Directions in Interpretability

Ruth Fong

CVPR 2022, Human-Centered AI Tutorial

June 20, 2022

Slides and links available at ruthfong.com

PRINCETON UNIVERSITY

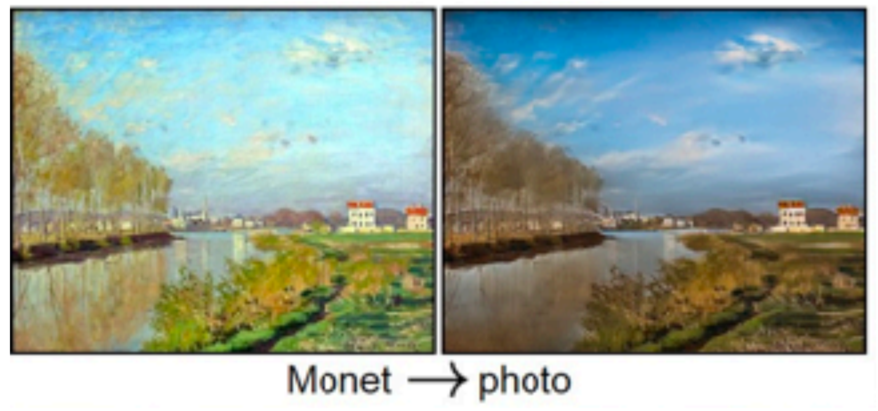
What is interpretability?

Research focused on explaining **complex AI systems** in a **human-interpretable** way.

Why interpretability?

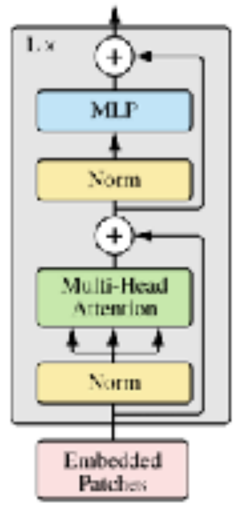
- 🔬 Science
- 🤝 Trust
- 🤖 Learning

An incomplete retrospective: the first decade of deep learning



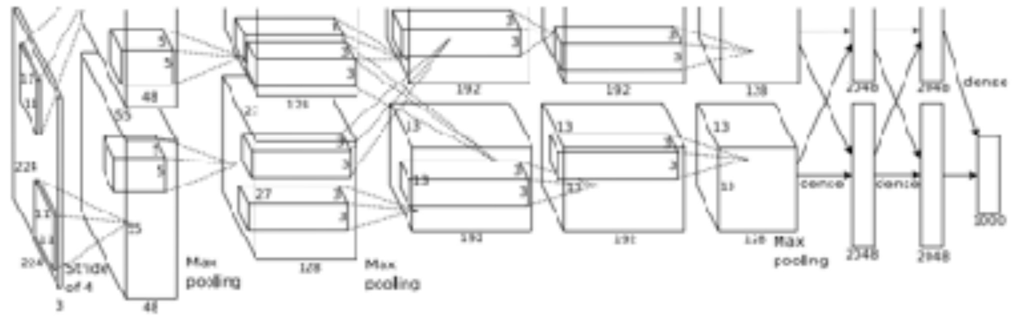
IMAGENET

GANs (2014-2018)
GAN, ProGAN, CycleGAN



Transformers (2017-now)
Transformer, BERT, ViT

2012 — 2022



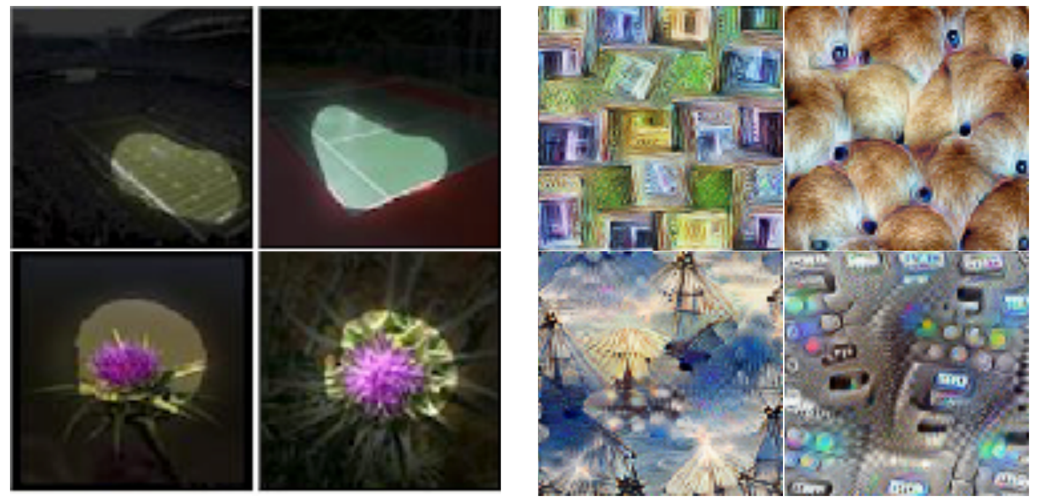
CNNs (2012-2016)
AlexNet, VGG16, GoogLeNet, ResNet50

Self-supervised learning (2016-now)
Colorization, MOCO, SWaV

Diffusion models (2020-now)
DDPM, DALL-E 2, Imagen

[Krizhevsky et al., NeurIPS 2012; Zhu* & Park* et al., ICCV 2017; Zhang et al., ECCV 2016; Dosovitskiy* et al., ICLR 2021; Ramesh et al., arXiv 2022]

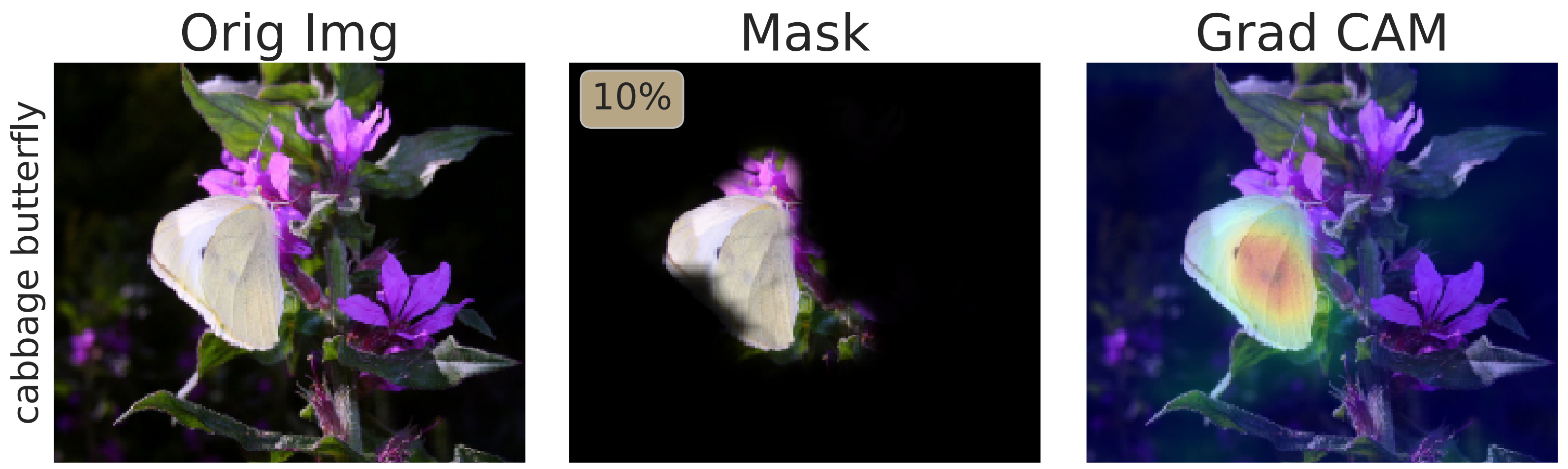
An incomplete retrospective: the first decade of interpretability



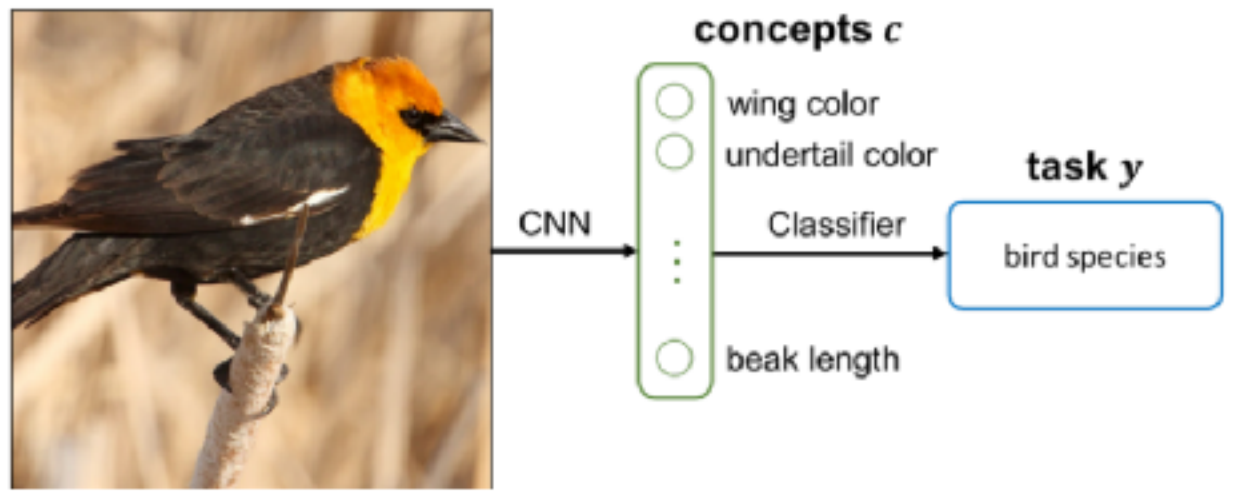
Feature visualization (2013-2018)
Activation Max., Feature Inversion,
Net Dissect, Feature Vis.

2012

2022



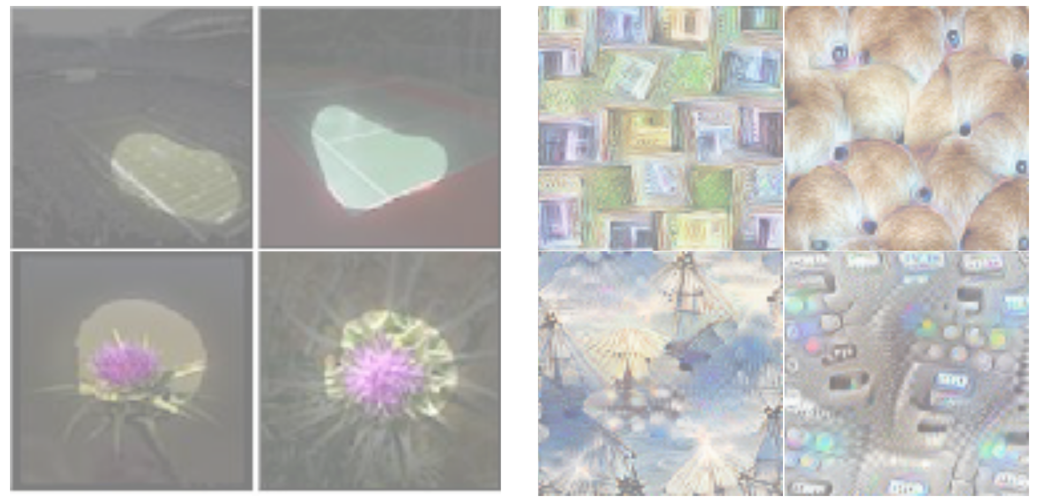
Attribution heatmaps (2013-2019)
Gradient, Grad-CAM,
Occlusion, Perturbations, RISE



Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet,
ProtoTree

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019;
Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]

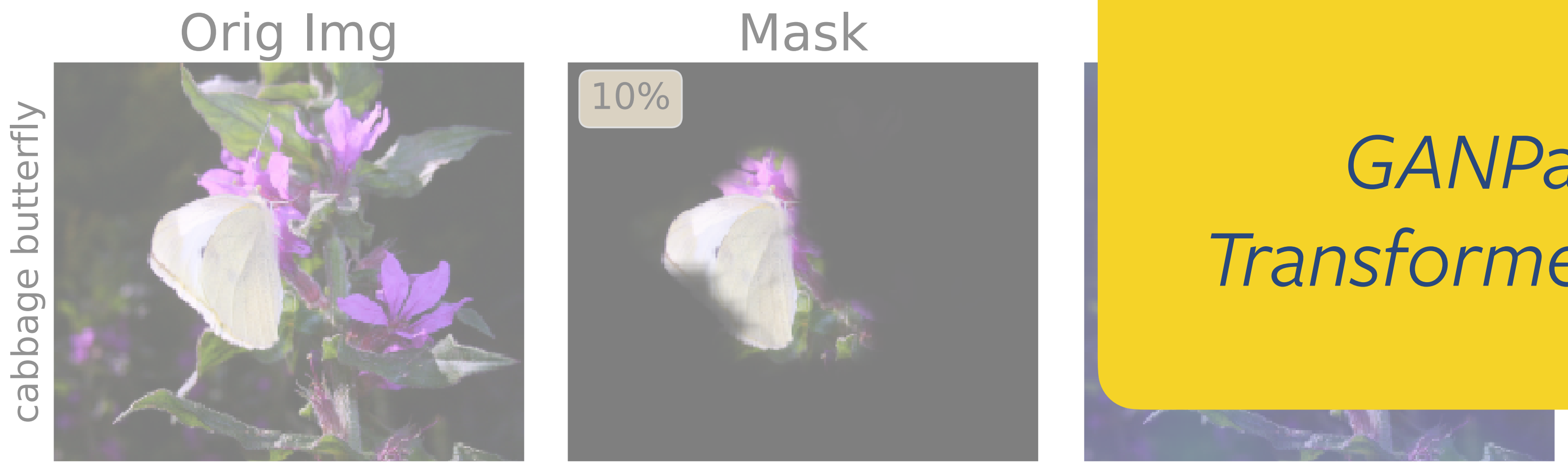
An incomplete retrospective: the first decade of interpretability



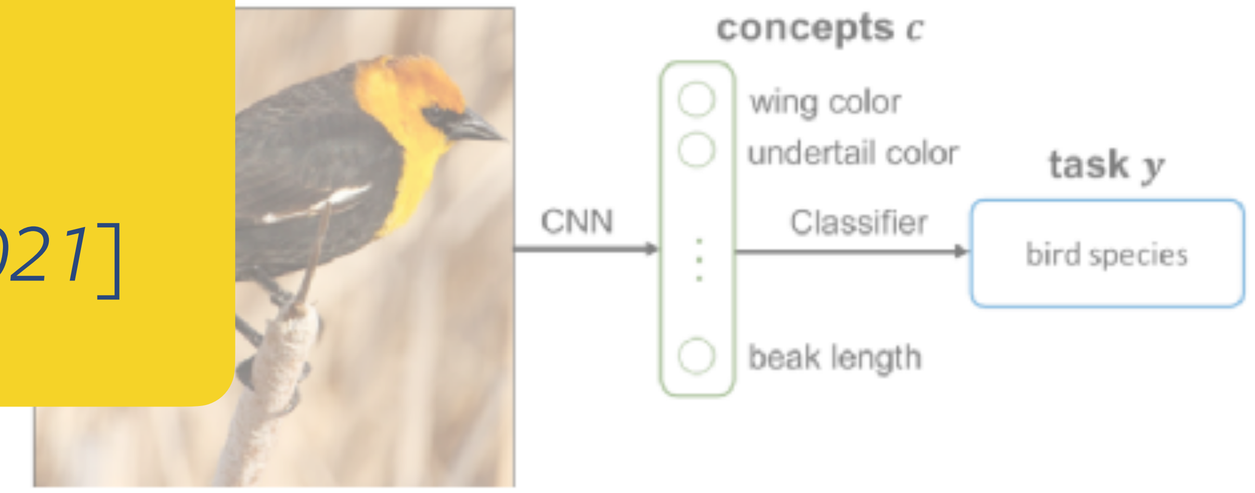
Primarily focused on understanding and approximating **CNNs**

Exceptions:
GANPaint [*Bau et al., ICLR 2019*]
Transformer Circuits [*Elhage et al., 2021*]

2012



Attribution heatmaps (2013-2019)
Gradient, Grad-CAM,
Occlusion, Perturbations, RISE



2022

Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet,
ProtoTree

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019; Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]

Directions for the next decade of interpretability

1. Develop interpretability methods for **diverse domains**
    - Beyond CNN classifiers: self-supervised learning, generative models, etc.

2. Center **humans** throughout the development process
    - In design, co-develop methods with real-world stakeholders.
    - In evaluation, measure human interpretability and utility of methods.
    - In deployment, package interpretability tools for the wider community.

Roadmap

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.
4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

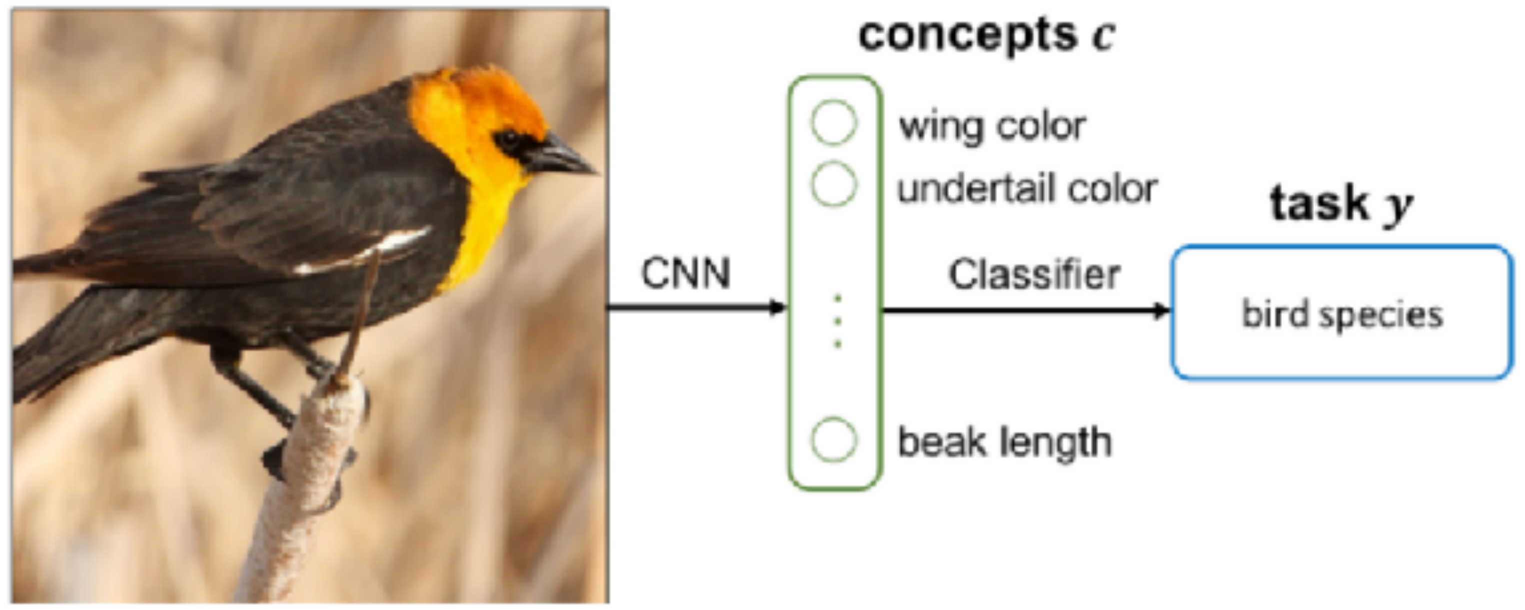
Roadmap

Sunnie S. Y. Kim

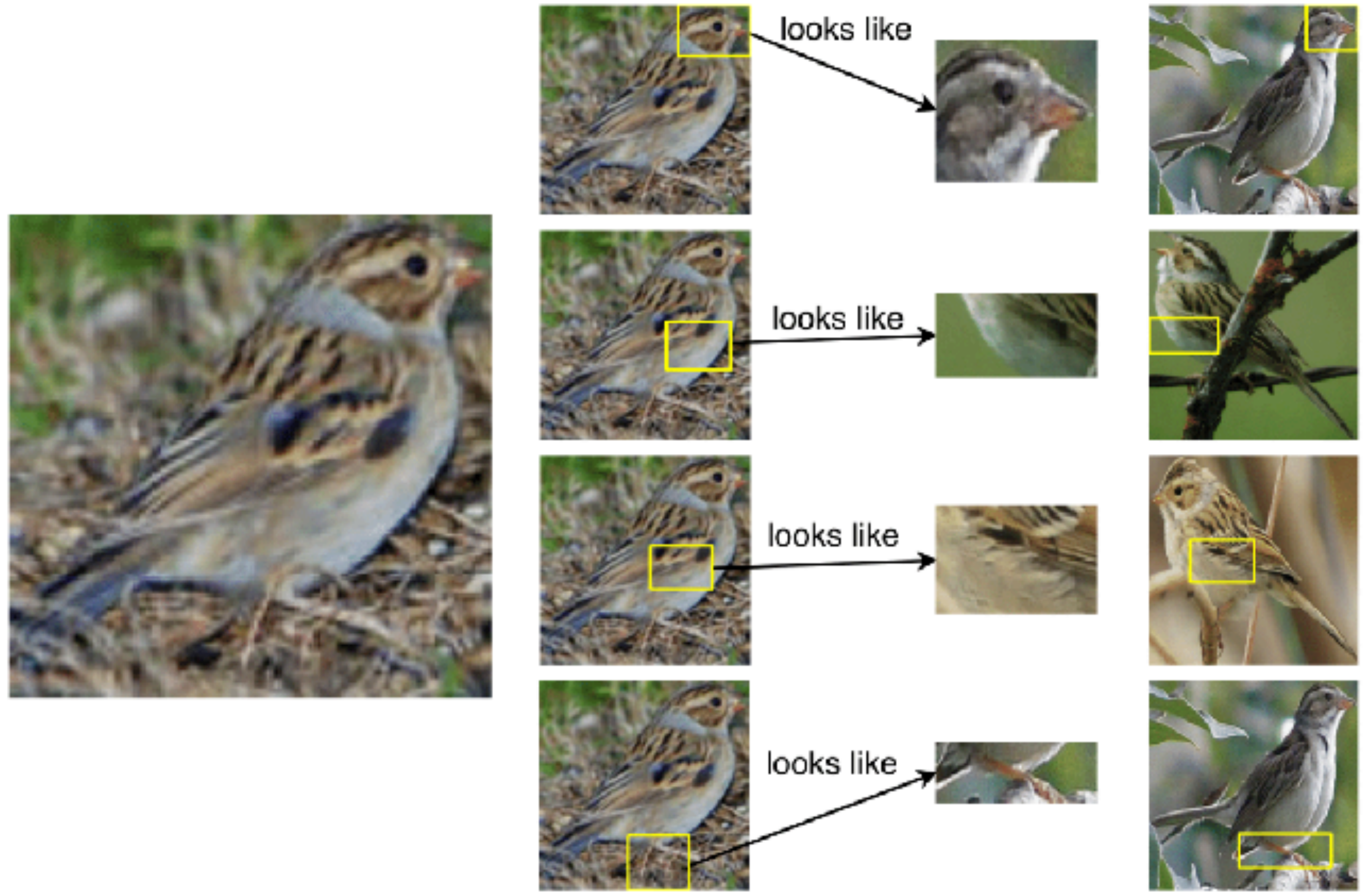
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.
4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

Explanation form factors: Why did the model predict Y?

