PHAROS
THE GREEK AI FACTORY

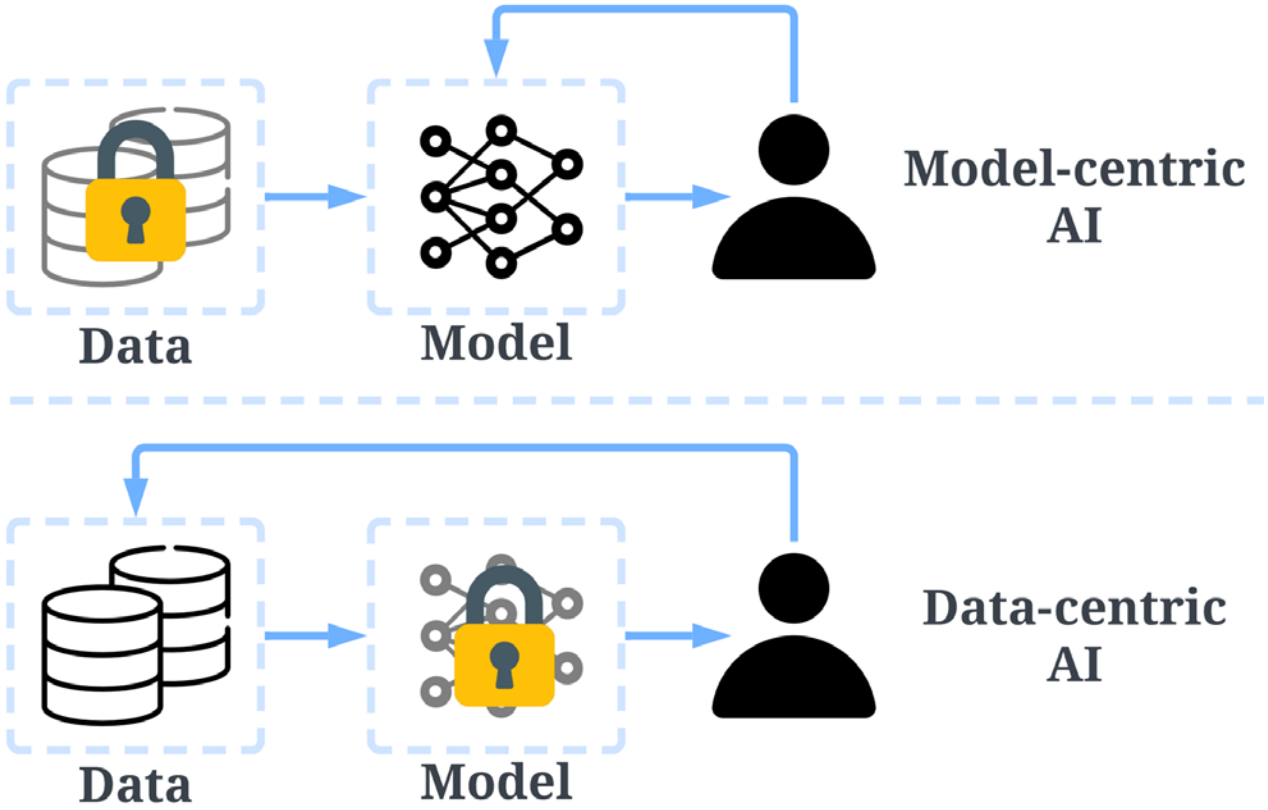
Data Generation Techniques for Medical Data

Ioannis Vlahavas, Prof.
Vasileios Kochliaridis, PhD Candidate
Zoi Katsantoni, MSc Graduate

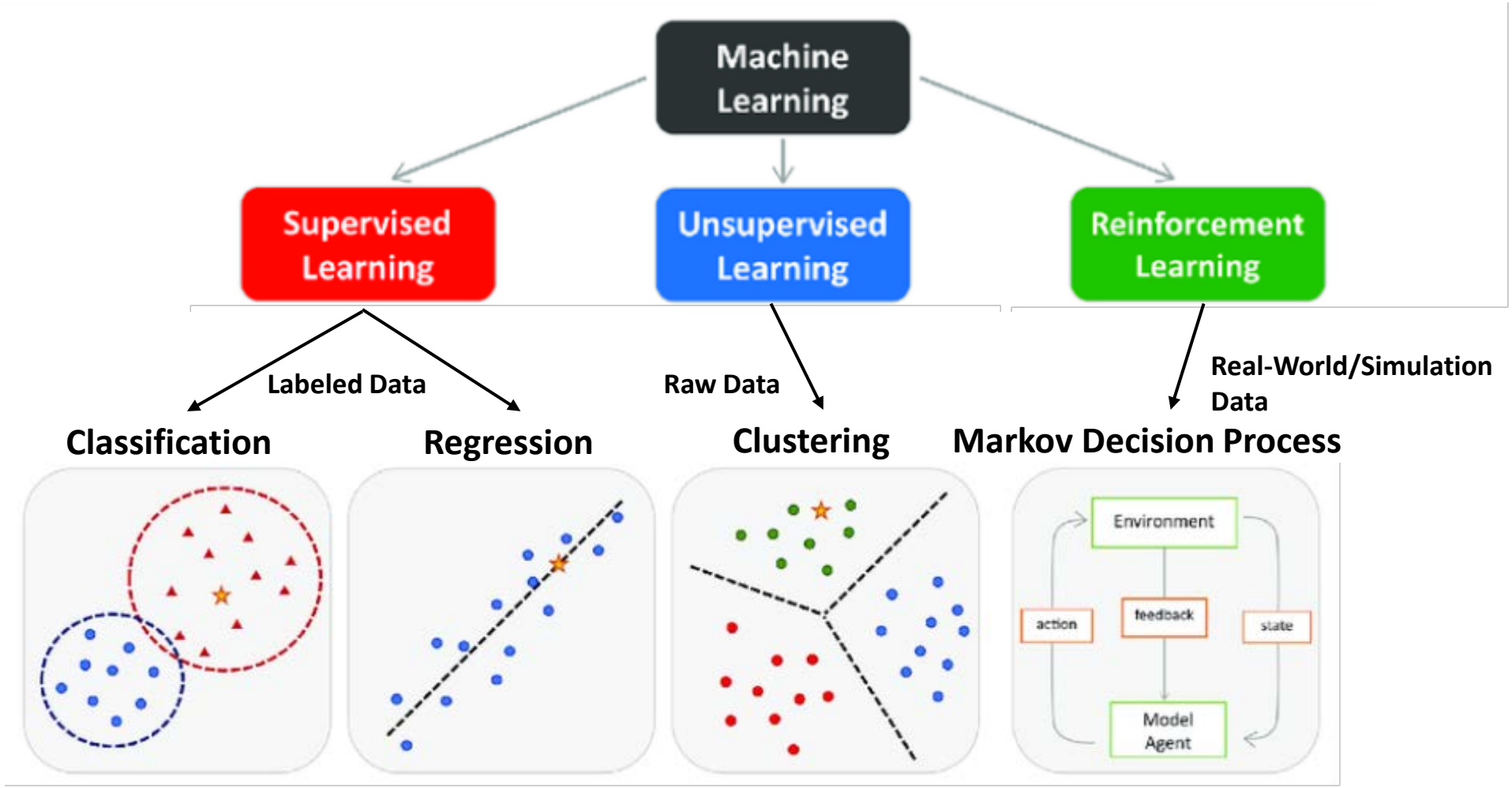
Department of Informatics, AUTH

Data-Centric AI: Motivation and Scope

- **Goal**: Shift from model-centric to data-centric ML
- Data quality as a key driver of model performance
- Requires large, high-quality datasets, that ensure:
  - Reliability
  - Diversity
  - Representativeness



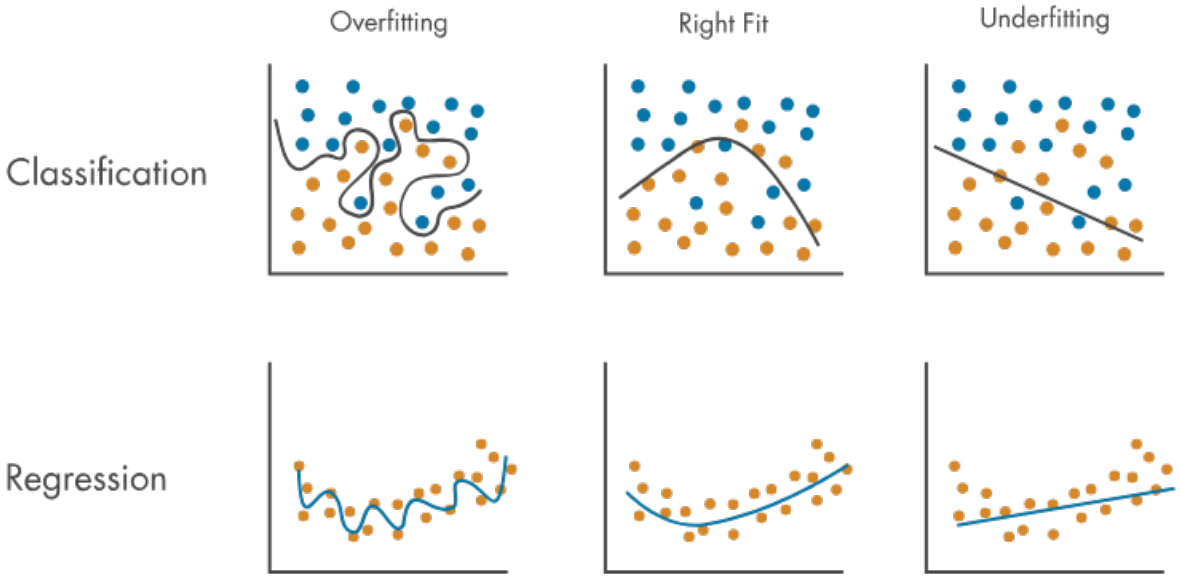
Machine Learning (ML) Fields



Machine learning learns from data

The Role of Data in Machine Learning Performance

- ML algorithms learn patterns directly from data
- Model performance depends heavily on the quality of the training data
- **High-Quality Data:**
  - Large datasets → Robustness
  - Diversity → Generalization
- **Low-Quality Data:**
  - Small datasets → Underfitting/Overfitting issues
  - Noise → Poor predictions
  - Class Imbalances → Poor performance on minority classes



