CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Big Data Processing

L02: Big Data Motivation

Dr. Ignacio Castineiras
Department of Computer Science

Outline

1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

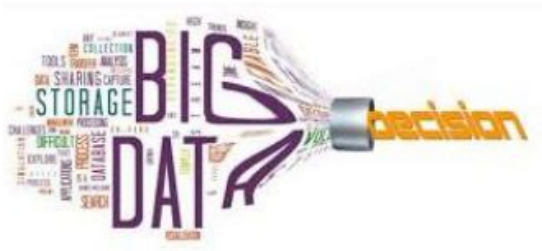
Outline

1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

Big Data Processing - Upside

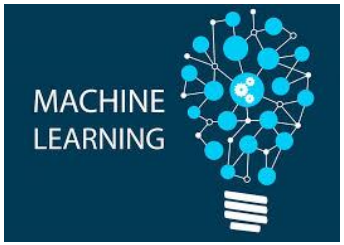
Artificial Intelligence: my view

Big Data



+

Machine Learning



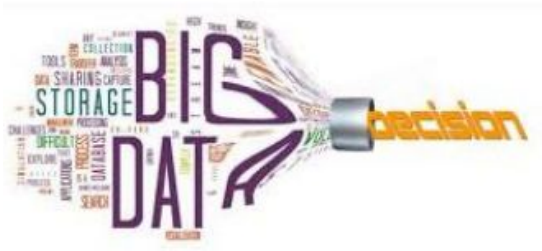
+

Combinatorial Optimisation

Big Data Processing - Upside

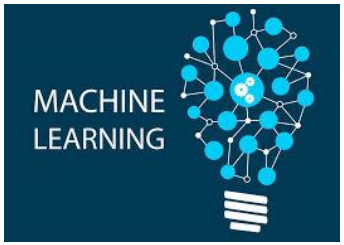
Artificial Intelligence: my view

Big Data



+

Machine Learning



+

Combinatorial Optimisation

Let me try to convince you of this with a couple of examples.

Big Data Processing - Upside

Example1:

Google HashCode'18: Self-driving rides
https://storage.googleapis.com/coding-competitions.appspot.com/HC/2018/hashcode2018_qualification_task.pdf

Big Data Processing - Upside

Example1:

Google HashCode'18: Self-driving rides
https://storage.googleapis.com/coding-competitions.appspot.com/HC/2018/hashcode2018_qualification_task.pdf

This is a problem we tackled 2 years ago, as part of the CIT Programming Society

Big Data Processing - Upside

Problem:

Given some passenger petitions given in advance,
can we schedule our taxi fleet rides to maximise the service?

Big Data Processing - Upside

More in detail...

1. We are managing a company of taxis.

➢ Typically, each taxi driver would like to maximise its own revenue.

➢ However, let's assume we are in a model with self-driving cars. On it, we don't care about the revenue of a particular taxi; instead we try to maximise the revenue produced by the whole fleet of taxis.

Big Data Processing - Upside

More in detail…

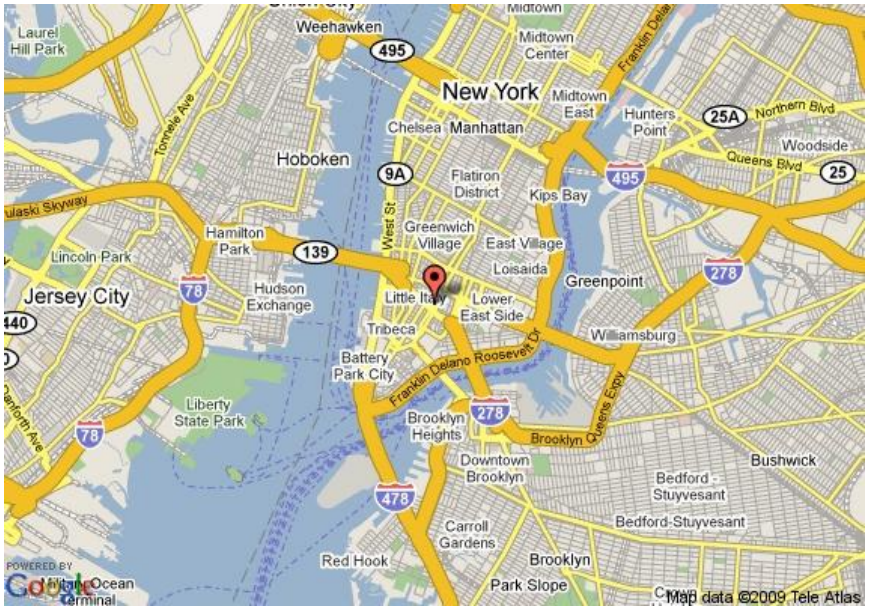
2. Let's imagine our taxis operate in a city.

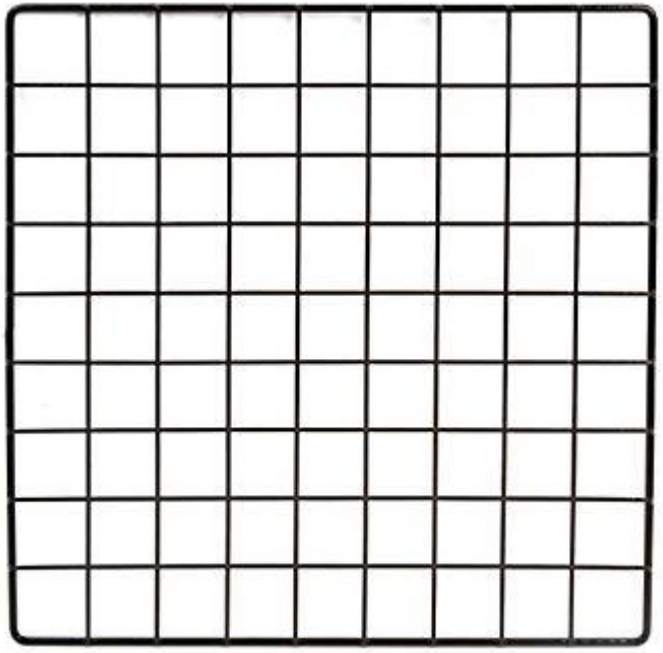
FREENOW

Big Data Processing - Upside

More in detail...

