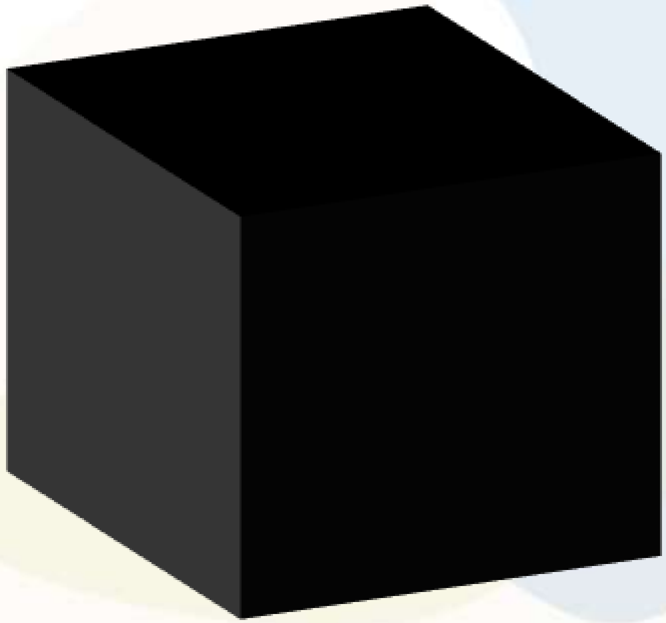
Looking Inside the Black Box

Andrew Doyle
@crocodoyle

McGill Centre for Integrative Neuroscience

Danger!

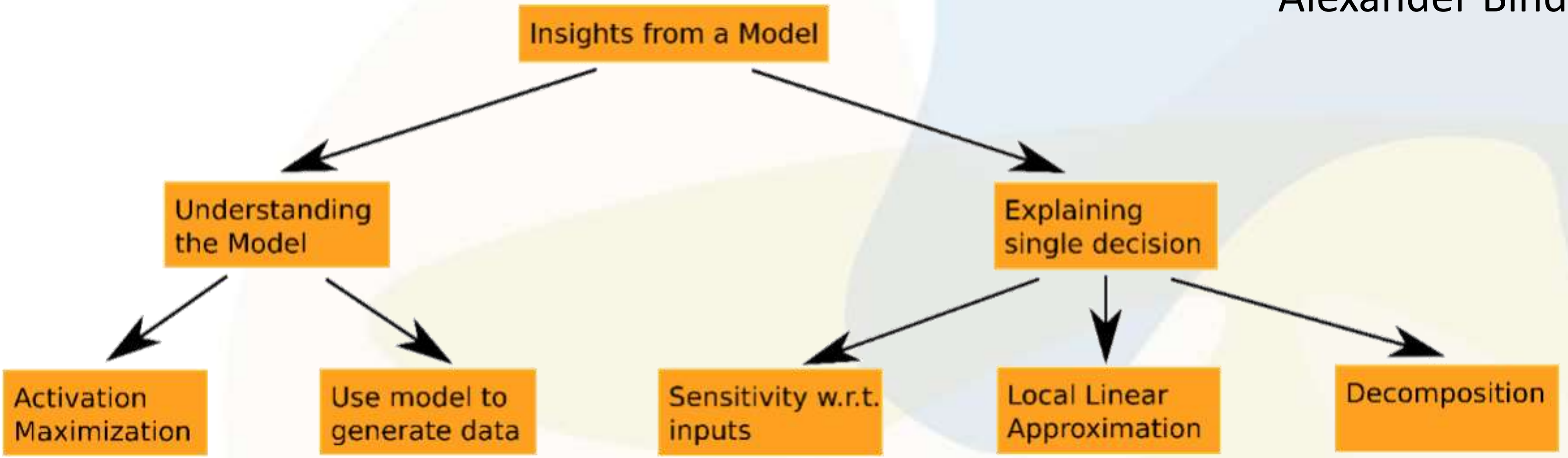


If you gaze long into an abyss, the abyss also gazes into you.

Nietzsche, Friedrich. “Beyond Good and Evil.” 1886.

Interpretability

Alexander Binder



Binder, Alexander. "Explaining Decisions of Neural Networks and Layer-wise Relevance Propagation" OHBM 2018. https://www.pathlms.com/ohbm/courses/8246/sections/12542/video_presentations/115841

Ask GPT-2

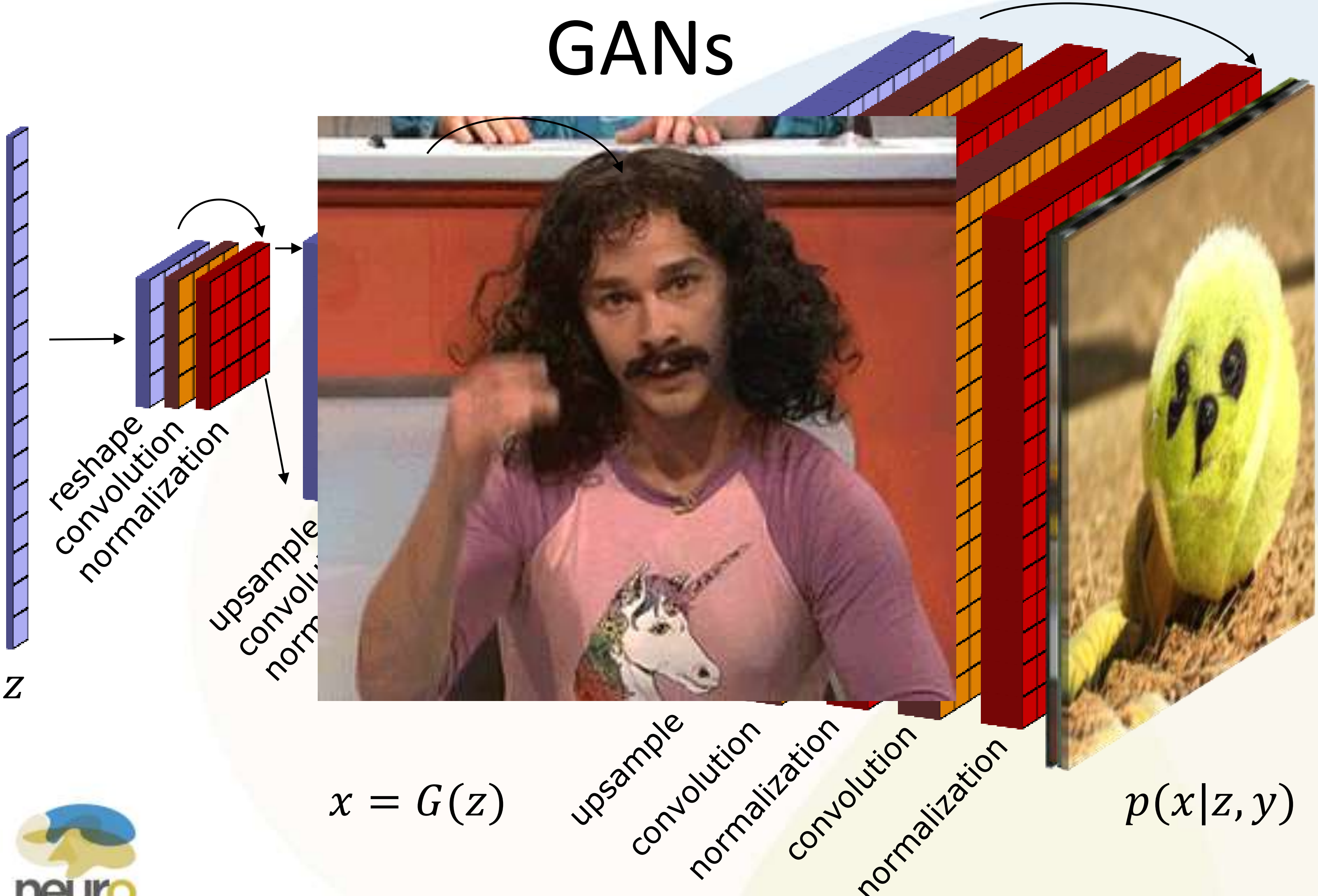
Q: How can we interpret the predictions of deep neural networks?

A: Deep learning is not perfect (and there's nothing stopping us from improving the algorithm), but the current state in the marketplace implies that machine learning is being more widely adopted than it has in the past. If you think about it - how important is it to get a computer to do all the computations, while leaving the user and content completely unaware of what's going on? This is probably exactly the goal of deep learning!

talktotransformer.com/

github.com/openai/gpt-2

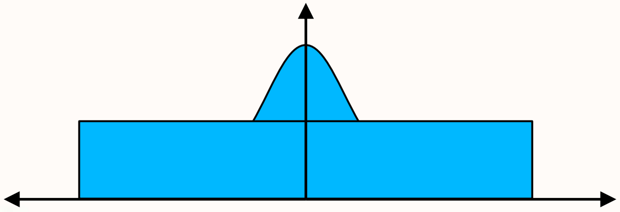
GANs



Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis." *arXiv preprint arXiv:1809.11096* (2018).

BigGAN

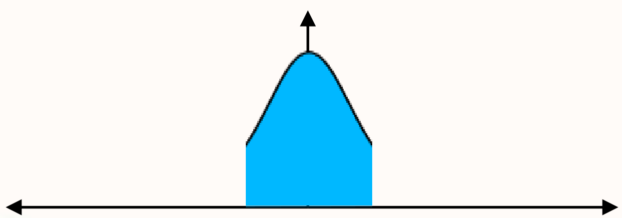
$p(z)$



Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis." *arXiv preprint arXiv:1809.11096* (2018).

BigGAN

$p(z)$



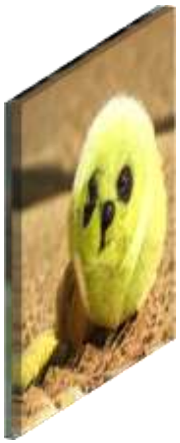
Truncation Trick

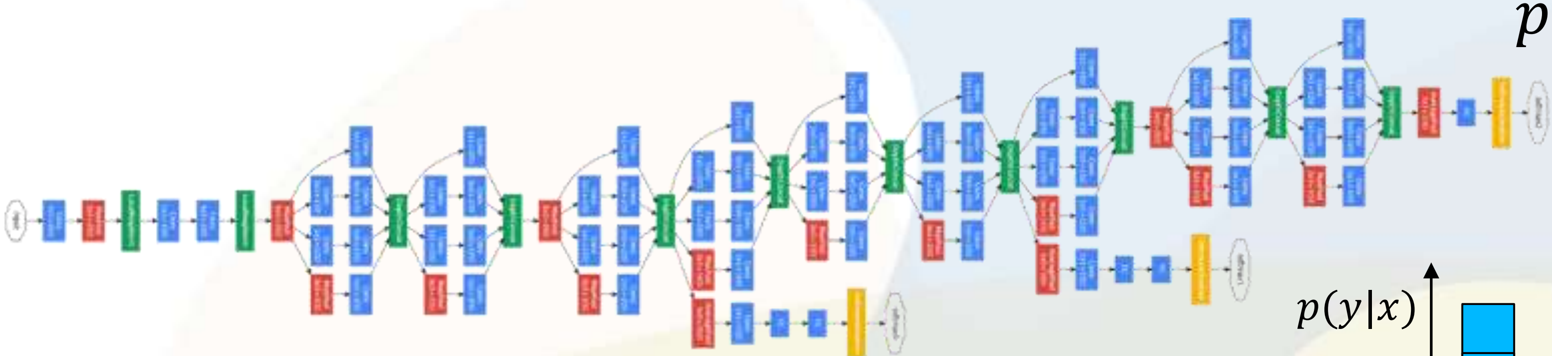
colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/biggan_generation_with_tf_hub.ipynb

Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis." *arXiv preprint arXiv:1809.11096* (2018).

