Feature selection in high-dimensional dataset using MapReduce

Claudio Reggiani
Yann-Aël Le Borgne
Gianluca Bontempi

BNAIC 2017 – November 8, 2017

mRMR

mRMR

$$\max_{x_j \in X - S_{m-1}}$$

candidate feature *j*

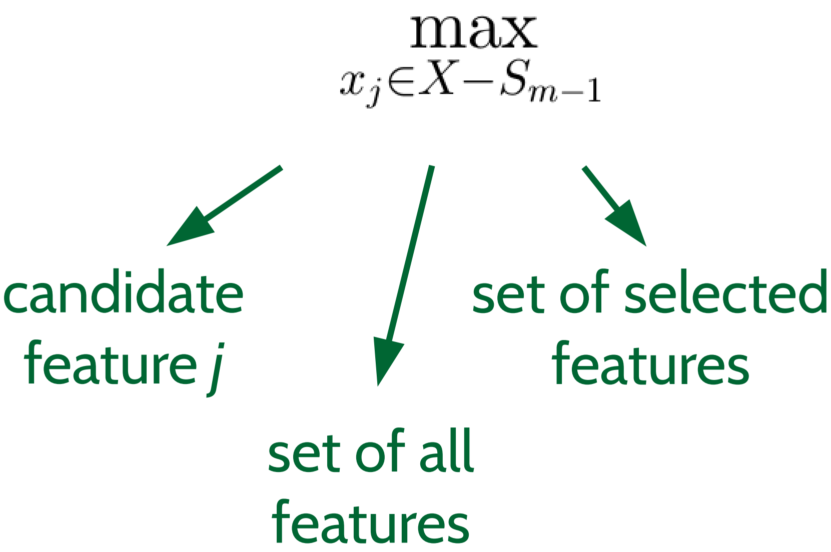
mRMR

$$\max_{x_j \in X - S_{m-1}}$$

candidate feature *j*

set of all features

mRMR



$$\max_{x_j \in X - S_{m-1}}$$

candidate feature *j*

set of all features

set of selected features

mRMR

$$\max_{x_j \in X - S_{m-1}} \left[\overbrace{\qquad\qquad}^{\text{maximal relevance}} \quad \overbrace{\qquad\qquad\qquad\qquad}^{\text{minimal redundancy}} \right]$$

candidate feature *j*

set of all features

set of selected features

mRMR

$$\max_{x_j \in X - S_{m-1}} \left[\overbrace{I(x_j;c)}^{\text{maximal relevance}} \quad \overbrace{\qquad\qquad\qquad\qquad}^{\text{minimal redundancy}} \right]$$

candidate feature *j*

set of selected features

set of all features

mRMR

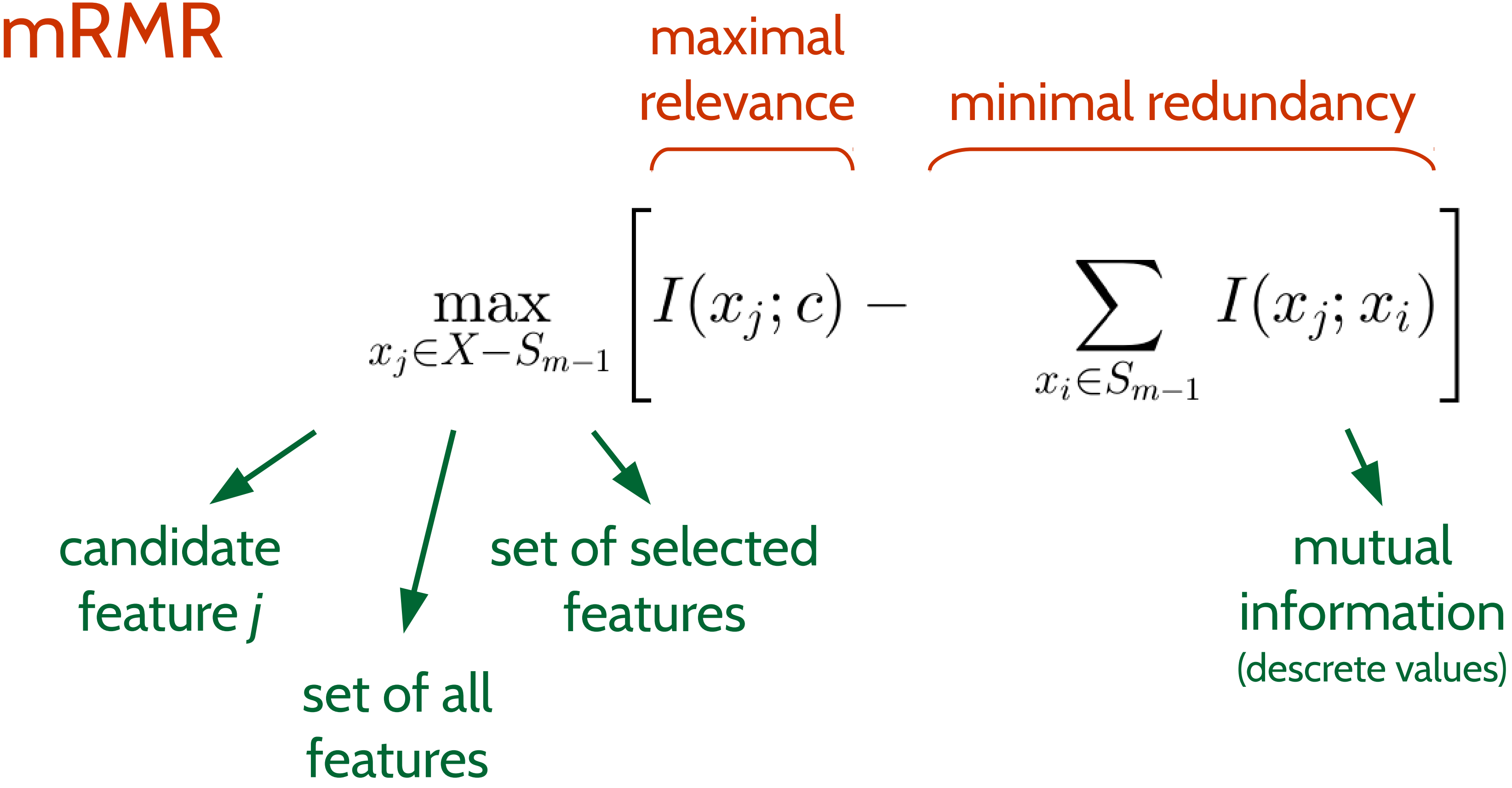
$$\max_{x_j \in X - S_{m-1}} \left[\overbrace{I(x_j; c)}^{\text{maximal relevance}} - \overbrace{\sum_{x_i \in S_{m-1}} I(x_j; x_i)}^{\text{minimal redundancy}} \right]$$

candidate feature *j*

set of all features

set of selected features

mRMR



$$\max_{x_j\in X-S_{m-1}}\left[\overbrace{I(x_j;c)}^{\text{maximal relevance}} - \overbrace{\sum_{x_i\in S_{m-1}} I(x_j;x_i)}^{\text{minimal redundancy}}\right]$$

candidate feature *j*

set of all features

set of selected features

mutual information (descrete values)

mRMR

$$\max_{x_j \in X - S_{m-1}} \left[\overbrace{I(x_j; c)}^{\text{maximal relevance}} - \overbrace{\frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i)}^{\text{minimal redundancy}} \right]$$

candidate feature *j*

set of all features

set of selected features

mutual information (descrete values)

mRMR

$$\max_{x_j \in X - S_{m-1}} \left[\overbrace{I(x_j; c)}^{\text{maximal relevance}} - \overbrace{\frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i)}^{\text{minimal redundancy}} \right]$$

candidate feature *j*

set of all features

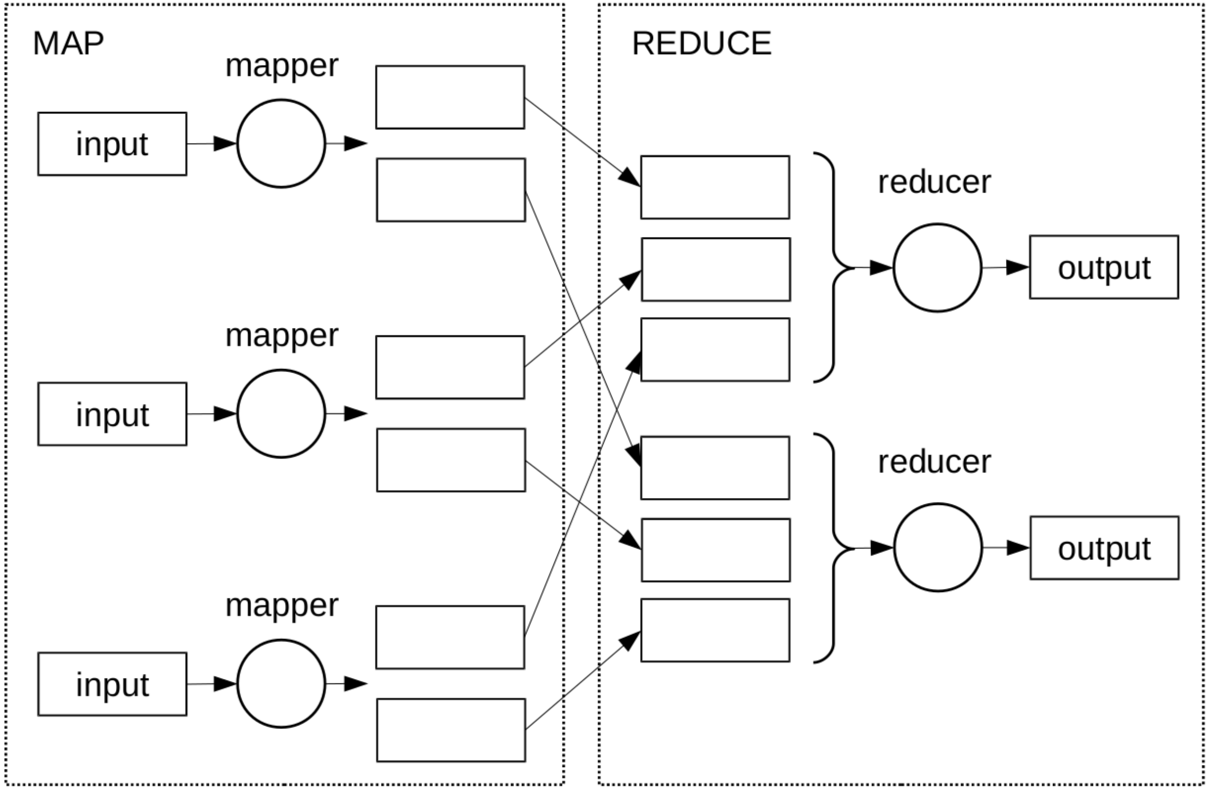
set of selected features

cardinality of selected features set

mutual information
(descrete values)

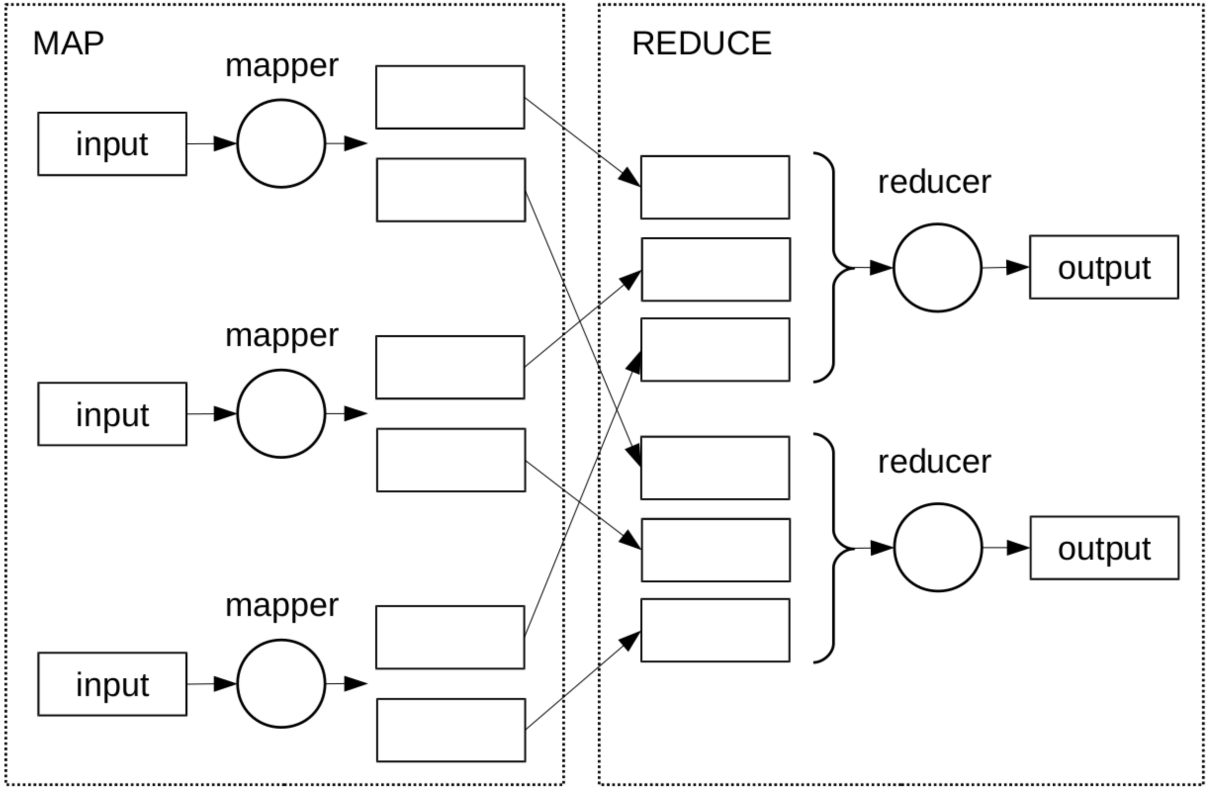
How do we cope with high-dimensional datasets?

