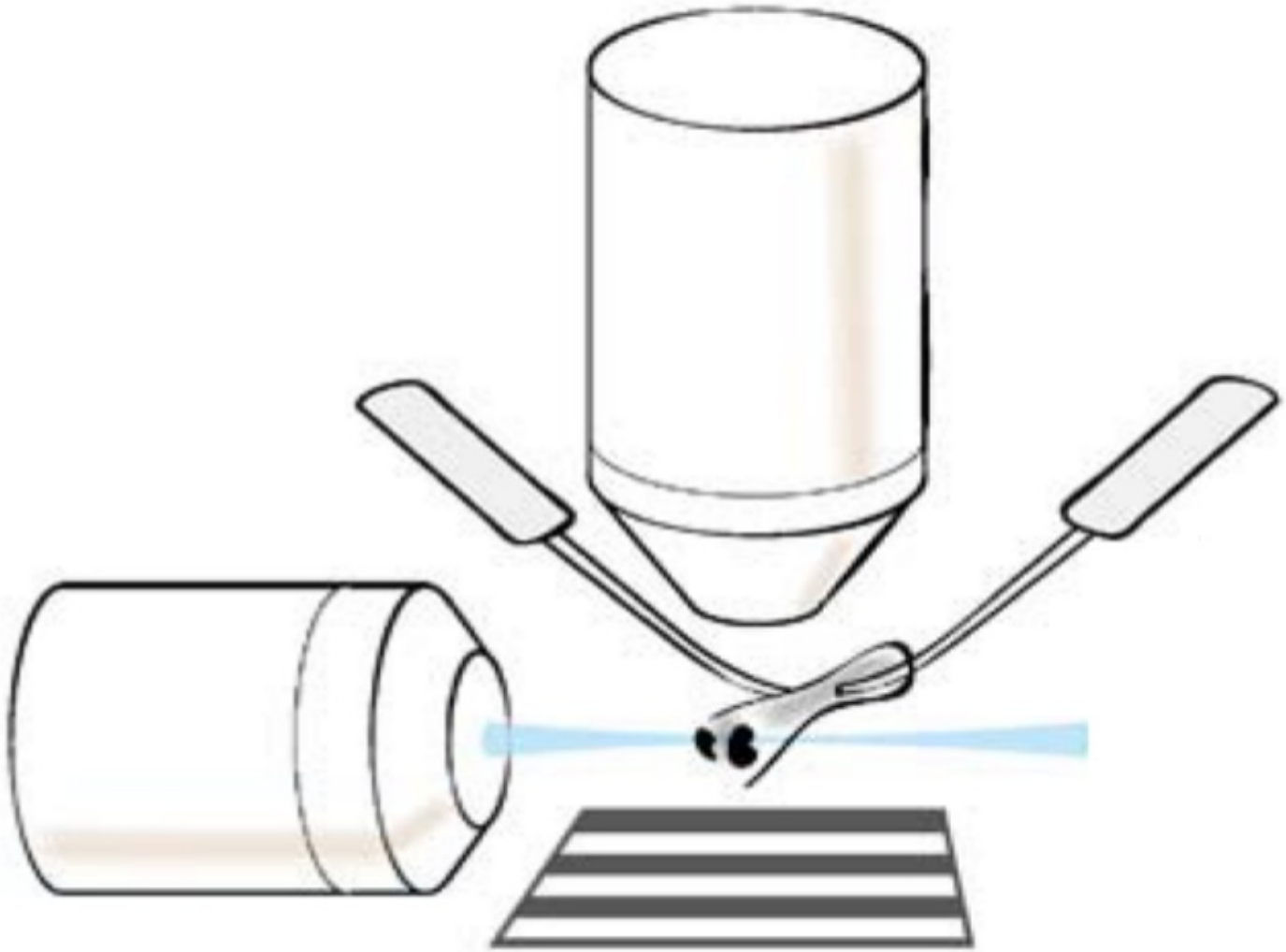
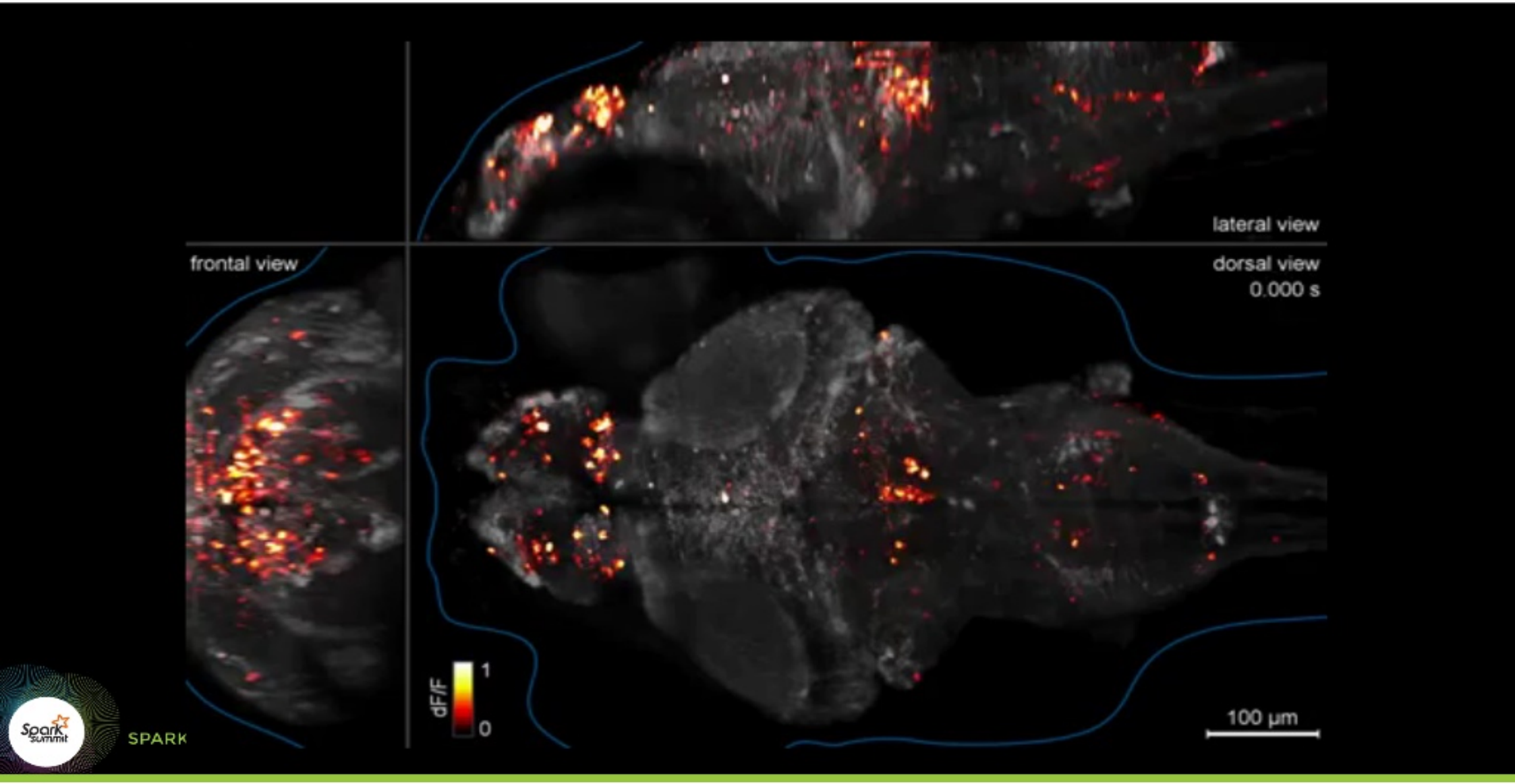
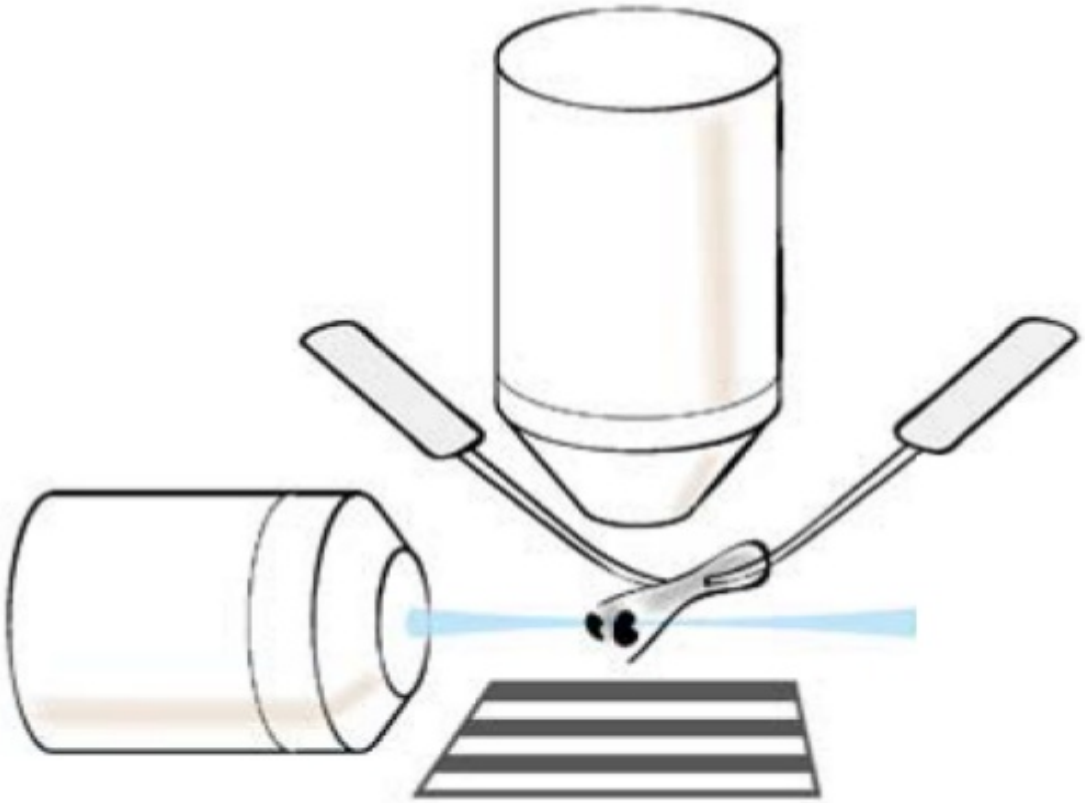
Bolt:
Building a distributed ndarray

