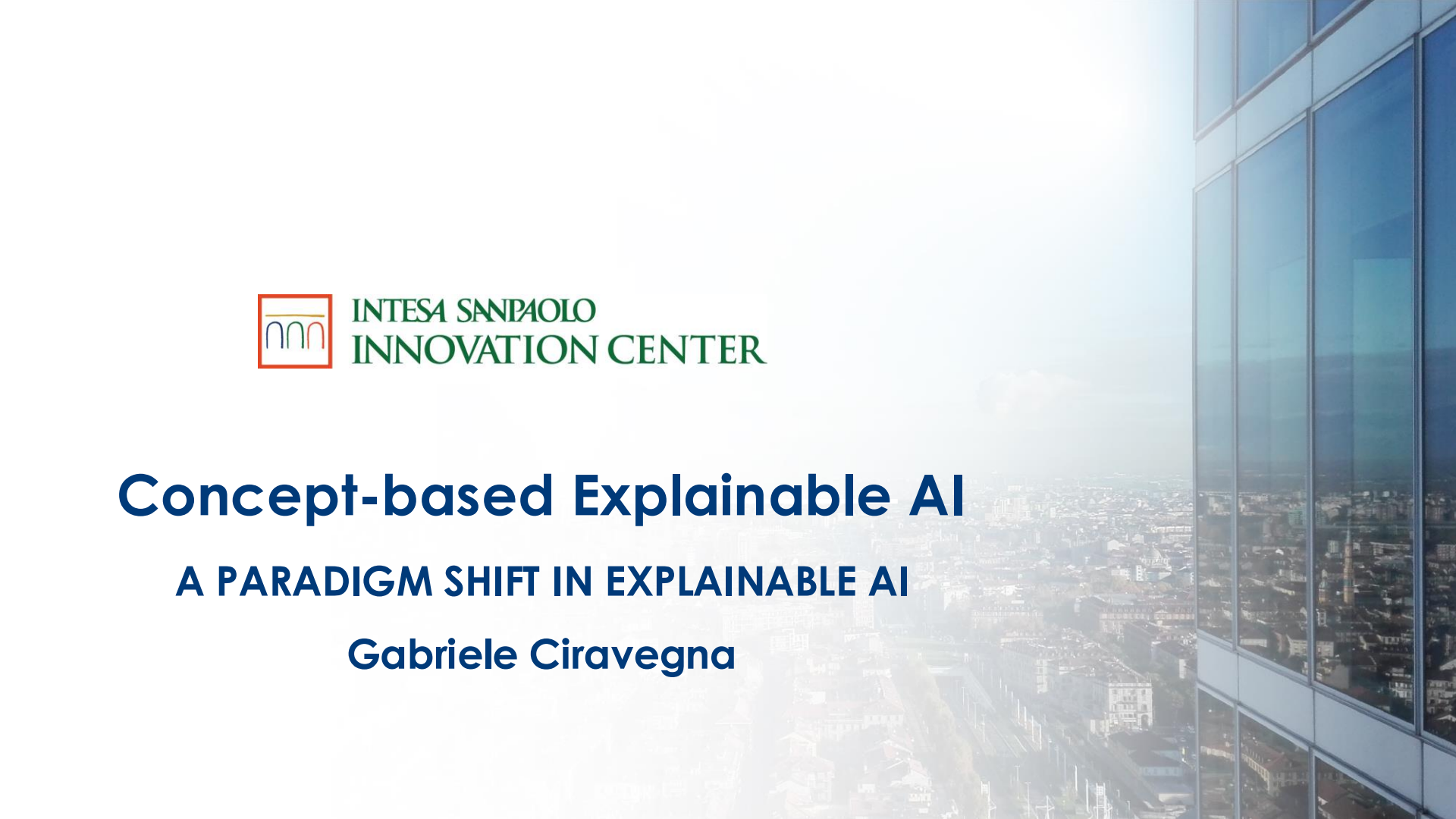
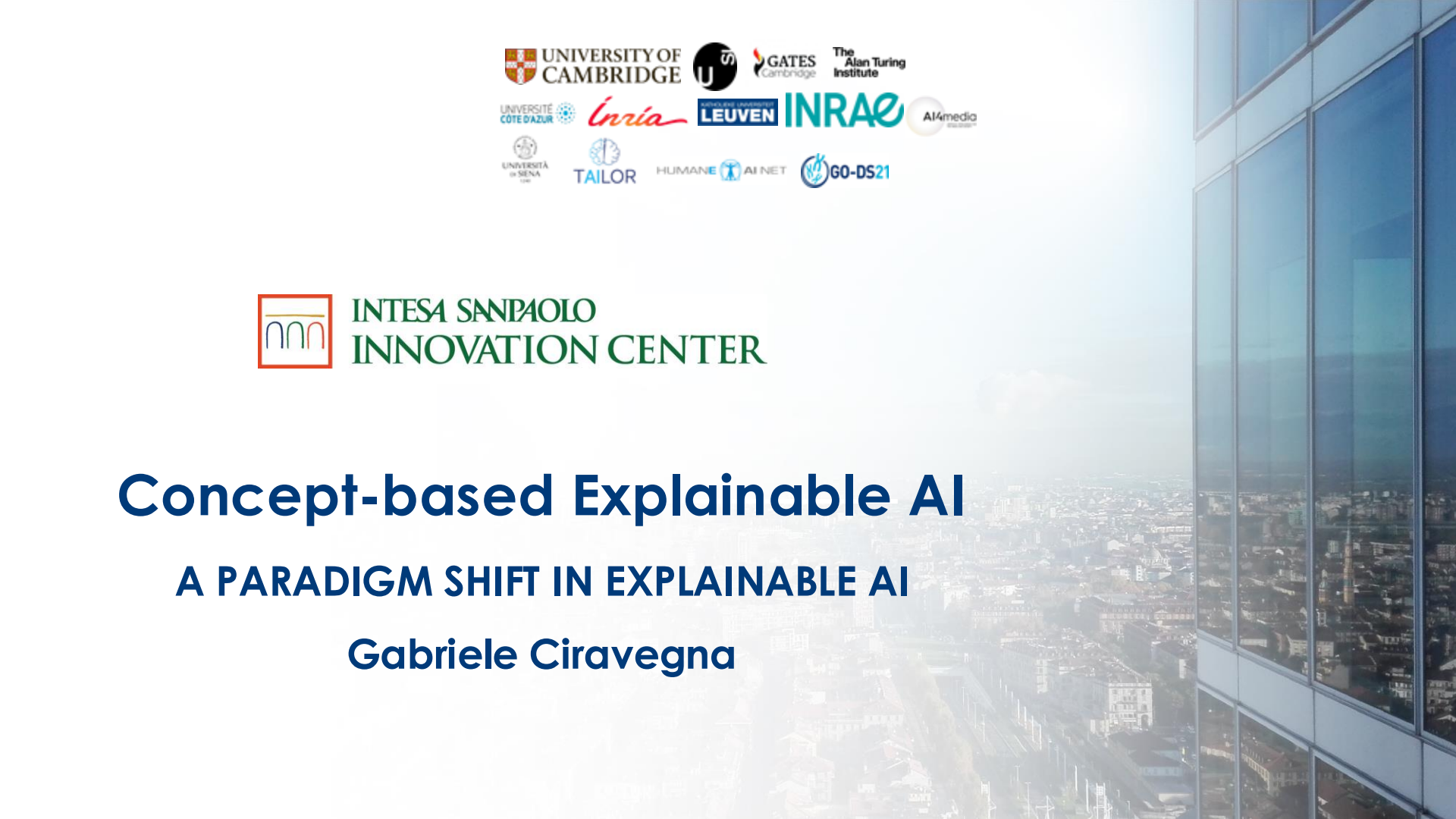


INTESA SANPAOLO
INNOVATION CENTER

Concept-based Explainable AI

A PARADIGM SHIFT IN EXPLAINABLE AI

Gabriele Ciravegna



Concept-based Explainable AI

A PARADIGM SHIFT IN EXPLAINABLE AI

Gabriele Ciravegna

Intesa Sanpaolo Innovation Center is the Intesa Sanpaolo Group company dedicated to the **frontiers of innovation**.

It explores:

- future scenarios and trends
- **develops multidisciplinary applied research projects**
- supports startups
- accelerates the business transformation of companies

According to the criteria of Open Innovation and Circular Economy, it favors the development of innovative ecosystems and spreads the culture of innovation, to make of Intesa Sanpaolo the driving force of a **more aware, inclusive and sustainable economy**.

INTESA SANPAOLO INNOVATION CENTER

Questionnaire

Questionnaire

1. Do you know XAI?

Questionnaire

1. Do you know XAI?
2. **Do you like XAI?**

Questionnaire

1. Do you know XAI?
2. Do you like XAI?
3. **Do you think XAI works well?**

Questionnaire

1. Do you know XAI?
2. Do you like XAI?
3. Do you think XAI works well?
4. **Do you think a normal user can understand why a model has taken a decision by means of an XAI explanation?**

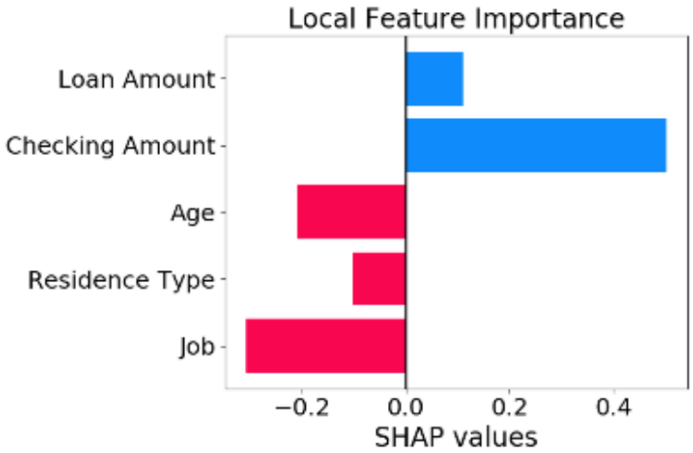
Outline

1. Motivation
2. Concept-based eXplainable AI (C-XAI)
3. Concept Bottleneck Models (CBMs)
4. Concept Embedding Models (CEMs)
5. Deep Concept Reasoner (DCR)

Motivation

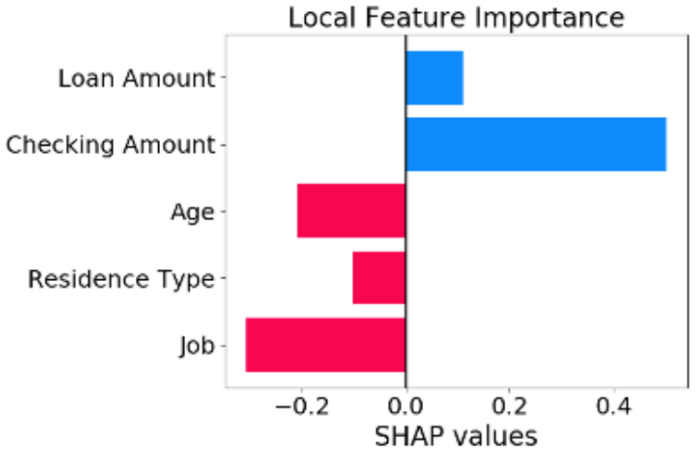
Explanations depends on the input data

Tabular data



Explanations depends on the input data

Tabular data



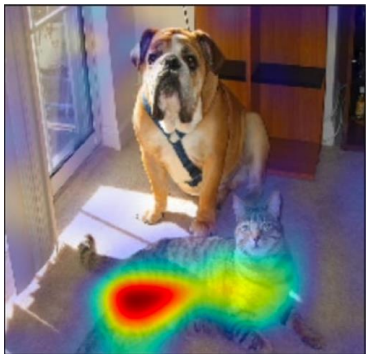
Structured data

- Each feature has a clear **semantic meaning**
- From feature importance we can easily define a **counterfactual example**
- **Global explanations** are meaningful as features keep the same meaning across different samples

Explanations depends on the input data

Textual data

Image data



INTESA SANPAOLO INNOVATION CENTER

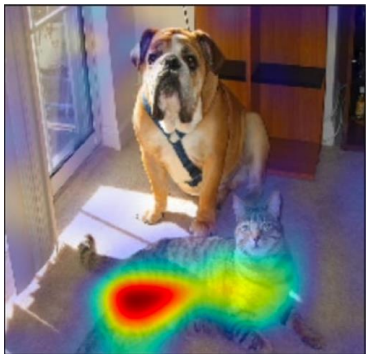
Explanations depends on the input data

Unstructured data

- features **lack semantic meaning**
- counterfactuals **cannot be defined** without an external model
- global explanations **cannot be extracted**

Textual data

Image data

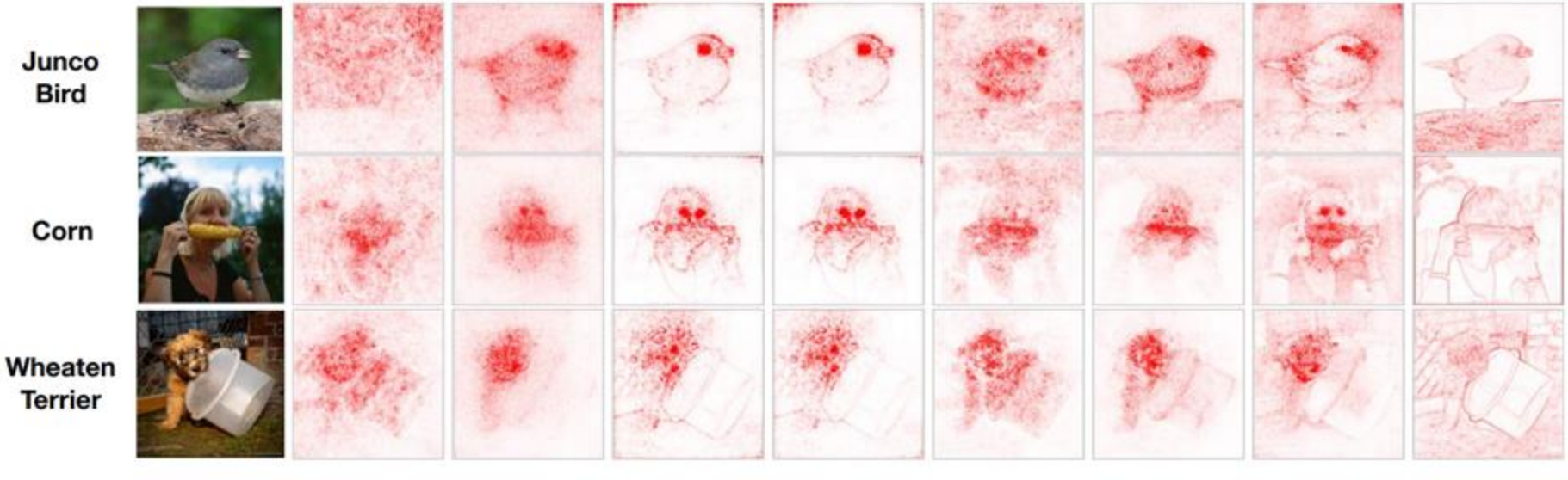


INTESA SANPAOLO INNOVATION CENTER

When considering unstructured data

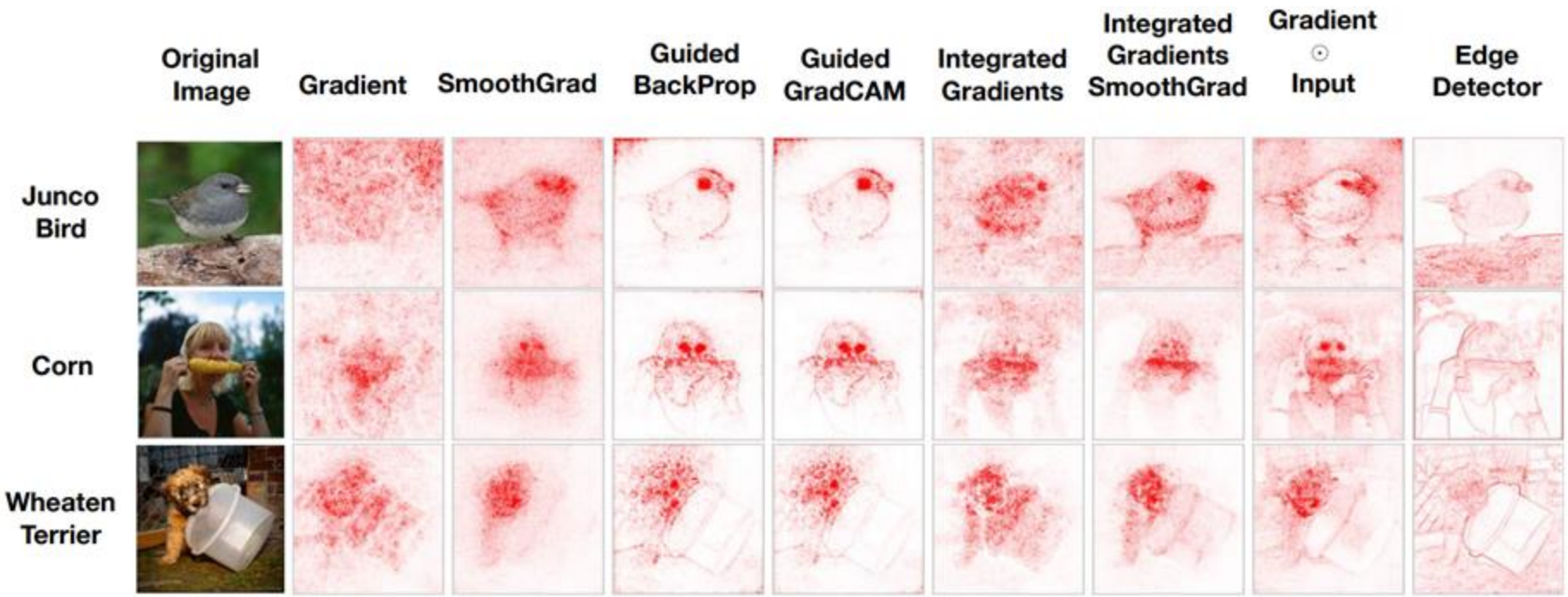
Standard XAI does not always work well

Which method is better? [1]



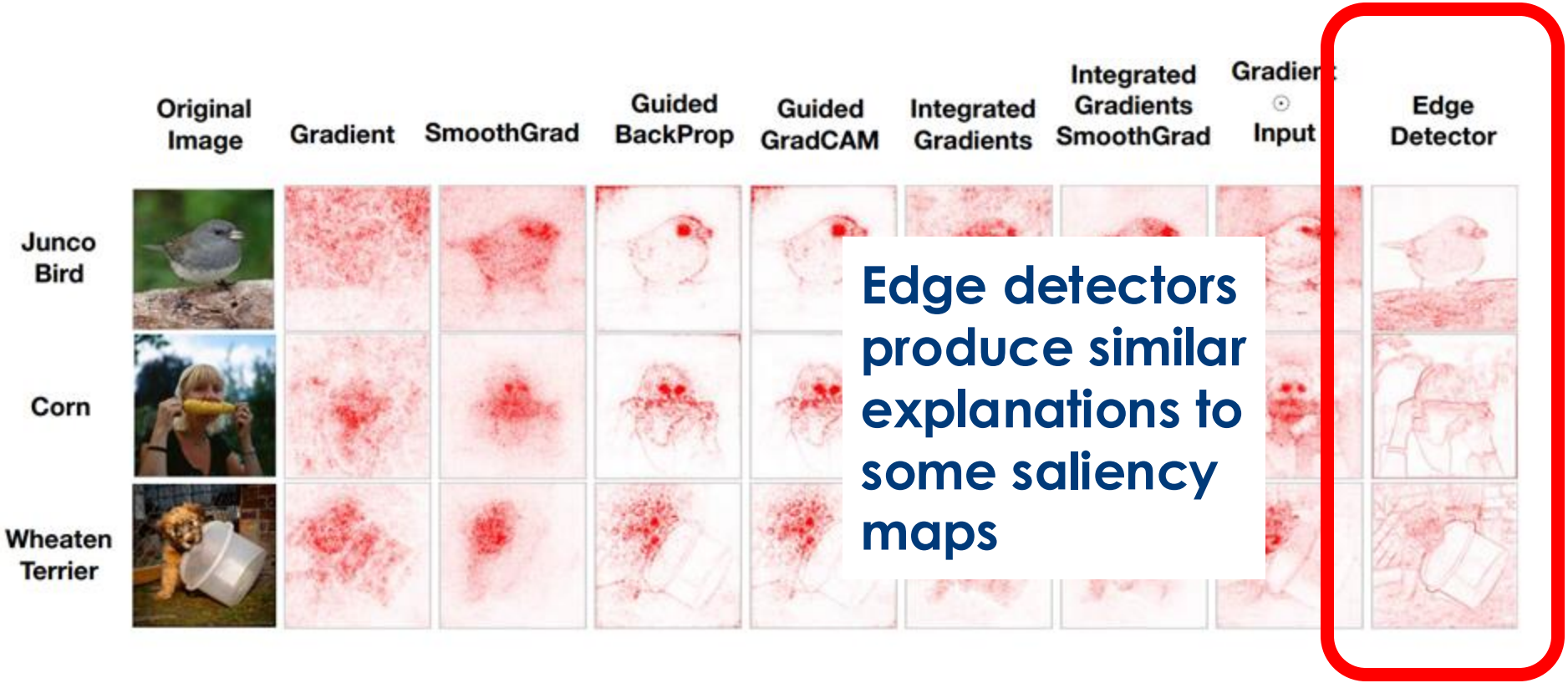
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Which method is better? [1]



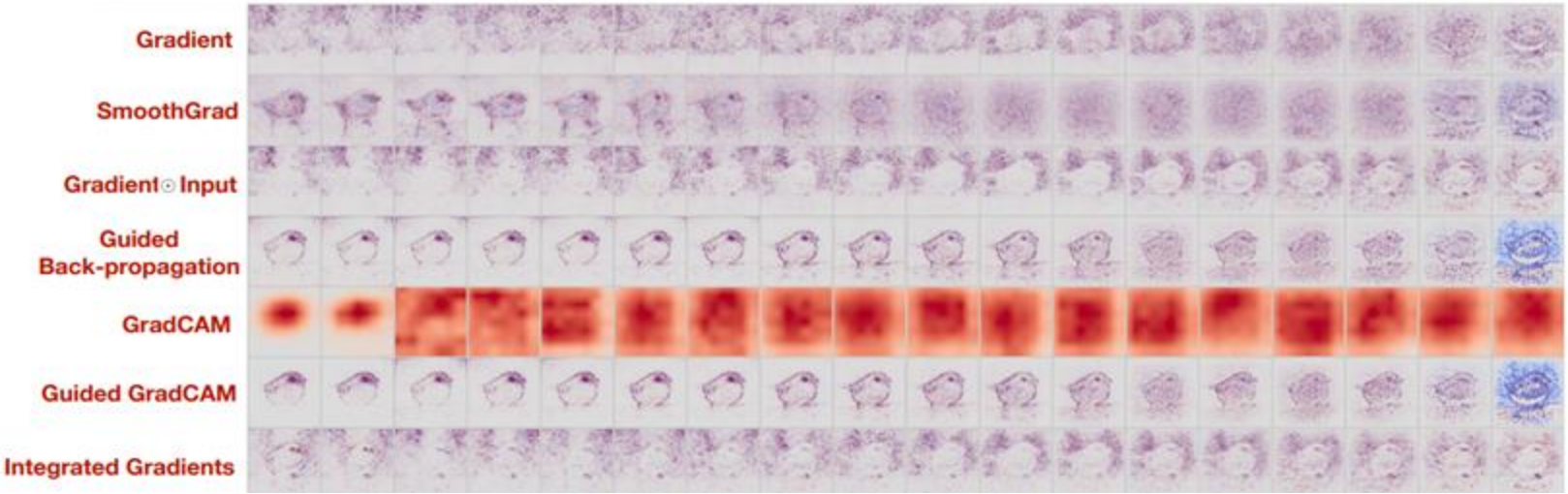
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Which method is better? [1]



