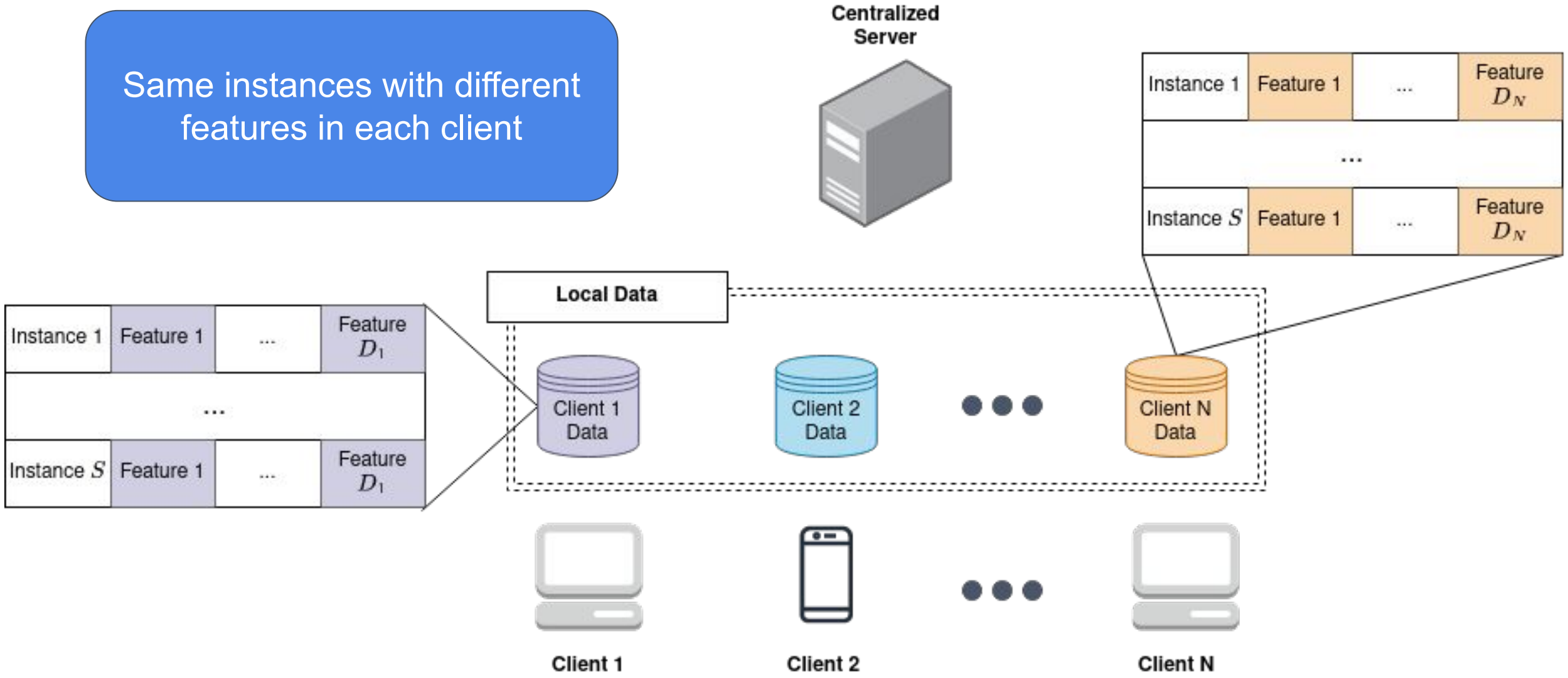
Horizontal Federated Learning

Same features with different instances in each client

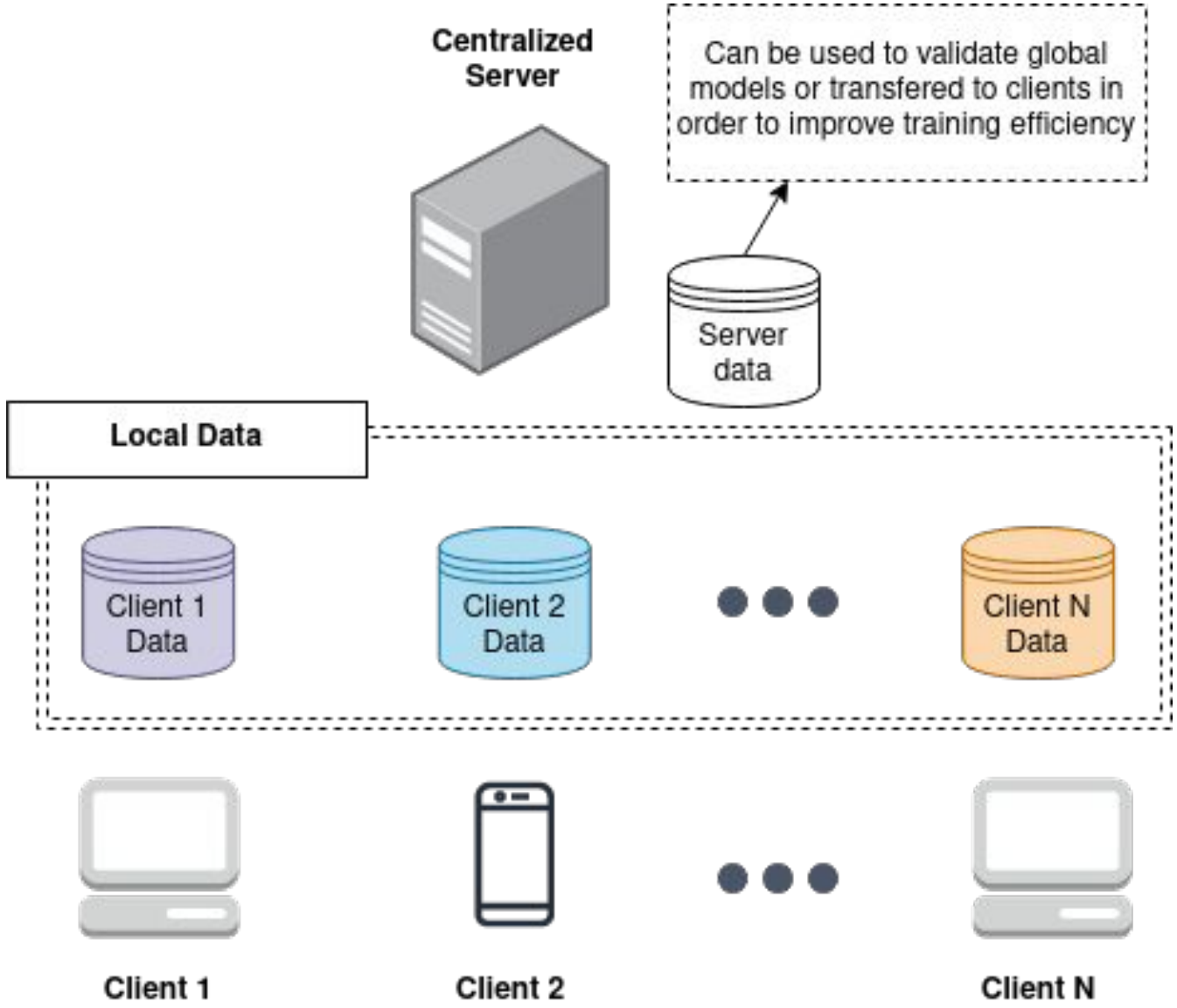


Vertical Federated Learning

Same instances with different features in each client



Federated Learning with public data



- Some problems considers a few amount of labeled and reliable data
- Improve clients models accuracy

