Directions in Interpretability

Ruth Fong

HEIBRiDS lecture

November 14, 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?

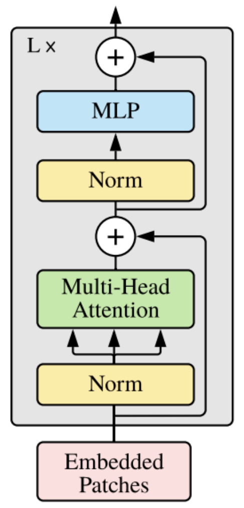
- 🔬 Science
- 🤝 Trust
- 🤖 Learning

An incomplete retrospective: the first decade of deep learning

Monet ⟶ photo

GANs (2014-2018)

GAN, ProGAN, CycleGAN



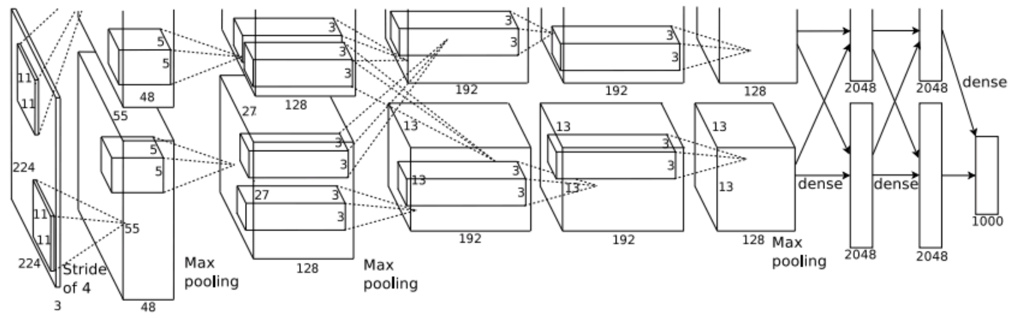
Transformers (2017-now)

Transformer, BERT, ViT

IMAGENET

2012

2022



CNNs (2012-2016)

AlexNet, VGG16,
GoogLeNet, ResNet50

Self-supervised learning (2016-now)

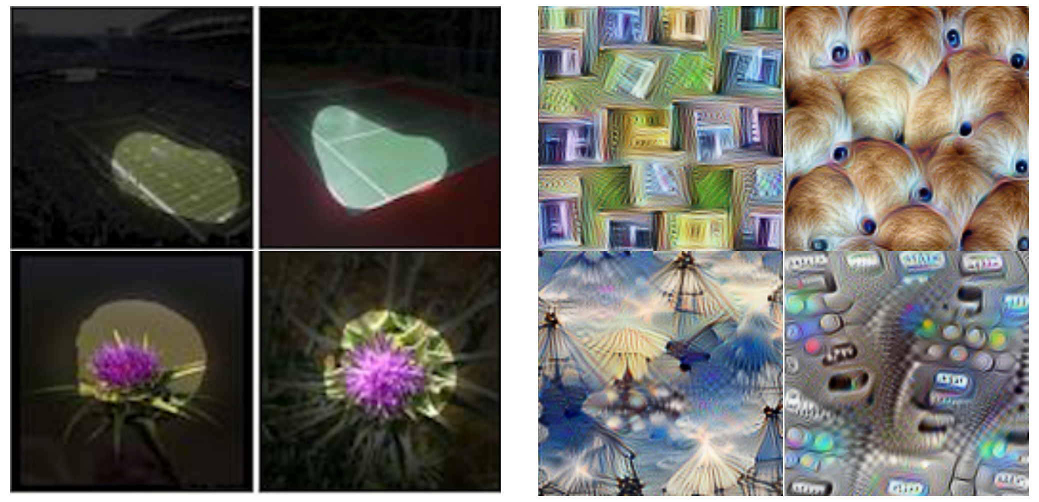
Colorization, MOCO, SWaV

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

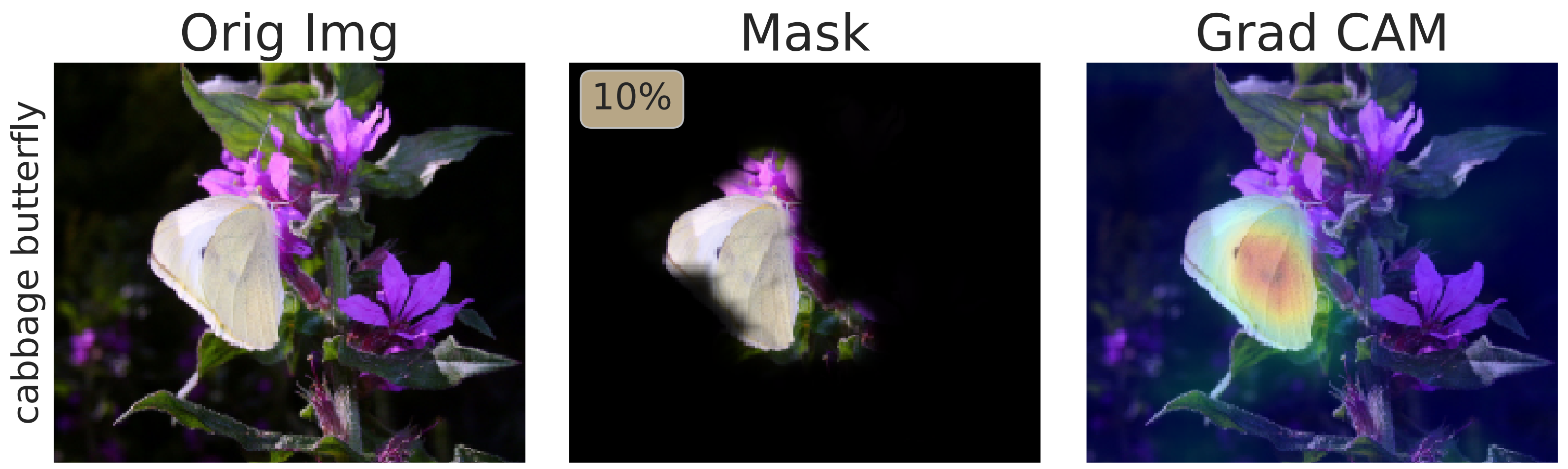


Feature visualization (2013-2018)

Activation Max., Feature Inversion,

Net Dissect, Feature Vis.

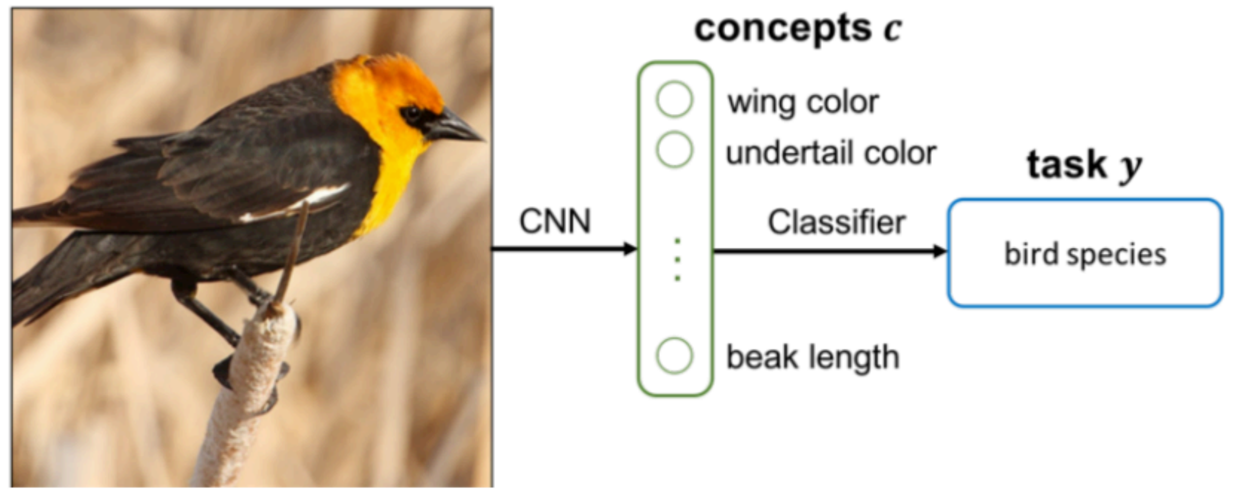
2012 — 2022



Attribution heatmaps (2013-2019)

Gradient, Grad-CAM,

Occlusion, Perturbations, RISE



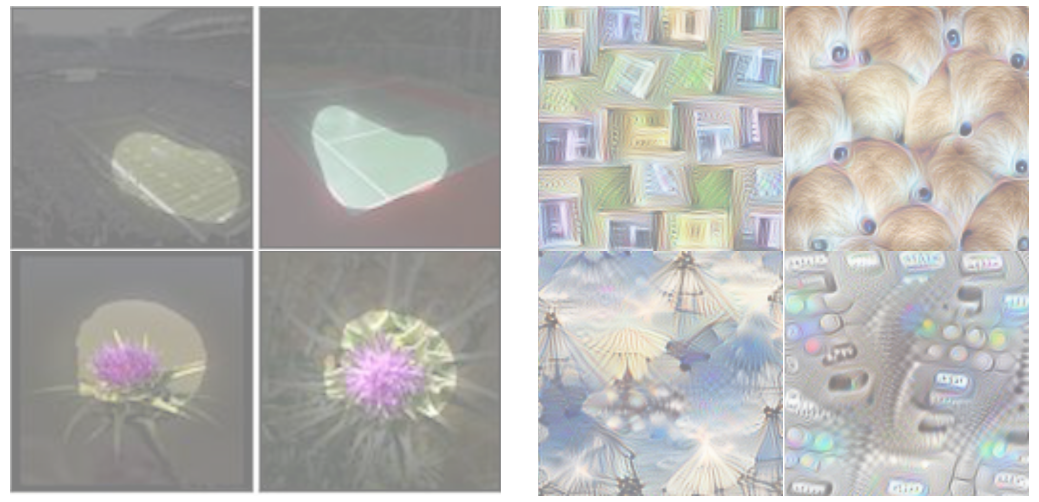
Interpretable-by-design (2020-now)

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

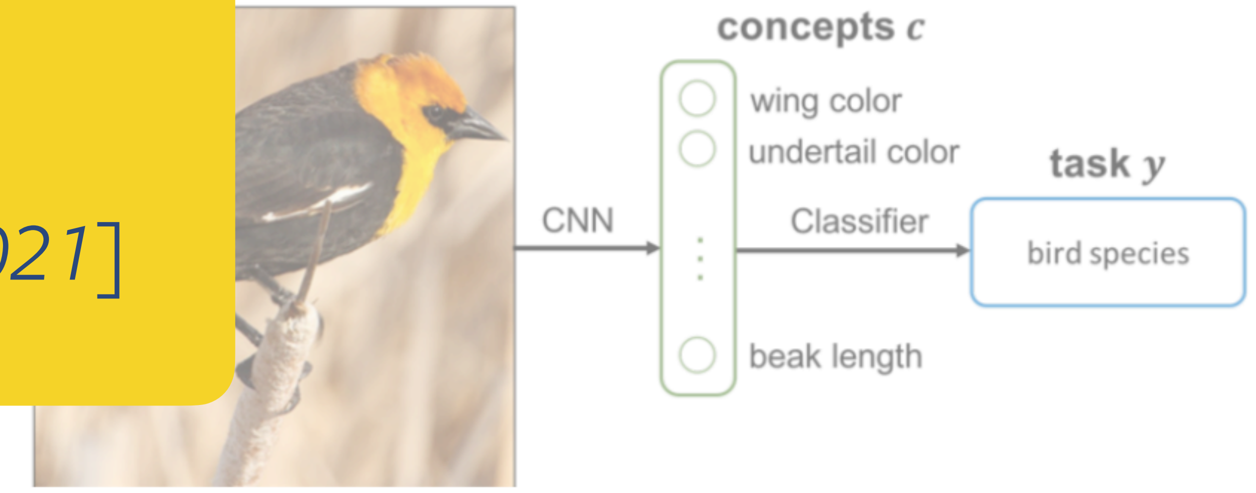


Primarily focused on understanding and approximating **CNNs**

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, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me Help the AI."*)

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.
   (+ *Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors in Concept-based Explanations.*)

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. **Interpretability** in ML + CV → **interdisciplinary** research (interpretability + X)
   (+ *Nicole Meister* and Dora Zhao* et al., arXiv 2022. Gender Artifacts in Visual Datasets.*)
   (+ *Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.*)

5. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
   (+ *Devon Ulrich and Ruth Fong, in prep. Interactive Visual Feature Search.*)

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, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me Help the AI."*)

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.
   (+ *Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors in Concept-based Explanations.*)

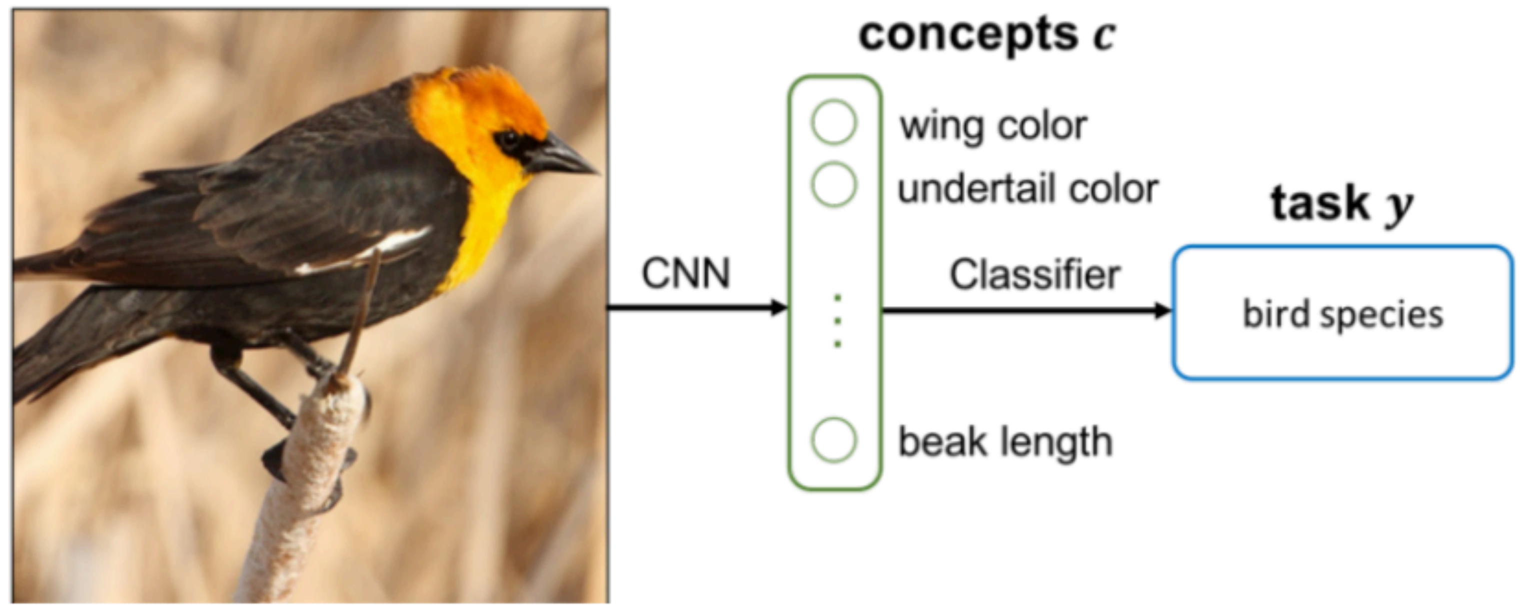
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Interpretability** in ML + CV → **interdisciplinary** research (interpretability + X)
   (+ *Nicole Meister* and Dora Zhao* et al., arXiv 2022. Gender Artifacts in Visual Datasets.*)
   (+ *Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.*)

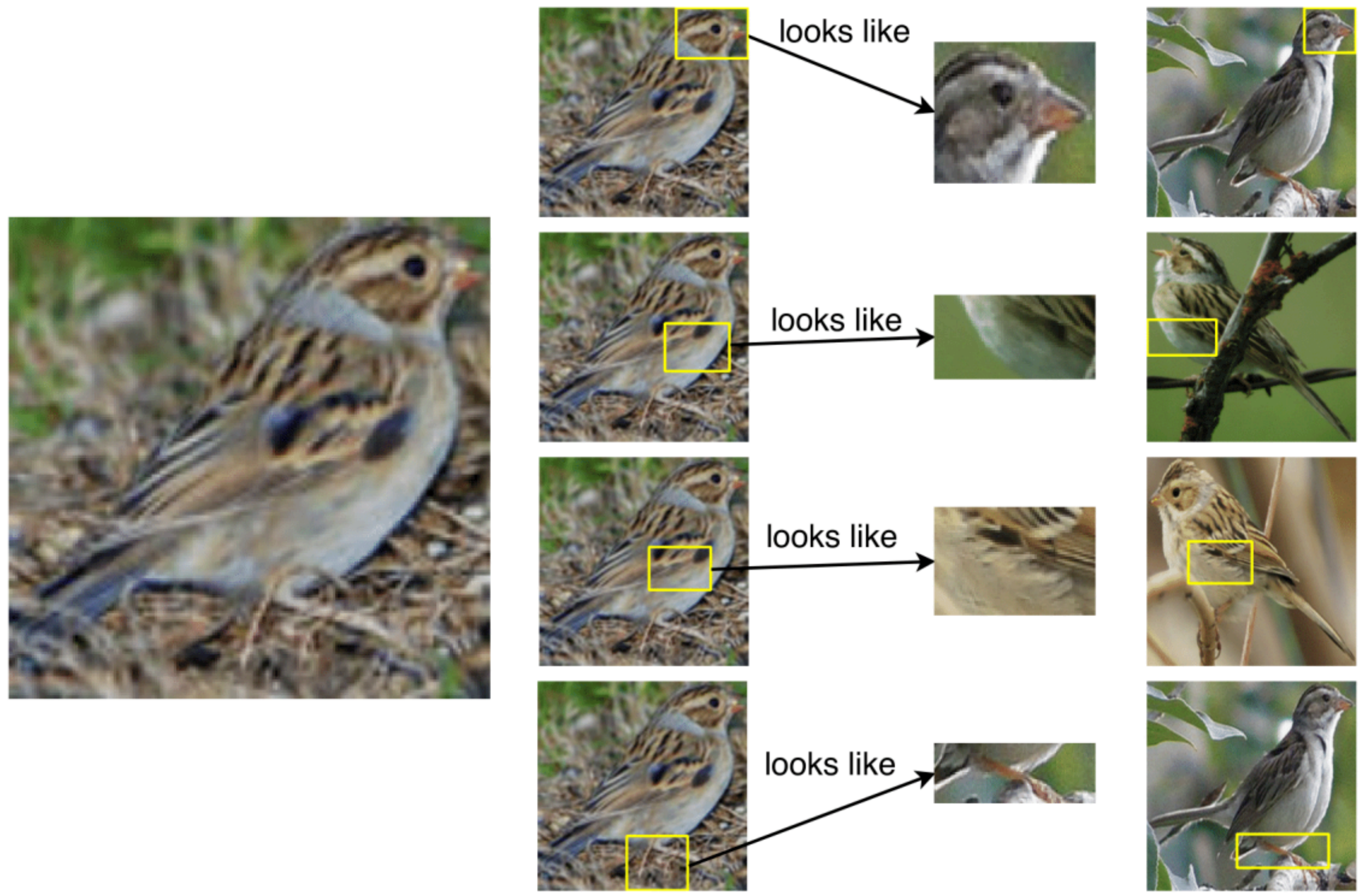
5. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
   (+ *Devon Ulrich and Ruth Fong, in prep. Interactive Visual Feature Search.*)

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)?

