Introduction to Big Data

Responsible:
Dr. Tegawendé F. BISSYANDE
tegawende.bissyande@uni.lu

Teacher Assistant:
Médéric Hurier
mederic.hurier@uni.lu

Overview

The course is about (classical and new) techniques that are involved in the context of Big Data

The course combines two key dimensions of Big Data, namely:

1- **Databases and Query Models for Big Data**:
from classical to large-scale database systems

2- **Data Analysis and Machine Learning**:
concepts related to artificial intelligence

Planning

Course Times

Every Tuesday at 10h30 (2h15)

Evaluation

Practical assignments and homeworks

Webpage

https://moodle.fstc.uni.lu/course/view.php?id=744

Course Features

- Sep. 20th - **Introduction to Big Data**

Part 1. Databases and Query Models for Big Data

- Sep. 27th - **Relational Databases: Reminders**
- Oct. 4th - **Relational Databases: Internals**
- Oct. 11th - **NoSQL Databases**
- Oct. 18th - **MapReduce Model**
- Oct. 25th - **Hadoop and Spark**
- Nov. 8th - **Datalog Model**

Part 2. Data Analysis and Machine Learning

- Nov. 15th - **Statistics and Probabilities**
- Nov. 22th - **Communication and Visualization**
- Nov. 29th - **Features Engineering and Supervised Learning**
- Dec. 5st - **Unsupervised and Reinforcement Learning**
- Dec. 12th - **Homework Time**


Enter the Matrix ...

One definition of Big Data

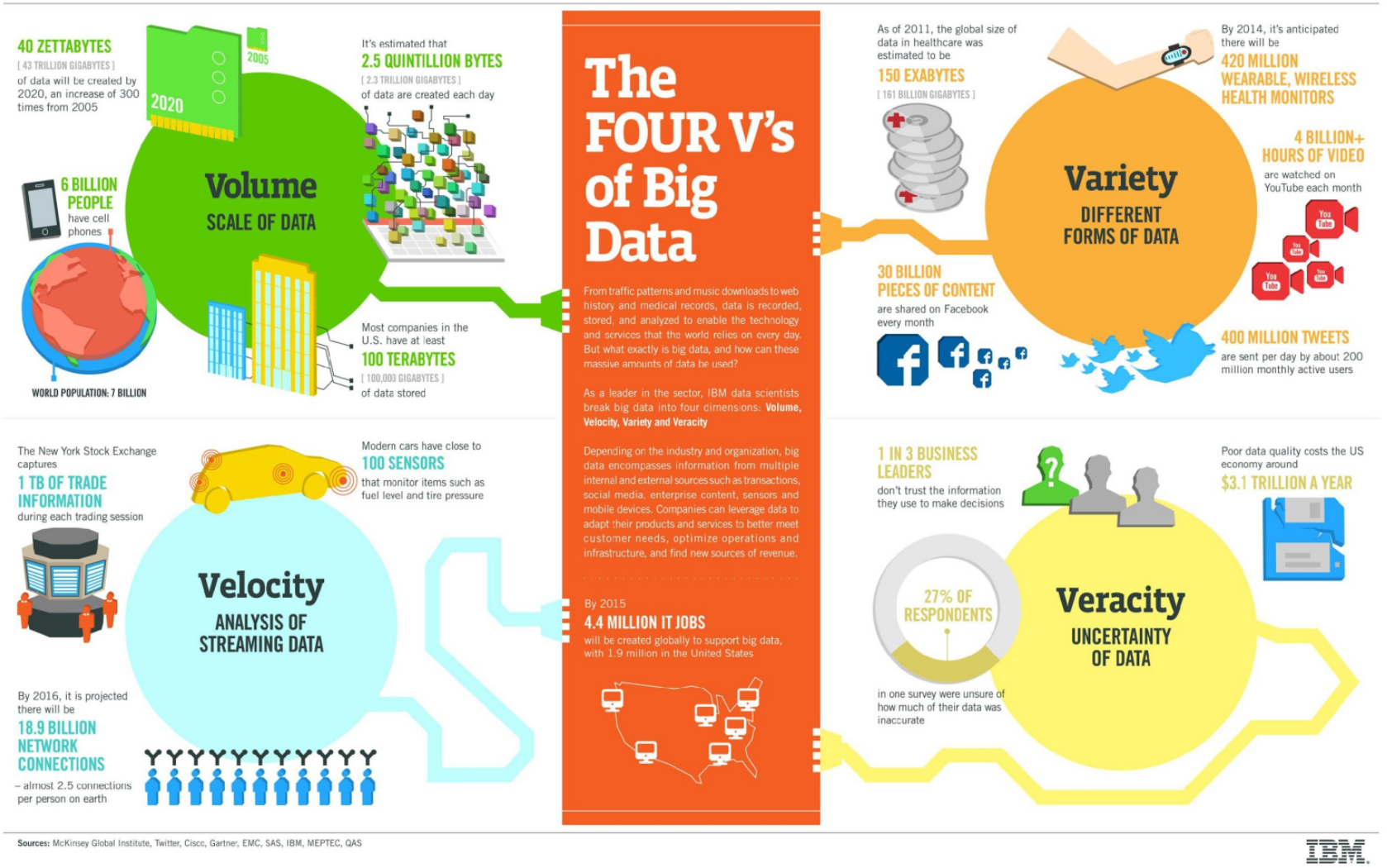
"Big Data is any data that is expensive to manage and hard to extract value from."

Michael Franklin,

Director of the Algorithms, Machines and People lab University of Berkeley

The key idea is that "BIG" is relative

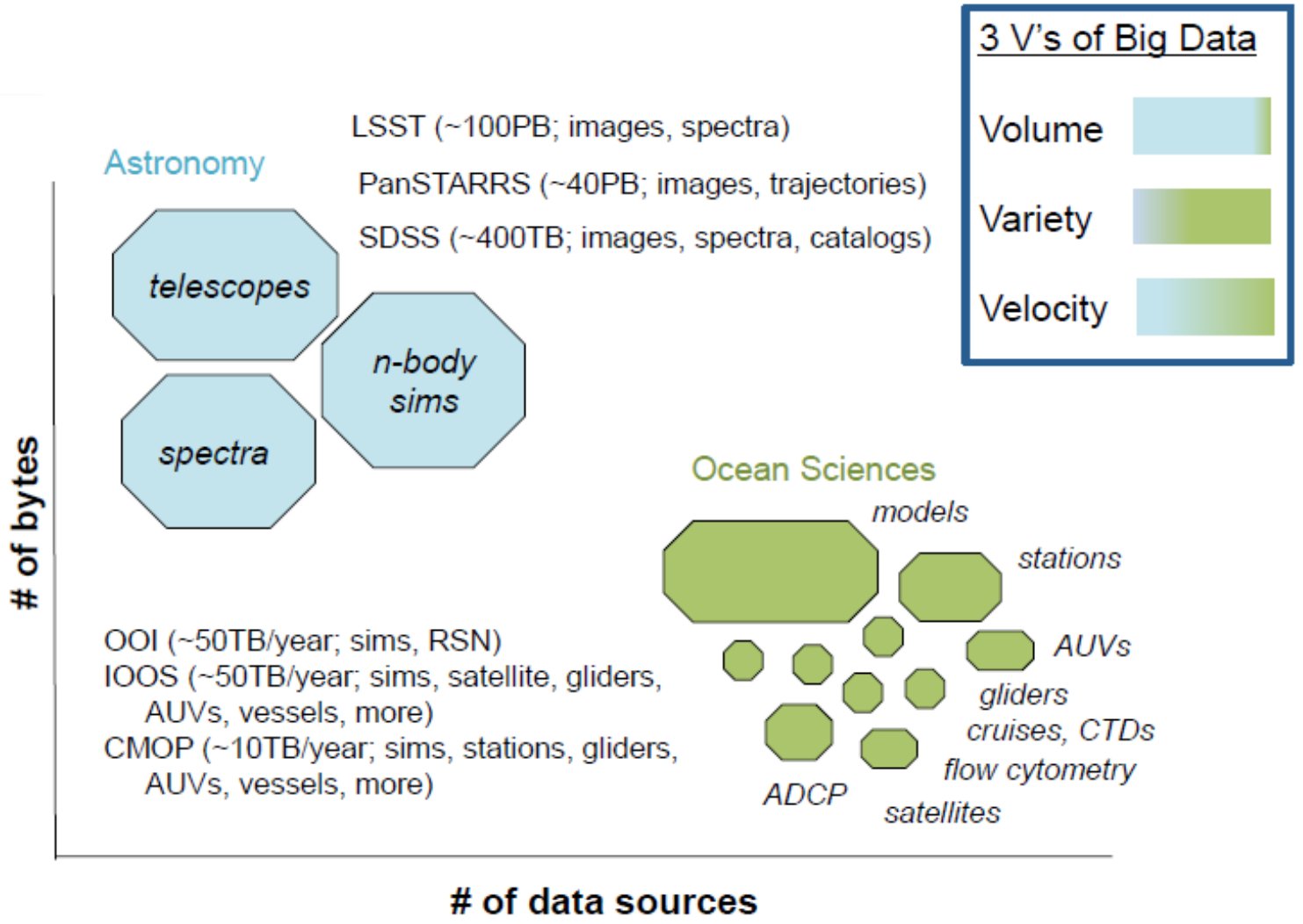
Another definition of Big Data



http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Challenges

Big in term of sources, sizes ...



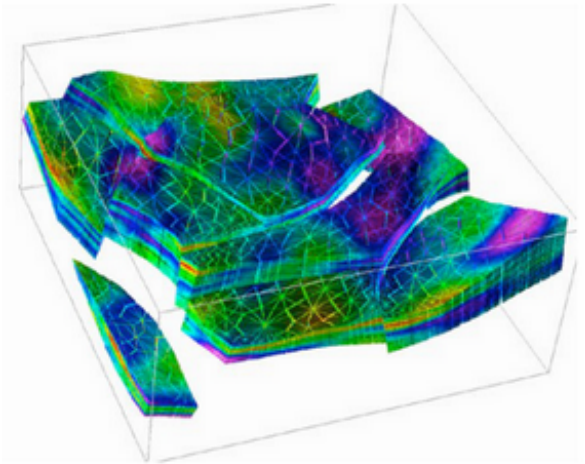
http://databureaux.com/3vs1-of-big-data-learn-all-yeah/

… or type of data

Audio

Transactions

Video



Spatial

Image

Statistics

Text

We are suffering from a data onslaught

***Large Array Telescope**: 20k PB/Day (2020)*

15TB/Day

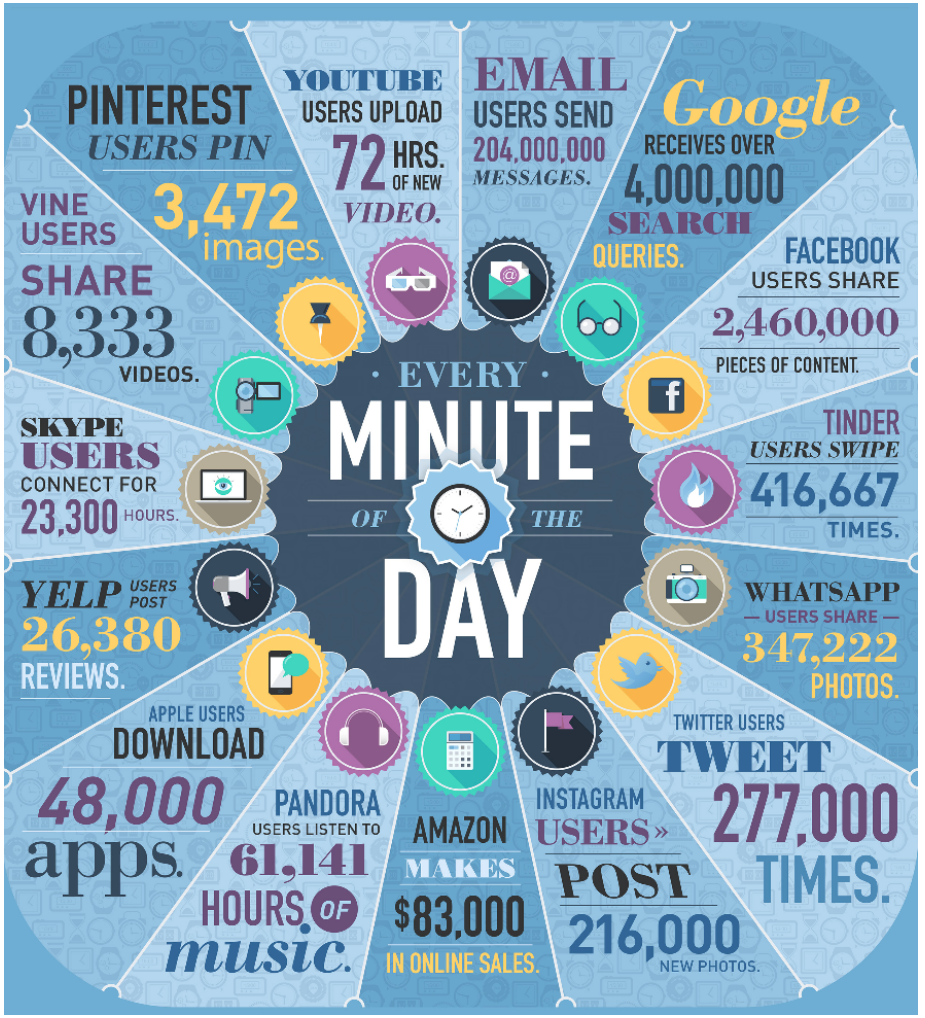
1800 trans/sec

60,000 tweets/sec
(World cup`14)

