

# Sequence Models for words and pixels

---



© A.A. Efros

Many slides from Steve Seitz's  
wonderful [5 min Lectures](#)

CS194: Intro to Computer Vision & Comp. Photography  
Alexei Efros, UC Berkeley, Fall 2024

---

# Michel Gondry train video

<http://www.youtube.com/watch?v=0S43lwBF0uM>

# “Amateur” by Lasse Gjertsen

---

<http://www.youtube.com/watch?v=JzqumbhfxRo>

# Generative AI

---

Generative models trained on **lots of data** have revolutionized computer science and beyond!



Text



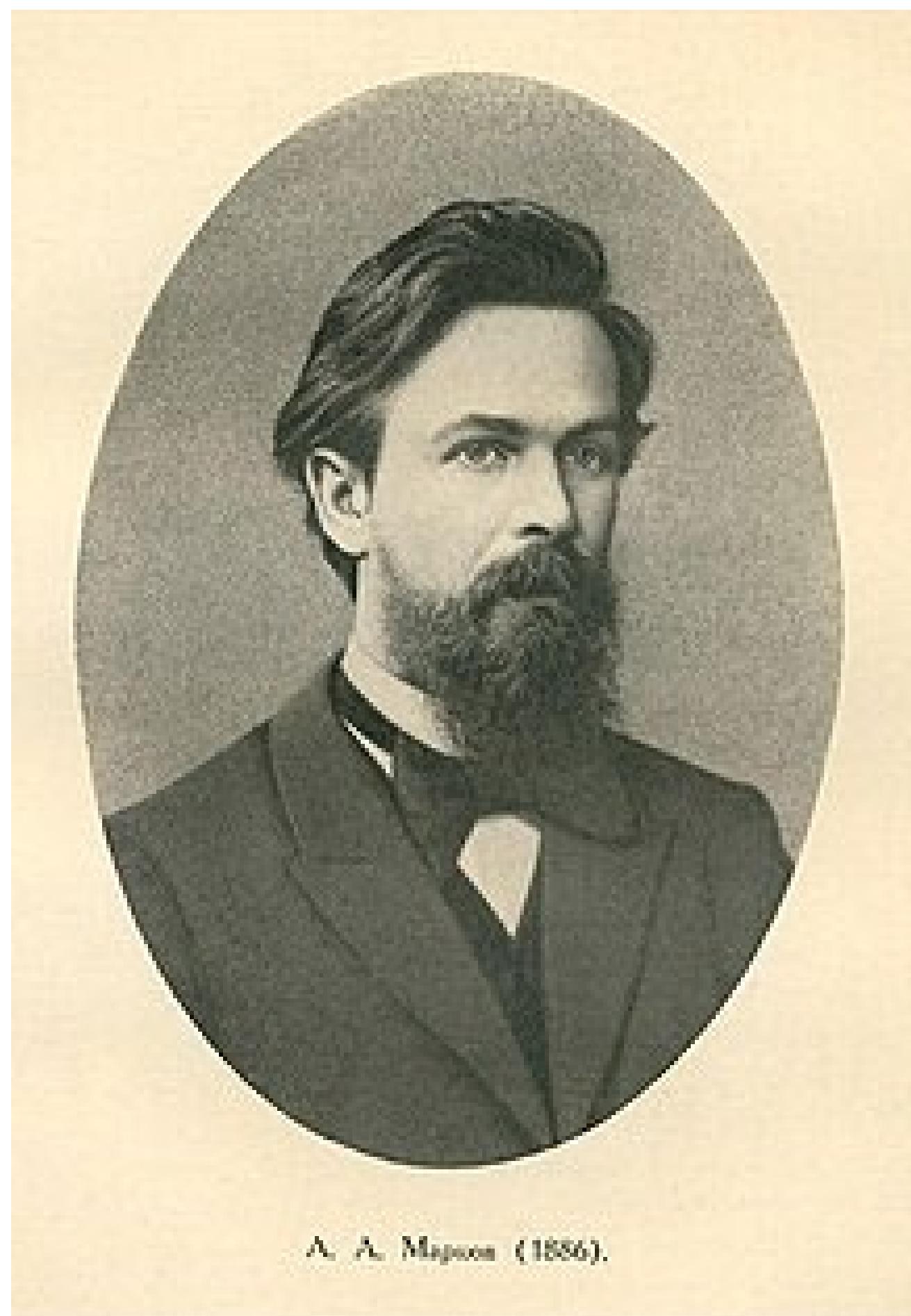
Pixels



Audio

# Ideas going back over a hundred years...

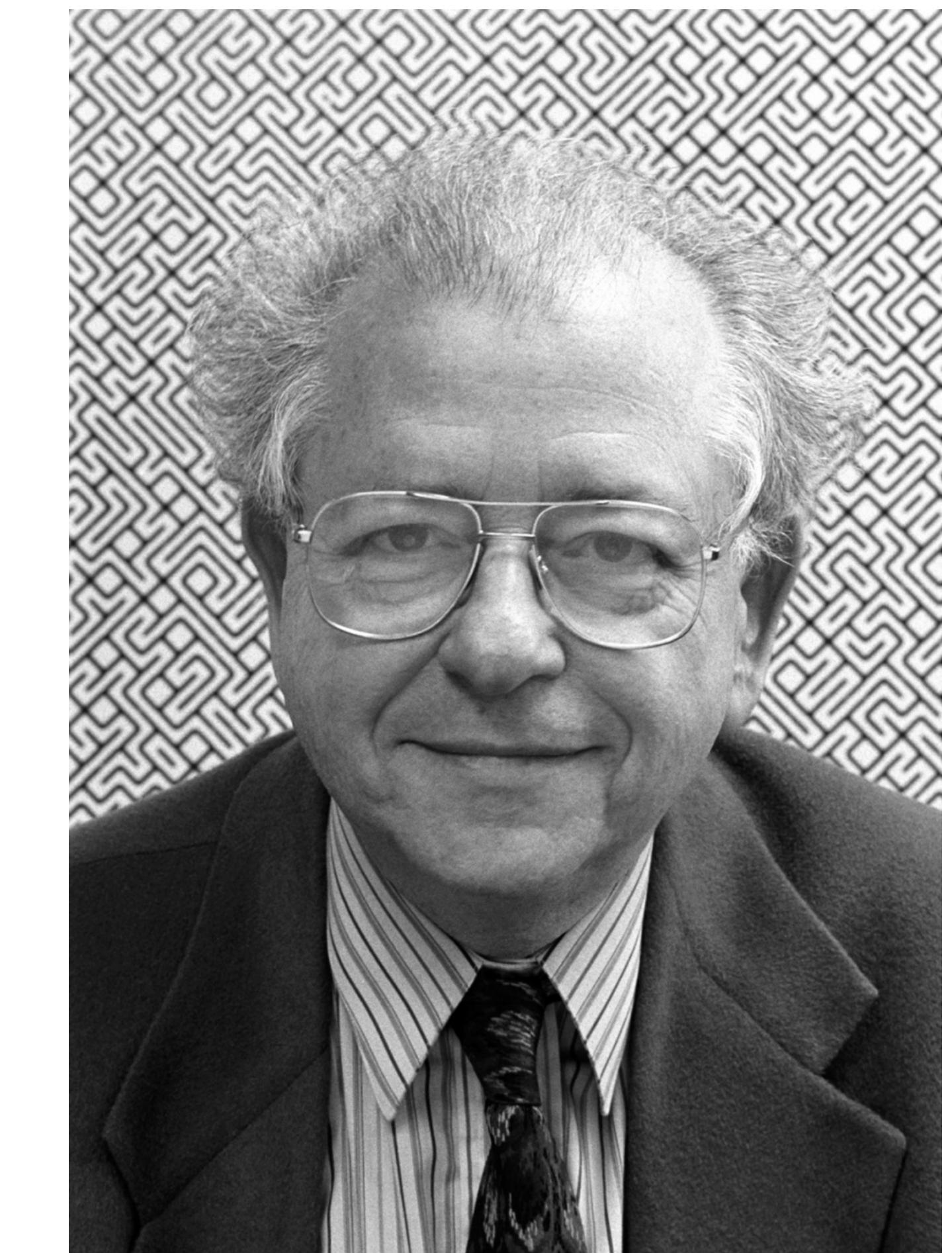
---



Andrey Markov  
(1856-1922)



Claude Shannon  
(1916-2001)



Bela Julesz  
(1928-2003)

# Weather Forecasting for Dummies™

---

Let's predict weather:

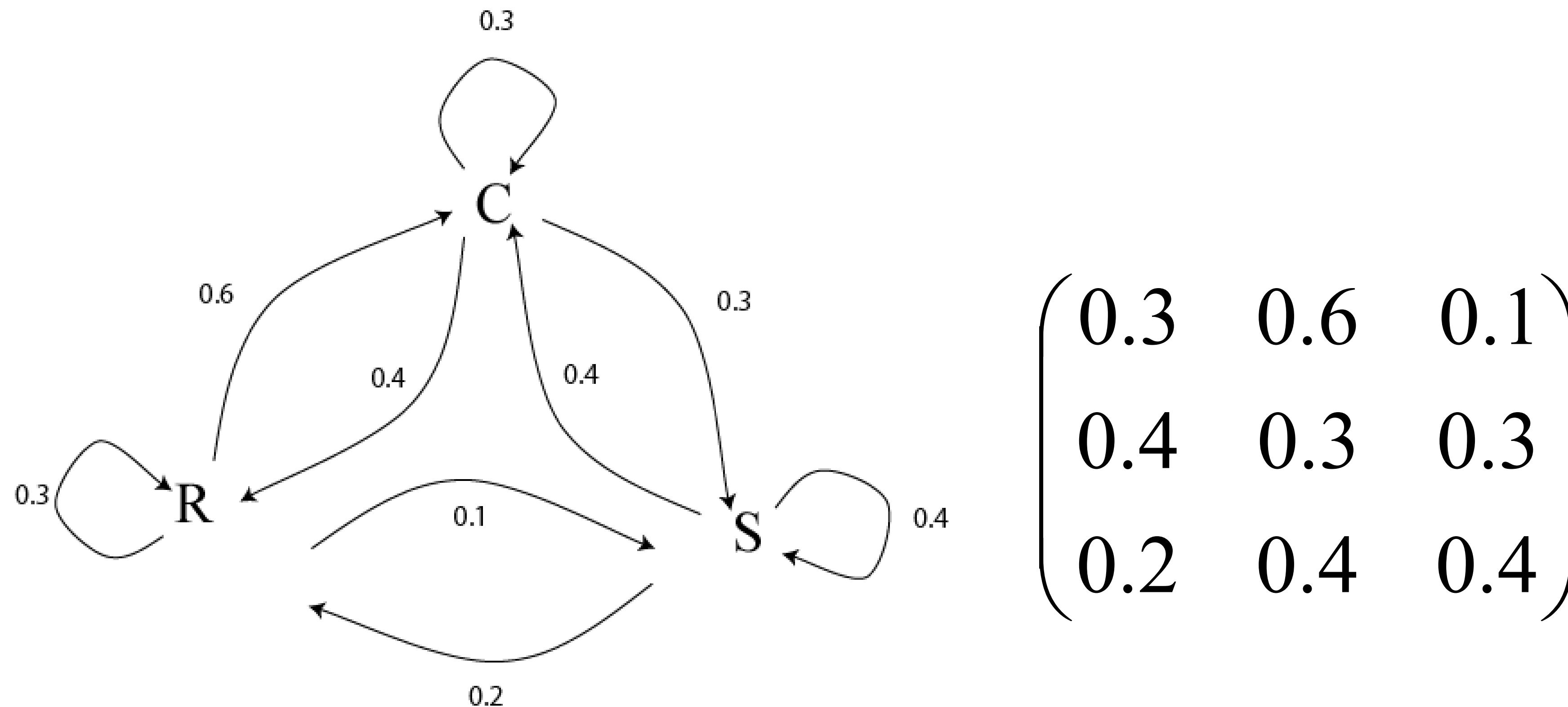
- Given today's weather only, we want to know tomorrow's
- Suppose weather can only be {Sunny, Cloudy, Raining}

The “Weather Channel” algorithm:

- Over a long period of time, record:
  - How often S followed by R
  - How often S followed by S
  - Etc.
- Compute percentages for each state:
  - $P(R|S)$ ,  $P(S|S)$ , etc.
- Predict the state with highest probability!
- It's a Markov Chain

# Markov Chain

---



What if we know today and yesterday's weather?

# Now, let's apply this to text

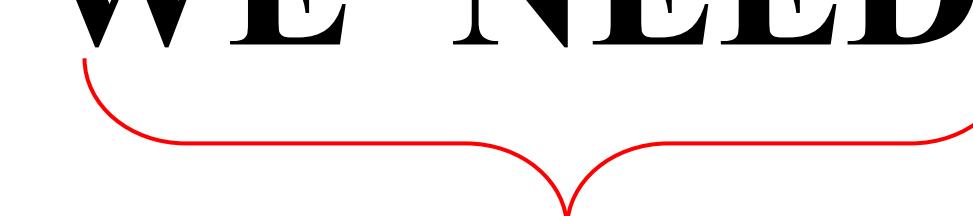
---

[Markov, 1913] statistical analysis of text

[Shannon, 1948] proposed a way to generate English-looking text using N-grams:

- Assume a generalized Markov model of language
- Use a large text to compute prob. distributions of each letter given N-1 previous letters
- Starting from a seed repeatedly sample this Markov chain to generate new letters
- Also works for whole words
- E.g.:

WE NEED TO EAT CAKE



## Results (using alt.singles corpus):

- “As I've commented before, really relating to someone involves standing next to impossible.”
- “One morning I shot an elephant in my arms and kissed him.”
- “I spent an interesting evening recently with a grain of salt”

Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still red

Bob Dylan, *Tangled up in Blue*

slide from Steve Seitz's [video](#)

Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still red

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Early one morning the sun was shining laying in bed  
her hair still red

Wondering if she had changed at all if

```
graph TD; sun[the sun] --- was[was]; was --- shining[shining]; shining --- laying[laying in bed]; hair[her hair] --- still[still red];
```

Early one morning the sun was shining

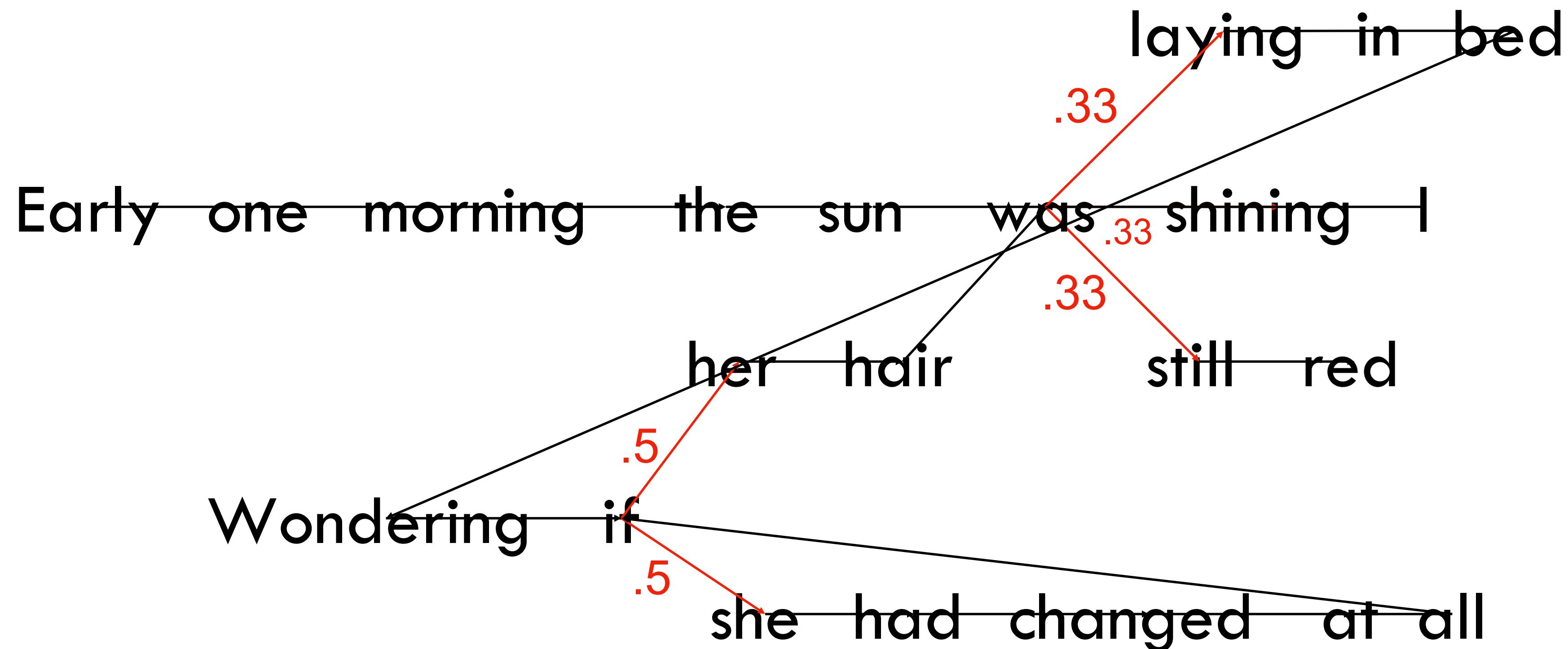
her hair still red

Wondering if she had changed at all if

laying in bed

```
graph TD; sun[sun] --- shining[shining]; her[her] --- hair[hair]; changed[changed] --- if1;if1[if]
```

Early one morning the sun was shining I  
laying in bed  
her hair still red  
Wondering if she had changed at all



## Language Model

$$P(x_n | x_{n-1})$$

Early one morning the sun was shining

Wondering if her hair was still red

she had changed at all

```
graph LR; was[was] --- shining[shining]; was --- hair[her hair]; was --- changed[changed]; was --- still[still]
```

$$P(x_n | x_{n-1}, x_{n-2})$$

Early one → morning  
one morning → the  
morning the → sun  
the sun → was  
sun was → shining  
was shining → I  
shining I → was  
I was → laying  
...

# Video Textures

Arno Schödl

Richard Szeliski

David Salesin

Irfan Essa

Microsoft Research, Georgia Tech

SIGGRAPH 2000

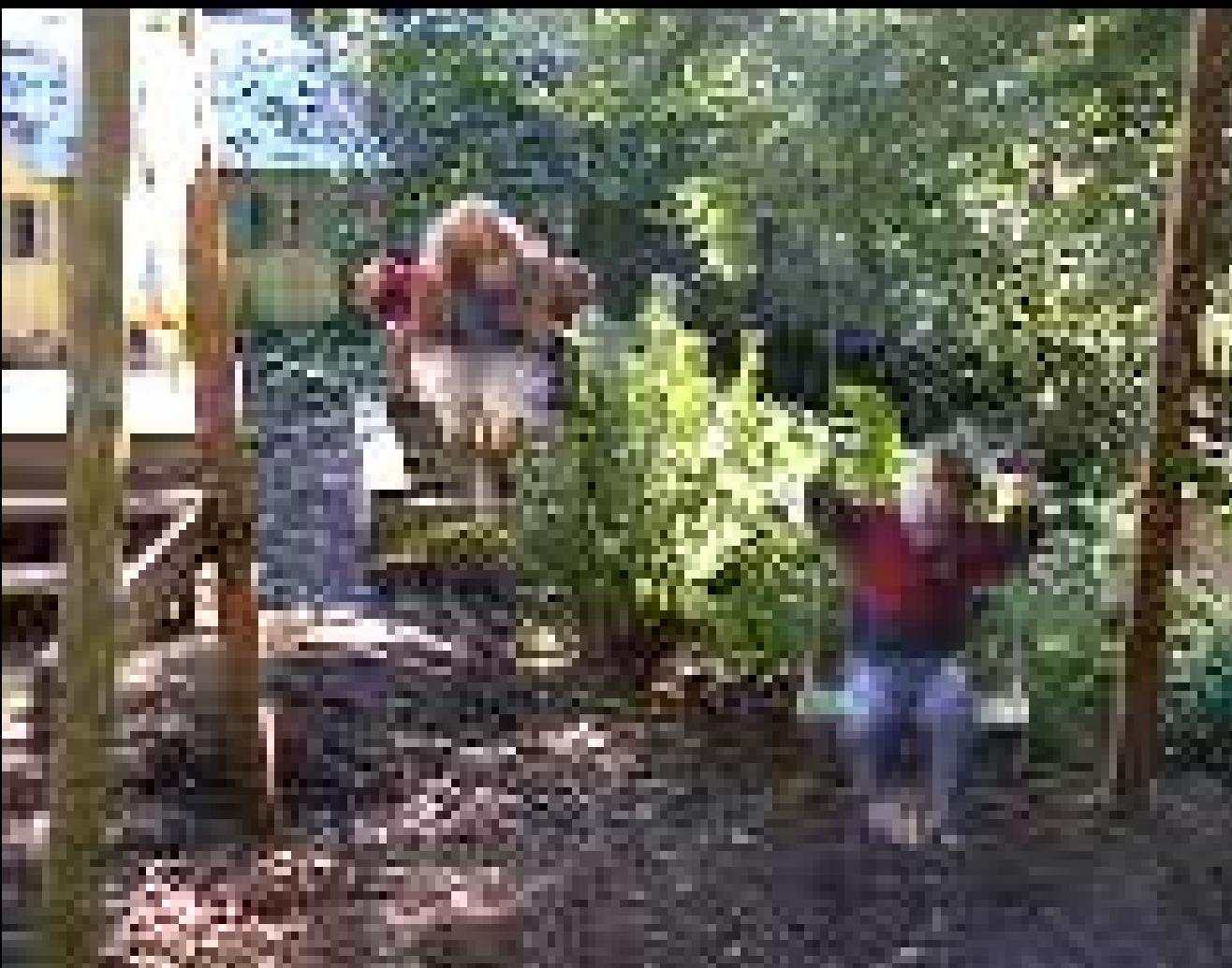
# Still photos



# Video clips



# Video textures



# Problem statement



video clip

video texture

# Our approach

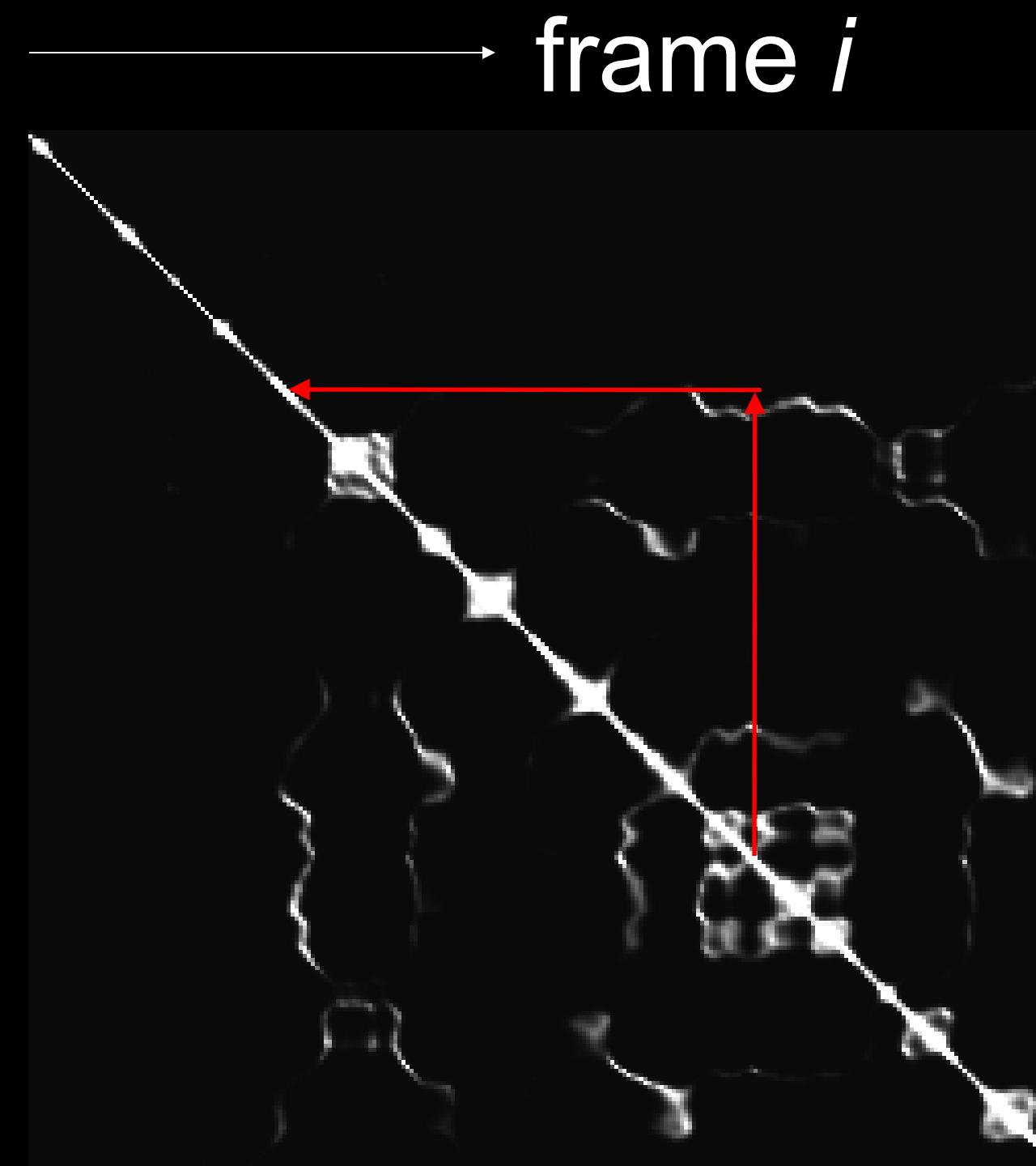


- How do we find good transitions?

# Finding good transitions

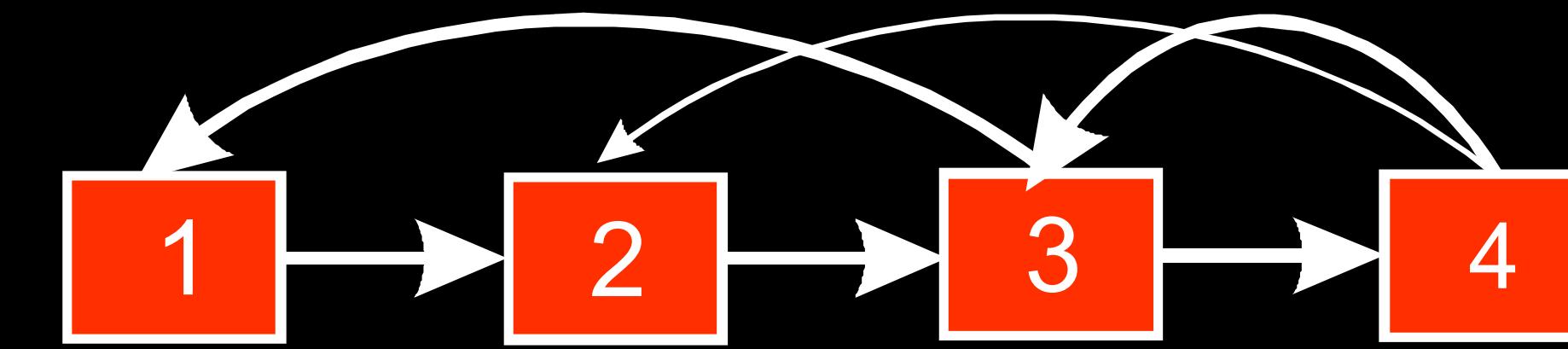
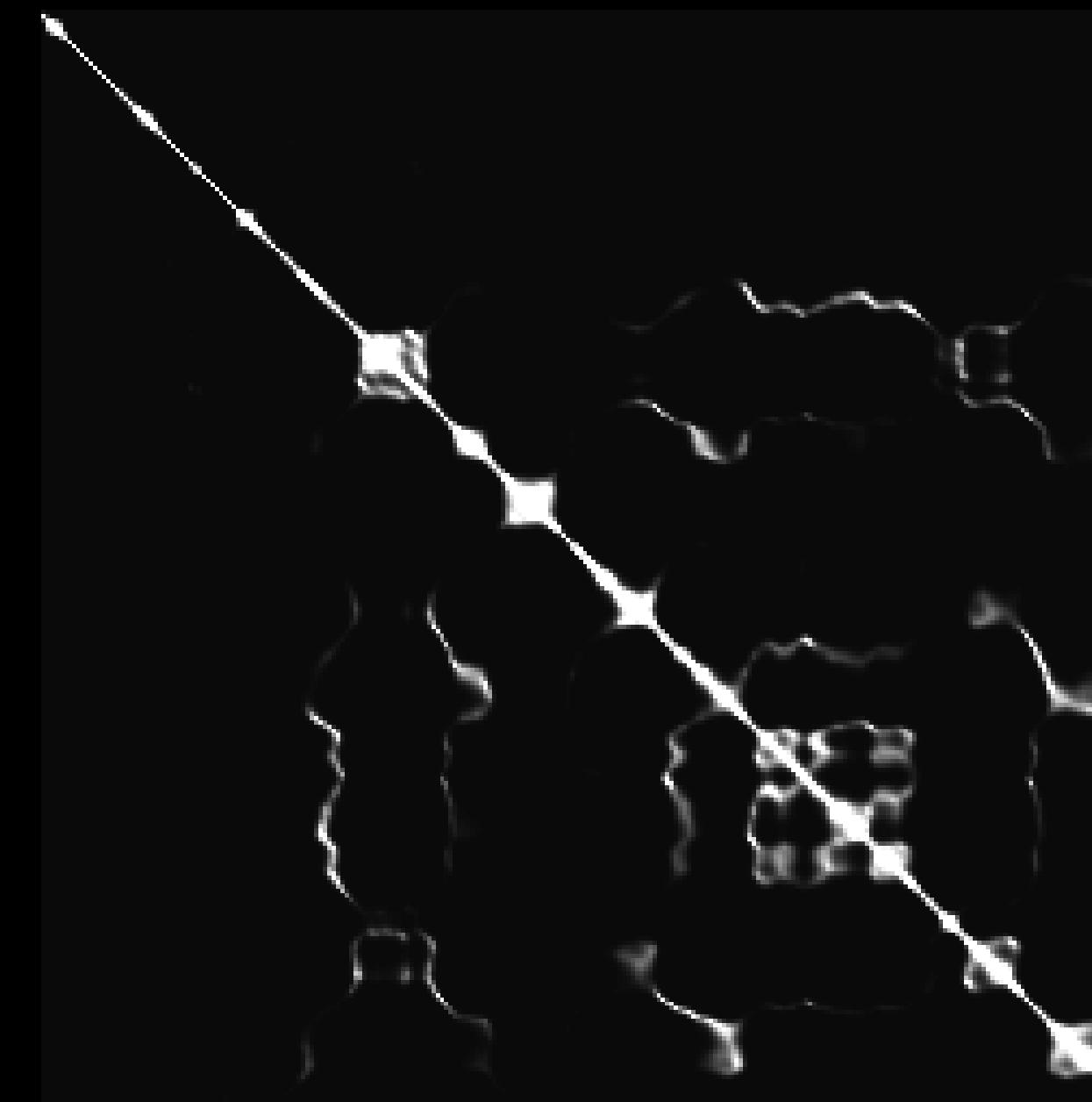
- Compute  $L_2$  distance  $D_{i,j}$  between all frames

vs.



Similar frames make good transitions

# Markov chain representation

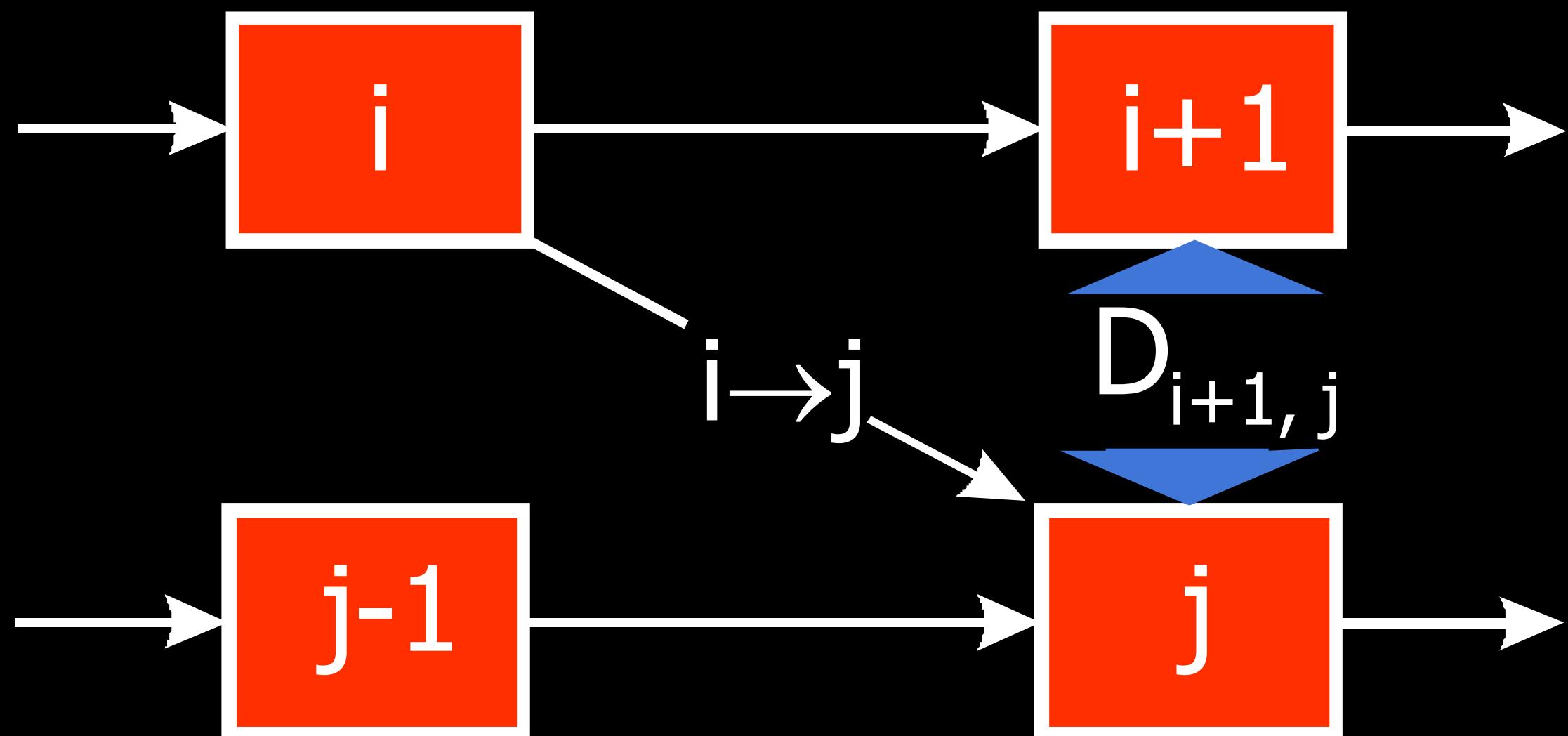


Similar frames make good transitions

# Transition costs

- Transition from  $i$  to  $j$  if successor of  $i$  is similar to  $j$ 
  - Cost function:  $C_{i \rightarrow j} = D_{i+1, j}$

.



