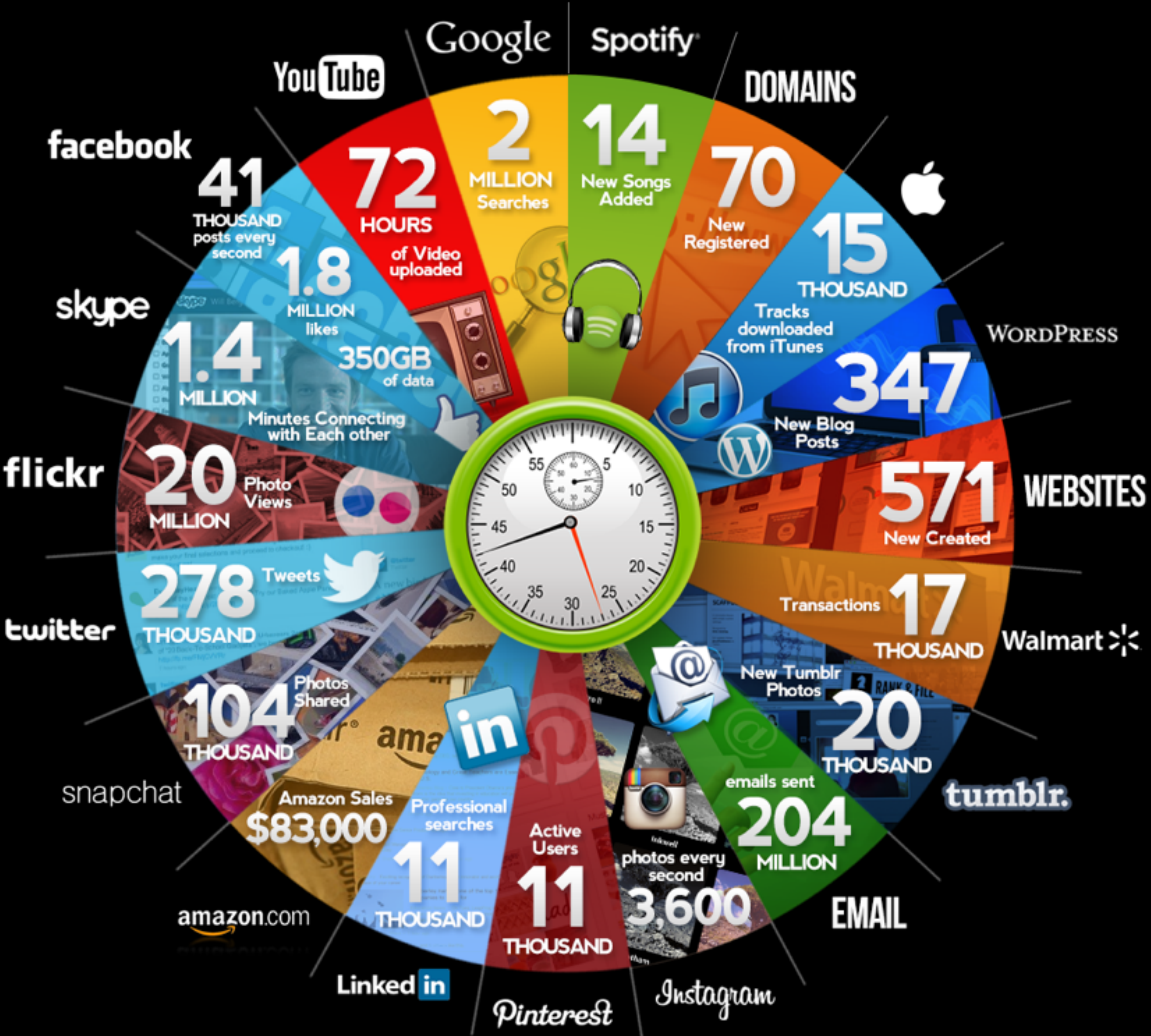
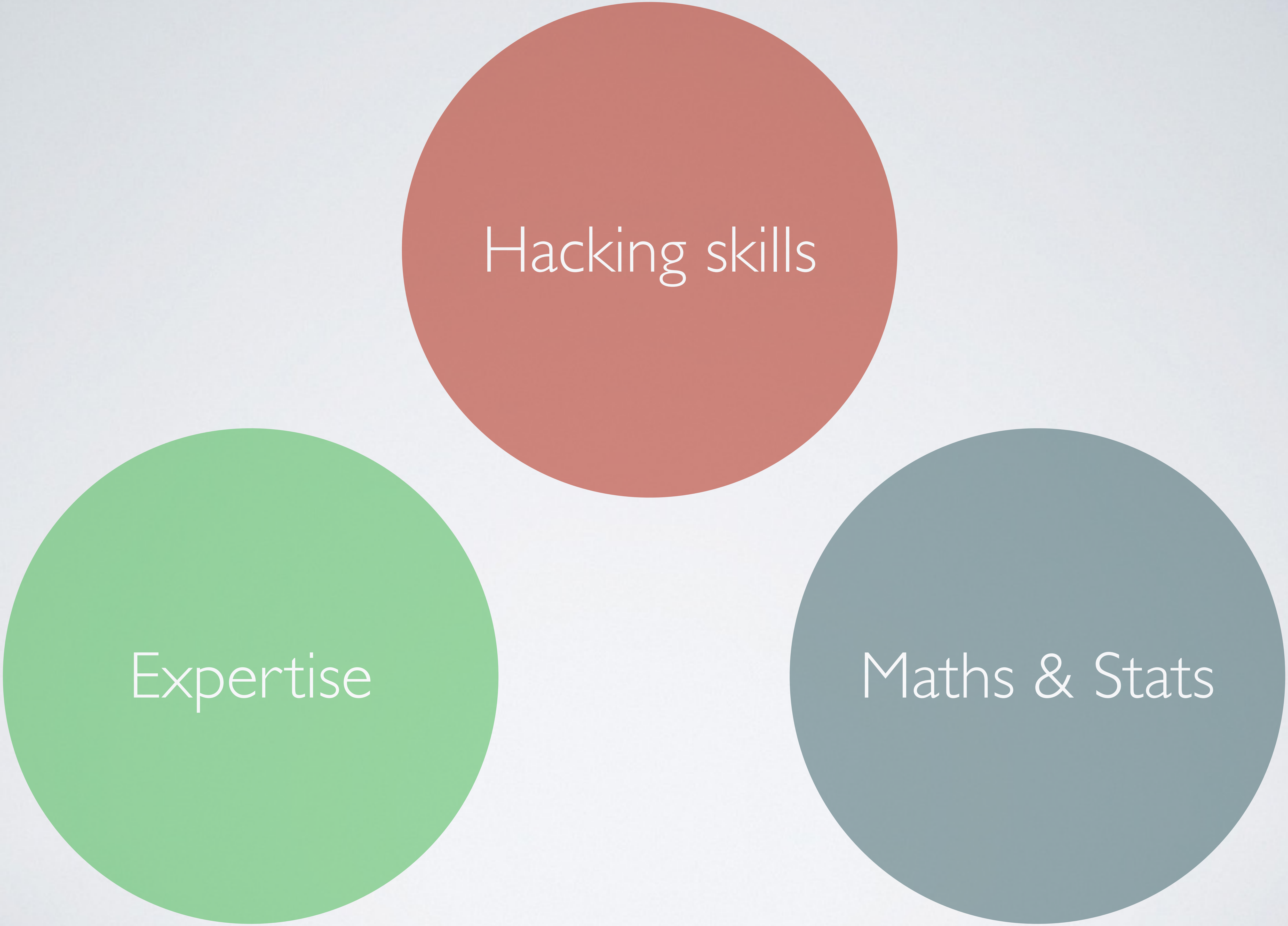
Data Science

joaquin vanschoren


Google
2 MILLION Searches
Spotify
14 New Songs Added
DOMAINS
70 New Registered
15 THOUSAND Tracks downloaded from iTunes
WordPress
347 New Blog Posts
WEBSITES
571 New Created
Walmart
Transactions 17 THOUSAND
tumblr.
New Tumblr Photos 20 THOUSAND
EMAIL
emails sent 204 MILLION
Instagram
photos every second 3,600
Pinterest
Active Users 11 THOUSAND
LinkedIn
Professional searches 11 THOUSAND
amazon.com
Amazon Sales $83,000
snapchat
Photos Shared 104 THOUSAND
twitter
278 THOUSAND Tweets
flickr
20 MILLION Photo Views
skype
1.4 MILLION Minutes Connecting with Each other
facebook
41 THOUSAND posts every second
1.8 MILLION likes
350GB of data
YouTube
72 HOURS of Video uploaded
55
5
50
10
45
15
40
20
35
30
25

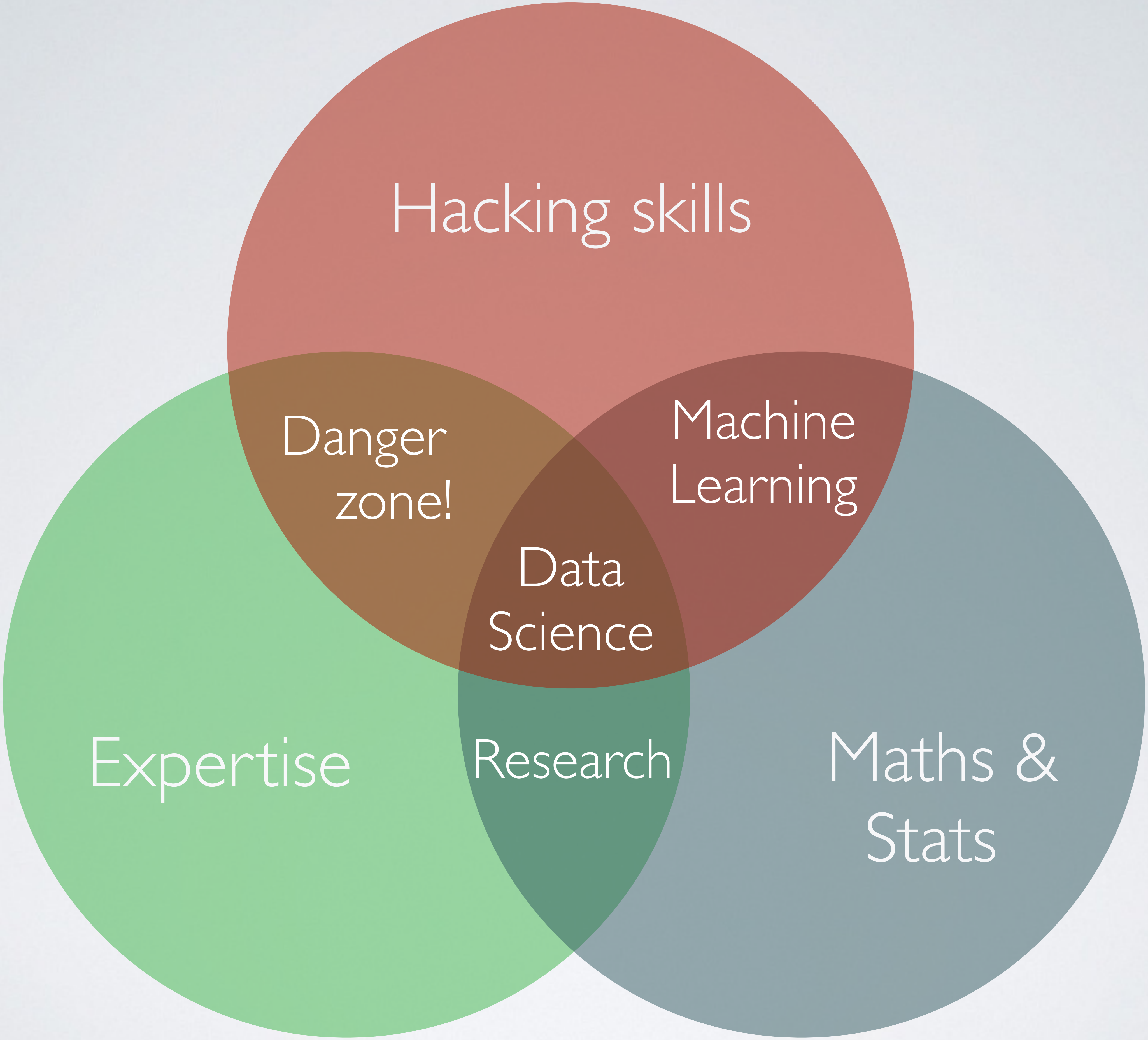
WHAT IS DATA SCIENCE?



Hacking skills

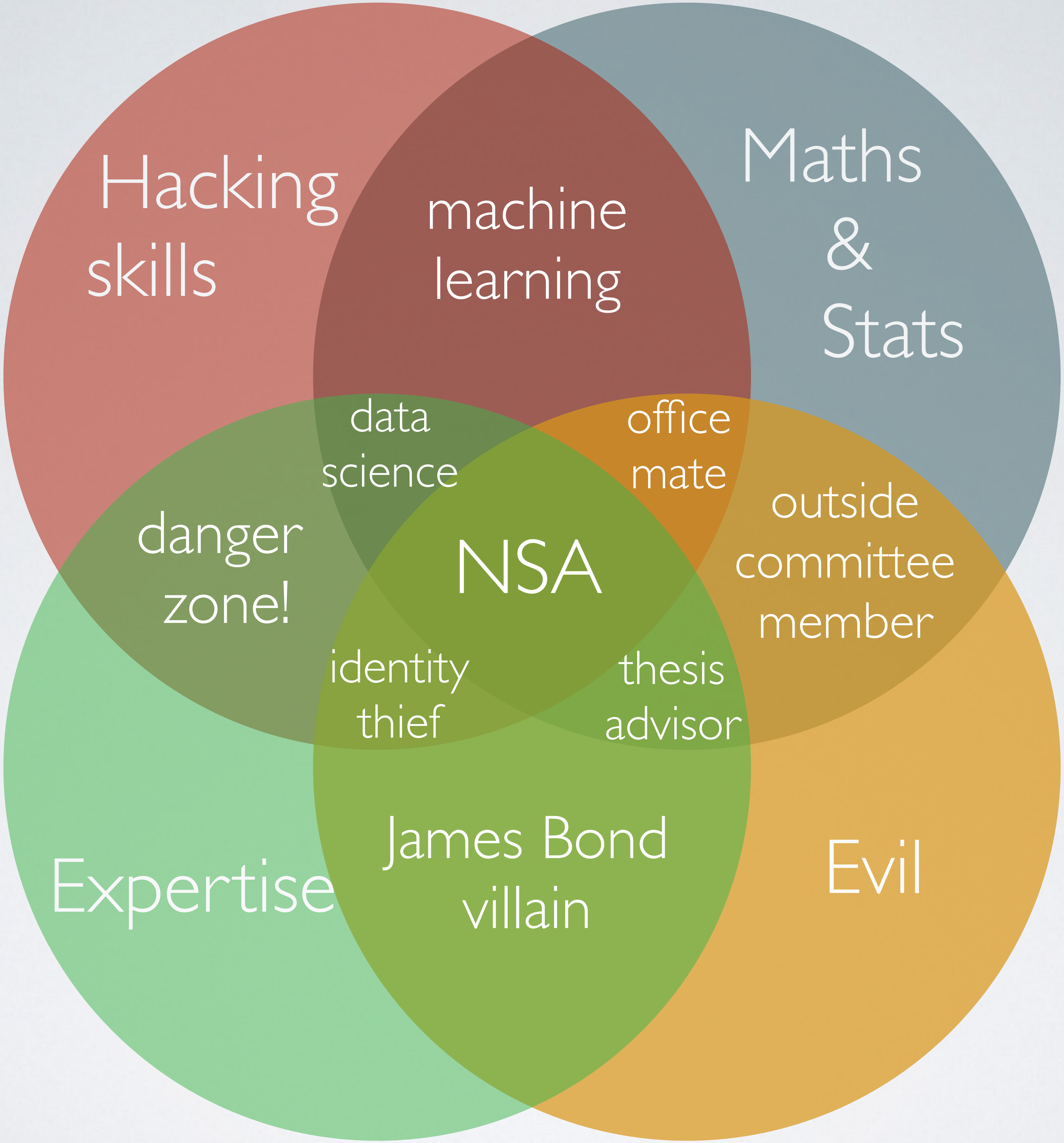
Expertise

Maths & Stats


Hacking skills
Danger zone!
Machine Learning
Data Science
Expertise
Research
Maths & Stats

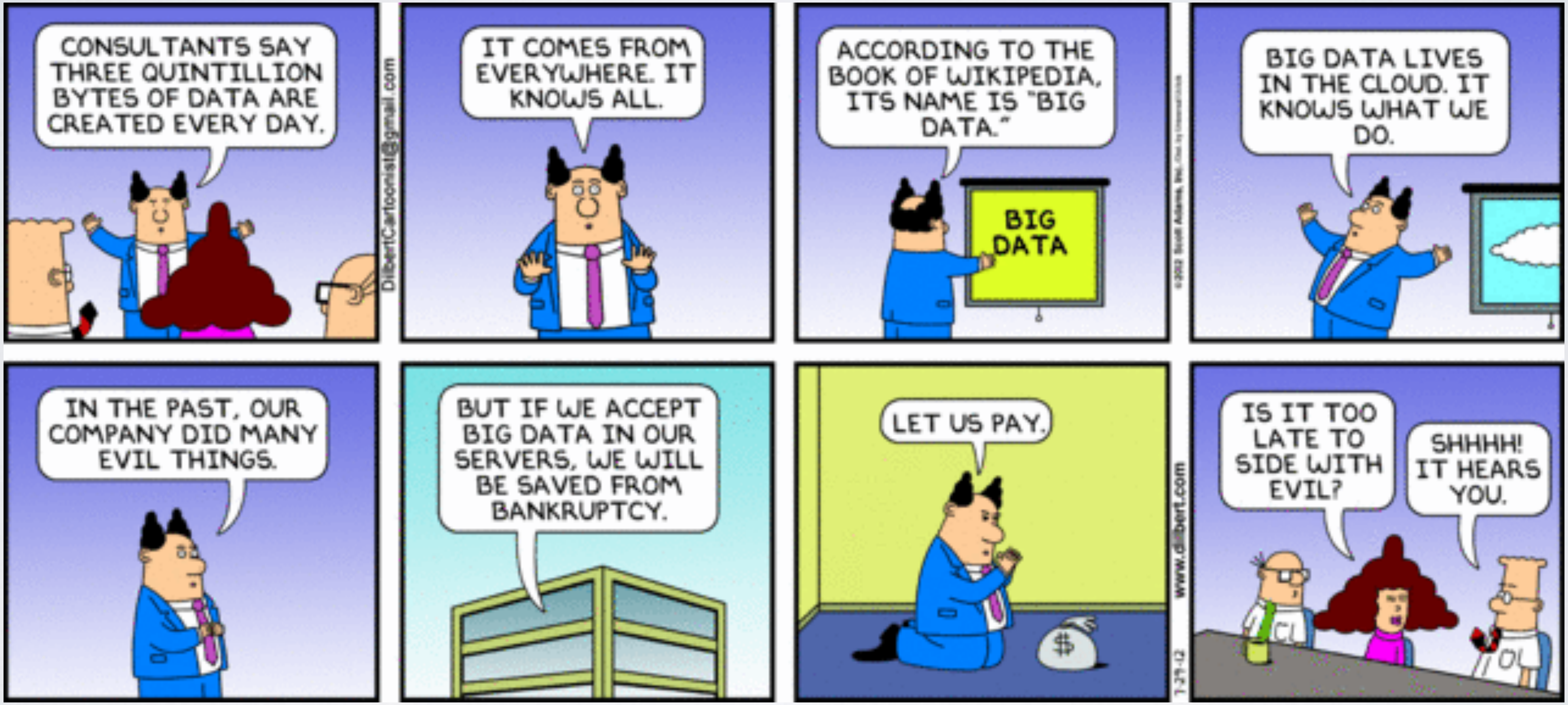
[Drew Conway]

(data)
science officer


Hacking skills
machine learning
Maths & Stats
data science
office mate
danger zone!
NSA
outside committee member
identity thief
thesis advisor
Expertise
James Bond villain
Evil

[Joel Grus]

THE HYPE


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
www.dilbert.com
7-29-12
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

THE HYPE

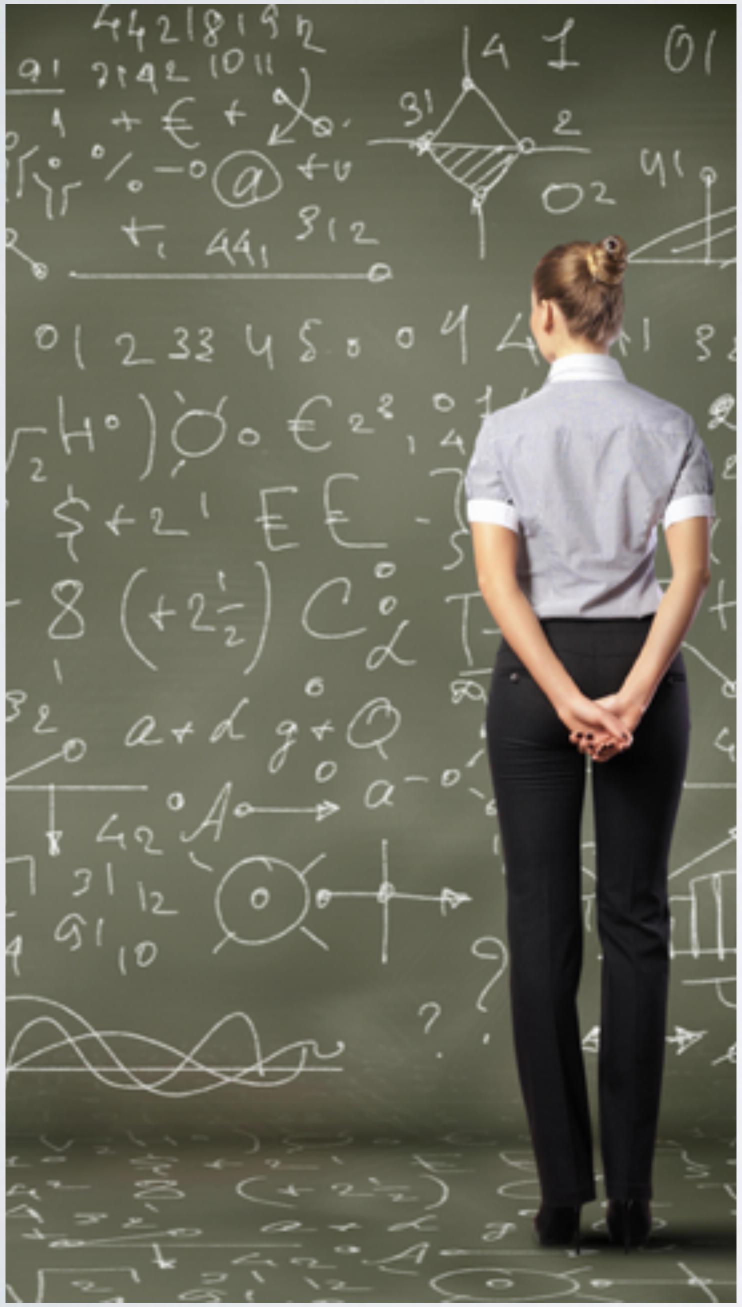
"Data Scientist: The Sexiest Job of the 21st Century"

– Harvard Business Review

"Whenever you read about data science or data analysis, it's about the ability to store petabytes of data, retrieve that data in nanoseconds, then turn it into a rainbow with a unicorn dancing on it."

– David Coallier

THE REALITY

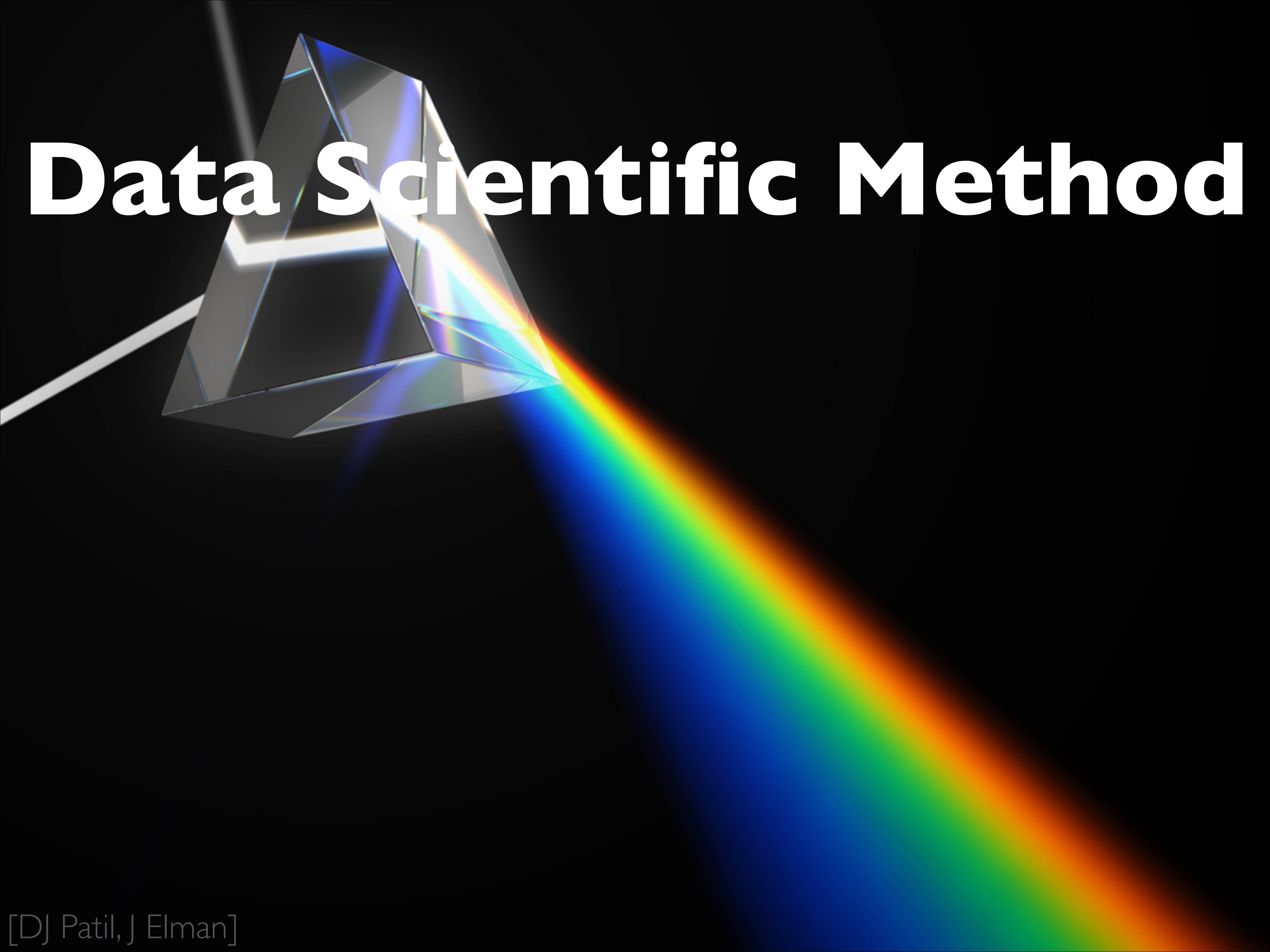


- You'll clean a lot of data. A LOT
- A lot of mathematics. Get over it
- Some days will be long. Get more coffee
- Not everything is about Big Data
- Most people don't care about data
- Spend time finding the right questions

[David Coallier]

Big Data and Open Data are fun, but what really matters is what you learn from it.