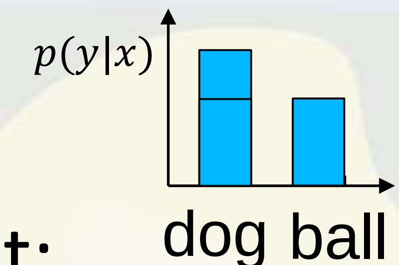
Inception Score

$x = G(z)$





$p(y|x)$



Test generated samples in InceptionNet:

- Low entropy in $p(y|x)$
- High entropy in $\int p(y|x = G(z))dz = p(y)$

$$IS(x) = e^{\mathbb{E}_x \boldsymbol{KL}(\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})||\boldsymbol{p}(\boldsymbol{y}))}$$

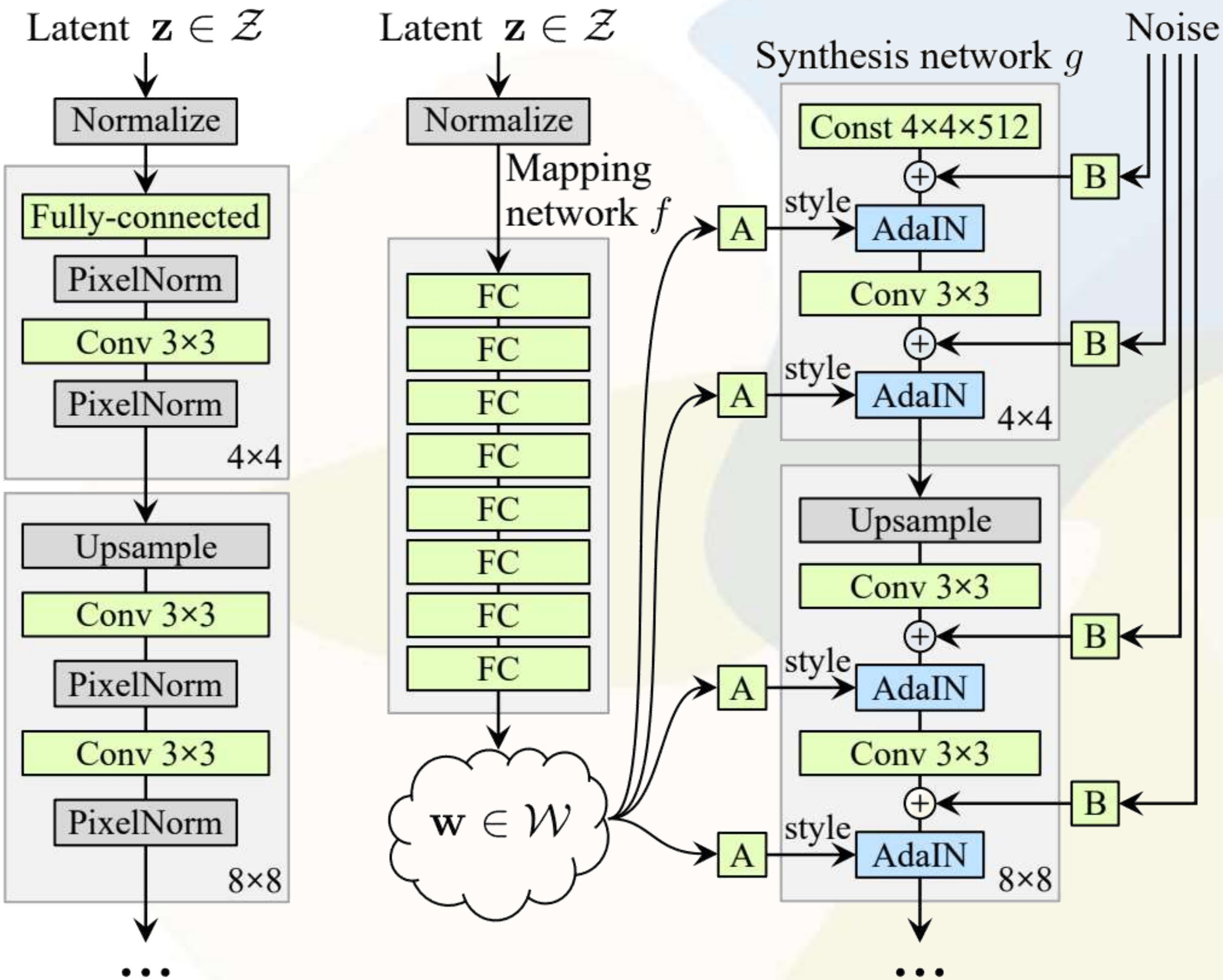
Salimans, Tim, et al. "Improved techniques for training GANs." NIPS, 2016.

Fréchet Inception Distance

$$FID(x) = (\mu_r - \mu_g)^2 + Tr\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right)$$

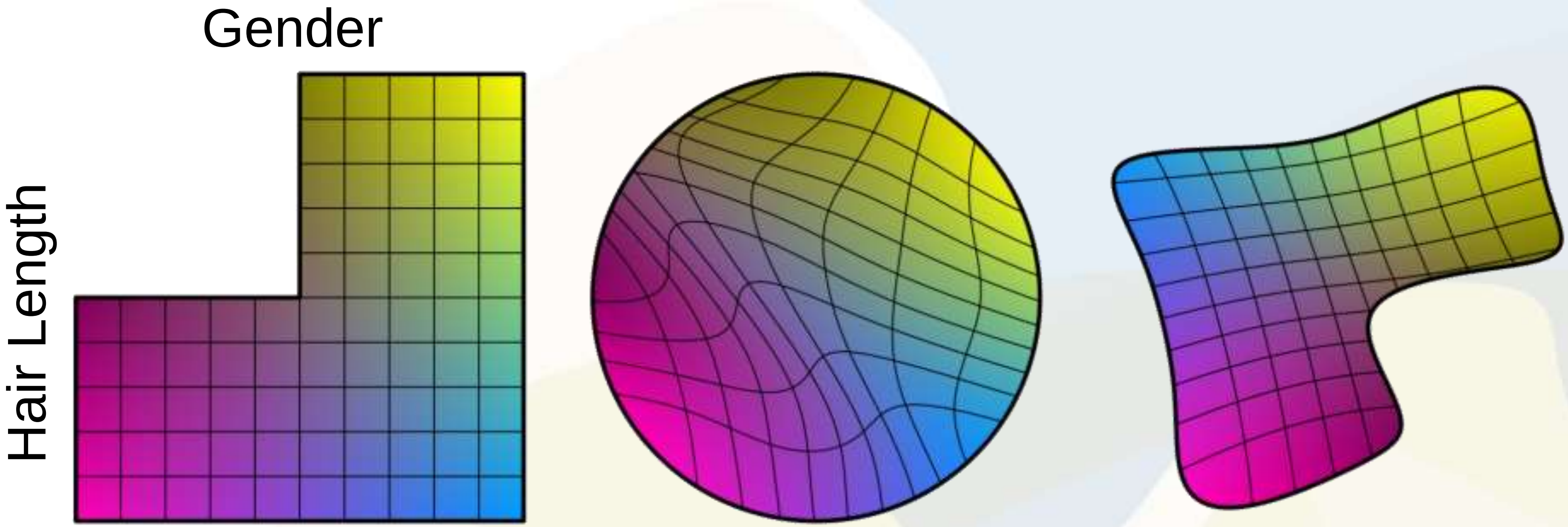
Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium." *Advances in Neural Information Processing Systems*. 2017.

StyleGAN



Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *arXiv preprint arXiv:1812.04948* (2018).

StyleGAN



(a) Distribution of features in training set

(b) Mapping from $\mathcal{Z}$ to features

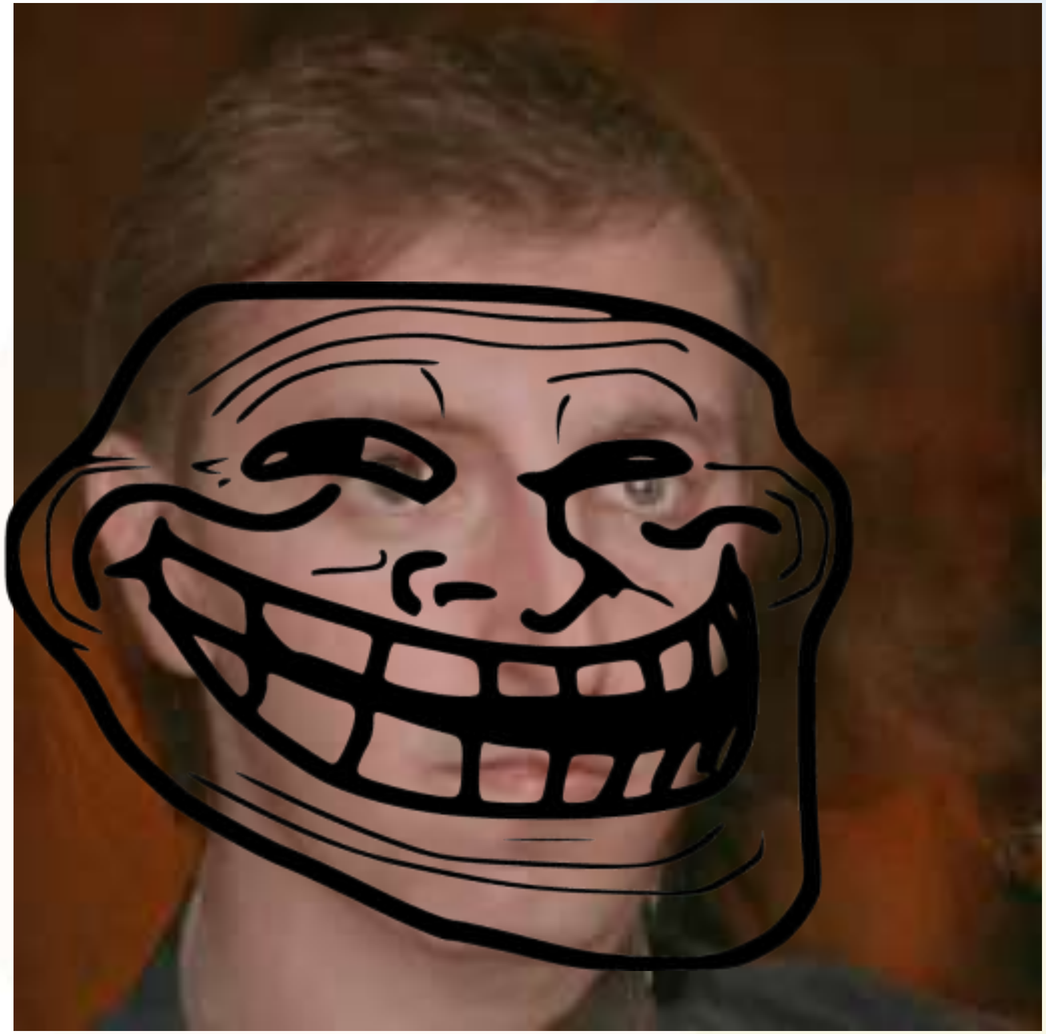
(c) Mapping from $\mathcal{W}$ to features

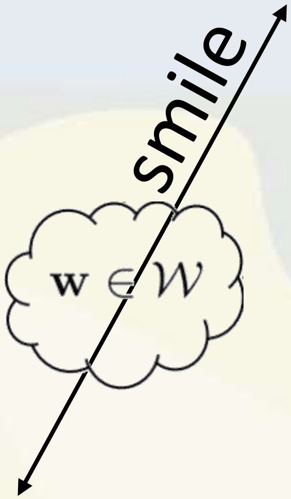
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *arXiv preprint arXiv:1812.04948* (2018).

StyleGAN

Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *arXiv preprint arXiv:1812.04948* (2018).

