



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**.



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?**

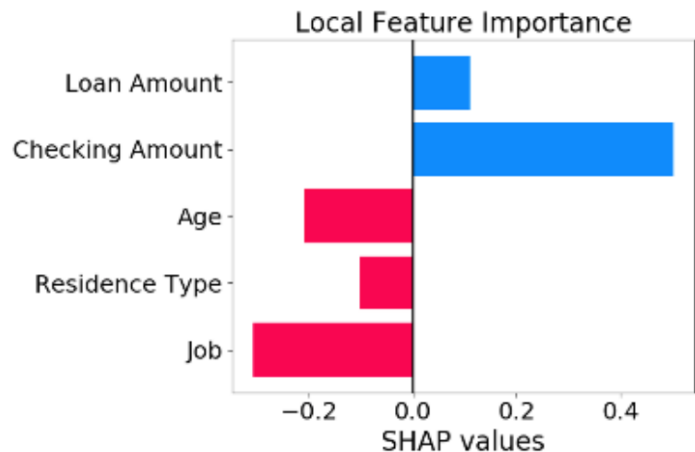
Outline

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

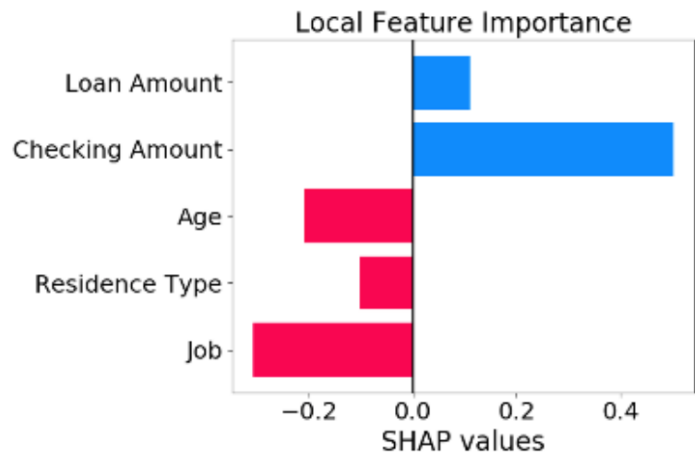
Explanations depends on the input data

Tabular data



Explanations depends on the input data

Tabular data



Structured data

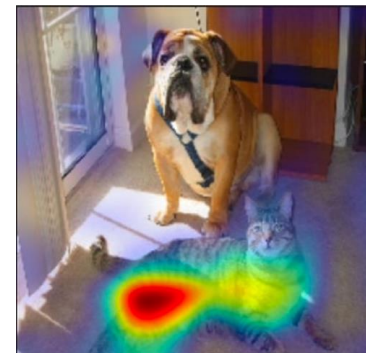
- Each feature has a clear **semantic meaning**
- From feature importance we can easily define a **counterfactual example**
- **Global explanations** are meaningful as features keep the same meaning across different samples

Explanations depends on the input data

Textual data

Successfully got married and went on my honeymoon with no major panic attacks. Hi reddit, I have found comfort in this community of fellow anxious people so I thought I'd share an update on my anxiety. I've had my entire life and it has become debilitating at times. It makes me scared to do things I want to do, such as travel. This year is a big year for me - I just got married last Saturday and I just got back from my honeymoon in California. Those two things alone were enough to keep me up at night anxiously worrying months before. BUT - I did it. It was a perfect trip and I'm so glad I didn't let my anxiety run it. I'm gonna list a few things that have helped me immensely and allowed me to enjoy my time this week. First, I went to my doctor and got a prescription for 10 0.5mg Klonopin pills. My was to use them for times of extreme anxiety (flying) only. These were very helpful to me. I was really anxious on my wedding morning and almost couldn't get out of my car to go to the hair salon, so I popped half a pill. About an hour into getting my hair and makeup done, I realized I was fine. The rest of the day was absolutely beautiful and I had no social anxiety and didn't feel any panic. I also took a whole pill the morning of my flight to California. I was perfectly relaxed on the plane and was pleasantly sleepy. I have no intentions of taking these pills on a regular day - just when I'm doing something I know will send me into a panic. Secondly is just simply not letting my anxiety stop me from doing things. For example, I was very anxious to get on our first flight and almost found myself saying "I can't do this." Had I actually not done it and gone home, I would have been even more scared of flying in the future. But, I did it anyway and I'm so glad I did. Once I was in the air, I turned on The Office and played games on my phone and before I knew it, we were landing. Now I'm excited to fly again and am already planning my next trip. I truly believe that exposure is helping my anxiety. Third, I have been taking CBD oil (both in oil form and through a vape pen) consistently. This has decreased my daily anxiety to a much lesser amount. Fourth, I cut out coffee for the most part and try to stay away from refined sugar. I noticed a correlation between my anxiety levels and what I was eating/drinking, so I just started cutting things out of my diet. I still eat sugar and drink an occasional latte, but it doesn't give me the same anxiety as when I was eating/drinking these things daily. I just really wanted to post and let people know that while it may seem like a never-ending battle, anxiety can definitely be helped. I am slowly getting better and getting to know myself in the process. Much love to my fellow anxiety-ridden folks.

Image data



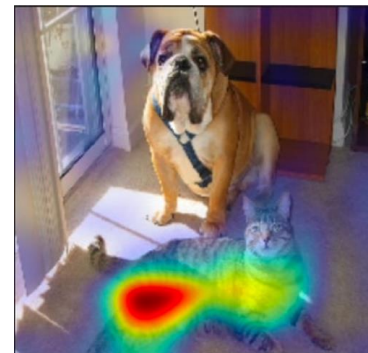
Explanations depends on the input data

Unstructured data

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

Textual data

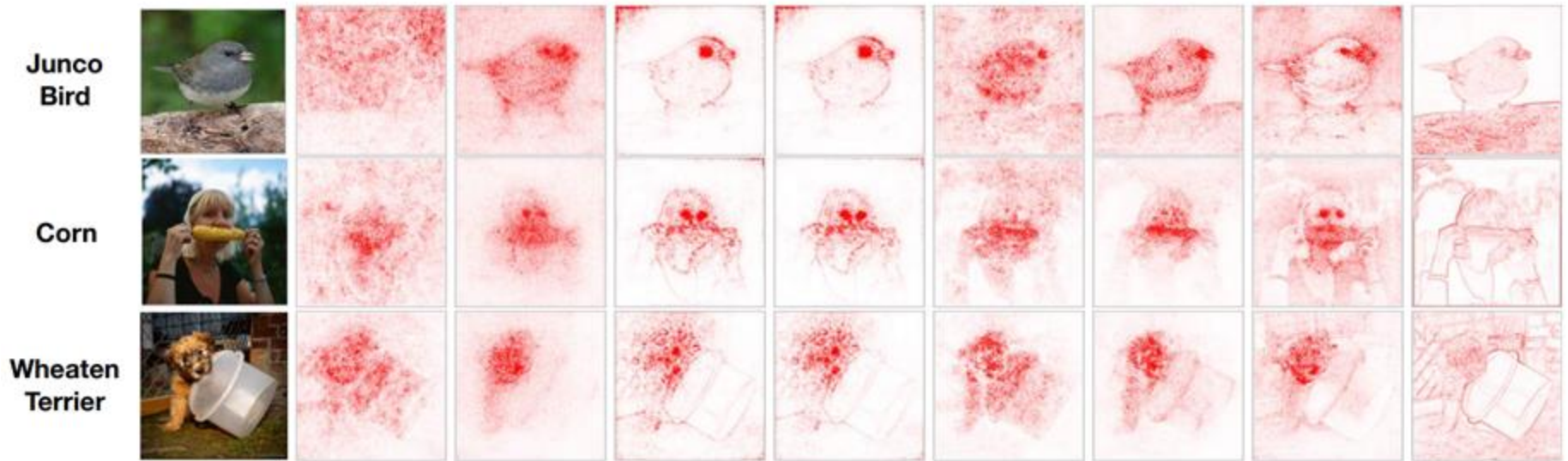
Successfully got married and went on my honeymoon with no major **panic attacks**! Hi reddit, I have found **comfort** in this community of **anxious** people so I thought I'd share an update on my **anxiety**. I've had **anxiety** my entire **life** and it **has become** debilitating at times. It makes me **scared** to do things I want to do, such as travel. This year is a big year **for** me - I just got married last Saturday and I just got **back from** my honeymoon in California. Those two things alone **were** enough to **keep** me up at night anxiously worrying **months** before. BUT - I did it. It was a perfect trip and I'm so glad I didn't let my **anxiety** run it. I'm gonna list a few things that have **helped** me immensely and allowed me to enjoy my **time** this week.​\First, I went to my doctor and got a prescription **for** 10 0.5mg Klonopin pills. My **anxiety** was to use them **for** times of extreme **anxiety** (flying) only. These were very **helpful** to me. I was really **anxious** on my wedding morning and almost couldn't **get** out of my car to **go** to the hair salon, so I **popped** **half** a pill. About an hour into getting my hair and makeup done, I realized I was fine. The rest of the day was absolutely beautiful and I **had** no social **anxiety** and don't feel any **panic**. I **also** took a whole pill the morning of my flight to California. I was perfectly relaxed on the plane and was **pleasantly** sleepy. I have no intentions of taking these pills **on** a regular day - just when I'm doing something I know will send me into a **panic**.​\Secondly is just simply not letting my **anxiety** stop me from doing things. **For** example, I was very **anxious** to **get** on our **first** flight and almost found myself saying "I can't **do** this." **Had** I actually not **done** it and gone home, I would have been even more **scared** of flying in the future. But, I did it anyway and I'm so glad I did. Once I was in the **air**, I turned on The Office and played **games** on my **phone** and before I knew it, we were landing. Now I'm excited to **fly** **again** and am **already** planning my next trip. I truly believe that exposure is helping my **anxiety**.​\Third, I have been taking CBD oil (both in oil form and through a vape pen) consistently. This **has** decreased my daily **anxiety** a **significant** amount.​\Fourth, I cut **out** coffee **for** the most part and try to stay away from refined sugar. I noticed a correlation between my **anxiety** levels and what I was eating/drinking, so I just started cutting things **out** of my **diet**. I **still** eat sugar and **drink** an occasional latte, but it doesn't give me the same **anxiety** as when I was eating/drinking these things **daily**.​\I just really wanted to **post** and let people know that while it may seem like a never-ending battle, **anxiety** can definitely be **helped**. I am slowly **getting** better and **getting** to know myself in the **process**. Much love to my **anxious** anxiety-ridden folks.



When considering unstructured data

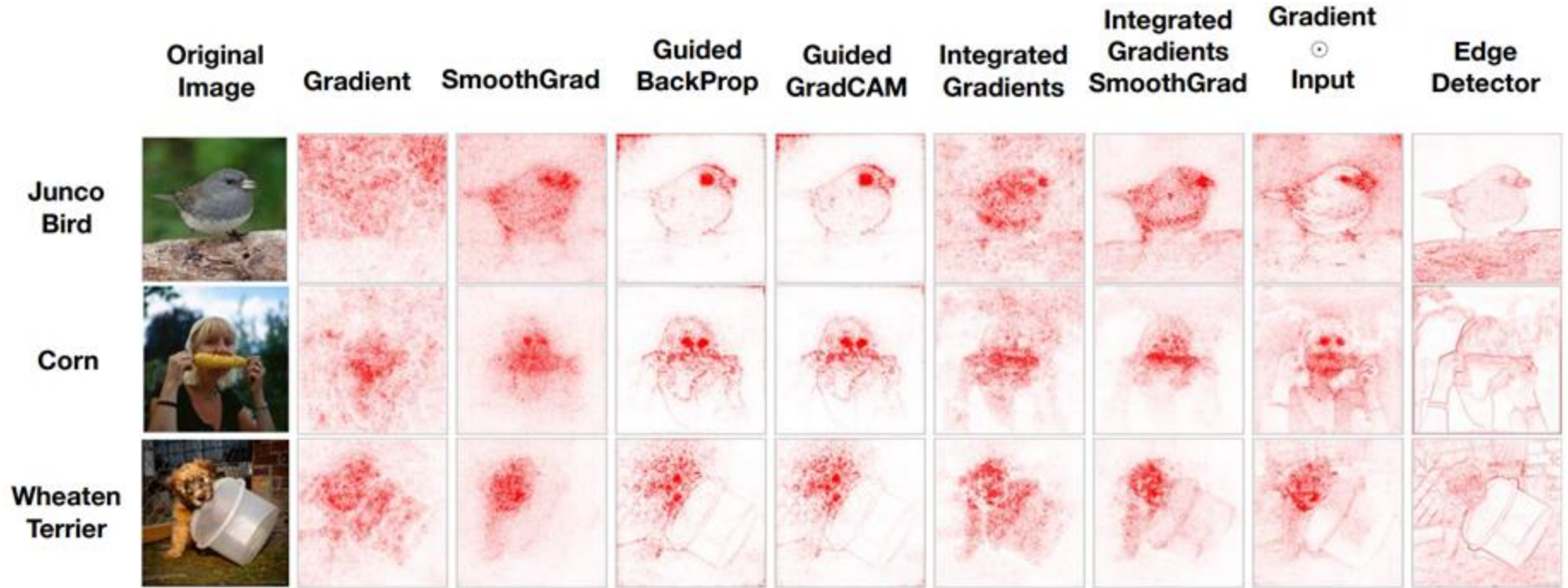
Standard XAI does not always work well

Which method is better? [1]



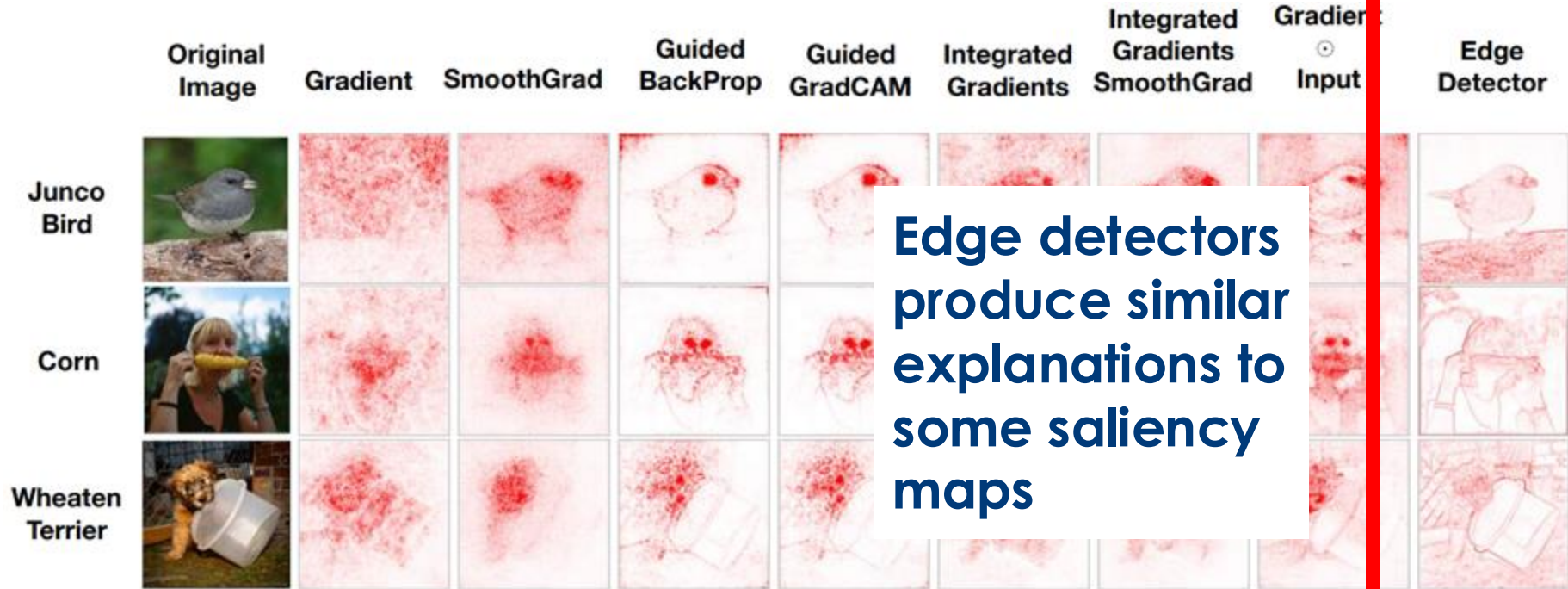
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Which method is better? [1]



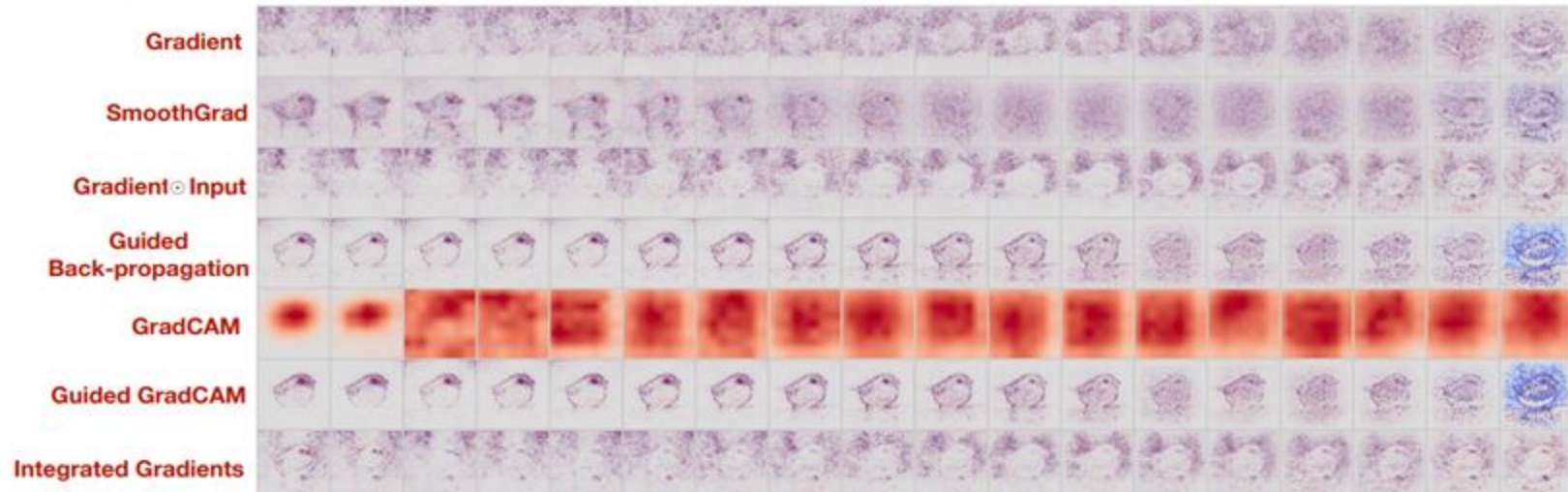
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Which method is better? [1]



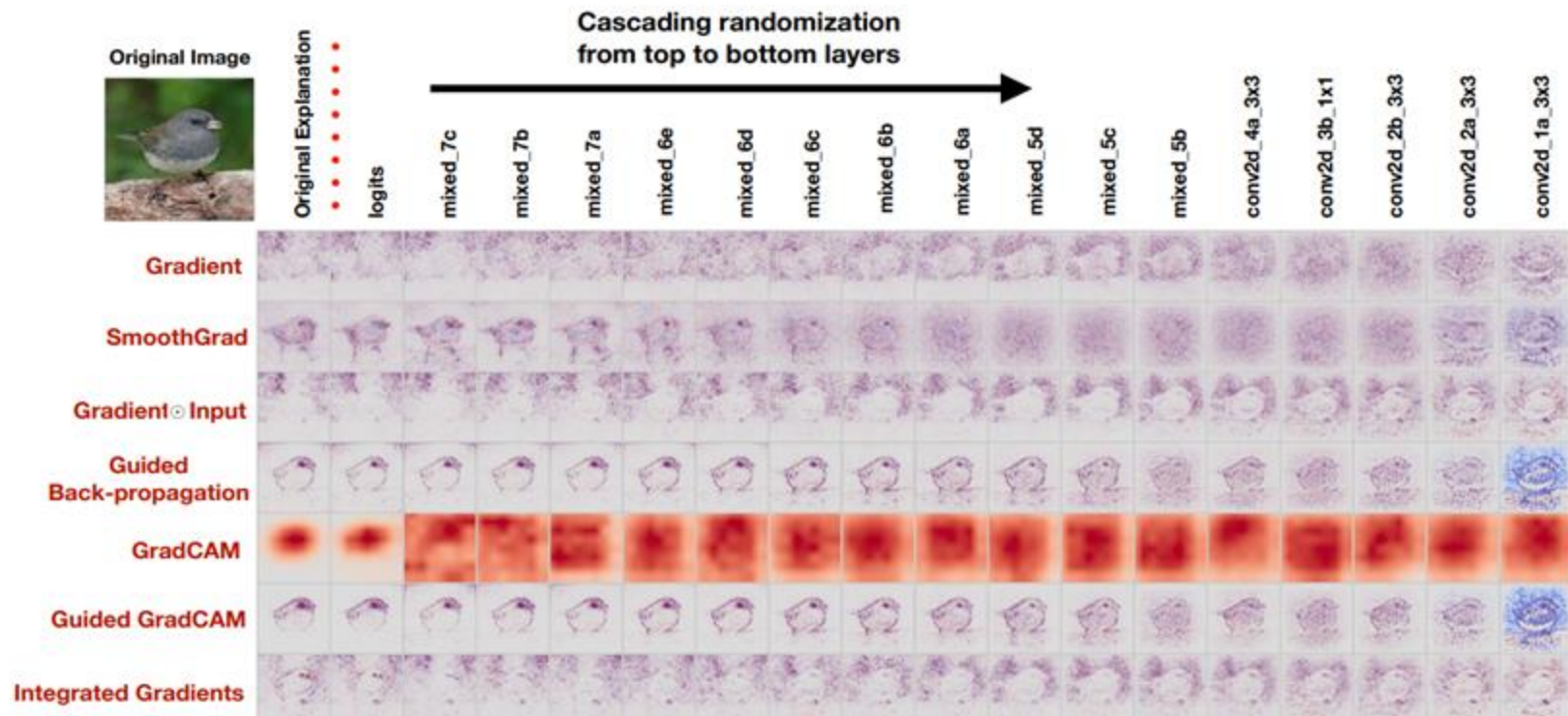
[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.