# Transition probabilities

- Probability for transition  $P_{i \rightarrow j}$  inversely related to cost:
  - $P_{i \rightarrow j} \sim \exp(-C_{i \rightarrow j} / \sigma^2)$



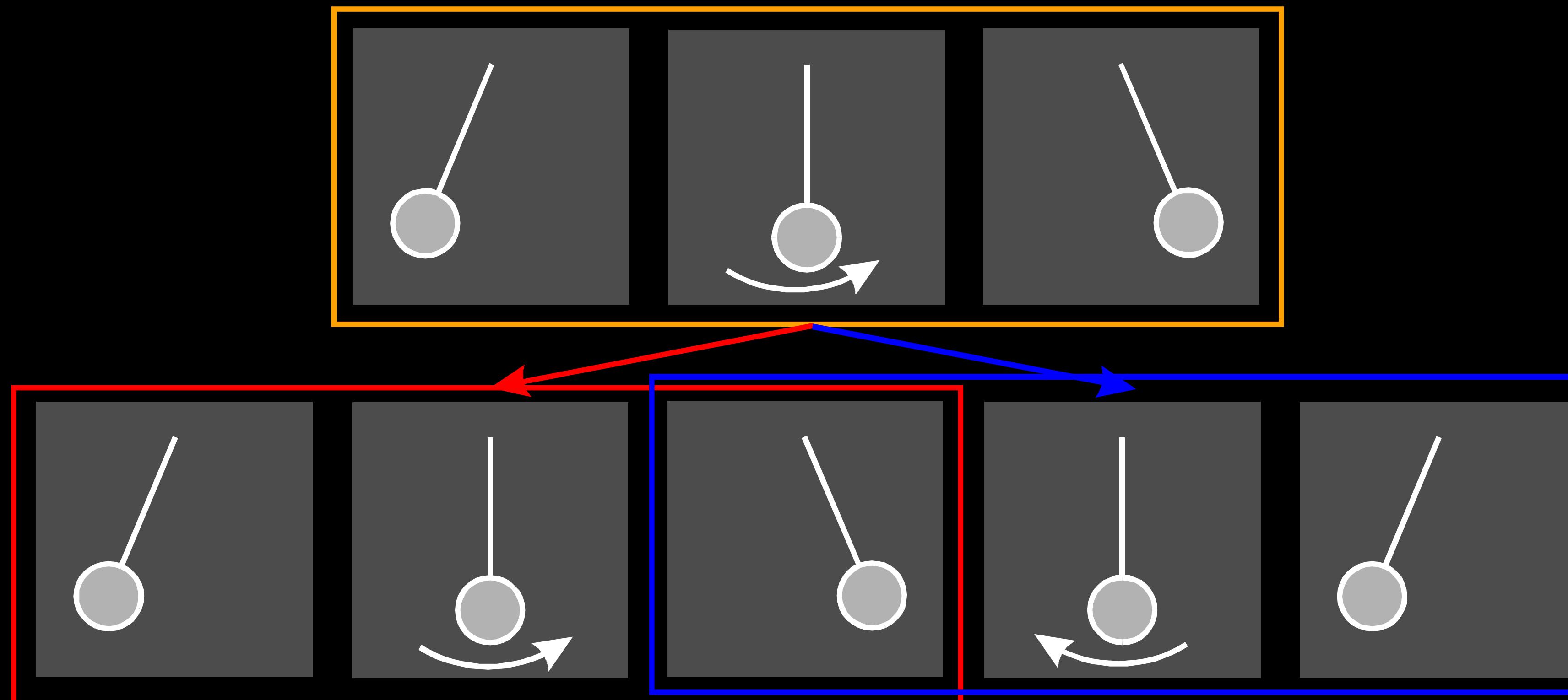
high  $\sigma$

low  $\sigma$

# Preserving dynamics



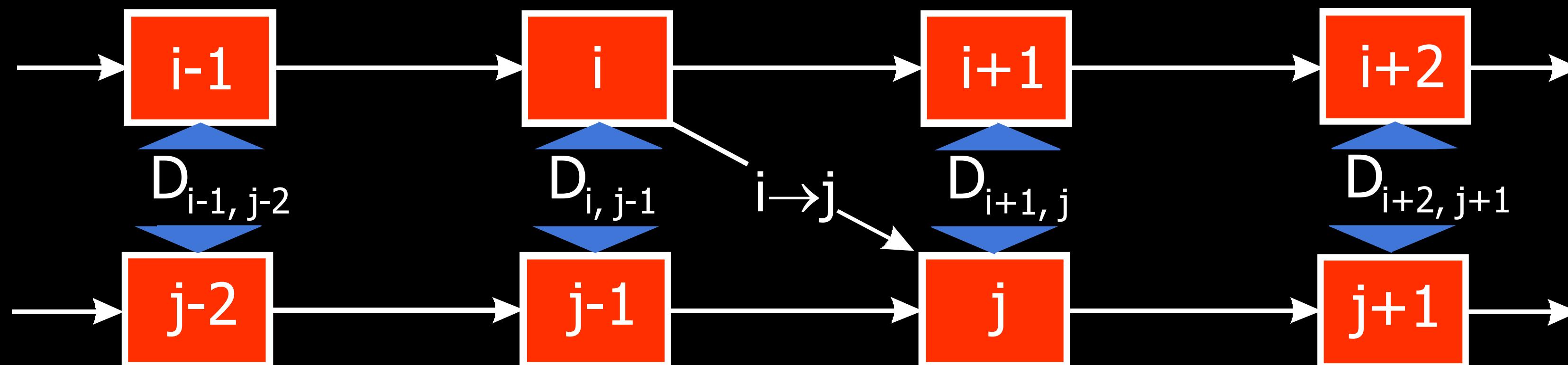
# Preserving dynamics



# Preserving dynamics

- Cost for transition  $i \rightarrow j$

- $$C_{i \rightarrow j} = \sum_{k=-N}^{N-1} w_k D_{i+k+1, j+k}$$



# Preserving dynamics – effect

- Cost for transition  $i \rightarrow j$ 
  - $C_{i \rightarrow j} = \sum_{k = -N}^{N-1} w_k D_{i+k+1, j+k}$



# User-controlled video textures



slow

variable

fast

User selects target frame range

# Video-based animation

- Like sprites  
computer games
- Extract sprites  
from real video
- Interactively control  
desired motion



©1985 Nintendo of America Inc.



# Video sprite extraction



blue screen matting  
and velocity estimation



# Video sprite control

- Augmented transition cost:

$$C_{i \rightarrow j}^{\text{Animation}} = \alpha C_{i \rightarrow j} + \beta \text{ angle}$$

Similarity term      Control term

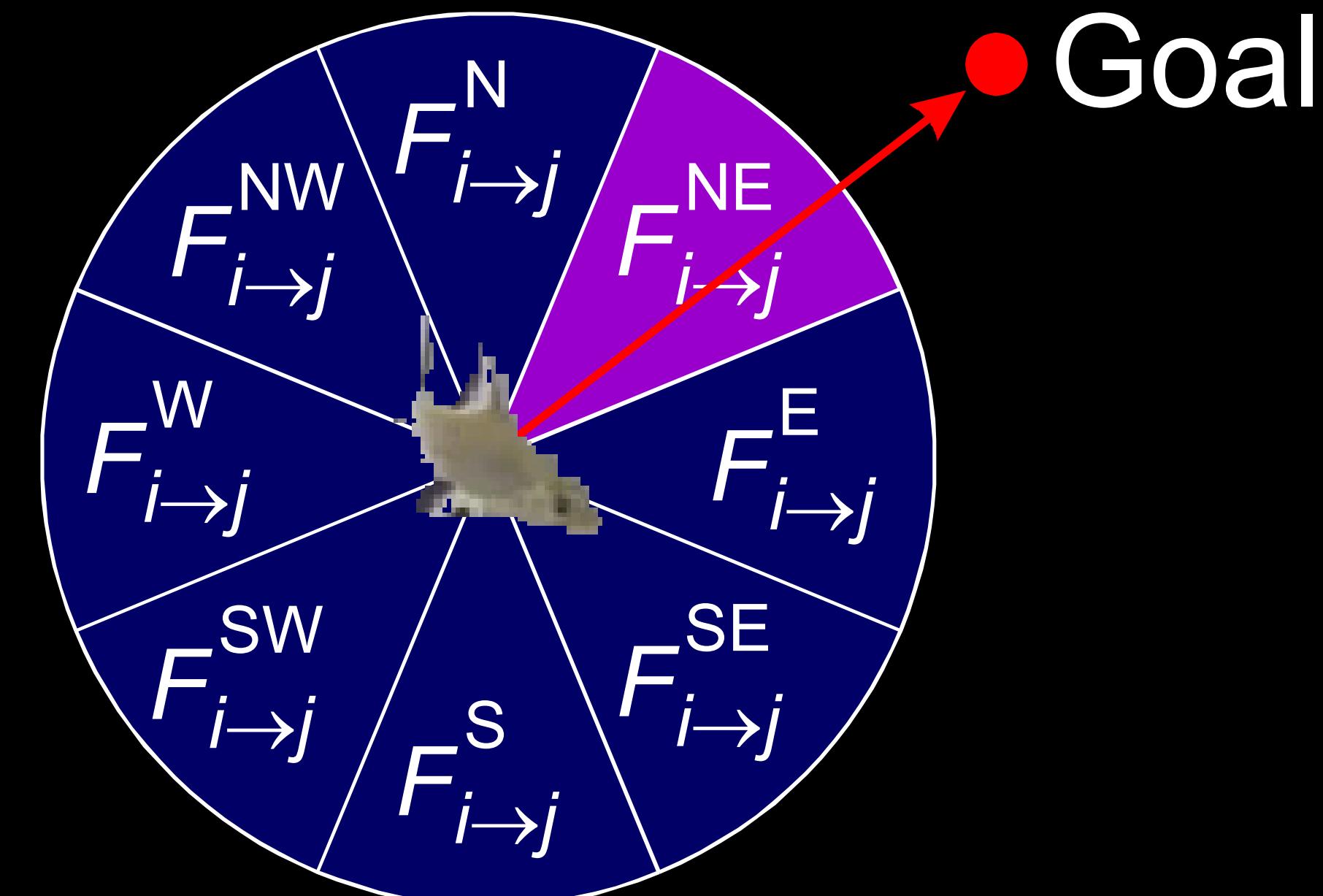
vector to  
mouse pointer

velocity vector

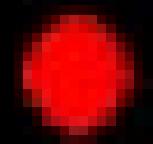
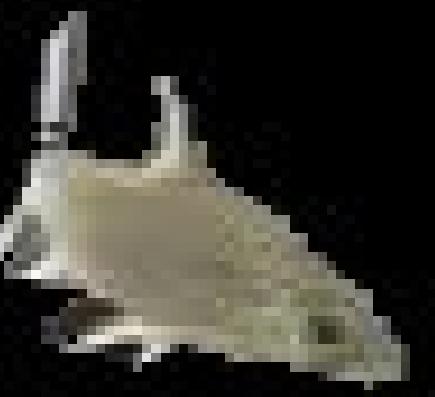
angle

# Video sprite control

- Need future cost computation
- Precompute future costs for a few angles.
- Switch between precomputed angles according to user input
- [GIT-GVU-00-11]



# Interactive fish



# Summary / Discussion

- Some things are relatively easy



# Discussion

- Some are hard



# Texture

- Texture depicts spatially repeating patterns
- Many natural phenomena are textures



radishes



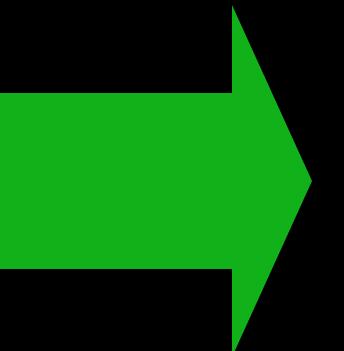
rocks



yogurt

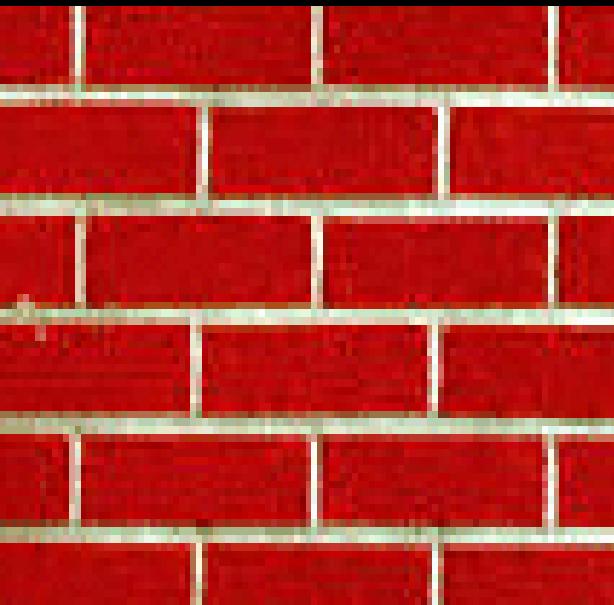
# Texture Synthesis

- Goal of Texture Synthesis: create new samples of a given texture
- Many applications: virtual environments, hole-filling, texturing surfaces

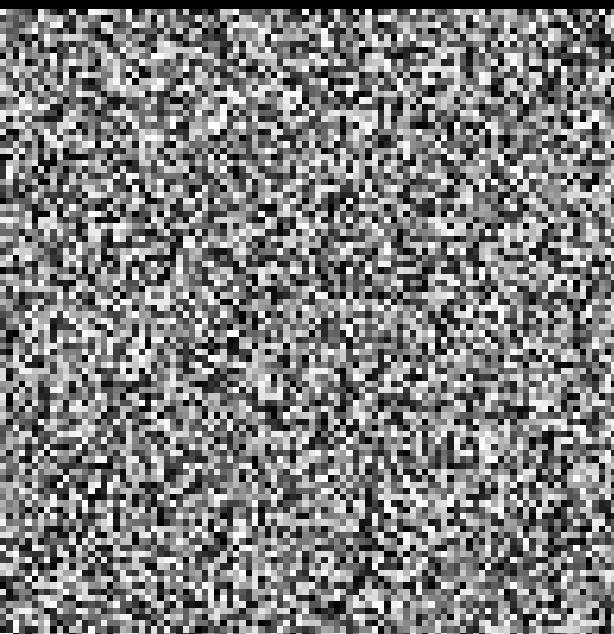


# The Challenge

- Need to model the whole spectrum: from repeated to stochastic texture



repeated

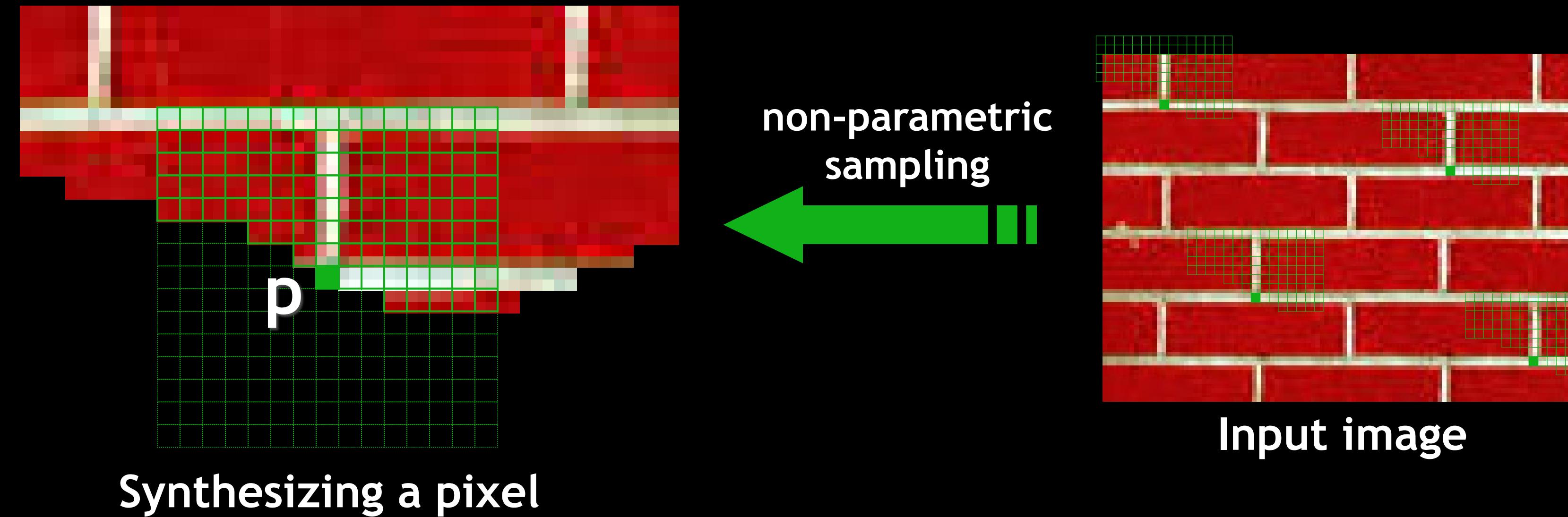


stochastic



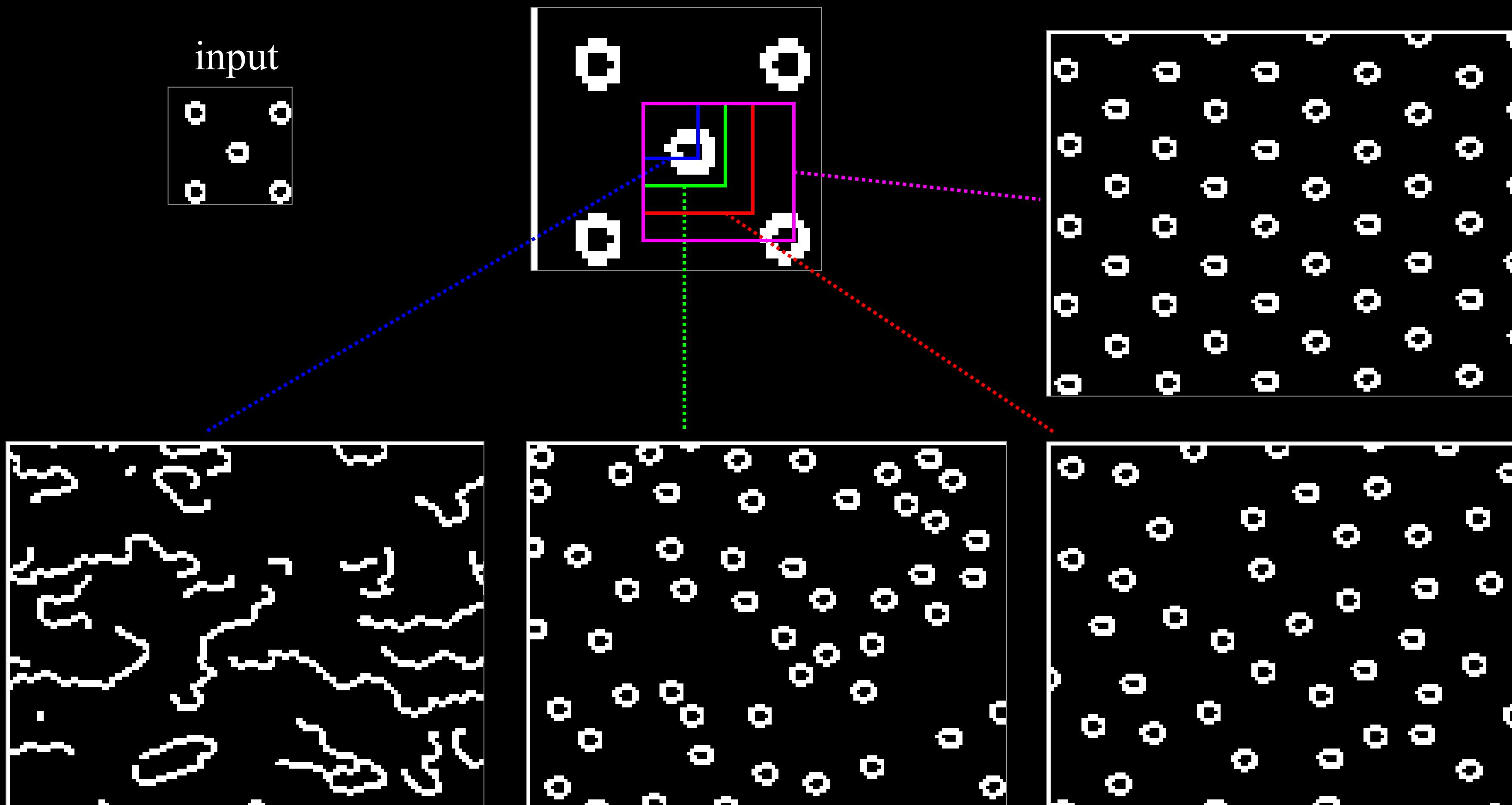
Both?

# Efros & Leung Algorithm (ICCV 1999)

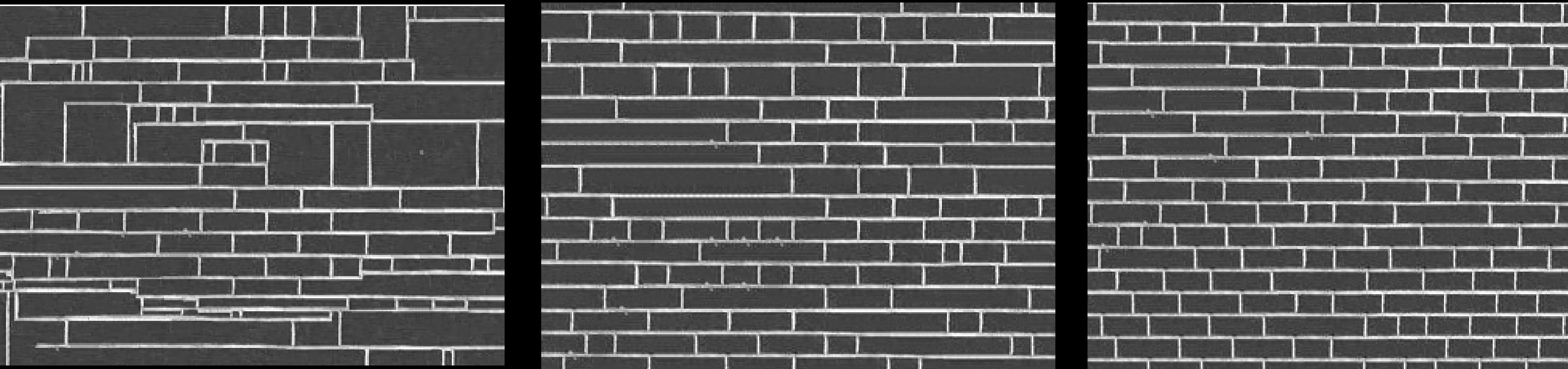
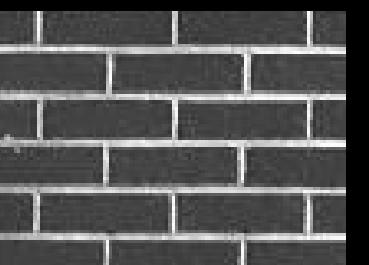
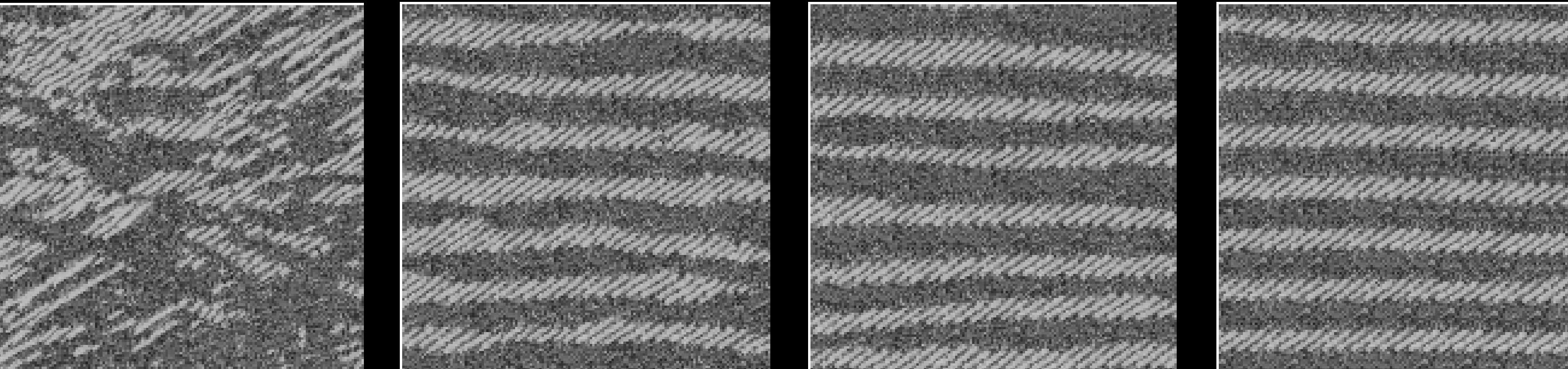
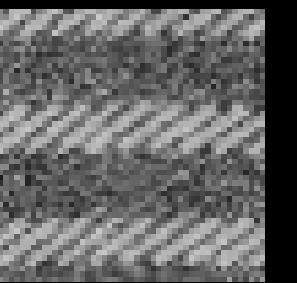


- Assuming Markov property, compute  $P(p|N(p))$ 
  - Building explicit probability tables infeasible
  - Instead, we *search the input image* for all similar neighborhoods — that's our pdf for  $p$
  - To sample from this pdf, just pick one match at random

# Neighborhood Window



# Varying Window Size

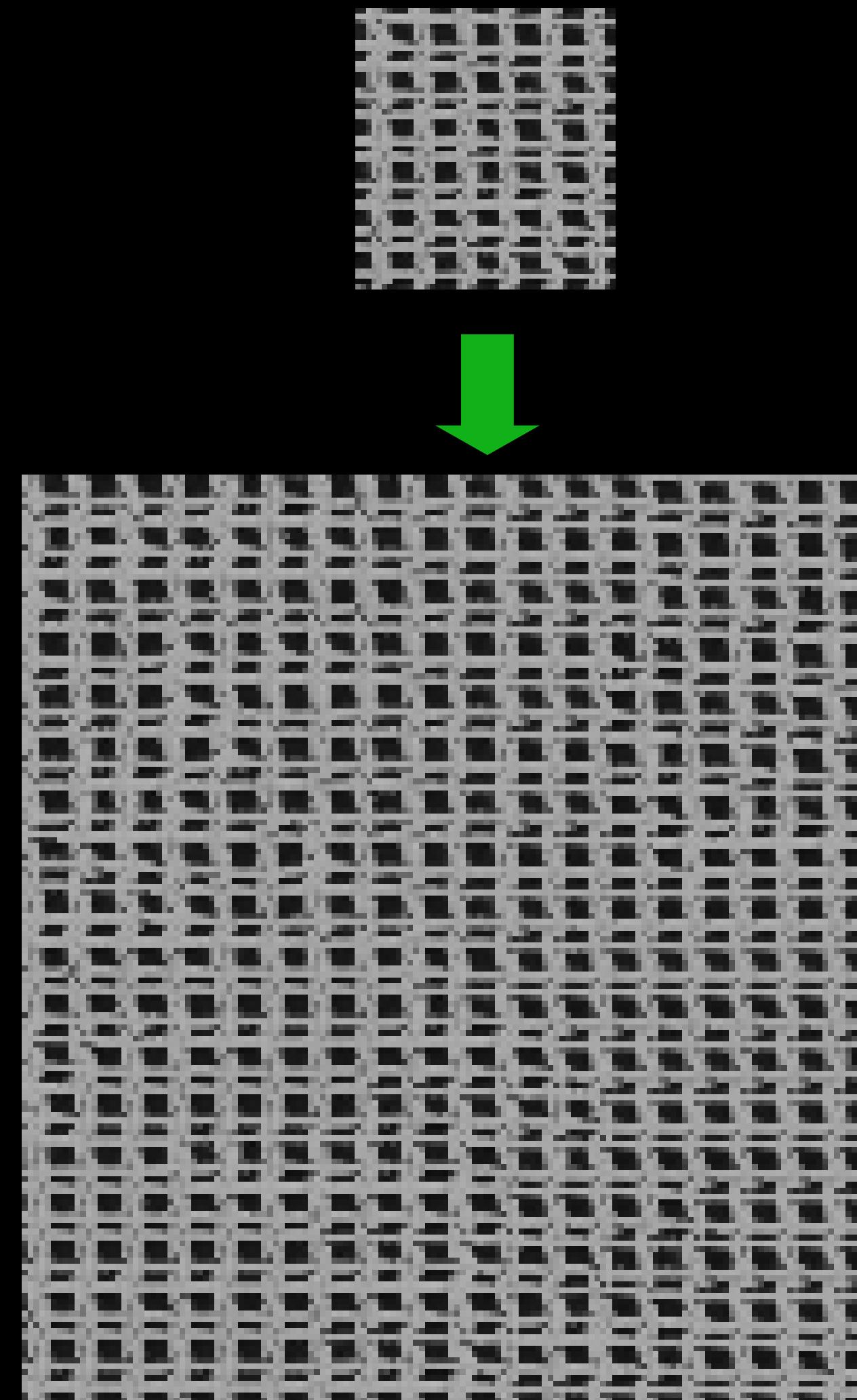


Increasing window size

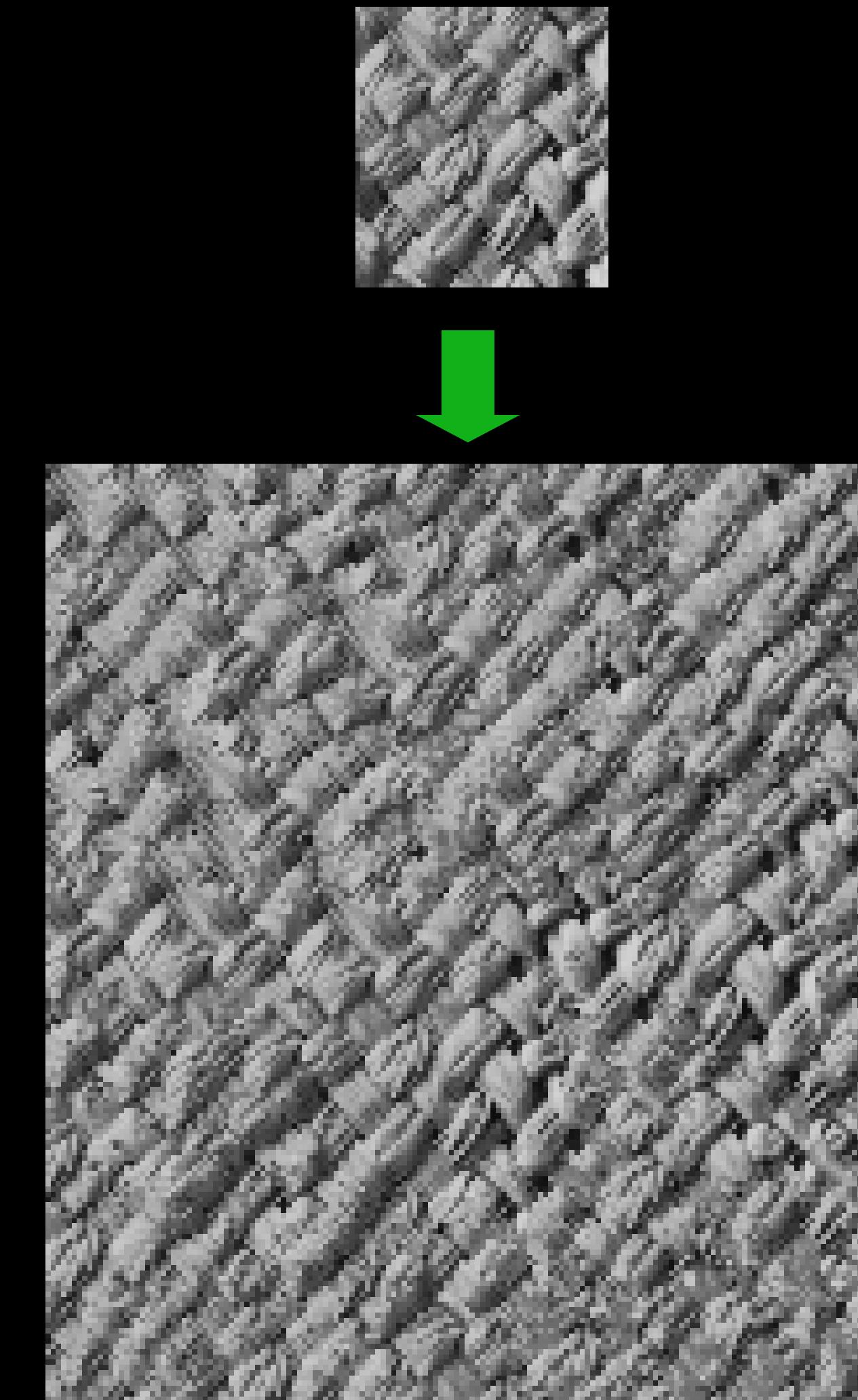


# Synthesis Results

french canvas

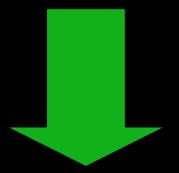


rafia weave

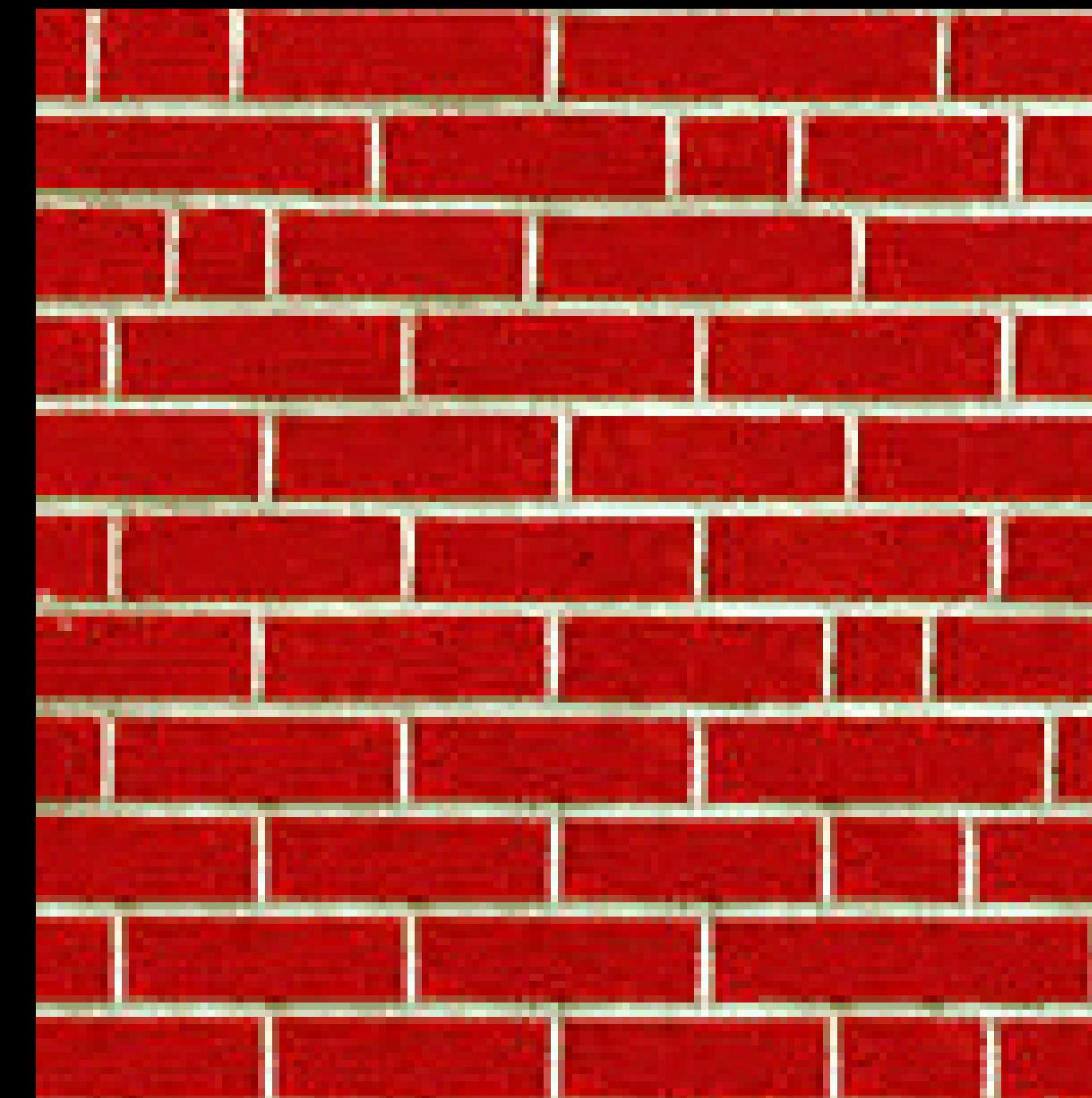
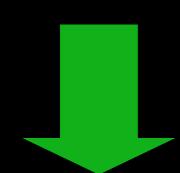
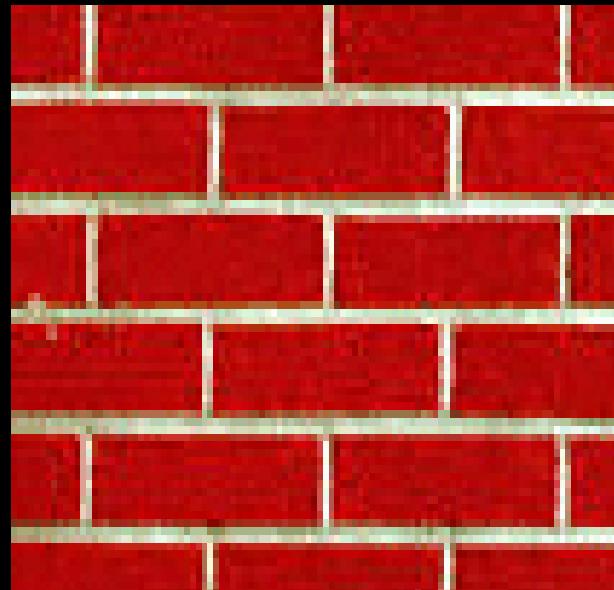


