BigData with Python

A gentle and simple introduction

Marcel Caraciolo

@marcelcaraciolo

Developer, Cientist, contributor to the Crab recsys project, works with Python for 6 years, interested at mobile, education, machine learning and dataaaaa!

Recife, Brazil - http://aimotion.blogspot.com

Disclaimer

Alguns slides foram retirados das apts do curso de Bill Howe - Introduction to Data Science

https://class.coursera.org/datasci-001/

About me

Co-founder of **Crab** - Python recsys library

Cientist Chief at Atepassar, e-learning social network

Co-Founder and Instructor of PyCursos, teaching Python on-line

Co-Founder of Pingmind, on-line infrastructure for MOOC's

Interested at Python, mobile, e-learning and machine learning!

What is BigData ?

Big Data

“Big Data is any data that is expensive to manage and hard to extract value from.”

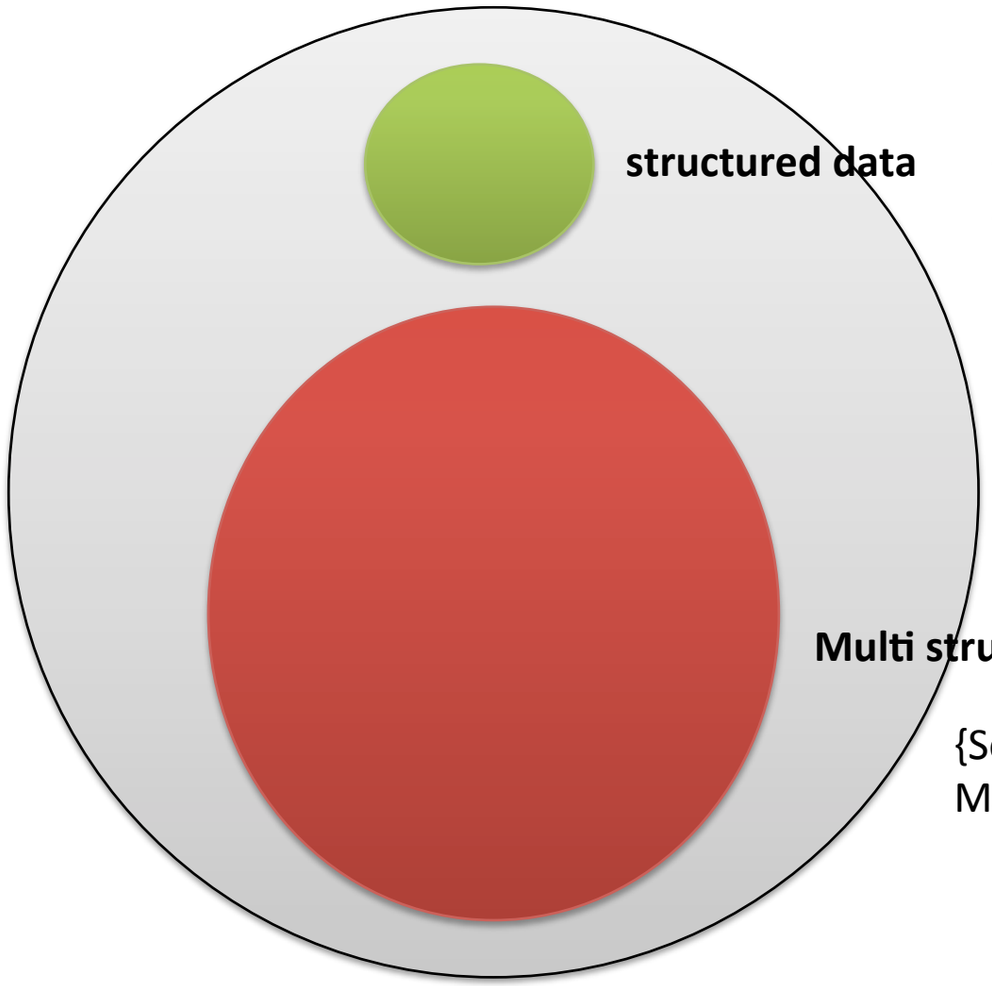
Michael Franklin Thomas M. Siebel
Professor of Computer Science Director of the Algorithms, Machines and People Lab University of Berkeley

Big Data

Erik Larson, 1989, Harper's magazine

"The keepers of big data say they do it for the consumer's benefit. But data have a way of being used for purposes other than originally intended."

Big Data, muitos dados.



structured data

{Relational data with well defined schemas}

Multi structured data

{Social data, blogs , click stream, Machine generated, xml etc…}

Challenges

Volume

the size of the data

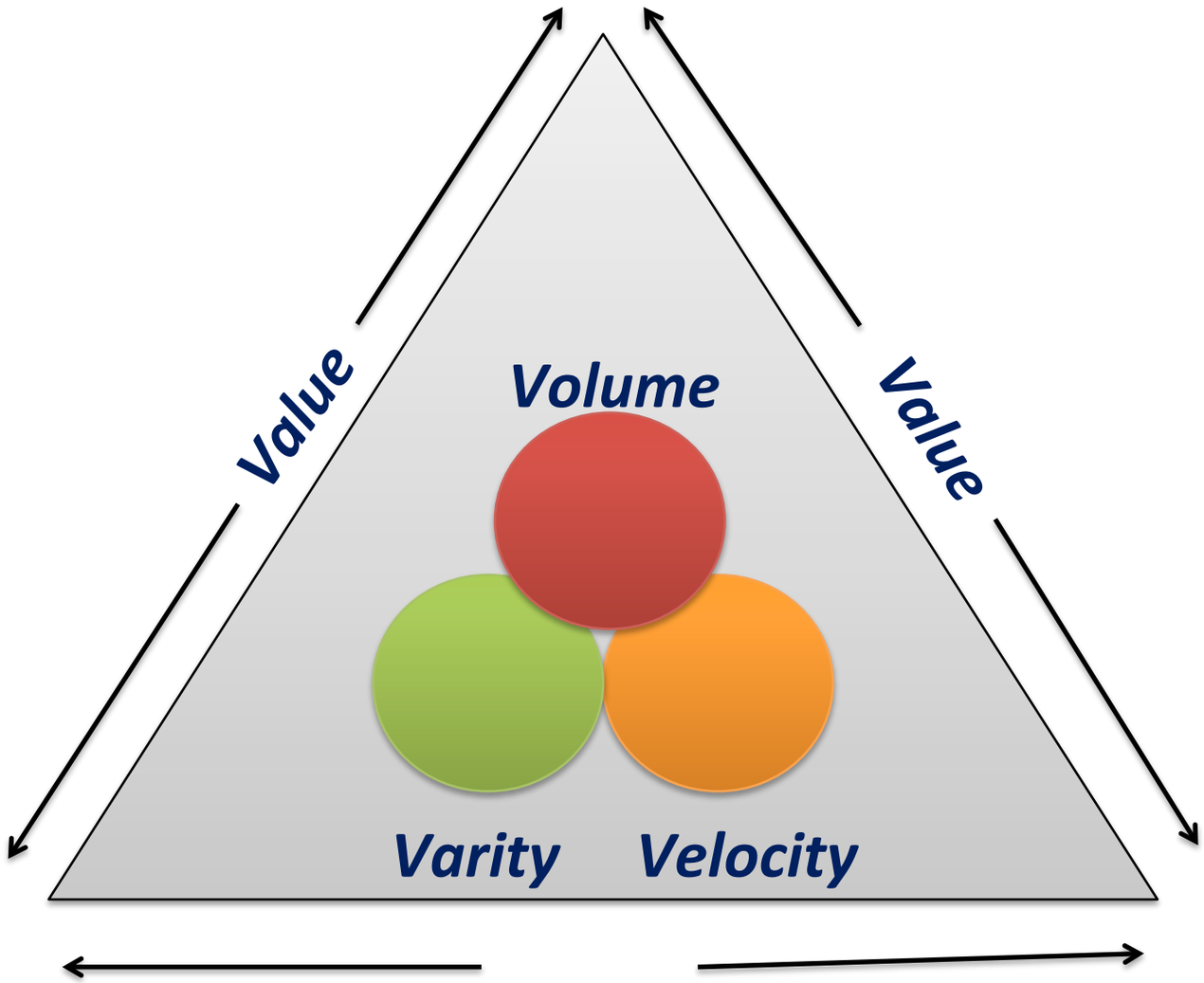
Velocity

the latency of data processing relative to the growing demand for interactivity

Variety

the diversity of sources, formats, quality, structures

Big Data Dimensions (V3)



Where does big data comes from ?

“data exhaust” from customers

new and pervasive sensors

the ability to “keep everything”

Trends … Gartner

Mobile analytics

Mobility

App stores and Market place

Human computer interface

Big Data

Personal cloud

Multi touch UI

In memory computing

Advanced Analytics

Green data centre

Flash Memory

Social CRM

Solid state drive

HTML5

Context aware computing

Where does big data comes from ?

Car black boxes: Privacy nightmare or a safety measure?

February 15, 2013 | By Ronald D. White

Email Share +1 1 Tweet 0 Recommend 271

What if the black box in your new car becomes a tool to invade your privacy? What if, on the other hand, it winds up saving your life after an accident?

Those are some of the questions being raised this week over black box data event recorders in cars. Privacy advocates worried on Thursday that the data could be misused. Safety advocates argued on Friday that a watered-down version of the recorders would slow safety innovations.

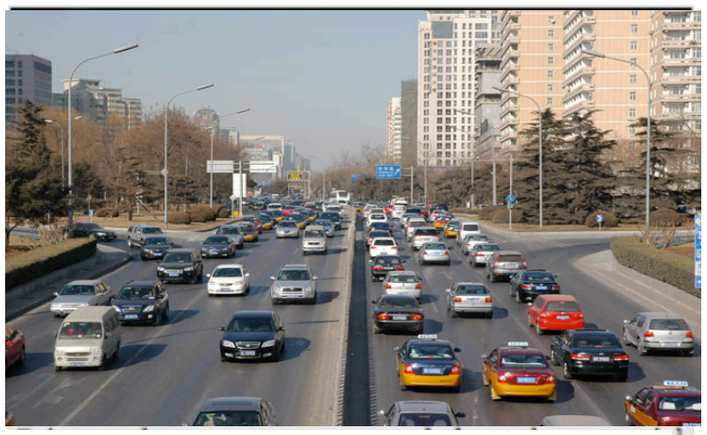
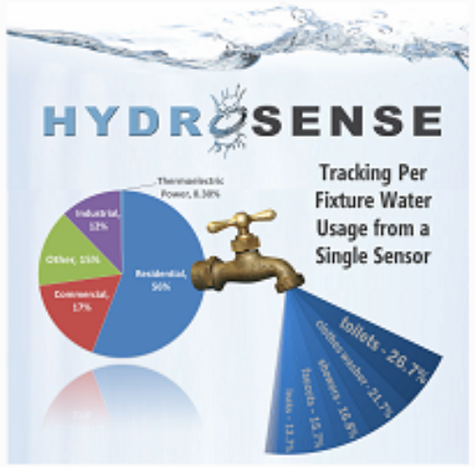


photo in public domain

Angeles...)

Where does big data comes from ?

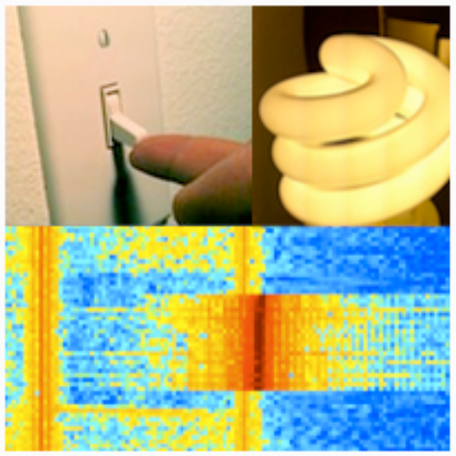


HydroSense

Water Fixture Usage with a Single Sensor

HydroSense is a pressure-based sensor that automatically determines water usage activity and flow down to the source (e.g., dishwasher, laundry, shower) from a single non-intrusive installation point.

Lead Researchers: Jon Froehlich, Eric Larson, Shwetak Patel



ElectriSense

Electrical Device Energy Usage with a Single Sensor

ElectriSense is a single plug-in sensor that provides whole home device level usage data. That is, using a single sensor plugged in anywhere in the home, ElectriSense can infer which electrical appliances are on and which off. This data could be used for numerous applications, for example, for providing home owners with itemized electrical bill that not only shows the total energy consumption but breaks the total on a per appliance basis (TV consumed 20 KWh, Lighting consumes 18 KWh and so on).

Lead Researchers: Sidhant Gupta, Shwetak Patel

The Problem…

Facebook

955 million active users as of March 2012,
1 in 3 Internet users have a Facebook account

More than 30 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) shared each month.

Holds 30PB of data for analysis, adds 12 TB of compressed data daily

The Problem…

Twitter

500 million users, 340 million daily tweets
1.6 billion search queries a day
7 TB data for analysis generated daily

The Problem…

Atepassar

180 thousand users, 300 thousand items
Recommend items daily for 8000 users
12 GB data for analysis generated daily

Traditional data storage, techniques & analysis tools just do not work at these scales !

Scalability

What does Scalable mean ?

Operationally:

In the past: **"Works even if data doesn't fit in main memory"**

Now: **"Can make use of 1000s of cheap computers"**

Algorithmically:

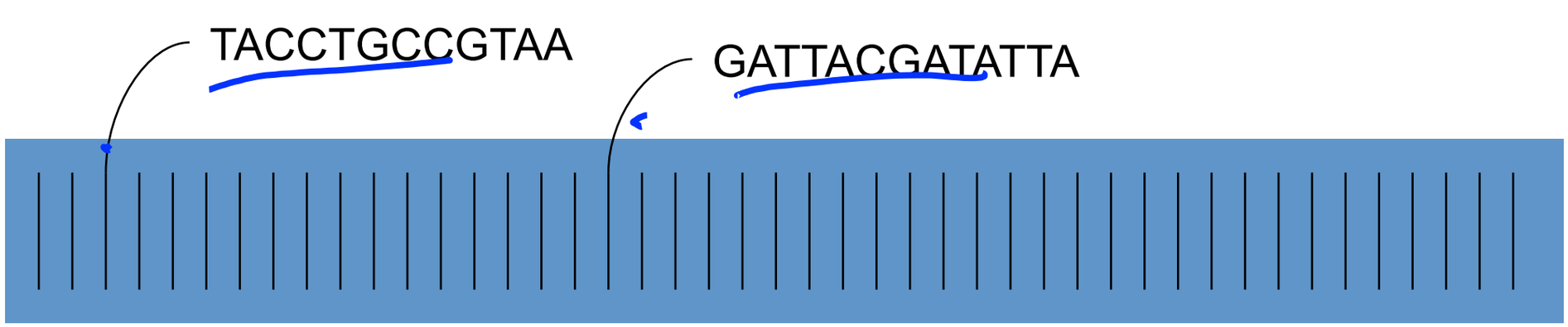
In the past: **If you have N data items, you must do no more than *Nm* operations -- "polynomial time algorithms"**

Now: **If you have N data items, you must do no more than *Nm*/*k* operations, for some large *k* -- "polynomial time algorithms must be parallelized"**

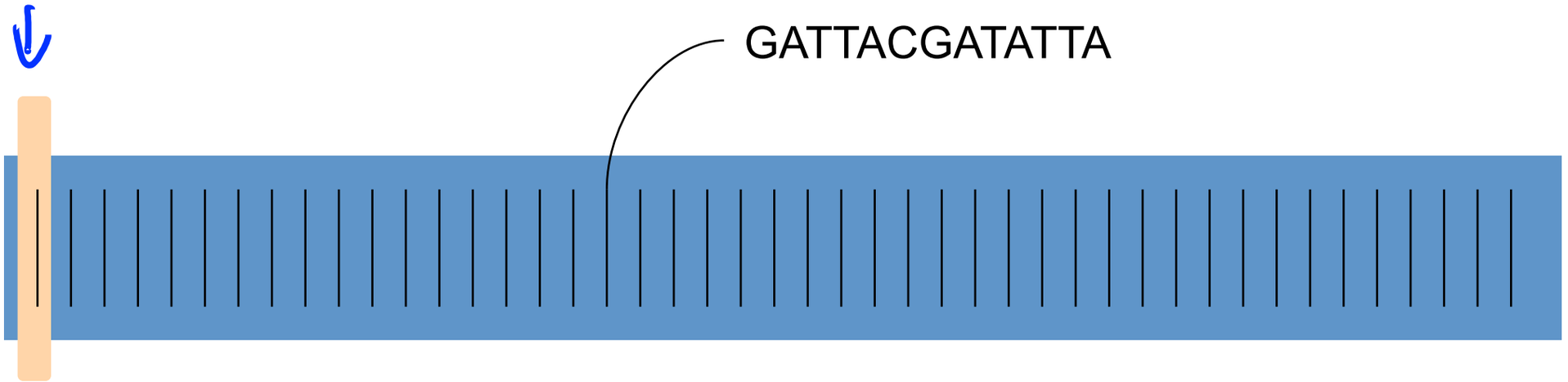
Soon: **If you have N data items, you should do no more than *N * log(N)* operations**

Example

Given a set of DNA sequences, find all sequences equal to “GATTACGATATTA”.

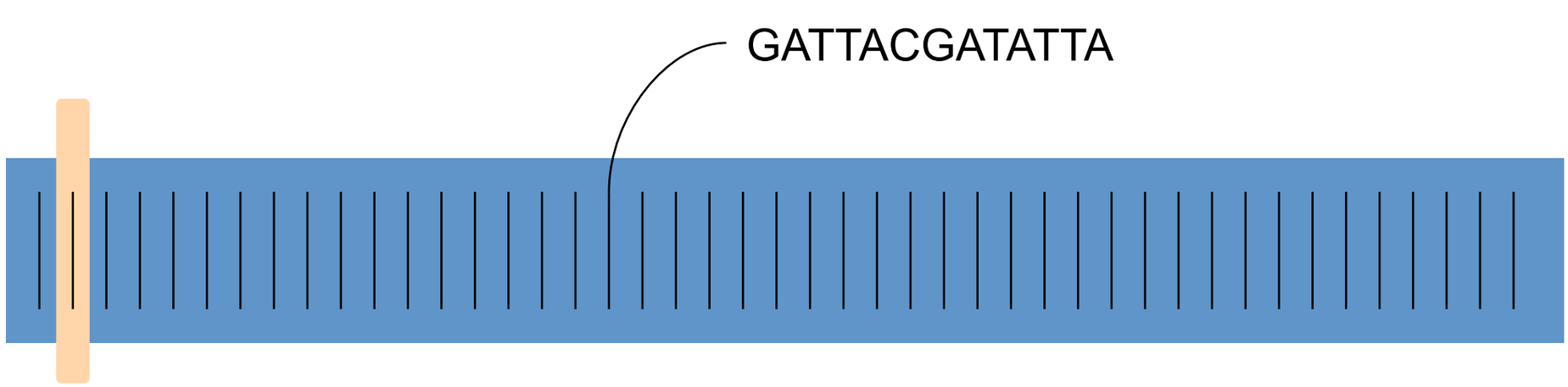

TACCTGCCGTAA
GATTACGATATTA

Time = 0



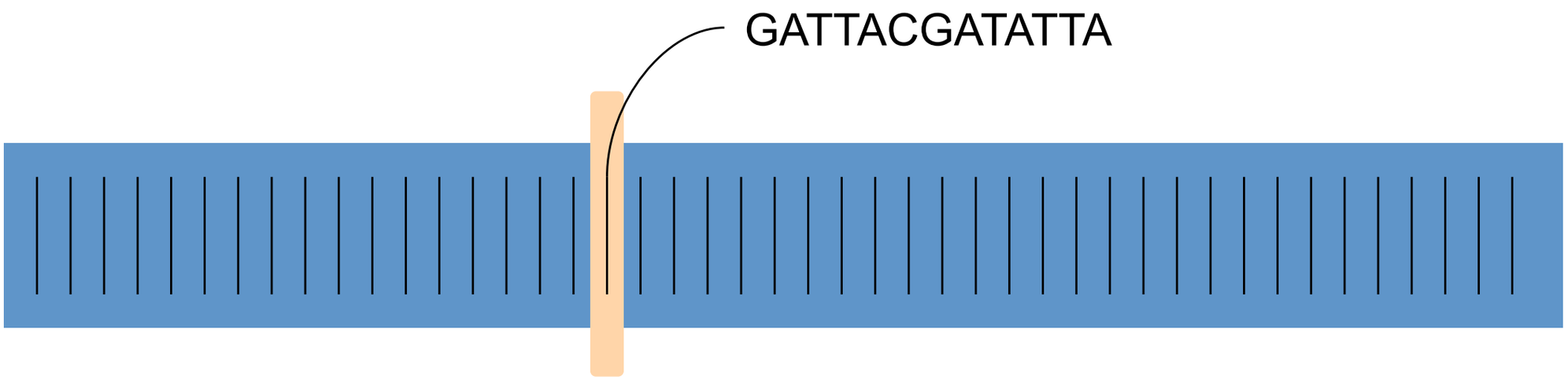
TACCTGCCGTAA = GATTACGATATTA?

Time = 1

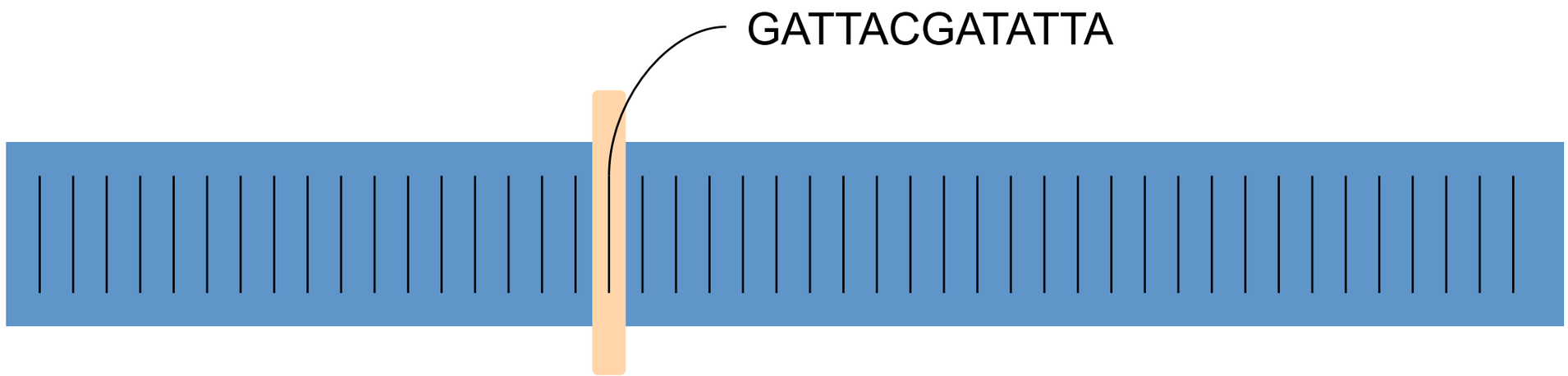


CCCCCAATGAC = GATTACGATATTA?

