Directions in Interpretability

Ruth Fong

MICCAI 2022, Workshop on Interpretability of Machine Learning in Medical Image Computing (iMIMIC)

September 22, 2022

Slides and links available at ruthfong.com

PRINCETON UNIVERSITY

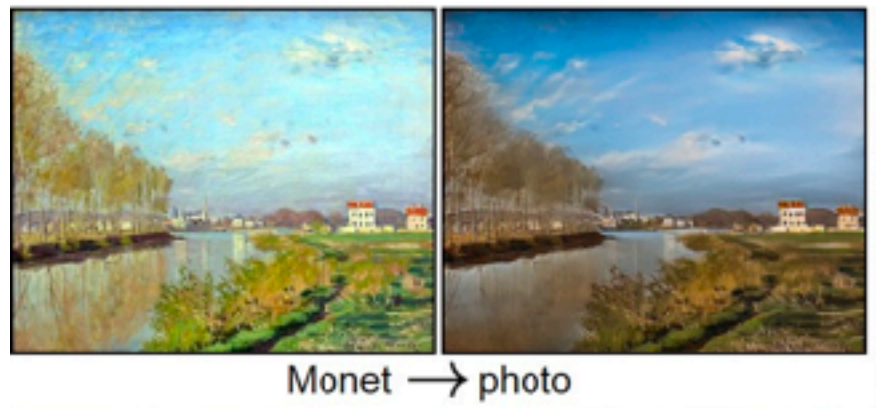
What is interpretability?

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

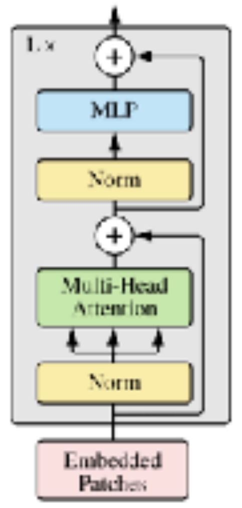
Why interpretability?

- 🔬 Science
- 🤝 Trust
- 🤖 Learning

An incomplete retrospective: the first decade of deep learning



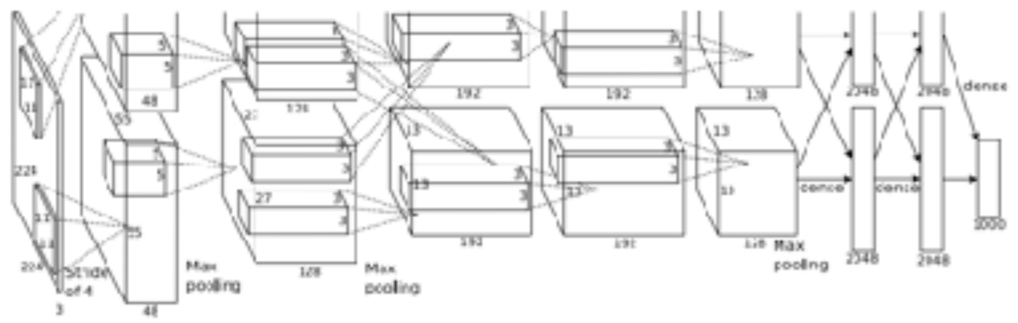
GANs (2014-2018)
GAN, ProGAN, CycleGAN



Transformers (2017-now)
Transformer, BERT, ViT

IMAGENET

2012 ⟷ 2022



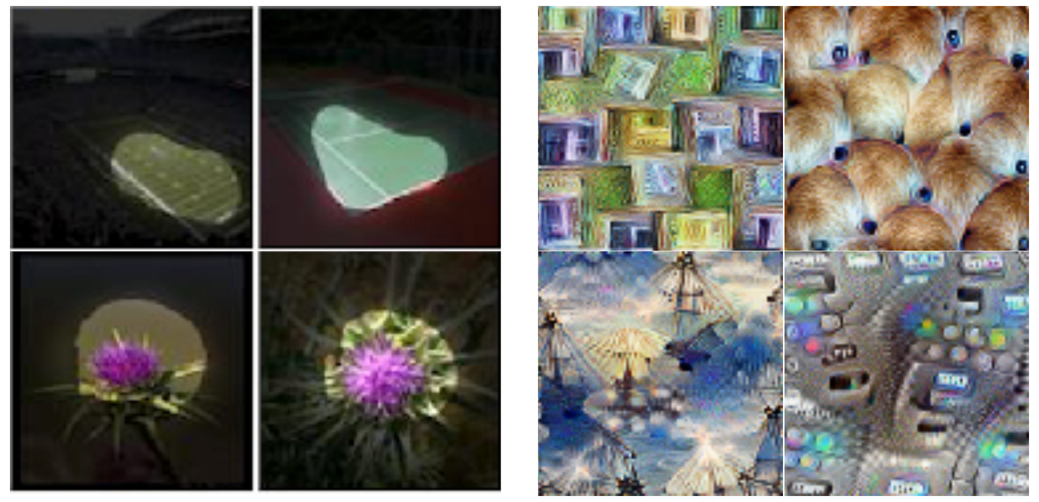
CNNs (2012-2016)
AlexNet, VGG16, GoogLeNet, ResNet50

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



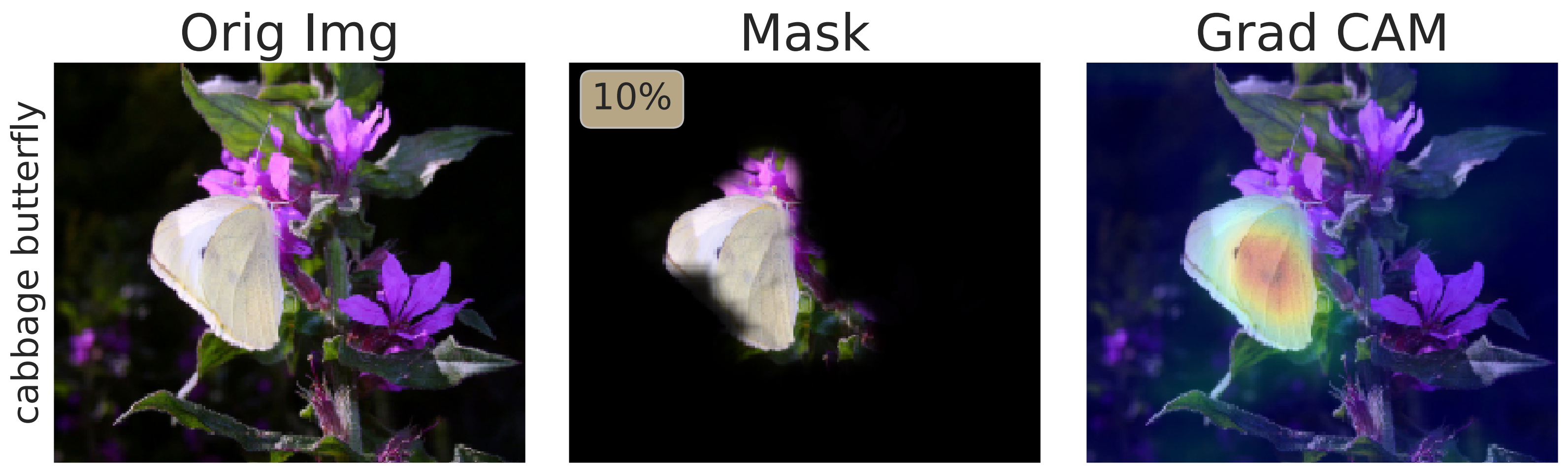
Feature visualization (2013-2018)

Activation Max., Feature Inversion,

Net Dissect, Feature Vis.

2012

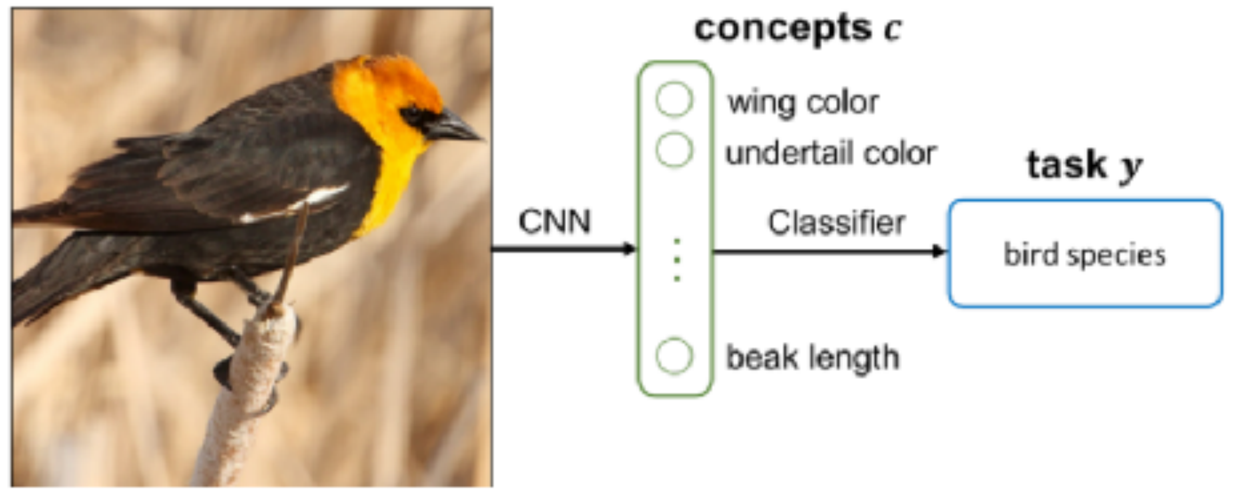
2022



Attribution heatmaps (2013-2019)

Gradient, Grad-CAM,

Occlusion, Perturbations, RISE



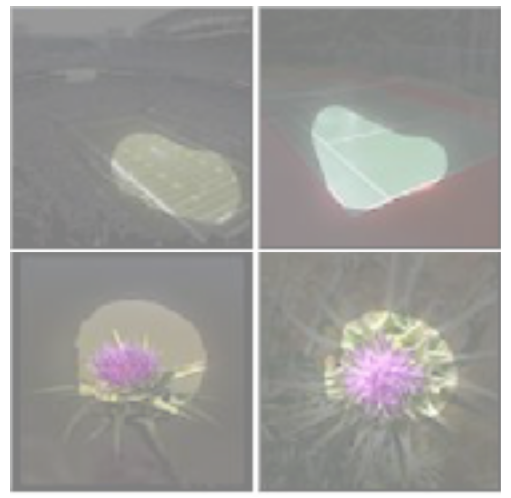
Interpretable-by-design (2020-now)

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

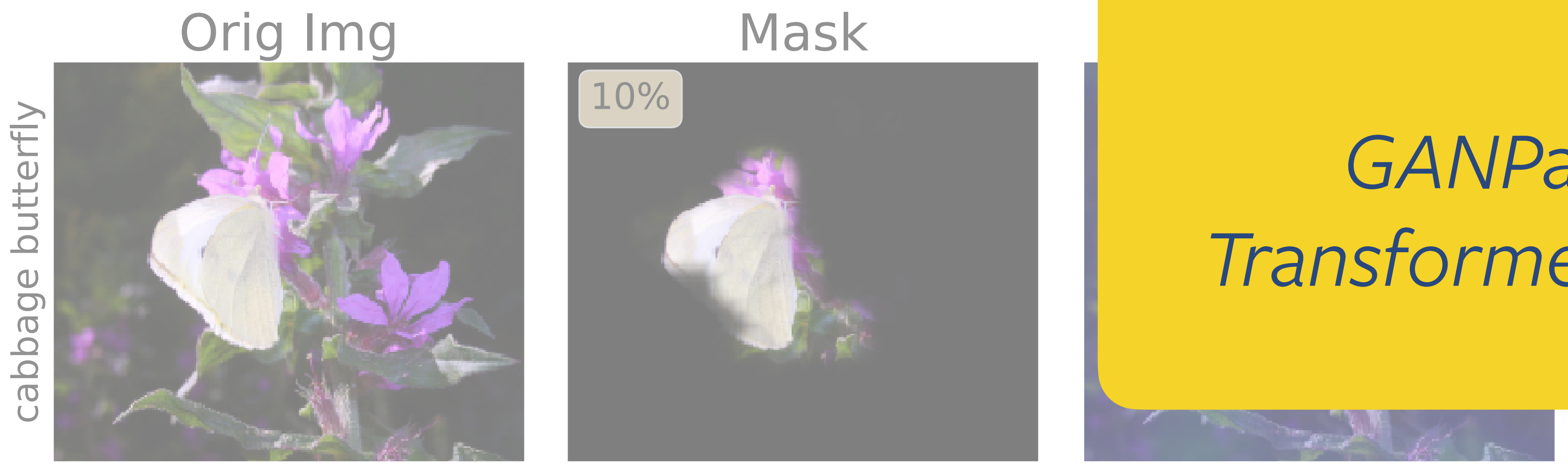


Primarily focused on understanding and approximating **CNNs**

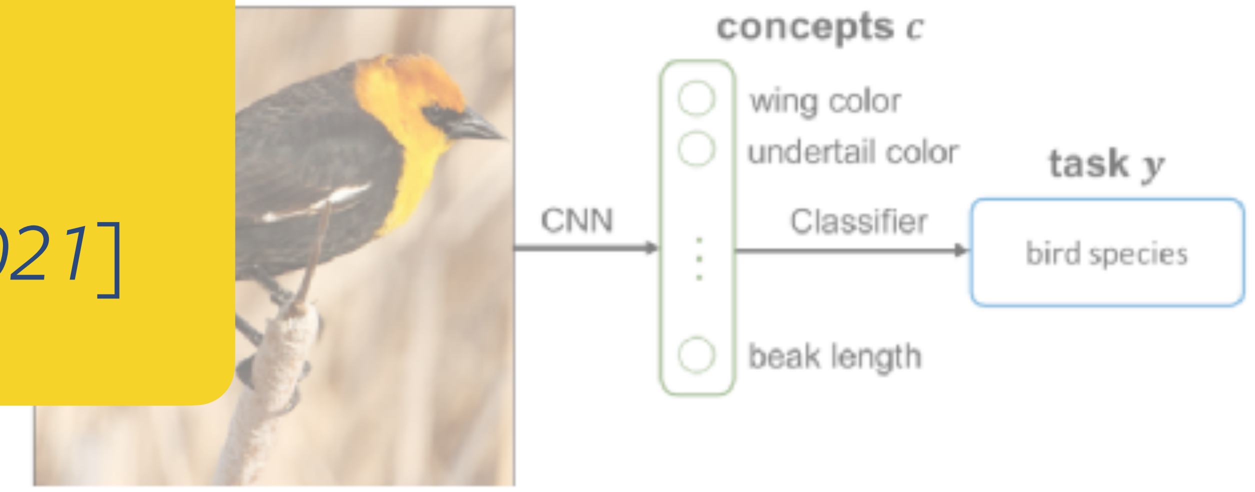
Exceptions:
GANPaint [*Bau et al., ICLR 2019*]
Transformer Circuits [*Elhage et al., 2021*]

