Reverse Engineering the visual system

PI : Thomas Serre

Victor Boutin

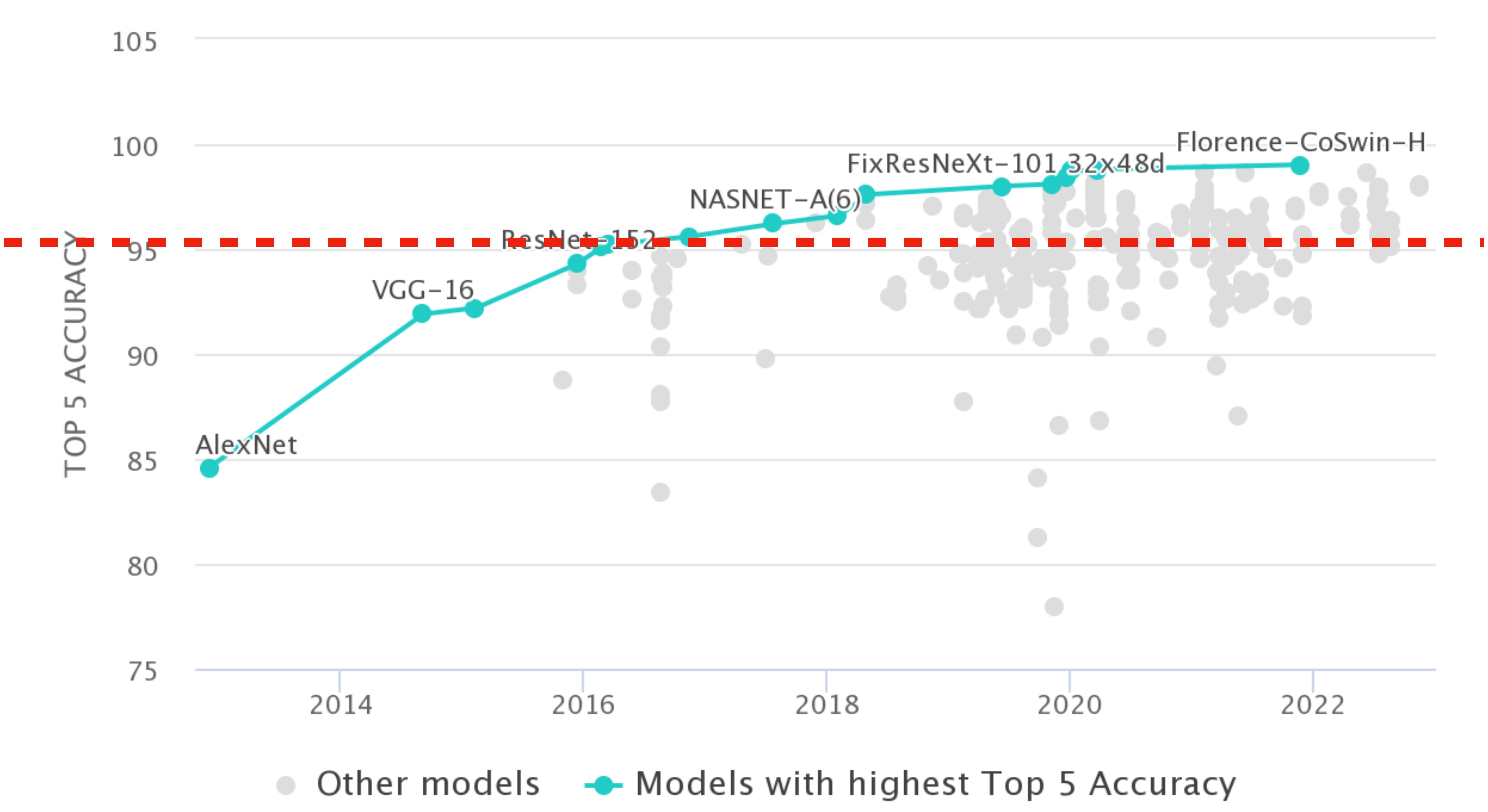
16/11/2023

ANITI

Université de Toulouse

Artificial vision = Biological vision ?

Claimed to be
« human level »



SOURCE : PAPERSWITHCODE.COM

Artificial vision = Biological vision ?

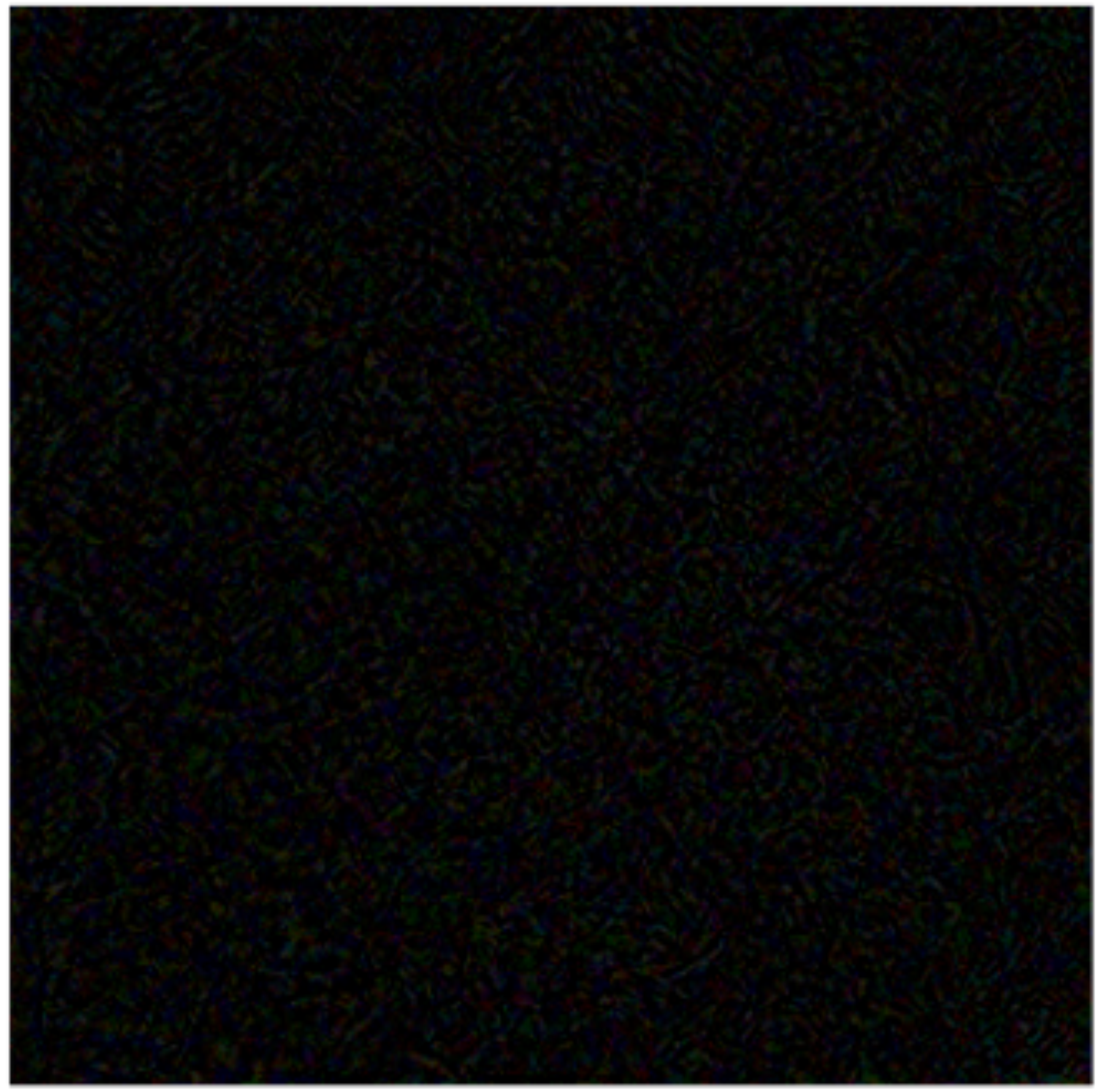
Prediction : Cat

SZEGEDY ET AL 2013

Artificial vision = Biological vision ?

Prediction : Cat

+ 0.25 x

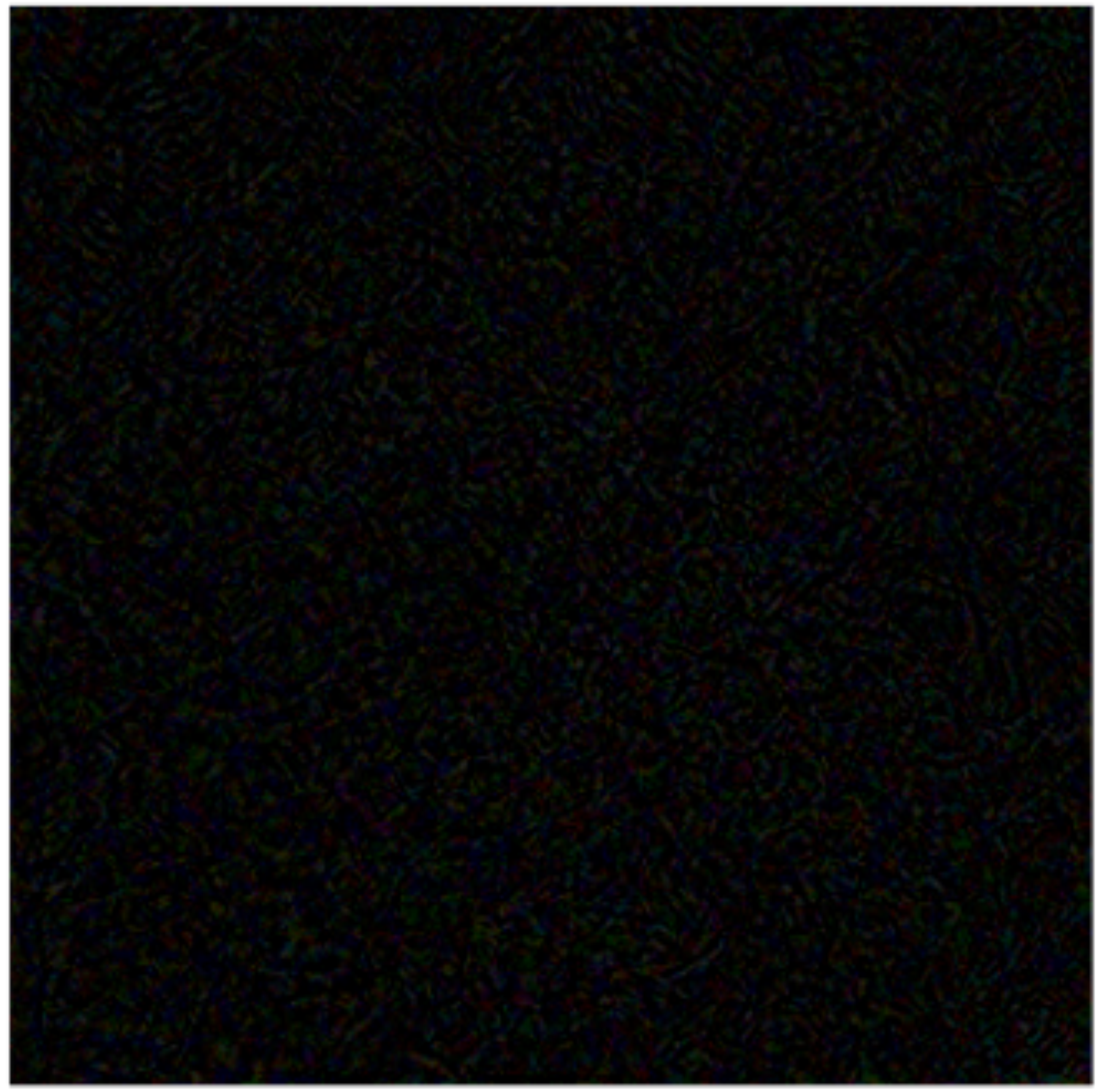


SZEGEDY ET AL 2013

Artificial vision = Biological vision ?

