Directions in Interpretability

Ruth Fong

HEIBRiDS lecture
November 14, 2022
Slides and links available at ruthfong.com
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

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

GANs (2014-2018)
- GAN
- ProGAN
- CycleGAN

Transformers (2017-now)
- Transformer
- BERT
- ViT

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.

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]

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 $\rightarrow$ **human-centered** evaluation
   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** $\rightarrow$ explanations via **labelled attributes and unlabelled features**
   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 $\rightarrow$ interpretability of **self-supervised** models
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Interpretability** in ML + CV $\rightarrow$ **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 $\rightarrow$ **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

1. **Automated** evaluation of interpretability $\rightarrow$ **human-centered** evaluation  
   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** $\rightarrow$ explanations via **labelled attributes and unlabelled features**  
   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 $\rightarrow$ interpretability of **self-supervised** models  
   Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning.

4. **Interpretability** in ML + CV $\rightarrow$ **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 $\rightarrow$ **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)

**Counterfactual** explanations (e.g. SCOUT)

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

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

Explanation
(not part of model design)
Interpretable-by-design models

Explanation
(produced as part of model design)
Current metrics focus on heatmap evaluation

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

- 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)?
1. Within method → **Cross-method comparison**

2. Automated evaluation → **Human-centered evaluation**

3. Intuition-based reasoning → **Falsifiable hypothesis testing**
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

**heatmap**
**post-hoc**
**interpretable-by-design**

[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]
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

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).

---

**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: Looks like
- 2: Looks like
- 3: Looks like
- 4: Looks like

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

**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.**

**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?

![Graph showing required accuracy gain for low-risk, medium-risk, and high-risk settings.]

- Low-risk: +6.2%
- Medium-risk: +8.2%
- High-risk: +10.9%

(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?

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
...

Interview
Merlin app
Heatmaps
Examples

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
   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**
   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
   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

Predict present or absence of attribute
Linearly combine with attribute weights

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

Predict present or absence of attribute
Linearly combine with attribute weights

**Con:** Problems with predicting fractional values
- hard to interpret
- can encode hidden information

CNN

1 1 0 ...
\[\times \times \times\]

\[\sum\]

sheepdog

attribute weights for sheepdog

[+1.2 fur +0.7 paw -0.6 tree ...]

[Koh*, Nguyen*, Tang* et al., ICML 2020]
ELUDE: Explanation via a Labelled and Unlabelled DEcomposition 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**

---

**Feature activations for sheepdog**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>8.2</td>
</tr>
<tr>
<td>$f_2$</td>
<td>4.5</td>
</tr>
<tr>
<td>$f_3$</td>
<td>-7.6</td>
</tr>
</tbody>
</table>

**Feature weights for sheepdog**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>fur</td>
<td>+1.2</td>
</tr>
<tr>
<td>paw</td>
<td>+0.7</td>
</tr>
<tr>
<td>tree</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

**Ground-truth presence/absence of attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>fur</td>
<td>1</td>
</tr>
<tr>
<td>paw</td>
<td>1</td>
</tr>
<tr>
<td>tree</td>
<td>0</td>
</tr>
</tbody>
</table>

[Vikram V. Ramaswamy et al., arXiv 2022. ELUDE.]
Attributes only: % of model explained via labelled attributes decreases as task complexity increases

<table>
<thead>
<tr>
<th>Task</th>
<th>% Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-way scene classification (indoor vs. outdoor)</td>
<td>95.7</td>
</tr>
<tr>
<td>16-way scene classification (home/hotel, workplace, etc.)</td>
<td>46.2</td>
</tr>
<tr>
<td>365-way scene classification (airfield, bowling alley, etc.)</td>
<td>28.8</td>
</tr>
</tbody>
</table>

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

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

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?

