Directions in Interpretability

Ruth Fong

Explainability in Machine Learning, Tübingen, Germany

March 28, 2023

Slides and links available at ruthfong.com

PRINCETON UNIVERSITY

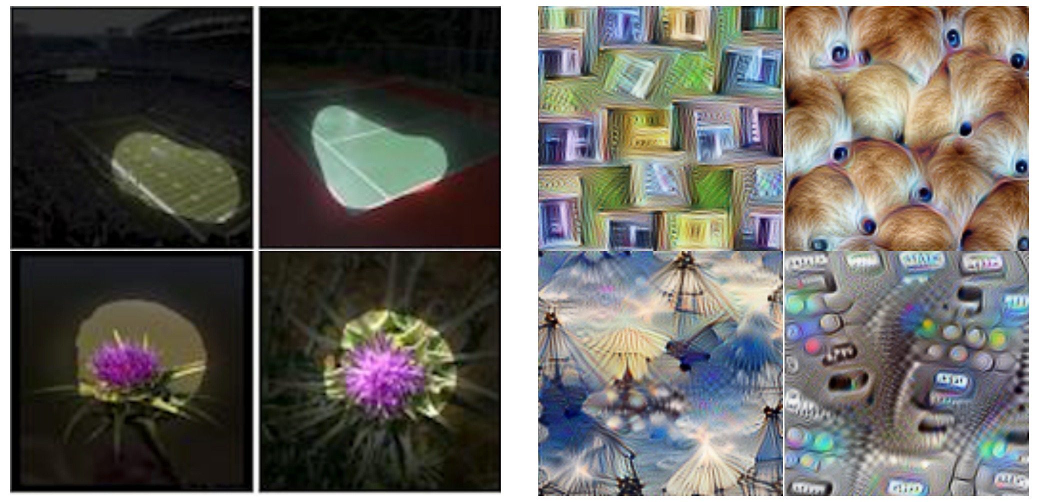
What is interpretability?

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

Why interpretability?

- 🔬 Science
- 🤝 Trust
- 🤖 Learning

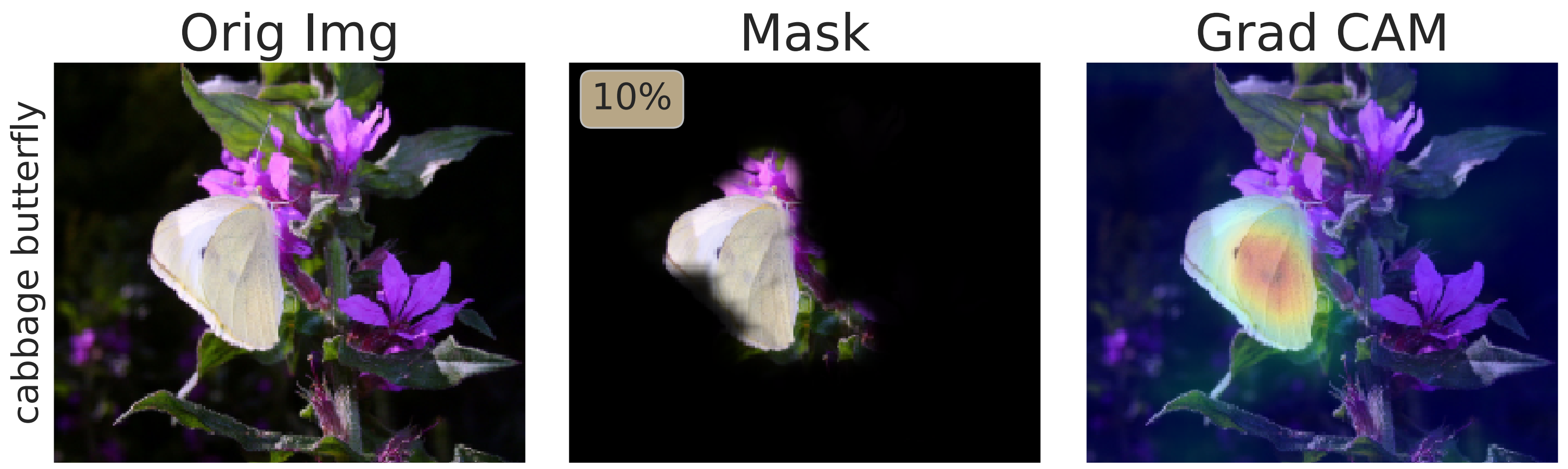
An incomplete retrospective: the first decade of interpretability



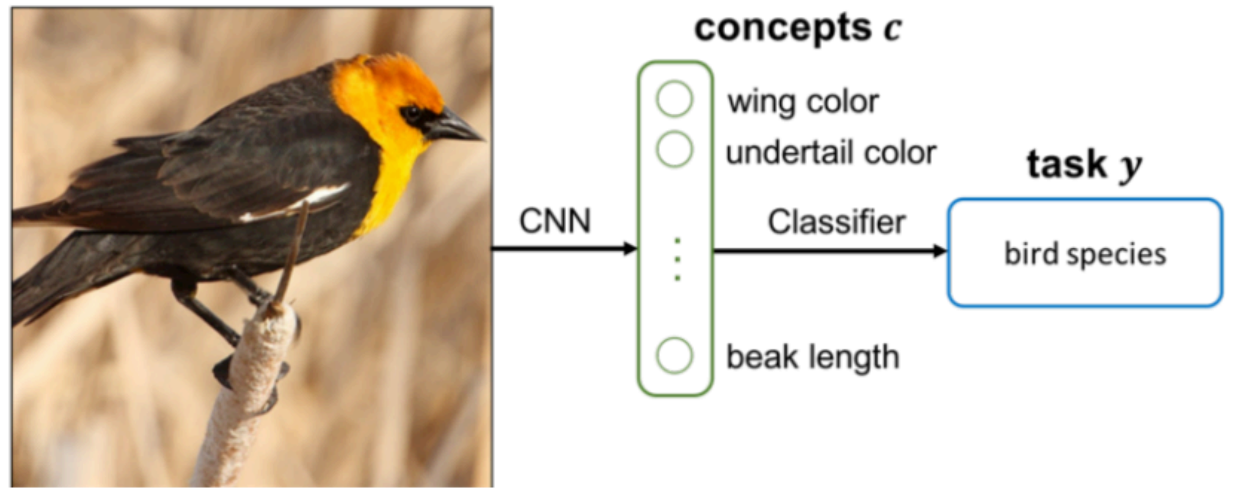
Feature visualization (2013-2018)
Activation Max., Feature Inversion,
Net Dissect, Feature Vis.

2012

2022



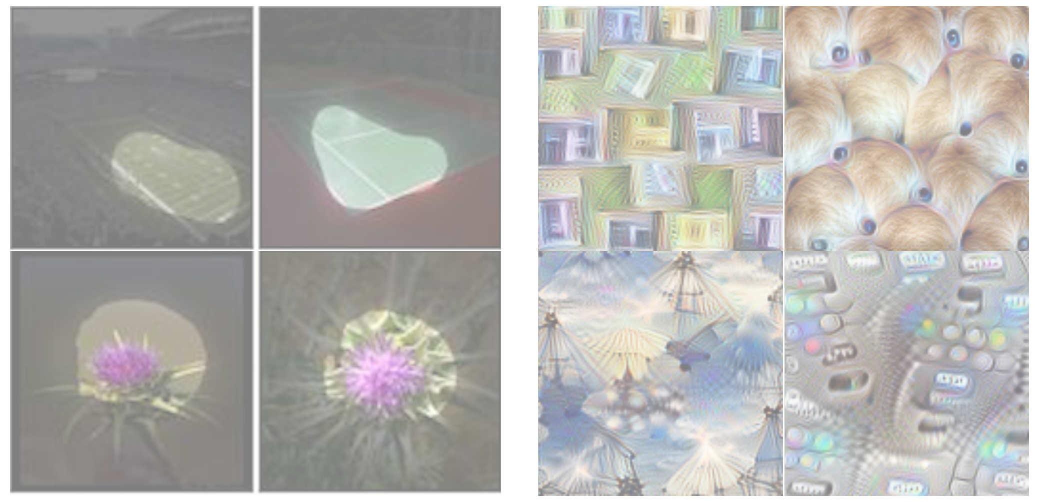
Attribution heatmaps (2013-2019)
Gradient, Grad-CAM,
Occlusion, Perturbations, RISE



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]

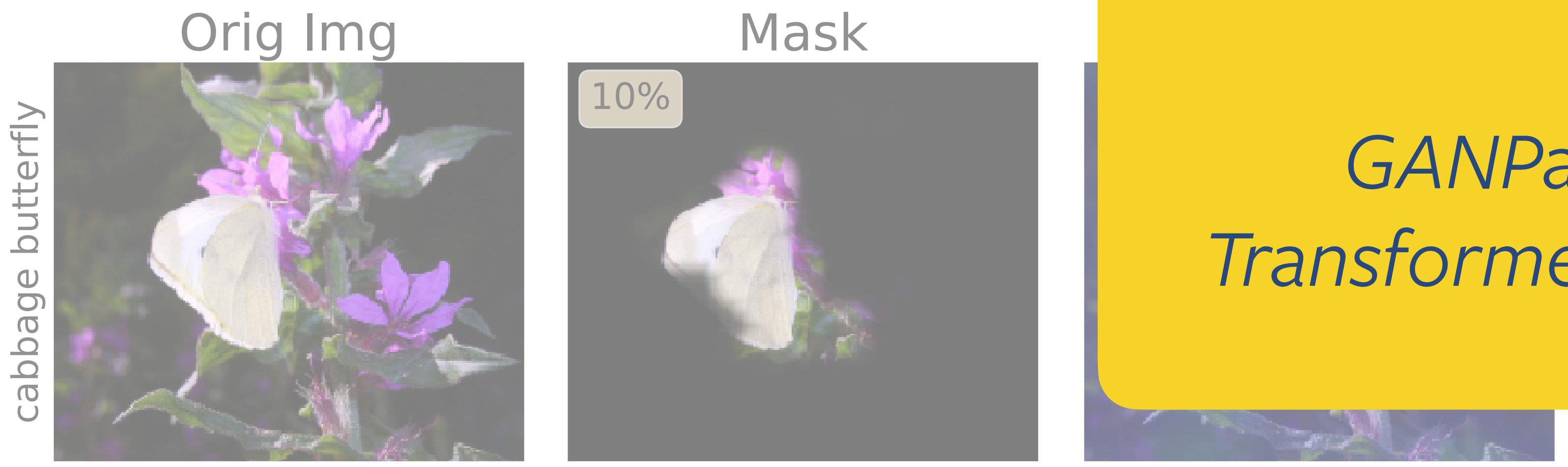
An incomplete retrospective: the first decade of interpretability

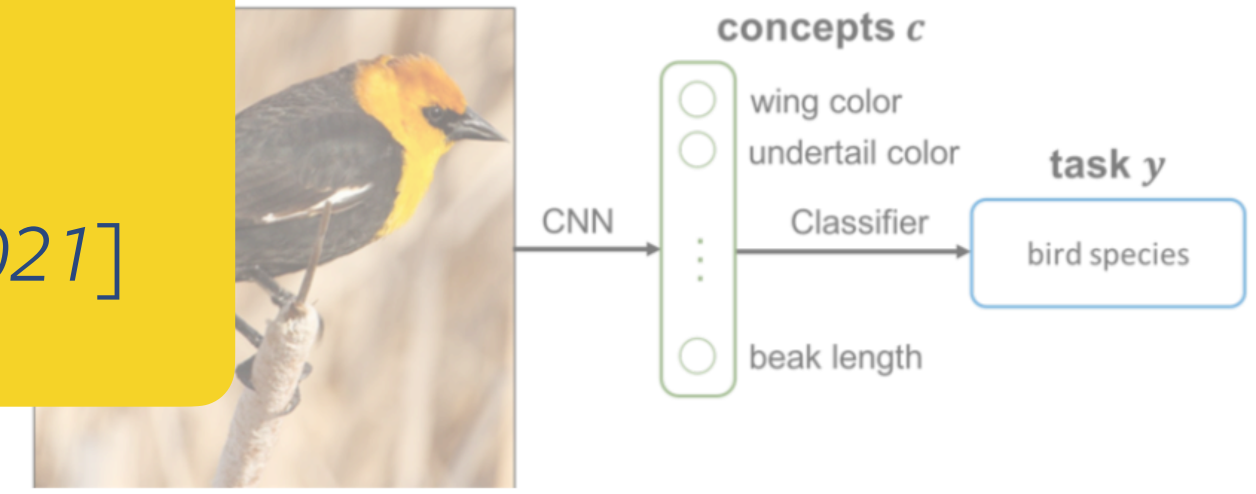


Primarily focused on understanding and approximating **CNNs**

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

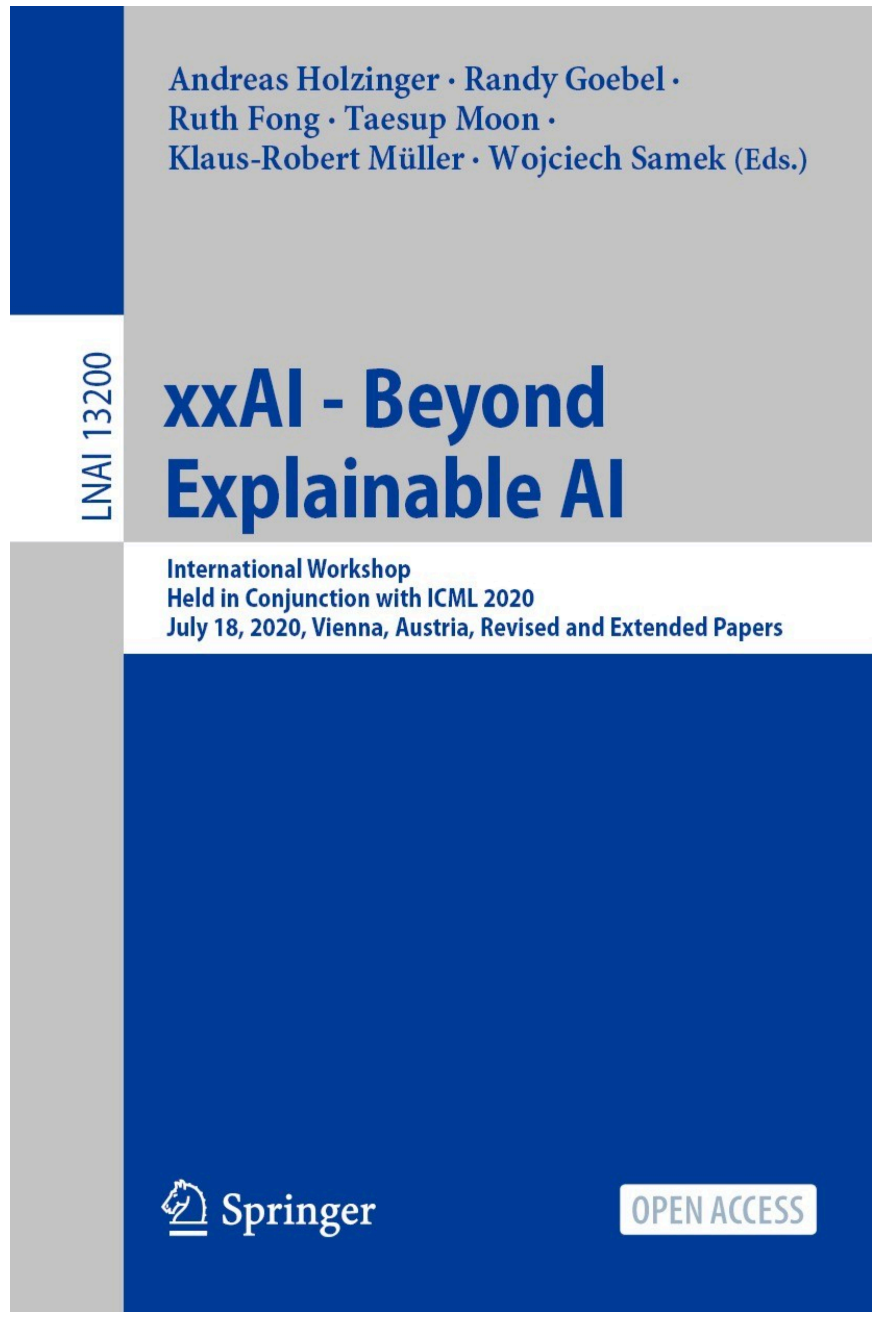
Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet,
ProtoTree

[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.



ICML 2020 workshop on XXAI

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. Interpretability by **ML researchers** → **user-oriented** interpretability
   Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández, CHI 2023.
   "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.
3. Explanations via **heatmaps** → explanations via **concepts**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky, CVPR 2023.
   Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability.
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, arXiv 2022. Interactive Visual Feature Search.*)

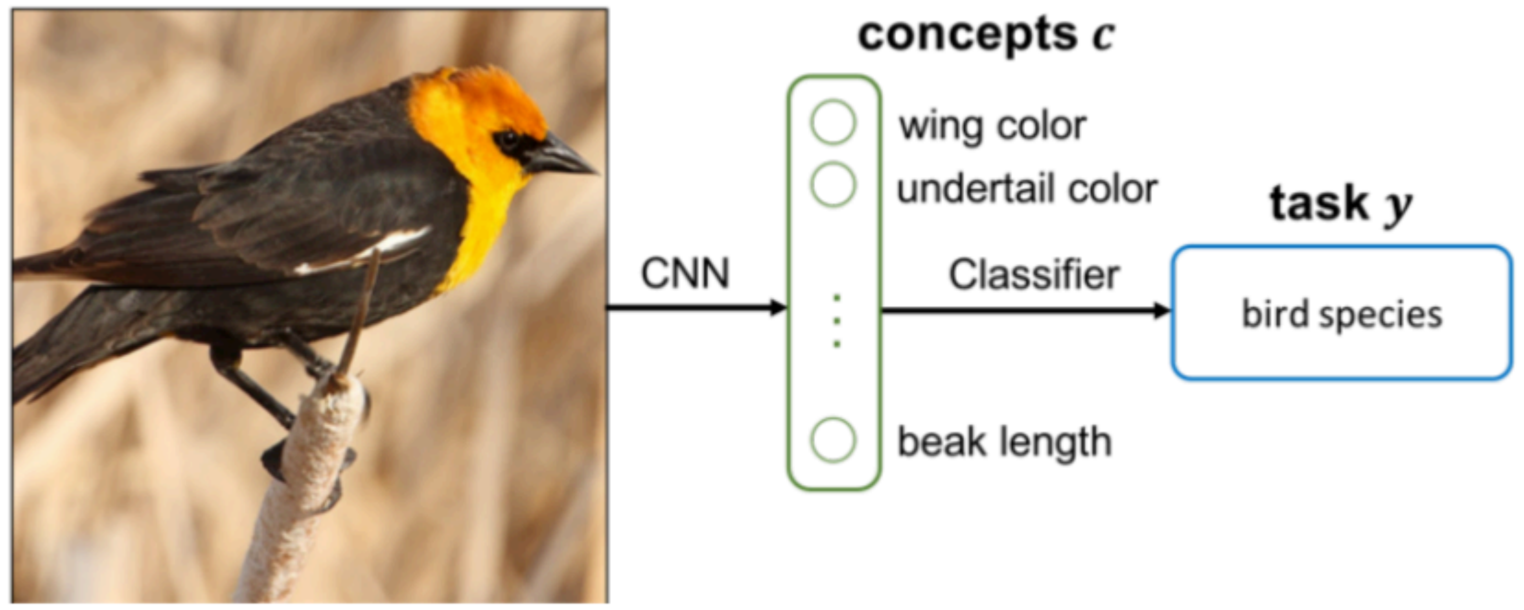
Roadmap

Sunnie S. Y. Kim

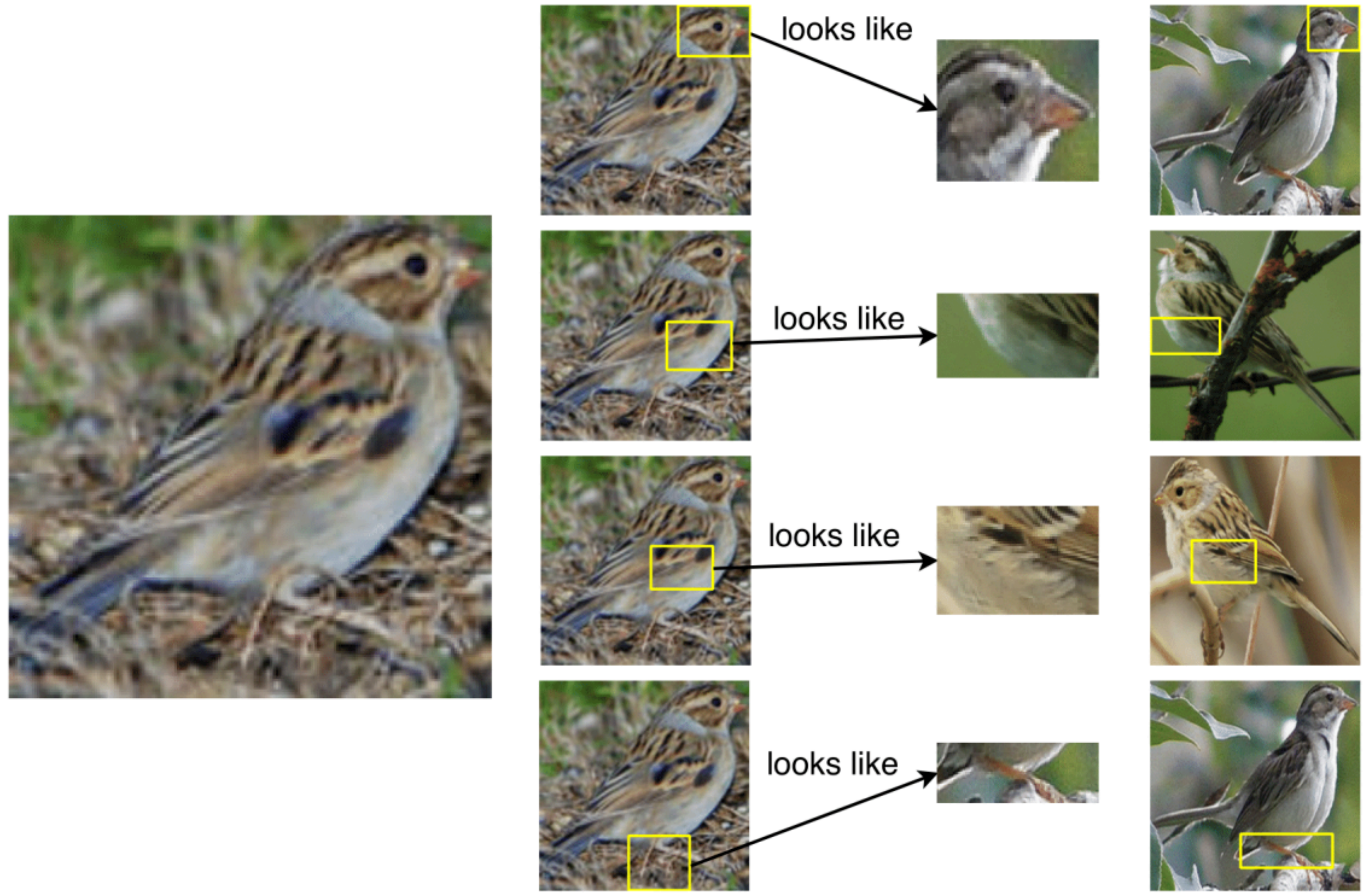
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Interpretability by **ML researchers** → **user-oriented** interpretability
   Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández, CHI 2023.
   "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.
3. Explanations via **heatmaps** → explanations via **concepts**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky, CVPR 2023.
   Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability.
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, arXiv 2022. 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)?