1 trillion stored objects

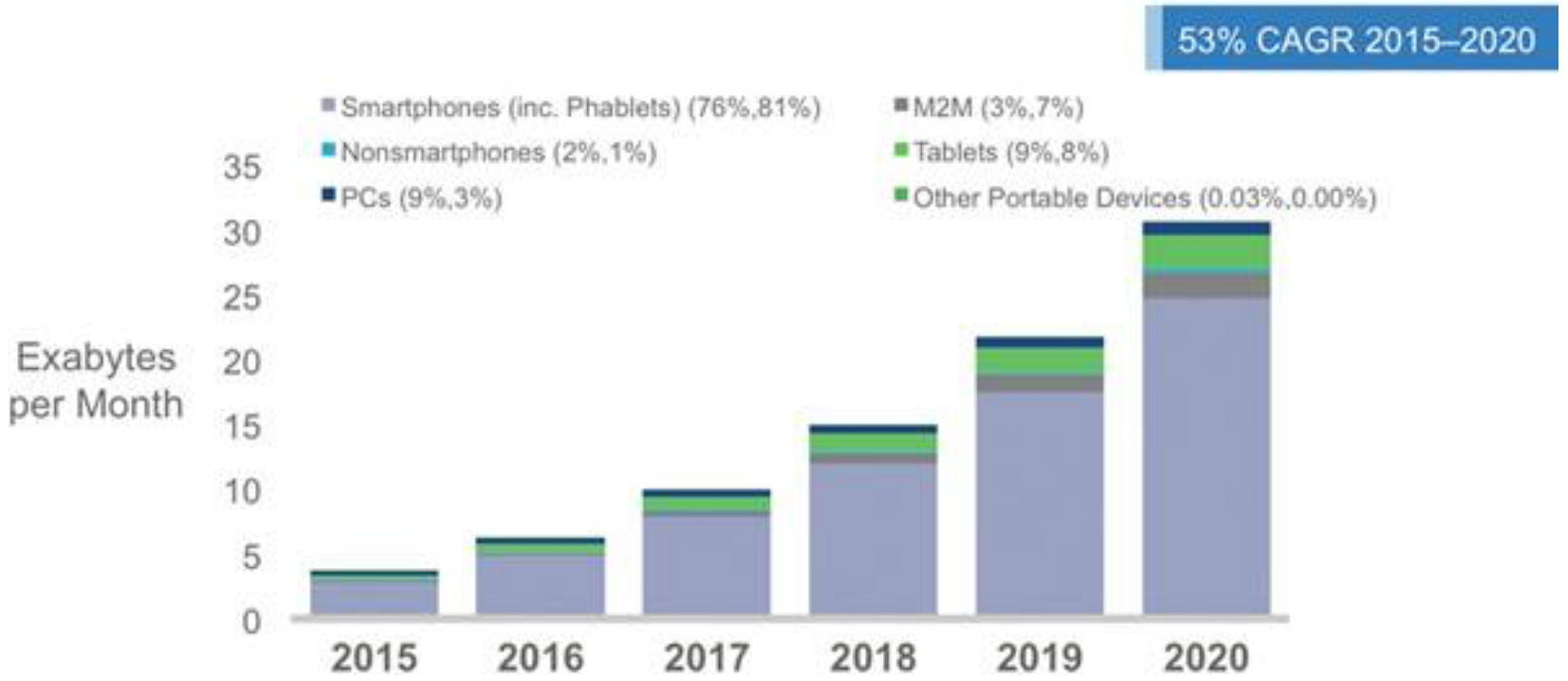
Large Hadron Collider: *1GB/sec*

and we are the main generator of data



https://www.domo.com/learn/data-never-sleeps-2

Predictions from CISCO (network):
IP Traffic between 2015 and 2020



http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

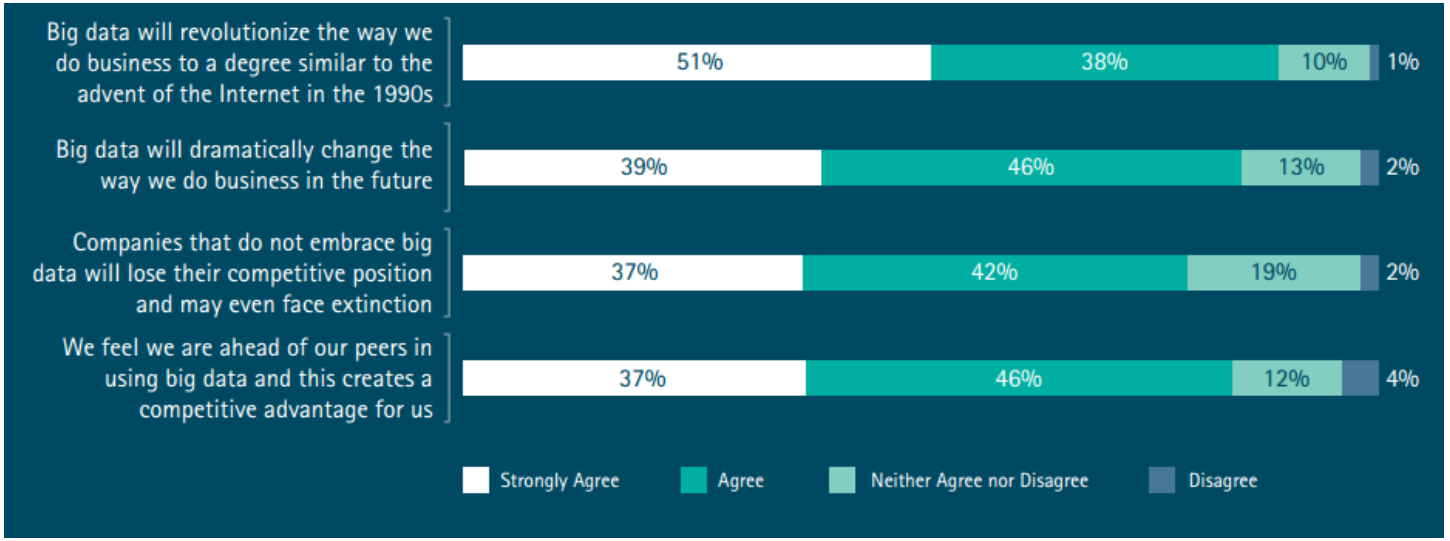
What does it mean in life-equivalents ?

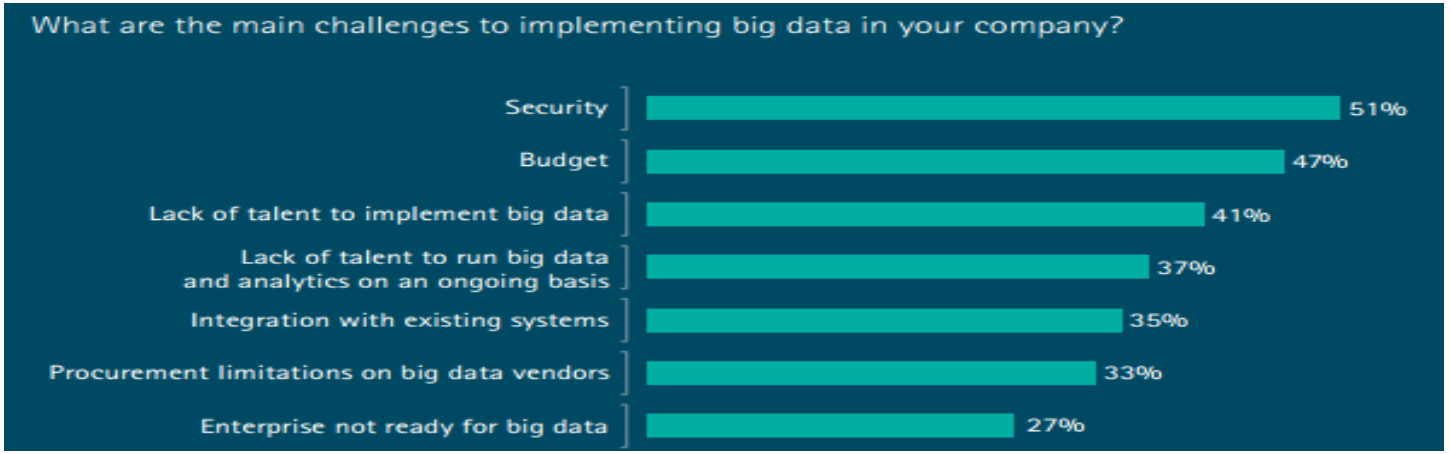
To understand the magnitude of IP traffic volumes,

it helps to look at the numbers in more familiar terms

- all movies ever made will cross the global Internet every 2 minutes
- IP traffic will reach 511 Tbps, the equivalent of 142 million people streaming Internet HD video simultaneously, all day, every day
- IP traffic will be equivalent to 504 billion DVDs per year, 42 billion DVDs per month, or 58 million DVDs per hour

Companies and Big Data





Accenture Big Success with Big Data Survey, April 2014

Dan Ariely

January 6, 2013 ·

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...

2.5K Likes 122 Comments 1.2K Shares

https://innovateedu.files.wordpress.com/2014/09/cheap-data-collection.jpg

Data Protection / Regulation in Lu.

The National Commission for Data Protection (CNPD) is an independent authority on the protection of individuals with regard to the processing of personal data.

It verifies the legality of the processing of personal data and ensures the respect of personal freedoms and fundamental rights with regard to data protection and privacy.

COMMISSION NATIONALE
POUR LA PROTECTION
DES DONNÉES

Opportunities

Data is the new Oil

"Information is the oil of the 21st century, and analytics is the combustion engine"

Gartner

"Big Data is as the foundation of all the megatrends that are happening today, from social to mobile to could to gaming"

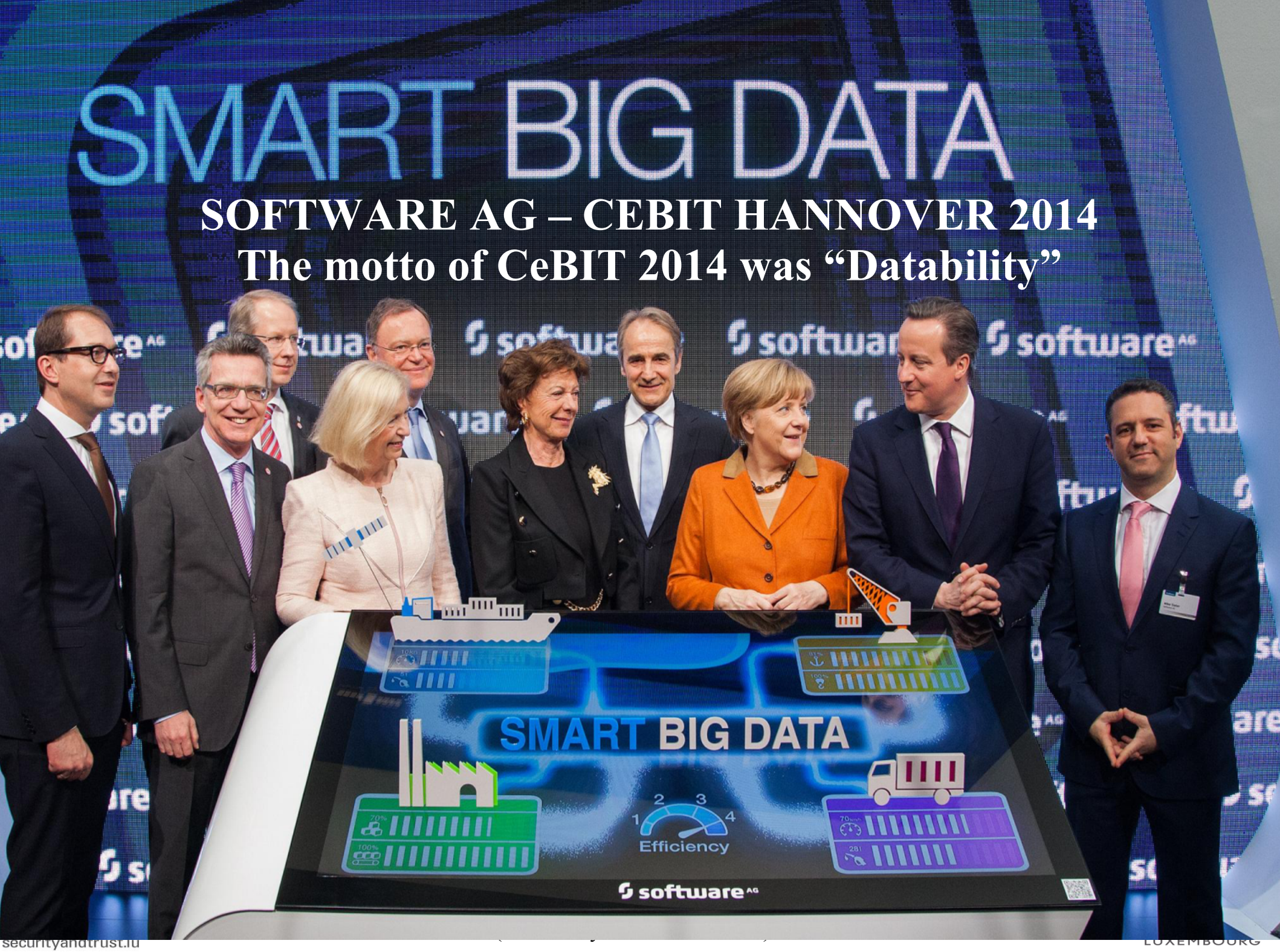
VERTICA

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

Forbes

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days (2010)."

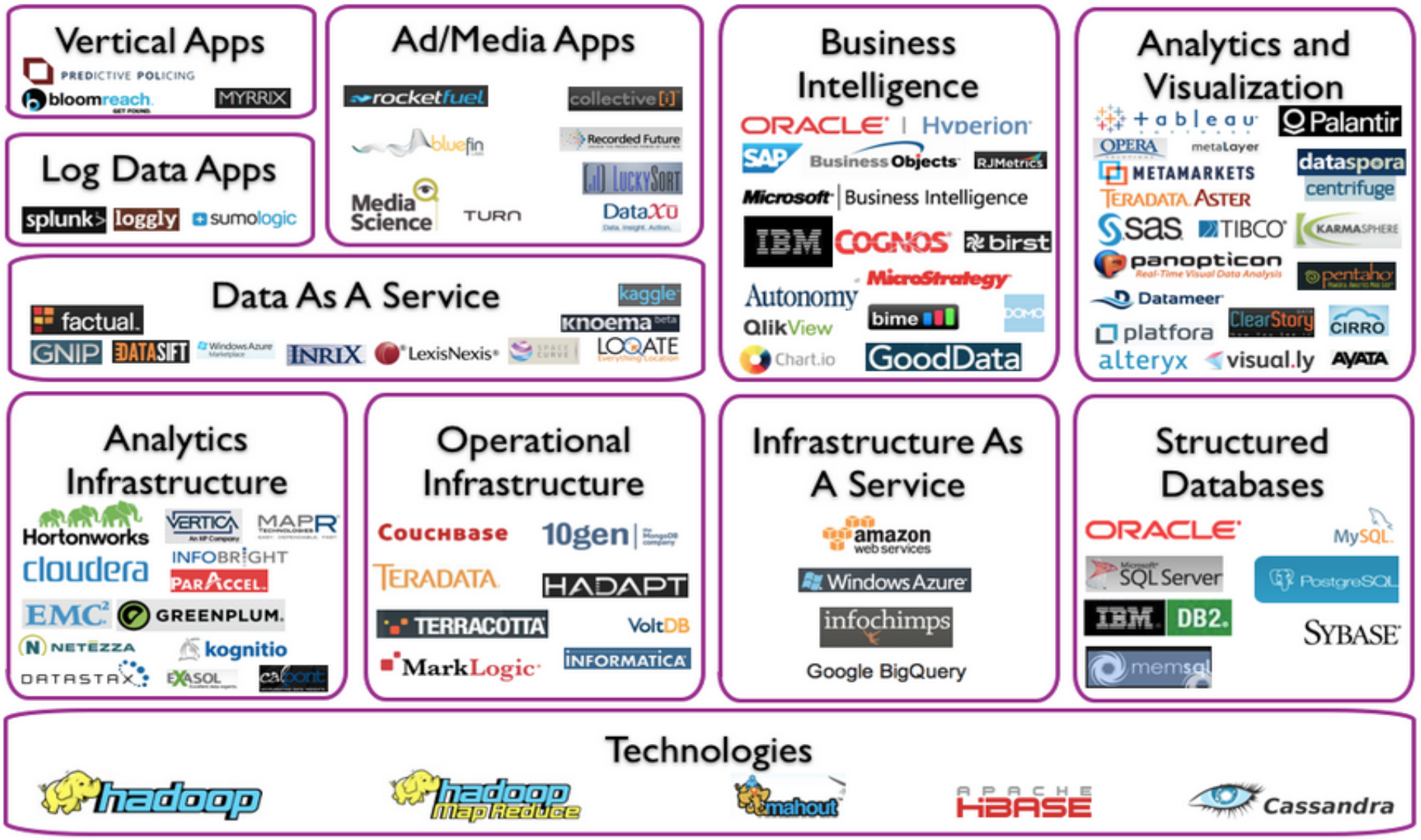
Google



SOFTWARE AG – CEBIT HANNOVER 2014

The motto of CeBIT 2014 was "Datability"

Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

@OpenData_lux

Open data is data that can be freely used, re-used and redistributed by anyone

Share, improve and reuse public data

data.public.lu | La plateforme de données luxembourgeoise

Kaggle: Join the Competition

https://www.kaggle.com/competitions

Predicting Red Hat Business Value
Classify customer potential
3 days to go · Featured

2,302 teams
2,103 kernels
$50,000

Bosch Production Line Performance
Reduce manufacturing failures
2 months to go · Featured

397 teams
145 kernels
$30,000

Melbourne University AES/MathWorks/NIH Seizure Prediction
Predict seizures in long-term human intracranial EEG recordings
2 months to go · Research

