Image Analysis with Artificial Intelligence

Eleftherios Trivizakis, PhD

Computational BioMedicine Laboratory (CBML)
Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)

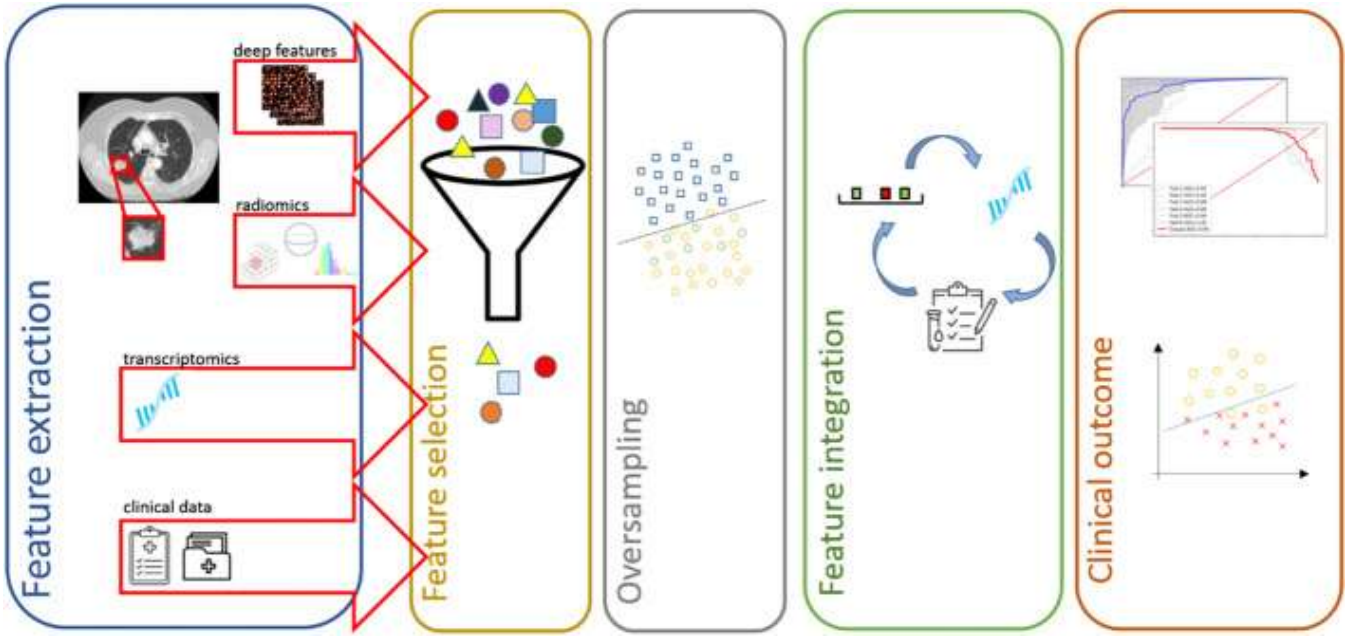
Contact: trivizakis@hmu.gr

Image Analysis

- Machine Learning Analysis
- Data Science Pitfalls
- Deep Learning: Supervised, Unsupervised, and Self-supervised Learning
- Visual Representation Learning
- Transformers Beyond Text
- Explainable AI (XAI)

Machine Learning-based Data Analysis – Step by Step

1. Raw **data acquisition**
2. Feature **extraction**
3. Feature **selection**
4. Balancing the **biases**
5. (Multi-Modal) Feature **integration**
6. Machine learning **classification**

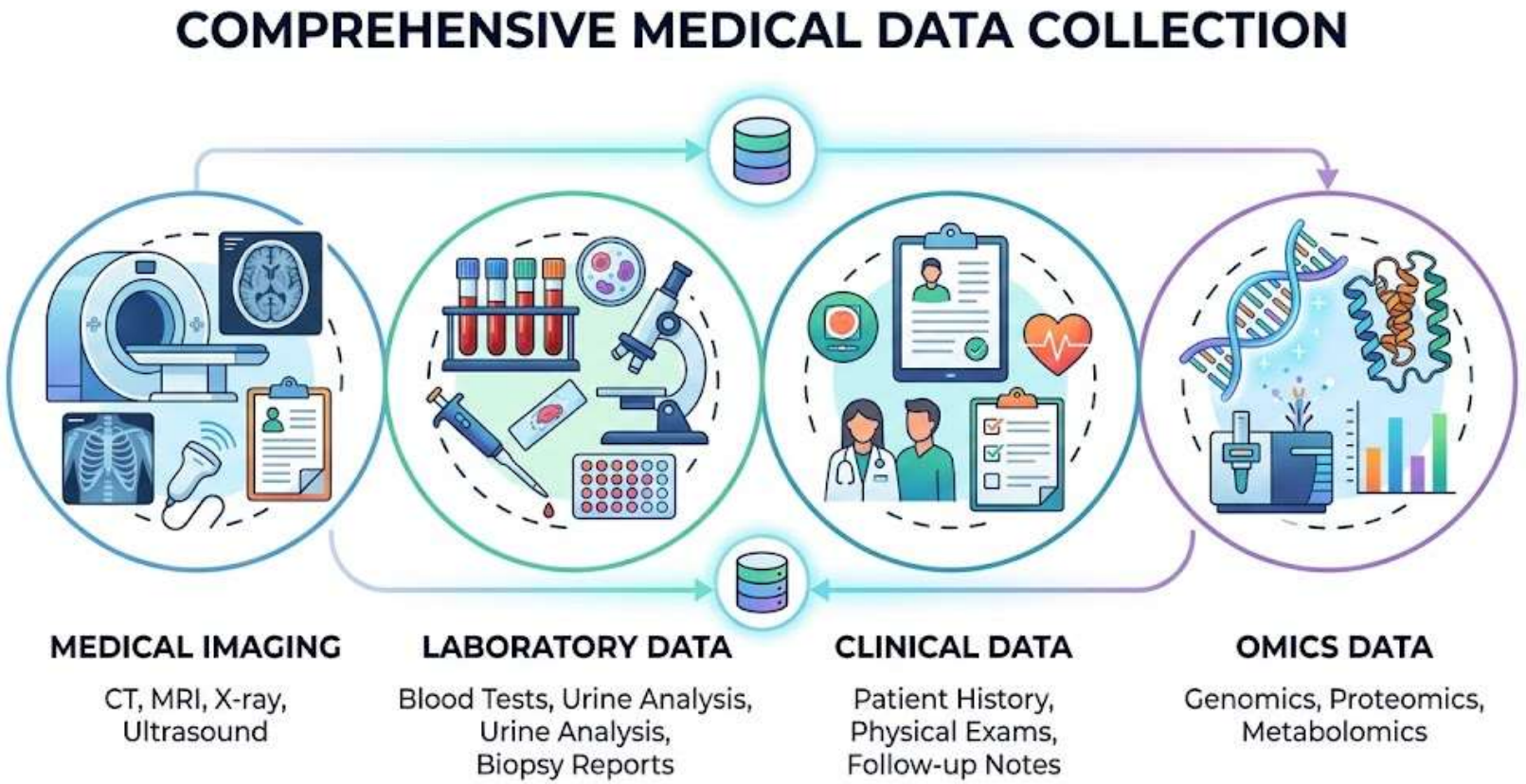


Trivizakis, E., Koutroumpa, N. M., Souglakos, J., Karantanas, A., Zervakis, M., & Marias, K. (2023). Radiotranscriptomics of non-small cell lung carcinoma for assessing high-level clinical outcomes using a machine learning-derived multi-modal signature. *BioMedical Engineering OnLine*, *22*(1), 125.

1. Data Collection

Raw Data

- Medical images
- Segmentation masks
- Clinical data
- Laboratory data
- -omics

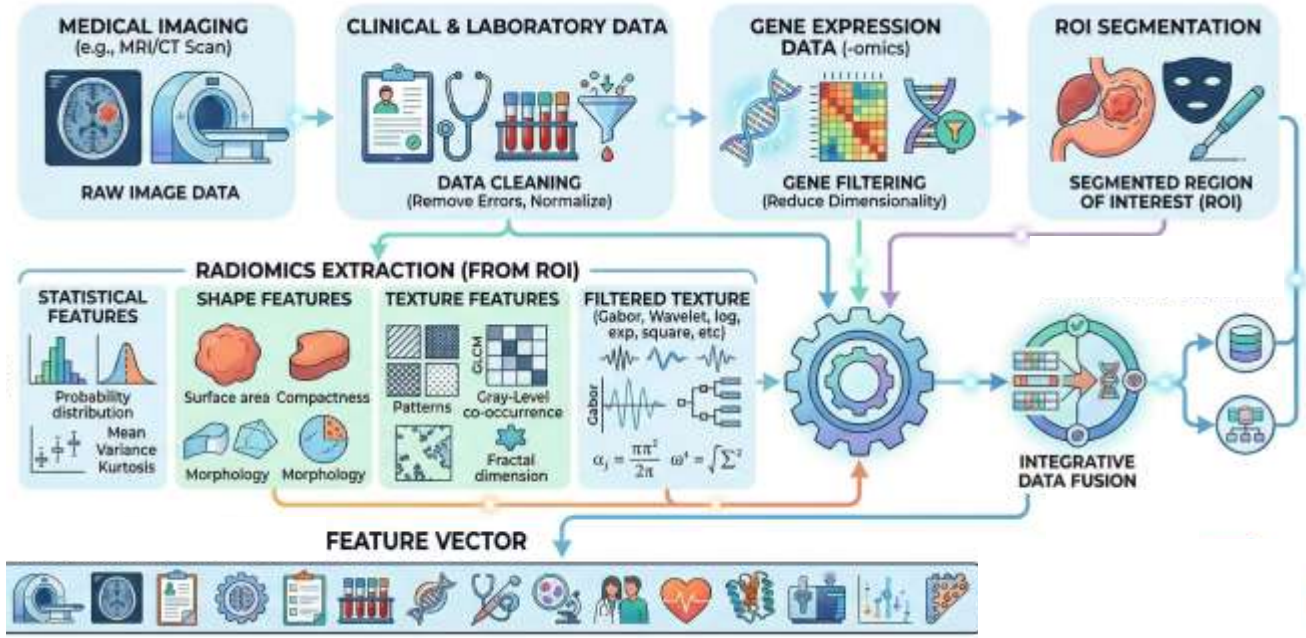


Usually, GBs to TBs of data!