Time = 17



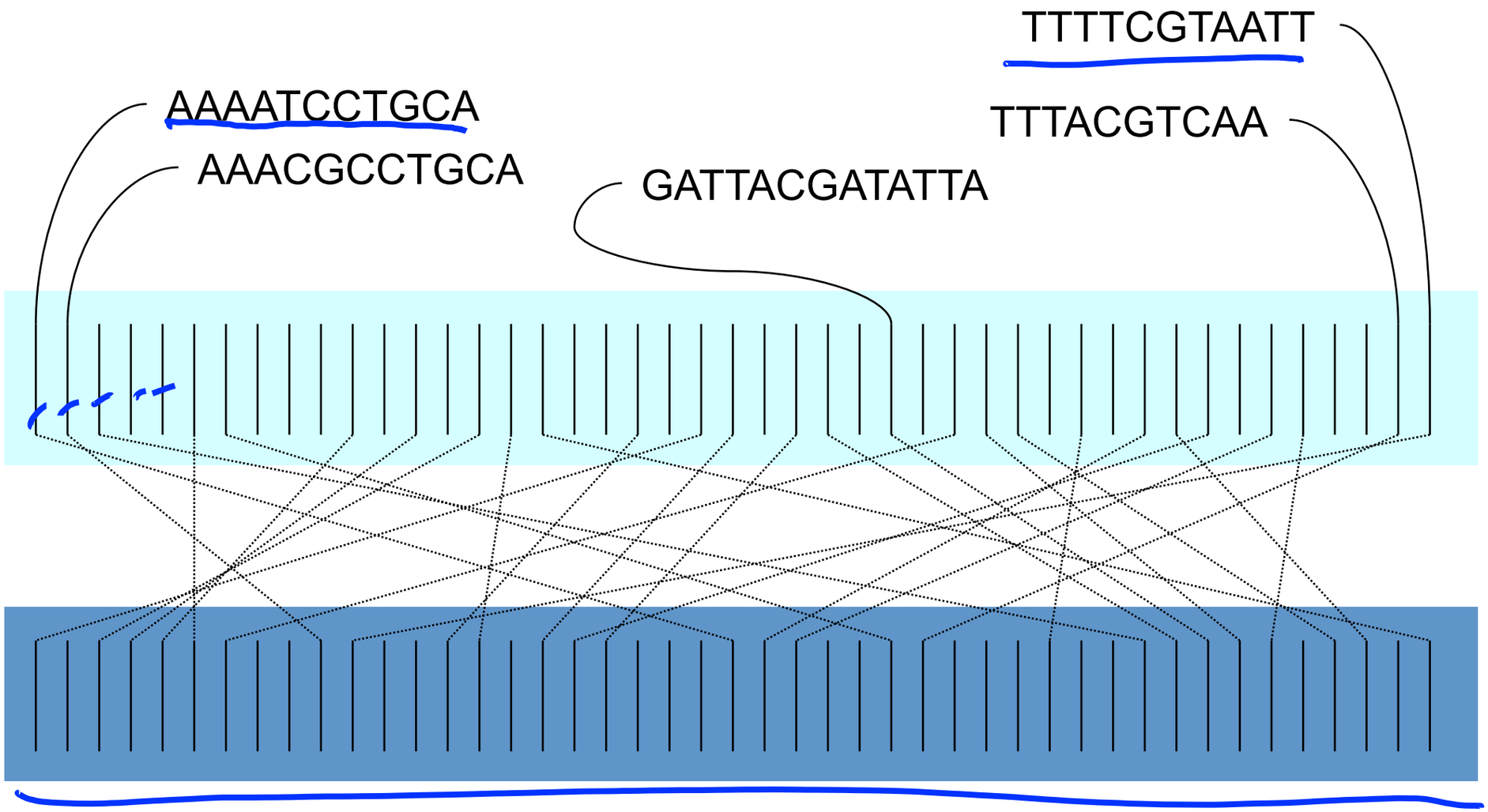
GATTACGATATTA contains GATTACGATATTA?



40 records, 40 comparions

N records, *N* comparisons

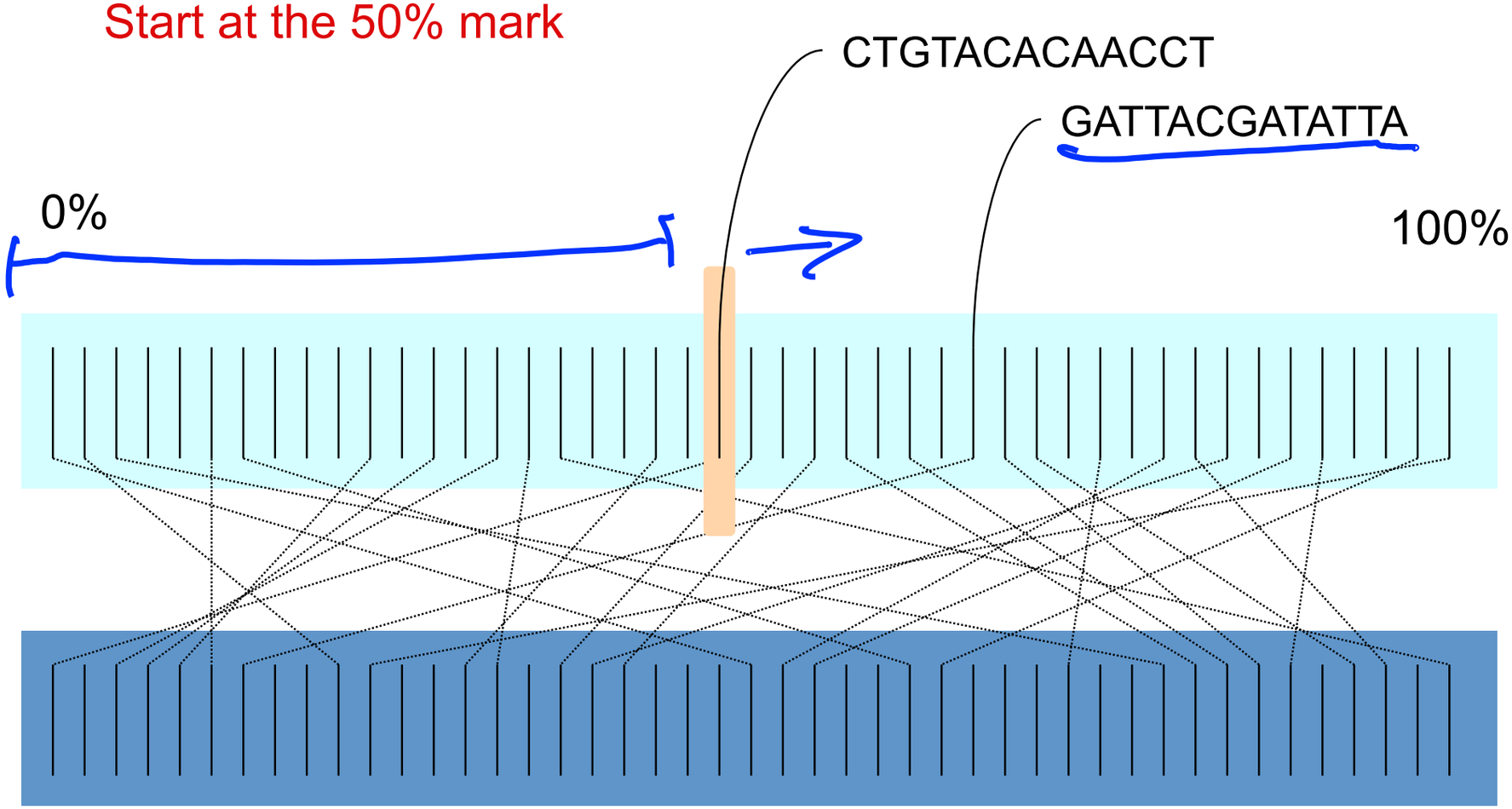
The algorithmic complexity is order N: O(*N*)



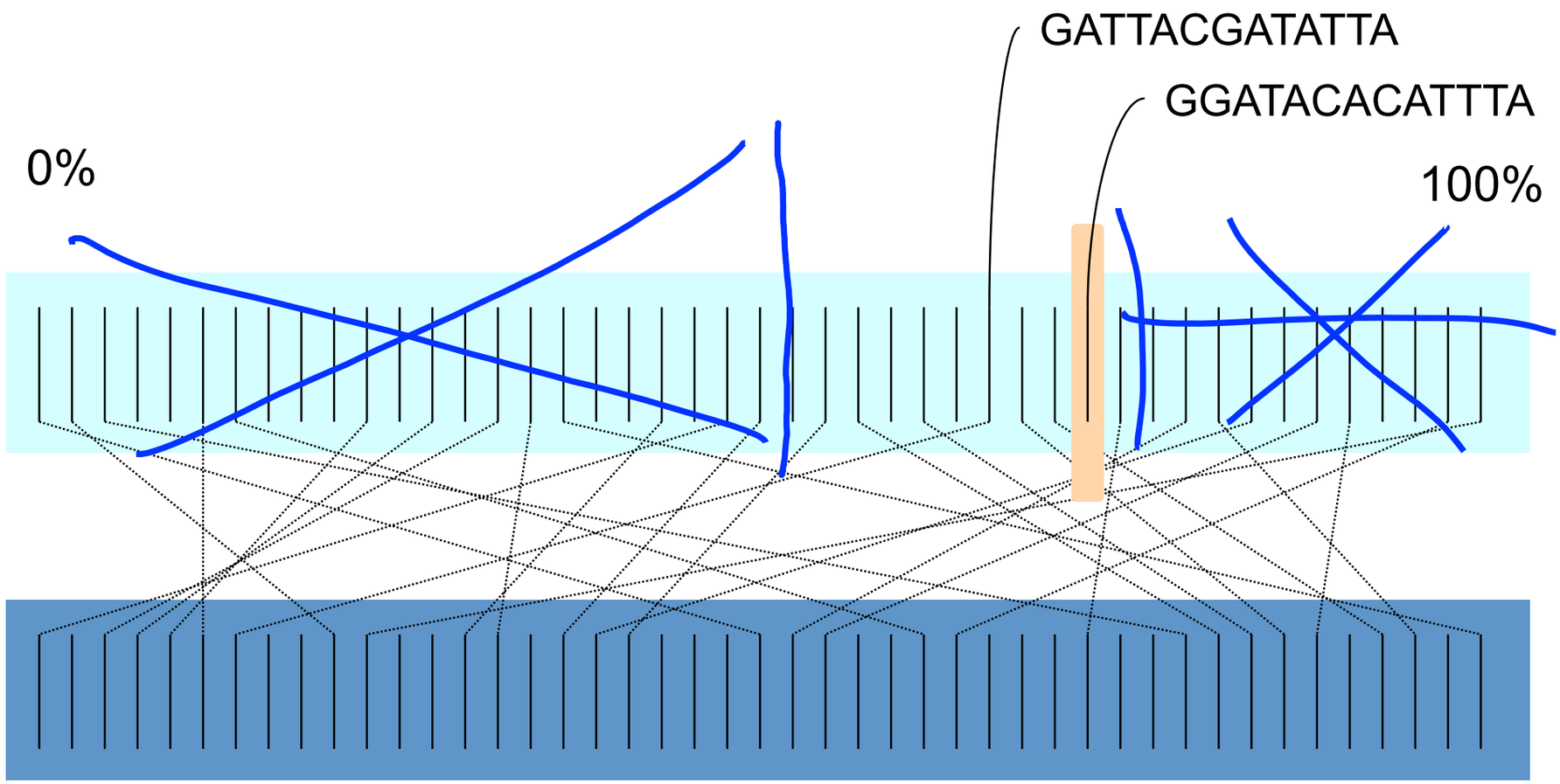
What if we sort the sequences ?

Time = 0

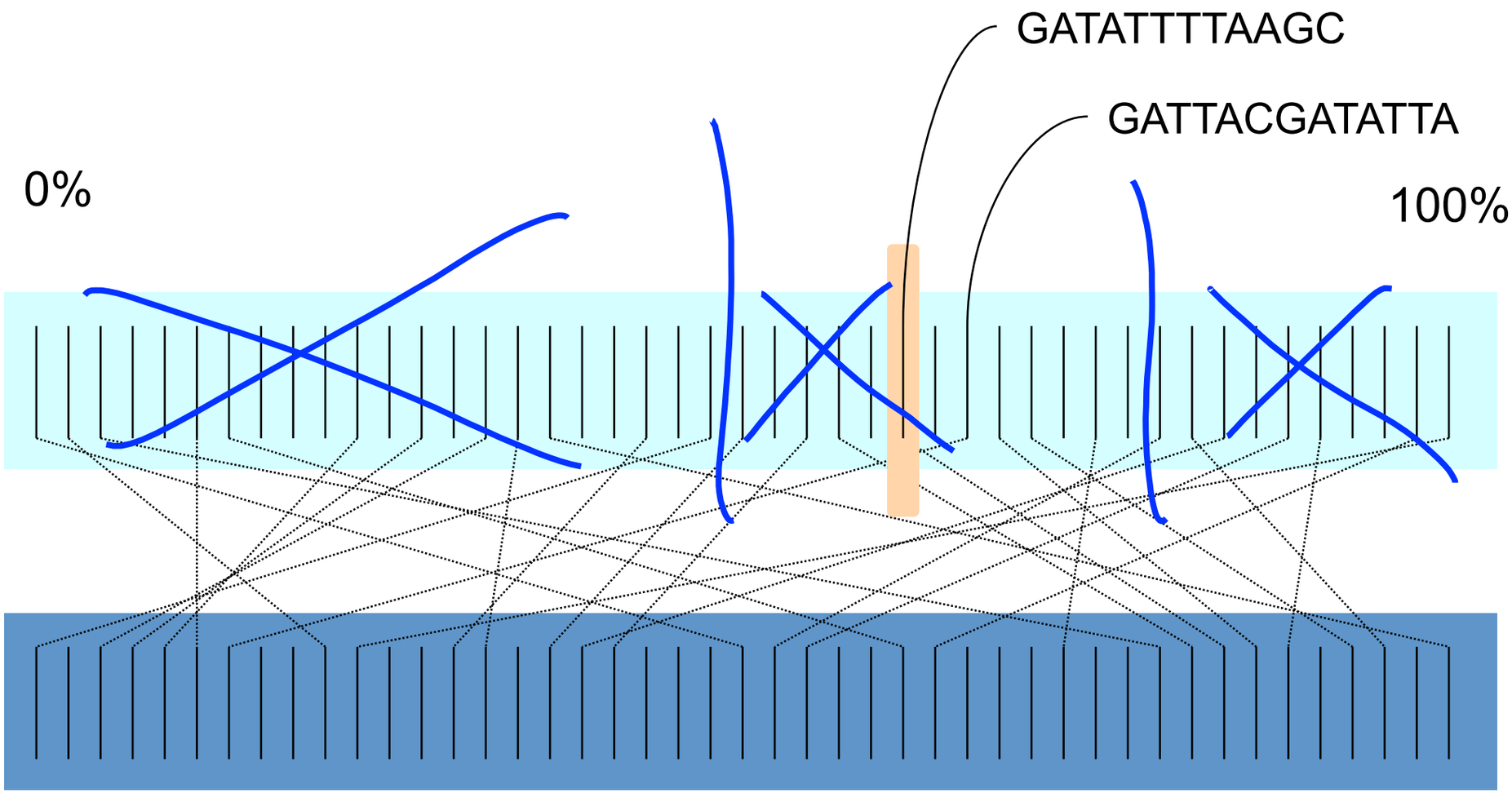
Start at the 50% mark



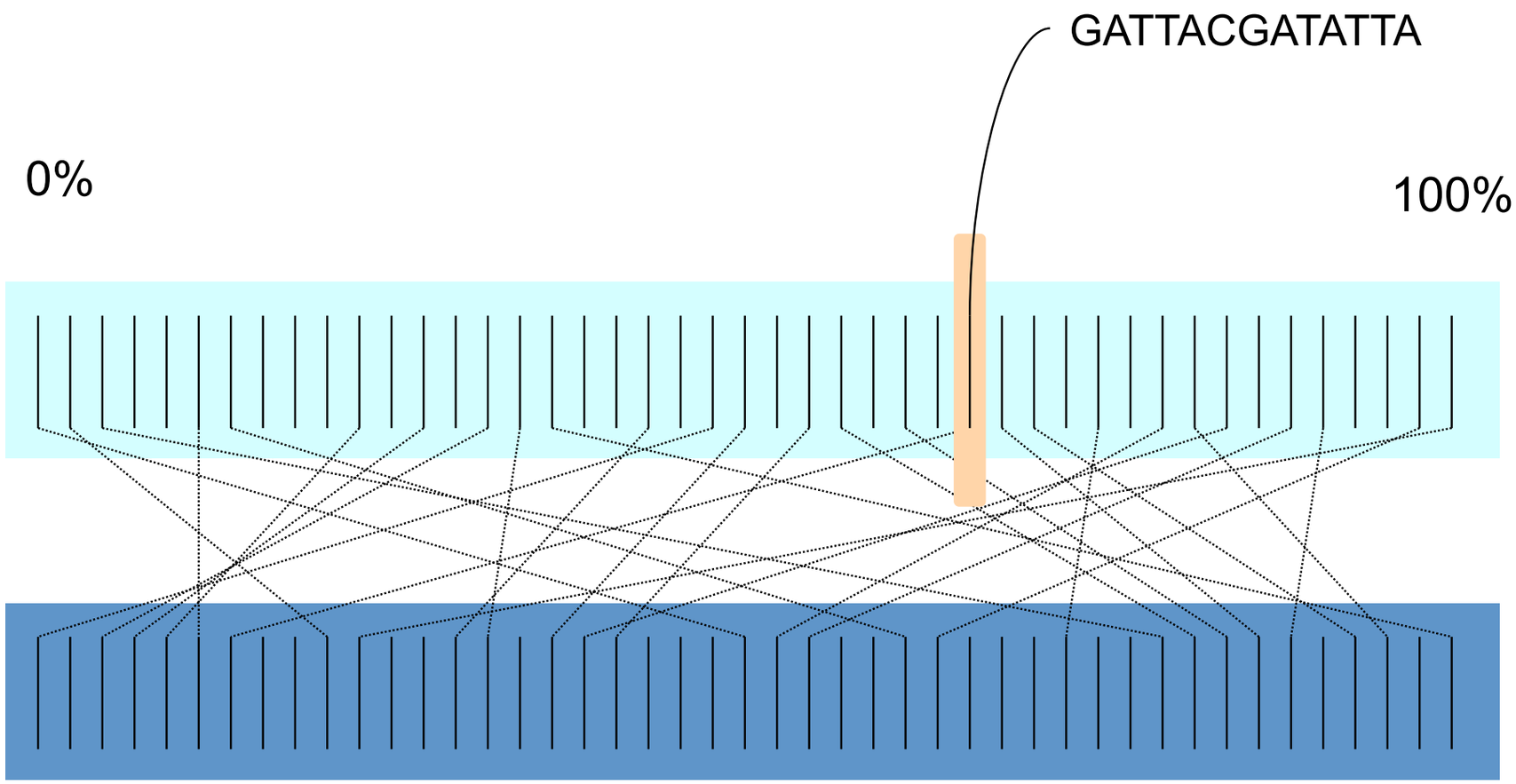
CTGTACACAACCT < GATTACGATATTA



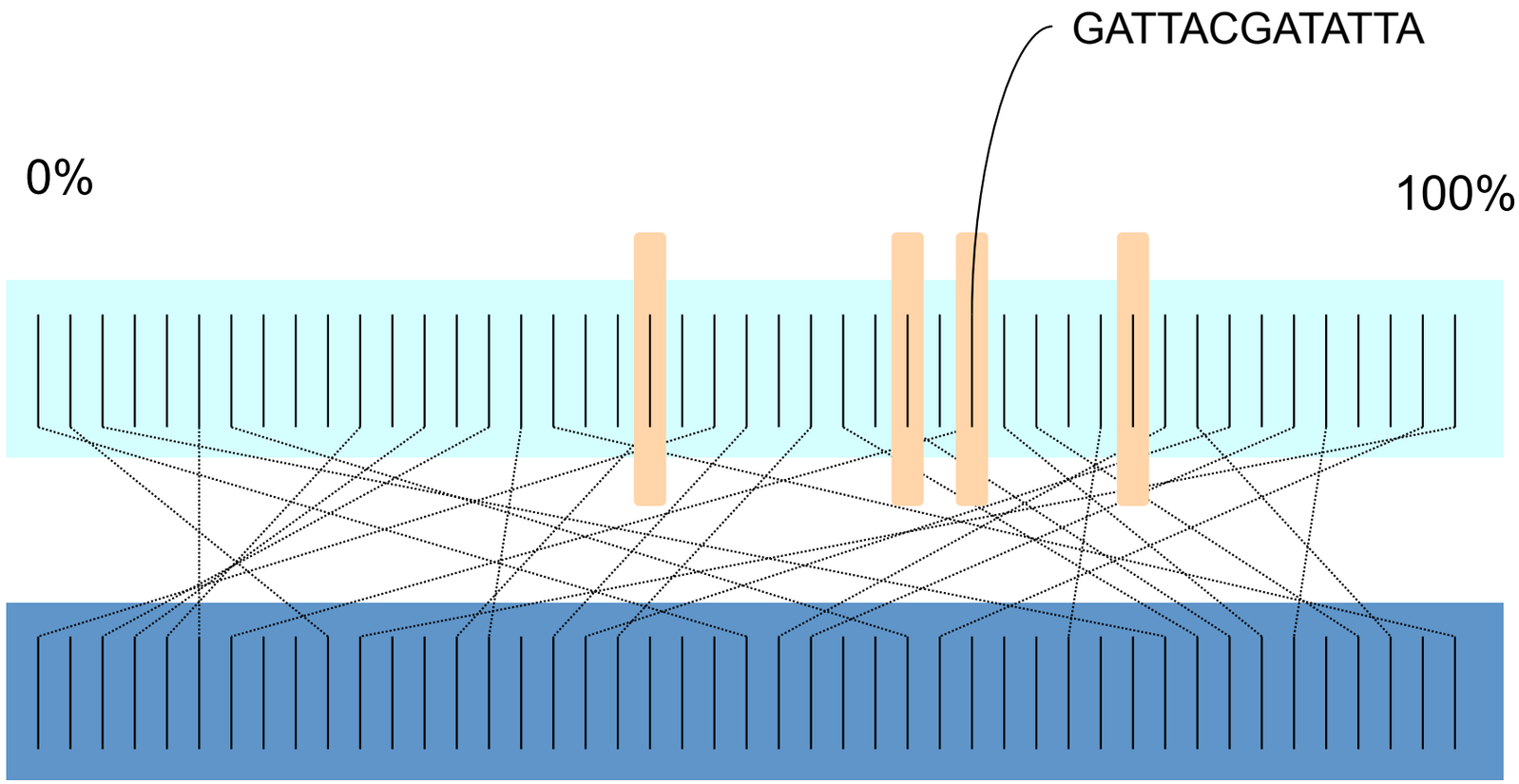
GGATACACATTTA > GATTACGATATTA



GATATTTTAAGC < GATTACGATATTA



GATTACGATATTA = GATTACGATATTA



How many comparisons did we do?