Heatmap explanations
(e.g. Grad-CAM)



Concept-based explanations
(e.g. Concept Bottleneck)



Prototype explanations
(e.g. ProtoPNet)

Why Cardinal (L) and not
Summer Tanager (R)?

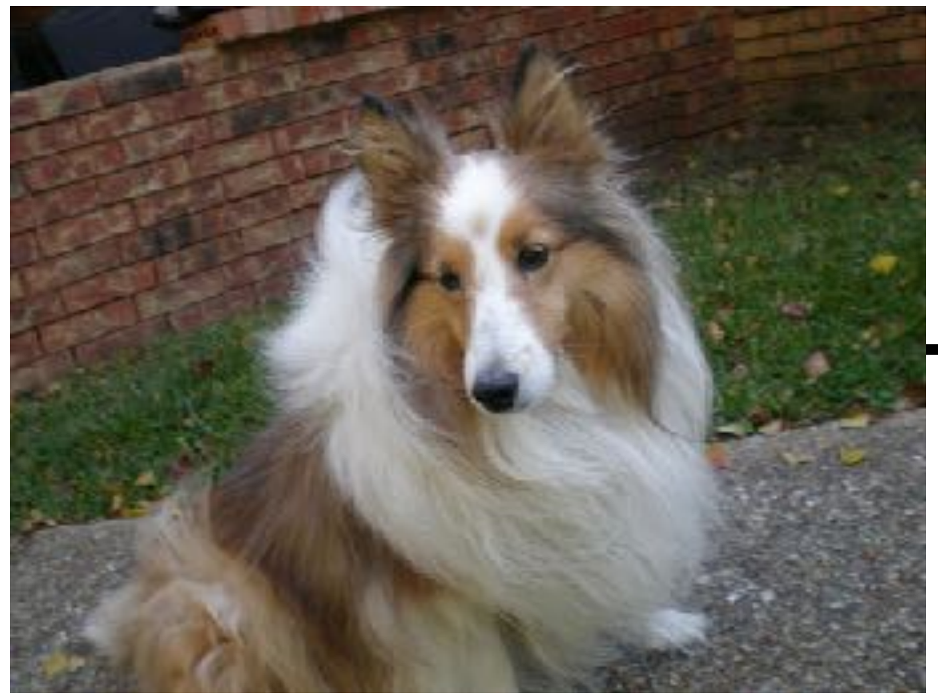
Counterfactual explanations
(e.g. SCOUT)

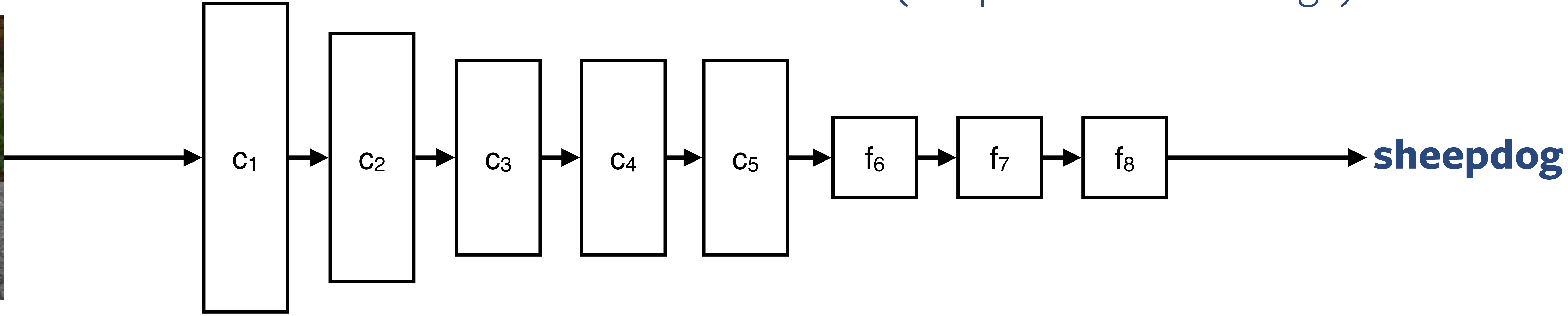
[Selvaraju et al., ICCV 2017; Koh*, Nguyen*, Tang* et al., ICML 2020; Chen* & Li* et al., NeurIPS 2019; Wang & Vasconcelos, CVPR 2020]

Post-hoc explanations

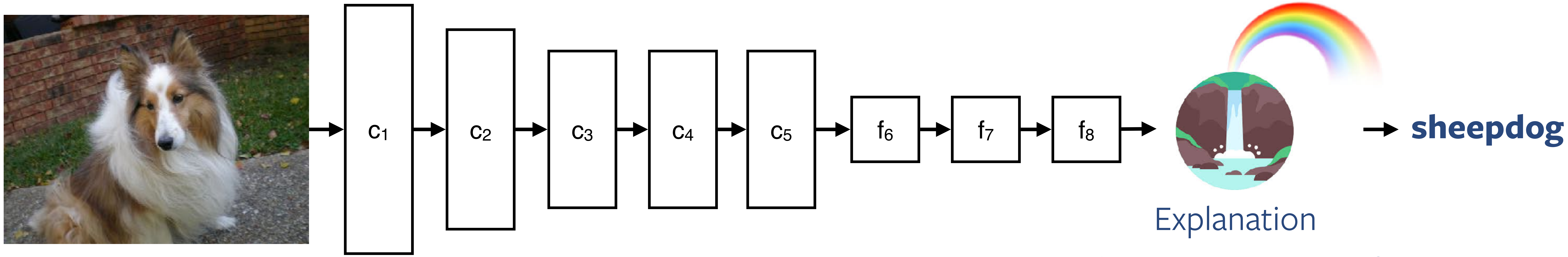
Explanation

(not part of model design)





Interpretable-by-design models



Explanation

(produced as part of model design)

Current metrics focus on heatmap evaluation

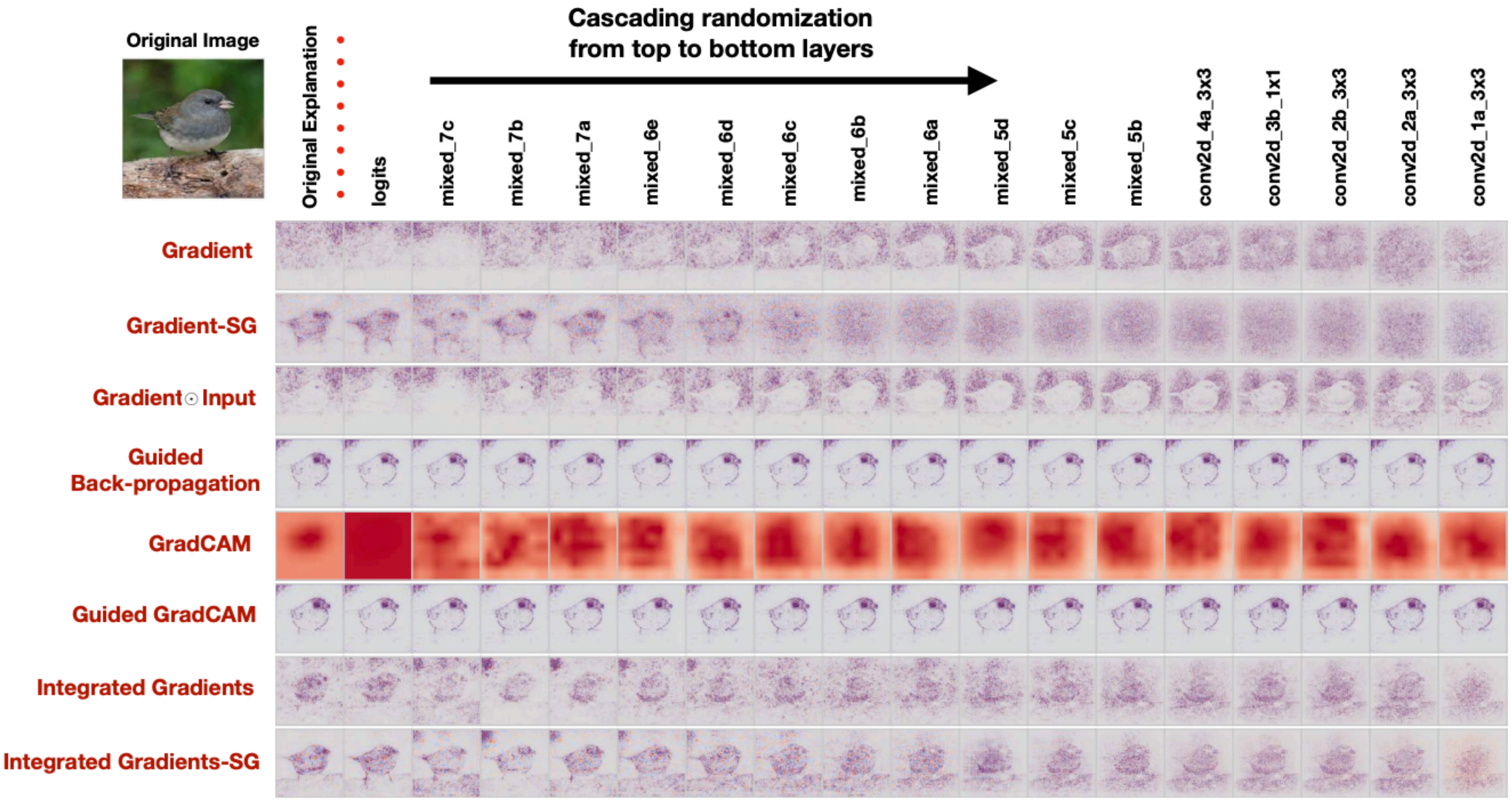
- Weak localization performance [Zhang et al., ECCV 2016]
- Perturbation analysis
  - Deletion game [Samek et al., TNNLS 2017]
  - Retrain classifiers with removed features [Hooker et al., NeurIPS 2019]
- **Sensitivity to…**
  - **output neuron [Rebuffi*, Fong*, Ji* et al., CVPR 2020]**
  - **model parameters [Adebayo et al., NeurIPS 2018]**
- …

Automatic

Selectivity to output class

[Mahendran & Vedaldi, ECCV 2016; Rebuffi et al., CVPR 2020]

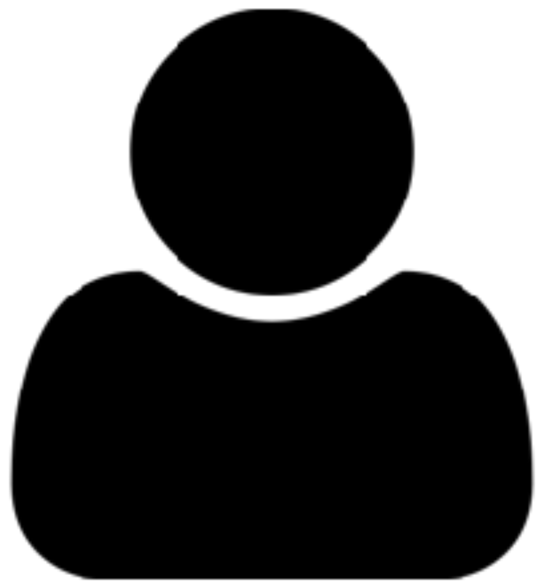
Sensitivity to model parameters (a.k.a. sanity checks)



[Adebayo et al., NeurIPS 2018]

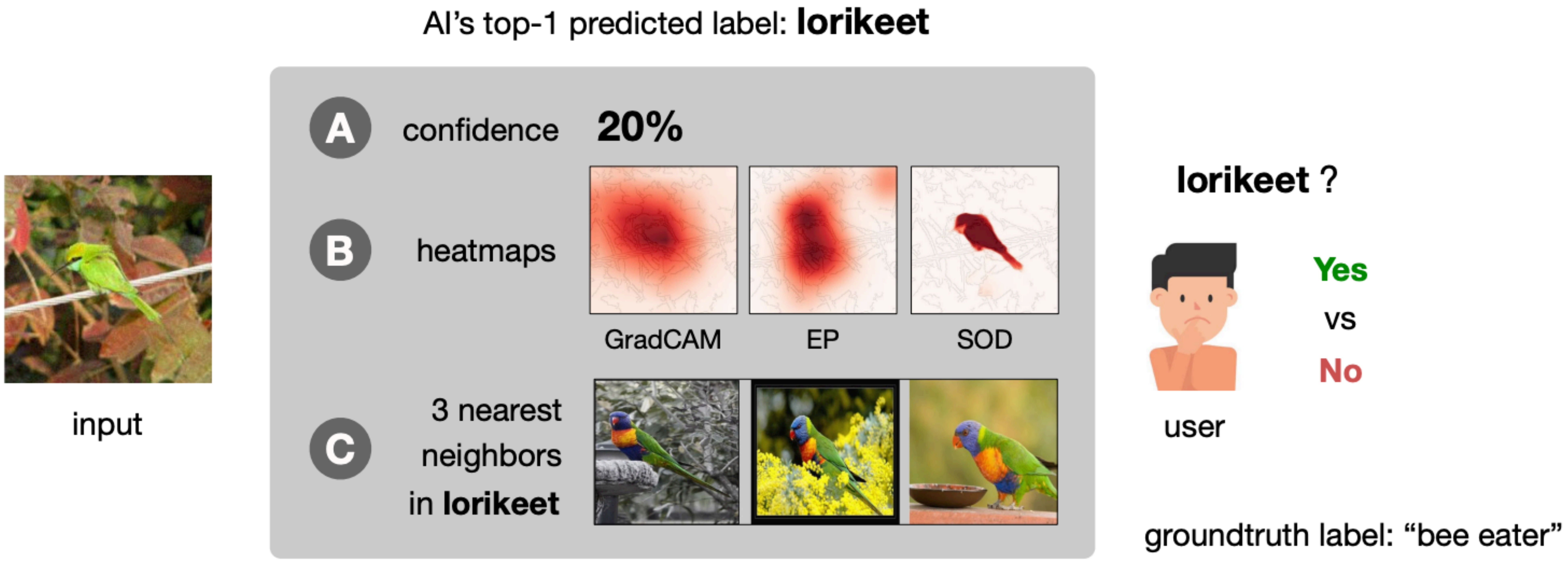
Current metrics focus on heatmap evaluation

- Sheng & Huang, HCOMP 2020
  Guess the incorrectly predicted label
- **Nguyen et al., NeurIPS 2021**
  Is this prediction correct?
- Colin* & Fel* et al., arXiv 2021
  What did the model predict (choose one of two)?



Human

Is this prediction correct?



[Nguyen et al., NeurIPS 2021]

HIVE: Evaluating the Human Interpretability of Visual Explanations

1. Within method → **Cross-method comparison**
2. Automated evaluation → **Human-centered evaluation**
3. Intuition-based reasoning → **Falsifiable hypothesis testing**

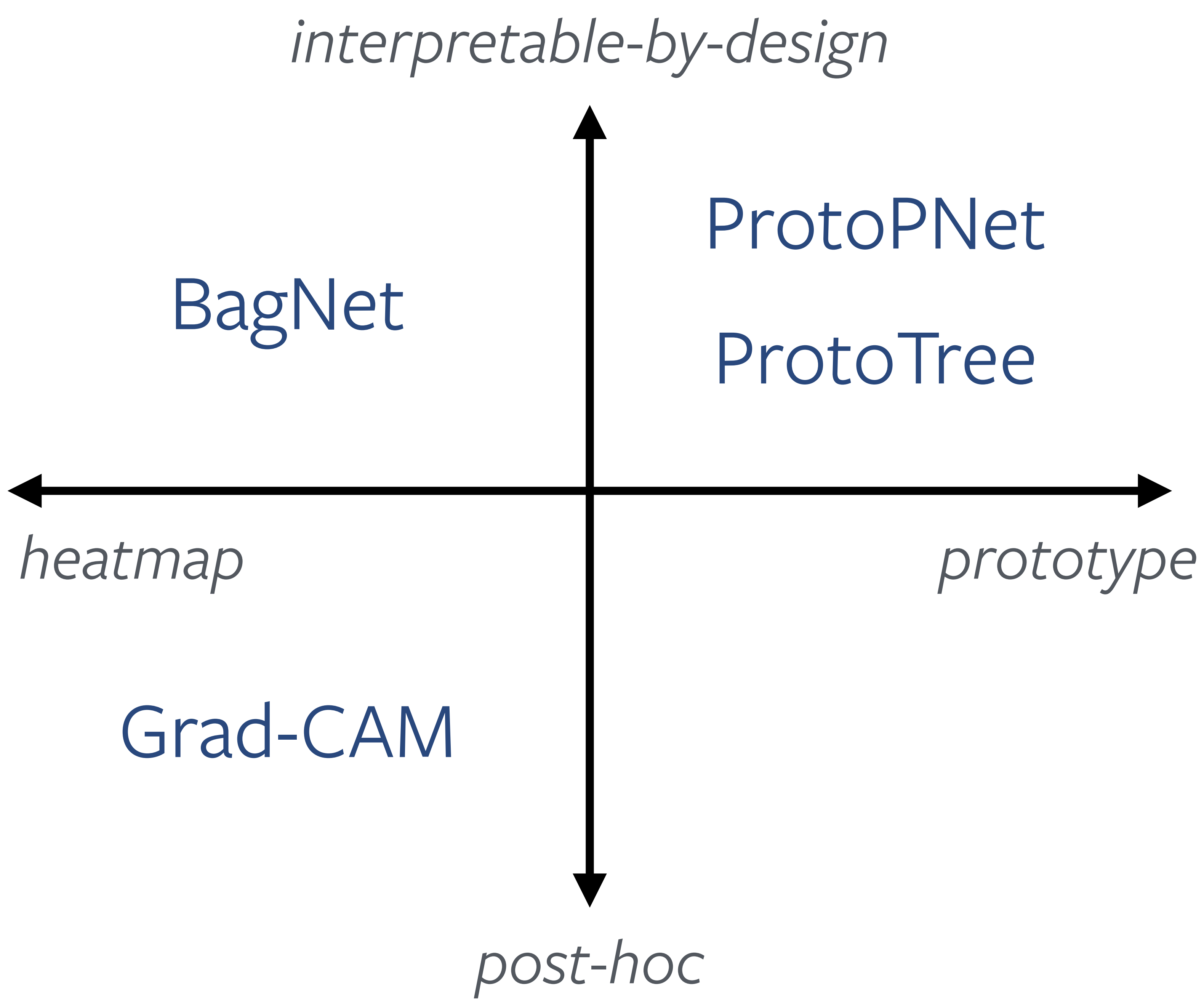
Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
HIVE: Evaluating the Human Interpretability of Visual Explanations.