2. Transforming raw data to meaningful features

Feature extraction

- Statistical
- Shape
- Texture
- Filtered Texture (Gabor, wavelet, log, exp,square, etc)
- Radiomics over ROI

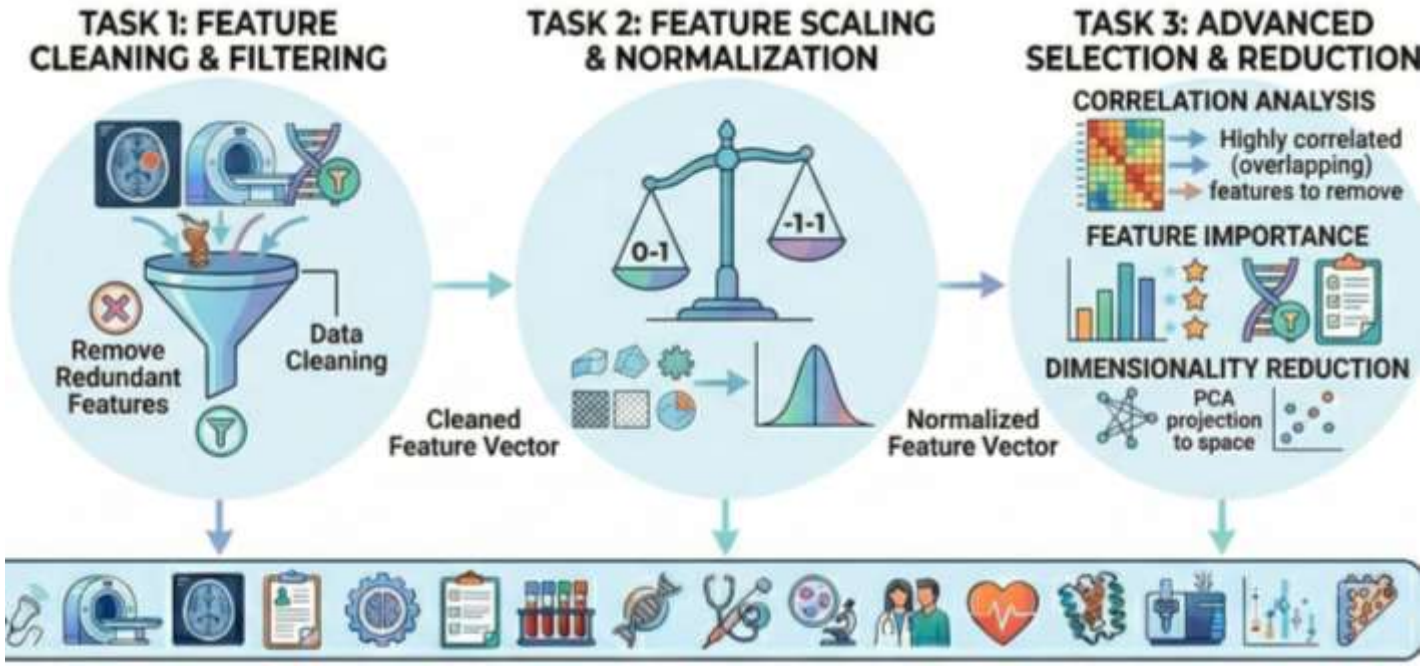


Introducing: the "curse of dimensionality"!!!
-Thousands of features!

3. Find the best features for the examined task

Feature Selection

- ANOVA
- LASSO Ridge
- LASSO Regression
- Minimum Redundancy Maximum Relevance (mRMR)
- Variance Threshold

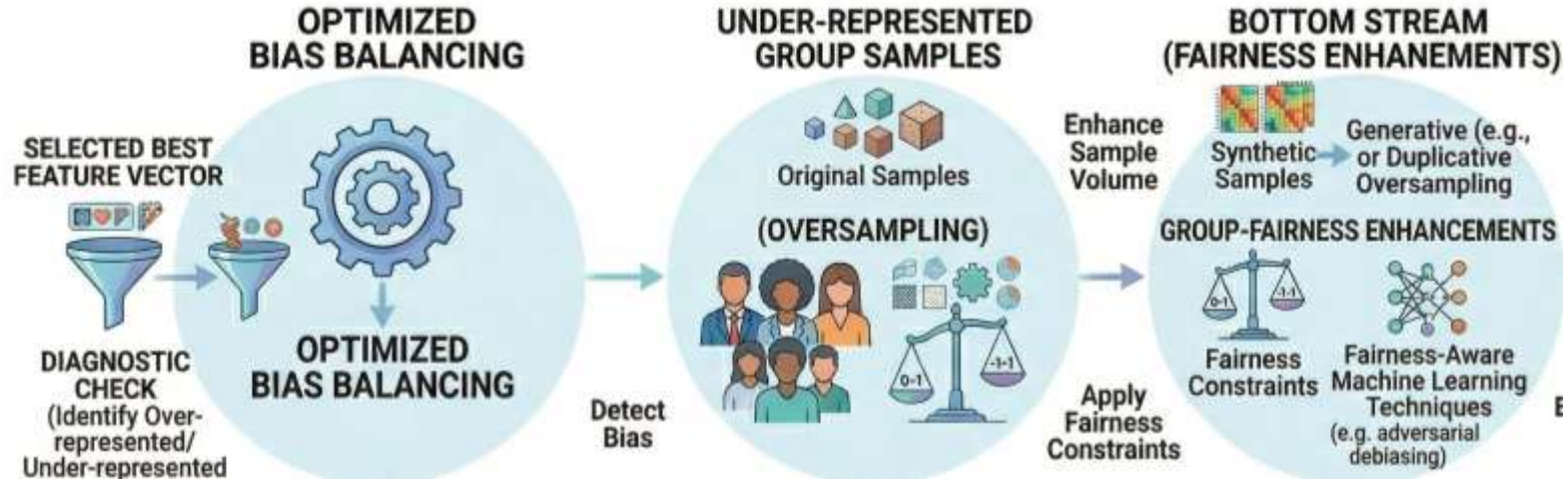


Lifting the "curse of dimensionality"
-Reducing Overfitting

4. Fixing the imbalances!

Balancing the biases

- Detect biases
- Oversampling
- Generative sampling
- Fairness enhancements

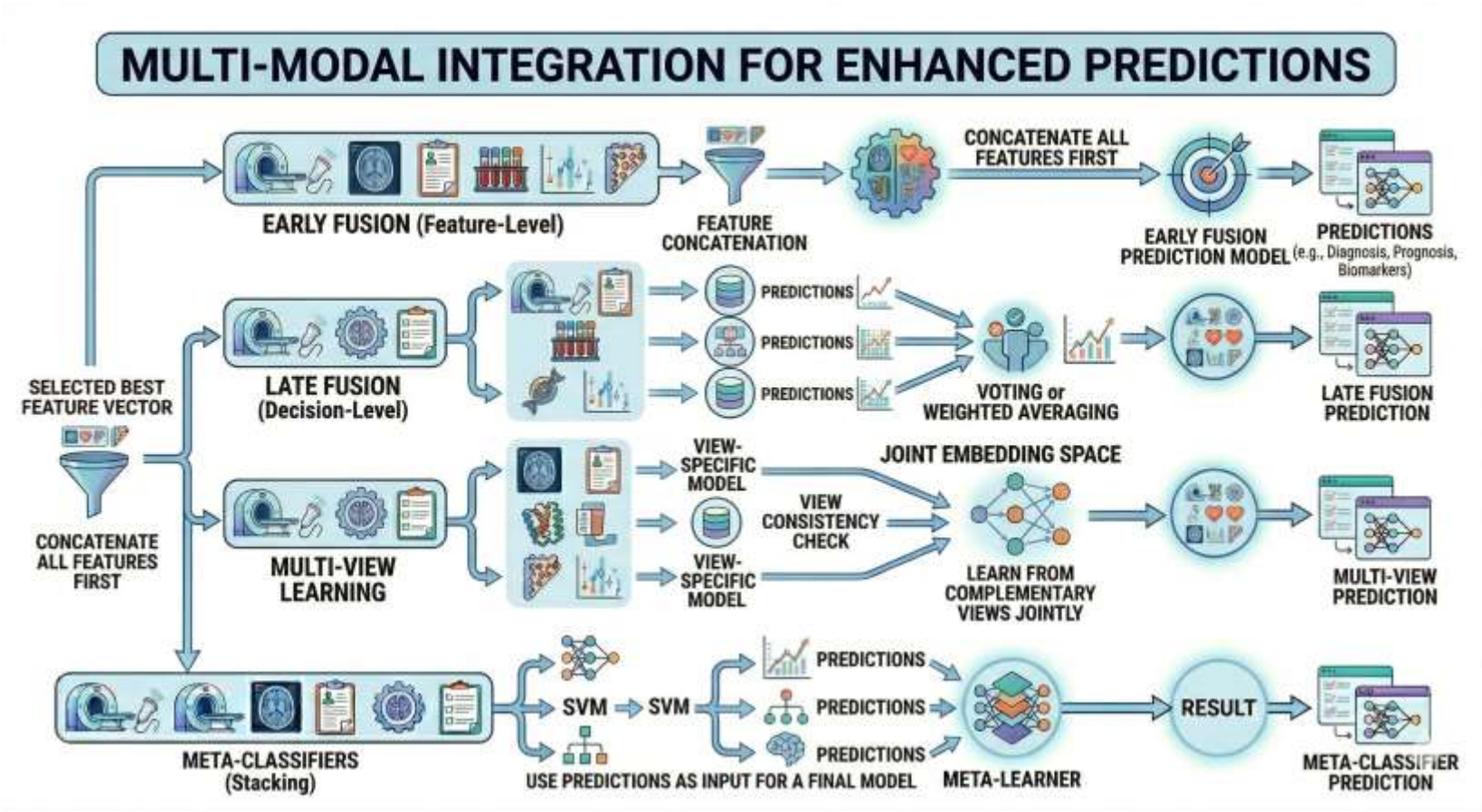


Improves unprivileged group convergence!

5. Combine diverse types of feature vectors

Multi-Modal Integration

- Early fusion
  - Feature-level
- Late fusion
  - Decision-level
- Multi-view
  - Distinct data sources
- Meta-classifiers
  - Multi-classifiers

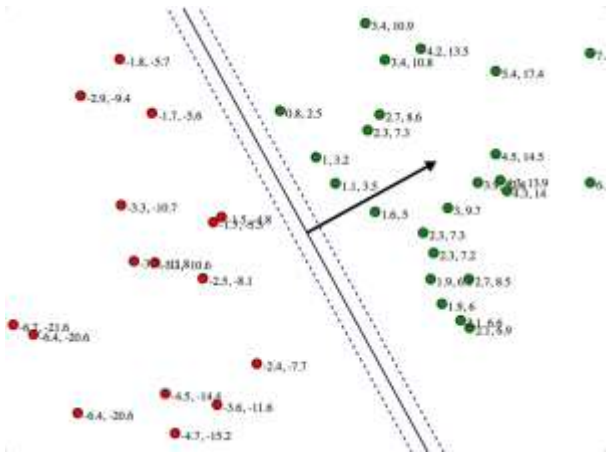


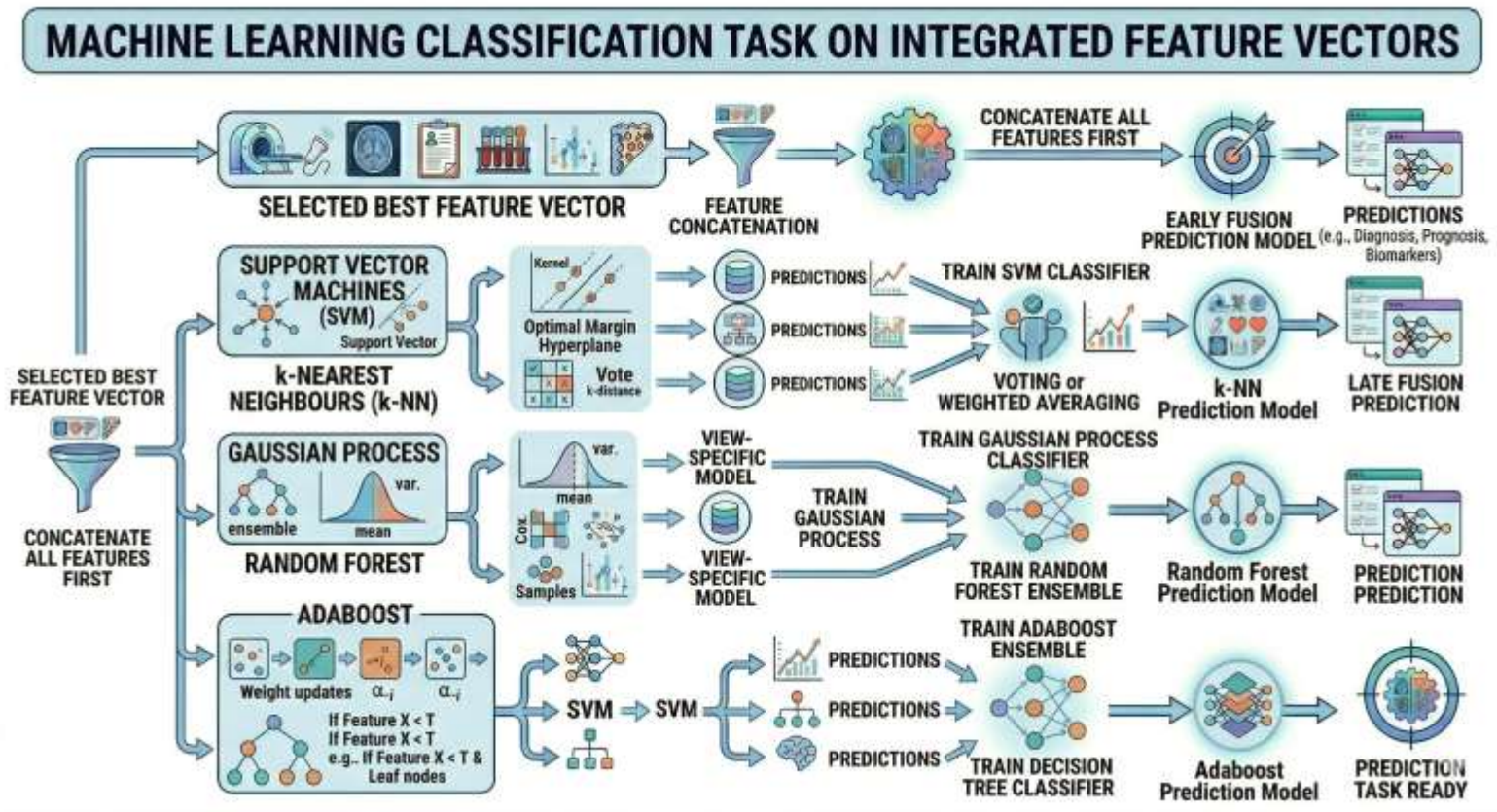
Use all data sources for improved predictions!