Prediction : Cat

\+ **0.25 x**

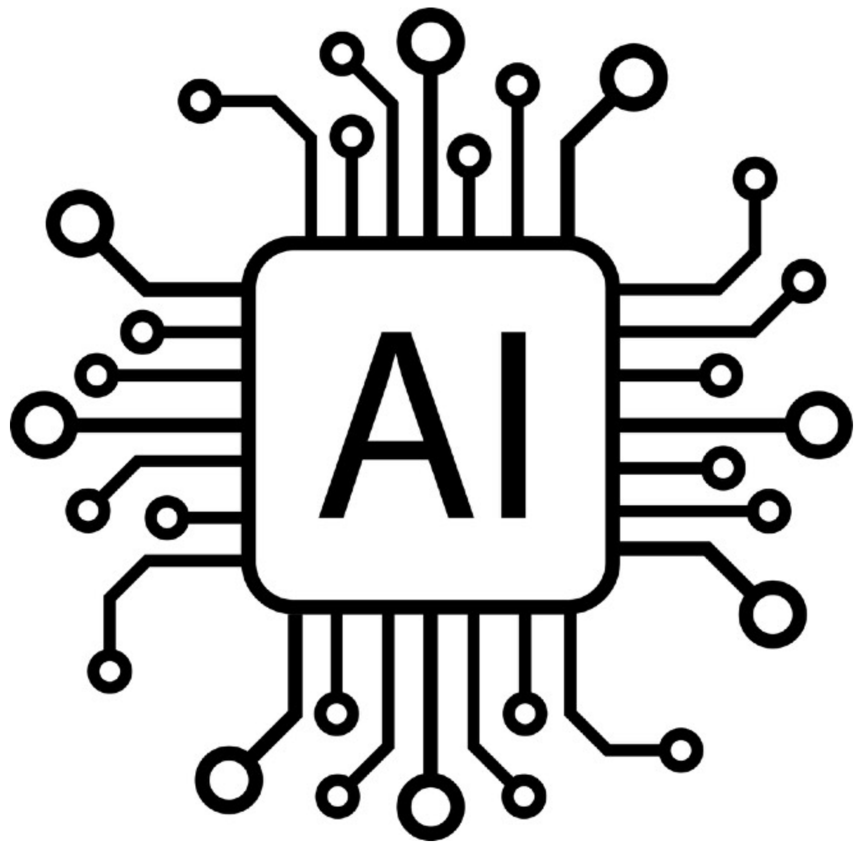


=

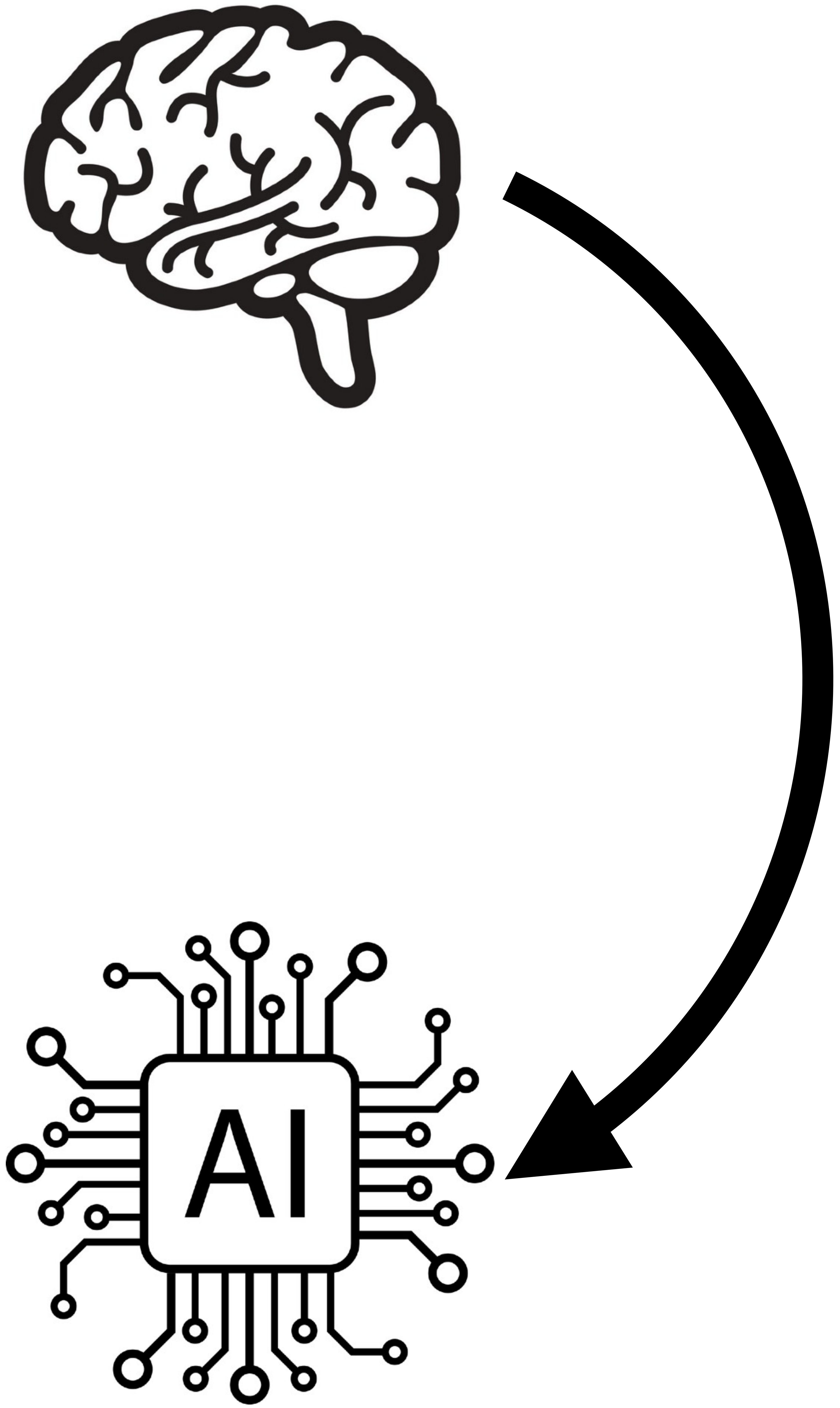
Prediction : Ostrich

SZEGEDY ET AL 2013

Reverse Engineering the Visual System



Reverse Engineering the Visual System

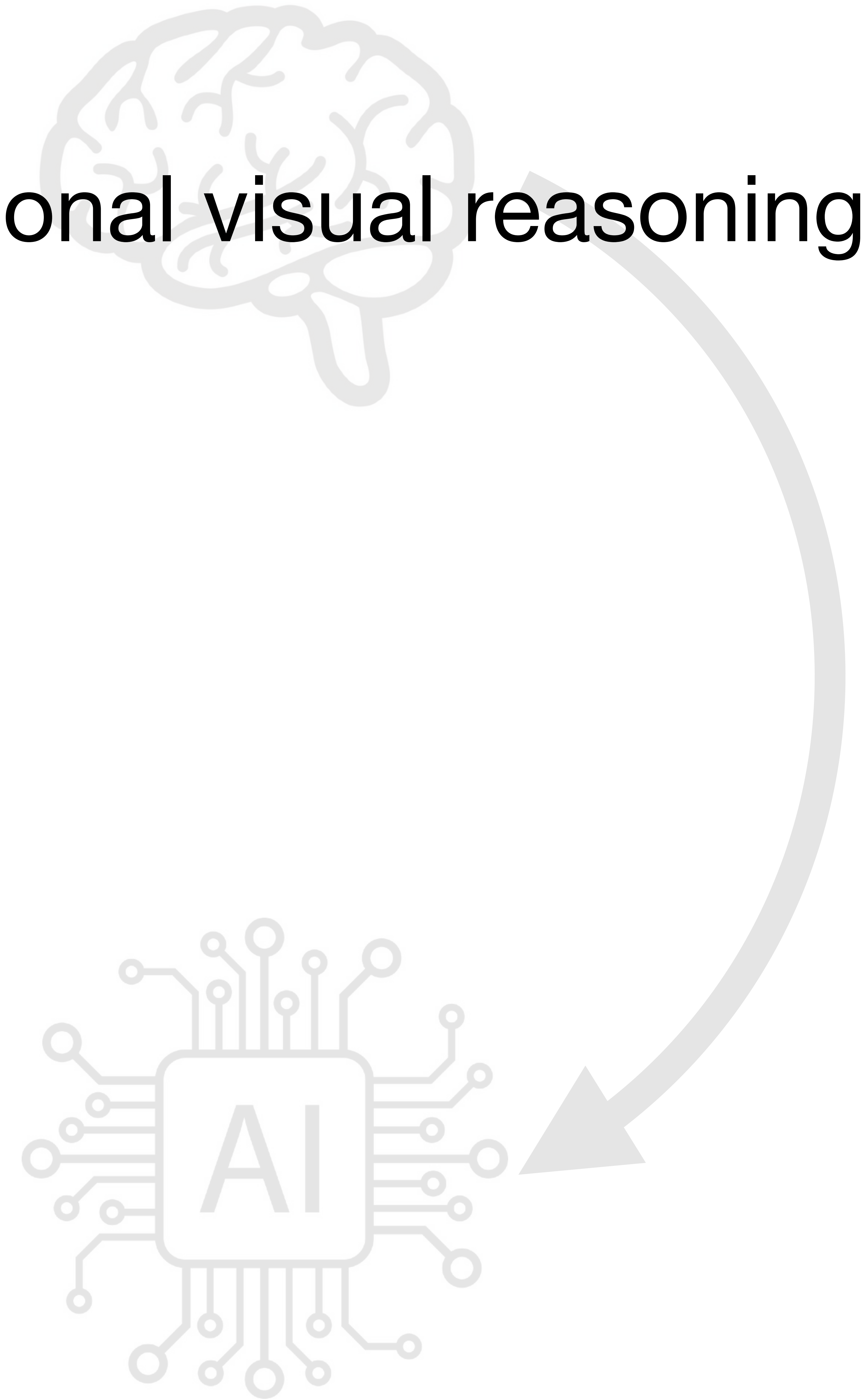


Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

Reverse Engineering the Visual System

- Benchmark for compositional visual reasoning

  ZERROUG ET AL 2022



Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

Reverse Engineering the Visual System

- Benchmark for compositional visual reasoning

Zerroug et al 2022

- Neuro plausible attentional mechanisms

Vaishnav et al 2022

Train AI on tasks
cognitive scienc
highlight key co
mechanisms

Reverse Engineering the Visual System

- Benchmark for compositional visual reasoning

  ZERROUG ET AL 2022

- Neuro plausible attentional mechanisms

  VAISHNAV ET AL 2022

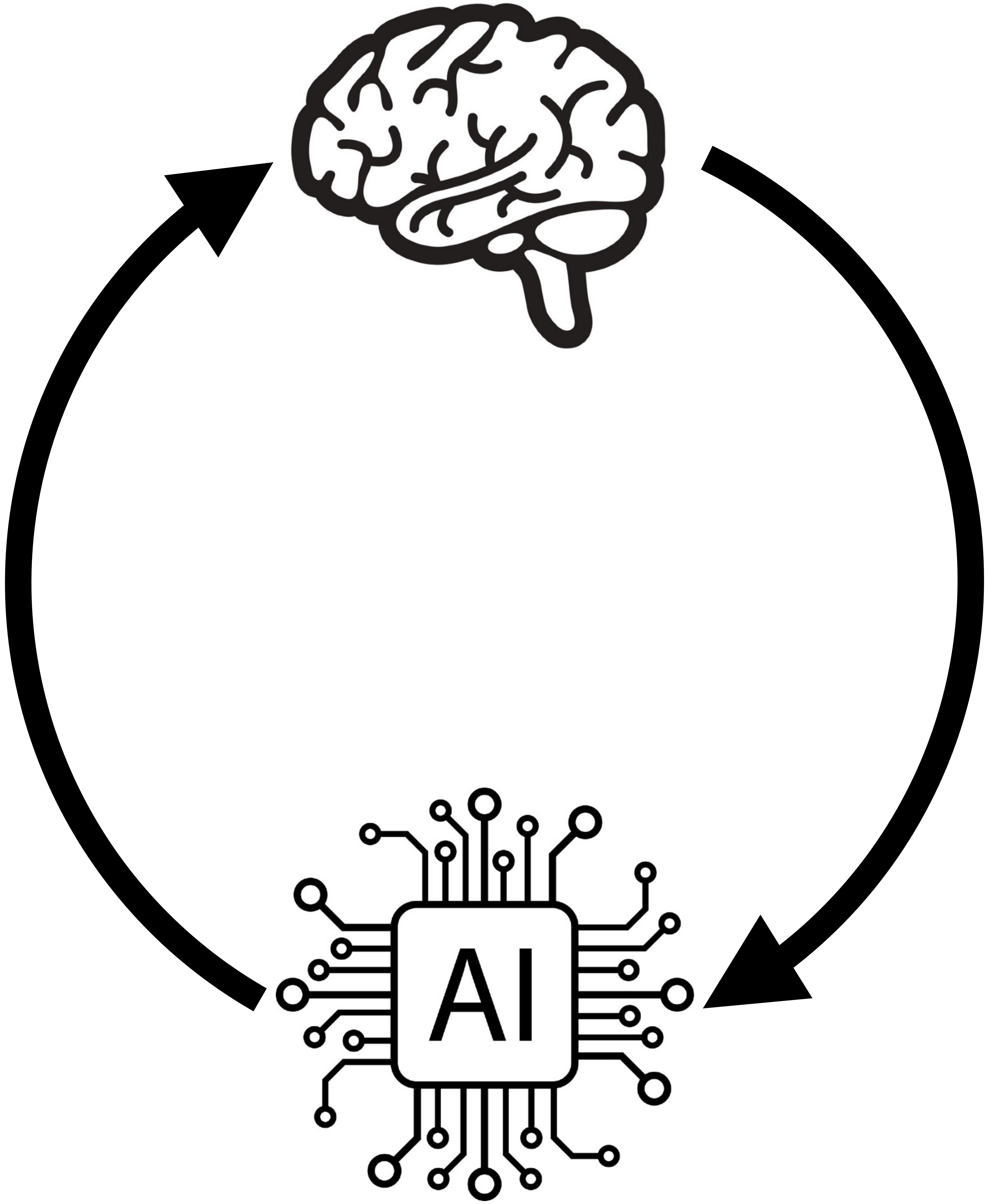
- Leveraging binding by synchrony in complex networks (with VanRullen)

  MUZELLEC ET AL 2023

Train AI on tasks
cognitive scienc
highlight key co
mechanisms

Reverse Engineering the Visual System

We test « humanness » of AI using XAI and metrics from cognitive science



Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

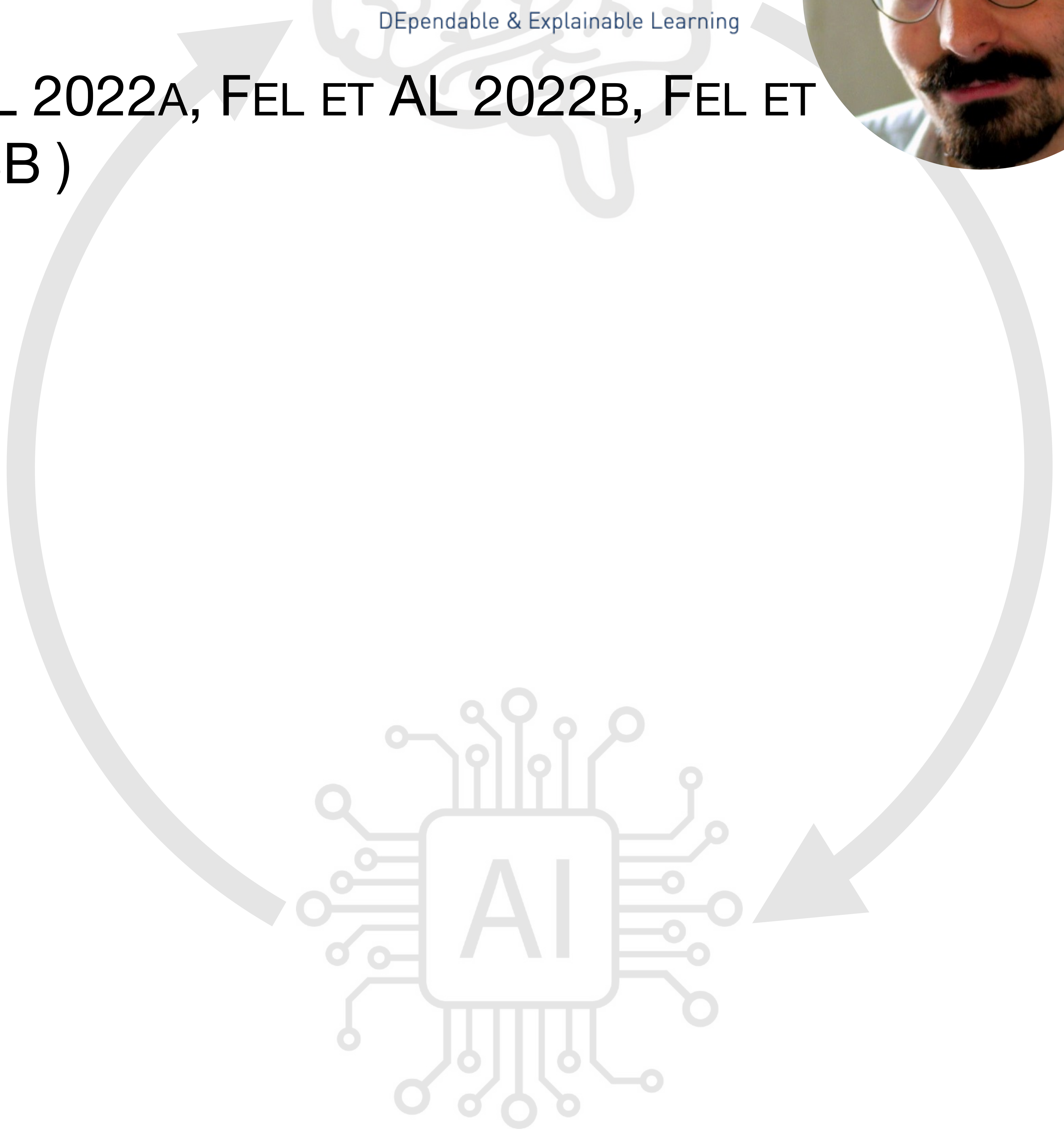
Reverse Engineering the Visual System

- Improving XAI methods

(Fel et al 2021, Fel et al 2022a, Fel et al 2022b, Fel et al 2023a, Fel et al 2023B)