StyleGAN



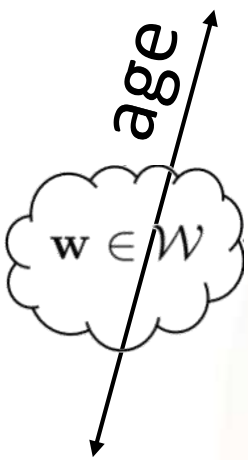


github.com/Puzer/stylegan-encoder

StyleGAN

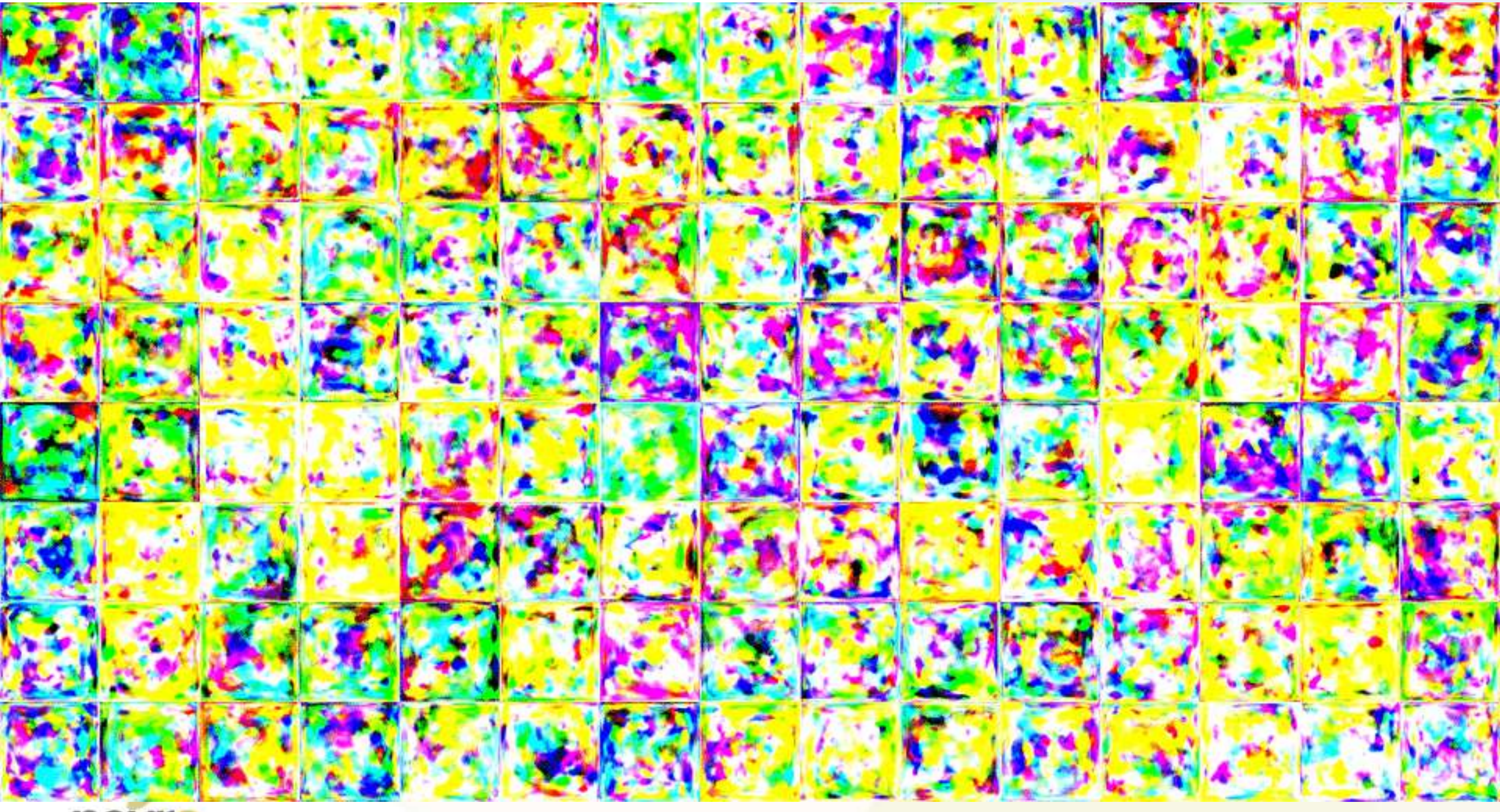
github.com/Puzer/stylegan-encoder

StyleGAN



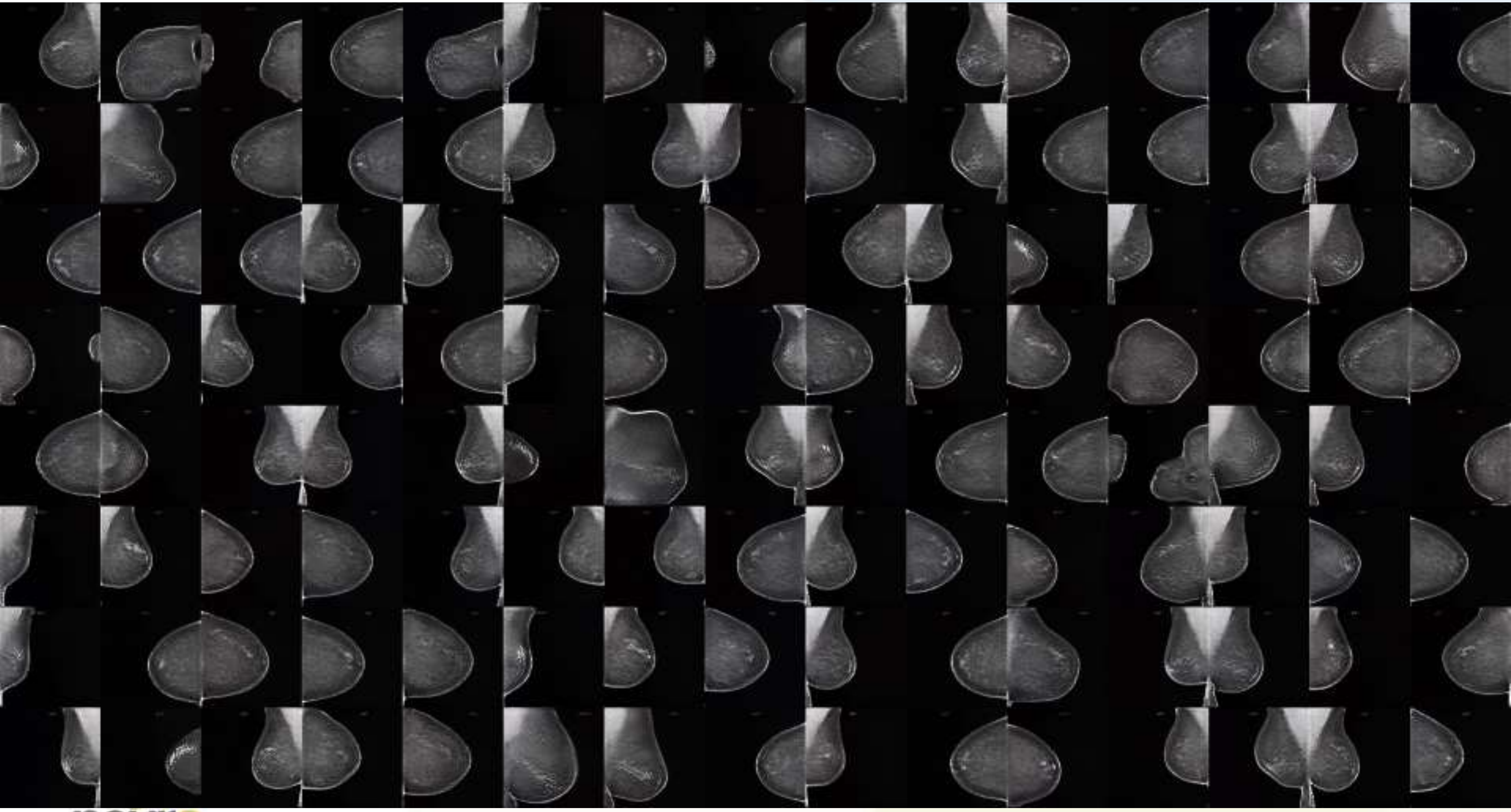
github.com/Puzer/stylegan-encoder

StyleGAN



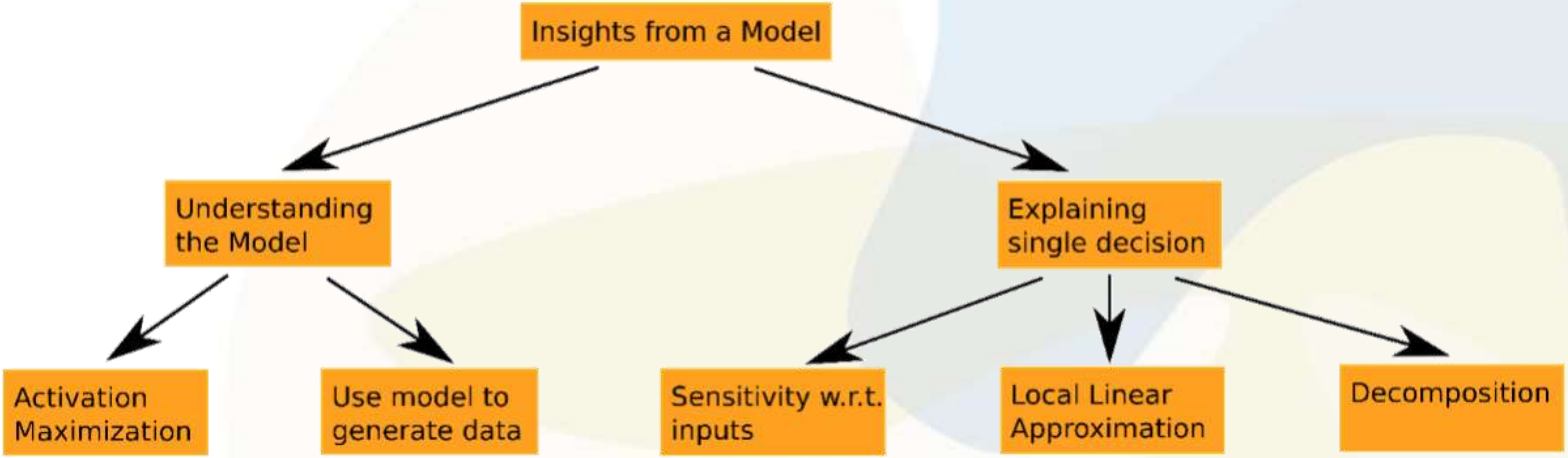
Mo et al. “What is the Effectiveness of Synthetic GAN Mammographic Images for Training a Breast Density DCNN Model?” RSNA 2019

StyleGAN



Mo et al. “What is the Effectiveness of Synthetic GAN Mammographic Images for Training a Breast Density DCNN Model?” RSNA 2019

