Multimedia Analytics

ICS 491

End of Course

Tue Dec 5	Multimedia Analytics	
Thu Dec 7	Course Overview	
Fri Dec 15		Final Project Infographic and Code

Multimedia Analytics

- Image processing
- Video processing
- Text processing
- Speech processing

Central Theme: How to Represent Data as Numbers?

If we can properly represent data as numbers, we have all the tools of math at our disposal to do whatever we want with the data

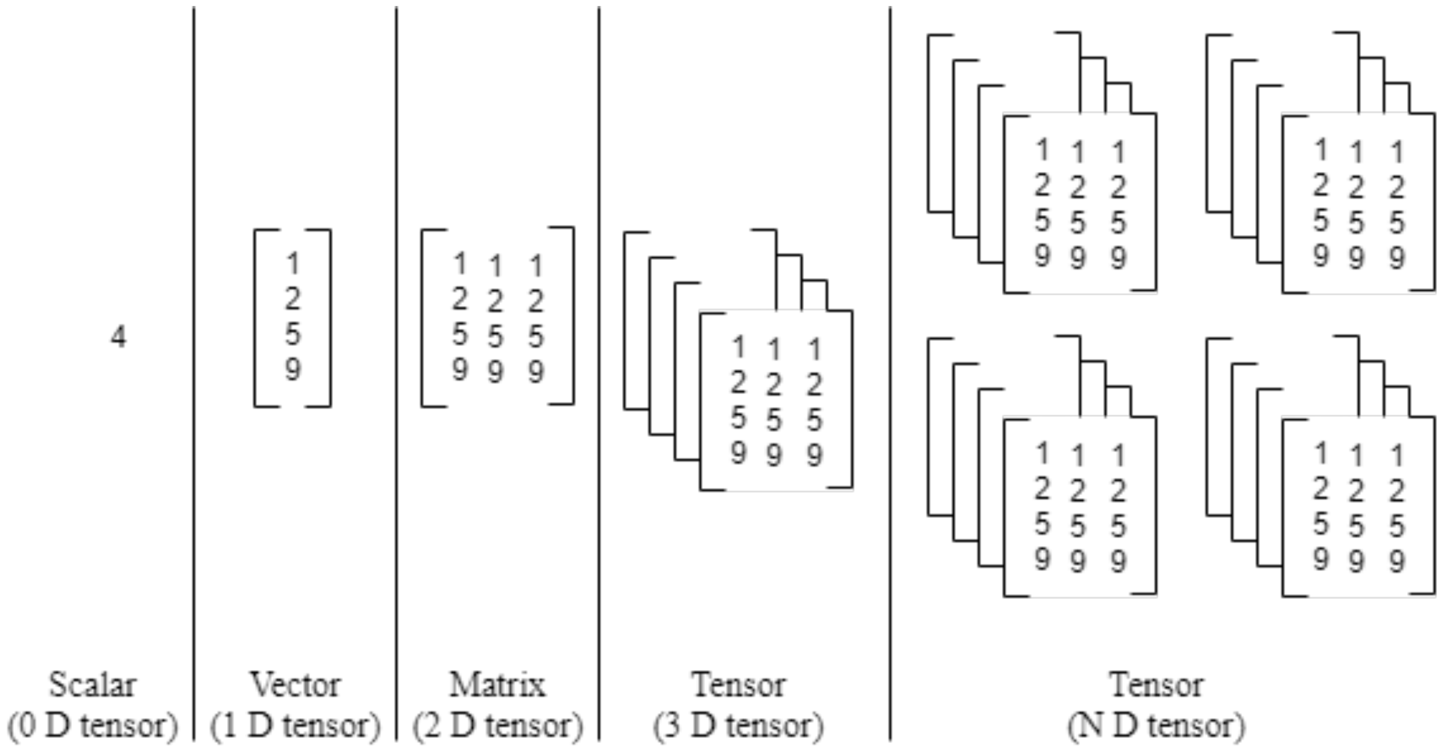
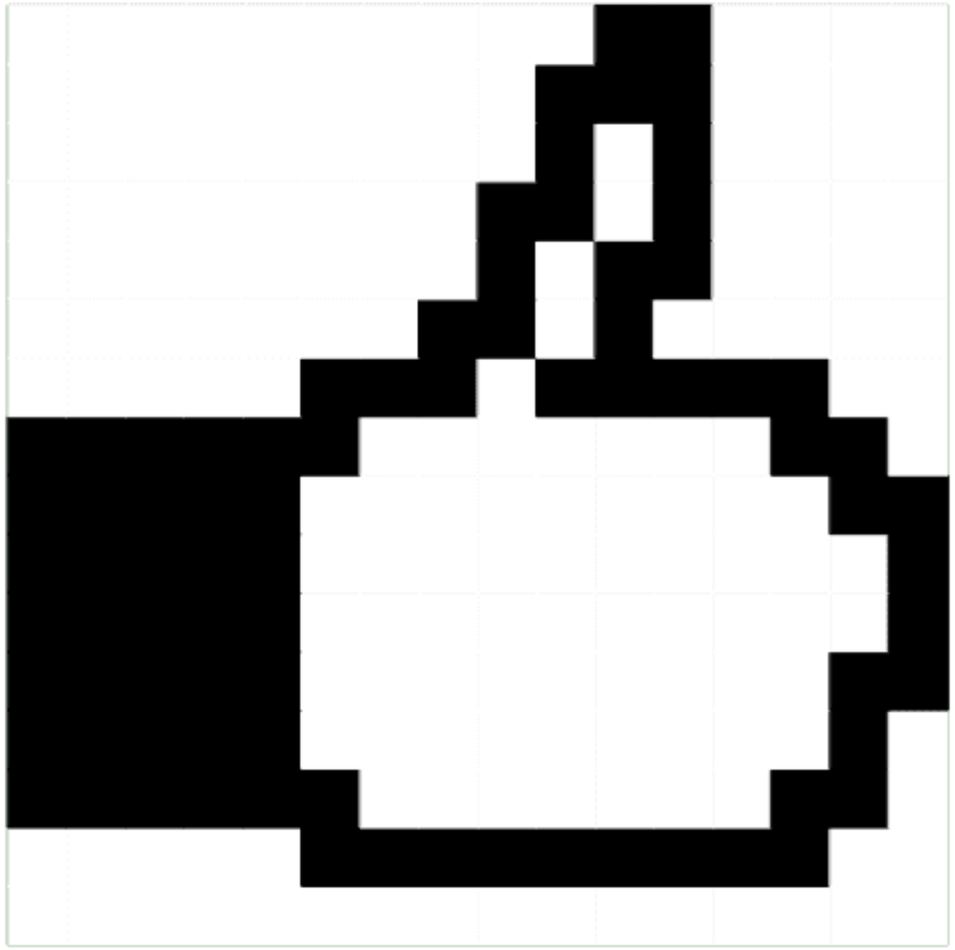


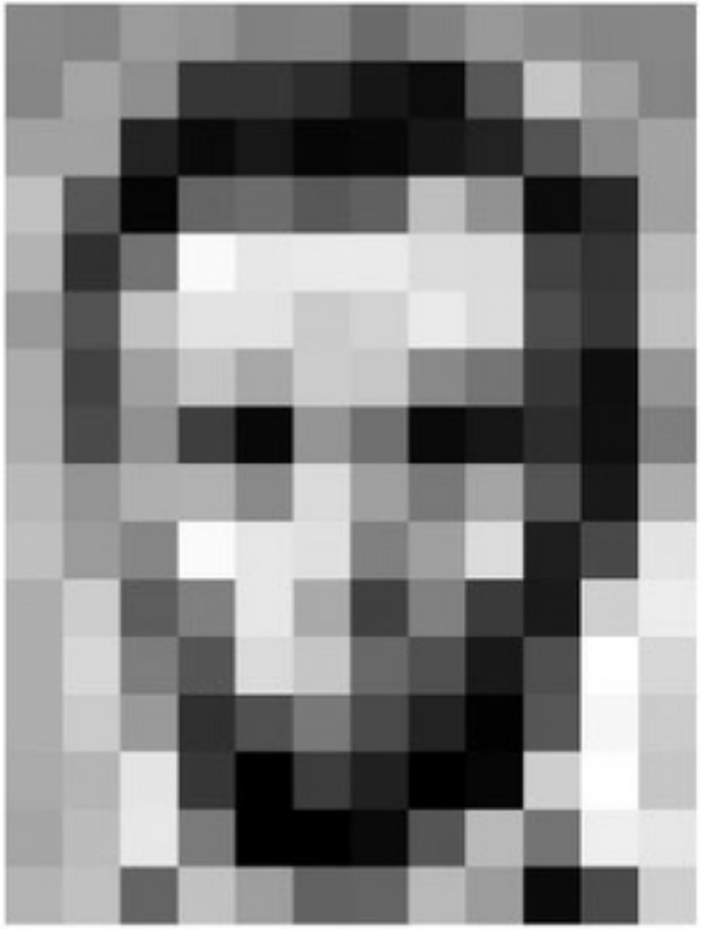
Image Processing

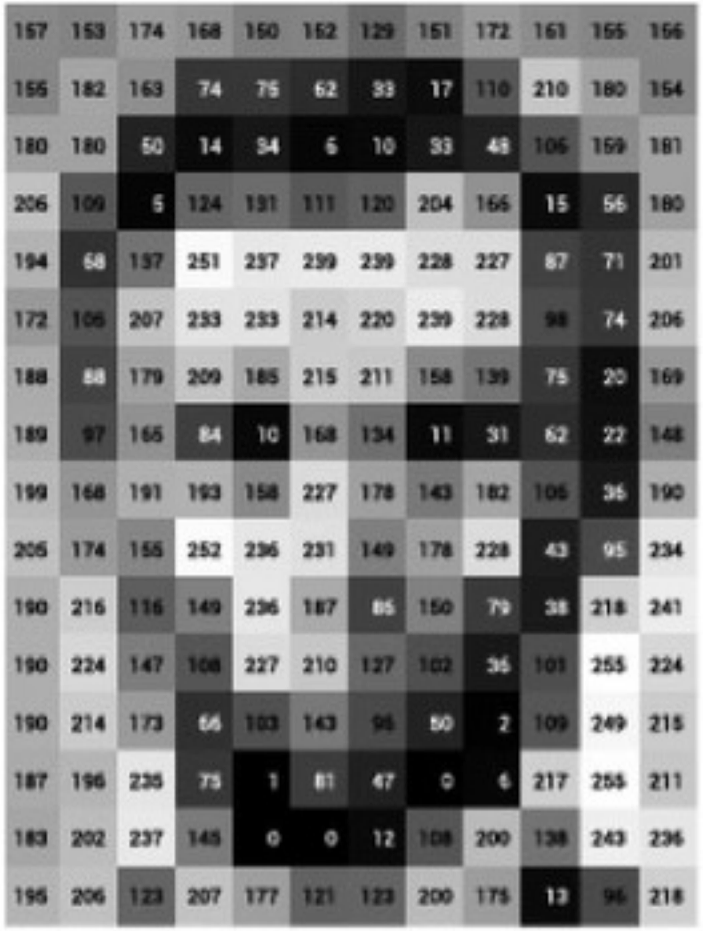
Images are a Matrix of Pixels

0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
0	0	0	0	0	1	1	1	0	1	1	1	1	1	0	0
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0
1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



