Feature Attribution:

Introduction to the Classic Methodologies

2024-07-19

Li Peng-Hsuan 李朋軒

Agenda

- Why feature attribution
- Classic methodologies—the general, the good and the sound
- The SOTA and the limit

Agenda

- Why feature attribution
- Classic methodologies—the general, the good and the sound
- The SOTA and the limit

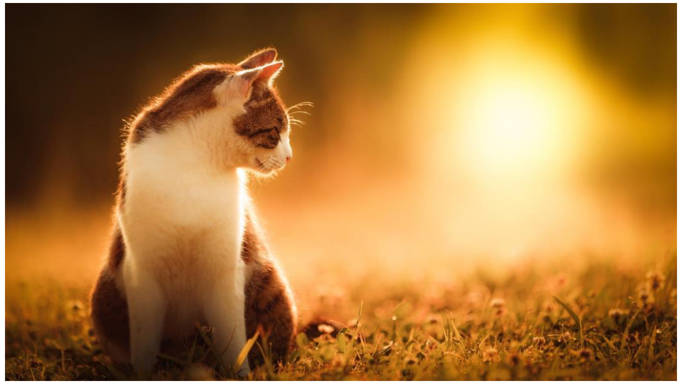
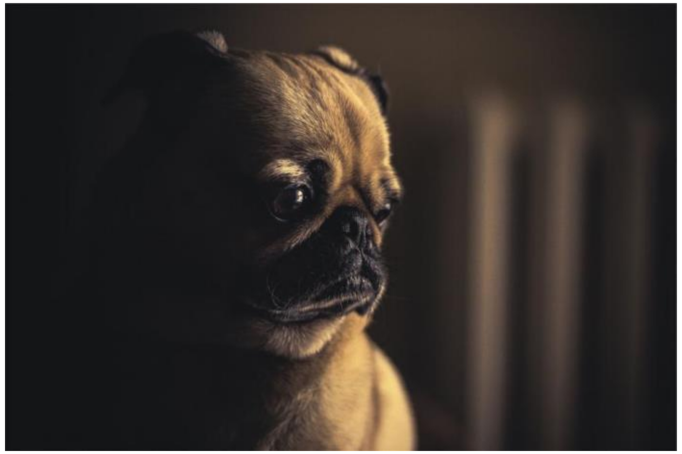
Why Feature Attribution

- Model validation
- Knowledge discovery

Cats and Dogs





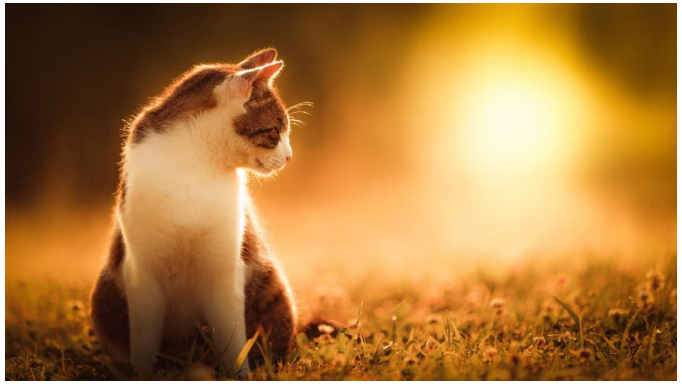
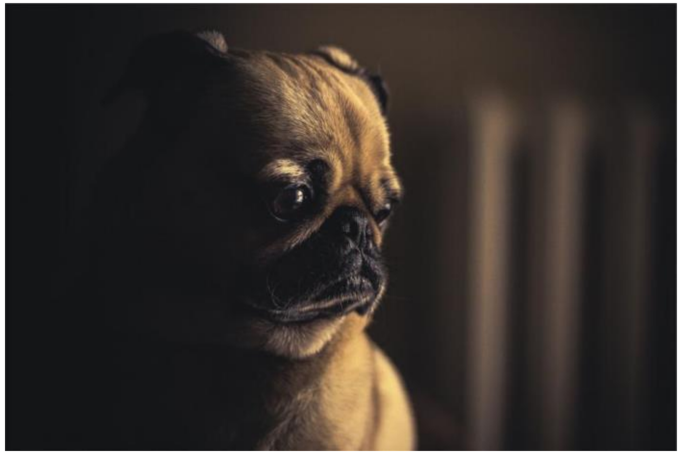