Types of Data Used in Machine Learning

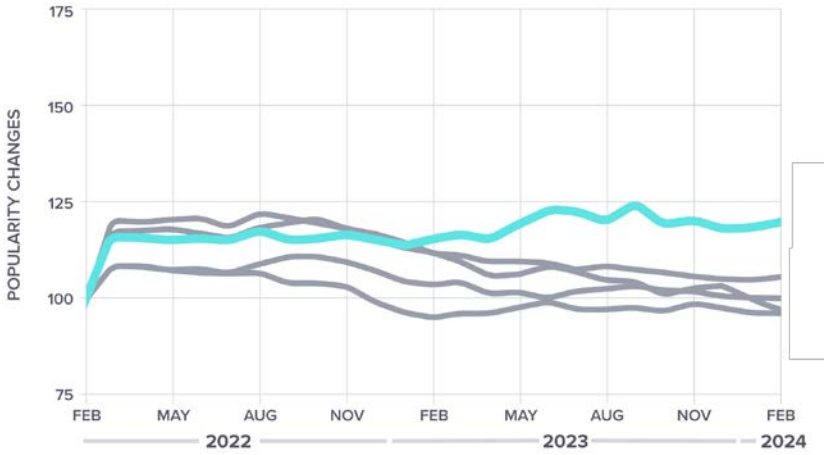
Tabular Data

ID	TOTAL ACTIONS	ACTION 1	ACTION 2	TOTAL TIME
10	120	80	40	0:50:05
11	255	130	125	1:40:03
12	180	100	80	1:20:19
13	305	205	100	1:58:58
14	71	50	21	0:35:41
15	418	310	108	2:08:18
16	222	150	72	1:32:58

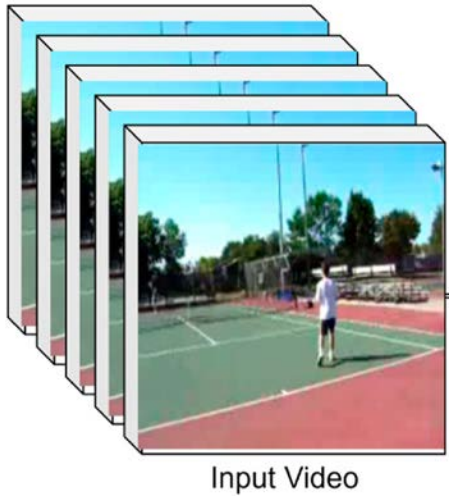
Visual Data

Natural Language (Text)

Time-Series

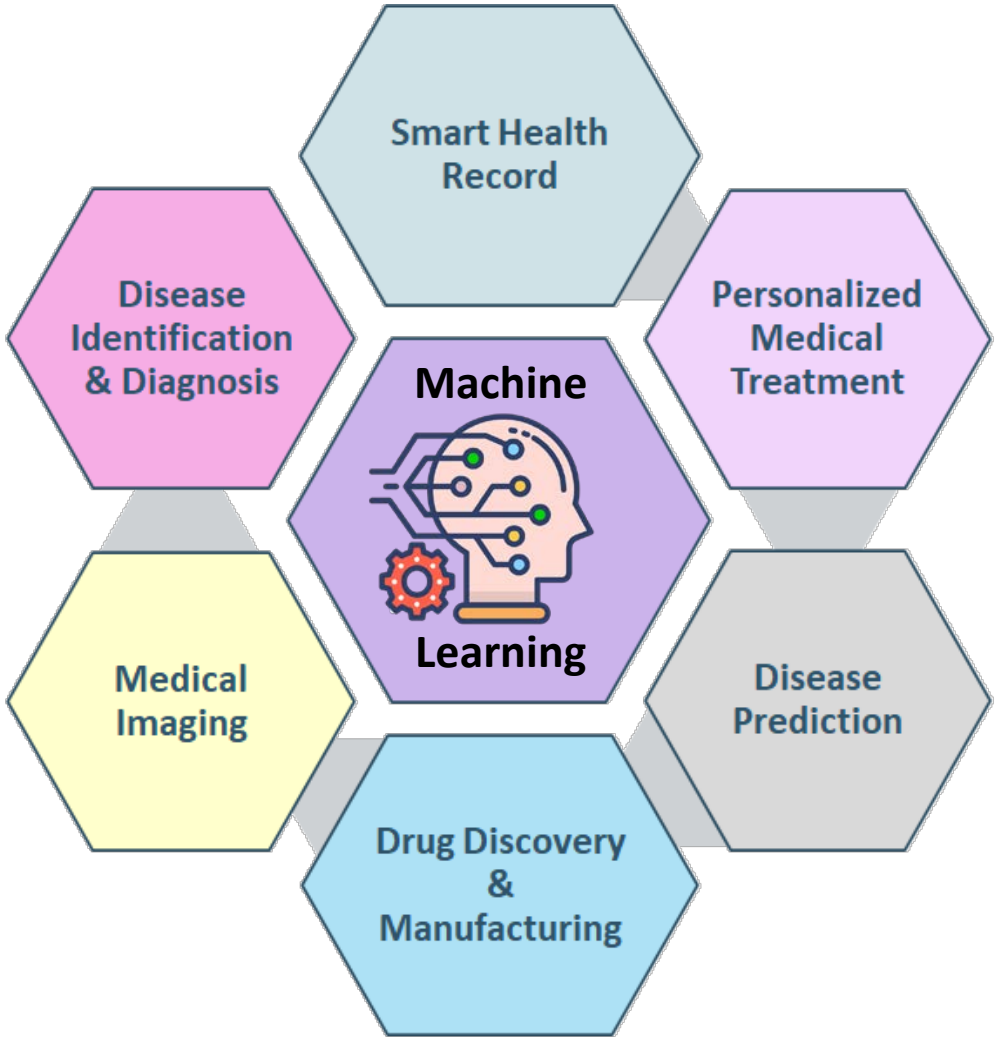


Video Streams



Sound/Speech

Machine Learning for Medical and Healthcare Applications

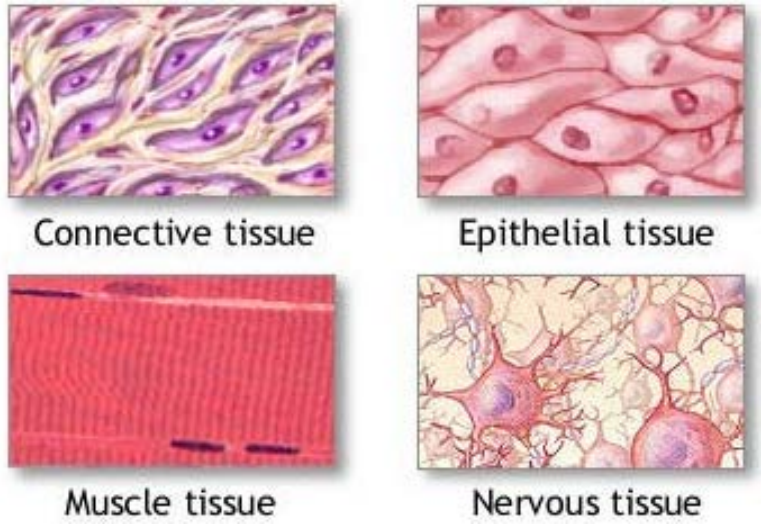


Electronic Health Records as a Data Source

- Electronic Health Records (EHRs) are a major source of healthcare data
- Patient information:
  - Demographics
  - Diagnoses
  - Medications
  - Laboratory results
  - Procedures
  - Vital signs
- Ideal for ML applications:
  - Disease prediction
  - Patient stratification
  - Outcome forecasting
  - Clinical decision support

Medical Imaging Data in AI

- Provide rich visual information of a patient:
  - Anatomy
  - Tissue structure
  - Disease-related abnormalities
- Medical Images:
  - X-Rays
  - MRI
  - CT
  - PET
  - Ultrasound
- ML applications:
  - Detection
  - Classification
  - Segmentation
  - Prognosis
- Medical imaging data is often high-dimensional and requires substantial storage and computational resources

