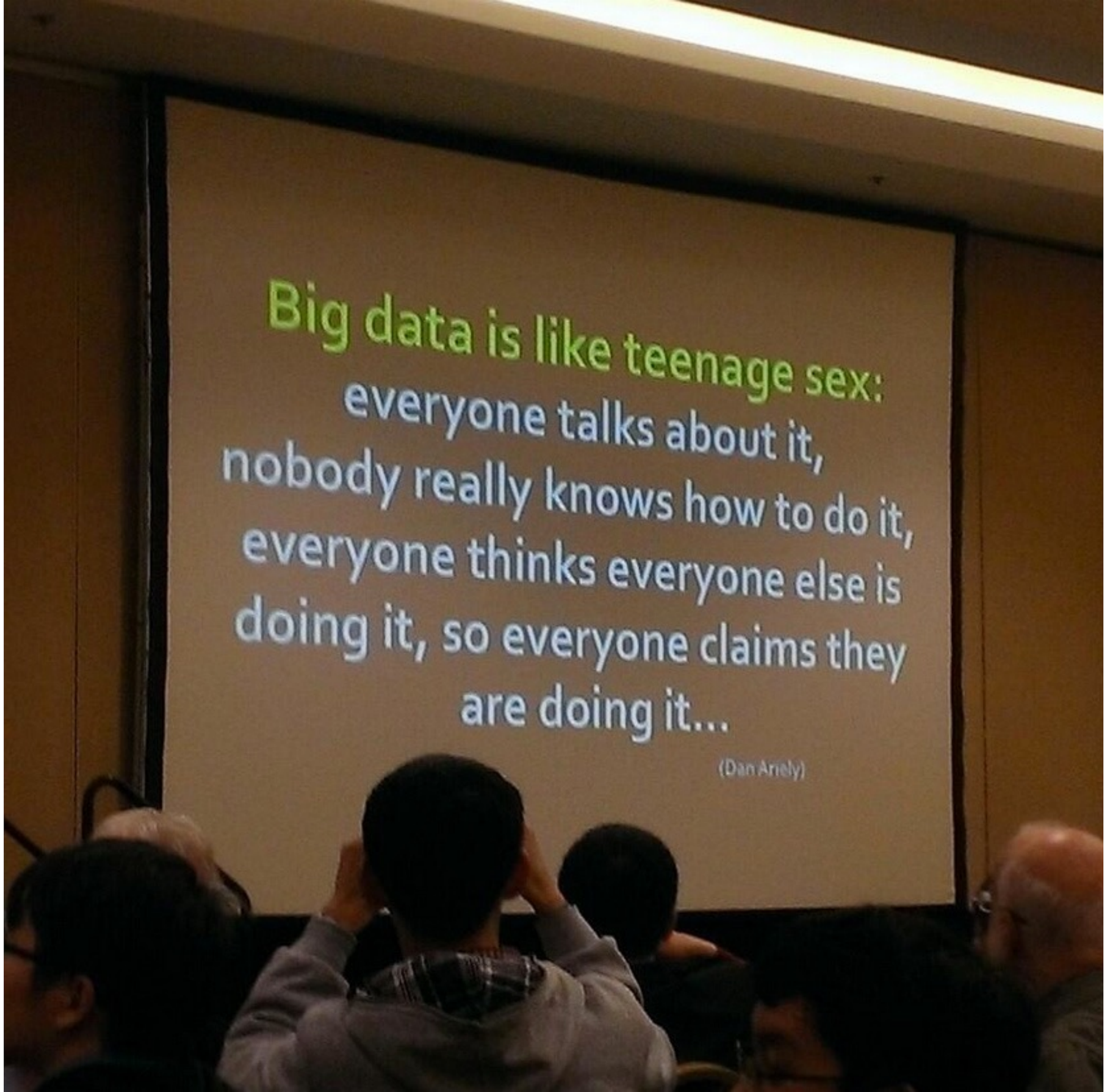
Introduction to Data Analysis

Ilya Ezepov


Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...
(Dan Ariely)

About me

Moscow Exchange, Intern

Yandex

Yandex, Data Analyst

BASED ON YOUR INTERNET HISTORY, YOU MIGHT BE DUMB ENOUGH TO ENJOY EXTREME SPORTS.

Dilbert.com DilbertCartoonist@gmail.com

CLICK HERE TO BUY A TICKET TO BASE JUMP FROM THE INTERNATIONAL SPACE STATION.

2-2-13 ©2013 Scott Adams, Inc./Dist. by Universal Uclick

I THINK THE INTERNET IS TRYING TO KILL ME.

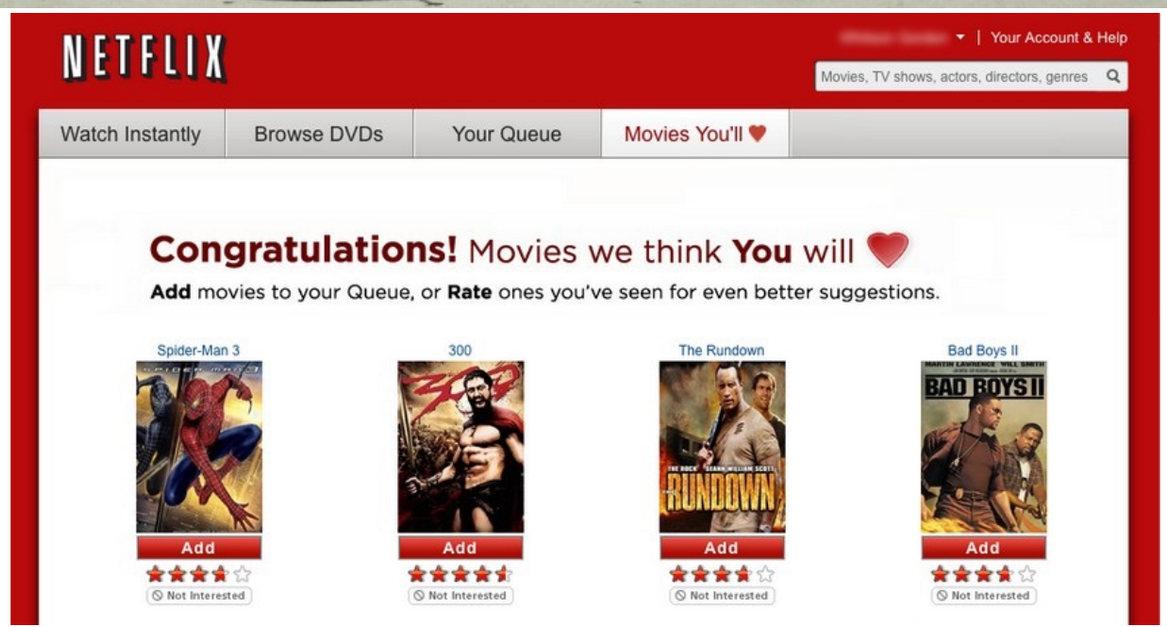
WE CALL IT "MACHINE LEARNING."

About DA

- Exploratory data analysis
- Data visualisation
- Hypothesis testing
- Predictions making
- Big Data analysis
- Map-Reduce
- Programming
- Reinforcement learning
- Deep learning
- Autoencoders
- Scalable data storages
- Data cleaning
- Data clustering
- Text analysis
- DA as a sport (Kaggle)
- Real-time applications

About DA

kaggle



Kaggle

As of July 2015, Kaggle claims approximately 332,000 data scientists on its job boards.

Kaggle

As of July 2015, Kaggle claims approximately 332,000 data scientists on its job boards.

Idea: In 1998 Rob McEwen asked data scientist for $500,000 to find best places to mine gold.

Kaggle

As of July 2015, Kaggle claims approximately 332,000 data scientists on its job boards.

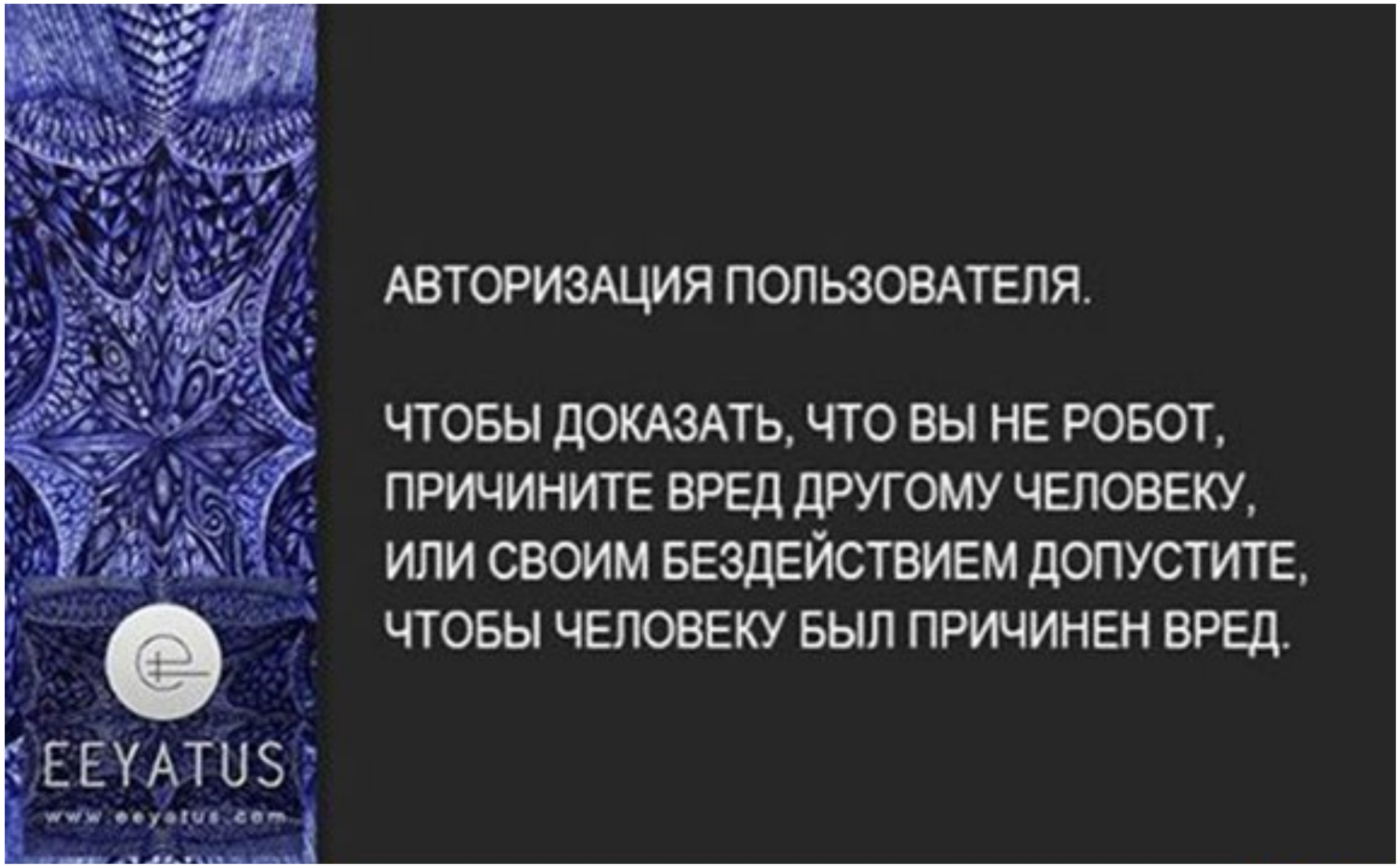
Idea: In 1998 Rob McEwen asked data scientist for $500,000 to find best places to mine gold. In a year he got $3 billion .

How to learn all this???

How to learn all this???

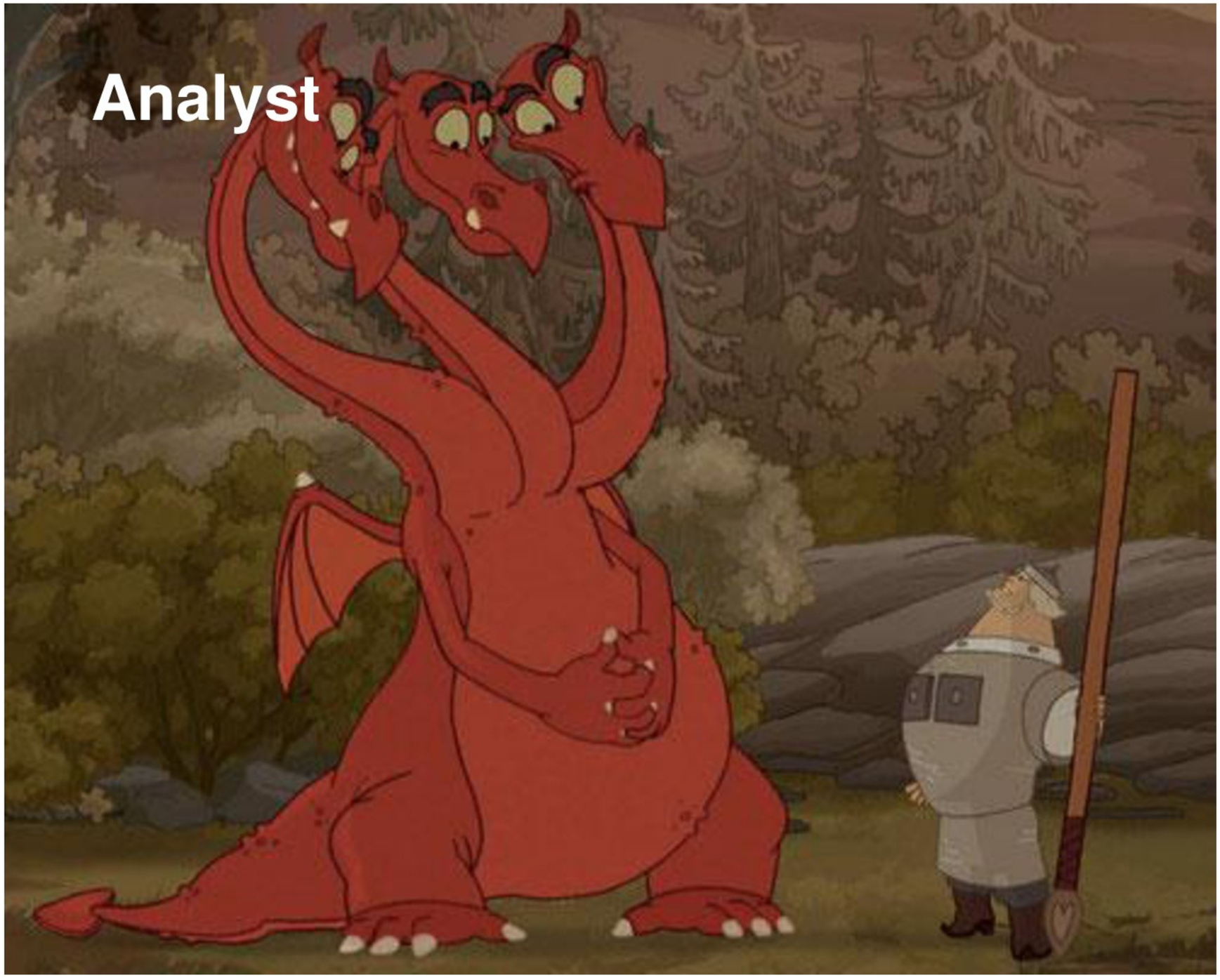
"Self-education is, I firmly believe, the only kind of education there is."