cat / dog

cat / dog





cat / dog

cat / dog

cat

cat

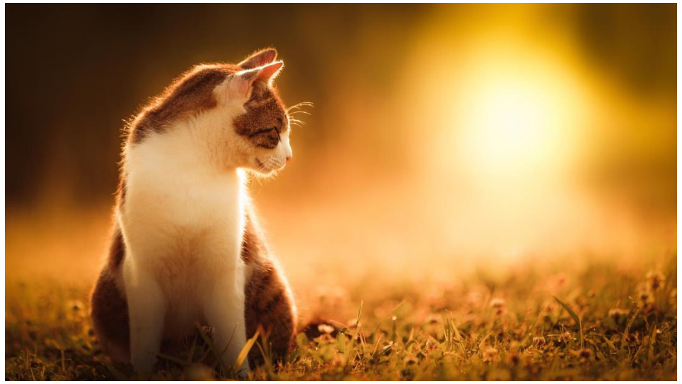
dog dog

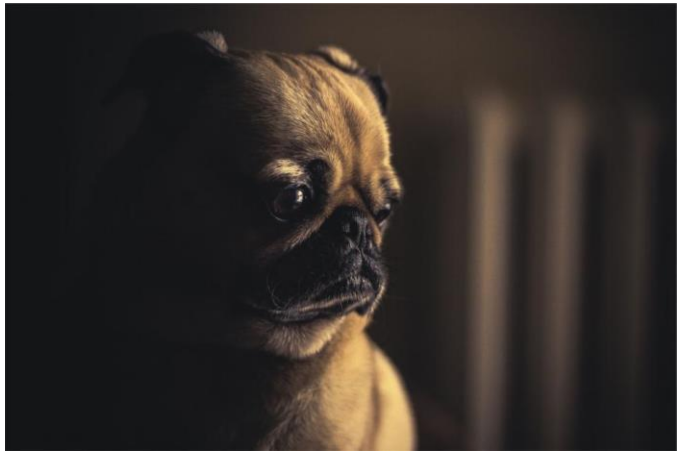
cat cat

dog

dog

cat cat





dog

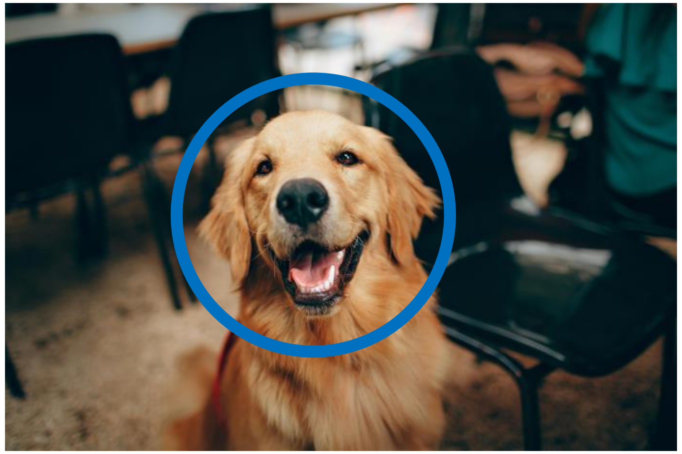
dog

cat / dog

important feature to me

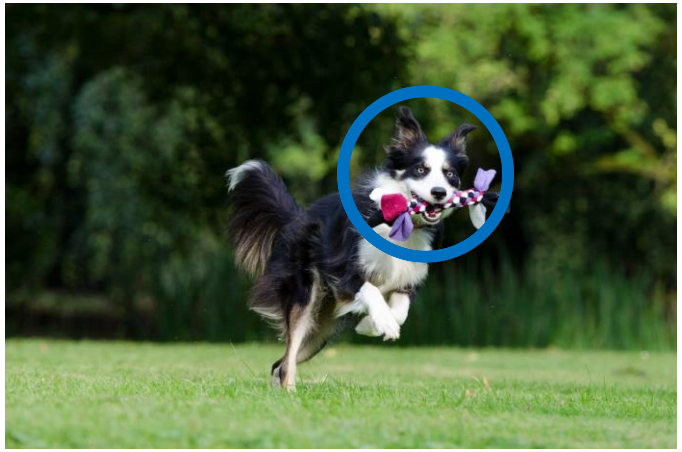
cat

dog

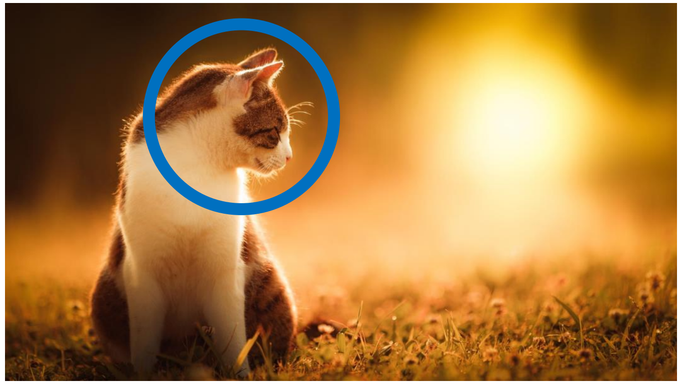


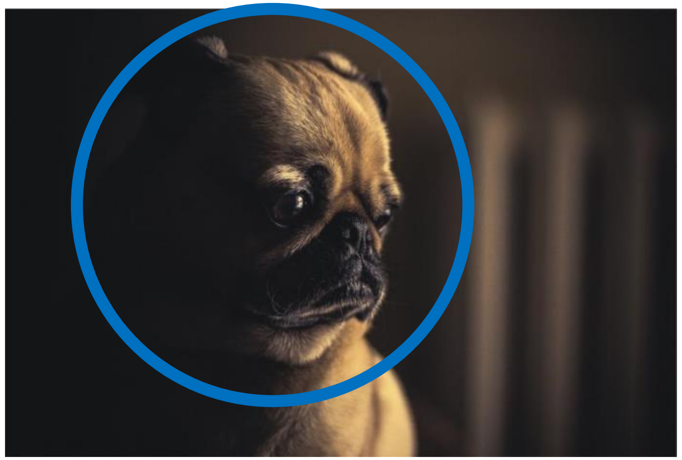
cat

dog



cat

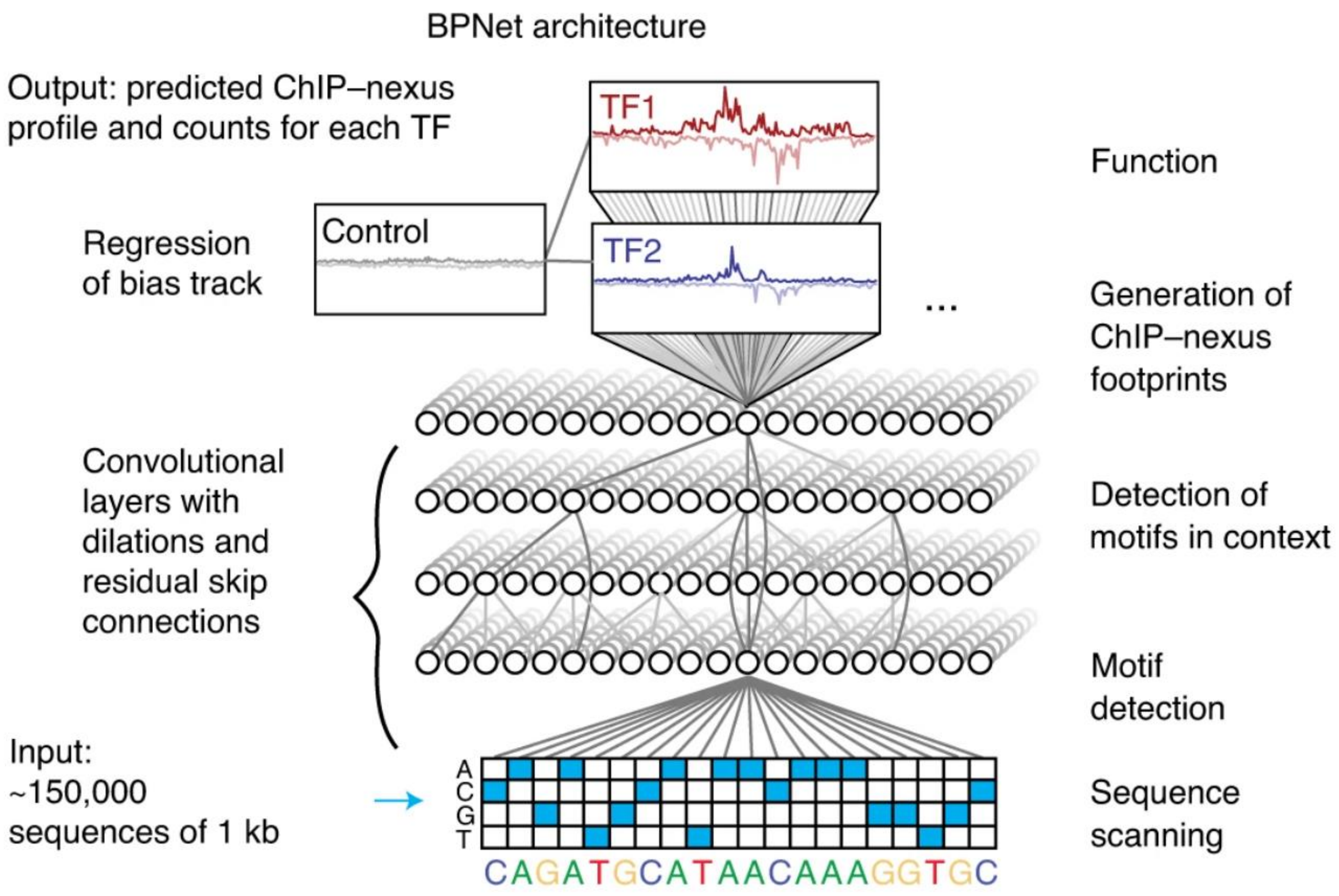




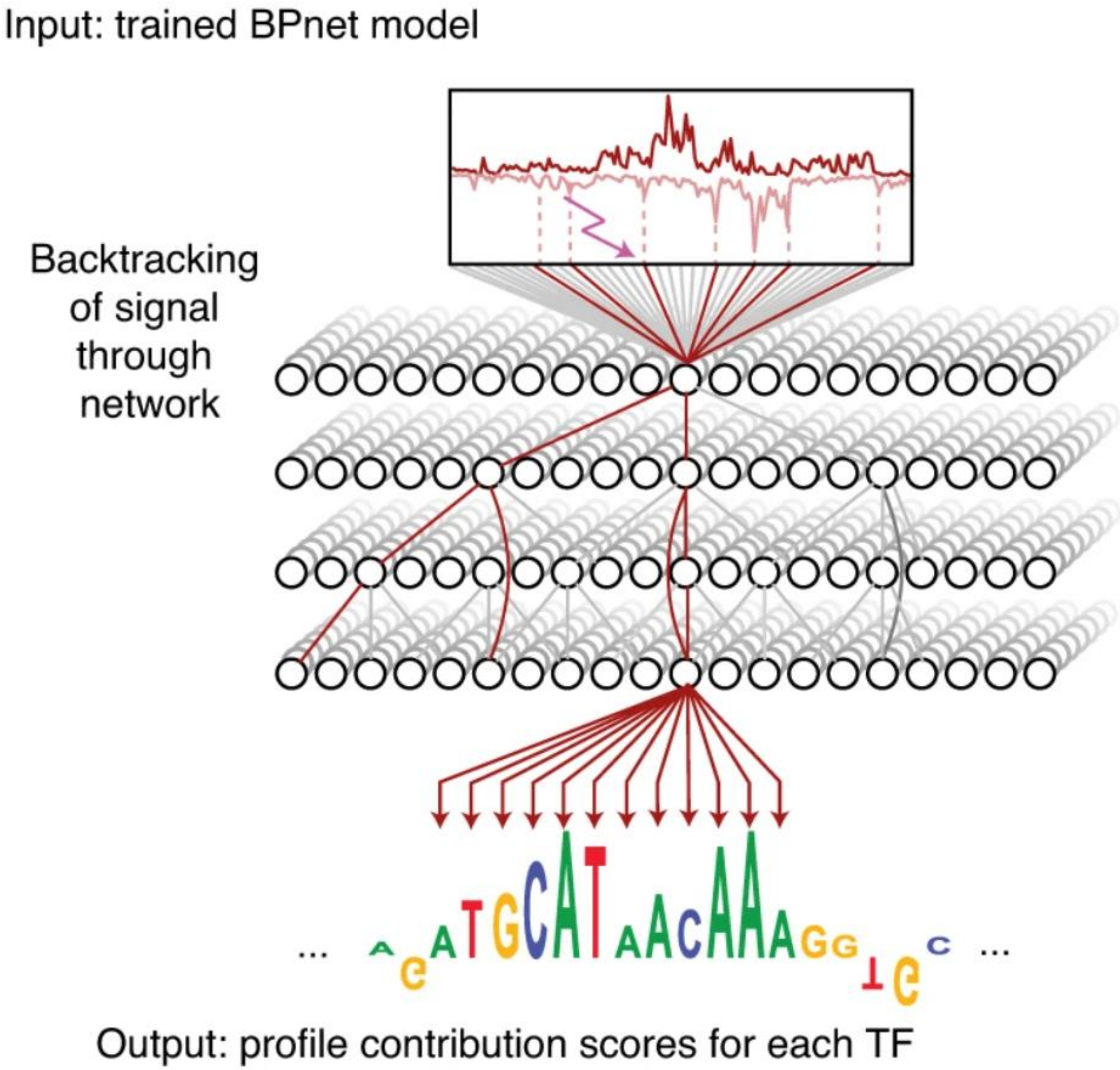
dog


cat / dog
important feature to me
cat
dog
cat
dog
cat
dog

Sequence & Motif



Base-resolution models of transcription-factor binding reveal soft motif syntax.
https://doi.org/10.1038/s41588-021-00782-6