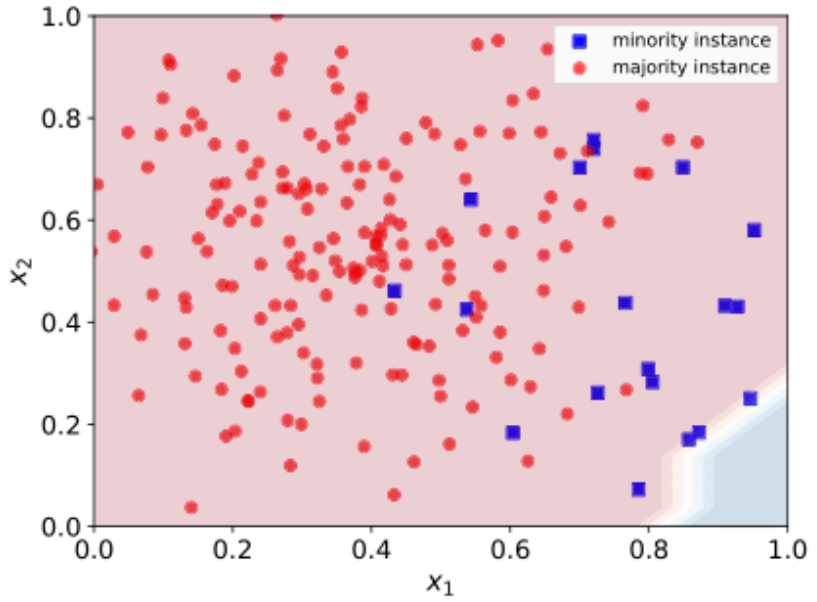


Challenges in Medical Data Collections

- Medical data is often complex, sensitive, incomplete, and expensive to annotate
- Bias:
  - Population → each person is unique or shares similar characteristics with a group of people
  - Institutional → variability in scanners, imaging protocols
- Require combining multiple data types to capture the full patient condition
- Understanding the nature of healthcare data is essential before developing or evaluating machine learning systems
- Other important challenges: Privacy, Ethics, and Regulatory Constraints

Class Imbalances

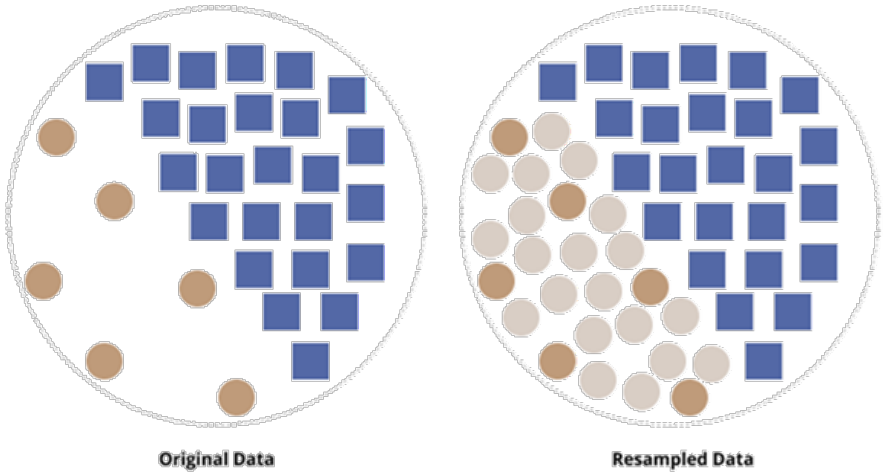
- Imbalanced classes can directly affect the performance of a model
- A model tends to overfit towards the majority class
- A model tends to completely disregard minority classes
- **Clinical data contains bias and class imbalances**

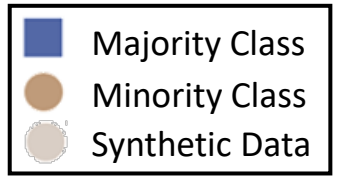


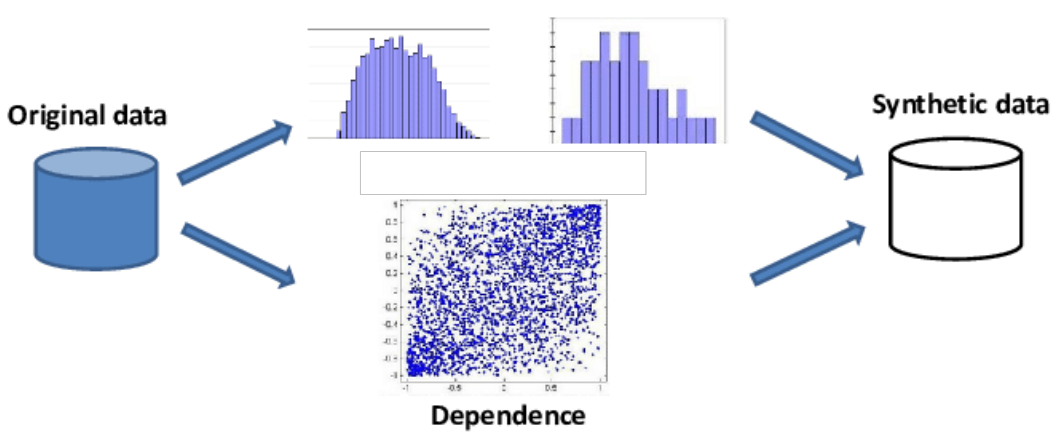
Example: The prevalence of diabetes in Greece is estimated between 9.6% and 12% among adults. If a system predicts that a person is negative to diabetes, it can be over 88% accurate without any clinical investigation

Low-Resource Learning

- Rare diseases affect small patient populations, making data collection inherently difficult
  - ➢ Low-Resource learning techniques are required
- **Active Learning**: a learning algorithm that can interactively query a domain expert to label new instances with the desired outputs
  - ➢ Requires unlabeled datasets
  - ➢ The use of human expert might be expensive, especially in medical applications
- **Low-Cost Solution**: Synthetic data generation

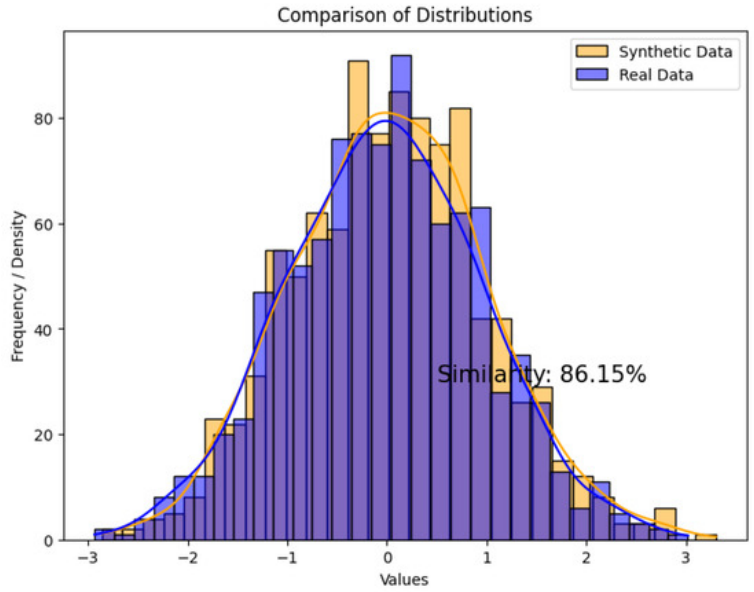


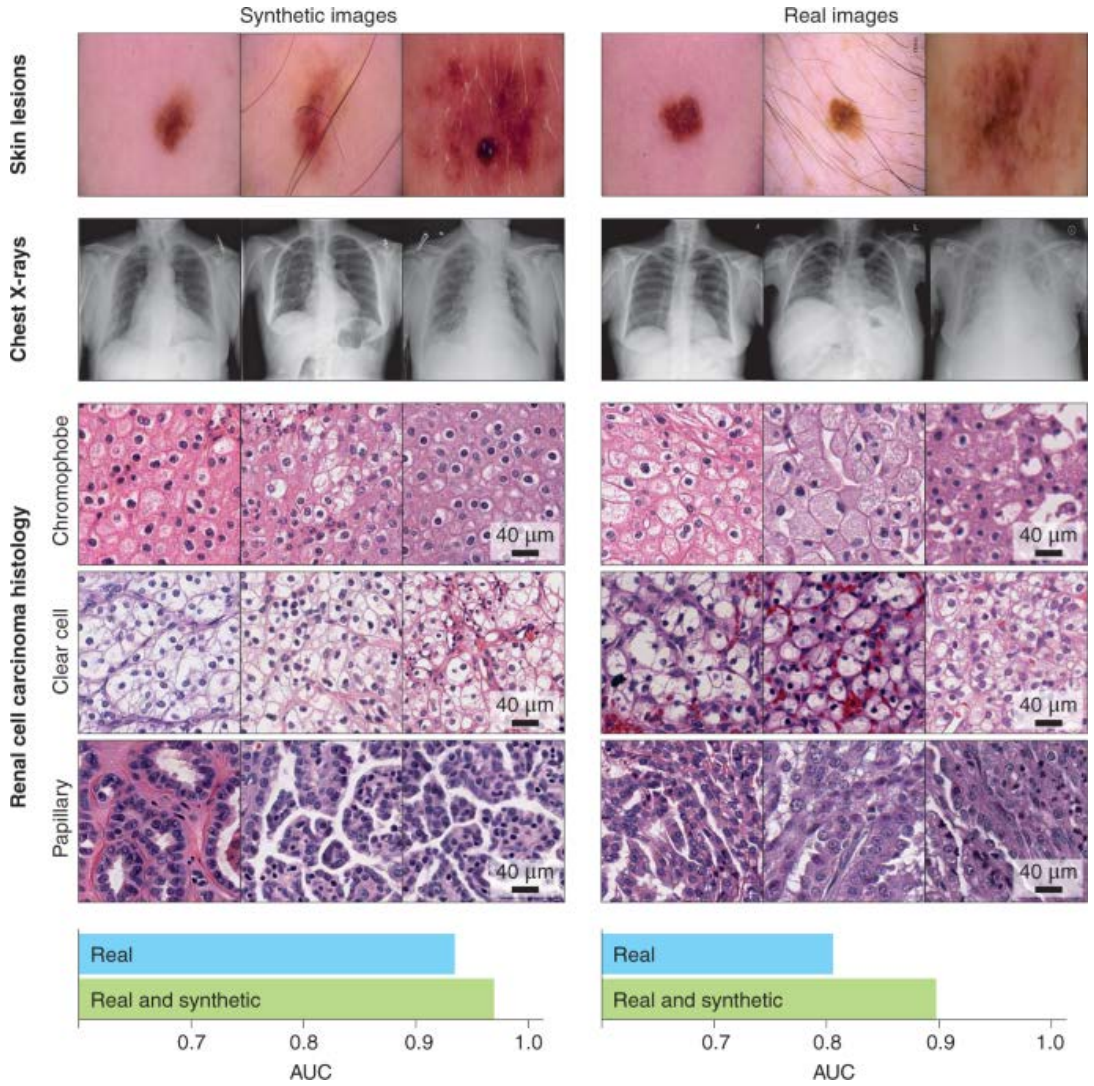




Synthetic Data Generation for Healthcare

- Training an ML classifier (e.g., SVM, Random Forest, etc.) directly using low-resource datasets will likely introduce decision bias
- A better approach:
  1. Train a model to generate synthetic instances of the minority class
  2. Generate enough synthetic instances to remove imbalance
  3. Train the ML classifier