40 records, only 4 comparisons

N records, log(*N*) comparisons

This algorithm is O(log(*N*)) Far better scalability

Relational Databases

Databases are good at "Needle in Haystack" problems:

- **Extracting small results from big datasets**
- **Transparently provide "old style" scalability**
- **Your query will always* finish, regardless of dataset size.**

- **Indexes are easily built and automatically used when appropriate**

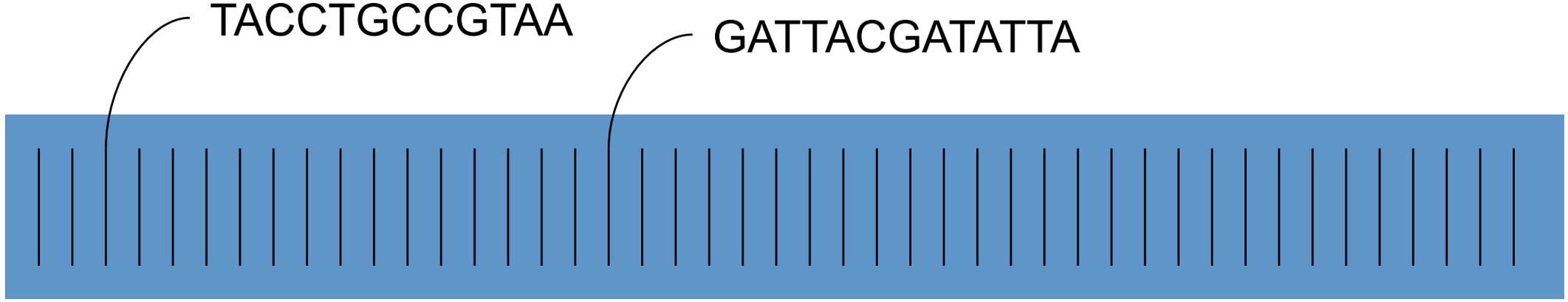
```
CREATE INDEX seq_idx ON sequence(seq);

SELECT seq
  FROM sequence
 WHERE seq = 'GATTACGATATTA';
```

Example

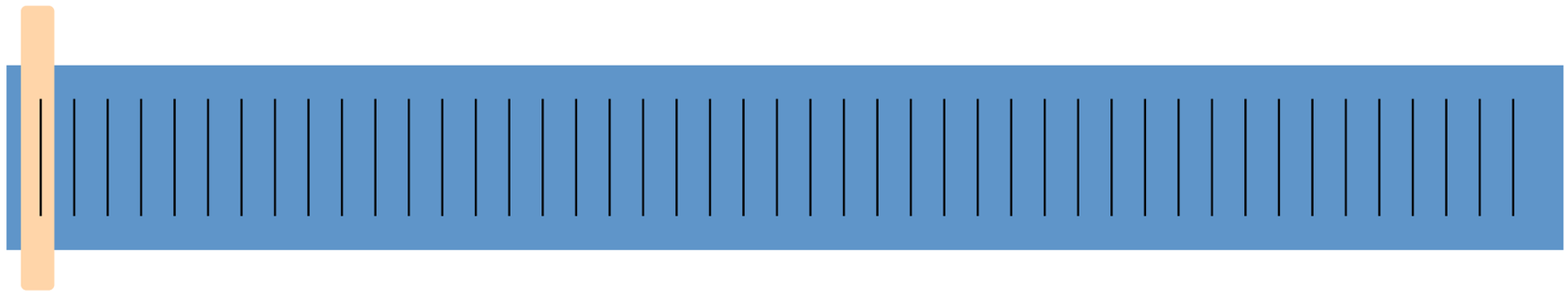
New task: Read Trimming

Given a set of DNA sequences,
trim the final n bps of each sequence

Generate a new dataset

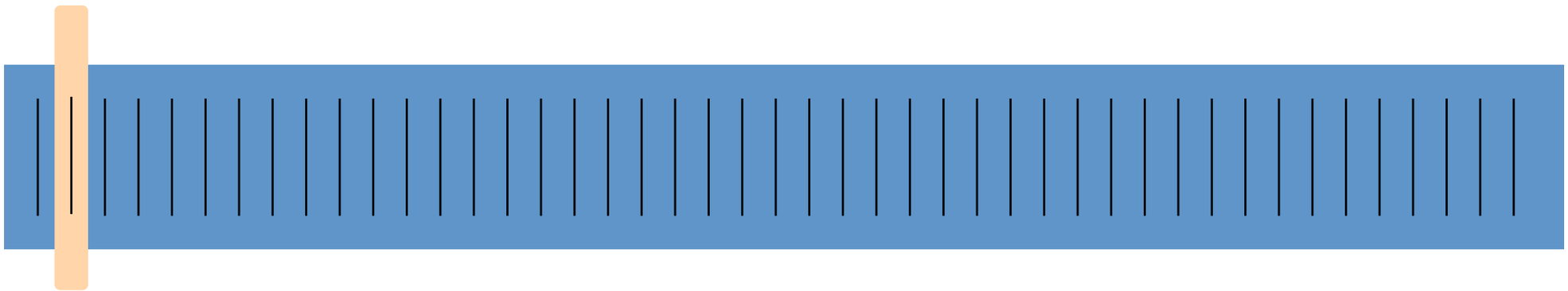

TACCTGCCGTAA
GATTACGATATTA

Time = 0



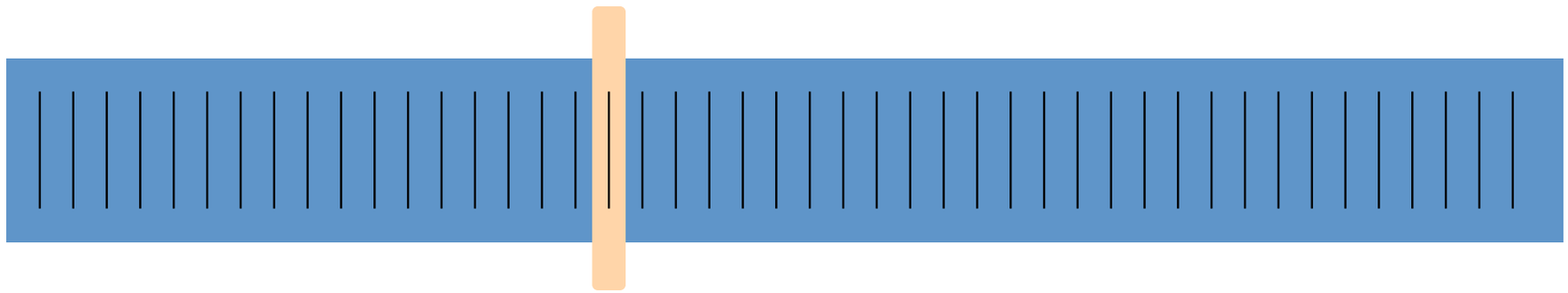
TACCTGCCGTAA becomes TACCT

Time = 1

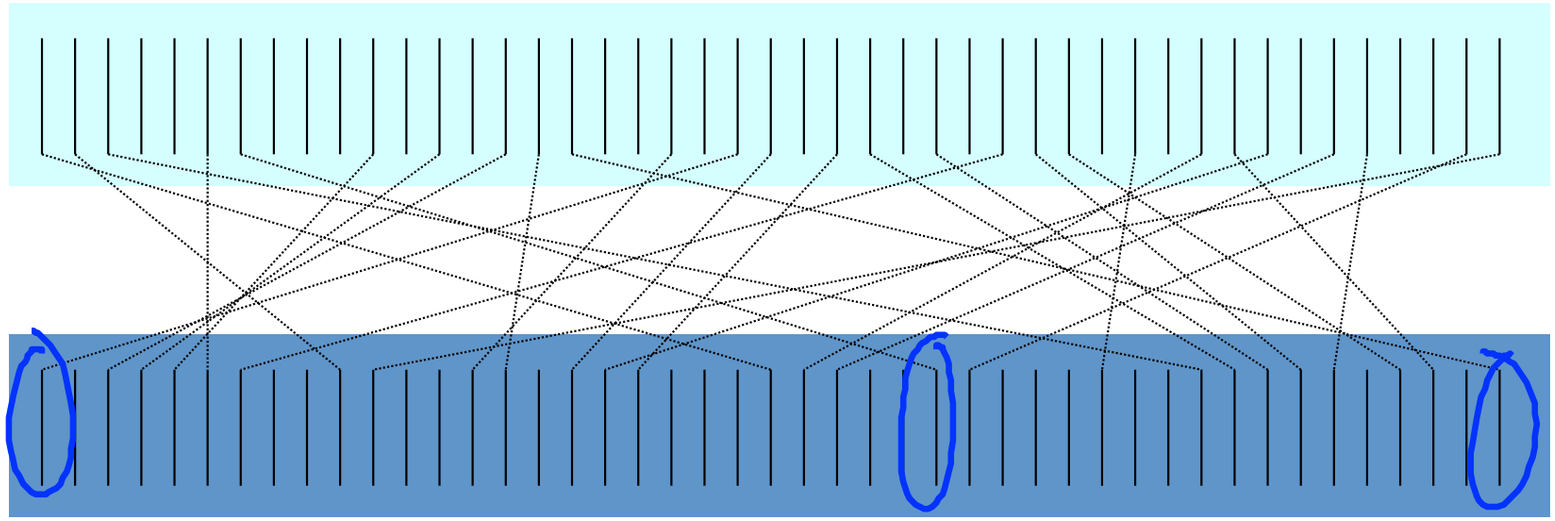


CCCCCAATGAC becomes CCCCC

Time = 17



GATTACGATATTA becomes GATTA

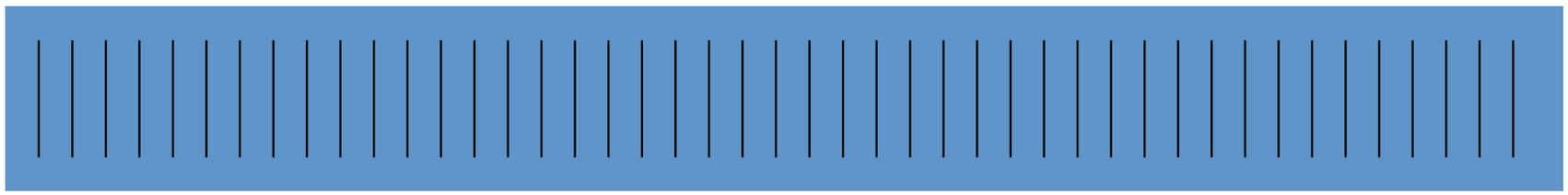


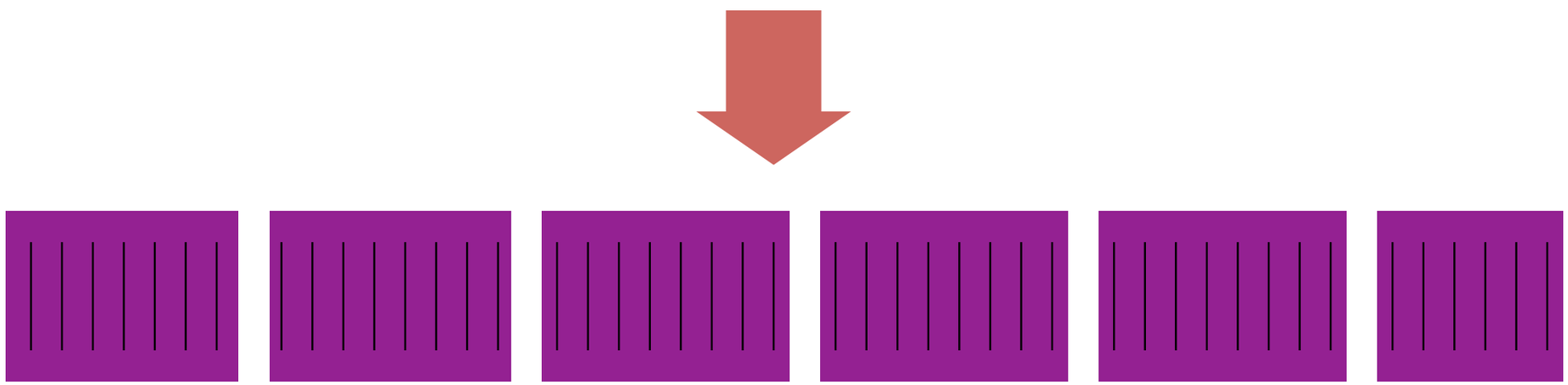
Can we use an index?

No. We have to touch every record no matter what.

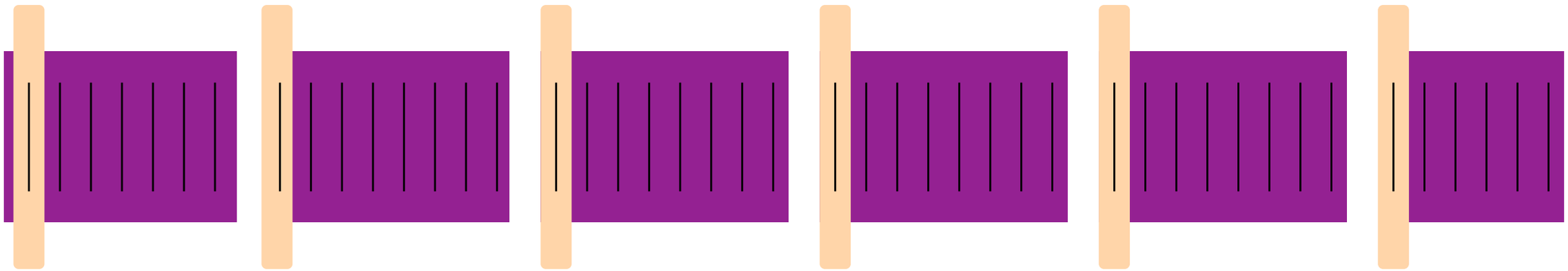
The task is fundamentally O(*N*)

Can we do any better?

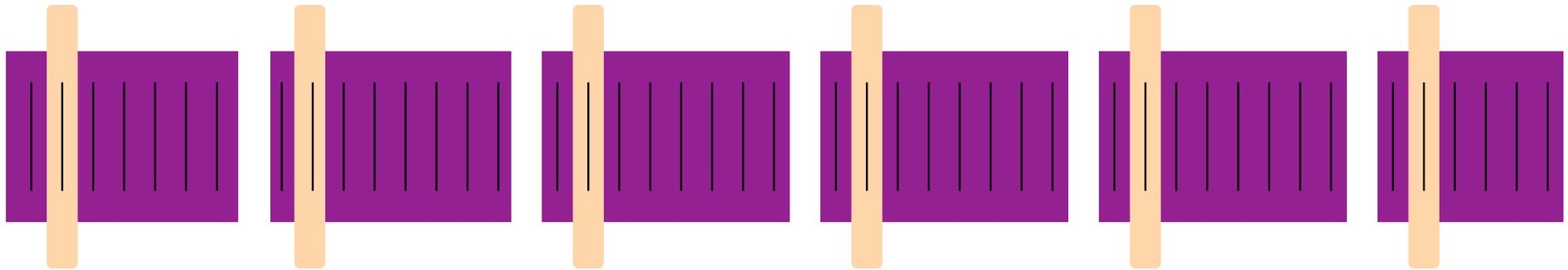




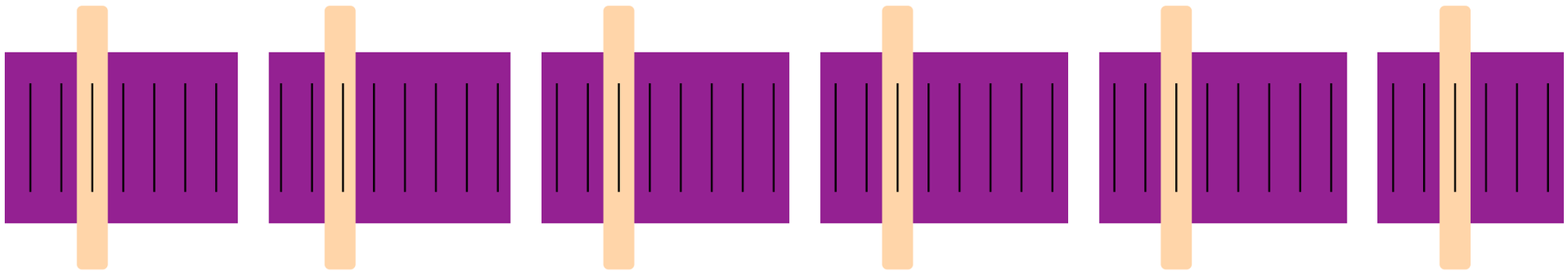
Time = 0



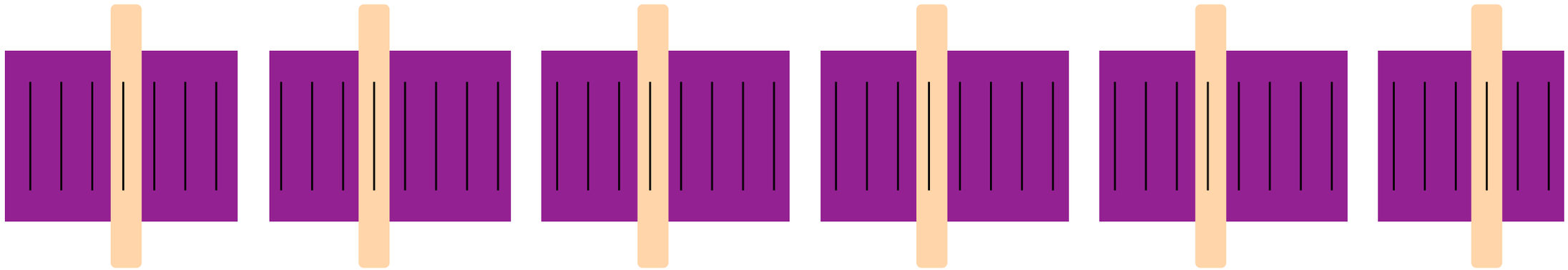
Time = 1



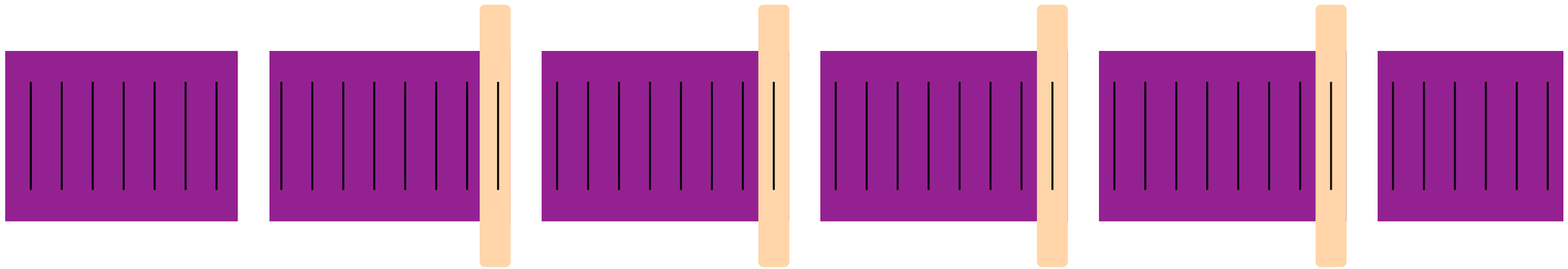
Time = 2



Time = 3



Time = 7

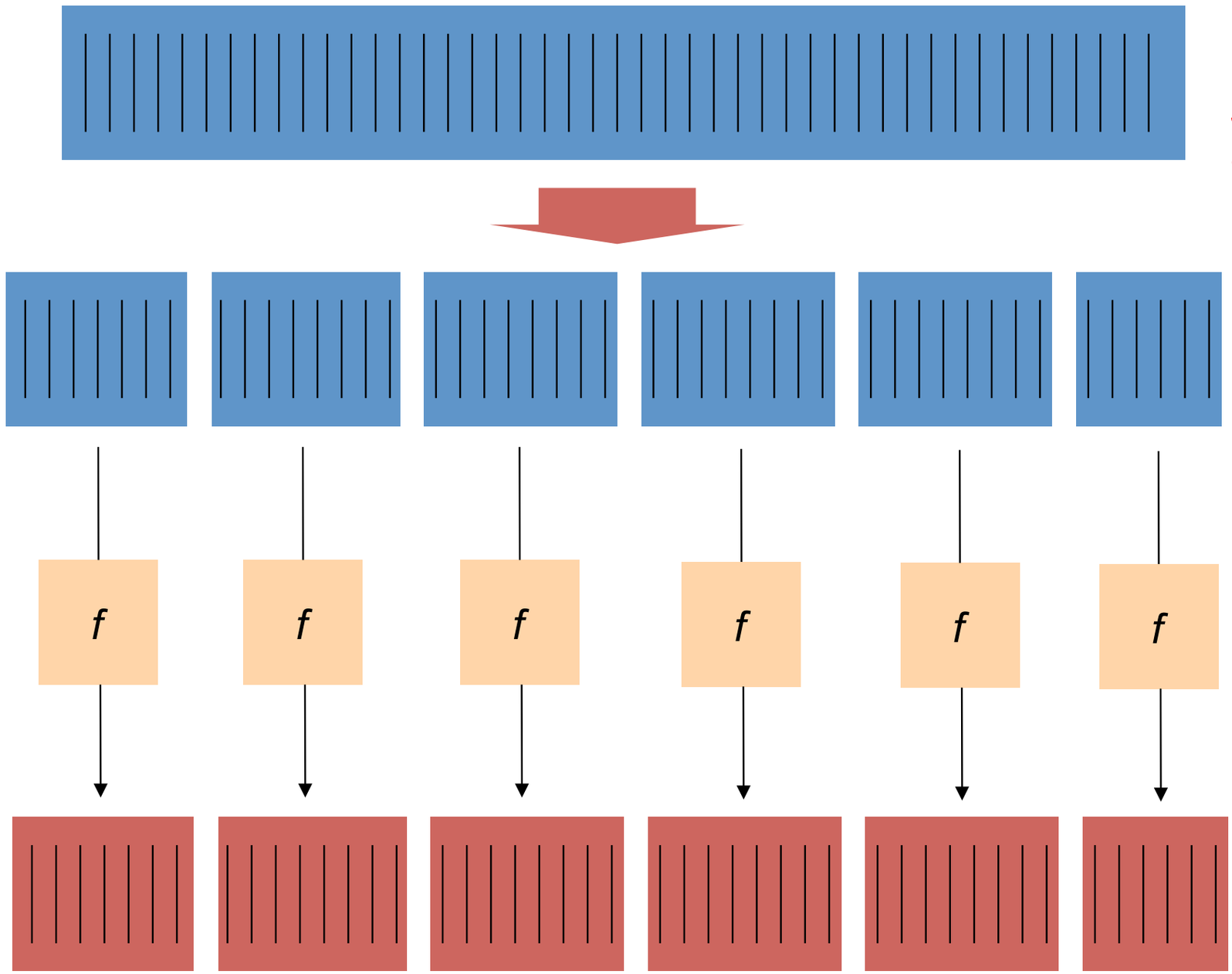


How much time did this take?