# More Results

white bread



brick wall

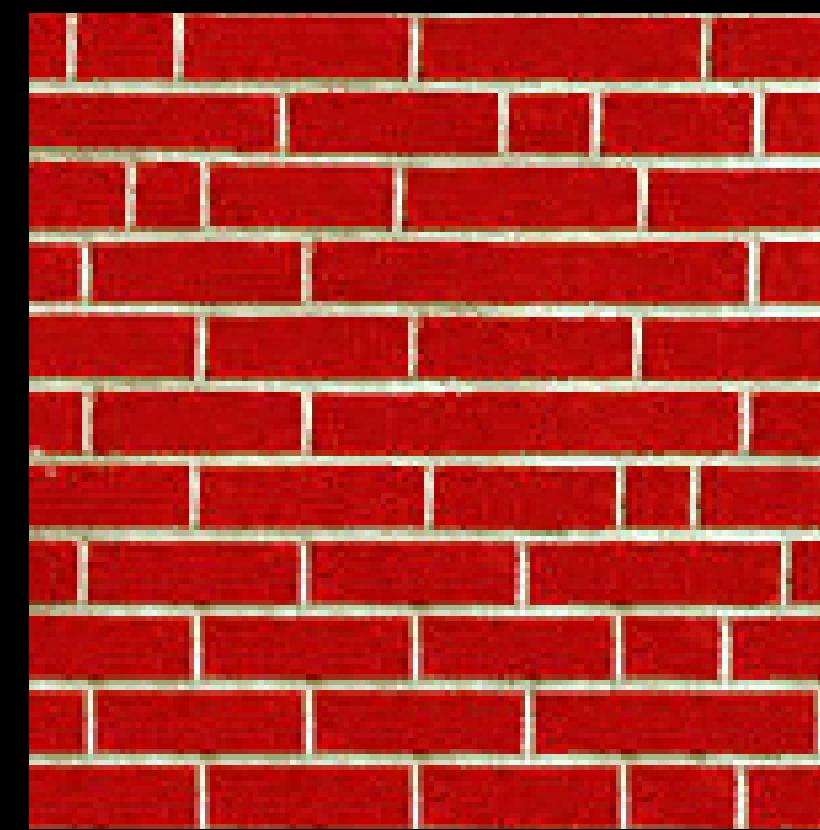
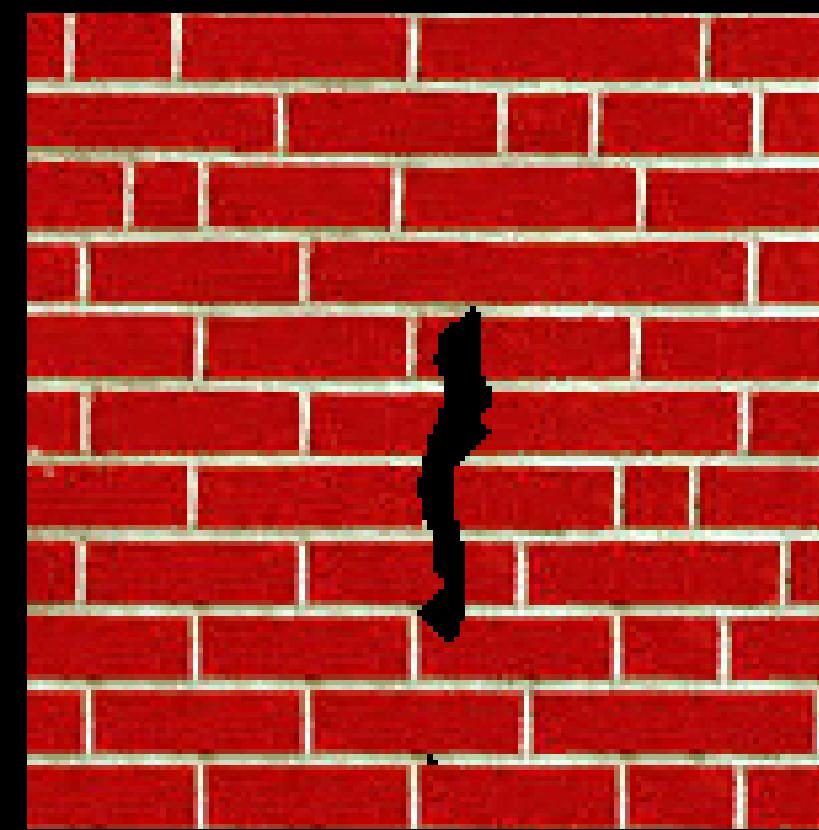
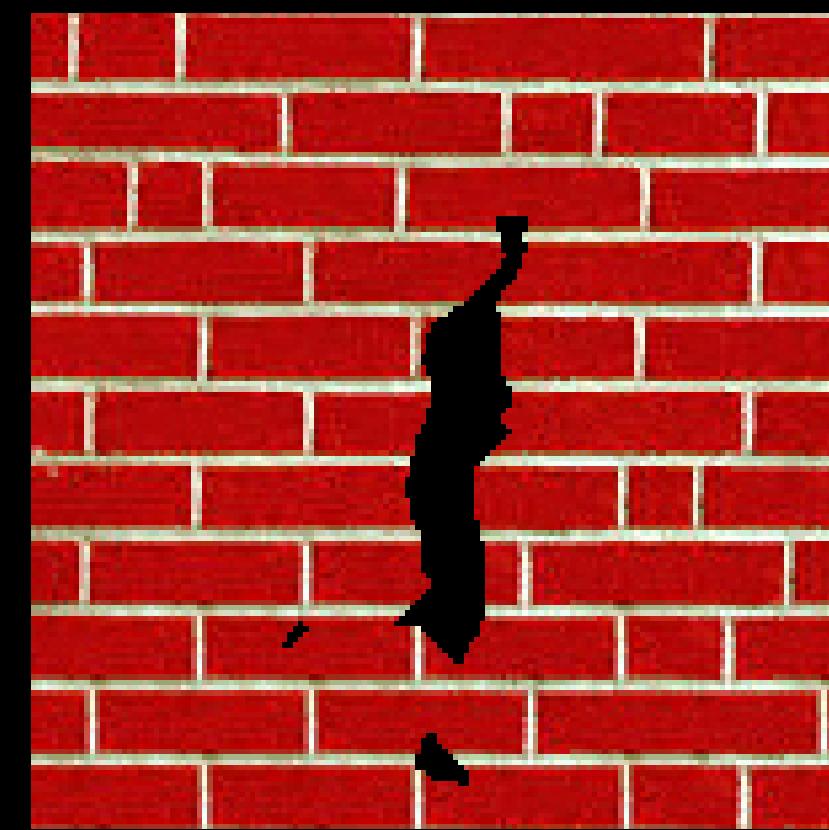


# Homage to Shannon

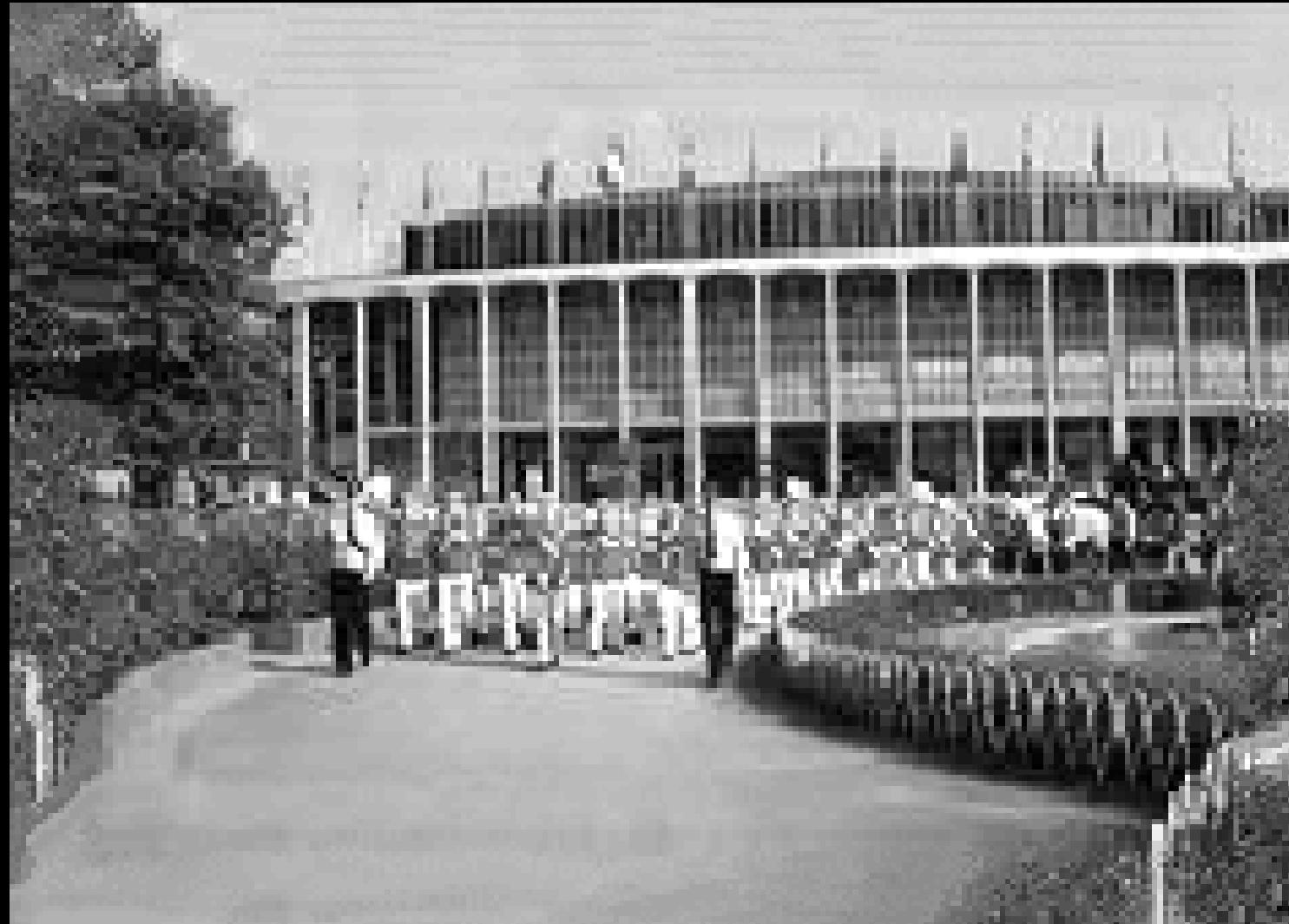
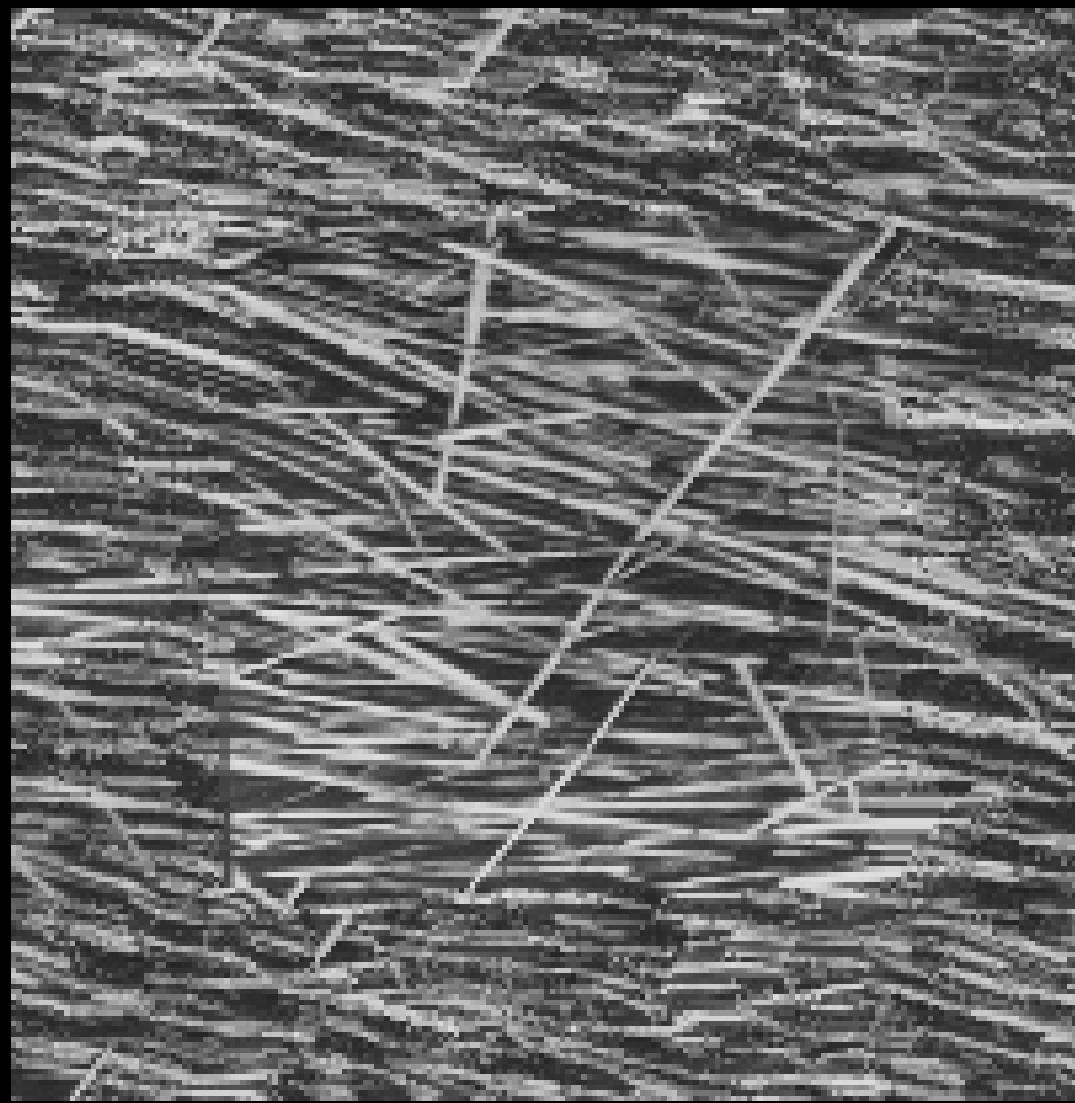
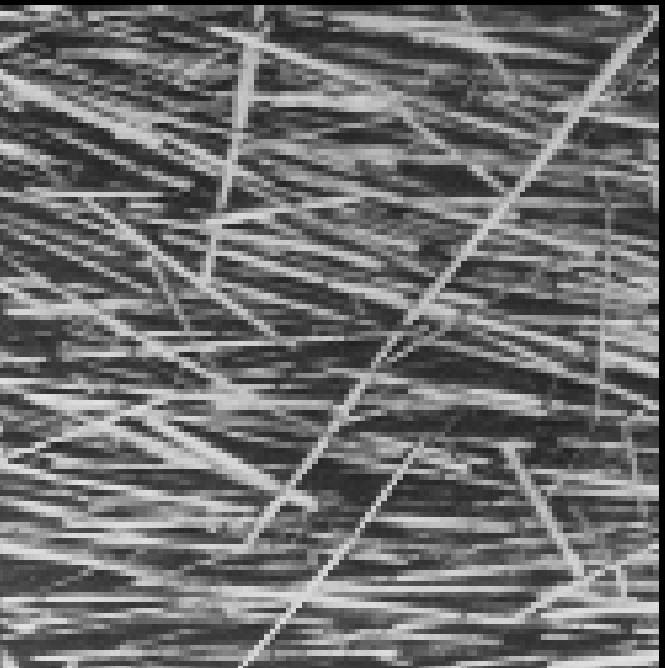
uring his anniversary speech, Senator Dick Gephardt was fearful riff on the looming crisis. He only asked, "What's your position?" A heartfelt sigh story about the emerging charges against Clinton. "Bothering people about continuing," Gephardt began, patiently observing, that the legal system was getting with this latest tangled

ithaim. them . "Whephartfe lartifelintomimen  
fel ck Clrtioout omain thartfelins. f ait s anento  
the ry onst wartfe lck Gephntoomimeationl sigab  
Chiooufit Clinut Cll riff on. hat's yordn, parut tly  
ons ycontonsteht wasked, paim i sahe loo' riff on  
nskoneploourtfeas leil A nst Clit, "Wleontongal s  
k Cirtioouirtfepe ong pme abegal fartfenstemem  
tiensteneltorydt telemepminsverdt was agemen  
ff ons artientont Cling peme asurtfe atih, "Boui s  
hal s fartfelt sig pedrl'dt ske abounutie aboutioo  
stfaonewwas yous aboonhardt thatins fain, ped,  
ains. them, pabout wasy arfiut cōuitly d, ln A h  
ble emthringbooreme agas fa bontinsyst Clinut  
ory about continst Clipequinist Cloke agatiff out  
stome zinemen fly ardt beoraboul n, thenly as t G  
cons fairmerne Diontonit wat coutlyohgans as fan  
ien, phrtfaul, "Wbaut cout congal cōmininga  
mifmst Cliiy abon 'al coountha.emungairt tf oun  
Whe looorystan loontieph. intly on, theoplegatick  
aul tatiesonly atie Diontiomt wal s f tbegae ener  
mthahcat's enenhmas fan, "intchthow abons w

# Hole Filling



# Extrapolation



# Image Analogies

Aaron Hertzmann<sup>1,2</sup>

Chuck Jacobs<sup>2</sup>

Nuria Oliver<sup>2</sup>

Brian Curless<sup>3</sup>

David Salesin<sup>2,3</sup>

<sup>1</sup>New York University

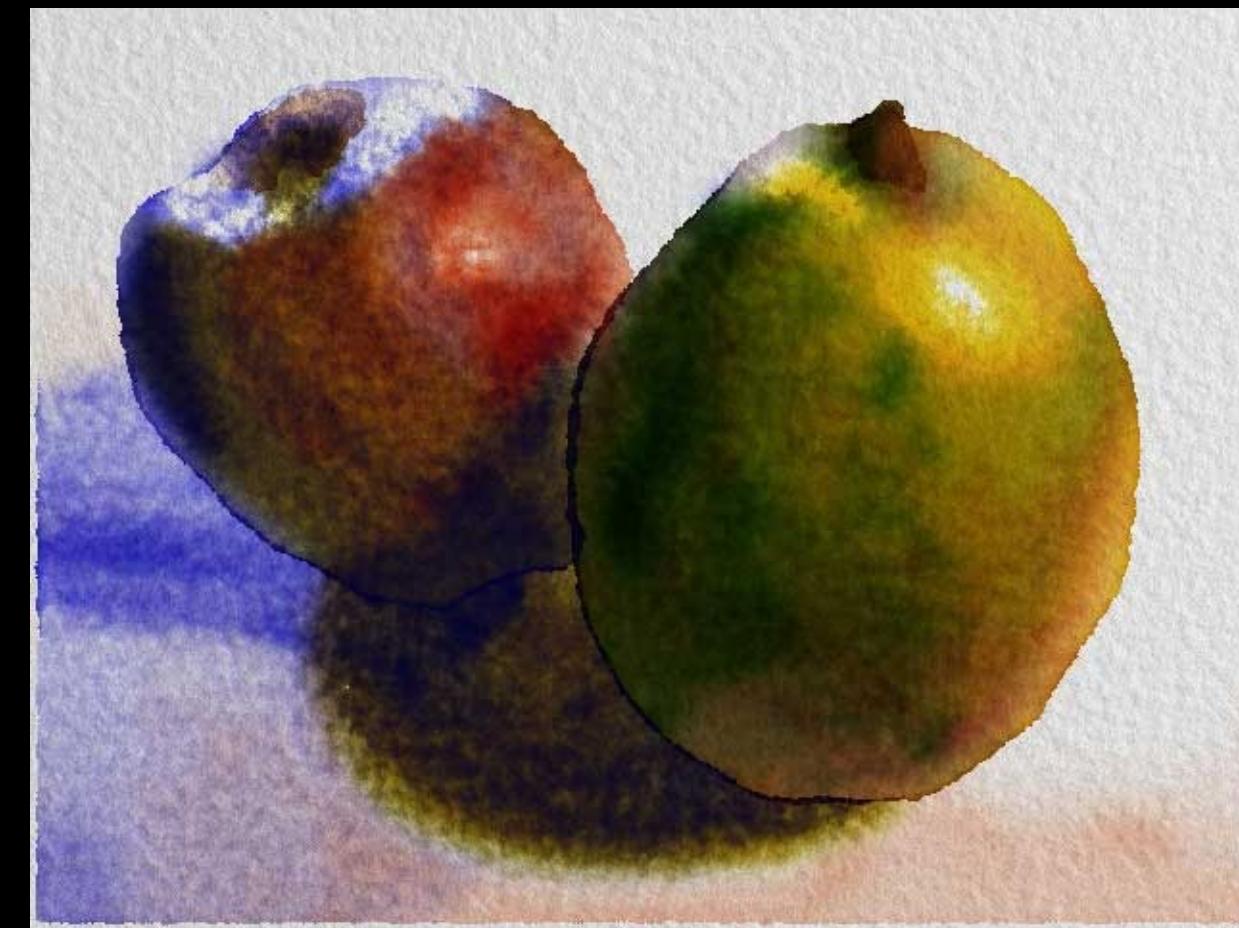
<sup>2</sup>Microsoft Research

<sup>3</sup>University of Washington

# Image Analogies



A



A'



B

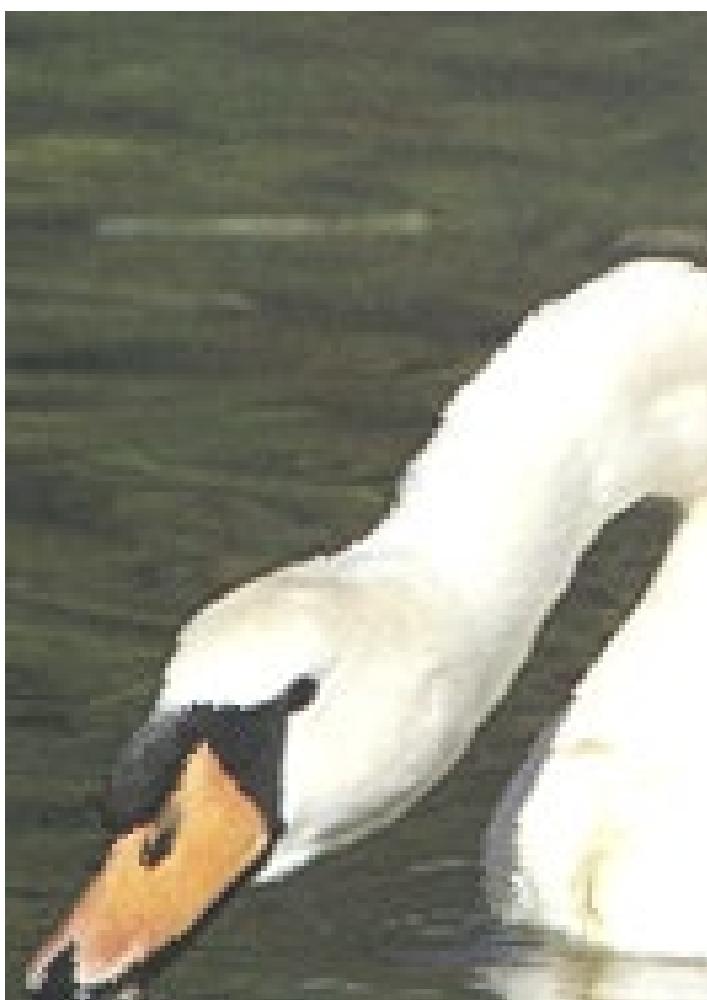


B'



# Image Analogies

Goal: Process an image by example



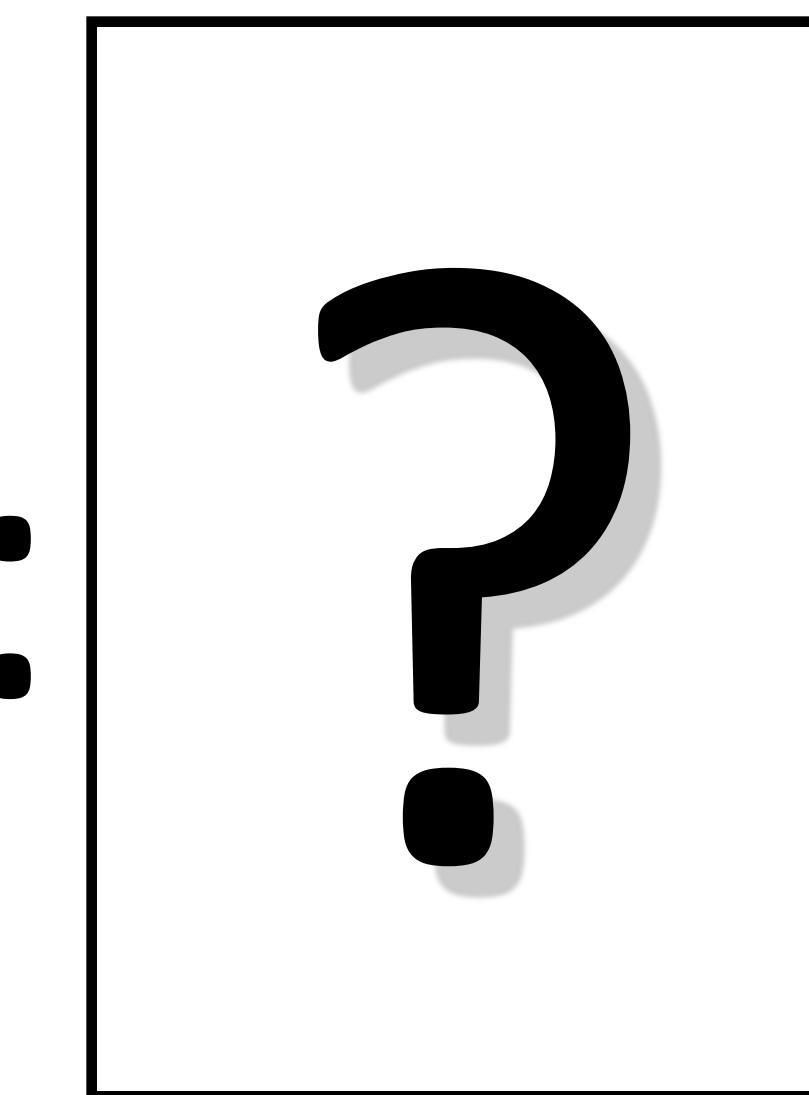
A



A'



B

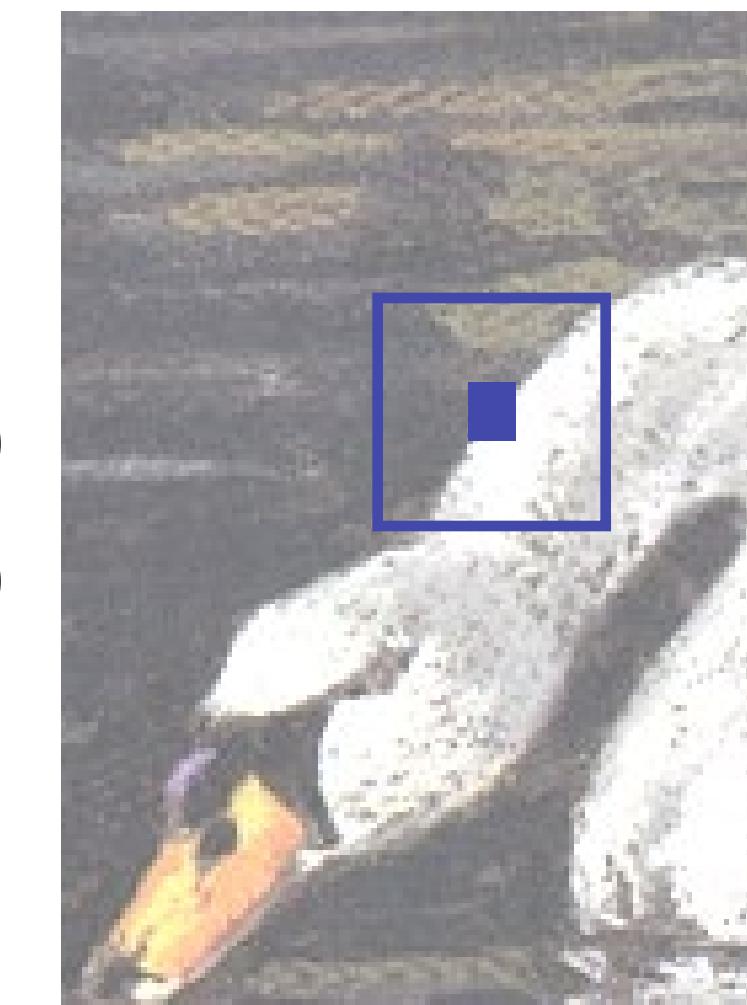


B'