Data Scientific Method



[DJ Patil, J Elman]

START WITH A QUESTION

Based on an observation

ANALYSE CURRENT DATA

Create an Hypothesis

CREATE FEATURES, EXPERIMENT

Test Hypothesis

ANALYSE RESULTS

Won’t be pretty, repeat

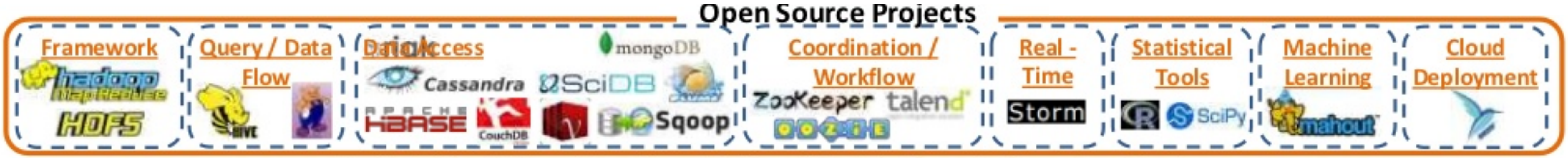
LET DATA FRAME THE CONVERSATION

Data gives you the **what**
Humans give you the **why**

CONVERSE

- What data is missing? Where can we get it?
- Automate data collection
- Clean data, then clean it more
- Visualize data: the brain *sees*
- Merge various sources of information
- Reformulate hypotheses
- Reformulate questions

DATA SCIENCE TOOLS



Open Source Projects

Framework

Query / Data Flow

Data Access

Coordination / Workflow

Real - Time

Statistical Tools

Machine Learning

Cloud Deployment

R

modelling, testing, prototyping

lubridate, zoo: dates, time series

reshape2: reshape data

ggplot2: visualize data

RCurl, RJSONIO: find more data

HMisc: miscellaneous

DMwR, mlr: machine learning

Forecast: time series forecasting

garch: time series modelling

quantmod: statistical financial trading

xts: extensible time series

igraph: study networks

maptools: read and view maps

PYTHON

scientific computing

numpy: linear algebra
scipy: optimization, signal/image processing, …
scikits: toolkits for scipy
scikit-learn: machine learning toolkit
statsmodels: advanced statistic modelling
matplotlib: plotting
NLTK: natural language processing
PyBrain: more machine learning
PyMC: Bayesian inference
Pattern: Web mining
NetworkX: Study networks
Pandas: easy-to-use data structures

OTHER

mongoDB

CouchDB

hadoop

mahout

D3.js

<canvas>

Processing.js

view the demo

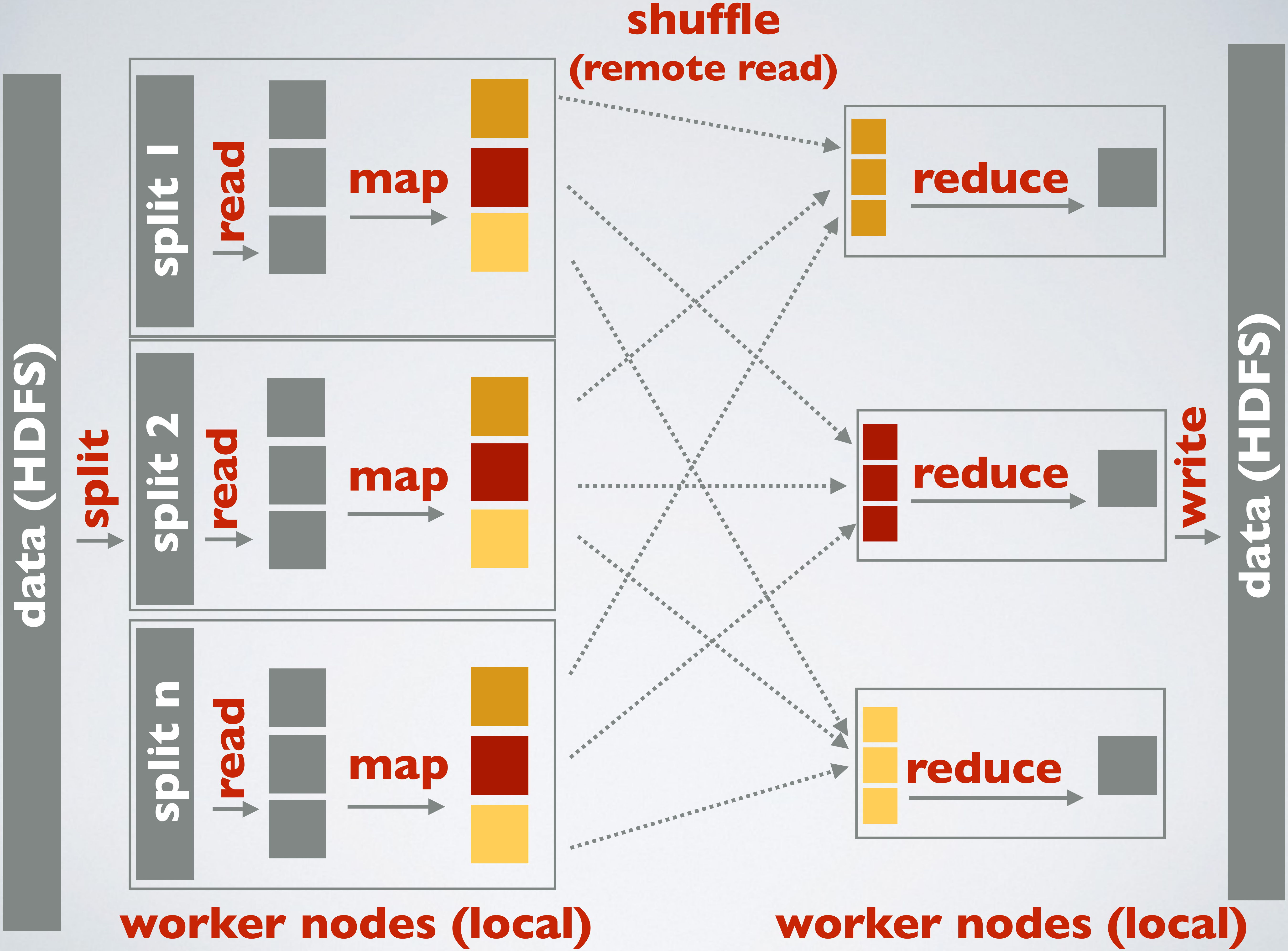
WEKA
The University
of Waikato

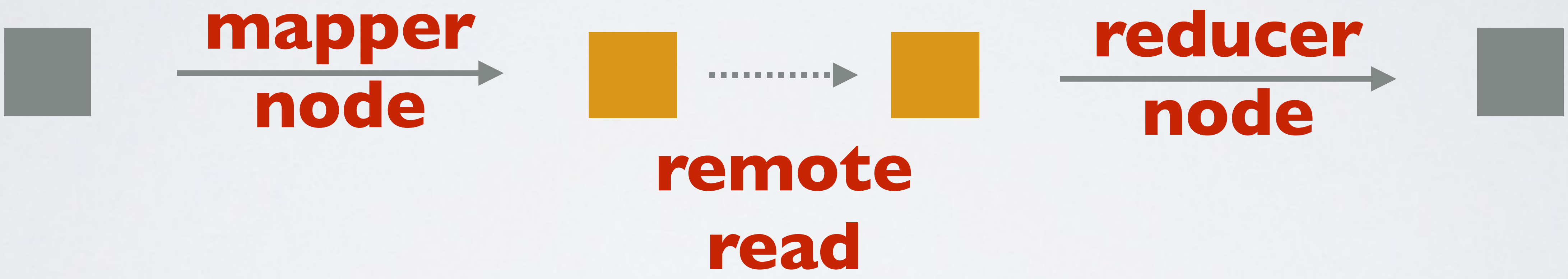
RAPID|MINER

KNIME

MapReduce

google.com/datacenters

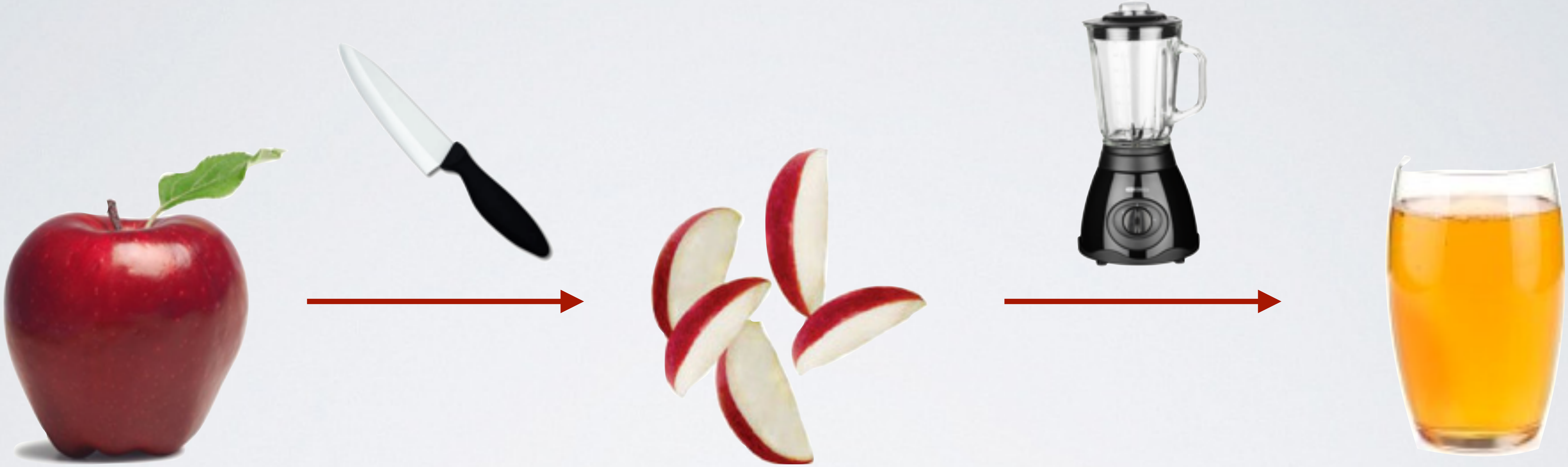

shuffle
(remote read)
data (HDFS)
split
split 1
read
map
split 2
read
map
split n
read
map
reduce
reduce
reduce
write
data (HDFS)
worker nodes (local)
worker nodes (local)

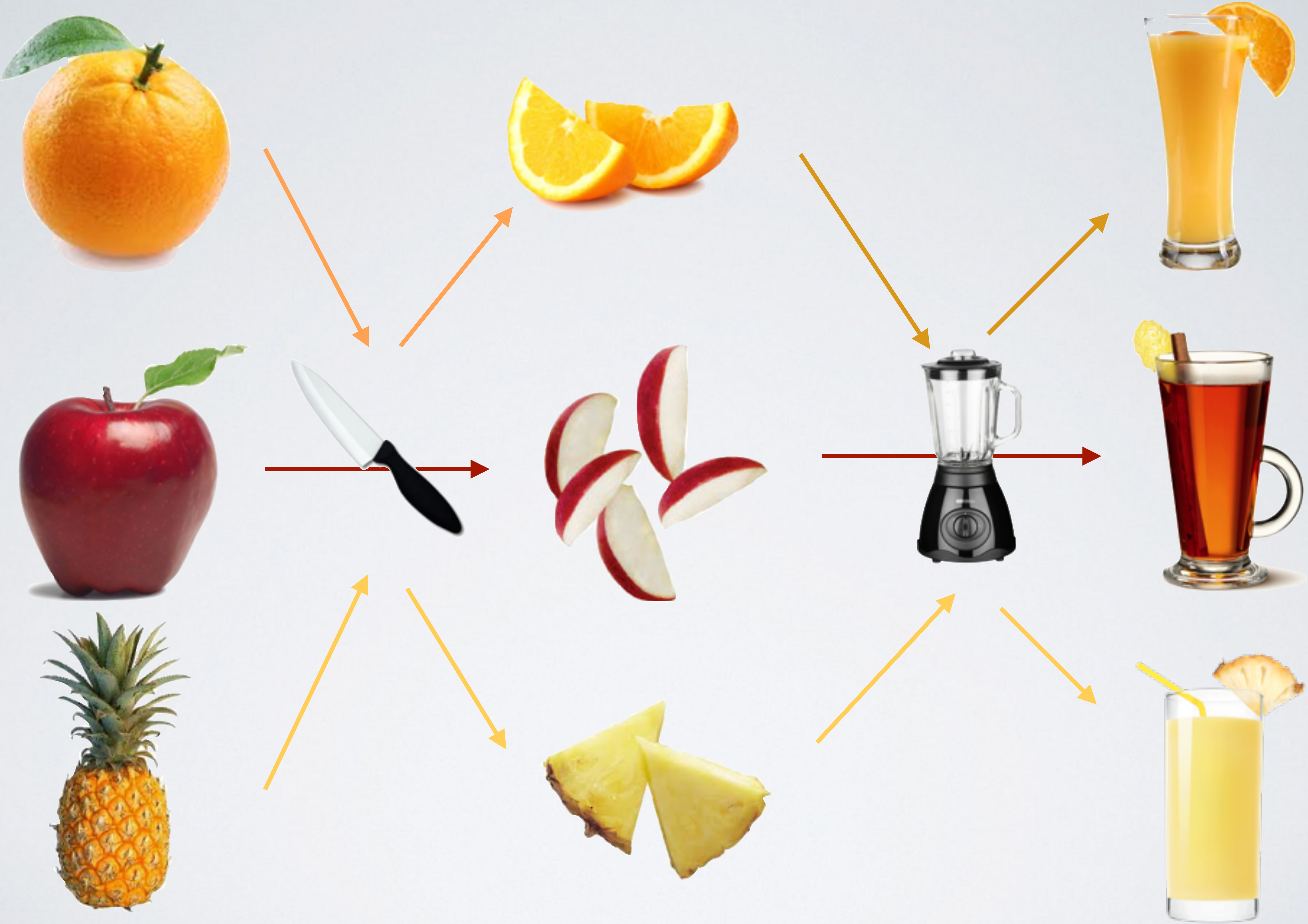


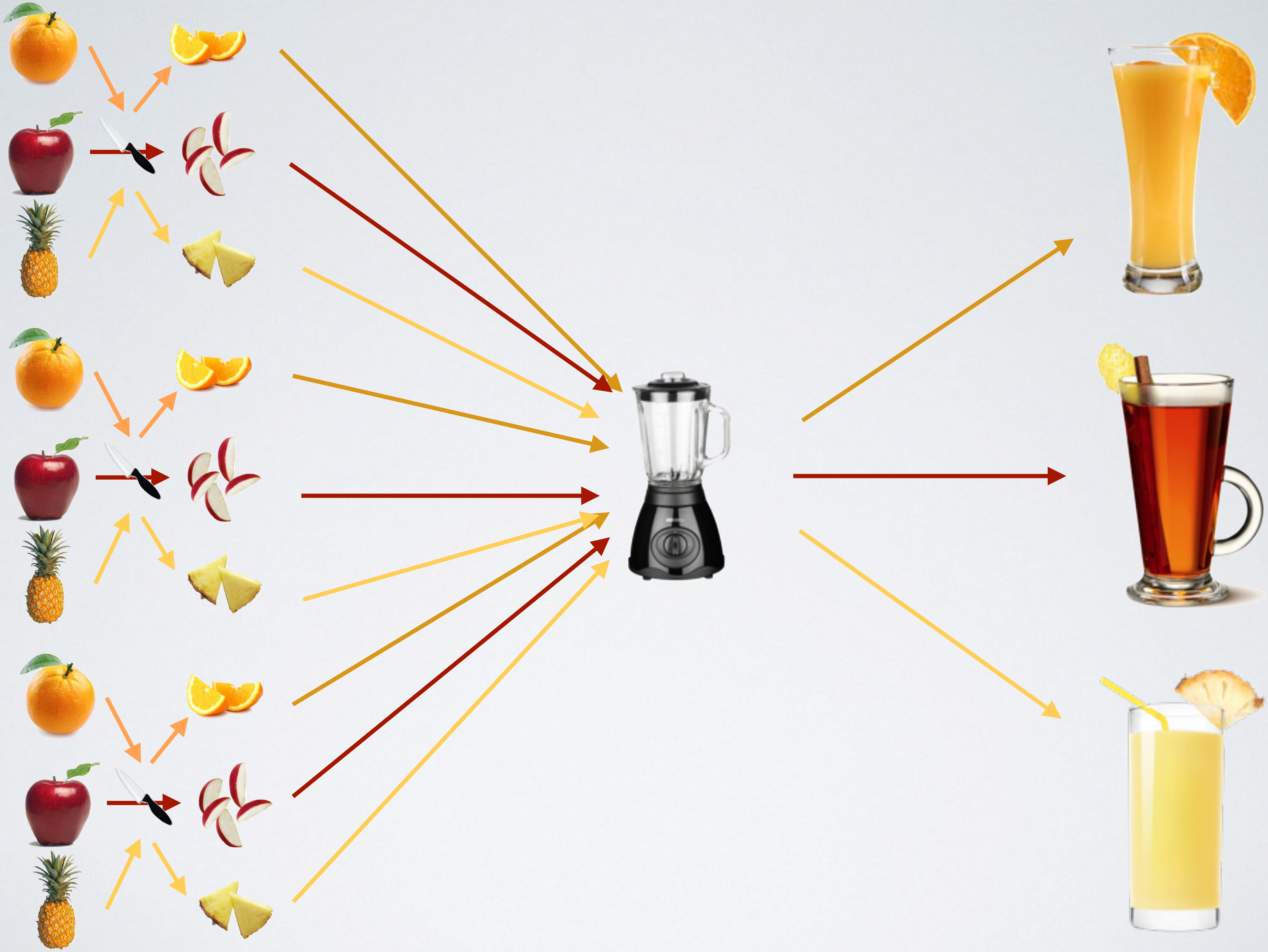
mapper node

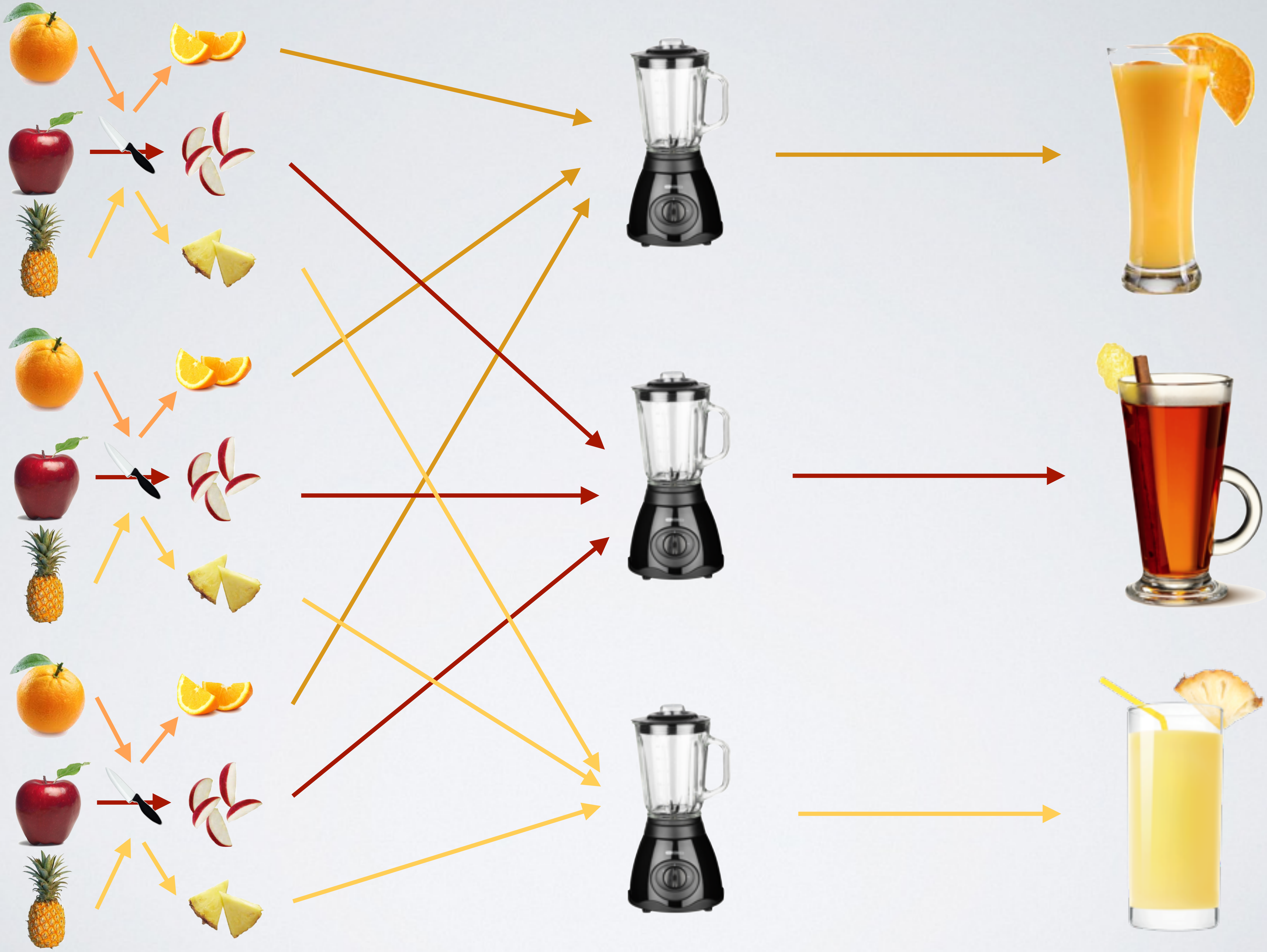
remote read

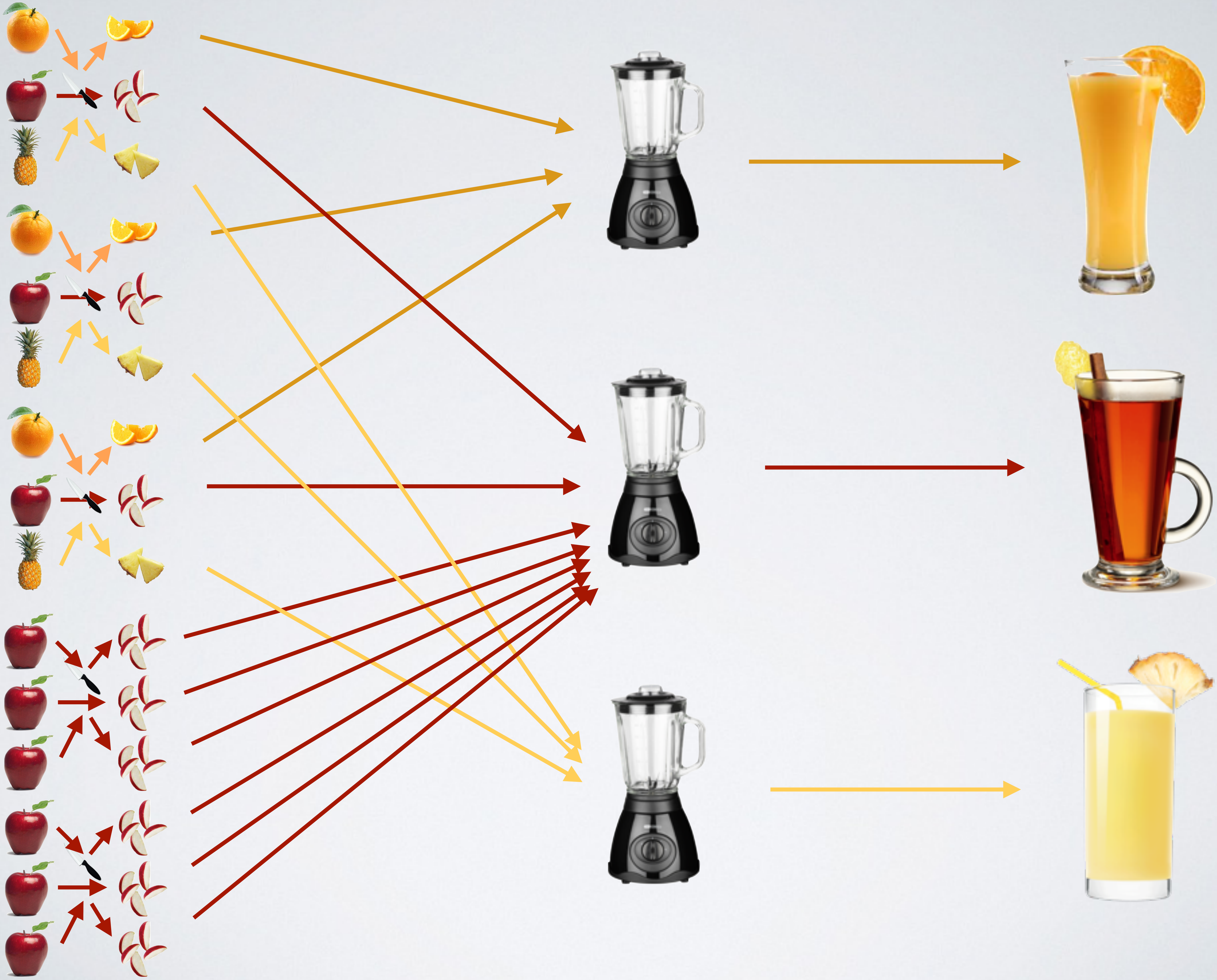
reducer node

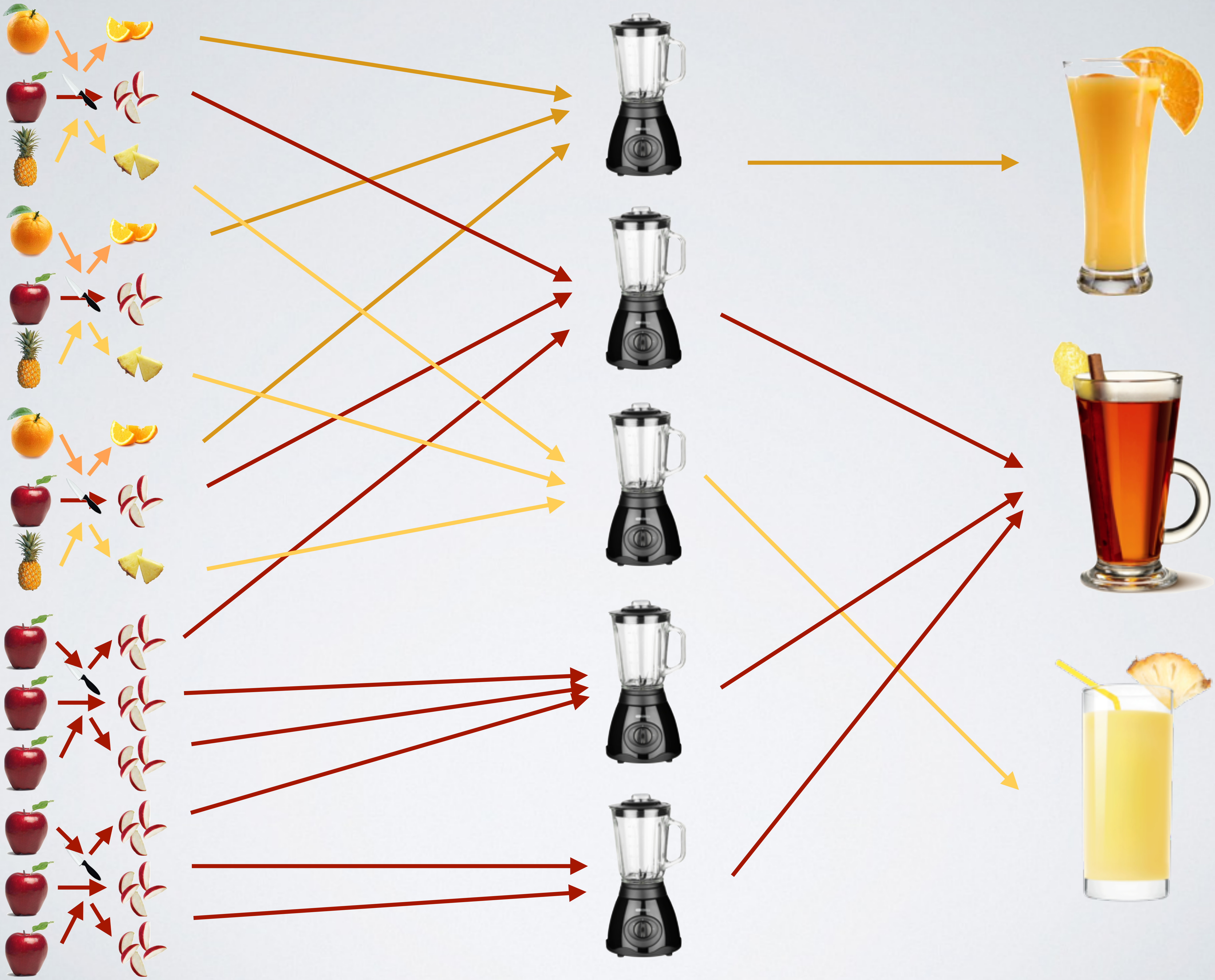




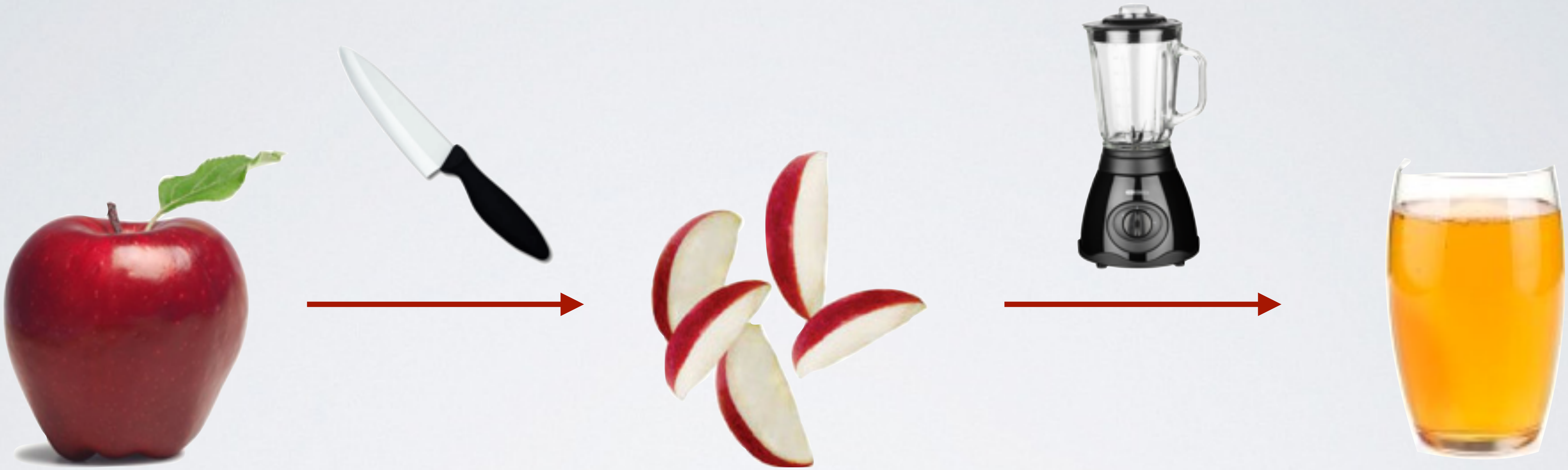








Mapper **Reducer**



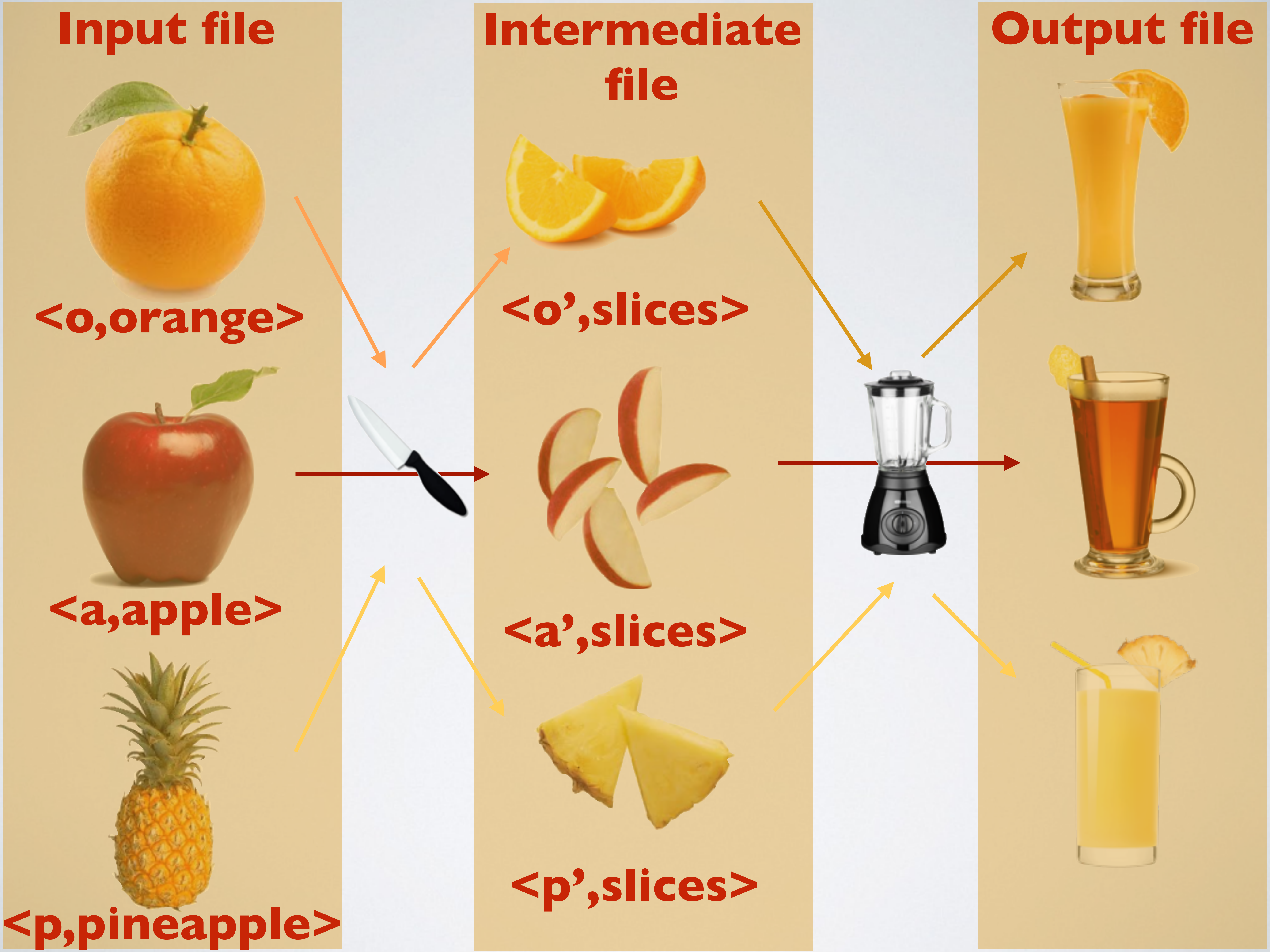
Input file

Intermediate file (local)