2. The streets of this city can be modelled as a grid.





Big Data Processing - Upside

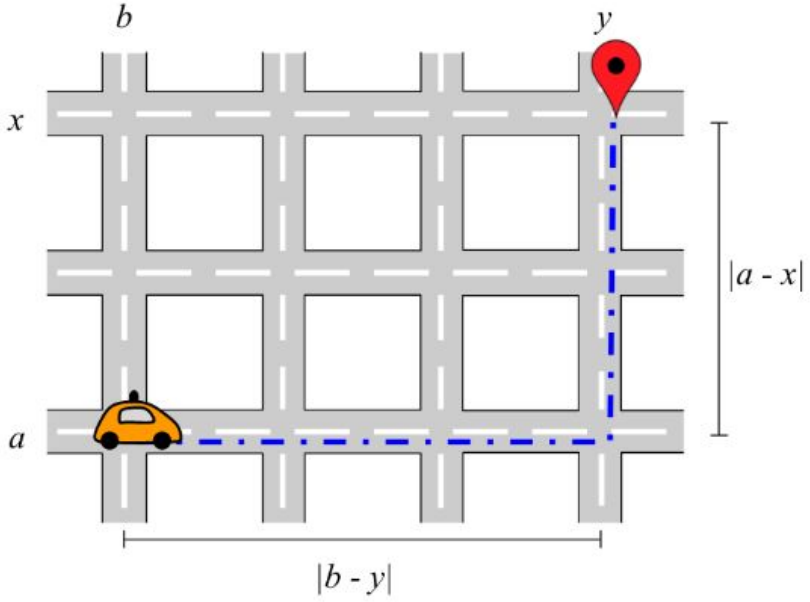
More in detail…

3. Let's imagine we have a set of customers, who have requested our taxi service in advance!

Big Data Processing - Upside

More in detail...

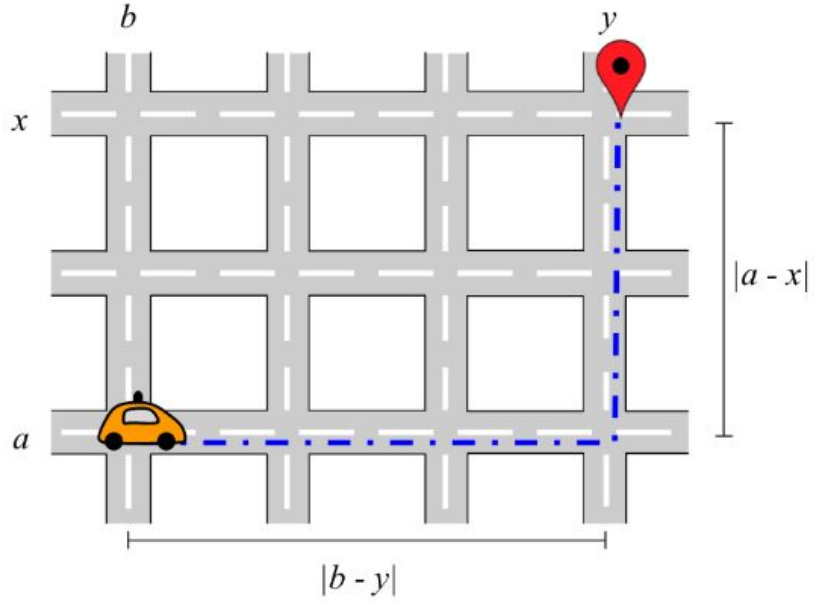
4. Each of these petitions can be represented via:
   a. The point to depart from.
   b. The point to reach to.
   c. The time of departure.



Big Data Processing - Upside

More in detail...

5. Likewise, the revenue for a trip can be modelled as the Euclidean distance between the start and end points, with a bonus for trips made on time.



Big Data Processing - Upside

More in detail...

6. Let's revisit the problem statement now:
   - You are the manager of Free Now.
   - You control a fleet of taxis.
   - It's early in the morning, and you know in advance all the customer petitions Free Now is going to have during the day.

<u>The problem is:</u>

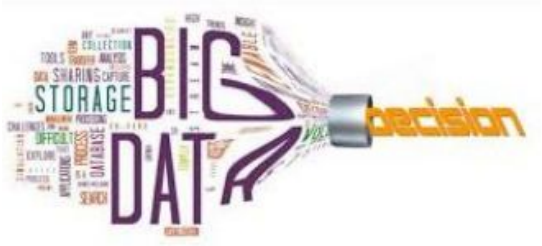
➢ Create an algorithm to schedule the customer trips taken by the taxis, with the goal of maximising the revenue made by the entire fleet of taxis.

Big Data Processing - Upside

Example1:

Clearly this is a Combinatorial Optimisation Problem, and it thus belong to the third component of the Artificial Intelligence view described before:

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

Example1:

But, wait a moment...

Did we just claim that, early in the morning,
we know **in advance**
all the taxi petitions we are going to receive during the day??!!

FREENOW

Big Data Processing - Upside

Example1:

But, wait a moment...

Did we just claim that, early in the morning,
we know **in advance**
all the taxi petitions we are going to receive during the day??!!

Is this science fiction?