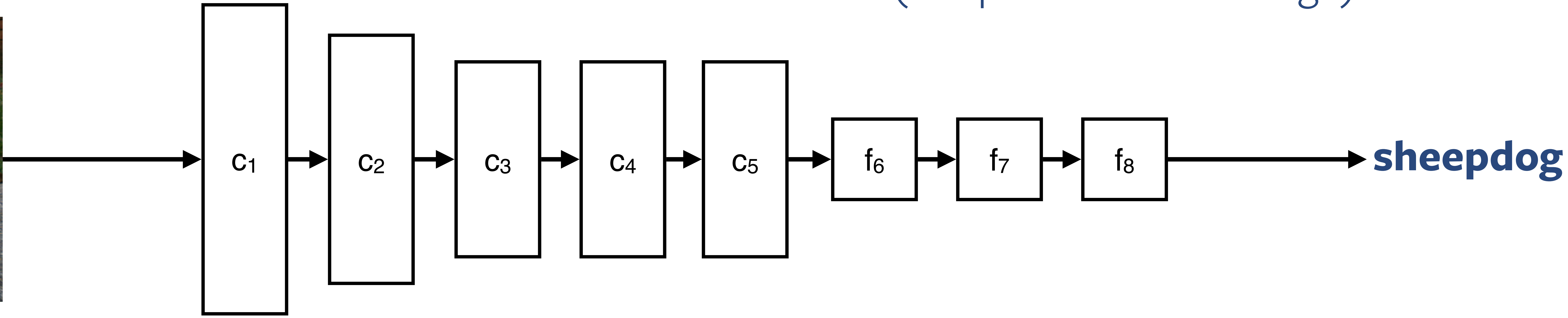
Counterfactual explanations
(e.g. SCOUT)

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

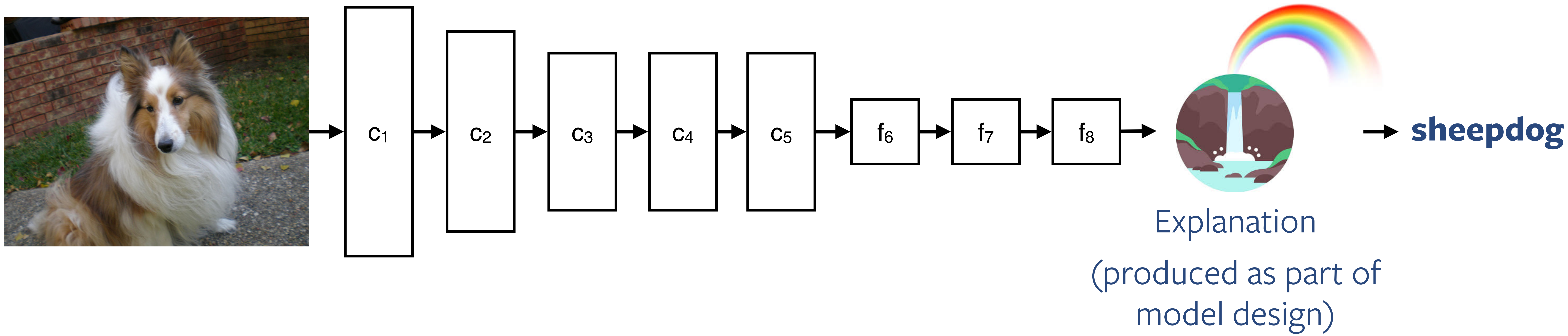
Post-hoc explanations

Explanation

(not part of model design)

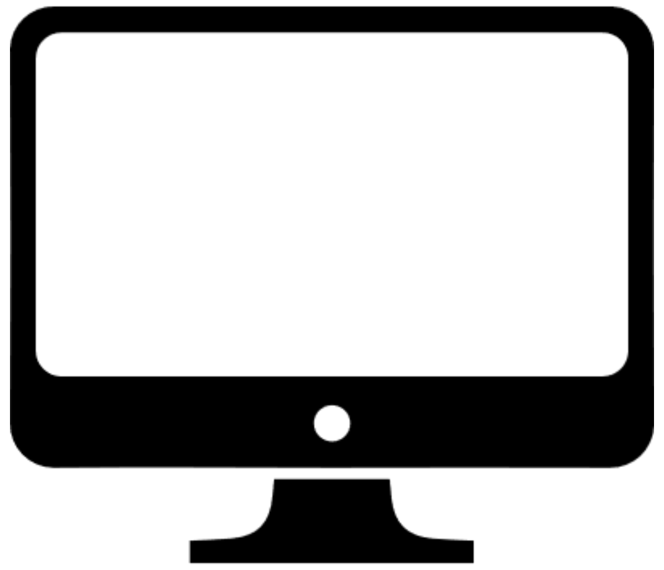


Interpretable-by-design models



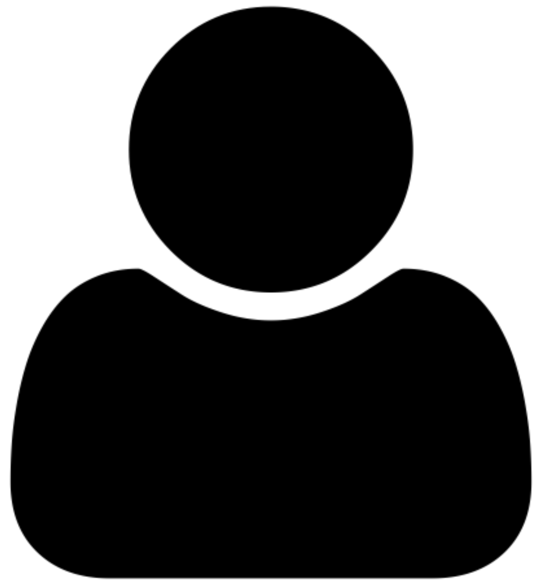
Current metrics focus on heatmap evaluation

- Weak localization performance [Zhang et al., ECCV 2016]
- Perturbation analysis
  - Deletion game [Samek et al., TNNLS 2017]
  - Retrain 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

- 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

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**

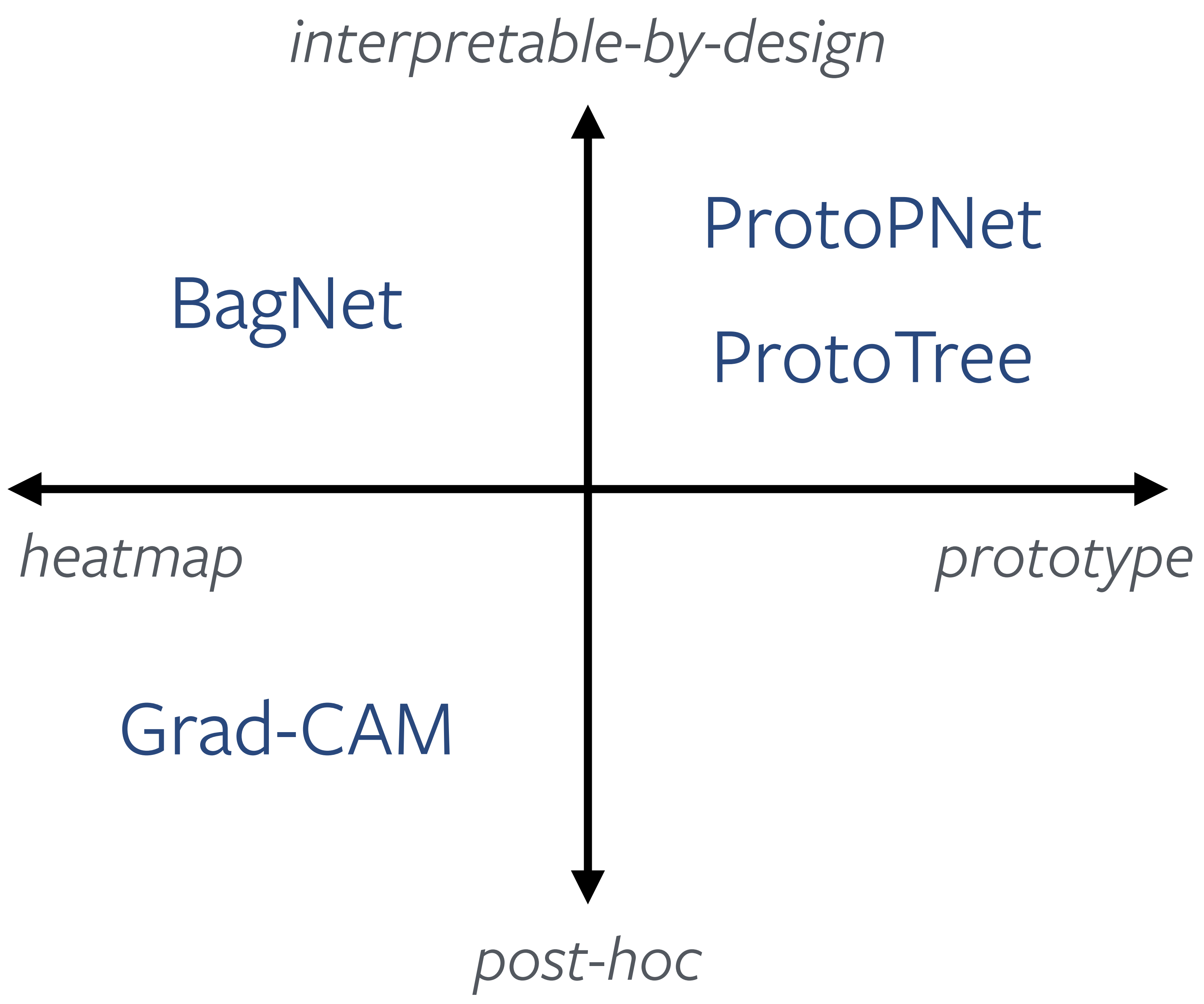
Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
HIVE: Evaluating the Human Interpretability of Visual Explanations.

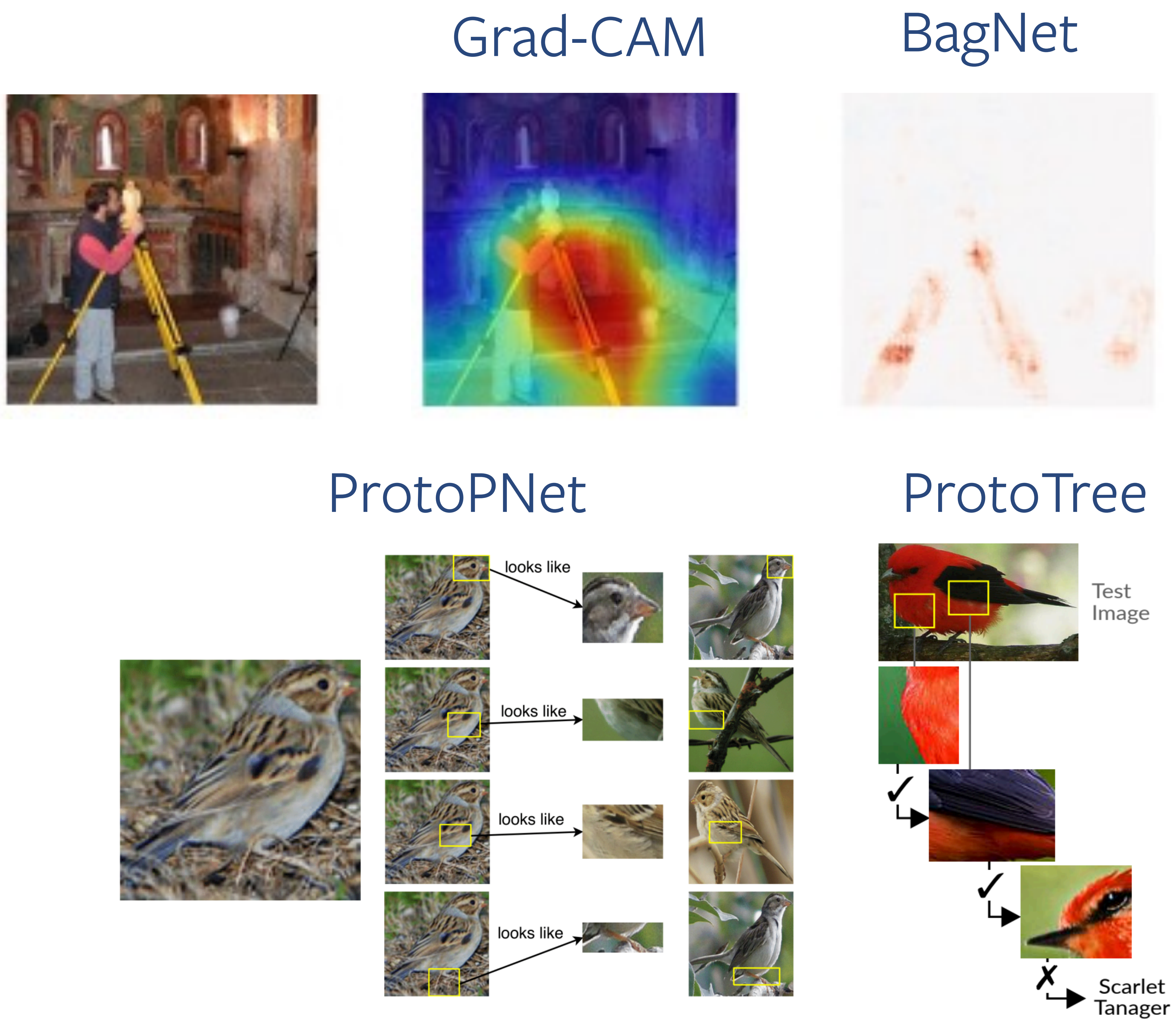
Our contributions

- 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., ECCV 2022. HIVE.]

1. Cross-method comparison





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

2. Human-centered evaluation

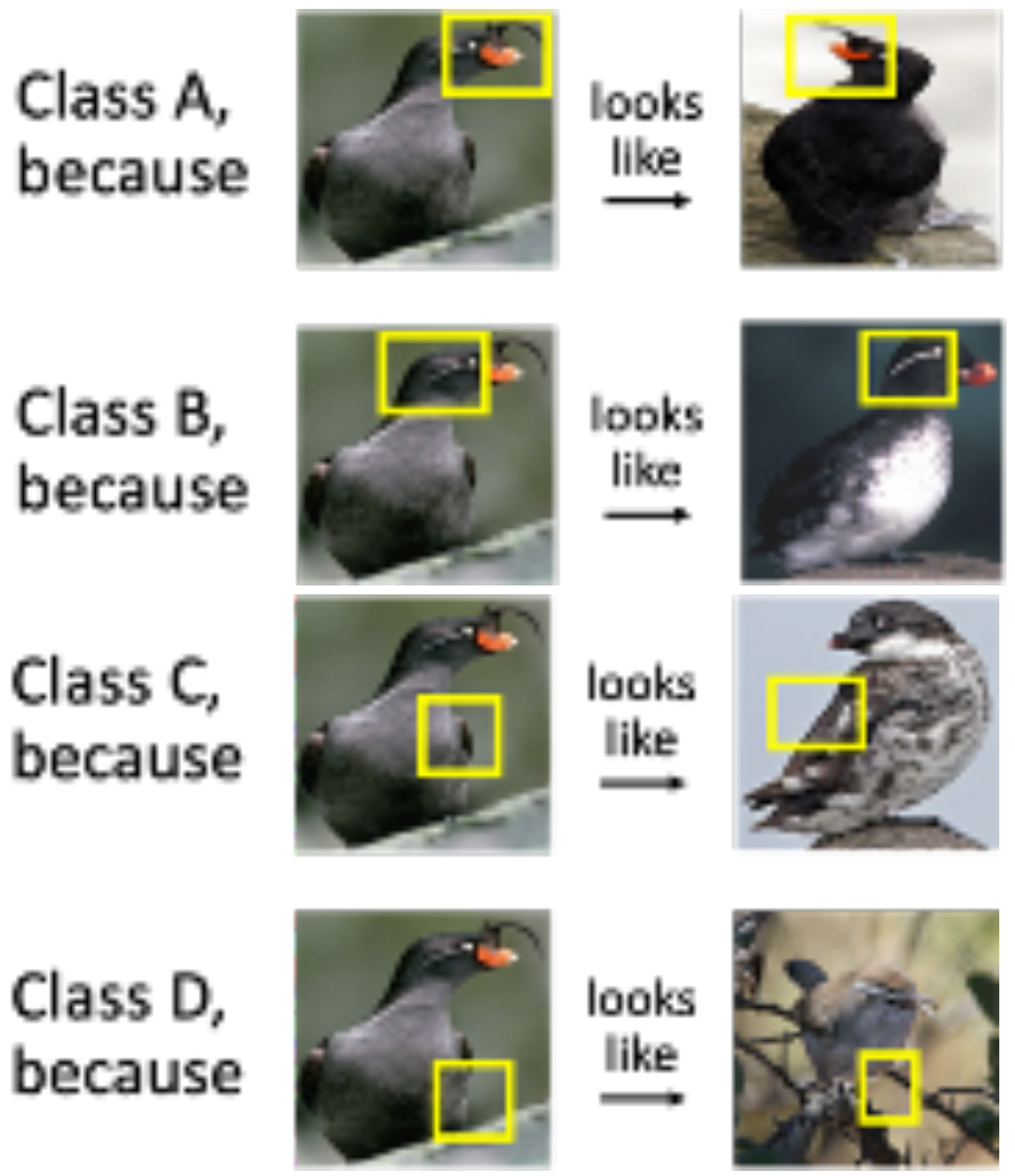
Agreement task

How confident are you in the model's prediction?

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

Distinction task

Which class do you think is correct?



[Sunnie S. Y. Kim et al., ECCV 2022. HIVE.; Chen* & Li* et al., NeurIPS 2019]

ProtoPNet and ProtoTree only

2. Human-centered evaluation

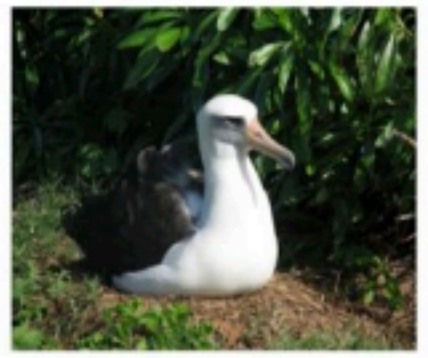
Agreement task

How confident are you in the model's prediction?

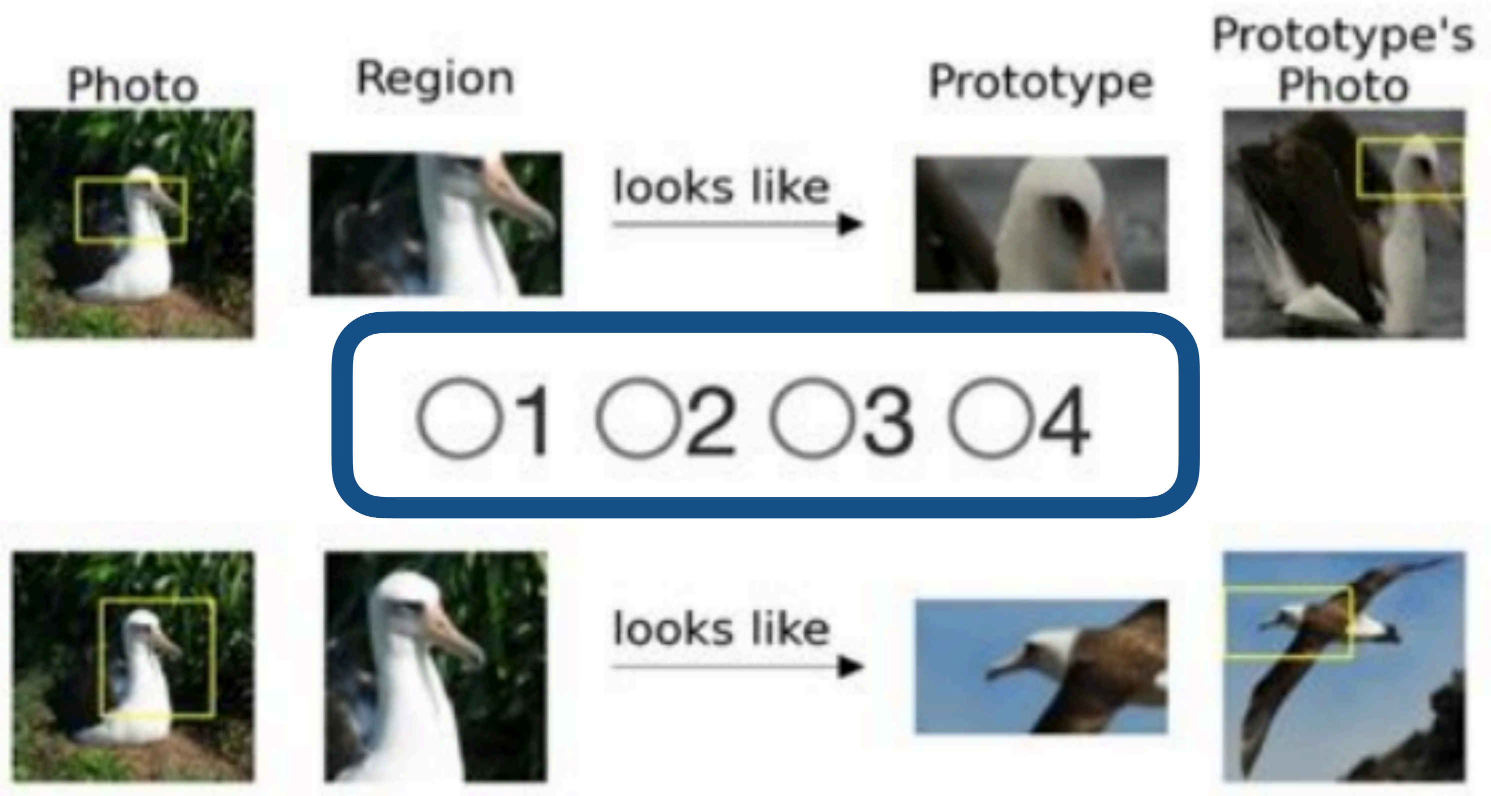
Finding #1: Prototype similarities often **do not align** with human notions of similarity.

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)



Shown below is the model's explanation for its prediction (all prototypes and their source photos are from **Species 2**).



○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., ECCV 2022. HIVE.; Chen* & Li* et al., NeurIPS 2019]

2. Human-centered evaluation

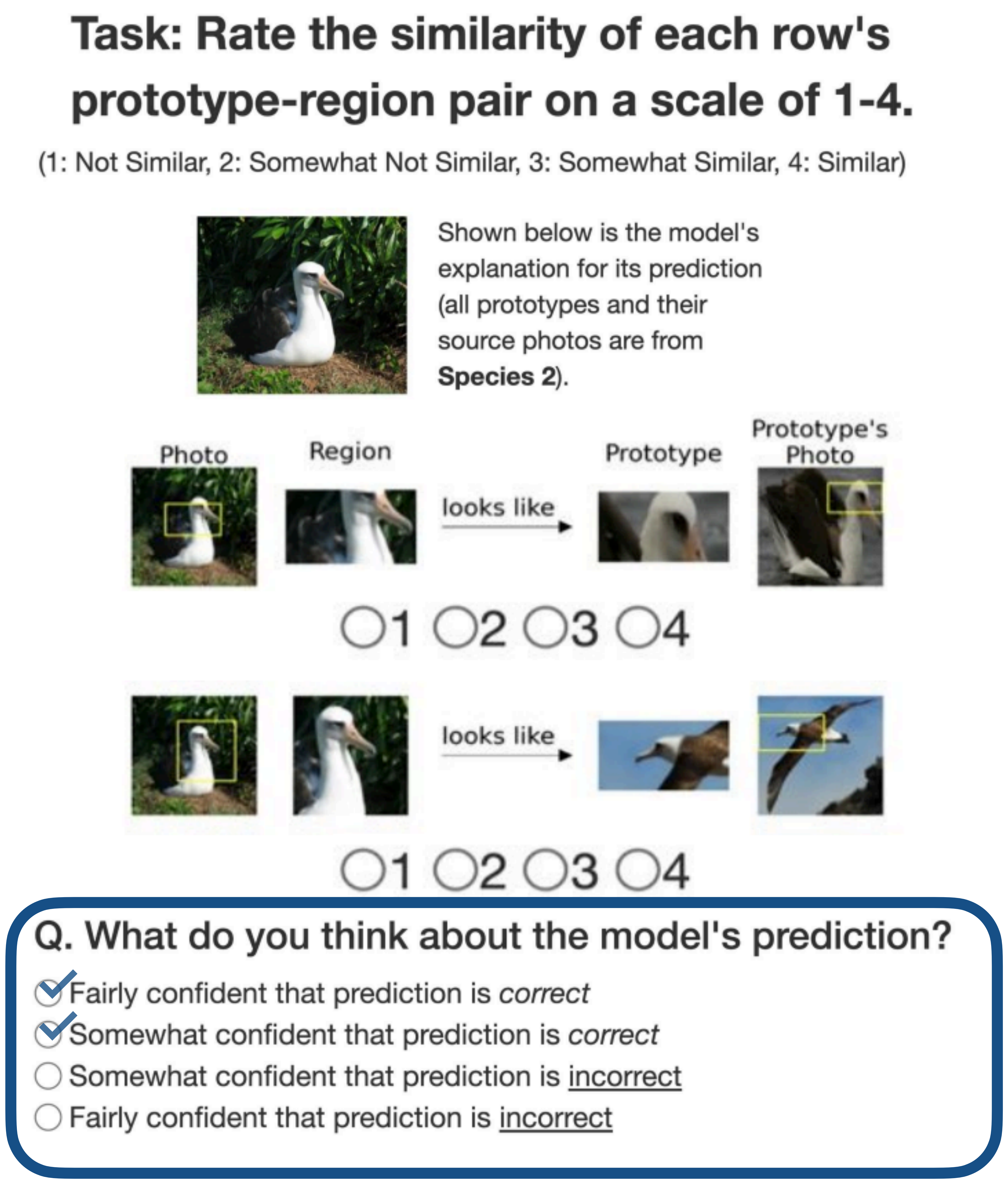
Agreement task

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**.

More than 50% were fairly or somewhat confident that a prediction is correct (even for incorrect predictions).