# Non-parametric sampling



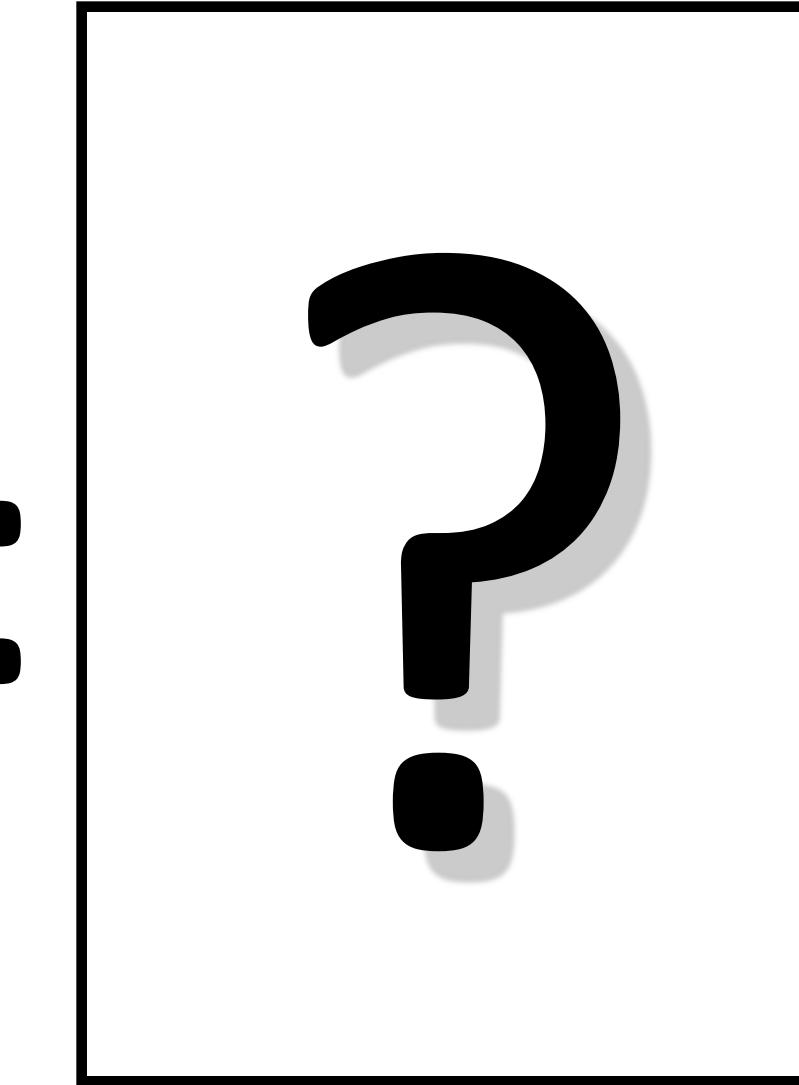
A



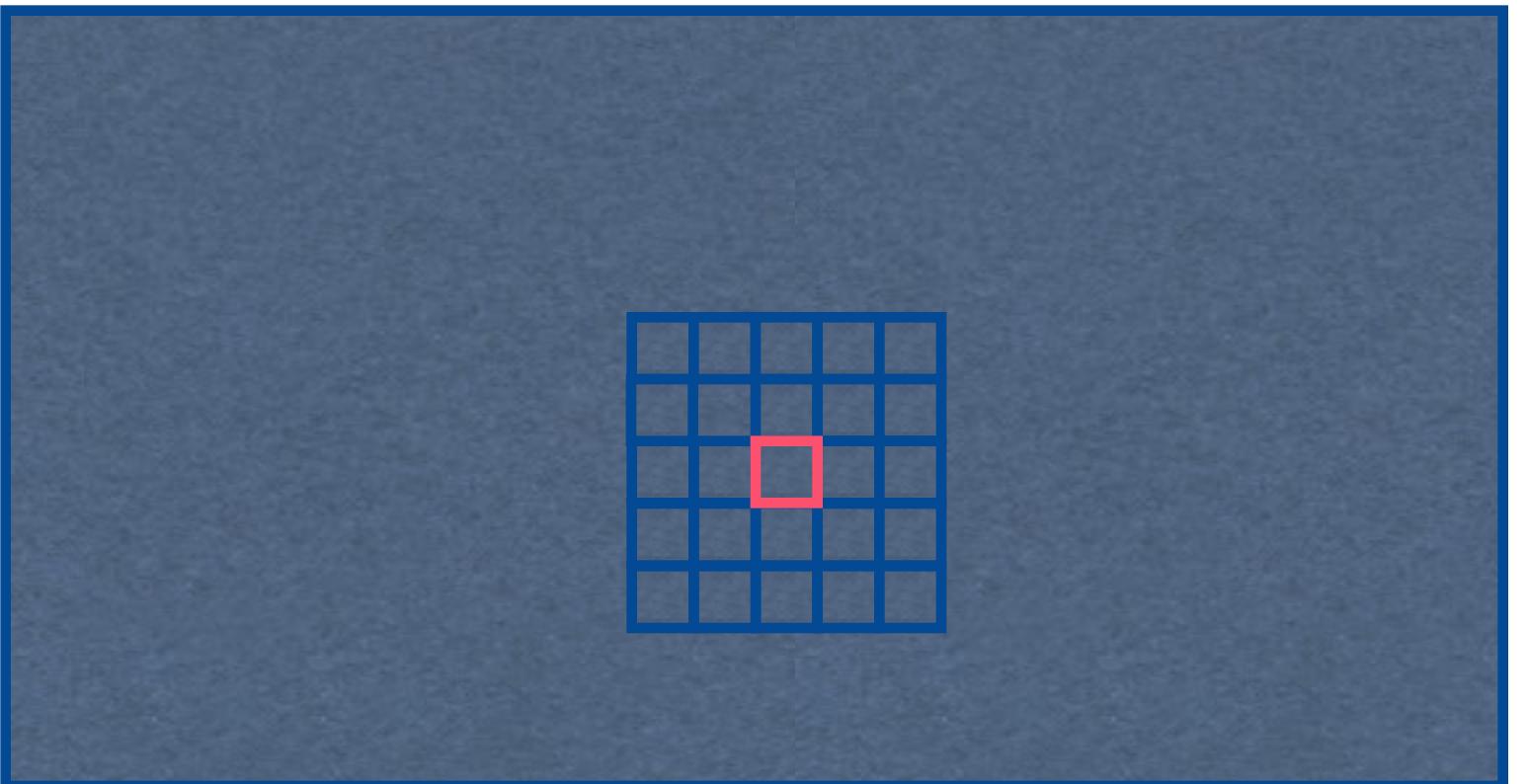
A'



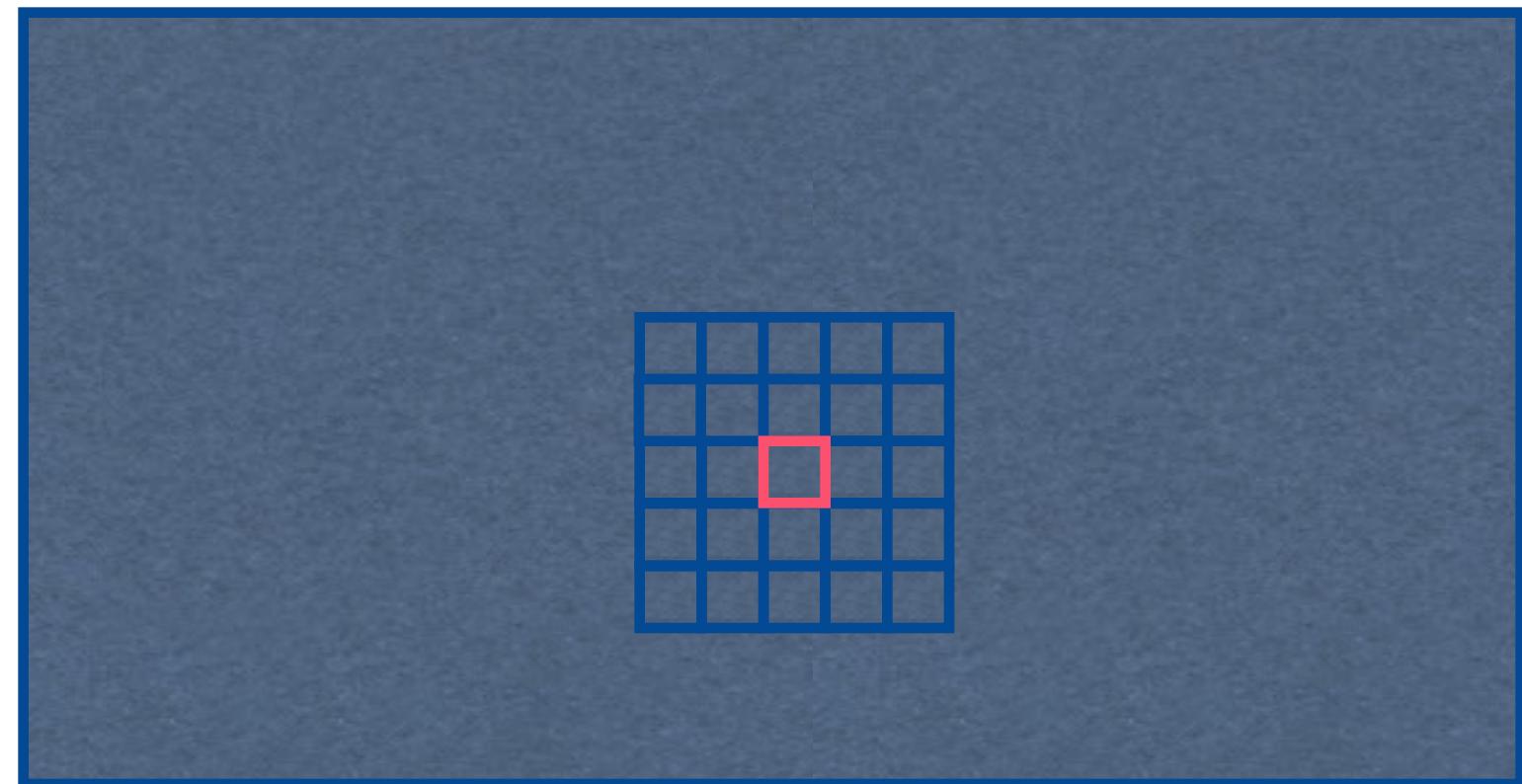
B



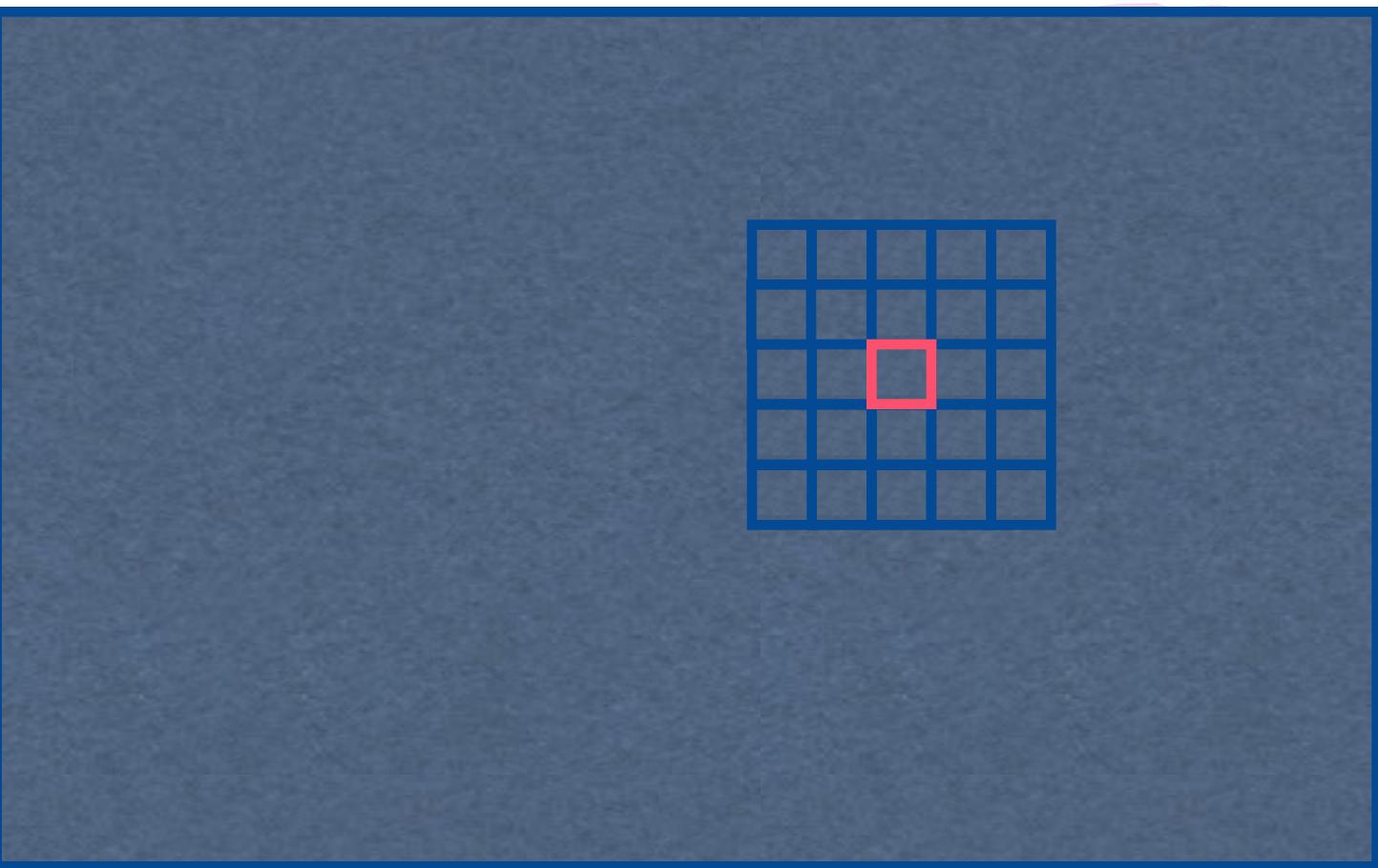
B'



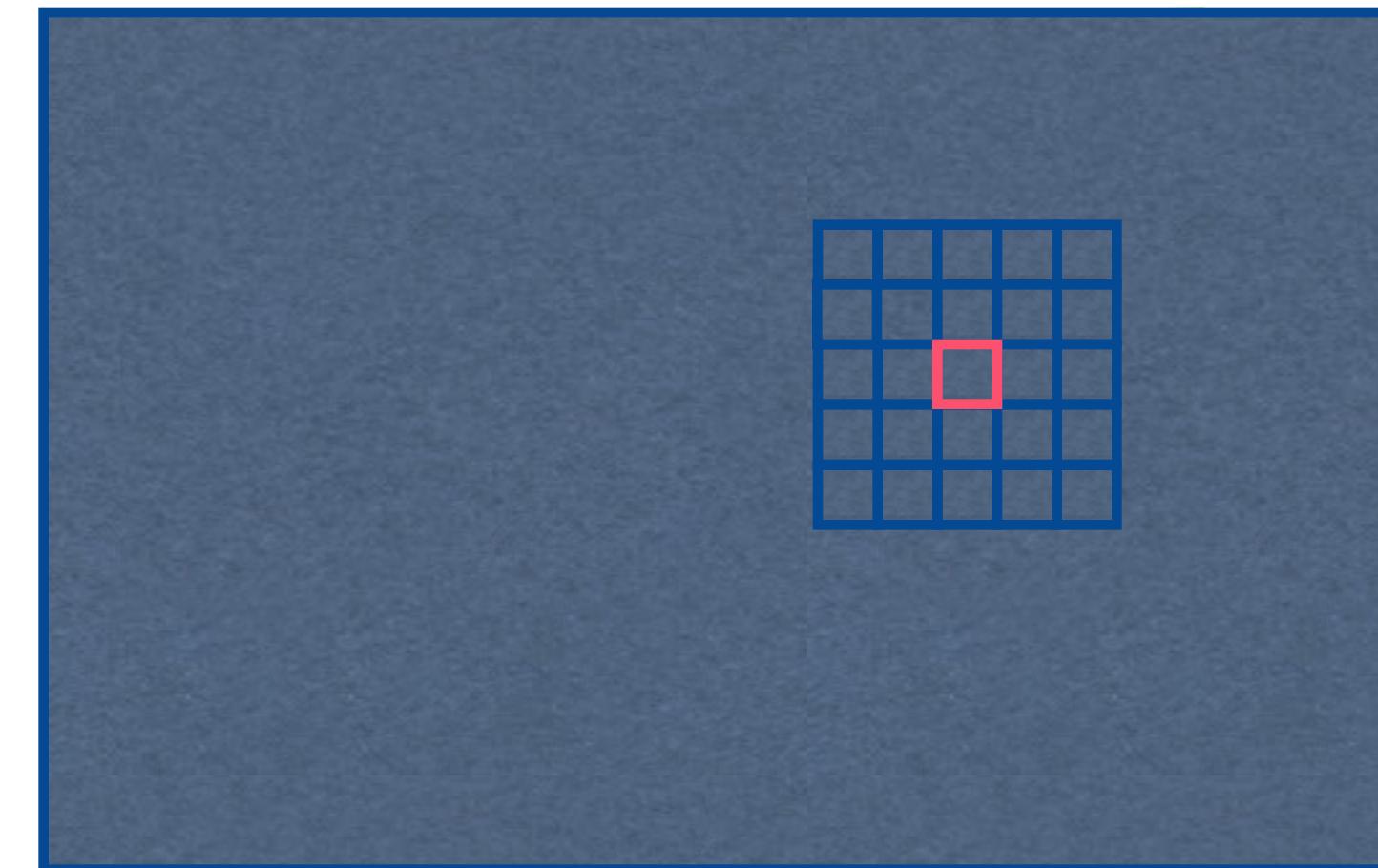
A



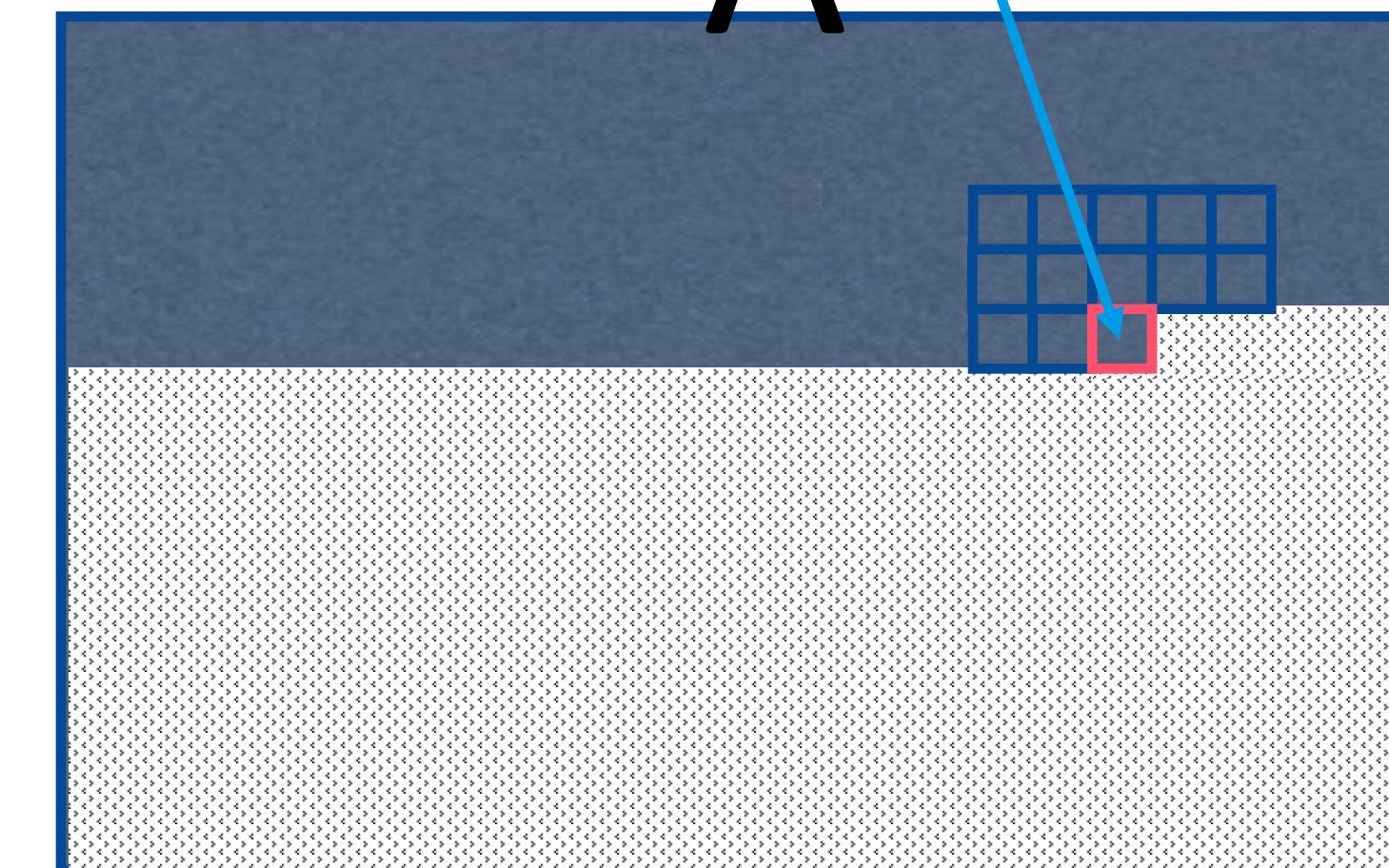
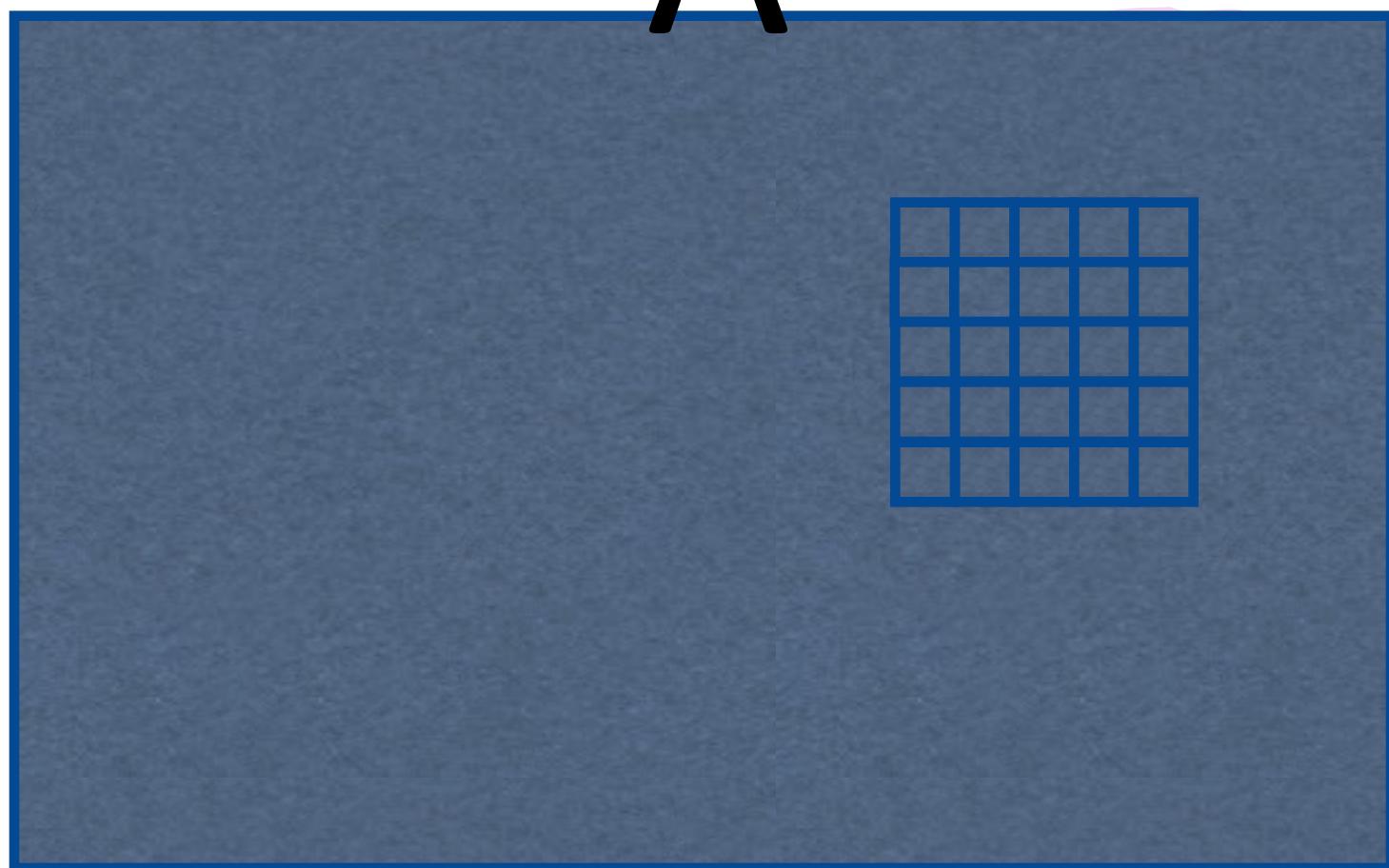
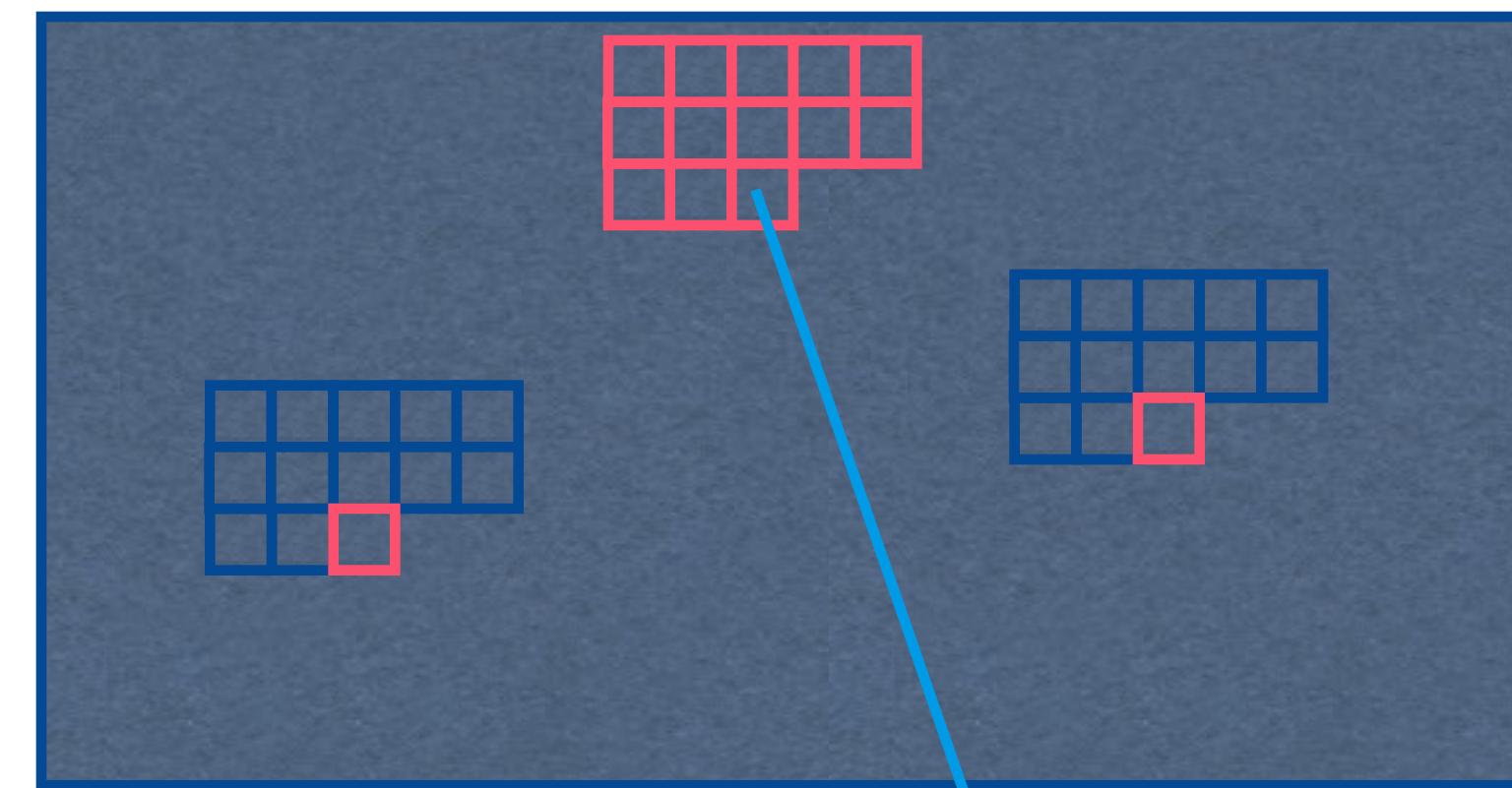
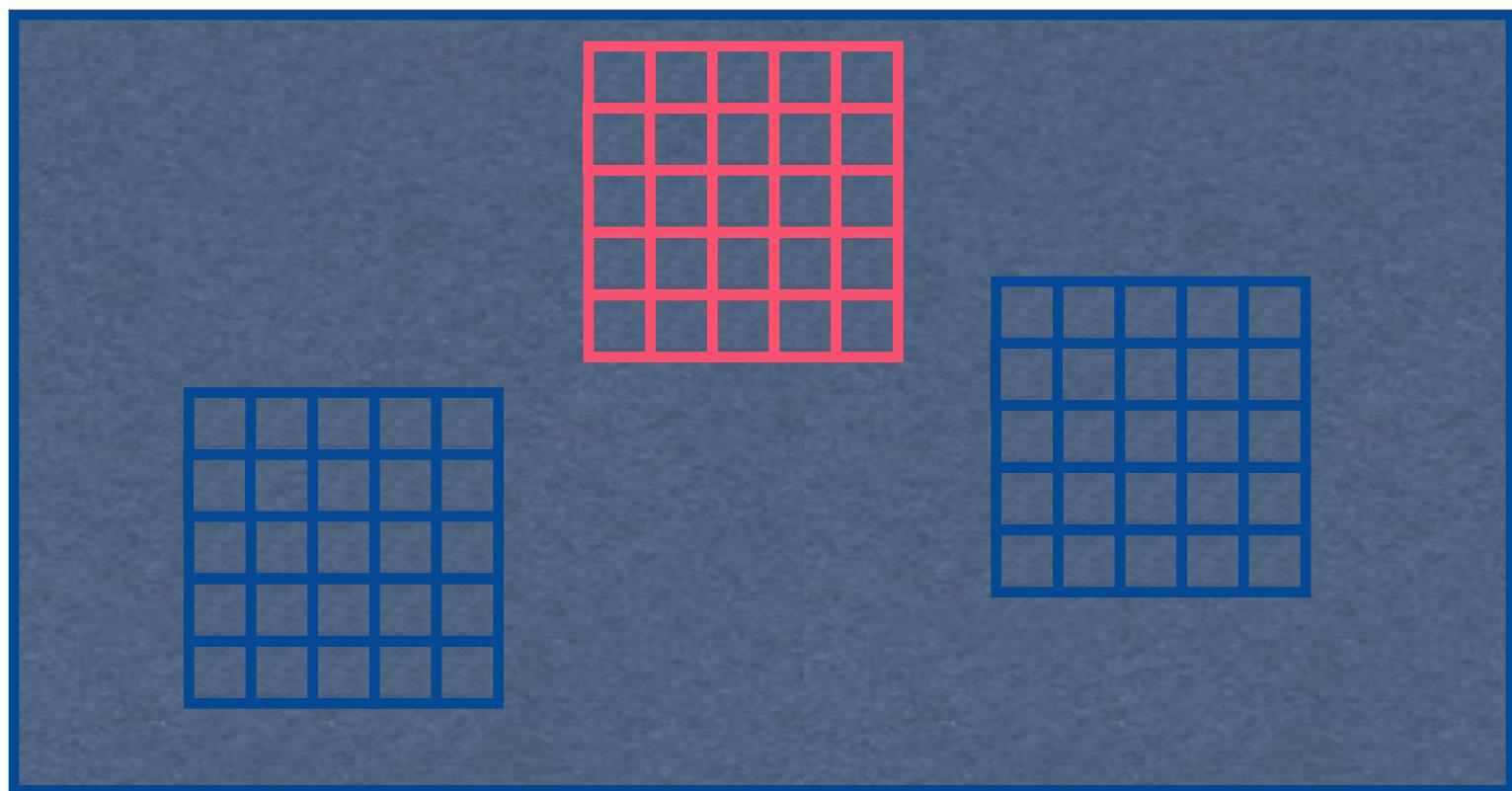
A'



B



B'



A

A'

B

B'

# Blur Filter



Unfiltered source ( $A$ )



Filtered source ( $A'$ )



Unfiltered target ( $B$ )

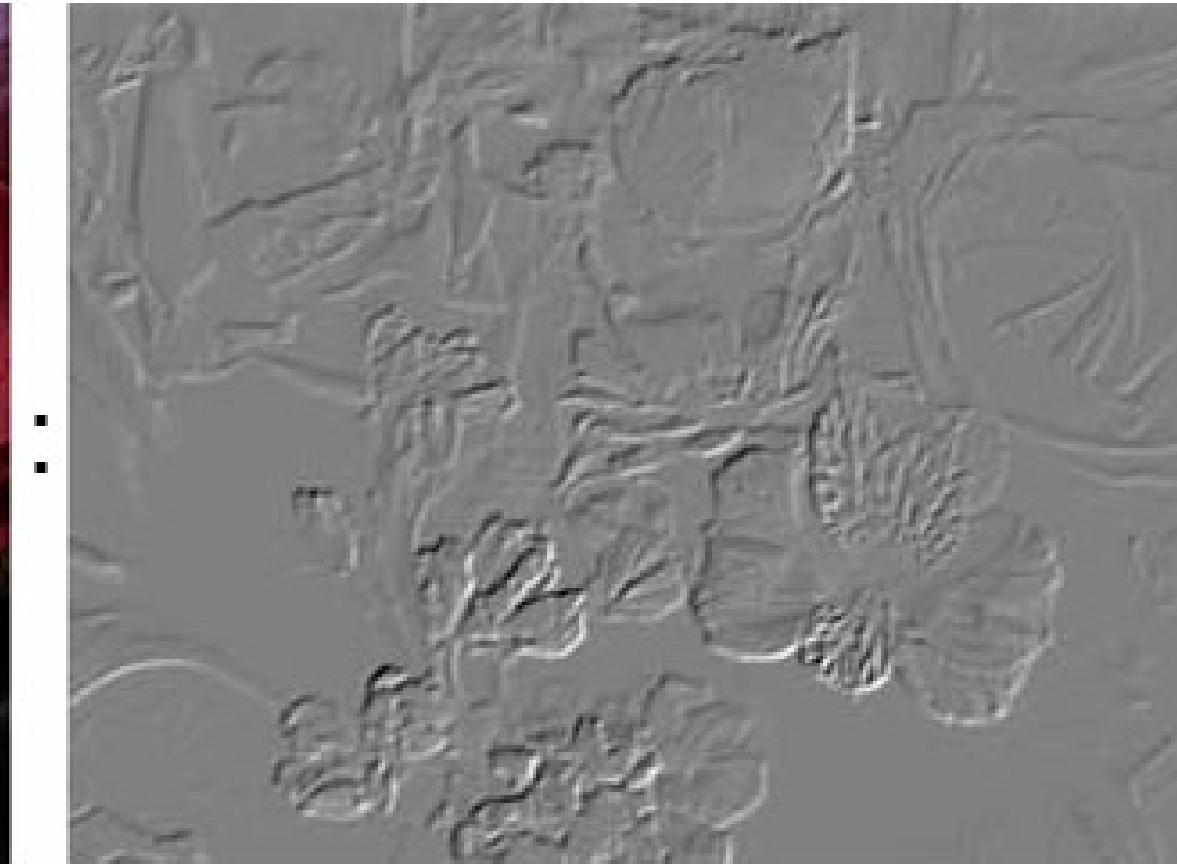


Filtered target ( $B'$ )