Our contributions

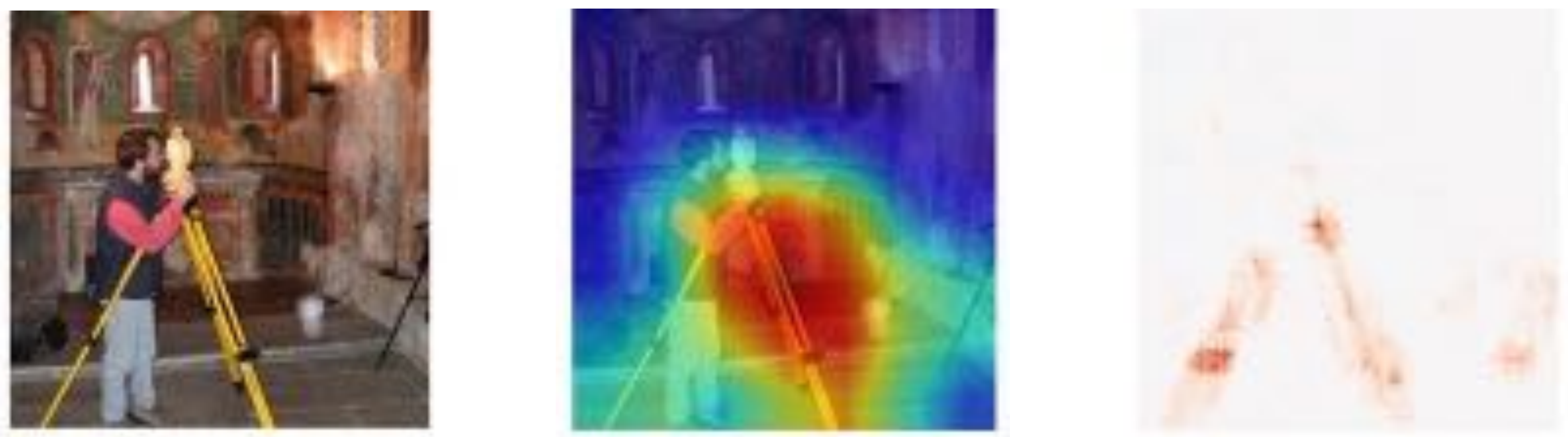
- Novel human study design for evaluating 4 diverse interpretability methods
  - **First human study** for interpretable-by-design and prototype methods
- Quantify the utility of explanations in distinguishing between **correct and incorrect predictions**
- Quantify how users would trade off between **interpretability and accuracy**
- **Open-source** HIVE studies to encourage reproducible research

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]

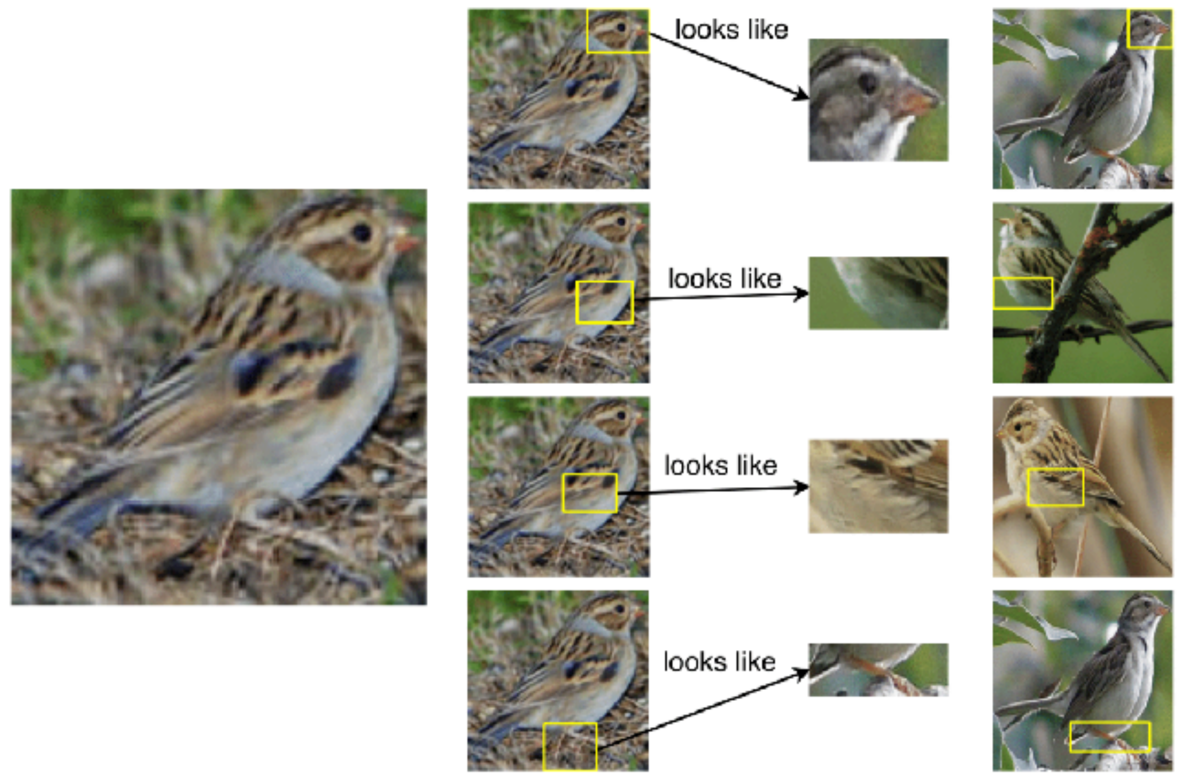
1. Cross-method comparison



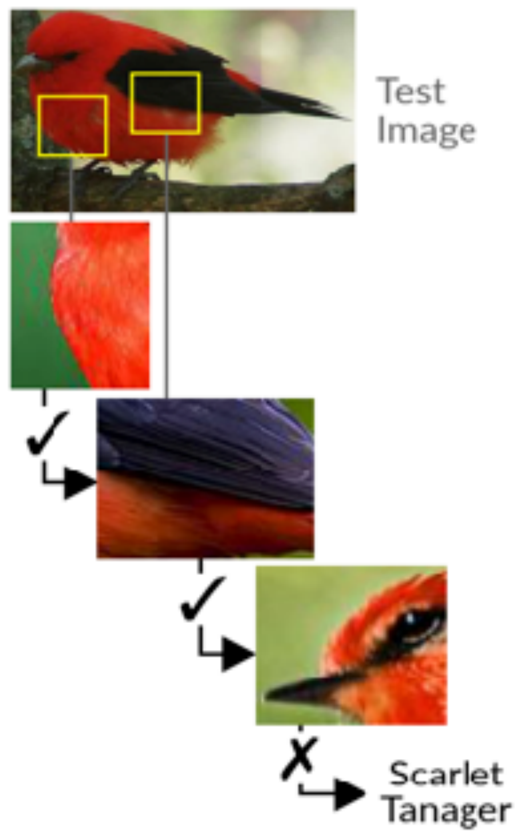
Grad-CAM BagNet



ProtoPNet



ProtoTree



[Selvaraji et al., ICCV 2017; Brendel & Bethge, ICLR 2019; Chen* & Li* et al., NeurIPS 2019, Nauta et al., CVPR 2021]

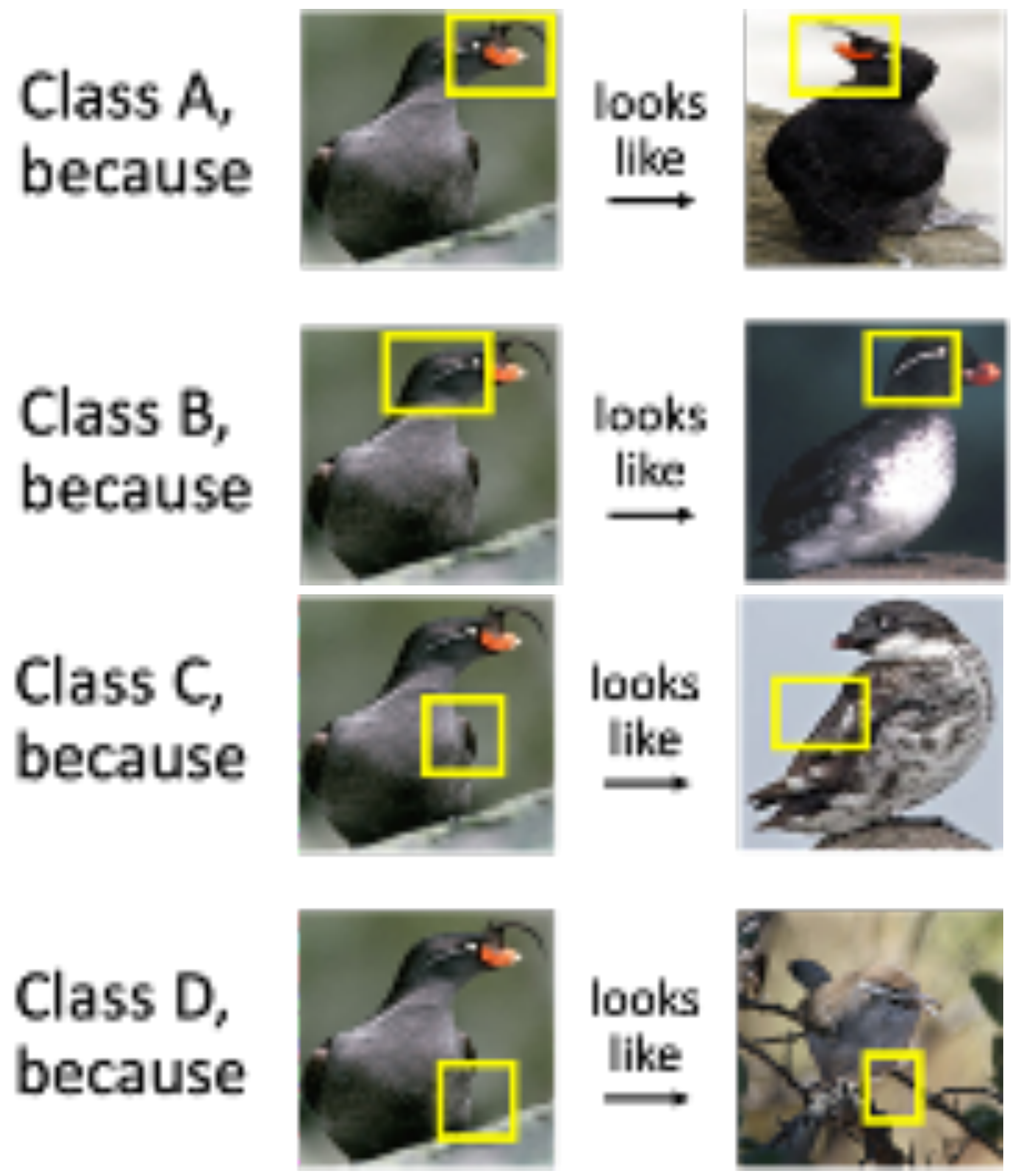
2. Human-centered evaluation

Agreement task

How confident are you in the model's prediction?

Distinction task

Which class do you think is correct?



Experimental set-up: AMT studies with N=50 participants each

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]

2. Human-centered evaluation

Agreement task

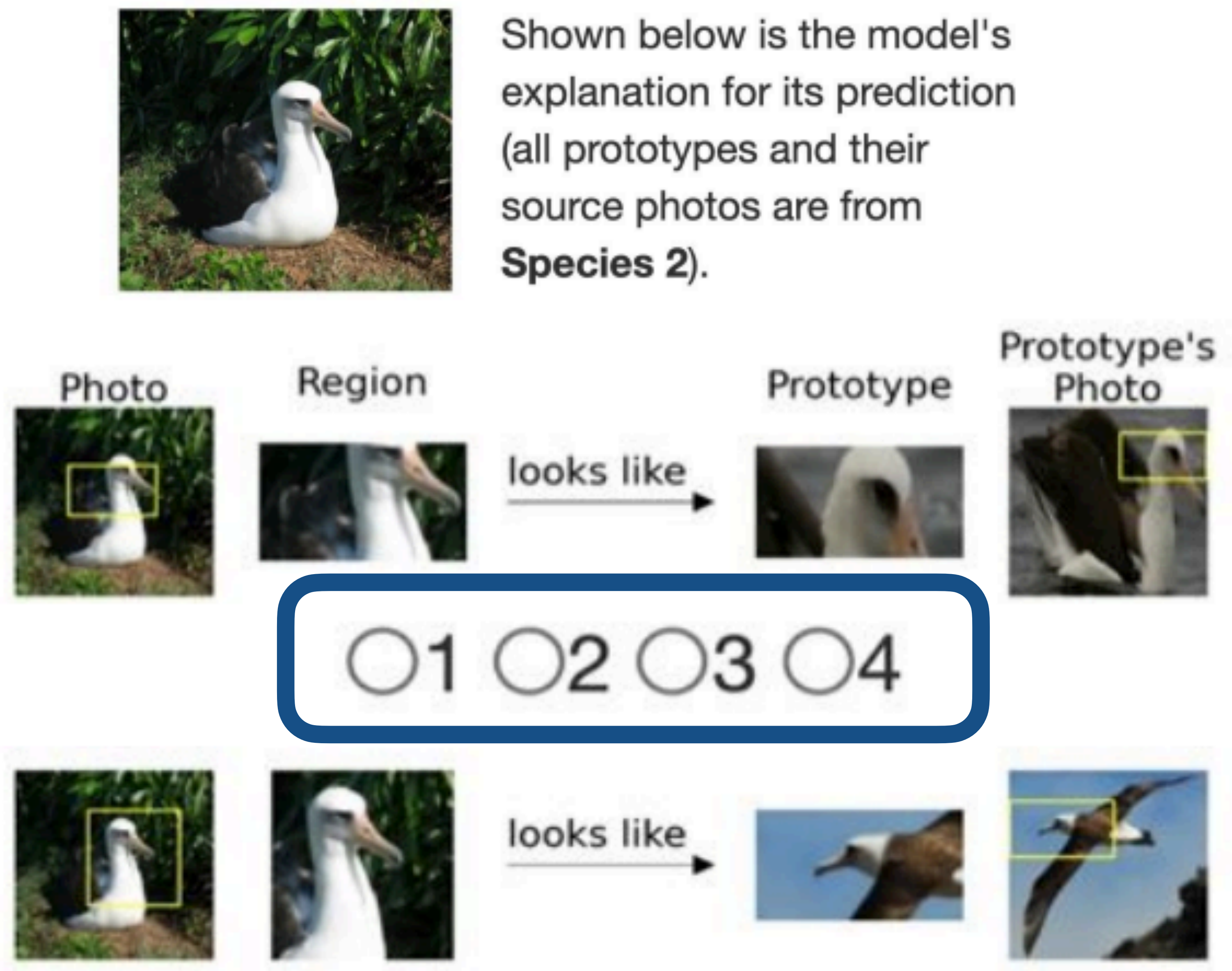
How confident are you in the model's prediction?

Finding #1: Prototype similarities often **do not align** with human notions of similarity.

ProtoPNet and ProtoTree only

Task: Rate the similarity of each row's prototype-region pair on a scale of 1-4.

(1: Not Similar, 2: Somewhat Not Similar, 3: Somewhat Similar, 4: Similar)



◯1 ◯2 ◯3 ◯4

Q. What do you think about the model's prediction?

- ◯ Fairly confident that prediction is *correct*
- ◯ Somewhat confident that prediction is *correct*
- ◯ Somewhat confident that prediction is incorrect
- ◯ Fairly confident that prediction is incorrect

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]

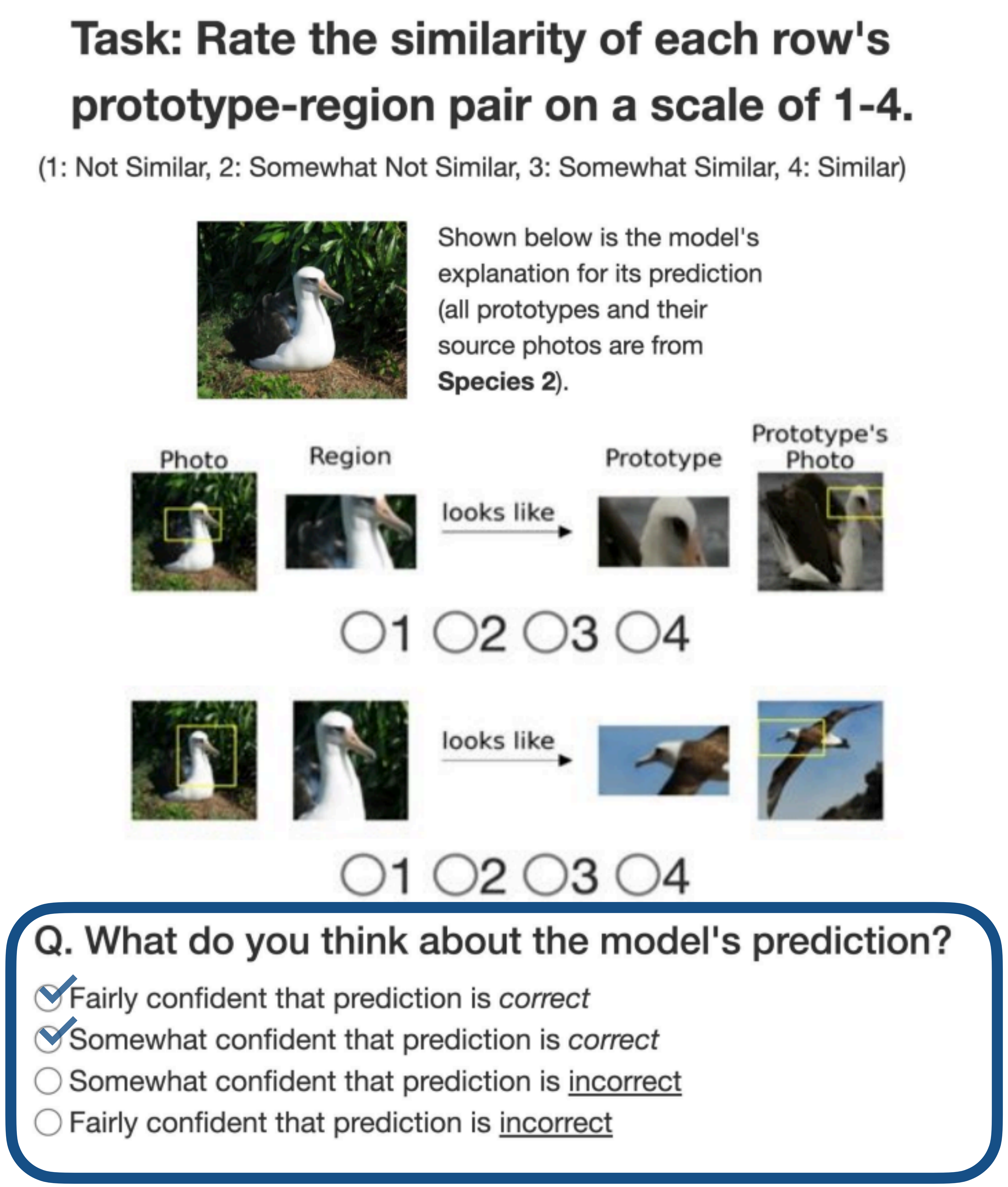
2. Human-centered evaluation

Agreement task

How confident are you in the model's prediction?

Finding #1: Prototype similarities often **do not align** with human notions of similarity.

Finding #2: Agreement task reveals **confirmation bias**.