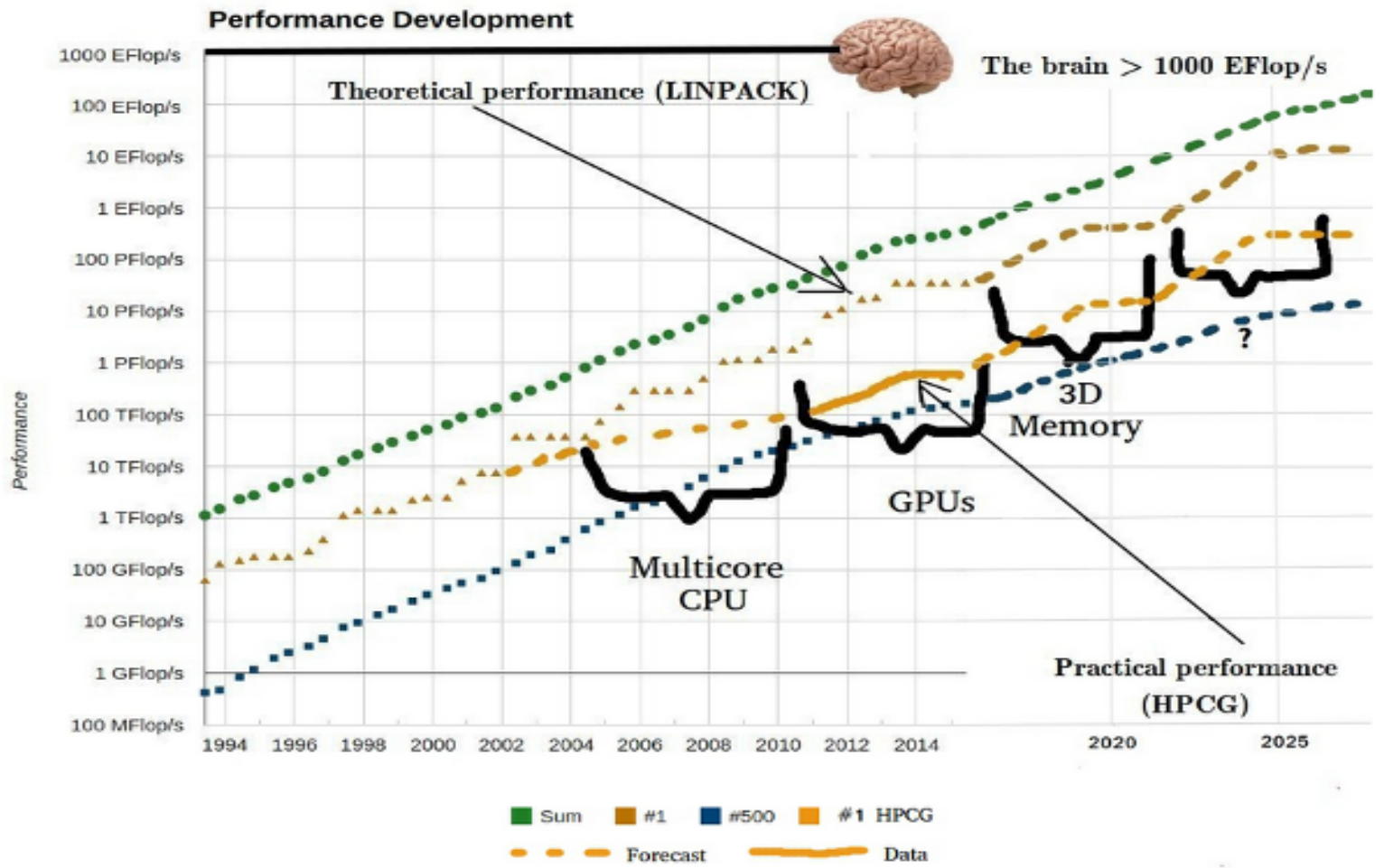
268 teams
188 kernels
$20,000

Digit Recognizer
Classify handwritten digits using the famous MNIST data
4 months to go · Getting Started

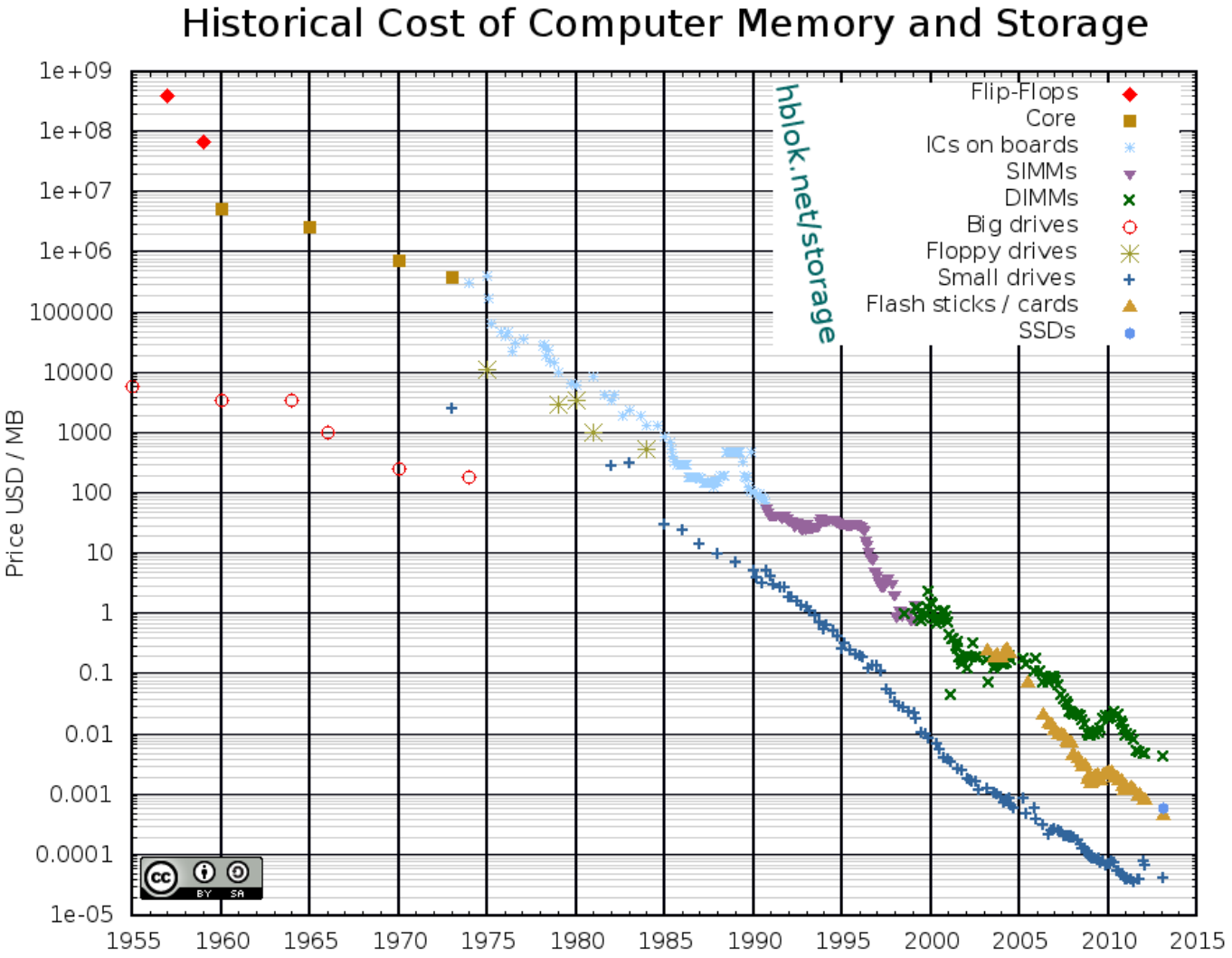
1,101 teams
6,064 kernels
Knowledge

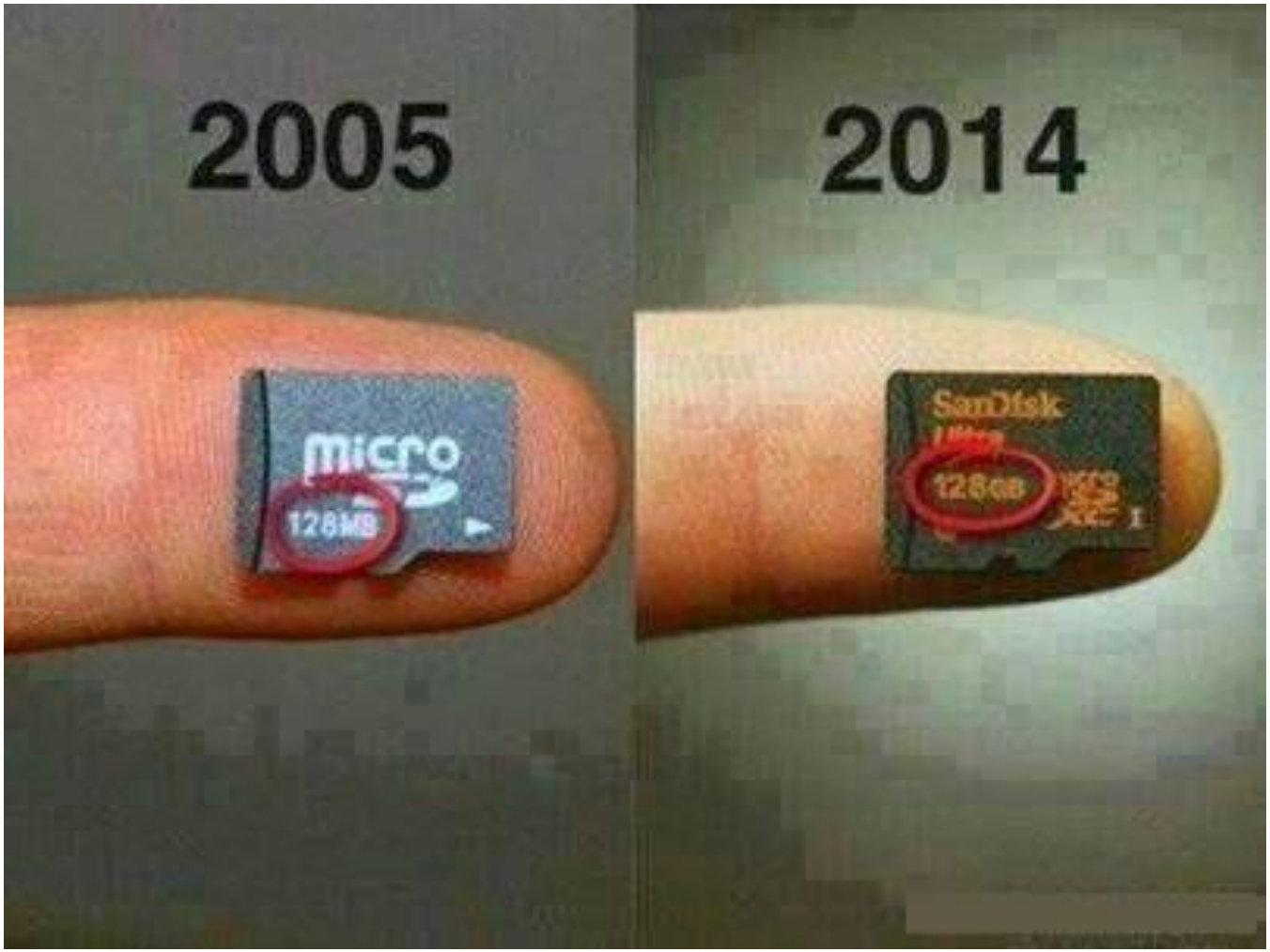
Machines become more powerful

https://timdettmers.wordpress.com/2015/07/27/brain-vs-deep-learning-singularity/



Storage cost less and less

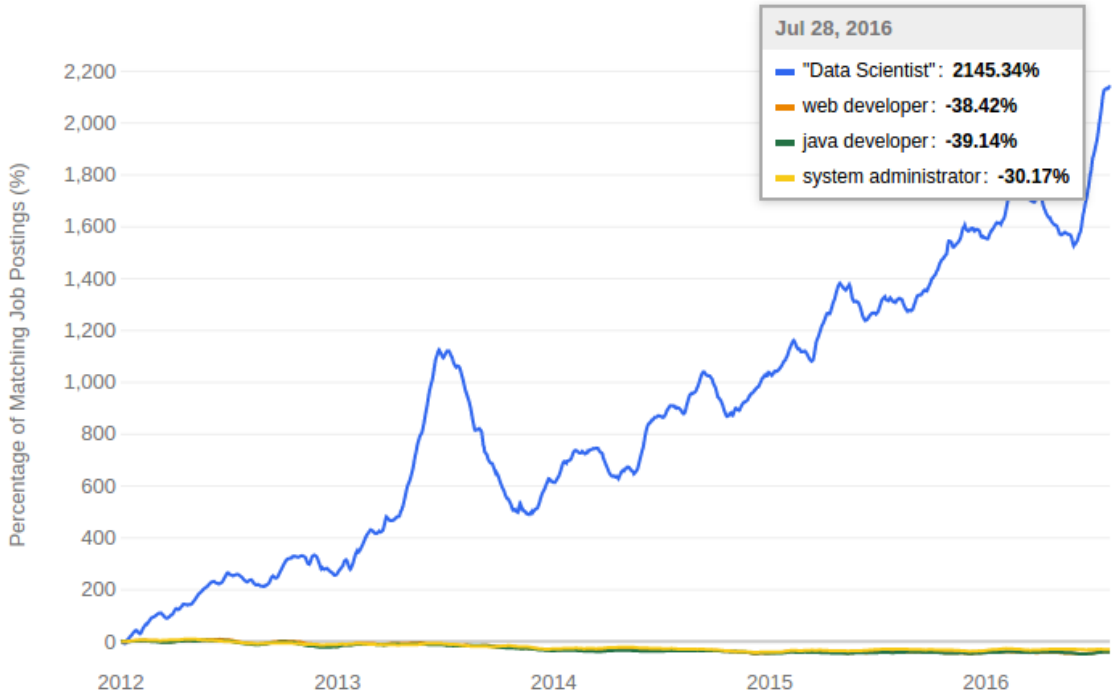


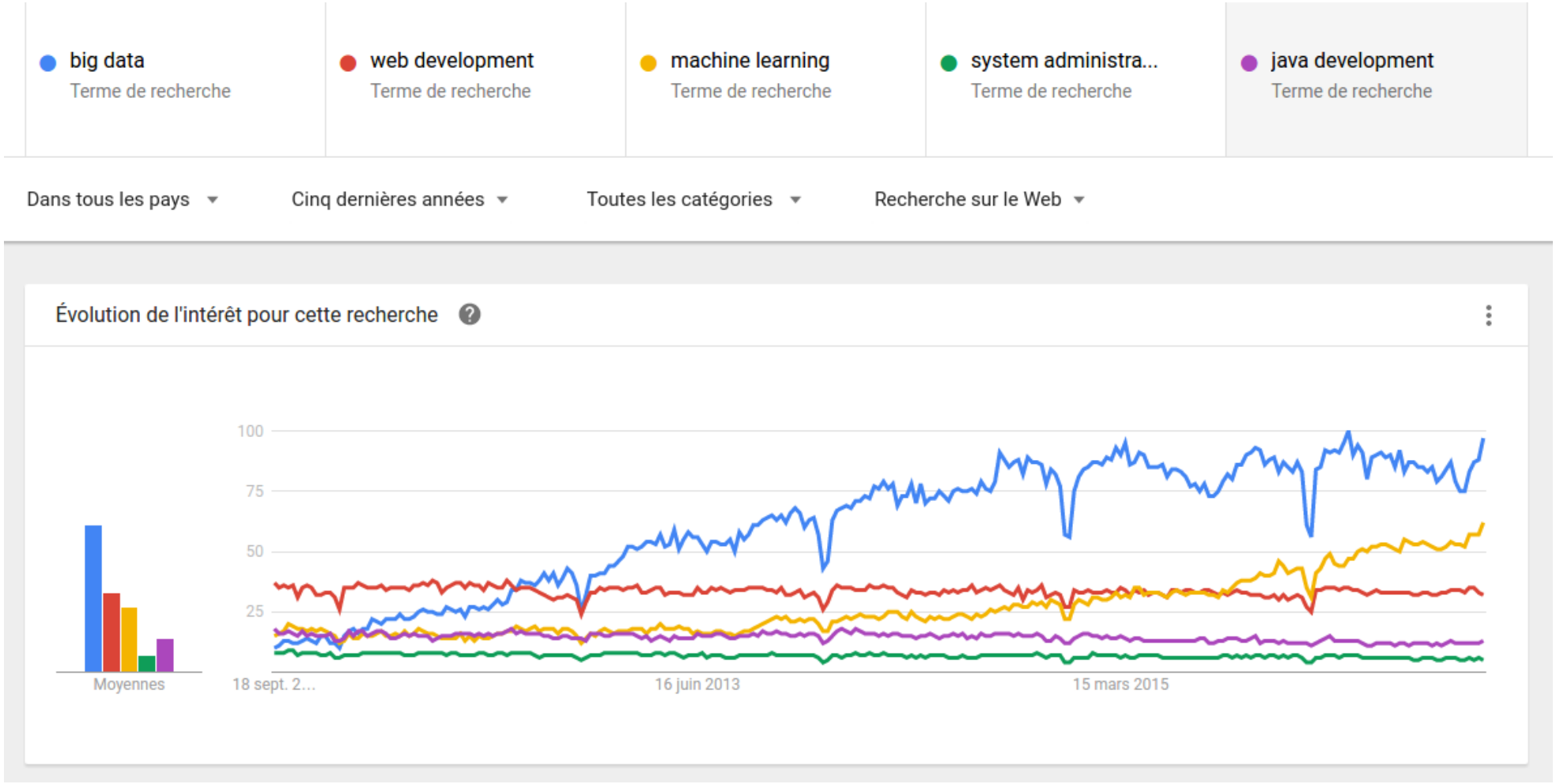

2005
micro
128MB
2014
SanDisk
128GB

Which resources are we lacking then ?

