

Forward feature selection in high-dimensional dataset using MapReduce

Claudio Reggiani

Yann-Aël Le Borgne

Gianluca Bontempi

BNAIC 2017 – November 8, 2017



Minimal redundancy, maximal relevance (mRMR)

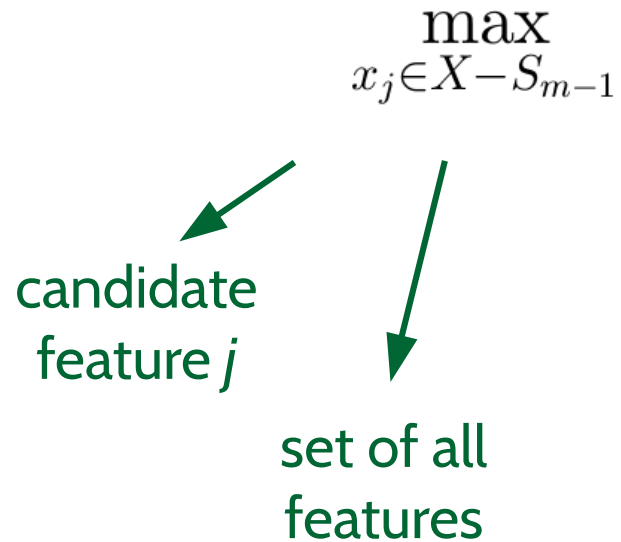
Minimal redundancy, maximal relevance (mRMR)

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

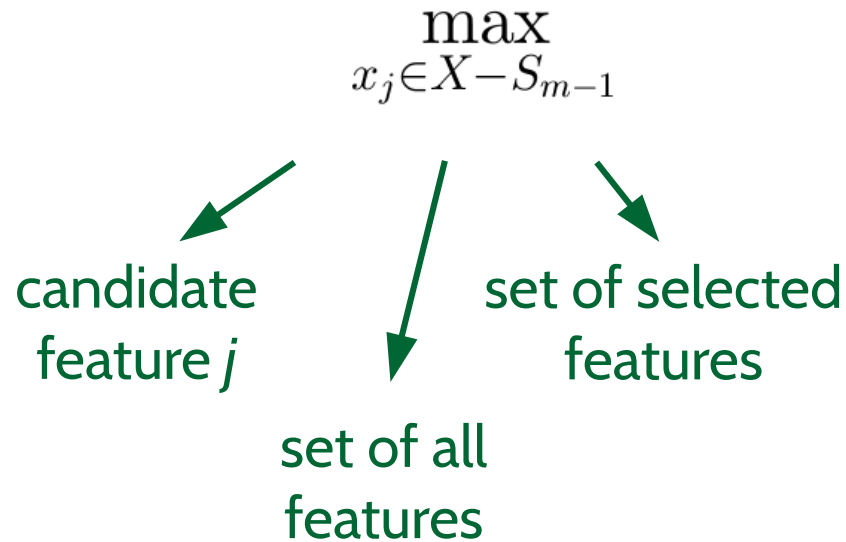
candidate
feature j



Minimal redundancy, maximal relevance (mRMR)



Minimal redundancy, maximal relevance (mRMR)



mRMR

maximal
relevance

minimal redundancy

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

$$\left[\begin{array}{l} \text{maximal} \\ \text{relevance} \end{array} \right]$$

candidate
feature j

set of selected
features

set of all
features

mRMR

maximal
relevance

minimal redundancy

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

candidate
feature j

set of selected
features

set of all
features

mRMR

maximal
relevance

minimal redundancy

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

candidate
feature j

set of selected
features

set of all
features

mRMR

maximal
relevance

minimal redundancy

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

candidate
feature j

set of all
features

set of selected
features

mutual
information
(discrete values)

mRMR

maximal
relevance

minimal redundancy

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

candidate
feature j

set of all
features

set of selected
features

mutual
information
(discrete values)

mRMR

maximal
relevance

minimal redundancy

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

candidate
feature j

set of all
features

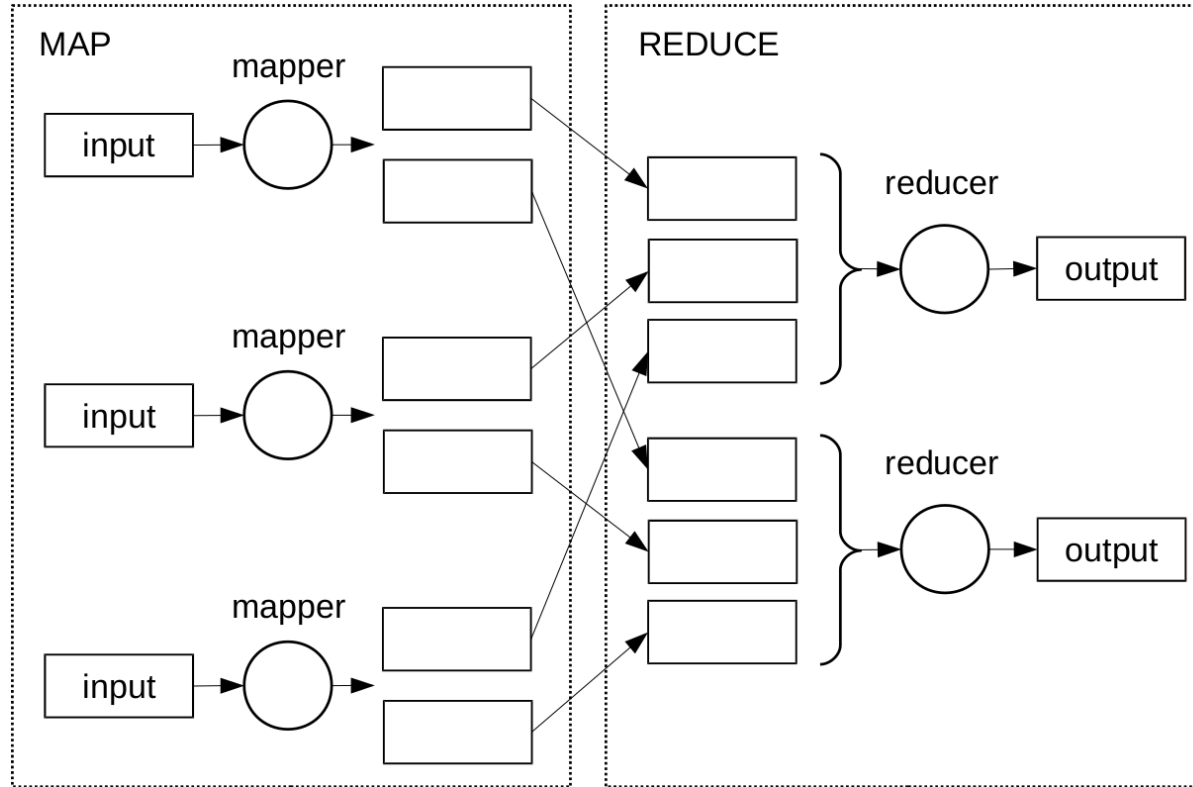
set of selected
features

cardinality
of selected
features set

mutual
information
(discrete values)

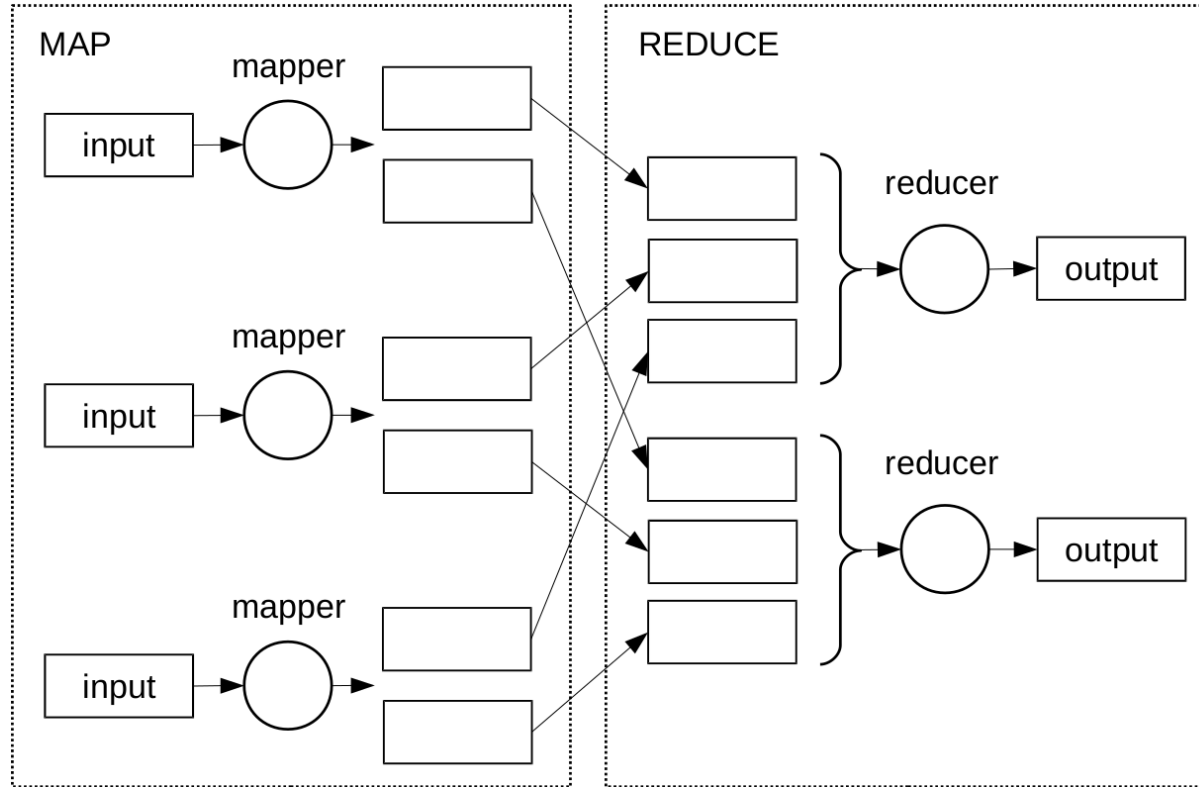
How do we cope with high-dimensional datasets?

