An Overview of Concept-based XAI

Elena Baralis, Eliana Pastor, Gabriele Ciravegna,
Tania Cerquitelli, Eleonora Poeta

ECML PKDD 2025, Porto - September 15th, 2025

Politecnico di Torino

CENTAI

Future Artificial Intelligence Research

ECML PKDD 2025

Meet the instructors and contributors

Eleonora Poeta
PhD student at Politecnico di Torino, Italy.

Gabriele Ciravegna
Researcher at CENTAI Institute, Italy.

Eliana Pastor
Assistant professor at Politecnico di Torino, Italy.

Tania Cerquitelli
Full professor at Politecnico di Torino, Italy.

Elena Baralis
Full professor at Politecnico di Torino, Italy.

Tutorial structure

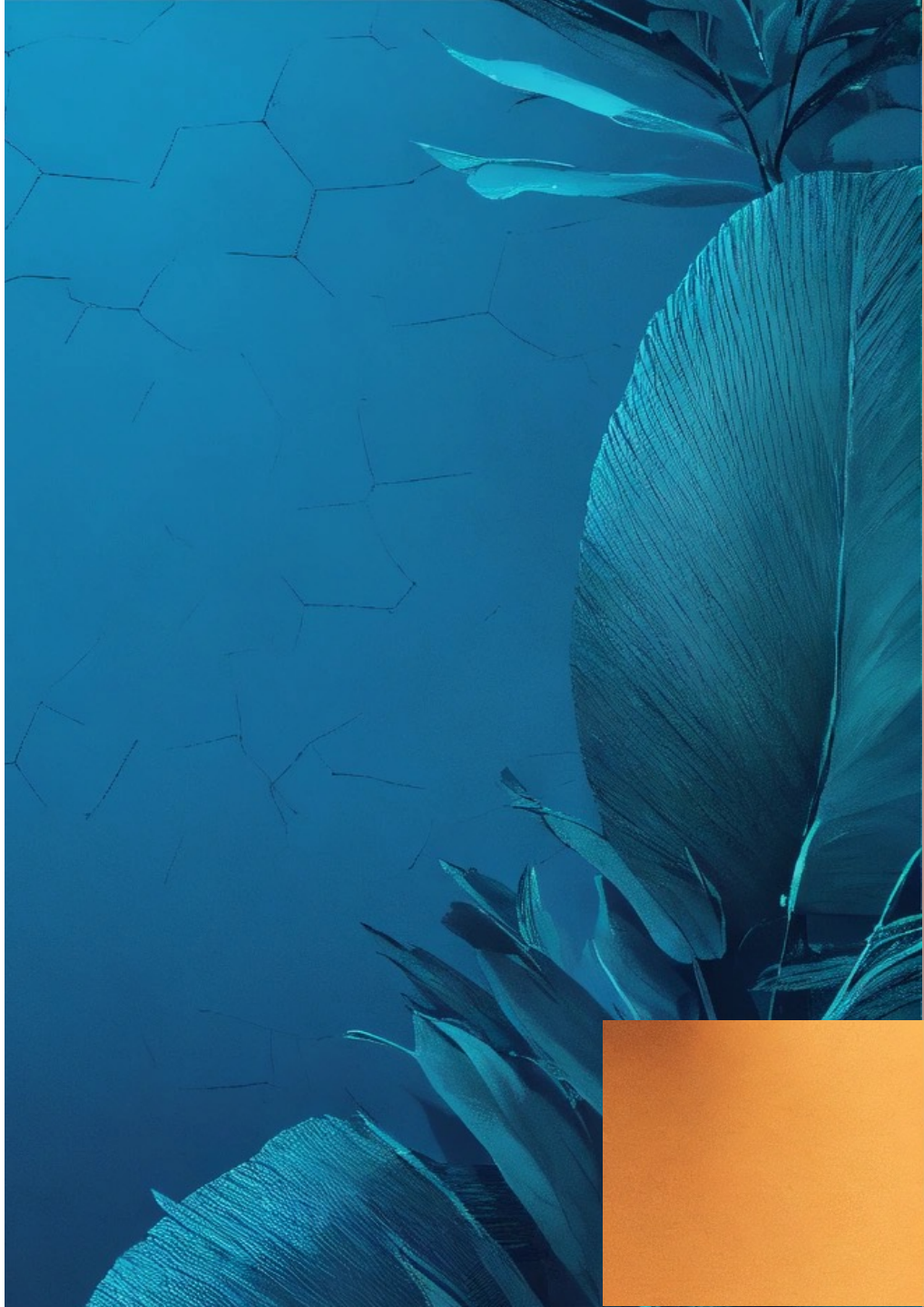
https://sites.google.com/view/cxai-tutorial

Material

- Slides
- Hands-on tutorial - repository
  - T-CAV
  - CBM
  - Label-Free CBM

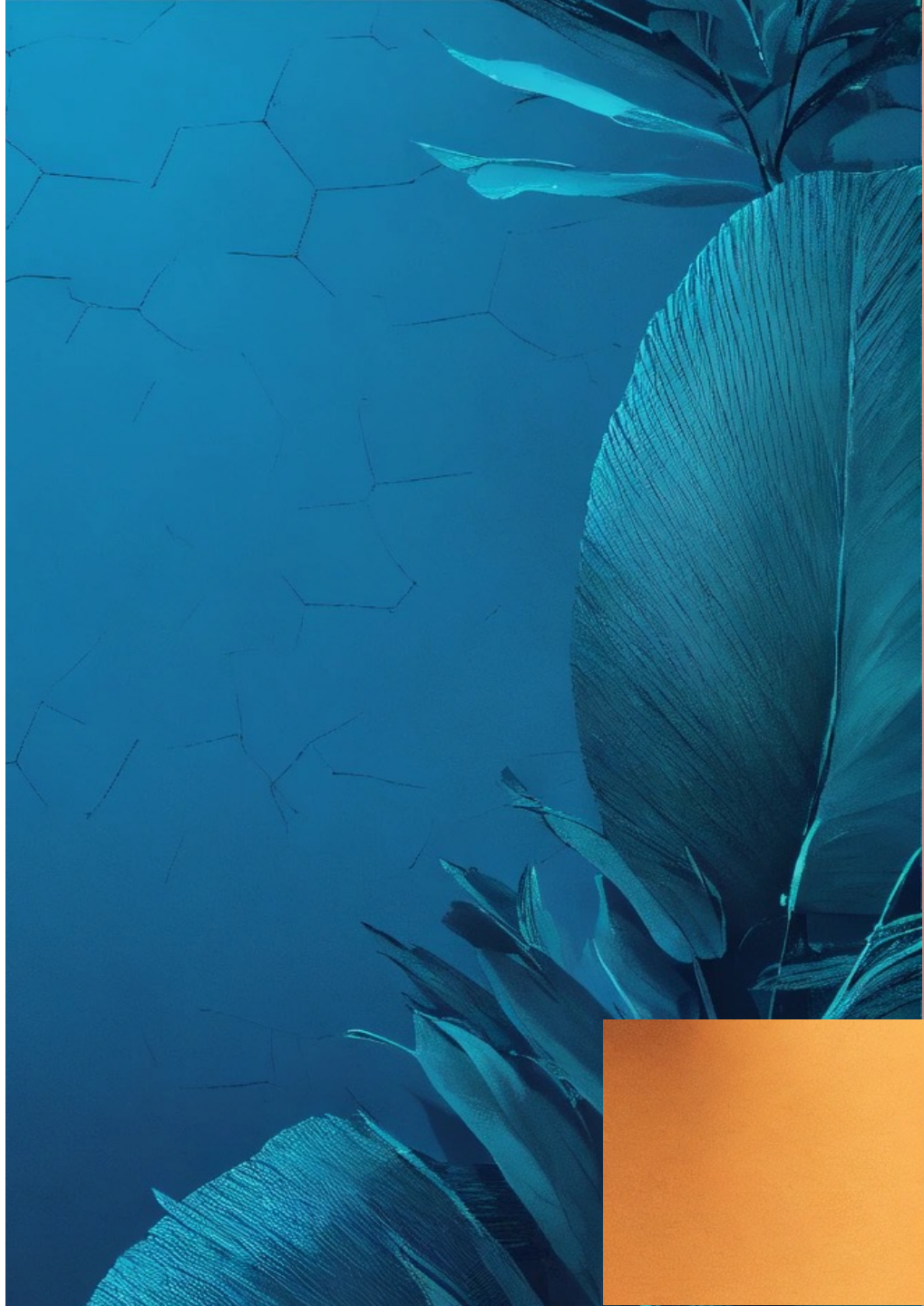
Tutorial structure

- Introduction to XAI techniques
- Taxonomy of C-XAI
- Post-hoc vs By-design C-XAI approaches
- Guidance for selecting C-XAI approaches
- Hands-on session



Tutorial structure

- **Introduction to XAI techniques**
- Taxonomy of C-XAI
- Post-hoc vs By-design C-XAI approaches
- Guidance for selecting C-XAI approaches
- Hands-on session



Why Explainability Matters in AI

- Deep learning systems are everywhere
- These systems influence critical decisions
- Demand is growing for AI that is ***trustworthy*** and ***understandable***

Call for transparency

Increasing adoption of AI for medical diagnosis

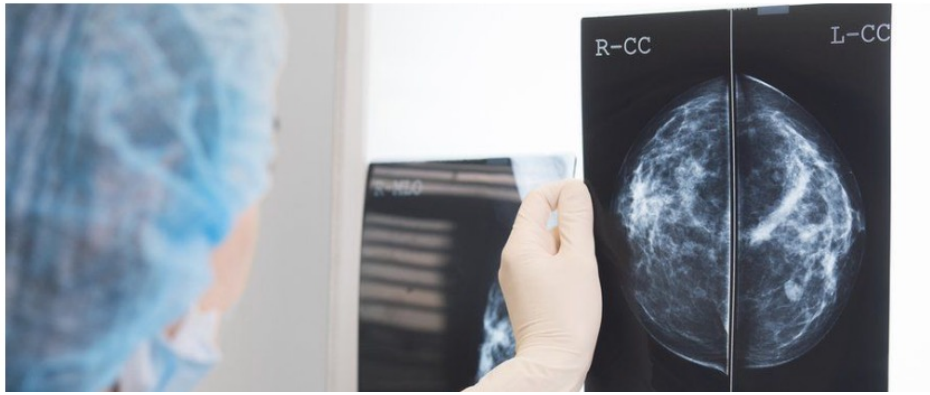
Accurate results, even ‘outperforming doctors’

But...

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

2 January 2020



Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin + Show authors

Nature **577**, 89–94 (2020) | Cite this article

98k Accesses | **1288** Citations | **3927** Altmetric | Metrics

Call for transparency

Response article

Researchers call for ***transparent and reproducible AI research***

Matters Arising | Published: 14 October 2020

Transparency and reproducibility in artificial intelligence

Benjamin Haibe-Kains ✉, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush & Hugo J. W. L. Aerts

Nature **586**, E14–E16 (2020) | Cite this article

15k Accesses | **54** Citations | **520** Altmetric | Metrics

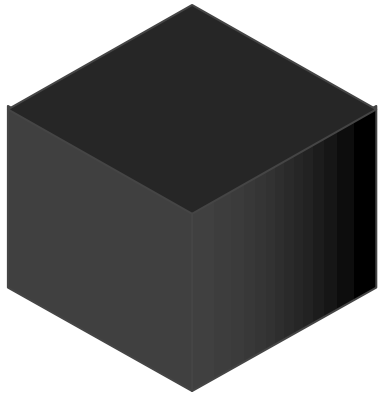
*'[…] The lack of details of the methods and algorithm code undermines its scientific value. Here, we identify obstacles that hinder **transparent and reproducible AI research** and provide solutions to these obstacles with implications for the broader field.'*

Transparency & Explainability

Most AI models are black boxes and lack interpretability

Opaqueness of a model can be at **multiple levels**, involving

- Data
- Model/System/Algorithm
- Learned function and pattern → reasons behind its functioning
- Intention and business model of the AI product



These elements should be transparent - clear, disclosed - to the end users

A definition of Explainability/Interpretability

*The ability to **explain** or to present the **reasoning** behind the decisions **of AI systems** or its technical process **in understandable terms to humans***

Multiple terms – not monolithic concepts

- **Interpretability**: Focus more on transparency, models that are **inherently interpretable**
- **Explainability**: Focus more on models that are not comprehensible by design to humans
- + understandability, comprehensibility, intelligibility, mental fit

Often used interchangeably

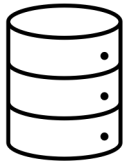
Molnar, Christoph. “Interpretable machine learning”. (2020)

A taxonomy of Explainable AI

- **At which step of the ML Pipeline?**
  - Pre-modeling, modelling, post-modelling explainability
- **Is the explanation method general?**
  - Model dependent vs model agnostic
- **What do we explain?**
  - The global model, subgroups, a single prediction
- **How do we represent the explanations?**
  - Feature importance, rules, visualization-based explanation, explanation by example
- **How can we derive explanations?**
  - Explaining by removing, local models, gradient-based

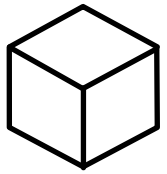
Explainability Stages

Explainability involves the entire AI development pipeline



Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering



Explainable modeling

Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

After model development
- Explaining predictions and behavior of trained models

Generalizability of Explainability

Is the explanation method general?

Model dependent solutions

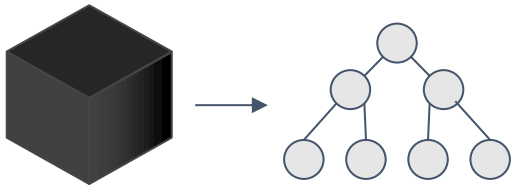
- Only applicable for specific models
- Rely on the model structure/properties

Model agnostic solutions

- Applicable to any model
- Rely on the model as an oracle (model predictions, output probabilities)

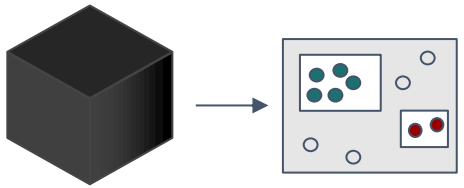
Scope of Explainability

What do we explain?



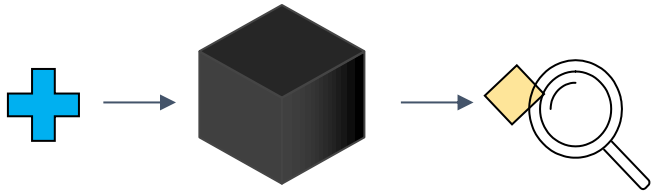
Global

How the model works globally



Subgroup

How the model behaves in data subgroups



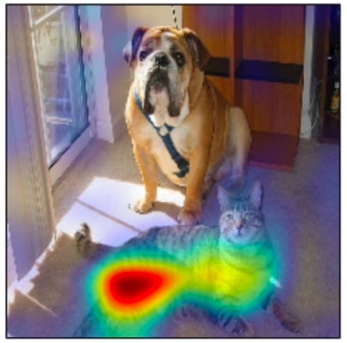
Individual/local

Explaining the reasons behind individual predictions

Multiple forms of explanation representation

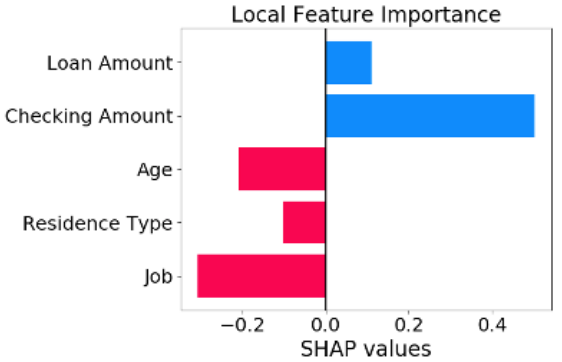
How do we represent the explanations?

Input attributions



(c) Grad-CAM 'Cat'

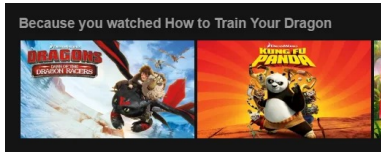
I am really happy

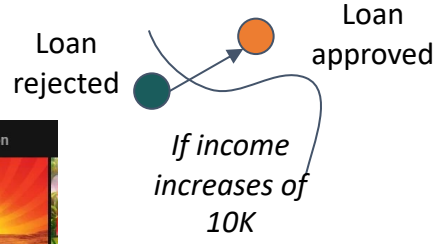


Rules

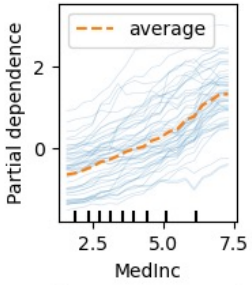
If Country is US, married, work hours > 45 → Income>50K

Explanations by example



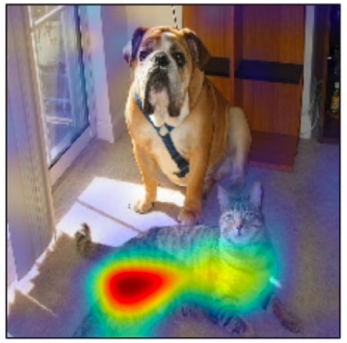


Visualizations



Multiple forms of explanation representation

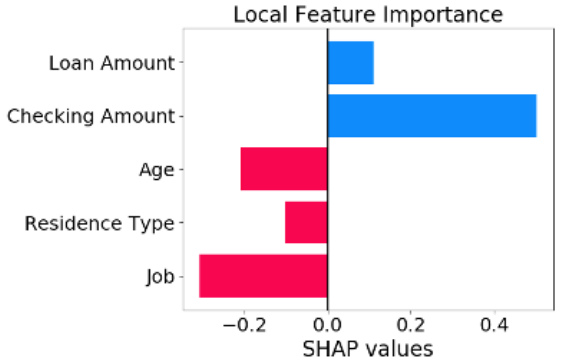
How do we represent the explanations?



(c) Grad-CAM 'Cat'

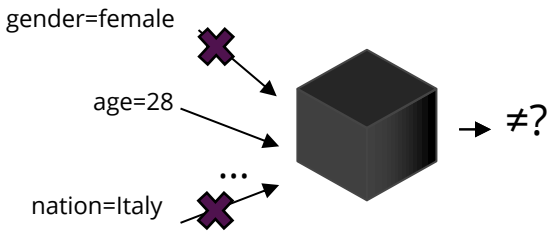
Input attributions

I am really happy

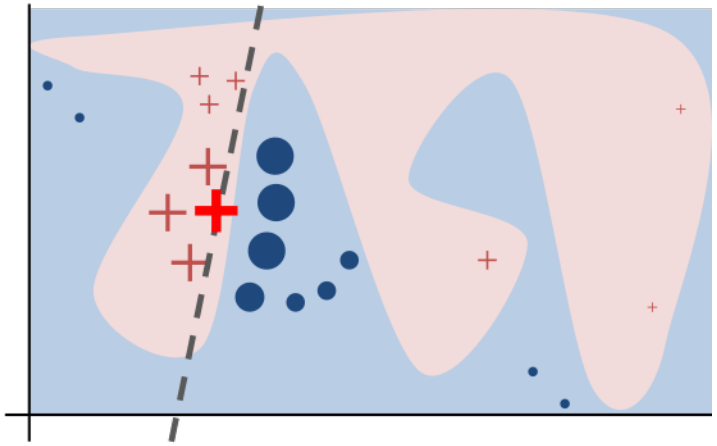


Identify which input features (e.g., pixels, words, attributes) contribute most to the model's output

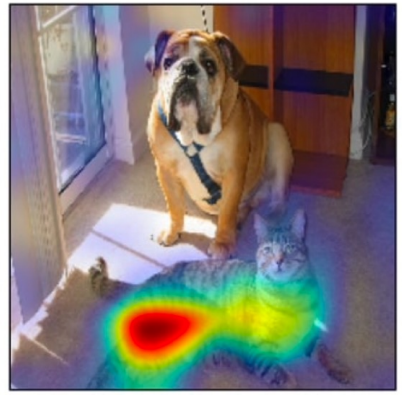
Multiple methodologies to derive input attribution explanations



Explaining by removing
SHAP, ..

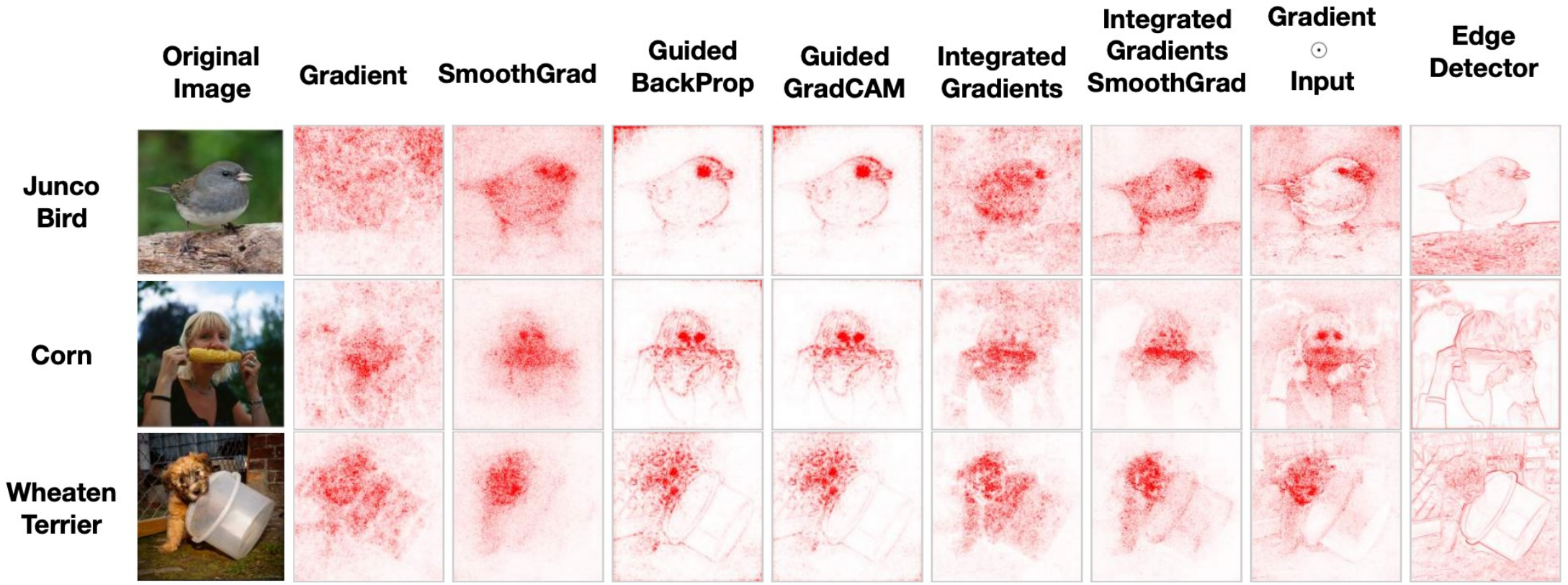


Local models
LIME, ..


(c) Grad-CAM 'Cat'

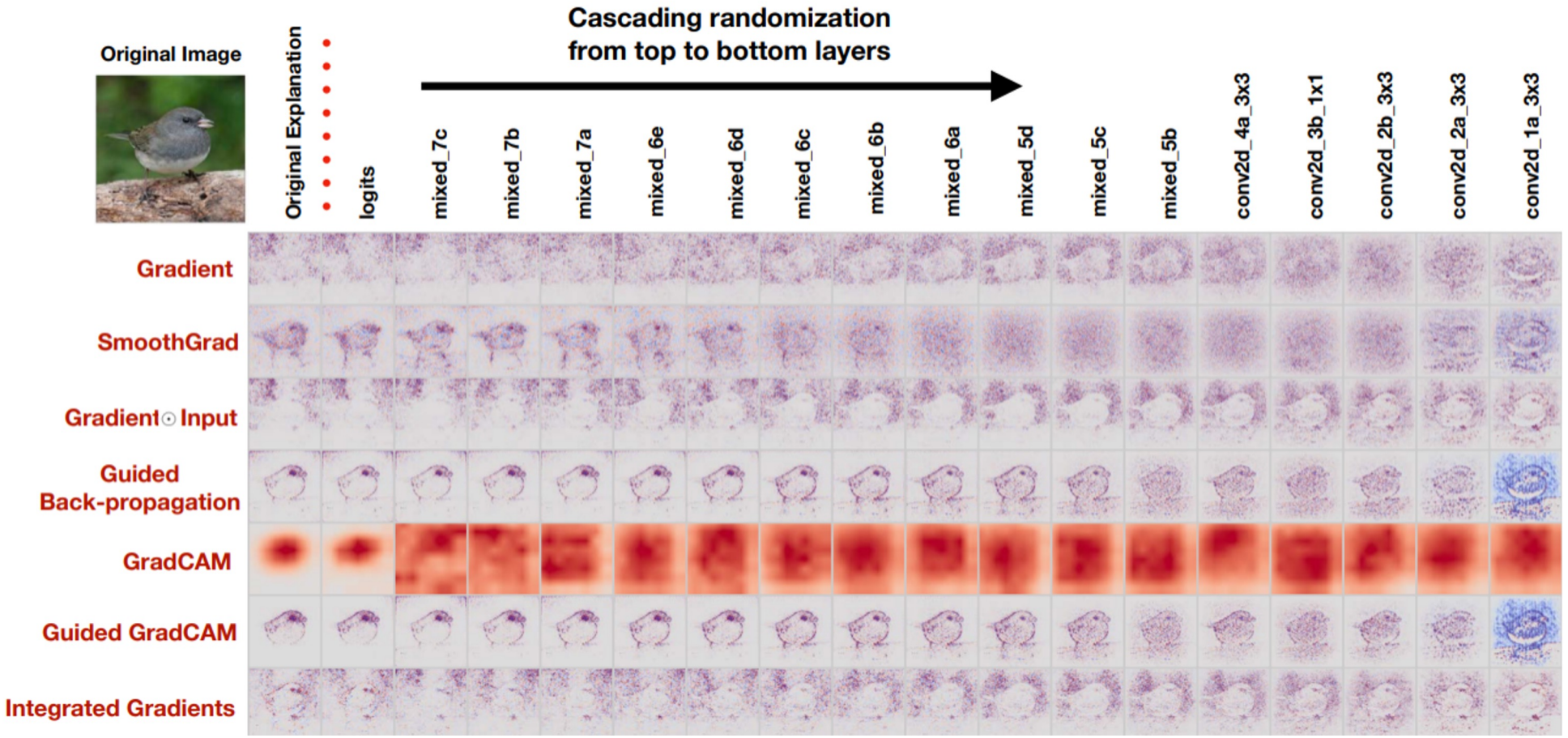
Gradient-based
IntegratedGradient, GradCAM, ..

Standard Explainable AI does not always work well - Edge detectors?



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS 2018.

Standard Explainable AI does not always work well - Faithful to model behavior?



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS 2018.

Standard Explainable AI may be difficult to interpret - Which class are we explaining?

Evidence for Siberian Husky

Evidence for Transverse Flute

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" Nature machine intelligence (2019)

The **where** vs the **what**

Showing ***where*** a network is looking

does not tell us what the network is seeing in a given input

Rudin, Cynthia. "Stop explaining black box machine learning models […]" Nature machine intelligence (2019)
Achtibat, Reduan, et al. "From attribution maps to human-understandable explanations" Nature machine intelligence (2023)

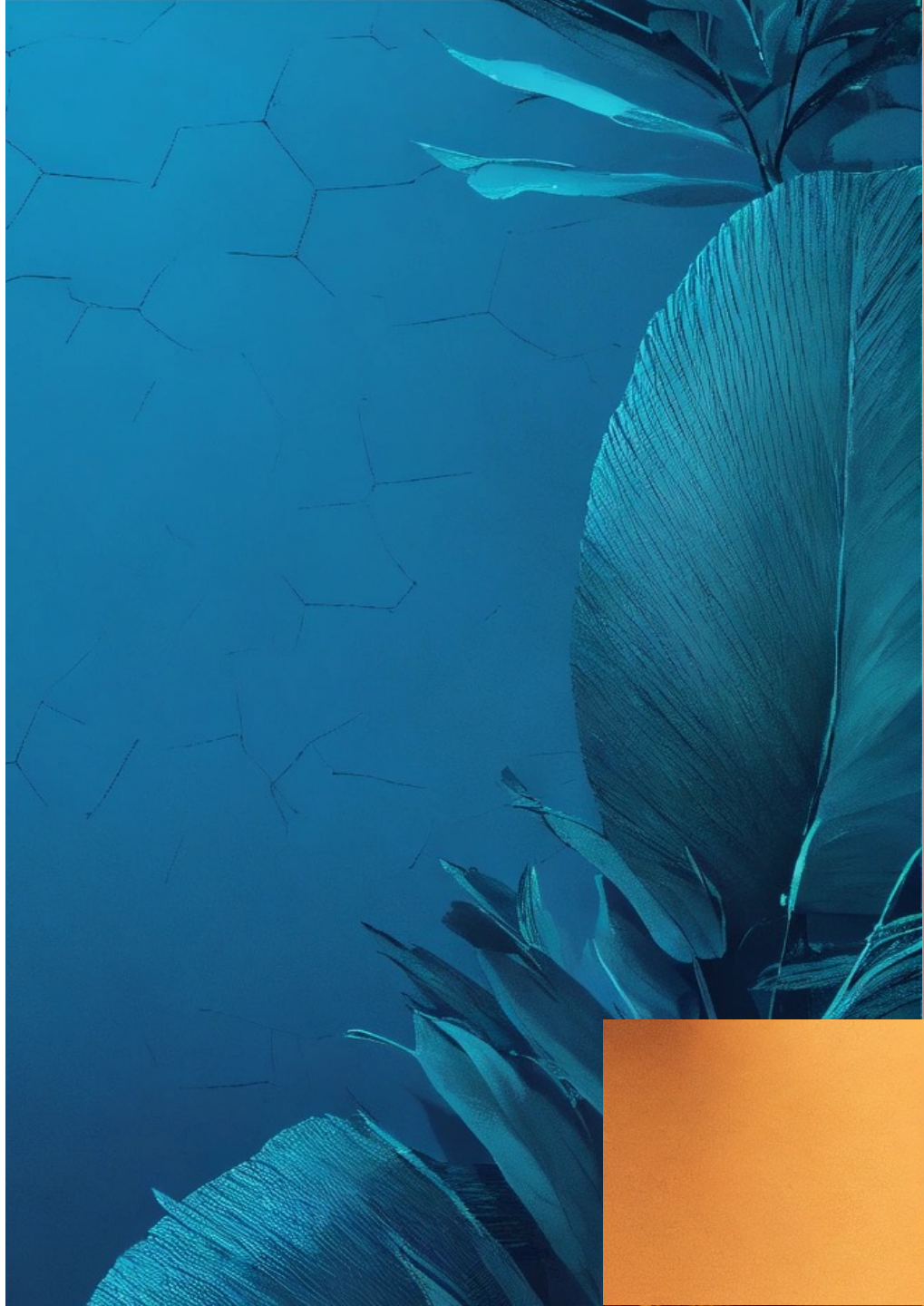
Beyond Input Attribution

Concept-Based XAI (C-XAI)

Explain model behavior via ***high-level, human-interpretable*** **concepts**

Tutorial structure

- Introduction to XAI techniques
- **Taxonomy of C-XAI**
- Post-hoc vs By-design C-XAI approaches
- Evaluation, Resources and applications
- Hands-on session



Taxonomy of C-XAI

- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations , Node-Concept Association , Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**

What is a concept?