Output file

<a,apple>

<a’,slices>



Input file

<o,orange>

<a,apple>

<p,pineapple>

Intermediate file

<o',slices>

<a',slices>

<p',slices>

Output file

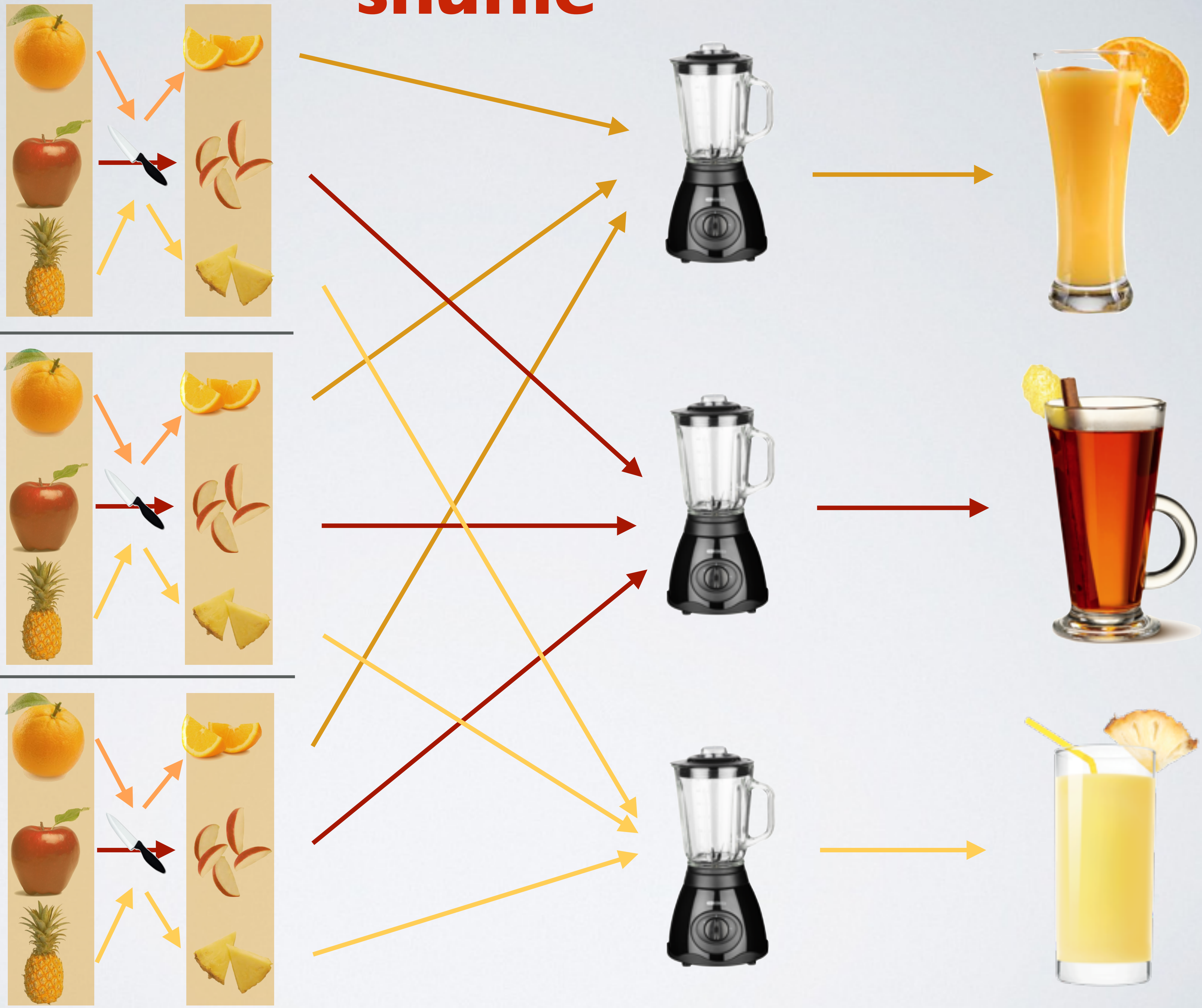
split

shuffle

split 0

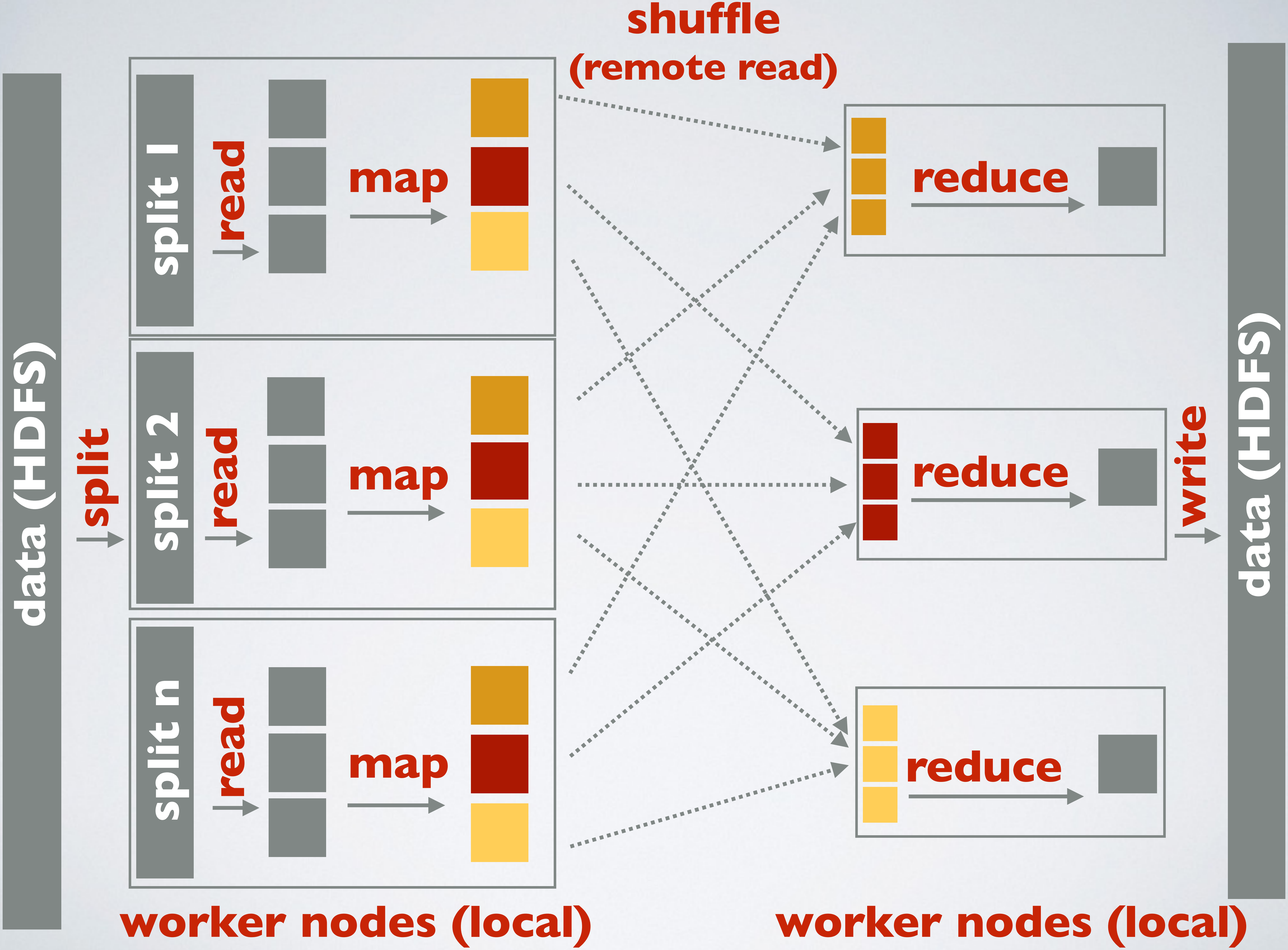
split 1

split 2



1 mapper/split

1 reducer/key(set)



shuffle
(remote read)

data (HDFS)

split

split 1

read

map

split 2

read

map

split n

read

map

reduce

reduce

reduce

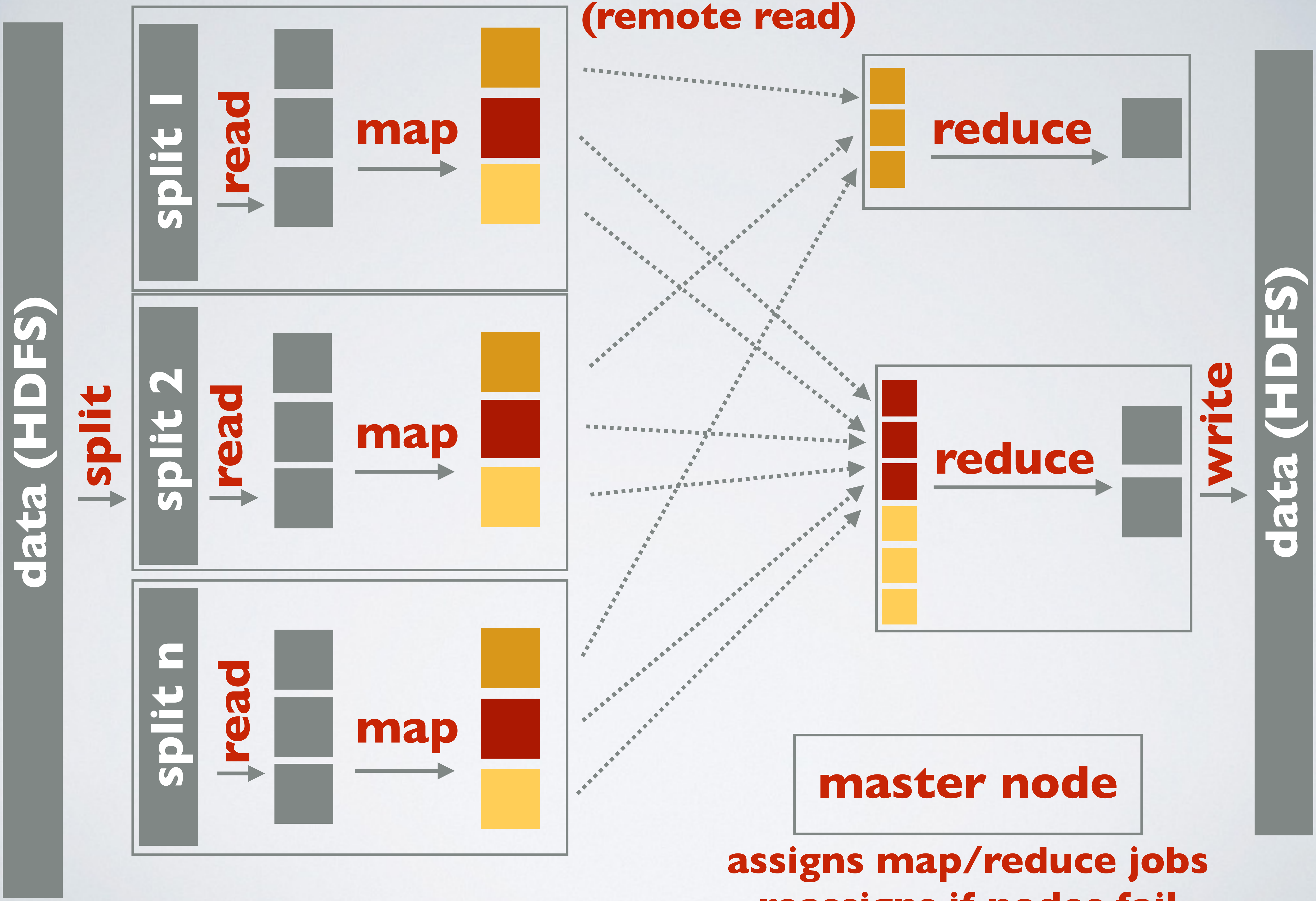
write

data (HDFS)

worker nodes (local)

worker nodes (local)

shuffle + parallel sort
(remote read)



data (HDFS)

split

split 1

read

map

split 2

read

map

split n

read

map

reduce

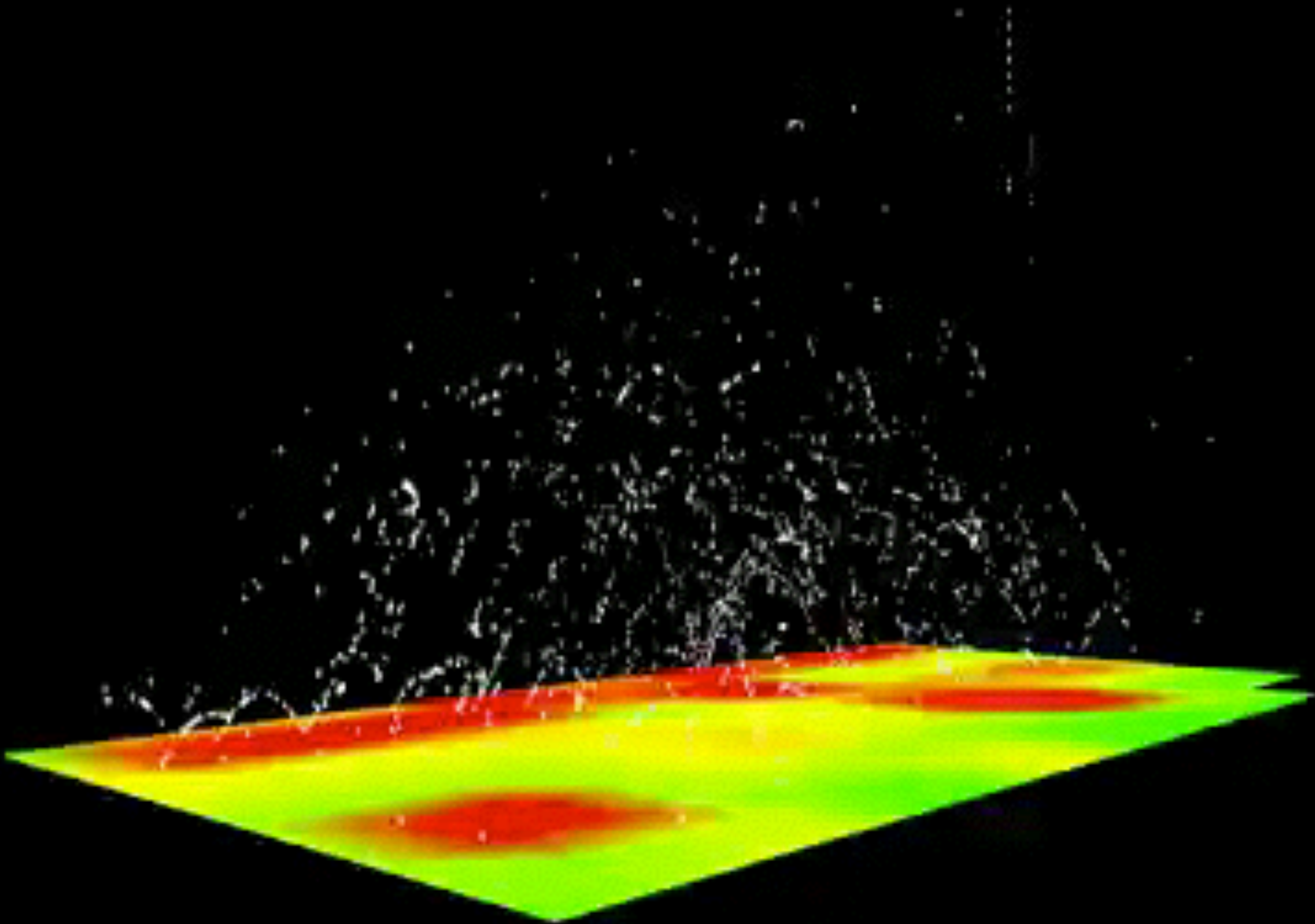
reduce

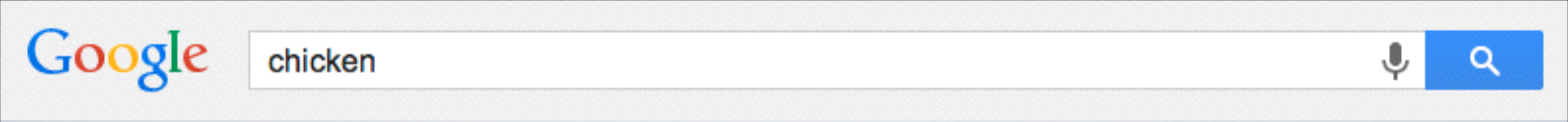
write

data (HDFS)

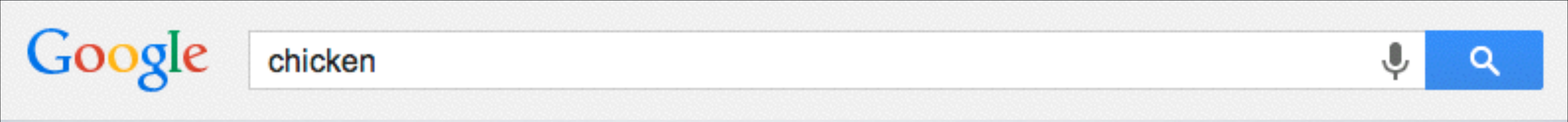
master node

assigns map/reduce jobs
reassigns if nodes fail





chicken



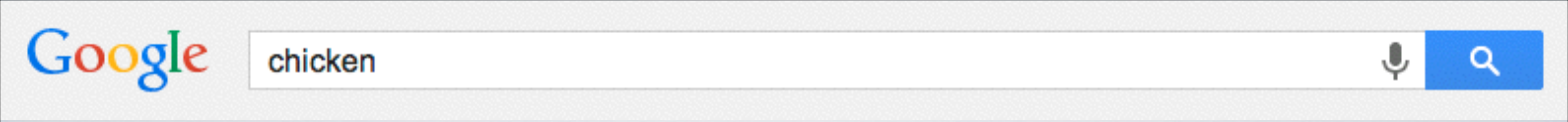
chicken

Map phase (3 parallel tasks)

- map$_1$ => (“why”,(doc$_1$,1)), (“did”,(doc$_1$,2)), (“the”,(doc$_1$,3)), (“chicken”,(doc$_1$,4)), (“cross”,(doc$_1$,5)), (“the”,(doc$_1$,6)), (“road”,(doc$_1$,7))
- map$_2$ => (“the”,(doc$_2$,1)), (“chicken”,(doc$_2$,2)), (“and”,(doc$_2$,3)), (“egg”,(doc$_2$,4)), (“problem”, (doc$_2$,5))
- map$_3$ => (“kentucky”,(doc$_3$,1)), (“fried”,(doc$_3$,2)), (“chicken”,(doc$_3$,3))

Intermediate shuffle & sort phase

- (“why”, <(doc$_1$,1)>),
- (“did”, <(doc$_1$,2)>),
- (“the”, <(doc$_1$,3), (doc$_1$,6), (doc$_2$,1)>)
- (“chicken”, <(doc$_1$,4), (doc$_2$,2), (doc$_3$,3)>)
- (“cross”, <(doc$_1$,5)>)
- (“road”, <(doc$_1$,7)>)
- (“and”, <(doc$_2$,3)>)
- (“egg”, <(doc$_2$,4)>)
- (“problem”, <(doc$_2$,5)>)
- (“kentucky”, <(doc$_3$,1)>)
- (“fried”, <(doc$_3$,2)>)



chicken

Intermediate shuffle & sort phase

- *("why", <(doc$_1$,1)>),*
- *("did", <(doc$_1$,2)>),*
- *("the", <(doc$_1$,3), (doc$_1$,6), (doc$_2$,1)>)*
- *("chicken", <(doc$_1$,4), (doc$_2$,2), (doc$_3$,3)>)*
- *("cross", <(doc$_1$,5)>)*
- *("road", <(doc$_1$,7)>)*
- *("and", <(doc$_2$,3)>)*
- *("egg", <(doc$_2$,4)>)*
- *("problem", <(doc$_2$,5)>)*
- *("kentucky", <(doc$_3$,1)>)*
- *("fried", <(doc$_3$,2)>)*

Reduce phase (11 parallel tasks)

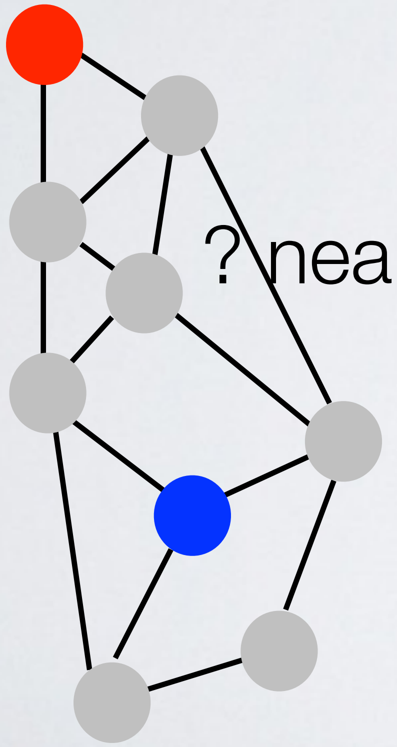
- *("why", <(doc$_1$,<1>)>),*
- *("did", <(doc$_1$,<2>)>),*
- *("the", <(doc$_1$, <3,6>), (doc$_2$, <1>)>)*
- *("chicken", <(doc$_1$,<4>), (doc$_2$,<2>), (doc$_3$,<3>)>)*
- *("cross", <(doc$_1$,<5>)>)*
- *("road", <(doc$_1$,<7>)>)*
- *("and", <(doc$_2$,<3>)>)*
- *("egg", <(doc$_2$,<4>)>)*
- *("problem", <(doc$_2$,<5>)>)*
- *("kentucky", <(doc$_3$,<1>)>)*
- *("fried", <(doc$_3$,<2>)>)*

Nearest bar

Input

graph

(node,label)



? nearest ● within distance *d*?

Nearest bar

Input

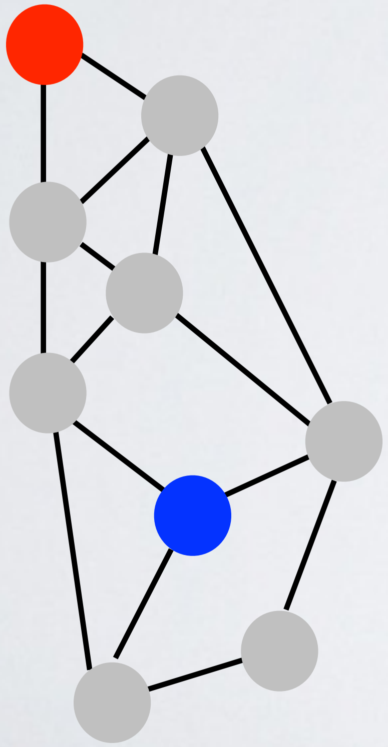
graph

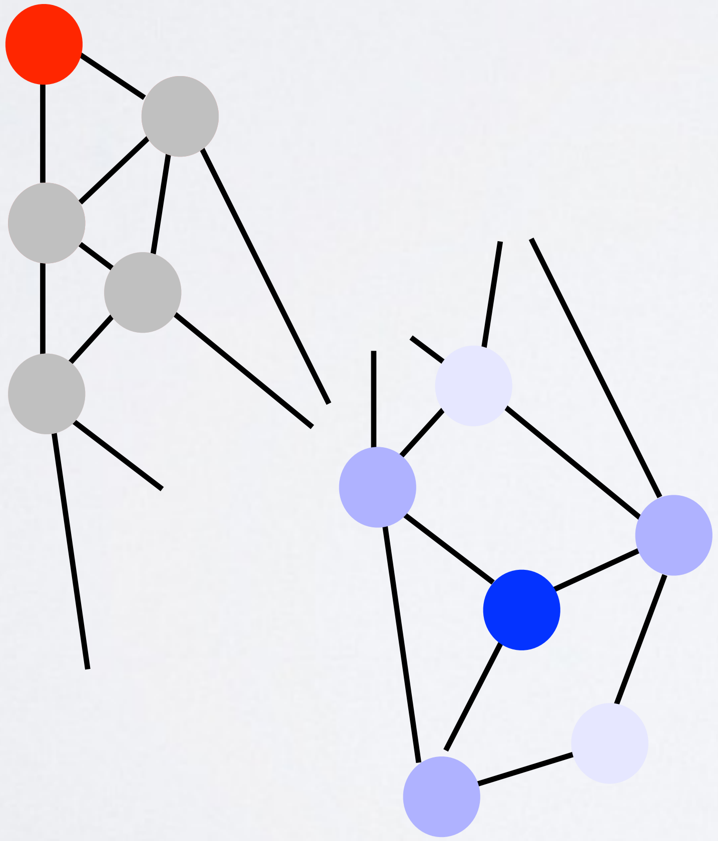
(node,label)

Map

∀●, search graph with radius *d*

< ● ,{● ,*distance*} >





Nearest bar

Input

graph

(node,label)

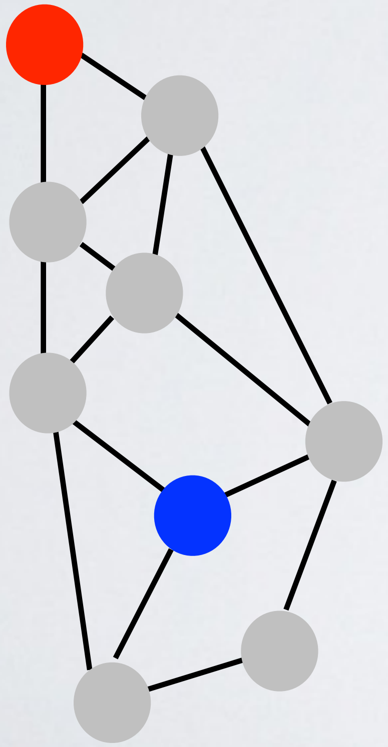
Map

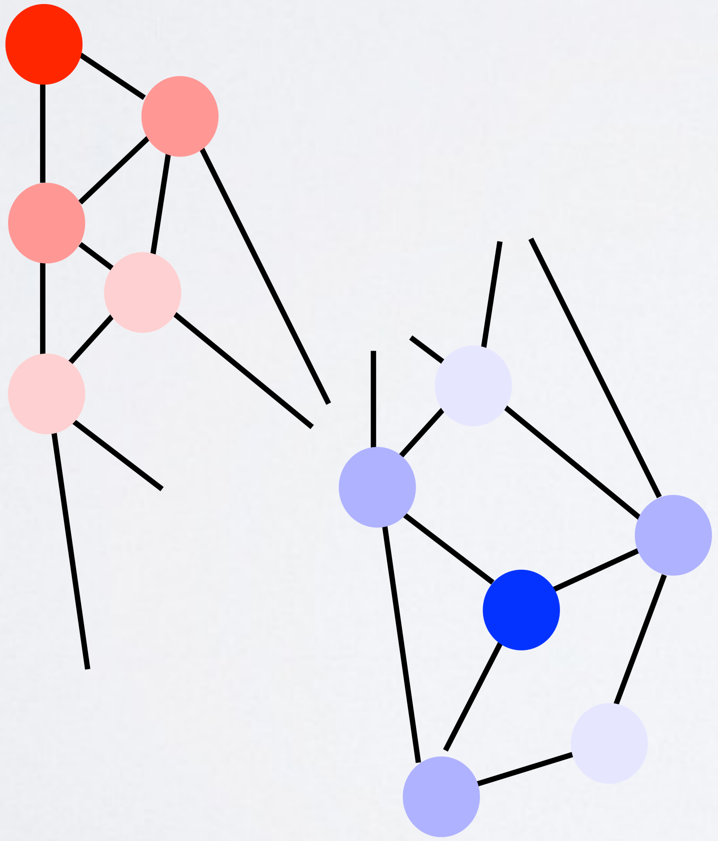
∀●, search graph

< ● ,{● ,*distance*} >

Shuffle/ Sort

by ● id





Nearest bar

Input

graph

(node,label)

Map

∀🔴, search graph

< ●, {🔴 ,*distance*} >

Shuffle/ Sort

by ● id

Reduce

< ● ,[{🔴,*distance*},

{🔵,*distance*}] >

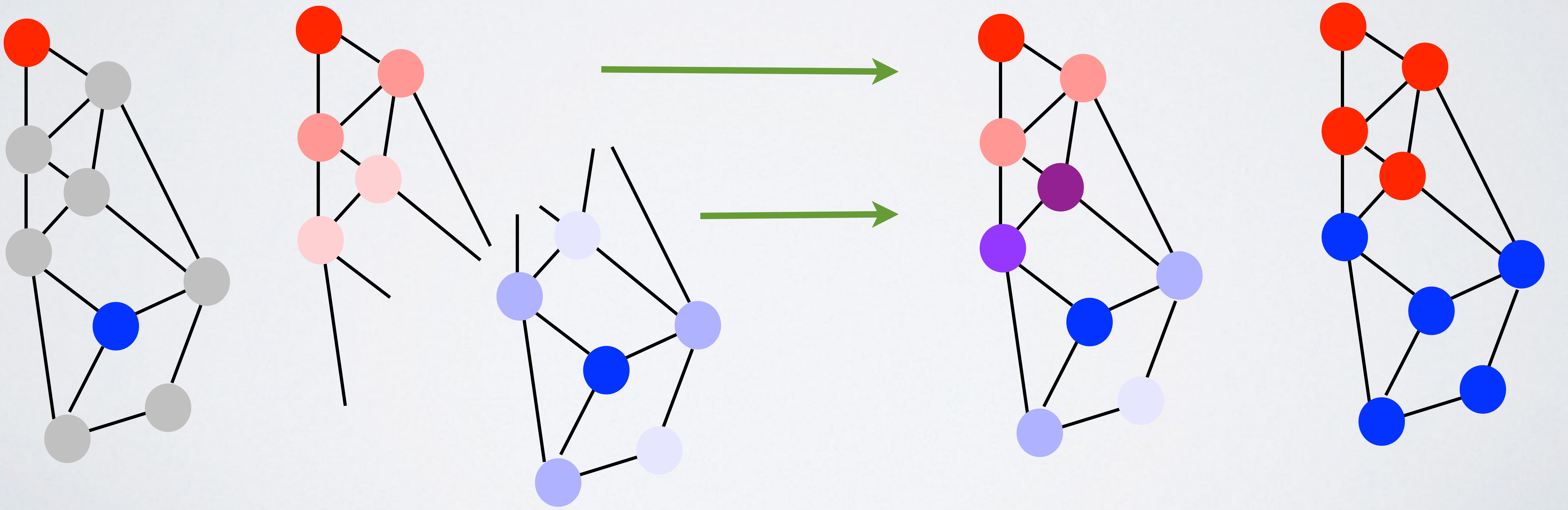
-> min()