2012

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]

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.
2. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

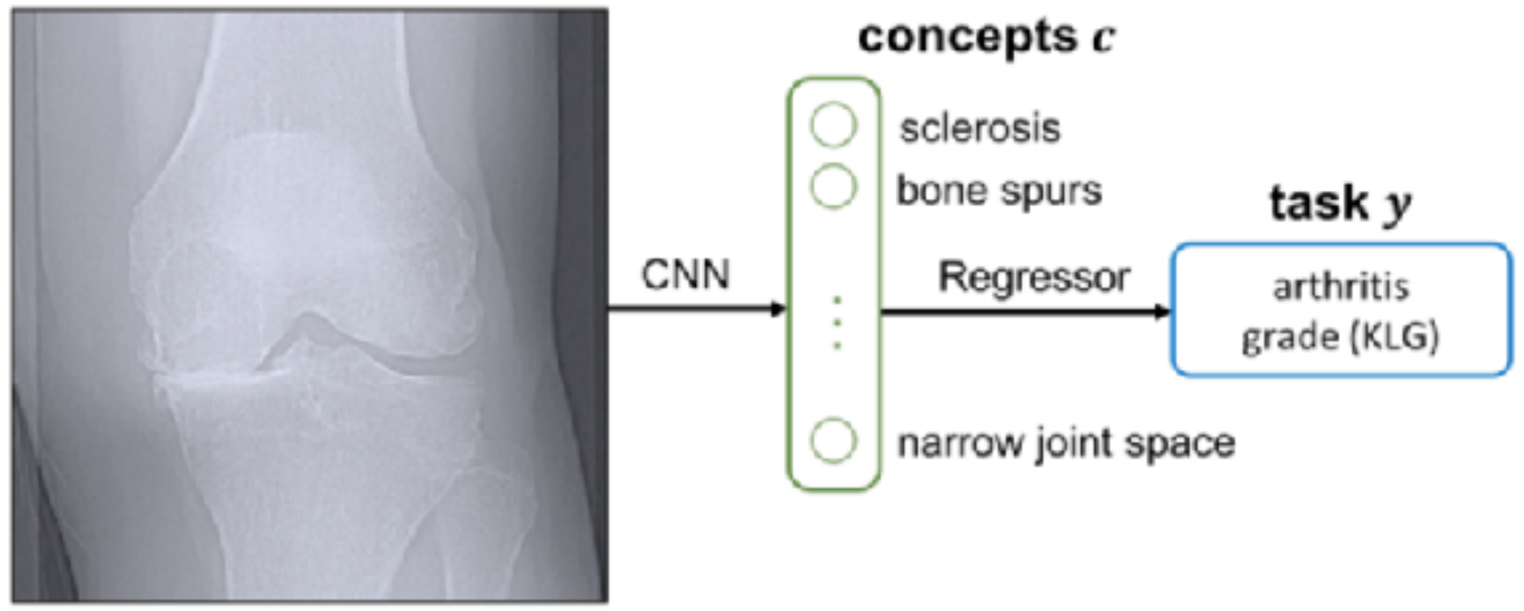
Roadmap

Sunnie S. Y. Kim

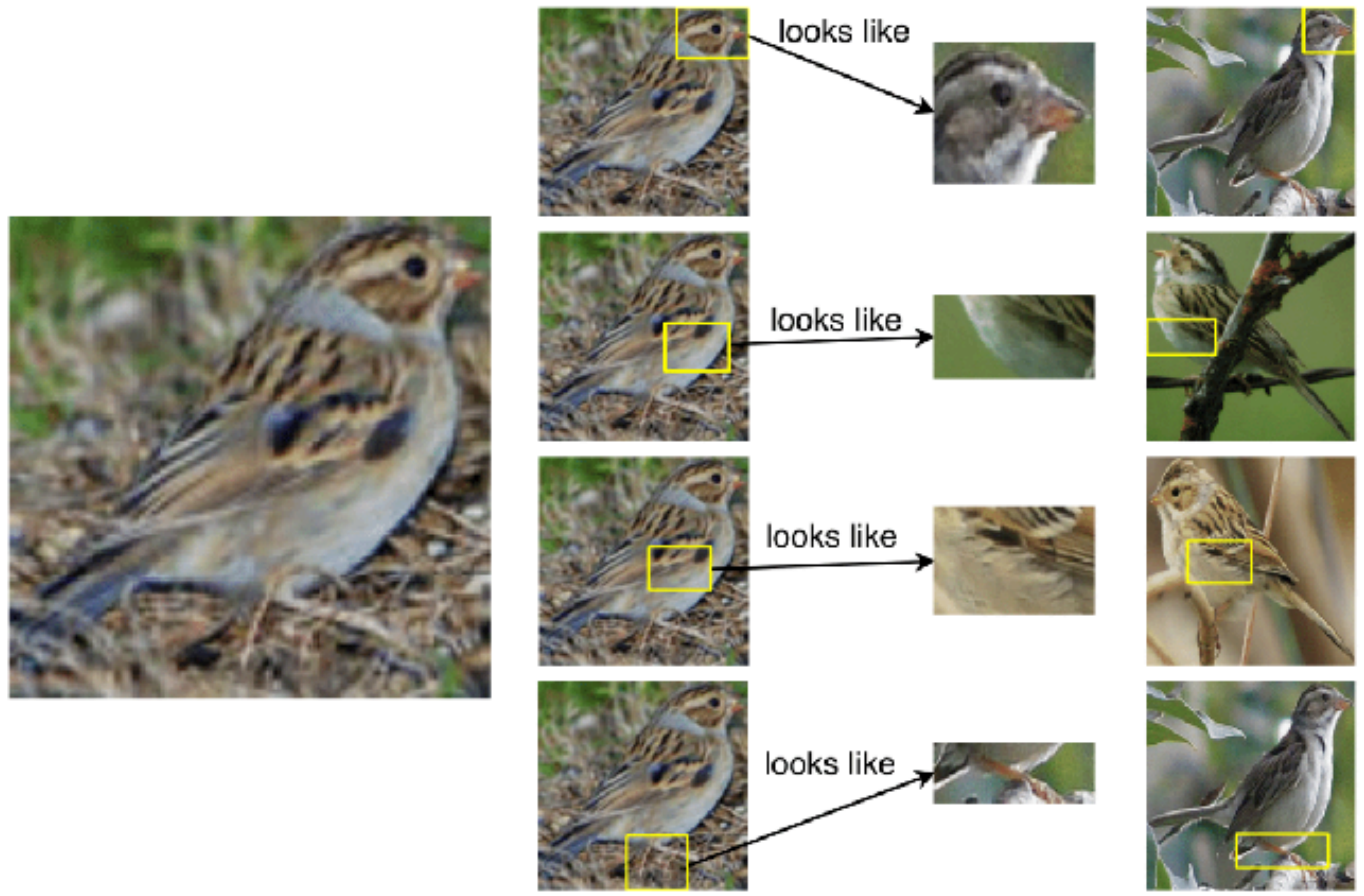
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. **Static** visualizations → **interactive** visualizations
   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)



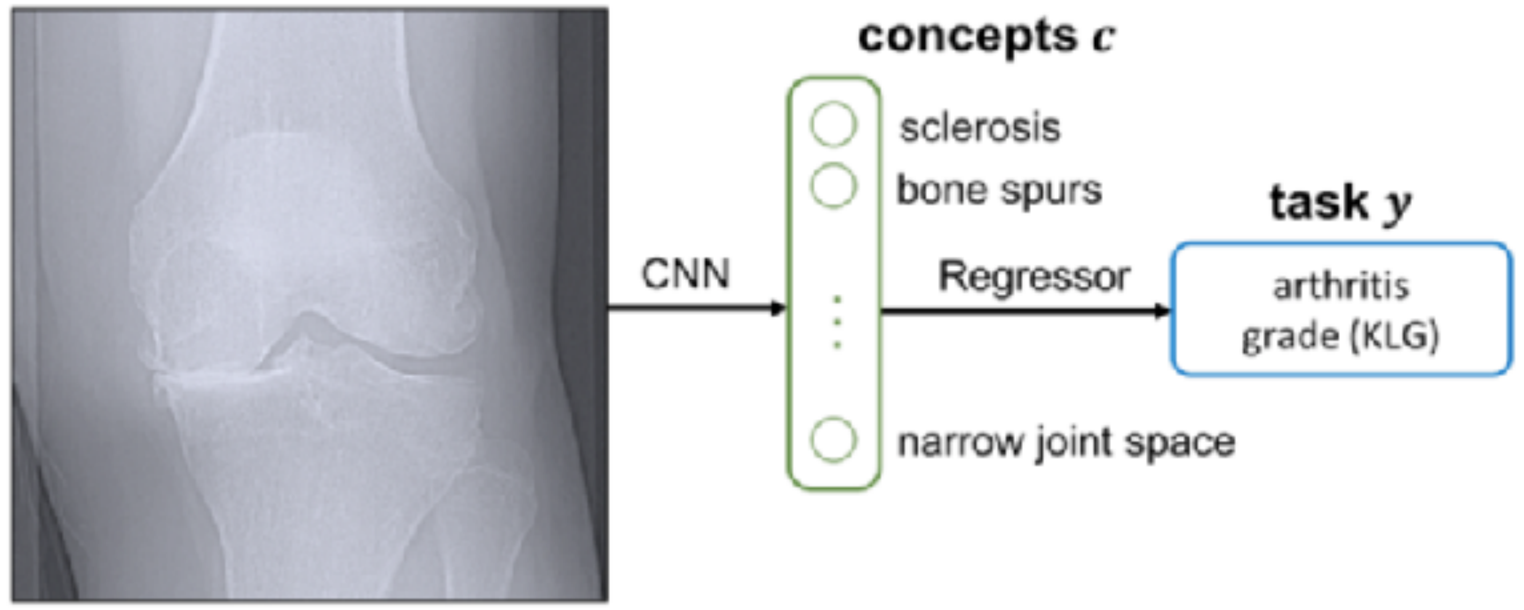
Prototype explanations
(e.g. ProtoPNet)

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

Counterfactual explanations
(e.g. SCOUT)

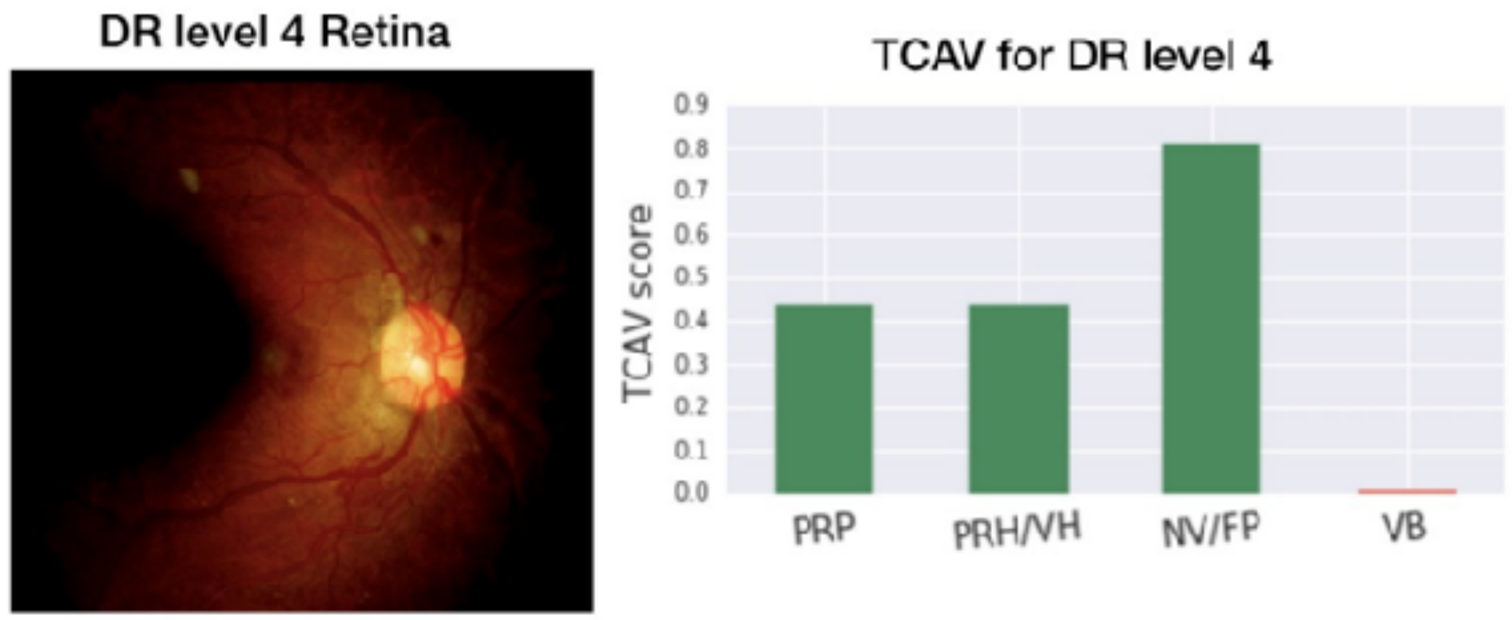
[Selvaraju et al., ICCV 2017; Koh*, Nguyen*, Tang* et al., ICML 2020; Chen* & Li* et al., NeurIPS 2019; Wang & Vasconcelos, CVPR 2020]

Explanation form factors: Why did the model predict Y?