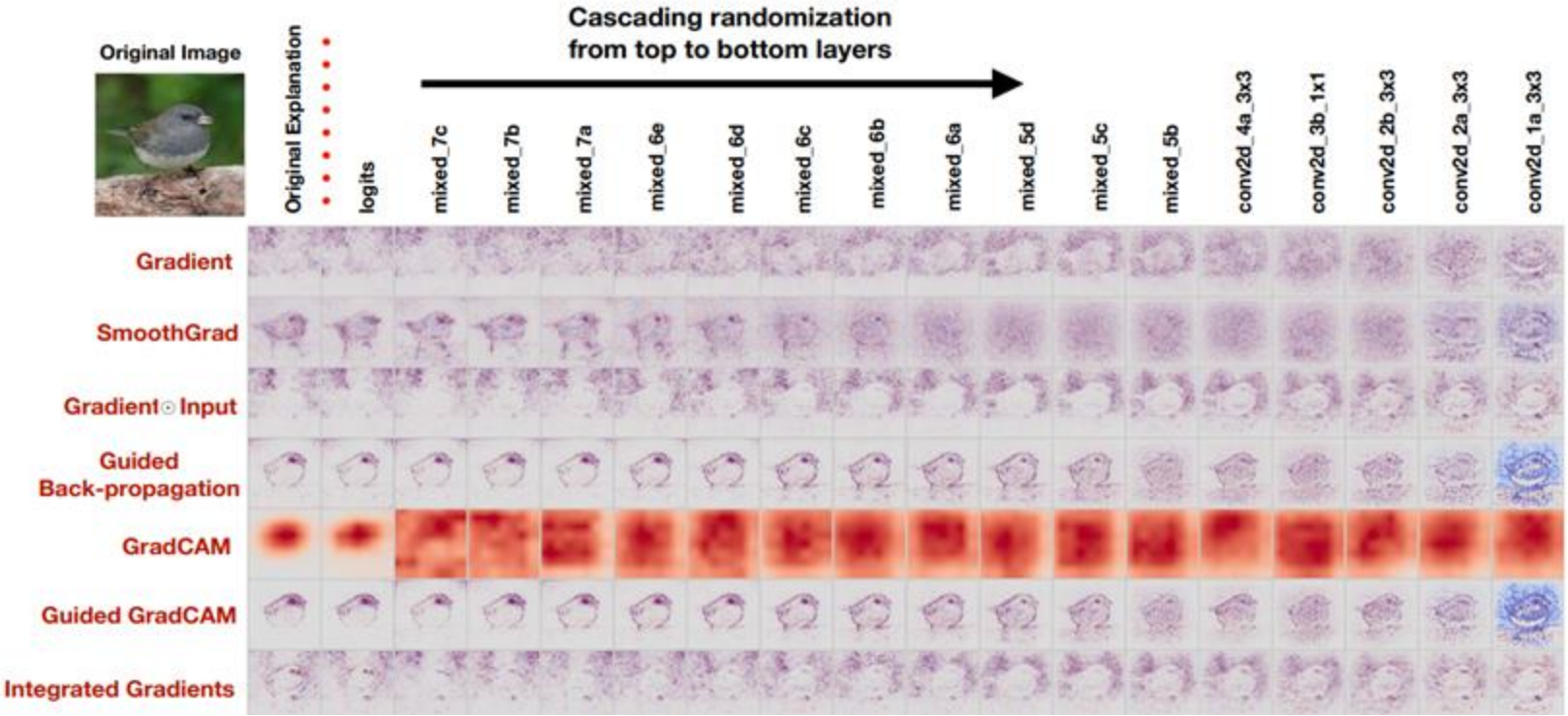
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Towards where are we randomizing? [1]



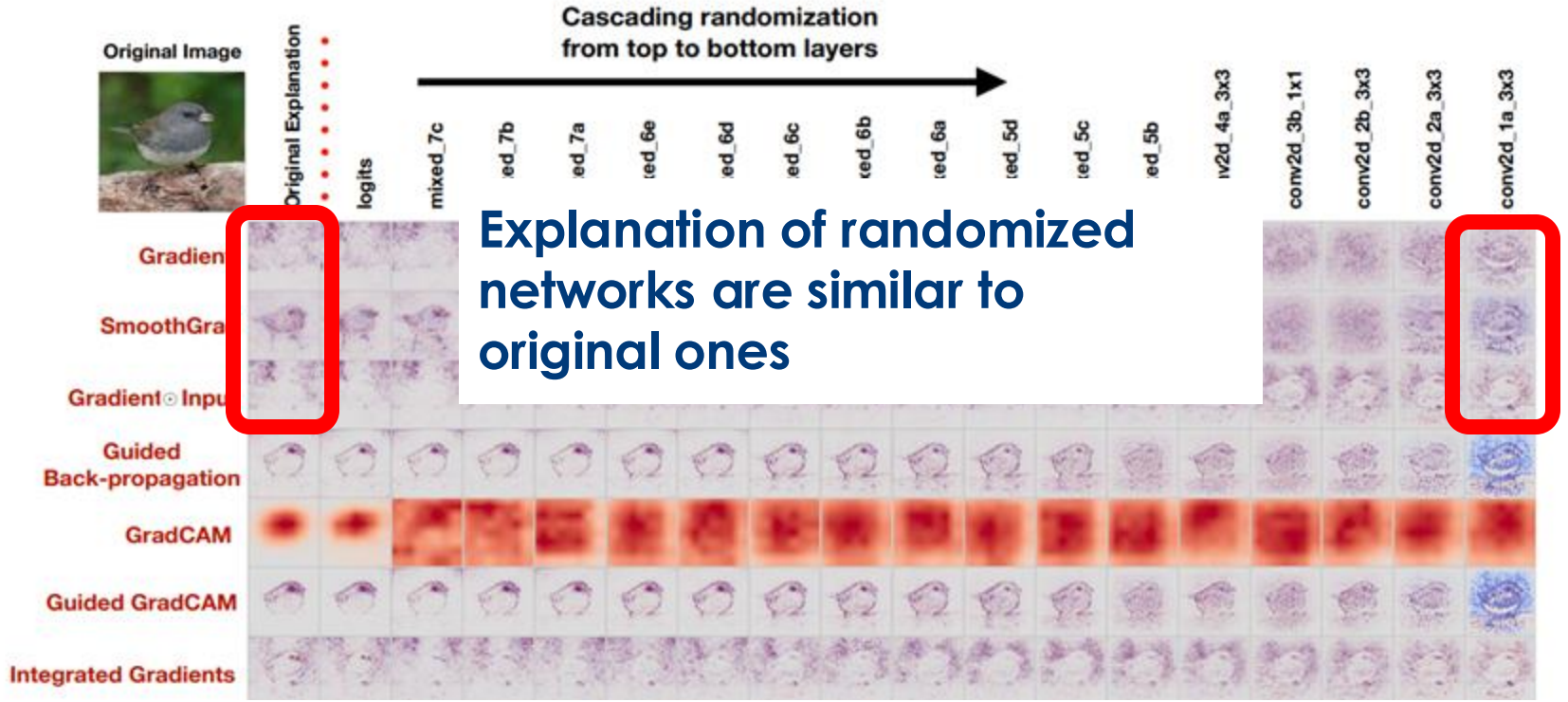
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Towards where are we randomizing? [1]



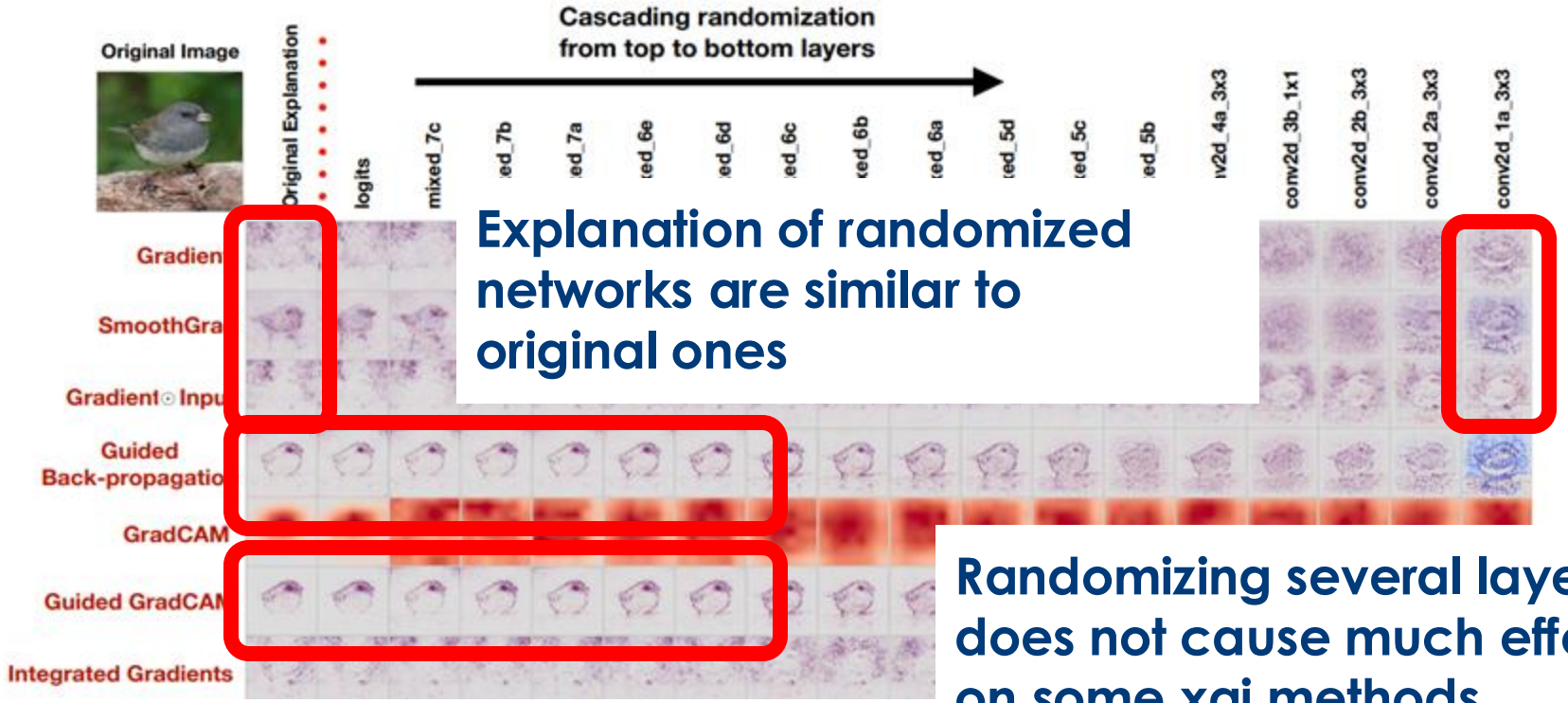
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Towards where are we randomizing? [1]



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

Towards where are we randomizing? [1]



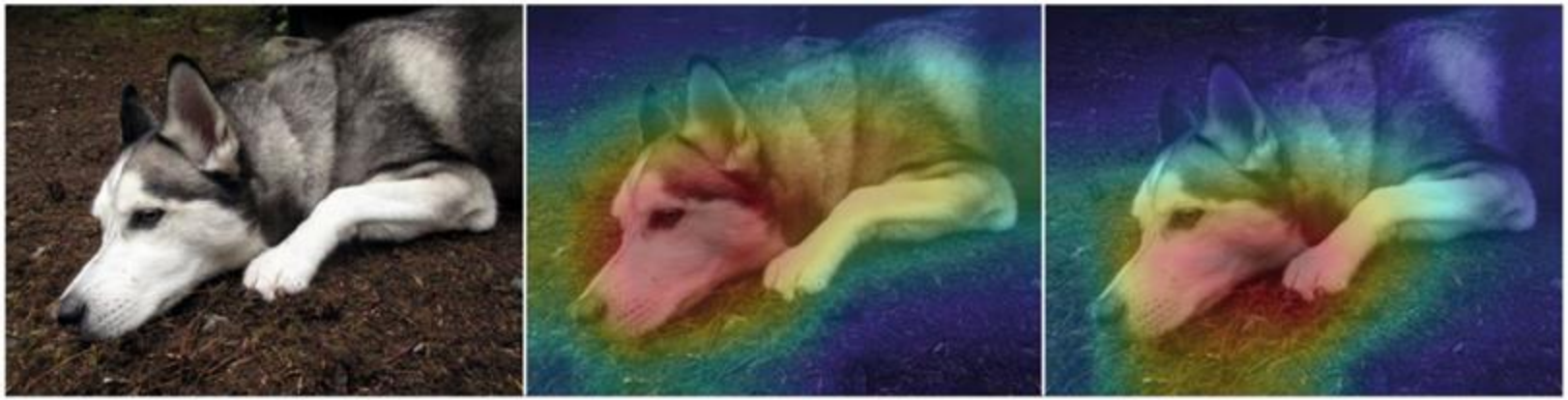
Explanation of randomized networks are similar to original ones

Randomizing several layers does not cause much effect on some xai methods

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

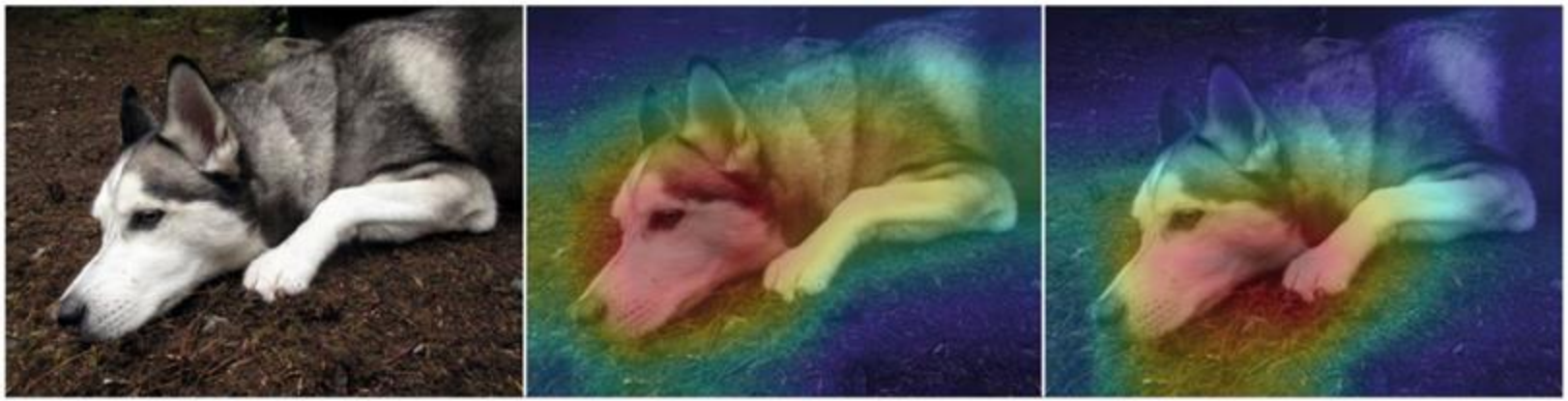
Standard xai methods are more input-dependent than model dependent

Which class are we explaining? [3]



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

Which class are we explaining? [3]



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

Why XAI explanations are difficult to understand?

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)

Standard users may not understand standard XAI explanations

Concept-based XAI (C-XAI)

Poeta, Eleonora, et al. "Concept-based Explainable Artificial Intelligence: A Survey", ACM Computing Surveys (2025)

WHAT IS THIS?

APPLE

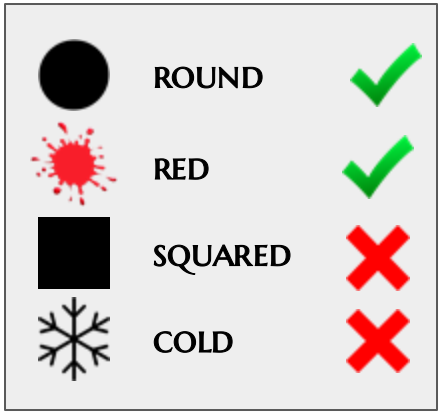
PREDICTION

WHAT IS THIS?

APPLE

PREDICTION

WHY?

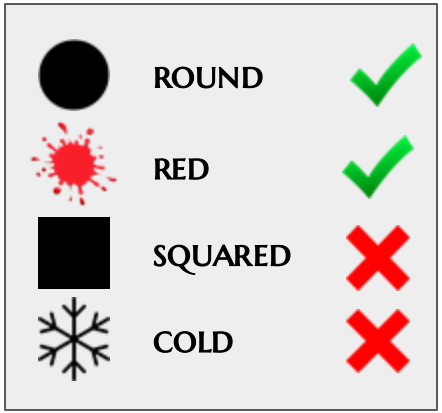


WHAT IS THIS?

APPLE

PREDICTION

WHY?



CONCEPT-BASED EXPLANATION

Some Definitions

What is a Concept?

What is a Concept?

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

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

What is a Concept?

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

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

Let’s try to be more concrete…

Different types of concepts

Symbolic Concepts

Human-defined attributes

"BEAK"

Different types of concepts

Symbolic Concepts

Human-defined attributes

Unsupervised Concept Basis

Cluster of similar samples

"BEAK"

Different types of concepts

Symbolic Concepts

Human-defined attributes

Unsupervised Concept Basis

Cluster of similar samples

Prototypes

(Part-of) a training sample

"Beak"

Different 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 of a main class

"Beak"

"A bird with bright feathers"

Different types of concepts

Symbolic Concepts

Human-defined attributes

Requires further annotations

"Beak"

Unsupervised Concept Basis

Cluster of similar samples

Can be extracted from the data

Prototypes

(Part-of) a training sample

Textual Concepts

Textual representation of a main class

Requires an LLM (or further annotations)

"A bird with bright feathers"

What is a Concept-based Explanation?

Different Types of Explanations too!

1. **Class-Concept Relations**

   Relation among a concept and an output class of a model

$$Beak \rightarrow Parrot$$

Different Types of Explanations too!

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

Different Types of Explanations too!

1. **Class-Concept Relations**

   Relation among a concept and an output class of a model

$$Beak \rightarrow Parrot$$

2. **Node-Concept Association**

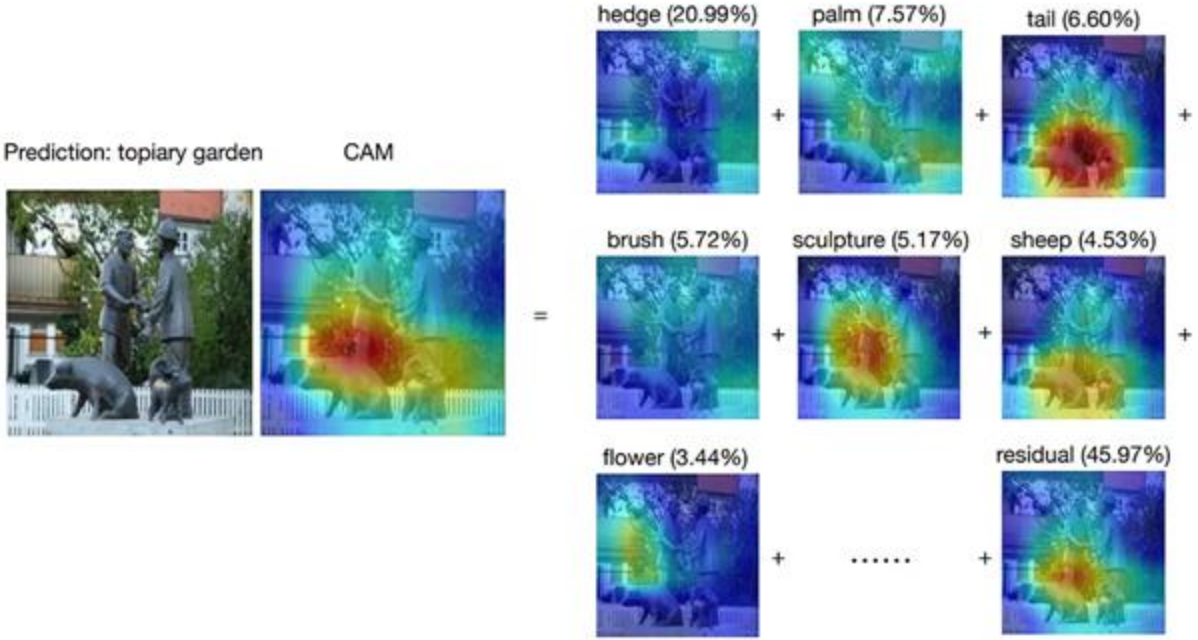
   Explicit association of a concept with a hidden node of the network

3. **Concept-Visualization**

   Visualization of a learnt concept in terms of the input features

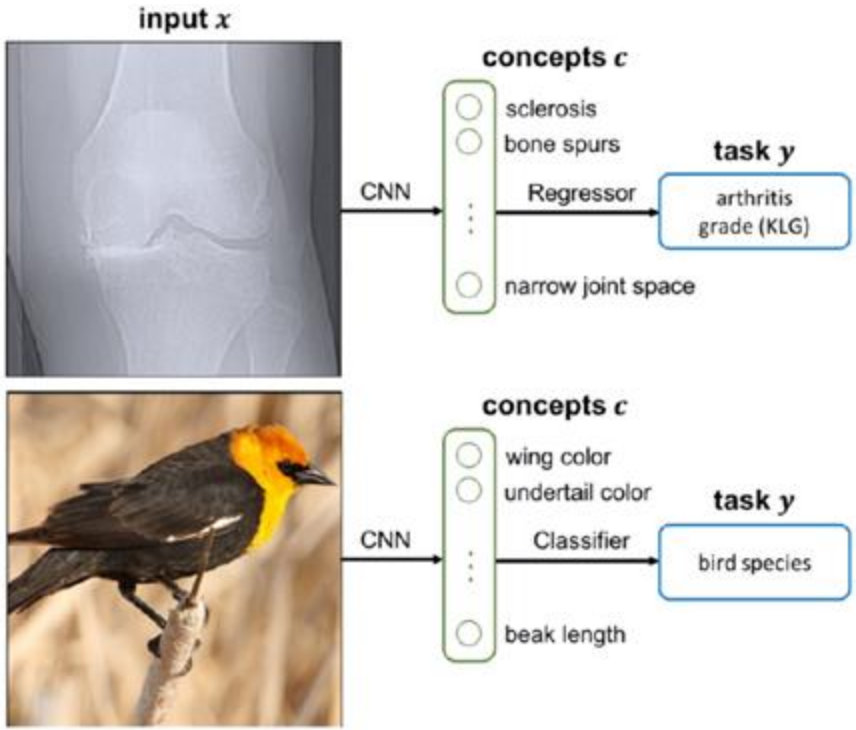
Post-hoc or Explainable-by-design?

Post-hoc Concept-Based Explanations

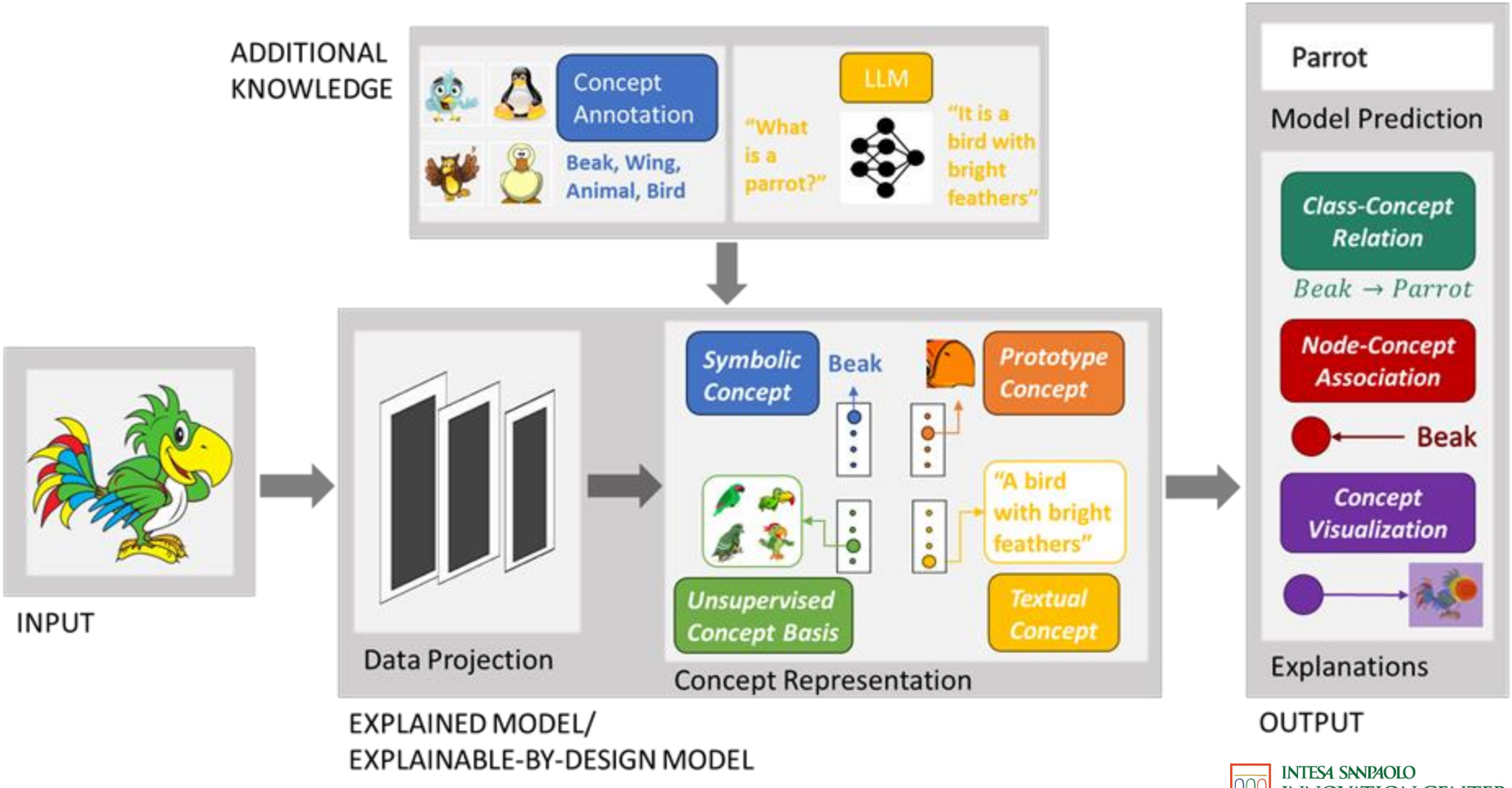


Zhou, Bolei, et al. "Interpretable basis decomposition for visual explanation." *ECCV 2018*.