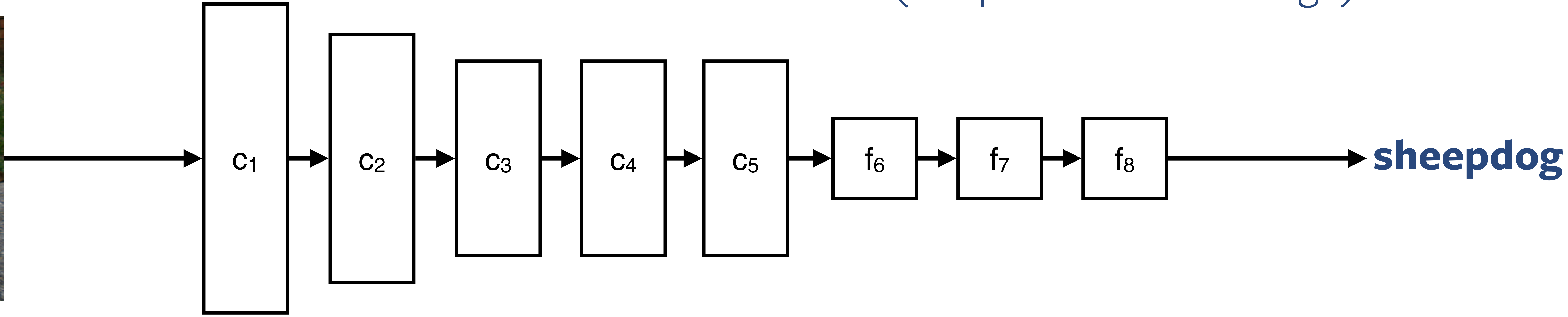
Counterfactual explanations
(e.g. SCOUT)

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

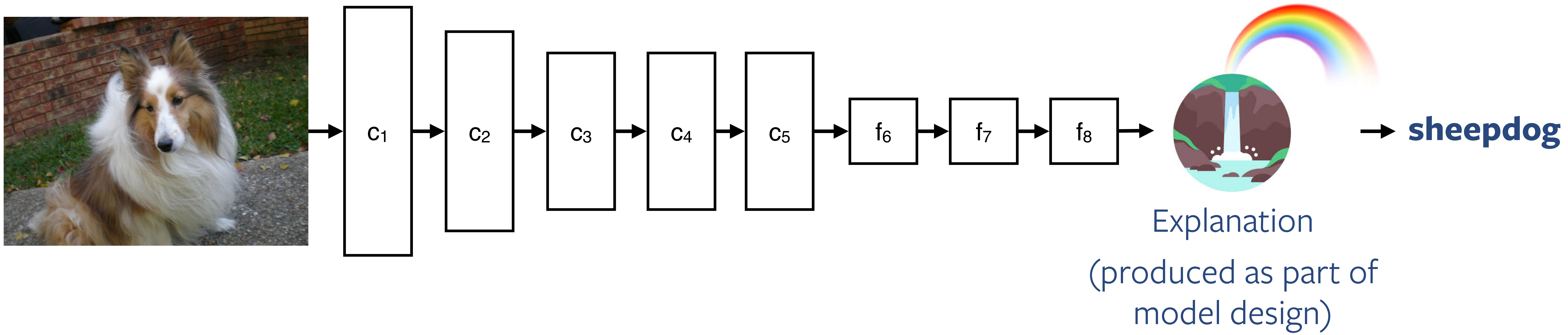
Post-hoc explanations

Explanation

(not part of model design)

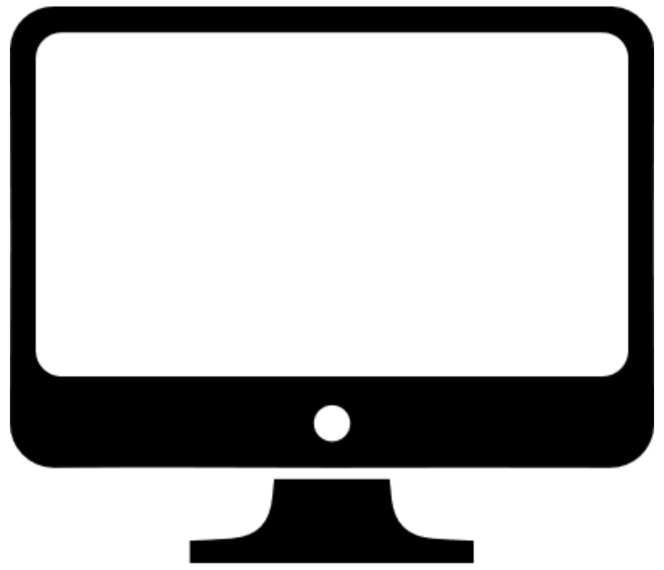


Interpretable-by-design models



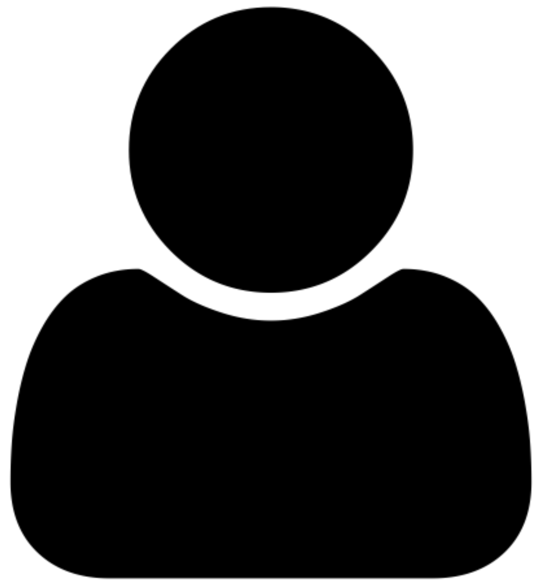
Current metrics focus on heatmap evaluation

- 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**

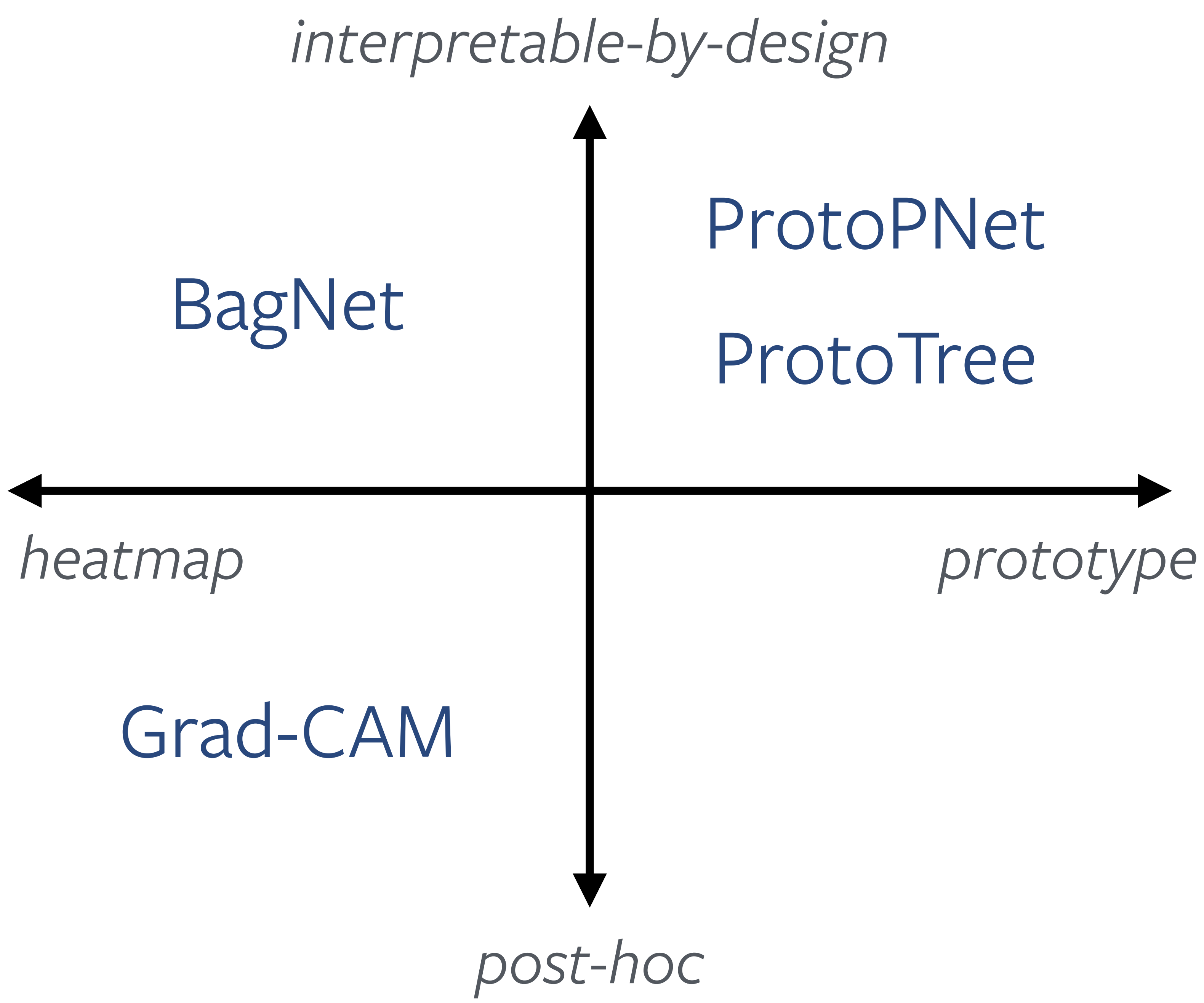
Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
HIVE: Evaluating the Human Interpretability of Visual Explanations.

Our contributions

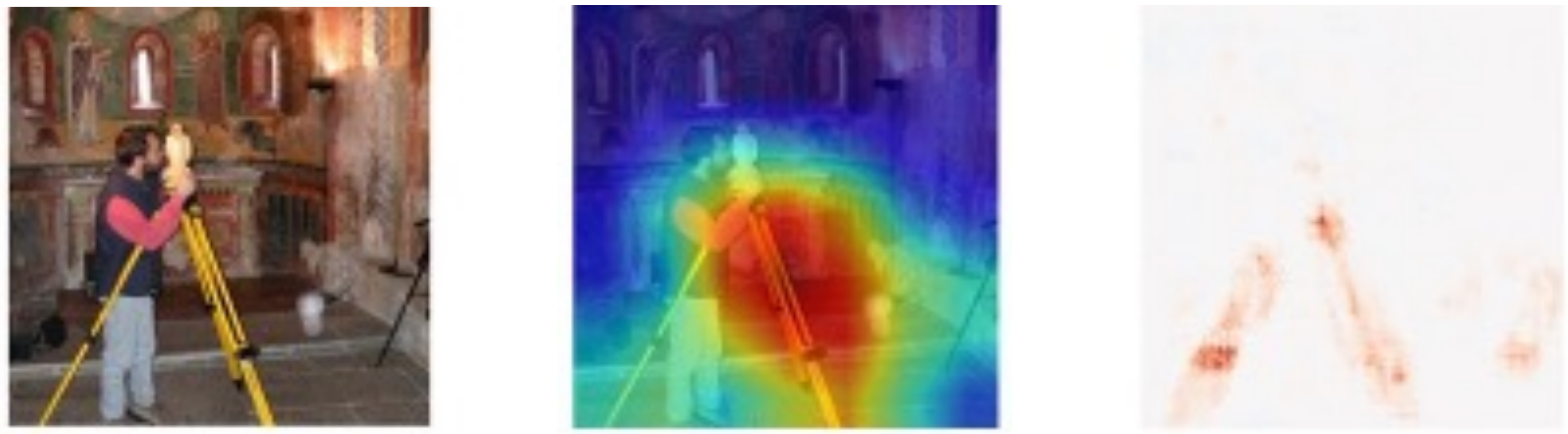
- Novel human study design for evaluating 4 diverse interpretability methods
  - **First human study** for interpretable-by-design and prototype methods
- Quantify the utility of explanations in distinguishing between **correct and incorrect predictions**
- Quantify how users would trade off between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

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

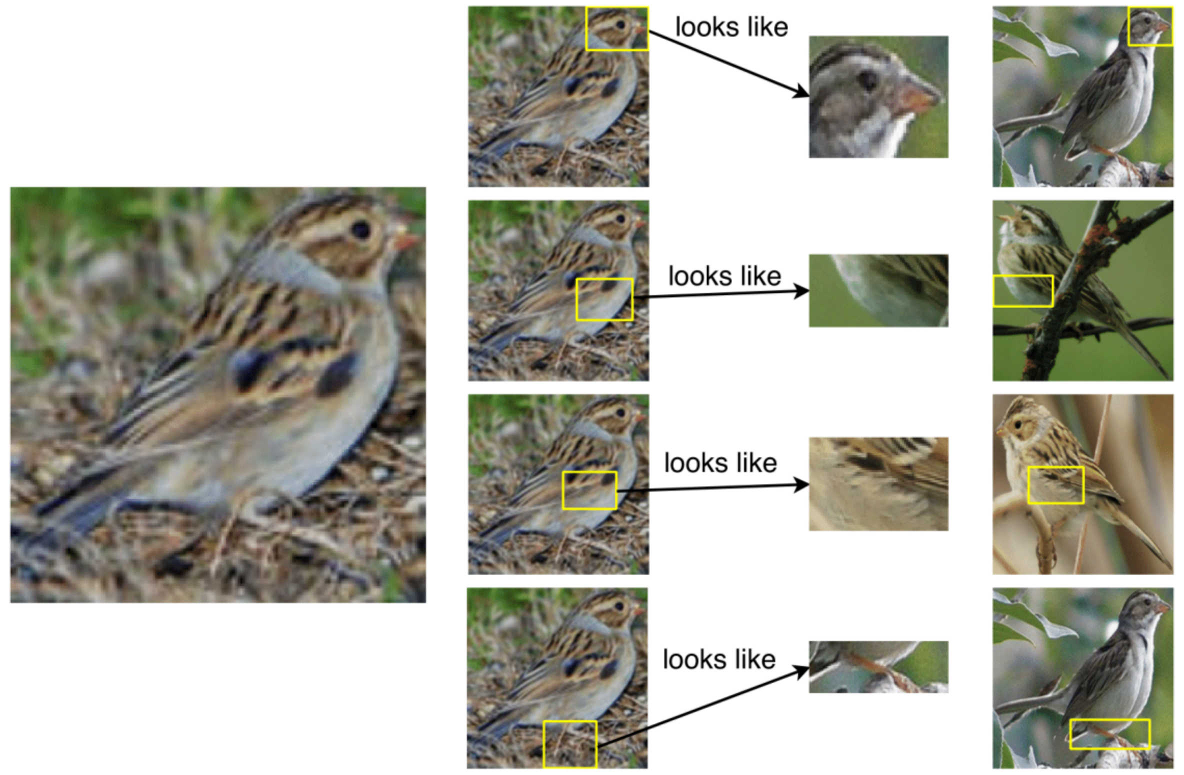
1. Cross-method comparison

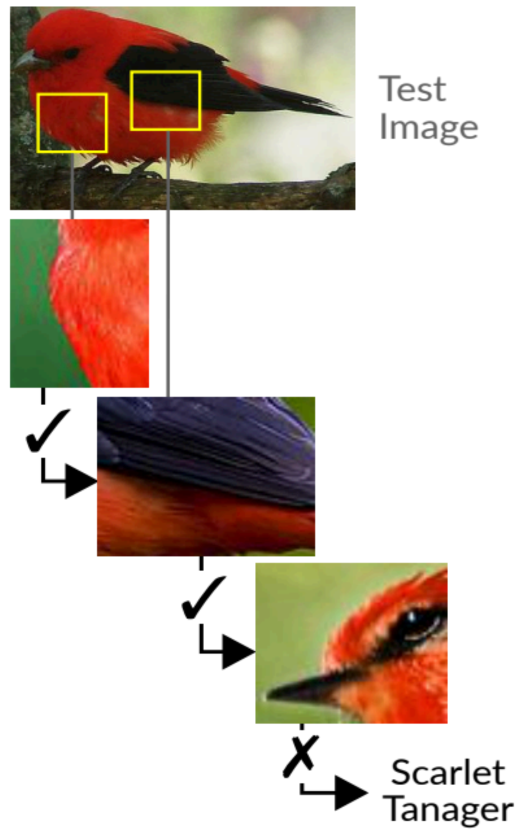


Grad-CAM BagNet



ProtoPNet ProtoTree





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

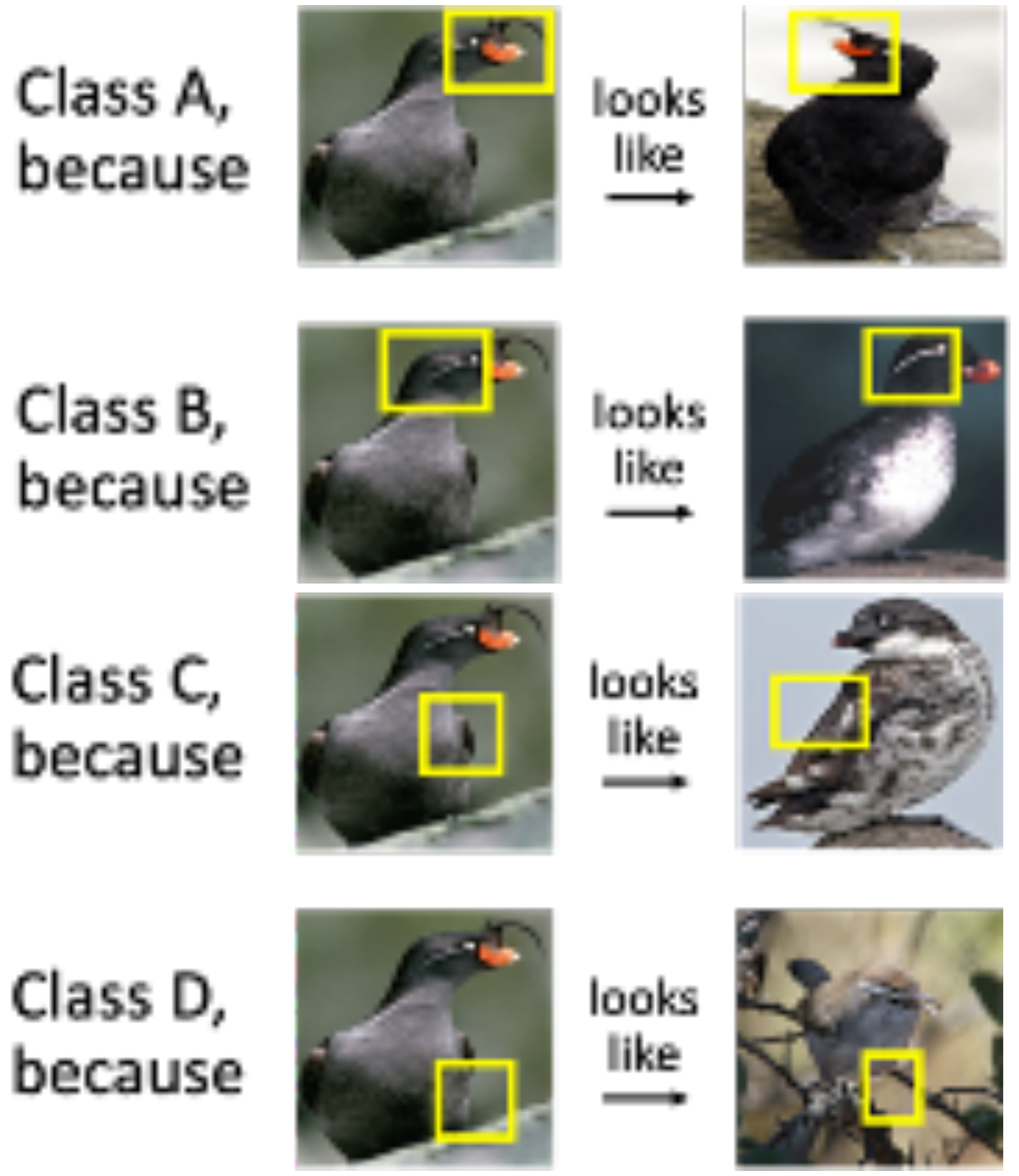
2. Human-centered evaluation

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

ProtoPNet and ProtoTree only

2. Human-centered evaluation

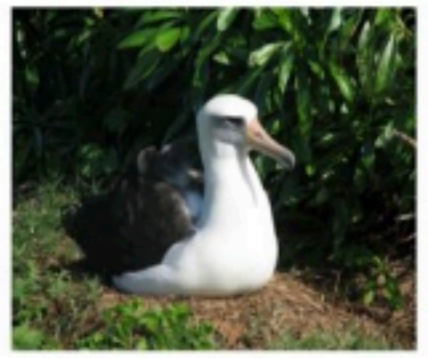
Agreement task

How confident are you in the model’s prediction?