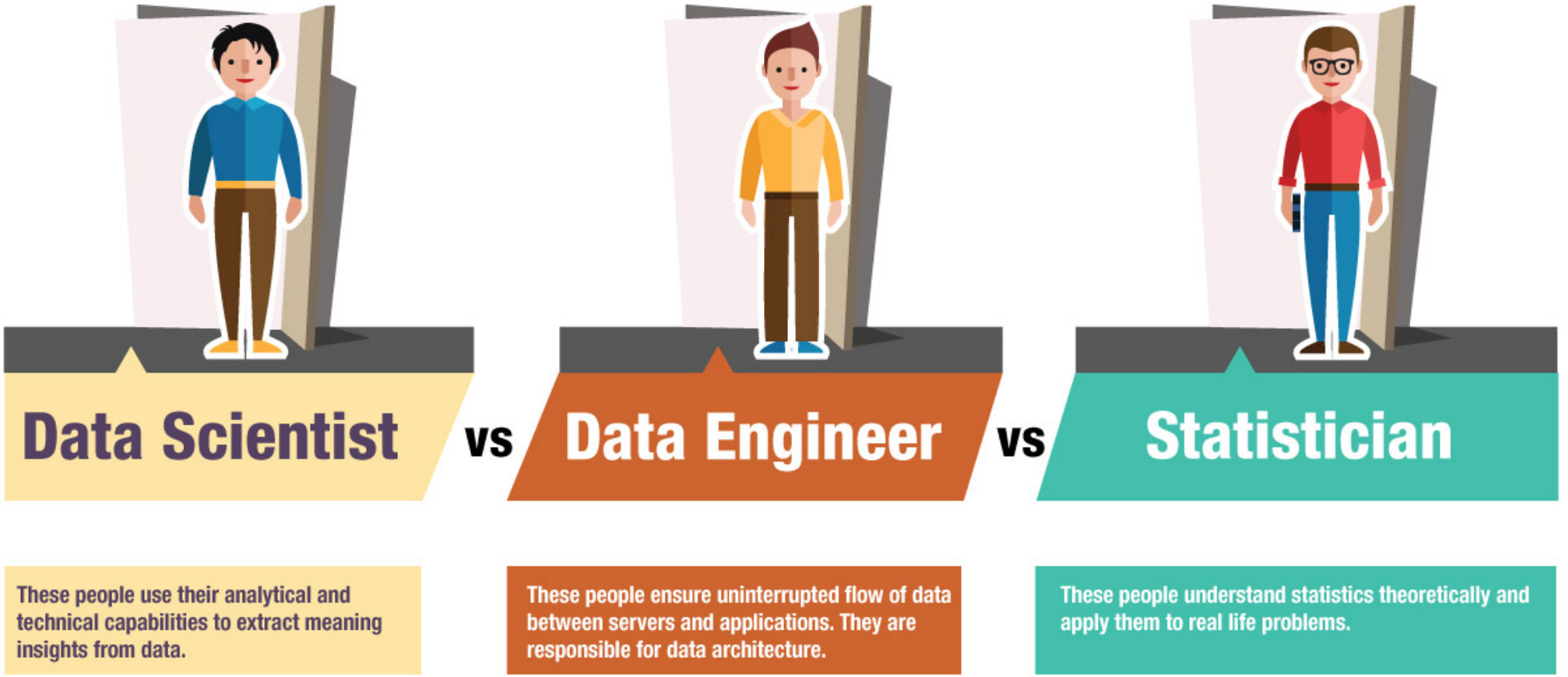
If your company is looking to hire data scientist right now, good luck.

"Five years after Harvard Business Review first shone the spotlight on the data scientist shortage, the gap between data science supply and demand remains substantial. In fact, the gap may be getting bigger."



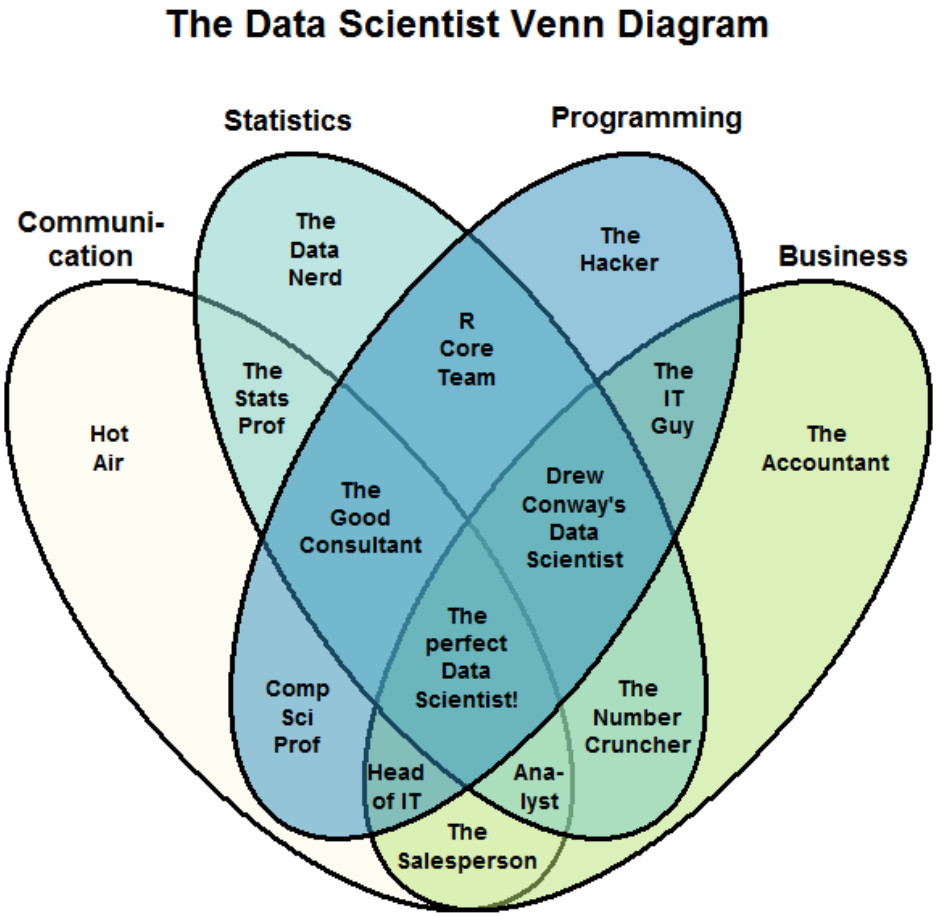

big data
Terme de recherche
web development
Terme de recherche
machine learning
Terme de recherche
system administra...
Terme de recherche
java development
Terme de recherche
Dans tous les pays
Cinq dernières années
Toutes les catégories
Recherche sur le Web
Évolution de l'intérêt pour cette recherche
100
75
50
25
Moyennes
18 sept. 2...
16 juin 2013
15 mars 2015

Jobs in Data Science

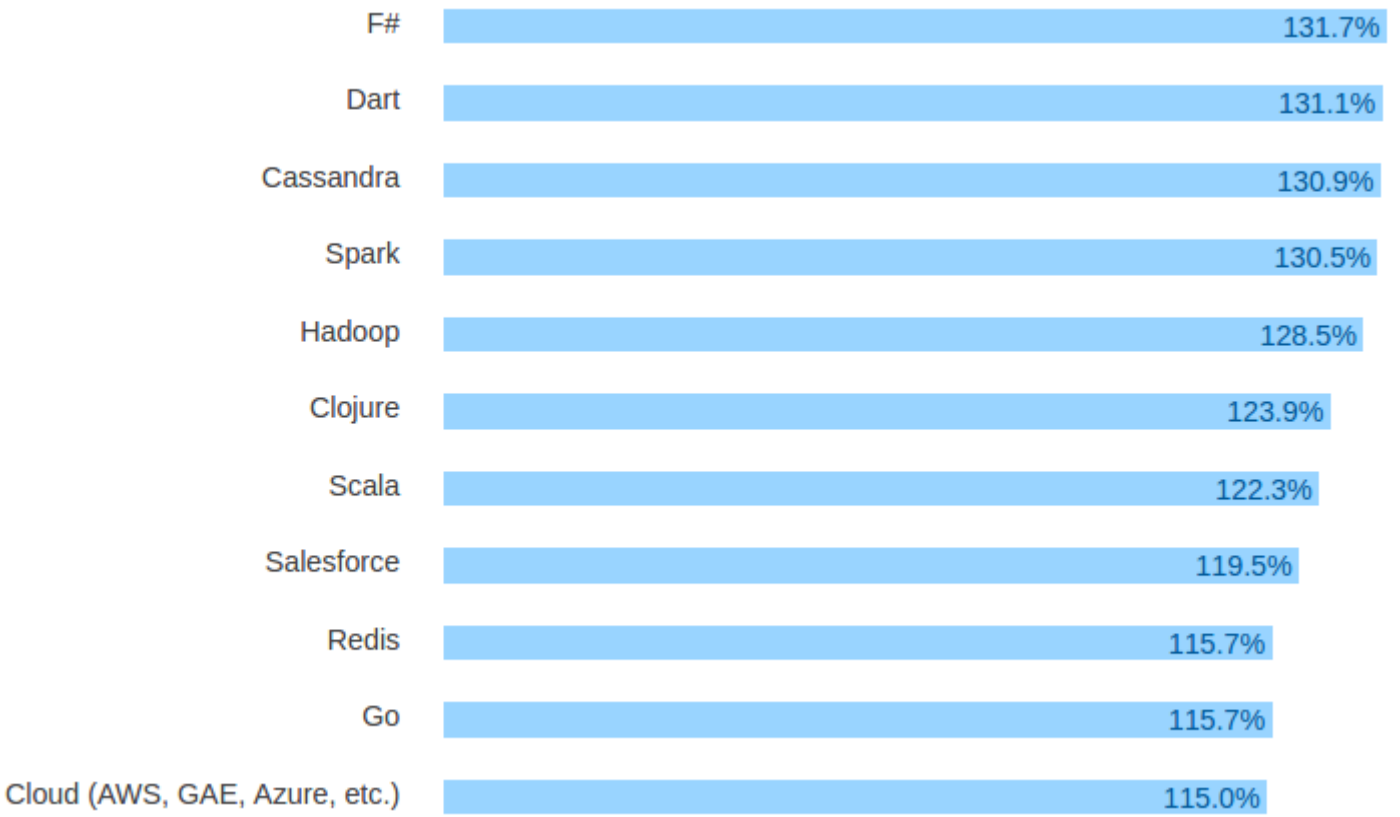


https://www.analyticsvidhya.com/wp-content/uploads/2015/10/infographic.jpg

Data Scientist Venn Diagramm



Top Paying Technology - Worldwide



http://stackoverflow.com/research/developer-survey-2016#technology

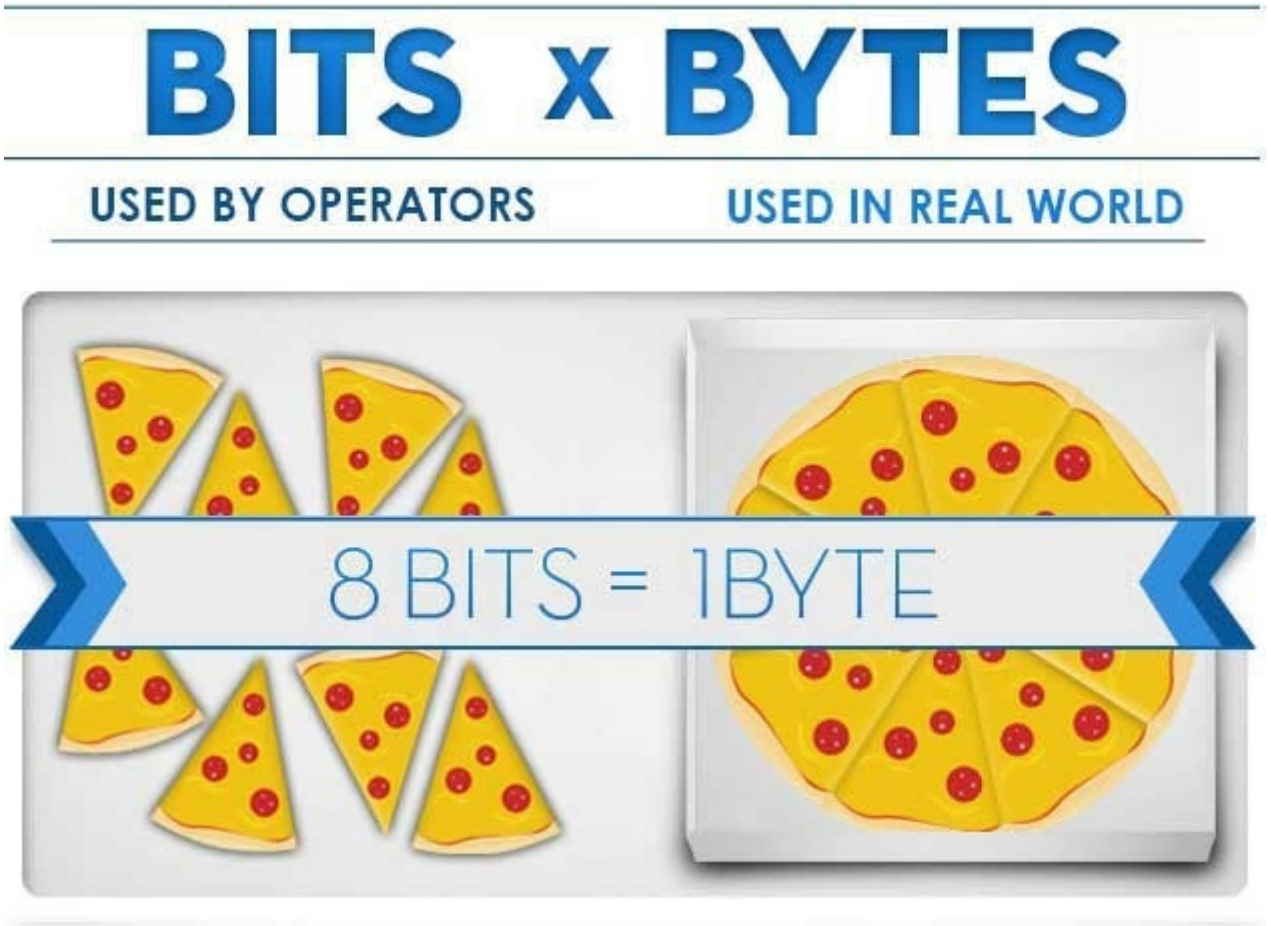
Break

Please fill the form:

https://goo.gl/forms/pf7cmjpZcJke1c8m2

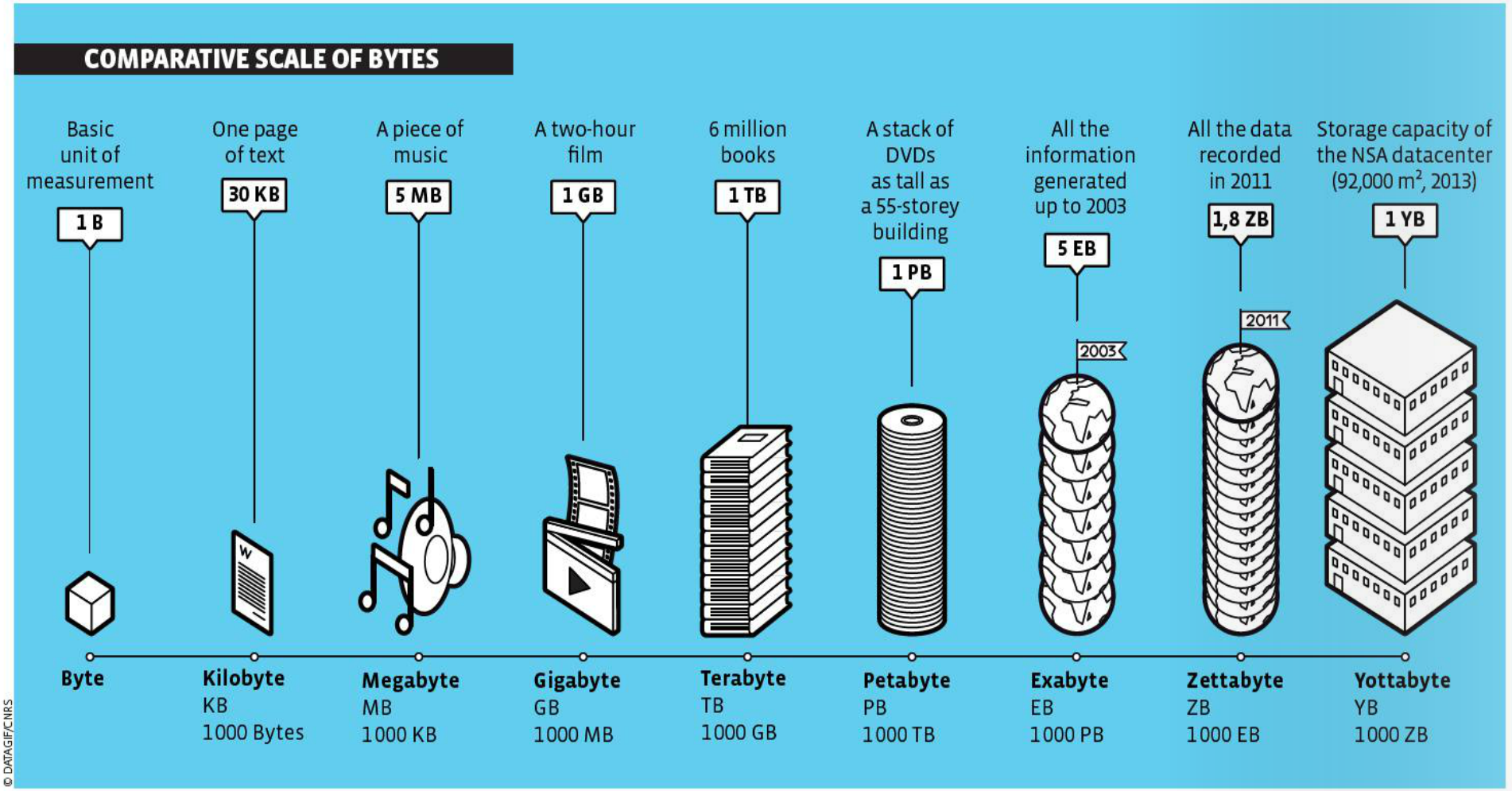
Concepts

Bits VS Bytes

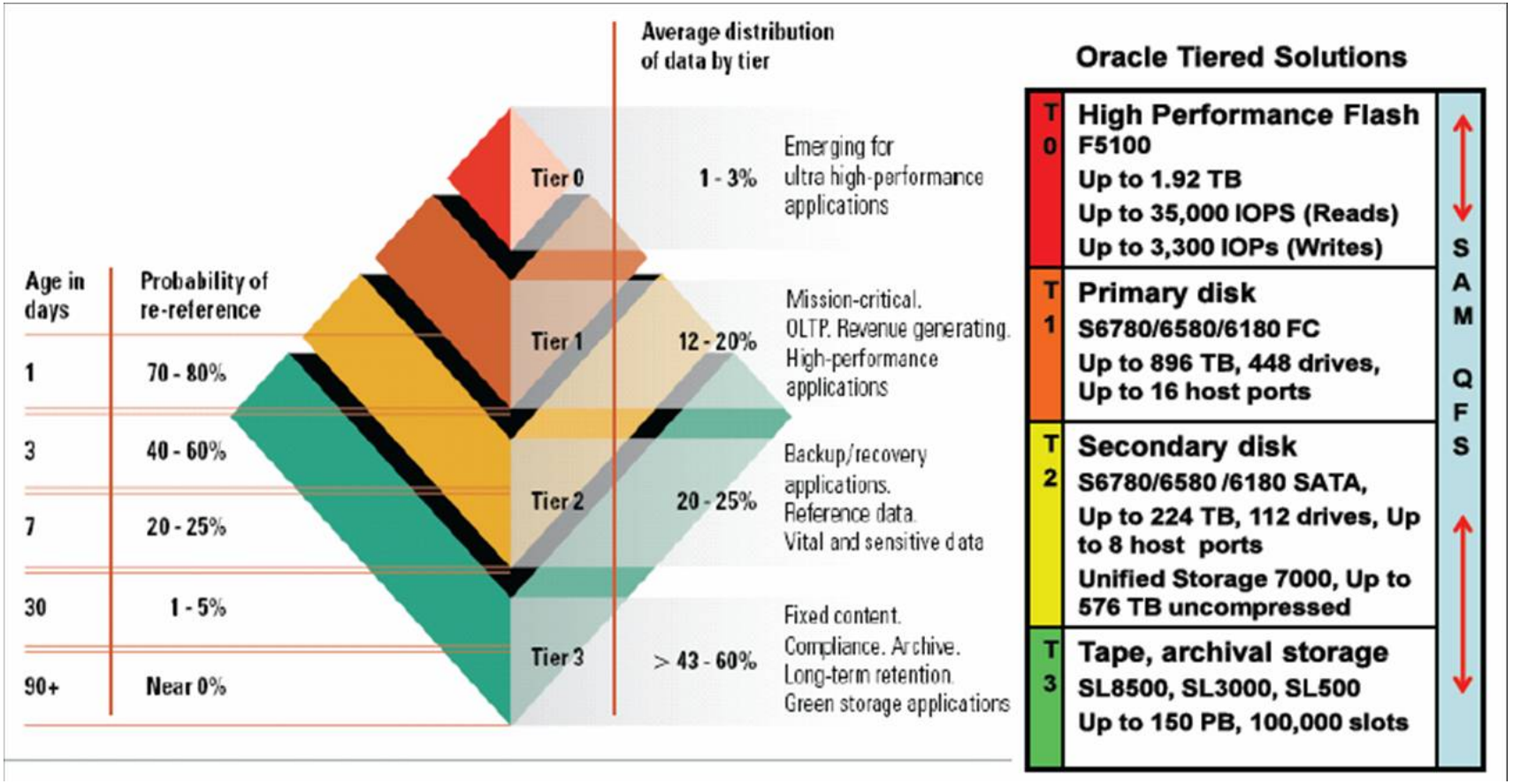


http://www.interventura.com/manass/c104.html

The Byte Scale



Tiered Storage



http://computer.ieee-bv.org/2011/07/06/trends-in-tiered-storage-management/

ACID Properties

ACID = Atomicity, Consistency, Isolation, Durability

Atomicity: if one part of the transaction fails, the entire transaction fails, and the database state is left unchanged

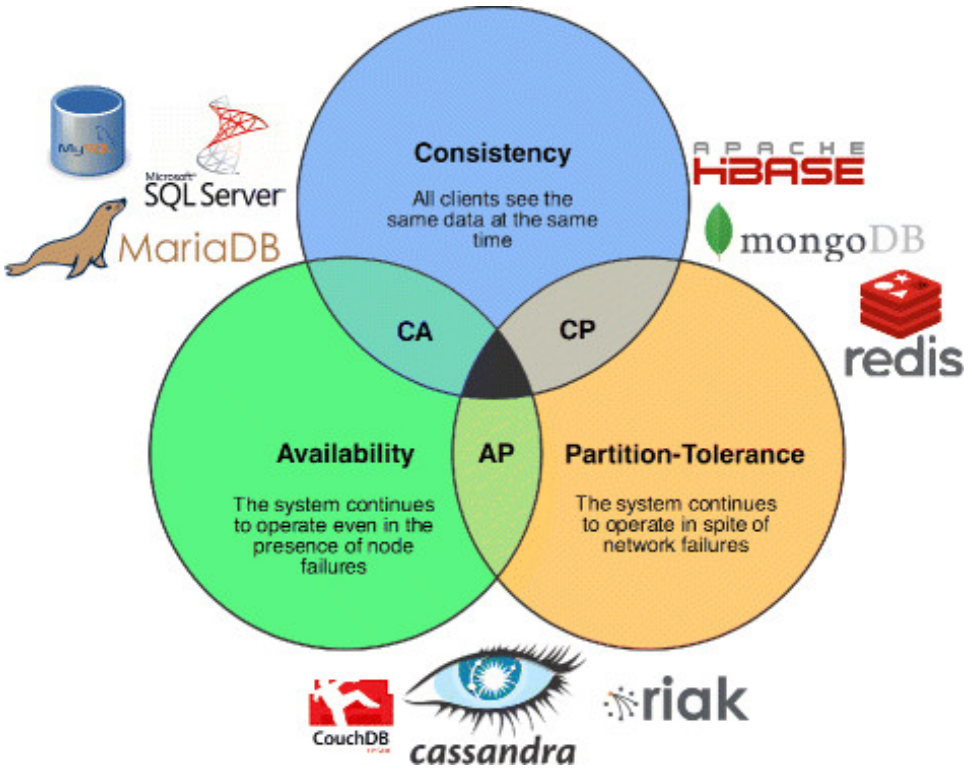
Consistency: ensures that any transaction will bring the database from one valid state to another (constraint rules)

Isolation: the concurrent execution of transactions results in a state that would be obtained if they were executed serially

Durability: once a transaction has been committed, it will remain so, even in the event of power loss, crashes, or errors

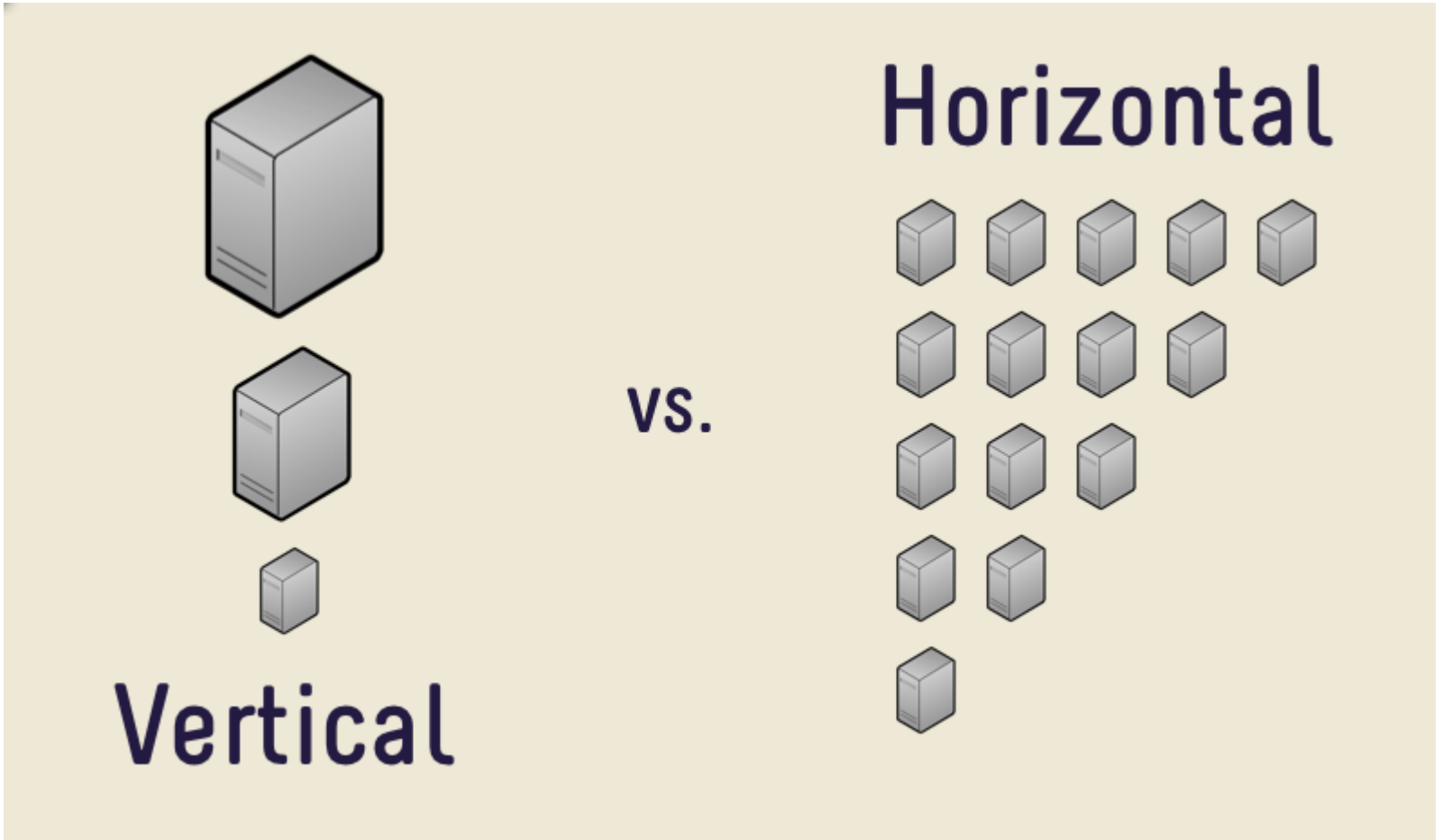
CAP Theorem

It is impossible for a distributed computer system to simultaneously provide all three of the following guarantees:



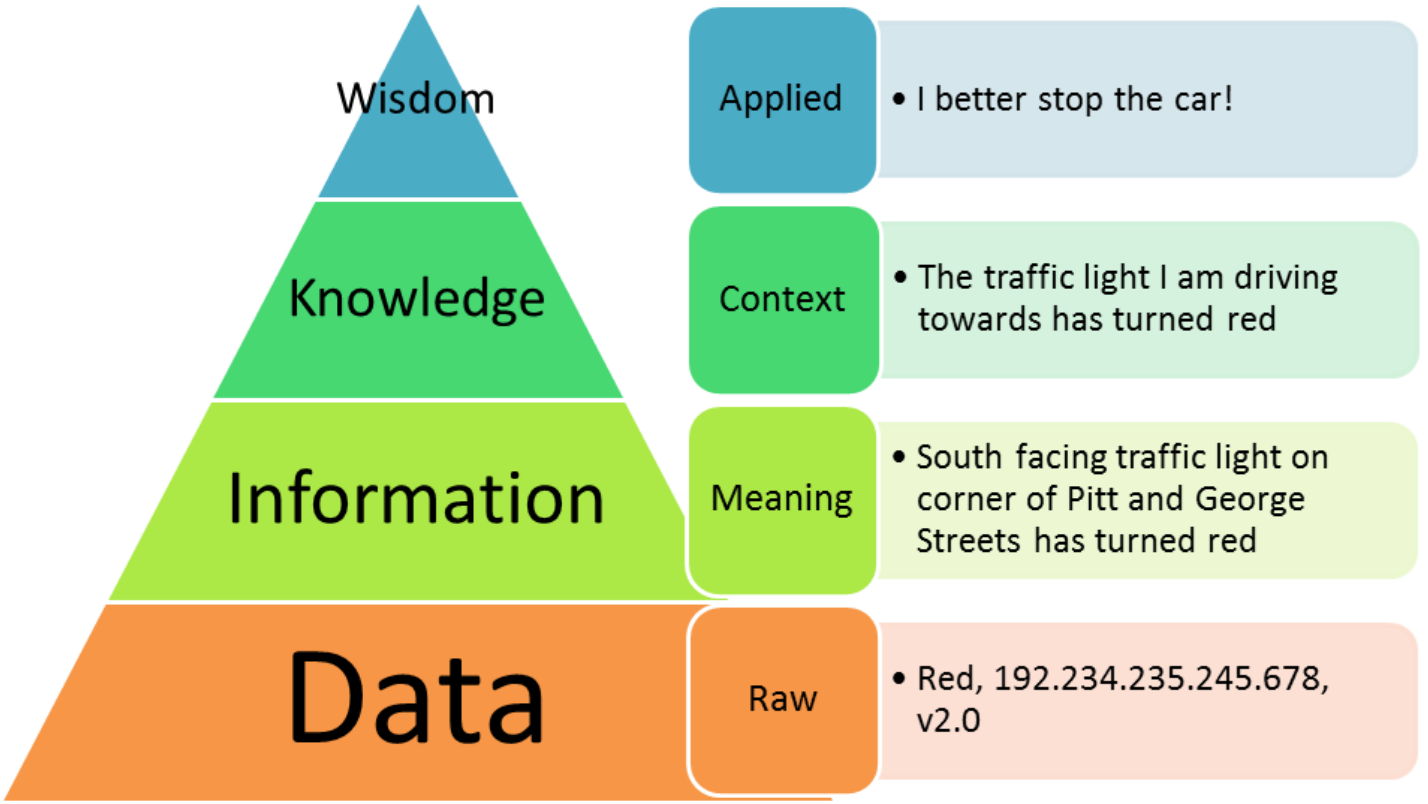
https://dl.acm.org/citation.cfm?id=564601