Output

< ● , 🔴 >

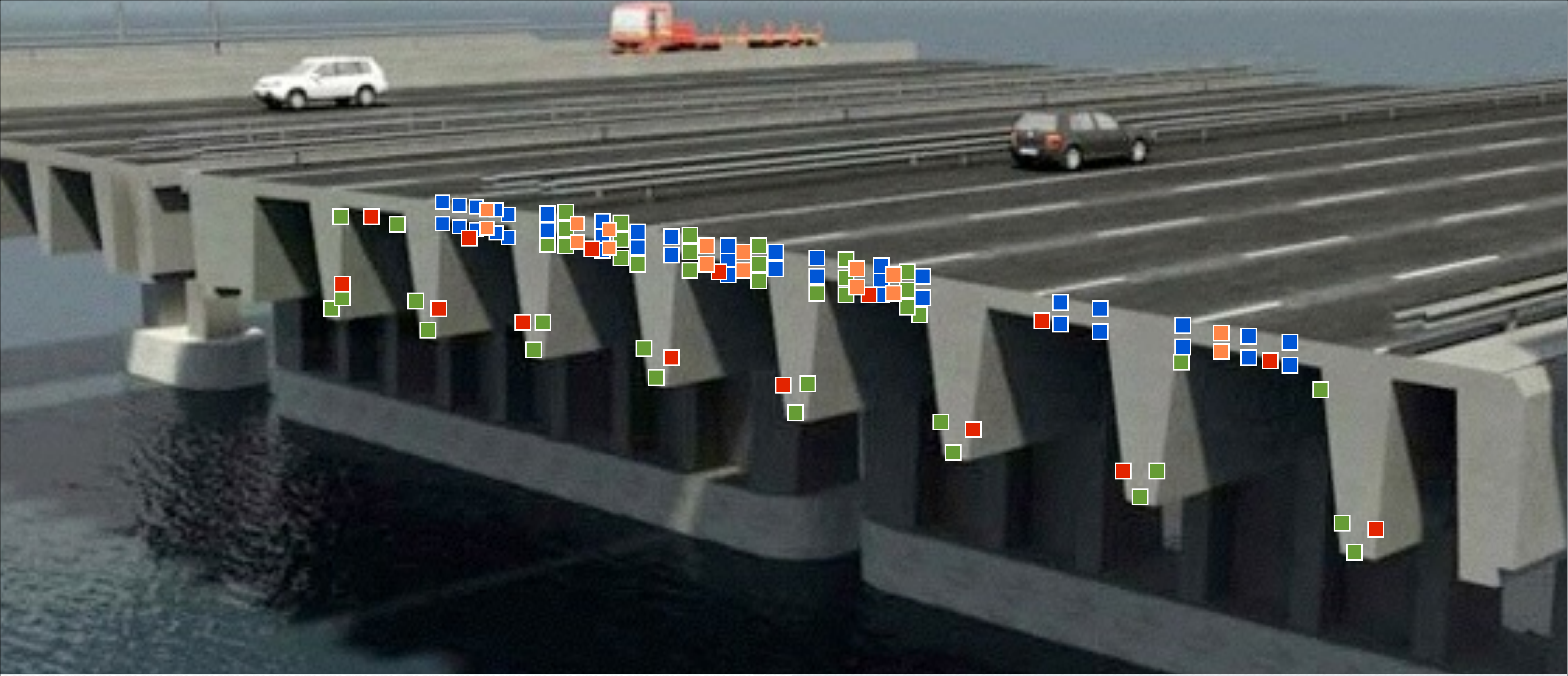
< ● , 🔵 >

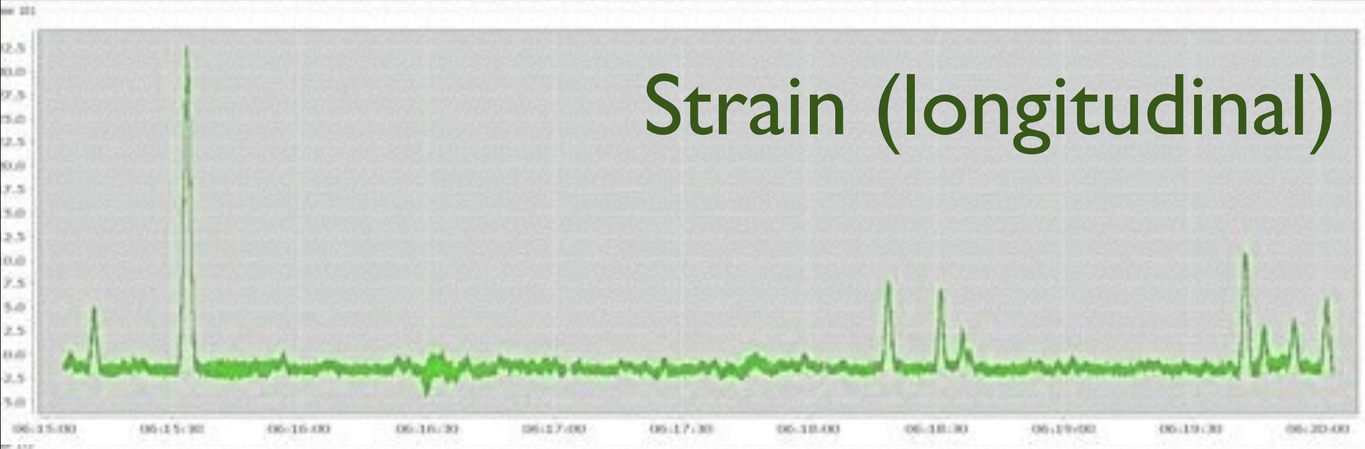
marked graph



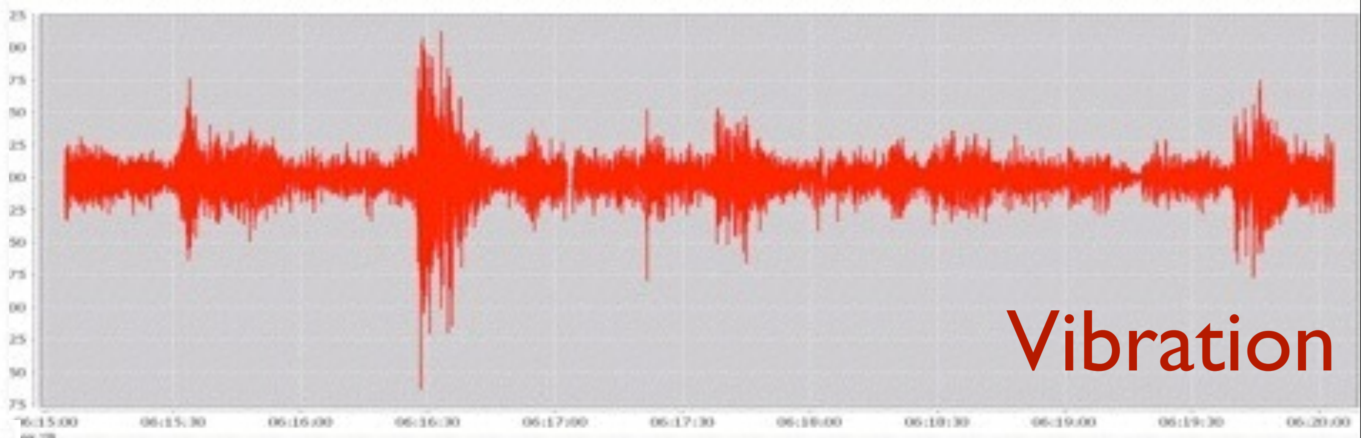
EXAMPLES

Sensor data

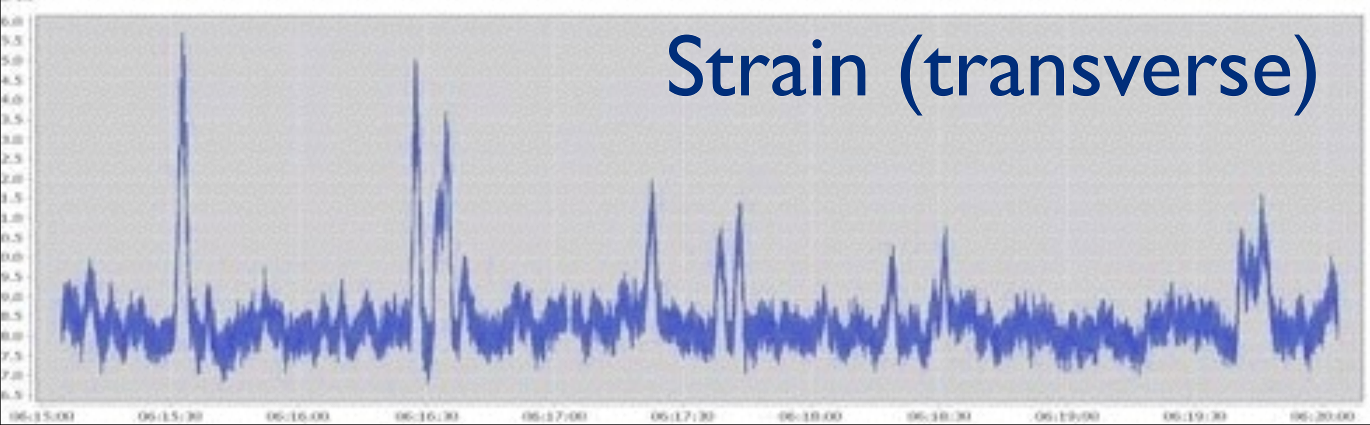




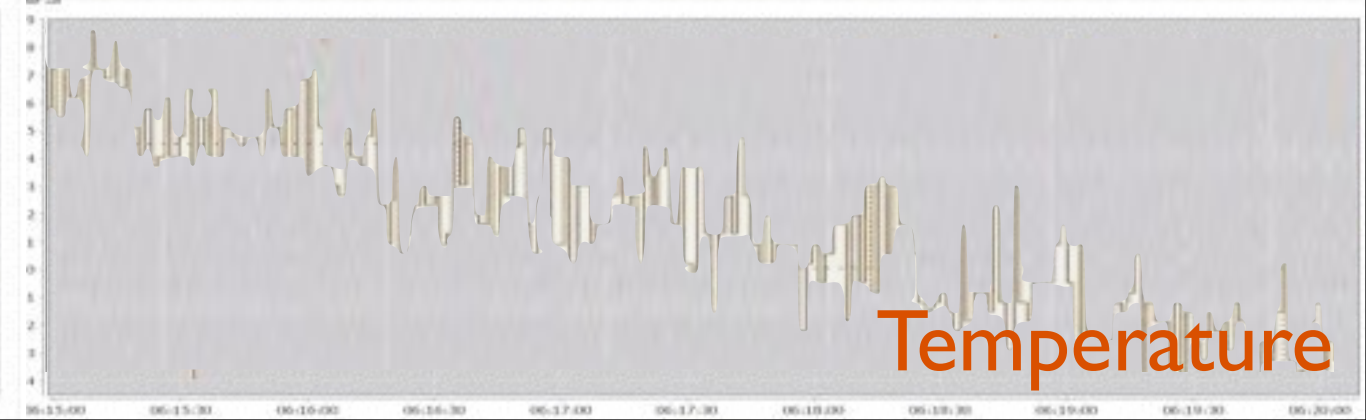
Strain (longitudinal)



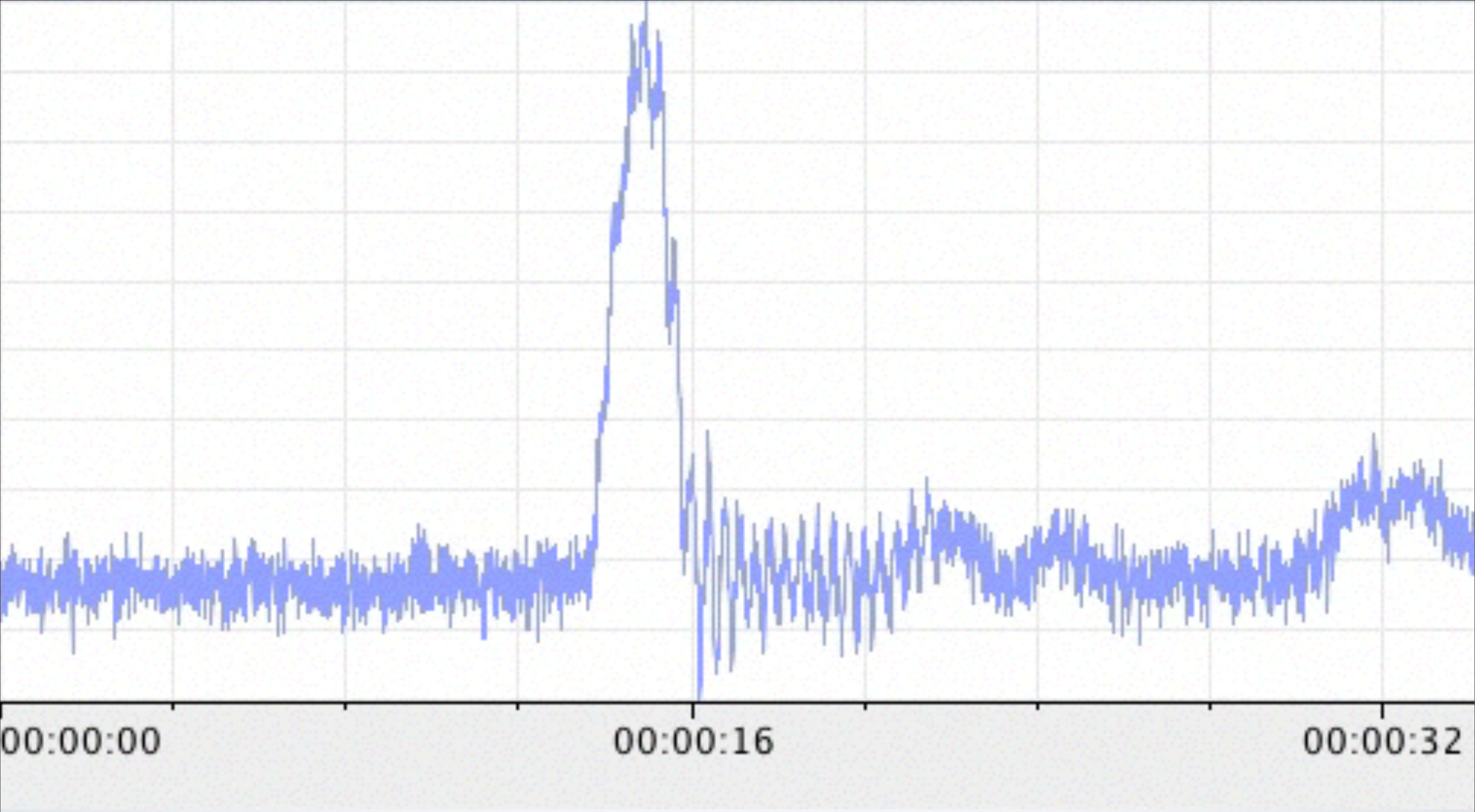
Vibration



Strain (transverse)



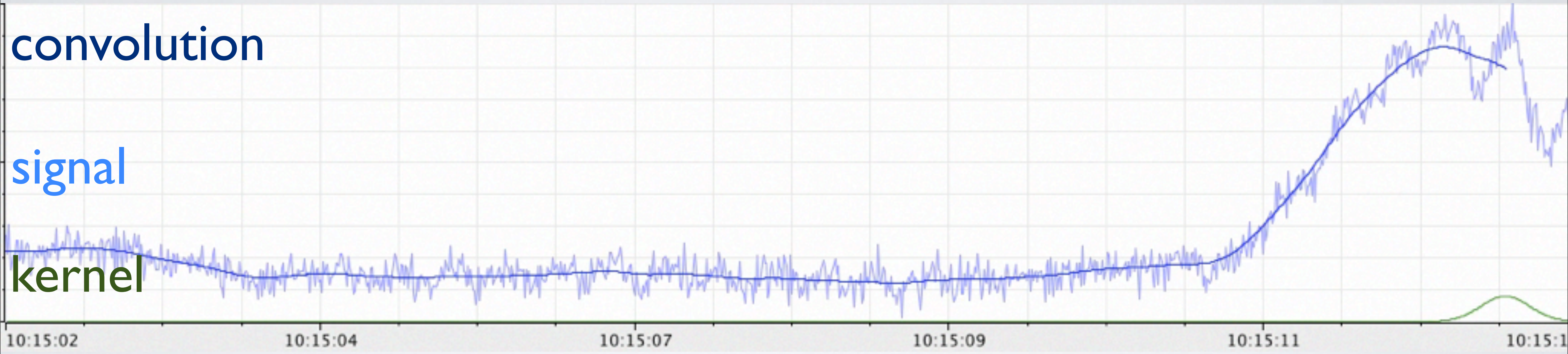
Temperature



NOISE

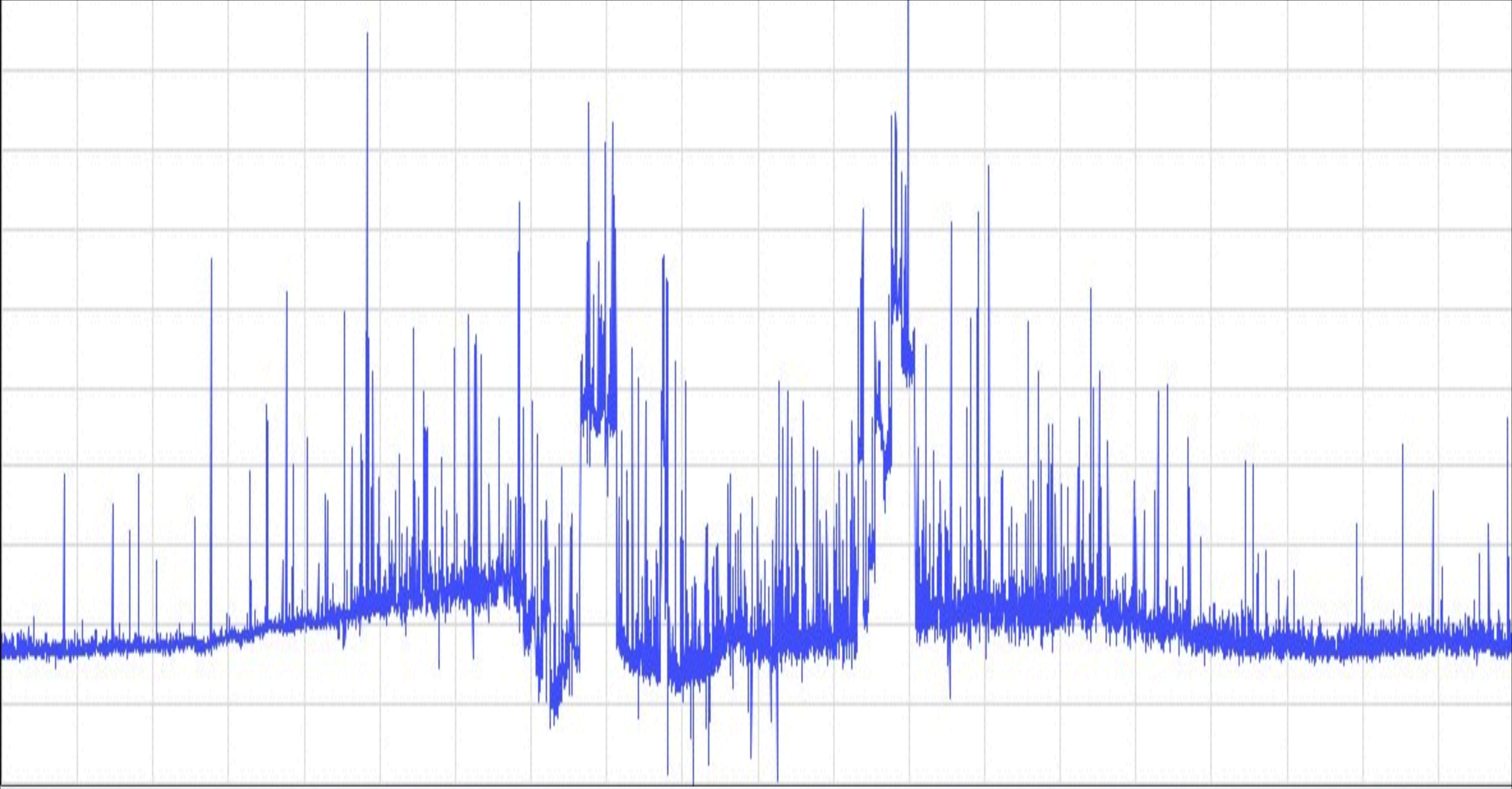
$$g * h \equiv \int_{-\infty}^{\infty} g(\tau)h(t-\tau)\, d\tau$$

$$(r * s)_j \equiv \sum_{k=-M/2+1}^{M/2} s_{j-k}\, r_k$$

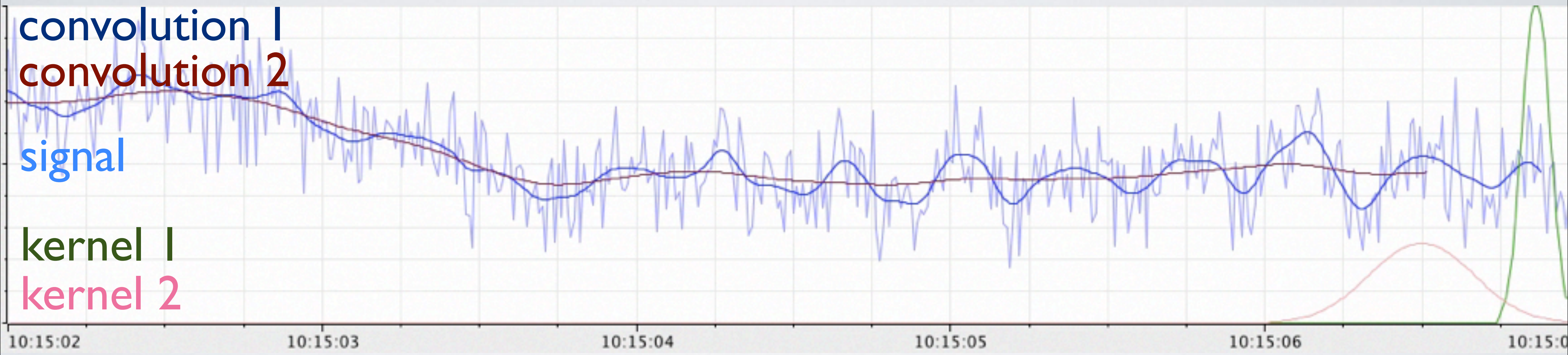


MATHS

Convolution



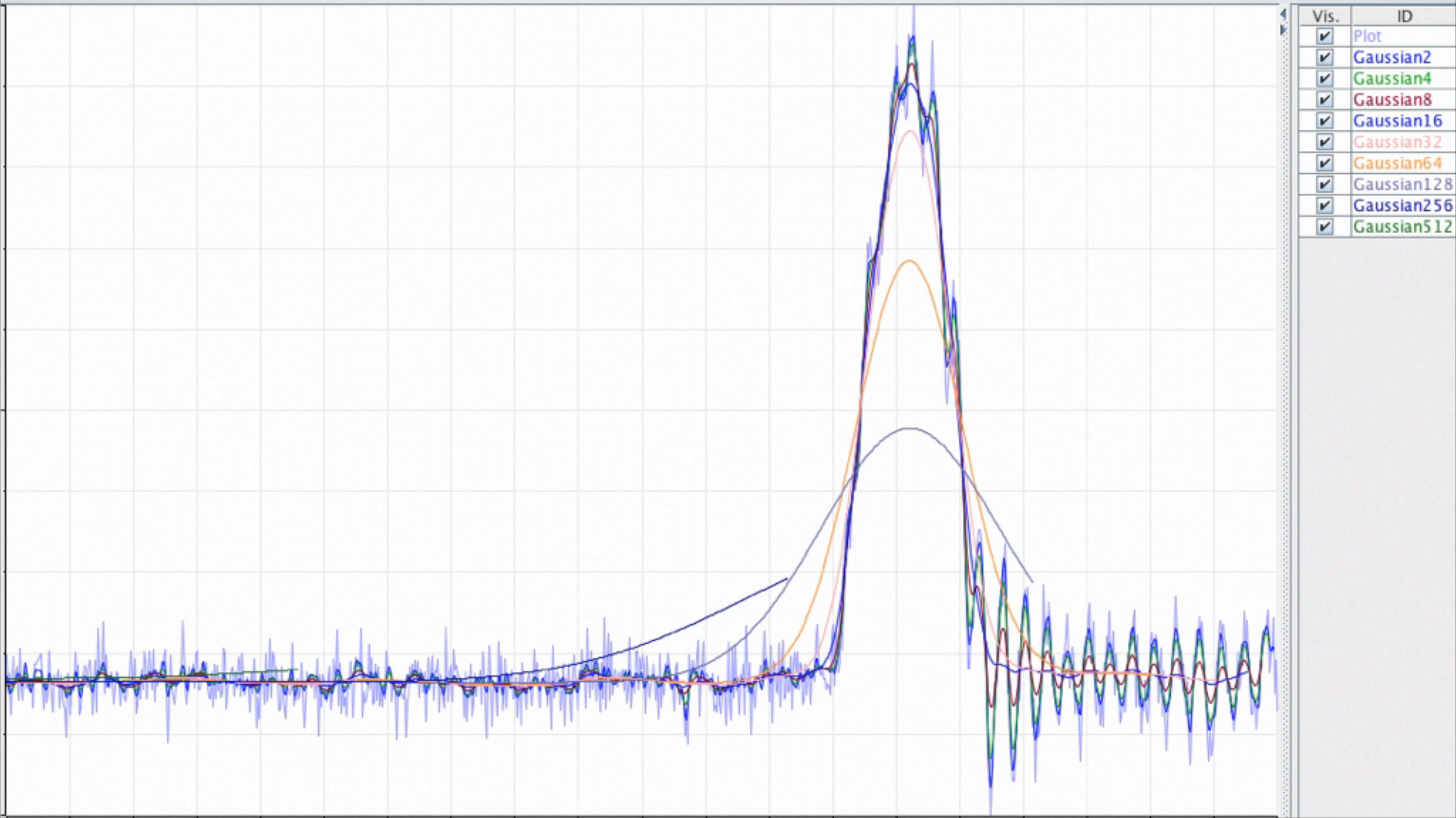
COMPLEXITY



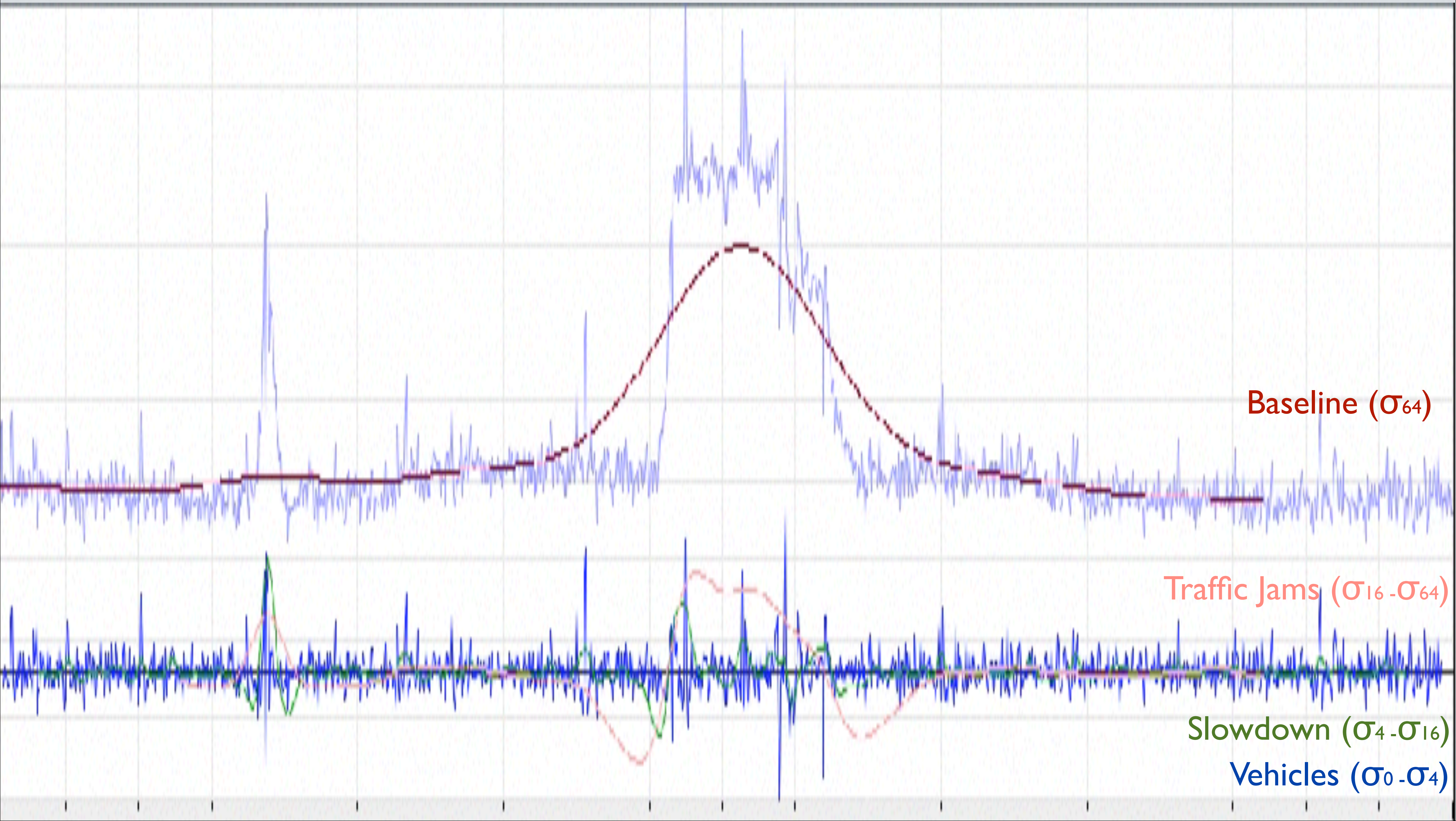
MATHS, AGAIN

scale space decomposition

SCALE-SPACE


Vis.
ID
✓ Plot
✓ Gaussian2
✓ Gaussian4
✓ Gaussian8
✓ Gaussian16
✓ Gaussian32
✓ Gaussian64
✓ Gaussian128
✓ Gaussian256
✓ Gaussian512