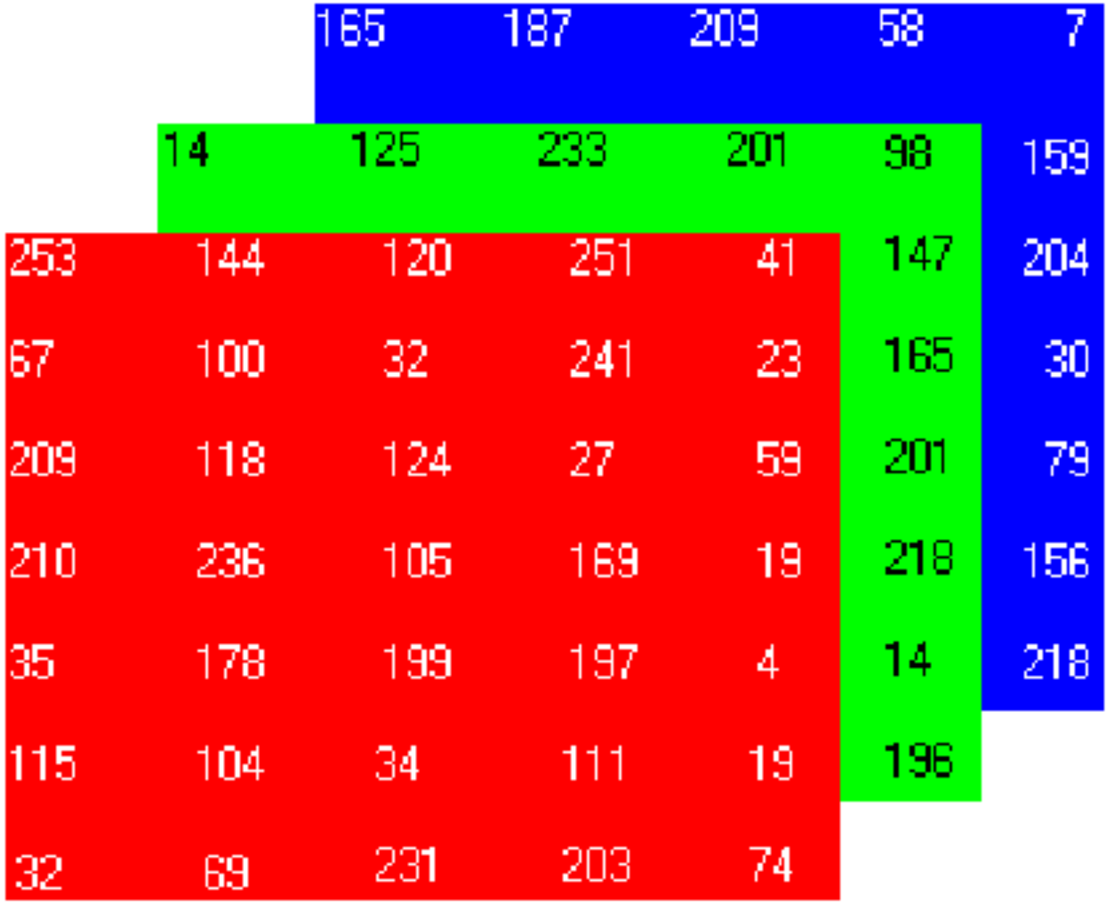
Images are a Matrix of Pixels





157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Images are a Matrix of Pixels



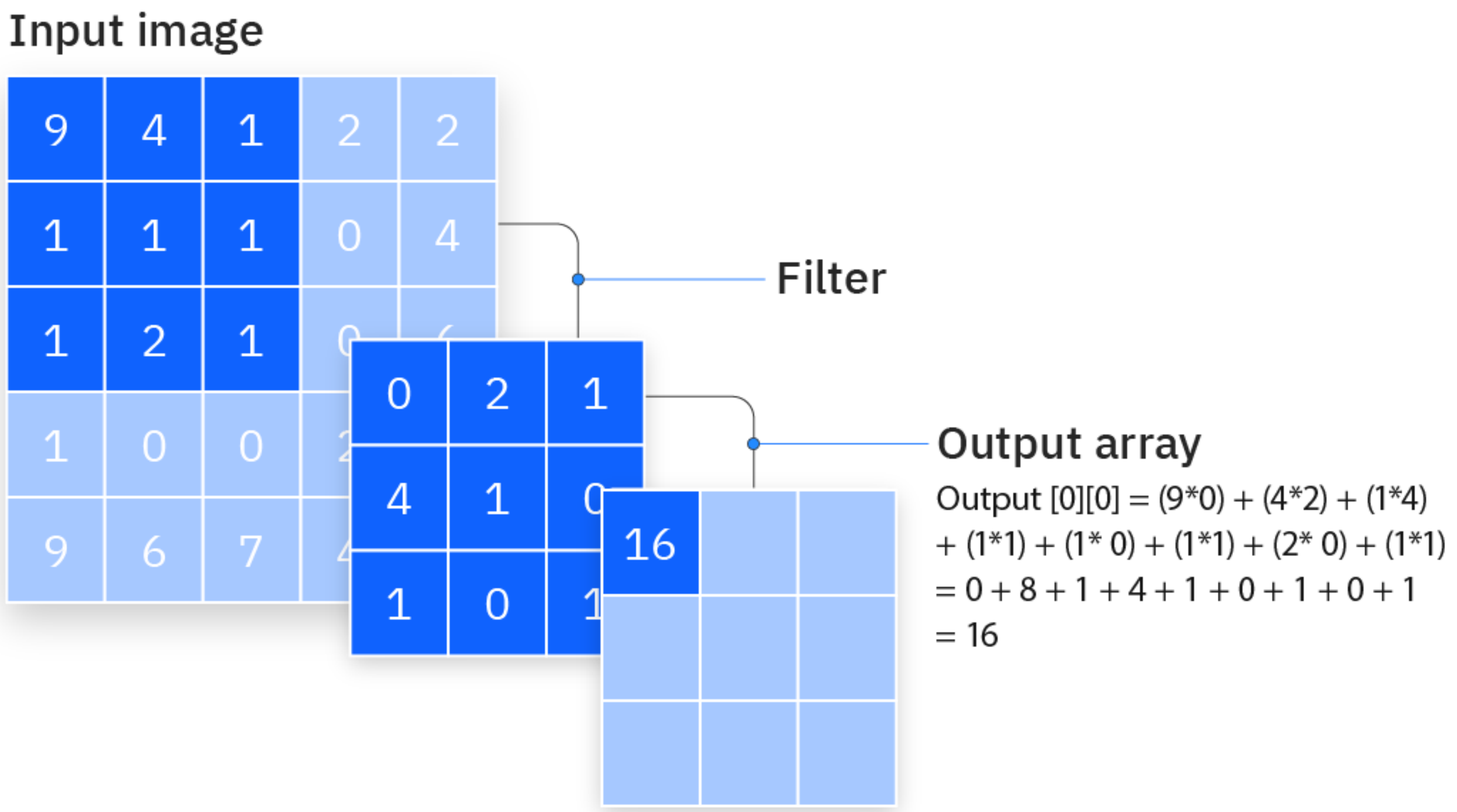
How do PhotoShop Effects Work?

Original Image

Kernel

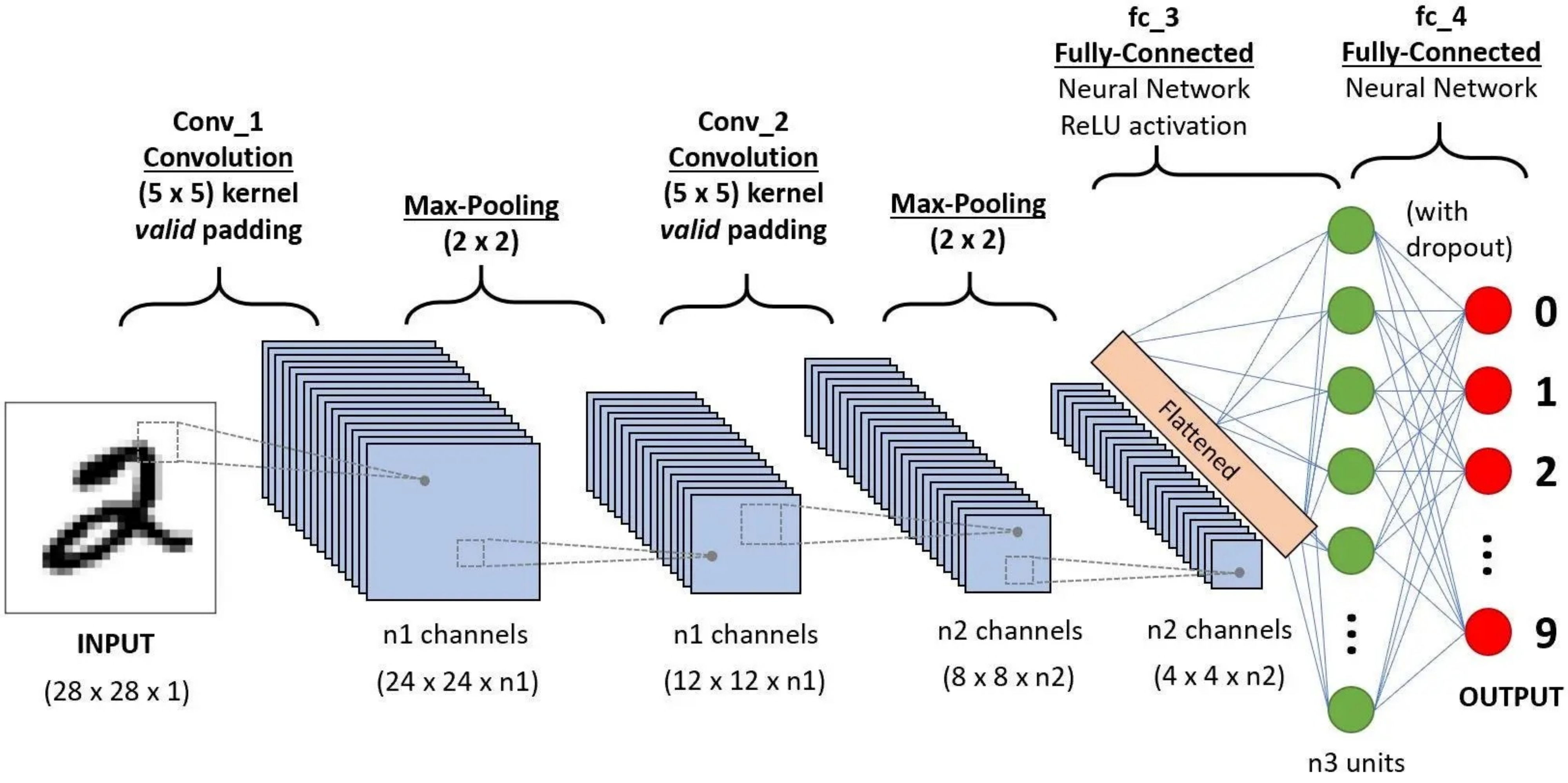
$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Filtered Image