Towards where are we randomizing? [1]



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

Towards where are we randomizing? [1]

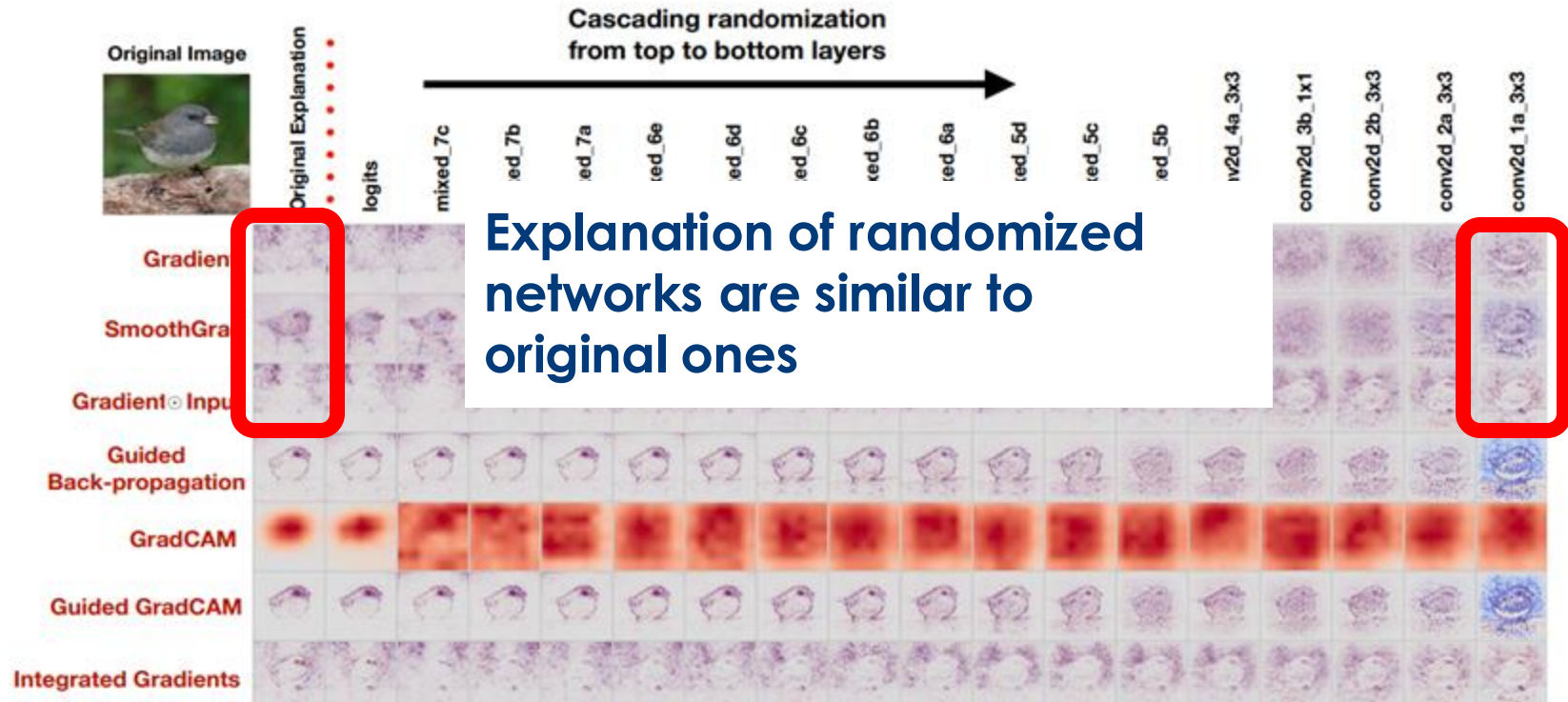


[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.



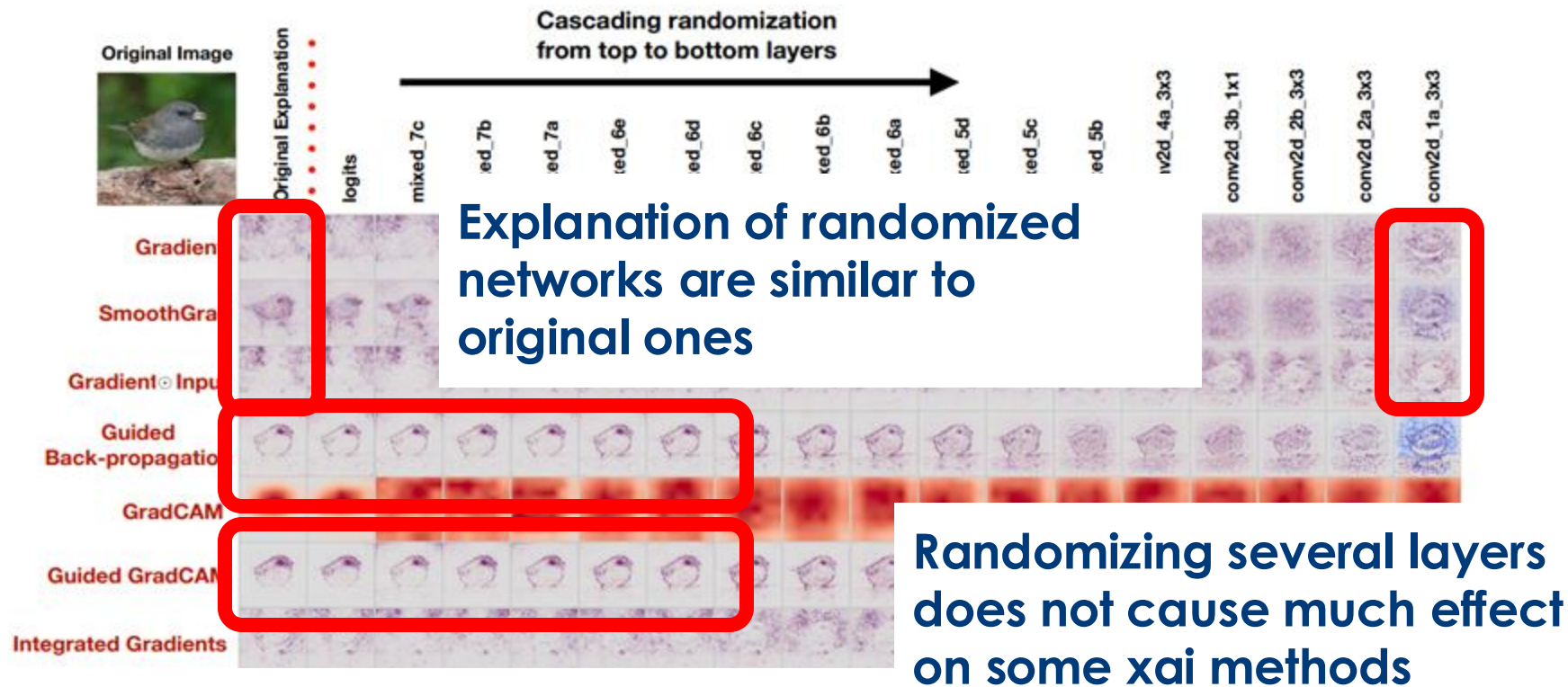
INTESA SANPAOLO
INNOVATION CENTER

Towards where are we randomizing? [1]



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

Towards where are we randomizing? [1]

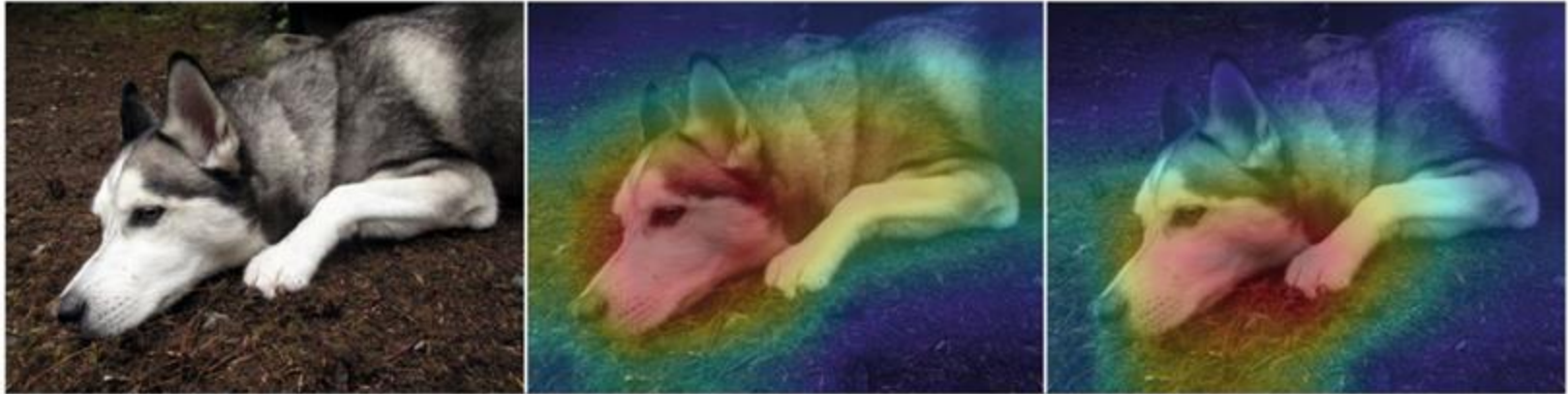


[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Neurips 2018.



Standard xai methods are more input-dependent than model dependent

Which class are we explaining? [3]



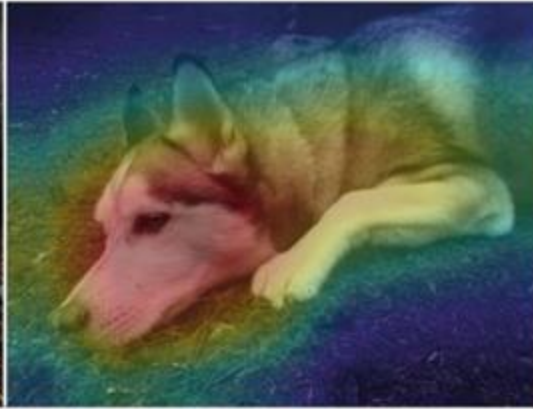
[3] Rudin, Cynthia. "Stop explaining black box machine learning models ..."
Nature machine intelligence (2019)

Which class are we explaining? [3]

“Siberian Husky”



“Transverse Flute”



[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] Ahtibat, 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



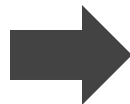
WHAT IS
THIS?



APPLE

PREDICTION

WHY?



	ROUND	
	RED	
	SQUARED	
	COLD	



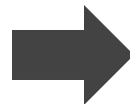
WHAT IS
THIS?



APPLE

PREDICTION

WHY?



	ROUND	
	RED	
	SQUARED	
	COLD	

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)



INTESA SANPAOLO
INNOVATION CENTER

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”



INTESA SANPAOLO
INNOVATION CENTER

Different types of concepts

Symbolic Concepts

Human-defined attributes

REQUIRES FURTHER
ANNOTATIONS

“BEAK”

Unsupervised Concept Basis

Cluster of similar samples

Prototypes

(Part-of) a training sample

CAN BE EXTRACTED
FROM THE DATA



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 → Parrot

Different Types of Explanations too!

1. Class-Concept Relations

Relation among a concept and an output class of a model

Beak → *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 → *Parrot*

2. Node-Concept Association

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



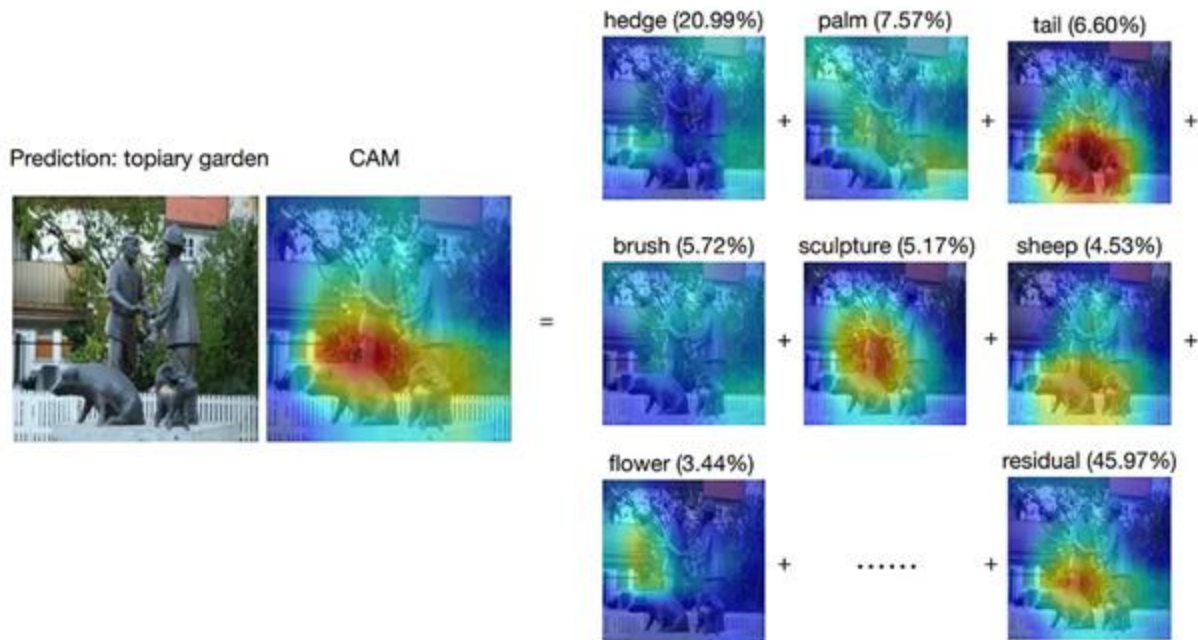
3. Concept-Visualization

Visualization of a learnt concept in terms of the input features



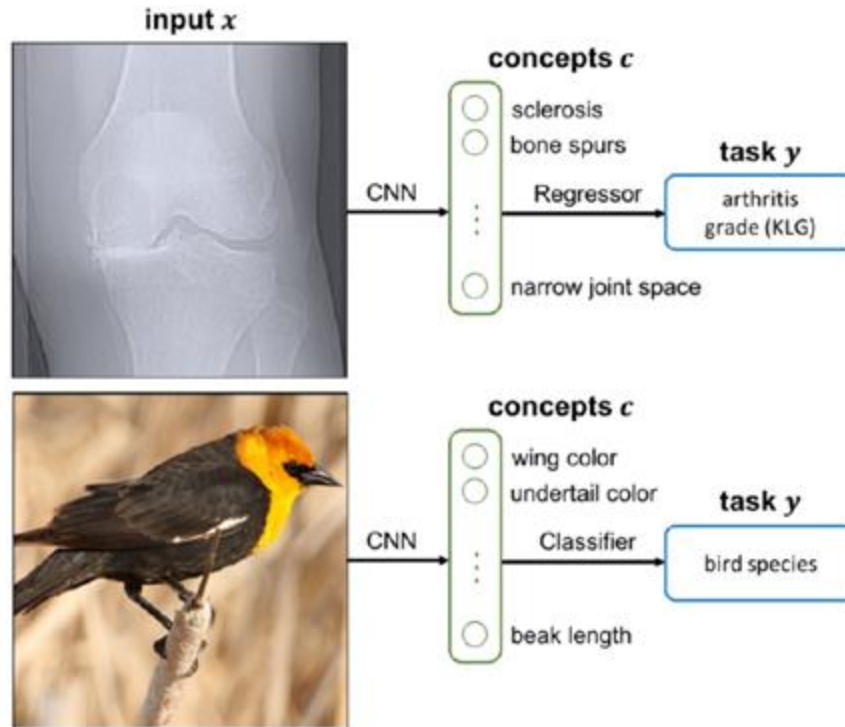
Post-hoc or Explainable-by-design?

Post-hoc Concept-Based Explanations



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

Explainable-by-design Concept-based Models

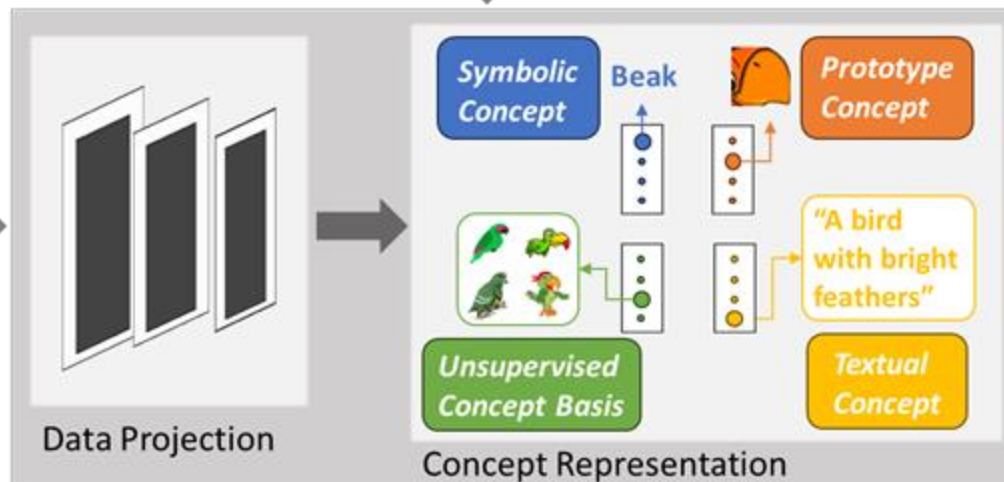


Koh, Pang W, et al. "Concept bottleneck models." *ICML 2020*.