Isaac Asimov

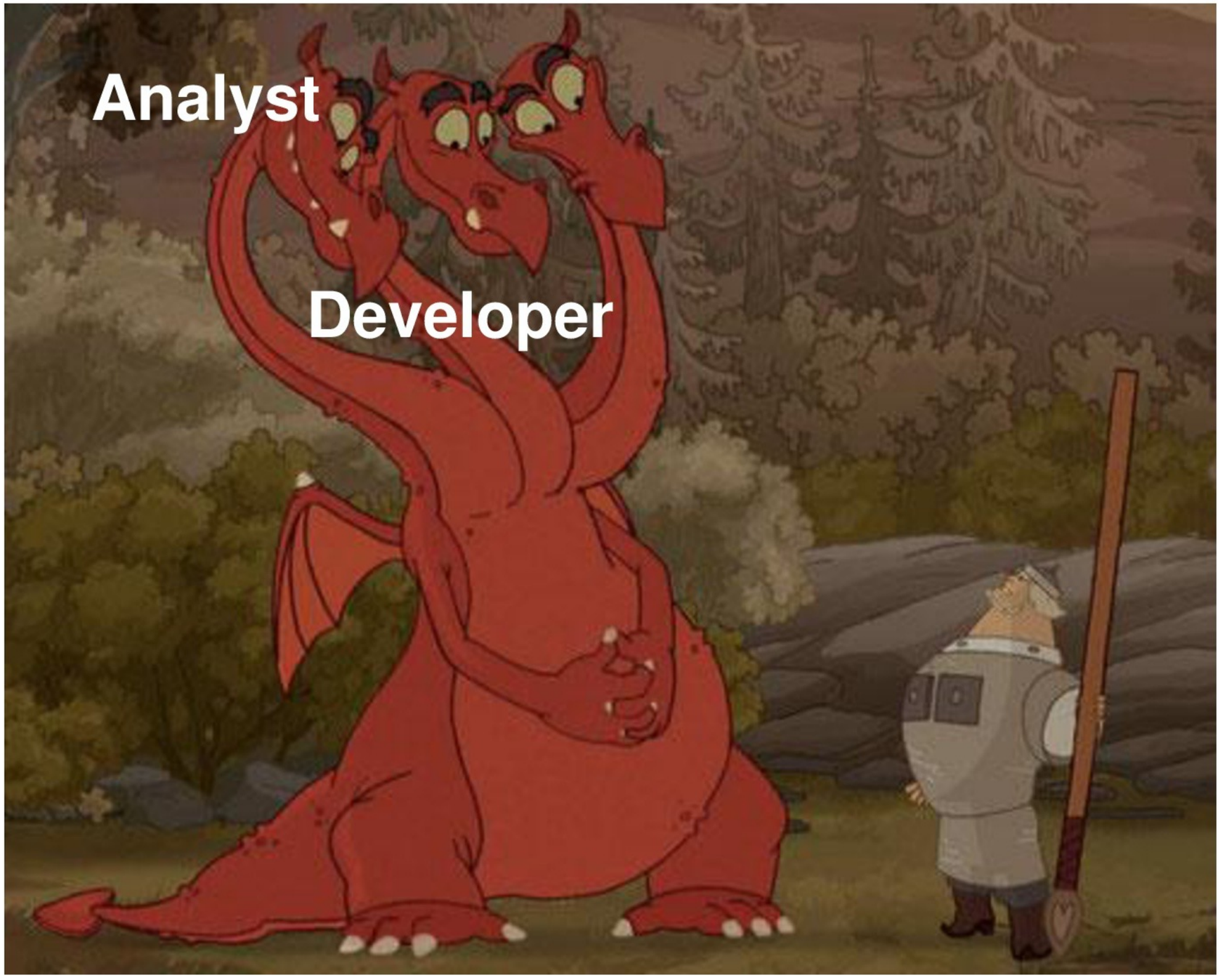

EEYATUS
www.eeyatus.com
АВТОРИЗАЦИЯ ПОЛЬЗОВАТЕЛЯ.
ЧТОБЫ ДОКАЗАТЬ, ЧТО ВЫ НЕ РОБОТ,
ПРИЧИНИТЕ ВРЕД ДРУГОМУ ЧЕЛОВЕКУ,
ИЛИ СВОИМ БЕЗДЕЙСТВИЕМ ДОПУСТИТЕ,
ЧТОБЫ ЧЕЛОВЕКУ БЫЛ ПРИЧИНЕН ВРЕД.

Types of data scientist

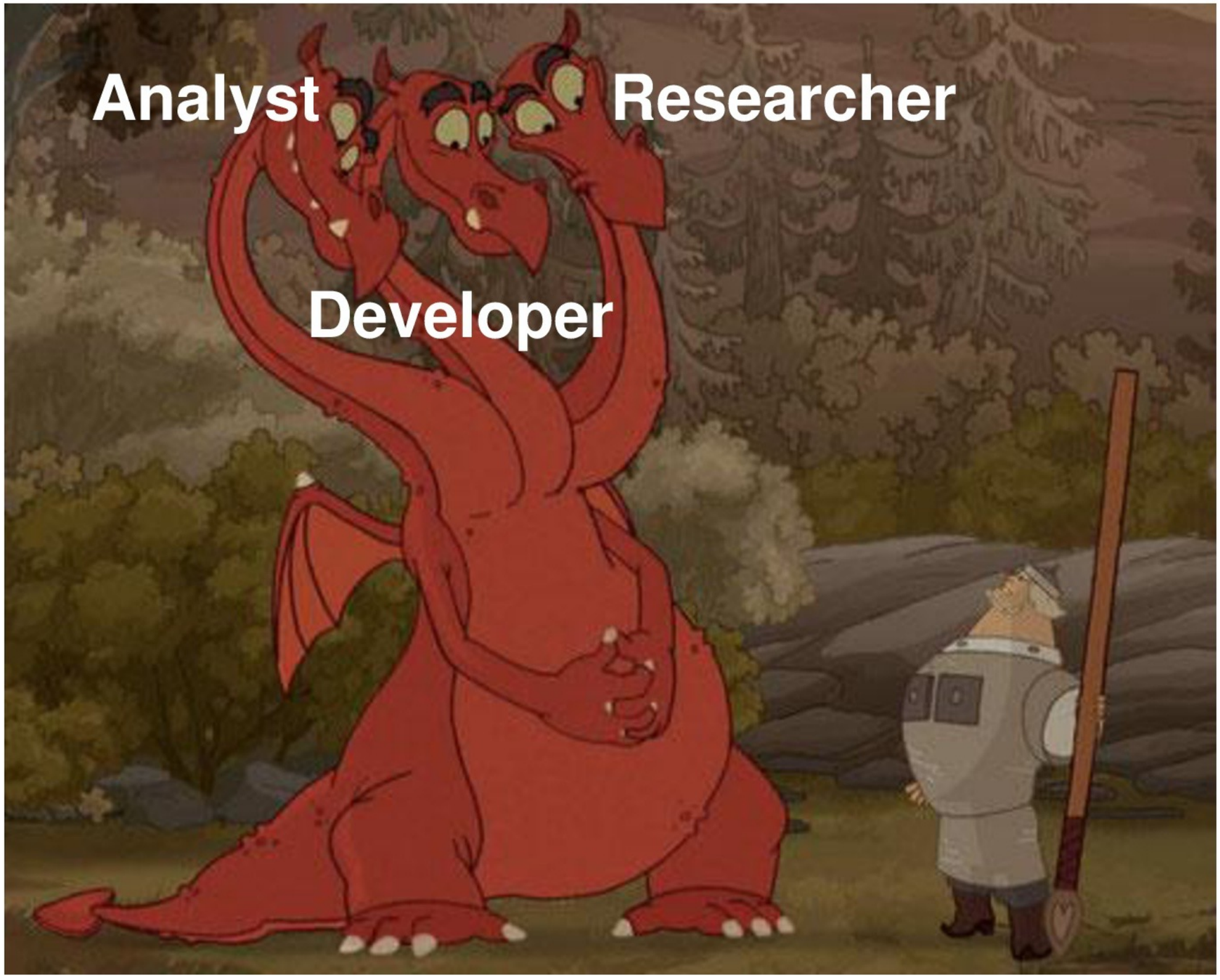
Types of data scientist



Types of data scientist



Types of data scientist



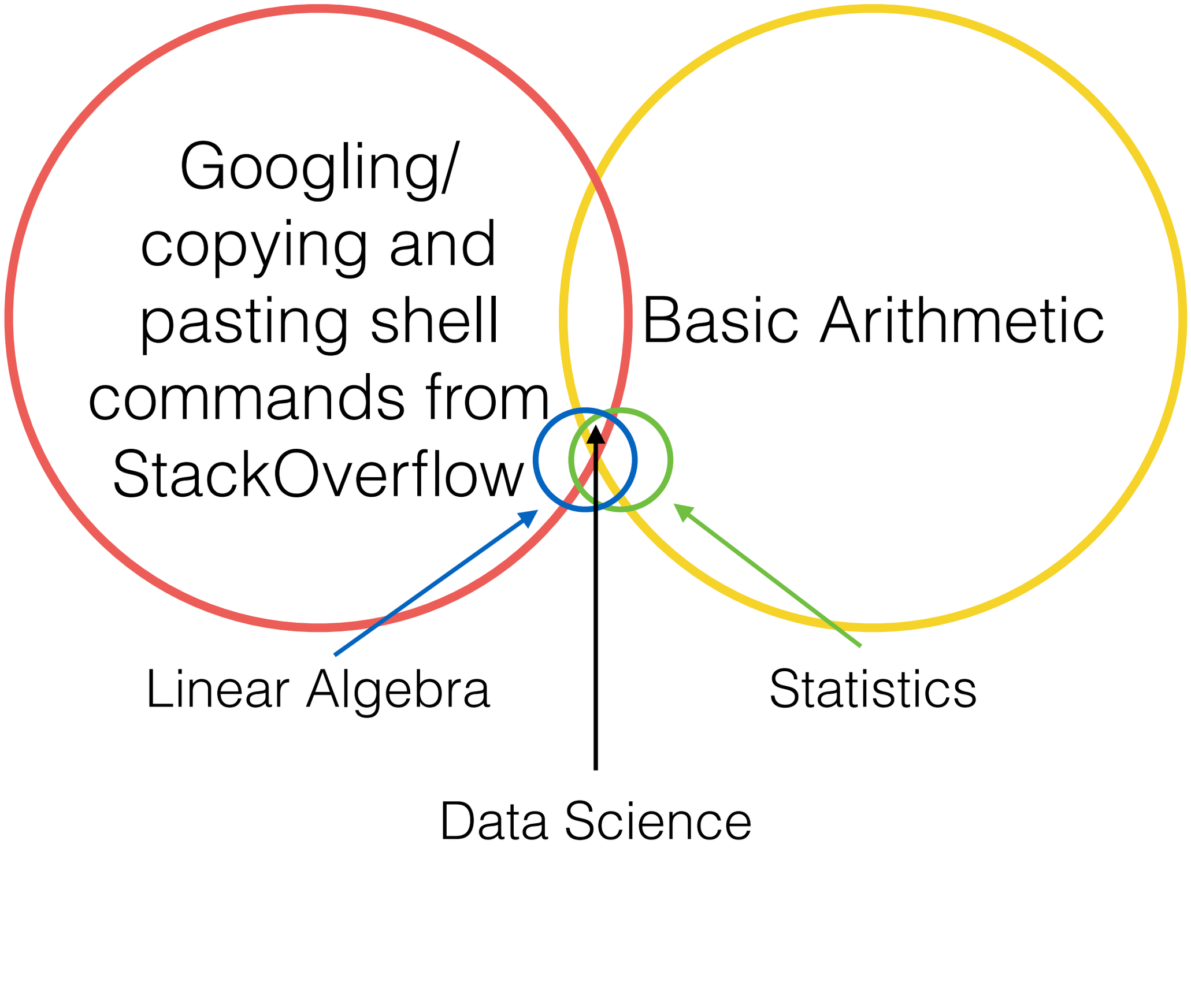
Math

Math skills

- Theory of probability and mathematical statistics
- Calculus and algebra
- Calculus of variations and optimal control

- Not that much for analyst and developer ways

- No upper limit for researcher way


Googling/
copying and
pasting shell
commands from
StackOverflow
Basic Arithmetic
Linear Algebra
Statistics
Data Science

Math skills

- **Moscow is** (one of) **the best place to study!**
- A lot of great universities: MSU, HSE, MIPT, etc.
- Yandex School of Data Analysis
- IUM
- Big and active community: meet-ups, hackathons, lectures, courses, challenges, etc.