6. Differentiate among class distributions – Features to decisions

Machine learning classification

- Support Vector Machines
- k-Nearest Neighbors
- Gaussian Process
- Random Forest
- Adaboost
- Decision Tree





The machine adapts to the input data!

Data Science Pitfalls

Fundamental Measurement

- Data and outcome **leakage**, often caused by poor split hygiene or inclusion of predictive "future" information.

Dataset Composition

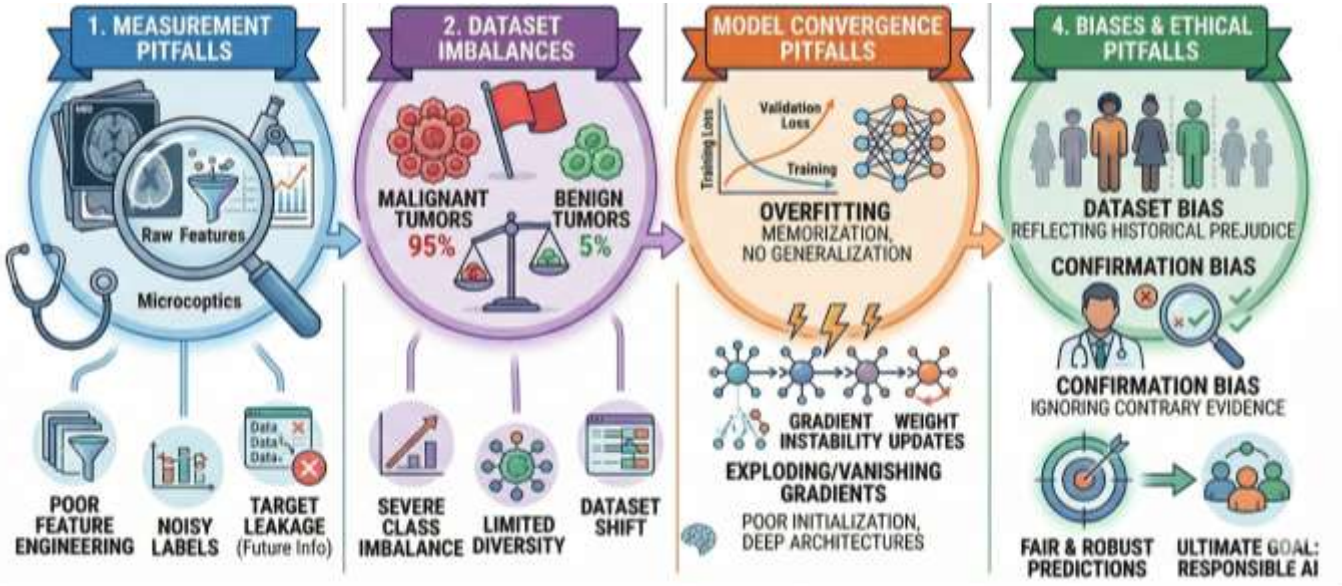
- Severe class **imbalance** and **shifts**, leading to models that perform well on training data but poorly in the real world.

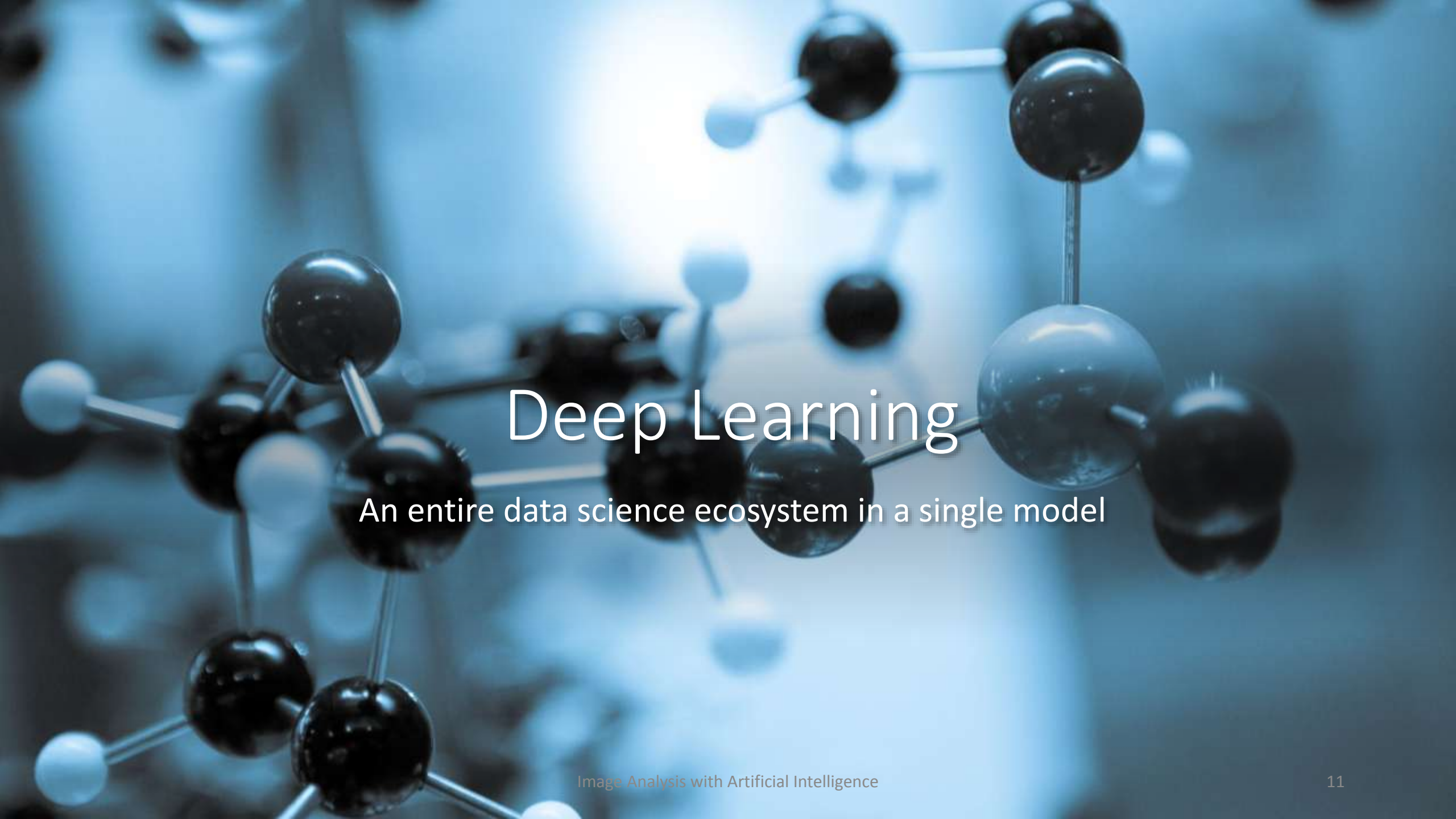
Model Training & Evaluation

- The classic **overfitting** and the less common **underfitting**, along with failure to use a **"true unseen set"**.

Bias and Ethics

- Dataset **biases**, resulting in unfair models that reflect historical prejudice.

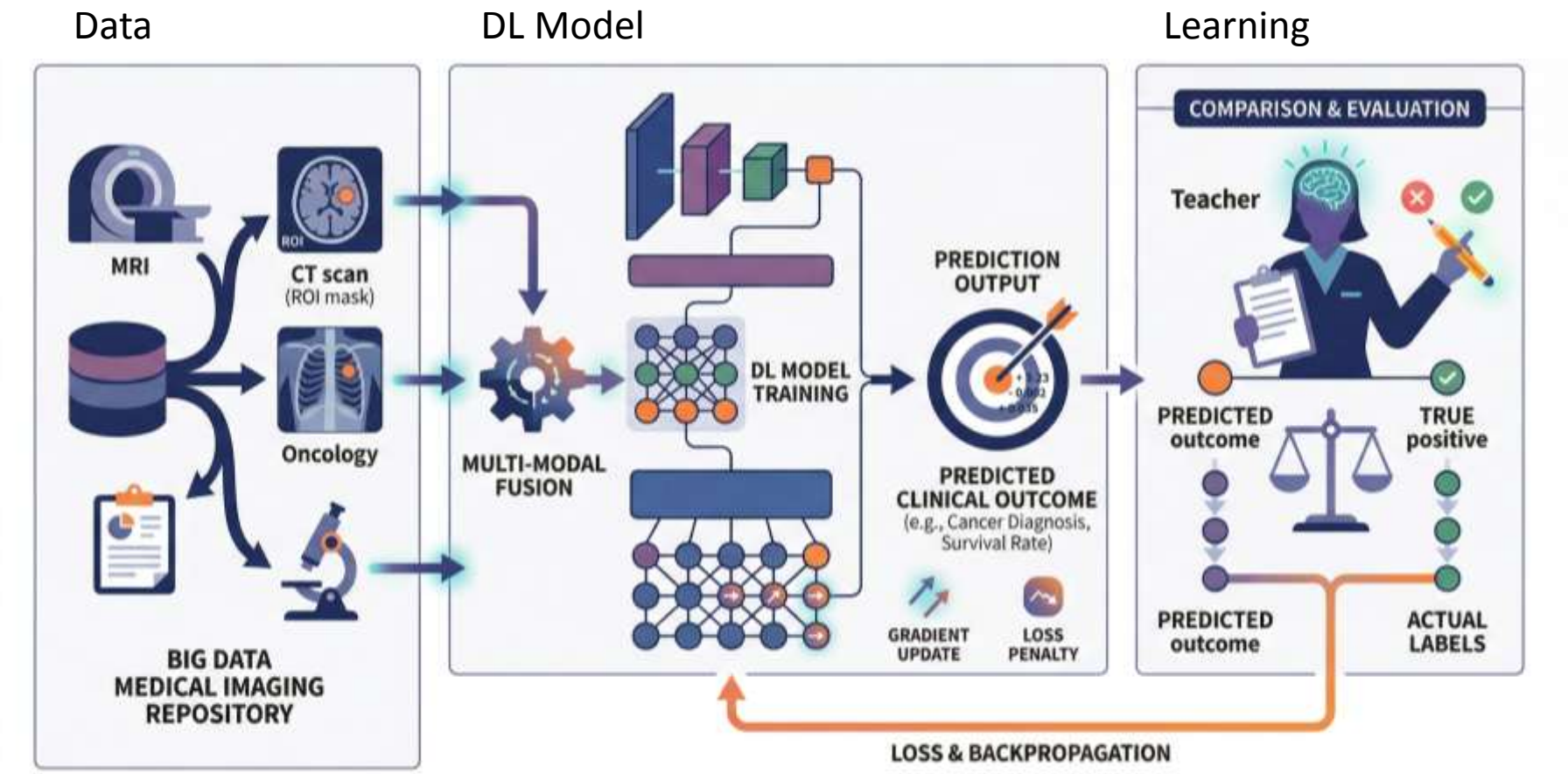




Deep Learning

An entire data science ecosystem in a single model

DL Models Learn the Entire Data Analysis Stack from Raw Data



Supervised Learning

Learn from labeled data!