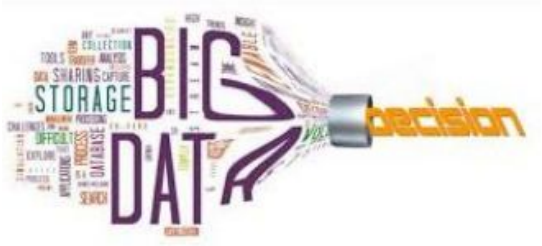
FREENOW

Big Data Processing - Upside

Example1:

Let's start reasoning about our problem with the other two Artificial Intelligence components in mind:

Big Data



+

Machine Learning

+

Combinatorial Optimisation

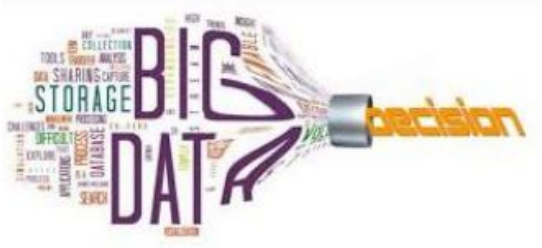
Big Data Processing - Upside

Example1:

Let's start reasoning about our problem with the other two Artificial Intelligence components in mind:

This is where the story becomes interesting for you as Data Enginners

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

Example1:

What would you do if you were working for Free Now, and you had to estimate in advance the amount of taxi petitions you are going to receive during the day?

Big Data Processing - Upside

Example1:

What would you do if you were working for Free Now, and you had to estimate in advance the amount of taxi petitions you are going to receive during the day?

I tell you: struggling! That's what you will do :)

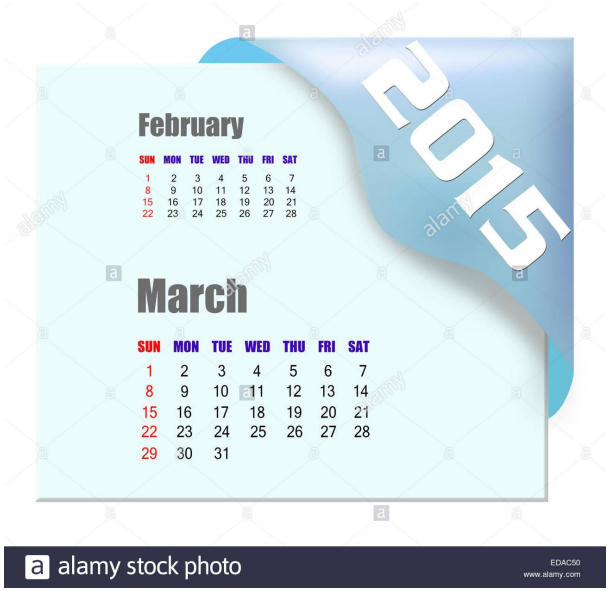
Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- Maybe you want to take a look at Free Now log files from past years.

➢ How many petitions did we receive in that year?



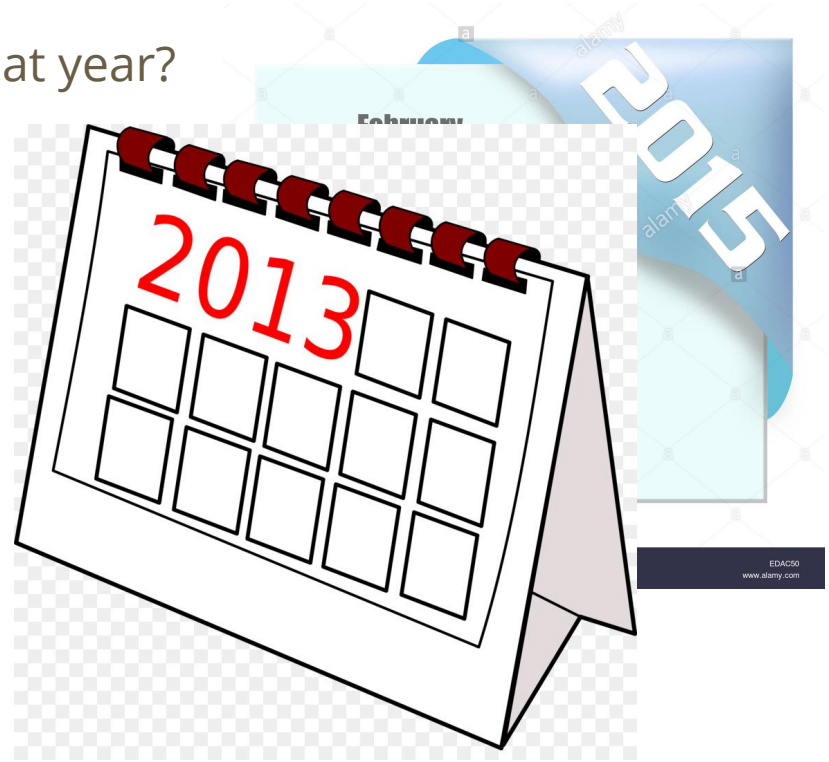
Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- Maybe you want to take a look at Free Now log files from past years.

➢ How many petitions did we receive in that year?
➢ And in that other year?