ADDITIONAL
KNOWLEDGE



INPUT



EXPLAINED MODEL/
EXPLAINABLE-BY-DESIGN MODEL

Parrot

Model Prediction

*Class-Concept
Relation*

Beak → Parrot

*Node-Concept
Association*

Beak

*Concept
Visualization*



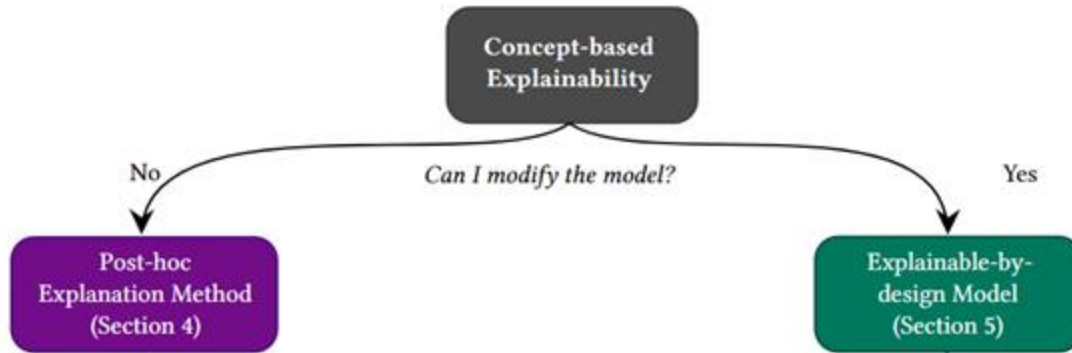
Explanations

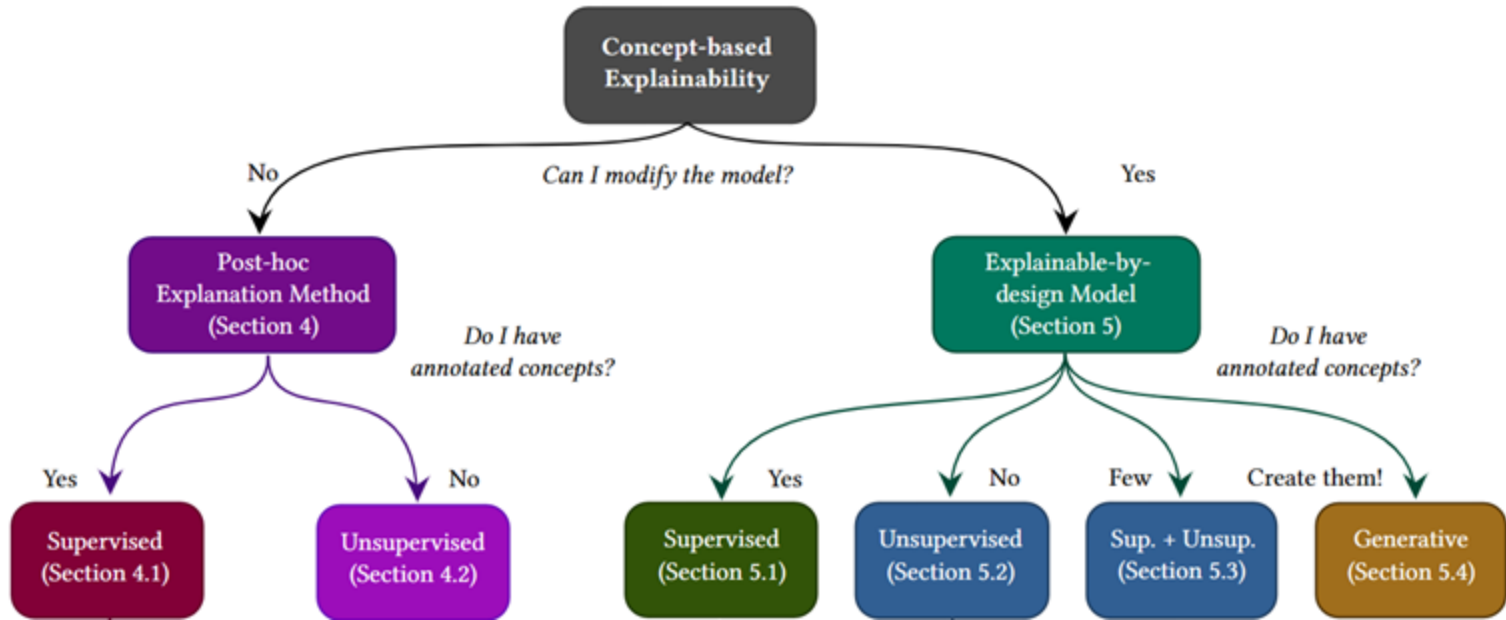
OUTPUT

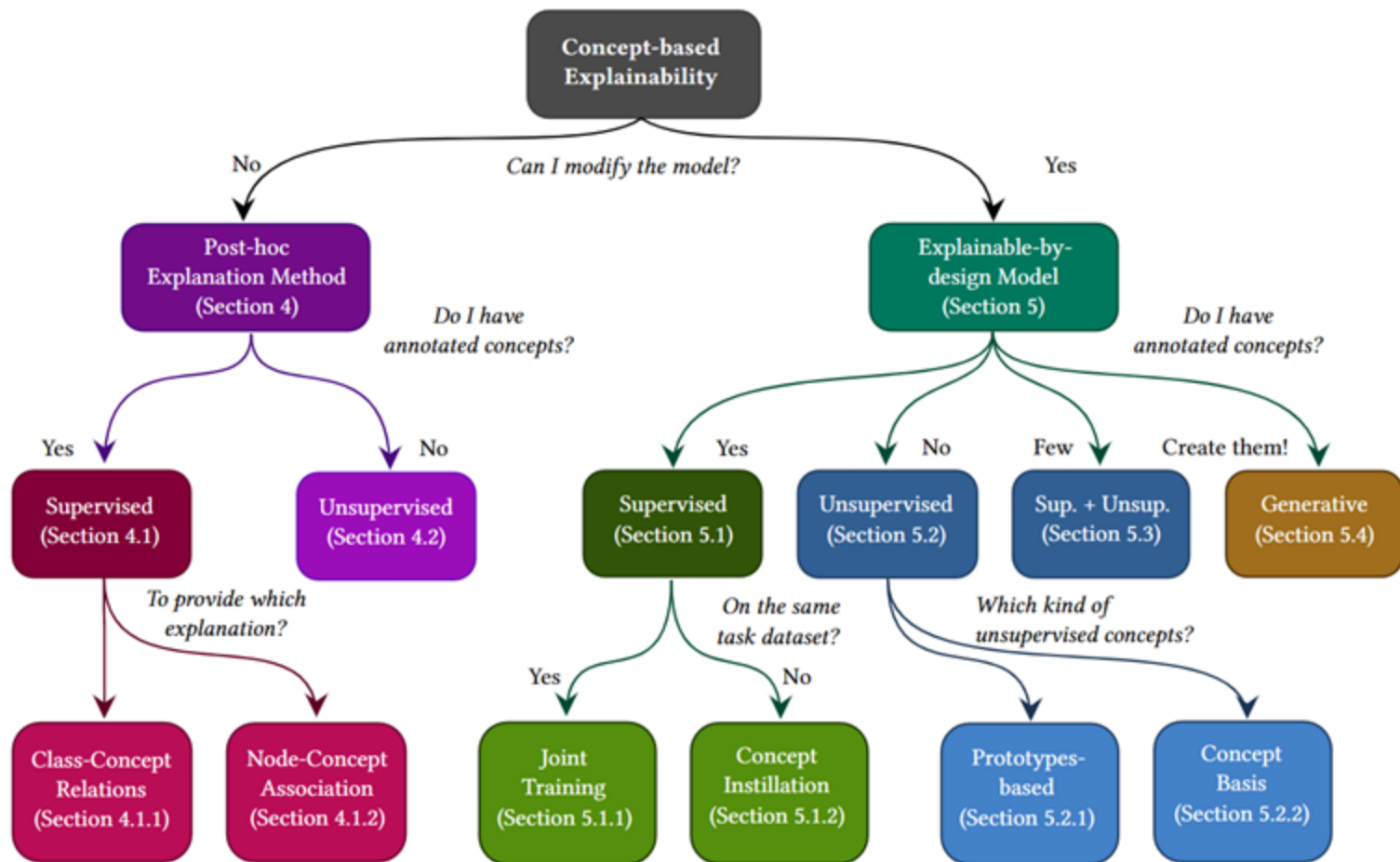


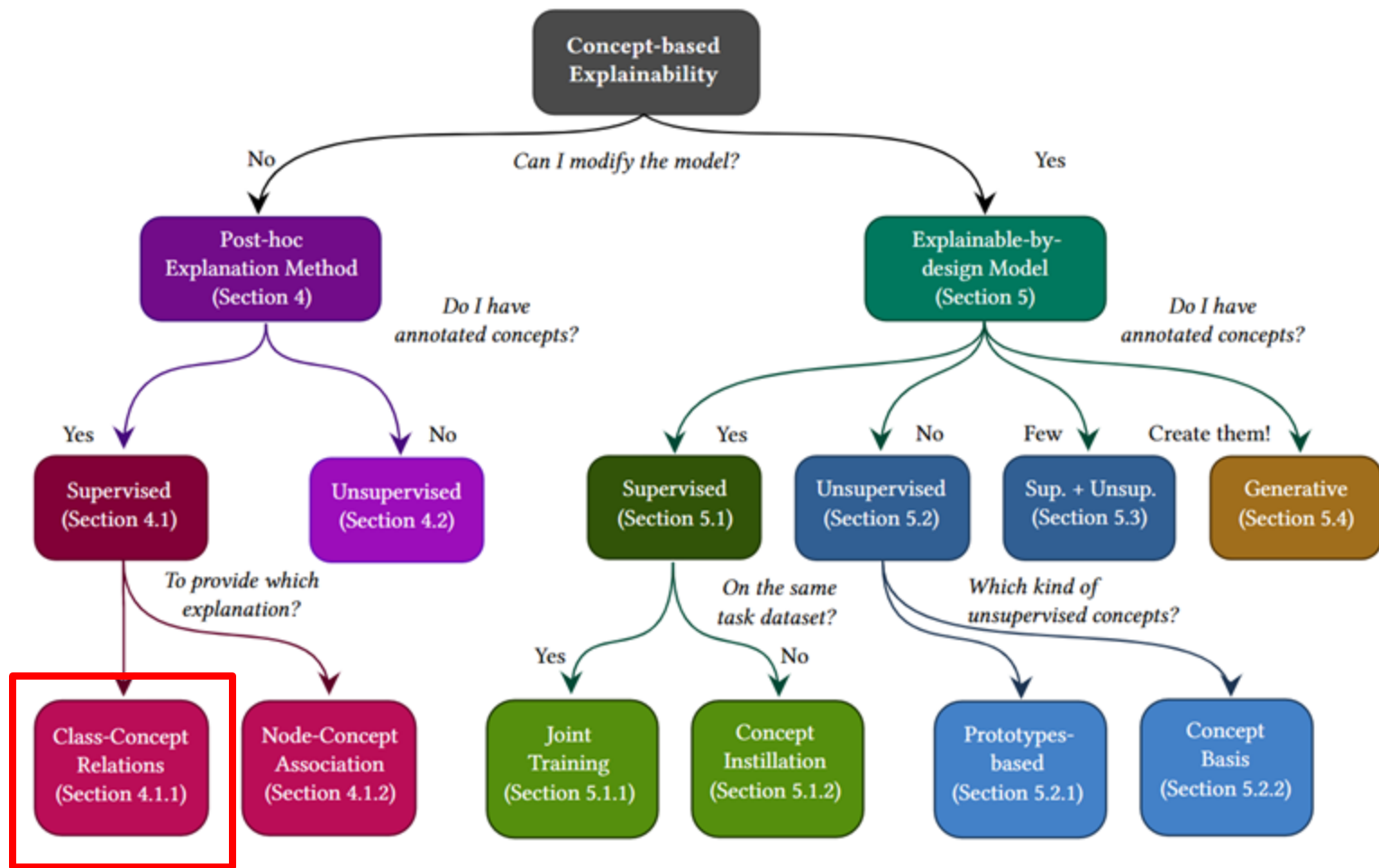
INTESA SANPAOLO
INNOVATION CENTER

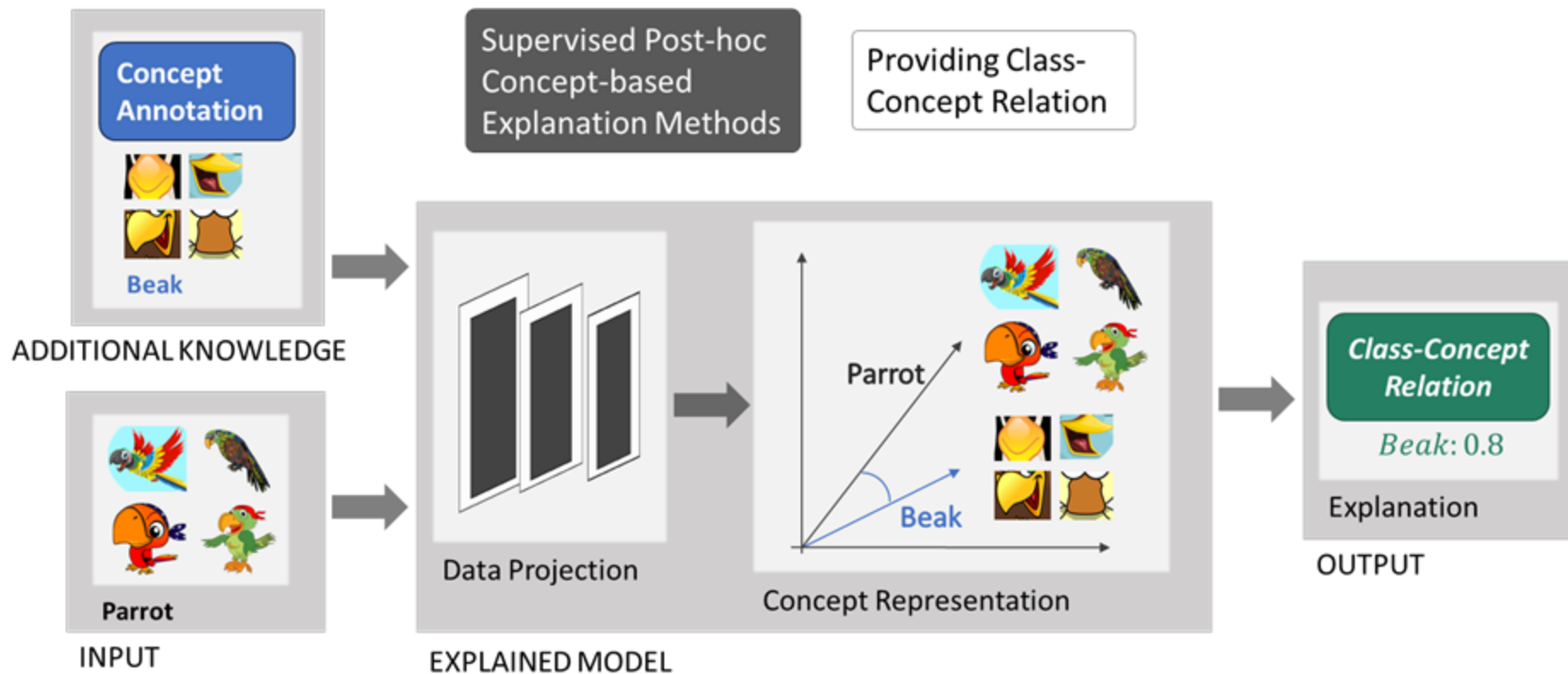
**How do i choose a
C-XAI method?**



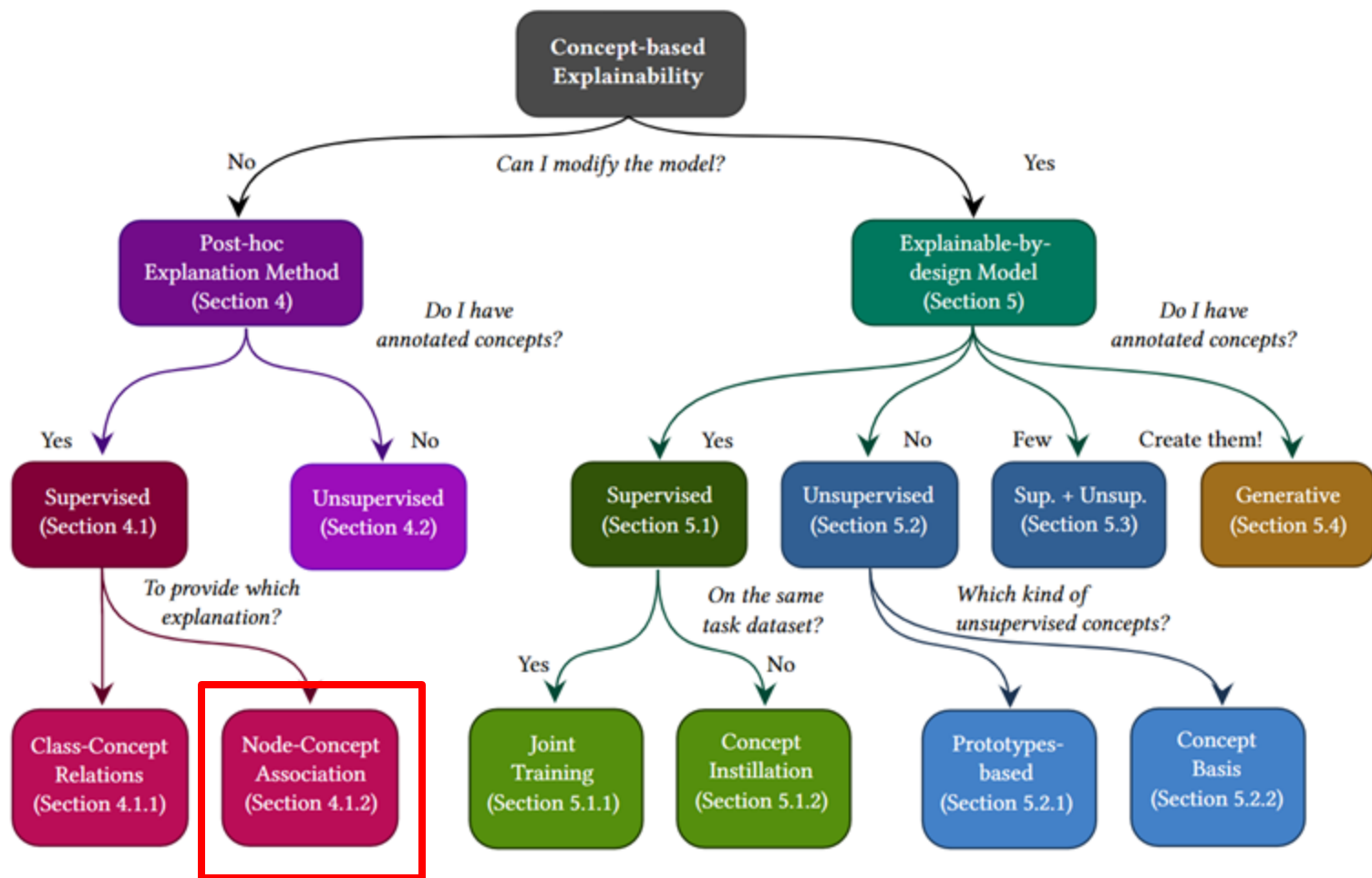


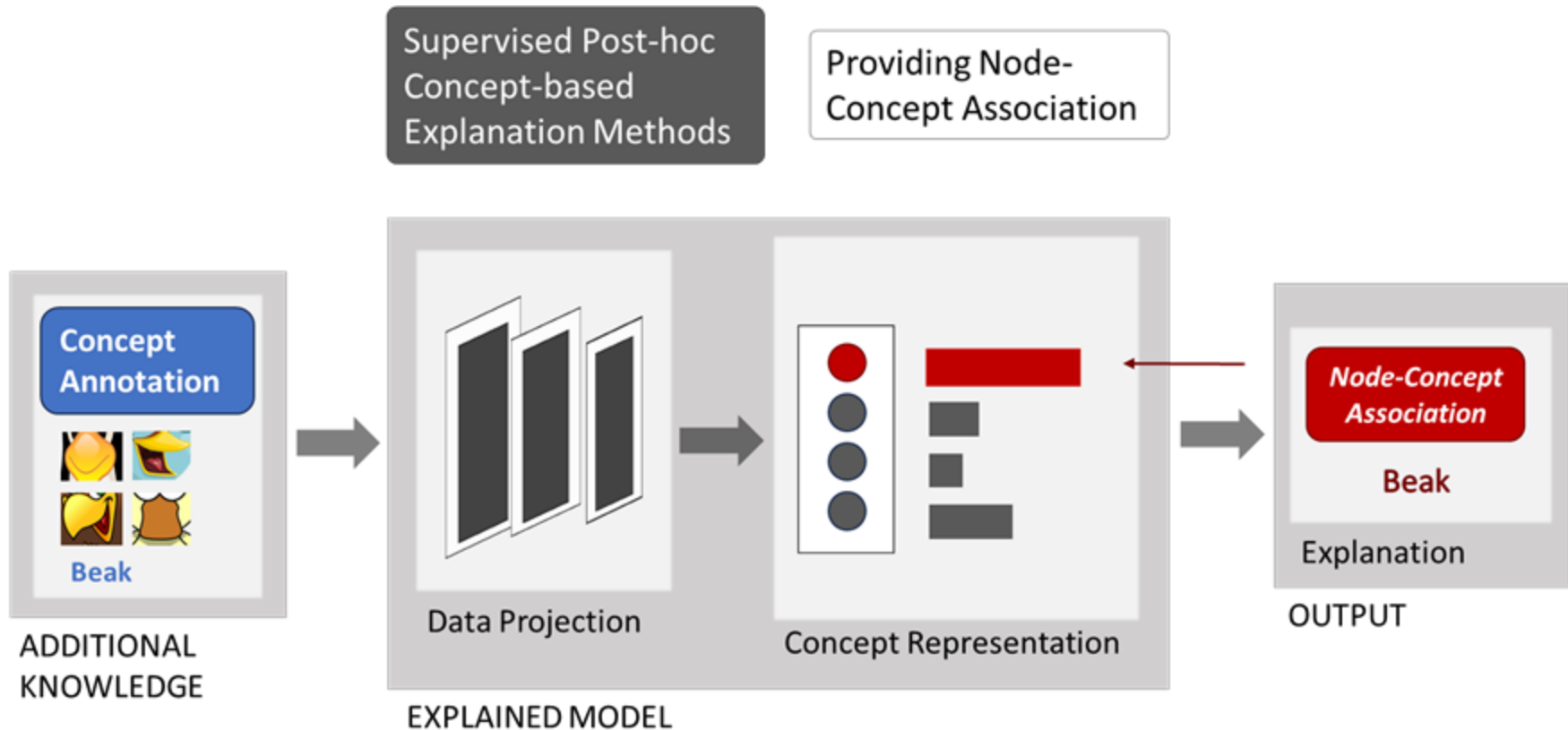




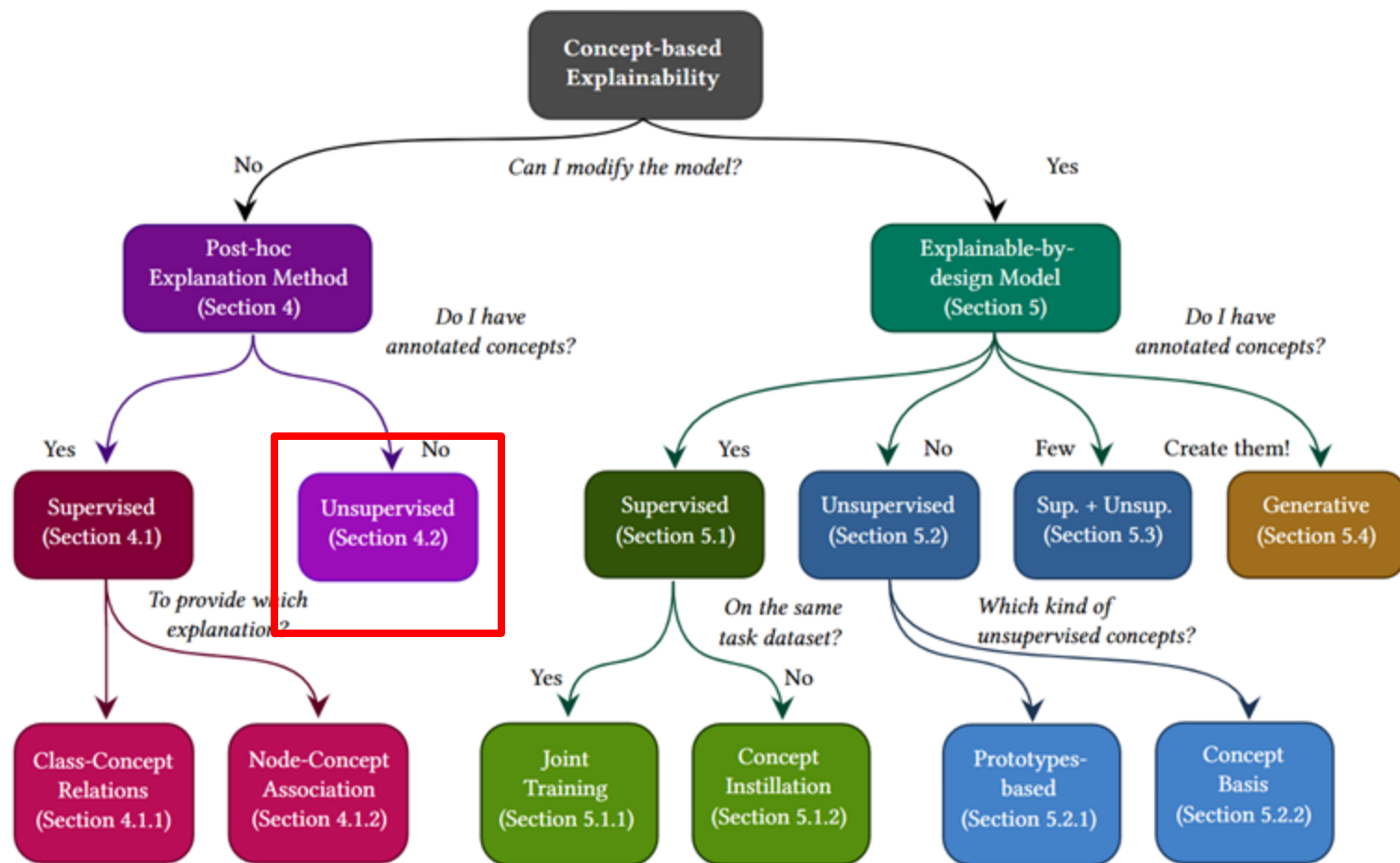


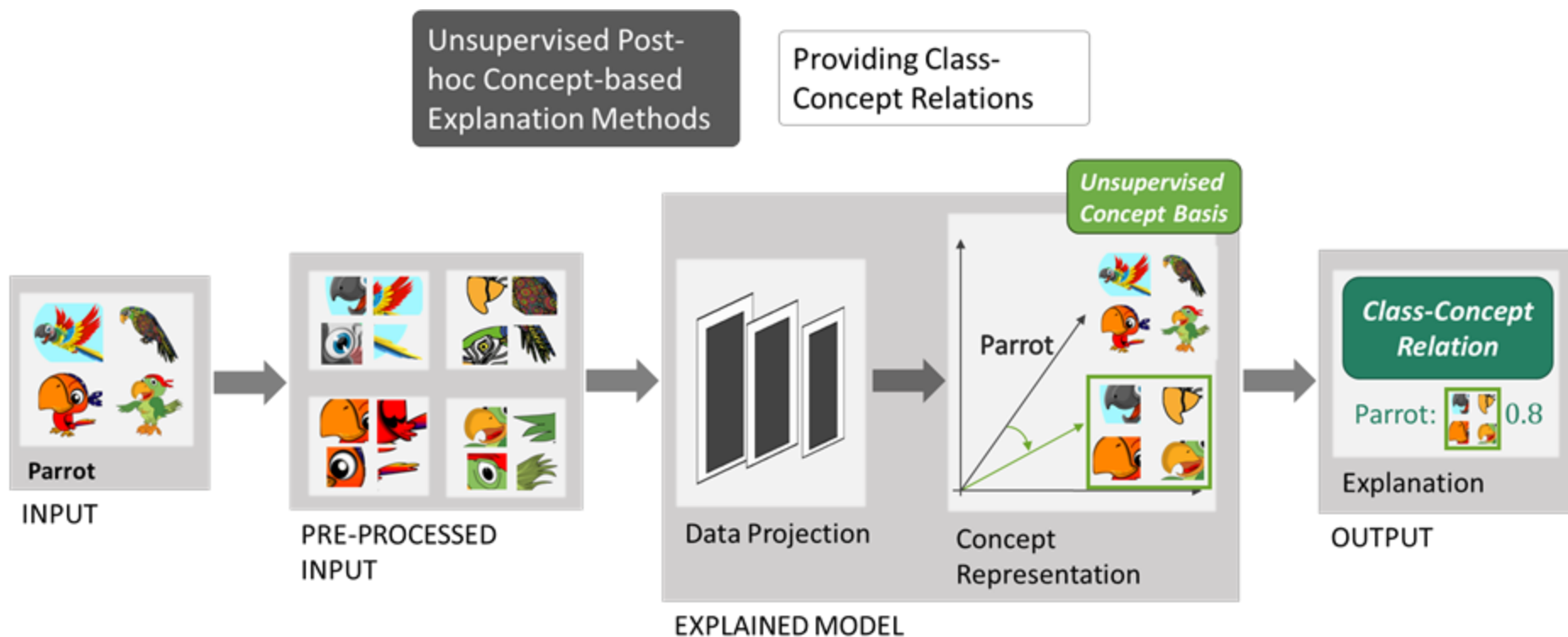
E.g., Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. (2018).