Brief introduction to MapReduce

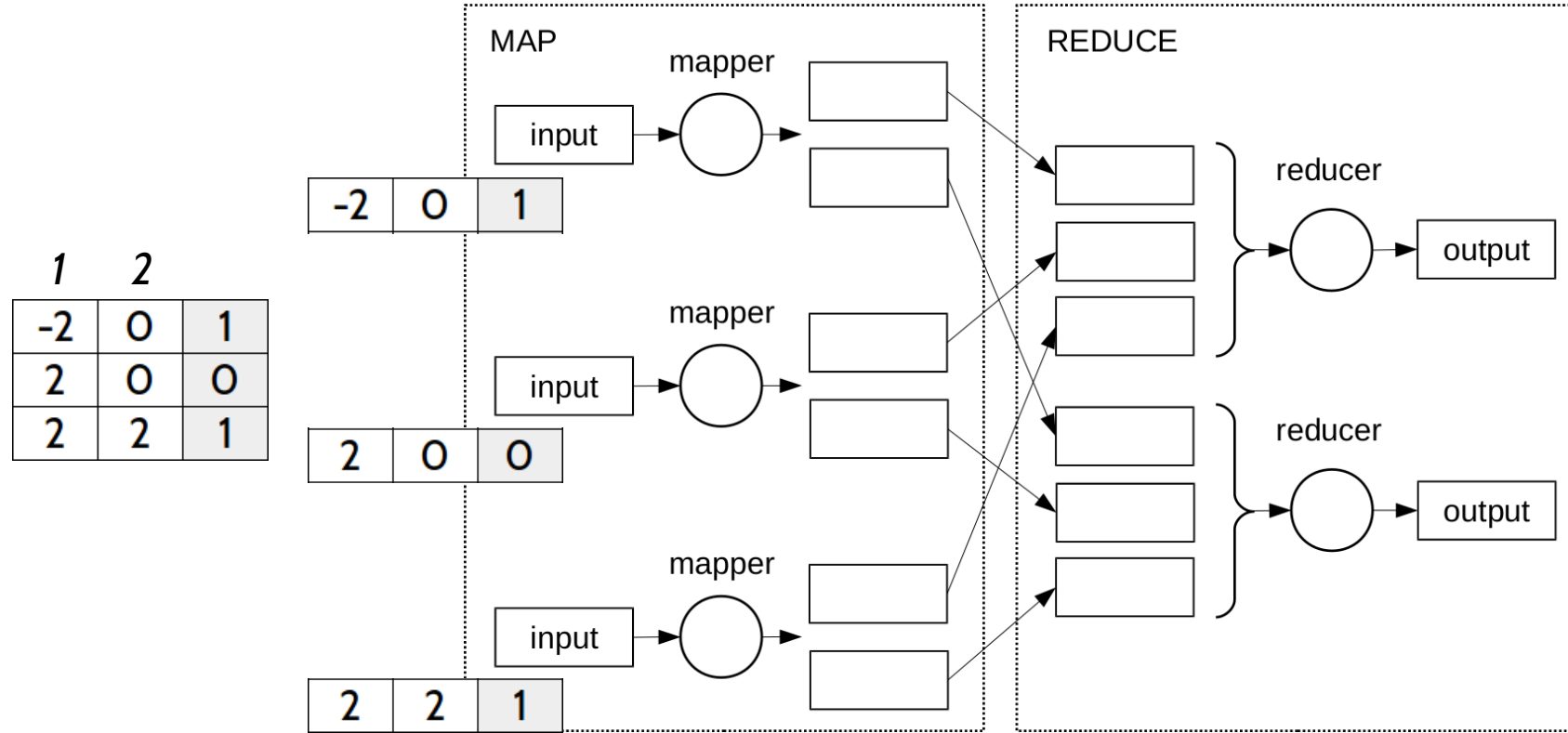


Brief introduction to MapReduce

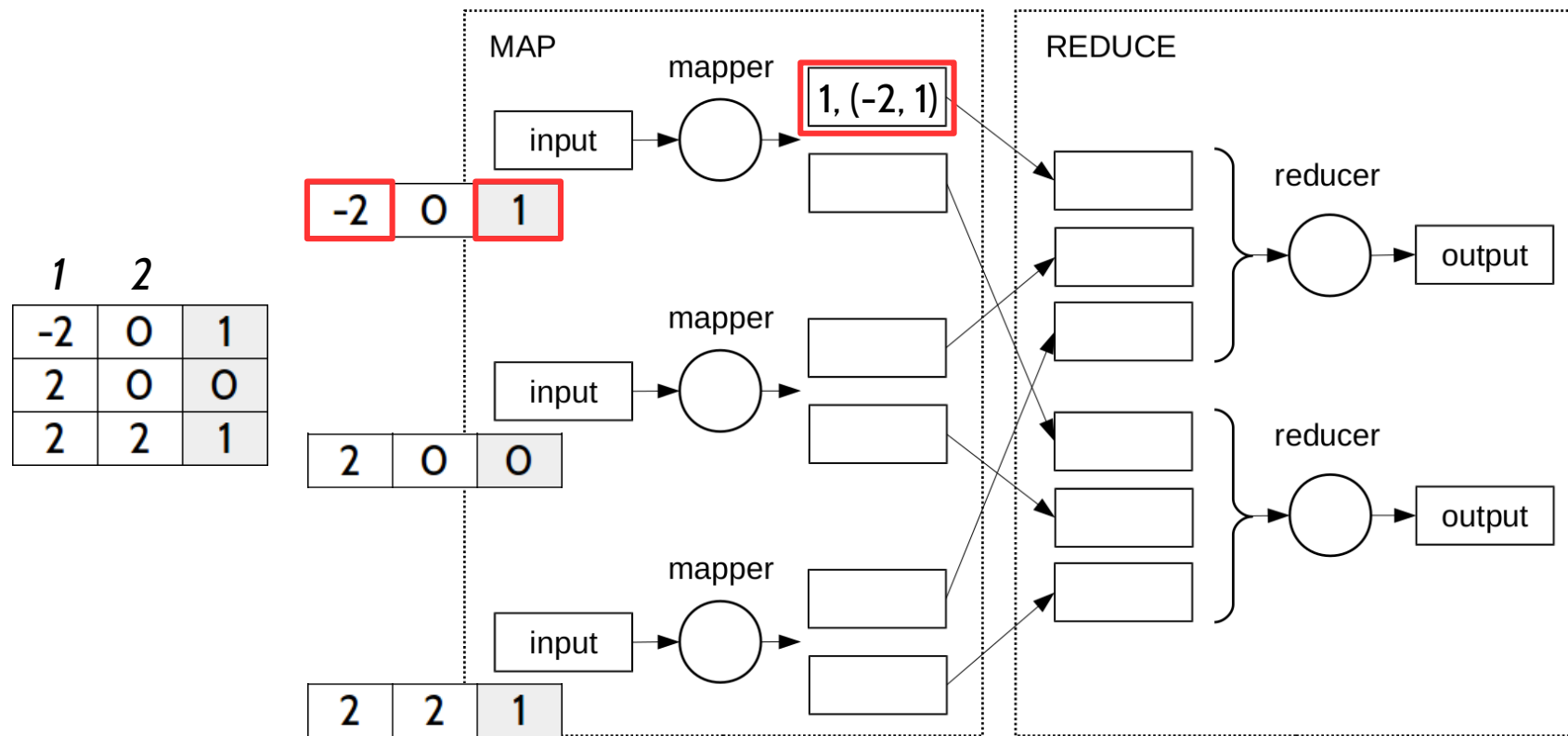
1	2	
-2	0	1
2	0	0
2	2	1



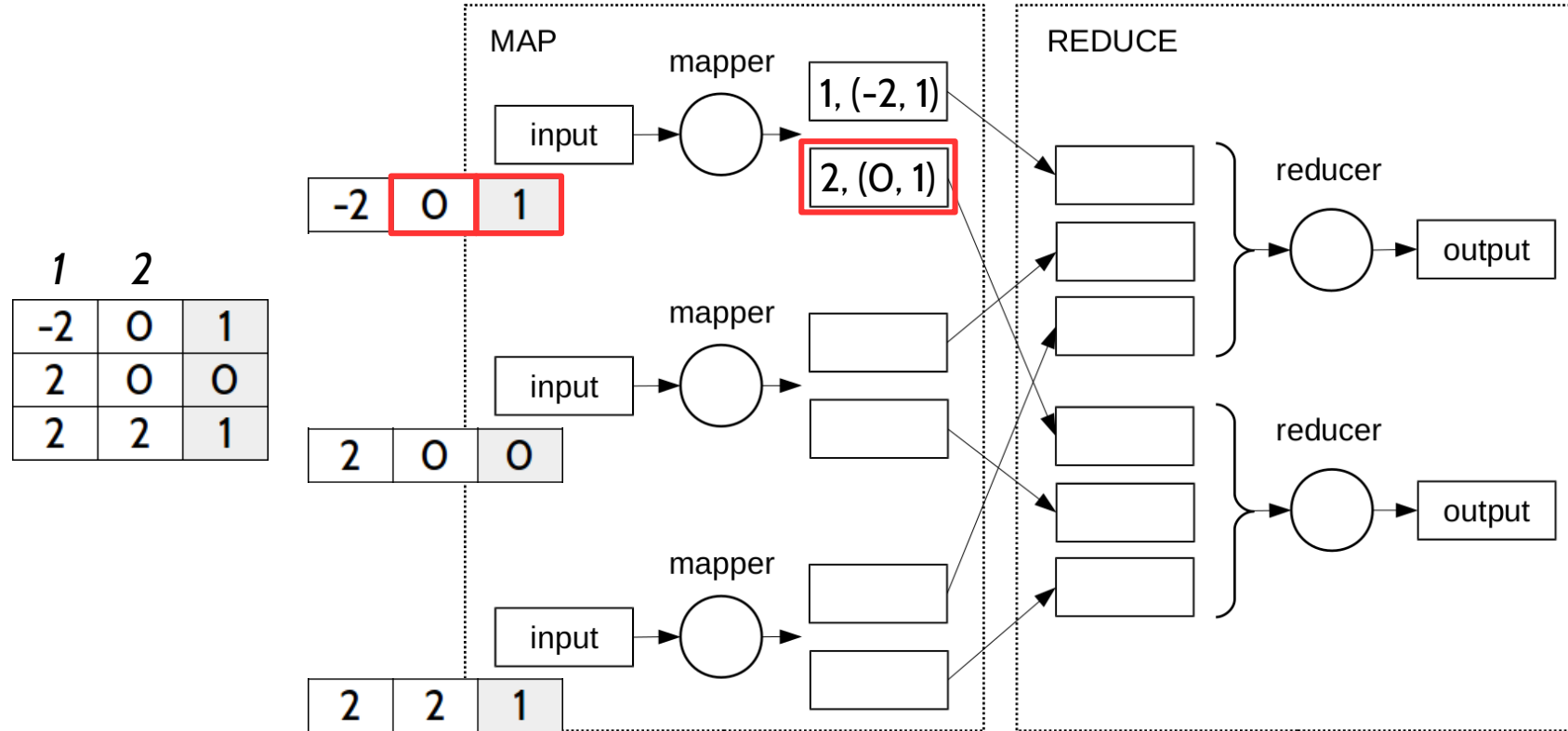
Brief introduction to MapReduce



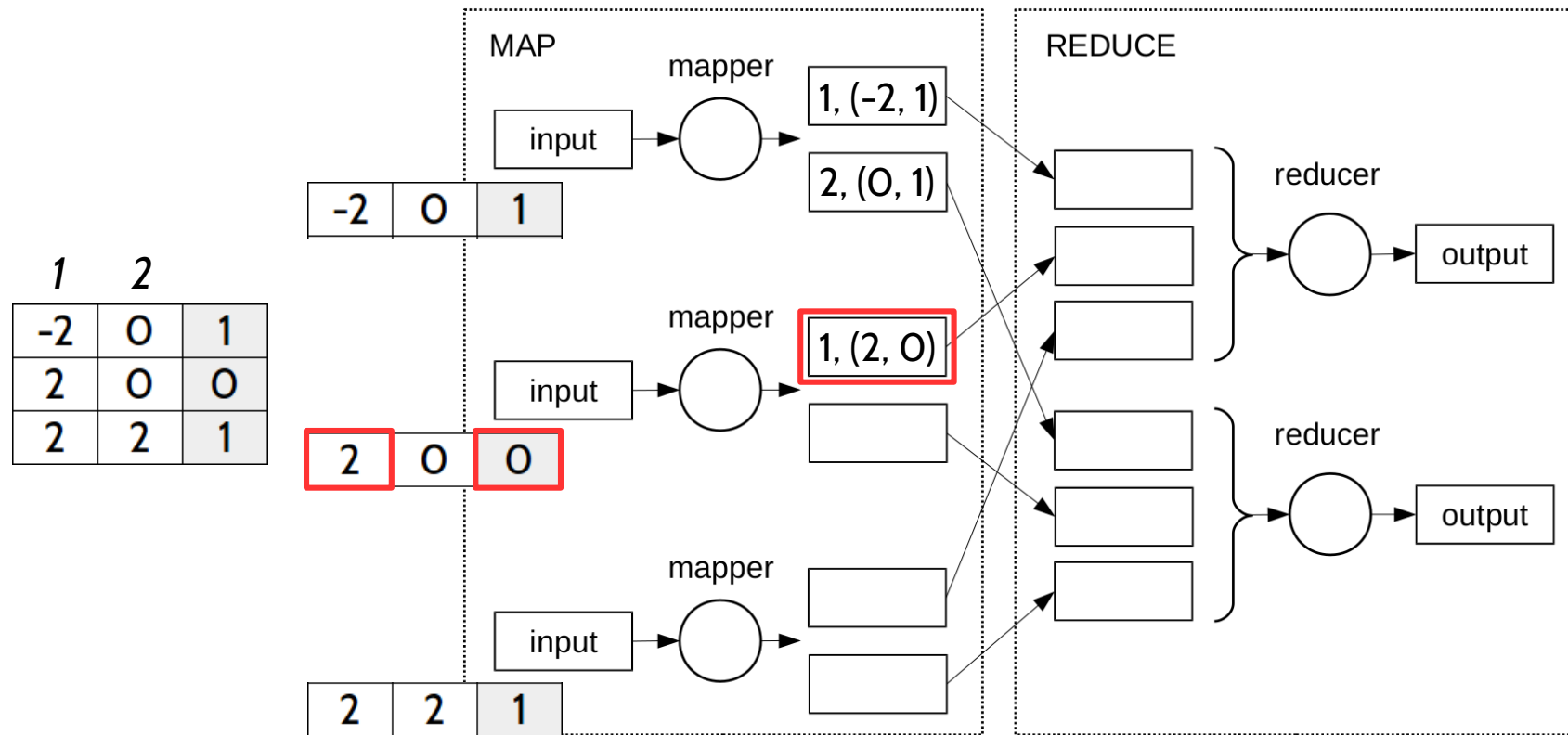
Brief introduction to MapReduce