- **Definition**
  - Training a model on a labeled dataset (match x-->y)
- **Deep Learning Context**
  - The primary driver of DL breakthroughs (e.g., in medical image classification)
- **Key Concept**
  - The loss function compares the model's prediction y' to the real label y.
- **Examples**
  - Image Classification: predict "Cancer" or "Benign" tumor
  - Object Detection: Identify the bounding boxes of objects in an image
  - Region of Interest Segmentation: Pixel-based classification ("Background"Vs. "Tumor" pixels)

Unsupervised Learning

Learn from unstructured data!

- **Definition**
  - Finding hidden patterns or structures in an unlabeled dataset (only input x)
- **Deep Learning Context**
  - Leverages **massive** unlabeled data, often acting as a **foundational step** prior to supervised methods
- **Key Concept**
  - The model must **reconstruct or represent** the input data without explicit guidance
- **Examples:**
  - Clustering: Grouping similar data points (e.g. identifying different tumor types)
  - Representation Learning: Compressing data to essential features (e.g. using Autoencoders).
  - Generative Models: Creating new data resembling the input distribution (e.g. GANs, Diffusion Models).

Self-supervised Learning

Learning with generated labeled data... by itself!?!

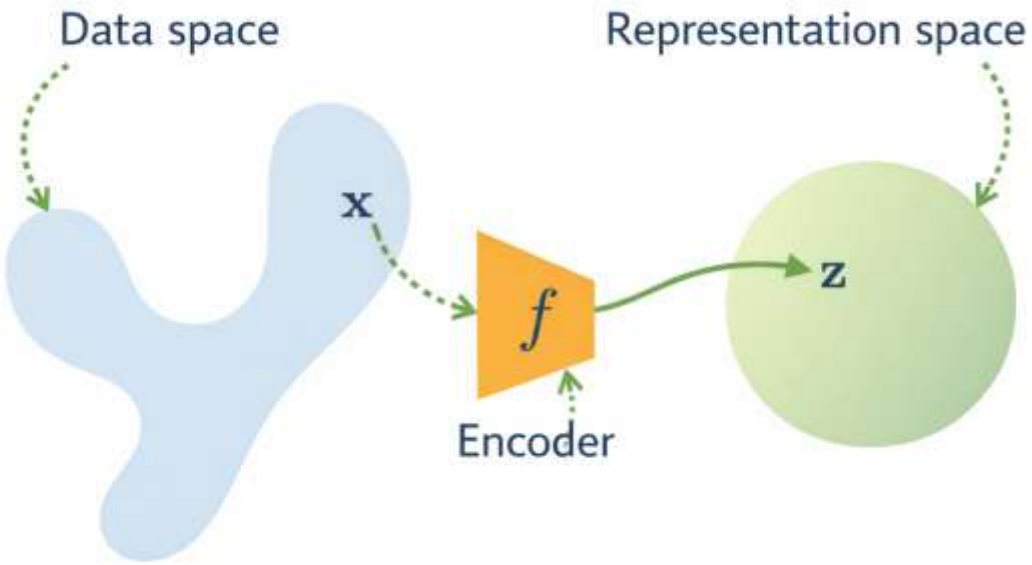
- **Definition**
  - Developing a model using automatically generated labels from the **data itself (e.g. match noisy x->x)**
- **Deep Learning Context**
  - The engine behind modern AI (e.g., transformers for language and vision).
- **Key Concept**
  - "Pre-training" on a huge unlabeled corpus to learn powerful, general features before fine-tuning for a specific task.
- **How it works**
  - Create a pretext task by hiding part of the data and asking the model to predict it (e.g., masking words).

Visual Representation Learning and SimCLR

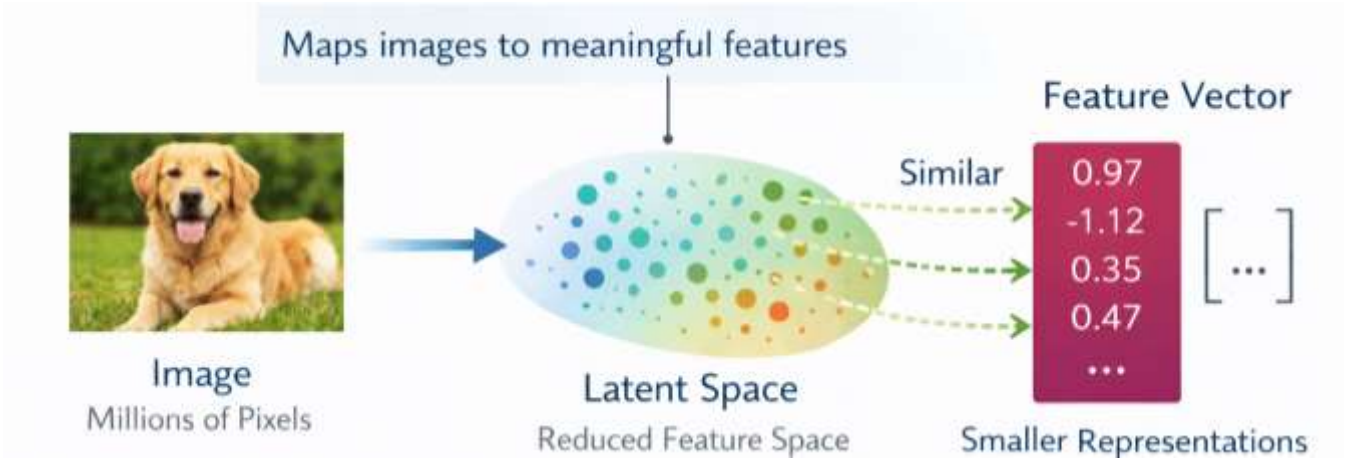
"What if a model could learn to see without ever being told what it's looking at?"

Grigoris Kalliatakis, PhD

Computational BioMedicine Laboratory (CBML)
Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)

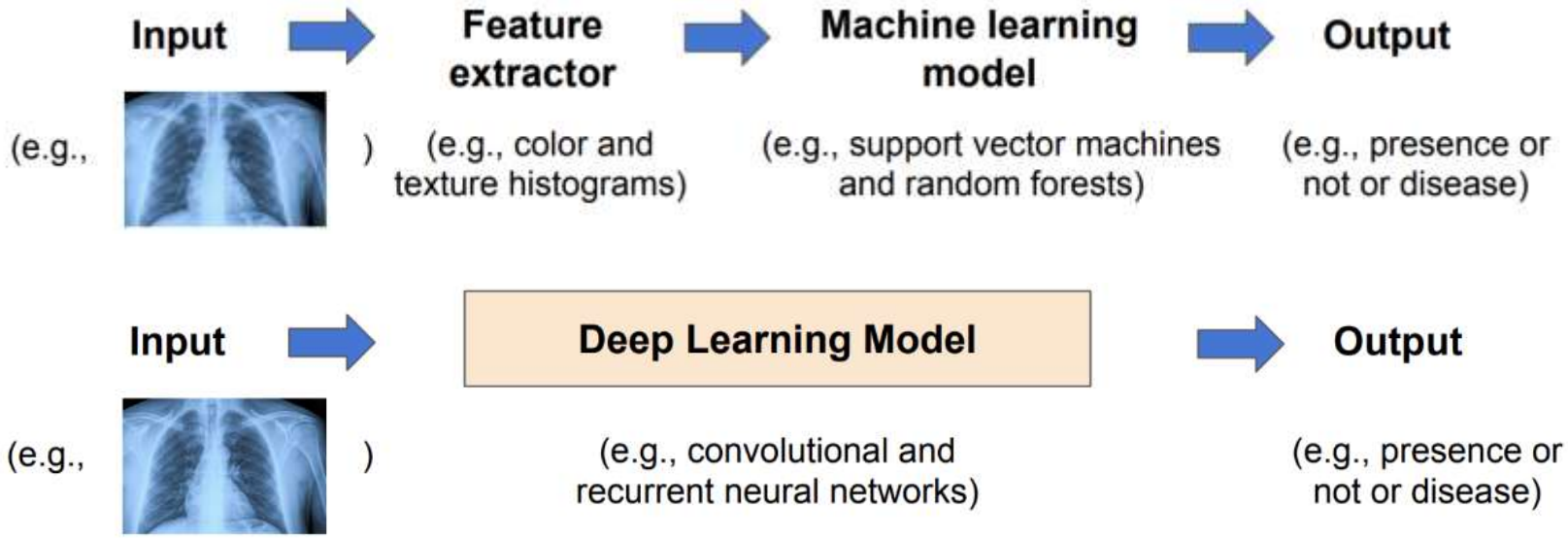


From pixels to representations



- Representation learning maps images to meaningful features.
- It creates a latent space, which is a compact description of an image's structure.
- It reduces dimensionality, converting millions of pixels into a smaller feature space.
- It enables generalization since similar images have similar representations.

Why representation learning?

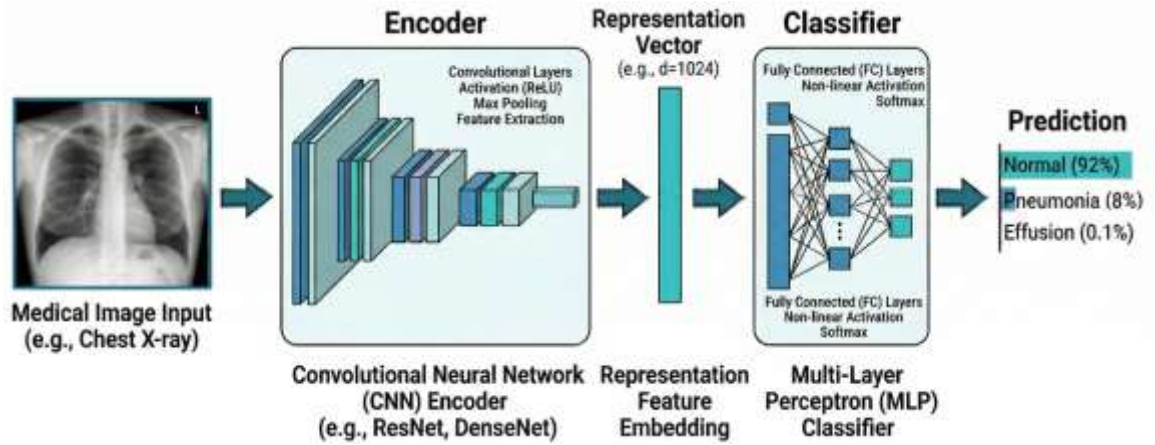


- Traditional computer vision relied on handcrafted features
- Deep Learning learns representations directly from data
- End-to-end learning improves performance and scalability

What is a representation?