DEEL
DEpendable & Explainable Learning





We test « humanness » of AI using XAI and metrics from cognitive science

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

Reverse Engineering the Visual System

- Improving XAI methods

(Fel et al 2021, Fel et al 2022a, Fel et al 2022b, Fel et al 2023a, Fel et al 2023b)



- https://serre-lab.github.io/Lens/

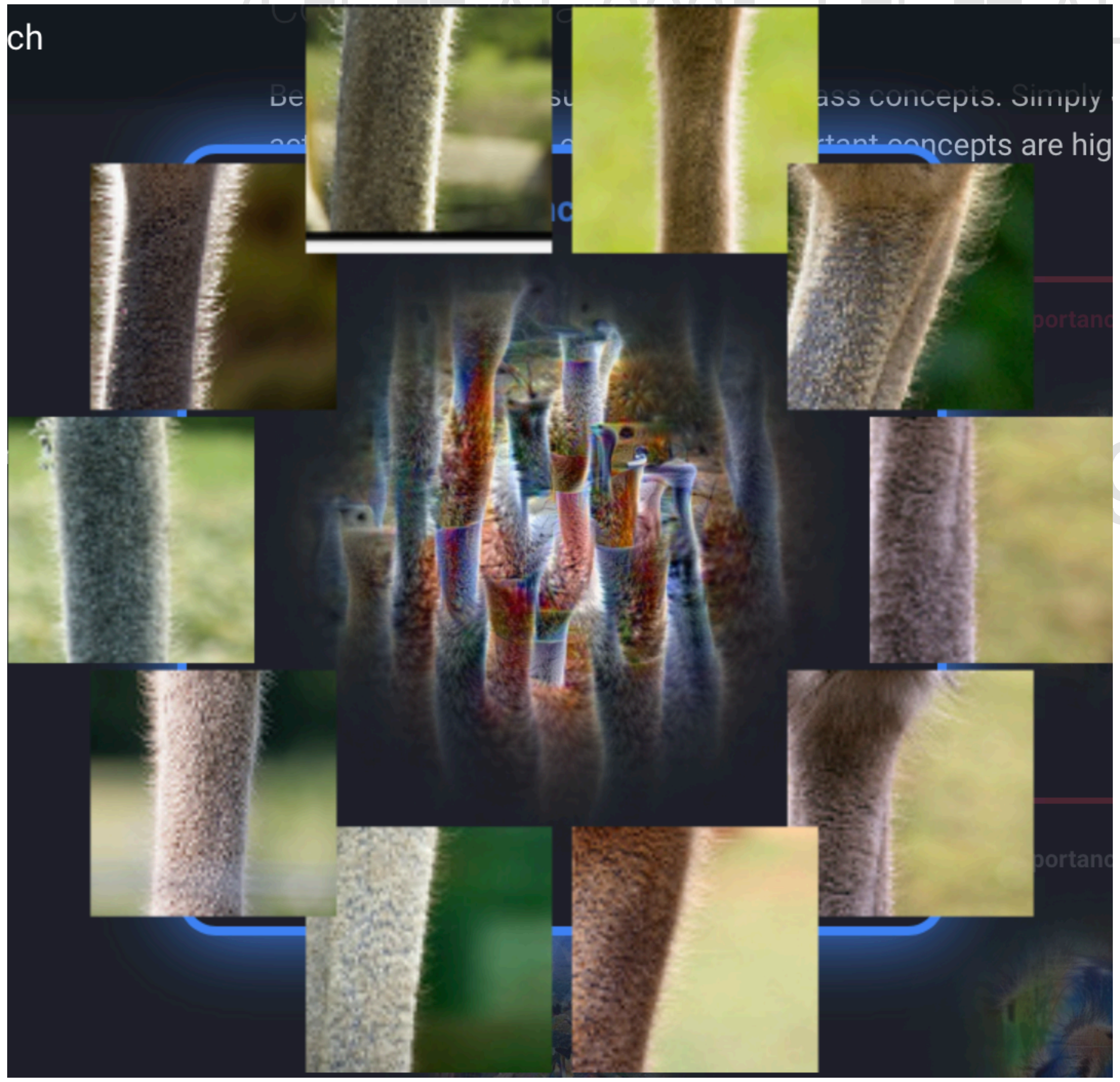
We test « humanness » of AI

cognitive science

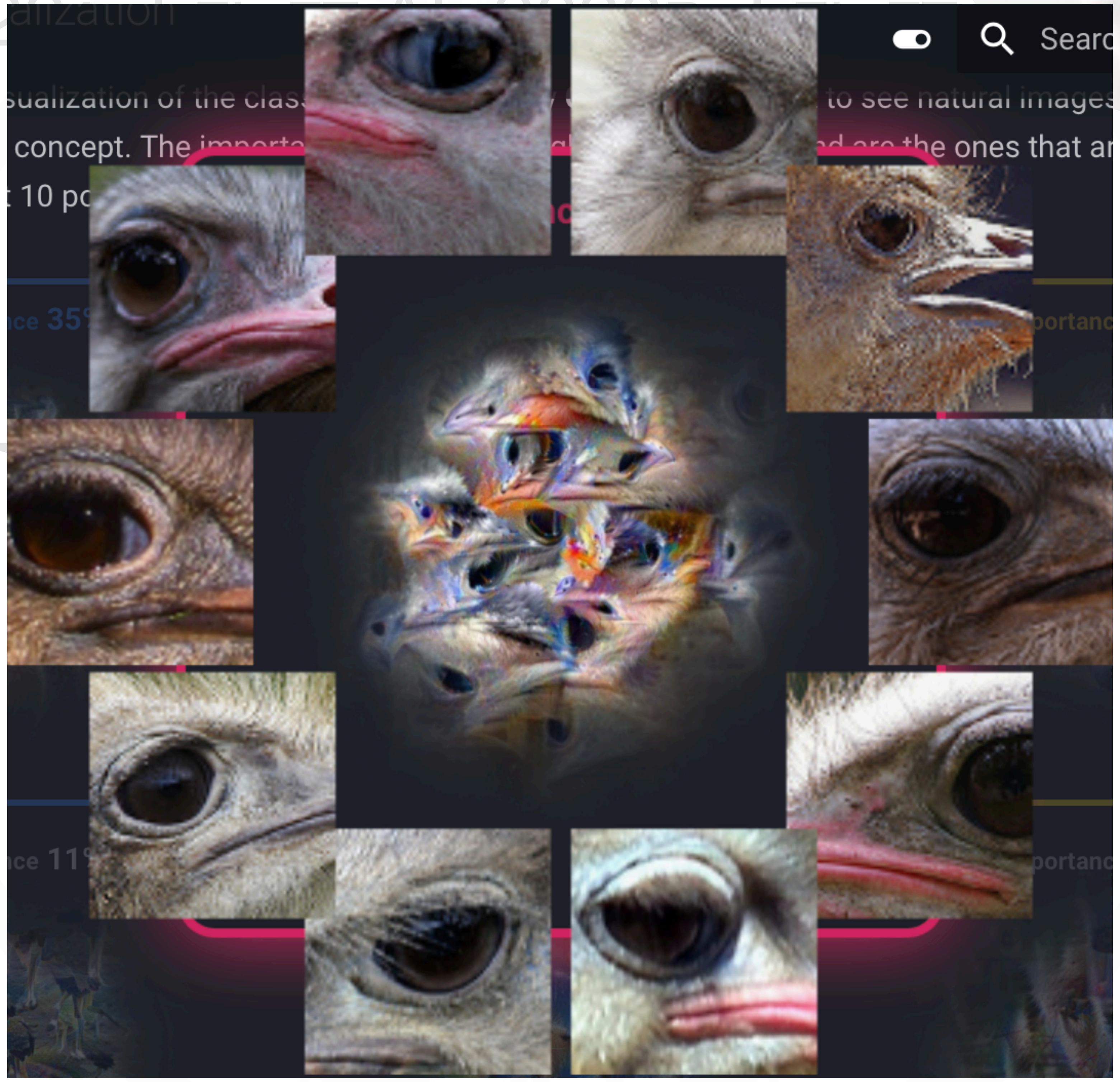
Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

Reverse Engineering the Visual System

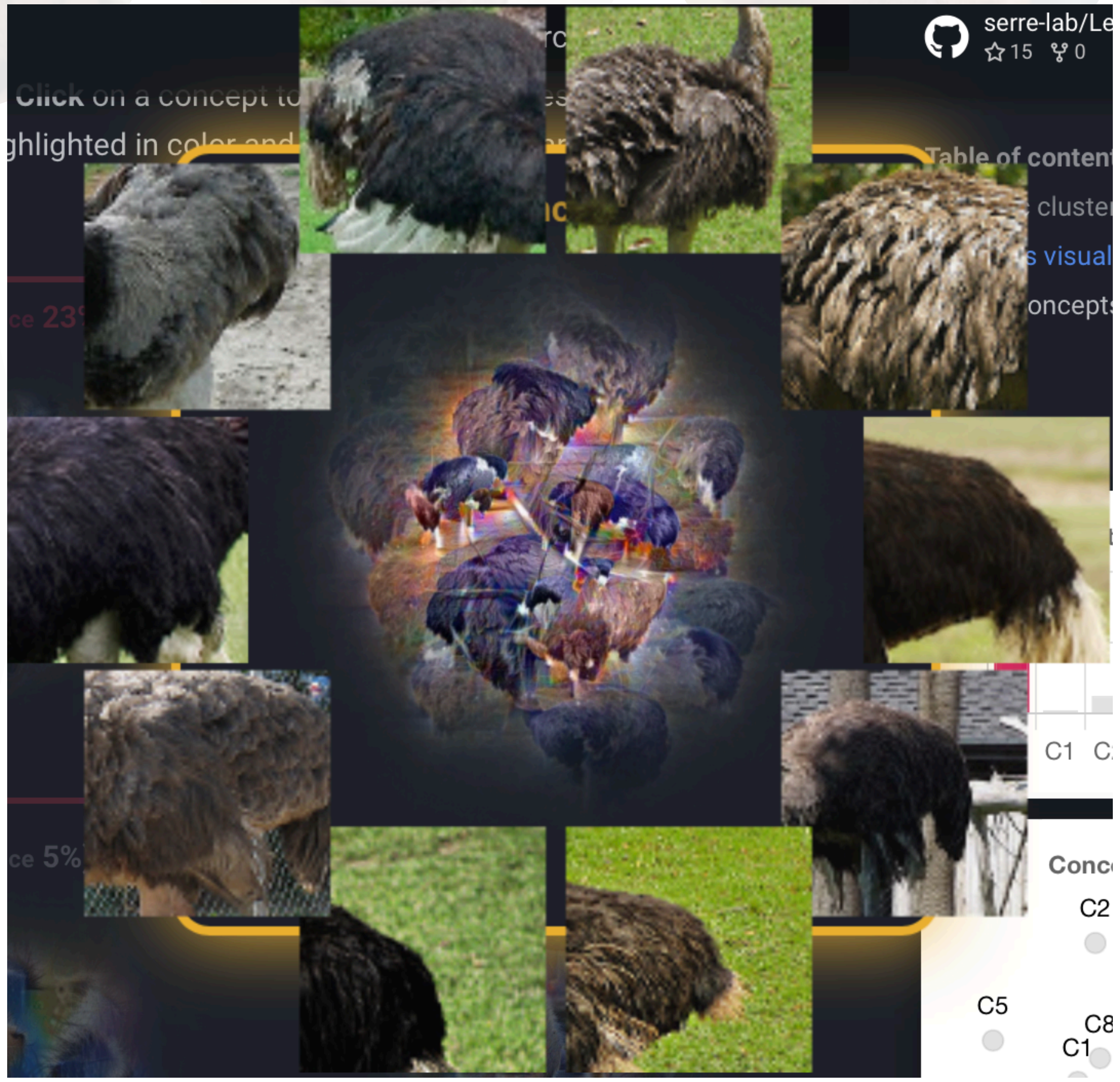
Concept 1 (35%)



Concept 2 (23%)



Concept 3 (20%)



Reverse Engineering the Visual System

DEEL
DEpendable & Explainable Learning



- Improving XAI methods

  (FEL ET AL 2021, FEL ET AL 2022A, FEL ET AL 2022B, FEL ET AL 2023A, FEL ET AL 2023B)

We test « humanness » of AI

- https://serre-lab.github.io/Lens/

cognitive science

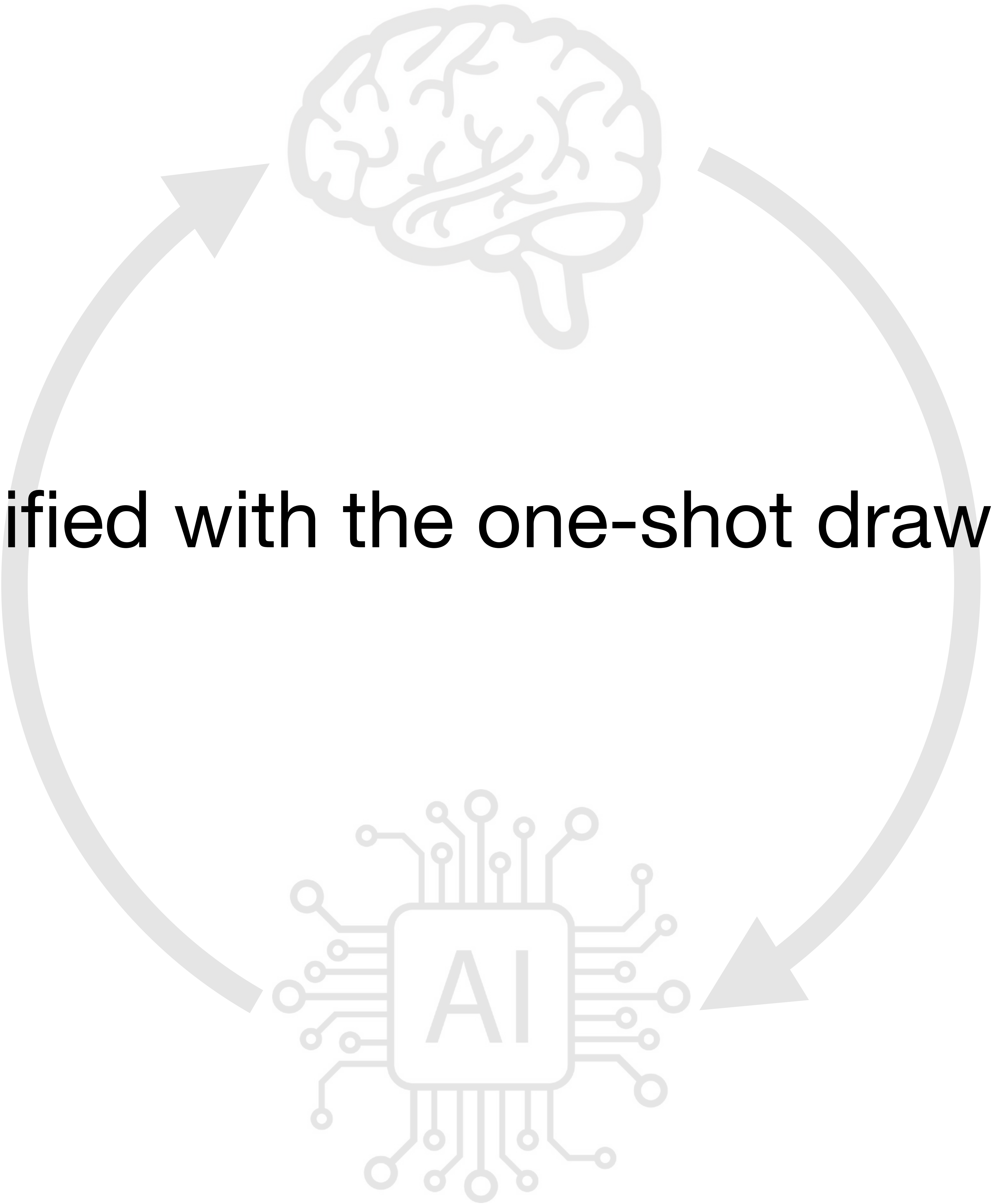
Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