Brief introduction to MapReduce



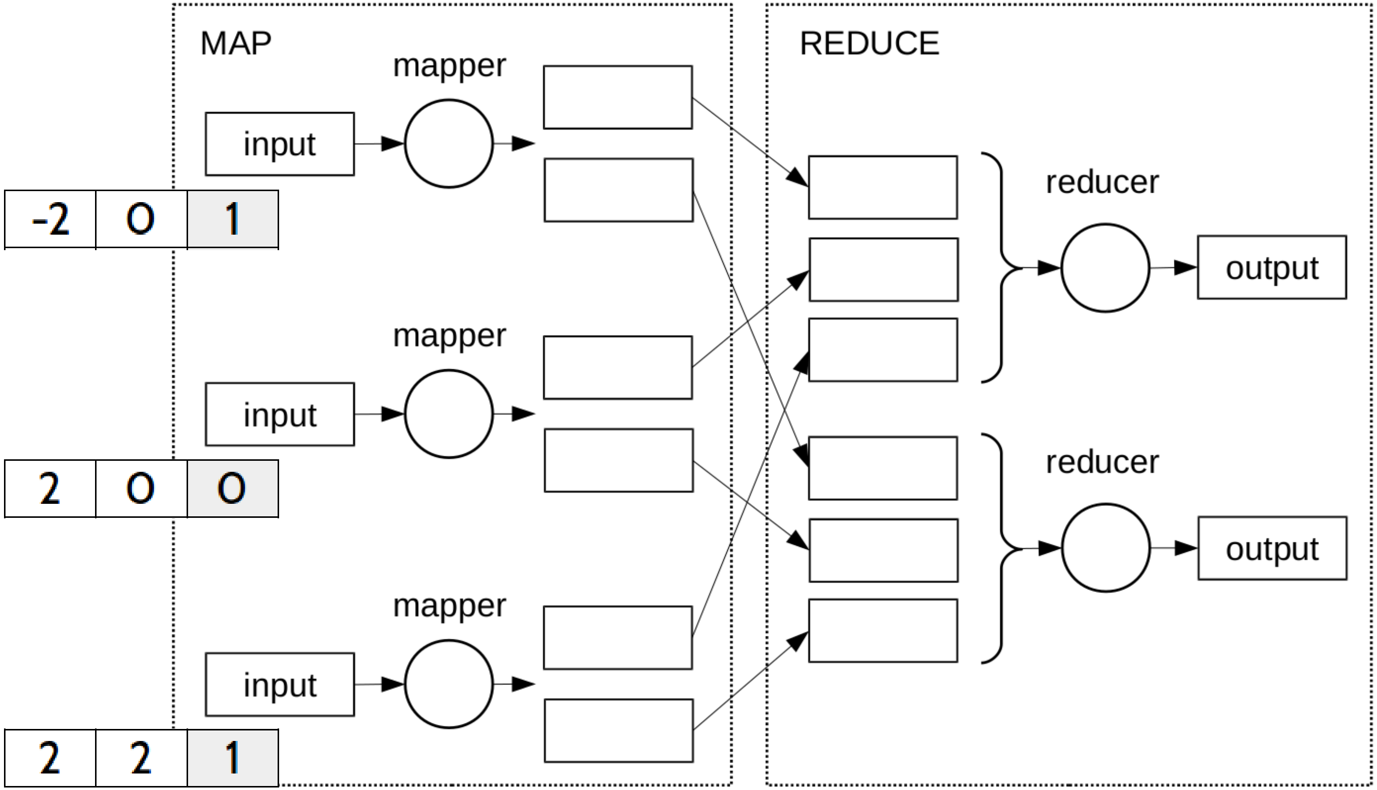
Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



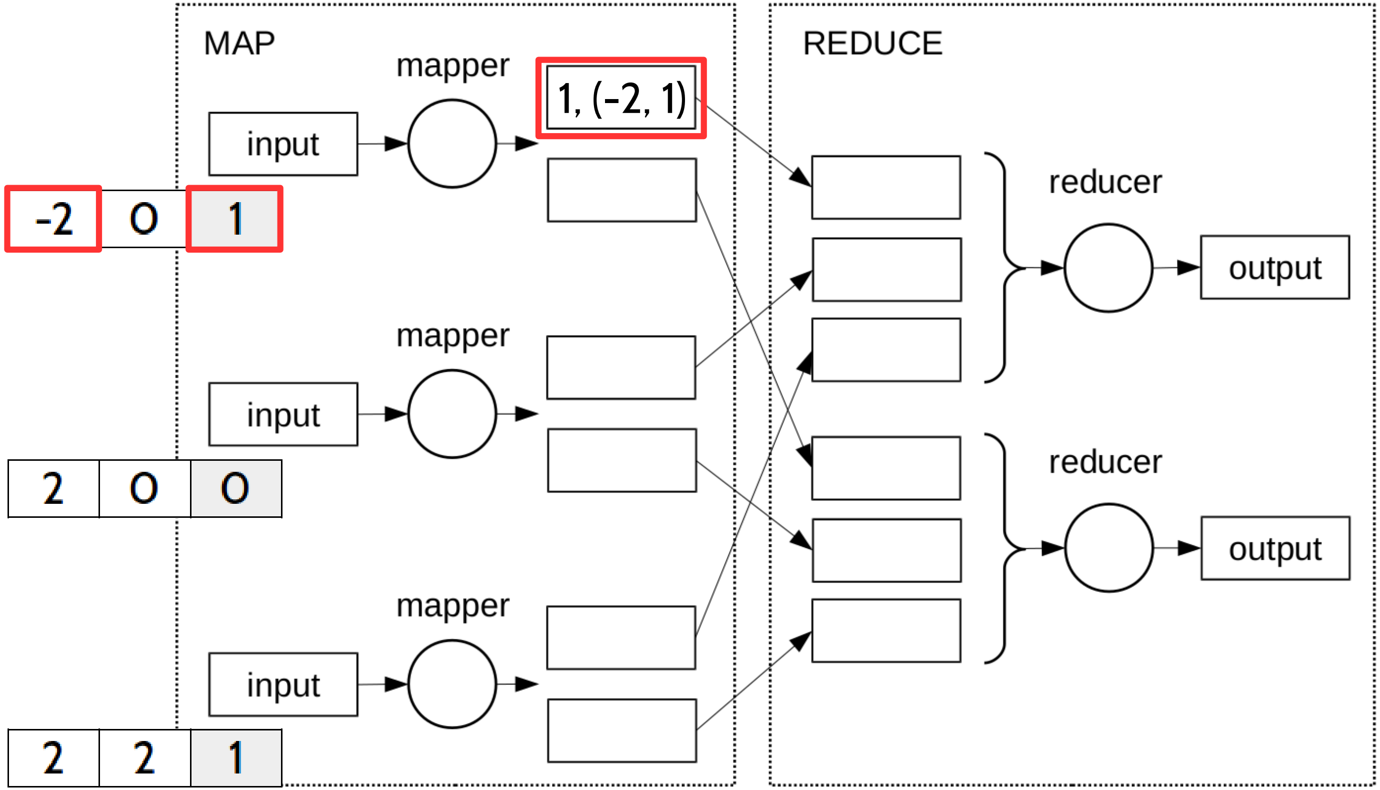
-2	0	1

2	0	0

2	2	1

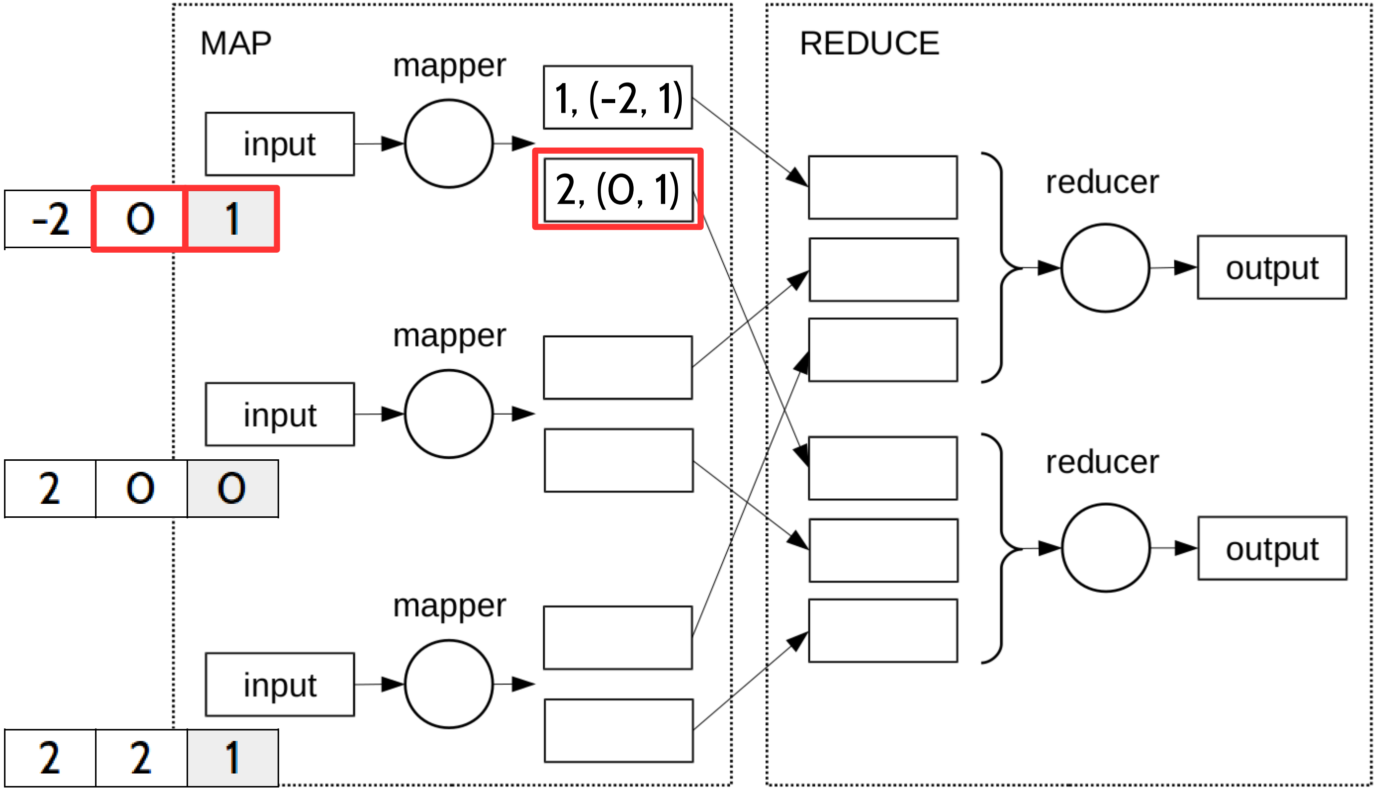
Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



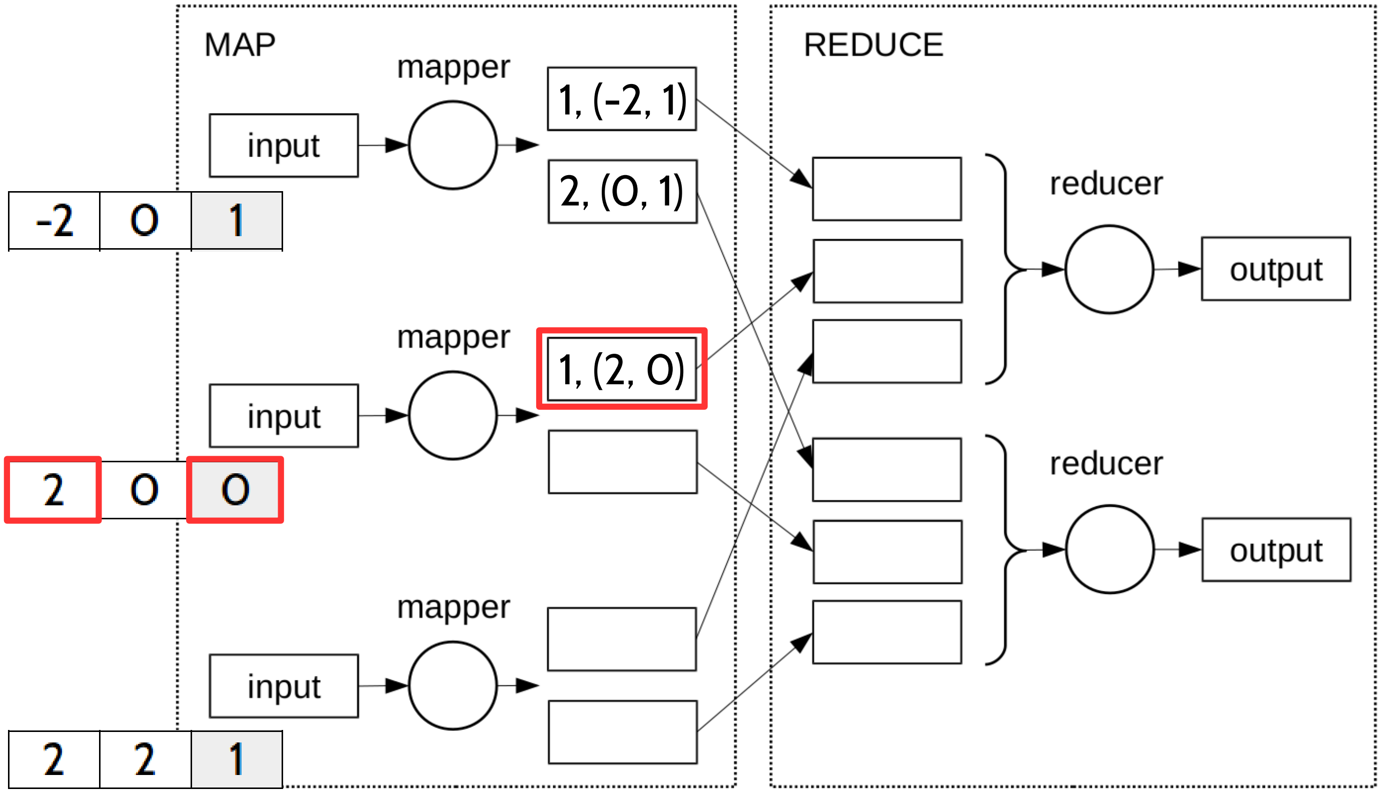
Brief introduction to MapReduce

1	*2*	
-2	0	1
2	0	0
2	2	1



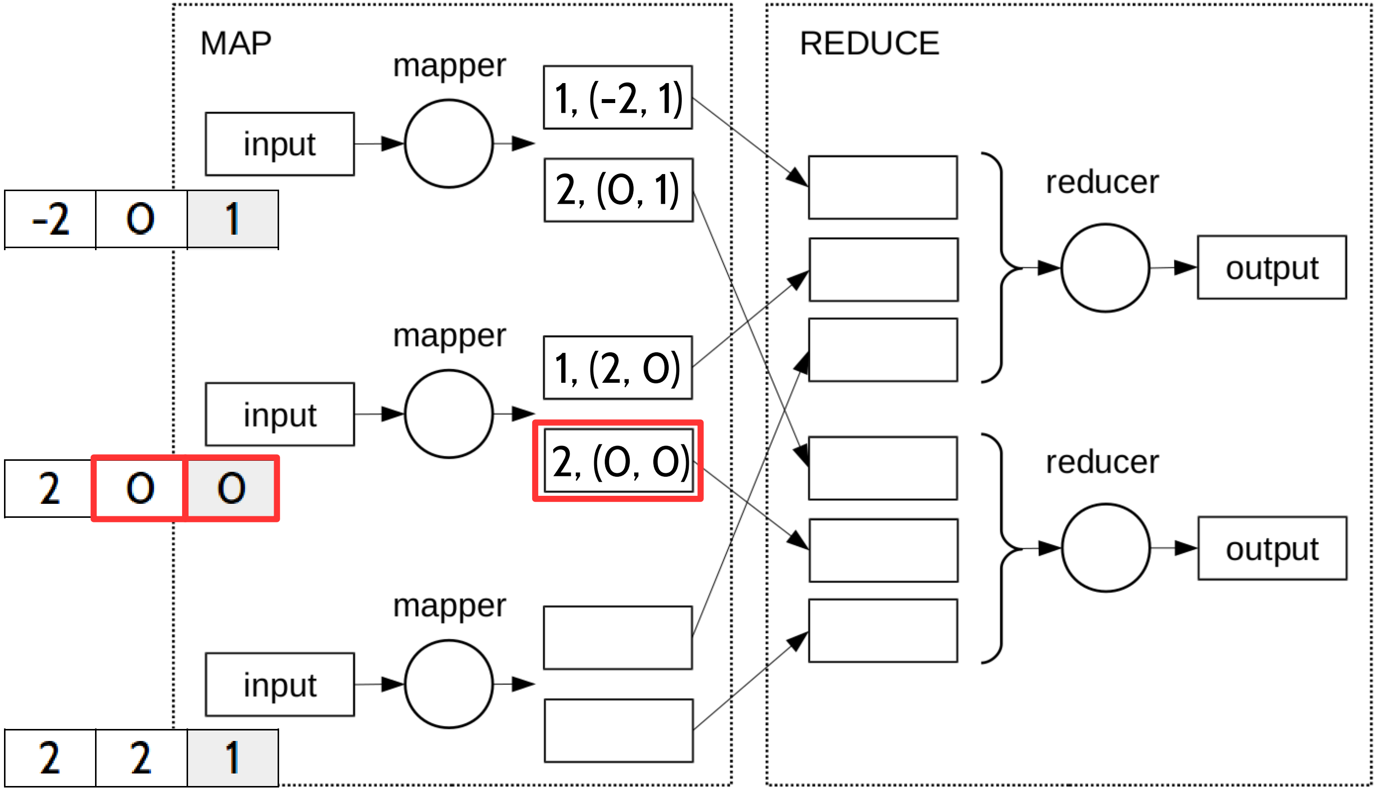
Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