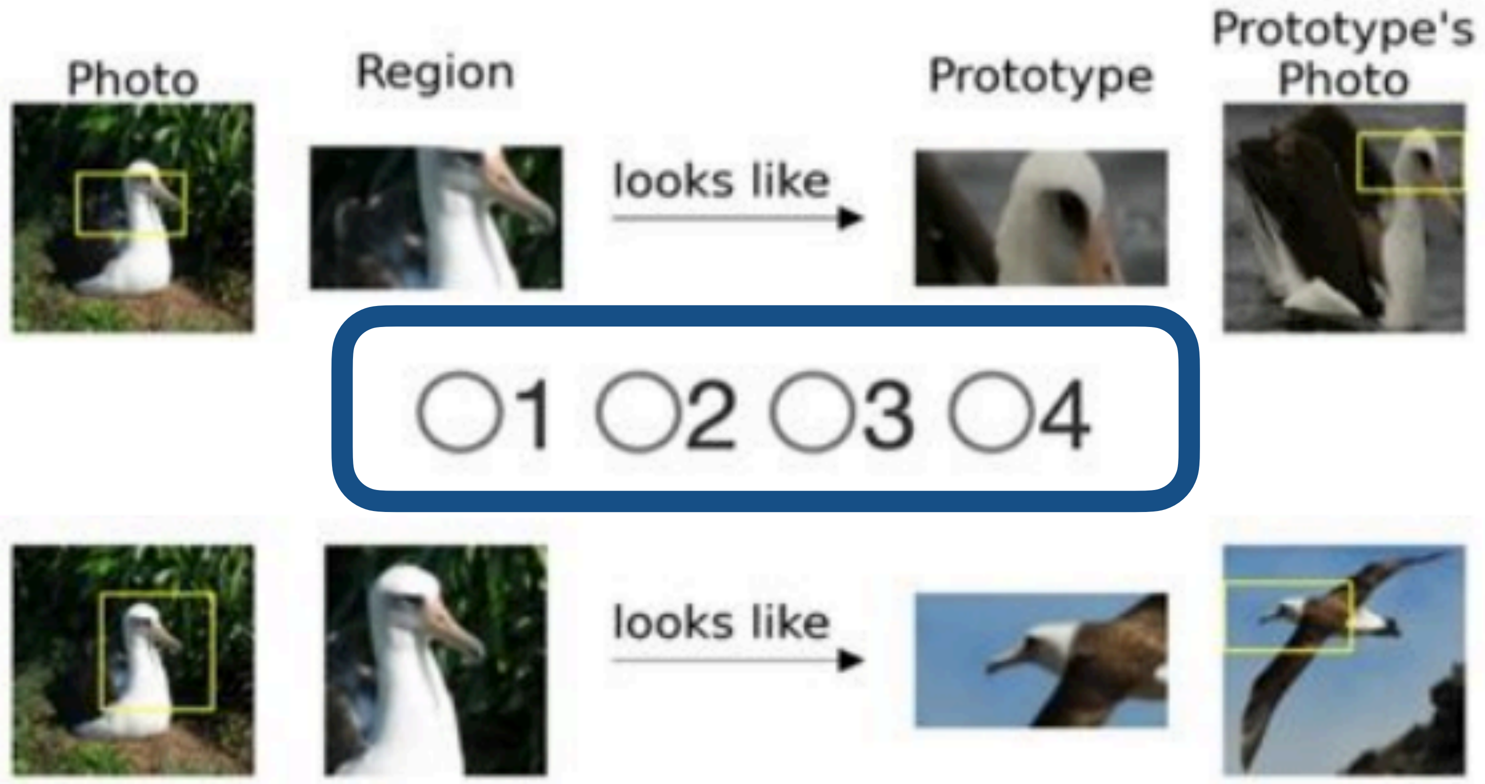
Finding #1: Prototype similarities often **do not align** with human notions of similarity.

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 <u>incorrect</u>
- ○ Fairly confident that prediction is <u>incorrect</u>

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

2. Human-centered evaluation

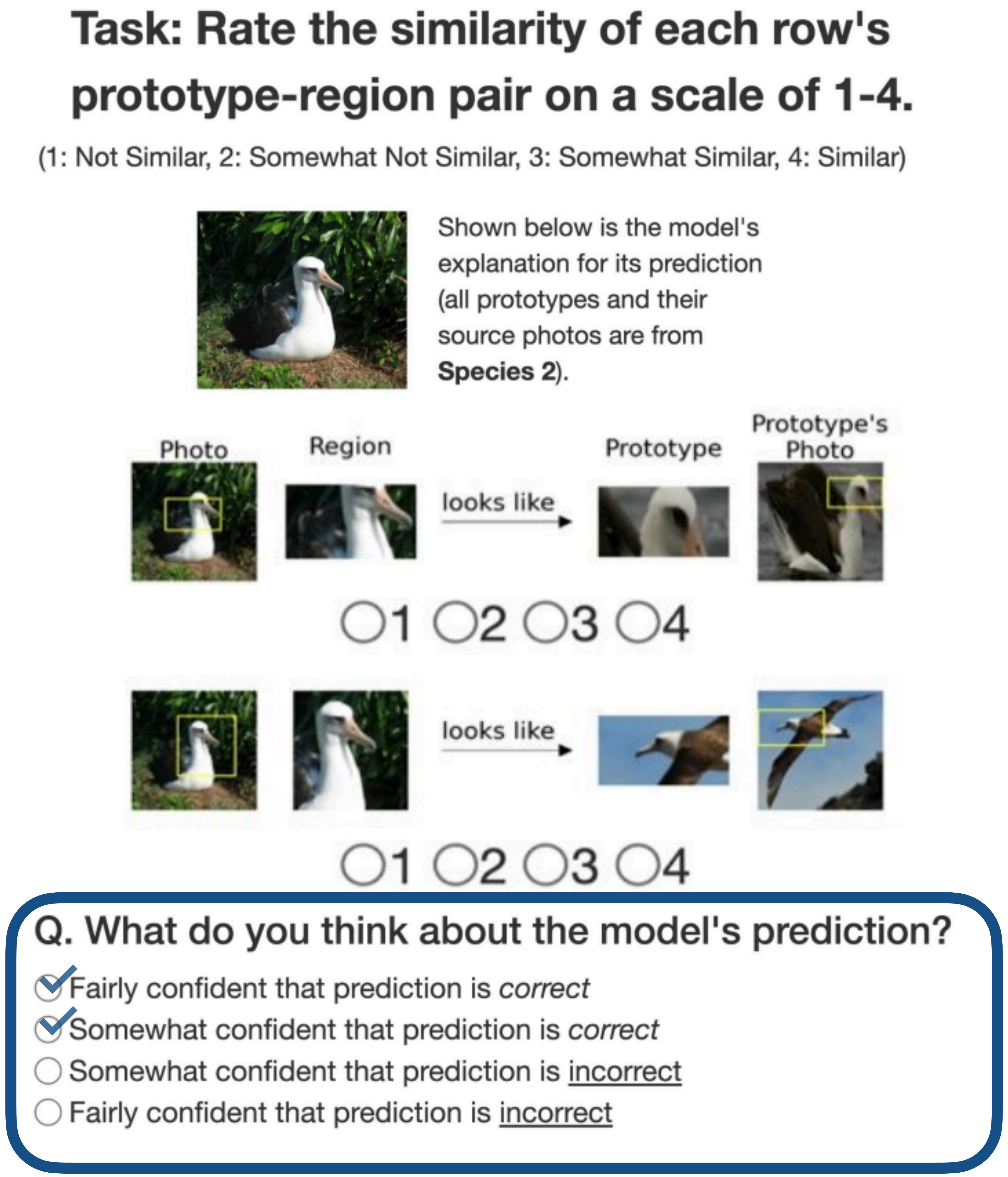
Agreement task

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

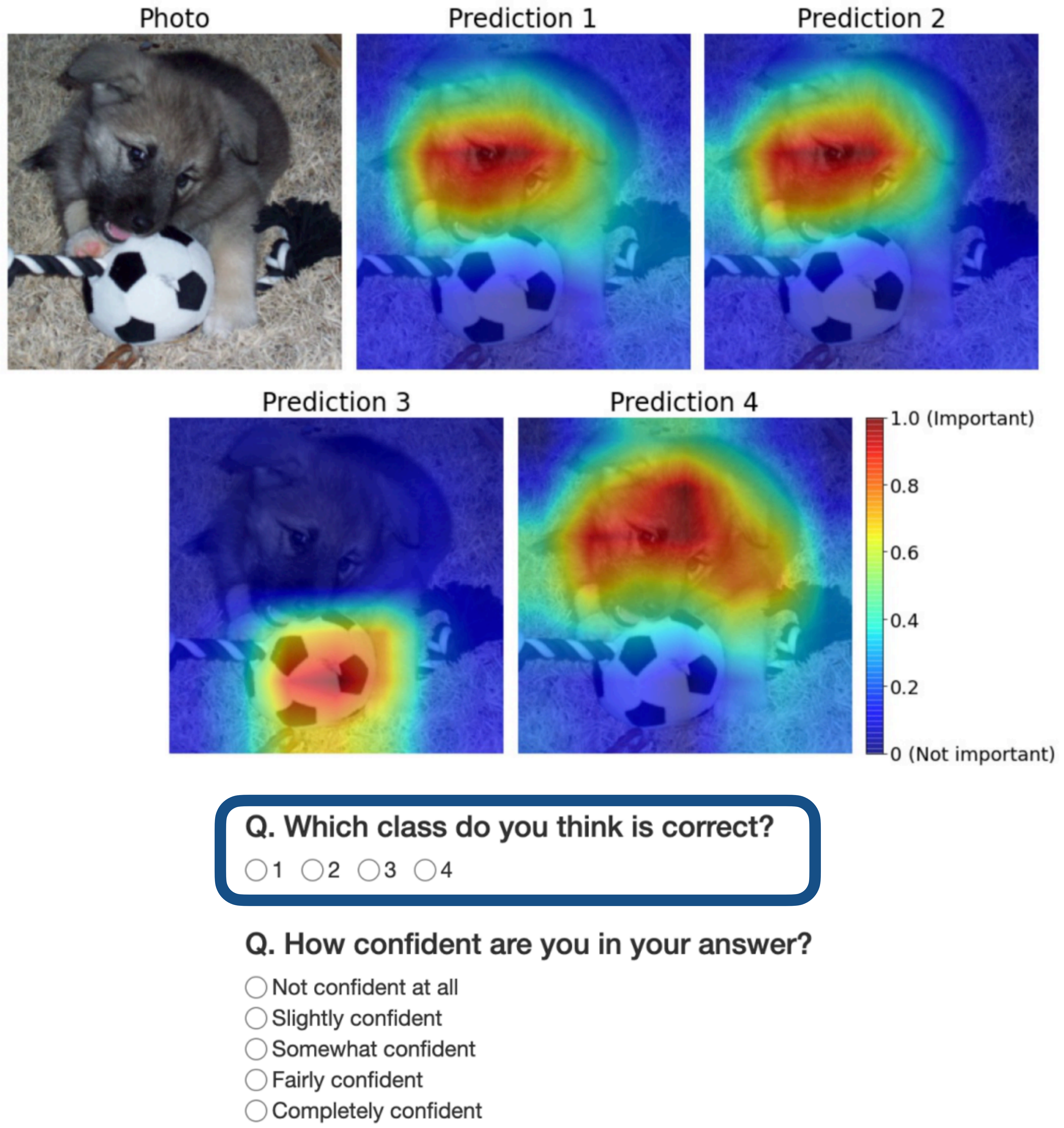
Distinction task

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.

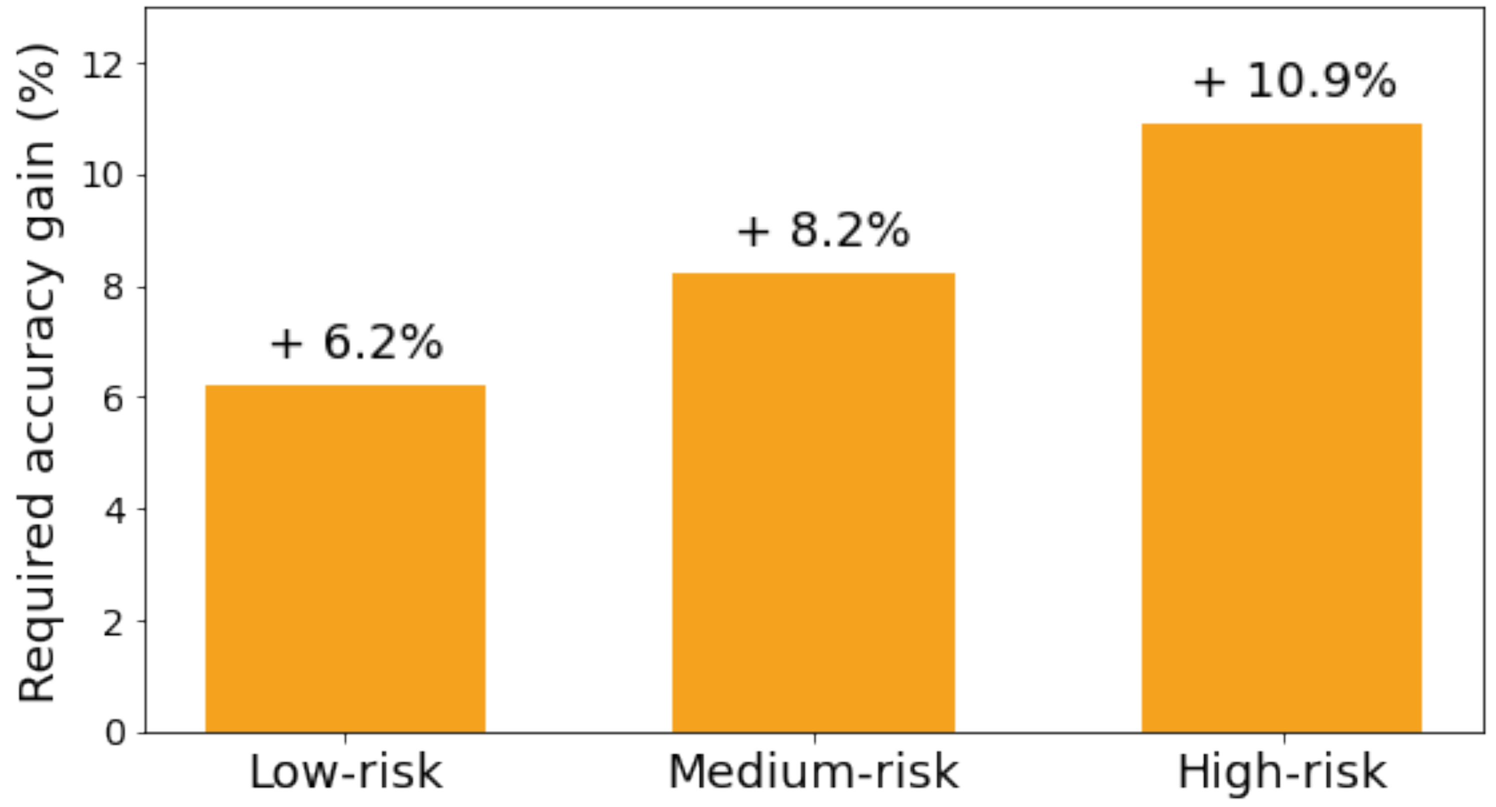
Finding #2: Agreement task reveals **confirmation bias**.

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?



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

[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

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. Interpretability by **ML researchers** → **user-oriented** interpretability
   Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández, CHI 2023.
   "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.
3. Explanations via **heatmaps** → explanations via **concepts**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky, CVPR 2023.
   Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability.
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, arXiv 2022. Interactive Visual Feature Search.*)

Understanding real AI end-users’ XAI needs, uses, and perceptions

	Prior work		Our work	
Who is studied		• **No humans**, or • **MTurkers** considering **hypothetical** AI use		**Real end-users** of an AI app
How it’s studied		• **Automated evaluation**, or • **Short experiments**		**In-depth interviews**

[Sunnie S. Y. Kim, et al., CHI 2023. “Help Me Help the AI.”]

Understanding real AI end-users' XAI needs, uses, and perceptions

Research questions

1. What are end users' XAI **needs** in real-world AI applications?
2. How do end-users **intend to use** XAI explanations?
3. How are existing XAI approaches **perceived** by end-users?

Ideal research setting

1. Real-world AI use by end-users with a diverse domain and AI knowledge base
2. Domain with significant AI and XAI research