Concept Bottleneck

Knee x-rays → knee osteoarthritis



TCAV

Retinal fundus imaging → diabetic retinopathy

Non-heatmap form factors (e.g. concept-based explanations) are more suitable for fine-grain tasks in medical imaging

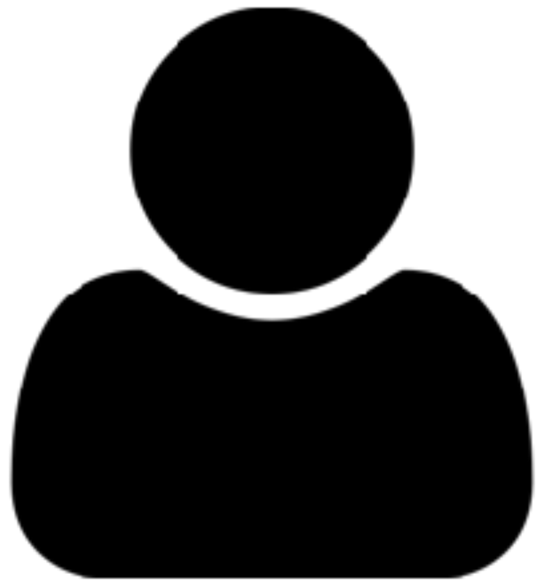
[Koh*, Nguyen*, Tang* et al., ICML 2020. Concept Bottleneck; Kim et al., ICML 2018. TCAV.]

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.

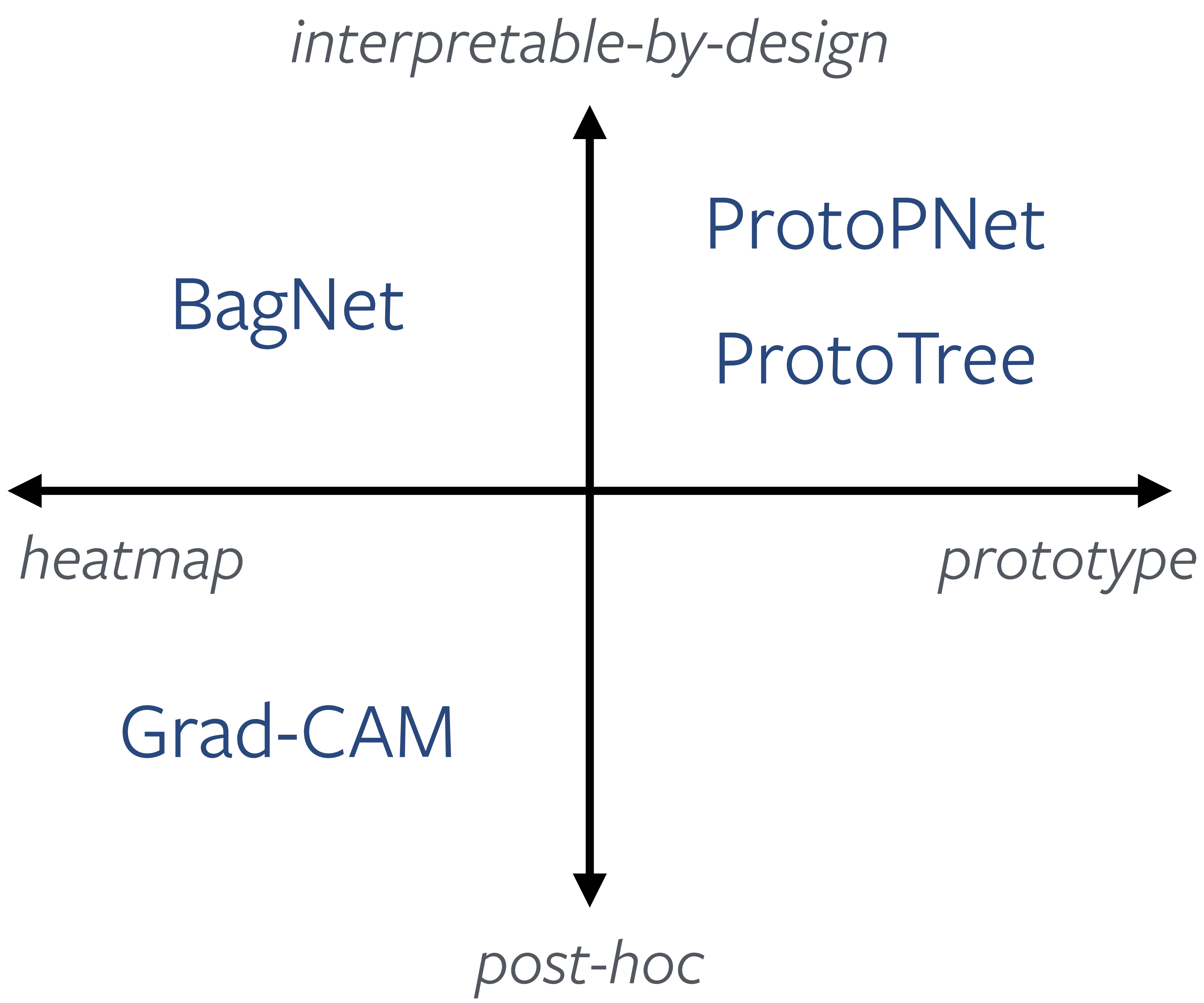
Our contributions

- 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

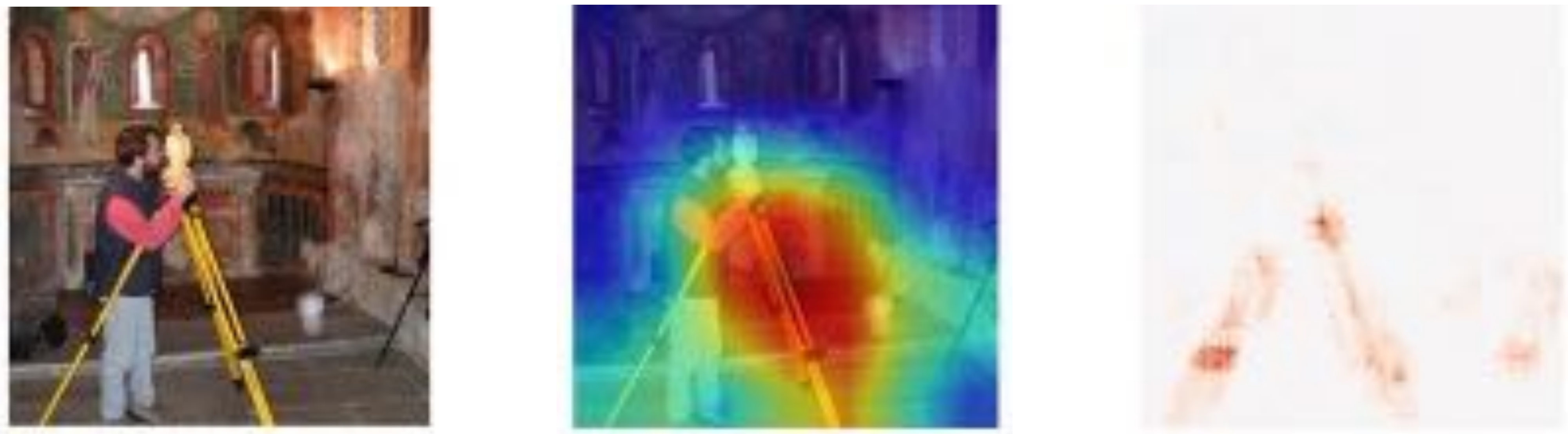
[Sunnie S. Y. Kim et al., ECCV 2022. HIVE.]

Follow up: Ramaswamy et al., arXiv 2022.
Overlooked factors in concept-based explanations:
Dataset choice, concept salience, and human capability.