7 cycles

40 records, 6 workers

O(*N*/*k*)



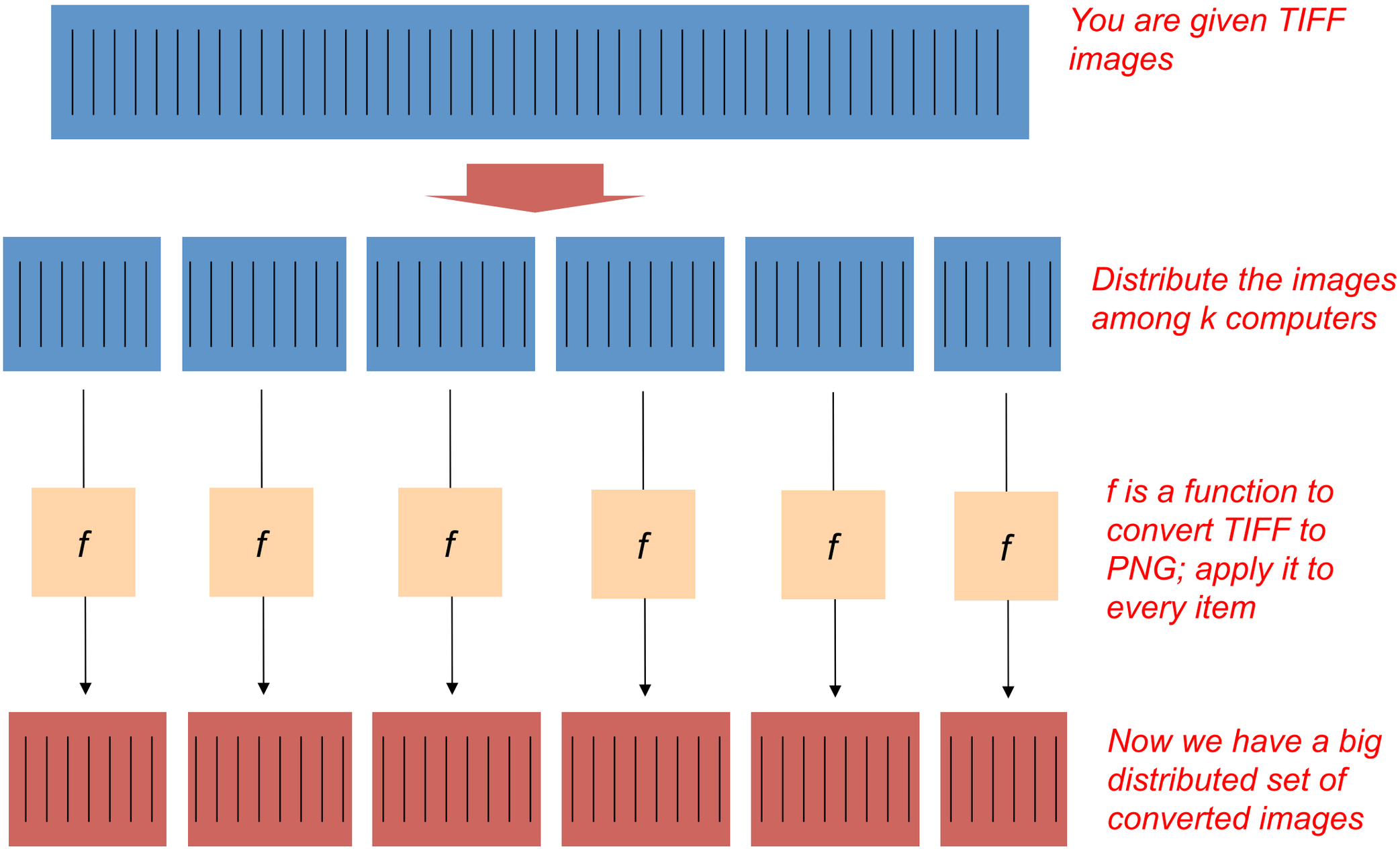
You are given short "reads": genomic sequences about 35-75 characters each

Distribute the reads among k computers

f is a function to trim a read; apply it to every item

Now we have a big distributed set of trimmed reads

Convert 405k TIFF images to PNG



You are given TIFF images

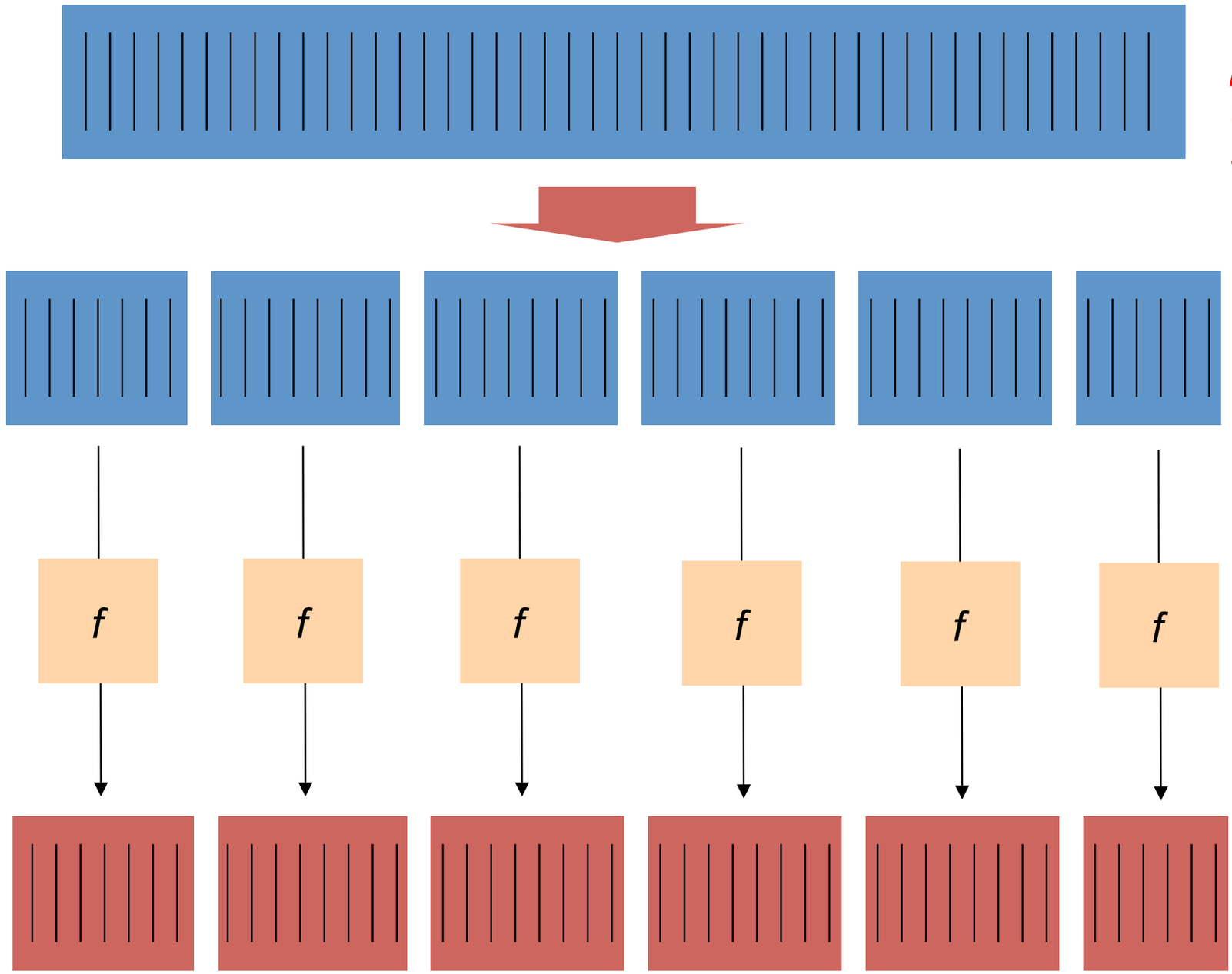
Distribute the images among k computers

f is a function to convert TIFF to PNG; apply it to every item

Now we have a big distributed set of converted images

http://open.blogs.nytimes.com/2008/05/21/the-new-york-times-archives-amazon-web-services-timesmachine/

Run thousands of simulations



You have sets of parameters for thousands of small simulations

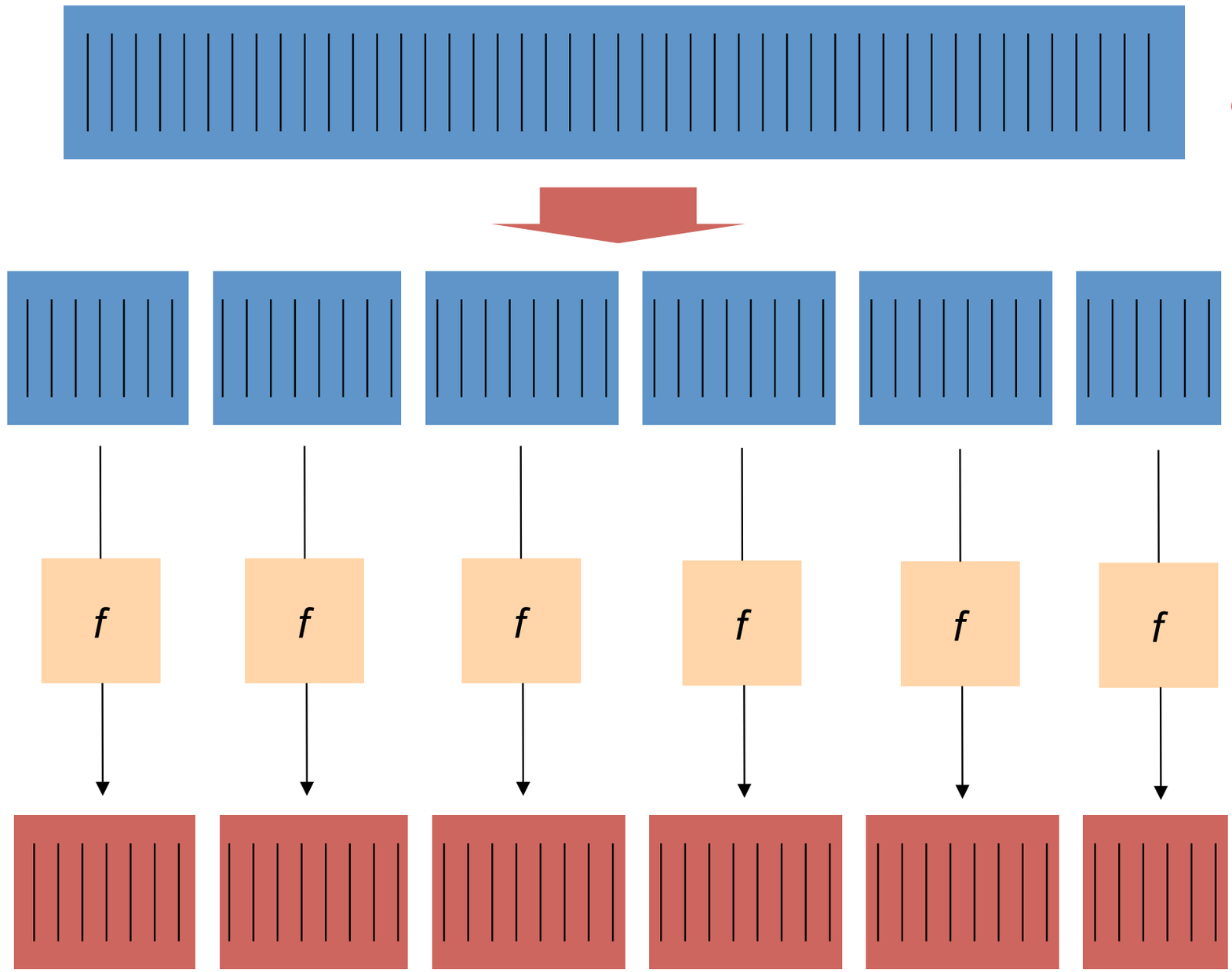
Divide the parameter sets among k computers

f is runs the simulation and produces some output; apply it to every item

Now we have a big distributed set of simulation results

http://escience.washington.edu/get-help-now/dave-williams-simulating-muscle-dynamics-cloud

Find the most common word in each document



You have millions of documents

Distribute the documents among k computers

f finds the most common word in a single document

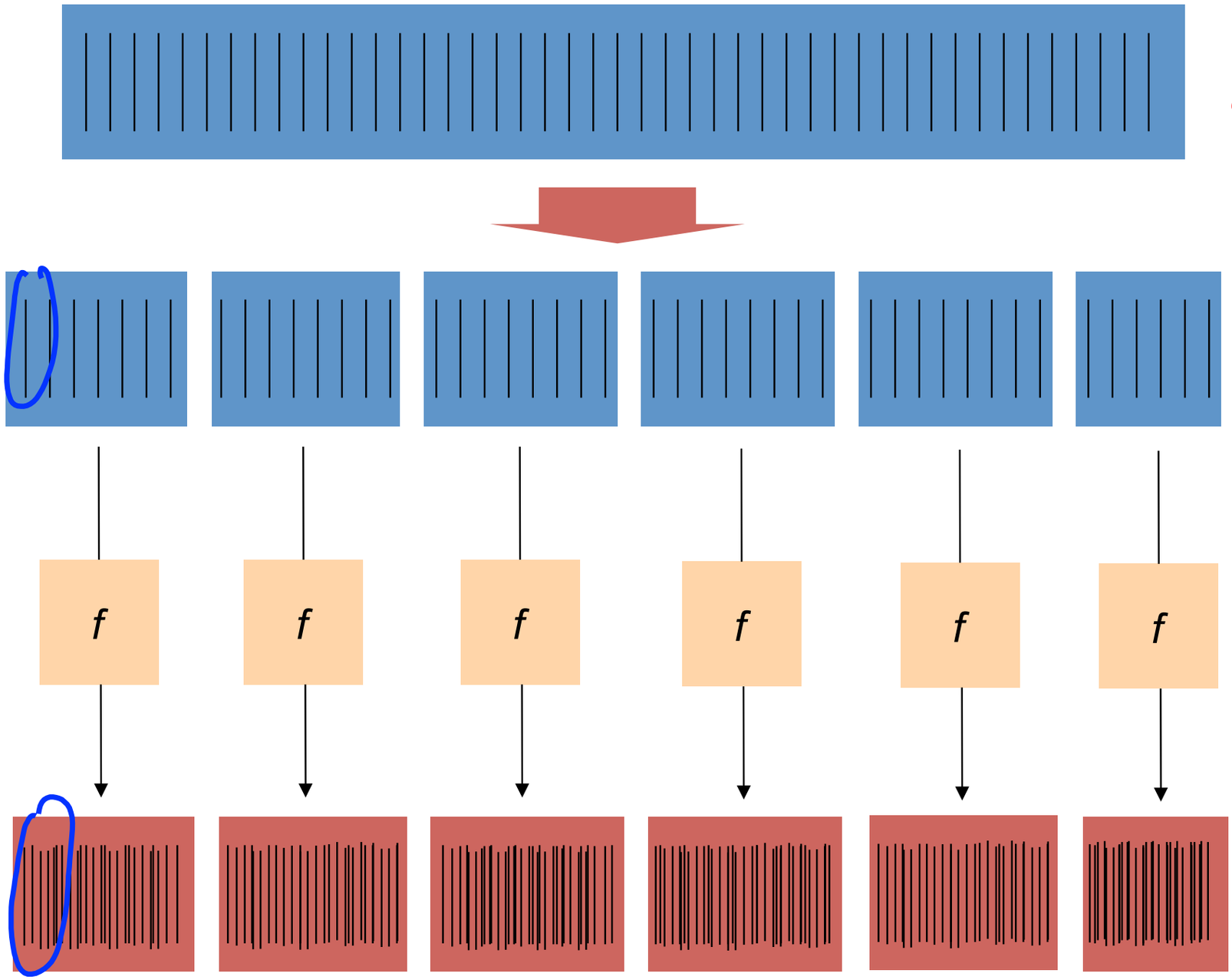
Now we have a big distributed list of (doc_id, word) pairs

Consider a slightly more general program to compute the word frequency of every word in a singe document

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

(people, 2)
(government, 6)
(assume, 1)
(history, 2)
…



You have millions of documents

Distribute the documents among k computers

For each document f returns a set of (word, freq) pairs

Now we have a big distributed list of sets of word freqs.

There's a pattern...

A function that *maps* a read to a trimmed read

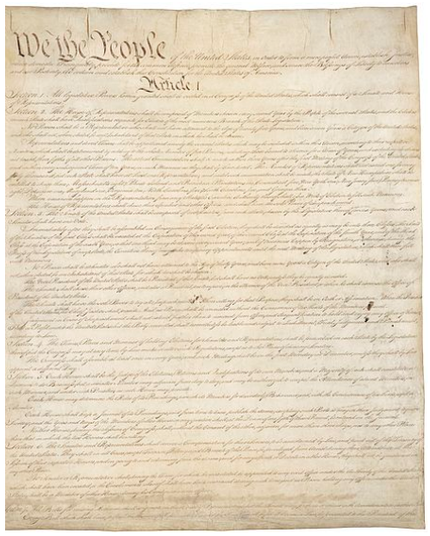
A function that *maps* a TIFF image to a PNG image

A function that *maps* a set of parameters to a simulation result

A function that *maps* a document to its most common word

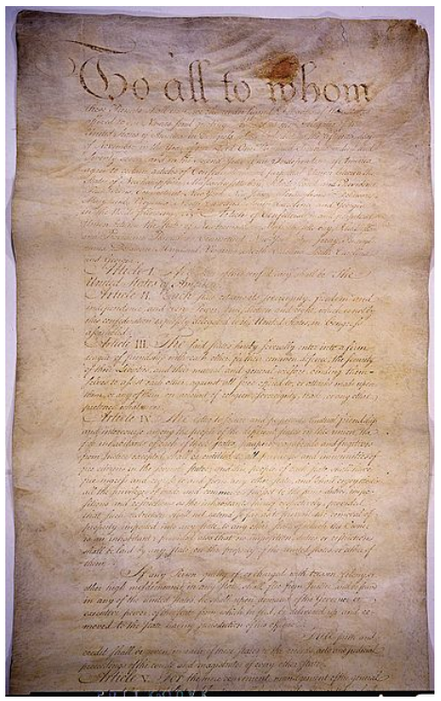
A function that *maps* a document to a histogram of word frequencies

What if we want to compute the word frequency across *all* documents?



