Concept-based Explainable AI

Explainable and Trustworthy AI

Gabriele Ciravegna

OUTLINE

1. Motivation
2. Concept-based eXplainable AI (C-XAI)
3. Testing with Concept Activation Vectors (T-CAV)
4. Concept Bottleneck Models (CBM)
5. Concept Embedding Models (CEM)

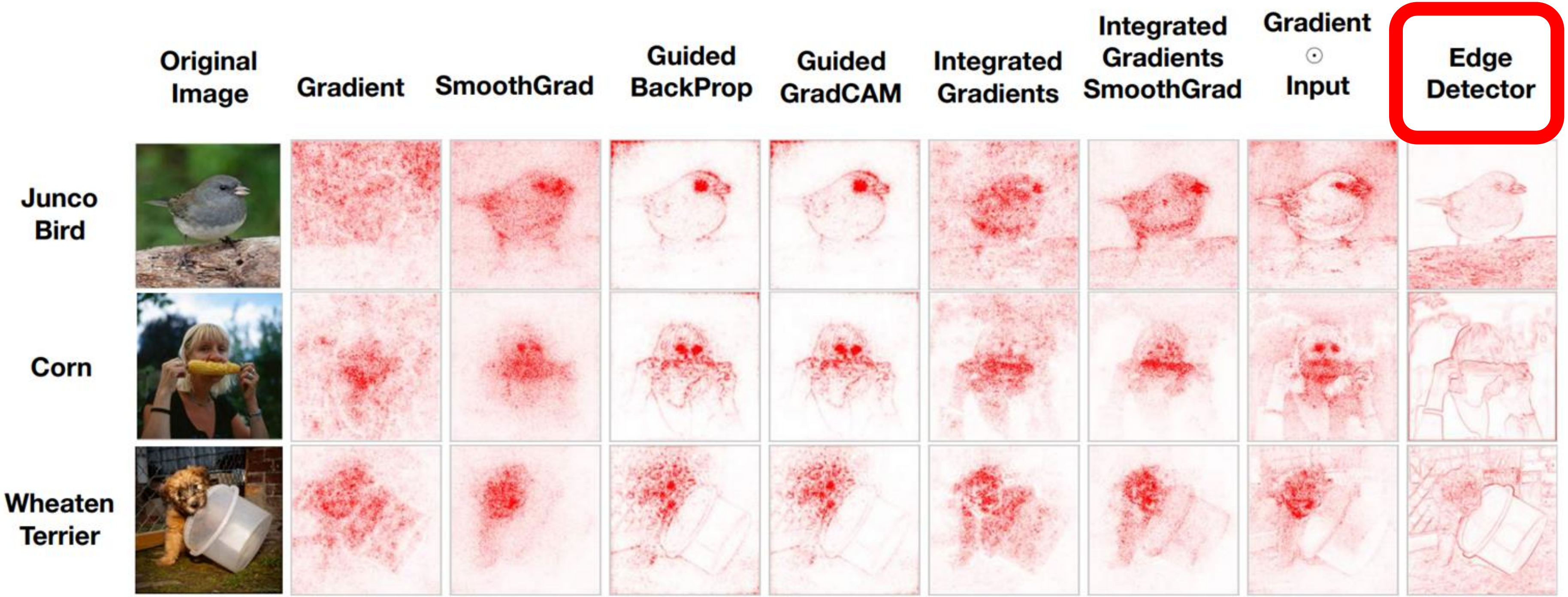
C-XAI PART I

C-XAI PART II

1. Motivation

Standard Explainable AI does **not** always work well

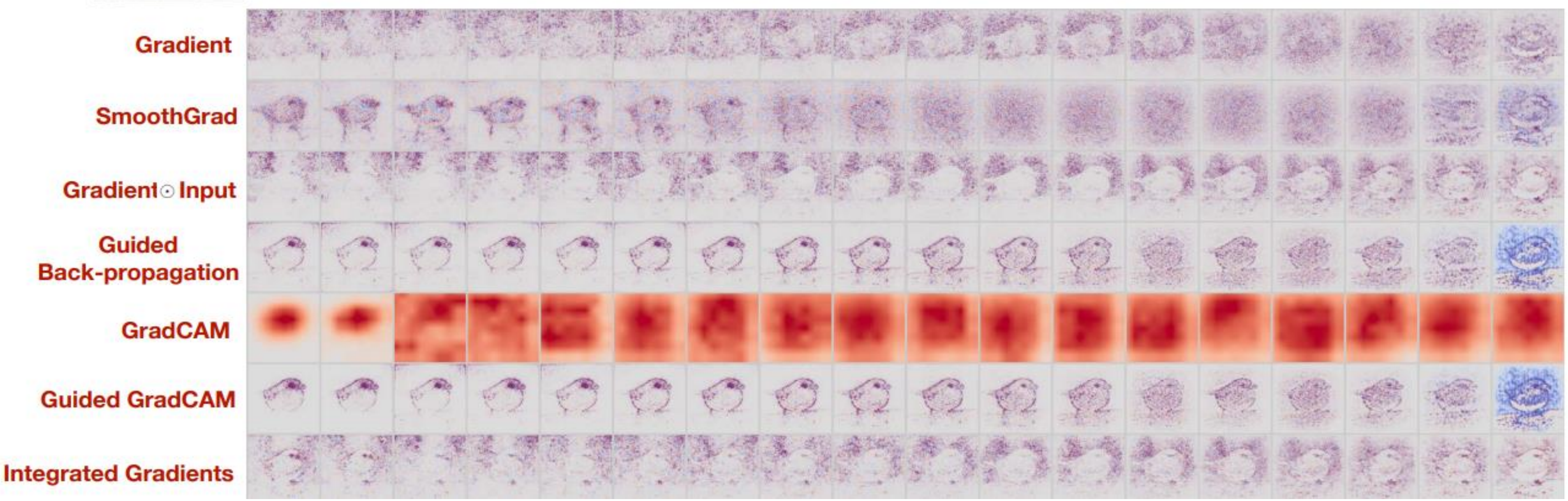
Which method is better? [1]



[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

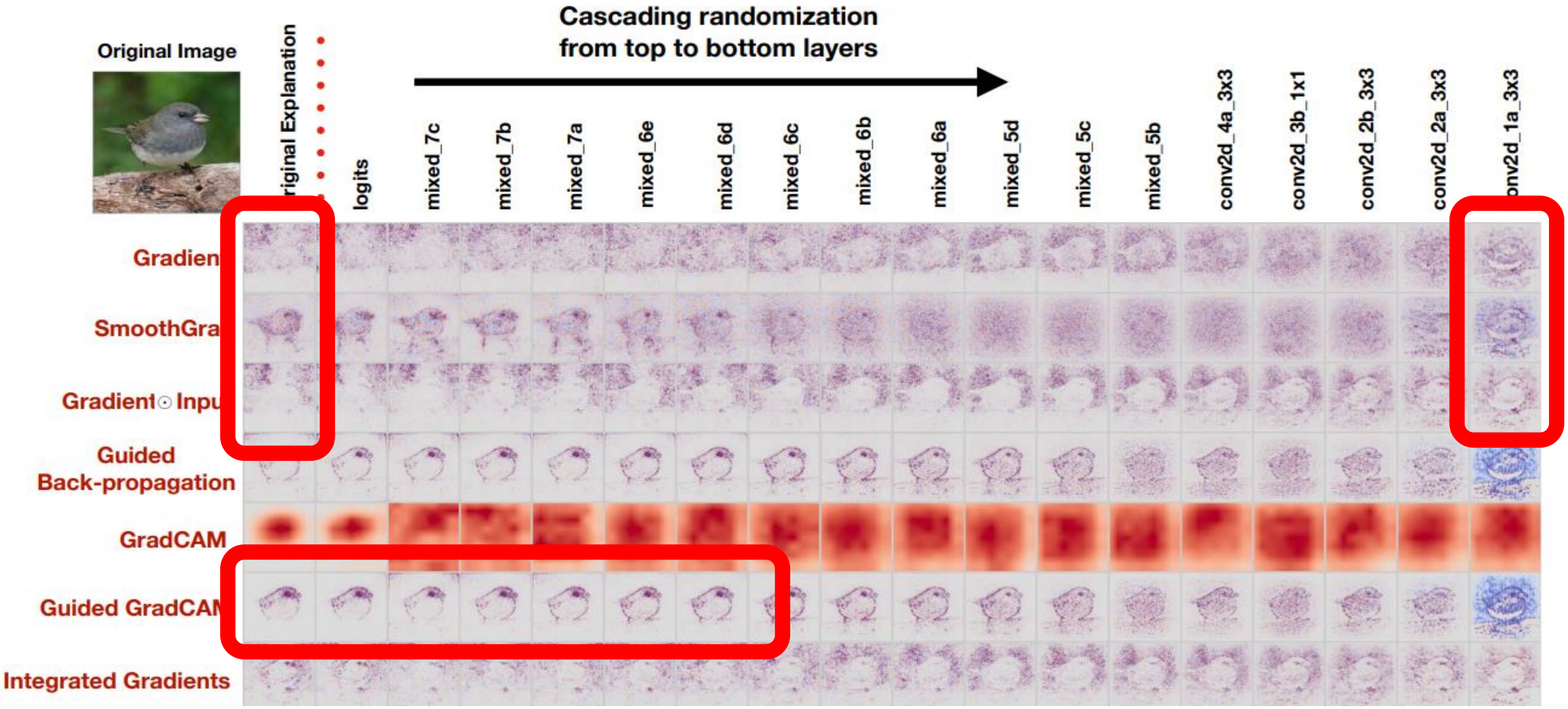
- It is not easy to assess which explanation method is better by only looking at the saliency maps
- Edge detectors produce similar explanations to some saliency maps (particularly those considering the input values, e.g., Gradient x Input)

Towards where are we randomizing? [1]



[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Towards where are we randomizing? [1]



[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

- Randomizing a few layers does not have almost any effect on the explanation
- The explanation of a completely randomized network is still similar to the original one
- It is difficult to understand which layer is being randomized

❌ Some explanation methods are more input-dependent than model dependent

Which class are we explaining? [3]

“Siberian Husky” “Transverse Flute”

[3] Rudin, Cynthia. "Stop explaining black box machine learning models …” Nature machine intelligence (2019)

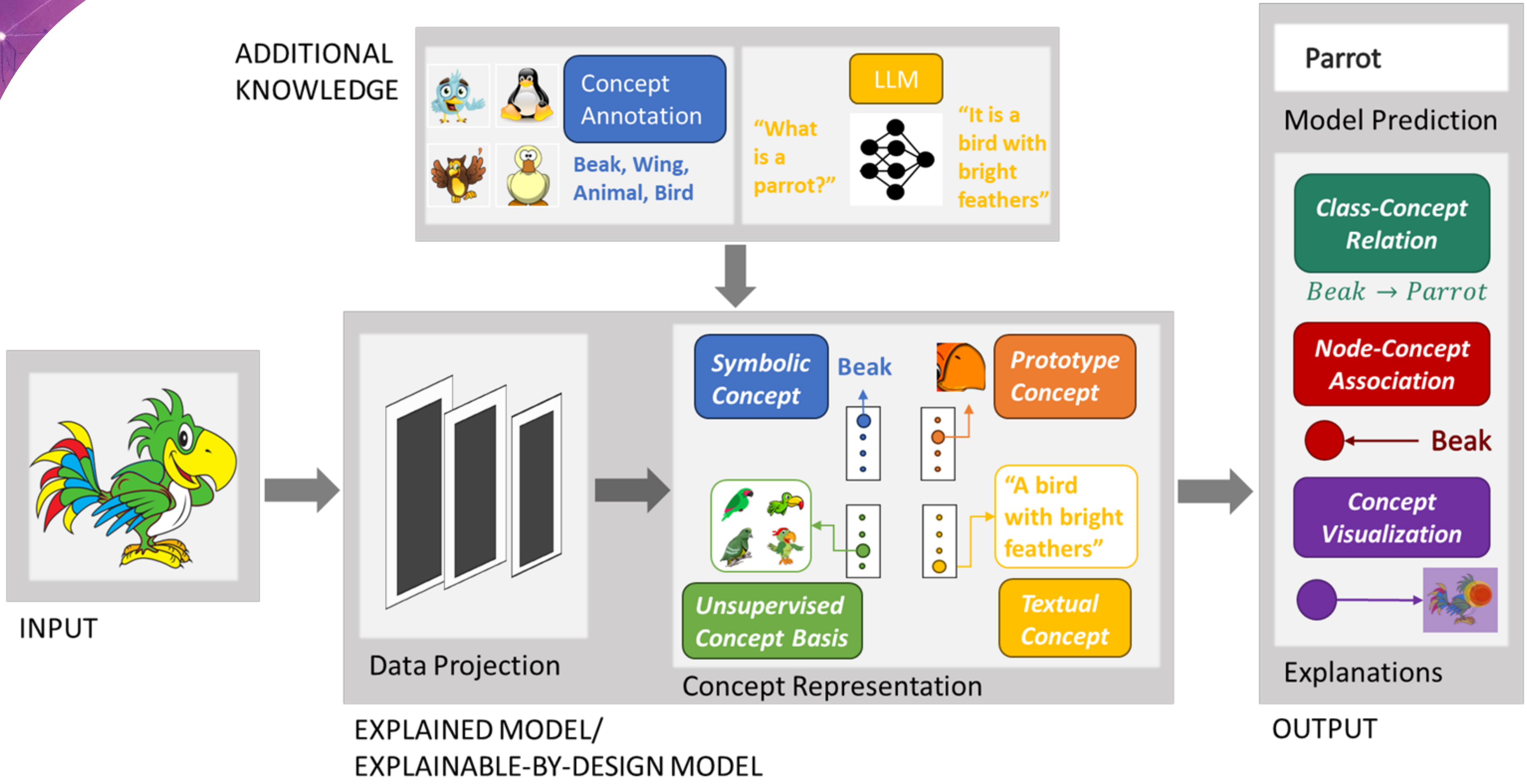
- It is difficult to determine the explained class only looking at the saliency maps
- Saliency maps of very different classes can be still similar

Why XAI explanations are difficult to understand?