Machine learning resources

- **Andrew Ng course @ Coursera**
- https://www.coursera.org/learn/machine-learning
- 11 weeks x (~40 min videos, quiz, assignment)
- Assignments on Octave/MATLAB :(
- Video & quiz + assignment on python/R just for yourself is absolutely ok
- The classic machine learning course: the best place to start (with low math level)

Machine learning resources

- **Yandex video lectures (russian)**
- https://yandexdataschool.ru/edu-process/courses
- 24 x 90 min
- High math level

Machine learning resources

- **Yandex & HSE @ Coursera (russian)**

- https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie

- 7 weeks x 60 min

- Simplified version of Yandex Data School

Machine learning resources

- **Mining Massive Datasets @ Coursera**

- https://www.coursera.org/course/mmds
- http://mmds.org

- 7 weeks x (~200 min videos)

- A lot of material

- Not only about machine learning, but also about general data analysis

Machine learning resources

- **MIT AI course**

- http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/

- 22 x 45 min

- ML course from Electrical Engineering and Computer Science department

Machine learning resources

- **Deep Learning @ Udacity**

- https://www.udacity.com/course/deep-learning--ud730

- Brief and clean introduction

- Assignments on TensorFlow (new python library)

Machine learning resources

- **Deep Learning**

- http://deeplearning.net/tutorial/index.html
- http://deeplearning.net/reading-list/

- Python tutorials (with Theano, python lib for tensor calculations)

- Deep understanding of deep learning

Machine learning resources

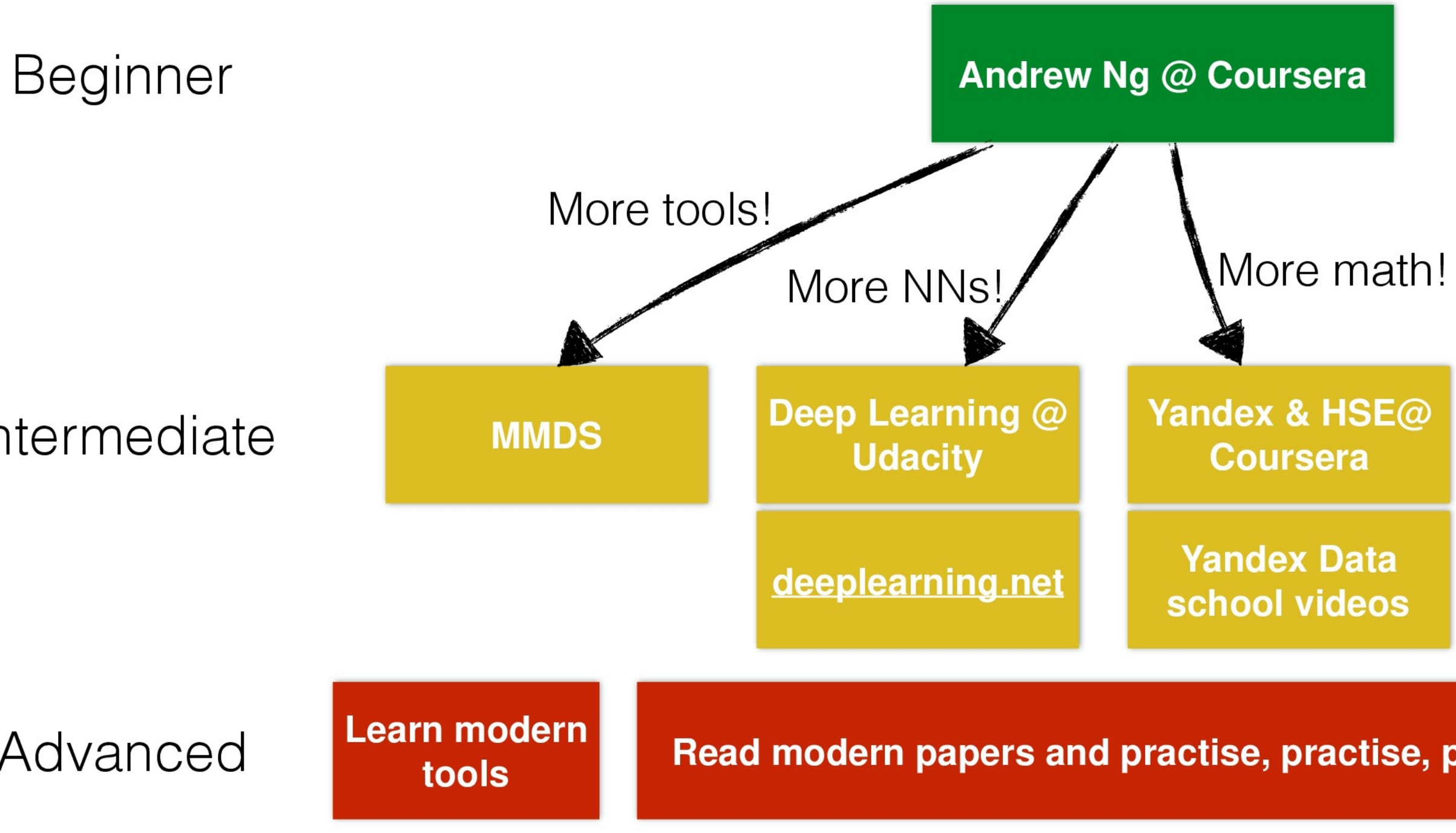
- **Reinforcement learning (David Silver)**

- http://www.youtube.com/watch?v=2pWv7GOvuf0&list=PL5X3mDkKaJrL42i_jhE4N-p6E2Ol62Ofa

- 10 x 90 min

- Great place to learn one the most hot field of machine learning

Machine learning resources



Programming

The best language?

Useful languages:

Fast languages:

Sep 2016	Sep 2015	Change	Programming Language	Ratings	Change
1	1		Java	18.236%	-1.33%
2	2		C	10.955%	-4.67%
3	3		C++	6.657%	-0.13%
4	4		C#	5.493%	+0.58%
5	5		Python	4.302%	+0.64%
6	7	↑	JavaScript	2.929%	+0.59%
7	6	↓	PHP	2.847%	+0.32%
8	11	↑	Assembly language	2.417%	+0.61%
9	8	↓	Visual Basic .NET	2.343%	+0.28%
10	9	↓	Perl	2.333%	+0.43%
11	13	↑	Delphi/Object Pascal	2.169%	+0.42%
12	12		Ruby	1.965%	+0.18%
13	16	↑	Swift	1.930%	+0.74%
14	10	⇓	Objective-C	1.849%	+0.03%
15	17	↑	MATLAB	1.826%	+0.65%
16	34	⇑	Groovy	1.818%	+1.31%
17	14	↓	Visual Basic	1.761%	+0.23%
18	19	↑	R	1.684%	+0.64%
19	44	⇑	Go	1.625%	+1.37%
20	18	↓	PL/SQL	1.443%	+0.36%

The most important language

Fast languages

- C++ or Java is must-know for developer …
- And absolutely ok not to know for analyst
- So, experience with one is a plus, but not worth it to start learning

Useful languages

- Python or R: let the holy war begin!

Useful languages

- Python or R: let the holy war begin!
- My experience: ~2 years of R programming with complete switching to python
- Python is much wider:
  - A lot of machine learning libraries
  - Fast calculations (via numpy)
  - Web development
  - Game development
  - Enterprise development

Where to learn python?

- Great intro:
  - https://www.codecademy.com/learn/python

- Another one:
  - http://learnpythonthehardway.org/book/

- Next try to solve real problems (math, financial modelling, web-parsing, Kaggle, etc.)

Strange languages

“PHP is a minor evil perpetrated and created by incompetent amateurs, whereas Perl is a great and insidious evil perpetrated by skilled but perverted professionals.”

Jon Ribbens

Tech

Linux/UNIX

- Linux is great. But how to start?

- Virtual Machine: resources needed; pointless
- Setup Jupyter notebook on Amazon AWS t2.micro
- Run a web server on Digital Ocean
- For mac users: just open Terminal app

- Intro to bash: https://www.codecademy.com/learn/learn-the-command-line

Version control systems

- Learn Git, because:
  - Data scientist often work in teams
  - There is a lot of great stuff on github
  - It's must-known to work in big companies
  - Sometimes it's good even for personal long-term big projects
  - And it's cool to have popular github account

- Use on of these:
  - https://www.codecademy.com/learn/learn-git
  - https://www.codeschool.com/learn/git
  - https://try.github.io

“Data Scientist:
The Sexiest Job of the 21st Century.”

–Harvard Business Review

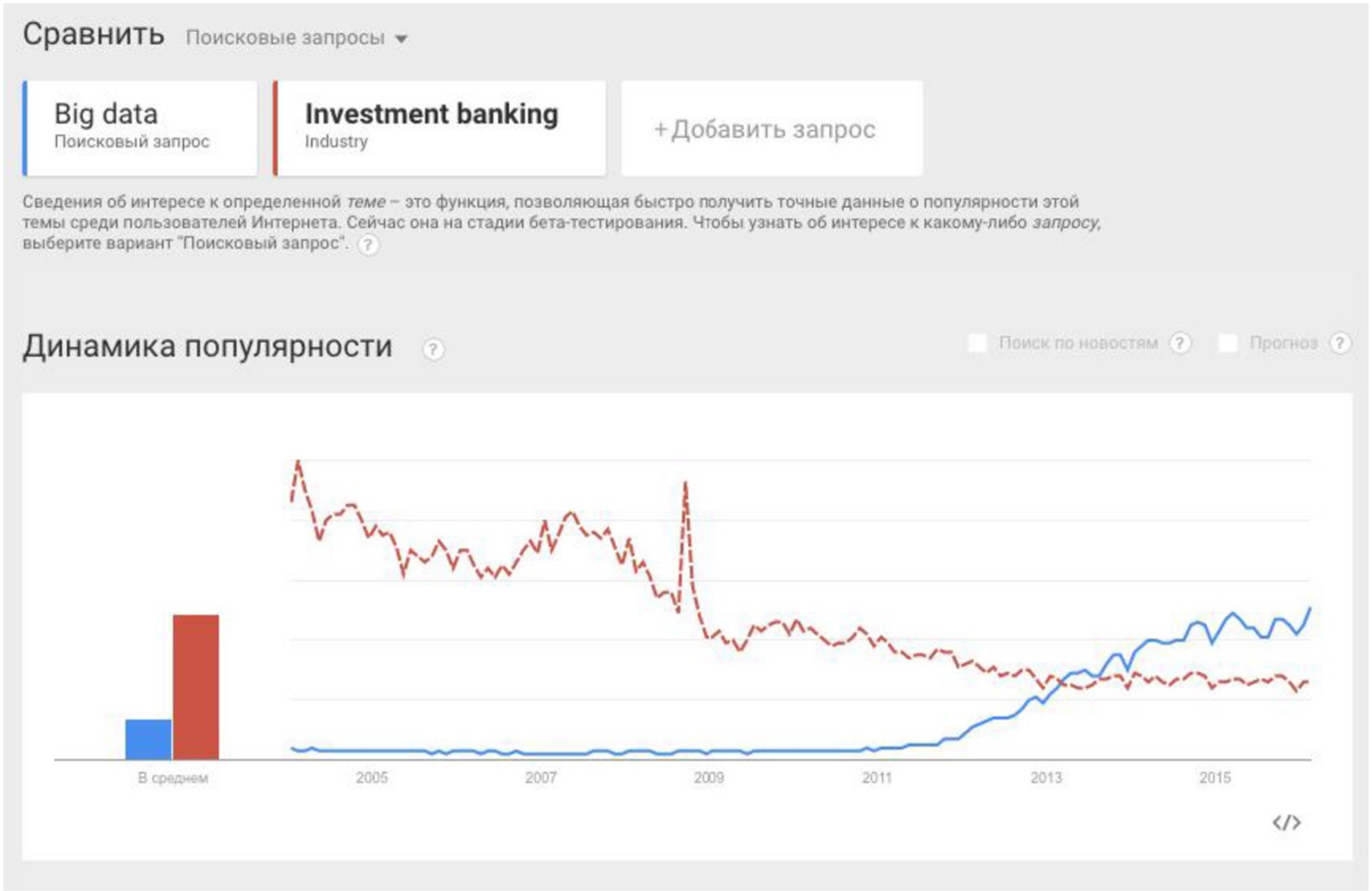
Pt. II

Introduction to Big Data

Agenda

- Limitation of classical data analysis
- Distributed filesystems
- Data Centres
- MapReduce computational model
- Few words about Big Data world
- Study plan

Popularity



trends.google.com


amazon.com®

amazon.com®

item-item collaborative filtering patent:

100x increase in sales

Classical data analysis

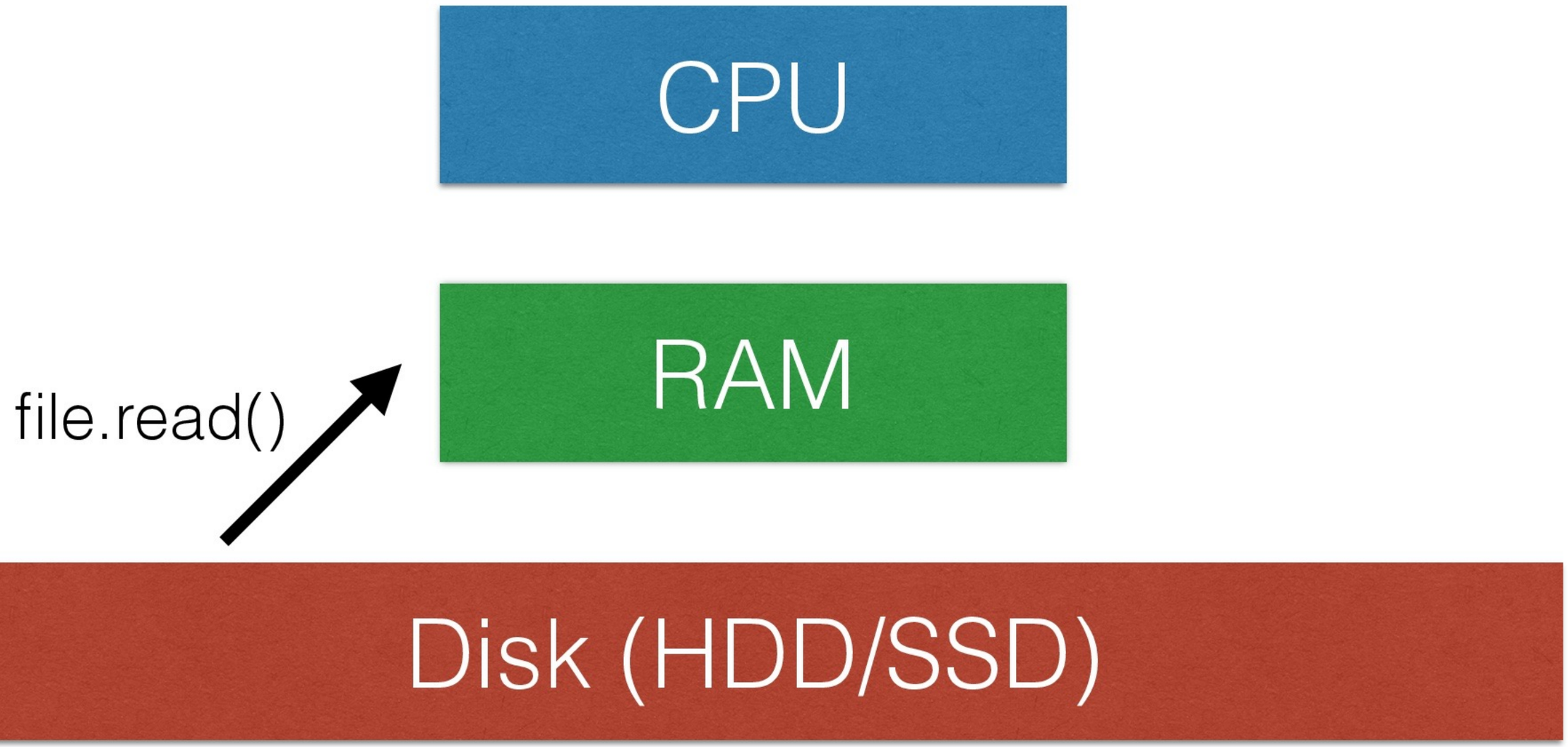
Super simplified computer

CPU

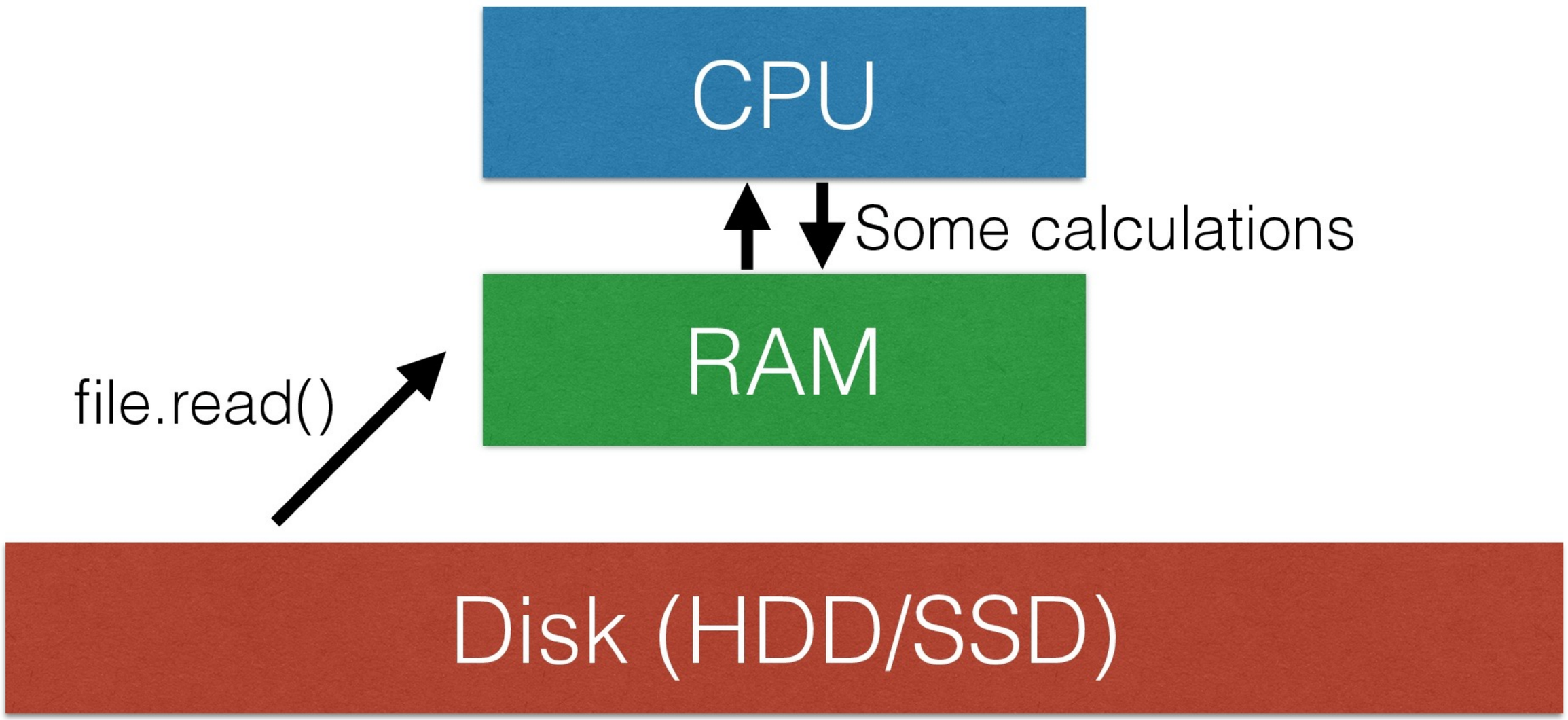
RAM ~8GB

Disk (HDD/SSD) ~1 TB

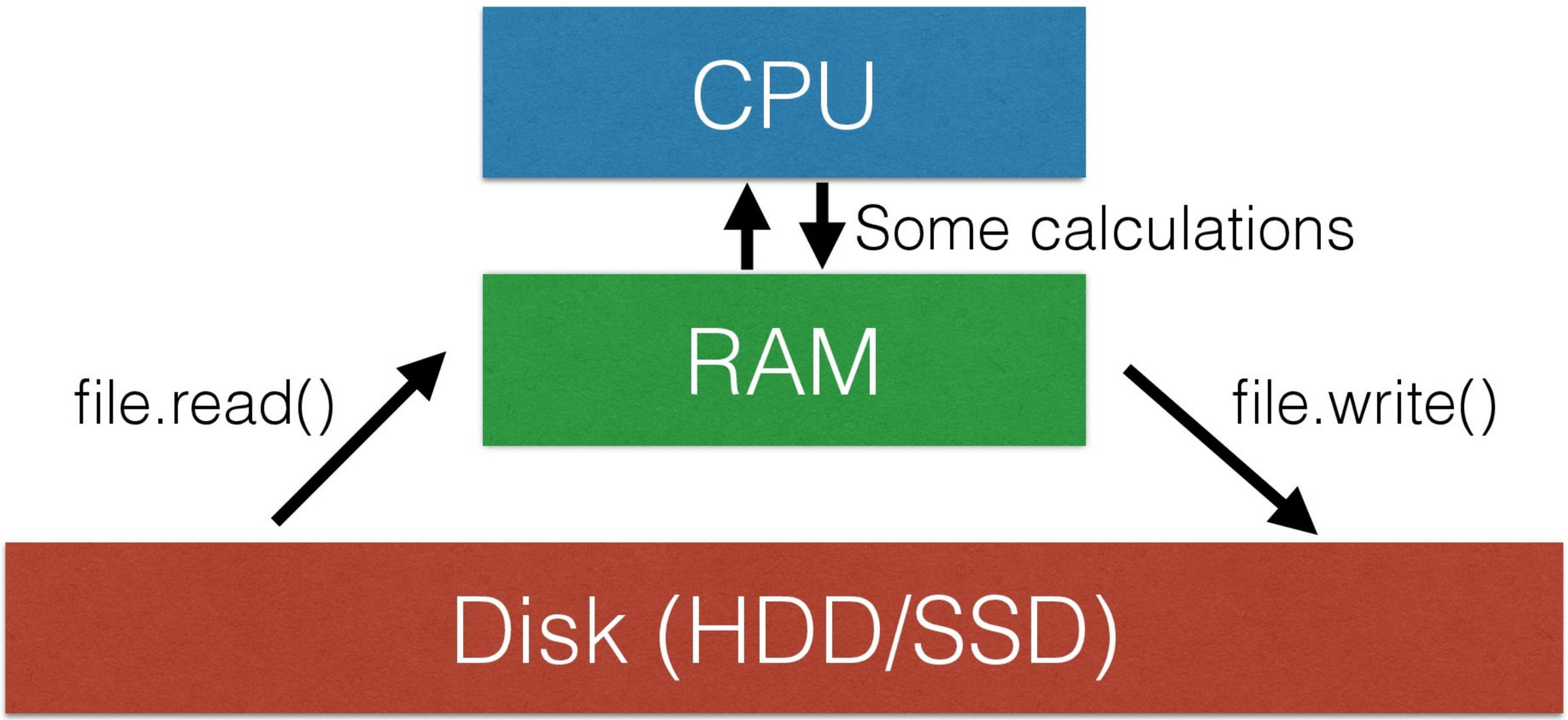
Super simplified computer



Super simplified computer



Super simplified computer



The Internet

- There are 993,059,597 websites online right now

http://www.internetlivestats.com/
http://www.webperformancetoday.com/2012/05/24/average-web-page-size-1-mb/

The Internet

- There are 993,059,597 websites online right now
- 10 pages per site => 10 billion pages

http://www.internetlivestats.com/
http://www.webperformancetoday.com/2012/05/24/average-web-page-size-1-mb/

The Internet

- There are 993,059,597 websites online right now
- 10 pages per site => 10 billion pages
- 1 MB per page

http://www.internetlivestats.com/
http://www.webperformancetoday.com/2012/05/24/average-web-page-size-1-mb/

The Internet

- There are 993,059,597 websites online right now
- 10 pages per site => 10 billion pages
- 1 MB per page
- 10 000 TB ! (~ 10 PB)
- Wikipedia is “only” 10 TB

http://www.internetlivestats.com/
http://www.webperformancetoday.com/2012/05/24/average-web-page-size-1-mb/

Let’s count words!

- Read web page into memory

Let’s count words!

- Read web page into memory
- Count words on the page

Let’s count words!

- Read web page into memory
- Count words on the page
- Add numbers to dictionary {(word, count)}

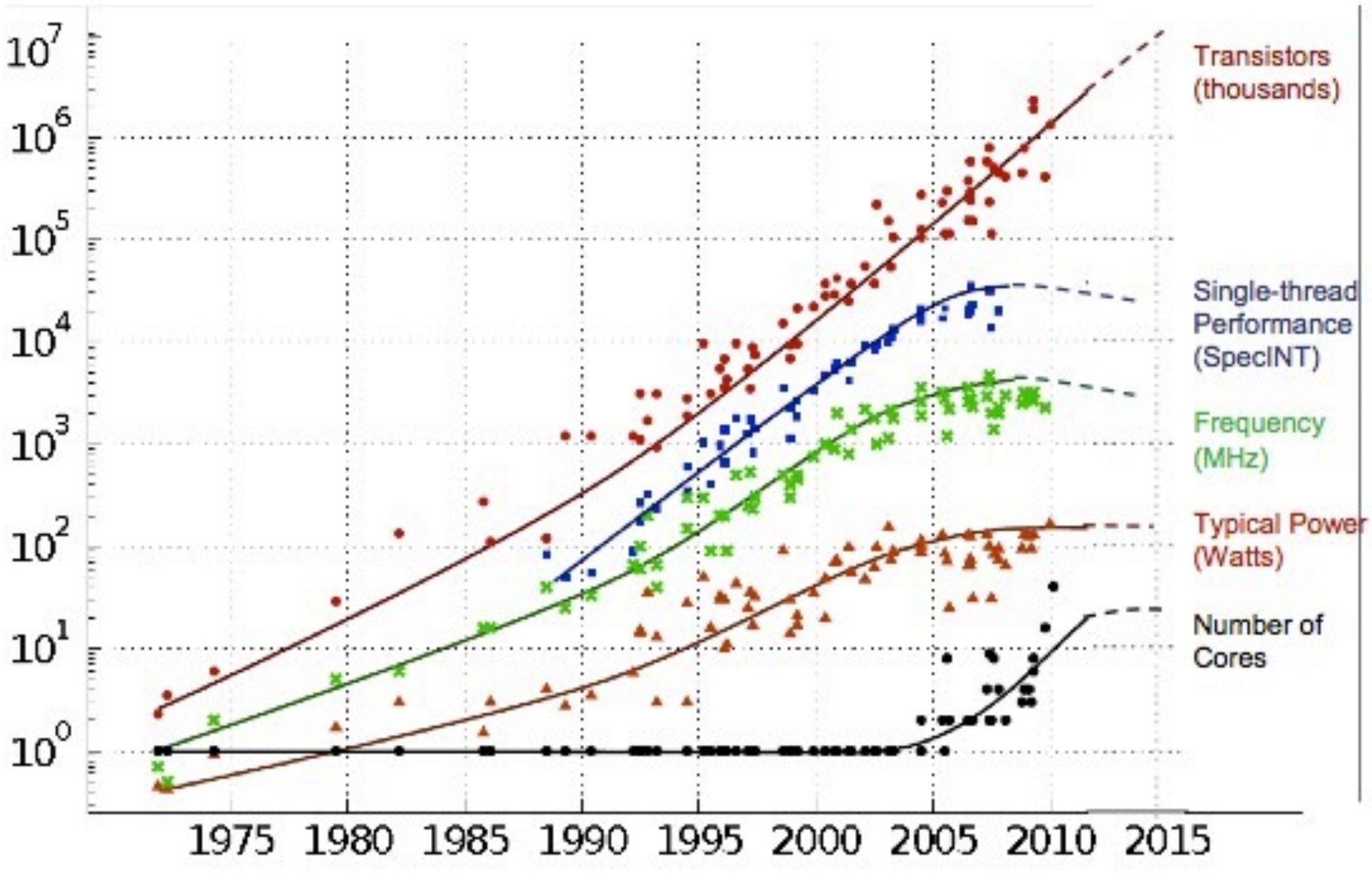
Let’s count words!