# Edge Filter



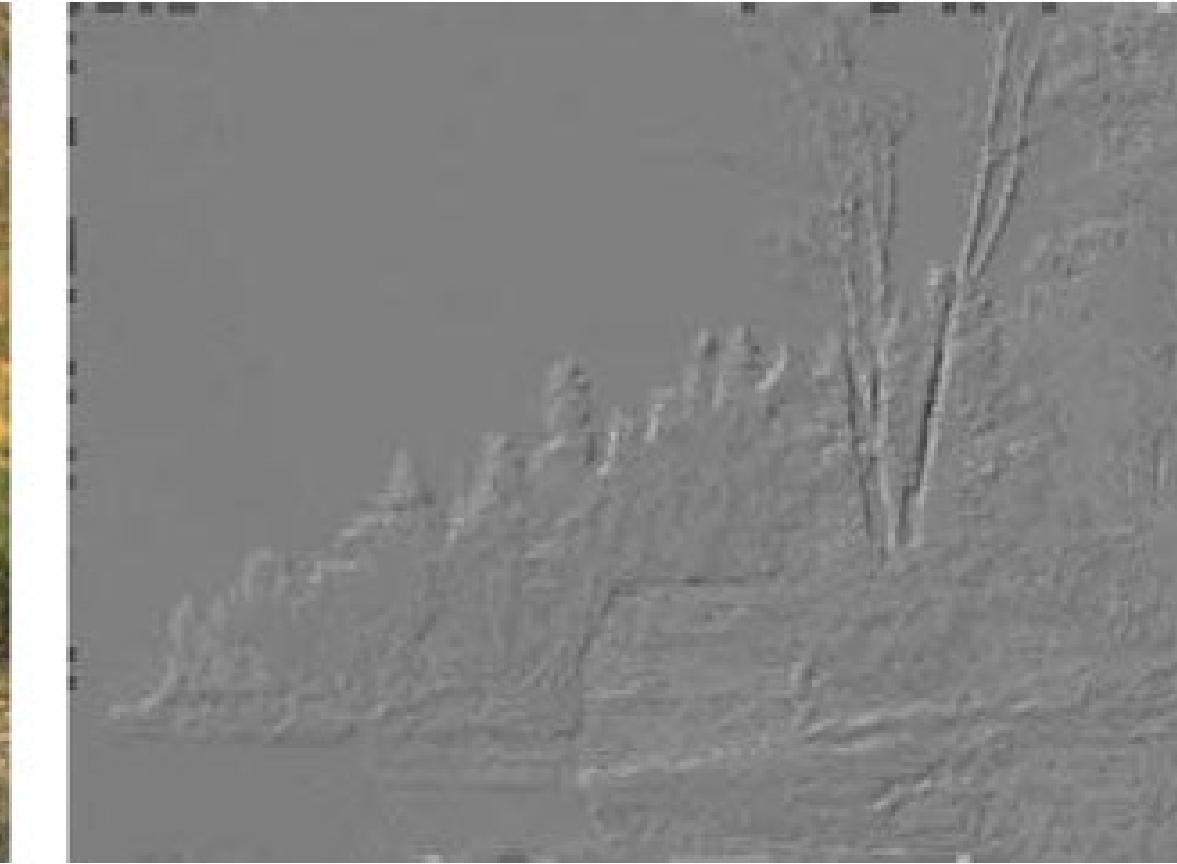
Unfiltered source (*A*)



Filtered source (*A'*)



Unfiltered target (*B*)



Filtered target (*B'*)

# Artistic Filters



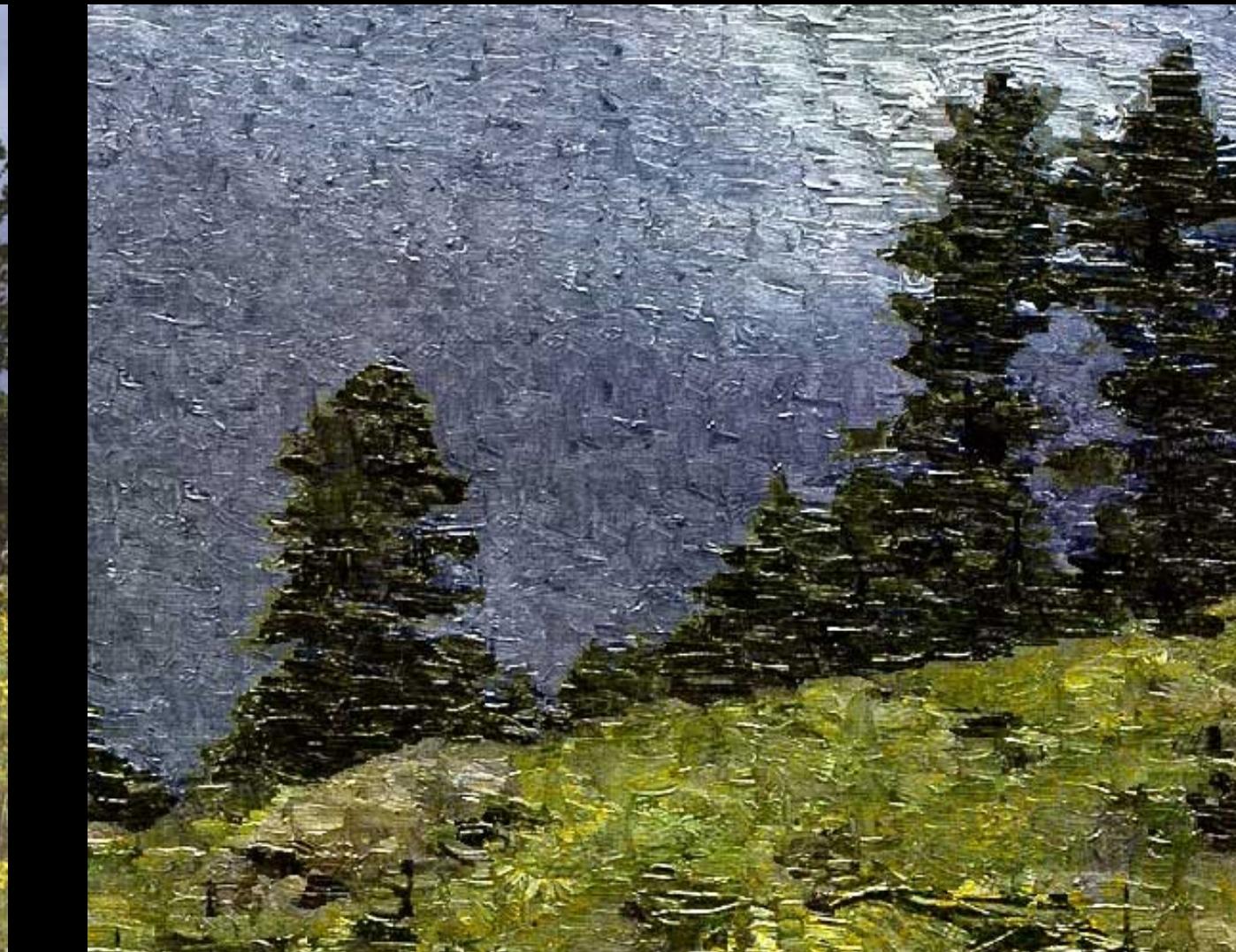
A



A'



B



B'

# Colorization



**Unfiltered source ( $A$ )**



**Unfiltered target ( $B$ )**



**Filtered source ( $A'$ )**

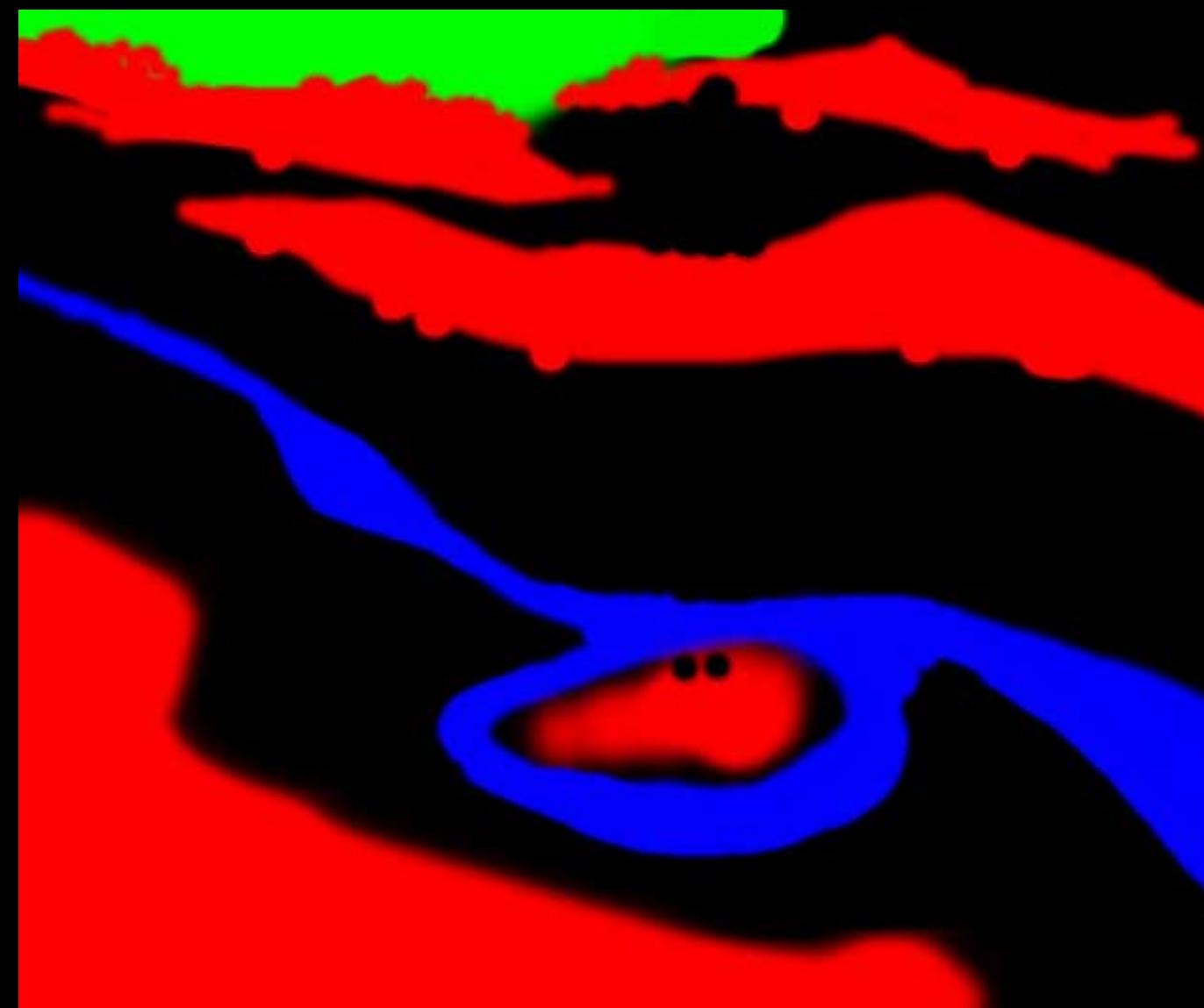


**Filtered target ( $B'$ )**

# Texture-by-numbers



A



B



A'



B'

# Visual Prompting via Image Inpainting

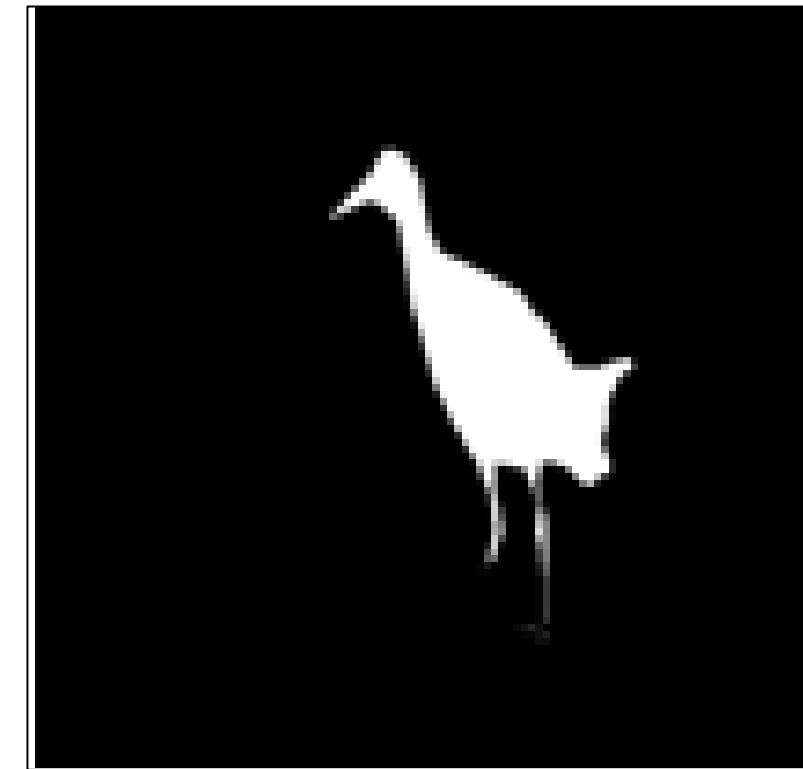
Amir Bar\*, Yossi Gandelsman\*,

Trevor Darrell, Amir Globerson, Alexei A Efros

NeurIPS 2022



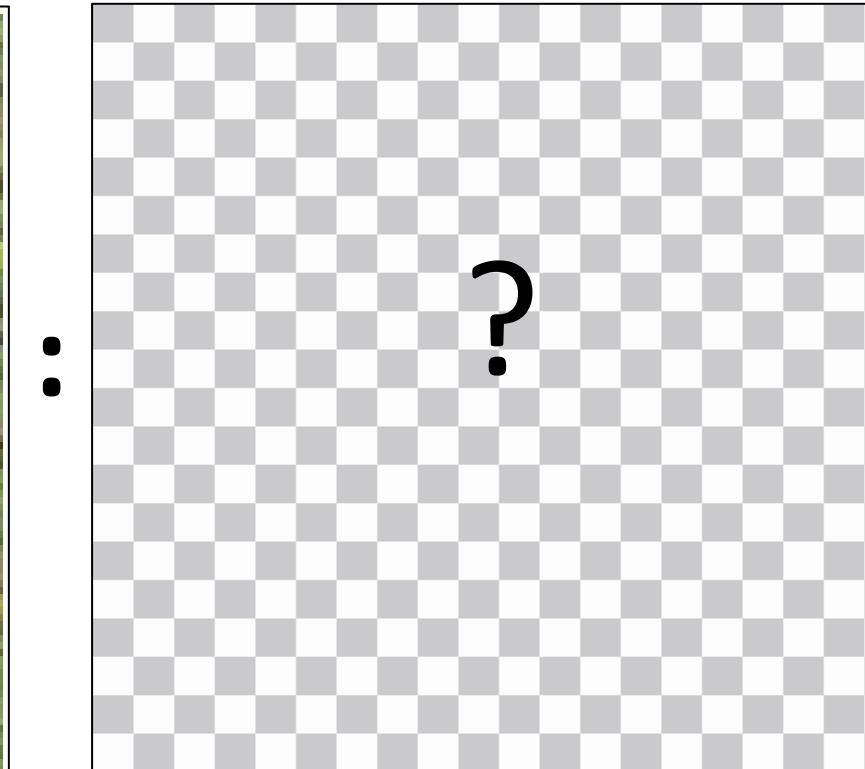
:



::

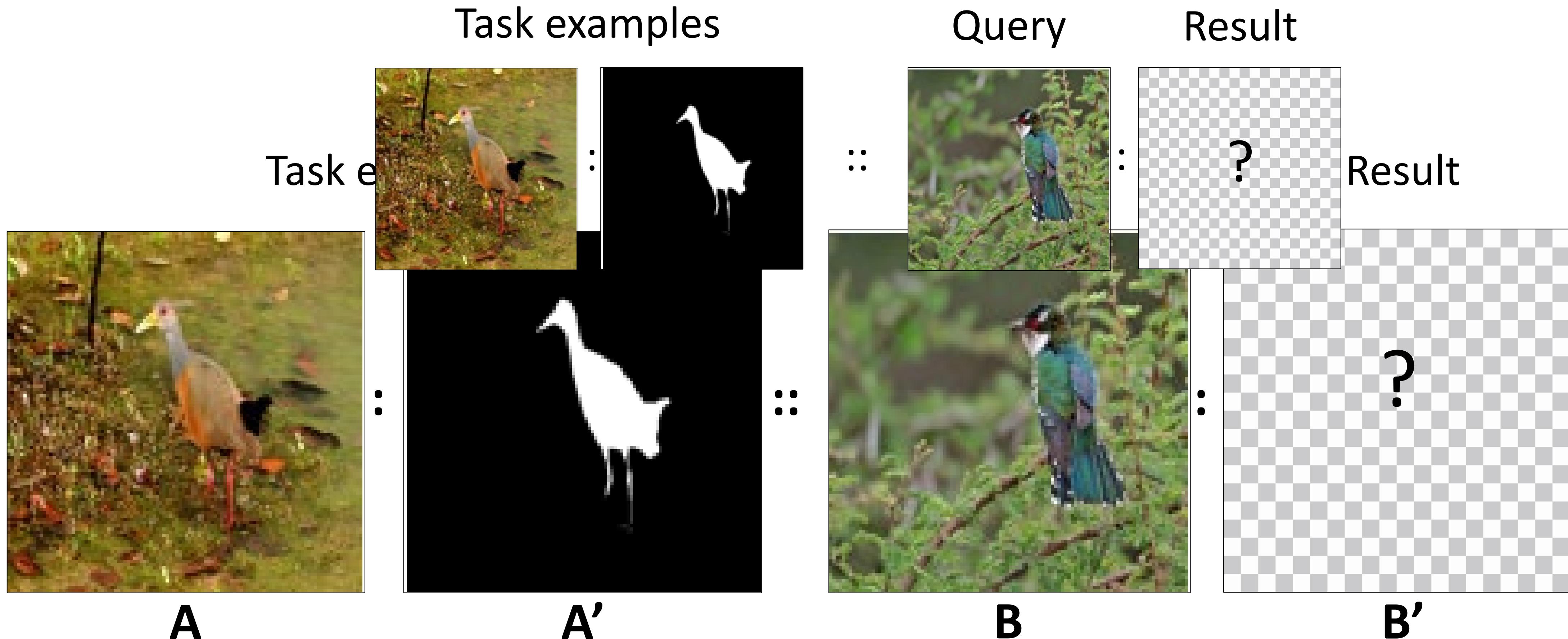


:



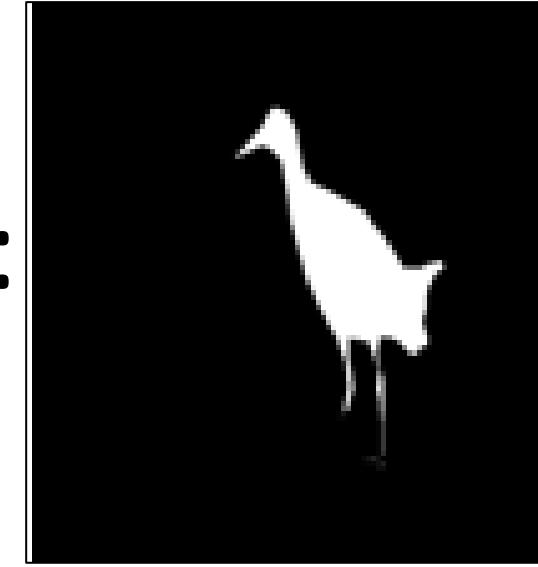
\* Equal contribution

# Visual Prompting

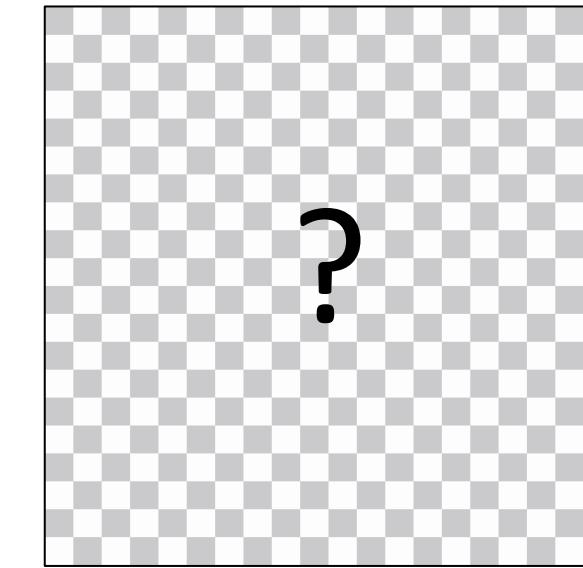
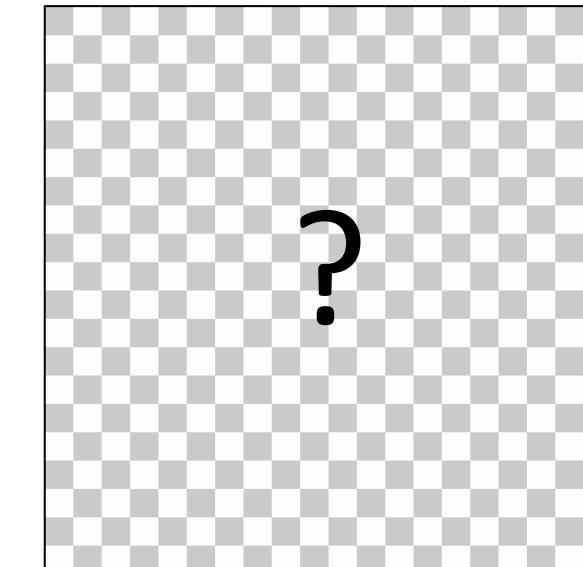
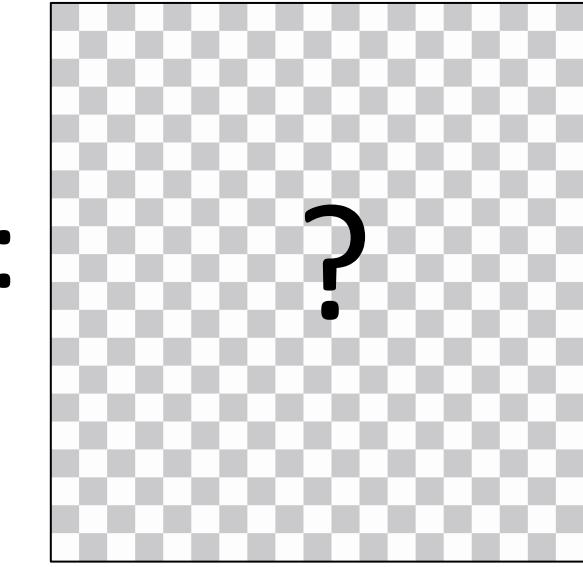


# Visual Prompting

Task examples



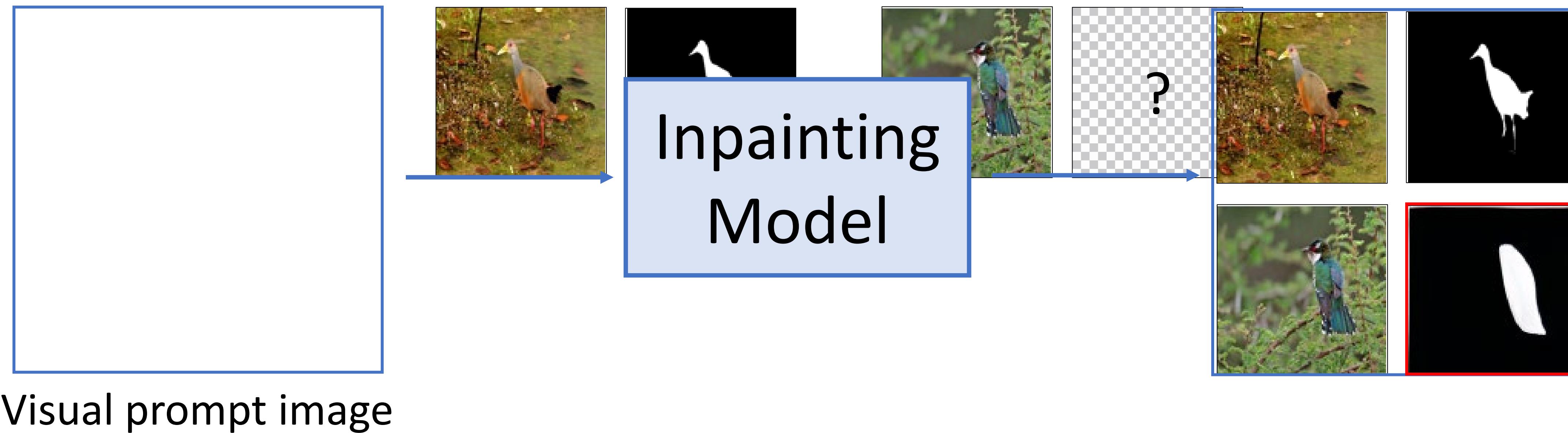
::



Query

Result

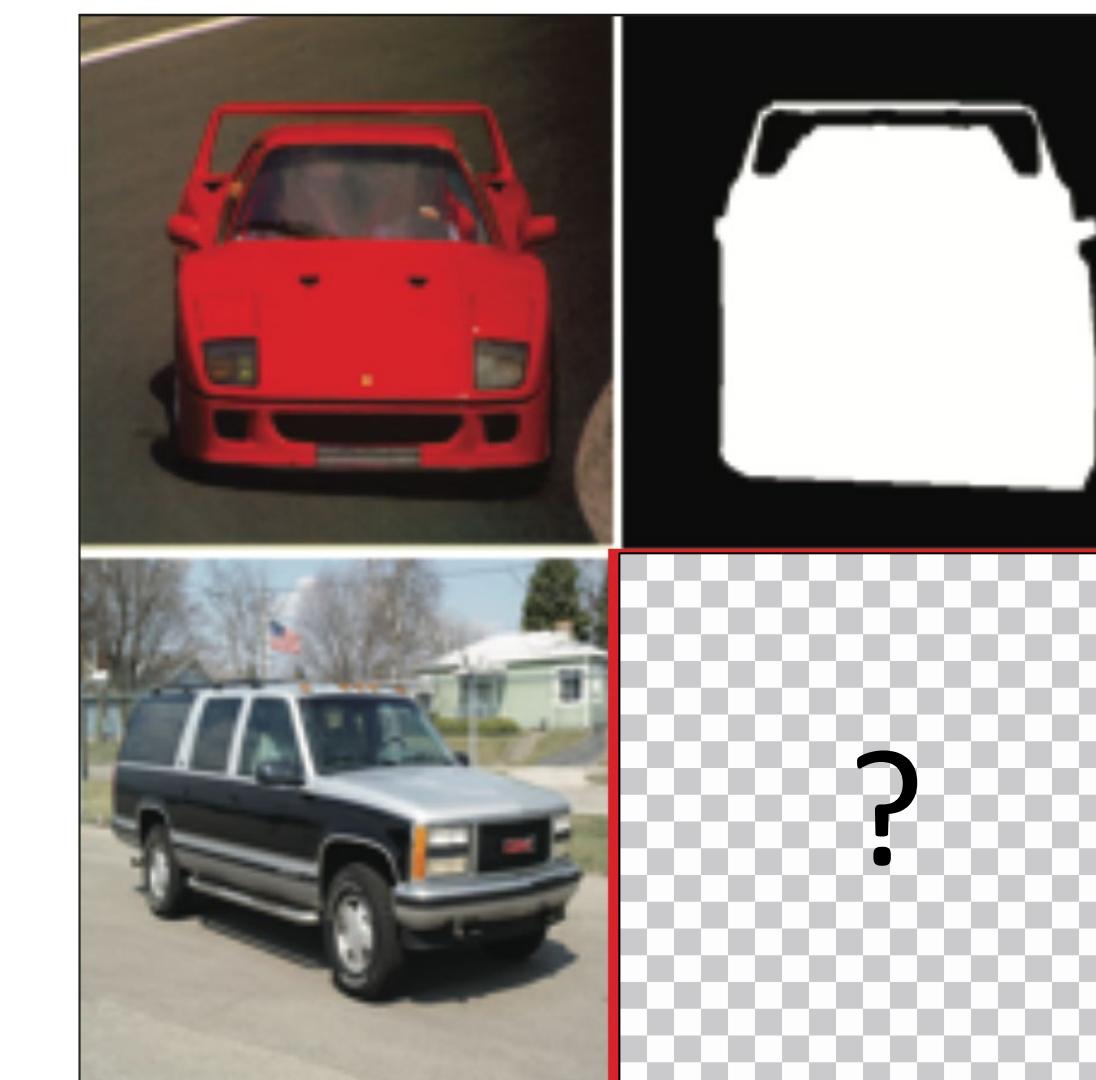
# Inpainting models to the rescue!