Explainable-by-design Concept-based Models

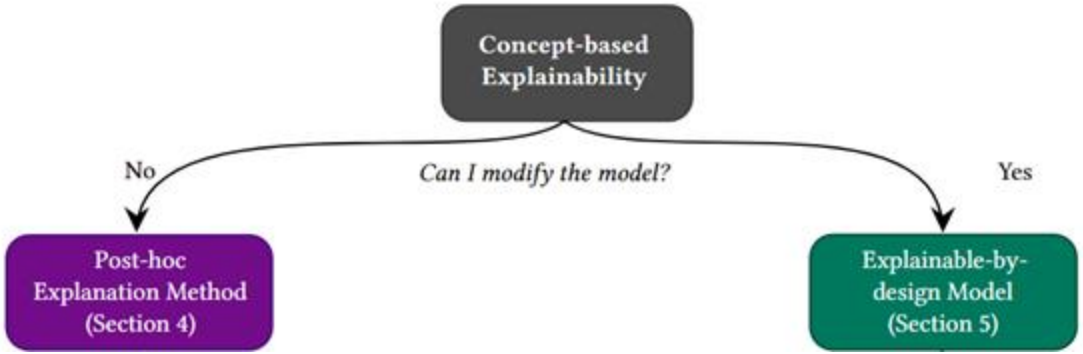


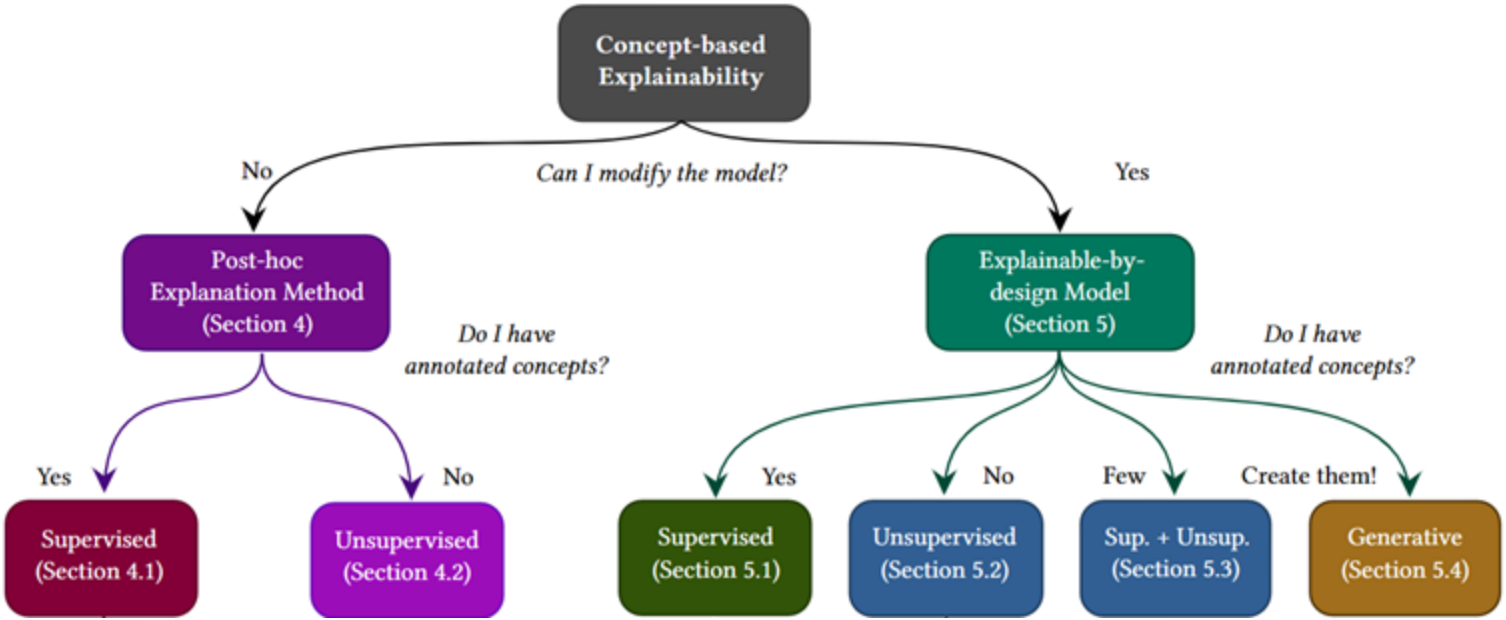
Koh, Pang W, et al. "Concept bottleneck models." *ICML 2020*.

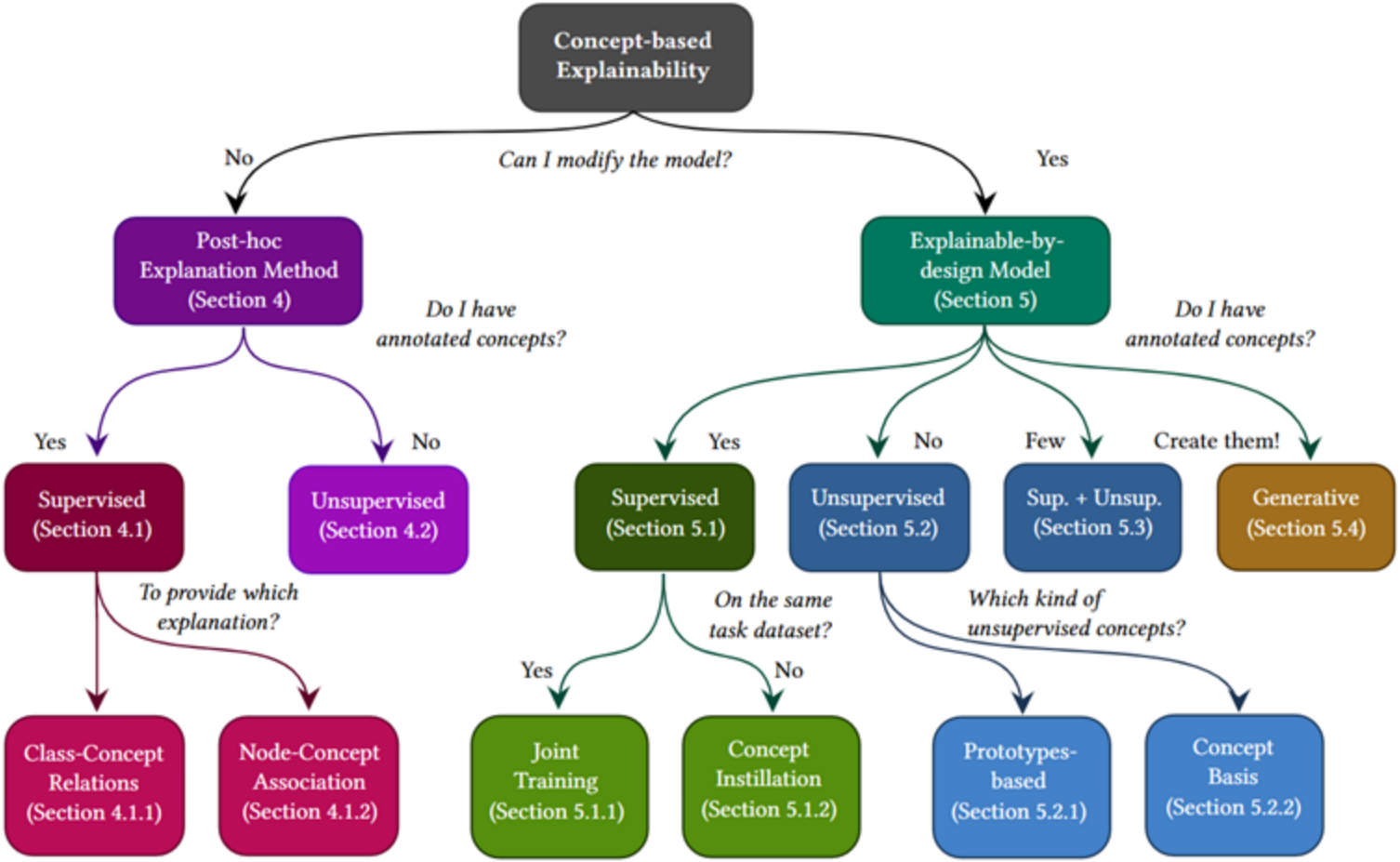


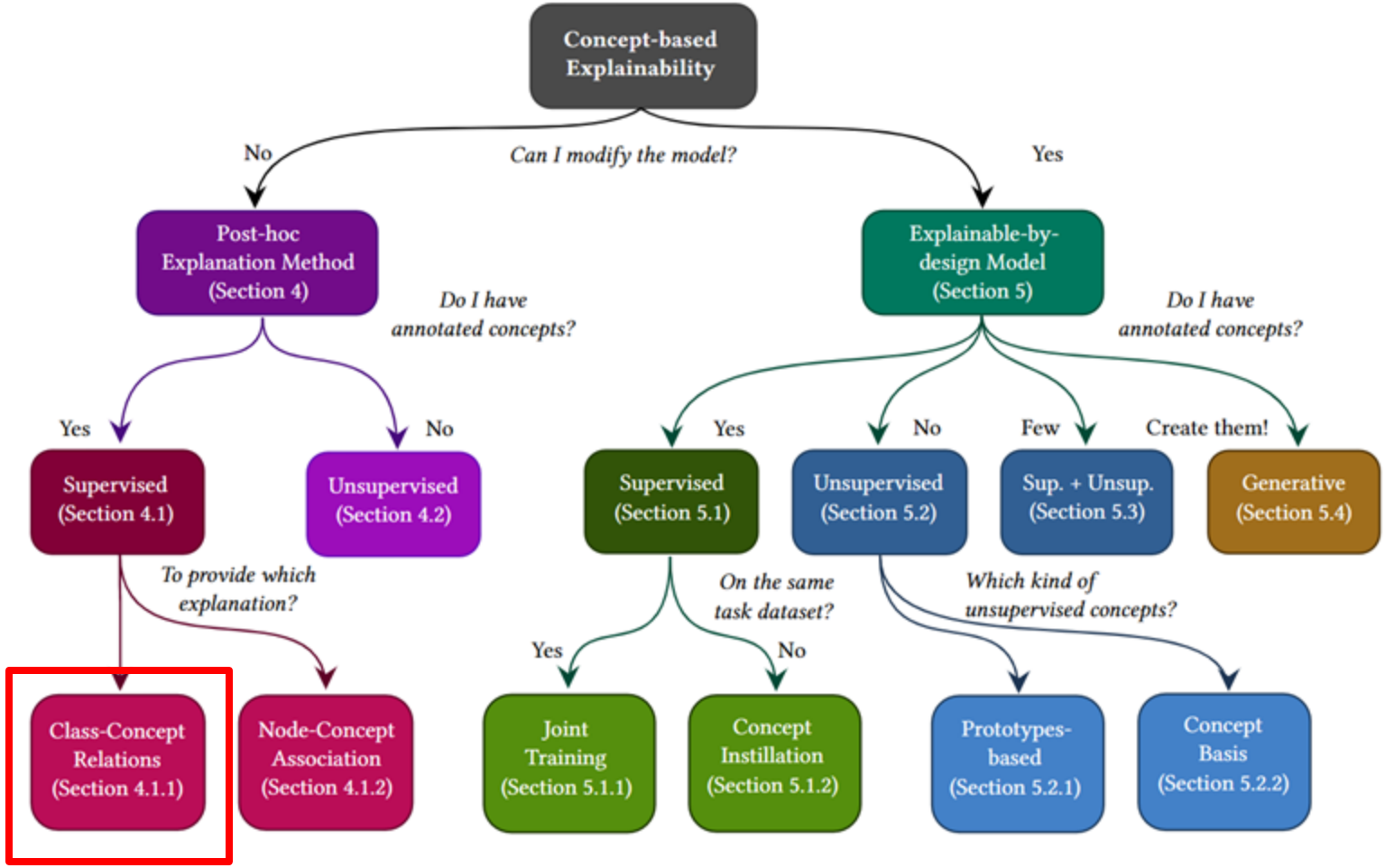
INTESA SANPAOLO INNOVATION CENTER

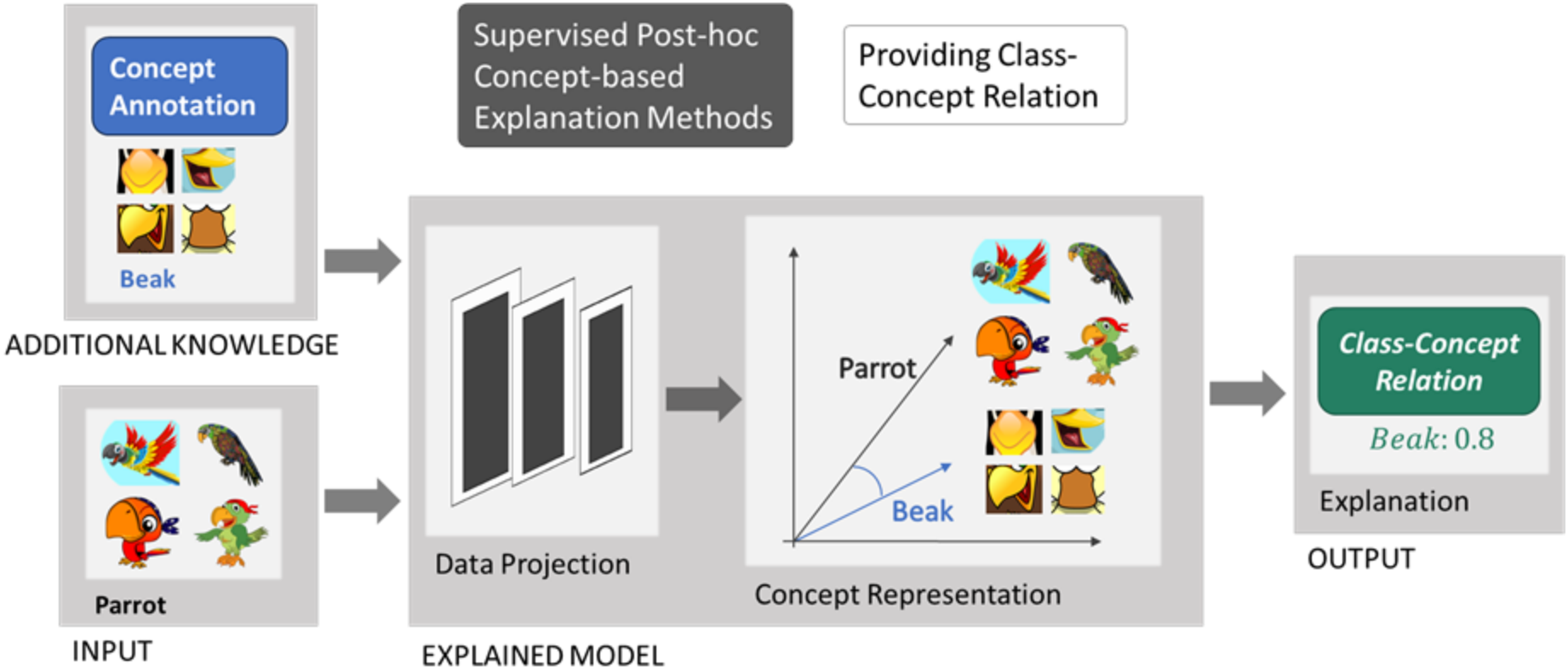
How do i choose a C-XAI method?


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)

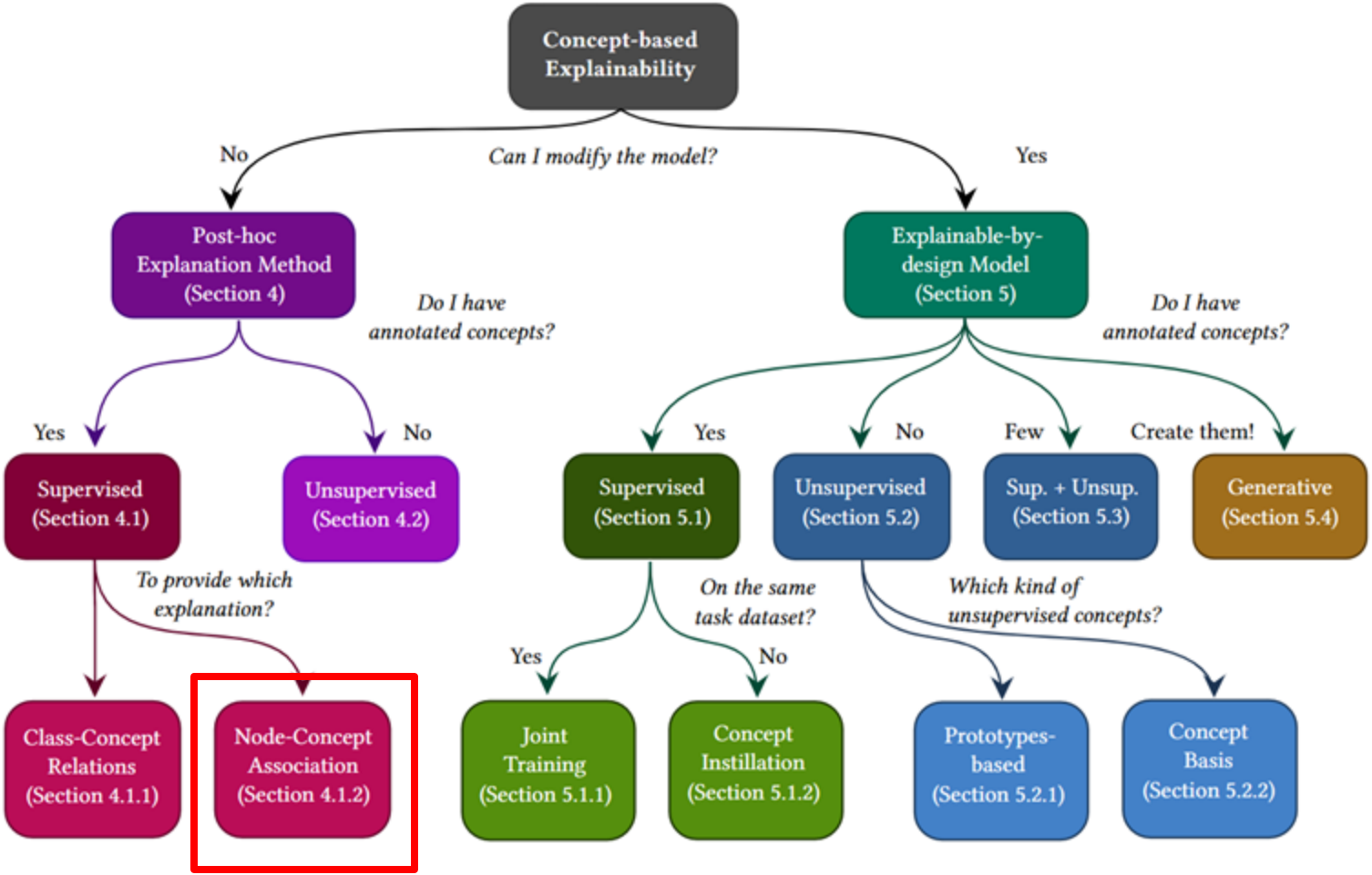


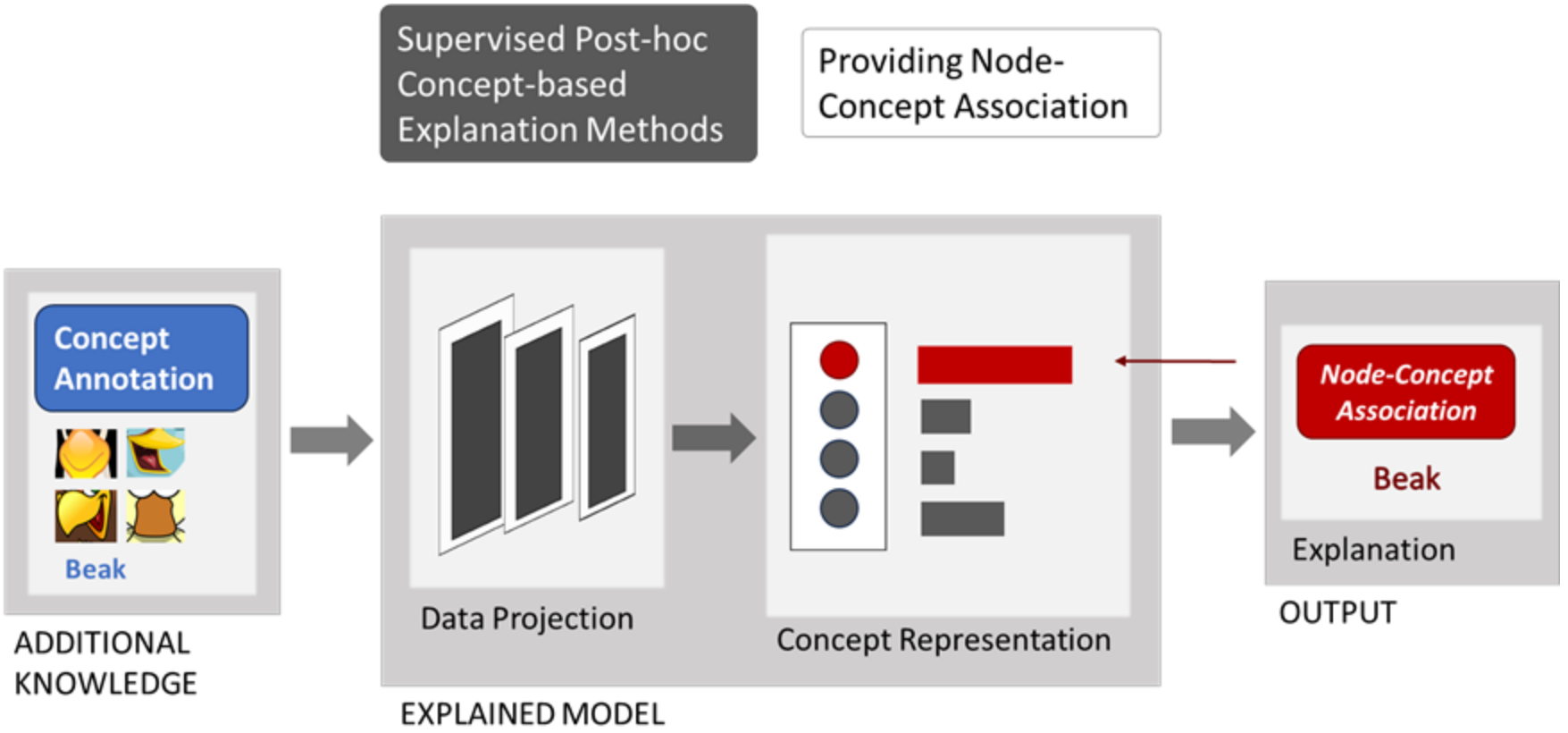

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Supervised (Section 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)


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)

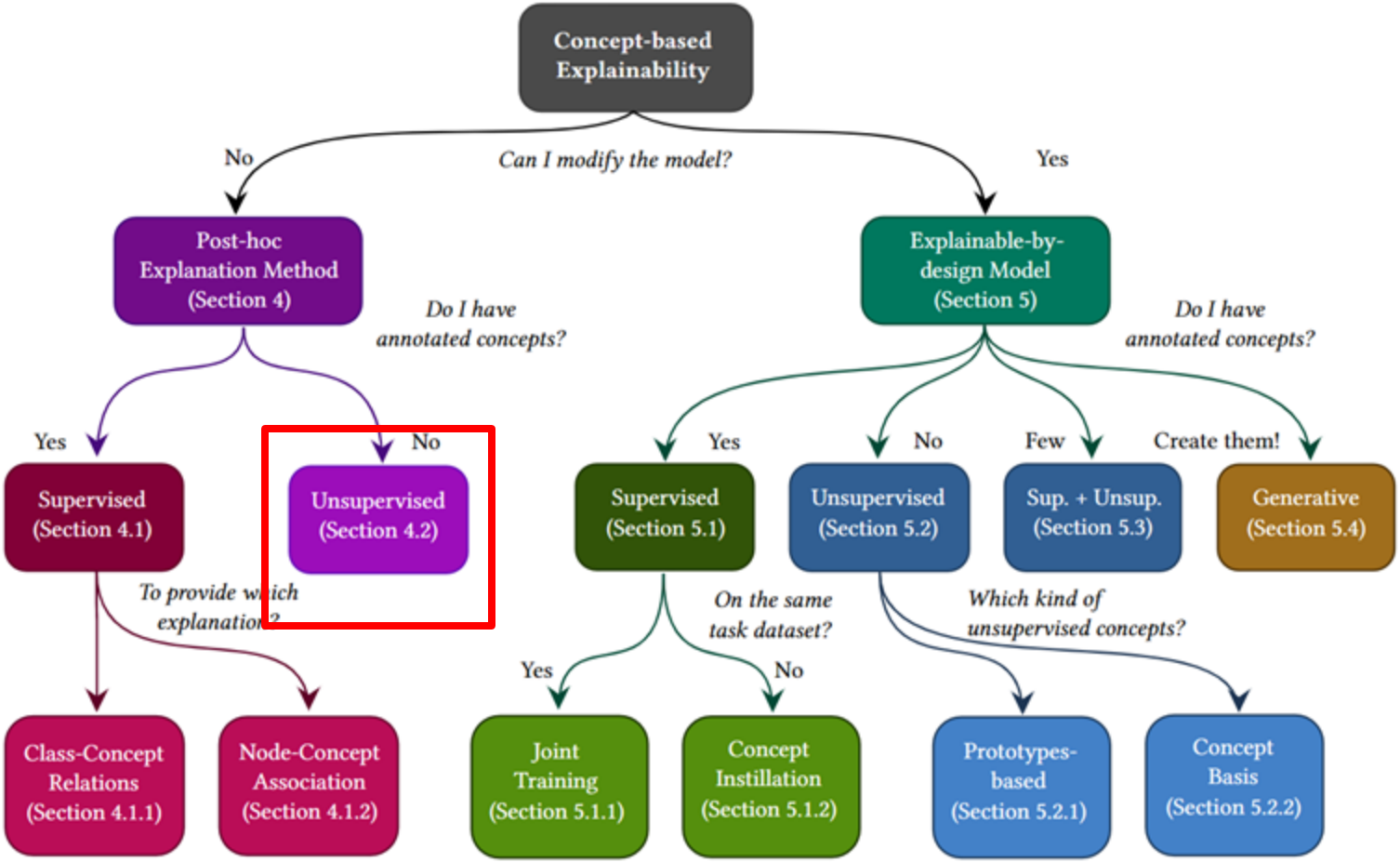


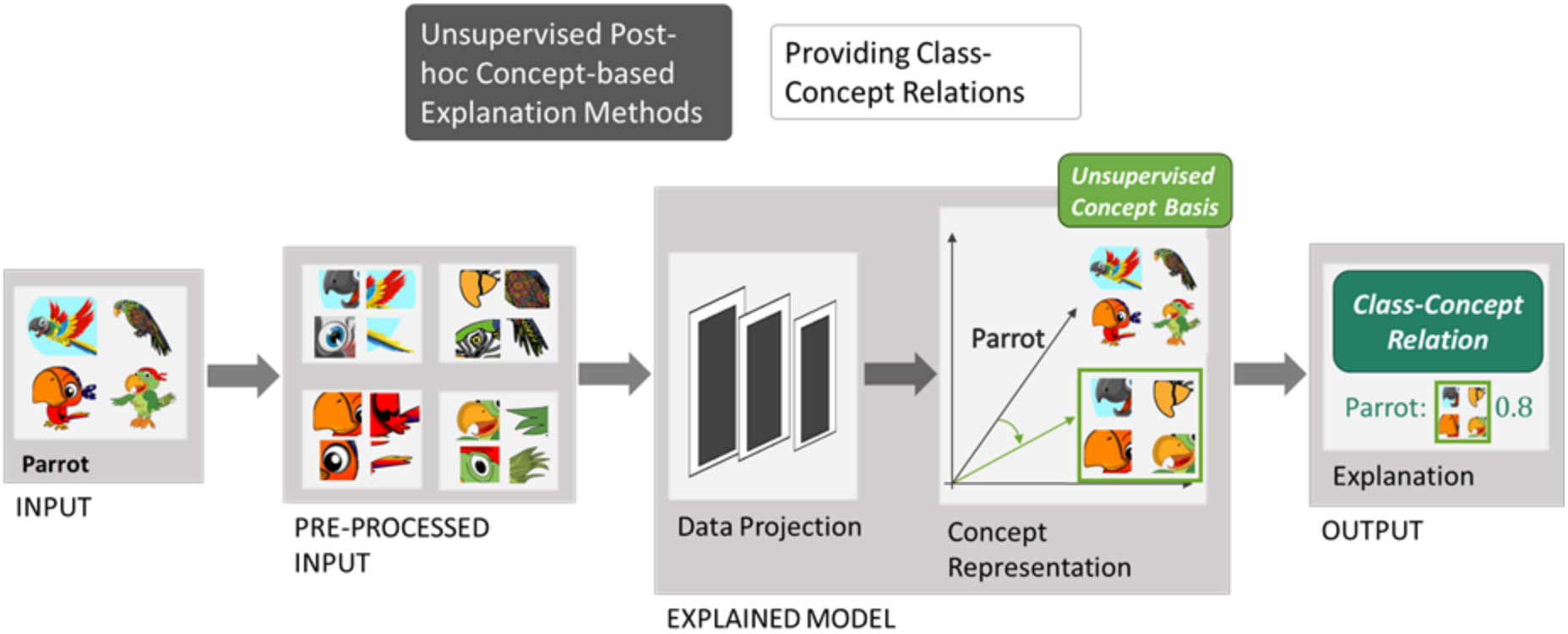
E.g., Kim, Been, et al. "*Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).*" International conference on machine learning. (2018).