“Showing **where** a network is looking
does not tell us **what** the network is seeing
in a given input” [3, 4]

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

2. Concept-based Explainable AI (C-XAI)


ADDITIONAL KNOWLEDGE
Concept Annotation
Beak, Wing, Animal, Bird
LLM
"What is a parrot?"
"It is a bird with bright feathers"
INPUT
Data Projection
Symbolic Concept
Beak
Prototype Concept
"A bird with bright feathers"
Unsupervised Concept Basis
Textual Concept
Concept Representation
EXPLAINED MODEL/
EXPLAINABLE-BY-DESIGN MODEL
Parrot
Model Prediction
Class-Concept Relation
Beak → Parrot
Node-Concept Association
Beak
Concept Visualization
Explanations
OUTPUT

What is a Concept?

"A concept can be any abstraction, such as a colour, an object, or even an idea"[9]

Let's try to be more concrete…

[9] Molnar, Christoph. "Interpretable machine learning". (2020)

Different types of concepts

1. **Symbolic Concepts**

   Human-defined attributes

   “BEAK”

2. **Unsupervised Concept Basis**

   Cluster of similar samples

3. **Prototypes**

   (Part-of) a training sample

4. **Textual Concepts**

   Textual representation of a main class

   “A bird with bright feathers”

Symbolic Concepts

"BEAK"

- 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
  - Image-level annotation
    - Annotate the presence for each image of a concept
    - 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 (a.k.a, the latent space)

- Not built to resemble human-defined concepts
  - Still capture abstractions more understandable to humans than individual features or pixels
  - E.g., a cluster of green birds.

- Clustering algorithms must employed to extract unsupervised concepts

Prototypes

- Explanation by Example
  - It will be better explained in the remaining of the course
- Representative examples of peculiar traits of the training samples
  - Entire samples
  - Parts of a training sample (e.g., a particular type of beak)
- The set of prototypes should be
  representative of the whole data set

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

Textual Concepts

- Textual descriptions of main classes
    - From an individual description, distinctive pieces are extracted
    - Each piece embodies a characteristic of the corresponding class
    - It can be shared among different classes (e.g., a bird with bright feathers)

- Provided at training time by means of an external generative model
    - It requires a Large-Language Models LLMs
      with knowledge of the given task

- Employed in the form of a numerical embedding
    - of the corresponding text

"A bird
with bright
feathers"

Concept-based Explanations

1. **Class-Concept Relations**
   Relation among a concept and an output class of a model

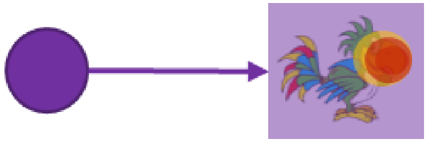
$$Beak \rightarrow Parrot$$

2. **Node-Concept Association**

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

3. **Concept-Visualization**

   Visualization of a learnt concept in terms of the input features



Class-Concept Relations

- Relationship between a specific concept and an output class of the model $Beak \rightarrow Parrot$
  - Concept importance
  - Logic rule involving multiple concepts and their connection to an output class

- Can be applied to all type of concepts:
  - E.g., with prototypes, we have parrot := 0.8 prototype_1 + 0.2 prototype_2

Node-Concept Association

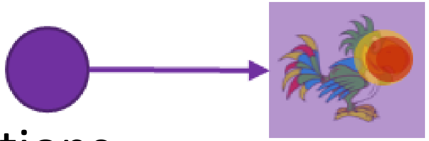
- Assign a concept to an internal unit (or a filter) of a network
- It enhances the transparency of deep learning models
  - highlighting what internal units see in a given sample.

- 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

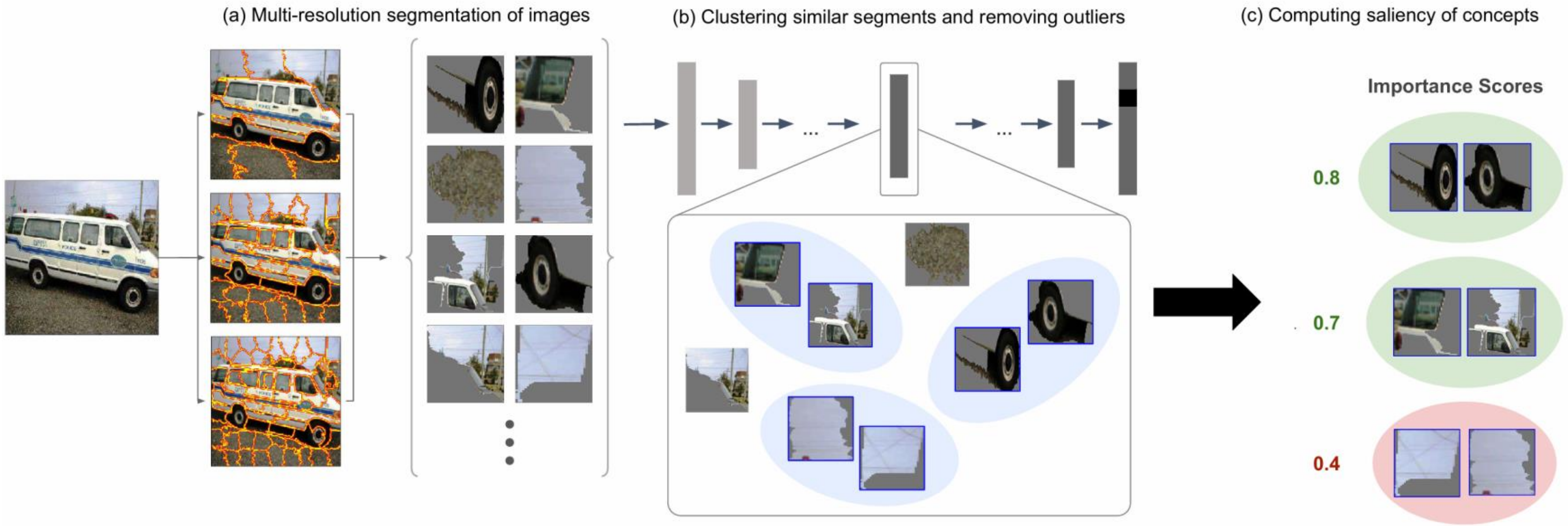
- Crucial when non-symbolic concepts are employed
  - Need to understand which unsupervised attributes or prototypes the network has learned.



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

Post-hoc or Explainable-by-design?

Post-hoc Concept-based Explanations



Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. Towards automatic concept-based explanations. NeurIPS 2019

Post-hoc Concept-based Explanation methods

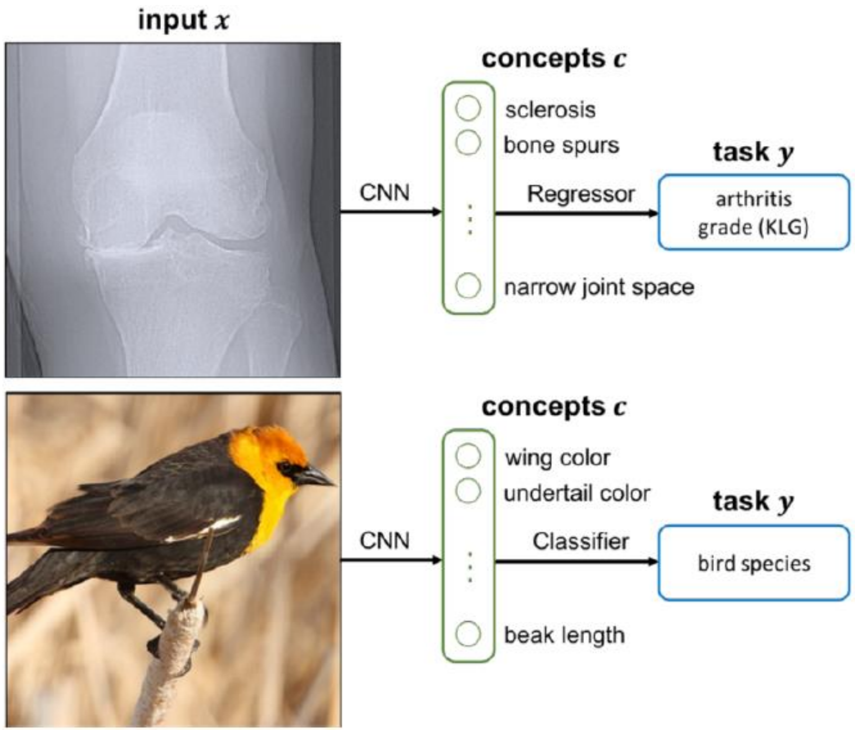
- Standard pipeline:
  - Project samples representing the concepts in the model latent space
  - Analyze their relationship to the prediction (or the hidden node activations)

- Concepts employed can be supervised or unsupervised
  - Prototypes and generative concept have not been employed so far

- Pros:
  - They don’t compromise the learning capacity of a model
  - They 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)