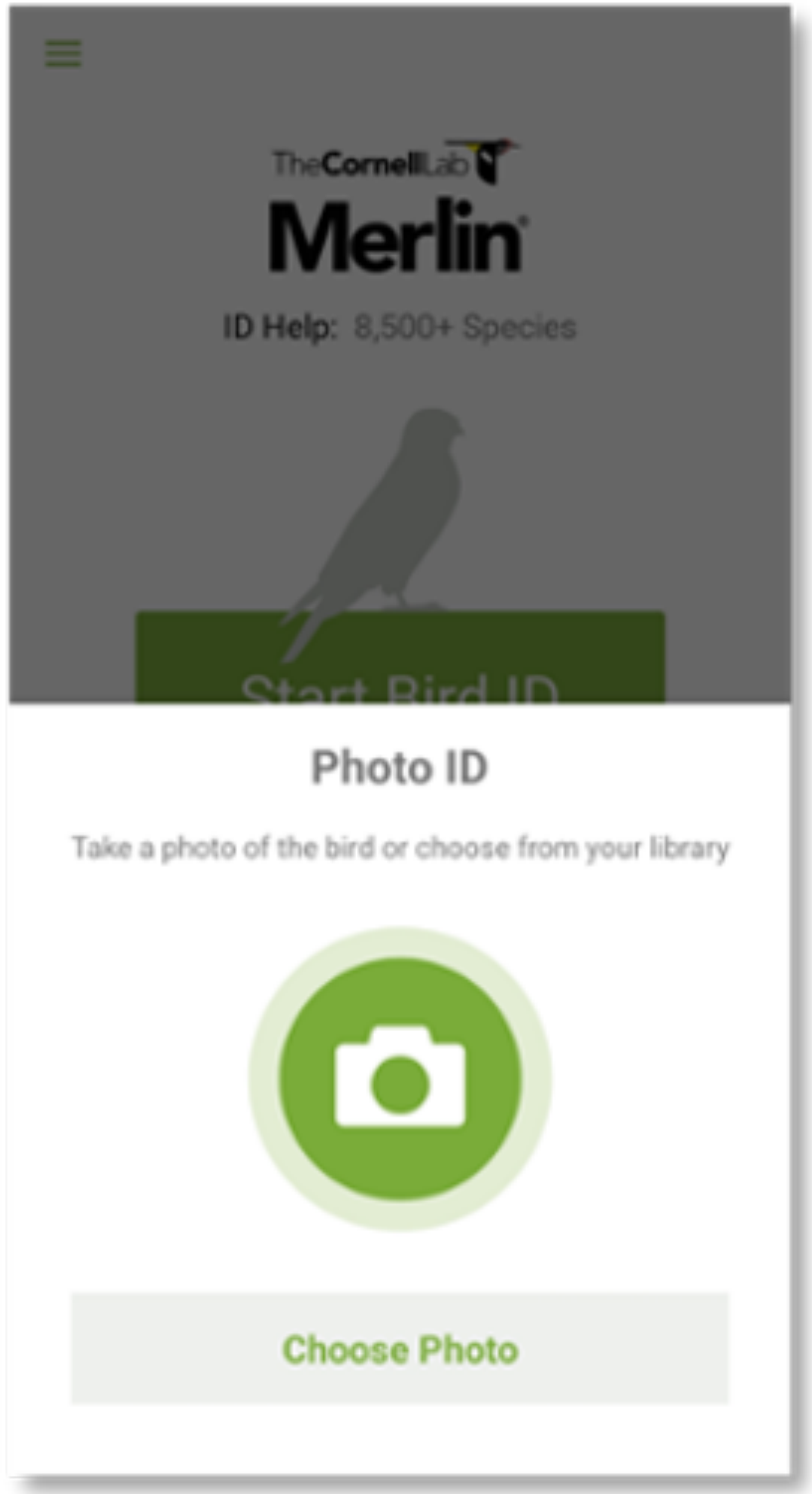
Our work

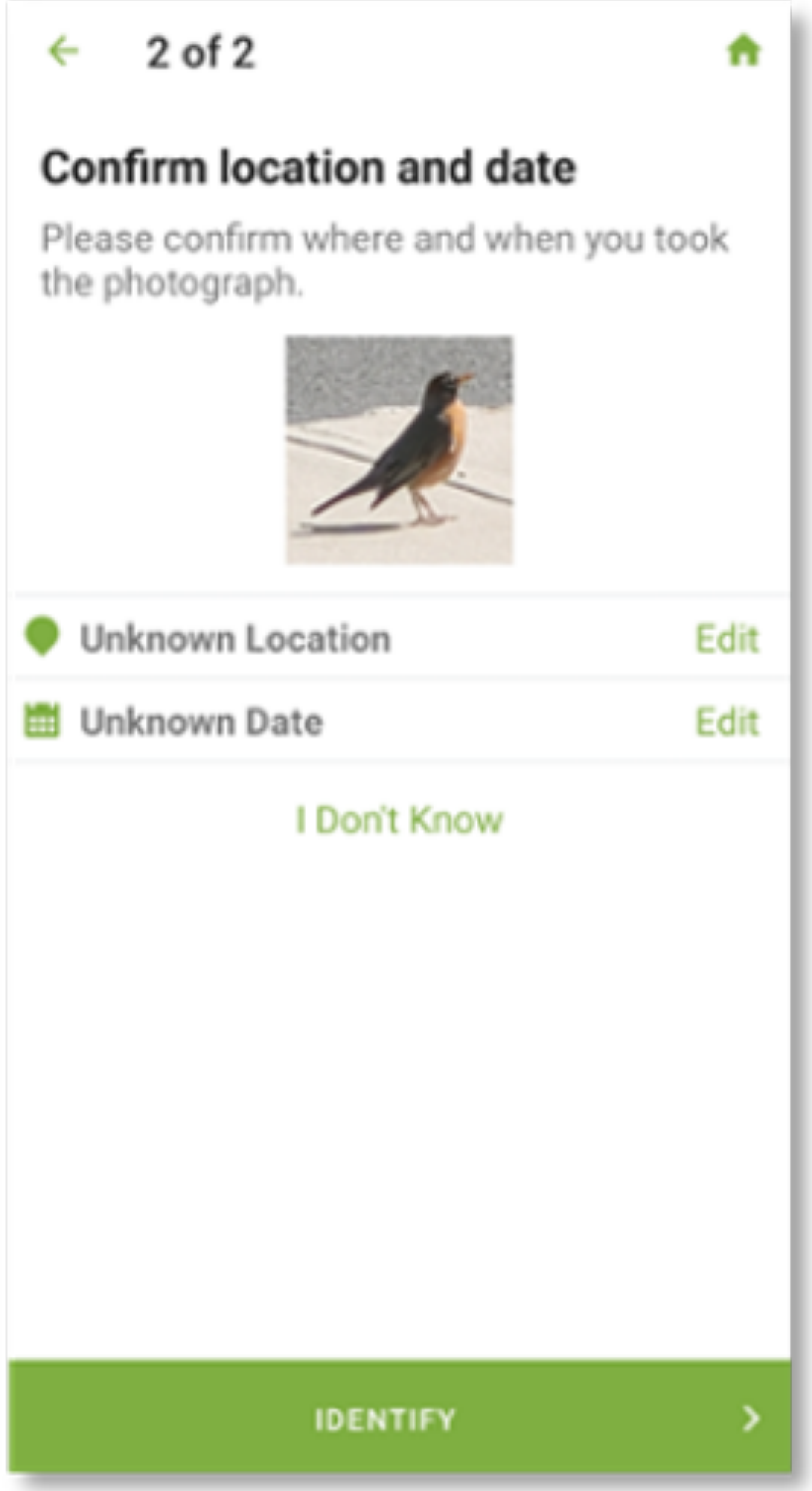
Real end-users of an AI app

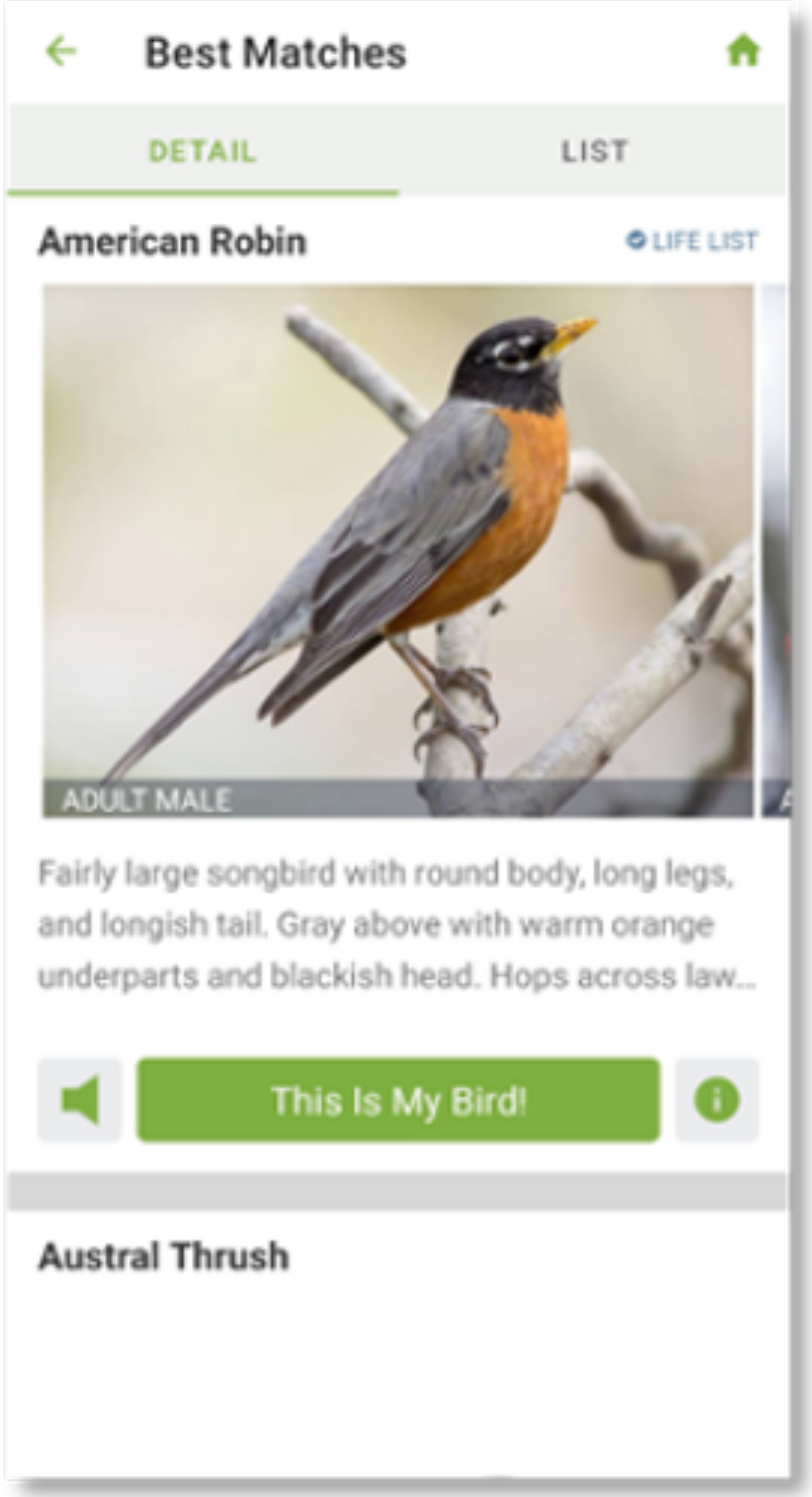
In-depth interviews

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

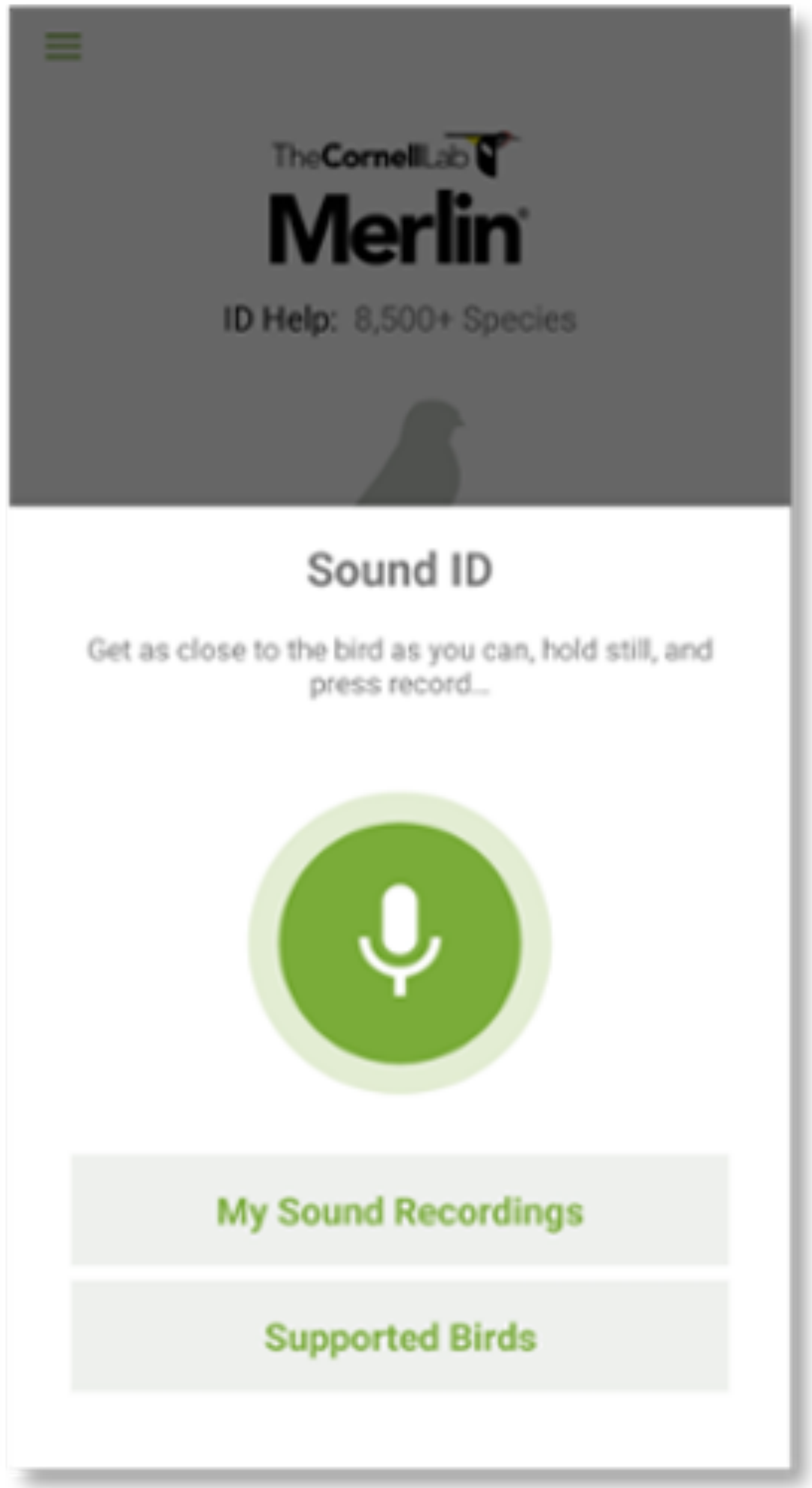
Merlin Photo ID

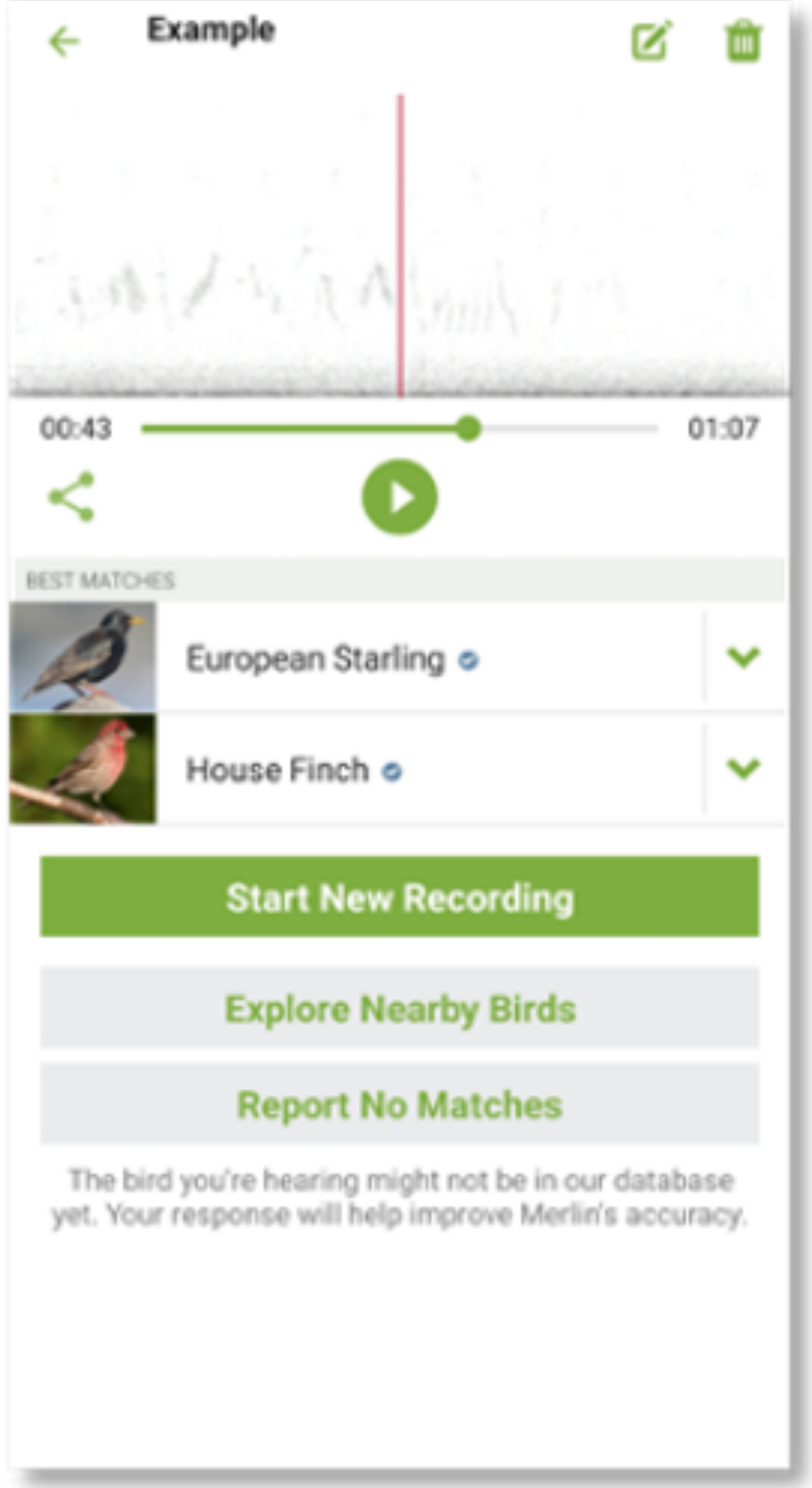


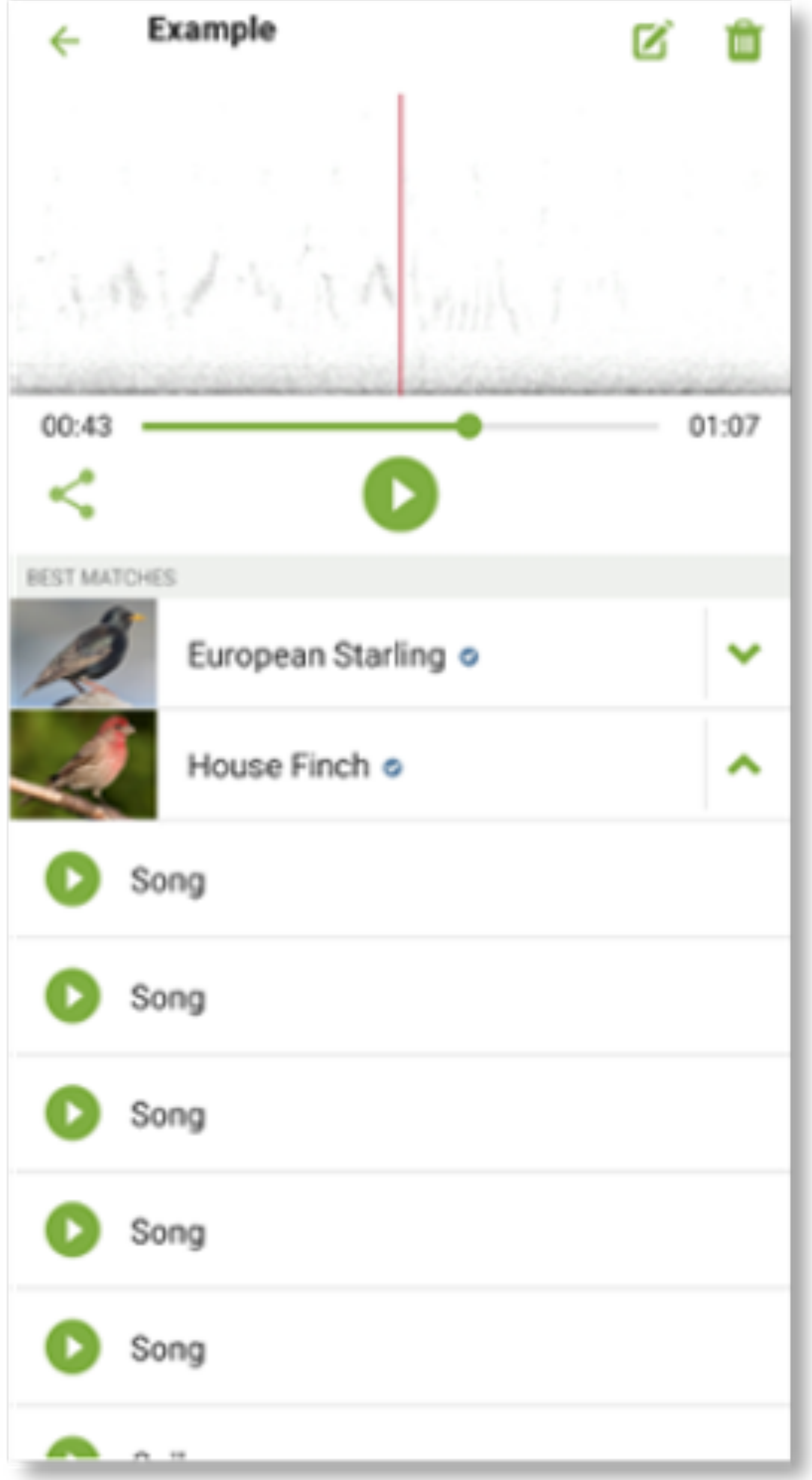


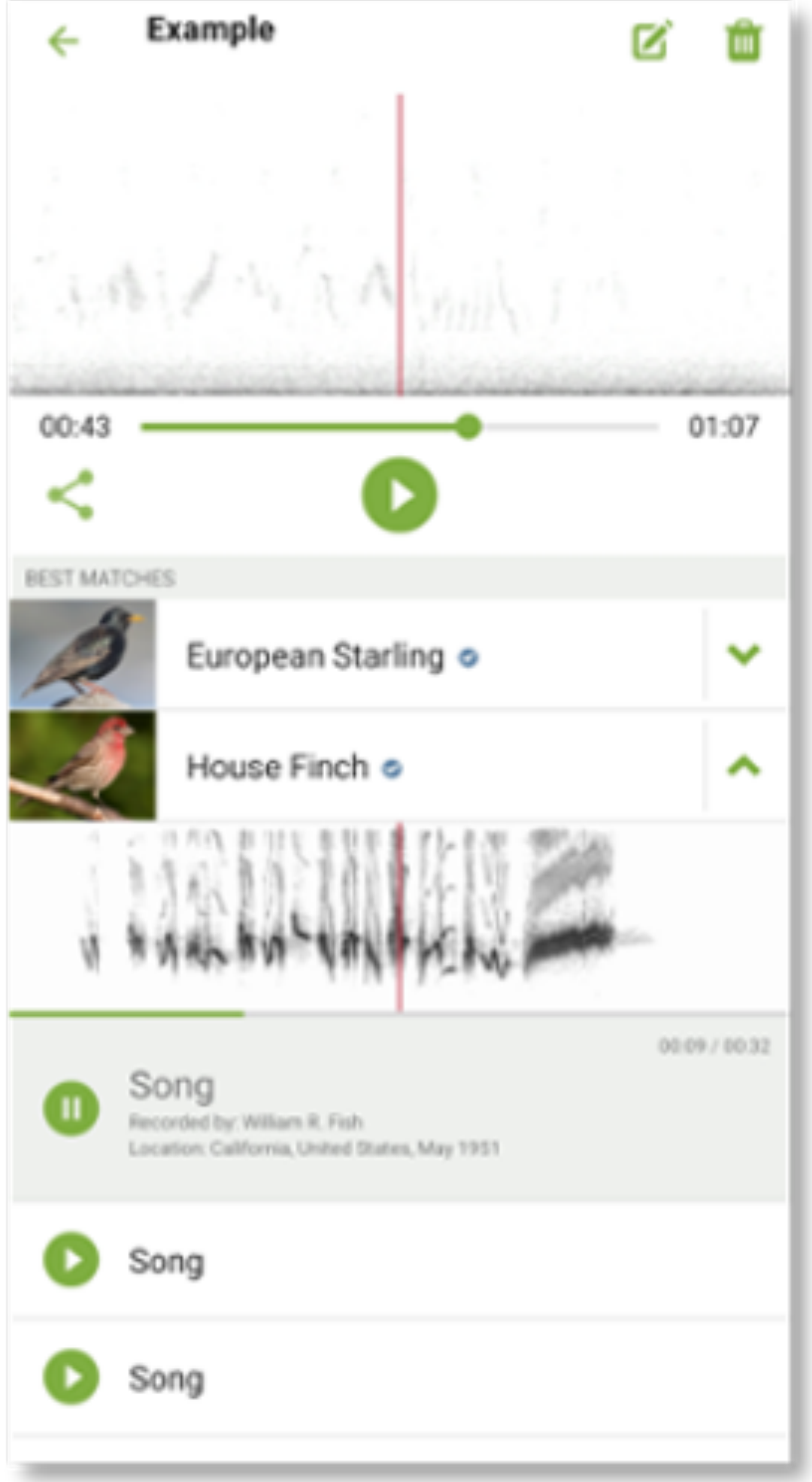


Merlin Sound ID







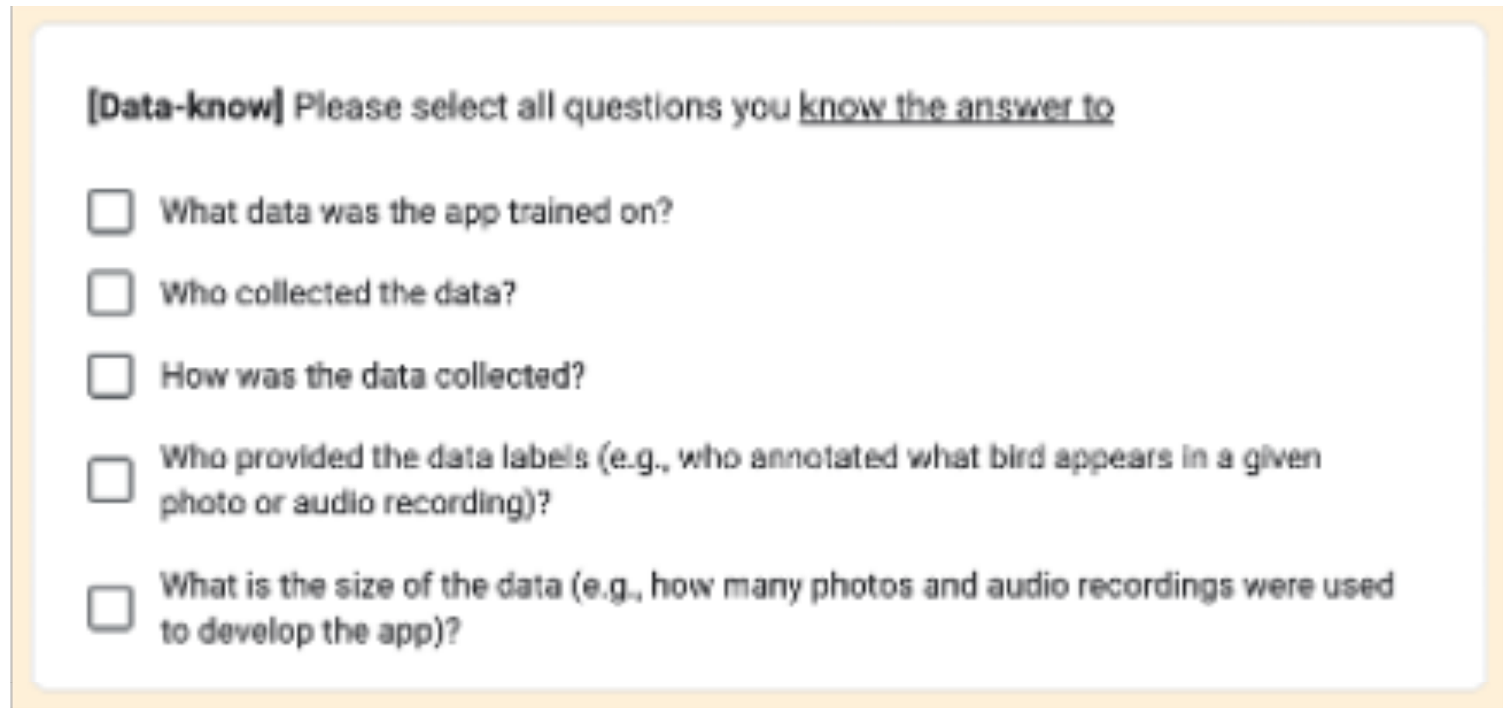


Methods

1. Recruited participants

	Low-AI	Medium-AI	High-AI
Low-domain	P7, P12, P16	P8, P14	P11, P13
Medium-domain	P2, P20	P1, P4, P10	P6
High-domain	P5, P17	P3, P9, P15	P18, P19