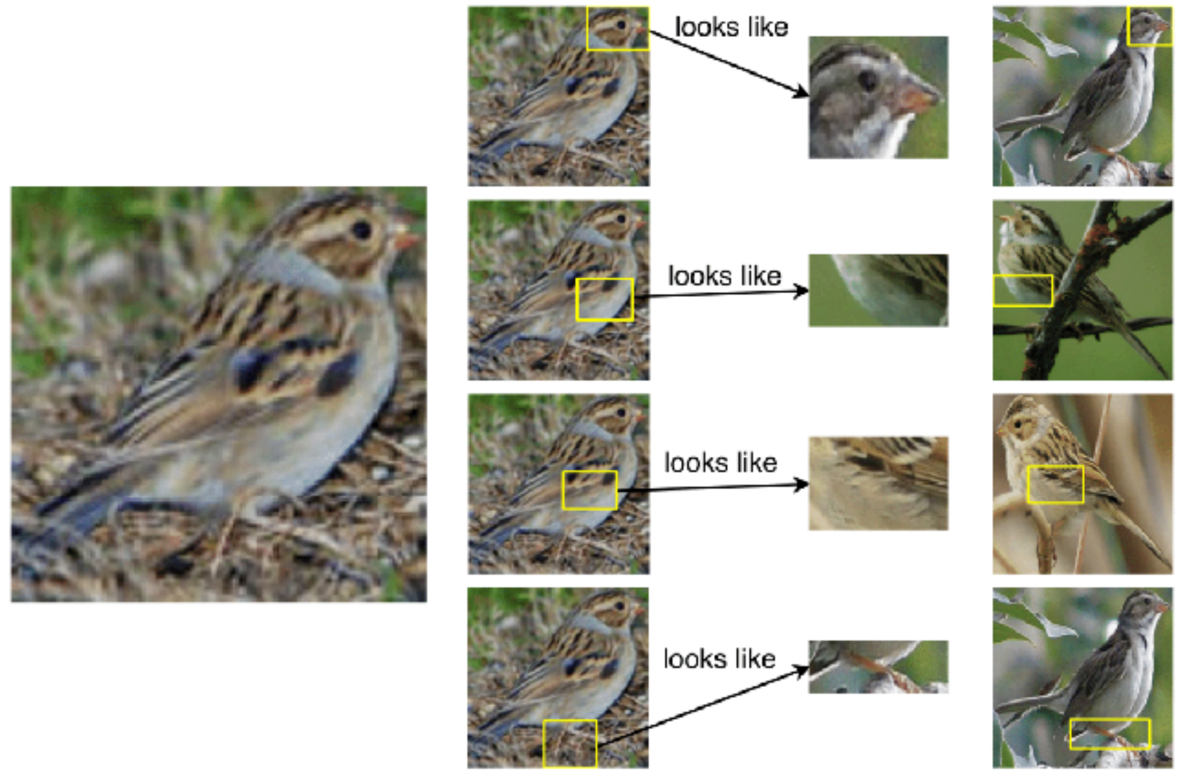
1. Cross-method comparison



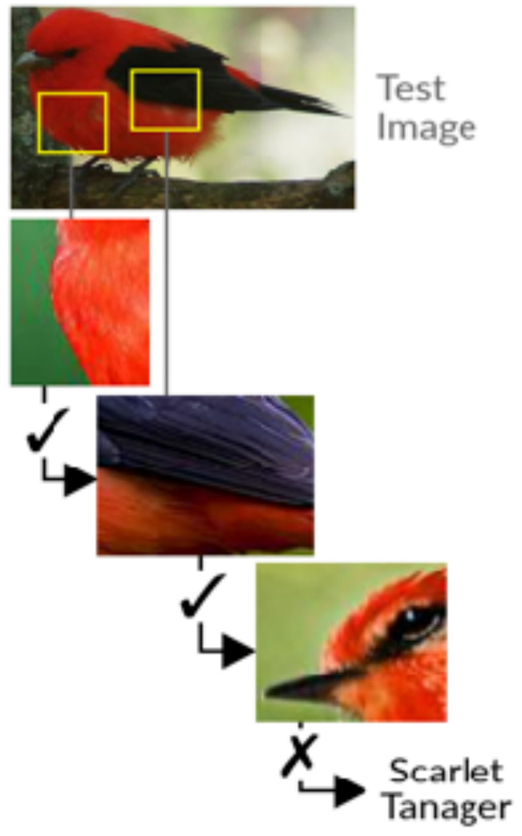
Grad-CAM BagNet



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

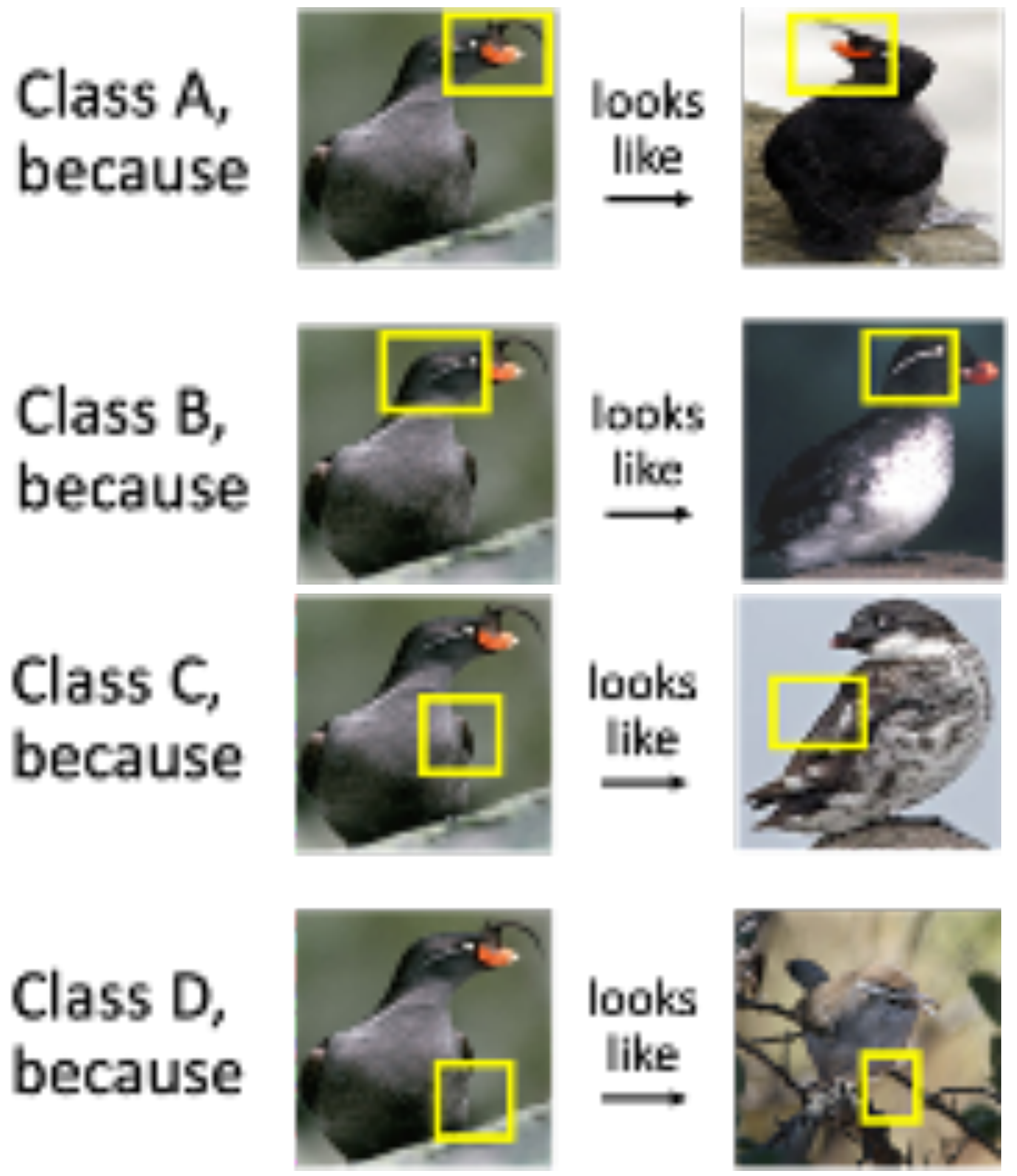
Agreement task

How confident are you in the model's prediction?

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]

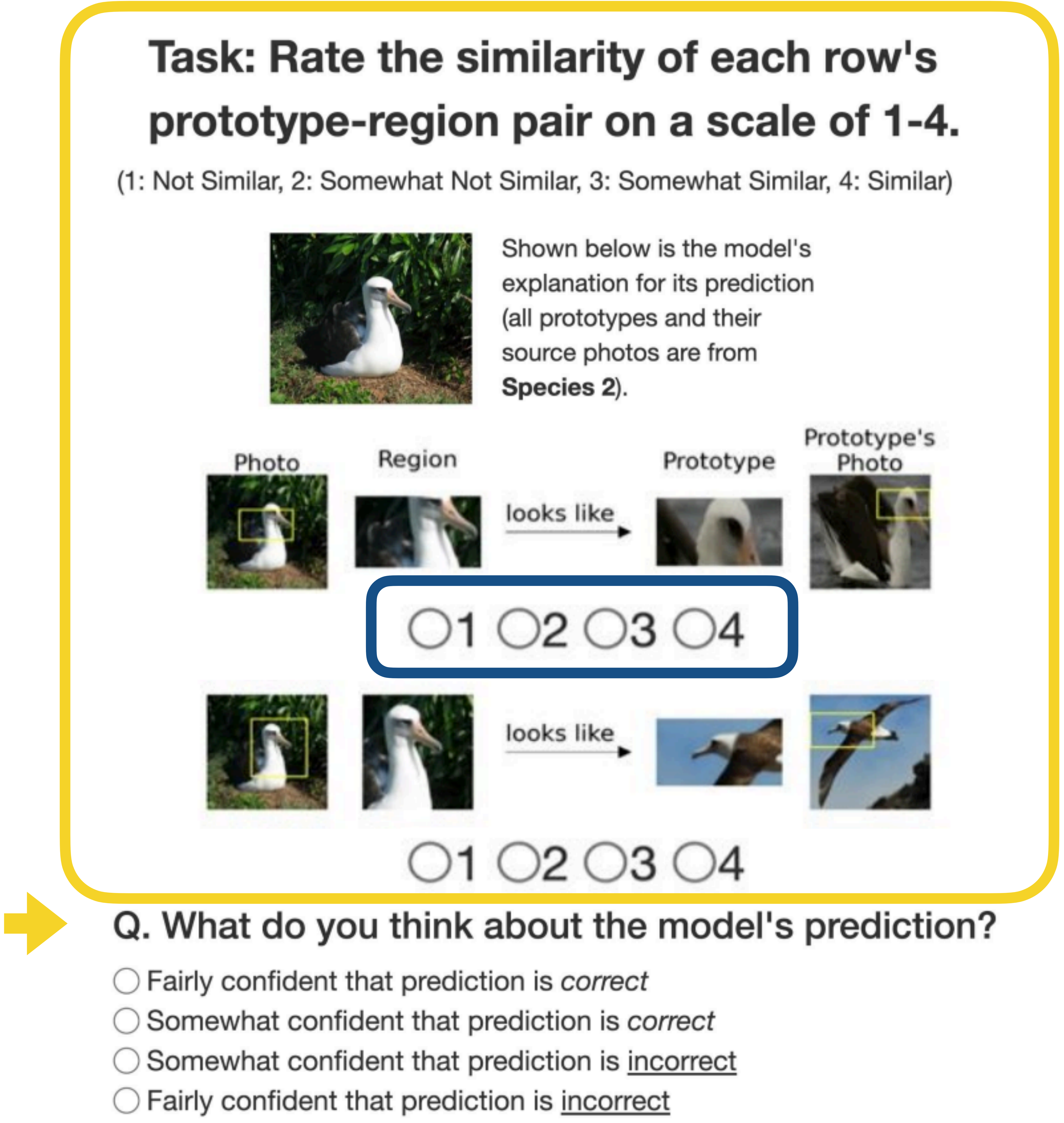
ProtoPNet and ProtoTree only

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.