Explainable-by-design Concept-based Models



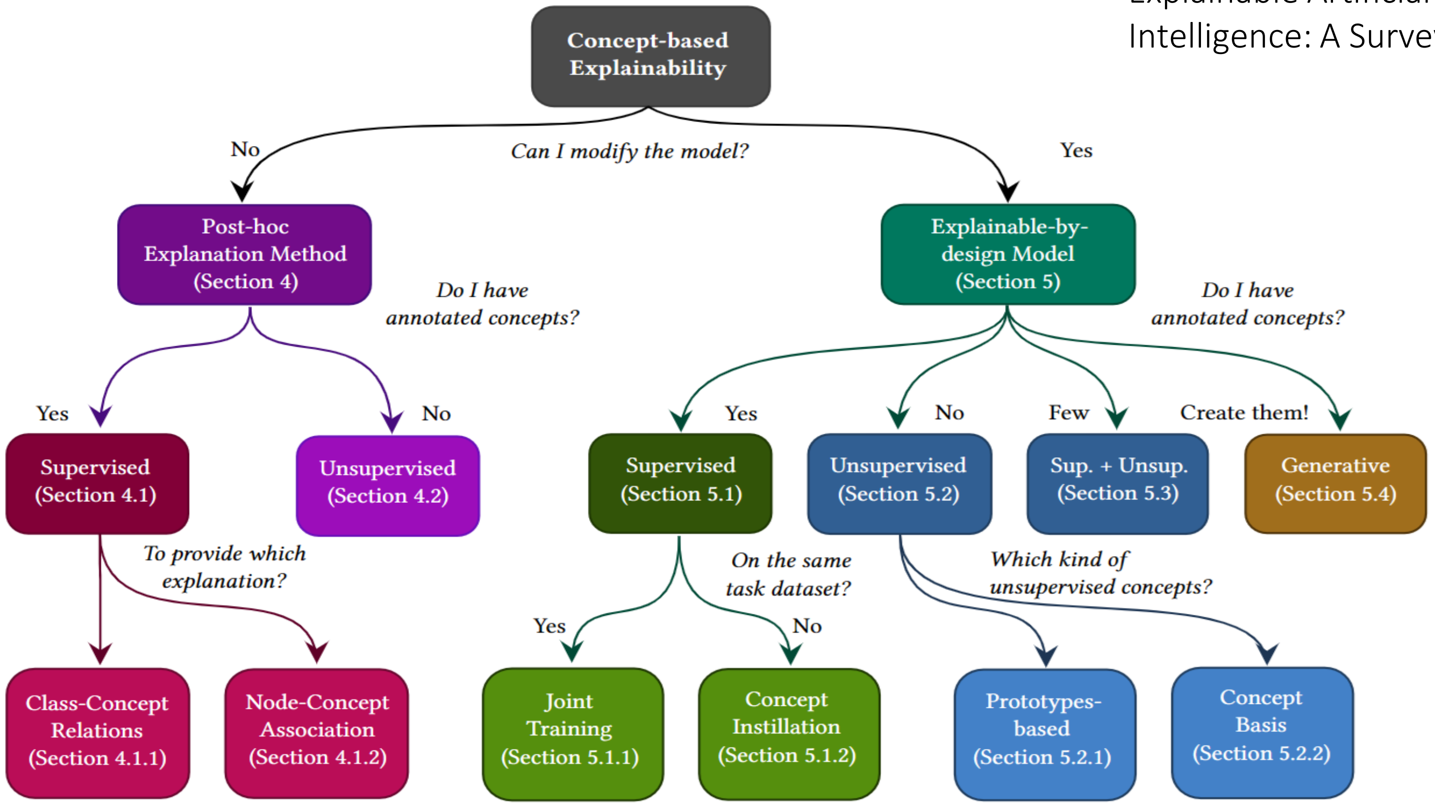
Koh, Pang W, et al. "Concept bottleneck models." *ICML 2020*.

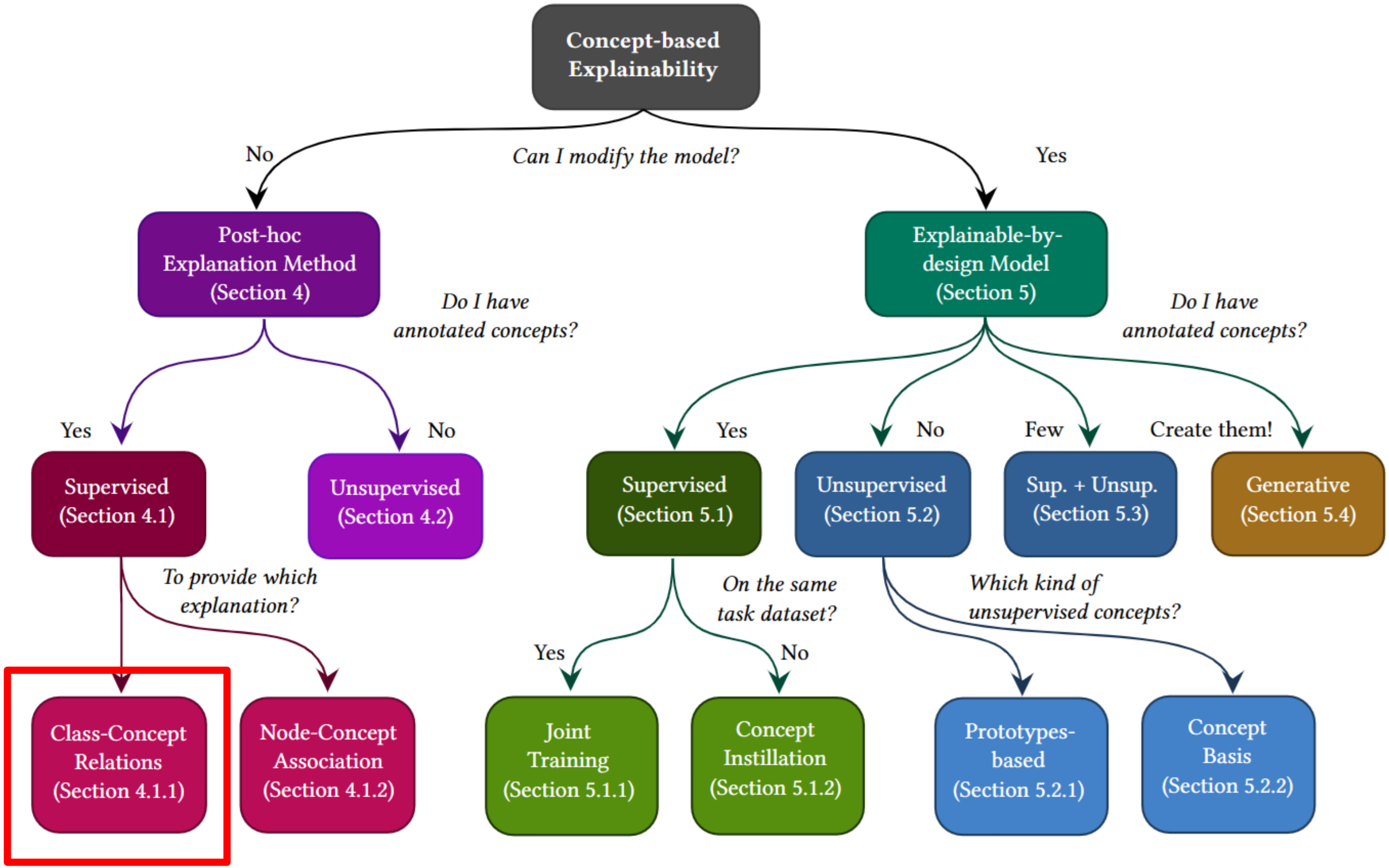
Explainable-by-design Concept-based Models

- Neural models with an explicit concept representation as an intermediate layer
- Predicted concepts influence the task predictions
- All types of concepts and explanation can be employed
- 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

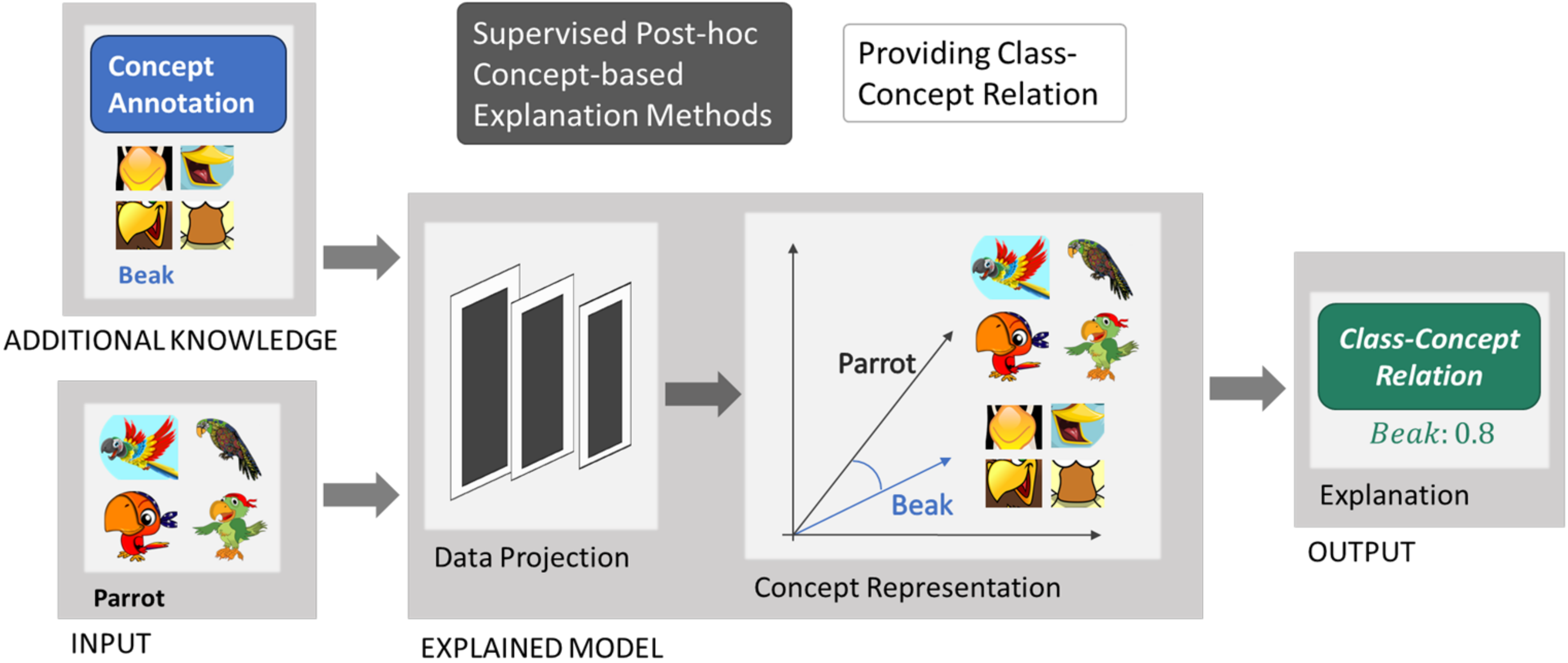
C-XAI Taxonomy

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





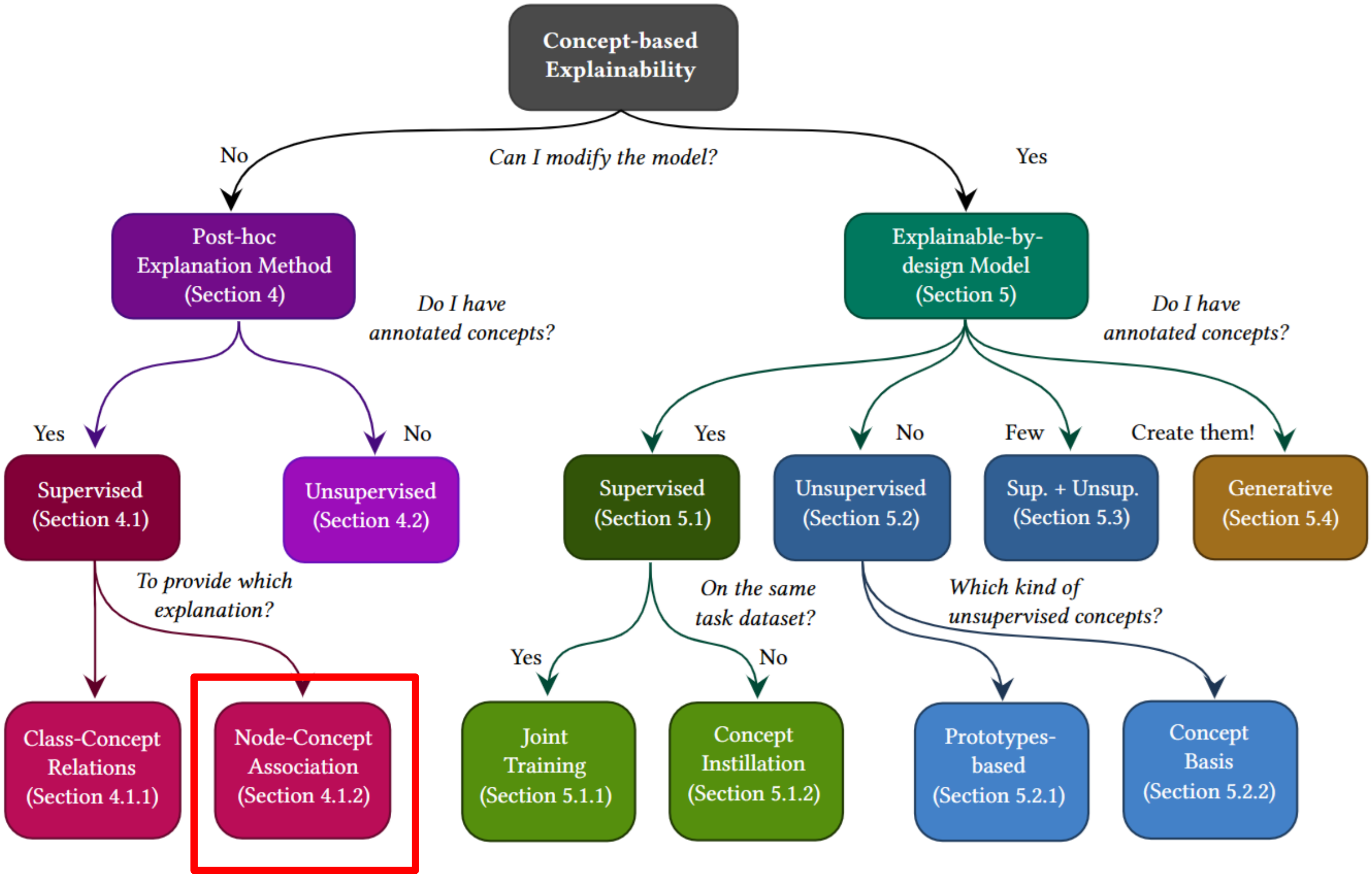
Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)