A concept can be any abstraction, such as a colour, an object, or even an idea

Molnar, Christoph. “Interpretable machine learning”. (2020)

Taxonomy of C-XAI

- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations , Node-Concept Association , Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**

Types of concepts

- **Symbolic Concepts**
  Human-defined attributes
- **Unsupervised Concept Basis**
  Cluster of similar samples
- **Prototypes**
  (Part-of) a training sample
- **Textual Concepts**
  Textual representation

“BEAK” **“GREEN”** **“WINGS”**

“A bird with bright feathers”

Poeta E., Ciravegna G., Pastor E., et al., "Concept-based Explainable Artificial Intelligence: A Survey" (2024)

Symbolic Concepts

“BEAK” **“GREEN”** **“WINGS”**

- **Human-defined attributes or abstractions**
  - Of the final classes
  - e.g., bird --> the beak of the bird, the color of the bird

- **Require auxiliary data & annotations**
  - **Istance-level annotation**
    - Annotate the presence of a concept for each instance
    - More expensive
  - **Class-level annotation**
    - All samples belonging to a class are annotated as having a certain attribute
    - Less expensive but less precise (e.g., attribute could not be visible)

Unsupervised Concept Basis

- **Cluster of similar samples**
  - Extracted from the network representation (the latent space)
  - e.g., a cluster of green birds.

- Can capture abstractions more understandable to humans than individual features or pixels

- Clustering algorithms employed to extract unsupervised concepts

Prototypes

- **Explanation by example**

- **Representative examples** of peculiar traits of the training samples
  - **Entire samples**
  - **Parts of a training sample** (e.g., a hooked beak)

- Different from unsupervised concept basis
  - Represent a single example instead of a group of examples

Textual Concepts

"A bird with bright feathers"

- **Textual descriptions of main classes**
  - From an individual description, distinctive parts are extracted
- **Provided** at training time **by an external generative model**
  - It requires a Large-Language Models with knowledge of the given task

Taxonomy of C-XAI

- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations, Node-Concept Association, Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**

Types of concept-based Explanations

- **Class-Concept Relations**

Relation among a concept and an output class of a model

$$Beak \rightarrow Parrot$$

- **Node-Concept Association**

Explicit association of a concept with a hidden node of the network

- **Concept-Visualization**

Visualization of a learnt concept in terms of the input features

Poeta E., et al., "Concept-based Explainable Artificial Intelligence: A Survey" (2024)

Class-Concept Relations

$Beak \rightarrow Parrot$

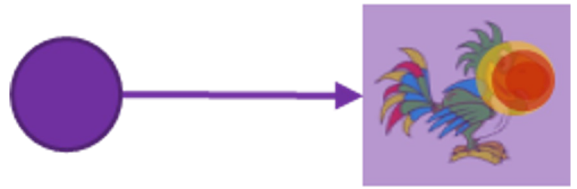
- **Relationship between a concept and an output class of the model**
  - Concept importance
  - Logic rule involving multiple concepts and their connection to an output class

- Can be applied to all type of concepts
  - e.g., symbolic, parrot = 0.3 beak + 0.2 green + 0.3 wings + 0.2 tropical
  - e.g., with prototypes, parrot := 0.8 prototype_1 + 0.2 prototype_2

Node-Concept Association

- **Assign a concept to an internal unit (or a filter) of a network**
- It **enhances the transparency** of deep learning models
  - highlighting what internal units see in a given sample
- It can be defined **post-hoc**
  - by considering the hidden units maximally activating on input samples representing a concept
- It can also be forced **during training**
  - by requiring a unit to predict a concept

Concept Visualization



- **Highlight the input features that best represent a specific concept**
  - Similar to saliency map but for concepts

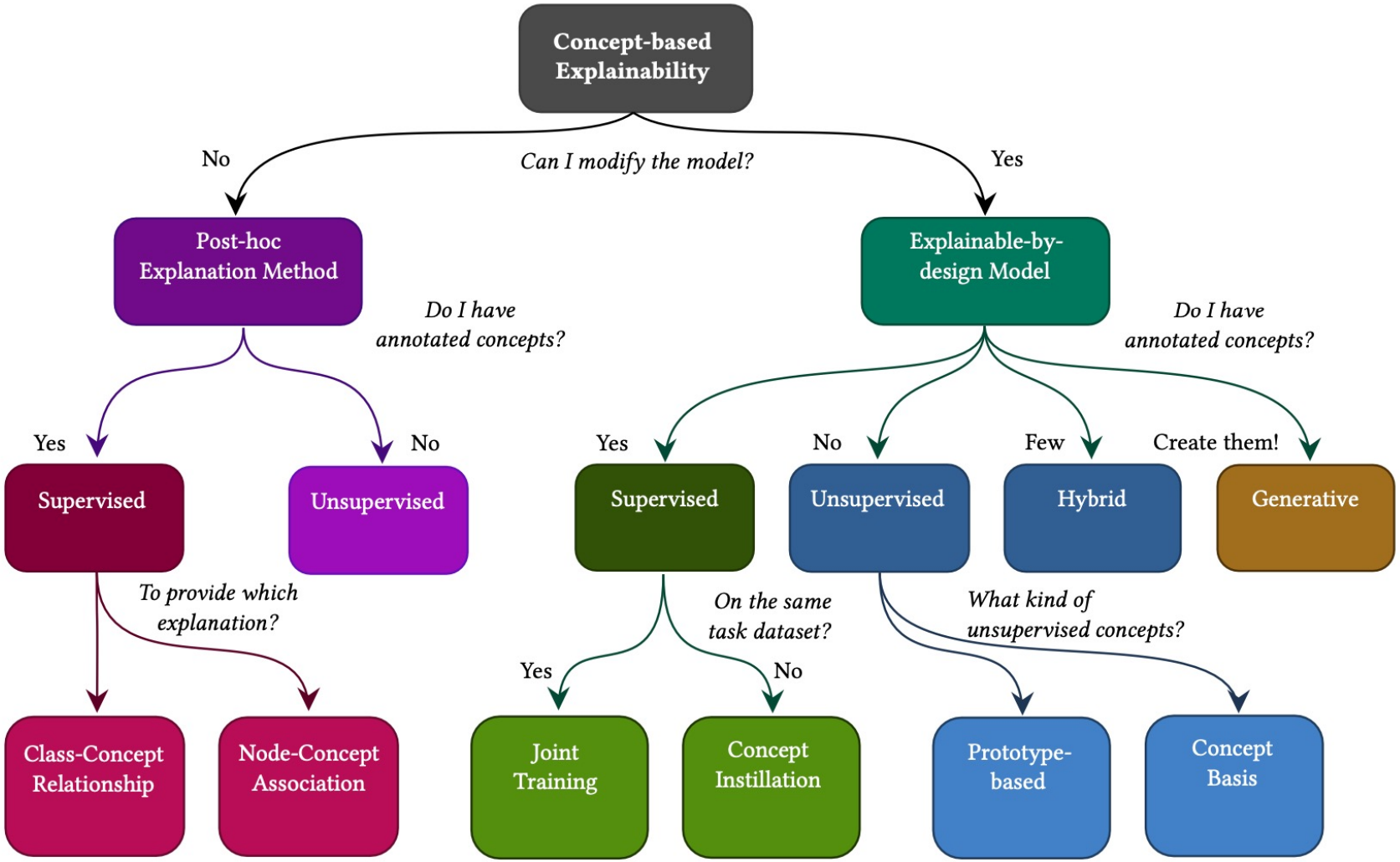
- **Crucial when non-symbolic concepts are employed**
  - Need to understand which **unsupervised concept basis or prototypes** the network has learned

- Often combined with one of the previous explanations
  - Enable understanding the concepts associated with a specific class or node

Taxonomy of C-XAI

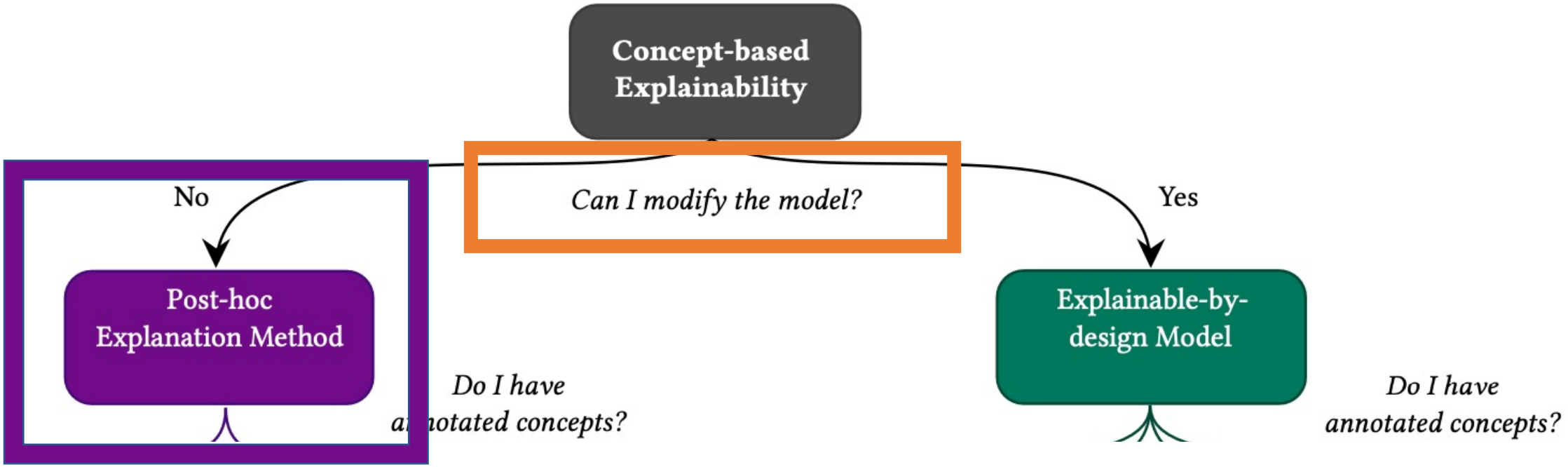
- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations , Node-Concept Association , Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**

C-XAI Taxonomy



Poeta E., et al., "Concept-based Explainable Artificial Intelligence: A Survey" (2024)

C-XAI Taxonomy



Poeta E., et al., "Concept-based Explainable Artificial Intelligence: A Survey" (2024)

Post-hoc Concept-based Explanation methods

Explain via concepts an **already trained model** that we cannot (or do not want to) modify

Underlying idea is that **a network automatically learns to recognize** some **concepts**

Post-hoc Concept-based Explanation methods

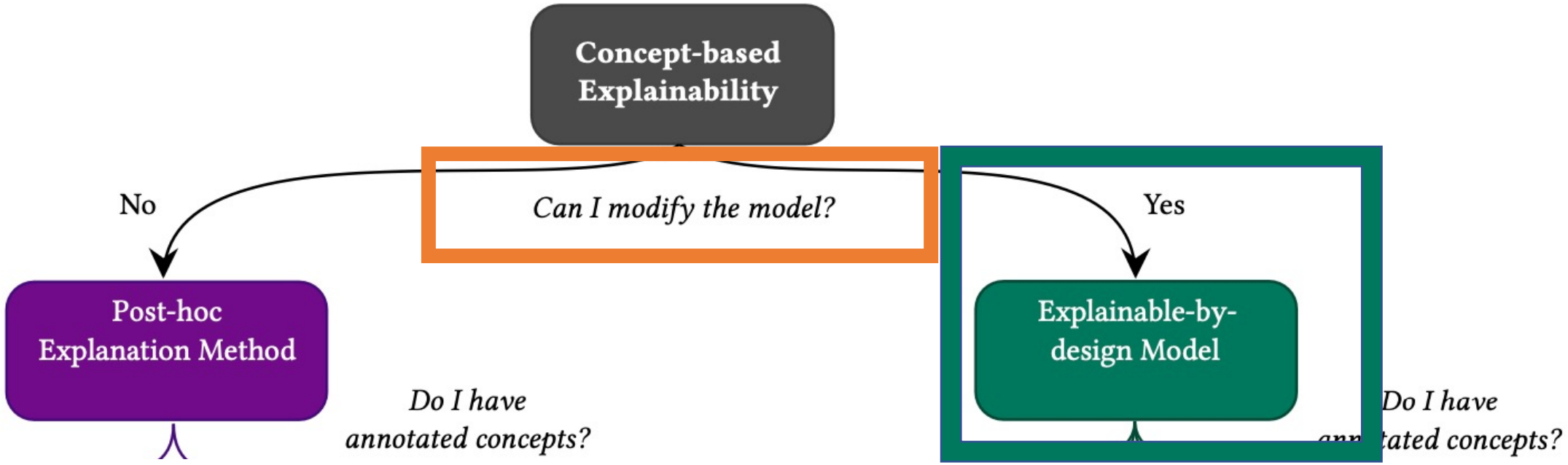
Pros.

- They **do not compromise the learning capacity** of a model
- They can provide more interpretable explanations than standard post-hoc methods

Cons.

- **Cannot ensure the network really knows the concepts**
  - It has not been trained for that
  - Not interpretable by design

C-XAI Taxonomy



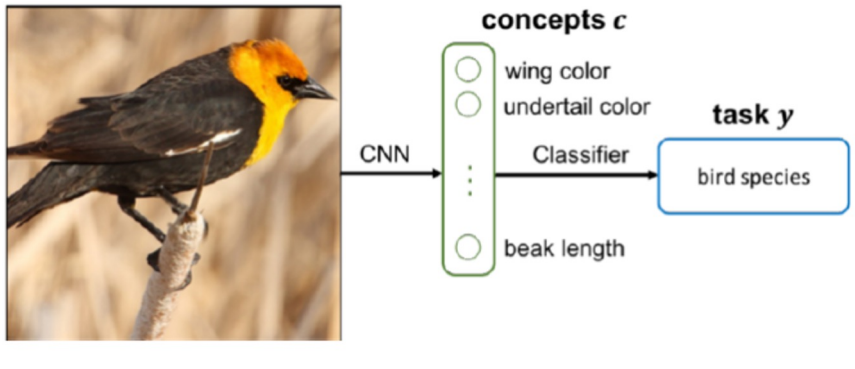
Poeta E., et al., "Concept-based Explainable Artificial Intelligence: A Survey" (2024)

Explainable-by-design Concept-based Models

Directly **train concept-aware models**

- Neural models with an **explicit concept representation**, typically as an intermediate layer

- Task predictions are influenced by the predicted concepts



Explainable-by-design Concept-based Models

Pros.

- They can be regarded as **inherently transparent models** as they provide **node-concept association** by-design

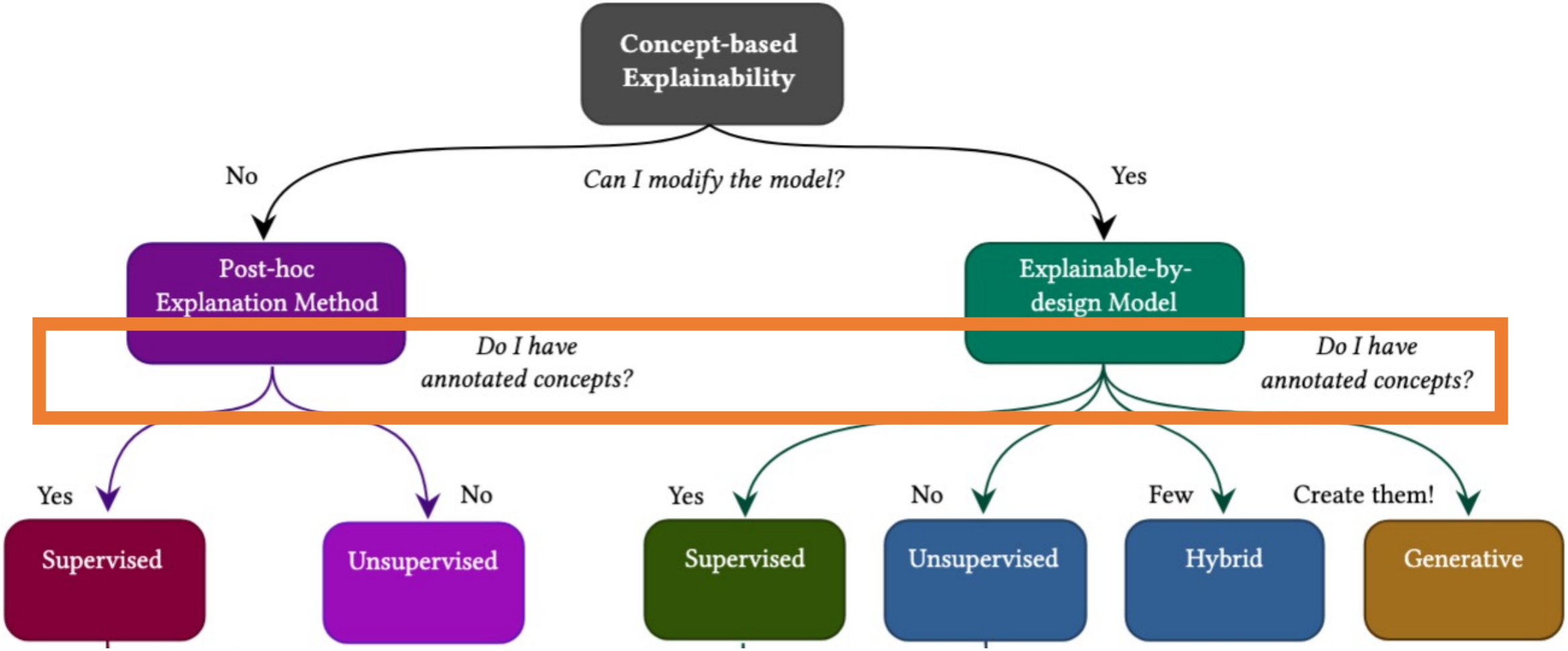
Cons.

- They need ad-hoc training
- Predicting concepts **might reduce network task performance**

Taxonomy of C-XAI

- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations , Node-Concept Association , Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**

C-XAI Taxonomy



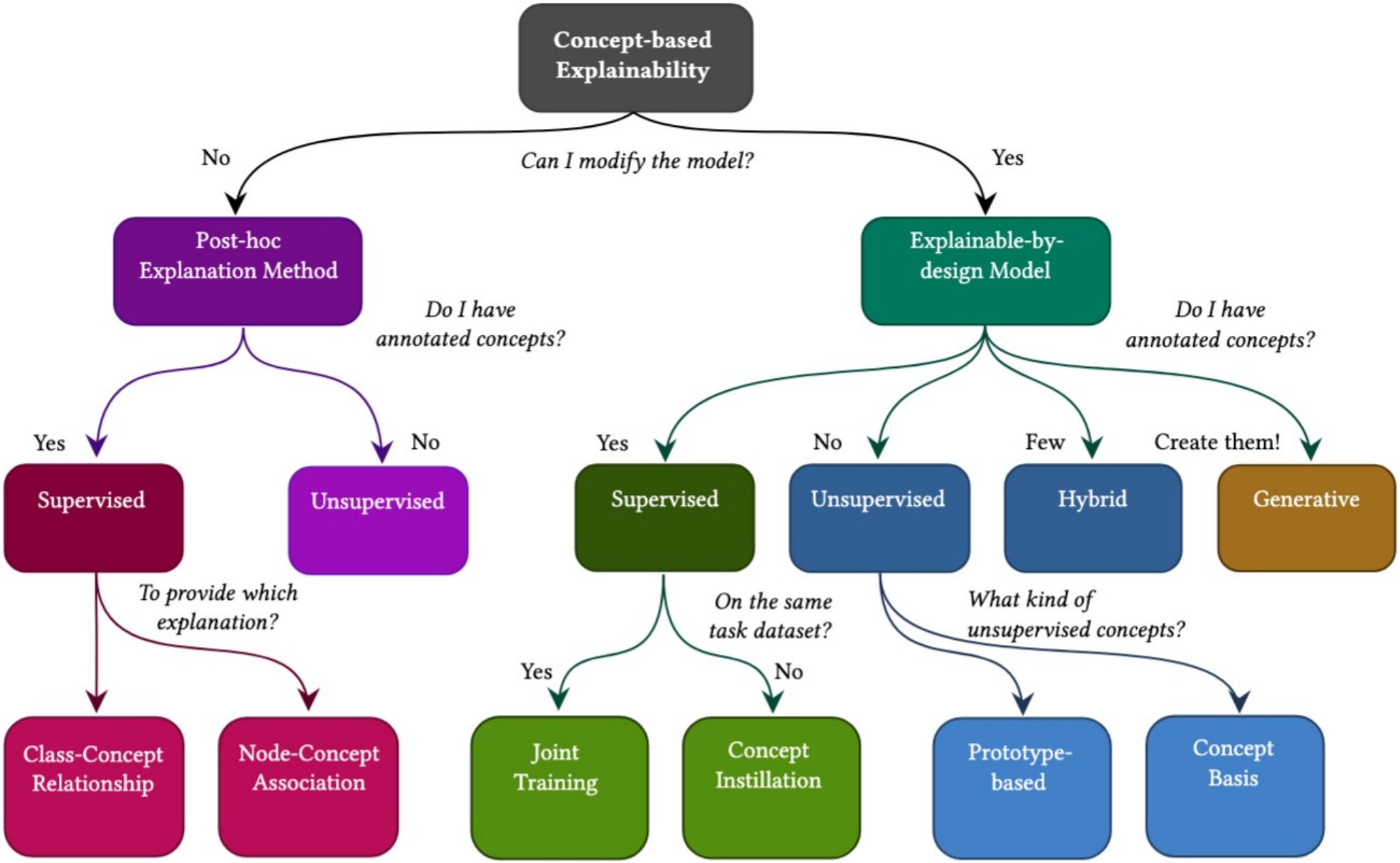
Concept annotation

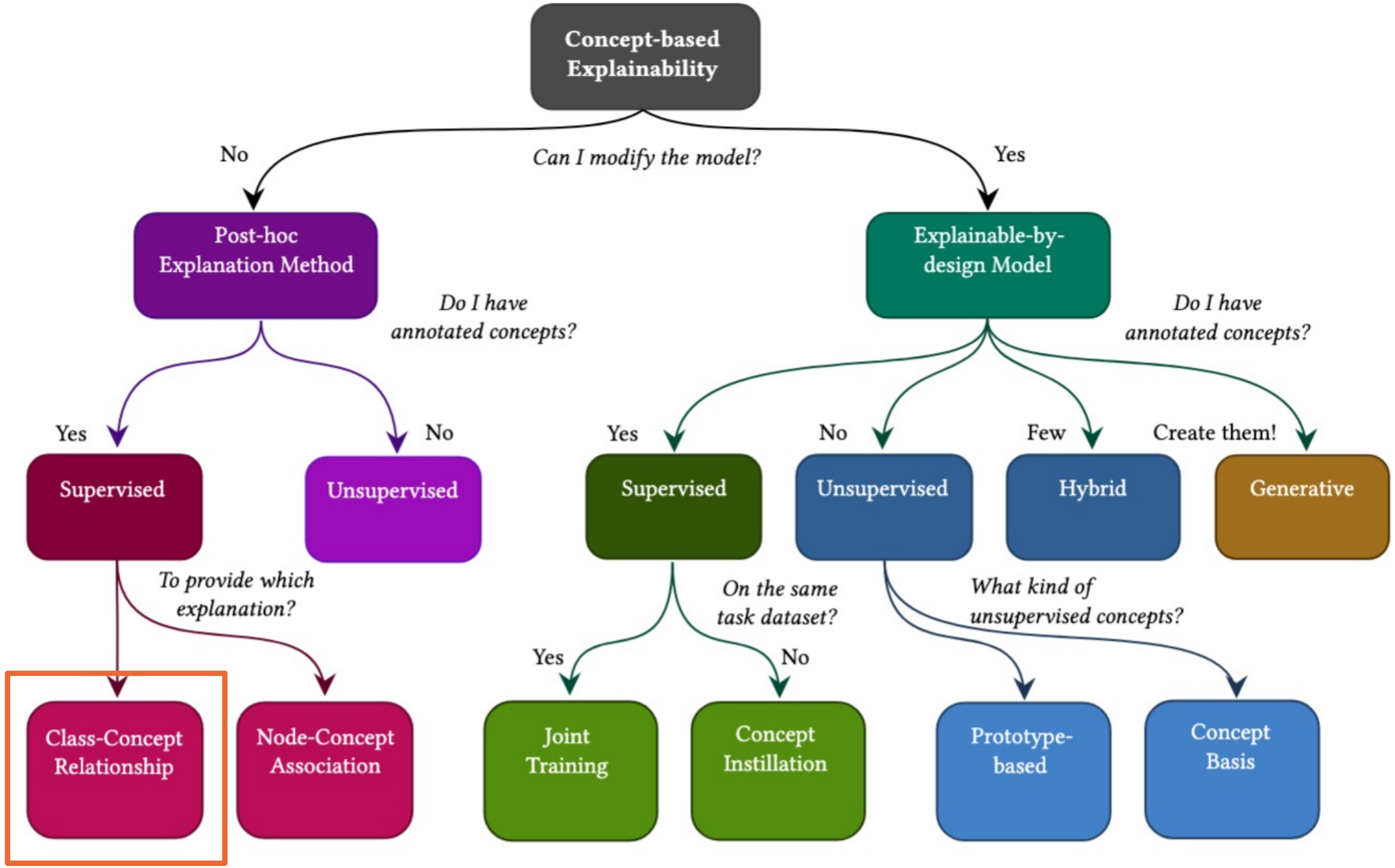
Do I have annotations?

- **Supervised methods** — *Yes*
  Use dataset **annotated with concepts** related to the task
- **Unsupervised approaches** — *No*
  Extract concepts from the same data **automatically**
- **Hybrid (by-design) approaches** — *Few*
  Leverage **few supervised** concepts **and unsupervised** (extracted) ones;
- **Generative methods** — *Generate!*
  Create concepts by means of an external generative model

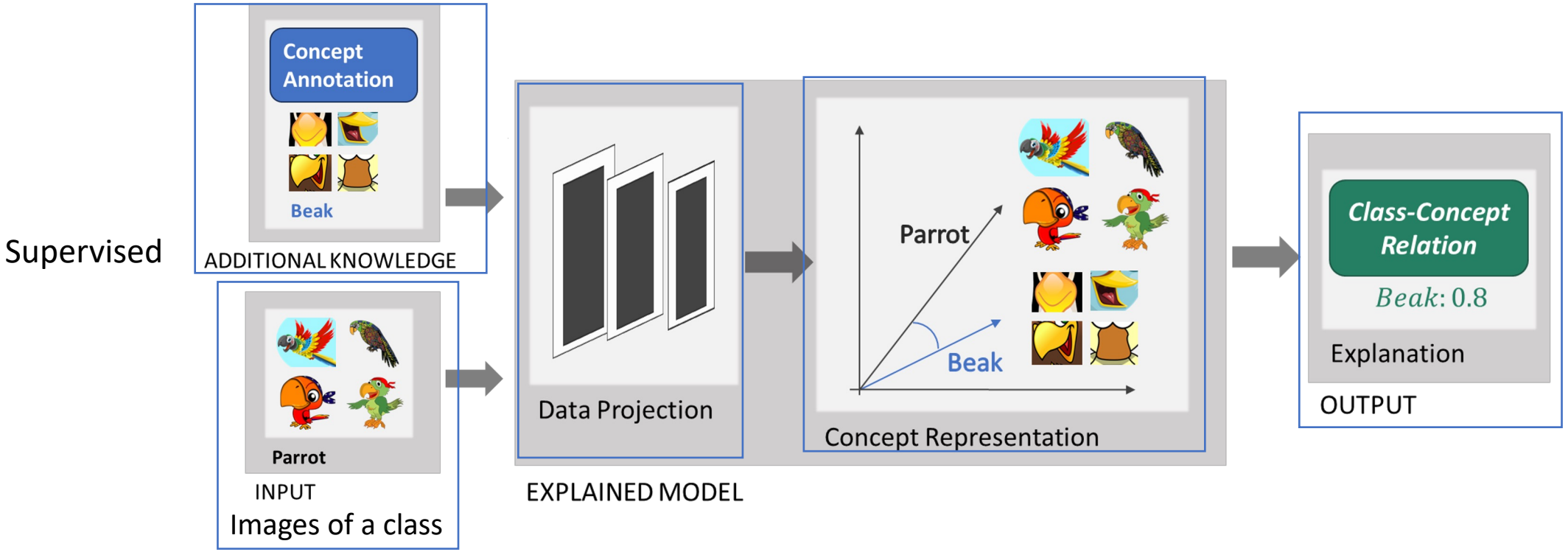
Taxonomy of C-XAI

- **Definition of concept**
- **Types of concepts**
  - **Symbolic Concepts, Unsupervised Concept Basis, Prototypes, Textual Concepts**
- **Types of concept-based explanations**
  - **Class-Concept Relations , Node-Concept Association , Concept-Visualization**
- **Stages of Explainability**
  - Post-hoc, By design
- **Annotation availability**
  - Supervised/annotated concepts, Unsupervised, Hybrid, Generated
- **Overview of C-XAI approaches**


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

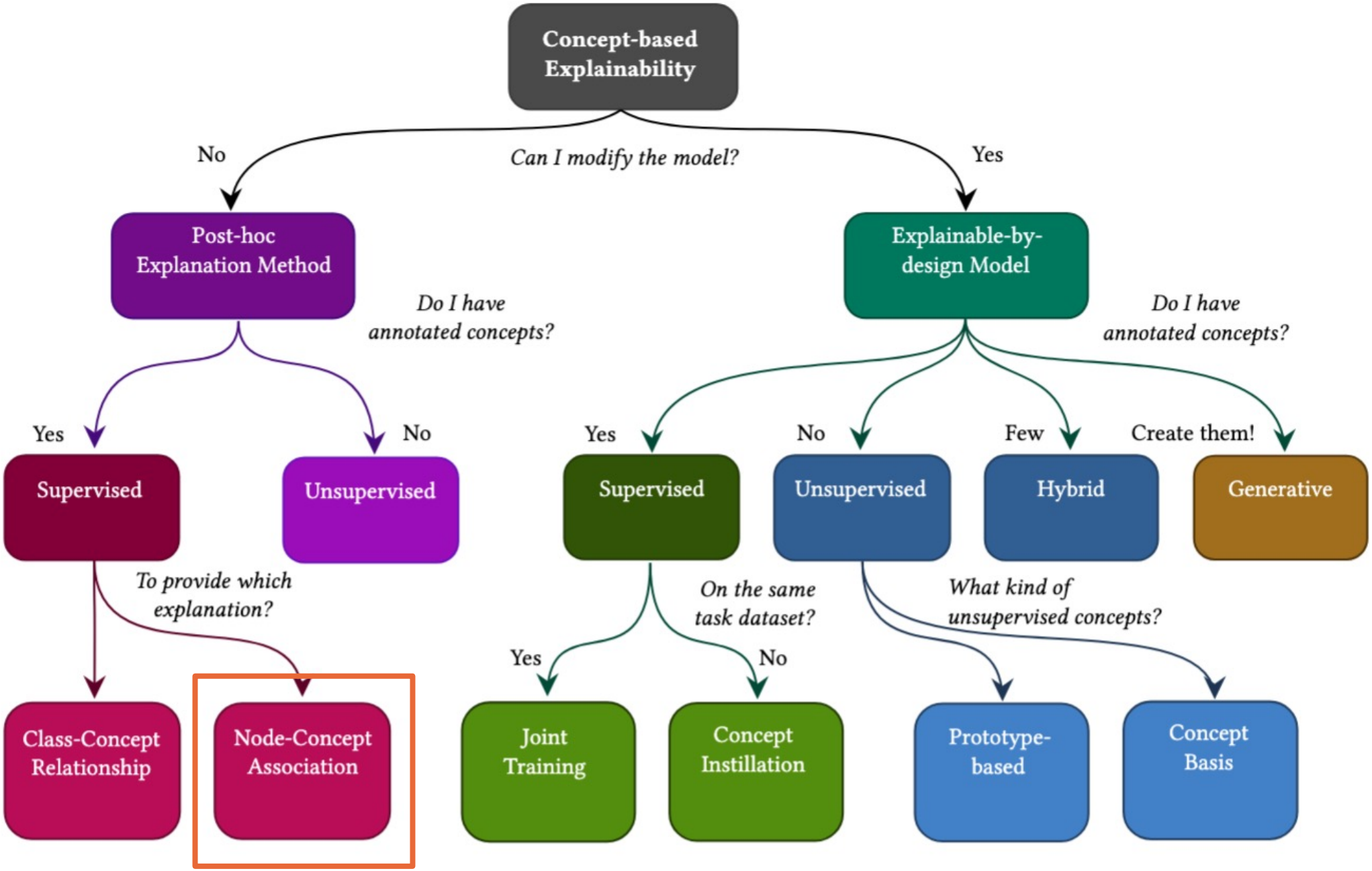

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Post-hoc Supervised method providing class-concept relations

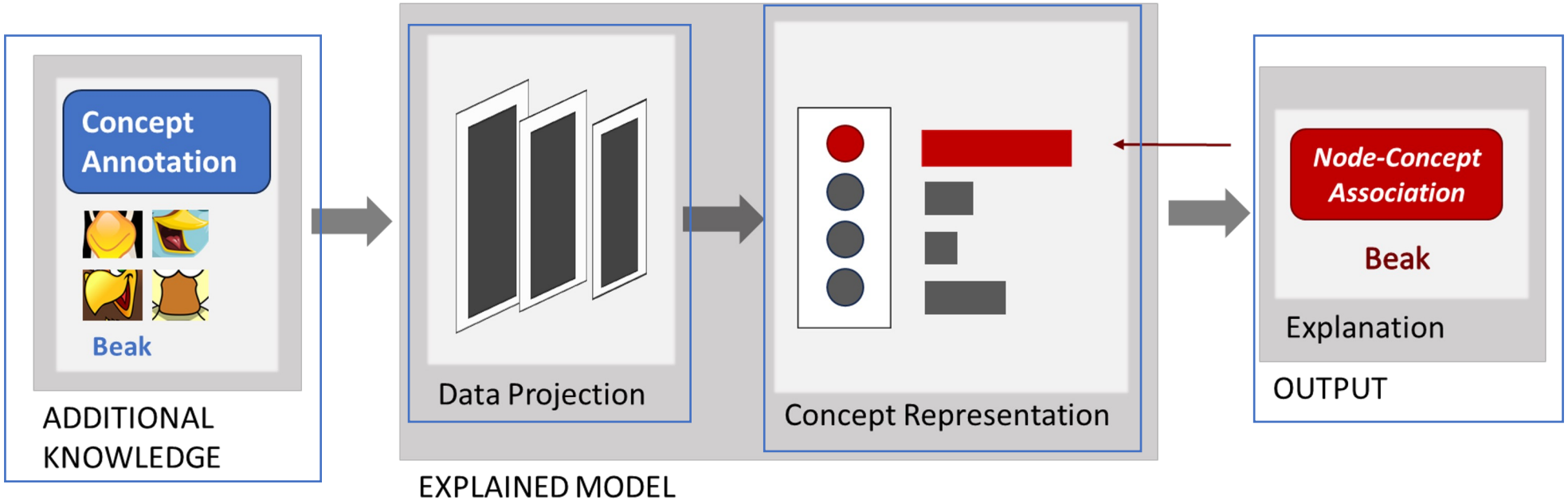


- **Analyze the relationship of the representation of the concept data** in the model latent space **with the output class**

e.g., TCAV - Kim, Been, et al. "*Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV).*" ICML (2018).