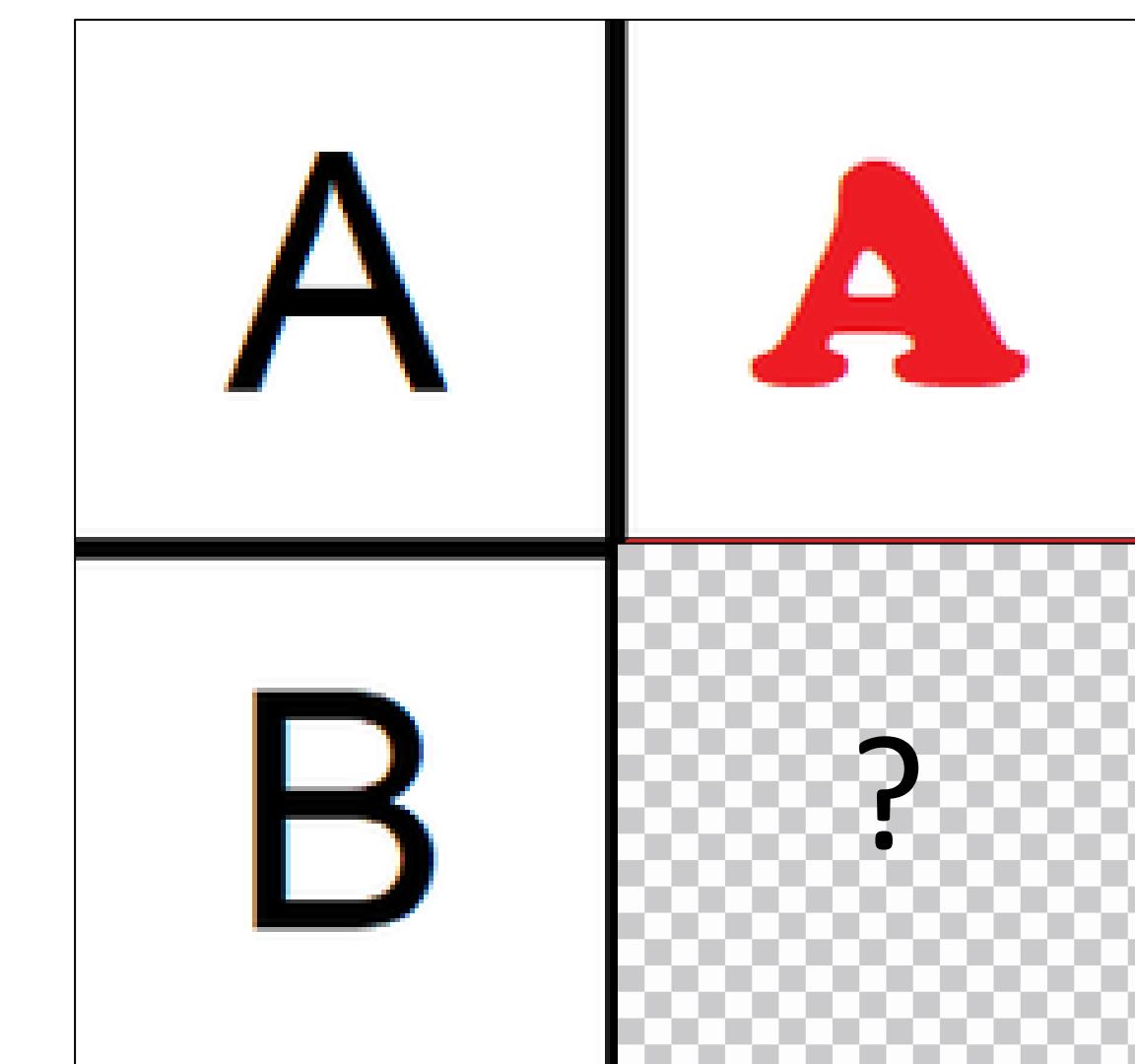
# Wide range of tasks



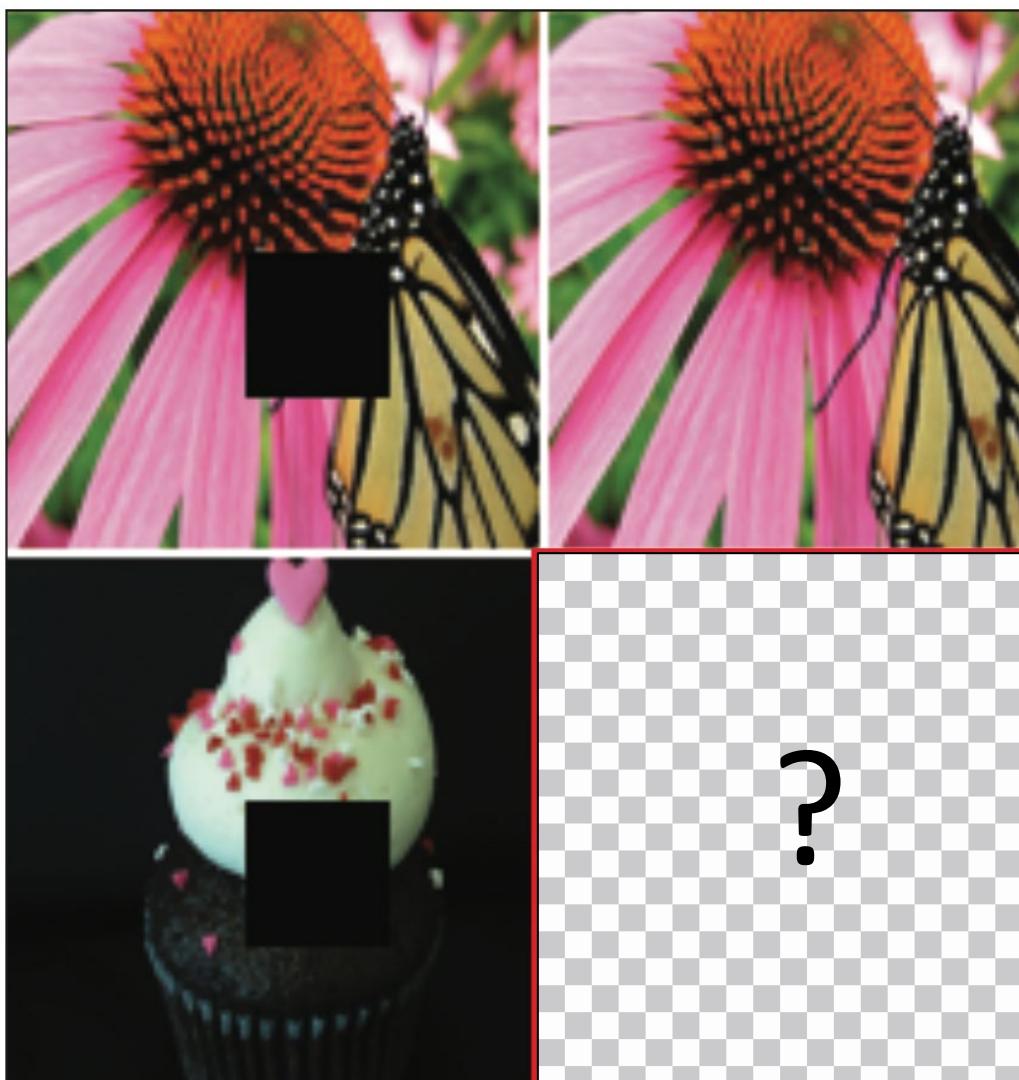
Colorization



Segmentation



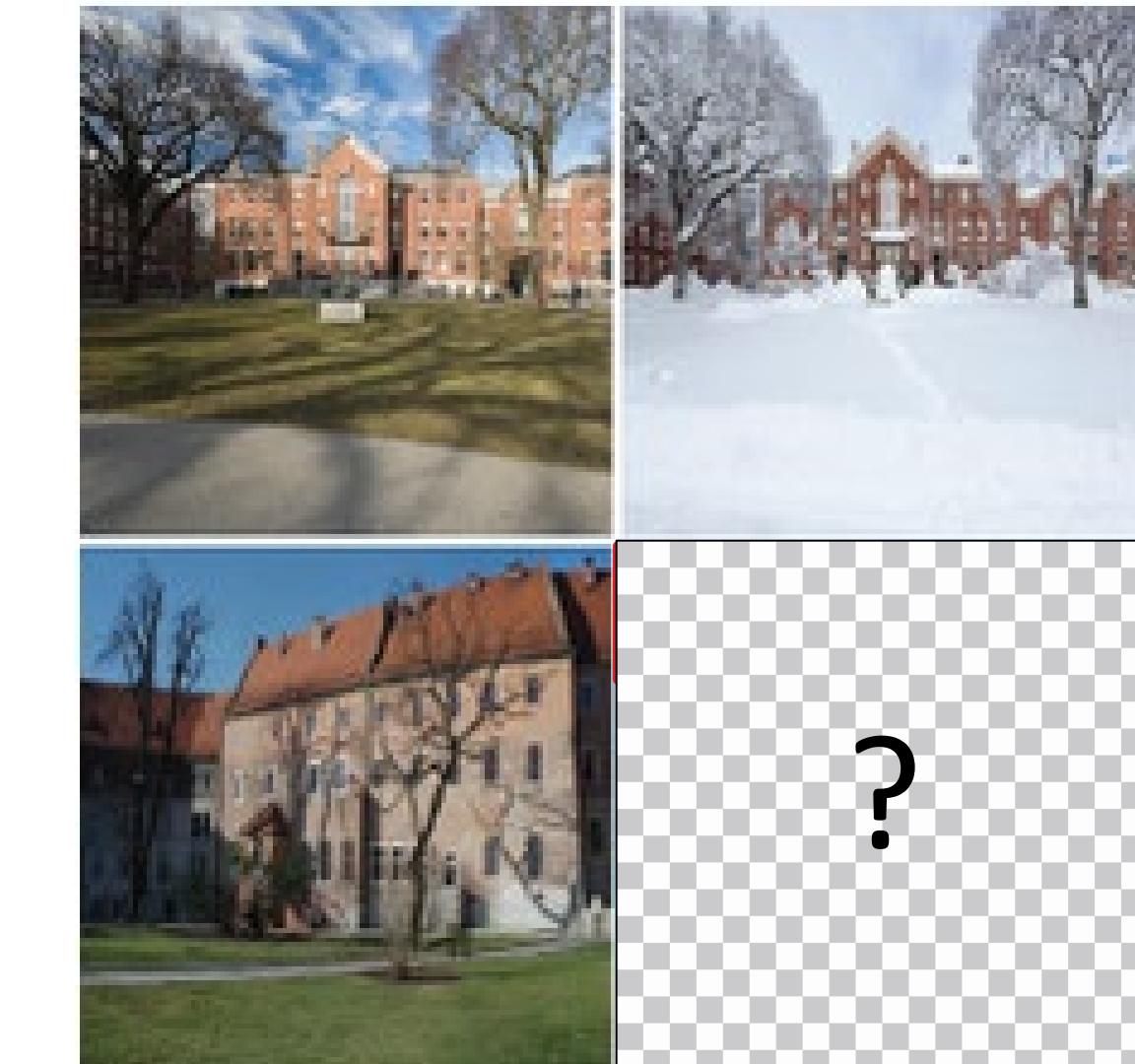
Font Style Transfer



Inpainting



Edge Detection



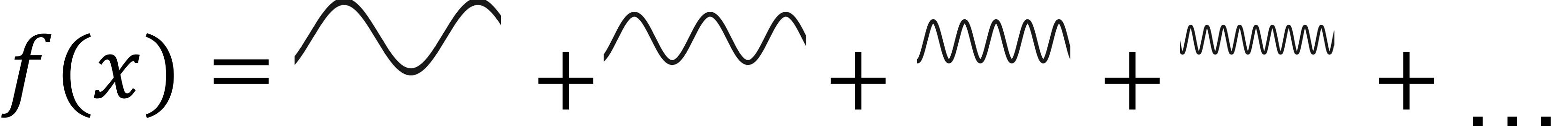
Style Transfer

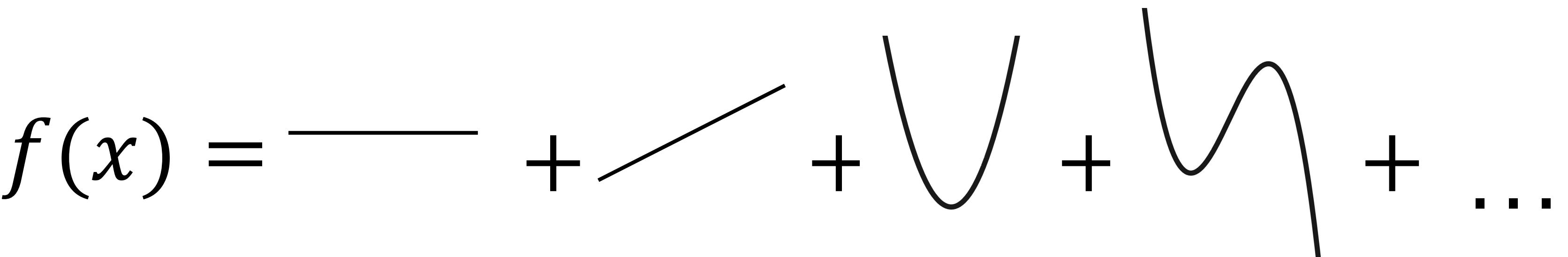
Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still red

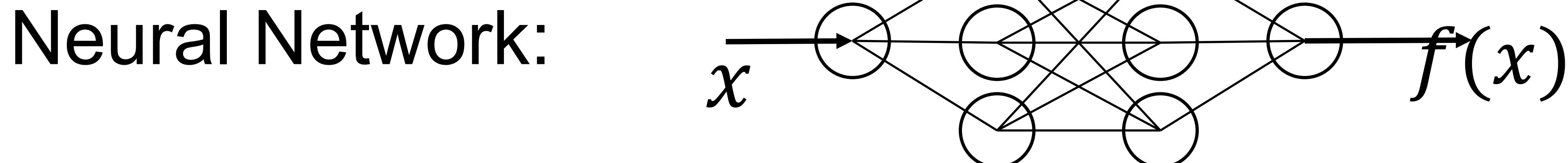
$$P(x_n | x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, x_{n-8}, x_{n-9}, x_{n-10}, x_{n-11}, x_{n-12}, x_{n-13})$$

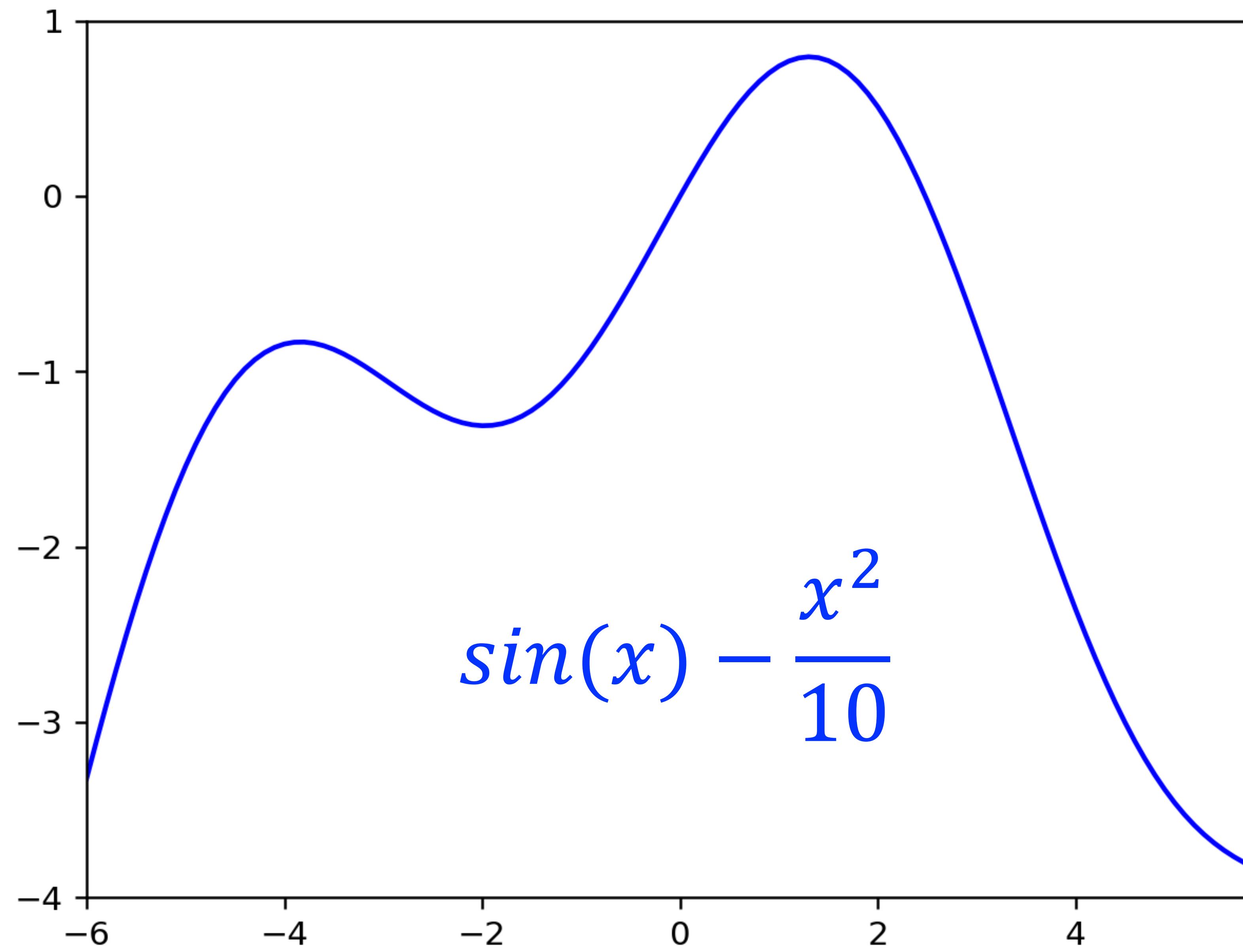
$10^{70}$  combinations

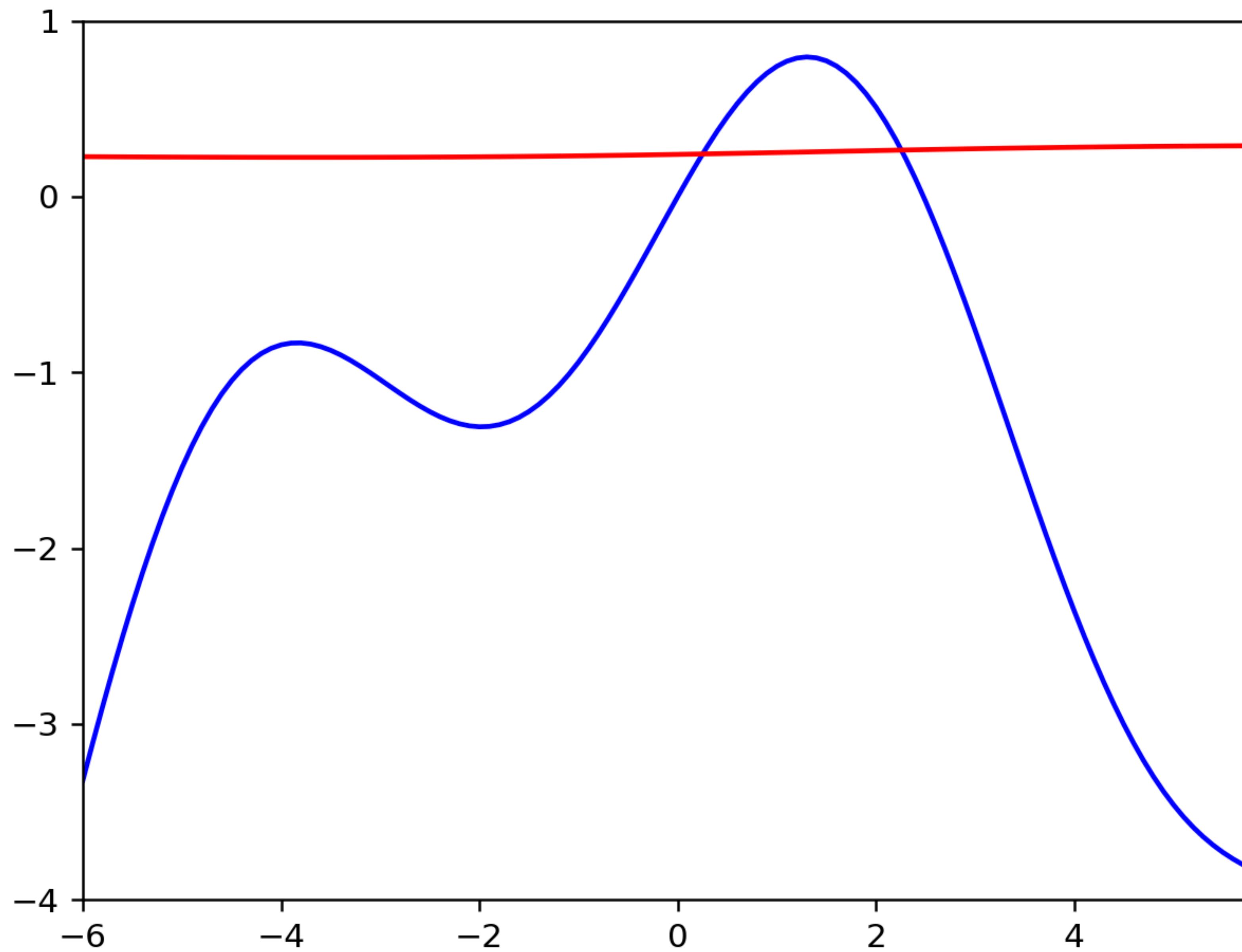
# Function Approximation

Fourier Series:  $f(x) =$   + ...

Taylor Series:  $f(x) =$   + ...







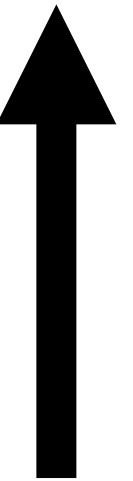


e Seitz's [video](#)

$$P(x_n | x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, \dots)$$

Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still red

red

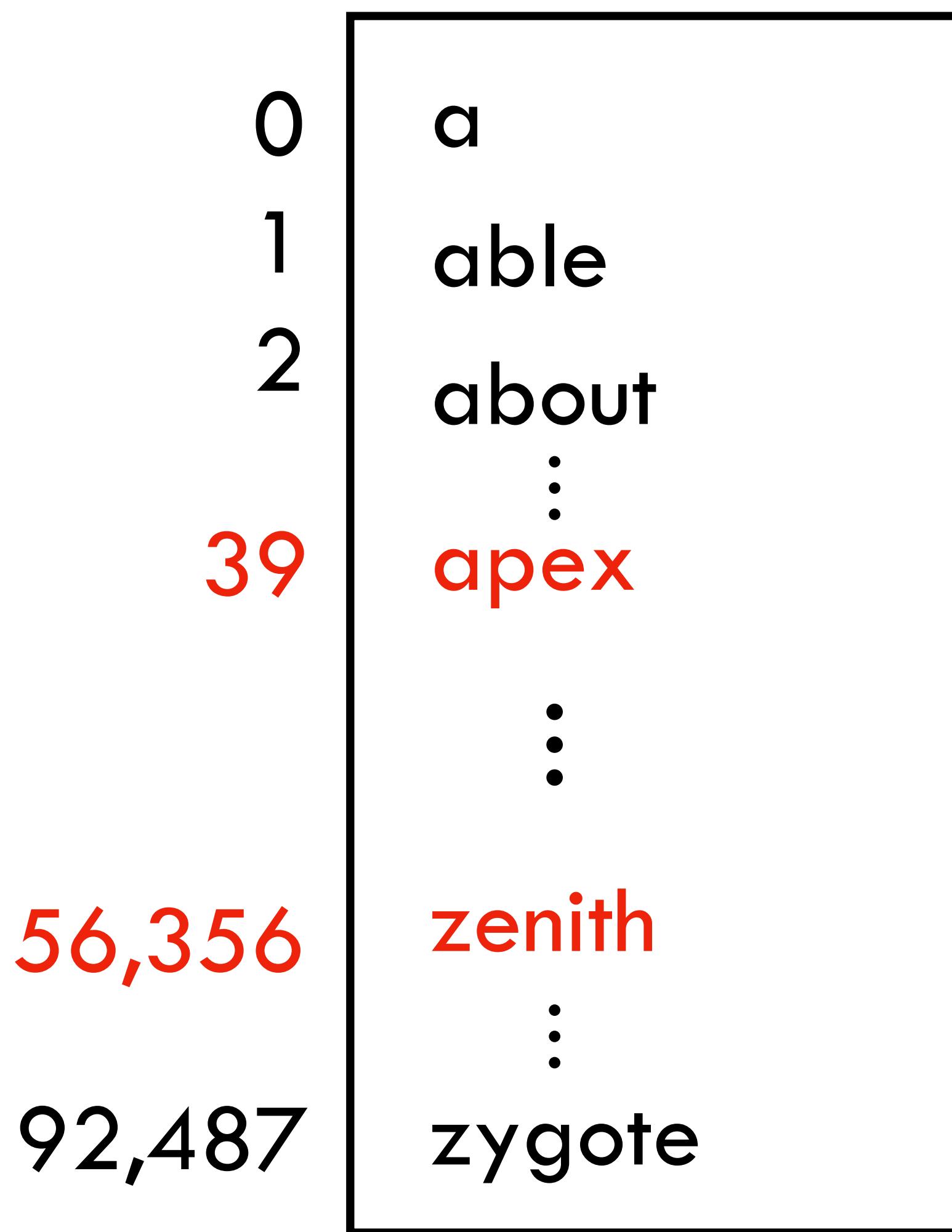


Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still

red

# neural network

Early one morning the sun was shining was laying in bed wondering if she had changed at all if her hair was still

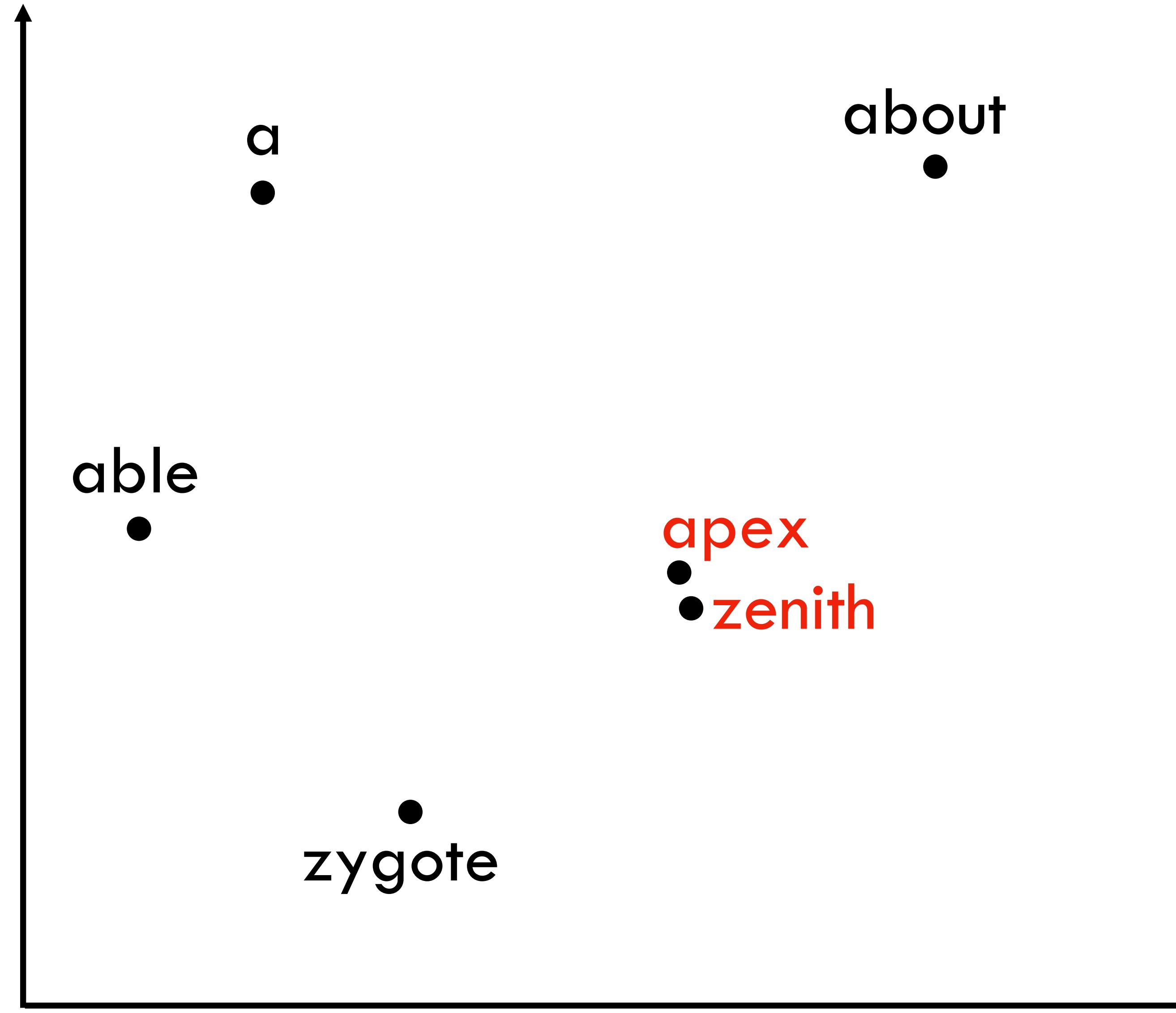


# word2vec

[Collobert & Weston 2008; Mikolov et al. 2013]

house, where the professor lived without his wife and child; or so he said jokingly sometimes: “Here’s where I live. My house.” His daughter often added, without resentment, for the visitor’s information, “It started out to be for me, but it’s really his.” And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked “Kitty” and half full of eternal milk; but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter’s preference was for the store-bought gimmicks and appliances, the toasters and carpet sweepers of Lilliput, but she knew that most adult visitors would

Deep  
Net



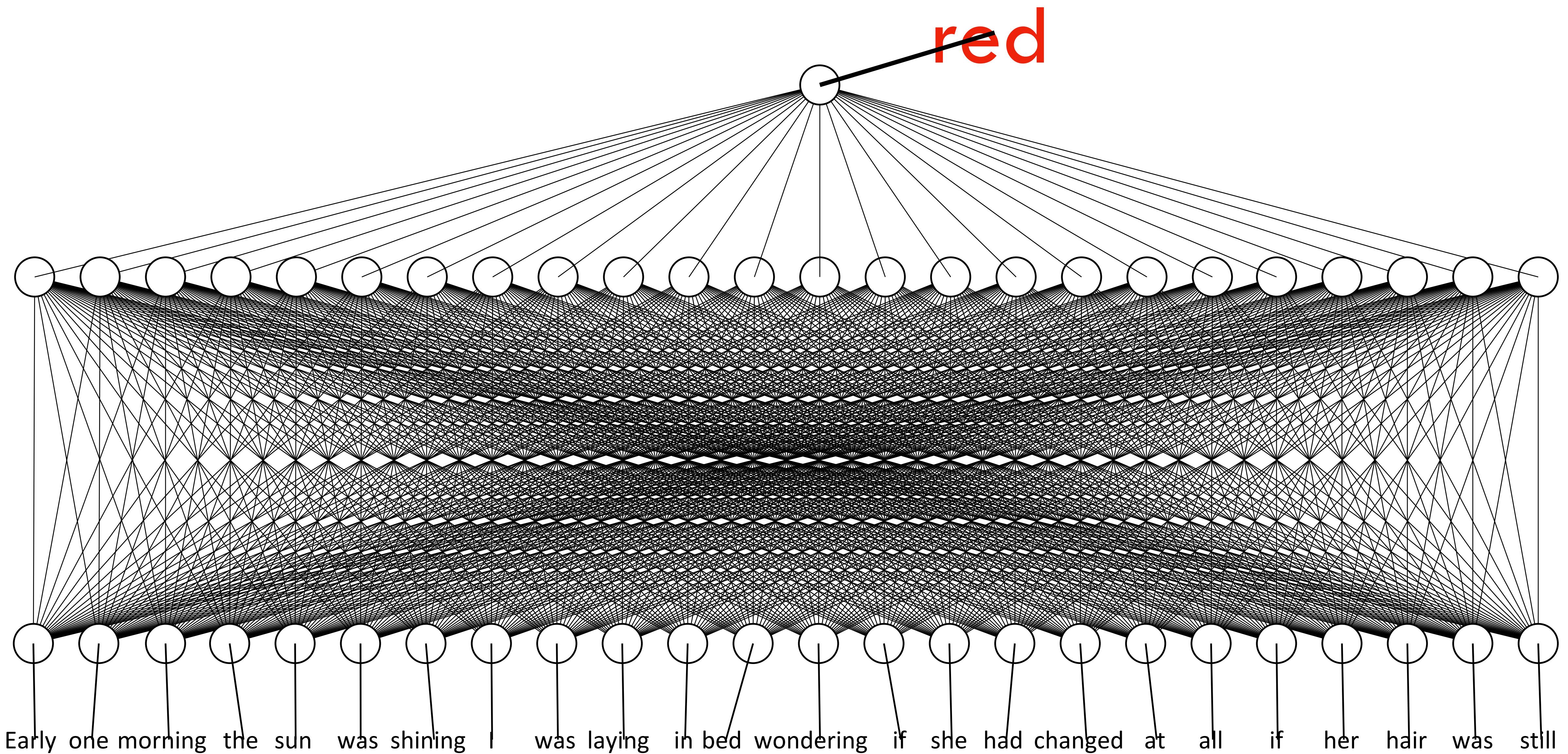
**Word Embedding** (e.g., word2Vec, GloVe)

slide from Steve Seitz's [video](#)

red

# neural network

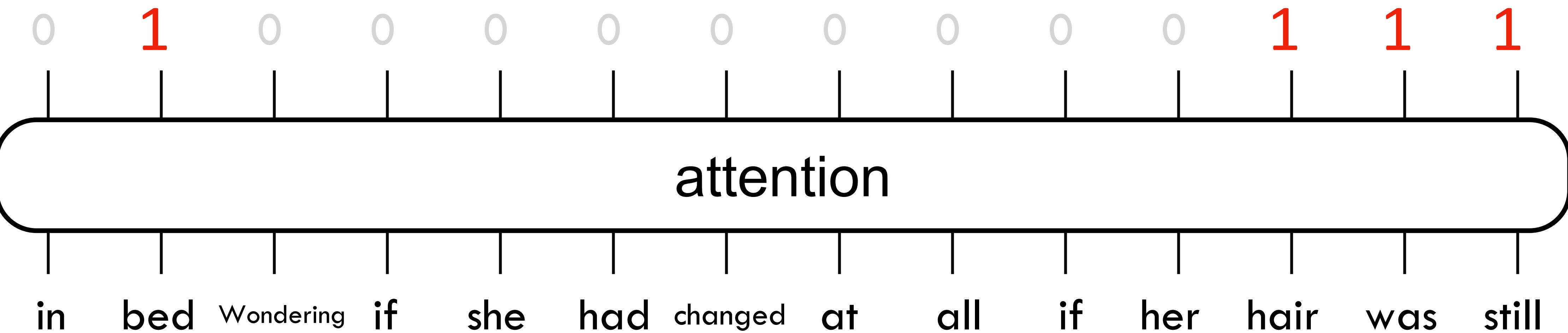
Early one morning the sun was shining was laying in bed wondering if she had changed at all if her hair was still



Early one morning the sun was shining I was laying in bed  
Wondering if she had changed at all if her hair was still ?

bed

hair was still red



slide from Steve Seitz's [video](#)

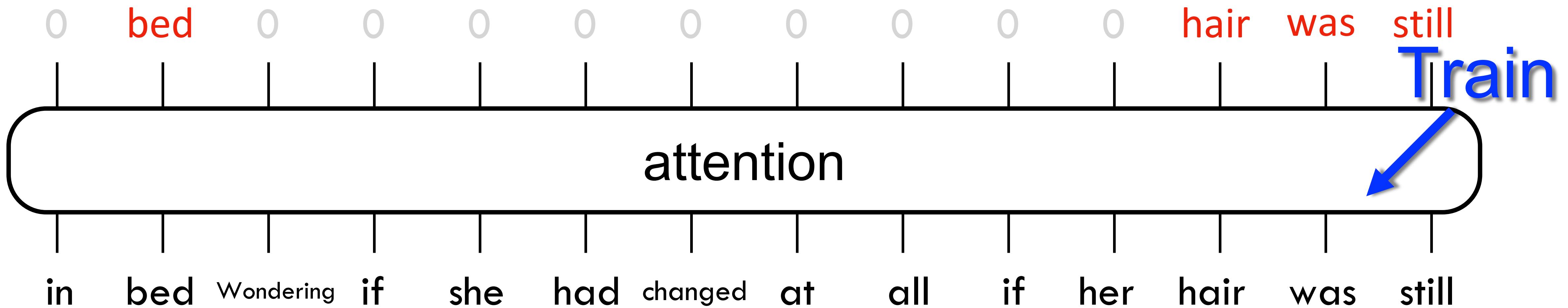
red

next word prediction

0 bed 0 0 0 0 0 0 0 hair was still

attention

in bed Wondering if she had changed at all if her hair was still



Two roads diverged in a yellow wood  
And sorry I could not travel both  
And be one traveler, long I stood  
And looked down as far as I could  
To where it bent in the undergrowth;

Robert Frost, *Road Not Taken*

slide from Steve Seitz's [video](#)