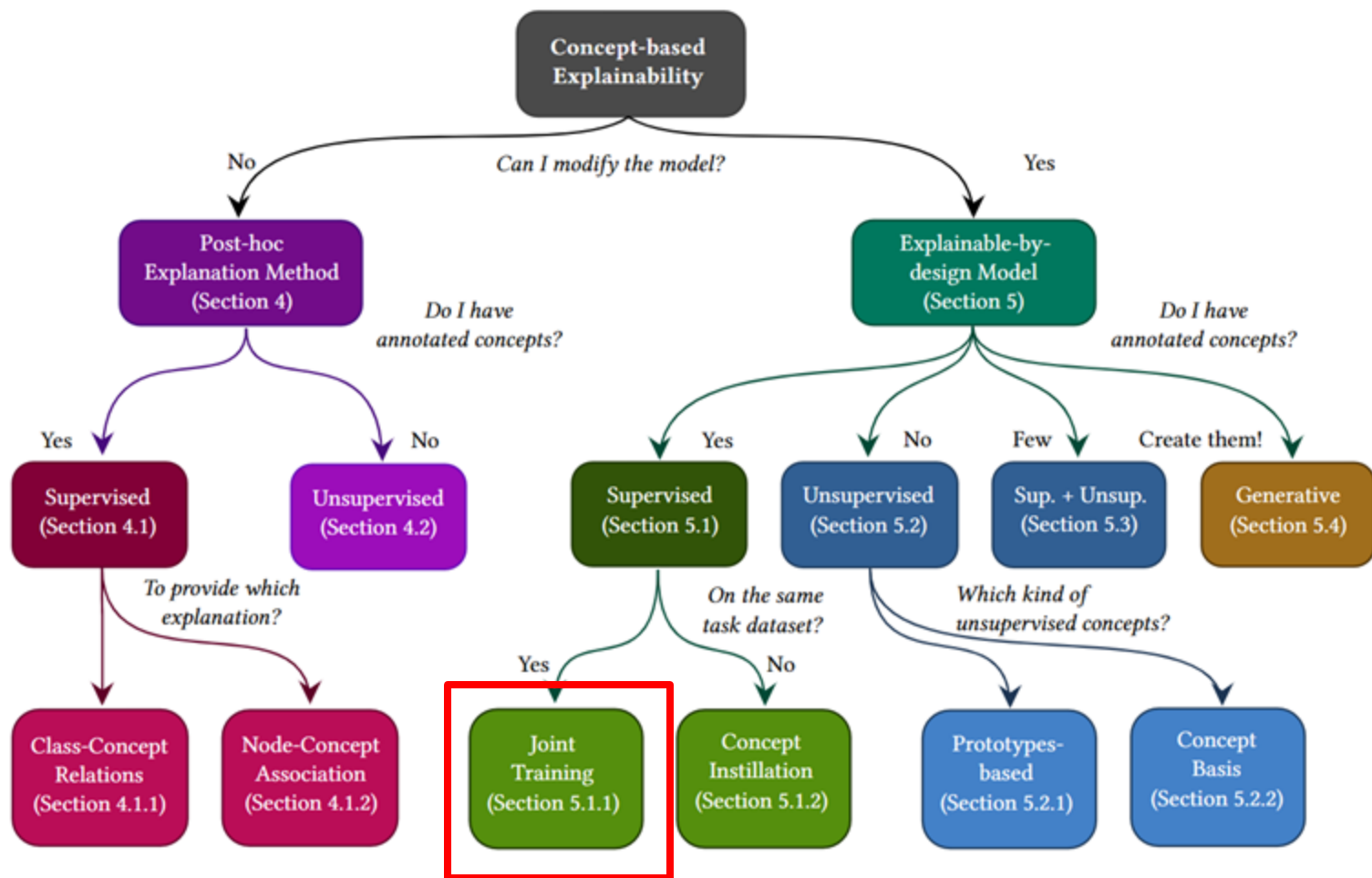


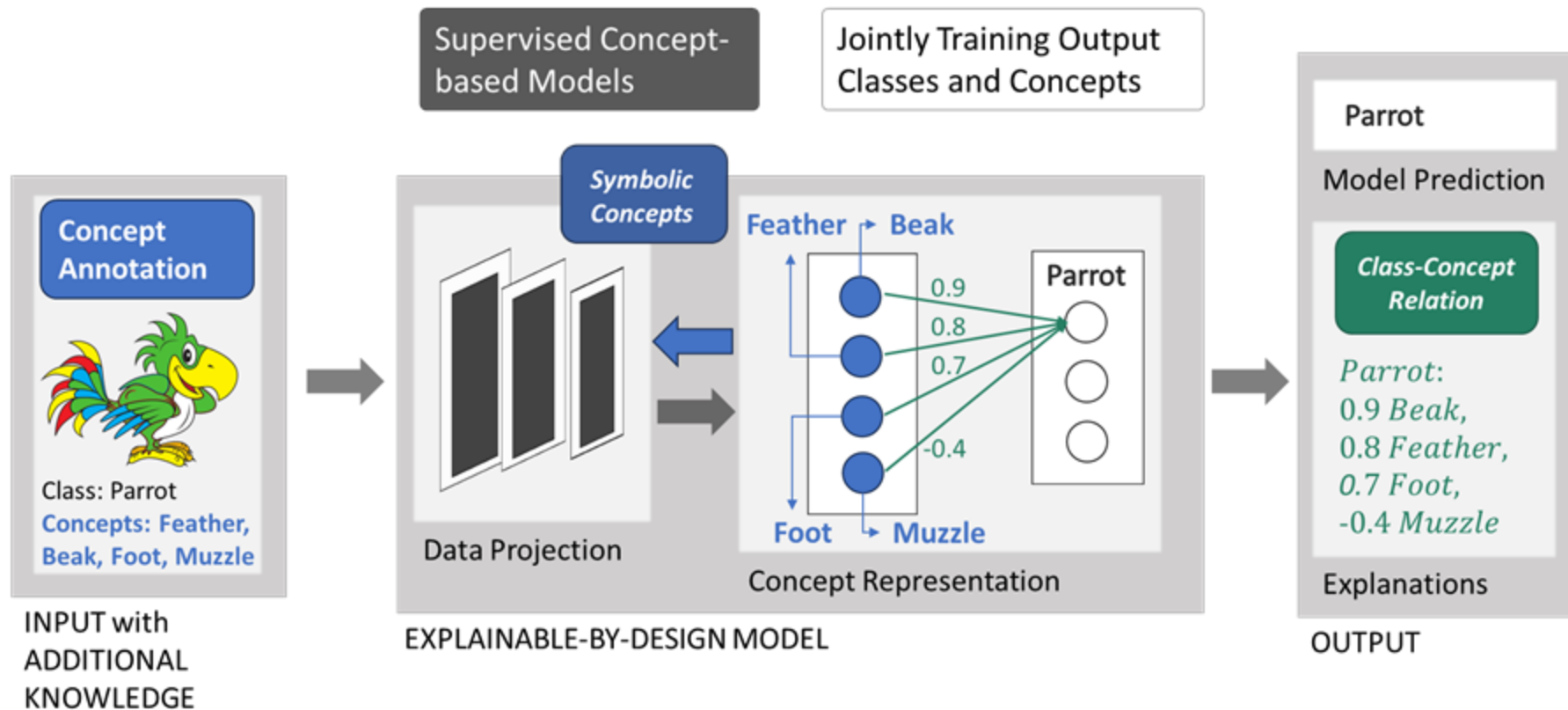
E.g., Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." CVPR. 2017



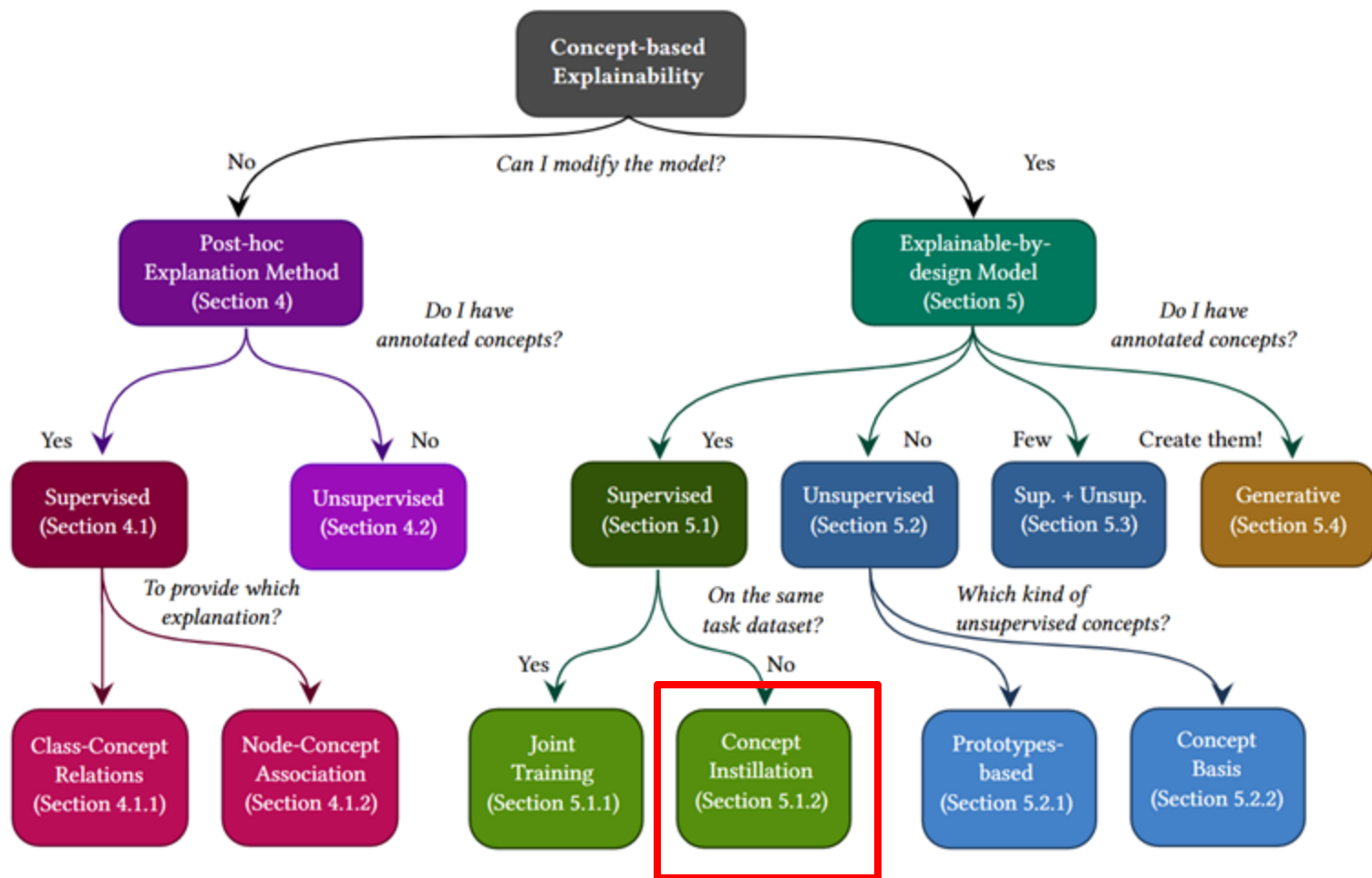


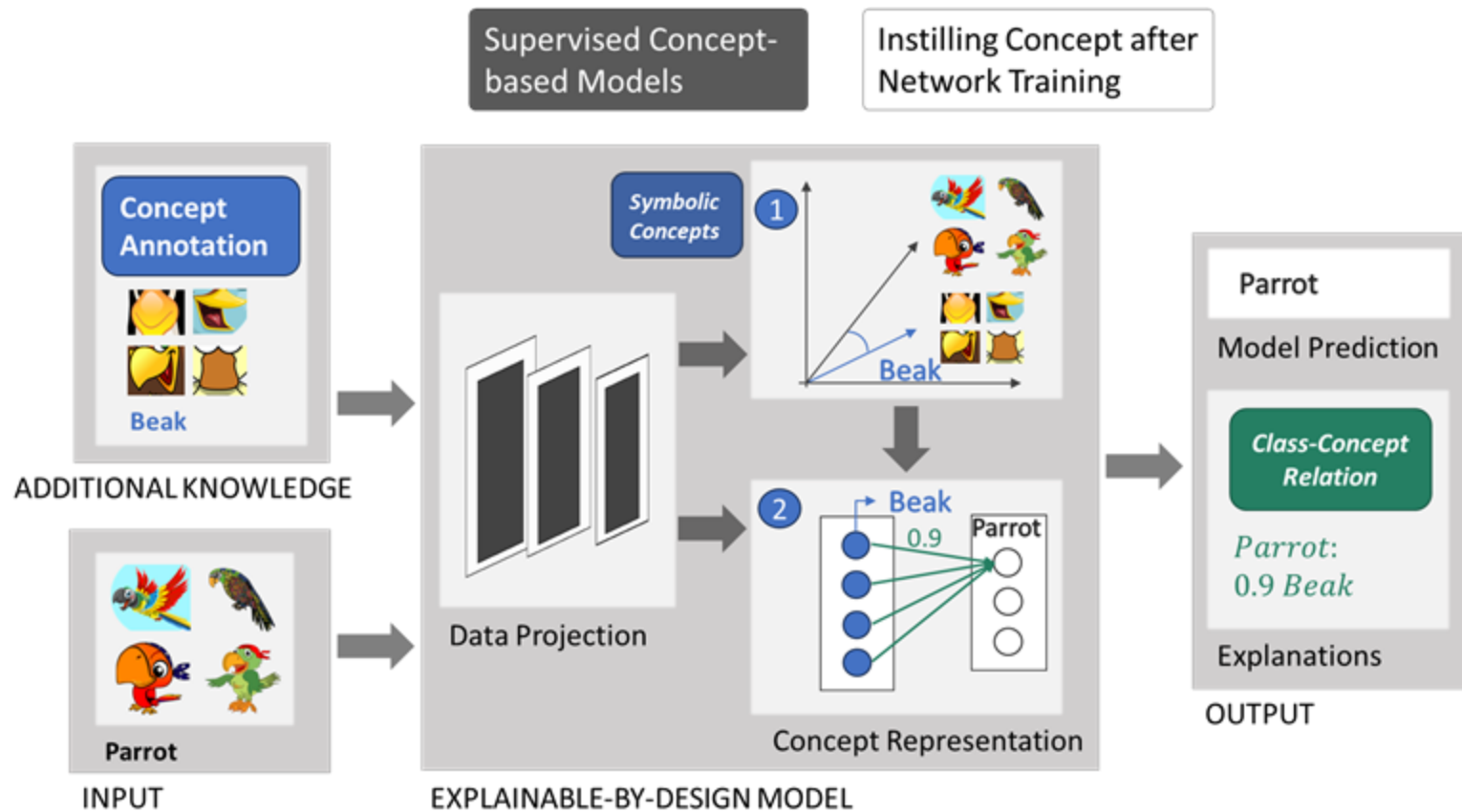
E.g., Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." NeurIPS (2019).





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

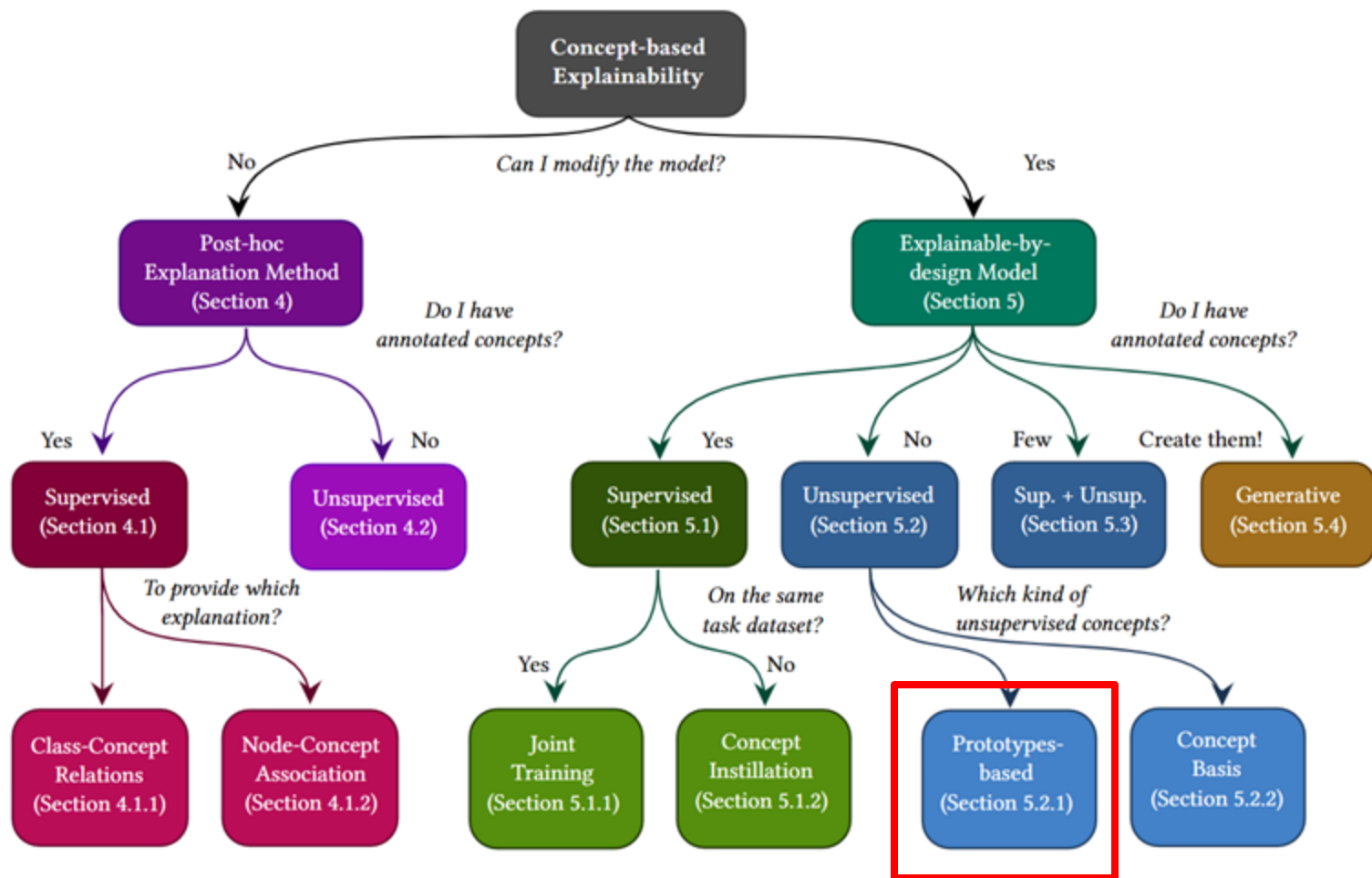


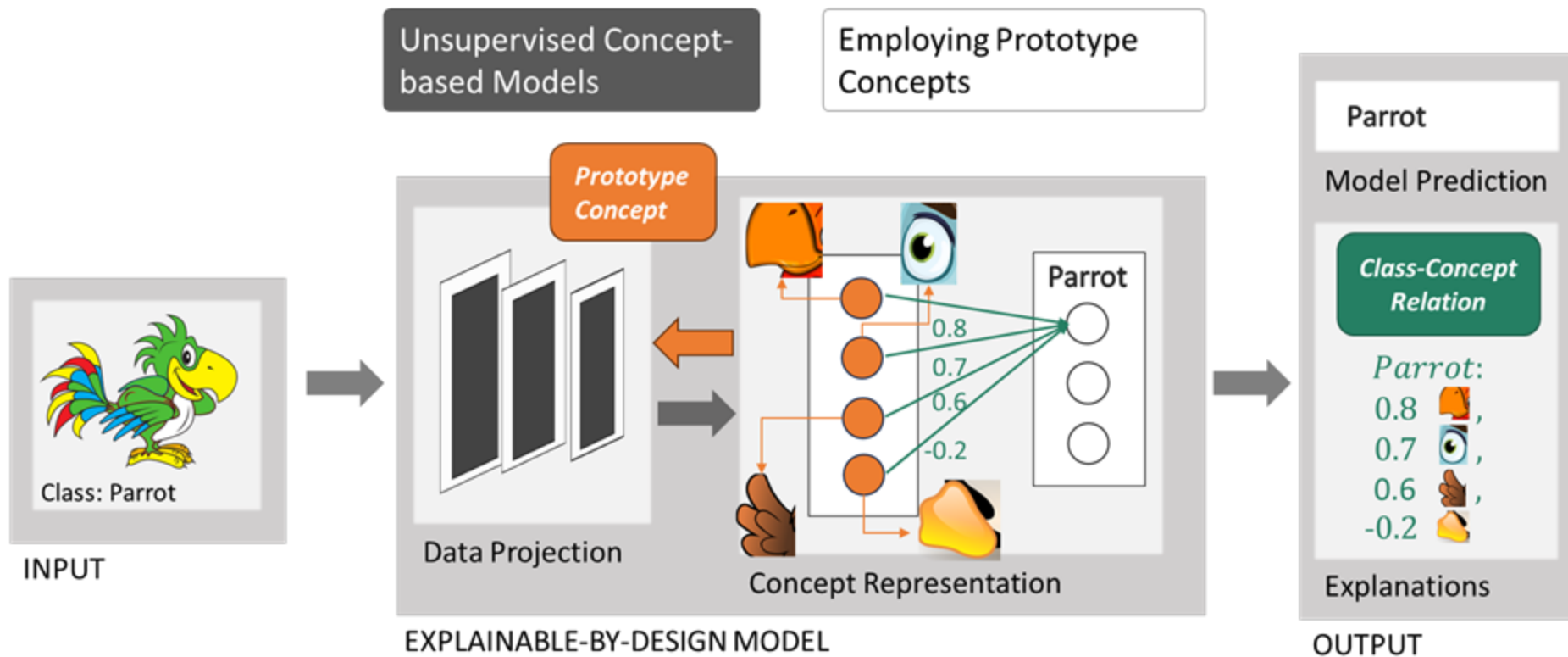


E.g., Chen, Zhi, et al. "Concept whitening for interpretable image recognition." Nature Machine Intelligence (2020).