- Read web page into memory
- Count words on the page
- Add numbers to dictionary {(word, count)}
- Repeat!

WordCount algorithm

- 10 000 TB to process
- 100 MB/s HDD read speed
- 1200+ days
- Even with infinite RAM and top CPU the run time is more than 3 years!

Maybe faster CPU/HDD/RAM?

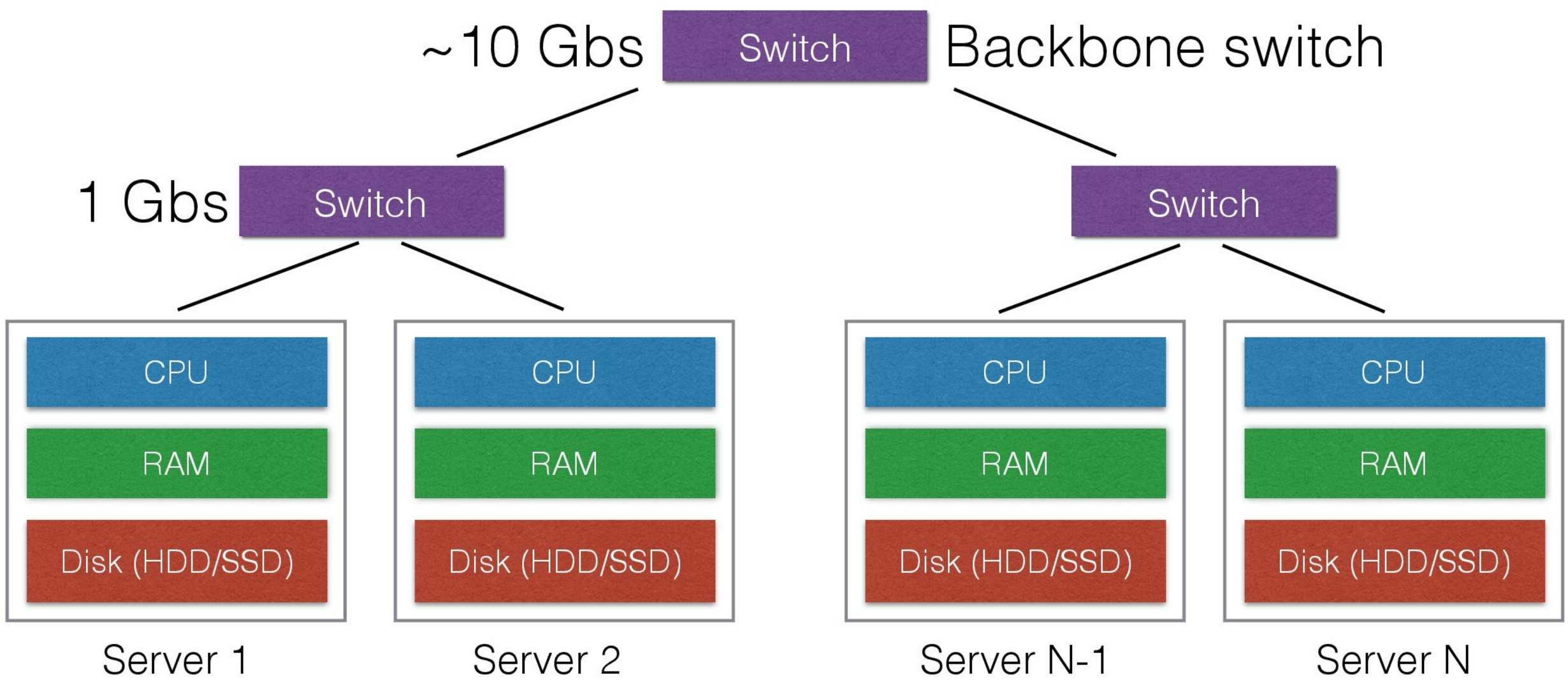


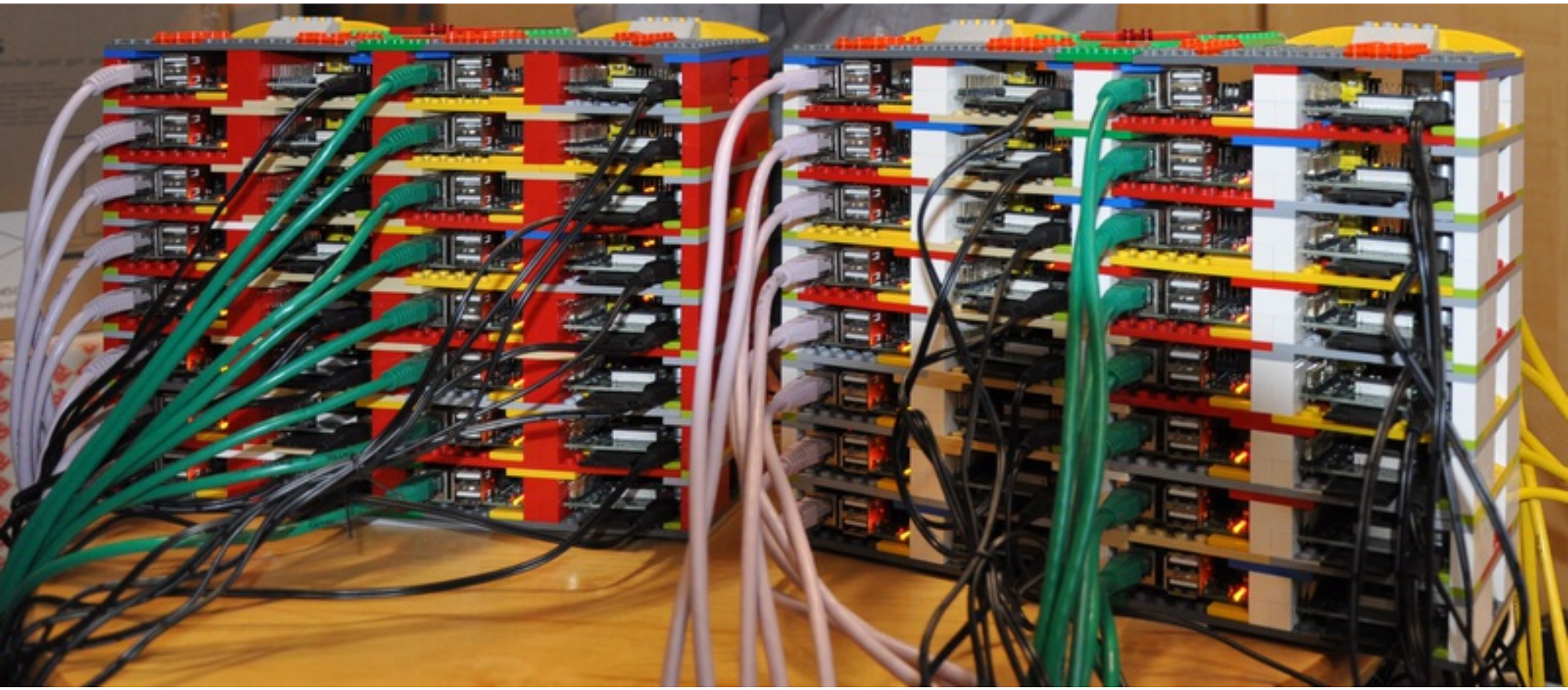
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

More machines

Parallelisation

- The only possible way is to use many machines
- Machines are connected to racks (2-50 machines in rack)





64 Raspberry PI


google.com/datacenters

Node failures

- Hard-loaded computer fails every 3 years (1000 days)
- 10K servers in data center …
- 10 failures/day

Node failures

- Hard-loaded computer fails every 3 years (1000 days)
- 10K servers in data center …
- 10 failures/day

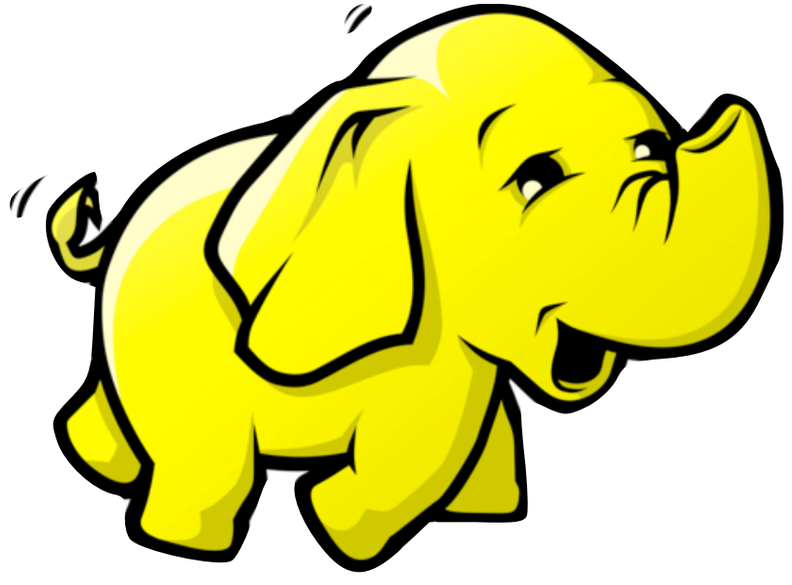
- How to save the data?
- How to deal with computations?

Distributed File System

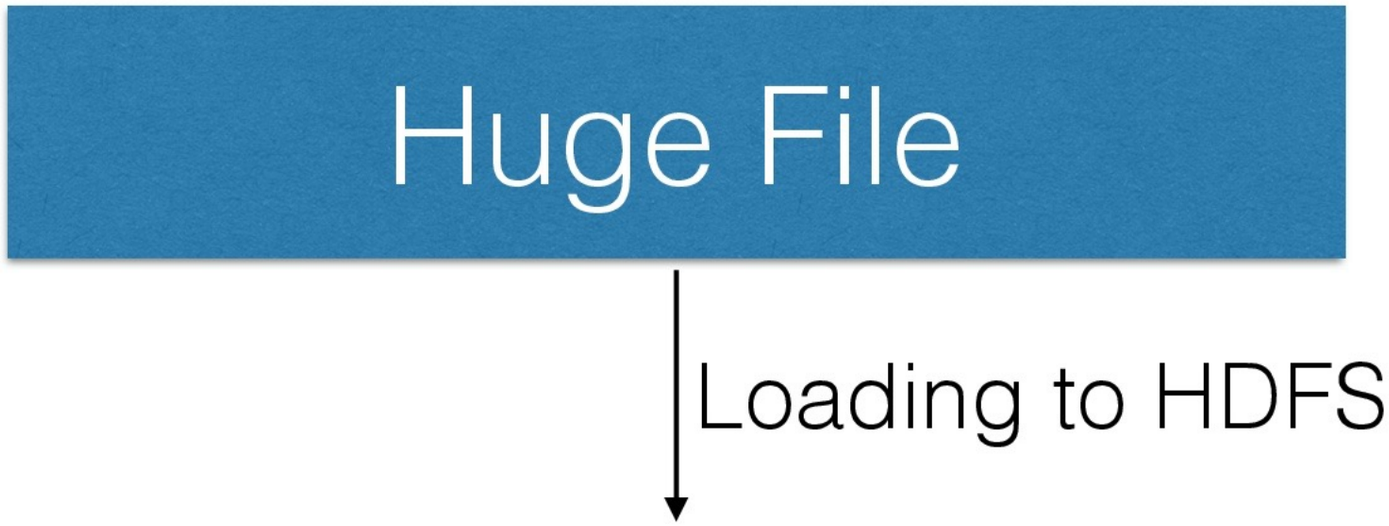
- It's not good to store huge files on single server
- Backups are needed

- Most famous realisations:
  - GFS (Google File System), closed
  - HDFS (Hadoop Distributed File System), open-source, part of Hadoop project