US Constitution

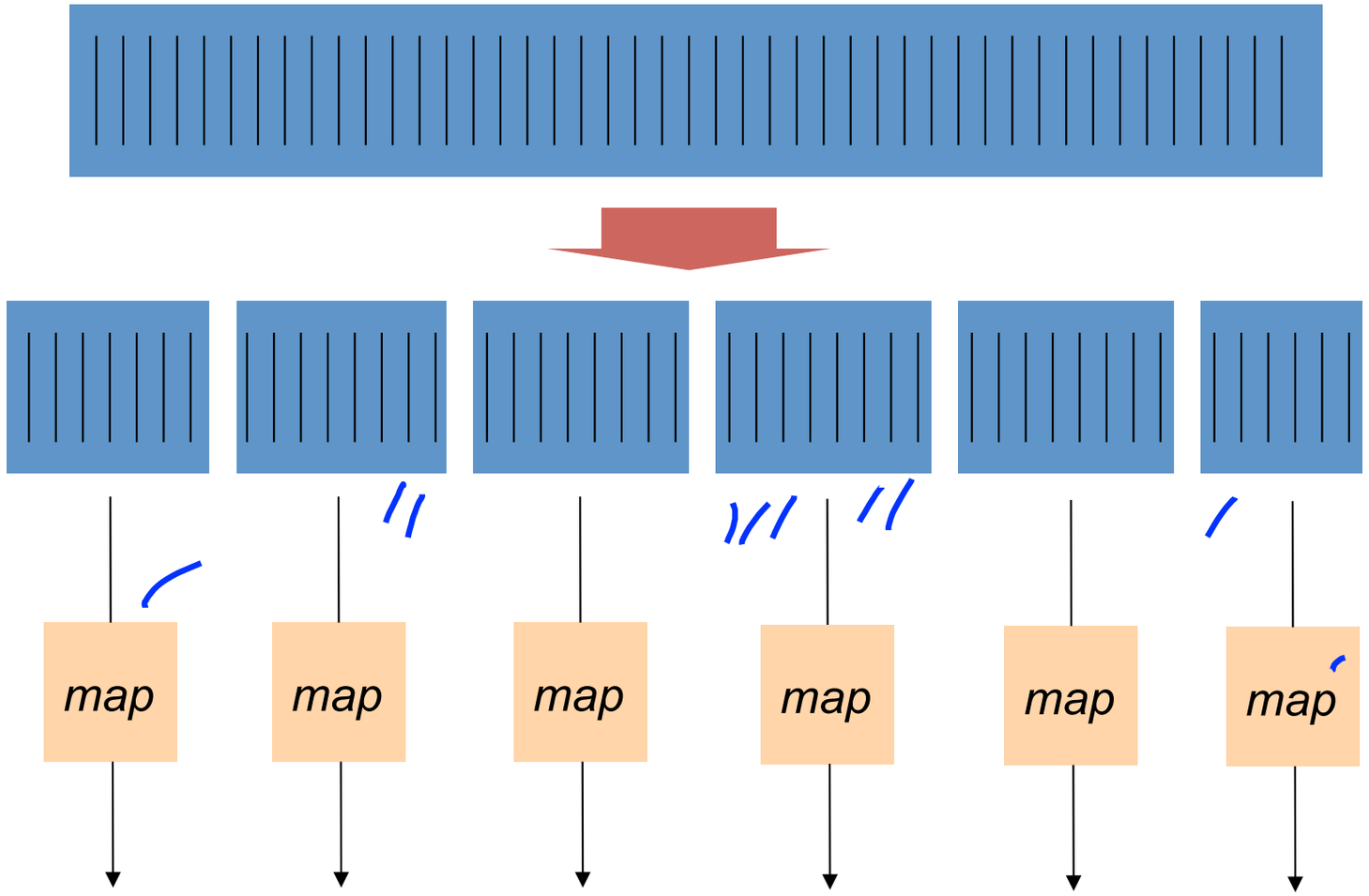
Declaration of Independence



Articles of Confederation

(people, 78)
(government, 123)
(assume, 23)
(history, 38)
…

Compute the word frequency across 5M docs



You have millions of documents

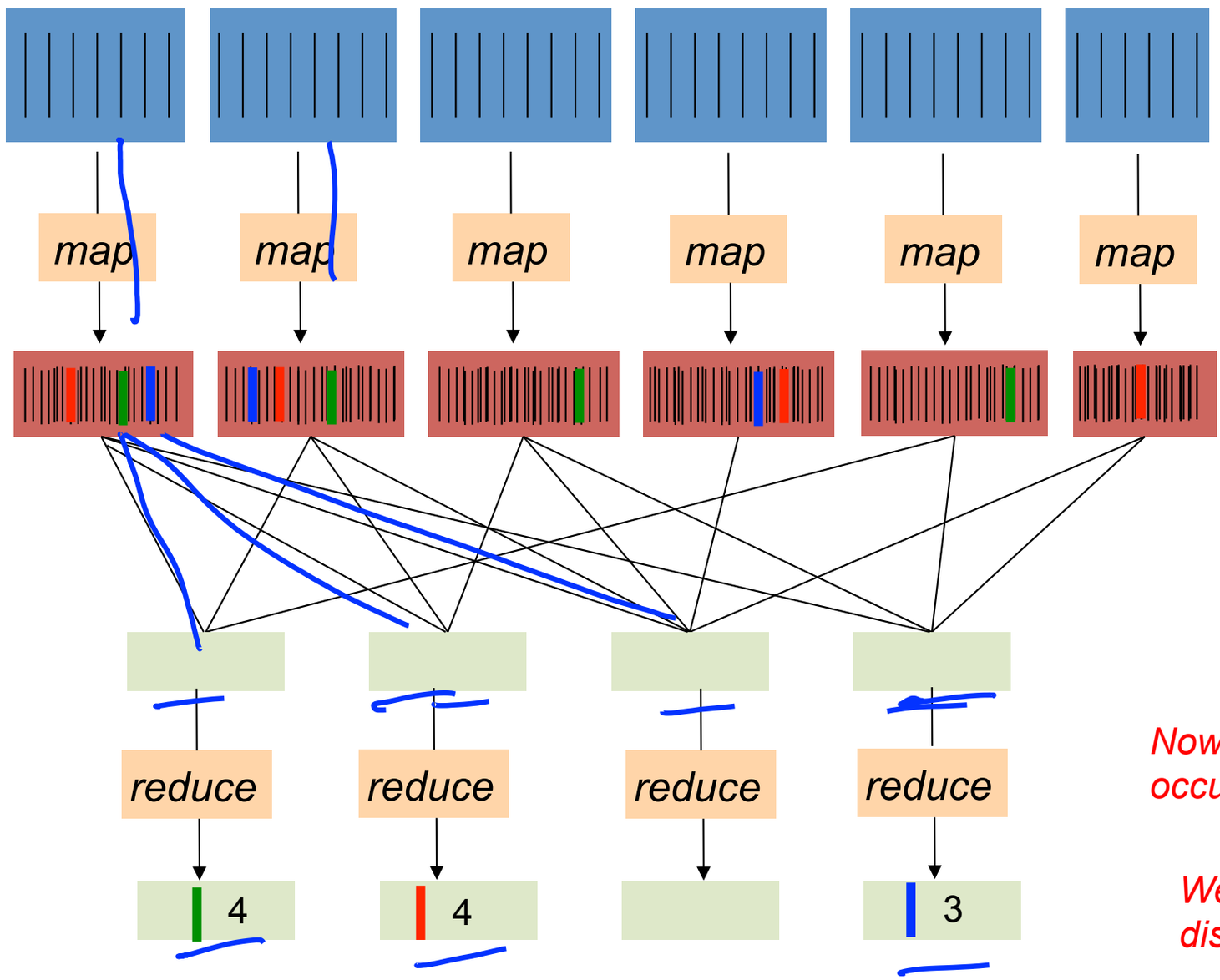
Distribute the documents among k computers

For each document, return a set of (word, freq) pairs

Now what?

- *But we don't want a bunch of little histograms – we want one big histogram.*
- *How can we make sure that a single computer has access to every occurrence of a given word regardless of which document it appeared in?*

Compute the word frequency across 5M docs



Distribute the documents among k computers

For each document, return a set of (word, freq) pairs

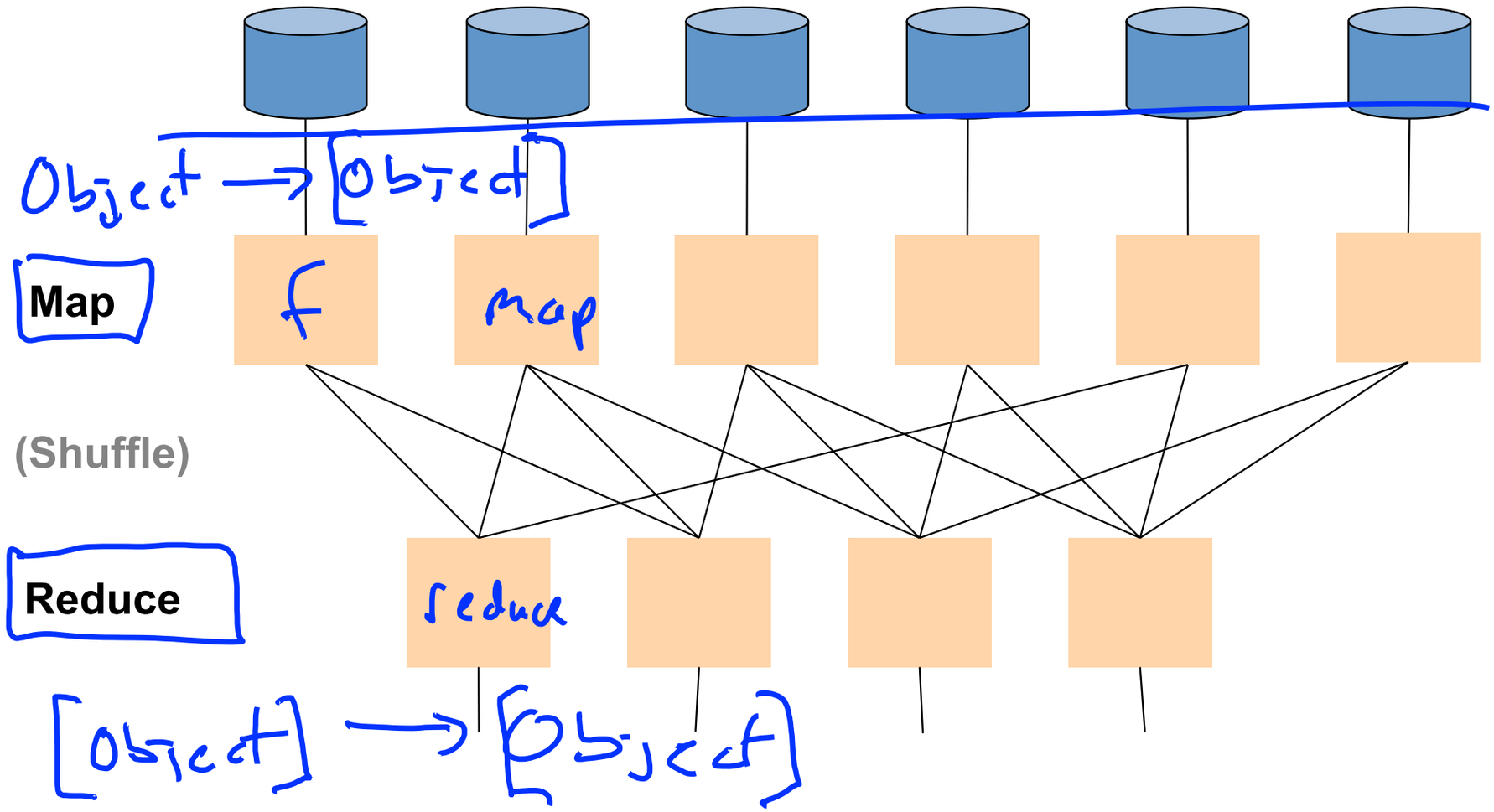
Now we have a big distributed list of sets of word freqs.

Now just count the occurrences of each word

We have our distributed histogram

Compute the word frequency across 5M docs

Some distributed algorithm…



Compute the word frequency across 5M docs

MapReduce Programming Model

- Input & Output: each a set of key/value pairs
- Programmer specifies two functions:

map (in_key, in_value) -> list(out_key, intermediate_value)

- Processes input key/value pair
- Produces set of intermediate pairs

reduce (out_key, list(intermediate_value)) -> list(out_value)

- Combines all intermediate values for a particular key
- Produces a set of merged output values (usually just one)

Compute the word frequency across 5M docs

```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
   for each word w in input_value:
       EmitIntermediate(w, 1);
```

```
reduce(String output_key, Iterator intermediate_values):
  // output_key: word
  // output_values: ????
   int result = 0;
   for each v in intermediate_values:
       result += v;
   Emit(result);
```

Example

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

Word length histogram example

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

How many “big”, “medium”, and “small” words are used?

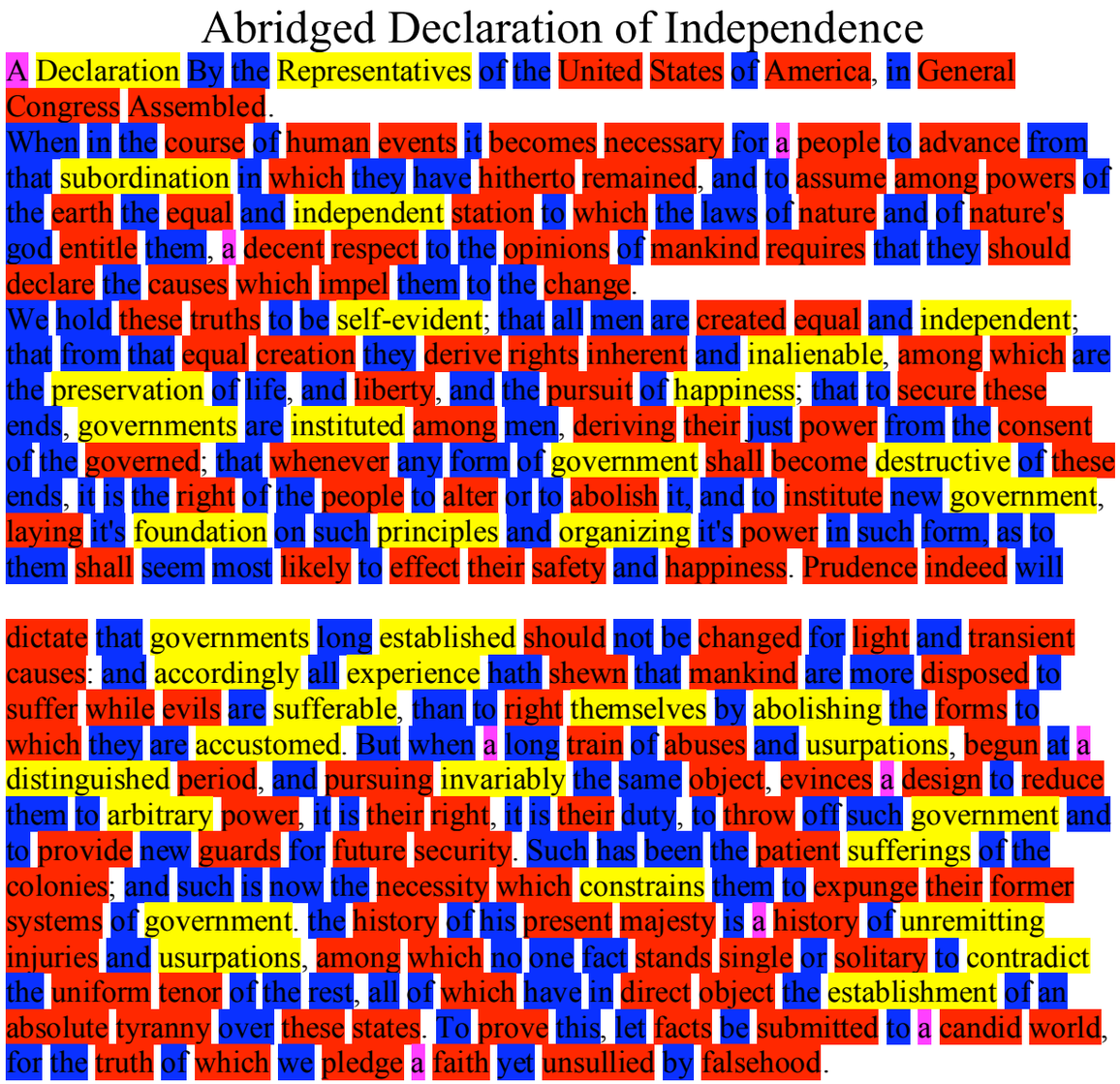
Word length histogram example

Big = Yellow = 10+ letters

Medium = Red = 5..9 letters

Small = Blue = 2..4 letters

Tiny = Pink = 1 letter



Word length histogram example

Split the document into chunks and process each chunk on a different computer

Chunk 1

Chunk 2

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

Word length histogram example

Abridged Declaration of Independence

Map Task 1
(204 words)

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

(key, value)

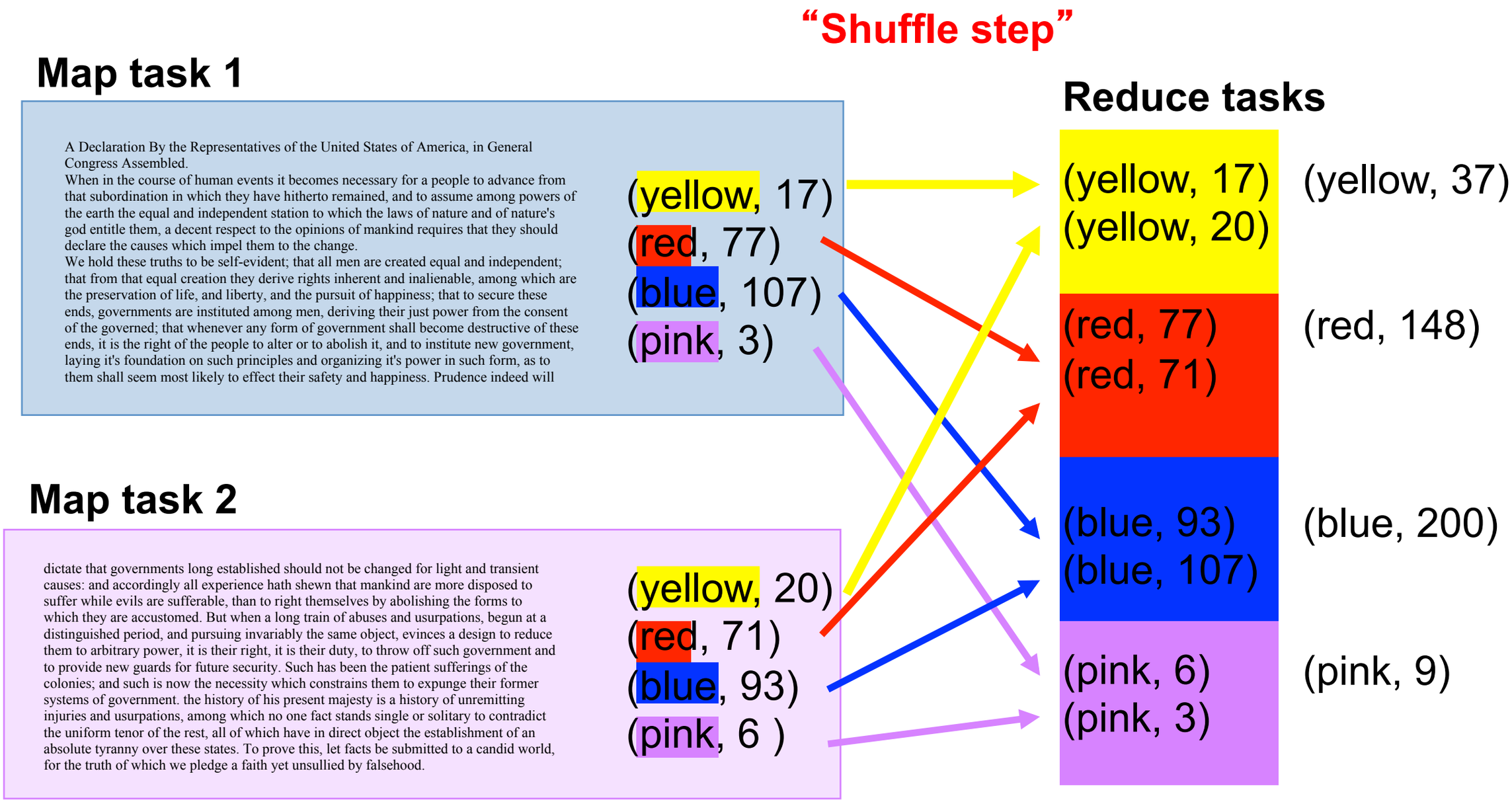
(yellow, 17)
(red, 77)
(blue, 107)
(pink, 3)

Map Task 2
(190 words)

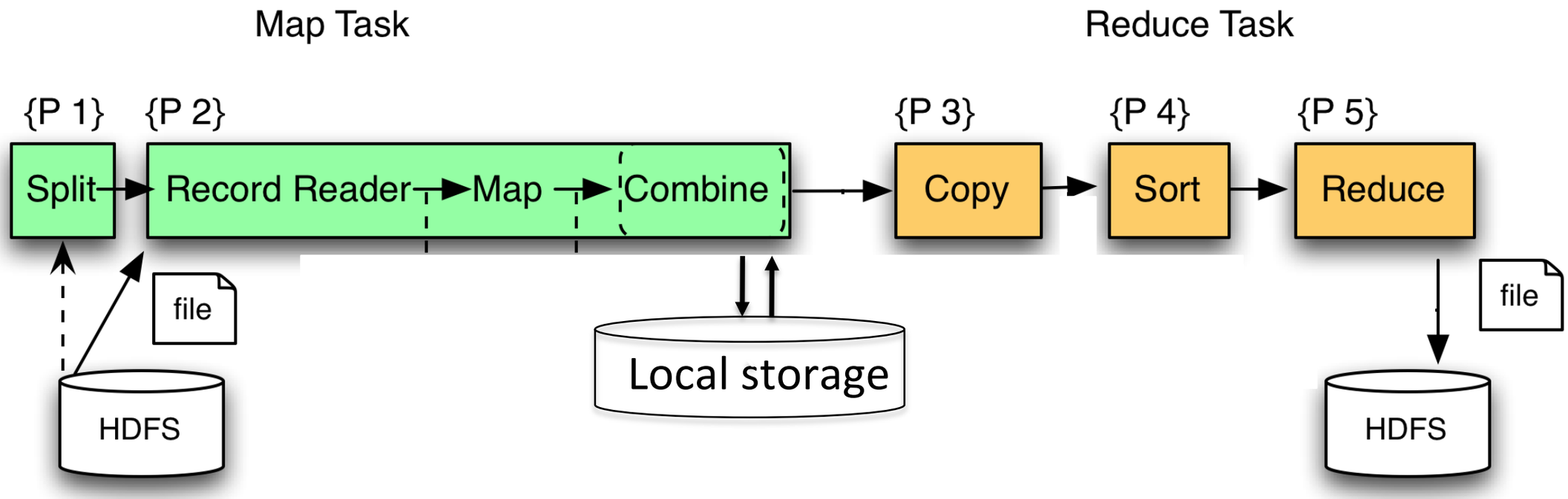
dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

(yellow, 20)
(red, 71)
(blue, 93)
(pink, 6)

