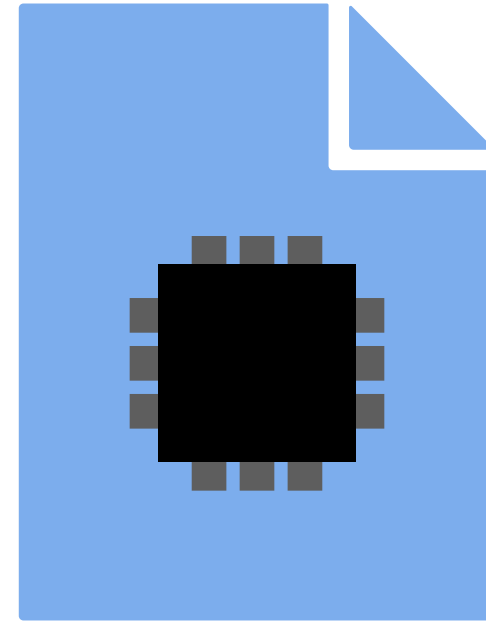


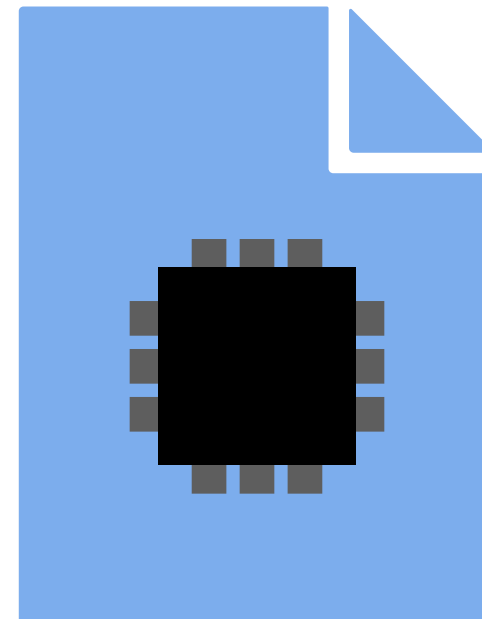
Pure Tensor Program Rewriting via Access Patterns

Gus Henry Smith, Andrew Liu, Steven Lyubomirsky, Scott Davidson, Joseph McMahan, Michael Taylor, Luis Ceze, Zachary Tatlock

University of Washington

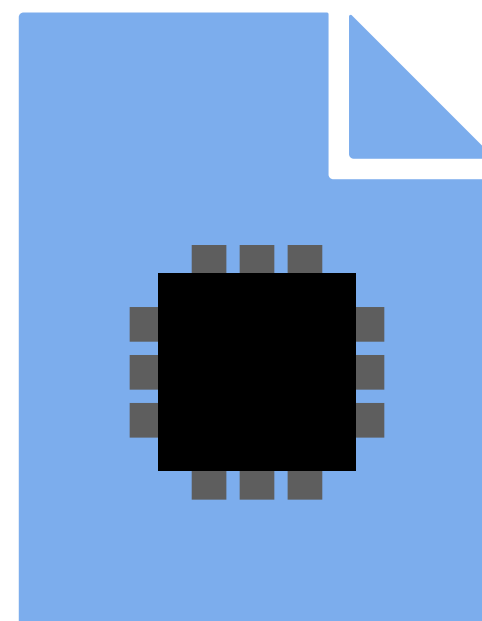
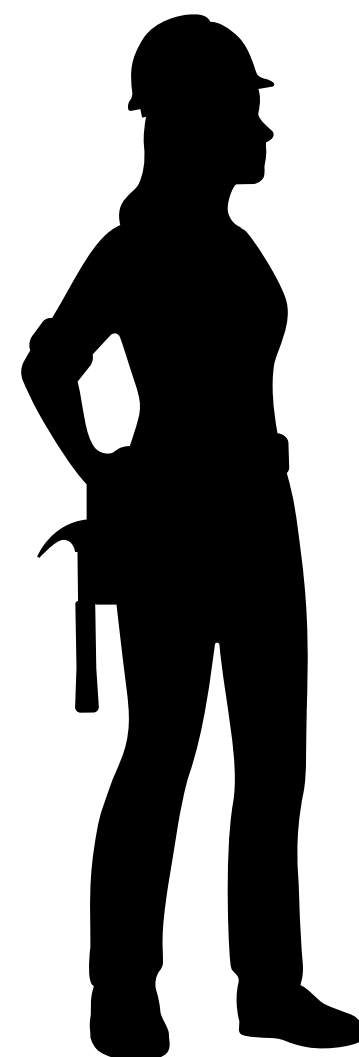


It reads an entire weight array
of shape rows by cols.



It reads an entire weight array
of shape rows by cols.

It then pushes n vectors of length
rows through the array.





It reads an entire weight array
of shape rows by cols.

It then pushes n vectors of length
rows through the array.

It computes the dot product of
every vector with every column
of the weights.



It reads an entire weight array
of shape `rows by cols`.

It then pushes `n` vectors of length
`rows` through the array.

It computes the dot product of
every vector with every column
of the weights.

Finally, it writes out `n` vectors of
length `cols`.

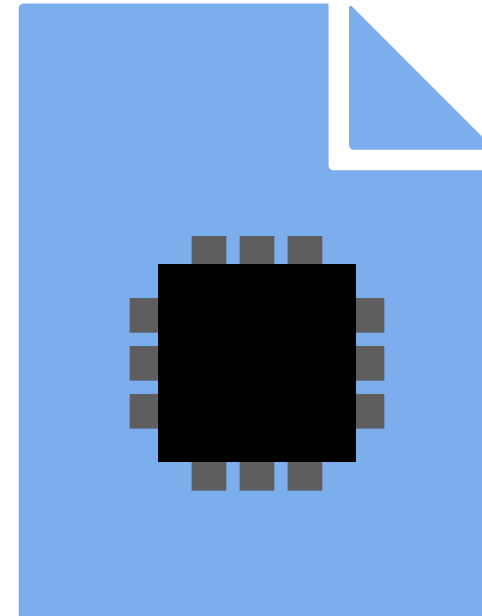
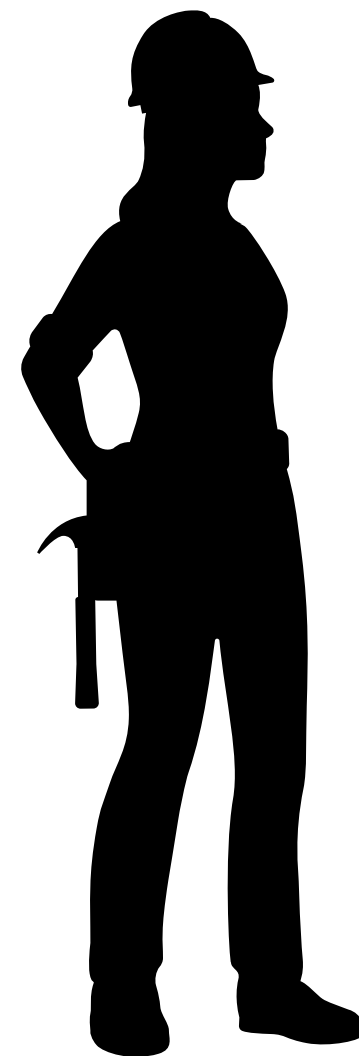
...but how do I compile to it?

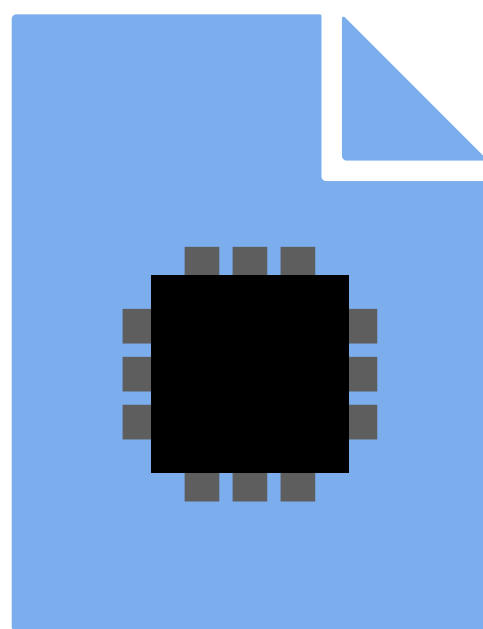
It reads an entire weight array of shape rows by cols.

It then pushes n vectors of length rows through the array.

It computes the dot product of every vector with every column of the weights.

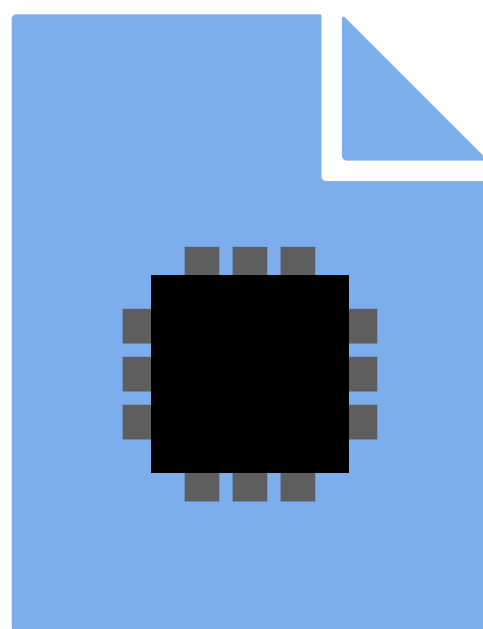
Finally, it writes out n vectors of length cols.





<custom compiler>





<custom compiler>



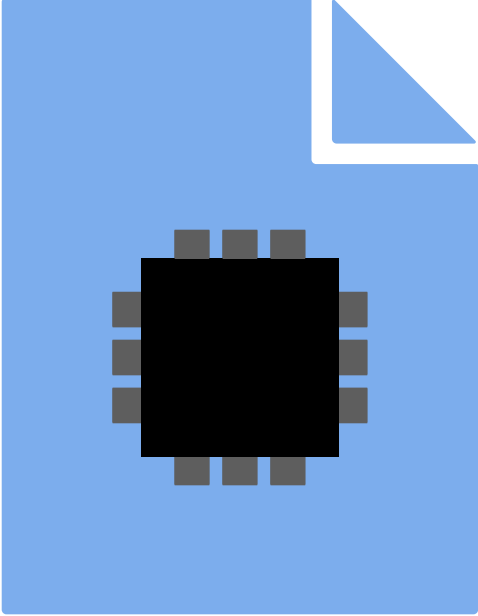
Building backends is hard, even for compiler engineers!





It reads an entire weight array
of shape rows by cols.

It then pushes n vectors of length
rows through the array.



It computes the dot product of
every vector with every column
of the weights.

Finally, it writes out n vectors of
length cols.

Given so much detail about how the
hardware functions, could a compiler
map to it automatically?



It reads an entire weight array
of shape `rows by cols`.

It then pushes `n` vectors of length
`rows` through the array.

It computes the dot product of
every vector with every column
of the weights.

Finally, it writes out `n` vectors of
length `cols`.

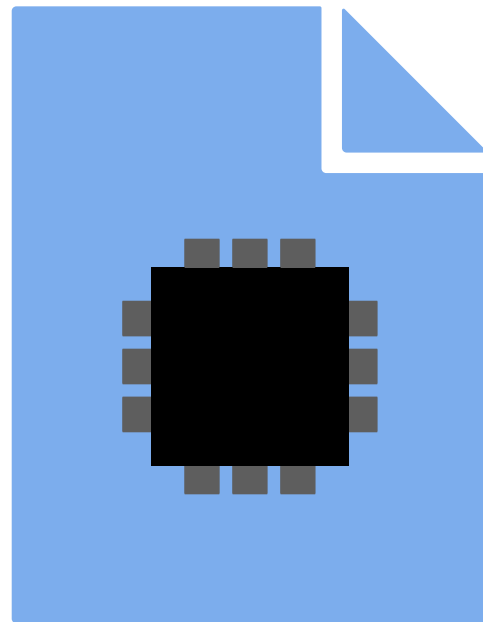
Can we compile her description of the hardware into a pattern, and search the workload for this pattern?

It reads an entire weight array of shape rows by cols.

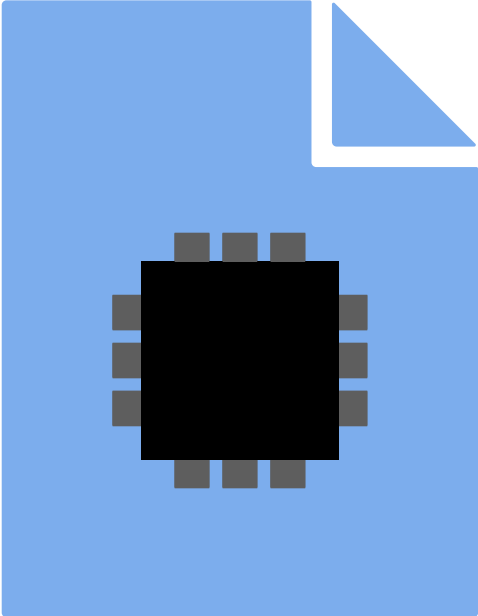
It then pushes n vectors of length rows through the array.

It computes the dot product of every vector with every column of the weights.

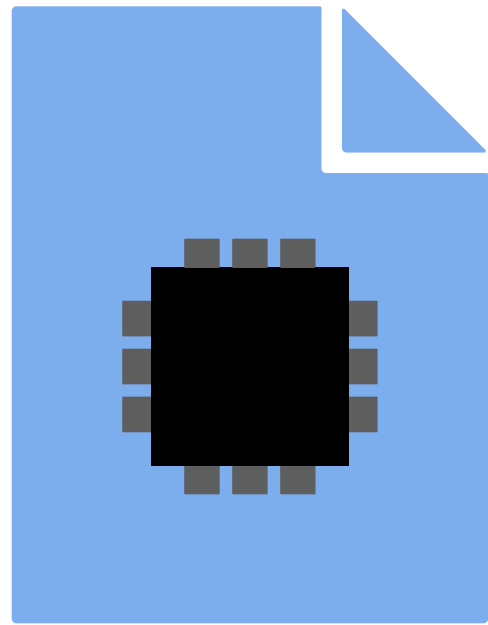
Finally, it writes out n vectors of length cols.



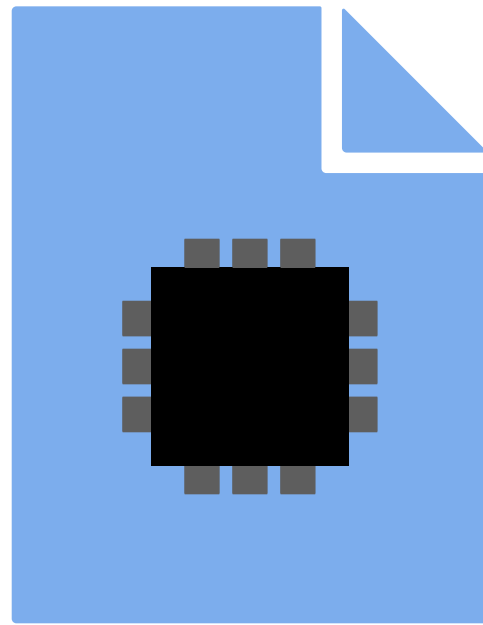
Can we compile her description of the hardware into a pattern, and search the workload for this pattern?



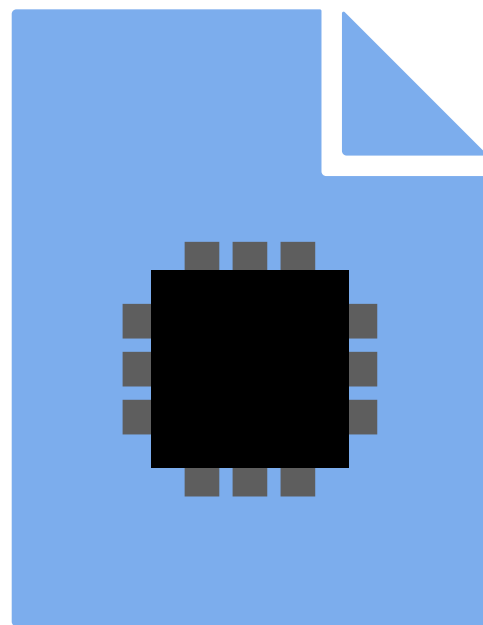
Can we compile her description of the hardware into a pattern, and search the workload for this pattern?



Can we compile her description of the hardware into a pattern, and search the workload for this pattern?



Can we compile her description of the hardware into a pattern, and search the workload for this pattern?



Hardware mapping is a
program rewriting problem!

...but current IRs are not up
to the task.

Three requirements for a hardware mapping IR:

Three requirements for a hardware mapping IR:

1. The language must be **pure**, enabling equational reasoning in term rewriting.

Three requirements for a hardware mapping IR:

1. The language must be **pure**, enabling equational reasoning in term rewriting.
2. The language must be **low-level**, letting us reason about hardware.

Three requirements for a hardware mapping IR:

1. The language must be **pure**, enabling equational reasoning in term rewriting.
2. The language must be **low-level**, letting us reason about hardware.
3. The language must **not use binding**, making term rewriting much easier.

Three requirements for a hardware mapping IR:

1. The language must be expressive enough to represent the hardware. Binding structures—such as lambdas—provide expressiveness.

2. The language must be simple enough to be mapped to hardware.

3. The language must **not use binding**, making term rewriting much easier.

Three requirements for a hardware mapping IR:

1. The language

Binding structures—such as lambdas—provide expressiveness.

However, they are difficult to deal with in term rewriting: for example, rewrites must explicitly ensure that they do not introduce name conflicts.

ing.

2. The language

3. The language must **not use binding**, making term rewriting much easier.

Three requirements for a hardware mapping IR:

1. The language

Binding structures—such as lambdas—provide expressiveness.

However, they are difficult to deal with in term rewriting: for example, rewrites must explicitly ensure that they do not introduce name conflicts.

2. The language

Thus, we seek to avoid using binding altogether!

3. The language must **not use binding**, making term rewriting much easier.

Three examples of IRs from TVM:

Pure?

Low-level?

Can avoid
binding?

Three examples of IRs from TVM:

	Pure?	Low-level?	Can avoid binding?
Relay	✓	✗	✓

Three examples of IRs from TVM:

	Pure?	Low-level?	Can avoid binding?
Relay	✓	✗	✓
TE	✓	✓	✗

Three examples of IRs from TVM:

	Pure?	Low-level?	Can avoid binding?
Relay	✓	✗	✓
TE	✓	✓	✗
TIR	✗	✓	✗

Three examples of IRs from TVM:

	Pure?	Low-level?	Can avoid binding?
Relay	✓	✗	✓
TE	✓	✓	✗
TIR	✗	✓	✗



Three examples of IRs from TVM:



	Pure?	Low-level?	Can avoid binding?
Relay	✓	Current tensor IRs fall short on our requirements!	
TE	✓	✓	✗
TIR	✗	✓	✗

We present our core abstraction, **access patterns**.

We present our core abstraction, **access patterns**.

Around them, we design **Glenside**, a pure, low-level,
binder-free tensor IR.

We present our core abstraction, **access patterns**.

Around them, we design **Glenside**, a pure, low-level,
binder-free tensor IR.

Finally, we demonstrate how Glenside **enables low-level
tensor program rewriting**.



Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- Case Studies
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation



Outline

- **Motivating Example: Matrix Multiplication**
- Access Pattern Definition
- Case Studies
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation

We want to represent matrix multiplication in a way that

We want to represent matrix multiplication in a way that

1. is pure,

We want to represent matrix multiplication in a way that

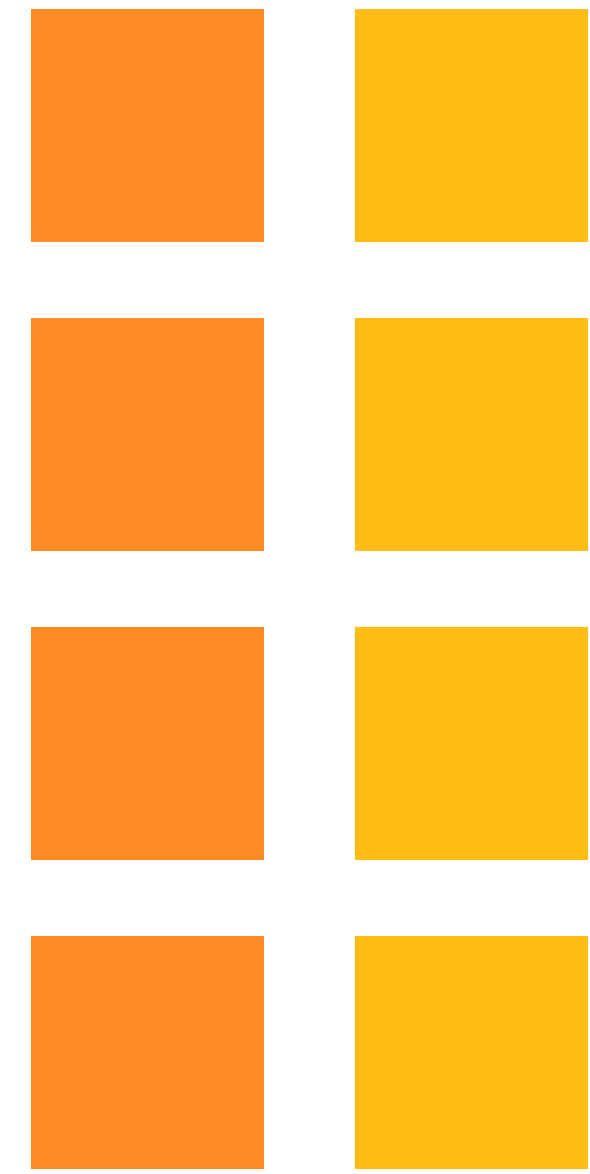
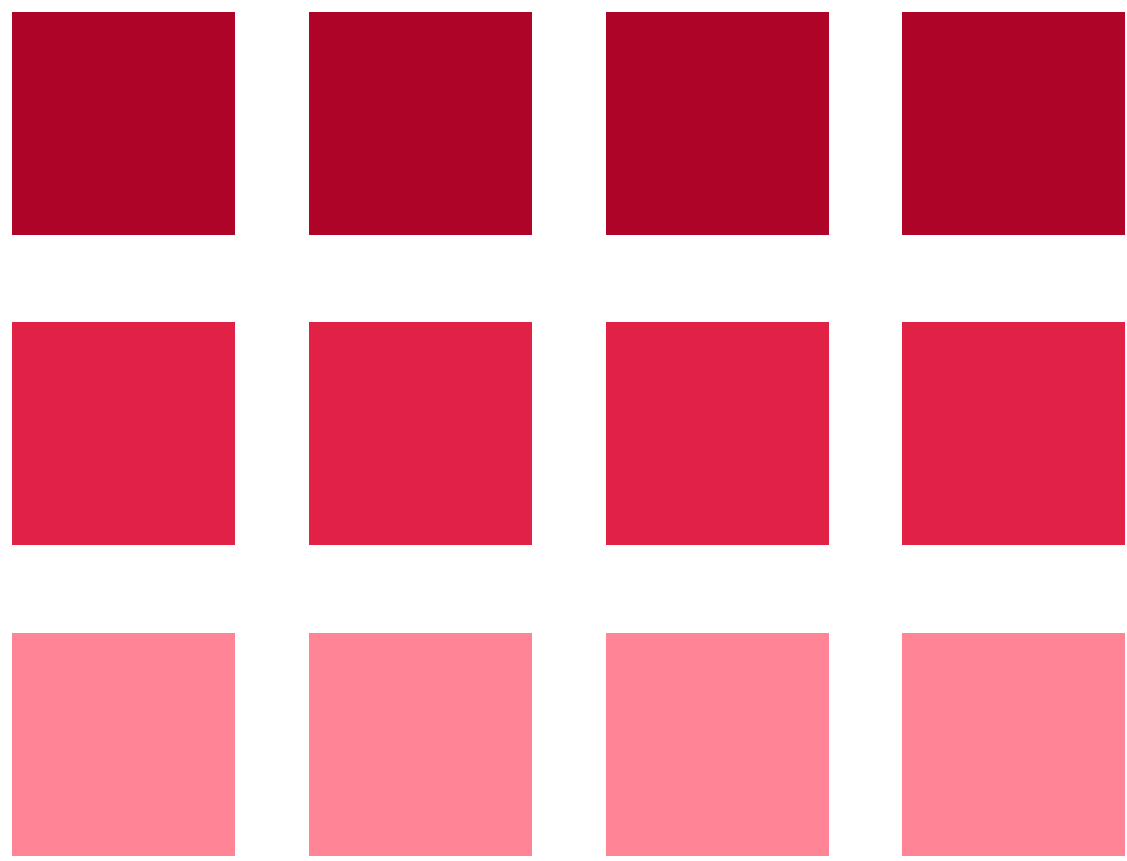
1. is pure,