[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Chen* & Li* et al., NeurIPS 2019]

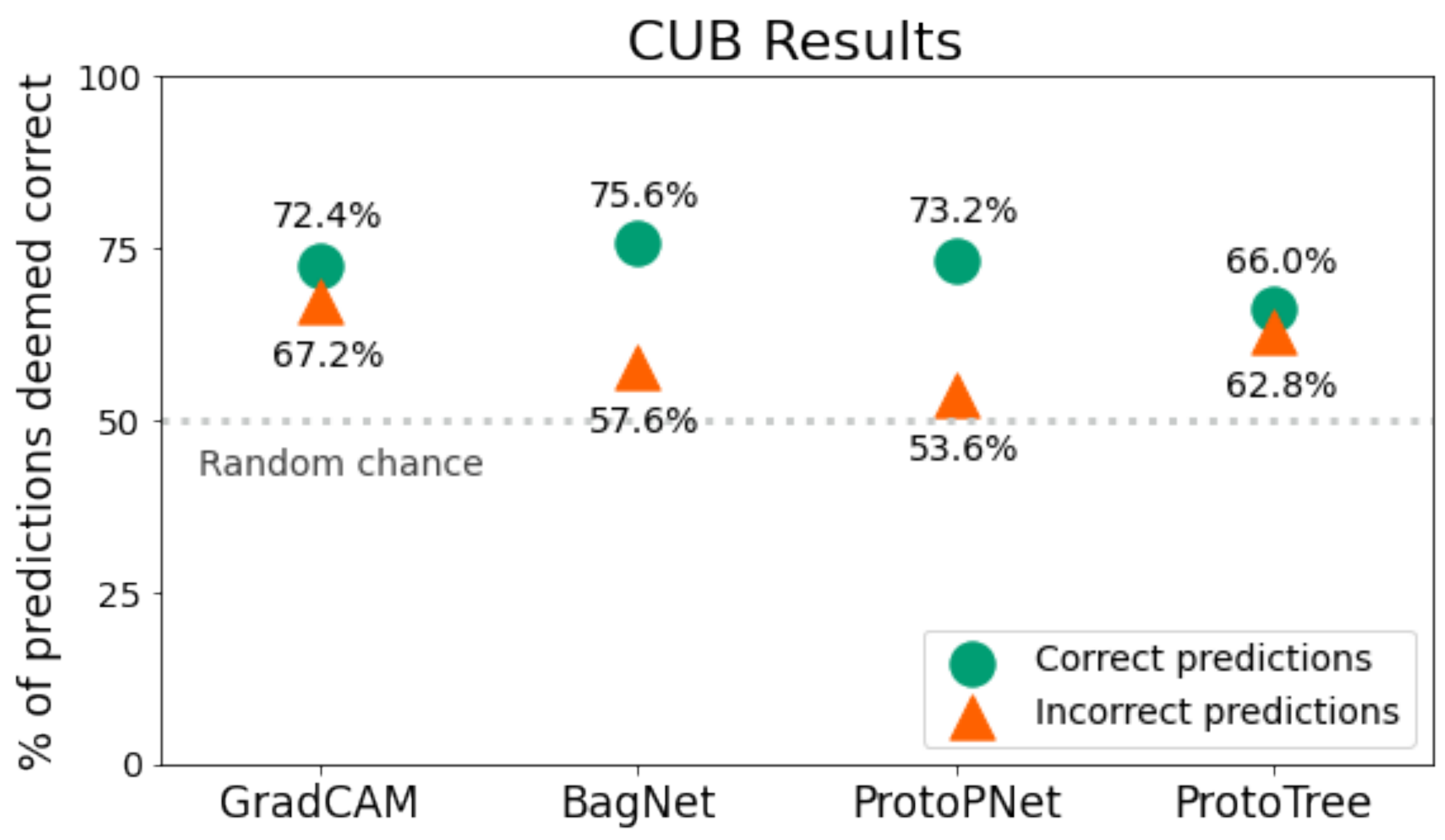
2. Human-centered evaluation

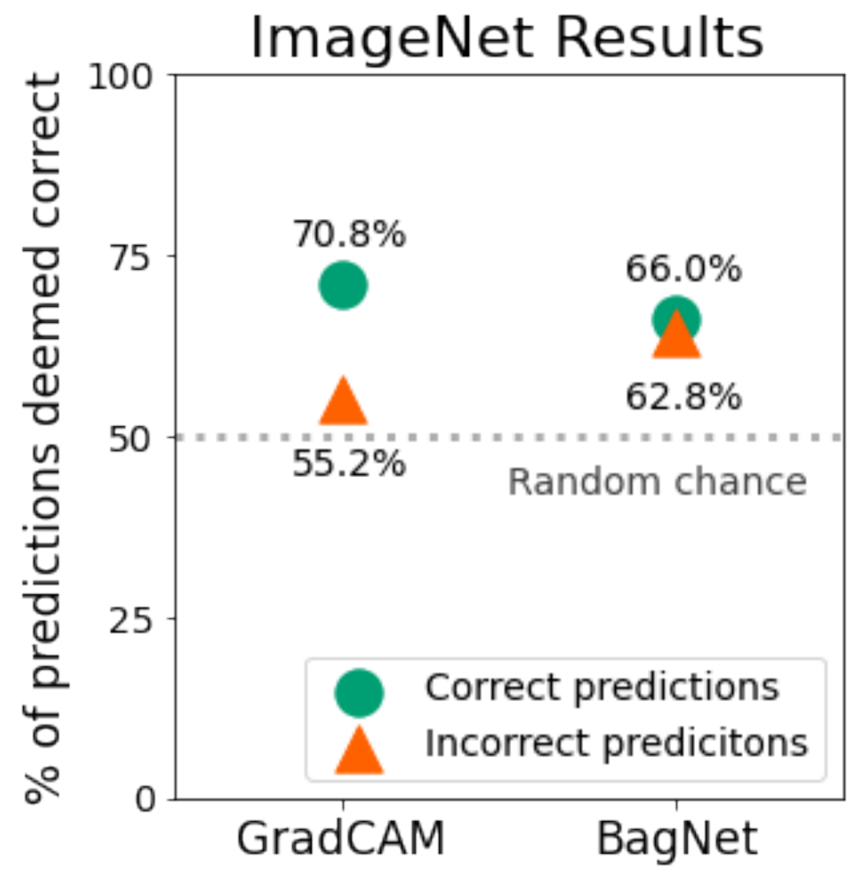
Agreement task

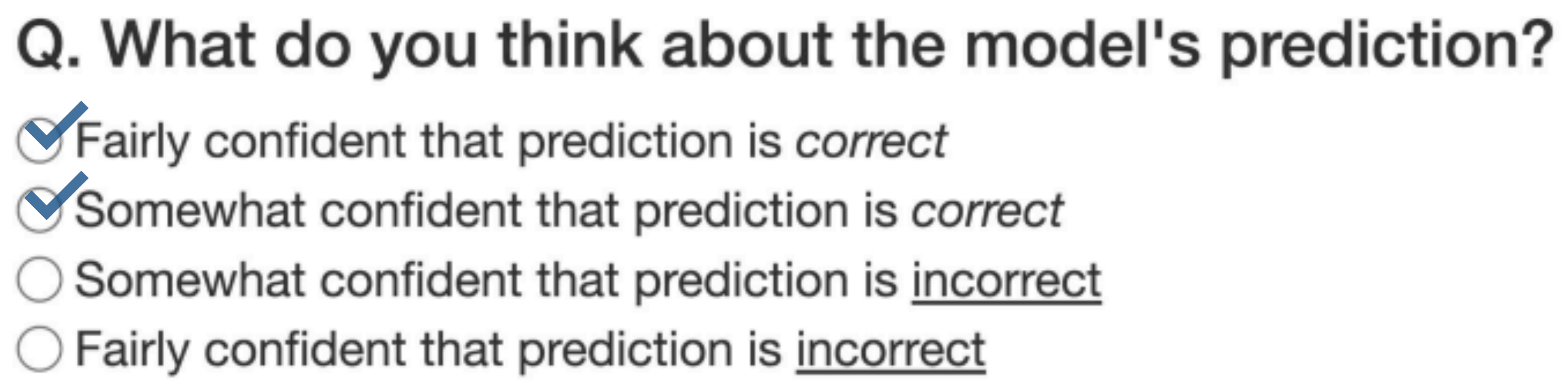
How confident are you in the model's prediction?

Finding #1: Prototype similarities often **do not align** with human notions of similarity.

Finding #2: Agreement task reveals **confirmation bias**.







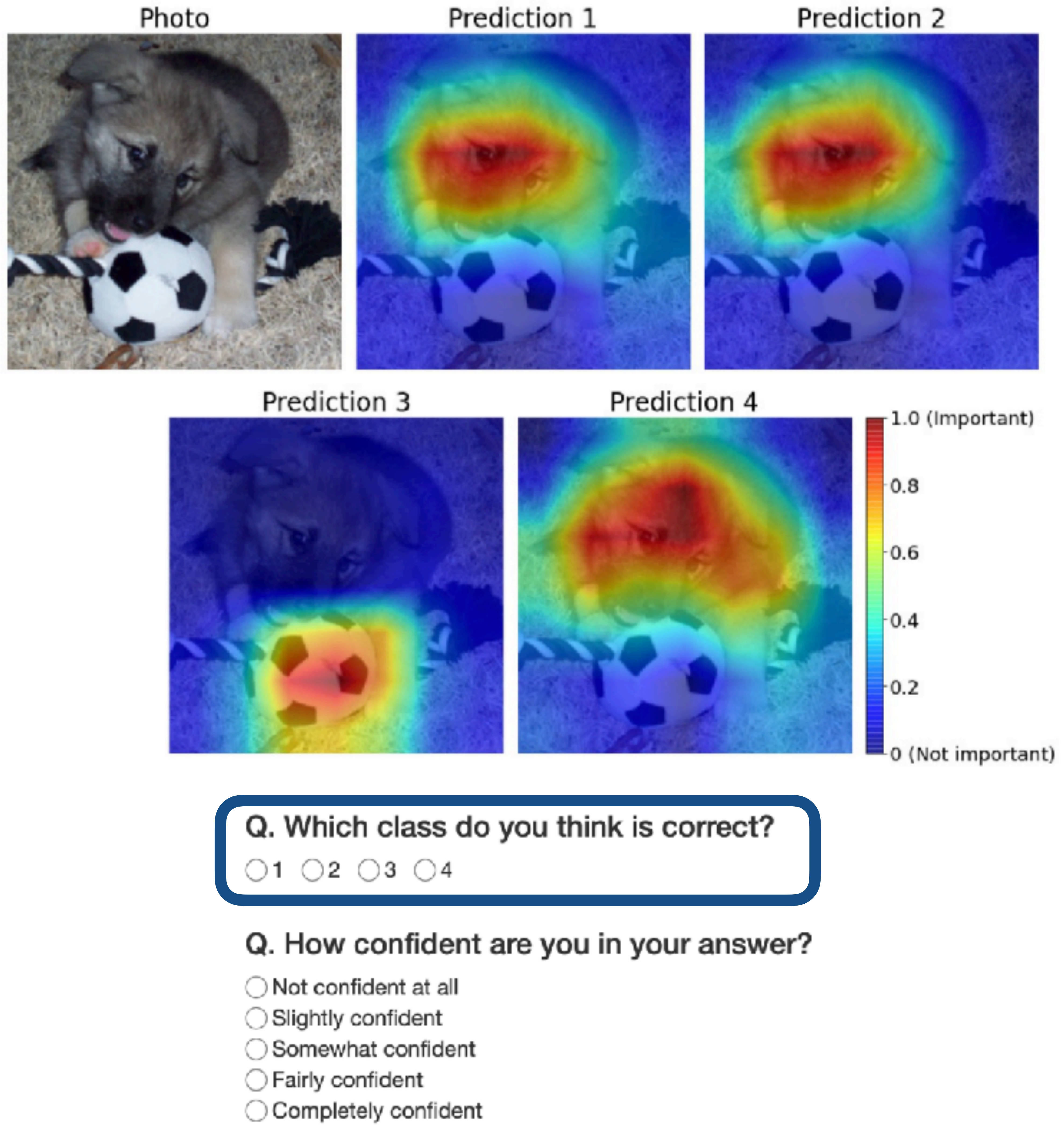
[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]

2. Human-centered evaluation

Distinction task

Which class do you think is correct?

Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.



[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.; Selvaraju et al., ICCV 2017]

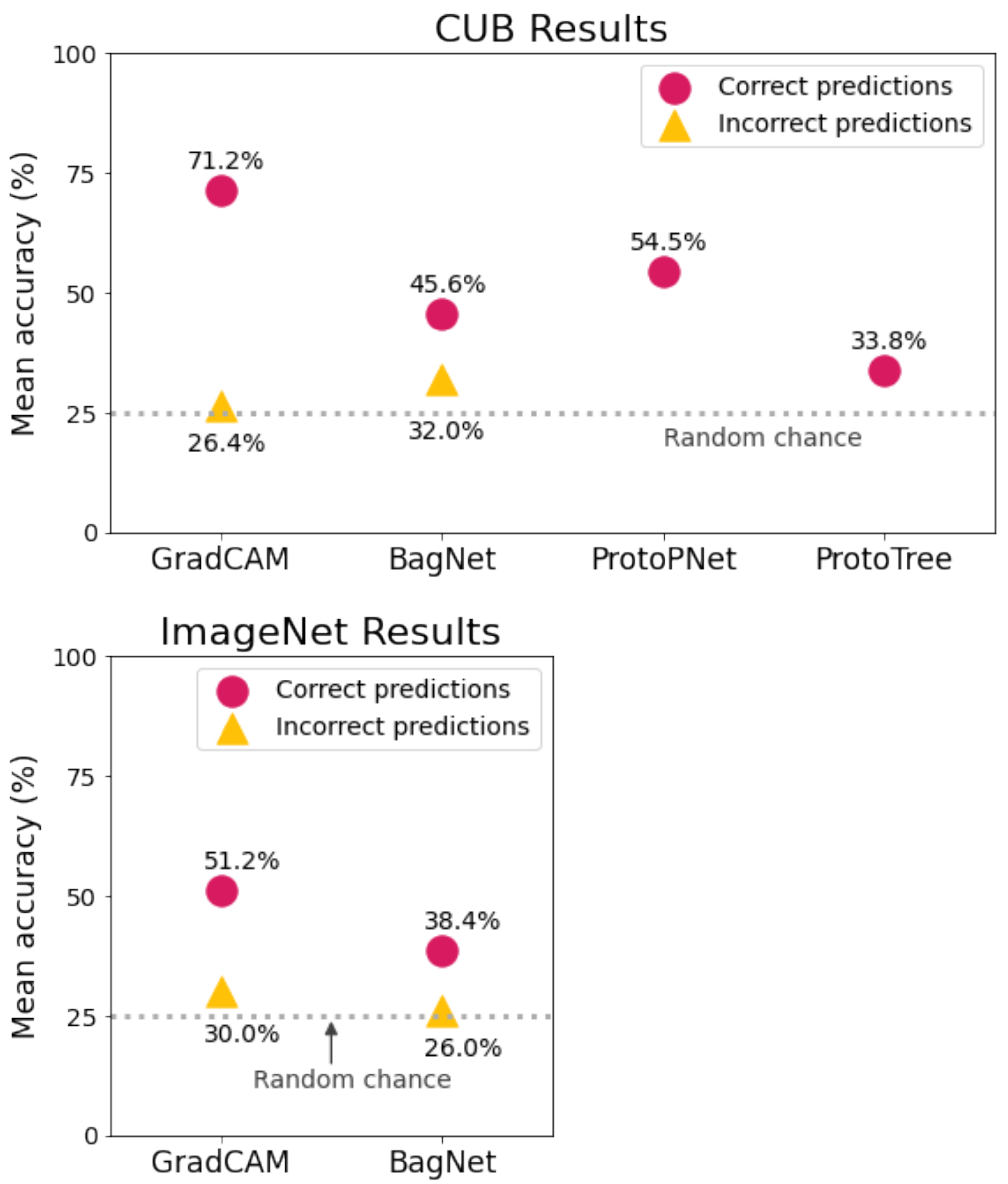
2. Human-centered evaluation

Distinction task

Which class do you think is correct?

Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.

Goal: Interpretability should help humans identify and explain model errors.



Q. Which class do you think is correct?

◯1 ◯2 ◯3 ◯4

Q. How confident are you in your answer?

◯ Not confident at all
◯ Slightly confident
◯ Somewhat confident
◯ Fairly confident
◯ Completely confident

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]

3. Falsifiable hypothesis testing

Finding #1: Prototype similarities often **do not align** with human notions of similarity.

Finding #2: Agreement task reveals **confirmation bias**.

Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]

3. Falsifiable hypothesis testing

Finding #1: Prototype similarities often **do not align** with human notions of similarity.

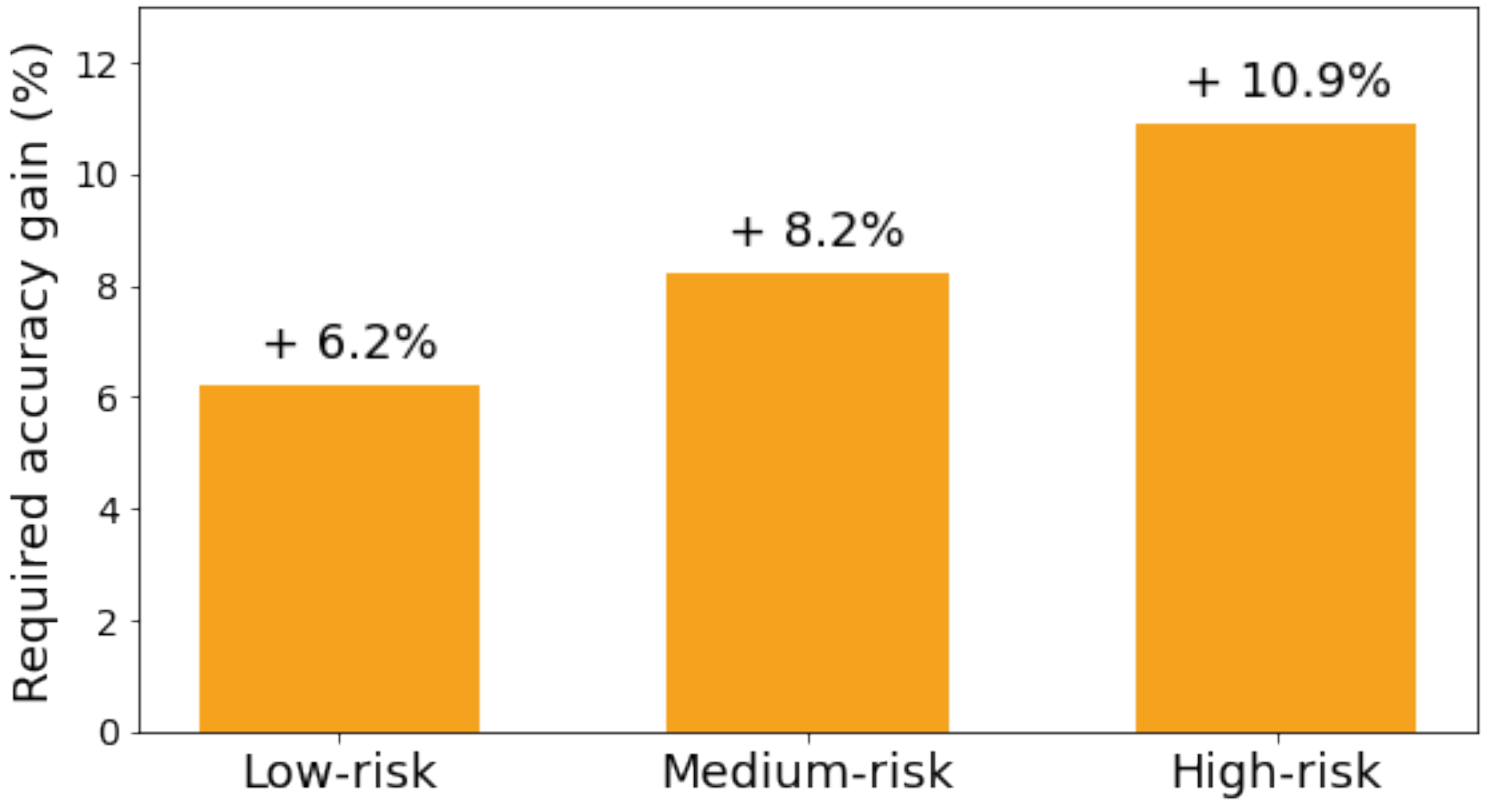
Finding #2: Agreement task reveals **confirmation bias**.

Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.

Finding #4: Participants prefer interpretability over accuracy, esp. in high-risk settings.

Interpretability-accuracy tradeoff

Q: What is the minimum accuracy of a baseline model that would convince you to use it over a model with explanations?



(e.g. educational purposes) (e.g. biodiversity monitoring) (e.g. veterinary medicine)

[Sunnie S. Y. Kim et al., arXiv 2021. HIVE.]

Challenges for human evaluation

- Skill cost: web development skills
- Financial cost: budget for AMT experiments
- Time cost: human study design and iteration (e.g. task feasibility, IRB approval, quality control)

Takeaway: As a research community, invest in and reward human evaluation studies (like dataset development).