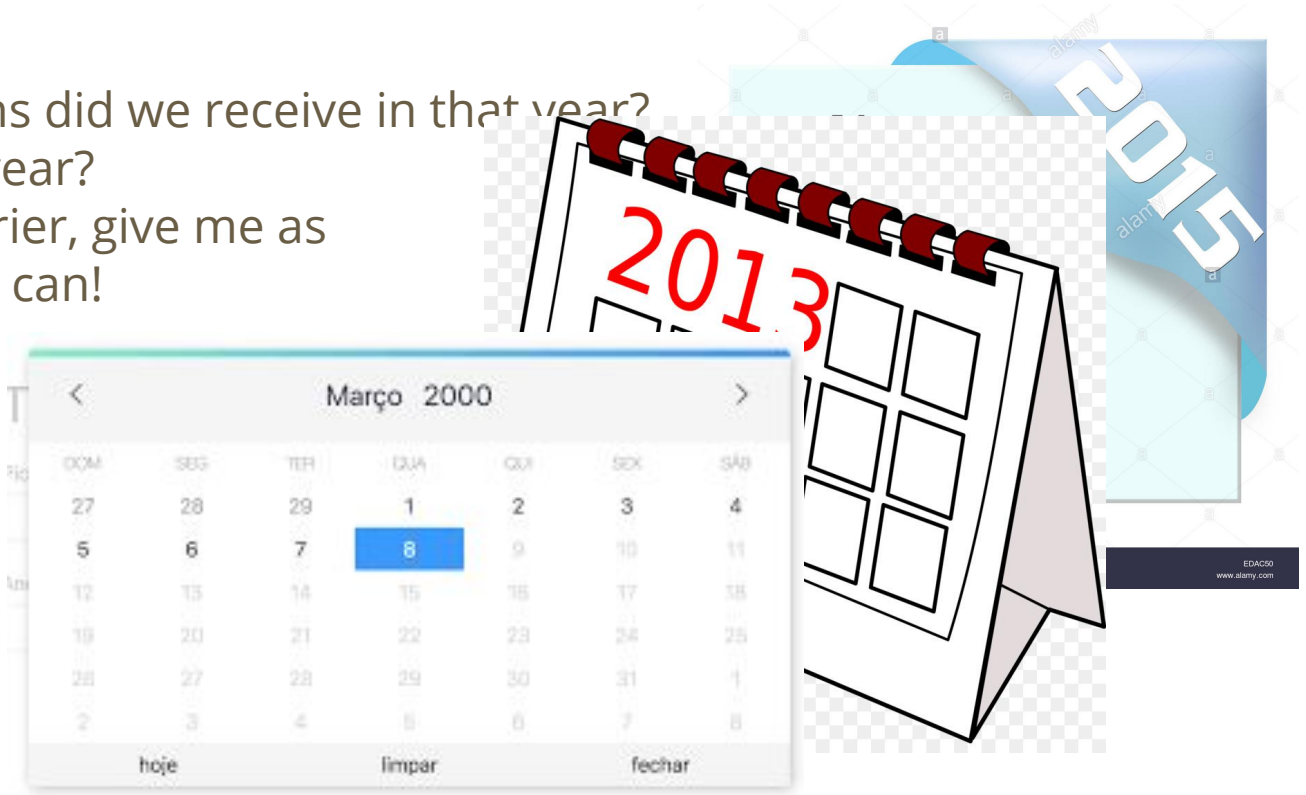
Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- Maybe you want to take a look at Free Now log files from past years.

➢ How many petitions did we receive in that year?
➢ And in that other year?
➢ The more the merrier, give me as many years as you can!



Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- But that's not enough.
  What if the weather also has an impact in the amount of taxis being taken?

➢ What was the weather like in these last days?

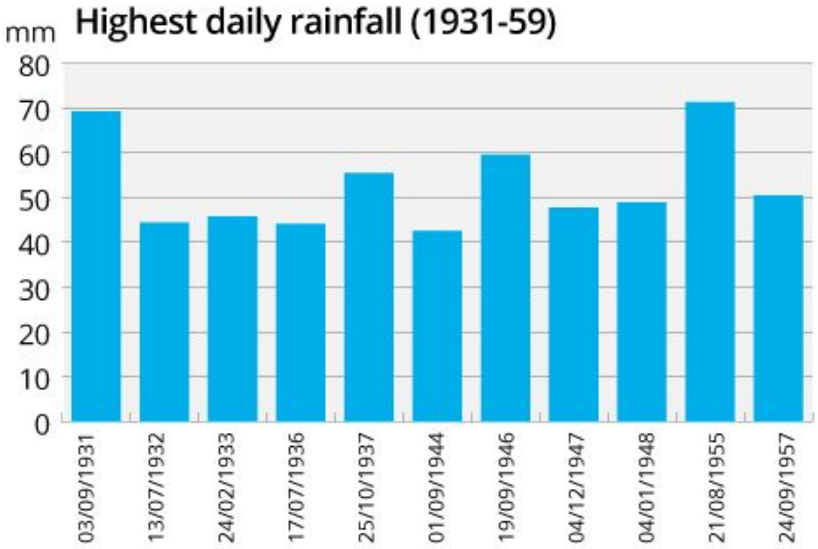
Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- But that's not enough.
  What if the weather also has an impact in the amount of taxis being taken?

➢ What was the weather like in these last days?
➢ And, by these last days I mean the last thousands of days :)



Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- Is that enough? Maybe the income of the people matters too...

➢ How many people own a car these days?

Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- Is that enough? Maybe the income of the people matters too...

➢ How many people own a car these days?
➢ And how many did in the past?

Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- And is it that all?

➢ What about climate awareness?

ZERO % EMISSION

Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- And is it that all?

➢ What about the last bus monthly-pass offer?

Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- And is it that all?

➢ And what about...?
➢ And also...?
➢ Don't forget about...?

Big Data Processing - Upside

Example1:

It is certainly not an easy task, but:

- And is it that all?

➢ And what about…?
➢ And also…?
➢ Don’t forget about…?

I told you: struggling! That’s what you will do :)

Big Data Processing - Upside

Example1:

In other words,
you want to put your hands into tons of data

Big Data Processing - Upside

Example1:

In other words,
you want to put your hands into tons of data
and make sense of it!

Big Data Processing - Upside

Example1:

This is the role of the Big Data component!
You analyse the data, looking for patterns, correlations, aggregations, and ultimate conclusions.