- Harmonizing machines and humans with XAI

  (FEL ET AL 2022C)

Reverse Engineering the Visual System

We test « humanness » of AI using XAI and metrics from cognitive science



Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

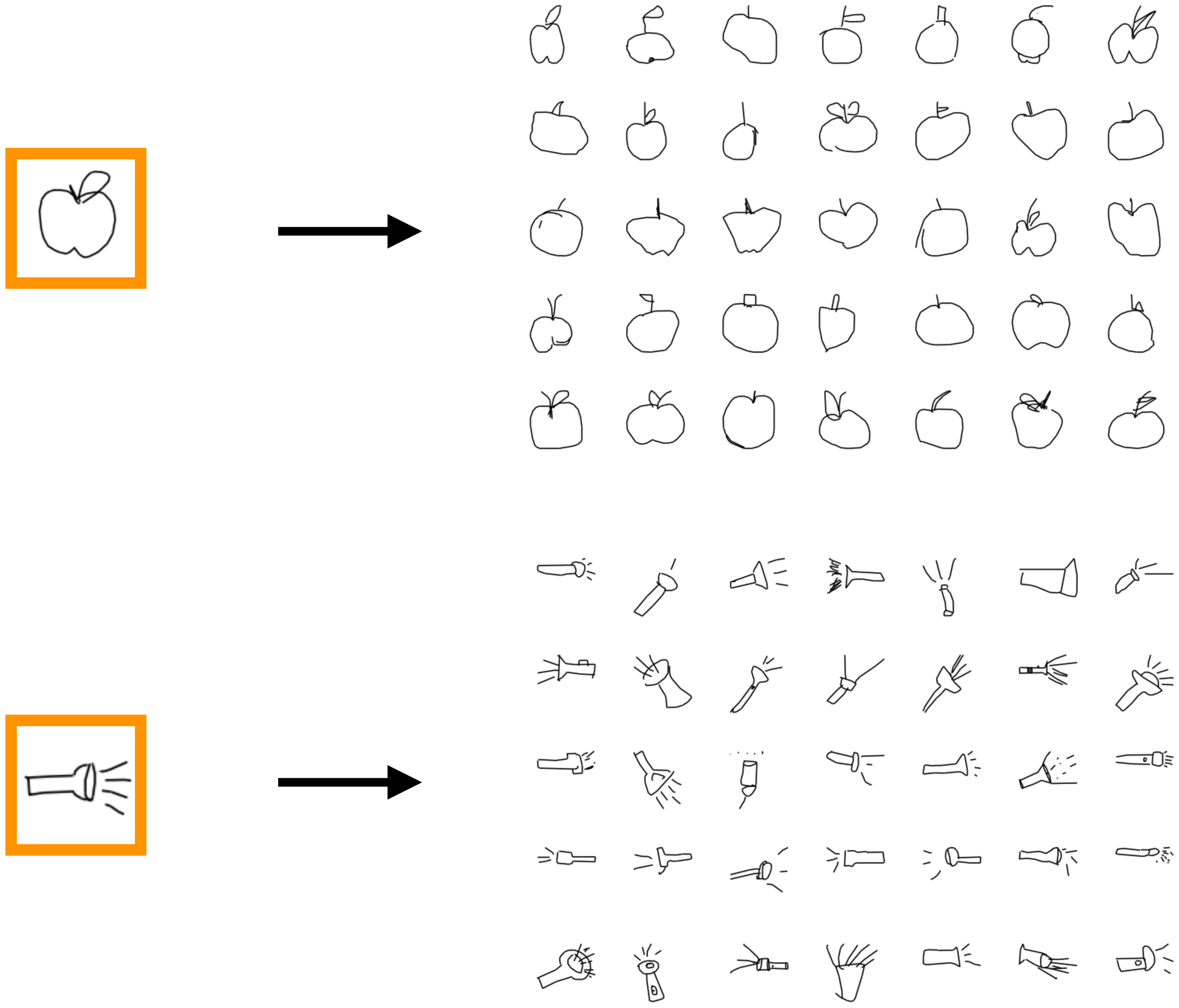
The chair exemplified with the one-shot drawing project …

One-Shot Drawing Task (Lake et al 2015)

One-Shot Drawing Task (Lake et al 2015)

Exemplars

Variations

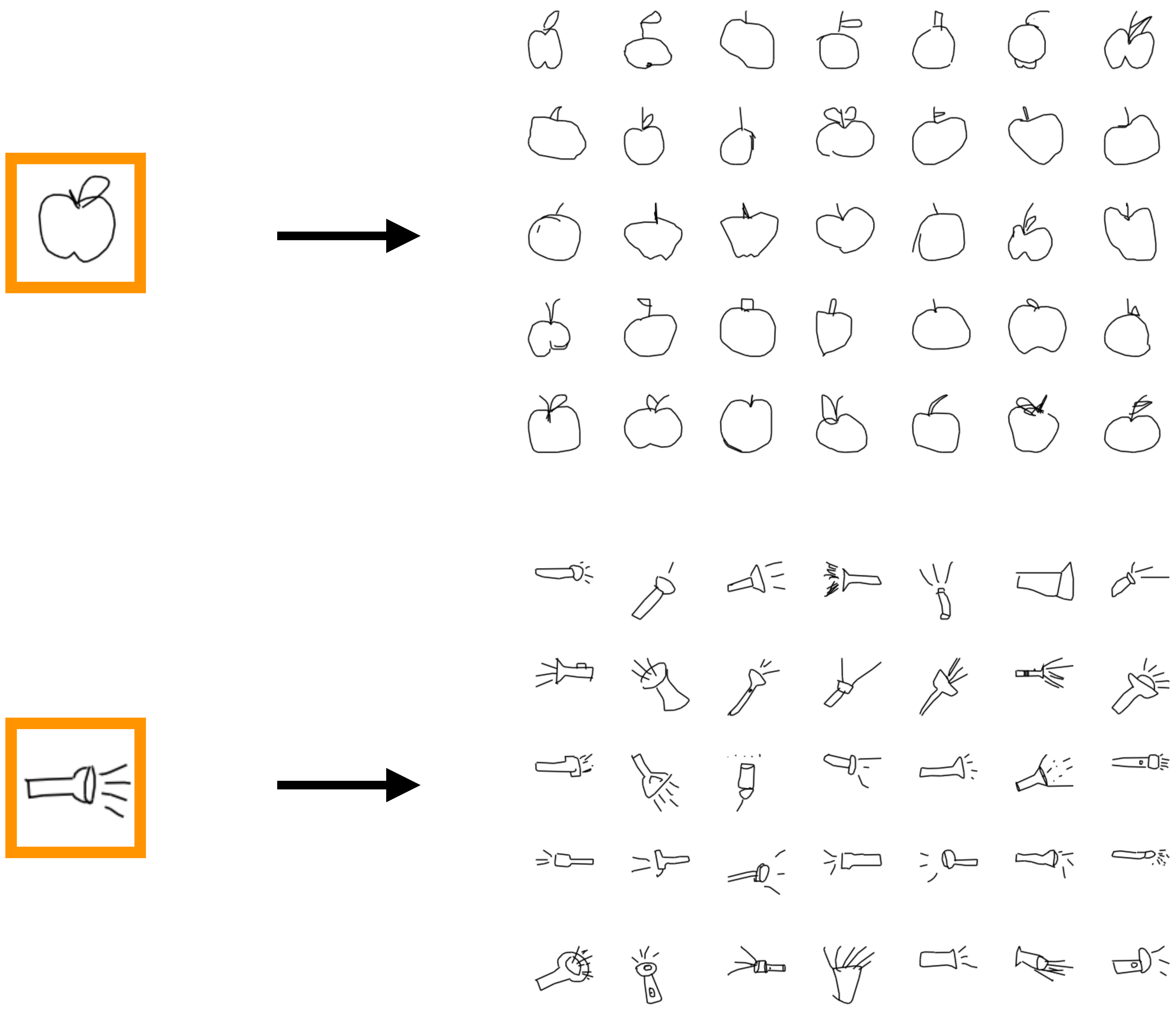


One-Shot Drawing Task (Lake et al 2015)

Training

Exemplars

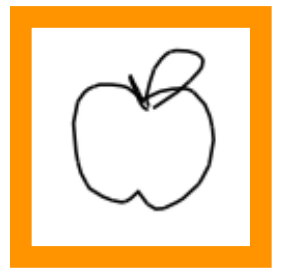
Variations

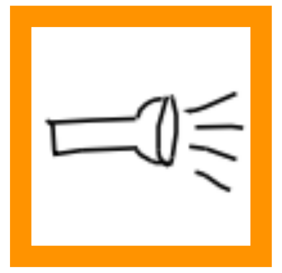


One-Shot Drawing Task (Lake et al 2015)