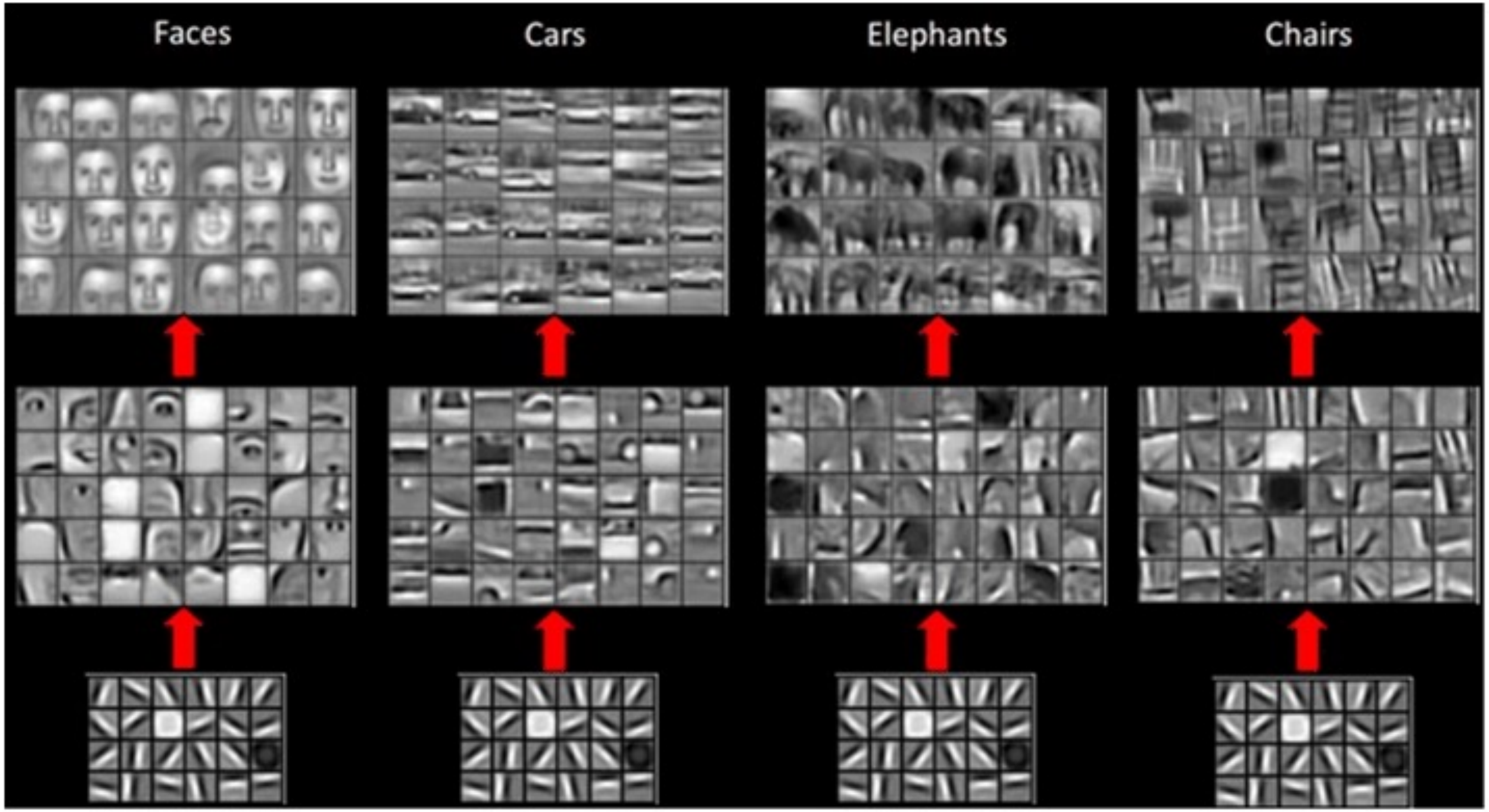

Input image
Filter
Output array
Output [0][0] = (9*0) + (4*2) + (1*4)
+ (1*1) + (1* 0) + (1*1) + (2* 0) + (1*1)
= 0 + 8 + 1 + 4 + 1 + 0 + 1 + 0 + 1
= 16

https://setosa.io/ev/image-kernels/

The Convolutional Neural Network



Visualizing Feature Maps



Explainable AI for Images: Saliency Maps

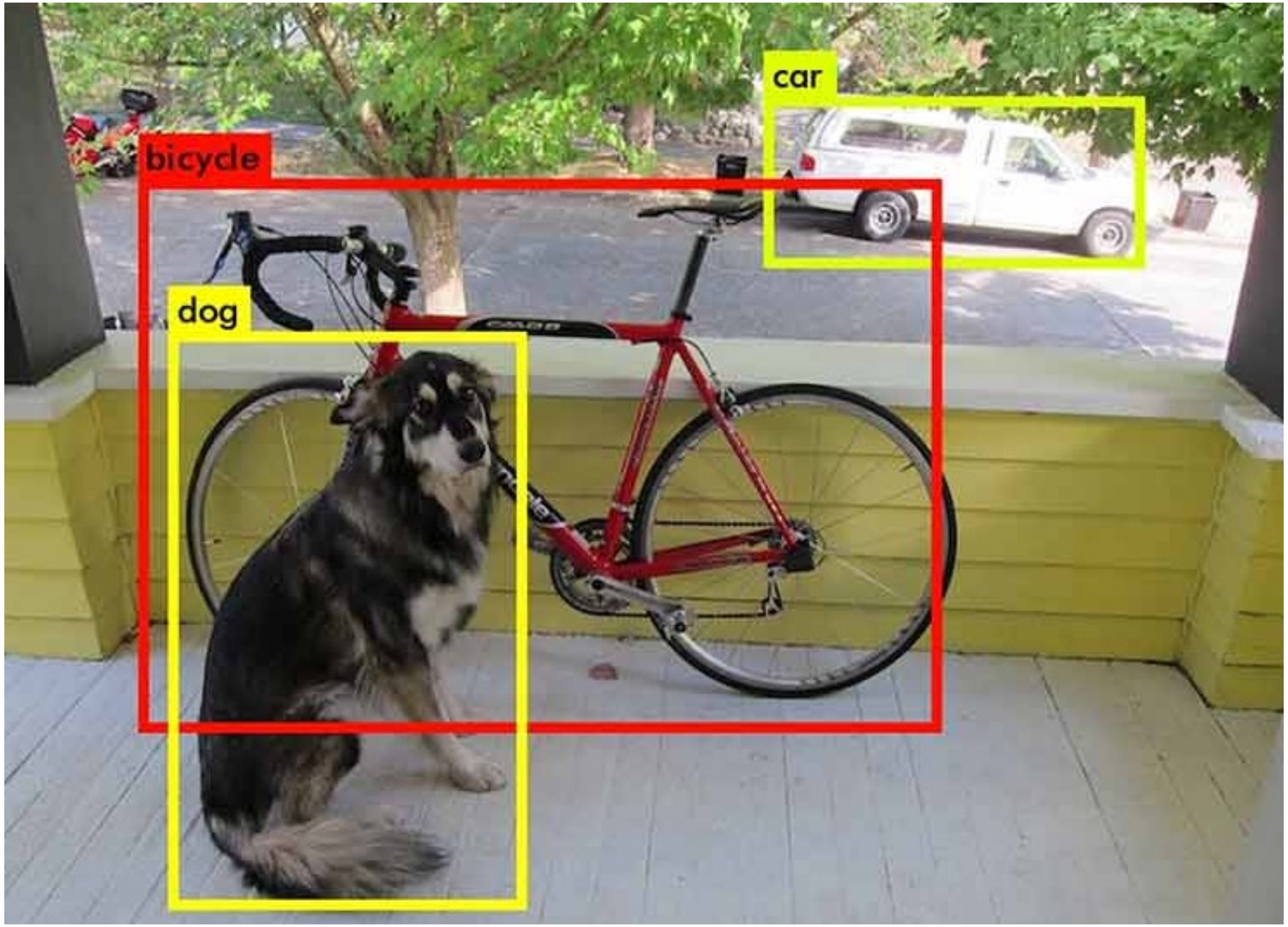
(a) Husky classified as wolf (b) Explanation

$$D = \left| \frac{\Delta probability}{\Delta pixel} \right|$$

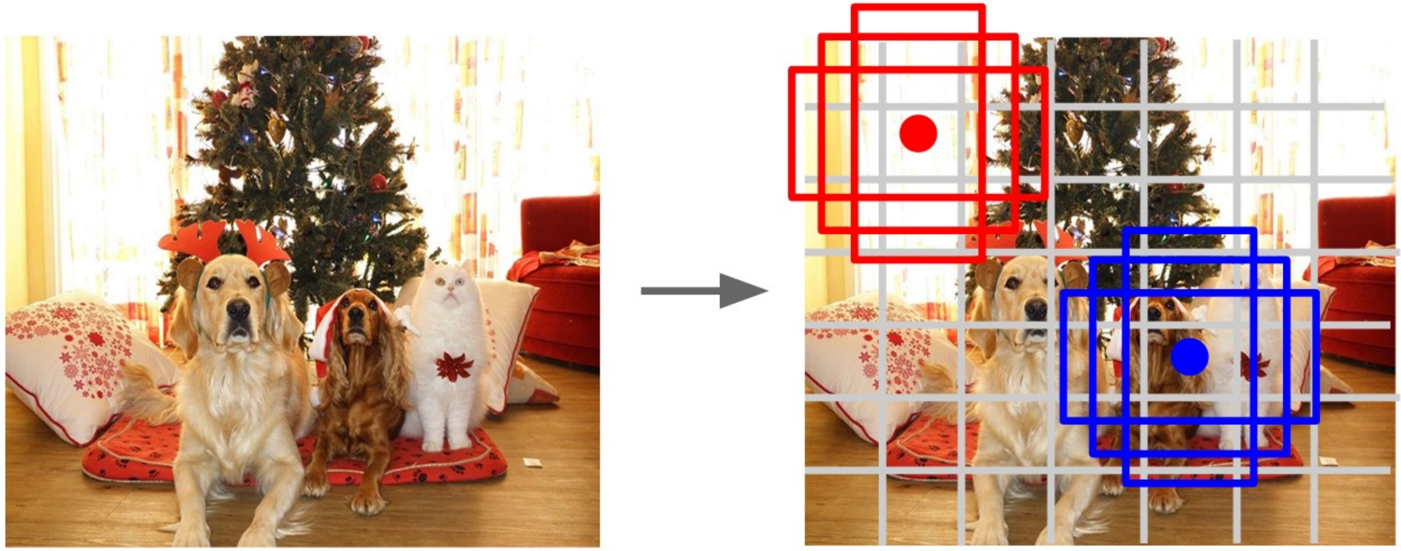
- Consider each pixel value in turn: R, G, B, then the next pixel.
- Make a copy of the image array before you change anything!
- Make the pixel value larger or smaller by various amounts. Each time, find the CNN's prediction with the changed value, and calculate the value of D.
- Repeat the previous step a few times, and calculate the pixel's saliency: the average value of D.
- Store the saliency of each pixel in a list, so that we can visualize it later.