SCALE-SPACE DECOMPOSITION



145 sensors

100Hz

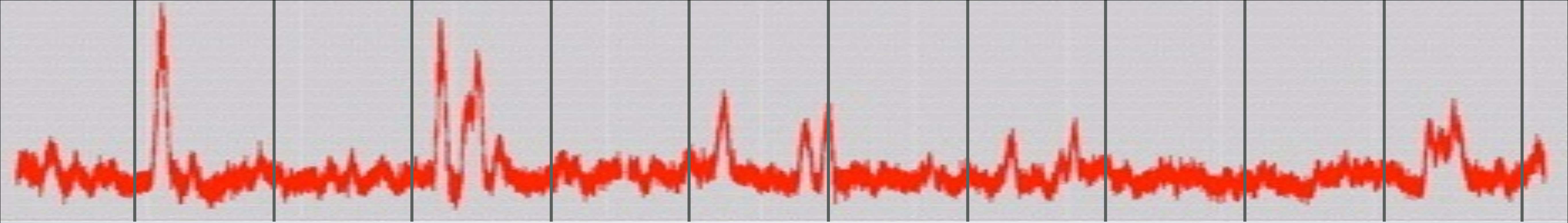
5GB/day

2TB/year

50MB/s disk I/O

VOLUME

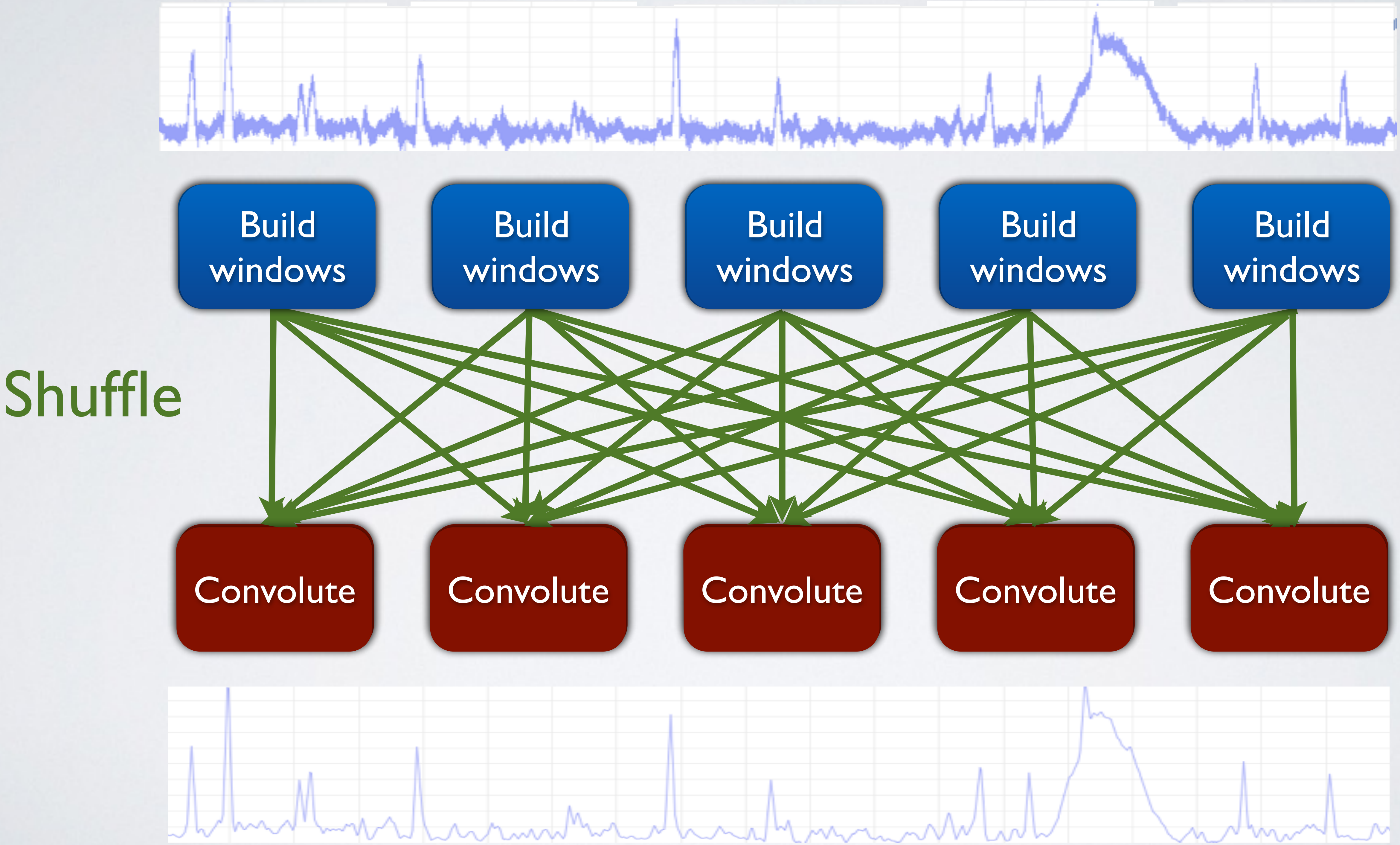
MapReduce



data: 2008-10-24 06:15:04.559, -6.293695, -1.1263204, 2.985364, 43449.957, 2.3577218, 38271.21
question: min, mean, max signal over all strain sensors?

```
  public void map(LongWritable key, Text value, Context context) {
          String values[] = value.toString().split("\t");
          Text time = new Text(values[0]);
          for(int i = 1; i <= nrStressSensors; i++)
             context.write(time, new Text(values[i]));
 }
  public void reduce(Text key, Iterable<Text> values, Context context) {
        //init; sum, min, max, count = 0
        Double d;
        for (Text v : values) {
             d = Double.valueOf(v.toString());
          sum += d;
          min = Math.min(min, d);
          max = Math.max(max, d);
          count++;
        }
        context.write(new Text(key+" min"), new Text(Double.toString((min))));
        context.write(new Text(key+" max"), new Text(Double.toString((max))));
        context.write(new Text(key+" avg"), new Text(Double.toString((sum/count))));
}
```

CONVOLUTION



Build windows

Build windows

Build windows

Build windows

Build windows

Shuffle

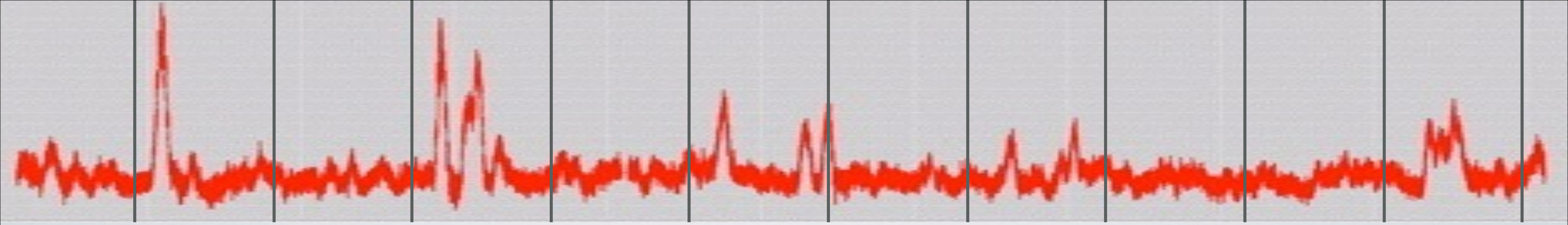
Convolute

Convolute

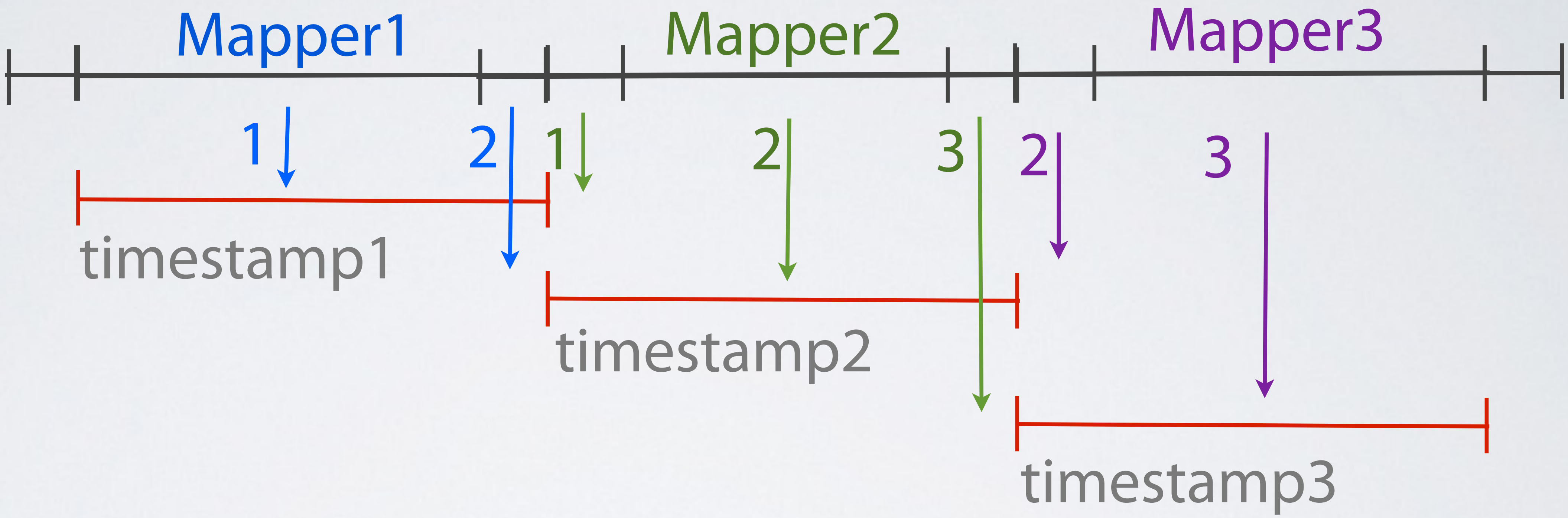
Convolute

Convolute

Convolute



Map
(window)



Mapper1

Mapper2

Mapper3

timestamp1

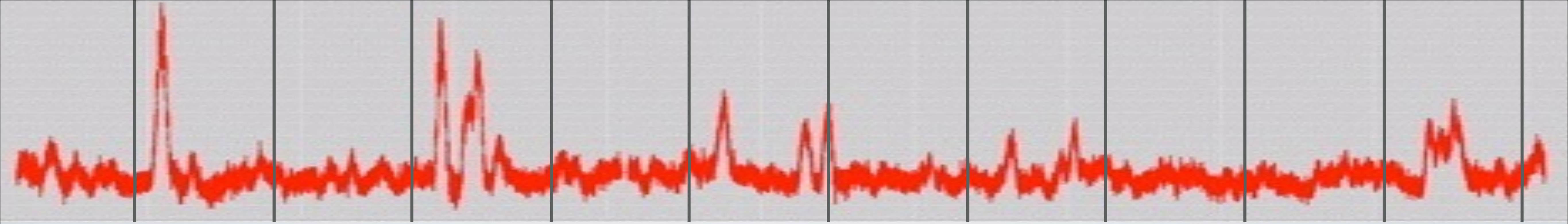
timestamp2

timestamp3

Reduce
(convolute)

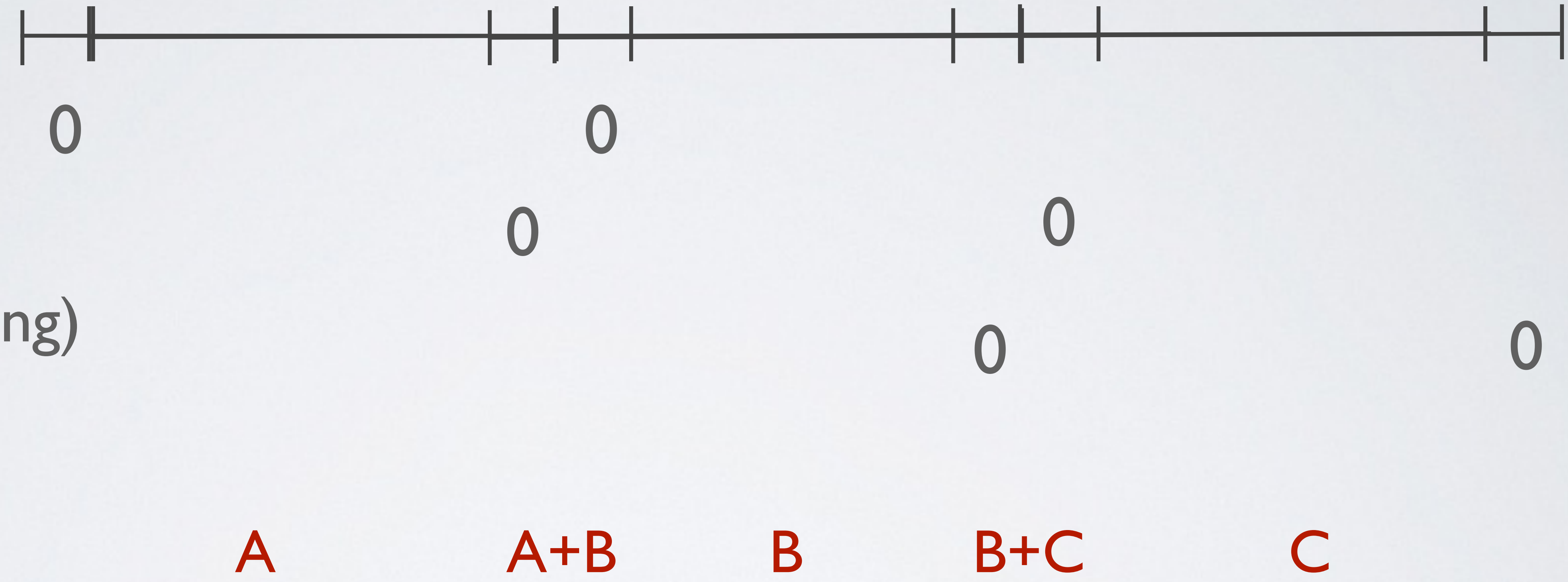
Emit only unpolluted data

OVERLAP-CONVOLUTE



Map
(convolute
with 0-padding)

Reduce
(add)

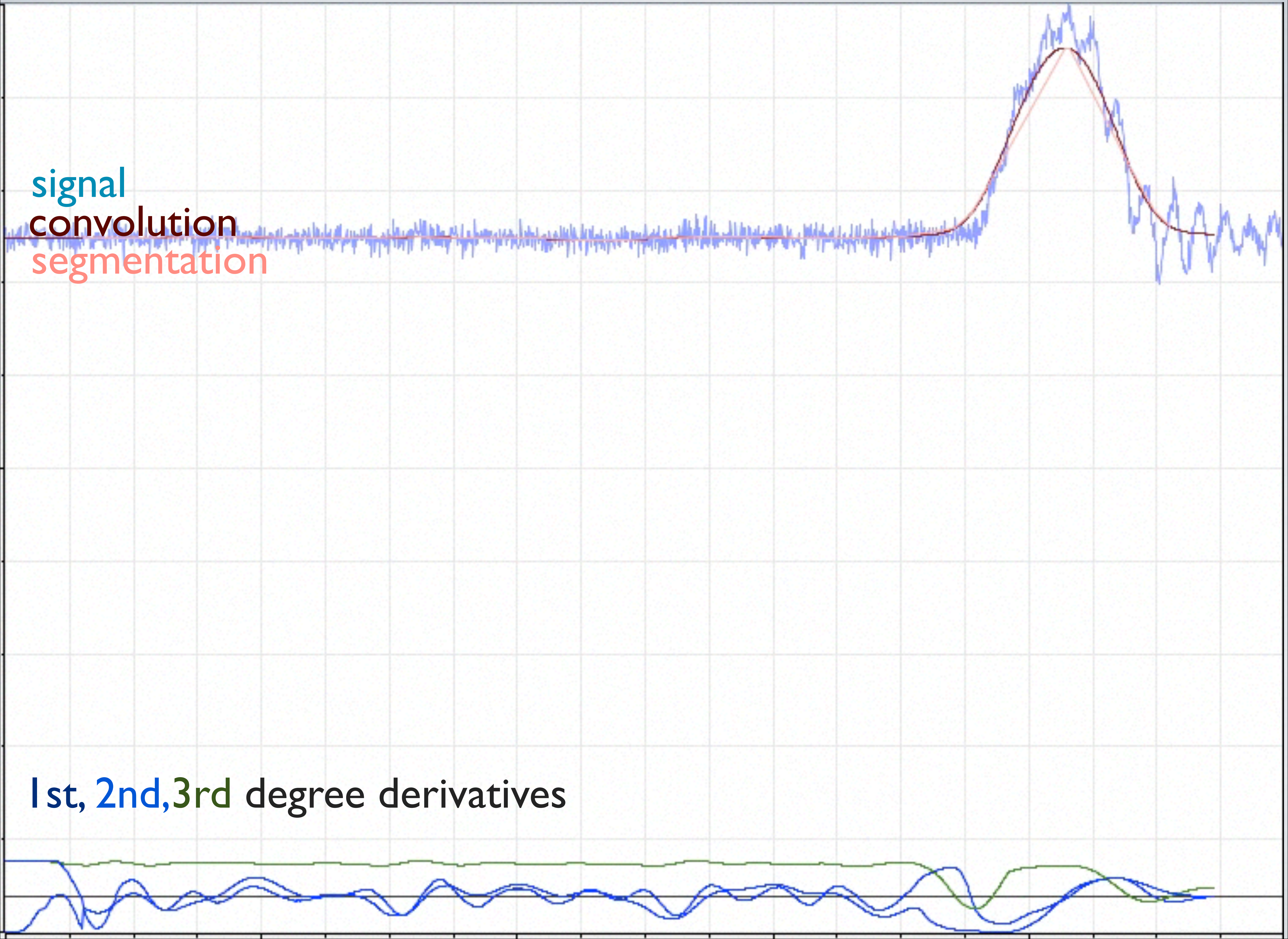


Add values in overlapping regions

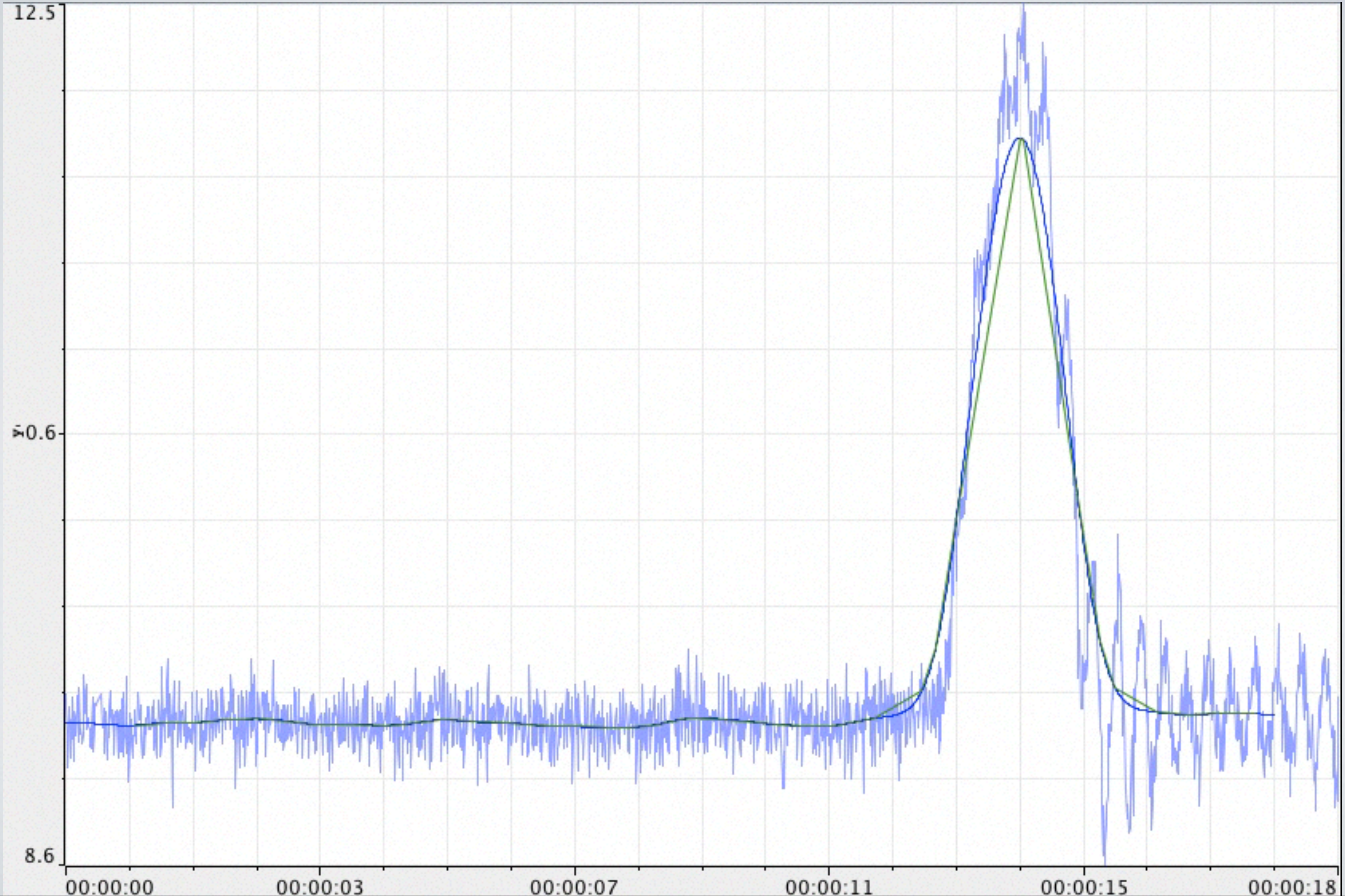
CONVOLUTE-ADD

SEGMENTATION

- You don’t need 100Hz data for everything
- Approximate signal with linear segments
- Key points: 0-crossings of 1st, 2nd, 3rd derivative
- Maths: derivative of smoothed signal = convolution with derivative of kernel

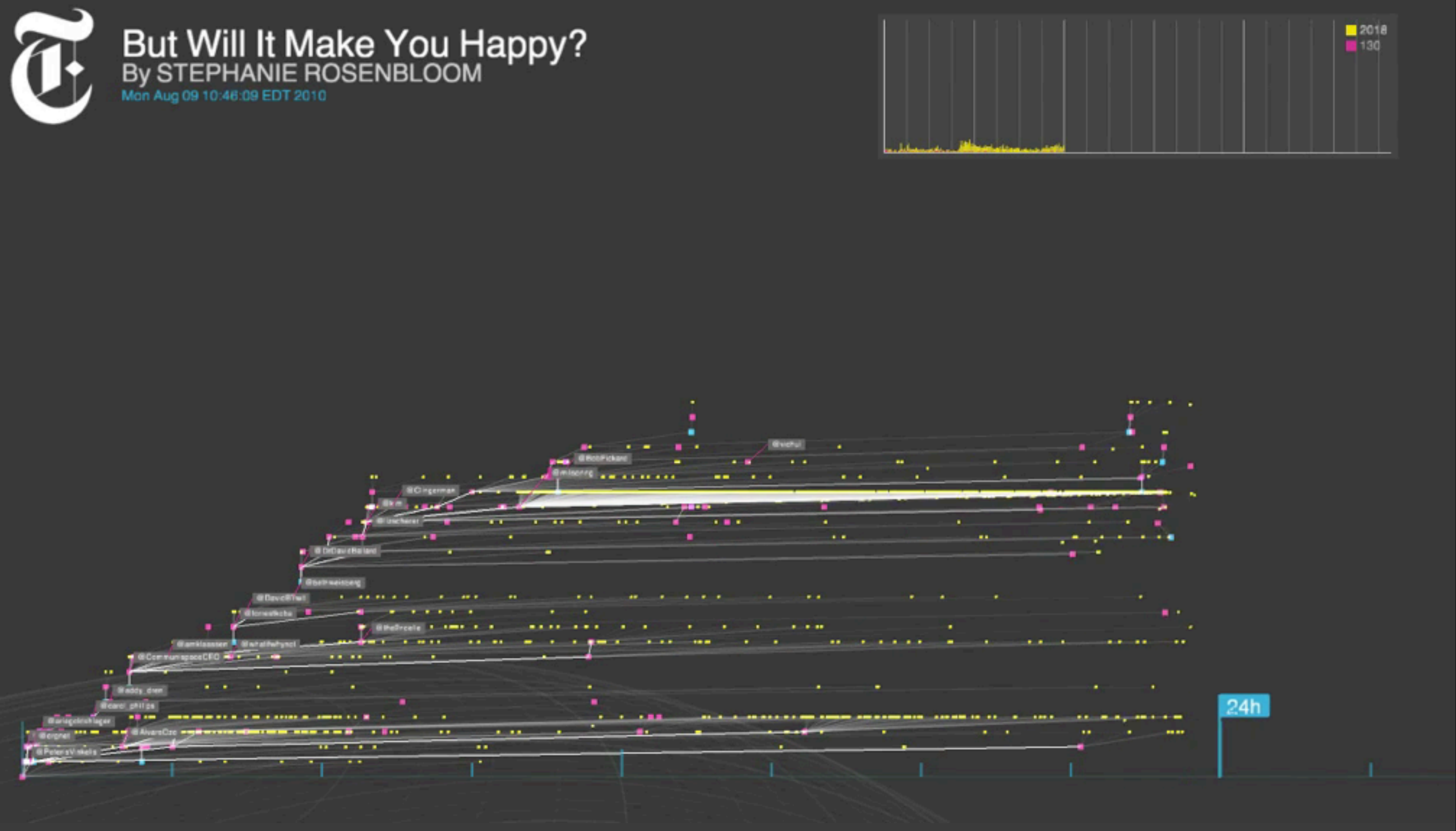

signal
convolution
segmentation
1st, 2nd,3rd degree derivatives

SEGMENTATION RESULT



Twitter data

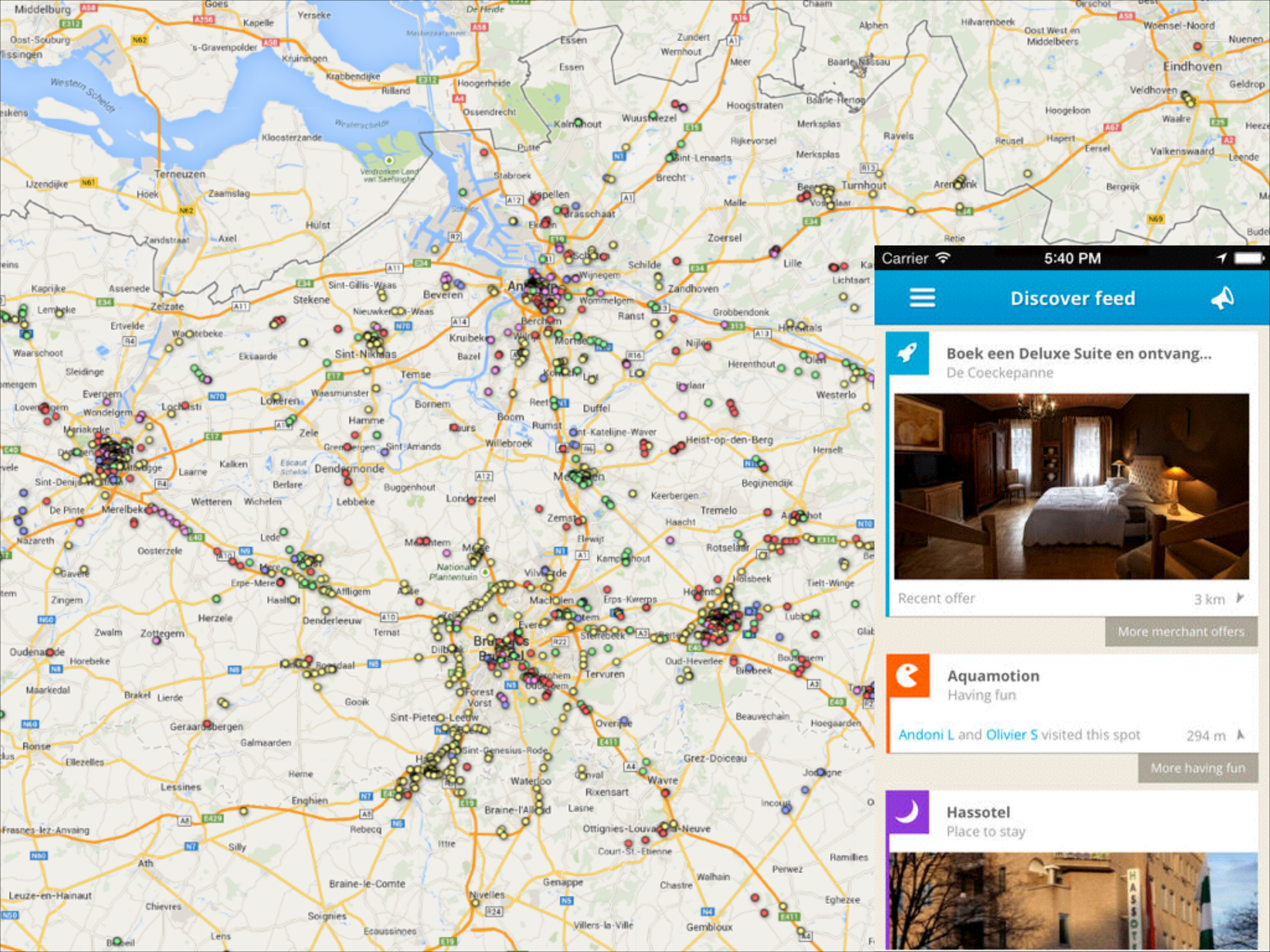
TRACKING NEWS STORIES



But Will It Make You Happy?

By STEPHANIE ROSENBLOOM

Mon Aug 09 10:46:09 EDT 2010



Carrier 5:40 PM

Discover feed

Boek een Deluxe Suite en ontvang...