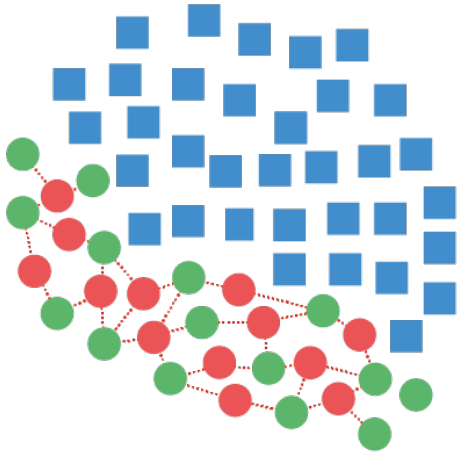




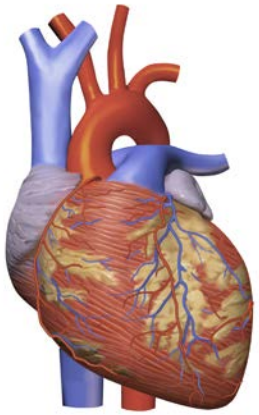
Categories of Synthetic Data Generation Techniques

Collected Data

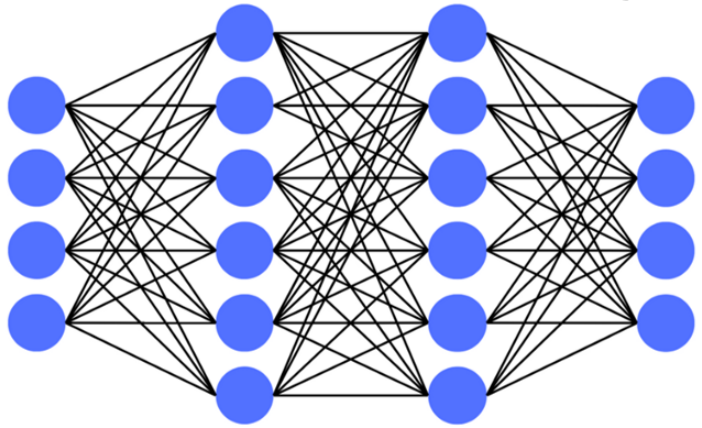
Statistical Methods



Simulation/Modeling

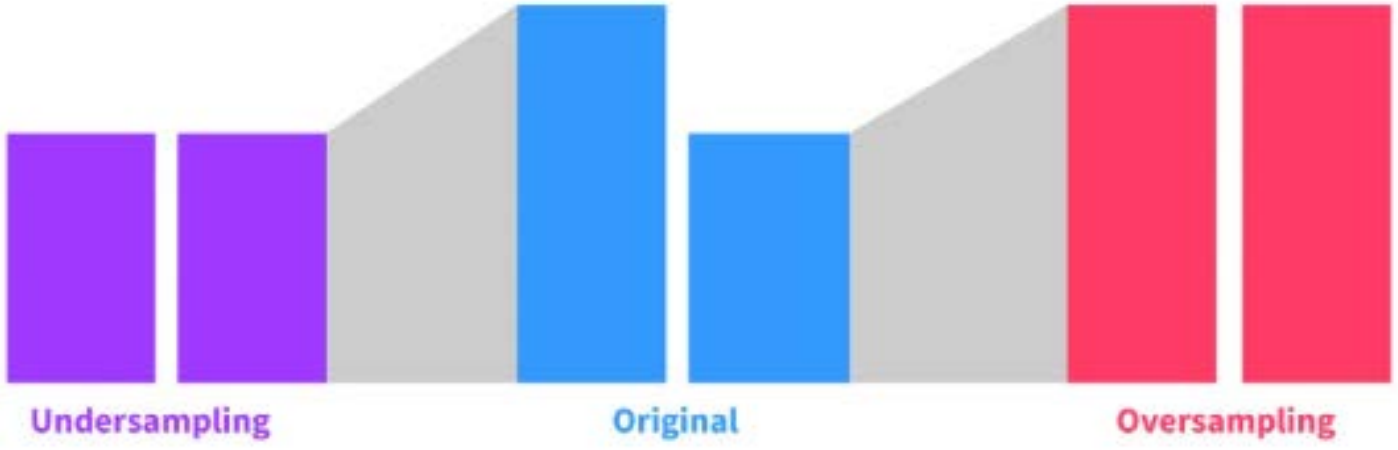


Generative Modeling



Statistical Resampling Techniques

- **Oversampling Methods**: generate synthetic data of the minority class.
  - Random oversampling
  - SMOTE-based Methods
  - ADASYN
- **Undersampling Methods:** exclude samples of the majority class.
  - Random undersampling.
  - Tomek's links.
  - Repeated Edited Nearest Neighbors.

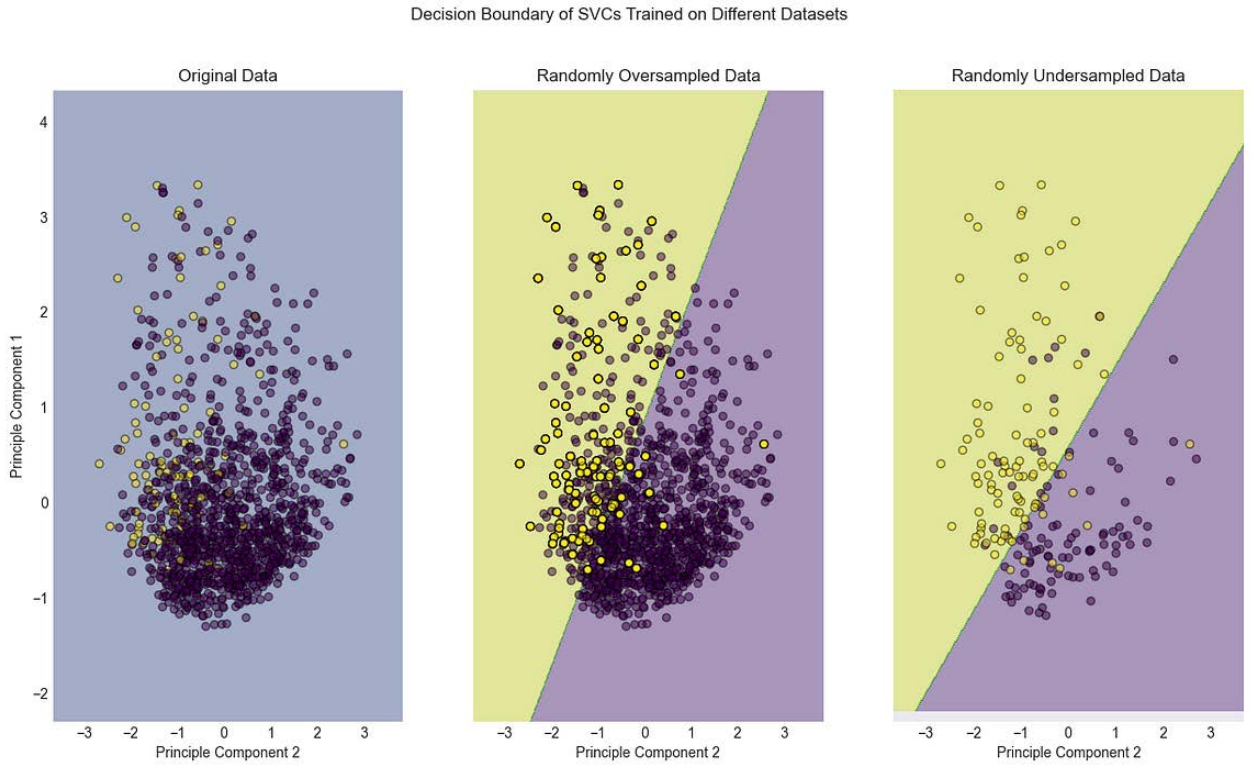


Random Oversampling & Undersampling

- **Oversampling** increases the number of samples in the minority class in a random manner.
  - It is often done by duplicating existing minority-class instances or generating slight variations of them
  - It can improve sensitivity for rare cases, especially in medical diagnosis tasks
  - **Drawback**: it may increase the risk of overfitting, because repeated samples do not add much new information
- **Undersampling** reduces the number of samples in the majority class to balance the dataset.
  - This helps prevent the model from being dominated by the most frequent class during training
  - It can reduce training time and simplify the learning process
  - **Drawback**: it may discard useful instances by removing potentially important majority-class representatives

