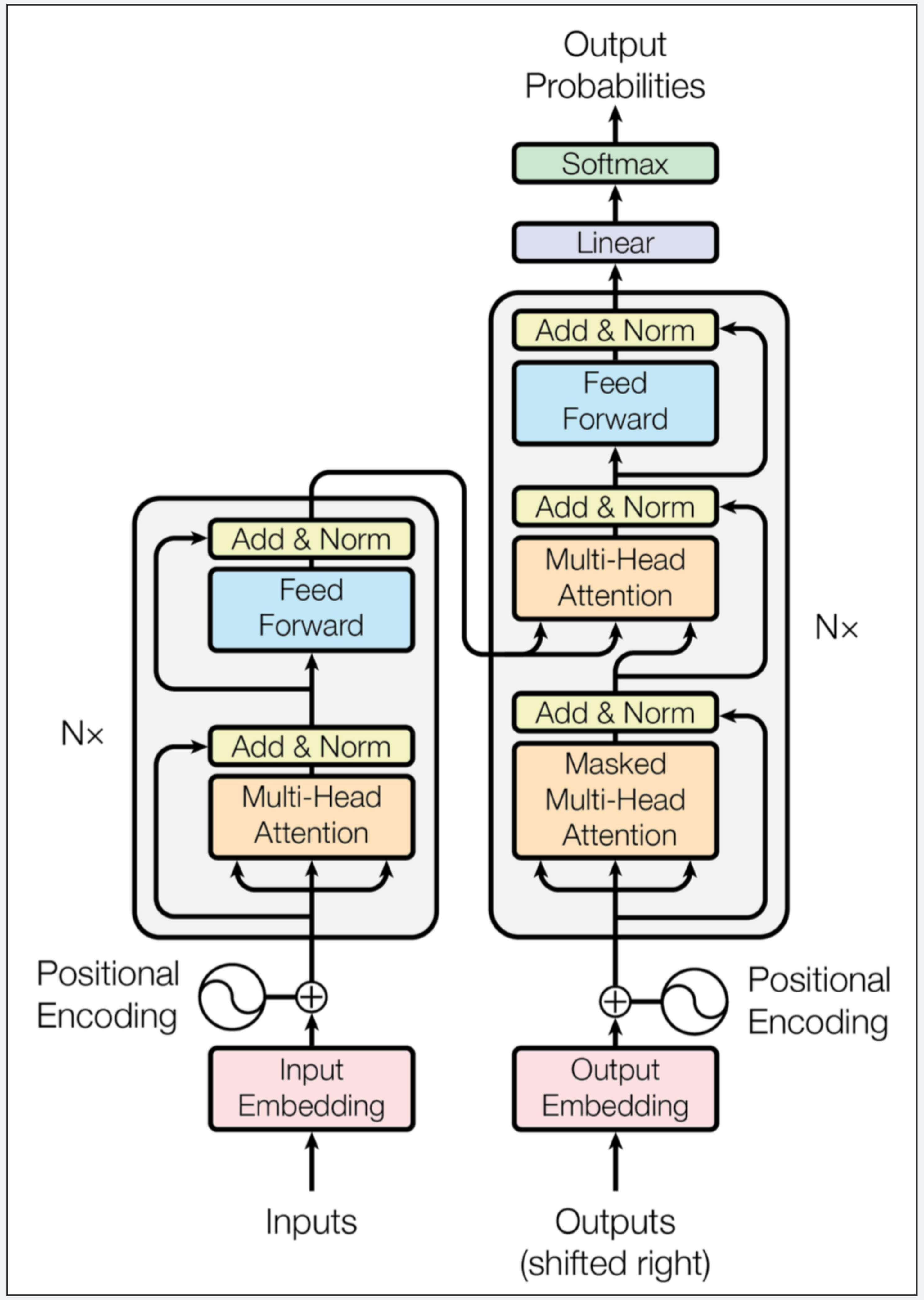
Cyclic

Learning Linear Groups in Neural Networks

Friday, September 8, 2023

Emmanouil Theodosis
etheodosis@g.harvard.edu
manosth.github.io/
github.com/manosth/

Deep learning is empirical…



Transformer

Problems

- How to design?
- Functional understanding?
- Interpretability?

… and not very efficient

ChatGPT

- GPT-1 : 117M
- GPT-2 : 1.5B
- GPT-3 : 175B
- **GPT-4: 170T**

https://chat.openai.com

- LLaMA-2 7B : 7B
- LLaMA-2 13B : 13B
- **LLaMA-2 70B: 70B**

https://ai.meta.com/llama/

- LaMDA: 137B
- **PaLM : 540B**

https://bard.google.com

- **Claude 2: 175B**

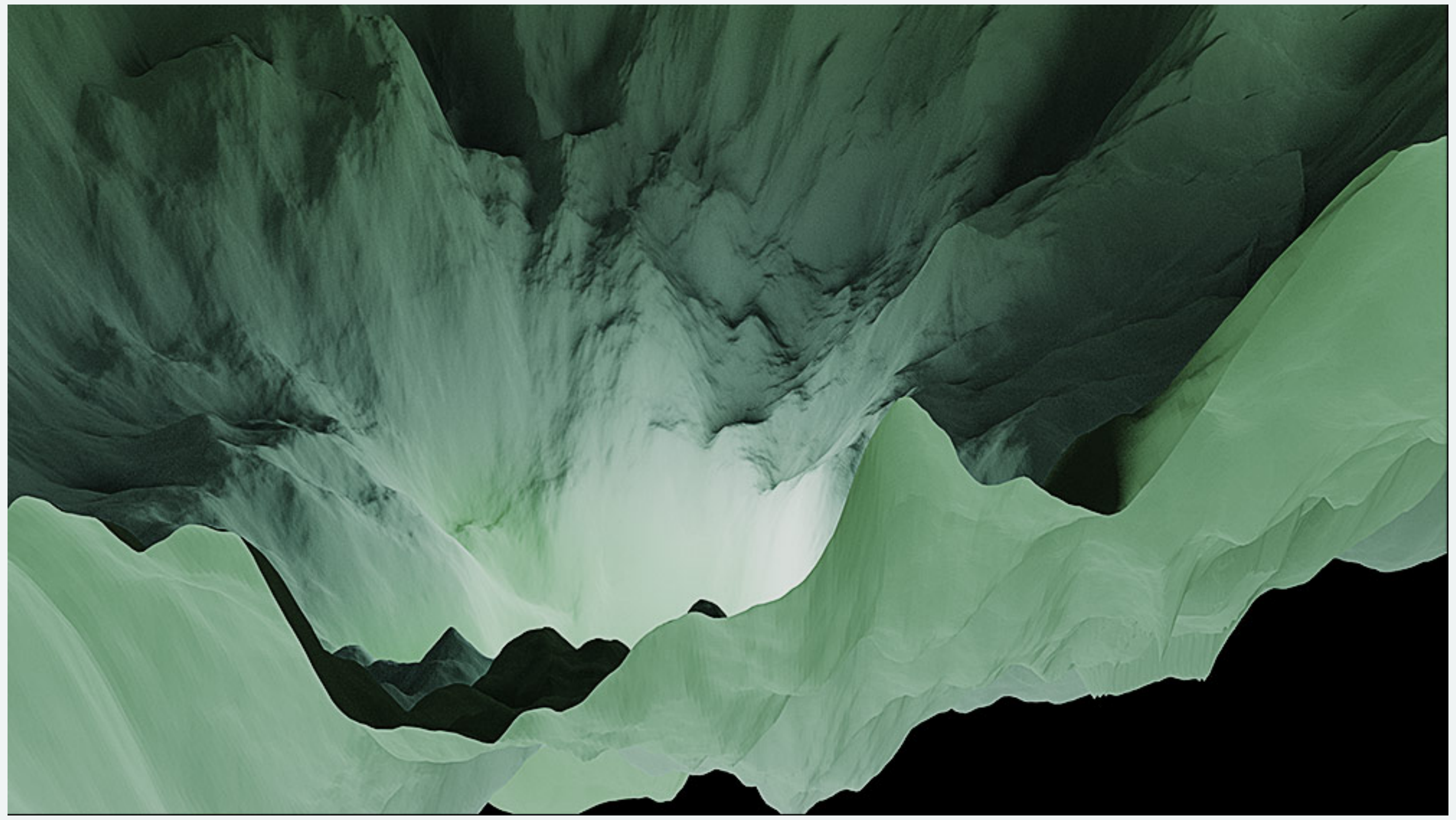
https://claude.ai/

So?

Overparametrized models

- more parameters than data
- neurons memorize data points
- the rest interpolate

`result:` hope for the best.



Use domain knowledge

- in a systematic way
- constrain the parameter space

`result:` similar (or better!) performance and more efficient.

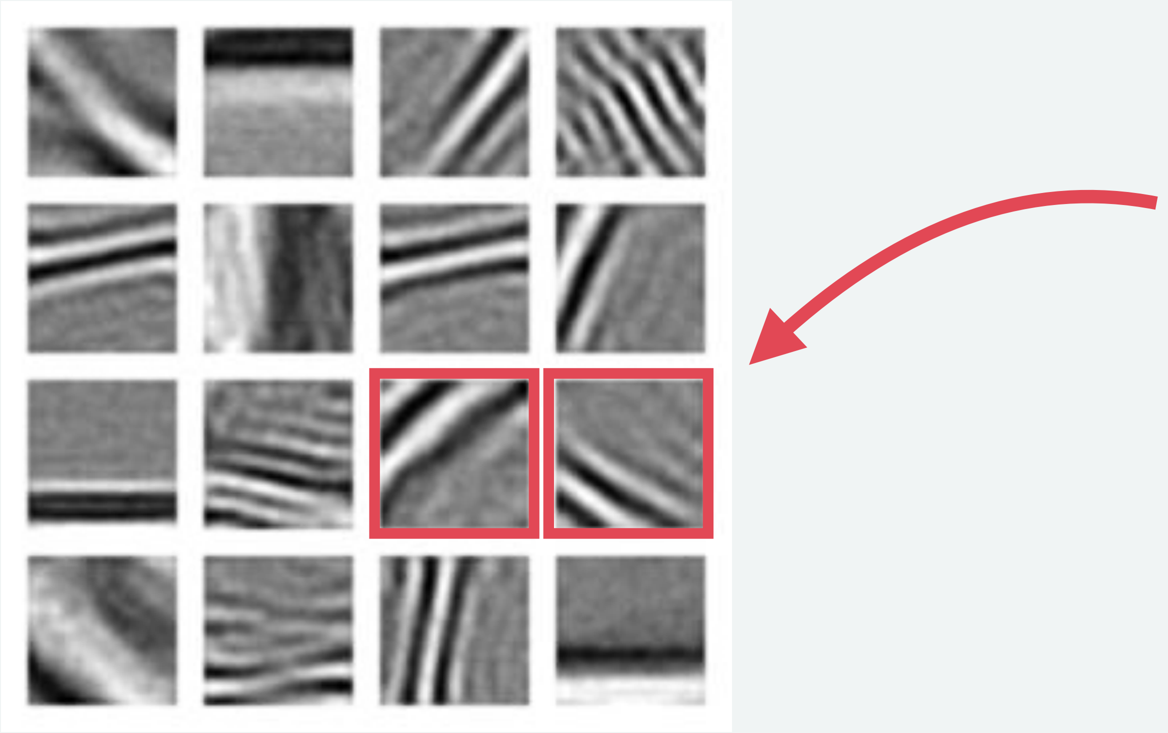
Enter equivariance

Classification has invariants

$=$ Person on a chair

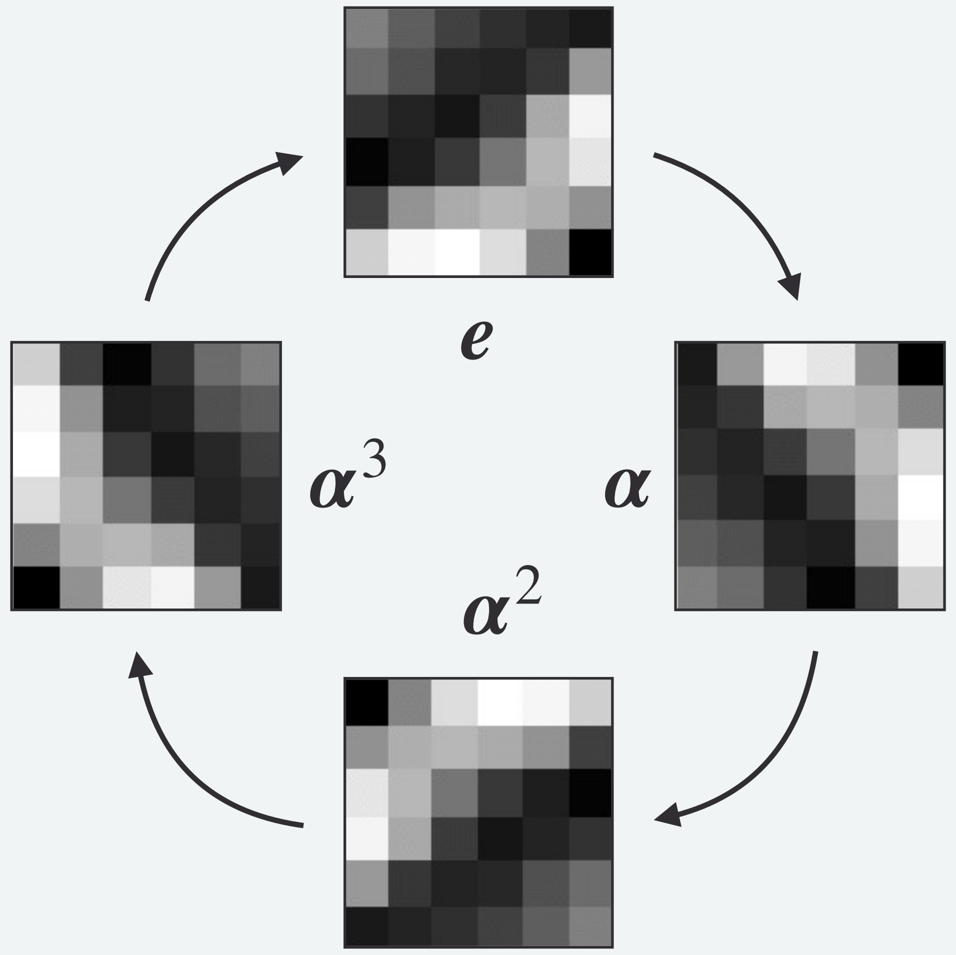
$=$ Person on a chair

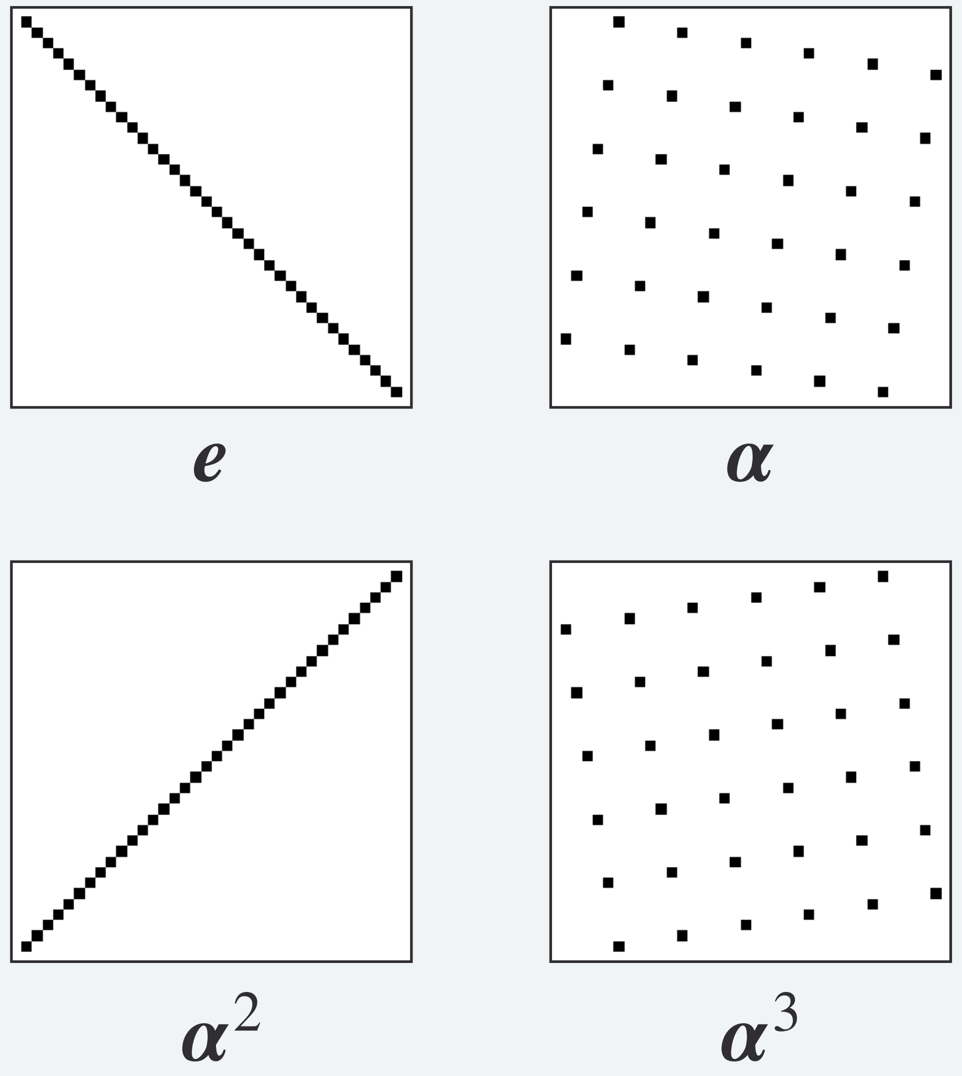
Filters have symmetries



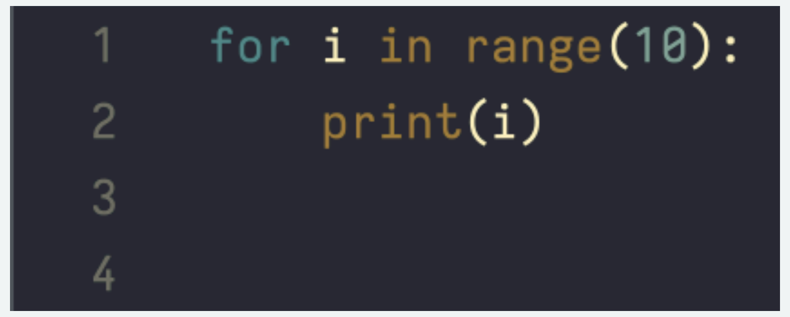
different orientations

Group equivariant CNNs

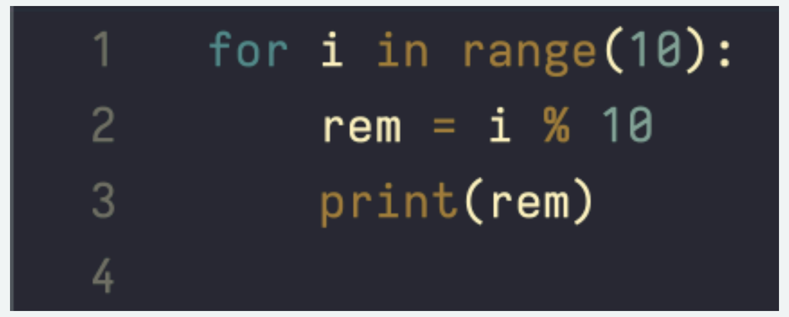




How do we generalize?