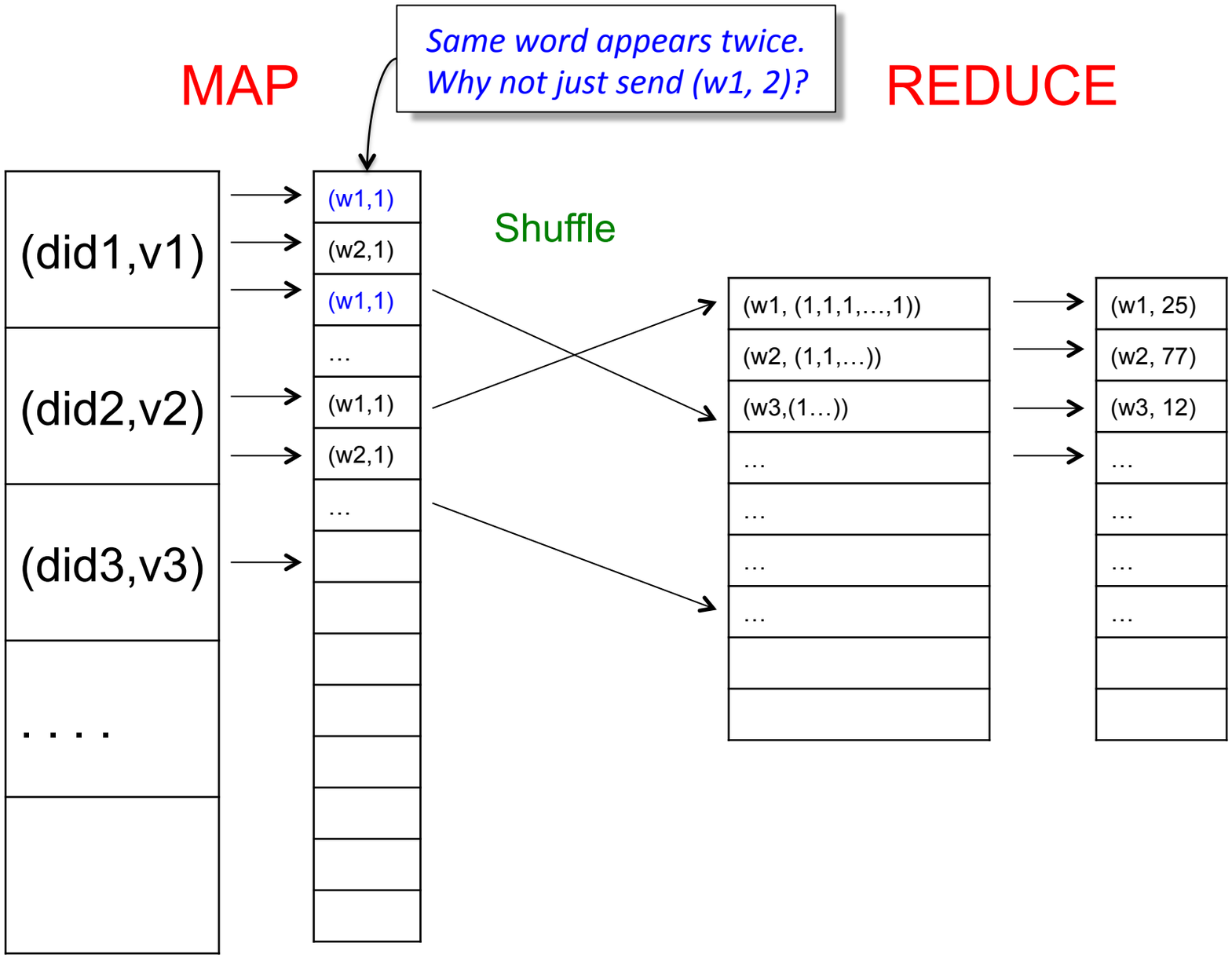
Word length histogram example



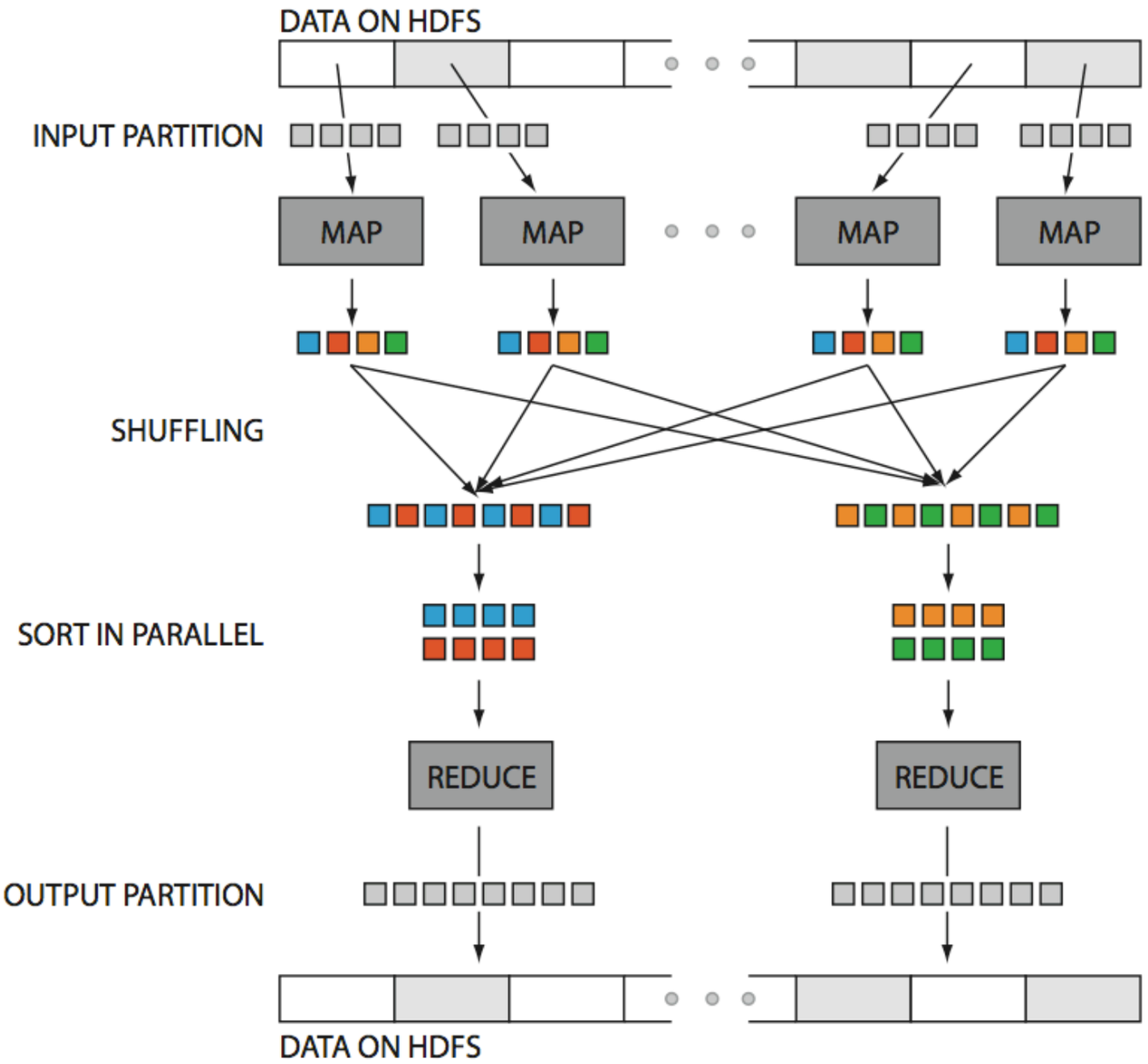
MR Phases


Map Task
Reduce Task
{P 1}
{P 2}
{P 3}
{P 4}
{P 5}
Split
Record Reader
Map
Combine
Copy
Sort
Reduce
file
HDFS
Local storage
file
HDFS

MR Phases

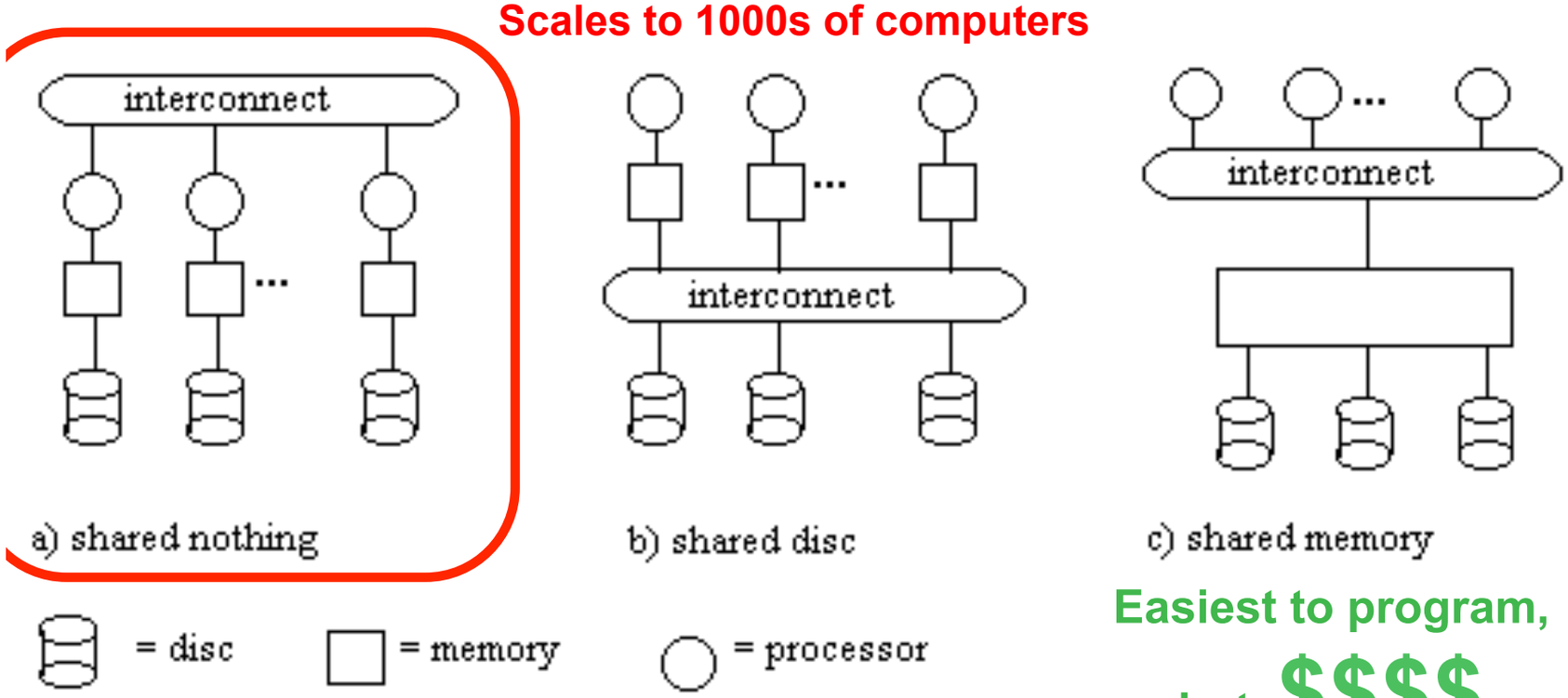

MAP
Same word appears twice.
Why not just send (w1, 2)?
REDUCE
(did1,v1)
(did2,v2)
(did3,v3)
. . . .
(w1,1)
(w2,1)
(w1,1)
…
(w1,1)
(w2,1)
…
Shuffle
(w1, (1,1,1,…,1))
(w2, (1,1,…))
(w3,(1…))
…
…
…
…
(w1, 25)
(w2, 77)
(w3, 12)
…
…
…
…

MR Phases


DATA ON HDFS
INPUT PARTITION
MAP
MAP
MAP
MAP
SHUFFLING
SORT IN PARALLEL
REDUCE
REDUCE
OUTPUT PARTITION
DATA ON HDFS

Taxonomy of Parallel Architectures

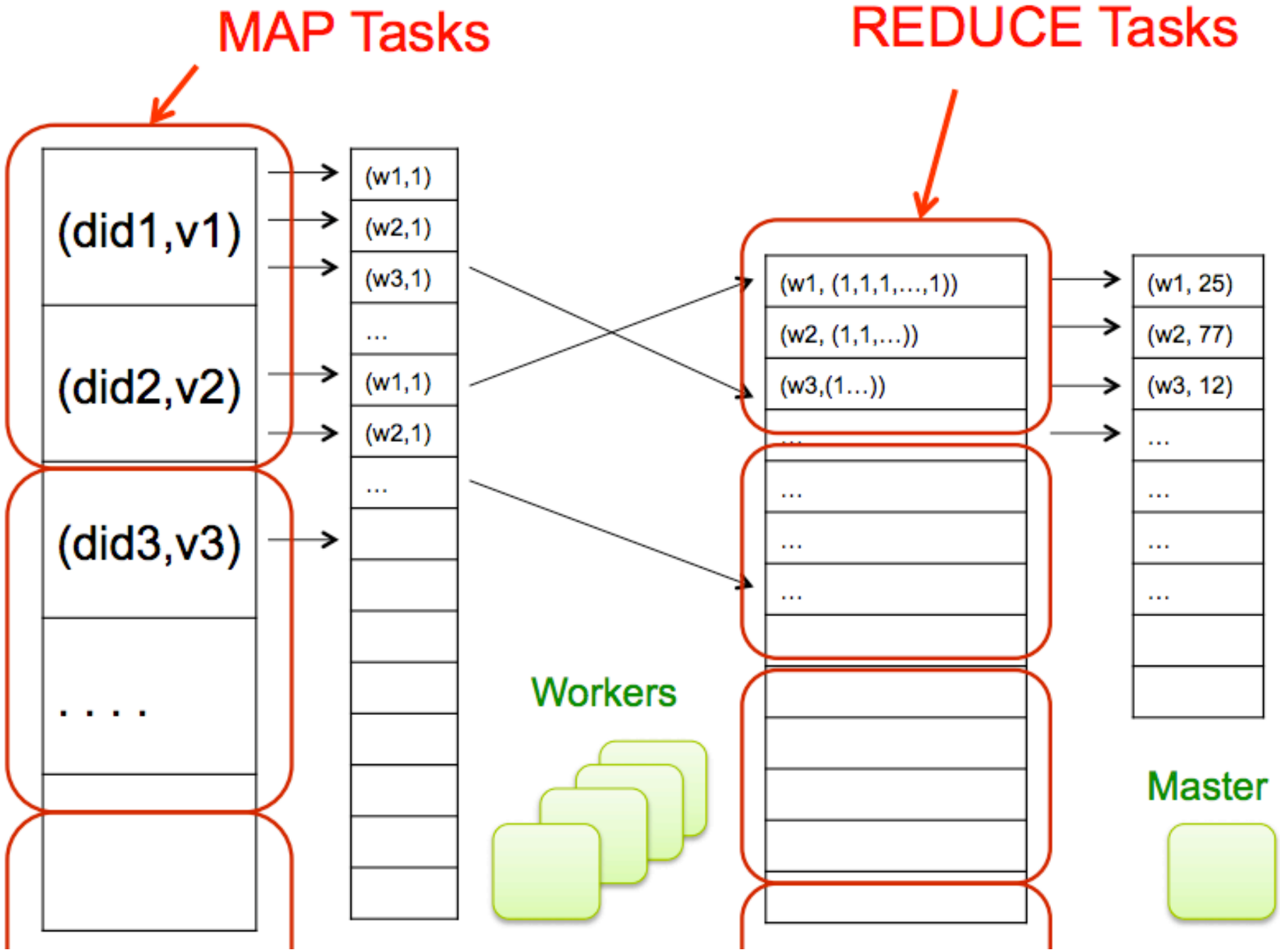
Scales to 1000s of computers



Easiest to program,

but $$$$

Fig. 3.1 Logical multi-processor database designs (diagram after [DEWI92])


MAP Tasks
REDUCE Tasks
(did1,v1)
(did2,v2)
(did3,v3)
. . . .
(w1,1)
(w2,1)
(w3,1)
…
(w1,1)
(w2,1)
…
(w1, (1,1,1,…,1))
(w2, (1,1,…))
(w3,(1…))
…
…
…
…
…
(w1, 25)
(w2, 77)
(w3, 12)
…
…
…
…
Workers
Master

Implementation

- There is one master node
- Master partitions input file into *M splits*, by key
- Master assigns *workers* (=servers) to the *M map tasks*, keeps track of their progress
- Workers write their output to local disk, partition into *R regions*
- Master assigns workers to the *R reduce tasks*
- Reduce workers read regions from the map workers' local disks

Large-Scale Data Processing

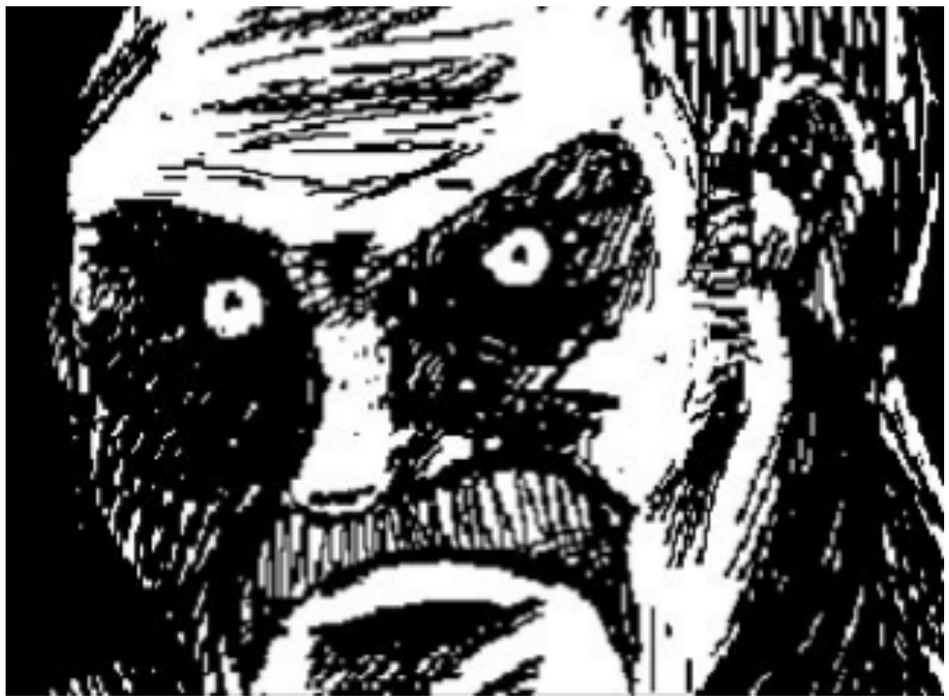
- Many tasks process big data, produce big data
- Want to use hundreds or thousands of CPUs
  - ... but this needs to be easy
  - Parallel databases exist, but they are expensive, difficult to set up, and do not necessarily scale to hundreds of nodes.

- MapReduce is a *lightweight* framework, providing:
  - **Automatic parallelization and distribution**
  - **Fault-tolerance**
  - **I/O scheduling**
  - **Status and monitoring**

MapReduce Contemporaries

- Dryad (Microsoft)
  - Relational Algebra
- Pig (Yahoo)
  - Near Relational Algebra over MapReduce
- HIVE (Facebook)
  - SQL over MapReduce
- Cascading
  - Relational Algebra
- Clustera
  - U of Wisconsin
- Hbase
  - Indexing on HDFS

Talk is cheap,
Show me the code



Examples

Input:

tweet1, (“I love pancakes for breakfast”)

tweet2, (“I dislike pancakes”)

tweet3, (“What should I eat for breakfast?”)

tweet4, (“I love to eat”)

Desired output:

“pancakes”, (tweet1, tweet2)

“breakfast”, (tweet1, tweet3)

“eat”, (tweet3, tweet4)

“love”, (tweet1, tweet4)

…

Examples

Employee

Name	SSN
Sue	999999999
Tony	777777777

Assigned Departments

EmpSSN	DepName
999999999	Accounts
777777777	Sales
777777777	Marketing

Emplyee ⋈ Assigned Departments

Name	SSN	EmpSSN	DepName
Sue	999999999	999999999	Accounts
Tony	777777777	777777777	Sales
Tony	777777777	777777777	Marketing

Examples

Key idea: Lump all the tuples together into one dataset

Employee

Name	SSN
Sue	999999999
Tony	777777777

Assigned Departments

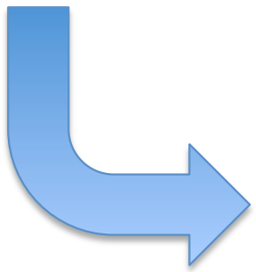
EmpSSN	DepName
999999999	Accounts
777777777	Sales
777777777	Marketing

Employee, Sue, 999999999
Employee, Tony, 777777777
Department, 999999999, Accounts
Department, 777777777, Sales
Department, 777777777, Marketing

What is this for?

Examples

Employee, Sue, 999999999
Employee, Tony, 777777777
Department, 999999999, Accounts
Department, 777777777, Sales
Department, 777777777, Marketing



key=999999999, value=(Employee, Sue, 999999999)
key=777777777, value=(Employee, Tony, 777777777)
key=999999999, value=(Department, 999999999, Accounts)
key=777777777, value=(Department, 777777777, Sales)
key=777777777, value=(Department, 777777777, Marketing)

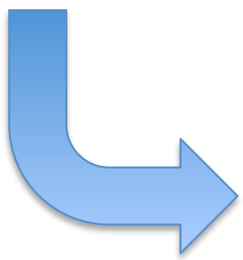
why do we use this as the key?