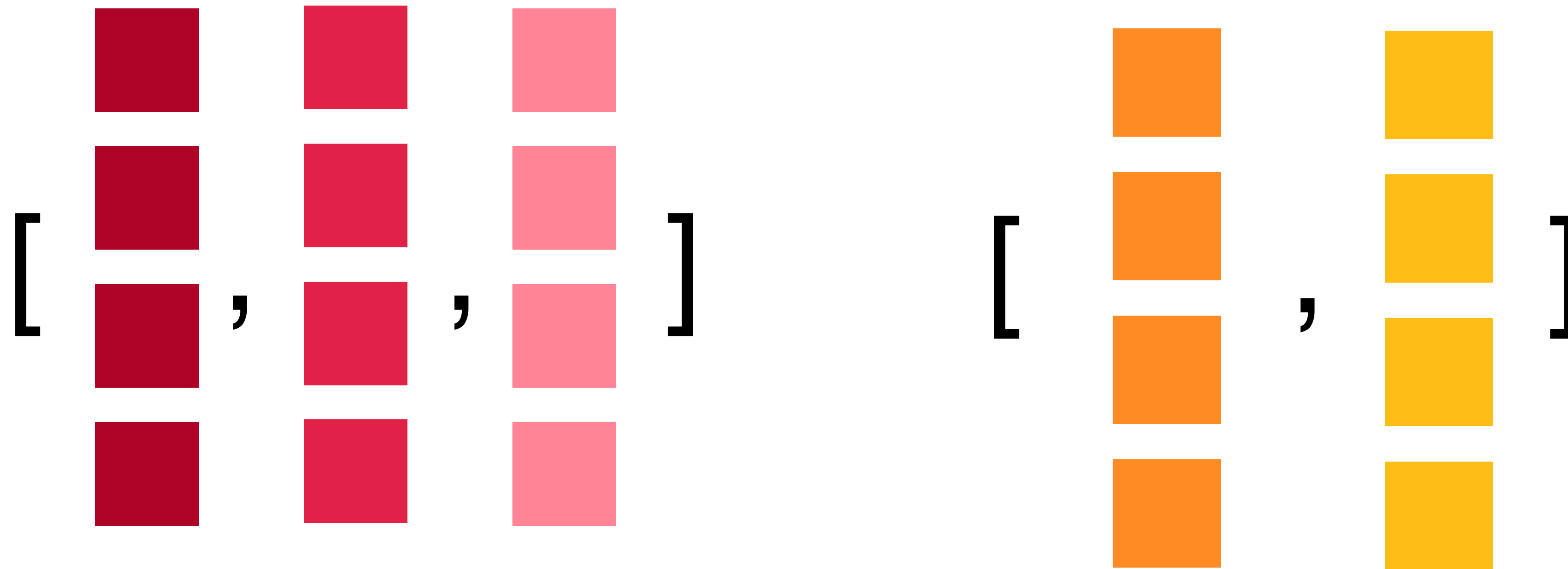
2. is low-level, and

We want to represent matrix multiplication in a way that

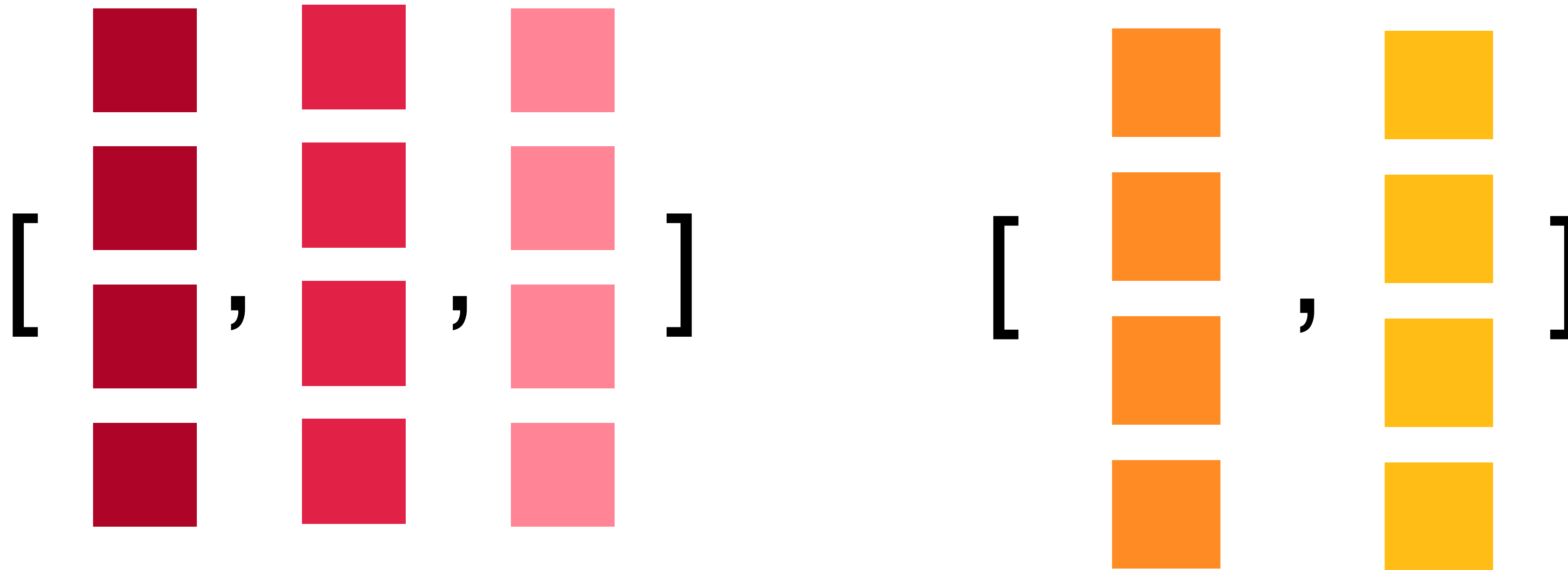
1. is pure,
2. is low-level, and
3. avoids binding.

Given matrices A and B , pair each row of A with each column of B , compute their dot products, and arrange the results back into a matrix.

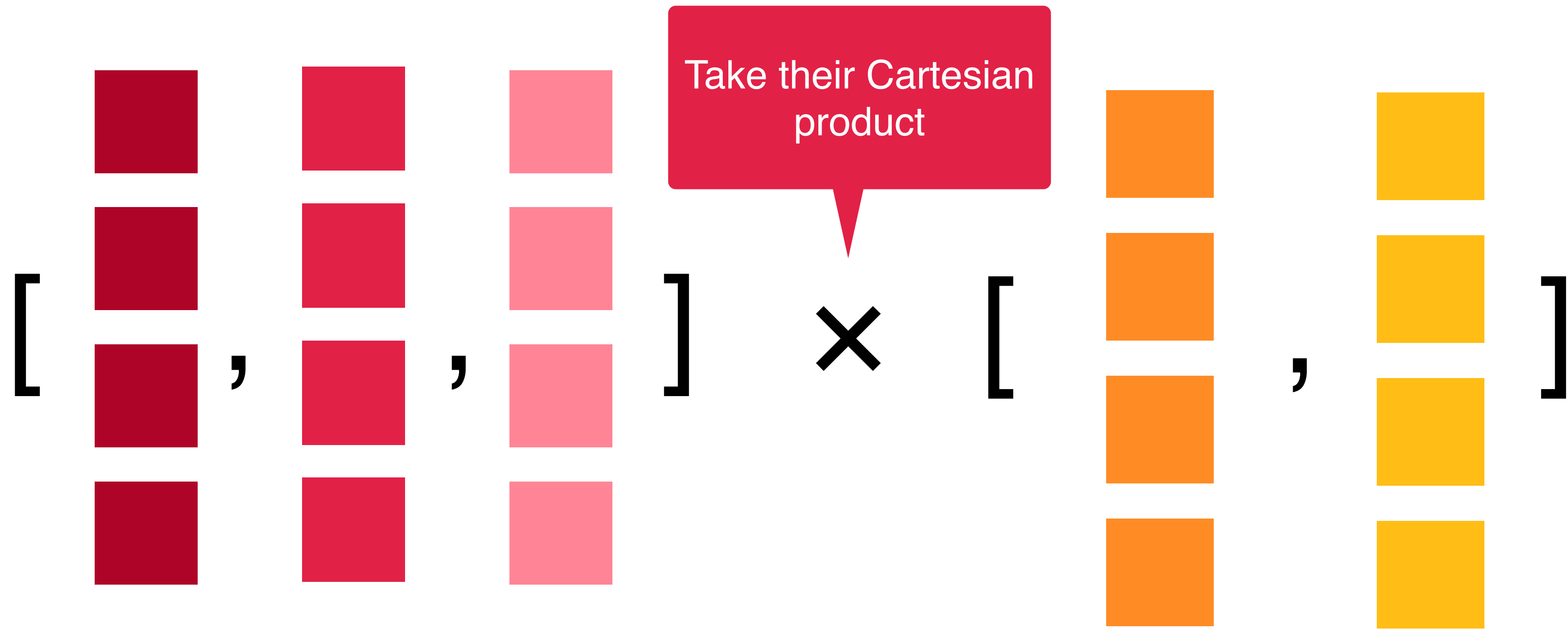


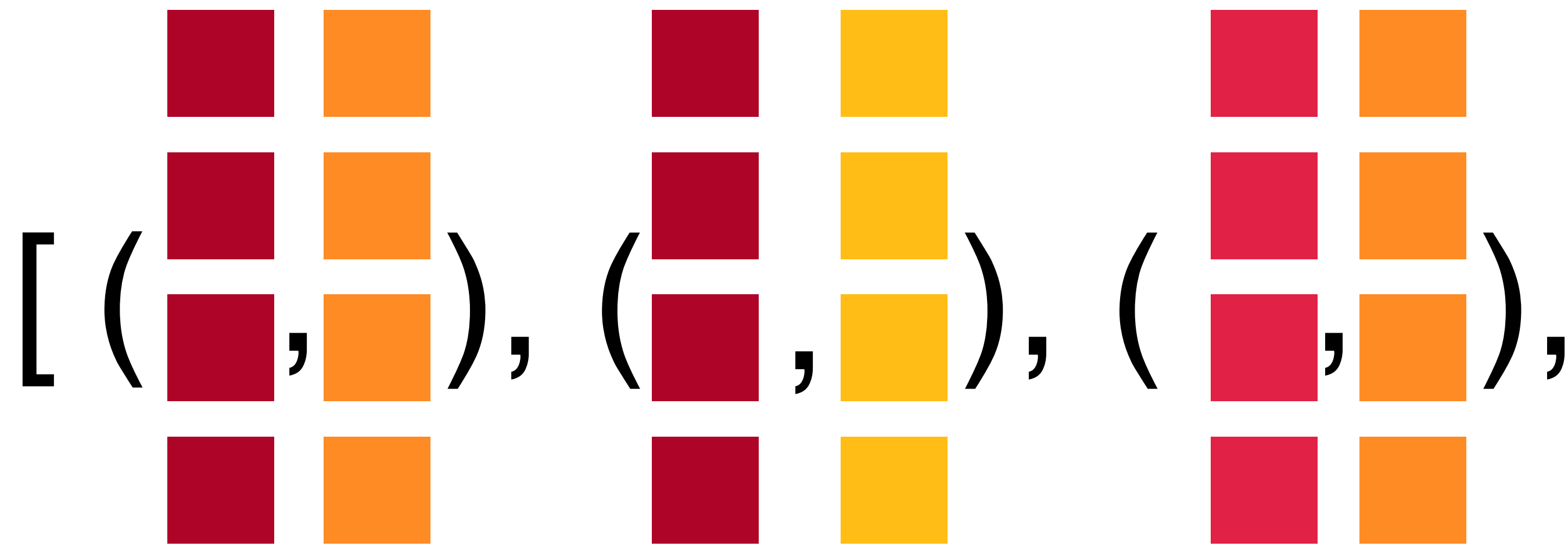


View matrices as lists of rows/columns

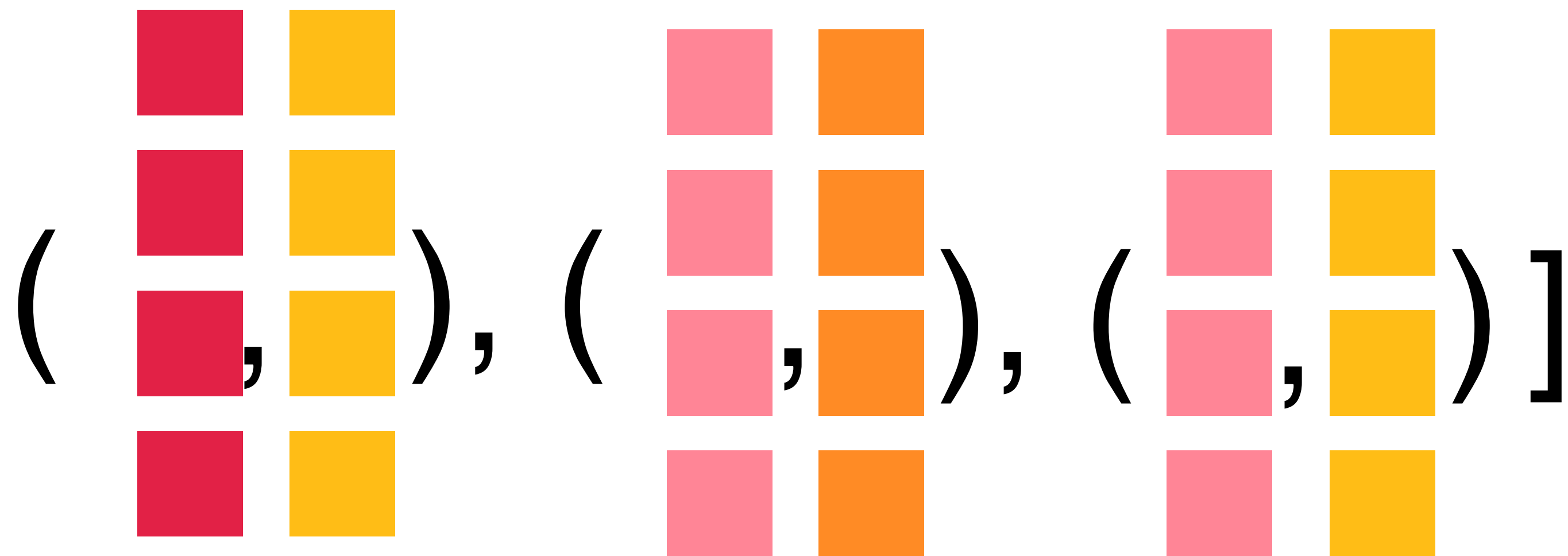


View matrices as lists of rows/columns



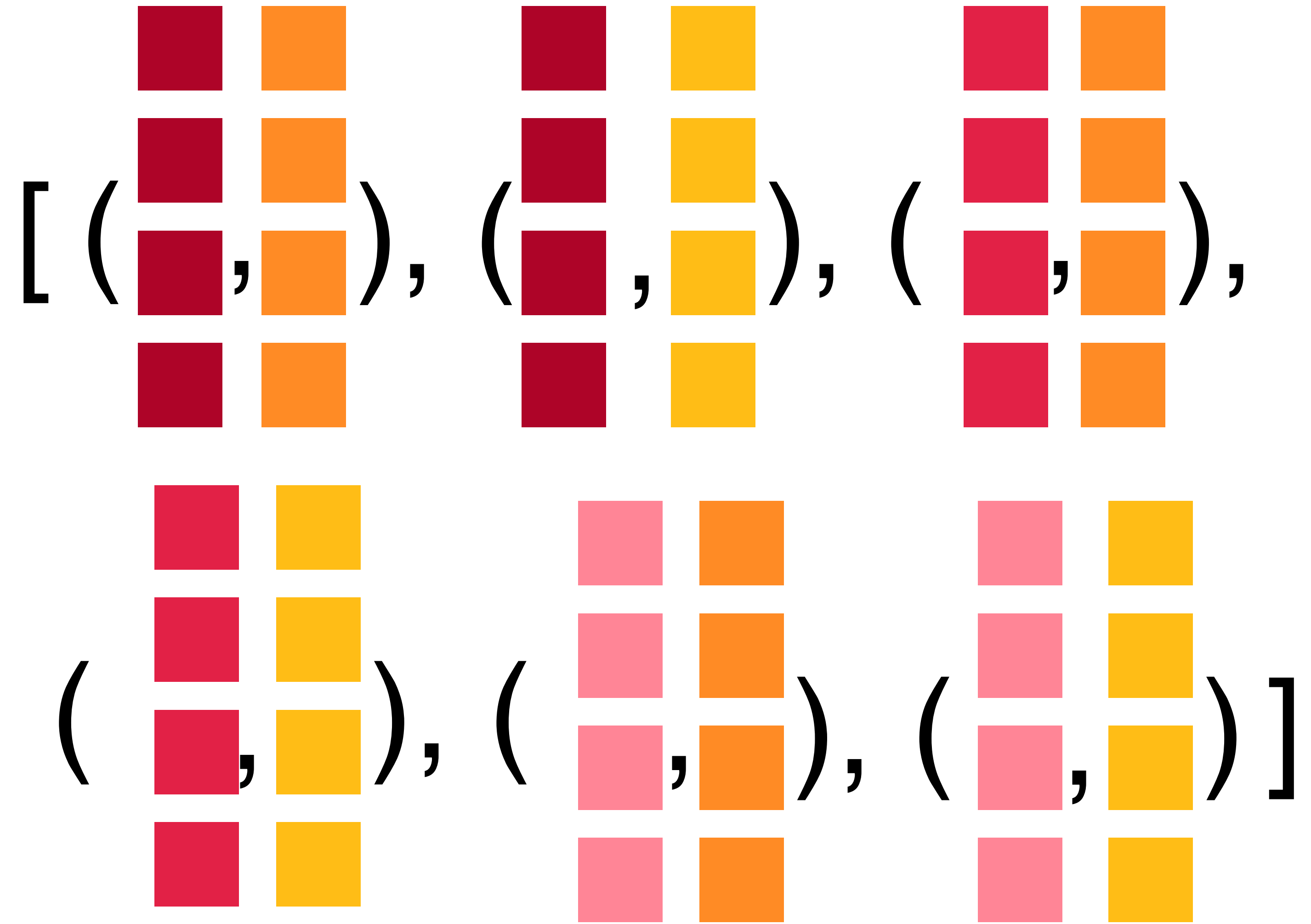


Every row paired with every column



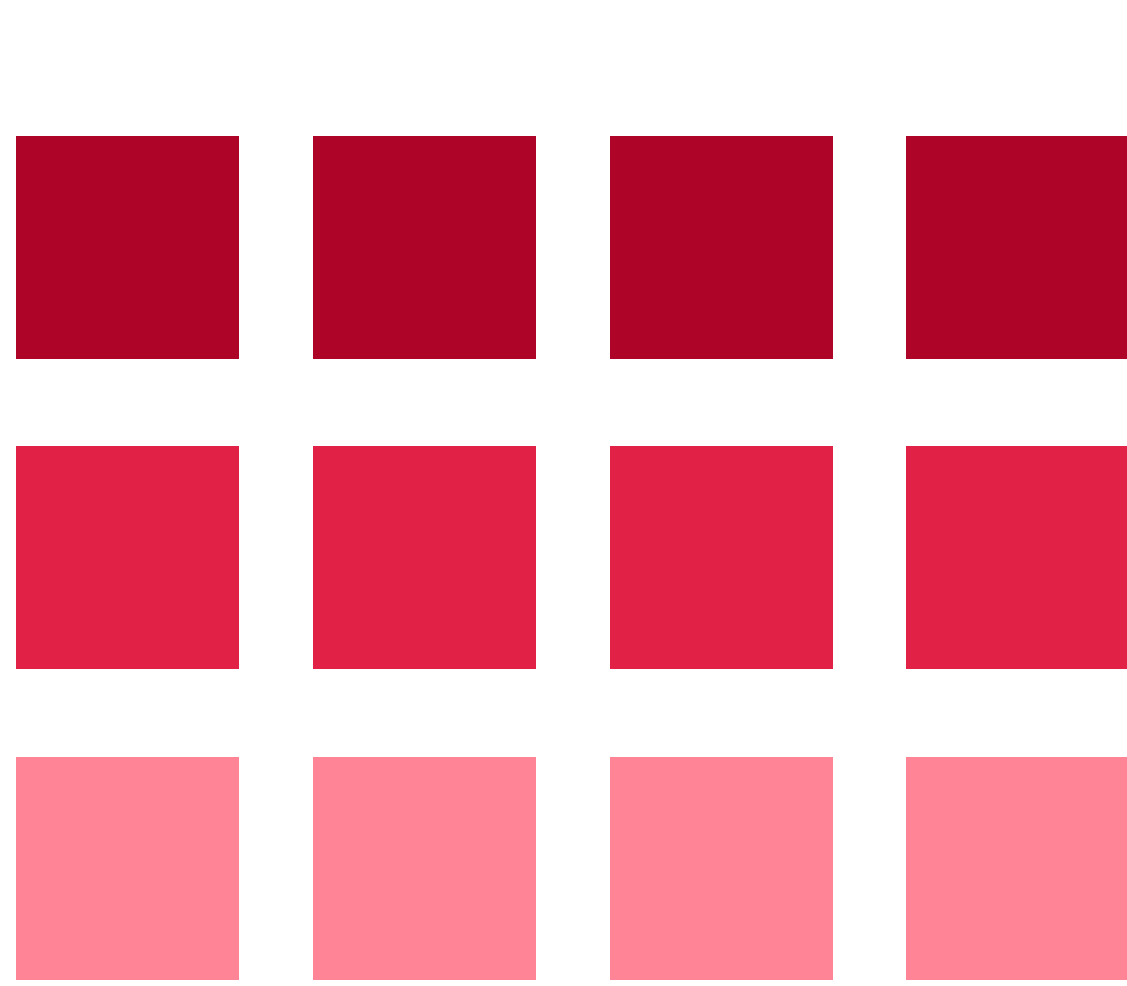
map dotProd

Map dot product operator over every row-column pair



[■, ■, ■,
■, ■, ■]

But there's a problem!



×



=

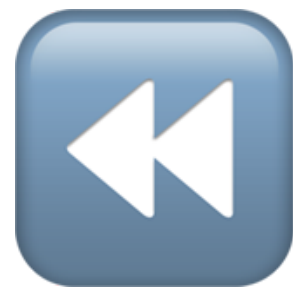


≠

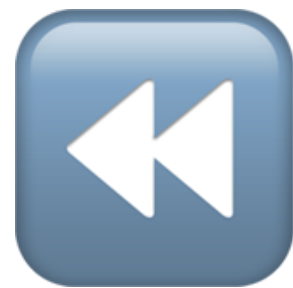
The values are correct,
but the shape is
missing!