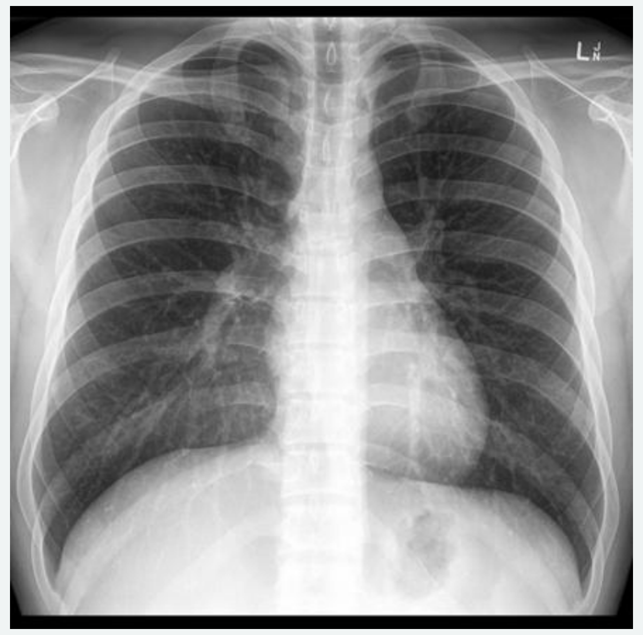


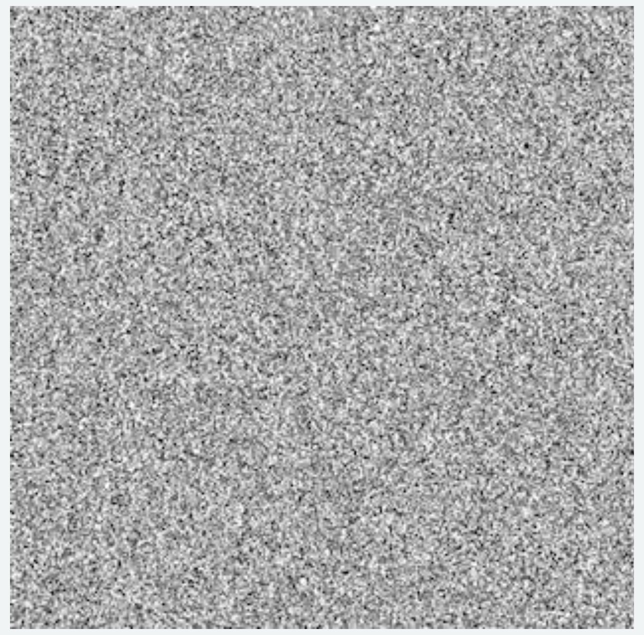
```python
for i in range(10):
    print(i)
```



```python
for i in range(10):
    rem = i % 10
    print(rem)
```

code





privacy

GCT

GCG

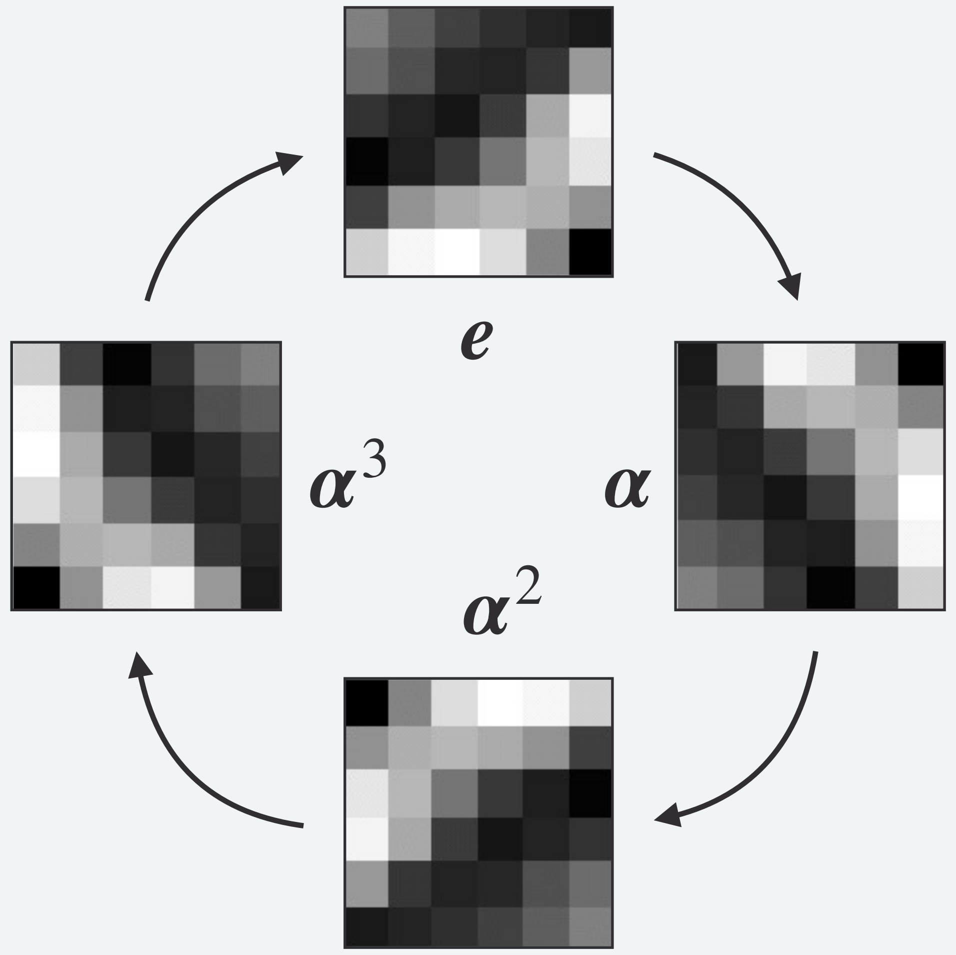
Alanine

···AATGCTACT···

···AATGCGACT···

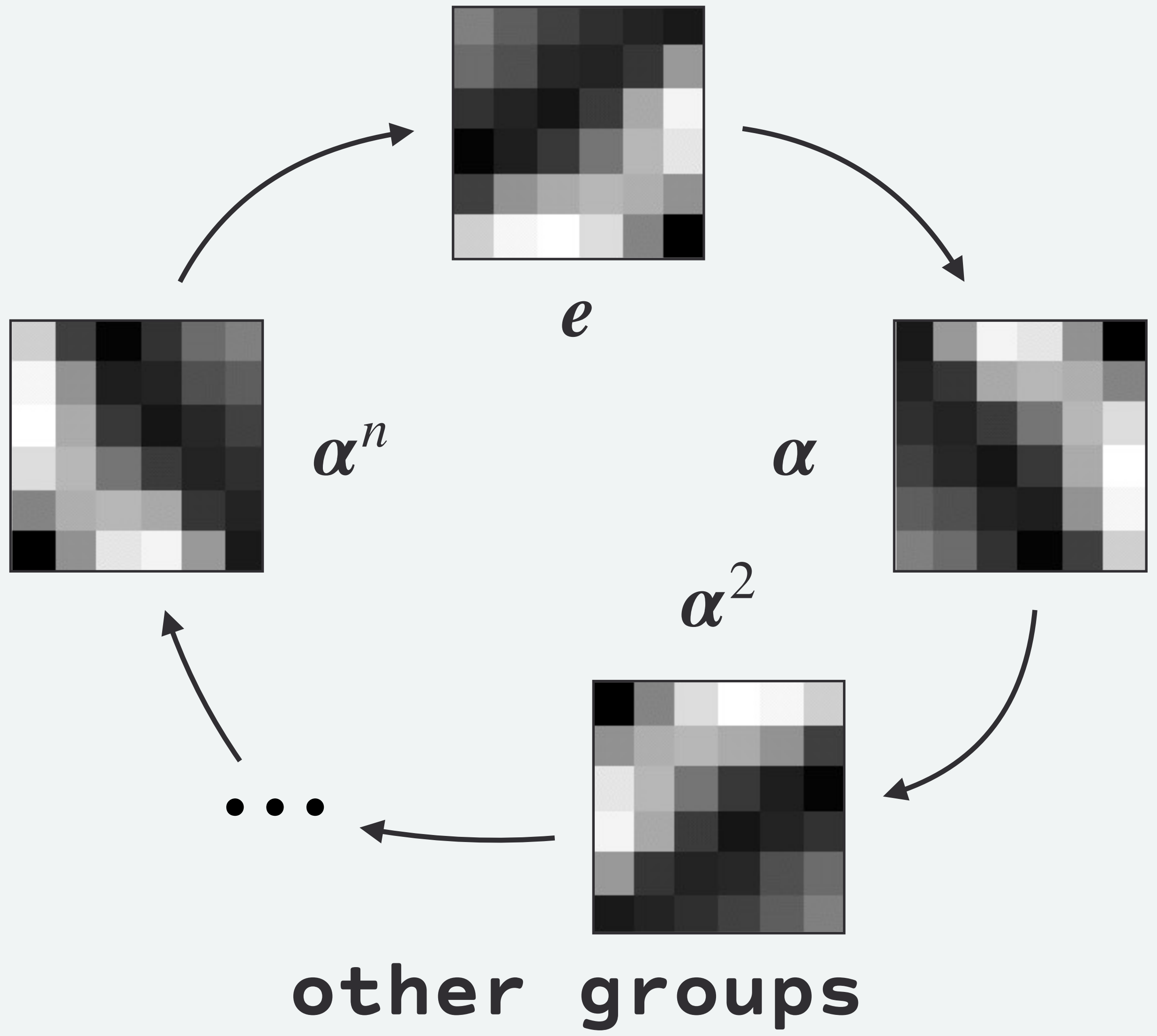
biology

Our work

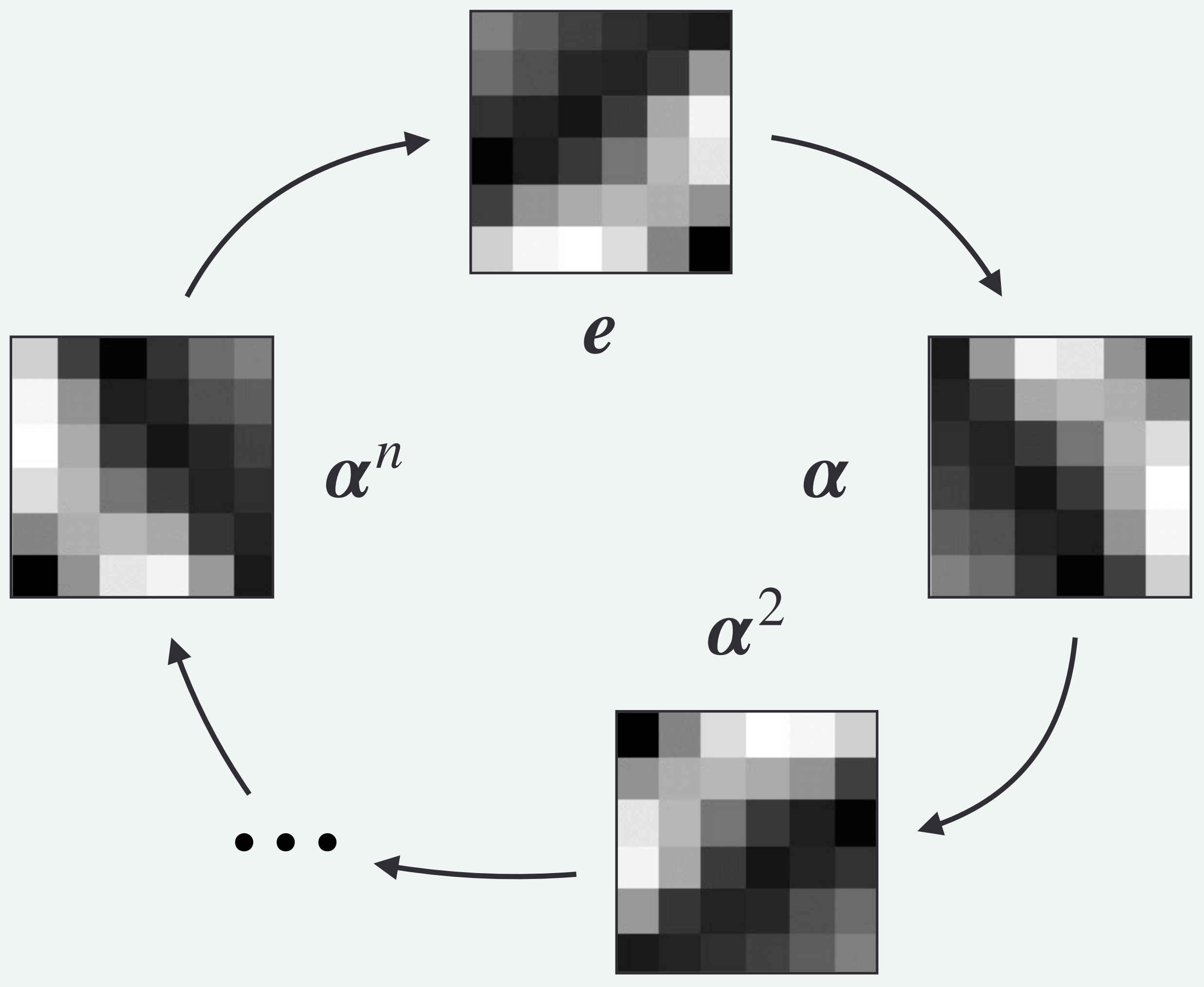


other groups

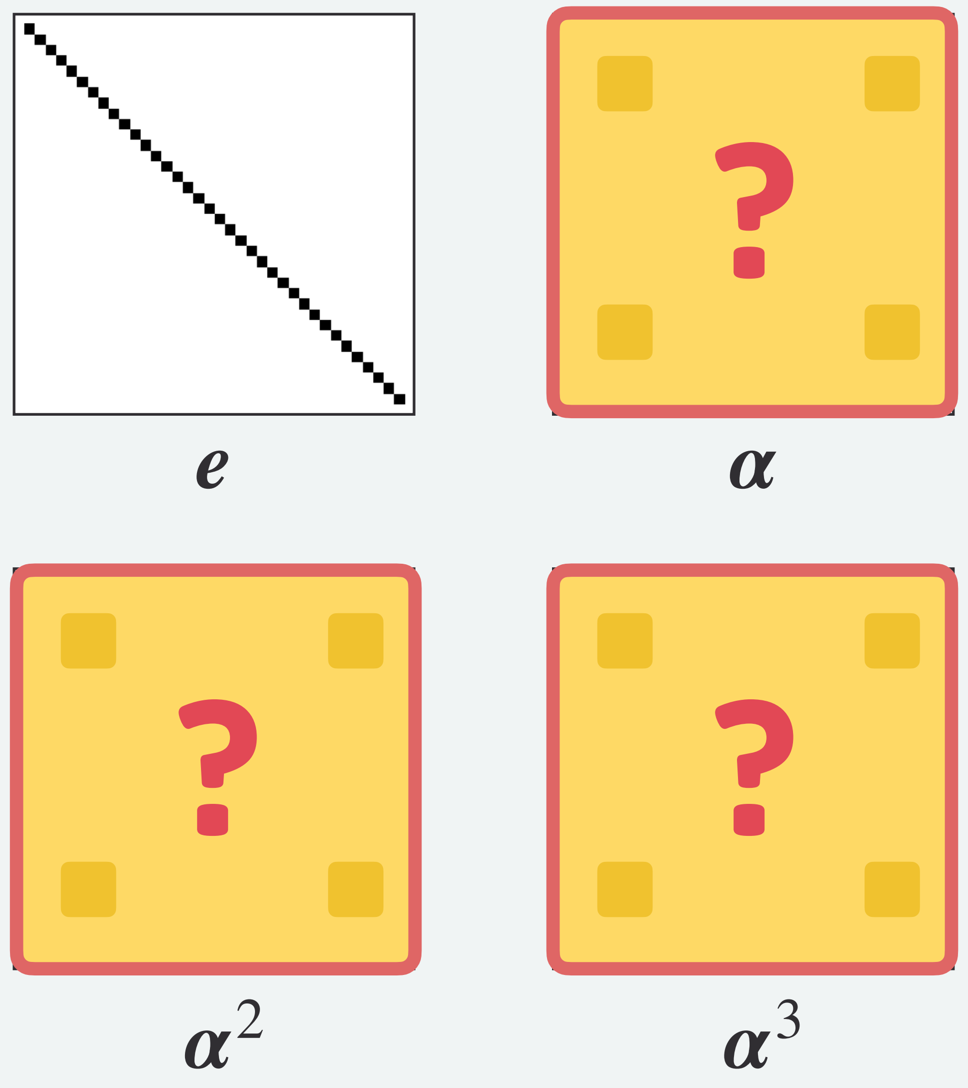
Our work



Our work

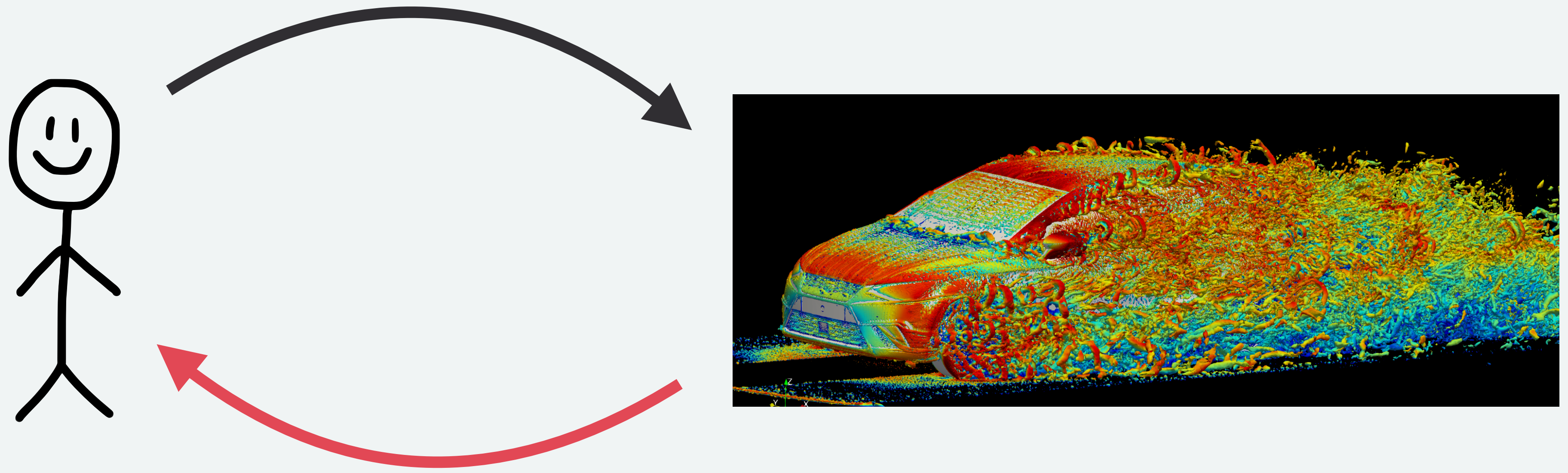


other groups

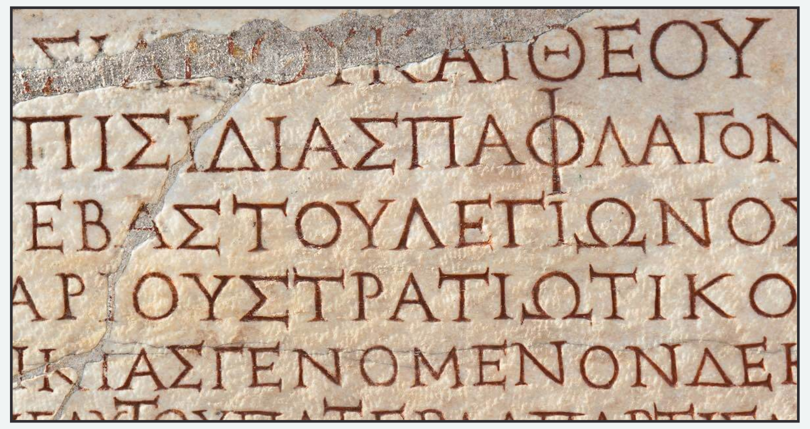


other actions

Benefits



bidirectional information flow



ΚΑΙΘΕΟΥ
ΠΙΣΙΔΙΑΣΠΑΦΛΑΓΟΝ
ΕΒΑΣΤΟΥΛΕΓΙΩΝΟΣ
ΑΡ ΟΥΣΤΡΑΤΙΩΤΙΚΟ
ΙΑΣΓΕΝΟΜΕΝΟΝΔΕ

САНКТЪПІТЕРБУРХЪ.