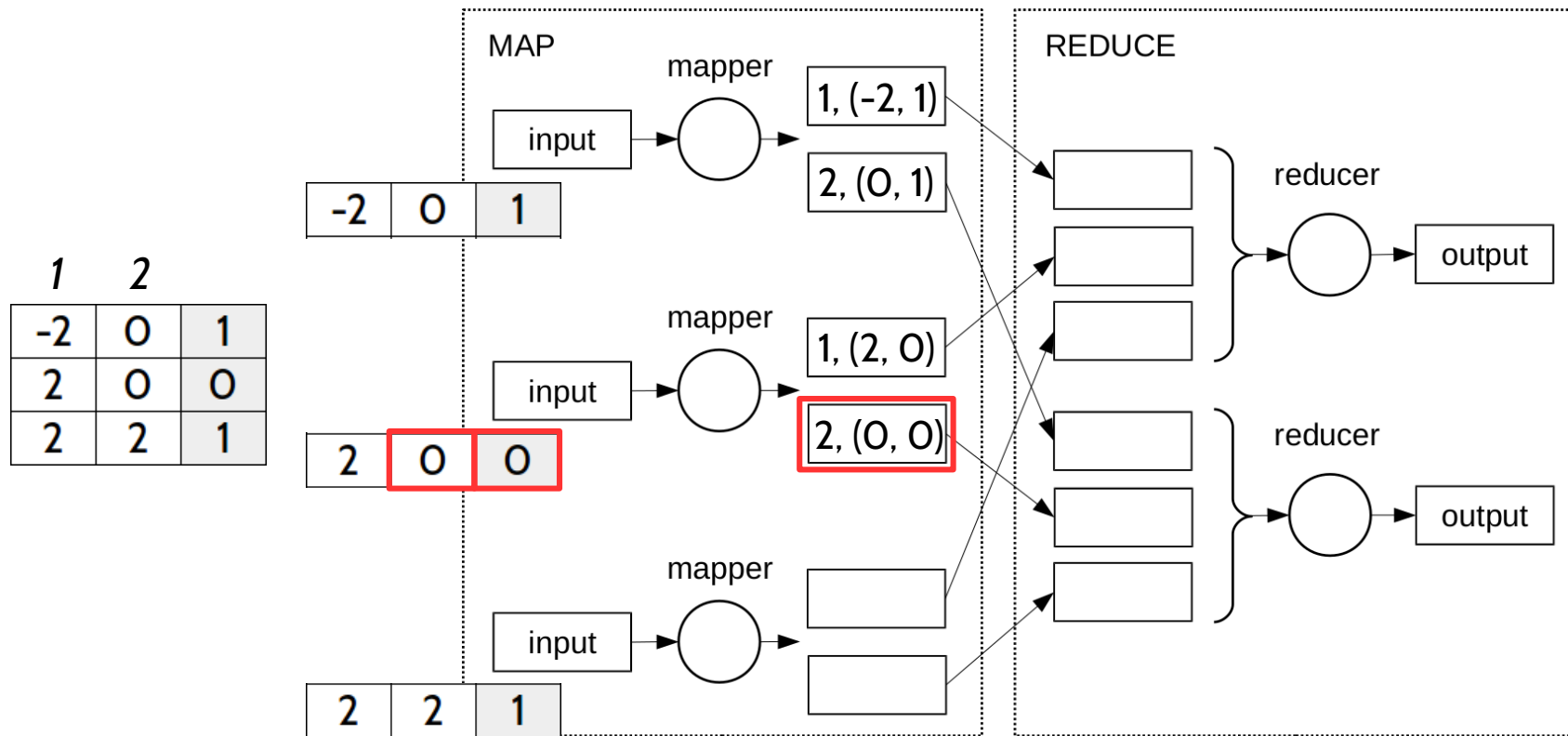
Brief introduction to MapReduce



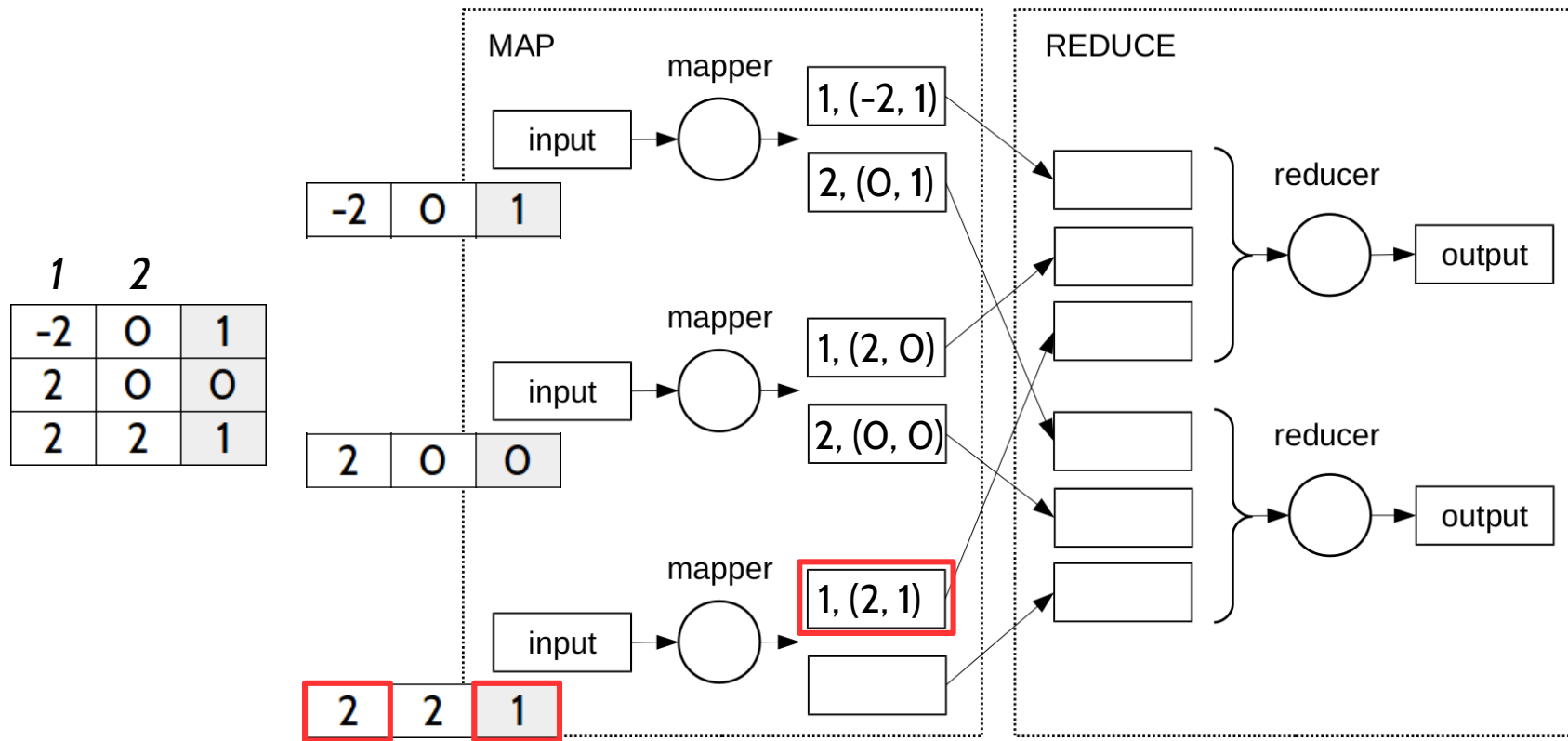
Brief introduction to MapReduce



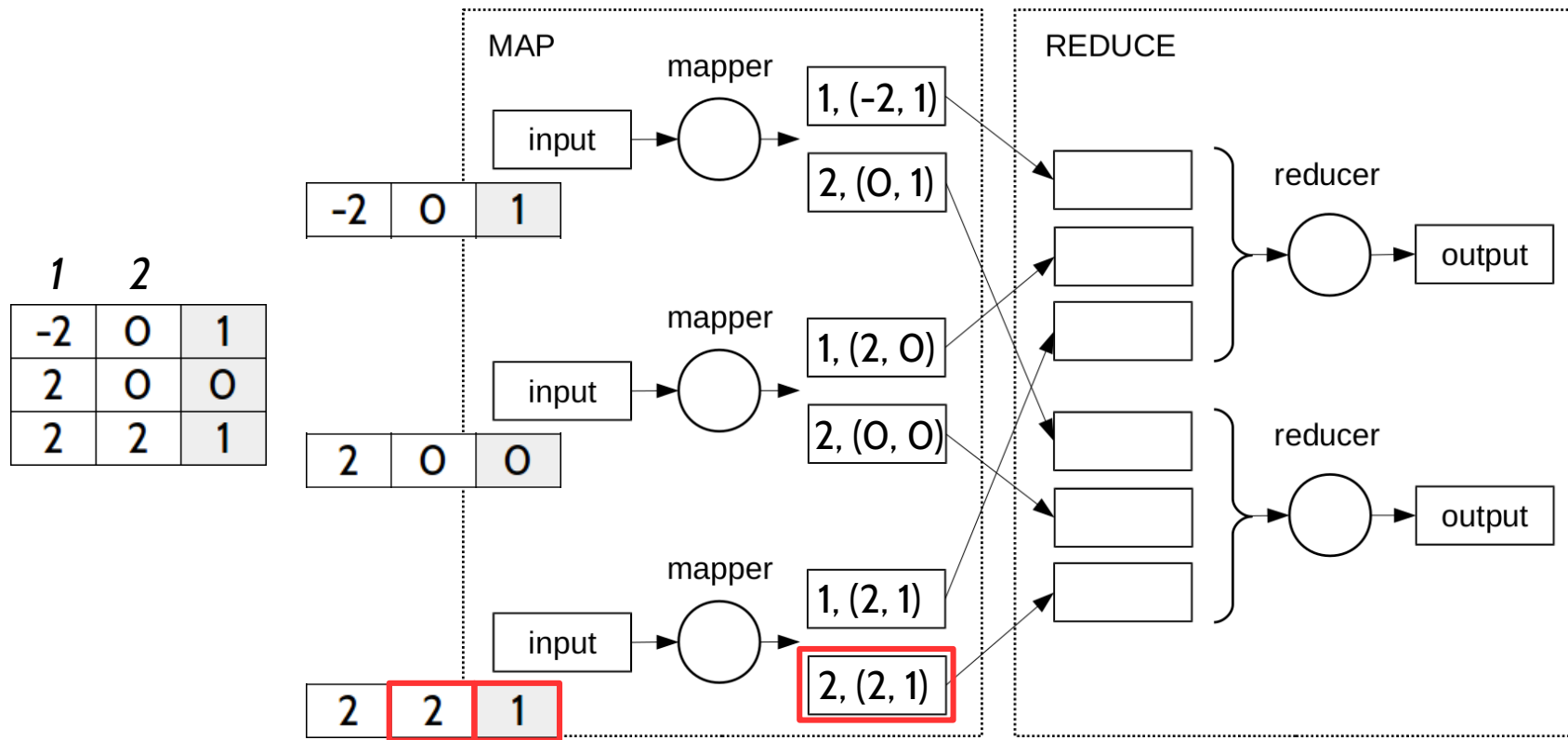
Brief introduction to MapReduce



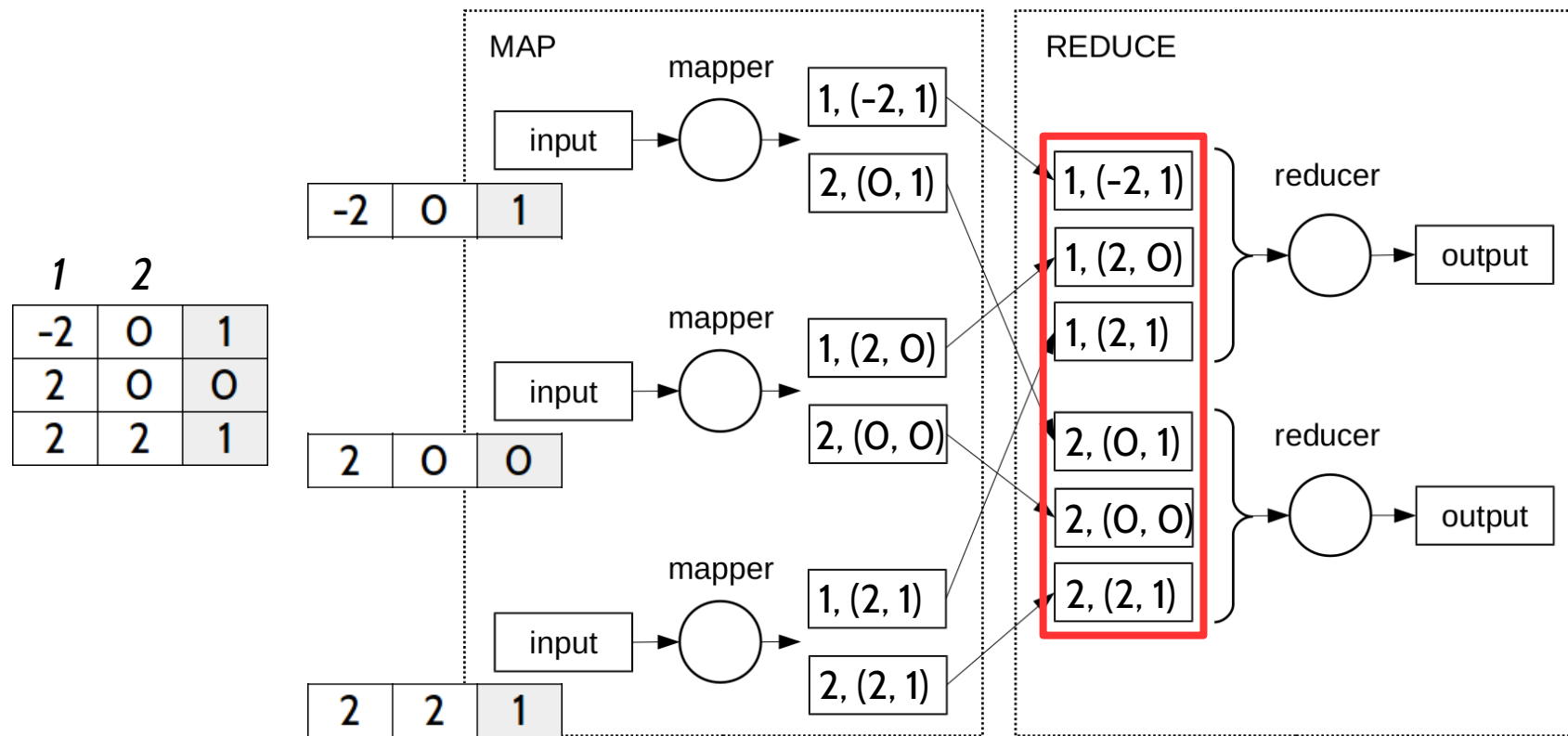
Brief introduction to MapReduce