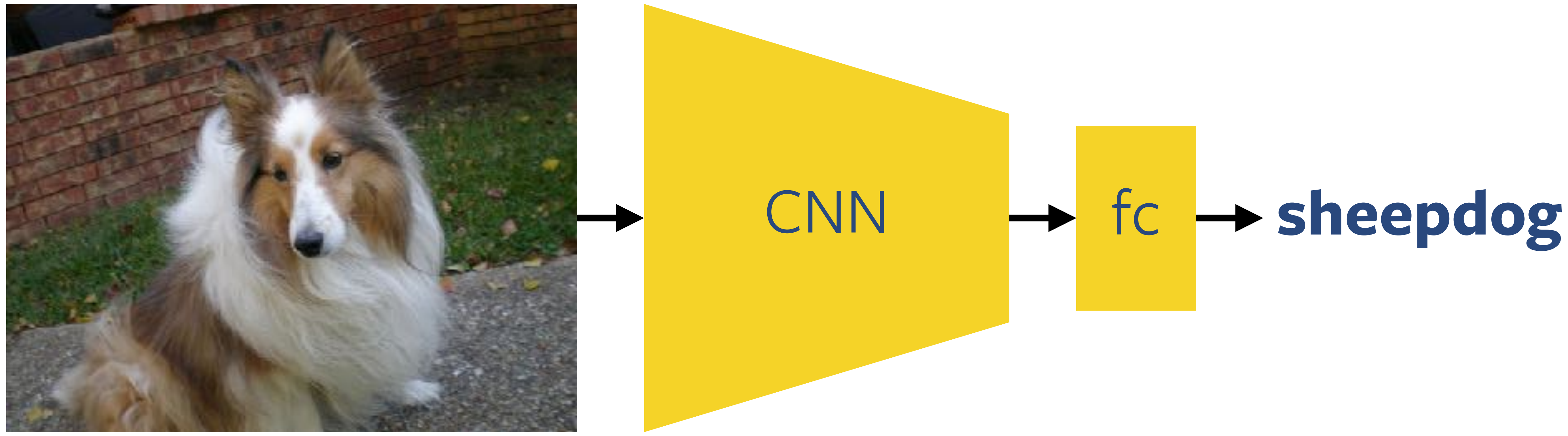
Roadmap

Vikram V. Ramaswamy

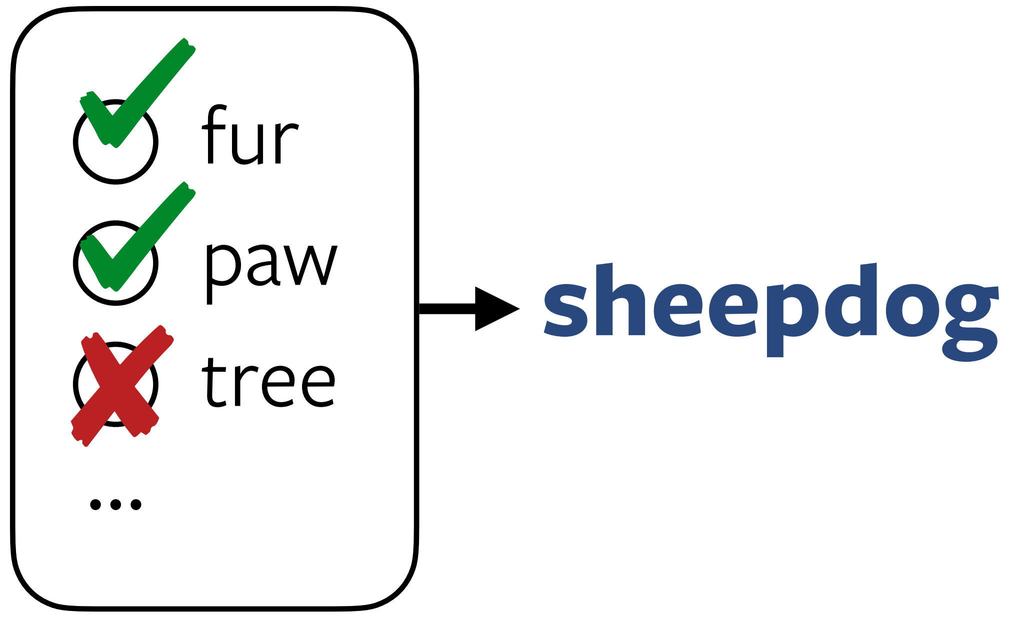
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.
4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

Concept-based explanations

Why did the model predict **sheepdog**?

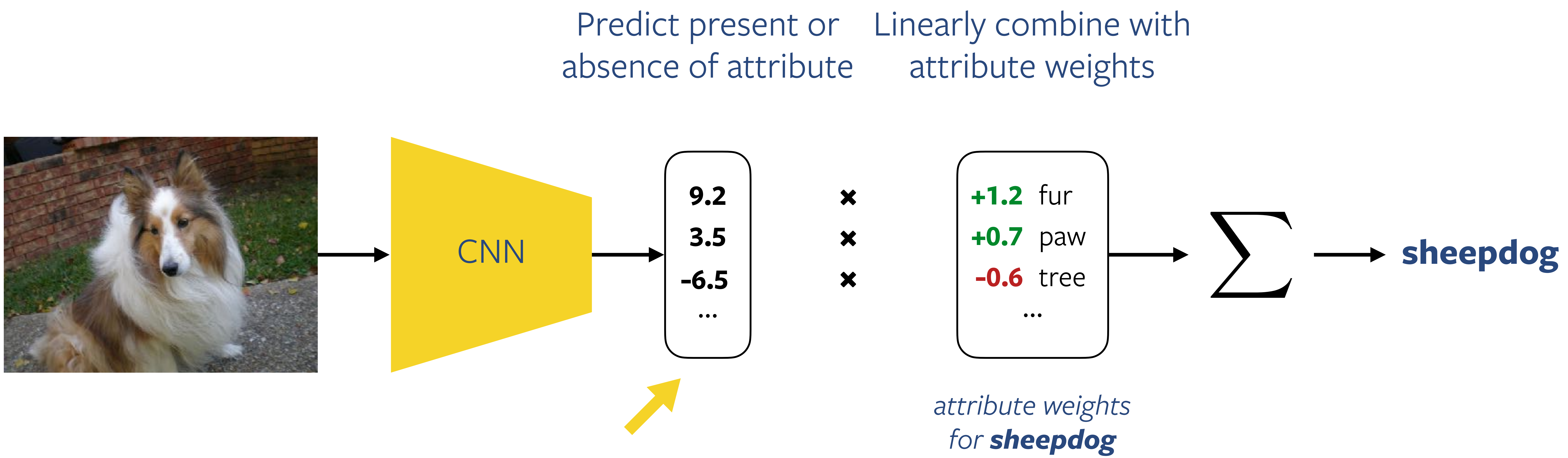


Concept-based explanation



Pro: Labelled concepts are interpretable to humans

Concept Bottleneck: Linear Combination of Labelled Attributes

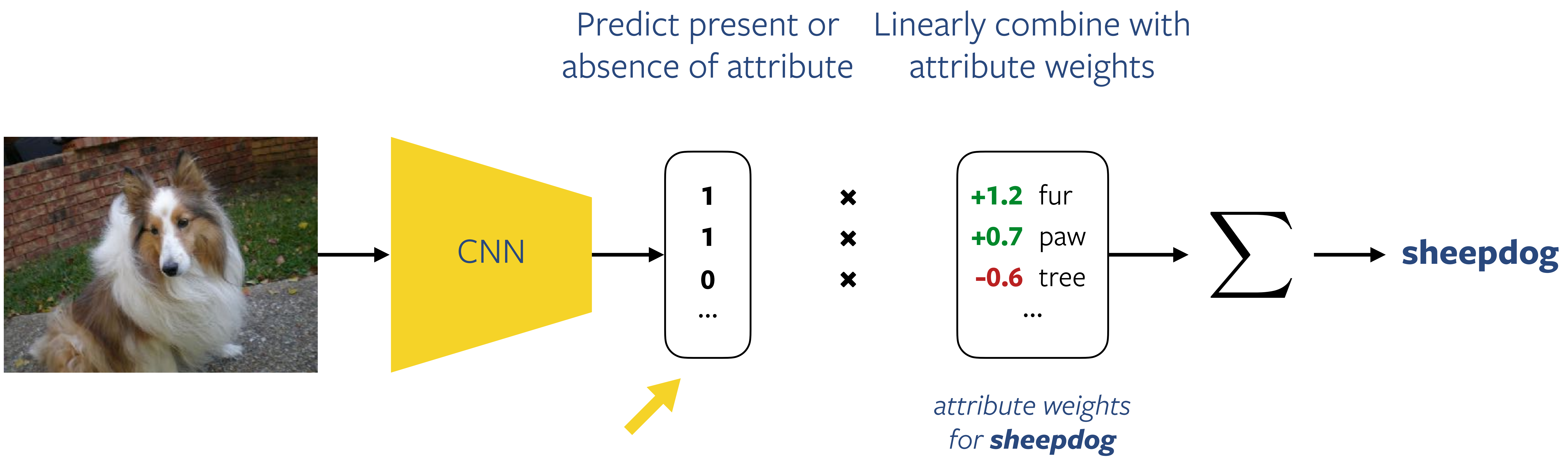


Con: Problems with predicting fractional values

- hard to interpret
- can encode hidden information

[Koh*, Nguyen*, Tang* et al., ICML 2020]

Concept Bottleneck: Linear Combination of Labelled Attributes

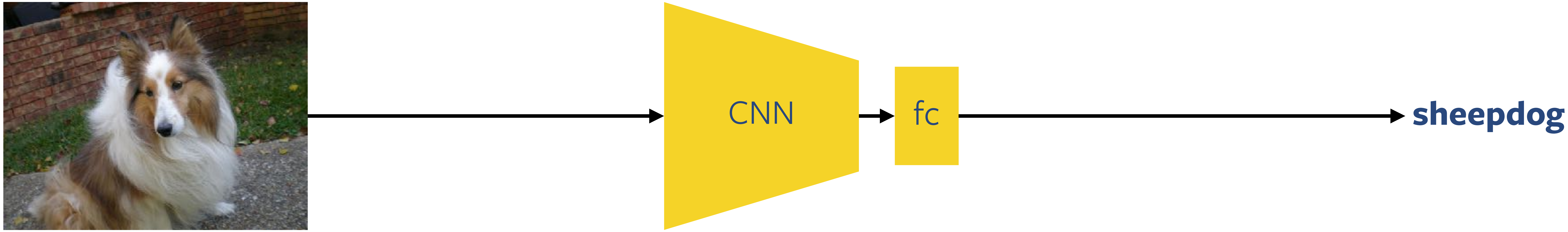


Con: Problems with predicting fractional values

- hard to interpret
- can encode hidden information

[Koh*, Nguyen*, Tang* et al., ICML 2020]

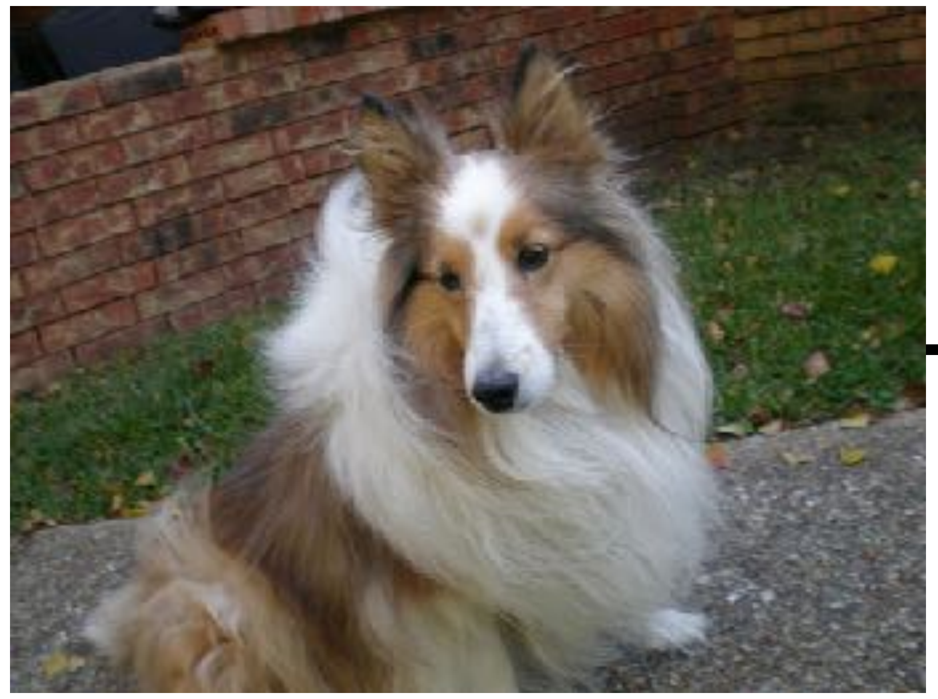
ELUDE: **E**xplanation via a **L**abelled and **U**nlabelled **DE**composition of features



Goal: Approximate behavior of original CNN

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

ELUDE: Decomposition of labelled and unlabelled features



Goal: Approximate behavior of original CNN

1. Linearly combine **ground-truth, labelled attributes**
2. Learn remaining **unlabelled features as low-rank space**

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

Attributes only: % of model explained via labelled attributes decreases as task complexity increases

Task	% Explained
2-way scene classification (indoor vs. outdoor)	95.7
16-way scene classification (home/hotel, workplace, etc.)	46.2
365-way scene classification (airfield, bowling alley, etc.)	28.8

Without fractional values encoding hidden information, attribute-only approaches are limited.

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

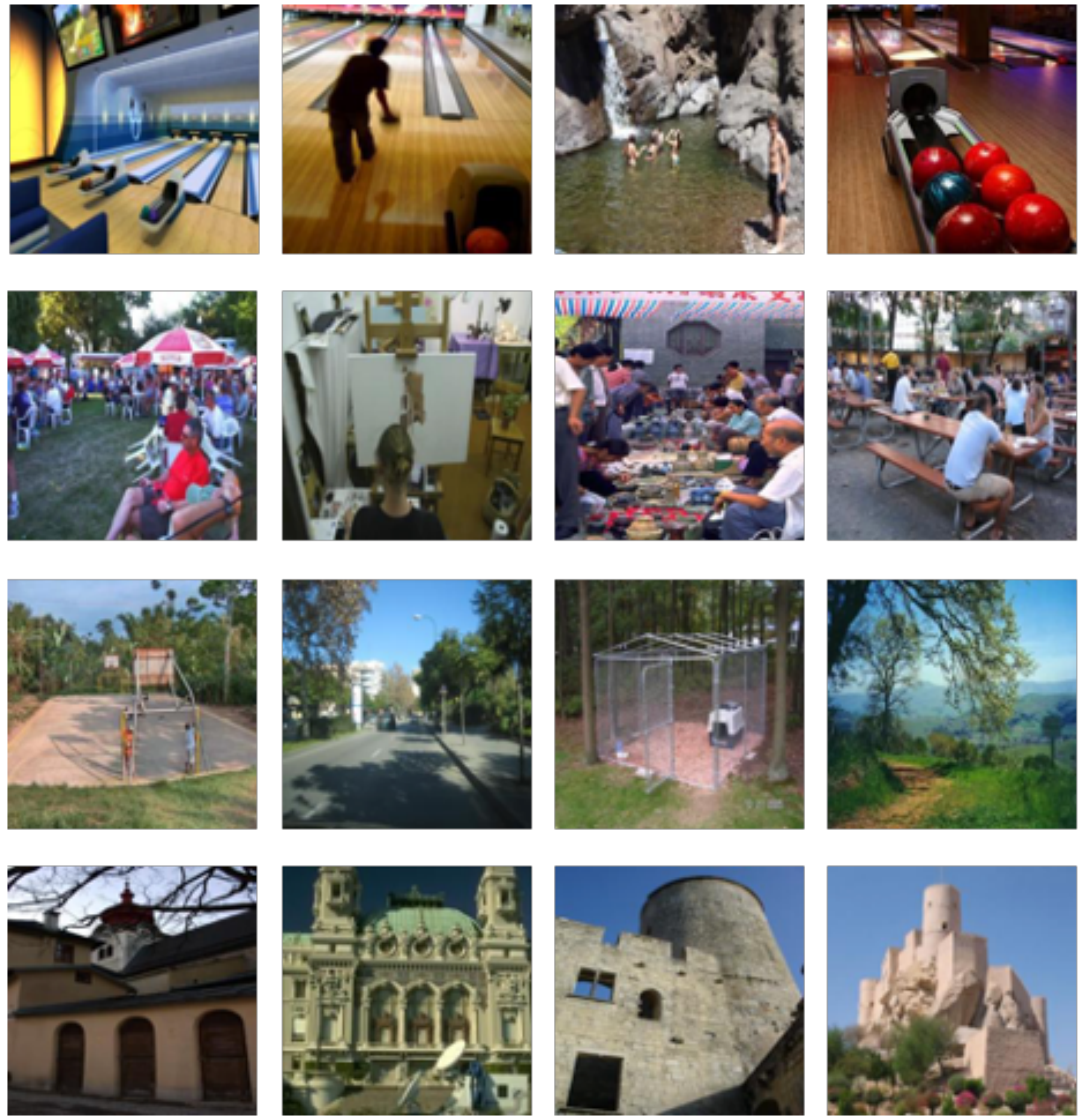
Attributes only: % of model explained via labelled attributes decreases as task complexity increases

Scene group	TPR
home/hotel	99.0
comm-buildings/towns	93.5
water/ice/snow	60.6
forest/field/jungle	40.2
workplace	14.2
shopping-dining	12.4
cultural/historical	6.5
cabins/gardens/farms	4.7
outdoor-transport	3.2
indoor-transport	0.0
indoor-sports/leisure	0.0
indoor-cultural	0.0
mountains/desert/sky	0.0
outdoor-manmade	0.0
outdoor-fields/parks	0.0
industrial-construction	0.0

Without fractional values encoding hidden information, attribute-only approaches are limited.

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

Features + attributes: Unlabelled features correspond to human-interpretable concepts



bowling alleys?

people eating?

outdoor sports fields?

castle-like buildings?

Scene group	TPR
home/hotel	99.0
comm-buildings/towns	93.5
water/ice/snow	60.6
forest/field/jungle	40.2
workplace	14.2
shopping-dining	12.4
cultural/historical	6.5
cabins/gardens/farms	4.7
outdoor-transport	3.2
indoor-transport	0.0
indoor-sports/leisure	0.0
indoor-cultural	0.0
mountains/desert/sky	0.0
outdoor-manmade	0.0
outdoor-fields/parks	0.0
industrial-construction	0.0

attributes only

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

Challenges for concept-based methods

- Attributes-only approaches are incomplete
- Develop more methods to explain the "remainder"
  - Interpretable Basis Decomposition (IBD) [Zhou et al., ECCV 2018]
  - Automatic Concept-based Explanations (ACE) [Ghorbani et al., NeurIPS 2019]
  - ConceptSHAP [Yeh et al., NeurIPS 2020]
- Ensure that concept-based explanations are truly human-interpretable

Takeaway: Be realistic about the benefits and limitations of an interpretability method and work towards addressing the limitations.

Roadmap

Iro Laina

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.
4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

Supervised Learning

( , sheepdog)