Beyond Classification

Object Detection



YOLO Algorithm for Object Detection



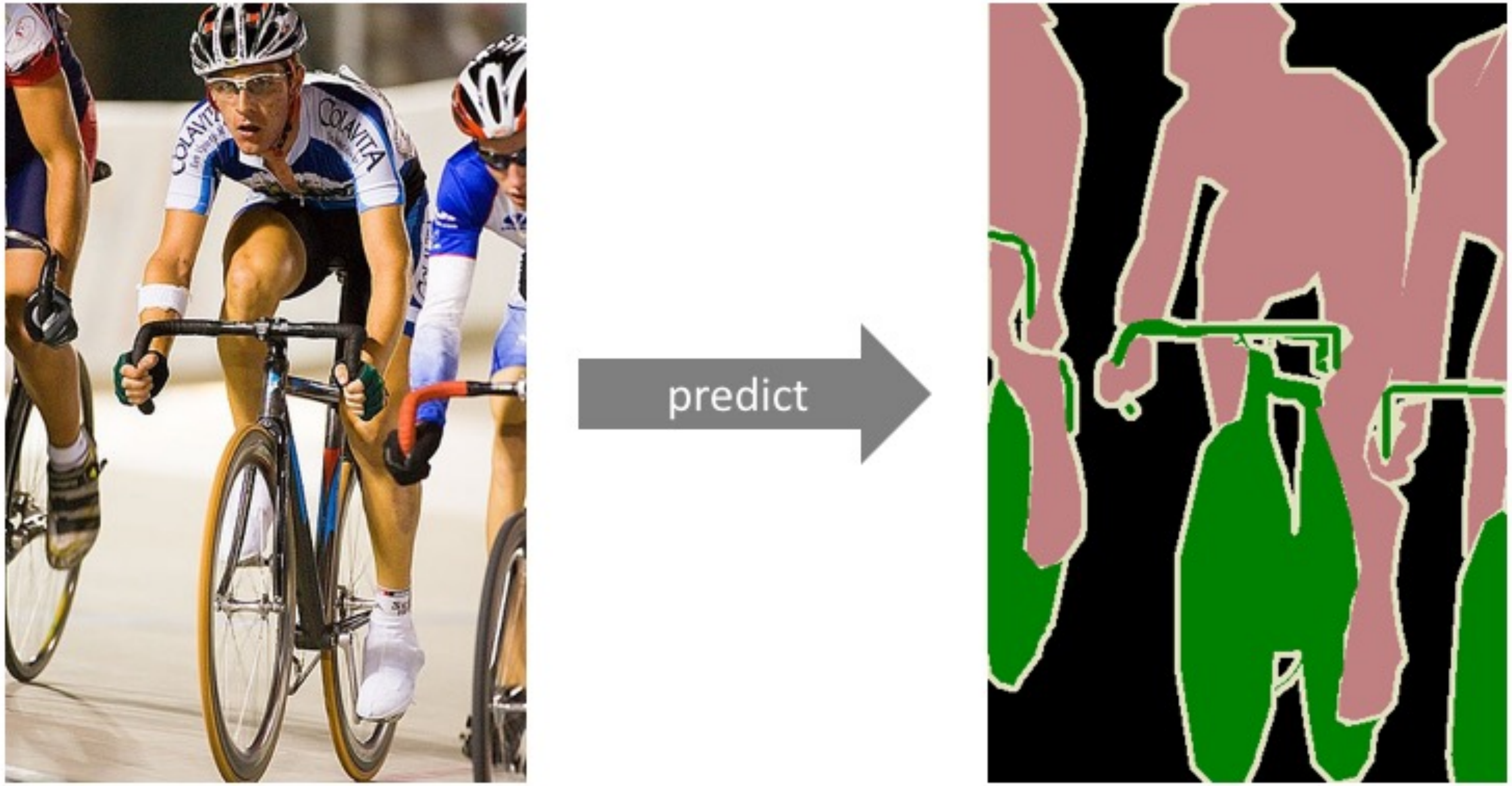
Divide image into a grid

Use a set of base boxes per grid

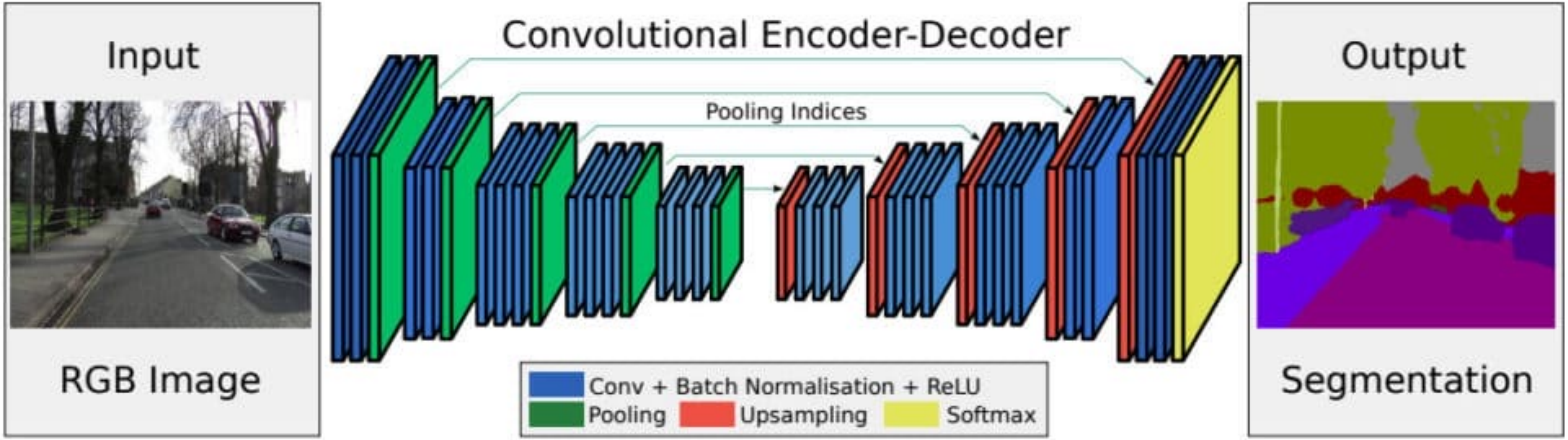
Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: box coordinates and confidence scores
- Predict scores for each of C classes (including background as a class)
- Looks a lot like a Region Proposal Network, but category-specific

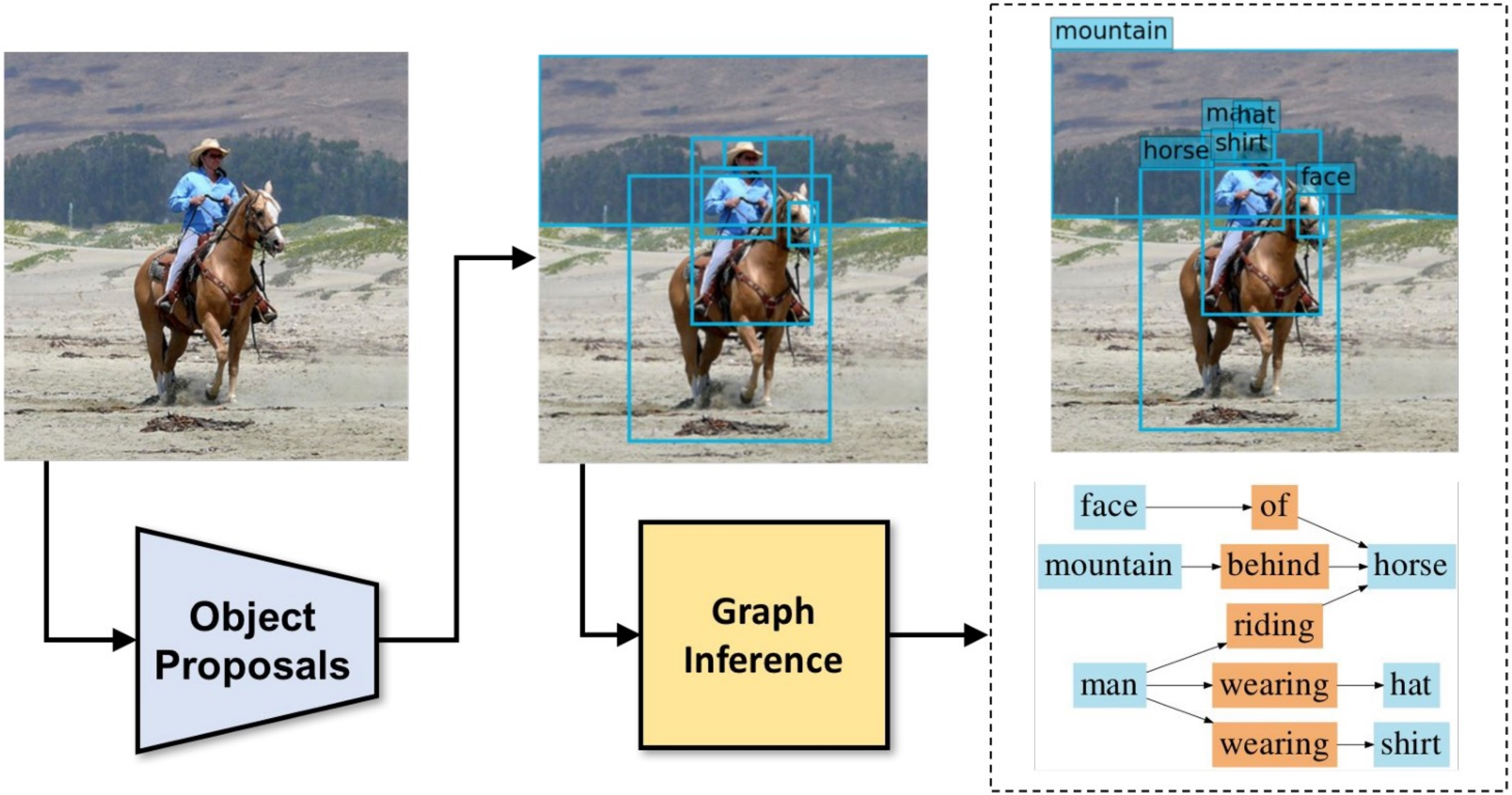
Segmentation



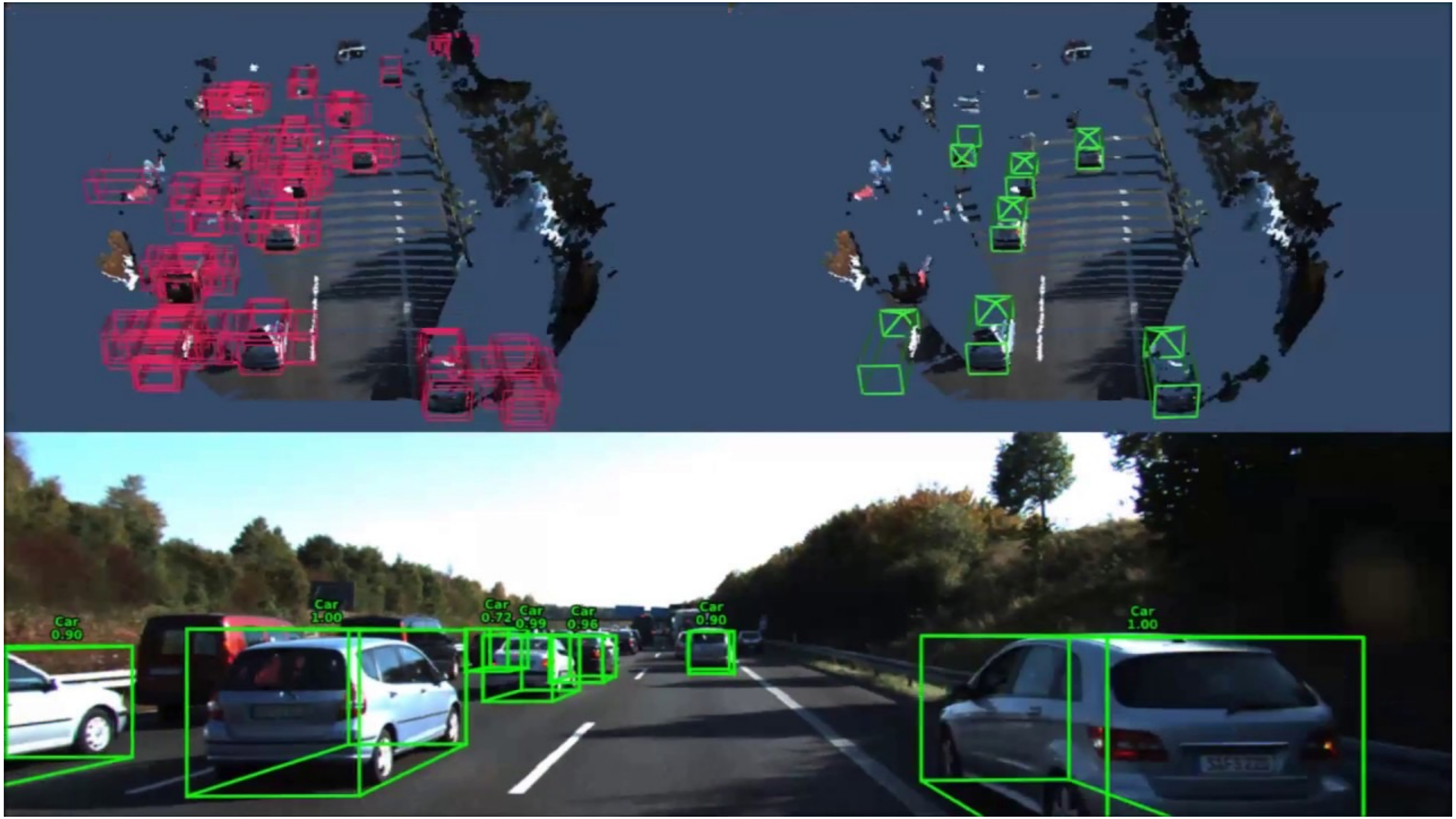
Person
Bicycle
Background


Input
RGB Image
Convolutional Encoder-Decoder
Pooling Indices
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax
Output
Segmentation