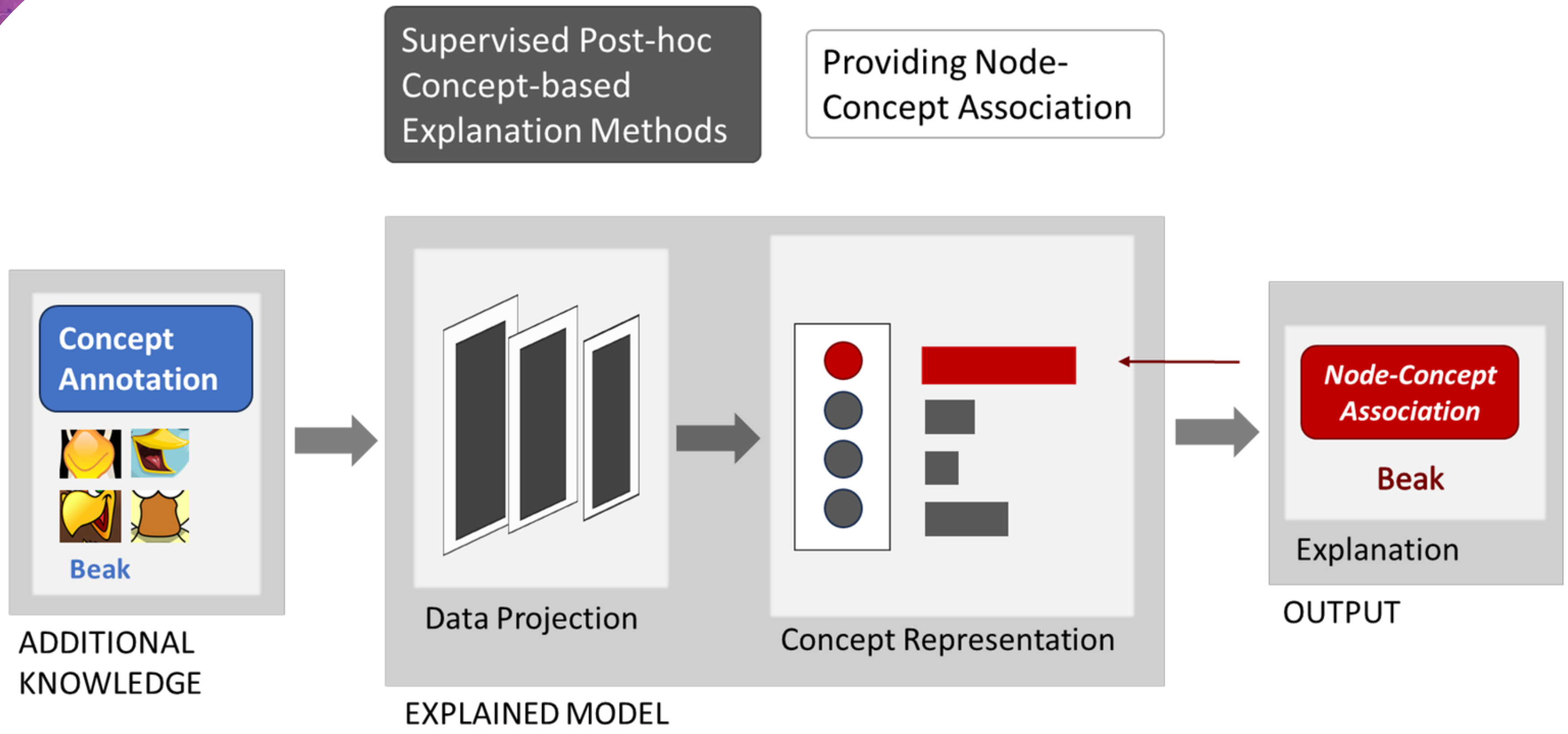


E.g., Kim, Been, et al. "*Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).*" International conference on machine learning. (2018).

Post-hoc supervised method providing class-concept relations

- Take a pre-trained model
- Require a set of data annotated with concepts
- Analyze the projection of these data into the model latent space
- They correlate the projection with those of the output classes


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Do I have annotated concepts?
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)



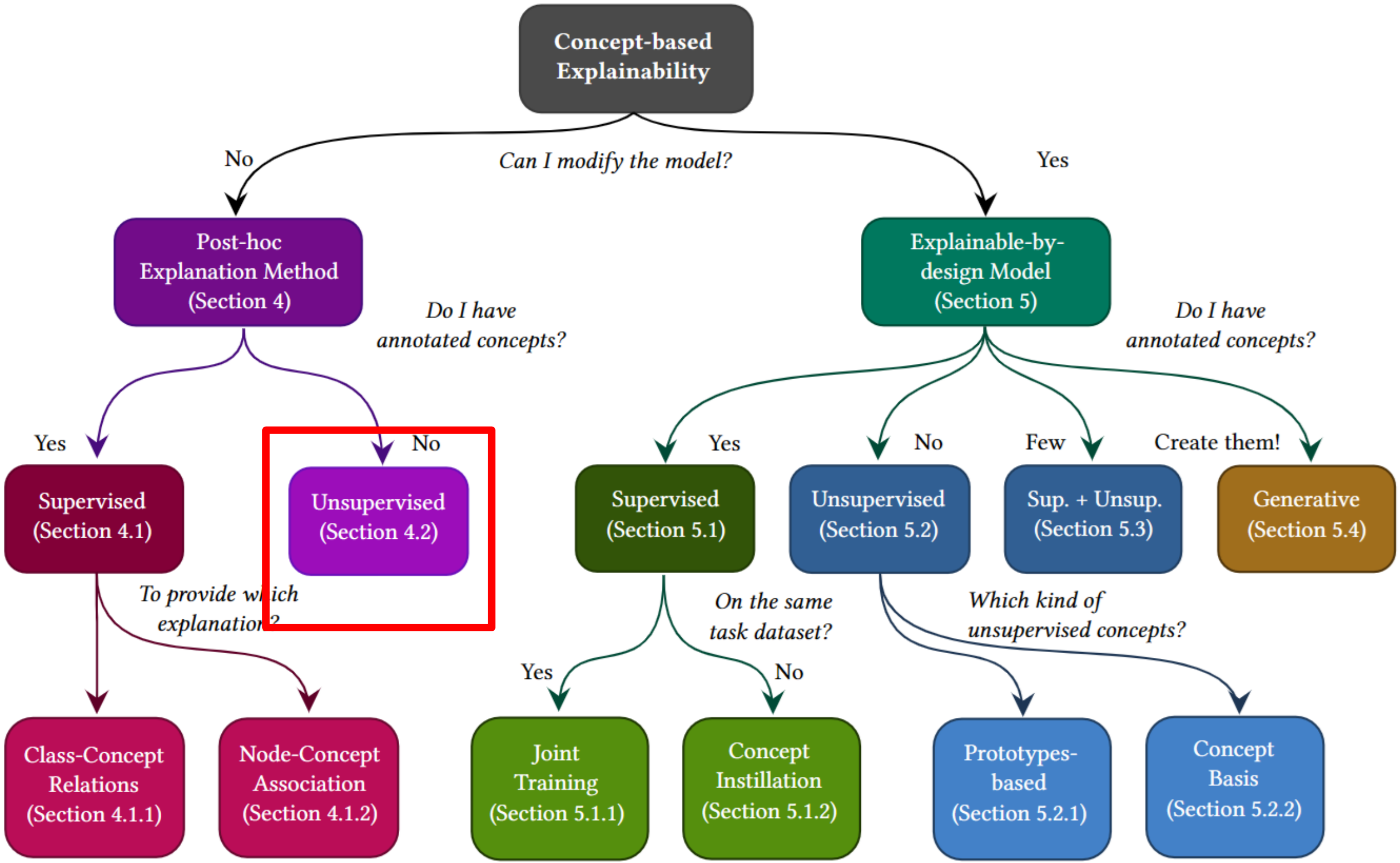
E.g., Bau, David, et al. "*Network dissection: Quantifying interpretability of deep visual representations.*" CVPR. 2017

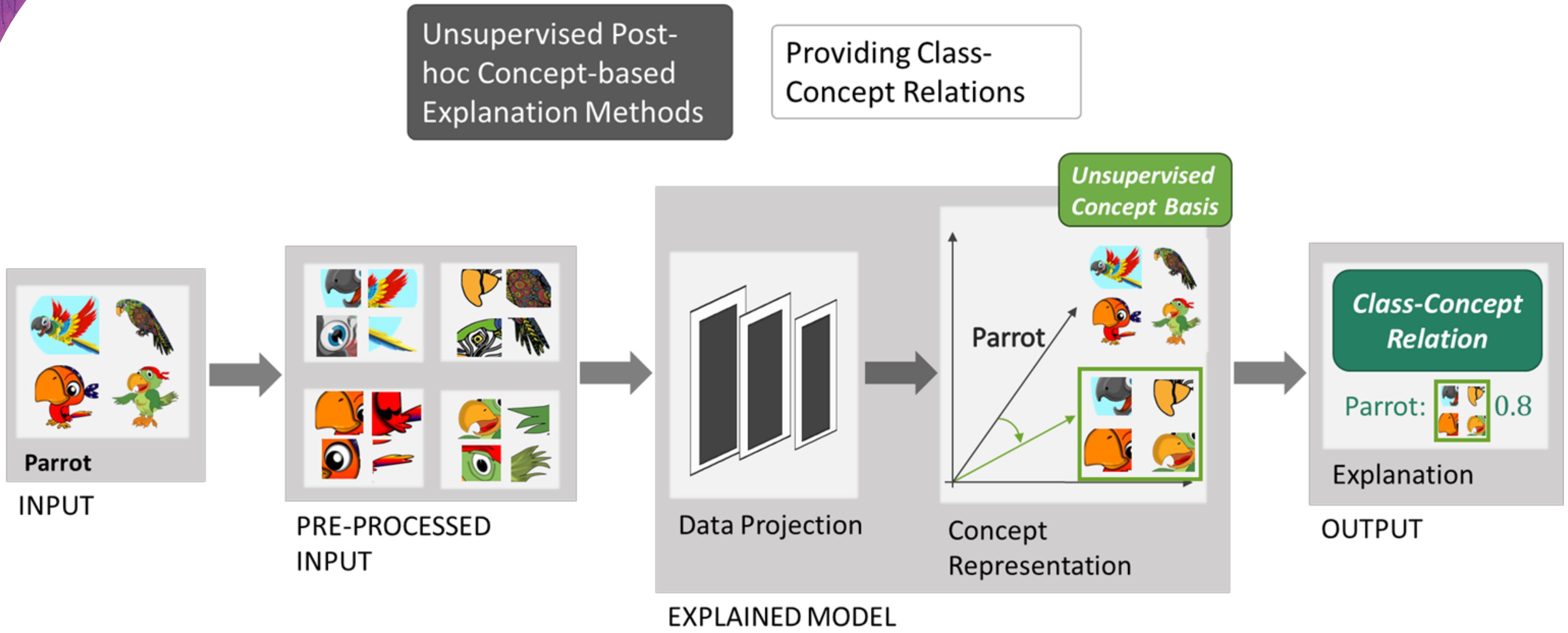
Post-hoc supervised method providing node-concept association

- Similarly to methods providing class-concept relations
  - Take a pre-trained model
  - Require a set of data annotated with concepts