Input: trained BPnet model
Backtracking of signal through network
Output: profile contribution scores for each TF

Agenda

- Why feature attribution
- Classic methodologies—the general, the good and the sound
- The SOTA and the limit

The General, the Good and the Sound

- The *shapley* methodology
- The *inversion* methodology
- The *gradient* methodology

The General


我們需要能從跨國合作中獲利的點子
賣武器收保護費!
調高關稅!
公平分配合作效益?
Shapley
Lloyd Shapley, 1951

Awesome Cooperative Development Plan

....

...........................legalize marijuana trade...............

......drill oil from the Arctic Ocean................................

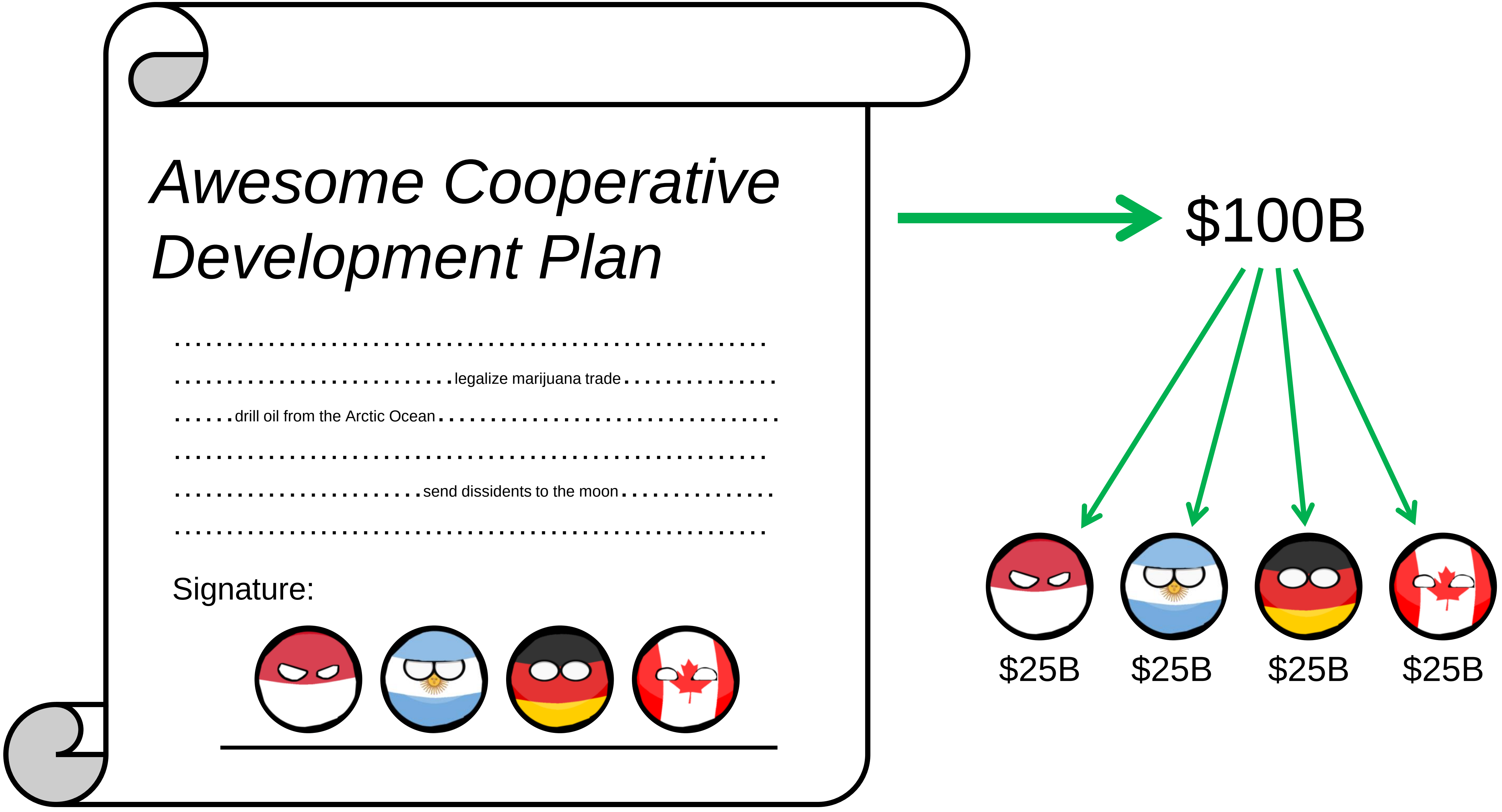
....