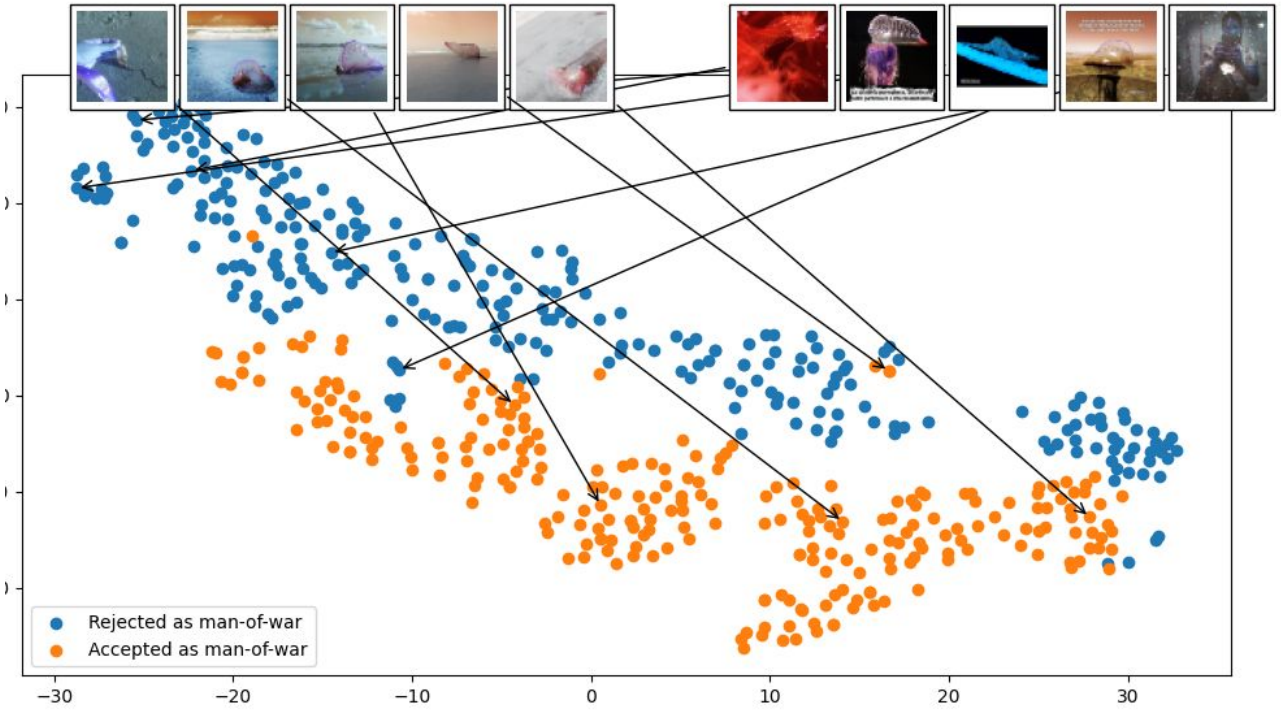
Hypothesis and Experiments

- Hypothesis 1: We can use data visualization techniques to aid data interpretation in the RESMA project

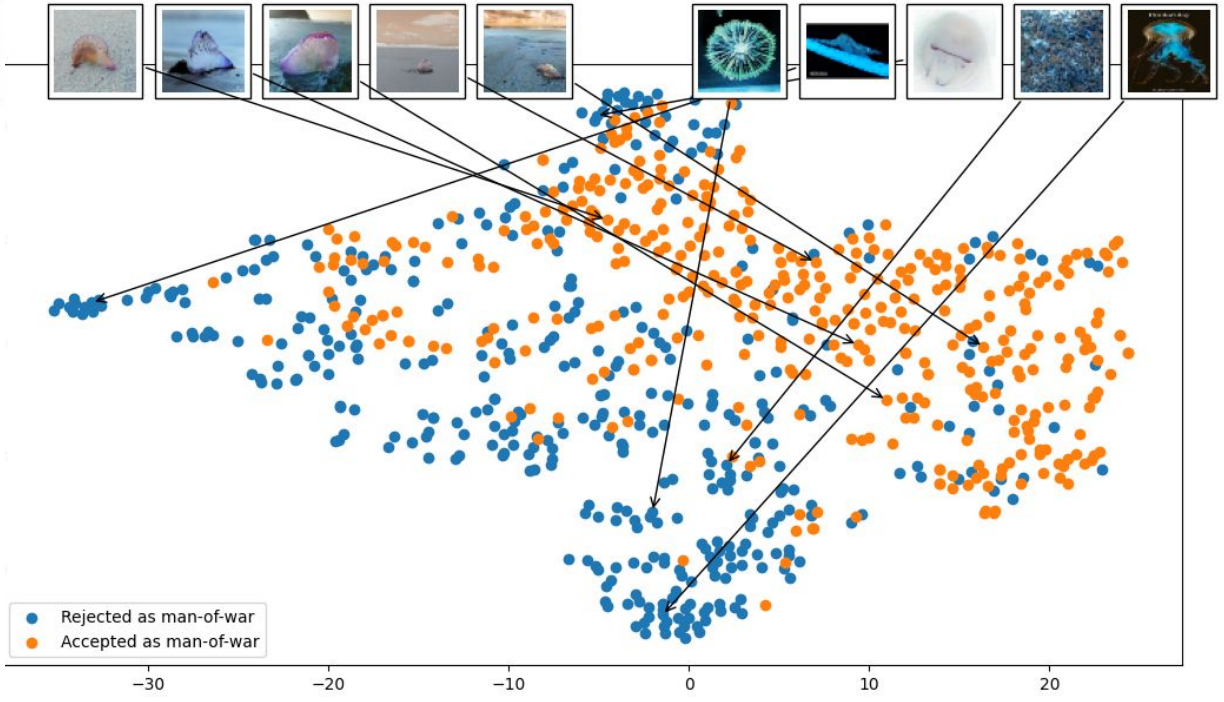
- Hypothesis 2: We can apply Federated Learning to aid data interpretation in the RESMA project