[
■, ■, ■,
■, ■, ■]

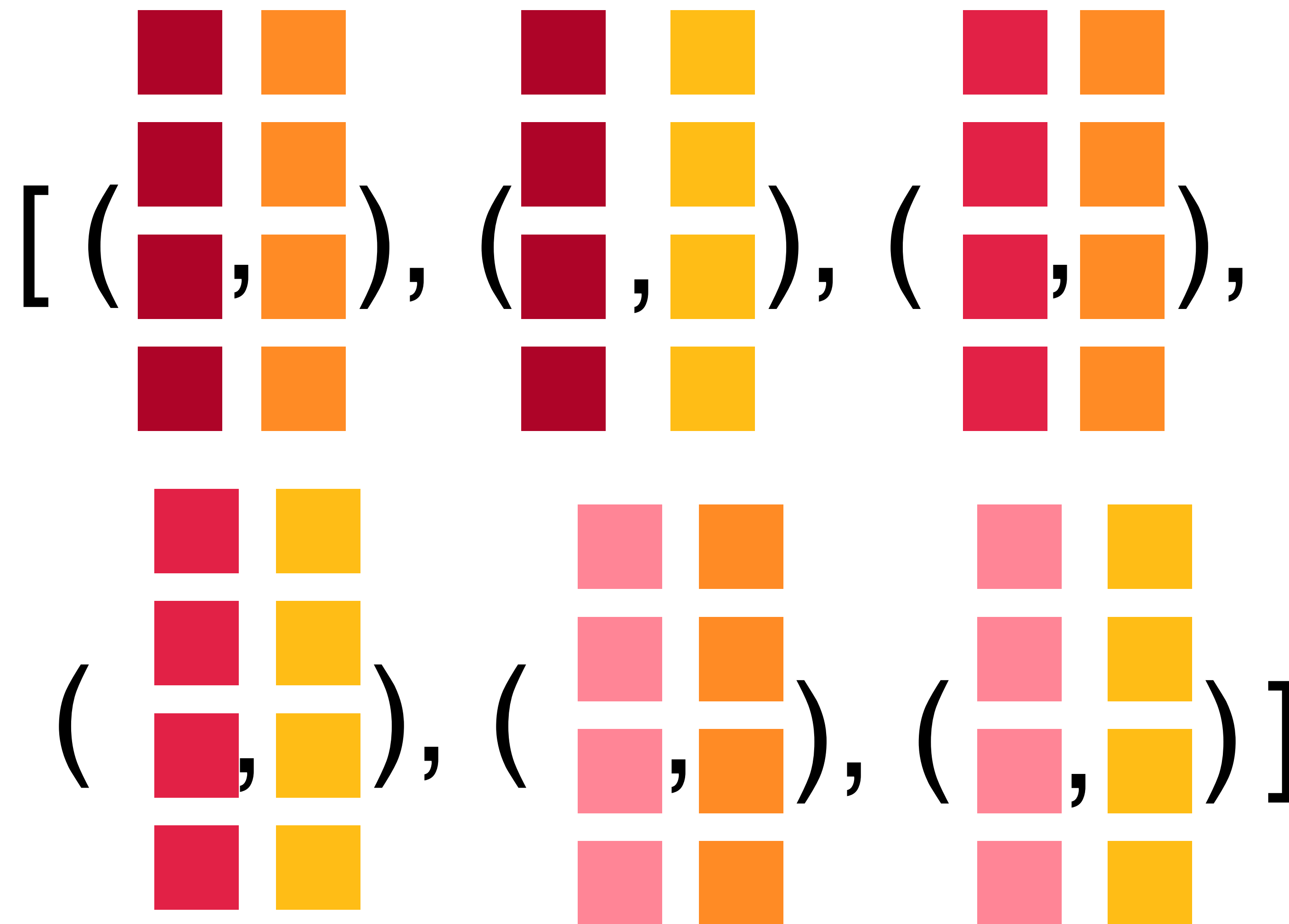
[■, ■, ■,
■, ■, ■]

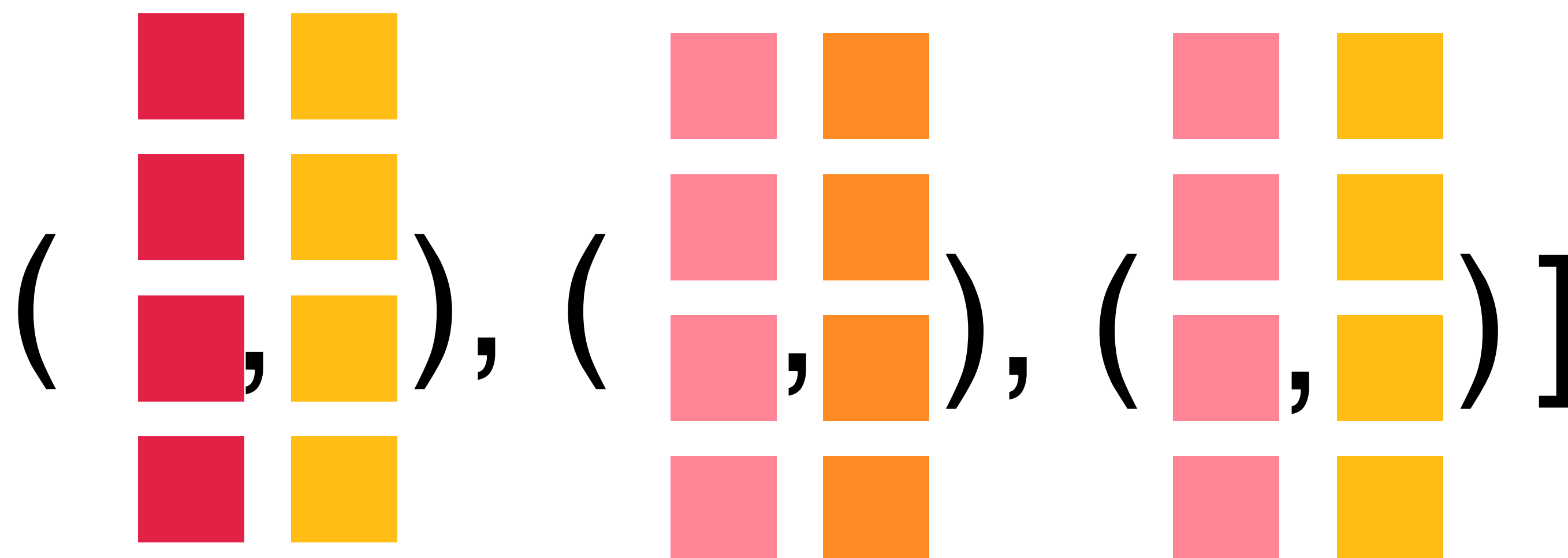
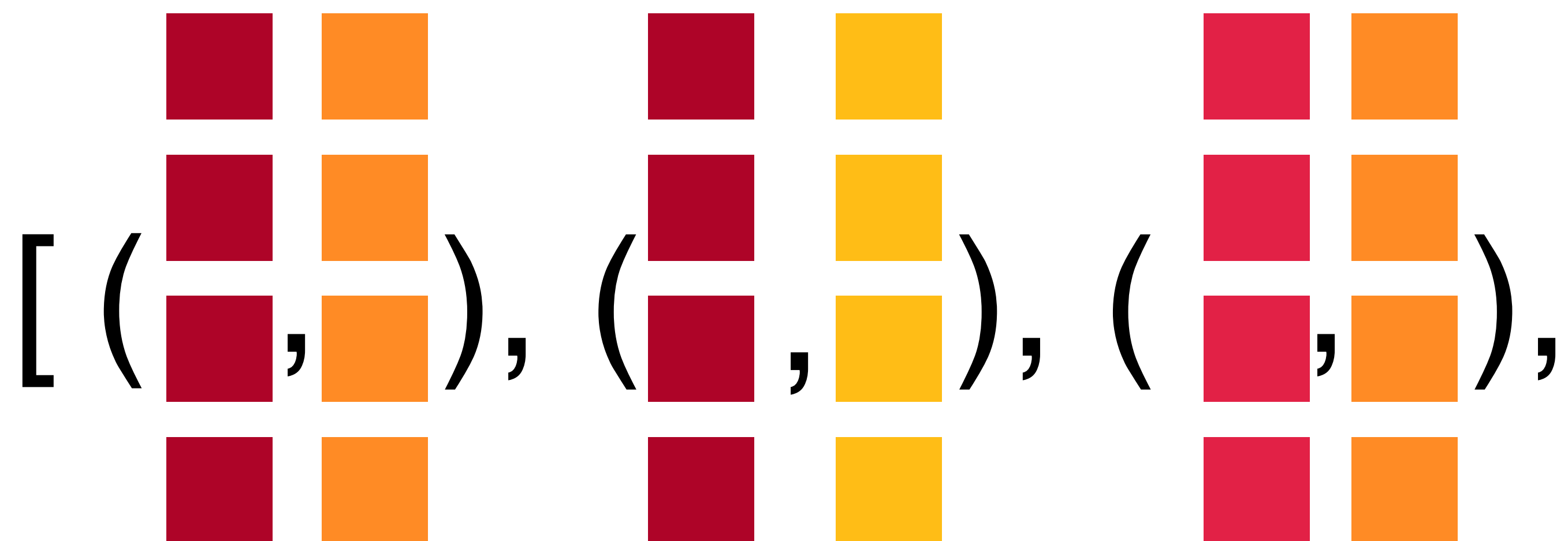
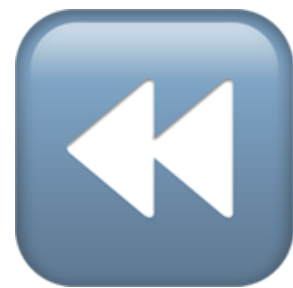


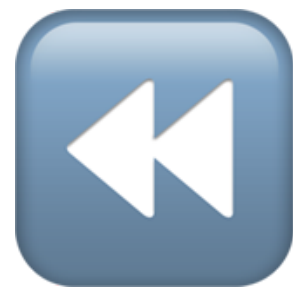
[■, ■, ■,
 ■, ■, ■]



map dot-product







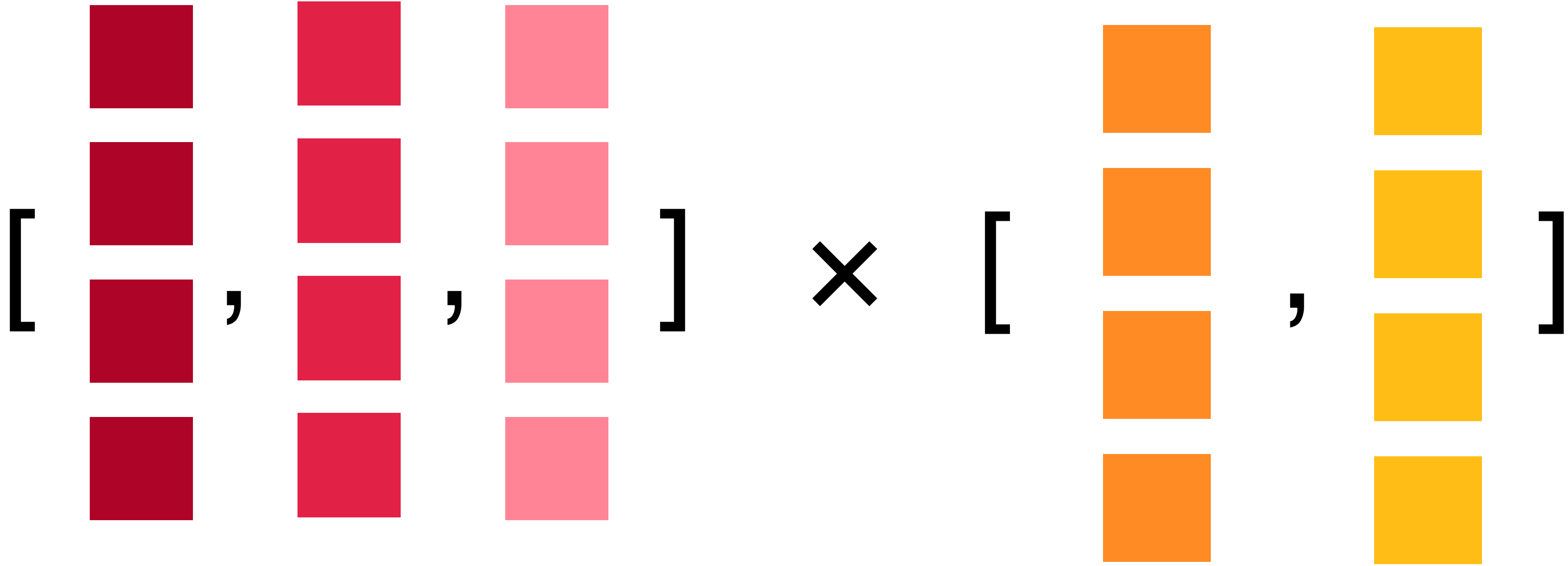
$$\begin{bmatrix} \text{dark red} & \text{red} & \text{light pink} \\ \text{dark red} & \text{red} & \text{light pink} \\ \text{dark red} & \text{red} & \text{light pink} \\ \text{dark red} & \text{red} & \text{light pink} \end{bmatrix} \times \begin{bmatrix} \text{orange} & \text{yellow} \\ \text{orange} & \text{yellow} \\ \text{orange} & \text{yellow} \end{bmatrix}$$



$$\begin{bmatrix} \text{dark red square} & \text{red square} & \text{light red square} \\ \text{dark red square} & \text{red square} & \text{light red square} \\ \text{dark red square} & \text{red square} & \text{light red square} \\ \text{dark red square} & \text{red square} & \text{light red square} \end{bmatrix} \times \begin{bmatrix} \text{orange square} & \text{yellow square} \\ \text{orange square} & \text{yellow square} \\ \text{orange square} & \text{yellow square} \\ \text{orange square} & \text{yellow square} \end{bmatrix}$$

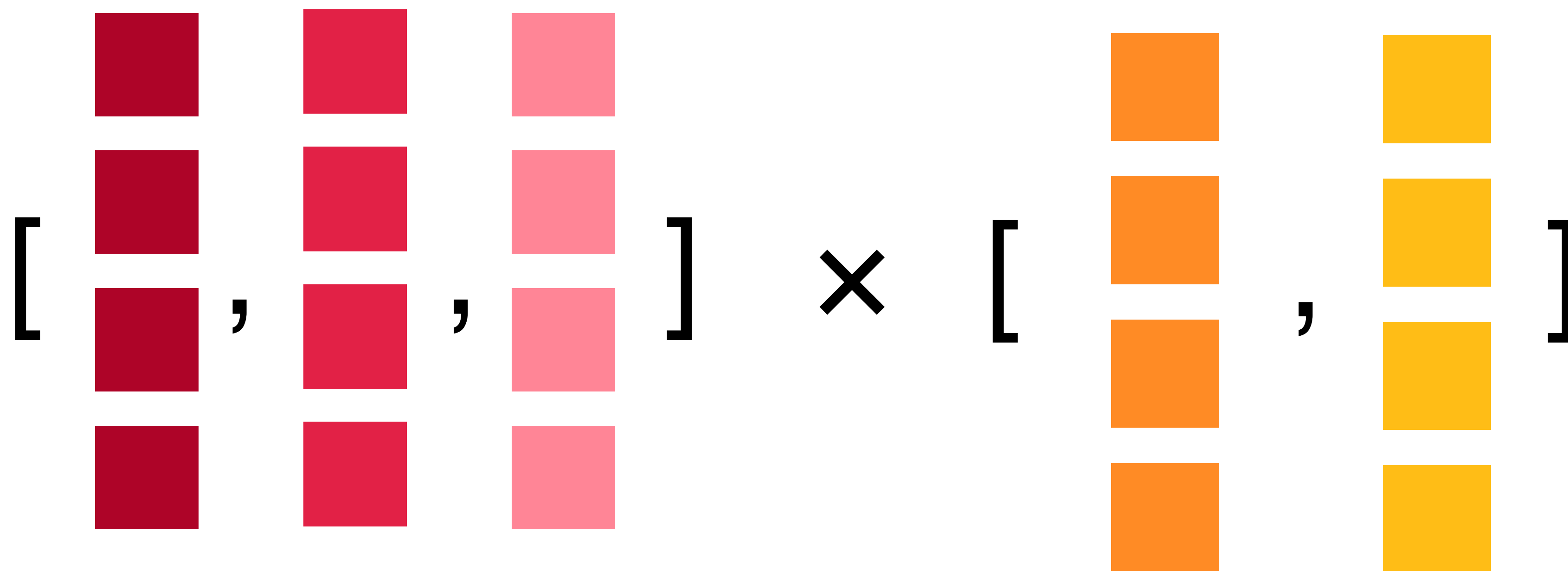


Shape information is present here...





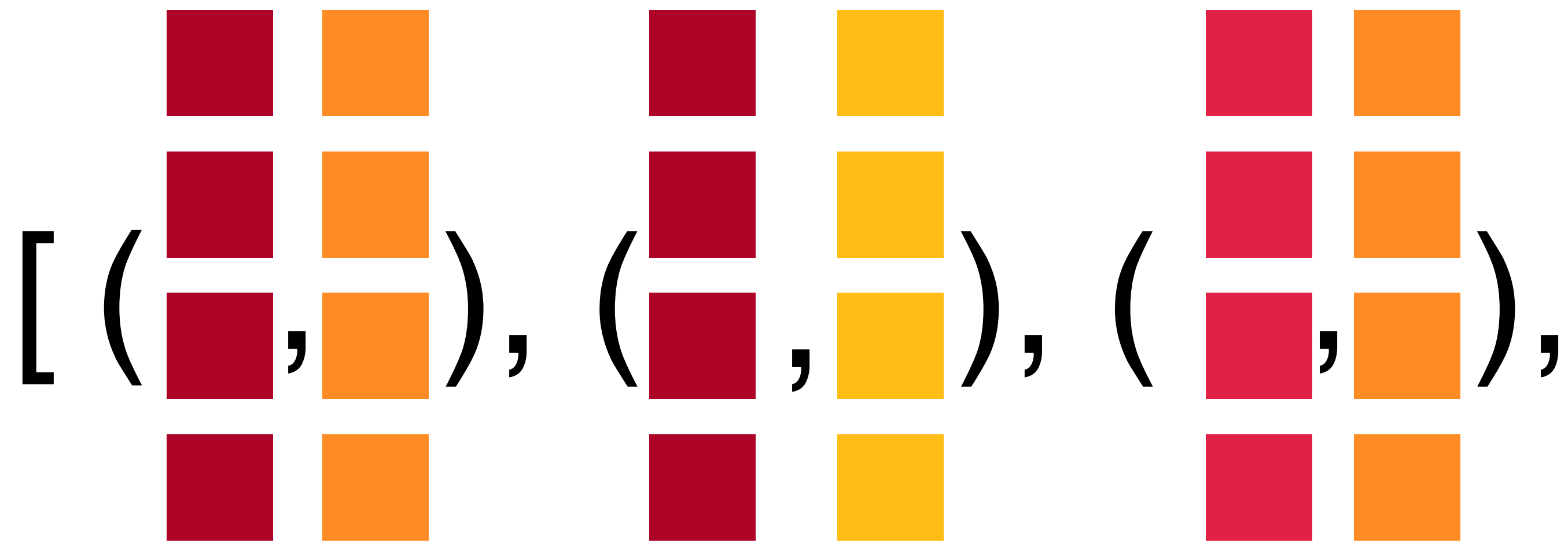
Shape information is present here...



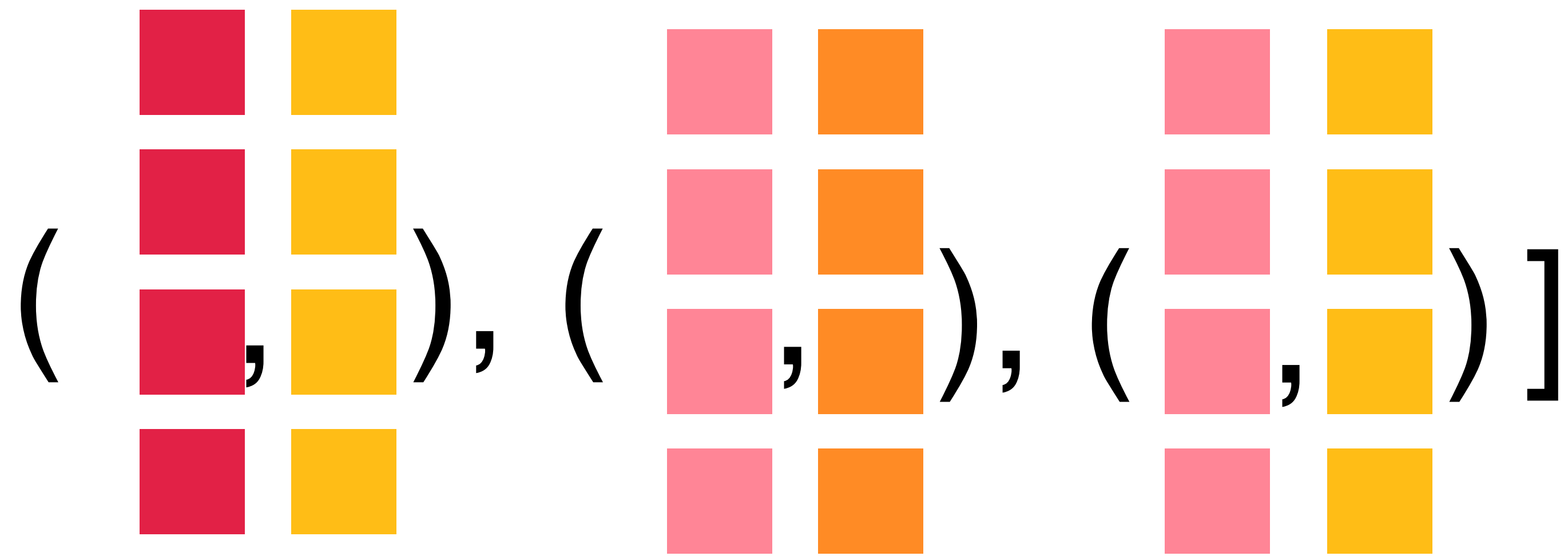


$\left[\left(\begin{array}{cc} \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \end{array} \right), \left(\begin{array}{cc} \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \end{array} \right), \left(\begin{array}{cc} \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \\ \color{red}\blacksquare & \color{orange}\blacksquare \end{array} \right), \right.$

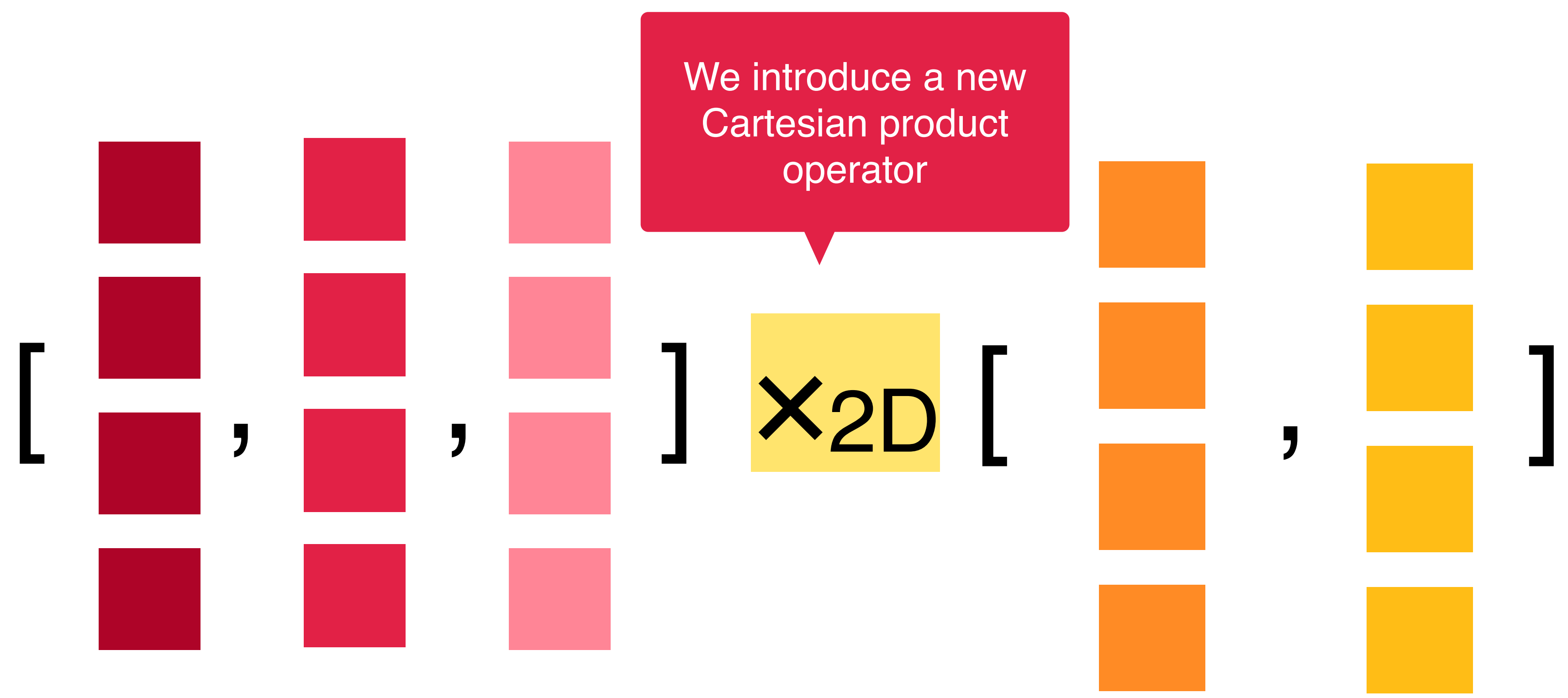
$\left. \left(\begin{array}{cc} \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \\ \color{red}\blacksquare & \color{yellow}\blacksquare \end{array} \right), \left(\begin{array}{cc} \color{pink}\blacksquare & \color{orange}\blacksquare \\ \color{pink}\blacksquare & \color{orange}\blacksquare \\ \color{pink}\blacksquare & \color{orange}\blacksquare \\ \color{pink}\blacksquare & \color{orange}\blacksquare \end{array} \right), \left(\begin{array}{cc} \color{pink}\blacksquare & \color{yellow}\blacksquare \\ \color{pink}\blacksquare & \color{yellow}\blacksquare \\ \color{pink}\blacksquare & \color{yellow}\blacksquare \\ \color{pink}\blacksquare & \color{yellow}\blacksquare \end{array} \right) \right]$

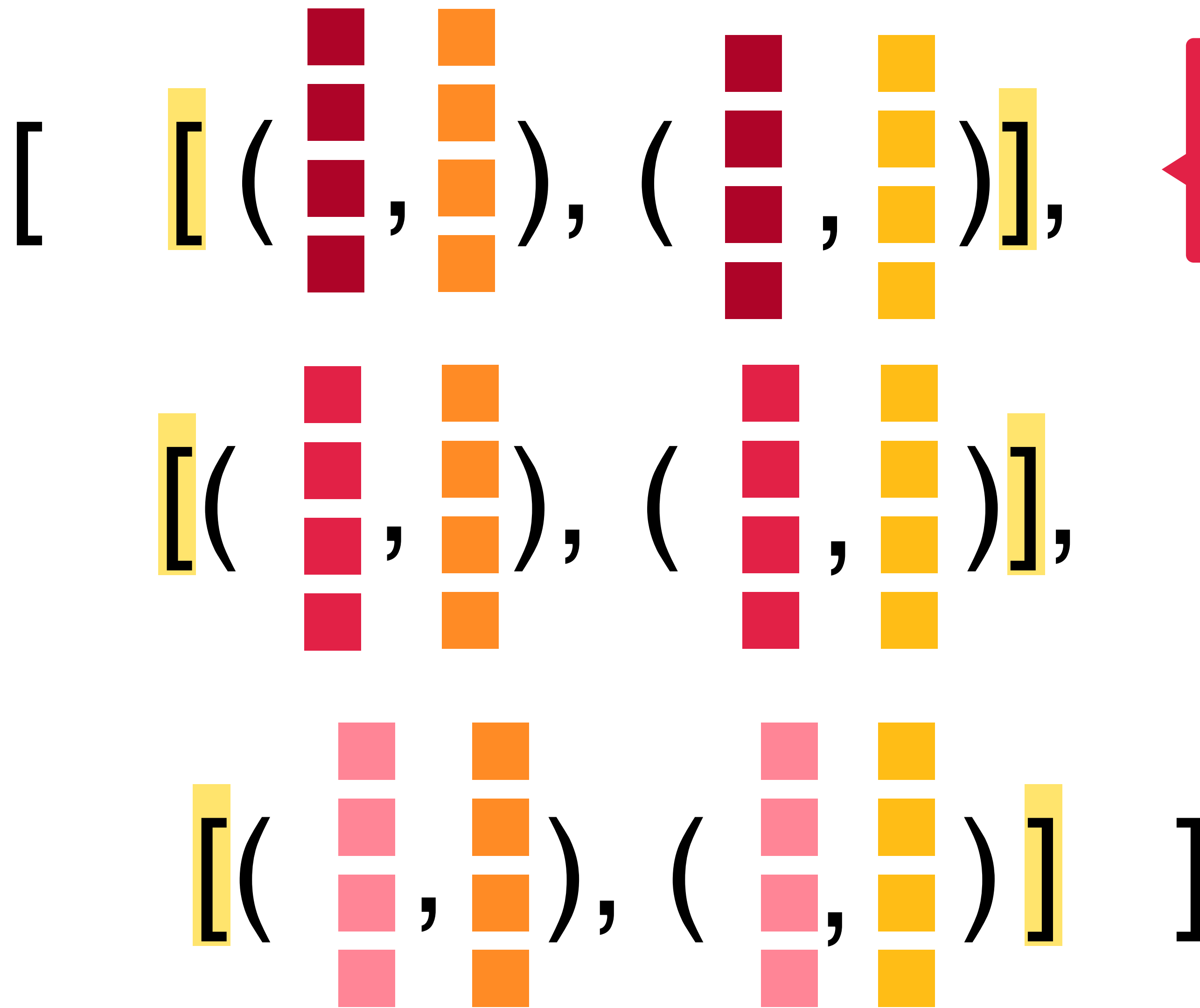


...but absent here!