Example: Classification using SVM

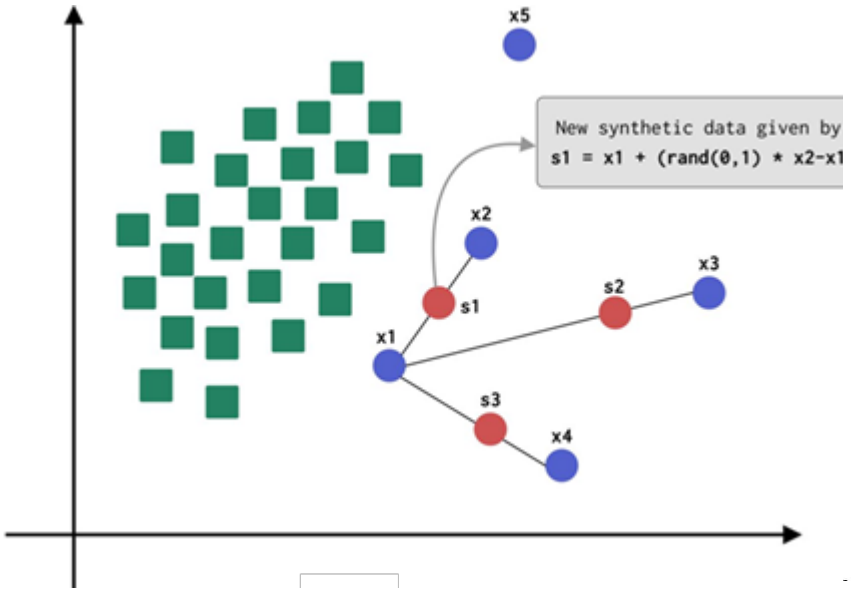
- **SVM Classifier**: its goal is to maximize the margin between the two classes
- The SVM trained on the original dataset seems to underperform → it ignores all the minority samples (yellow samples)
- The SVMs trained on oversampled and undersampled datasets are less biased
- The decision boundaries of SVCs trained on over-sampled and under-sampled dataset differ



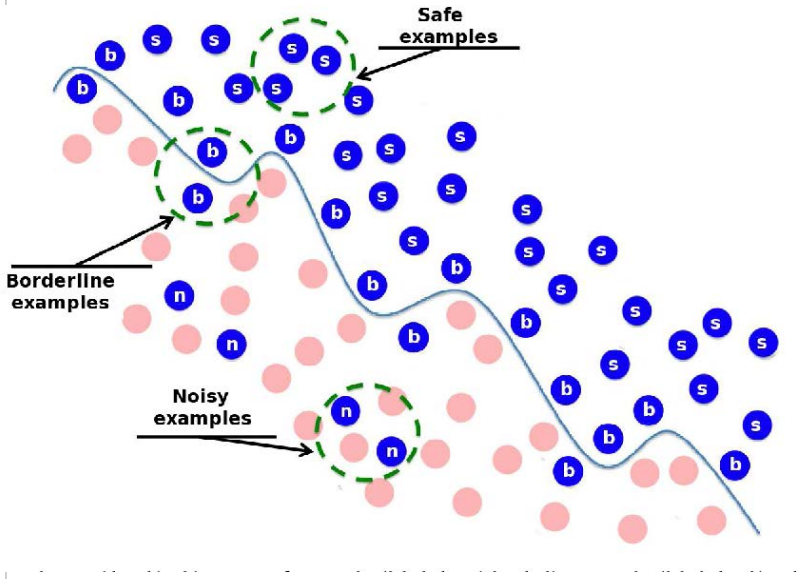
Synthetic Minority Oversampling Technique (SMOTE)

- Creates new data points by **interpolating between existing ones**
- Algorithmic Steps:
  1. Identifies the Minority Class m_i
  2. Finds nearest neighbor for each sample that belongs in m_i
  3. Generates a new sample by interpolating k random neighbors
- Variations:
  - SMOTENC
  - SMOTEN
  - Borderline SMOTE
  - K-Means-SMOTE
  - SVM-SMOTE

SMOTE

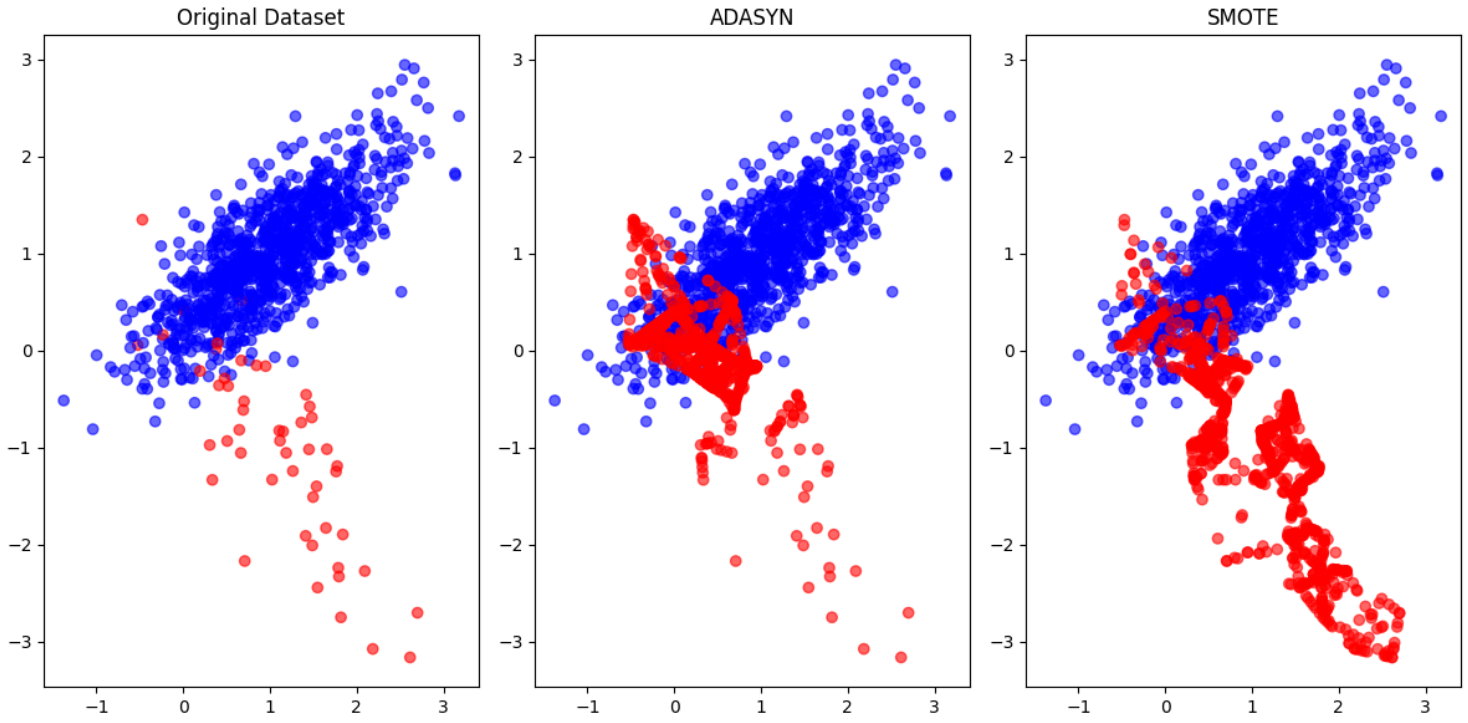


Borderline SMOTE



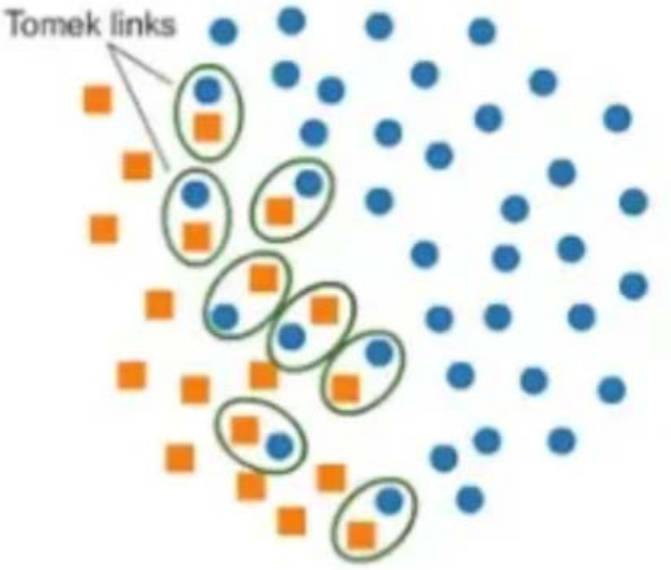
Adaptive Synthetic (ADASYN) Algorithm

- Similar to SMOTE, but it generates different number of samples depending on an estimate of the local distribution of the class to be oversampled
- **Core idea**: uses a weighted distribution to focus on generating synthetic samples in **areas where the minority class is hardest to learn** → where minority class is sparsely represented



Tomek's Links

- Attempts to **clean decision boundaries** in classification problems by removing Tomek links
- A pair of instances (x_1, x_2) is a Tomek link if:
  1. x_1 belongs to class c_1, while x_2 belongs to c_2
  2. The distance $dist(x_1, x_2)$, which is the distance between x_1 and x_2 is minimal
  3. There exists no example x_3 such that $dist(x_1, x_3) < dist(x_1, x_2)\ AND\ dist(x_2, x_3) < dist(x_1, x_2)$
- Tomek's links algorithm is usually combined with an upsampling technique, such as SMOTE