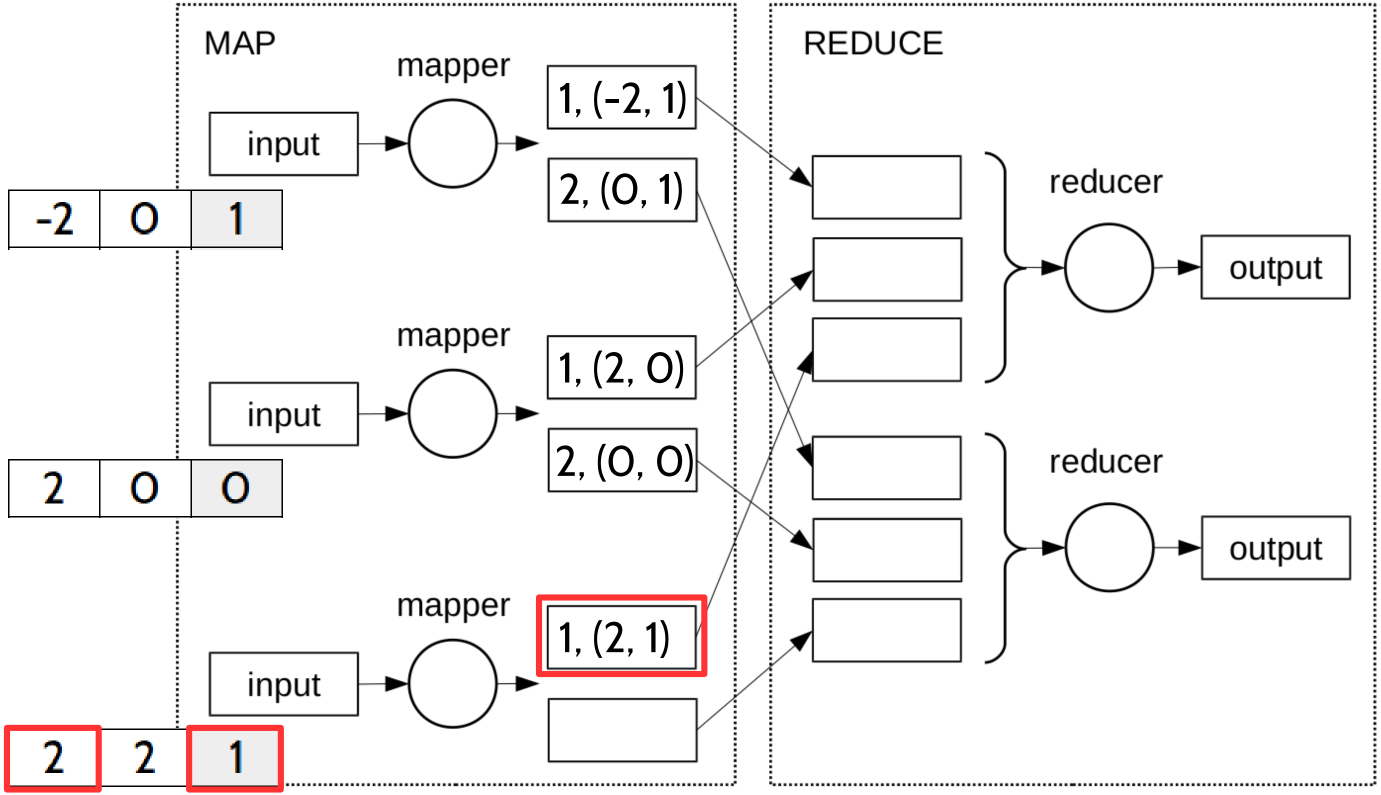
Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



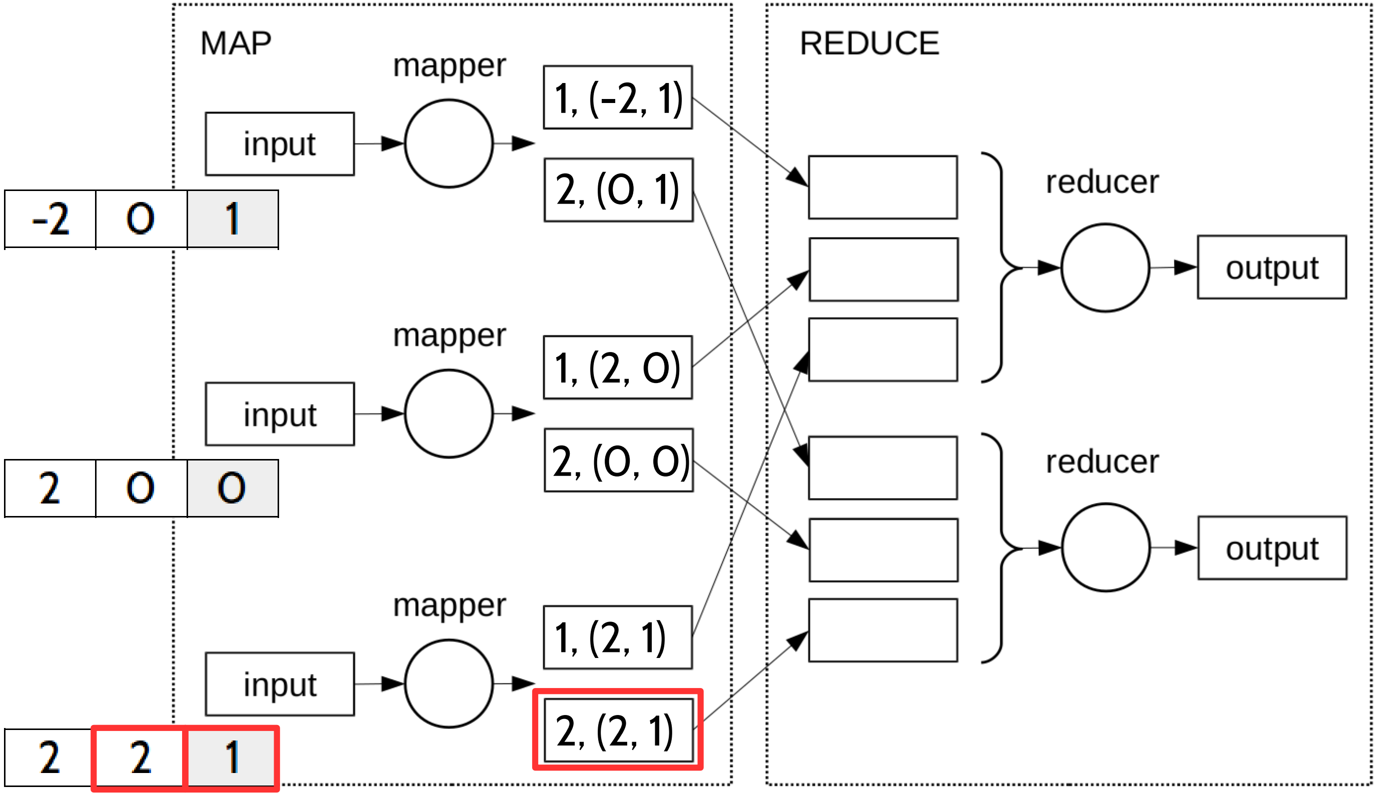
Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1

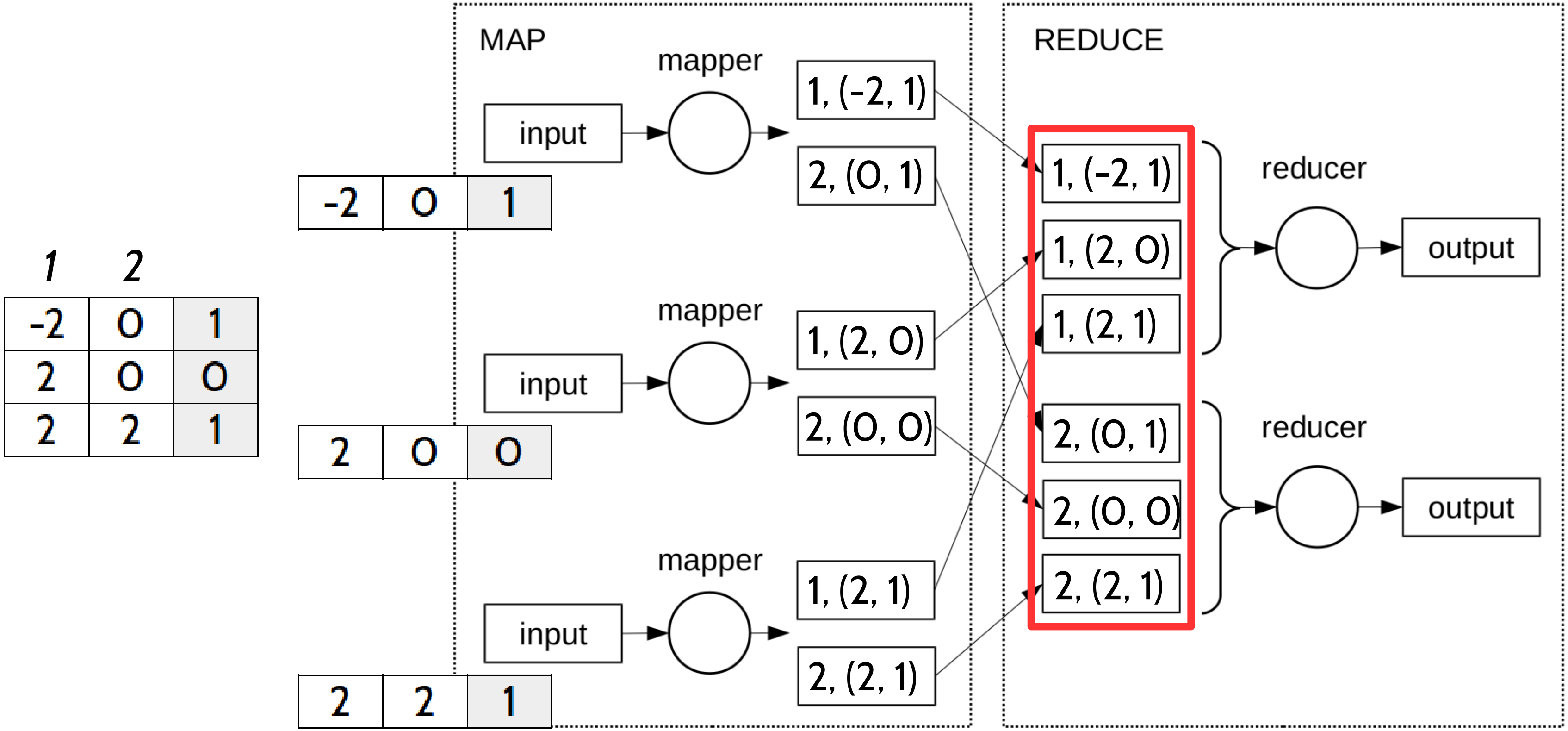


Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



Brief introduction to MapReduce



1	2	
-2	0	1
2	0	0
2	2	1

-2 0 1

2 0 0

2 2 1

MAP

input → mapper → 1, (-2, 1) / 2, (0, 1)

input → mapper → 1, (2, 0) / 2, (0, 0)

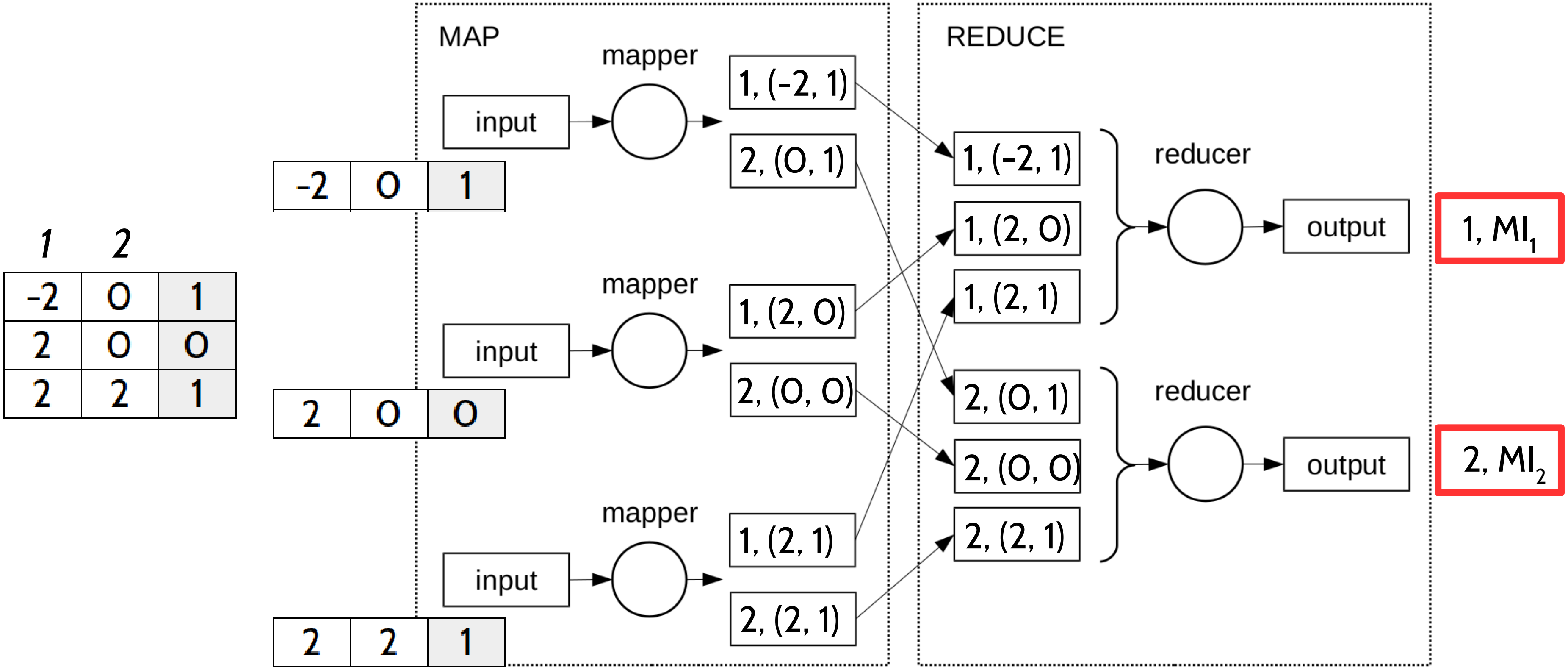
input → mapper → 1, (2, 1) / 2, (2, 1)