[Sunnie S. Y. Kim et al., ECCV 2022. HIVE.; Chen* & Li* et al., NeurIPS 2019]

2. Human-centered evaluation

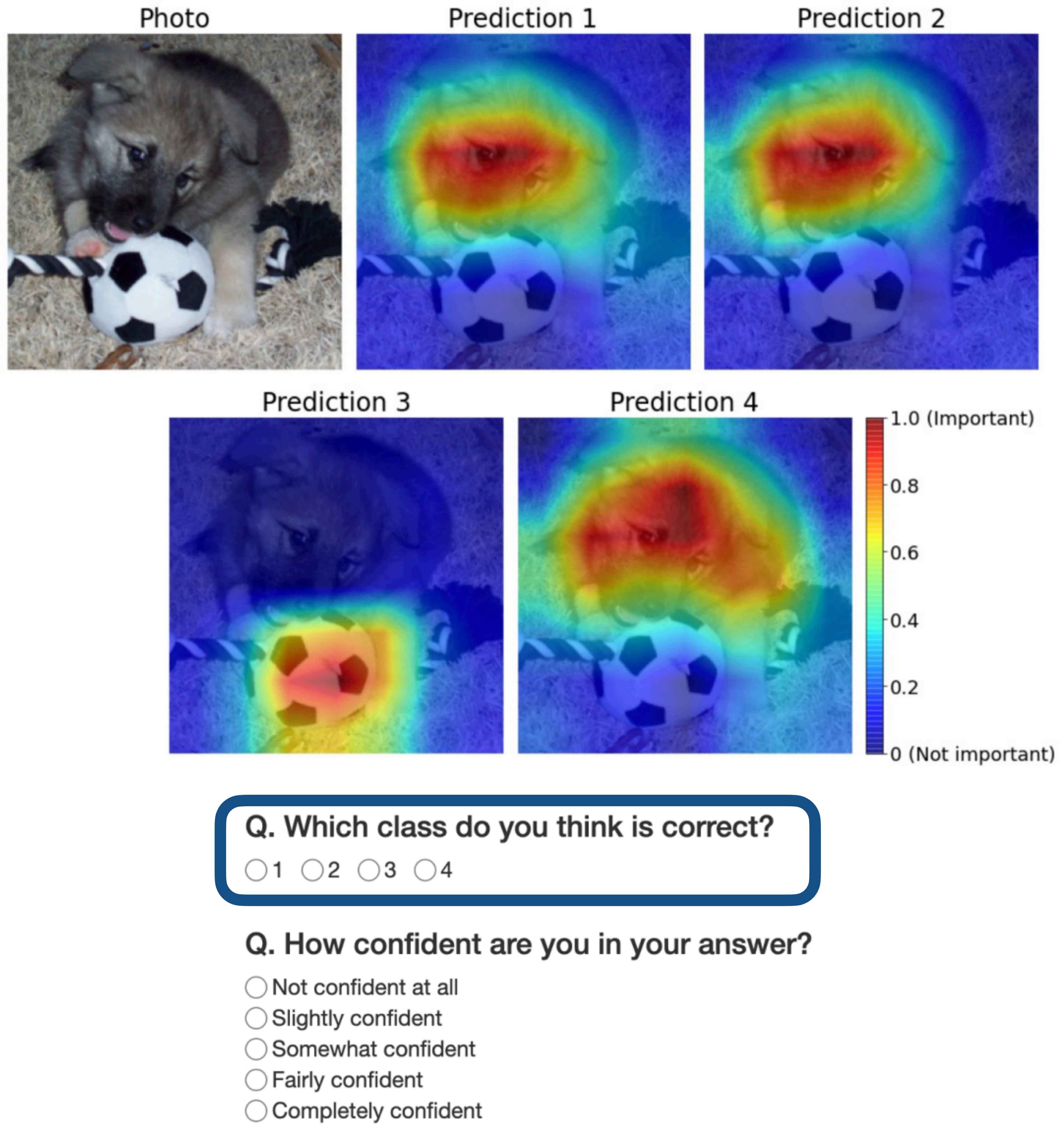
Distinction task

Which class do you think is correct?

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

For incorrect predictions, correctly answered around 25% of the time (**random guessing**).

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



[Sunnie S. Y. Kim et al., ECCV 2022. HIVE.; Selvaraju et al., ICCV 2017]

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., ECCV 2022. 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**.

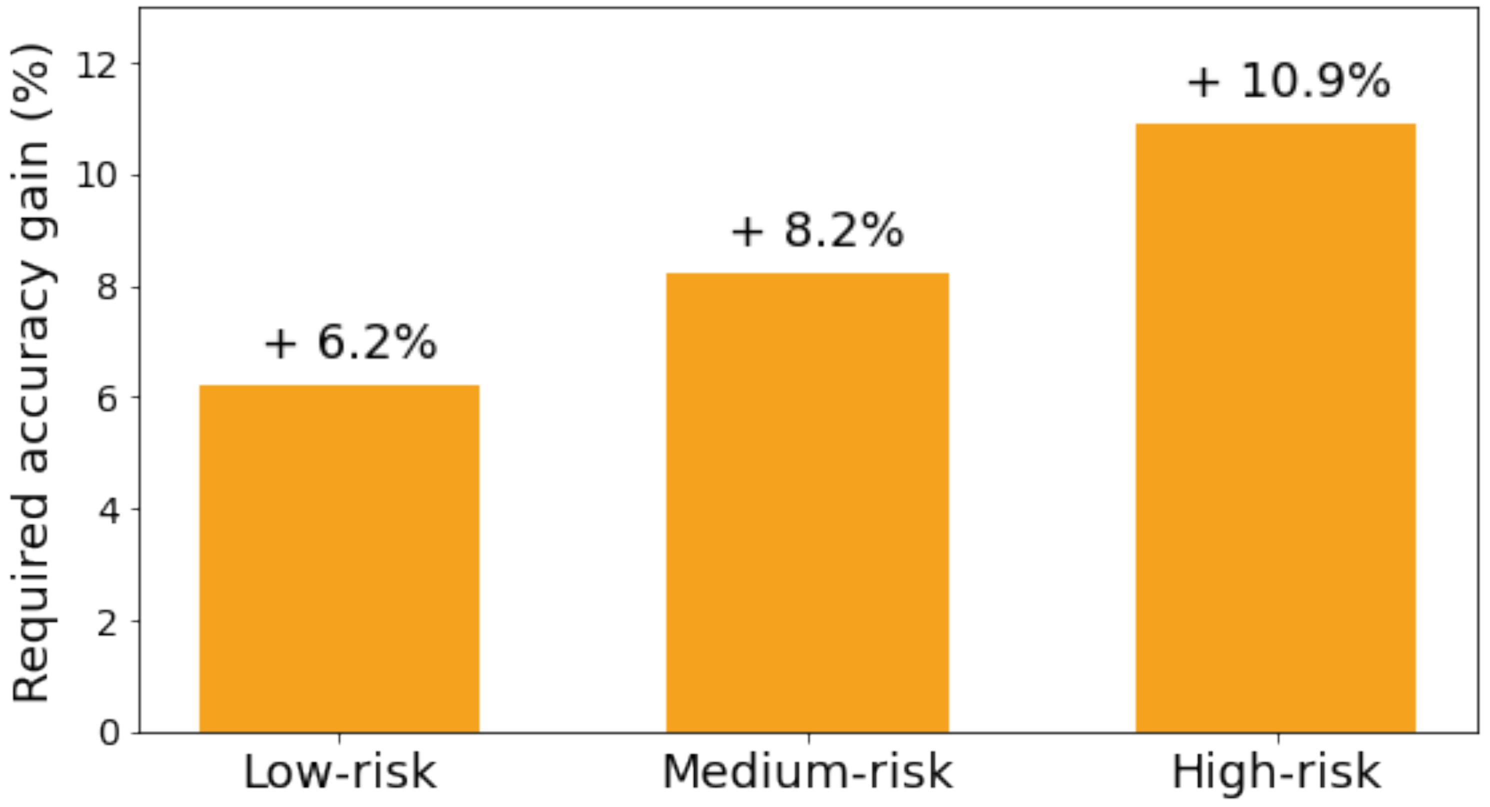
Finding #3: Participants struggle to identify the **correct class**, esp. for incorrect predictions.

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

Follow up: Kim et al., arXiv 2022.
"Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.

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?

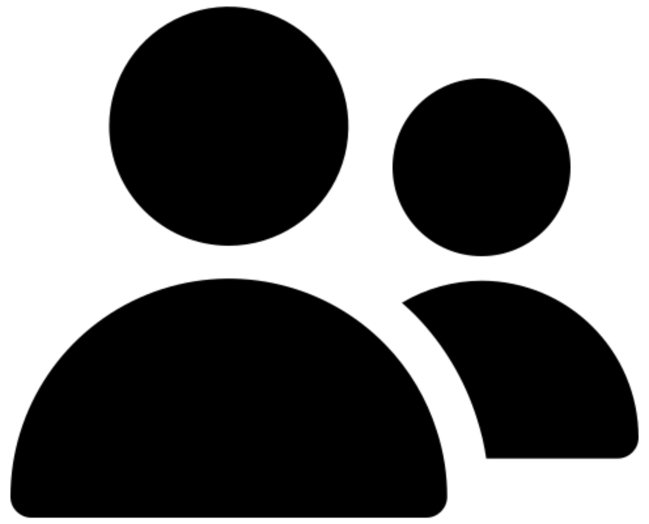


(e.g. educational purposes)
(e.g. biodiversity monitoring)
(e.g. veterinary medicine)

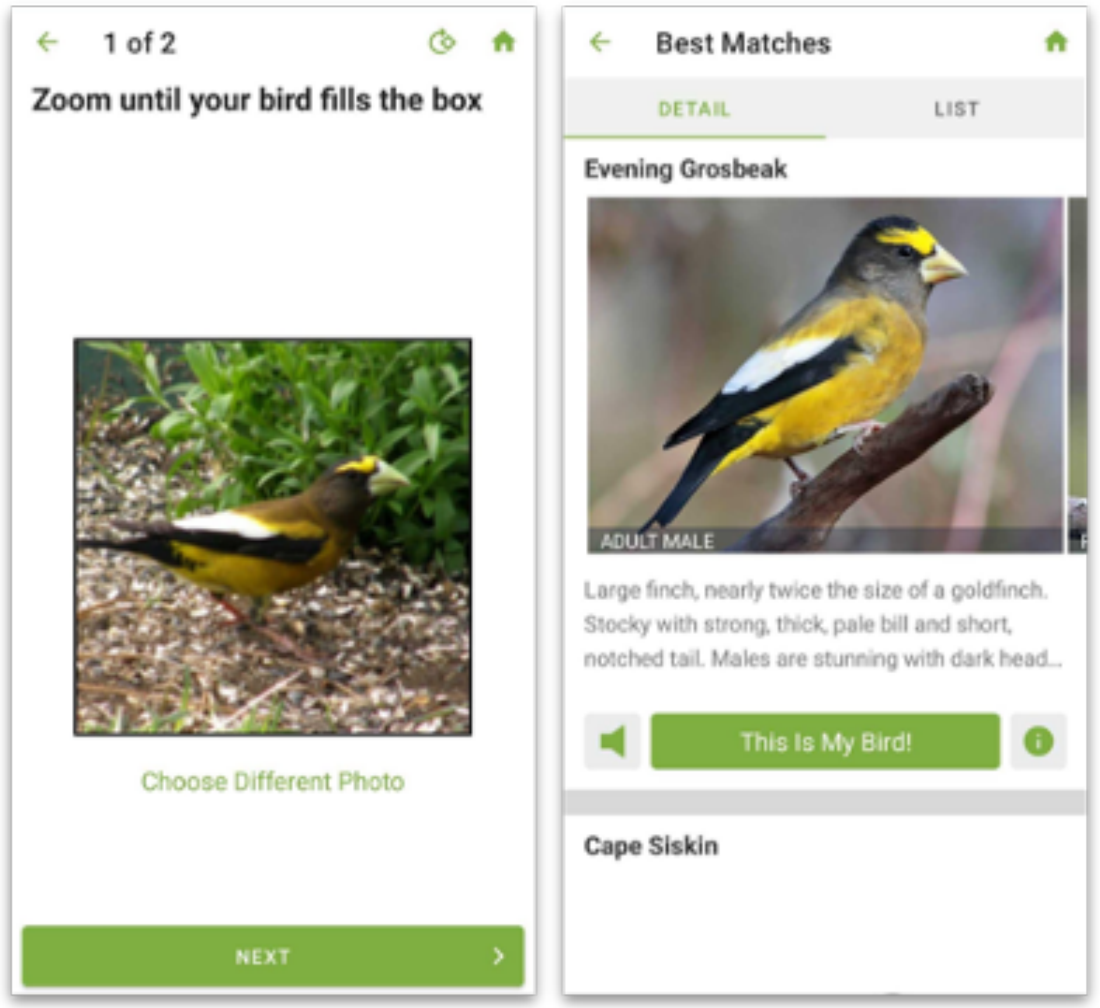
[Sunnie S. Y. Kim et al., ECCV 2022. HIVE.]

Follow up: “Help Me Help the AI” — interview study with Merlin users

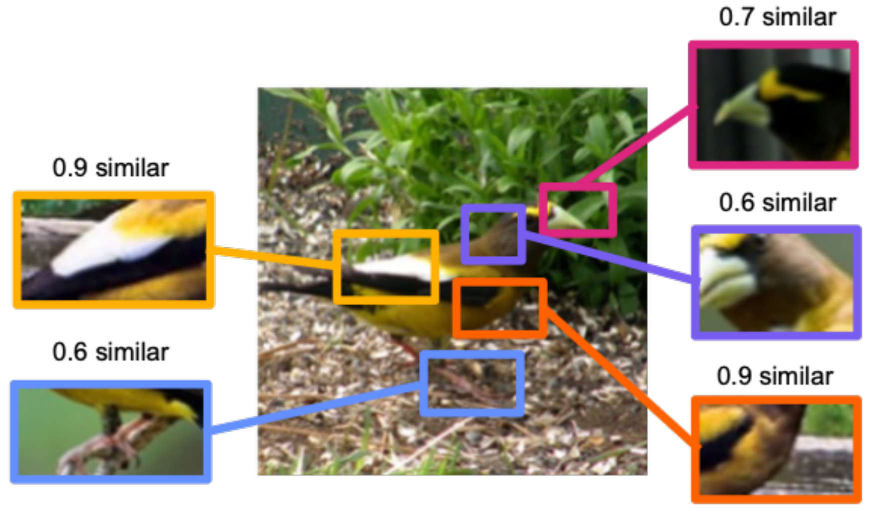
What **kind of explanation** best explains this prediction?



Interview



Merlin app



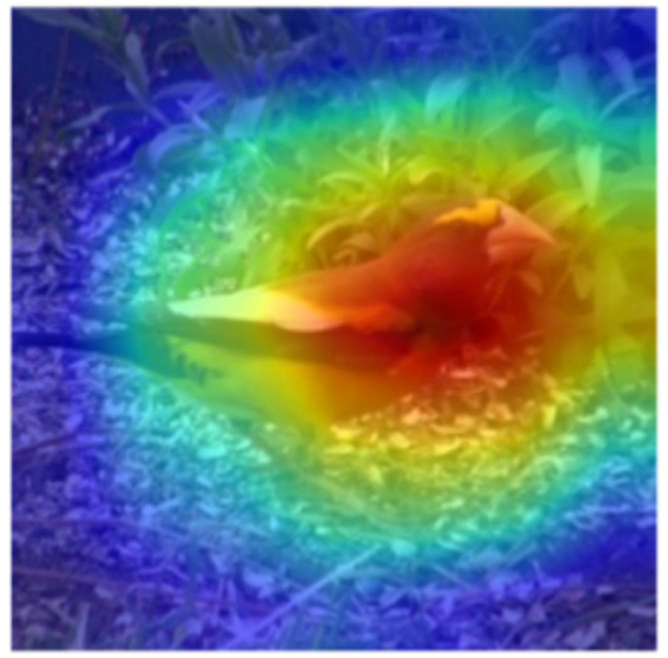
Prototypes

Score for Evening Grosbeak
= 1.7

= ~~- 1.2 long beak~~
+ 1.1 yellow beak
+ 0.8 black feathers
- 0.7 white body
+ 0.5 yellow body
~~+ 0.1 round body~~

...

Concepts



Heatmaps

Examples

Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández, arXiv 2022.
“Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction.

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, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me Help the AI."*)

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.
   (+ *Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors in Concept-based Explanations.*)

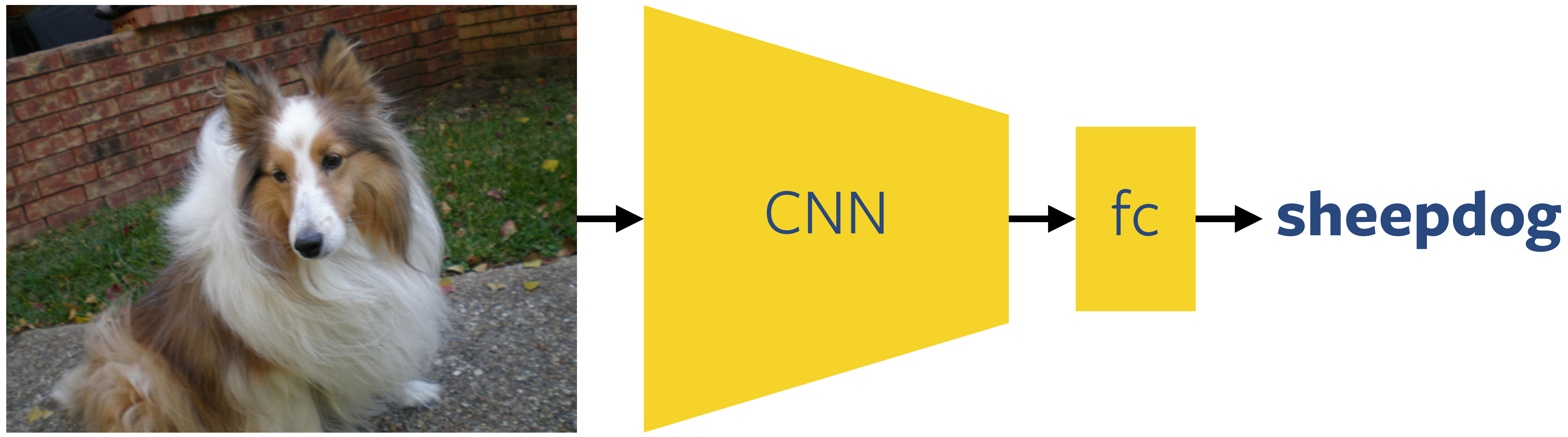
3. Interpretability of **supervised** models → interpretability of **self-supervised** models
   Iro Laina, Ruth Fong, Andrea Vedaldi, NeurIPS 2020.
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Interpretability** in ML + CV → **interdisciplinary** research (interpretability + X)
   (+ *Nicole Meister* and Dora Zhao* et al., arXiv 2022. Gender Artifacts in Visual Datasets.*)
   (+ *Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.*)

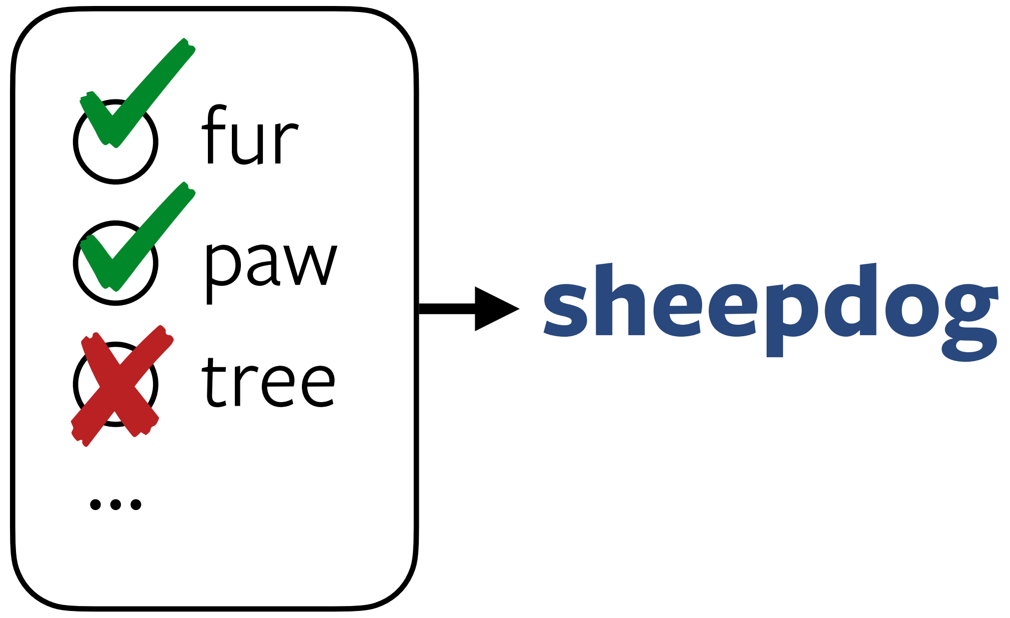
5. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
   (+ *Devon Ulrich and Ruth Fong, in prep. Interactive Visual Feature Search.*)

Concept-based explanations

Why did the model predict **sheepdog**?