De Coeckepanne

Recent offer 3 km

More merchant offers

Aquamotion

Having fun

Andoni L and Olivier S visited this spot 294 m

More having fun

Hassotel

Place to stay

6AM
New York City
Residence
Food
Arts & Entertainment
College & University
Nightlife Spot
Great Outdoors
Shop & Service
Professional & Other Places
Travel & Transport
foursquare

Geospacial data

KEEP
CALM
AND
ANALYZE
BIG DATA

OPEN DATA
OPEN SCIENCE

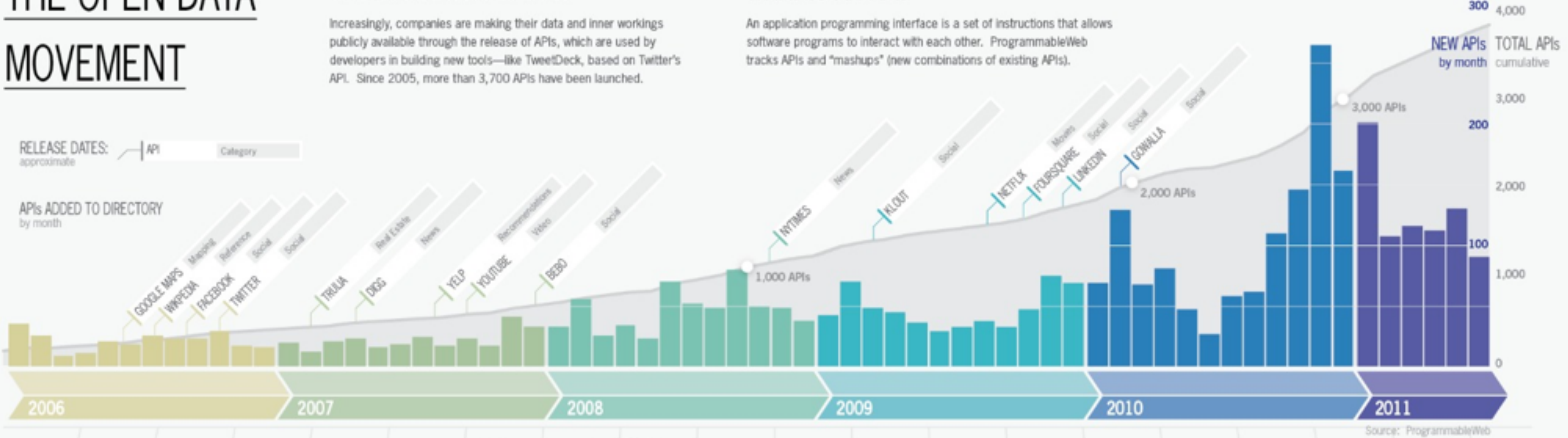
THE OPEN DATA MOVEMENT

THE EVOLUTION OF APIs

Increasingly, companies are making their data and inner workings publicly available through the release of APIs, which are used by developers in building new tools—like TweetDeck, based on Twitter's API. Since 2005, more than 3,700 APIs have been launched.

WHAT IS AN API?

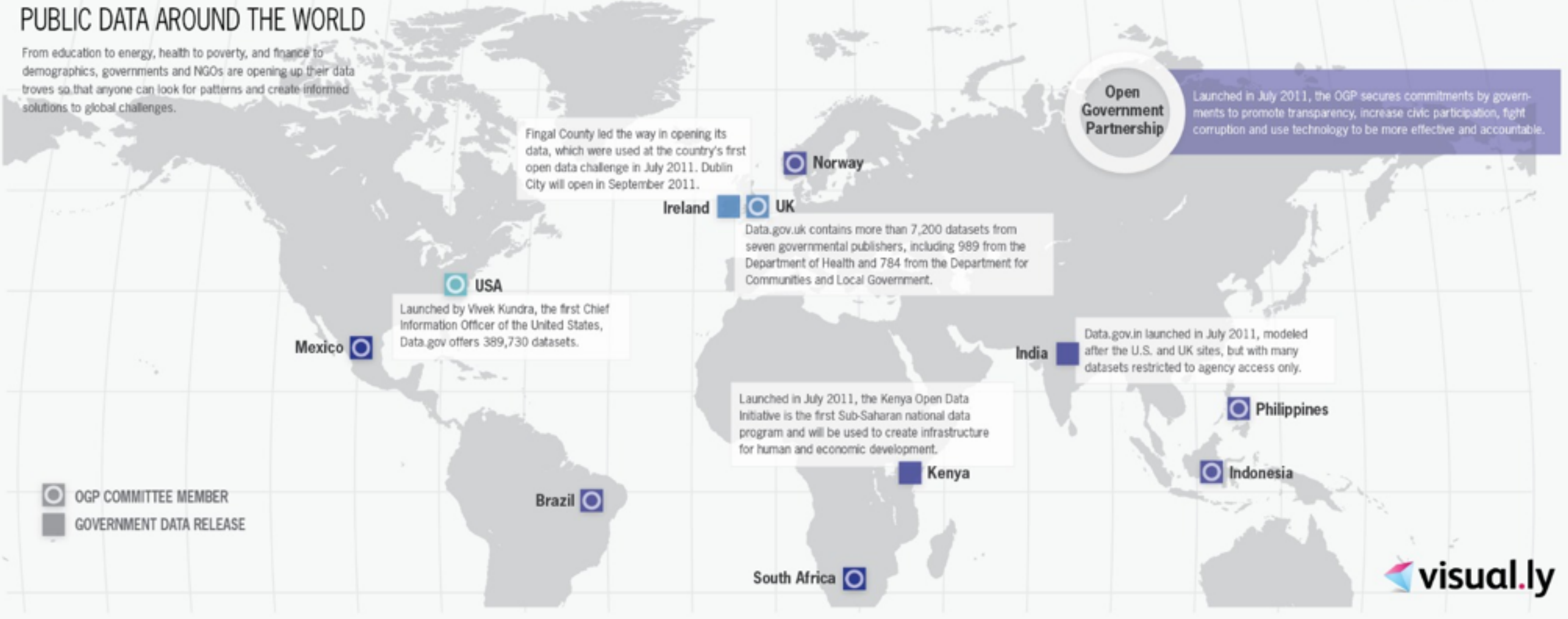
An application programming interface is a set of instructions that allows software programs to interact with each other. ProgrammableWeb tracks APIs and "mashups" (new combinations of existing APIs).



Source: ProgrammableWeb

PUBLIC DATA AROUND THE WORLD

From education to energy, health to poverty, and finance to demographics, governments and NGOs are opening up their data troves so that anyone can look for patterns and create informed solutions to global challenges.



Open Government Partnership

Launched in July 2011, the OGP secures commitments by governments to promote transparency, increase civic participation, fight corruption and use technology to be more effective and accountable.

Fingal County led the way in opening its data, which were used at the country's first open data challenge in July 2011. Dublin City will open in September 2011.

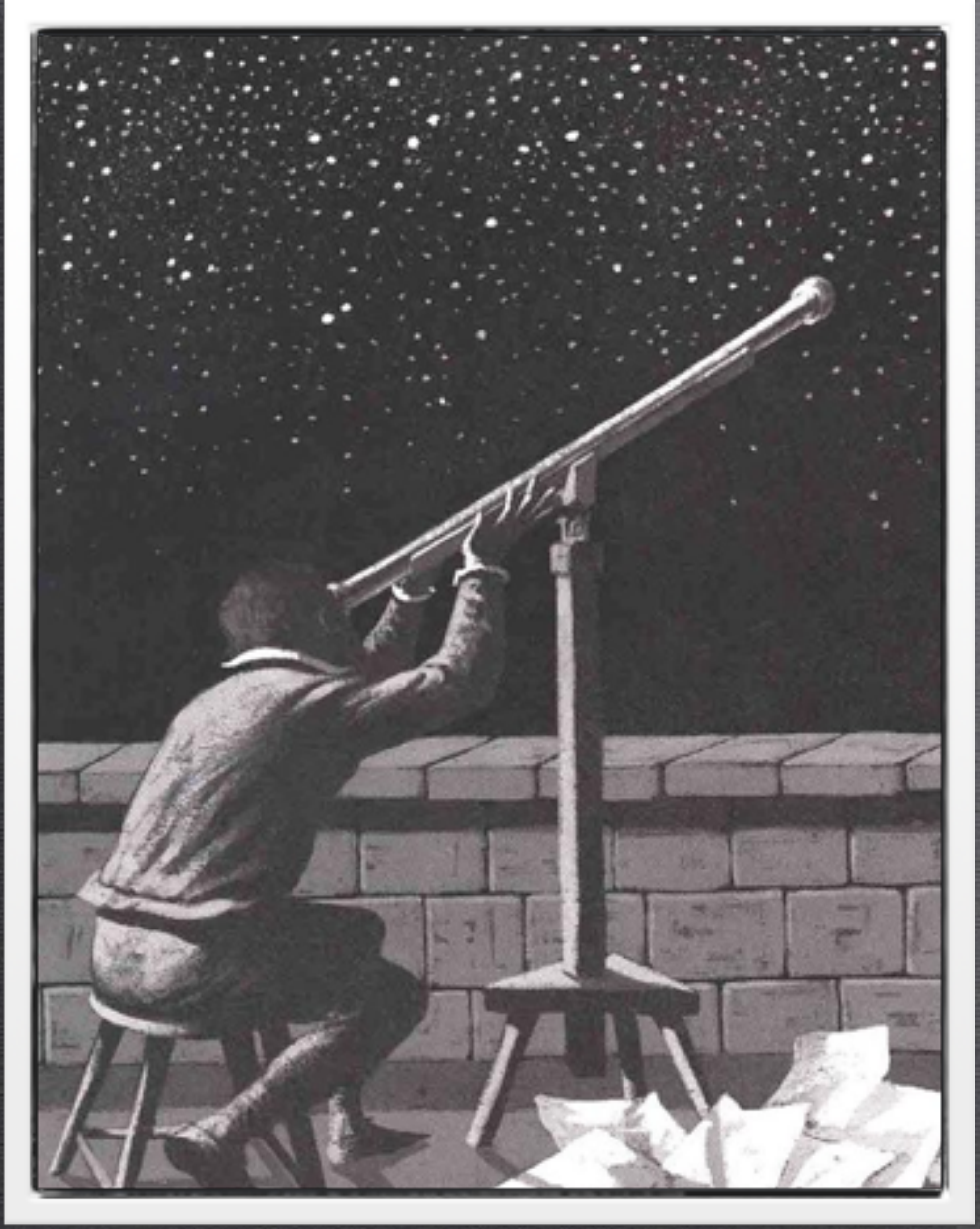
Data.gov.uk contains more than 7,200 datasets from seven governmental publishers, including 989 from the Department of Health and 784 from the Department for Communities and Local Government.

Launched by Vivek Kundra, the first Chief Information Officer of the United States, Data.gov offers 389,730 datasets.

Data.gov.in launched in July 2011, modeled after the U.S. and UK sites, but with many datasets restricted to agency access only.

Launched in July 2011, the Kenya Open Data Initiative is the first Sub-Saharan national data program and will be used to create infrastructure for human and economic development.

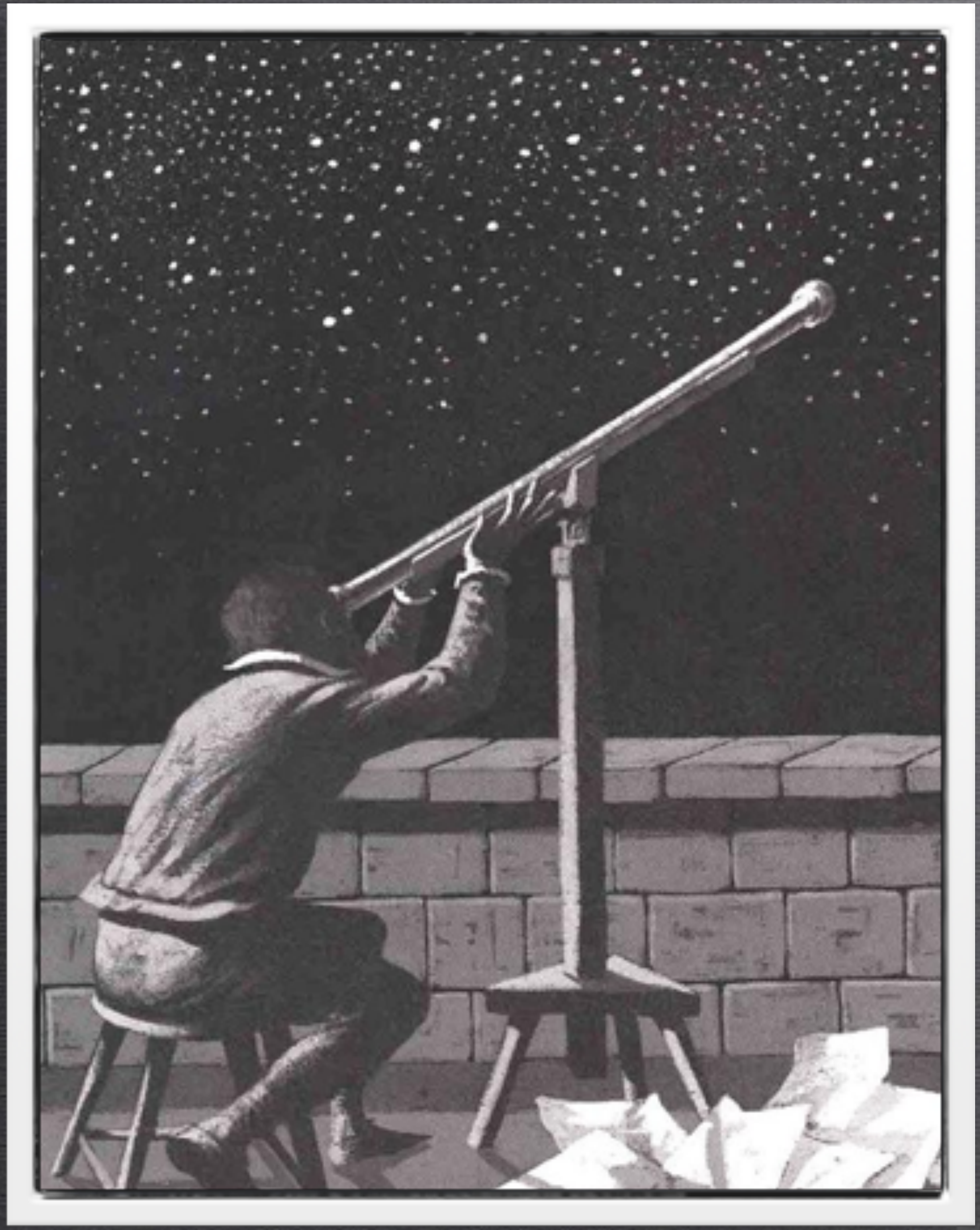
visual.ly



1609

Galileo Galilei discovers Saturn's rings

What did he do?



1450

Printing press

1609

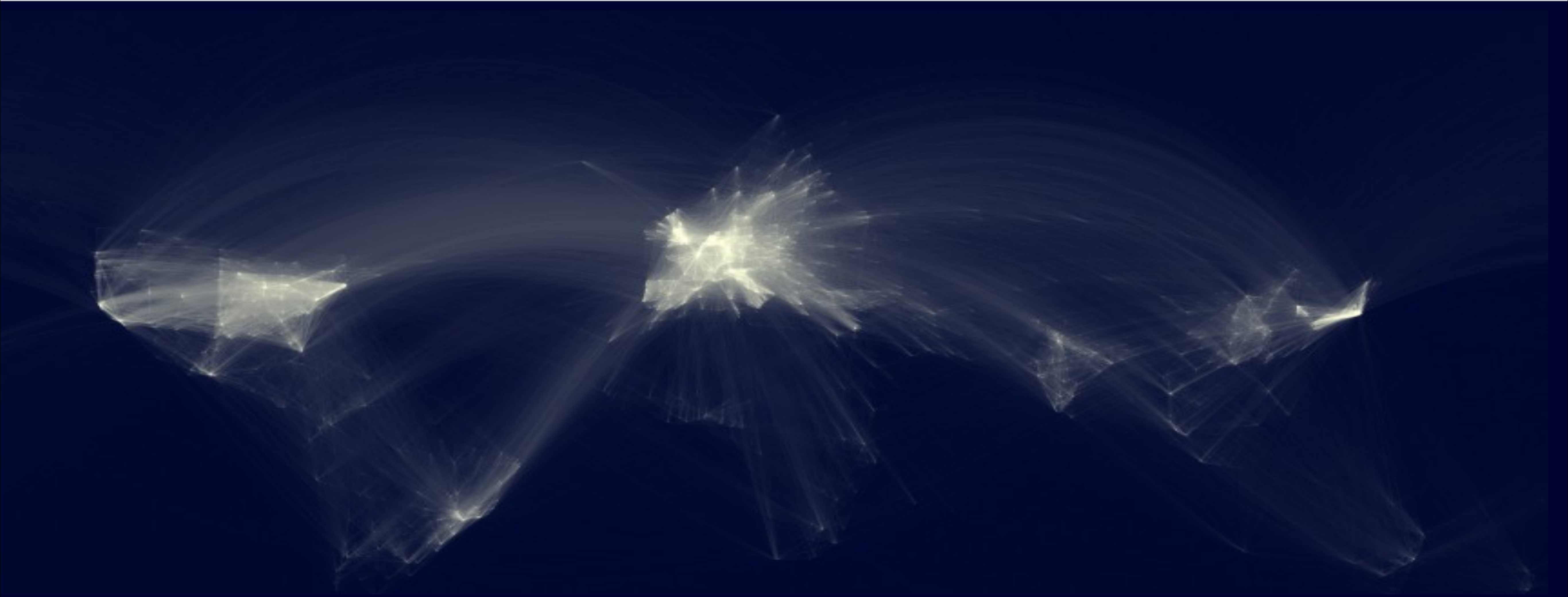
Galileo Galilei
anagrams

18th century

Scientific revolution

Journal accepted as best
way to advance science

today

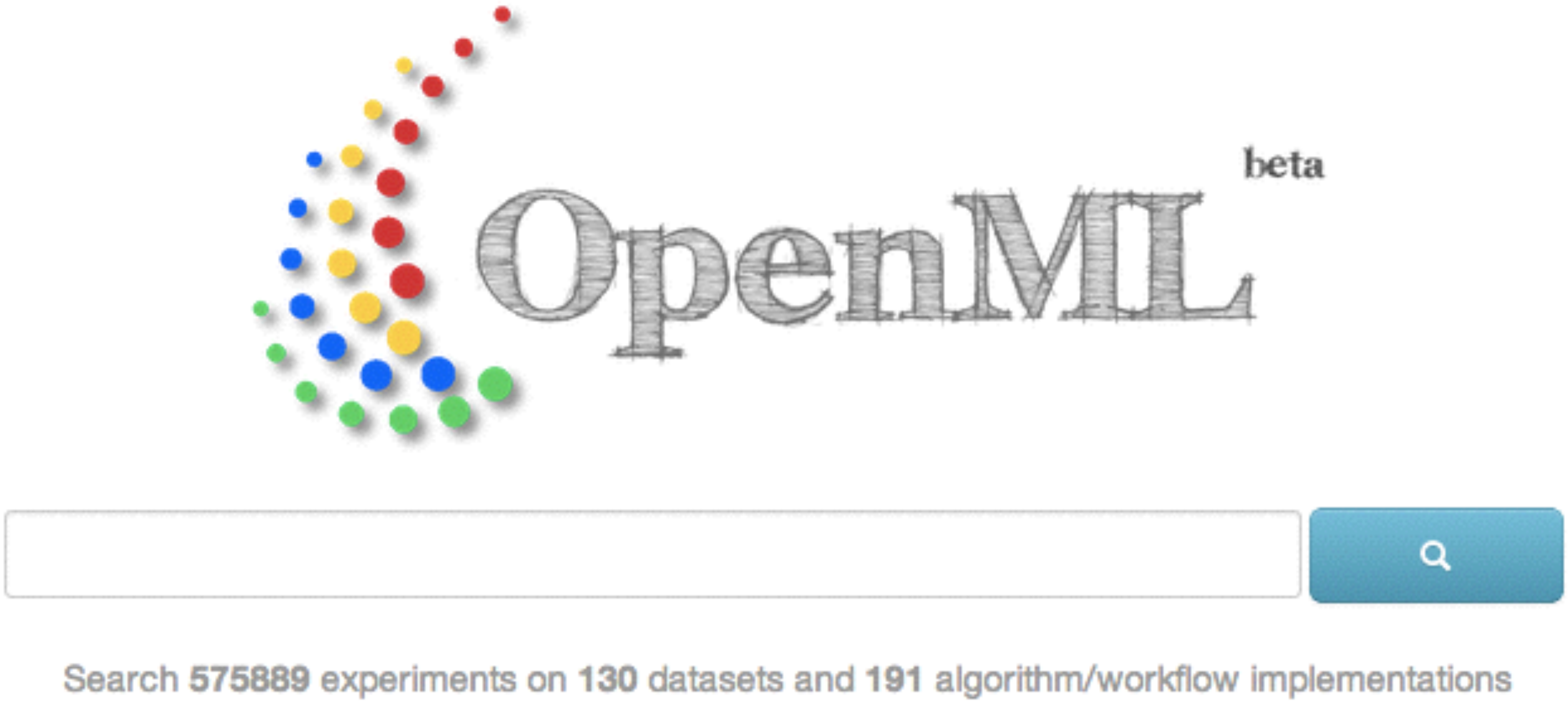


Are journals still the best we can do?

We have the internet, but publish results on paper?

openml.org

An open science platform for machine learning

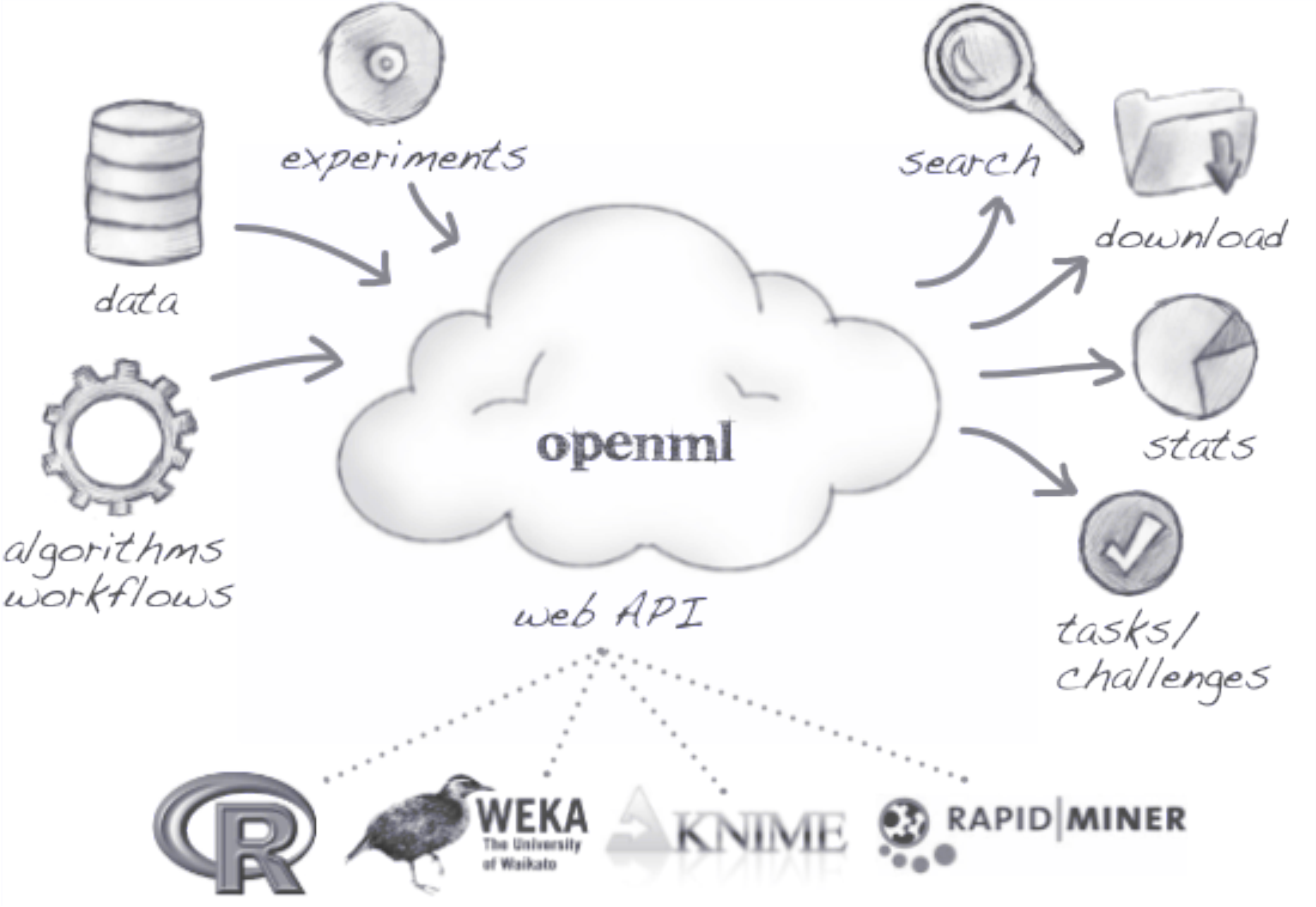


Search 575889 experiments on 130 datasets and 191 algorithm/workflow implementations

openml.org

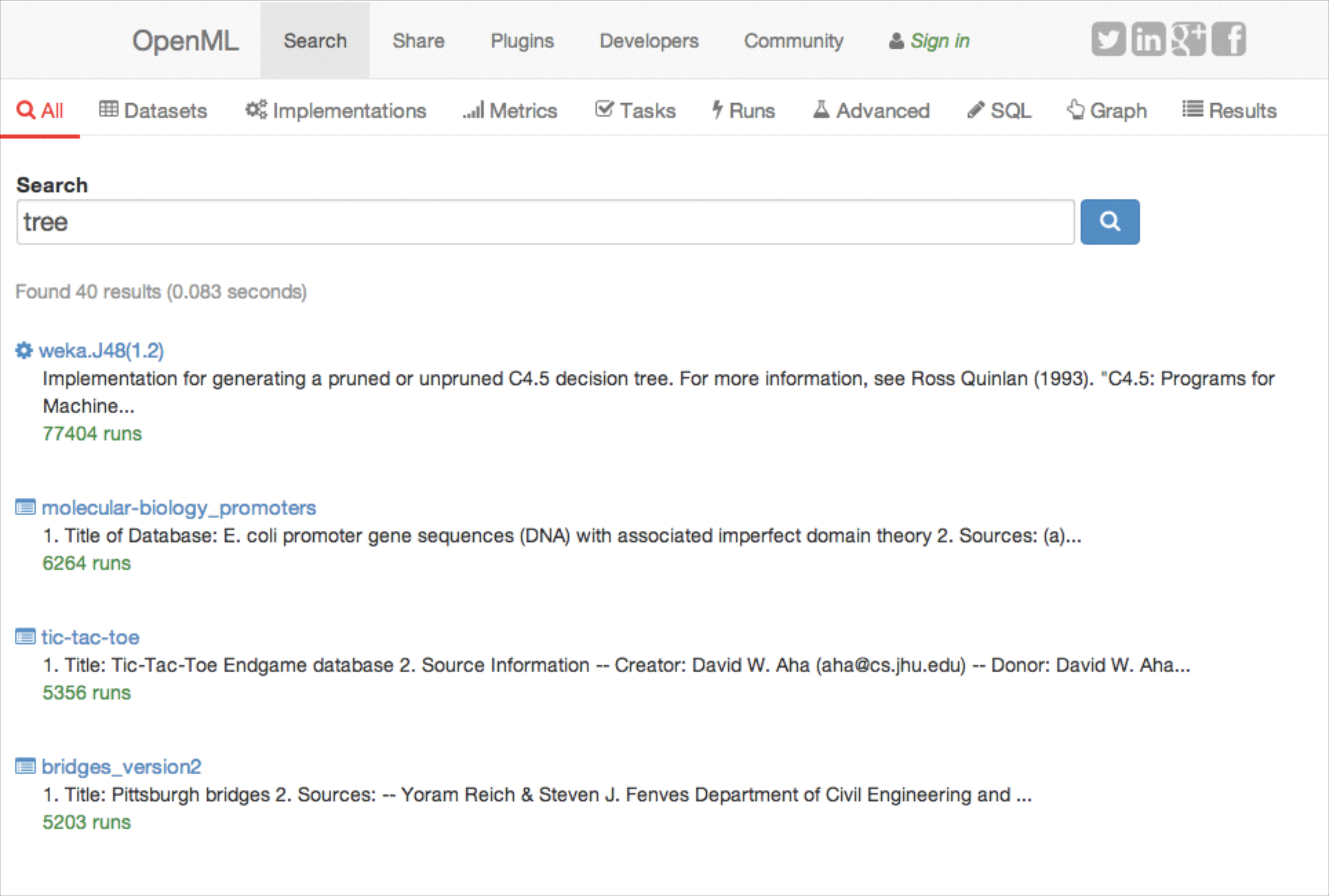
Share results

Search results



Integrated in machine learning tools

Search: Free text



Search

tree

Found 40 results (0.083 seconds)

weka.J48(1.2)

Implementation for generating a pruned or unpruned C4.5 decision tree. For more information, see Ross Quinlan (1993). "C4.5: Programs for Machine...

77404 runs

molecular-biology_promoters

1. Title of Database: E. coli promoter gene sequences (DNA) with associated imperfect domain theory 2. Sources: (a)...

6264 runs

tic-tac-toe

1. Title: Tic-Tac-Toe Endgame database 2. Source Information -- Creator: David W. Aha (aha@cs.jhu.edu) -- Donor: David W. Aha...