REDUCE

1, (-2, 1)
1, (2, 0)
1, (2, 1)
→ reducer → output

2, (0, 1)
2, (0, 0)
2, (2, 1)
→ reducer → output

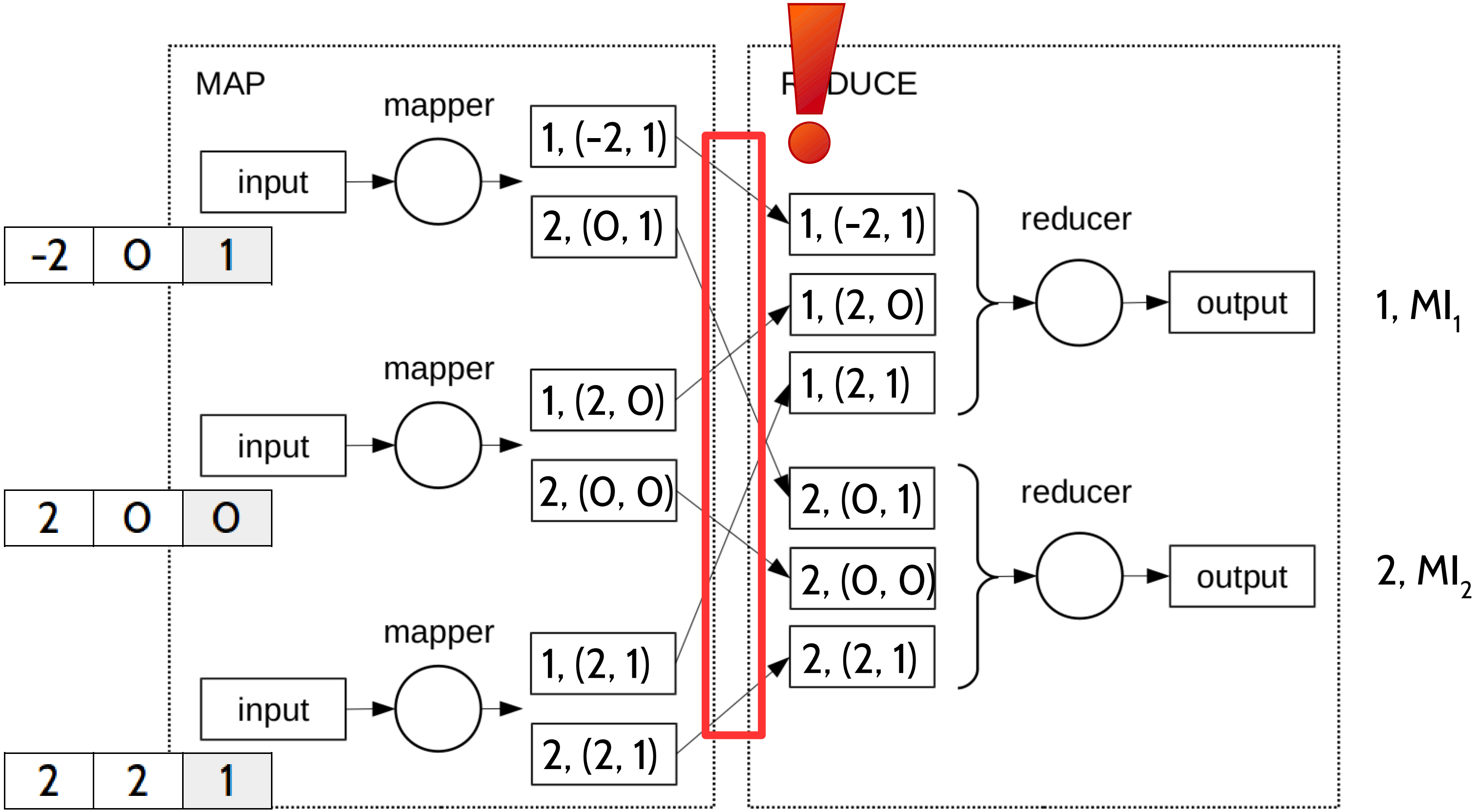
Brief introduction to MapReduce



1	*2*	
-2	0	1
2	0	0
2	2	1

Brief introduction to MapReduce

1	2	
-2	0	1
2	0	0
2	2	1



Brief introduction to MapReduce

-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1

...

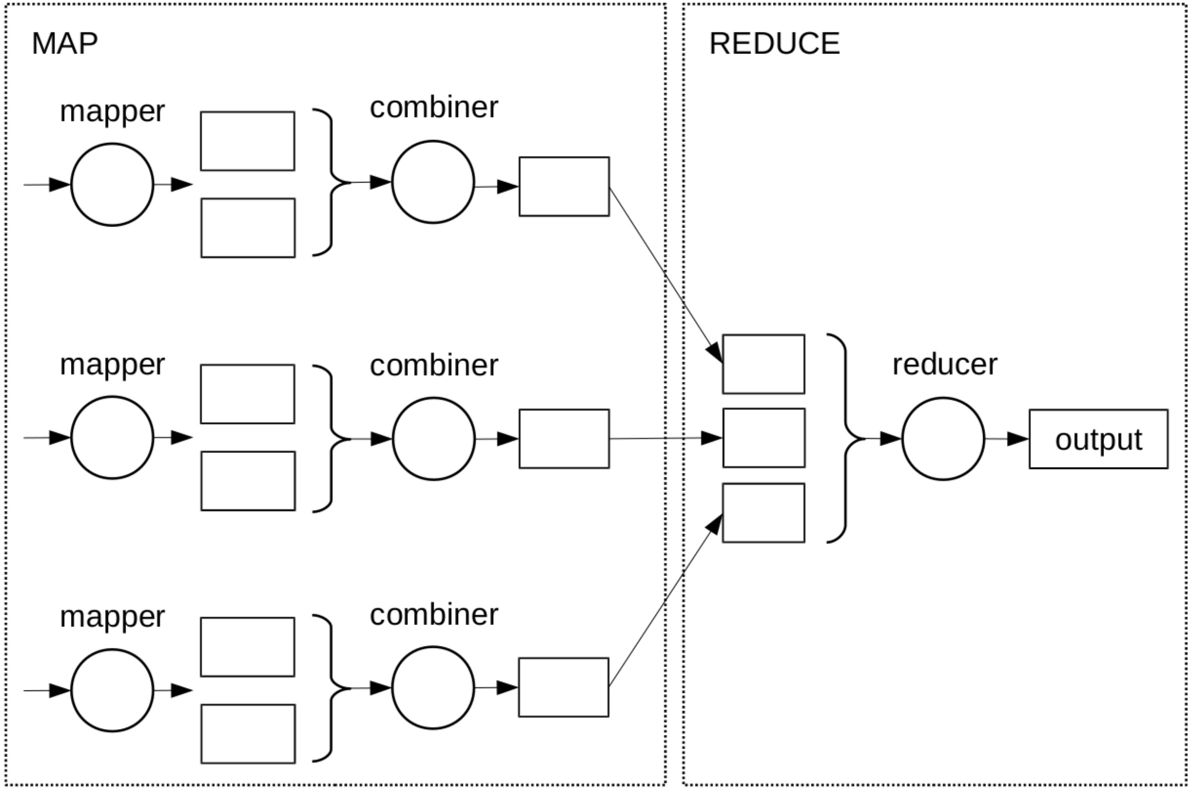
-2	0	1
2	0	0
2	2	1

Brief introduction to MapReduce

-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1

...

-2	0	1
2	0	0
2	2	1

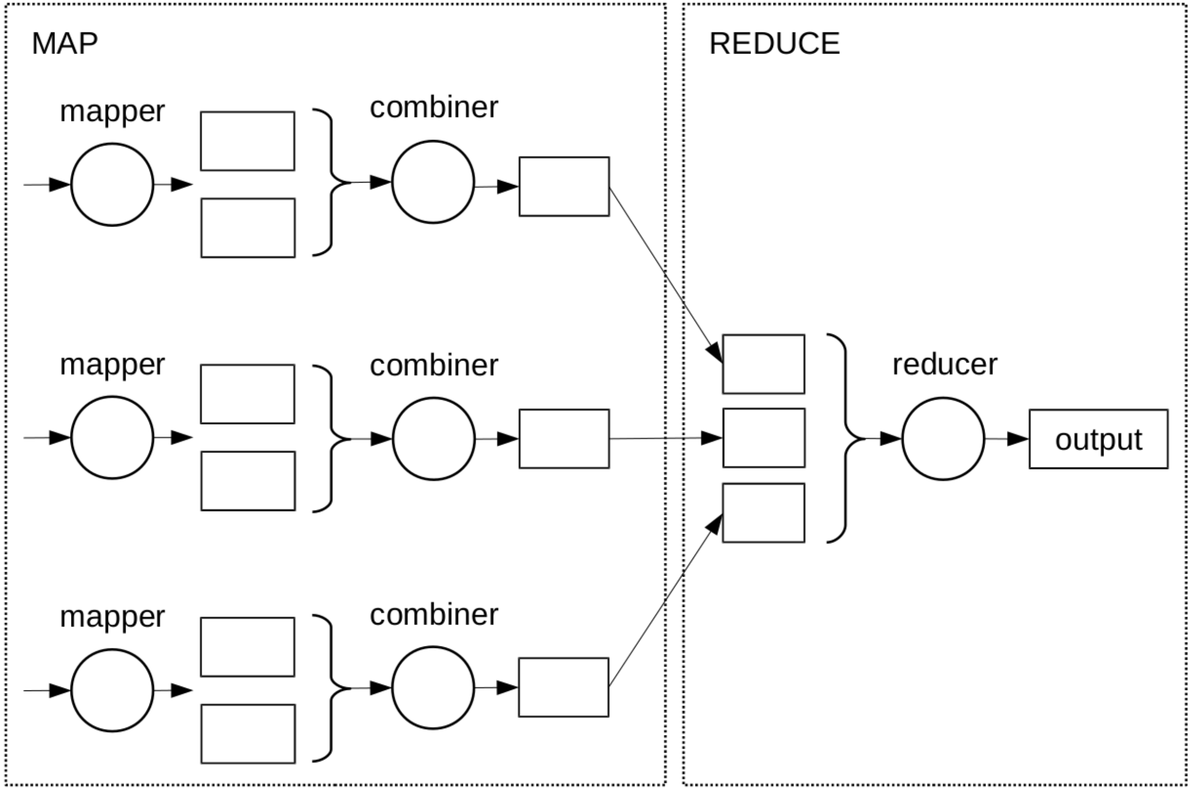


Brief introduction to MapReduce

-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1

...

-2	0	1
2	0	0
2	2	1

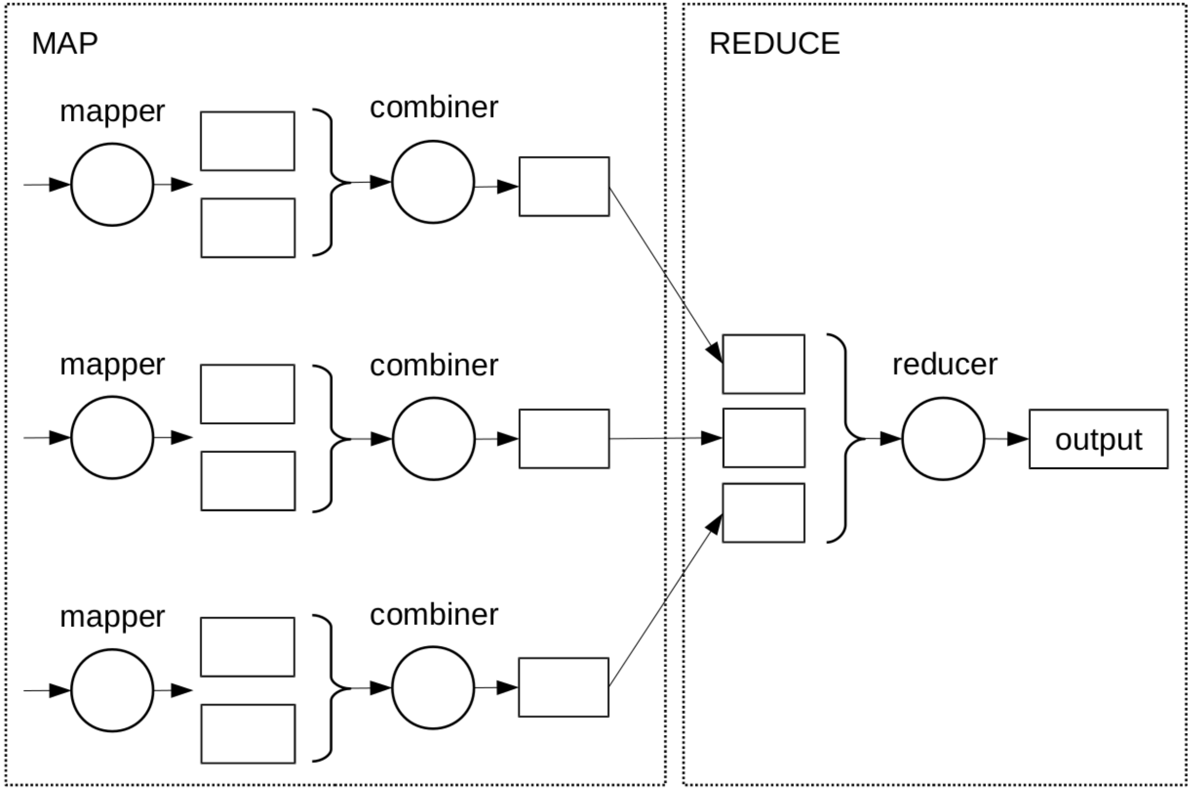


Brief introduction to MapReduce

-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1

...

-2	0	1
2	0	0
2	2	1



1, (-2, 1)

1, (-2, 1)

1, (-2, 1)

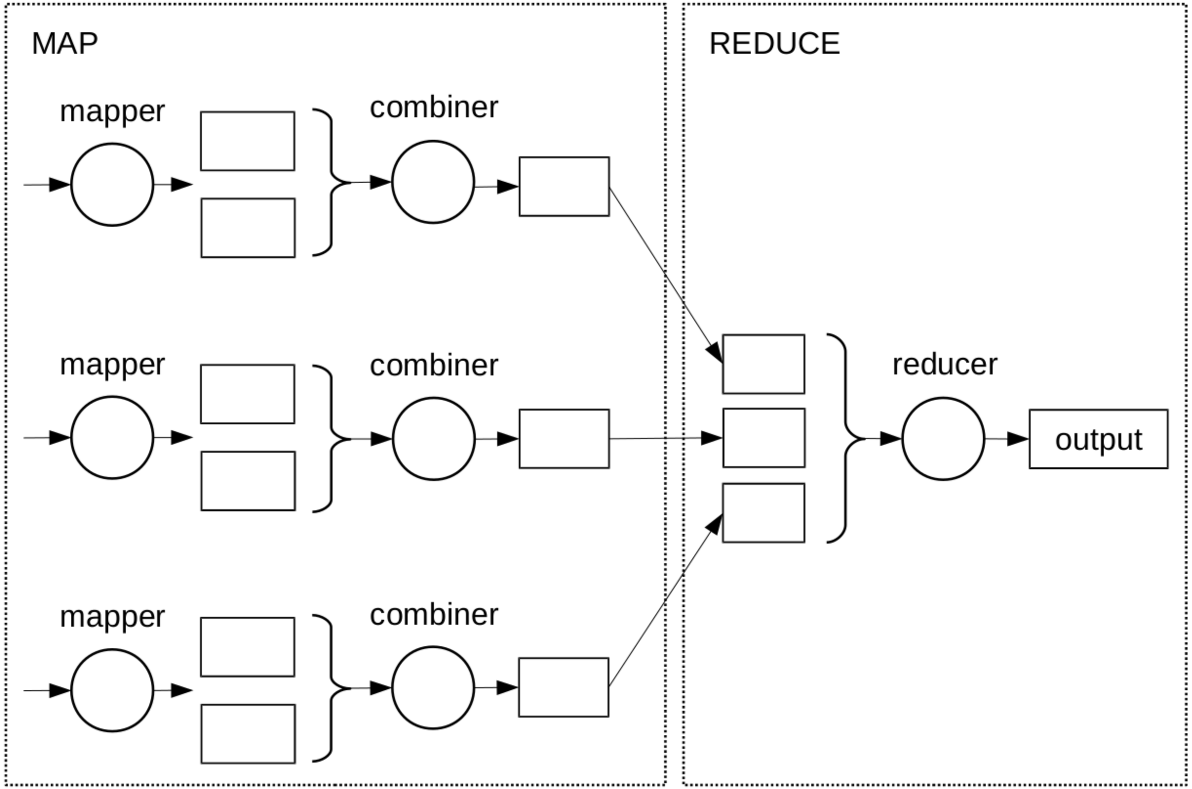
1, (-2, 1)

Brief introduction to MapReduce

-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1
-2	0	1
2	0	0
2	2	1

...