Examples

key=999999999, values=[(Employee, Sue, 999999999),
(Department, 999999999, Accounts)]

Sue, 999999999, 999999999, Accounts

key=777777777, values=[(Employee, Tony, 777777777),
(Department, 777777777, Sales),
(Department, 777777777, Marketing)]



Tony, 777777777, 777777777, Sales
Tony, 777777777, 777777777, Marketing

Examples

Order(orderid, account, date)

1, aaa, d1
2, aaa, d2
3, bbb, d3

LineItem(orderid, itemid, qty)

1, 10, 1
1, 20, 3
2, 10, 5
2, 50, 100
3, 20, 1

Map

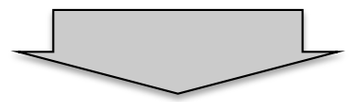
tagged with relation name

Order
- 1, aaa, d1 → 1 : “Order”, (1,aaa,d1)
- 2, aaa, d2 → 2 : “Order”, (2,aaa,d2)
- 3, bbb, d3 → 3 : “Order”, (3,bbb,d3)

Line
- 1, 10, 1 → 1 : “Line”, (1, 10, 1)
- 1, 20, 3 → 1 : “Line”, (1, 20, 3)
- 2, 10, 5 → 2 : “Line”, (2, 10, 5)
- 2, 50, 100 → 2 : “Line”, (2, 50, 100)
- 3, 20, 1 → 3 : “Line”, (3, 20, 1)

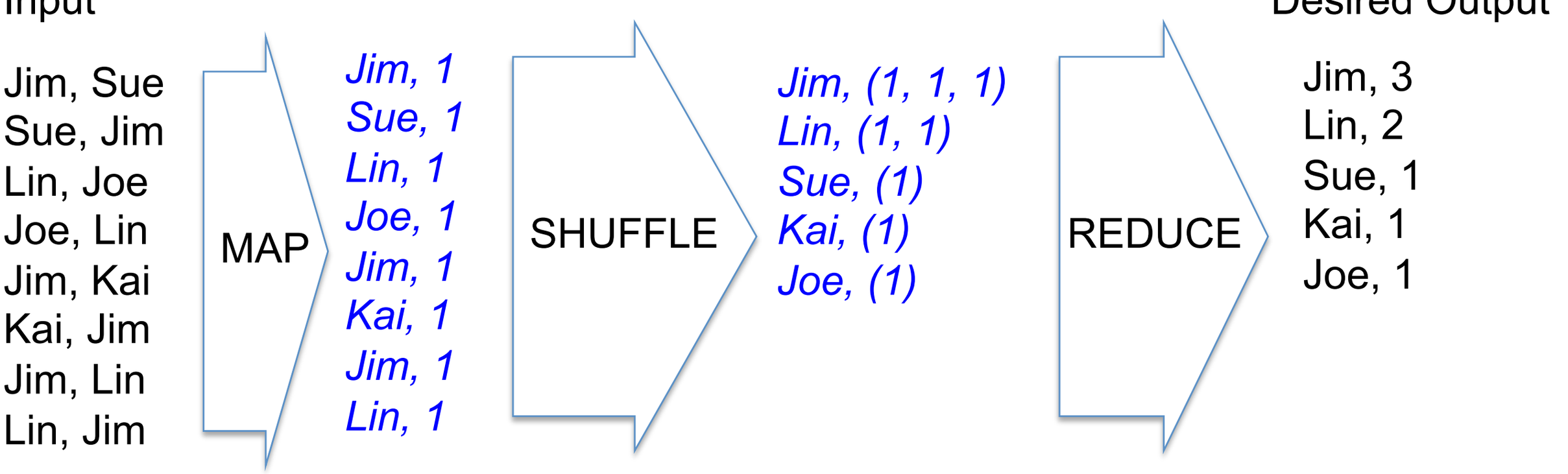
Reducer for key 1

“Order”, (1,aaa,d1)
“Line”, (1, 10, 1)
“Line”, (1, 20, 3)



(1, aaa, d1, 1, 10, 1)
(1, aaa, d1, 1, 20, 3)

Examples


Input
Jim, Sue
Sue, Jim
Lin, Joe
Joe, Lin
Jim, Kai
Kai, Jim
Jim, Lin
Lin, Jim
MAP
Jim, 1
Sue, 1
Lin, 1
Joe, 1
Jim, 1
Kai, 1
Jim, 1
Lin, 1
SHUFFLE
Jim, (1, 1, 1)
Lin, (1, 1)
Sue, (1)
Kai, (1)
Joe, (1)
REDUCE
Desired Output
Jim, 3
Lin, 2
Sue, 1
Kai, 1
Joe, 1

Meeting Hadoop

Cluster Computing

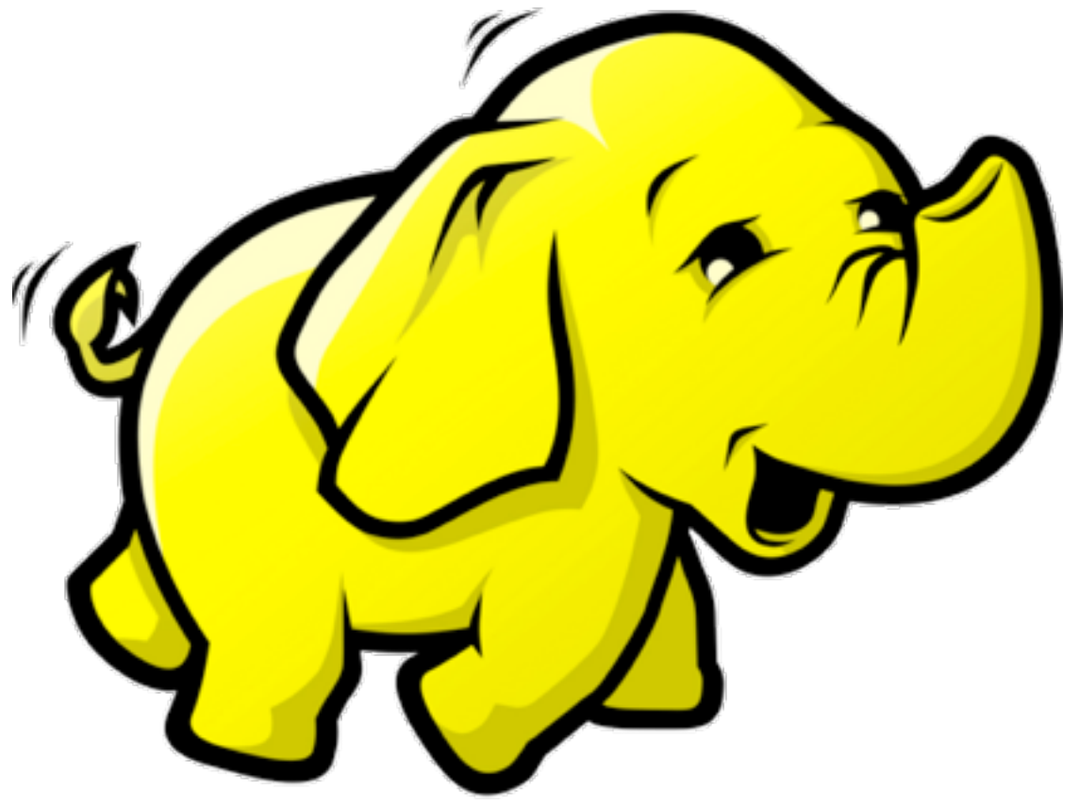
- Large number of commodity servers, connected by high speed, commodity network
- Rack: holds a small number of servers
- Data center: holds many racks

Cluster Computing

- Massive parallelism:
  - 100s, or 1000s, or 10000s servers
  - Many hours
- Failure:
  - If medium-time-between-failure is 1 year
  - Then 10000 servers have one failure / hour

Distributed File System (DFS)

- For very large files: TBs, PBs
- Each file is partitioned into *chunks*, typically 64MB
- Each chunk is replicated several times (≥3), on different racks, for fault tolerance
- Implementations:
  - Google's DFS: GFS, proprietary
  - Hadoop's DFS: HDFS, open source



Hadoop

What is Hadoop …

*Flexible and available architecture for large scale distributed **batch** processing on a network of commodity hardware.*

Apache top level project

http://hadoop.apache.org/

500 contributors

It has one of the strongest eco systems with large no of sub projects

Yahoo has one of the biggest installation Hadoop
Running 1000s of servers on Hadoop

Inspired by …

{Google GFS + Map Reduce + Big Table}

Architecture behind Google's

Search Engine

Creator of Hadoop project

Doug Cutting
Co-founder of
Apache Hadoop

Use cases … What is Hadoop used for

Big/Social data analysis

Text mining, patterns search

Machine log analysis

Geo-spacitial analysis

Trend Analysis

Genome Analysis

Drug Discovery

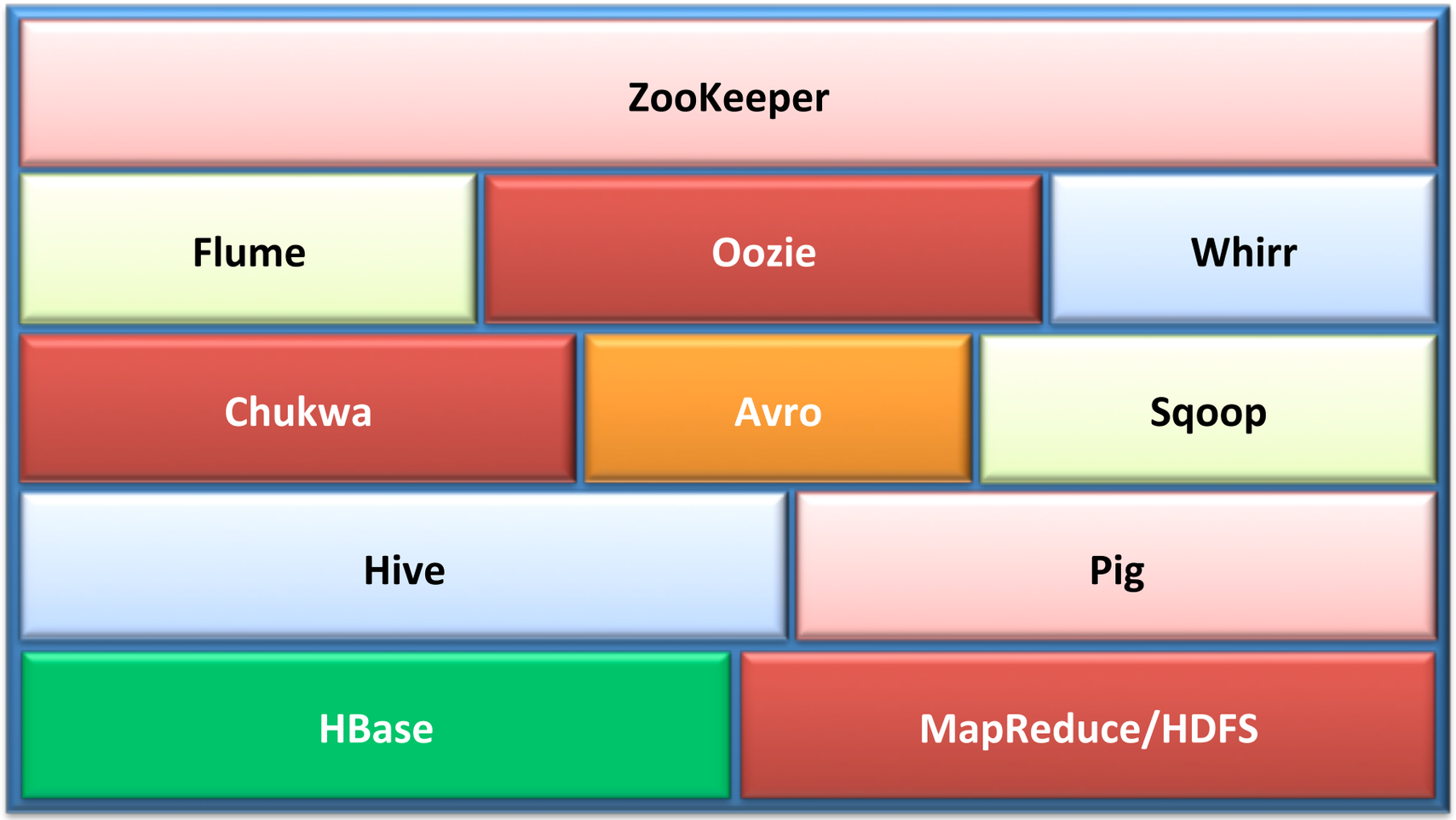
Fraud and compliance management

Video and image analysis

Who uses Hadoop … long list

- *Amazon/A9*
- *Facebook*
- *Google*
- *IBM*
- *Disney*
- *Last.fm*
- *New York Times*
- *Yahoo!*
- *Twitter*
- *Linked in*

Hadoop ecosystem …



Hadoop distribution …

cloudera

Hortonworks

InfoSphere BigInsights

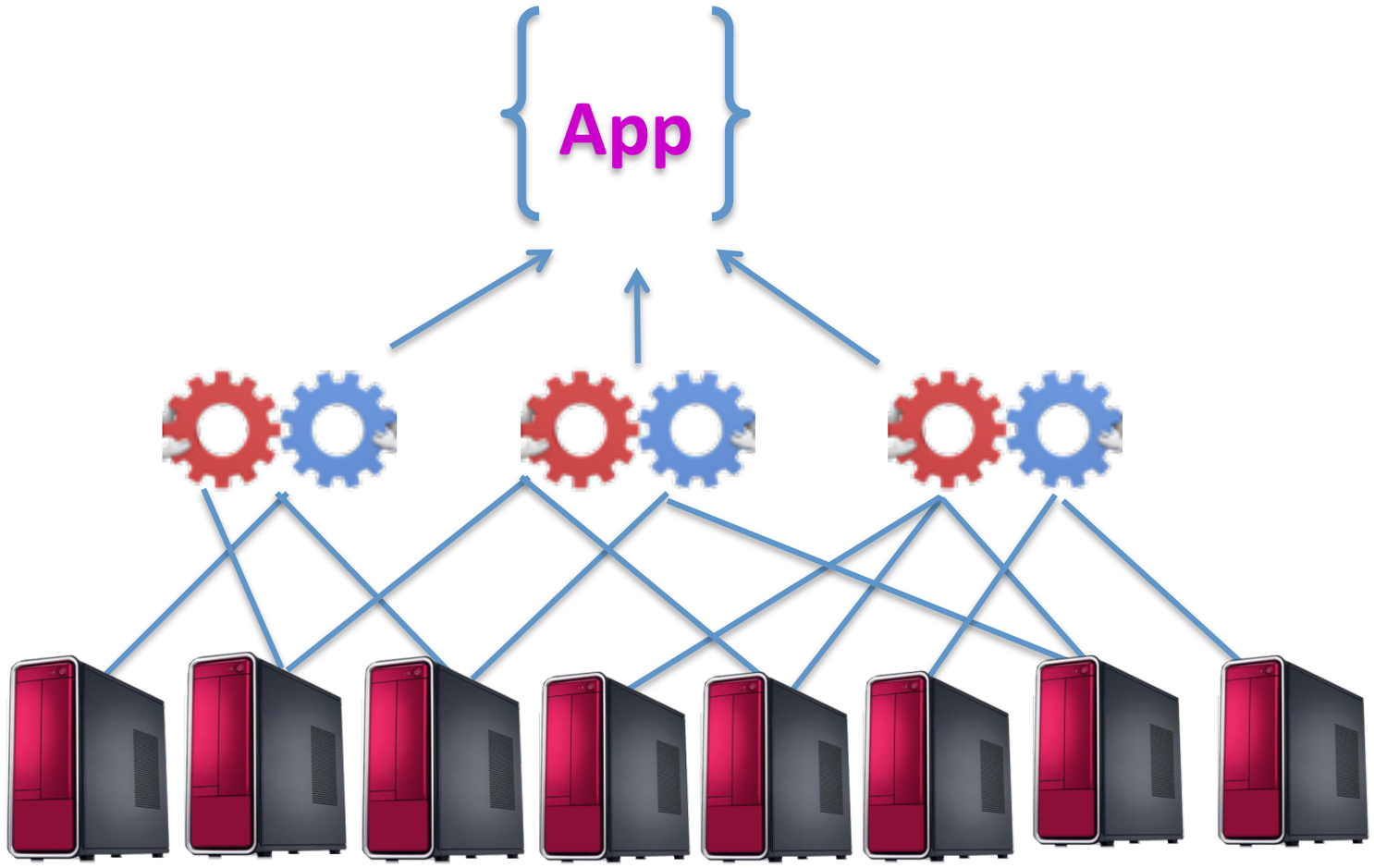
MAPR TECHNOLOGIES

EASY. DEPENDABLE. FAST.

Scale up

Traditional Databases

Scale out



Hadoop distributed file system

Hadoop Components …

HDFS

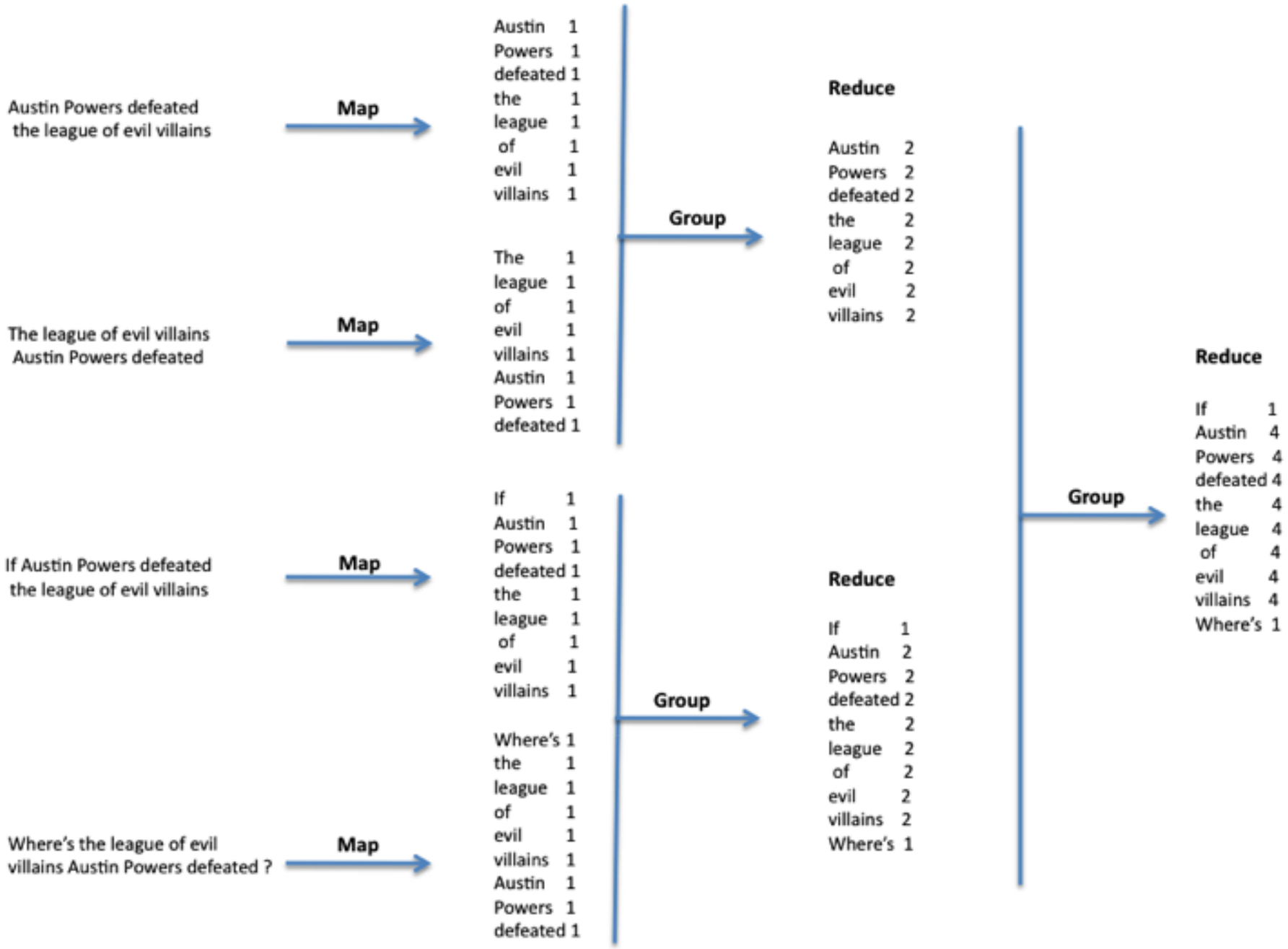
Map Reduce

Job tracker

Task Tracker

Name Node

MapReduce



http://labs.google.com/papers/mapreduce.html

Exemplo de MapReduce

```
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")

class MRWordFreqCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            yield (word.lower(), 1)

    def reducer(self, word, counts):
        yield (word, sum(counts))

if __name__ == '__main__':
    MRWordFreqCount().run()
```

Projeto MrJob

Criado pela Equipe de Engenharia do Yelp

Totalmente Open-Source

Todo em Python

Utiliza Map-Reduce para Processamento

Permite rodar tanto no Amazon EMR como no Hadoop

Objetivos do MrJobs

- ✔ Se você quer aprender MapReduce, ele é para você
- ✔ Se você tem um problema cavalar e precisa de muito processamento e não está afim de mexer em Hadoop
- ✔ Se você já tem um cluster Hadoop e quer rodar scripts Python
- ✔ Se você quer migrar seu código Python do Hadoop para o EMR
- ✖ Se você não quer escrever Python (Impossível!), não é para você!

Passos importantes

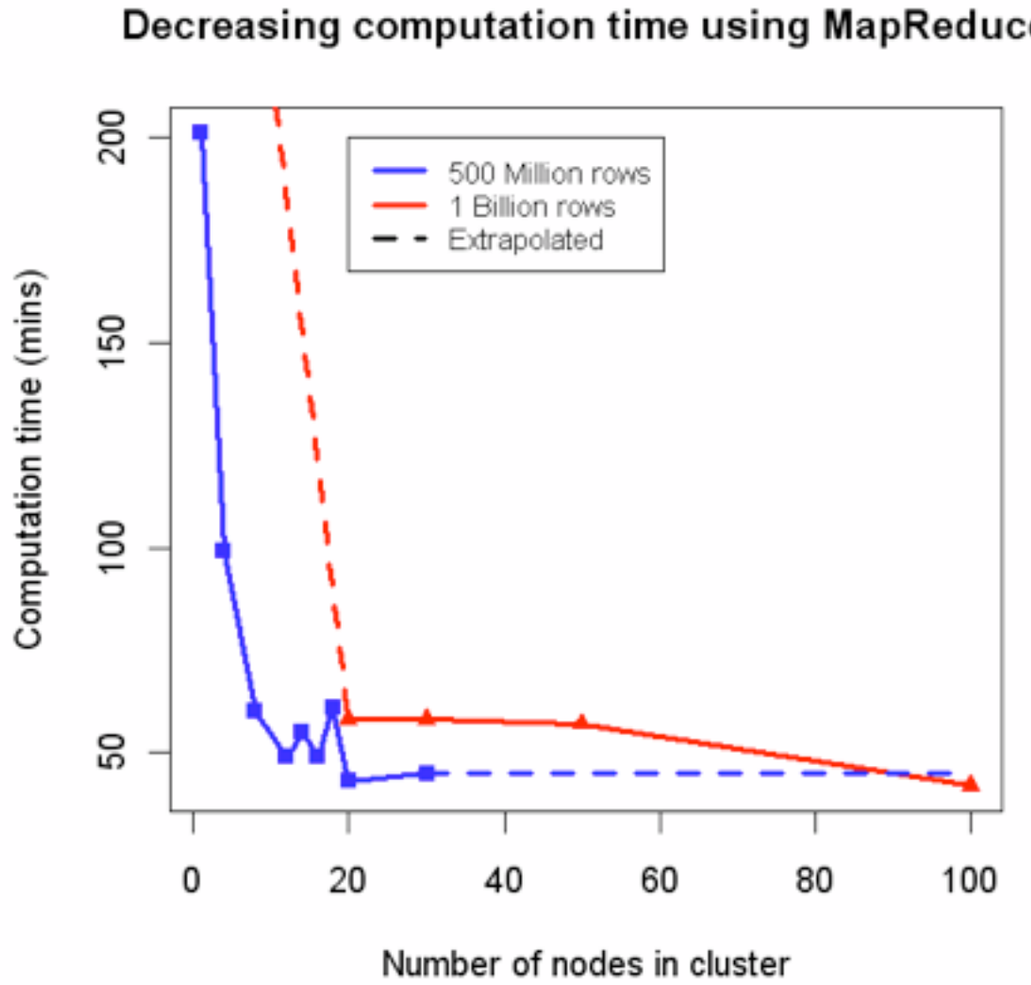
sudo easy_install mrjob

Vamos a uma demo...