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)

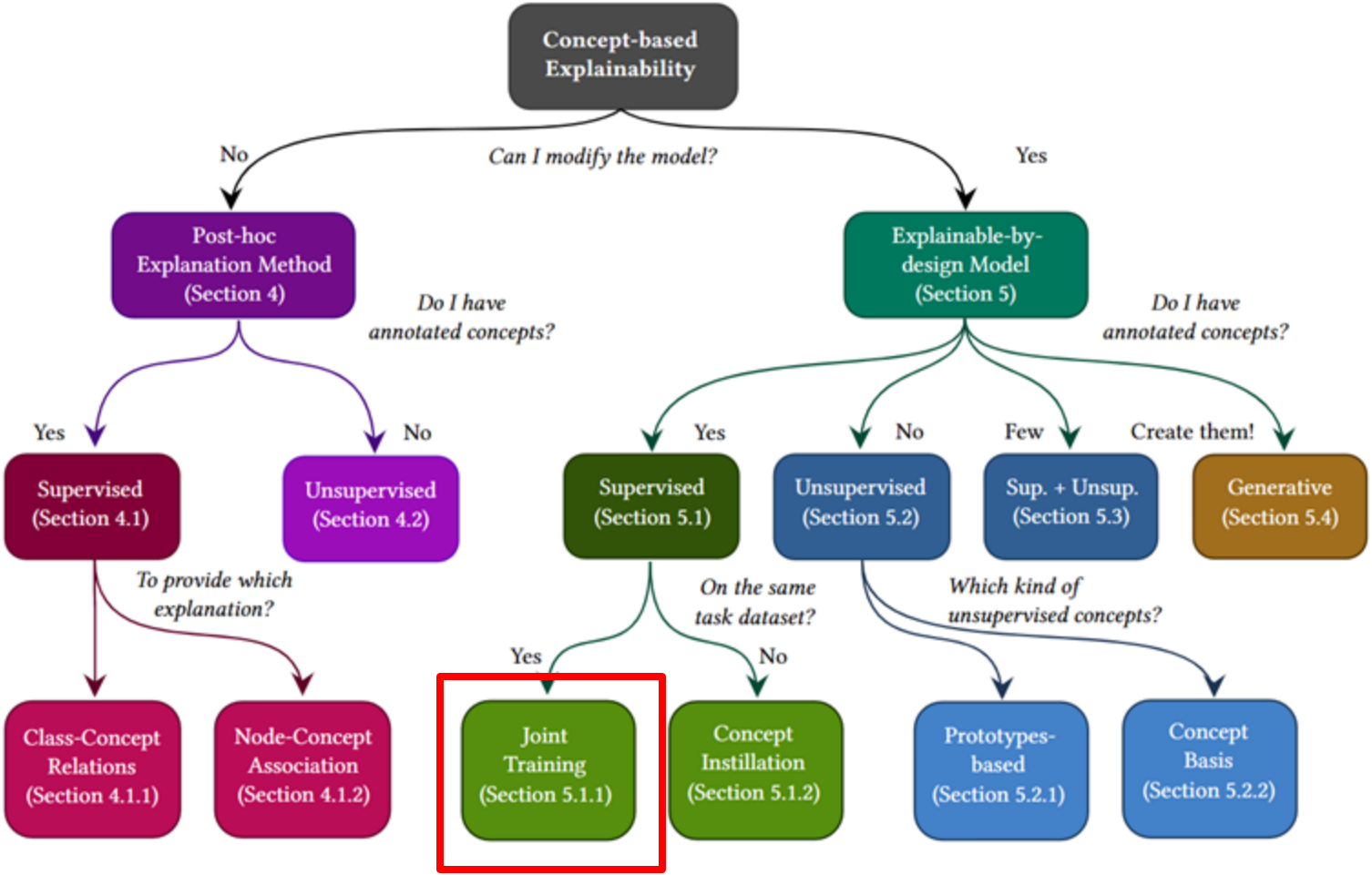


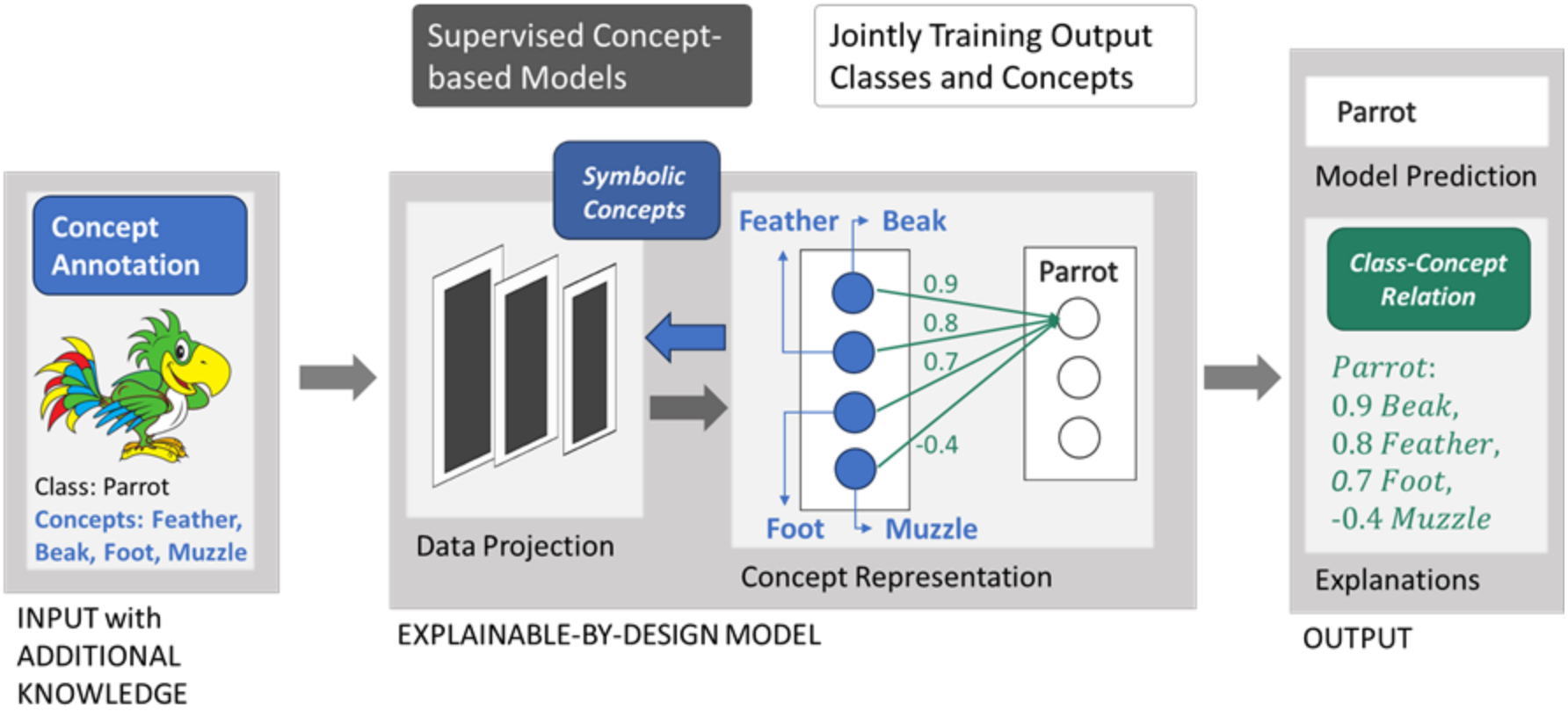
E.g., Bau, David, et al. "*Network dissection: Quantifying interpretability of deep visual representations.*" CVPR. 2017


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (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)

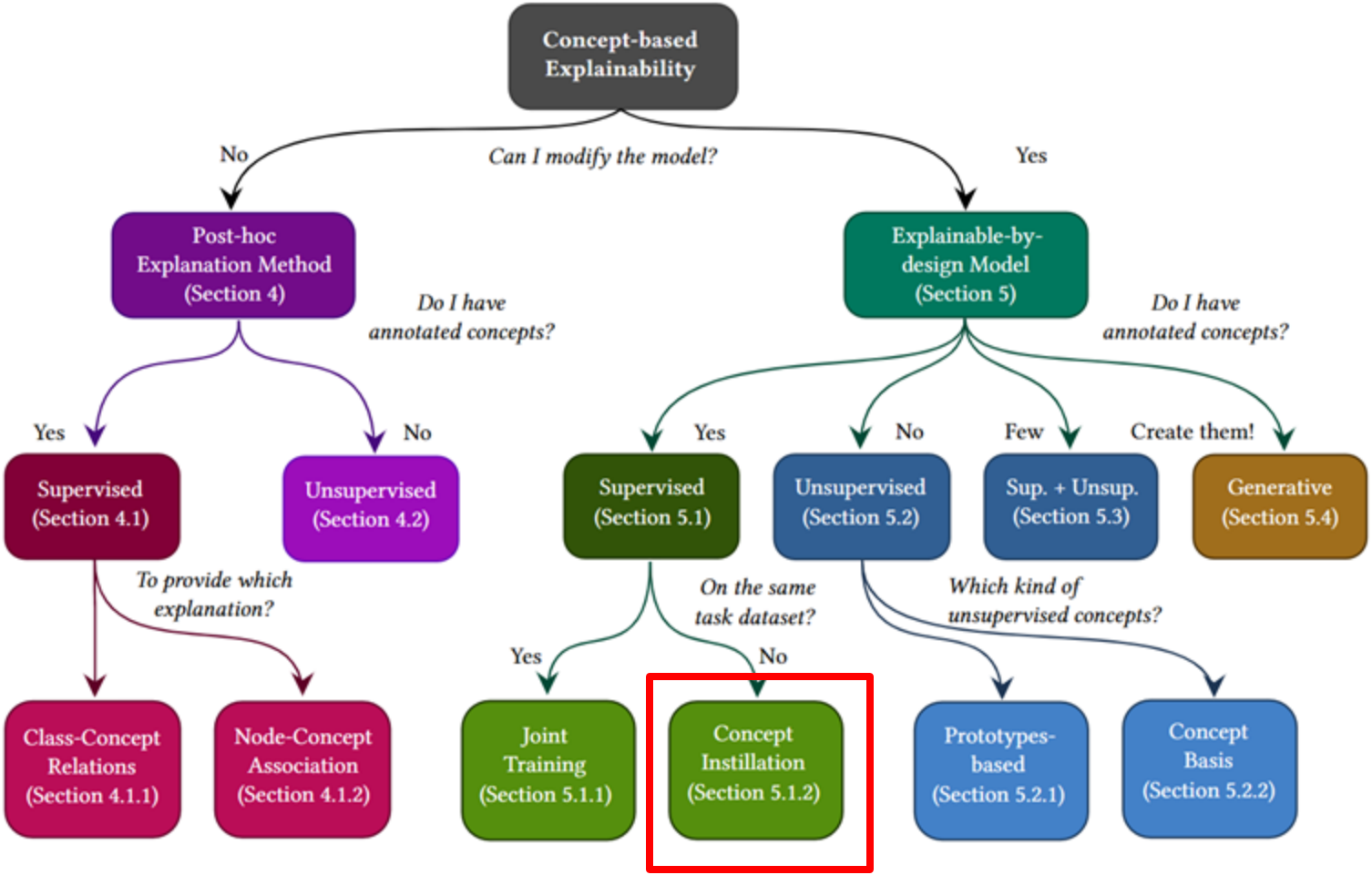


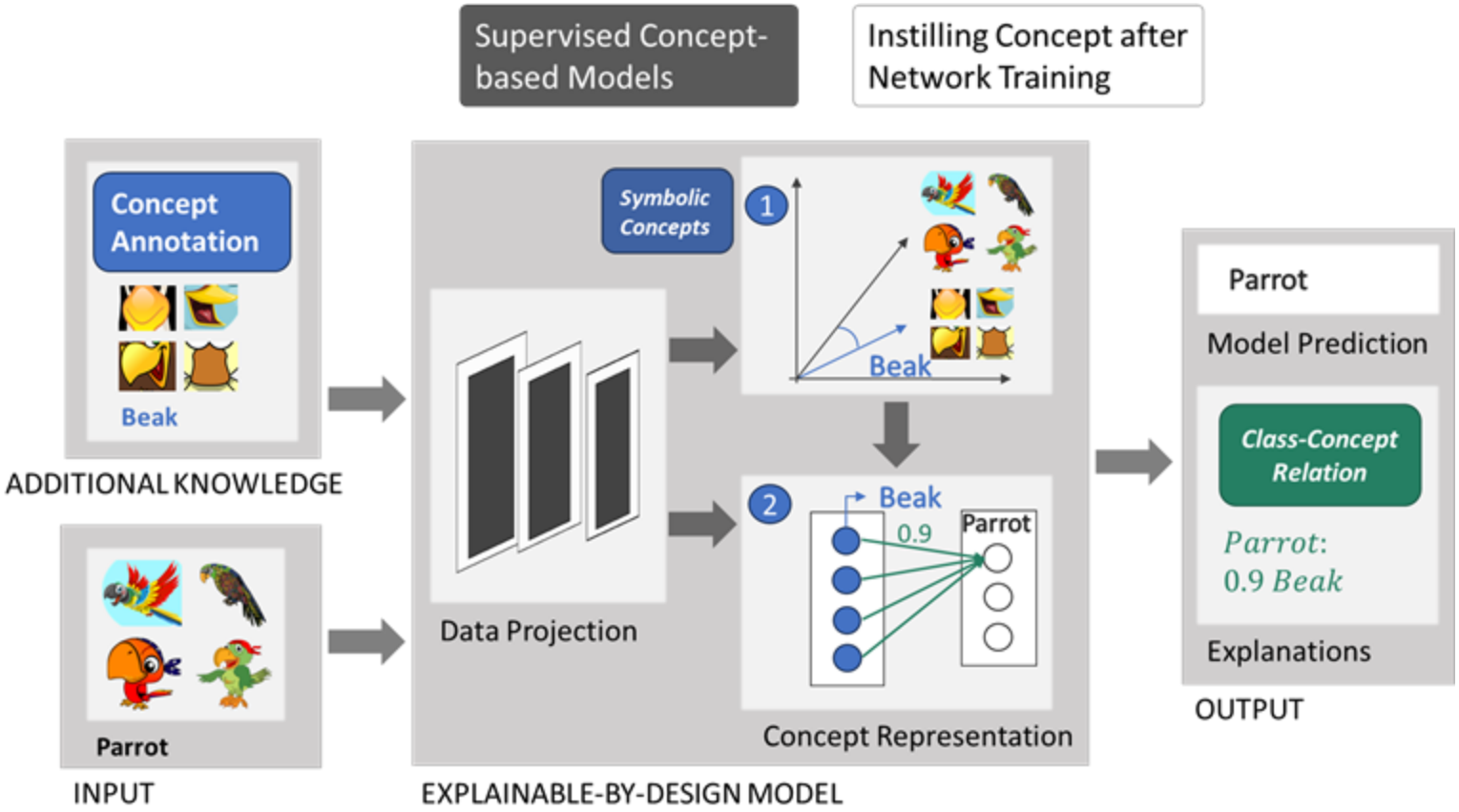
E.g., Ghorbani, Amirata, et al. "*Towards automatic concept-based explanations.*" NeurIPS (2019).


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)

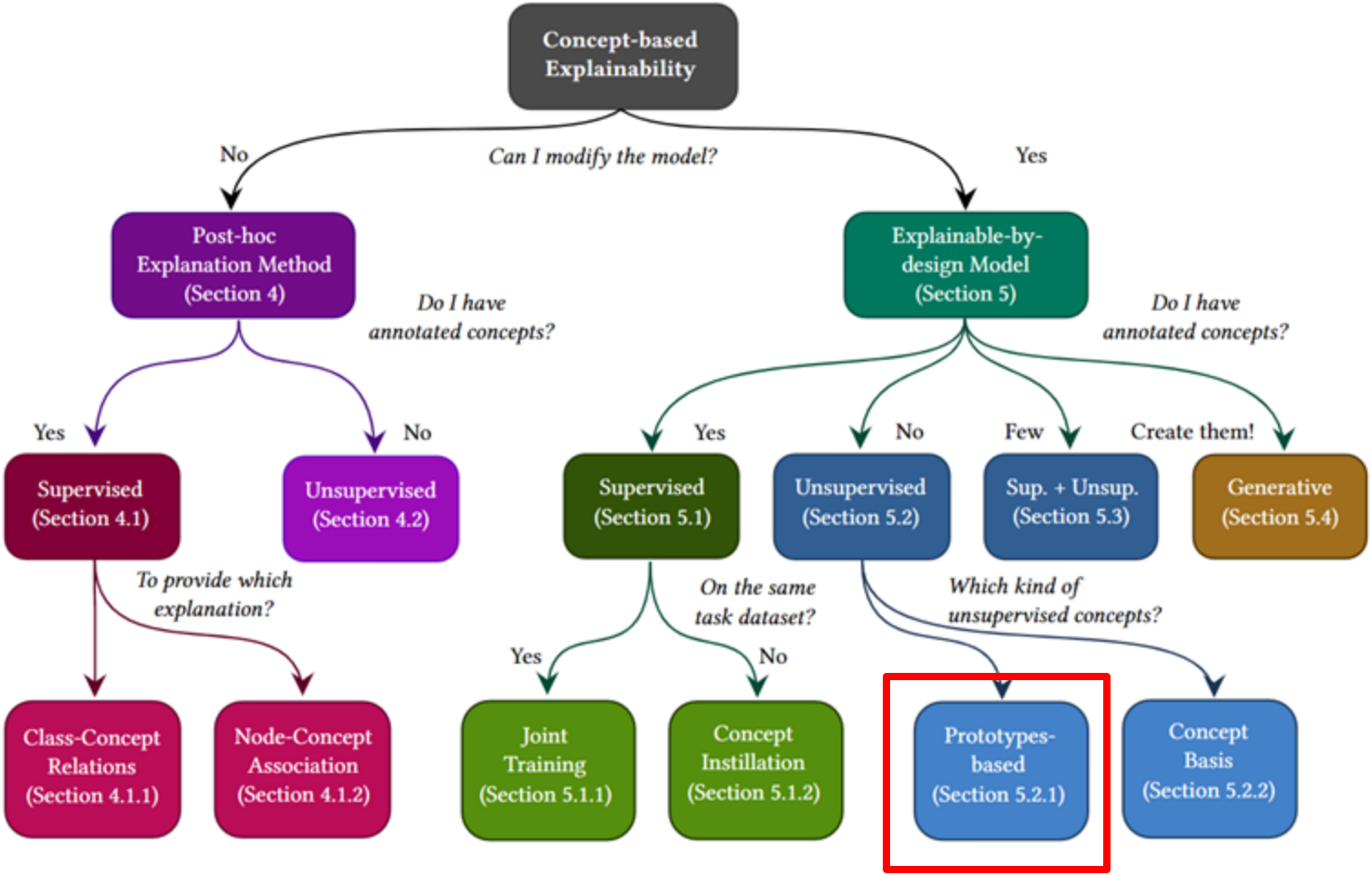


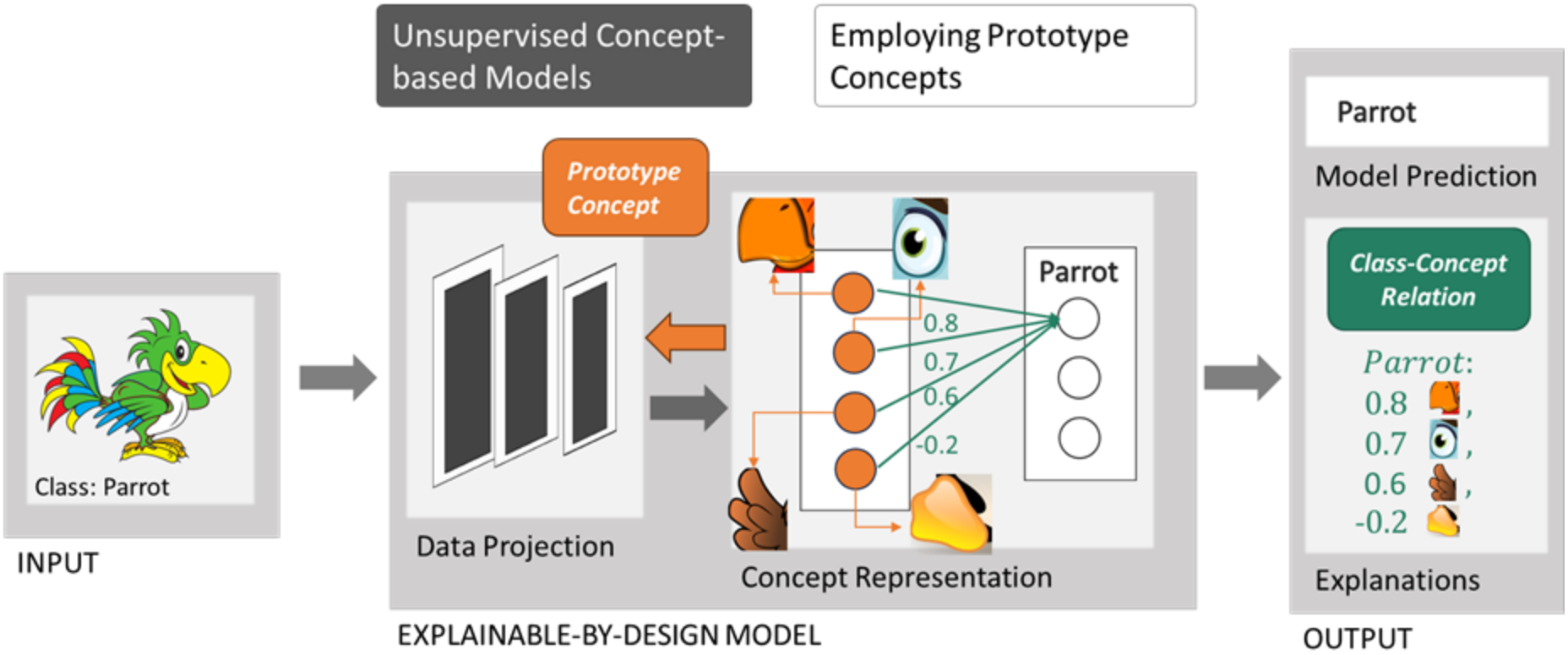
E.g., Koh, Pang Wei, et al. "*Concept bottleneck models.*" ICML (2020)


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)