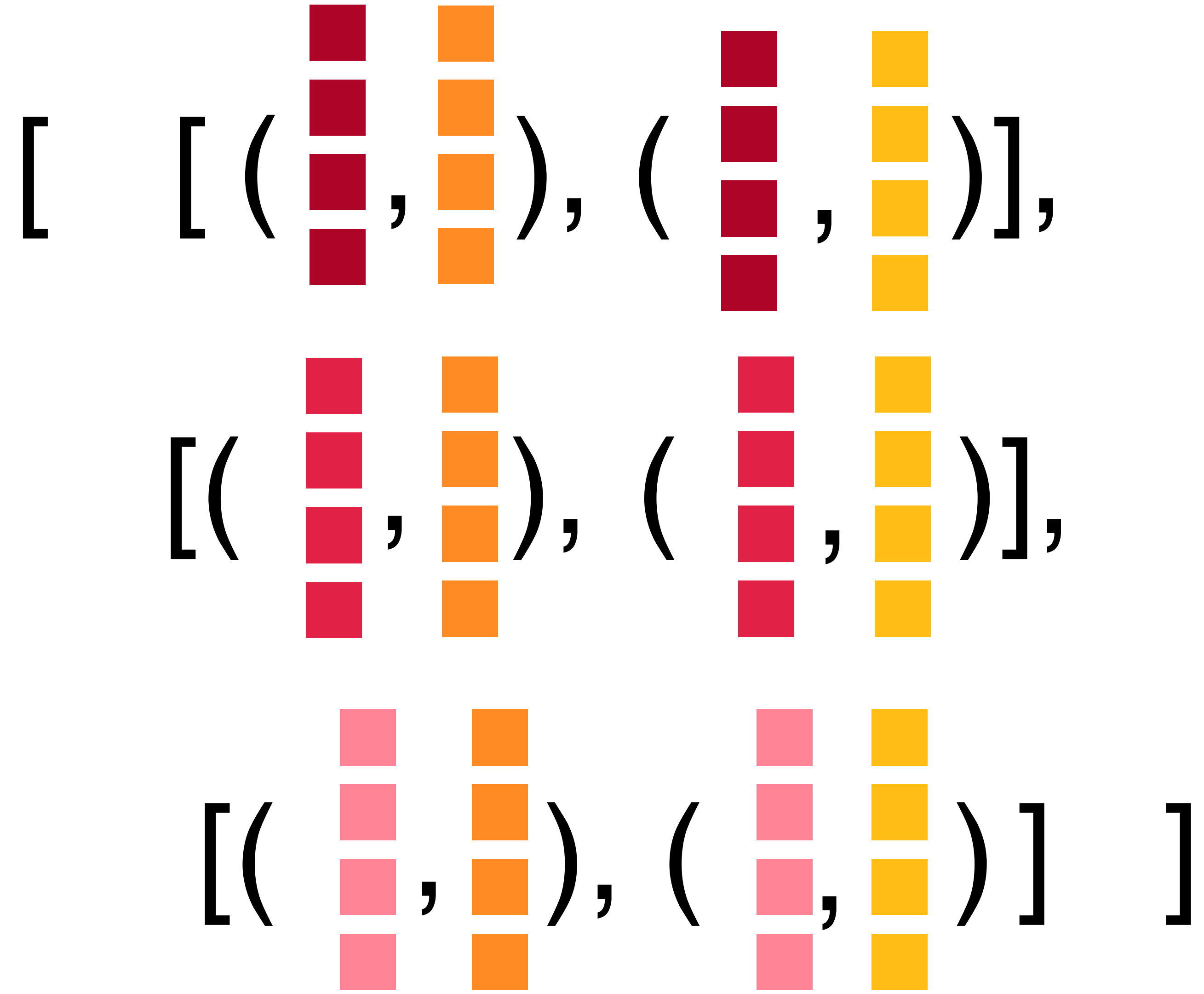
Cartesian product destroys
our shape information!





2D Cartesian product operator preserves shape info

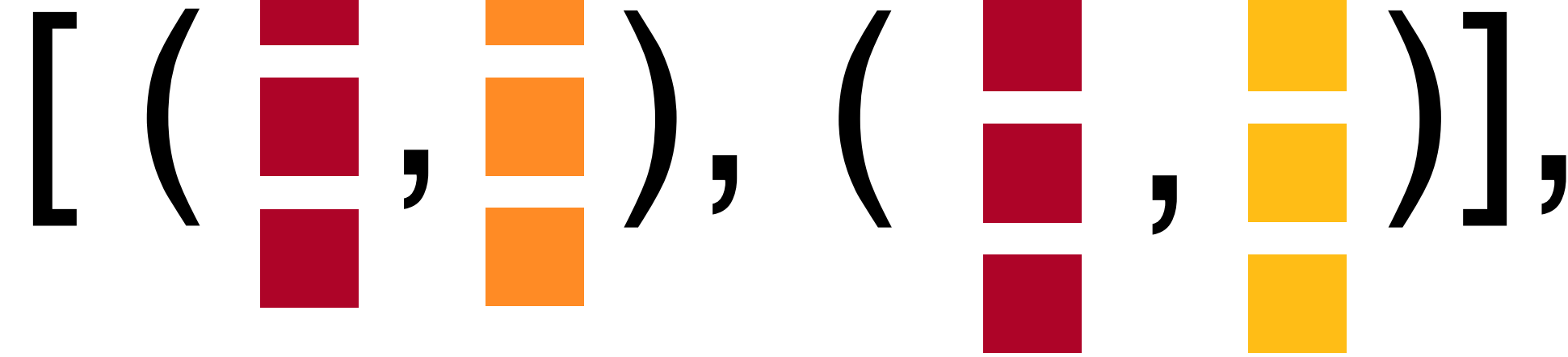
map dotProd



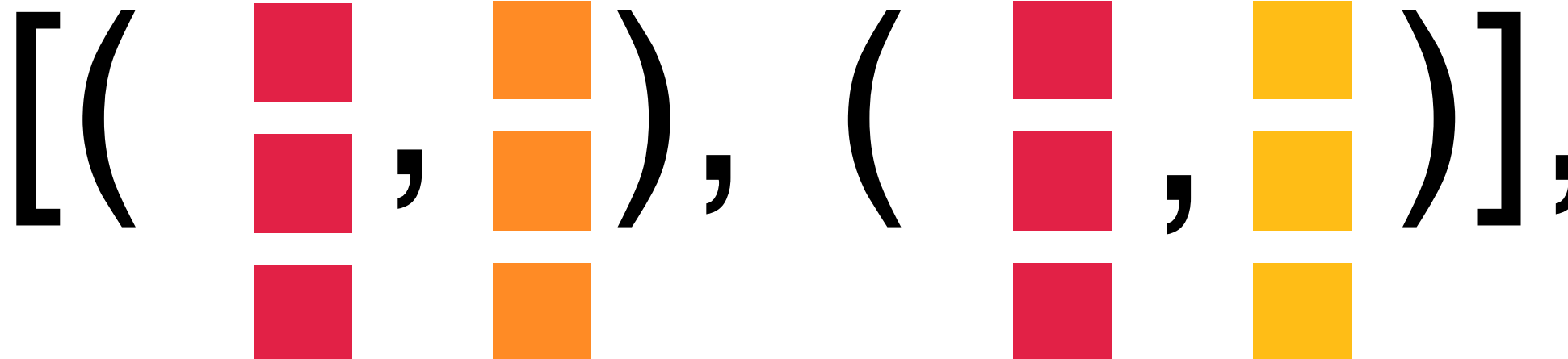


[

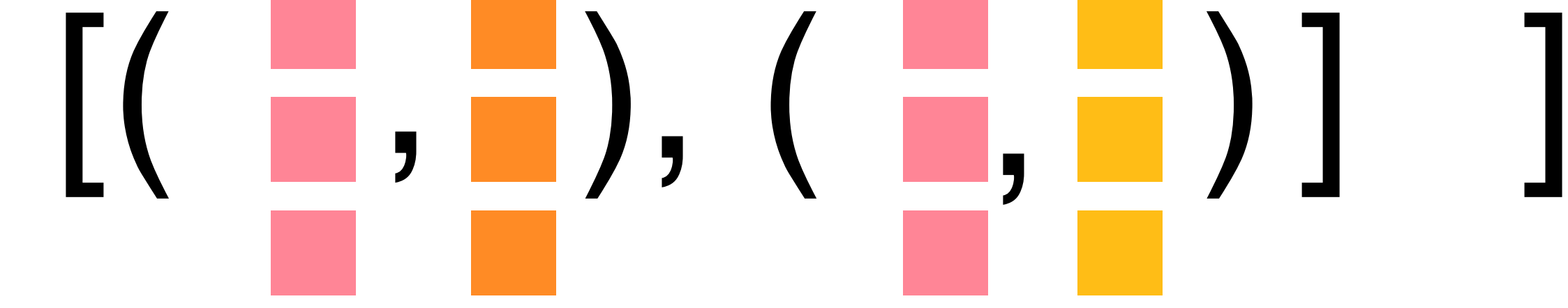
dotProd



dotProd



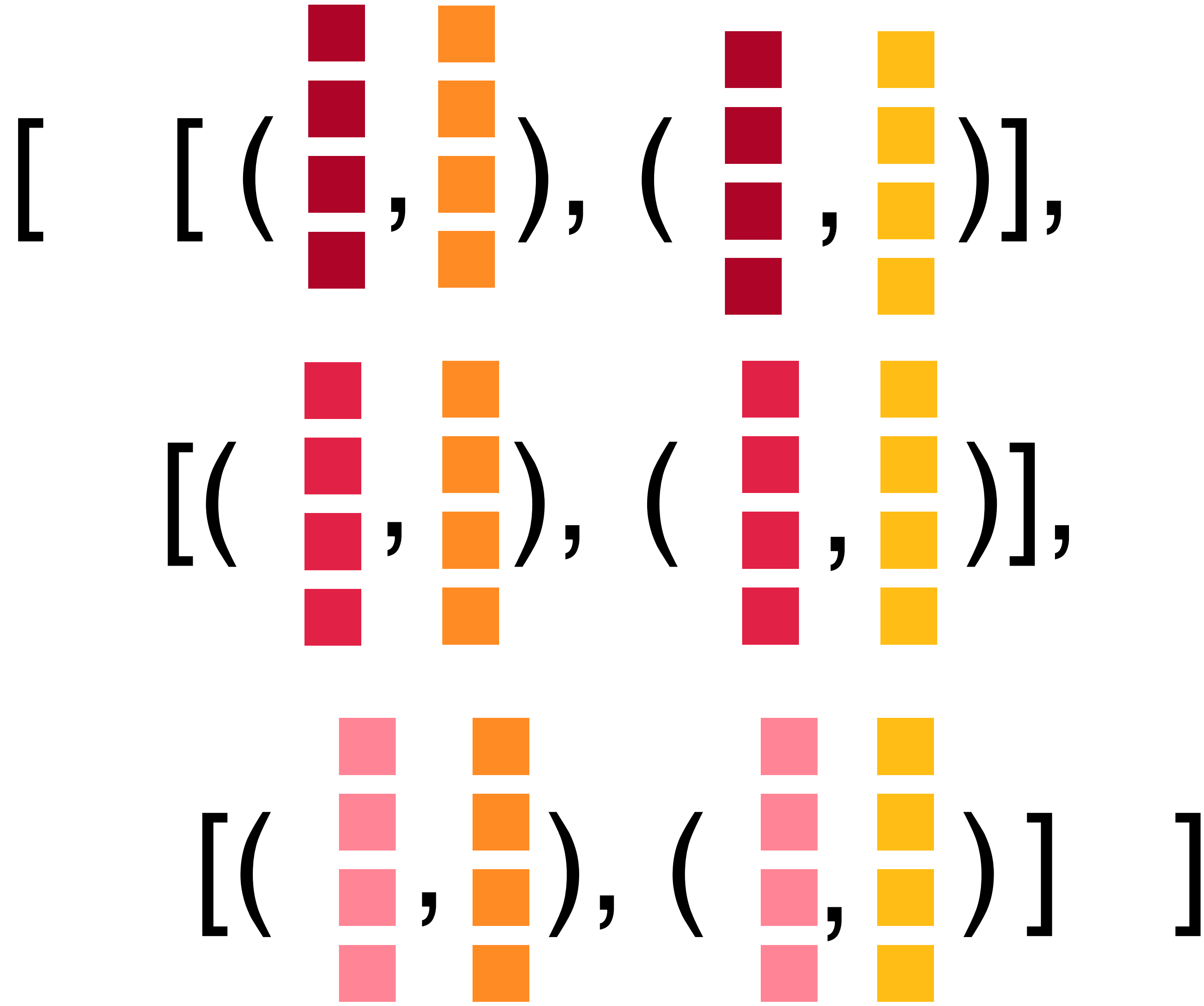
dotProd

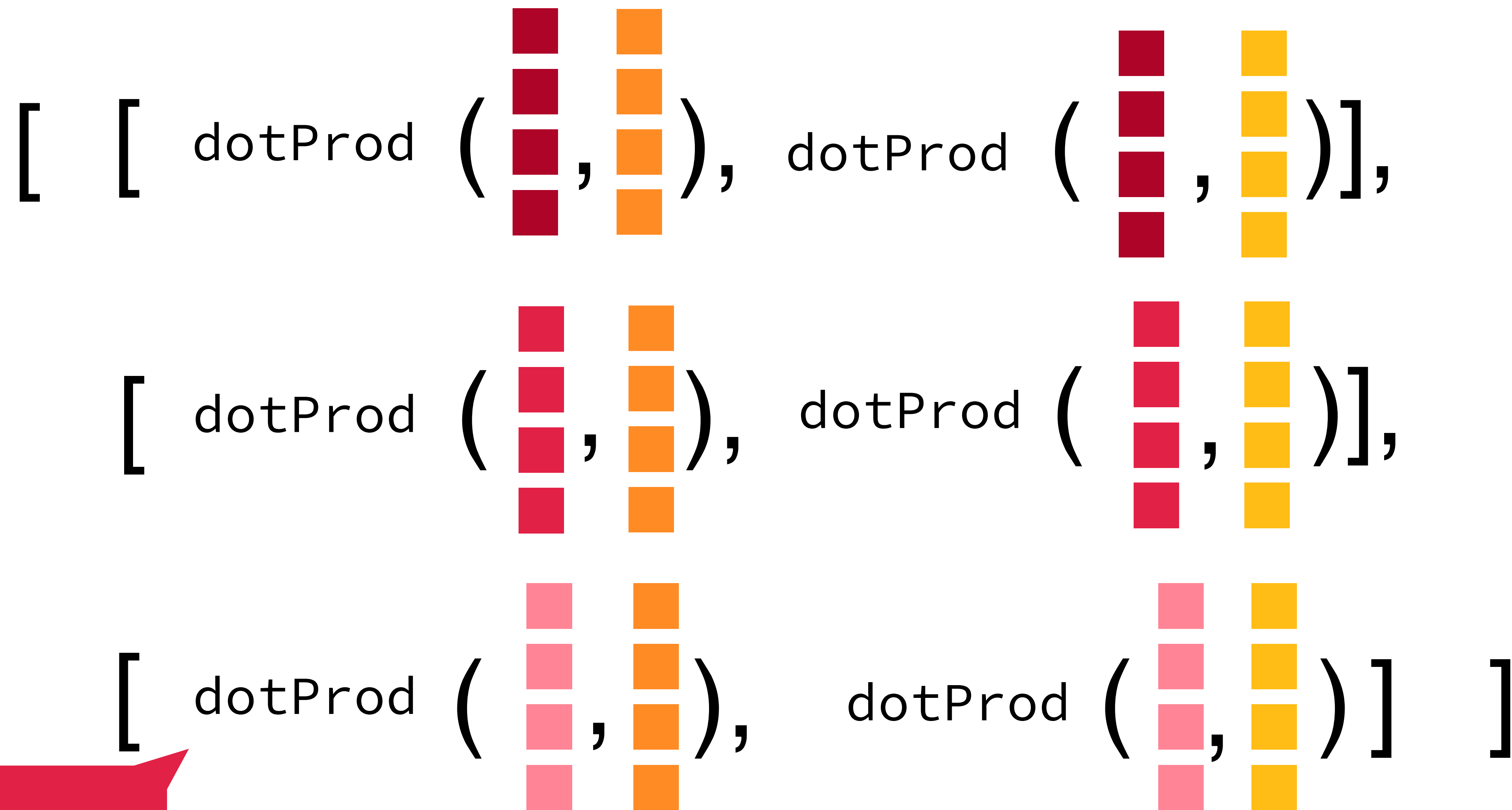


But now, map operator maps over wrong dimension!

map2D dotProd

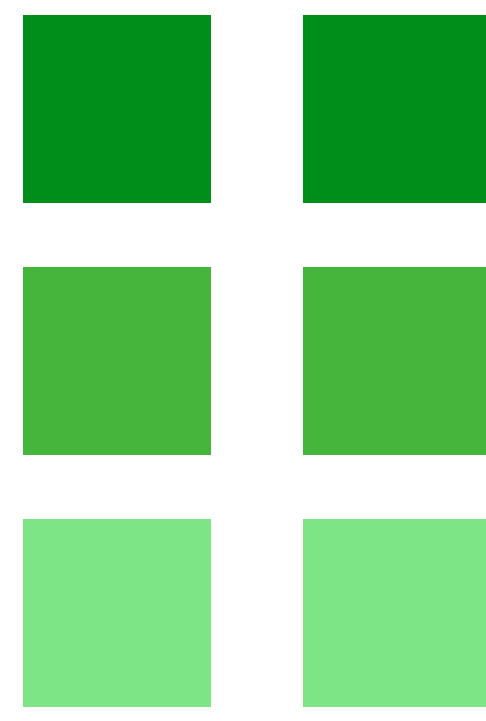
We also need a new
map operator

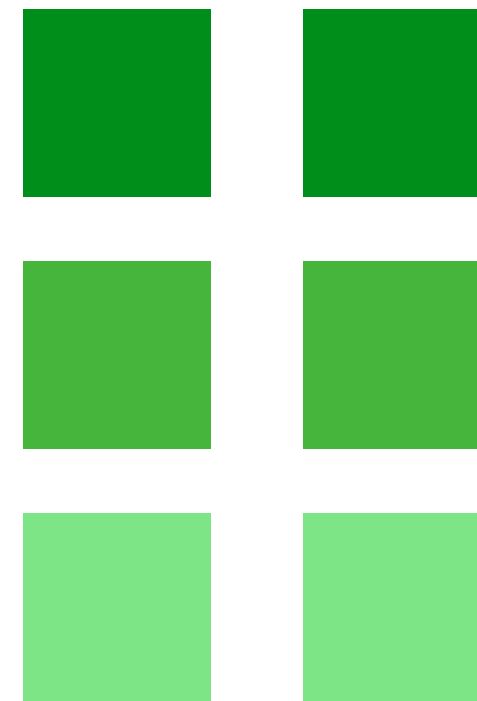




2D map operator maps over correct dimension

[[[■, ■],
[[■, ■],
[[■, ■]]]





Shape information is preserved!

\times_{2D} and map2D hard-code which dimensions are **iterated over** and which dimensions are **computed on**...

\times_{2D} and map2D hard-code which dimensions are **iterated over** and which dimensions are **computed on**...

...but if tensor shapes change, we'll need entirely new operators!

\times_{2D} and map2D hard-code which dimensions are **iterated over** and which dimensions are **computed on**...

...but if tensor shapes change, we'll need entirely new operators!

Can we encode this in the tensor itself?

\times_{2D} and map2D hard-code which dimensions are **iterated over** and which dimensions are **computed on**...

...but if tensor shapes change, we'll need entirely new operators!

Can we encode this in the tensor itself?

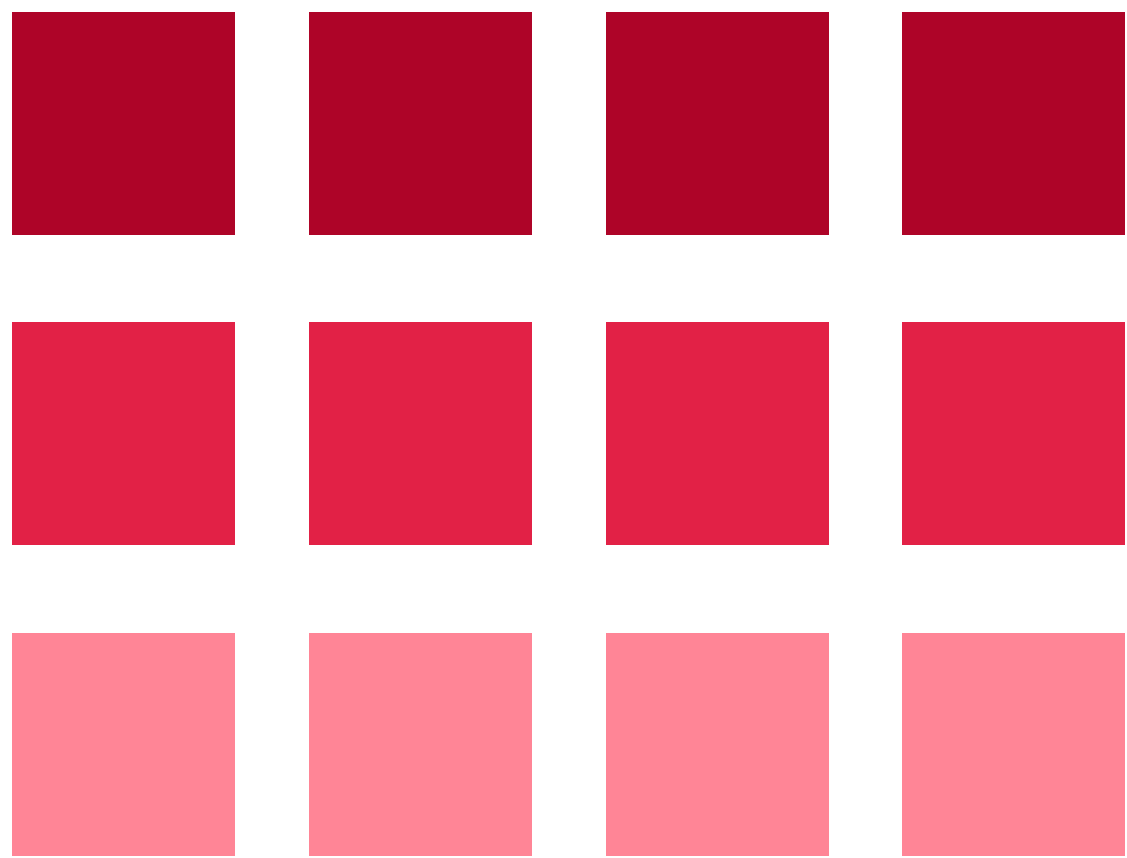
(Yes! This is what access patterns do!)