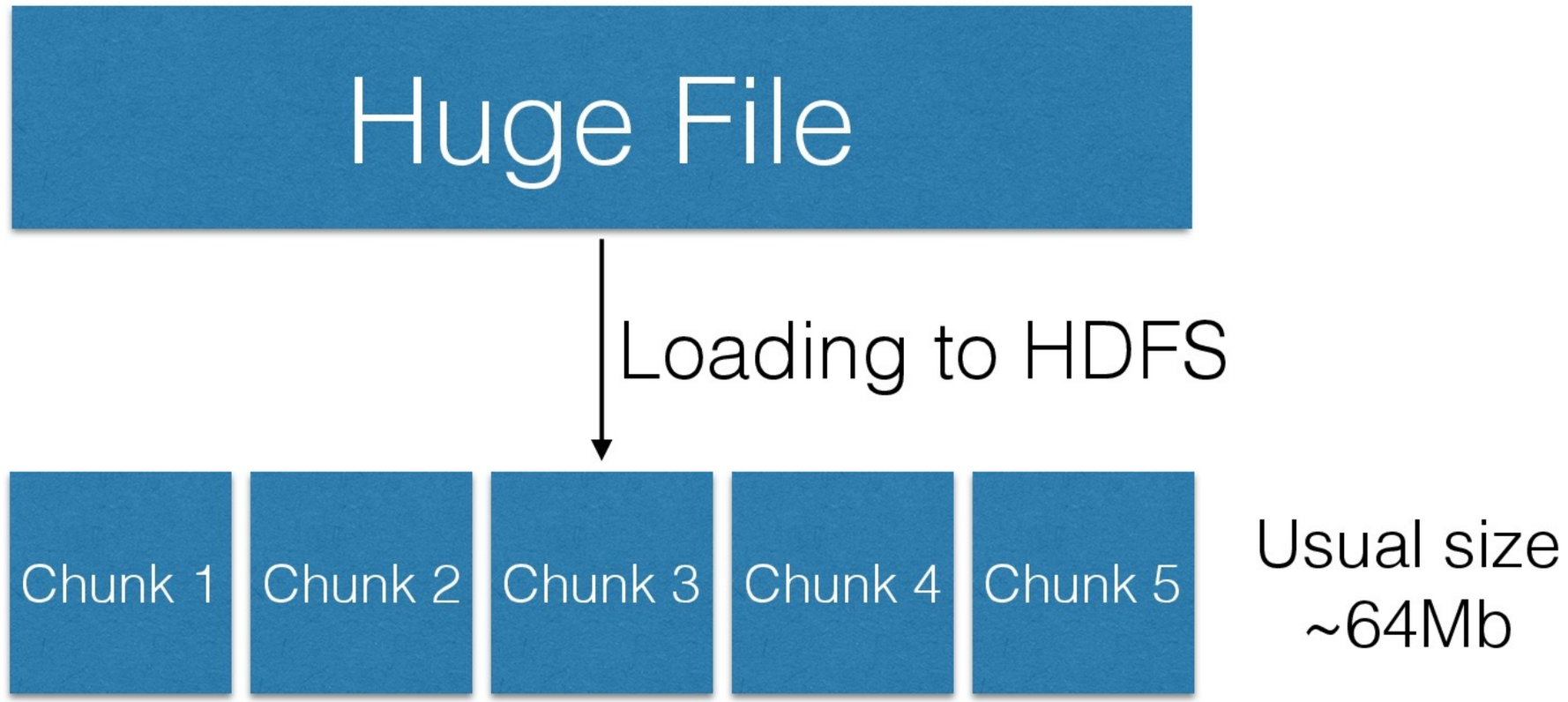
- We will talk about second one



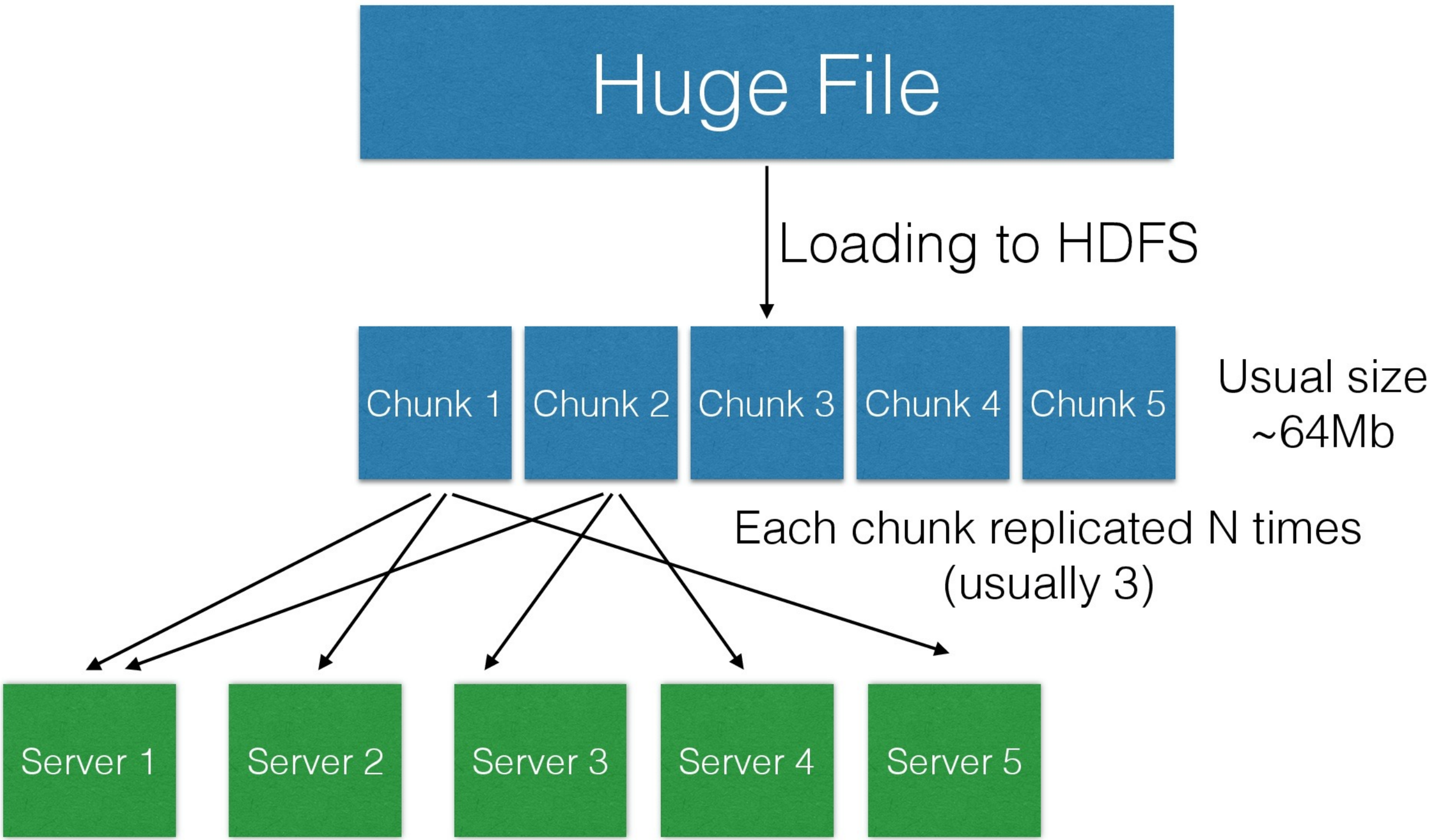
Distributed File System



Distributed File System



Distributed File System



Distributed File System

Huge File

Loading to HDFS

Chunk 1 | Chunk 2 | Chunk 3 | Chunk 4 | Chunk 5

Usual size
~64Mb

Each chunk replicated N times
(usually 3)

Server 1 | Server 2 | Server 3 | Server 4 | Server 5

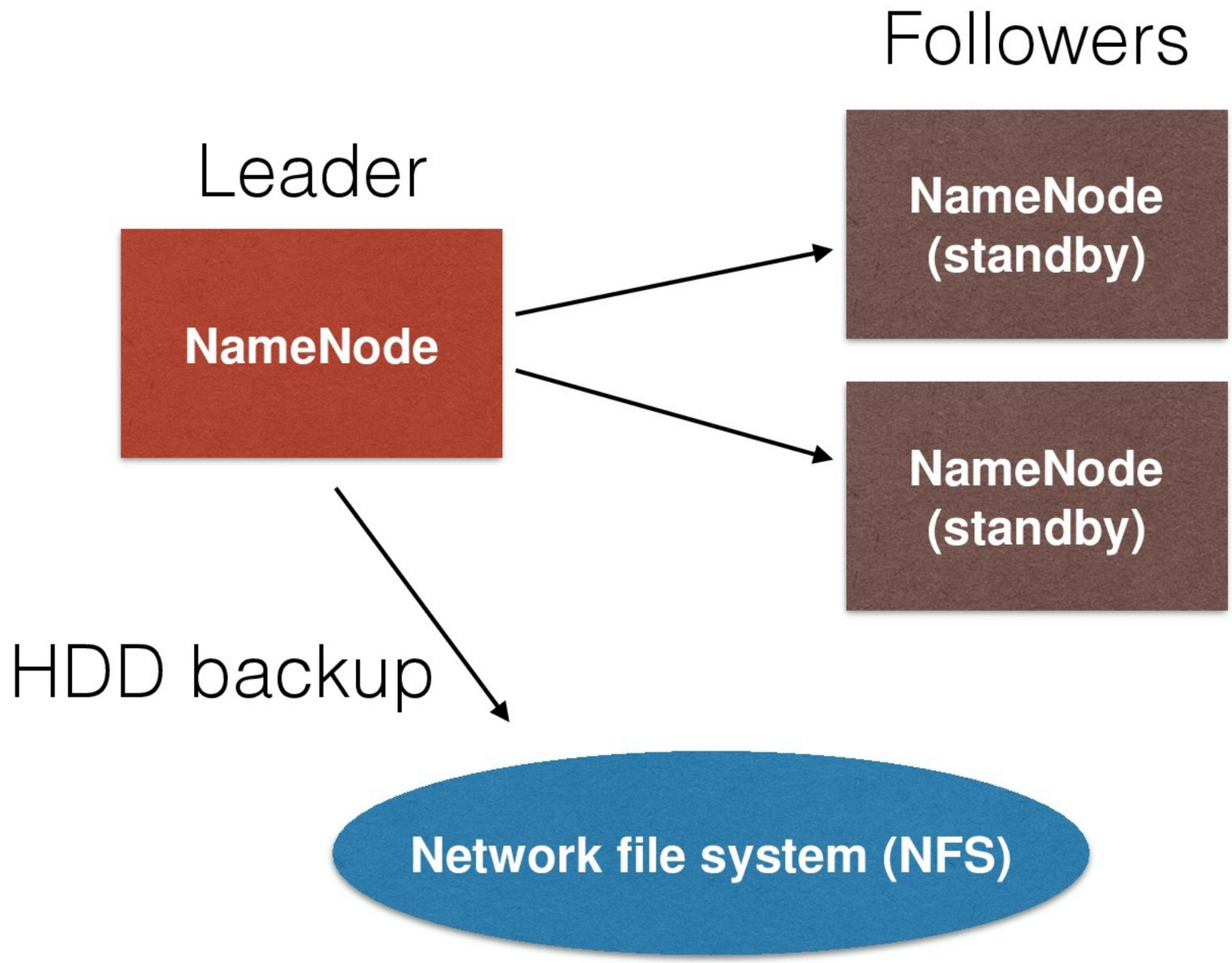
NameNode

Knows where to find chunks

Where are points of failure?

1. Backbone switch (no connection between racks)
2. Rack switch (no connection in the rack)
3. Servers's HDD
4. NameNode network connection
5. NameNode HDD

NameNode


Followers
Leader
NameNode
NameNode (standby)
NameNode (standby)
HDD backup
Network file system (NFS)

Computations

Parallel sorting

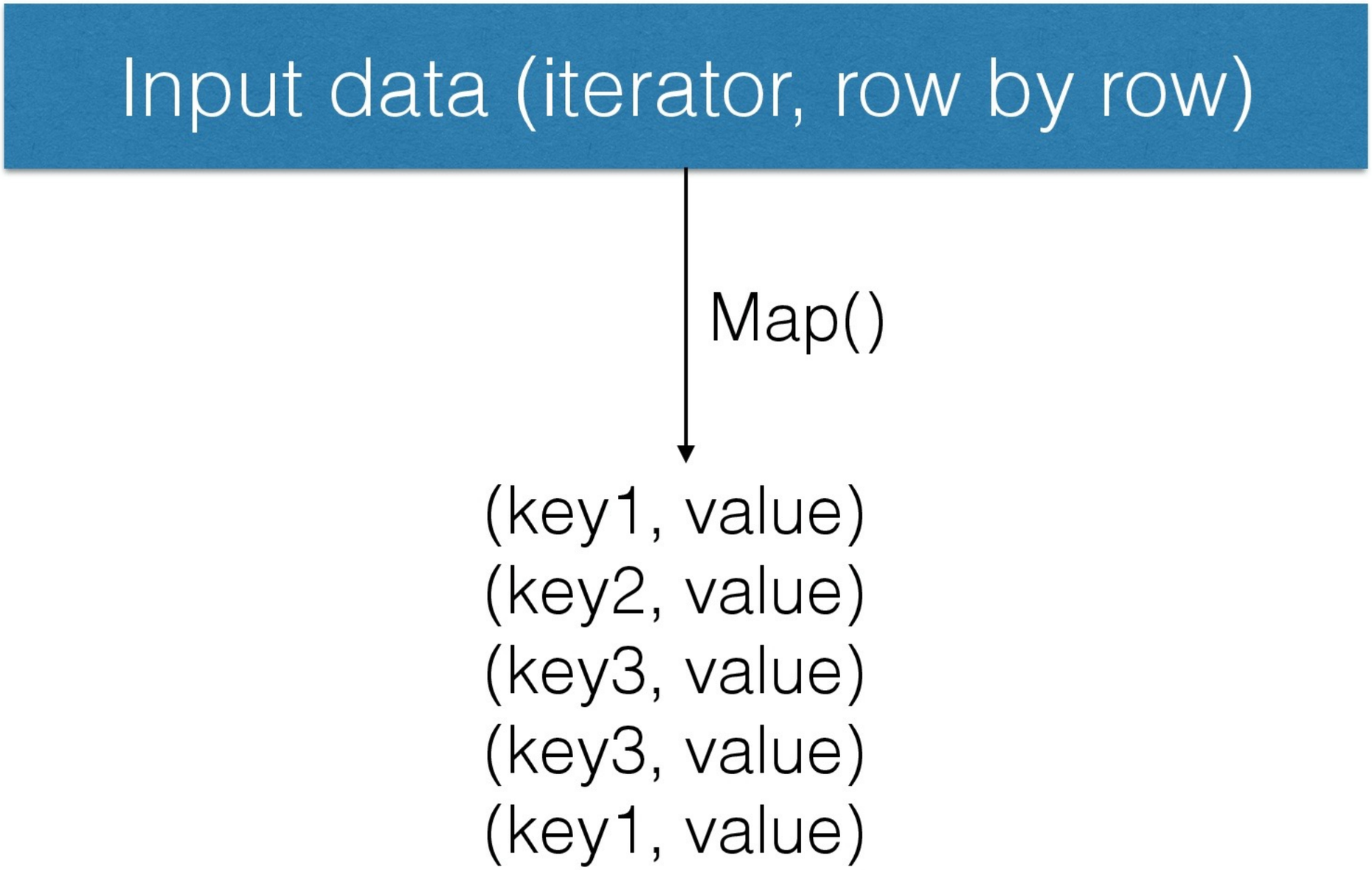
- Various algorithms (e.g. Merge Sort)
- Google results (on 10k machines data center):
  - 2007: 1 PB / 12.13 hours
  - 2008: 1 PB / 6.03 hours
  - 2010: 1 PB / 2.95 hours
  - 2011: 1 PB / 0.55 hours
  - 2012: 50 PB / 23 hours
  - Why did not they go on?