Jason Wittenbach
Janelia Research Campus (HHMI)
Freeman Lab

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO




lateral view
frontal view
dorsal view
0.000 s
dF/F
1
0
100 µm



t, (x, y, z)

time ~ 10^4

space ~ 10^7

~ 10^{11} elements

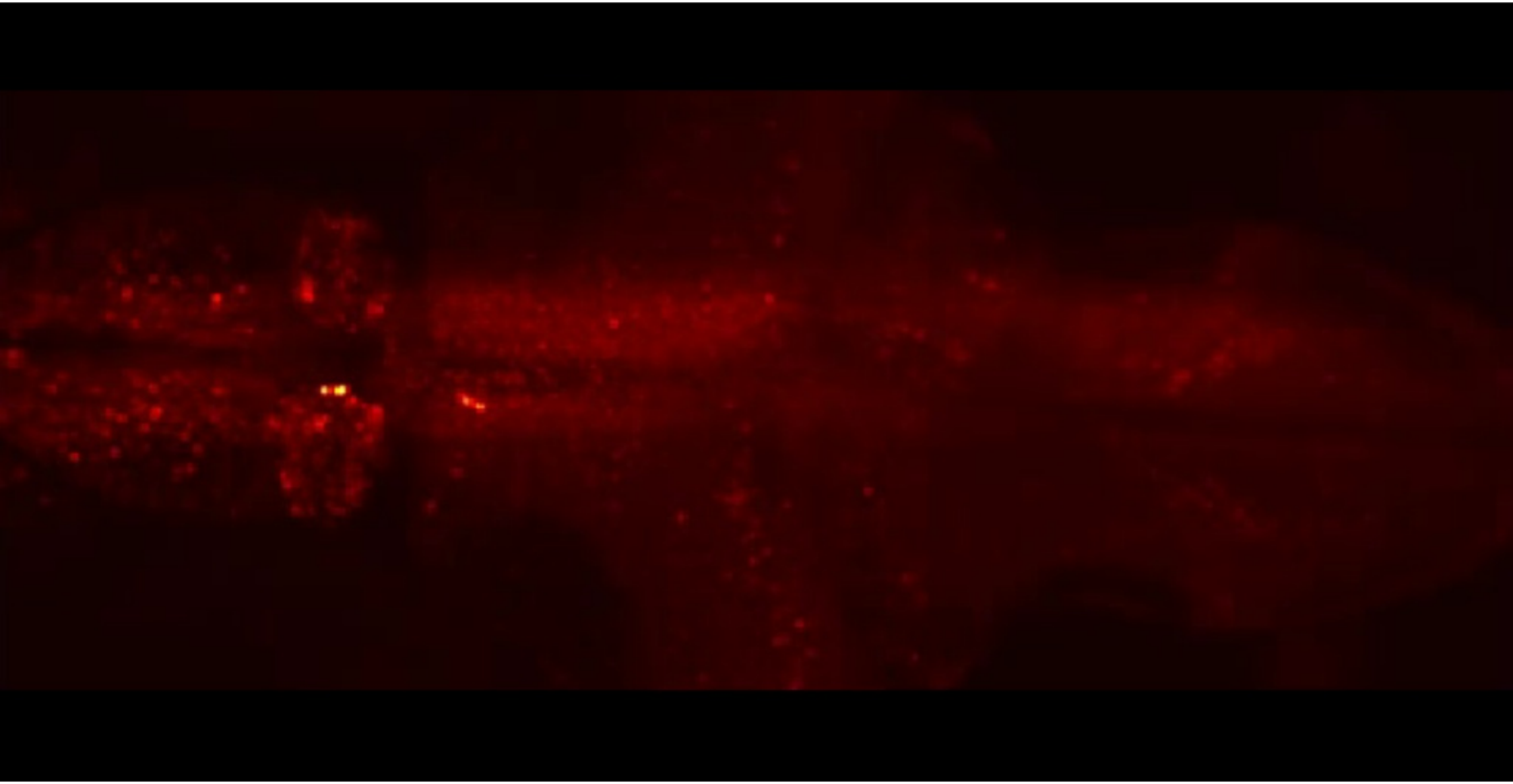
~ 1 TB

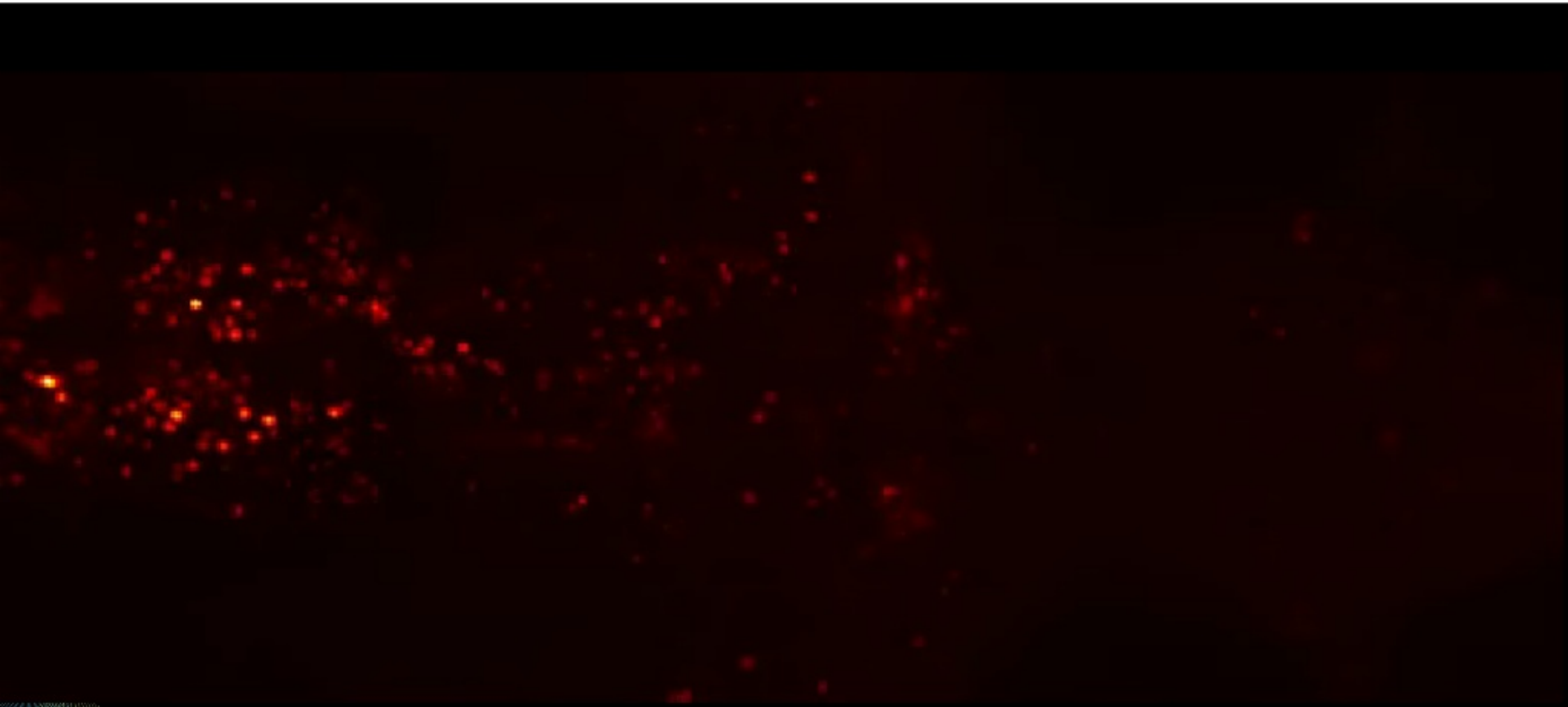
(n, k)