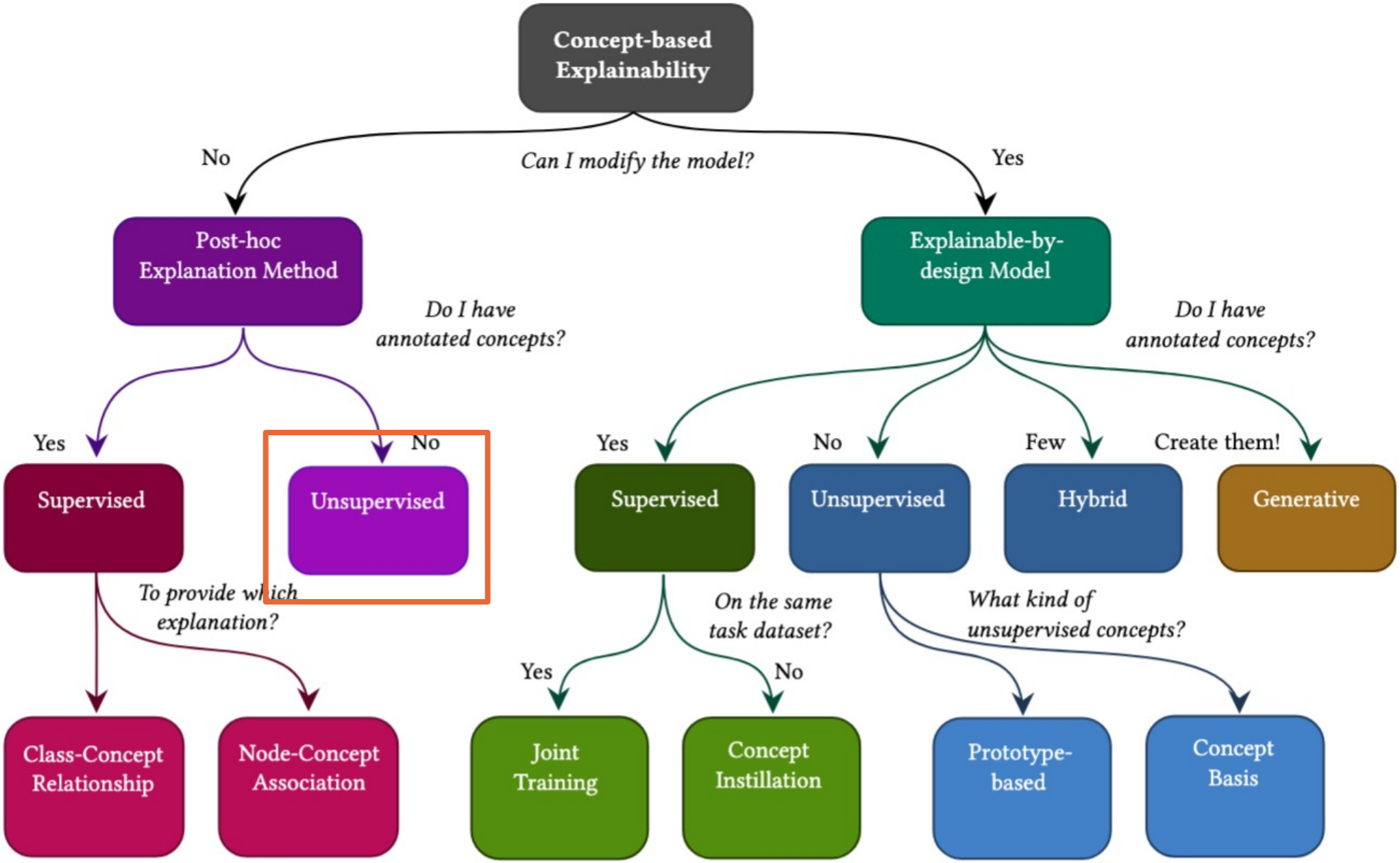

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Post-hoc Supervised method providing node-concept association

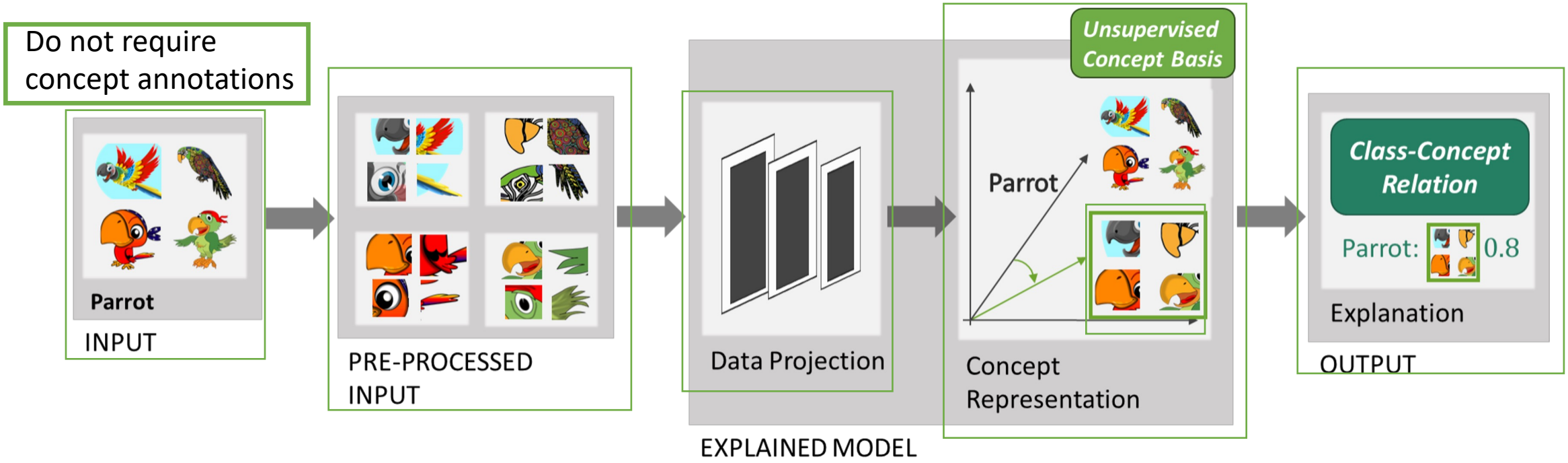


- **Analyze the activations** of the hidden nodes **when fed with the concept annotation data**
- They **associate to each node the concept** for which it is activated the most

e.g., Network dissection - Bau, David, et al. "*Network dissection: Quantifying interpretability of deep visual* representations." CVPR (2017).

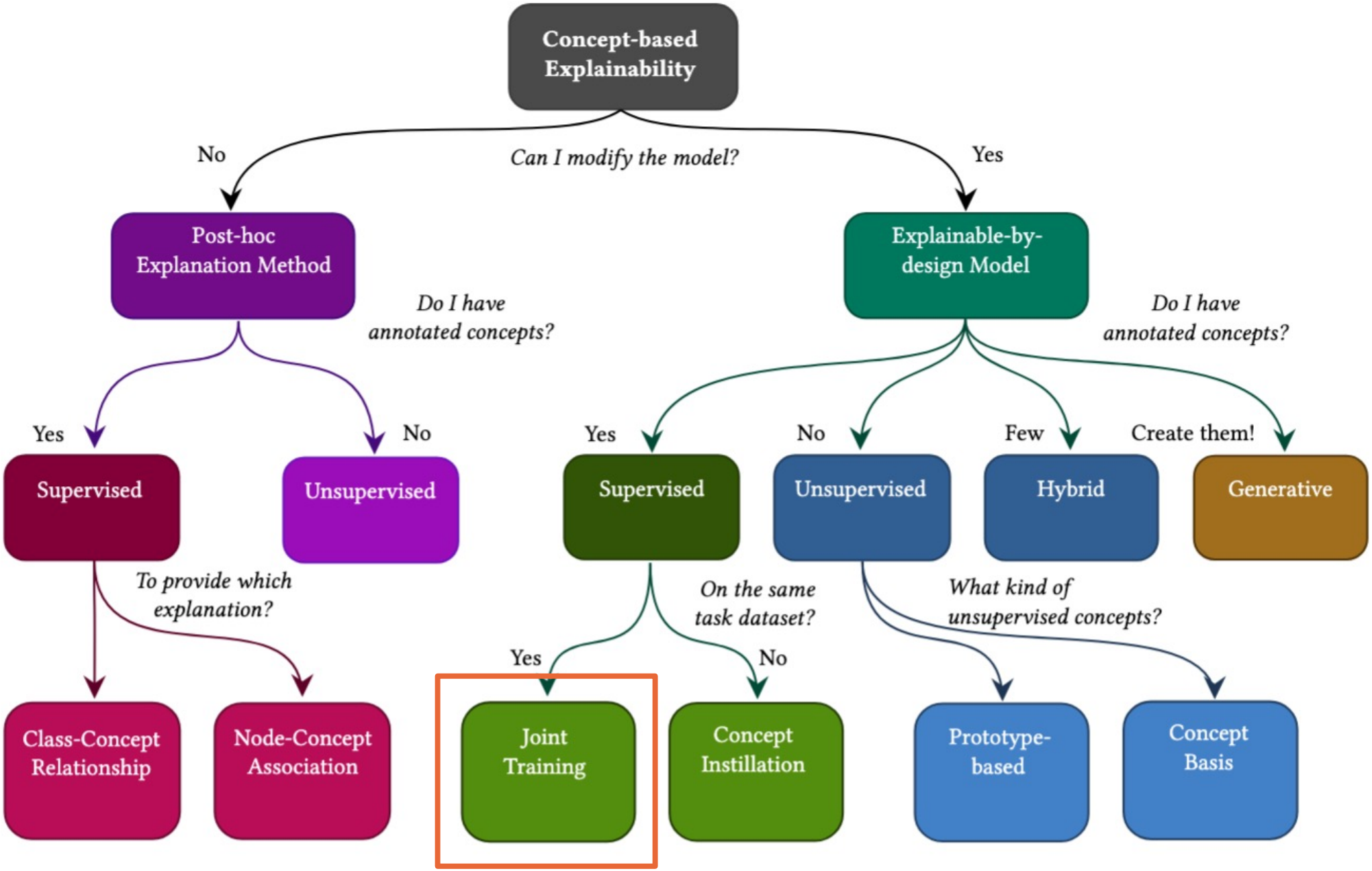

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Post-hoc Unsupervised method providing class-concept relation association

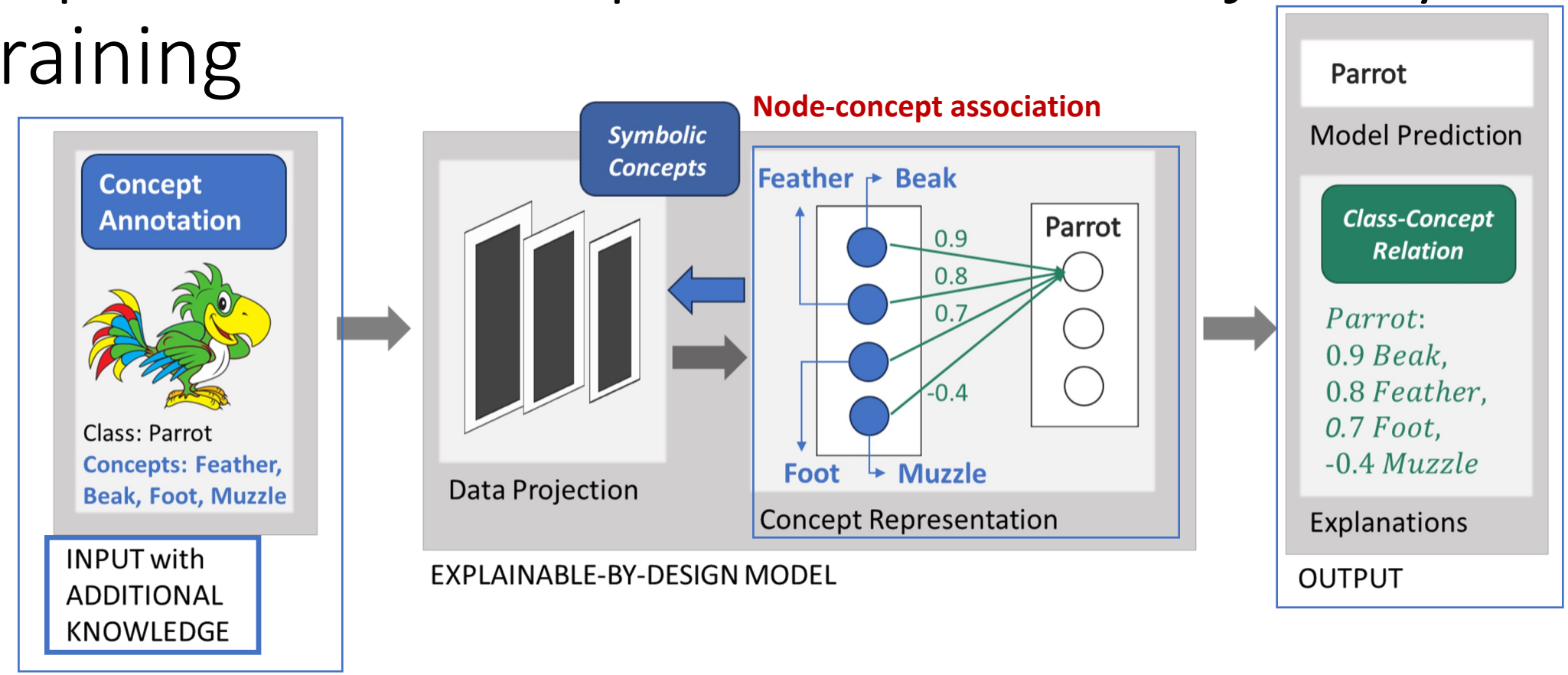


- Pre-process the inputs in parts of the samples
- Clusterize projections of the processed inputs in the latent space → clusters are **unsupervised concepts**
- **Analyze the relation of unsupervised concepts with output classes/predictions**

e.g., Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." NeurIPS (2019).

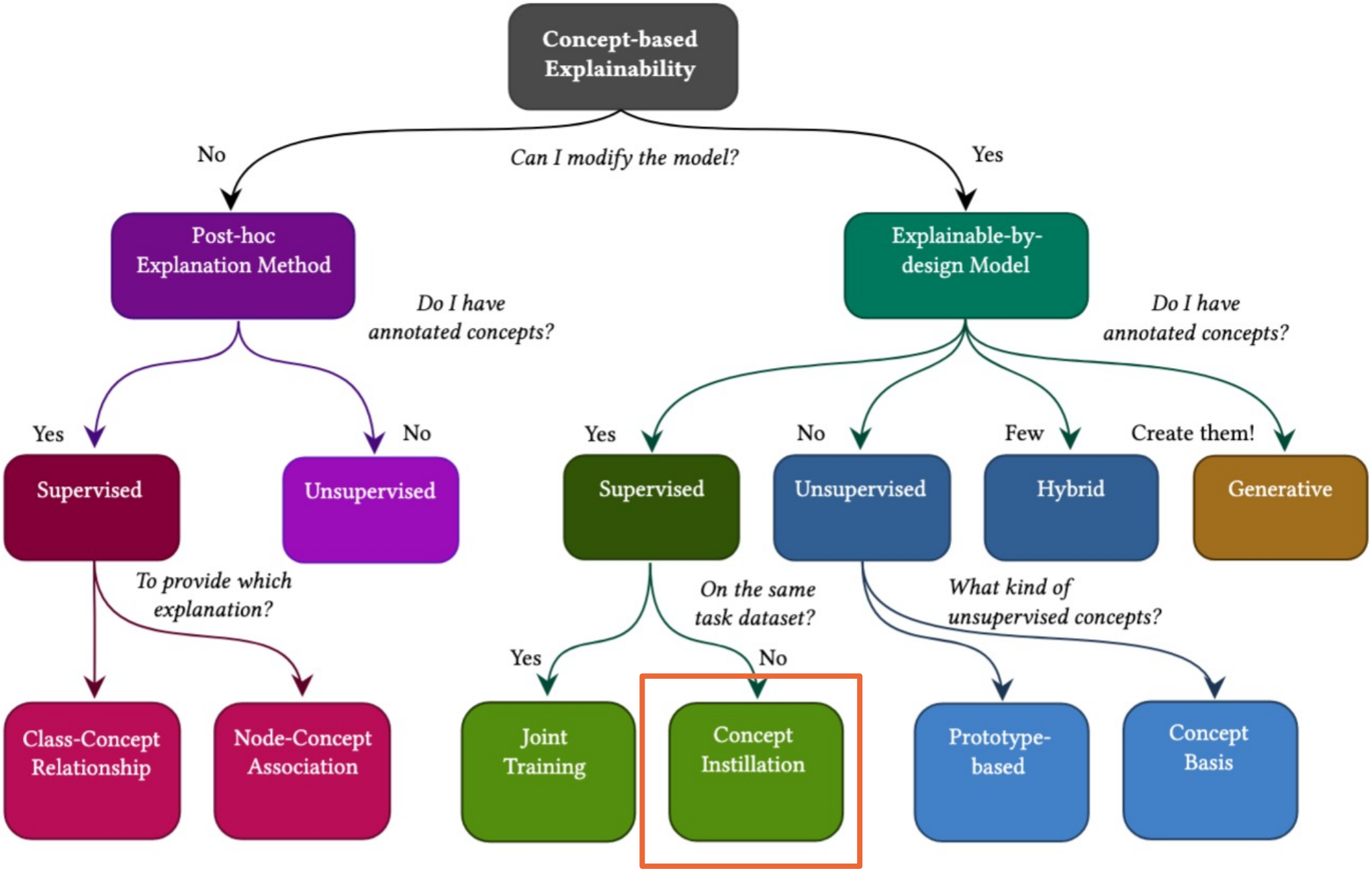

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Supervised Concept-based models jointly training

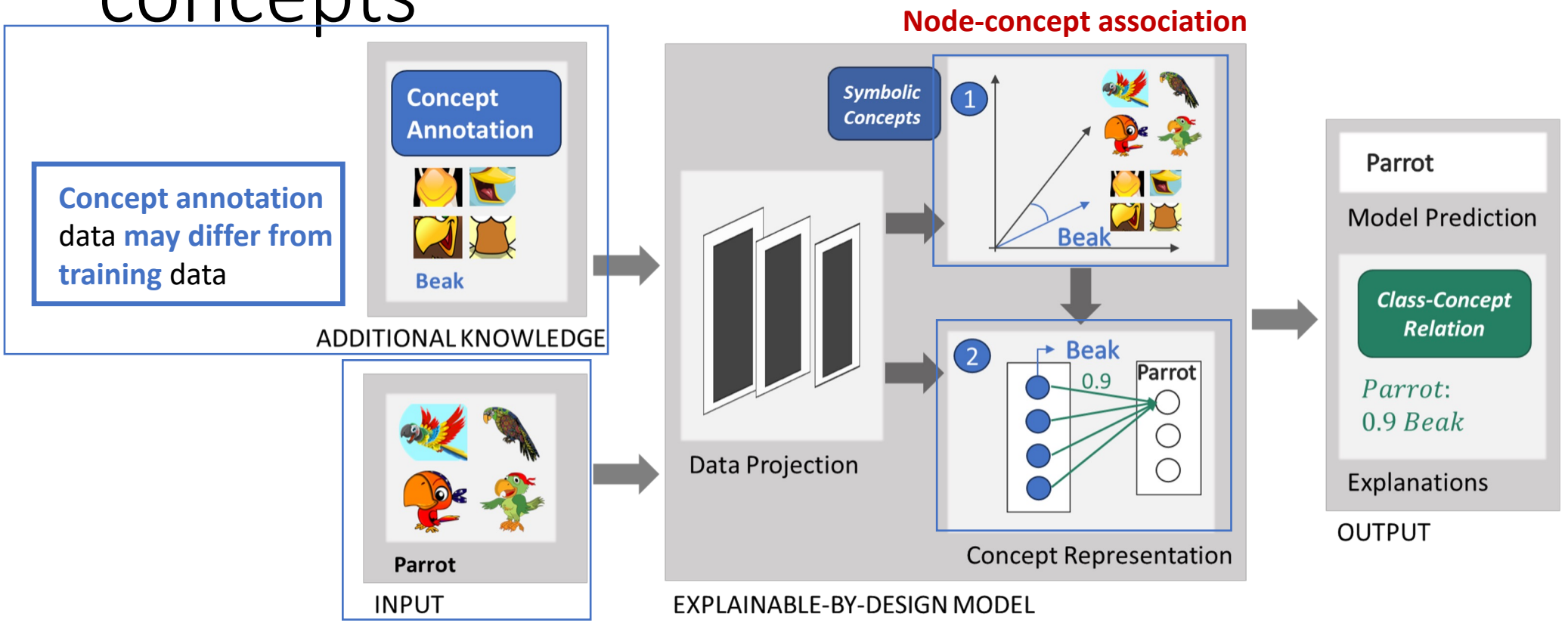


- **Train** a model with a **hidden layer predicting the concepts**
- The **predicted concepts** are used **to make the final prediction**
  - If the task predictor is a white box model also **extract class-concept relations**

e.g., CBM - Koh, Pang Wei, et al. "*Concept bottleneck models*." ICML (2020)

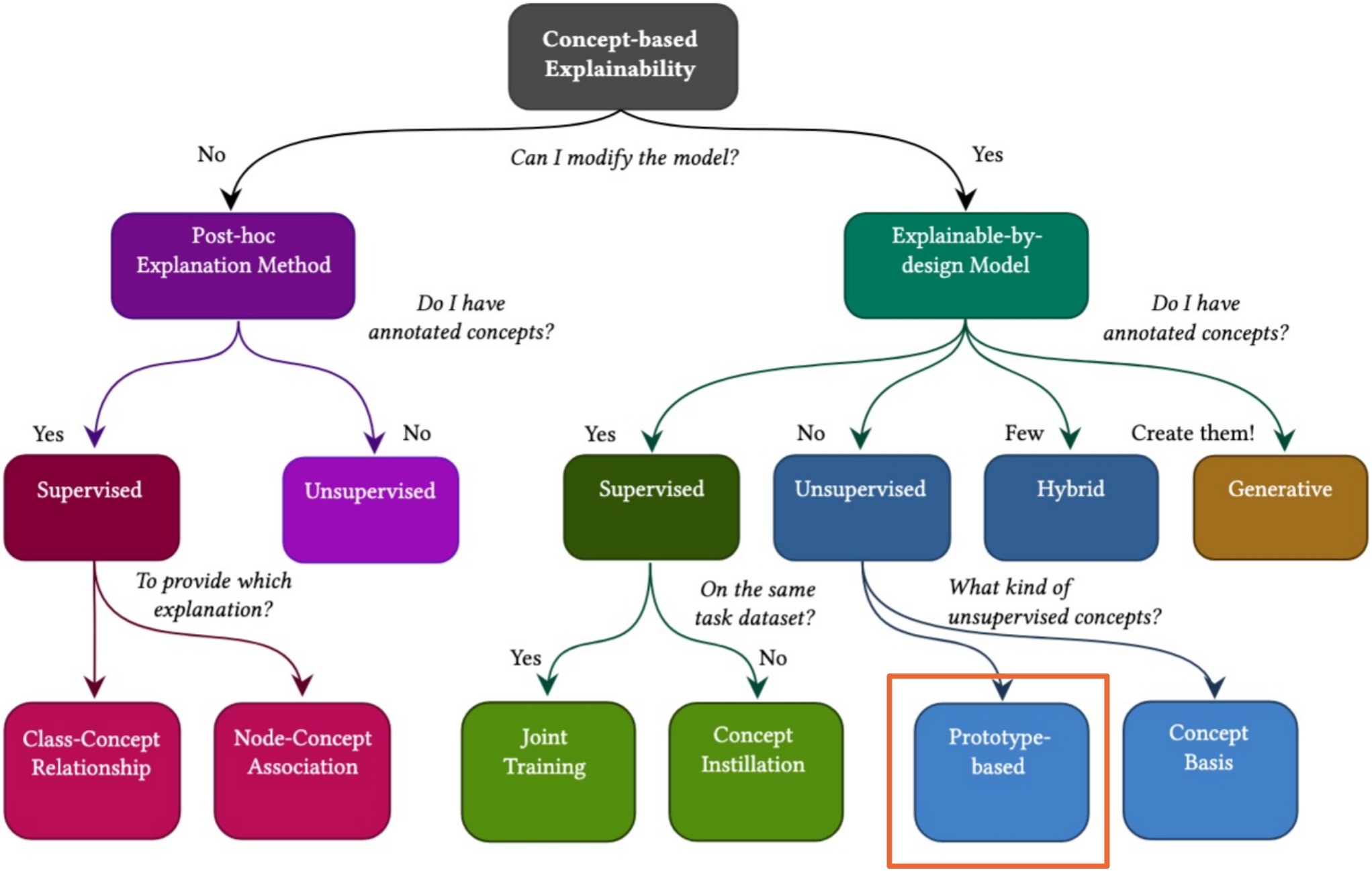

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Supervised Concept-based models instilling concepts

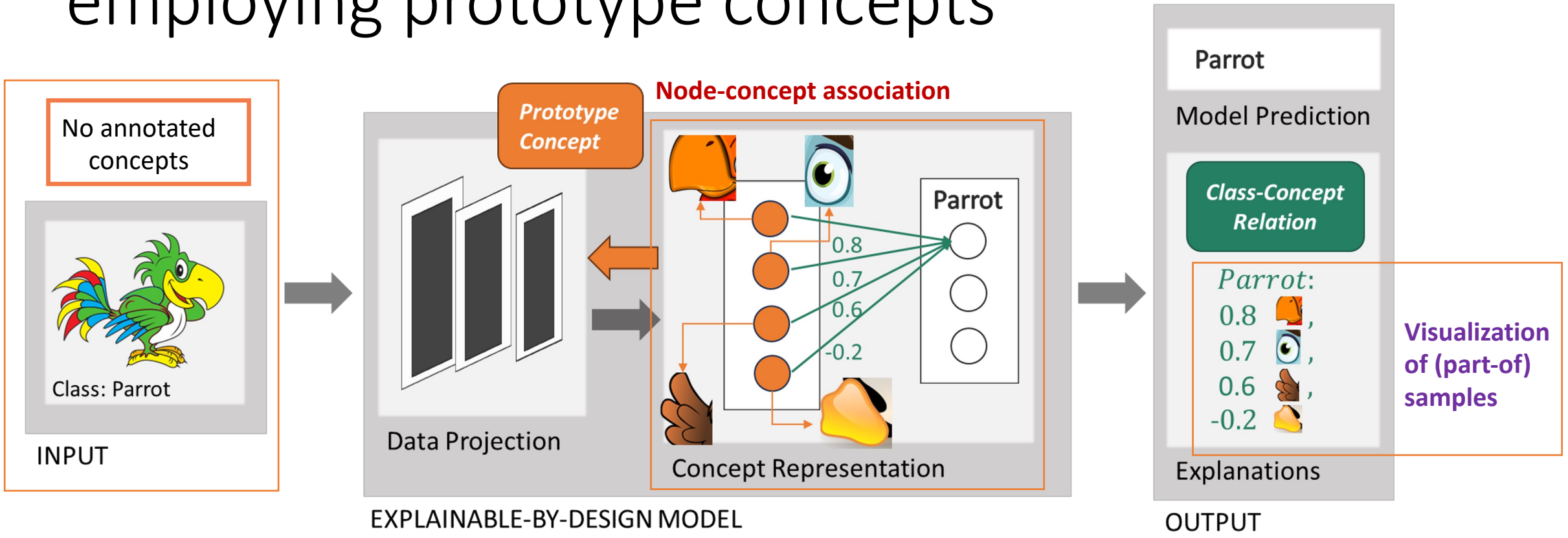


- **Turn a black-box model into an explainable-by-design** one:
  - **Concept instillation:** fine-tune a layer to predict the concepts
  - Keep **training the top of the network to predict the classes**

e.g., Chen, Zhi, et al. "*Concept whitening for interpretable image recognition*." Nature Machine Intelligence (2020).