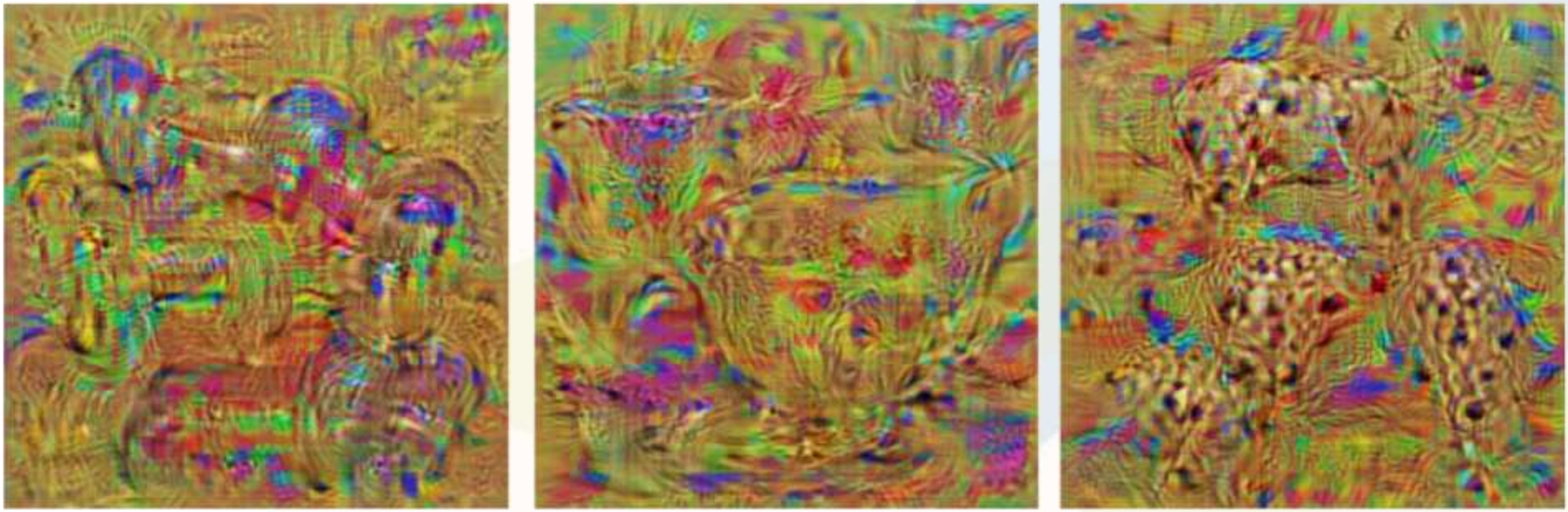
Interpretability



BigGAN
StyleGAN

Class Appearance Models

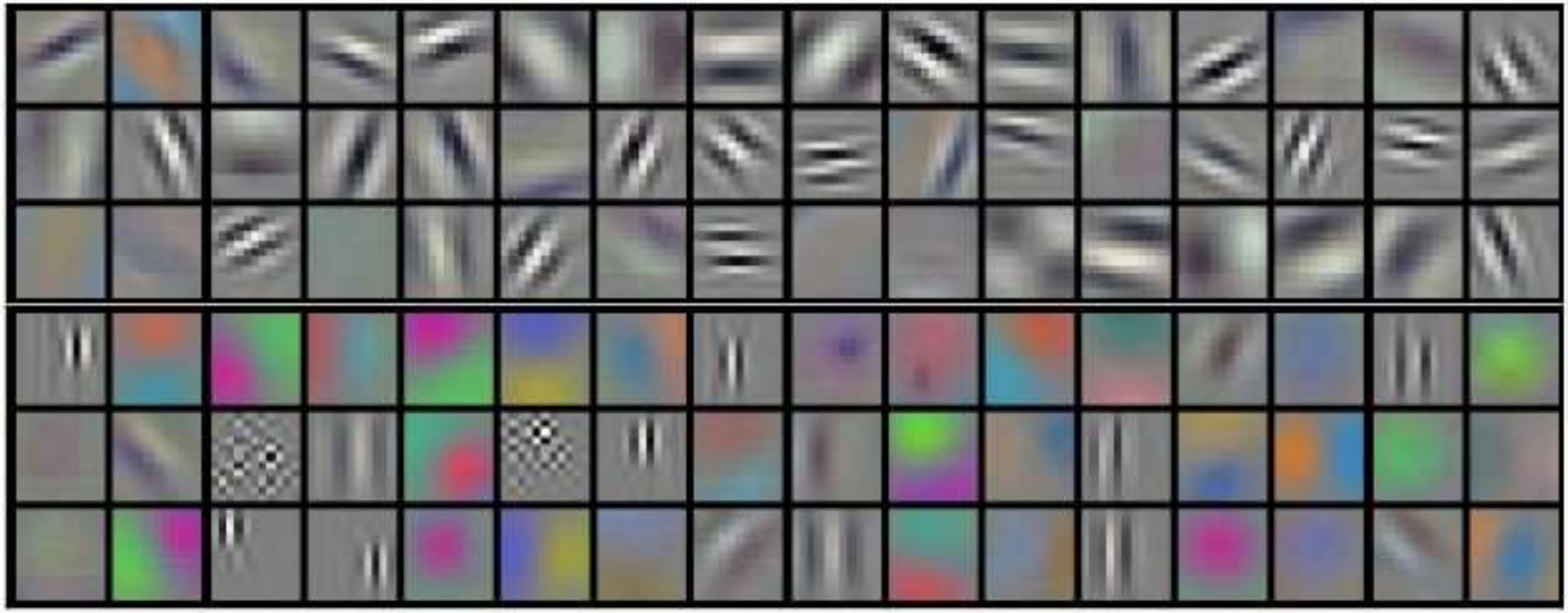
$$\underset{x}{\operatorname{argmax}}\, p(y = c|x) - \lambda||x||_2^2$$



dumbbell **cup** **dalmatian**

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013)

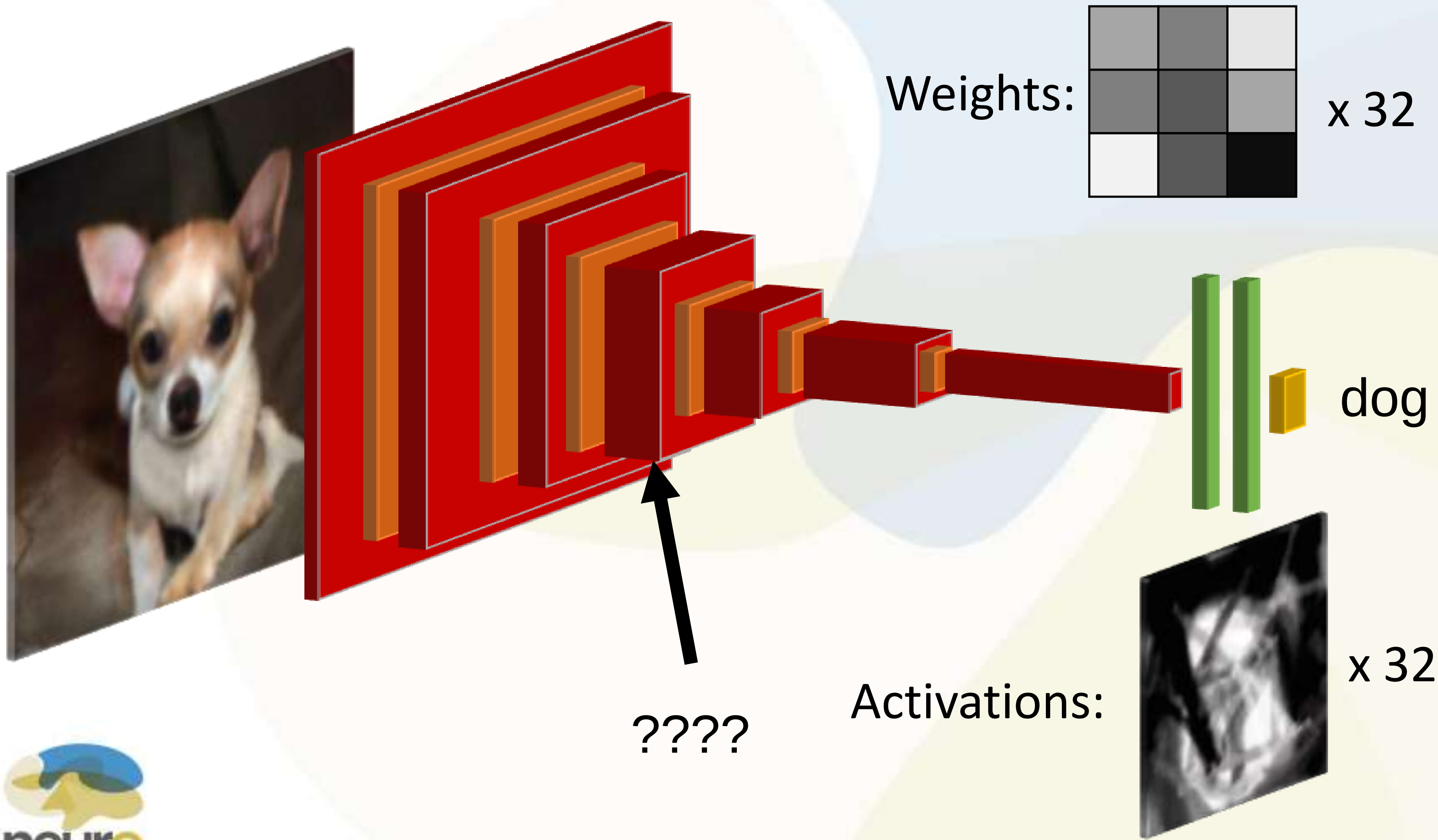
Visualizing Filters



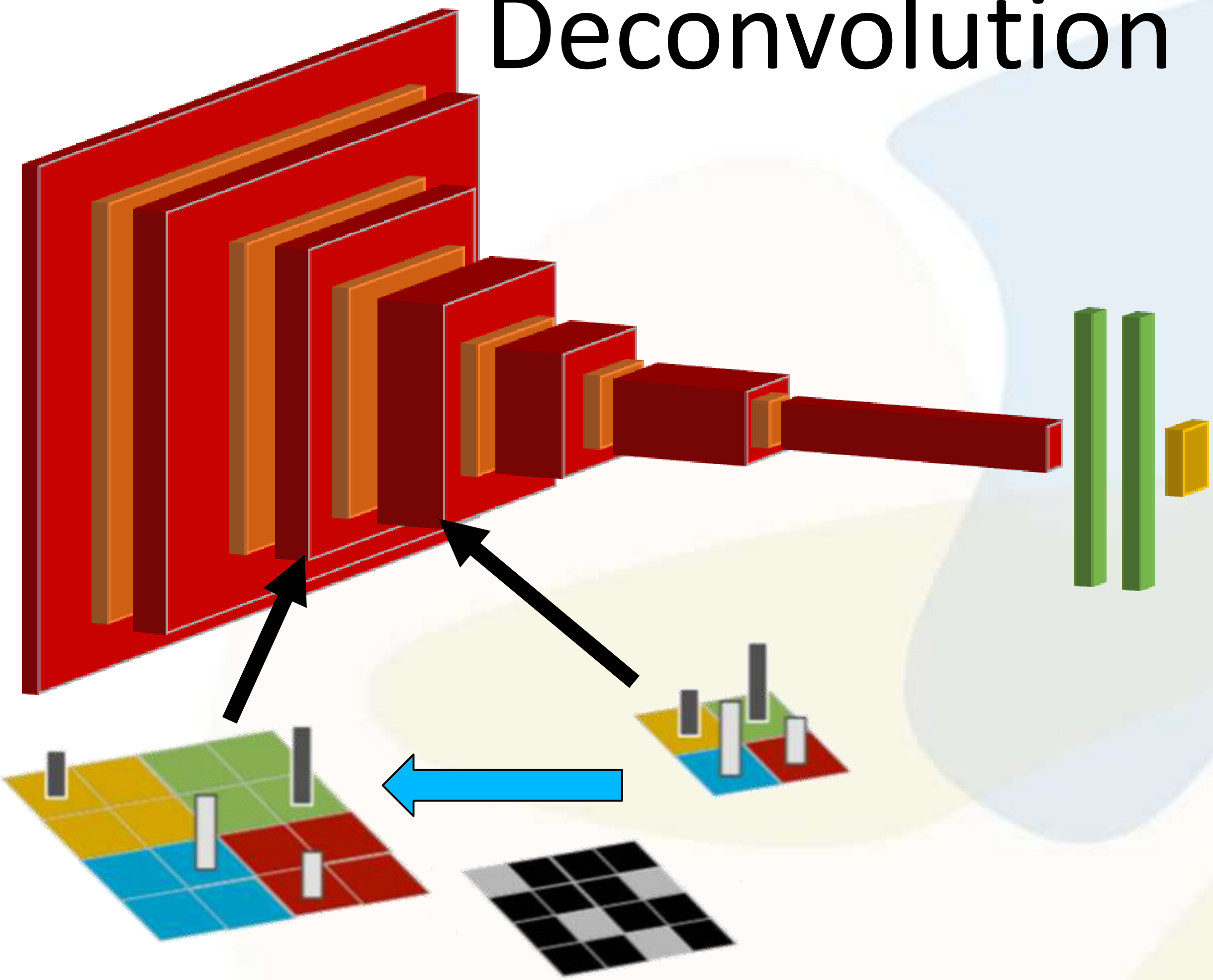
Filter weights for layer 1 of AlexNet

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Visualizing Filters



Deconvolution



Saved max pool locations

What input x does this filter respond to?