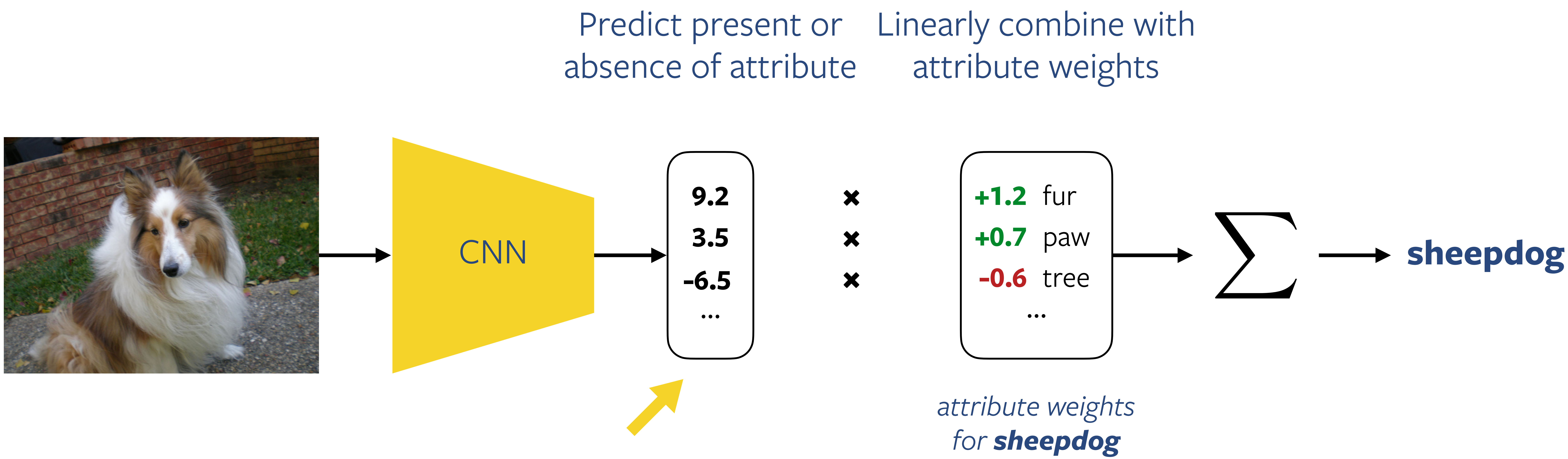


Concept-based explanation



Pro: Labelled concepts are interpretable to humans

Concept Bottleneck: Linear Combination of Labelled Attributes

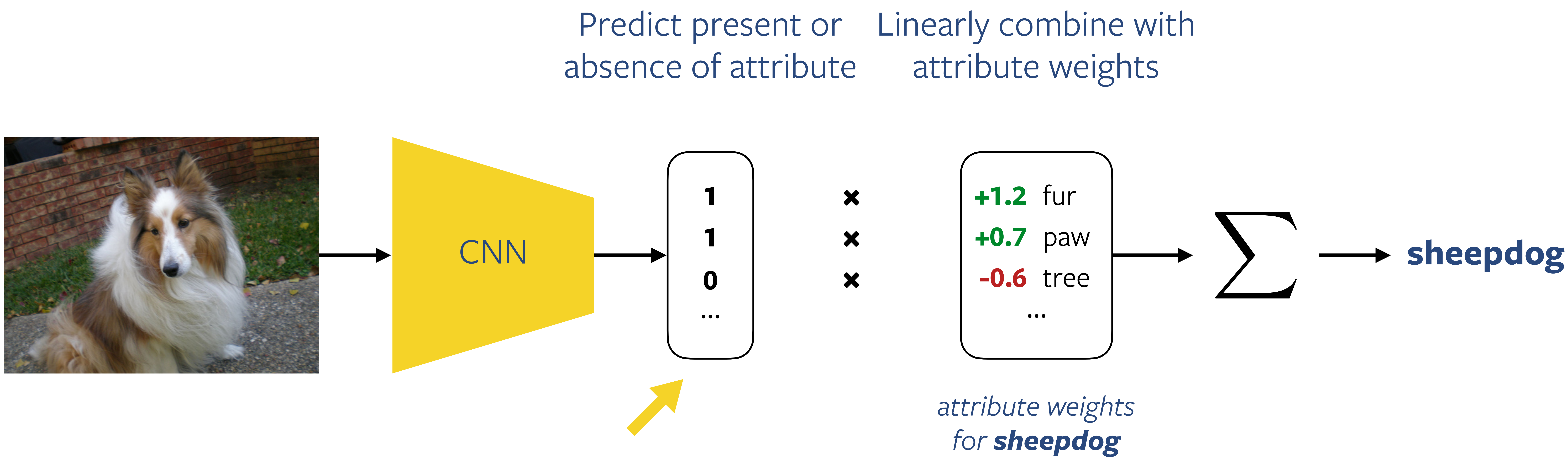


Con: Problems with predicting fractional values

- hard to interpret
- can encode hidden information

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

Concept Bottleneck: Linear Combination of Labelled Attributes

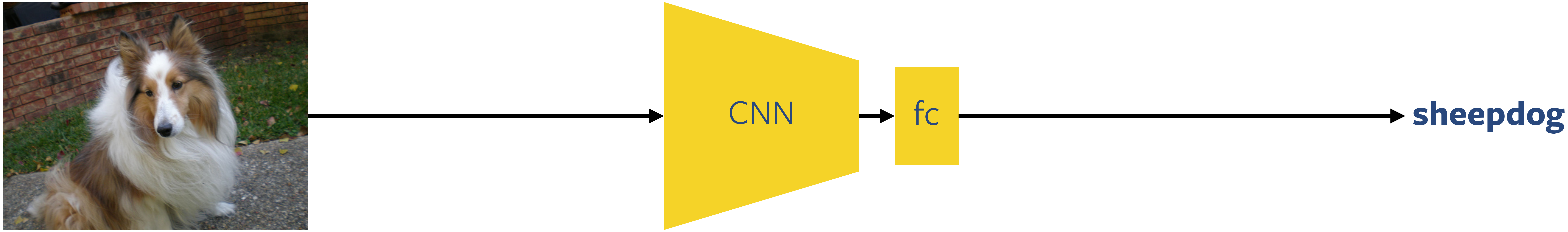


Con: Problems with predicting fractional values

- hard to interpret
- can encode hidden information

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

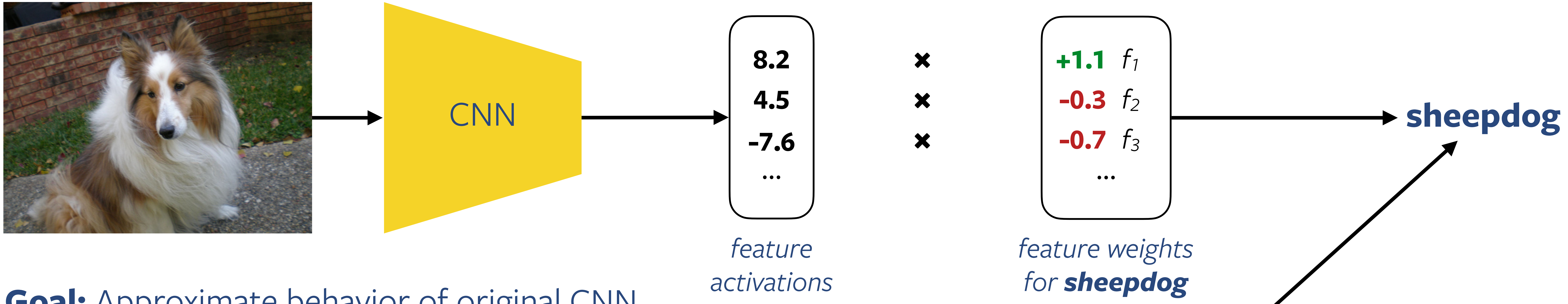
ELUDE: **E**xplanation via a **L**abelled and **U**nlabelled **DE**composition of features



Goal: Approximate behavior of original CNN

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

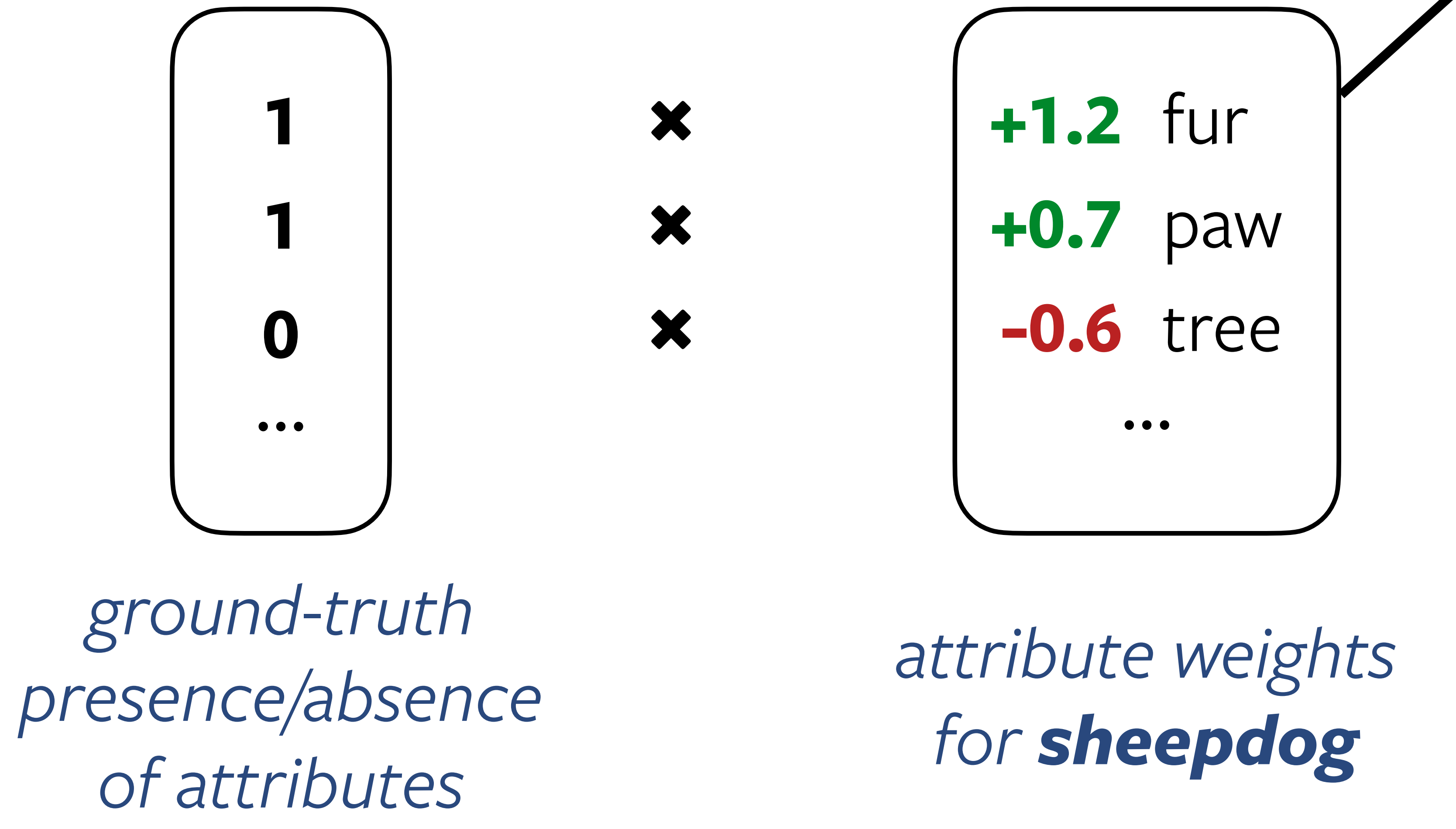
ELUDE: Decomposition of labelled and unlabelled features



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.]

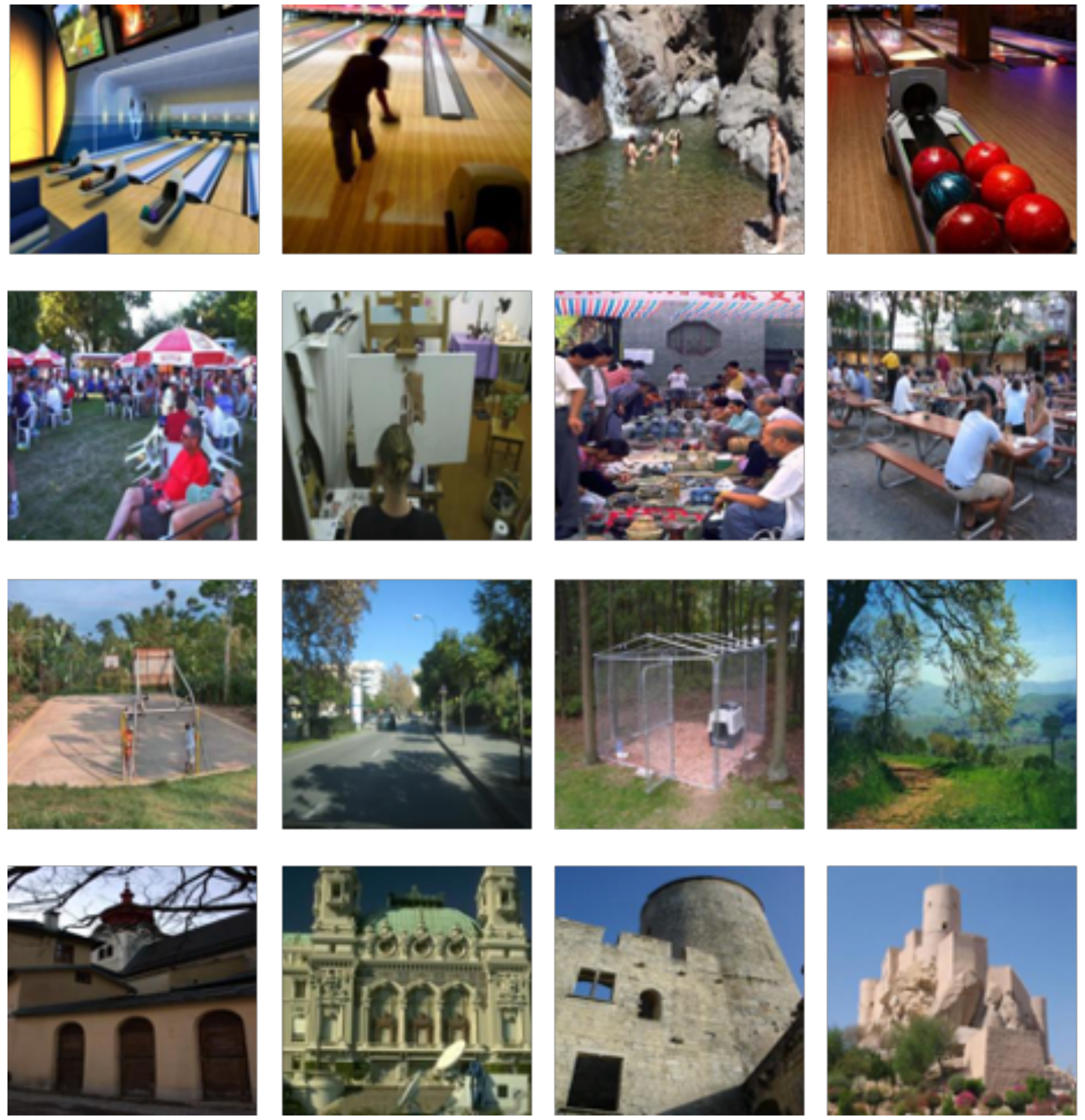
Attributes only: % of model explained via labelled attributes decreases as task complexity increases

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

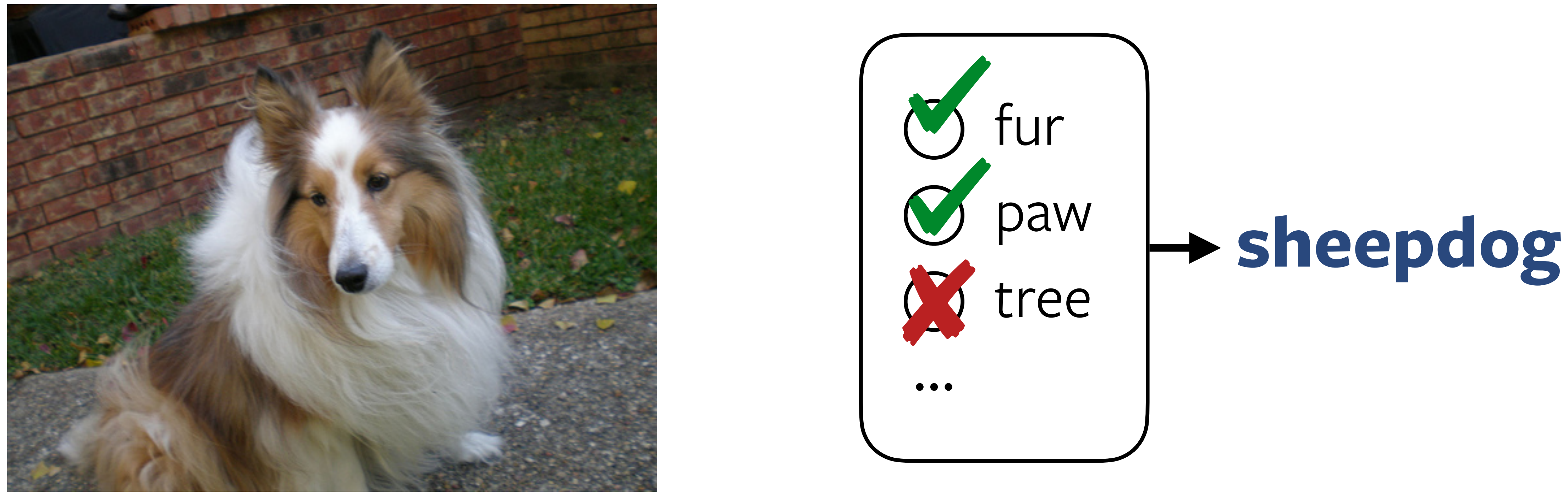
[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]

Follow up: Overlooked factors in concept-based explanations

Factor #1: Probe dataset choice matters (i.e. different datasets → different explanations).

Factor #2: Some concepts used in explanations are harder to learn than output classes.

Factor #3: Humans can reason with a small amount of concepts (i.e. max 32 concepts).



Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky, arXiv 2022.
Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability.

Follow up: Overlooked factors in concept-based explanations

Factor #1: Probe dataset choice matters (i.e. different datasets → different explanations).

Factor #2: Some concepts used in explanations are harder to learn than output classes.

Factor #3: Humans can reason with a small amount of concepts (i.e. max 32 concepts).

Suggestion: Choose a probe dataset with a similar distribution to that of the training dataset.

Training dataset: Places365

hockey arena

Probe dataset:

ADE20k
{grandstand, goal, ice rink, scoreboard}

Pascal
{plaything, road}

Concepts used to explain **hockey arena** differ based on probe dataset.

Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors.

Follow up: Overlooked factors in concept-based explanations

Factor #1: Probe dataset choice matters (i.e. different datasets → different explanations).

Factor #2: Some concepts used in explanations are harder to learn than output classes.

Factor #3: Humans can reason with a small amount of concepts (i.e. max 32 concepts).

Suggestion: Only use easily learnable concepts in concept-based explanations.

Training dataset: Places365

bathroom
(norm AP = 43.3)

Probe dataset: Broden

Concept	norm AP
toilet	39.9
shower	18.8
countertop	12.6
bathtub	11.1
screen door	9.6

The class **bathroom** is easier to learn than the concepts used to explain it.

Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors.

Follow up: Overlooked factors in concept-based explanations

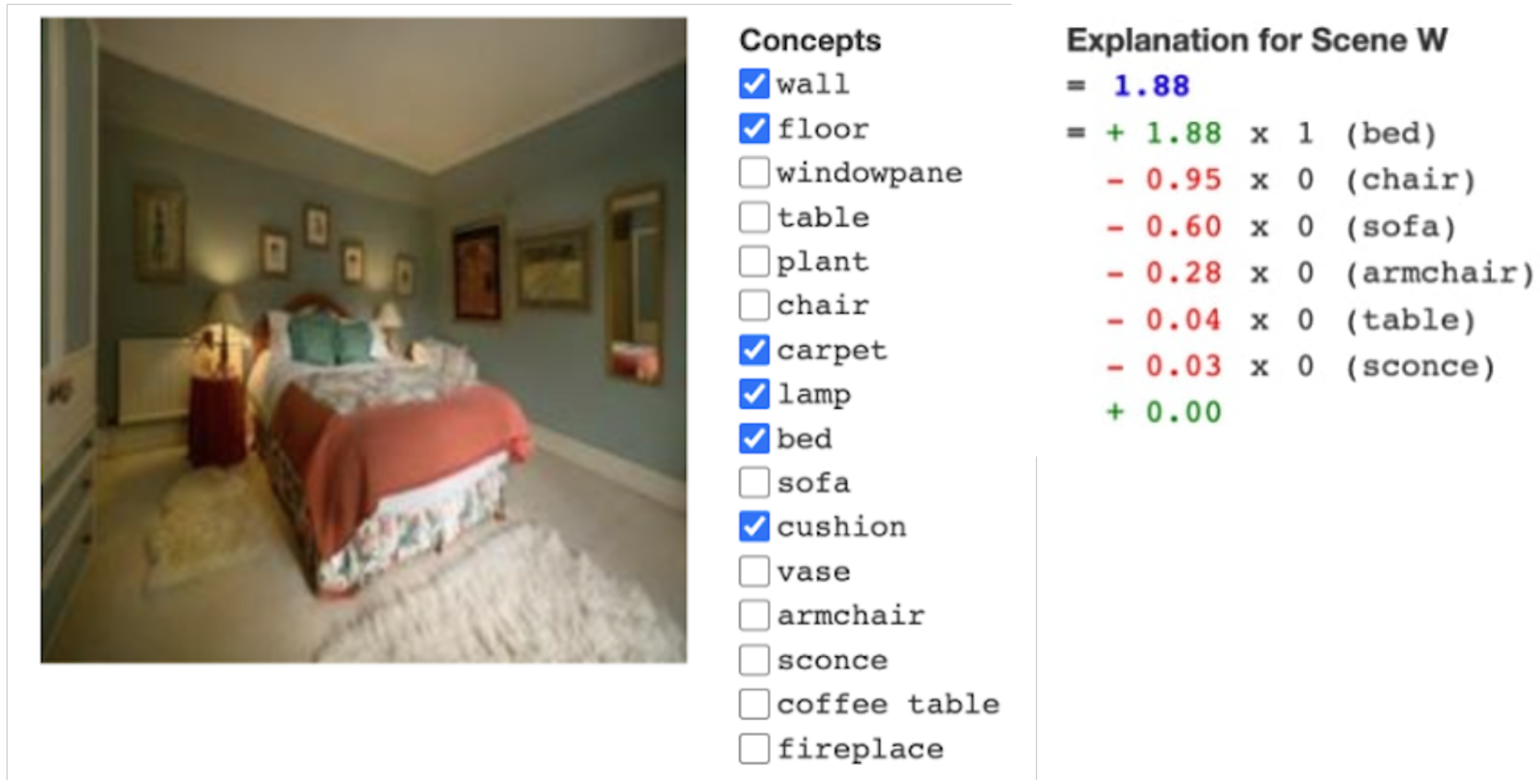
Factor #1: Probe dataset choice matters (i.e. different datasets → different explanations).

Factor #2: Some concepts used in explanations are harder to learn than output classes.

Factor #3: Participants can reason with a small amount of concepts (i.e. max 32 concepts).

1. Which scene do you think the model predicts?
2. How many concepts would you prefer?

AMT human study
(*N = 125 participants*)



Participants struggle to identify concepts as the number of concepts increases.
(71.7% for 8 concepts; 56.8% for 32 concepts)

Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors.