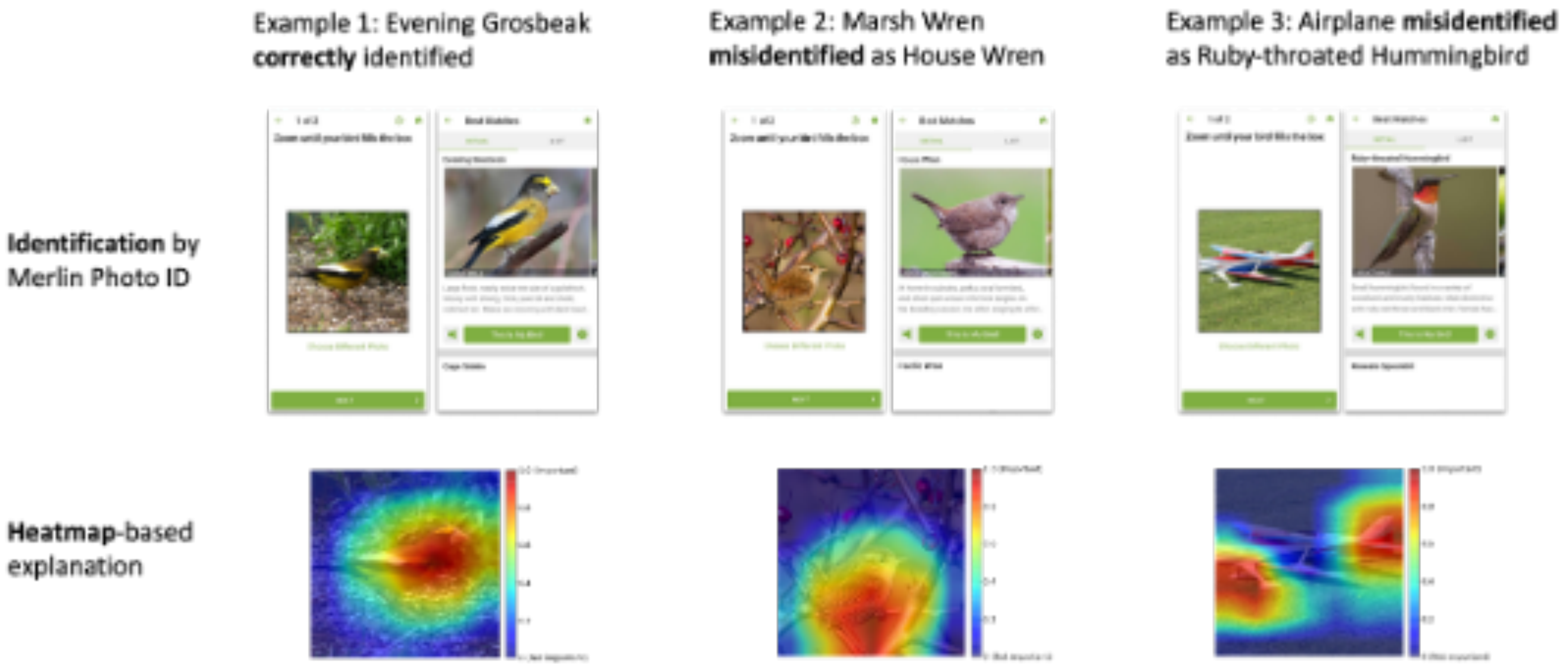
2. Conducted interviews



[Data-know] Please select all questions you know the answer to

- [ ] What data was the app trained on?
- [ ] Who collected the data?
- [ ] How was the data collected?
- [ ] Who provided the data labels (e.g., who annotated what bird appears in a given photo or audio recording)?
- [ ] What is the size of the data (e.g., how many photos and audio recordings were used to develop the app)?

3. Transcribed and analyzed interviews



[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

System details: Wanted by only AI experts and domain enthusiasts

High-AI background

"Would email the app developers and play with data/model myself"

Low-AI background

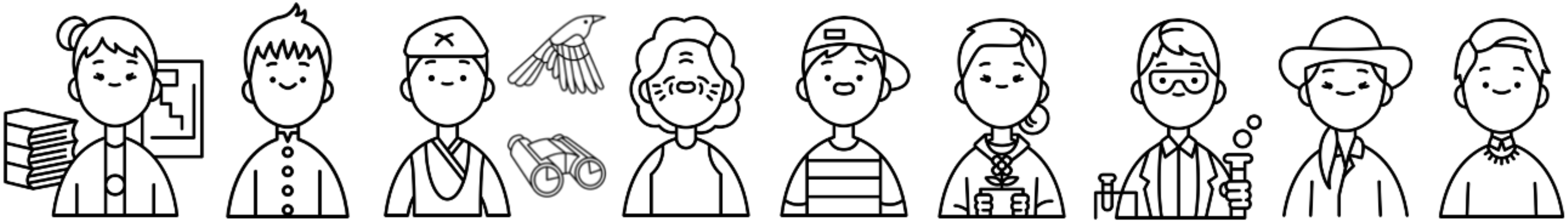
"Curious but wouldn't go out of my way"
"Don't want to ruin the mystique"

Low-AI background + High-domain interest

"Want to know how the AI distinguishes similar birds"

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Practically useful information: Wanted by everyone



"Want practically useful information that can improve collaboration with AI"

e.g. AI's capabilities and limitations, confidence, and detailed outputs

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Old and new uses of explanations

1. **Understand** the AI's outputs
2. **Calibrate trust** in the AI
3. **Learn** from the AI to perform the task better on their own
4. **Change behavior** to help the AI perform better
5. **Give feedback** to developers to improve the AI

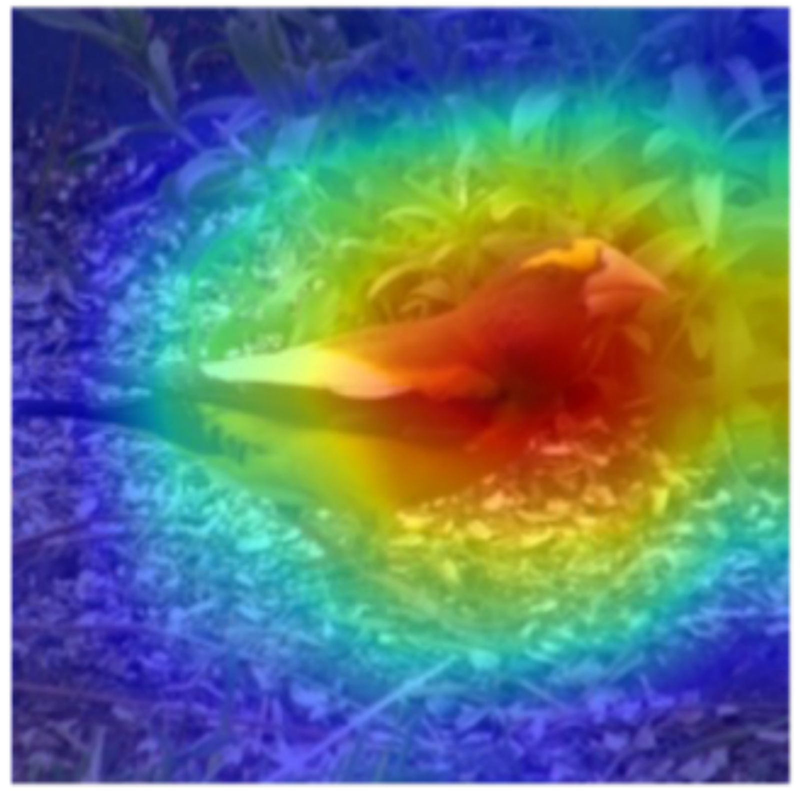
"Help Me Help the AI"

Do current XAI approaches satisfy end-users' needs and use goals?

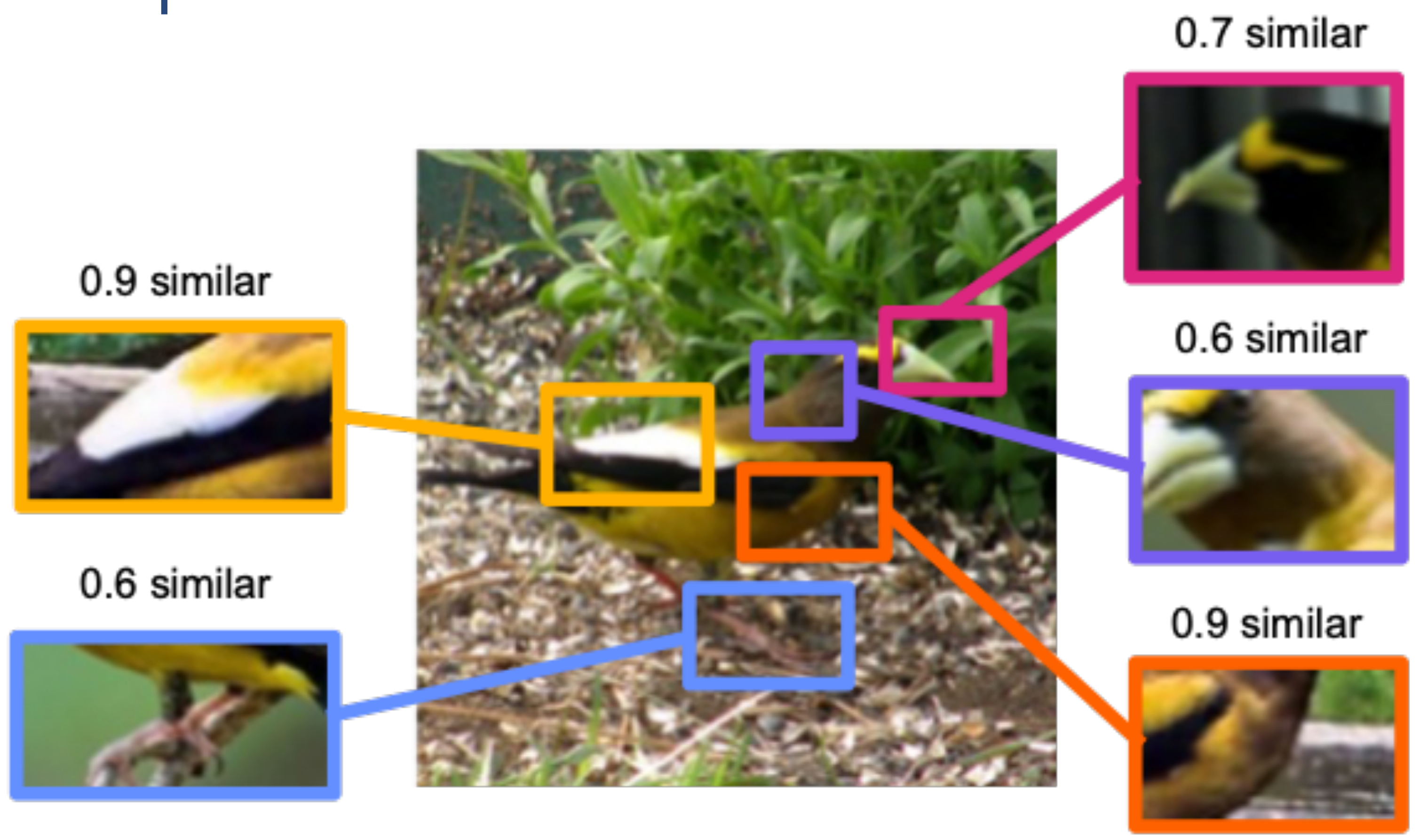
[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Perceptions of different explanation form factors

Examples



Heatmaps



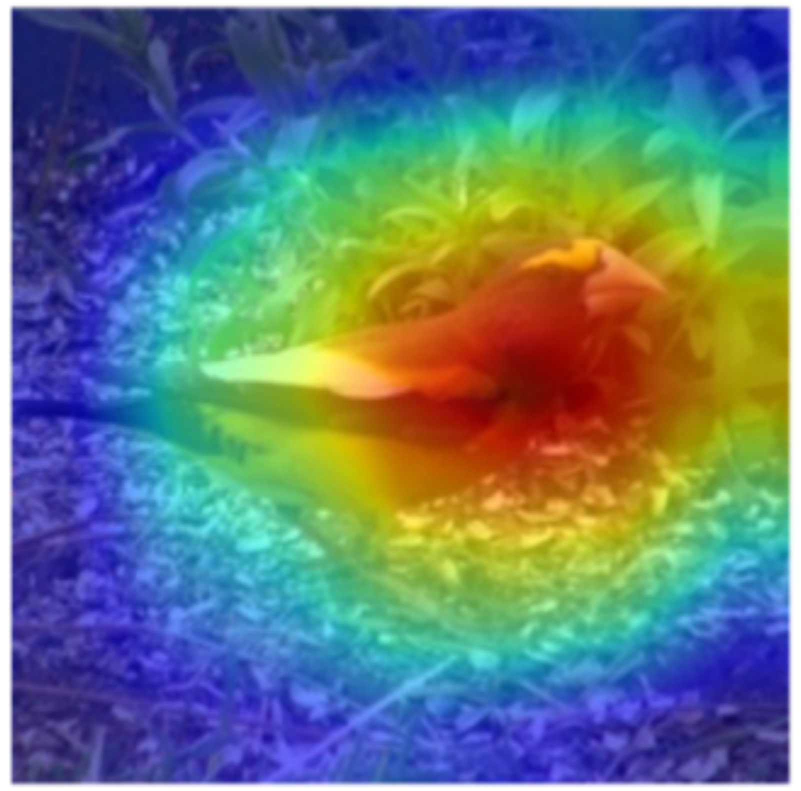
Prototypes

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

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Heatmap-based explanations



Heatmaps

- Intuitive, pleasing
- Helpful for spotting AI's mistakes

- Unintuitive, confusing
- Uninformative, too coarse
- Doesn't explain why certain parts are important
- Doesn't give actionable feedback

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Example-based explanations

Examples

- Intuitive, pleasing
- Helpful for verifying AI's outputs
- Allows end-users' moderation

- Uninformative, impression-based
- Doesn't add much to current examples in app
- Doesn't give actionable feedback

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Concept-based explanations

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~~
...

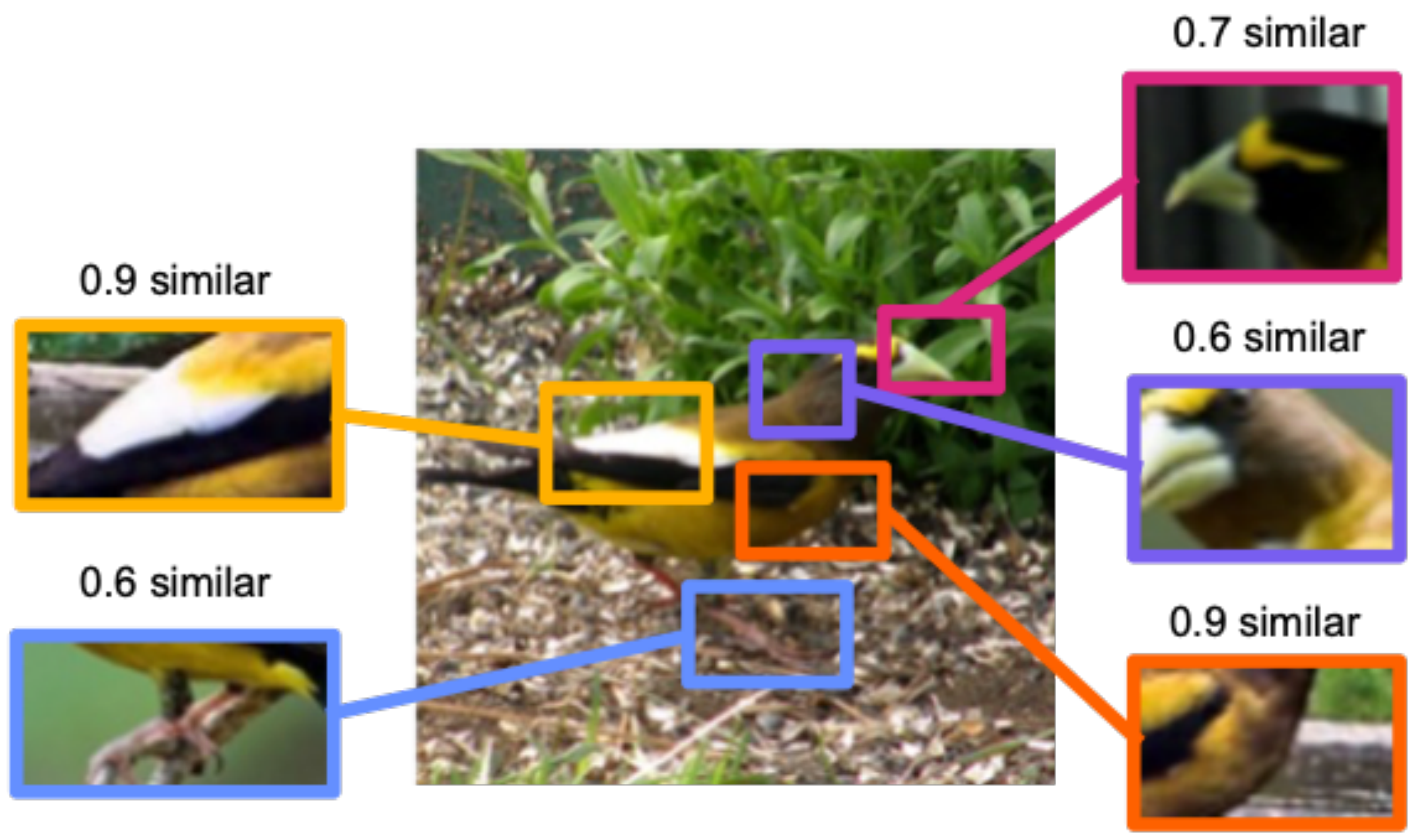
Concepts

- Parts-based form
- Resembles human reasoning and explanations
- Helpful for verifying AI's outputs
- Helpful for learning bird ID
- Numbers are helpful

- Current concepts are too generic
- Meaning of coefficients is unclear
- Numbers are overwhelming

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Prototype-based explanations



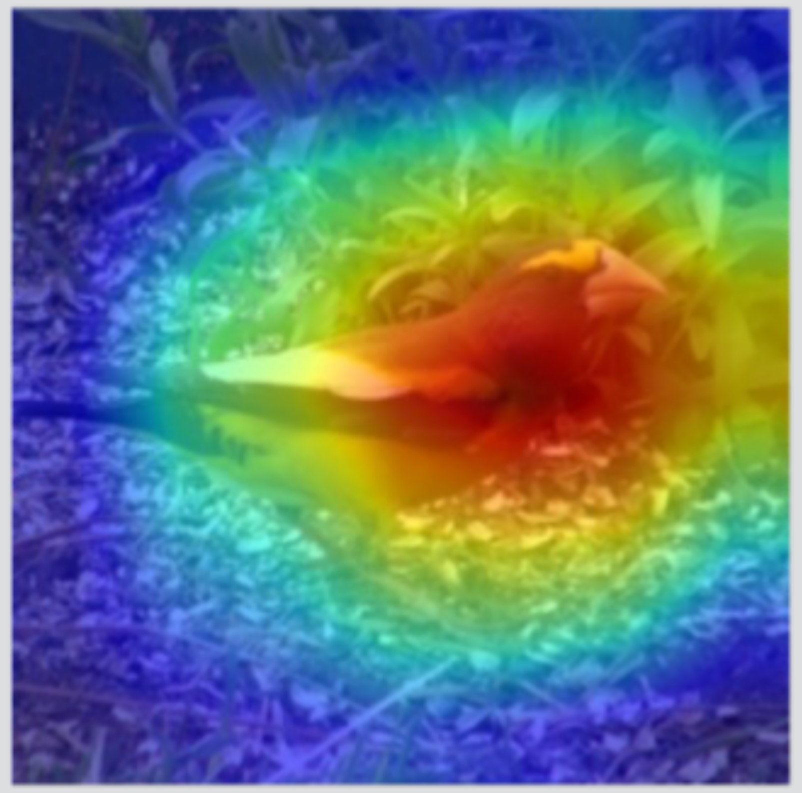
Prototypes

- Parts-based form
- Resembles human reasoning and explanations
- Intuitive, visual
- Helpful for verifying AI’s outputs
- Helpful for learning bird ID

- Cluttered
- Difficult to see on small screens
- Some prototypes are ambiguous and uninteresting

[Sunnie S. Y. Kim, et al., CHI 2023. “Help Me Help the AI.”]