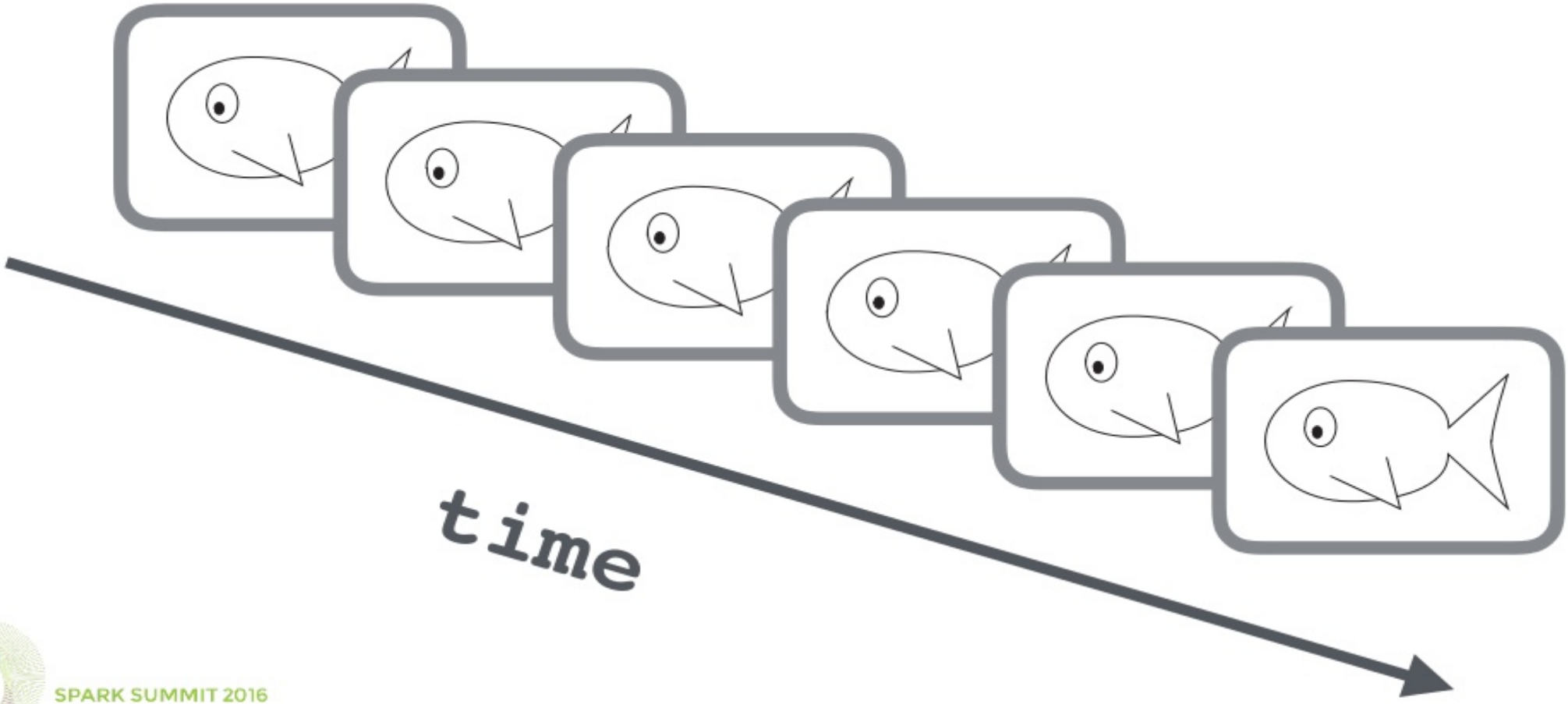
(x, y, t)

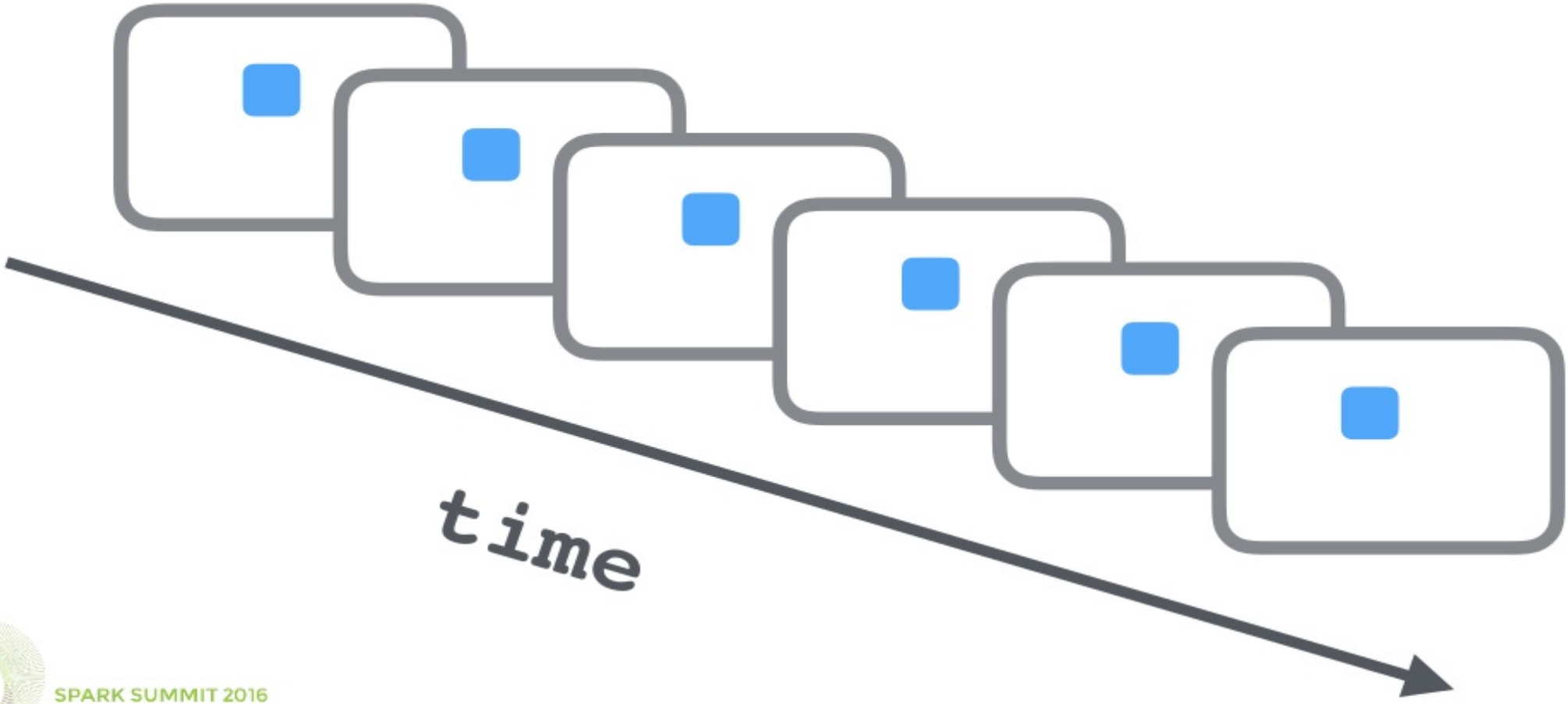
(x, y, z, t)