- Experiments
  - Apply Deep Learning for classification of man-of-war (physaliaphysalis)
  - Apply Autoencoder for data representation of man-of-war
  - Adapt the previous techniques for Federated Learning context

Classifier (CNN)

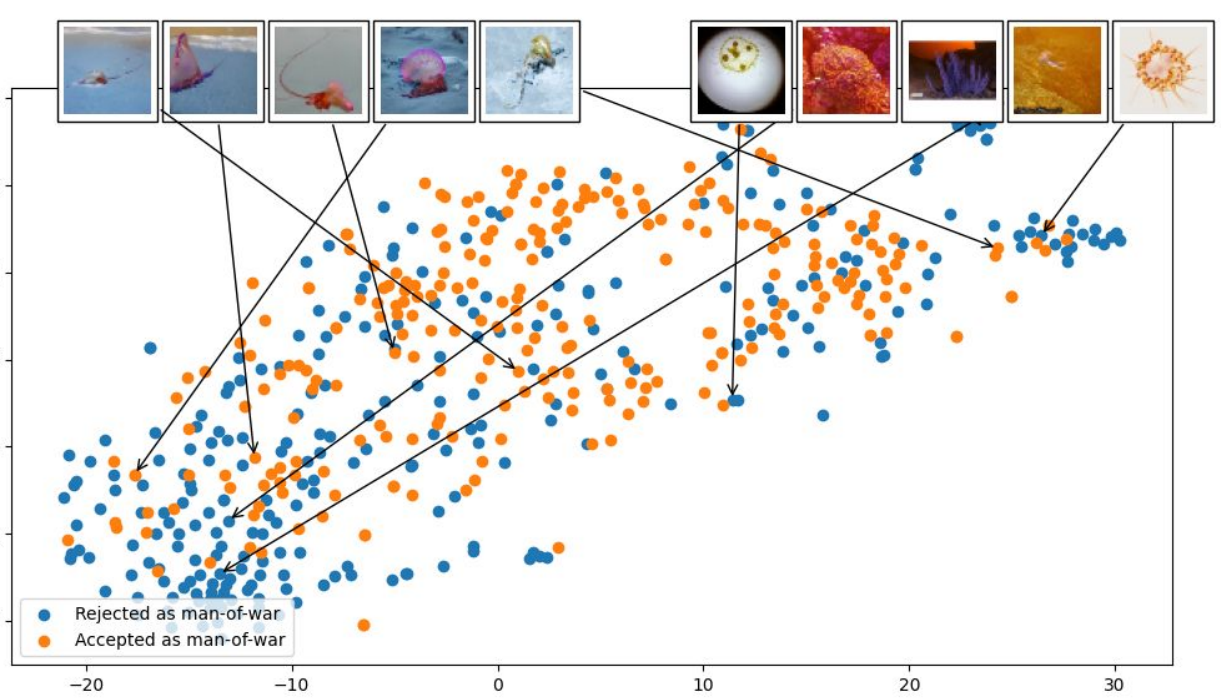


Centralized Learning

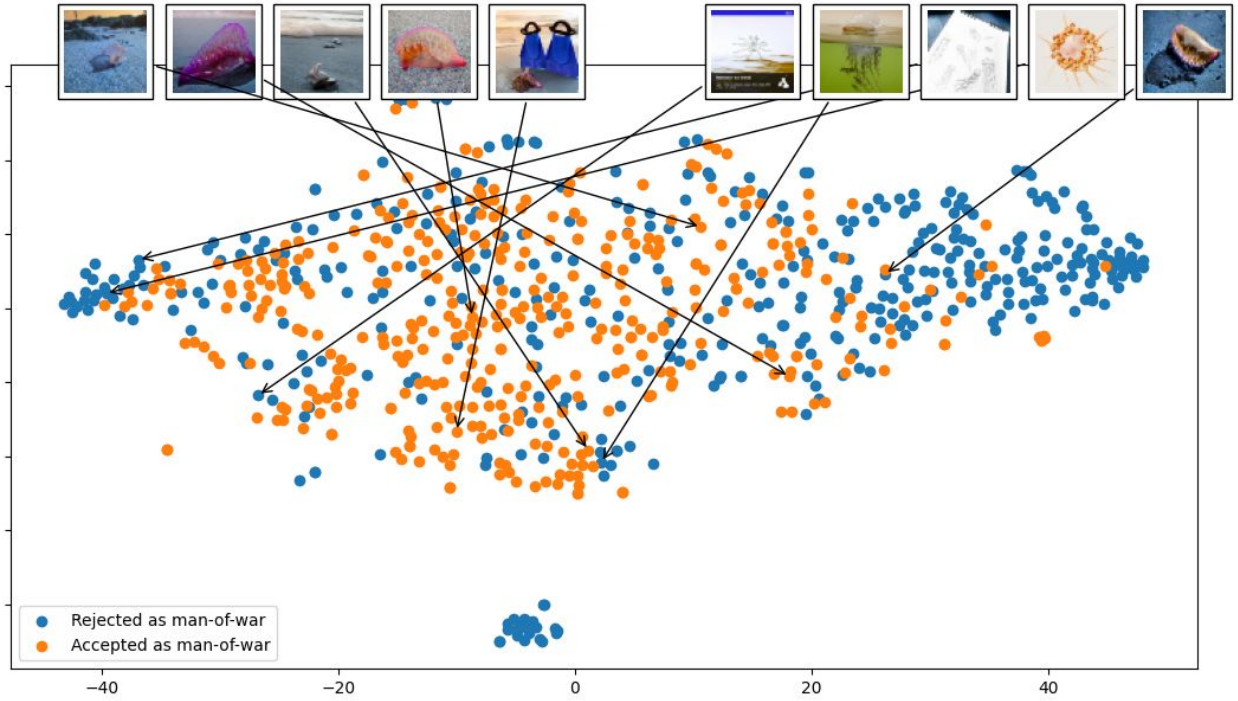


Federated Learning (10 clients)

Autoencoder

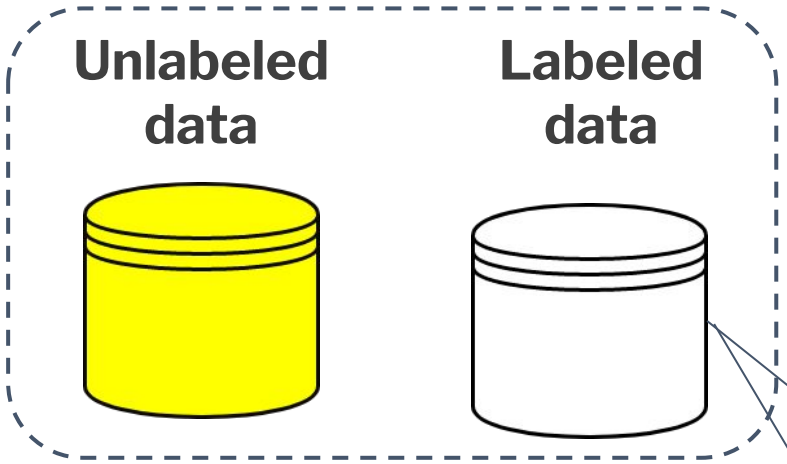


Centralized Learning



Federated Learning (10 clients)