XAI perceptions depend on AI background



Heatmaps

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

high-AI

"Intuitive"
"Helpful for representing info"

"Want to see more concepts and numbers"

low-AI

"Not intuitive"
"Related to weather?"

"Stuff like this would go right over my head and make no sense"

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Creator-consumer gap in XAI

	Creators High-AI	**End-users** High-AI	**End-users** Low-AI
			
XAI needs	Want AI system details	Curious 🤩	Not curious 🥱
		Want practically useful information for human-AI collaboration	
XAI uses	Understanding, Calibrating trust	Understanding, Calibrating trust, Learning from AI, Changing behavior to help AI, Giving feedback to developers	
XAI perceptions	Satisfied 👍	Satisfied 👍	Dissatisfied 👎

[Sunnie S. Y. Kim, et al., CHI 2023. "Help Me Help the AI."]

Challenges for human-centered XAI

Concerns about explanations

- Not faithful
- Difficult to digest
- Engender over-trust in AI

Takeaway: Explanations should be designed with end-users, answer “why” (not just “what”), and use multiple forms and modalities.

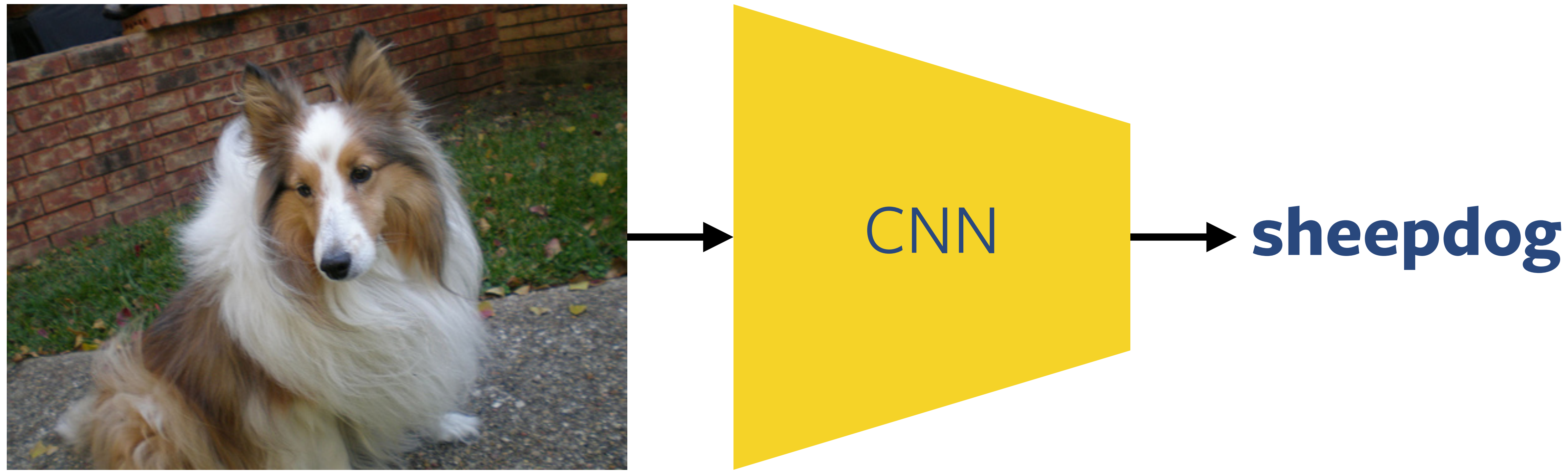
Roadmap

Vikram V. Ramaswamy

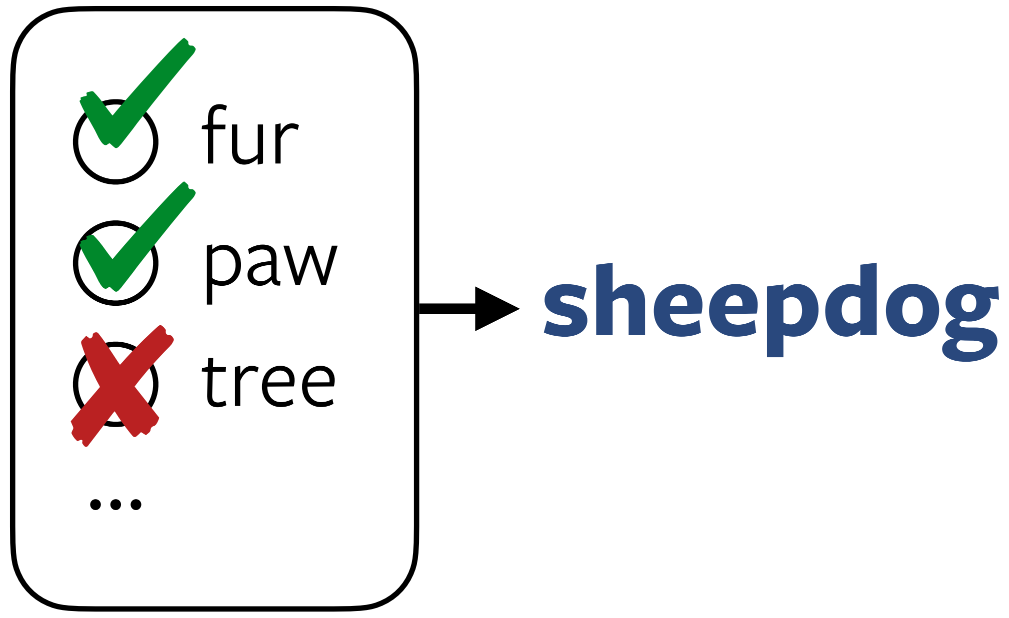
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
2. Interpretability by **ML researchers** → **user-oriented** interpretability
   Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández, CHI 2023.
   "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction.
3. Explanations via **heatmaps** → explanations via **concepts**
   Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky, CVPR 2023.
   Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability.
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, arXiv 2022. Interactive Visual Feature Search.*)

Concept-based explanations

Why did the model predict **sheepdog**?



Concept-based explanation



1.2 fur **+ 0.7** paw **- 0.6** tree =
score for **sheepdog**

Pro: Labelled concepts are interpretable to humans

Goal: Understand the effects of choices made by different concept-based explanations.

1. Effect of the **probe dataset** (i.e. dataset with labelled concepts)
2. Effect of the **concepts used** in an explanation (e.g. how easy-to-learn are concepts?)
3. Effect of **explanation complexity** (e.g. number of concepts used)

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

1. Effect of the **probe dataset**

Setup

- Model: Scene prediction classifier (Places365-trained ResNet18)
- Probe datasets: ADE20k and Pascal
  - Use all object and object-parts concepts
- Explanations: NetDissect and TCAV

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]
[Zhou et al., CVPR 2017, ADE20k; Everingham et al., IJCV 2010. Pascal; Bau* & Zhou* et al., CVPR 2017, NetDissect; Kim et al., ICML 2018, TCAV]

1. Effect of the **probe dataset**

NetDissect

- 123 neurons highly activated (i.e. used in explanations) by both datasets.
- Some correspond to similar concepts but **roughly 56%** (69 neurons) correspond to very different concepts.

Neuron	ADE20k label	ADE20k score	Pascal label	Pascal score
9	plant	0.082	potted-plant	0.194
181	plant	0.068	potted-plant	0.140
318	computer	0.079	tv	0.251
386	autobus	0.067	bus	0.200
435	runway	0.071	airplane	0.189
185	chair	0.077	horse	0.153
239	pool-table	0.069	horse	0.171
257	tent	0.042	bus	0.279
384	washer	0.043	bicycle	0.201
446	pool-table	0.193	tv	0.086

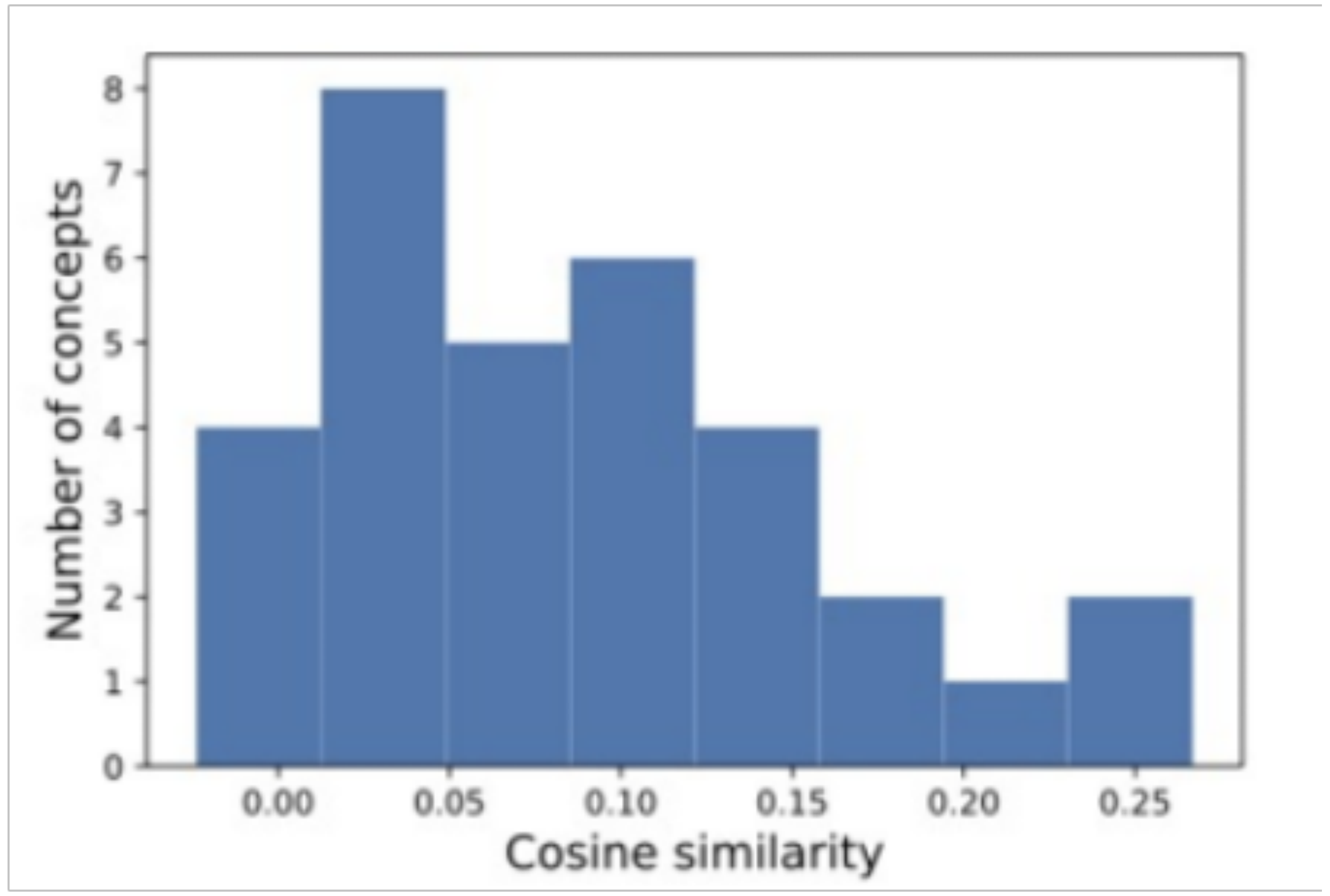
[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

1. Effect of the **probe dataset**

TCAV

- **Low cosine similarity** between TCAV vectors computed using Pascal or ADE20k.

Concept	ADE20k AUC	Pascal AUC	Cosine sim
ceiling	96.6	93.0	0.267
box	83.0	80.1	0.086
pole	89.0	79.3	0.059
bag	79.4	75.4	0.006
rock	92.6	82.8	-0.024
mean	92.0	88.1	0.087



[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

1. Effect of the **probe dataset**

Takeaway

- Probe dataset has a large impact on what explanations are generated.
- .**Suggestion:** Use probe datasets that are similar in distribution to training datasets.

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

2. Effect of the **concepts used**

Learnability of concepts

- What concepts should be labelled and used?.
- .**Assumption:** All concepts used in explanations are **easier to learn** than the target classes.
- Why does this matter?
  - Suppose we explain “bedroom” with “bed”.
  - We expect the model to first learn the concept “bed” and use it to predict the class “bedroom”.
  - But, this isn’t possible if “bed” is harder to learn than “bedroom”.

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

2. Effect of the **concepts used**

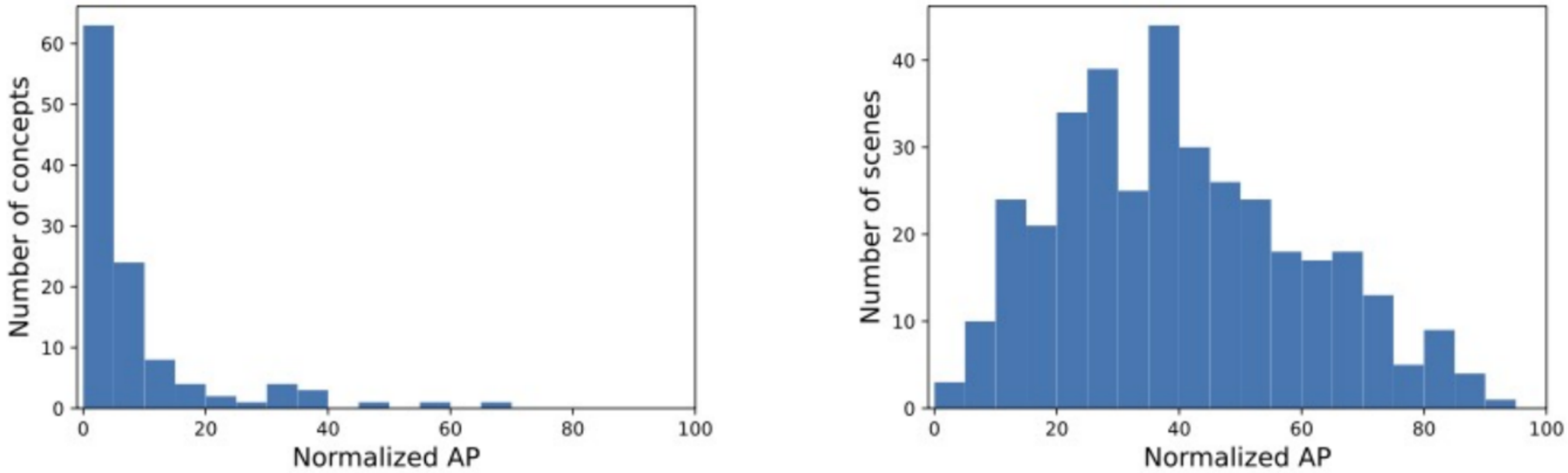
Setup

- Task datasets: Places365 (scenes) and CUB (birds).
- Probe datasets: Broden (textures, parts, objects, etc.) and CUB (bird attributes).
- Goal: Study how learnable concepts are to the target classes.
- Method: Measure learnability by training a linear classifier to predict concepts using features from pre-trained models and compare to blackbox model for target classes.
- Metric: Normalized AP (to compare across different base rates)

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]
[Bau* & Zhou*, CVPR 2017, Net Dissect; Wang et al., 2011. CUB.]

2. Effect of the **concepts used**

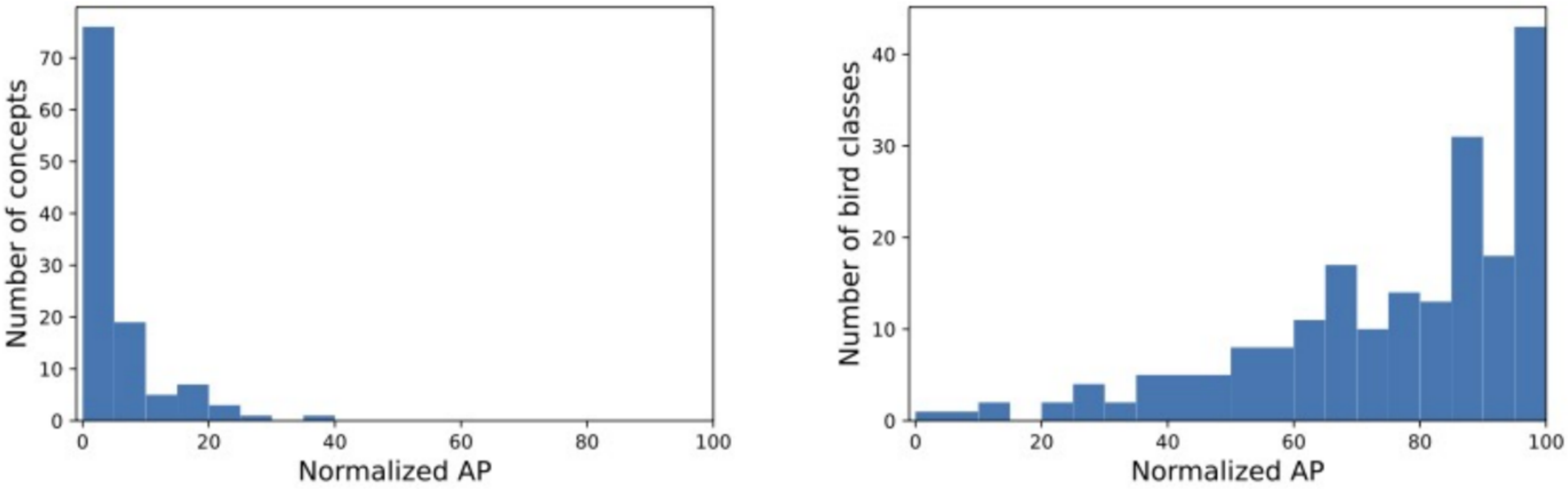
Learnability of Broden concepts vs. Places365 scenes



[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

2. Effect of the **concepts used**

Learnability of CUB concepts vs. CUB classes



[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

2. Effect of the **concepts used**

Learnability of Broden concepts for scene explanations (*red italics: scene is easier than concept*).

Scene	Concepts				
arena/perform 38.8	tennis court 74.0	grandstand 44.4	ice rink 40.7	*valley 19.0*	*stage 11.9*
art-gallery 27.4	binder 42.6	*drawing 10.8*	*painting 10.5*	*frame 2.5*	*sculpture 0.7*
bathroom 43.3	*toilet 39.9*	*shower 18.8*	*countertop 12.6*	*bathtub 11.1*	*screen door 9.6*
kasbah 50.2	ruins 64.3	*desert 17.3*	*arch 16.2*	*dirt track 8.9*	*bottle rack 4.2*
kitchen 33.9	*work surface 24.8*	*stove 18.2*	*cabinet 10.3*	*refrigerator 8.8*	*doorframe 2.8*
lock-chamber 36.5	water wheel 47.4	dam 43.7	*boat 16.1*	*embankment 4.8*	*footbridge 4.1*
pasture 19.2	cow 63.7	leaf 21.1	*valley 19.0*	*field 6.8*	*slope 4.1*

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

2. Effect of the **concepts used**

Takeaway

- Classes are often being explained using hard-to-learn concepts.
- Suggests that explanations are not **causal**.
- **Suggestion:**
  - **Simple fix:** Use only easy-to-learn concepts..
  - **But… not enough: why** are these methods learning non-causal explanations?**.**

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

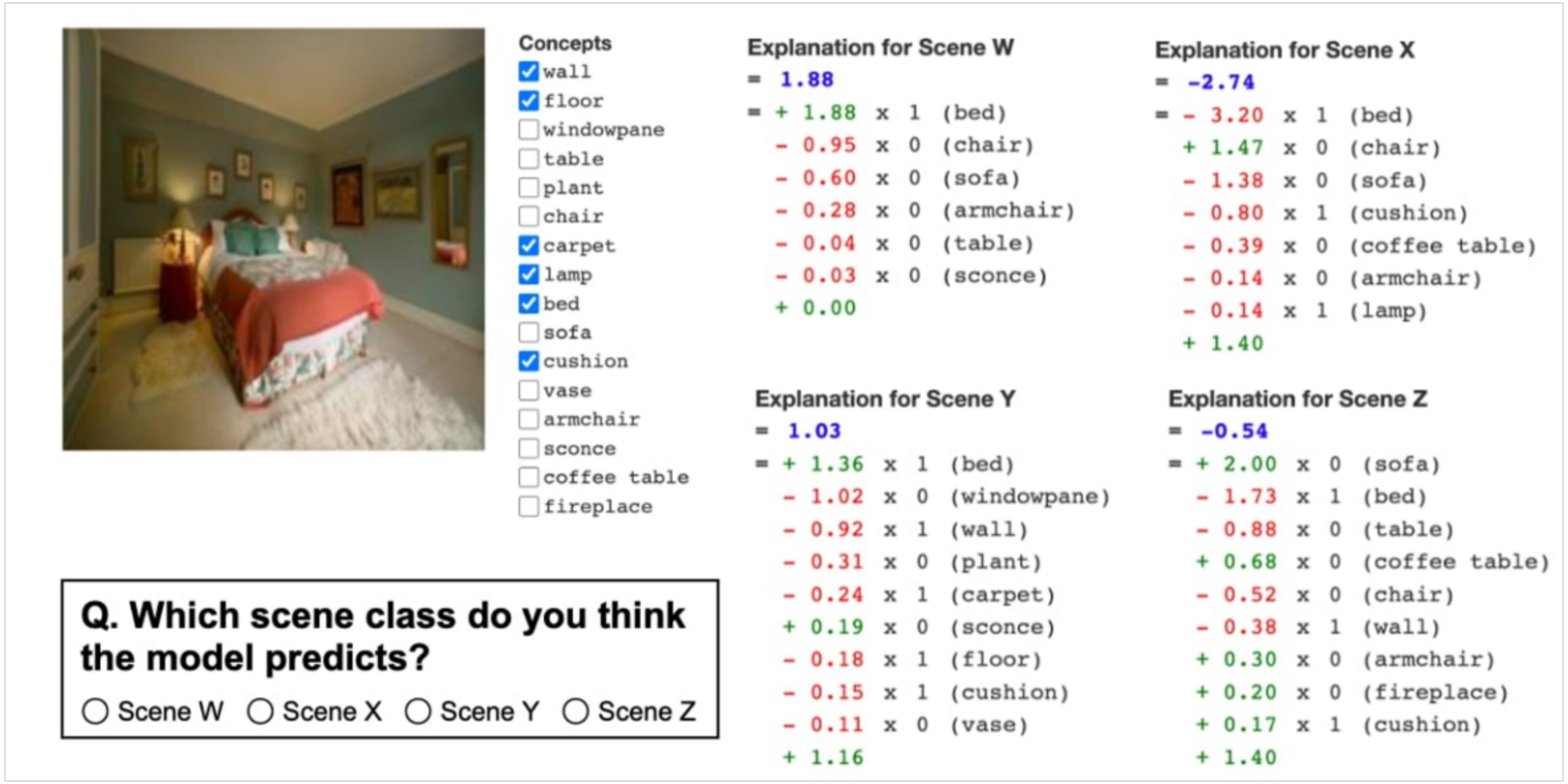
Research questions

- Can humans actually parse explanations?
- Current approaches use as many concepts as available: is this useful for humans?
- Goal: Understand if humans...
  - Can recognize concepts and predict scenes that the model would.
  - Reason about trade-offs between complexity of explanation and the “correctness” of an explanation.