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

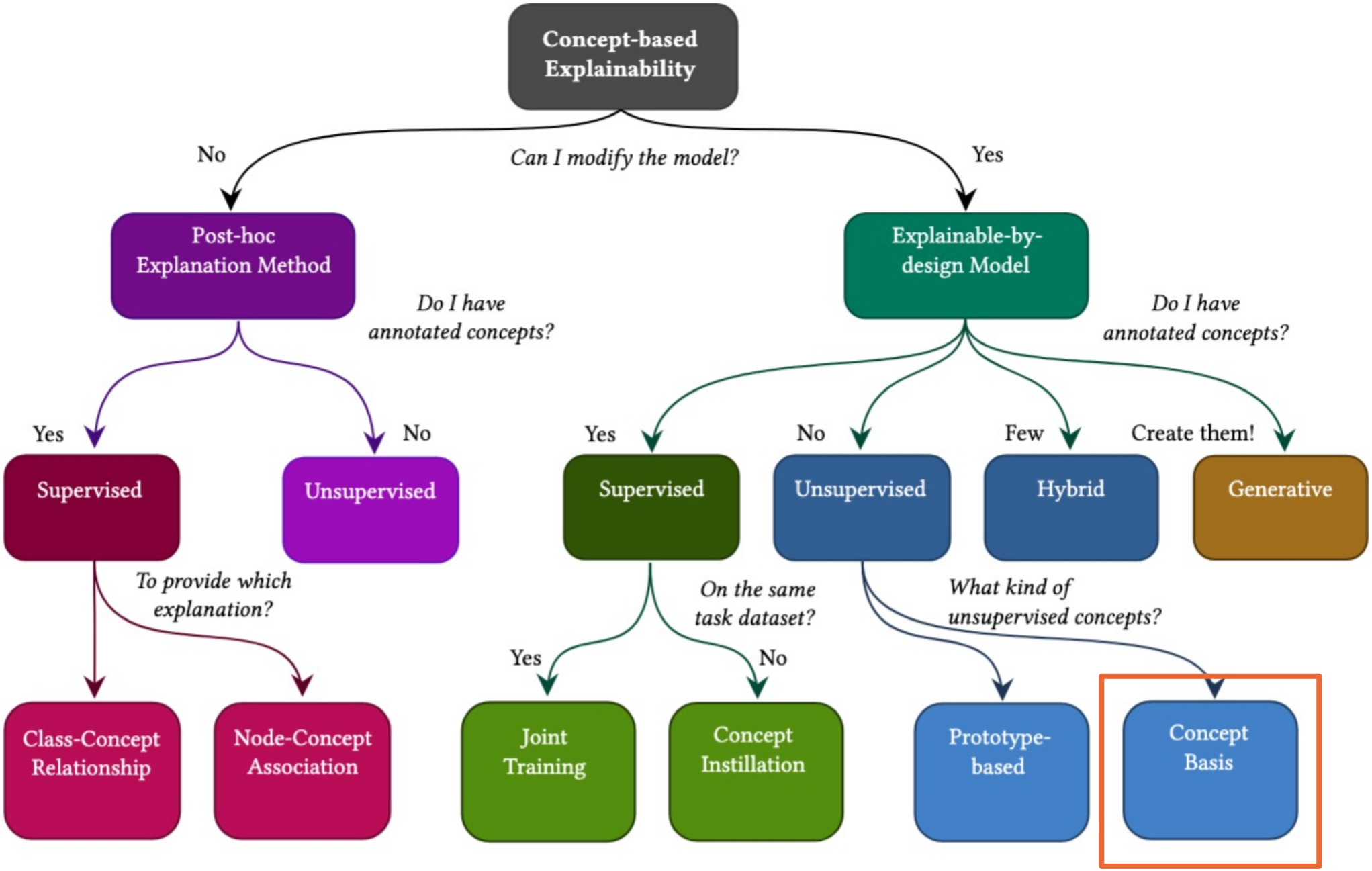
Unsupervised Concept-based models employing prototype concepts



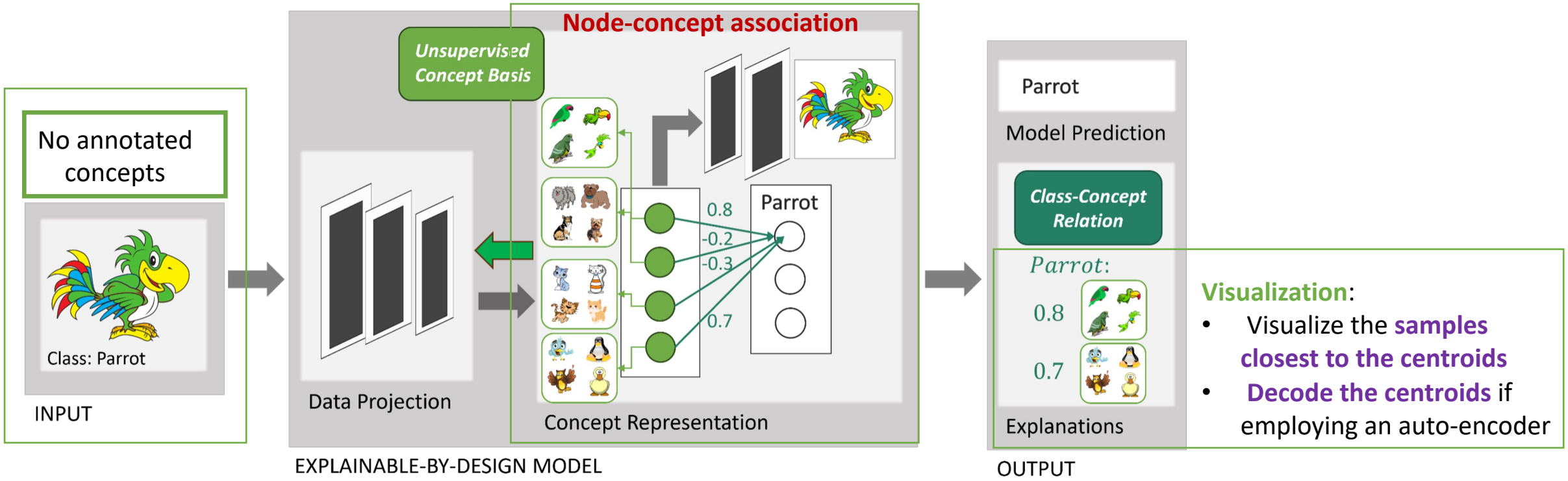
Train the network to:

- Learn to predict the output classes
- **Encode** in the hidden layers **representative (part-of) train examples**

e.g., Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." NeurIPS (2019).


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

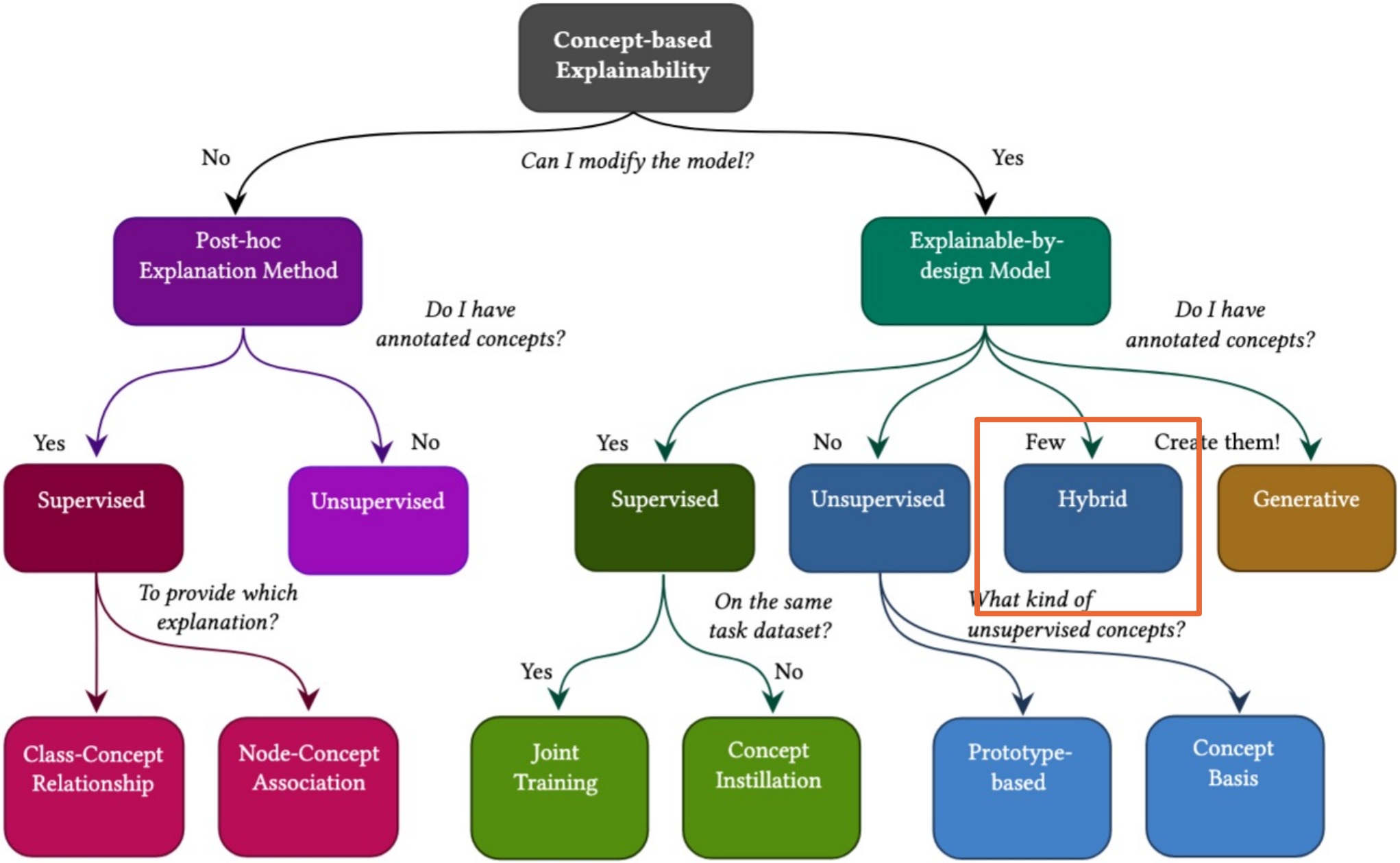
Unsupervised Concept-based models employing unsupervised concept basis



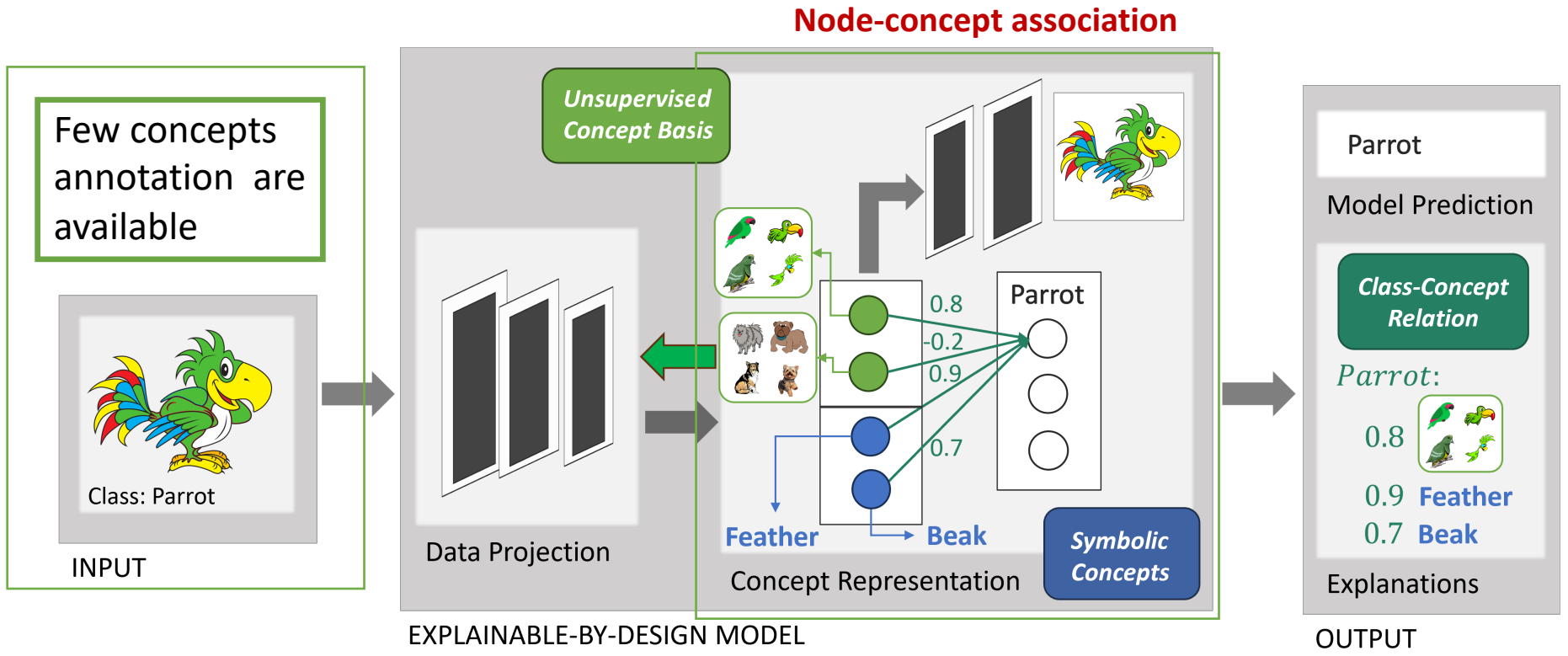
They train the network to:

- Learn to predict the output class
- **Create clusters of samples in the latent representation**, associating clusters to nodes

e.g., Alvarez Melis, David, and Tommi Jaakkola. "*Towards robust interpretability with self-explaining neural networks*." NeurIPS (2018).


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Crea te them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

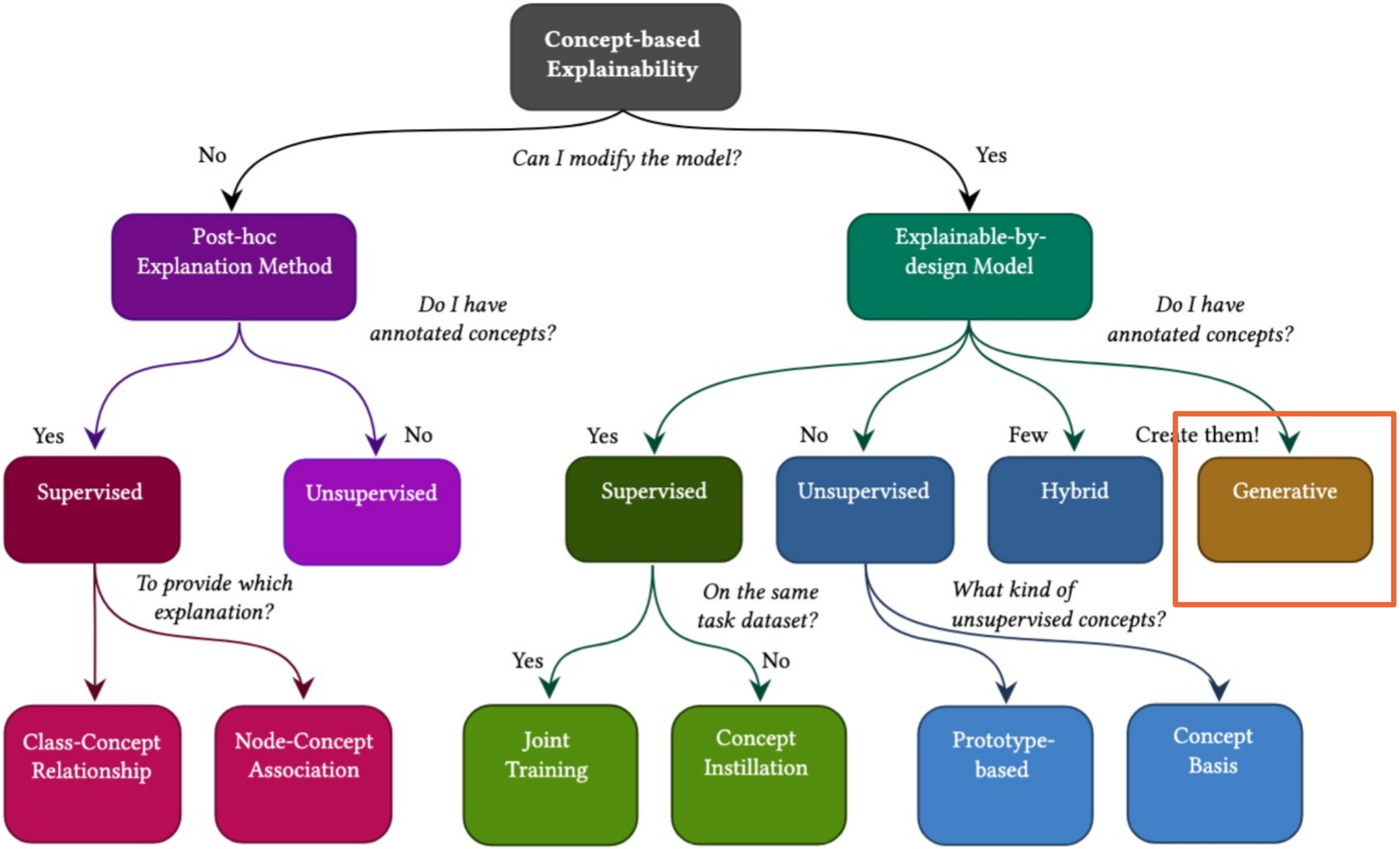
Hybrid Concept based Models



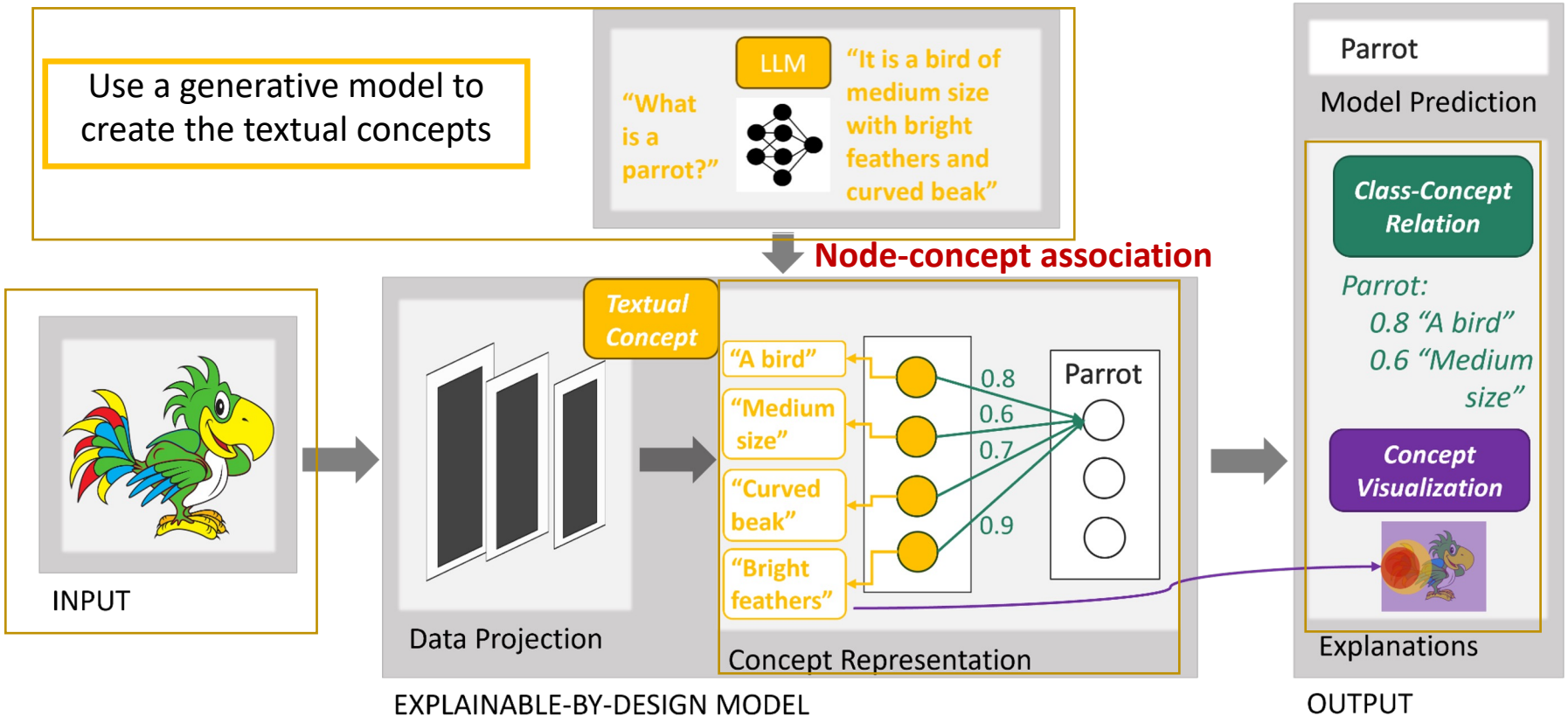
Train the network to:

- Learn to **predict** a **given set of concepts** with a **subset of neurons**
- Create **a clusterized representation** in the **remaining neurons**

e.g., Sawada, Yoshihide, and Keigo Nakamura. "*Concept bottleneck model with additional unsupervised concepts*." IEEE Access 10 (2022): 41758-41765.


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Generative concept-based models

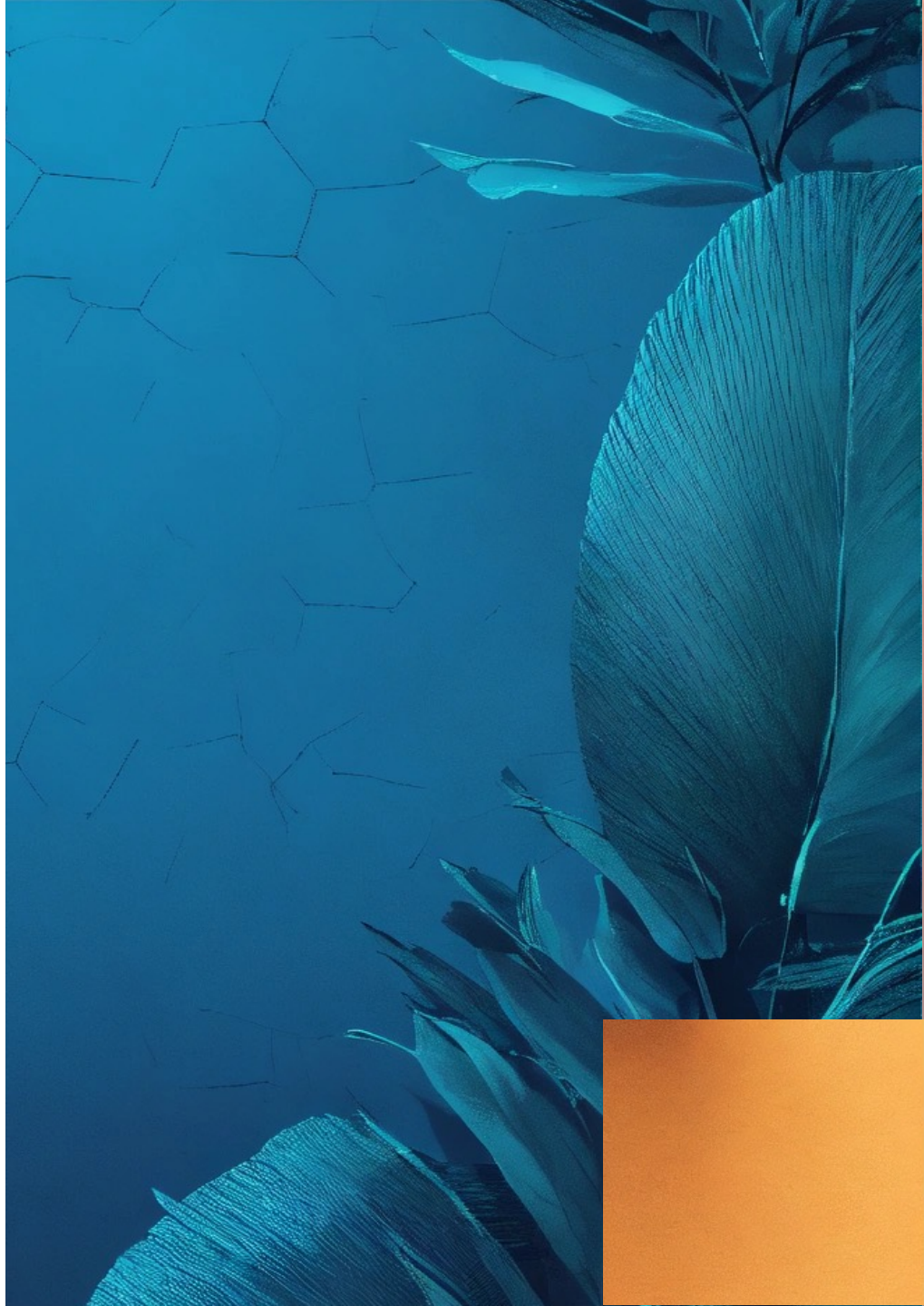


- The **embeddings of textual concepts** are **aligned to the latent input representation** to produce **concept scores**
- The **concept scores** are used to **provide the final classification**

e.g., Label Free - Oikarinen, Tuomas, et al. "*Label-free concept bottleneck models*." ICLR (2023).

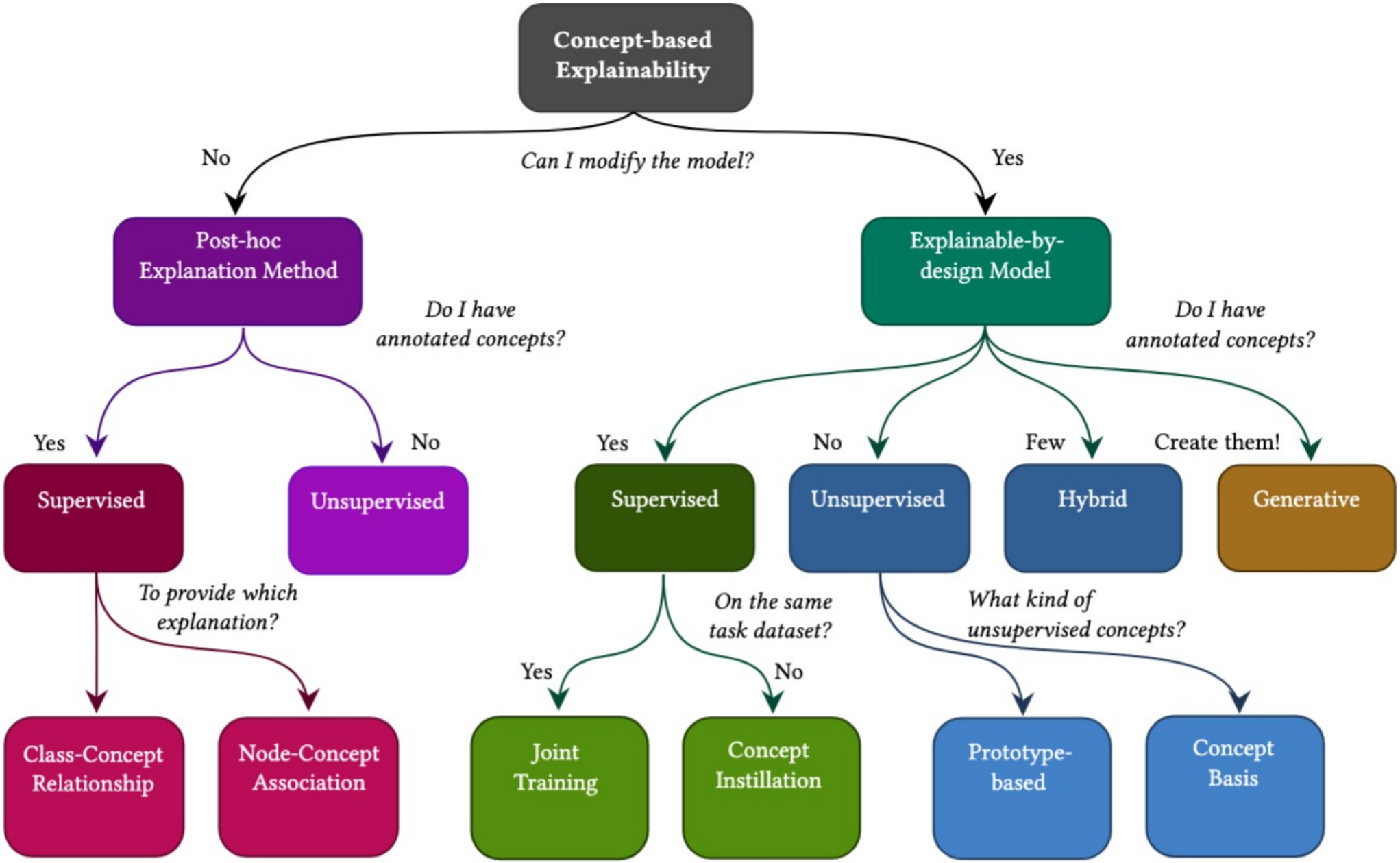
Tutorial structure

- Introduction to XAI techniques
- Taxonomy of C-XAI
- **Post-hoc vs By-design C-XAI approaches**
- Evaluation, Resources and applications
- Hands-on session



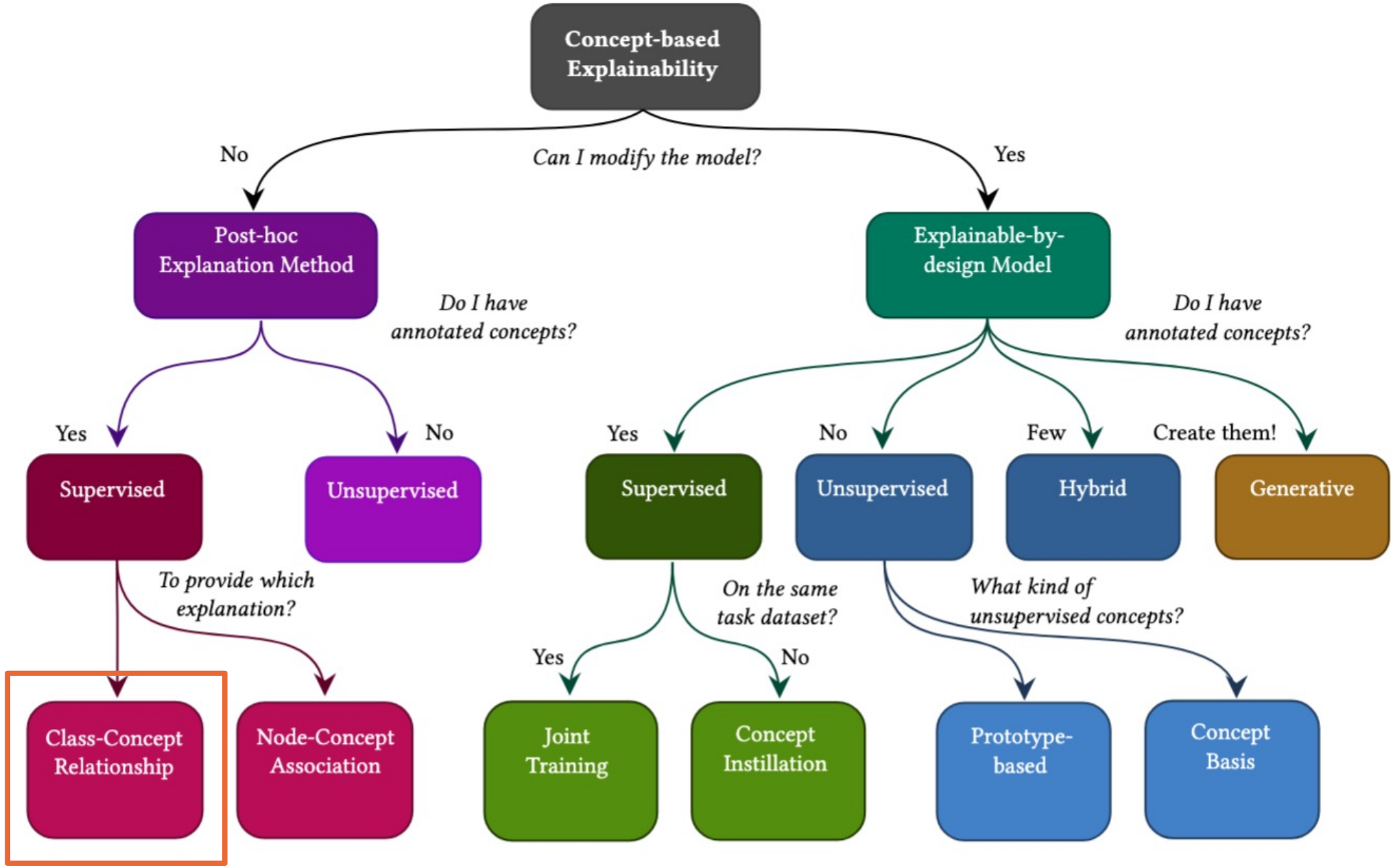
OUTLINE

1. Testing with Concept Activation Vector (T-CAV)
2. Concept Bottleneck Models (CBM)
3. Concept Embedding Models (CEM)
4. Self-Explainable Neural Network (SENN)
5. Label-free Concept Bottleneck Models (LF-CBM)


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Credits for some of these slides goes to Kim et al.