-2	0	1
2	0	0
2	2	1



1, (-2, 1)

1, (-2, 1)

1, (-2, 1)

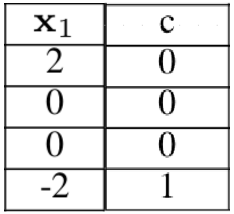
1, (-2, 1)

?

Data structure with cumulative property

Data structure with cumulative property

MAP

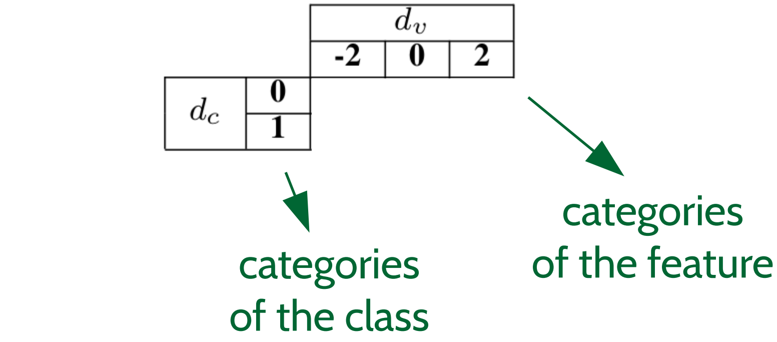


$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

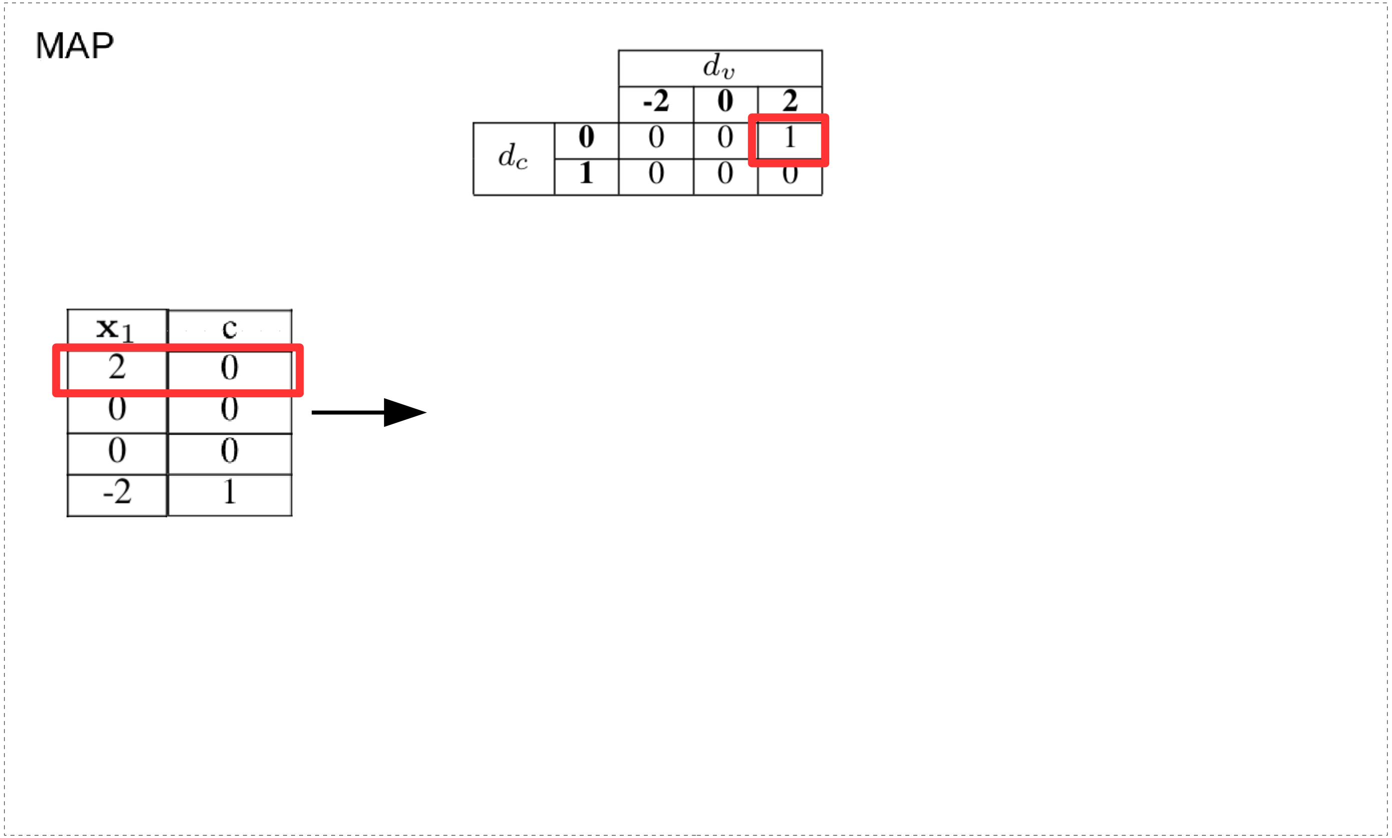
Data structure with cumulative property

MAP

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1



Data structure with cumulative property

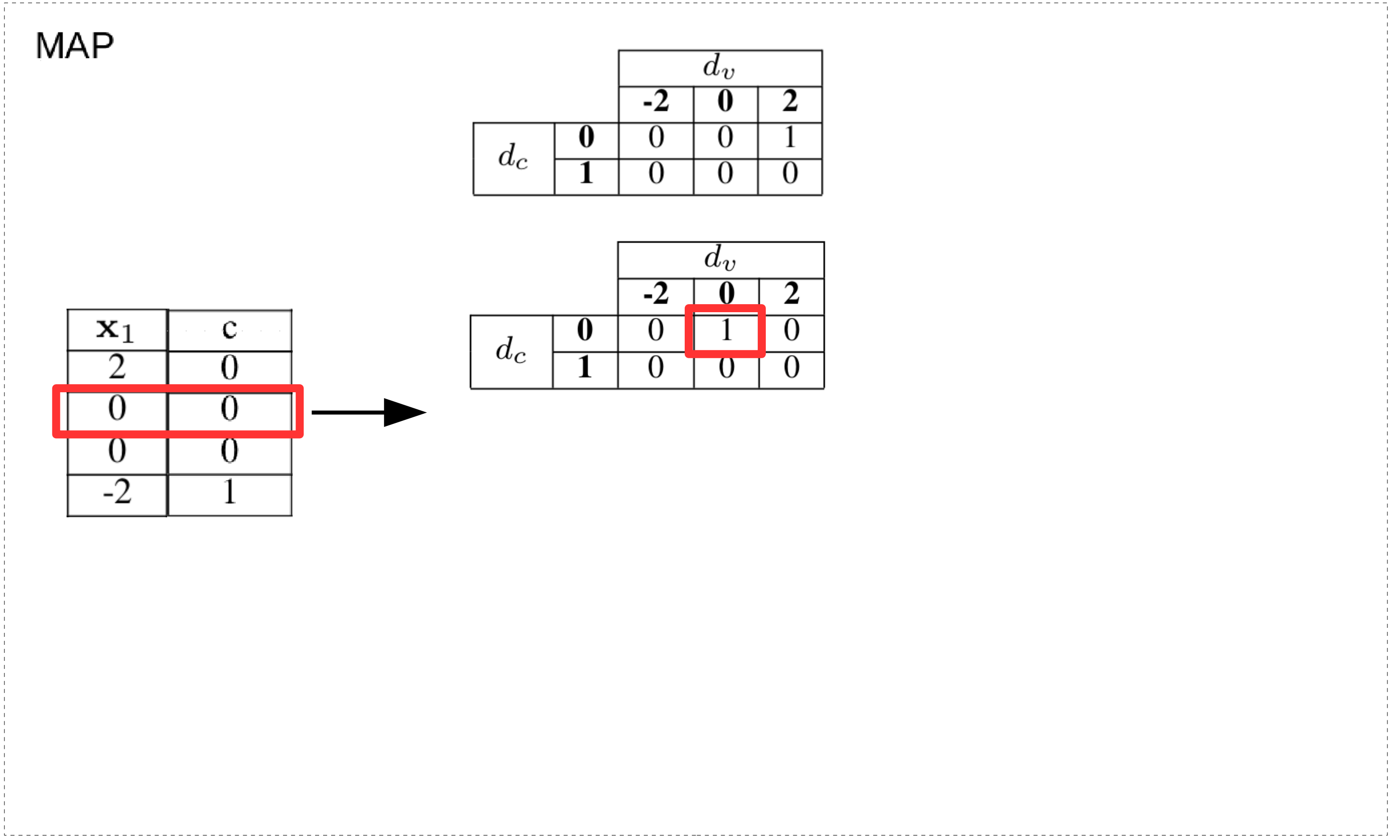


MAP

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

Data structure with cumulative property



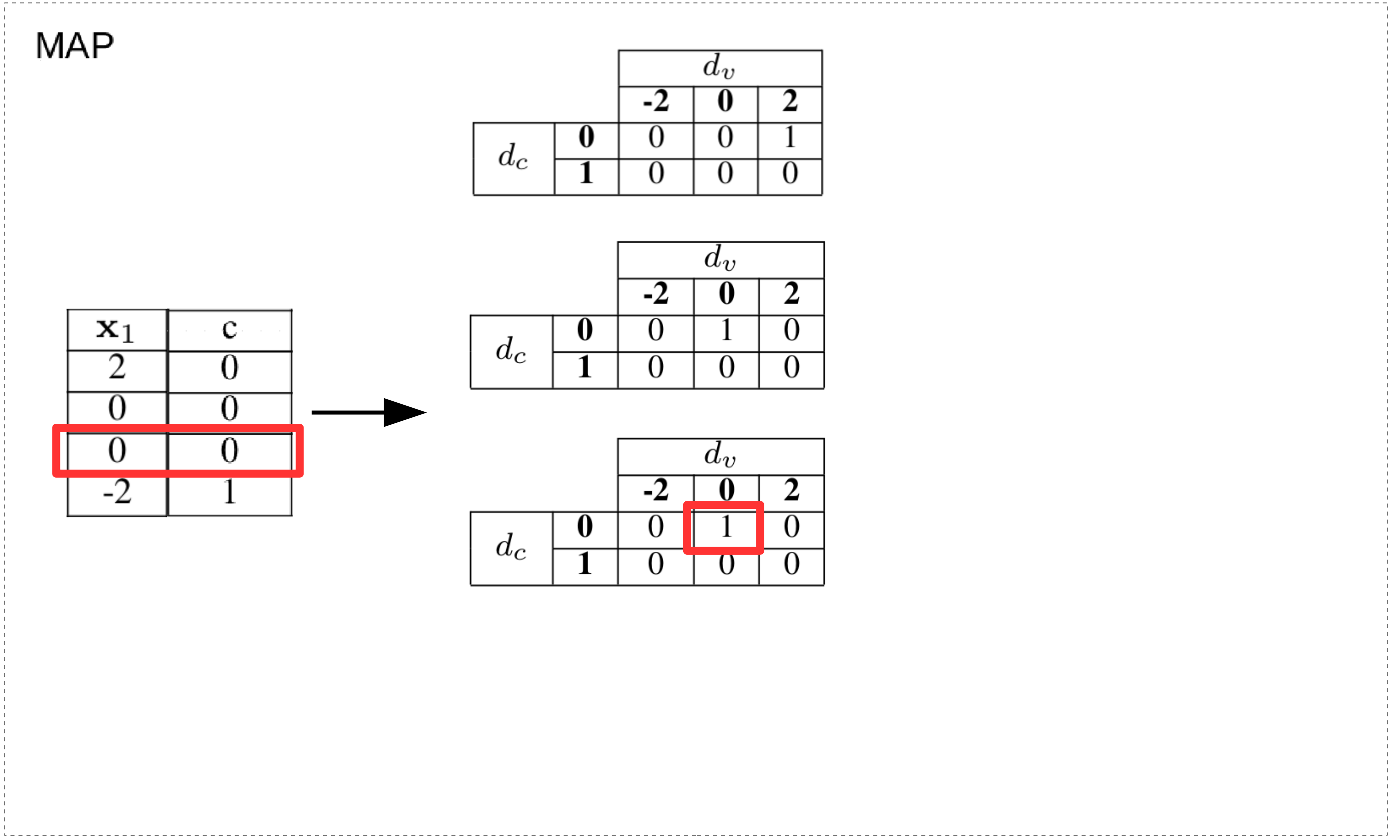
MAP

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

Data structure with cumulative property



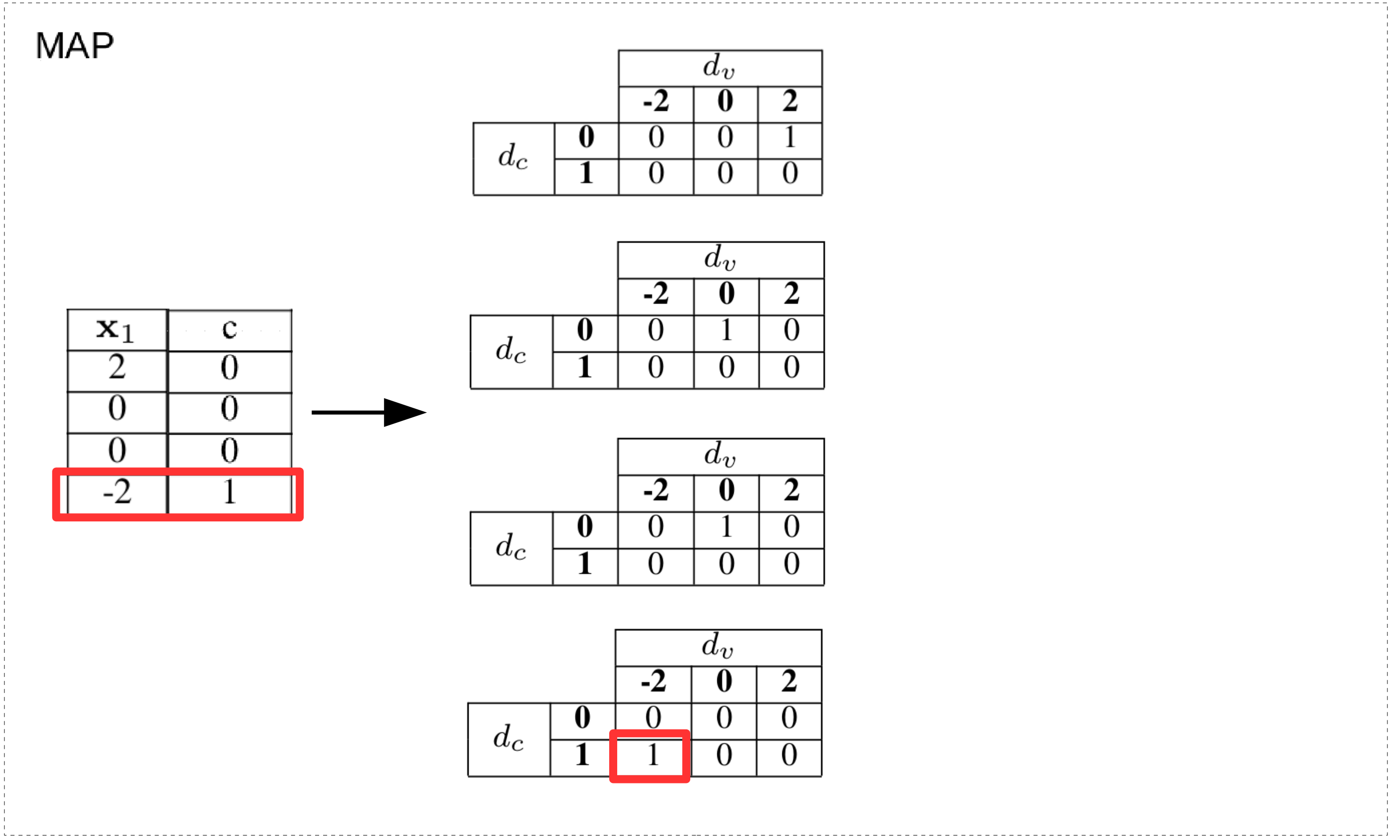
$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