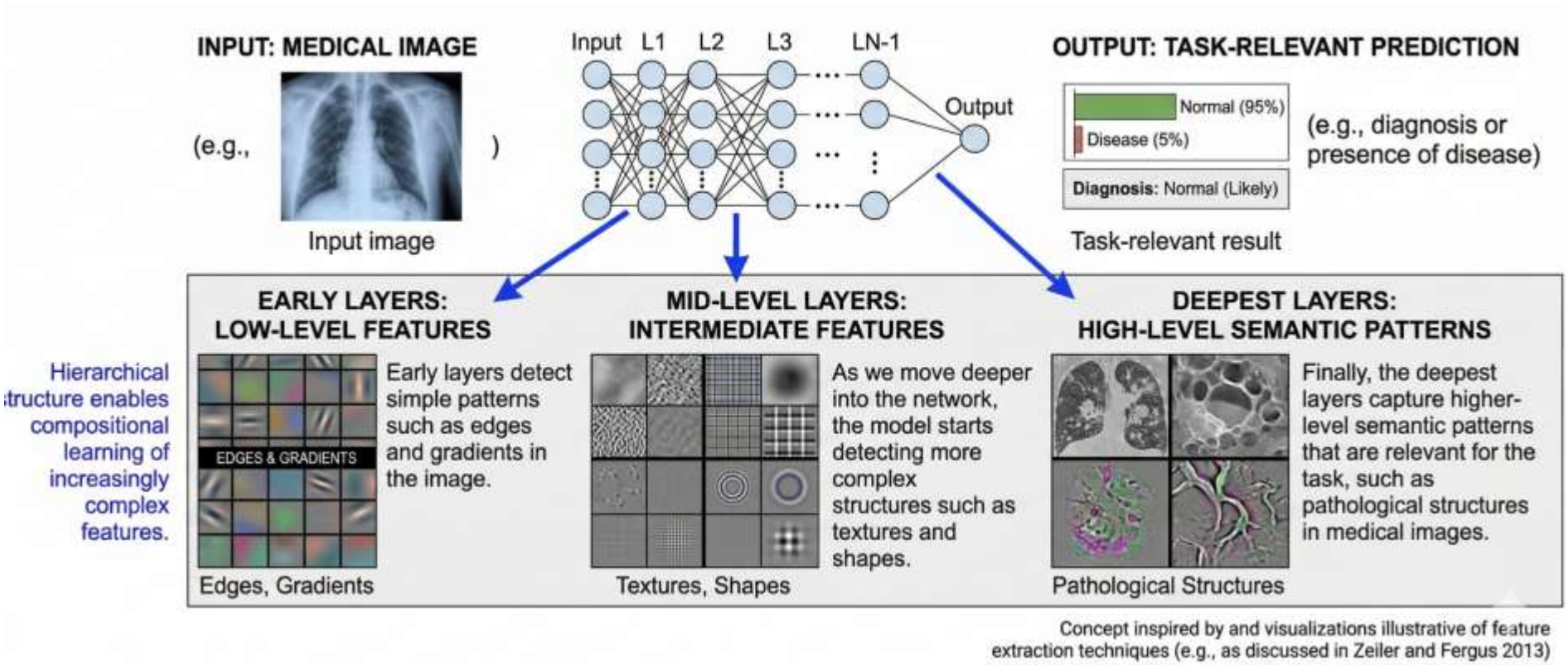
Terminology

Representations are called **"embeddings"**, and the networks that produce them are **"encoders"** or **"embedding models"**

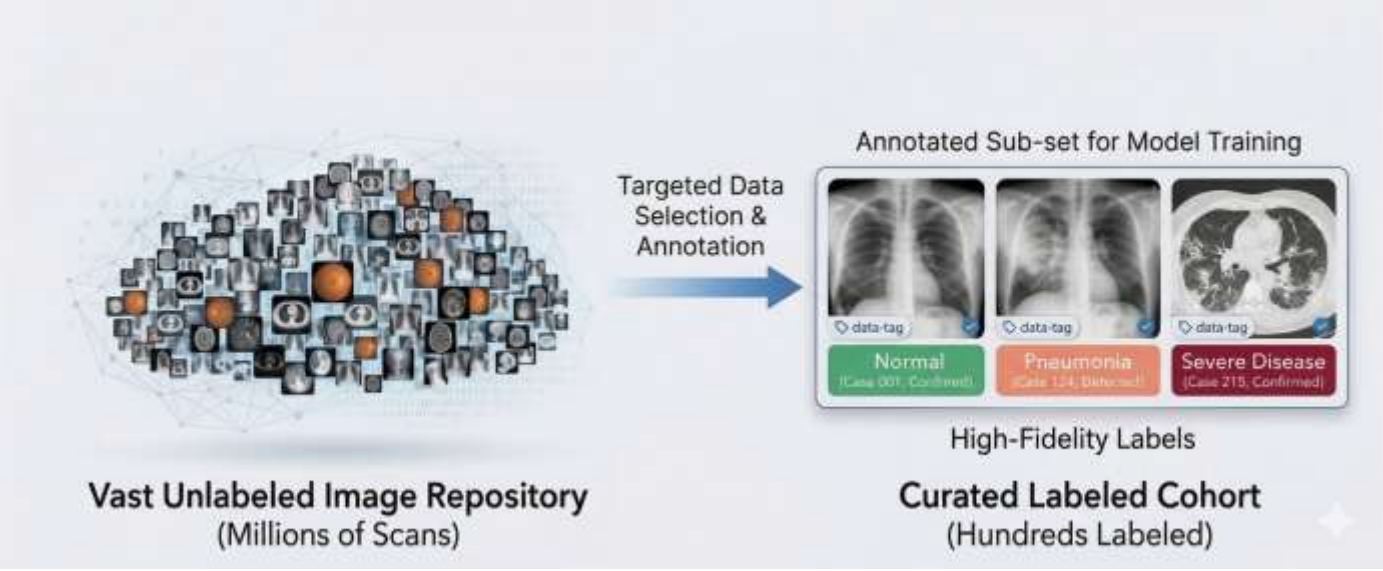


- **Compact encoding**: image summarised as a vector
- **Task-relevant**: captures disease patterns, morphology
- **Dimensionality reduction**: millions of pixels → smaller feature space
- **Generalisation**: similar images → similar embeddings

Deep Networks Learn Hierarchical Representations

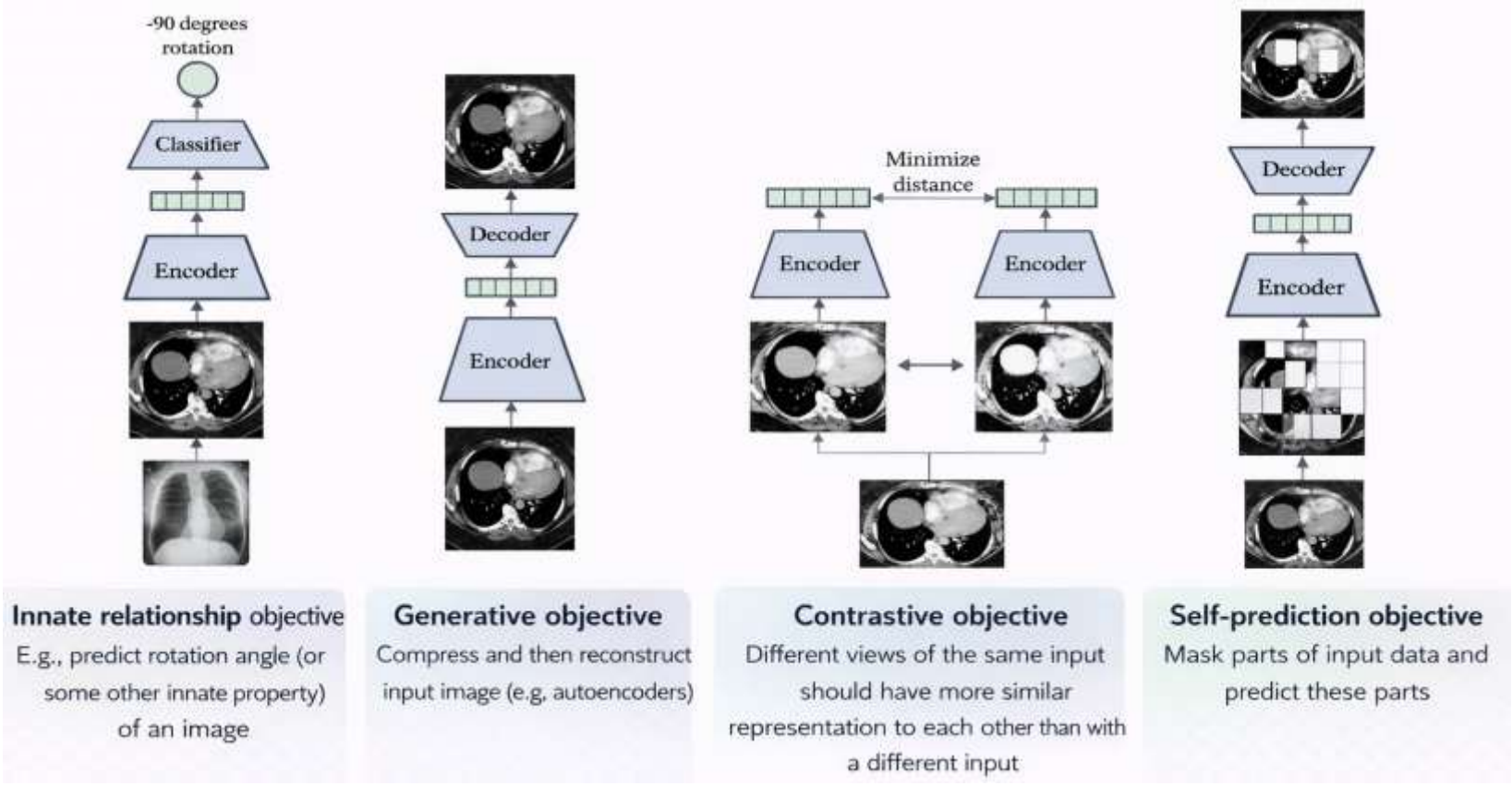


The label bottleneck in medical imaging



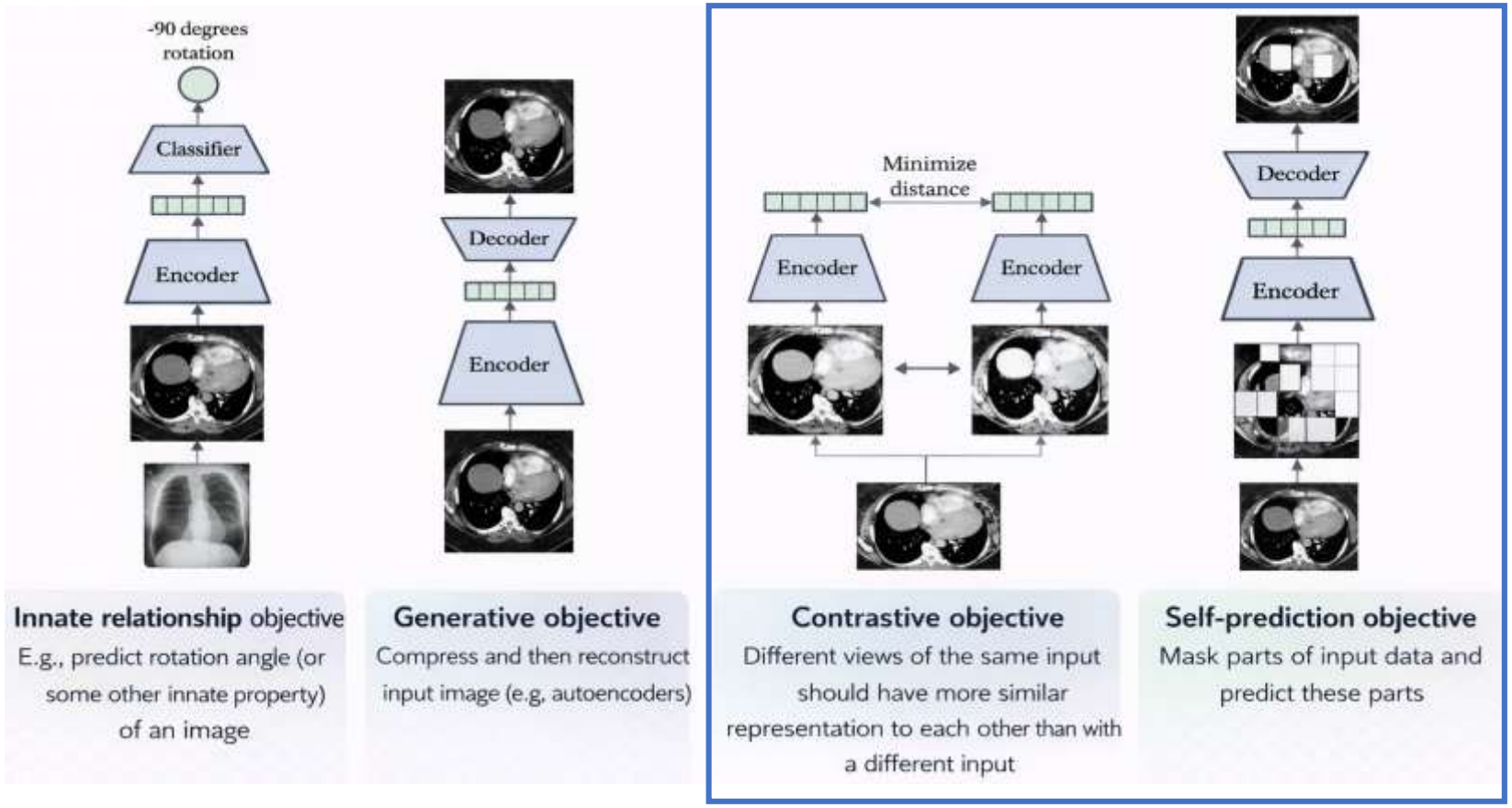
- Medical datasets are large but weakly labelled
- Expert annotation is expensive and slow
- Supervised learning does not scale
- Self-supervised learning: the way forward