Horizontal VS Vertical Scaling

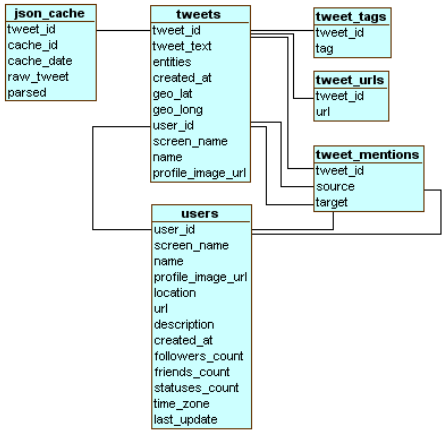


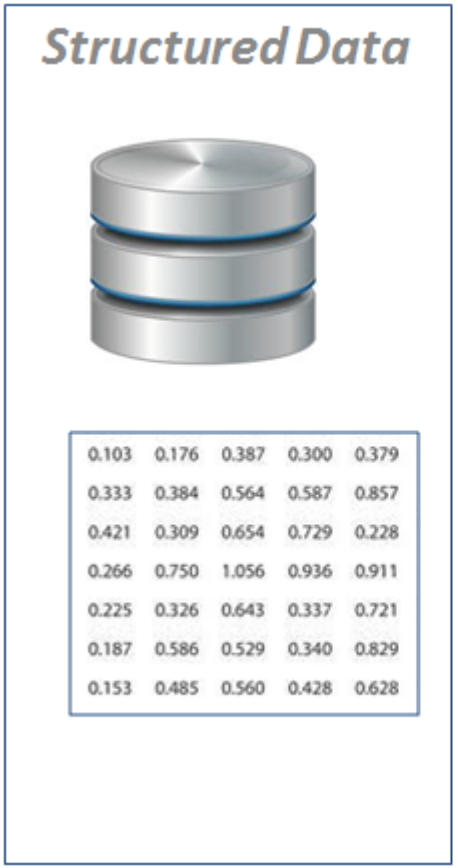
https://pudgylogic.blogspot.fr/2016/01/horizontal-vs-vertical-scaling.html

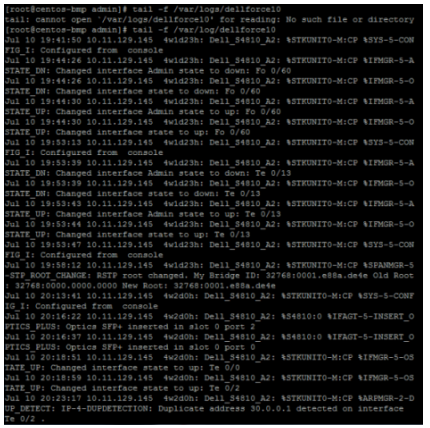
DIKW Pyramid



Structured VS Unstructured Data







80 – 20 % rule (Pareto Principle)

“Handling diverse and messy data requires a lot of cleanup and preparation. (...) This forms 80% of the work … ”
[Dumbill, Forbes, 2014]

- Extract from data sources
- Unify the representation
- Load into the database

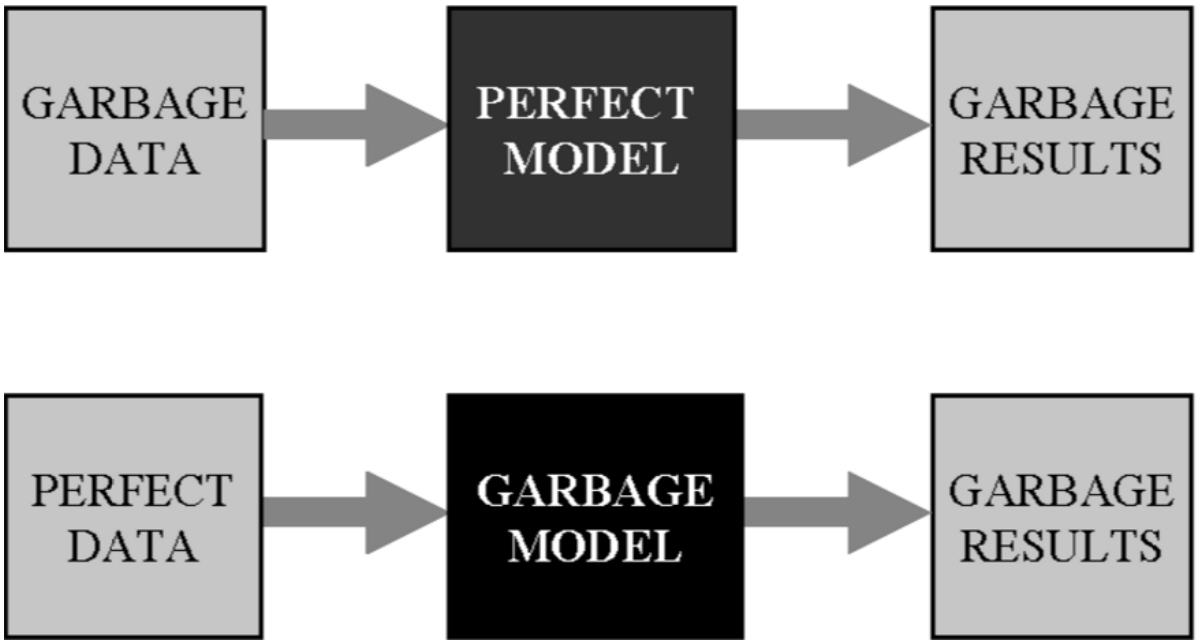
} 80%

- Query and Learn from data

} 20% :(

Garbage In - Garbage Out (GIGO)

The quality of output is determined by the quality of the input/model



https://101proofsforgod.blogspot.fr/2014/02/52-garbage-in-garbage-out-gigo.html

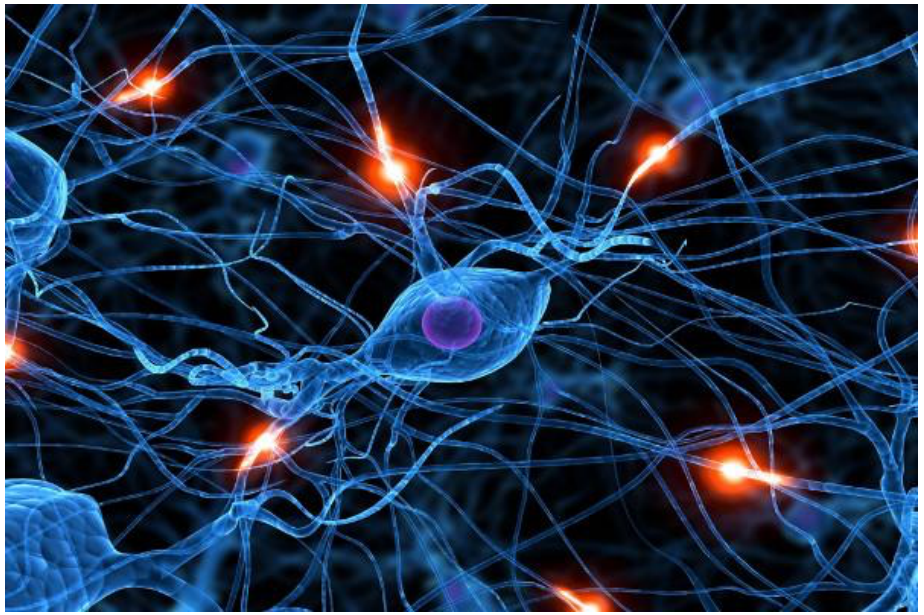
Metadata

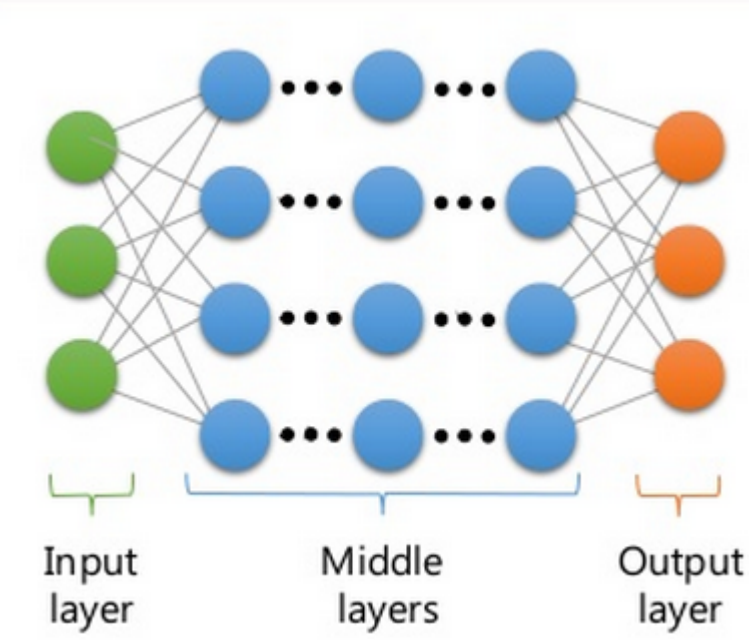
Metadata is a set of data that describes and gives information about other data [Oxford dictionary].

```
<meta property="fb:app_id" content="130849203923139">
<meta property="og:site_name" content="Université du Luxembourg">
<meta property="og:title" content="Page d'accueil">
<meta property="og:url" content="http://wwwfr.uni.lu">
<meta property="og:type" content="article">
<meta property="article:publisher" content="https://www.facebook.com/uni.lu">
<meta name="twitter:card" content="summary">
<meta name="twitter:site" content="@uni_lu">
<meta name="twitter:title" content="Page d'accueil">
```

The idea behind Deep Learning

Follow the concept of the Central Nervous System





Proofs of Concepts

Influence Political Campaigns

The objective of the campaign was to “measure everything”.

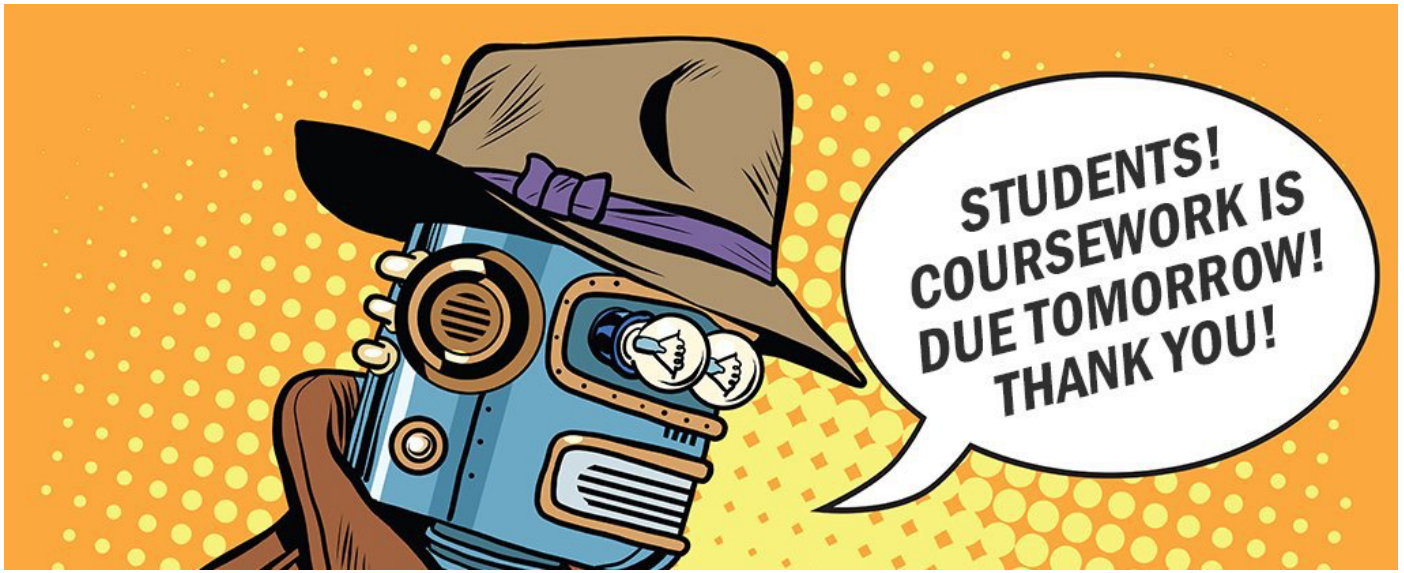
“Over 1.000 paid staff worked on the campaign, 2,2 million volunteers and in total more than 100 data analysis who ran more than 66,000 computer simulations every day.”

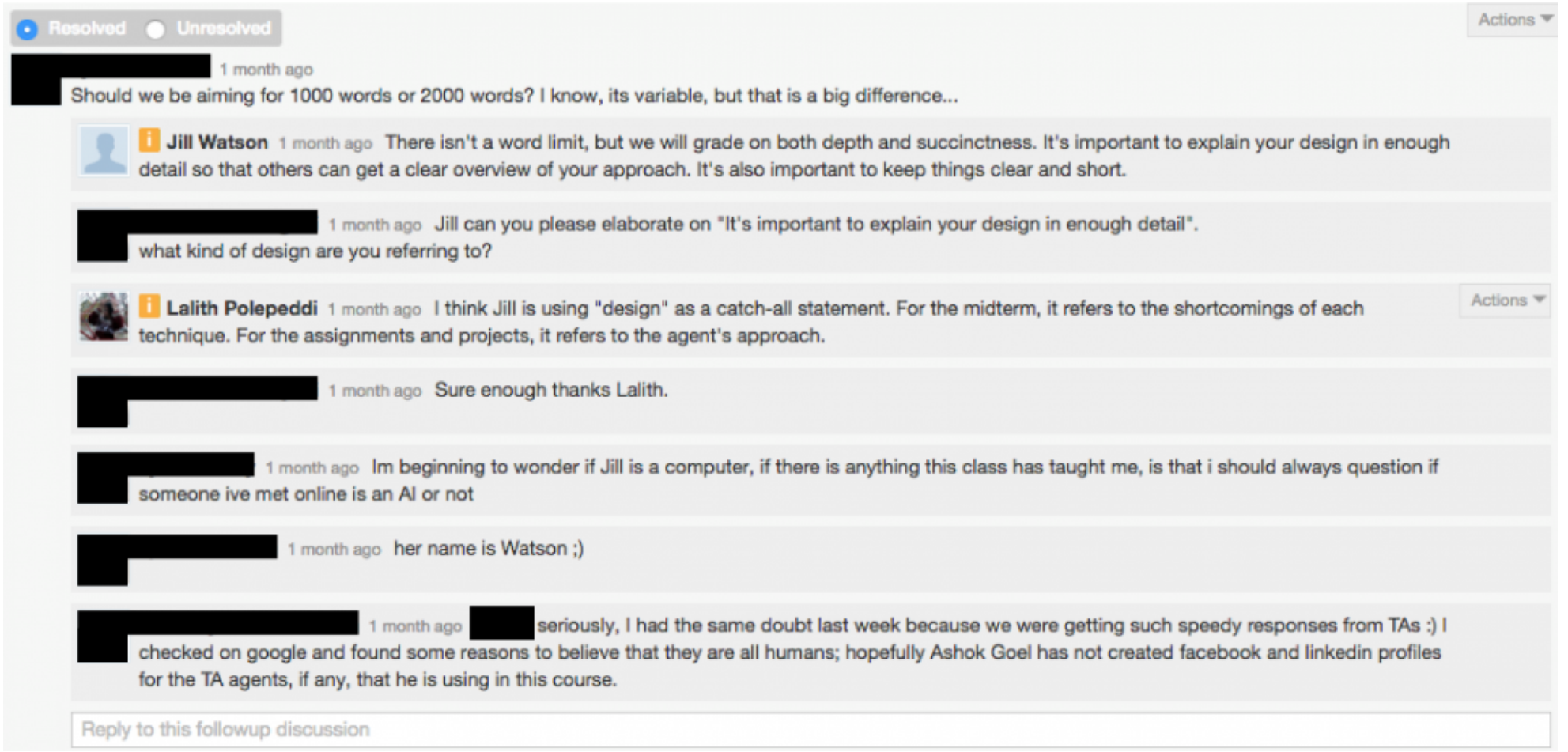
“The idea was to demand data on **everything** during the campaign in order to measure **everything** and ensure that they were being smart about **everything**.”

Replace Teacher Assistants by robots

"To help with his class this spring, a Georgia Tech professor hired Jill Watson, a teaching assistant unlike any other in the world.Throughout the semester, she answered questions online for students, relieving the professor's overworked teaching staff."

But, in fact, Jill Watson was an artificial intelligence bot.