x **y**

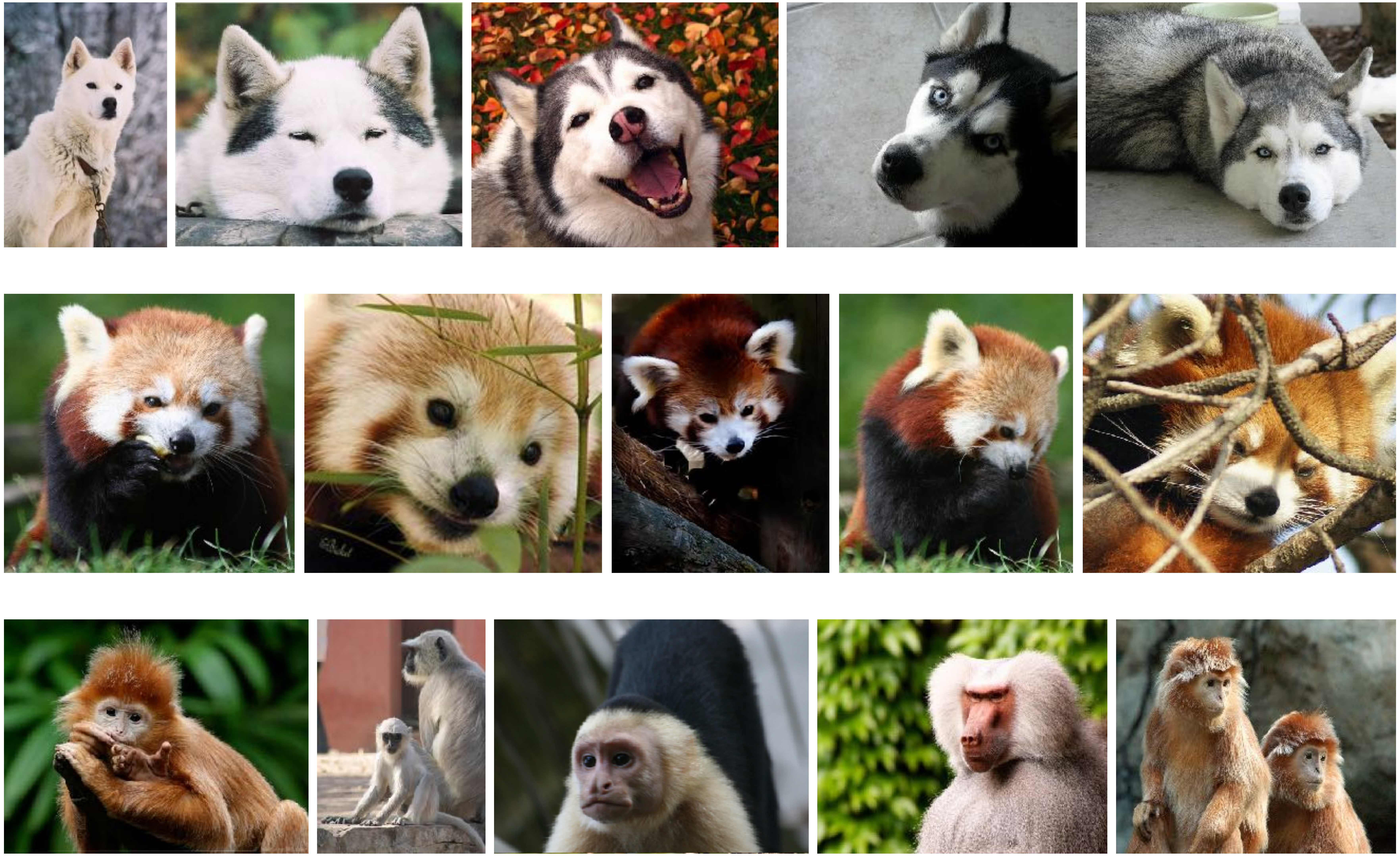
Self-Supervised Learning

$\mathbf{x}$

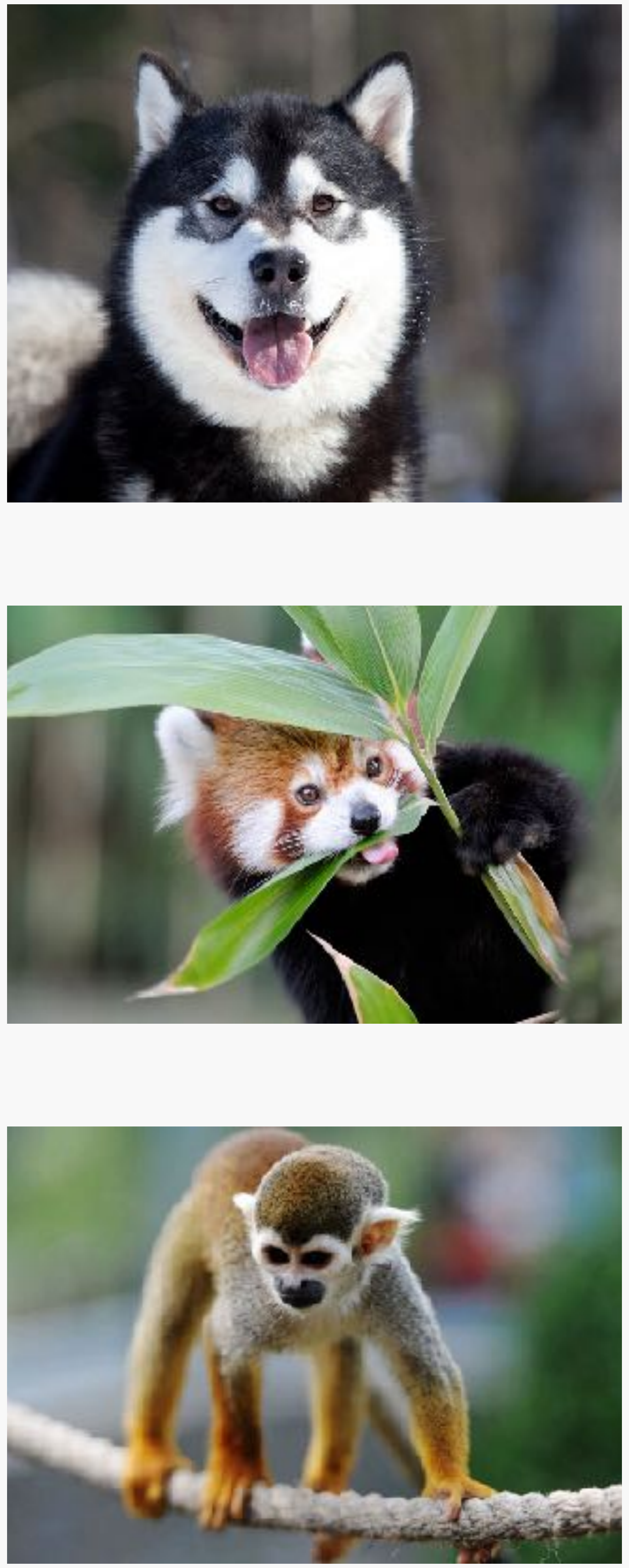
Visual Concept

Query

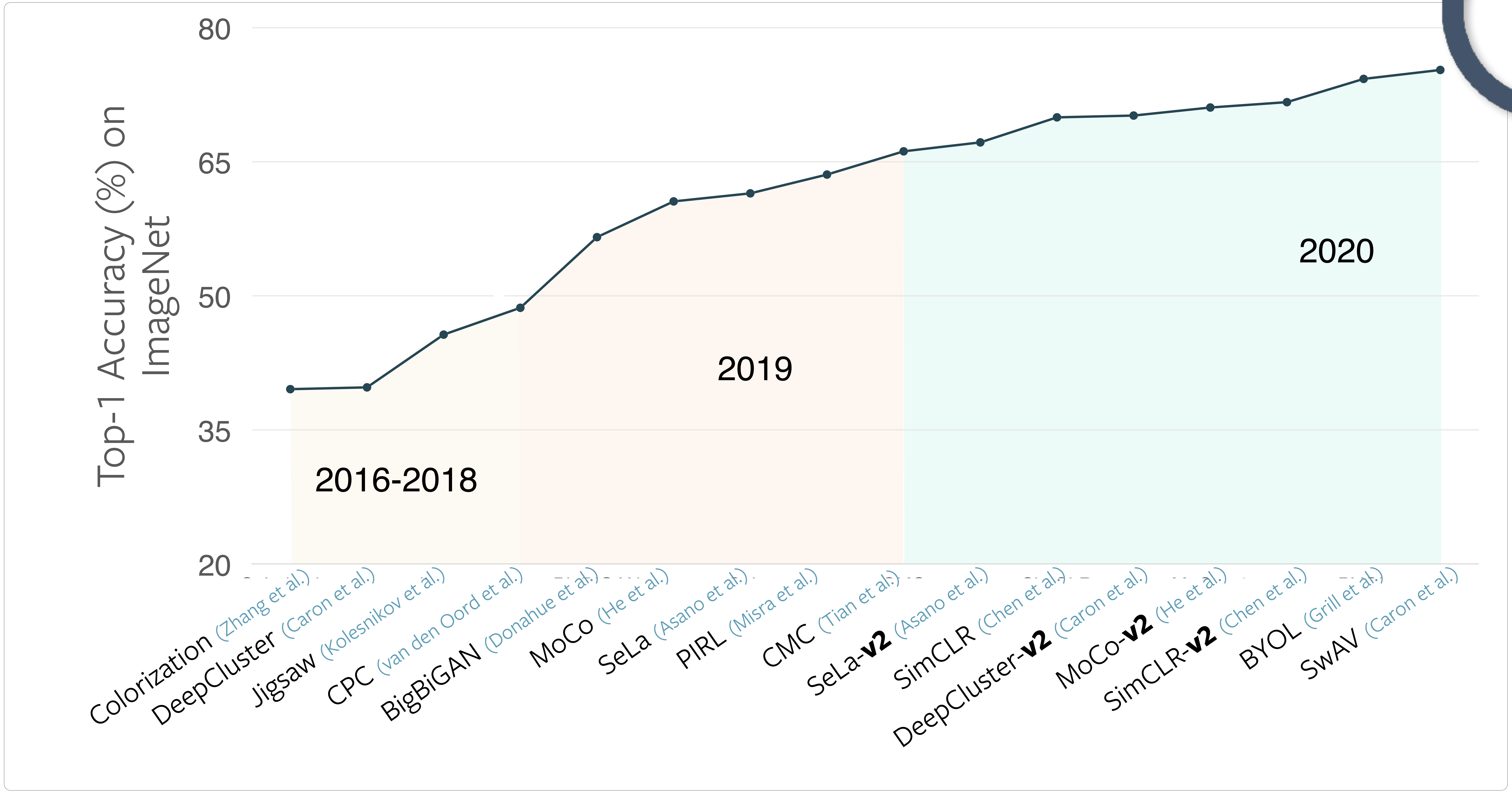
Visual Concept



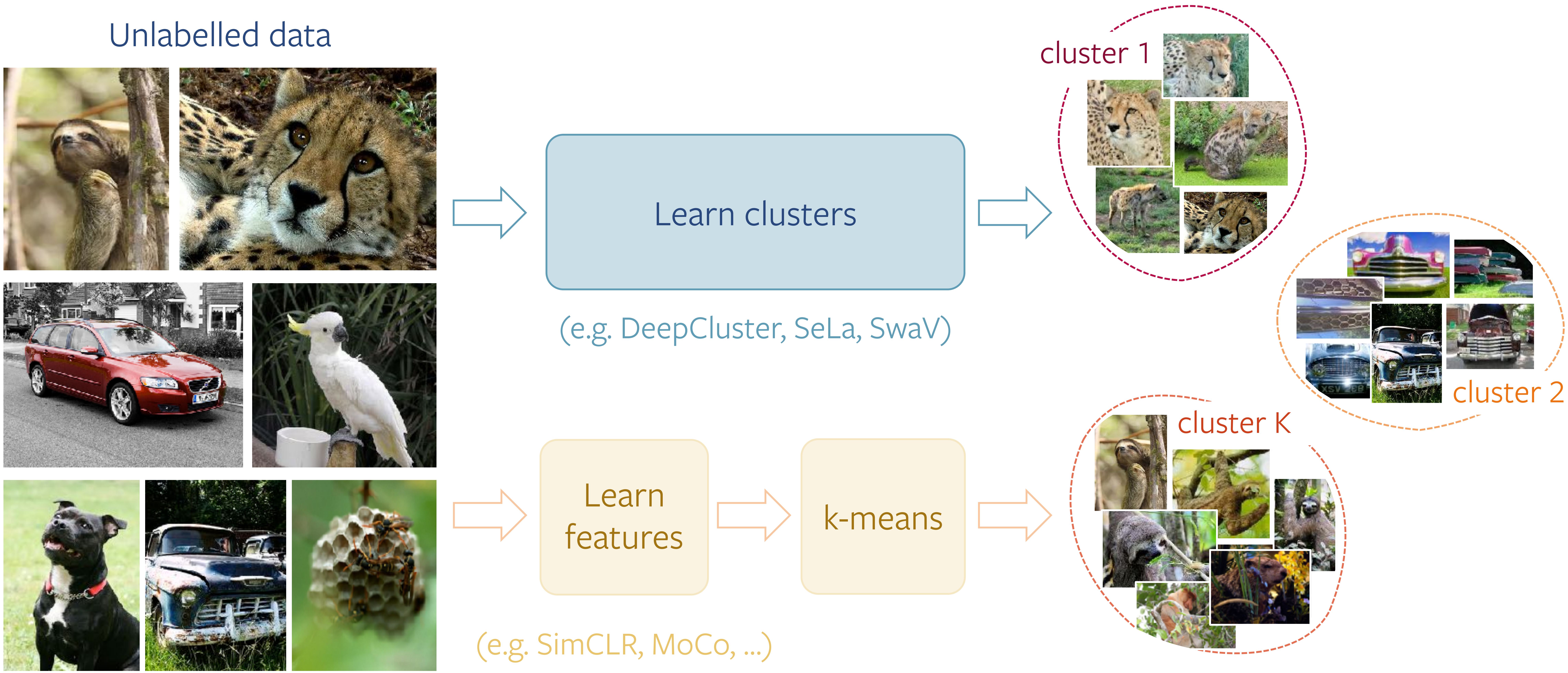
Query



Self-Supervised Learning



Self-Supervised Learning



Learnability

(A)

(B)

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]

Learnability

(A)

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]

Describability

dessert with chocolate sauce

(A)

(B)

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]

Describability

“ dessert with chocolate sauce

Manual

(A)

(B)

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]

Describability

“dessert with chocolate sauce

Manual

OR

Automatic

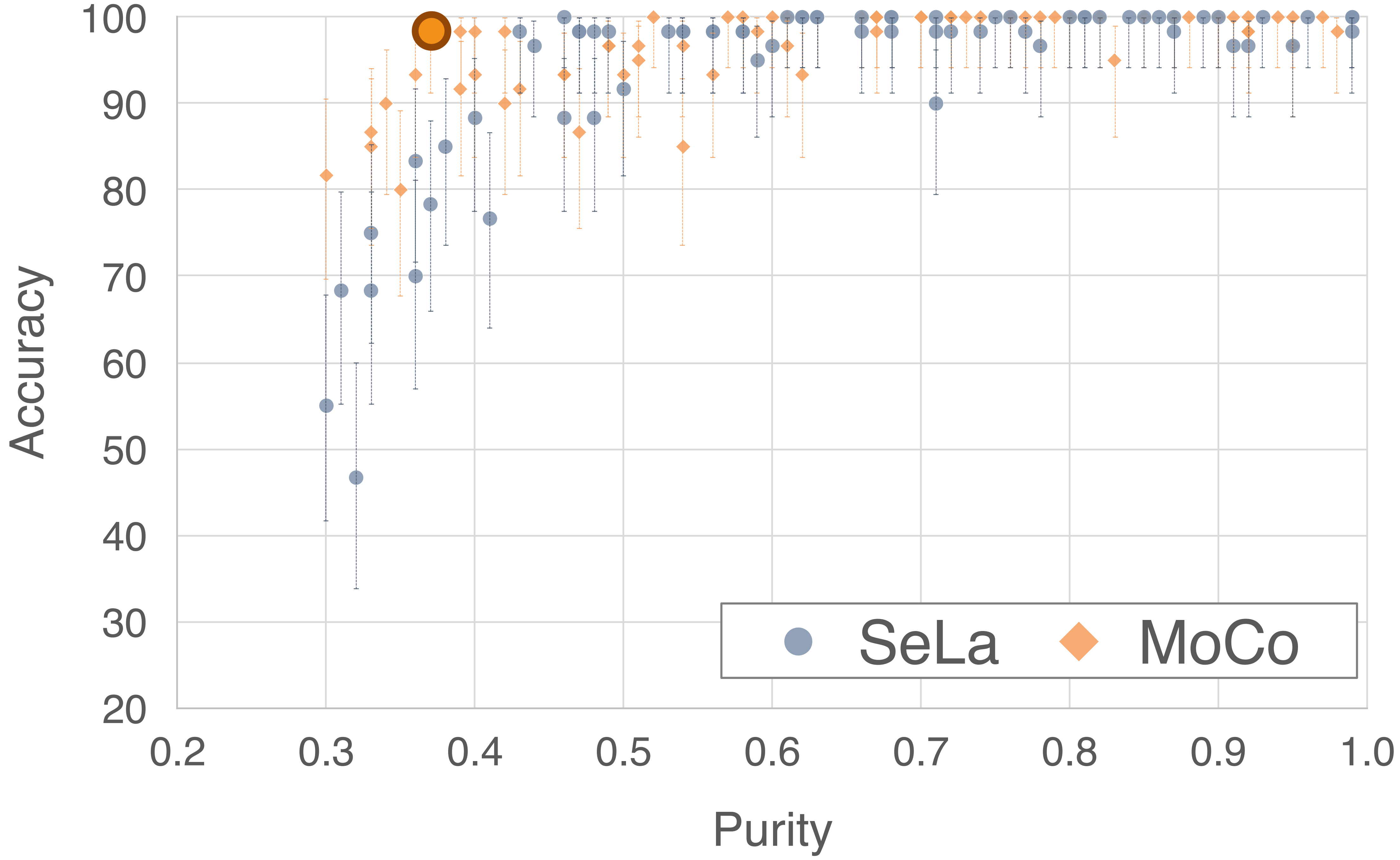
(A)

(B)

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]

Evaluation

Learnability



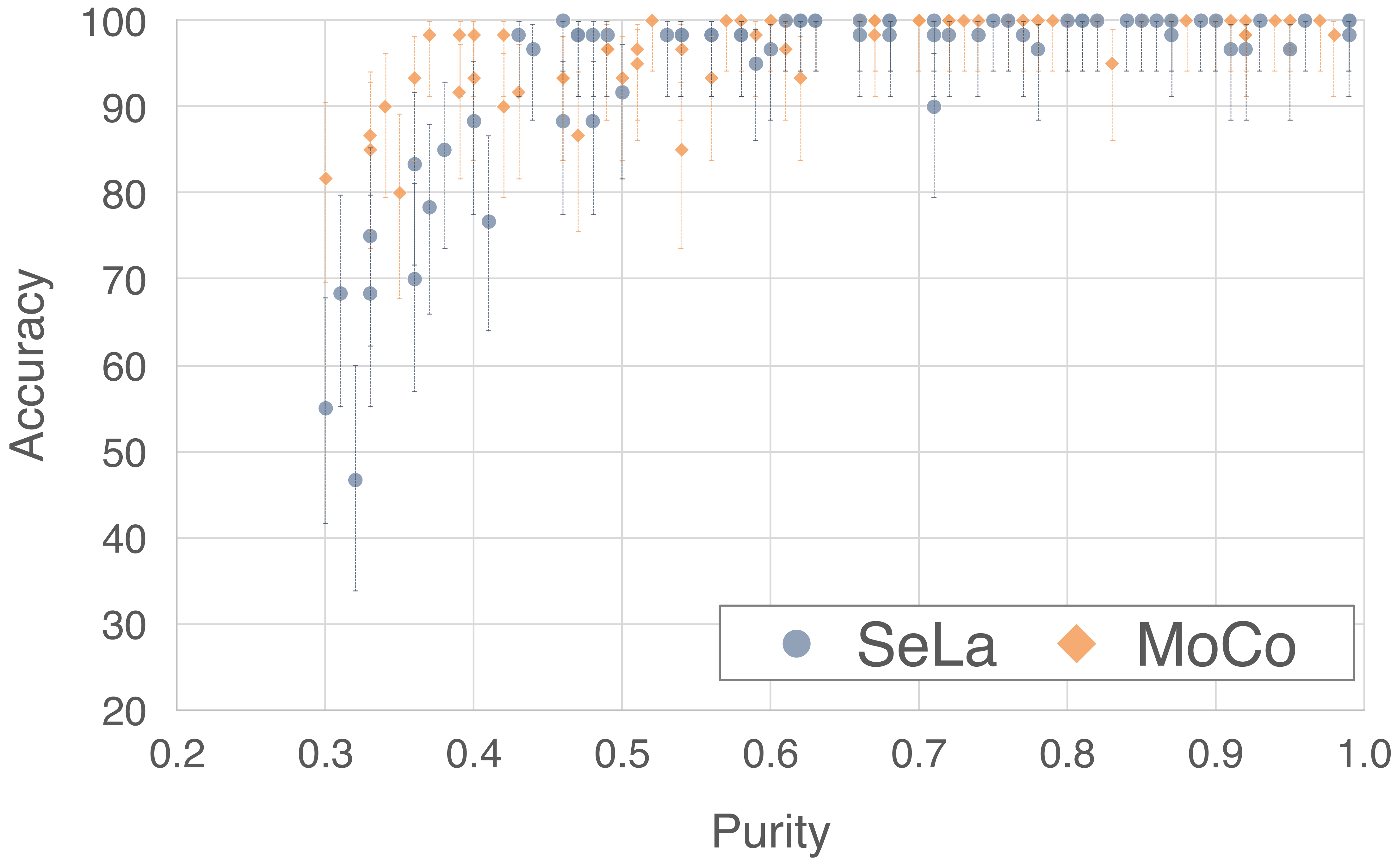
ImageNet cluster purity:
how correlated is a cluster's contents to a single ImageNet label?

***purity* = 1** → *cluster only contains images from a single ImageNet label*

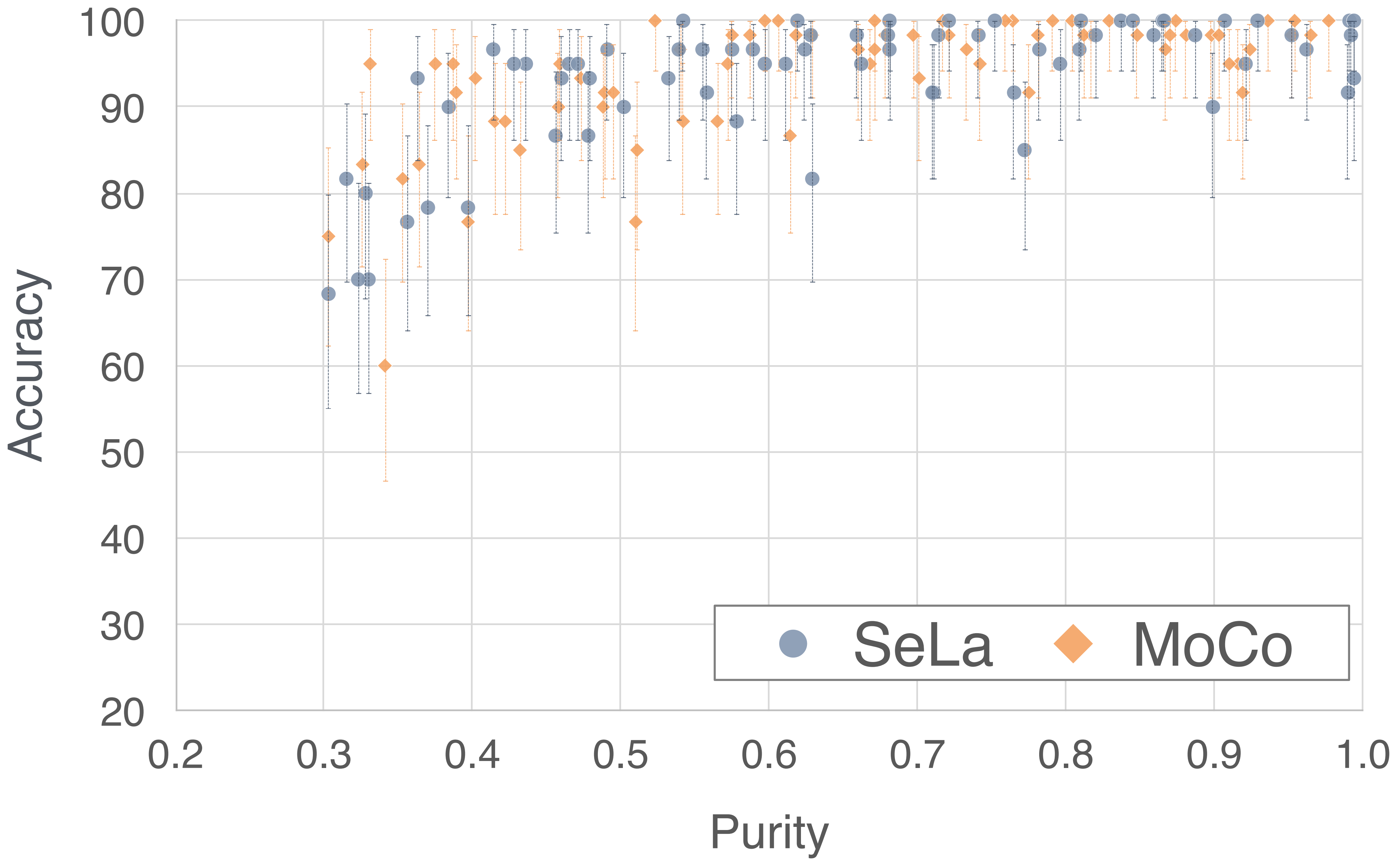
[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]
[Asano et al., ICLR 2020; He et al., CVPR 2020]

Evaluation

Learnability



Describability



[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]
[Asano et al., ICLR 2020; He et al., CVPR 2020]

Follow up: Laina et al., ICLR 2022.
Measuring the Interpretability of Unsupervised Representations via Quantized Reverse Probing.