Different representation learning paradigms



Different representation learning paradigms

Popular State-of-the-art approaches



Self-supervised learning landscape

Method	Key Idea	Architecture
SimCLR	Contrast augmented views	ResNet + MLP head
MoCo	Queue of negative examples	Momentum encoder
MAE	Mask and reconstruct image	Vision Transformer

- Learn representations from unlabelled images
- Different training strategies, same downstream goal
- Representations transfer to classification, segmentation, detection

Common thread: create a supervisory signal from the data itself — no human labels required

SimCLR: “Simple Framework for Contrastive Learning of Visual Representations”

1. **Stochastic Data Augmentation**

   Two random views per image from family T

2. **Base Encoder f(·)**

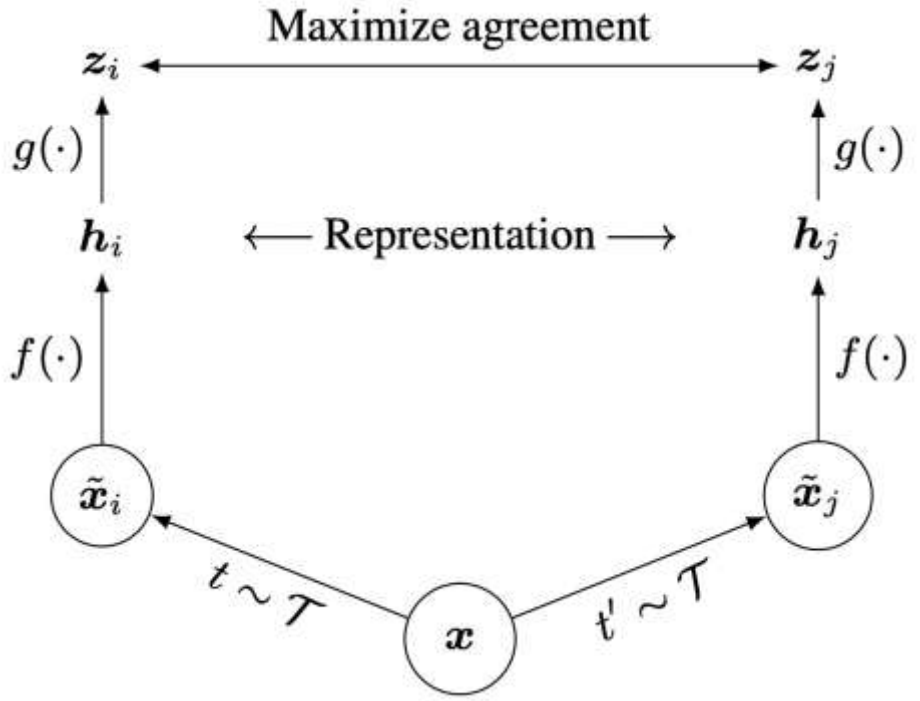
   ResNet-50 extracts representation h

3. **Projection Head g(·)**

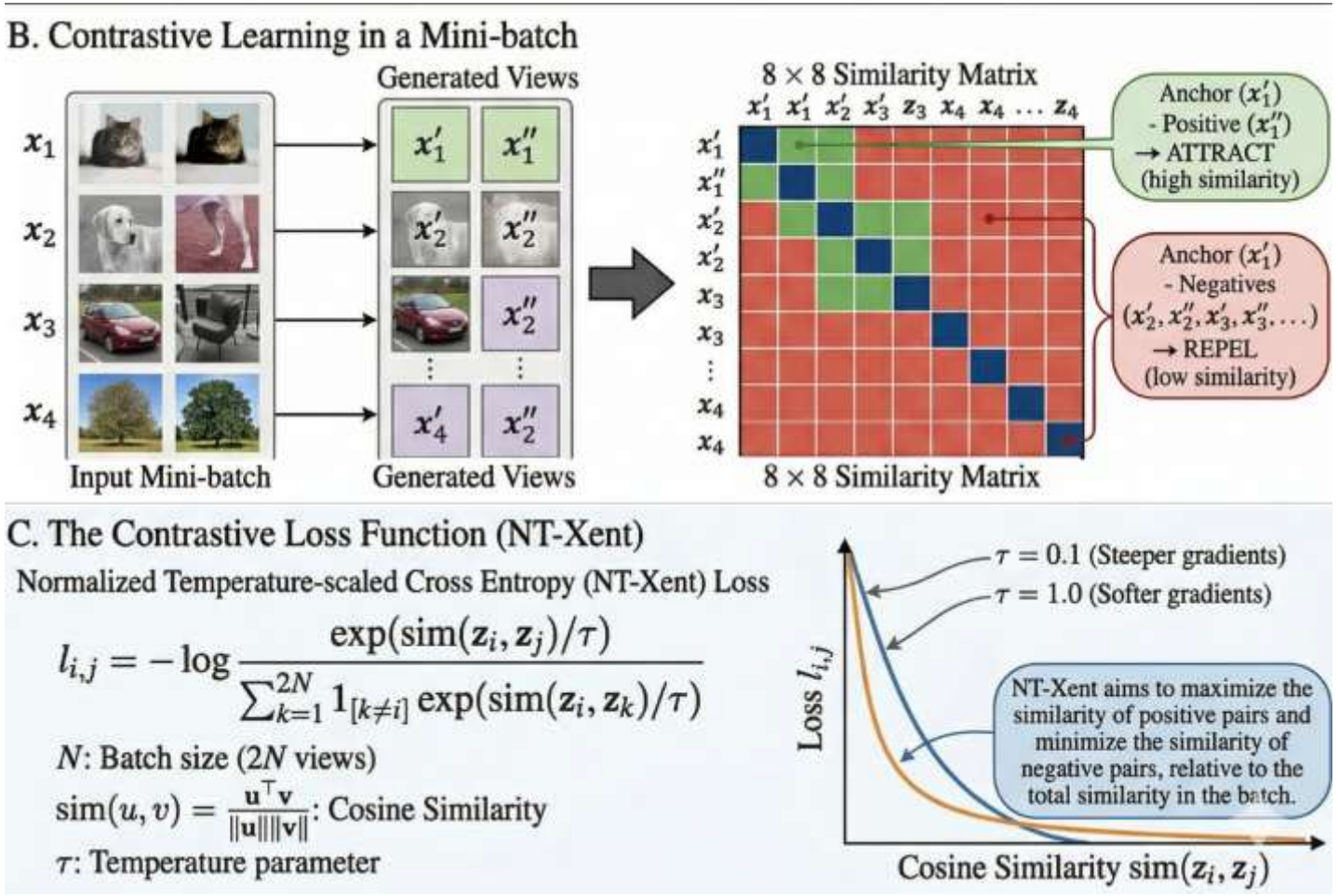
   MLP maps h → z for contrastive loss

4. **Contrastive Loss (NT-Xent)**

   Maximises agreement between positive pairs



Contrastive learning & contrastive loss



Data augmentation in SimCLR

Random Crop & Resize

Scale + position variation

Colour Distortion

Jitter, greyscale, hue shift

Gaussian Blur

Frequency-based variation

Composition of All

Combined is critical

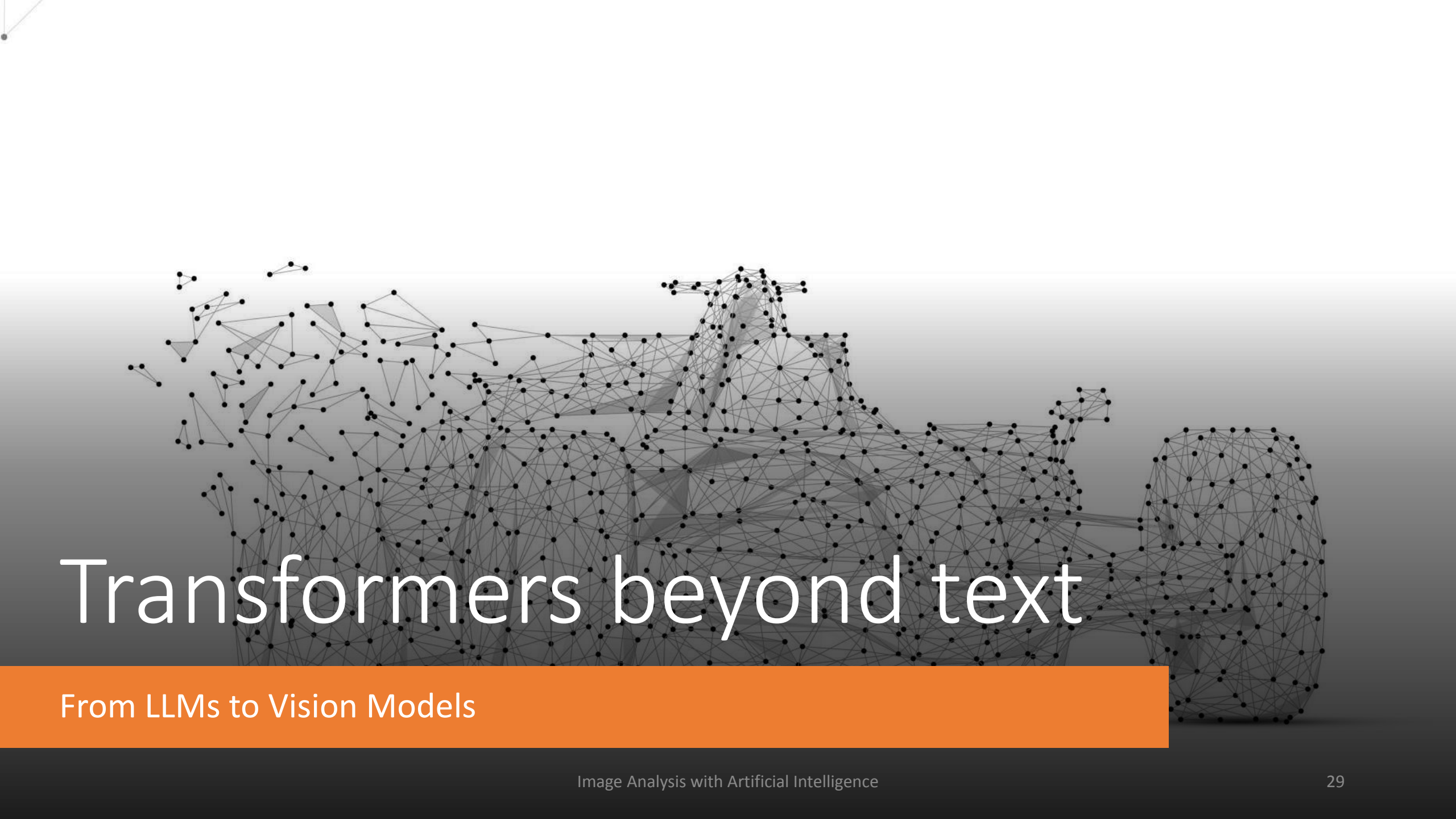
Key Finding (Chen et al., 2020)

Random crop + colour distortion is the most effective combination. Cropping alone creates views sharing colour distribution, allowing the network to match colour histograms (a shortcut). Adding colour distortion forces learning of higher-level semantic features.