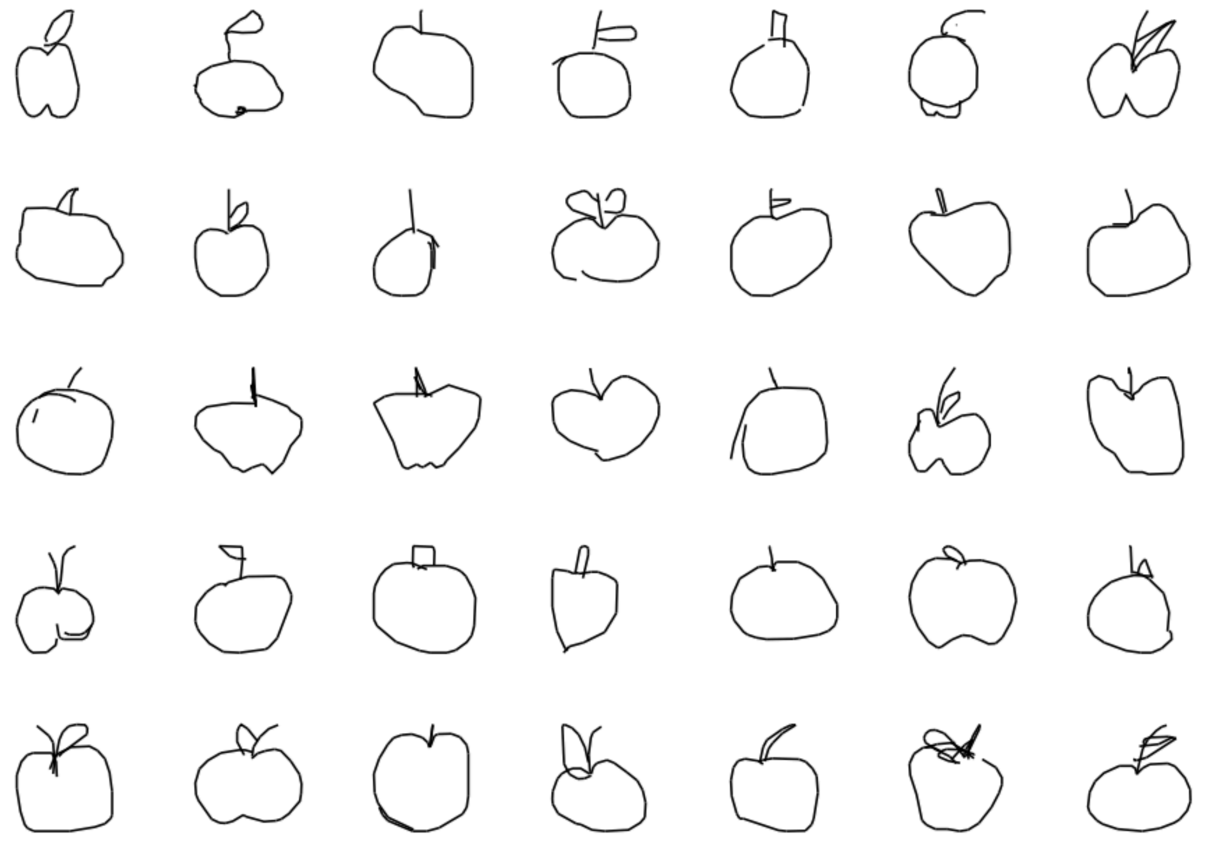
Training

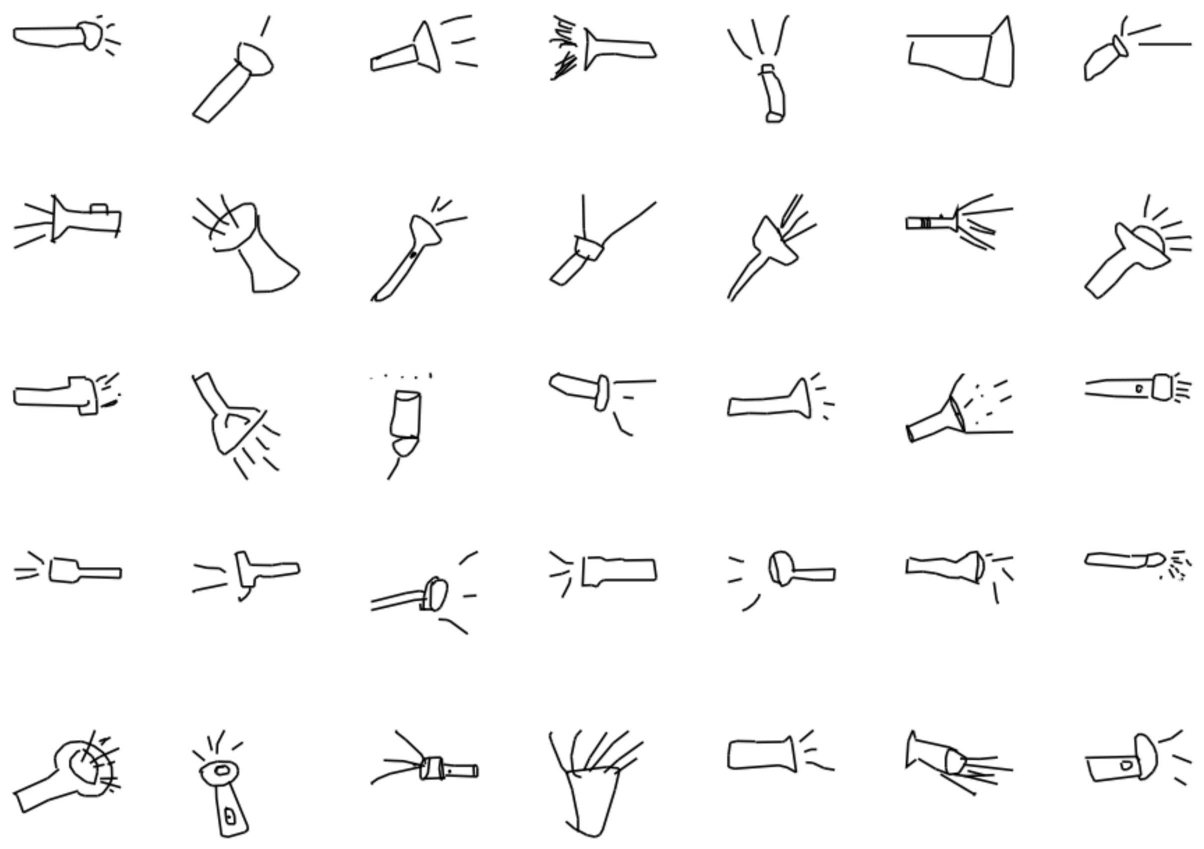
Exemplars





Variations





Testing

New exemplars

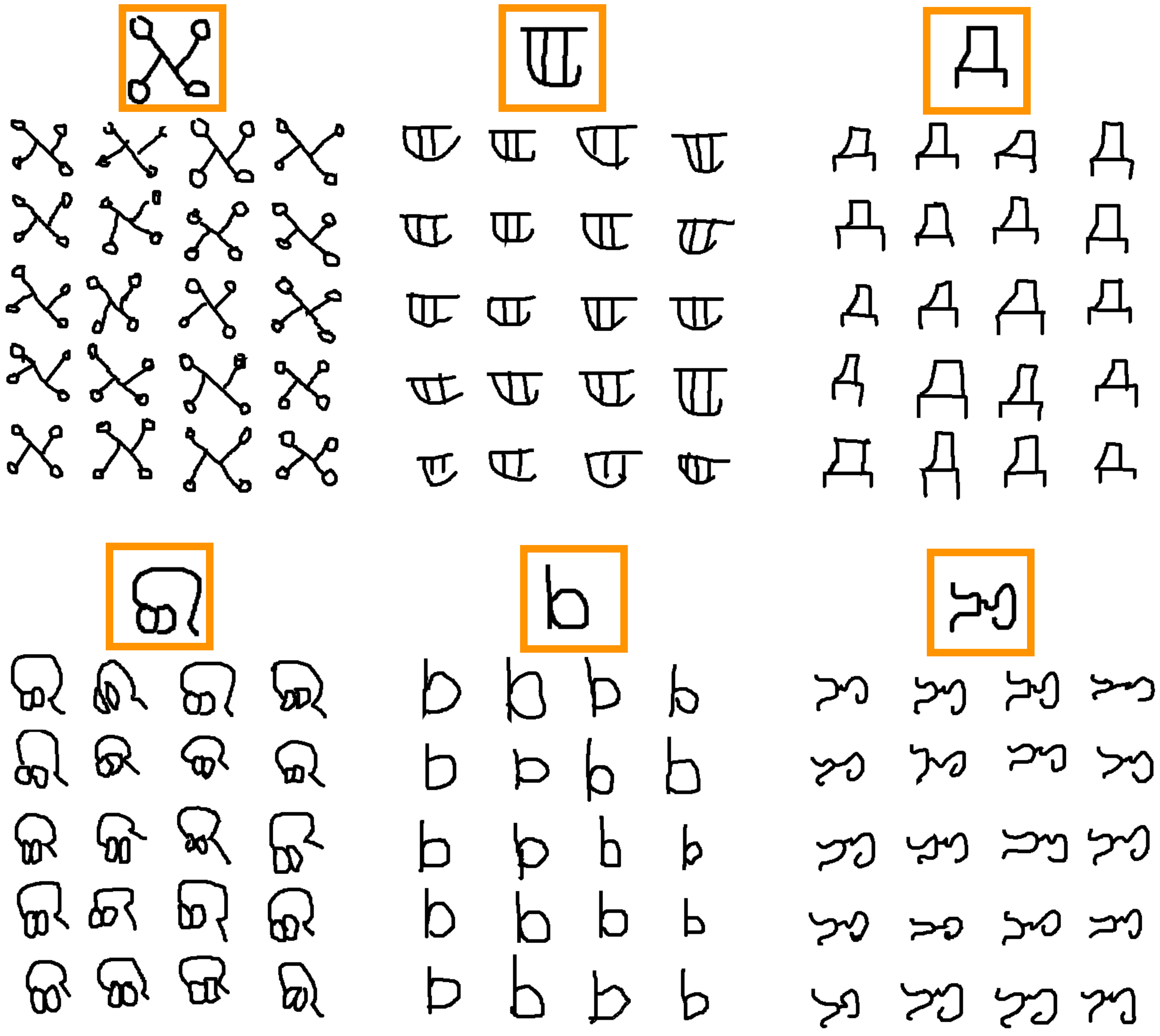
Variations

?

?

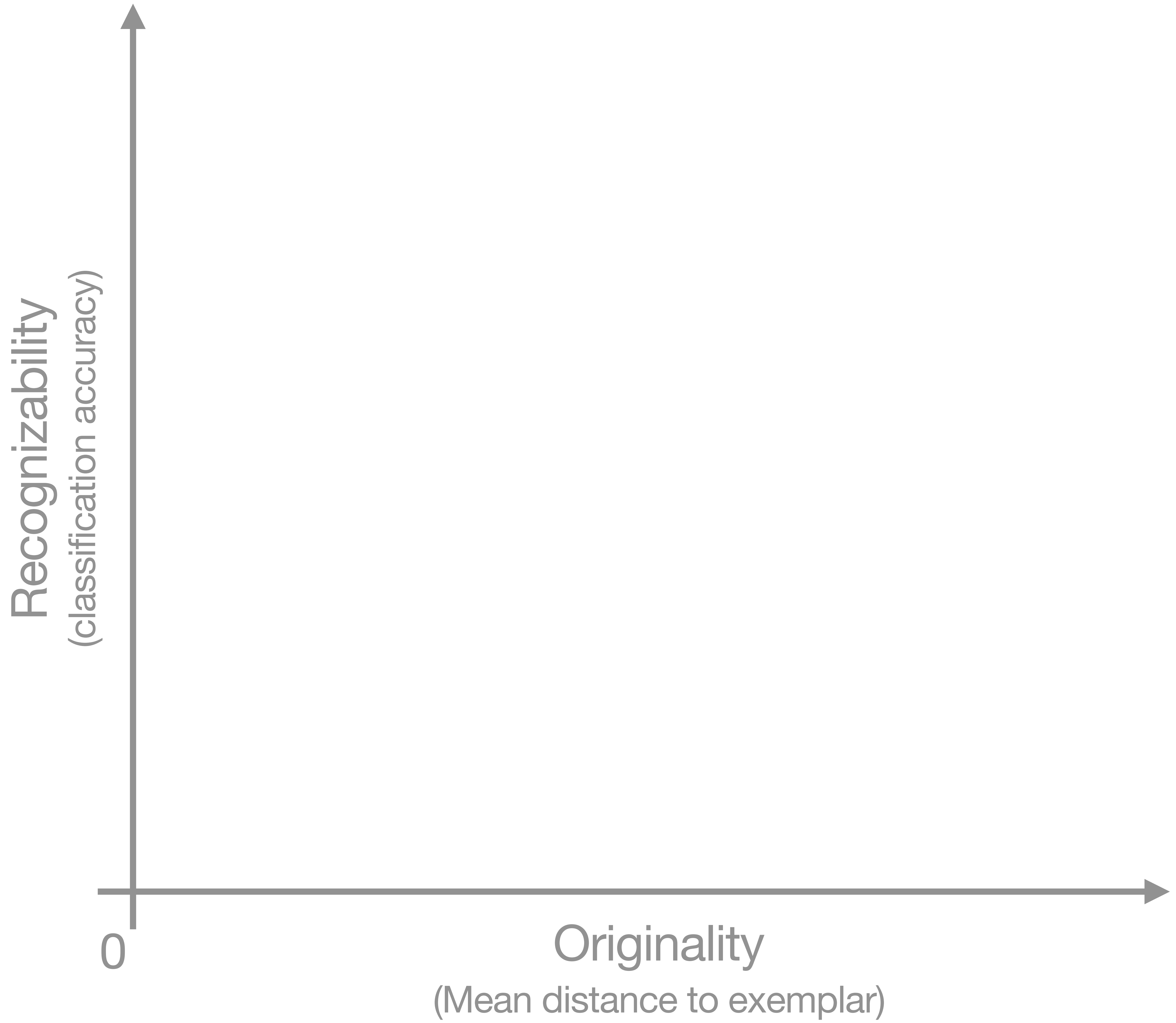
One-Shot Drawing Task (Lake et al 2015)

Omniglot (Lake et al 2015)



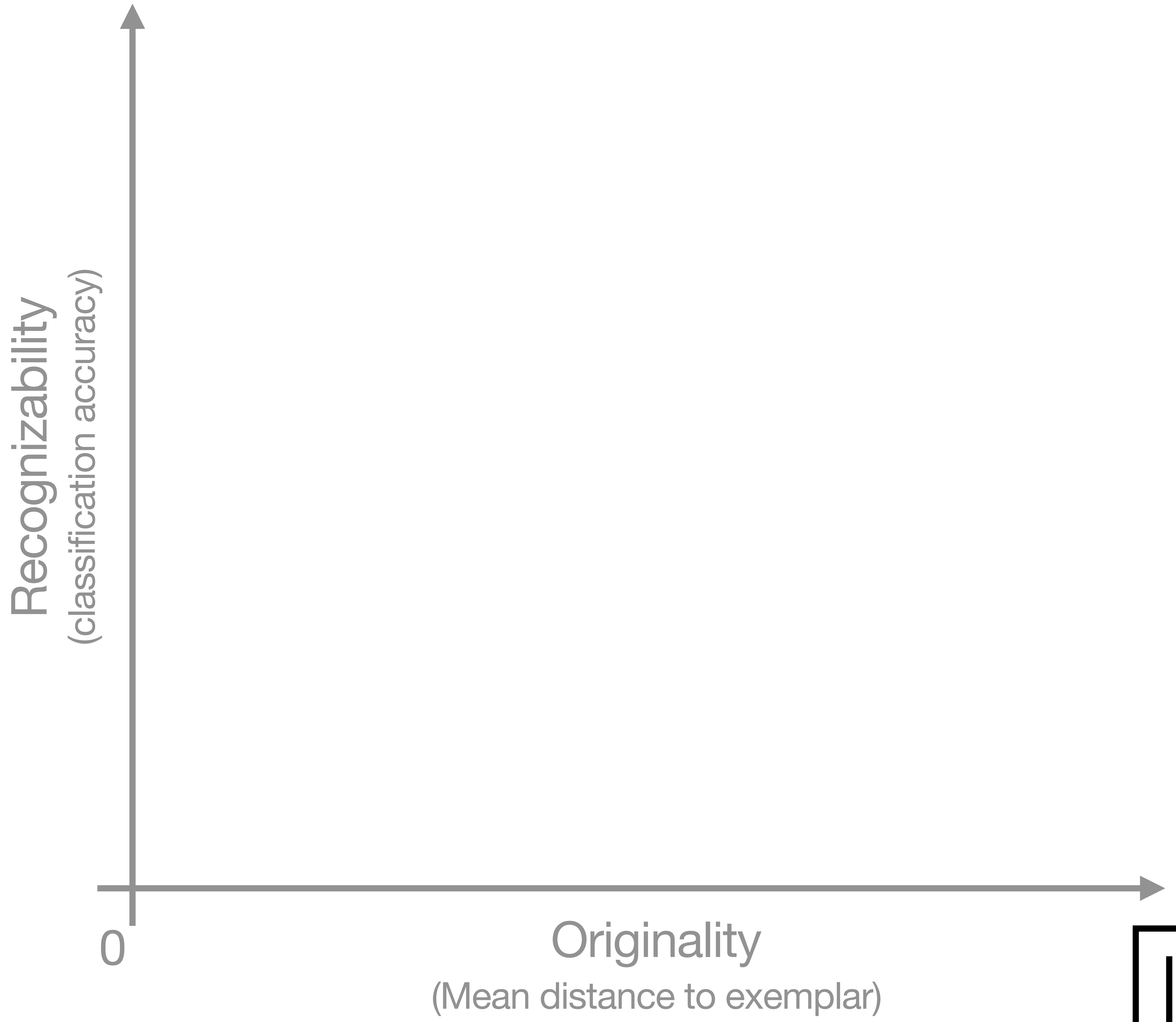
Quick, Draw ! (Ha et al 2017)

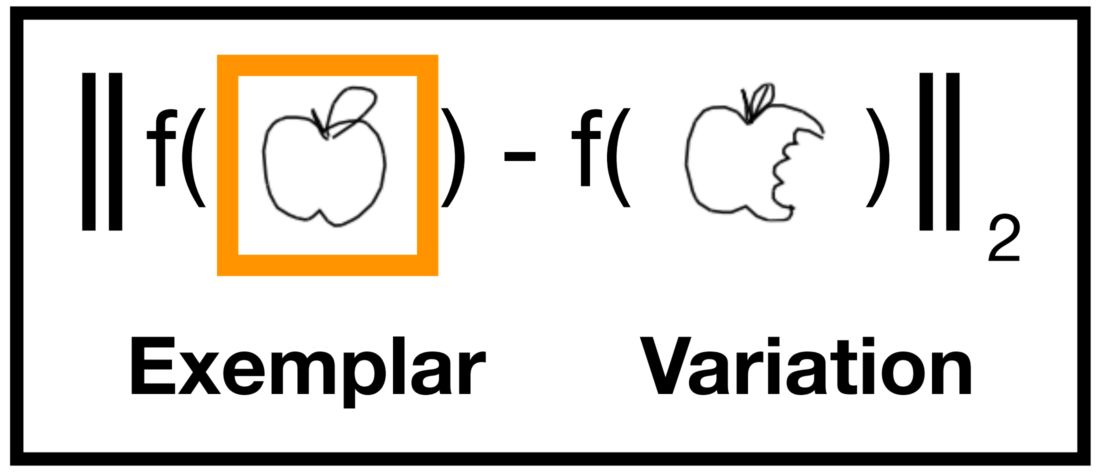
Task Evaluation : Originality vs Recognizability (Boutin et al 2022)