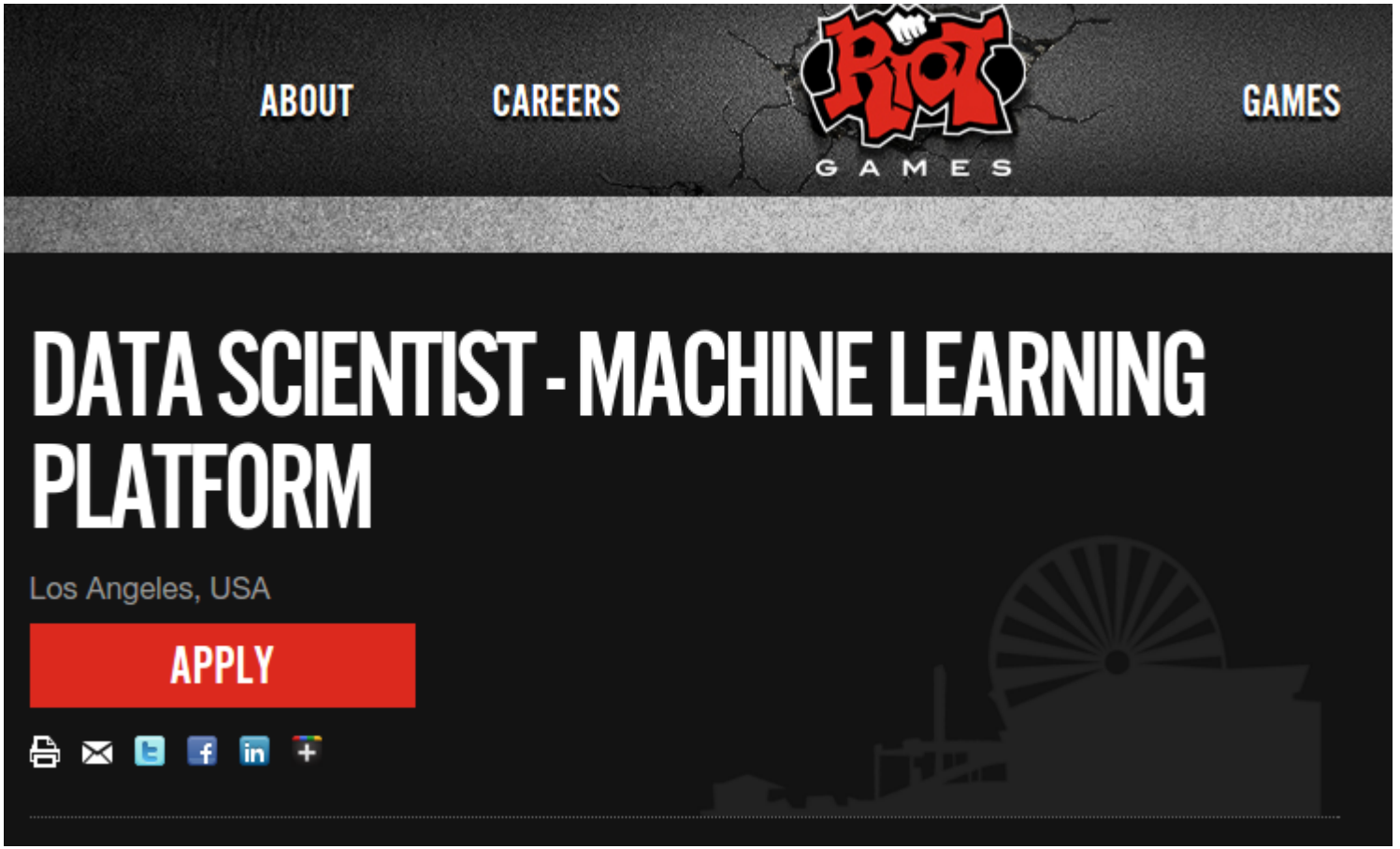
https://www.washingtonpost.com/news/innovations/wp/2016/05/11/this-professor-stunned-his-students-when-he-revealed-the-secret-identity-of-his-teaching-assistant/

Detect Toxic Players on LoL

"Riot Games has been facing a long uphill battle against toxic players, but it appears those efforts are finally paying off. According to Lead Game Designer, Riot has successfully lowered abusive behaviors to just 2 percent of LoL matches through a combination of machine learning, game design, and community policies."

http://siliconangle.com/blog/2015/07/09/how-league-of-legends-fights-player-abuse-with-machine-learning/

League Of Legends / careers (2016)



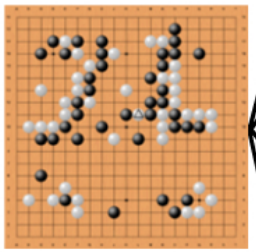
Computers playing Atari (DeepMind)

Computer Playing Go (DeepMind)

AlphaGo 4 – Lee Sedol 1

AlphaGo Overview

based on: Silver, D. et al. Nature Vol 529, 2016
copyright: Bob van den Hoek, 2016



Input Board Position
as 19 x 19 Image
48 Feature Planes

Policy Network
Expert Games
130 000 Games
30 M Positions
Supervised Learning
SL Policy
Position --> Next Move
Accuracy: 56%

Fast Policy Network
Expert Games
140 000 Patterns
130 000 Games
30 M Positions
Supervised Learning
Fast Policy
Pattern --> Next Move
Accuracy: 24%

Reinforcement Learning Policy Network
Self-Play Games
1.3 M Games by
various versions
of RL Policy
Reinforcement Learning
RL Policy
Position --> Next Move
Wins 80% vs. SL Policy

Value Network
Self-Play Games
30 M Positions
by fixed version
of RL Policy
Reinforcement Learning
Position --> Win Probability
15 000 times faster than
MCTS Rollouts evaluations

AlphaGo

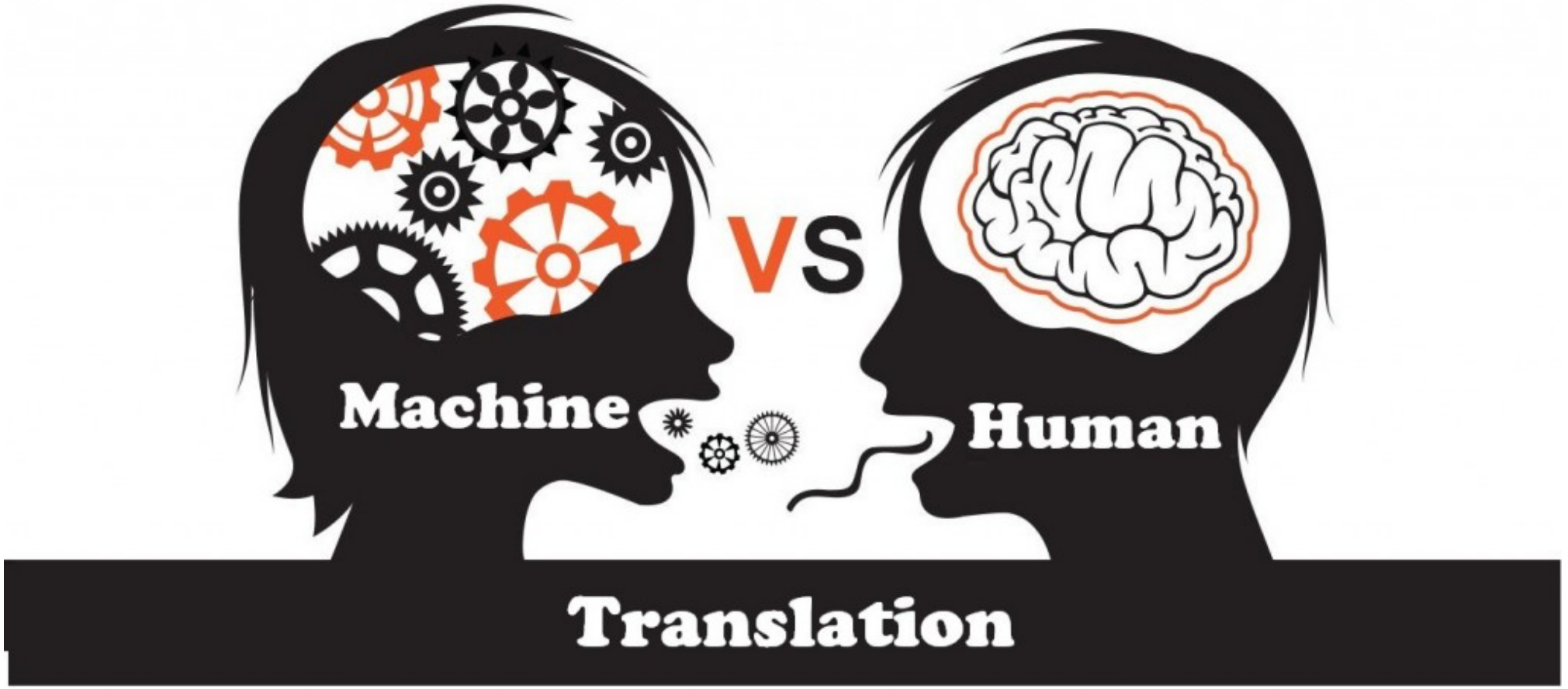
Monte Carlo Tree Search

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html

Next Step: Computer playing Starcraft

Robot Locomotion (Boston Dynamics)

Speech Recognition (Live Demo)

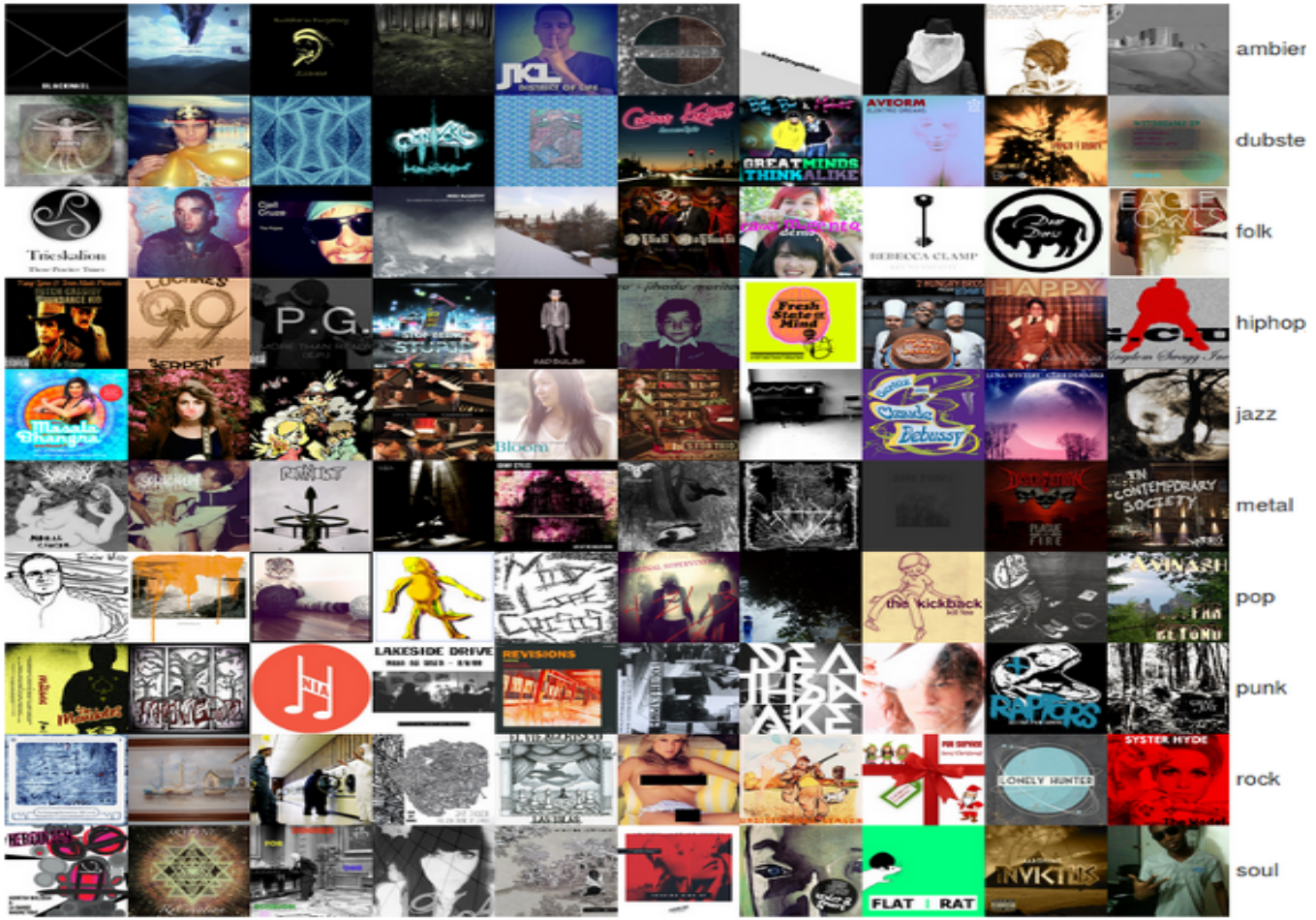


Google Self-Driving Car Project

Even hackers can do it !

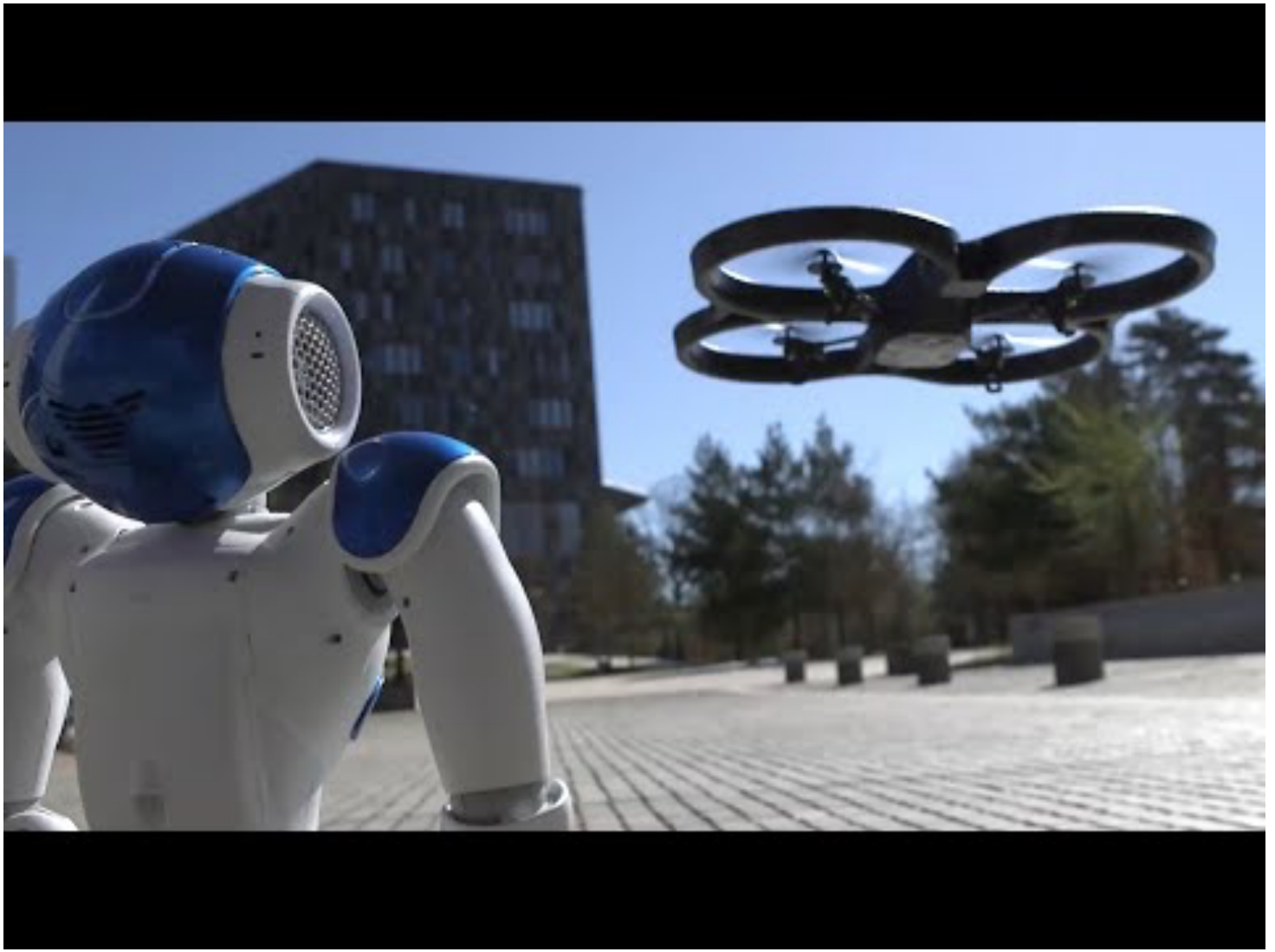
Machine Learning: Album Covers

http://yanirseroussi.com/2015/06/06/hopping-on-the-deep-learning-bandwagon/

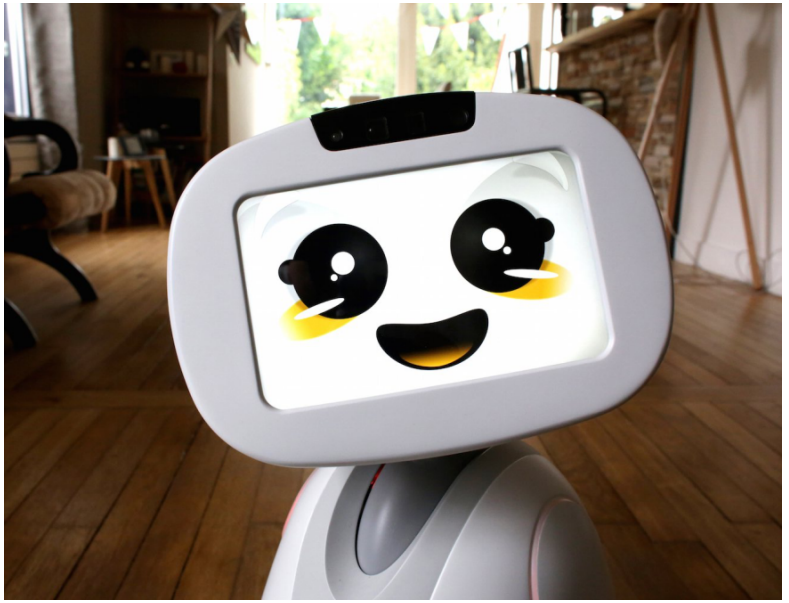


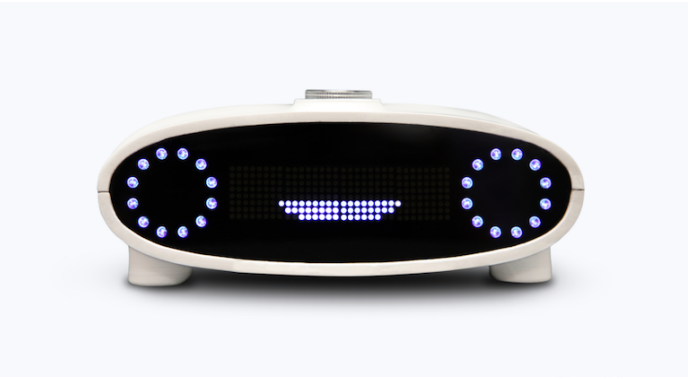
Collaboration between drones

Collaboration between Humans/Robots



Robot companions are coming !