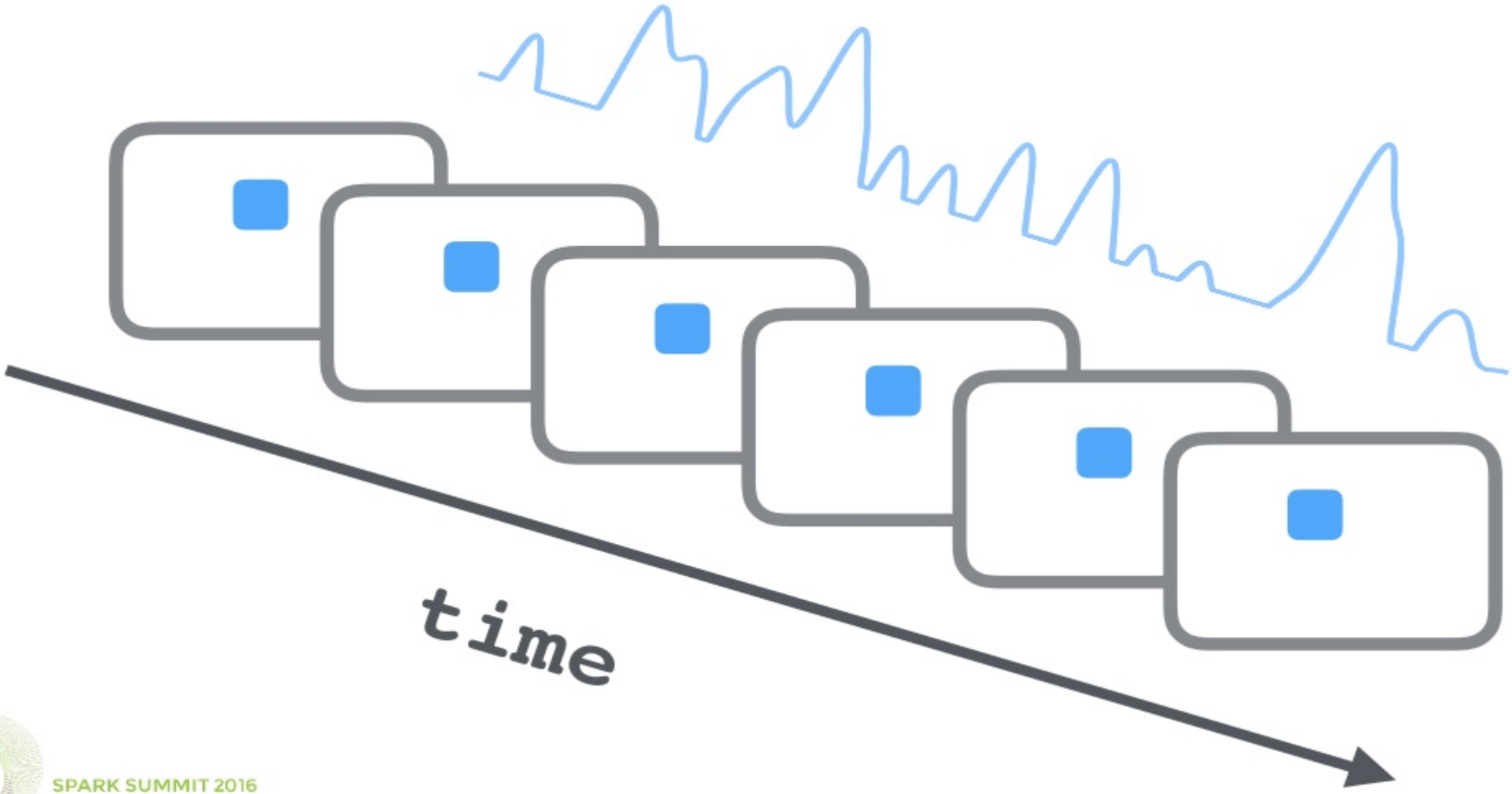
(x, y, z, c, t)