Data structure with cumulative property



MAP

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

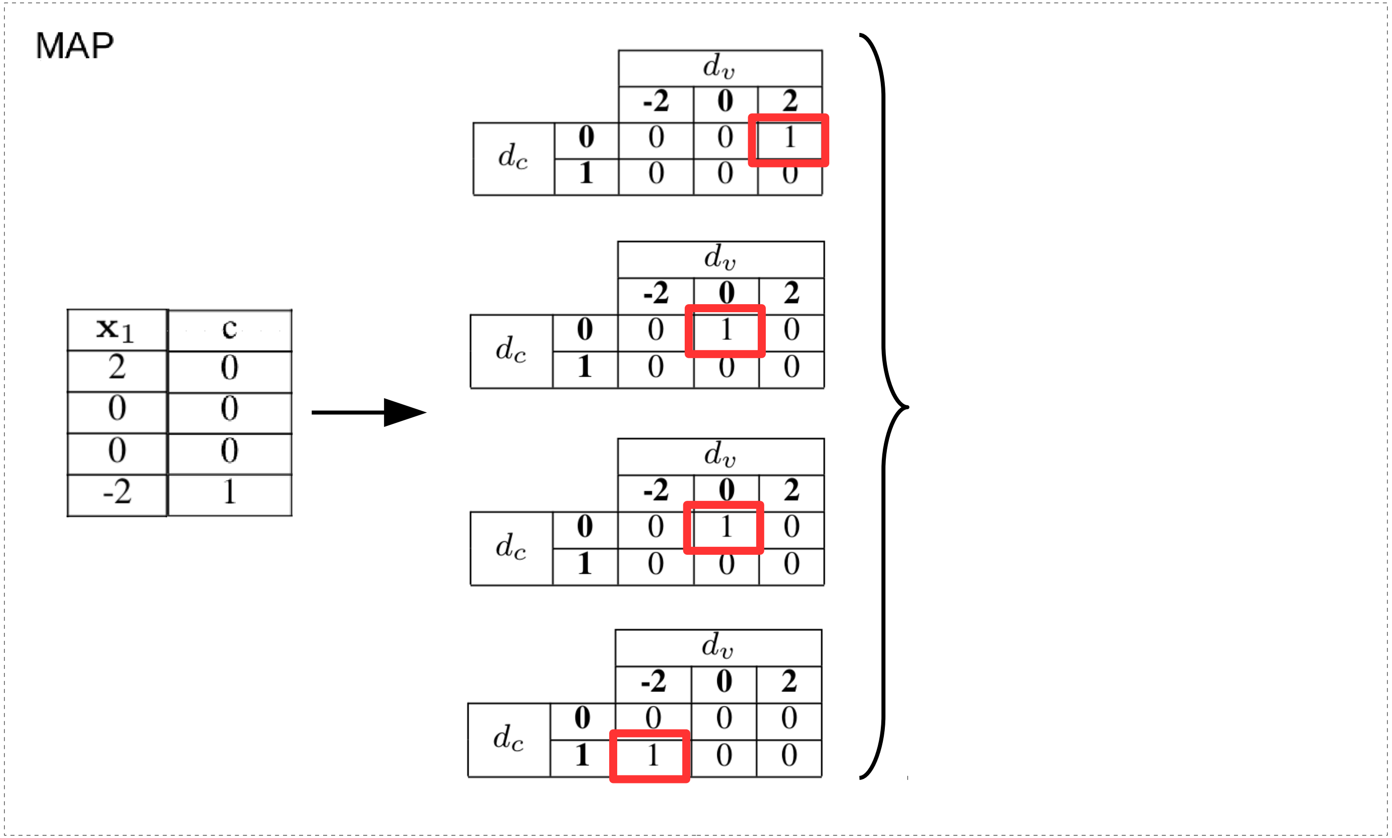
		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	0
	1	1	0	0

Data structure with cumulative property



$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

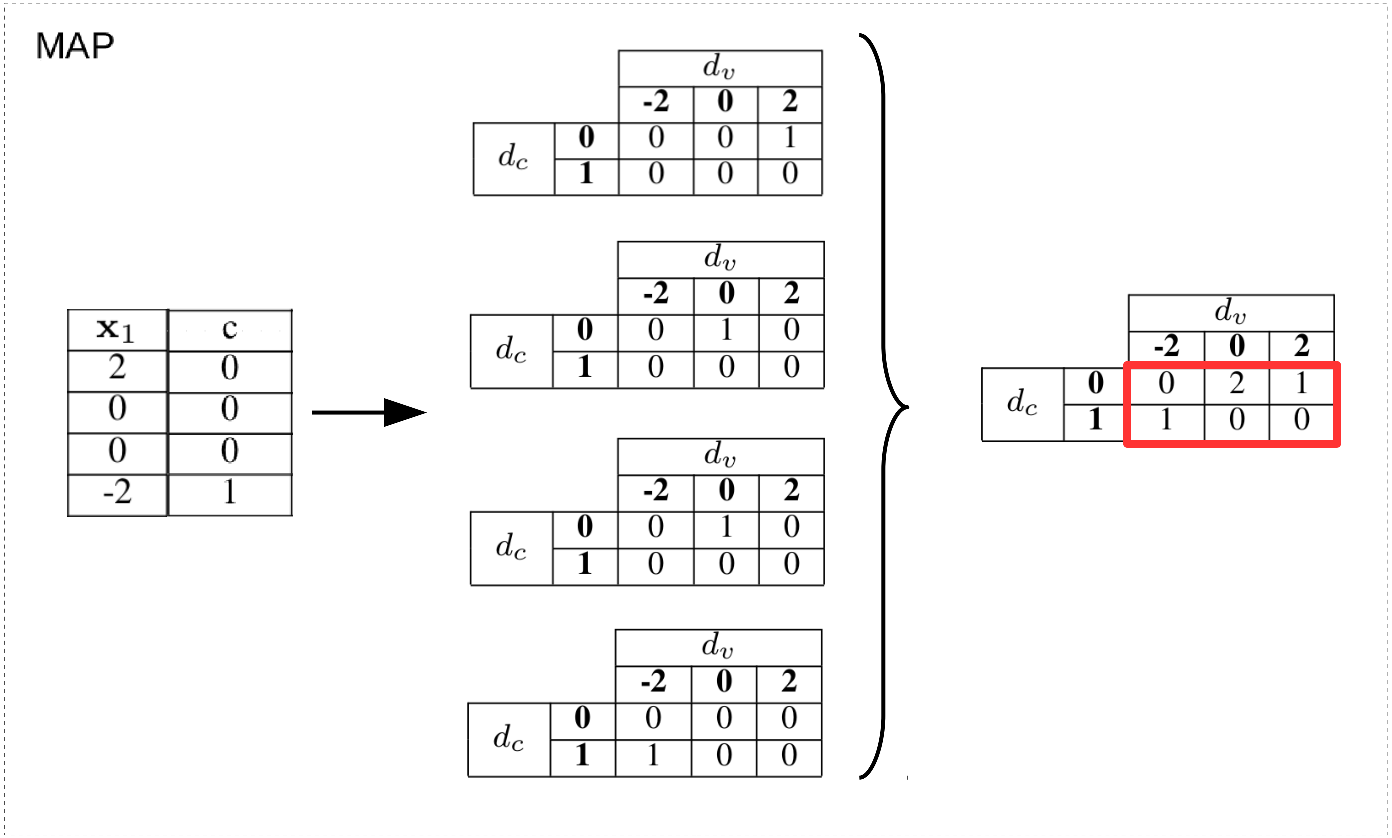
		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	0
	1	1	0	0

Data structure with cumulative property



MAP

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

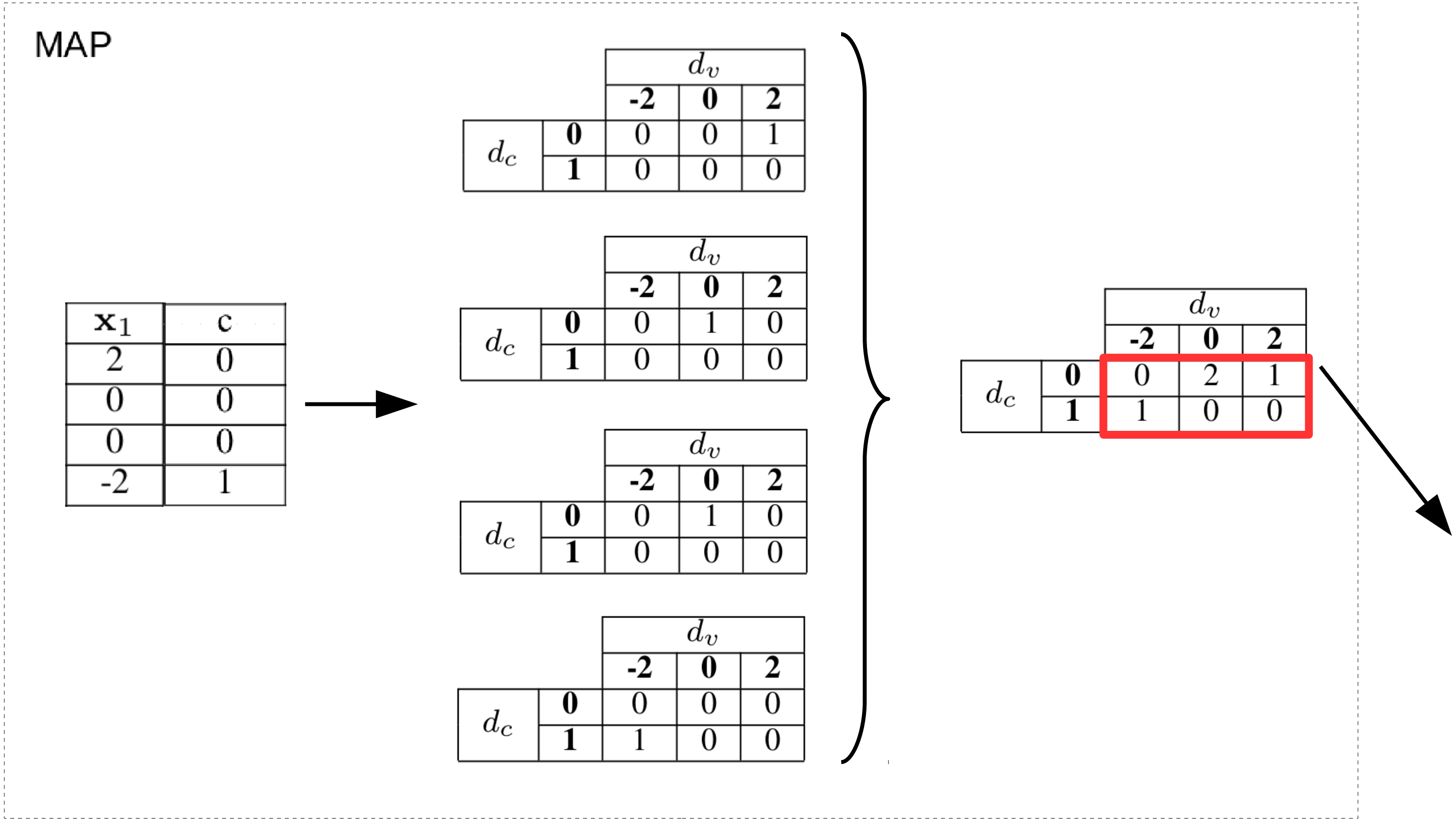
		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	0
	1	1	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	2	1
	1	1	0	0

Data structure with cumulative property



MAP

$\mathbf{x}_1$	c
2	0
0	0
0	0
-2	1

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	1
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

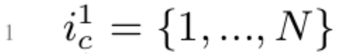
		d_v		
		-2	**0**	**2**
d_c	**0**	0	1	0
	1	0	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	0	0
	1	1	0	0

		d_v		
		-2	**0**	**2**
d_c	**0**	0	2	1
	1	1	0	0

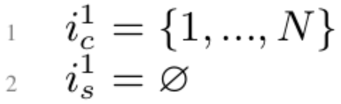
mRMR in MapReduce

mRMR in MapReduce



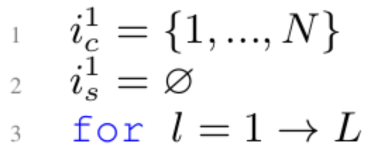
```
$i_c^1 = \{1, ..., N\}$
```

mRMR in MapReduce



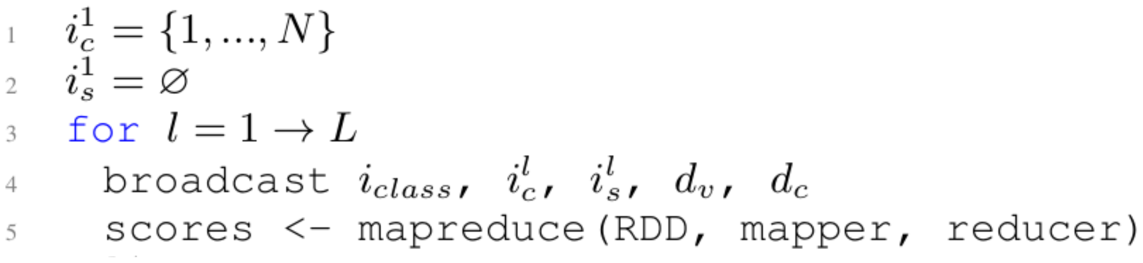
```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
```

mRMR in MapReduce

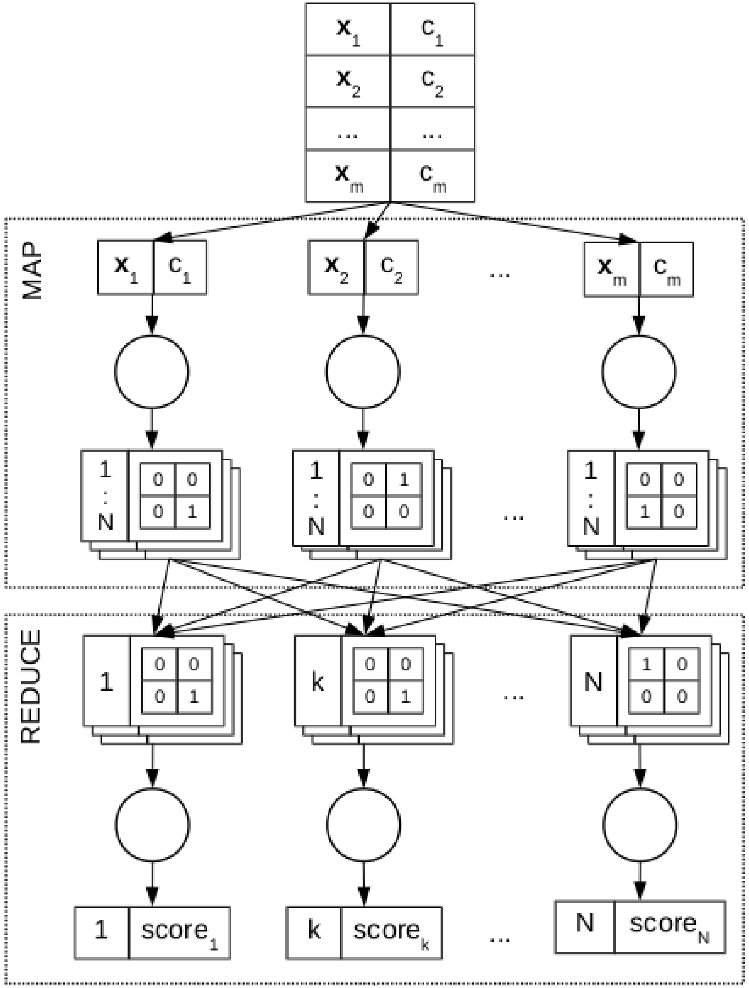


```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
for $l = 1 \rightarrow L$
```