Knowledge Scene Graph Prediction

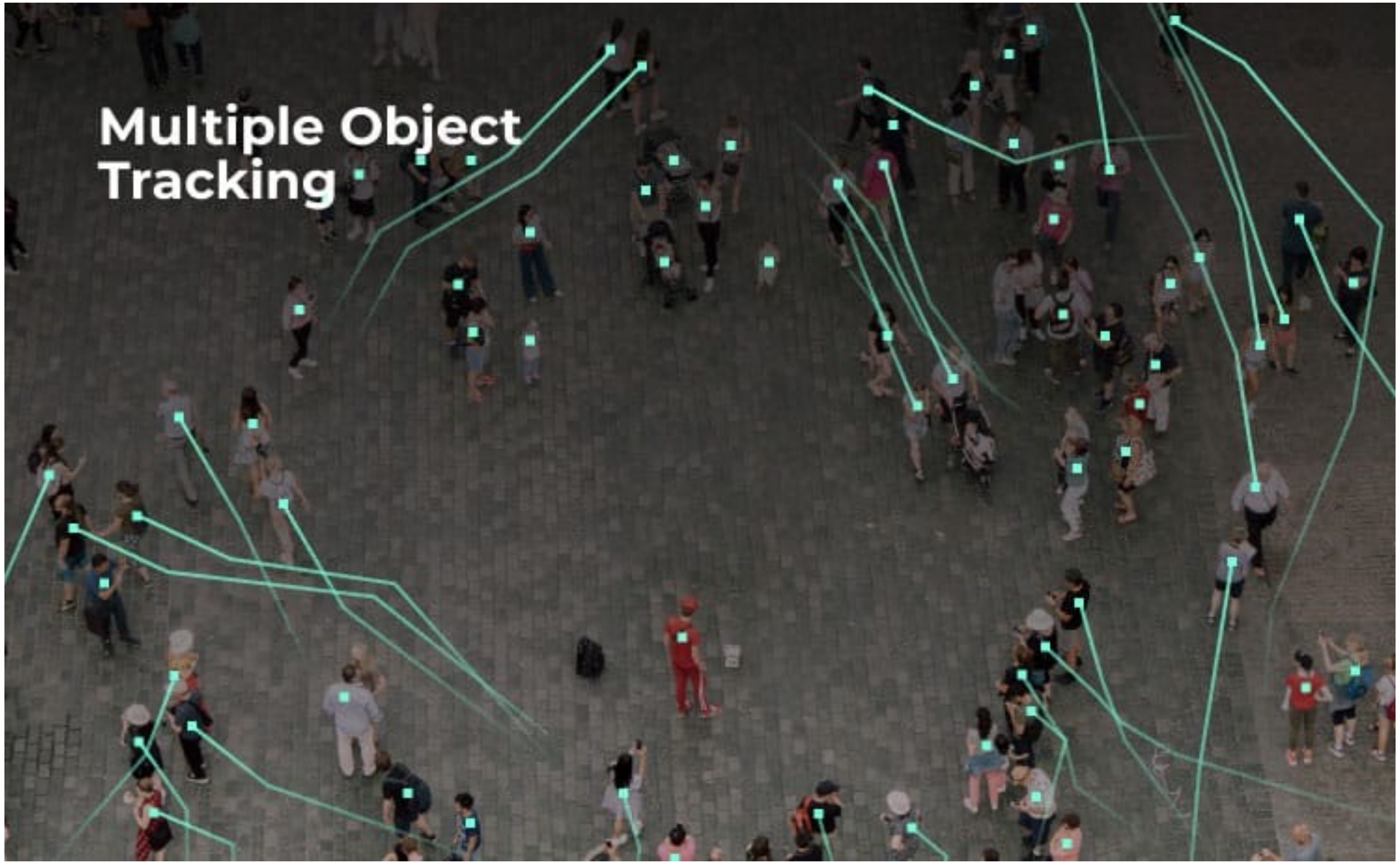


3D Object Detection



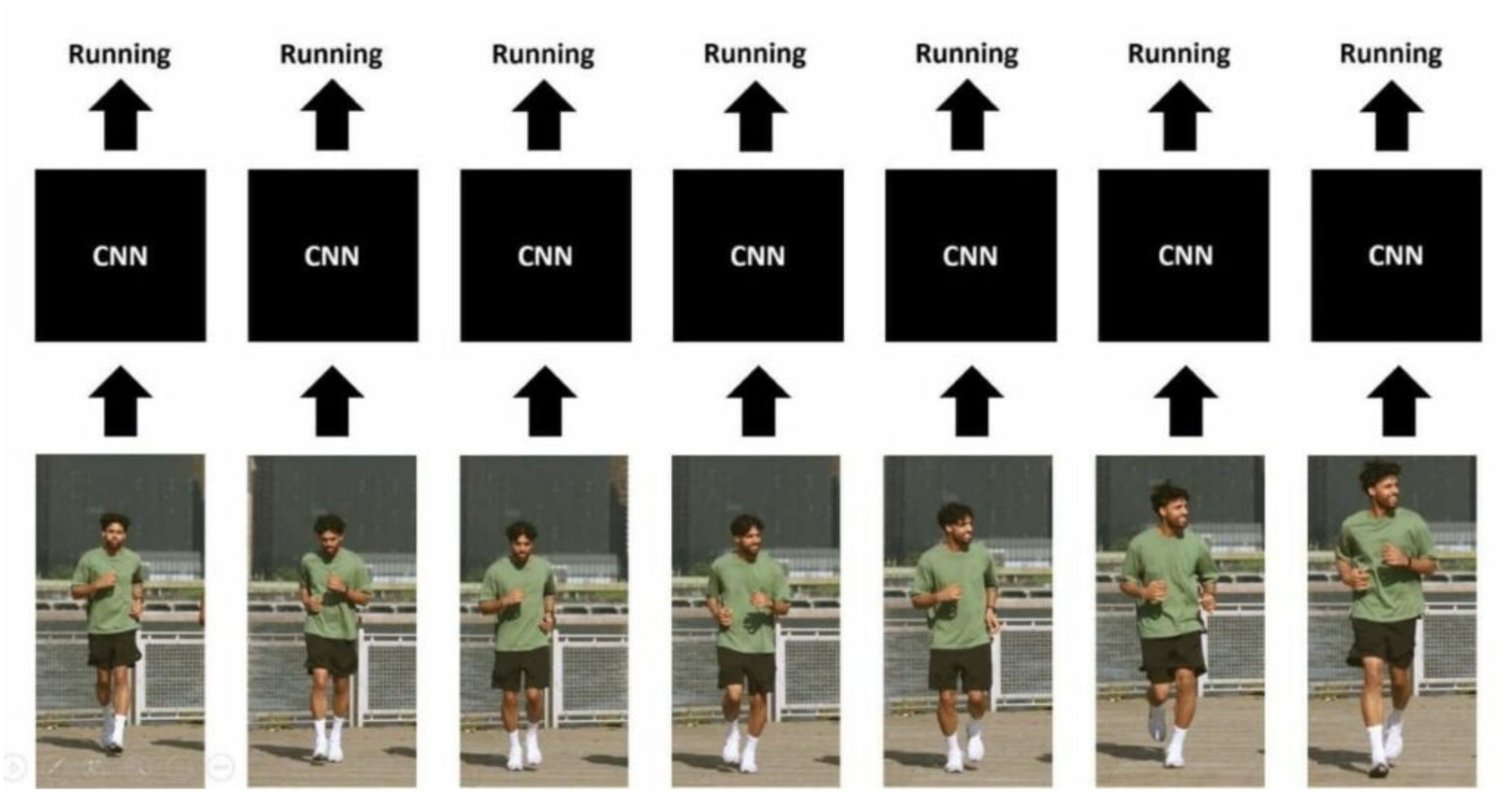
Pose Estimation

Object Tracking



Video Processing

Like Image Processing, But On a Per-Frame Basis

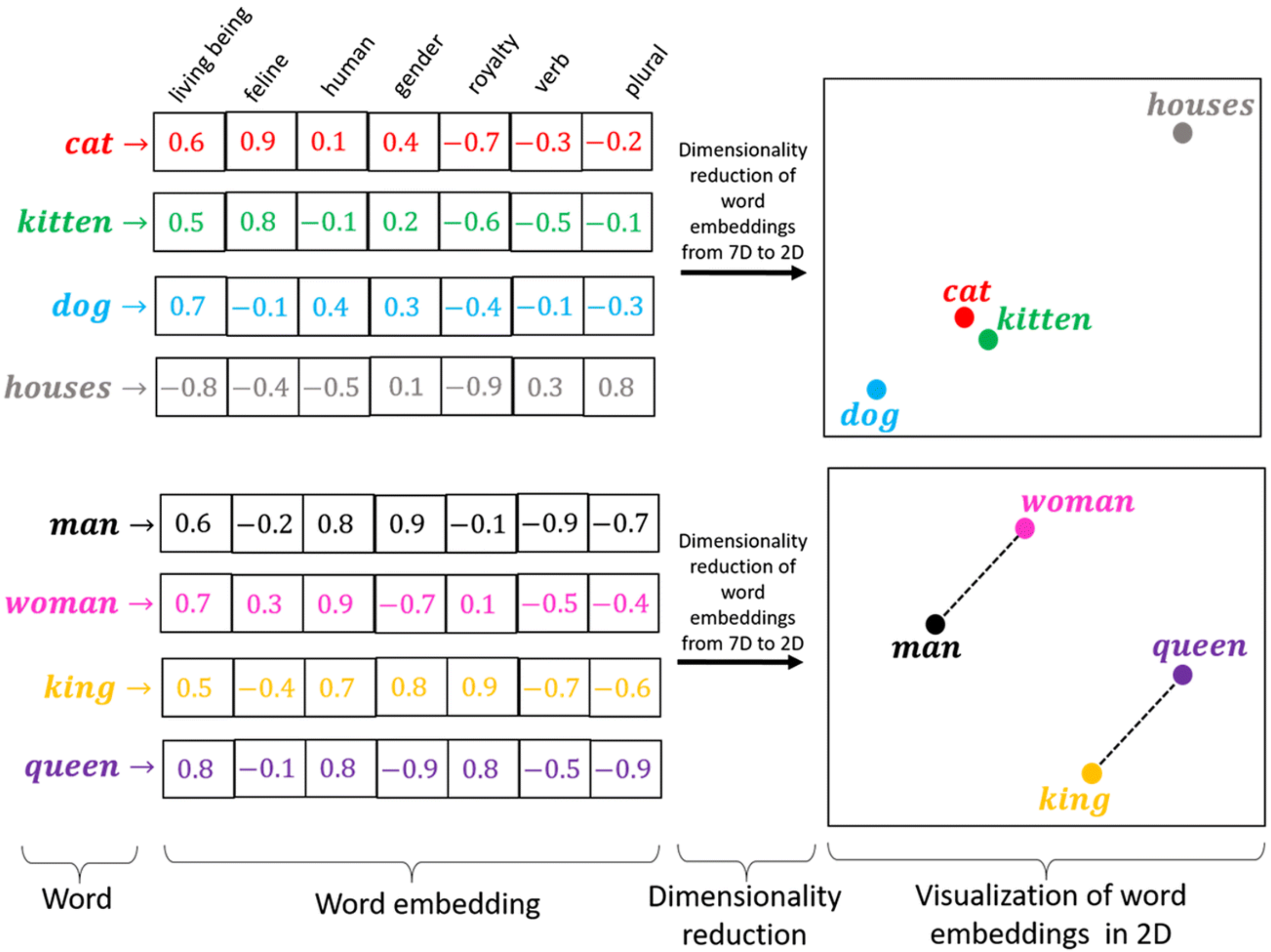


Text Processing

How to Represent Text as Numbers? Bag of Words

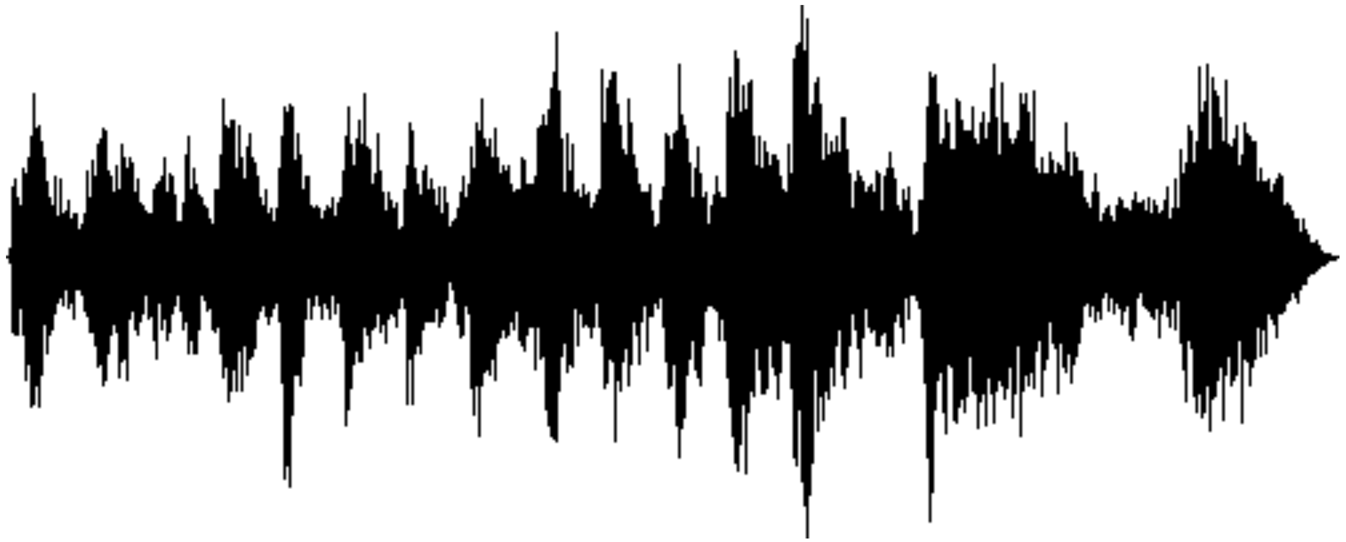