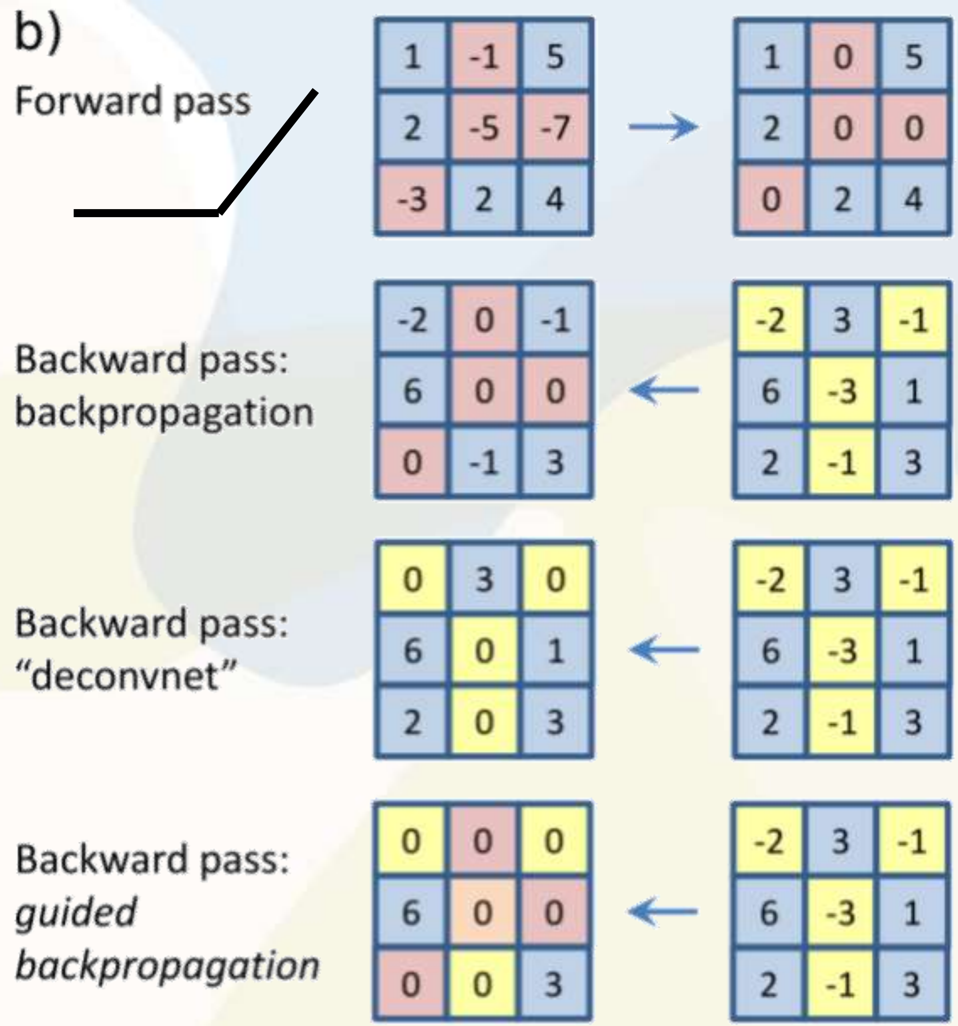
Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

Deconvolution

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

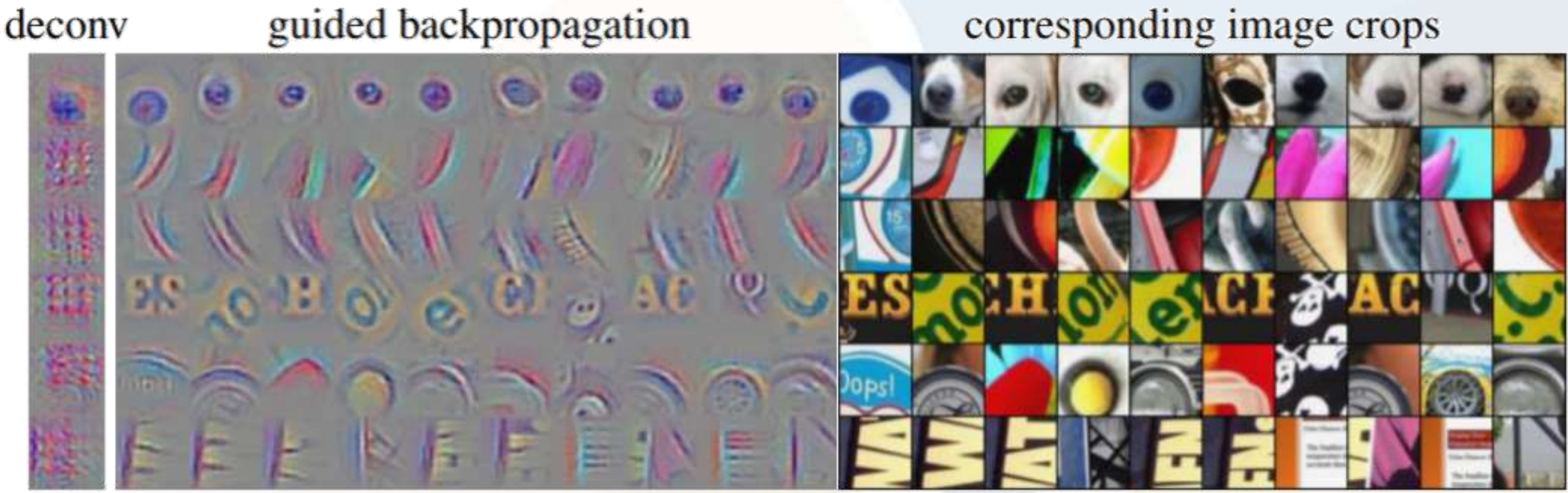
Guided Backprop

- Deconvolution fails without max pooling!
- Guided backprop changes how **ReLU** is handled



Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." *arXiv preprint arXiv:1412.6806* (2014).

Guided Backprop

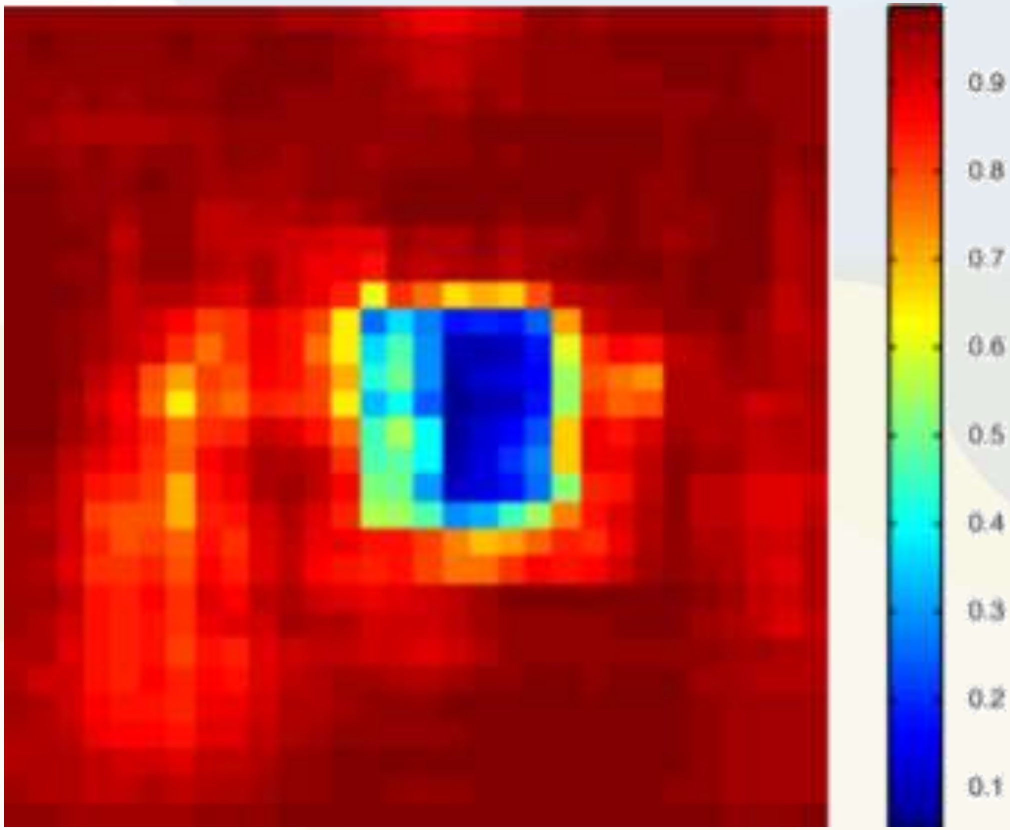


Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." *arXiv preprint arXiv:1412.6806* (2014).

Input Importance

Occlusion Testing

x

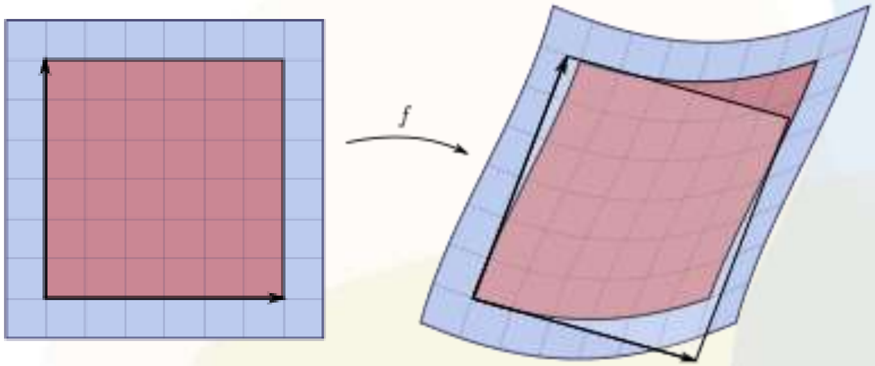


p(y=pomeranian | x) when occluded

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

Saliency

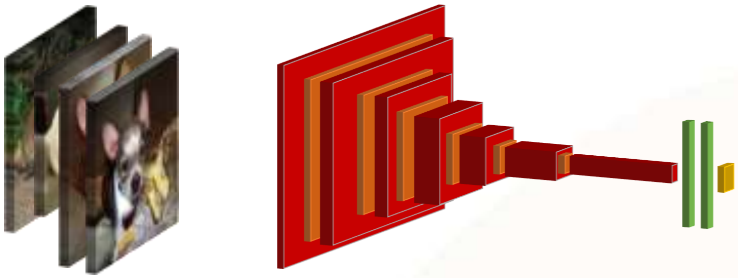
- Jacobian:



$$J = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial x_1} & \cdots & \frac{\partial \hat{y}_1}{\partial x_k} \\ \cdots & & \cdots \\ \frac{\partial \hat{y}_m}{\partial x_1} & \cdots & \frac{\partial \hat{y}_m}{\partial x_k} \end{bmatrix}$$

- Best linear approximation for non-linear mapping $p(y|x)$ near $\boldsymbol{x}$

Backprop

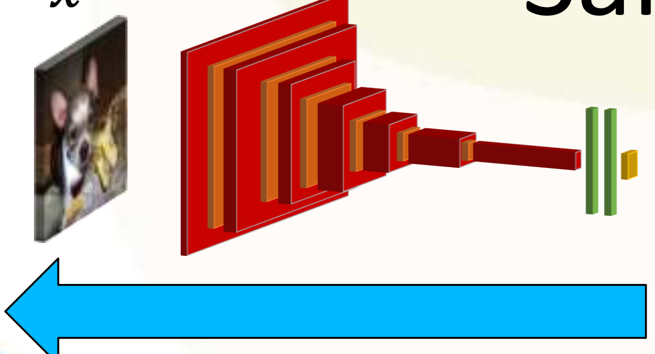


$$\hat{y} = p(y|x) \longrightarrow L(y, \hat{y})$$

$$\nabla_\theta L(y, \hat{y}) = \left[\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \dots, \frac{\partial L}{\partial w_n}\right]^T$$