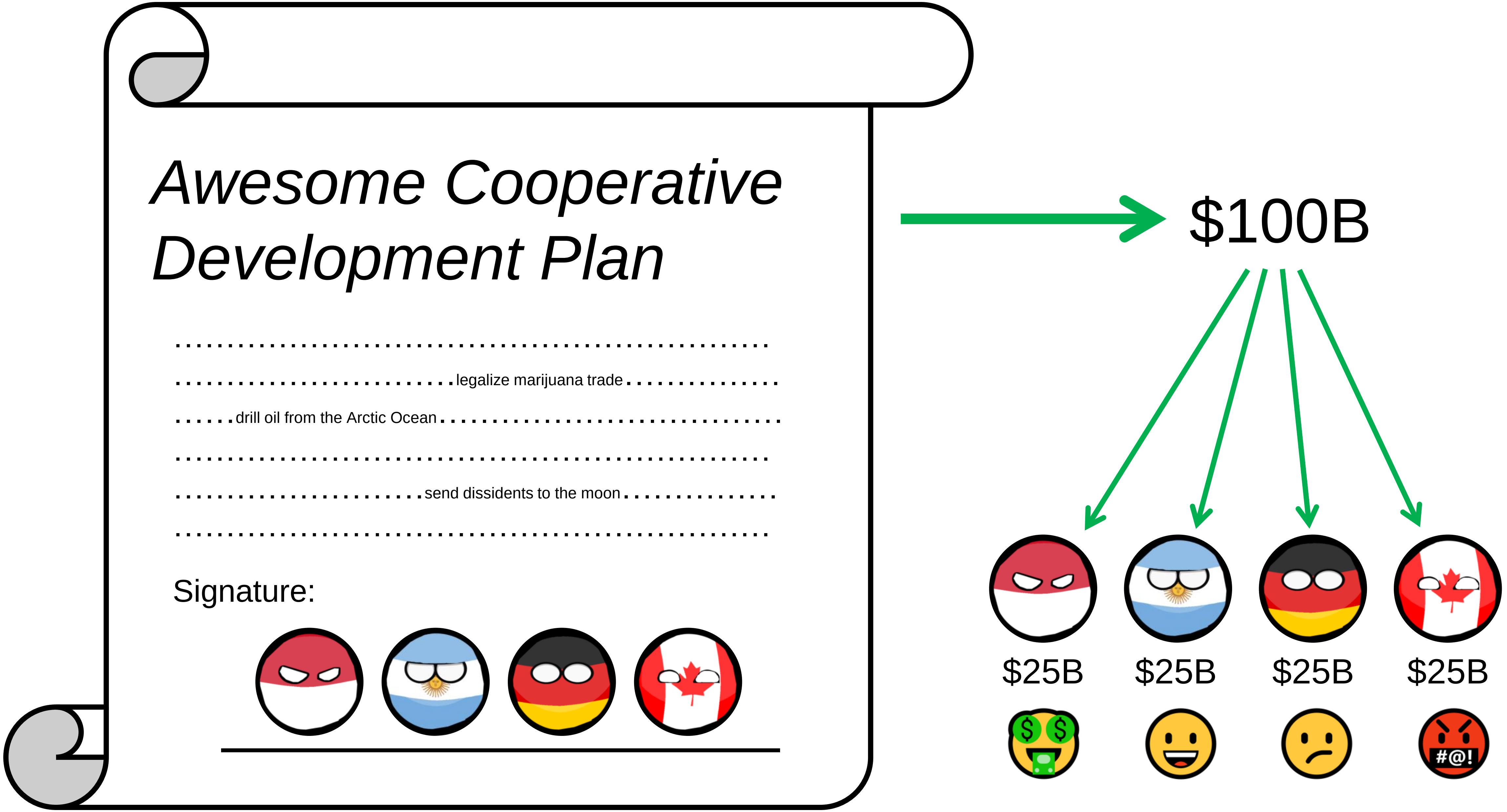
........................send dissidents to the moon...............

....

Signature:

→ $100B


Awesome Cooperative Development Plan
legalize marijuana trade
drill oil from the Arctic Ocean
send dissidents to the moon
Signature:
$100B
$25B
$25B
$25B
$25B



Awesome Cooperative Development Plan
legalize marijuana trade
drill oil from the Arctic Ocean
send dissidents to the moon
Signature:
$100B
$25B
$25B
$25B
$25B
#@!

Awesome Cooperative Development Plan

legalize marijuana trade

drill oil from the Arctic Ocean

send dissidents to the moon

Signature:

~~$100B~~

$50B

I contributed $50B!

Awesome Cooperative Development Plan

legalize marijuana trade

drill oil from the Arctic Ocean

send dissidents to the moon

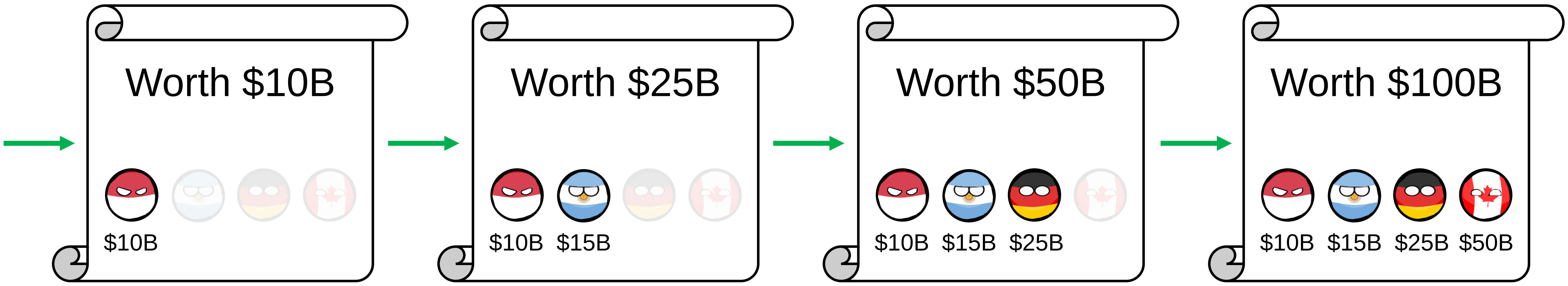
Signature:

~~$100B~~

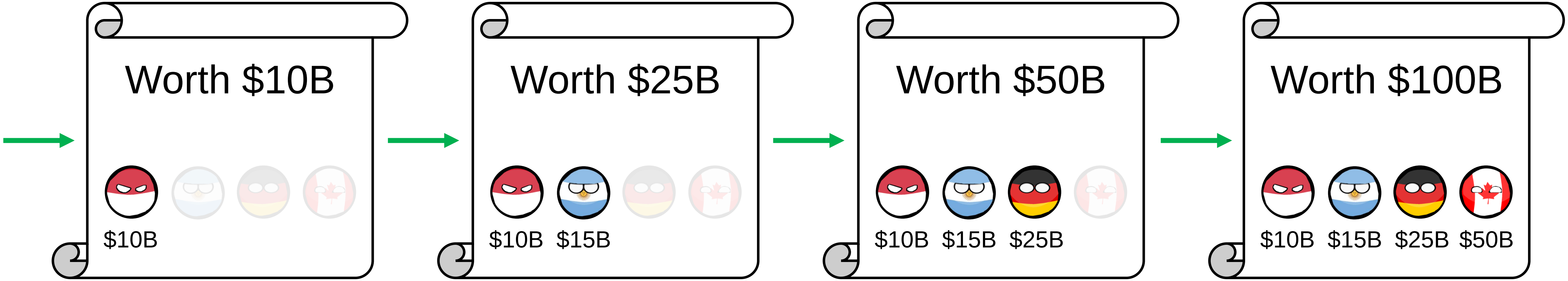
→ $10B

We contributed $90B!