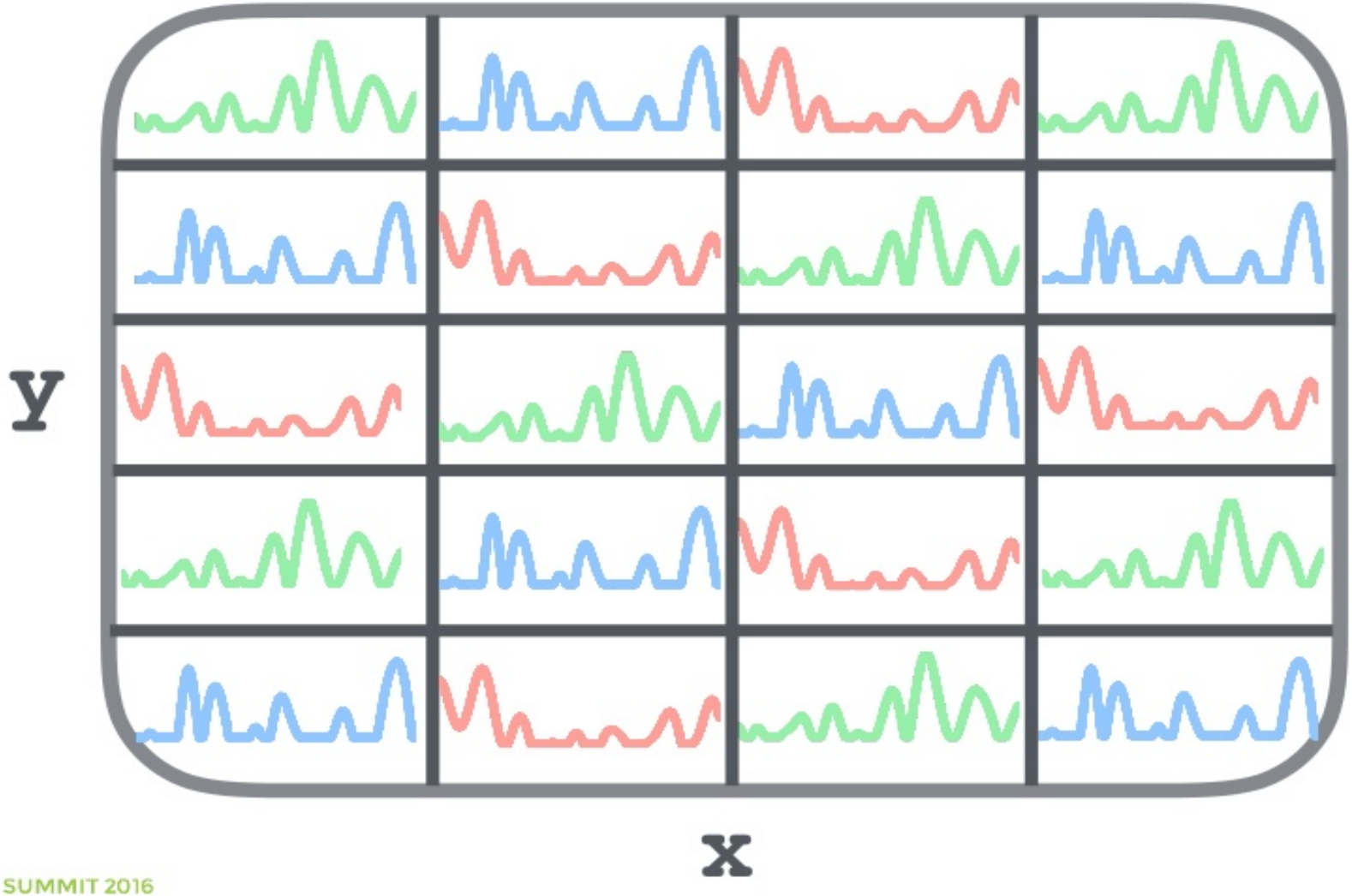





time


time


time


y
x

neuroscience

astronomy

geospatial

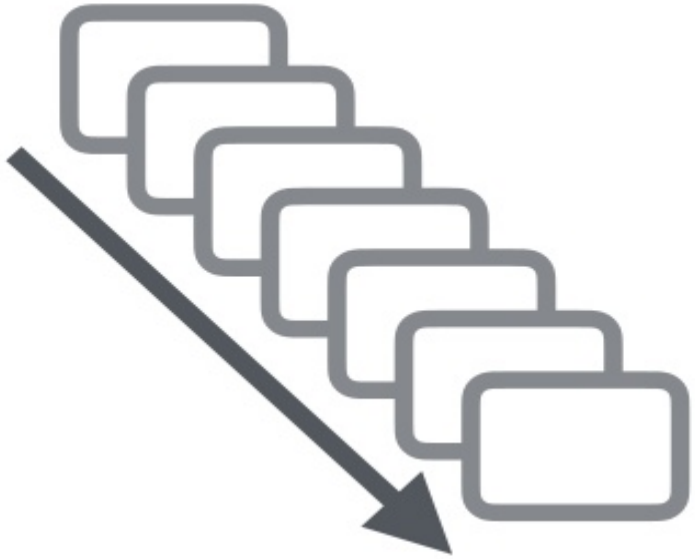
climate science

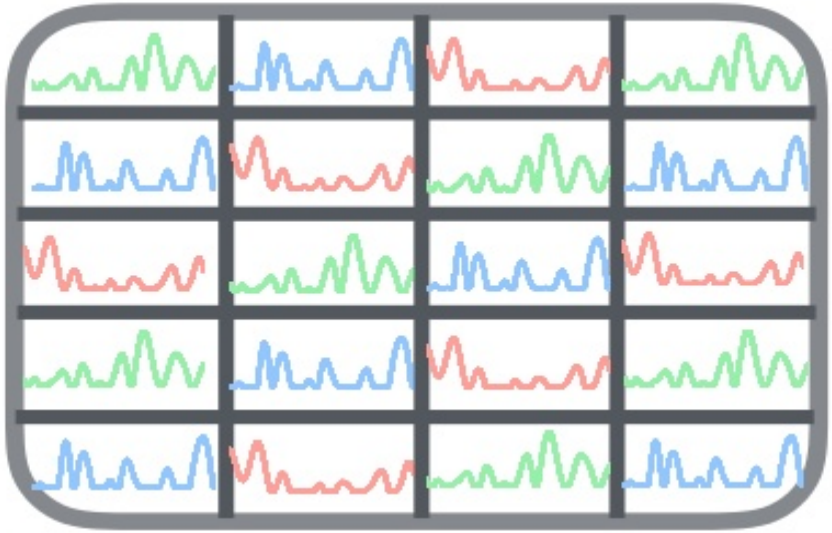
- a distributed ndarray
- built on PySpark
- conforms to NumPy API

```
data.mean(axis=0)

data.T

data[2, 4:10]
```



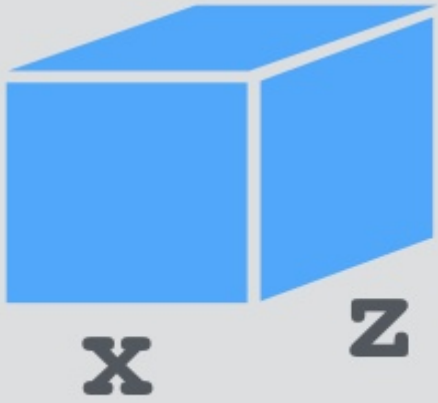


(t,| x, y, z)

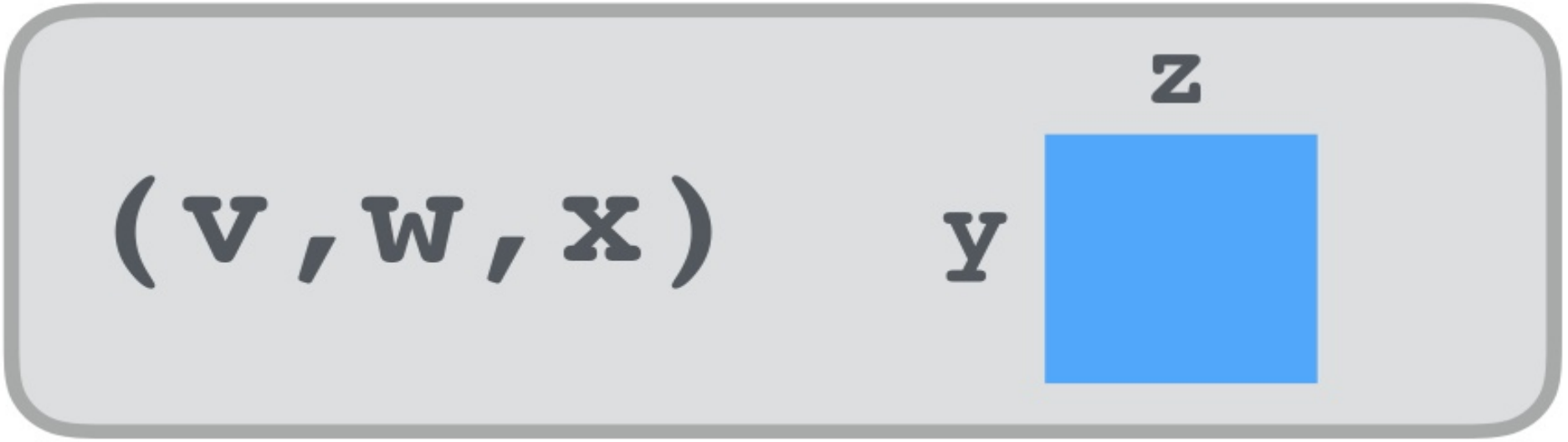
(x, y, z,| t)

(v, w, | x, y, z)