Train

red

next word prediction

0 bed 0 0 0 0 0 0 0 hair was still

attention

in bed Wondering if she had changed at all if her hair was still

Train

brown

next word prediction

0 bed 0 0 0 0 0 0 0 hair was still

attention

in bed Wondering if she had changed at all if her hair was still

Train

brown

next word prediction

0 bed 0 0 0 0 0 0 0 hair was still

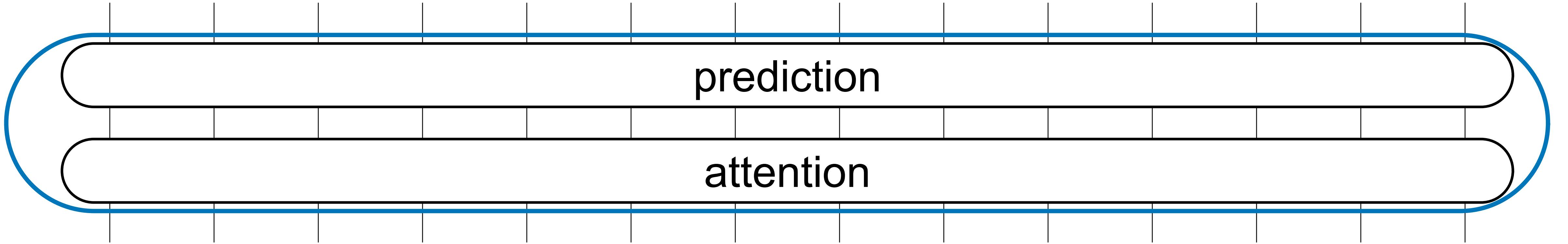
in bed Wondering if she had changed at all if her hair was still

attention

red

# Transformer

in bed Wondering if she had changed at all if her hair was still

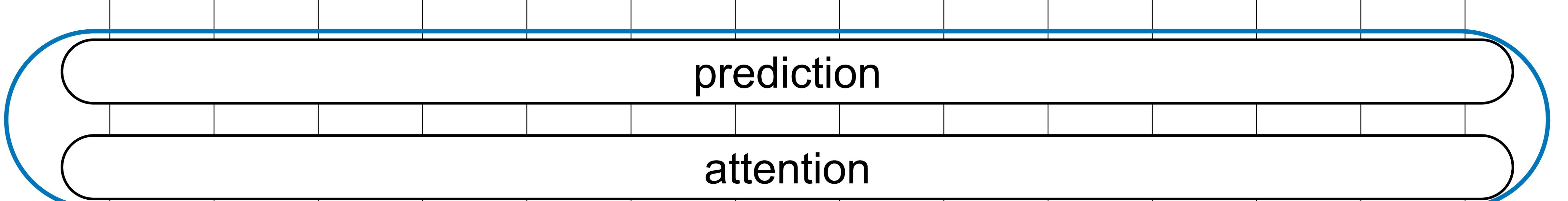


prediction

attention

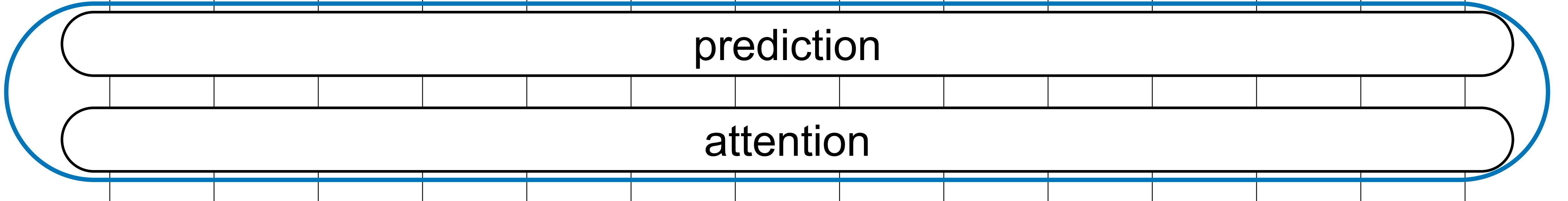
⋮

**96 (GPT-3) 118 (Palm)**



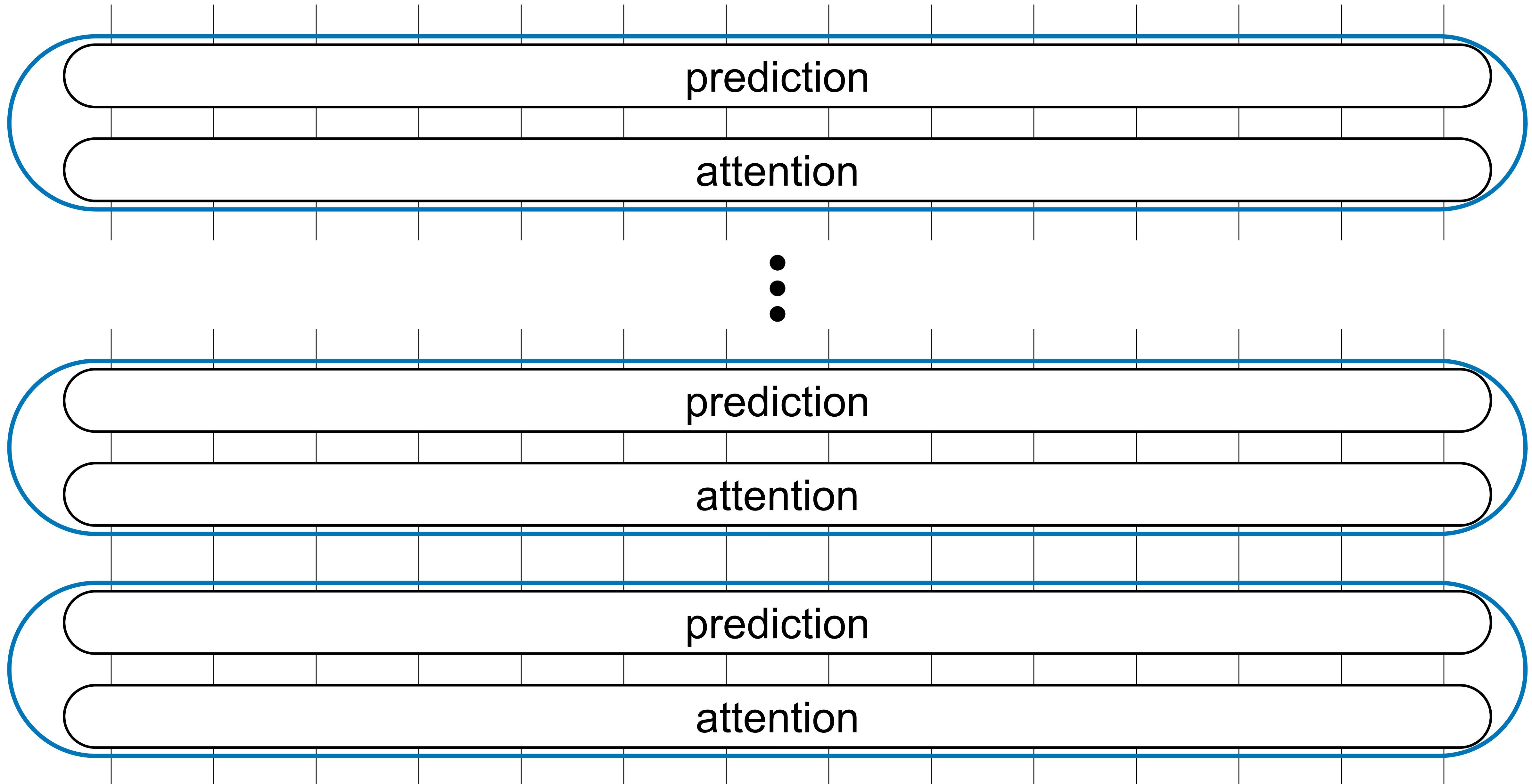
prediction

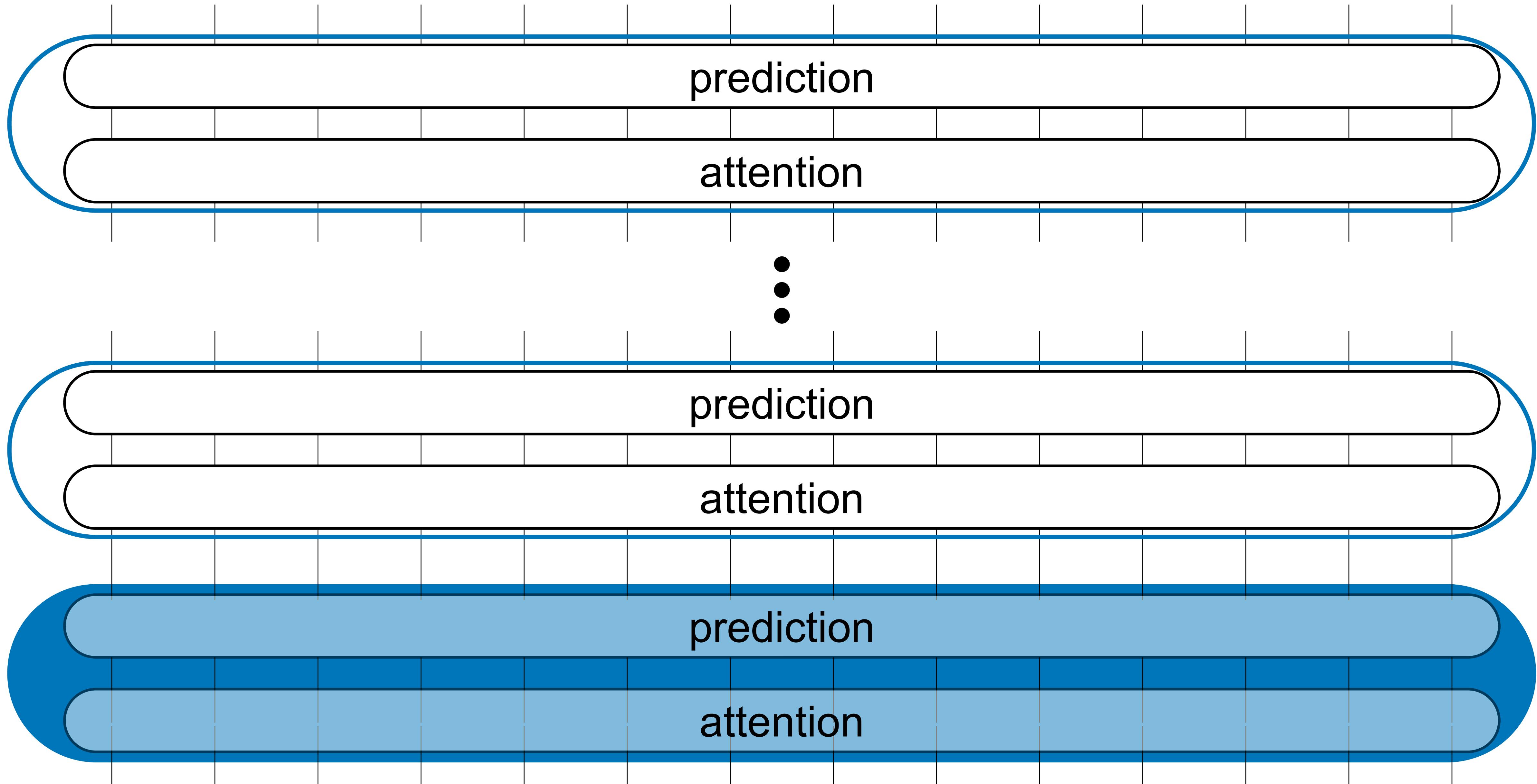
attention



prediction

attention

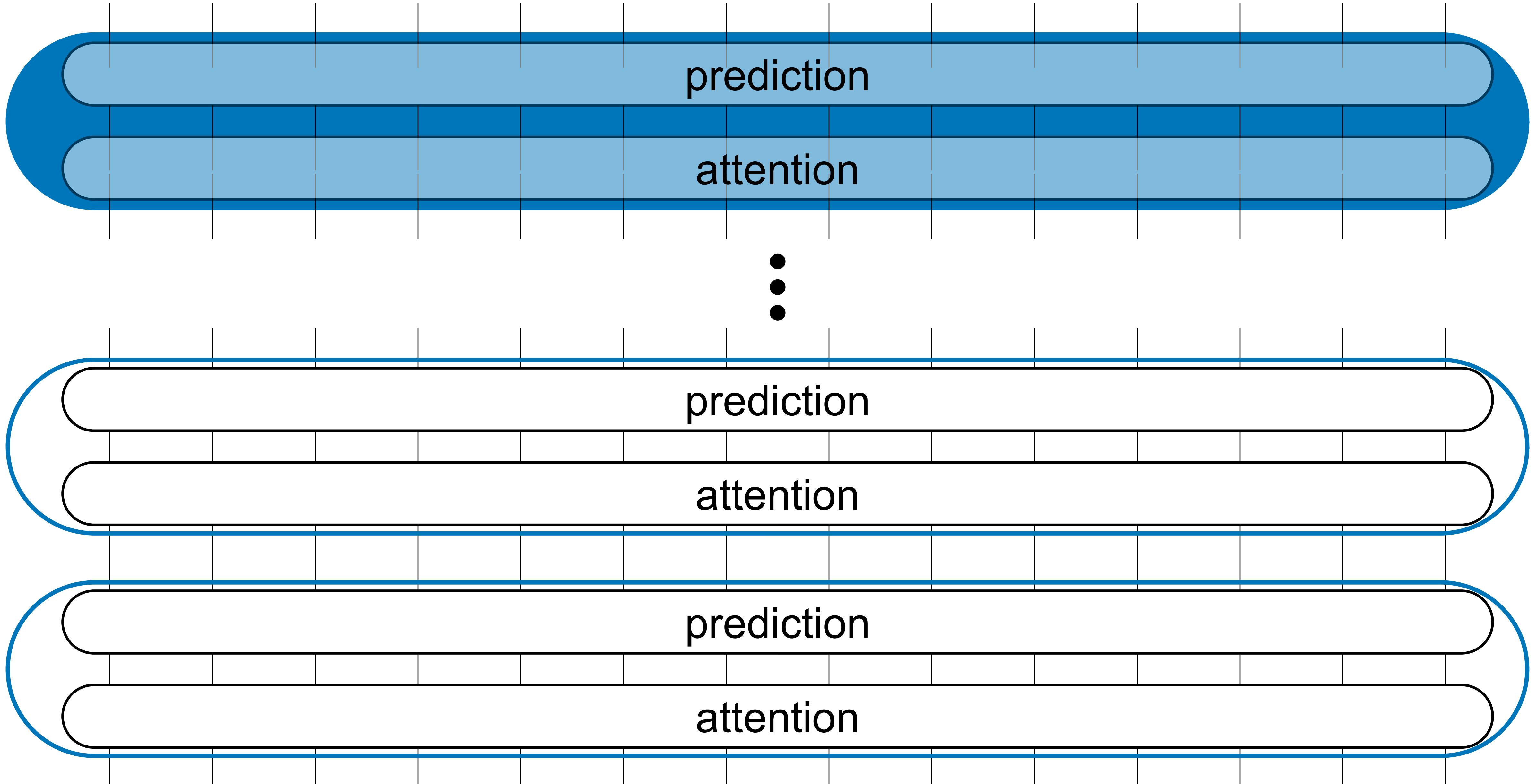


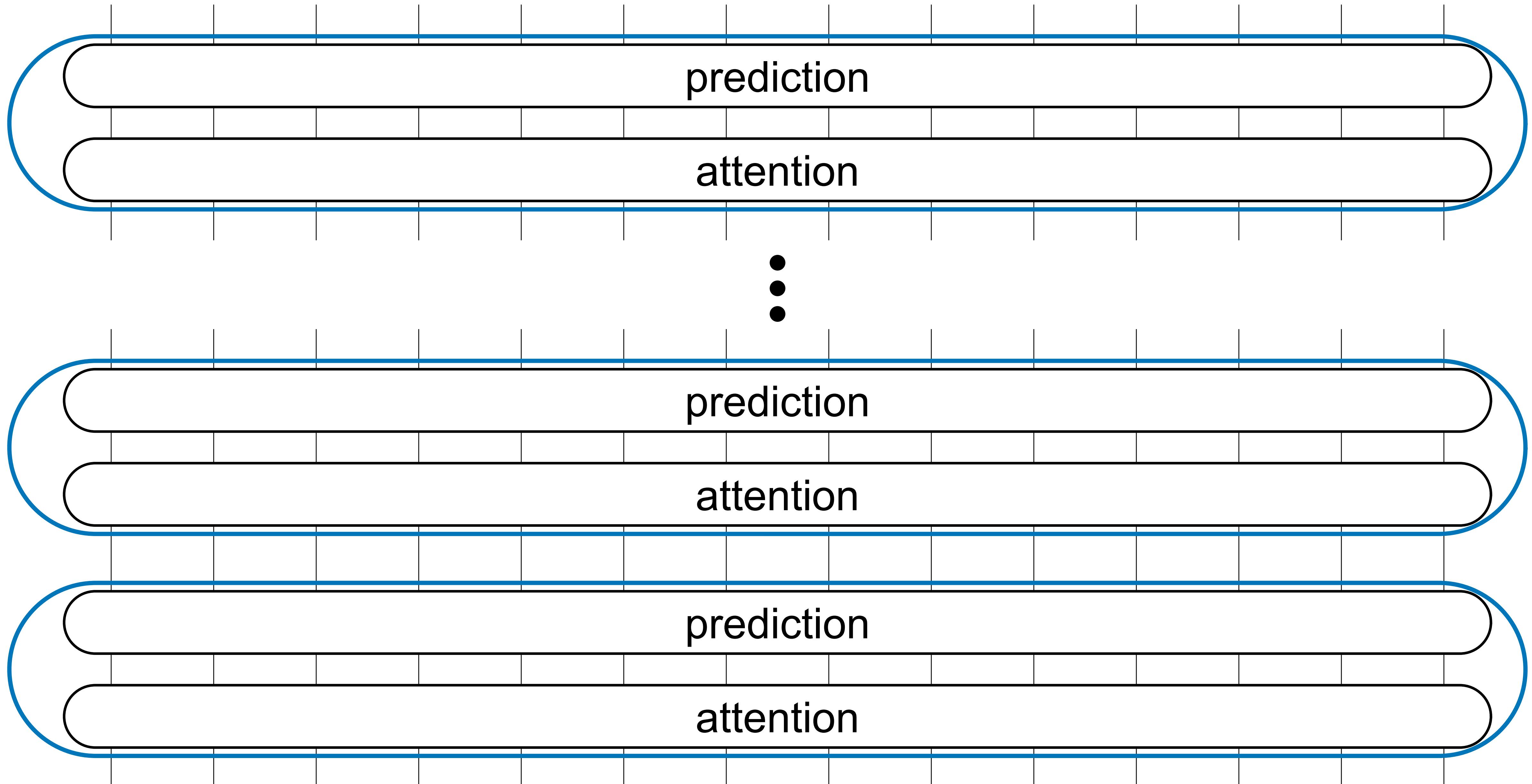


**Syntax**

slide from Steve Seitz's [video](#)

# Semantics





How much data  
to train?

All of it...

**355 years**

a month

The 16th President was

The capital of Zimbabwe is

Frank Zappa's middle name is

Napoleon was born on this date

The prime factorization of 19456721434 is

Queen Victoria's maiden name was

US per-capita income in 1957 was

The lat long coordinates of Rome are

The 16th President was Abraham Lincoln

The capital of Zimbabwe is Harare

Frank Zappa's middle name is Vincent

Napoleon was born on this date 1769

The prime factorization of 19456721434 is  $2 \times 3 \times 3 \times 17$

Queen Victoria's maiden name was Alexandrina Victoria

US per-capita income in 1957 was \$2,974

The lat long coordinates of Rome are 41.894722, 12.48

a pattern of characters that looks like a star

```
•   O   •  
○ ○ ○  
•   O   •
```

**a pattern of characters that looks like a vertical line**

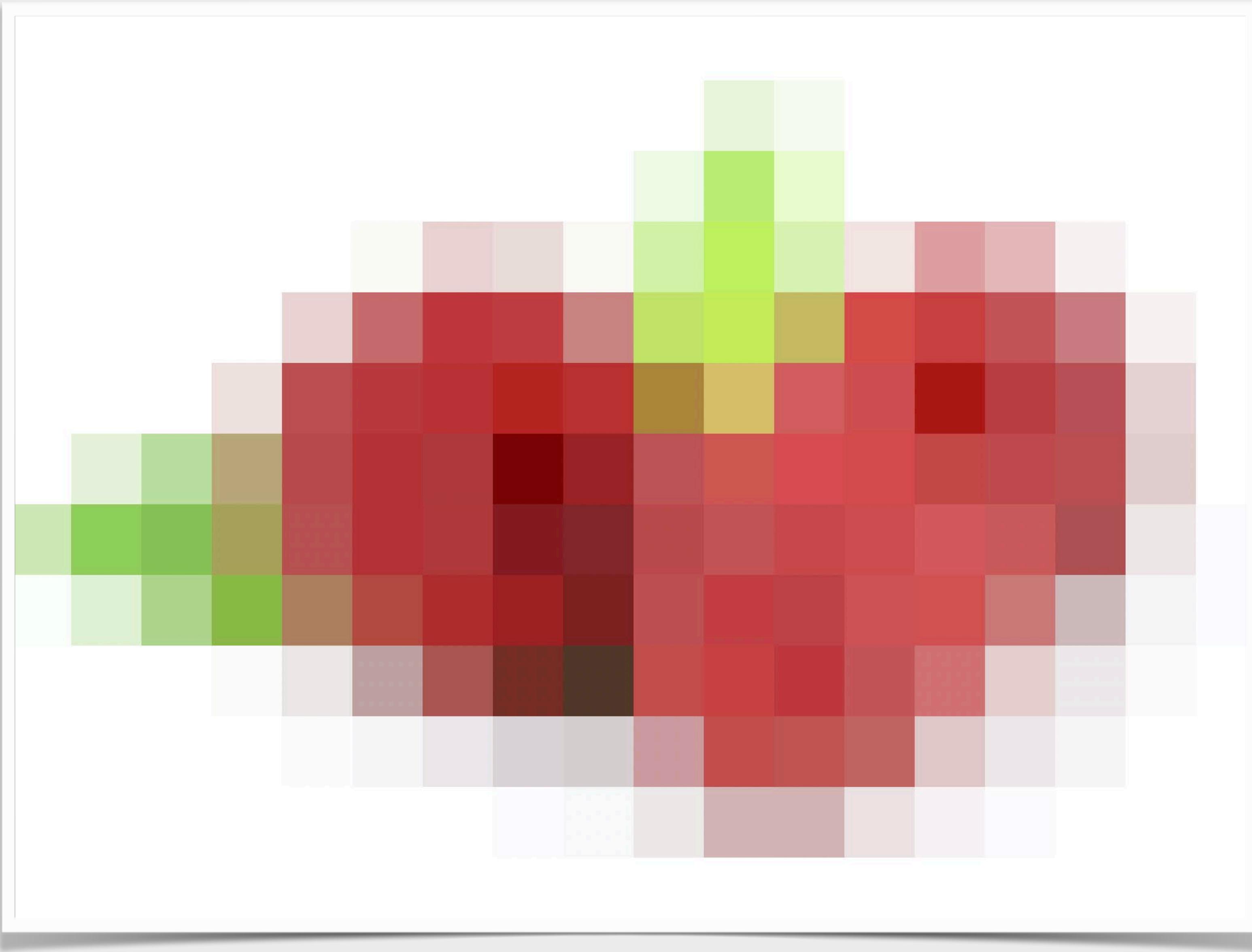
- O •
- O •
- O •
- O •
- O •

a pattern of characters that looks like a triangle

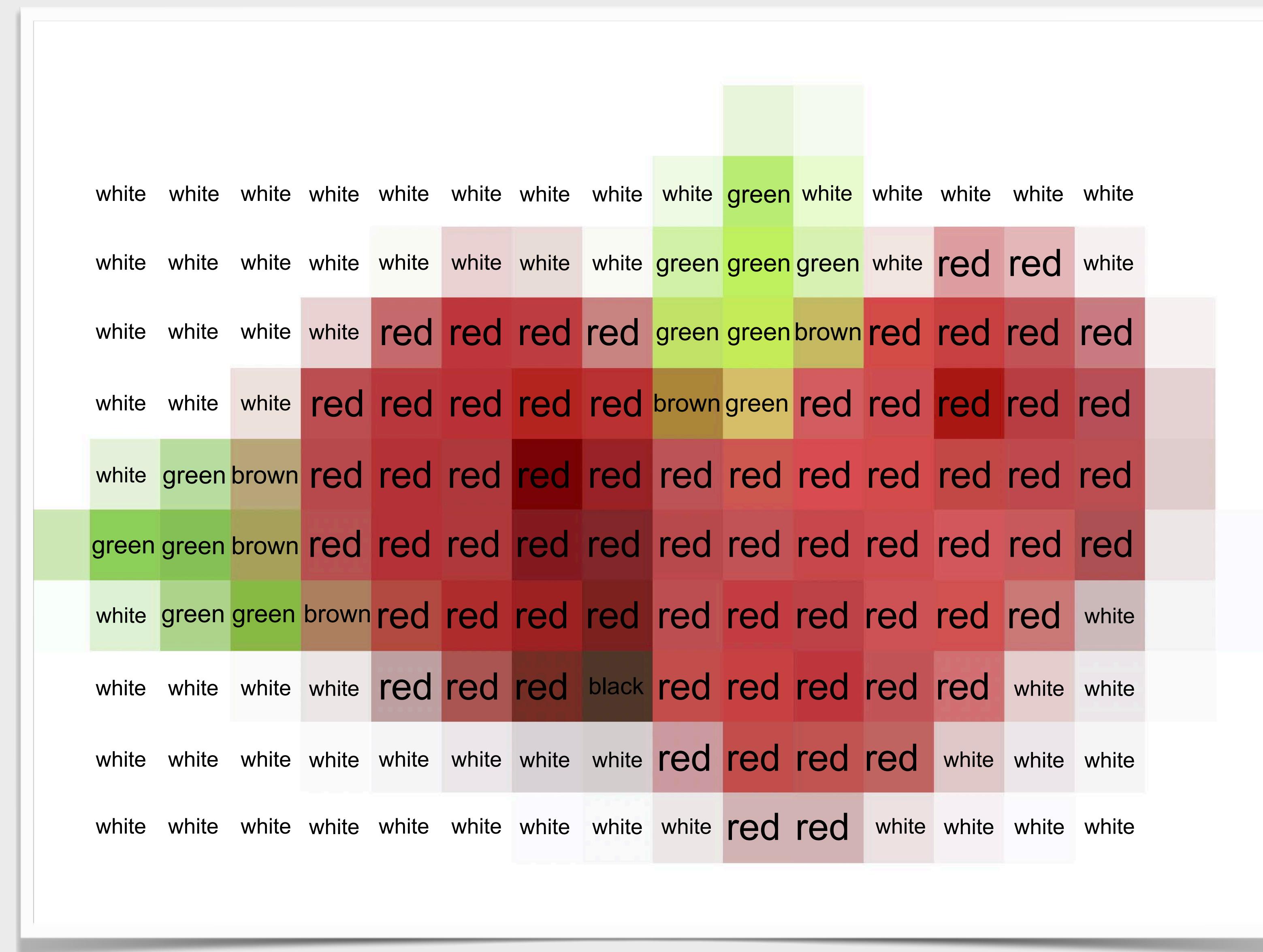
- O    •
- O    O    •
- O    O    O    •
- O    O    O    O    •
- O    O    O    O    O    •



slide from Steve Seitz's [video](#)



slide from Steve Seitz's [video](#)



white white white white white white white white green white white white white white  
white white white white white white white white green green green white red red white  
white white white white red red red red green green brown red red red red  
white white white red red red red red brown green red red red red red  
white green brown red  
green green brown red  
white green green brown red white  
white white white white red red red black red red red red red white white  
white white white white white white white white red red red red white white white  
white white white white white white white white white red red white white white white

(255,0,0)

white white white white white white white white white green white white white white white  
white white white white white white white white green green green white red red white  
white white white white red red red red green green brown red red red red  
white white white red red red red red brown green red red red red red  
white green brown red  
green green brown red  
white green green brown red white  
white white white white red red red black red red red red red white white  
white white white white white white white white red red red red white white white  
white white white white white white white white white red red white white white white

raspberries

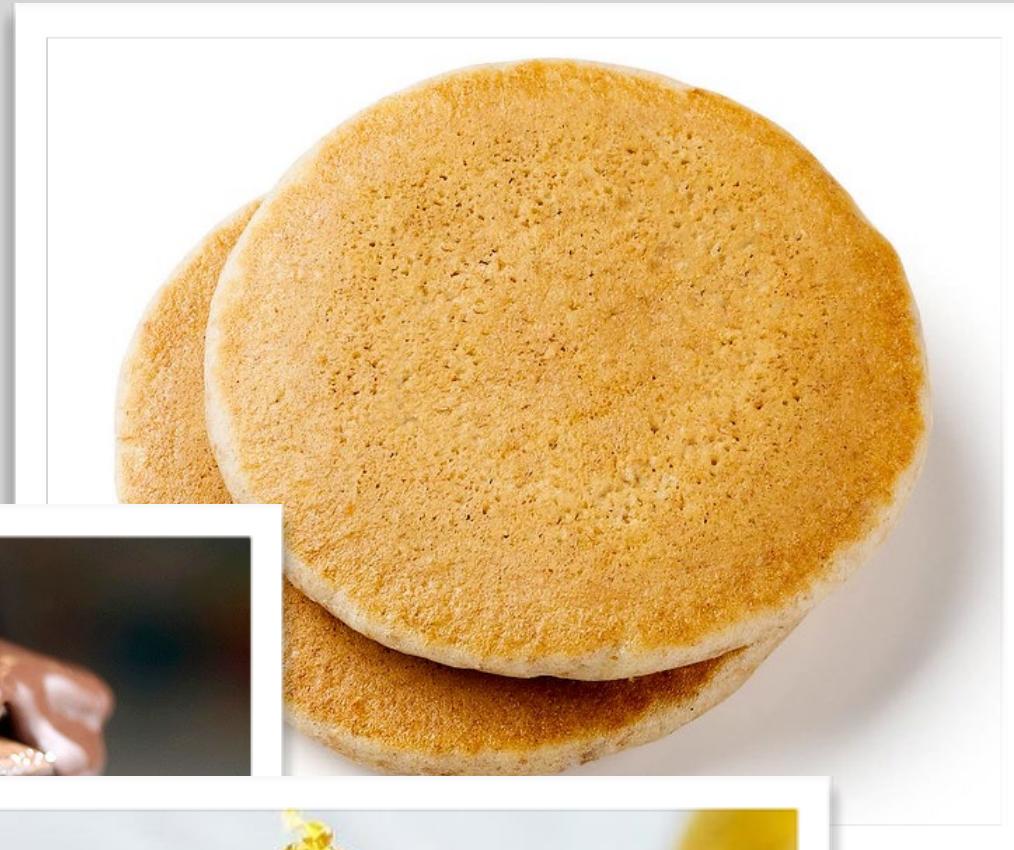
1

Billion

pancakes

sunsets

slide from Steve Seitz's [video](#)



white

Large Language Model

A  
image  
raspberry

slide from Steve Seitz's [video](#)

white

Large Language Model

A image white  
raspberry

slide from Steve Seitz's [video](#)

red

large Language Model

A      image    white    white  
raspberry

slide from Steve Seitz's [video](#)

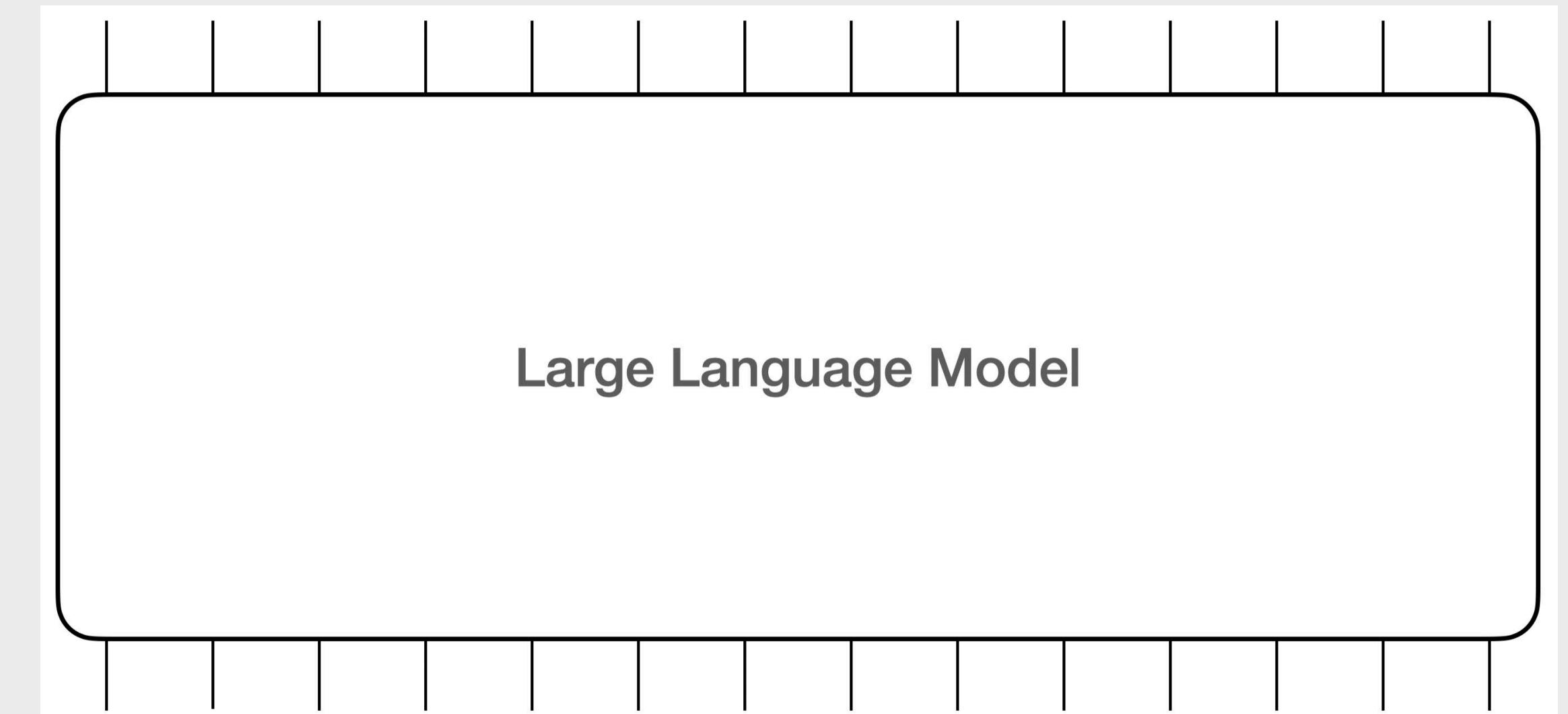
# Large Language Model

A image white white red red red white white green green green white  
raspberry

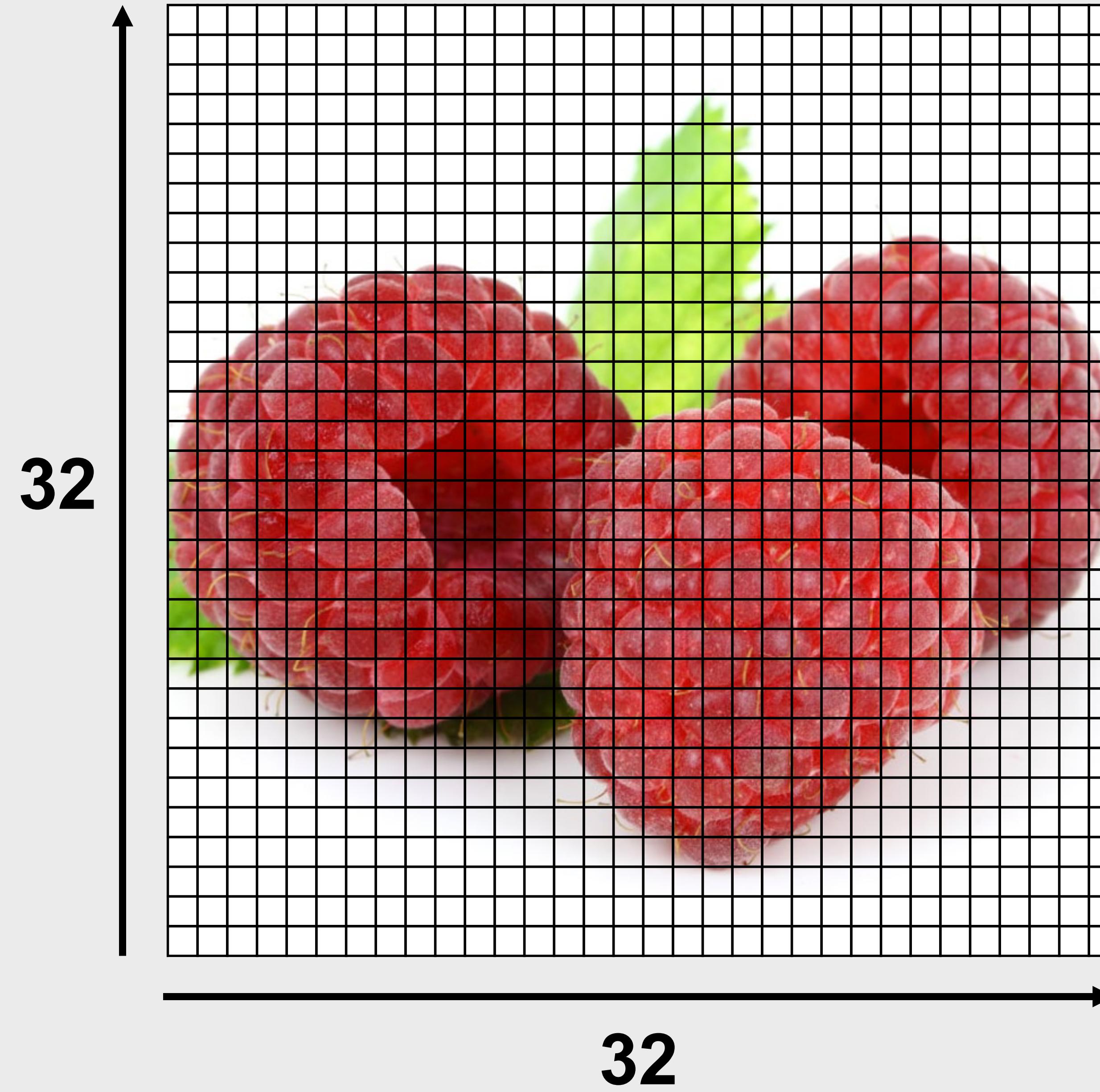
slide from Steve Seitz's [video](#)