	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

How to Represent Text as Numbers? word2vec



Speech Processing

How to Represent Audio as Numbers?
Waveforms are just 1D arrays!



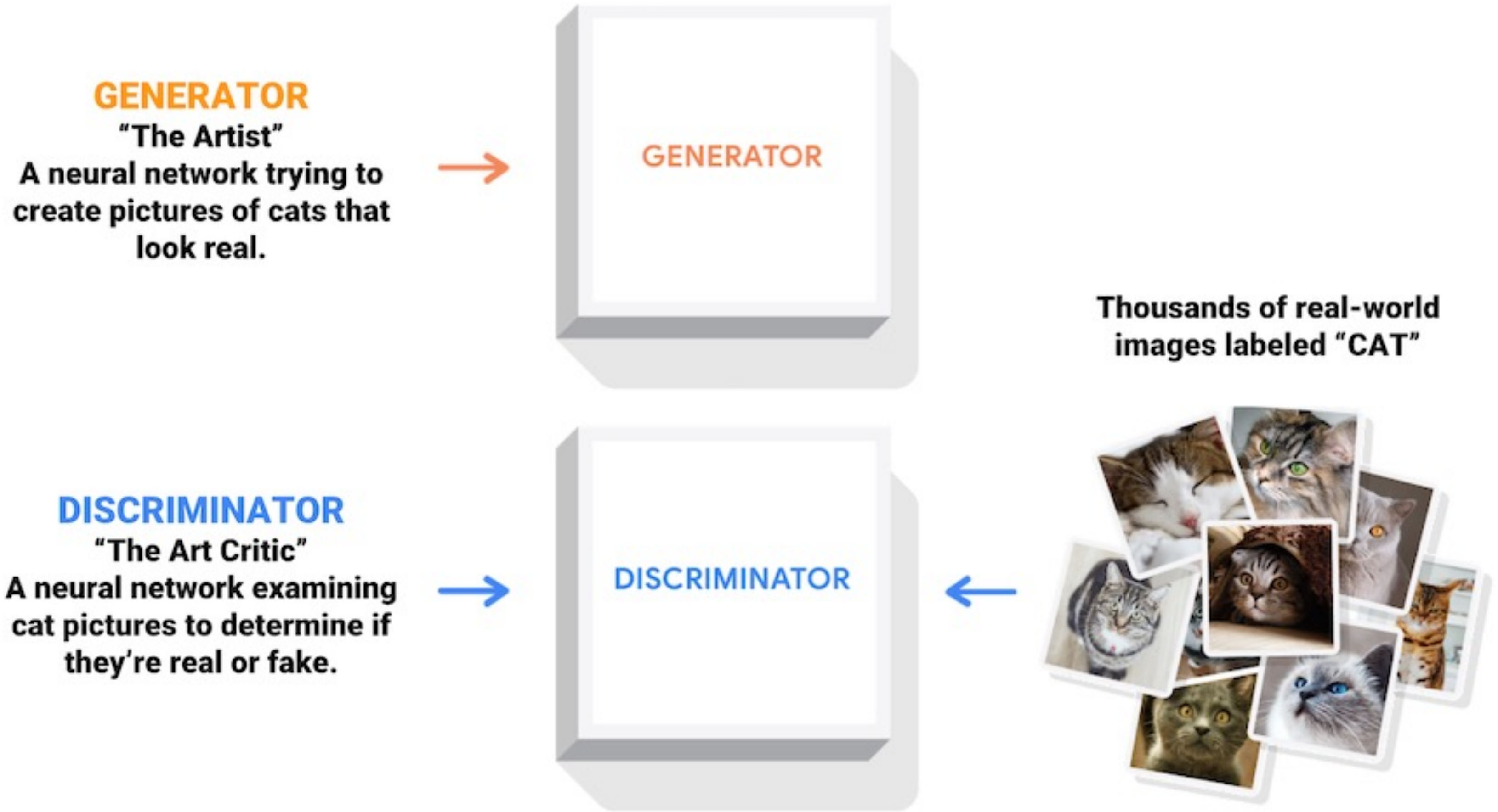
2	8	7	6	0

Generating Synthetic Data (e.g., DeepFakes)

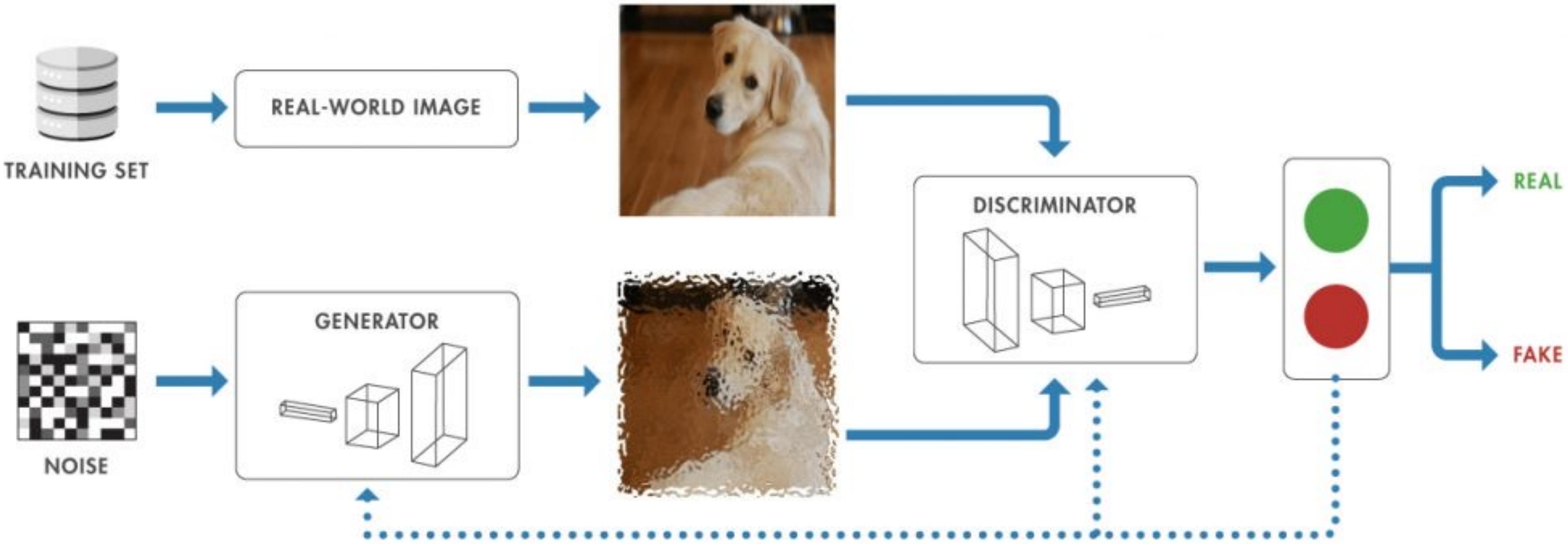
- GANS

- Diffusion Models

Generative Adversarial Network (GAN)



Generative Adversarial Network (GAN)