(v,w) y

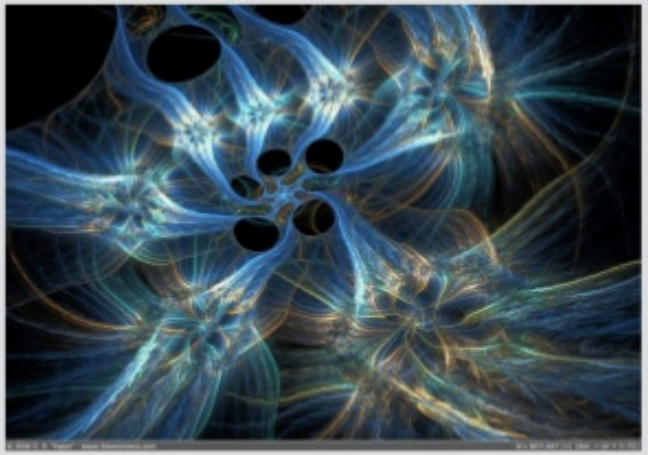


(v, w, x, | y, z)



```
(v, | w, x, y, z)
```

```
(v)
```



indexing slicing

apply-along-axis

indexing slicing

apply-along-axis

transpose reshape

indexing slicing

apply-along-axis

transpose reshape

map reduce filter

chunking padding

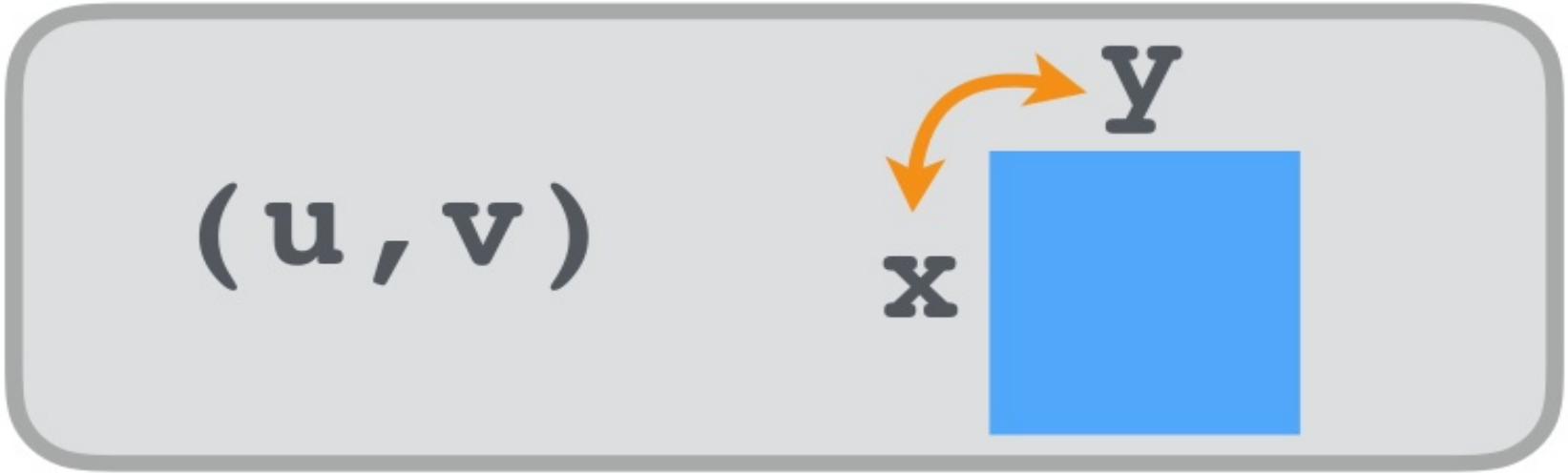
indexing → filter
map

apply-along-axis

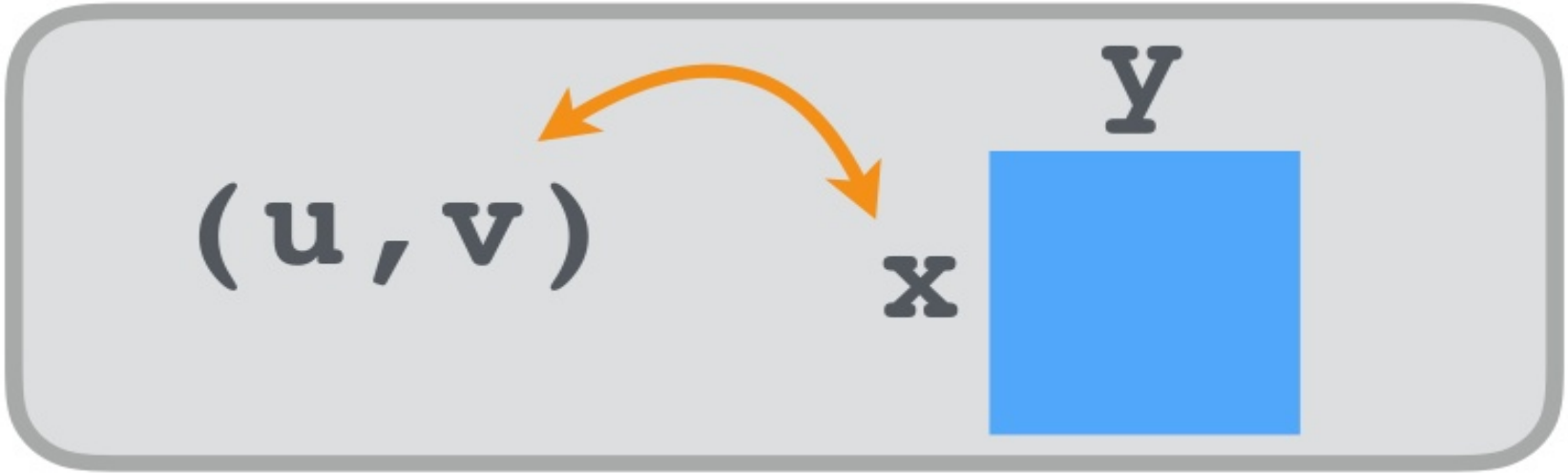
map reduceByKey

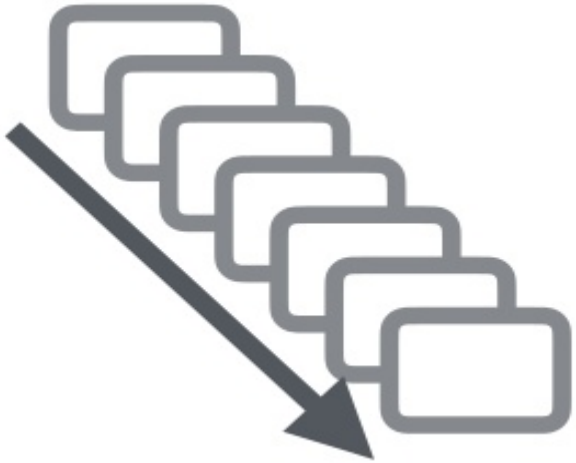
transpose → map

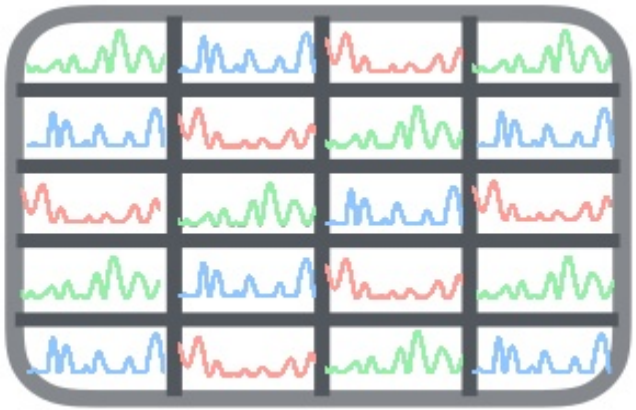
transpose → map



transpose → shuffle





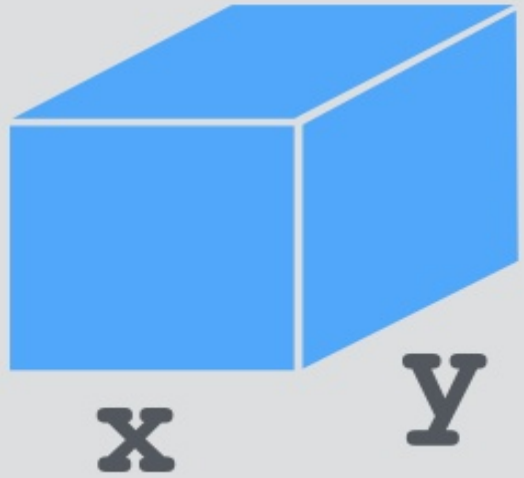


(t | x,y,z)

(x,y,z | t)

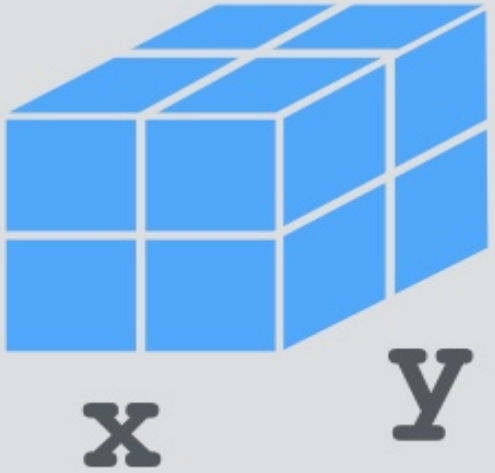
(t | x, y, z)

(t)