Limitations of Statistical Techniques for High-Dimensional Data

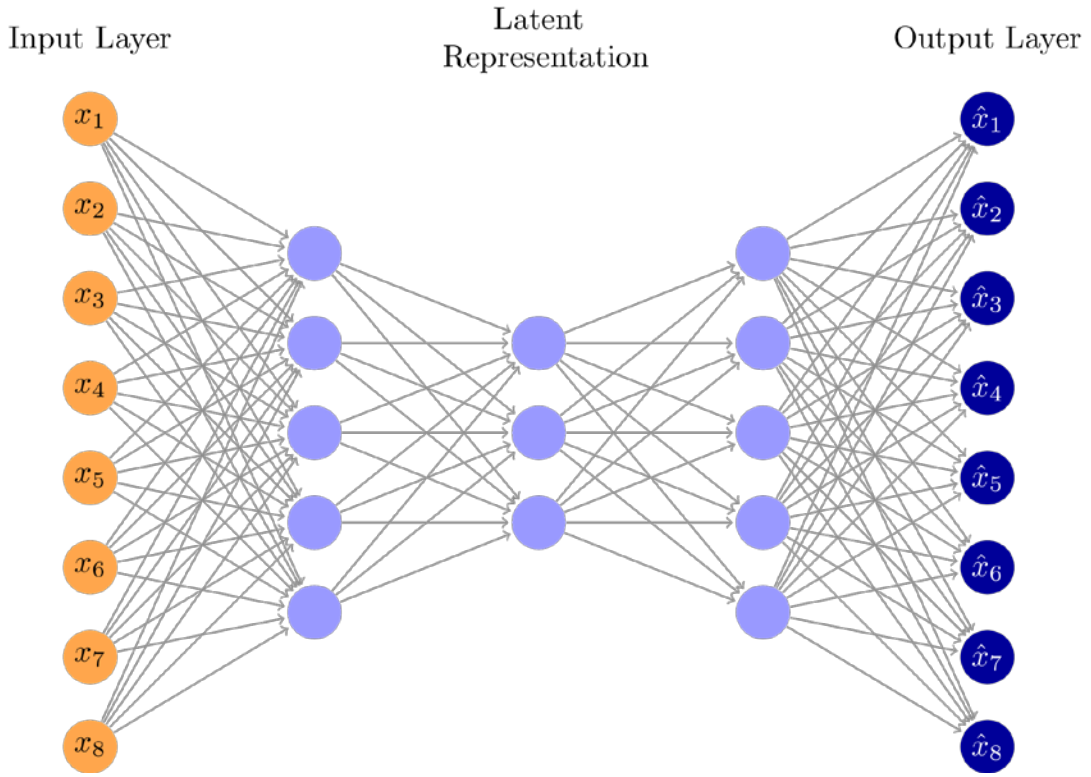
- Statistical resampling methods are mainly designed for simpler feature spaces (such as Tabular data)
- Their effectiveness decreases when data become high-dimensional, complex, and highly correlated
  - Not ideal for medical datasets
  - Medical images, physiological signals, clinical text, and multimodal records contain rich structures with high-correlated features that are difficult to generate with simple resampling
- Moreover, oversampling or interpolation in high-dimensional spaces may generate unrealistic or low-quality samples
  - For instance, images might become blurred
  - Time-series might become inconsistent

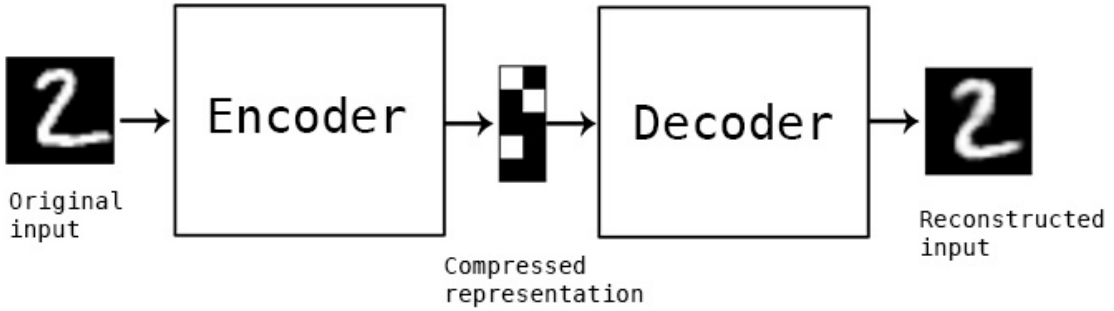
Signals, Text, and Visual Data Need Generative Models

- Signals, text, and visual data are high-dimensional and contain complex underlying patterns
- These data types include important spatial, temporal, and semantic relationships that simple statistical methods cannot model well
- **Moreover**, medical signals such as ECG or EEG depend on time-dependent structure and waveform consistency
- **Generative Models**: ML models that learn the underlying distribution of the data, so they can produce new synthetic samples that follow similar patterns, structures, and variations as the original dataset
- Methods:
  - Probabilistic Modeling
  - Autoencoders
  - Generative Adversarial Networks (GANs)
  - Diffusion Models

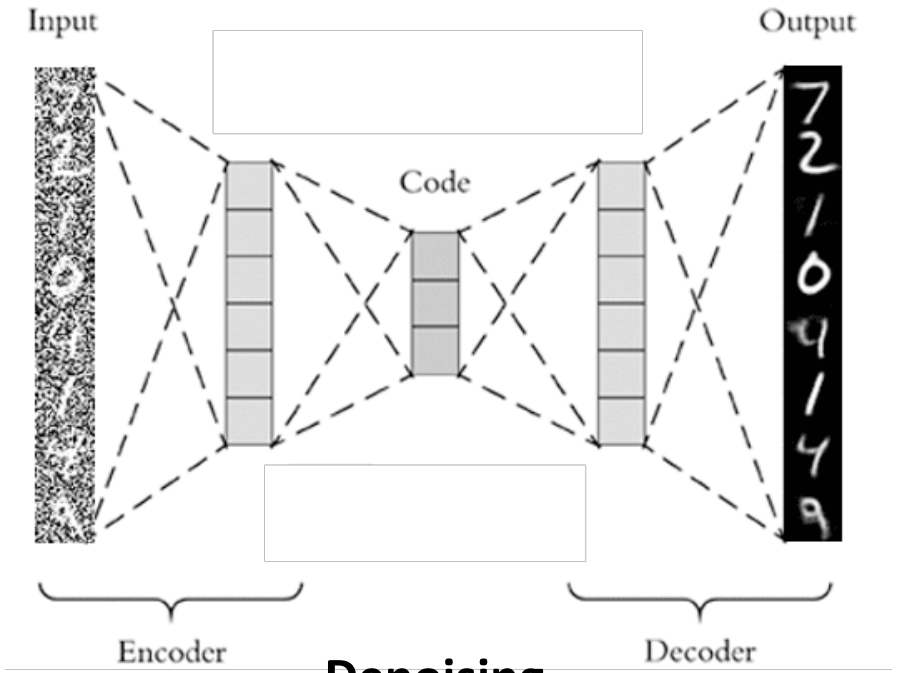
Autoencoders

- Autoencoders are a special type of neural networks that compress data into latent representations (e.g., small vectors) and then learn to reconstruct it to closely match the original input
- They are composed of 2 components:
  a) **Encoder:** captures important features by reducing dimensionality.
  b) **Decoder**: rebuilds the data from this compressed representation.

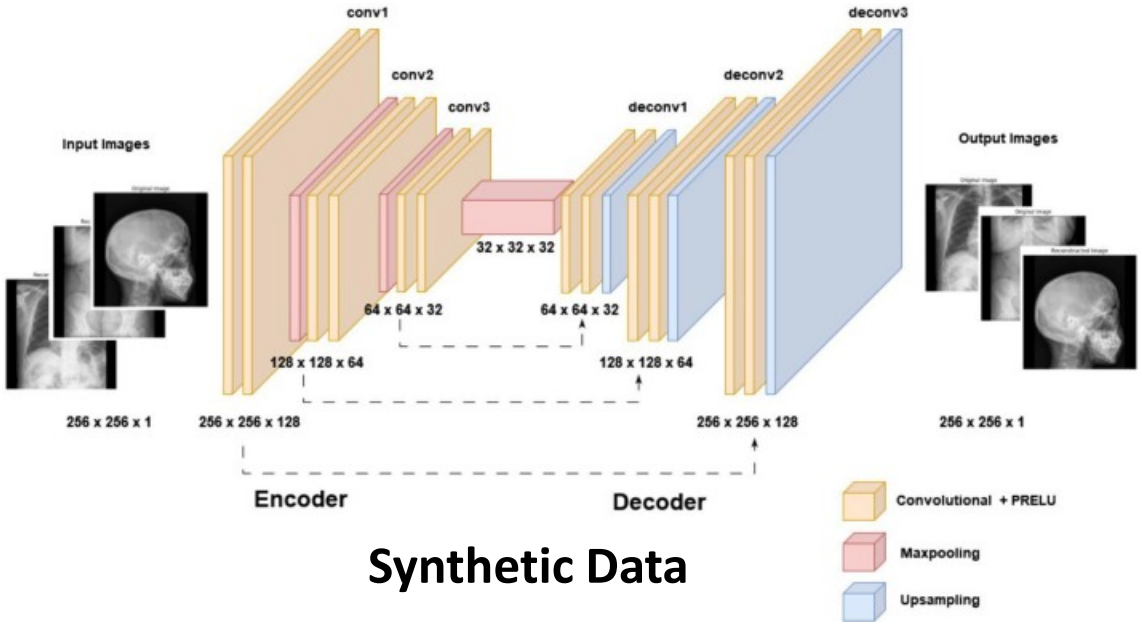




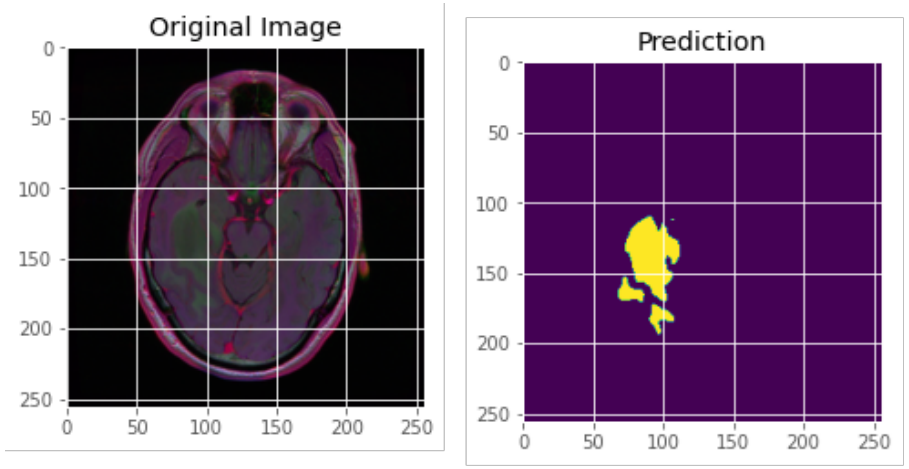
Examples of Autoencoder Applications in Medical Data