- Analyze the activations of the hidden nodes when fed with these data

- They associate to each node the concept for which they activate the most (on average)

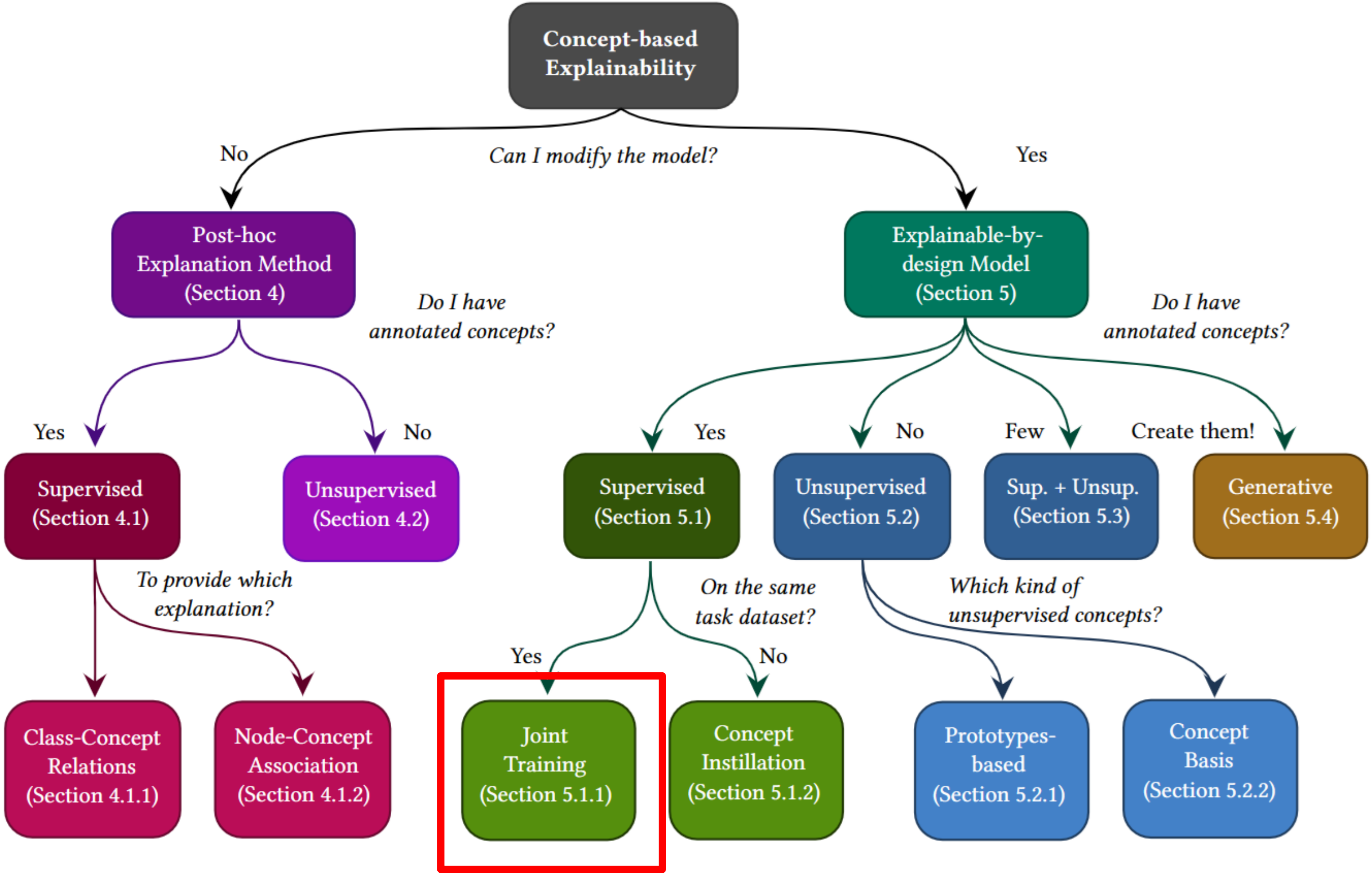

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
To provide which explanation?
On the same task dataset?
Which kind of unsupervised concepts?
Yes
No
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)



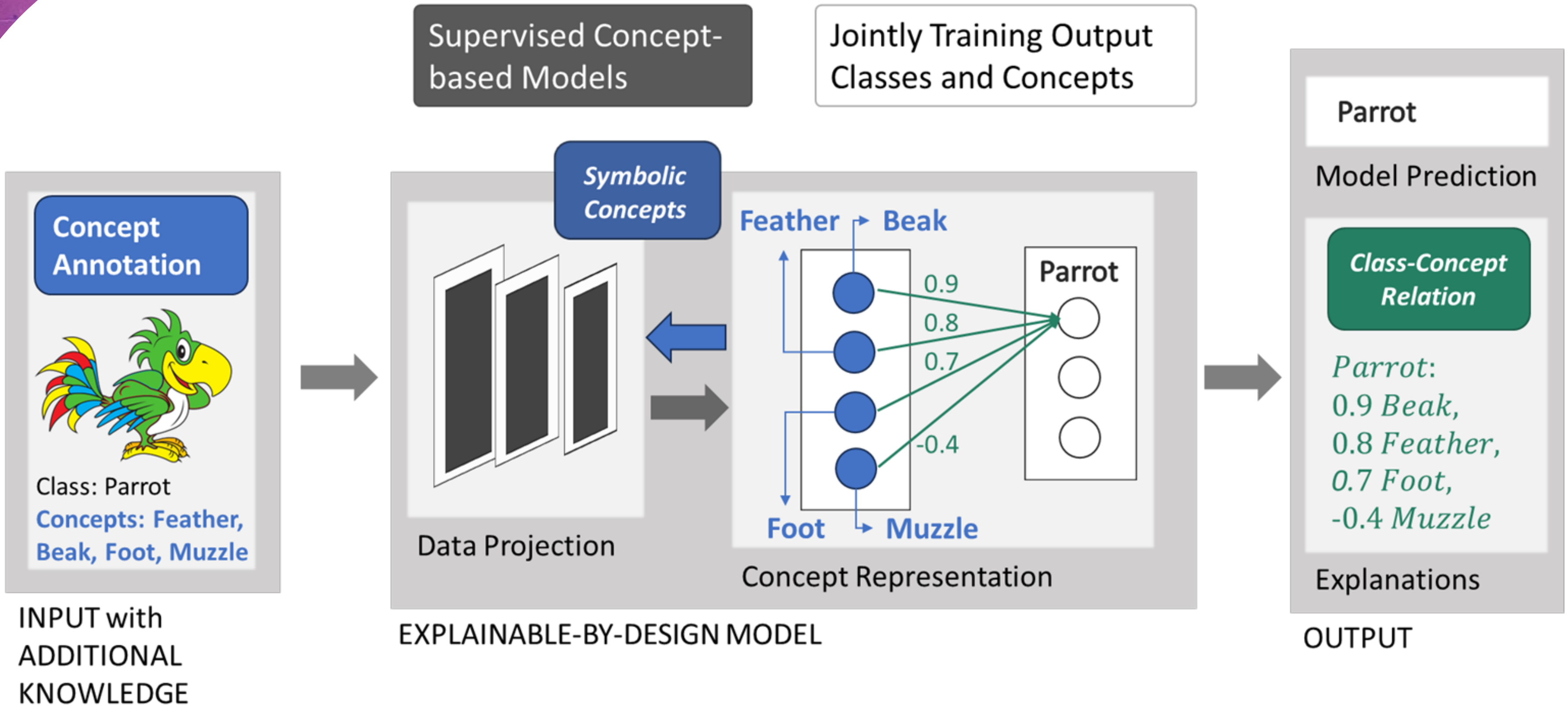
E.g., Ghorbani, Amirata, et al. "*Towards automatic concept-based explanations.*" NeurIPS (2019).

Post-hoc unsupervised method providing class-concept relation association

- Similarly to supervised methods take a pre-trained model
  - BUT: They don’t require a set of data annotated with concepts
- They split input data into smaller crops
- Analyze the projections of the crops in the latent space of the model
- They clusterize projections --> clusters are unsupervised concepts
- They analyze the correlation of unsupervised concepts with output classes/predictions



Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Do I have annotated concepts?
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)



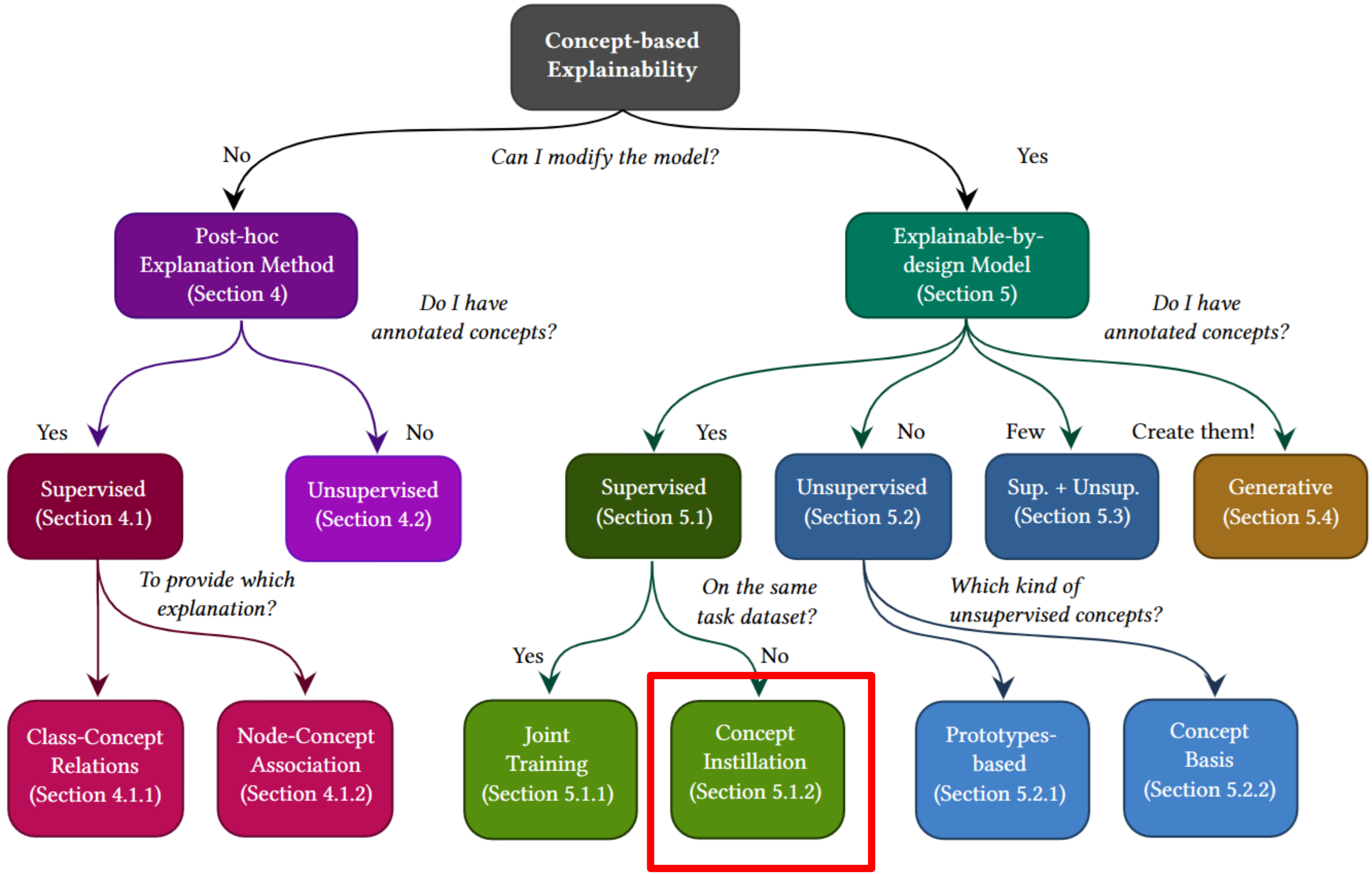
E.g., Koh, Pang Wei, et al. "*Concept bottleneck models*." ICML (2020)