Application Example 1



Helps experts to filter, curate and identify relevant data

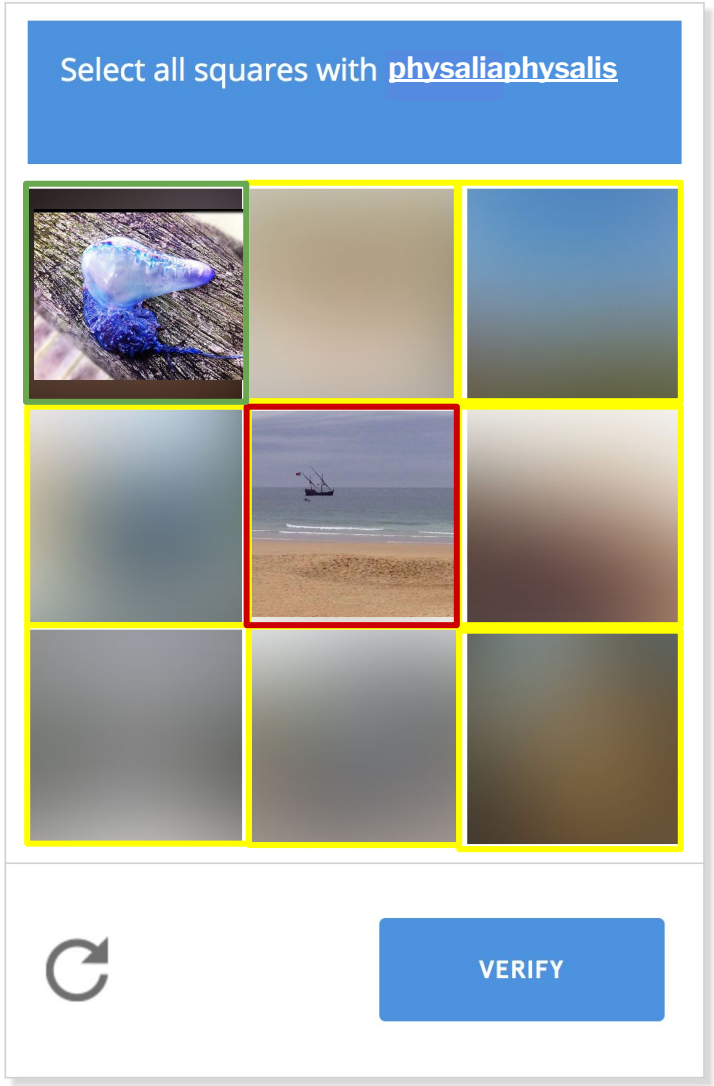
Machine learning techniques

Positive label candidates

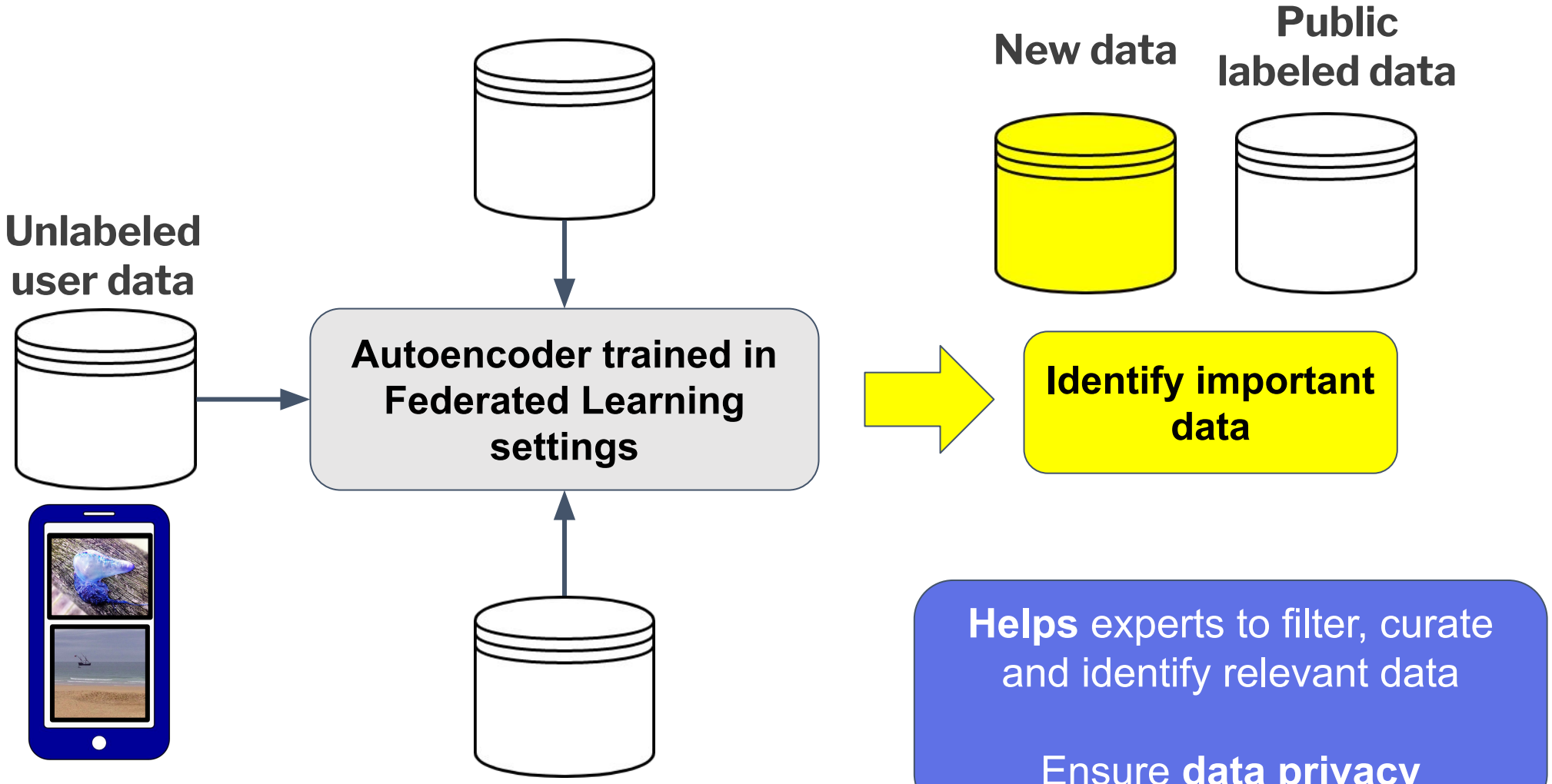