Denoising



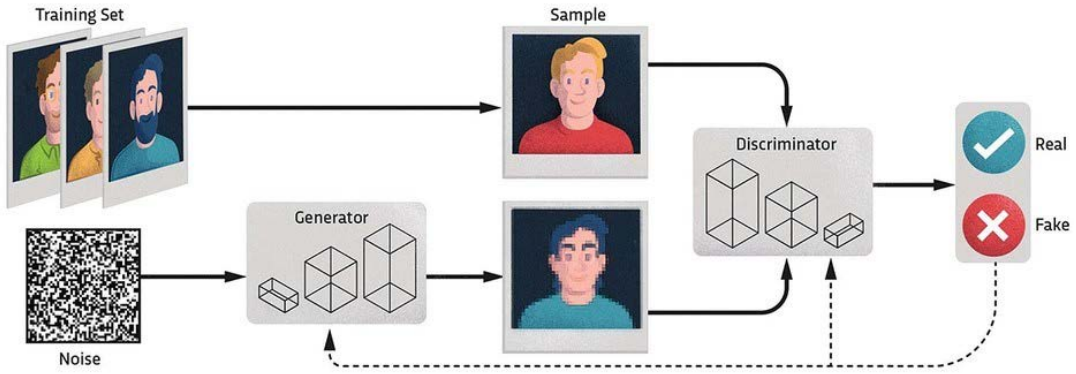
Synthetic Data

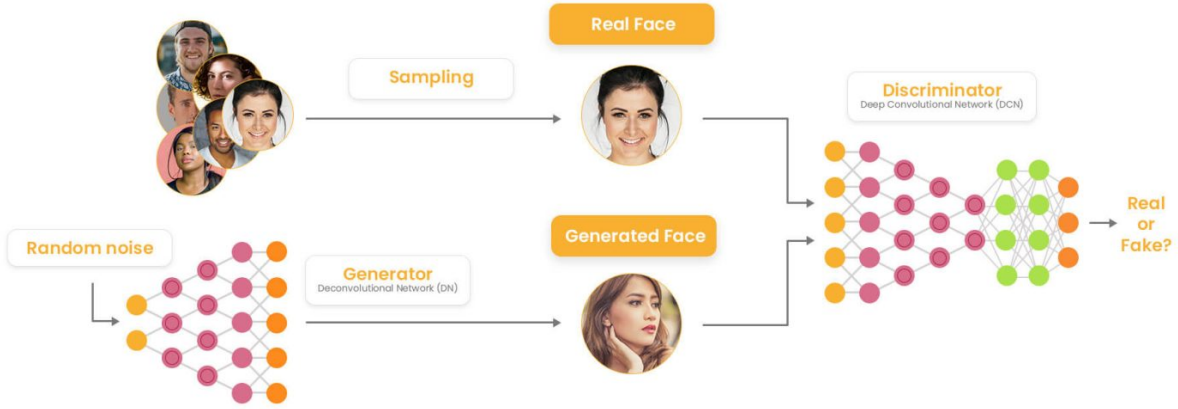


Semantic Segmentation

Generative Adversarial Networks (GANs)

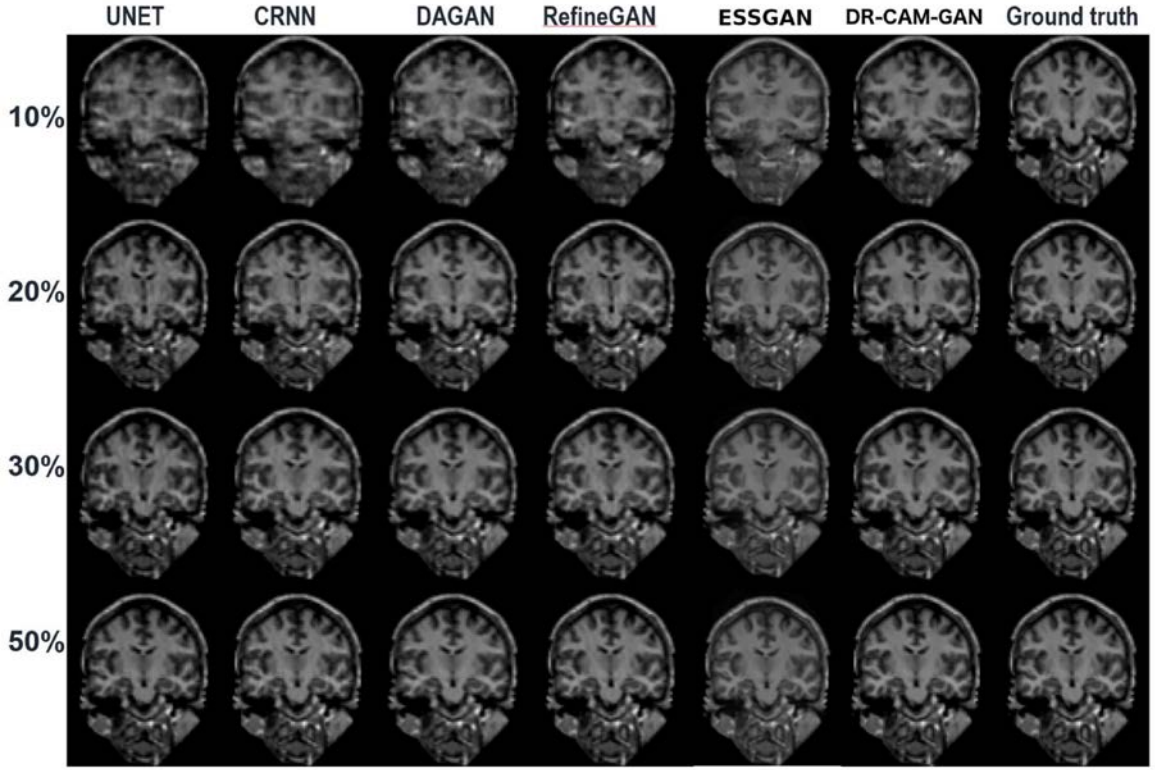
- Inspired by Adversarial Learning in Game Theory.
- Idea: When 2 opponents compete, they are constantly getting better by finding ways to outsmart their opponent.
- Goals: Synthetic Text, Video, Audio generation.
- Involves the use of 2 networks:
  - **Generator**: Generates synthetic data
  - **Discriminator**: Classifies whether an instance is real or synthetic