https://cloud.google.com/blog/big-data/2016/02/history-of-massive-scale-sorting-experiments-at-google

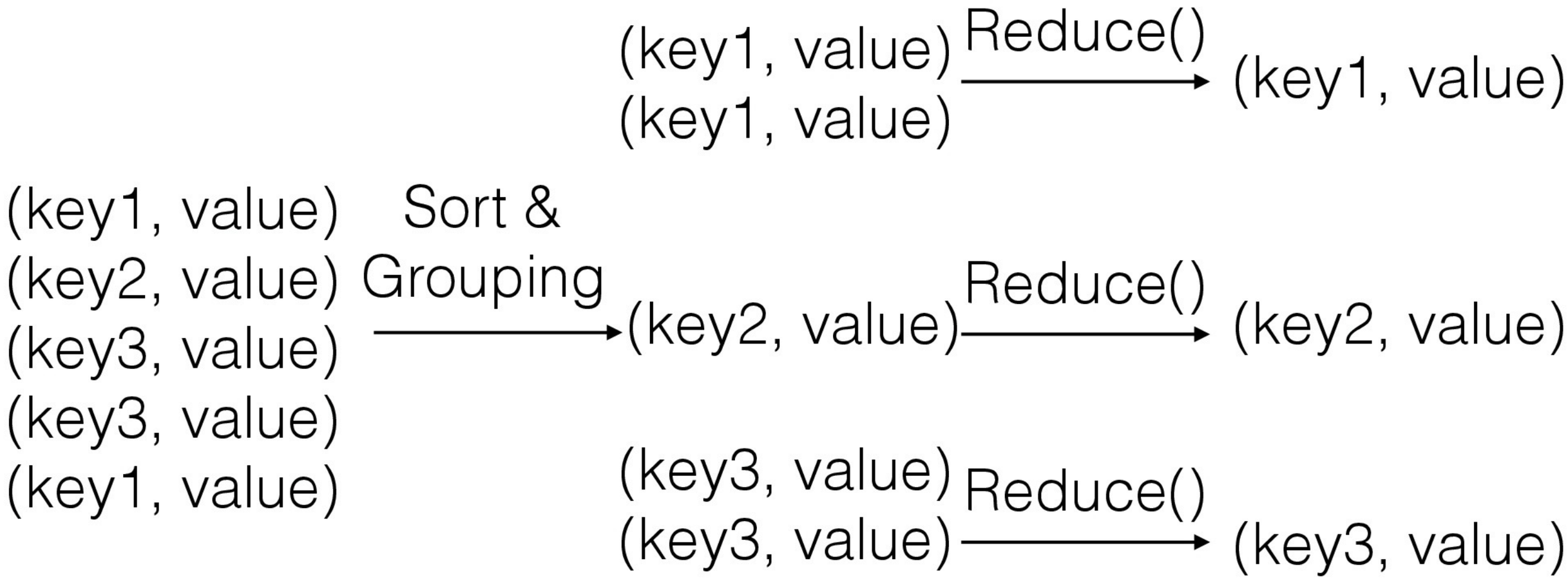
MapReduce

- Largle-scale computational model
- Released by Google (known since 1995)
- Natively parallelised
- Various problems could be solved
- **Must-know on any data scientist interview**

Step one : Map


Input data (iterator, row by row)
Map()
(key1, value)
(key2, value)
(key3, value)
(key3, value)
(key1, value)

Step two : Reduce



Example

We need to calculate revenue by city of the international shop

Shop	Category	Value	Price	Revenue
Moscow	closes	1	12	12
London	closes	1	8	8
Moscow	music	2	5	10
Moscow	toys	12	5	60
Paris	music	4	100	400
London	closes	1	4	4
Paris	music	6	6	36

Example

Map: Take row and return (city, revenue)
Reduce: Sum all values for the key

Shop	Revenue
Moscow	12
London	8
Moscow	10
Moscow	60
Paris	400
London	4
Paris	36

Map → (Moscow, 12) (London, 8) (Moscow, 10) (Moscow, 60) (Paris, 400) (London, 4) (Paris, 36)

Sort → (London, 8) (London, 4) (Moscow, 12) (Moscow, 10) (Moscow, 60) (Paris, 400) (Paris, 36)

Reduce → (London, 12) (Moscow, 82) (Paris, 436)

More examples

- Revenue by category
- Revenue by shop and category
- Mean revenue by the shop
- Uniq stores
- Histogram of sales

Single machine map reduce

```
$cat data.txt | python mapper.py | sort | python reducer.py
```

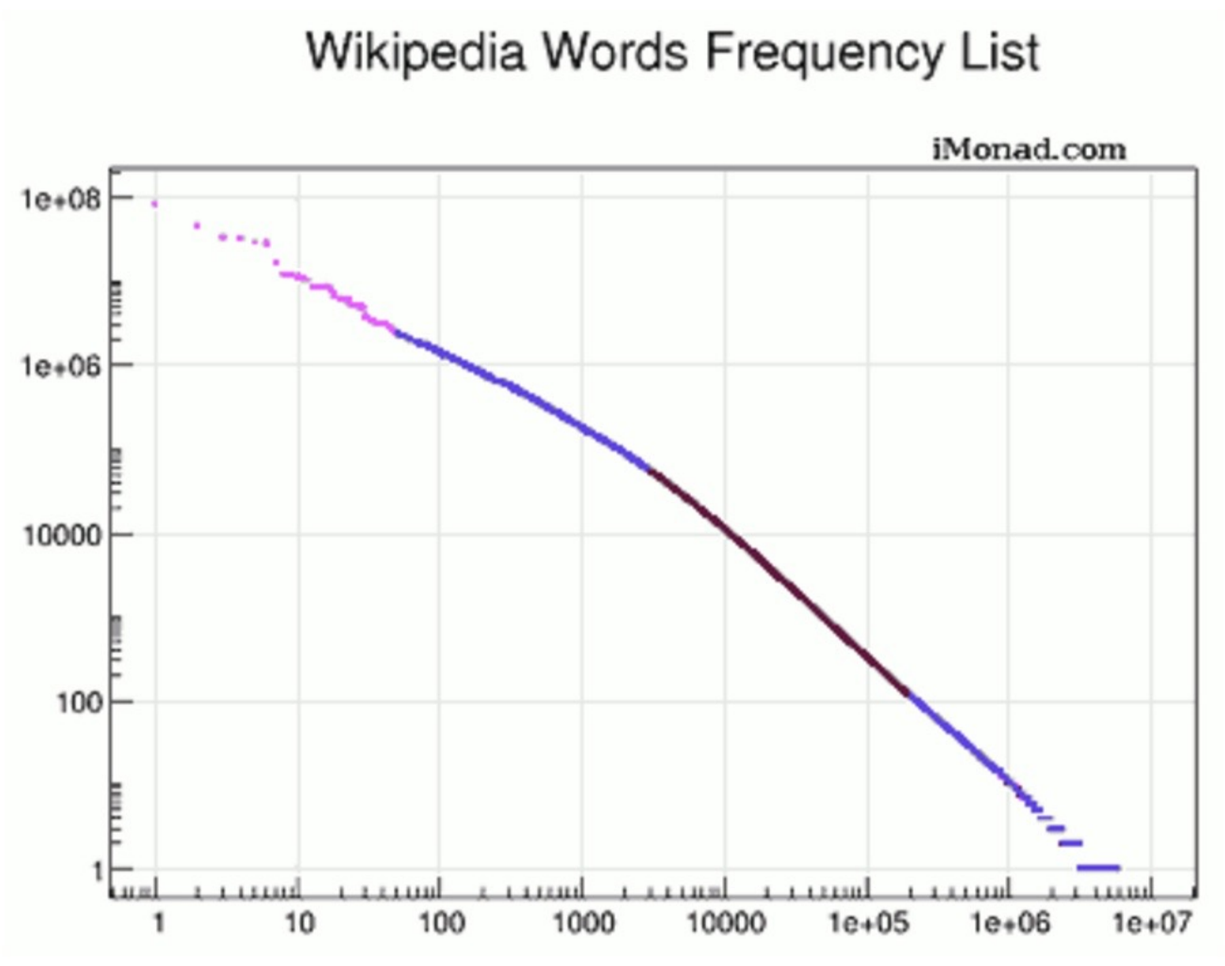
```
mapper.py:
import sys
for row in sys.stdin:
    print processor(row)
```

```
reducer.py:
import sys
for rows in grouped(sys.stdin):
    print processor(rows)
```

More about WordCount

- “Hello World” of MapReduce: Word count
  (on 10000 machines 1200 days become 4 hours)

- The one problem - “monsters”
  some reducers will get ~1e8 (key, value) pairs



Zipf’s law:
TF ~ 1/Rank

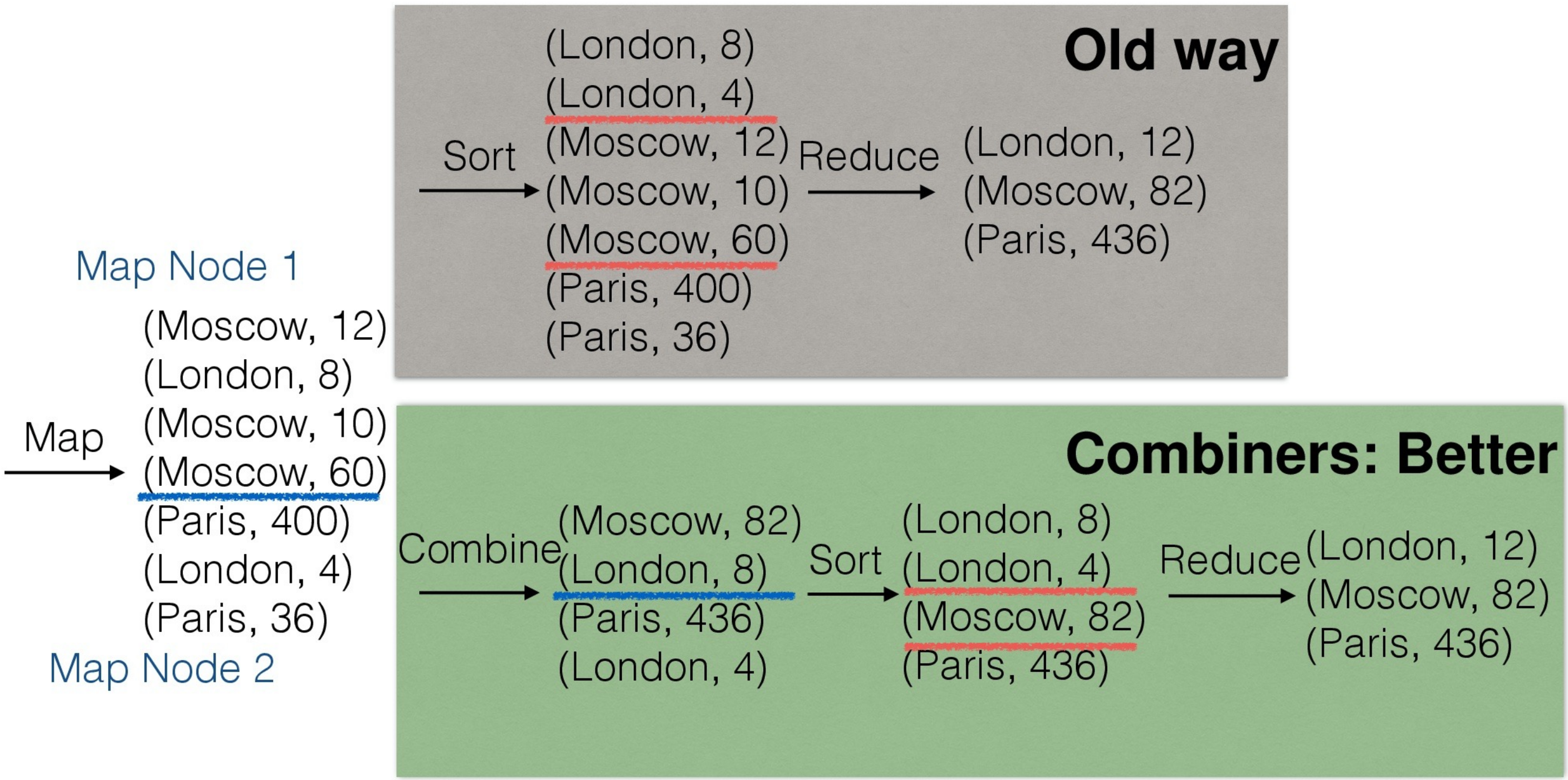
Map aggregators

- Old WordCount map: `yield (word, 1)`
- New WordCount map:

```
for row in rows:
      counter[word] += 1
for key in counter:
     yield (key, counter[key])
```

Combiners

- Combiners are reduce function (usually) run on the map node after mapping, before sorting



Restrictions on combiners

- Commutative and associative
  - f(a,b) = f(b,a)
  - f(a,f(b,c)) = f(f(a,b),c)

- Sum, prod
- Mean- why? how to solve the problem?
- Median, quantiles - why? what to do?