Big Data

+

Machine Learning

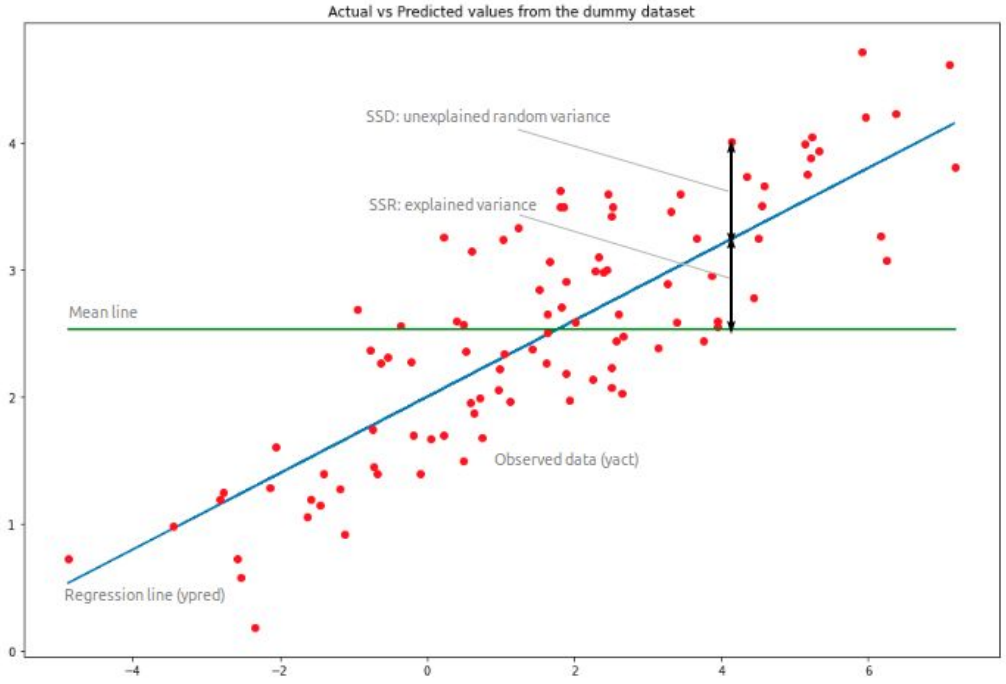
+

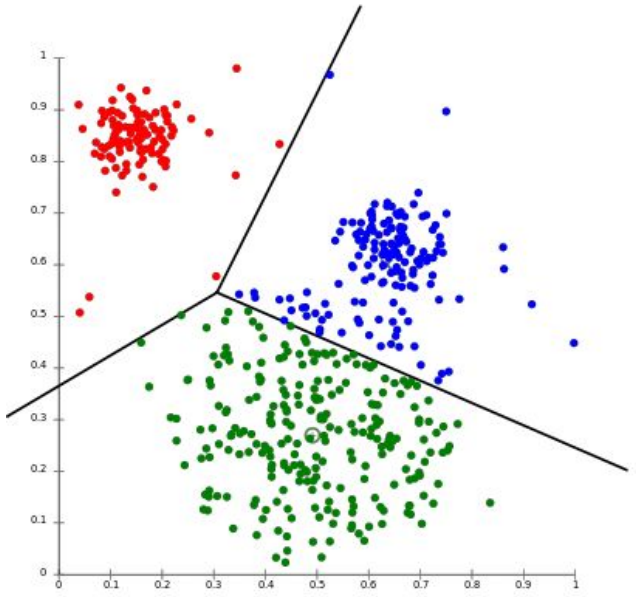
Combinatorial Optimisation

Big Data Processing - Upside

Example1:

Then you pass on these patterns, correlations, aggregations, and conclusions so as to build prediction models.



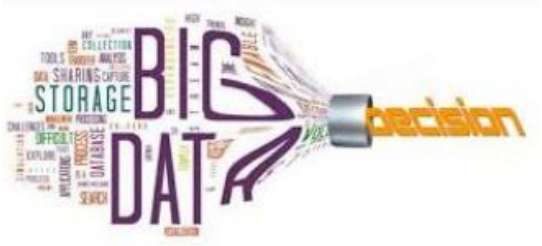


Big Data Processing - Upside

Example1:

This is done by the Machine Learning component!

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

Example1:

Now, you collect novel data,
perhaps obtained every new hour, or even every new minute,
or even every new second.

Big Data Processing - Upside

Example1:

This is the role of the Big Data Streaming Analysis component again!

Big Data

+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

Example1:

This is the role of the Big Data Streaming Analysis component again!
This component keeps up to date with the data being received and processes it, whatever the ingestion pace is.

Big Data

+

Machine Learning

+

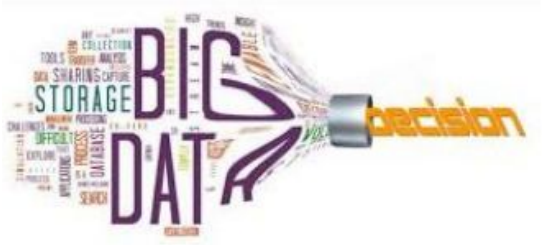
Combinatorial Optimisation

Big Data Processing - Upside

Example1:

And this novel data, once analysed, can be fed into our model predictor previously created, so as to provide us with an updated prediction for the novel data just received.

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

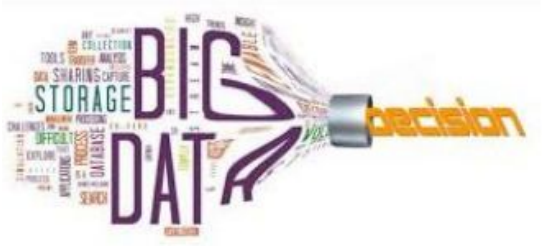
And now, and only now...

Big Data Processing - Upside

Example1:

Finally, with this prediction for the novel data
we can use combinatorial optimisation so as to enable
better decisions.

Big Data



+

Machine Learning

+

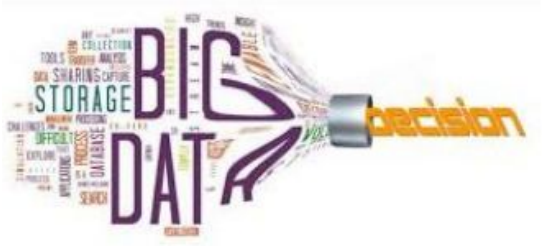
Combinatorial Optimisation

Big Data Processing - Upside

Example1:

In our case, once we can predict the taxi customer petitions for the new day, we can come back to our Google Hash Code'18 constraint optimisation problem and solve it maximising the revenue generated by the taxi fleet.