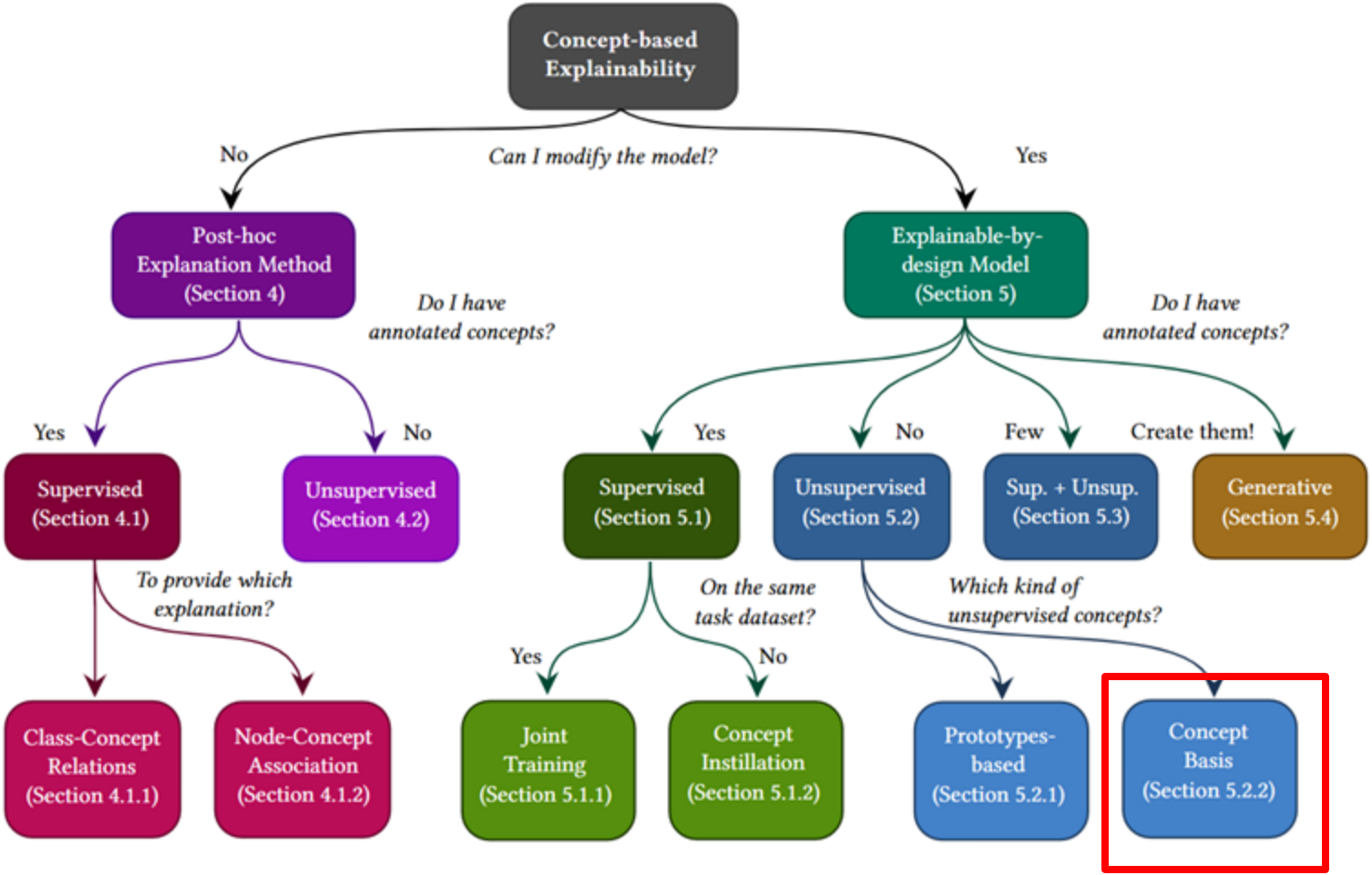


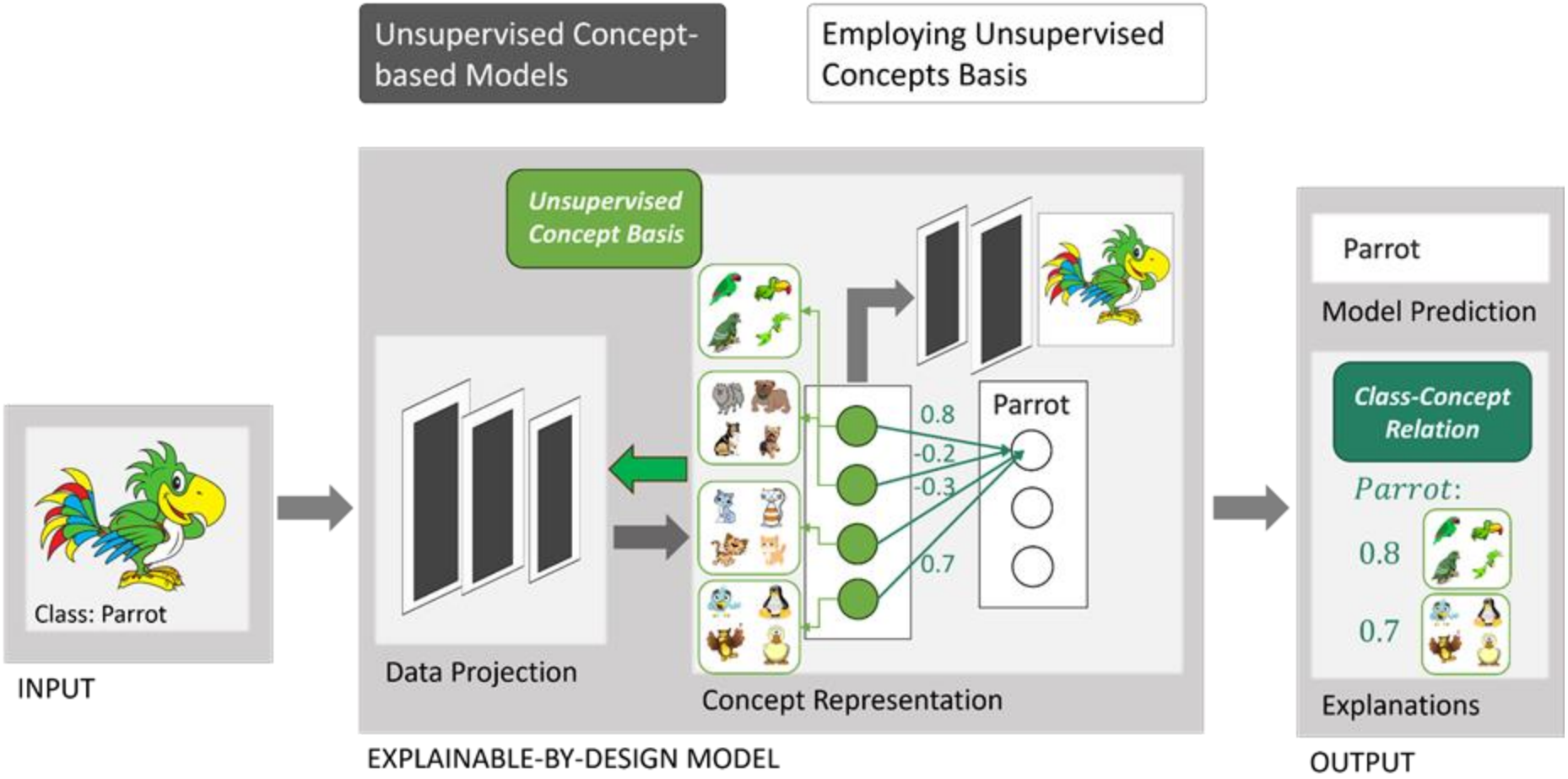
E.g., Chen, Zhi, et al. "*Concept whitening for interpretable image recognition*." Nature Machine Intelligence (2020).


Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (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)



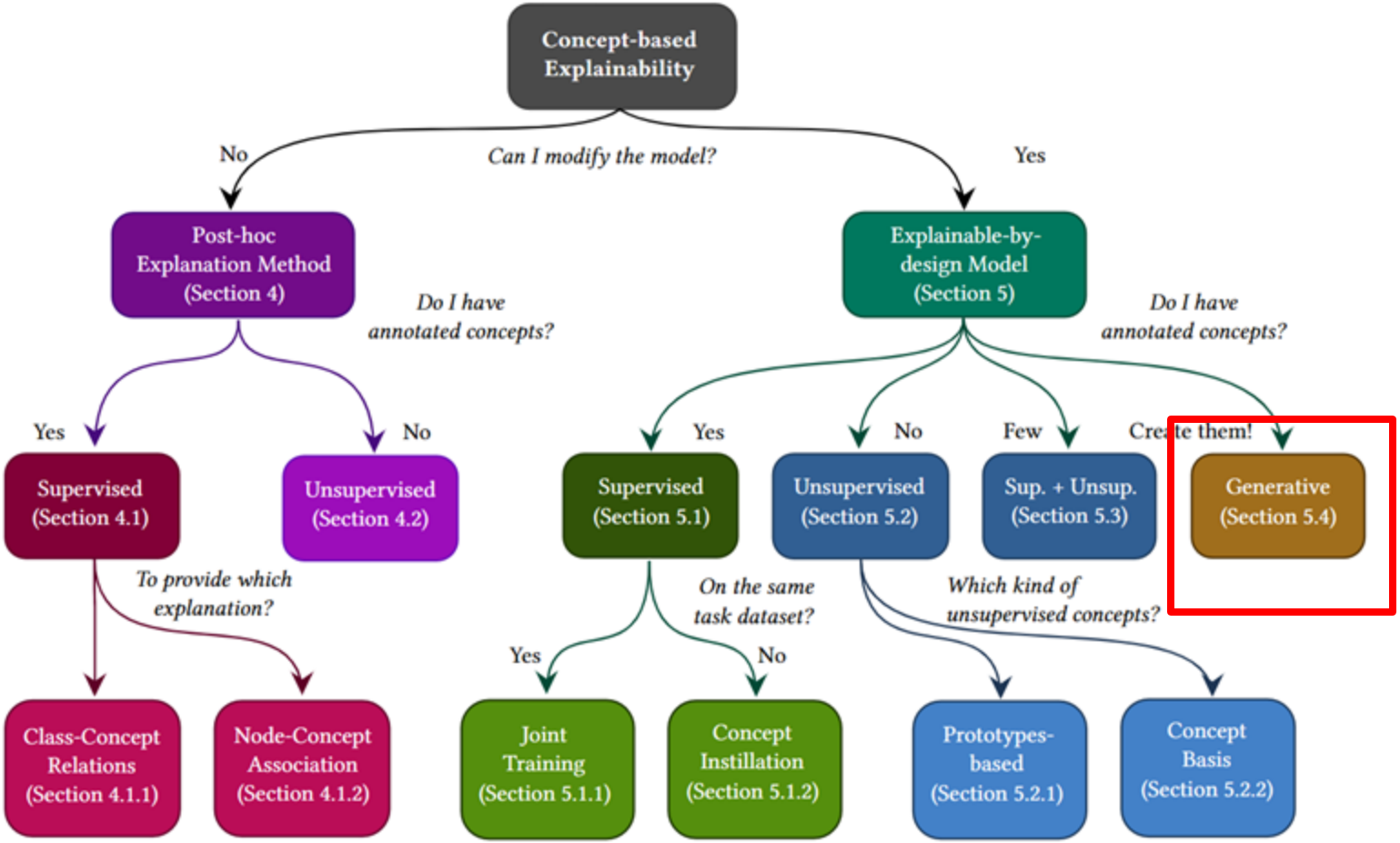
E.g., Chen, Chaofan, et al. "*This looks like that: deep learning for interpretable image recognition*." Neurips 2019.

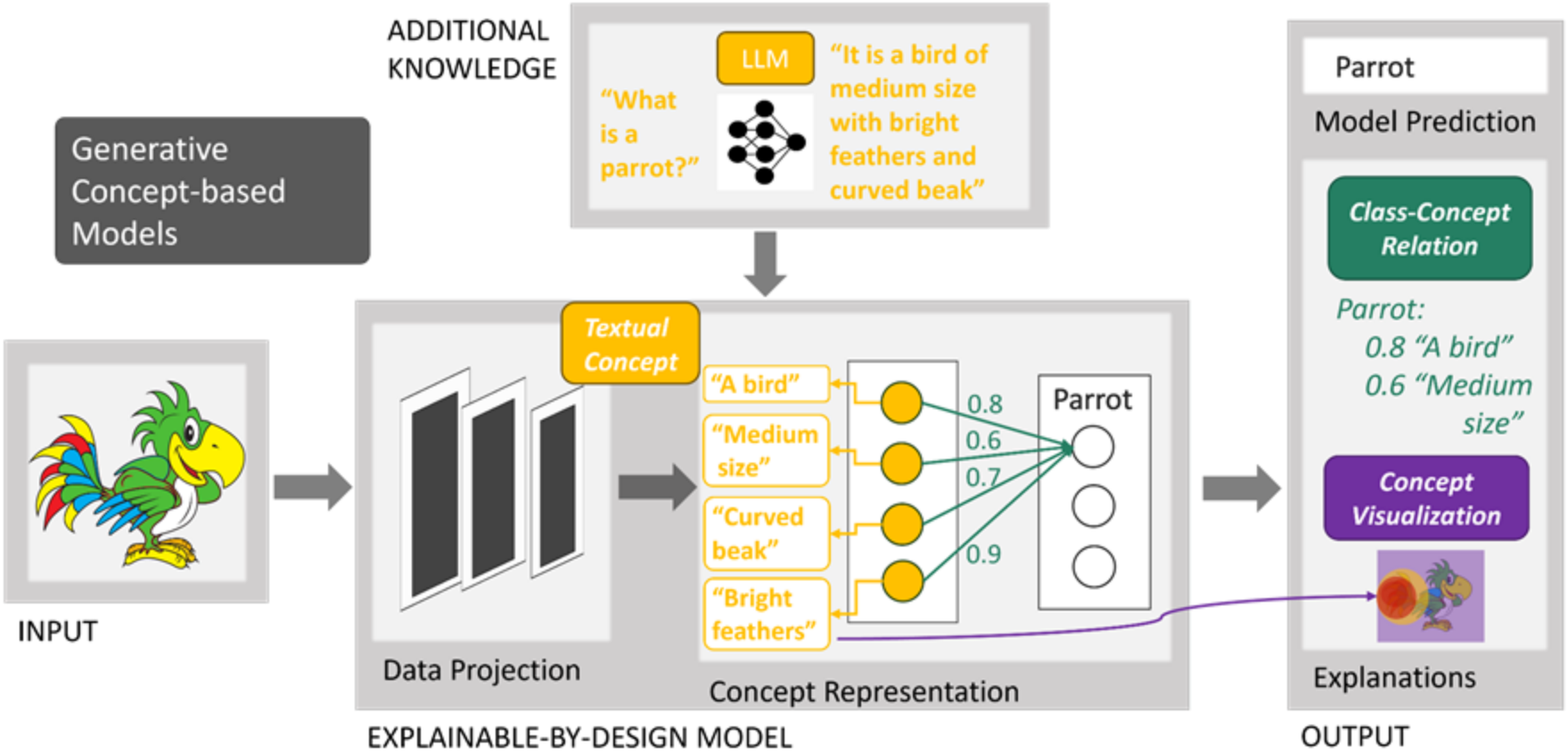

Concept-based Explainability
Can I modify the model?
No
Yes
Post-hoc Explanation Method (Section 4)
Explainable-by-design Model (Section 5)
Do I have annotated concepts?
Do I have annotated concepts?
Yes
No
Yes
No
Few
Create them!
Supervised (Section 4.1)
Unsupervised (Section 4.2)
Supervised (Section 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.

INTESA SANPAOLO INNOVATION CENTER


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)



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

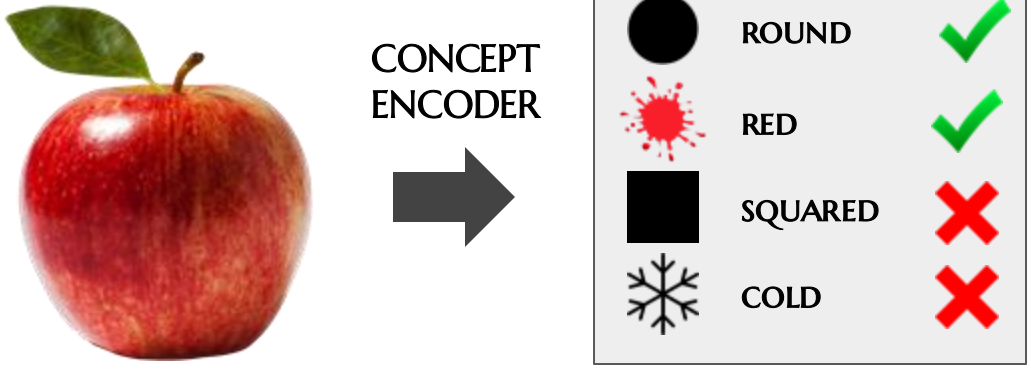
Let's see how CBMs work more in details…

Concept Bottleneck Models (CBMs)

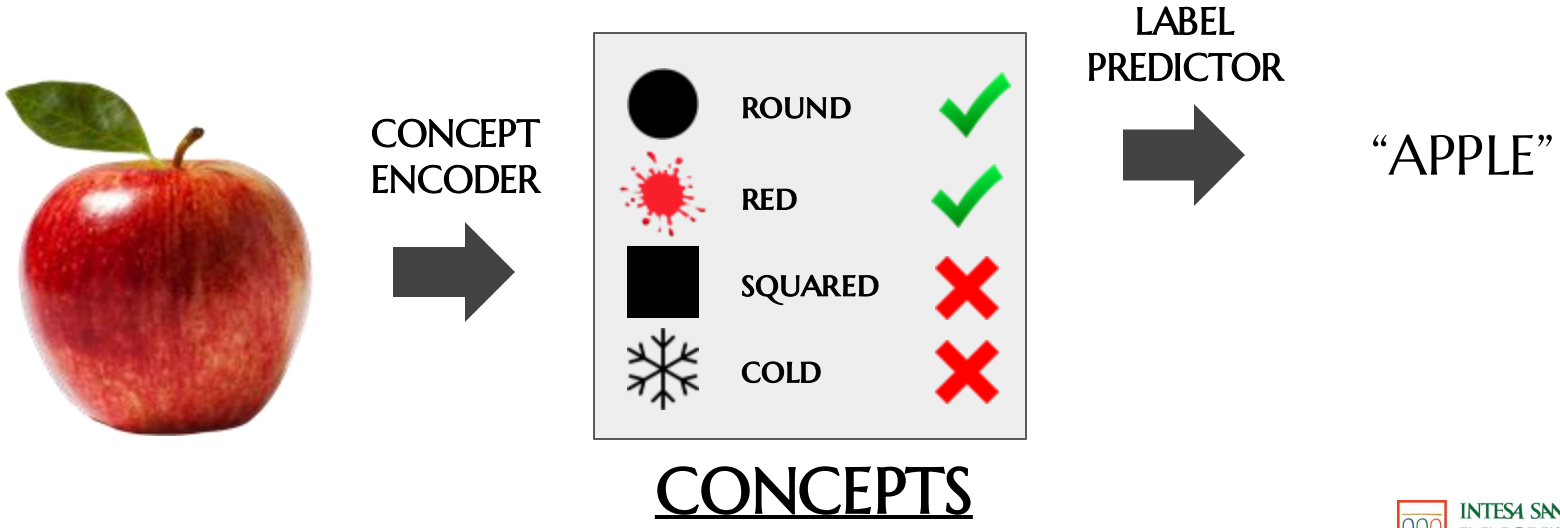
Koh, Pang Wei, et al. "Concept bottleneck models." *ICML* 2020.

CBMs Explicitly Represents Concepts

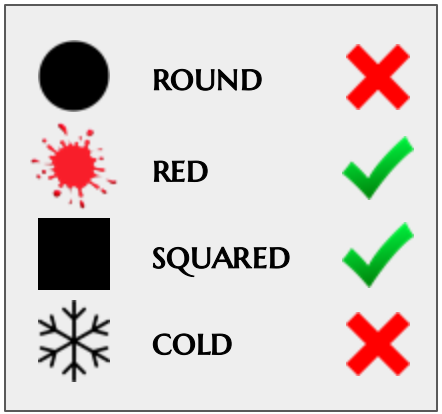
CBMs Explicitly Represents Concepts



CBMs Explicitly Represents Concepts



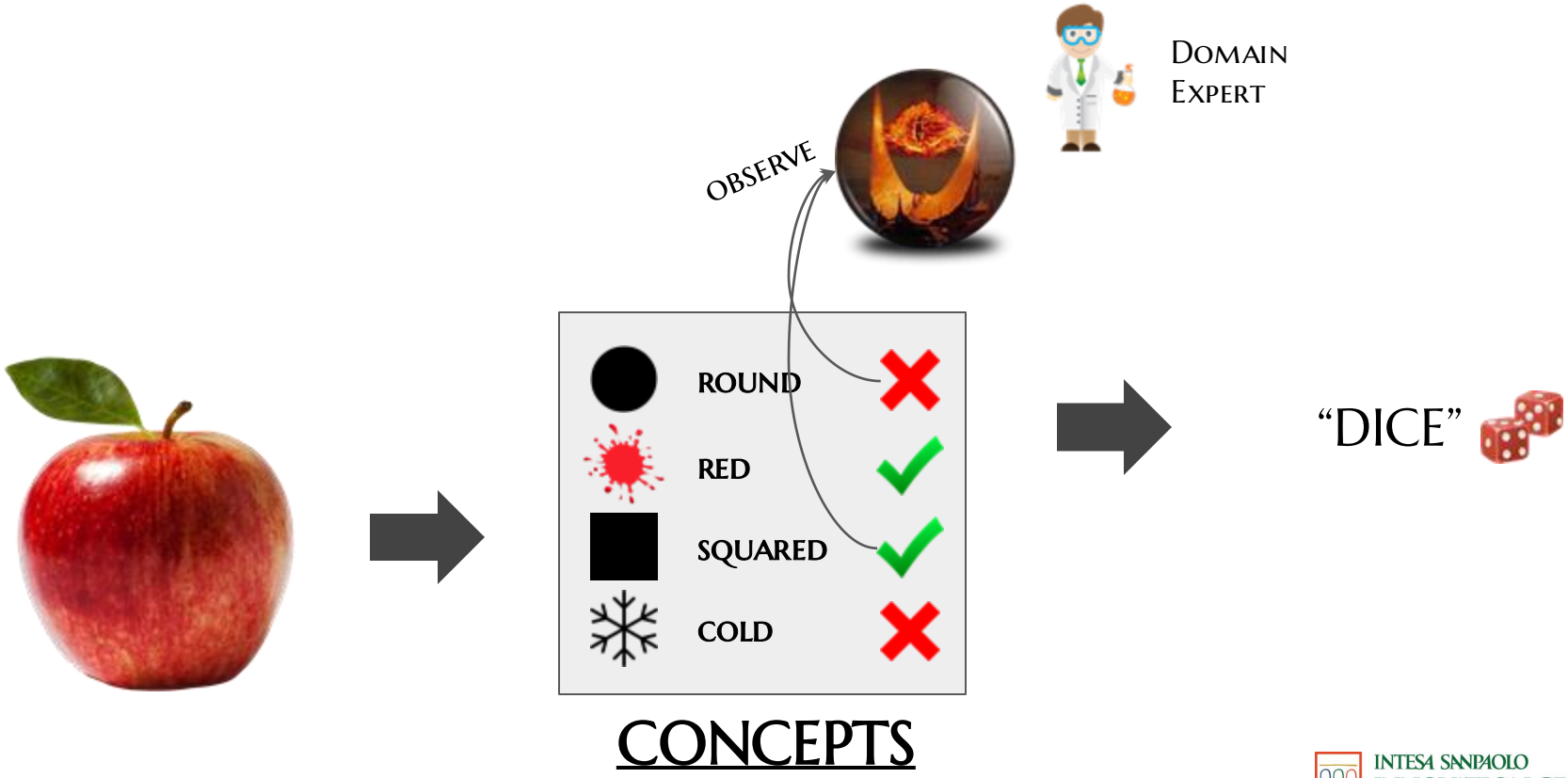
CBMs Allows Interactions!