РЕЛЯЦІЯ

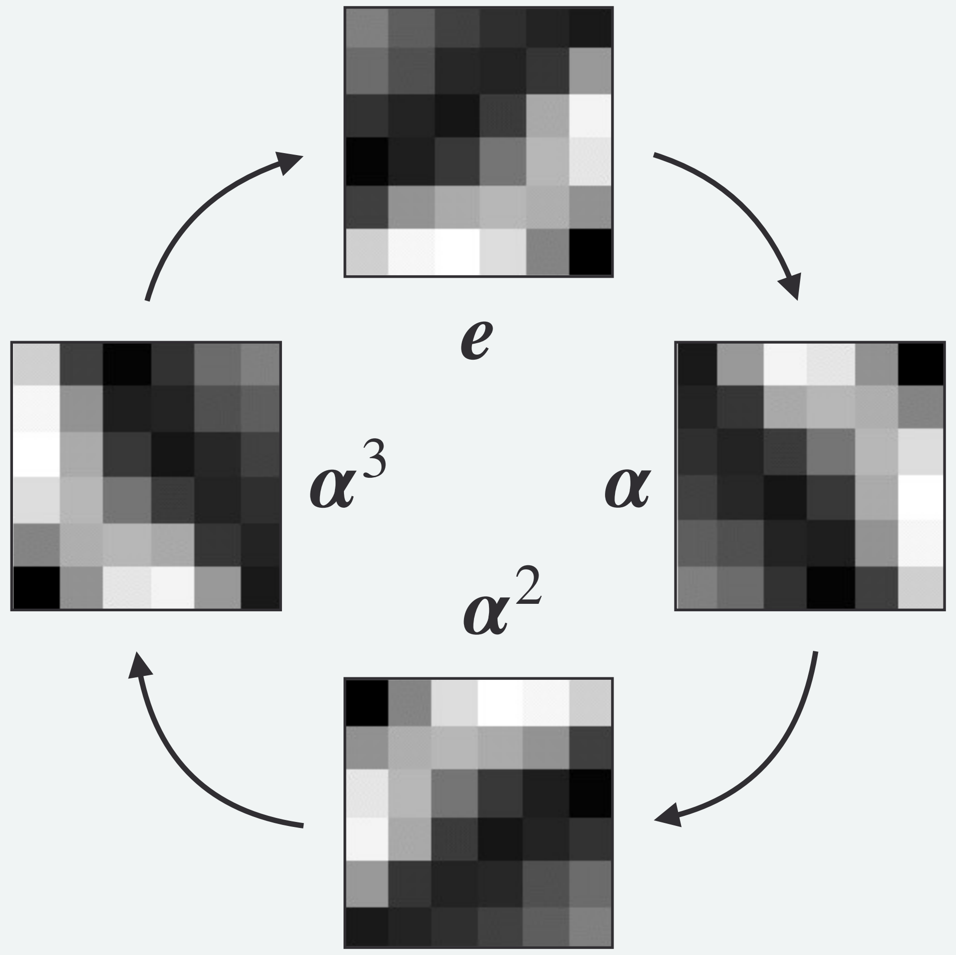
О СЛУЧІВШЕІСЯ МОРСКОИ БАТАЛІИ,
МЕЖДУ РОССІИСКОЮ АВАНГАРДІЕЮ,
И ШВЕЦКОЮ ЭСКВАДРОЮ.

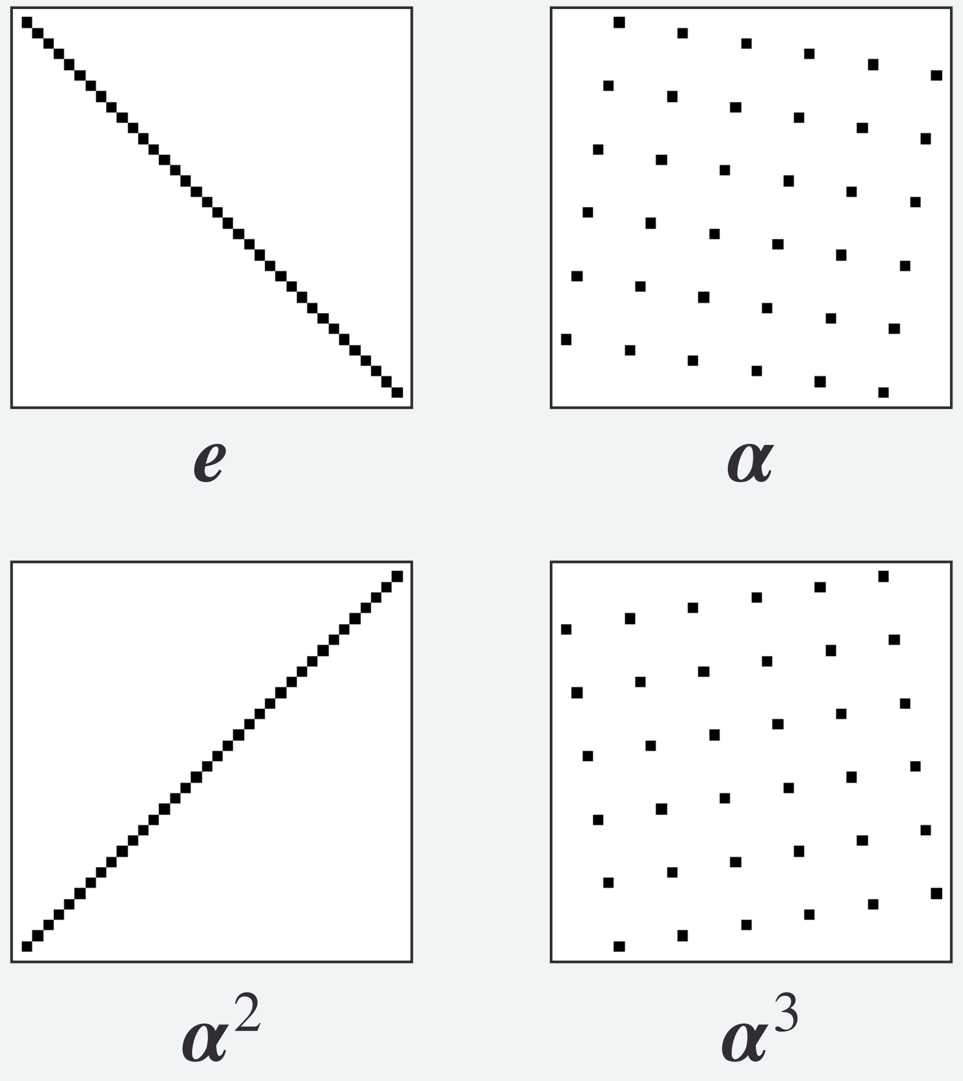
Іюля въ 25 день, [то есть въ воскресеніе] по полудни стала слышать не малая стрѣлба въ морѣ, и по томъ съ брантъ вахта репортовали Господіна Адмірала, что та стрѣлба была отъ крейсеровъ непріятелскіхъ. и по томъ въ 14 - парусахъ швецкои віце адміралъ Ллін пошелъ въ море. чего для много размышляли, для чего такая стрѣлба. и хотя невѣдали для чего, однакожъ положили за не просто. для чего Господінъ Адміралъ отправілъ на море Господіна Шаудбеінахта [по прошенію его] къ блізжнімъ островамъ отъ непріятеля, съ тритцатью пятью скампавѣями. и когда помянутои господінъ шаудбеінахтъ прібылъ къ своему караулу и осмотрѣлъ, что швецкои адміралъ ватрангъ, съ шестью караблями лінѣйными, и съ тремя фрегатами стоітъ на старомъ мѣстѣ. а помянутои віце адміралъ идетъ на зюйдъ остъ. при которомъ и бомбардірскіе галіоты. чего для болѣе стали думать, что или

efficient models

Approach

Elements of group theory





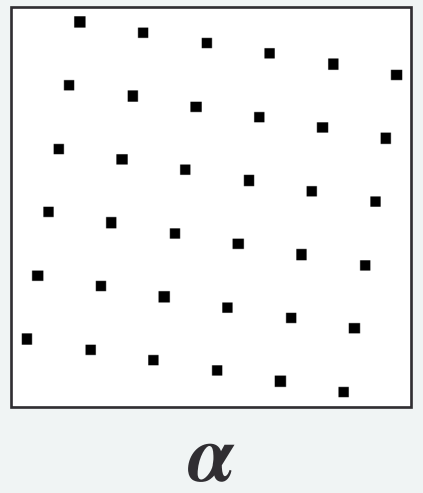
Abstract group

High level interactions

$$\{e, \alpha, \alpha^2, \alpha^3\}$$

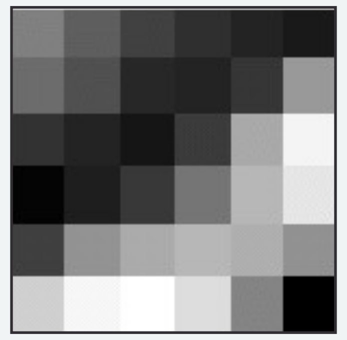
Group representation

“Implementation” matrices



Data space

Where the group acts



Elements of group theory

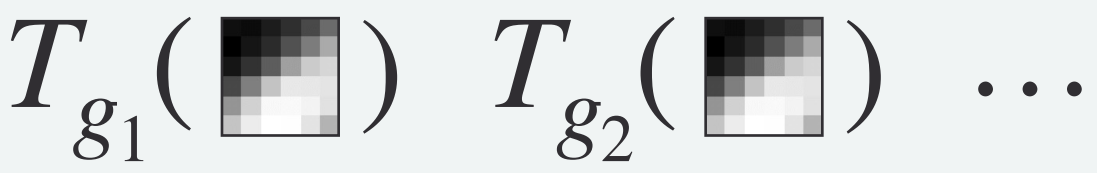
Cyclic groups

What are they?

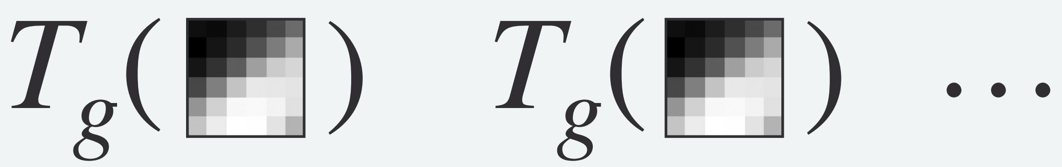
$$C_4 = \{\boldsymbol{e}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^2, \boldsymbol{\alpha}^3\}$$

Why do we care?

- **short:** computation
- **long:** finite*, abelian, generator

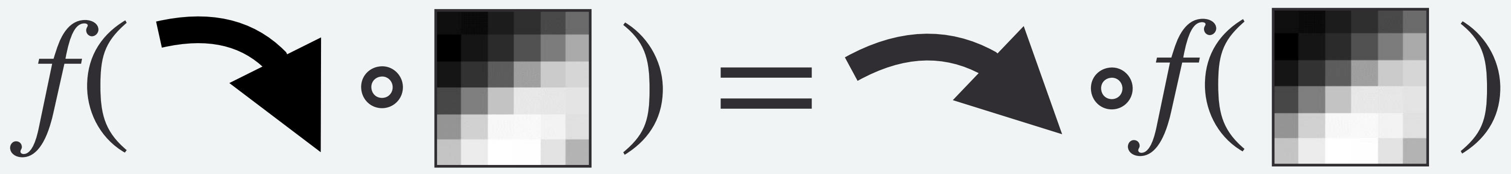


$$\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^l = \boldsymbol{\alpha}^l \cdot \boldsymbol{\alpha}^k$$



Equivariance

What does it mean?



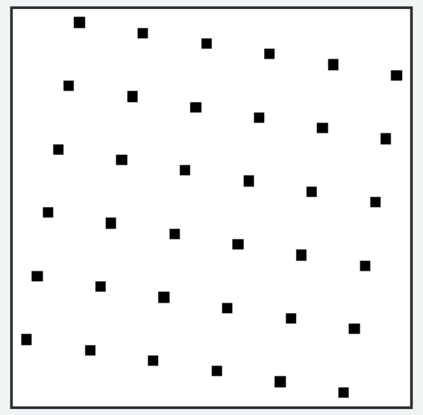
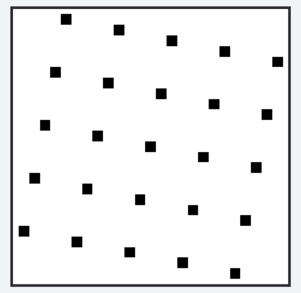
(A bit more formally $f(T(x)) = T'(f(x))$.)

Example:

$f(x) = x^2$ **vs** $f(\alpha \cdot x) = \alpha^2 \cdot x^2$

Here $T(u) = \alpha \cdot u$ and $T'(u) = \alpha^2 \cdot u$.

So where do groups come in?

$T =$  $T' =$ 

Represent the “same” transformation!

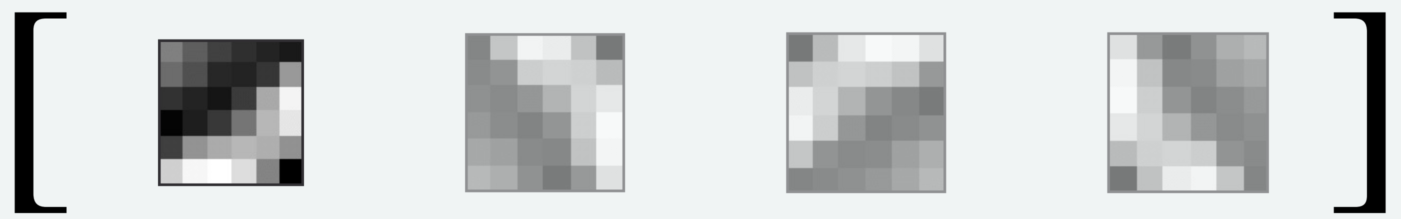
Main idea

Theorem[1]: For equivariance of a neural network with respect to a group G each layer needs to be a G-convolution.

$$(f * g)(u) = \sum_{v \in G} f(uv^{-1})g(v) \qquad \left(\text{vs } (f * g)(u) = \sum_{v \in \mathbb{Z}} f(u - v)g(v)\right)$$

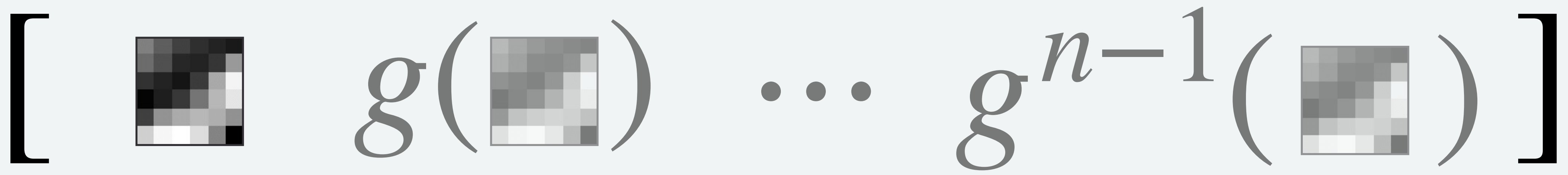
For C_4

Convolution with the rotated set



What about C_n?

Cardinality and action change

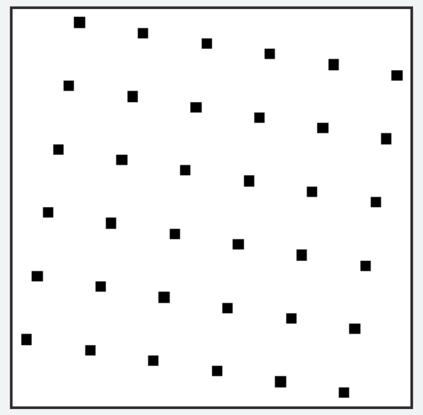


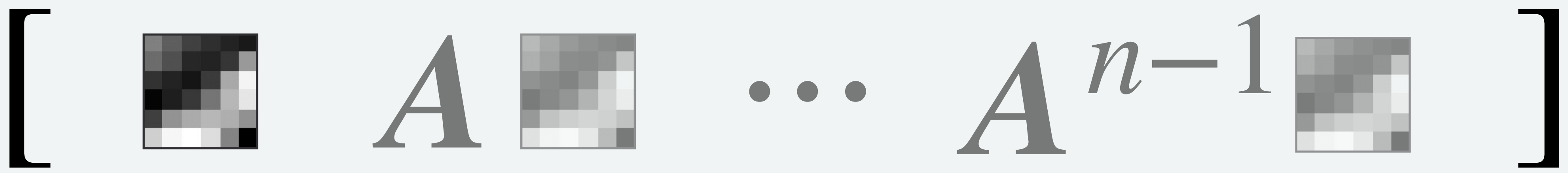
How to represent g?