Big Data



+

Machine Learning

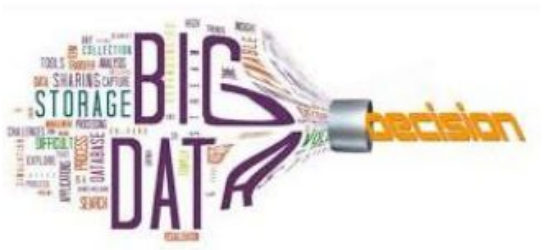
+

Combinatorial Optimisation

Big Data Processing - Upside

Example1:

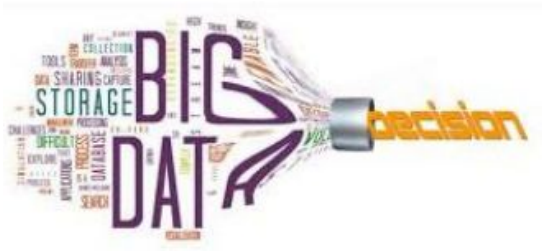
And this is just one example of how Artificial Intelligence can be used as part of a use-case providing us with a better taxi service, that ultimate reduces the number of private cars and improves the CO2 levels.



Big Data Processing - Upside

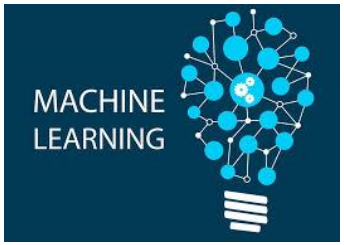
Artificial Intelligence: my view

Big Data



+

Machine Learning



+

Combinatorial Optimisation

Let me try to convince you of this with a couple of examples.

Big Data Processing - Upside

Example2:

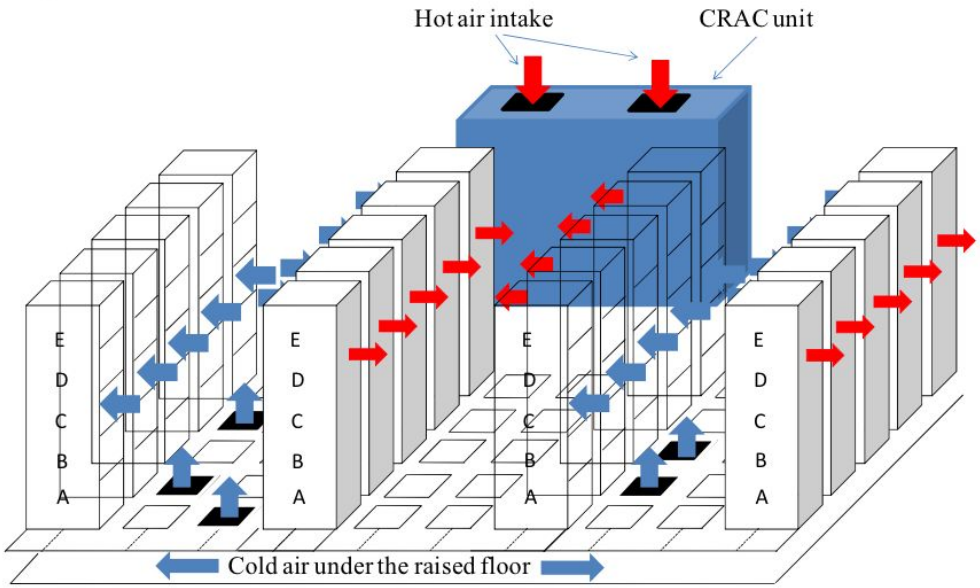
Years ago I worked for the European Union funded research project GENiC: Globally Optimised Energy Efficient Data Centres
http://projectgenic.eu/

Big Data Processing - Upside

Example2:

The project fell under the research areas of:

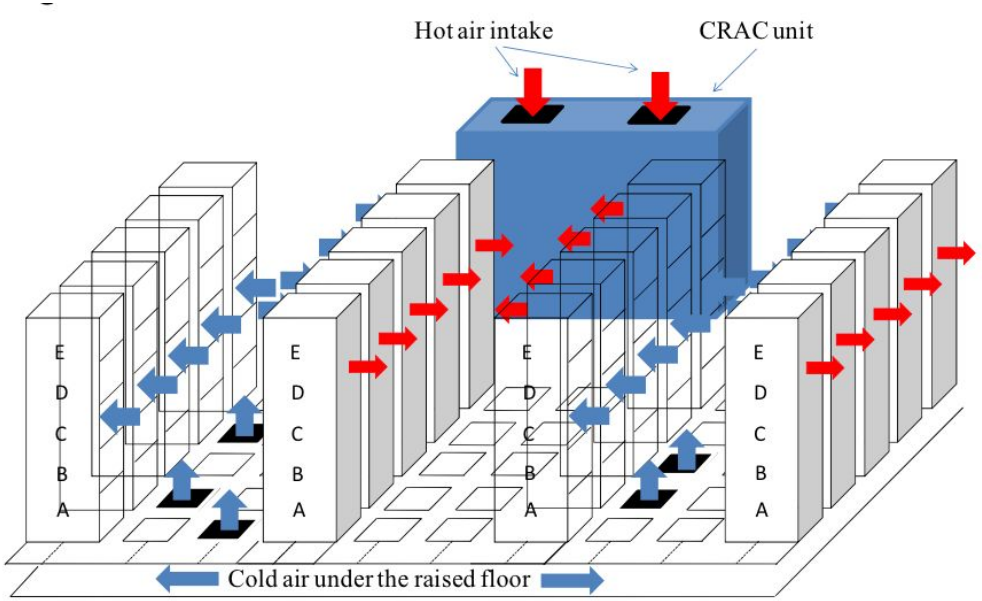
- Green computing.
- Sustainable Data Centres.
- Renewable energy sources.



Big Data Processing - Upside

Example2:

- On it, we had a set of servers (which play the role of the fleet of taxis).