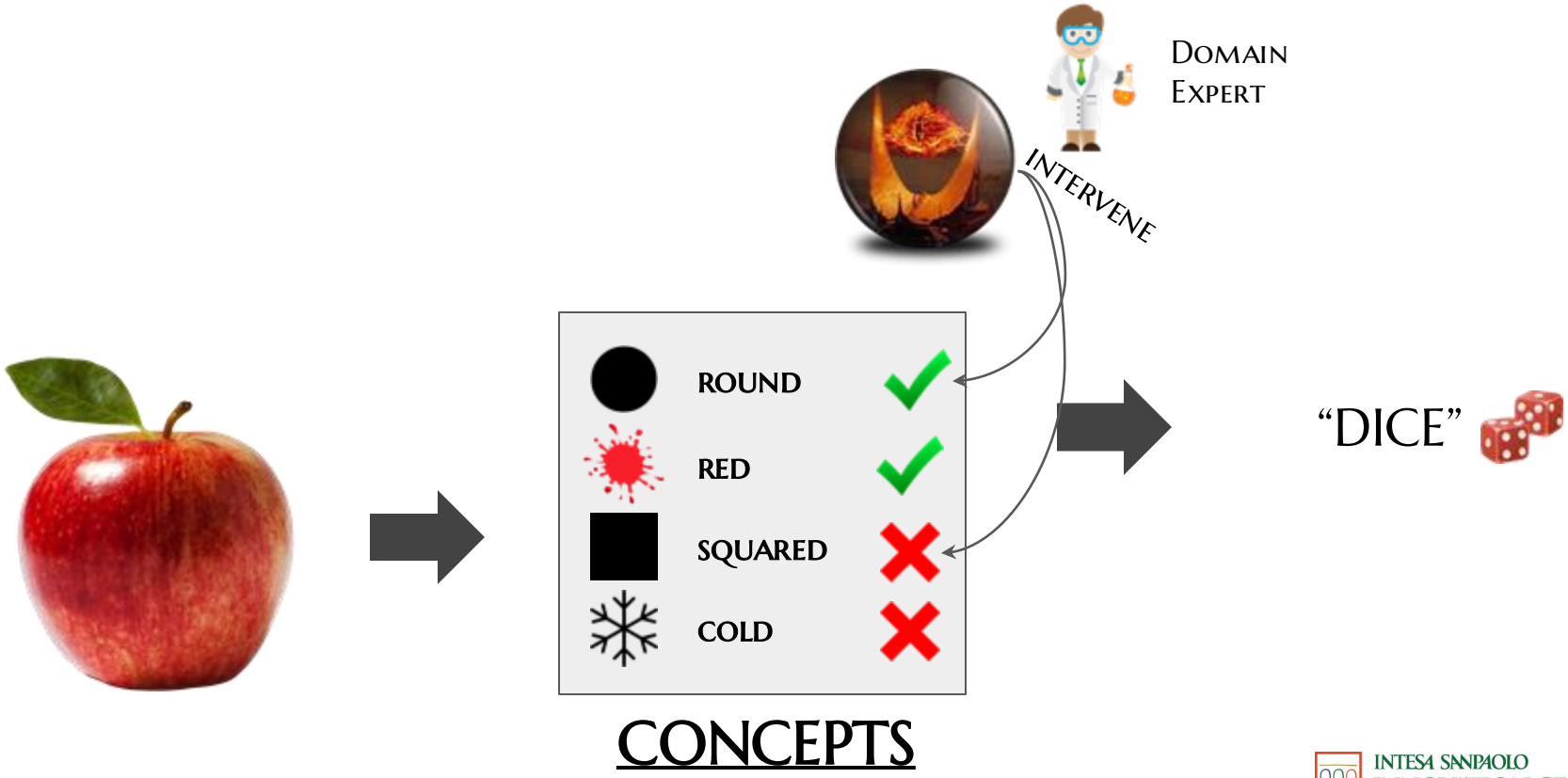
CONCEPTS

"DICE"

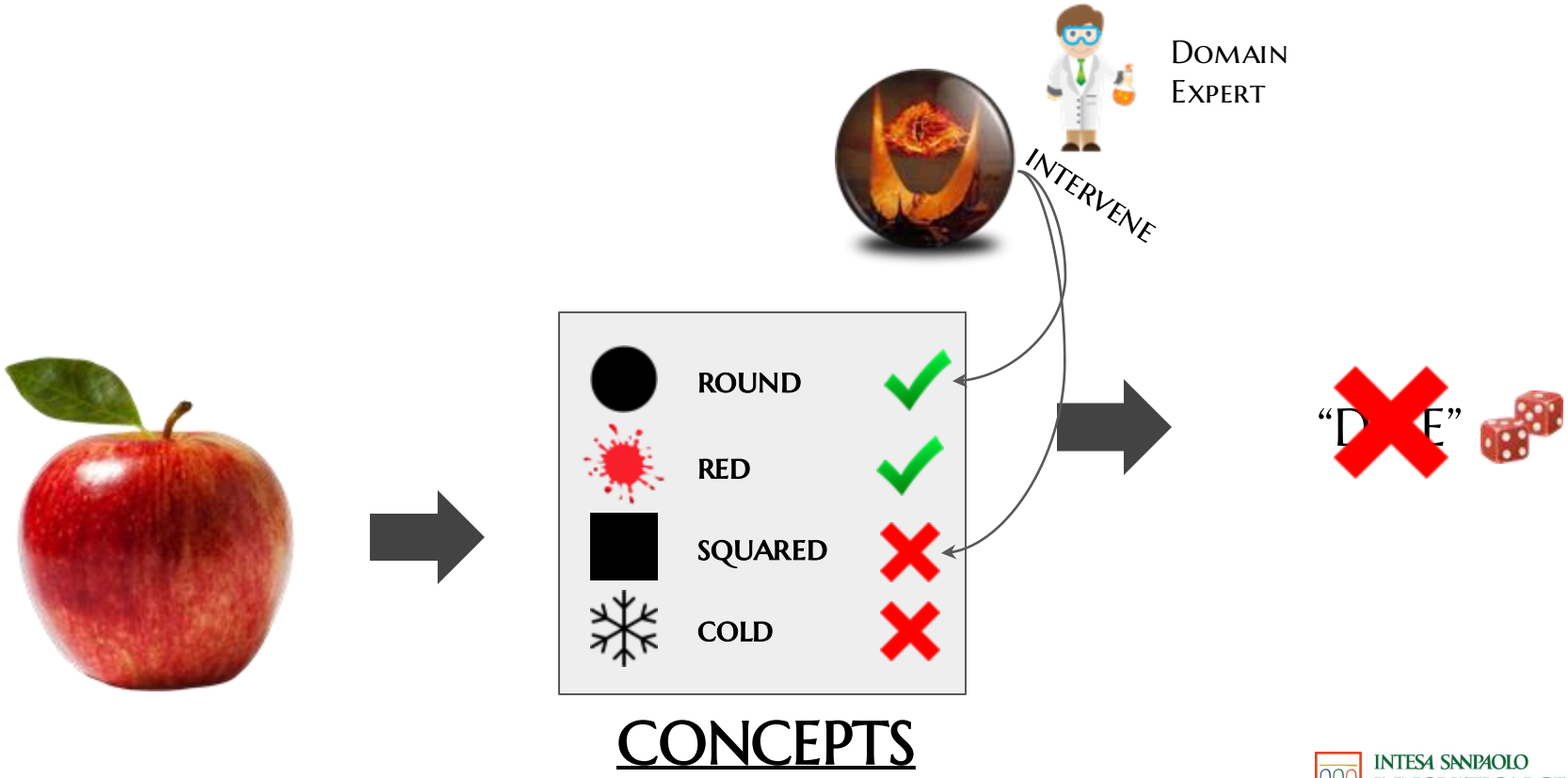
CBMs Allows Interactions!



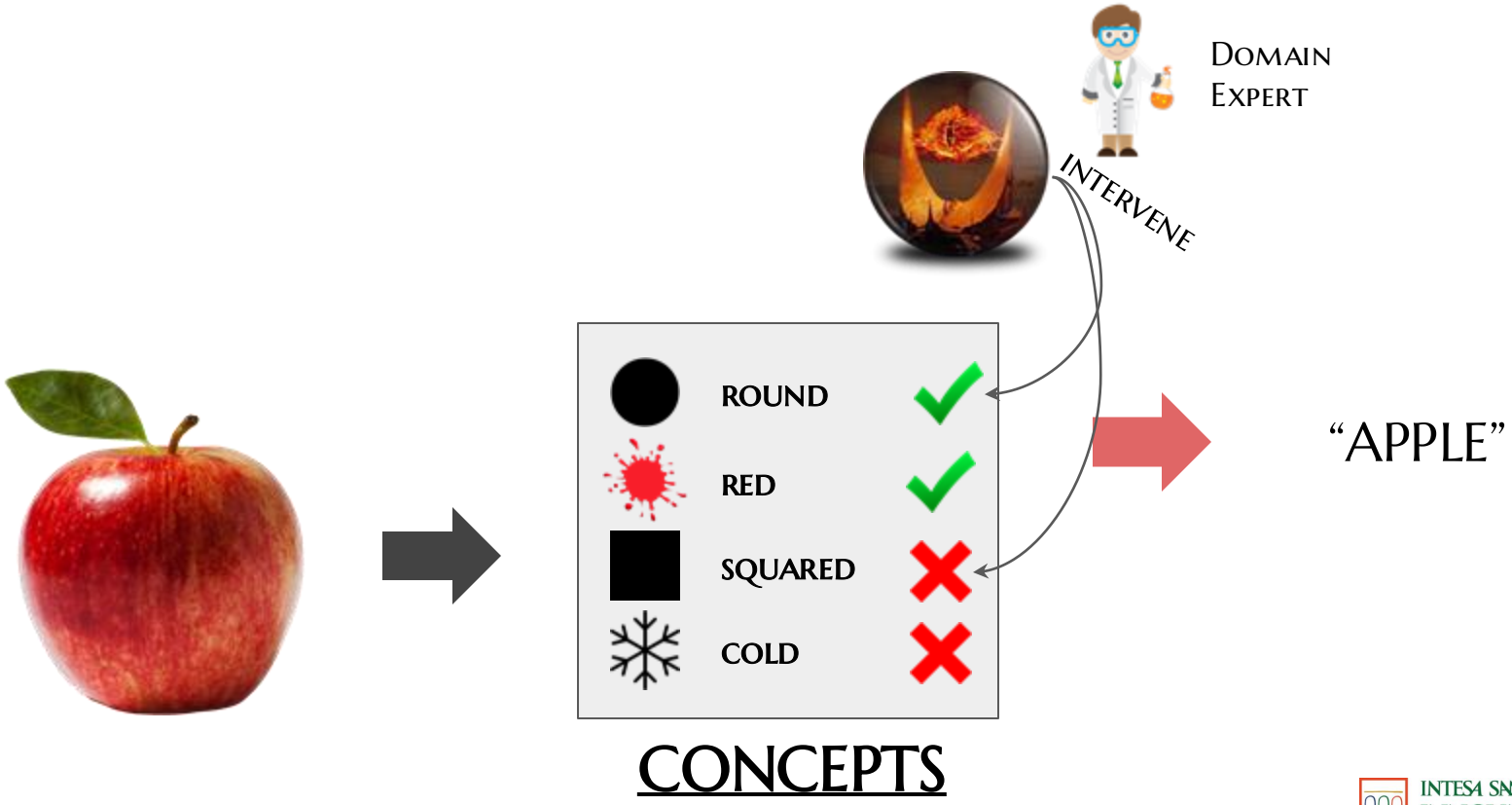
CBMs Allows Interactions!



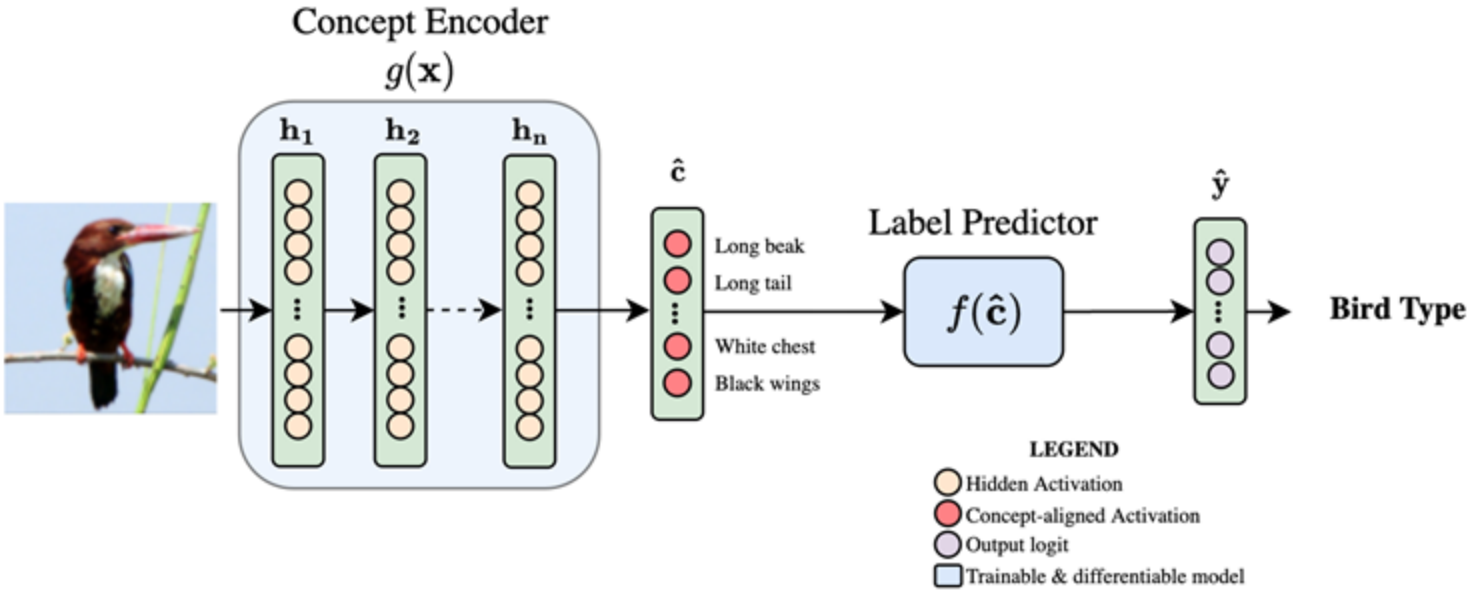
CBMs Allows Interactions!



CBMs Allows Interactions!



Concept bottleneck models architecture



Concept bottleneck models are not the solution

Poor Trade-offs

Struggle to compromise between accuracy and explainability

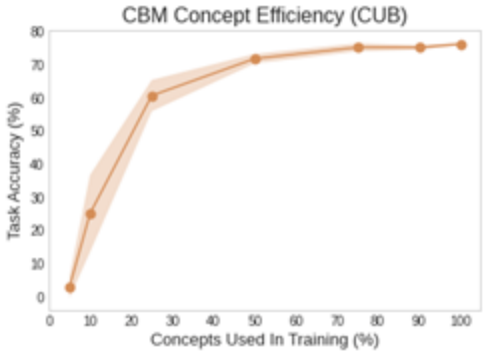
Concept bottleneck models are not the solution

Poor Trade-offs

Struggle to compromise between accuracy and explainability

Low Concept Efficiency

CBMs do not scale in real-world conditions

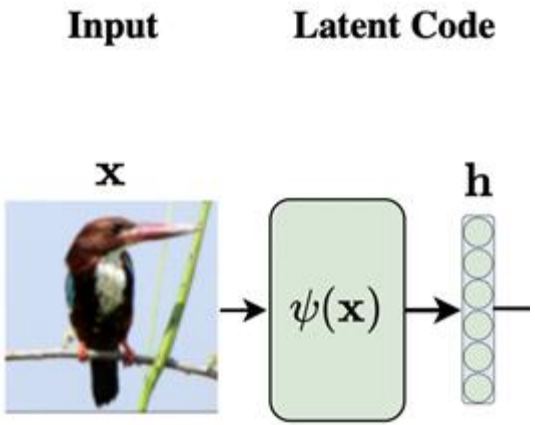


Concept Embedding Models (CEM)

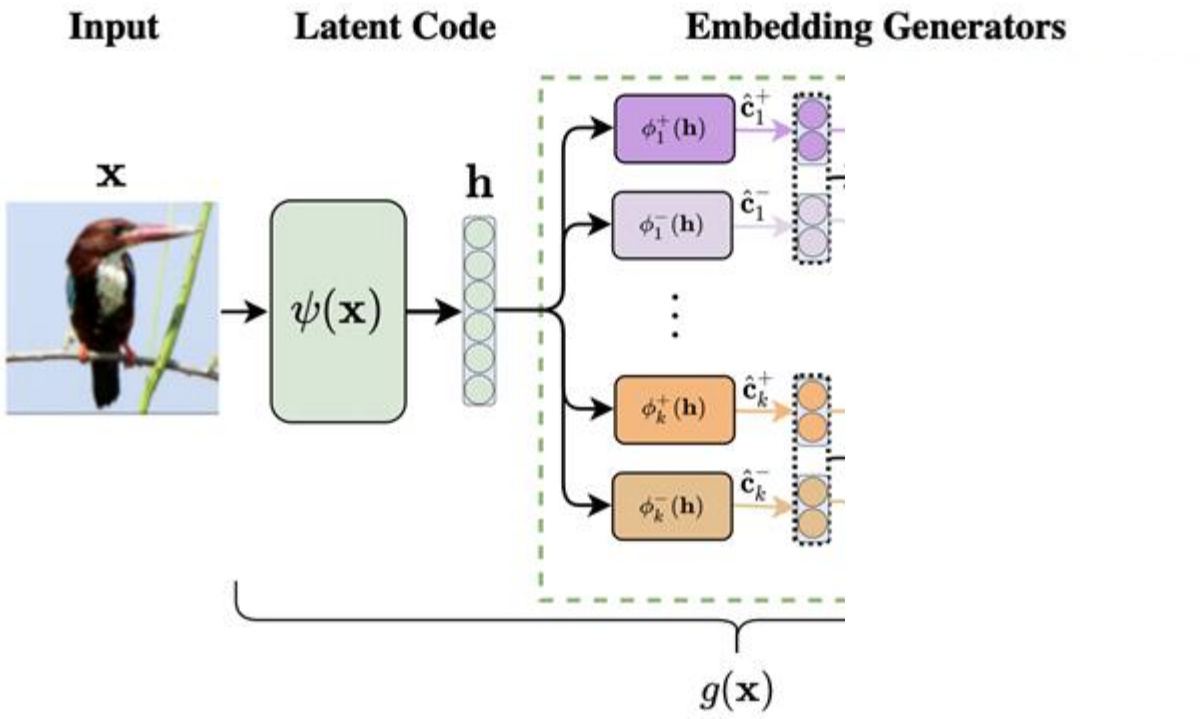
Espinosa Zarlenga, Mateo, et al. "Concept embedding models: Beyond the accuracy-explainability trade-off." *Advances in Neural Information Processing* (2022).

Solution: Concept Embedding Models

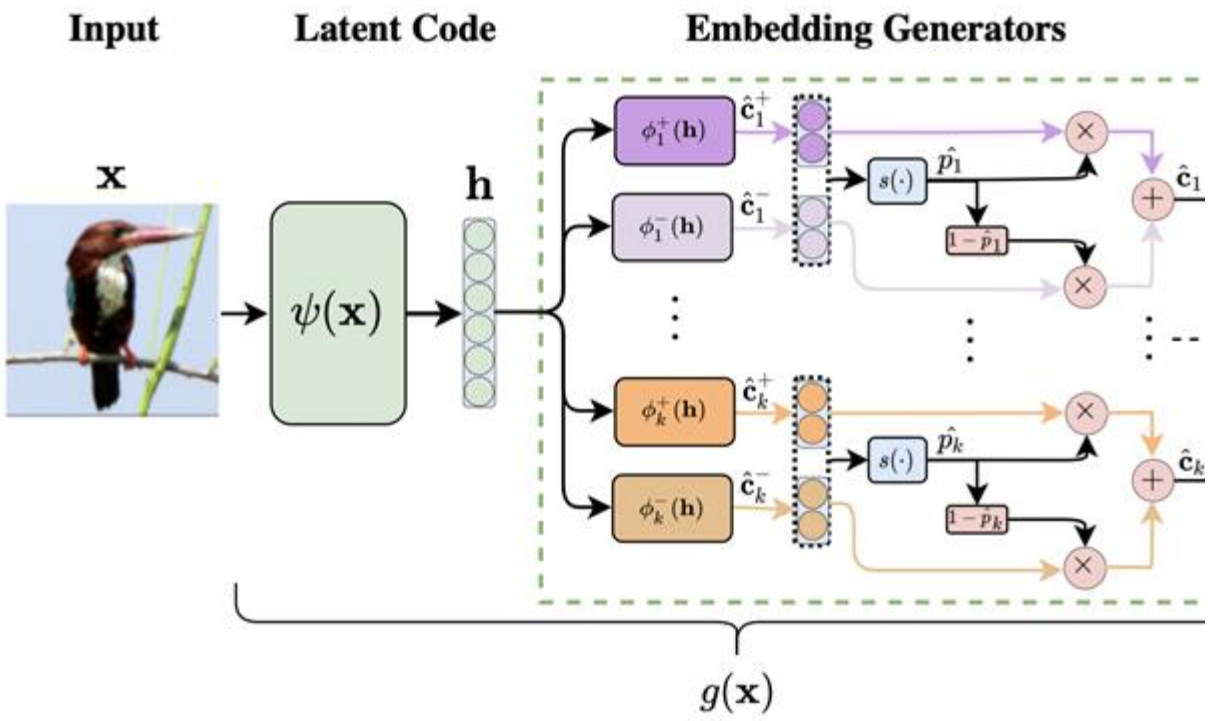
Solution: Concept Embedding Models



Solution: Concept Embedding Models

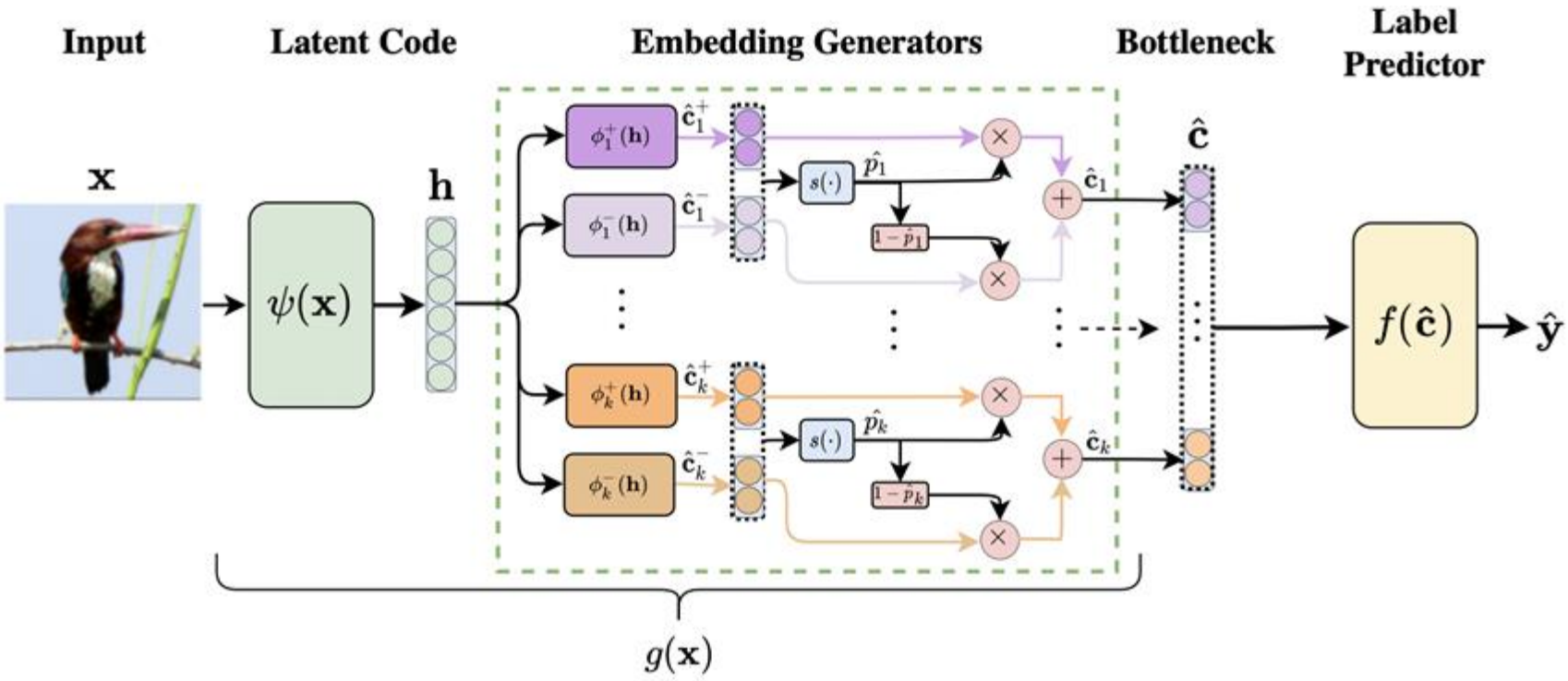


Solution: Concept Embedding Models



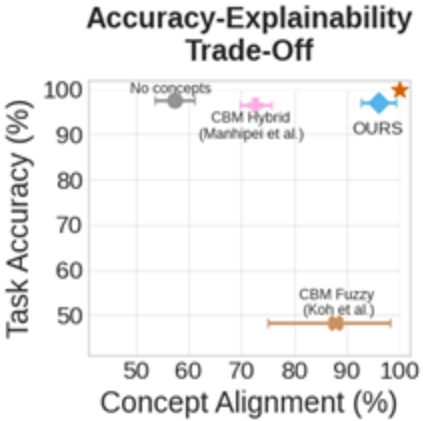
INTESA SANPAOLO INNOVATION CENTER

Solution: Concept Embedding Models



CEM Advantages

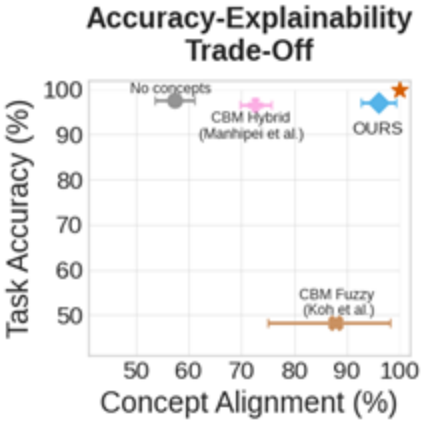
CEM Advantages

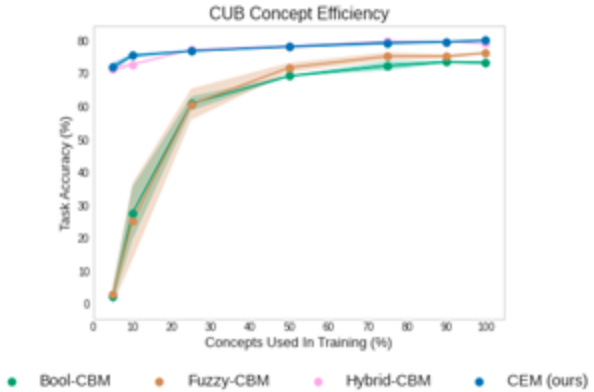


Beyond Trade-offs

CEMs overcome the current accuracy-explainability trade-off

CEM Advantages





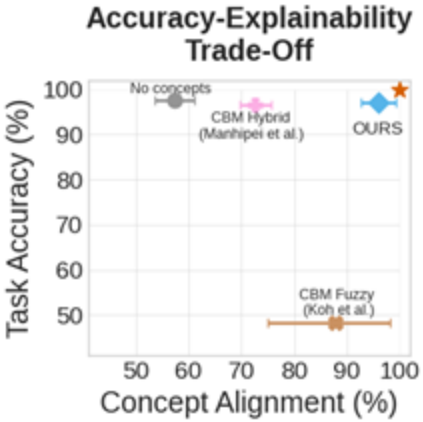
Beyond Trade-offs

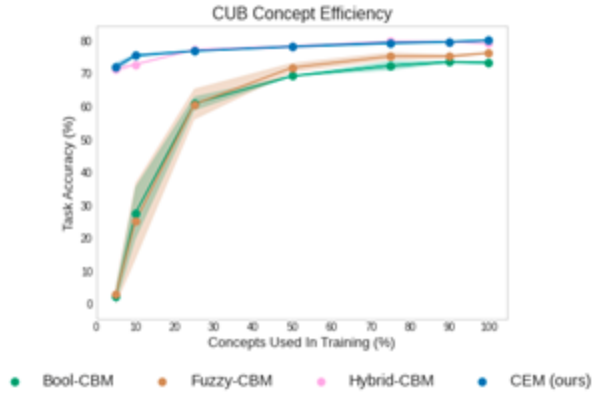
CEMs overcome the current accuracy-explainability trade-off

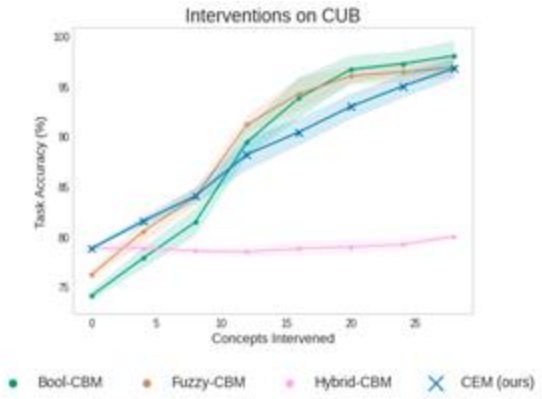
High Concept Efficiency

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

CEM Advantages







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

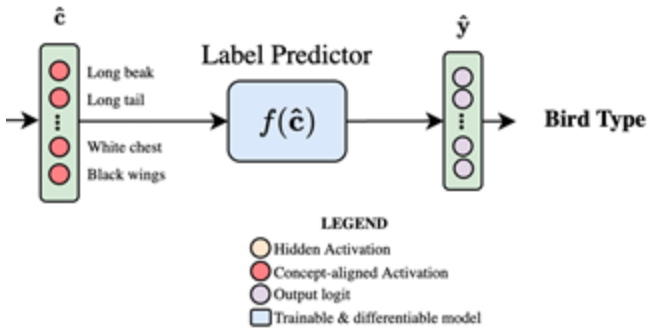
INTESA SANPAOLO INNOVATION CENTER

Have we lost something?