Known positive labels

Known negative labels

Unknown labels



Application Example 2



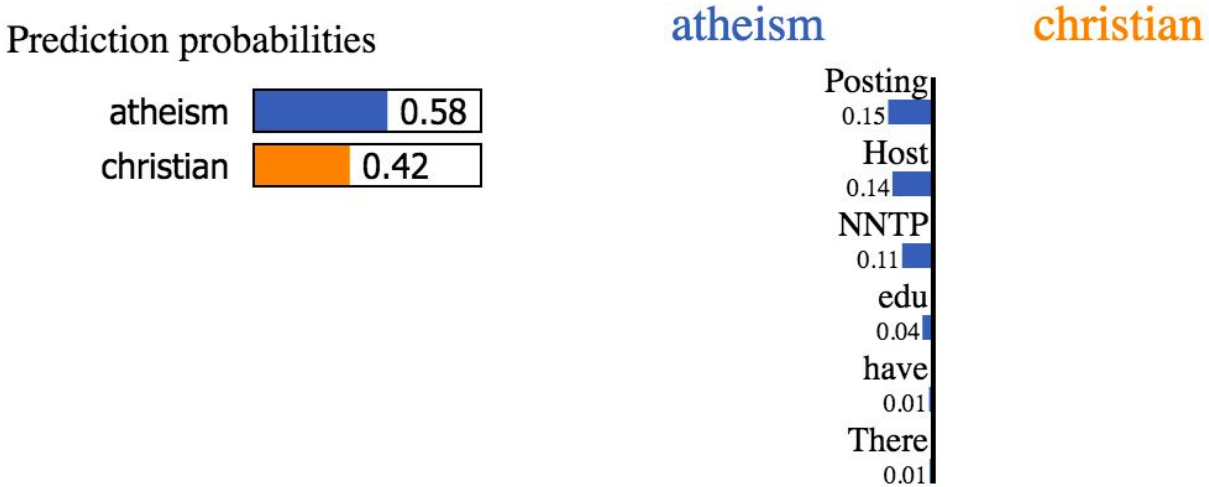
Helps experts to filter, curate and identify relevant data