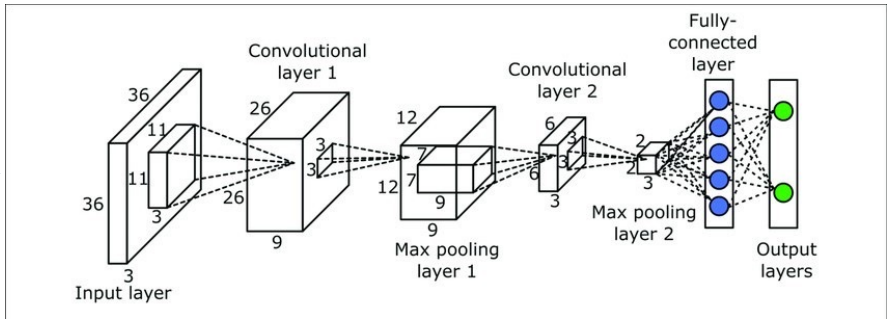
1. Testing with Concept Activation Vectors (T-CAV)


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Example: Post-training explanation

Given Image
(Cash-machine)

Trained ML model
(e.g. GoogleNet)

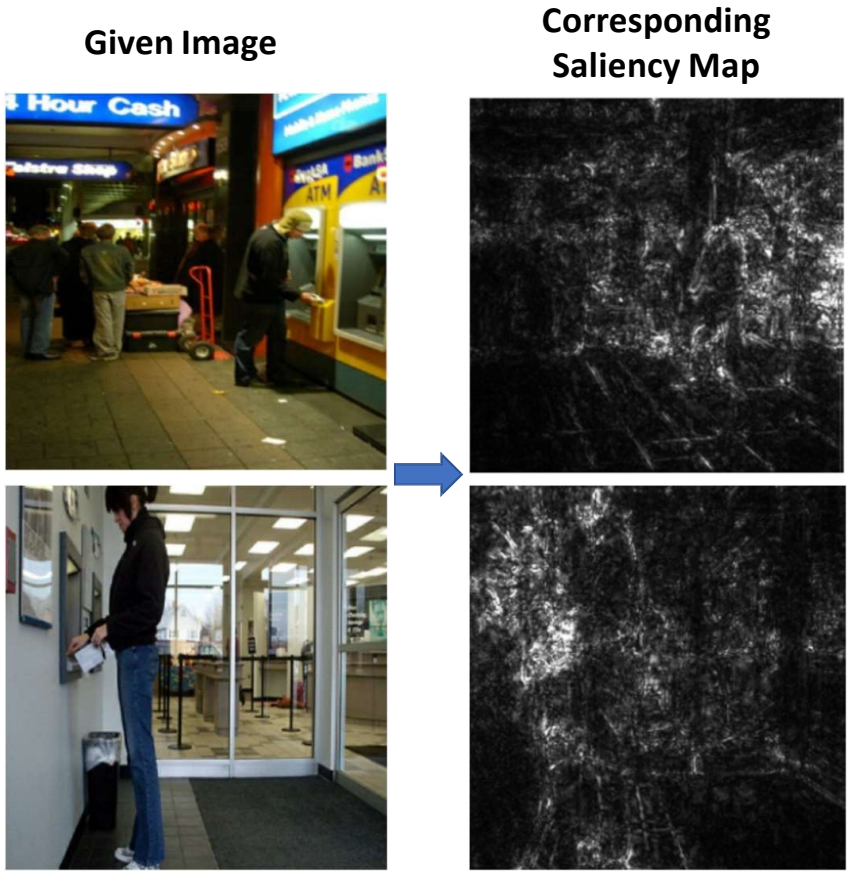


Predicted as
Cash-machine

$p(z)$

- **Why was this a cash machine?**

Problem Objective



- Did the **'human'** concept matters?
- Did the **'paper'** concept matters?
- Did the **'ATM'** or **'Cash'** concept matters?

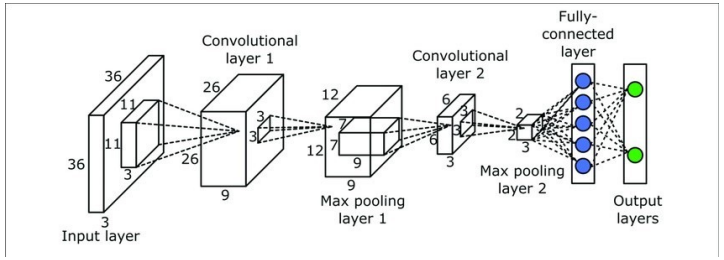
TCAV objective:

Quantitatively measure how important are **"user- chosen concepts"**

TCAV Example 1 (Zebra)

Given Image
(Zebra)

Trained ML model
(e.g. GoogleNet)



Predicted as
Zebra

$p(z)$

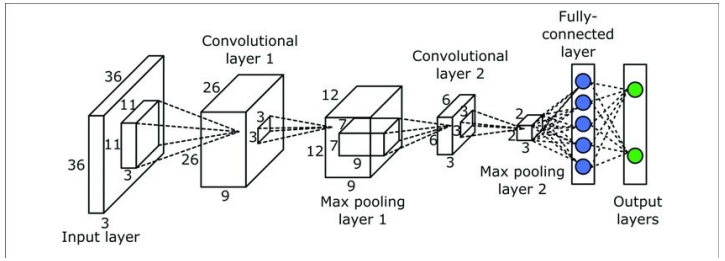
Was **Stripe concept** important to this **zebra** image classifier?

TCAV tells that **Stripe has a positive importance** for the classification of **zebras**

TCAV Example 2 (Doctor)

Given Image
(Doctors)

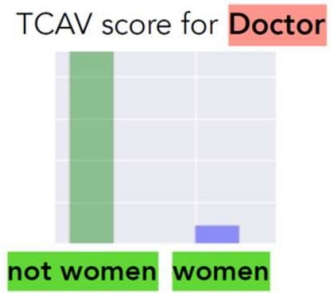
Trained ML model
(e.g. GoogleNet)



Predicted as
Doctor

$p(z)$

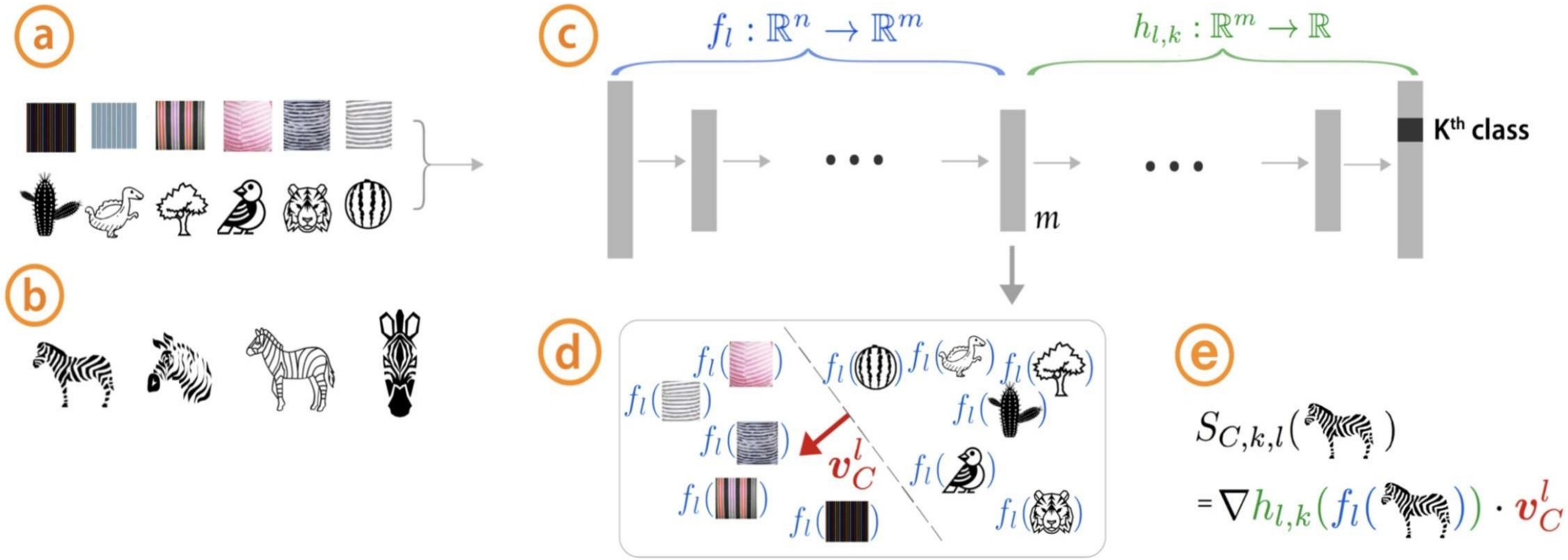
Was **Woman concept** important to this **doctor** image classifier?



TCAV tells that **Woman has a negative importance** for the classification of doctors

BIAS IDENTIFICATION!

TCAV: Overview



TCAV components

a) A dataset annotated with both **examples of concepts** and **random images**

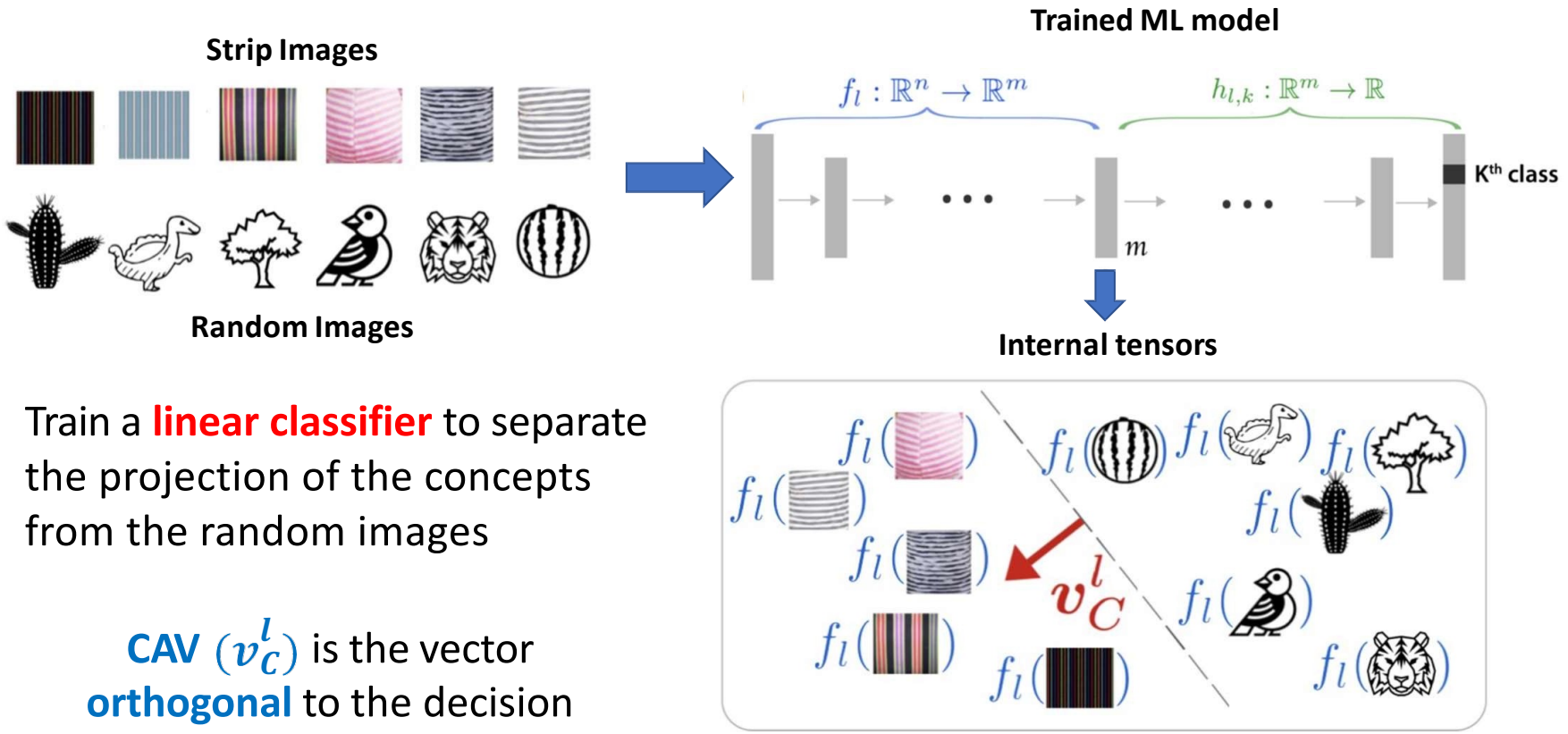
b) The dataset with the **original classes**

c) The **model** to explain

d) The Concept Activation Vectors (CAV)

e) The TCAV score showing the **influence** of a concept on a given class

TCAV: (1) How to define CAV?



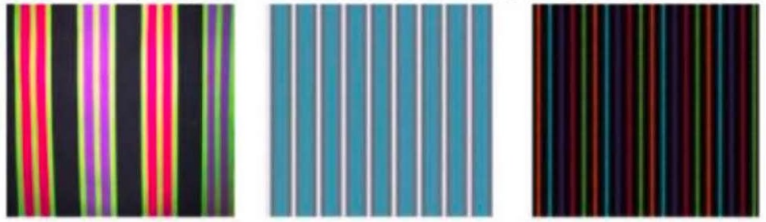
Train a **linear classifier** to separate the projection of the concepts from the random images

CAV (v_C^l) is the vector **orthogonal** to the decision boundary

Sorting Images with CAVs

- Given a set of images (e.g., belonging to the same class)
- Compute the cosine similarity between
  - the latent representation of an image $f_l(x)$
  - the CAV v_C^l of the selected concept

CEO concept: most similar striped images

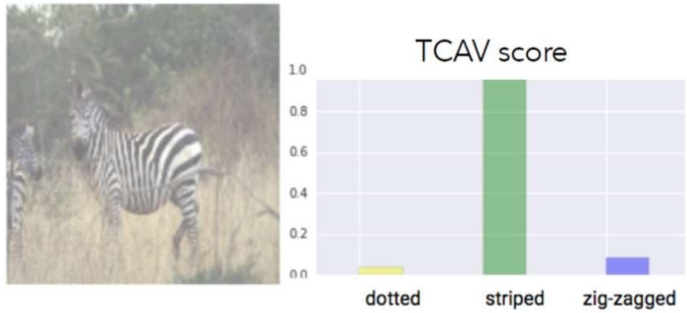


Model Women concept: most similar necktie images

CEO concept: least similar striped images

Model Women concept: least similar necktie images

TCAV: (2) How to compute TCAV scores?



zebra-ness → $\frac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$

striped CAV →

$$S_{C,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon} = \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l, \quad (1)$$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

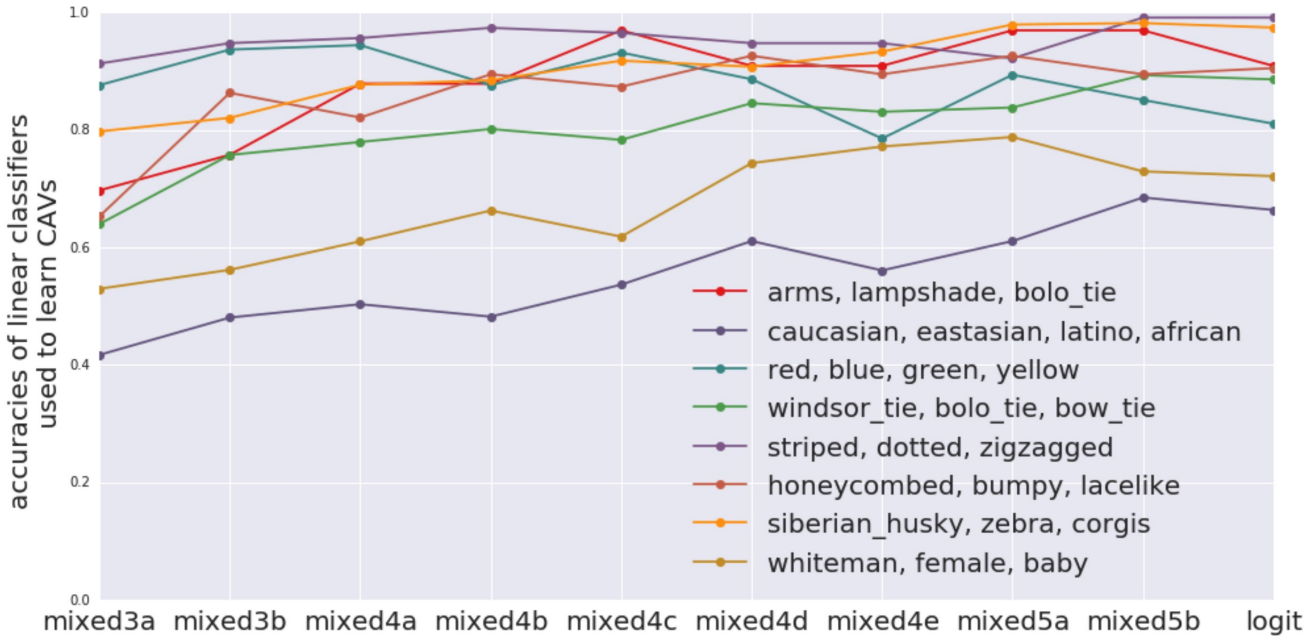
Directional derivative with CAV:

- $S_{C,k,l}(x) > 0$: *positive influence*
- $S_{C,k,l}(x) < 0$: *negative influence*

The TCAV score is the number of class samples having a positive directional derivative w.r.t. the CAV

When and where can concept be learnt?

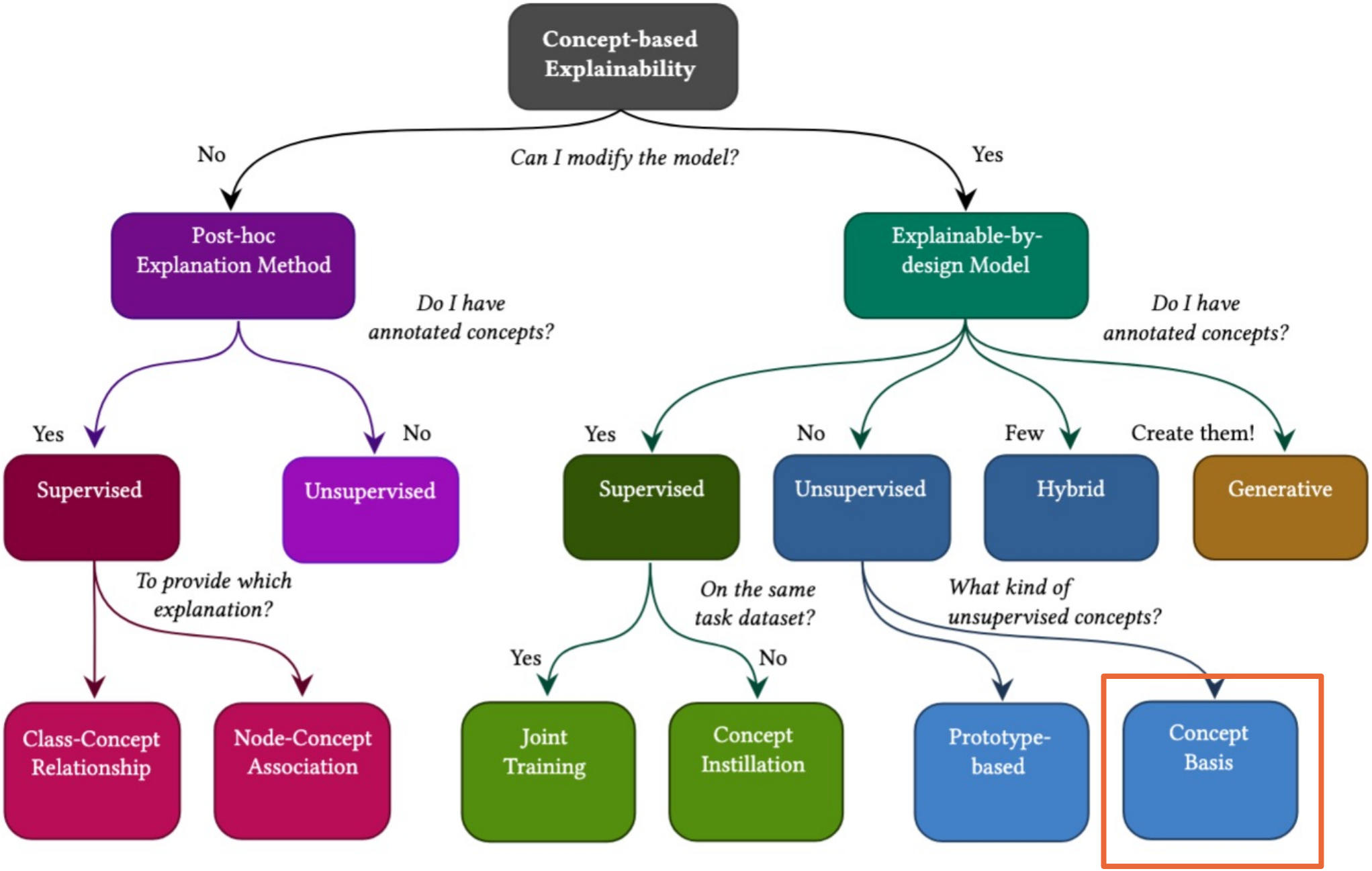
- Accuracy of the «linear probe»
  - *High* implies the network **has automatically learnt** a concept
  - *Low* implies the network **does not use** that concept for predicting the final class



- Simpler concepts have high accuracy throughout the NN
- High-level concepts can be detected better at higher layers

Credits for some of these slides goes to Alvarez Melis & Jaakkola

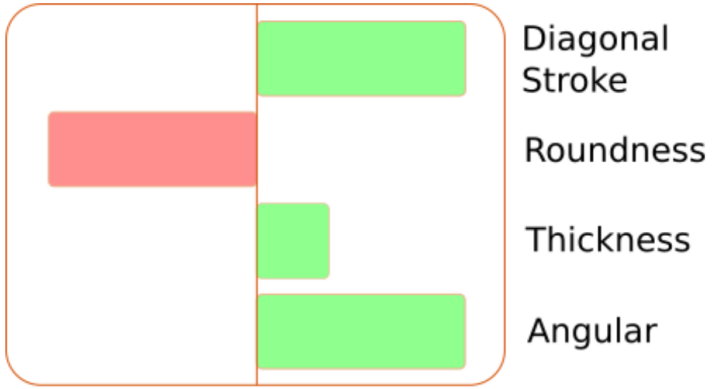
2. Self Explainable Neural Networks (SENNs)


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Let's Learn Concepts by Design

- Instead of explaining a network in terms of concepts
- Let's use a linear model working on top of interpretable features:
  - Linear model: $f(x) = \sum_i^k \theta_i x_i$
  - Linear model working on top of Basis Concepts: $f(x) = \sum_i^k \theta_i \, h_i(x)$

x

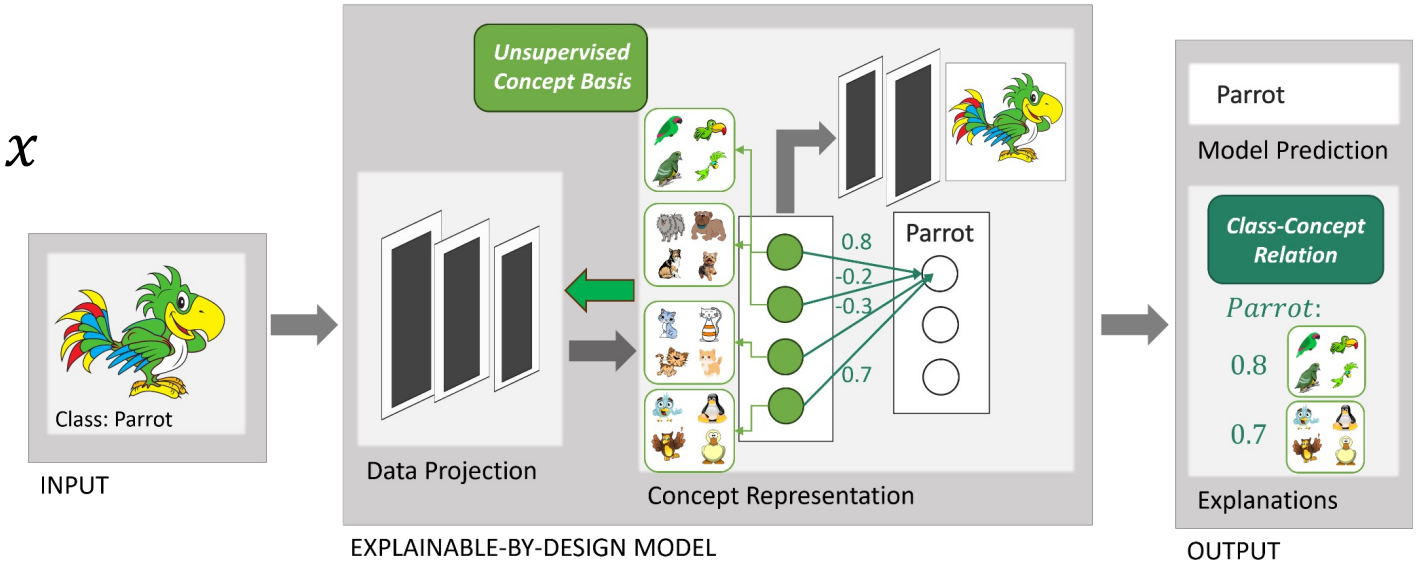


$h(x)$

How are Concepts learnt?

- **Fidelity**: the representation of x should preserve relevant information of the input
  - Autoencoder:

$$\mathrm{x}' = h_{dec}\big(h(x)\big)$$



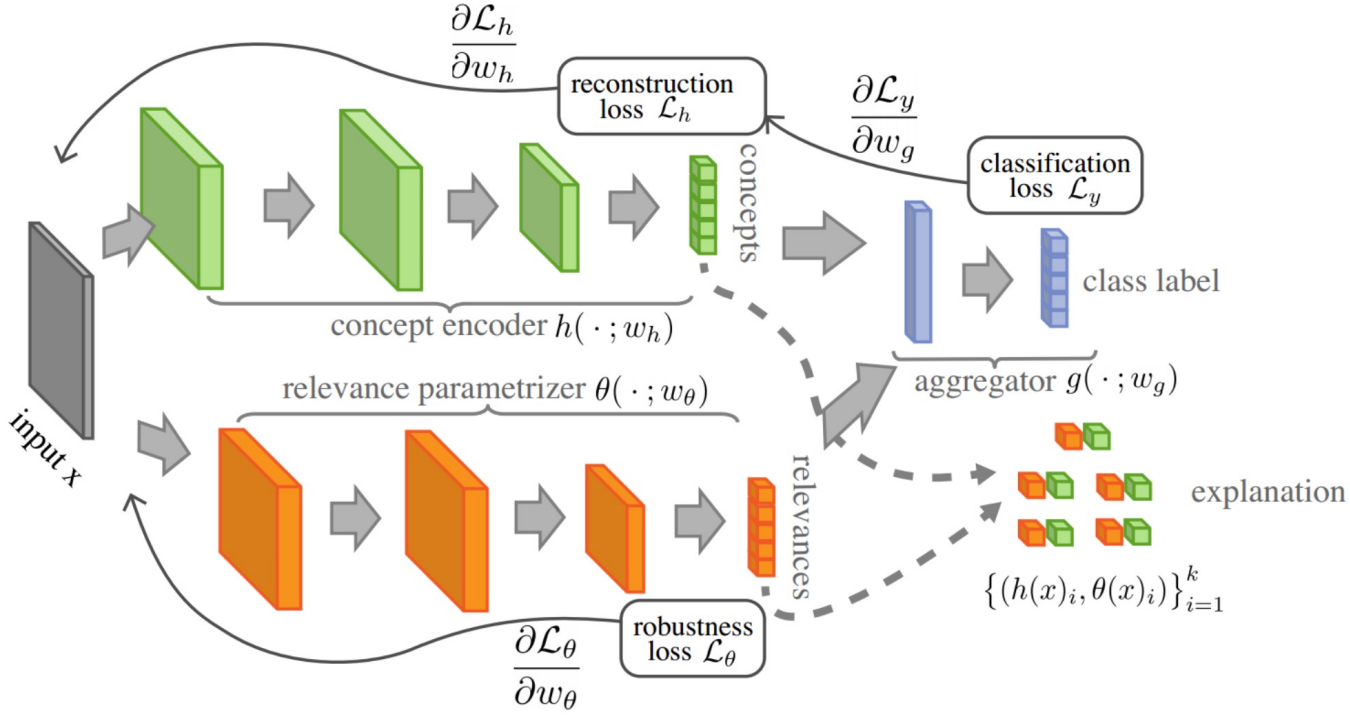
- **Sparsity:** the classes should be represented by few concepts:

  Sparse Auto Encoder:

$$L_{SAE} = \big|\big|x - x'\big|\big|^2 + \lambda KL(h(x)||\rho)$$

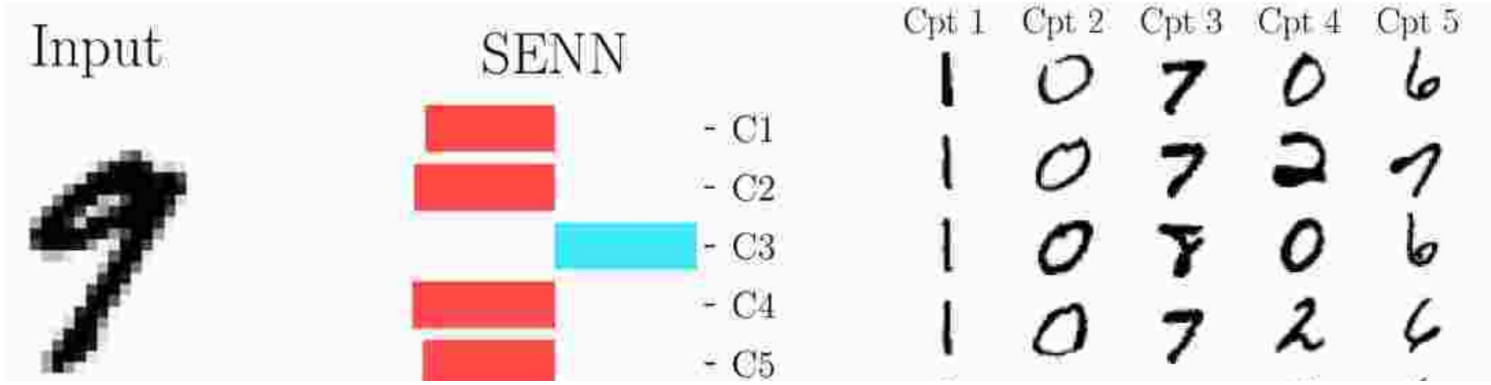
Beyond Linear Models

- Relevance scores θ_i are fixed globally for all samples. Let's predict them!
  - Linear model working on top of Basis Concepts:
    $$f(x) = \sum_i^k \theta_i \, h_i(x)$$
  - Linear model Predicting local weights:
    $$f(x) = \sum_i^k \theta_i(x) \, h_i(x)$$