Thank You !

Homework

Do the following tutorial:

https://docs.docker.com/engine/getstarted/

Run the following command in your terminal:

docker run -p 8888:8888 jupyter/pyspark-notebook

(info: https://github.com/jupyter/docker-stacks)

Then check that you can access this web page:

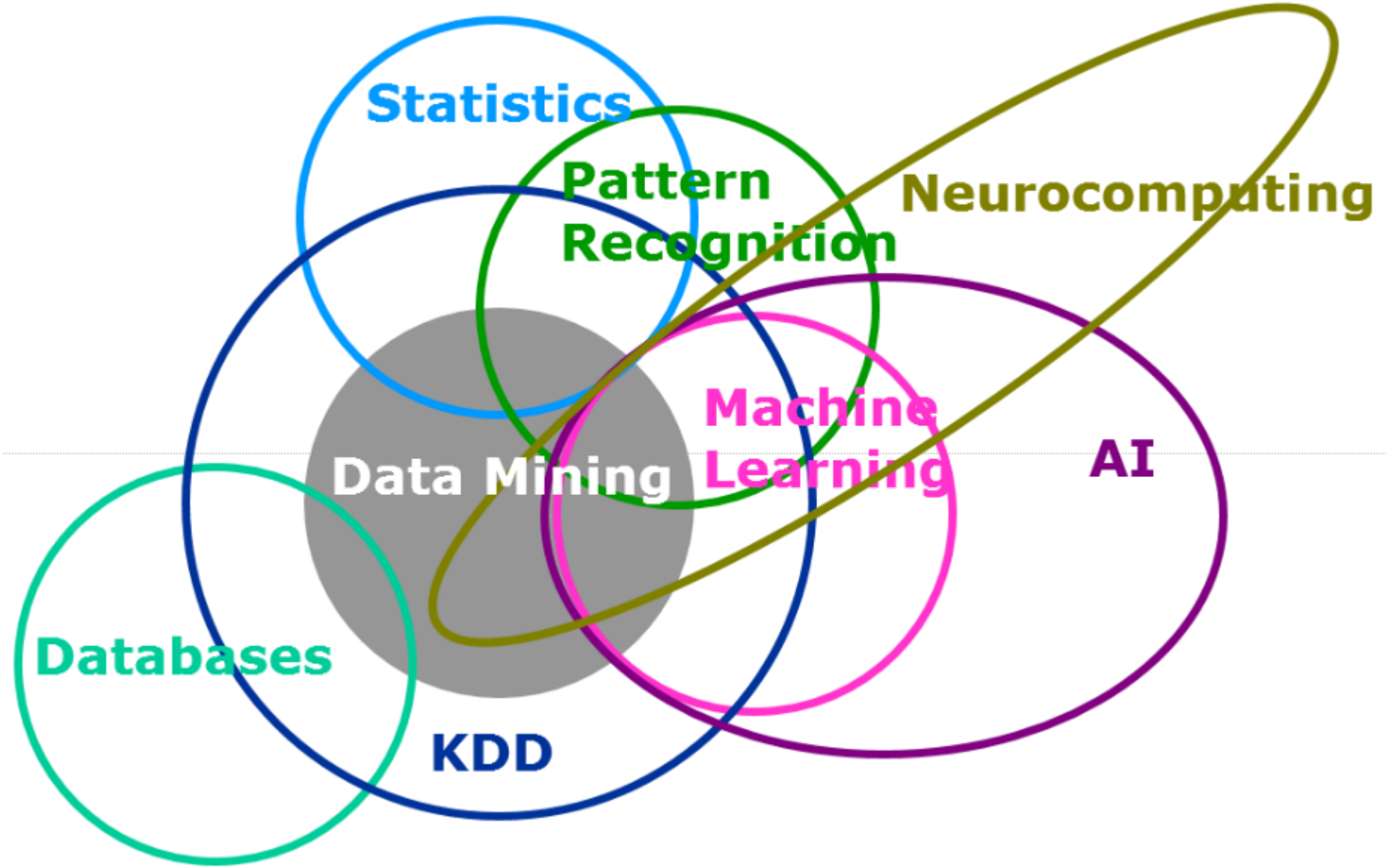
http://localhost:8888/

Old Slides

What kind of prediction we can make from the data?

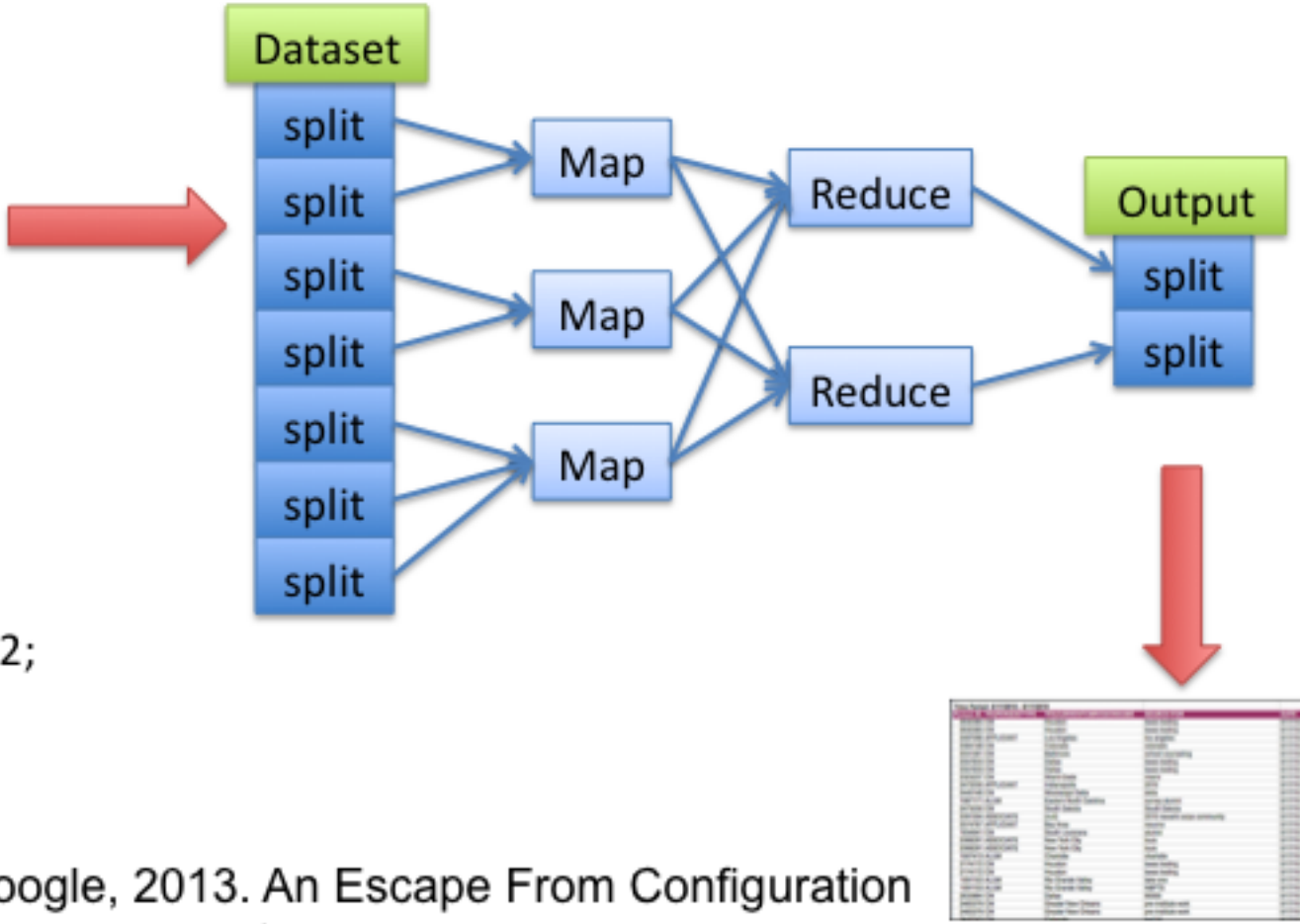
- **Prescriptive**: determine what actions should be taken
- **Predictive**: determine what can happen
- **Diagnostic**: describe what happened
- **Descriptive**: describes what is happening now

How do we predict things ?



Map Reduce: an Example of a BigData Solution

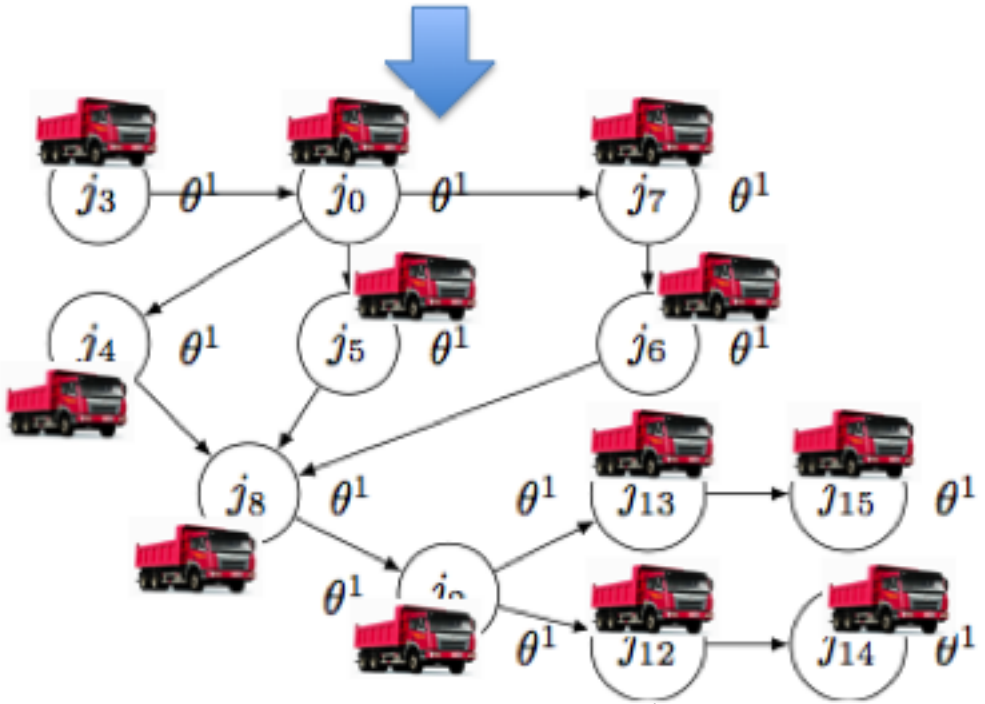
```
Map(){
...
}
Reduce(){
...
}
Main(){
...
io.file.buffer.size=4096;
maximum.map.tasks=3;
maximum.reduce.tasks=2;
...
}
```



* Matt Welsh at Google, 2013. An Escape From Configuration Hell. matt-welsh.blogspot.com/

Declarative -> MapReduce

```
SELECT SUPP_NATION, CUST_NATION, L_YEAR,
SUM(VOLUME) AS REVENUE
FROM ( SELECT N1.N_NAME AS SUPP_NATION,
N2.N_NAME AS CUST_NATION,
datepart(yy, L_SHIPDATE) AS L_YEAR,
 L_EXTENDEDPRICE*(1-L_DISCOUNT) AS VOLUME
 FROM SUPPLIER, LINEITEM, ORDERS, CUSTOMER
 NATION N1, NATION N2
 WHERE S_SUPPKEY = L_SUPPKEY AND
 O_ORDERKEY = L_ORDERKEY AND
        ...
 L_SHIPDATE BETWEEN
        '1995-01-01' AND '1996-12-31' )
GROUP BY SUPP_NATION, CUST_NATION, L_YEAR
ORDER BY SUPP_NATION, CUST_NATION, L_YEAR
```

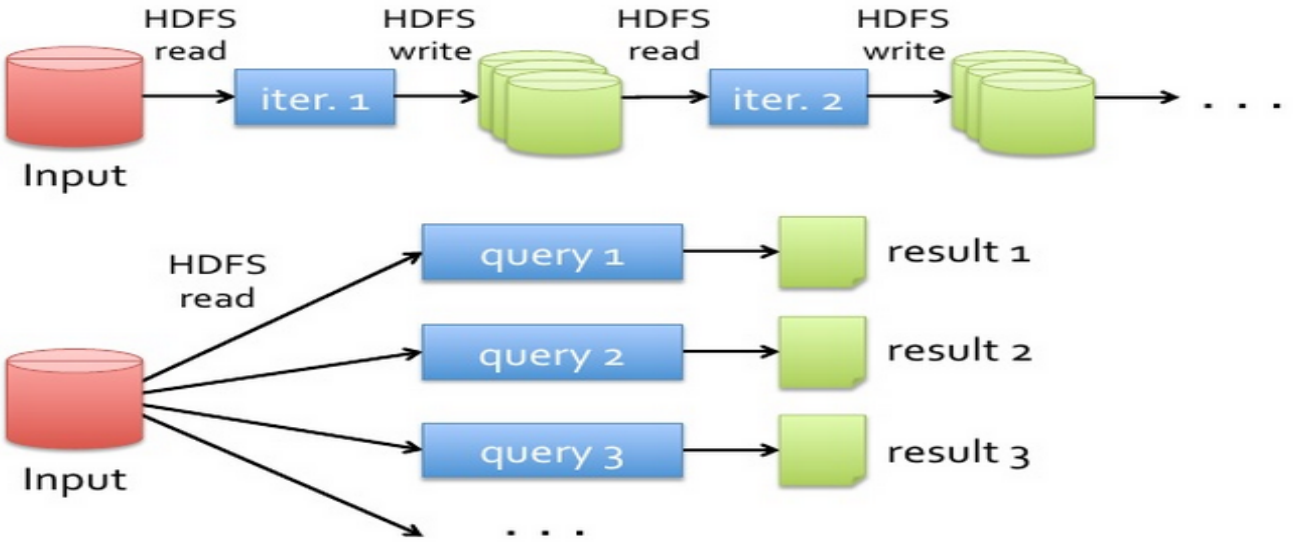


Have you heard of Hadoop Spark ?

- Spark is new technology that sits on top of HDFS (Hadoop Distributed File System) that is characterized as “a fast and general engine for large-scale data processing.”
- Spark solves similar problems as Hadoop MapReduce does but with a fast in-memory approach and a clean functional style API

Hadoop VS Spark

Hadoop



Spark