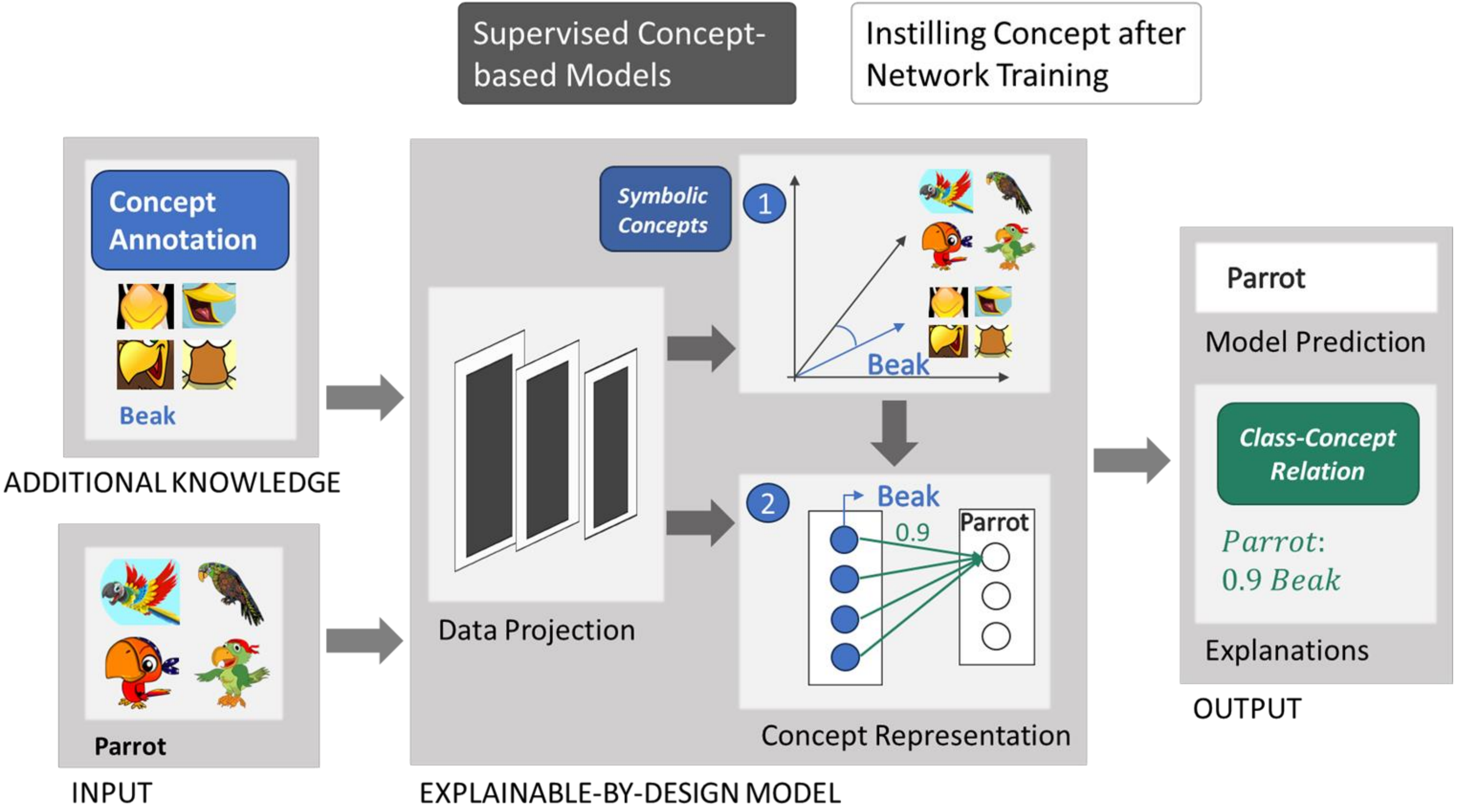
Supervised Concept-based models jointly training

- They train a model from scratch with a hidden layer predicting these concepts
  - Node-concept association by-design

- The predicted concepts are used to make the final prediction
  - If the «task predictor» is a white box model you can also extract class-concept relations

- Pros:
  - Very intuitive explanation («I see a beak, feathers and not a muzzle, it is a bird»)
  - They allow concept interventions and interacting with the model
- Cons:
  - They require a set of data annotated with both classes and concepts


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Do I have annotated concepts?
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)

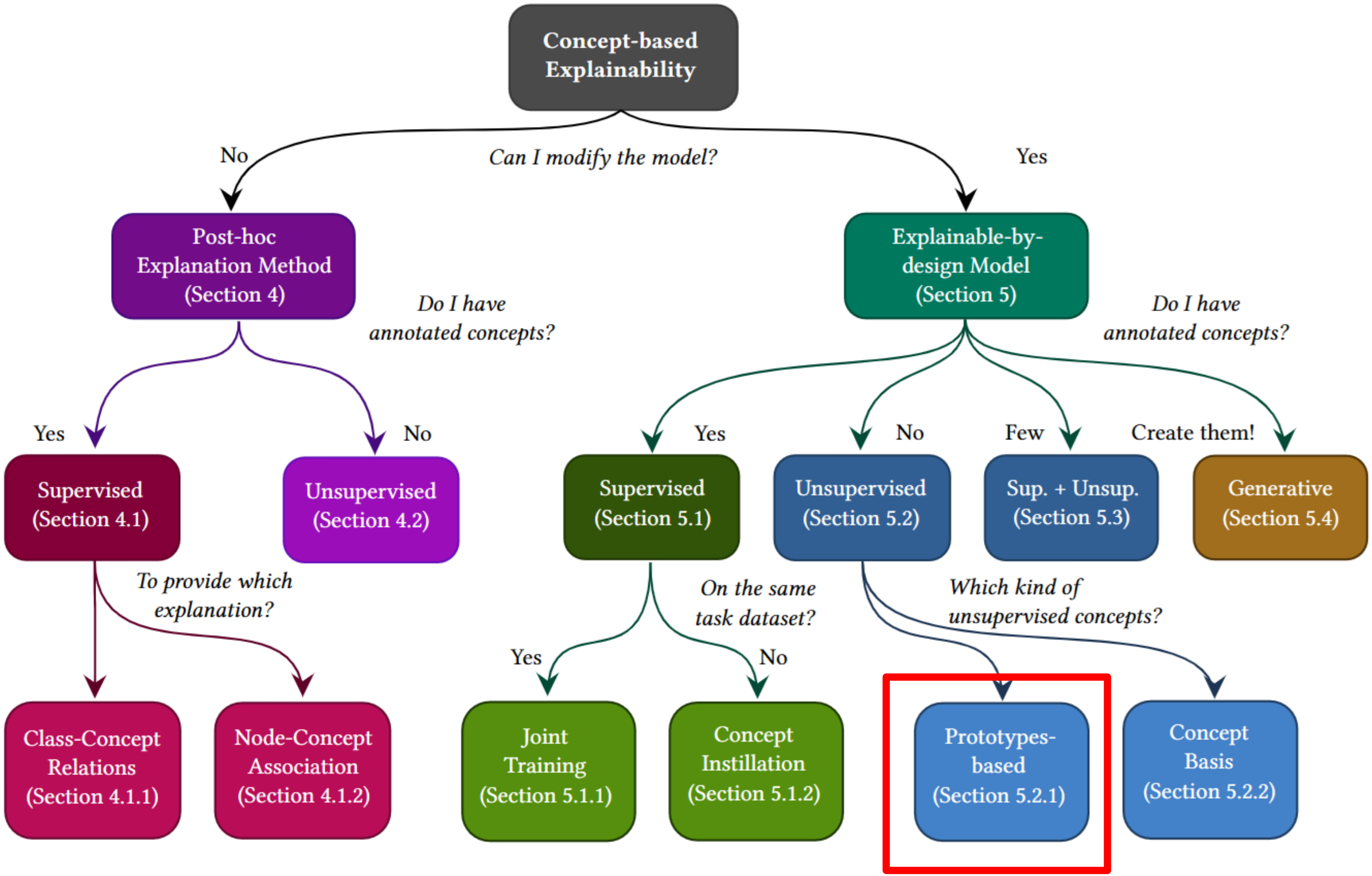


E.g., Chen, Zhi, et al. "*Concept whitening for interpretable image recognition.*" Nature Machine Intelligence (2020).

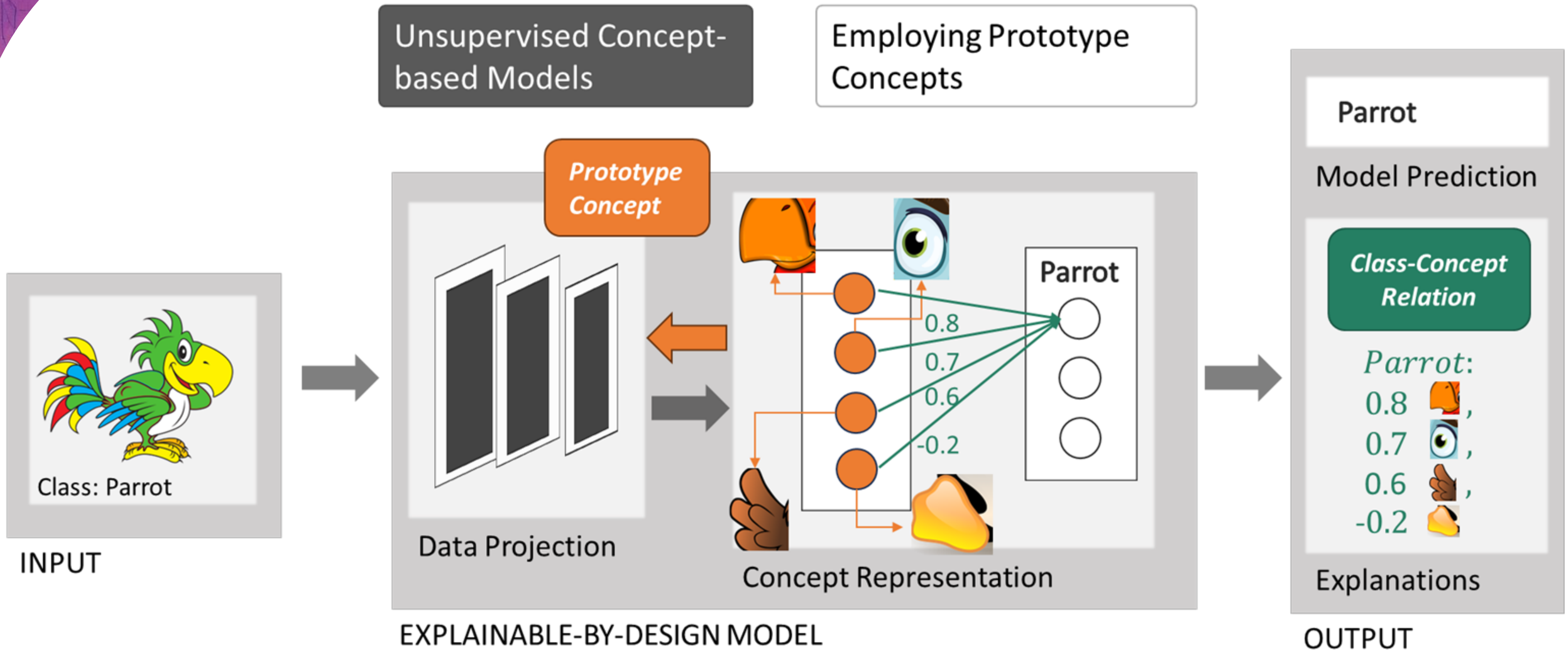
Supervised Concept-based models instilling concepts

- Differently from joint-training models:
  - Take a pre-trained model
  - The set of data annotated with the concepts may not be the same of the training data

- They turn a black-box model into an explainable-by-design one:
  - They fine-tune a certain layer to predict for the given concepts
  - They keep training the top of the network to predict the original classes