Findings

ImageNet cluster purity

SeLa: cluster 393 (0.668)
a newborn baby lying on a bed

SeLa: cluster 332 (0.542)
a snake on a hand

98.3%   100.0%

MoCo: cluster 2335 (0.459)
view of the mountains from the lake

93.3%   95.0%

[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]
[Asano et al., ICLR 2020; He et al., CVPR 2020]

Challenges for novel frontiers in deep learning

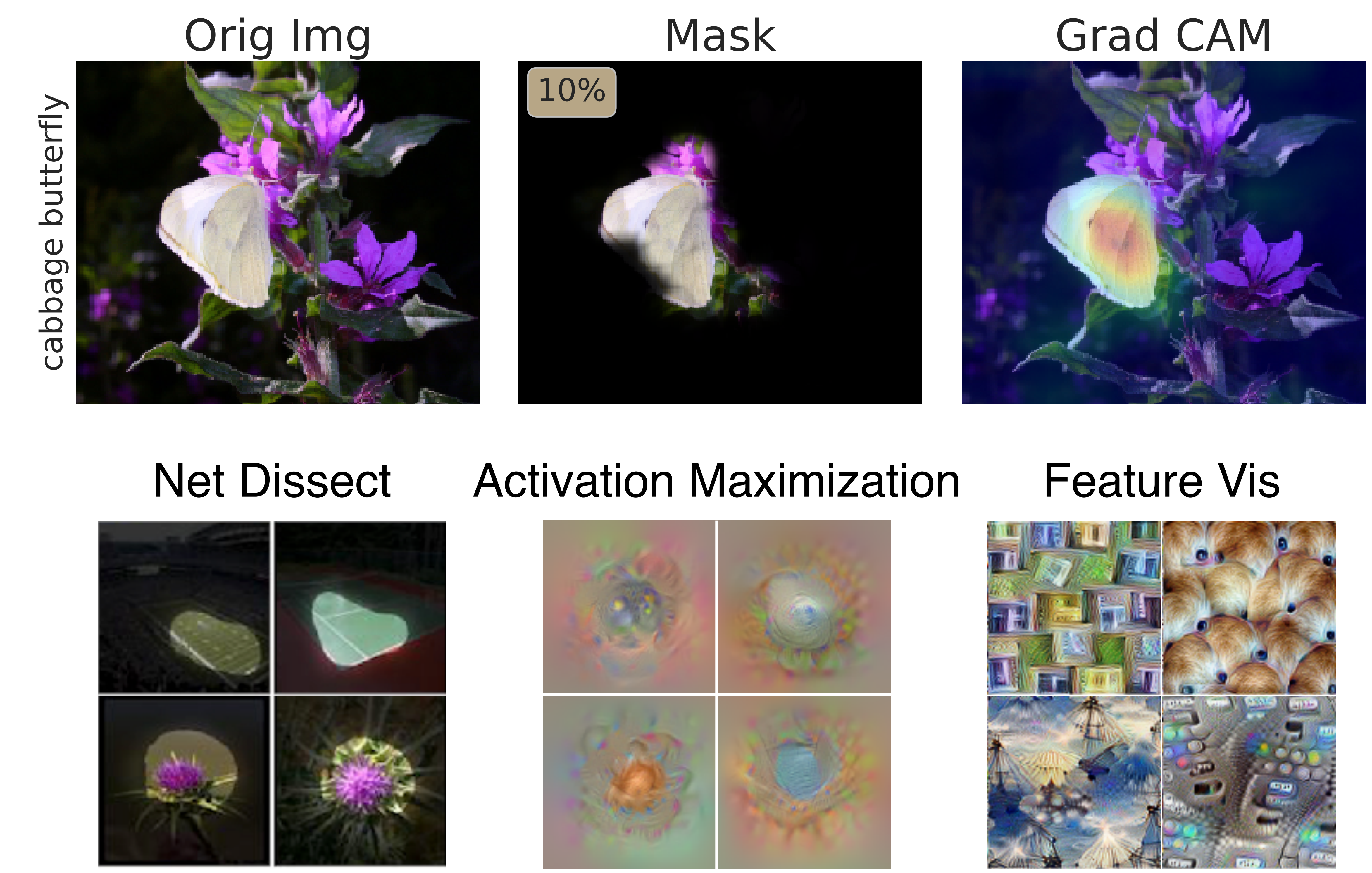
- Need to contextualize interpretability to the novel frontiers
- Lack of access to standardized implementations

Takeaway: Collaboration and buy-in from novel research areas is crucial for interpretability in those frontiers.

Roadmap

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2021.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.
4. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

Interpretability Tools

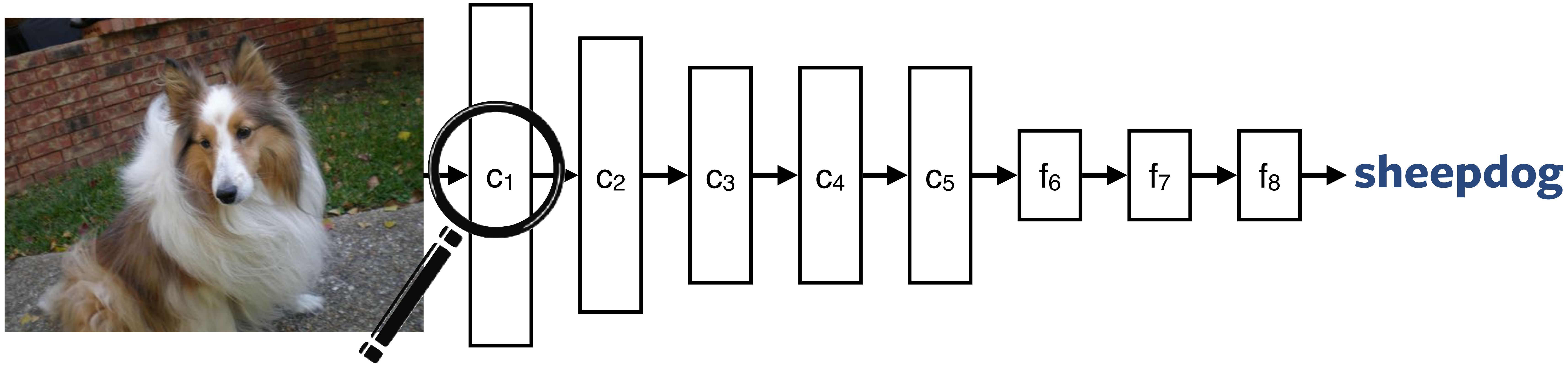


Current tools render **static images**.

Future tools should be **interactive**!

[Fong et al., ICCV 2019; Selvaraju et al., ICCV 2017; Bau et al., CVPR 2017; Mahendran & Vedaldi, IJCV 2016; Olah et al., Distill 2018; Fong et al., VISxAI 2021]

Interpretability: Interactive, Exploratory, Easy-to-use



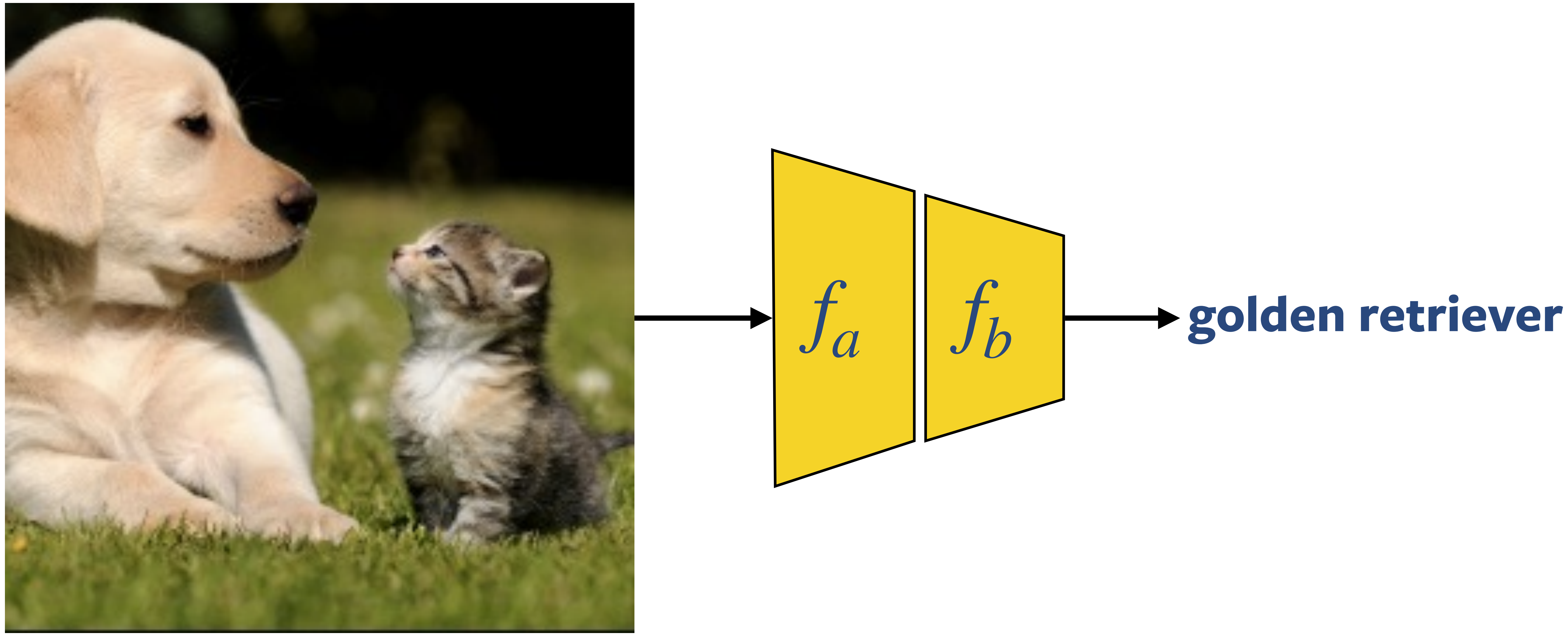
How can we **easily explore** hypotheses about the model?

Acknowledgement: Chris Olah

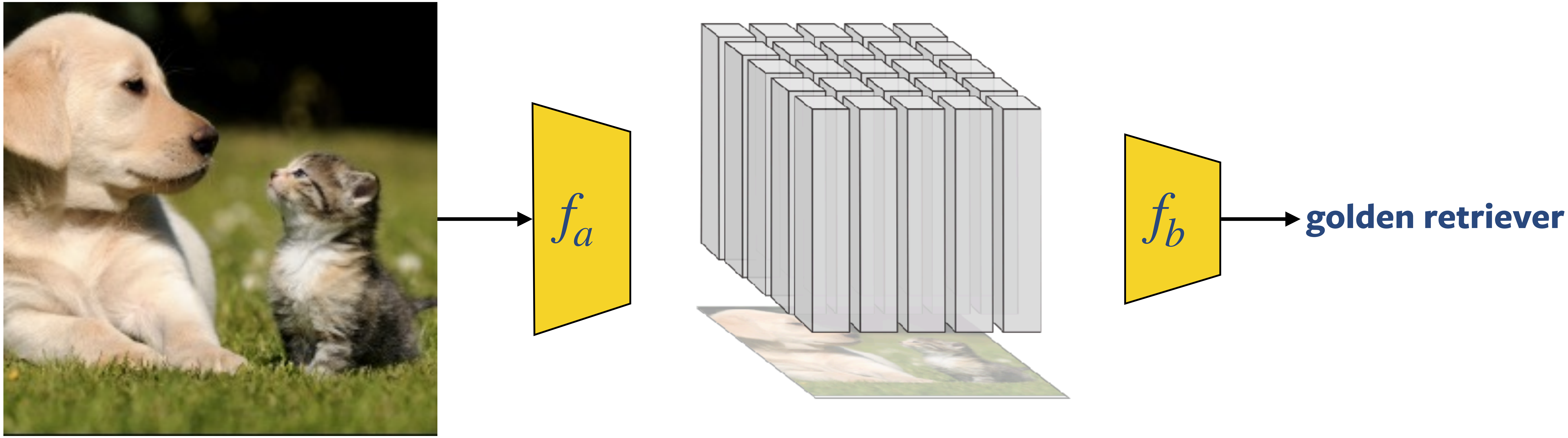
Interactive Similarity Overlays

Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
Interactive Similarity Overlays.

Spatial Activations



Spatial Activations

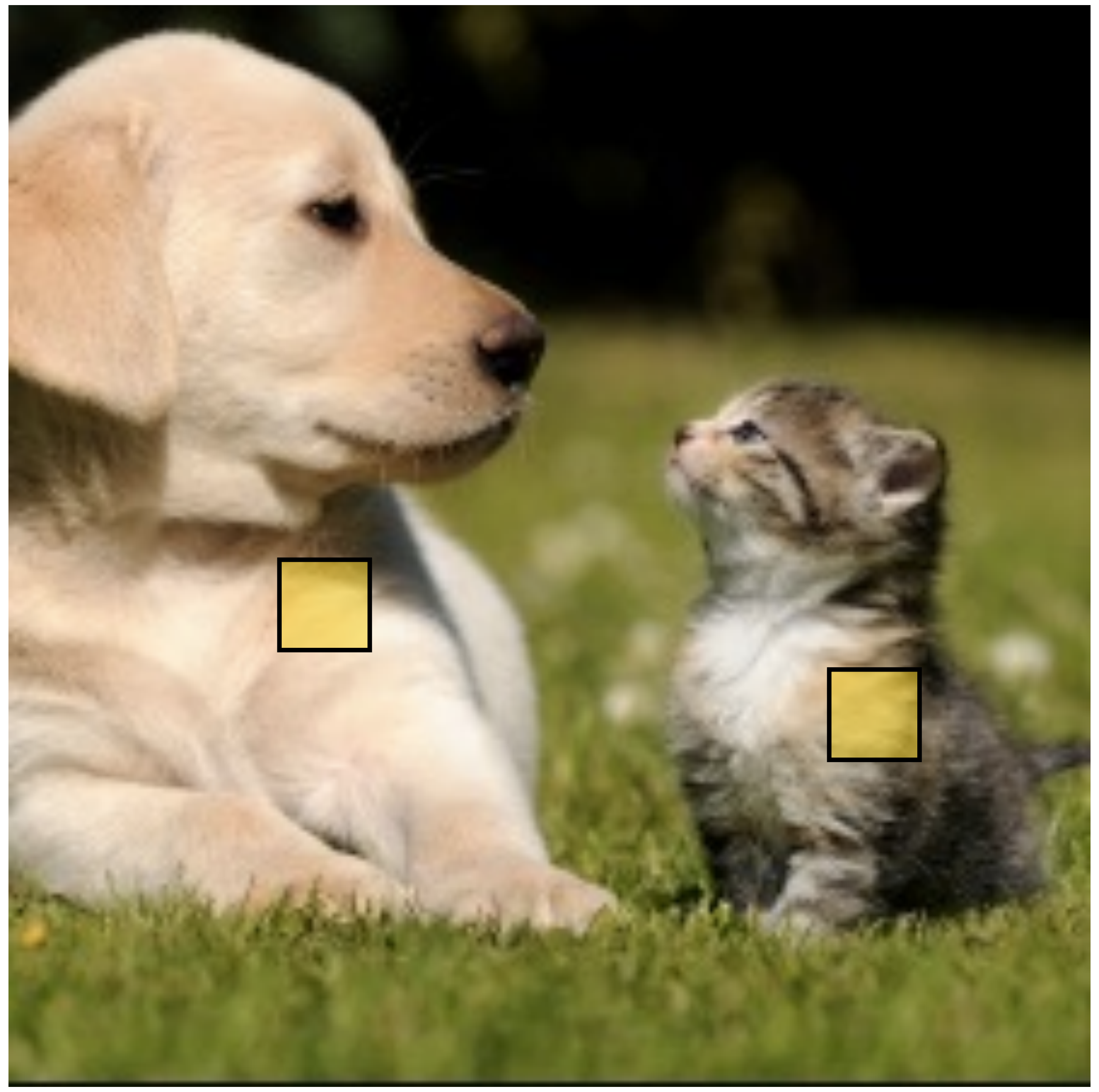


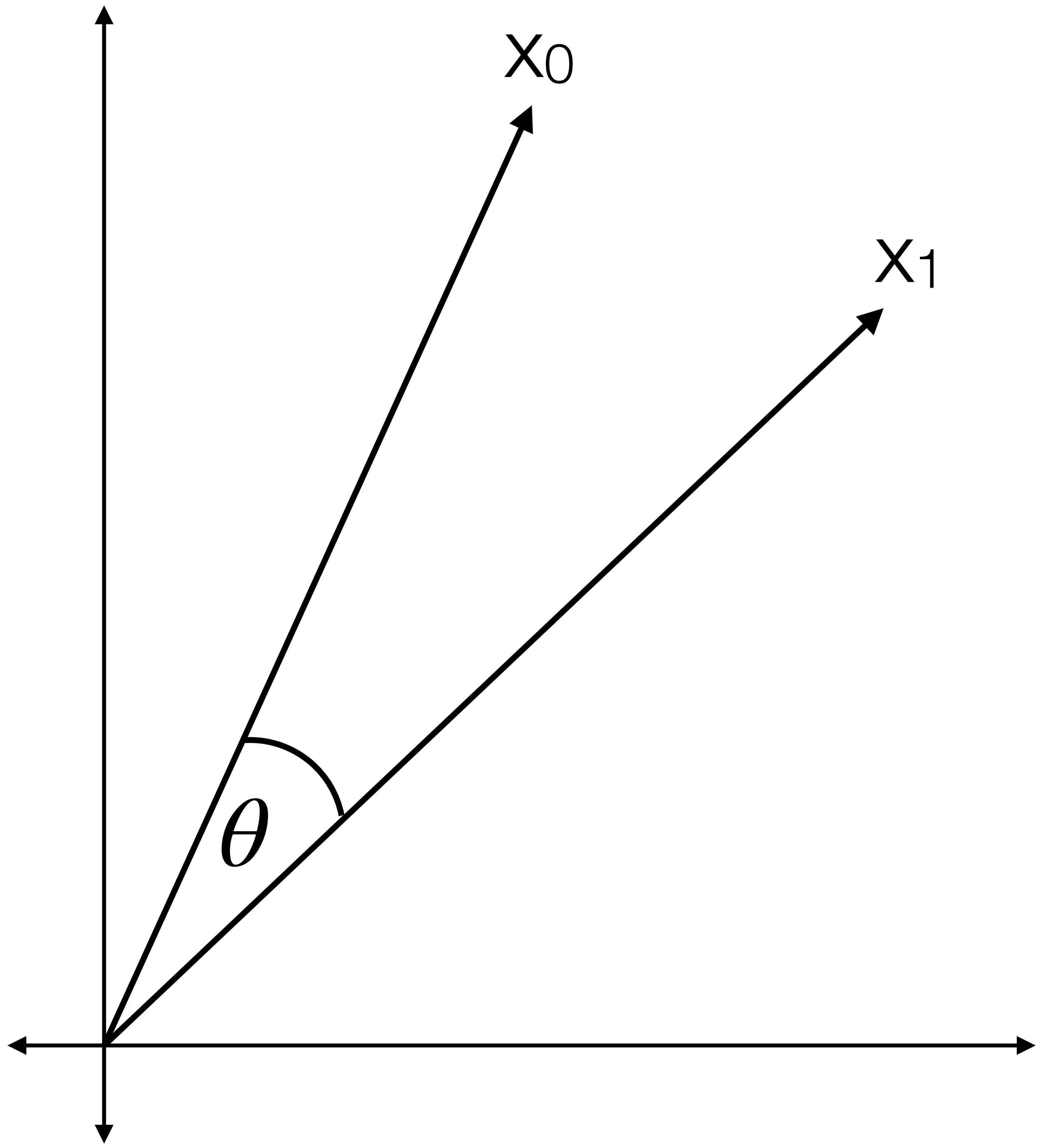
[Olah et al., Distill 2018]

Interactive Similarity Overlays

$a_{6,5}$ = [17.7, 0, 103.4, 6.81, 0, 0, 0, 0, 32.0, 0, 0, 0, ...]