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, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me Help the AI."*)
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.
   (+ *Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors in Concept-based Explanations.*)
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. **Interpretability** in ML + CV → **interdisciplinary** research (interpretability + X)
   (+ *Nicole Meister* and Dora Zhao* et al., arXiv 2022. Gender Artifacts in Visual Datasets.*)
   (+ *Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.*)
5. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
   (+ *Devon Ulrich and Ruth Fong, in prep. Interactive Visual Feature Search.*)

Supervised Learning

( , sheepdog)

x **y**

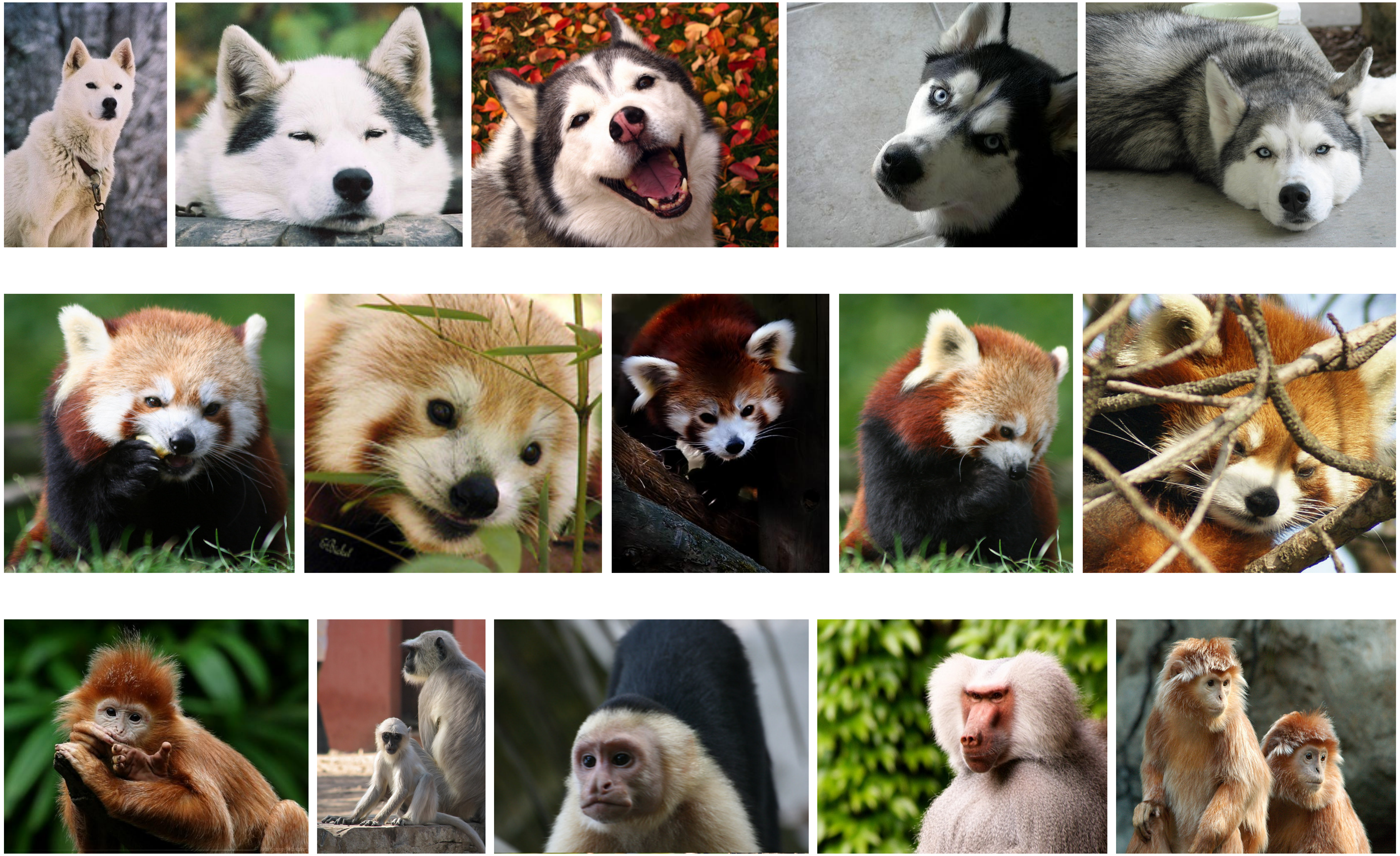
Self-Supervised Learning

x

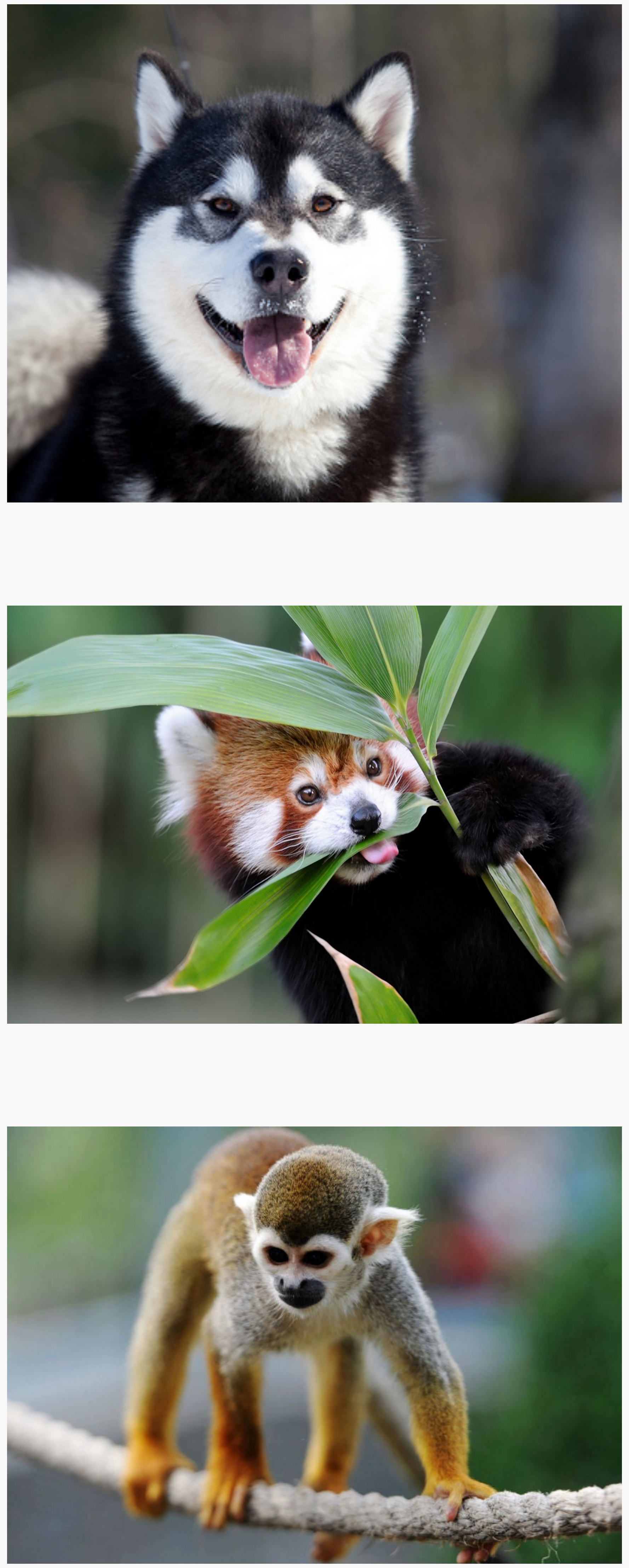
Visual Concept

Query

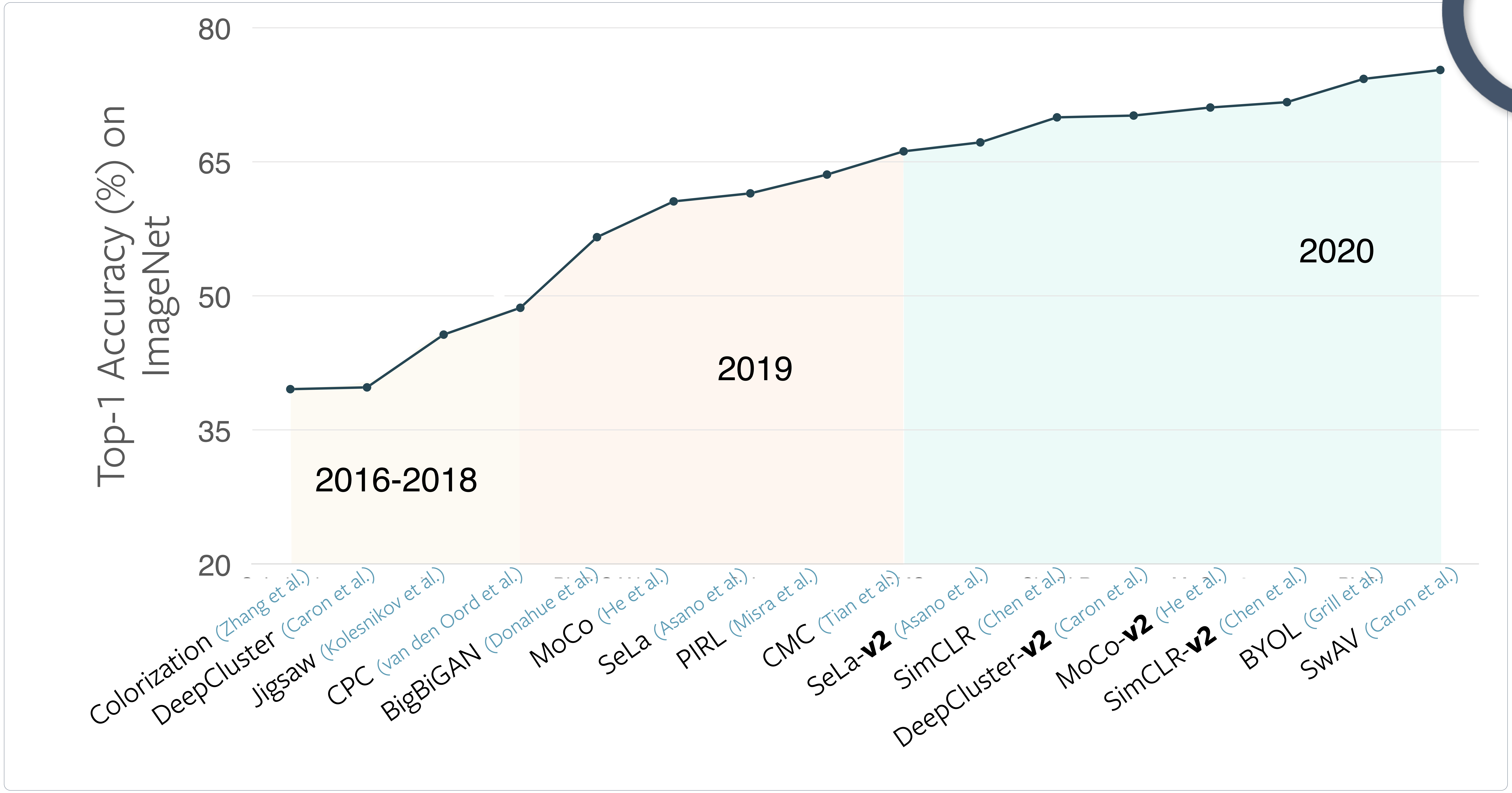
Visual Concept



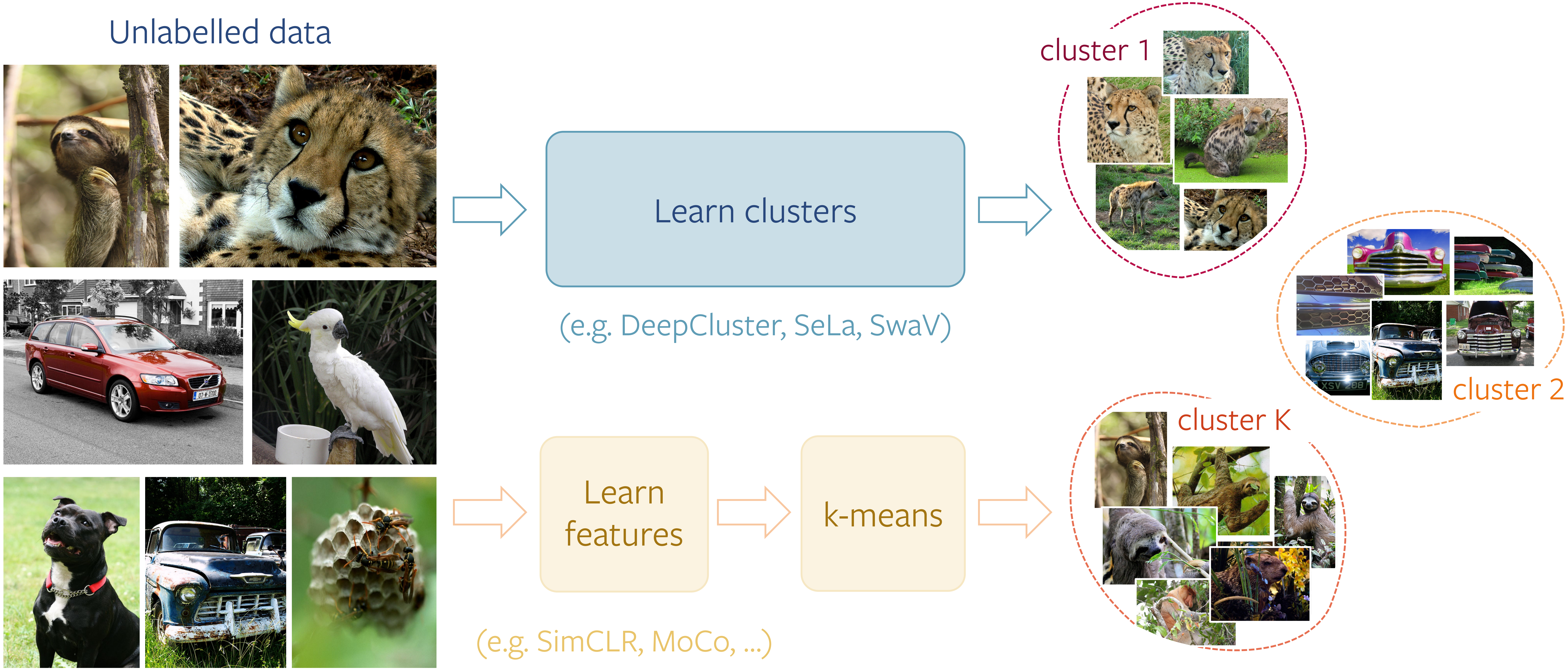
Query



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

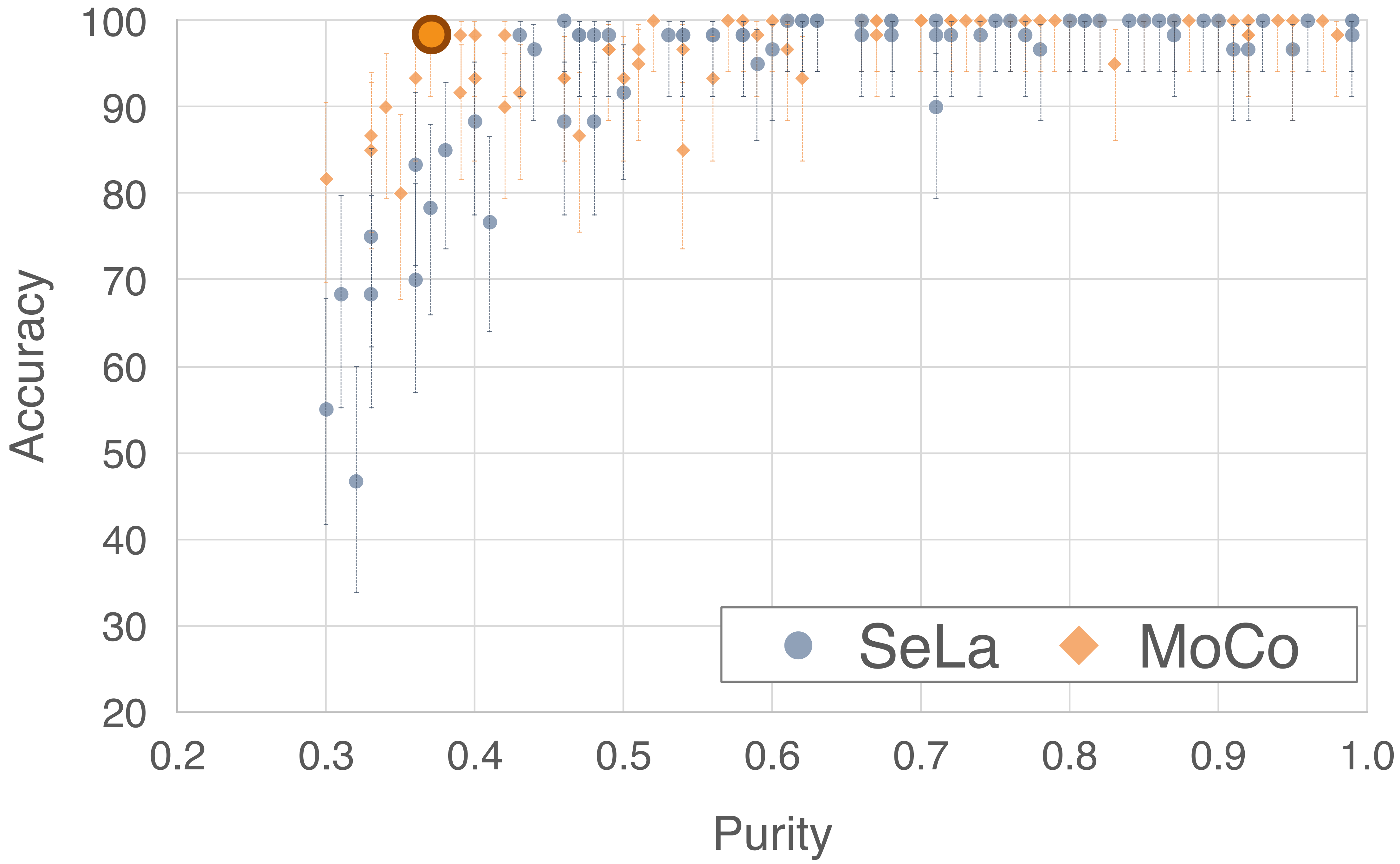
(A)

(B)

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

Evaluation

Learnability 

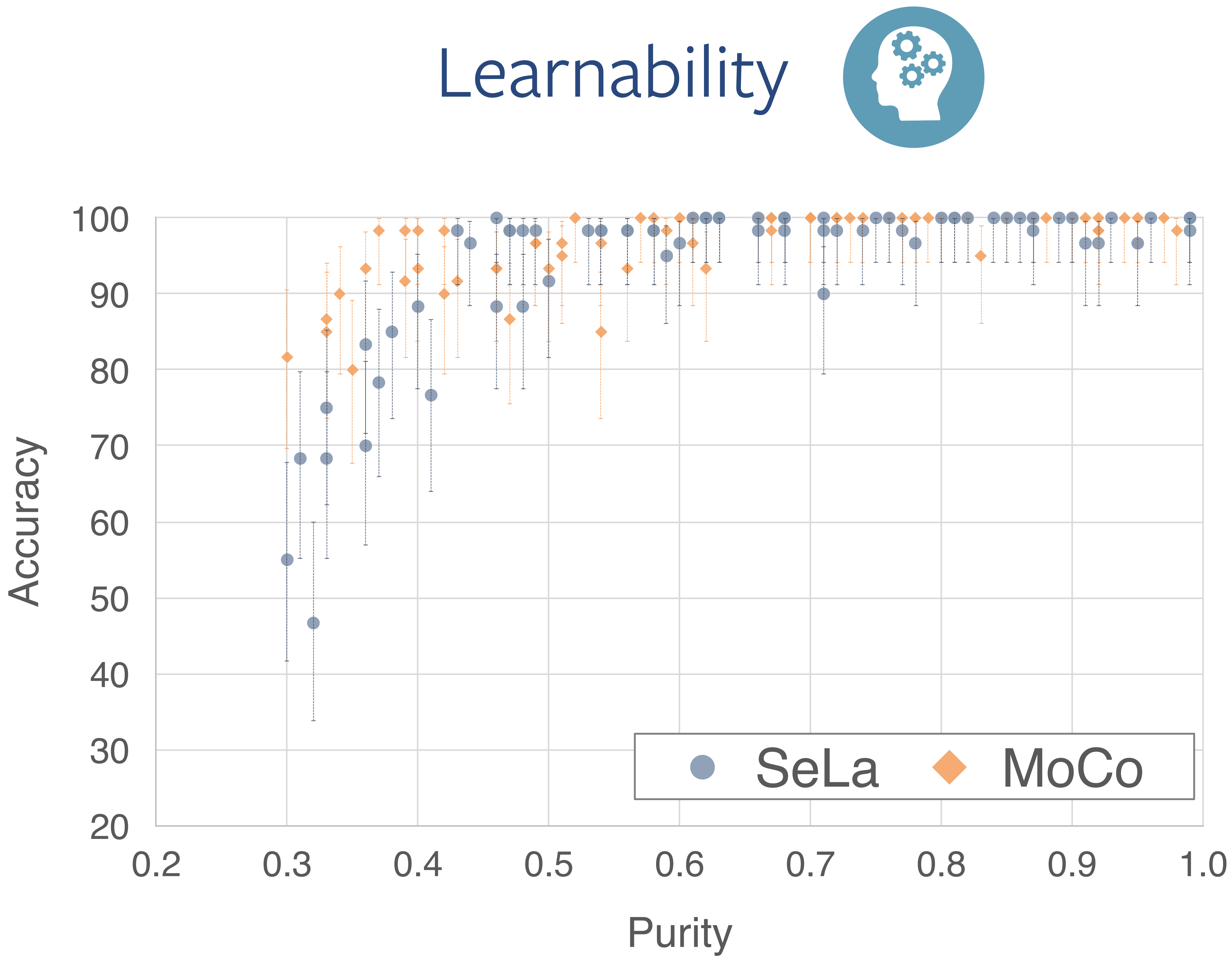


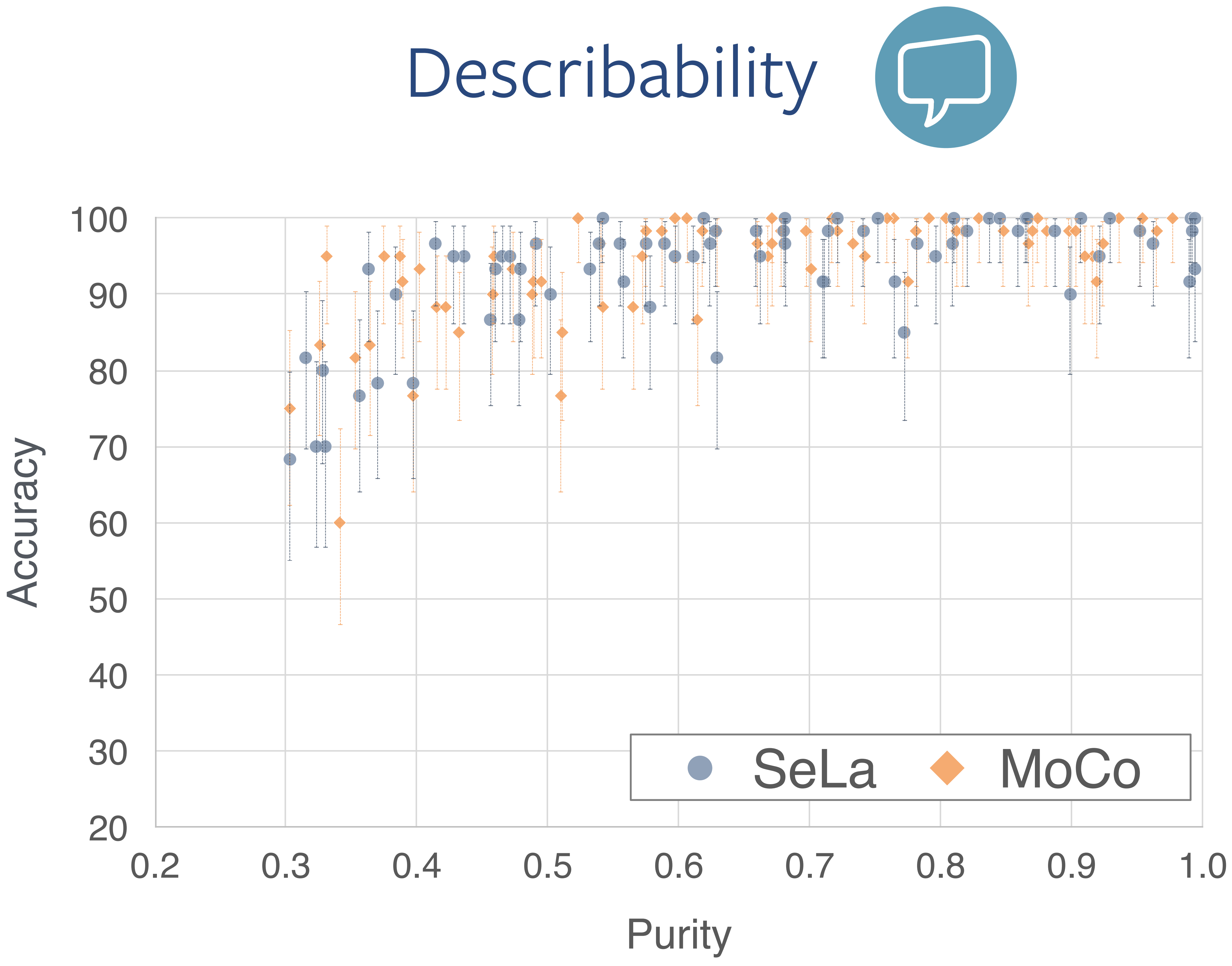
ImageNet cluster purity:
how correlated is a cluster's contents
to a single ImageNet label?