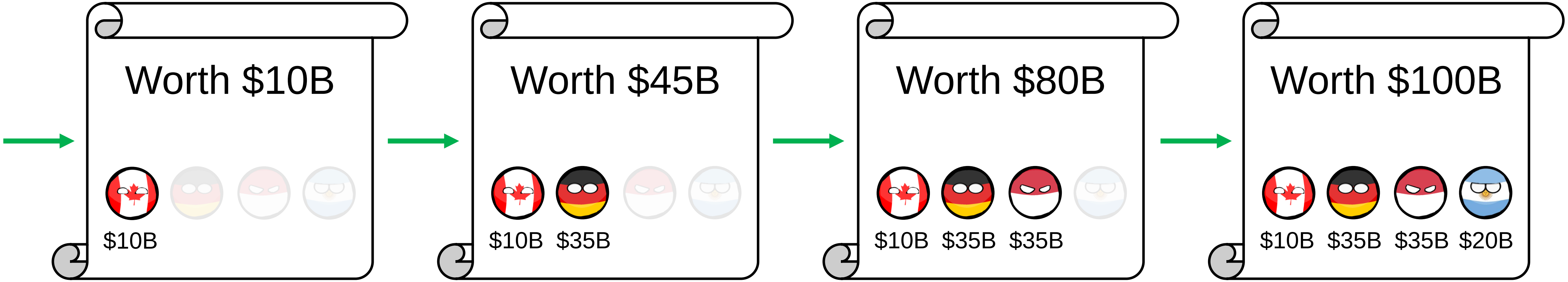
Signature order

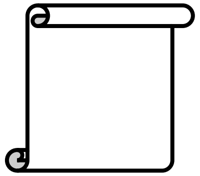


Order 1

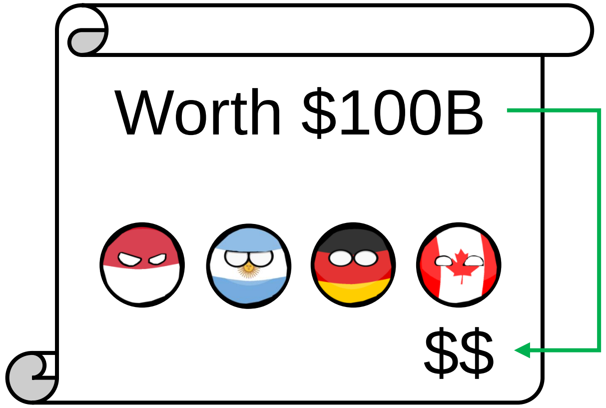


Order 2



shapley-value( ; ) = average( $$)

Economics vs. ML



attribute

vs.