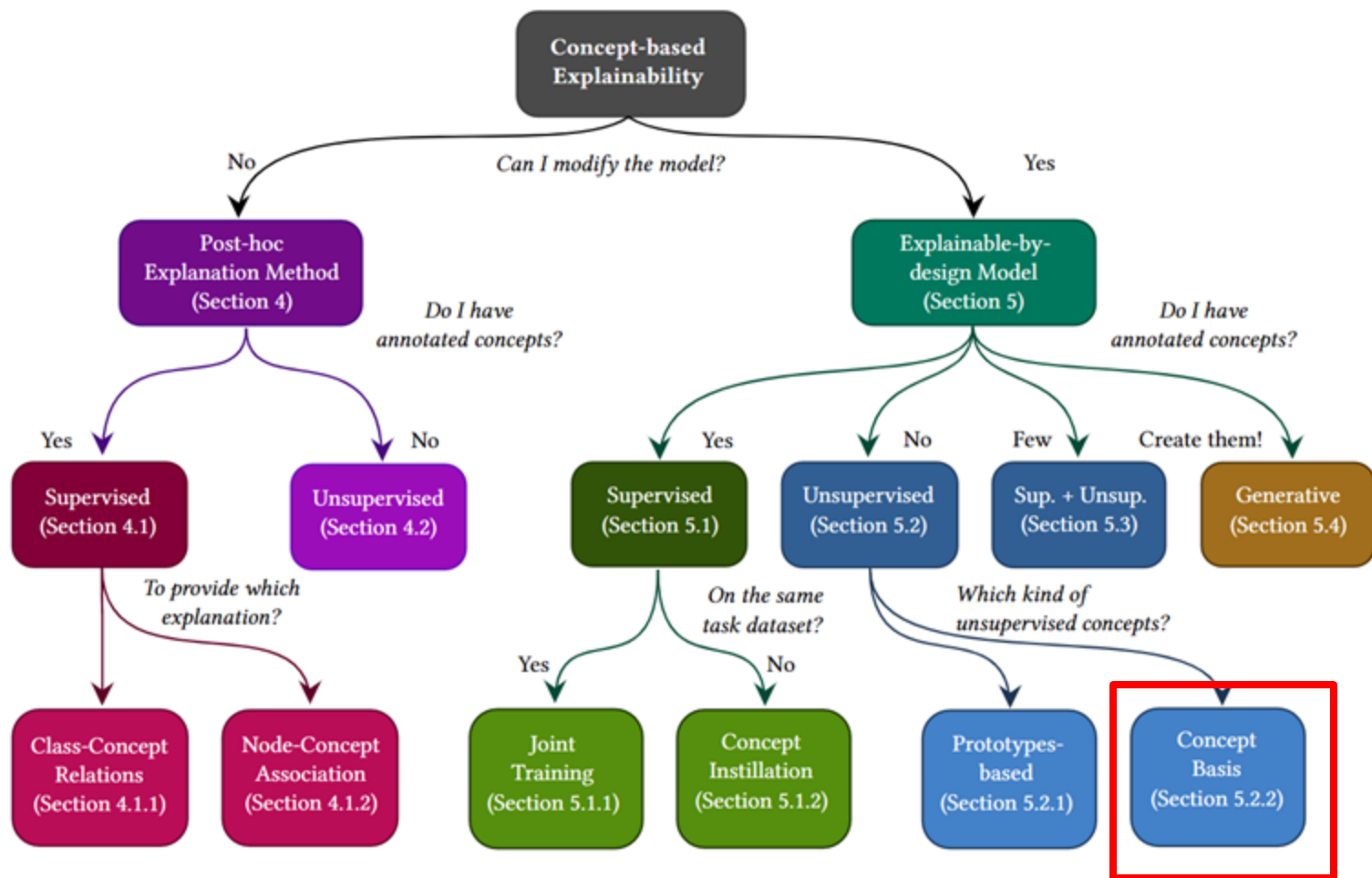


INTESA SANPAOLO
INNOVATION CENTER



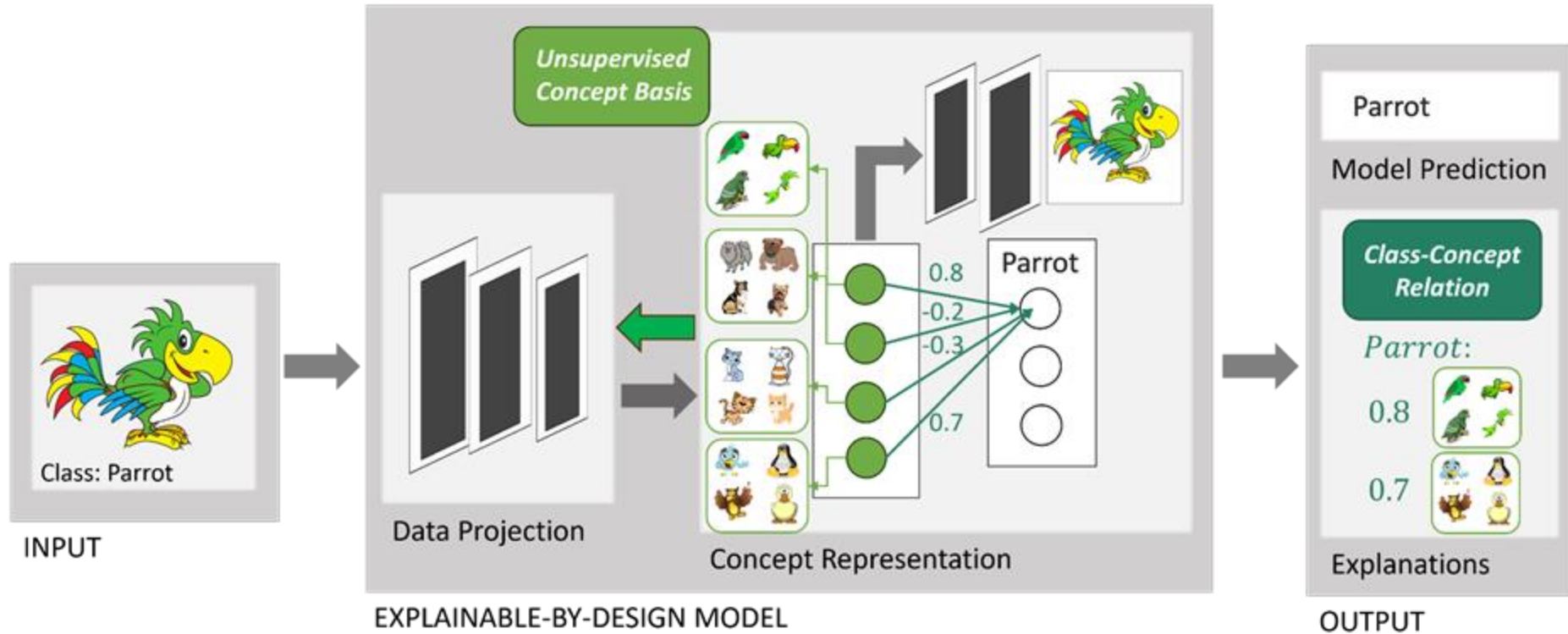


E.g., Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Neurips 2019.

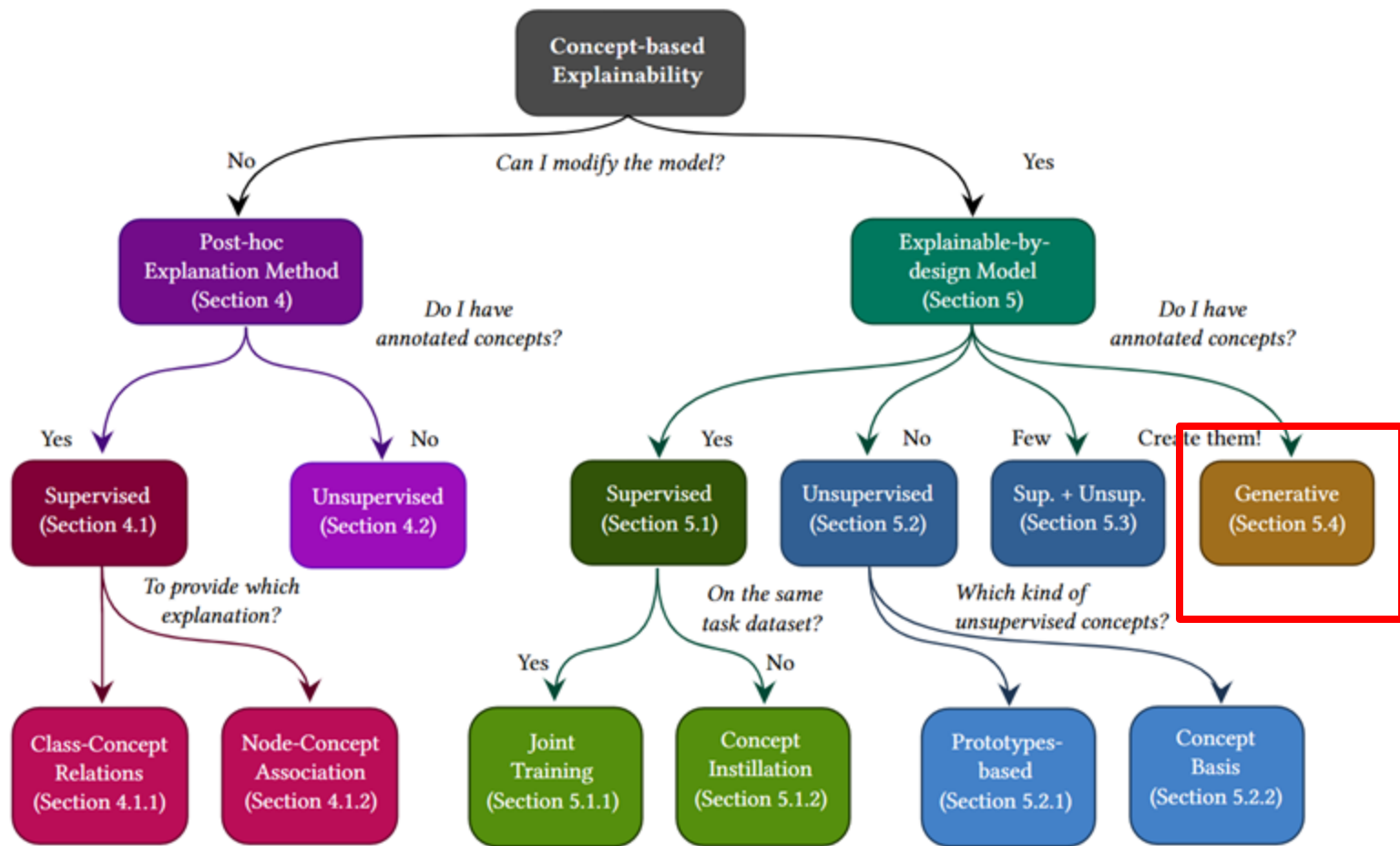


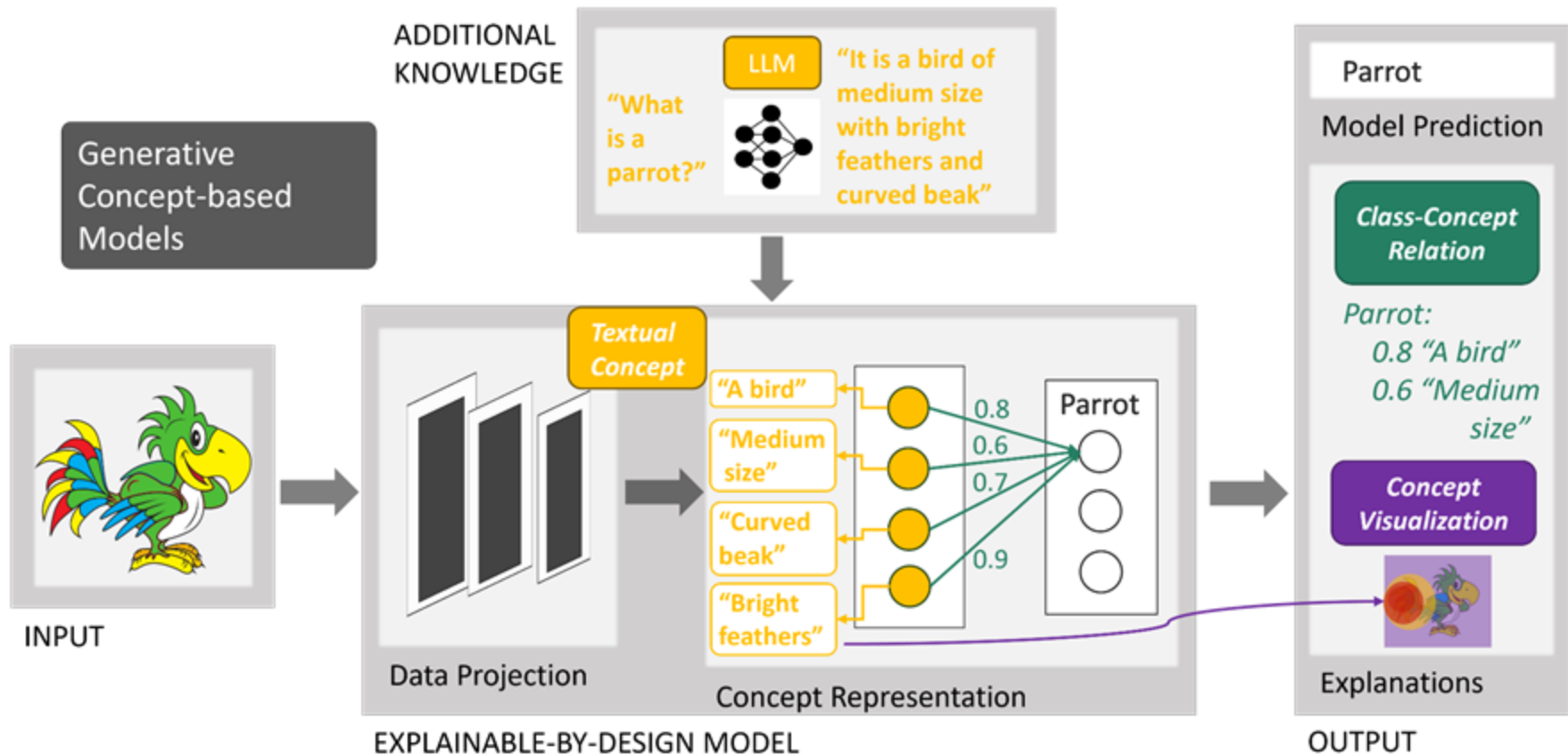
Unsupervised Concept-
based Models

Employing Unsupervised
Concepts Basis



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





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



INTESA SANPAOLO
INNOVATION CENTER

**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



CONCEPT
ENCODER



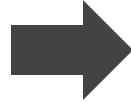
	ROUND	✓
	RED	✓
	SQUARED	✗
	COLD	✗

CONCEPTS

CBMs Explicitly Represents Concepts



CONCEPT
ENCODER



	ROUND	
	RED	
	SQUARED	
	COLD	

CONCEPTS

LABEL
PREDICTOR



“APPLE”



INTESA SANPAOLO
INNOVATION CENTER

CBMs Allows Interactions!



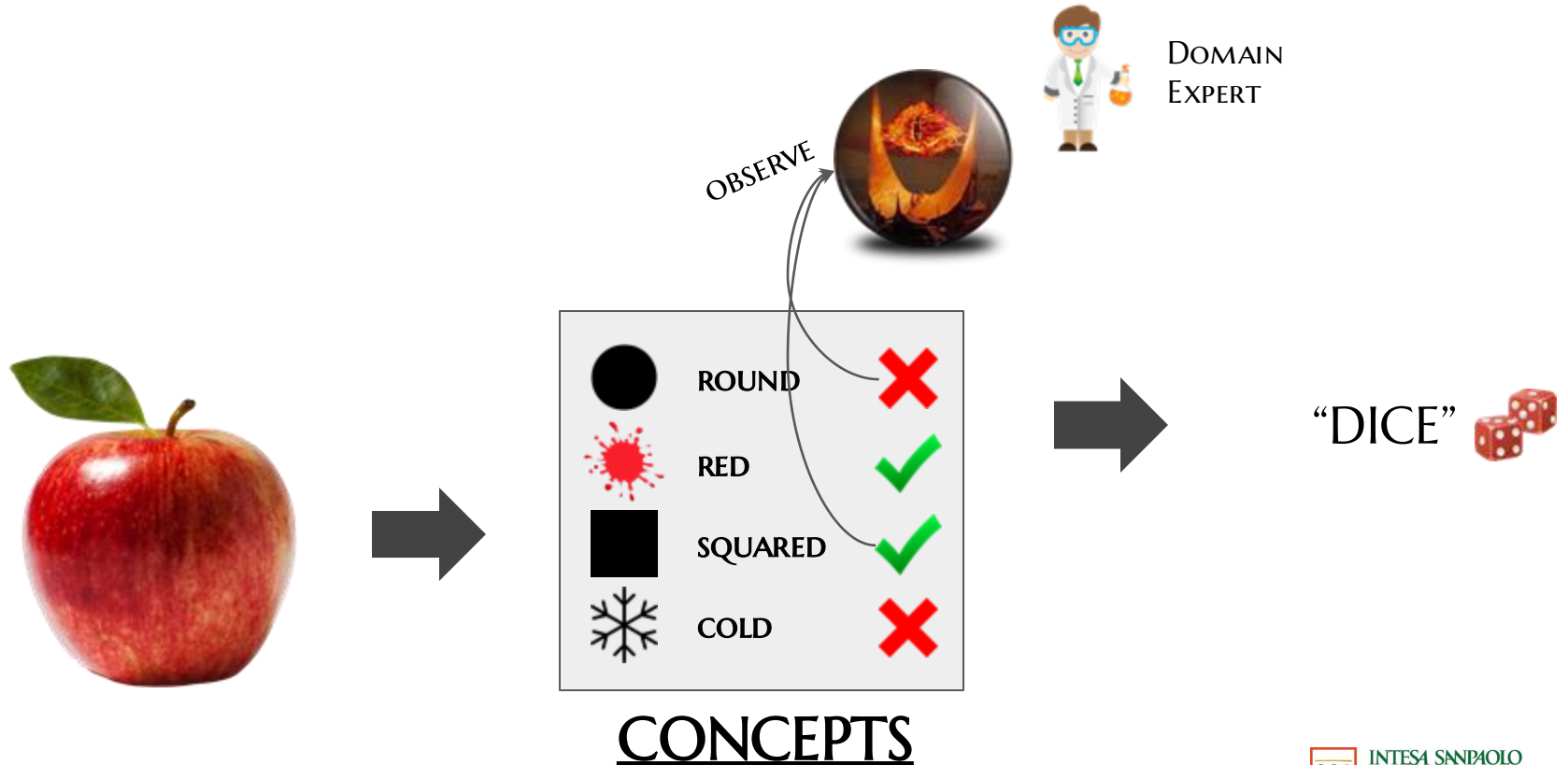
	ROUND	
	RED	
	SQUARED	
	COLD	



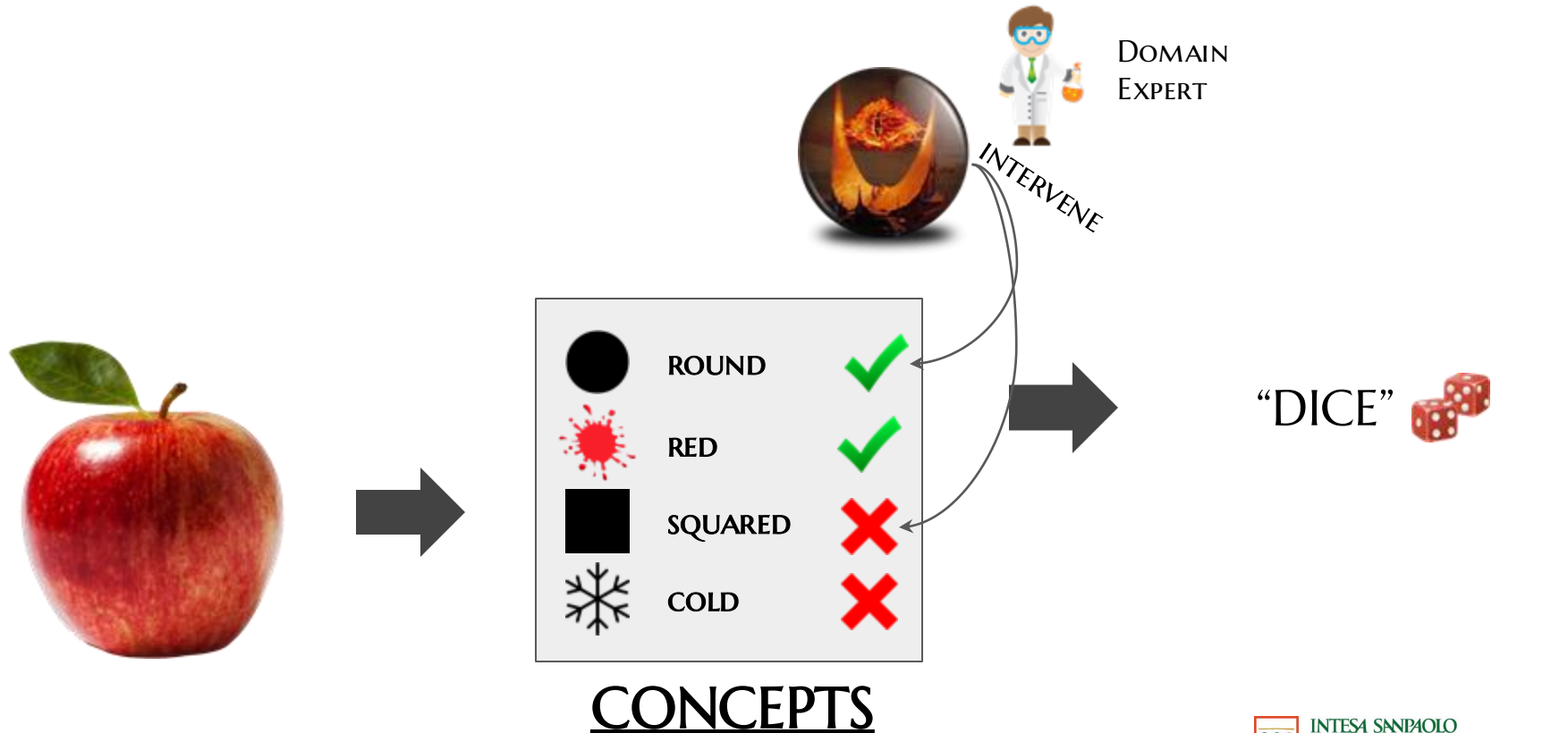
“DICE” 