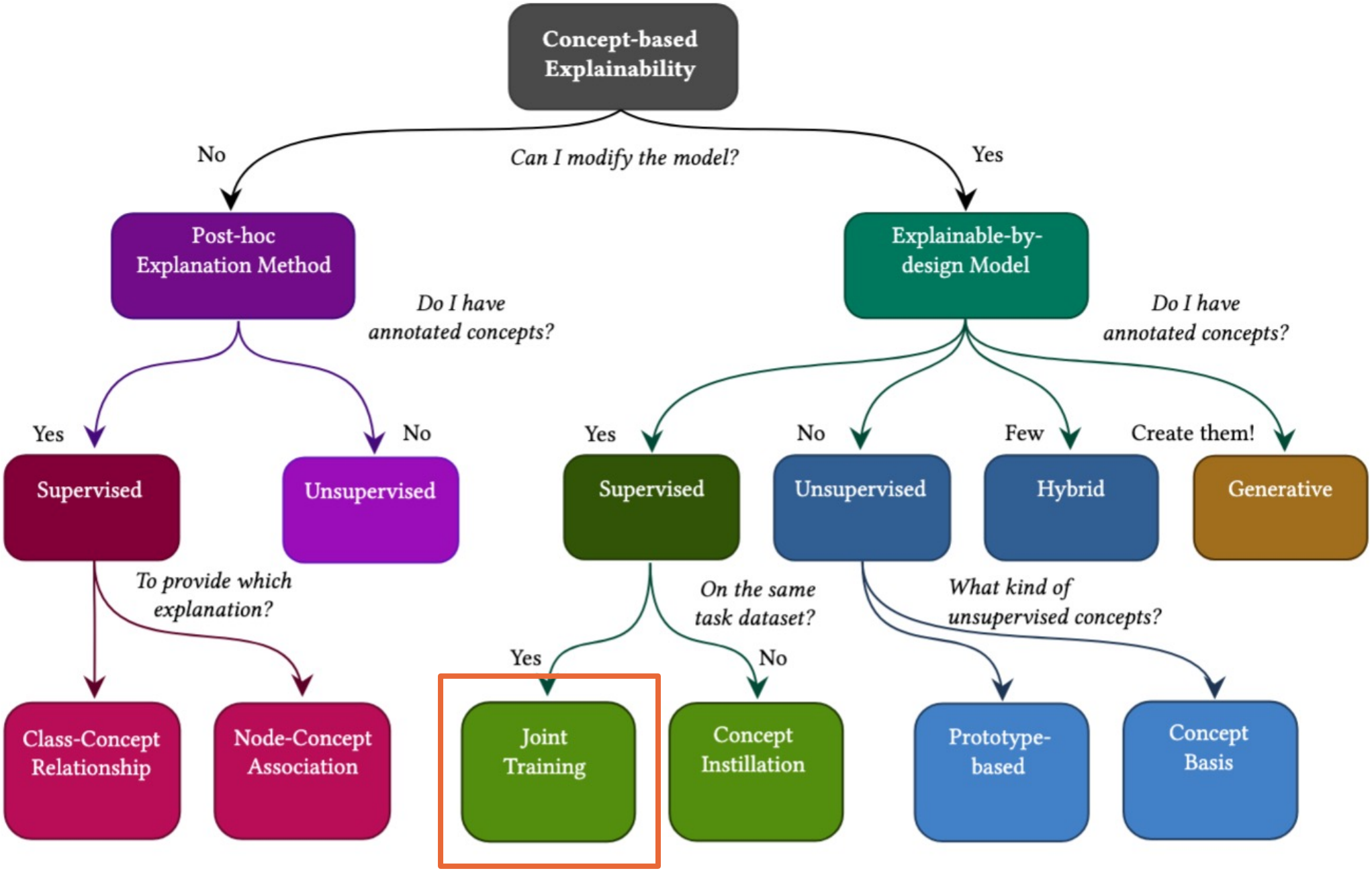
Interpreting Unsupervised Basis Concept

- $h(x)$ is faithful and sparse
- But it still requires to be interpreted (similarly to mechint)
- Let's use its «concept dictionary»

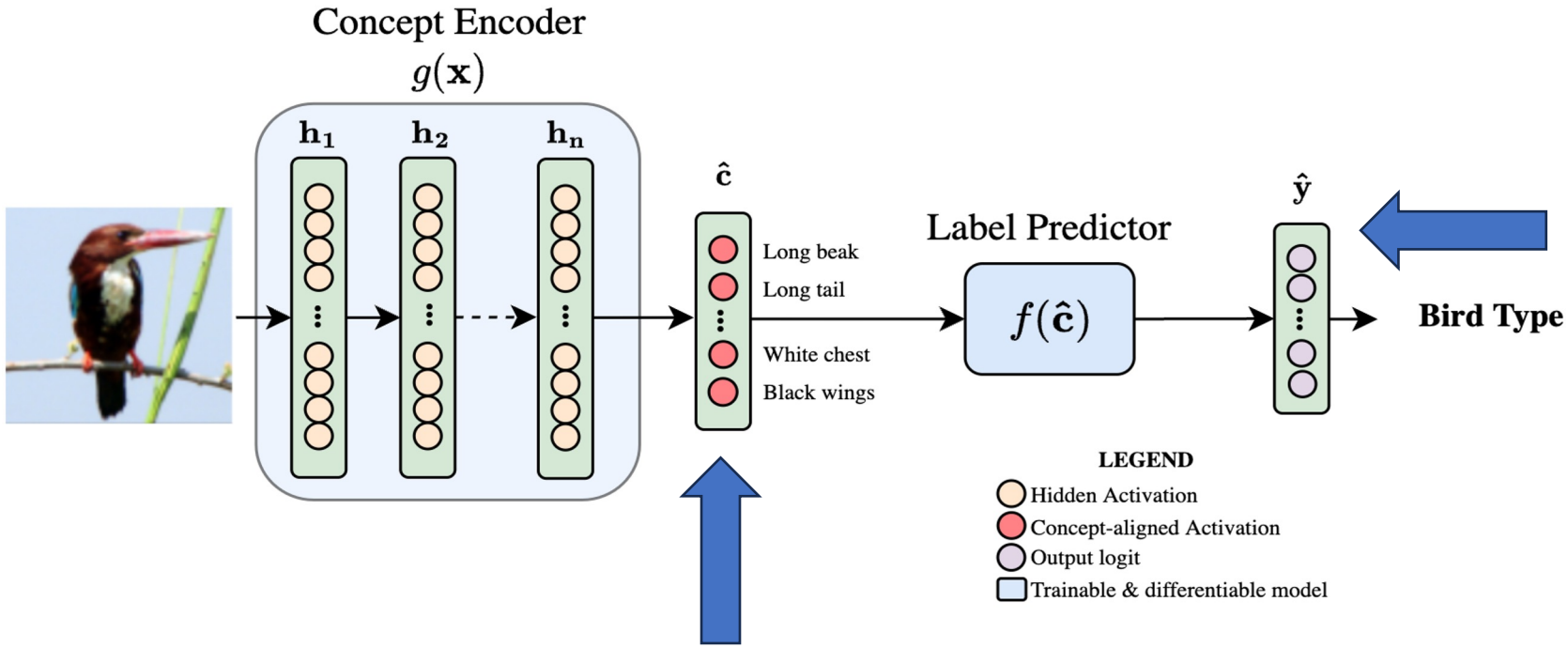


Credits for some of these slides goes to Koh et al.

3. Concept Bottleneck Models (CBMs)


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Concept bottleneck models architecture



Task loss: $L_y(f(c_i); y_i)$

Concept loss: $L_c(g(x_i); c_i)$

Different training strategy

- Indipendent: $\hat{f} = \arg min_{\mathrm{f}} \sum_i L_y(f(c_i), y_i)$

  $\hat{g} = \arg \min_g \sum_i L_c(g(x_i), c_i)$

  f is trained using the truth concepts

- Sequential: $\hat{f} = \arg min_f \sum_i L_y\big(f\big(g(x_i)\big), y_i\big)$

  g is trained first as above, then freezed

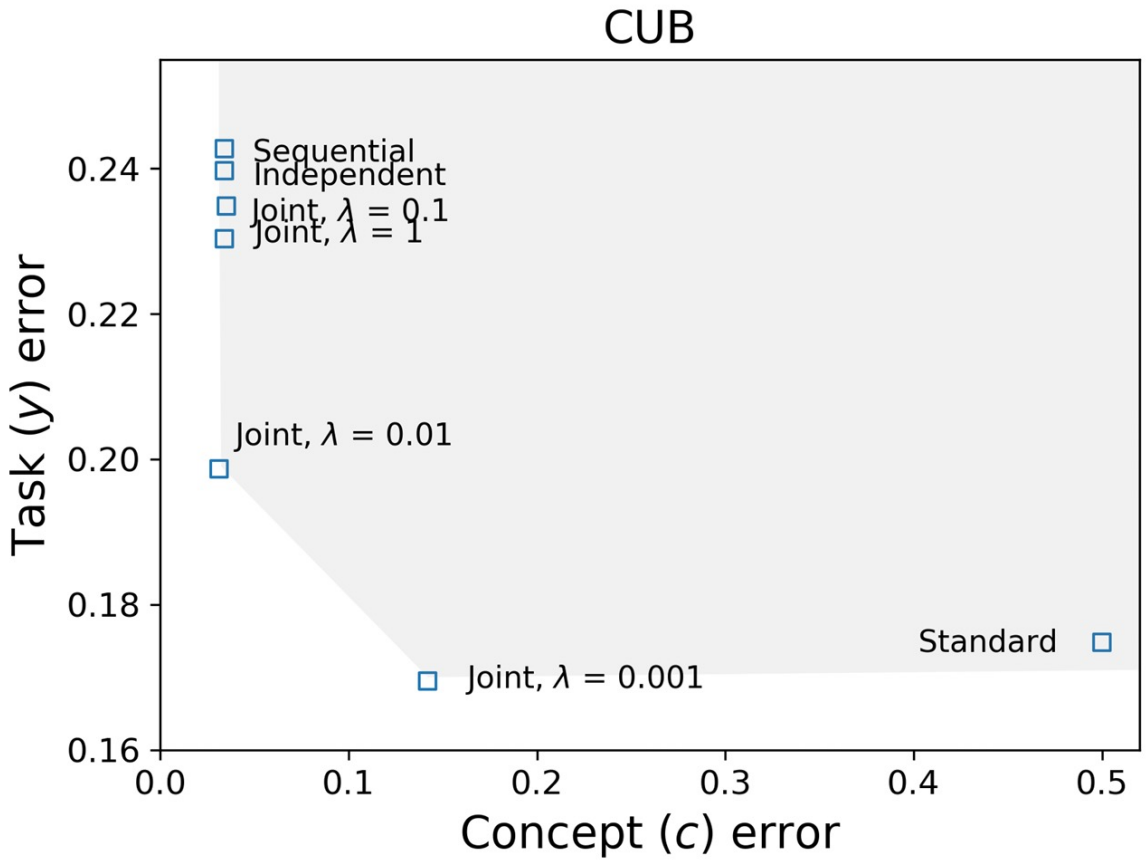
- Joint: $\hat{f}, \hat{g} = \arg min_{\mathrm{f}} \sum_i L_y(f(c_i), y_i) + \lambda \arg min_g \sum_i L_c(g(x_i), c_i)$

  f,g trained together for some $\lambda > 0$

- Standard: $\hat{f}, \hat{g} = \arg min_{\mathrm{f}} \sum_i L_y(f(c_i), y_i)$

  It ignores the concepts loss

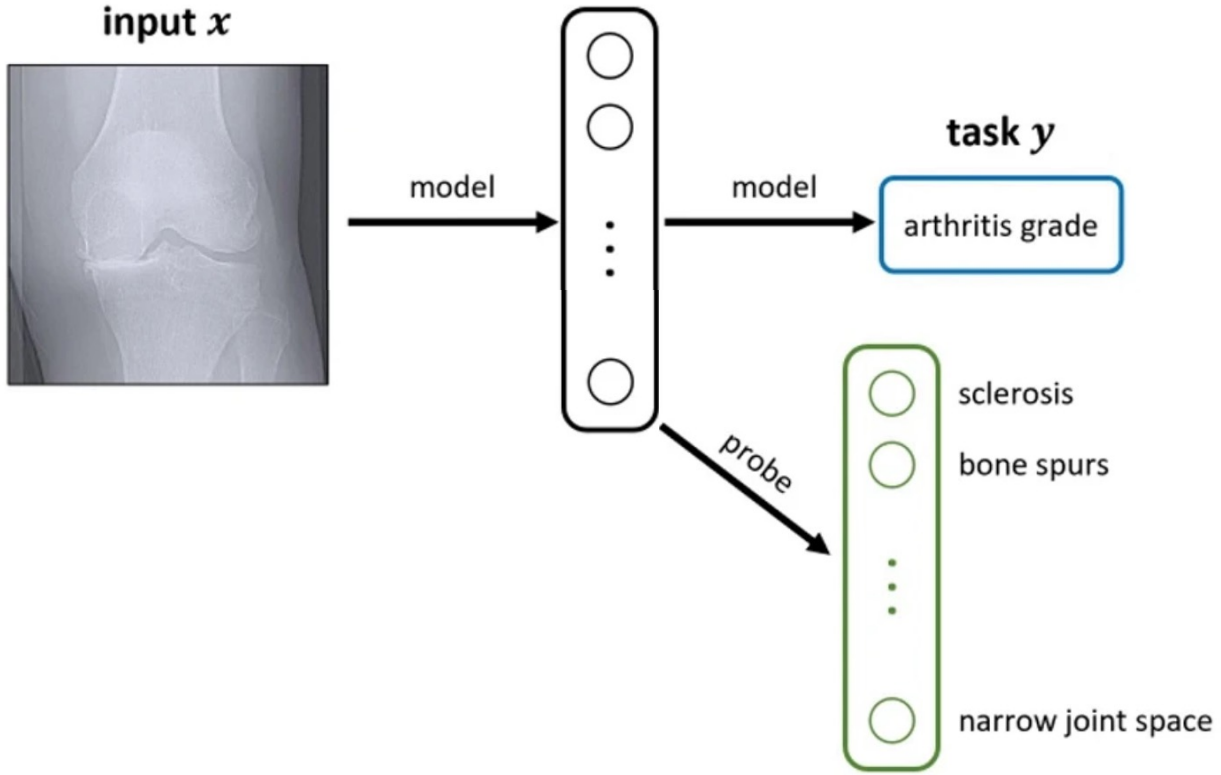
Different interpretability/performance trade-offs



- **Sequential** and **indipendent** are the more «trustworthy» beacause they ensure no concept leakage
- **Joint** strategy provides better task accuracy
  - Different trade-offs according to the λ value
- **Standard** model still has higher accuracy on average

Explictly concept training ensure model learns the concepts

Standard E2E trained model

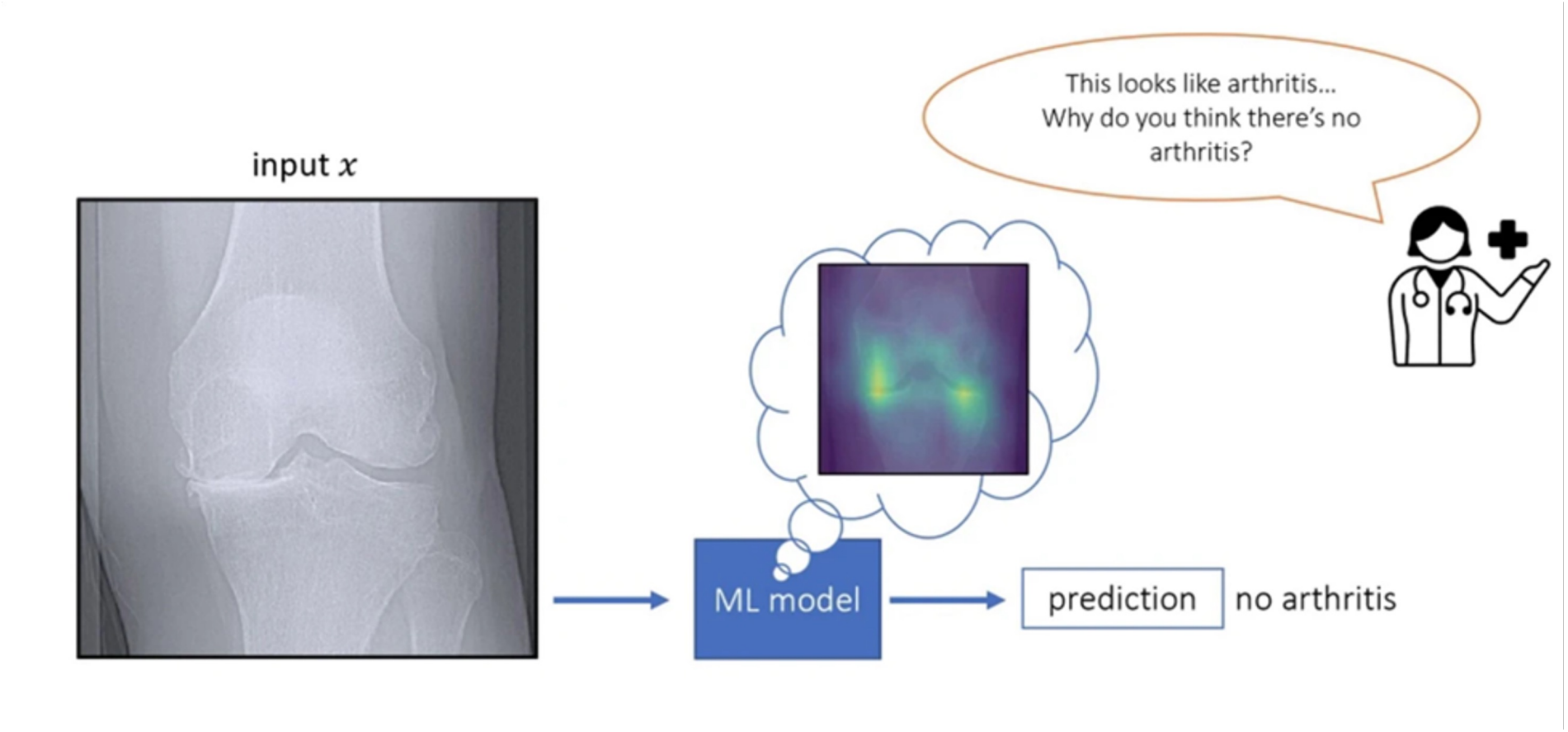


Method	X-Ray Concept Error (↓)
Independent	0.53
Sequential	0.53
Joint	0.54
TCAV [Probe]	0.68

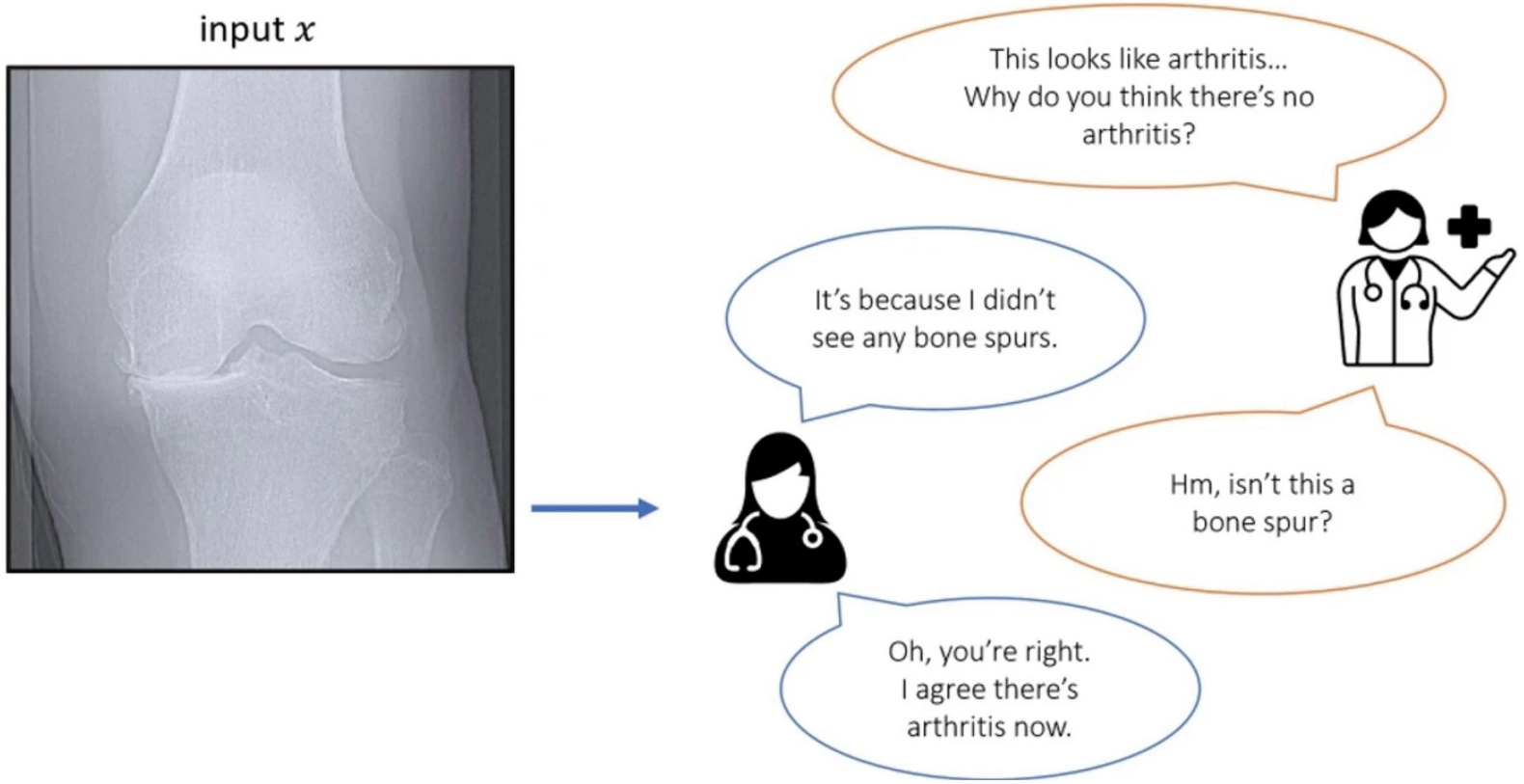
In a trained model, identifying some concepts may not be possible, because it might not have learnt them automatically

→ Only by explicitly training a model we can ensure it represents all concepts!

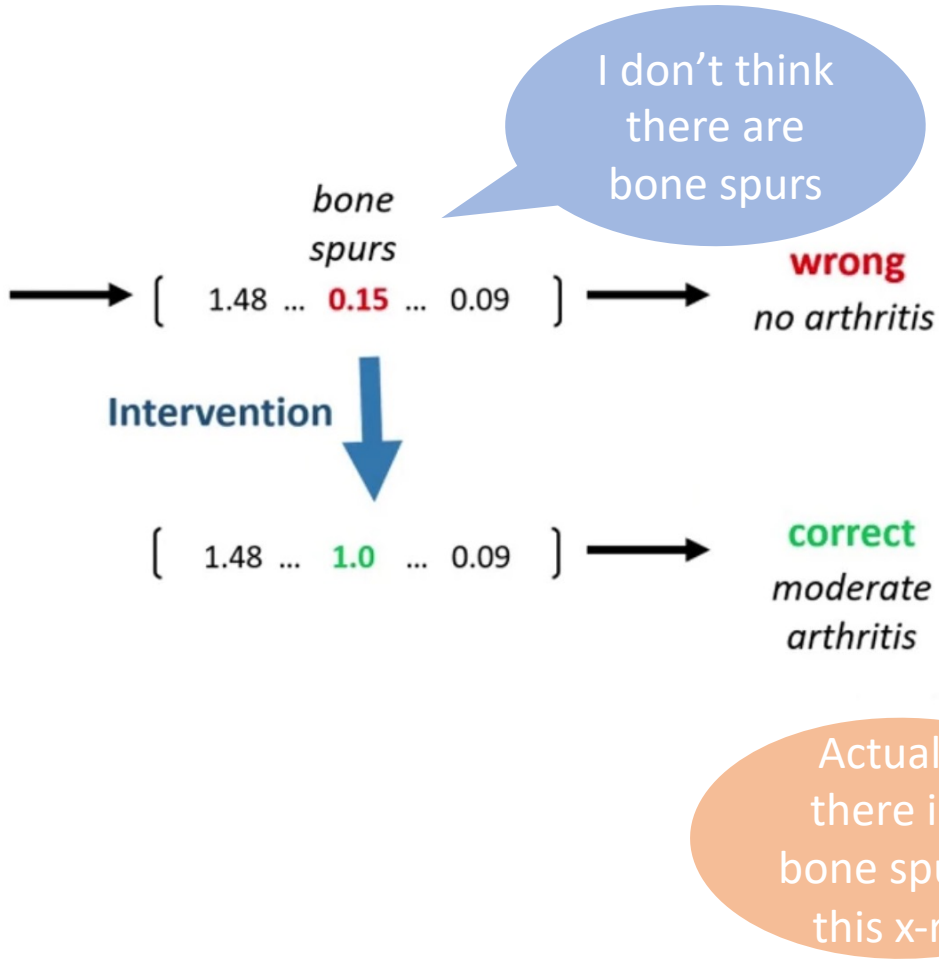
End-2-End models are difficult to interact with



Ideal: Interact through high-level concepts



Importance of Concept Intervention

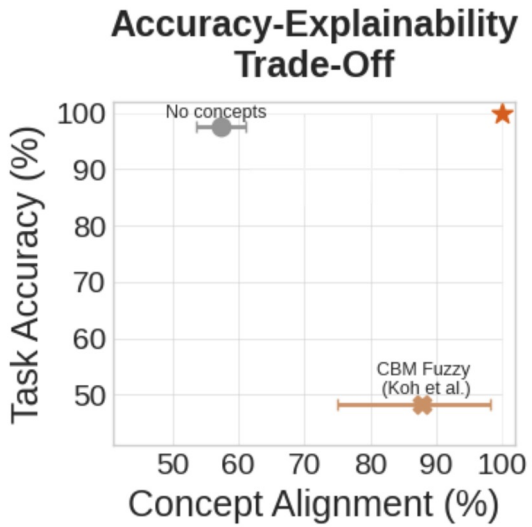


CBM Drawbacks

Poor Trade-offs

—

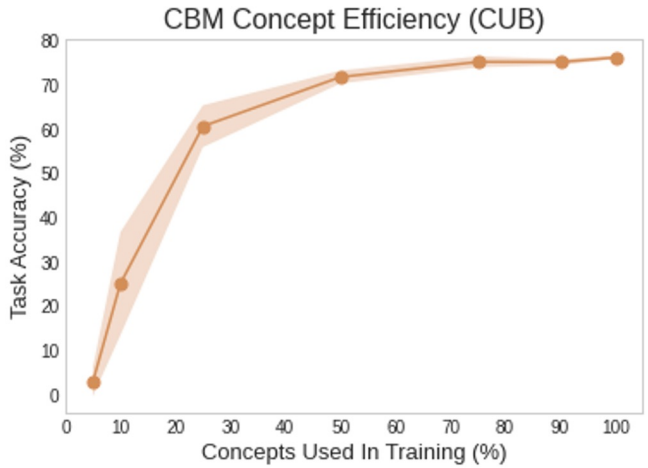
Struggle to compromise between accuracy and explainability

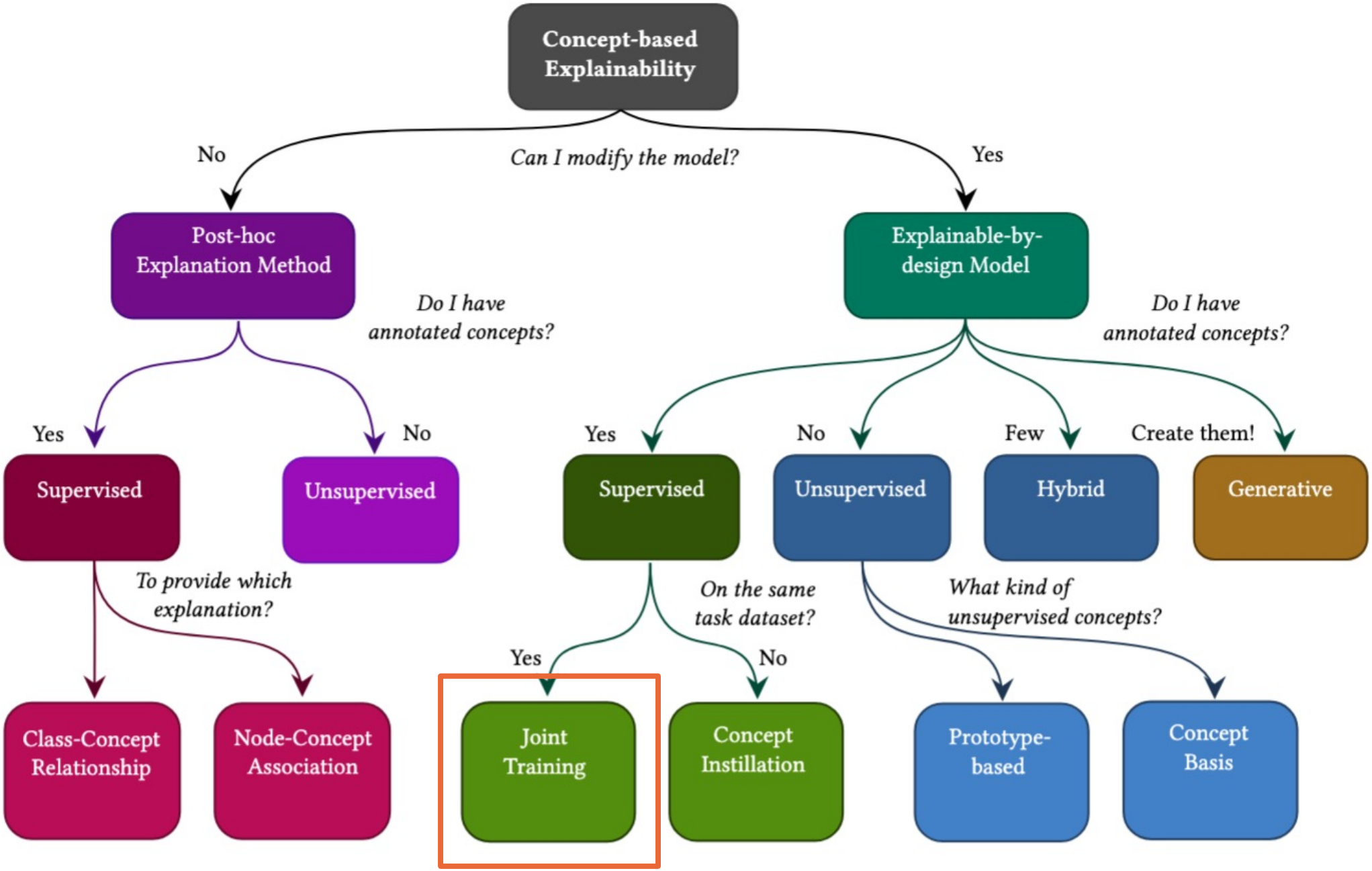


Low Concept Efficiency

—

CBMs do not scale in real-world conditions

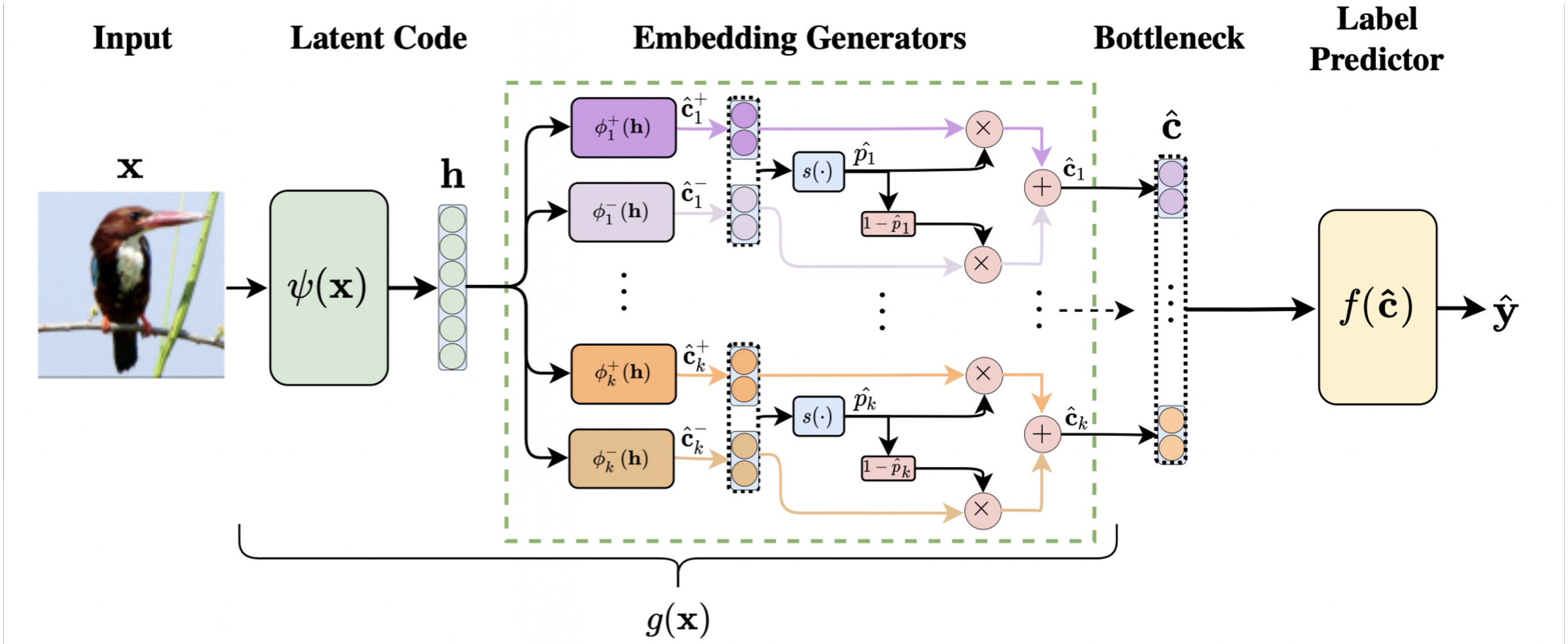



Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

Credits for some of these slides goes to Barbiero et al.