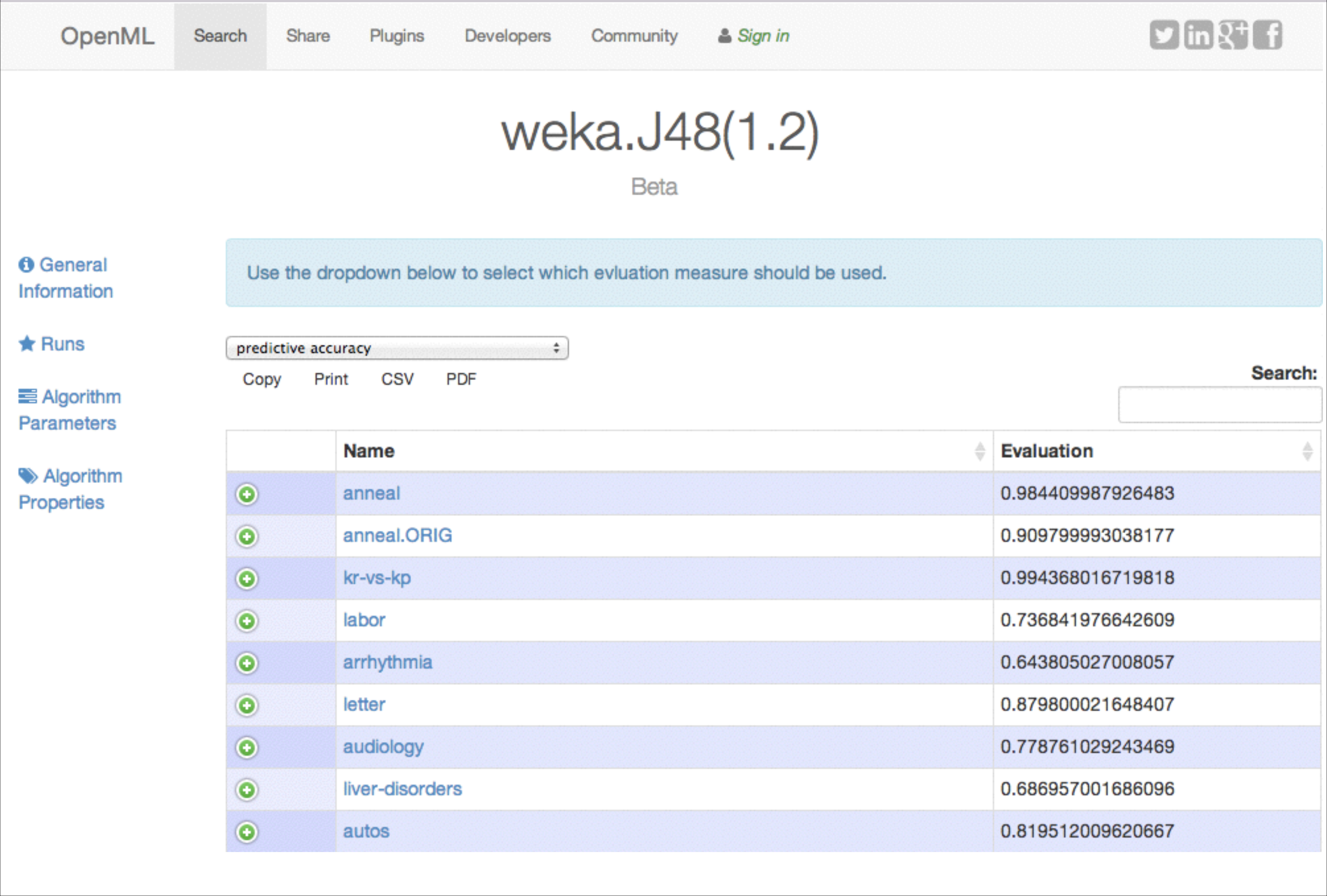
5356 runs

bridges_version2

1. Title: Pittsburgh bridges 2. Sources: -- Yoram Reich & Steven J. Fenves Department of Civil Engineering and ...

5203 runs

Search: Algorithm detail



weka.J48(1.2)

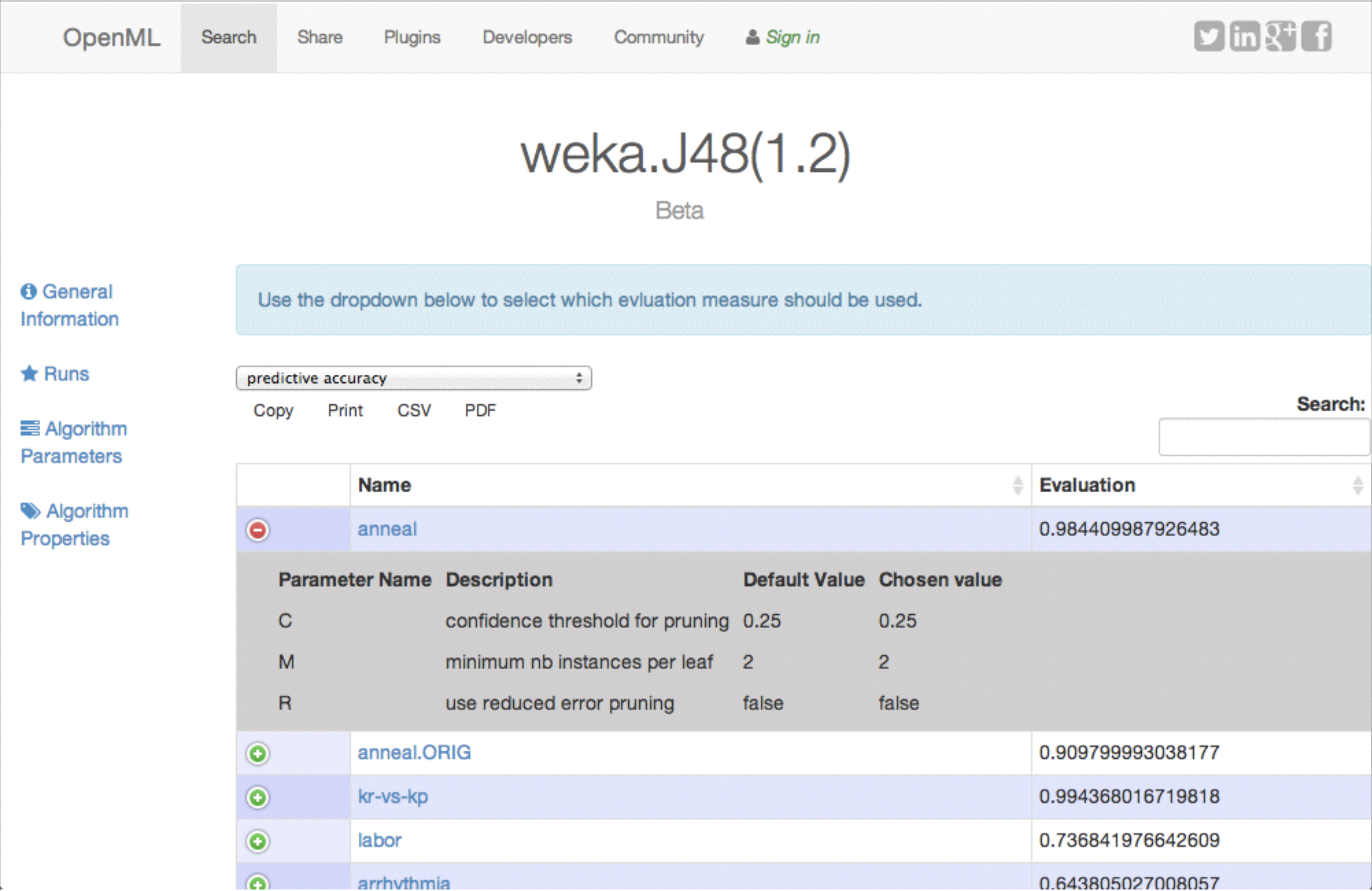
Beta

Use the dropdown below to select which evluation measure should be used.

predictive accuracy

	Name	Evaluation
⊕	anneal	0.984409987926483
⊕	anneal.ORIG	0.909799993038177
⊕	kr-vs-kp	0.994368016719818
⊕	labor	0.736841976642609
⊕	arrhythmia	0.643805027008057
⊕	letter	0.879800021648407
⊕	audiology	0.778761029243469
⊕	liver-disorders	0.686957001686096
⊕	autos	0.819512009620667

Search: Run detail



weka.J48(1.2)

Beta

Use the dropdown below to select which evluation measure should be used.

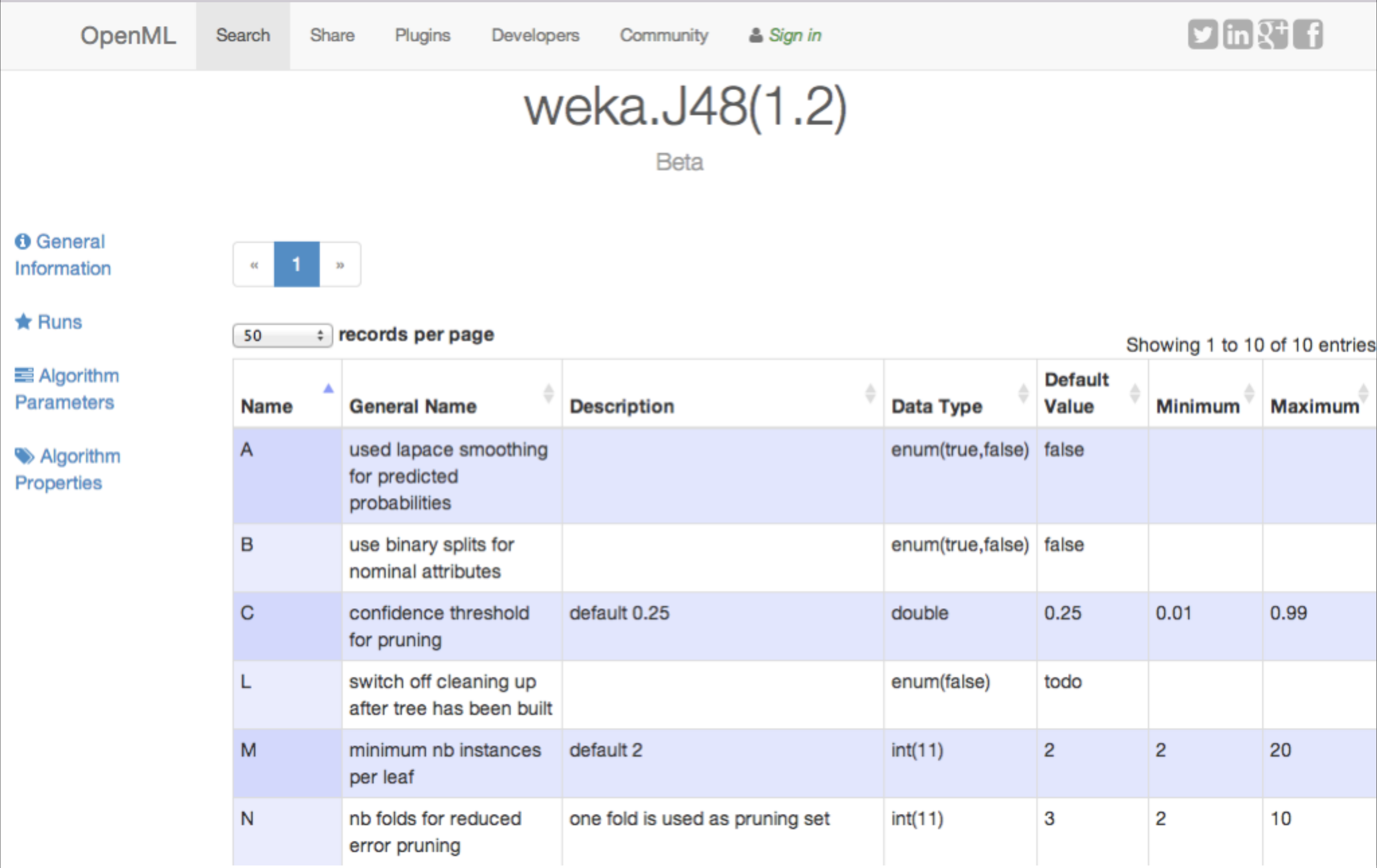
predictive accuracy

	Name	Evaluation
⊖	anneal	0.984409987926483

Parameter Name	Description	Default Value	Chosen value
C	confidence threshold for pruning	0.25	0.25
M	minimum nb instances per leaf	2	2
R	use reduced error pruning	false	false

	Name	Evaluation
⊕	anneal.ORIG	0.909799993038177
⊕	kr-vs-kp	0.994368016719818
⊕	labor	0.736841976642609
⊕	arrhythmia	0.643805027008057

Search: Algorithm parameters



weka.J48(1.2)

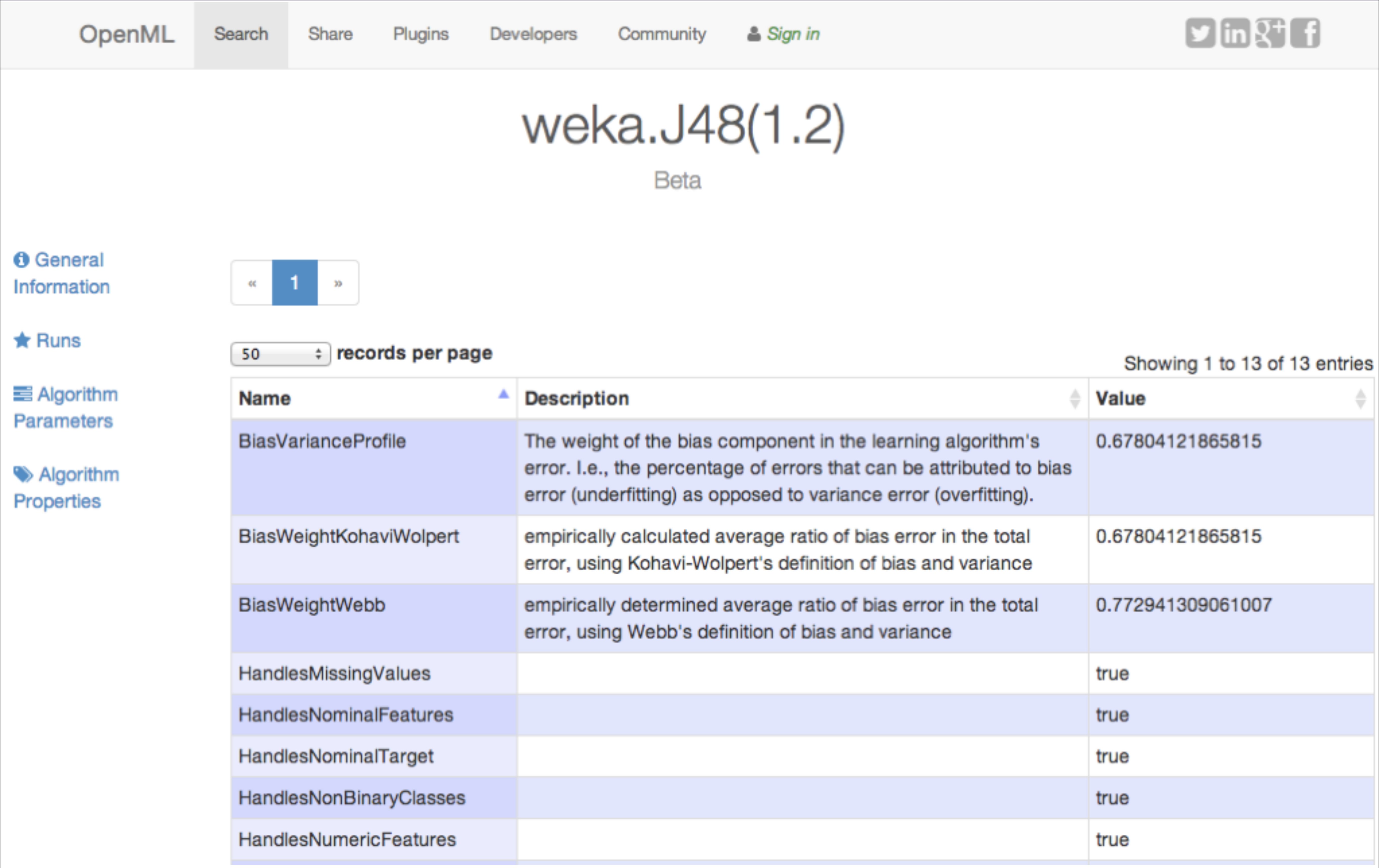
Beta

records per page

Showing 1 to 10 of 10 entries

Name	General Name	Description	Data Type	Default Value	Minimum	Maximum
A	used lapace smoothing for predicted probabilities		enum(true,false)	false		
B	use binary splits for nominal attributes		enum(true,false)	false		
C	confidence threshold for pruning	default 0.25	double	0.25	0.01	0.99
L	switch off cleaning up after tree has been built		enum(false)	todo		
M	minimum nb instances per leaf	default 2	int(11)	2	2	20
N	nb folds for reduced error pruning	one fold is used as pruning set	int(11)	3	2	10

Search: Algorithm properties



weka.J48(1.2)

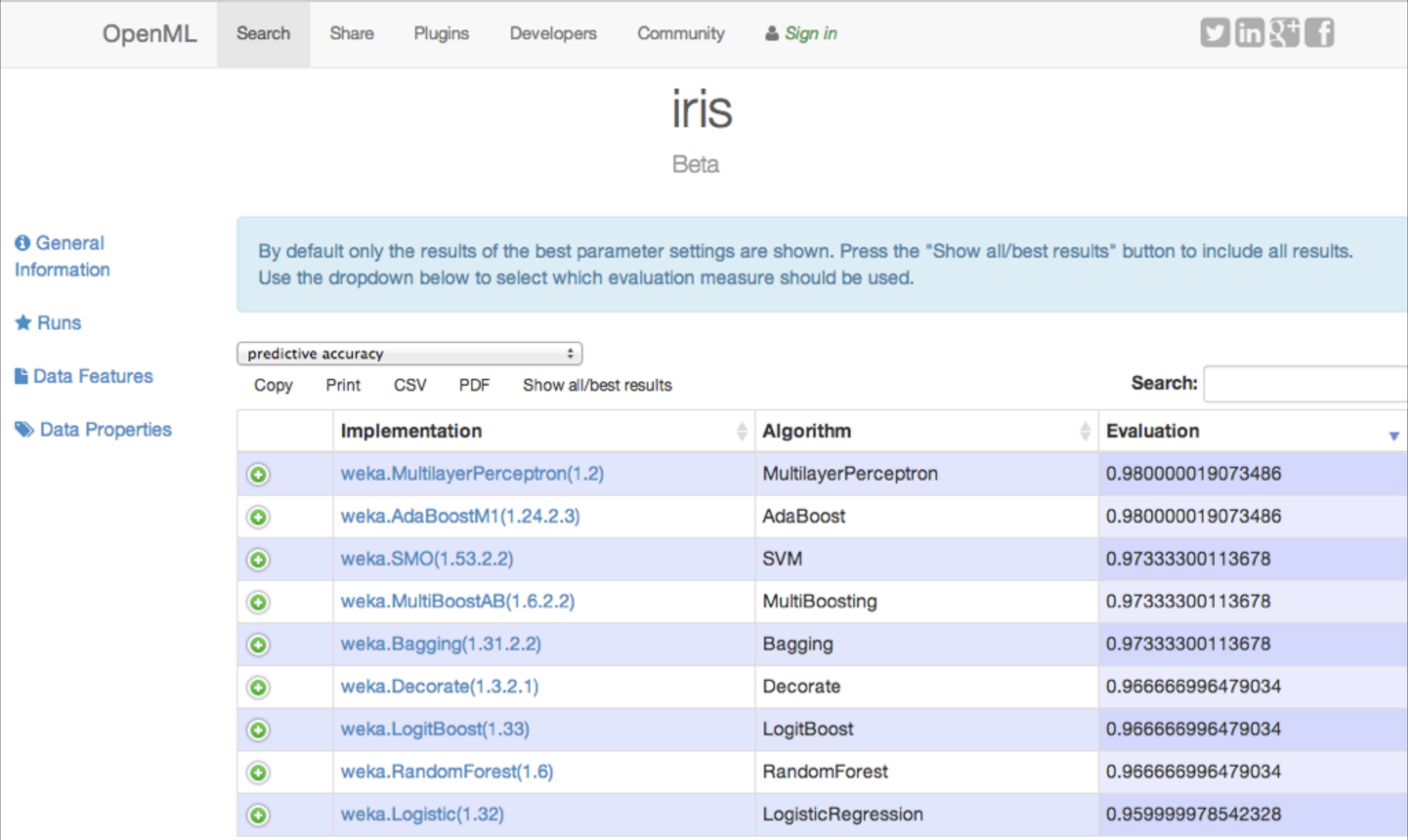
Beta

records per page

Showing 1 to 13 of 13 entries

Name	Description	Value
BiasVarianceProfile	The weight of the bias component in the learning algorithm's error. I.e., the percentage of errors that can be attributed to bias error (underfitting) as opposed to variance error (overfitting).	0.67804121865815
BiasWeightKohaviWolpert	empirically calculated average ratio of bias error in the total error, using Kohavi-Wolpert's definition of bias and variance	0.67804121865815
BiasWeightWebb	empirically determined average ratio of bias error in the total error, using Webb's definition of bias and variance	0.772941309061007
HandlesMissingValues		true
HandlesNominalFeatures		true
HandlesNominalTarget		true
HandlesNonBinaryClasses		true
HandlesNumericFeatures		true

Search: Dataset detail



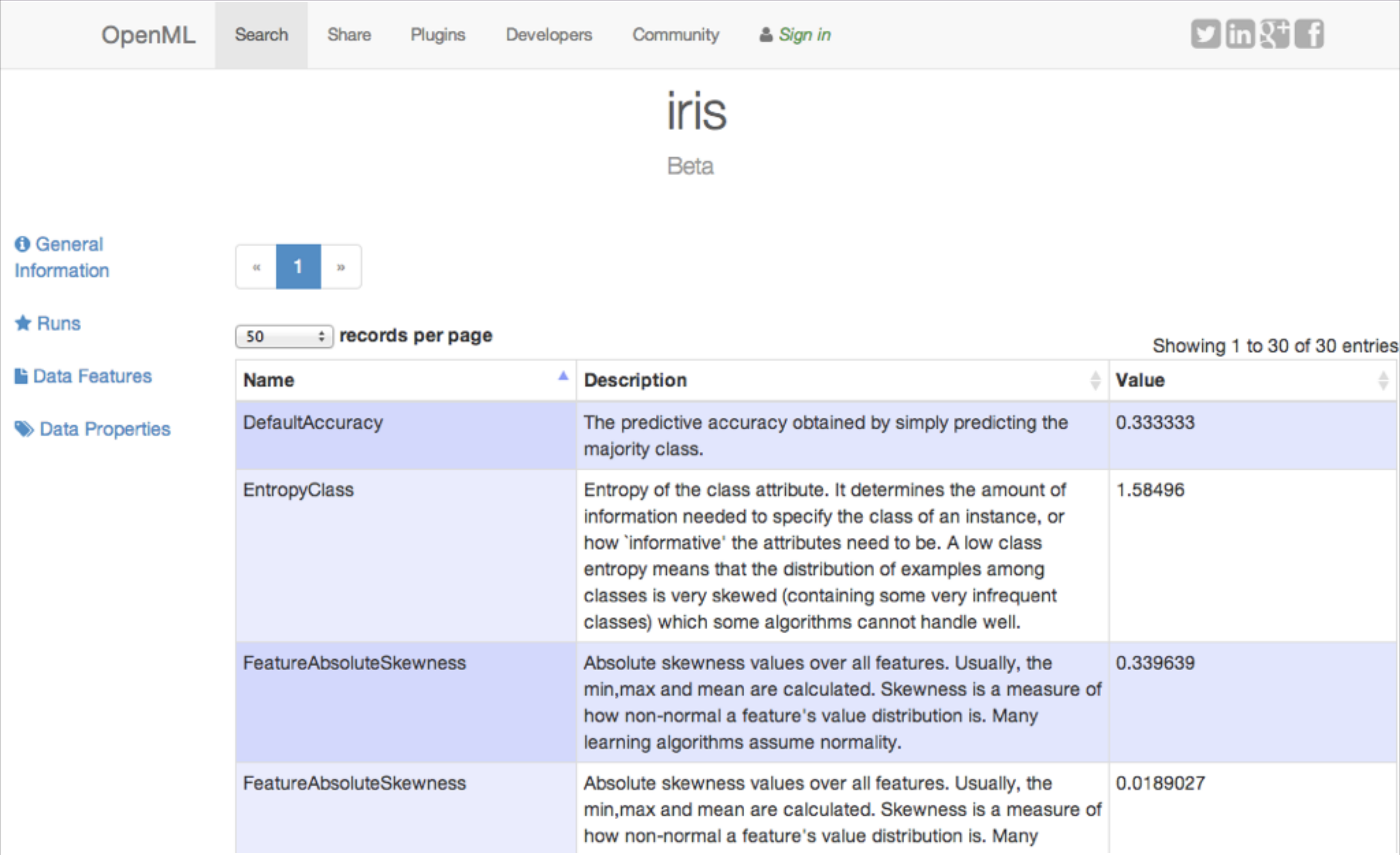
iris

Beta

By default only the results of the best parameter settings are shown. Press the "Show all/best results" button to include all results. Use the dropdown below to select which evaluation measure should be used.

	Implementation	Algorithm	Evaluation
⊕	weka.MultilayerPerceptron(1.2)	MultilayerPerceptron	0.980000019073486
⊕	weka.AdaBoostM1(1.24.2.3)	AdaBoost	0.980000019073486
⊕	weka.SMO(1.53.2.2)	SVM	0.97333300113678
⊕	weka.MultiBoostAB(1.6.2.2)	MultiBoosting	0.97333300113678
⊕	weka.Bagging(1.31.2.2)	Bagging	0.97333300113678
⊕	weka.Decorate(1.3.2.1)	Decorate	0.966666996479034
⊕	weka.LogitBoost(1.33)	LogitBoost	0.966666996479034
⊕	weka.RandomForest(1.6)	RandomForest	0.966666996479034
⊕	weka.Logistic(1.32)	LogisticRegression	0.959999978542328

Search: Dataset properties



iris

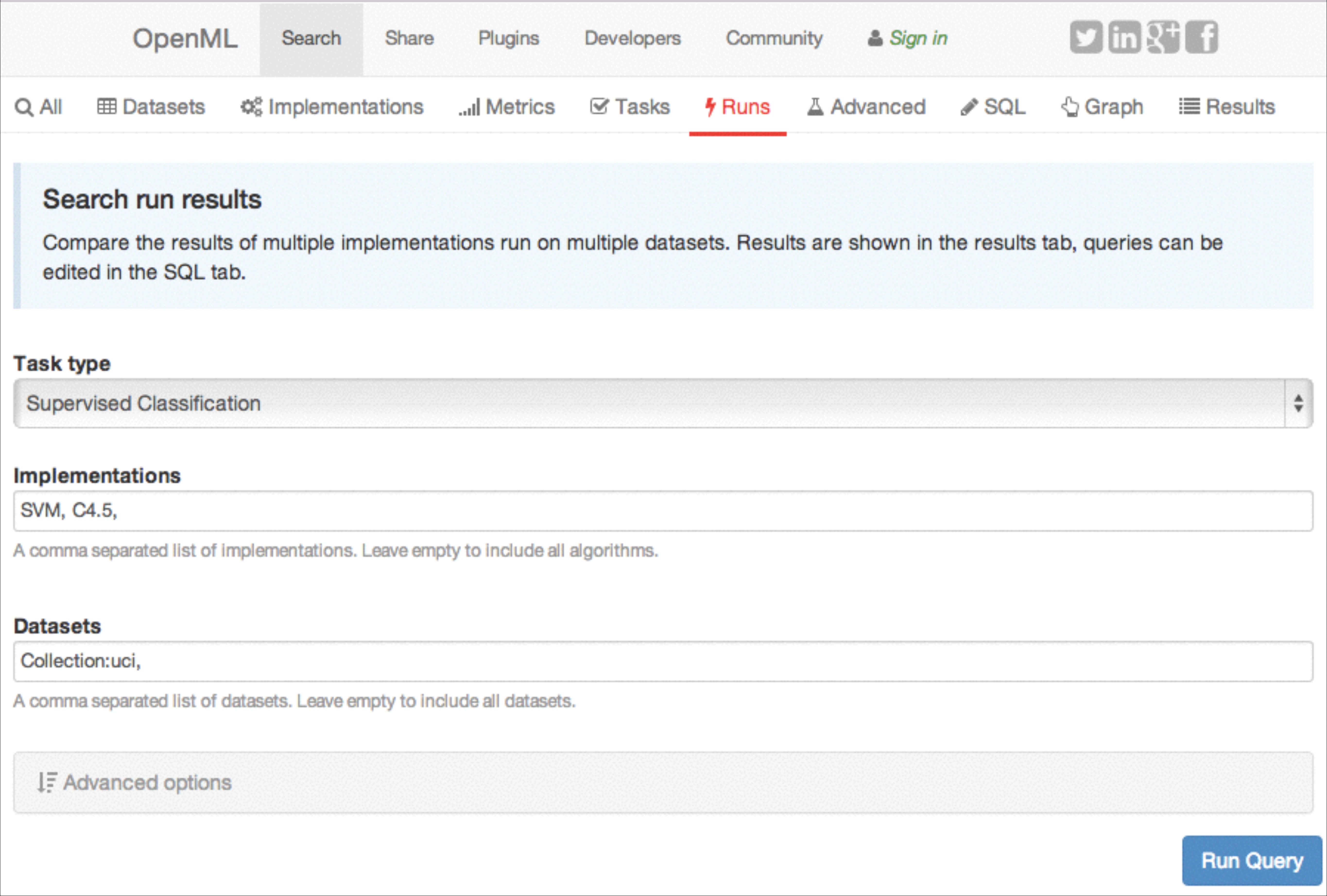
Beta

50 records per page

Showing 1 to 30 of 30 entries

Name	Description	Value
DefaultAccuracy	The predictive accuracy obtained by simply predicting the majority class.	0.333333
EntropyClass	Entropy of the class attribute. It determines the amount of information needed to specify the class of an instance, or how `informative' the attributes need to be. A low class entropy means that the distribution of examples among classes is very skewed (containing some very infrequent classes) which some algorithms cannot handle well.	1.58496
FeatureAbsoluteSkewness	Absolute skewness values over all features. Usually, the min,max and mean are calculated. Skewness is a measure of how non-normal a feature's value distribution is. Many learning algorithms assume normality.	0.339639
FeatureAbsoluteSkewness	Absolute skewness values over all features. Usually, the min,max and mean are calculated. Skewness is a measure of how non-normal a feature's value distribution is. Many	0.0189027

Search: Quick comparisons



Search run results

Compare the results of multiple implementations run on multiple datasets. Results are shown in the results tab, queries can be edited in the SQL tab.