Outline

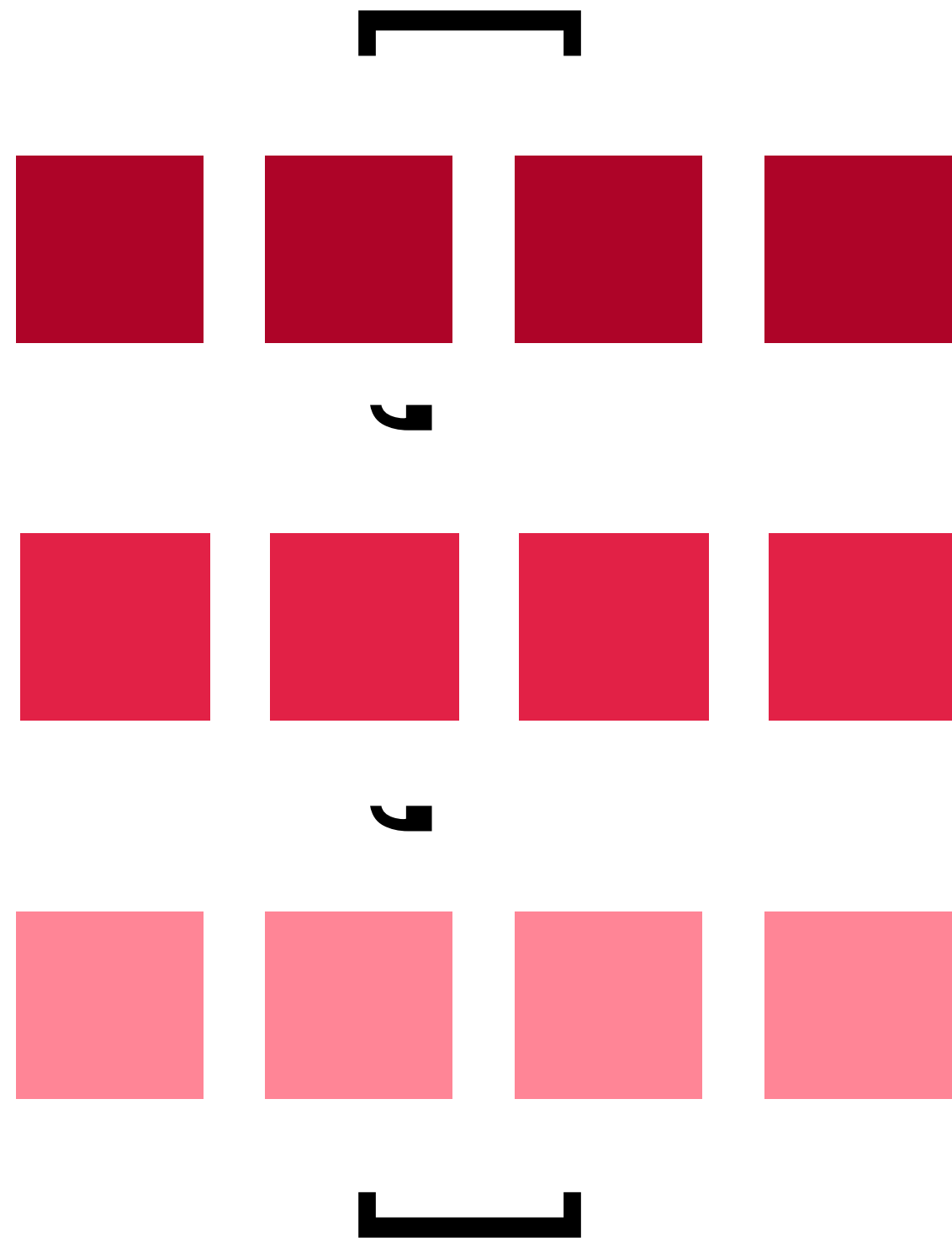
- Motivating Example: Matrix Multiplication
- **Access Pattern Definition**
- Case Studies
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation

A **tensor** looks like...



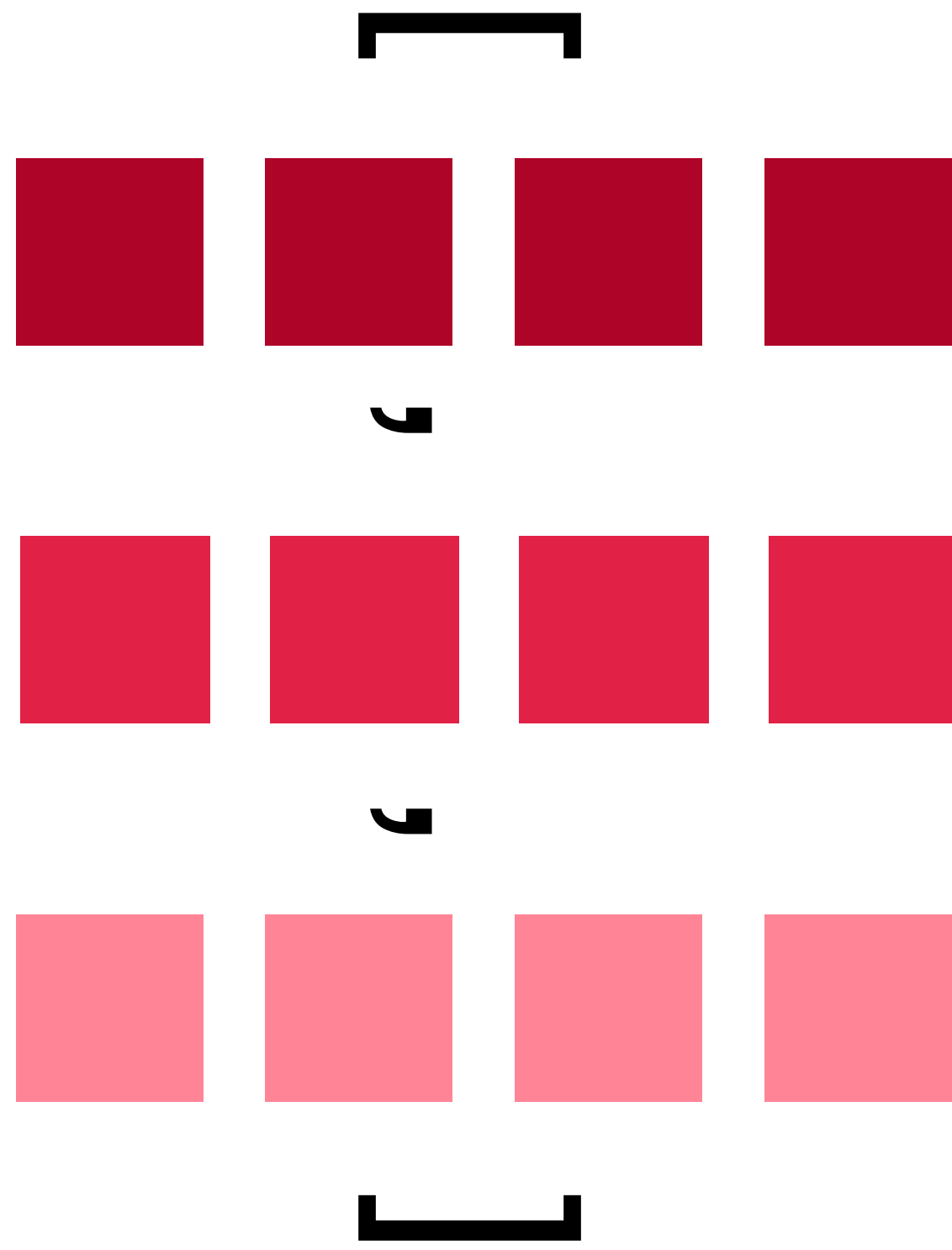
(3, 4)

An **access pattern** looks like...

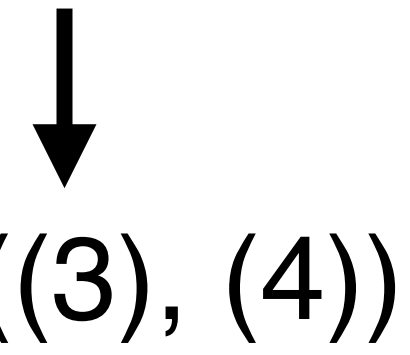


((3), (4))

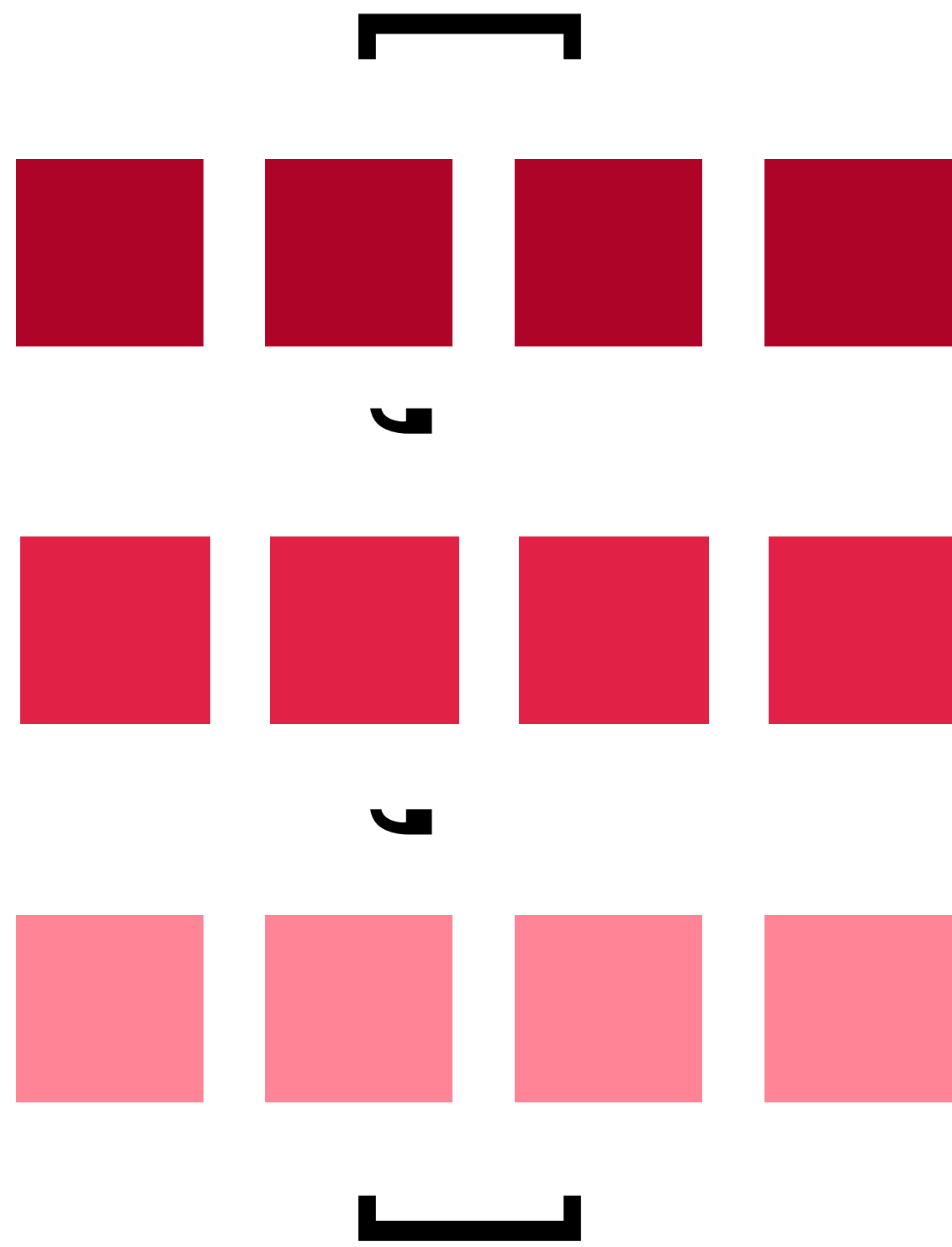
An **access pattern** looks like...



access dimensions
(iterated over)



An **access pattern** looks like...



access dimensions
(iterated over)

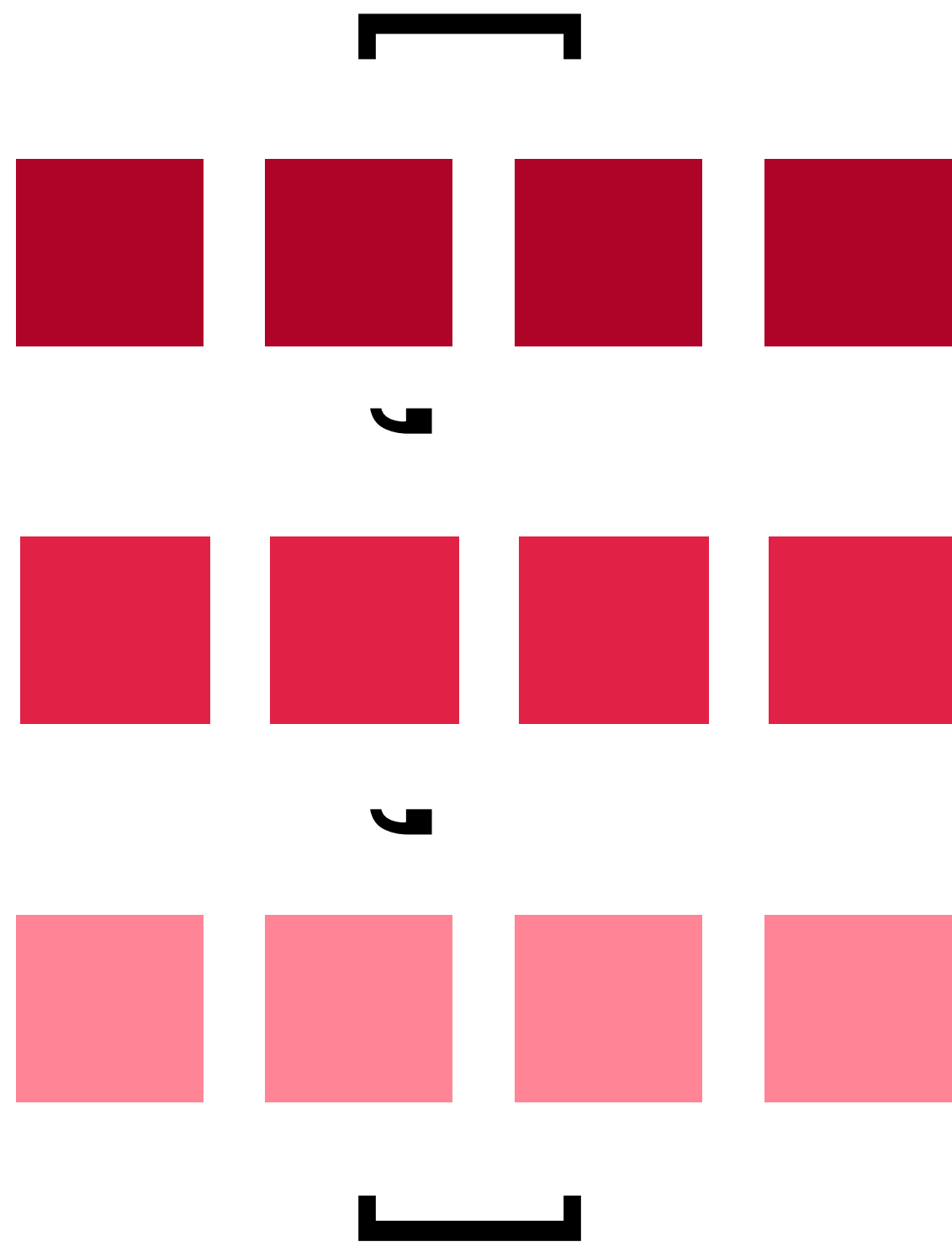


((3), (4))



compute dimensions
(computed on)

An **access pattern** looks like...



A 3-length vector of 4-length vectors

access dimensions
(iterated over)

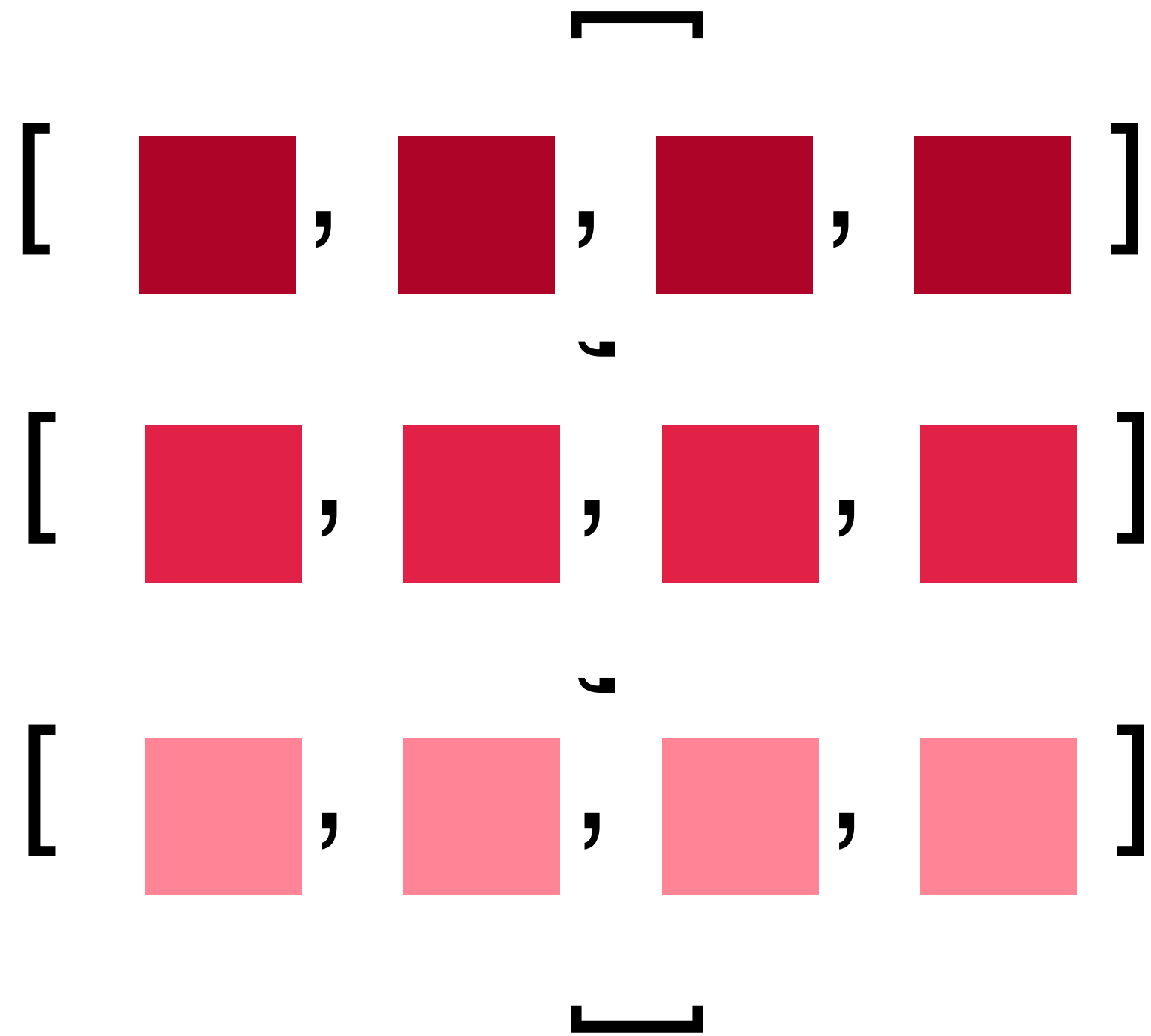


$((3), (4))$



compute dimensions
(computed on)

An **access pattern** looks like...



access dimensions
(iterated over)

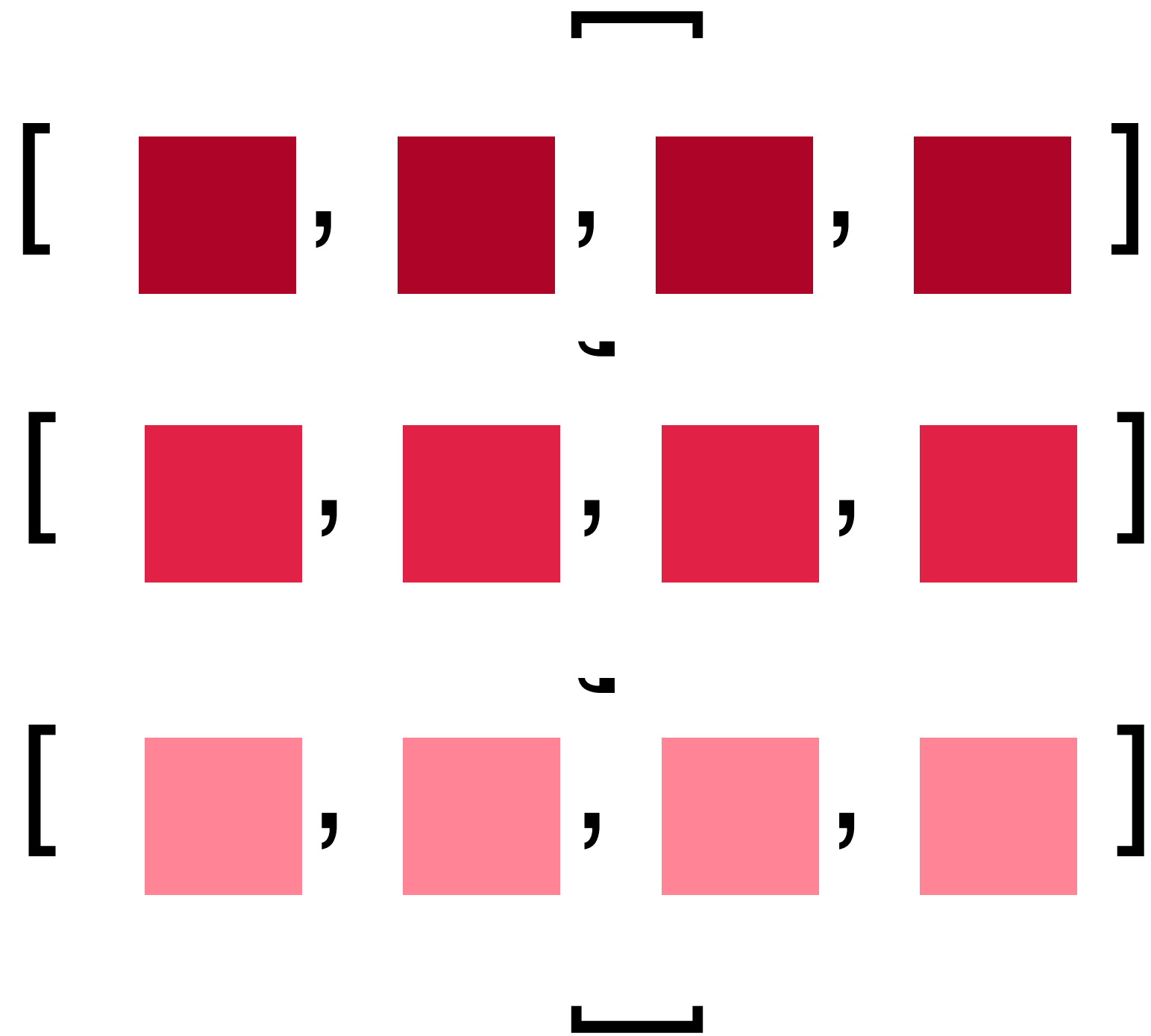


$((3,4), ())$



compute dimensions
(computed on)

An **access pattern** looks like...