CONCEPTS

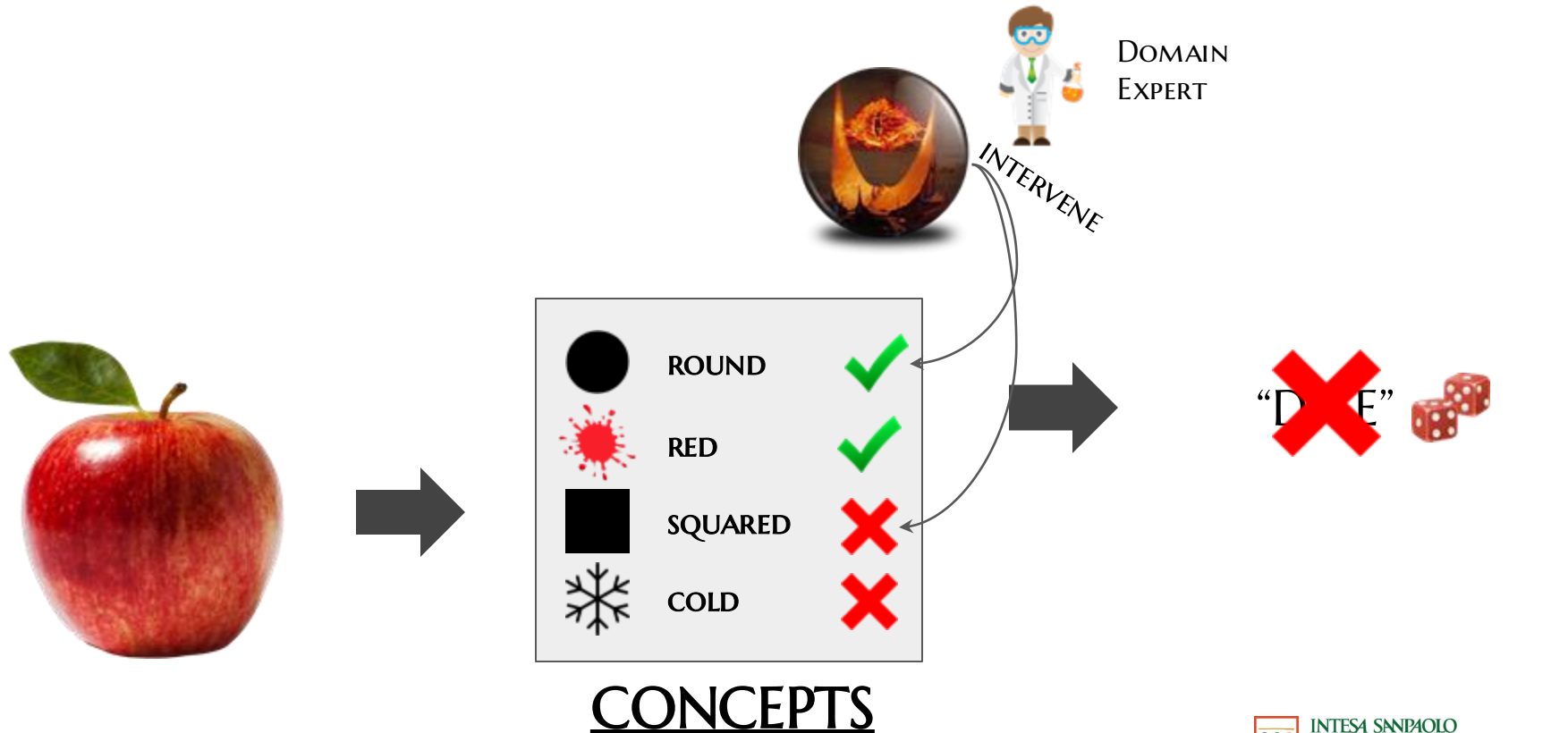
CBMs Allows Interactions!



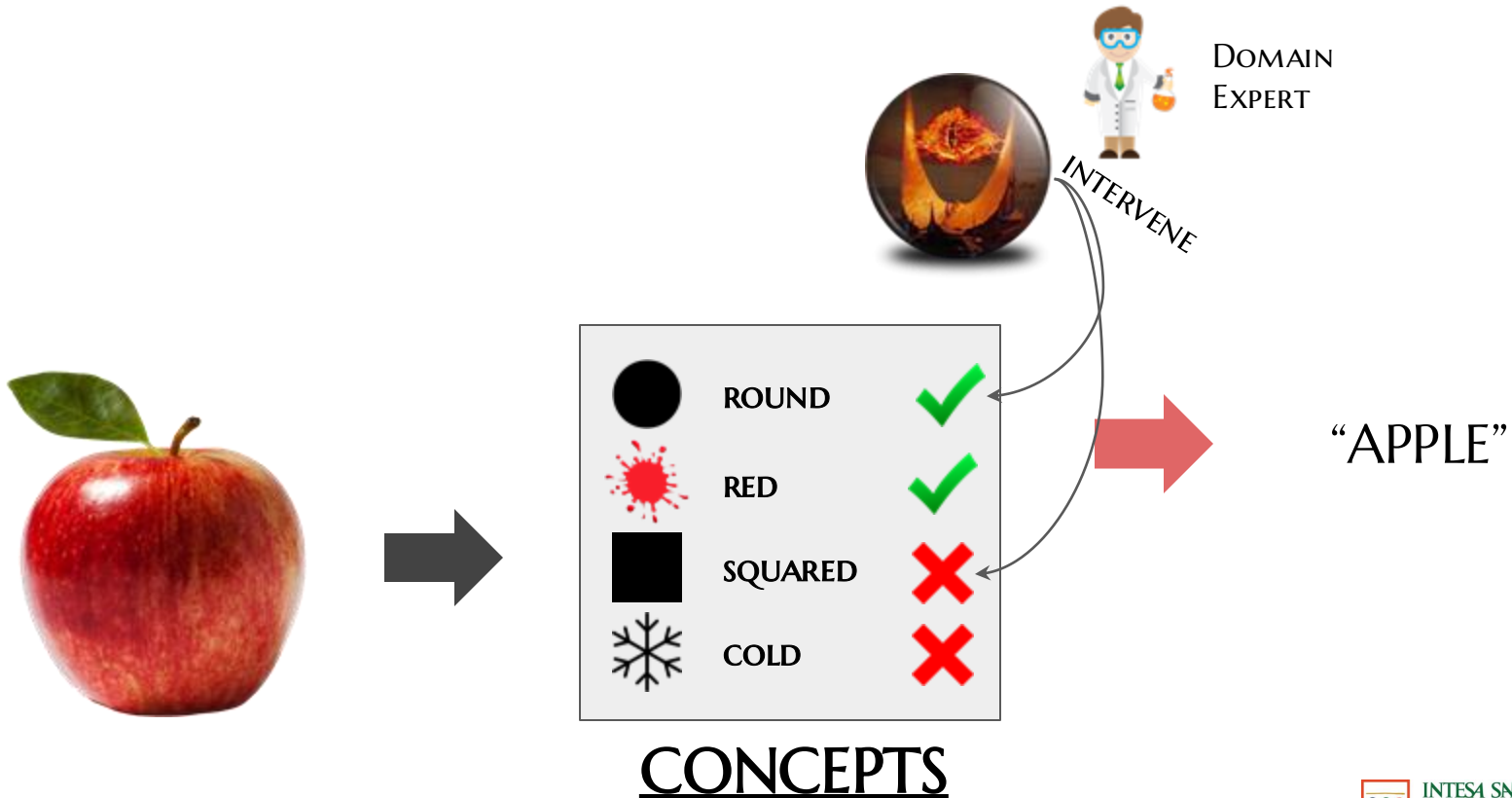
CBMs Allows Interactions!



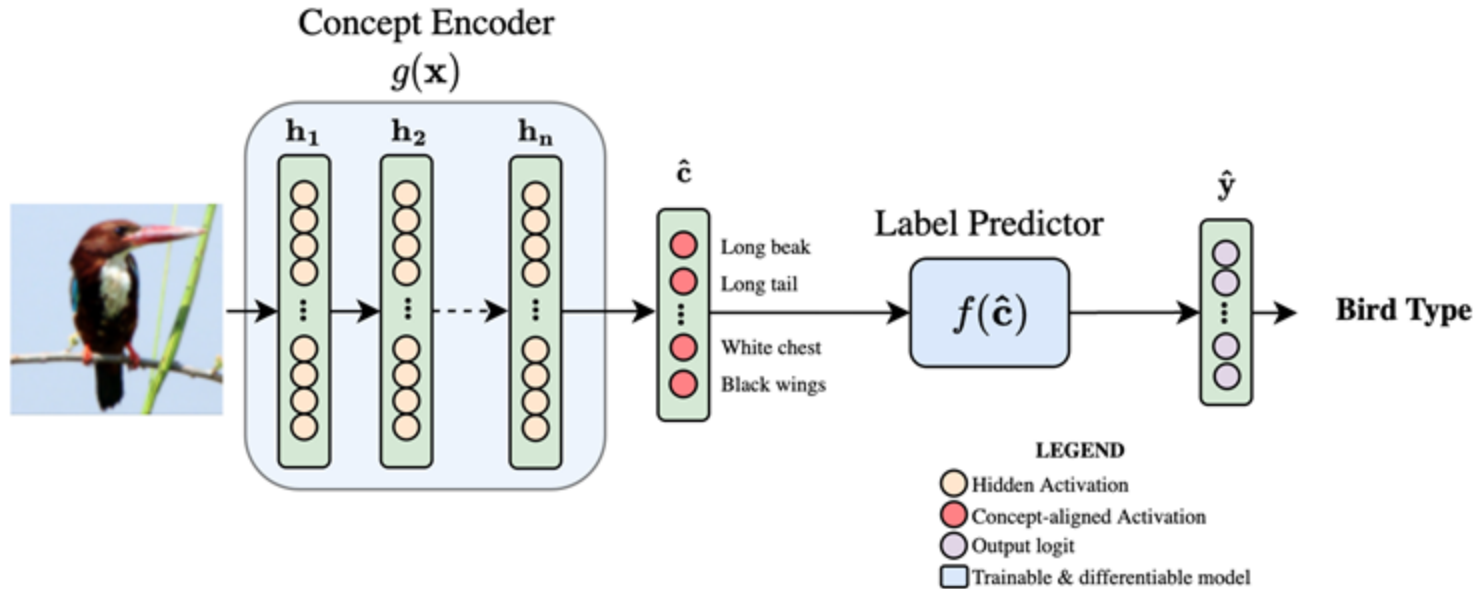
CBMs Allows Interactions!



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

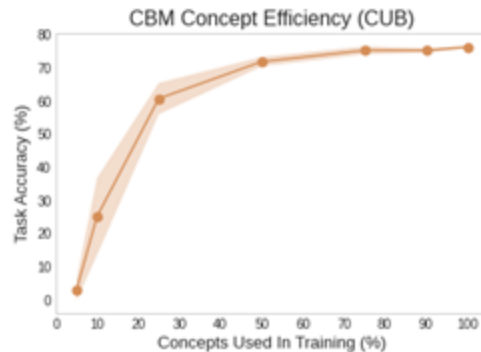
POOR TRADE-OFFS

STRUGGLE TO COMPROMISE BETWEEN
ACCURACY AND EXPLAINABILITY



LOW CONCEPT EFFICIENCY

CBMs DO NOT SCALE IN REAL-WORLD
CONDITIONS

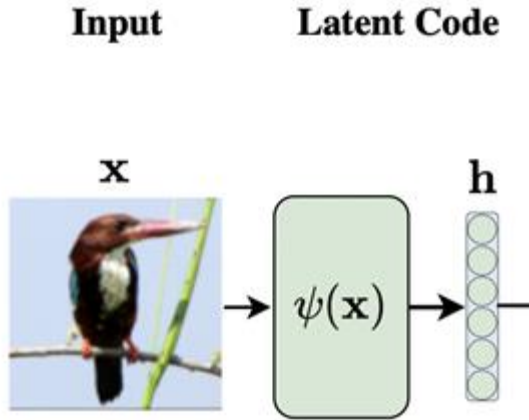


CONCEPT EMBEDDING MODELS (CEM)

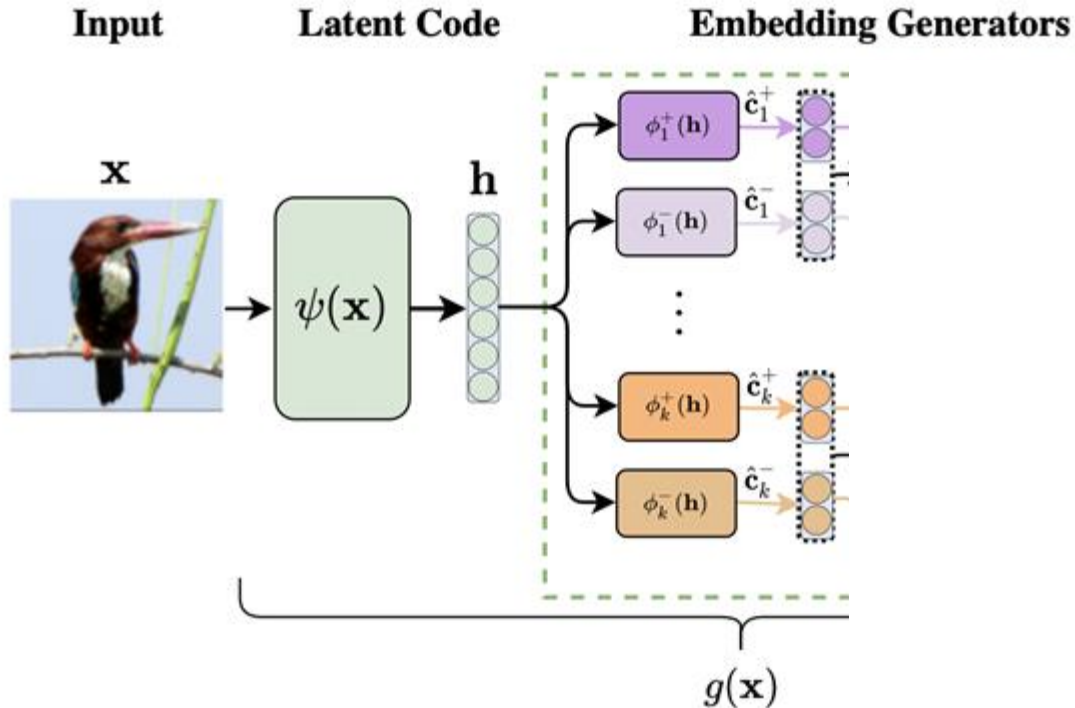
ESPINOSA ZARLENGA, MATEO, ET AL. "CONCEPT EMBEDDING MODELS: BEYOND THE ACCURACY-EXPLAINABILITY TRADE-OFF." *ADVANCES IN NEURAL INFORMATION PROCESSING* (2022).

Solution: Concept Embedding Models

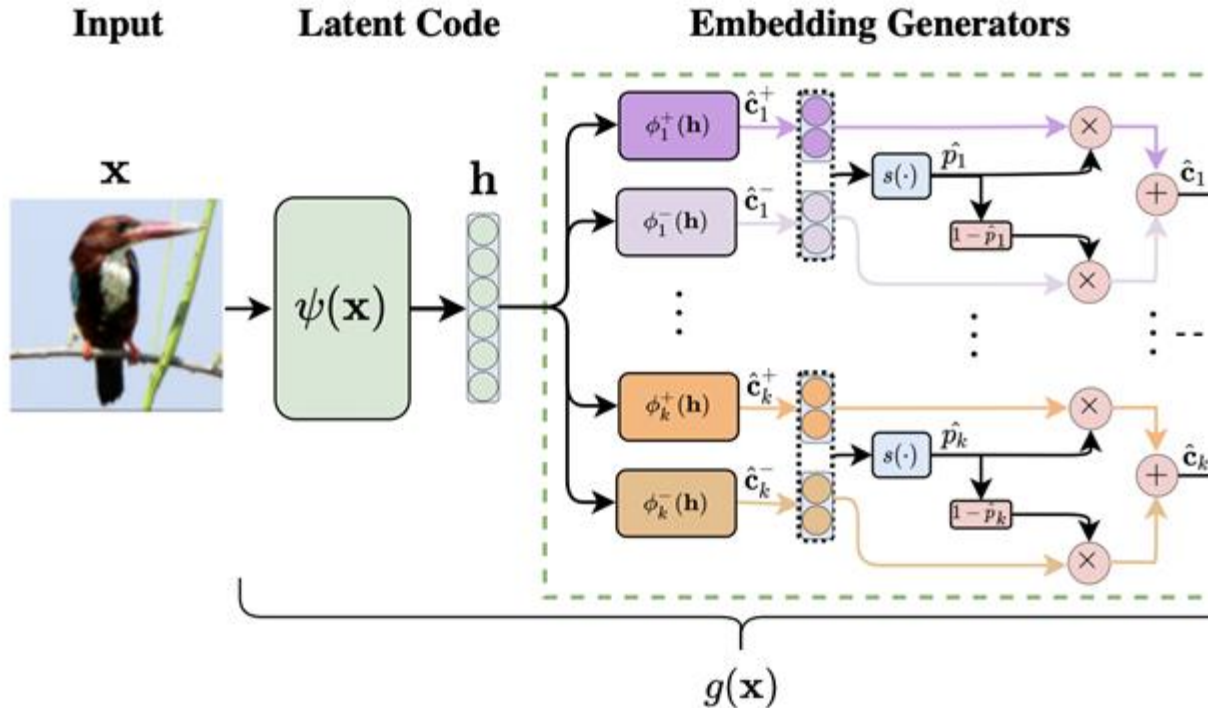
Solution: Concept Embedding Models



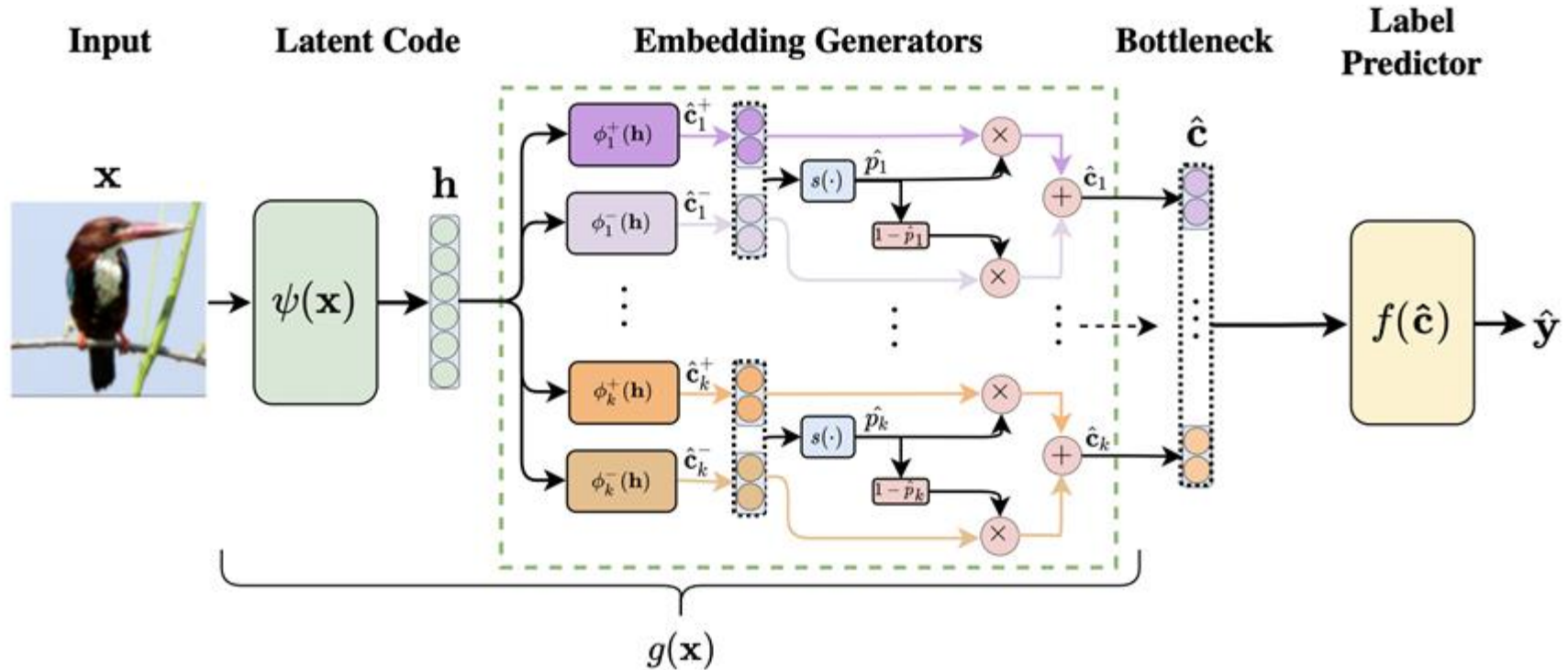
Solution: Concept Embedding Models



Solution: Concept Embedding Models

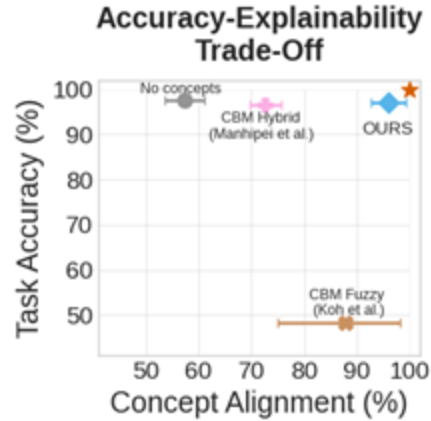


Solution: Concept Embedding Models



CEM Advantages

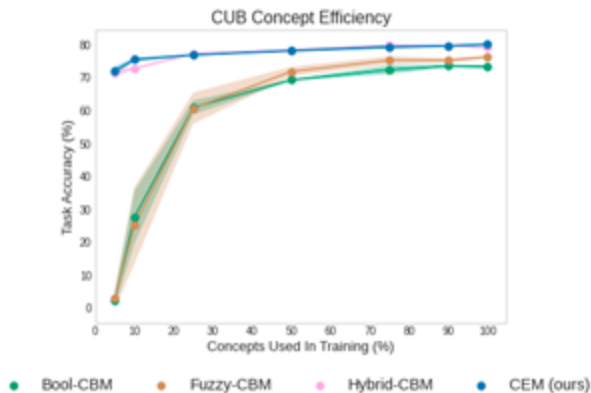
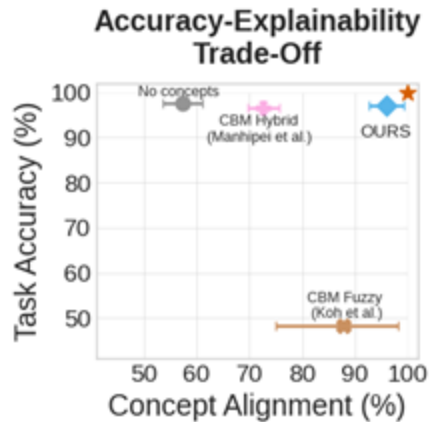
CEM Advantages