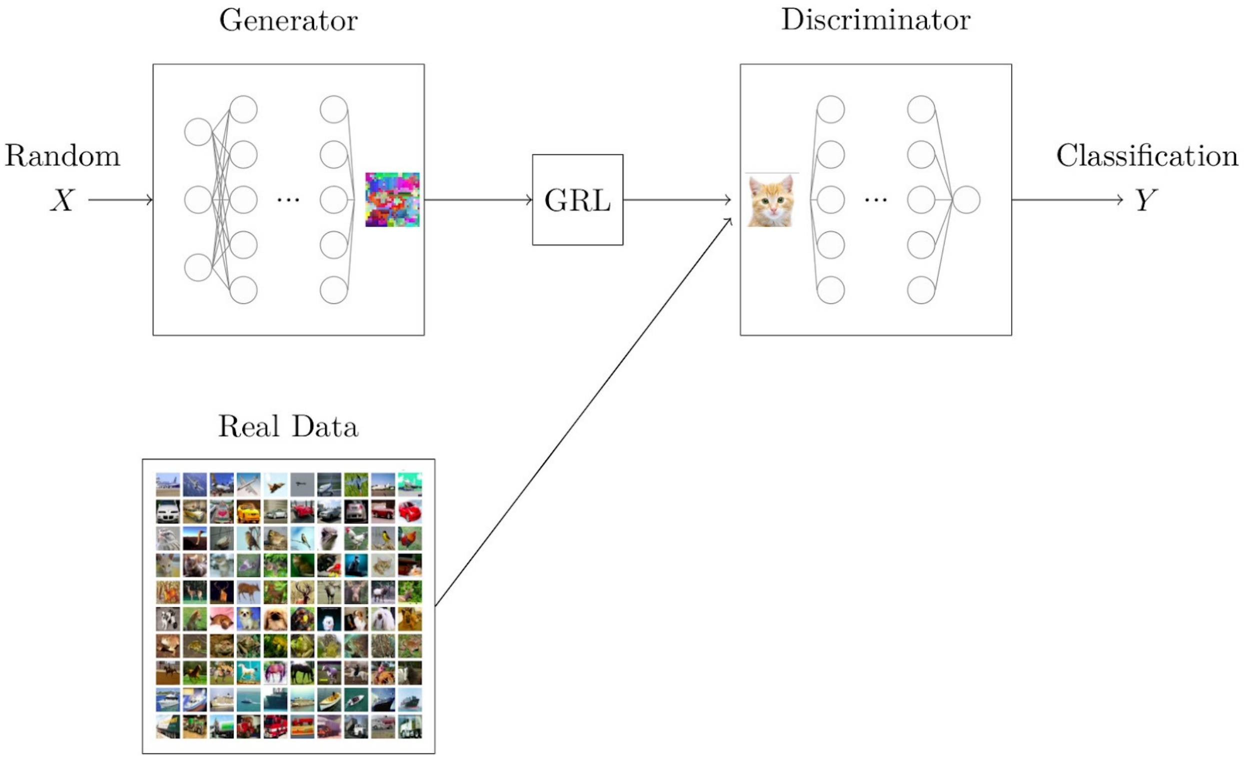


GAN Training Process

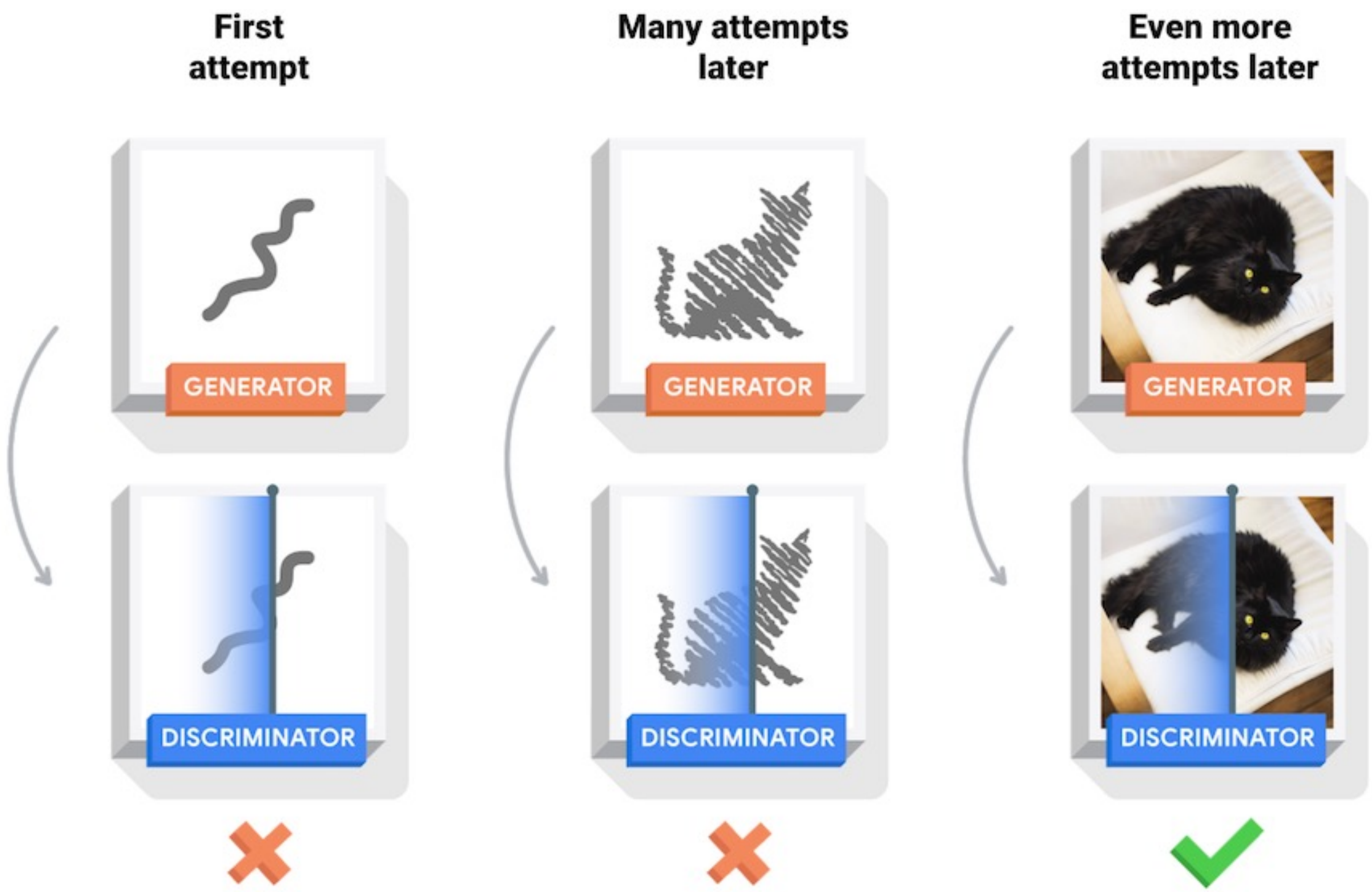
- Generator G produces noise
- Discriminator D learns to classify noise vs. real
- D tells G how to make noise look more real
- G starts generating real-looking images

While True:

  - D gets confused, tries harder to distinguish real vs. fake images
  - G gets better at generating fake images
  - D gets better at identifying fake images



GAN Training Process



Multiple Interacting Neural Networks

This is the first time in this class where we build a single system with **multiple neural networks** which **interact with each other**

This is a recurring theme in many new areas of deep learning

GAN Loss Function

$$\underbrace{\min_G \max_D V(D, G)}_{\text{Minimize loss for Generator; Maximize loss for Discriminator}}$$

$$V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} \underbrace{[\log D(x)]}_{\substack{\text{Discriminator} \\ \text{output for} \\ \text{real data x}}} + \mathbb{E}_{z \sim p_z(z)} [\log(1 - \underbrace{D(G(z))}_{\substack{\text{Discriminator output} \\ \text{for generated fake} \\ \text{data G(z)}}})]$$

GAN Loss Function

$$\min_G \max_D V(D,G)$$

For Discriminator:

$$V(D,G) = \overbrace{\mathbb{E}_{x\sim p_{data}(x)}[\log \underbrace{D(x)]}_{\text{Discriminator output for real data x: as close to 1 as possible}}}^{\text{Maximize to get D(X) as close to 1}} + \overbrace{\mathbb{E}_{z\sim p_z(z)}[\log(1 - \underbrace{D(G(z))}_{\text{Discriminator output for generated fake data G(z): as close to 0 as possible}})]}^{\text{Maximize to get D(G(z)) as close to 0}}$$

GAN Loss Function

For Generator (only cares about generated images):

$$\overbrace{\mathbb{E}_{z\sim p_z(z)}[\log(1-\underbrace{D(G(z))}_{\text{Discriminator output for generated fake data G(z): as close to 1 as possible}})]}^{\text{Minimize to get D(G(z)) as close to 1}}$$

GAN Training Process

<u>Alternate between:</u>

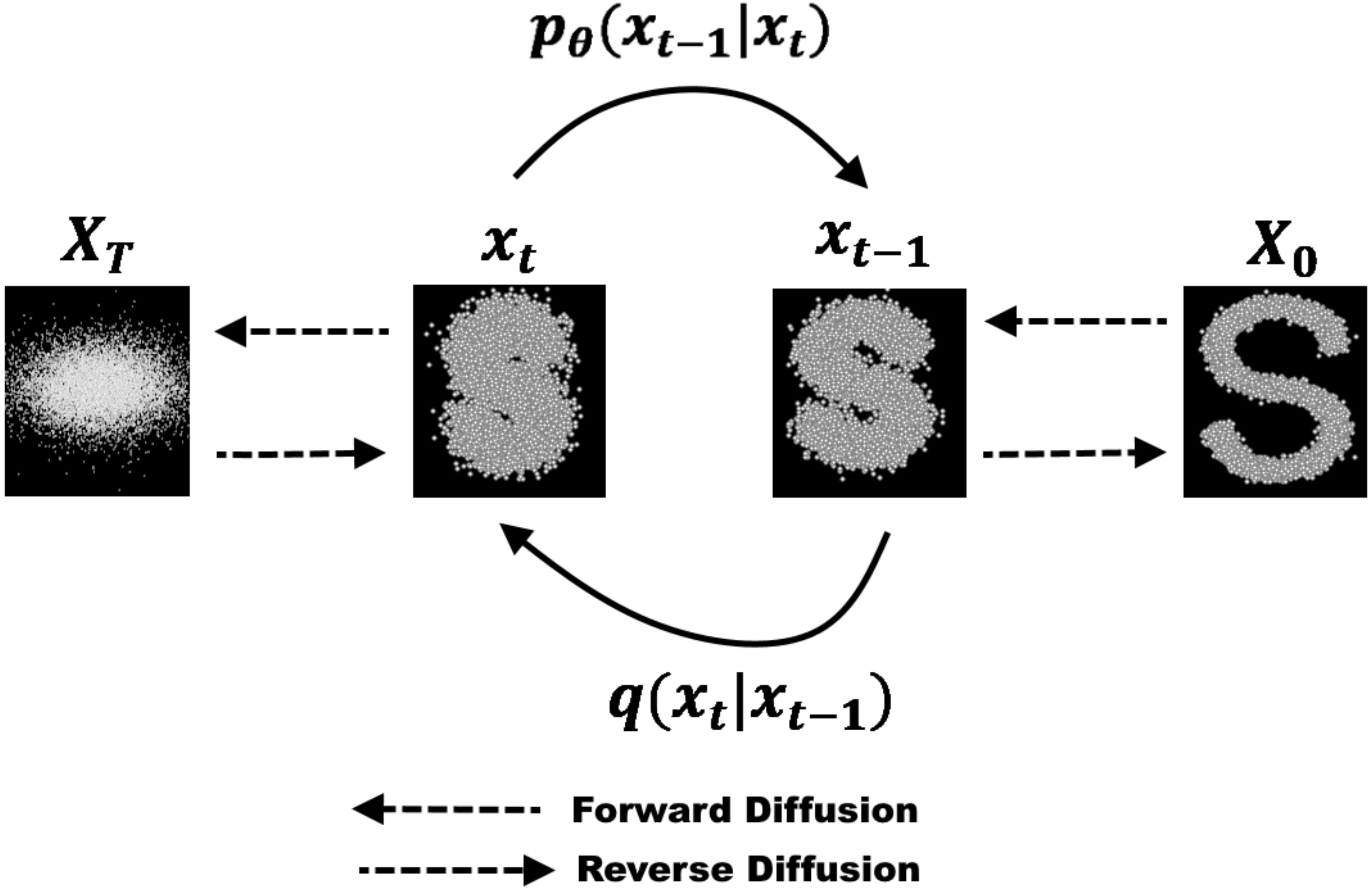
- Gradient ascent for Discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