Saliency

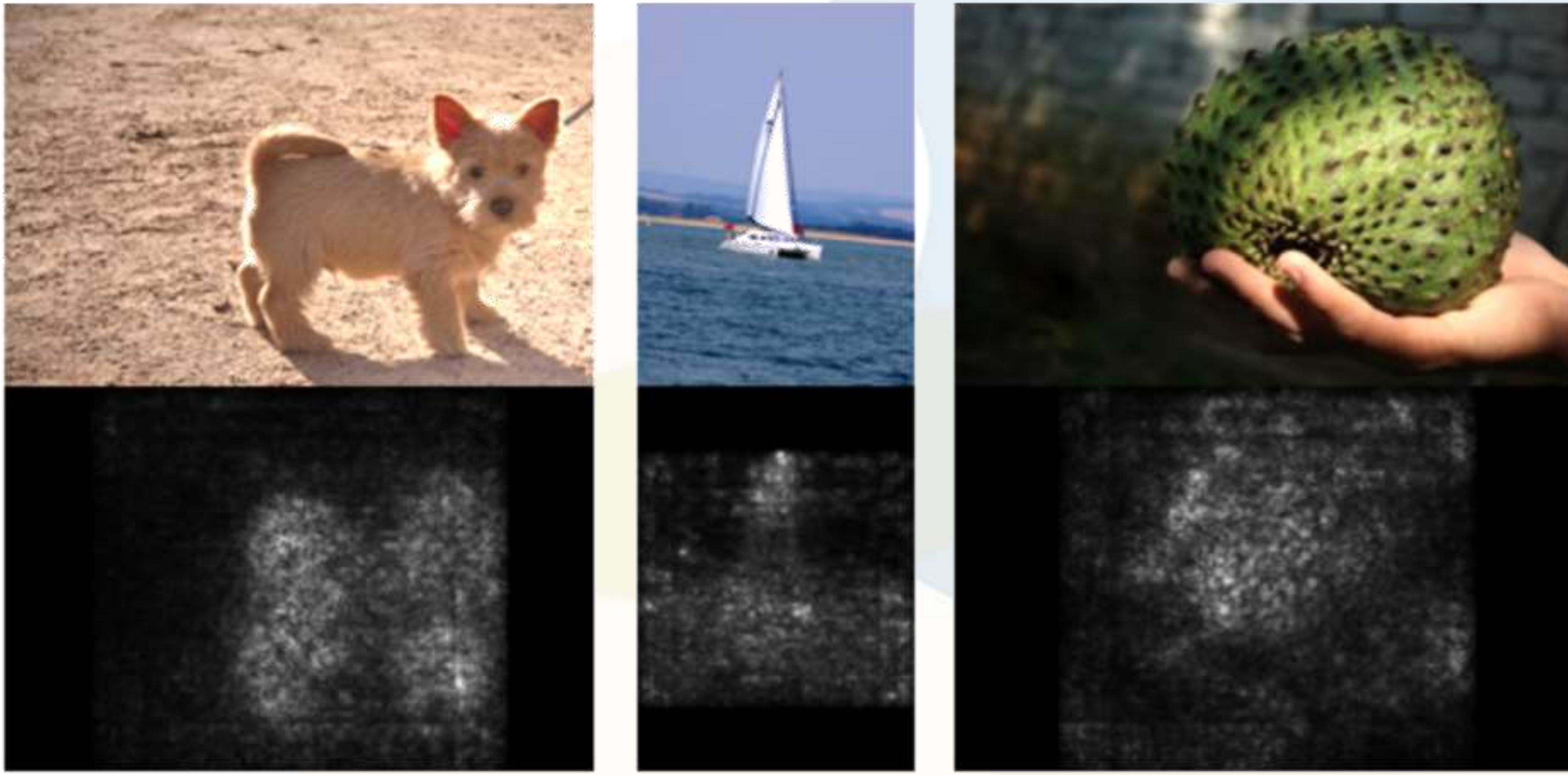
x



$\hat{y}$

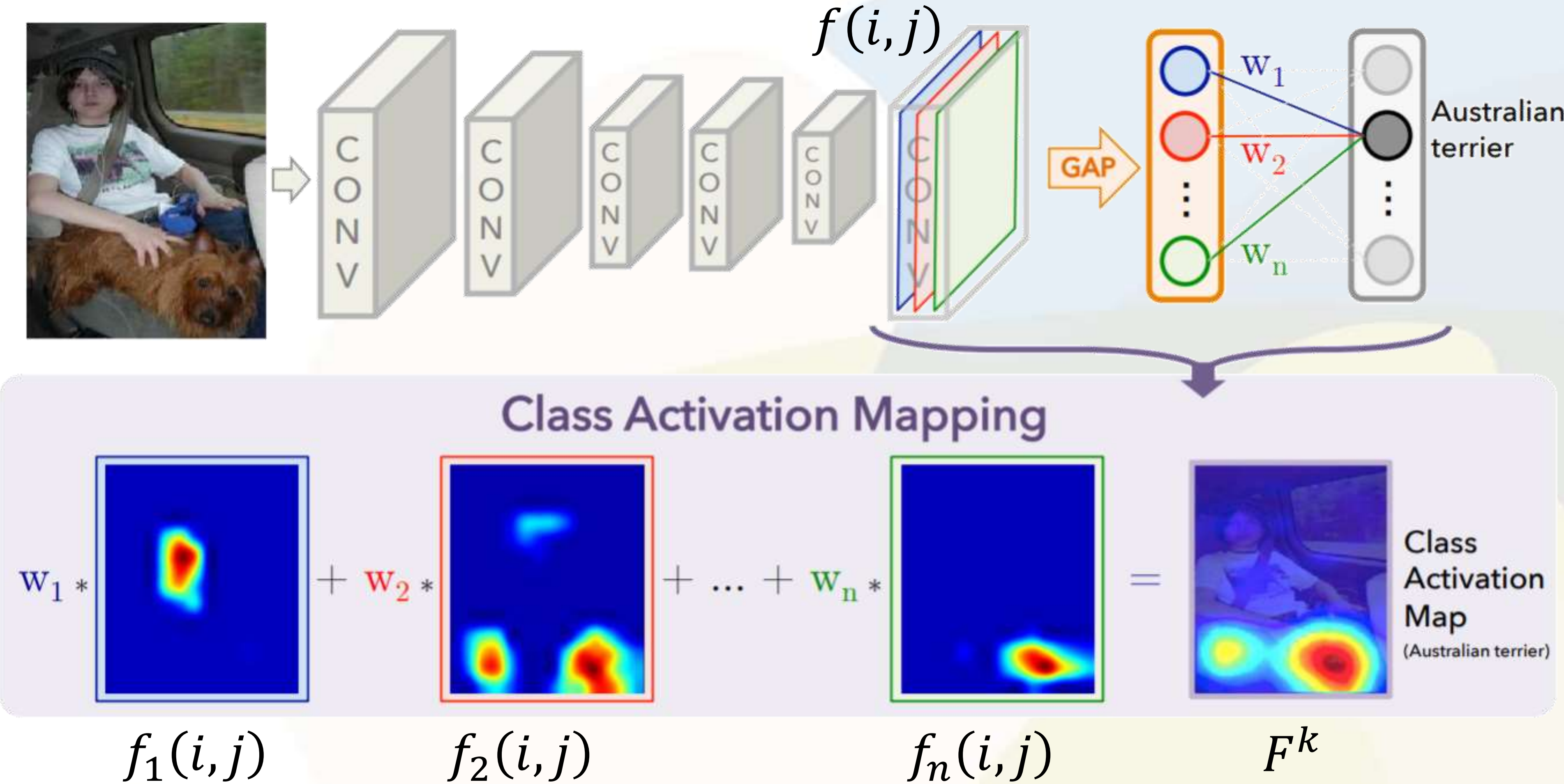
$$\begin{bmatrix} \frac{\partial \hat{y}_1}{\partial x_1} & \cdots & \frac{\partial \hat{y}_1}{\partial x_k} \\ \cdots & & \cdots \\ \frac{\partial \hat{y}_m}{\partial x_1} & \cdots & \frac{\partial \hat{y}_m}{\partial x_k} \end{bmatrix}$$

Saliency



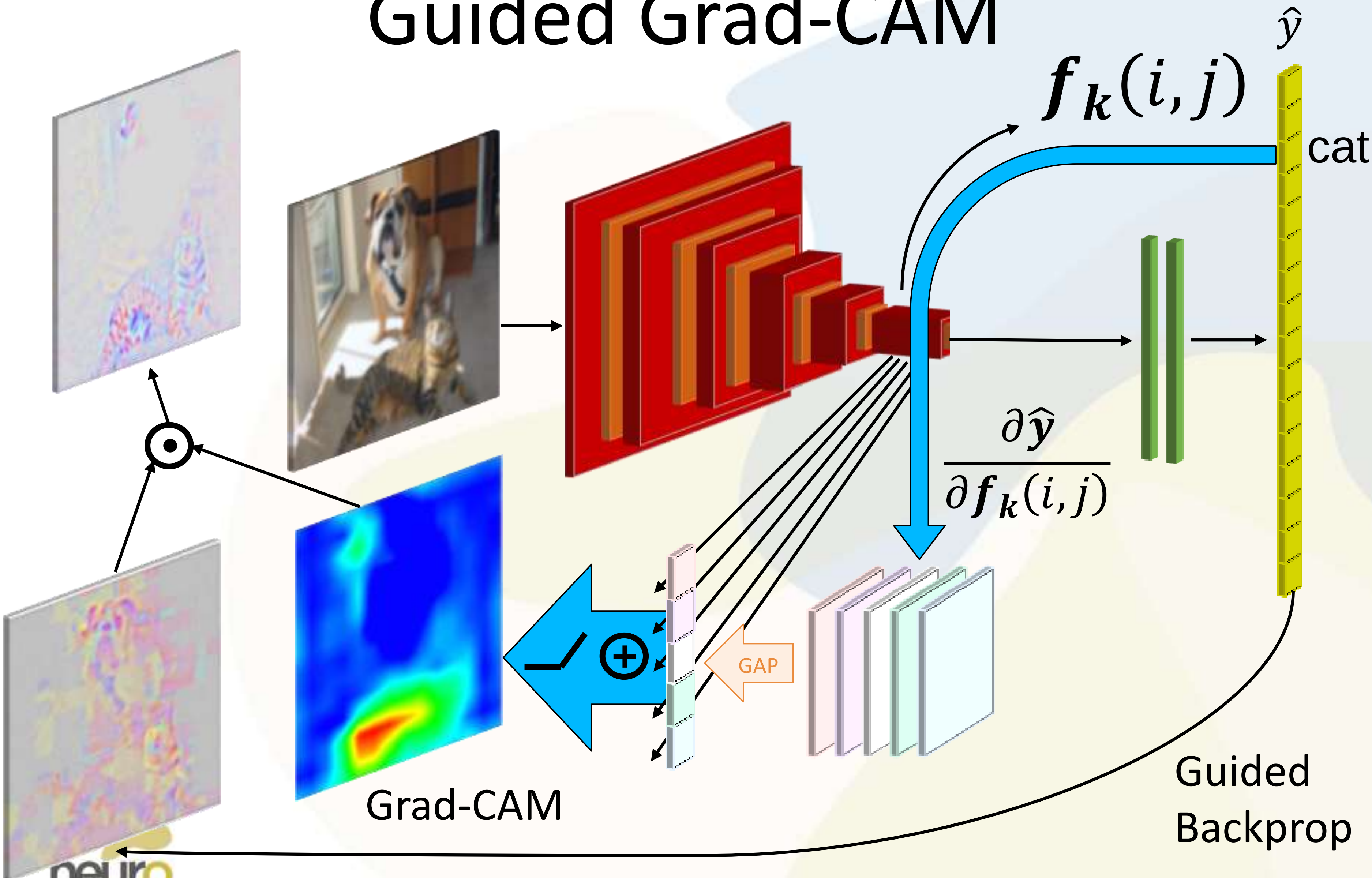
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013)

Class Activation Mapping



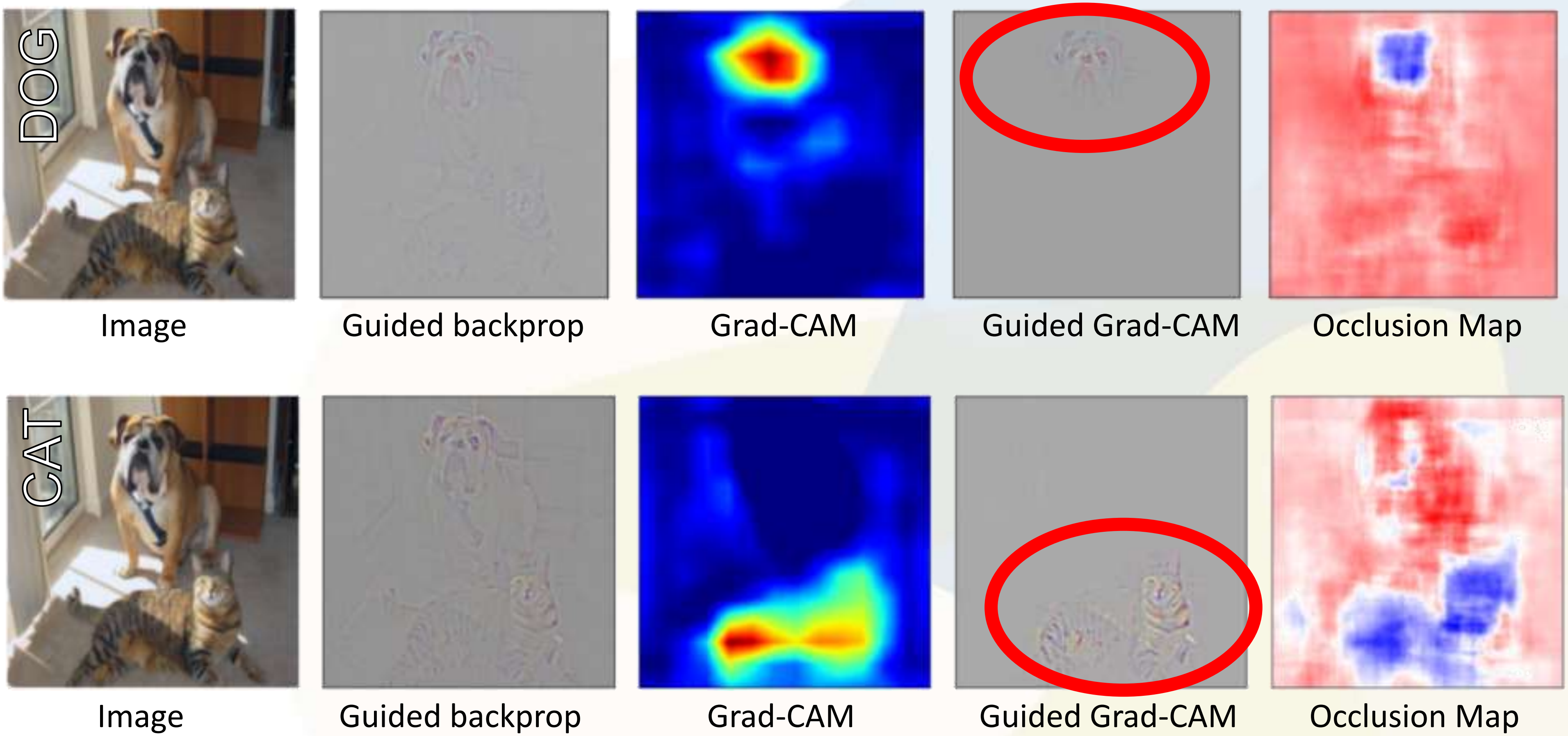
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2929).