BEYOND TRADE-OFFS

CEMs OVERCOME THE CURRENT
ACCURACY-EXPLAINABILITY TRADE-
OFF

CEM Advantages



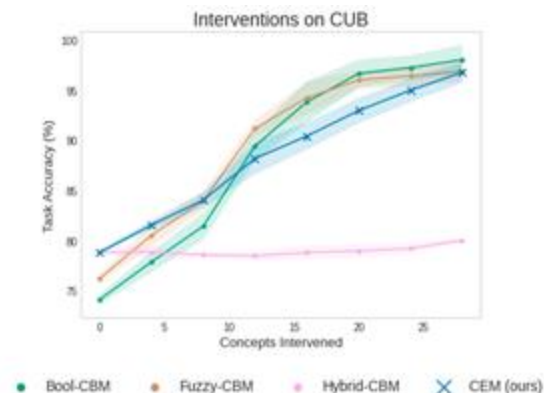
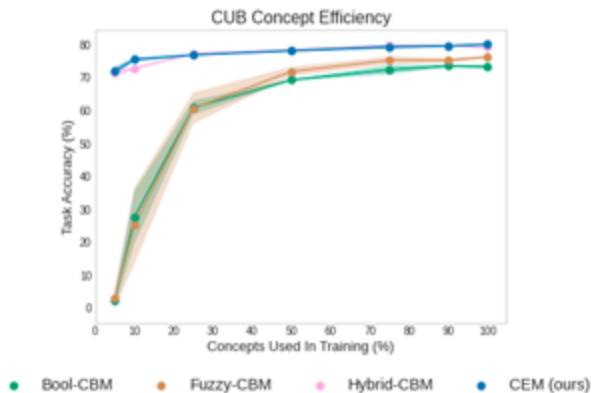
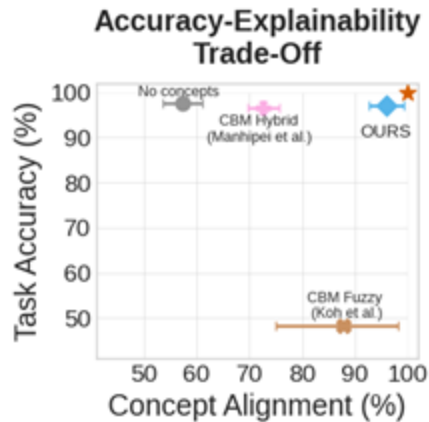
BEYOND TRADE-OFFS

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

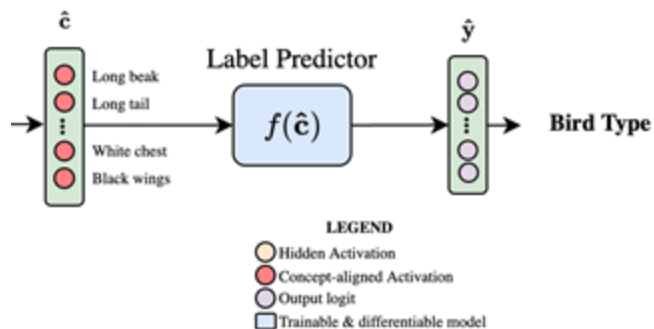
CEMs ARE RESPONSIVE TO
CONCEPT INTERVENTIONS

Have we lost something?

Have we lost something?

Interpretability

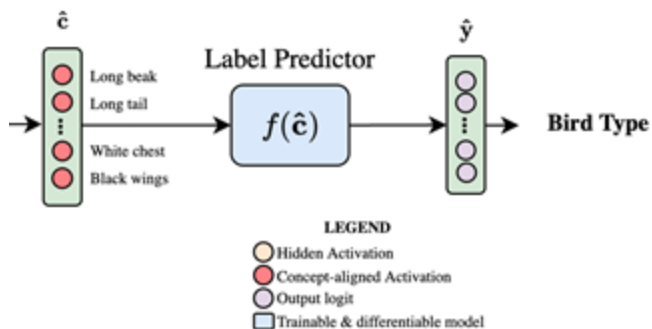
CBM: INTERPRETABLE



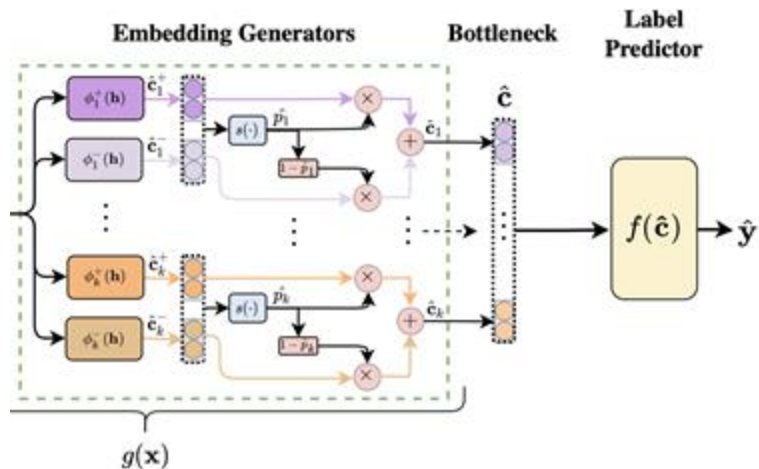
Have we lost something?

Interpretability

CBM: INTERPRETABLE



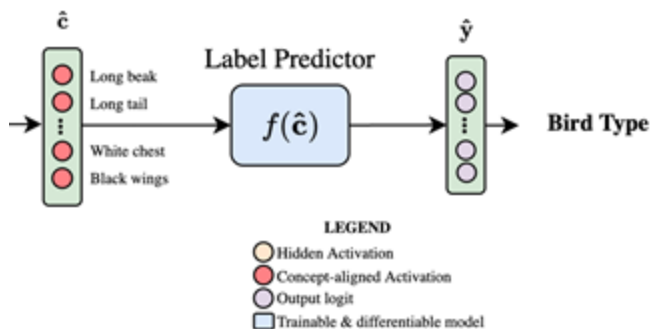
CEM: NON-INTERPRETABLE



Have we lost something?

Interpretability

CBM: INTERPRETABLE



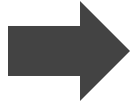
CEM: NON-INTERPRETABLE

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

Can we create an Interpretable Model over CE?



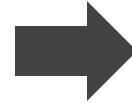
CONCEPT
ENCODER



	ROUND	
	RED	
	SQUARED	
	COLD	

CONCEPTS

CONCEPT
PREDICTOR



ROUND & RED
→ APPLE

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

“RED”



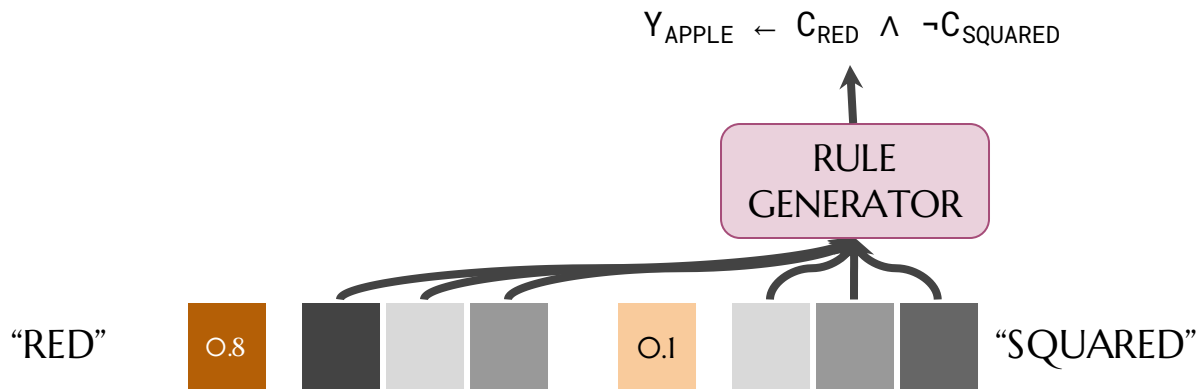
“SQUARED”



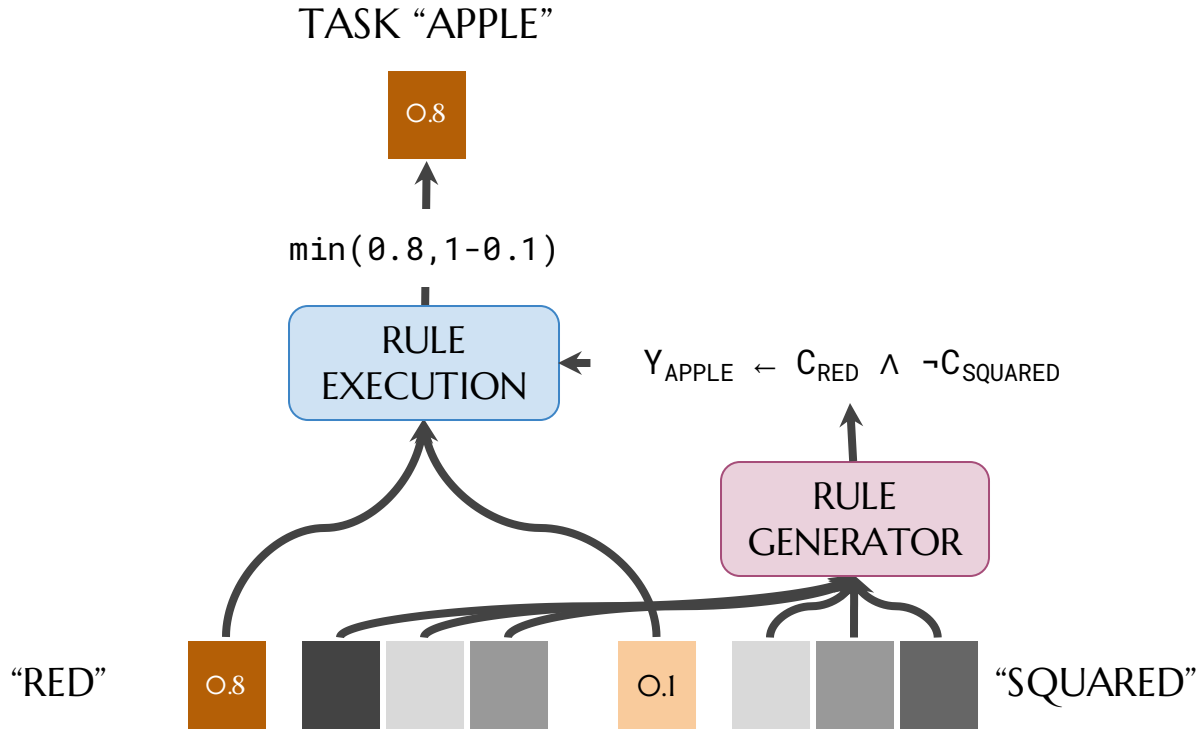
INTESA SANPAOLO
INNOVATION CENTER

DEEP CONCEPT REASONER

1. GENERATES SYNTACTIC RULES WITH CONCEPT EMBEDDINGS

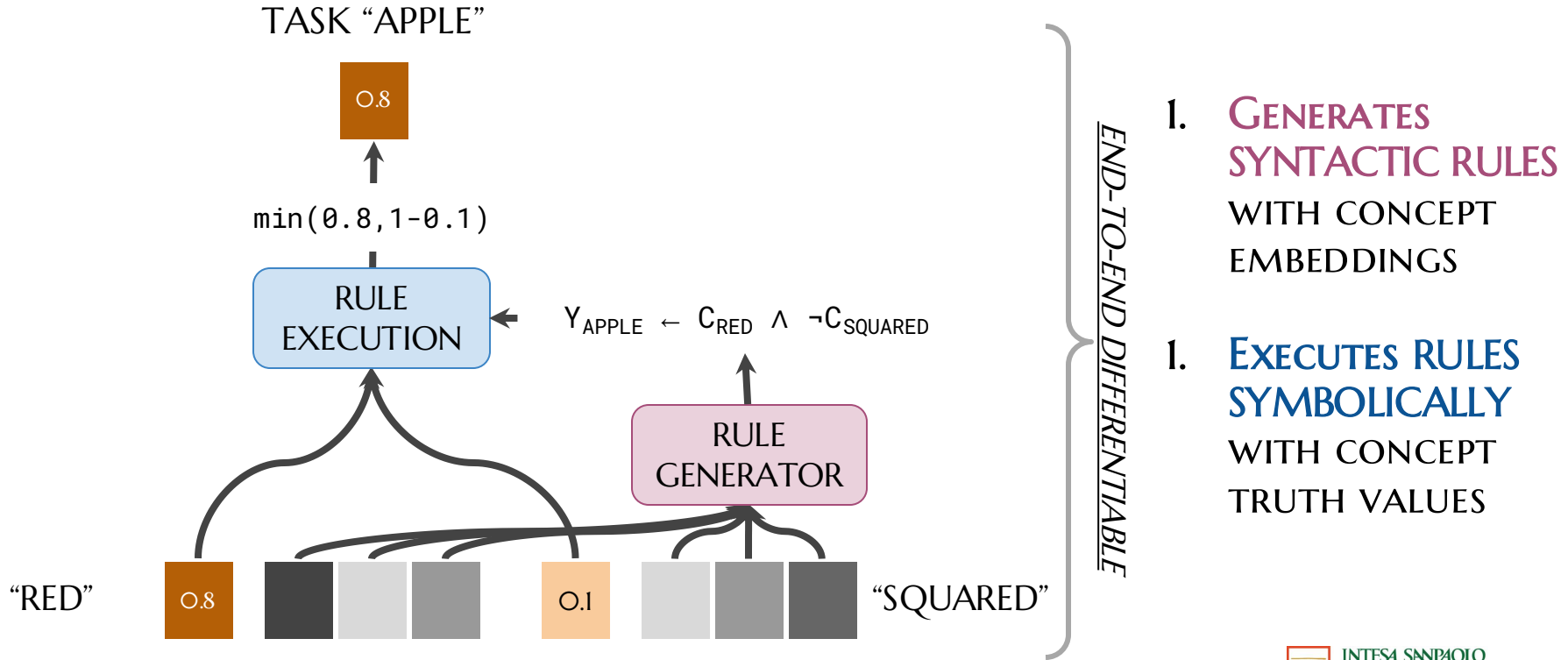


DEEP CONCEPT REASONER

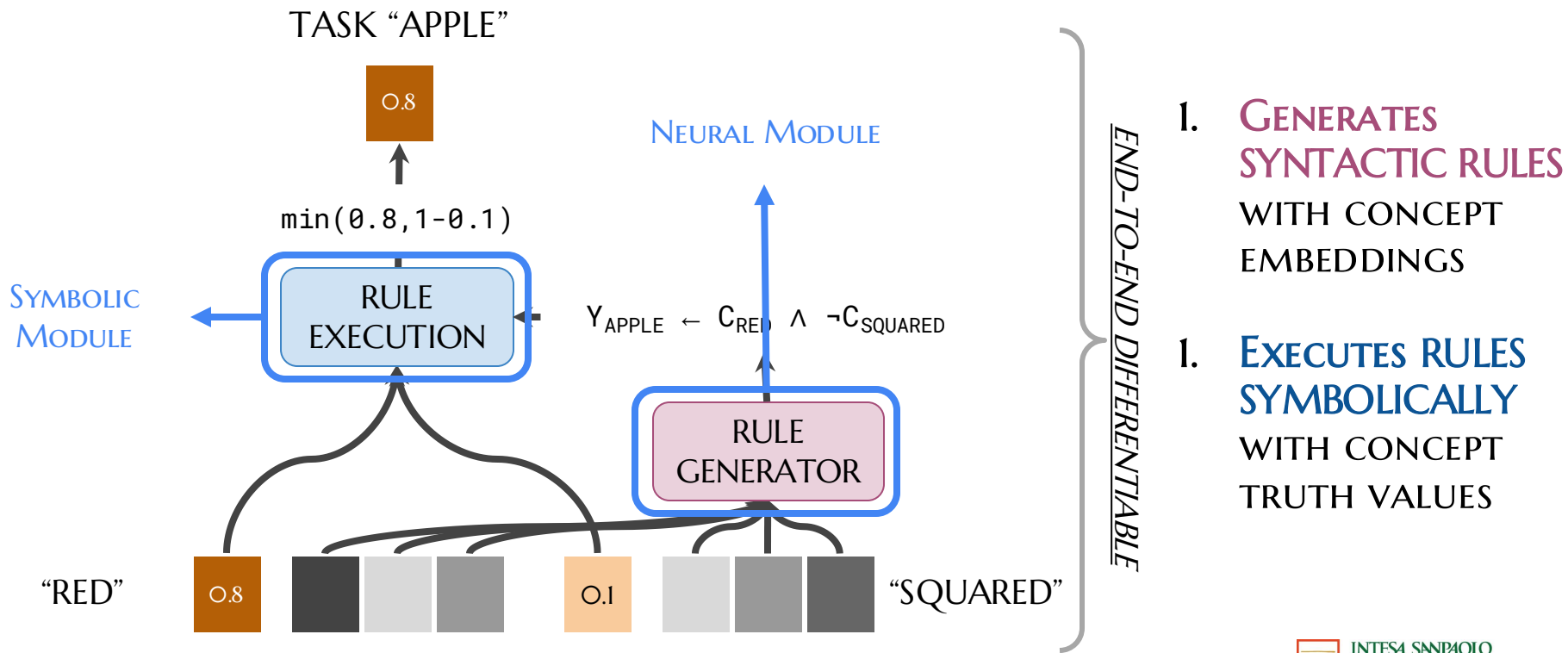


1. **GENERATES SYNTACTIC RULES** WITH CONCEPT EMBEDDINGS
1. **EXECUTES RULES SYMBOLICALLY** WITH CONCEPT TRUTH VALUES

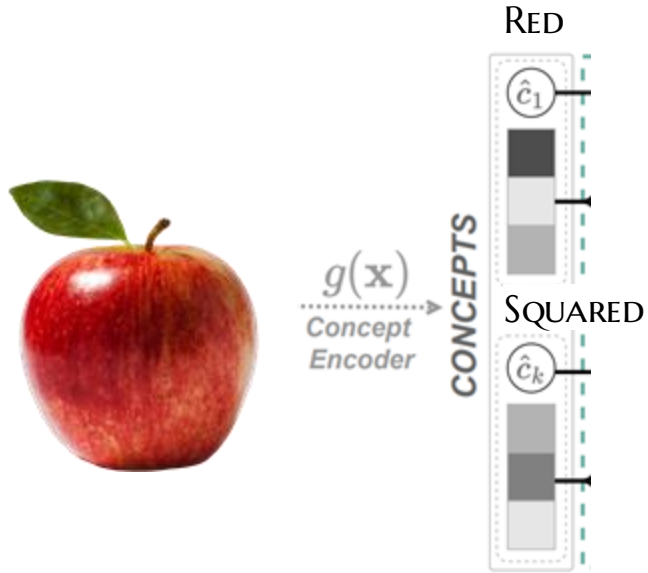
DEEP CONCEPT REASONER



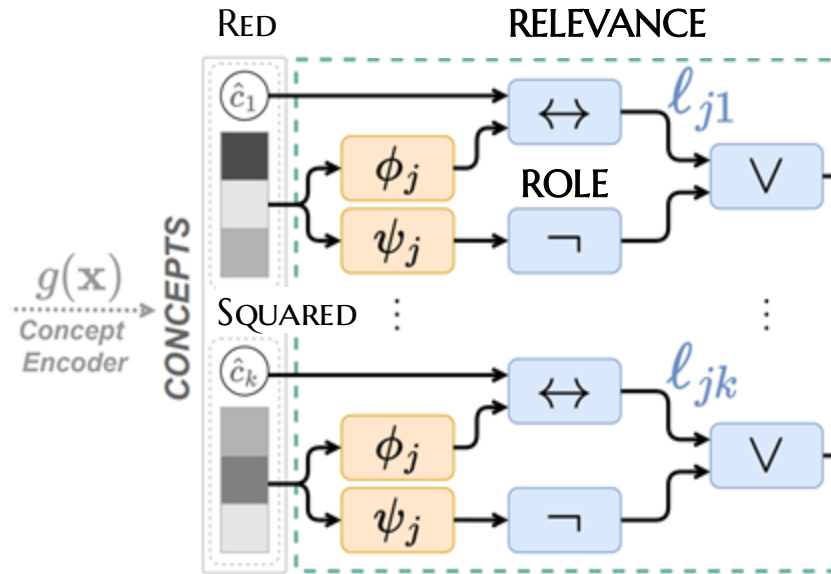
DEEP CONCEPT REASONER



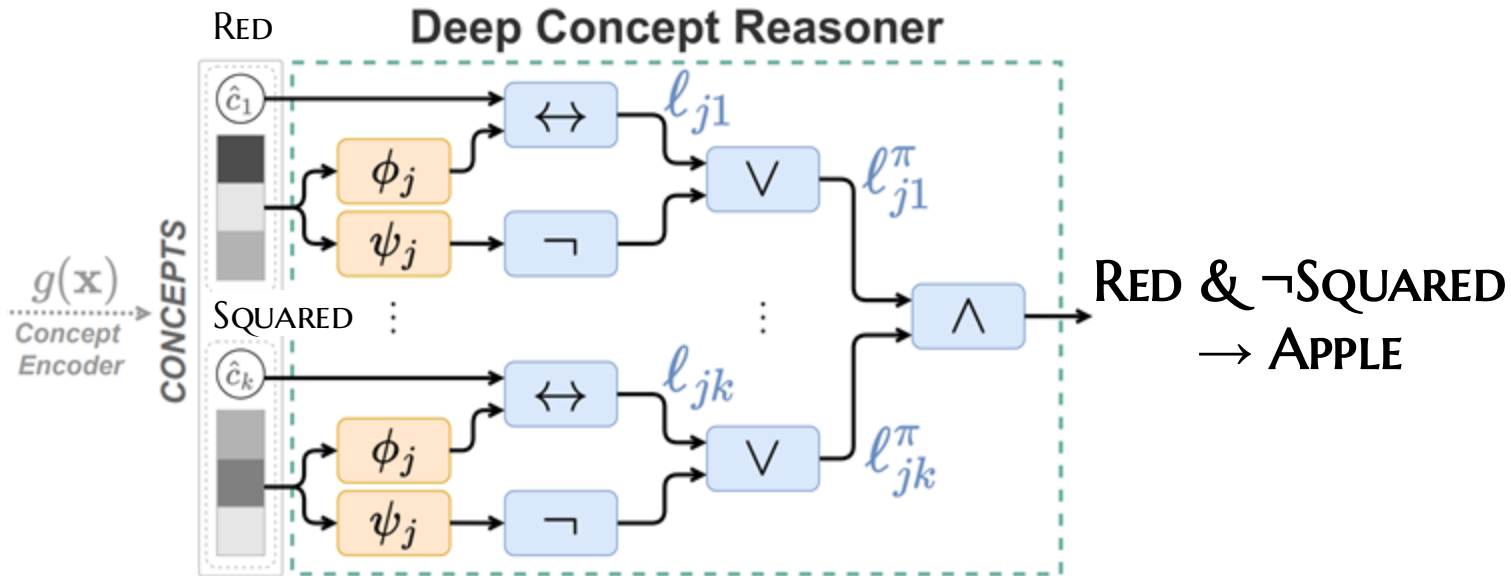
Deep Concept Reasoner (DCR)