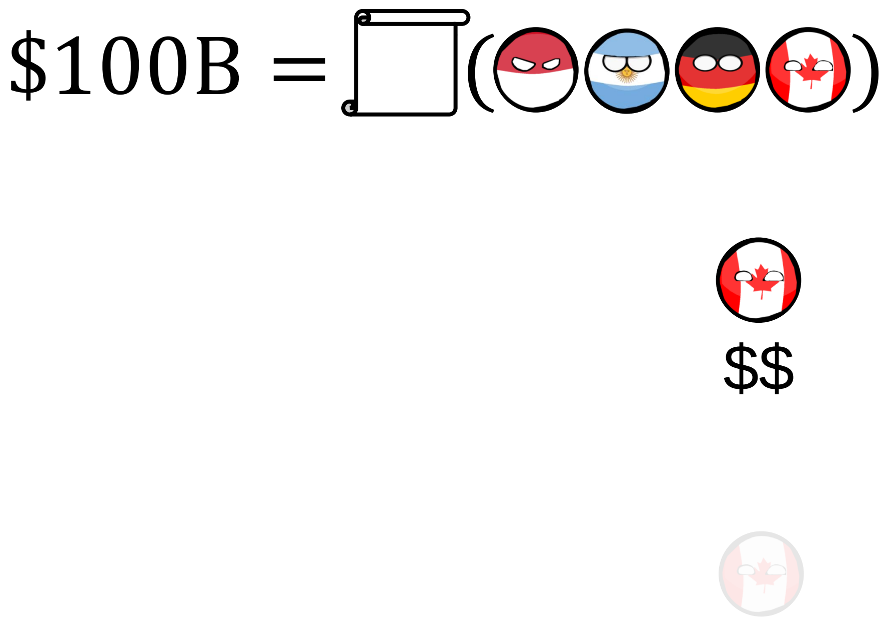
$$y = f(\boldsymbol{x})$$

attribute

(feature) x_i

(contribution) c_i

The *Shapley* Methodology



$$y = f(\boldsymbol{x})$$

attribute

(feature) x_i

(contribution) c_i

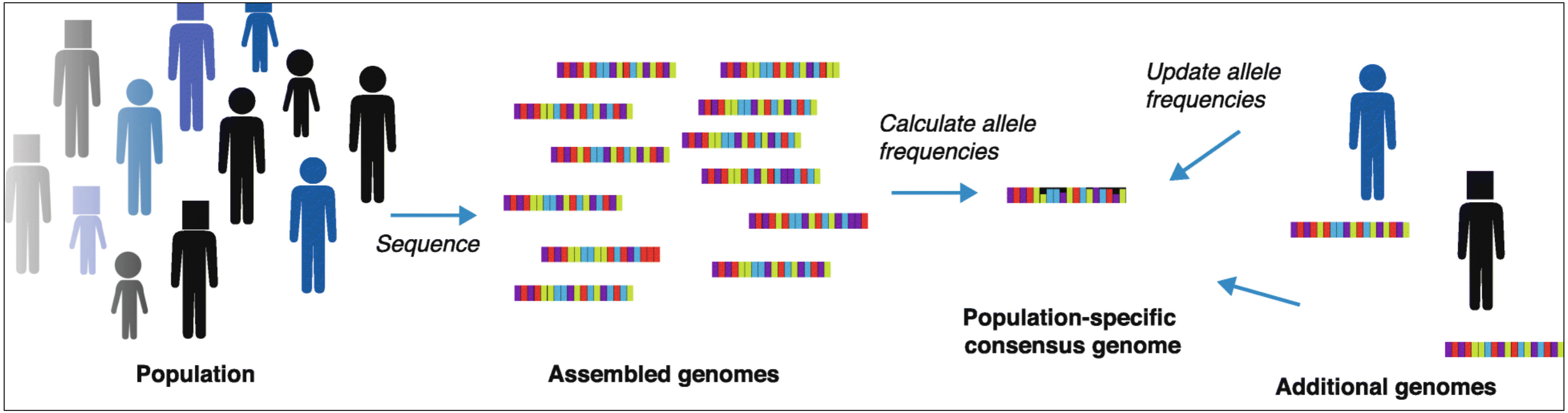
(reference) r_i

Sample $\boldsymbol{x}$

Reference $\boldsymbol{r}$

Sample x

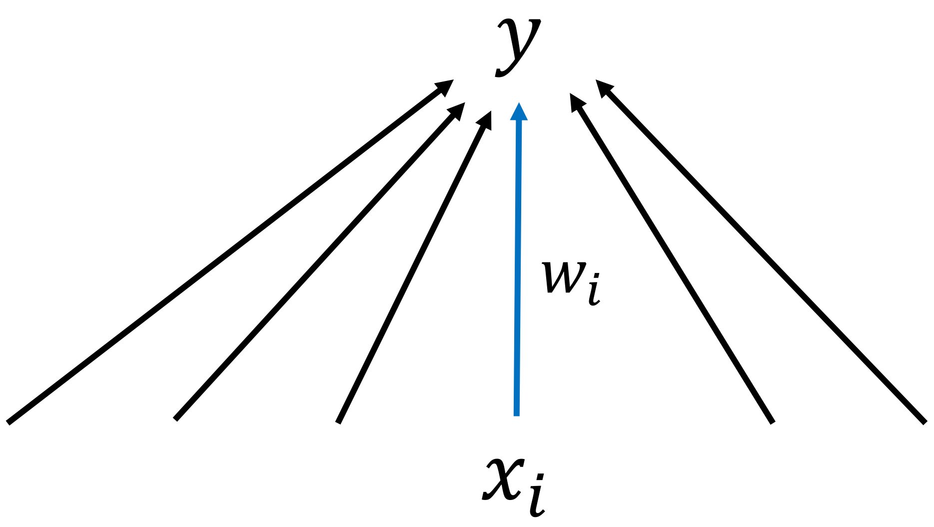
Reference r



Is it time to change the reference genome?
https://doi.org/10.1186/s13059-019-1774-4