Use invertible matrices

$$\rho : G \to \mathrm{GL}_d(\mathbb{R})$$

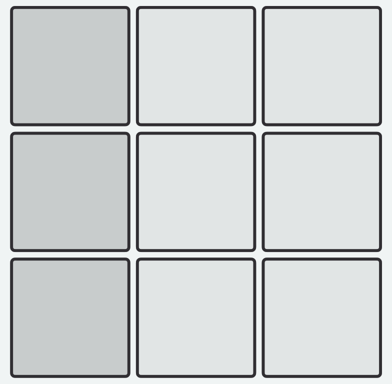
$g = \boldsymbol{A} =$



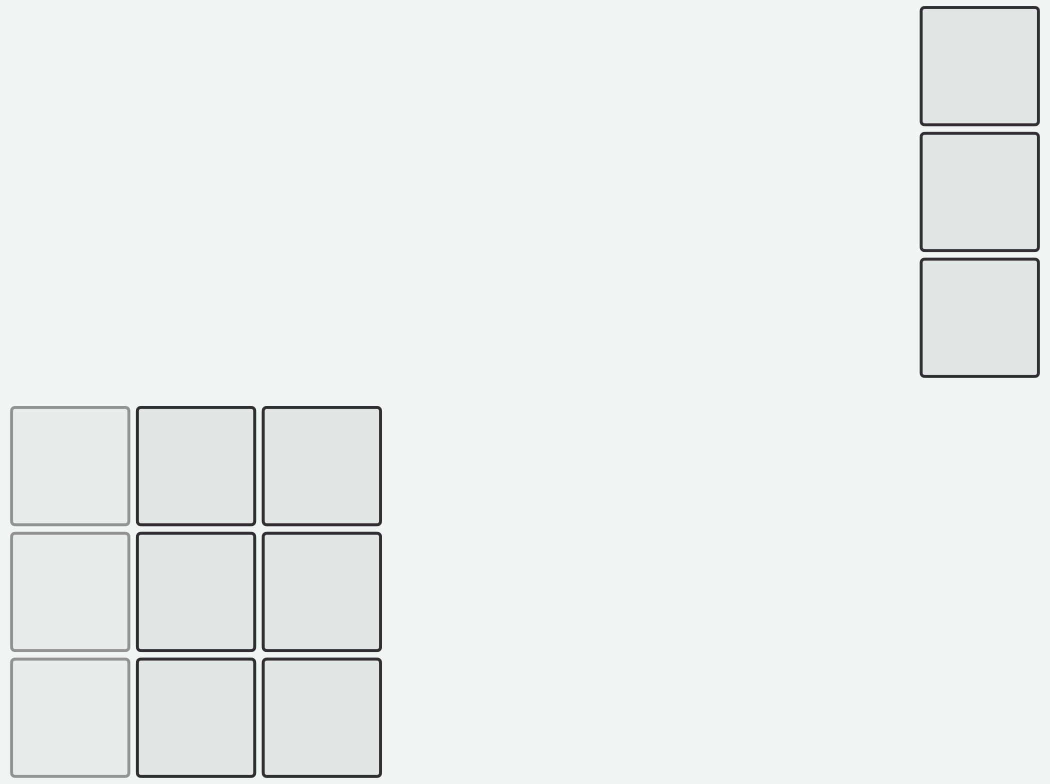


[1] Risi Kondor and Shubhendu Trivedi. "On the generalization of equivariance and convolution in neural networks to the action of compact groups". In International Conference on Machine Learning, 2018.