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

2. Human-centered evaluation

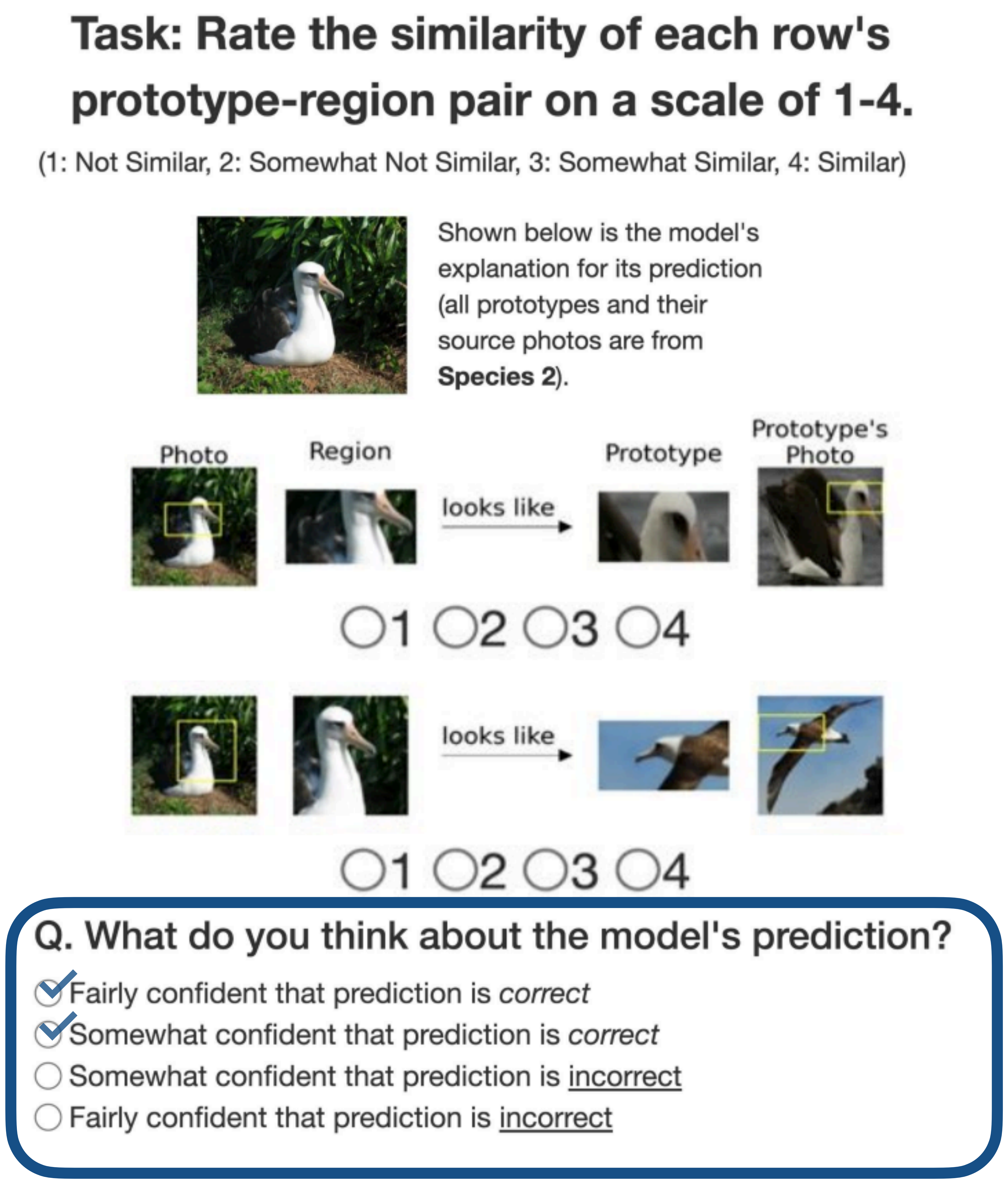
Agreement task

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

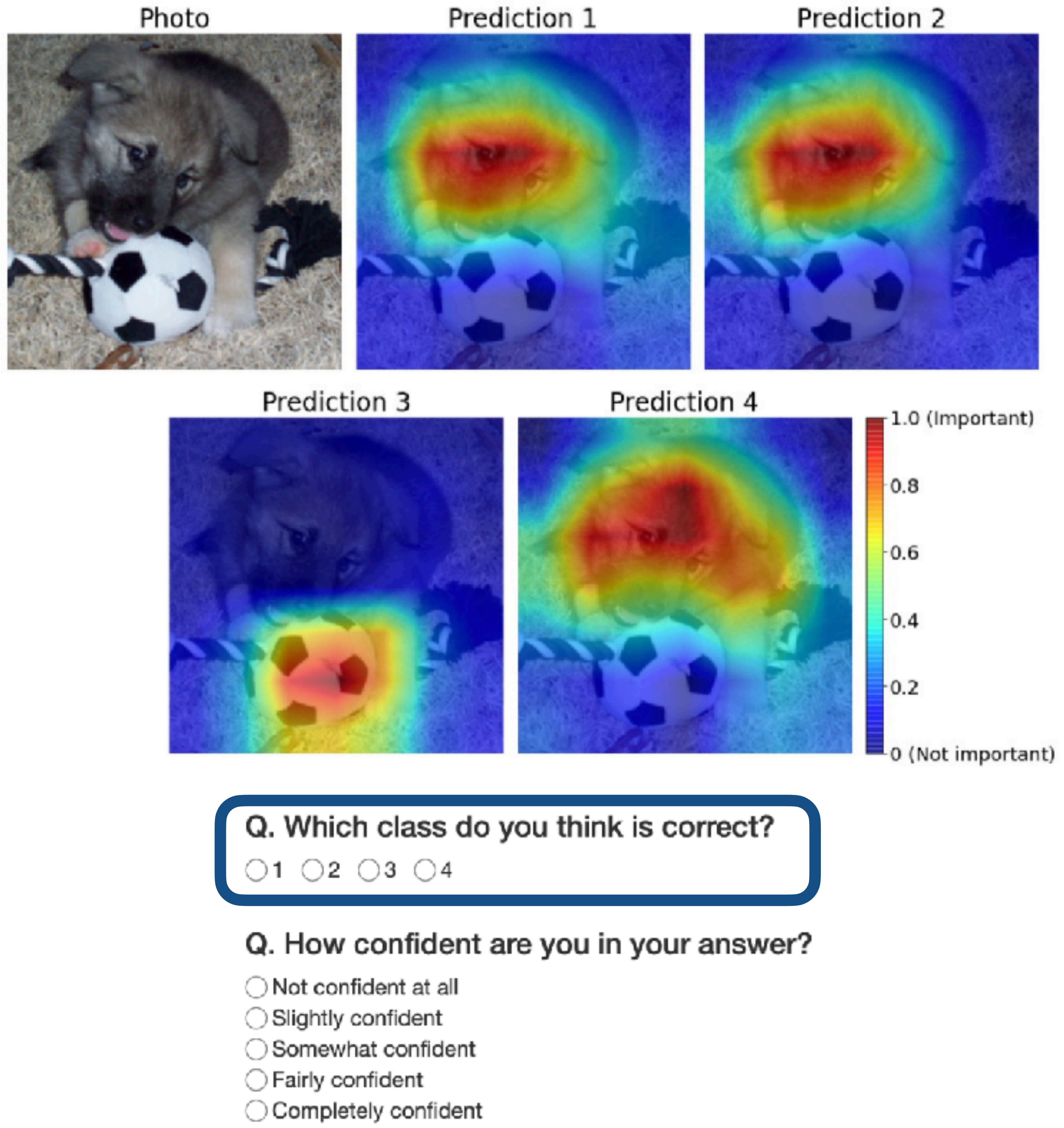
Distinction task

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.

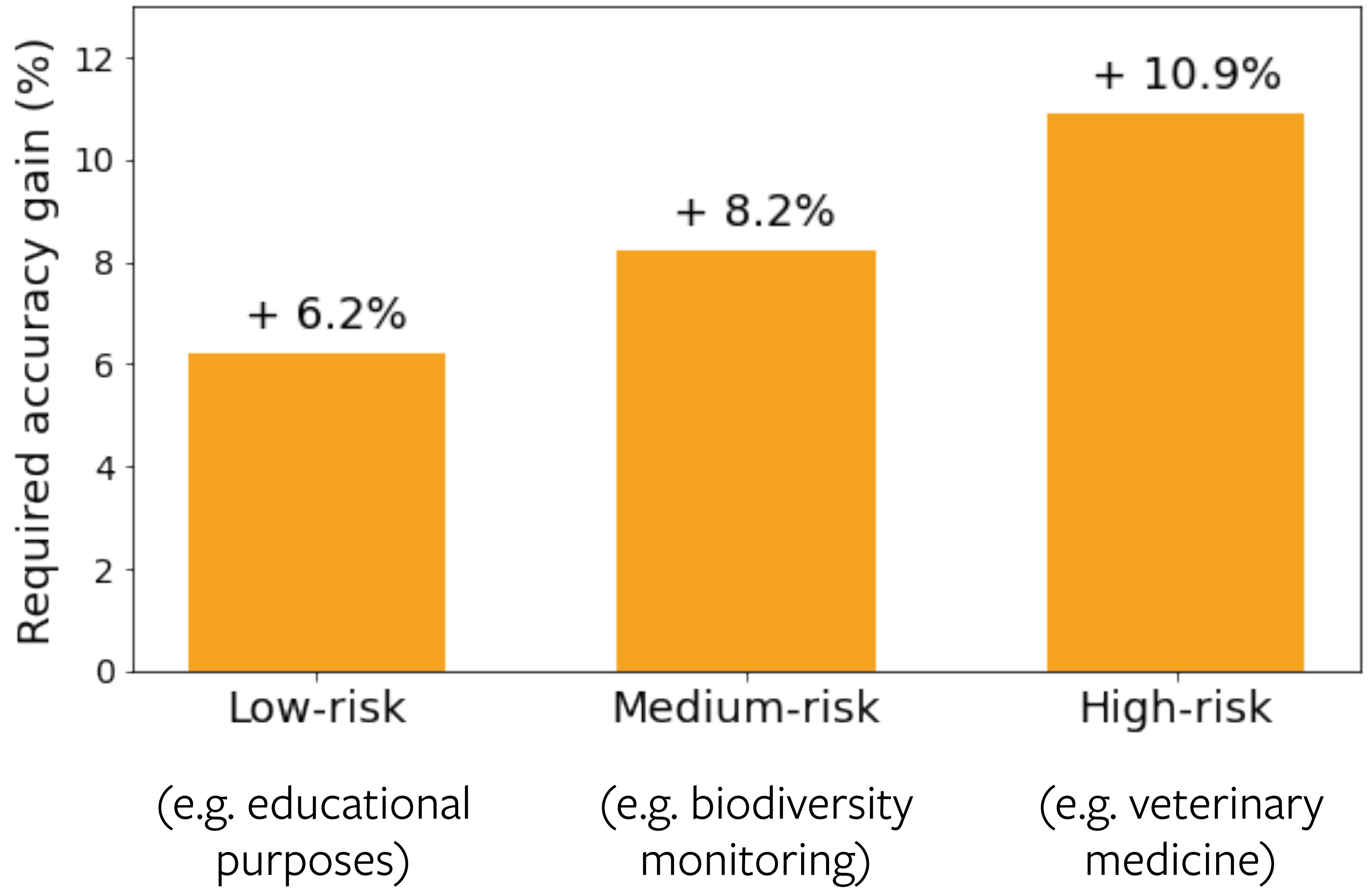
Finding #2: Agreement task reveals **confirmation bias**.

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.

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?



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

Challenges for human evaluation

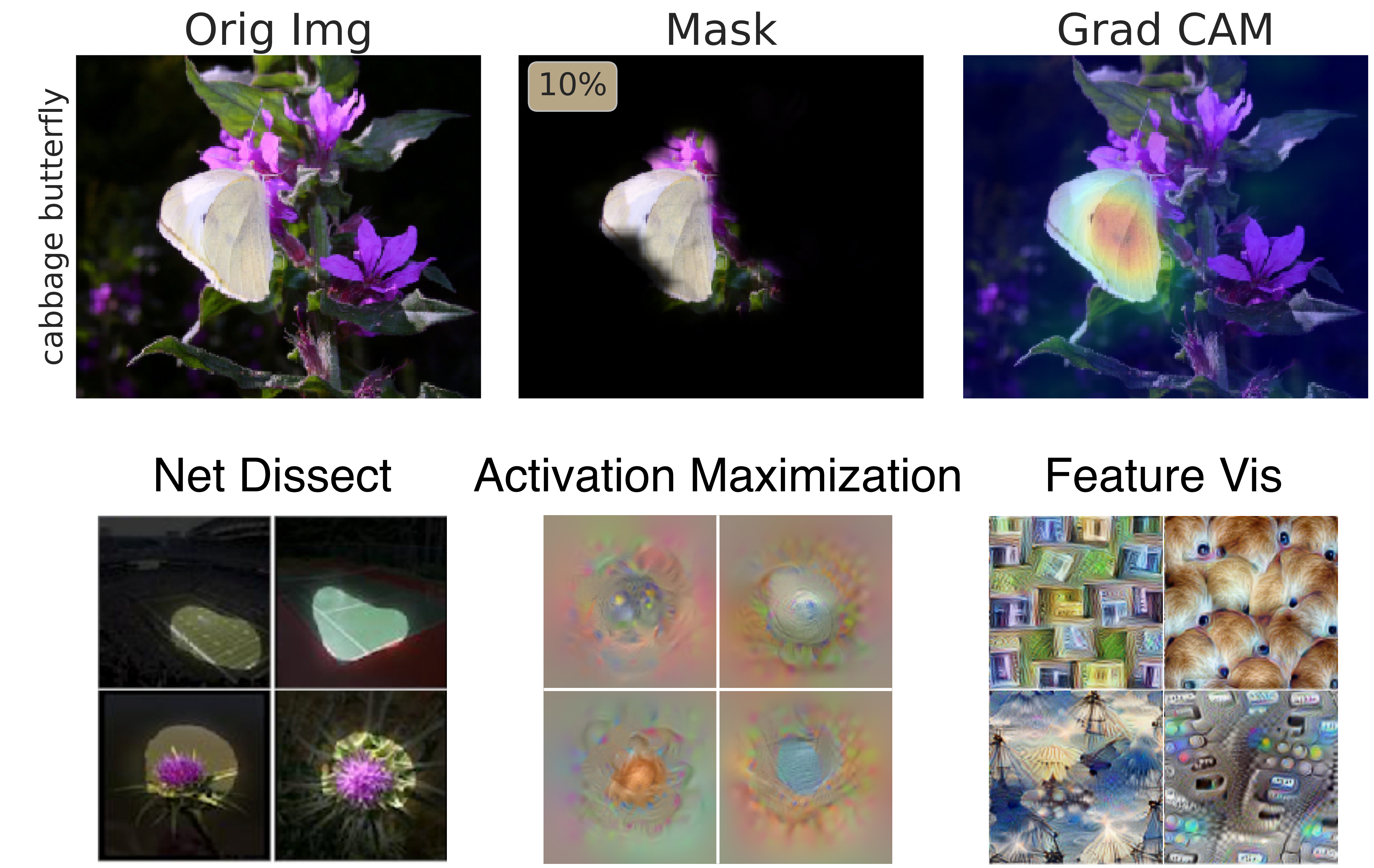
- Skill cost: web development skills
- Financial cost: budget for AMT experiments
- Time cost: human study design and iteration (e.g. task feasibility, IRB approval, quality control)

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

Roadmap

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. **Static** visualizations → **interactive** visualizations
   Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
   Interactive Similarity Overlays.

Interpretability Tools

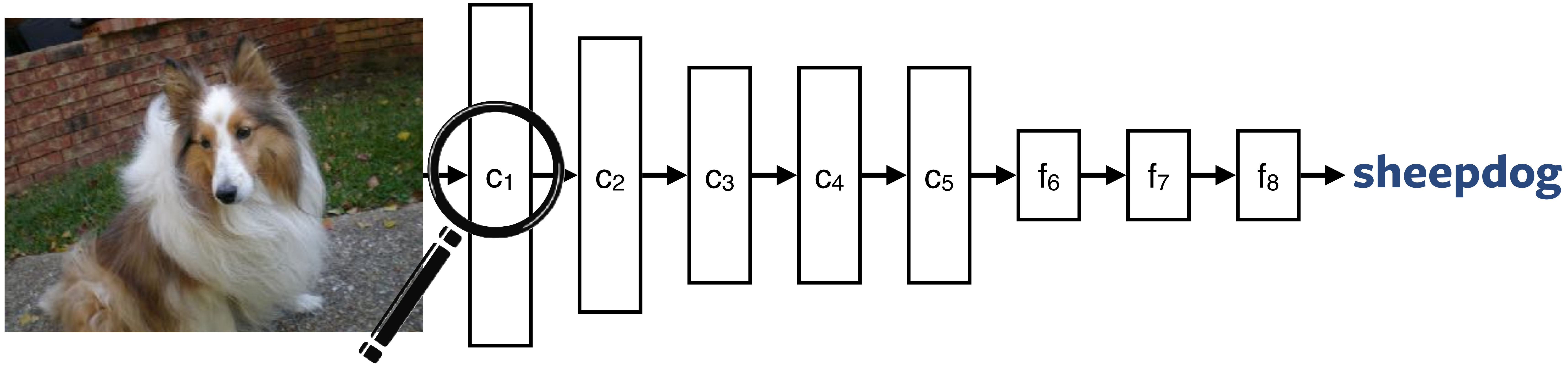


Current tools render **static images**.