mRMR in MapReduce

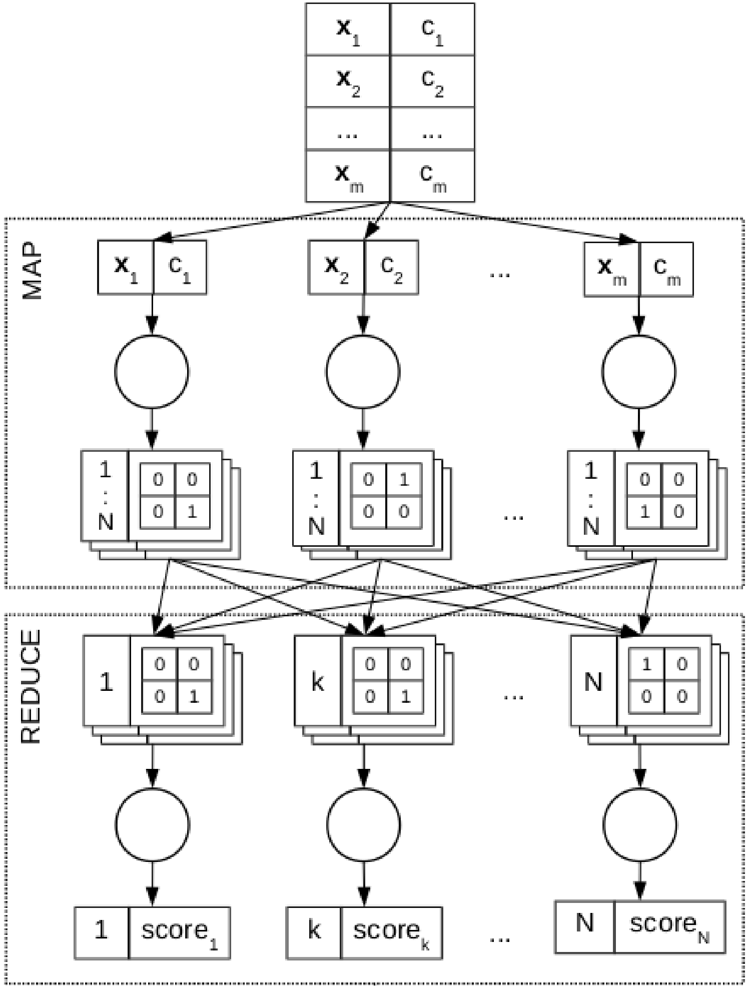


```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
for $l = 1 \rightarrow L$
  broadcast $i_{class}$, $i_c^l$, $i_s^l$, $d_v$, $d_c$
  scores <- mapreduce(RDD, mapper, reducer)
```



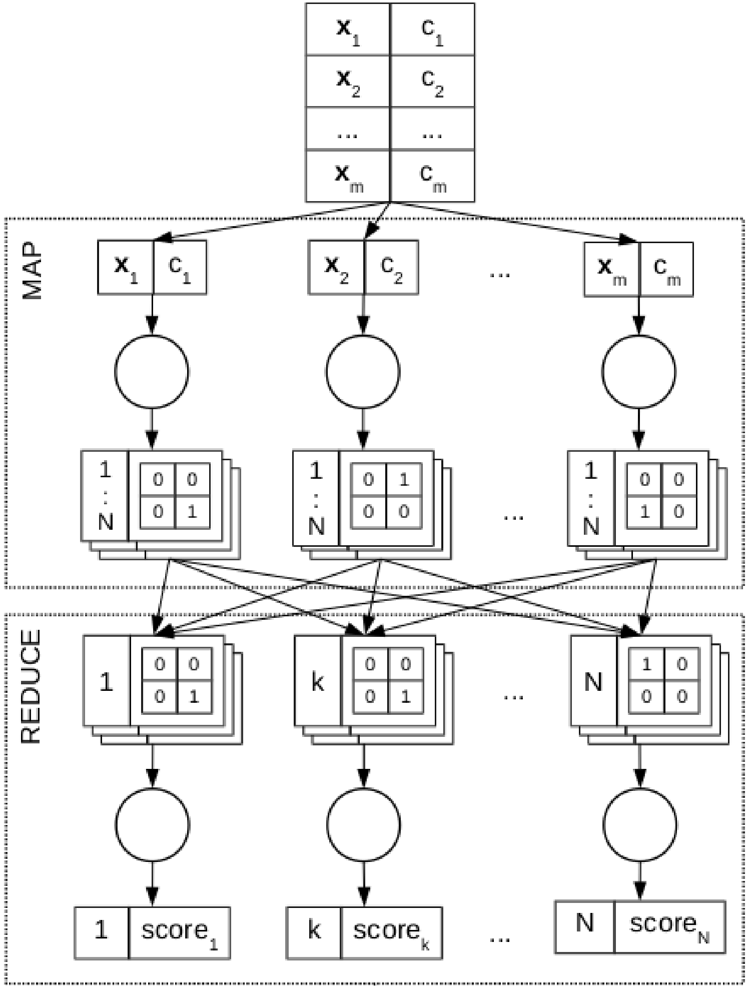
mRMR in MapReduce

```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
for $l = 1 \rightarrow L$
  broadcast $i_{class}$, $i_c^l$, $i_s^l$, $d_v$, $d_c$
  scores <- mapreduce(RDD, mapper, reducer)
  $k^*$ $\leftarrow$ collectArgmax(scores)
```



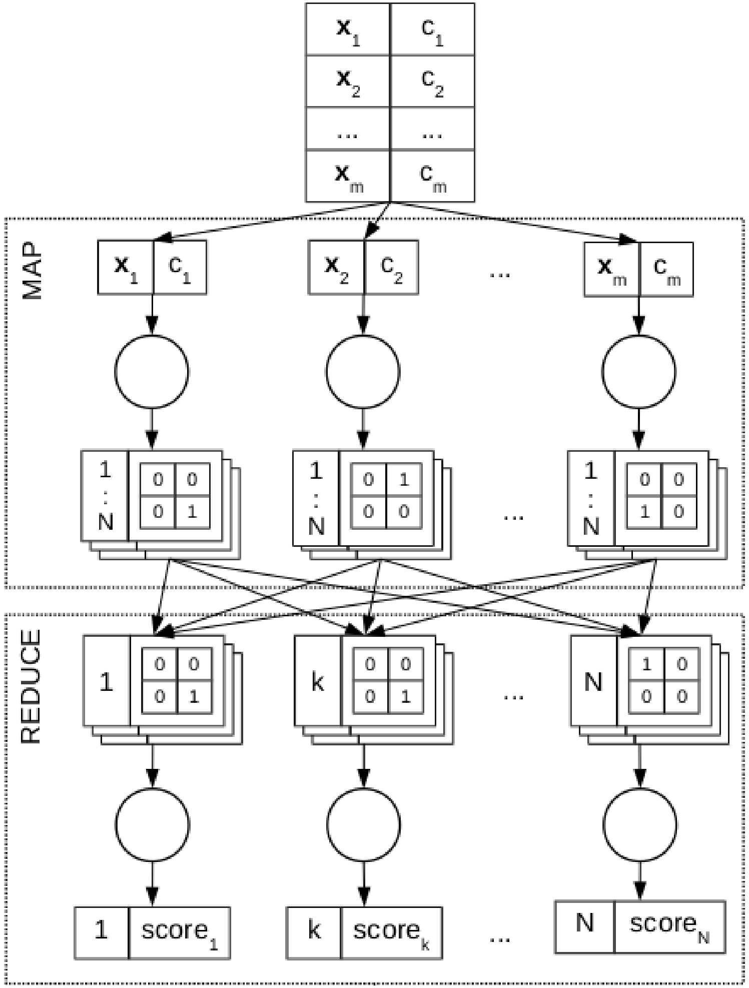
mRMR in MapReduce

```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
for $l = 1 \to L$
  broadcast $i_{class}$, $i_c^l$, $i_s^l$, $d_v$, $d_c$
  scores <- mapreduce(RDD, mapper, reducer)
  $k^* \leftarrow$ collectArgmax(scores)
  $i_c^{l+1} \leftarrow i_c^l \backslash k^*$
  $i_s^{l+1} \leftarrow i_s^l \cup k^*$
```

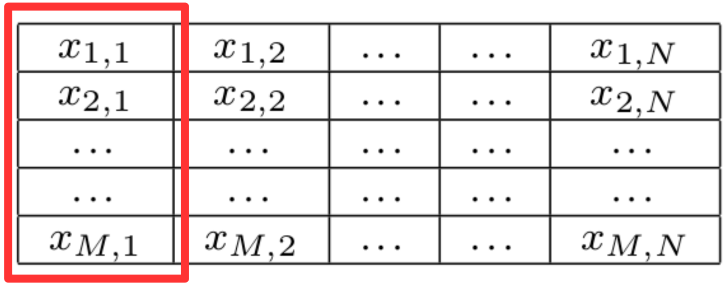


mRMR in MapReduce

```
$i_c^1 = \{1, ..., N\}$
$i_s^1 = \varnothing$
for $l = 1 \to L$
  broadcast $i_{class}$, $i_c^l$, $i_s^l$, $d_v$, $d_c$
  scores <- mapreduce(RDD, mapper, reducer)
  $k^* \leftarrow$ collectArgmax(scores)
  $i_c^{l+1} \leftarrow i_c^l \backslash k^*$
  $i_s^{l+1} \leftarrow i_s^l \cup k^*$
output $i_s^L$
```

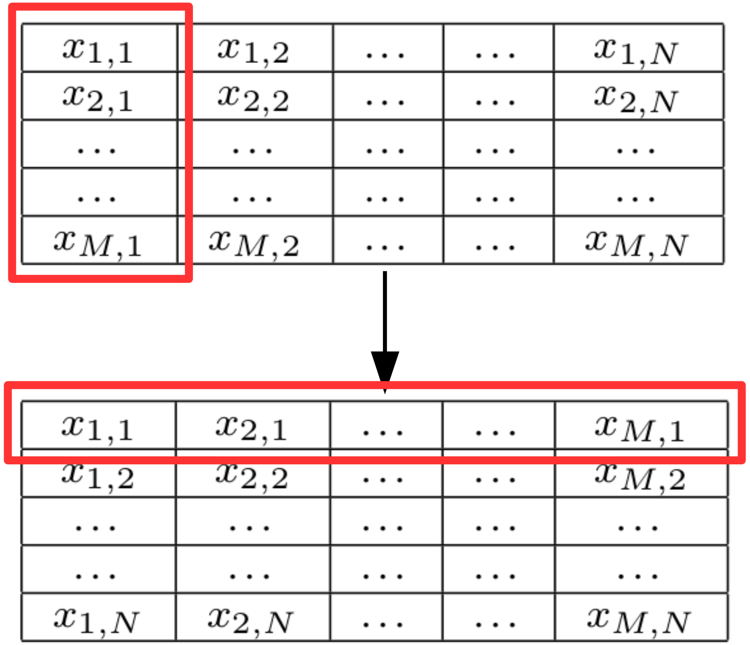


mRMR in MapReduce, alternative layout



$x_{1,1}$	$x_{1,2}$	$\dots$	$\dots$	$x_{1,N}$
$x_{2,1}$	$x_{2,2}$	$\dots$	$\dots$	$x_{2,N}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{M,1}$	$x_{M,2}$	$\dots$	$\dots$	$x_{M,N}$

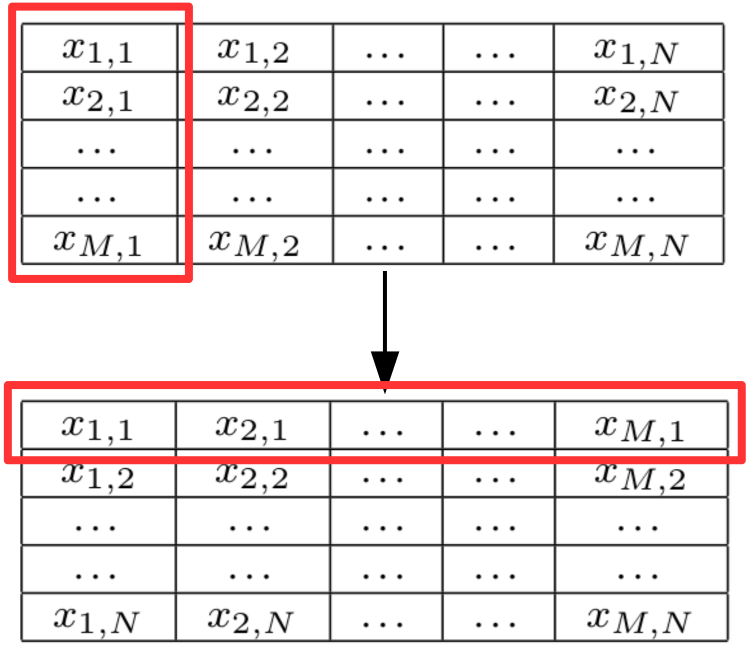
mRMR in MapReduce, alternative layout



$x_{1,1}$	$x_{1,2}$	$\dots$	$\dots$	$x_{1,N}$
$x_{2,1}$	$x_{2,2}$	$\dots$	$\dots$	$x_{2,N}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{M,1}$	$x_{M,2}$	$\dots$	$\dots$	$x_{M,N}$