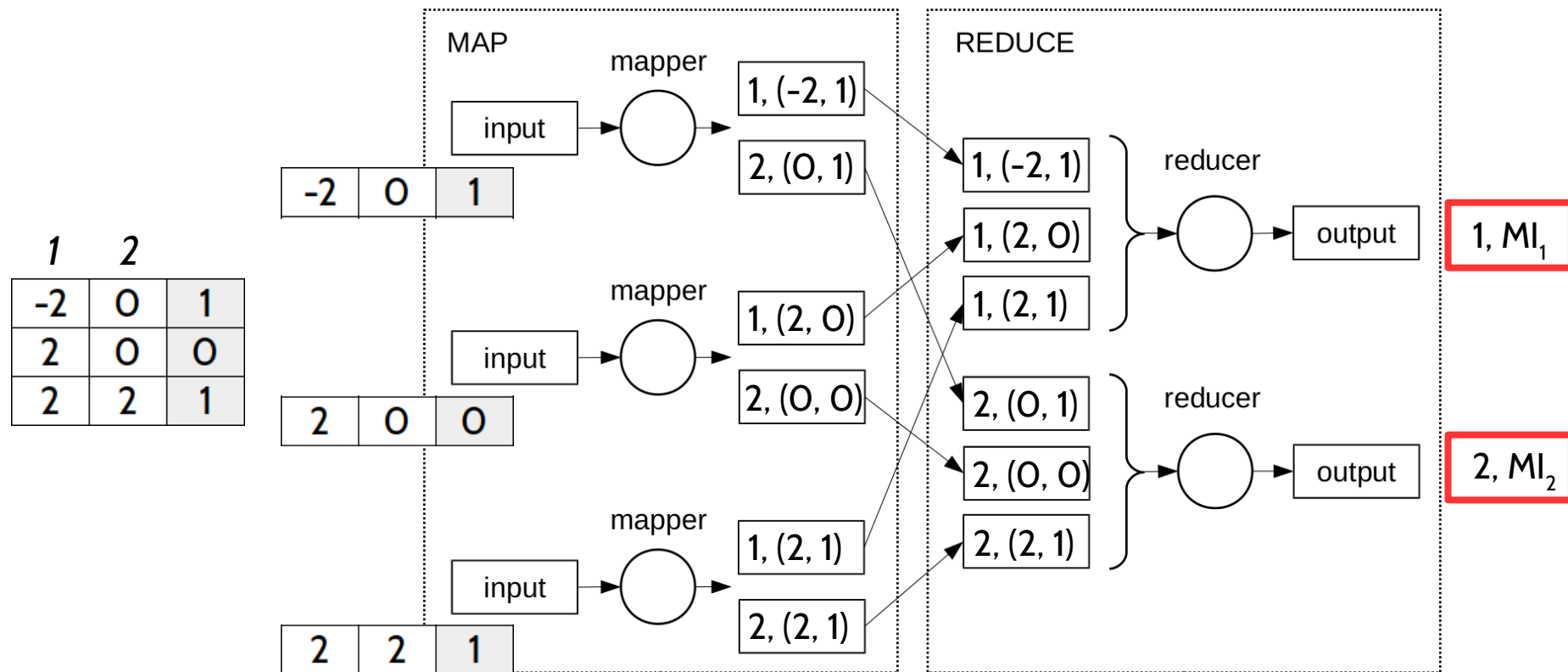
Brief introduction to MapReduce



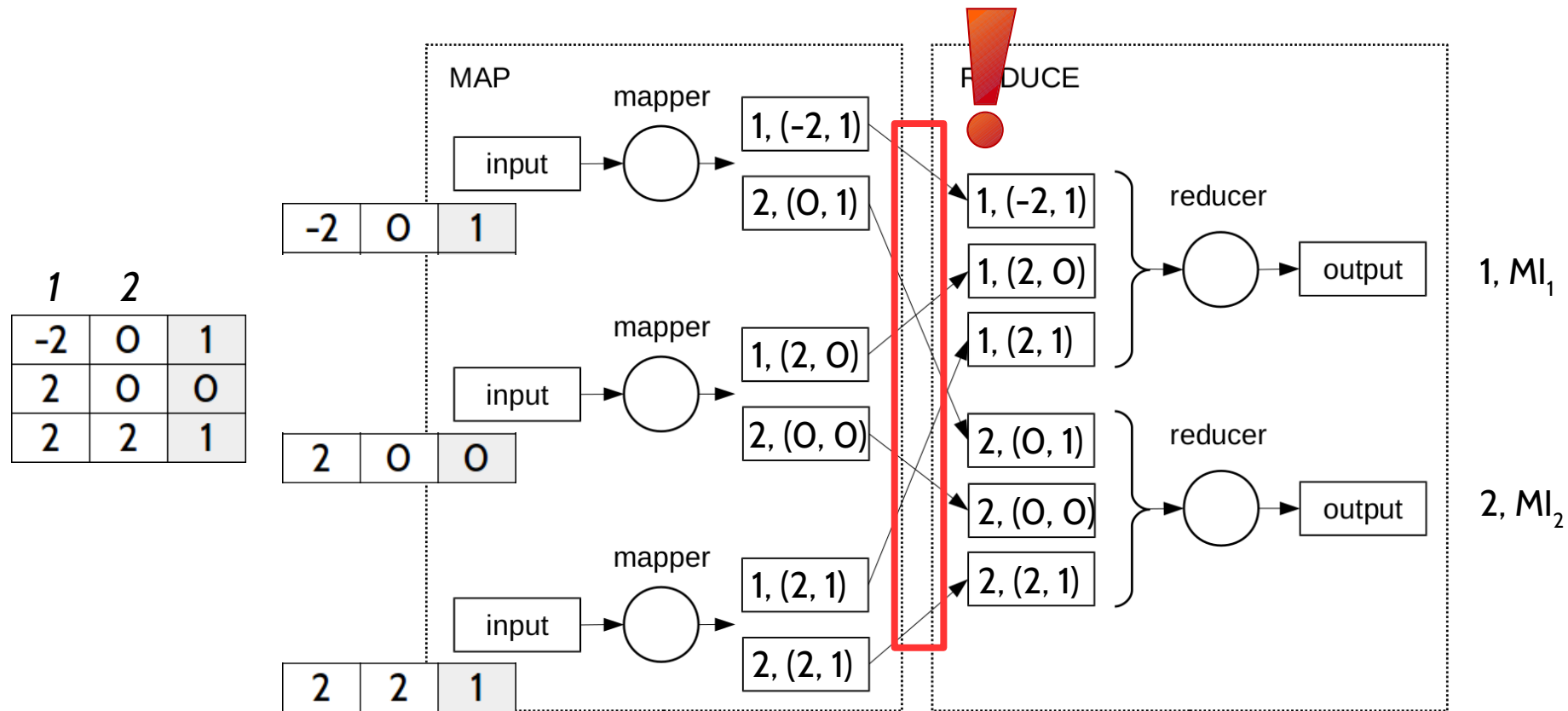
Brief introduction to MapReduce



Brief introduction to MapReduce



Brief introduction to MapReduce



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

...

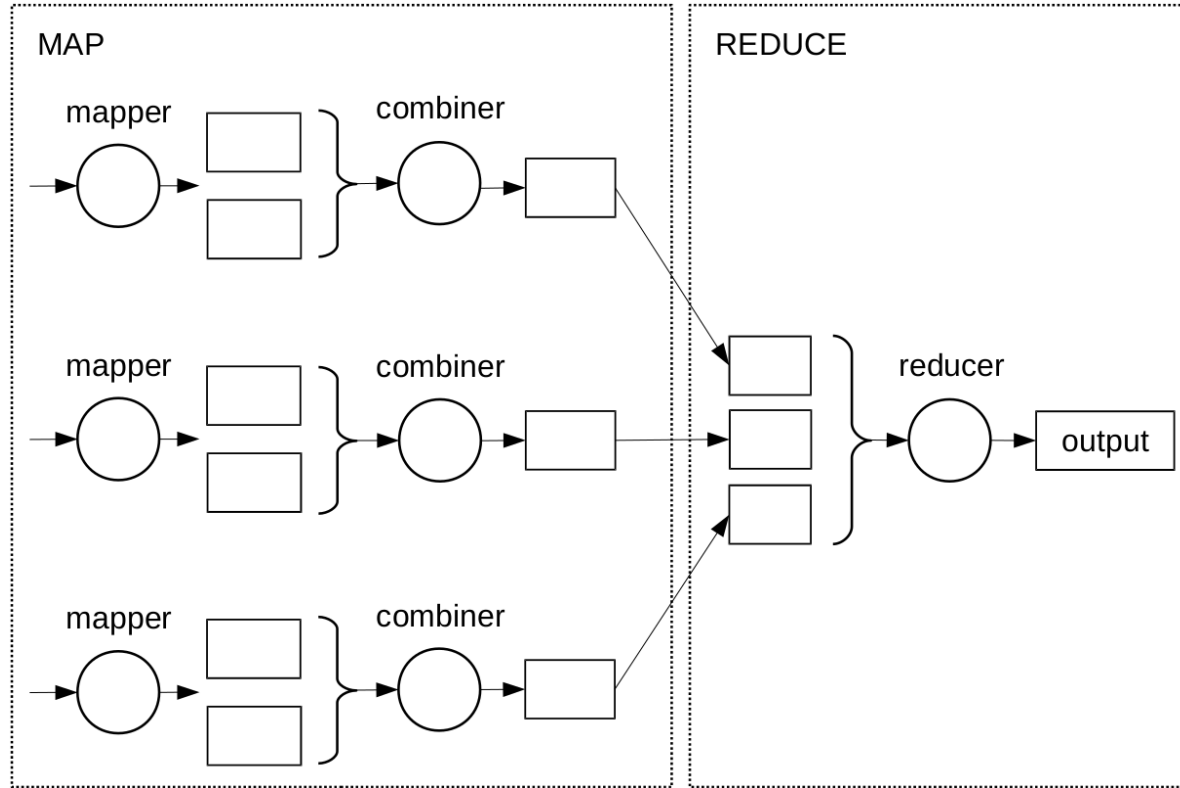