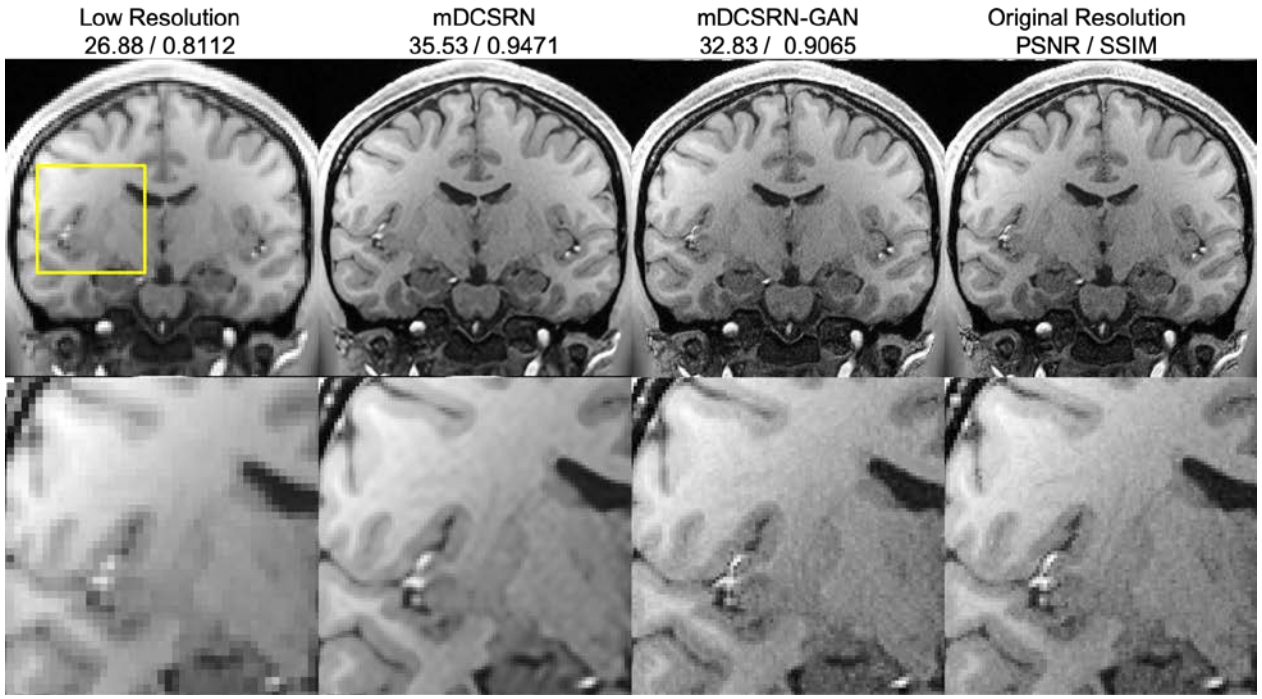


Examples of GAN Applications in Medical Data

Super Resolution

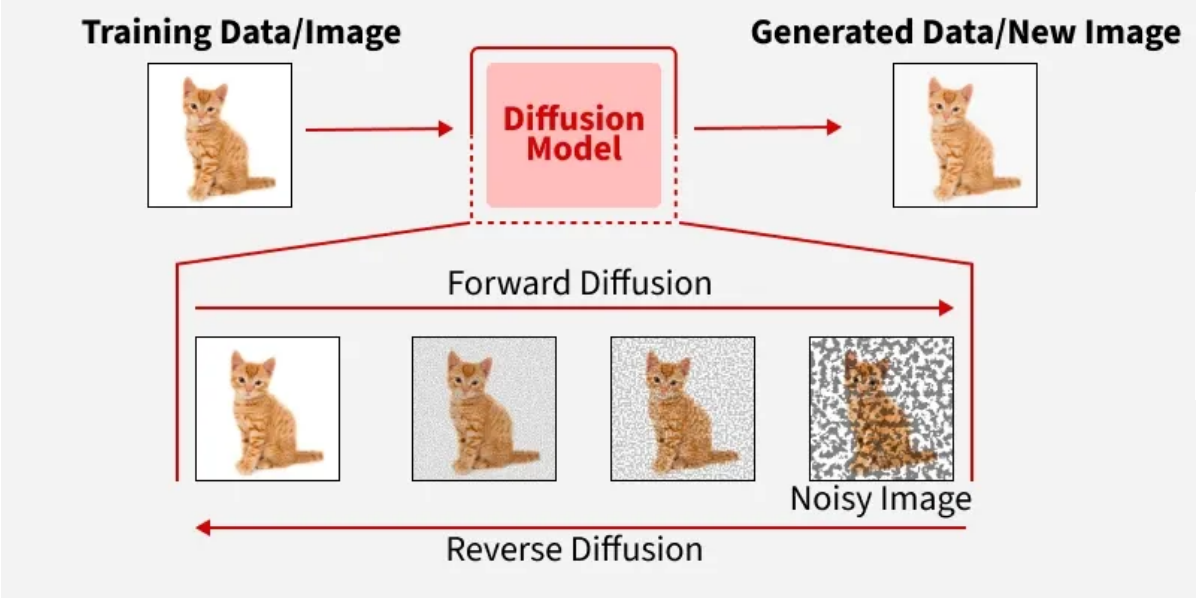


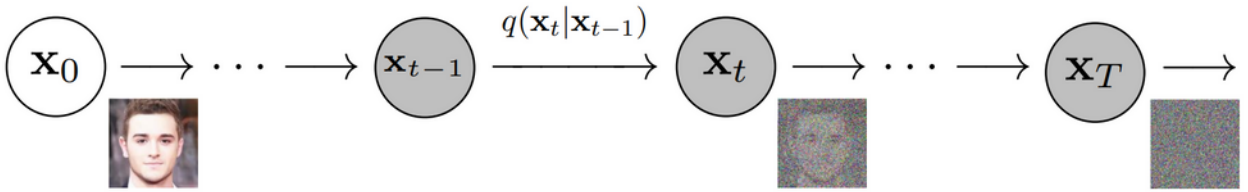
Super Resolution



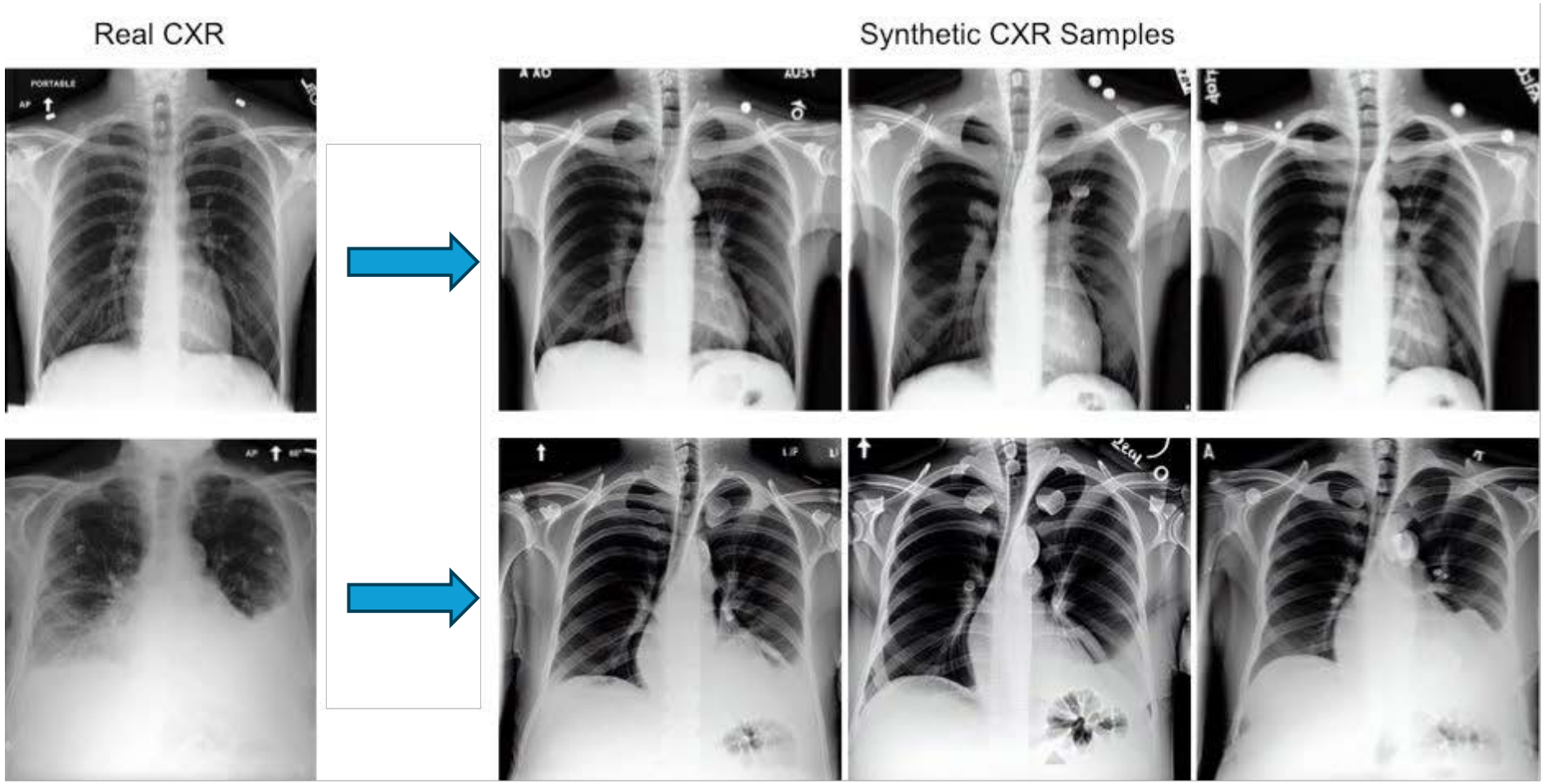
Diffusion Models

- Diffusion models create new data like images, audio or even video by starting with random noise and gradually turning it into something meaningful
- They work by simulating a diffusion process where data is slowly corrupted by noise during training and then learning to reverse this process (using a neural network) step by step
- By doing so the model learns how to generate high quality samples from scratch





Examples of Diffusion Applications in Medical Data



Diffusion vs GANs

- **GANs** generate data through an adversarial game between a generator and a discriminator, whereas **diffusion models** generate data by starting from random noise and gradually denoising it through many iterative steps
- A key advantage of diffusion models is that they are generally more stable to train and are often associated with better mode coverage than GANs, which can suffer from mode collapse and fail to represent parts of the true data distribution
- A major disadvantage of diffusion models is that they are usually computationally more expensive, because they often require tens to hundreds of denoising steps to generate a single sample
- For this reason, diffusion models are often viewed as theoretically and empirically stronger in sample quality and distribution coverage, while GANs are often preferred when fast generation is more important

Conclusions

- Medical labeled data are sensitive, complex and often expensive to acquire
- ML algorithms require large amounts of data, to be robust and efficient
- Synthetic data generation is a good low-cost technique that addresses privacy concerns, biases and class imbalances
- Statistical approaches can be used for data generation, but only in low-dimensionality domains
- Generative approaches can efficiently learn the underlying data distribution of a dataset and generate realistic data

PHAROS
THE GREEK AI FACTORY

Data Generation Techniques for Medical Data

Ioannis Vlahavas, Prof.
Vasileios Kochliaridis, PhD Candidate
Zoi Katsantoni, MSc Graduate

Department of Informatics, AUTH