Texto da posse de Obama em 2009.

PyCon-PE

Desempenho MapReduce



Mais informações

http://packages.python.org/mrjob/

https://github.com/Yelp/mrjob

Distributed Computing with **mrJob**

https://github.com/Yelp/mrjob

Elsayed et al: Pairwise Document Similarity in Large Collections with MapReduce

Map - make user the key

(Alice,Matrix,5)	→	Alice (Matrix,5)
(Alice,Alien,1)	→	Alice (Alien,1)
(Alice,Inception,4)	→	Alice (Inception,4)
(Bob,Alien,2)	→	Bob (Alien,2)
(Bob,Inception,5)	→	Bob (Inception,2)
(Peter,Matrix,4)	→	Peter (Matrix,4)
(Peter,Alien,3)	→	Peter (Alien,3)
(Peter,Inception,2)	→	Peter (Inception,2)

Reduce - create inverted index

Alice (Matrix,5) Alice (Alien,1) Alice (Inception,4)	→	Alice (Matrix,5)(Alien,1)(Inception,4)
Bob (Alien,2) Bob (Inception,5)	→	Bob (Alien,2)(Inception,5)
Peter (Matrix,4) Peter (Alien,3) Peter (Inception,2)	→	Peter (Matrix,4)(Alien,3)(Inception,2)

Distributed Computing with **mrJob**

https://github.com/Yelp/mrjob

Elsayed et al: Pairwise Document Similarity in Large Collections with MapReduce

Map - emit all cooccurred ratings

Alice (Matrix,5)(Alien,1) (Inception,4)	→	Matrix,Alien (5,1) Matrix,Inception (5,4) Alien,Inception (1,4)
Bob (Alien,2)(Inception,5)	→	Alien,Inception (2,5)
Peter (Matrix,4)(Alien,3) (Inception,2)	→	Matrix,Alien (4,3) Matrix,Inception (4,2) Alien,Inception(3,2)

Reduce - compute similarities

Matrix,Alien (5,1) Matrix,Alien (4,3)	→	Matrix,Alien (-0.47)
Matrix,Inception (5,4) Matrix,Inception (4,2)	→	Matrix,Inception (0.47)
Alien,Inception (1,4) Alien,Inception (2,5) Alien,Inception (3,2)	→	Alien,Inception (-0.63)

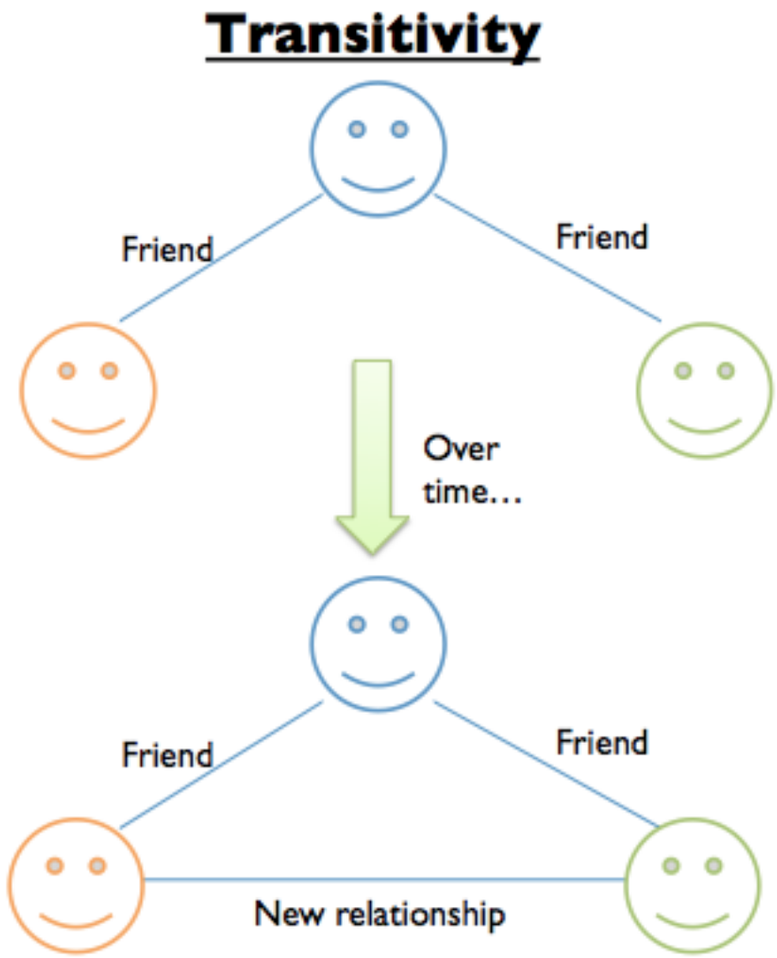
Atepassar Recommendations

http://atepassar.com

Problema: Como recomendar novos amigos ?

Atepassar Recommendations

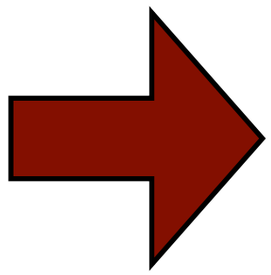
http://atepassar.com



Atepassar Recommendations

http://atepassar.com

marcel;jonas,maria,jose,amanda
maria;carol,fabiola,amanda,marcel
amanda;paula,patricia,maria,marcel
carol;maria,jose,patricia
fabiola;maria
paula;fabio,amanda
patricia;amanda,carol
jose;marcel,carol
jonas;marcel,fabio
fabio;jonas,paula
carla



"marcel" [["carol", 2], ["fabio", 1], ["fabiola", 1], ["patricia", 1], ["paula", 1]]
"maria" [["jose", 2], ["patricia", 2], ["jonas", 1], ["paula", 1]]
"patricia" [["maria", 2], ["jose", 1], ["marcel", 1], ["paula", 1]]
"paula" [["jonas", 1], ["marcel", 1], ["maria", 1], ["patricia", 1]]
"amanda" [["carol", 2], ["fabio", 1], ["fabiola", 1], ["jonas", 1], ["jose", 1]]
"carol" [["amanda", 2], ["marcel", 2], ["fabiola", 1]]
"fabio" [["amanda", 1], ["marcel", 1]]
"fabiola" [["amanda", 1], ["carol", 1], ["marcel", 1]]
"jonas" [["amanda", 1], ["jose", 1], ["maria", 1], ["paula", 1]]
"jose" [["maria", 2], ["amanda", 1], ["jonas", 1], ["patricia", 1]]

$python friends_recommender.py - r emr --num-ec2-instances 5 facebook_data.csv > output.dat

Atepassar Recommendations

http://atepassar.com

marcel;jonas,maria,jose,amanda
maria;carol,fabiola,amanda,marcel
amanda;paula,patricia,maria,marcel
carol;maria,jose,patricia
fabiola;maria
paula;fabio,amanda
patricia;amanda,carol
jose;marcel,carol
jonas;marcel,fabio
fabio;jonas,paula
carla

marcel;jonas,maria,jose,amanda
carol;maria,jose,patricia
fabio;jonas,paula
…

iola", 1], ["patricia", 1], ["paula", 1]]
1], ["paula", 1]]
, 1], ["paula", 1]]
1], ["patricia", 1]]
", 1], ["jonas", 1], ["jose", 1]]
la", 1]]
el", 1]]
1], ["paula", 1]]
1], ["patricia", 1]]

$python friends_recommender.py - r emr --num-ec2-instances 5 facebook_data.csv > output.dat

Atepassar Recommendations

http://atepassar.com

Celery - para agendamento dos jobs coletores e executores.

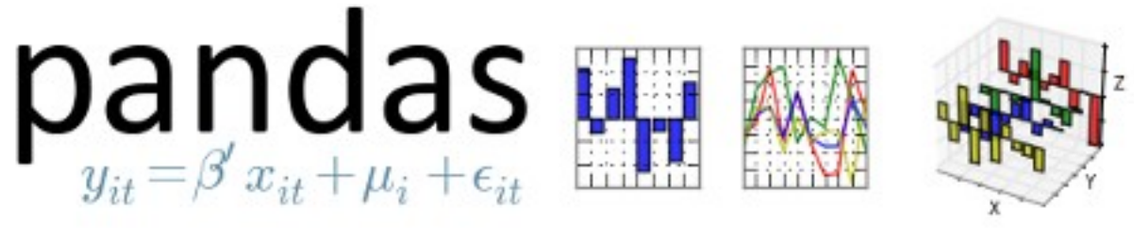
mrJob - para mapreduce e acesso ao Hadoop

MongoDb - para armazenamento das recomendações

Boto - acesso aos files do S3.

A melhor parte!

Projetos interessantes

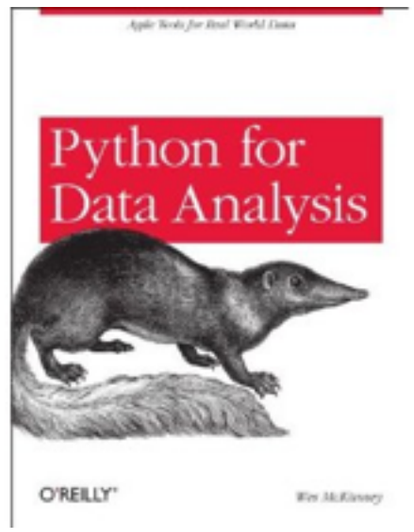


Pandas: a data analysis library for Python,
poised to give R a run for its money... http://pandas.pydata.org/

Estrutura de dados para manipulação rápida

- slicing
- indexing
- subseting

Handling missing data
Agregações, Séries Temporais



Projetos interessantes

disco
massive data - minimal code

http://discoproject.org/

Outro framework para computação distribuída com Python com MapReduce.

Criado pelo Instituto Nokia.

Backend dele é escrito em Erlang (funcional, concorrente e bem escalável!)

Não utiliza o FileSystem mais usado HDFS e sim um novo padrão por eles (DDFS).

Projetos interessantes

Python & MongoDb

http://api.mongodb.org/python/2.0/examples/map_reduce.html

MongoDb - Banco de Dados Não relacional (NoSQL)

Possui suporte nativo built-in para fazer MapReduce.

Escrever o código em JS e não é muito legível e fica preso ao Mongo ...

```
>>> reduce = Code("function (key, values) {"
...               "  var total = 0;"
...               "  for (var i = 0; i < values.length; i++) {"
...               "    total += values[i];"
...               "  }"
...               "  return total;"
...               "}")
```

Projetos interessantes

Dumbo

https://github.com/klbostee/dumbo/wiki/Short-tutorial

Uma das primeiras bibliotecas em cima do MapReduce e Python.

Complicado para começar e está desatualizada :(

Projetos interessantes

http://pydoop.sourceforge.net/docs/index.html

Um wrapper em Python em cima do Hadoop para computação distribuída.

Legal, mas dá um trabalho para configurar.

Projetos interessantes

https://developers.google.com/appengine/docs/python/dataprocessing/

Mapreduce com Python na Google AppEngine

Ainda experimental e fica "preso" à plataforma AppEngine.

Projetos interessantes

http://scikit-learn.org/stable/

Algoritmos de aprendizagem de máquina

Supervisionados & Não supervisionados

Pré-processamento, extração de dados

Avaliação de classificadores, Pipeline,
seleção de atributos.

Projetos interessantes

NLTK

Natural Language Toolkit

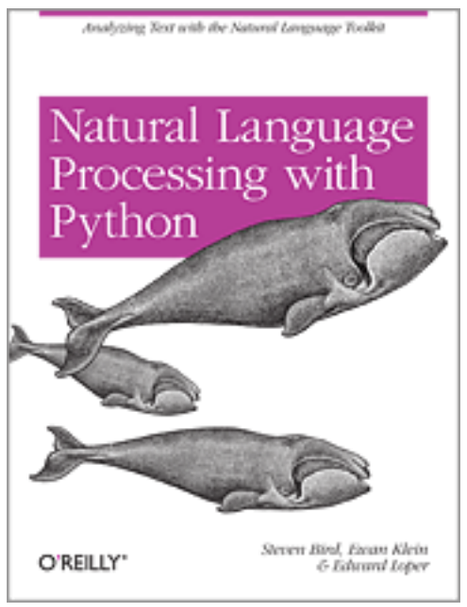
Reference Guide

http://nltk.org/

Processamento de linguagem natural

Várias ferramentas para tokenização, pos tagging, named entity recognition, classificadores, etc.

Vários corpus disponíveis!



Projetos interessantes

NLTK

Natural Language Toolkit

Reference Guide

http://nltk.org/

Proc

Vári
pos
classificadores, etc.

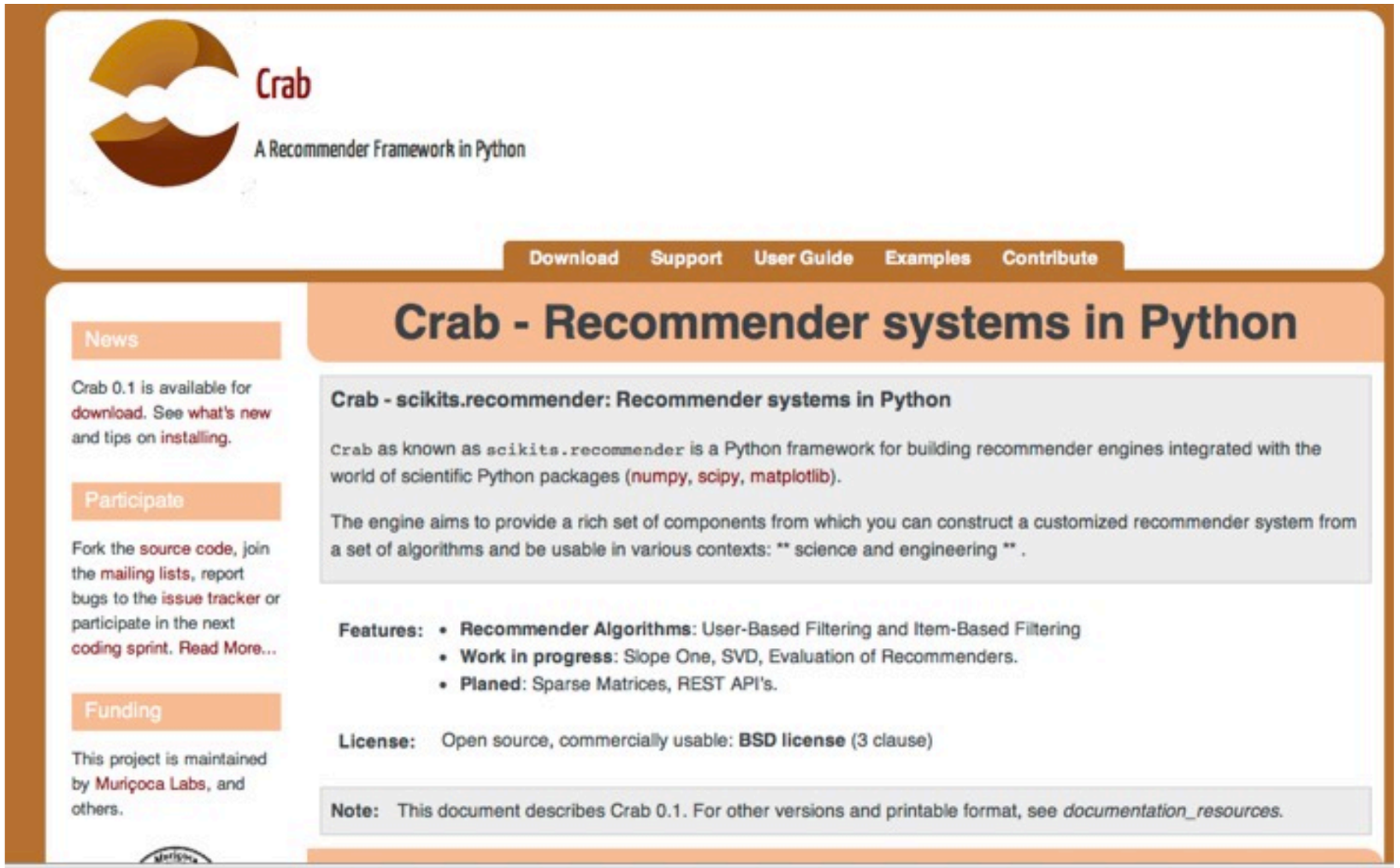
Vários corpus disponíveis!

Pipeline for distributed Natural Language Processing, made in Python

https://github.com/NAMD/pypln

Projetos interessantes

Our Project's Home Page



http://muricoca.github.com/crab

Future **Releases**

✔ Planned Release 0.13

New home for python-recsys:
https://github.com/python-recsys/crab

New commiters: igormedeiros

✔ Planned Release 0.14

Support to Item-Based Recommenders using MapReduce with MrJob

Join **us**!

1. Read our Wiki Page

   https://github.com/muricoca/crab/wiki/Developer-Resources

2. Check out our current sprints and open issues

   https://github.com/muricoca/crab/issues

3. Forks, Pull Requests mandatory

4. Join us at irc.freenode.net #muricoca or at our discussion list

   http://groups.google.com/group/scikit-crab

Vários outros ...

Numpy

milk

NetworkX

Orange

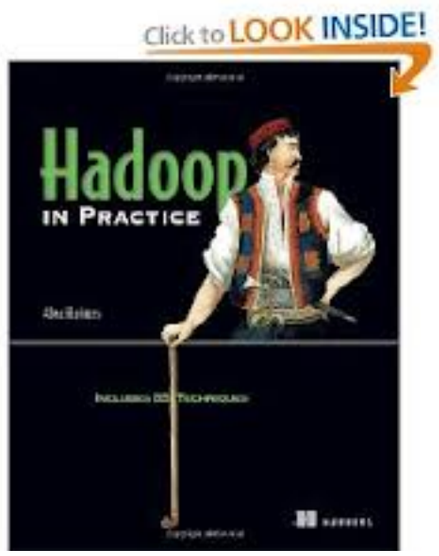
Matplotlib

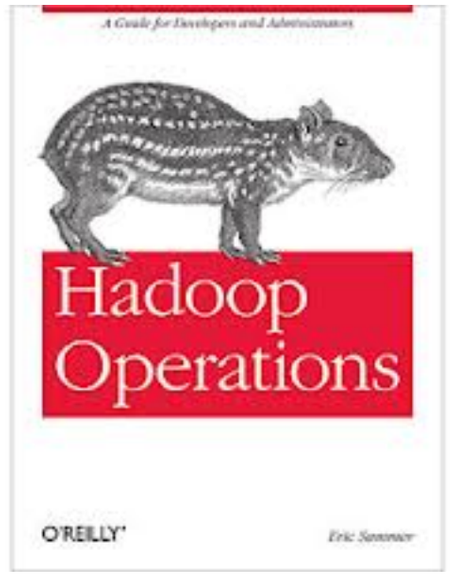
Scipy

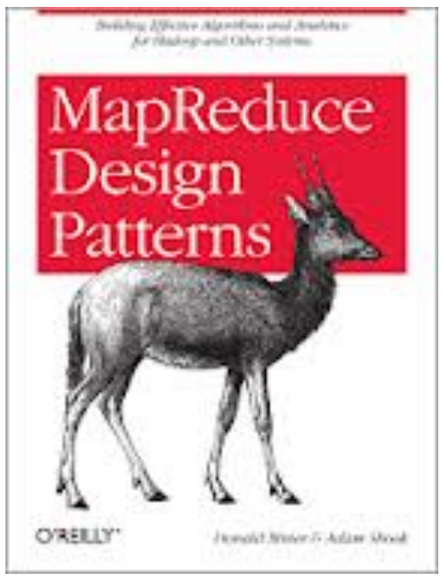
StatsModels

PyBrain

Livros recomendados







Livros recomendados

For free...

http://shop.oreilly.com/product/0636920022640.do?cmp=il-radar-ebooks-big-data-now-radar

Livros recomendados

For free...

http://infolab.stanford.edu/~ullman/mmds/book.pdf

Artigos recomendados

For free...

http://aimotion.blogspot.com.br/2012/10/atepassar-recommendations-recommending.html

http://aimotion.blogspot.com.br/2012/08/introduction-to-recommendations-with.html

Artigos recomendados

For free...

http://aimotion.blogspot.com.br/2012/10/atepassar-recommendations-recommending.html

http://aimotion.blogspot.com.br/2012/08/introduction-to-recommendations-with.html

https://github.com/marcelcaraciolo/recsys-mapreduce-mrjob

THE BIGGEST PLONE EVENT IN THE WORLD AND THE BIGGEST PYTHON EVENT IN BRAZIL

Plone® Conf
Brasília 2013

www.ploneconf.org

python brasil [9]
9ª Conferência Brasileira de Python

www.pythonbrasil.org.br

BigData with Python

A gentle and simple introduction

Marcel Caraciolo

@marcelcaraciolo

Developer, Cientist, contributor to the Crab recsys project, works with Python for 6 years, interested at mobile, education, machine learning and dataaaaa!

Recife, Brazil - http://aimotion.blogspot.com