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



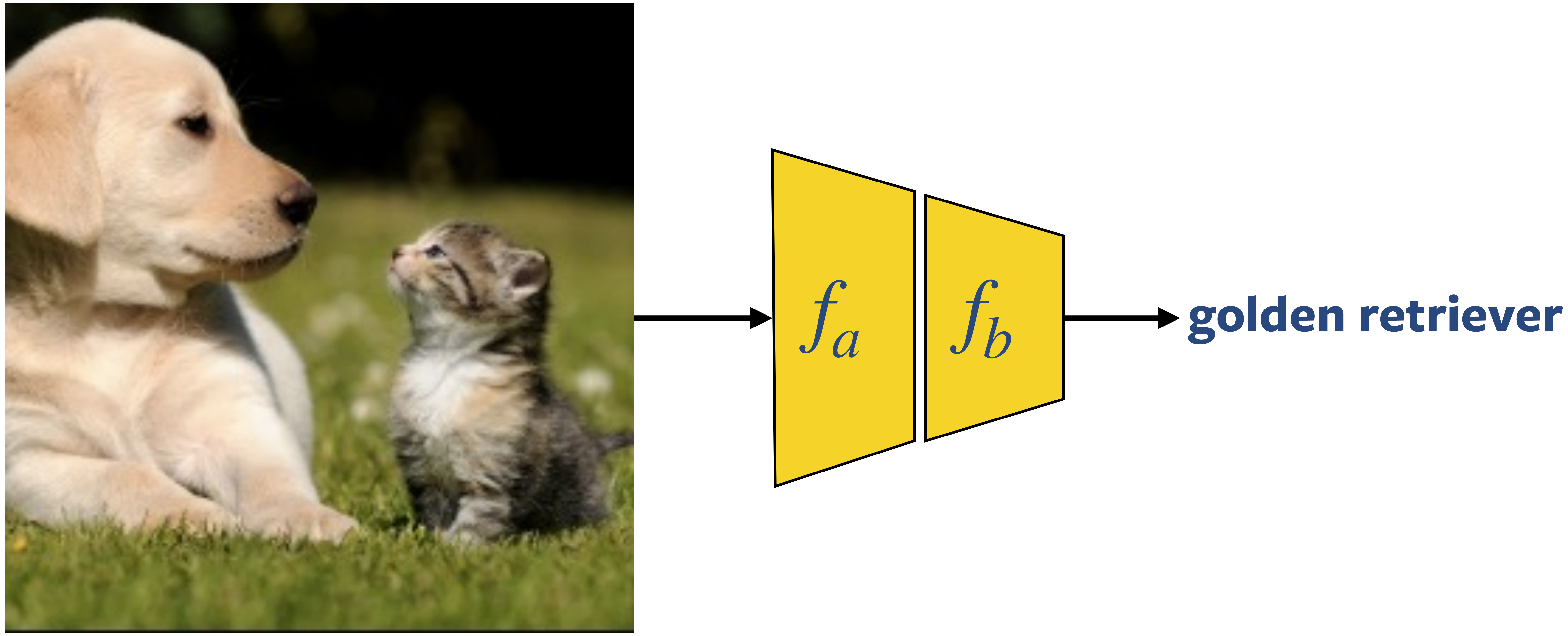
How can we **easily explore** hypotheses about the model?

Acknowledgement: Chris Olah

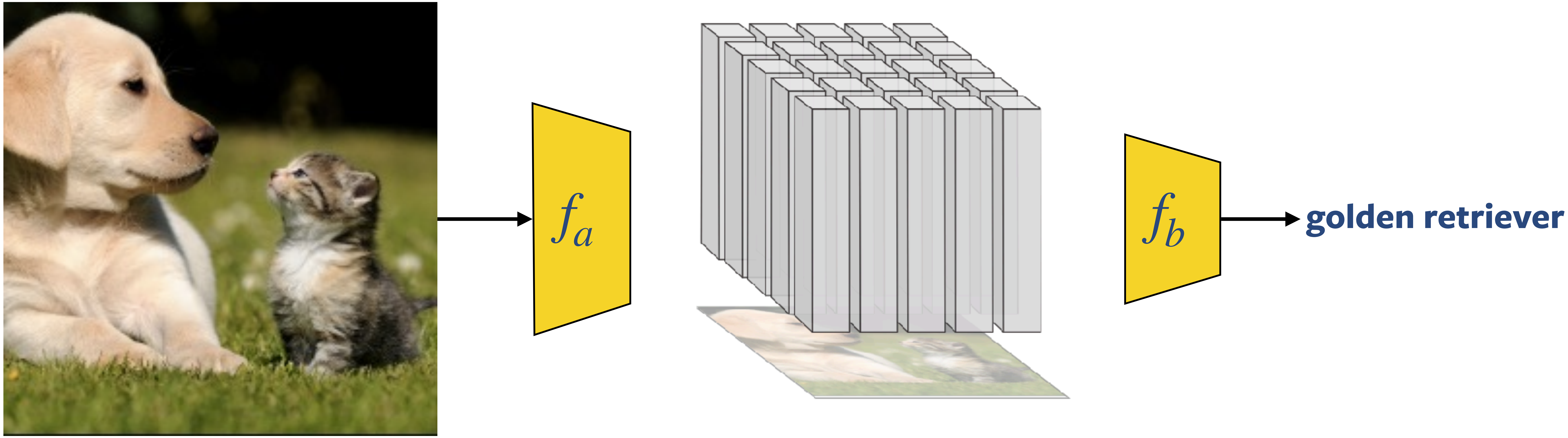
Interactive Similarity Overlays

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

Spatial Activations



Spatial Activations



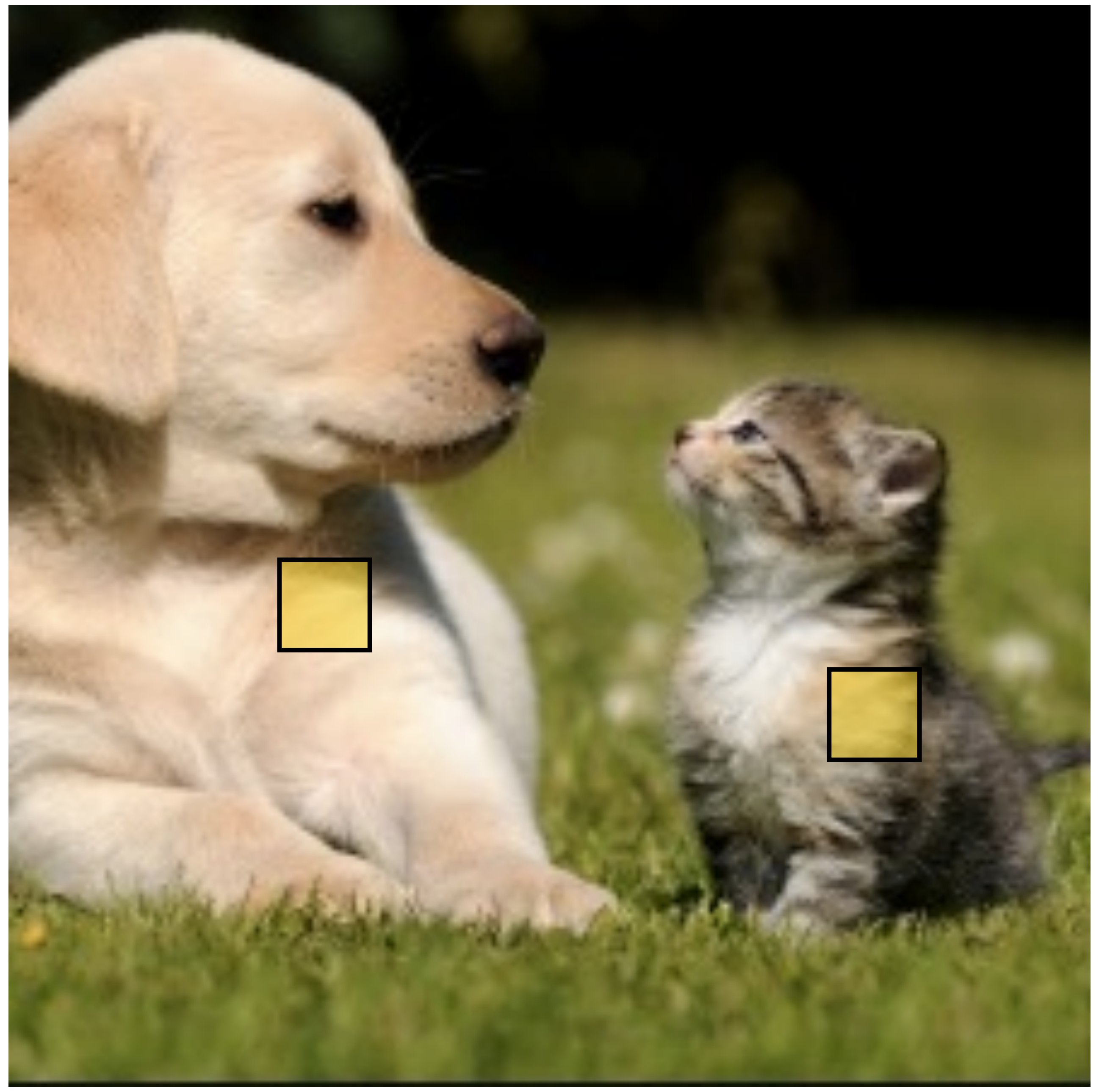
[Olah et al., Distill 2018]

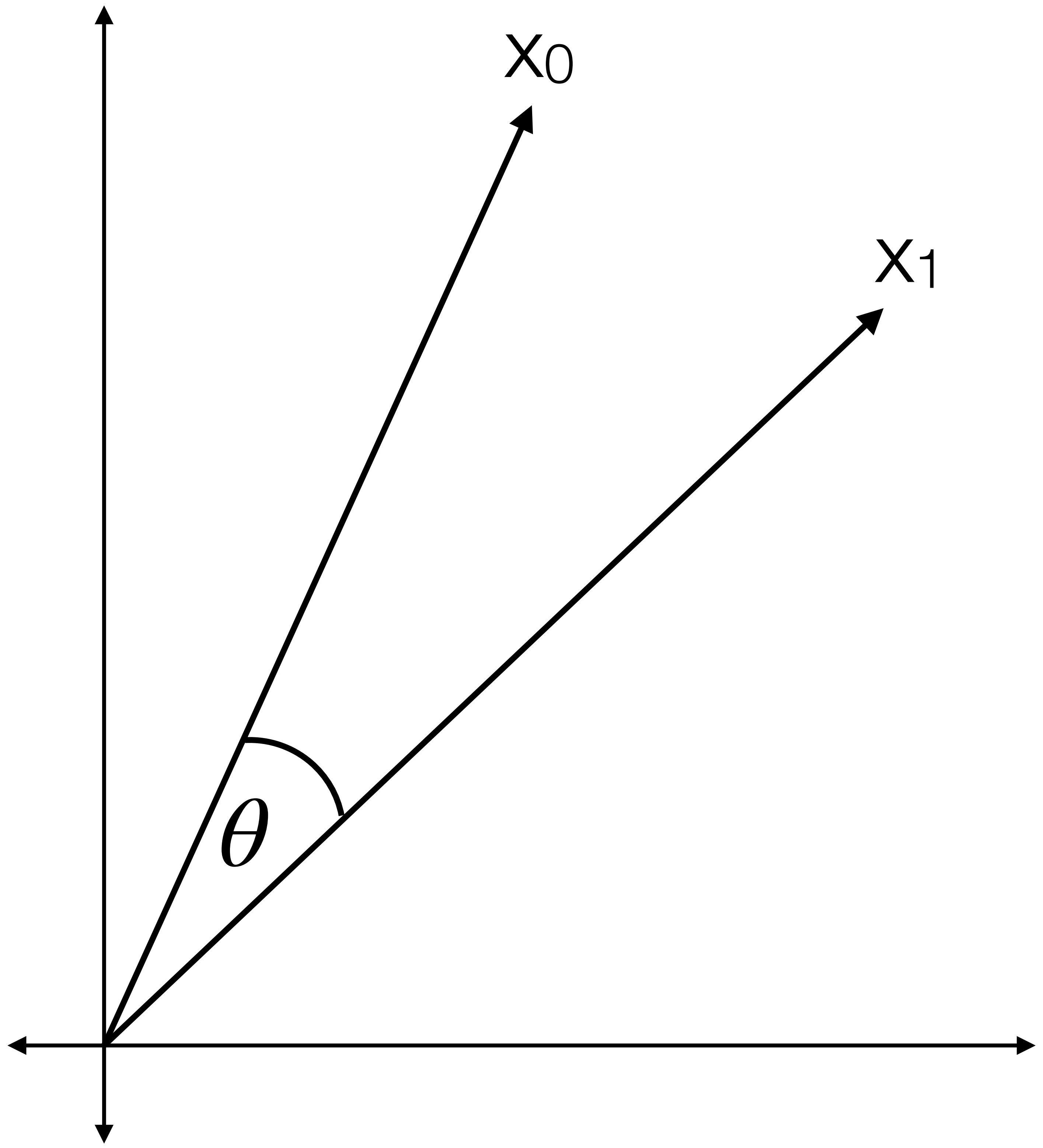
Interactive Similarity Overlays

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

[Olah et al., Distill 2018]