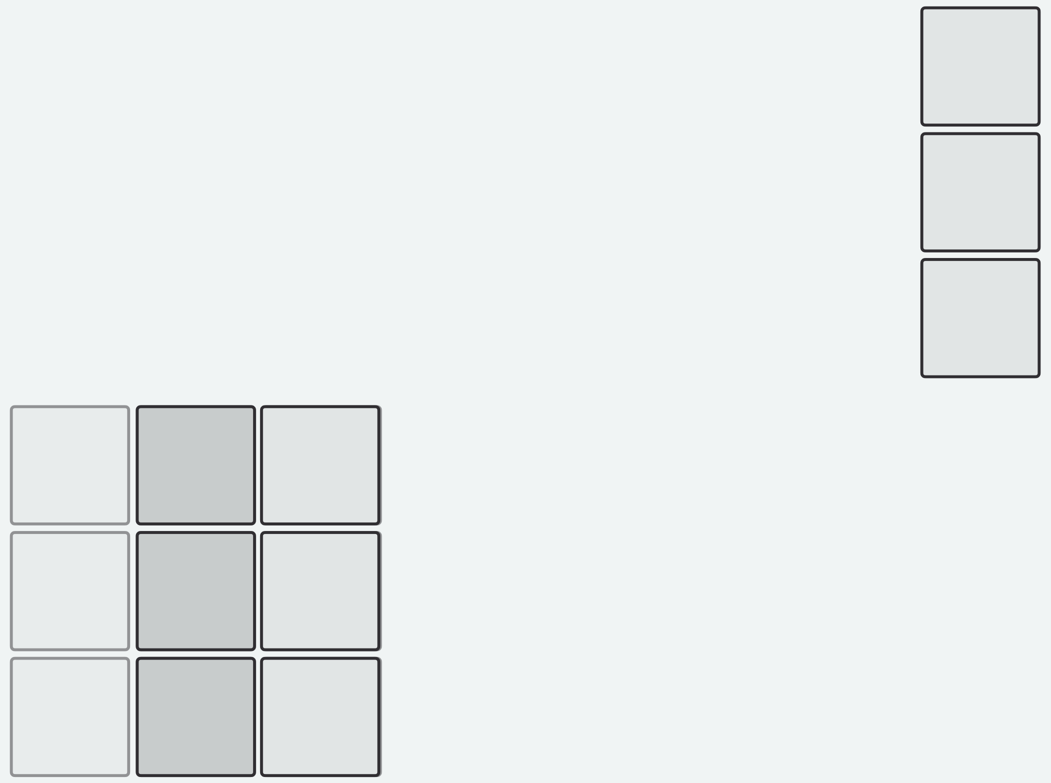
Vectorization



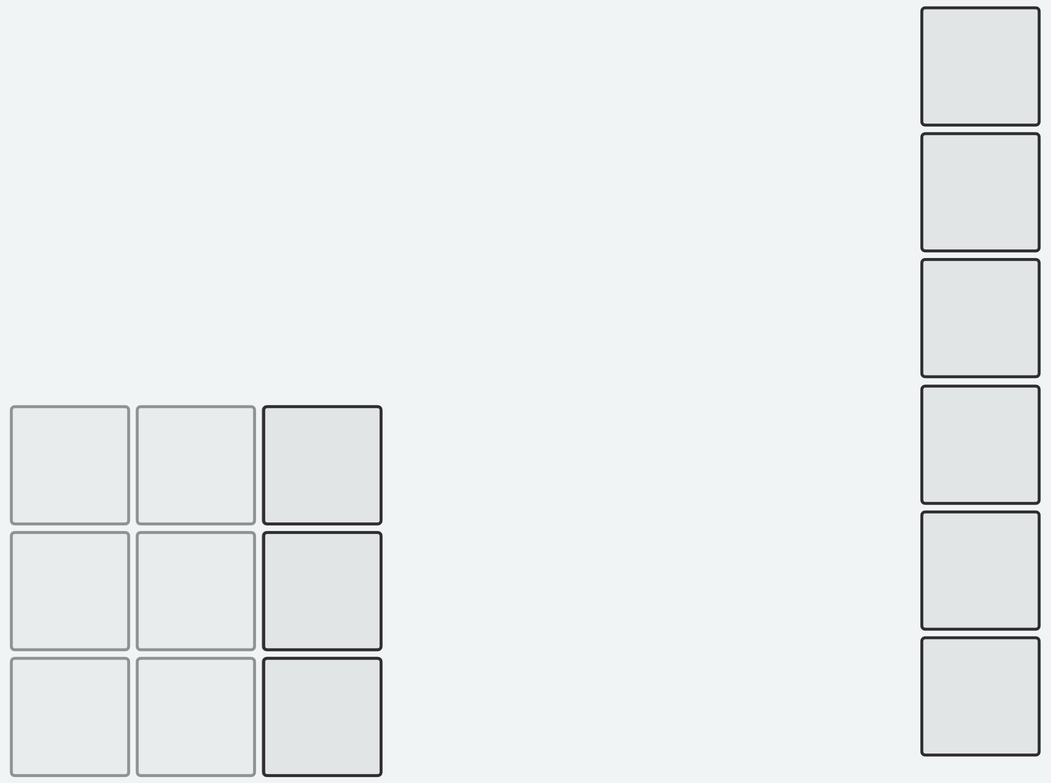
Vectorization



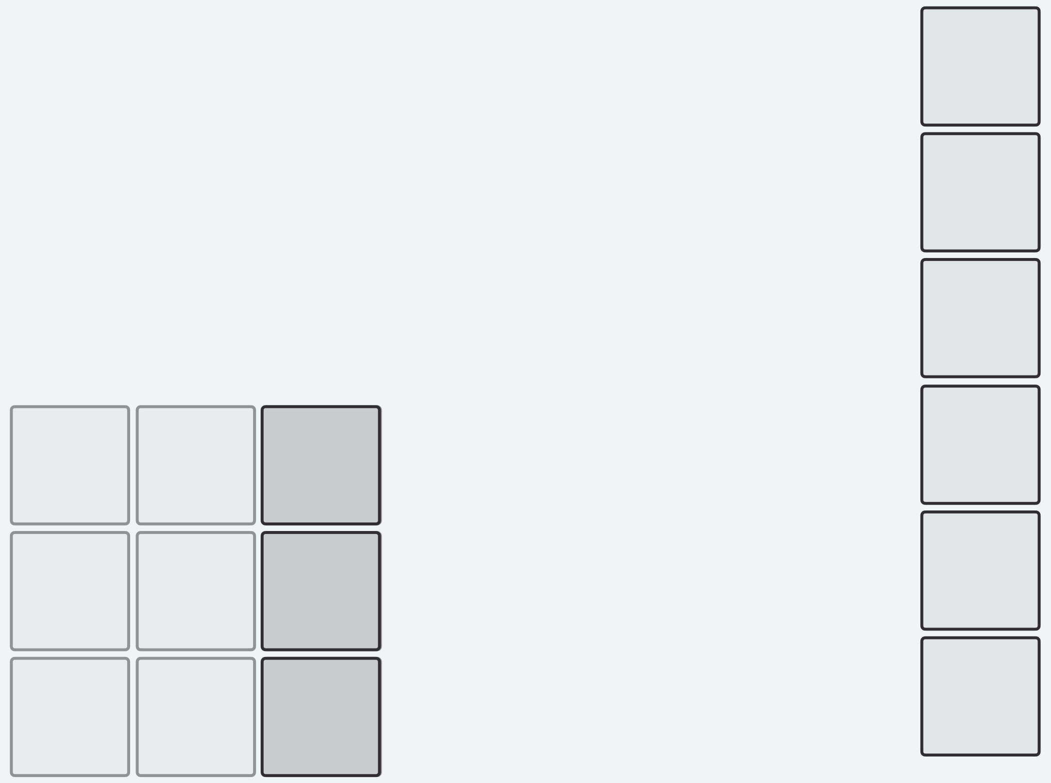
Vectorization



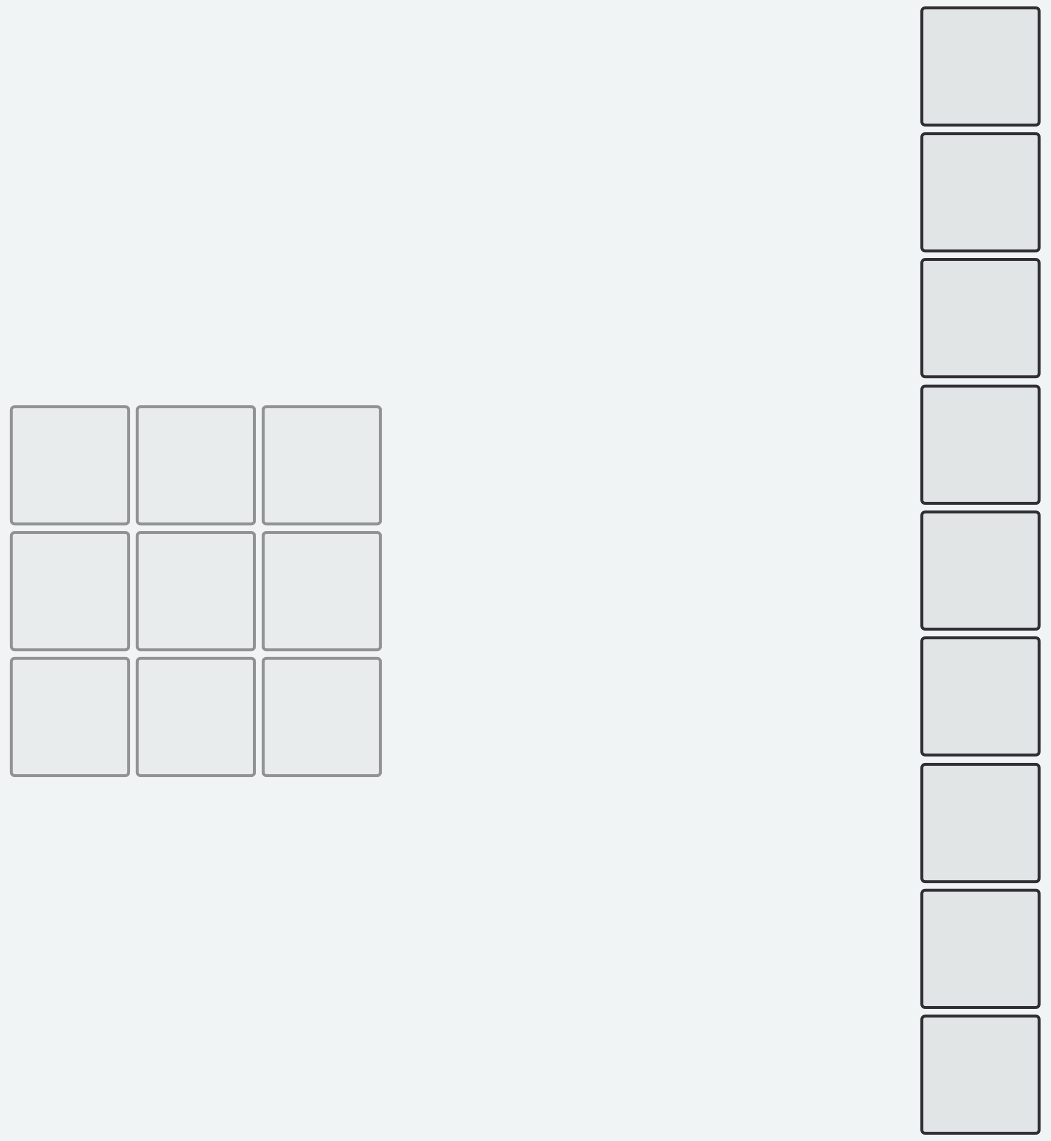
Vectorization



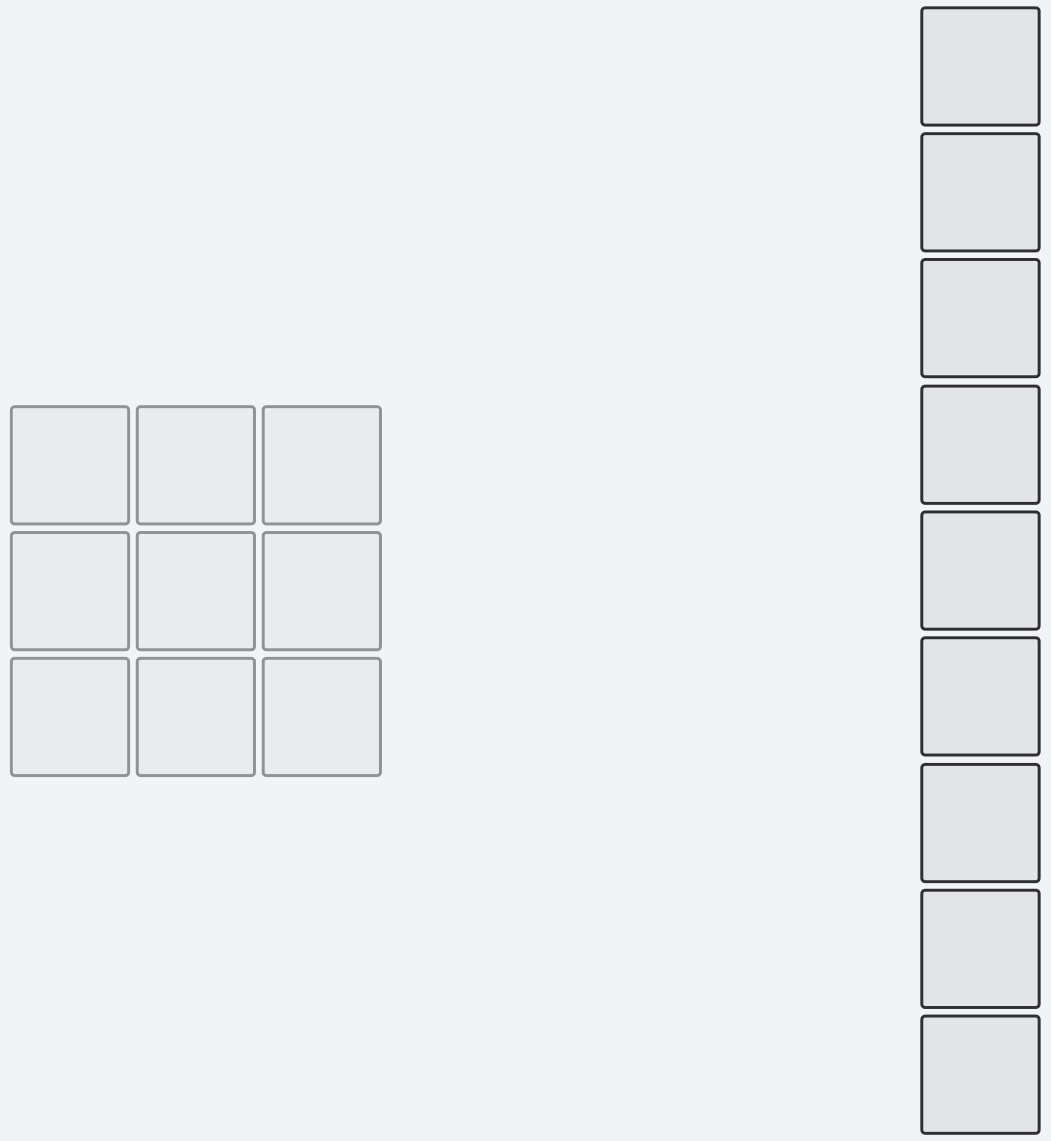
Vectorization

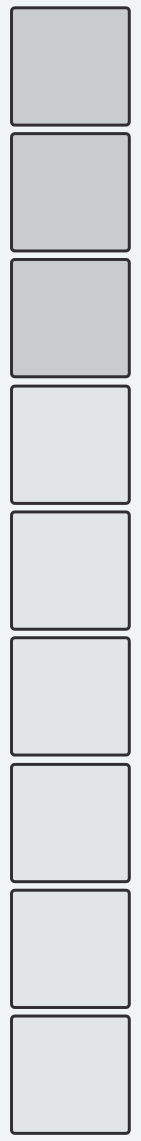


Vectorization

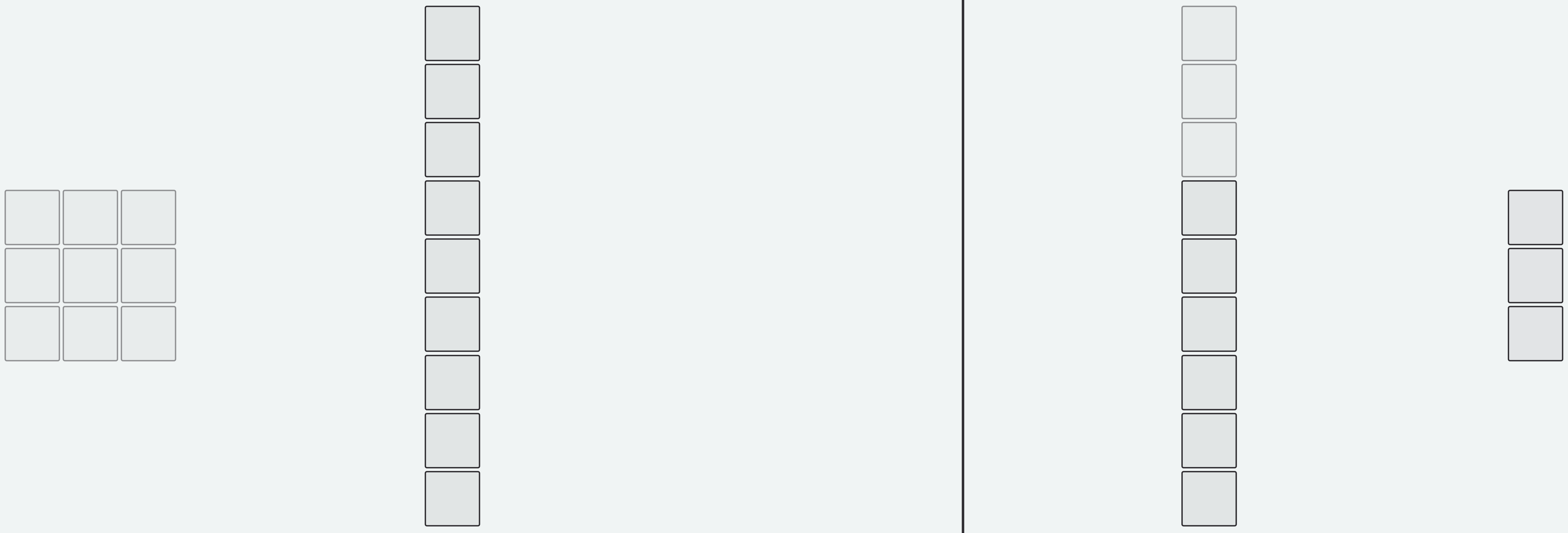


Vectorization

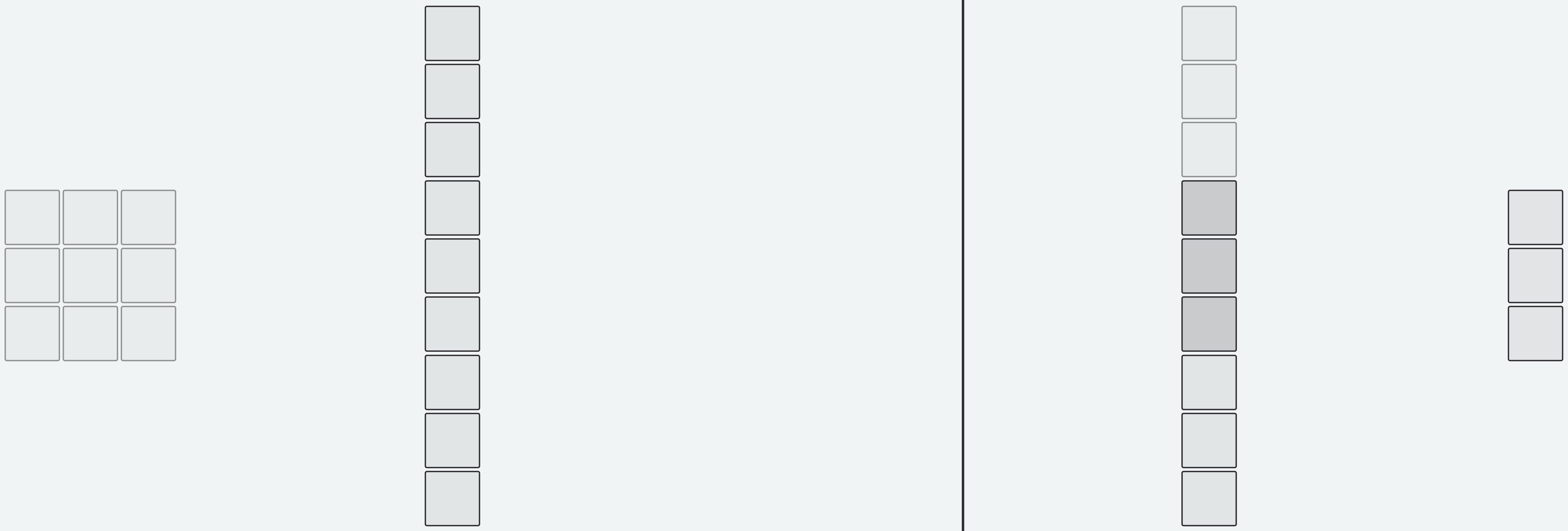




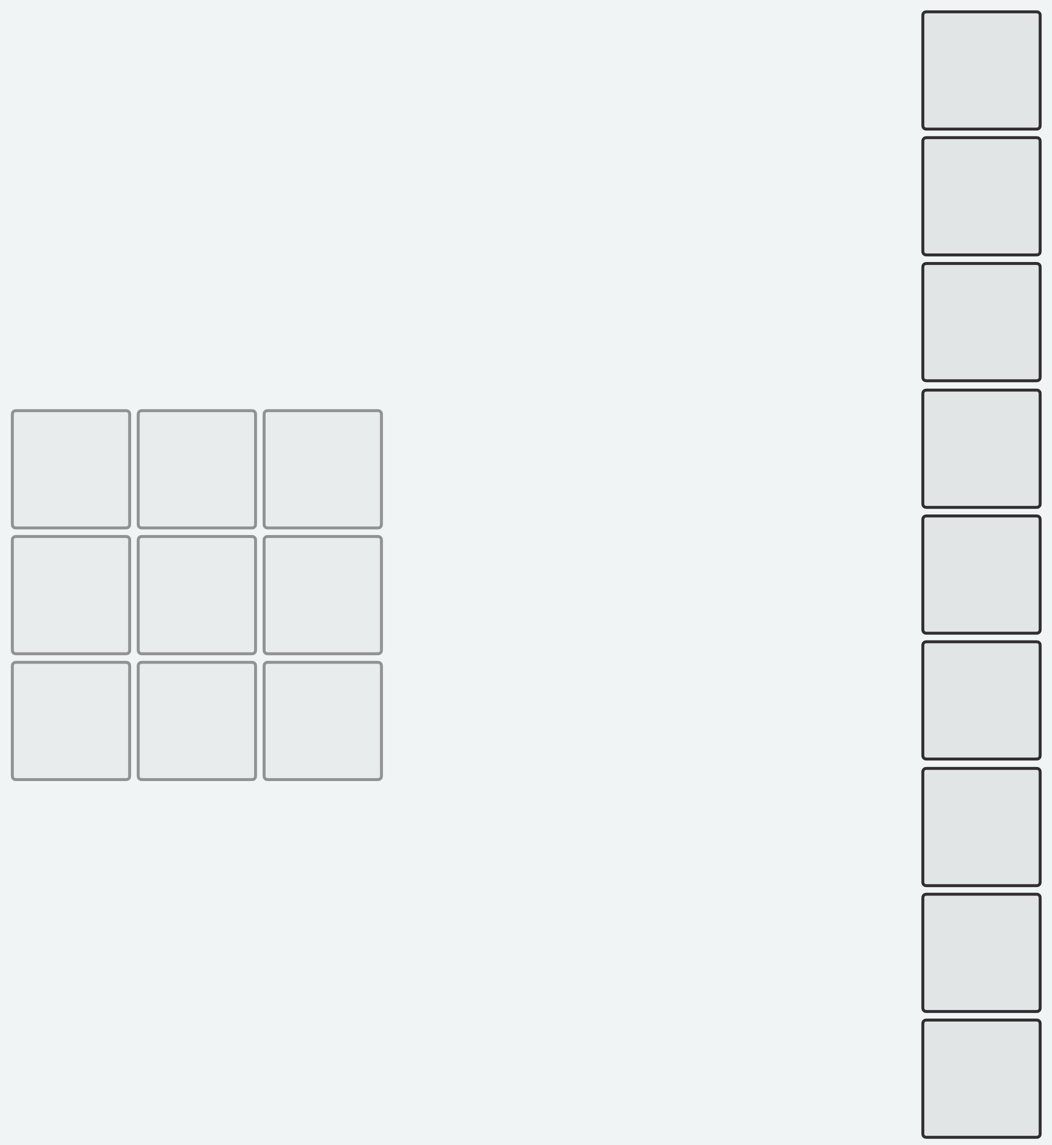
Vectorization