4. Concept Embedding Models (CEM)

Concept Embedding Models: overview



Concept Embedding workflow

1. $h = \psi(x)$: the latent space of the model

2. $\boldsymbol{c_i^+} = \phi_i^+(x)$: neural model dedicated to represent the i-th **positive** concept embedding

3. $\mathrm{p}_i = s([\boldsymbol{c_i^+}, \boldsymbol{c_i^-}])$: the *concept score* (i.e., probability of presence of the ith concept) is a function shared among concepts working on the concatenations of the concept embeddings

4. $\hat{\boldsymbol{c}}_1 = \mathrm{p_i}\mathbf{c_i^+} + (1 - \mathrm{p_i})\mathbf{c_i^-}$: the *concept embedding* is represented by the weigthed combination of the positive and negative concept embeddings according to its presence

5. $f([\hat{\boldsymbol{c}}_1, \ldots \hat{\boldsymbol{c}}_i, \ldots \hat{\boldsymbol{c}}_k])$: the task predictor works on the concatenation of all the concept embeddings

CEM: A neural-symbolic approach

Neural

—

Concepts are represented with: unsupervised **embeddings**

$c_i \in \mathbb{R}^k$

Symbolic (CBM)

—

Concepts are represented with: **supervised** scalars

$c_i \in [0,1]$

Neural Symbolic (CEM)

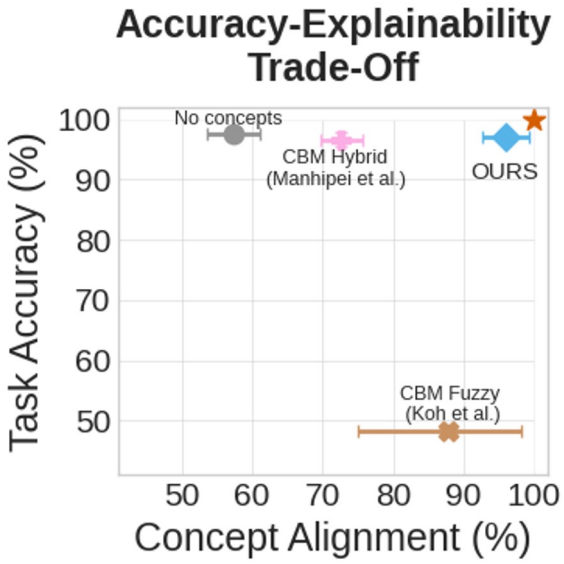
—

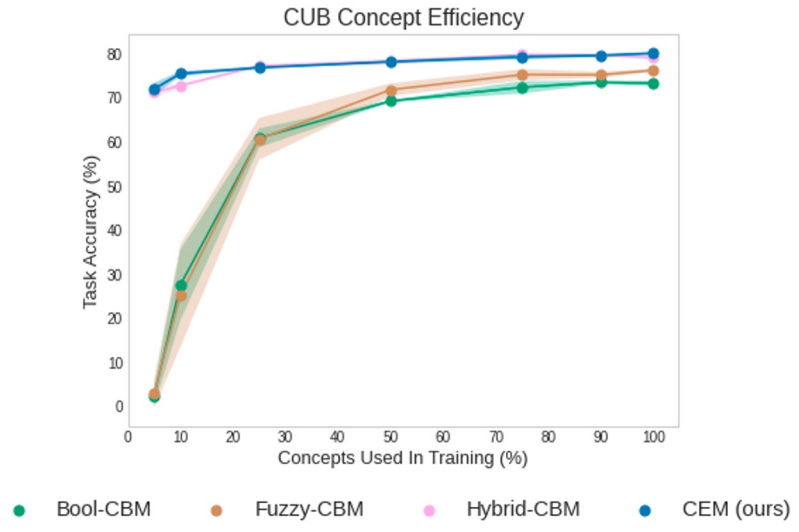
Concepts are represented with: pairs of **supervised embeddings**

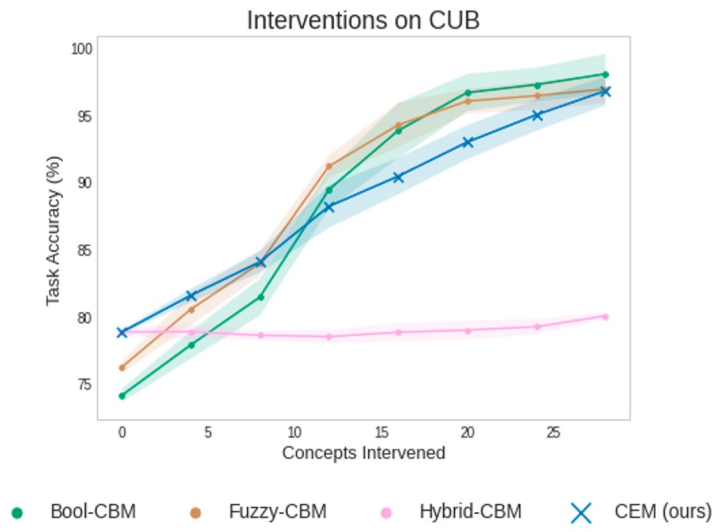
$c_i \in \mathbb{R}^k$

$c_i = \text{agg}(c_i^+, c_i^-)$

CEM Advanatages







Beyond Trade-offs

—

CEMs overcome the current accuracy-explainability trade-off

High Concept Efficiency

—

CEMs scale to real-world conditions where concept supervisions are scarce

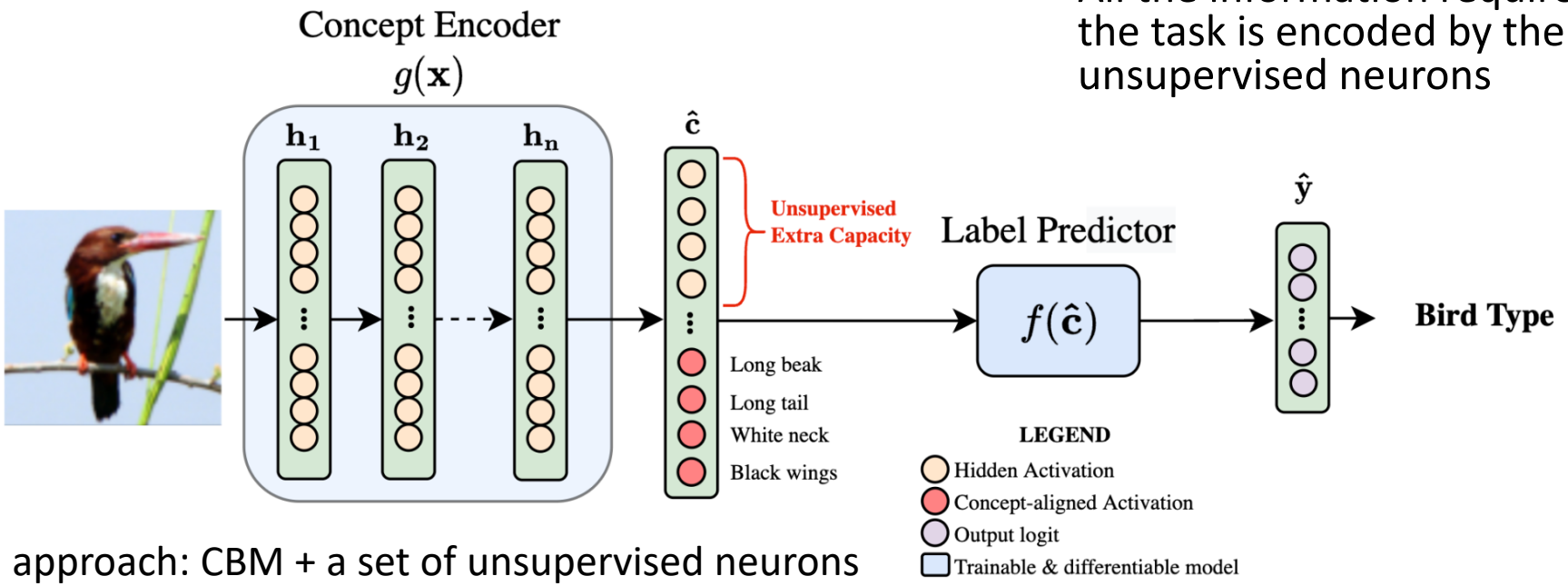
Effective interventions

—

CEMs are responsive to concept interventions

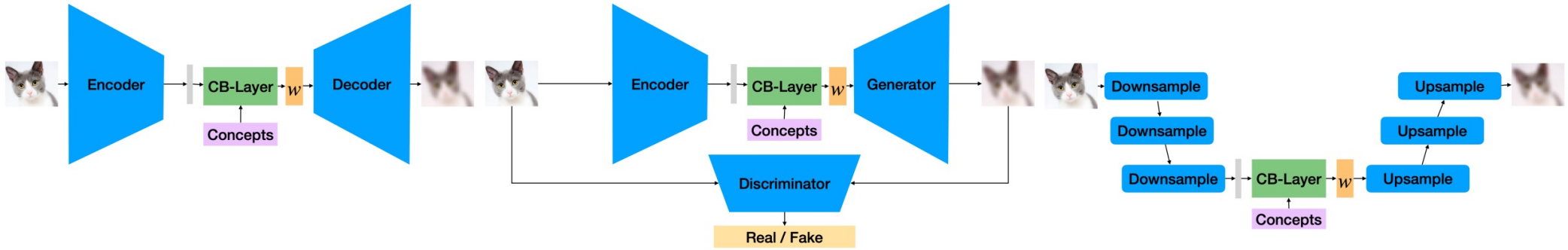
CEM vs Hybrid approach

- PROS:
  - Retain high accuracy
  - Has high concept efficiency like CEM
- CONS:
  - Prevent any effect of concept intervention
    - Changing the predicted scores has no effect on the task prediction
  - All the information required to predict the task is encoded by the unsupervised neurons



Hybrid approach: CBM + a set of unsupervised neurons

Generative models based on CEM



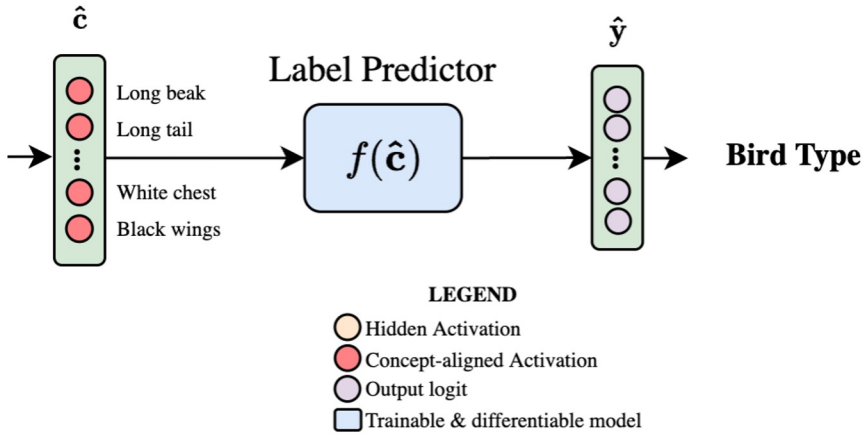
(a) VAE with CB layer (b) GAN with CB layer (c) Diffusion with CB layer

Ismail, Aya Abdelsalam, et al. "Concept bottleneck generative models." ICLR 2024.

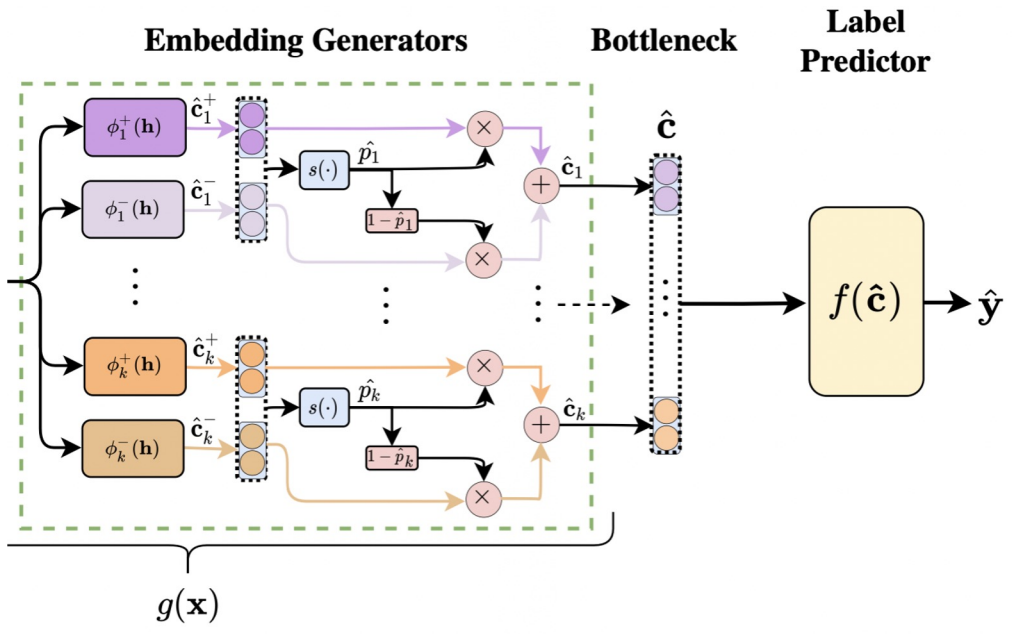
Have we lost something?

Interpretability

CBM: Interpretable



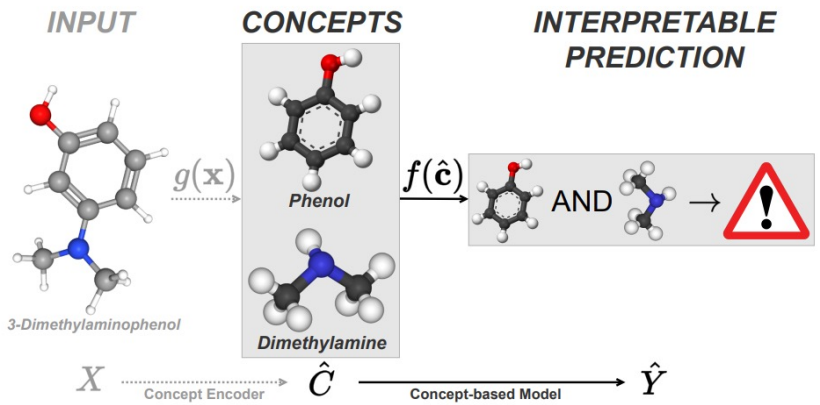
CEM: NON-Interpretable



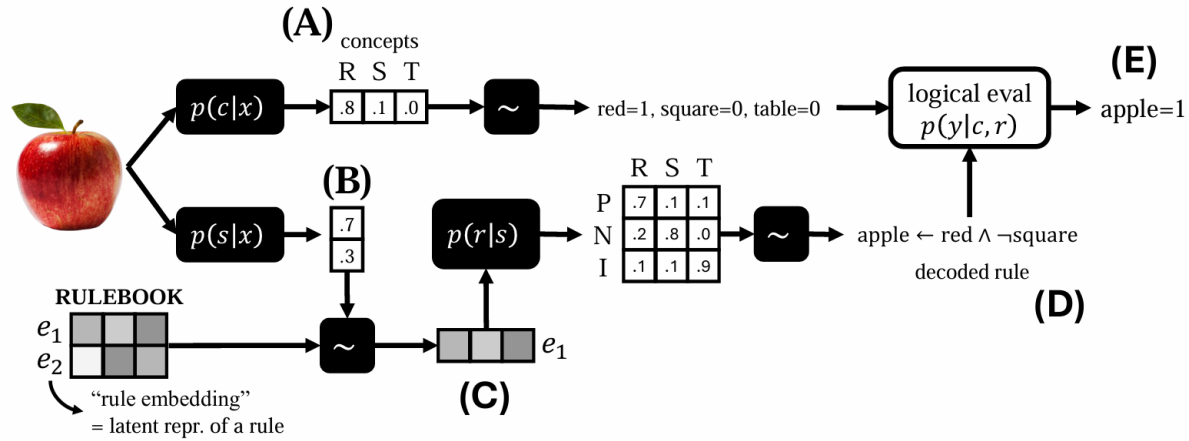
$$\hat{\mathbf{c}}_{\text{yellow}} = [2.3, 0.3, -3.5, \ldots]^T$$

Task-Interpretable CEMs

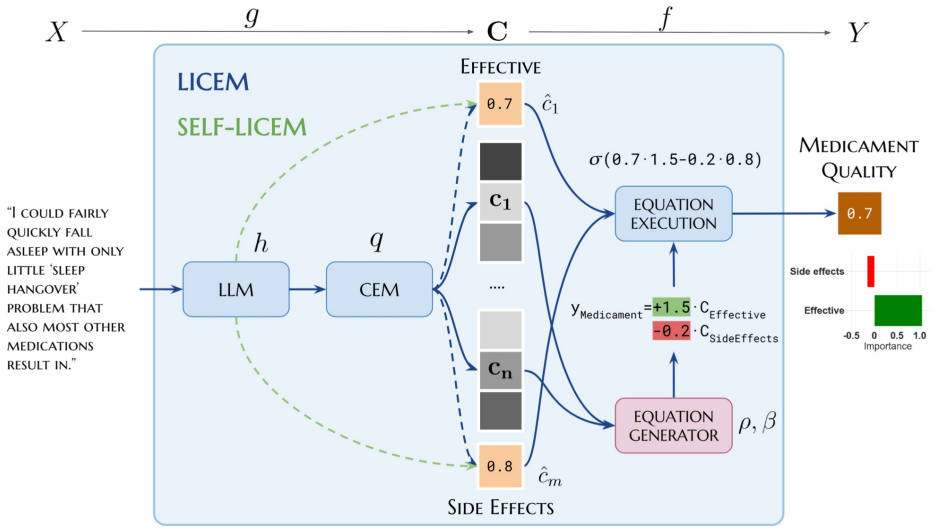
DCR[1]: Local, Rule-based Predictions



CMR[2]: Global, Rule-based Predictions



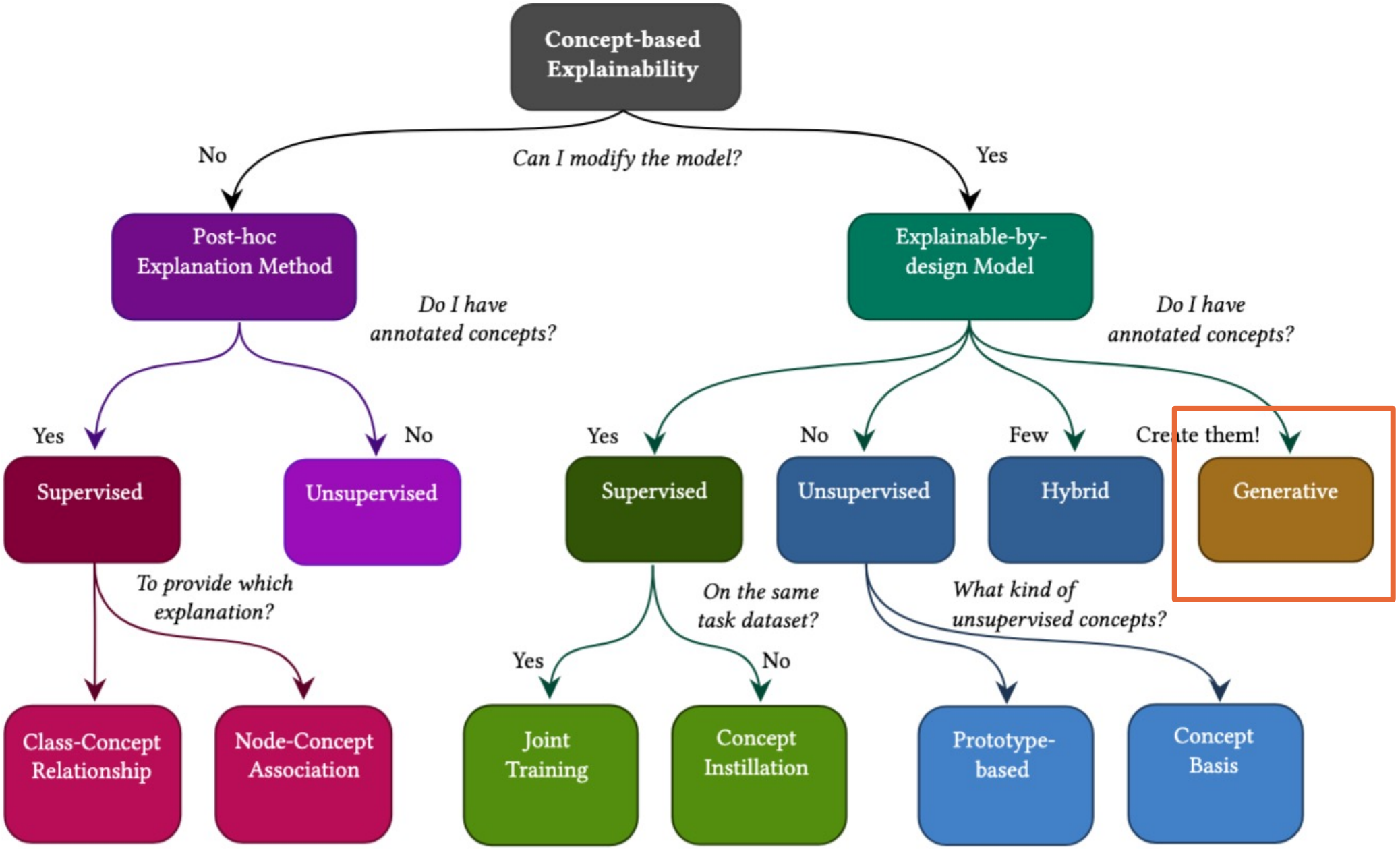
LICEM[3]: Local, Equation-based Predictions



1. Barbiero, Pietro, et al. "Interpretable Neural-Symbolic Concept Reasoning." International Conference on Machine Learning (2023).
2. Debot, David, et al. "Interpretable Concept-Based Memory Reasoning." The 38th Annual Conference on Neural Information Processing Systems (2024).
3. De Santis, Francesco, et al. "Linearly Interpretable Concept Embedding Model for Text Classification." Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2025).

Credits for some of these slides goes to Oikarinen et al.

5. Label-free CBM (LF-CBMs)


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method
Explainable-by-design Model
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised
Unsupervised
Supervised
Unsupervised
Hybrid
Generative
To provide which explanation?
On the same task dataset?
What kind of unsupervised concepts?
Yes
No
Class-Concept Relationship
Node-Concept Association
Joint Training
Concept Instillation
Prototype-based
Concept Basis

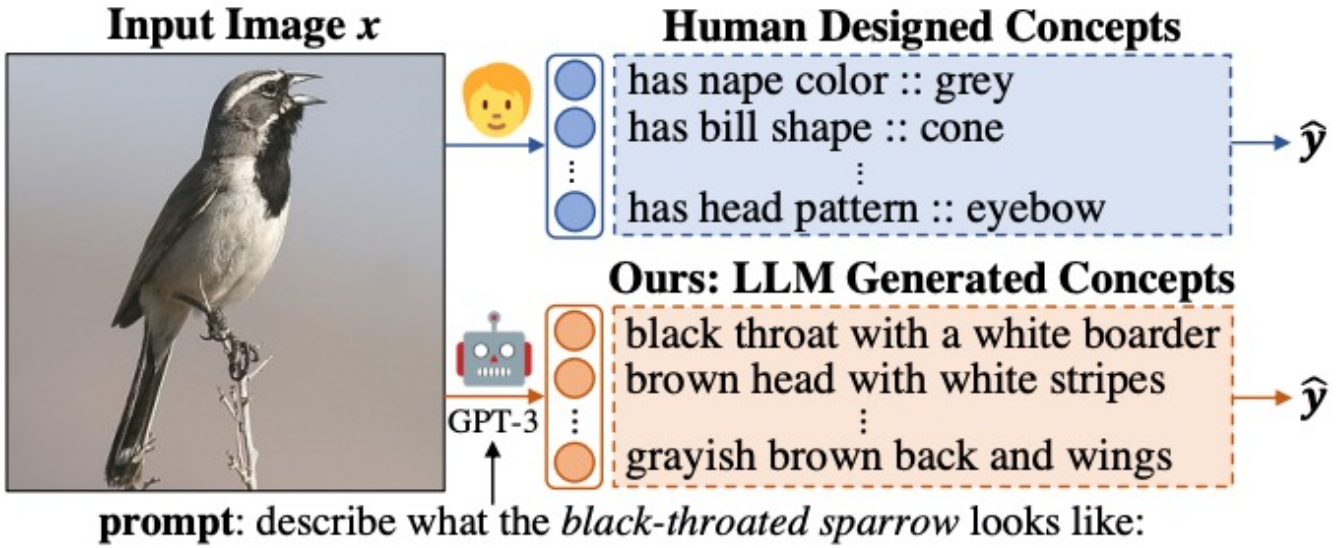
CBM Issues

- Strong extra annotation effort required from humans: Up to |C| * |N|
  - |C| # concepts
  - |N| # samples

- Low performance when using few concepts without embeddings
  - Unable to distinguish classes characterized by the same concepts
  - E.g. Plane vs Helicopter cannot be distinguished with only the concept fly

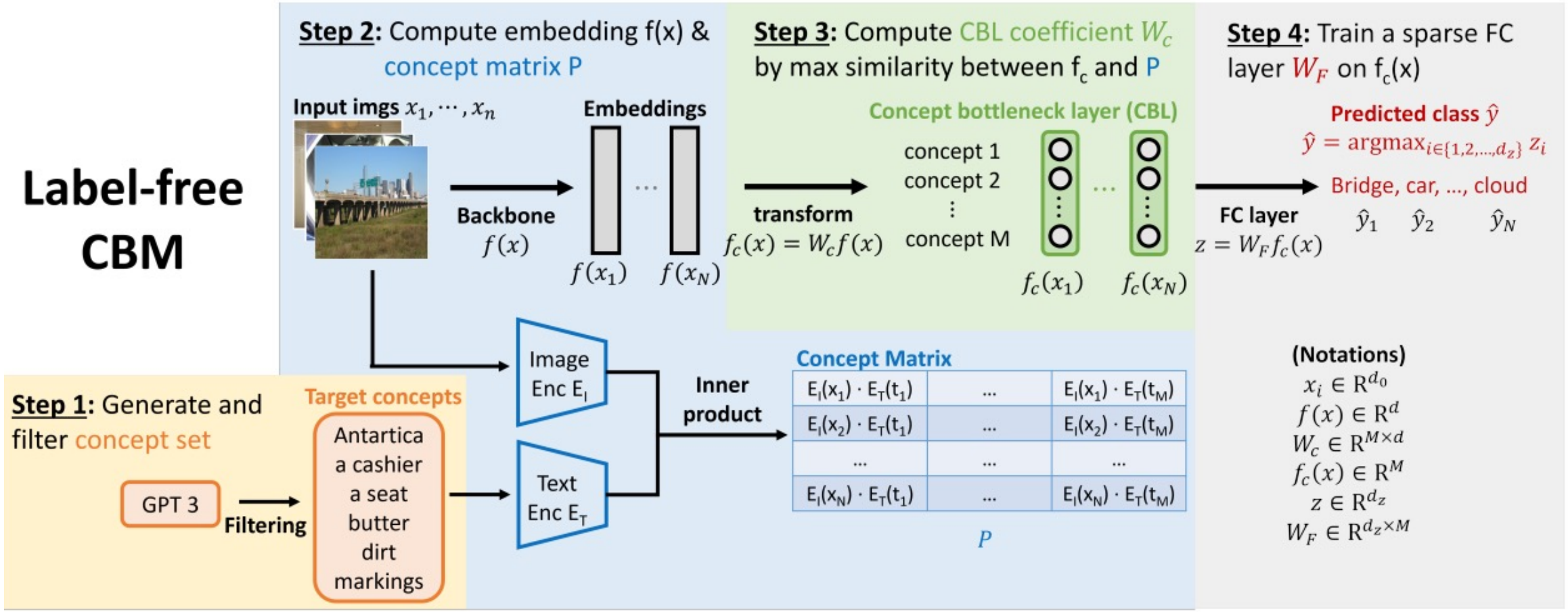
Let’s use a pretrained VLM!