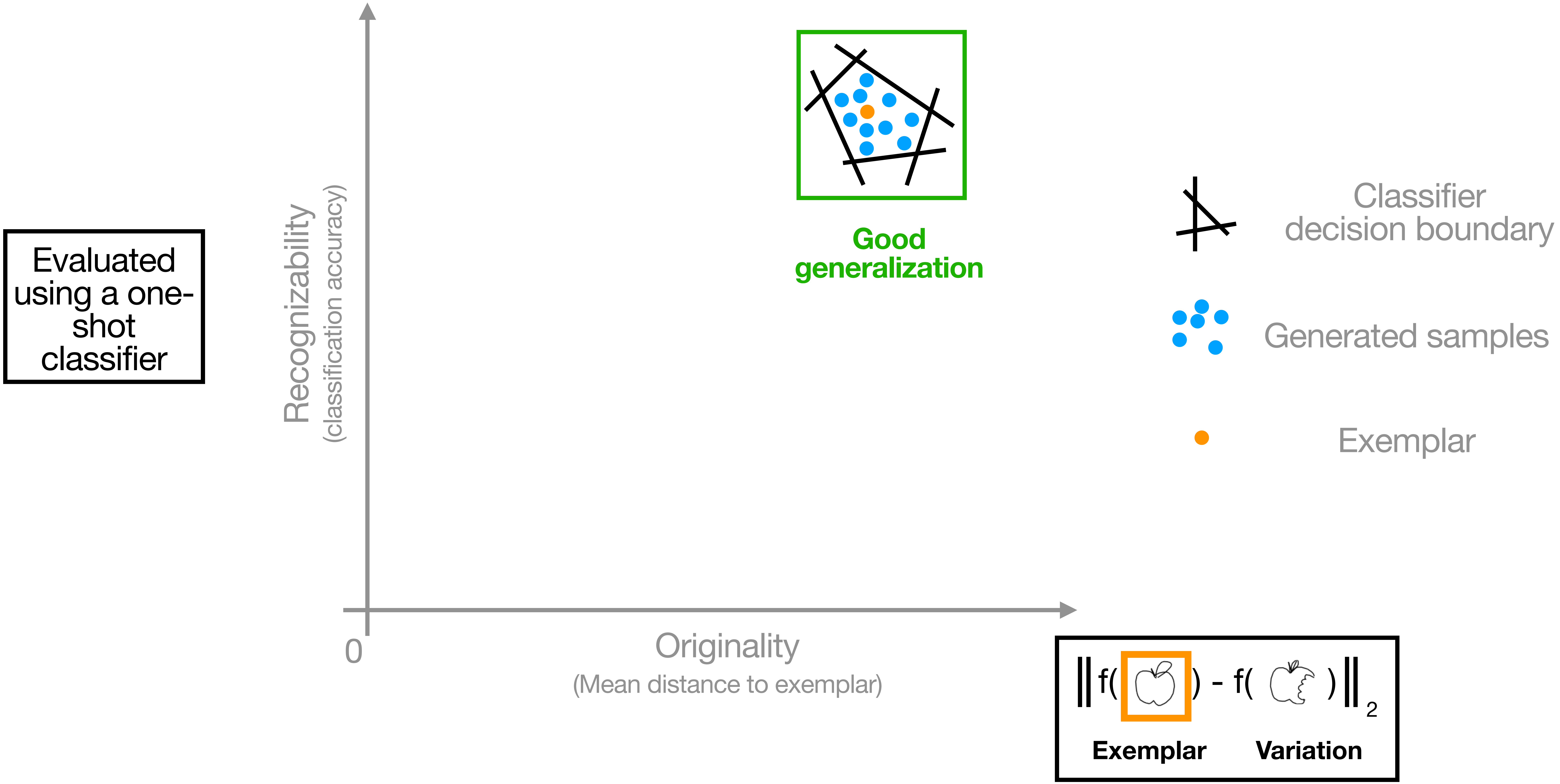
Task Evaluation : Originality vs Recognizability (Boutin et al 2022)

Evaluated using a one-shot classifier



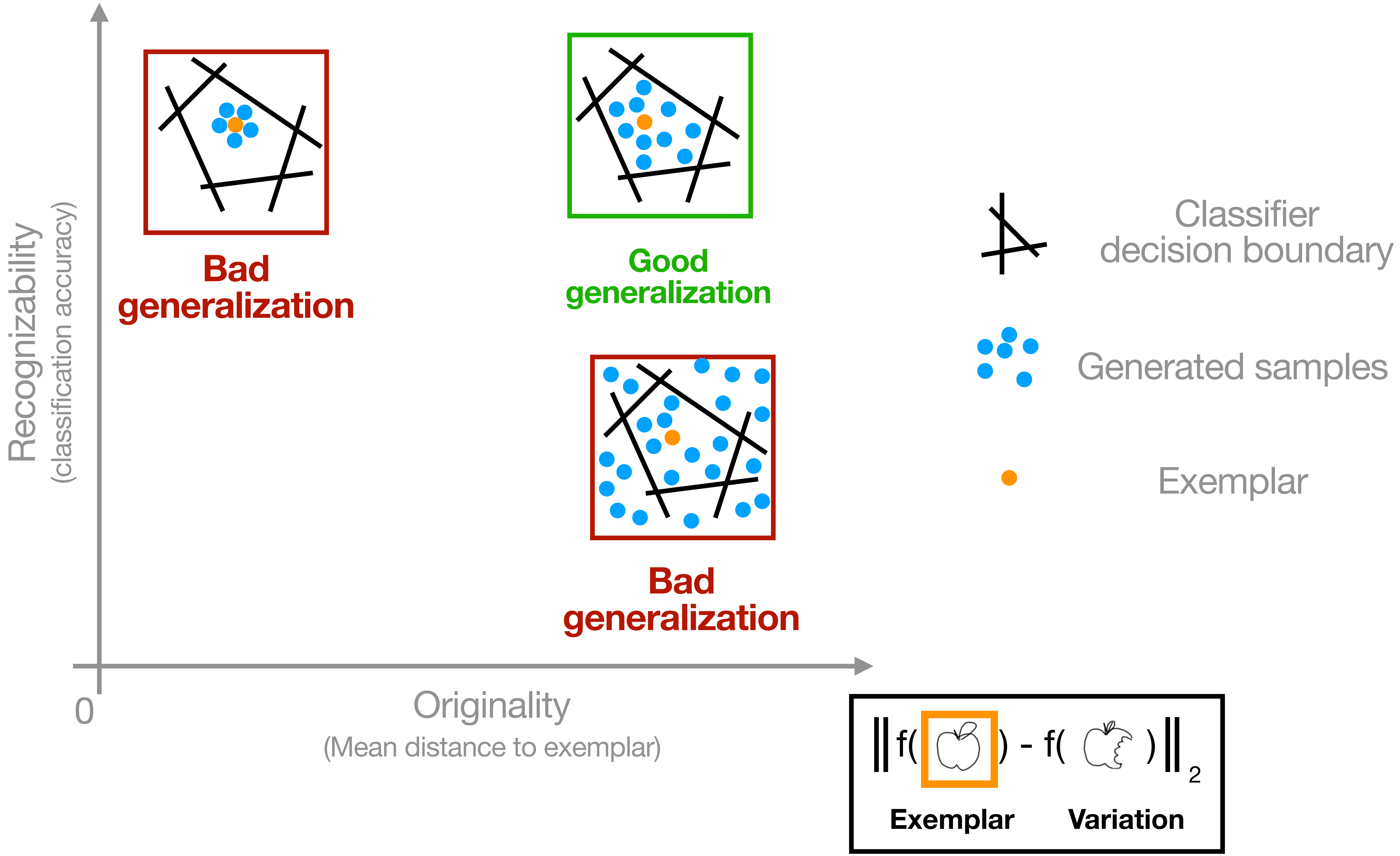


Task Evaluation : Originality vs Recognizability (BOUTIN ET AL 2022)

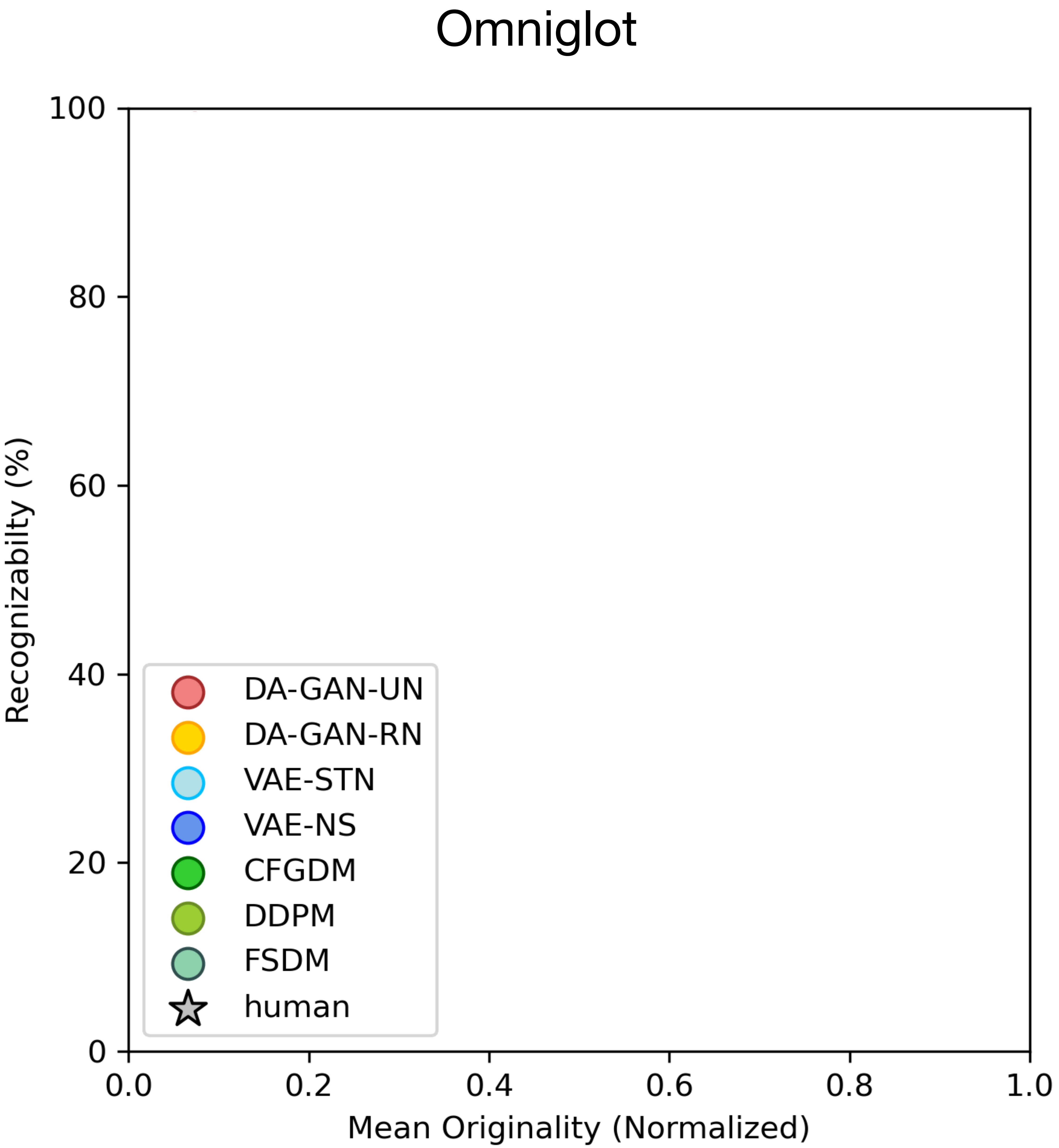


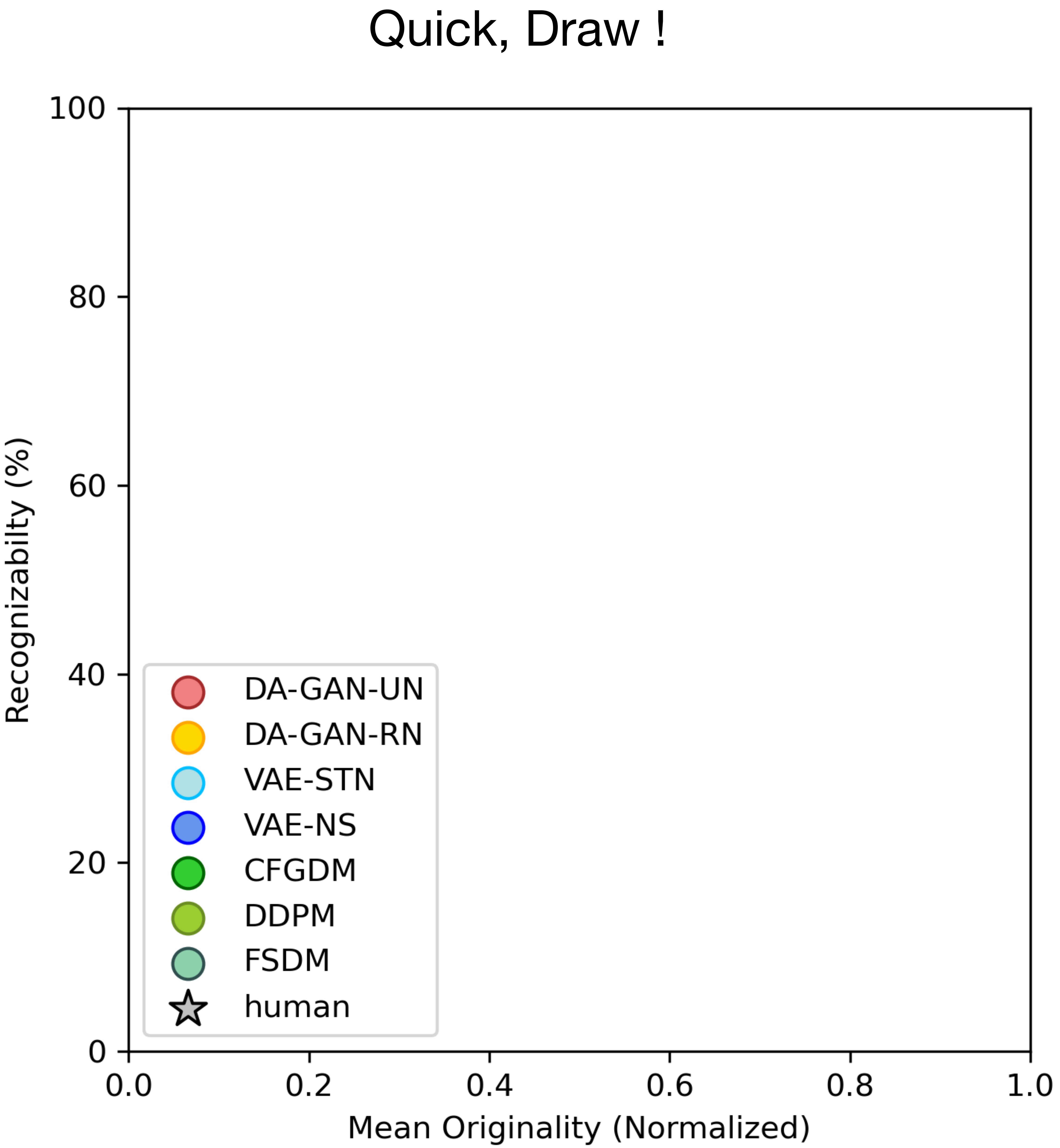
Task Evaluation : Originality vs Recognizability (BOUTIN ET AL 2022)