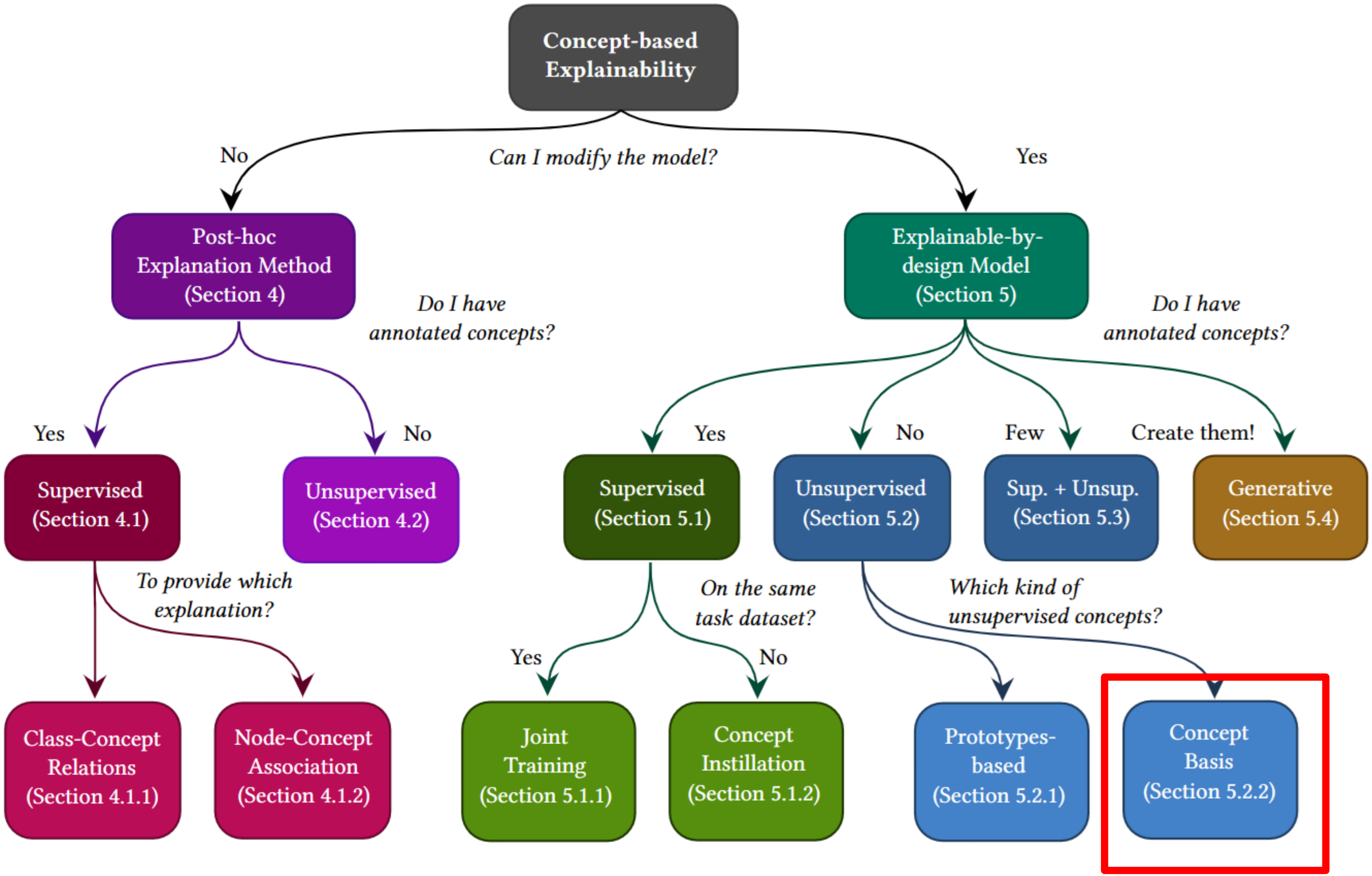
Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Do I have annotated concepts?
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)

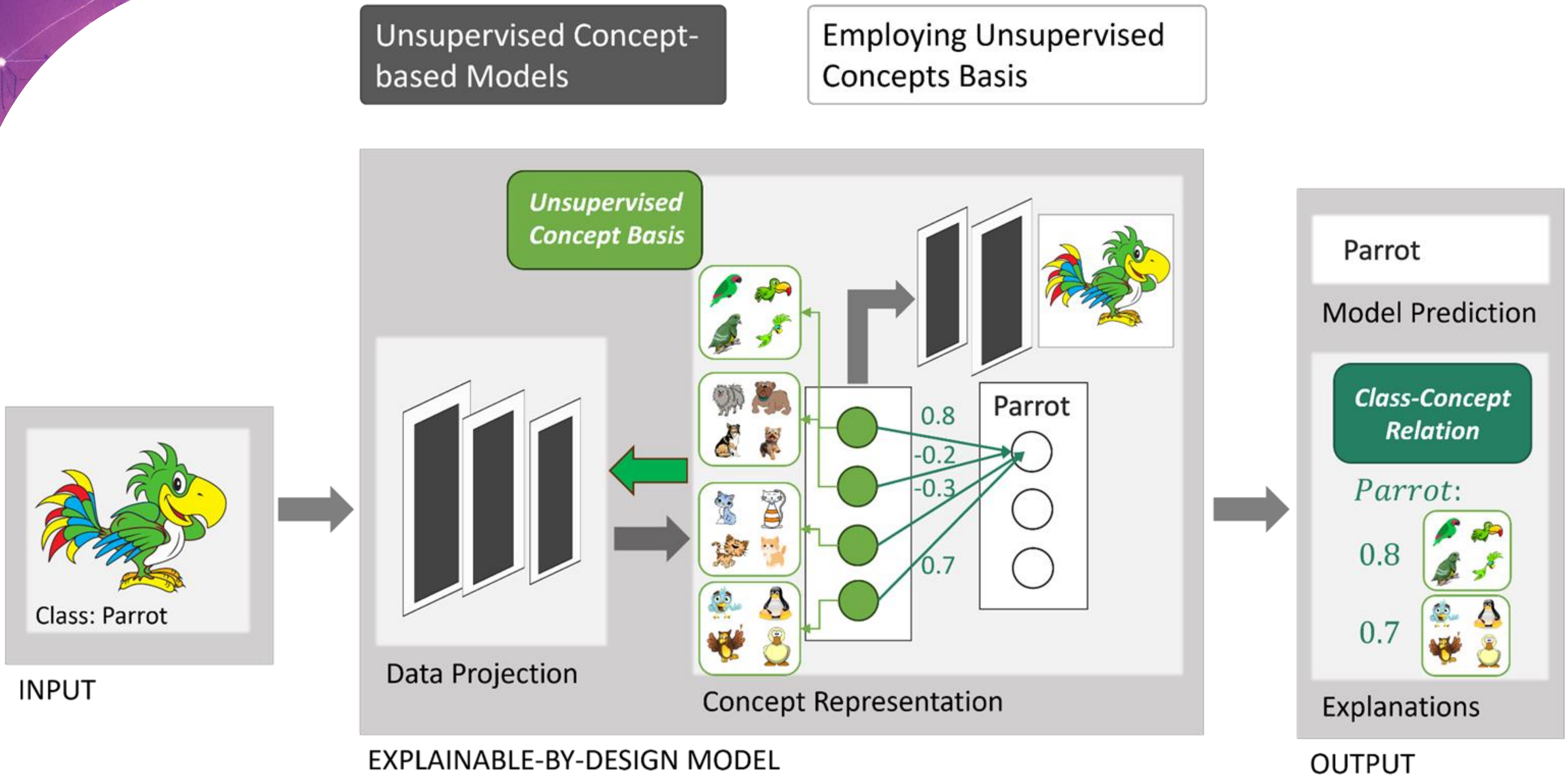


E.g., Chen, Chaofan, et al. "*This looks like that: deep learning for interpretable image recognition.*" NeurIPS 2019.

Unsupervised Concept-based models employing prototype concepts

- They don’t require annotated concepts
- They train the network to both:
  - Learn to predict the output class
  - Encode in the hidden layers the most representative training examples
- Again, explainable-by-design:
  - Node-concept association
  - Class-concept relations in case of a white-box task predictor
- To visualize the prototypes:
  - Check the (part of the) sample for which the protype activate the most

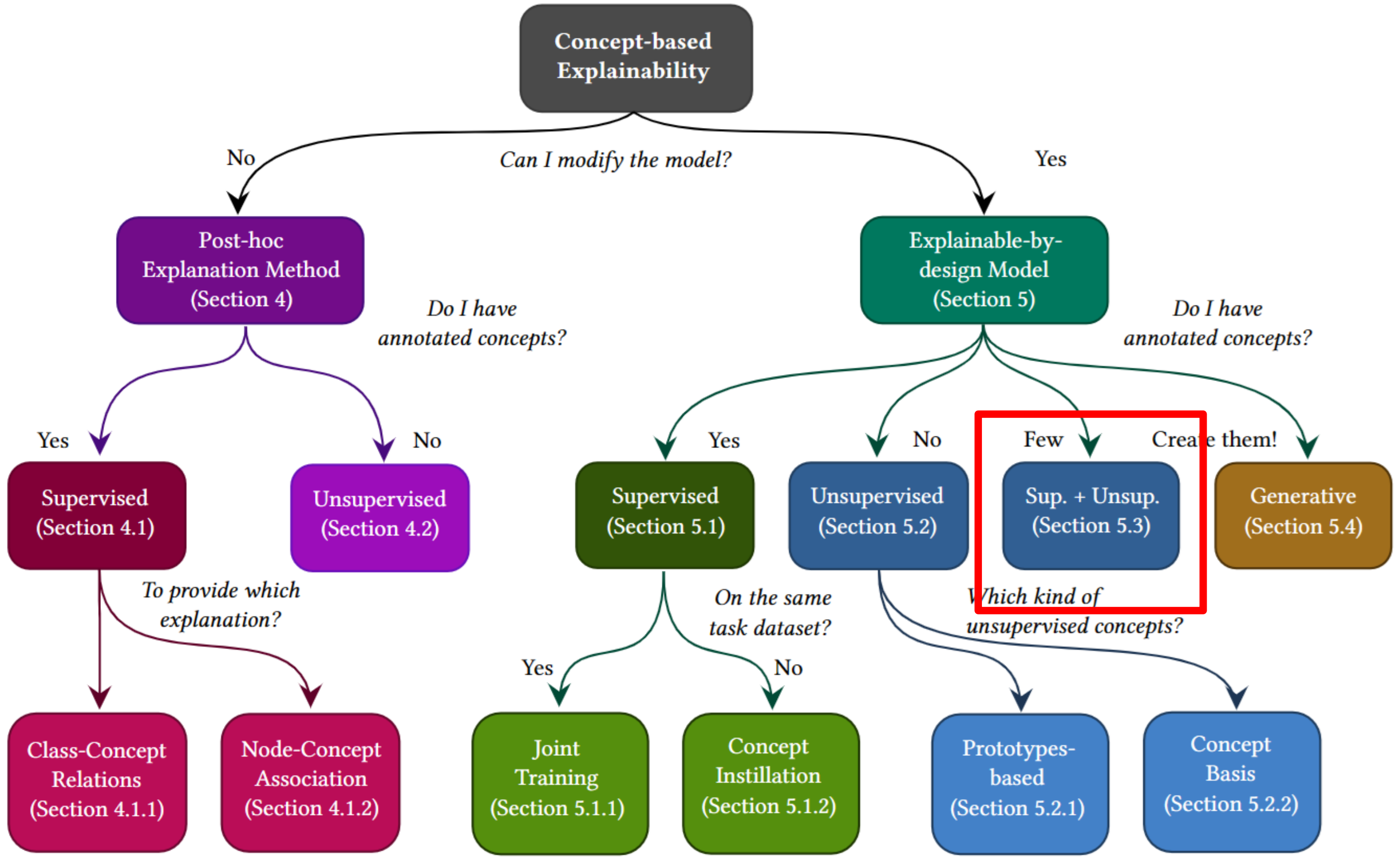

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
To provide which explanation?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Do I have annotated concepts?
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
On the same task dataset?
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Which kind of unsupervised concepts?
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)