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

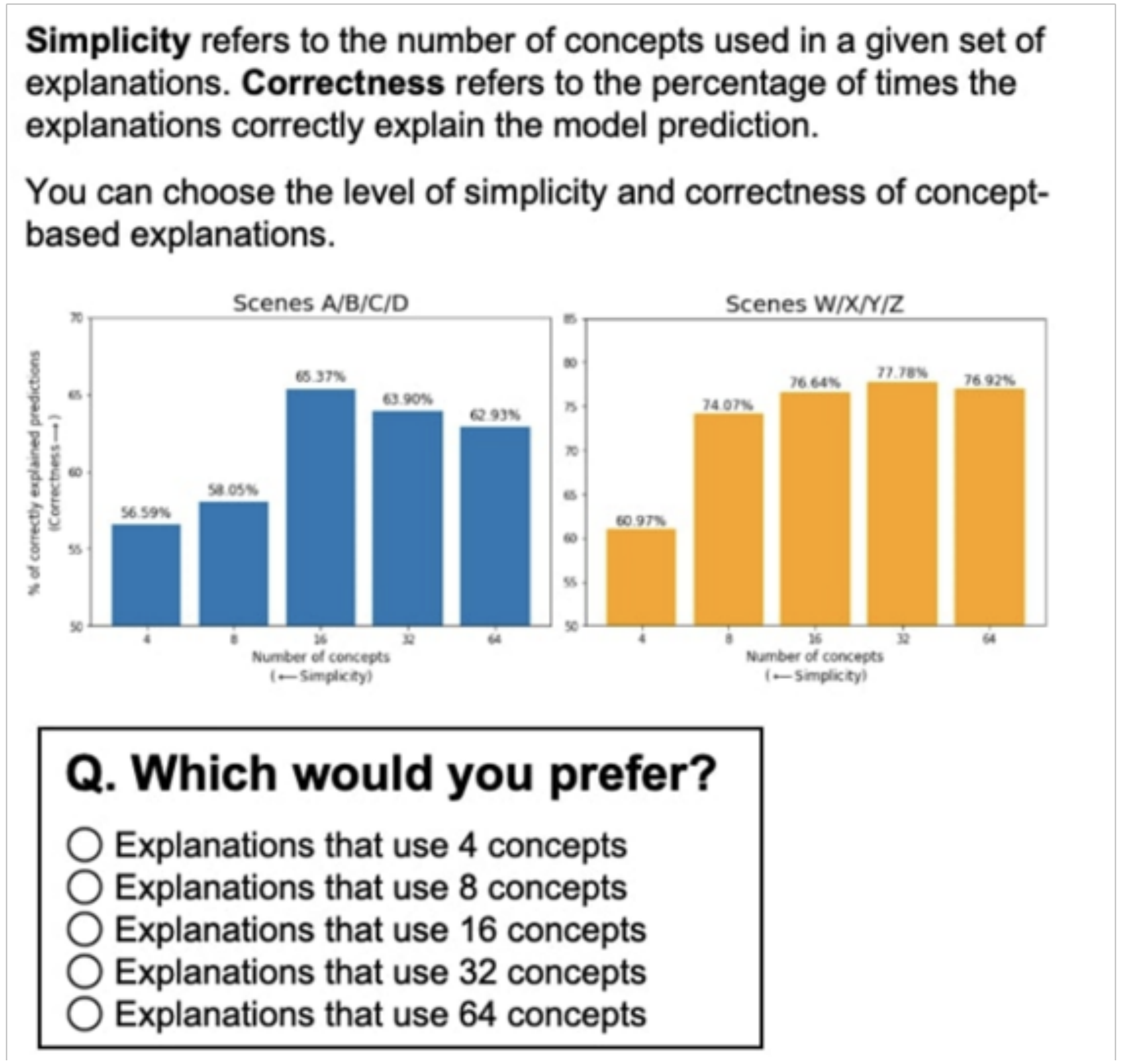
Task 1: Simulate model with explanations



[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

Task 2: Pick complexity of explanation

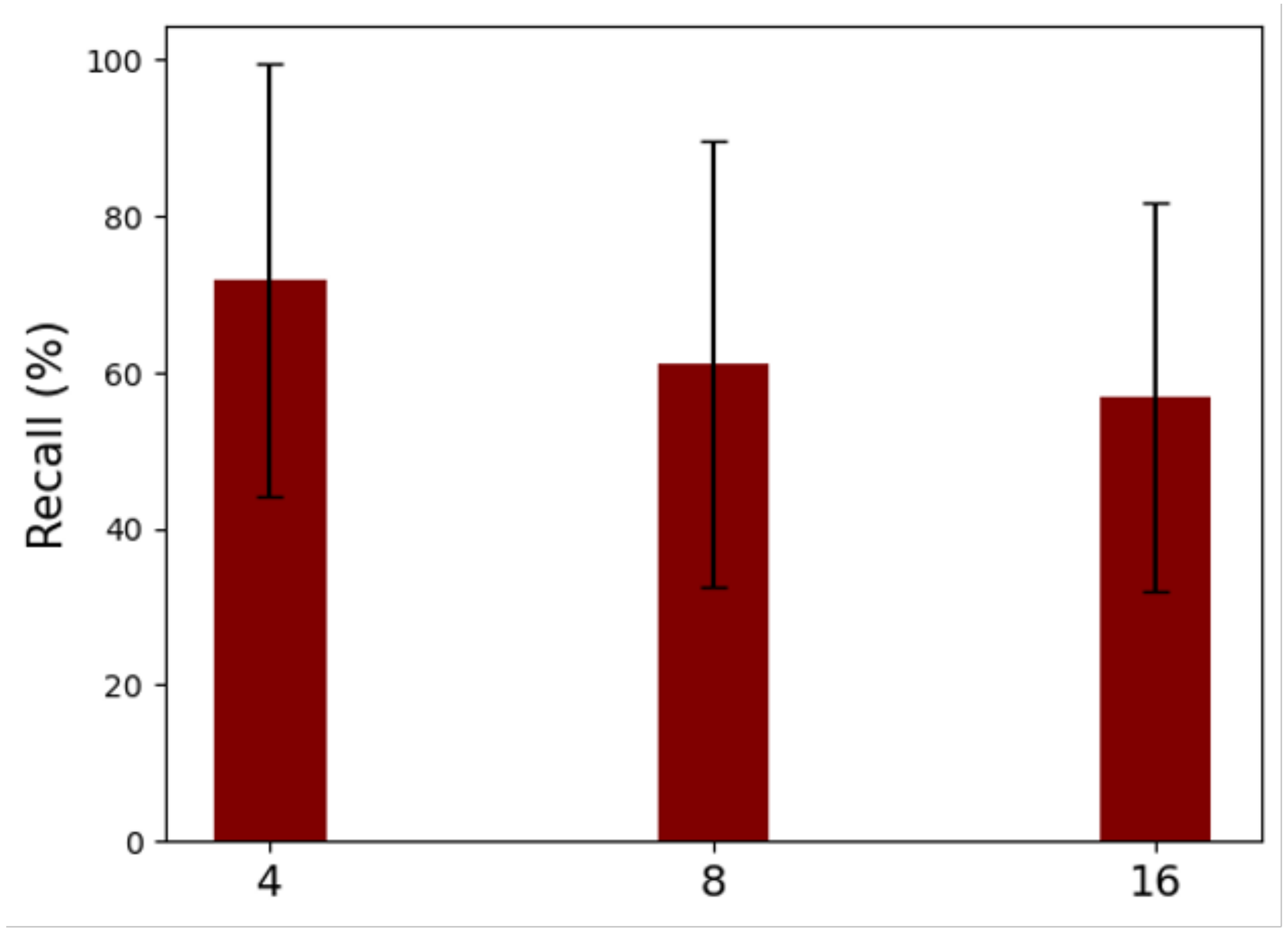


[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

Task 1: Simulate model with explanations

- When presented with more concepts, participants spend **more time on the task** but are **worse at recognizing concepts**.

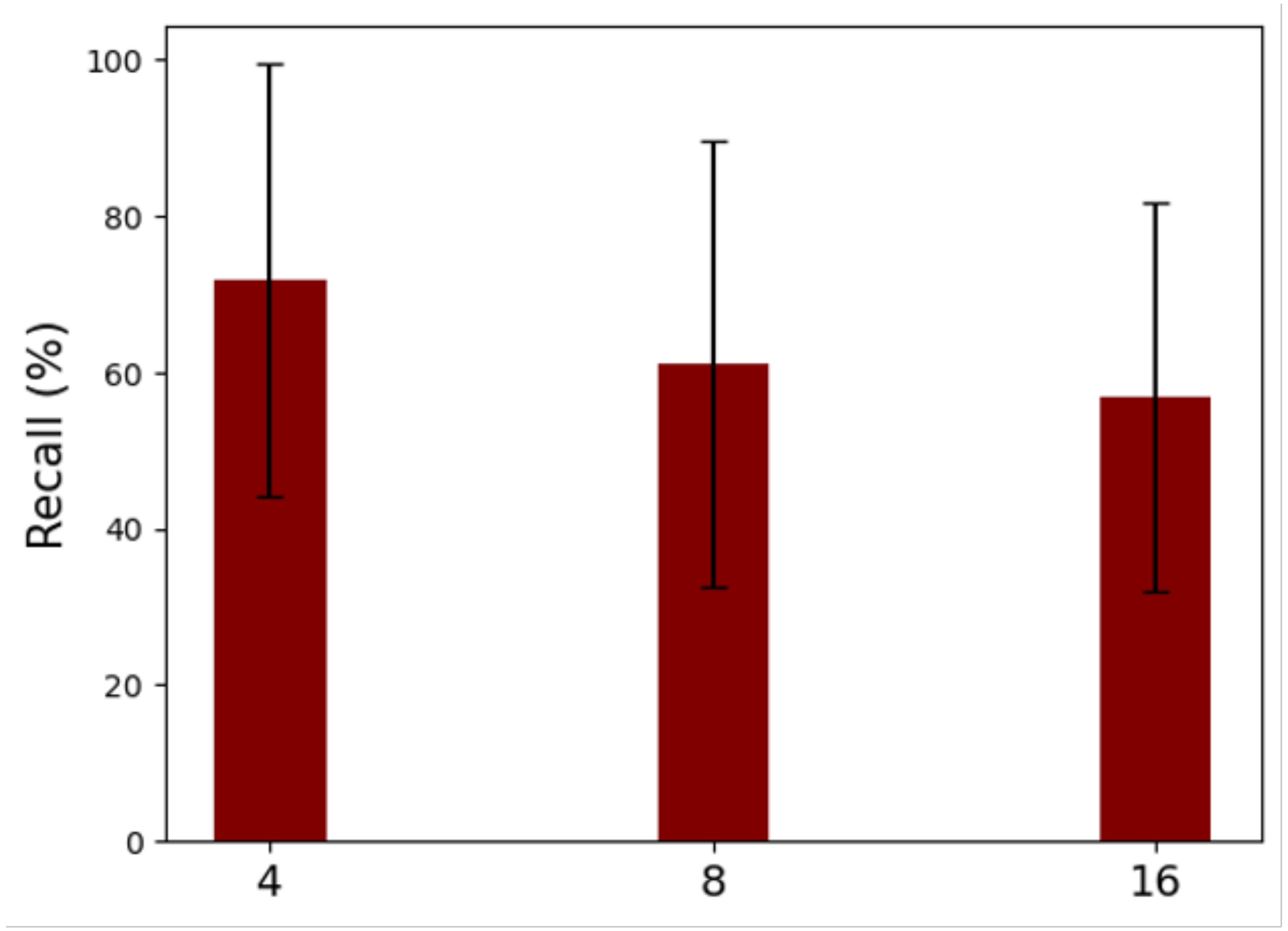


[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

Results:

- Task 1: When presented with more concepts, participants spend **more time on the task** but are **worse at recognizing concepts**.
- Task 2: Majority of participants prefer explanations with **≤ 32 concepts**.



[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

3. Effect of the **explanation complexity**

Takeaway

- Should consider the complexity of explanations and what users need from the explanation.
- Suggestion: Limit number of concepts within explanation.

[Vikram V. Ramaswamy, et al., CVPR 2023. Overlooked Factors.]

Challenges for concept-based methods

- Explanations are highly dependent on choice of probe datasets.
- Explanations often are composed of concepts that are harder-to-learn than target classes being explained.
- Humans have limited capacity for digesting complex explanations.

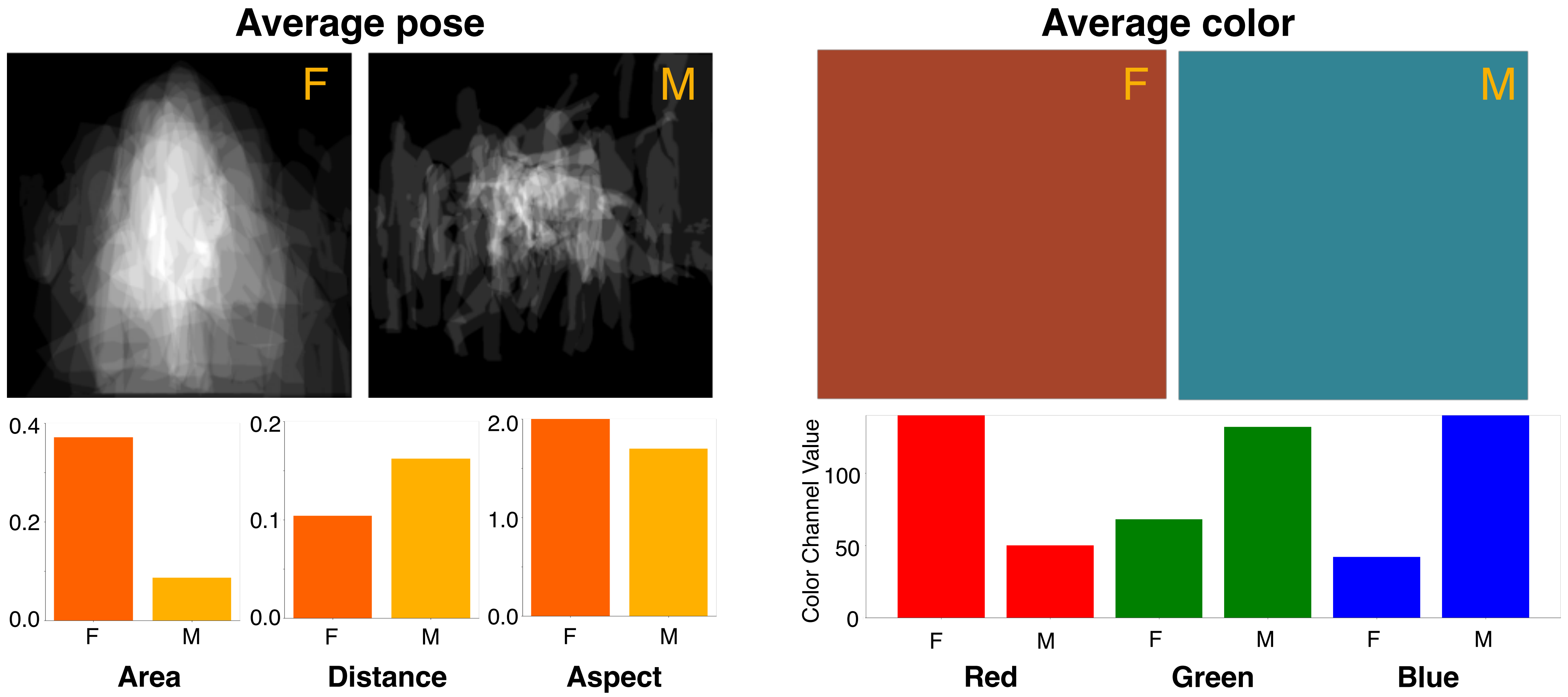
Takeaway: Be realistic about the limitations of concept-based methods (e.g. probe dataset, concept learnability, and explanation complexity) and work towards addressing the limitations.

Roadmap

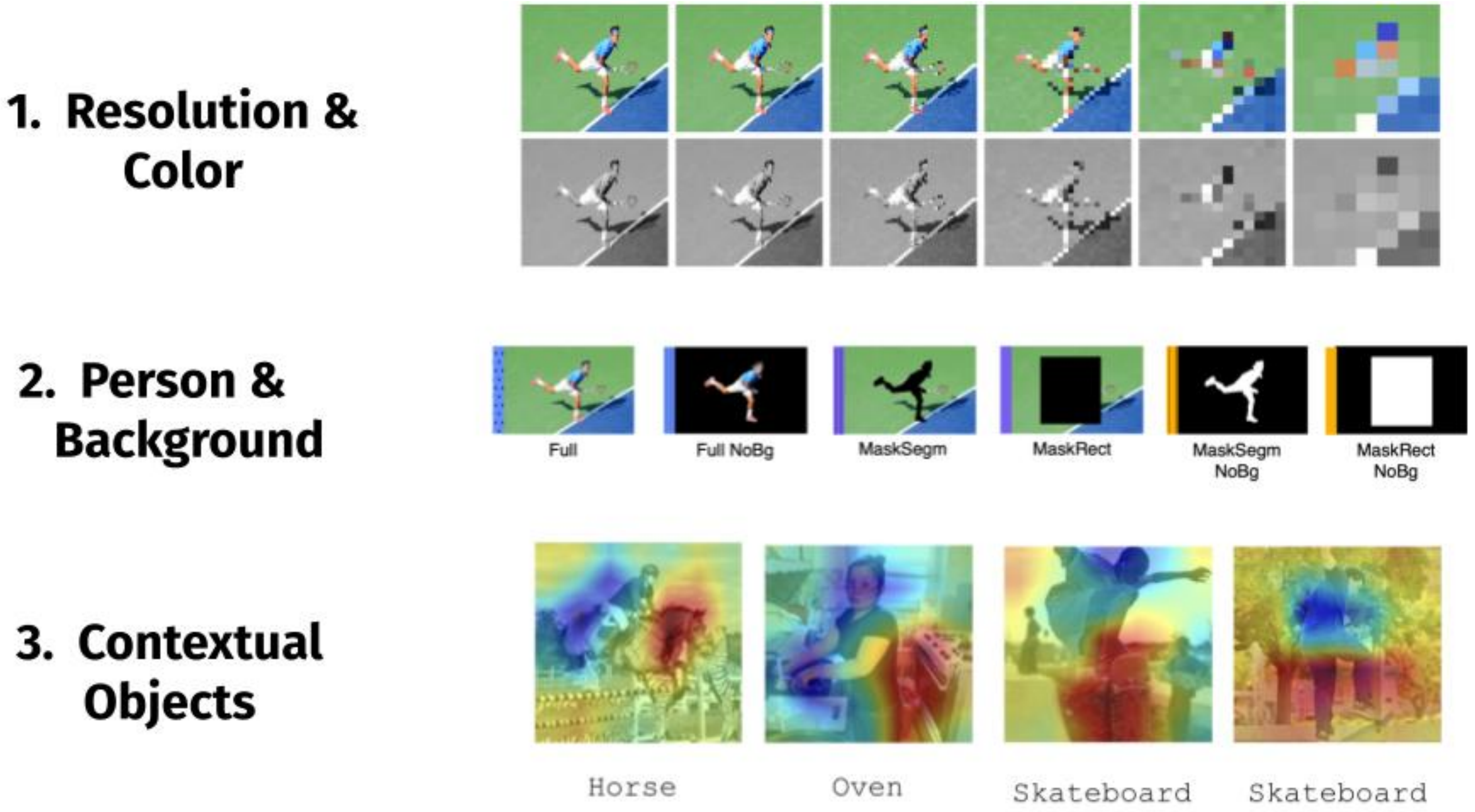
1. **Automated** evaluation of interpretability → **human-centered** evaluation
   Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, ECCV 2022.
   HIVE: Evaluating the Human Interpretability of Visual Explanations.
   (+ *Sunnie S. Y. Kim et al., arXiv 2022. "Help Me 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.*)

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.