-2	0	1
2	0	0
2	2	1

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

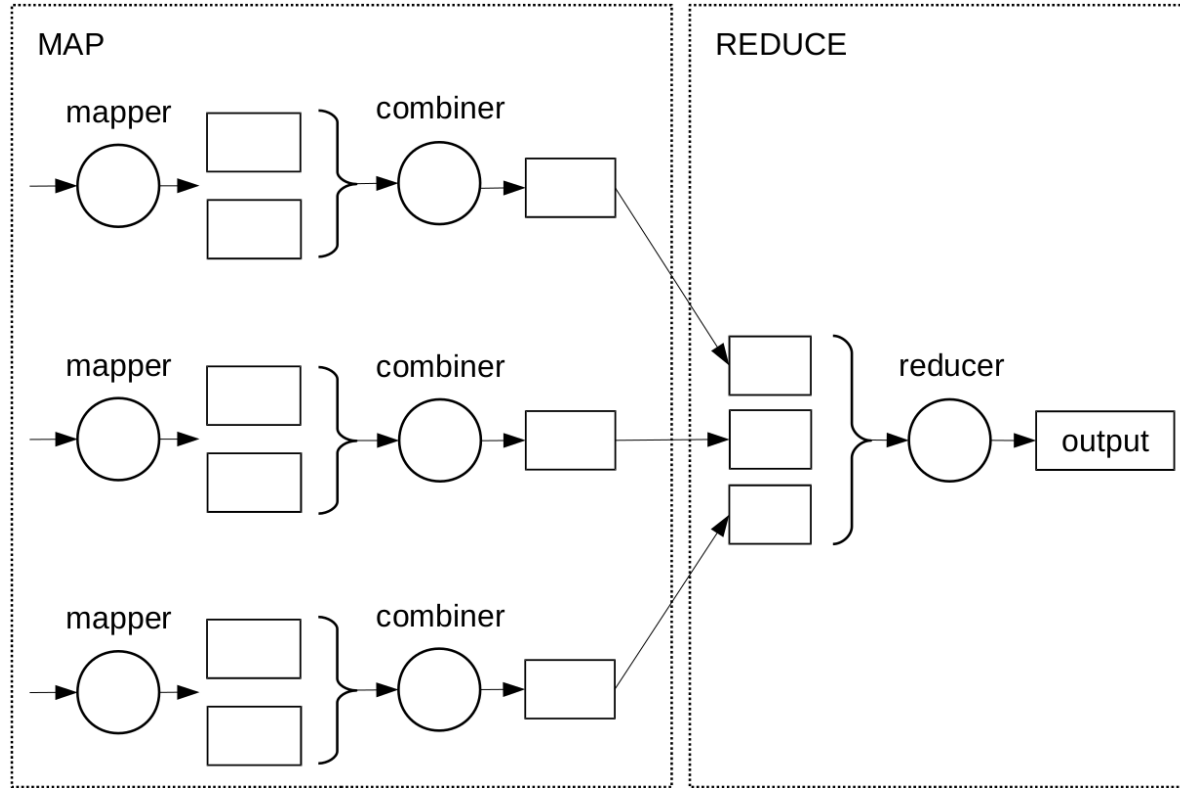
...

-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



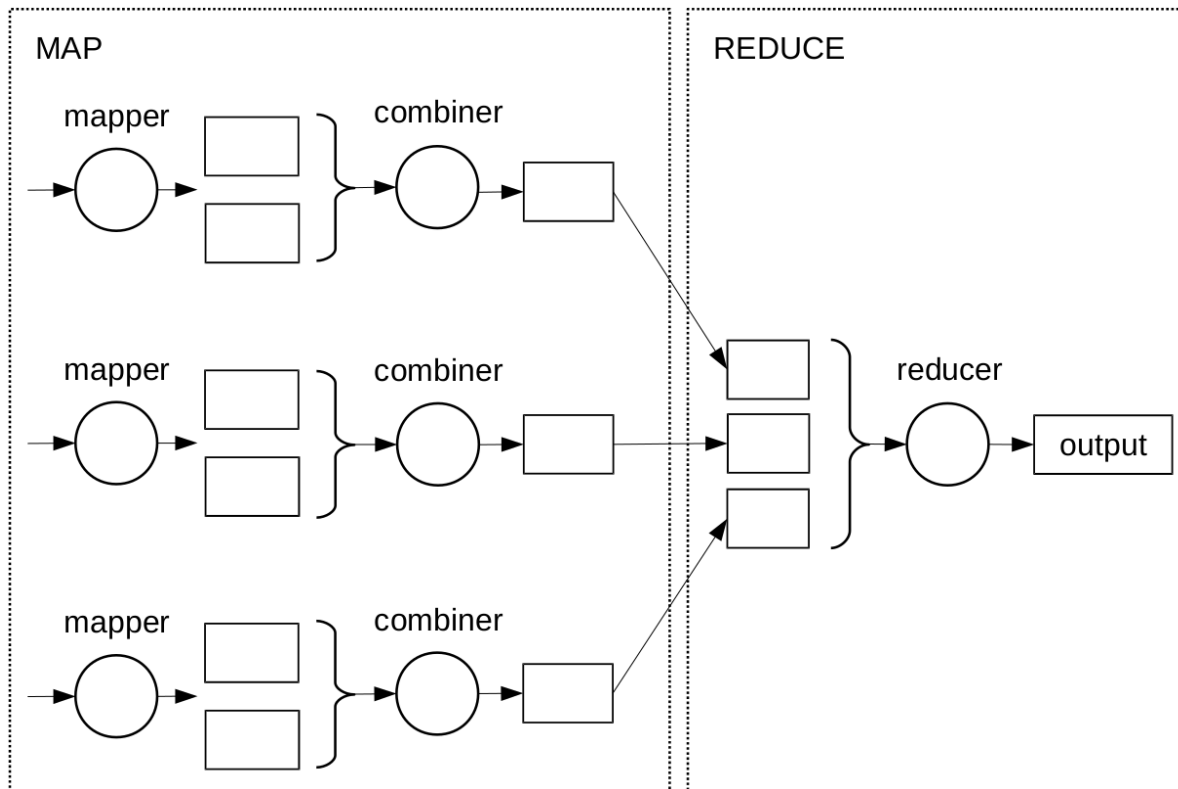
Brief introduction to MapReduce



-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)