Big Data Processing - Upside

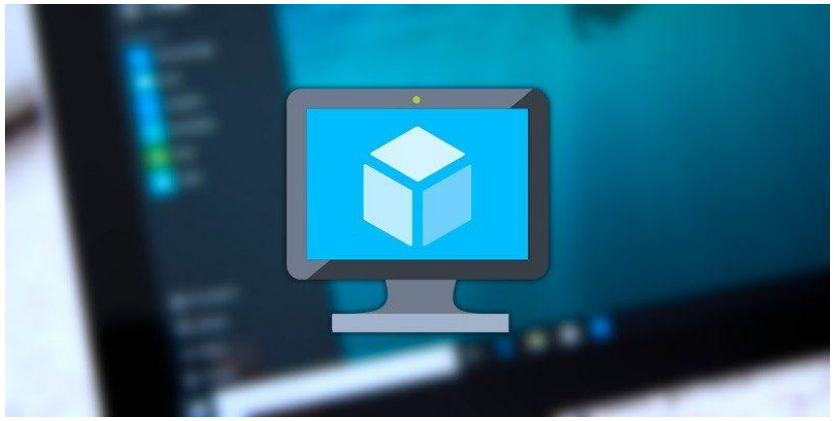
Example2:

- We had a set of customer petitions (e.g., using YouTube).

Big Data Processing - Upside

Example2:

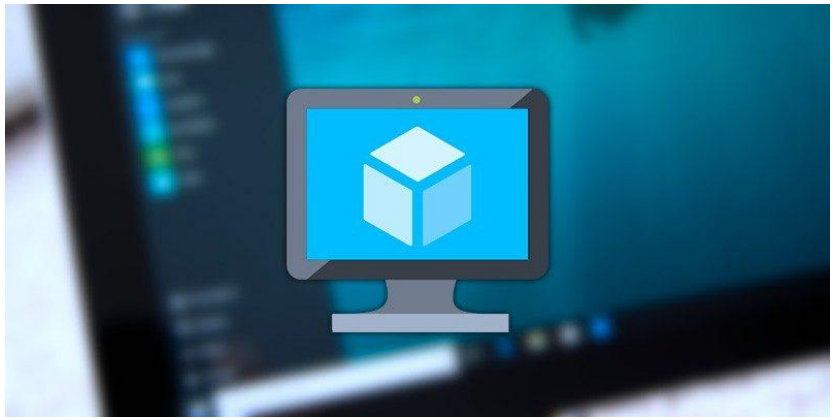
- Each customer petition was identified by the time and the amount of resources needed. This petition was processed by a Virtual Machine (VM) consolidated in a server.



Big Data Processing - Upside

Example2:

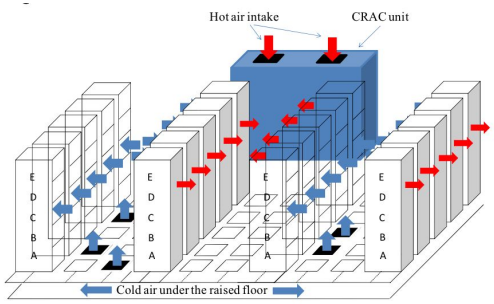
- Each customer petition was identified by the time and the amount of resources needed. This petition was processed by a Virtual Machine (VM) consolidated in a server.
- This played the role of our customer asking for taxis. A VM can satisfy multiple customer petitions (as long as it has the resources for it).



Big Data Processing - Upside

Example2:

- The goal was to minimise the power consumption for running the customer petitions by wisely consolidating them in concrete VMs of the Data Centre.

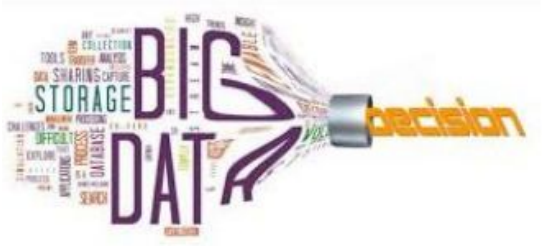


Big Data Processing - Upside

Example2:

Once again, this was clearly a Combinatorial Optimisation Problem, and it thus belong to the third component of the Artificial Intelligence view described before:

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

Example2:

However, at this stage I hope you have found the similarity between this example and the previous one.

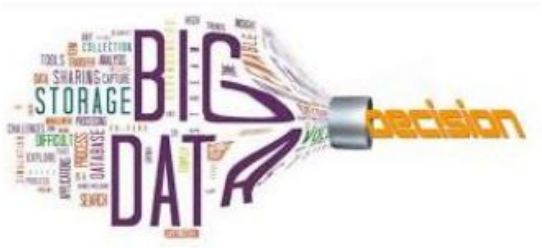
How on earth did we get to know the customer petitions **in advance**??!!

Big Data Processing - Upside

Example2:

In this case we skip the reasoning about the Data Engineering components for Big Data and Machine Learning, but you can just guess their role from the previous example.

Big Data



+

Machine Learning

+

Combinatorial Optimisation

Big Data Processing - Upside

When you look at these examples and the way
Artificial Intelligence in general
(and Data Engineering in particular)
are shaping our society
you might wonder...

Big Data Processing - Upside

When you look at these examples and the way
Artificial Intelligence in general
(and Data Engineering in particular)
are shaping our society
you might wonder...

When did all of this start?

Big Data Processing - Upside

This is a very difficult question to answer
as it has been a race with many hints.

But let me highlight the following one (see video from 1:20 to 3:00):

https://www.youtube.com/watch?v=x7qPAY9JqE4

Outline

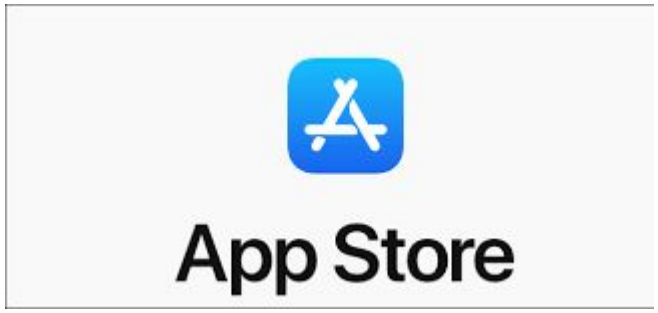
1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

Big Data Processing - Downside

- I remark the arrival of the SmartPhone as it comes with plenty of sensors:
  - Wifi Connection.
  - Accelerometer.
  - GPS.