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





[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

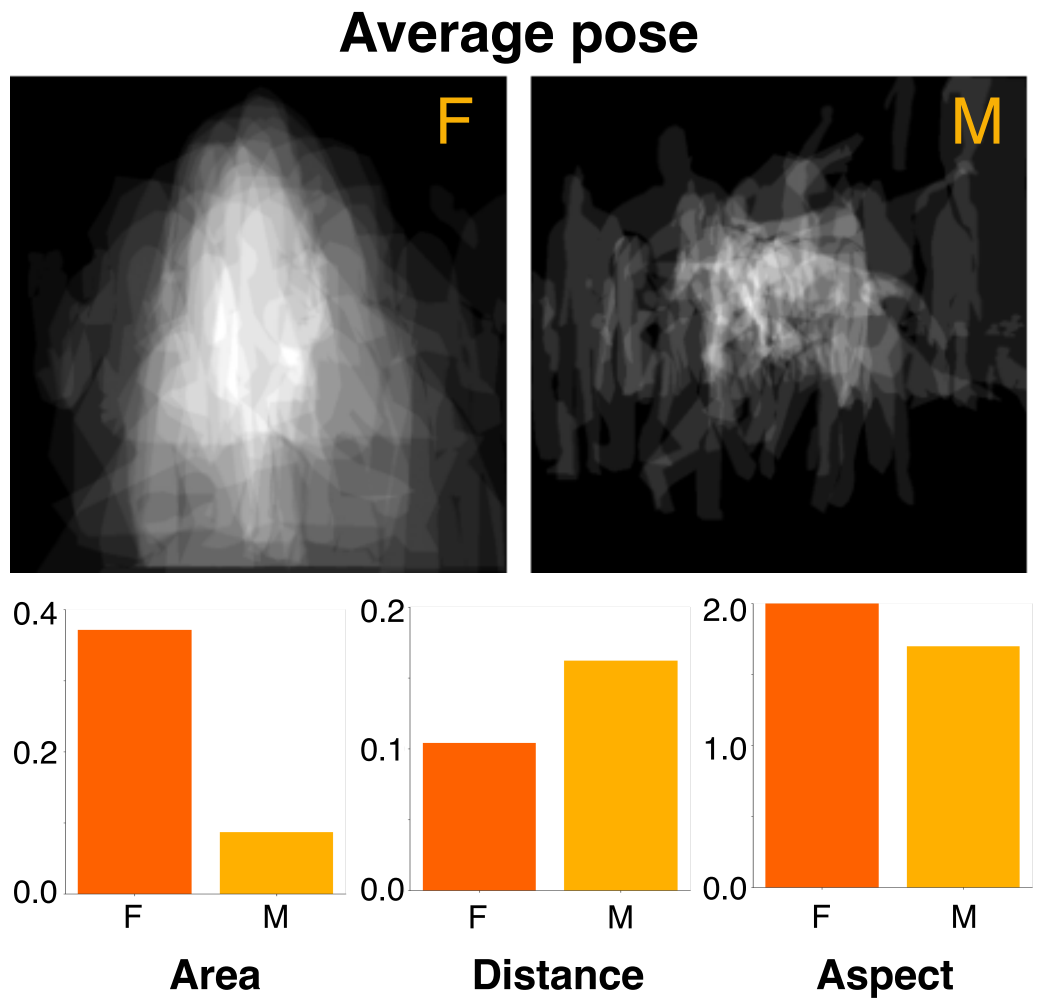
93.3% 95.0%

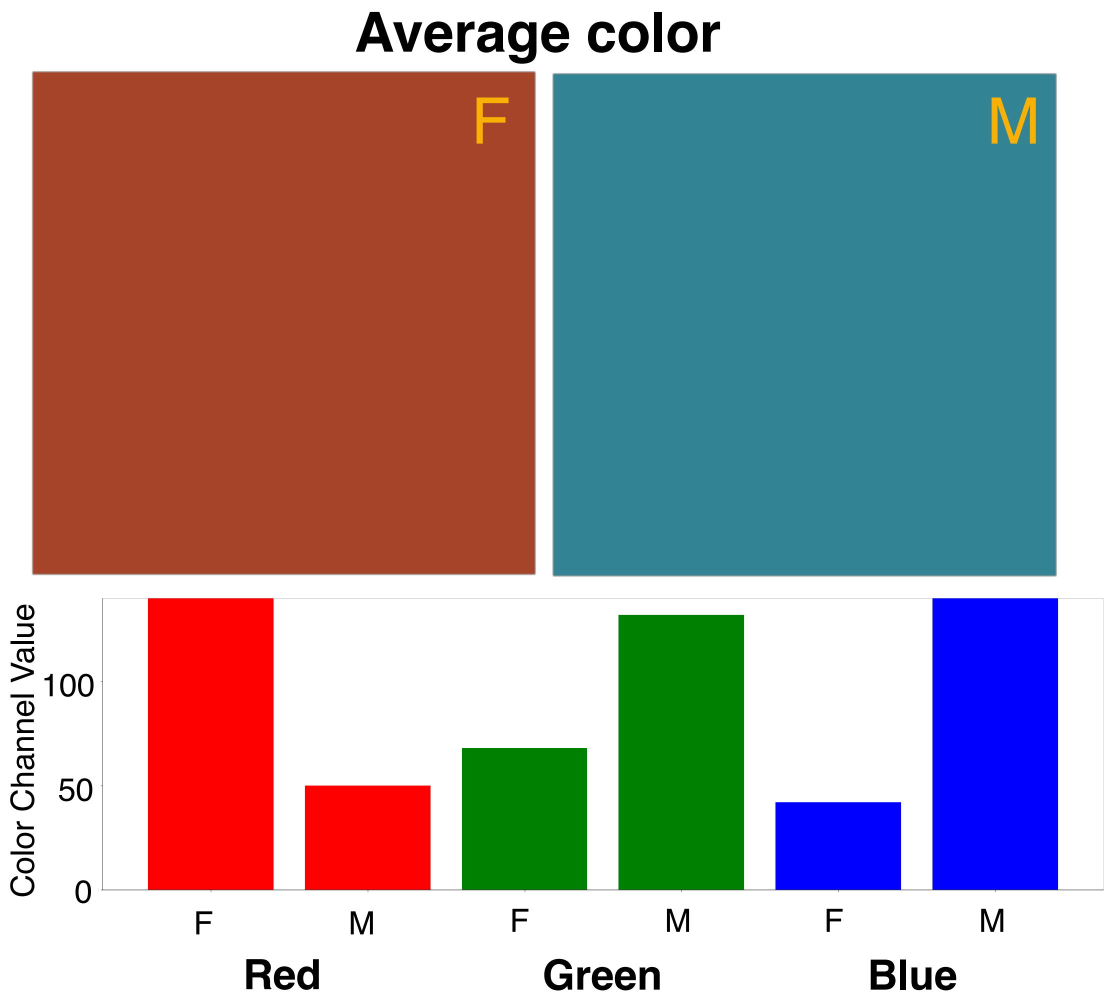
[Iro Laina, et al., NeurIPS 2020. Quantifying Learnability and Describability.]
[Asano et al., ICLR 2020; He et al., CVPR 2020]

ML fairness cross-talk: Gender artifacts in CV

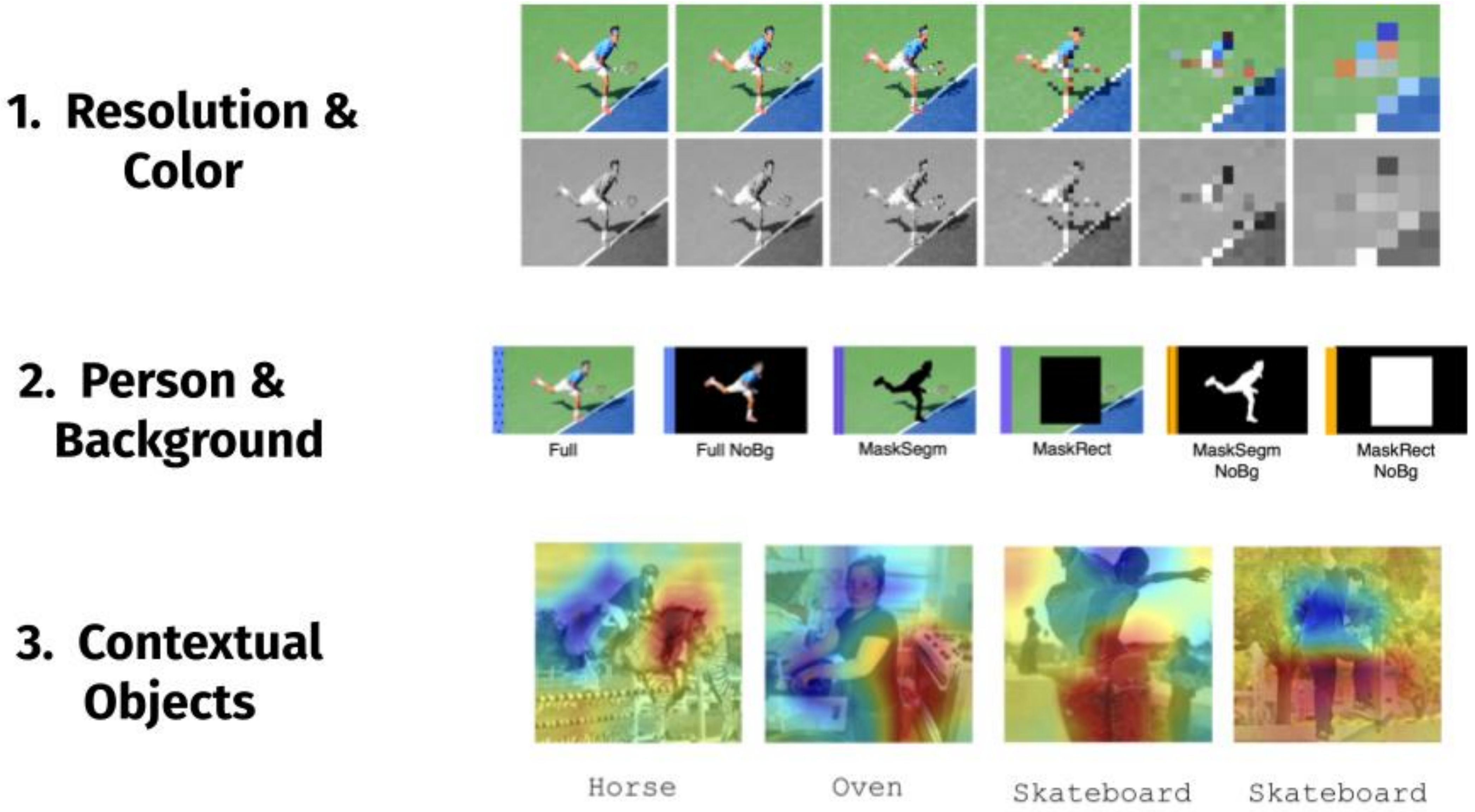
Nicole Meister

Dora Zhao





Differences in top 20 female vs. male predicted images.*



Gender artifacts are **everywhere** in visual datasets.

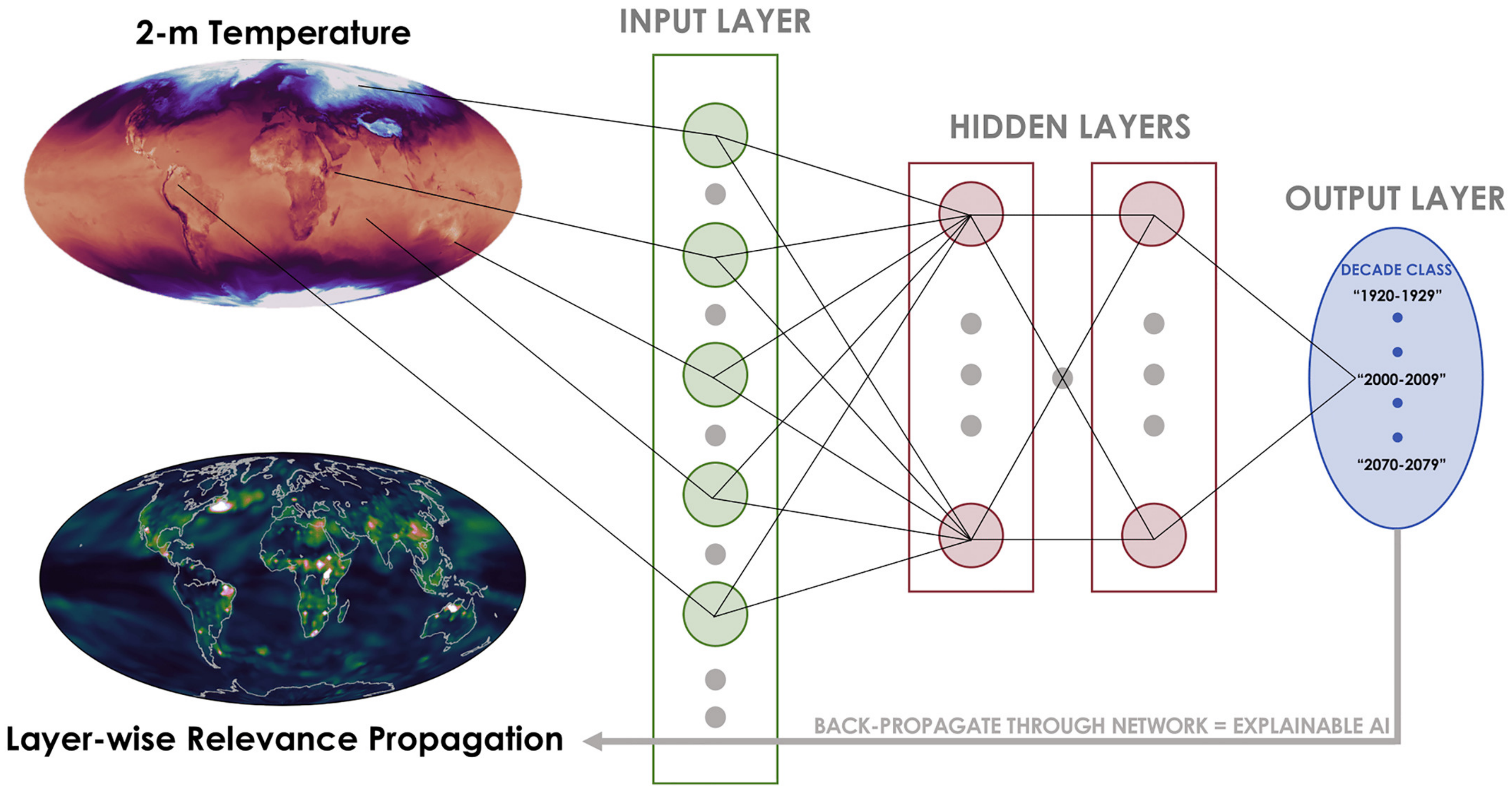
(* *binary perceived gender expression; we do not condone gender prediction.*)

Nicole Meister*, Dora Zhao*, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, arXiv 2022. Gender Artifacts in Visual Datasets.

Extending Interpretability to Geosciences

Indu Panigrahi

Elizabeth Barnes



Understand and improve
a coral reef fossil segmentation model
(our work)

Identify important regions in the world that
reliably predict seasonal climate
(Elizabeth Barnes' group at Colorado State)

Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.
Zachary M. Labe and Elizabeth A. Barnes, JAMES 2021. Detecting Climate Signals Using Explainable AI.

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, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me Help the AI."*)

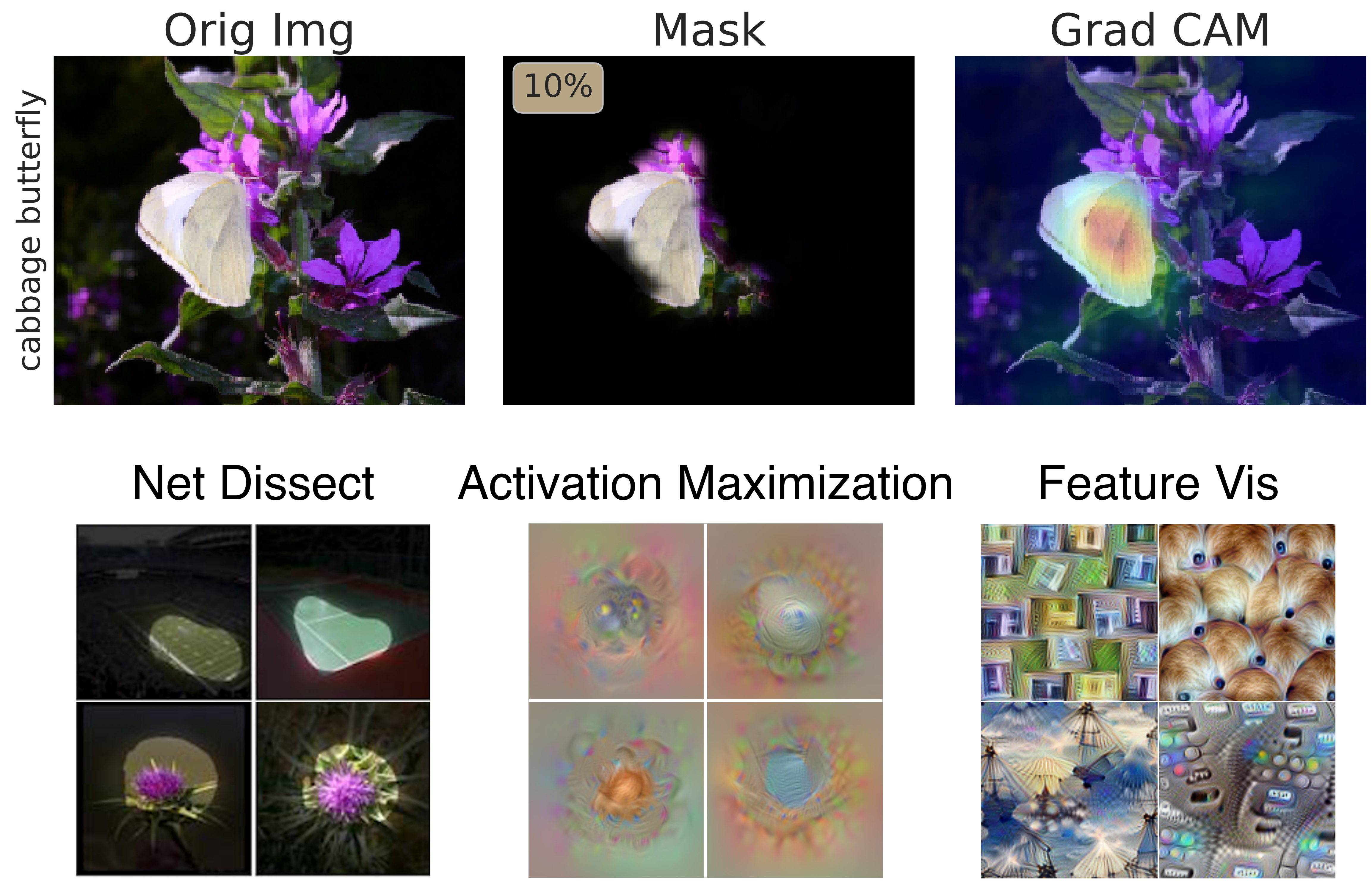
2. Explanations via **labelled attributes** → explanations via **labelled attributes and unlabelled features**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky, arXiv 2022.
   ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features.
   (+ *Vikram V. Ramaswamy et al., arXiv 2022. Overlooked Factors in Concept-based Explanations.*)

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. **Interpretability** in ML + CV → **interdisciplinary** research (interpretability + X)
   (+ *Nicole Meister* and Dora Zhao* et al., arXiv 2022. Gender Artifacts in Visual Datasets.*)
   (+ *Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.*)

5. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.
   (+ *Devon Ulrich and Ruth Fong, in prep. Interactive Visual Feature Search.*)

Interpretability Tools

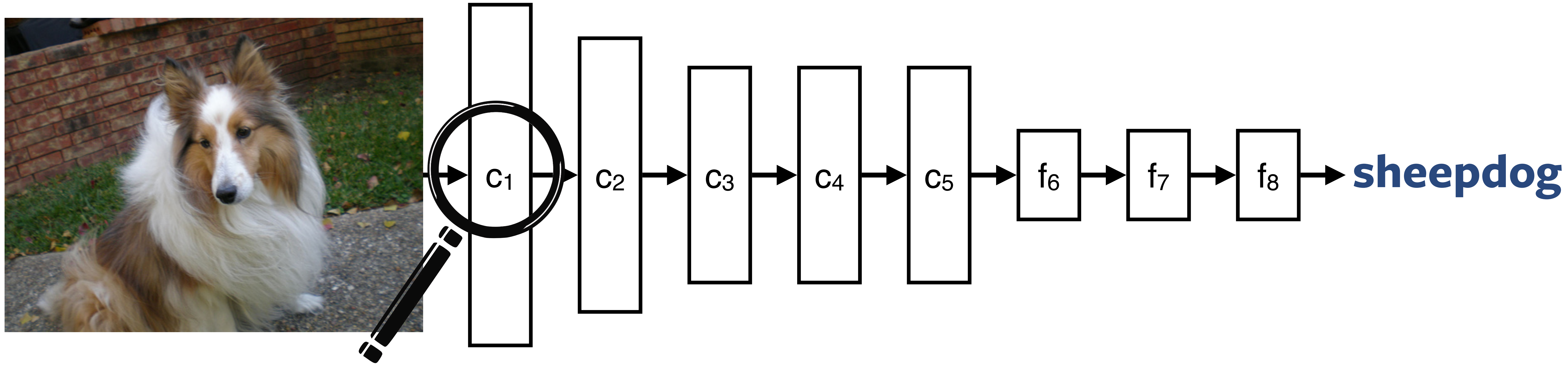


Current tools render **static images**.