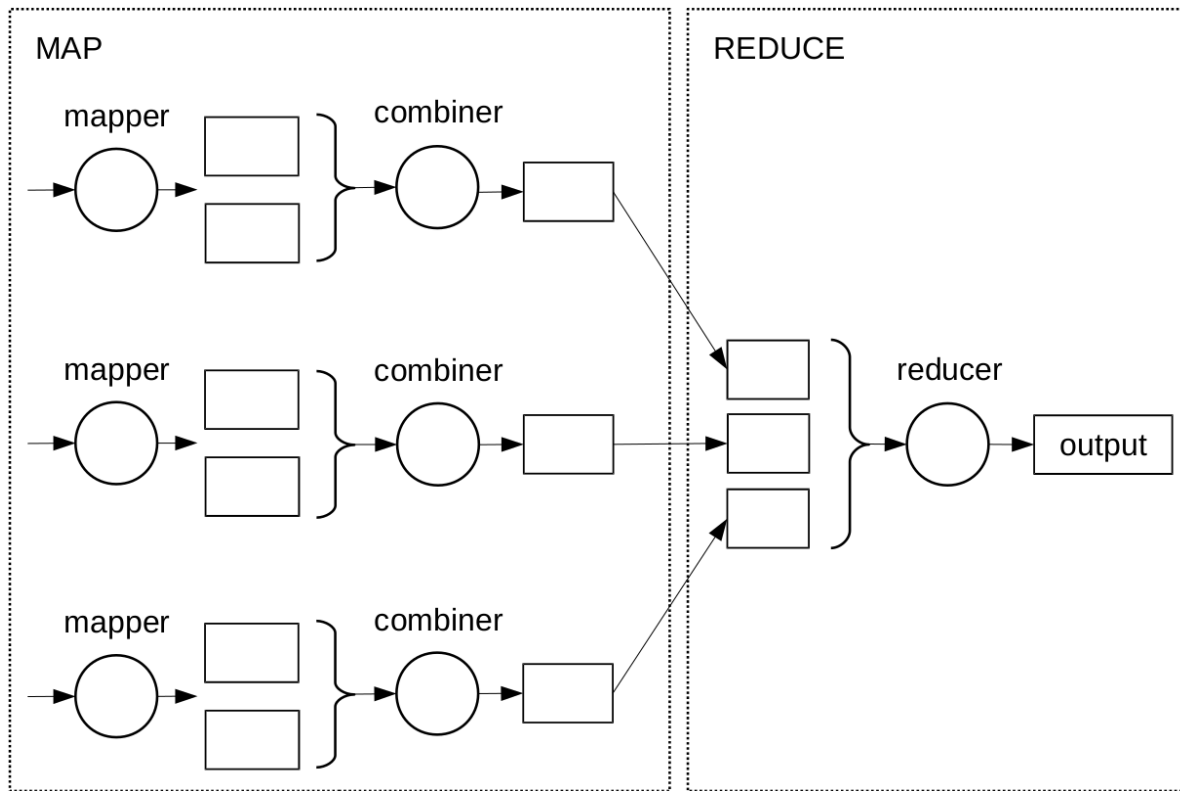
1, (-2, 1)

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

x_1	c
2	0
0	0
0	0
-2	1



d_c	0	d_v		
	1	-2	0	2

categories
of the class

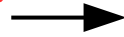
categories
of the feature

Data structure with cumulative property

MAP

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

x_1	c
2	0
0	0
0	0
-2	1



Data structure with cumulative property

MAP

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

Data structure with cumulative property

MAP

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



		d_v		
d_c	0	-2	0	2
	1	0	0	1

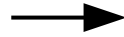
		d_v		
d_c	0	0	1	0
	1	0	0	0

		d_v		
d_c	0	0	1	0
	1	0	0	0

Data structure with cumulative property

MAP

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



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

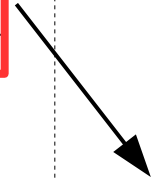
		d_v		
		-2	0	2
d_c	0	0	1	0
	1	0	0	0

		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

$$1 \quad i_c^1 = \{1, \dots, N\}$$

mRMR in MapReduce

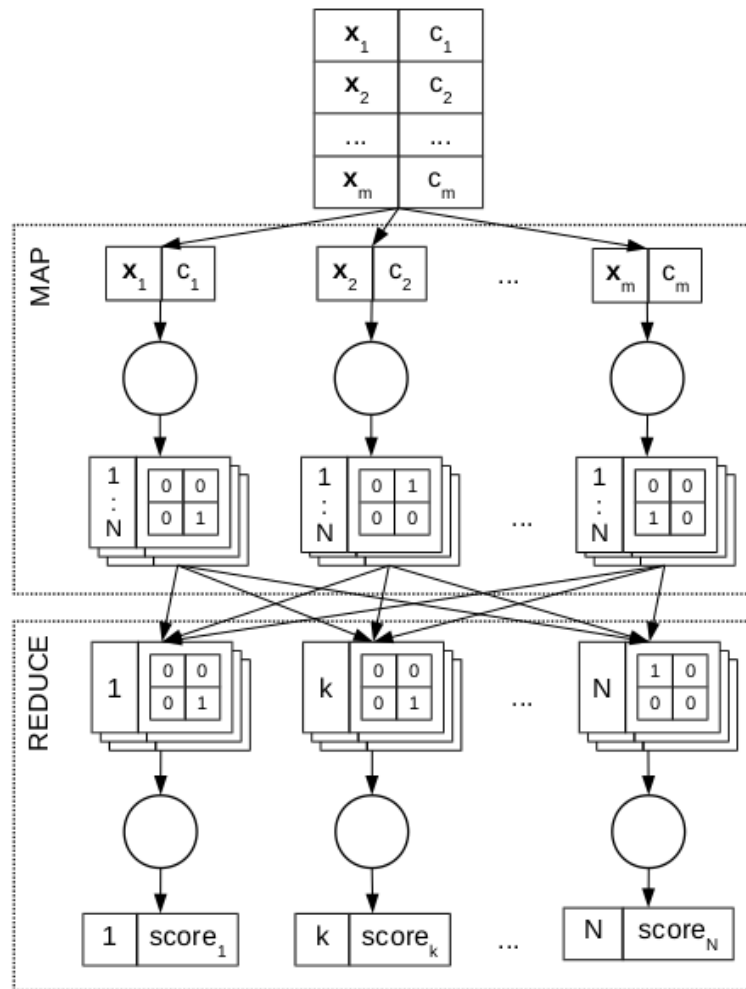
$$\begin{array}{ll} 1 & i_c^1 = \{1, \dots, N\} \\ 2 & i_s^1 = \emptyset \end{array}$$

mRMR in MapReduce

```
1   $i_c^1 = \{1, \dots, N\}$   
2   $i_s^1 = \emptyset$   
3  for  $l = 1 \rightarrow L$ 
```

mRMR in MapReduce

```
1  $i_c^1 = \{1, \dots, N\}$   
2  $i_s^1 = \emptyset$   
3 for  $l = 1 \rightarrow L$   
4   broadcast  $i_{class}, i_c^l, i_s^l, d_v, d_c$   
5   scores  $\leftarrow$  mapreduce(RDD, mapper, reducer)
```

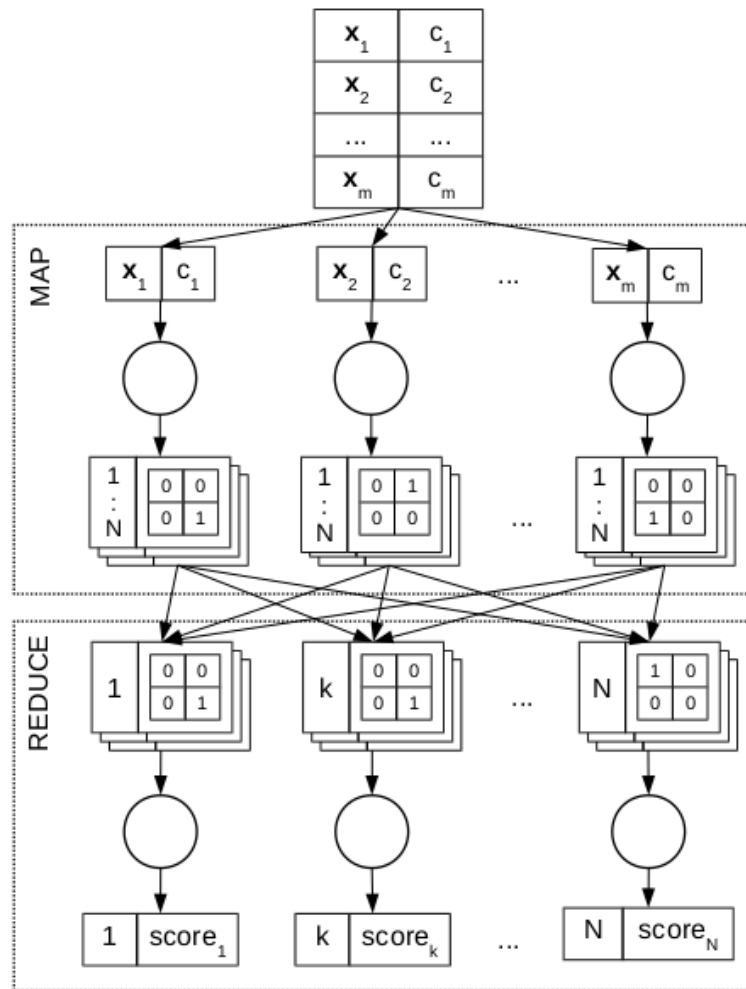


mRMR in MapReduce

```

1   $i_c^1 = \{1, \dots, N\}$ 
2   $i_s^1 = \emptyset$ 
3  for  $l = 1 \rightarrow L$ 
4    broadcast  $i_{class}, i_c^l, i_s^l, d_v, d_c$ 
5    scores  $\leftarrow$  mapreduce(RDD, mapper, reducer)
6     $k^* \leftarrow \text{collectArgmax}(\text{scores})$ 