Guided Grad-CAM



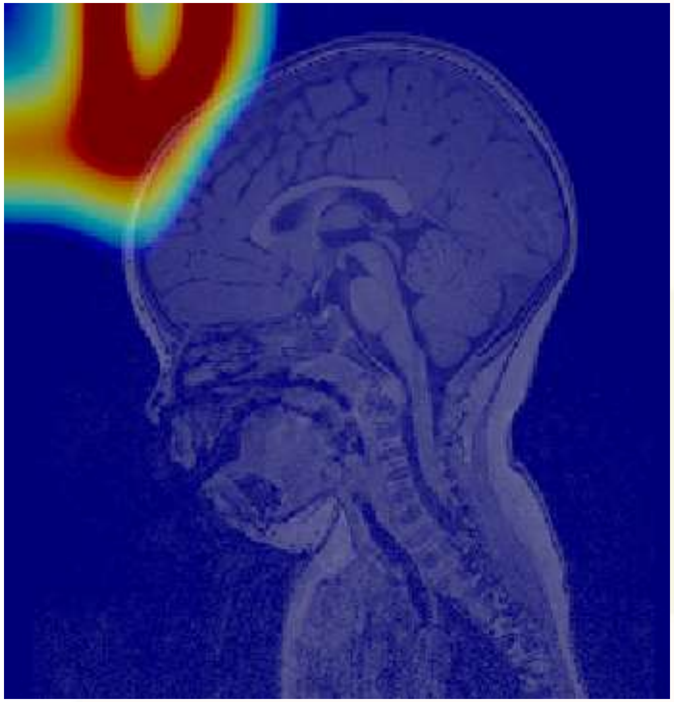
Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *ICCV*. 2017.

Guided Grad-CAM

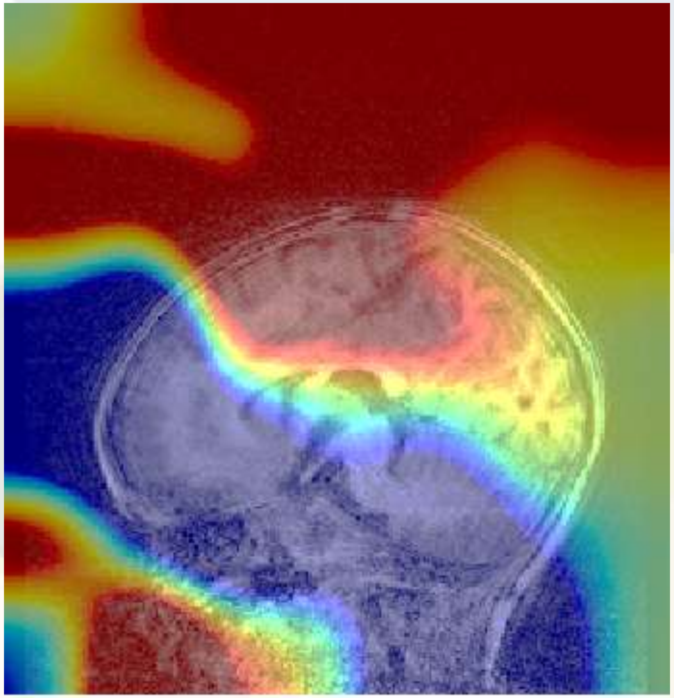


Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *ICCV*. 2017.

Guided Grad-CAM



PASS



FAIL

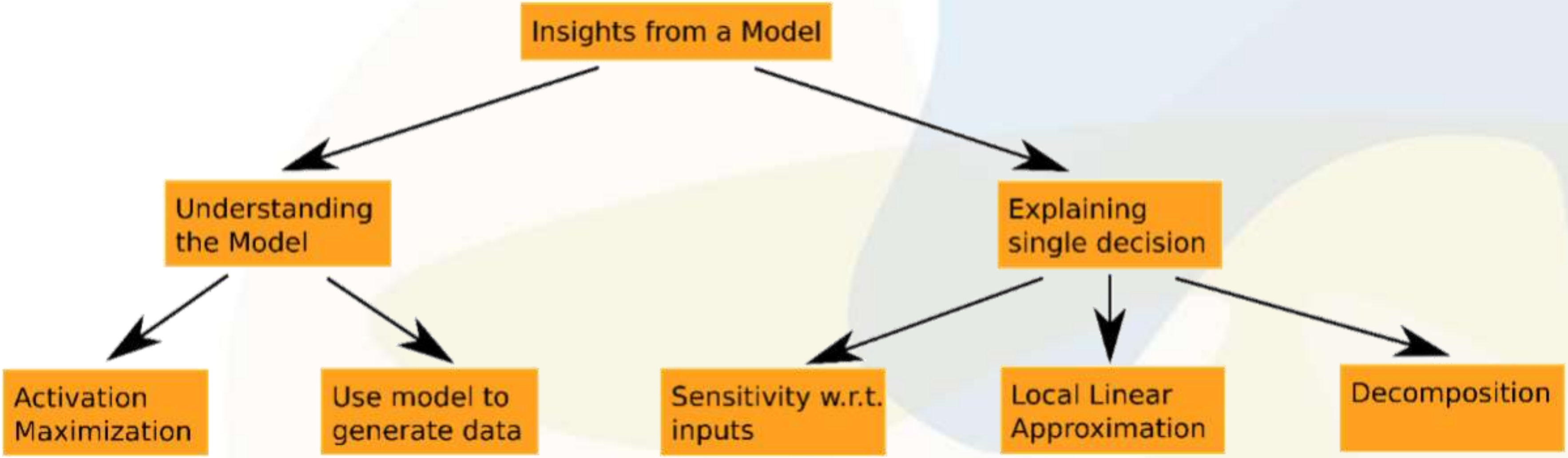
neuro

Institut et hôpital neurologiques de Montréal
Montreal Neurological Institute and Hospital

McGill

Centre universitaire de santé McGill
McGill University Health Centre

Interpreting Predictions



Deconvolution

Guided Backprop

Saliency

Grad-CAM

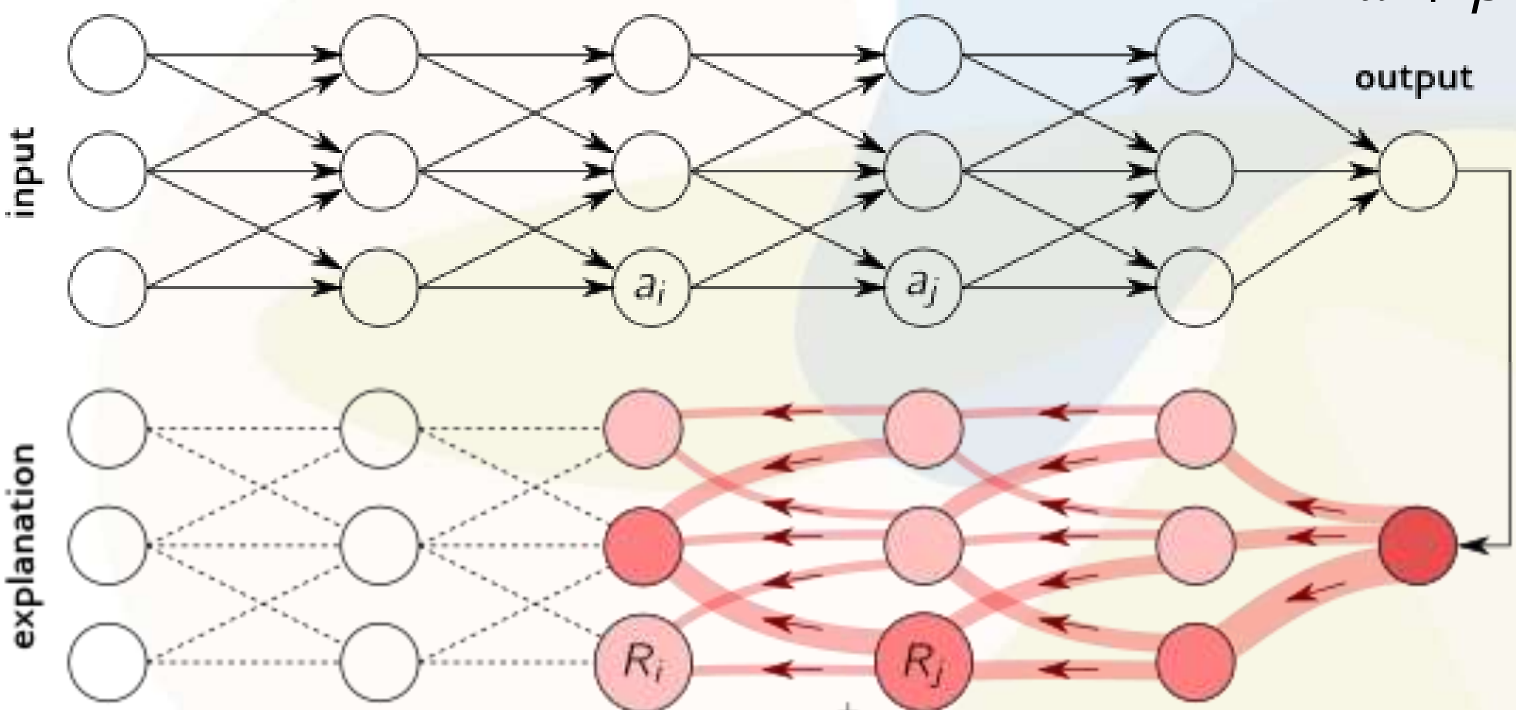
Negative Evidence

@LifeIsAJoke00, weheartit.com

Layer-wise Relevance Propagation

$$R_i^l = \sum_j \left(\alpha \cdot \frac{(a_i \cdot w_{ij})^+}{\sum_i (a_i \cdot w_{ij})^+} + \beta \cdot \frac{(a_i \cdot w_{ij})^-}{\sum_i (a_i \cdot w_{ij})^-}\right) \cdot R_j^{l+1}$$

$$\alpha + \beta = 1$$



$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS one* 10.7 (2015): e0130140.

LRP

Conservation $\sum_p R_p^l = \sum_p R_p^{l+1}$

Positivity $R_p > 0$

Continuity Small changes in input should produce small changes in relevance

Selectivity Removing relevant features should decrease prediction accuracy

Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS one* 10.7 (2015): e0130140.