The Good

Attribution for a linear function



$$y = f(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{w} + b$$

$$c_i = x_i w_i$$

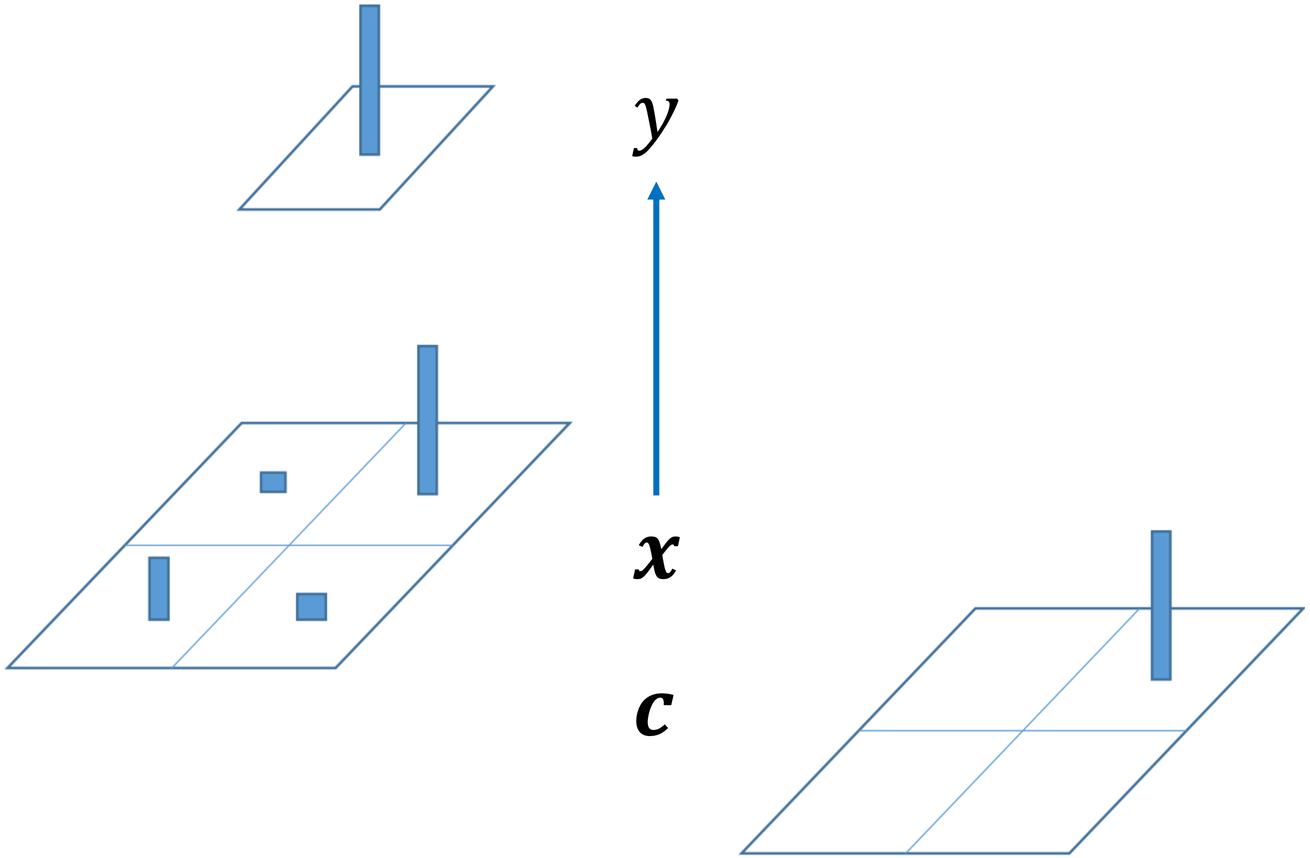
Attribution for ReLU

$$y = f(x) = \max(x, 0)$$

$$\begin{tikzcd} y \\ x \arrow[u, "{1[x \geq 0]}"'] \end{tikzcd}$$

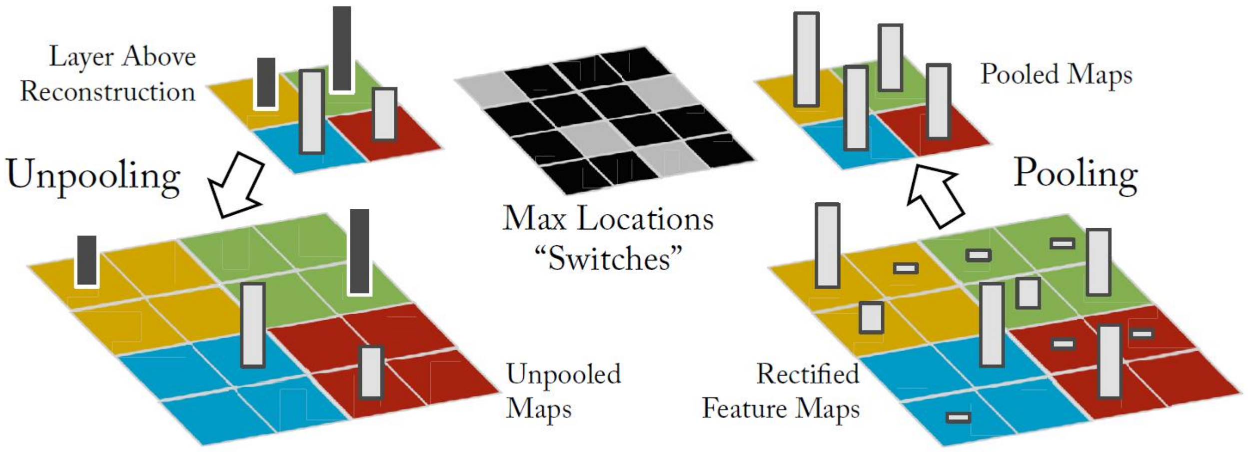
$$c = x \cdot 1[x \geq 0]$$

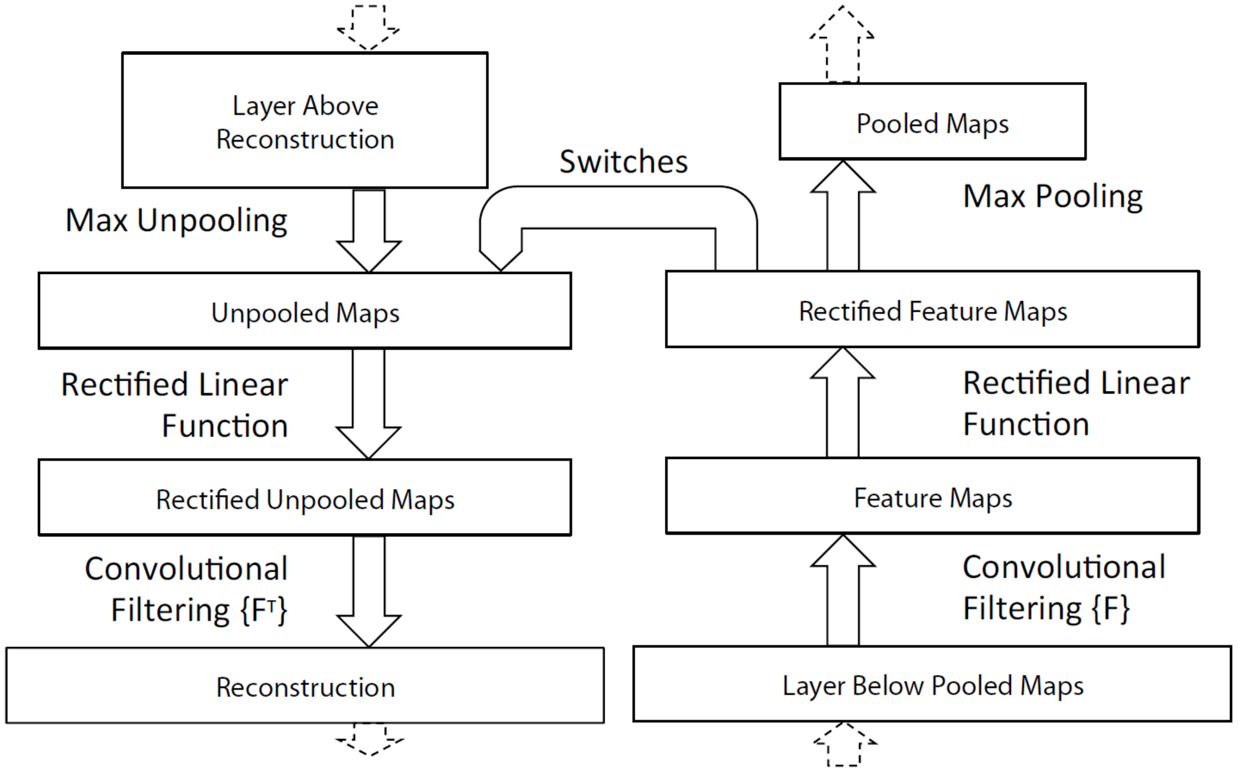
Attribution for Max-Pooling



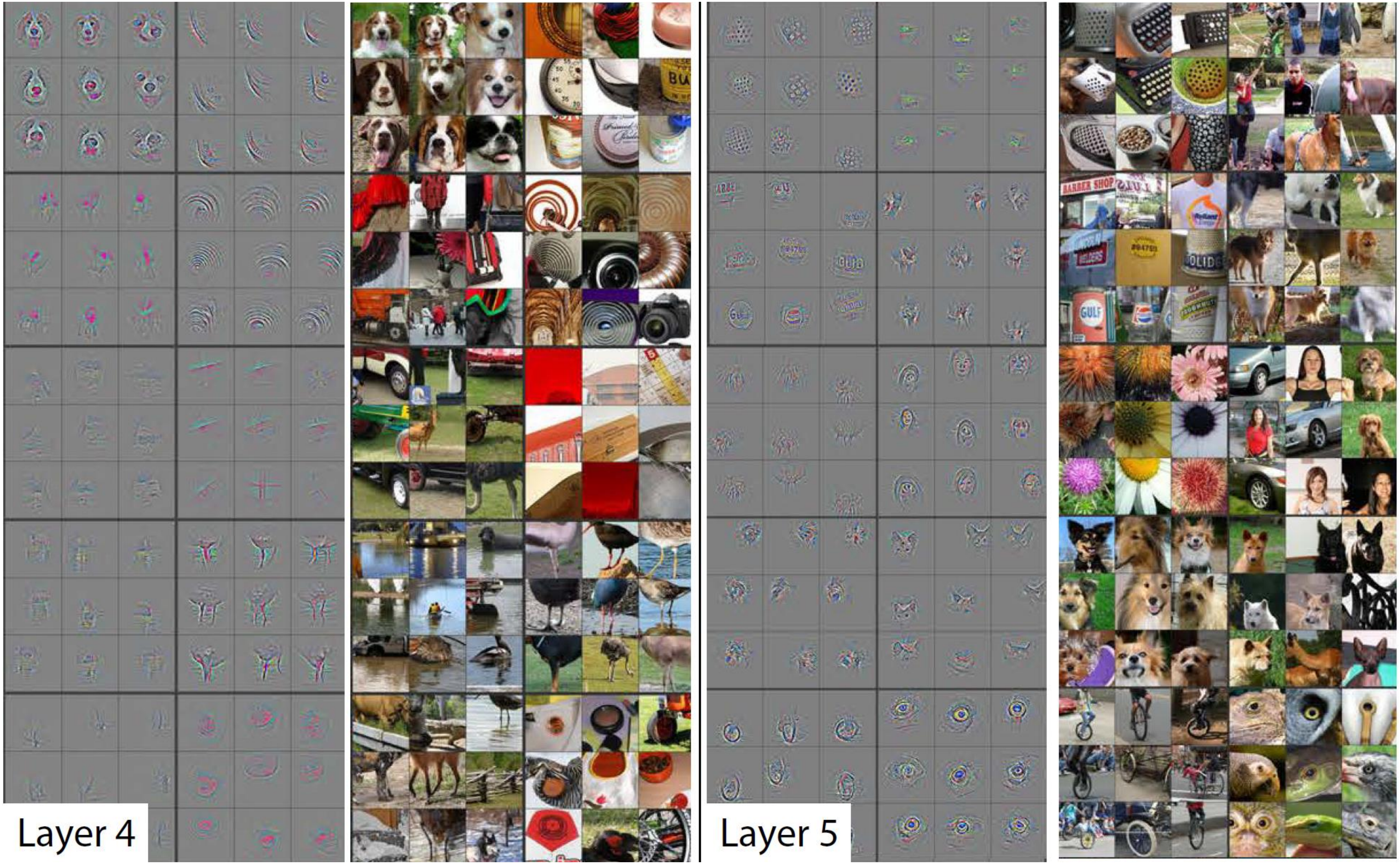
$$y = f(\boldsymbol{x}) = \max(\boldsymbol{x})$$

Attribution for CNN





Visualizing and Understanding Convolutional Networks.
https://doi.org/10.1007/978-3-319-10590-1_53


Layer 4
Layer 5

$$y = f(x) = g_3\left(g_2\big(g_1(x)\big)\right) = g_3 \circ g_2 \circ g_1(x)$$

$$x = f^{-1}(y) = g_1^{-1} \circ g_2^{-1} \circ g_3^{-1}(y)$$

$$y = f(x) = g_3\left(g_2\left(g_1(x)\right)\right) = g_3 \circ g_2 \circ g_1(x)$$

$$x = f^{-1}(y) = g_1^{-1} \circ g_2^{-1} \circ g_3^{-1}(y)$$

Hey! But functions are not generally invertible!