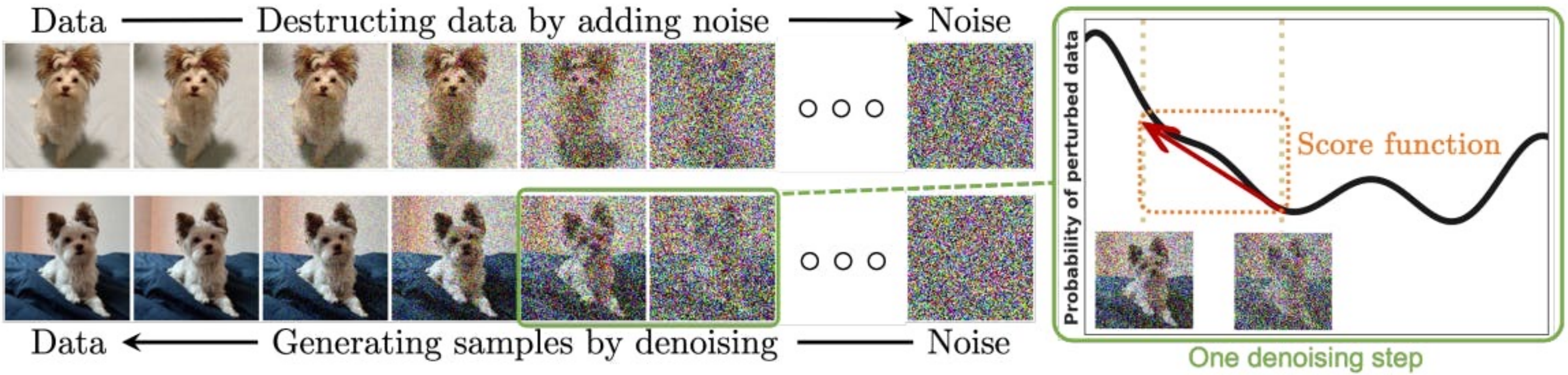
- Gradient descent for Generator

$$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

Diffusion Models



Diffusion Models

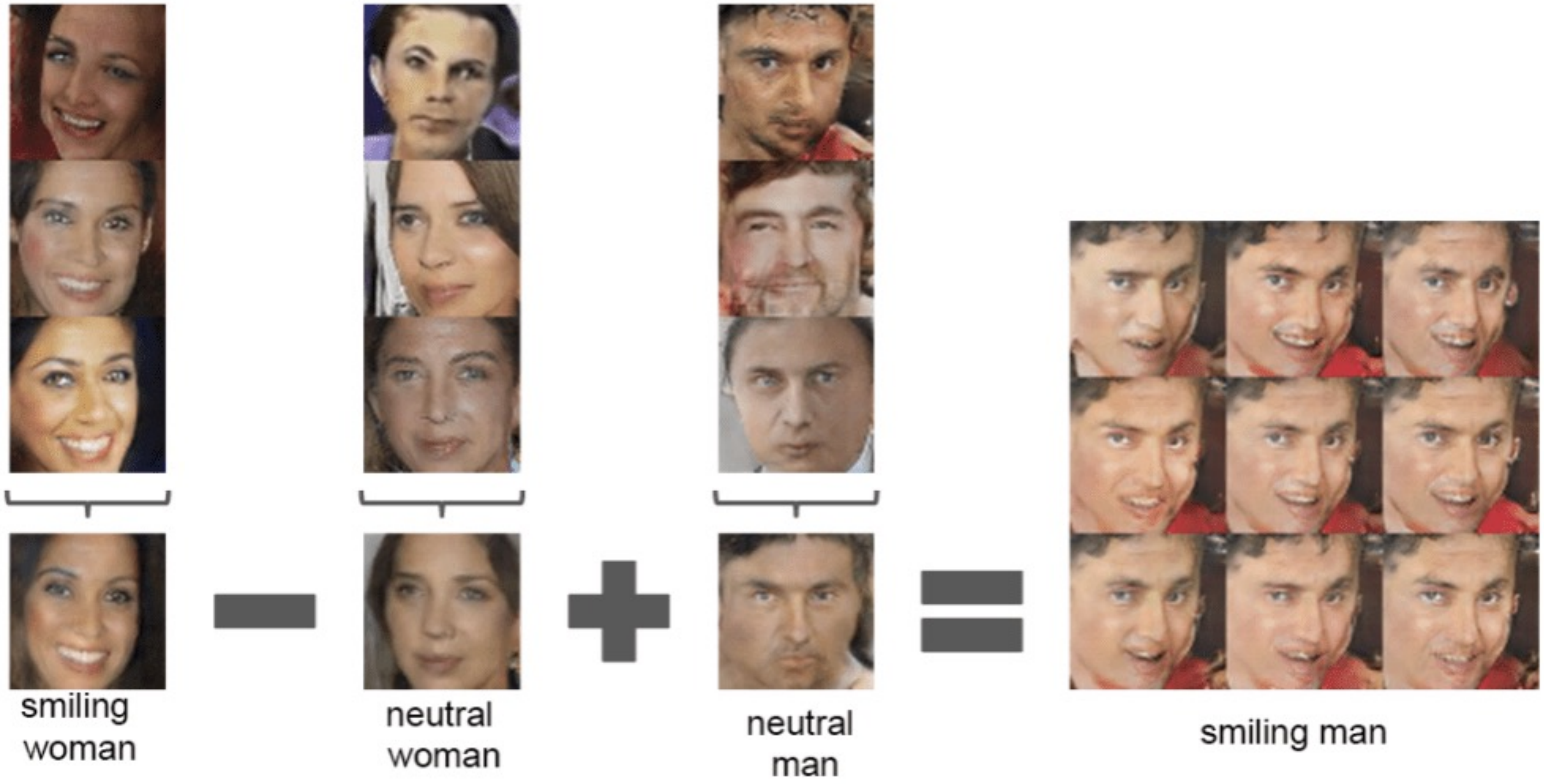


Applications of Generative Models

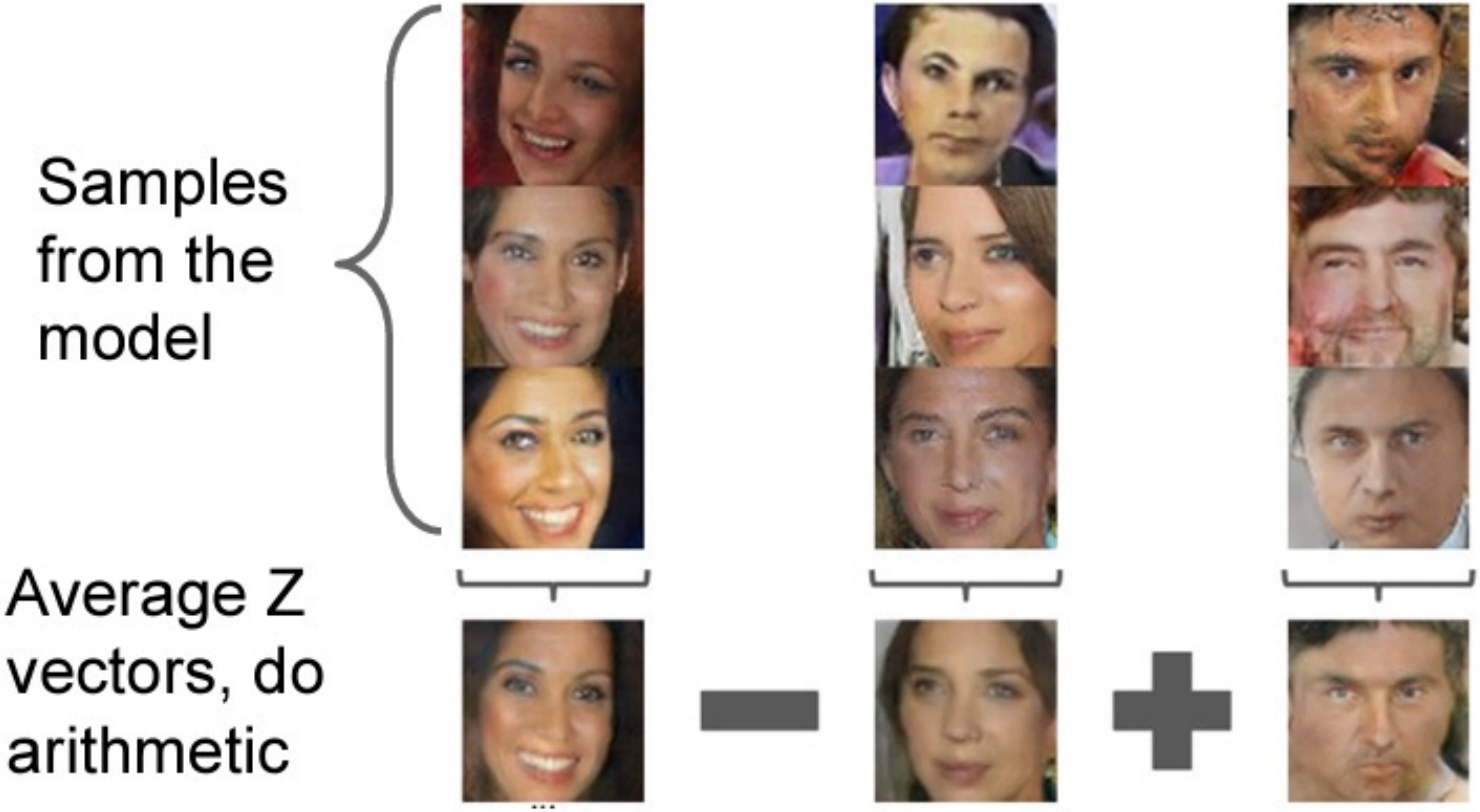
www.thispersondoesnotexist.com

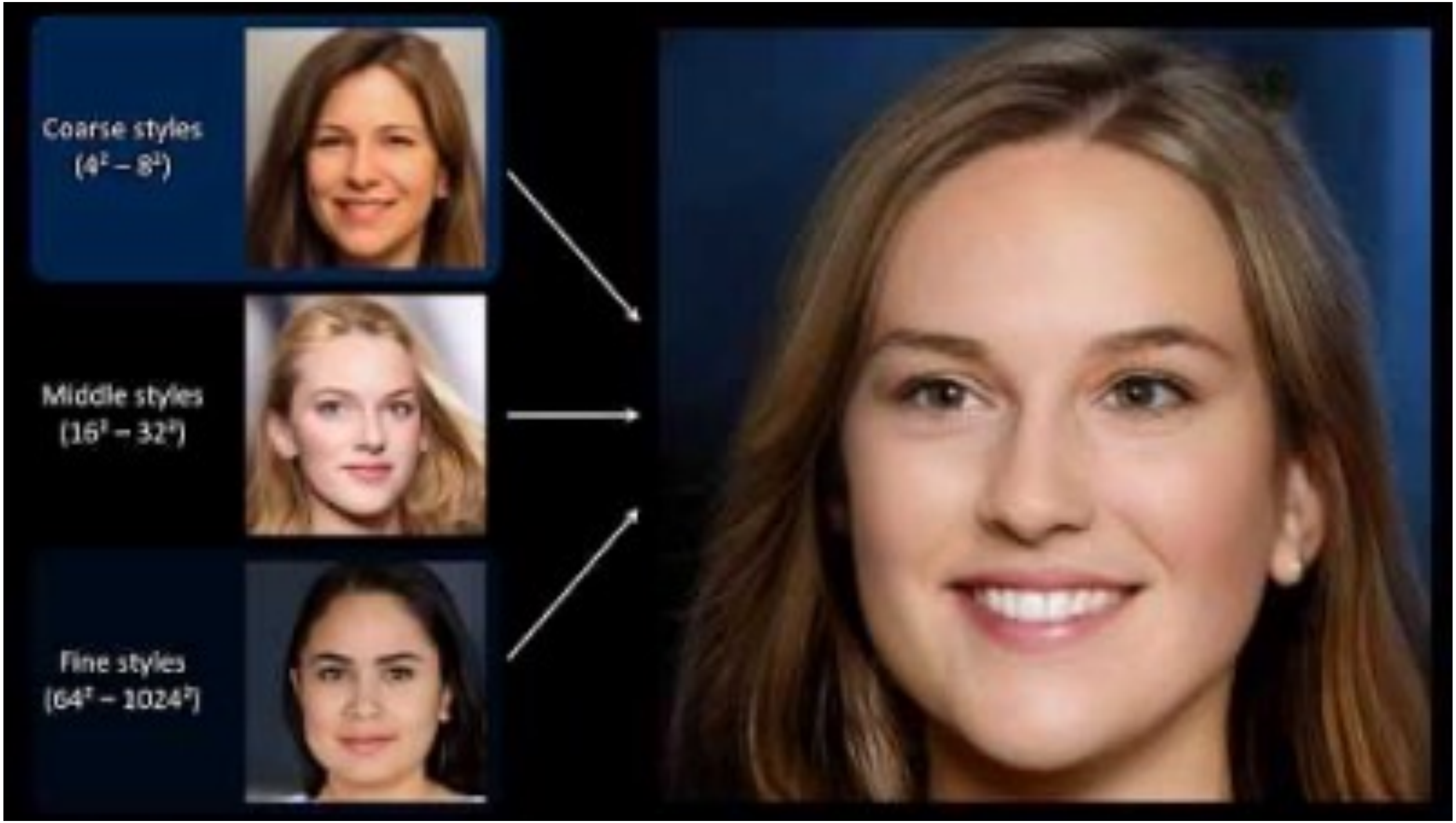
Interpolating between random points in latent space

Latent Space Math



Latent Space Math




Coarse styles
($4^2 - 8^2$)
Middle styles
($16^2 - 32^2$)
Fine styles
($64^2 - 1024^2$)

Obama Deepfake (2018)

Volodymyr Zelenskyy Deepfake (2022)

Text to Image