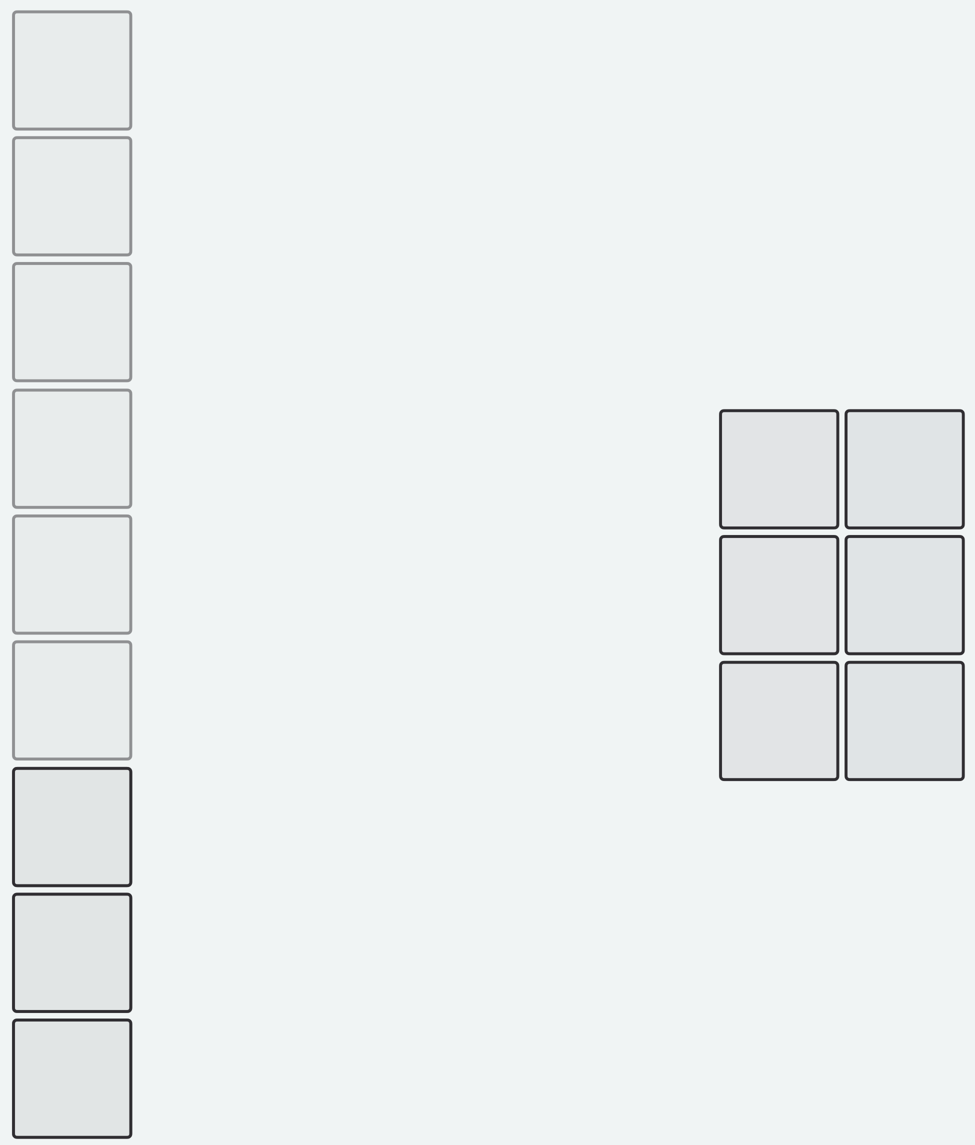


Vectorization

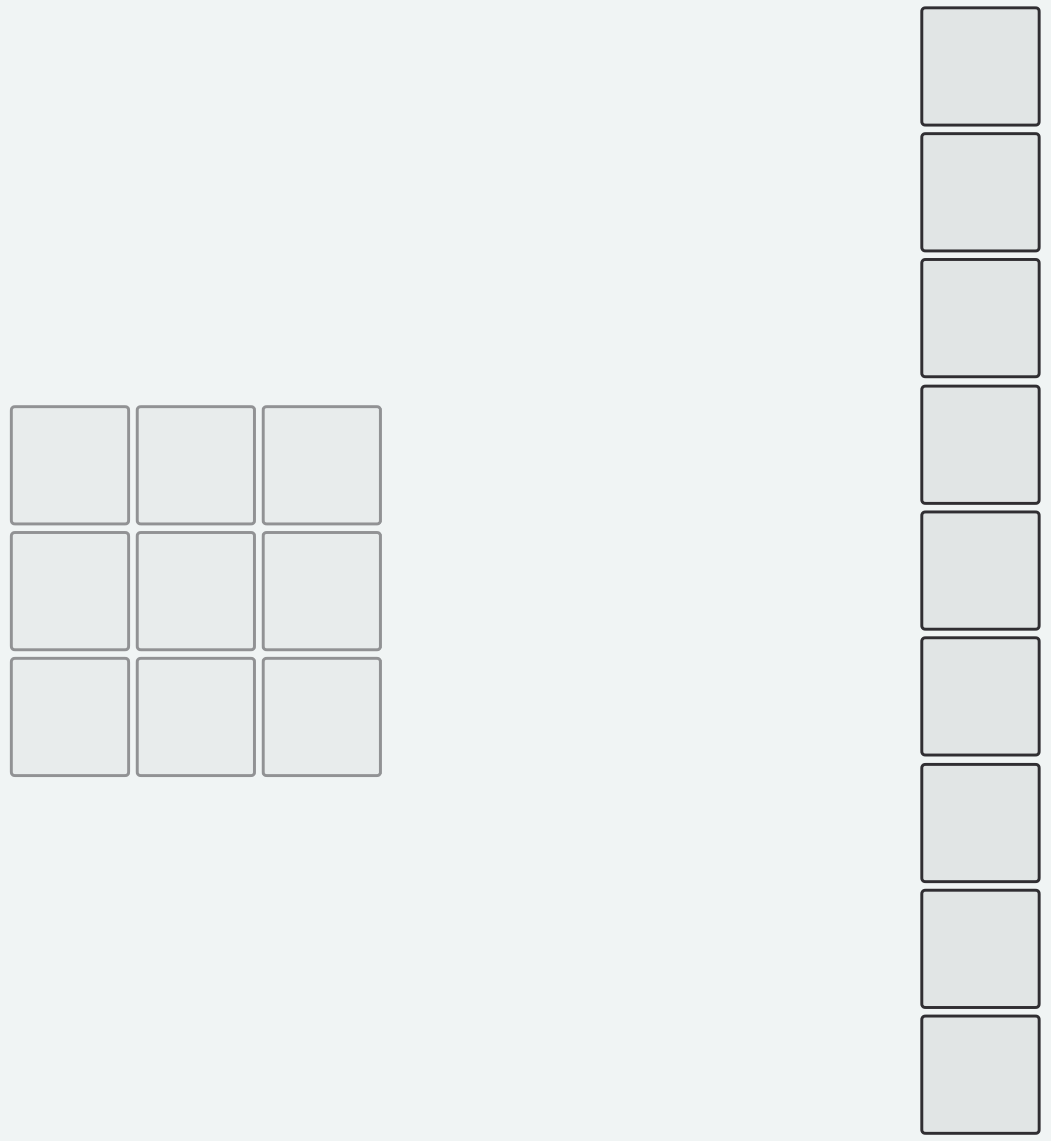


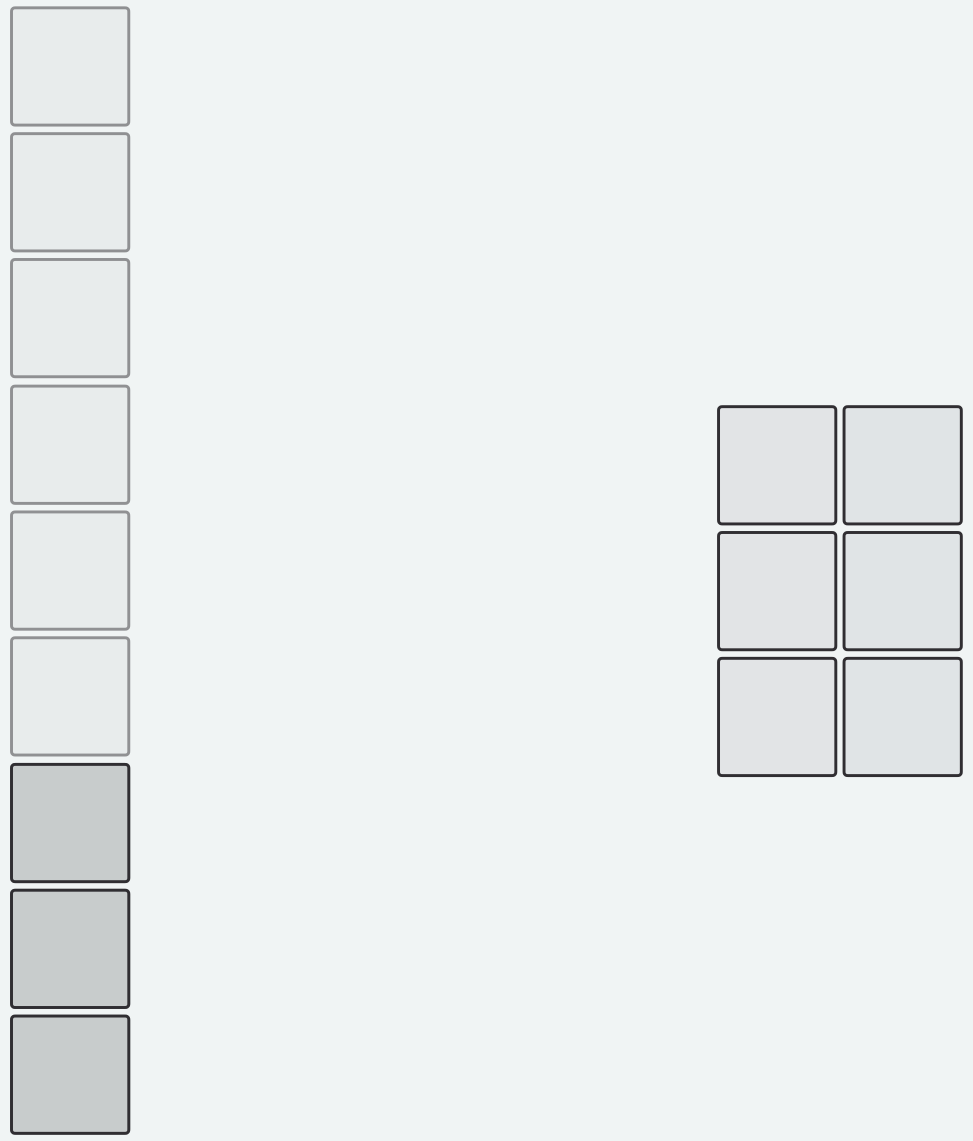
Vectorization



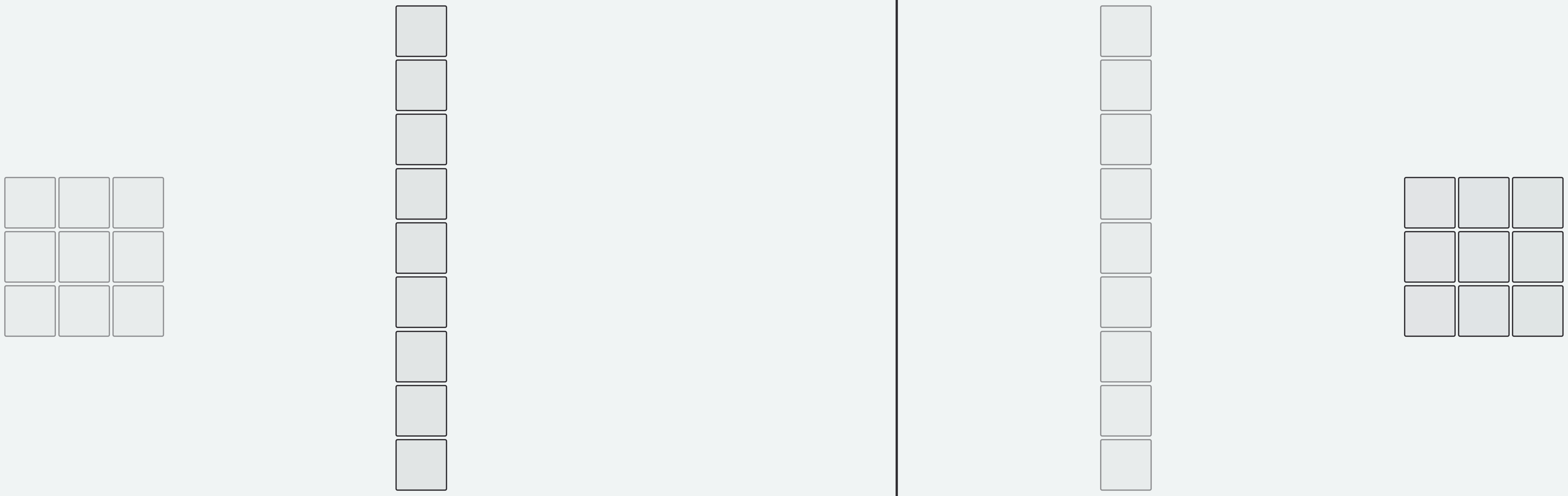


Vectorization



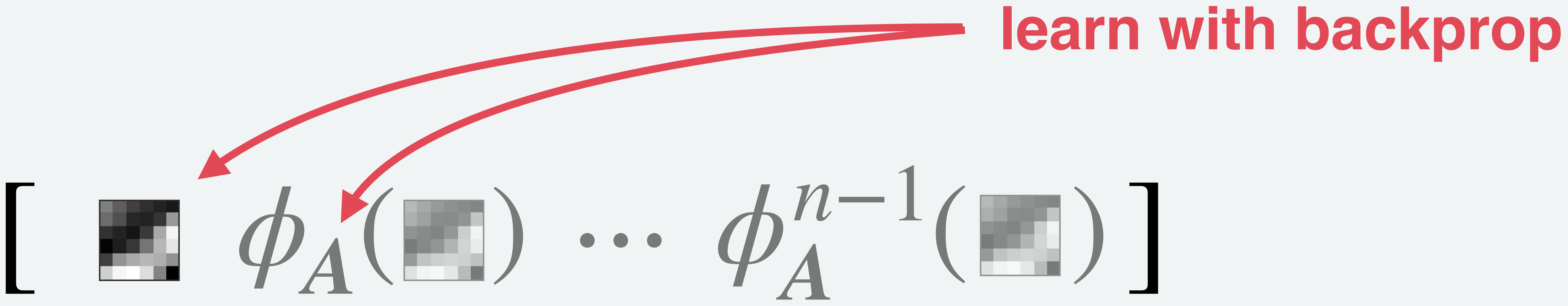


Vectorization



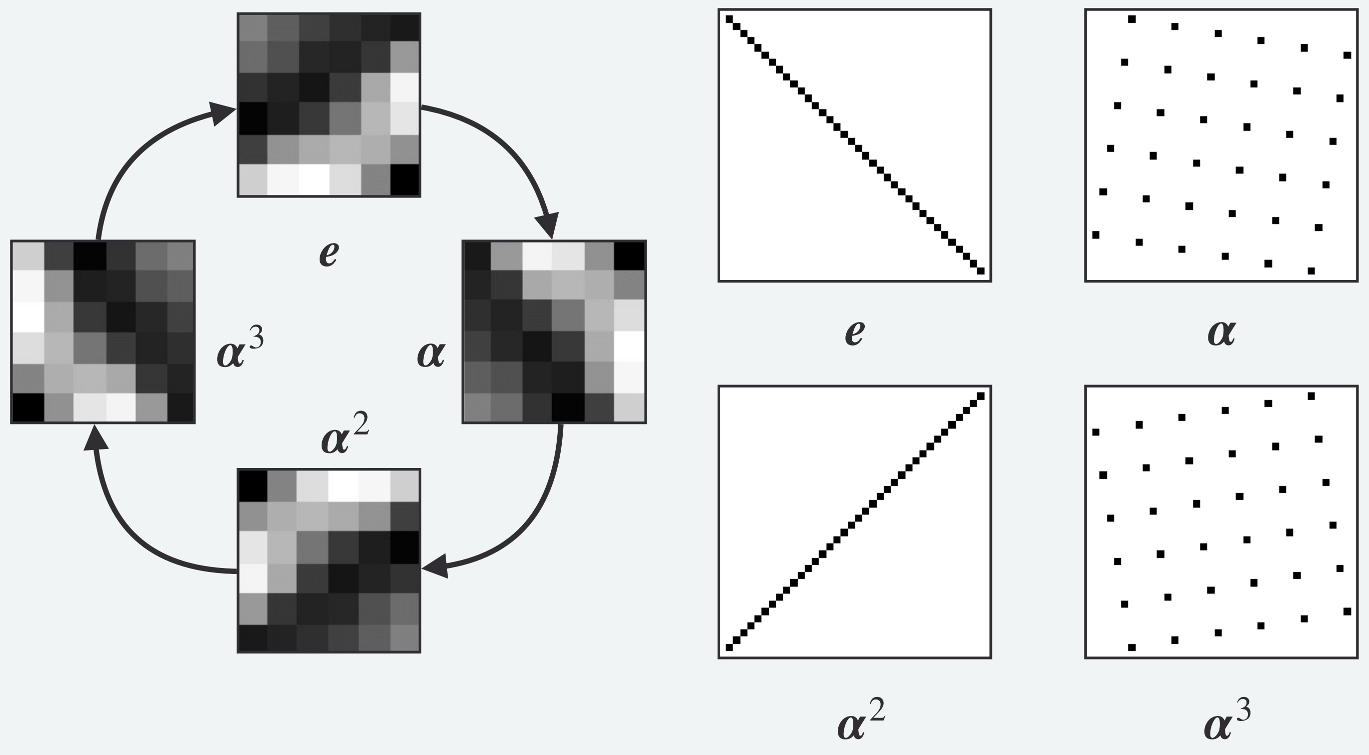
Final model

Filters

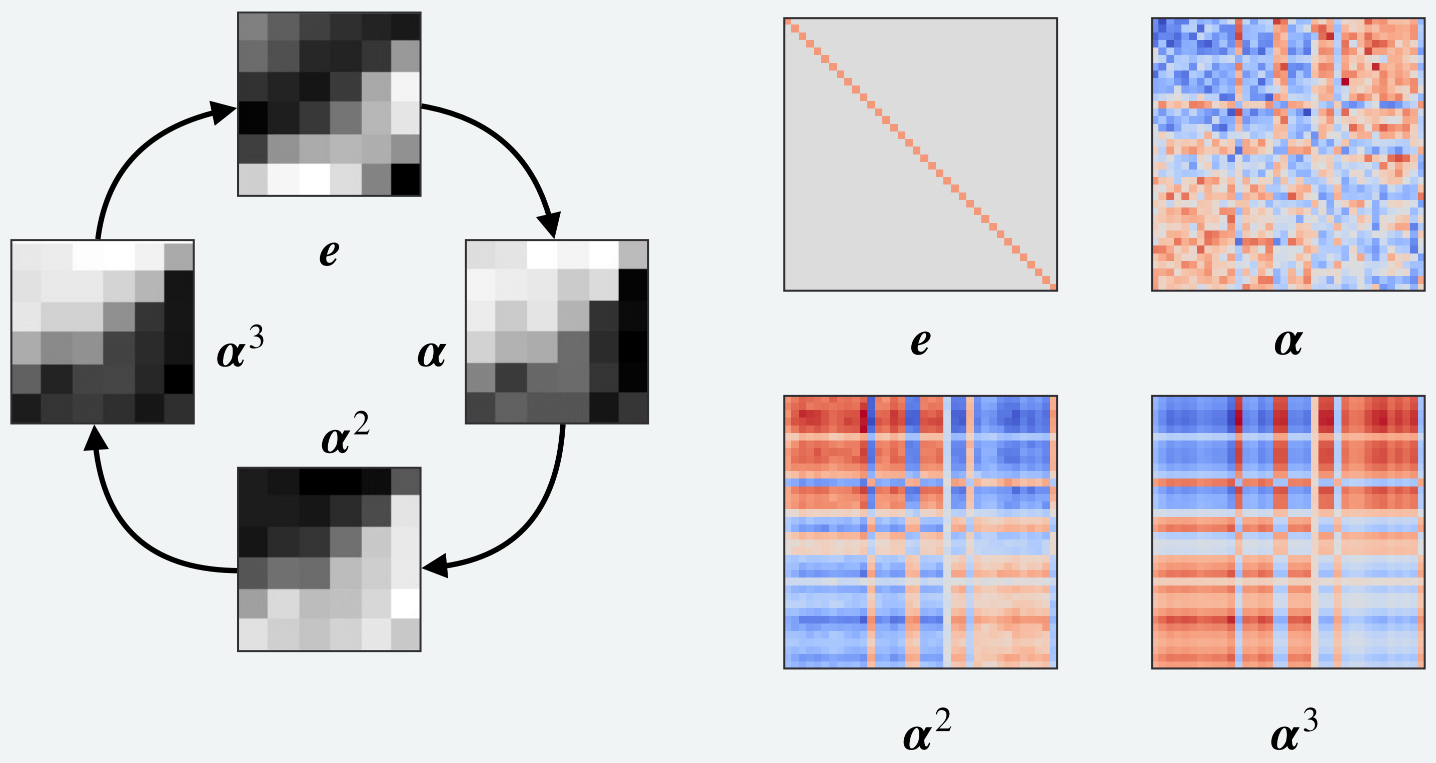


$$\phi_A : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$$
$$X \mapsto \mathrm{vec}^{-1}(A\,\mathrm{vec}(X))$$

Now we go from this...



... to this!