```

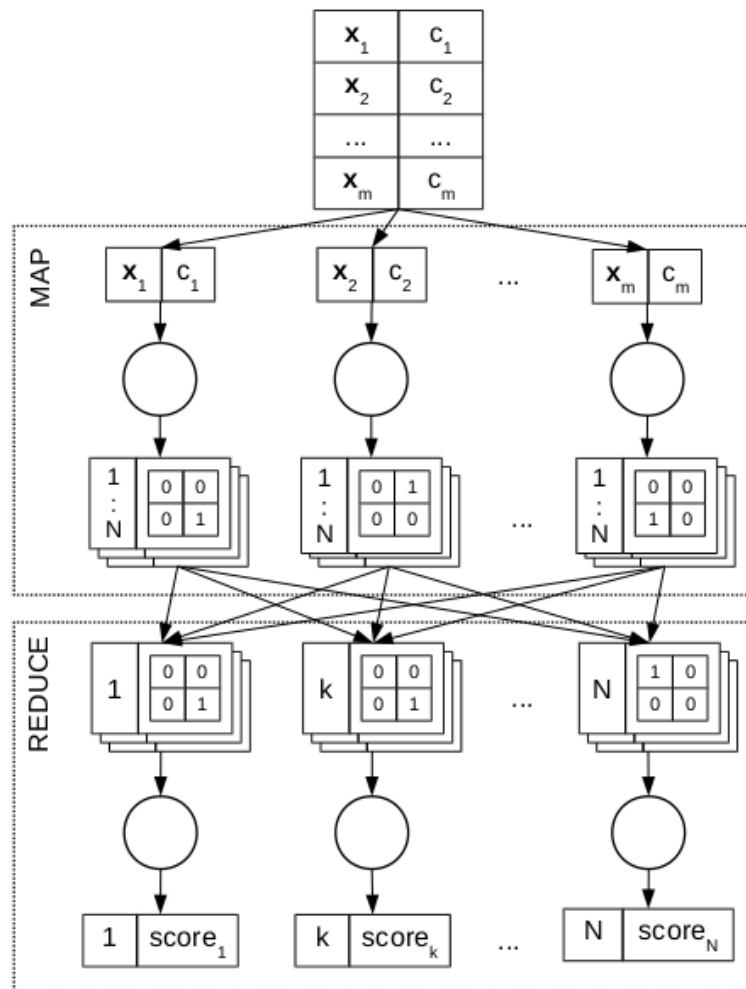


mRMR in MapReduce

```

1   $i_c^1 = \{1, \dots, N\}$ 
2   $i_s^1 = \emptyset$ 
3  for  $l = 1 \rightarrow L$ 
4    broadcast  $i_{class}, i_c^l, i_s^l, d_v, d_c$ 
5    scores  $\leftarrow$  mapreduce(RDD, mapper, reducer)
6     $k^* \leftarrow \text{collectArgmax}(\text{scores})$ 
7     $i_c^{l+1} \leftarrow i_c^l \setminus k^*$ 
8     $i_s^{l+1} \leftarrow i_s^l \cup k^*$ 

```

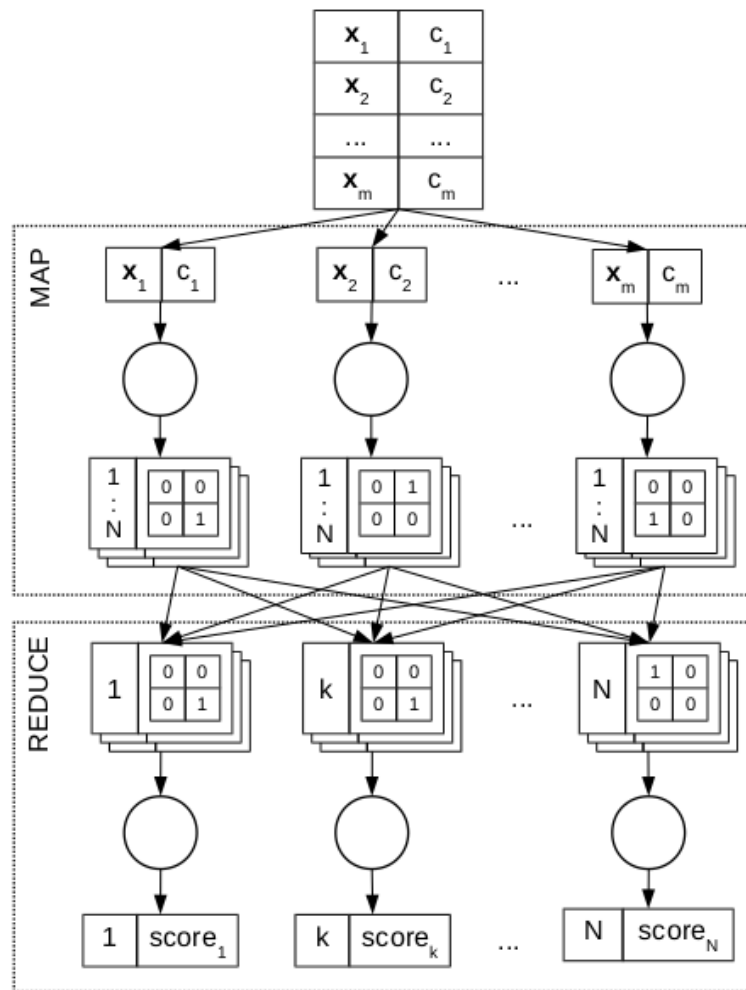


mRMR in MapReduce

```

1   $i_c^1 = \{1, \dots, N\}$ 
2   $i_s^1 = \emptyset$ 
3  for  $l = 1 \rightarrow L$ 
4    broadcast  $i_{class}, i_c^l, i_s^l, d_v, d_c$ 
5    scores  $\leftarrow$  mapreduce(RDD, mapper, reducer)
6     $k^* \leftarrow \text{collectArgmax}(\text{scores})$ 
7     $i_c^{l+1} \leftarrow i_c^l \setminus k^*$ 
8     $i_s^{l+1} \leftarrow i_s^l \cup k^*$ 
9  output  $i_s^L$ 

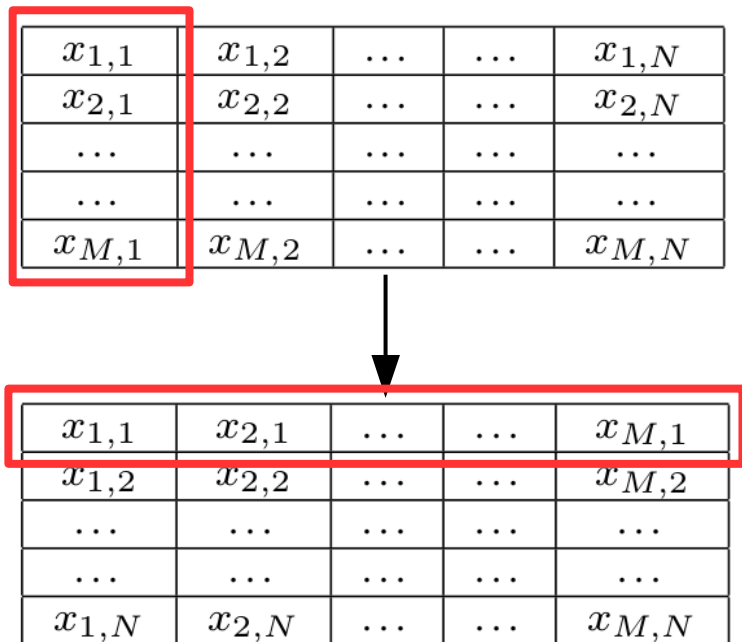
```



mRMR in MapReduce, alternative layout

$x_{1,1}$	$x_{1,2}$	$x_{1,N}$
$x_{2,1}$	$x_{2,2}$	$x_{2,N}$
...
...
$x_{M,1}$	$x_{M,2}$	$x_{M,N}$

mRMR in MapReduce, alternative layout

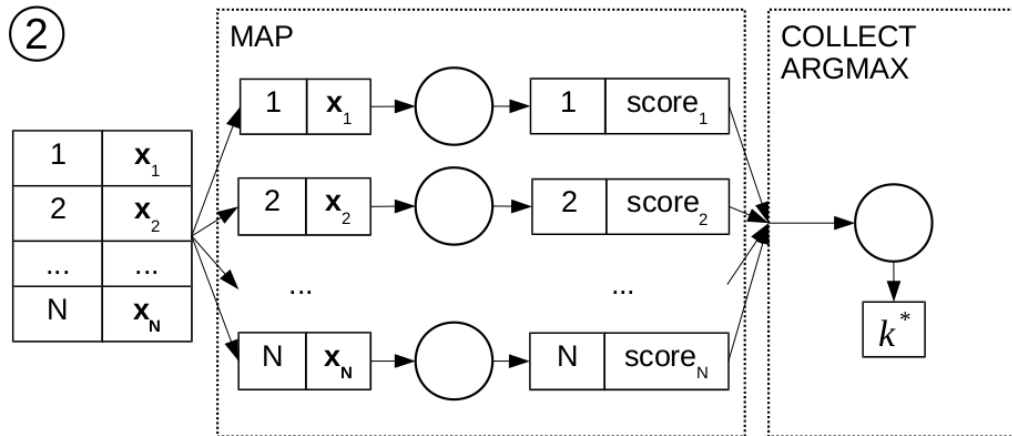


mRMR in MapReduce, alternative layout

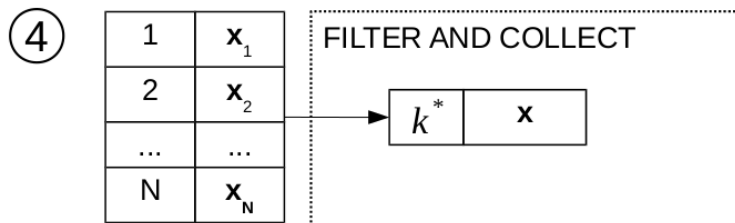
$x_{1,1}$	$x_{1,2}$	$x_{1,N}$
$x_{2,1}$	$x_{2,2}$	$x_{2,N}$
...
...
$x_{M,1}$	$x_{M,2}$	$x_{M,N}$