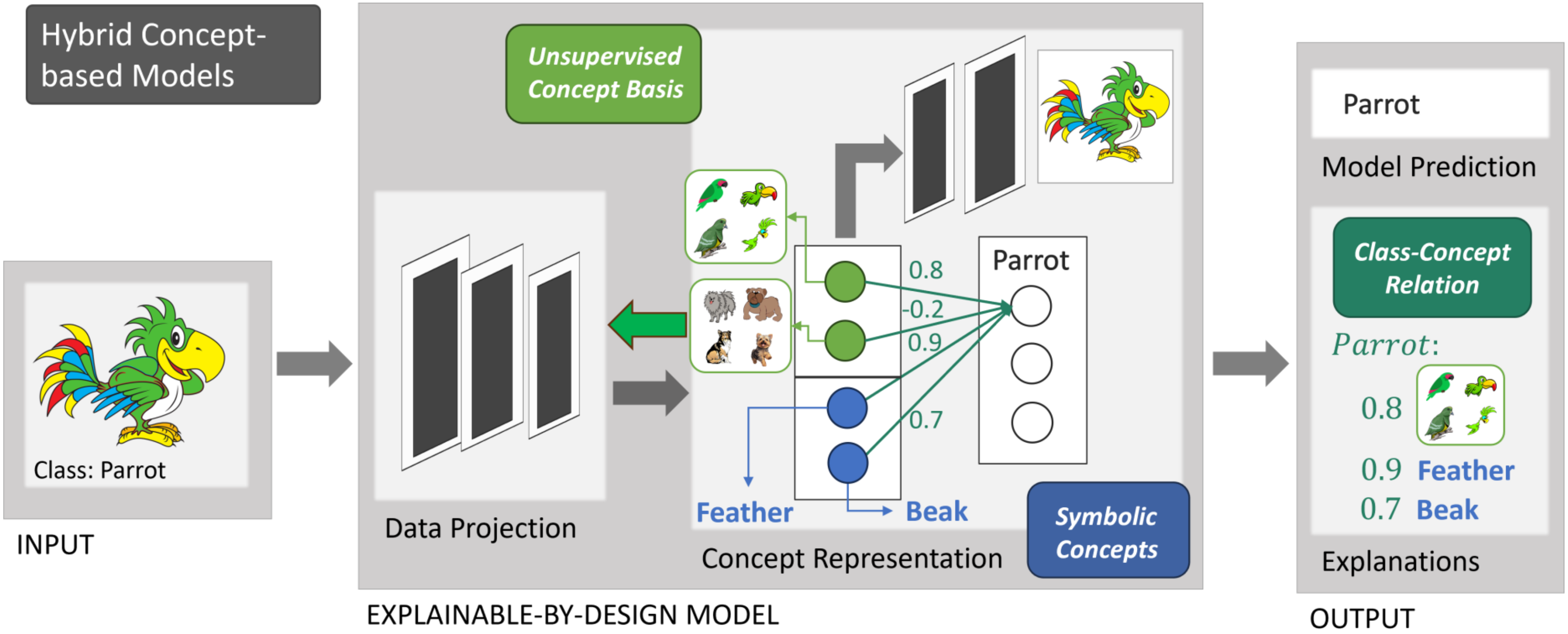
Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." Neurips 2018.

Unsupervised Concept-based models employing unsupervised concept basis

- They train the network to both:
  - Learn to predict the output class
  - Create cluster of samples in the latent representation
- Again, explainable-by-design:
  - Node-concept association
  - Class-concept relations in case of a white-box task predictor
- To characterize the unsupervised concepts:
  - Visualize the samples closest to the centroids
  - Decode the centroids if employing an auto-encoder



Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
To provide which explanation?
On the same task dataset?
Which kind of unsupervised concepts?
Yes
No
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)



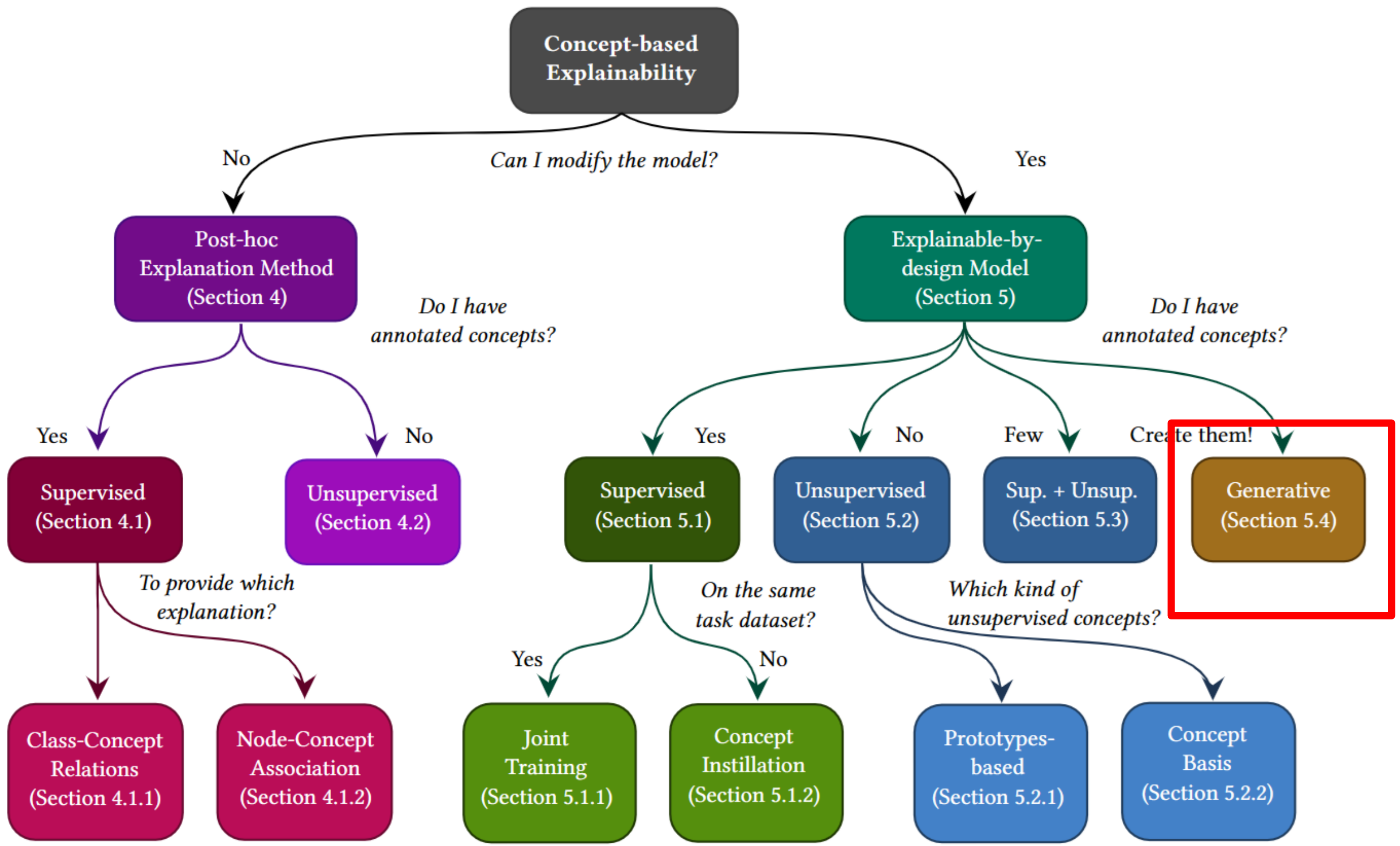
Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." Neurips 2018.

Hybrid Concept based Models

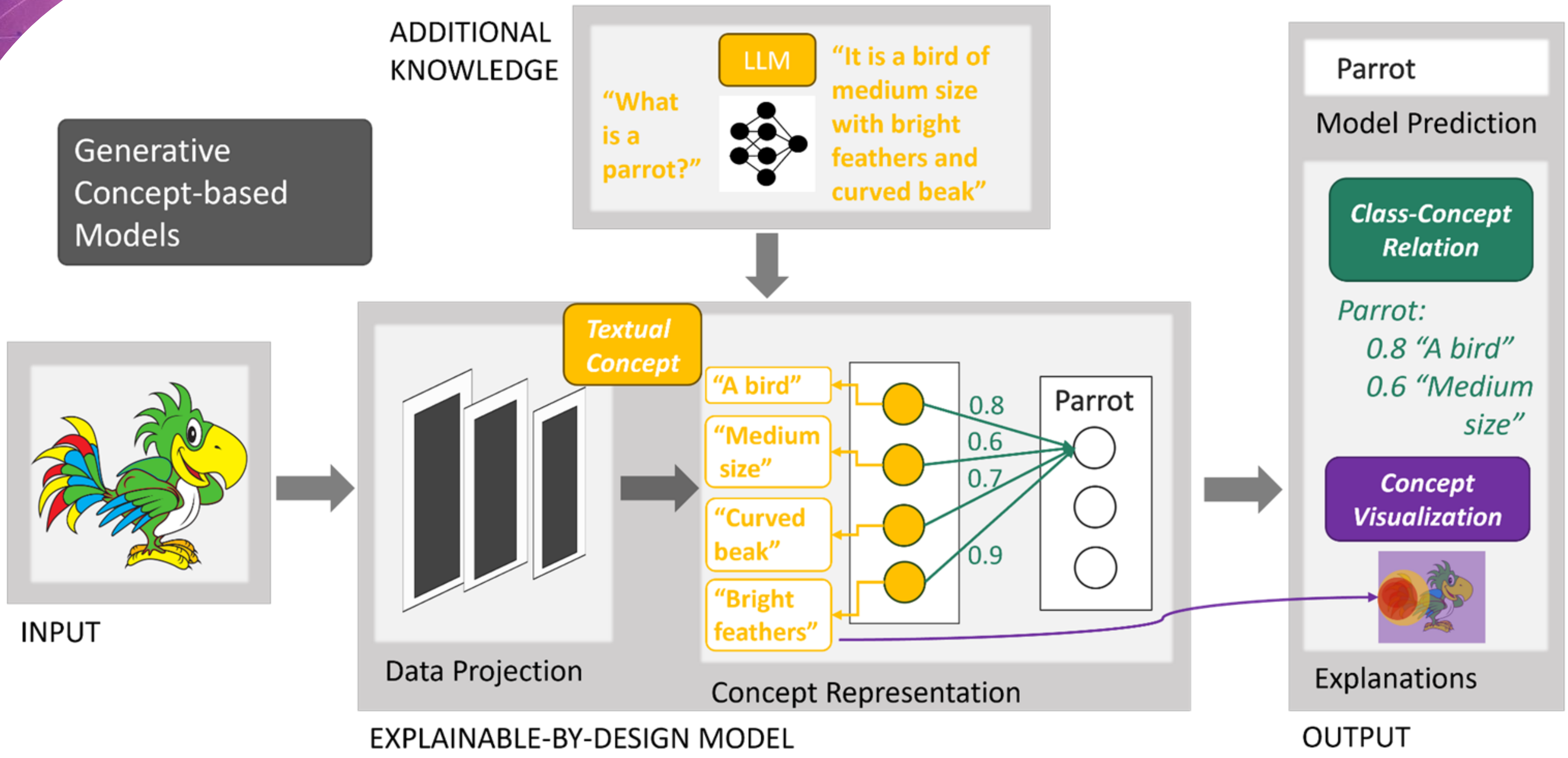
- They train the network to both:
  - Learn to predict a given set of concept with a subset of neurons
  - Create a clusterized representation in the remaining neurons

- Pros:
  - Overcome the accuracy trade-off of fully supervised models
  - Decrease annotation cost
  - Avoid «concept leakage»

- Cons:
  - Most of the information required to classify the classes is encoded in the unsupervised neurons
  - Concept interventions are less effective


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Yes
No
Few
Create them!
Supervised (Section 5.1)
Unsupervised (Section 5.2)
Sup. + Unsup. (Section 5.3)
Generative (Section 5.4)
To provide which explanation?
On the same task dataset?
Which kind of unsupervised concepts?
Class-Concept Relations (Section 4.1.1)
Node-Concept Association (Section 4.1.2)
Yes
No
Joint Training (Section 5.1.1)
Concept Instillation (Section 5.1.2)
Prototypes-based (Section 5.2.1)
Concept Basis (Section 5.2.2)

Generative Concept-based Models



Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." IEEE CVPR 2023.

Generative concept-based models

- They employ a generative model to create the concept labels
  - For each class they ask a description to an LLM
  - They decompose this description into small pieces
- Concept-based model over textual concepts
  - The corresponding embeddings are aligned to the latent input representation to produce concept scores
  - The scores are used to provide the final classification (possibly interpretable)
- Pros:
  - No concept labelling required
- Cons:
  - Per-class labelling
  - Require an external generative model with knowledge of the problem

C-XAI (Part II)

- We will see some real examples

- Post-hoc supervised method:
  - Testing with Concept Activation Vector (T-CAV)

- Explainable-by-design supervised models:
  - Concept Bottleneck Model (CBM)
  - Concept Embedding Model (CEM)