A (3,4)-shaped tensor
of scalars

access dimensions
(iterated over)

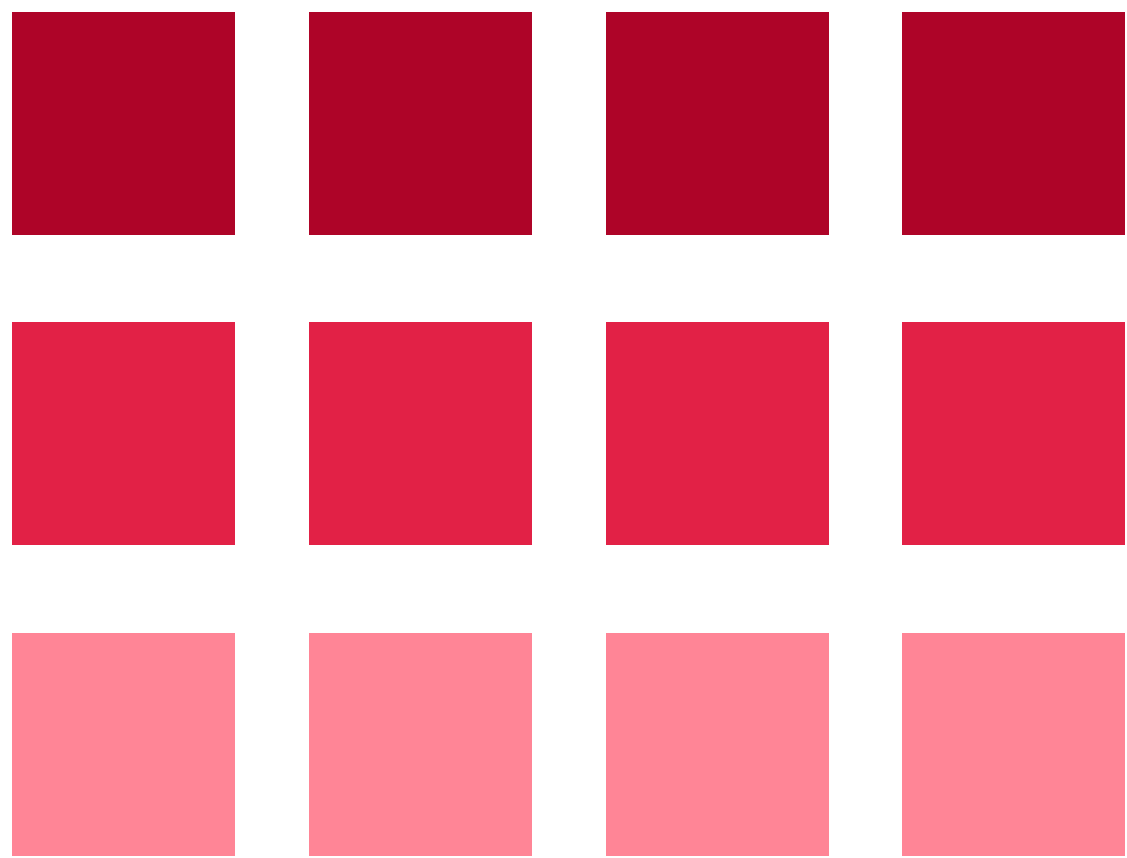


$((3,4), ())$



compute dimensions
(computed on)

An **access pattern** looks like...



access dimensions
(iterated over)

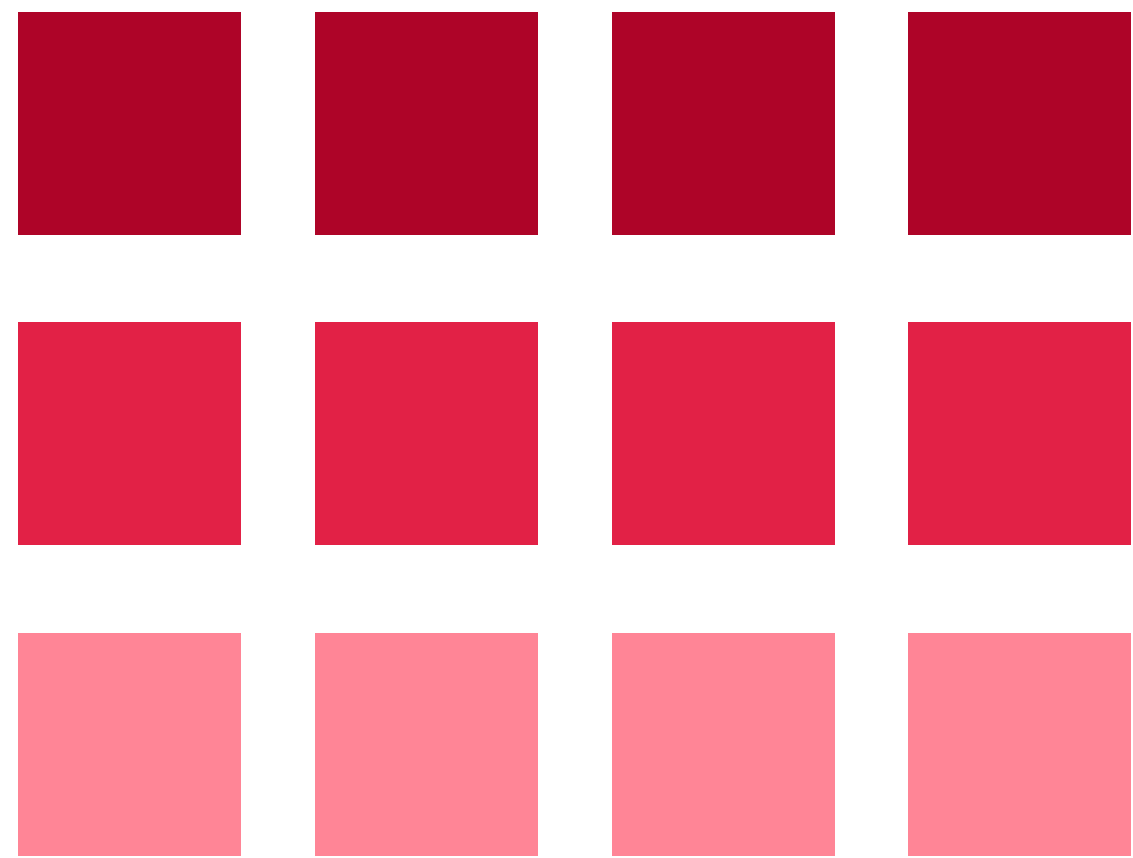


$((), (3,4))$



compute dimensions
(computed on)

An **access pattern** looks like...



A scalar-shaped tensor
of a single (3,4)-
shaped tensor

access dimensions
(iterated over)

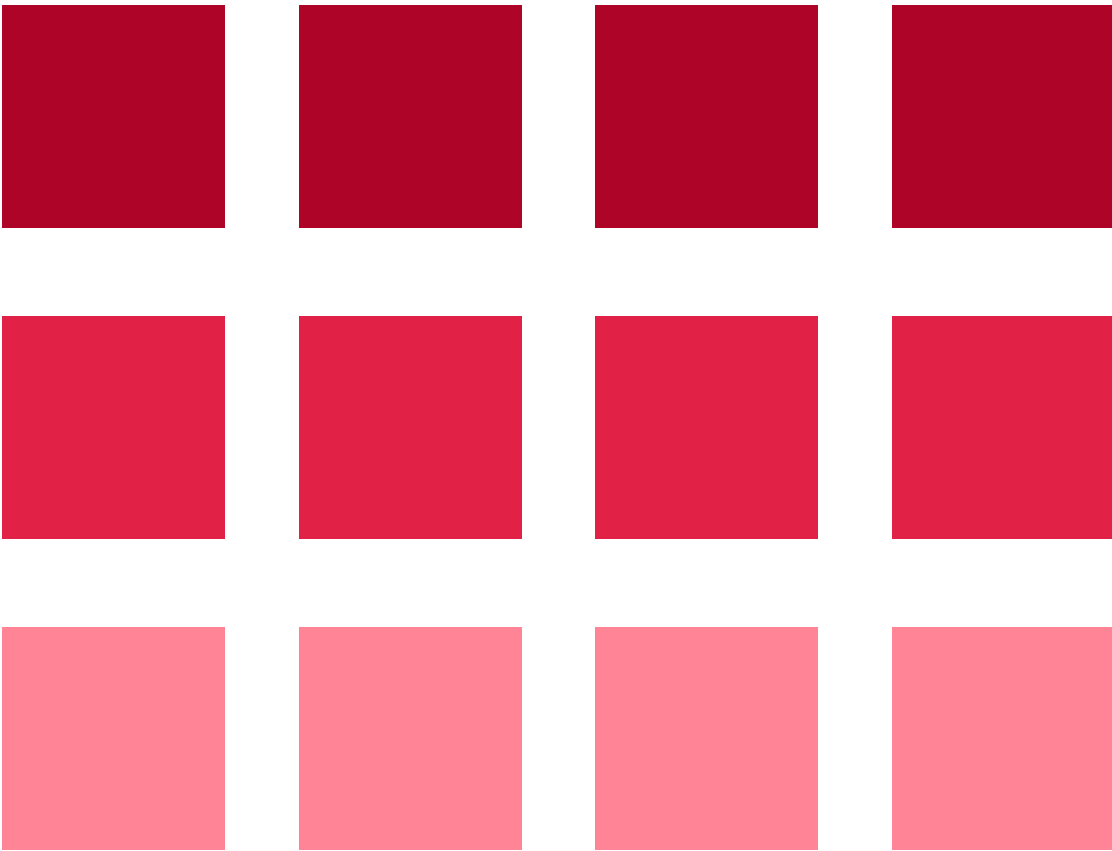


$((), (3,4))$

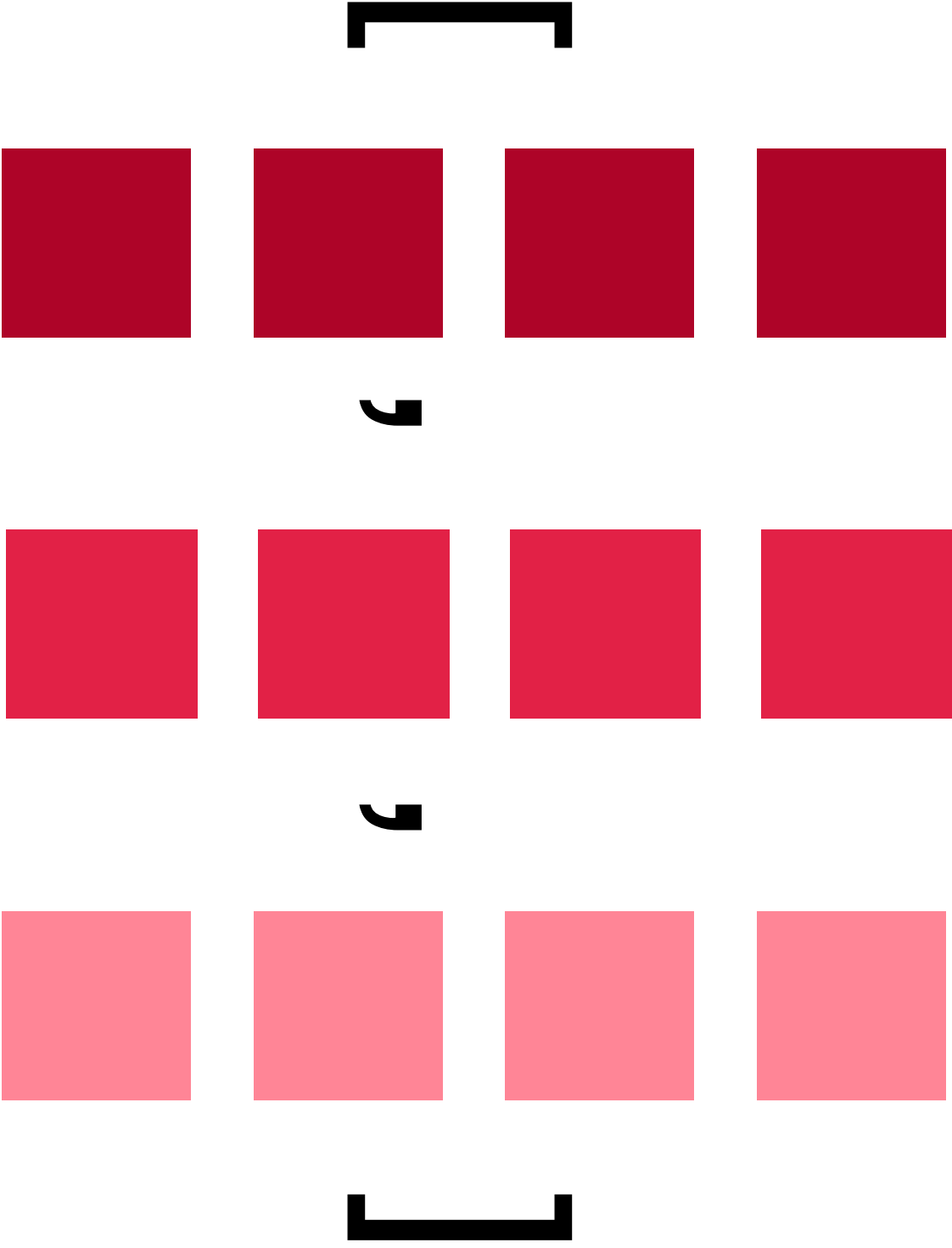


compute dimensions
(computed on)

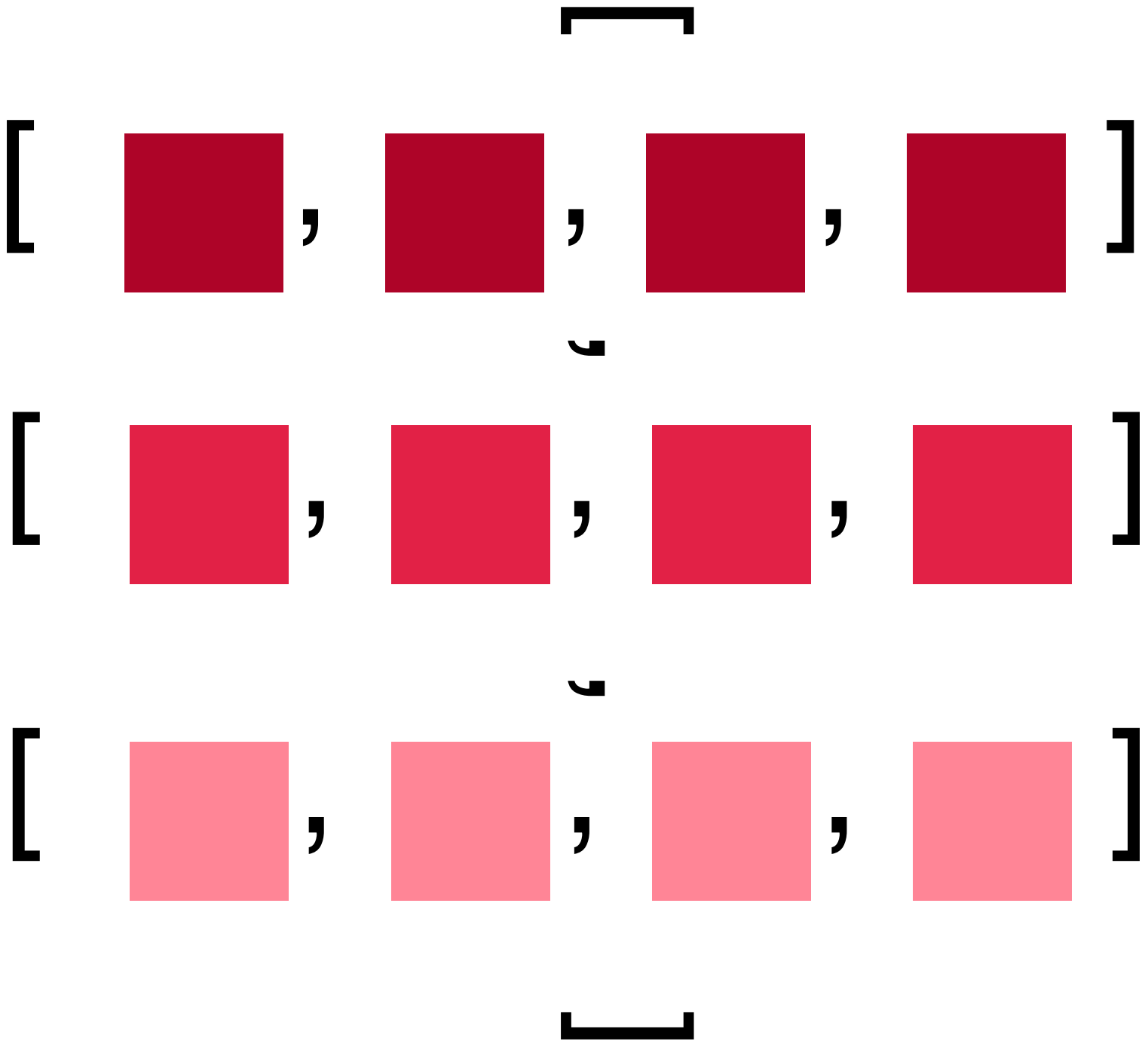
$((), (3,4))$



$((3), (4))$



$((3,4), ())$



Same tensor, three possible views!

Transformer	Input(s)	Output Shape
access	$((a_0, \dots), (\dots, a_n))$ and non-negative integer i	$((a_0, \dots, a_{i-1}), (a_i, \dots, a_n))$
cartProd	$((a_0, \dots, a_n), (c_0, \dots, c_p))$ and $((b_0, \dots, b_m), (c_0, \dots, c_p))$	$((a_0, \dots, a_n, b_0, \dots, b_m), (2, c_0, \dots, c_p))$
windows	$((a_0, \dots, a_m), (b_0, \dots, b_n)),$ window shape literal (s_0, \dots, s_l)	$((a_0, \dots, a_m, b'_0, \dots, b'_n), (w_0, \dots, w_n)),$ $(\lfloor (a_0 + 1) / s_0 \rfloor, \dots, \lfloor (a_m + 1) / s_l \rfloor)$
slice	$((a_0, \dots),$ dimension index l	(\dots, a_n) $- l$
squeeze	$((a_0, \dots), (\dots, a_n)),$ dimension index d ; we assume $a_d = 1$	$((a_0, \dots), (\dots, a_n))$ with a_d removed
flatten	$((a_0, \dots, a_m), (b_0, \dots, b_n))$	$((a_0 \cdots a_m), (b_0 \cdots b_n))$
reshape	$((a_0, \dots, a_m), (b_0, \dots, b_n)),$ access pattern shape literal $((c_0, \dots, c_p), (d_0, \dots, d_q))$	$((c_0, \dots, c_p), (d_0, \dots, d_q)),$ if $a_0 \cdots a_m = c_0 \cdots c_p$ and $b_0 \cdots b_n = d_0 \cdots d_q$

We can redefine common tensor and list operators with access pattern semantics, which gives us the **Glenside IR**—details in paper!

Table 1. Glenside’s access pattern transformers.



Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- Case Studies
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation



Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- **Case Studies**
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation

Given matrices A and B , pair each row of A with each column of B , compute their dot products, and arrange the results back into a matrix.

(access A 1) ; ((3), (4))

Access A as a list of its rows

`(access A 1)`

`; ((3), (4))`

(access A 1) ; ((3), (4))

(access B 1) ; ((4), (2))

<code>(access A 1)</code>	<code>;</code>	<code>((3), (4))</code>
<code>(transpose</code>	<code>;</code>	<code>((2), (4))</code>
<code> (access B 1)</code>	<code>;</code>	<code>((4), (2))</code>
<code>(list 1 0))</code>		

`(access A 1)`

`(transpose`

`(access B 1)`

`(list 1 0))`

`; ((3), (4))`

`; ((2), (4))`

`; ((4), (2))`

Access B as a list of
its rows, then
transpose into a list
of its columns

<code>(cartProd</code>	<code>;</code>	<code>((3, 2), (2, 4))</code>
<code> (access A 1)</code>	<code>;</code>	<code>((3), (4))</code>
<code> (transpose</code>	<code>;</code>	<code>((2), (4))</code>
<code> (access B 1)</code>	<code>;</code>	<code>((4), (2))</code>
<code> (list 1 0))</code>		

Create every row-column pair

```
(cartProd ; ((3, 2), (2, 4))  
  (access A 1) ; ((3), (4))  
  (transpose ; ((2), (4))  
    (access B 1) ; ((4), (2))  
    (list 1 0)))
```

```
(compute dotProd      ; ((3, 2), ())
  (cartProd           ; ((3, 2), (2, 4))
    (access A 1)      ; ((3), (4))
    (transpose        ; ((2), (4))
      (access B 1)    ; ((4), (2))
      (list 1 0))))
```

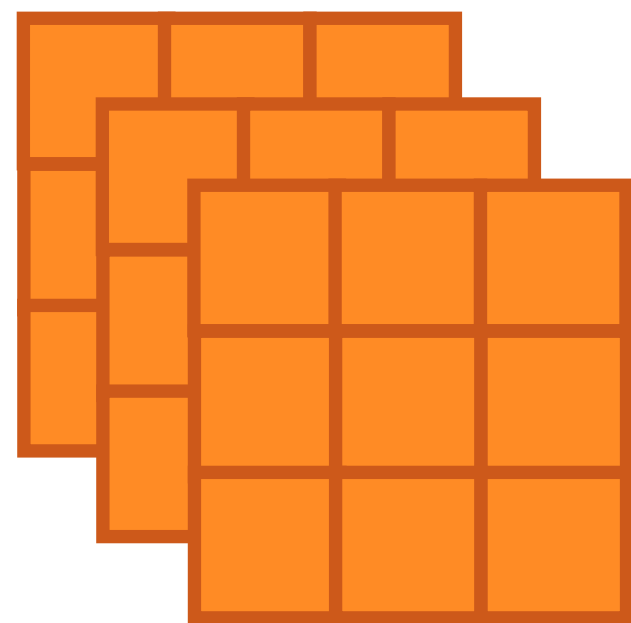
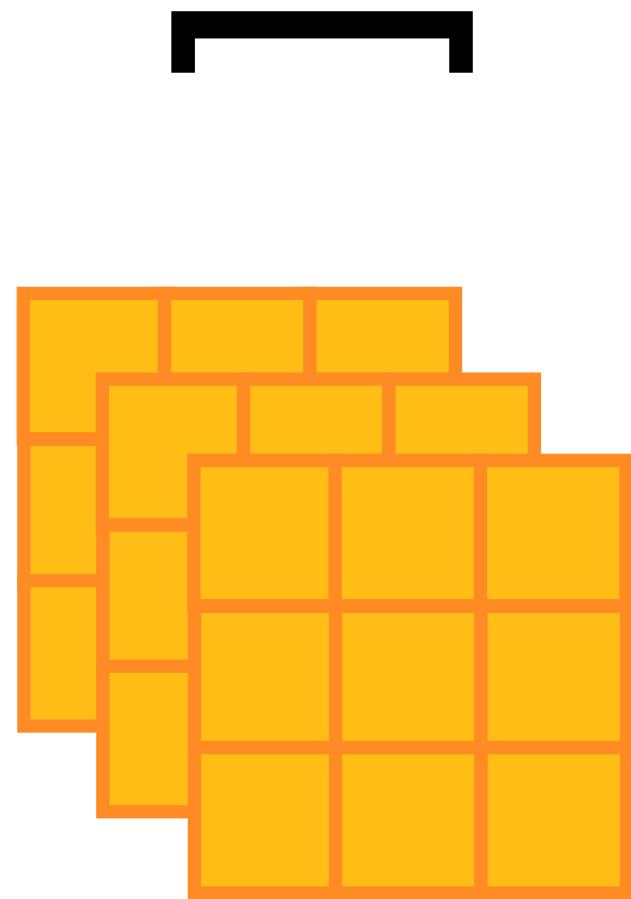
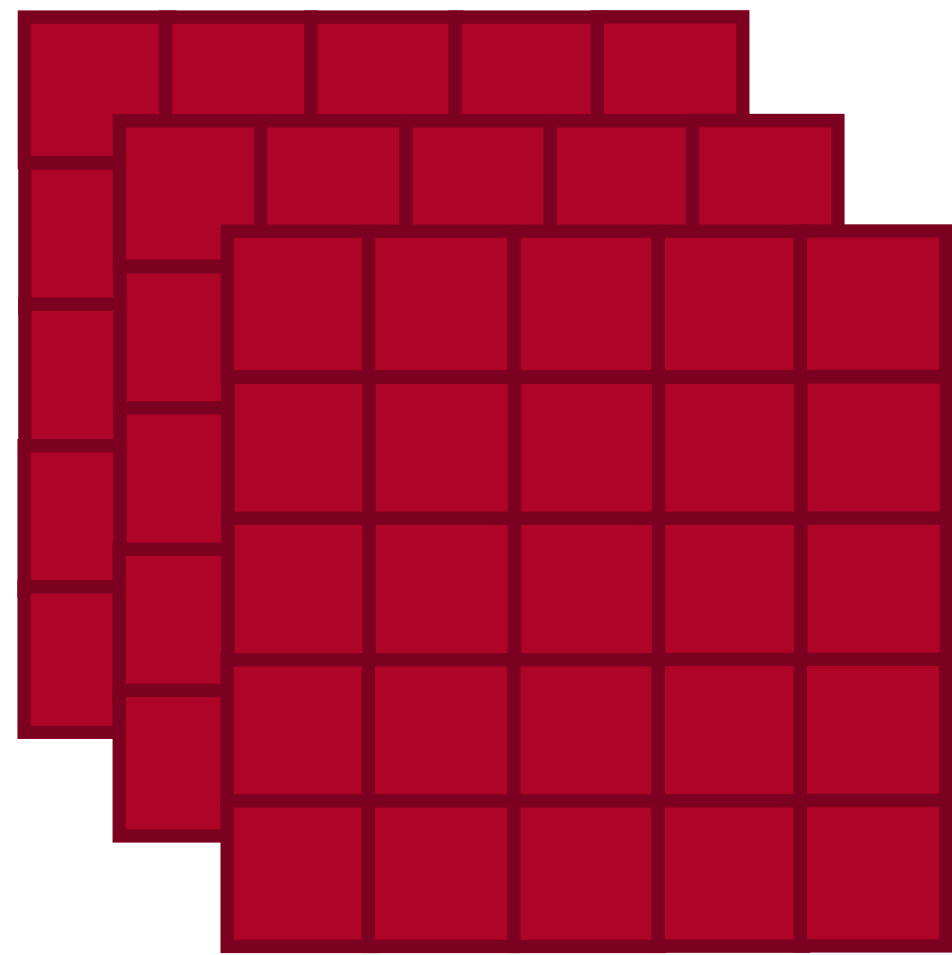
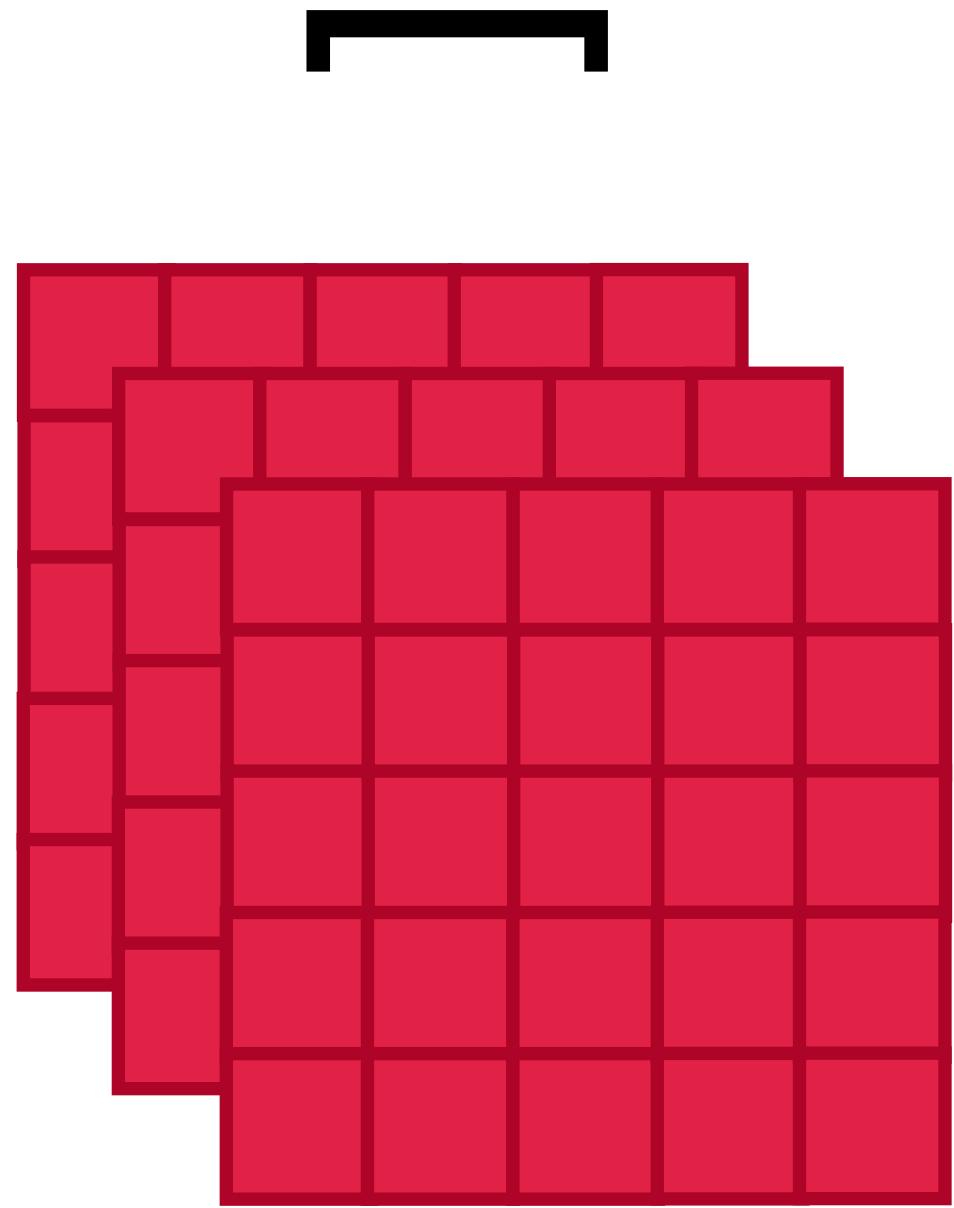
Compute dot product of every row–column pair

```
(compute dotProd ; ((3, 2), ())  
  (cartProd ; ((3, 2), (2, 4))  
    (access A 1) ; ((3), (4))  
    (transpose ; ((2), (4))  
      (access B 1) ; ((4), (2))  
      (list 1 0)))
```