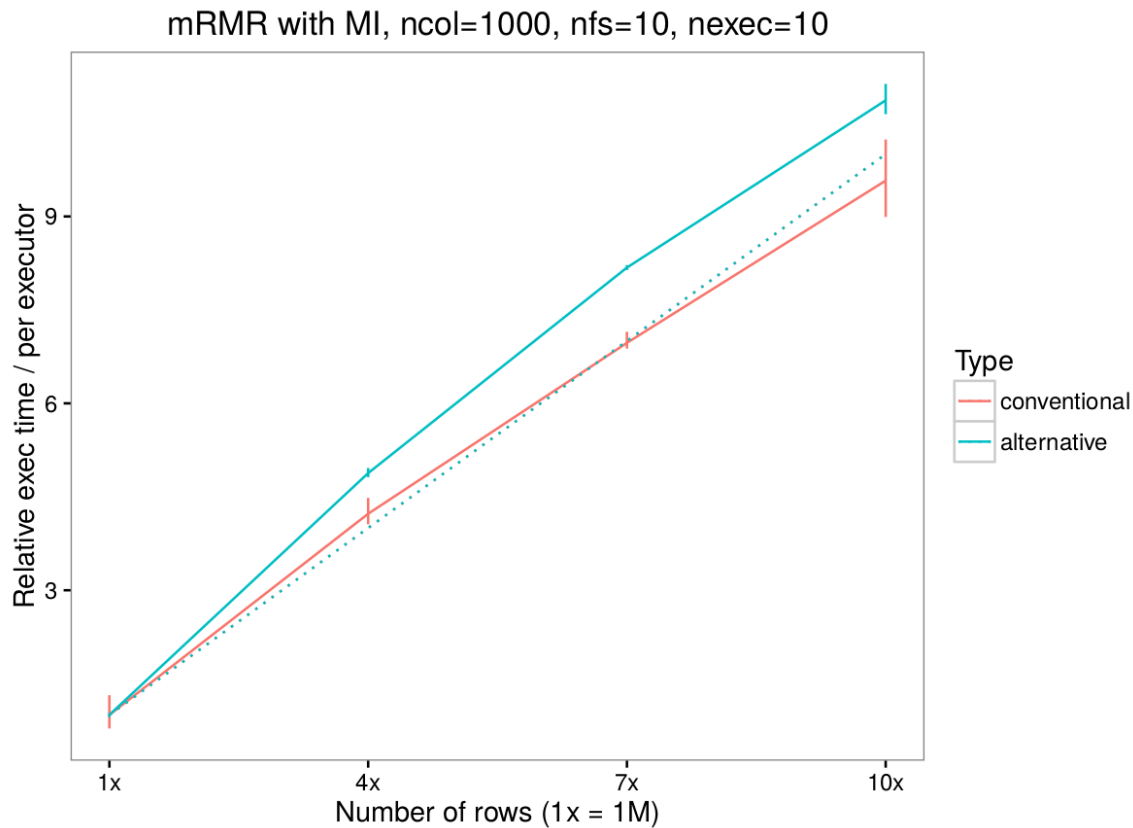
$x_{1,1}$	$x_{2,1}$	$x_{M,1}$
$x_{1,2}$	$x_{2,2}$	$x_{M,2}$
...
...
$x_{1,N}$	$x_{2,N}$	$x_{M,N}$



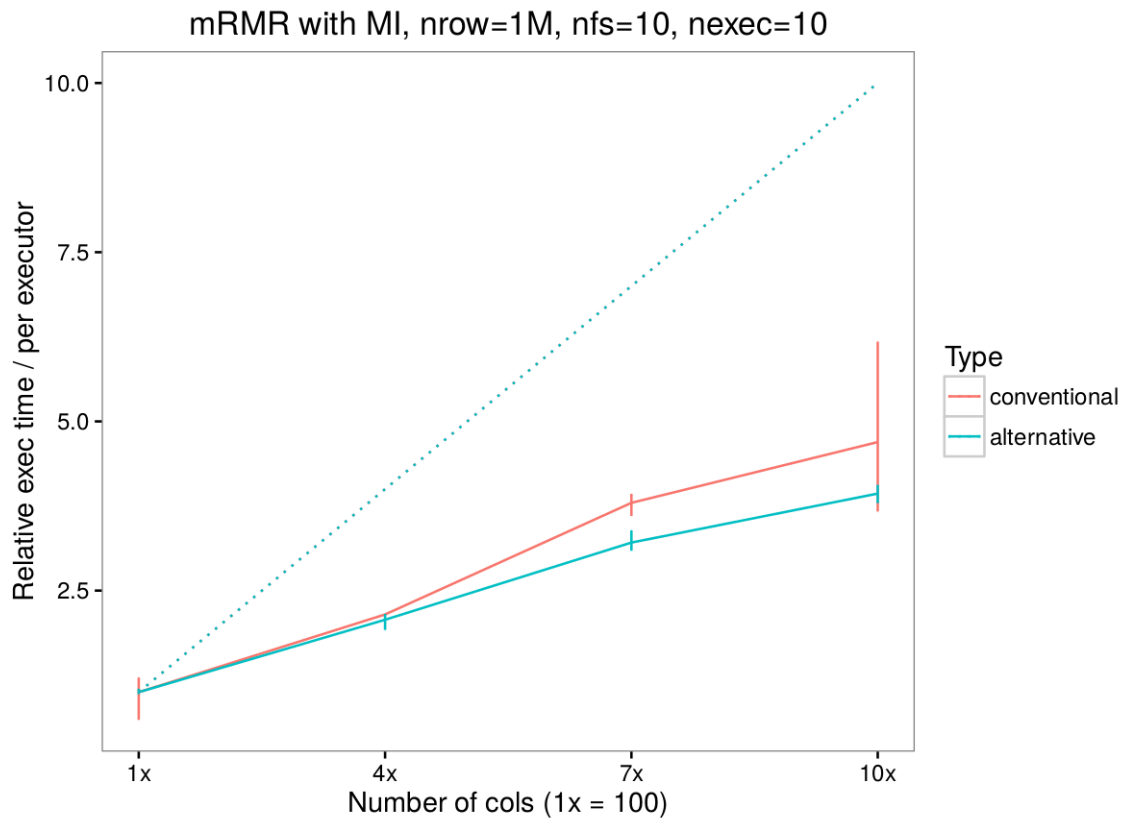
③ Broadcast: k^*



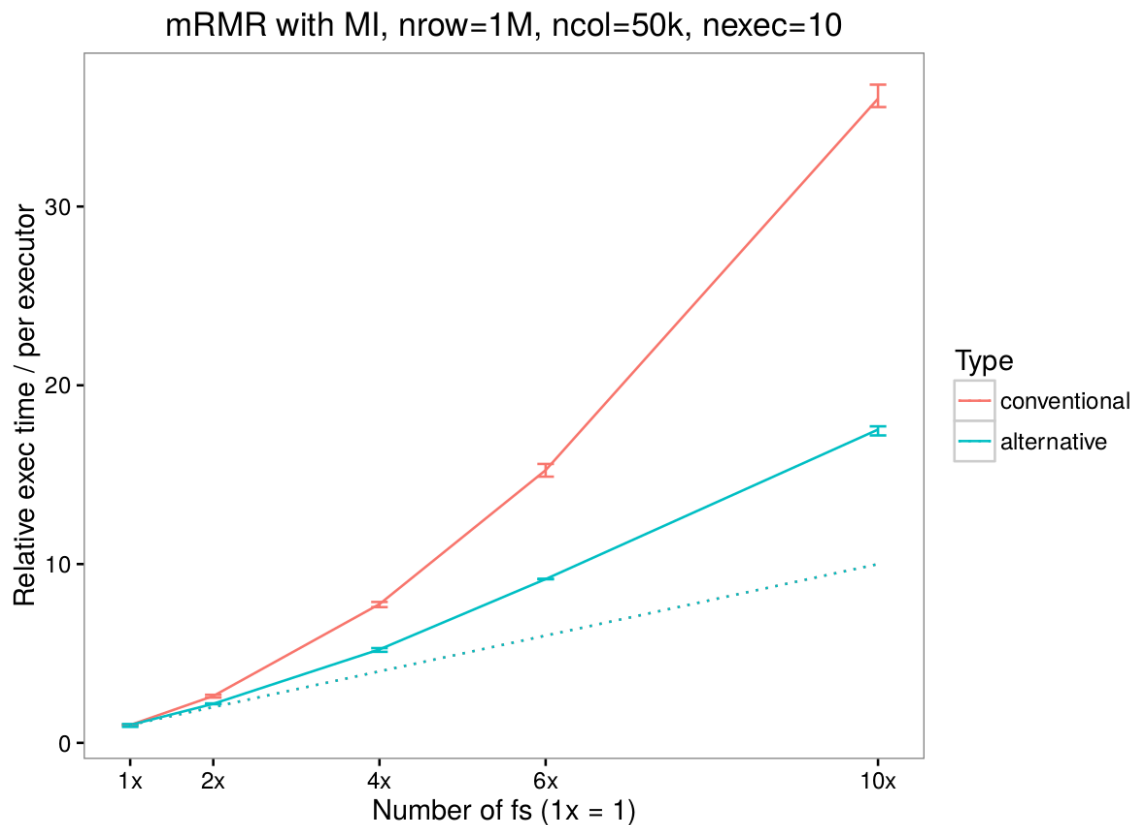
Scalability: #rows



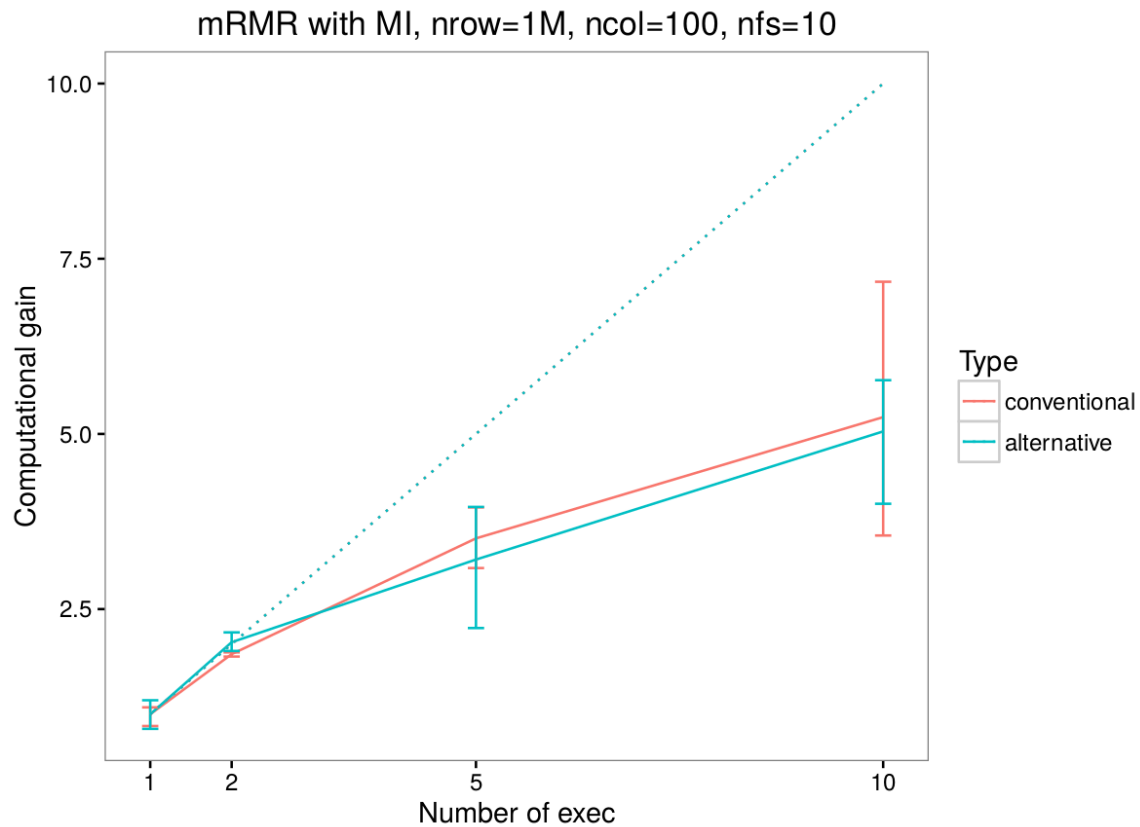
Scalability: #columns



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

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

```
1 function getResult:
2   arguments:
3     variableArray: Array[Double]
4     classArray: Array[Double]
5     selectedVariablesArray:
6       Array[Array[Double]]
7   return: Double
```




Yann-Aël
Le Borgne



Gianluca
Bontempi



github.com/creggian/spark-ifs



github.com/creggian/slides



claudioireggiani.com