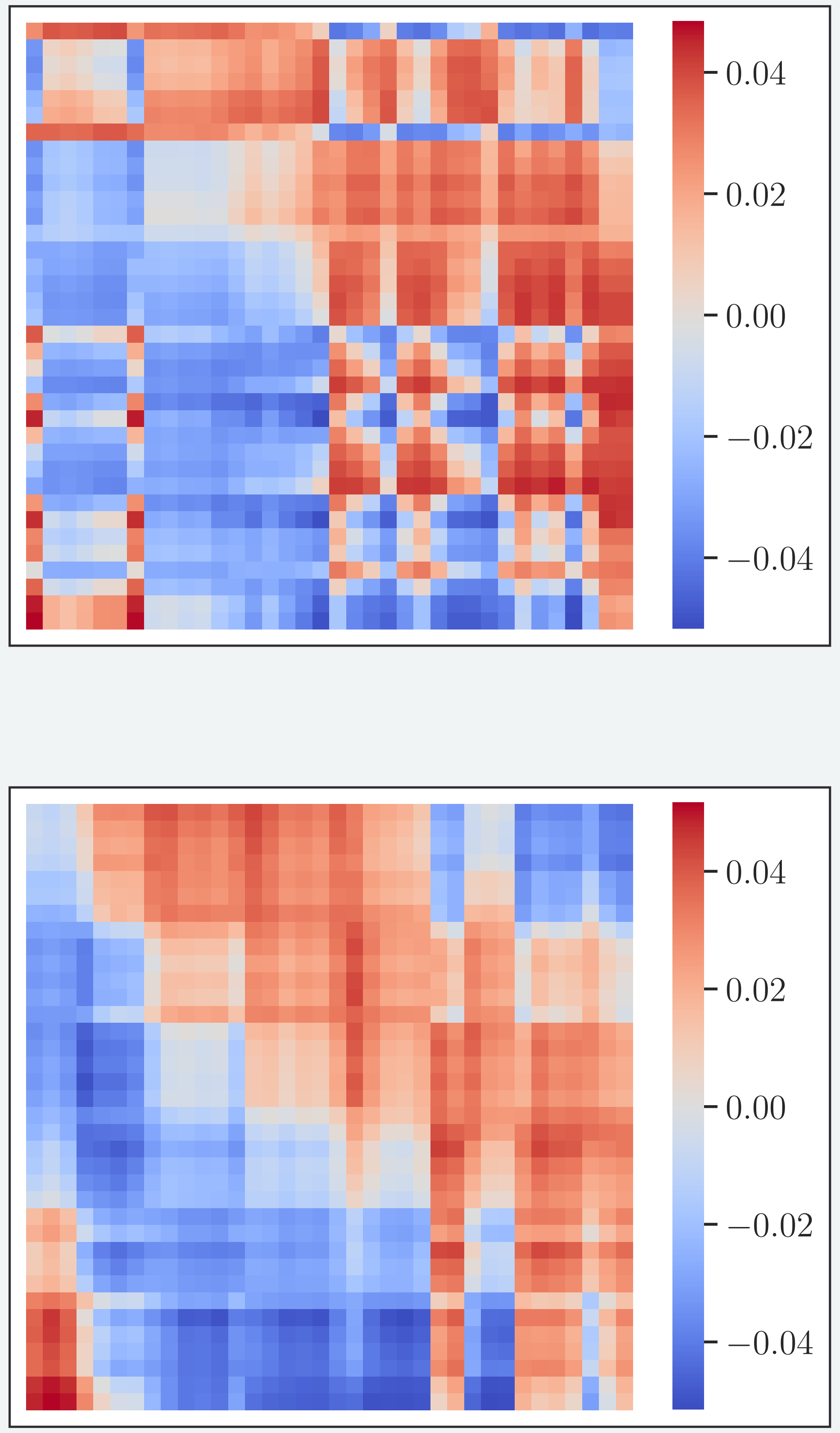
Invertibility loss

A needs to be invertible

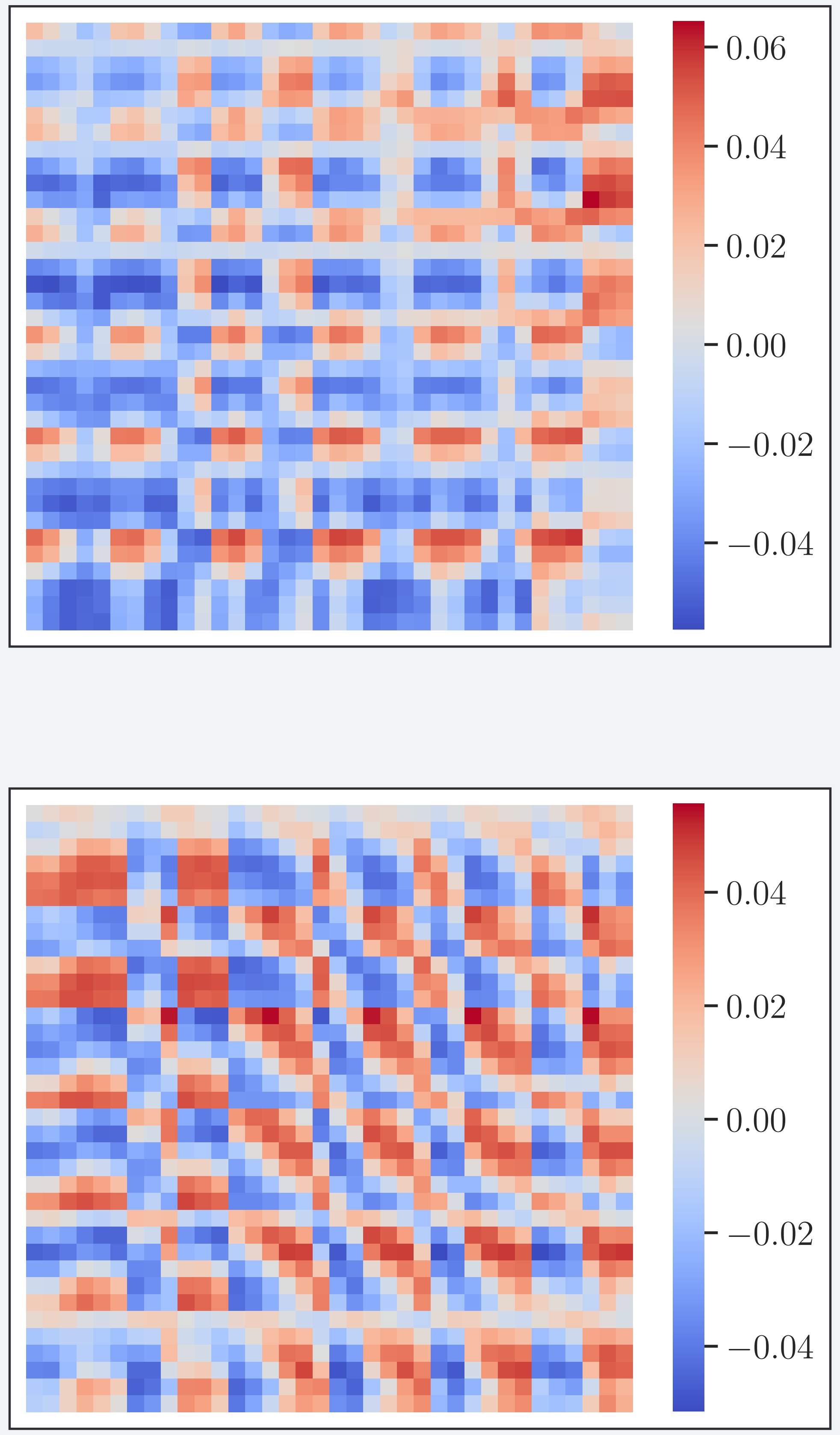
$$L = \mu\|A\tilde{A} - I\|_F$$

Experiments

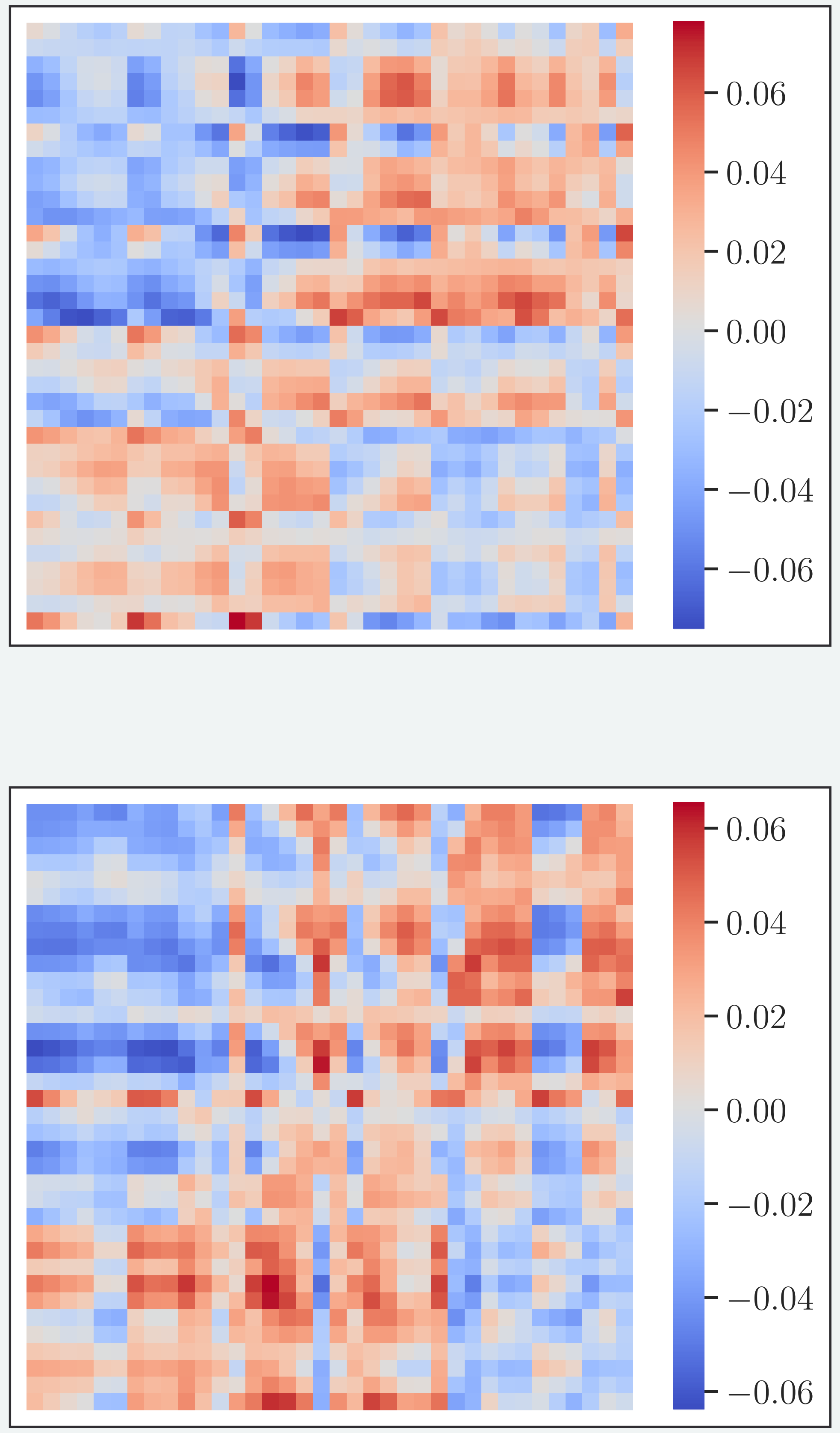
Group structures



Skew-symmetric

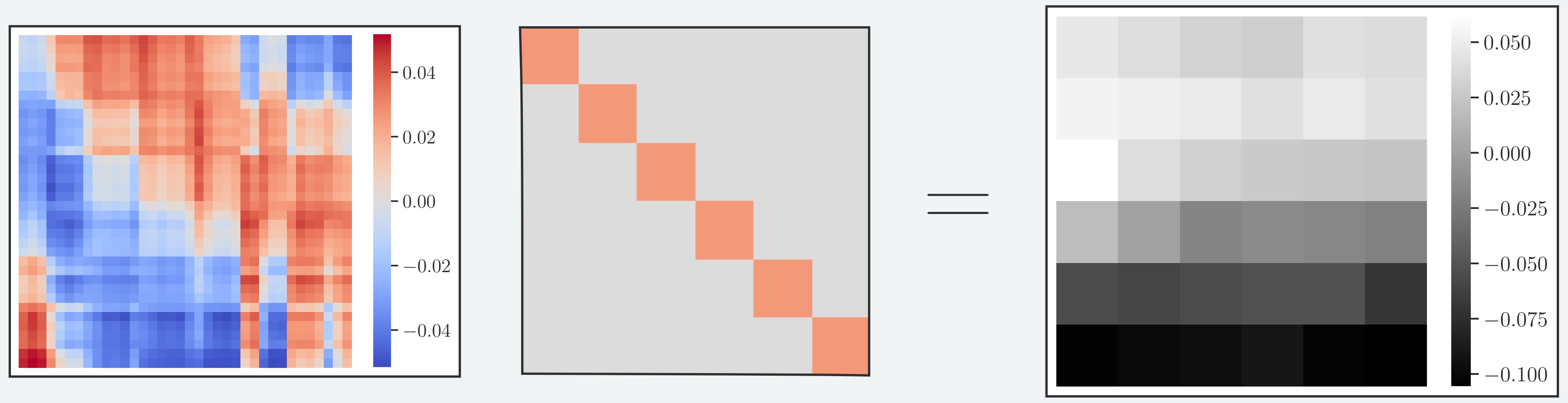


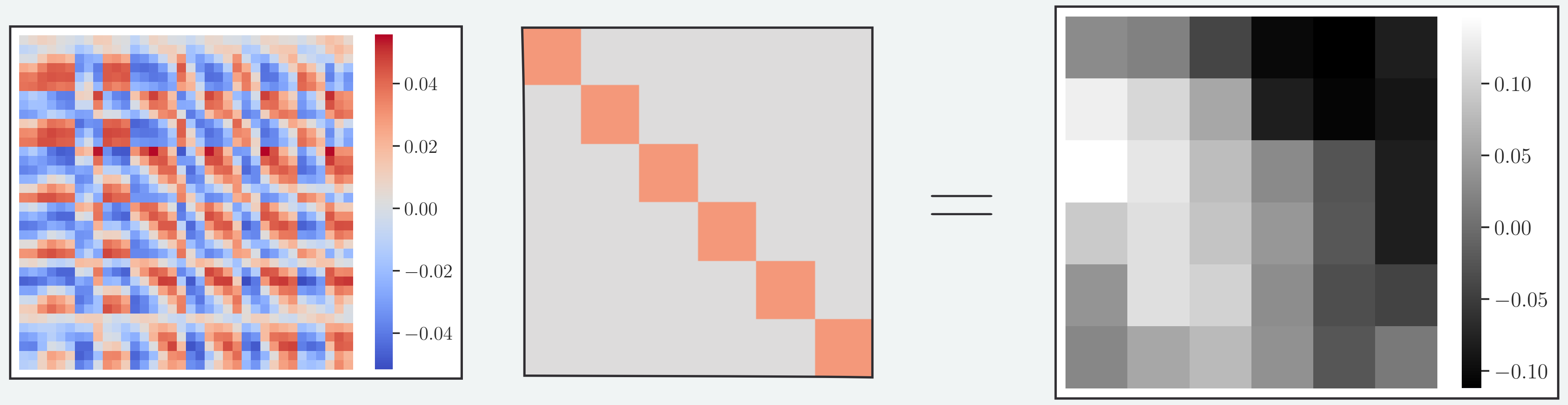
Toeplitz

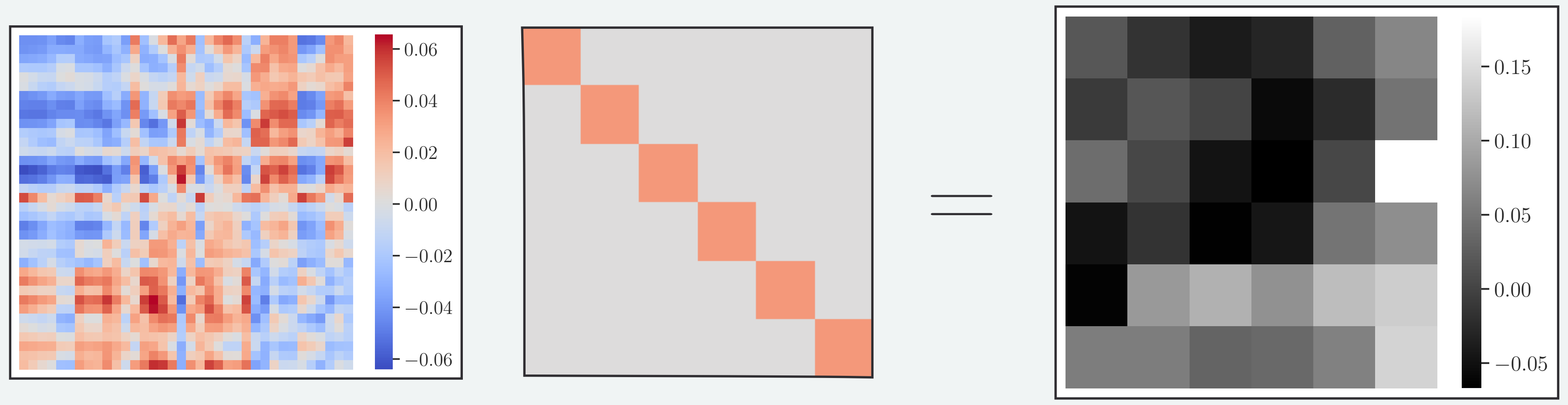


Multi-scale

Interpreting the structures





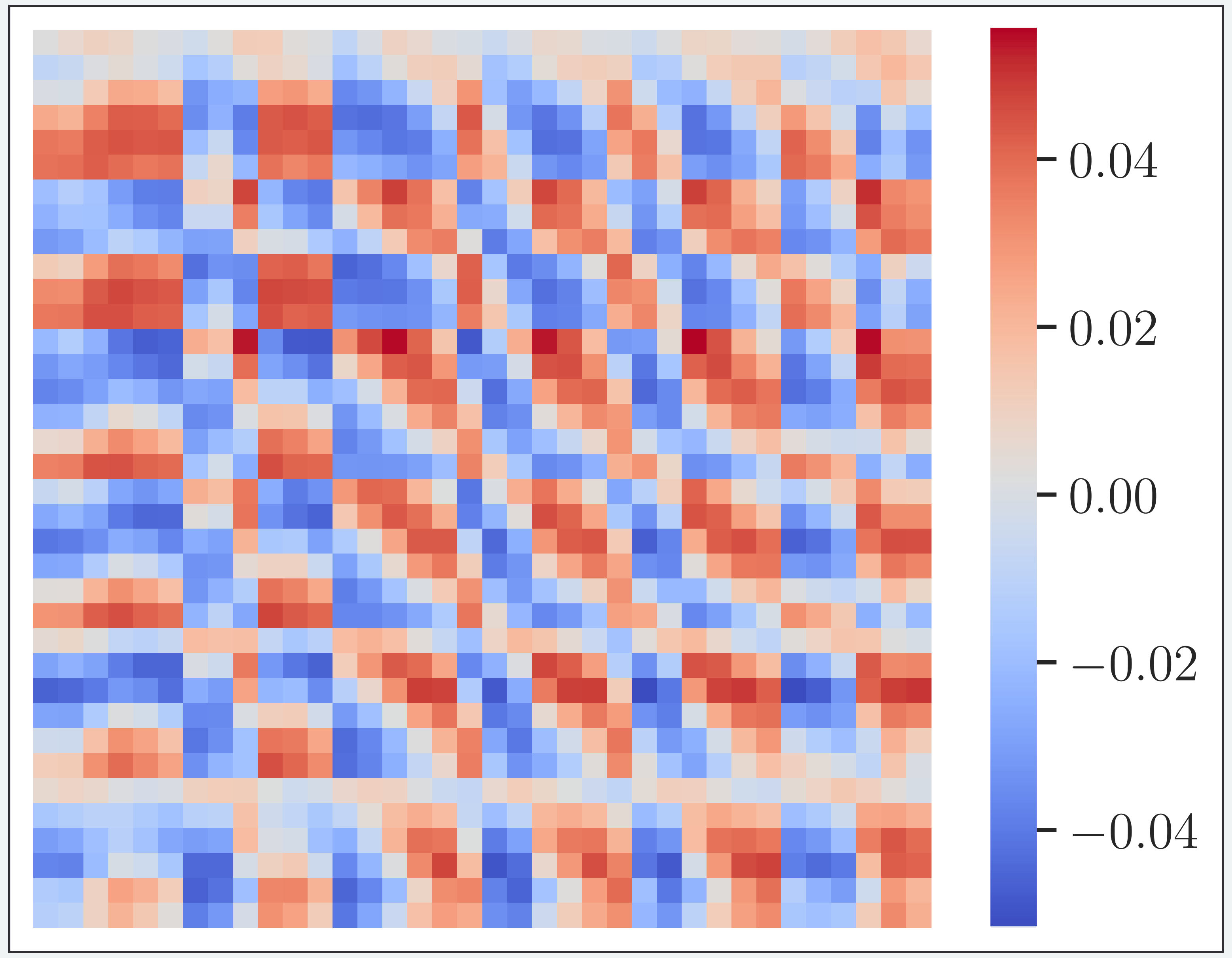


Skew-symmetric

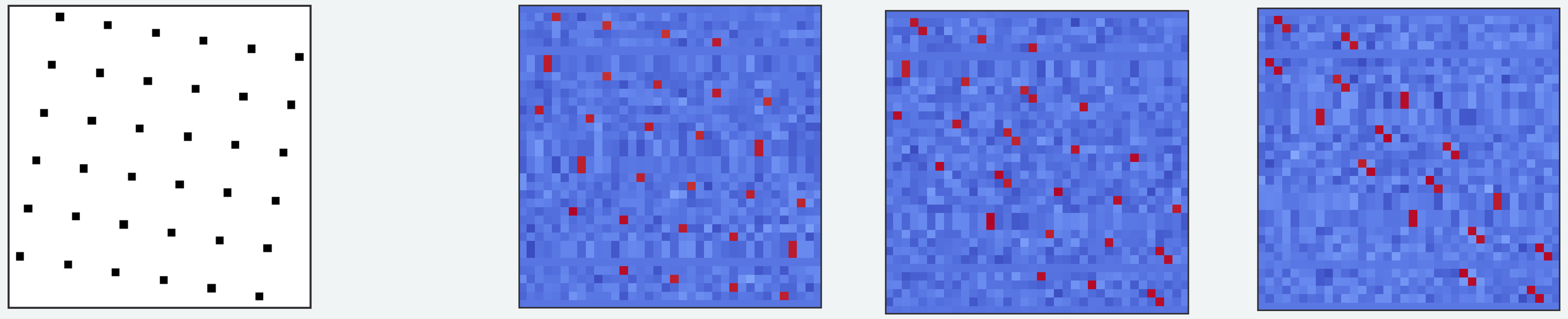
Toeplitz

Multi-scale

Interpreting the structures

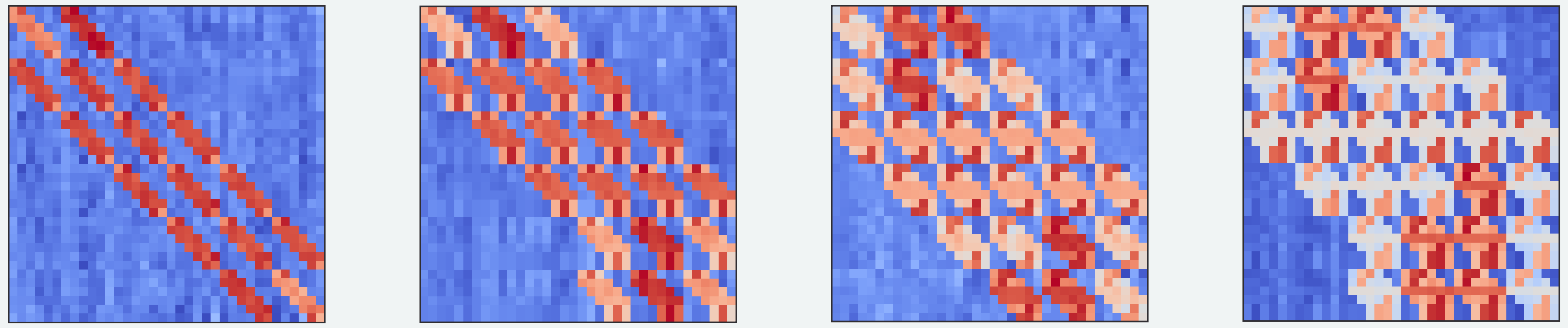


Rotations



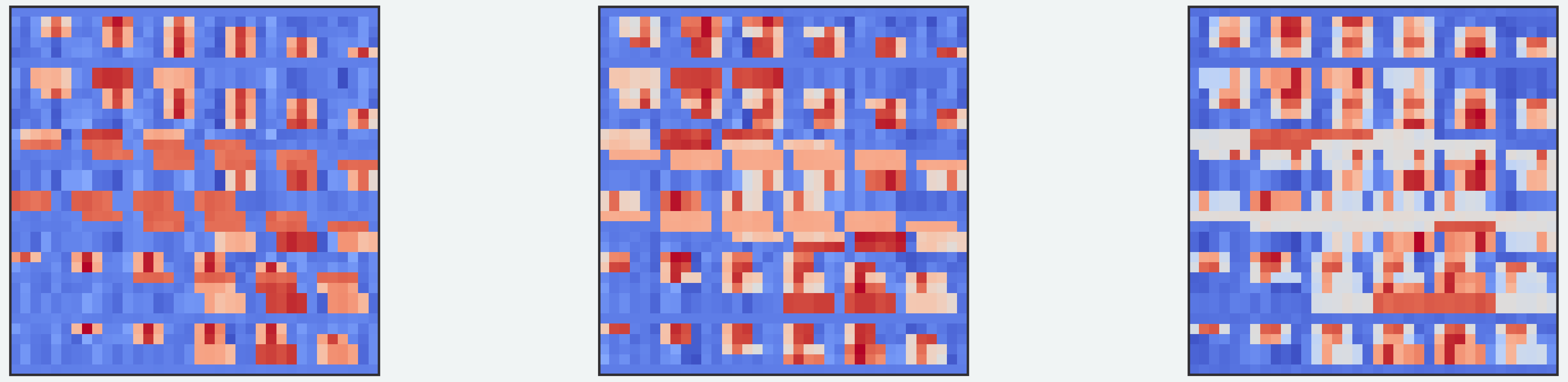
$\theta = 90^\circ$ $\theta = 60^\circ$ $\theta = 45^\circ$ $\theta = 30^\circ$

Pooling



$r = 3$ $r = 4$ $r = 5$ $r = 6$

… both!