Evaluated using a one-shot classifier



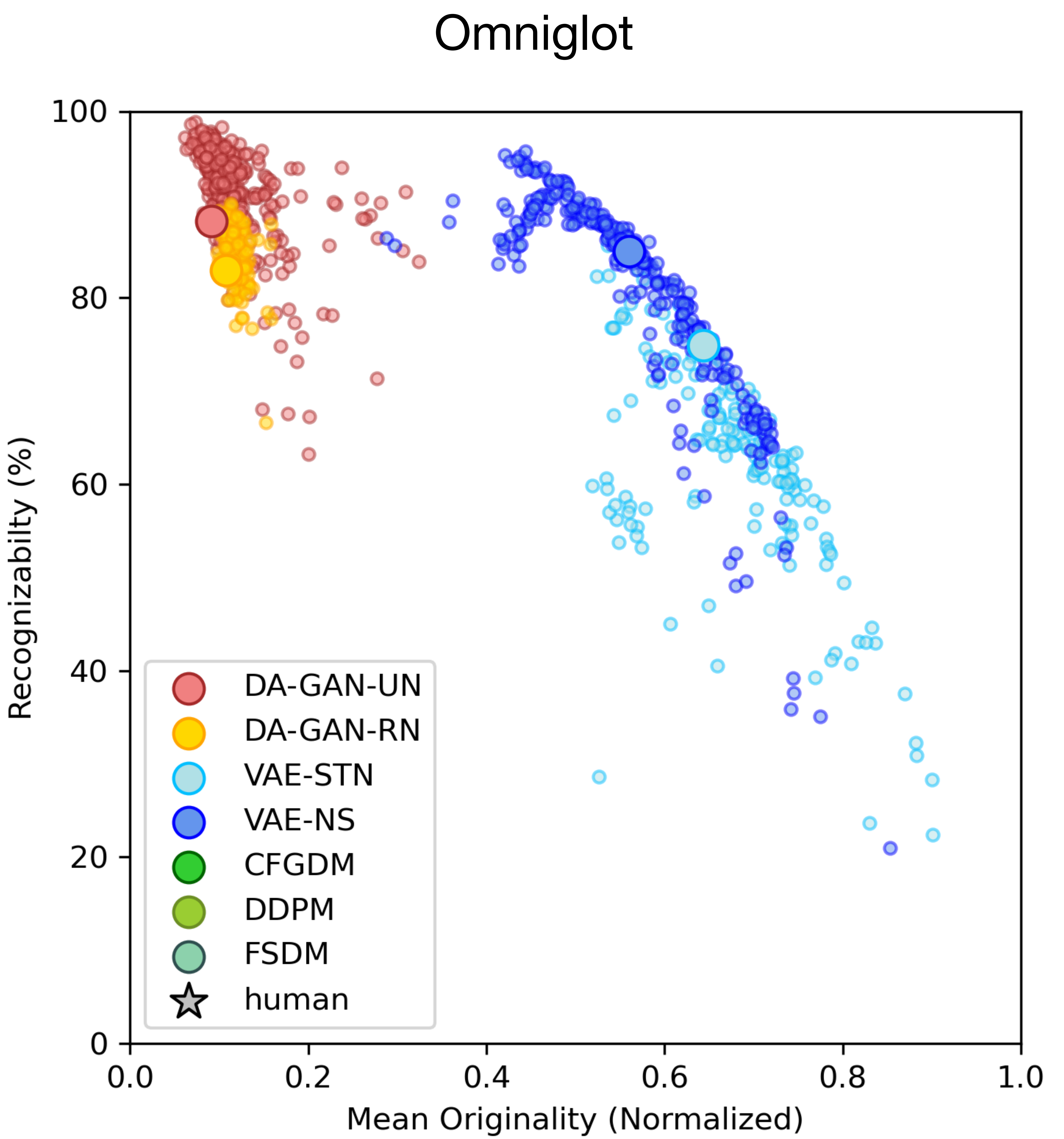
Models in the Originality vs. Recognizability Space (Boutin et al 2023)

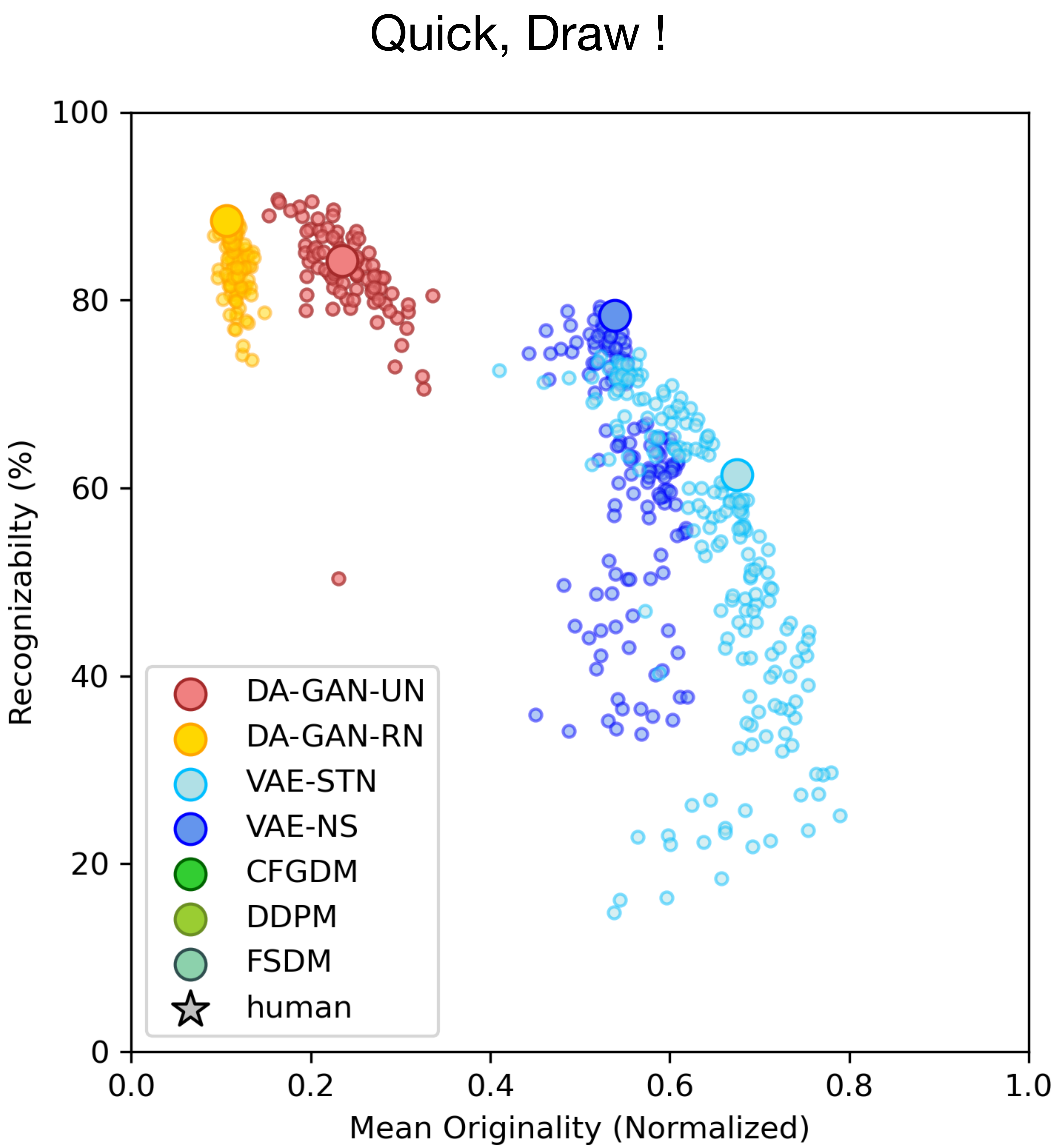




Models in the Originality vs. Recognizability Space (Boutin et al 2023)

Models in the Originality vs. Recognizability Space (BOUTIN ET AL 2023)





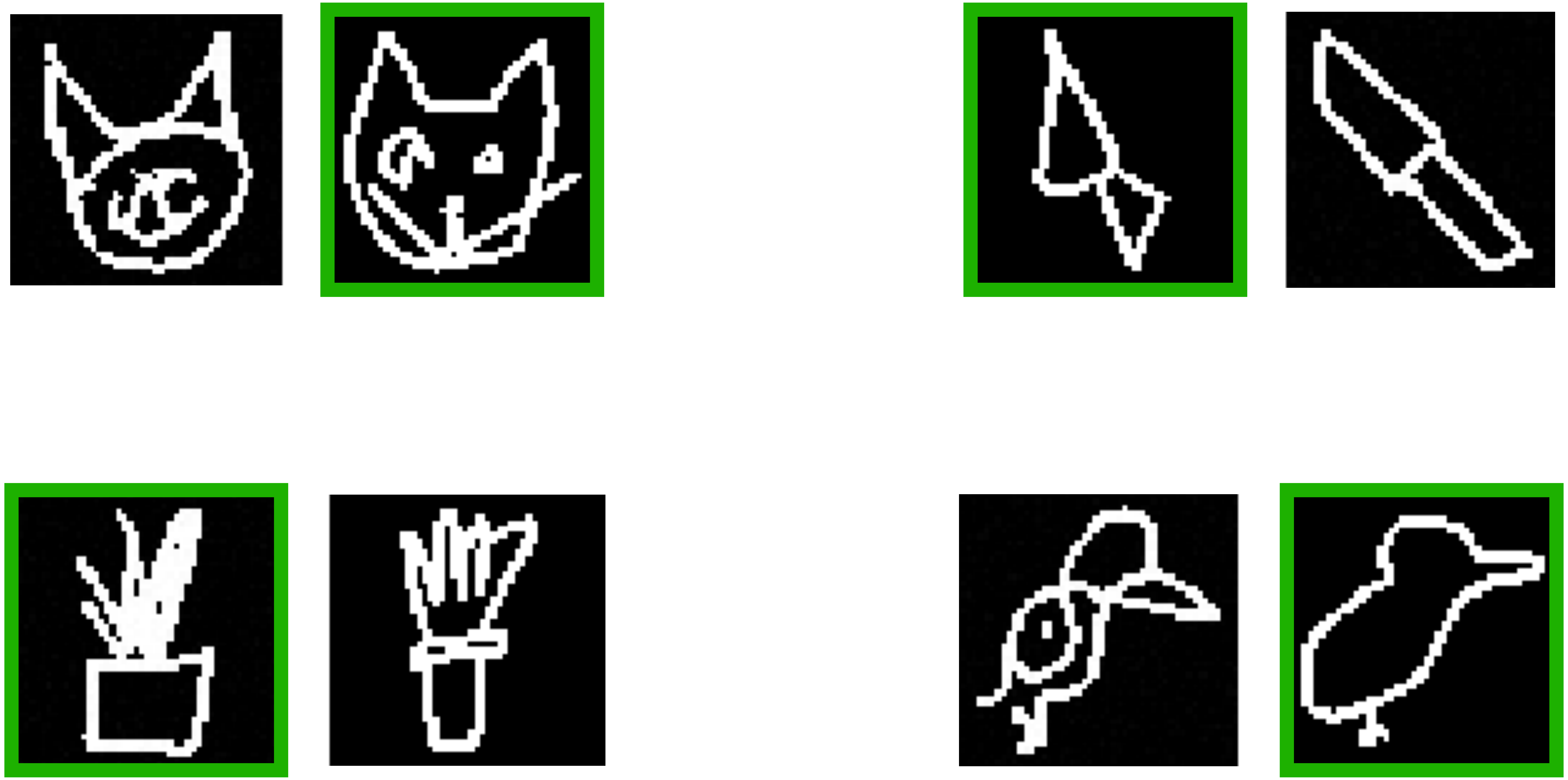
Models in the Originality vs. Recognizability Space (Boutin et al 2023)

Models in the Originality vs. Recognizability Space (Boutin et al 2023)

Models in the Originality vs. Recognizability Space (Boutin et al 2023)

Can you tell apart human from machine-generated samples ?

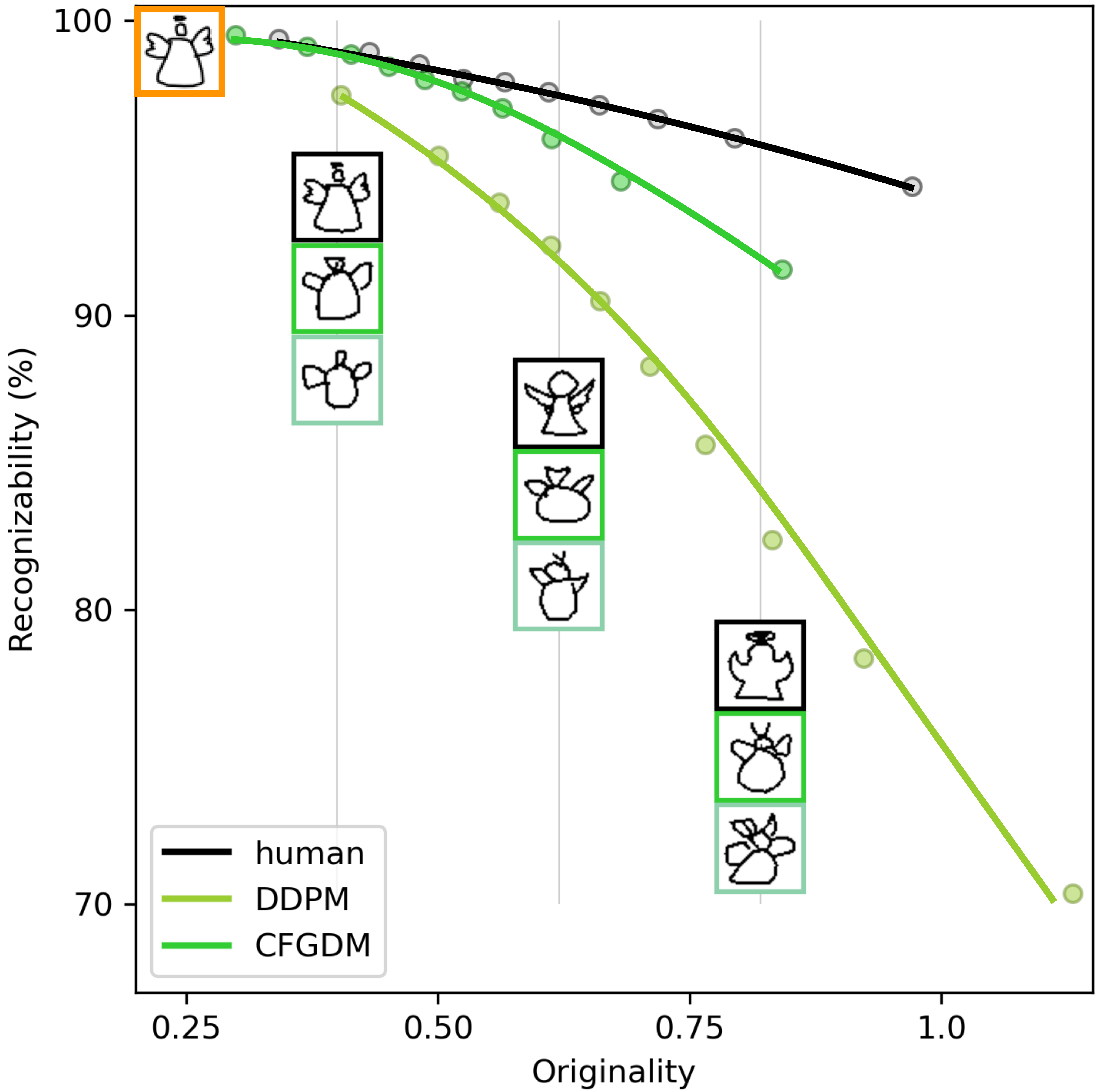
Can you tell apart human from machine-generated samples ?