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



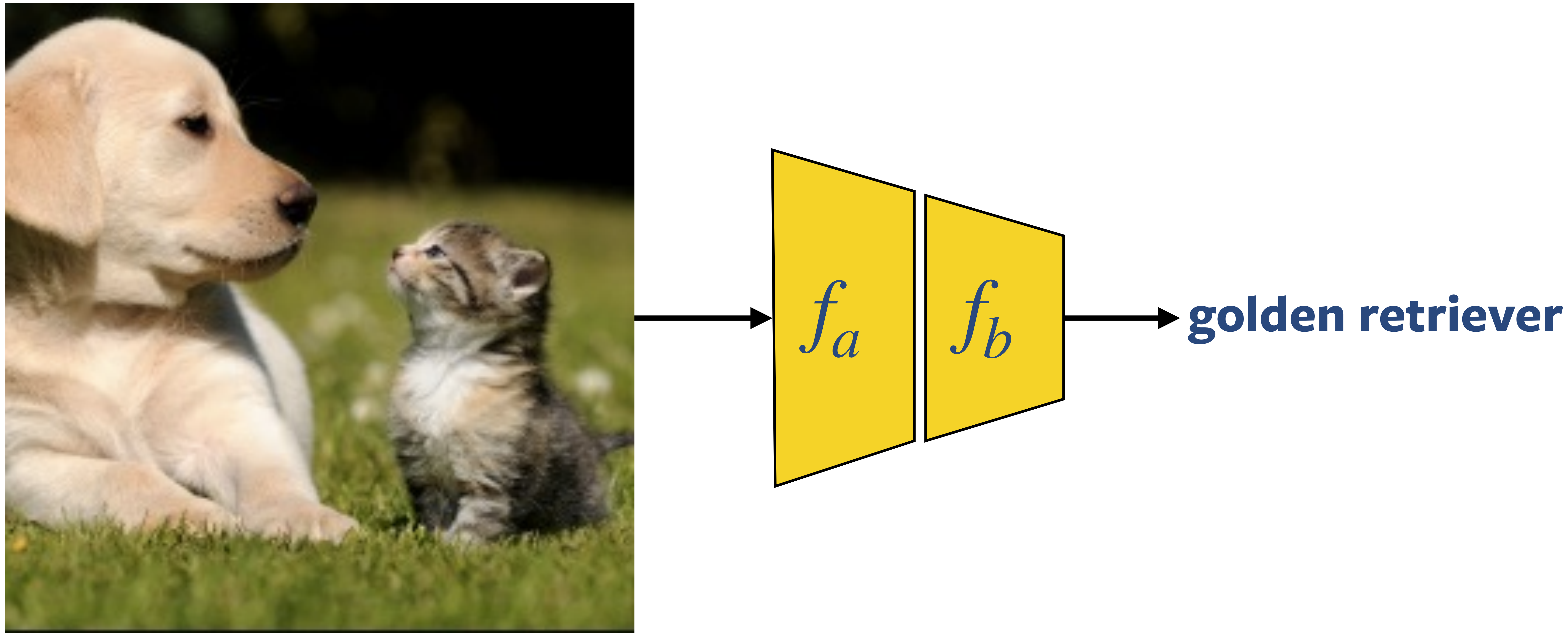
How can we **easily explore** hypotheses about the model?

Acknowledgement: Chris Olah

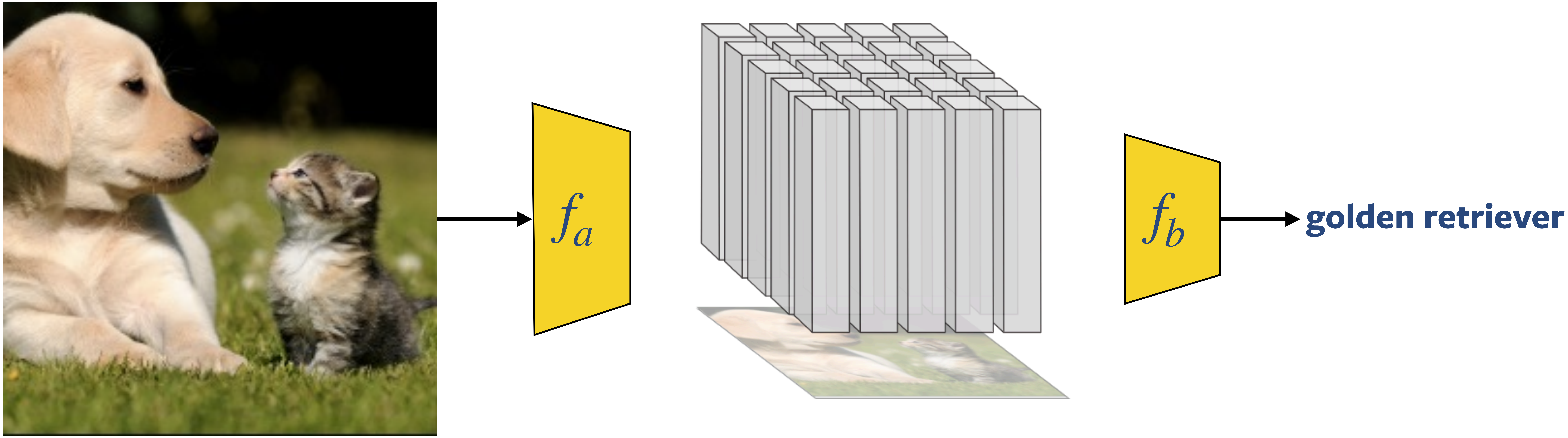
Interactive Similarity Overlays

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

Spatial Activations



Spatial Activations



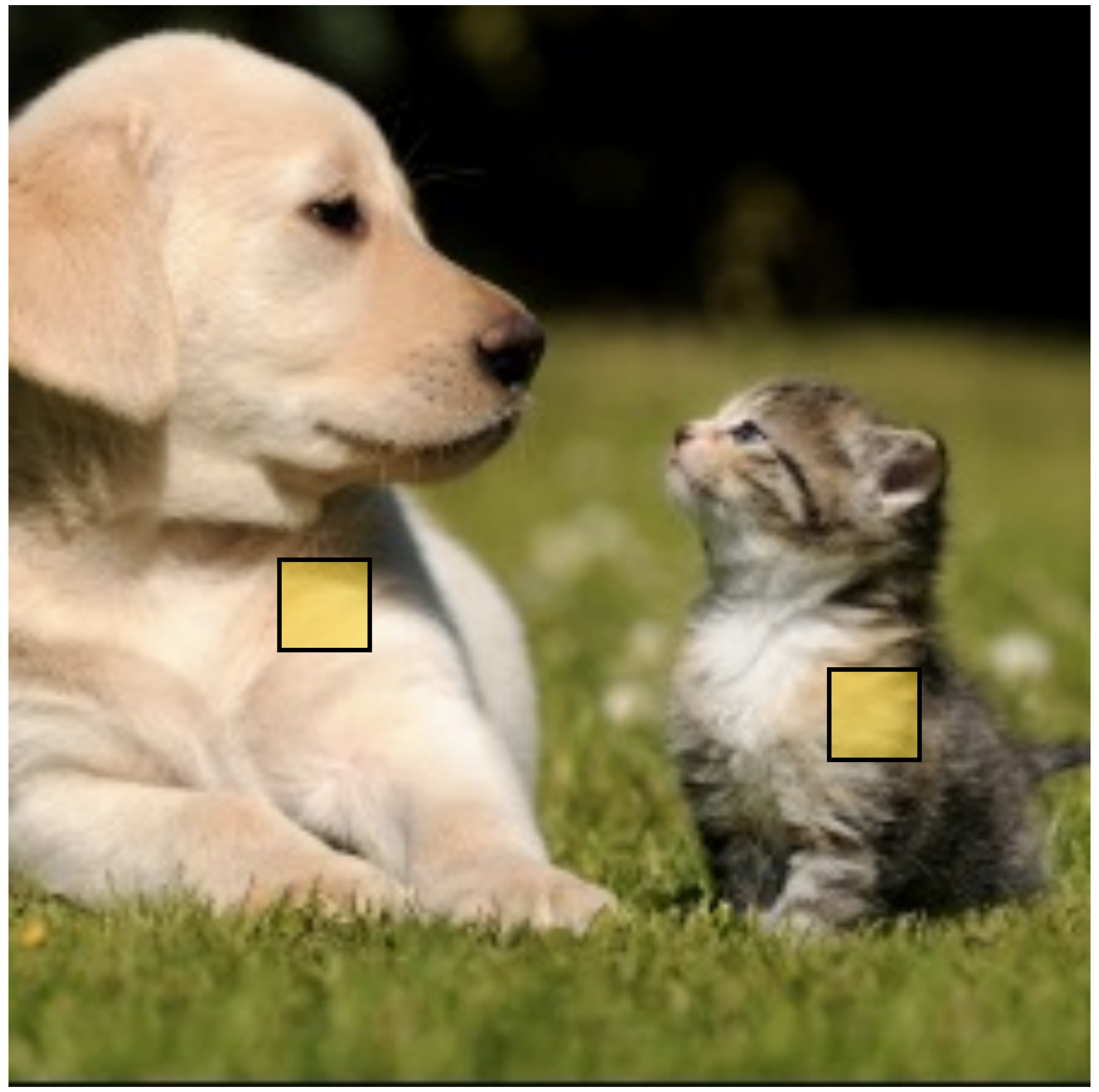
[Olah et al., Distill 2018]

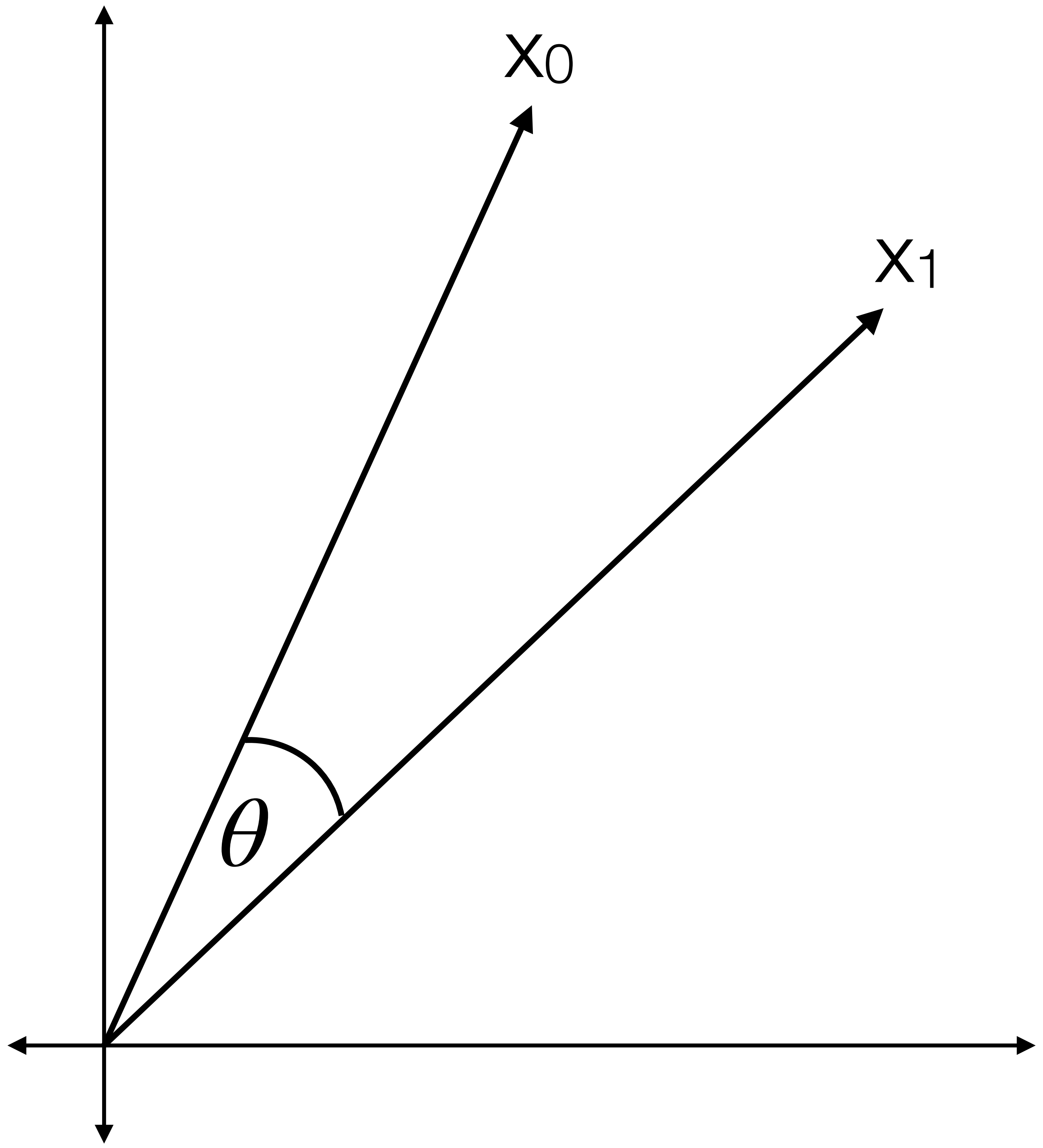
Interactive Similarity Overlays

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

[Olah et al., Distill 2018]

Interactive Similarity Overlays

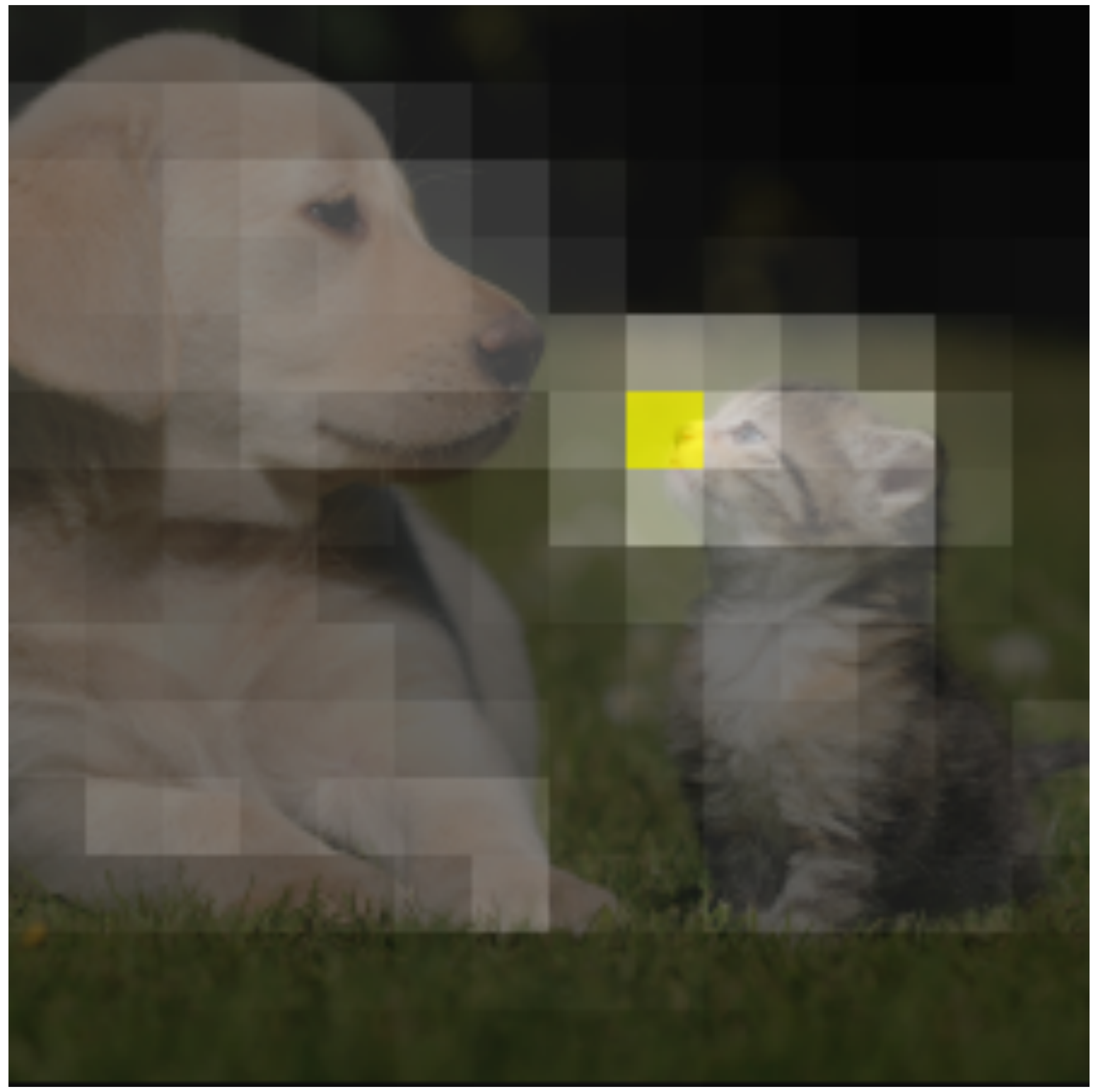




[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]

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.]

Interactive Similarity Overlays

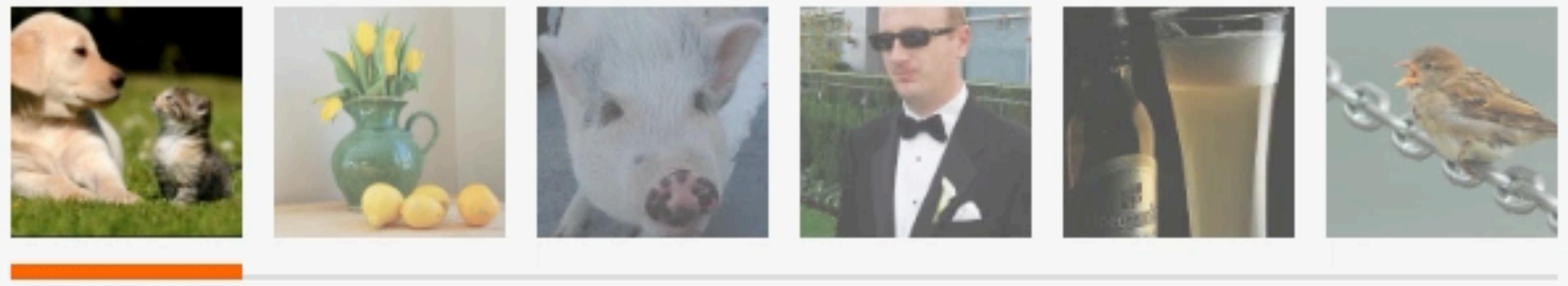
An interactive tool for understanding what neural networks consider similar and different.

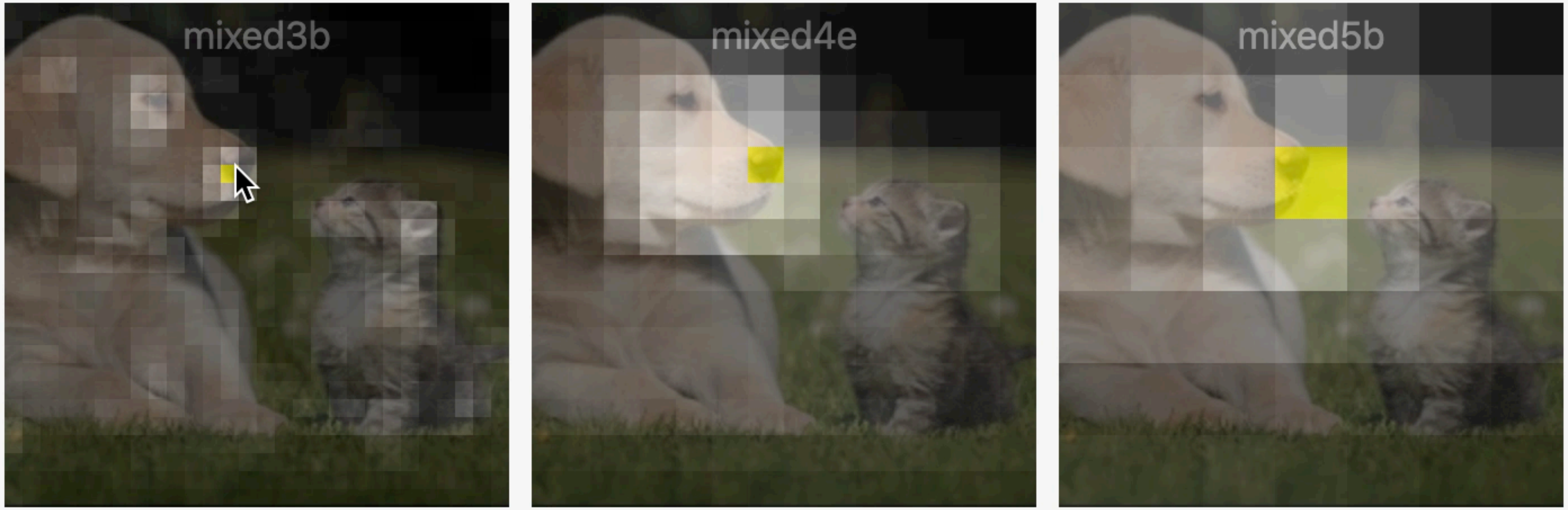
Hover over different parts of the above images. This interactive visualization shows how similar (or different) a neural network considers different image patches to the current image patch (highlighted in yellow). Try hovering over animal features (e.g., noses, eyes, faces) and background regions.

This article is best viewed in Google Chrome.