- No extra effort required by human annotators



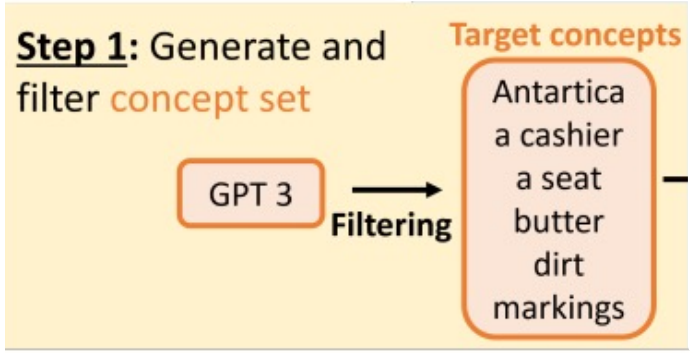
- Potentially infinite number of concepts
  --> no bottleneck in the representation

LF-CBM architecture



LF-CBM architecture

Label-free CBM

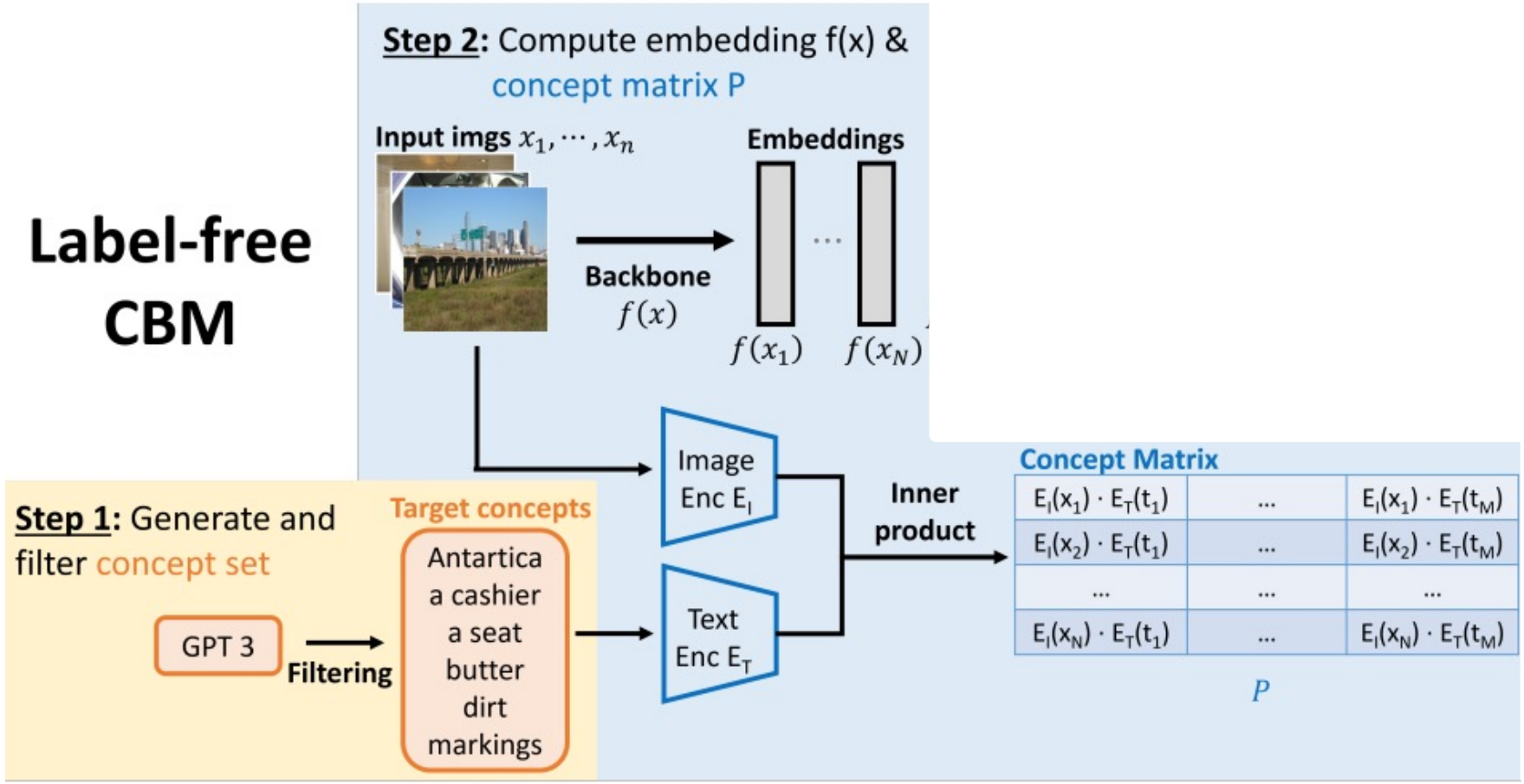


Step 1: Concept set Generation and Filtering

- Generation from GPT-3 prompted with:
  - «*List the most important features for recognizing something as a {class}:*»

- Filtering concepts with:
  - Excessively long descriptions
  - Too similar to classes
  - Too similar to each other
  - Not present in training data
  - Unable to learn with external model

LF-CBM architecture



Step 2: Concept matrix definition

Concept Matrix

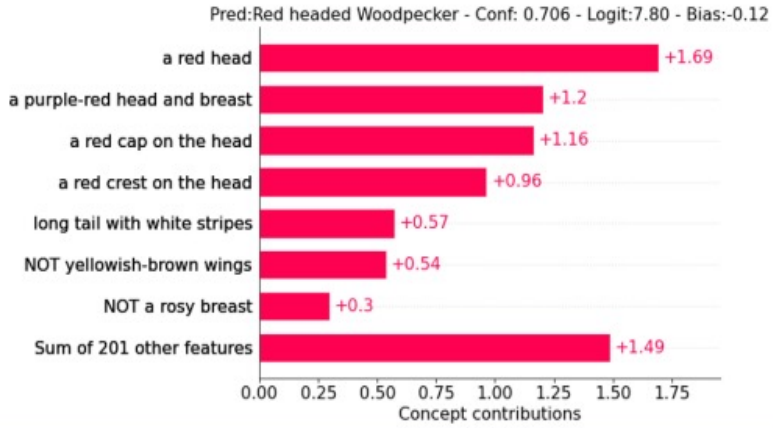
$E_I(x_1) \cdot E_T(t_1)$	...	$E_I(x_1) \cdot E_T(t_M)$
$E_I(x_2) \cdot E_T(t_1)$	...	$E_I(x_2) \cdot E_T(t_M)$
...	..	...
$E_I(x_N) \cdot E_T(t_1)$	..	$E_I(x_N) \cdot E_T(t_M)$

P

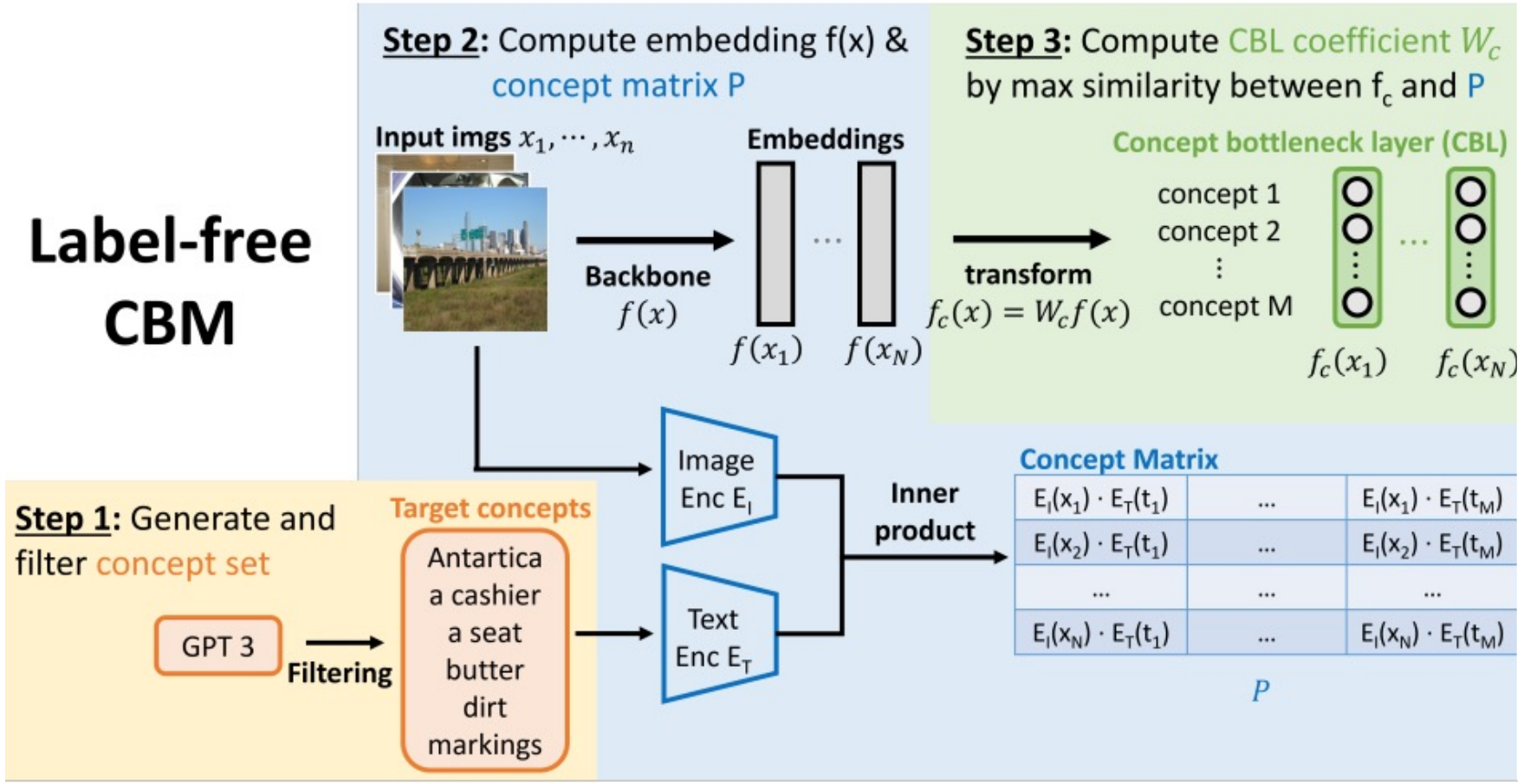
$E_I(x_i)$: CLIP embedding of the ith image

$E_T(t_j)$: CLIP embedding of the jth concept textual description

CUB



LF-CBM architecture

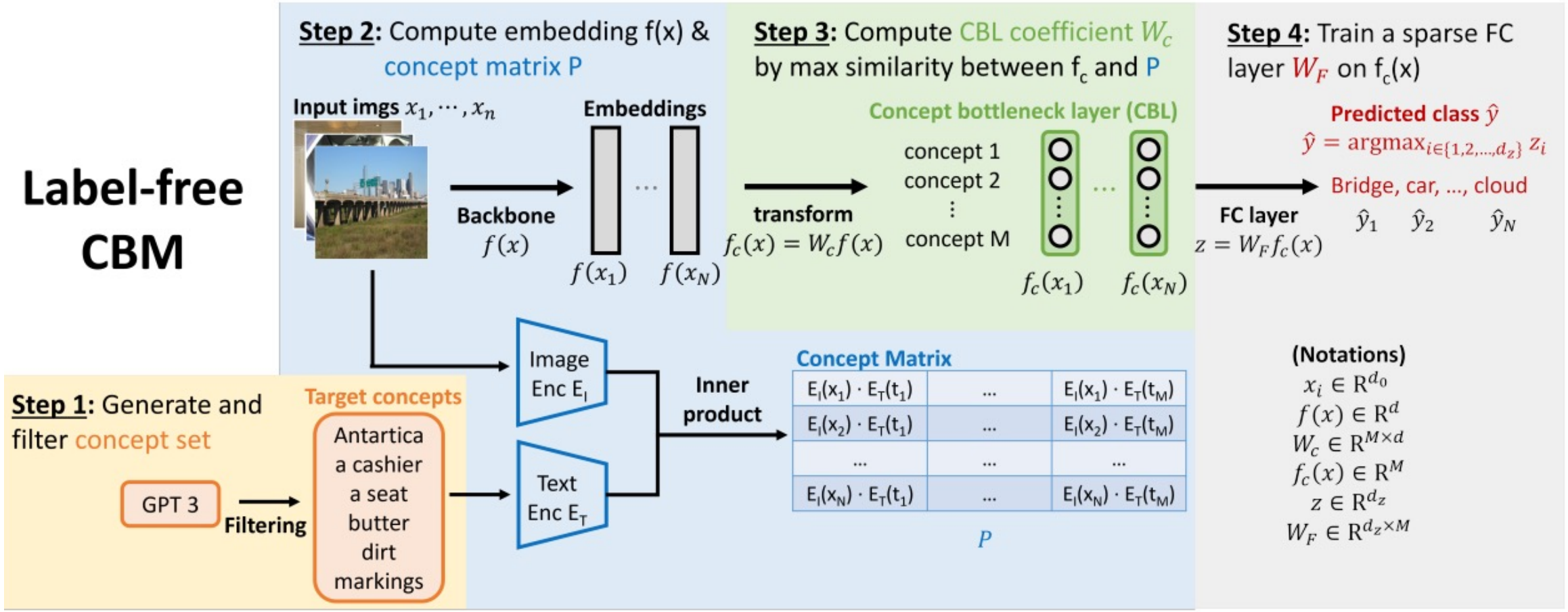


Step 3: Learn to predict the Concept Matrix

$$L(W_c) = \sum_{i=1}^{M} -\text{sim}(t_i, q_i) := \sum_{i=1}^{M} -\frac{\bar{q_i}^3 \cdot \bar{P_{:,i}}^3}{||\bar{q_i}^3||_2 ||\bar{P_{:,i}}^3||_2}.$$

- $t_i, P_{:,i},$: all CLIP concepts representation of a given concepts for all samples in the training set
- q_i: all predicted representation for a given concepts
  - $q_i = [W_c f(x_1), W_c f(x_2), \ldots, W_c f(x_N)]$
    - On top of the backbone image representation $f(x_N)$ that is also trained

LF-CBM architecture



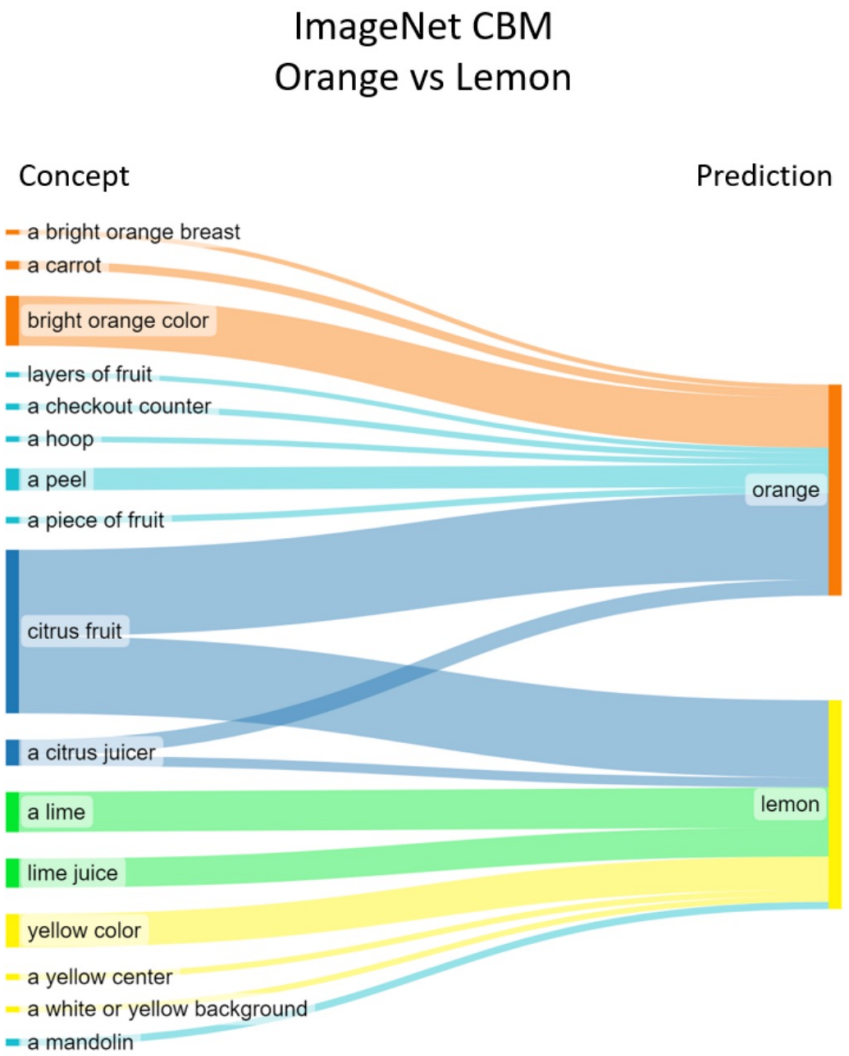
Step 4: Sparse FC layer on top of concept predictions

Final Prediction:

$$\hat{y} = W_F f_c(x)$$

Training:

- sequentially, after having trained the CBL
- $\min \sum_i^N L_{CE}\ (\hat{y}, y) + ||W_F||$



Performance

Model	Sparse final layer	Dataset				
		CIFAR10	CIFAR100	CUB200	Places365	ImageNet
Standard	No	88.80%*	70.10%*	76.70%	48.56%	76.13%
Standard (sparse)	Yes	82.96%	58.34%	**75.96%**	38.46%	**74.35%**
P-CBM	Yes	70.50%*	43.20%*	59.60%*	N/A	N/A
P-CBM (CLIP)	Yes	84.50%*	56.00%*	N/A	N/A	N/A
Label-free CBM **(Ours)**	Yes	**86.40%** ± 0.06%	**65.13%** ± 0.12%	74.31% ± 0.29%	**43.68%** ± 0.10%	71.95% ± 0.05%

Open Issues

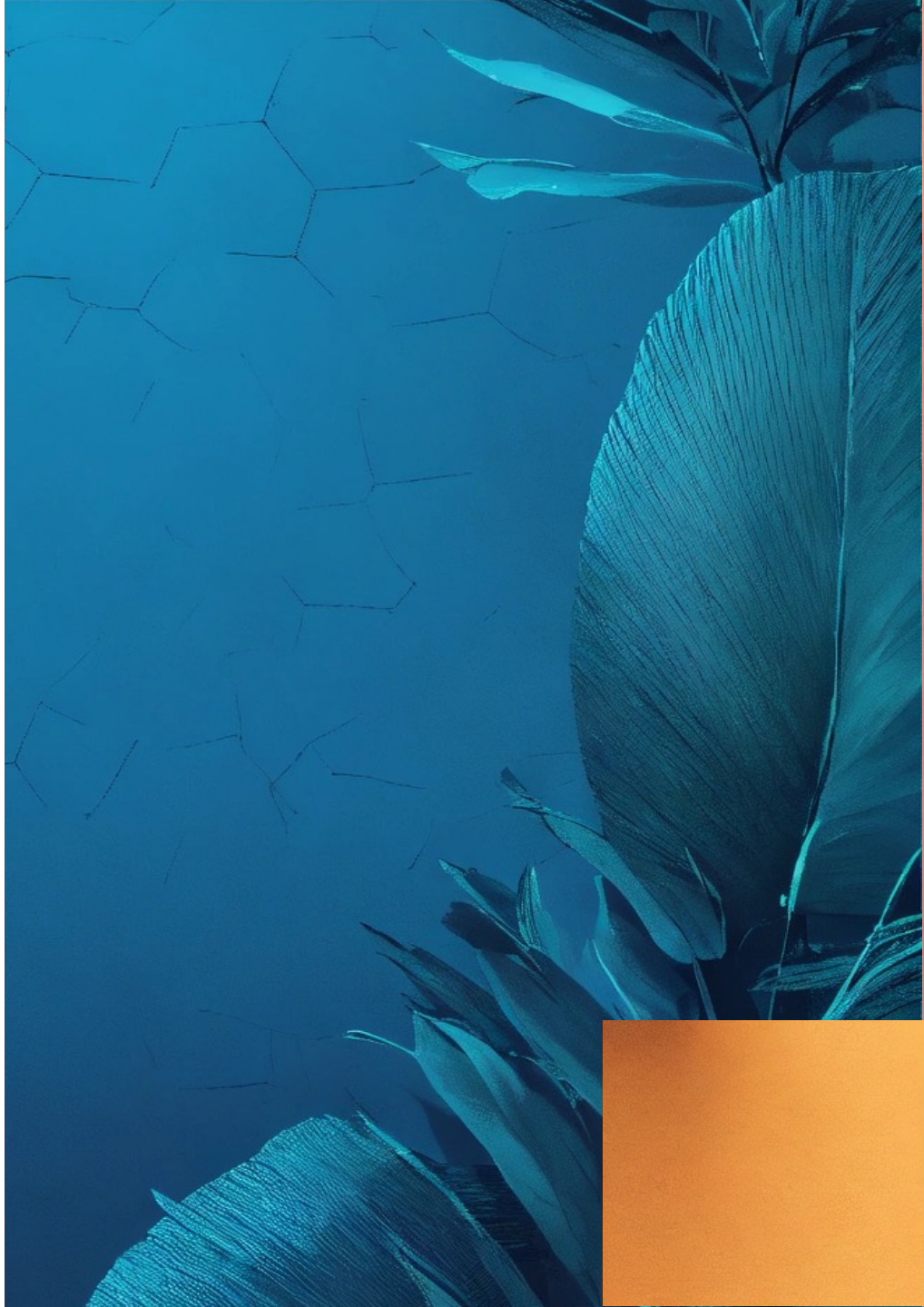
1. The employment of 3 Models to make one prediction is redundant
    1. GPT for concept list generation, CLIP for matrix description, Resnet/ViT for final prediction
2. LF-CBM concept accuracy is very low[1]

	MNIST E/O	MNIST Add.	CIFAR100	Skin	CUB200
BotCL (recon)	0.47 ± 0.01	0.41 ± 0.01	0.38 ± 0.03	0.47 ± 0.02	0.34± 0.01
BotCL (contr)	0.47 ± 0.02	0.45 ± 0.02	0.40 ± 0.04	0.44 ± 0.03	0.37± 0.02
SENN	0.61 ± 0.02	0.58 ± 0.01	0.44 ± 0.02	0.52 ± 0.02	0.41± 0.02
ProtoPNet	0.26 ± 0.01	0.24 ± 0.01	0.31 ± 0.01	0.16 ± 0.02	0.28± 0.03
LF-CBM	0.52± 0.01	0.50 ± 0.03	0.45 ± 0.01	0.58 ± 0.01	0.45 ± 0.01
LCBM (ours)	**0.88 ± 0.08**	**0.81 ± 0.04**	**0.60 ±≤0.01**	**0.58 ±≤0.01**	**0.55 ±≤0.01**

1. De Santis, Francesco, et al. "Towards Better Generalization and Interpretability in Unsupervised Concept-Based Models." ECML-PKDD 2025

Tutorial structure

- Introduction to XAI techniques
- Taxonomy of C-XAI
- **Evaluation, Resources and applications**
- Hands-on session



Metrics & Evaluation of Concept-based Models

- **Quantitative Evaluation**

Assesses how much concepts contribute to class predictions and task performance and their quality.

Automated, Replicable.

- **Qualitative Evaluation**

Examines the ***quality*** *of concept-based explanations* ***from a human perspective***.

Reflects **human understanding, hard to scale**.

Metrics & Evaluation of Concept-based Models

- **Quantitative Evaluation**

Two main categories

1. **Contribution to Prediction & Performance**
2. **Quality of Concepts**

Quantitative Evaluation

1. Contribution to Prediction & Performance

*How well do **concepts contribute** to class predictions and task performance?*

Concept effect on class prediction

Provide the **concept relevance** by means of the concept weights, connecting the concept to the final classes

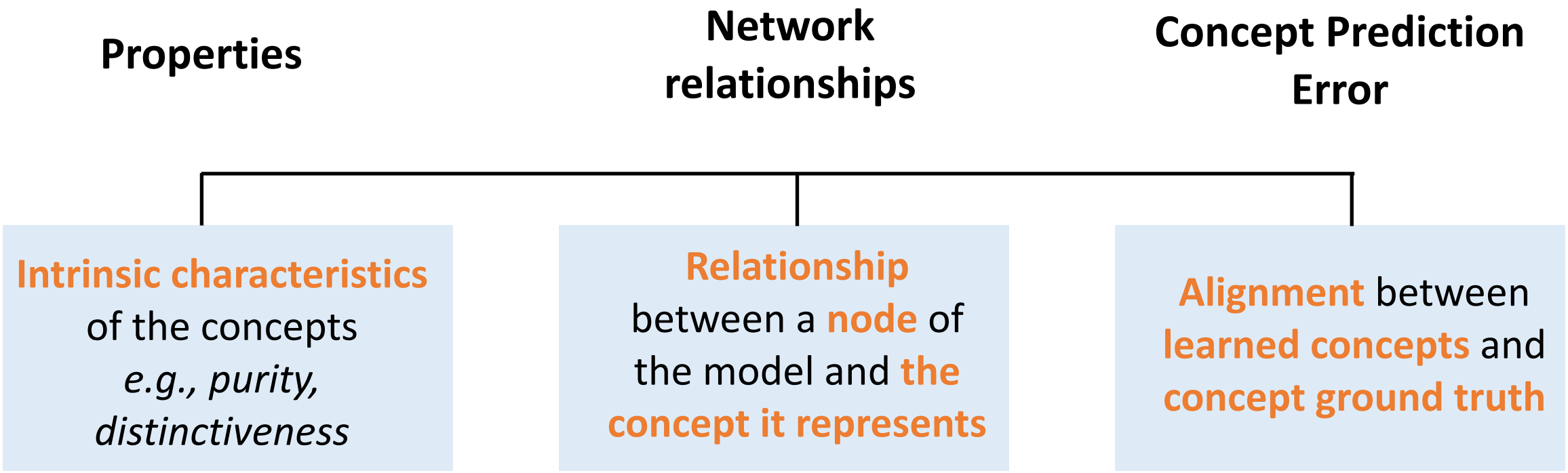
Concept effect on task performance

Measure the **concepts' contribution** to the predictive capacity of the model

Quantitative Evaluation

2. Quality of Concepts

*What is the **quality of the learned concepts**?*



Qualitative Evaluation

Human-centered assessment

Understandability & Trust

Plausibility, Trustability, Reasonability & Factuality

Coherence with Human Concepts

Assess if **learned concepts** group **similar** examples

Utility for decision-making

Practical usefulness and **effectiveness** in providing insights into the model

Quantitative

Strengths.

- Automated
- Scalable
- Reproducible

Limitations.

- Focuses on model-side importance, not human meaning
- Limited when concept labels are not available.

Qualitative

Strengths.

- Reflects real human understanding
- Captures trust and usefulness

Limitations.

- Costly, time-consuming, hard to scale.
- Lack of reproducibility or comparability due to subjectivity.

They are needed together!

Datasets for Concept-based xAI

Concept annotation data

Ground-truth annotation for (symbolic) concepts (e.g., color, texture, object parts)

Why concept datasets matter?

- **Enable supervised learning**

  Models can be trained to recognize concepts explicitly

- **Allow evaluation**

  Ground-truth concepts allow quantitative evaluation

- **Bridge human ↔ model reasoning**

  Ensure concepts are meaningful to users, not only to the network

Datasets for Concept-based xAI

Available for **different modalities**

 **Image**

CUB (200 bird species, 100+ attributes), BRODEN (1000+ visual concepts), OAI.

 **Text**

CEBaB, SST/IMDB.

 **Video**

BDD-OIA (driving actions, 21 concept types).

 **Tabular**

COMPAS (recidivism), MIMIC-II (patient survival).

Datasets for Concept-based xAI

Limitations.

- **Scarcity** of explicit concept labels

  Few have explicit annotation, most infer concepts indirectly

- **Limited** coverage **beyond images**

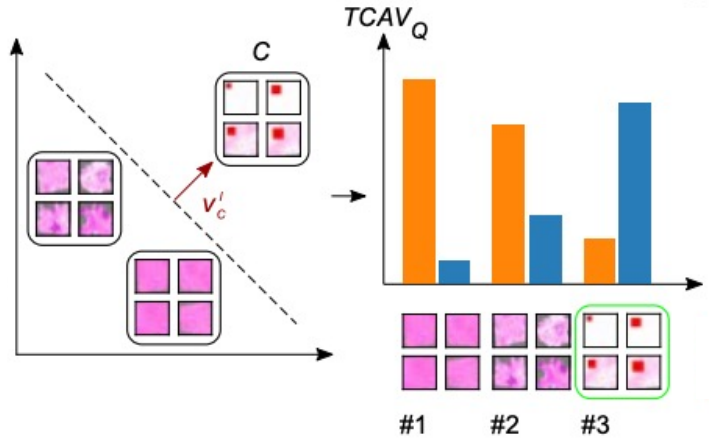
- **Lack of standardization** across modalities

Examples of application of C-XAI methods

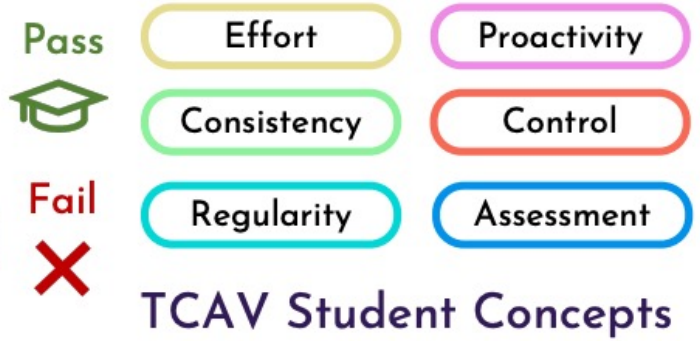
Medical.
Stronger adoption of concept-based XAI

Education.
Makes the model's reasoning much more *interpretable* for educators.

e.g., *TCAV* has been applied to skin cancer diagnosis



e.g., *TCAV* adapted to GNN to interpret student interaction patterns

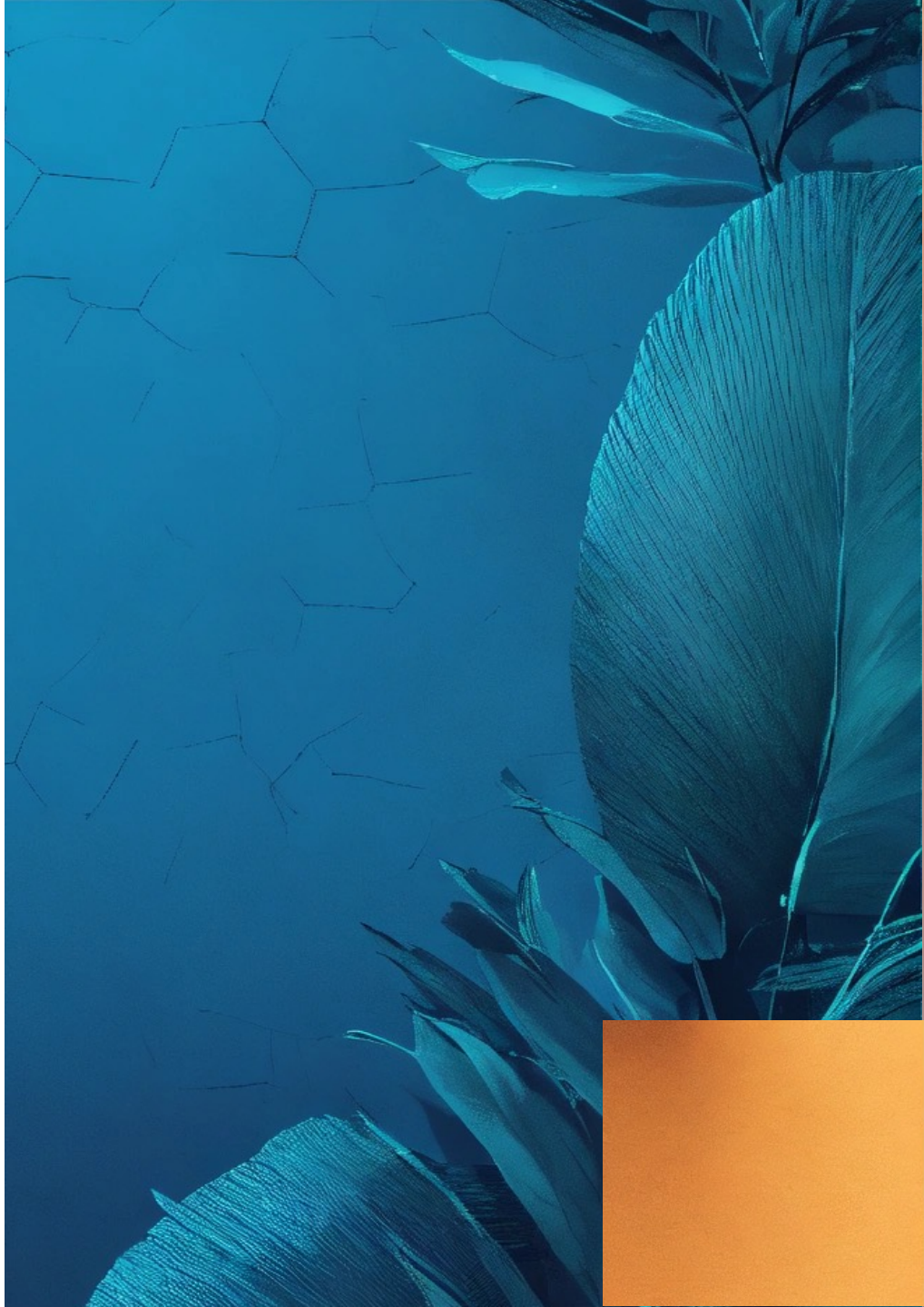


Trends and future directions

- Integration of **foundation and generative models** into C-XAI
  - Increasingly used to **support explanation tasks** and for **concept generation**
- Need for **standardize evaluation metrics**
- Need for **concept-based datasets**
  - Explicit, standardized, multi-modal concept datasets

Tutorial structure

- Introduction to XAI techniques
- Taxonomy of C-XAI
- Evaluation, Resources and applications
- **Hands-on session**



Hands-on time!

https://github.com/overview-cxai-tutorial-ecml-pkdd-2025/hands-on-cxai-ecml-pkdd

Github Repository

All notebooks will run in Google Colab –

no local setup required 😊

- **01** : Testing with Concept Activation Vectors (TCAV)
- **02** : Concept Bottleneck Model (CBM)
- **03** : Label Free Concept Bottleneck Model (LF-CBM)

Politecnico di Torino

CENTAI

Thank you!

Future Artificial Intelligence Research

ECML PKDD 2025