LIME

Local Interpretable Model-agnostic Explanations

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; L(f, g, \pi_x) + \Omega(g)$$

f : model

g : interpretable model version

ξ : explanation

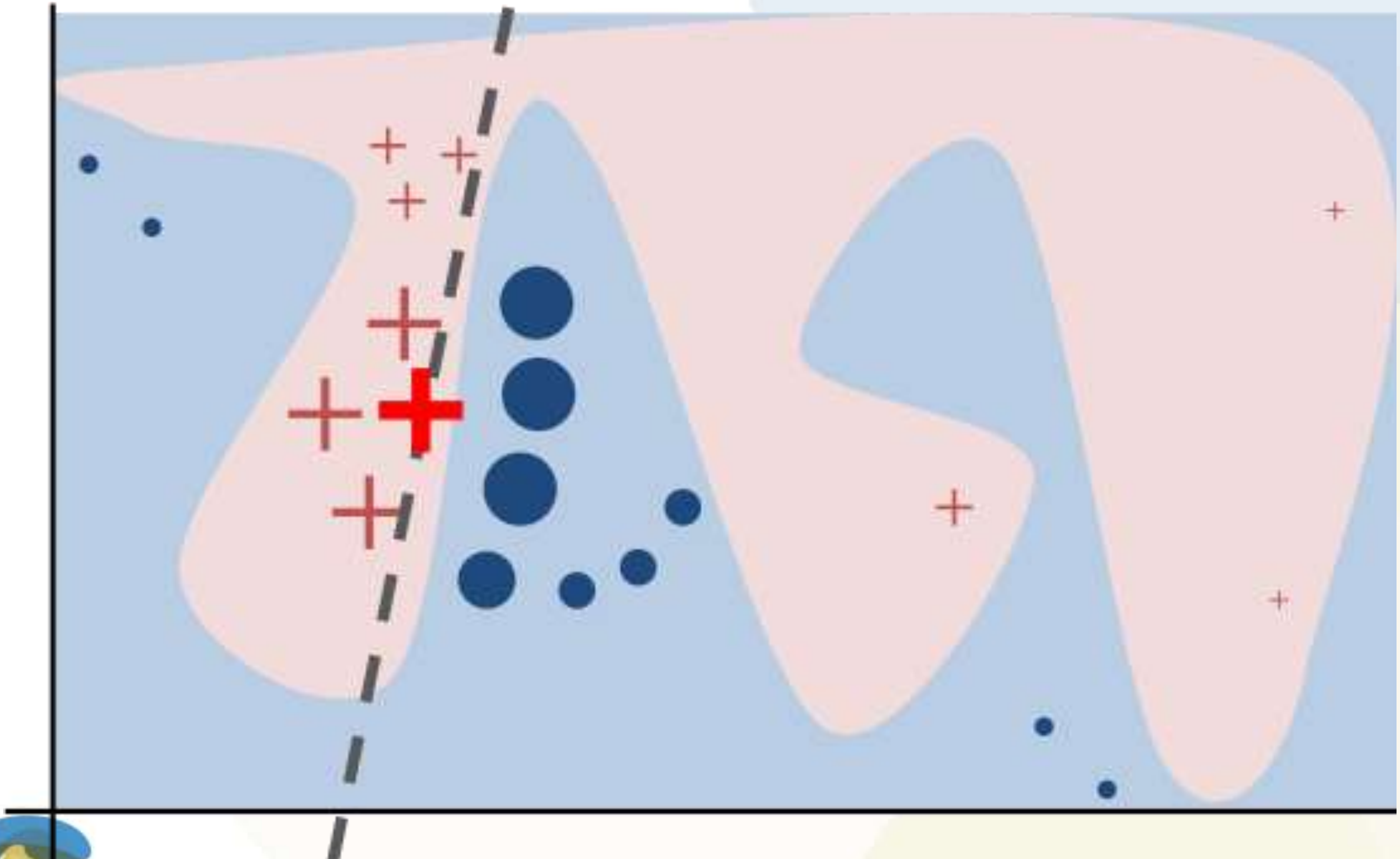
L : how bad g is at approximating real model f

π_x: proximity measure that determines what is "local"

Ω : complexity

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." *SIGKDD,* 2016.

LIME



https://github.com/marcotcr/lime

neuro
Institut et hôpital neurologiques de Montréal
Montreal Neurological Institute and Hospital

McGill

Centre universitaire de santé McGill
McGill University Health Centre

LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; L(f, g, \pi_x) + \Omega(g)$$

Ω : complexity, g : interpretable model, π_x: proximity, L : error

$$g(z') = w_g \cdot z'$$ ← Linear models

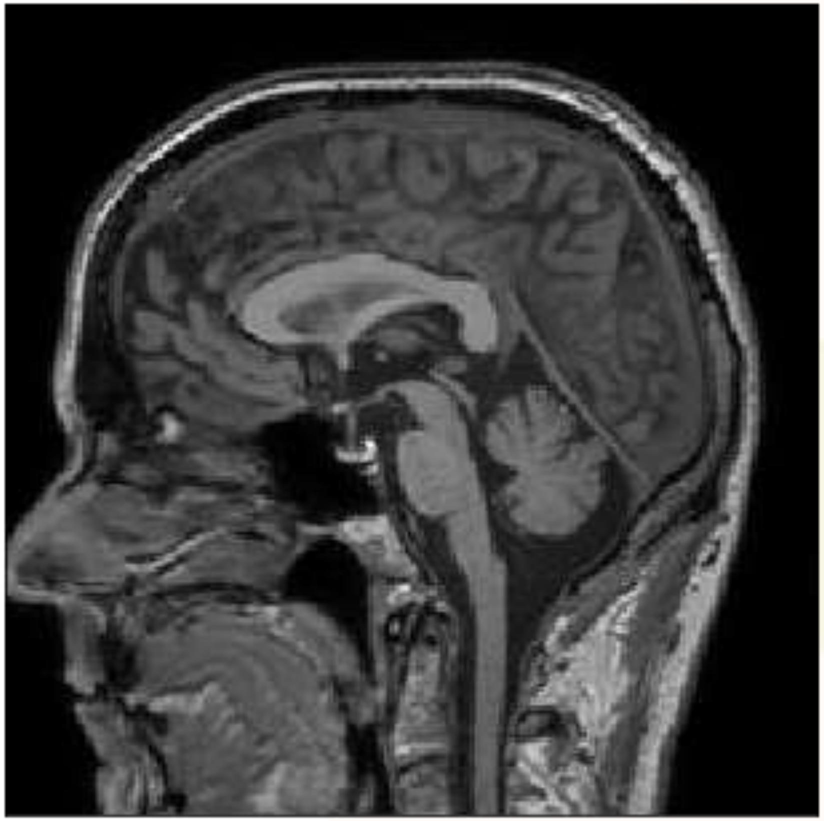
$$\pi_x(z) = e^{\frac{-(x-z)^2}{\sigma^2}}$$ ← Negative exponential of Euclidean distance

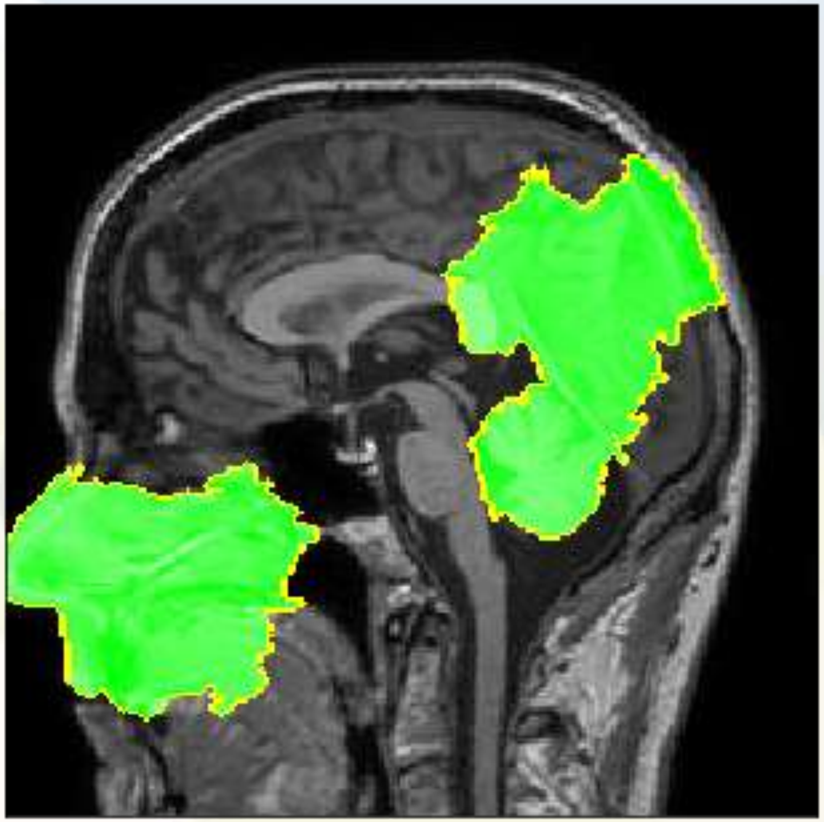
$$\Omega(g)$$ ← Choose K features

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \cdot (f(z) - g(z'))^2$$

https://github.com/marcotcr/lime

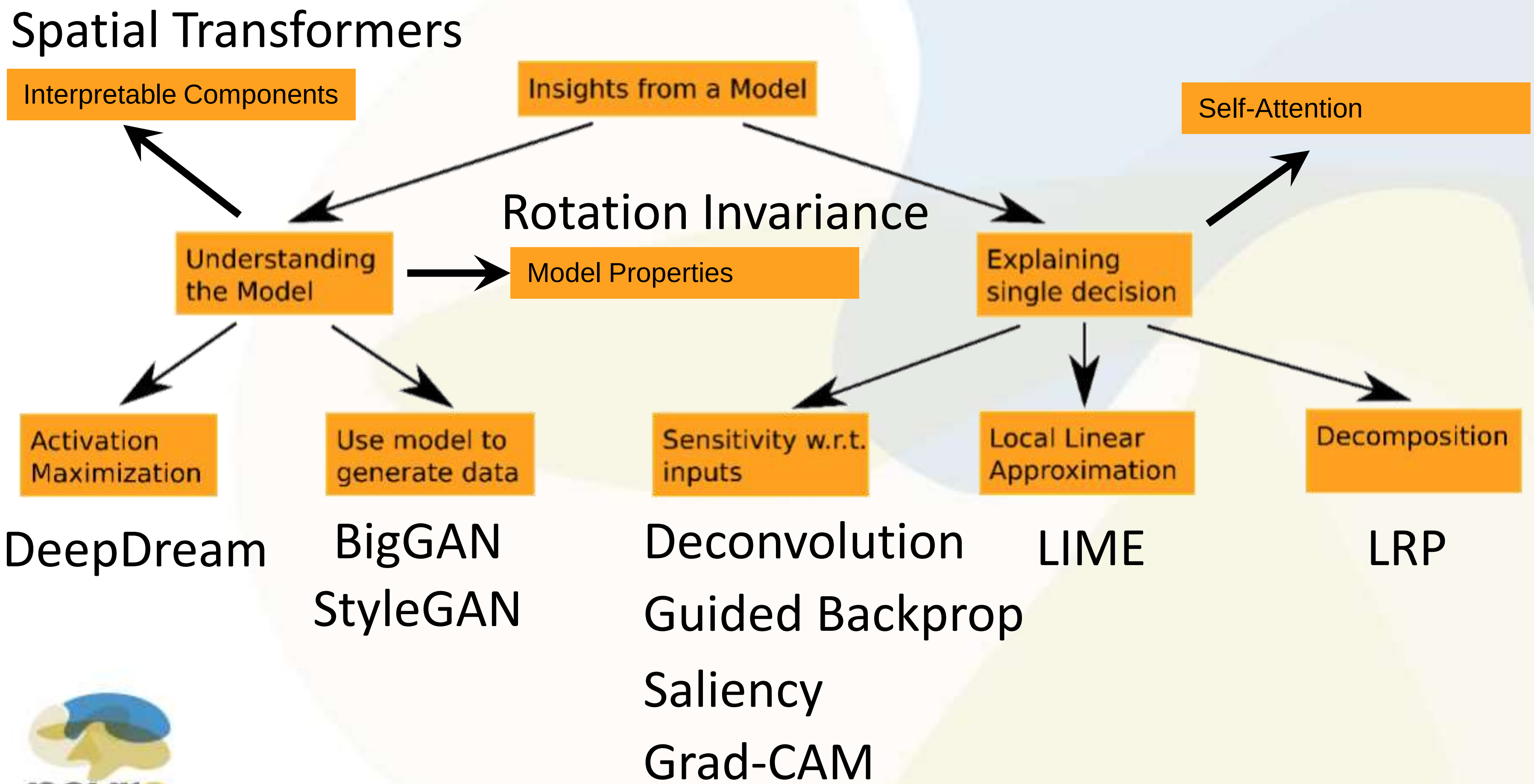
Defacing Detector





https://colab.research.google.com/drive/1EgdnWZeNqmzqEmnSR9PUnYXlTjeu1wAU

Interpreting Predictions



Interpretability

- Causality
- Transparency
- Simulatability
- Decomposability
- Algorithmic guarantees

Lipton, Zachary C. "The mythos of model interpretability." *arXiv preprint arXiv:1606.03490* (2016).

Danger!

- LIME: https://github.com/marcotcr/lime
- Saliency / Grad-CAM: https://github.com/raghakot/keras-vis
- StyleGAN: https://github.com/NVlabs/stylegan
- BigGAN: https://colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/biggan_generation_with_tf_hub.ipynb

Interpretability