The *Inversion* Methodology

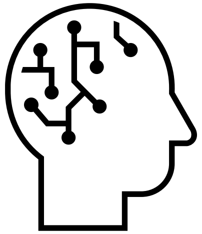
Given a model f

1. Break down composition $f = g_n \circ \cdots \circ g_2 \circ g_1$
2. Formulate inverse g_i^{-1}
3. Reduce inverse to linear $g_i^{-1}(y|\boldsymbol{x}) = g_i^L(y)$
4. Apply chain rule $\boldsymbol{c} = f^L(y) = g_1^L \circ g_2^L \circ \cdots \circ g_n^L(y)$

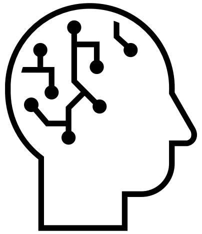
The Sound

Cat!

Still cat!



No cat 😓



How about using gradient as contribution?

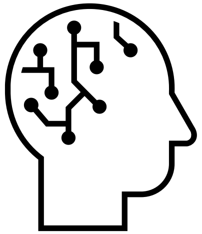
$$\boldsymbol{c} = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

How about using gradient as contribution?

$$\boldsymbol{c} = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_A} > \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_B}$$

Sun

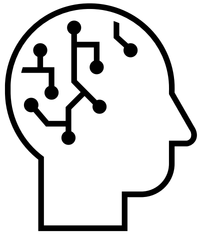


Totally sun

95%

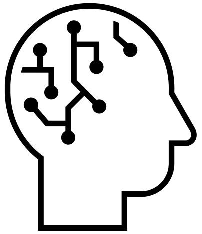
80%

Still sun



50%

Maybe not sun



😓

$$\frac{\partial f(\boldsymbol{x})}{\partial A} = 0$$

100%

$$\frac{\partial f\left(\boldsymbol{x}^{50\%}\right)}{\partial A} > 0$$

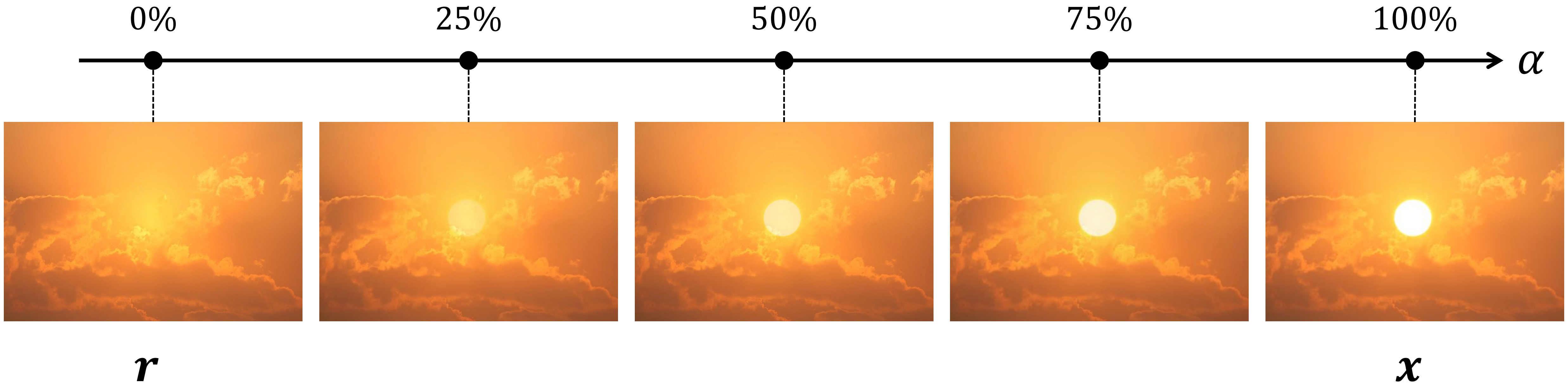
50%

How about adding up gradients at different percentages from reference?

$$\boldsymbol{c} = \sum_{\alpha} \frac{\partial f(\boldsymbol{x}^{\alpha})}{\partial \boldsymbol{x}}$$

The *Gradient* Methodology

$$\boldsymbol{c} = (\boldsymbol{x} - \boldsymbol{r}) \cdot \int_{\alpha=0}^{1} \frac{\partial f\big(\boldsymbol{r} + \alpha(\boldsymbol{x} - \boldsymbol{r})\big)}{\partial \boldsymbol{x}} d\alpha$$



Agenda

- Why feature attribution
- Classic methodologies—the general, the good and the sound
- The SOTA and the limit

The SOTA and the Limit

- SOTA methods
- The general limit to feature attribution
- Remarks

Shapley methodology	Inversion methodology	Gradient methodology
SHAP[1]	DeepLIFT[2]	Integrated Gradients[3]
For every function	For CNN	For deep networks
• Has variants with different niche • Kernel SHAP → general functions • Deep SHAP → CNN • Tree SHAP → decision trees	• Tailored inverse for CNN, for which it is empirically powerful • Not theoretically sound, no robustness guarantee for other model types	• Has soundness and robustness guarantee • Powerful for deep differentiable functions • Weak for non-differentiable functions, e.g., max-pooling • Has a tunable parameter for tradeoff between speed and accuracy

[1]
A Unified Approach to Interpreting Model Predictions.
https://arxiv.org/abs/1705.07874

[2]
Learning Important Features Through Propagating Activation Differences.
https://arxiv.org/abs/1704.02685

[3]
Axiomatic Attribution for Deep Networks.
https://arxiv.org/abs/1703.01365

$$y = f(\boldsymbol{x}) = x_1 + x_2 + x_3$$

$$c_1 = x_1$$

$$c_2 = x_2$$

$$c_3 = x_3$$

$$y = f(\boldsymbol{x}) = x_1 x_2 + x_3$$

$$c_1 = \,?$$

$$c_2 = \,?$$

$$c_3 = x_3$$

$$y = f(\boldsymbol{x}) = x_1 x_2 + x_3$$

$$c_1 = c_2 = \frac{x_1 x_2}{2}$$ 🤔

$$c_3 = x_3$$

$$y = f(\boldsymbol{x}) = {x_1}^{x_2} + x_3$$

$$c_1 = ???$$

🤯

$$c_2 = ???$$

$$c_3 = x_3$$

Various *sound*, *robust*, *powerful* methods exist for many popular model types.

A generally *correct* linear attribution doe not exist.

Large *generative* foundation models have started a new chapter.