Contrastive objective: Different views of the same input should have more similar representations than views from different inputs

Key limitations of SimCLR

- Massive Batch Size Requirements
  - SimCLR relies on having many "negative" examples to learn effectively.
  - Performance scales directly with batch size; the original paper used batches as large as 4,096 or 8,192.
  - Small batches lead to unstable training and poor representation learning.
- Extreme Computational Cost
  - Because of the large batch sizes, it requires significant hardware.
  - Training is time-intensive compared to supervised learning or more recent "lightweight" self-supervised methods.
- Heavy Reliance on Data Augmentation
  - The model's success is hyper-dependent on the composition of augmentations (specifically random cropping and colour jittering).
  - If the augmentations are too weak, the task becomes trivial; if they are too strong, the model loses semantic meaning.



Transformers beyond text

From LLMs to Vision Models

Natural Language Processing: Large Language Models (LLMs)

For decades, Recurrent Neural Networks (RNNs) and rule-based AI models were used to **read, understand, and analyze** text the way we (partially) do: one word at a time and left-to-right.

-But these AI models had a severe **"memory"** bottleneck.

The Breakthrough: "Attention is All You Need" (2017): The Transformer architecture flipped the script. It asked a radical question:

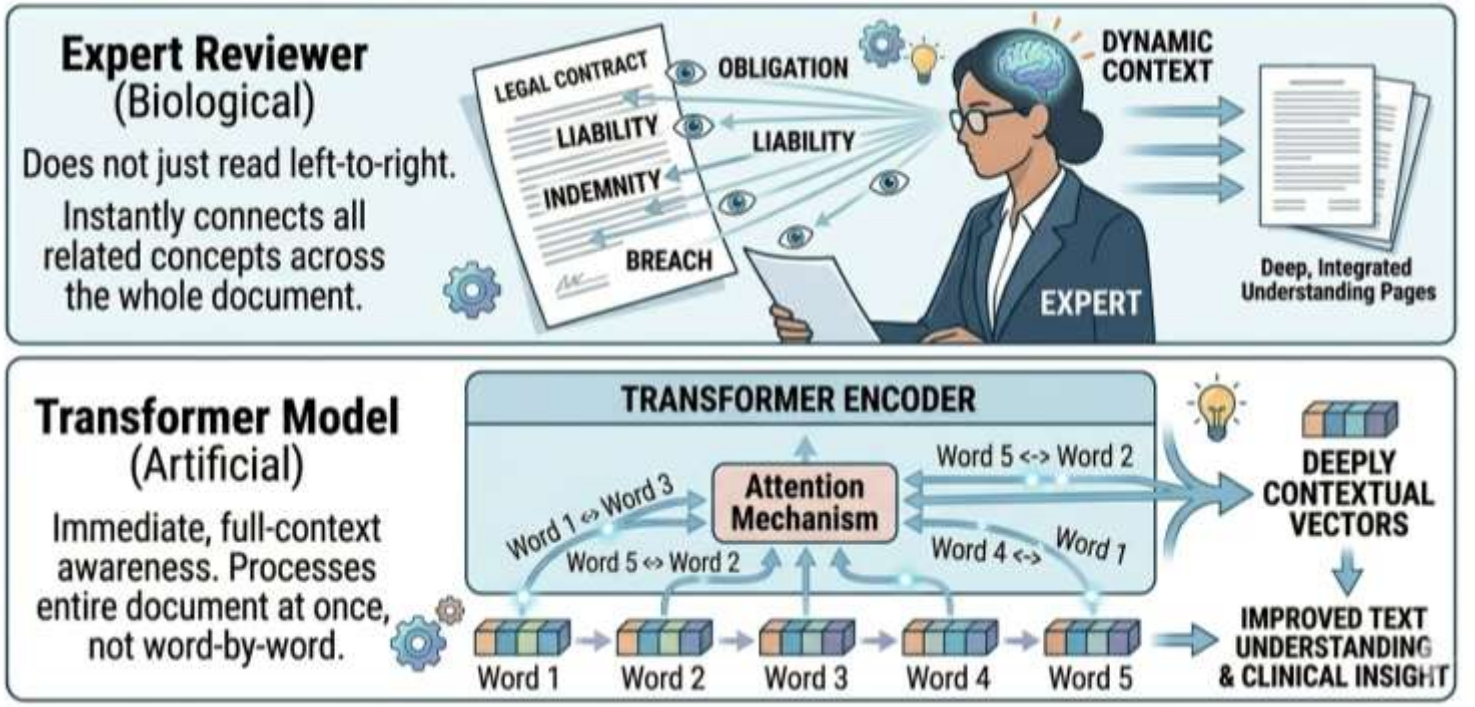
- What if we only use **attention mechanisms** and throw away the recurrent layers entirely?

Instead of processing word by word, the Transformer processes the **entire sentence** or paragraph in parallel.

How Self-Attention Works?

Think of a Transformer reading like an **expert reviewing a legal contract**. It doesn't just read left-to-right. For **every word** it encounters, it instantly looks back and forward across the entire document to find the most relevant **related concepts**.

This immediate, **full-context awareness** allows the model to build a **deeply dynamic and contextual understanding** of every single word based on its unique surroundings.



Transformers for Vision Transformers (ViT)

The Challenge: After conquering NLP, the question was: Can the Transformer's power of global context be applied to images? For computer science, images have always been the territory of **Convolutional Neural Networks (CNNs)**.

- **Why CNNs Dominate Vision:** CNNs are built on the principles of **locality** and **translation invariance**. A CNN's filter looks at a tiny 3x3 pixel neighborhood, identifies a small feature (like an edge or texture), and assumes that feature is useful anywhere in the image. This is a very efficient and powerful assumption (an "inductive bias").
- **The Problem with CNNs:** Because CNNs look locally, they struggle to connect features that are far apart. They can identify two eyes and a nose, but they might fail to realize they belong to the same face if they are separated in an unusual way. They lack the "big picture."

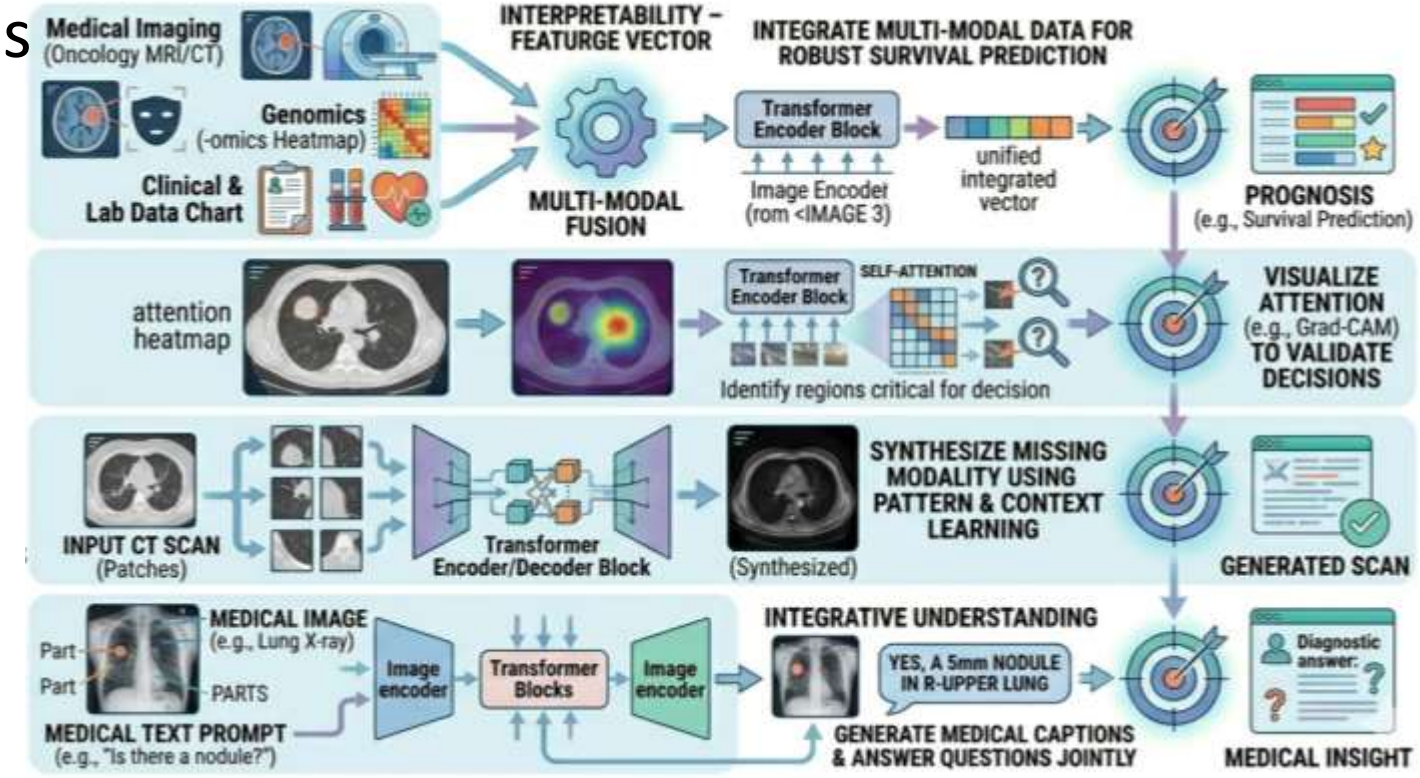
Introducing the Vision Transformer (ViT): The ViT group asked: Can we make an image look like a sentence?

- **Crop the image into square patches** (e.g., 16x16 pixels).
- **Treat each patch like a "word"** (a visual token).
- **Feed the sequence of visual tokens into a standard Transformer encoder.**
- **The Impact of Global Self-Attention:** By treating image patches like words, the ViT applies **Self-Attention across the entire image at once.**

A ViT doesn't just look at neighboring pixels; it learns how a patch containing a "cat's tail" in the corner relates to a patch containing a "cat's ear" in the opposite corner. This **Global Context awareness** allowed ViTs to surpass state-of-the-art CNNs on massive datasets, proving that attention could learn visual patterns without the built-in local assumptions of convolutions.