Task type

Supervised Classification

Implementations

SVM, C4.5,

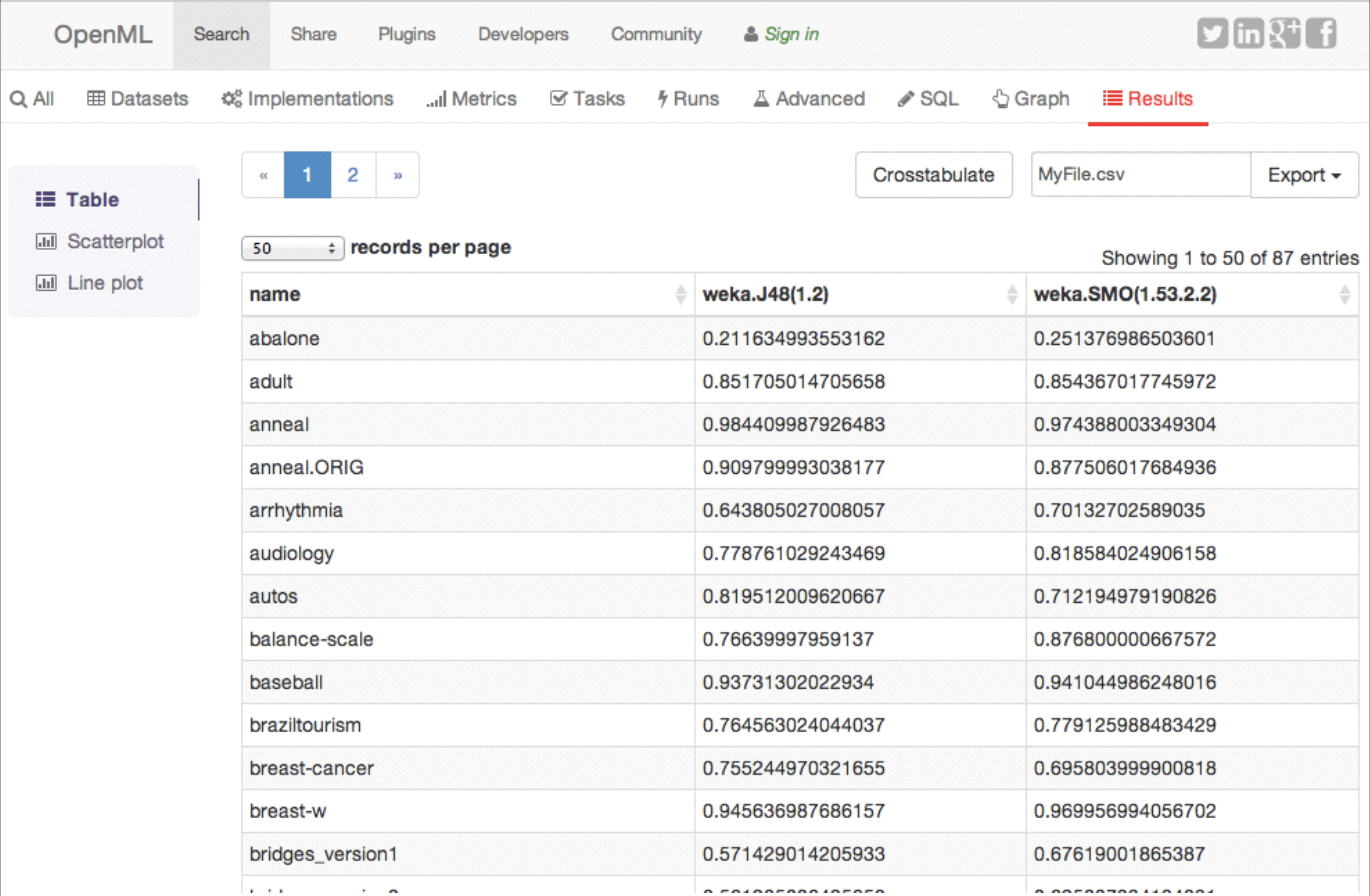
A comma separated list of implementations. Leave empty to include all algorithms.

Datasets

Collection:uci,

A comma separated list of datasets. Leave empty to include all datasets.

Search: Quick comparisons

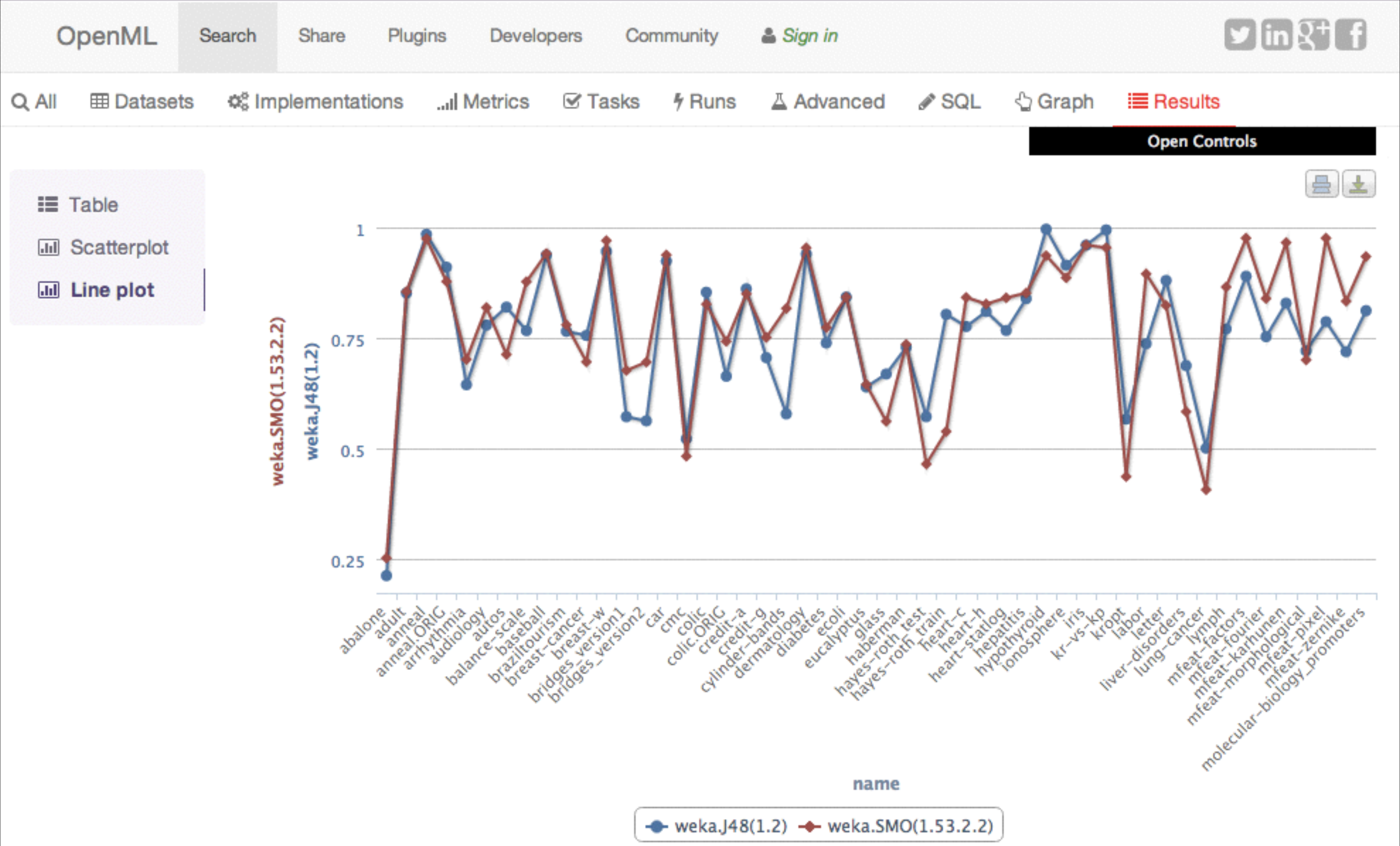


records per page

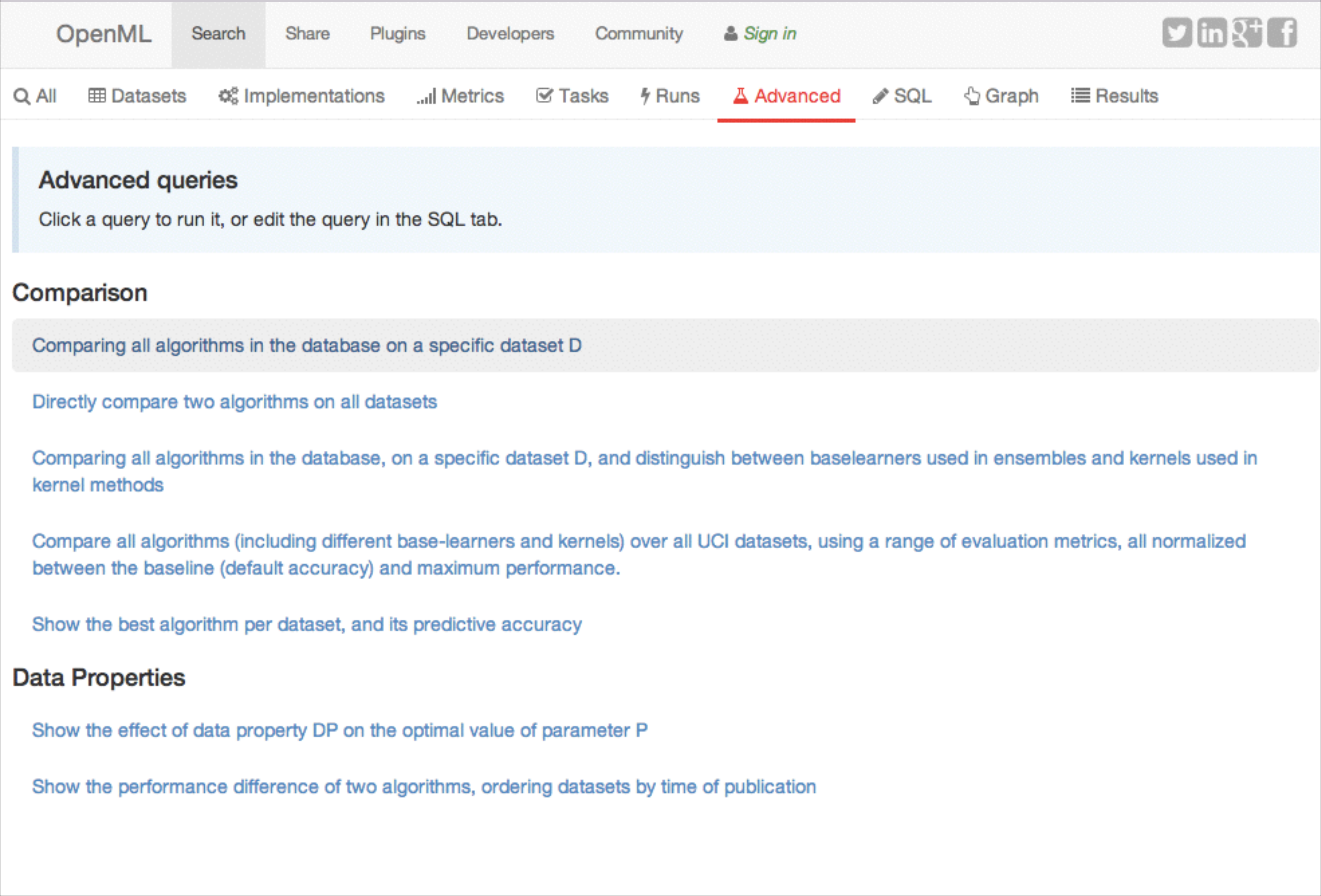
Showing 1 to 50 of 87 entries

name	weka.J48(1.2)	weka.SMO(1.53.2.2)
abalone	0.211634993553162	0.251376986503601
adult	0.851705014705658	0.854367017745972
anneal	0.984409987926483	0.974388003349304
anneal.ORIG	0.909799993038177	0.877506017684936
arrhythmia	0.643805027008057	0.70132702589035
audiology	0.778761029243469	0.818584024906158
autos	0.819512009620667	0.712194979190826
balance-scale	0.76639997959137	0.876800000667572
baseball	0.93731302022934	0.941044986248016
braziltourism	0.764563024044037	0.779125988483429
breast-cancer	0.755244970321655	0.695803999900818
breast-w	0.945636987686157	0.969956994056702
bridges_version1	0.571429014205933	0.67619001865387

Search: Visualizations



Search: Advanced queries



Advanced queries

Click a query to run it, or edit the query in the SQL tab.

Comparison

Comparing all algorithms in the database on a specific dataset D

Directly compare two algorithms on all datasets

Comparing all algorithms in the database, on a specific dataset D, and distinguish between baselearners used in ensembles and kernels used in kernel methods

Compare all algorithms (including different base-learners and kernels) over all UCI datasets, using a range of evaluation metrics, all normalized between the baseline (default accuracy) and maximum performance.

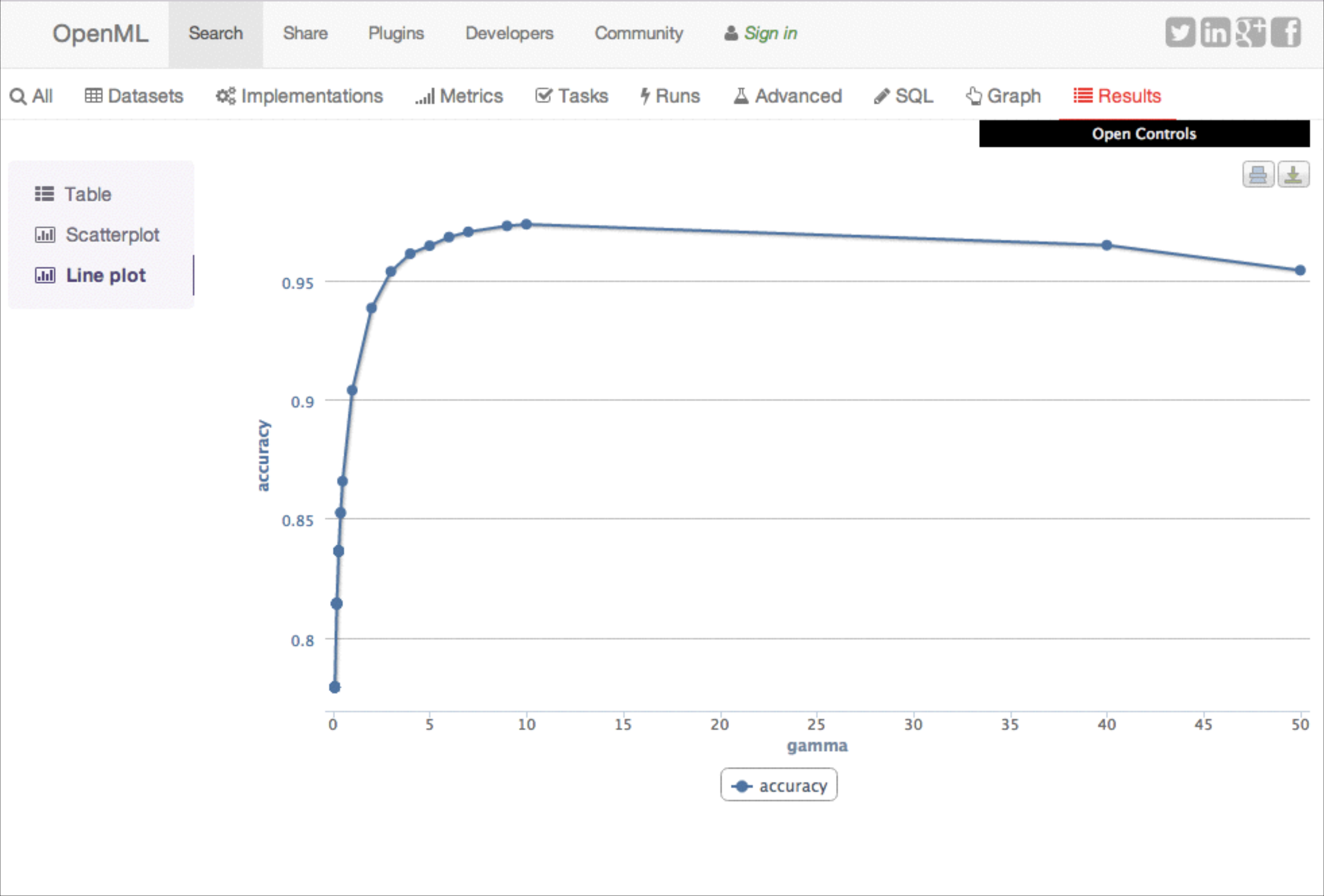
Show the best algorithm per dataset, and its predictive accuracy

Data Properties

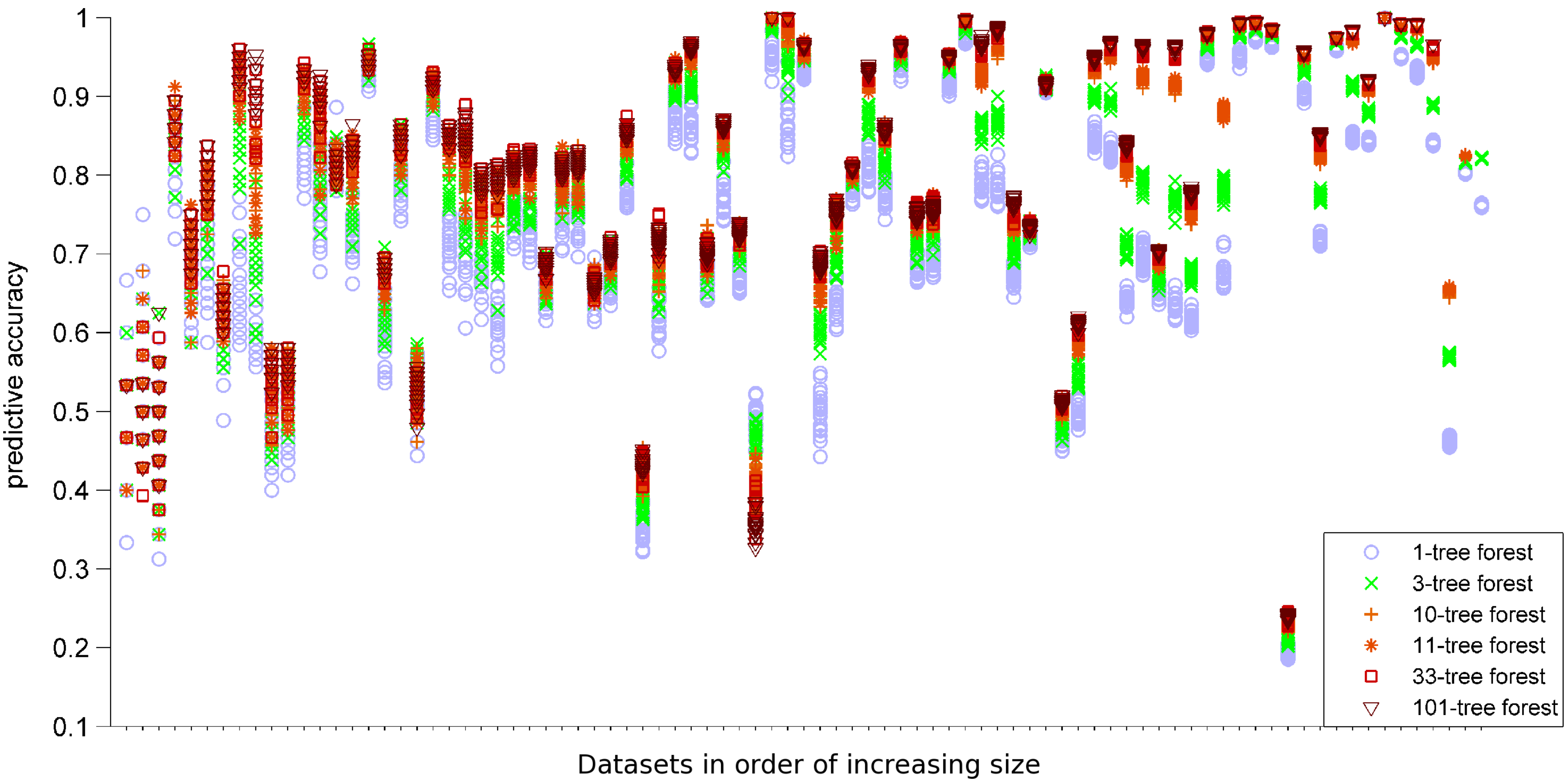
Show the effect of data property DP on the optimal value of parameter P

Show the performance difference of two algorithms, ordering datasets by time of publication

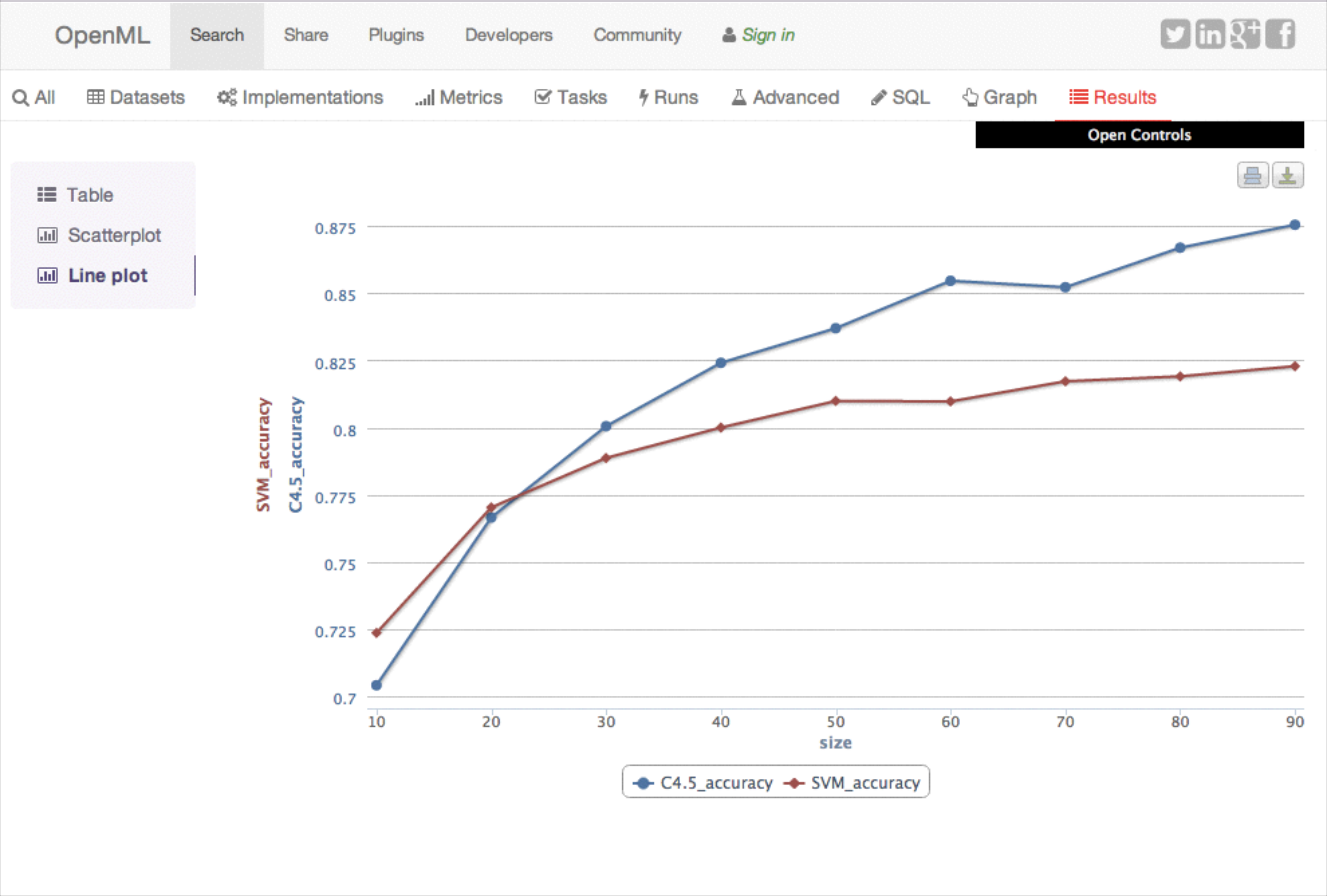
Search: Parameter effects



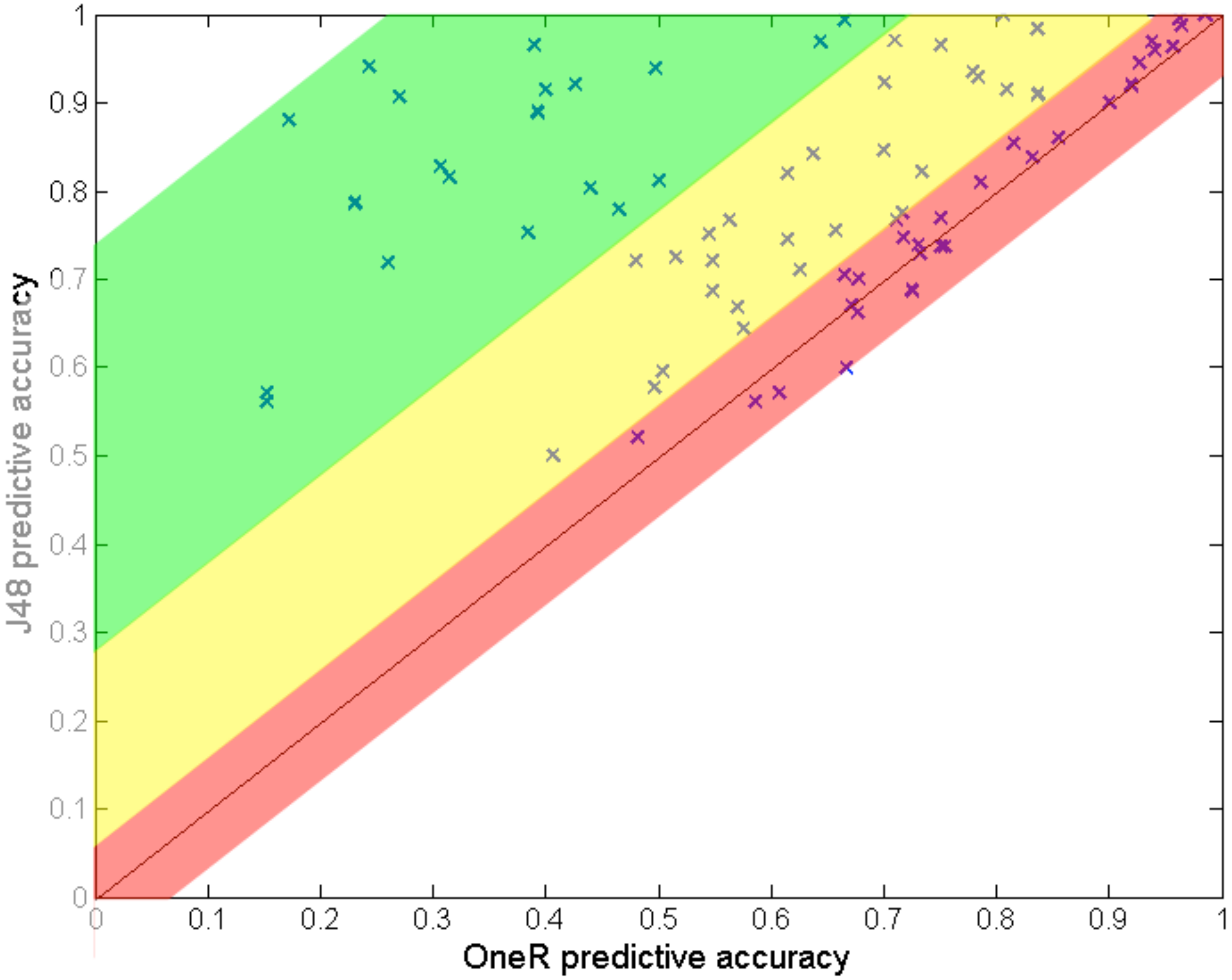
Search: Parameter effects

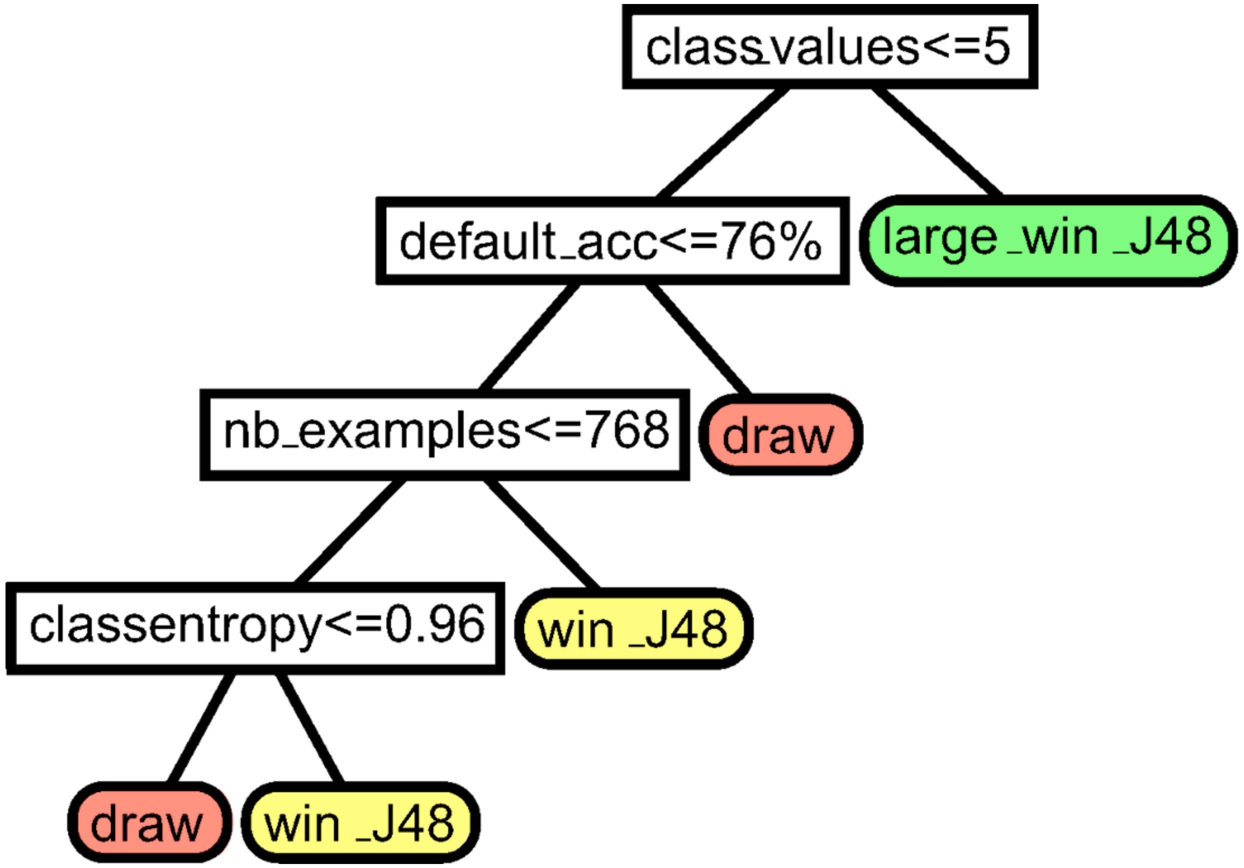


Search: Learning curves



Meta-models





class_values<=5

default_acc<=76%

large_win_J48

nb_examples<=768

draw

classentropy<=0.96

win_J48

draw

win_J48

🐦 @joavanschoren

joaquin.vanschoren@gmail.com