Have we lost something?

Interpretability

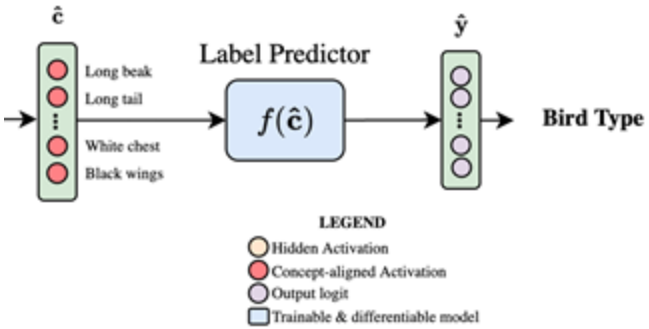
CBM: Interpretable



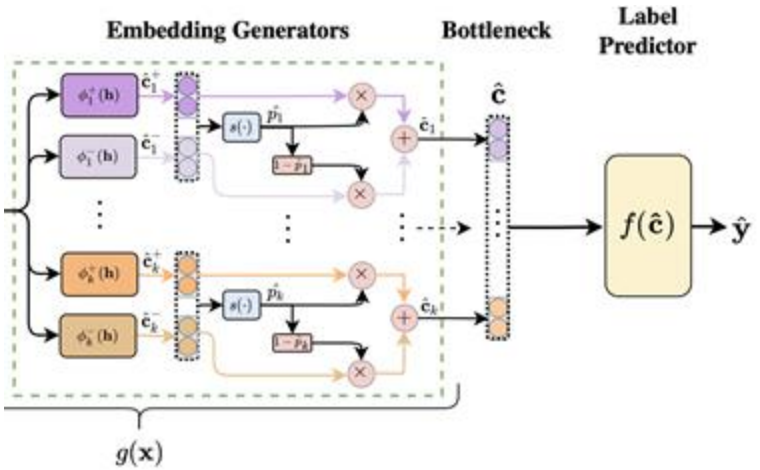
Have we lost something?

Interpretability

CBM: Interpretable



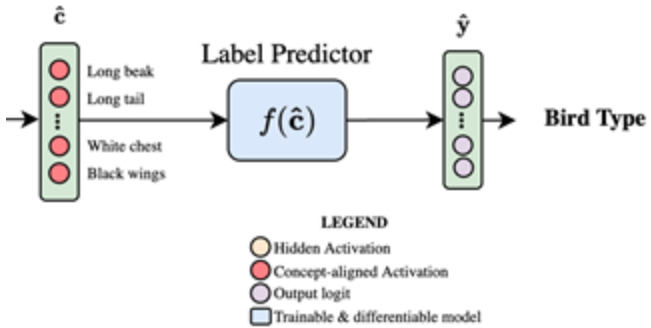
CEM: NON-Interpretable



Have we lost something?

Interpretability

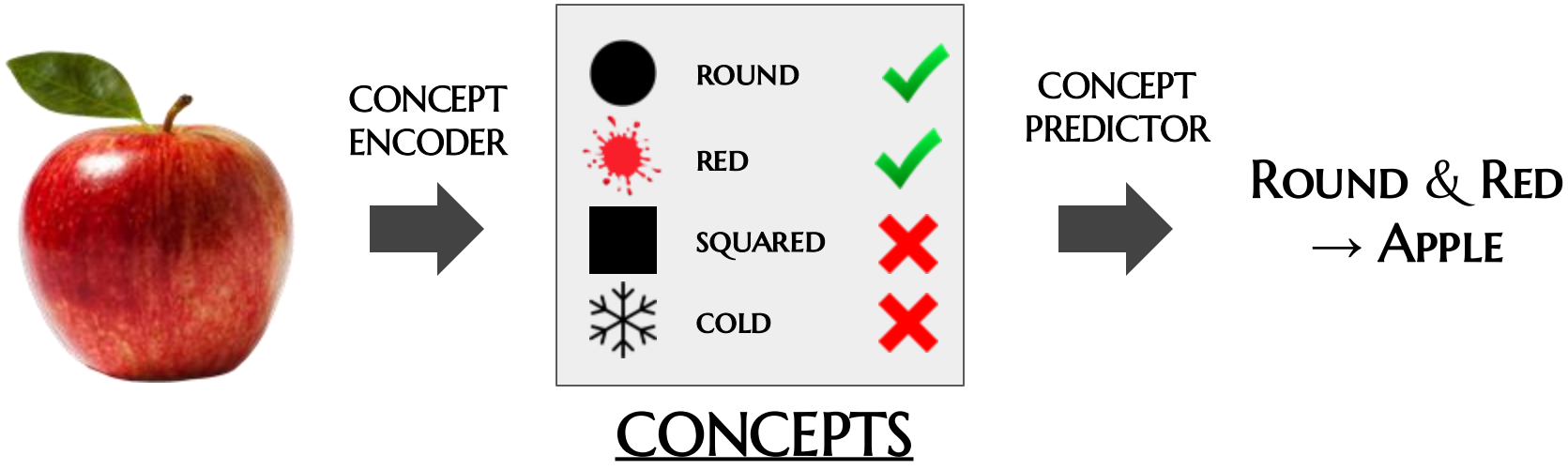
CBM: Interpretable



CEM: NON-Interpretable

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

Can we create an Interpretable Model over CE?



Deep Concept Reasoner (DCR)

Barbiero, Pietro, et al. "Interpretable Neural-Symbolic Concept Reasoning." International Conference on Machine Learning (2023).

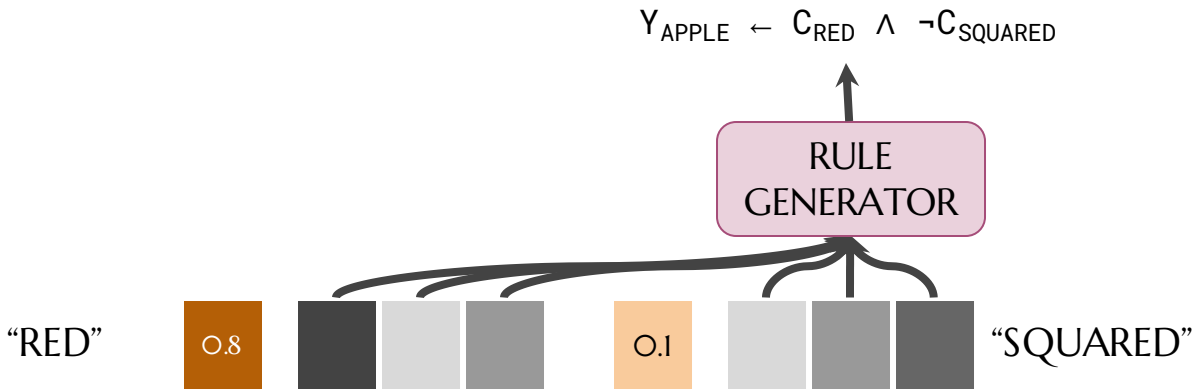
DEEP CONCEPT REASONER

DEEP CONCEPT REASONER

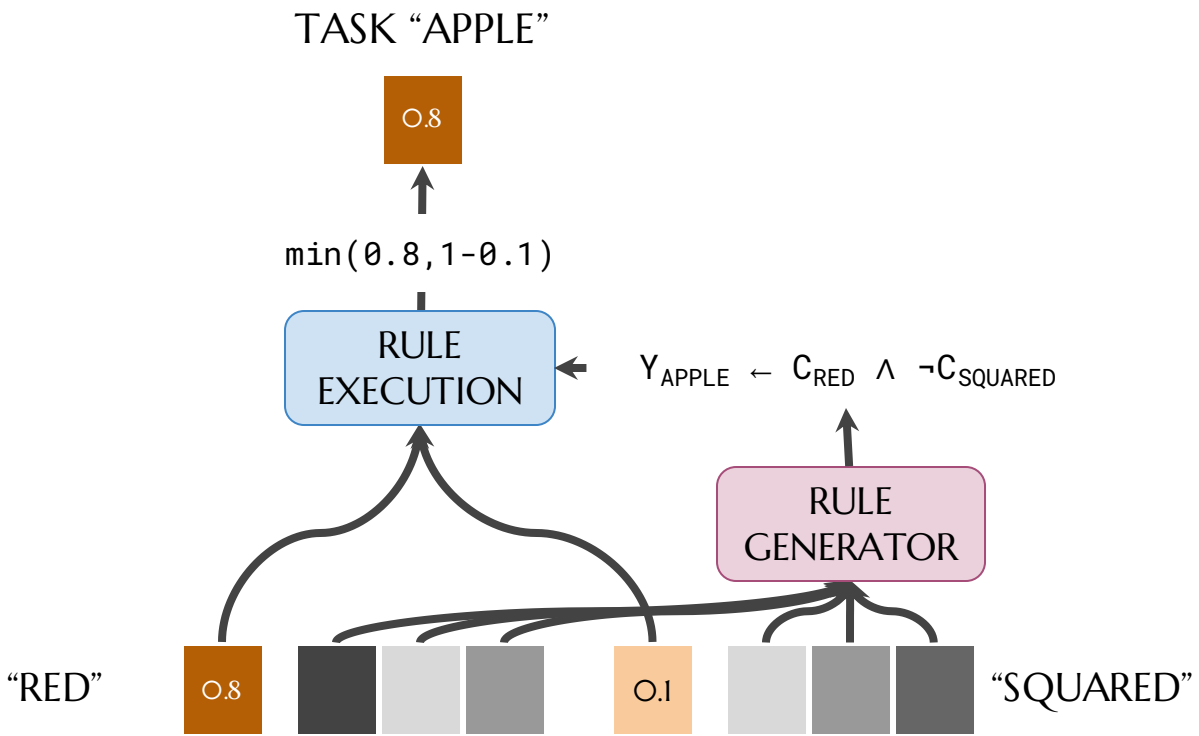
INTESA SANPAOLO INNOVATION CENTER

DEEP CONCEPT REASONER

1. GENERATES SYNTACTIC RULES WITH CONCEPT EMBEDDINGS



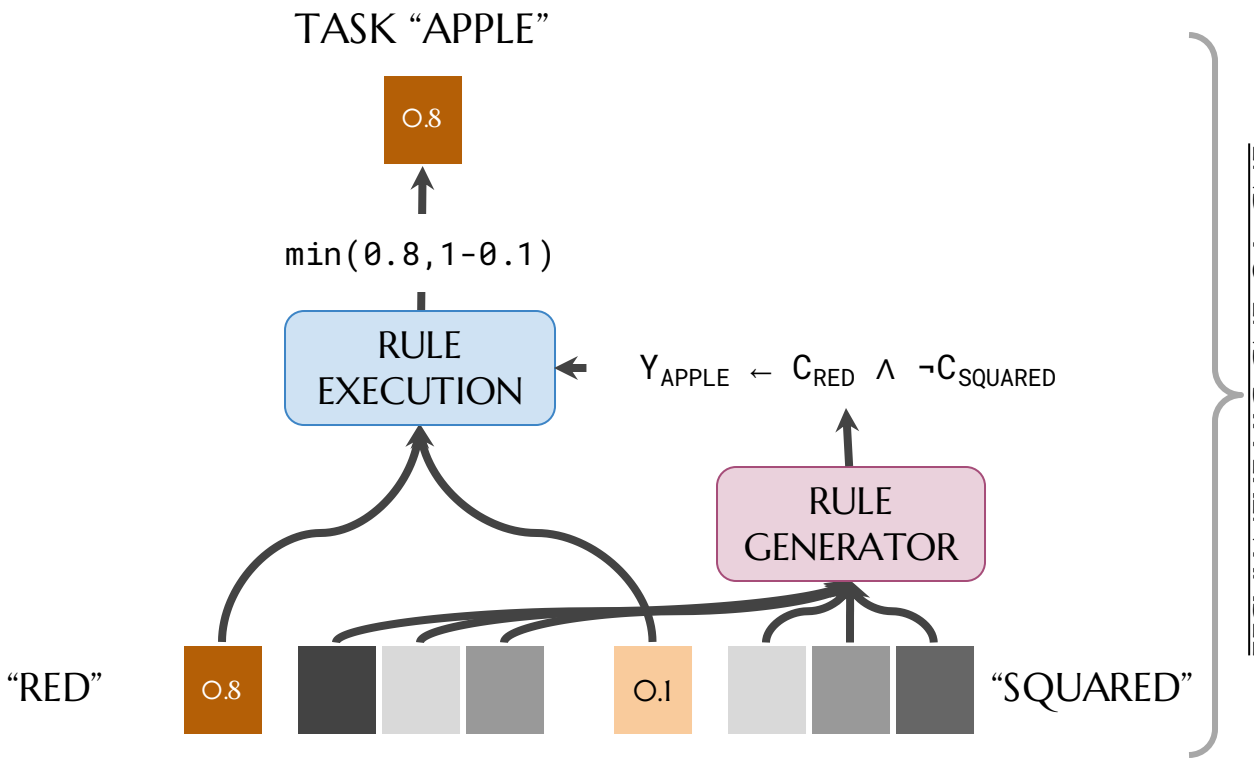
DEEP CONCEPT REASONER



1. GENERATES SYNTACTIC RULES WITH CONCEPT EMBEDDINGS

1. EXECUTES RULES SYMBOLICALLY WITH CONCEPT TRUTH VALUES

DEEP CONCEPT REASONER

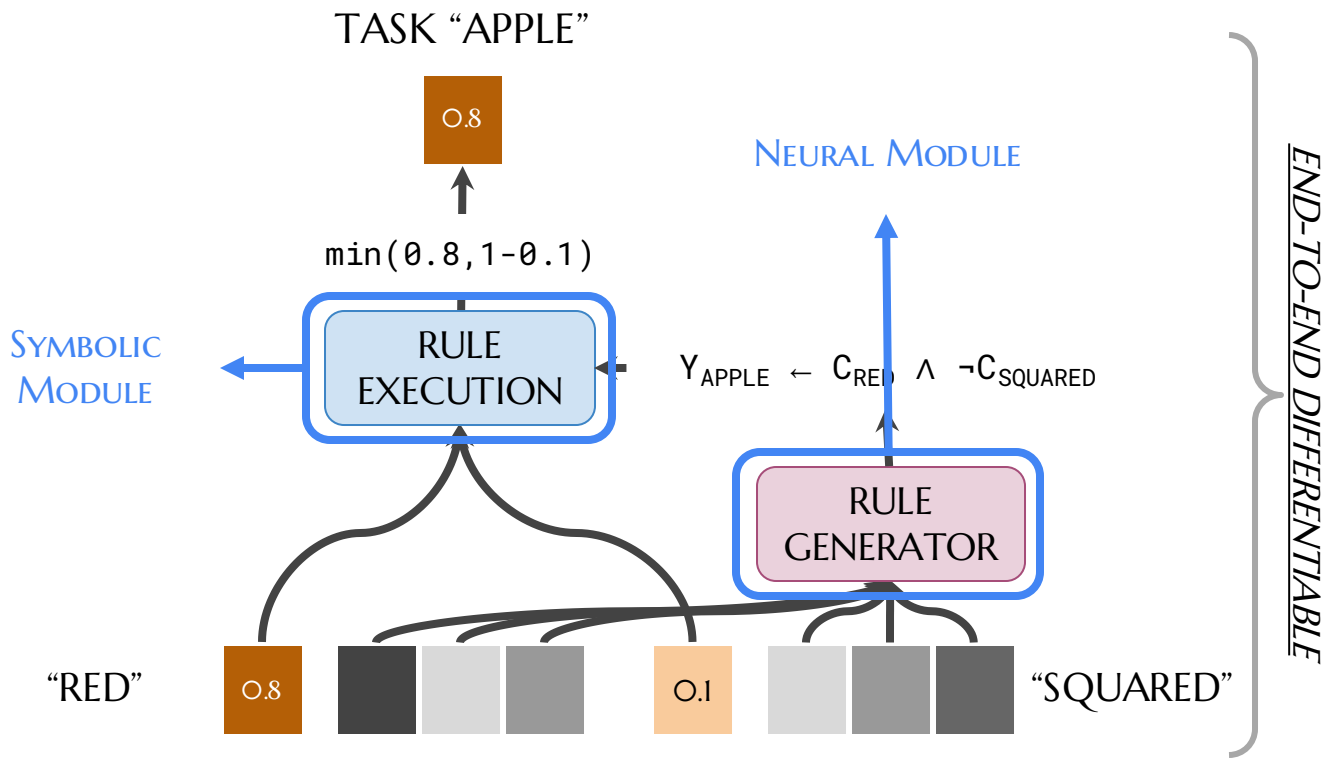


END-TO-END DIFFERENTIABLE

1. GENERATES SYNTACTIC RULES WITH CONCEPT EMBEDDINGS

1. EXECUTES RULES SYMBOLICALLY WITH CONCEPT TRUTH VALUES

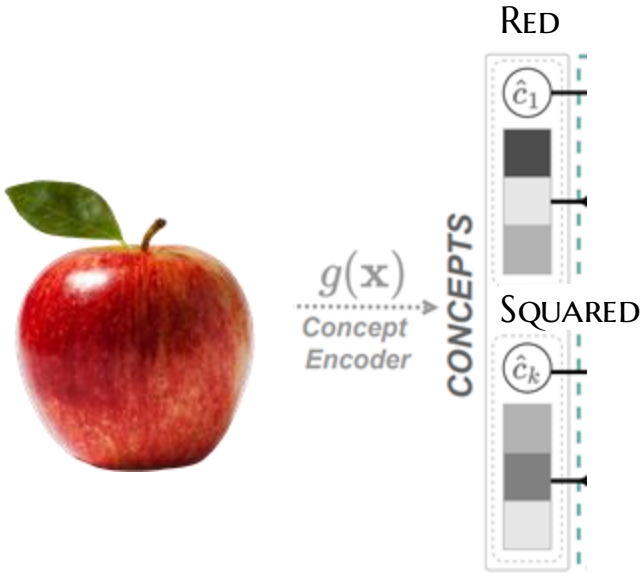
DEEP CONCEPT REASONER



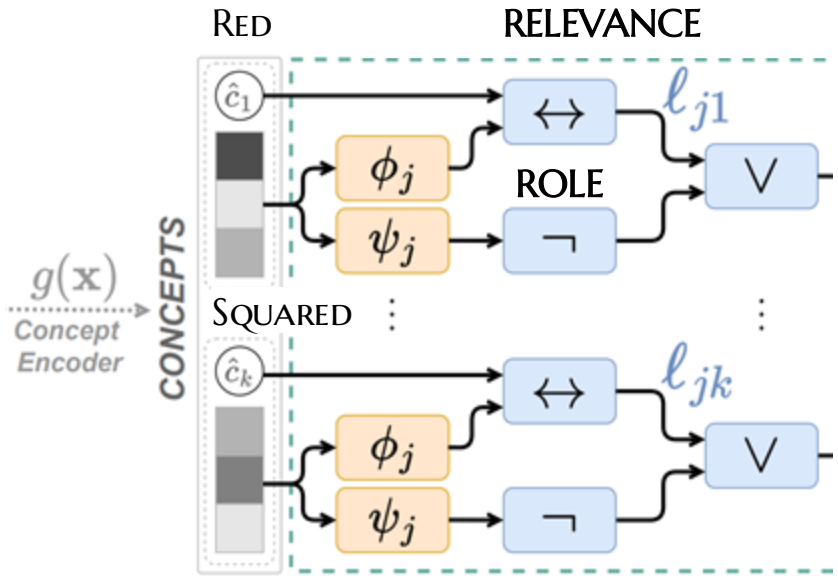
1. GENERATES SYNTACTIC RULES WITH CONCEPT EMBEDDINGS

1. EXECUTES RULES SYMBOLICALLY WITH CONCEPT TRUTH VALUES

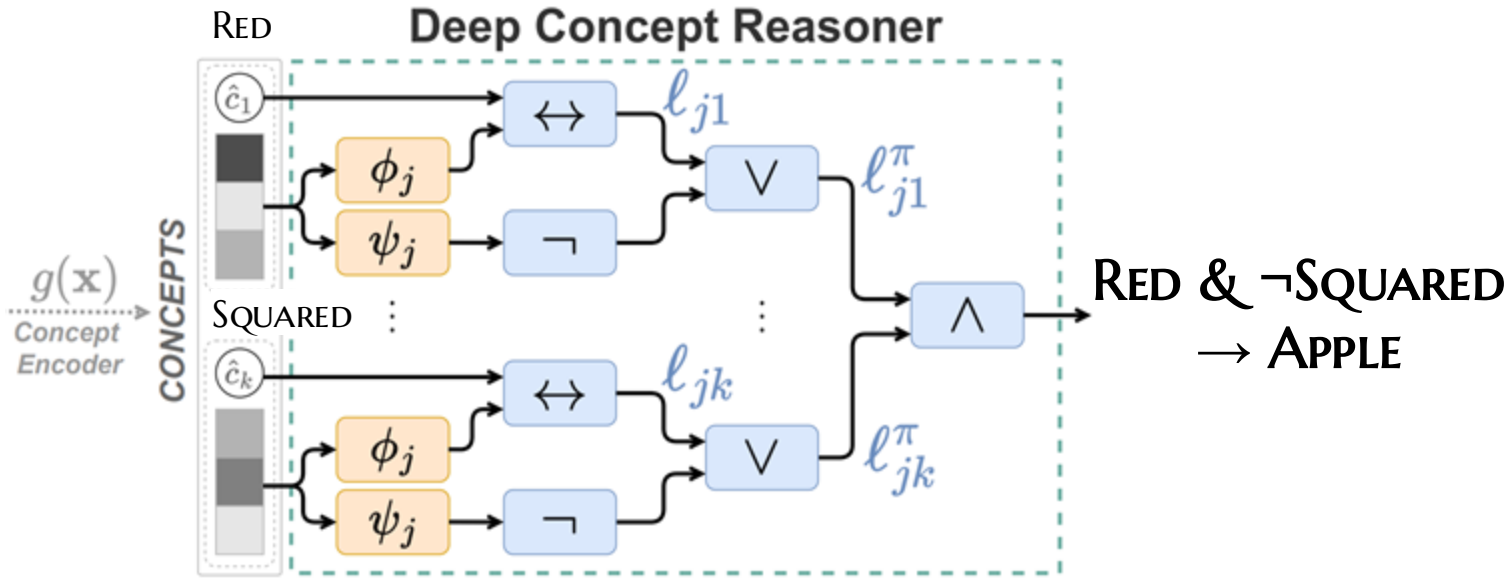
Deep Concept Reasoner (DCR)