Interactive Similarity Overlays

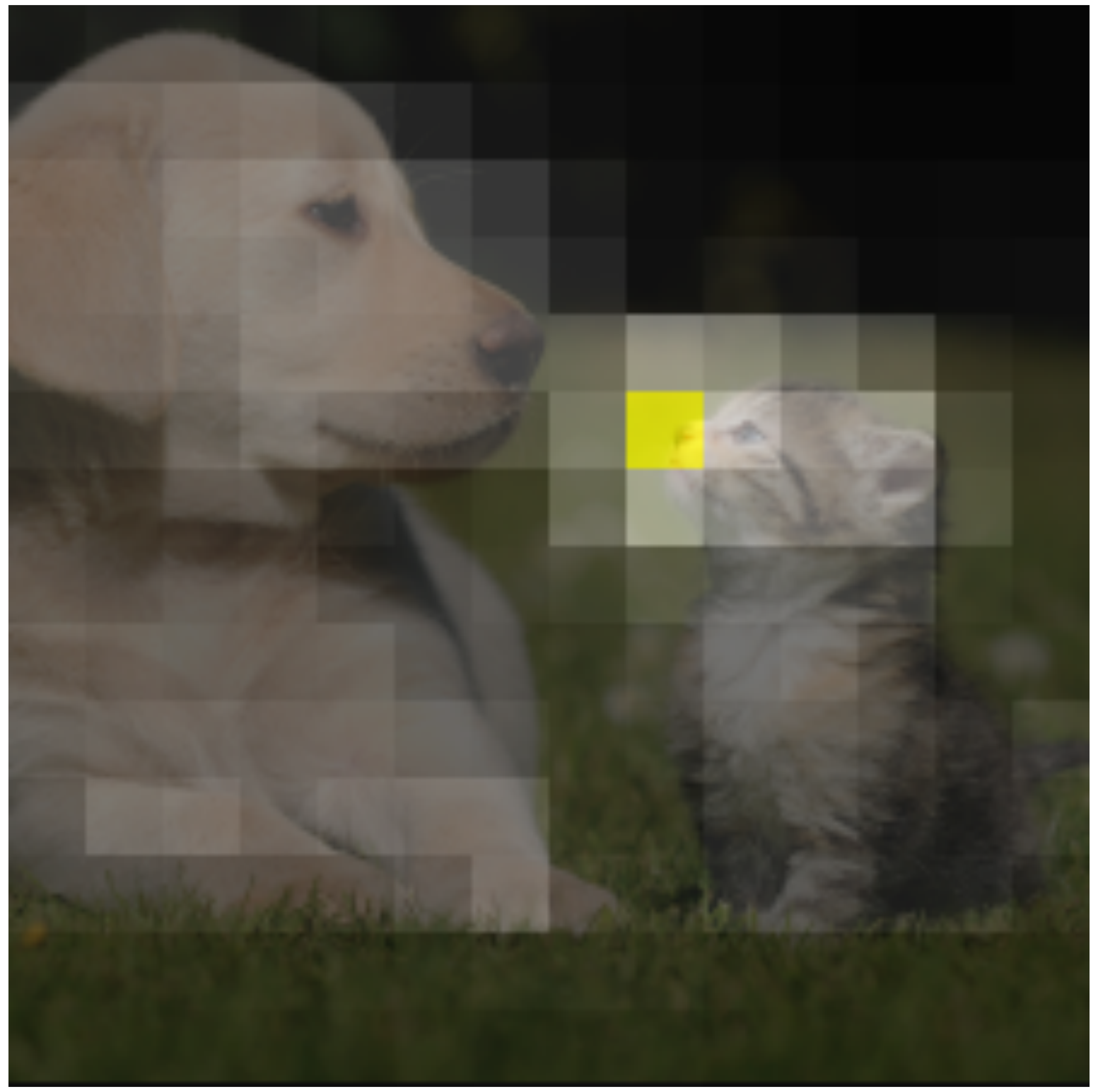




[Fong et al., VISxAI 2021. Interactive Similarity Overlays.]

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

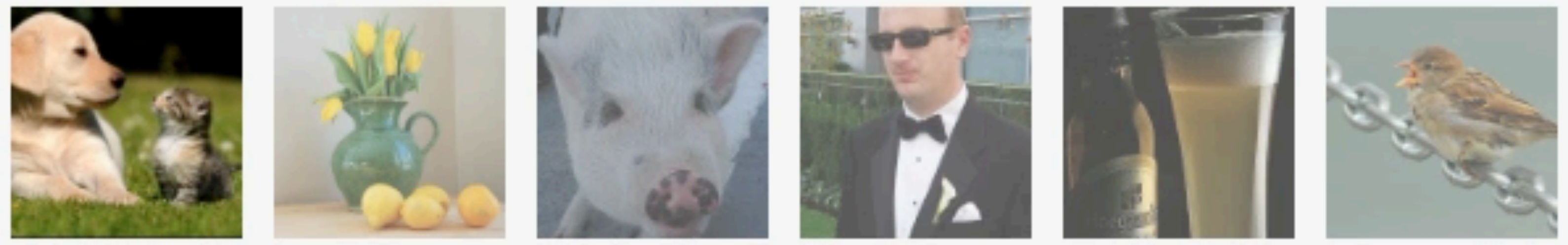
An interactive tool for understanding what neural networks consider similar and different.

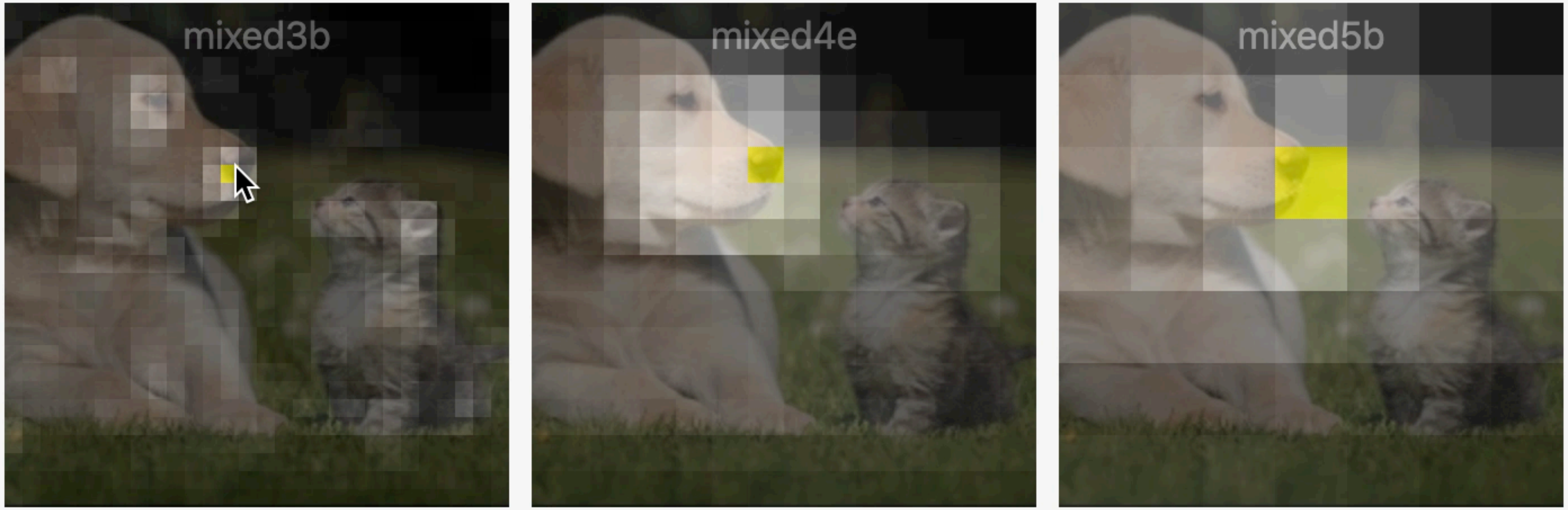
Hover over different parts of the above images. This interactive visualization shows how similar (or different) a neural network considers different image patches to the current image patch (highlighted in yellow). Try hovering over animal features (e.g., noses, eyes, faces) and background regions.

This article is best viewed in Google Chrome.

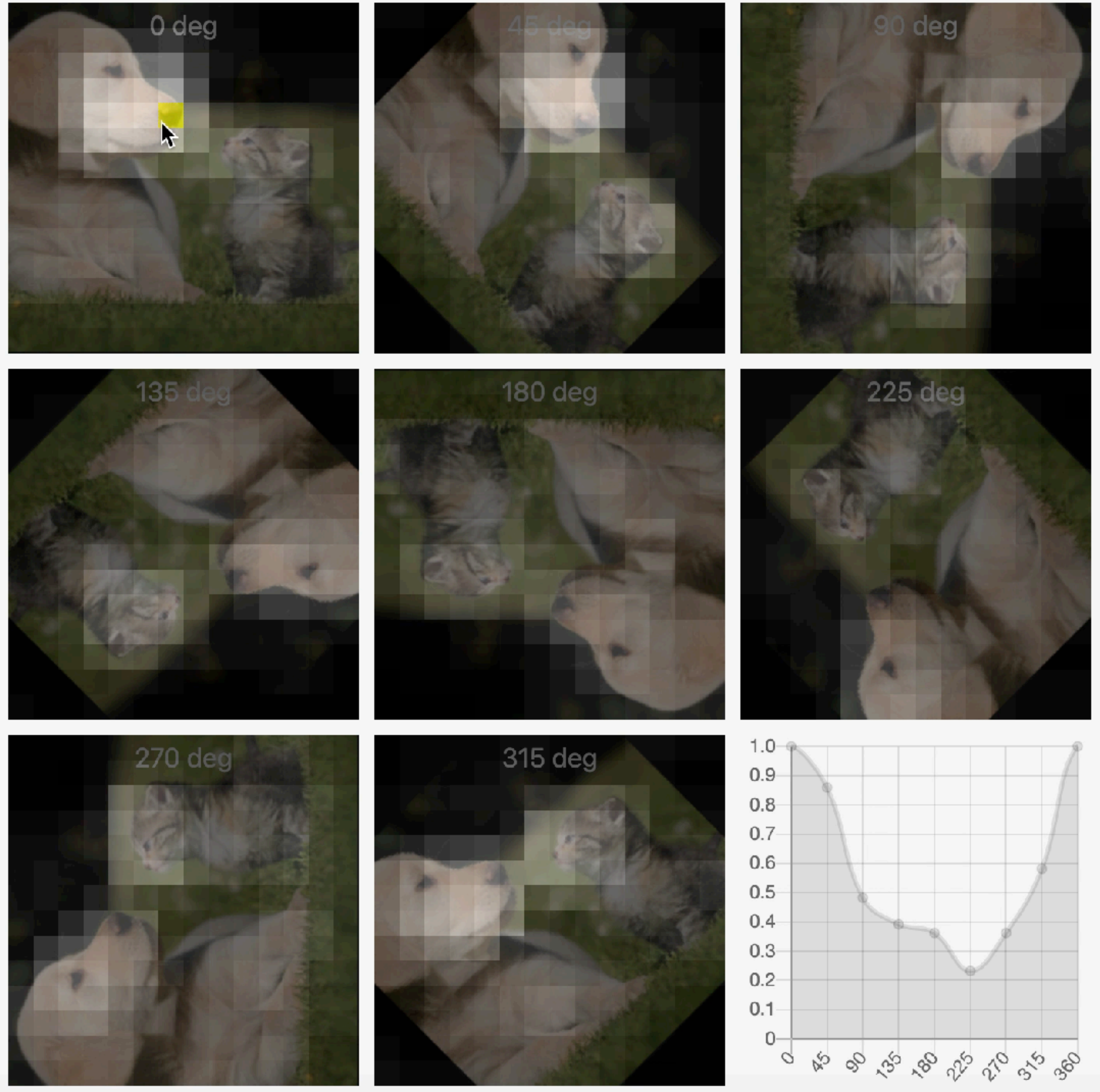
REPRODUCE IN A CO NOTEBOOK

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

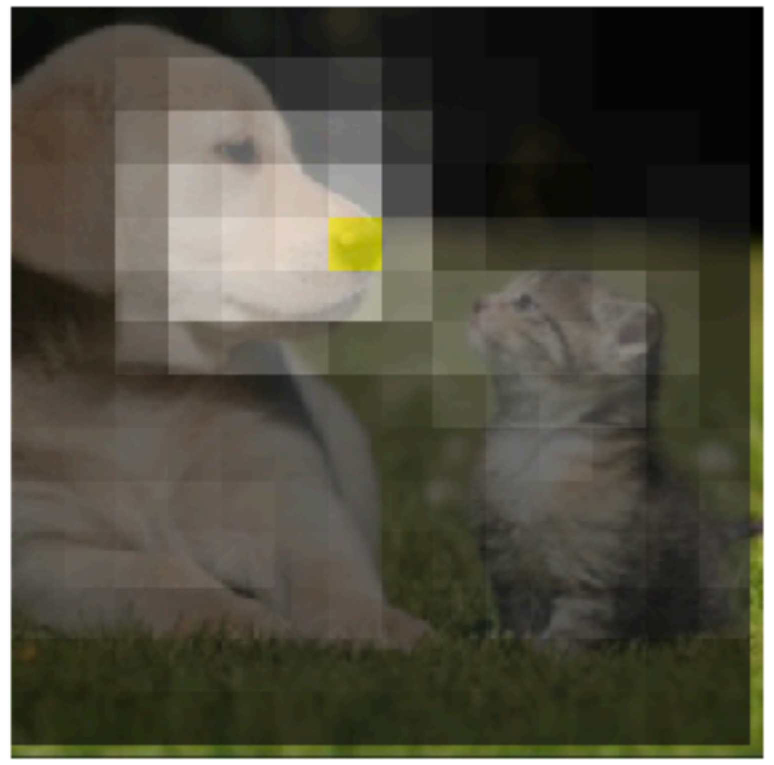
+ Code + Text Copy to Drive

```python
# 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()
```

```python
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])
```