Ensure **data privacy**

Opportunities

- Semi-supervised learning
  - Labeled and unlabeled data to train a single model
- Anomaly detection
  - Filter poor quality data
  - Identify exceptions
- Few-shot learning
  - Model trained with few data and able to generalize its representation
- Explainable AI (xAI)
  - White-Box algorithms and systems
  - Analyse Black-Box models decisions

Explainable AI for text classification



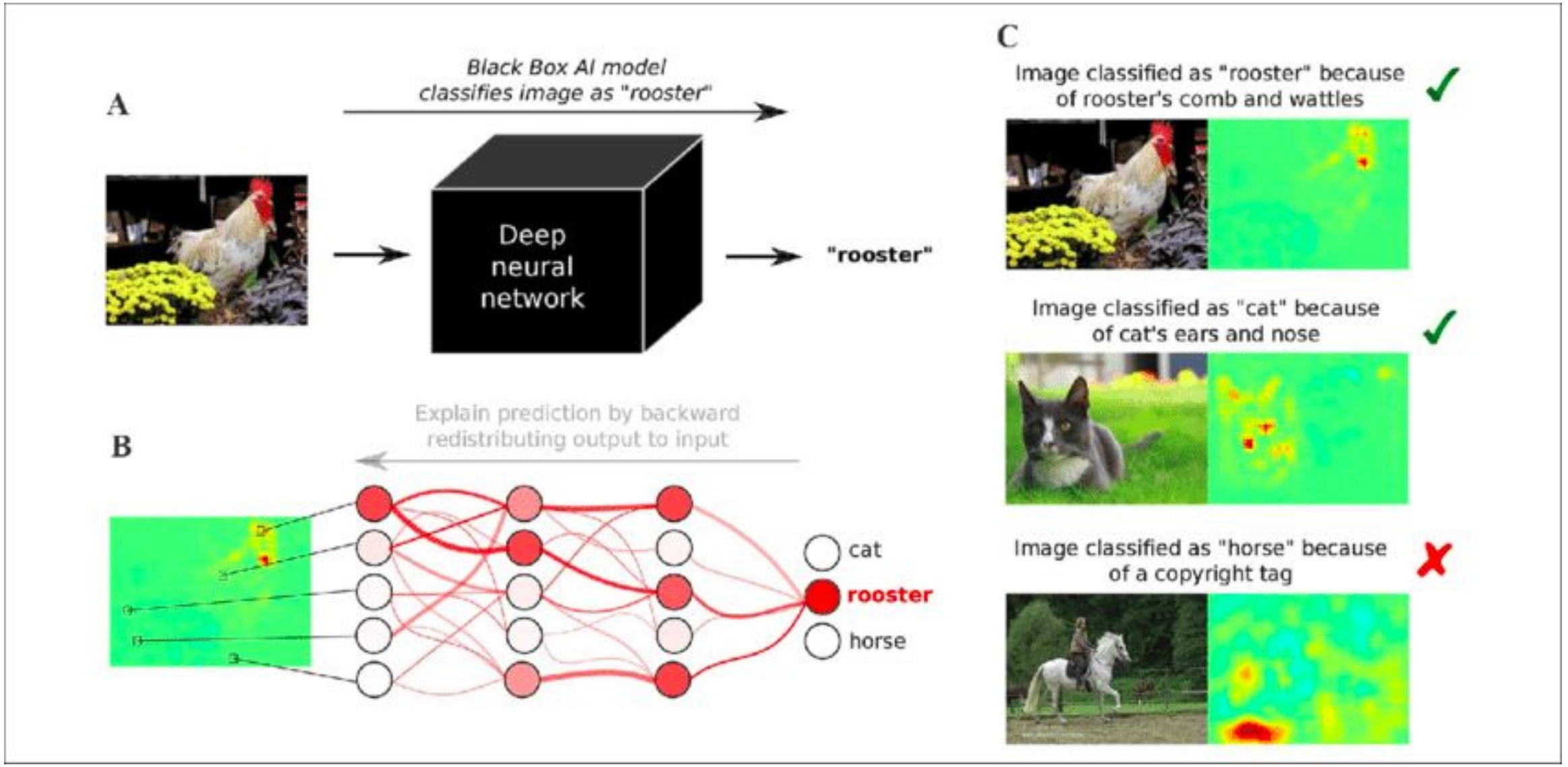
Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Explainable AI for image classification



Q&A

bruno.meyer@ufpr.br