Deep Concept Reasoner (DCR)

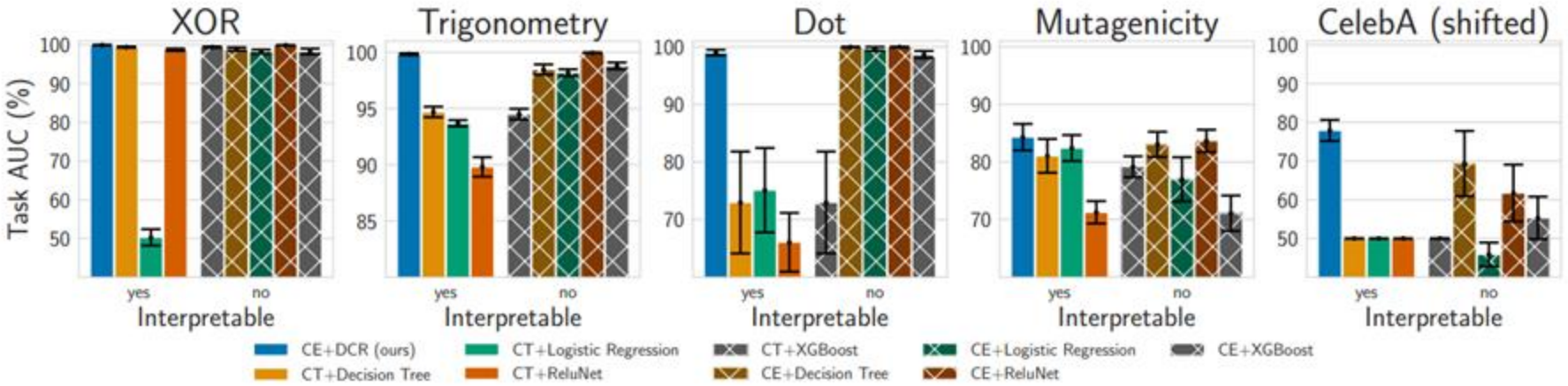


Deep Concept Reasoner (DCR)



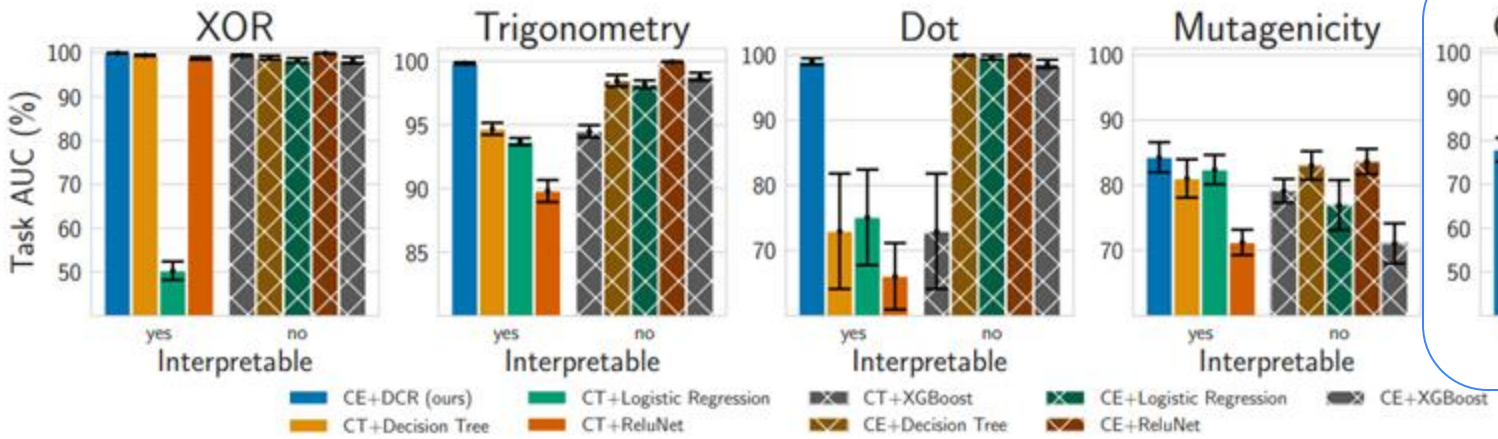
Does DCR Generalize Well?

Does DCR Generalize Well?



Does DCR Generalize Well?

Different set of test concepts!

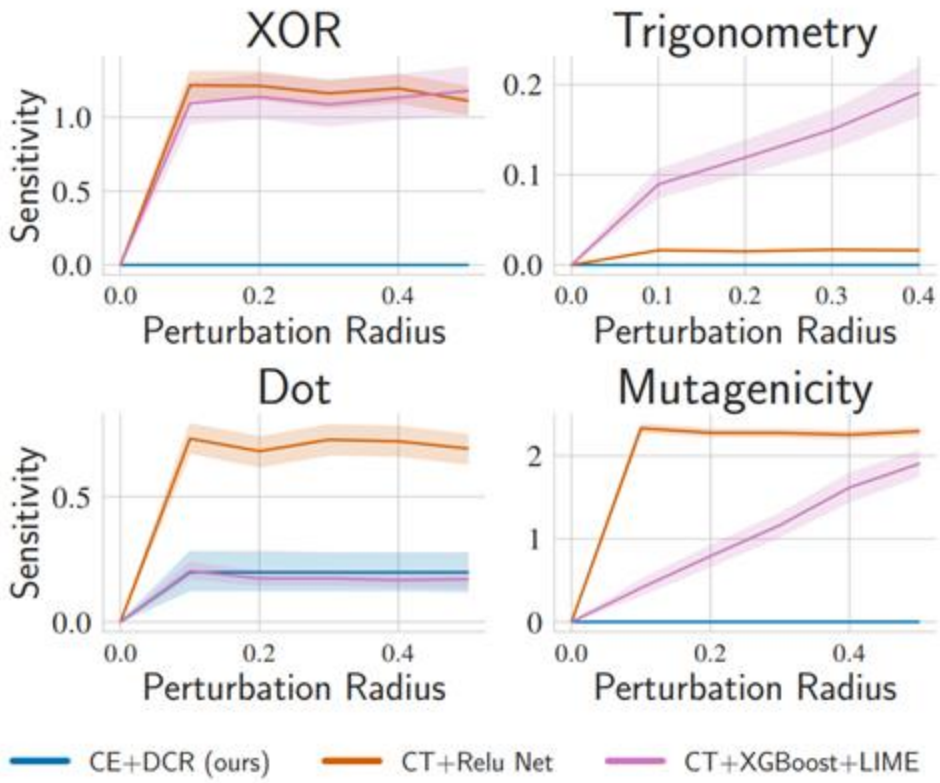


Are DCR Rules Semantically Meaningful?

Are DCR Rules Semantically Meaningful?

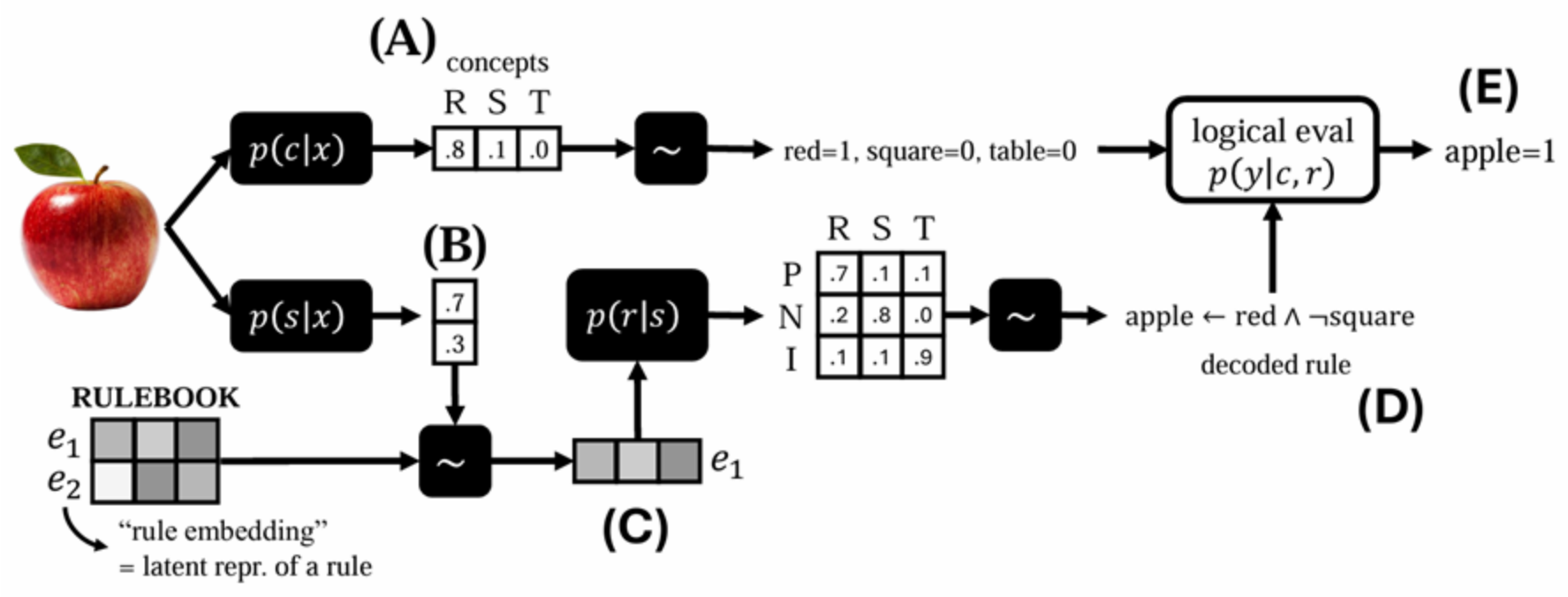
GROUND-TRUTH RULE	PREDICTED RULE	ERROR (%)
	XOR	
$y_0 \leftarrow \neg c_0 \wedge \neg c_1$	$y_0 \leftarrow \neg c_0 \wedge \neg c_1$	0.00 ± 0.00
$y_0 \leftarrow c_0 \wedge c_1$	$y_0 \leftarrow c_0 \wedge c_1$	0.00 ± 0.00
$y_1 \leftarrow \neg c_0 \wedge c_1$	$y_1 \leftarrow \neg c_0 \wedge c_1$	0.02 ± 0.02
$y_1 \leftarrow c_0 \wedge \neg c_1$	$y_1 \leftarrow c_0 \wedge \neg c_1$	0.01 ± 0.01
	Trigonometry	
$y_0 \leftarrow \neg c_0 \wedge \neg c_1 \wedge \neg c_2$	$y_0 \leftarrow \neg c_0 \wedge \neg c_1 \wedge \neg c_2$	0.00 ± 0.00
$y_1 \leftarrow c_0 \wedge c_1 \wedge c_2$	$y_1 \leftarrow c_0 \wedge c_1 \wedge c_2$	0.00 ± 0.00
	MNIST-Addition	
$y_{18} \leftarrow c'_9 \wedge c''_9$	$y_{18} \leftarrow c'_9 \wedge c''_9$	0.00 ± 0.00
$y_{17} \leftarrow c'_9 \wedge c''_8$	$y_{17} \leftarrow c'_9 \wedge c''_8$	0.00 ± 0.00
$y_{17} \leftarrow c'_8 \wedge c''_9$	$y_{17} \leftarrow c'_8 \wedge c''_9$	0.00 ± 0.00

Are rules Stable under perturbation?



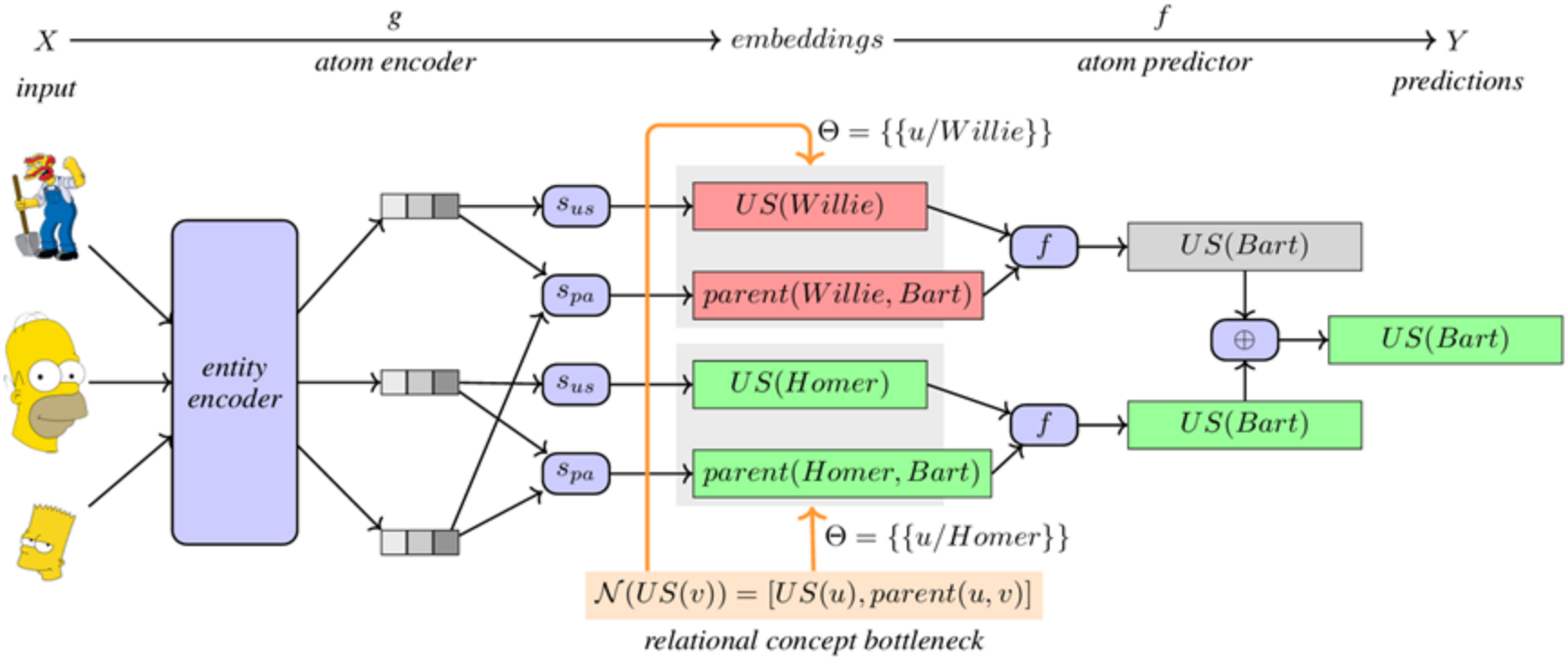
New works

DCR rules now have global validity with the "Rulebook"



Debot, David, et al. "Interpretable Concept-Based Memory Reasoning." NeurIPS 2024.

Relational Concept Bottleneck Models



Barbiero, Pietro, et al. "Relational Concept Bottleneck Models." NeurIPS 2024.

So far we have only studied logic rules…

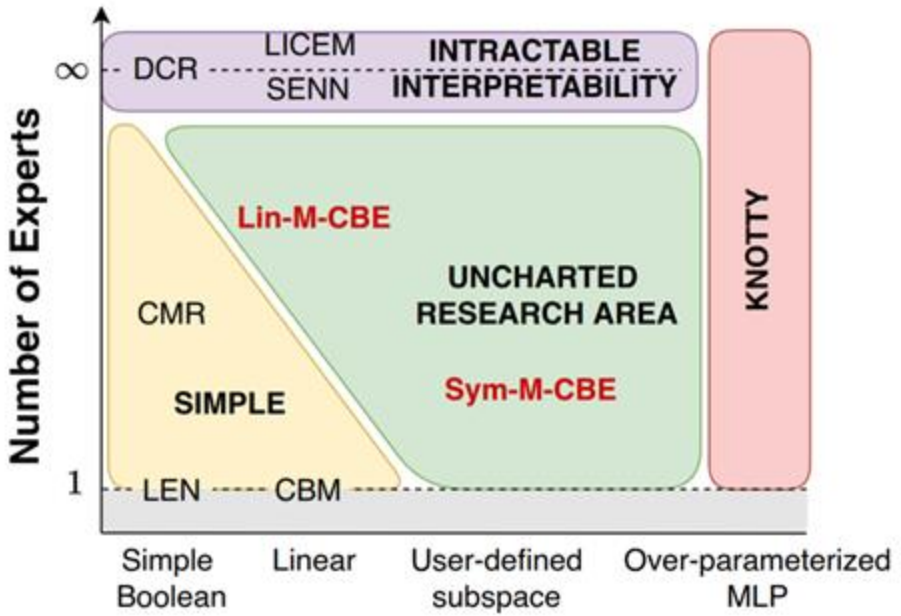
But there are many more functional forms still uncharted.

Varying:

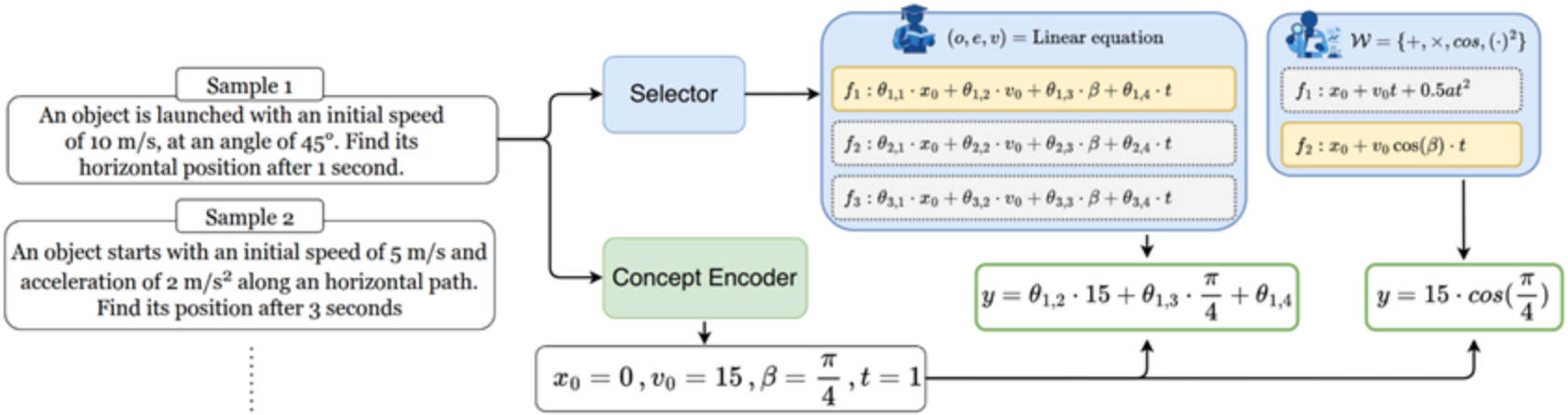
- the number of experts
- functional forms

Find:

- the optimal accuracy-interpretability trade-off



Let's generate any Equations!



Dynamic Selection — Selector routes inputs to specialized mathematical experts

User adaptation — each expert set of operators is defined by the user according to its expertise

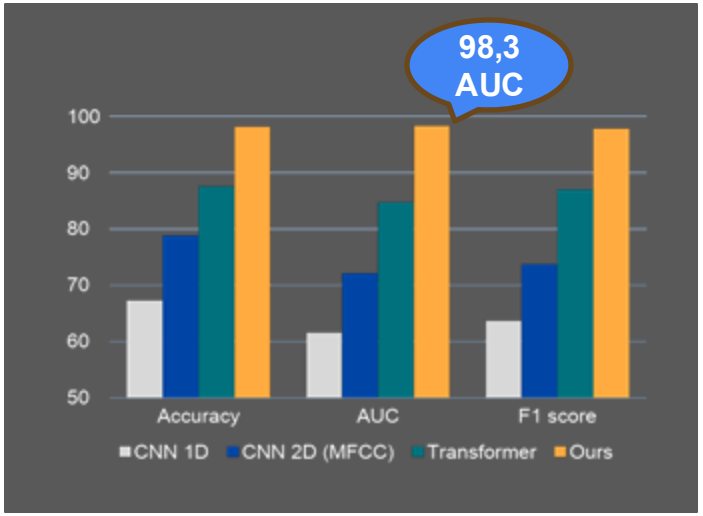
Flexible Forms — Supports both discrete classification and continuous regression tasks

De Santis, et al. "Mixture of Concept Bottleneck Experts", Spotlight at ICML 2026

INTESA SANPAOLO INNOVATION CENTER

Other works

AI4Voice: Non-invasive voice disorder analysis



TRANSFORMERS REACH HIGH ACCURACY
WITHOUT PERFORMING INVASIVE ANALYSIS

Koudounas, Alkis, et al. "Voice Disorder Analysis: a Transformer-based Approach." *INTERSPEECH*. ISCA, 2024.

INTESA SANPAOLO
INNOVATION CENTER

Conclusions

Why Concept-based Explanations?

Why Concept-based Explanations?

1. They **better resemble** the way humans reason and explain [5]

[5] Kim, Sunnie SY, et al. "Help Me Help the AI [...]" (2023)

Why Concept-based Explanations?

1. They **better resemble** the way humans reason and explain [5]
2. They are **more stable** under perturbation [6]

[5] Kim, Sunnie SY, et al. "Help Me Help the AI [...]" (2023)
[6] Alvarez Melis, et al. "Towards robust interpretability with SENN." (2018)

Why Concept-based Explanations?

1. They **better resemble** the way humans reason and explain [5]
2. They are **more stable** under perturbation [6]
3. They better **detect** model **biases** [7]

[5] Kim, Sunnie SY, et al. "Help Me Help the AI [...]" (2023)
[6] Alvarez Melis, et al. "Towards robust interpretability with SENN." (2018)
[7] Jain, Rishabh, et al. "Extending Logic Explained Networks to Text." (2022)

Why Concept-based Explanations?

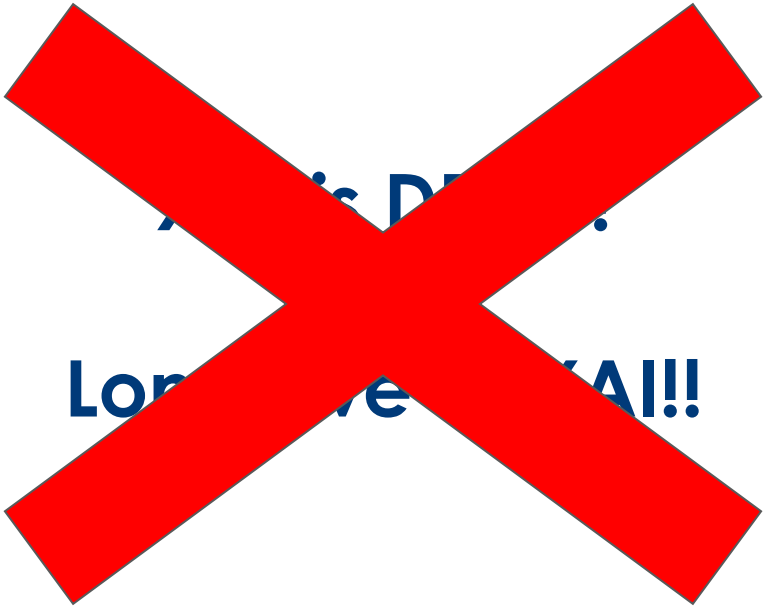
1. They **better resemble** the way humans reason and explain [5]
2. They are **more stable** under perturbation [6]
3. They better **detect** model **biases** [7]
4. They allow creating **more robust** models [8]

[5] Kim, Sunnie SY, et al. "Help Me Help the AI [...]" (2023)
[6] Alvarez Melis, et al. "Towards robust interpretability with SENN." (2018)
[7] Jain, Rishabh, et al. "Extending Logic Explained Networks to Text." (2022)
[8] Ciravegna, Gabriele, et al. "Logic explained networks." (2023)

XAI is DEAD!

XAI is DEAD!

Long Live C-XAI!!



INTESA SANPAOLO
INNOVATION CENTER

(Real) conclusions

- XAI is good for model debugging
  - It shows experts where the network is looking

(Real) conclusions

- XAI is good for model debugging
  - It shows experts where the network is looking

- C-XAI allows widening the audience
  - Employs higher-level symbols instead of raw features
  - Non-experts can understand the model decision process too

(Real) conclusions (2)

- Learning Concept highly enriches network interpretability & interactivity

(Real) conclusions (2)

- Learning Concept highly enriches network interpretability & interactivity

- Concept Embedding avoids performance trade-offs

(Real) conclusions (2)

- Learning Concept highly enriches network interpretability & interactivity
- Concept Embedding avoids performance trade-offs
- DCR/CMR allow interpretable task predictions

Acknowledgments

Eleonora Poeta

Eliana Pastor

Tania Cerquitelli

Elena Baralis

INTESA SANPAOLO
INNOVATION CENTER

Acknowledgments (2)

Pietro Barbiero

Francesco Giannini

Mateo Espinoza Zarlenga

Giuseppe Marra

INTESA SANPAOLO INNOVATION CENTER

References

Concept-based Explainable Artificial Intelligence: A Survey

Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off

Interpretable Neural-Symbolic Concept Reasoning

SURVEY

CEM

DCR

INTESA SANPAOLO INNOVATION CENTER

Thanks for your attention!

Any question?

Future work

1. Textual concepts could also be employed in post-hoc models
2. Improve counterfactual examples to assess model causality
3. Concept-based models could employ multi-modal representations
4. LLMs could be employed to create other type of concepts (not necessarily textual) for other type of models (not necessarily vision)

Applications

1. Generative Models
2. Out-of-Distributions Detectors
3. Medicine (E.g., CBM for X-RAY)
4. Education
5. Autonomous Driving?

INTESA SANPAOLO
INNOVATION CENTER