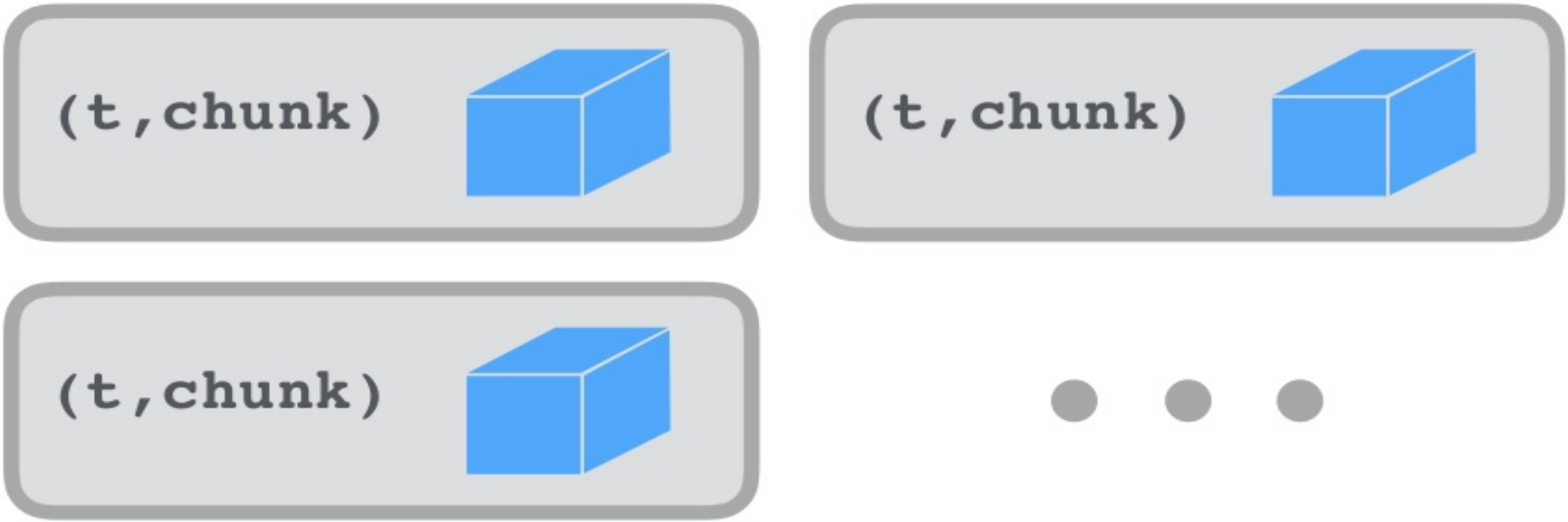


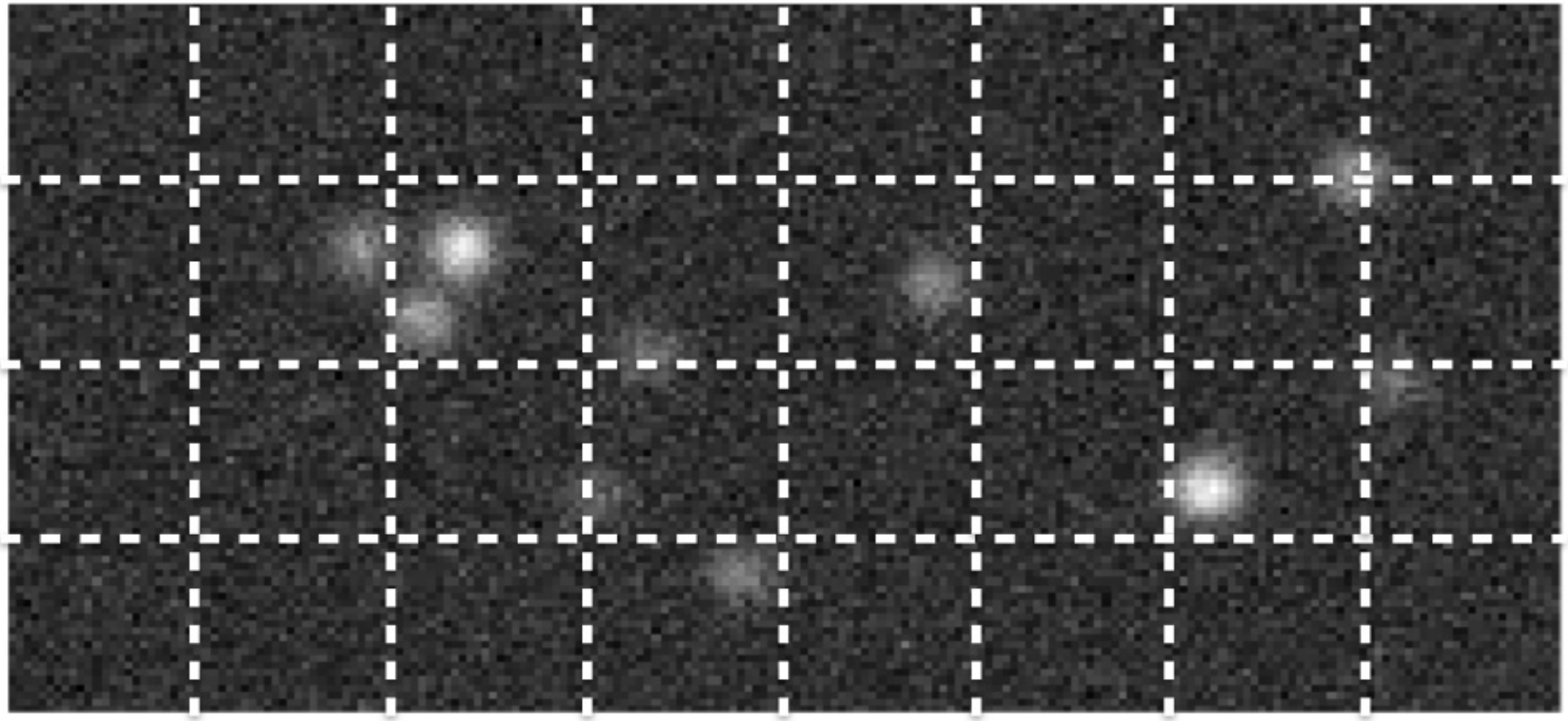
(t|x, y, z)

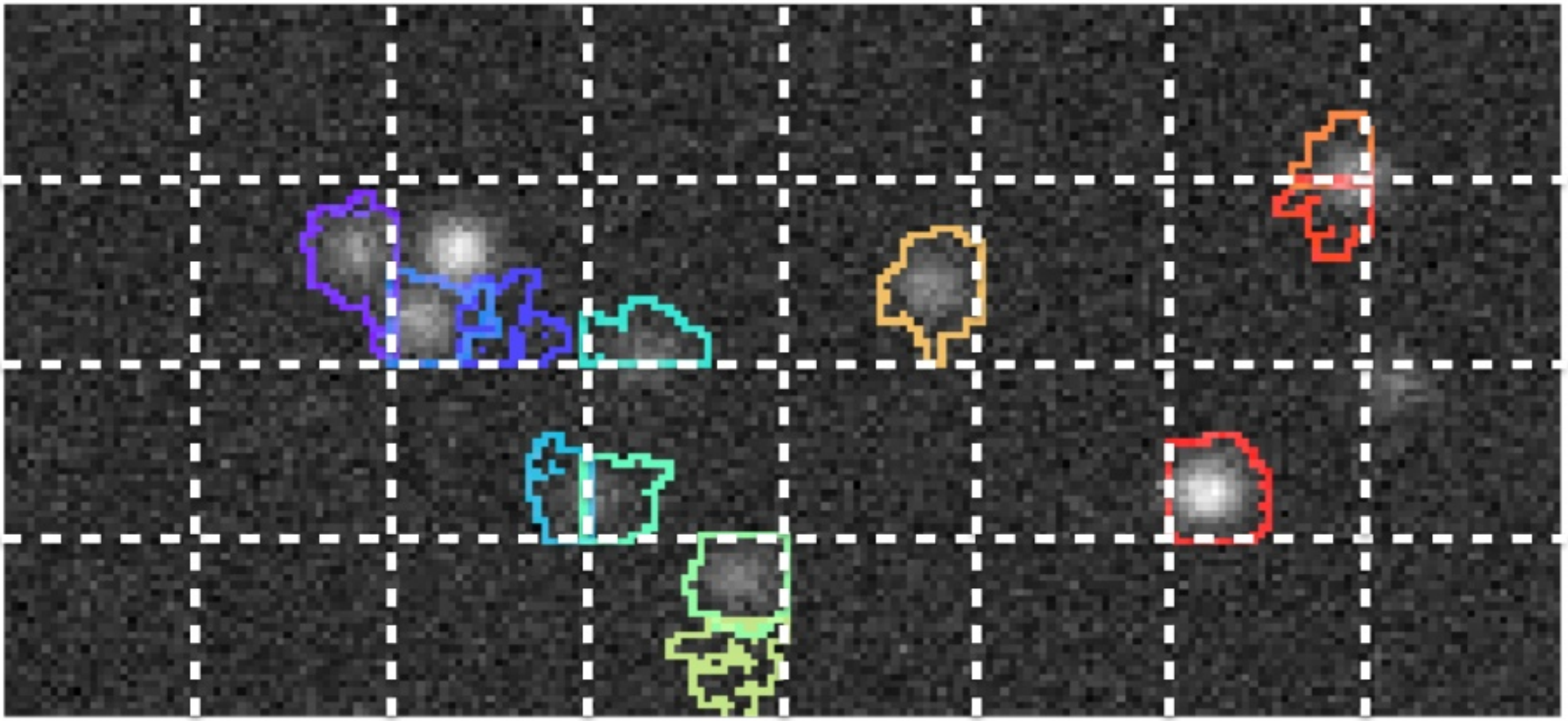
(t)