Human

Generalization Curves : Recognizability = f(Originality)

Generalization Curves : Recognizability = f(Originality)



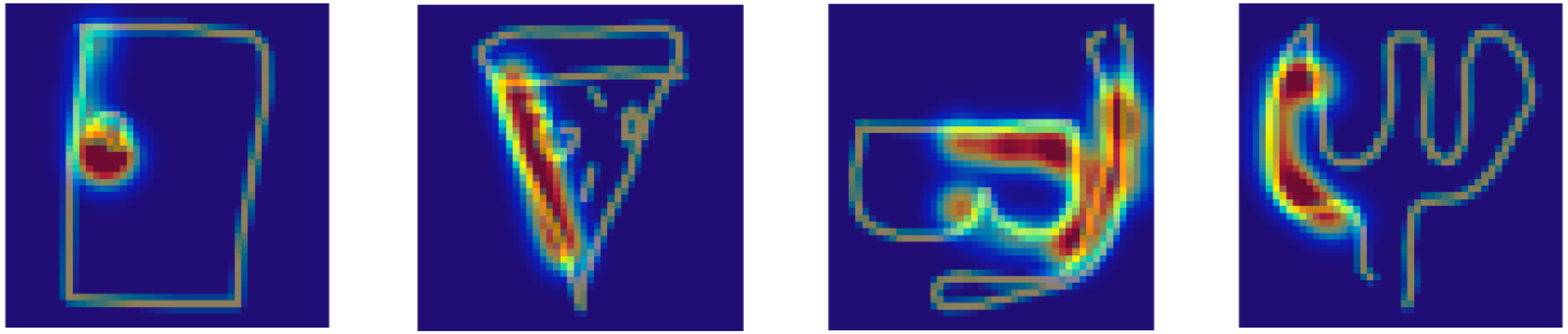
Important Features for Humans and Machines

Important Features for Humans and Machines

- Human importance maps collected using the ClickMe challenge (LINSLEY ET AL 2019)

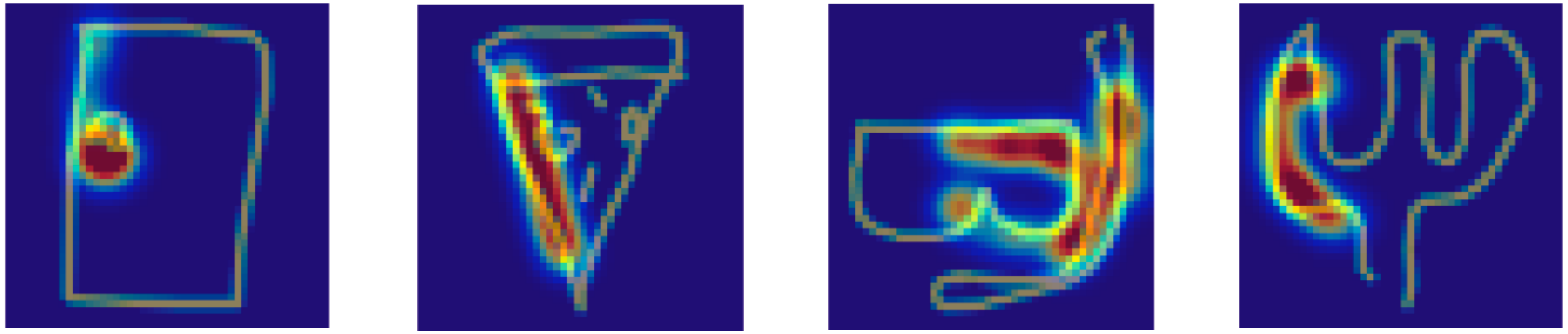
Important Features for Humans and Machines

- Human importance maps collected using the ClickMe challenge (LINSLEY ET AL 2019)



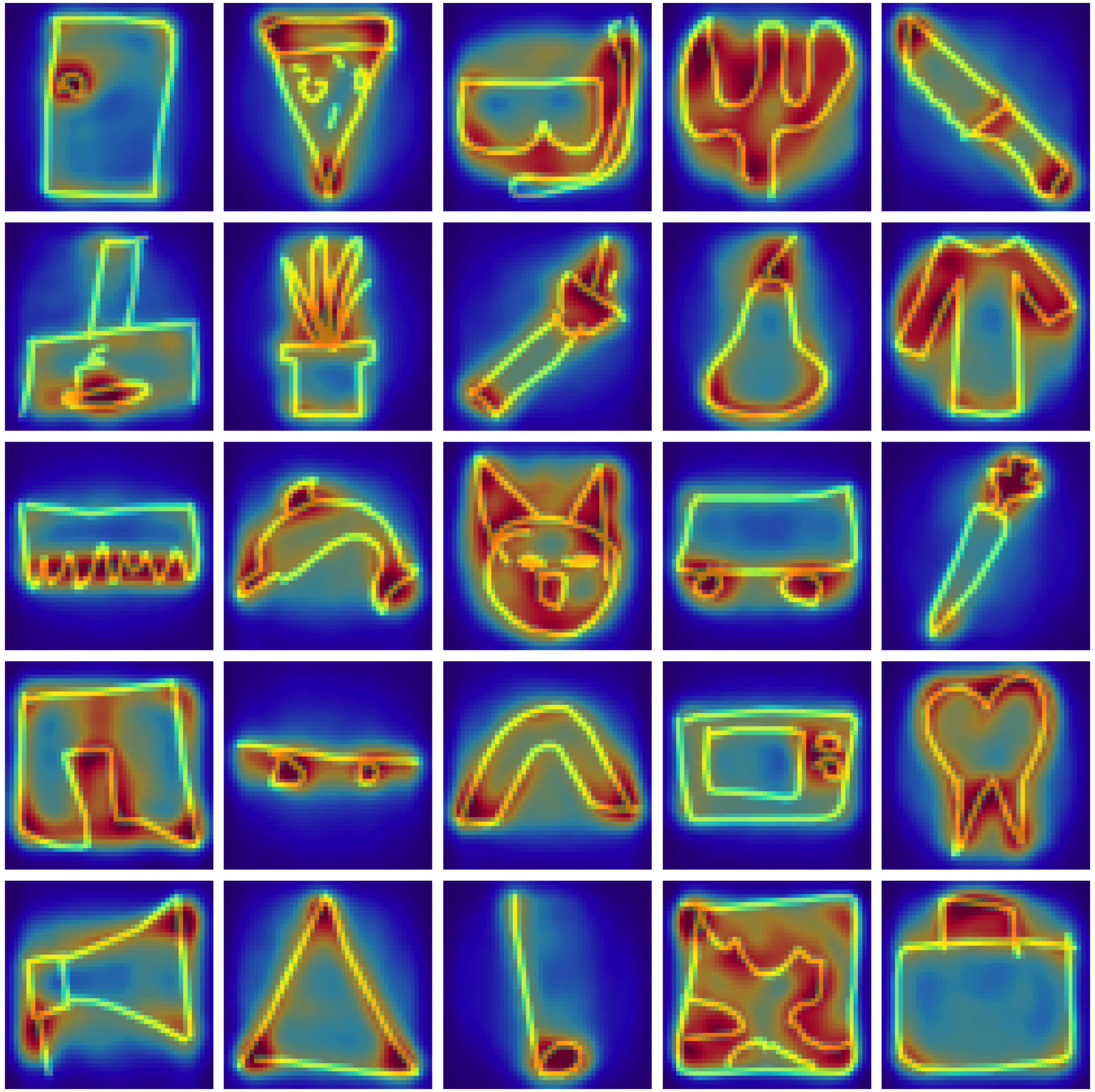
Important Features for Humans and Machines

- Human importance maps collected using the ClickMe challenge (LINSLEY ET AL 2019)

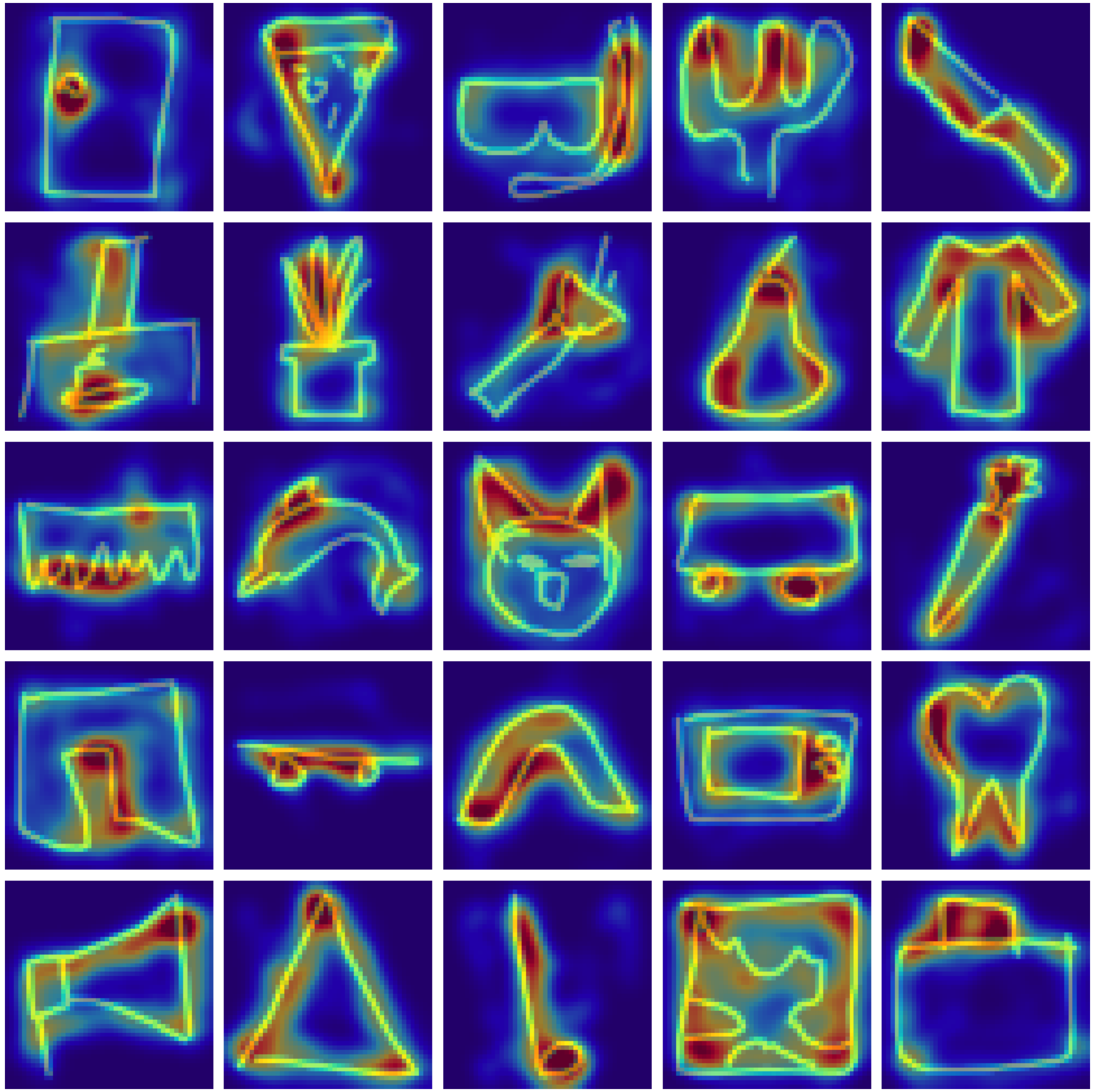


- **CFGDM** importance maps using attribution methods

Important Features for Humans and Machines



CFGDM



human

Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines

Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines

Diffusion models fail at modelling original drawings

Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines

Diffusion models fail at modelling original drawings

Different attentional strategies leveraged by humans and machines

Thank you for your attention …

2023

- Unlocking feature visualization for deeper networks with MAgnitude Constrained Optimization (Neurips 2023), T. Fel, T. Boissin, V. Boutin, A. Picard, P. Novello, J. Colin, D. Linsley, T. Rousseau, R. Cadène, L. Gardes & T. Serre
- A holistic approach to unifying automatic concept extraction and concept importance estimation (Neurips 2023), T. Fel, V. Boutin, M. Moayeri, R. Cadene, L. Bethune, L. Andeol, M. Chalvidal & T. Serre
- Learning functional transduction (Neurips 2023), M. Chalvidal, T. Serre & R.VanRullen
- Diffusion models as artists: Are we closing the gap between humans and machines? (ICML 2023), V. Boutin, T. Fel, L. Singhal, R. Mukherji, A. Nagaraj, J Colin & T. Serre
- CRAFT: Concept Recursive Activation FacTorization for explainability (CVPR 2023), T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadene & T. Serre
- GAMR: A Guided Attention Model for (visual) Reasoning (ICLR 2023), M Vaishnav & T. Serre

2022

- What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods (Neurips 2022), T. Fel, J. Colin, R. Cadene & T. Serre
- Harmonizing the object recognition strategies of deep neural networks with humans (Neurips 2022), T. Fel*, I.F. Rodriguez*, D. Linsley* & T. Serre
- A benchmark for compositional visual reasoning (Neurips 2022), A. Zerroug, M. Vaishnav, J. Colin, S. Musslick & T. Serre
- Diversity vs. recognizability: Human-like generalization in one-shot generative models (Neurips 2022), V. Boutin, L. Singhal, X. Thomas & T. Serre
- Meta-reinforcement learning with self-modifying networks (Neurips 2022), M. Chalvidal, T. Serre, R. VanRullen
- Understanding the computational demands underlying visual reasoning, (Neural Computation), M.Vaishnav, R. Cadene, A. Alamia, D. Linsley, R. VanRullen & T. Serre

2021

- Look at the variance! Efficient black-box explanations with Sobol-based sensitivity analysis (Neurips 2021), T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux & T. Serre.
- Go with the flow: Adaptive control for Neural ODEs (ICLR 2021), M. Chalvidal, M. Ricci, R. VanRullen, T. Serre
- Iterative VAE as a predictive brain model for out-of-distribution generalization (Neurips workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM)), V. Boutin, A. Zerroug, M. Jung, & Thomas Serre