Deep Concept Reasoner (DCR)

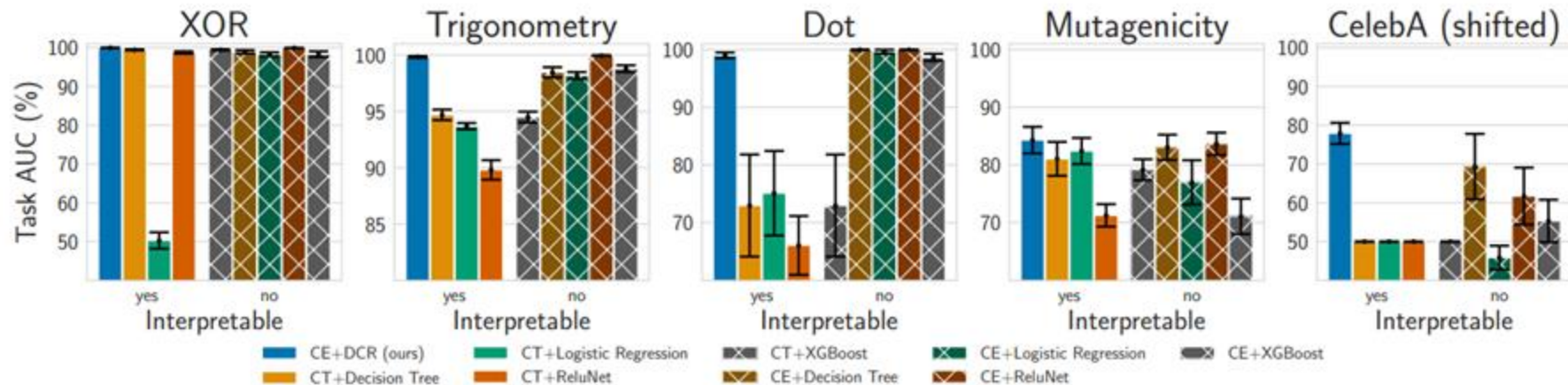


Deep Concept Reasoner (DCR)



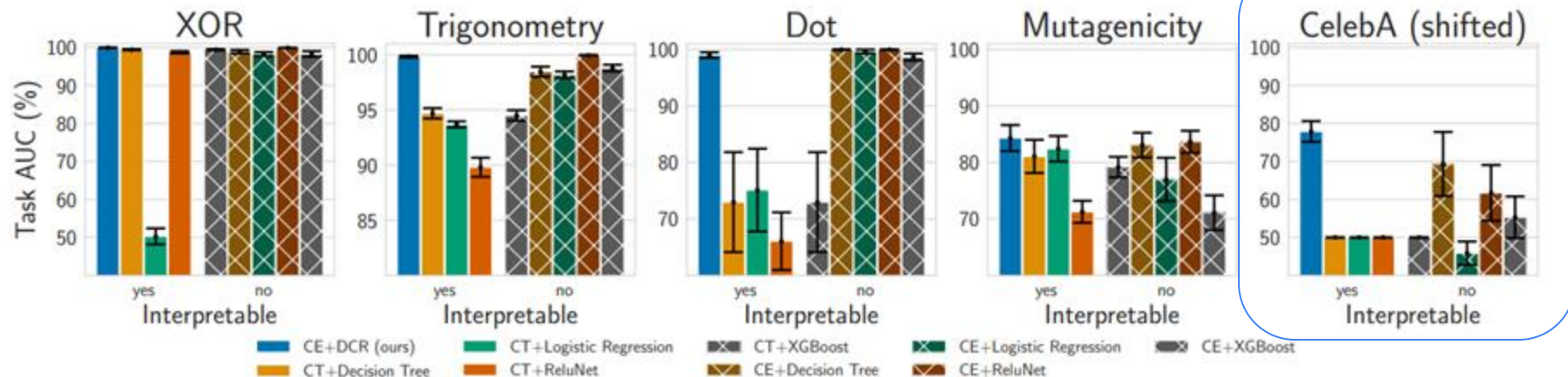
Does DCR Generalize Well?

Does DCR Generalize Well?



Does DCR Generalize Well?

DIFFERENT SET OF
TEST CONCEPTS!



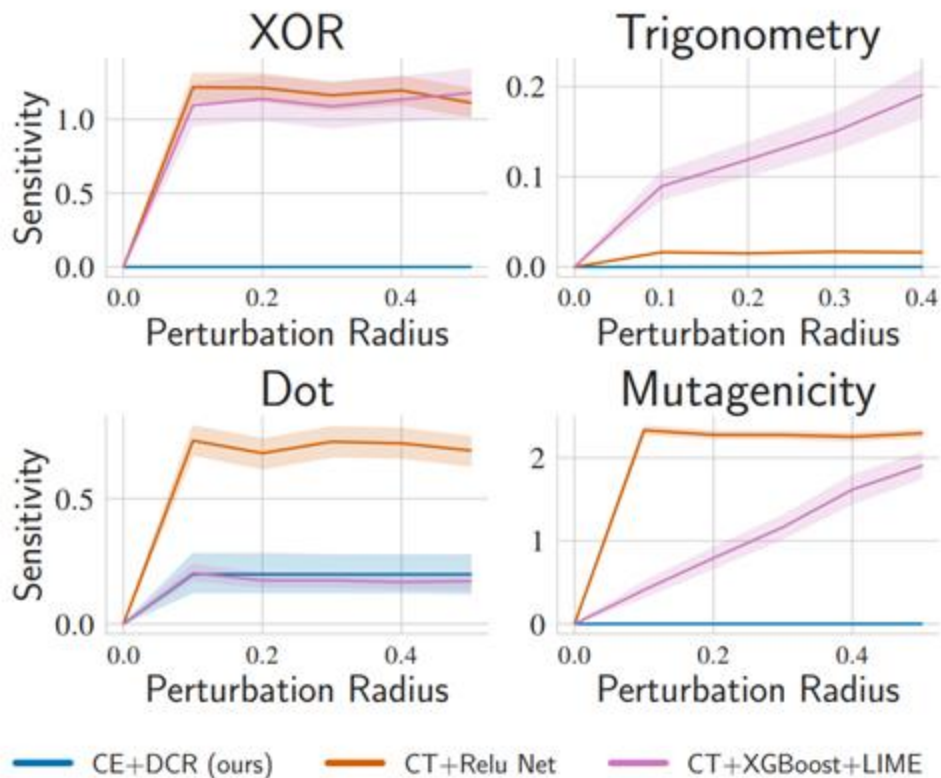
Are DCR Rules Semantically Meaningful?

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



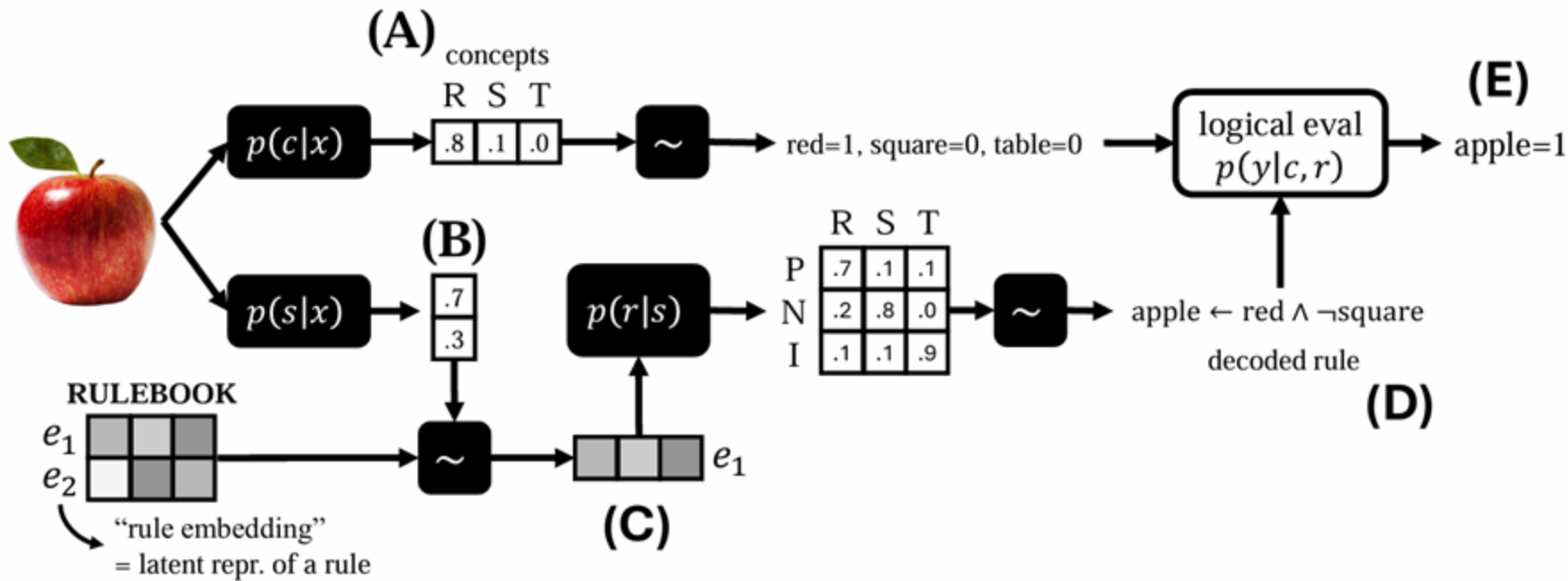
Are rules Stable under perturbation?



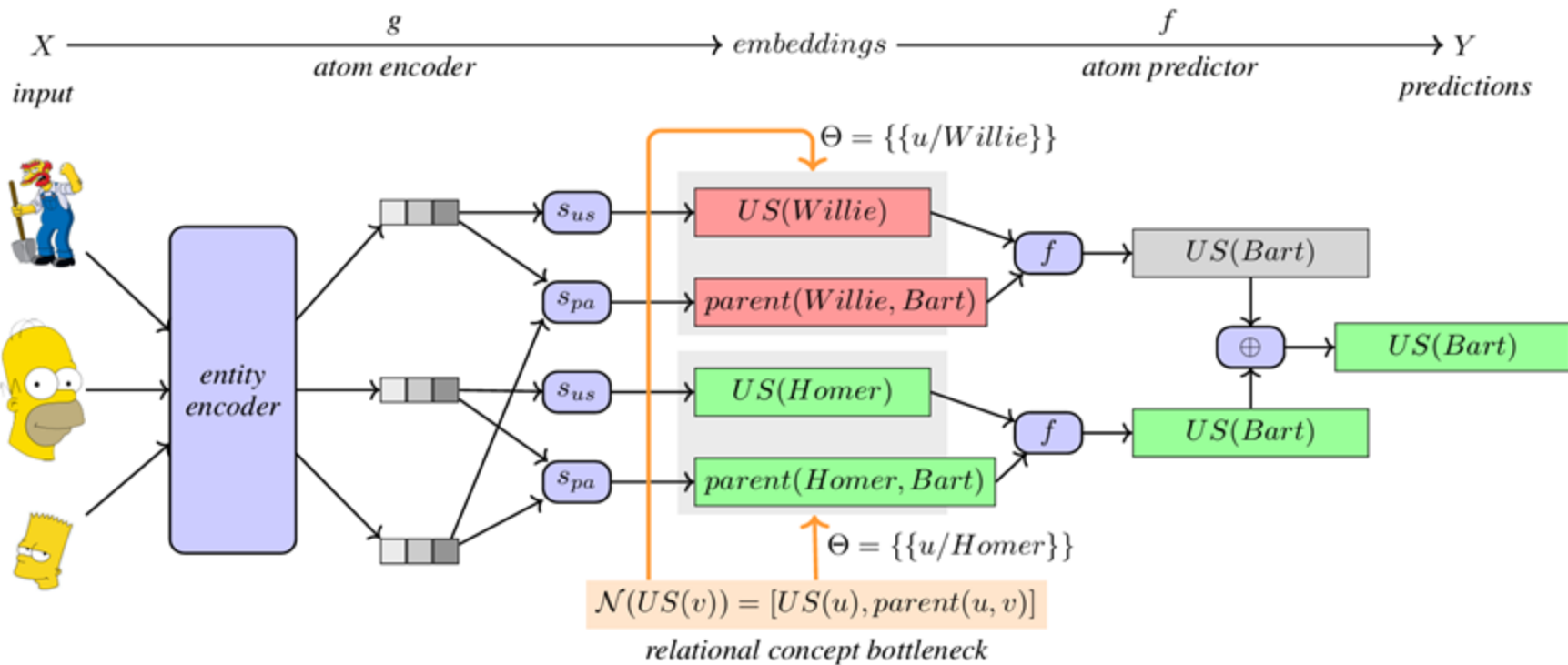
The background is a dark gray with a subtle, abstract geometric pattern. It consists of a network of thin, light gray lines connecting small dots, creating a series of interconnected triangles and polygons. The pattern is more dense on the left side and fades towards the right.

NEW WORKS

DCR rules now have global validity with the “Rulebook”



Relational Concept Bottleneck Models



So far we have only studied logic rules...

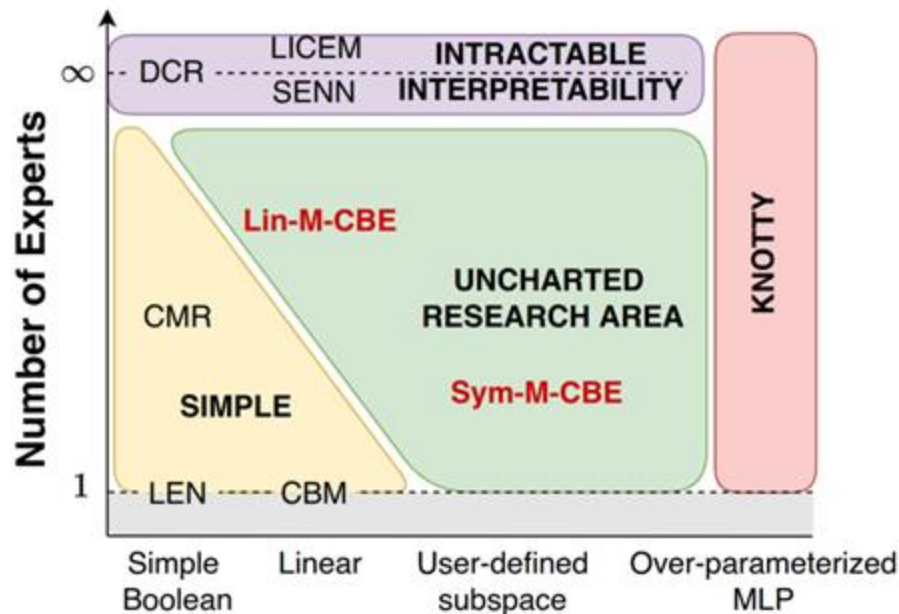
BUT THERE ARE MANY MORE
FUNCTIONAL FORMS STILL
UNCHARTED.

VARYING:

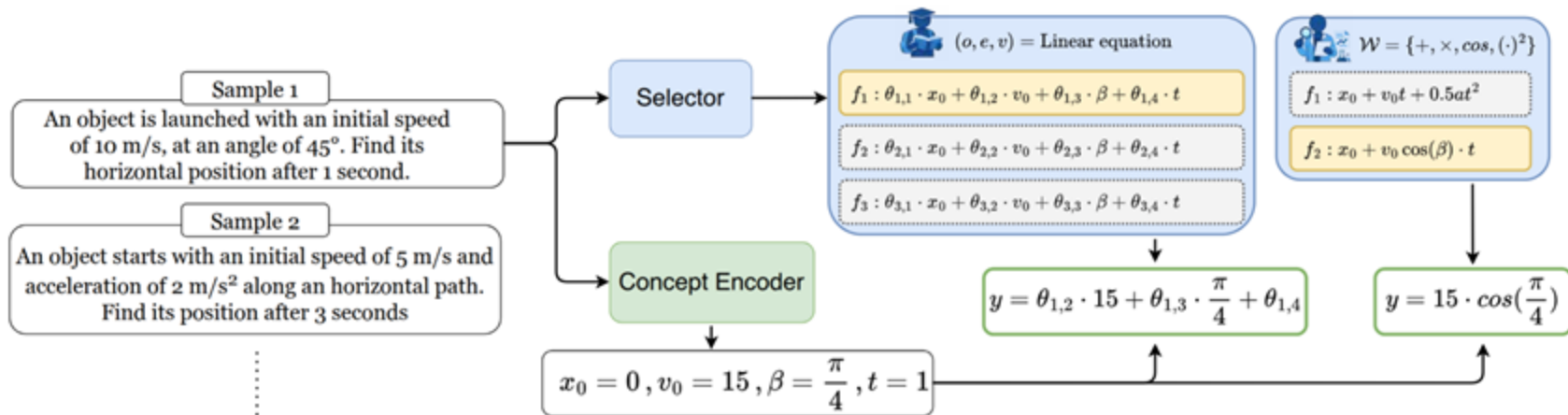
- THE NUMBER OF EXPERTS
- FUNCTIONAL FORMS

FIND:

- THE OPTIMAL ACCURACY-
INTERPRETABILITY TRADE-OFF



Let's generate any Equations!



DYNAMIC SELECTION — SELECTOR ROUTES INPUTS TO SPECIALIZED MATHEMATICAL EXPERTS

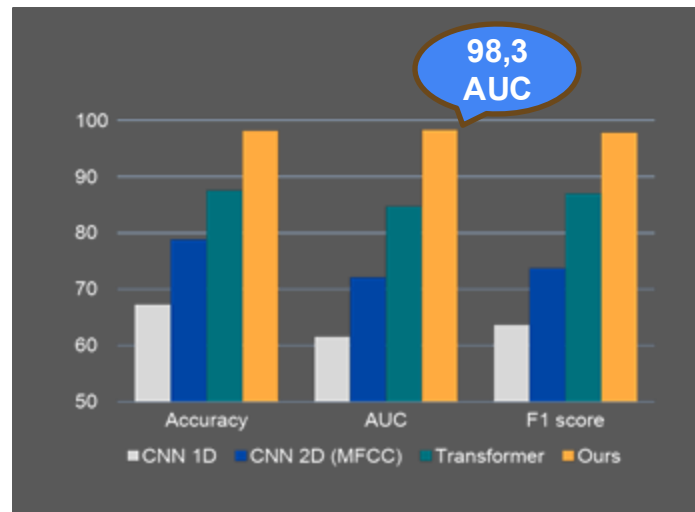
USER ADAPTATION — EACH EXPERT SET OF OPERATORS IS DEFINED BY THE USER ACCORDING TO ITS EXPERTISE

FLEXIBLE FORMS — SUPPORTS BOTH DISCRETE CLASSIFICATION AND CONTINUOUS REGRESSION TASKS

The background is a dark gray with a subtle, abstract geometric pattern. It consists of a network of thin, light gray lines connecting small dots, creating a series of interconnected triangles and polygons. The pattern is more dense on the left side and fades towards the right.

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.



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!!



(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



Acknowledgments (2)

PIETRO BARBIERO



FRANCESCO GIANNINI



MATEO ESPINOZA ZARLENGA



GIUSEPPE MARRA



INTESA SANPAOLO
INNOVATION CENTER

References

Concept-based Explainable Artificial Intelligence: A Survey



SURVEY

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



CEM

Interpretable Neural-Symbolic Concept Reasoning



DCR

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

The contents of this document are the property of Intesa Sanpaolo and are protected by copyright and trademark laws. All rights reserved. Therefore, without the prior formal consent of Intesa Sanpaolo, the aforementioned material cannot be copied, downloaded, reproduced, used on other Internet sites, modified, transferred, distributed or communicated to third parties except for personal use only; any commercial use is still prohibited. The contents not owned by Intesa Sanpaolo but by third parties indicated from time to time may be used only in compliance with the rights of such third parties and the rules established by them, to which reference is therefore made.

While taking care in preparing the contents, Intesa Sanpaolo accepts no responsibility for any errors or omissions of any kind and for any type of direct, indirect or accidental damage deriving from reading or using the information published or any form of content present in the document or for access or use of the material contained in other linked sites