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

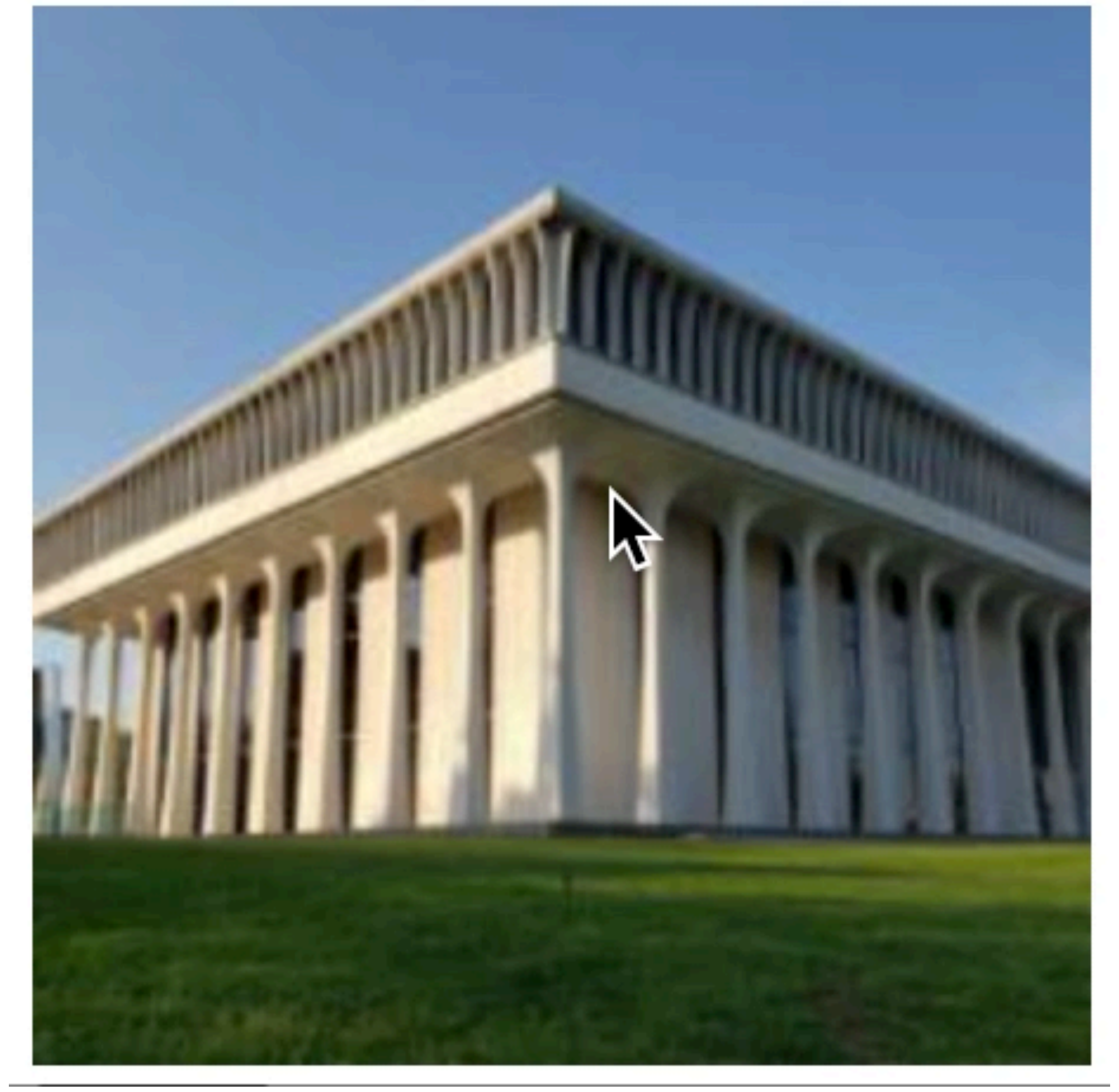


6,4

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.

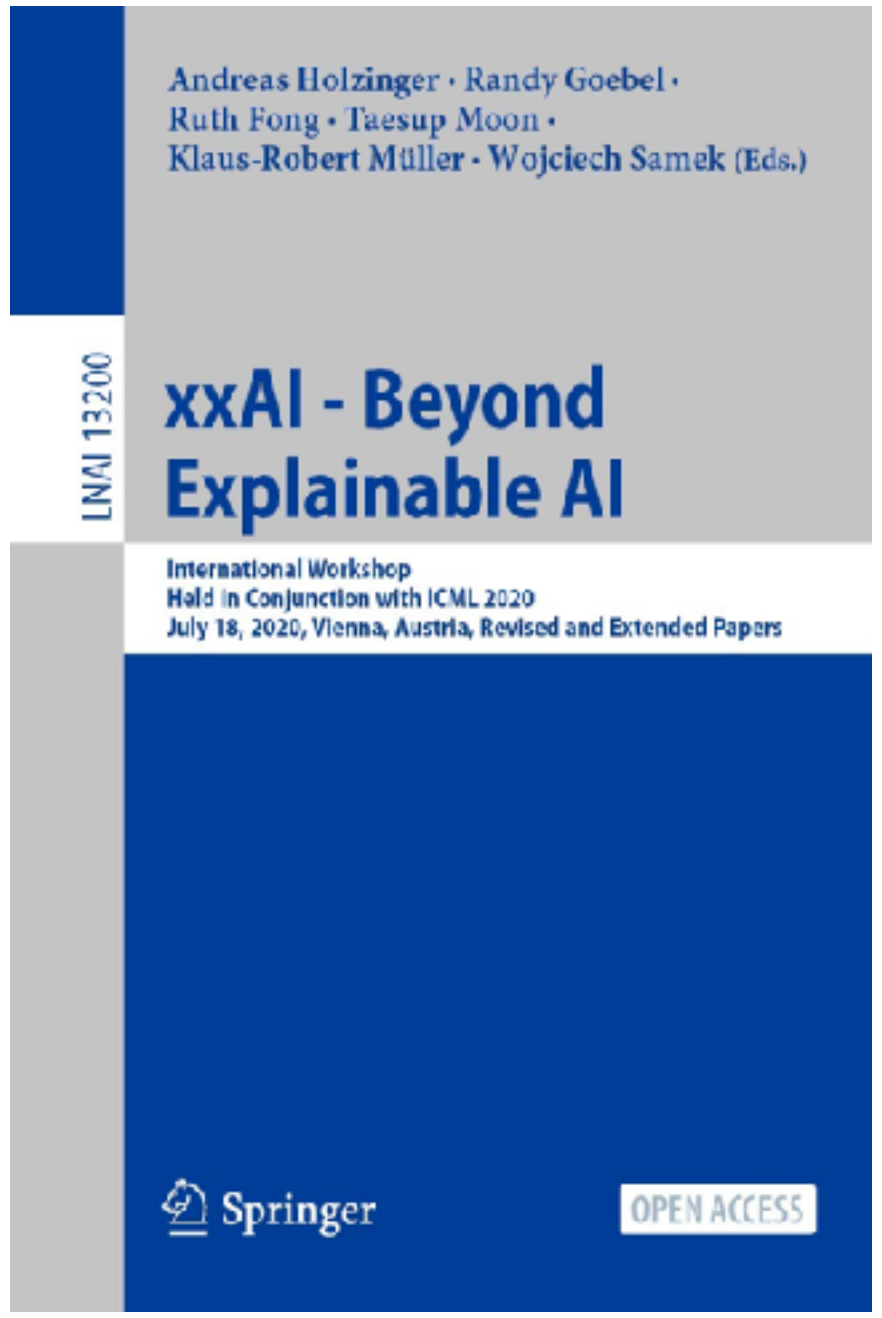
Takeaways from challenges in interpretability

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

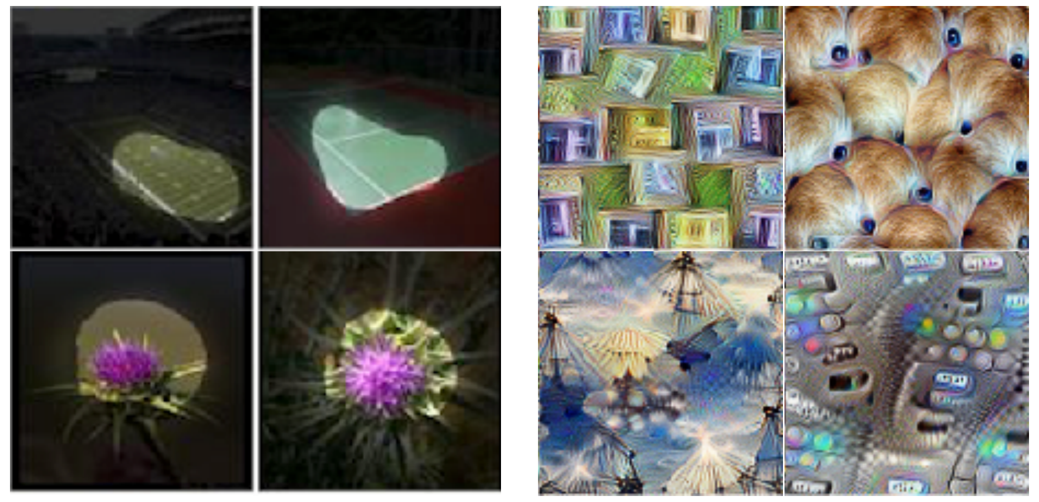
1. Develop interpretability methods for **diverse domains**
   - Beyond CNN classifiers: self-supervised learning, generative models, etc.

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]

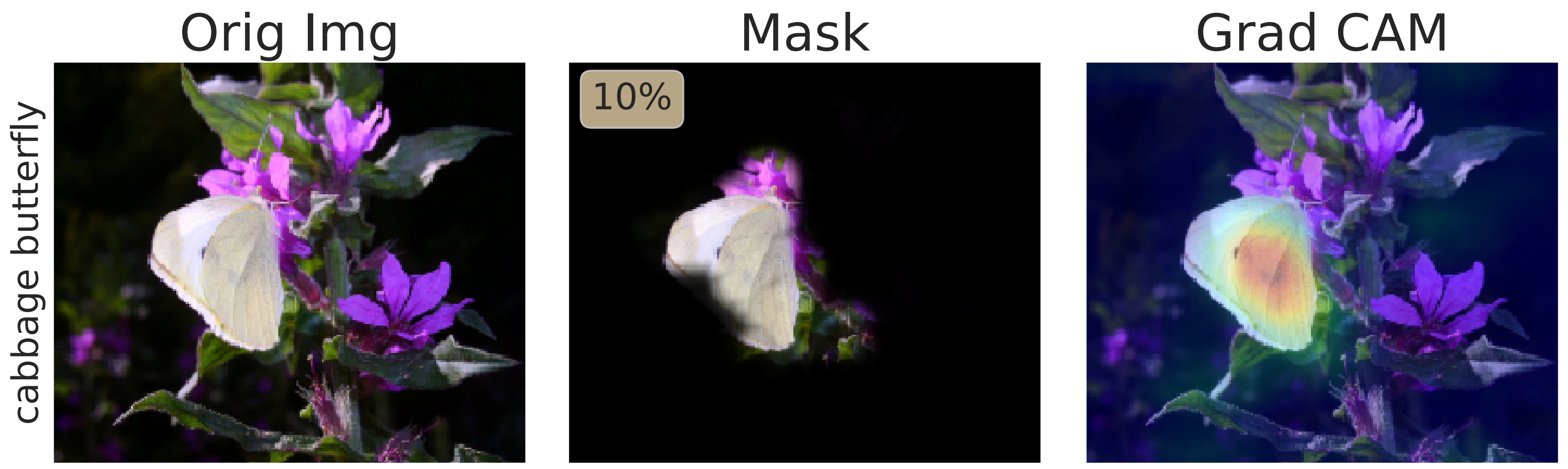
An incomplete retrospective: the first decade of interpretability



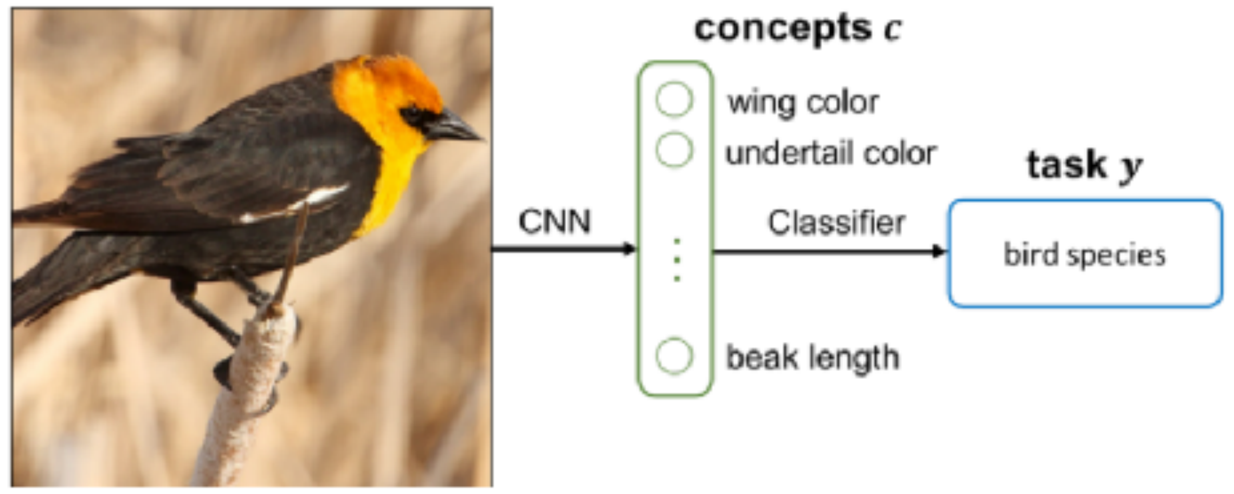
Primarily focused on understanding and approximating **CNNs**

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

UNIVERSITY OF OXFORD

RHODES TRUST

NSF

Open Philanthropy Project

bit.ly/vai-lg-postdoc

We're hiring postdocs!

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

Thank You