Join two tables?

Table 1 (key, value1)

key1, a
key2, b
key3, c

Table 2 (key, value2)

key1, x
key2, y
key4, z

Joined (key, value1, value2)

Inner
key1, a, x
key2, b, y

Left
key1, a, x
key2, b, y
key3, c, NaN

Right
key1, a, x
key2, b, y
key4, NaN, z

Outer
key1, a, x
key2, b, y
key3, c, NaN
key4, NaN, z

Environment

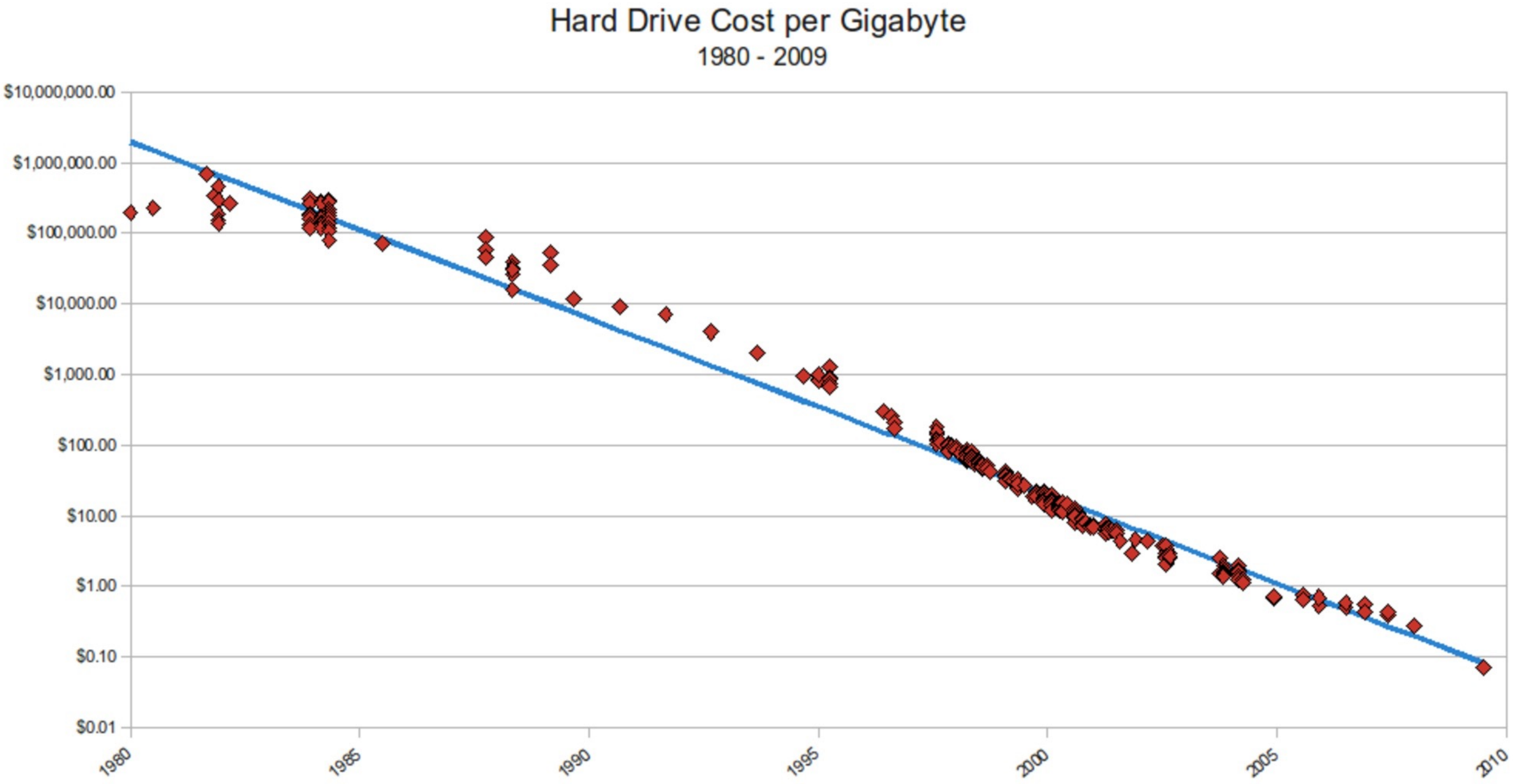
- Partitioning
- Scheduling
- Running processes near the data
- Grouping
- Handling failures
- Managing all inter-machine communications
  (you need just to specify two functions on any language)

What about node failure?

1. Map Node failure during running?
2. Reduce Node failure during running?
3. Master Node?

Few more words about Big Data

Why?


Hard Drive Cost per Gigabyte
1980 - 2009
$10,000,000.00
$1,000,000.00
$100,000.00
$10,000.00
$1,000.00
$100.00
$10.00
$1.00
$0.10
$0.01
1980
1985
1990
1995
2000
2005
2010

3V

1. **Volume**
2. **Variety**

   If something in the data may be wrong, it will
3. **Velocity**

   Yandex Real Time Crypta: 250k RPS, 15 TB/day

   (Wikipedia: 30-70k RPS, Reddit DDoS: 400k)

Correlations

- On the enormous amount of samples even weak correlations become meaningful

Observation: people buy beer with diapers

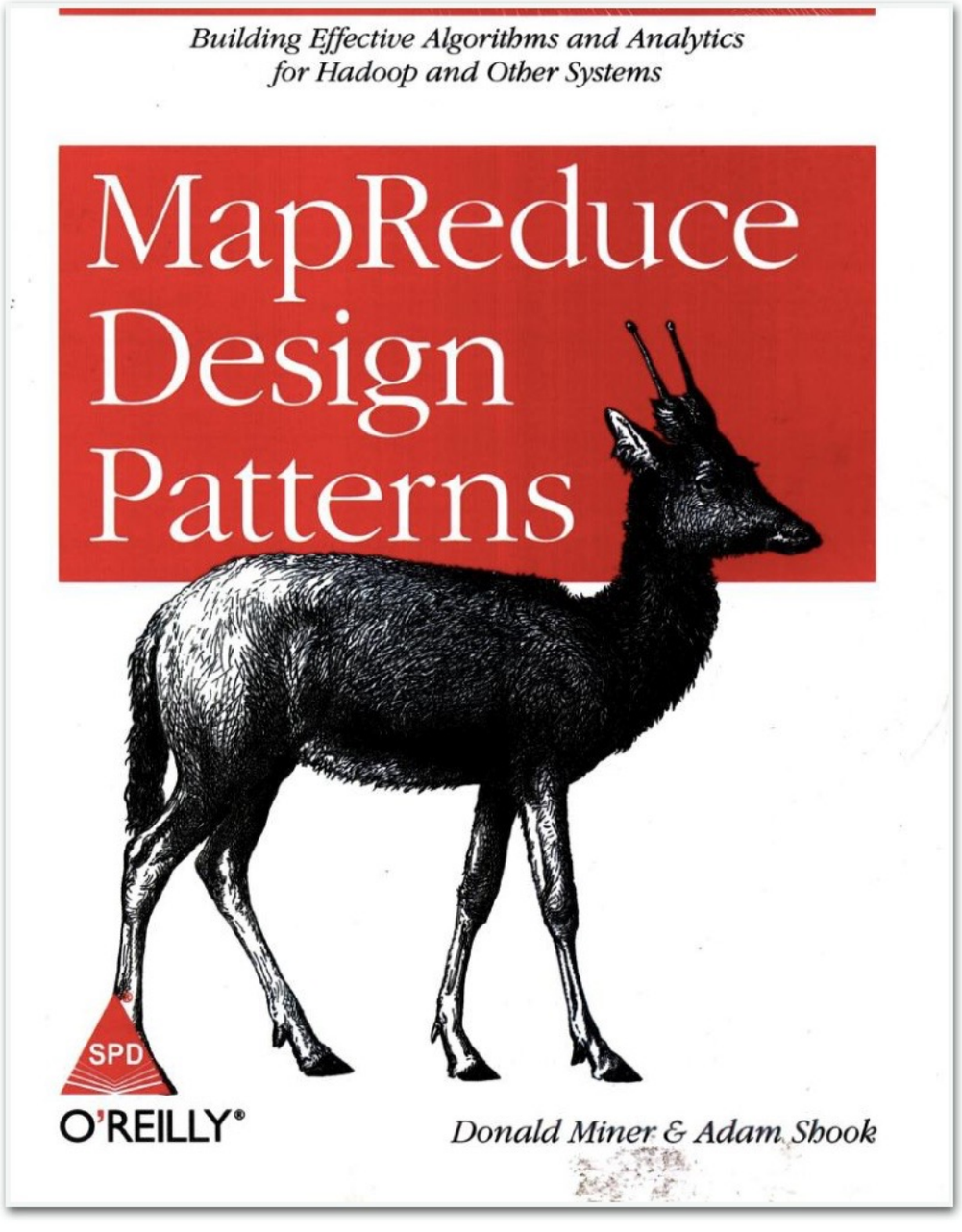
- The classic way: check p-value, use Granger causality test.
- The Big Data way: doesn't matter.
  Let's just make money on this correlation

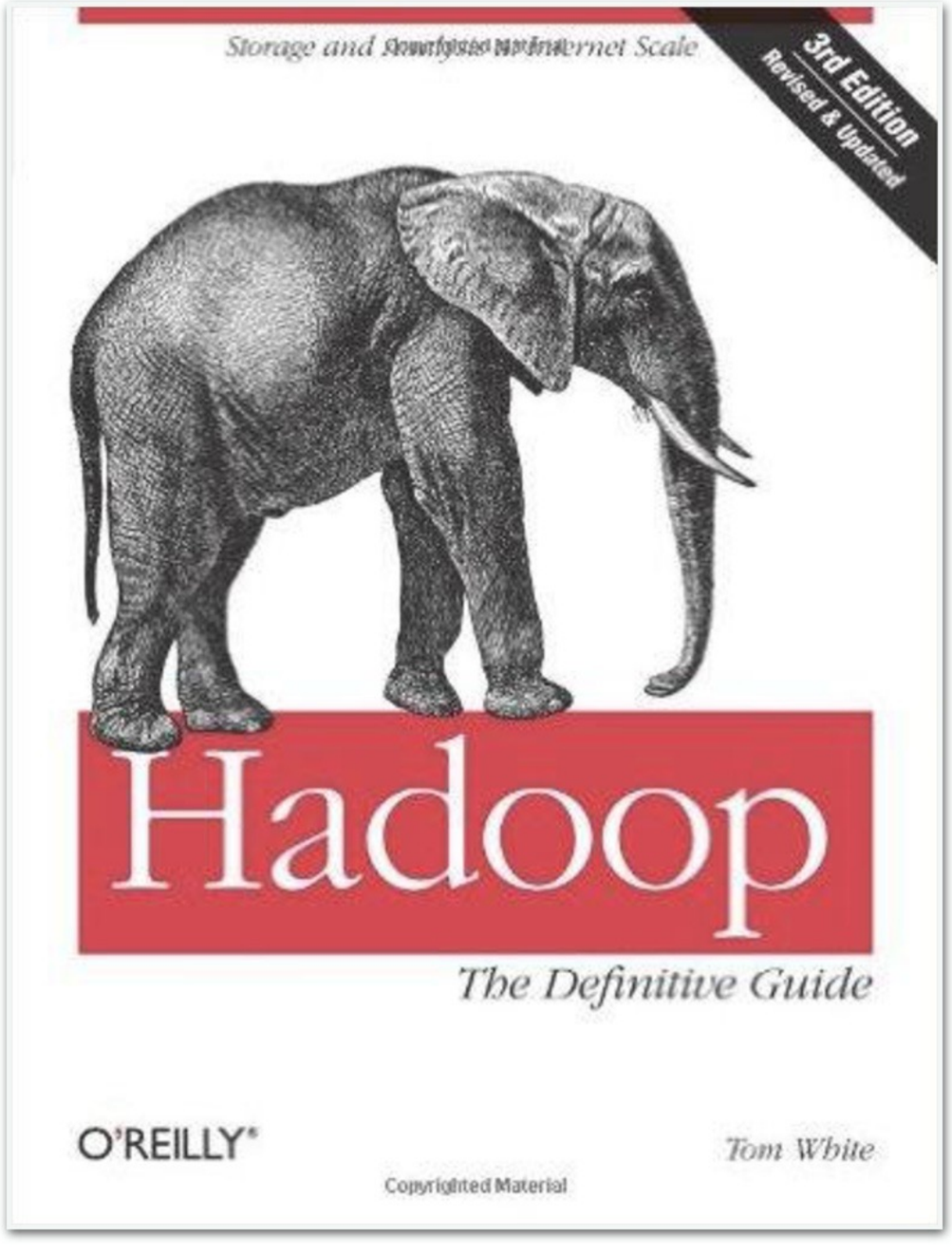
What to do next?

Read a popular book

Russian translation is also good

Read a tech book





Get familiar with Hadoop

- Check out the CDH by Cloudera
- Read about Hive & Pig
- Run local single-node pseudo cluster
- Play with Amazon AWS EMR (Elastic Map Reduce)

  10 machines for 0.15$/hour
- Run your own Hadoop on AWS

Study

- Mining Massive Data Sets @ Coursera, mmds.org
- Introduction to Hadoop and MapReduce @ Udacity
- Big Data Specialisation @ Coursera

“A real data scientist(TM) can implement algorithms, write proofs, setup Hadoop clusters, perform RCA, talk to clients, and doesn’t exist.”

Somewhere on Twitter