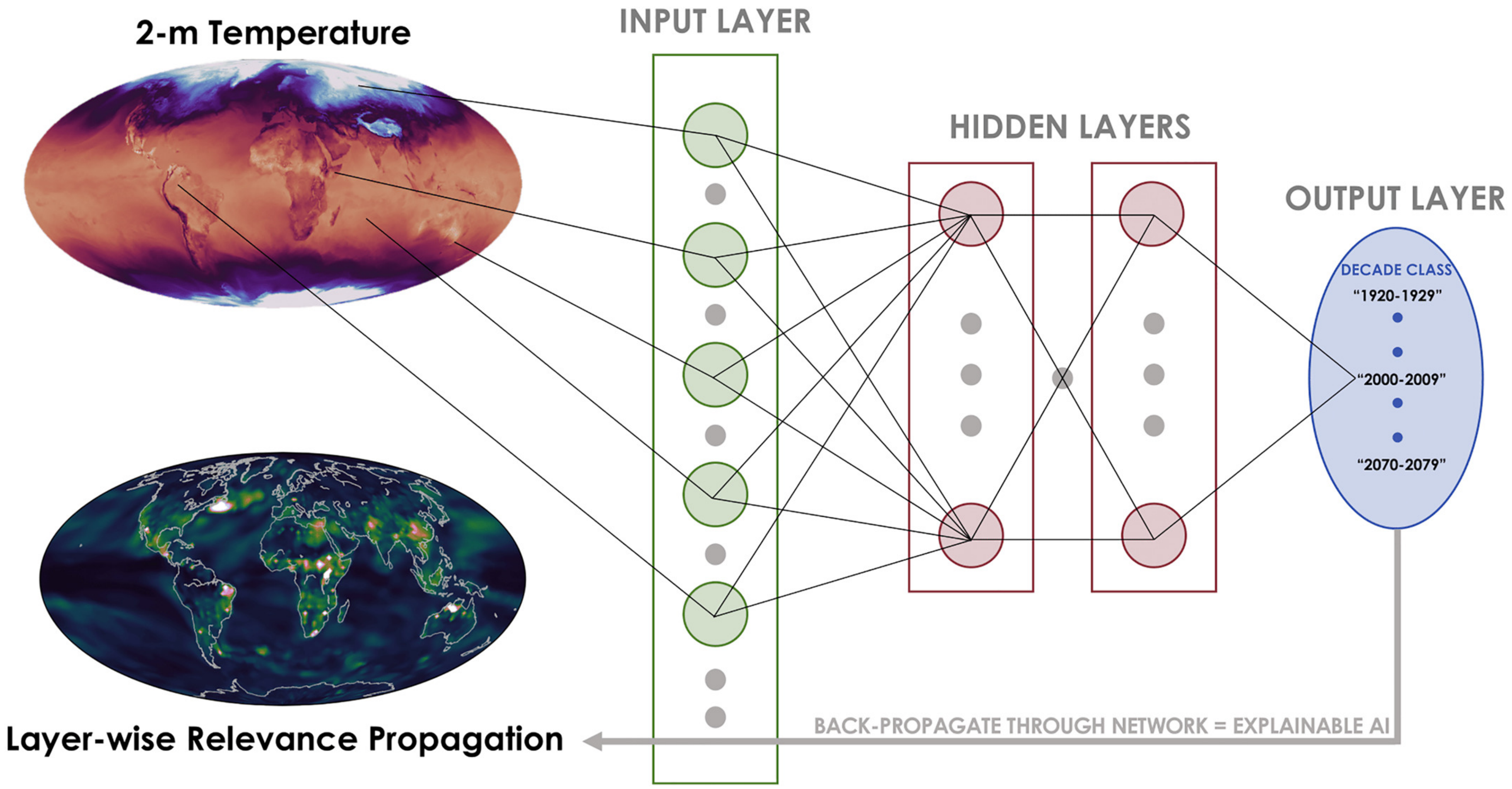
(** binary perceived gender expression; we do not condone gender prediction.*)

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)

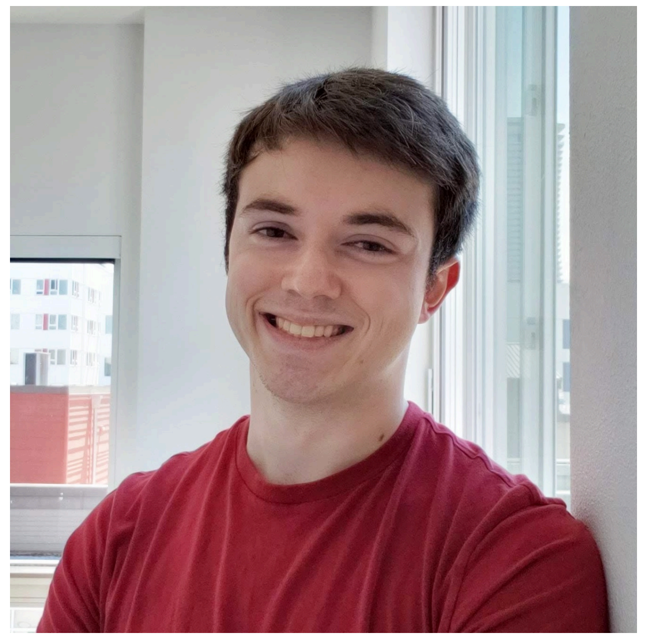
Indu Panigrahi et al., arXiv 2022. Improving Fine-Grain Segmentation via Interpretable Modifications.
Zachary M. Labe and Elizabeth A. Barnes, JAMES 2021. Detecting Climate Signals Using Explainable AI.

Interactive Similarity Overlays

bit.ly/interactive_overlay

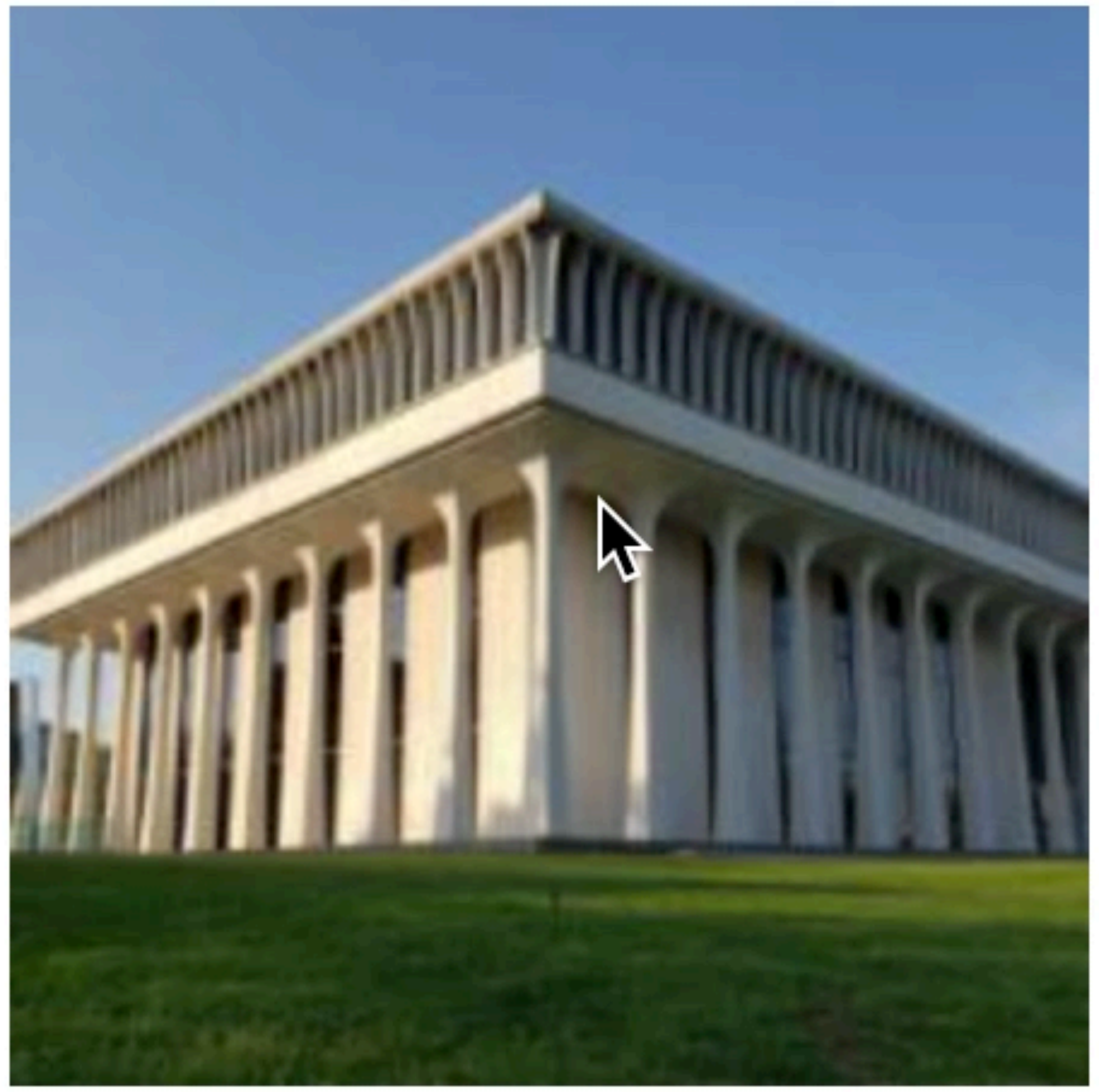
Ruth Fong, Alexander Mordvintsev, Andrea Vedaldi, Chris Olah, VISxAI 2021.
Interactive Similarity Overlays.

Preview: Interactive Visual Feature Search



bit.ly/interactive_search

Devon Ulrich



Devon Ulrich and Ruth Fong, arXiv 2022.
Interactive Visual Feature Search.
Acknowledgement: David Bau

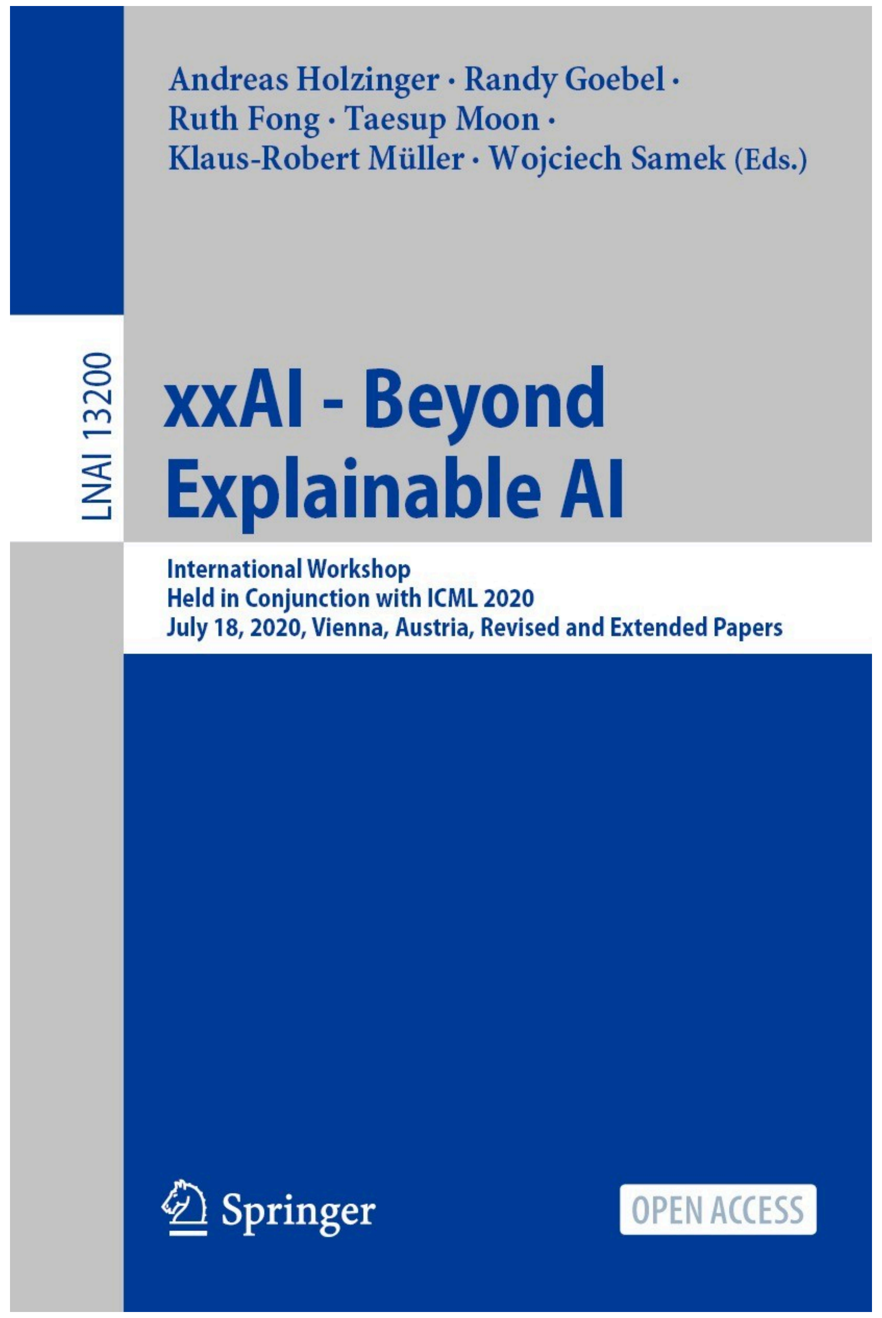
Takeaways from challenges in interpretability

- **Human studies:** As a research community, invest in and reward human evaluation studies (like dataset development).
- **Human-centered XAI:** Explanations should be designed with end-users, answer “why” (not just “what”), and use multiple forms and modalities.
- **Concept-based explanations:** Be realistic about the limitations of concept-based methods (e.g. probe dataset, concept learnability, and explanation complexity) and work towards addressing the limitations.

Directions for the next decade of interpretability

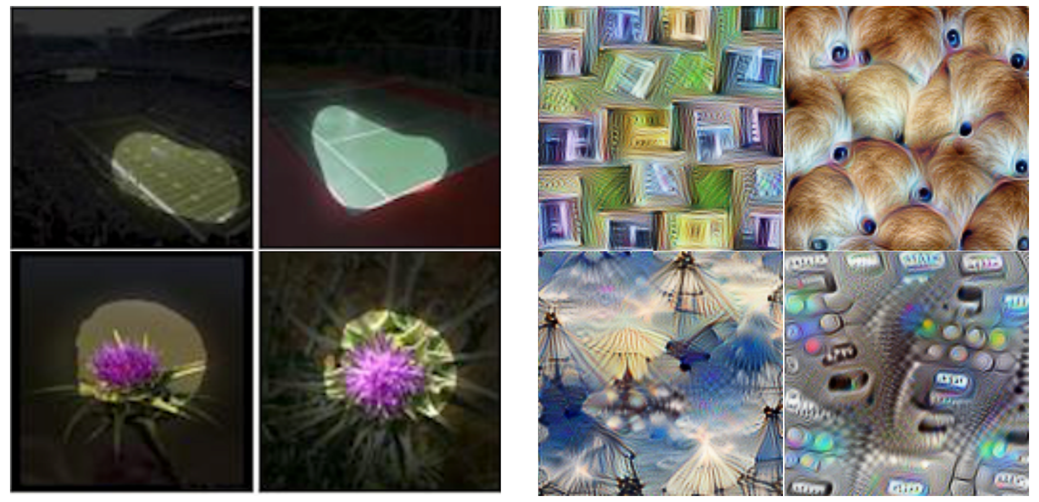
1. Develop interpretability methods for **diverse domains**
    - Beyond CNN classifiers: self-supervised learning, generative models, etc.

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

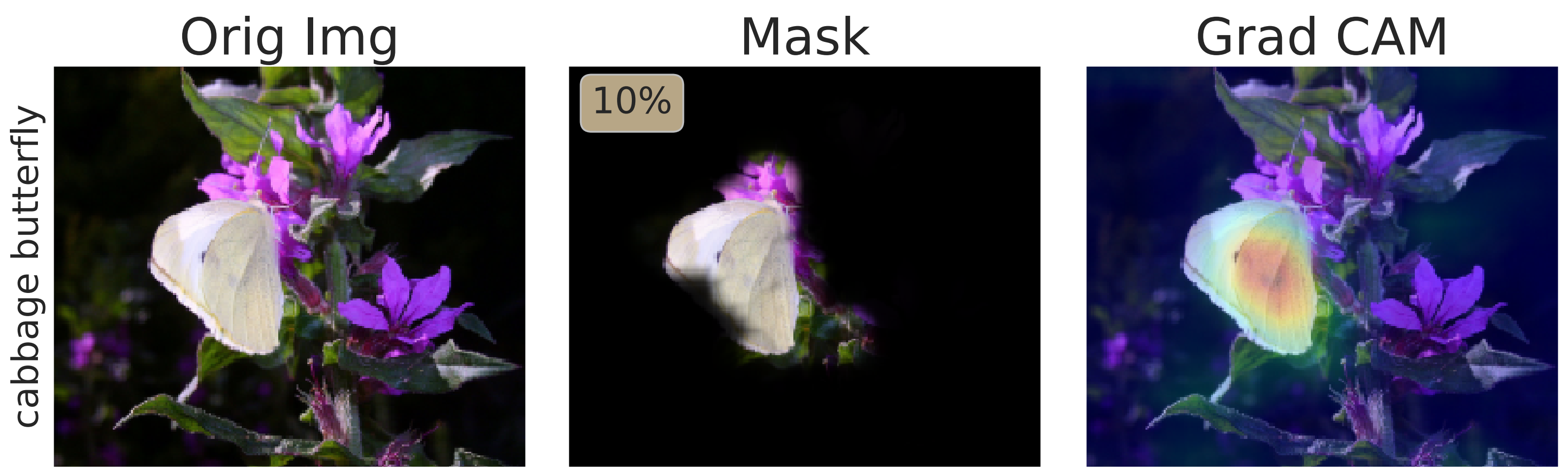
An incomplete retrospective: the first decade of interpretability



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

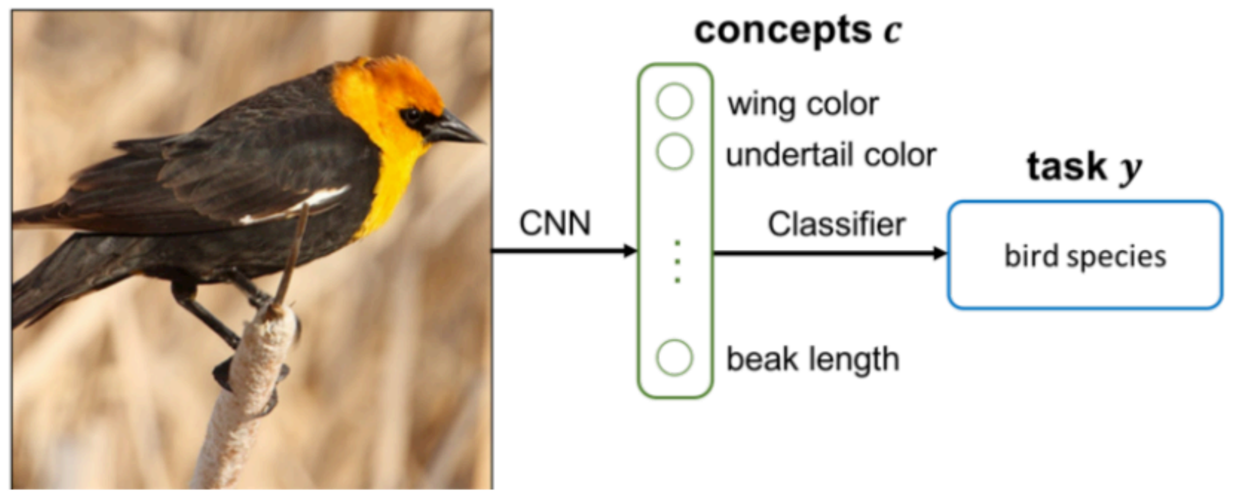
Primarily focused on understanding and approximating **CNNs**

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]

Into the future: the next decade of interpretability

???

2022 2032


Indu Panigrahi

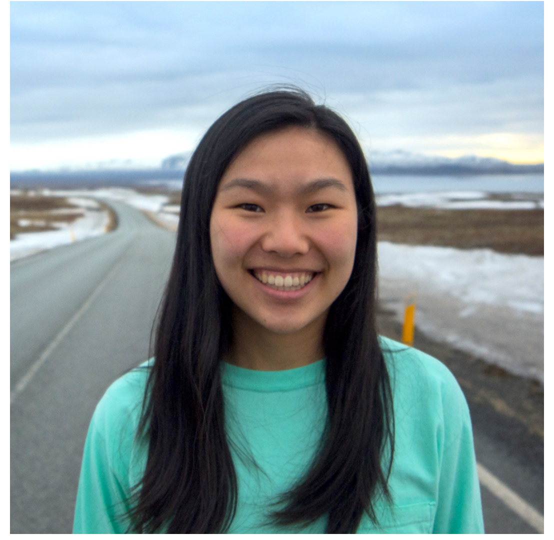

Devon Ulrich


Dora Zhao


Nicole Meister


Sunnie S. Y. Kim


Vikram V. Ramaswamy


Angelina Wang


Ryan A. Manzuk


Andrea Vedaldi


Elizabeth Anne Watkins


Andrés Monroy-Hernández


Chris Olah


Alex Mordvintsev


Adam C. Maloof


Olga Russakovsky

UNIVERSITY OF OXFORD

RHODES TRUST

NSF

Open Philanthropy Project

intel labs

SONY

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

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

Thank You