Interactive Similarity Overlays

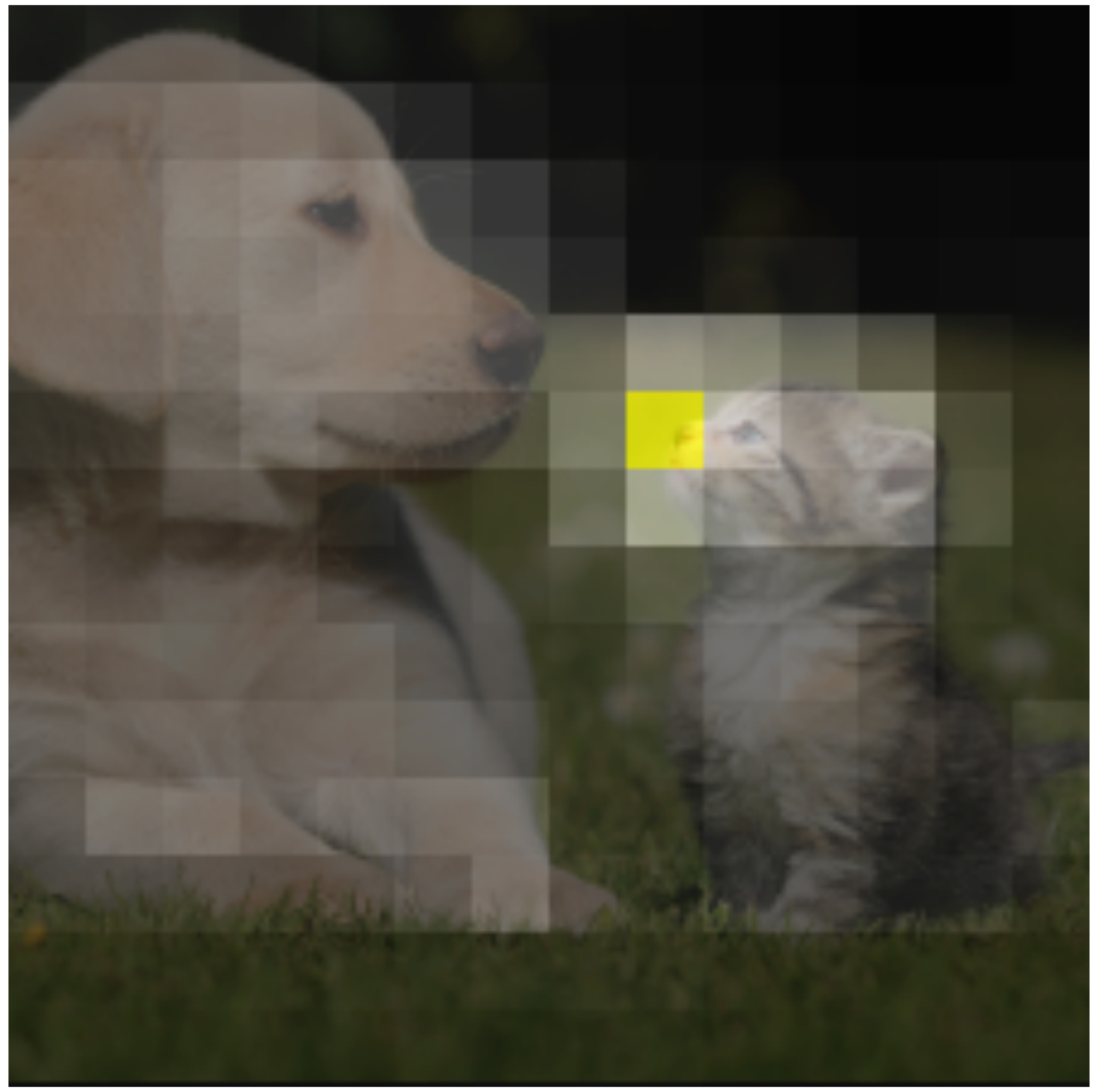




[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]

Live Demo: Interactive Similarity Overlays

bit.ly/interactive_overlay



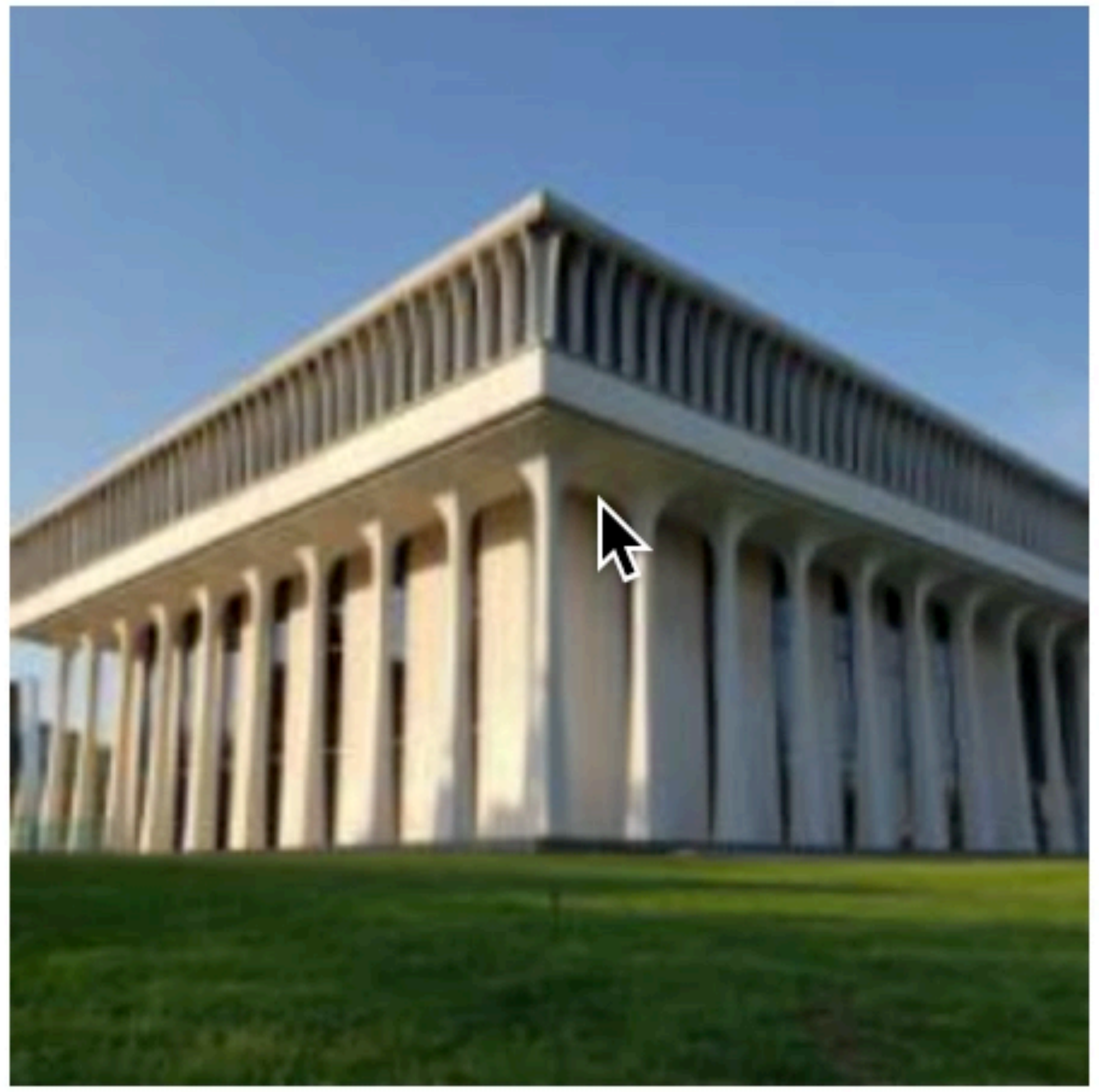
Interactive visualizations empower practitioners to easily explore model behavior.

[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]

Preview: Interactive Visual Feature Search

bit.ly/interactive_search

Devon Ulrich



Devon Ulrich and Ruth Fong, in prep 2022.
Interactive Visual Feature Search.

Acknowledgement: David Bau

Challenges for interactive visualizations

- Skills cost: web development skills
  - 📈 HuggingFace Spaces, Gradio, Streamlit
- Potential misuse: Intuition-based insights should be validated via quantitative experiments
- Poor incentives: software tooling for research is often not rewarded
- Inadequate publishing structures: Sparse publishing venues for interactive articles and/or visualizations
  - 📉 Distill journal hiatus
  - 📈 CVPR demo track
- Lack of cross-talk: HCI and AI communities are developing interpretability tools fairly independently

Takeaway: Relevant research communities should collectively invest in and reward software tooling for research, particularly interactive tools.

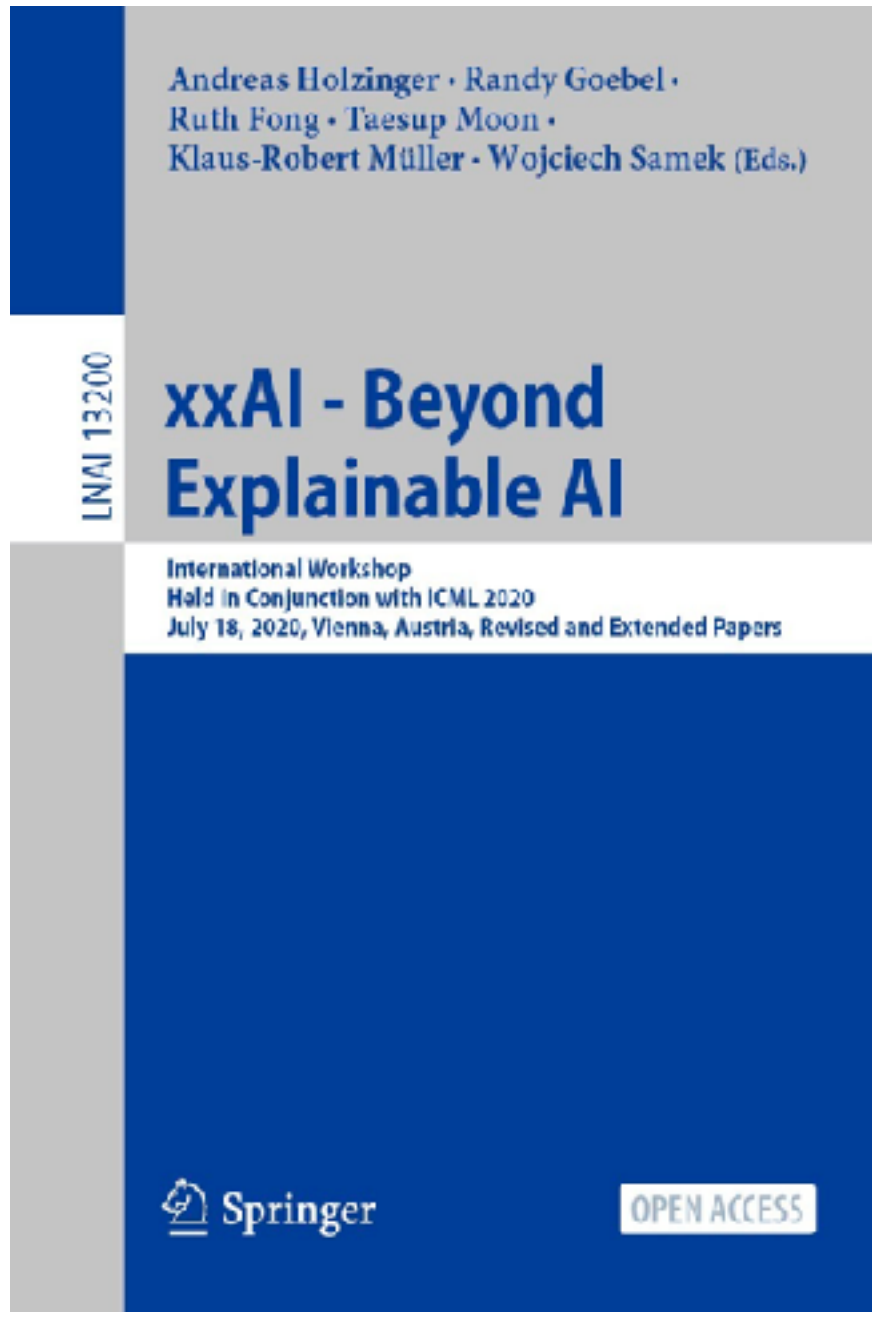
Takeaways from challenges in interpretability

- **Human studies:** As a research community, invest in and reward human evaluation studies (like dataset development).
- **(Concept-based) interpretability:** Be realistic about the benefits and limitations of an interpretability method and work towards addressing the limitations.
- **New frontiers:** Collaboration and buy-in from novel research areas is crucial for interpretability in those frontiers.
- **Interactive visualizations:** Relevant research communities should collectively invest in and reward software tooling for research, particularly interactive tools.

Directions for the next decade of interpretability

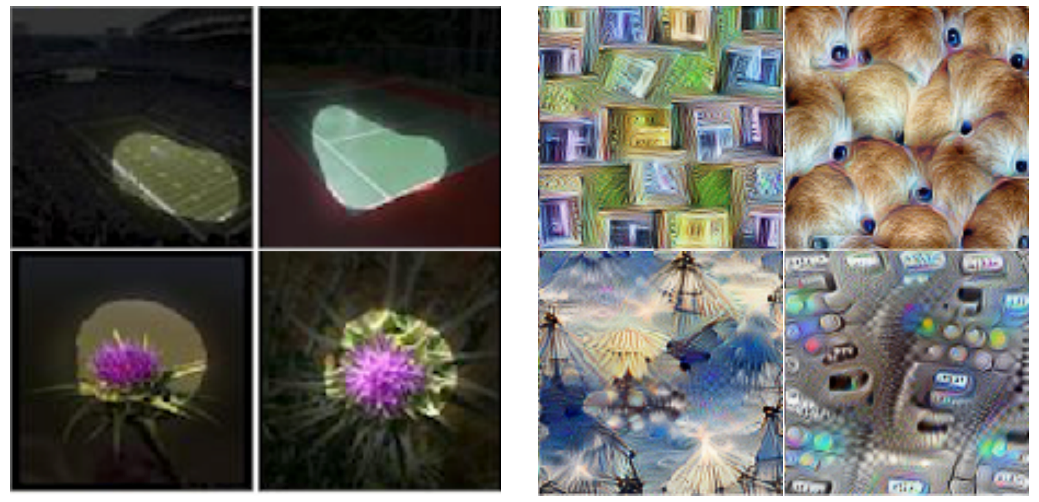
1. Develop interpretability methods for **diverse domains**
   - Beyond CNN classifiers: self-supervised learning, generative models, etc.

2. Center **humans** throughout the development process
   - In design, co-develop methods with real-world stakeholders.
   - In evaluation, measure human interpretability and utility of methods.
   - In deployment, package interpretability tools for the wider community.



ICML 2020 workshop on XXAI

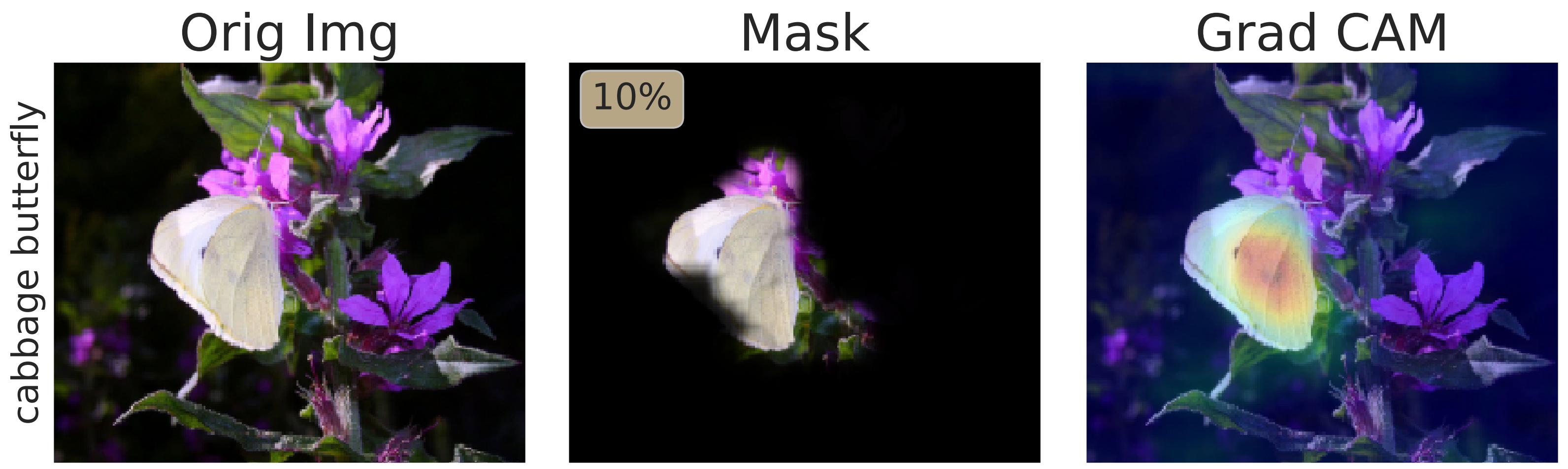
An incomplete retrospective: the first decade of interpretability



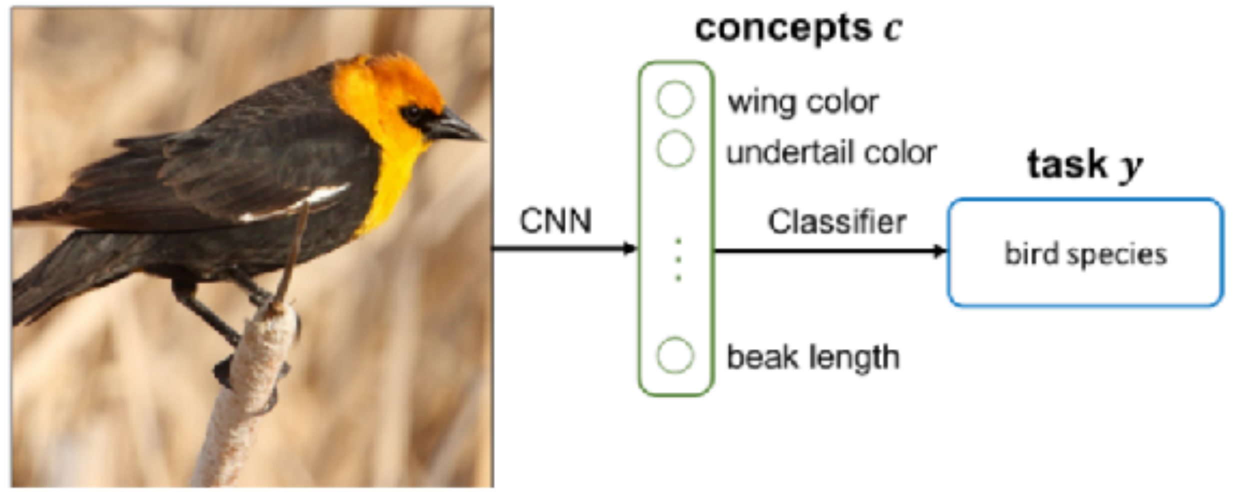
Primarily focused on understanding and approximating **CNNs**

Feature visualization (2013-2018)
Activation Max., Feature Inversion,
Net Dissect, Feature Vis.

2012 — 2022



Attribution heatmaps (2013-2019)
Gradient, Grad-CAM,
Occlusion, Perturbations, RISE



Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet,
ProtoTree

[Selvaraju et al., ICCV 2017; Fong* & Patrick* et al., ICCV 2019; Bau* & Zhou* et al., CVPR 2017; Olah et al., Distill 2017; Koh*, Nguyen*, Tang* et al., ICML 2020]

Into the future: the next decade of interpretability

???

2022 2032

Iro Laina

Devon Ulrich

Nicole Meister

Sunnie S. Y. Kim

Vikram V. Ramaswamy

Andrea Vedaldi

Chris Olah

Alex Mordvintsev

Olga Russakovsky

bit.ly/vai-lg-postdoc

We're hiring postdocs!

UNIVERSITY OF OXFORD

RHODES TRUST

NSF

Open Philanthropy Project

Talk acknowledgements: Brian Zhang, Sunnie S. Y. Kim, Vikram V. Ramaswamy, Olga Russakovsky

Thank You