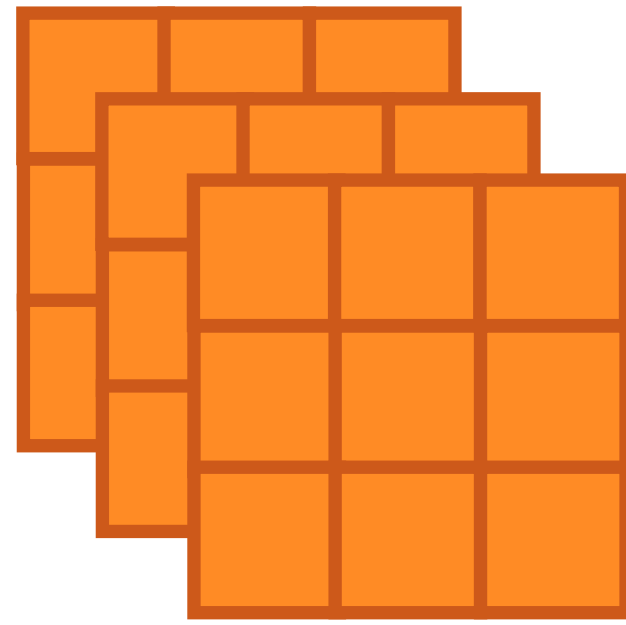
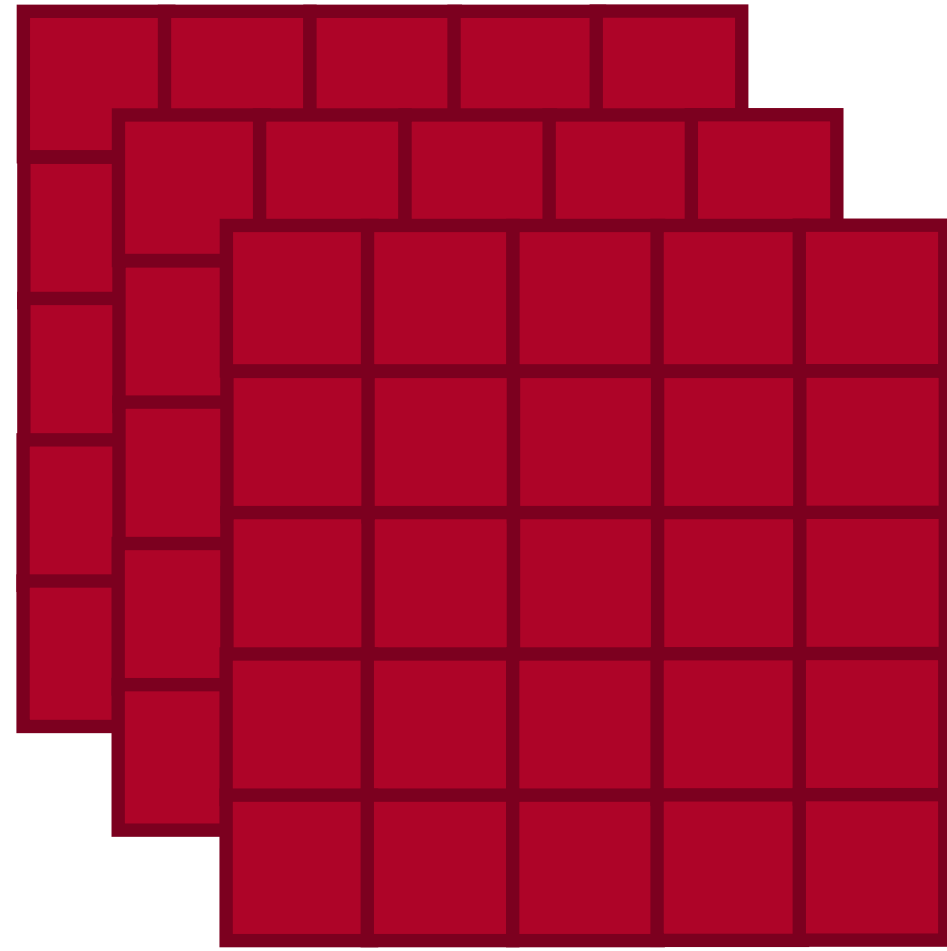


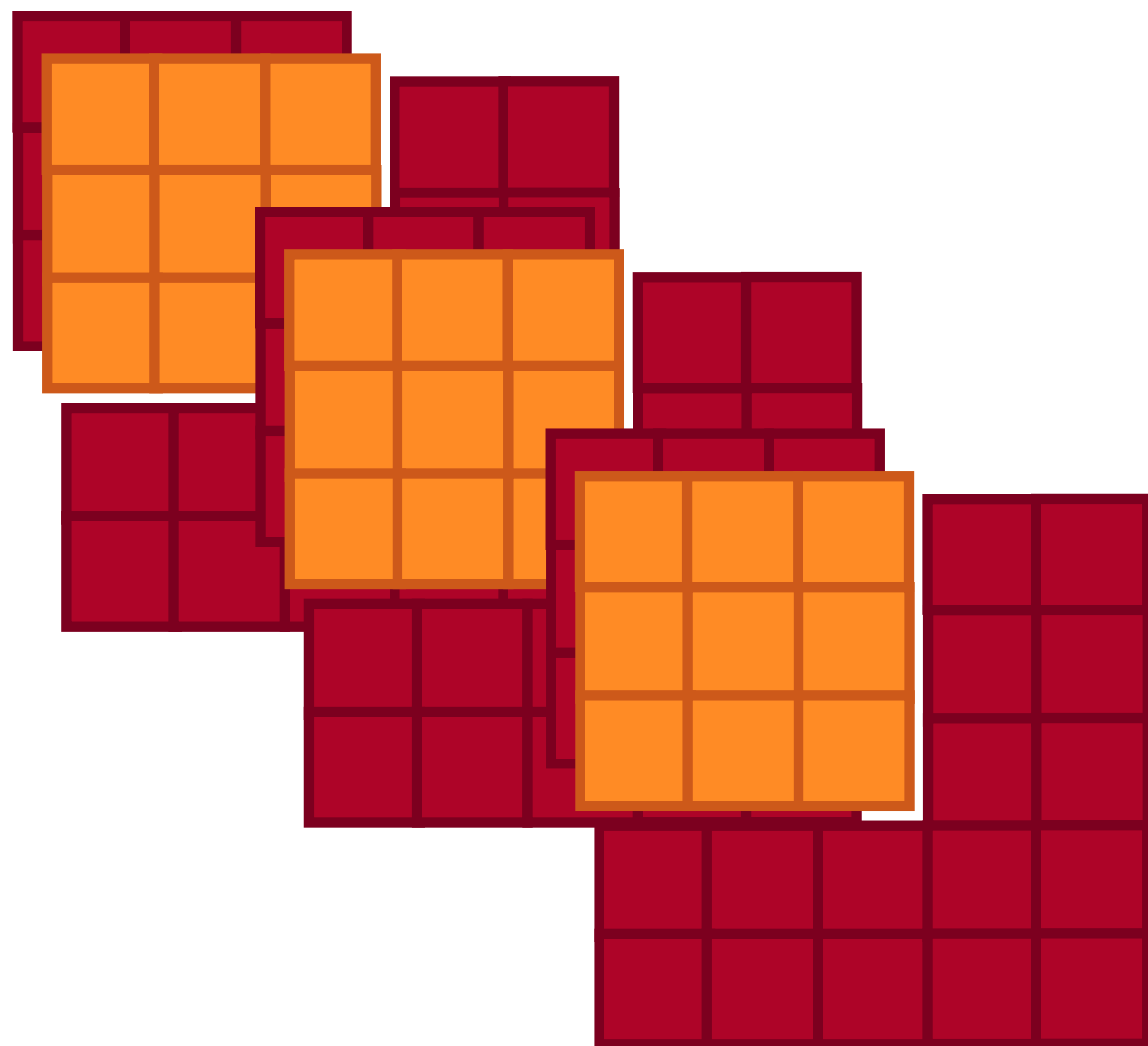
Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- **Case Studies**
 - Reimplementing Matrix Multiplication with Access Patterns
 - **Implementing 2D Convolution with Access Patterns**
 - Hardware Mapping as Program Rewriting
 - Flexible Hardware Mapping with Equality Saturation

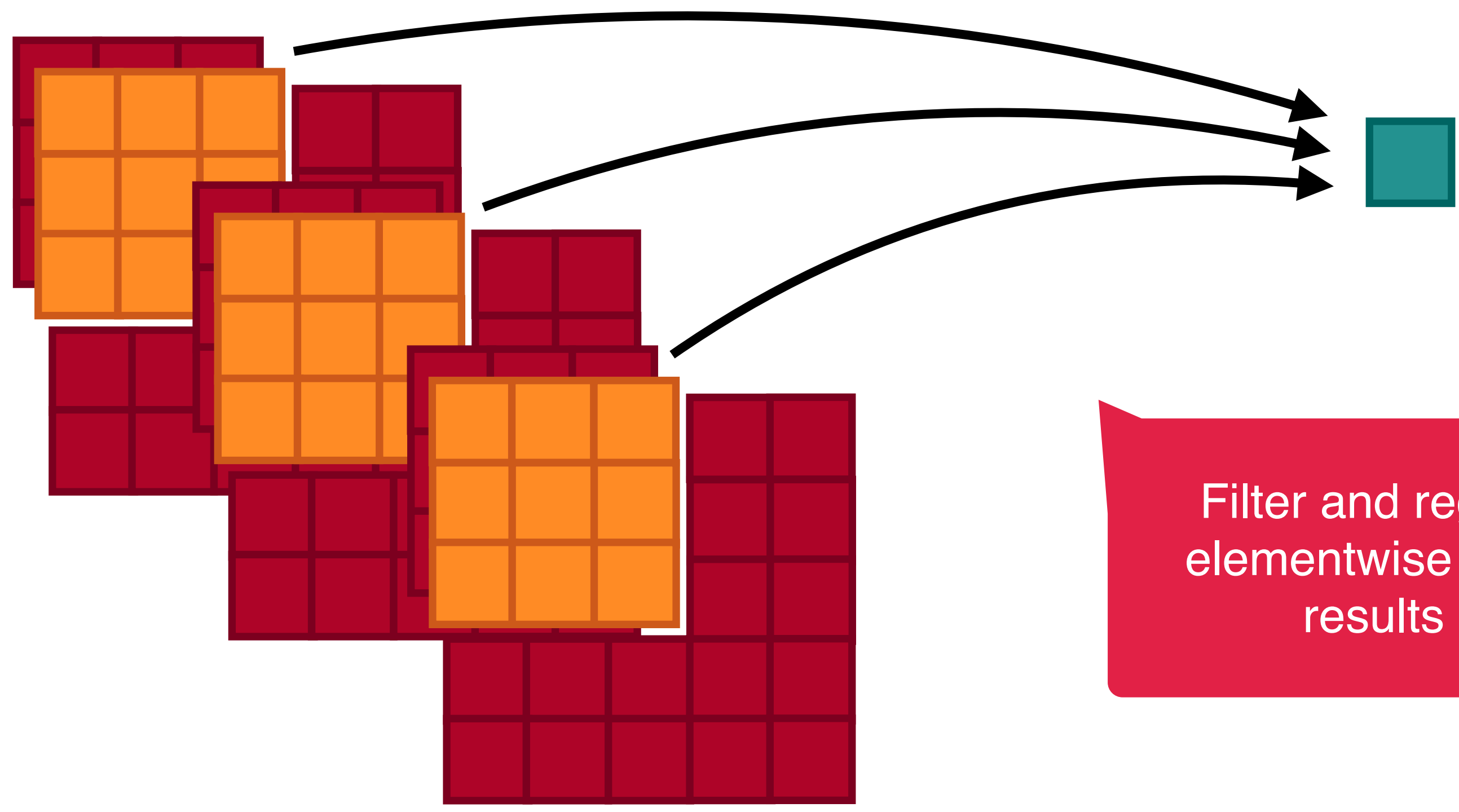


Inputs: a batch of image/activation tensors and a list of weight/filter tensors

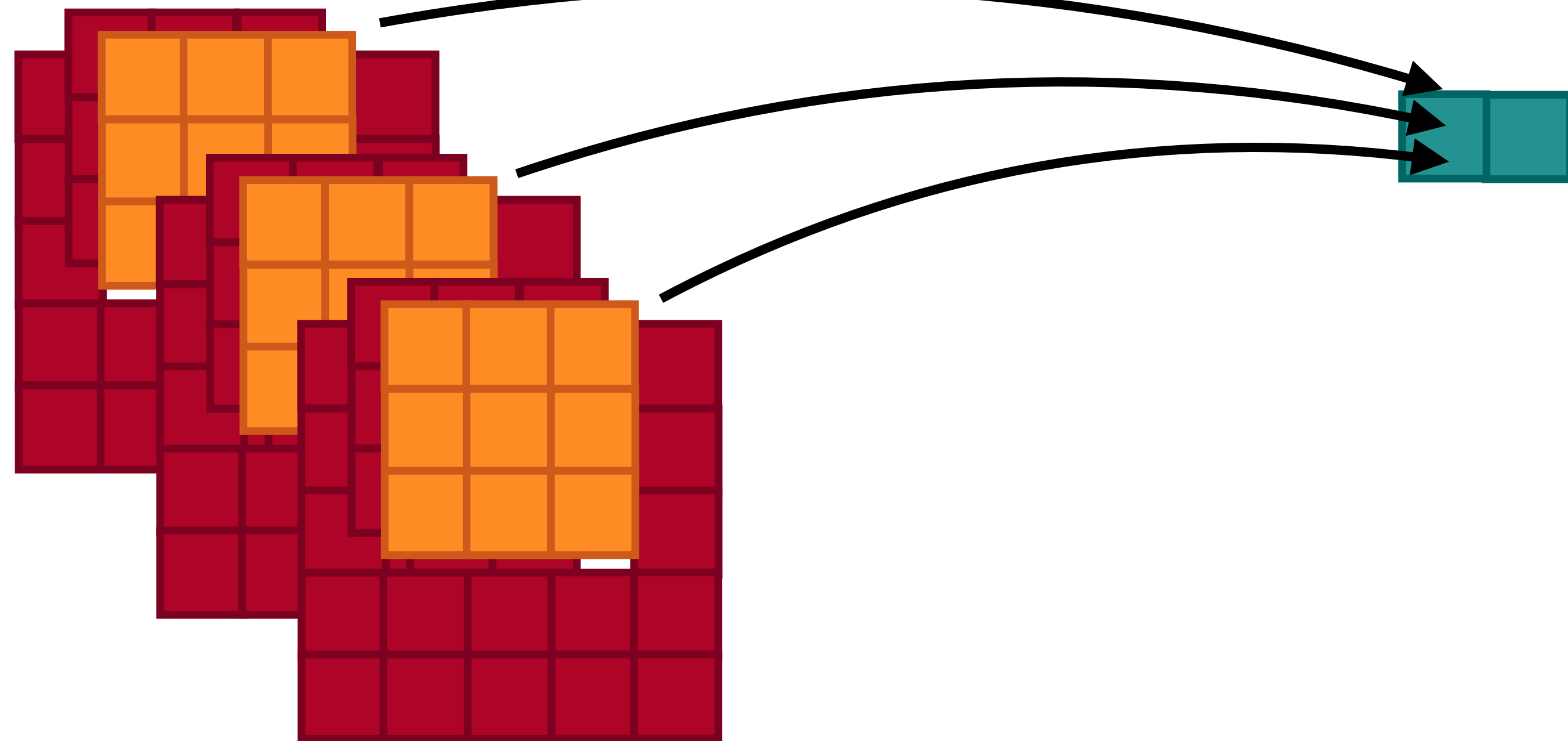


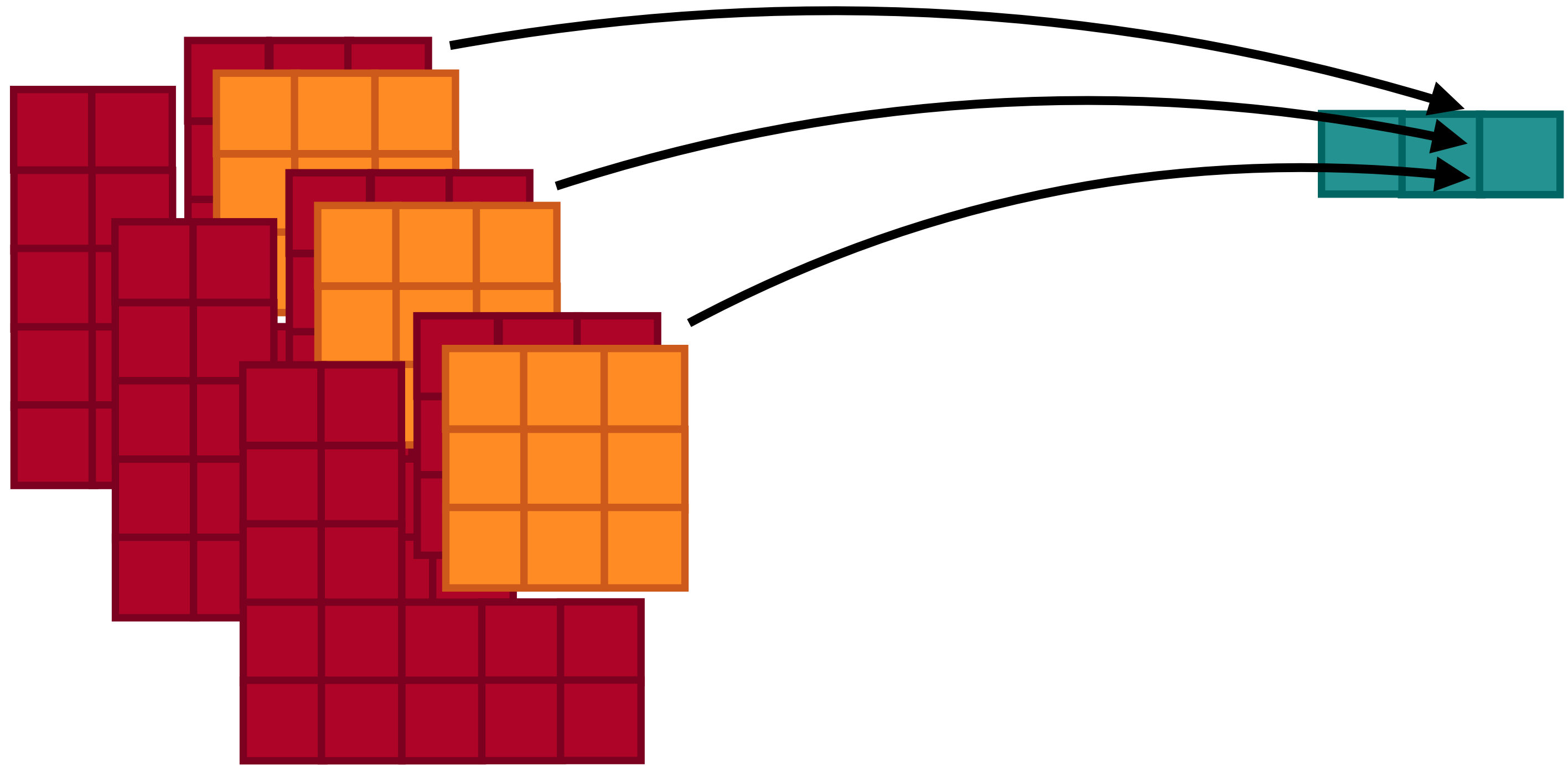


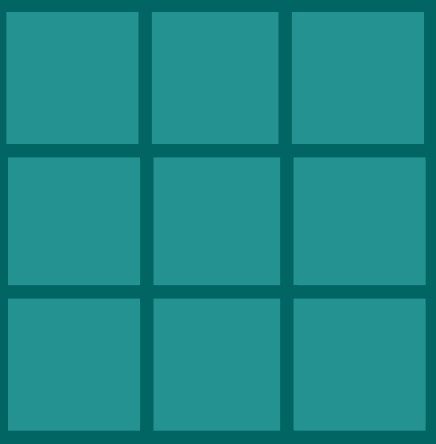
Filter and region of image are
elementwise multiplied and the
results are summed

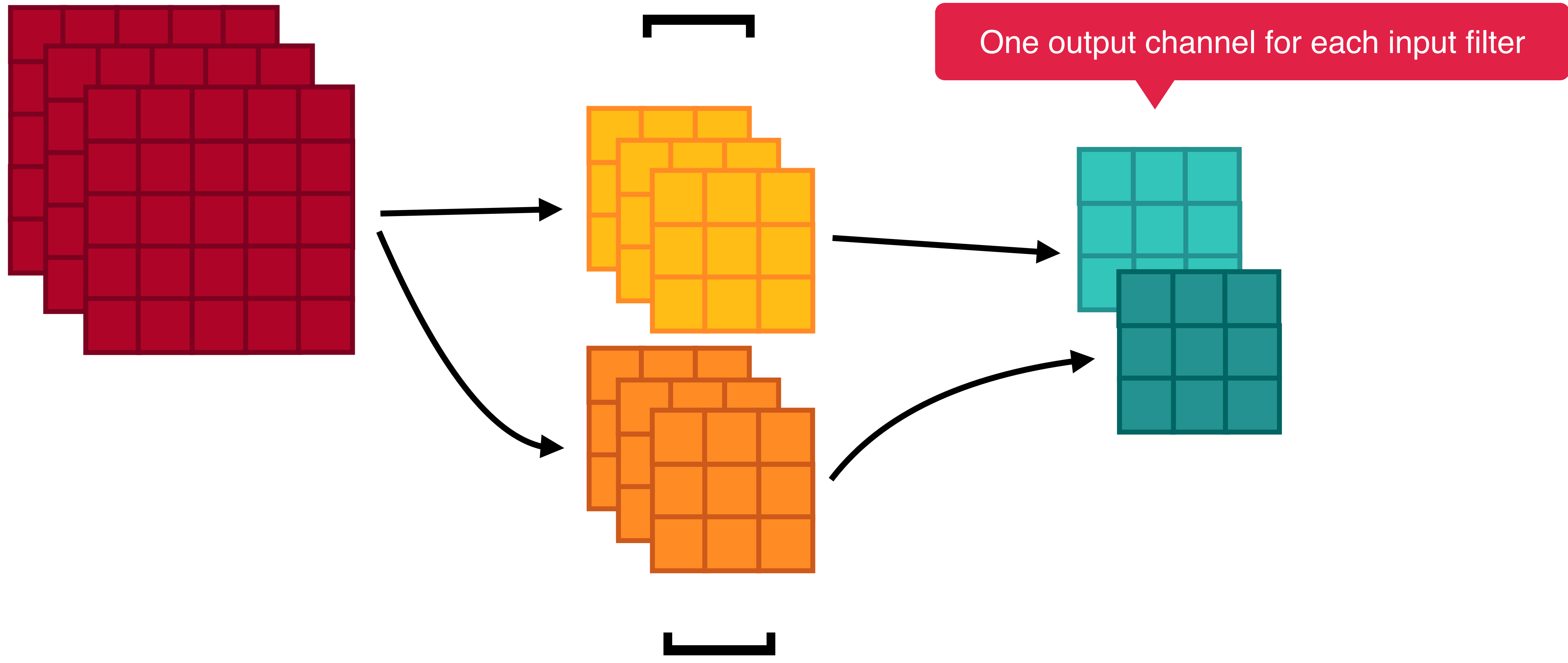


Filter and region of image are
elementwise multiplied and the
results are summed









One output channel for each input filter

Access weights as a list of 3D filters

(access weights 1)