Novel Medical Imaging Uses of Transformers

- Deep features & Radiomics as **Imaging Embeddings** for Transformers
- Enhanced **Ante-Hoc XAI**
- **Fill** Missing Modalities for Multi-Modal AI
- **VLMs**: Reports to Imaging



Demystifying the Black Box – eXplainable AI (XAI) in medical imaging

Manos Koutoulakis, PhDc

Computational BioMedicine Laboratory (CBML)
Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)

Explainable AI (XAI) – The problem

Input CT scan

Deep Neural Network
(CNN, Transformer – Millions of parameters)

Output:
92% Malignant

Clinicians: **How did it come up with this prediction?**

Interpretability VS Explainability

Interpretability
Model is **inherently**

Linear Regression

$y = 2.6x_1 - 1.2x_2$

You can read the coefficients Feature x_1 increases y by 2.3

Examples: Decision Tree, Linear regression

Explainability
Post-hoc method explains

Deep CNN + XAI

Saliency Maps

Highlights the specific image pixels that drove the model's prediction

Examples: IG, CAM, SmoothGrad, GradCAM

Three Families of XAI Methods

Gradient-Based
Fast & Pixel-level

- Backpropagation
- Saliency
- Integrated Gradients (IG)
- CAM Family (GradCAM, CAM, etc)
- DeepLIFT

Perturbation-based
Model Agnostic

- Masks input regions
- Occlusion
- Feature Ablation
- SHAP

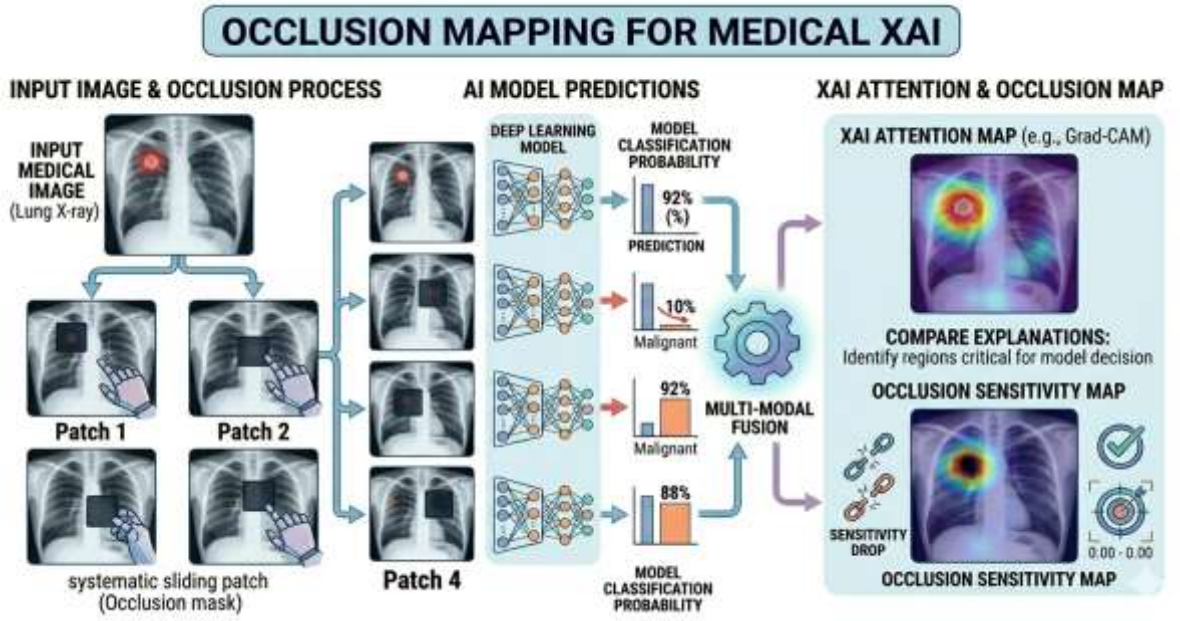
Attention-Based
For Transformers

- Reads Attention Maps
- Attention Rollout
- Cross-Attention
- DINO

Occlusion Sensitivity Mapping

This approach unveils the potential feature importance of a deep model using systemic or random perturbation/conditioning over the mage

- By occluding different voxels, patches, regions
- Occlusion maps **do not** take the feature maps into account, but only the different patches of the input image



PROS

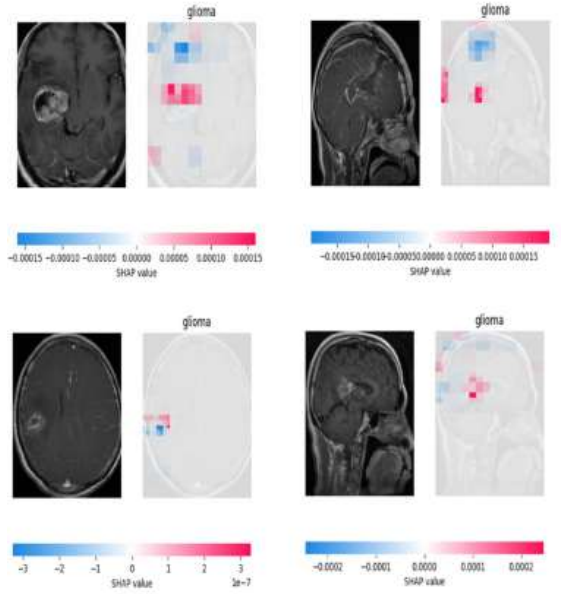
- **Simple** and **intuitive** methods to perform and interpret
- **Easy adaptation** to specific occlusion analysis
- **Easy comparison** with traditional clinical analysis (atlas-based methods), providing **transparent** visualization of the DL decision

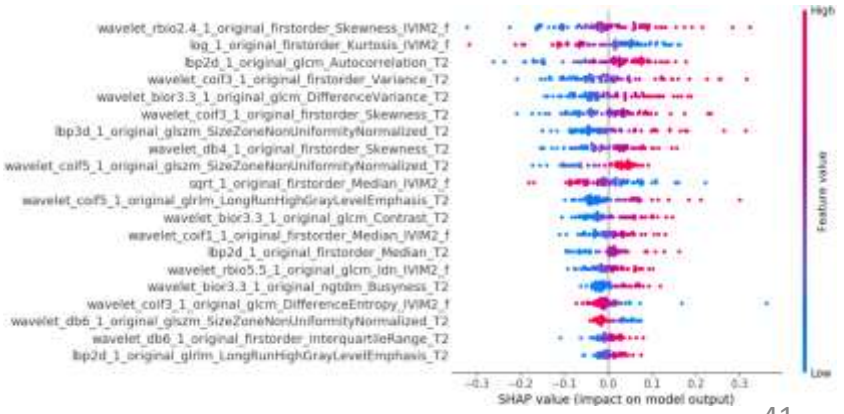
CONS

- **Large computational requirements** due to many forward & backward propagations
- **Too rigid** to follow anatomical/pathological structures present in the images

SHapley Additive exPlanation (SHAP)

- **<u>The Game Theory Foundation</u>**: SHAP is rooted in cooperative game theory, where it calculates the "Shapley value" by measuring the marginal contribution of every feature (the "players") toward the final model prediction (the "prize").
- **<u>Application to Imaging</u>**: In a medical imaging context, SHAP treats each voxel as a feature; it determines a voxel's attribution by calculating how the model's prediction changes when that specific voxel is included versus when it is replaced by a reference baseline.
- SHAP can also be used in **radiomics interpretability**!



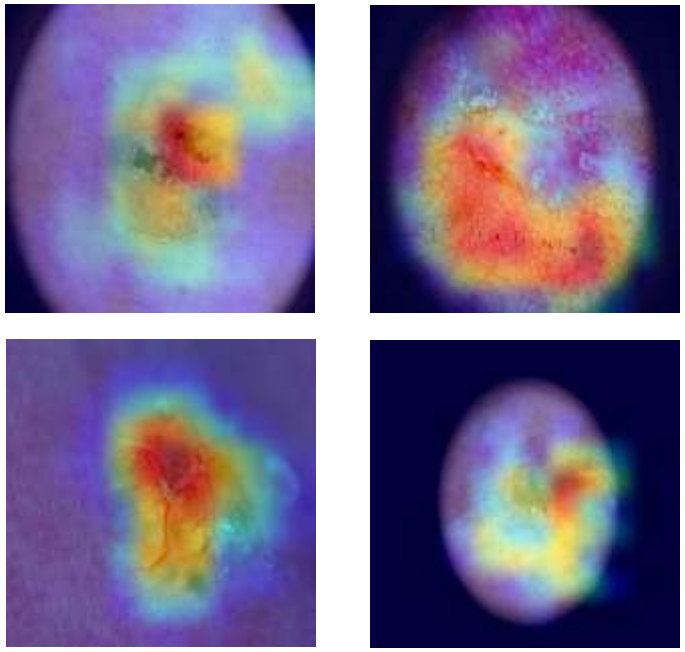


Ioannidis, G. S., Nikiforaki, K., Dimitriadis, A., Goumenakis, M., Trivizakis, E., Papanikolaou, N., ... & Marias, K. (2025, July). Explainable AI Radiomics in Prostate Cancer Aggressiveness Prediction using different quantitative Diffusion MRI models. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1-7). IEEE.

Gradient-weighted CAM (Grad-CAM)

A generalization of the CAM approach by employing the gradient of output of the network with respect to the activations of the feature maps to generate the attribution maps (heatmap)

- Class-specific producing a visualization of each class
- Employes a gradient-based weighted average of feature maps to create heatmaps



PROS

- Improves accuracy and interpretability of results over other gradient-based methods
- Unlike CAM, there is no requirement for specific CNN architecture

CONS

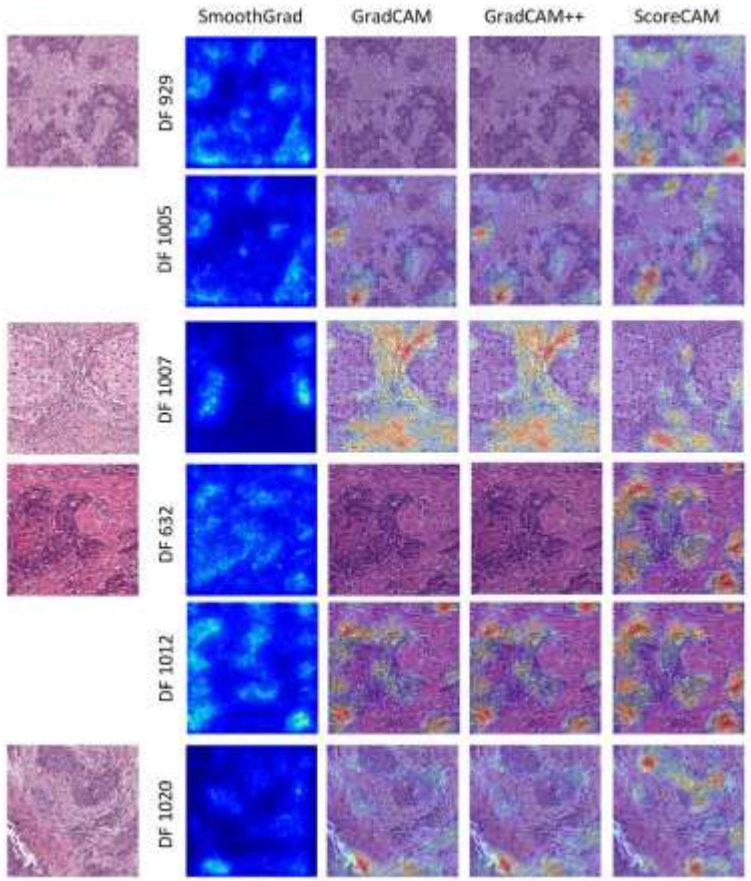
- Lack of robustness to changes in input image and unclear explanation of the basis for prediction in complex images
- Low specificity & resolution due to the low dimensionality of its attribution maps

XAI in Healthcare

Clinical Safety	Detect crucial model errors before harm
Trust Building	Clinicians need to verify reasoning
Regulatory Compliance	FDA deems explainability mandatory
Scientific Evolution	Biomarker Discovery

Challenges

Same model & same input could yield **different** attention maps!



- Method bias
- Interobserver bias
- Minimal clinician involvement
- Qualitative evaluation with limited objectivity
- Understudied confounding variables that affect XAI

Koutoulakis, E., Trivizakis, E., Markodimitrakis, E., Agelaki, S., Tsiknakis, M., & Marias, K. (2026). A critical review of explainable deep learning in lung cancer diagnosis. *Artificial Intelligence Review*, *59*(1), 28.

Thanks!

Any question?