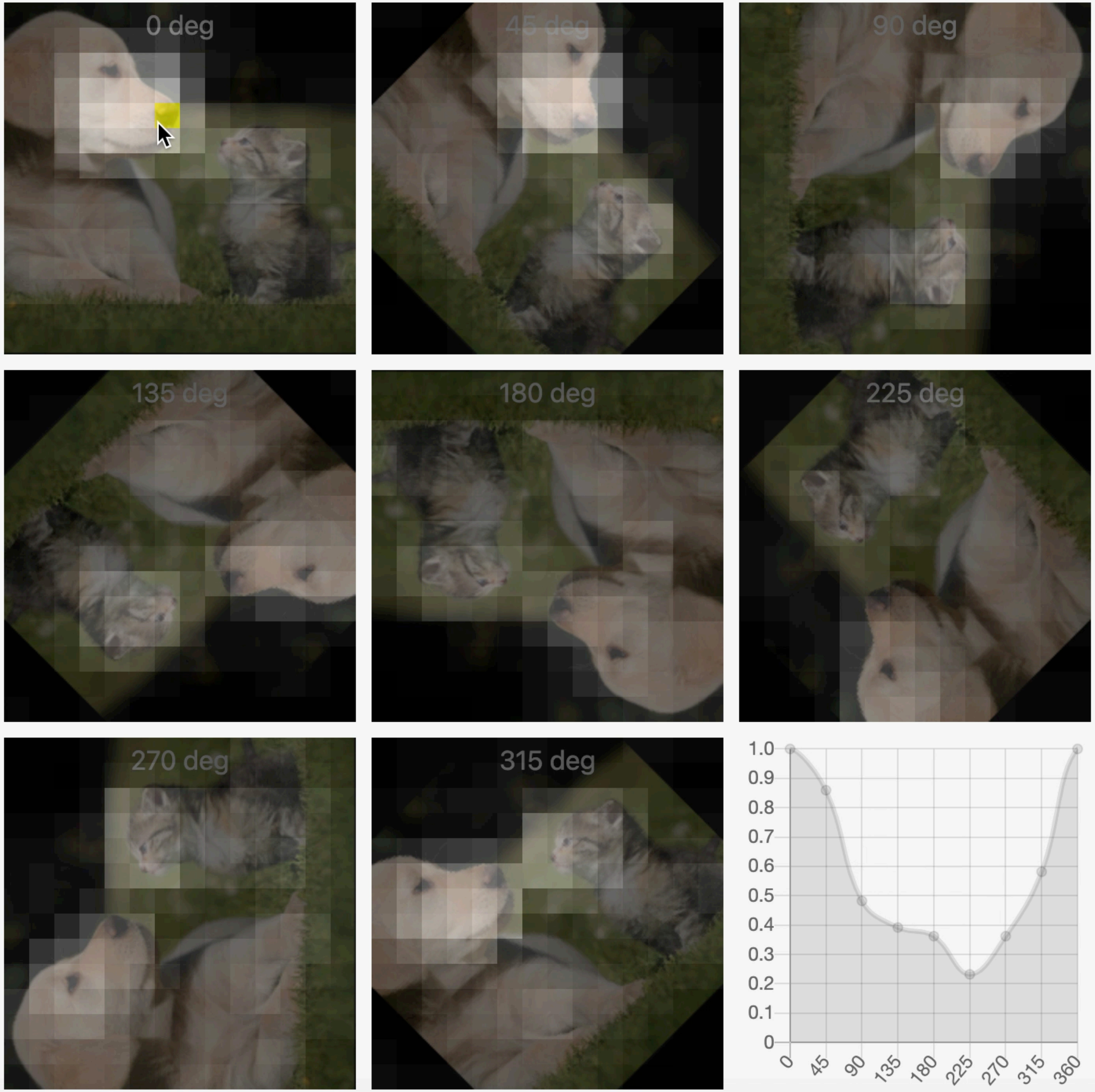
REPRODUCE IN A CO NOTEBOOK

Layers with different spatial resolutions.





The location of the highlighted image patch (in yellow) has been synchronized across images, such that the overlays show similarity scores with respect to each image's highlighted patch (i.e., no similarity scores were computed between images). Consider exploring edges in mixed3b layers and semantic features (e.g., objects and object parts, like noses and eyes) in mixed4e and mixed5b layers.


0 deg
45 deg
90 deg
135 deg
180 deg
225 deg
270 deg
315 deg
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
45
90
135
180
225
270
315
360

Interactive Overlays: Basic Examples (TensorFlow)

File Edit View Insert Runtime Tools Help Cannot save changes

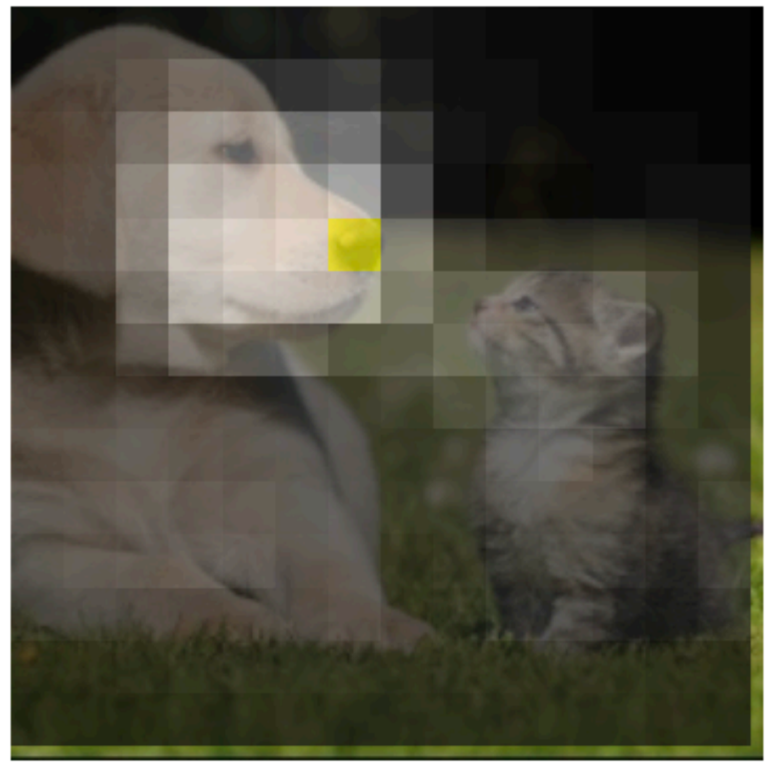
+ Code + Text Copy to Drive

```
# Get images
img_urls = ["https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/dog_cat.jpeg",
            "https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/flowers.jpeg",
            "https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/pig.jpeg",
            "https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/bowtie_guy.jpeg",
            "https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/beer.jpeg",
            "https://raw.githubusercontent.com/ruthcfong/interactive_overlay/master/images/chain.jpeg"]
imgs = [load(url) for url in img_urls]

model = models.InceptionV1()
model.load_graphdef()
```

```
acts = get_acts(model, imgs[0], "mixed4d")
grid = np.hstack(np.hstack(cossim_grid(acts, acts)))
colored_grid = add_color_index(grid, acts.shape[0])
```

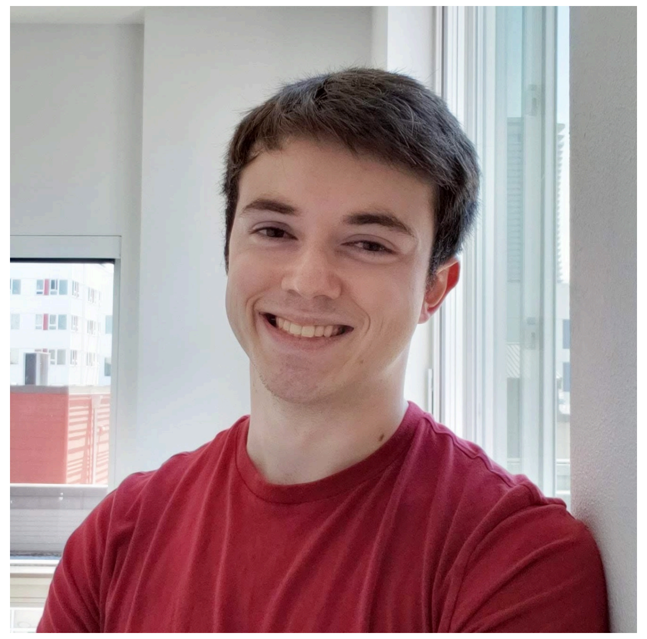
```
lucid_svelte.CossimOverlay({
    "image_url": _image_url(imgs[0]),
    "masks_url": _image_url(colored_grid),
    "size": 224,
    "N": acts.shape[0],
})
```



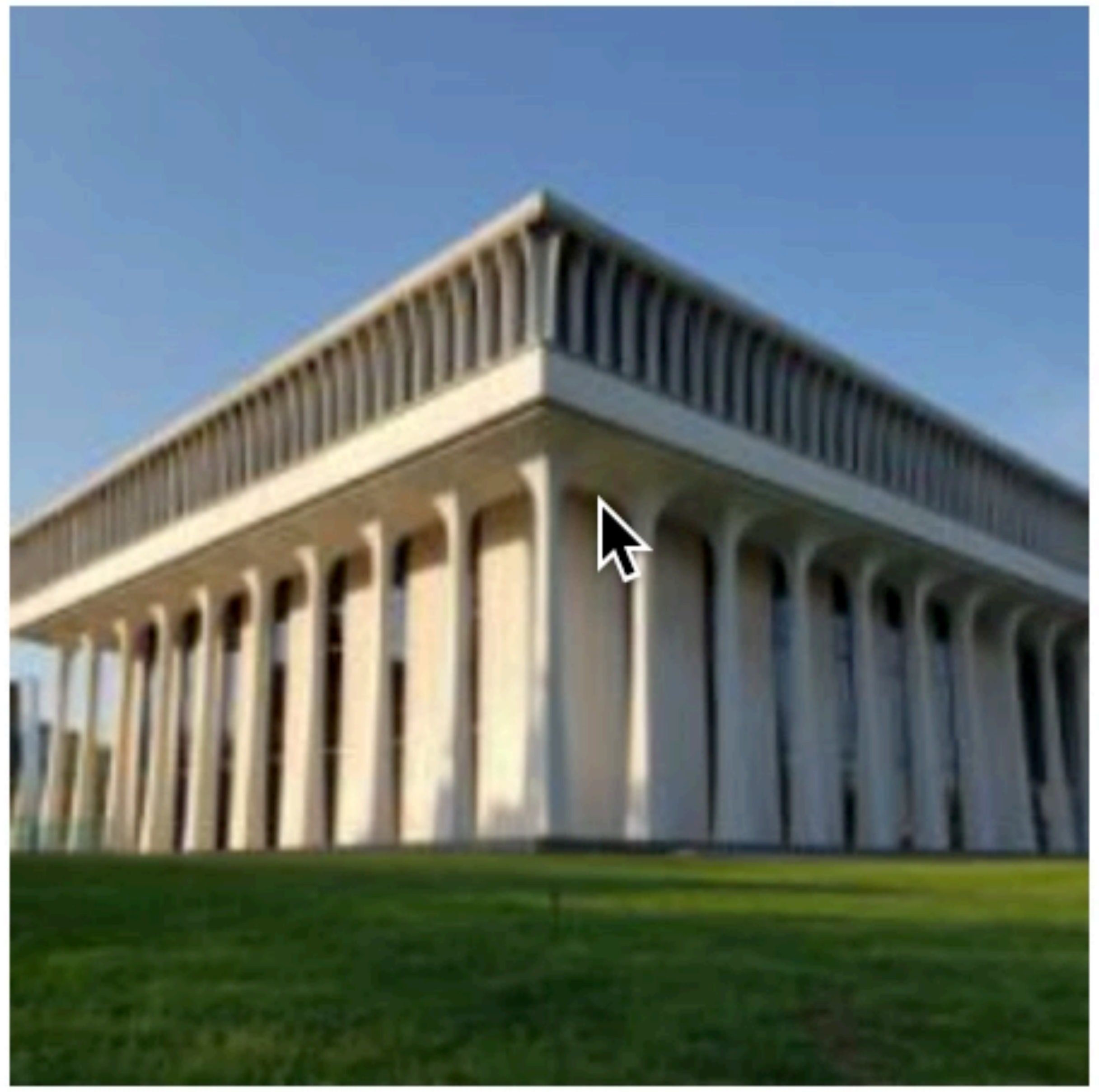
6,4

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.

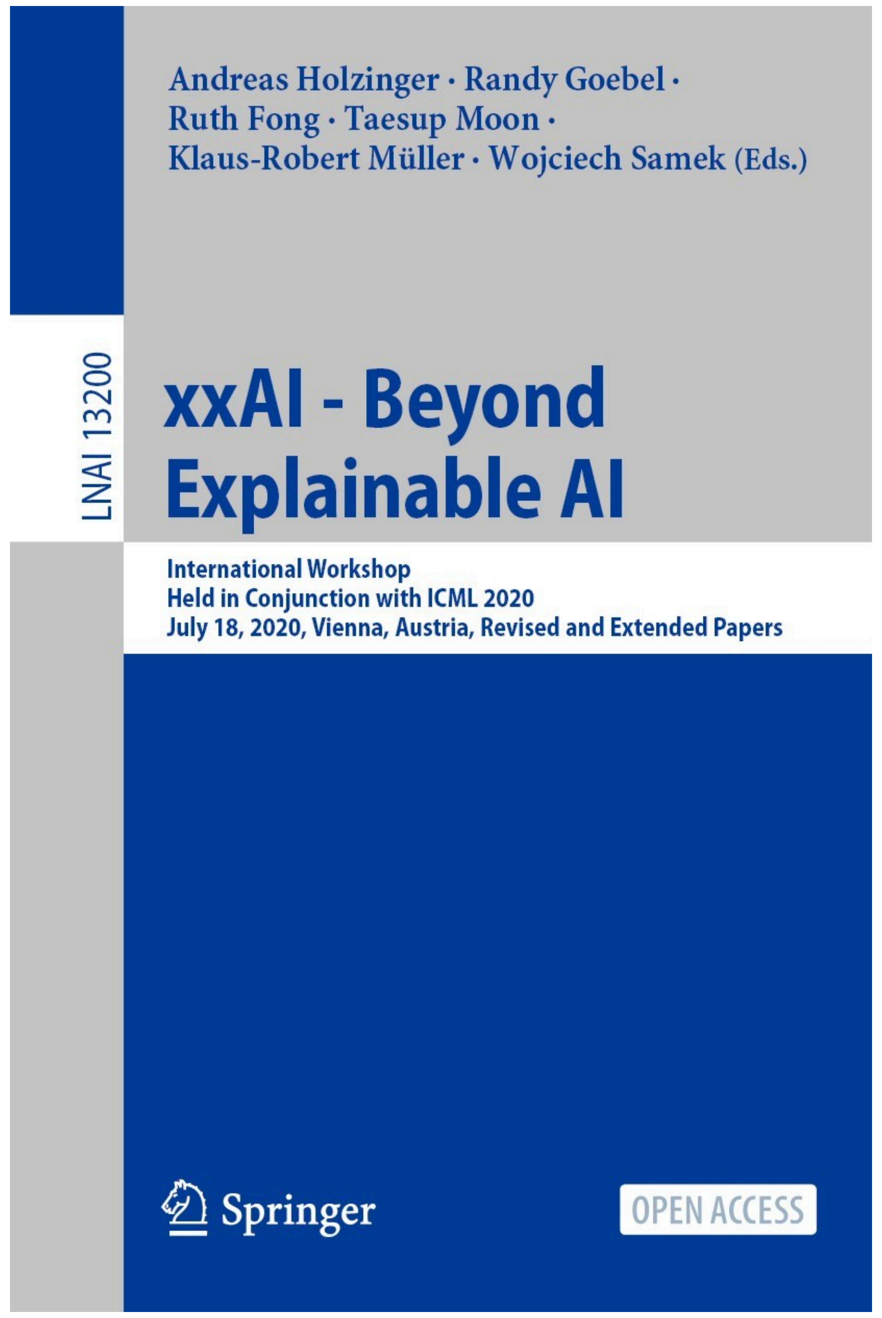
Takeaways from challenges in interpretability

- **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

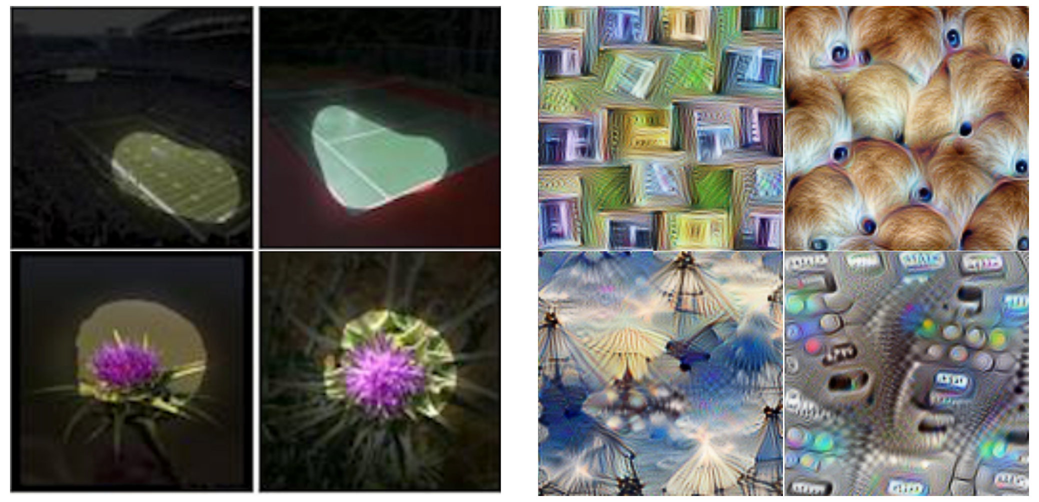
1. Develop interpretability methods for **diverse domains**
   - Beyond CNN classifiers: self-supervised learning, generative models, etc.

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

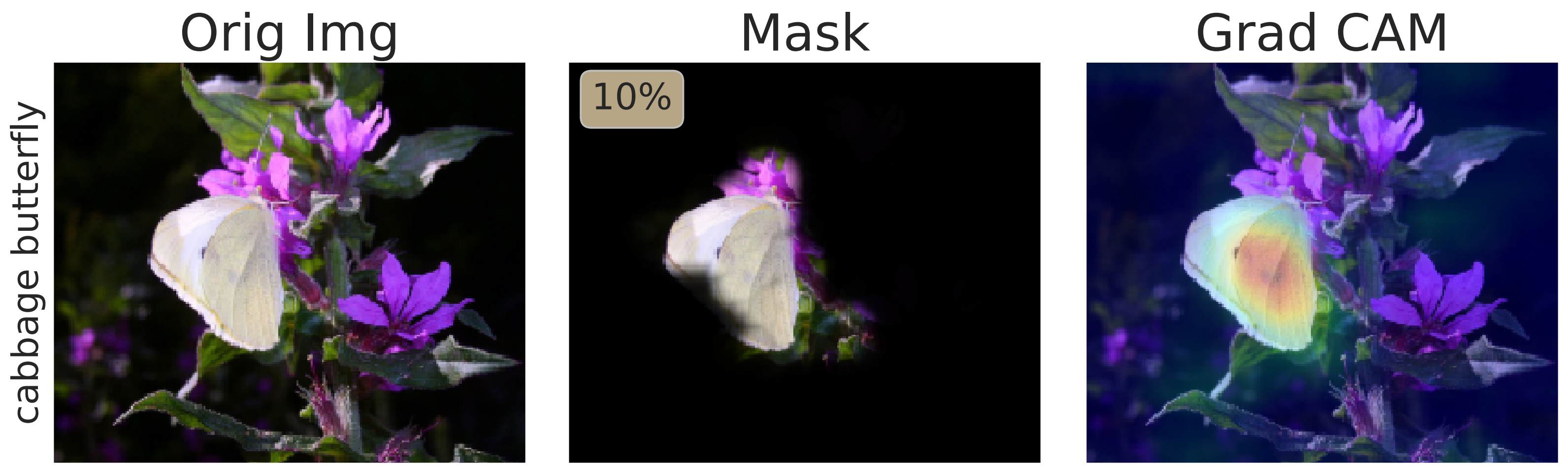


Primarily focused on understanding and approximating **CNNs**

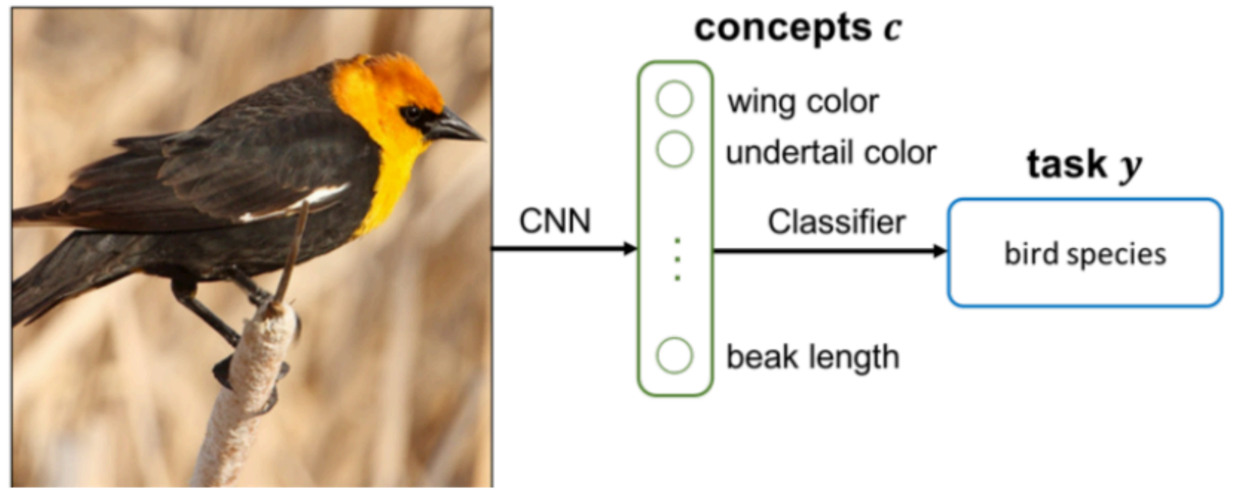
Feature visualization (2013-2018)
Activation Max., Feature Inversion,
Net Dissect, Feature Vis.

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

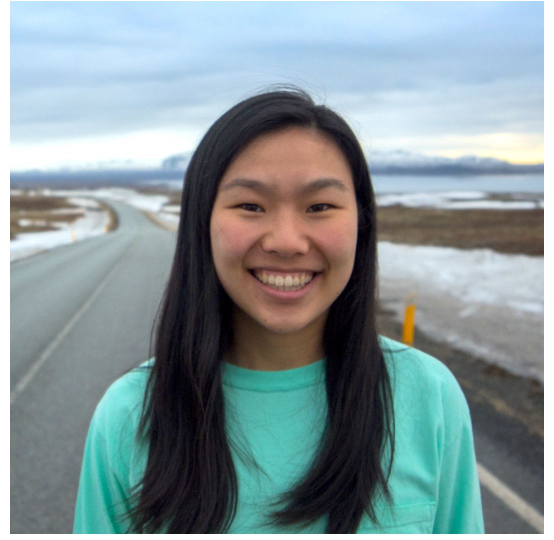

Devon Ulrich


Dora Zhao


Nicole Meister


Sunnie S. Y. Kim


Vikram V. Ramaswamy


Angelina Wang


Ryan A. Manzuk


Iro Laina


Andrea Vedaldi


Elizabeth Anne Watkins


Andrés Monroy-Hernández


Chris Olah


Alex Mordvintsev


Adam C. Maloof


Olga Russakovsky

UNIVERSITY OF OXFORD

RHODES TRUST

NSF

intel labs

Open Philanthropy Project

We're hiring postdocs!
bit.ly/vai-lg-postdoc

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

Thank You