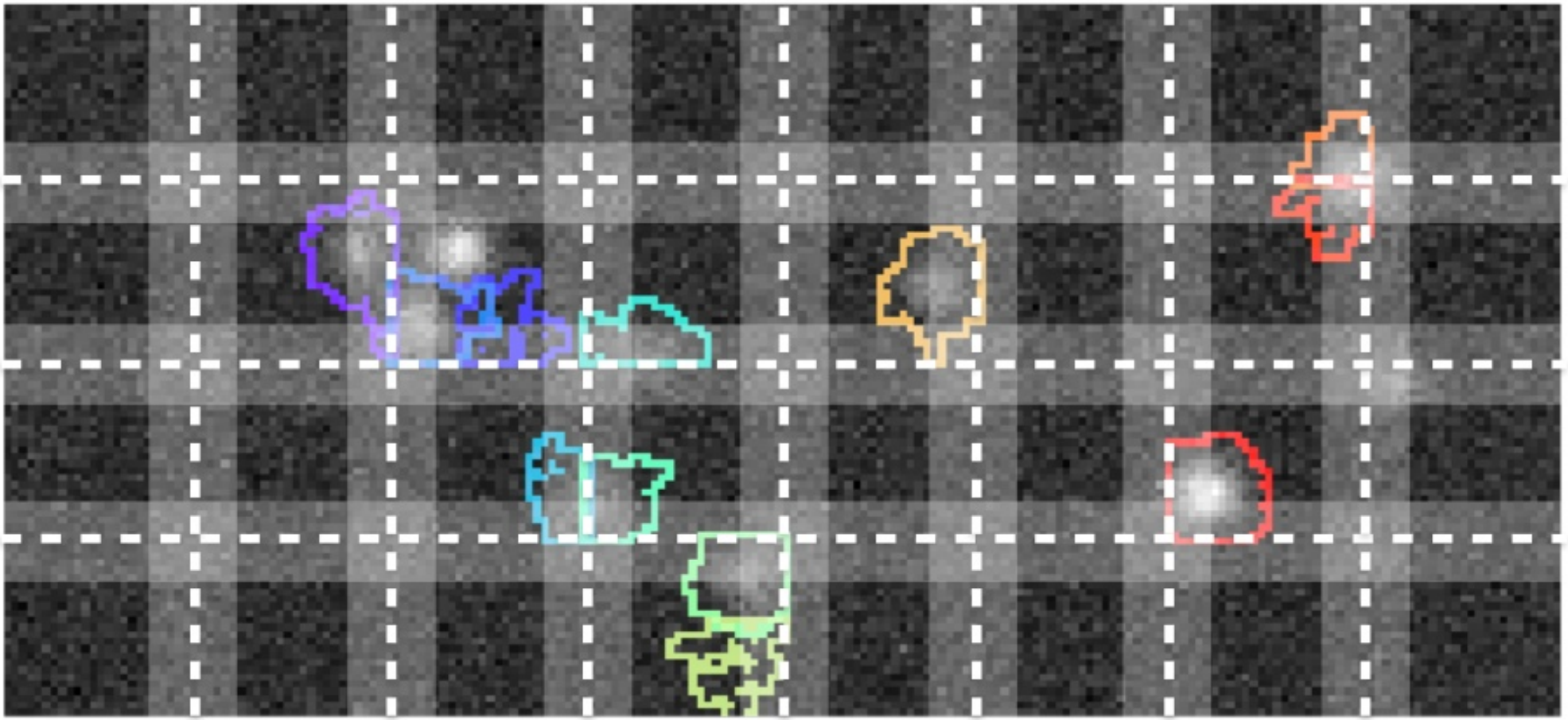


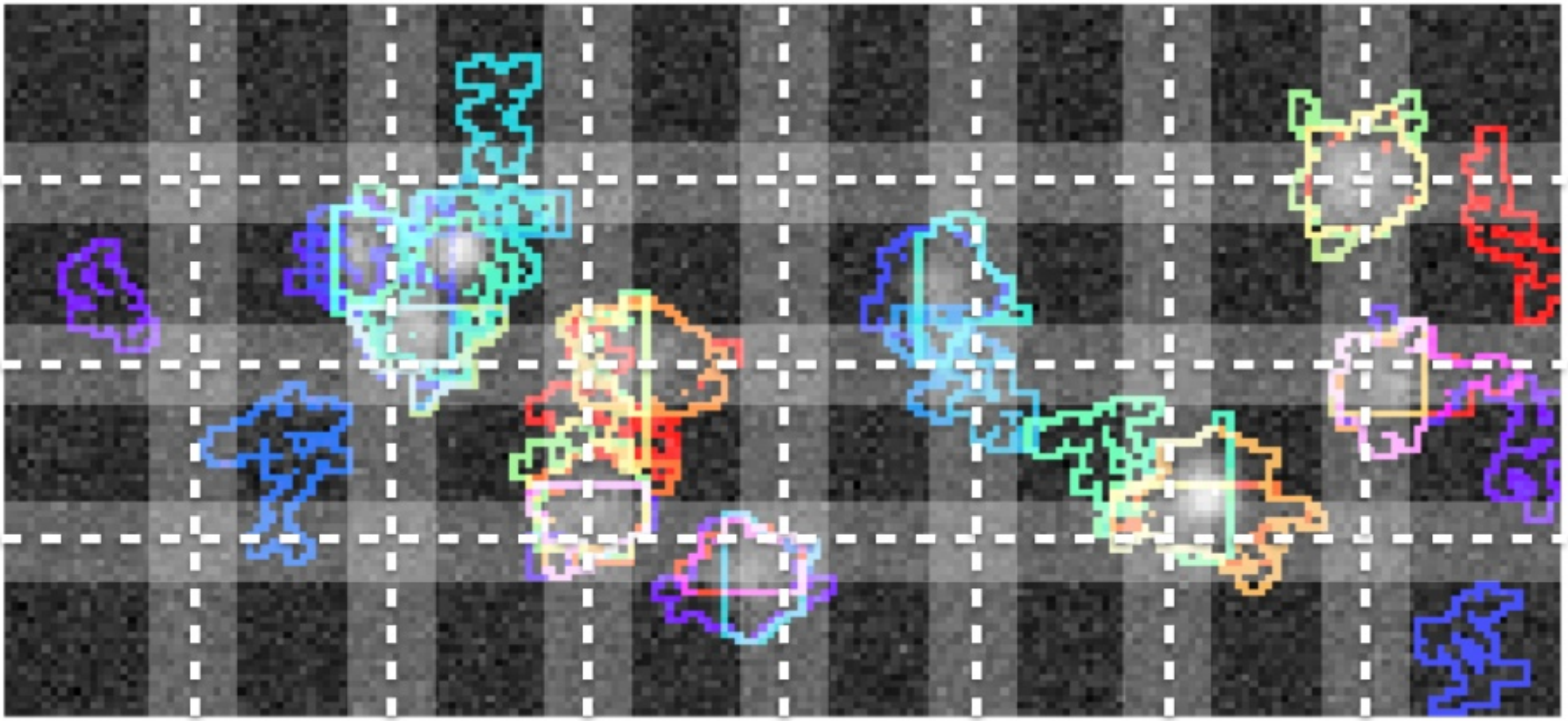
(t | x, y, z)

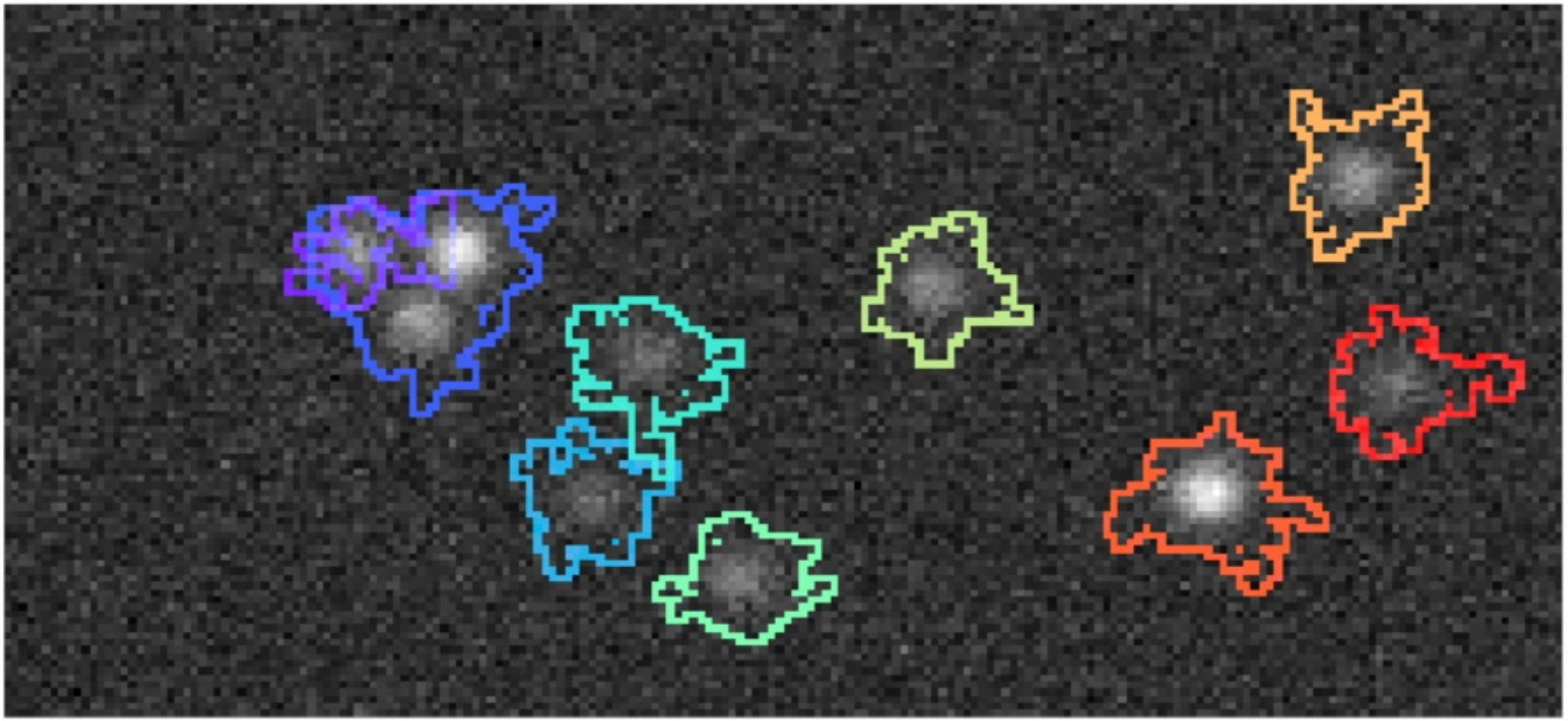












chunking

+

transpose

=

shuffle optimization

thunder

???

thanks

Freeman Lab

Jeremy Freeman
Nicholas Sofroniew
Andrew Osheroff

Janelia Scientific Computing

Ken Carlile
Robert Lines

join us!

GitHub

bolt-project
thunder-project

Twitter

@jsonWittenbach