![Attributes Table]

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

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).
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

**Probe dataset:**
- ADE20k
  - {grandstand, goal, ice rink, scoreboard}
- Pascal
  - {plaything, road}

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

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

**Probe dataset:** Broden

<table>
<thead>
<tr>
<th>Concept</th>
<th>norm AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>toilet</td>
<td>39.9</td>
</tr>
<tr>
<td>shower</td>
<td>18.8</td>
</tr>
<tr>
<td>countertop</td>
<td>12.6</td>
</tr>
<tr>
<td>bathtub</td>
<td>11.1</td>
</tr>
<tr>
<td>screen door</td>
<td>9.6</td>
</tr>
</tbody>
</table>

bathroom (norm AP = 43.3)

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

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?

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

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

1. **Automated** evaluation of interpretability → **human-centered** evaluation
   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
   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
   Quantifying Learnability andDescribability 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

\[(x, \text{sheepdog}, y)\]
Self-Supervised Learning
Visual Concept

Query
Self-Supervised Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorization</td>
<td>2016-2018</td>
</tr>
<tr>
<td>Jigsaw</td>
<td>2016-2018</td>
</tr>
<tr>
<td>CPC</td>
<td>2019</td>
</tr>
<tr>
<td>BigBiGAN</td>
<td>2019</td>
</tr>
<tr>
<td>MoCo</td>
<td>2019</td>
</tr>
<tr>
<td>SeLa</td>
<td>2019</td>
</tr>
<tr>
<td>CMC</td>
<td>2019</td>
</tr>
<tr>
<td>SeLa-v2</td>
<td>2020</td>
</tr>
<tr>
<td>SimCLR</td>
<td>2020</td>
</tr>
<tr>
<td>MoCo-v2</td>
<td>2020</td>
</tr>
<tr>
<td>SimCLR-v2</td>
<td>2020</td>
</tr>
<tr>
<td>BYOL</td>
<td>2020</td>
</tr>
<tr>
<td>SwAV</td>
<td>2020</td>
</tr>
</tbody>
</table>
Self-Supervised Learning

Unlabelled data

Learn clusters

(e.g. DeepCluster, SeLa, SwaV)

Learn features

(e.g. SimCLR, MoCo, ...)

k-means

cluster 1

cluster 2

cluster K
Learnability

Learnability

white animal in snow

Describability

“dessert with chocolate sauce”

[A] [B]

Describability

“dessert with chocolate sauce”

Manual

(A)

(B)

Describability

“dessert with chocolate sauce”

Manual OR Automatic

(A) cup of coffee

(B) ice cream with chocolate sauce

Evaluation

Learnability

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

\[ \text{purity} = 1 \rightarrow \text{cluster only contains images from a single ImageNet label} \]

[Asano et al., ICLR 2020; He et al., CVPR 2020]
Evaluation

Learnability

Describability

[Asano et al., ICLR 2020; He et al., CVPR 2020]
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

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

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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeLa: 393</td>
<td>98.3%</td>
</tr>
<tr>
<td>SeLa: 332</td>
<td>100.0%</td>
</tr>
<tr>
<td>MoCo: 2335</td>
<td>93.3%</td>
</tr>
</tbody>
</table>

[Asano et al., ICLR 2020; He et al., CVPR 2020]
Gender artifacts are everywhere in visual datasets.

(*binary perceived gender expression; we do not condone gender prediction.)
Extending Interpretability to Geosciences

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)

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
   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**
   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
   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
Spatial Activations

\[ f_a \quad f_b \quad \text{golden retriever} \]
Spatial Activations

[Olah et al., Distill 2018]
Interactive Similarity Overlays

\[ a_{6,5} = [17.7, 0, 103.4, 6.81, 0, 0, 0, 0, 32.0, 0, 0, 0, ...] \]
Interactive Similarity Overlays

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

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.
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.
```python
# Get images
img_urls = [
    "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 = getActs(model, imgs[0], "mixed4d")
grid = np.hstack(np.hstack(gramm_sim_grid(acts, acts)))
colored_grid = add_color_index(grid, acts.shape[0])
```

Preview: Interactive Visual Feature Search

bit.ly/interactive_search

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.
An incomplete retrospective: the first decade of interpretability

[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]

Primarily focused on understanding and approximating CNNs

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

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

Interpretable-by-design (2020-now)
Concept Bottleneck, ProtoPNet, ProtoTree
Into the future: the next decade of interpretability
Thank You