; ((O), (C, K_h, K_w))

Access activations as a batch of 3D images

(access activations 1)

; ((N), (C, H, W))

(access weights 1)

; ((O), (C, K_h, K_w))

Form windows over input images

(windows

(access activations 1)

; ((N), (C, H, W))

(access weights 1)

; ((O), (C, K_h , K_w))

(windows

(access activations 1) ; ((N), (C, H, W))

(shape C Kh Kw)

(shape 1 Sh Sw))

These parameters control
window shape and strides

(access weights 1) ; ((O), (C, Kh, Kw))

At each location in each new image, there is a (C, K_h, K_w) -shaped window

(windows	:	$((N, 1, H', W'), (C, K_h, K_w))$
(access activations 1)	:	$((N), (C, H, W))$
(shape C Kh Kw)		
(shape 1 Sh Sw))		
(access weights 1)	:	$((O), (C, K_h, K_w))$

Pair windows with filters

```
(cartProd ; ((N, 1, H', W', O), (2, C, Kh, Kw))
  (windows ; ((N, 1, H', W'), (C, Kh, Kw))
    (access activations 1) ; ((N), (C, H, W))
    (shape C Kh Kw)
    (shape 1 Sh Sw))
  (access weights 1)) ; ((O), (C, Kh, Kw))
```

Compute dot product of each window-filter pair

```
(compute dotProd  
  (cartProd  
    (windows  
      (access activations 1)  
      (shape C Kh Kw)  
      (shape 1 Sh Sw))  
    (access weights 1)))
```

```
; ((N, 1, H', W', O), ())  
; ((N, 1, H', W', O), (2, C, Kh, Kw))  
; ((N, 1, H', W'), (C, Kh, Kw))  
; ((N), (C, H, W))  
; ((O), (C, Kh, Kw))
```

```

(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (windows
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))
          (access weights 1)))
    1)
  (list 0 3 1 2))
; ((N, O, H', W'), ())
; ((N, 1, H', W', O), ())
; ((N, 1, H', W', O), (2, C, Kh, Kw))
; ((N, 1, H', W'), (C, Kh, Kw))
; ((N), (C, H, W))
; ((O), (C, Kh, Kw))

```

Remove and rearrange dimensions



Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- **Case Studies**
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - **Hardware Mapping as Program Rewriting**
 - Flexible Hardware Mapping with Equality Saturation

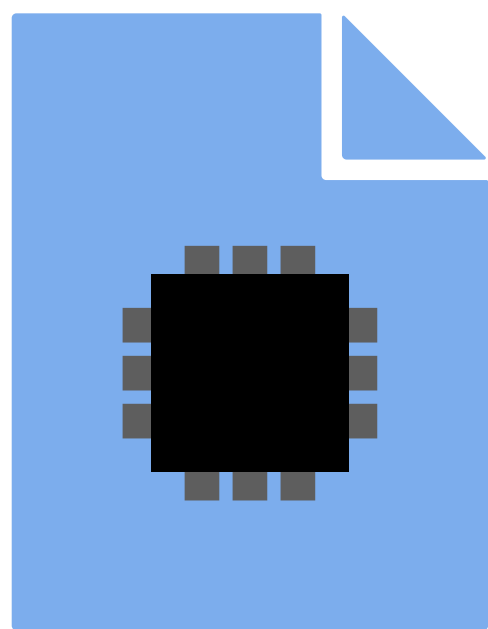


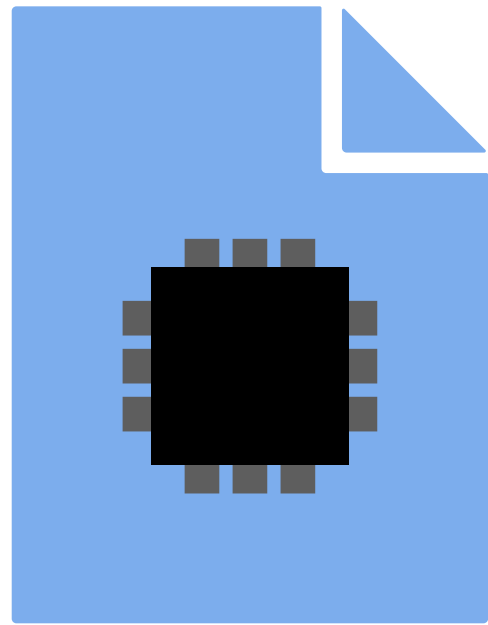
It reads an entire weight array
of shape `rows by cols`.

It then pushes `n` vectors of length
`rows` through the array.

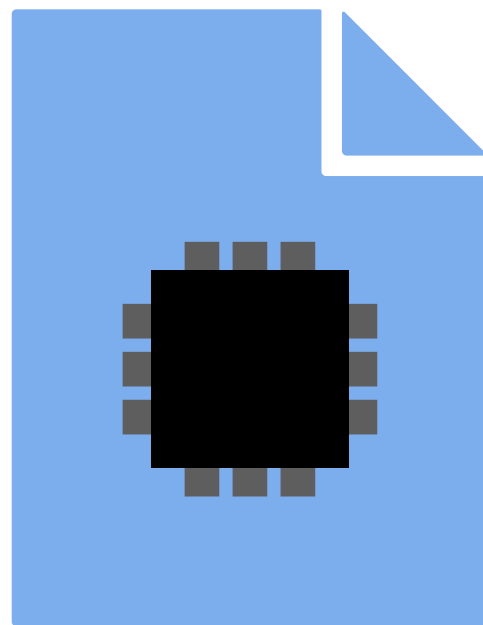
It computes the dot product of
every vector with every column
of the weights.

Finally, it writes out `n` vectors of
length `cols`.





Can we represent hardware as a searchable pattern?



```
(compute dotProd  
  (cartProd ?a0 ?a1))  
where ?a0 is of shape  
  ((?n), (?rows))  
and ?a1 is of shape  
  ((?cols), (?rows))
```

With Glenside, we can!

```
(compute dotProd  
  (cartProd ?a0 ?a1))
```

where ?a0 is of shape
 ((?n), (?rows))

and ?a1 is of shape
 ((?cols), (?rows))



We can directly rewrite to hardware invocations!

```
(systolicArray ?rows ?cols ?a0 ?a1)
```

```
(compute dotProd  
  (cartProd ?a0 ?a1))
```

where ?a0 is of shape
 ((?n), (?rows))



```
(systolicArray ?rows ?cols ?a0 ?a1)
```

and ?a1 is of shape
 ((?cols), (?rows))

```
(compute dotProd  
  (cartProd  
    (access A 1)  
    (transpose  
      (access B 1)  
      (list 1 0))))
```

```
(compute dotProd  
  (cartProd ?a0 ?a1))
```

where ?a0 is of shape
((?n), (?rows))



```
(systolicArray ?rows ?cols ?a0 ?a1)
```

and ?a1 is of shape
((?cols), (?rows))

```
(compute dotProd
```

```
  (cartProd
```

```
    (access A 1)
```

```
    (transpose
```

```
      (access B 1)
```

```
      (list 1 0))))
```

```
(compute dotProd  
  (cartProd ?a0 ?a1))
```

where ?a0 is of shape
 ((?n), (?rows))



```
(systolicArray ?rows ?cols ?a0 ?a1)
```

and ?a1 is of shape
 ((?cols), (?rows))

```
(systolicArray
```

```
  4 2
```

```
  (access A 1)
```

```
  (transpose
```

```
    (access B 1)
```

```
    (list 1 0))))
```




Outline

- Motivating Example: Matrix Multiplication
- Access Pattern Definition
- **Case Studies**
 - Reimplementing Matrix Multiplication with Access Patterns
 - Implementing 2D Convolution with Access Patterns
 - Hardware Mapping as Program Rewriting
 - **Flexible Hardware Mapping with Equality Saturation**

```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (windows
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))
        (access weights 1)))
    1)
  (list 0 3 1 2))
```

```
(compute dotProd
  (cartProd
    (access A 1)
    (transpose
      (access B 1)
      (list 1 0))))
```

Convolution and matrix multiplication have similar structure!

(transpose

(squeeze

(compute dotProd

(cartProd

(windows

(access activations 1)

(shape C Kh Kw)

(shape 1 Sh Sw))

(access weights 1)))

1)

(list 0 3 1 2))

(compute dotProd

(cartProd

(access A 1)

(transpose

(access B 1)

(list 1 0)))

```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (windows
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))
          (access weights 1)))
    1)
  (list 0 3 1 2))
```

Can we apply our hardware rewrite?

```
(compute dotProd
  (cartProd ?a0 ?a1))
where ?a0 is of shape
  ((?n), (?rows))
and ?a1 is of shape
  ((?cols), (?rows))
```

```
(transpose
 (squeeze
  (compute dotProd
   (cartProd
    (windows ; ((N, 1, H', W'), (C, Kh, Kw))
    (access activations 1)
    (shape C Kh Kw)
    (shape 1 Sh Sw))
   (access weights 1))) ; ((O), (C, Kh, Kw))
 1)
(list 0 3 1 2))
```

```
(compute dotProd
 (cartProd ?a0 ?a1))
where ?a0 is of shape
((?n), (?rows))
and ?a1 is of shape
((?cols), (?rows))
```

Our access pattern shapes do not pass the rewrite's conditions

```
(transpose
 (squeeze
 (compute dotProd
 (cartProd
 (windows ; ((?n), (?rows))
 (access activations 1)
 (shape C Kh Kw)
 (shape 1 Sh Sw))
 (access weights 1))) ; ((?cols), (?rows))
 1)
(list 0 3 1 2))
```

```
(compute dotProd
 (cartProd ?a0 ?a1))
where ?a0 is of shape
 ((?n), (?rows))
and ?a1 is of shape
 ((?cols), (?rows))
```

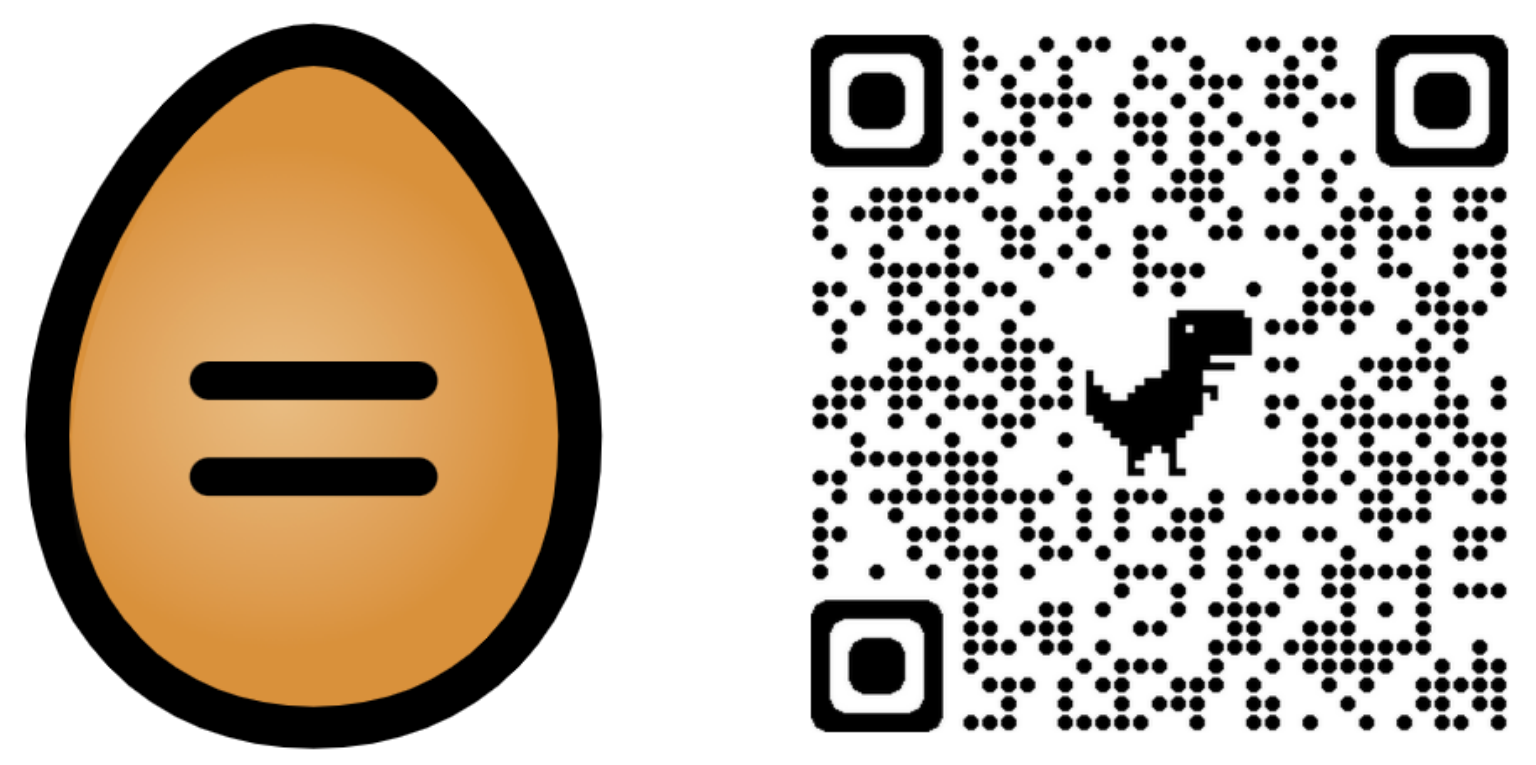
Can we flatten our access patterns?

?a → (reshape (flatten ?a) ?shape)

Flattens and immediately reshapes an access pattern

?a → (reshape (flatten ?a) ?shape)

Flattens and immediately reshapes an access pattern




```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (windows ; ((N, 1, H', W'), (C, Kh, Kw))
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))
        (access weights 1))) ; ((O), (C, Kh, Kw))
      1)
    (list 0 3 1 2))
```

```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (reshape (flatten (windows ; ((N, 1, H', W'), (C, Kh, Kw))
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))) ?shape0)
        (reshape (flatten (access weights 1)) ?shape1))) ; ((O), (C, Kh, Kw))
      1)
    (list 0 3 1 2))
```

```
(transpose
 (squeeze
  (compute dotProd
   (cartProd
    (reshape (flatten (windows ; ((N, 1, H', W'), (C, Kh, Kw))
     (access activations 1)
     (shape C Kh Kw)
     (shape 1 Sh Sw))) ?shape0)
    (reshape (flatten (access weights 1) ?shape1))) ; ((O), (C, Kh, Kw))
  1)
 (list 0 3 1 2))
```

But our access pattern shapes haven't changed!

```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (reshape (flatten (windows ; ((N, 1, H', W'), (C, Kh, Kw))
          (access activations 1)
            (shape C Kh Kw)
              (shape 1 Sh Sw))) ?shape0)
        (reshape (flatten (access weights 1) ?shape1))) ; ((O), (C, Kh, Kw))
      1)
    (list 0 3 1 2))
```

We need to “bubble” the reshapes to the top

These rewrites “bubble” reshape through cartProd and compute dotProd

```
(cartProd  
  (reshape ?a0 ?shape0)  
  (reshape ?a1 ?shape1)) → (reshape (cartProd ?a0 ?a1) ?newShape)
```

```
(compute dotProd  
  (reshape ?a ?shape)) → (reshape (compute dotProd ?a) ?newShape)
```

```
(transpose
  (squeeze
    (compute dotProd
      (cartProd
        (reshape (flatten (windows ; ((N, 1, H', W'), (C, Kh, Kw))
          (access activations 1)
          (shape C Kh Kw)
          (shape 1 Sh Sw))) ?shape0)
        (reshape (flatten (access weights 1)) ?shape1))) ; ((O), (C, Kh, Kw))
      1)
    (list 0 3 1 2))
```

```
(transpose
 (squeeze
  (reshape (compute dotProd
   (cartProd
    (flatten (windows ; ((N · 1 · H' · W'), (C · Kh · Kw))
     (access activations 1)
     (shape C Kh Kw)
     (shape 1 Sh Sw)))
    (flatten (access weights 1))) ?shape) ; ((O), (C · Kh · Kw))
  1)
 (list 0 3 1 2))
```

reshapes have been moved out, and the access patterns are flattened!

```

(transpose
 (squeeze
  (reshape (compute dotProd
            (cartProd
             (flatten (windows ; ((N · 1 · H' · W'), (C · Kh · Kw))
                        (access activations 1)
                        (shape C Kh Kw)
                        (shape 1 Sh Sw)))
             (flatten (access weights 1))) ?shape) ; ((O), (C · Kh · Kw))
  1)
 (list 0 3 1 2))

```

```

(compute dotProd
 (cartProd ?a0 ?a1))

```

where ?a0 is of shape
 ((?n), (?rows))
 and ?a1 is of shape
 ((?cols), (?rows))

Our rewrite can now map
 convolution to matrix
 multiplication hardware!

`?a → (reshape (flatten ?a) ?shape)`

`(cartProd
 (reshape ?a0 ?shape0)
 (reshape ?a1 ?shape1)) → (reshape (cartProd ?a0 ?a1) ?newShape)`

`(compute dotProd
 (reshape ?a ?shape)) → (reshape (compute dotProd ?a) ?newShape)`

These rewrites *rediscover* the im2col transformation!

In conclusion,

In conclusion,

we have presented **access patterns** as a new tensor representation,

In conclusion,

we have presented **access patterns** as a new tensor representation,

we have used them to build the **pure, low-level, binder free IR Glenside,**

In conclusion,

we have presented **access patterns** as a new tensor representation,

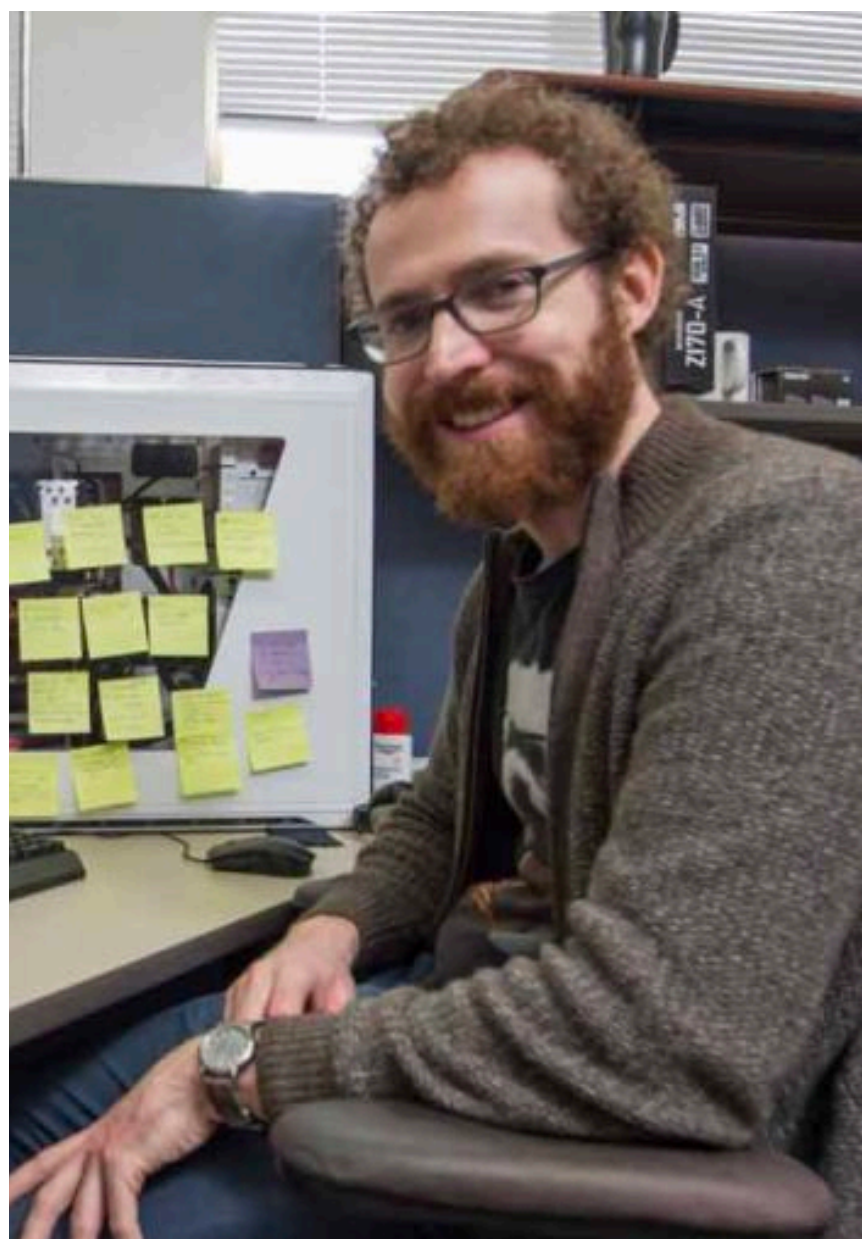
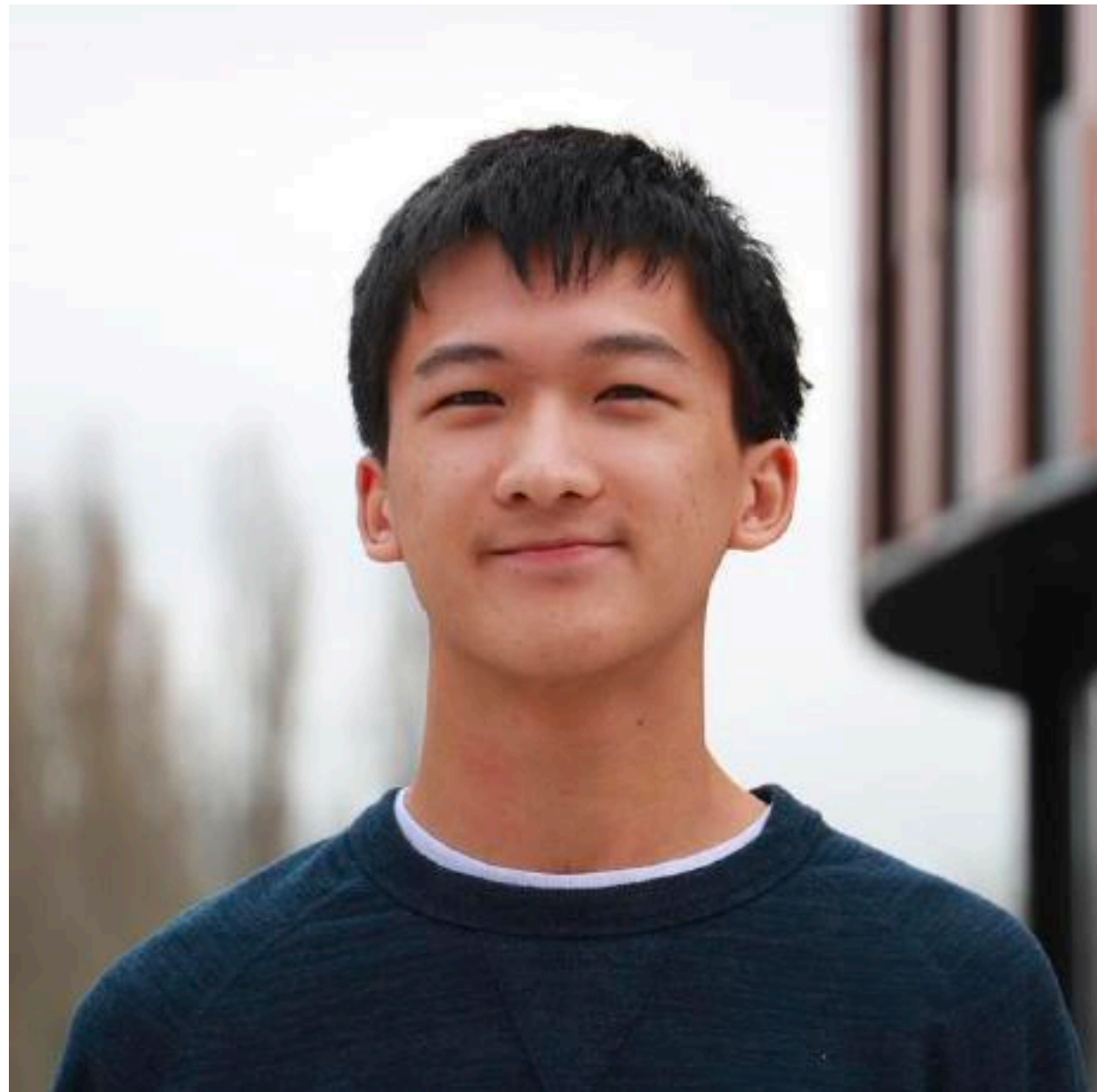
we have used them to build the **pure, low-level, binder free IR Glenside,**

and have shown how they **enable hardware-level tensor program rewriting.**

<https://github.com/gussmith23/glenside>

Glenside is an actively maintained Rust library! Try it out and open issues if you have questions!





Thank you!