$\theta = 60^\circ$

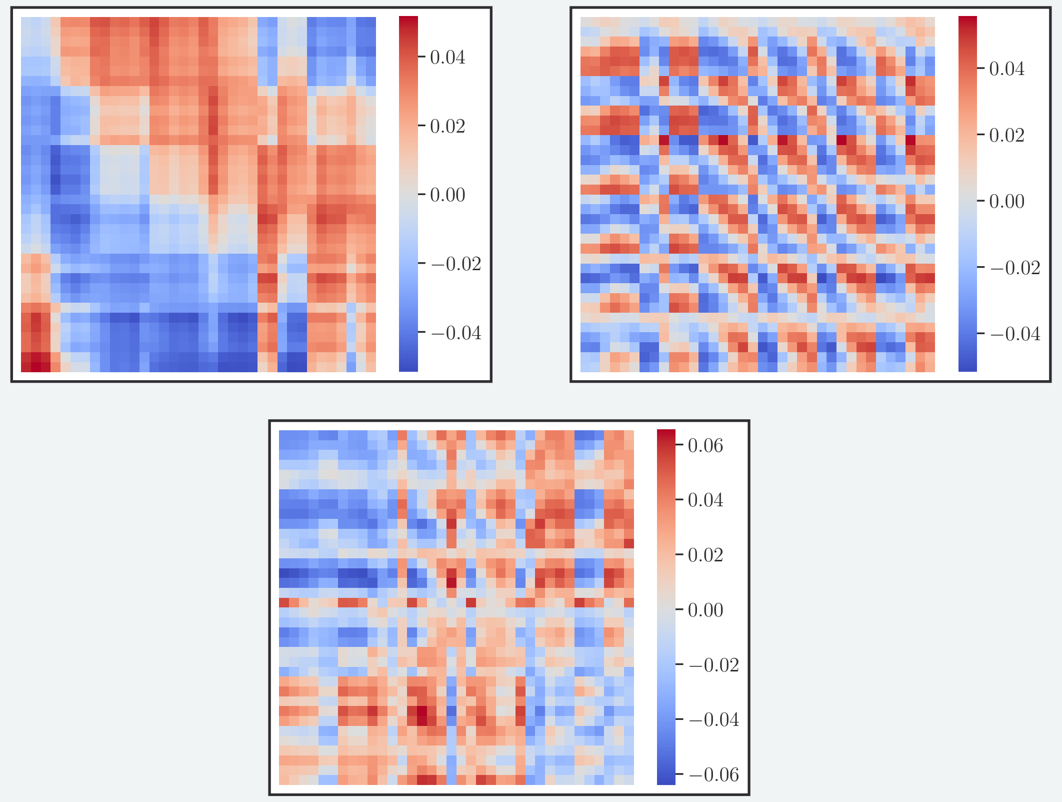
$r = 4$ $r = 5$ $r = 6$

Observations

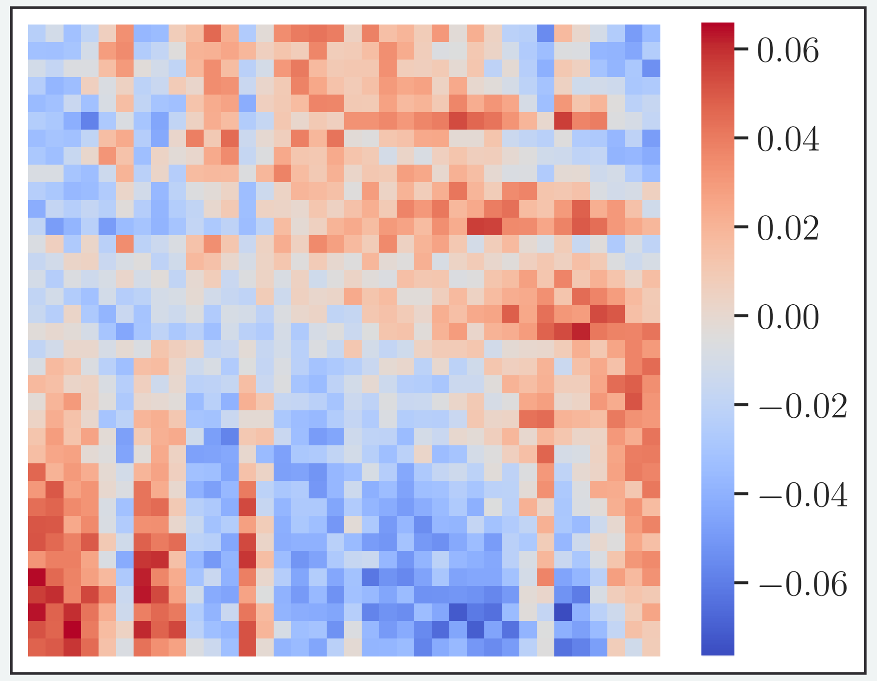
- aligned on grid
- acts on many pixels

More experiments

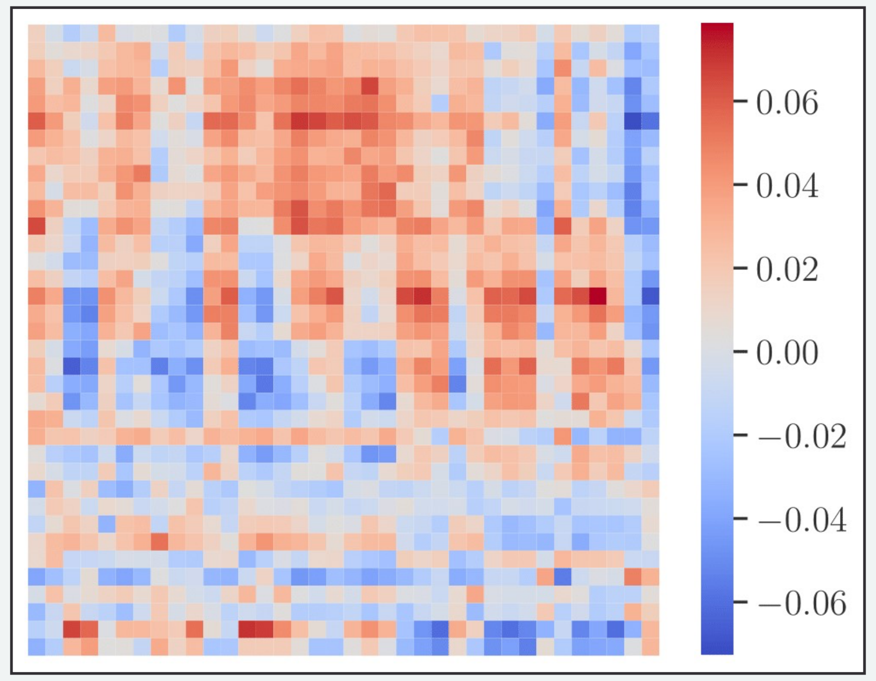
Actions are consistent



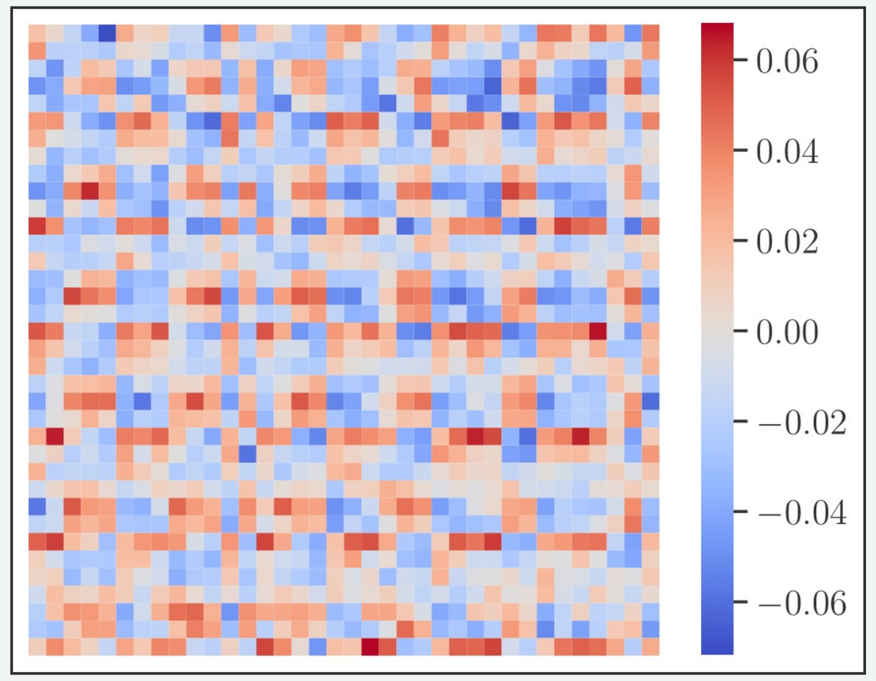
CIFAR10



MNIST

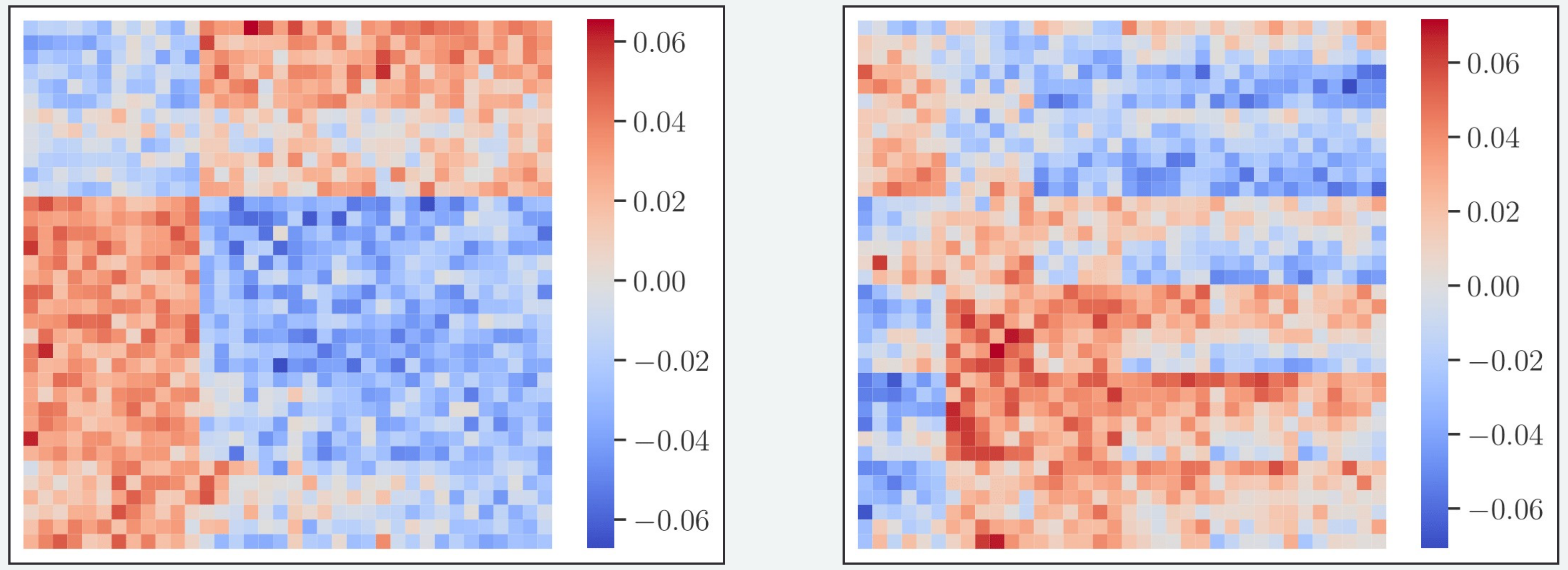


rotMNIST

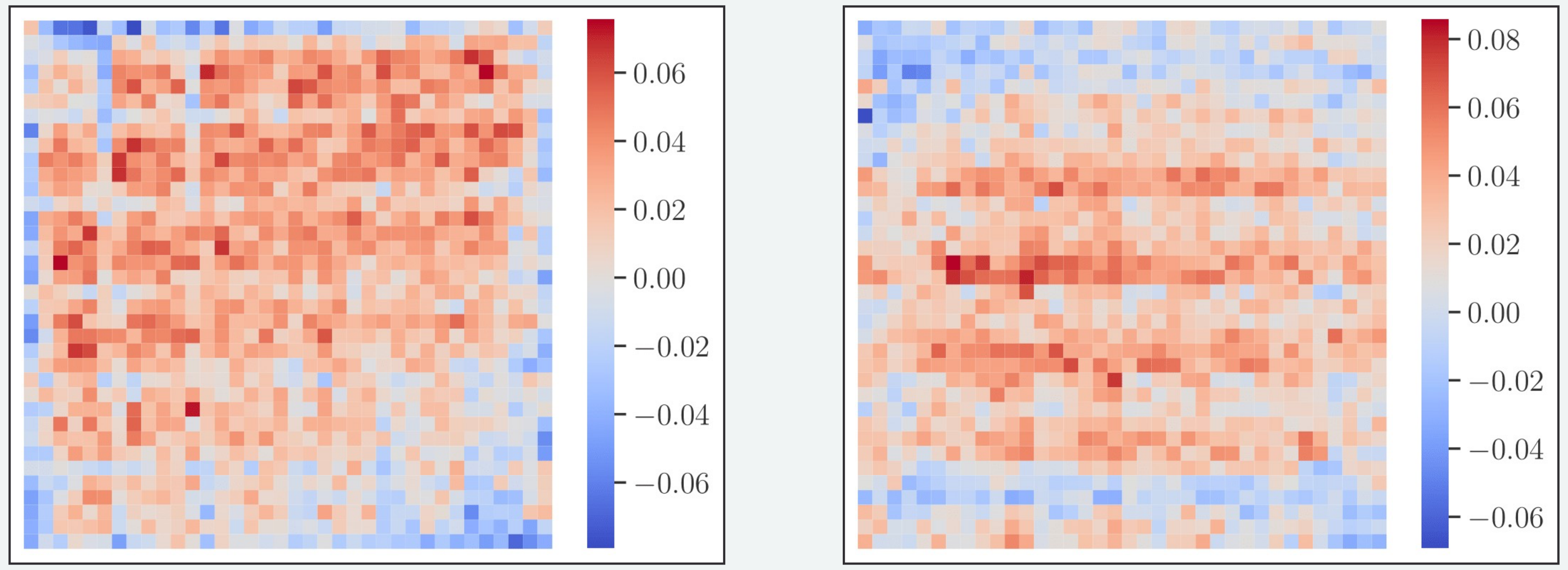


FashionMNIST

… and task dependent

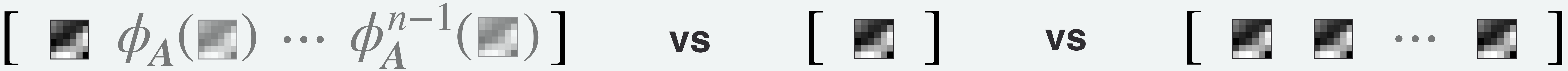


`half` (CIFAR10)



`sim` (MNIST)

Even more experiments



Knowledge transfer

MNIST	LGN	P4CNN	Single channel	Same channels
Accuracy	**98.86**	98.68	98.52	97.42

Fashion MNIST	LGN	P4CNN	Single channel	Same channels
Accuracy	89.23	88.57	85.79	**89.27**

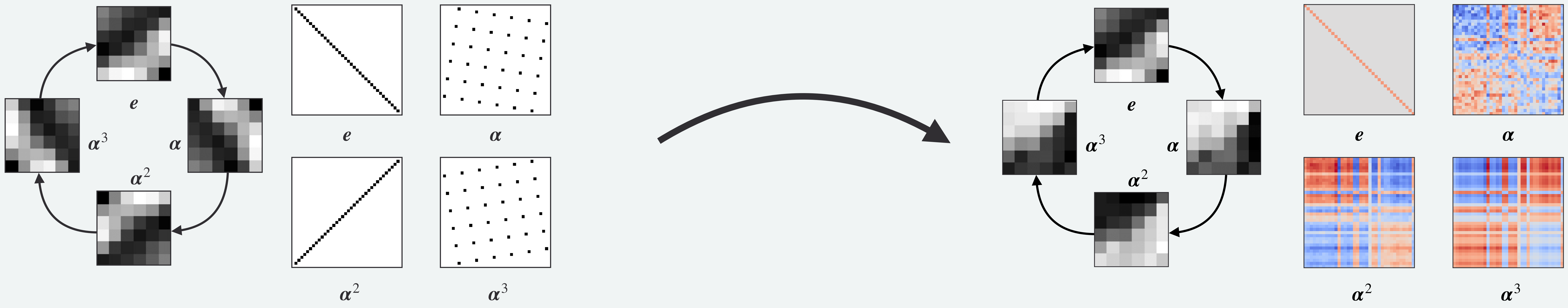
CNN comparison

	MNIST	rotMNIST	FashionMNIST	CIFAR10
LGCNN	99.27	**93.75**	**91.46**	**79.03**
P4CNN	**99.35**	92.96	91.44	75.26

(Based on ALL-CNN)

Concluding

Key takeaway



What the future holds

- non-cyclic groups
- apply to other domains
- systematic interpration

THANK YOU