**1,000,000s of pixels**

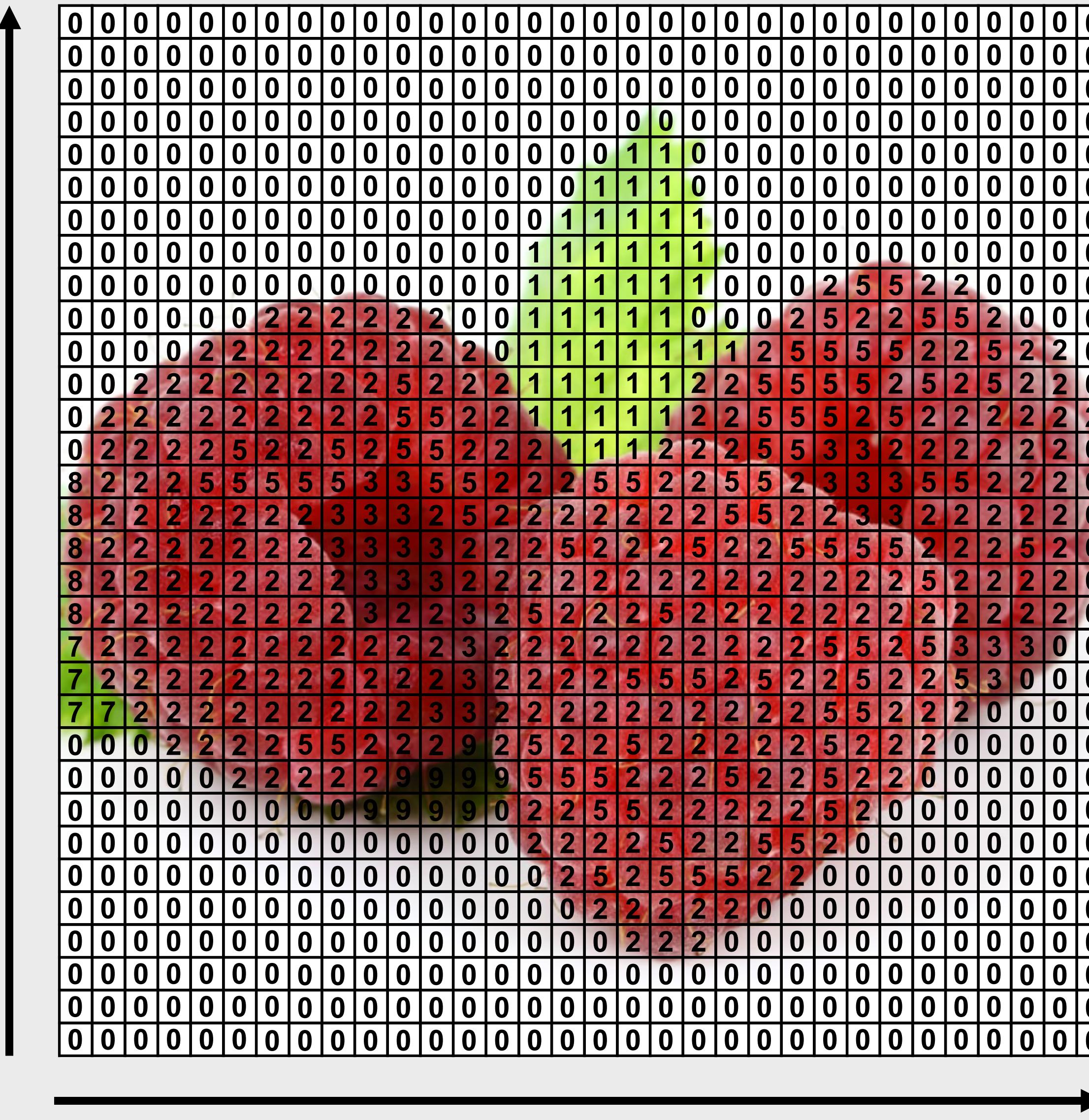


**1,000s of words**



slide from Steve Seitz's [video](#)

32



32

slide from Steve Seitz's [video](#)

$$32 \times 32 = 1024$$

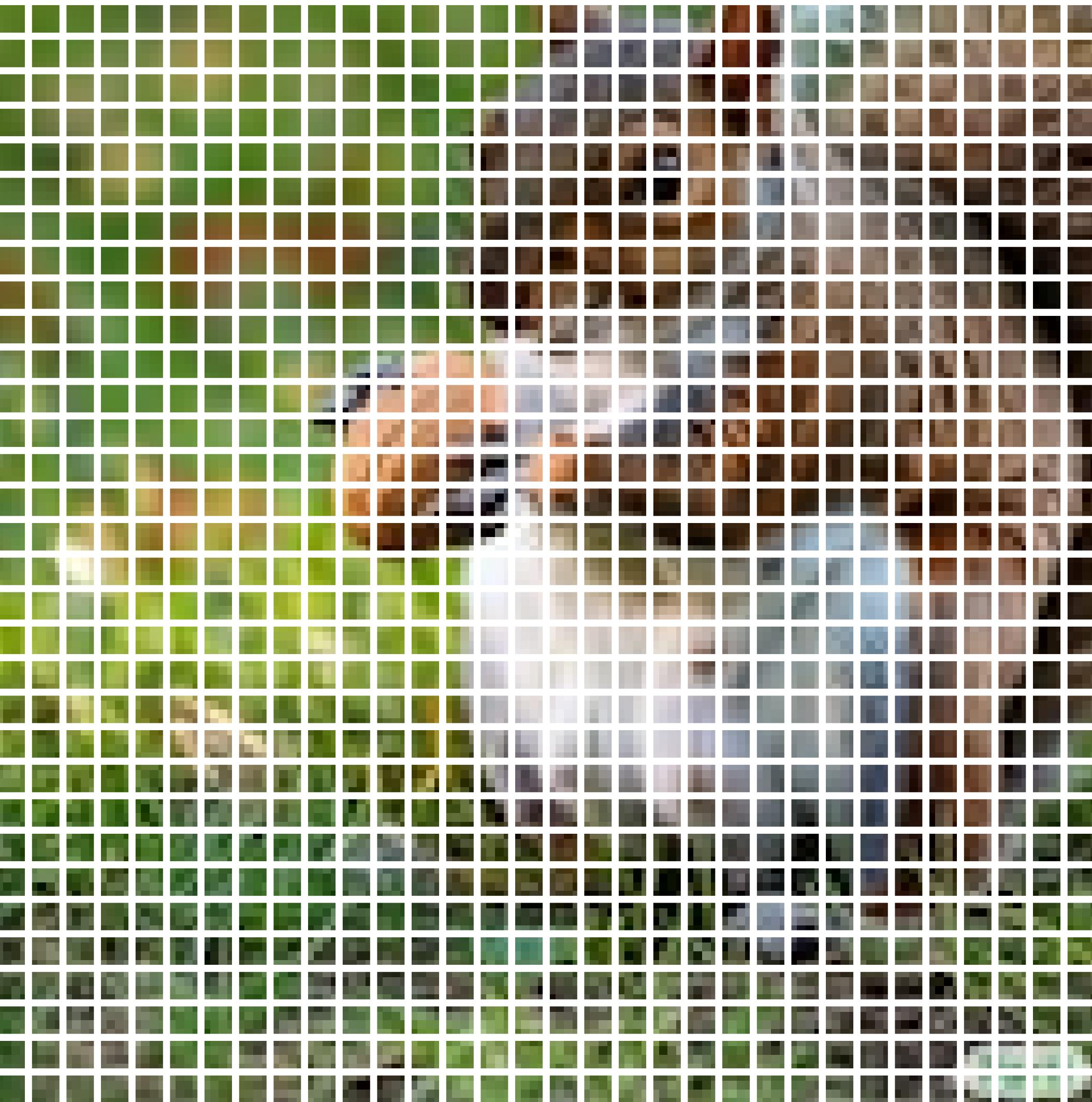
# Visual words

# squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9  
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1 1 6

# squirrel reaching for a nut

# squirrel reaching for a nut



squirrel reaching for a nut

slide from Steve Seitz's [video](#)



squirrel reaching for a nut

slide from Steve Seitz's [video](#)

**Up-sampled  
4x**



squirrel reaching for a nut

slide from Steve Seitz's [video](#)



Parti



Parti



Parti



Parti



Parti

squirrel reaching for a nut

Parti, <https://parti.research.google/>



Parti



Parti

squirrel reaching for a nut underwater

slide from Steve Seitz's [video](#)



Parti



Parti

fossil of a squirrel reaching for a nut

slide from Steve Seitz's [video](#)



Parti



Parti

squirrel made of toothpicks wearing sunglasses reaching for a nut

slide from Steve Seitz's [video](#)



Parti



Parti

DLSR photograph of a whimsical fantasy house shaped like a squirrel  
with windows and a door, in the forest

slide from Steve Seitz's [video](#)



Parti



Parti

Squirrel reaching for a nut. by Leonardo da Vinci

slide from Steve Seitz's [video](#)



Squirrel reaching for a nut. Van Gogh painting

slide from Steve Seitz's [video](#)



Part I



Part II

Intricately carved cathedral door of a squirrel reaching for a nut

slide from Steve Seitz's [video](#)



Parti



Parti

Squirrel reaching for a nut. Woodcut tessellation pattern by M.C. Escher

slide from Steve Seitz's [video](#)



Parti



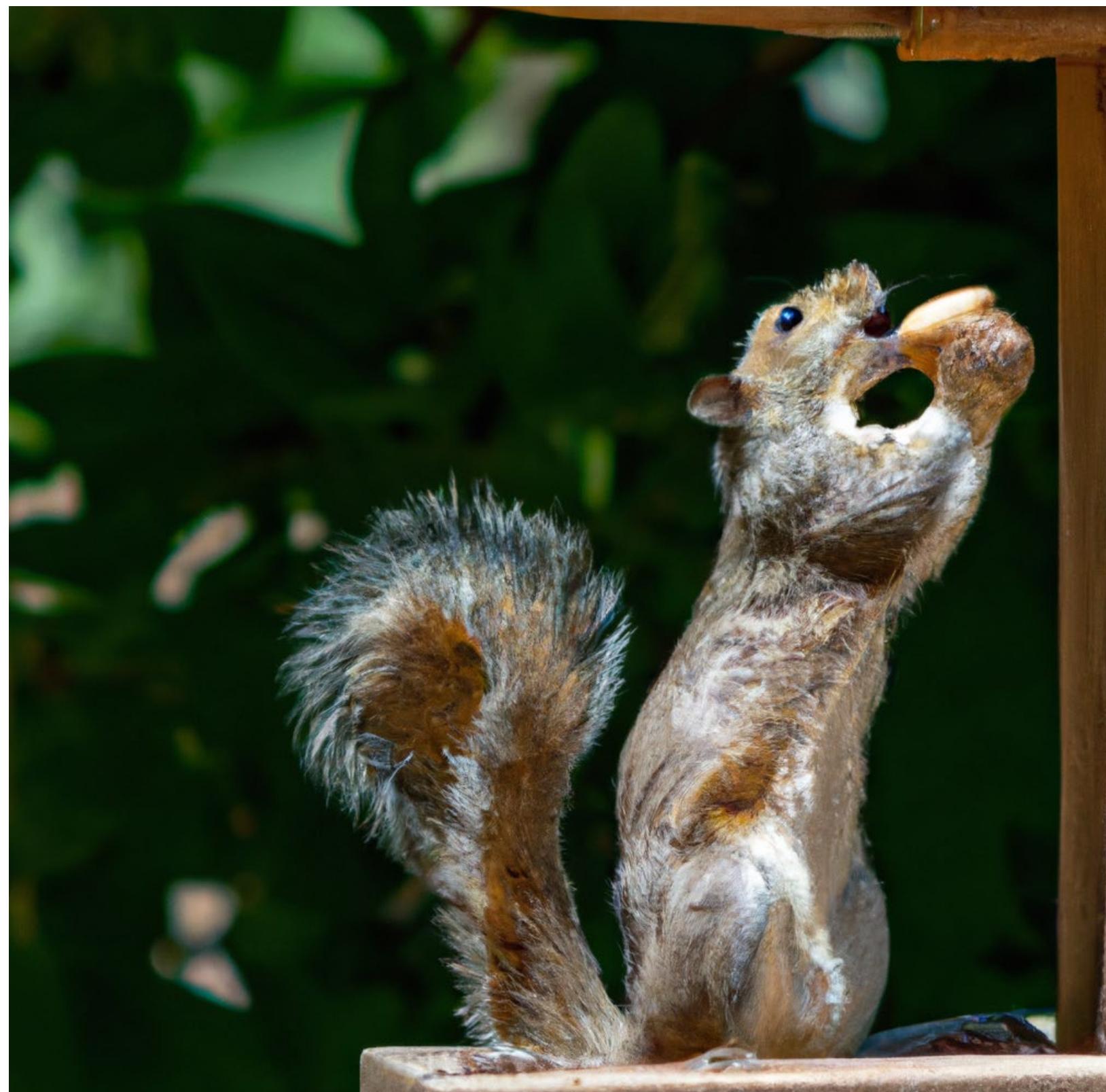
Parti

Squirrel reaching for a nut. Latte art

slide from Steve Seitz's [video](#)

# Algorithm vs. Data

Diffusion-based



Auto-regressive

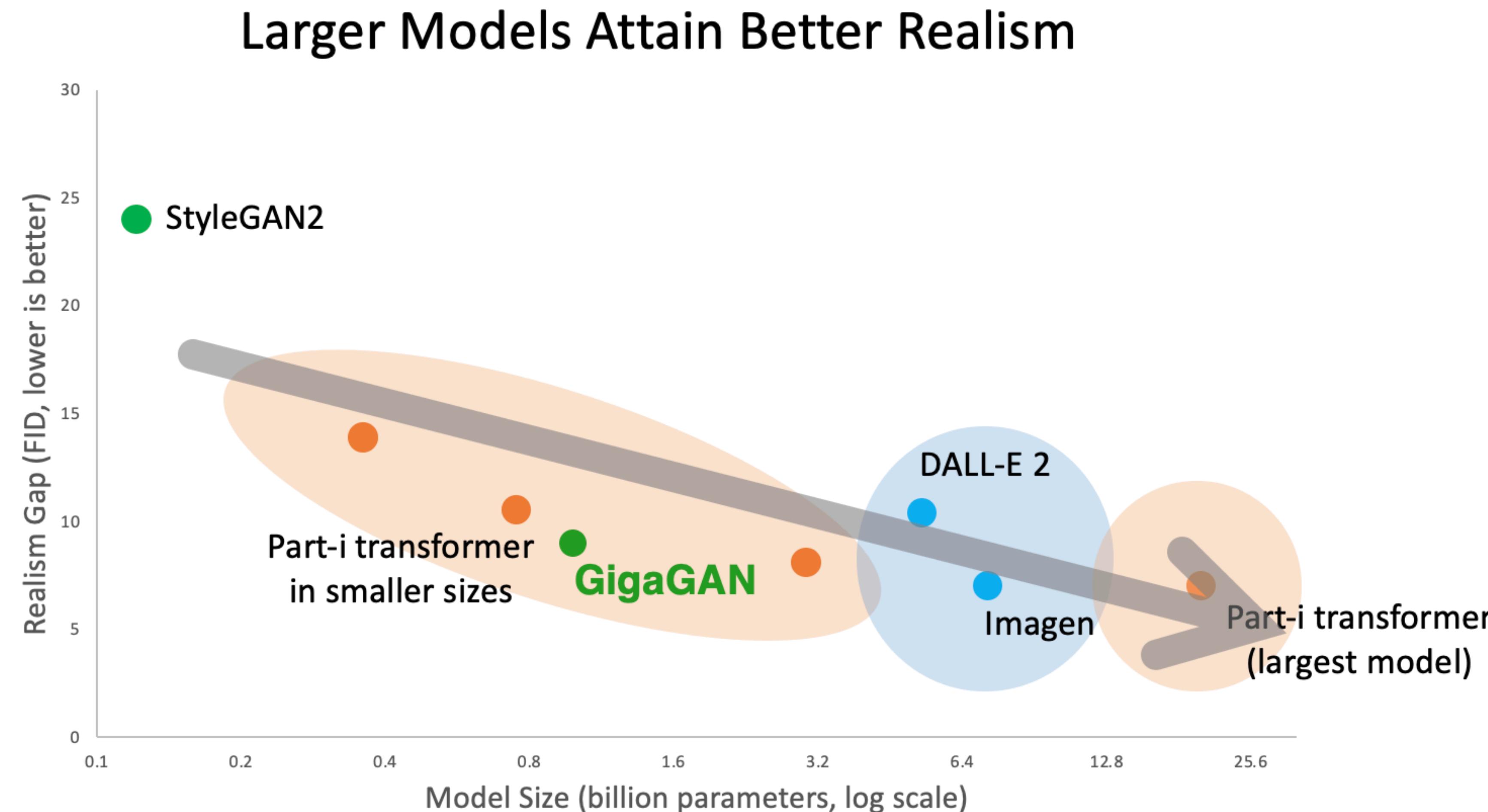


GAN-based



Prompt: “*squirrel reaching for a nut*”

# data capacity vs. image quality



Graph by Taesung Park

# Generative Magic

Prompt: “*a green creature made of leaves and vines bursting out of the ground ready to attack; detailed, best on artstation, raymond swanland, magic the gathering, epic, stunning, masterpiece*”

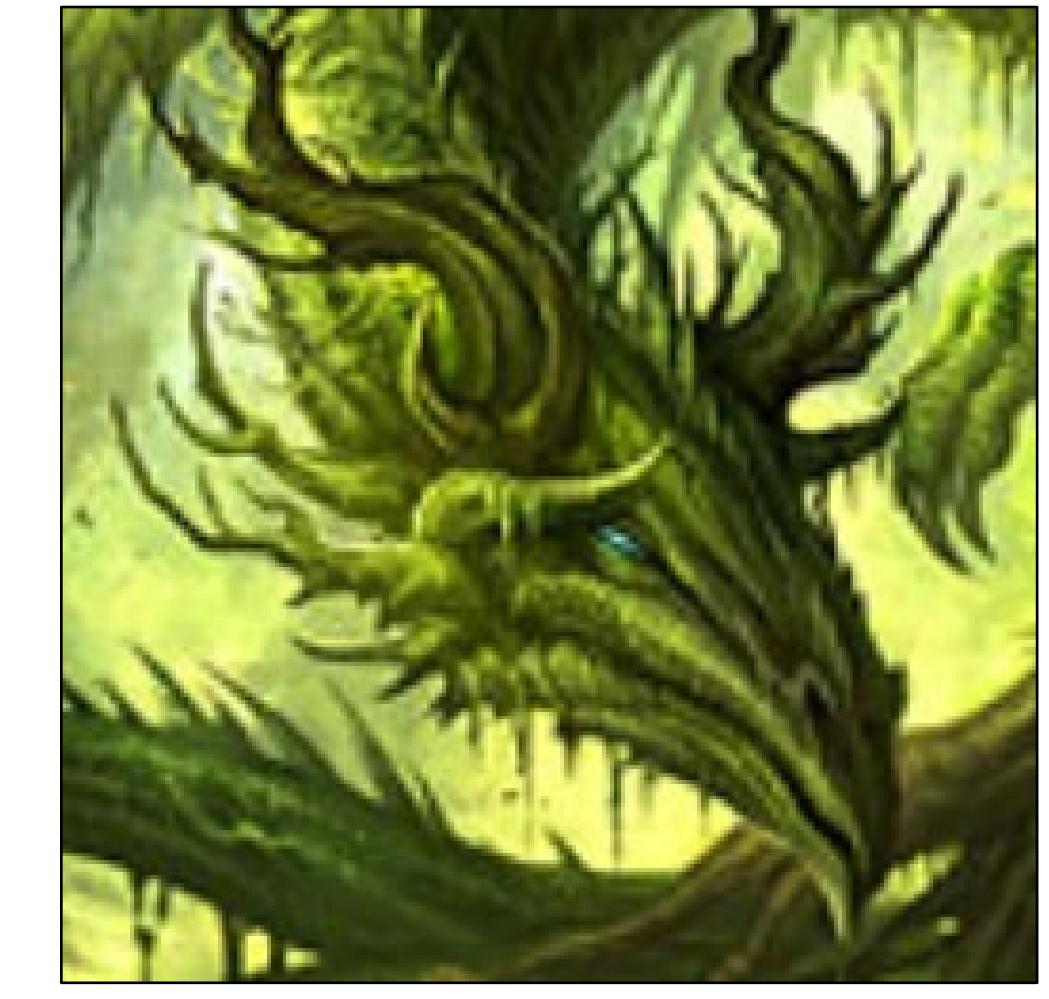
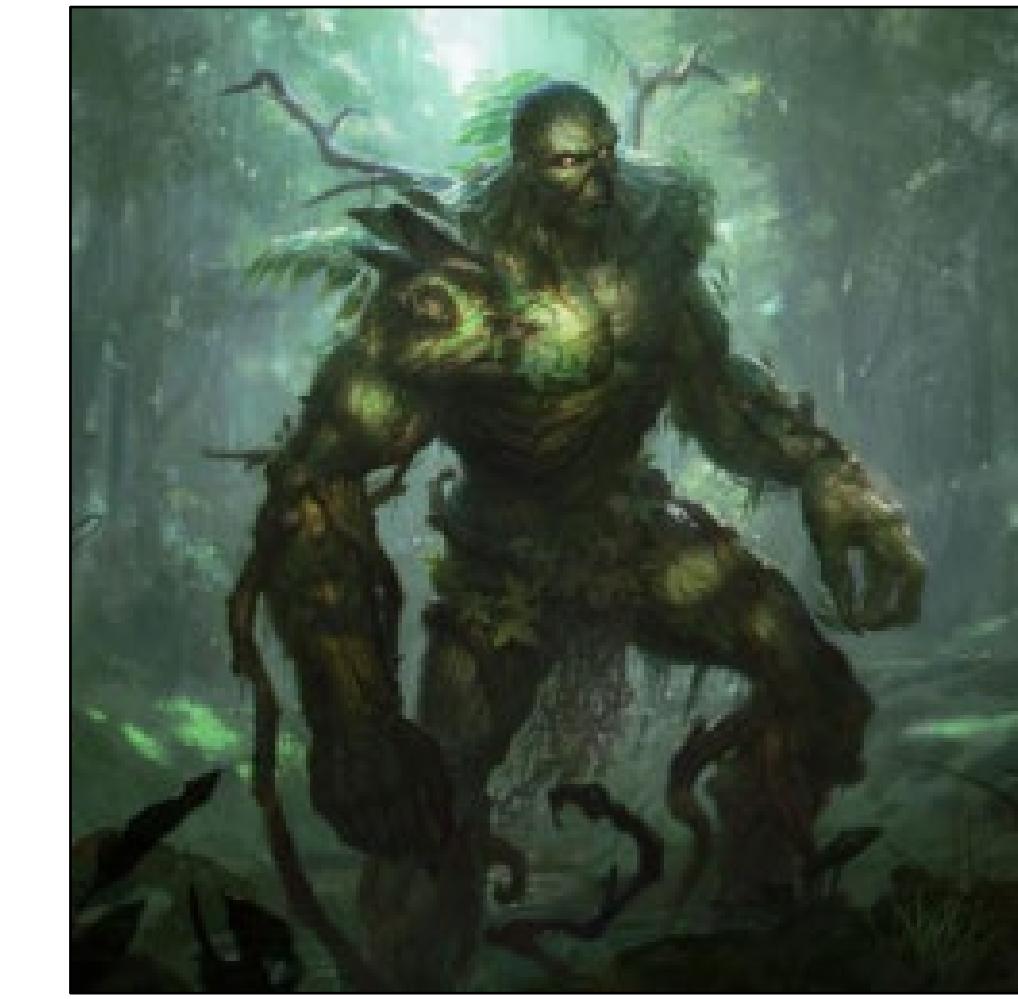
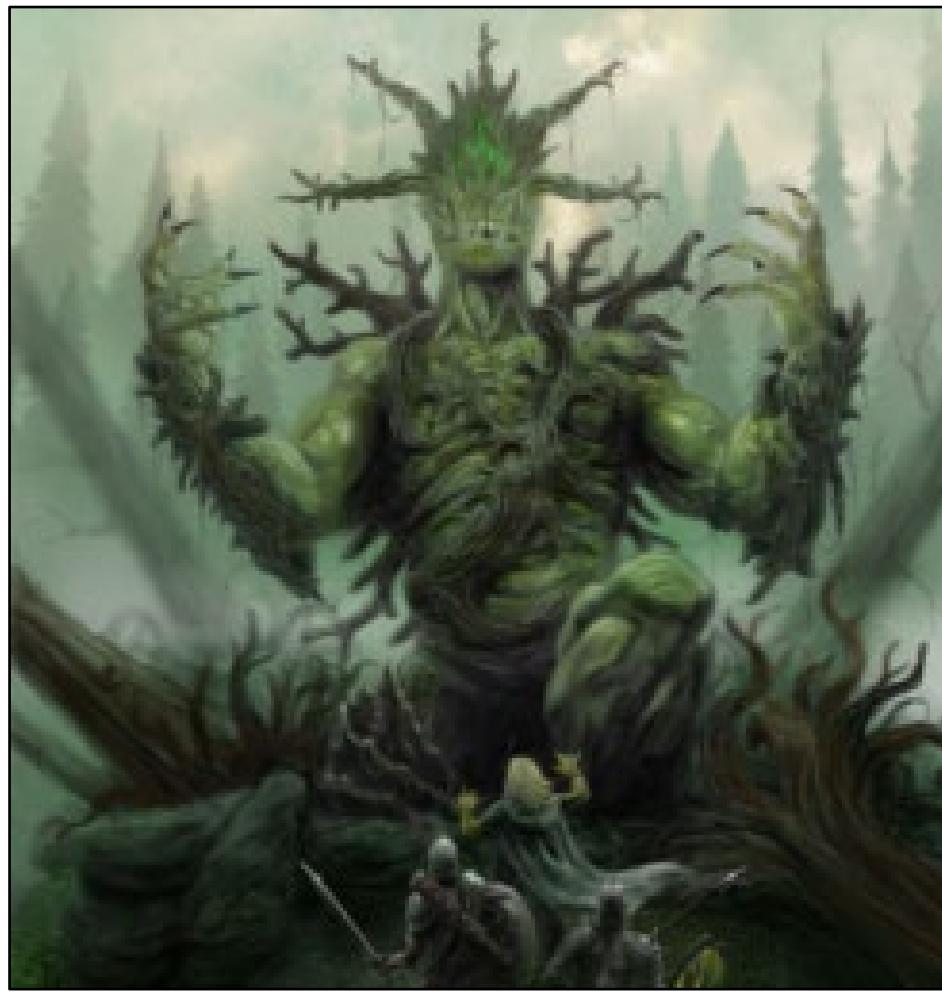
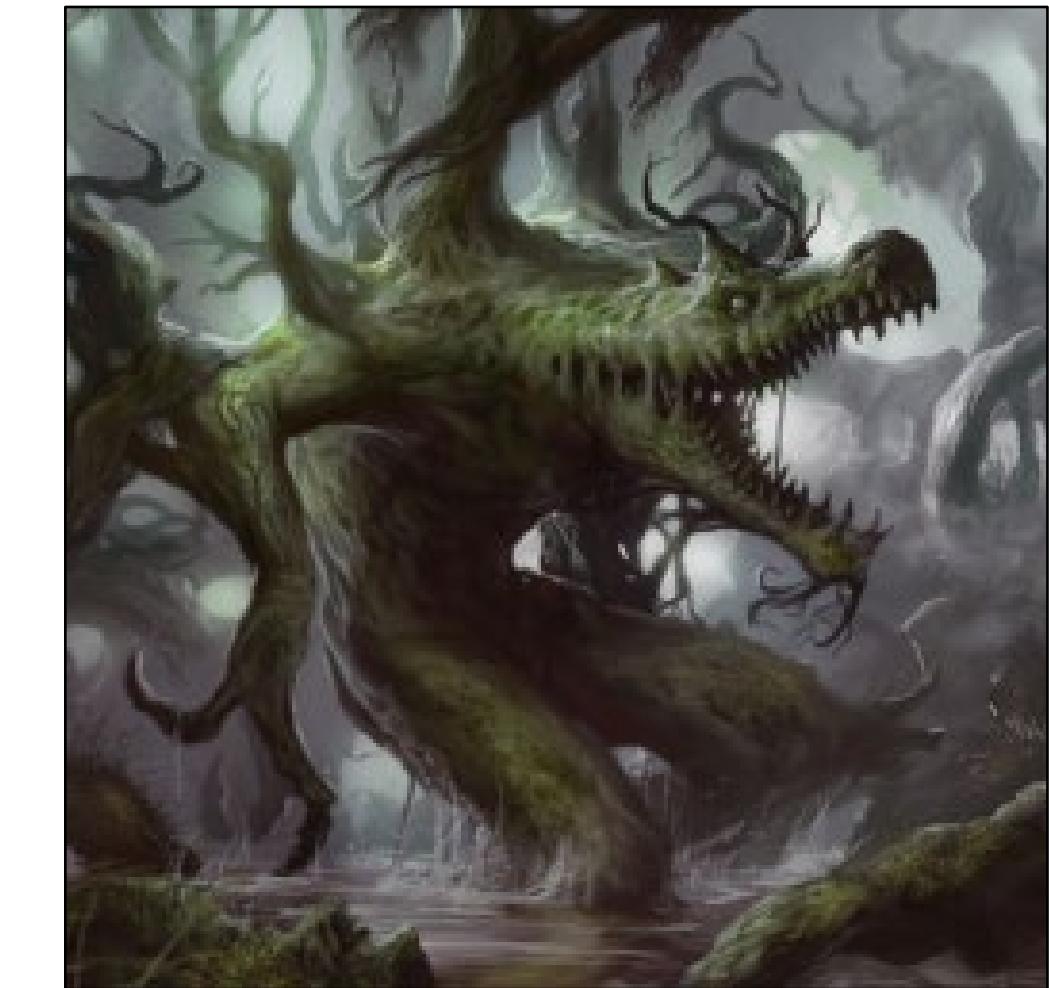
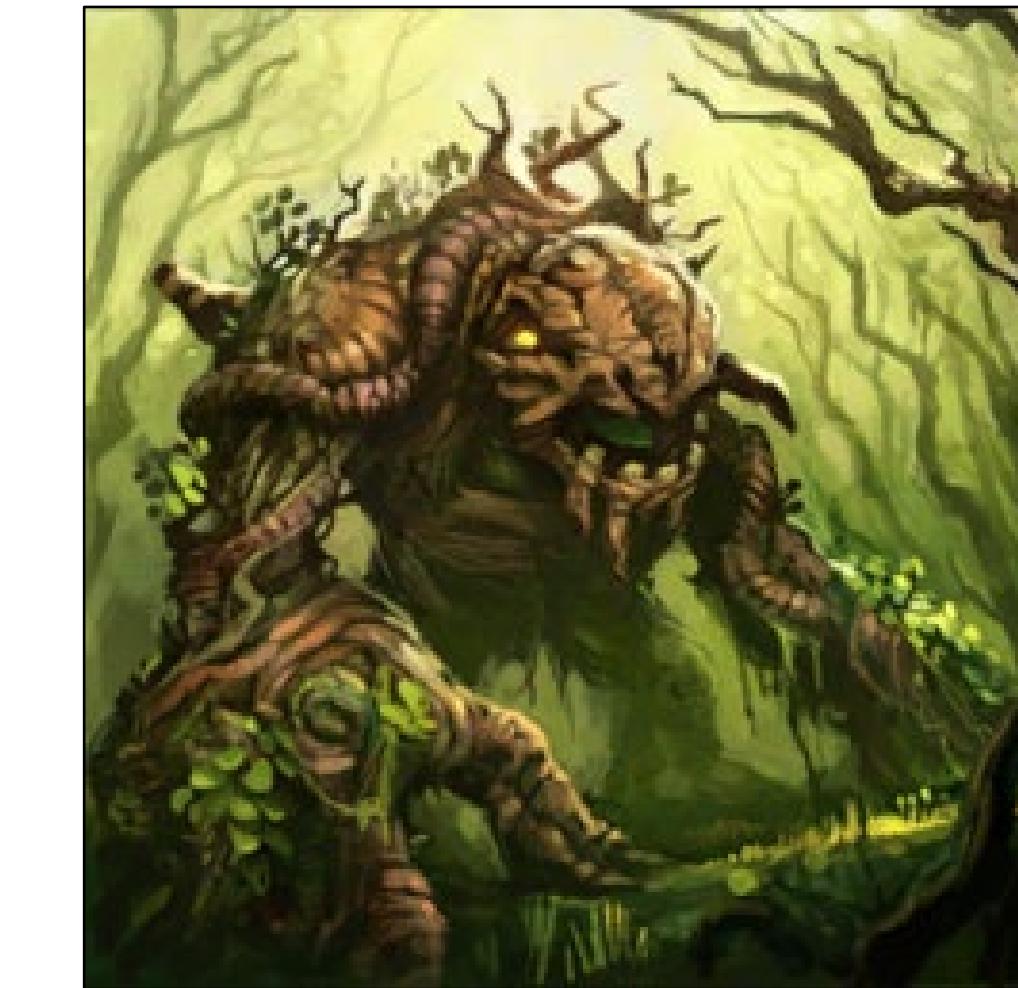
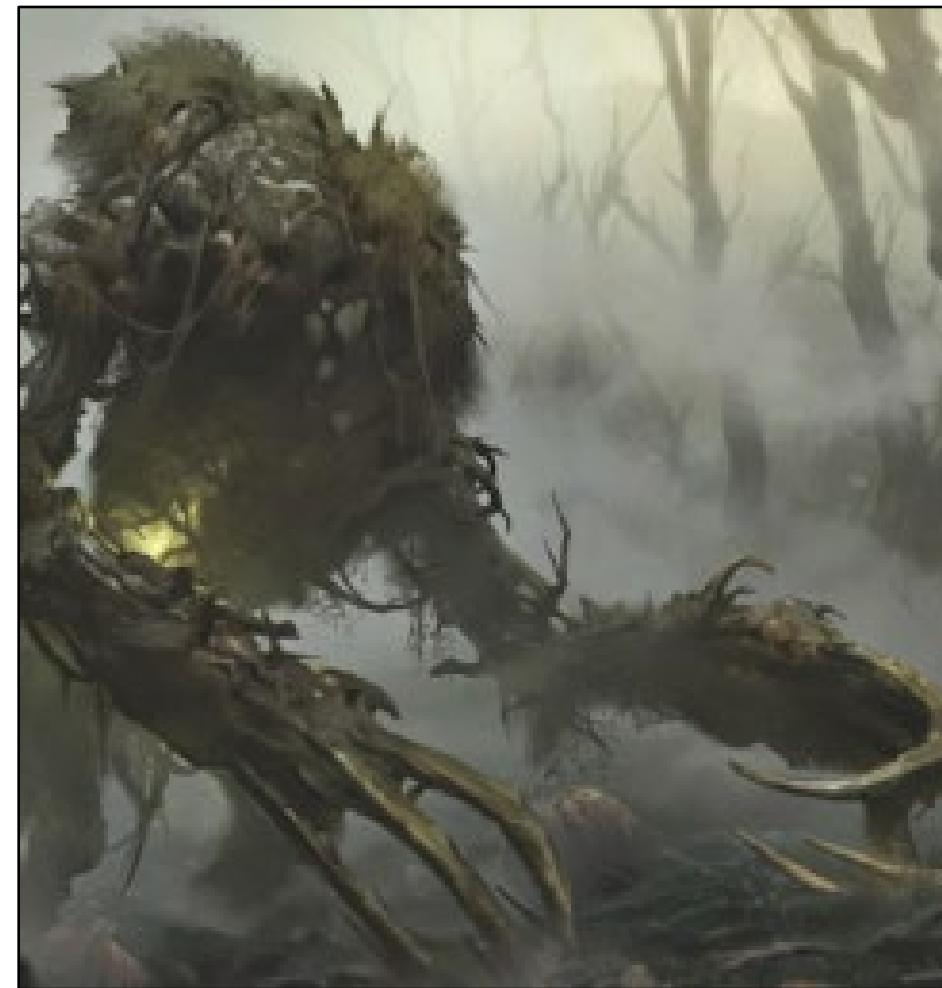


# “influences” from the training data



generated image

-----



to paraphrase Arthur C. Clarke:

*Interpolation in sufficiently  
high-dimensional space is  
indistinguishable from magic*