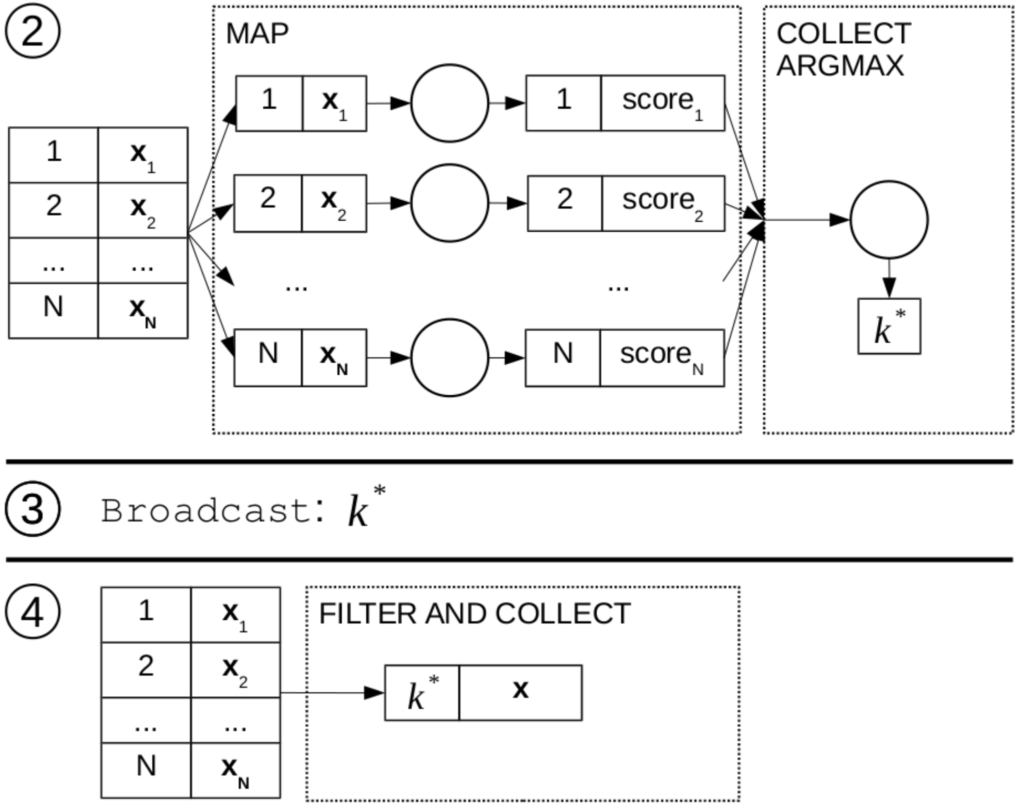
$x_{1,1}$	$x_{2,1}$	$\dots$	$\dots$	$x_{M,1}$
$x_{1,2}$	$x_{2,2}$	$\dots$	$\dots$	$x_{M,2}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{1,N}$	$x_{2,N}$	$\dots$	$\dots$	$x_{M,N}$

mRMR in MapReduce, alternative layout

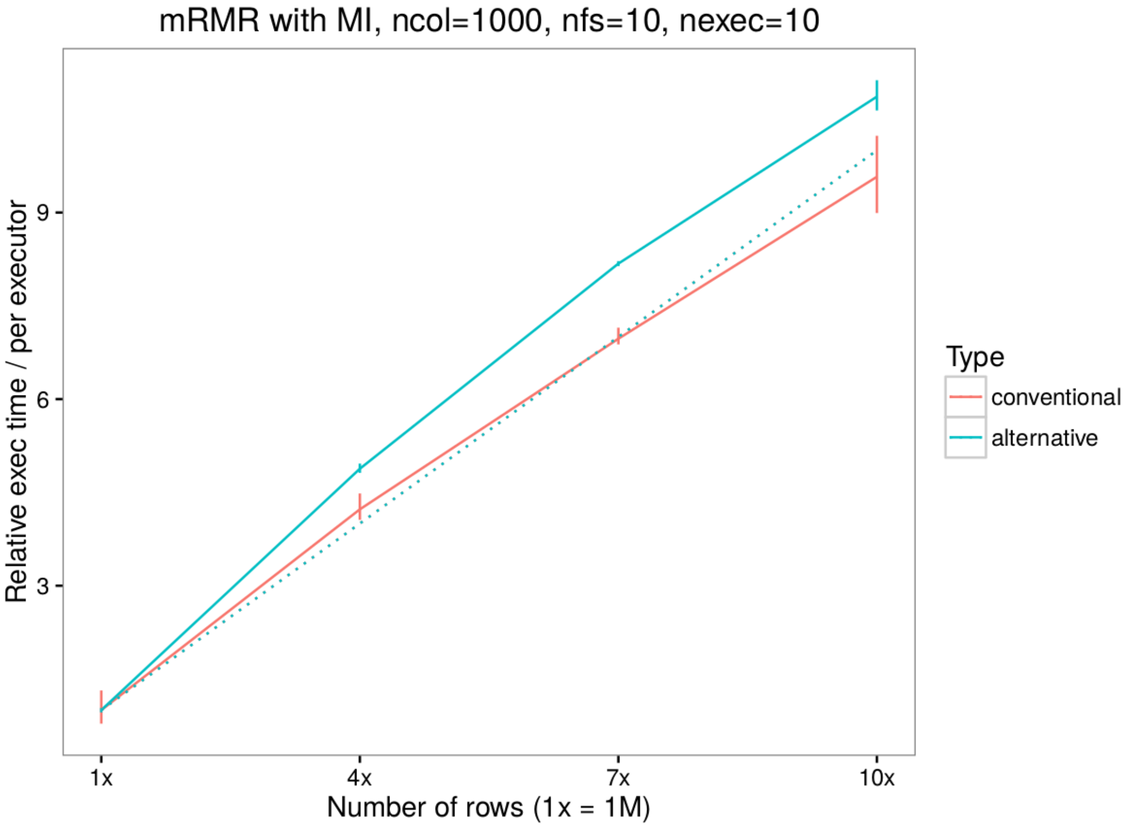


$x_{1,1}$	$x_{1,2}$	$\dots$	$\dots$	$x_{1,N}$
$x_{2,1}$	$x_{2,2}$	$\dots$	$\dots$	$x_{2,N}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{M,1}$	$x_{M,2}$	$\dots$	$\dots$	$x_{M,N}$

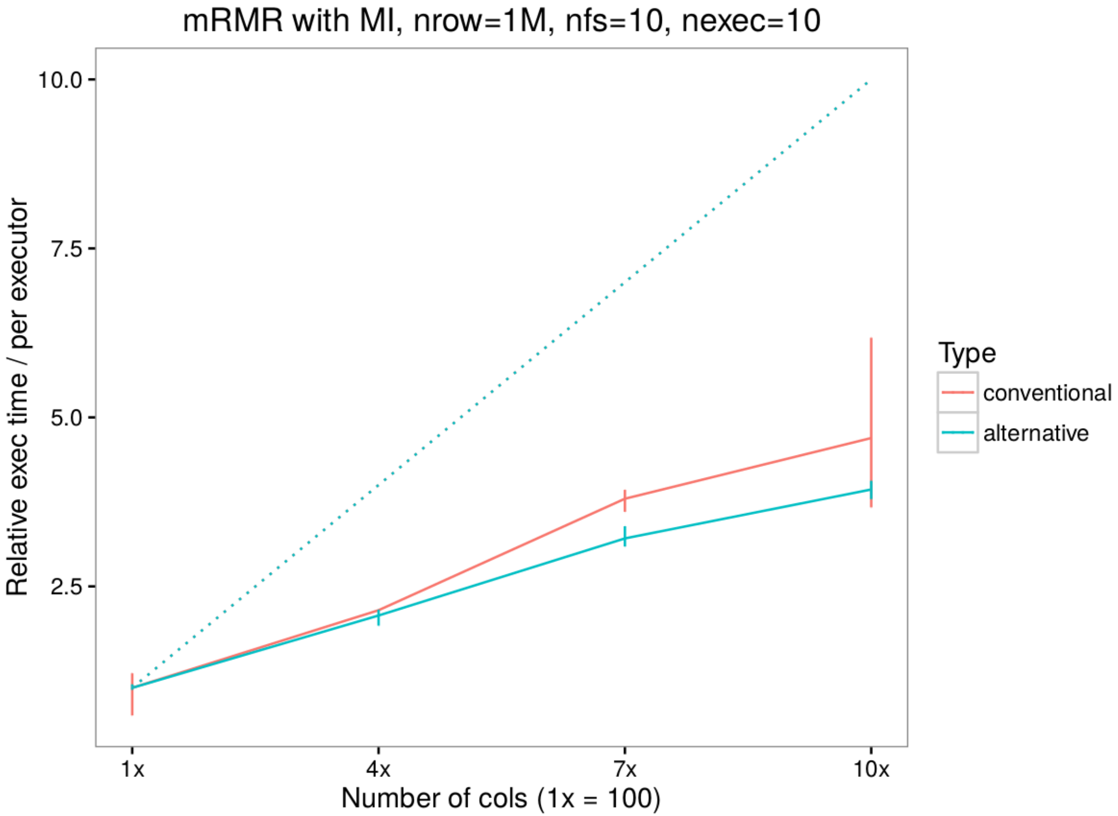
$x_{1,1}$	$x_{2,1}$	$\dots$	$\dots$	$x_{M,1}$
$x_{1,2}$	$x_{2,2}$	$\dots$	$\dots$	$x_{M,2}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{1,N}$	$x_{2,N}$	$\dots$	$\dots$	$x_{M,N}$



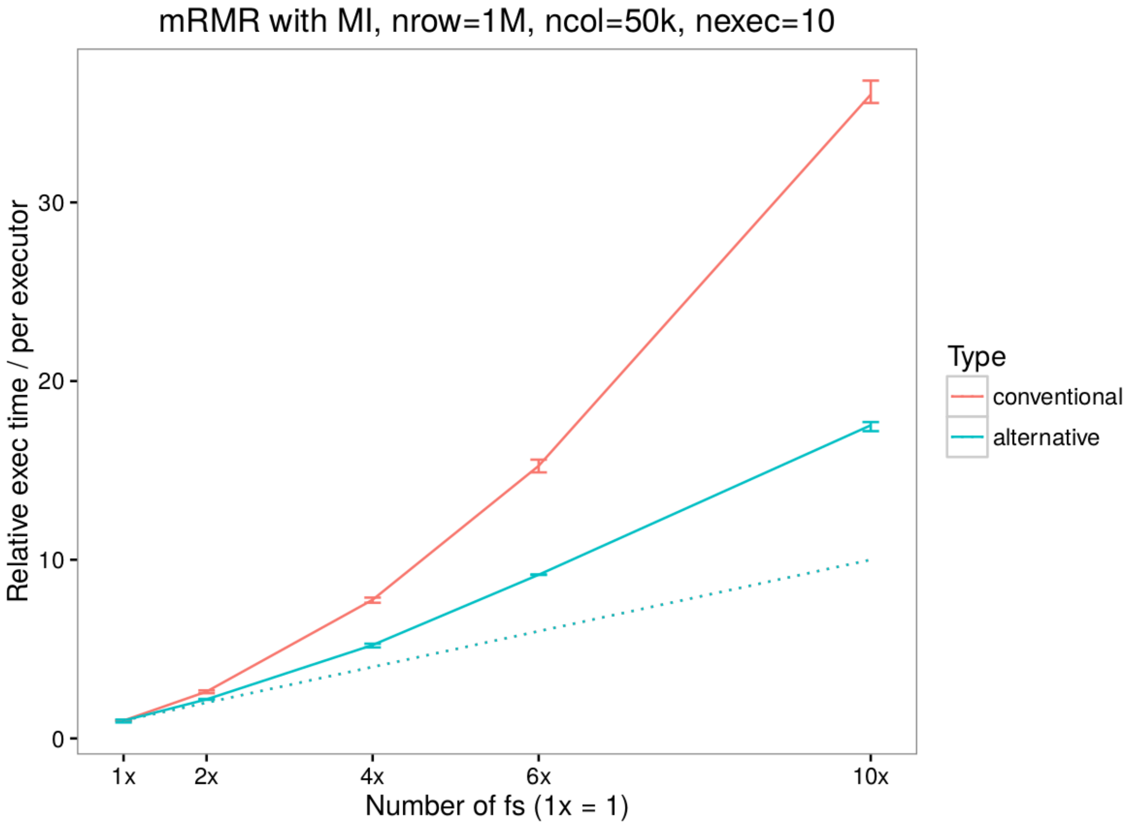
Scalability: #rows



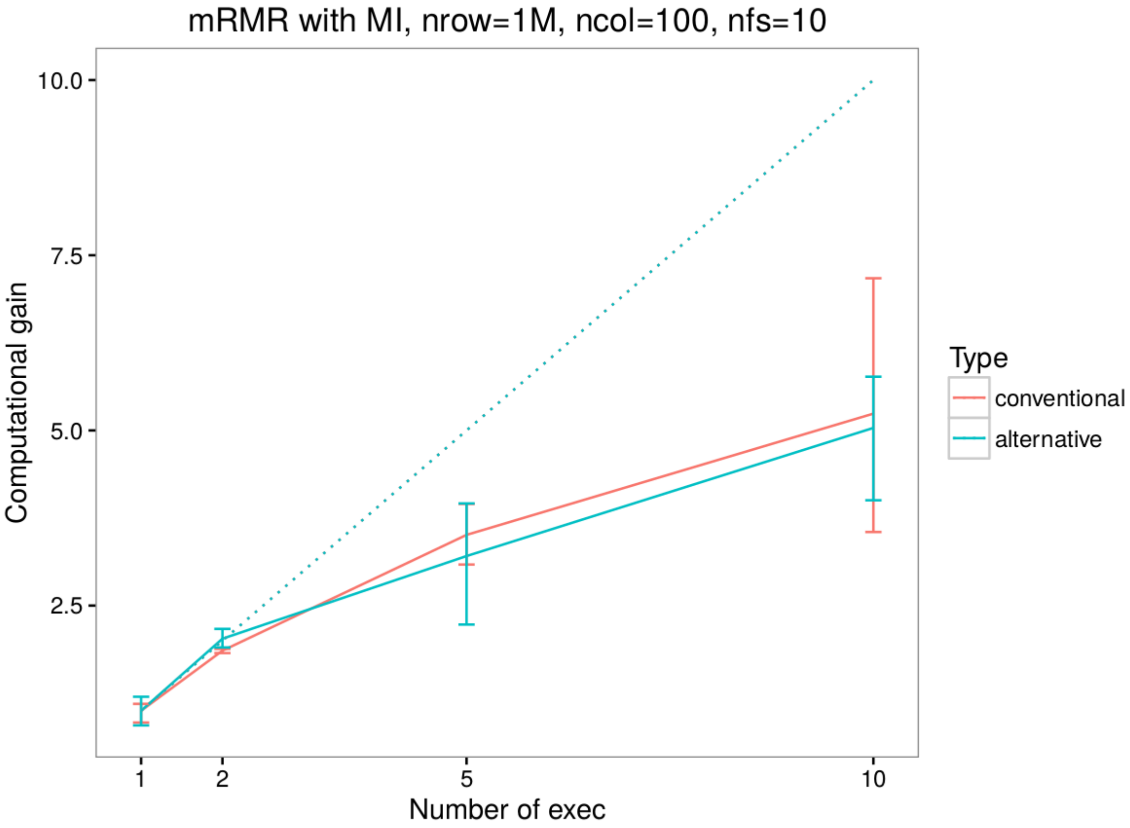
Scalability: #columns



Scalability: #selected features



Scalability: #nodes

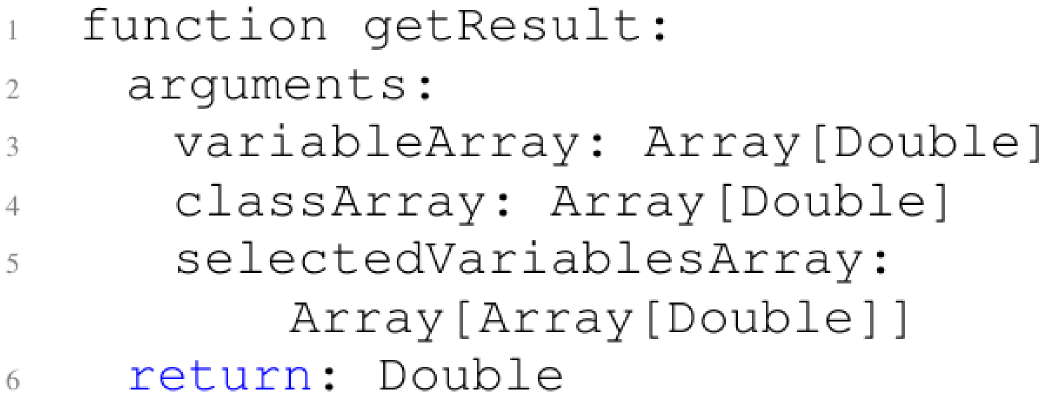


Customization

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]$$

Customization

$$\max_{x_j \in X - S_{m-1}} \underbrace{\left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]}$$



```
function getResult:
  arguments:
    variableArray: Array[Double]
    classArray: Array[Double]
    selectedVariablesArray:
         Array[Array[Double]]
  return: Double
```

Yann-Aël
Le Borgne

Gianluca
Bontempi

ULB Université Libre de Bruxelles

MLG Machine Learning Group

innoviris.brussels
empowering research

github.com/creggian/spark-ifs

github.com/creggian/slides

claudioreggiani.com