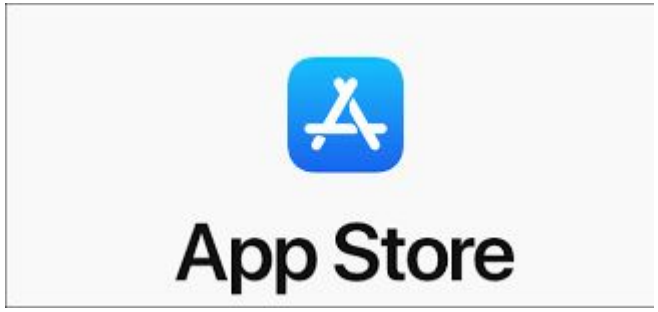
Big Data Processing - Downside

- Ever since people started downloading mobile apps.



Big Data Processing - Downside

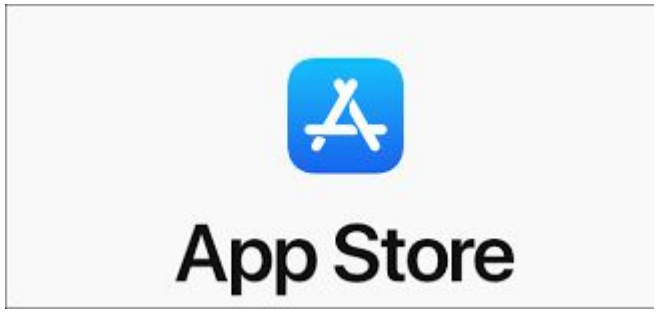
- Ever since people started downloading mobile apps.



- Most of these mobile apps are *free*
  (in the sense that one does not need to pay for downloading them).

Big Data Processing - Downside

- Ever since people started downloading mobile apps.



- Most of these mobile apps are *free*
(in the sense that one does not need to pay for downloading them).

- However, nothing is really for free, and we just *pay* the price of these apps with our data, with the information we generate.

Big Data Processing - Downside

- Just think for a second of 1 permission most apps require from us:

ACCESS_FINE_LOCATION

```
<manifest
    xmlns:android="http://schemas.android.com/apk/res/android"
    package="---"
    android:versionCode="1"
    android:versionName="1.0">

    <uses-permission android:name="android.permission.INTERNET" />
    <uses-permission android:name="android.permission.SYSTEM_ALERT_WINDOW"/>

    <uses-permission android:name="android.permission.ACCESS_FINE_LOCATION" />
    <uses-permission android:name="android.permission.ACCESS_COARSE_LOCATION" />
    <application
      android:name=".MainApplication"
      android:label="@string/app_name"
```

Big Data Processing - Downside

Basically, this means answering to the question:

Where are you?

Big Data Processing - Downside

- Of course accessing our location is crucial for the service offered by some apps:

Big Data Processing - Downside

- But, there are many other apps requiring ACCESS_FINE_LOCATION too. And, for most of them, it is unclear which service(s) they won't be able to offer us in case the permission is declined.

Big Data Processing - Downside

- But, there are many other apps requiring ACCESS_FINE_LOCATION too. And, for most of them, it is unclear which service(s) they won't be able to offer us in case the permission is declined.

Big Data Processing - Downside

- But, there are many other apps requiring ACCESS_FINE_LOCATION too. And, for most of them, it is unclear which service(s) they won't be able to offer us in case the permission is declined.

Big Data Processing - Downside

- But, there are many other apps requiring ACCESS_FINE_LOCATION too. And, for most of them, it is unclear which service(s) they won't be able to offer us in case the permission is declined.

Big Data Processing - Downside

- Why is it so much important to know your location?

Big Data Processing - Downside

- Why is it so much important to know your location?

Because knowing where you are 24/7
is a way of inferring the answer
to a much more powerful question:

Big Data Processing - Downside

- Why is it so much important to know your location?

Because knowing where you are 24/7
is a way of inferring the answer
to a much more powerful question:

Who are you?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

This resonates much more (as it makes the metaphor explicit)
if you are watching these slides in your mobile phone:

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

The examples presented here are from a talk of Chema Alonso,
Chief Data Officer (CDO) of Telefonica,
one of the strongest telecommunication companies in the world.

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

1. **Information derived from your location:**
   Where do you live?
   Where do you go on holidays?
   Where do you work?
   Where do you do your shopping?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

2. **Information derived from how your location changes over time:**
   Do you walk?
   Do you go by car?
   By bike perhaps?
   Maybe a combination of them?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

3. **Let's now cross your location with a database of points of interest:**
   These are databases with millions of points, updated automatically and fixed manually, everyday.

   Do you remember the movie *Dude, where's my car?*
   So, with Siri, this problem can no longer happen.

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

3. **Let's now cross your location with a database of points of interest:**
   The app detects that you are going at car speed.
   Then, you stop in a location which stands for a restaurant in its point of interests database.
   You spend about an hour there.
   Finally, when you go out and start walking again...

   ...the app informs you, automatically, where did you park your car.

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

4. **Crossing your location with the time of the day it is:**
   Where do you sleep?
   Do you sleep every night in the same place?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

4. **Crossing your location with the time of the day it is:**
   Do you go to the gym?
   With who?
   And how often?
   And for how long?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

4. **Crossing your location with the time of the day it is:**
   Do you play sports?
   Do you go to the pub?
   And what about restaurants? Which ones do you like?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

5. **And the list just goes on…**
   Do you go to the doctor?
   Do you go to the hospital?
   Do you visit somebody in jail?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

5. **And the list just goes on…**
   Coming back to where you work and the exact location of your office:
   - What is your role in your company?
   - What is your salary?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

5. **And the list just goes on…**
   Combining where do you live with a state property management database:
   - Do you own your place or are you renting?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

5. **And the list just goes on…**
   What is your close circle?
   Who do you stay with?

Big Data Processing - Downside

Let's think together of the information
that can be inferred from your location.

5. **And the list just goes on…**
   Do you go to political meetings?
   Or to demonstrations?
   Do you have a faith? Do you go to church?

Big Data Processing - Downside

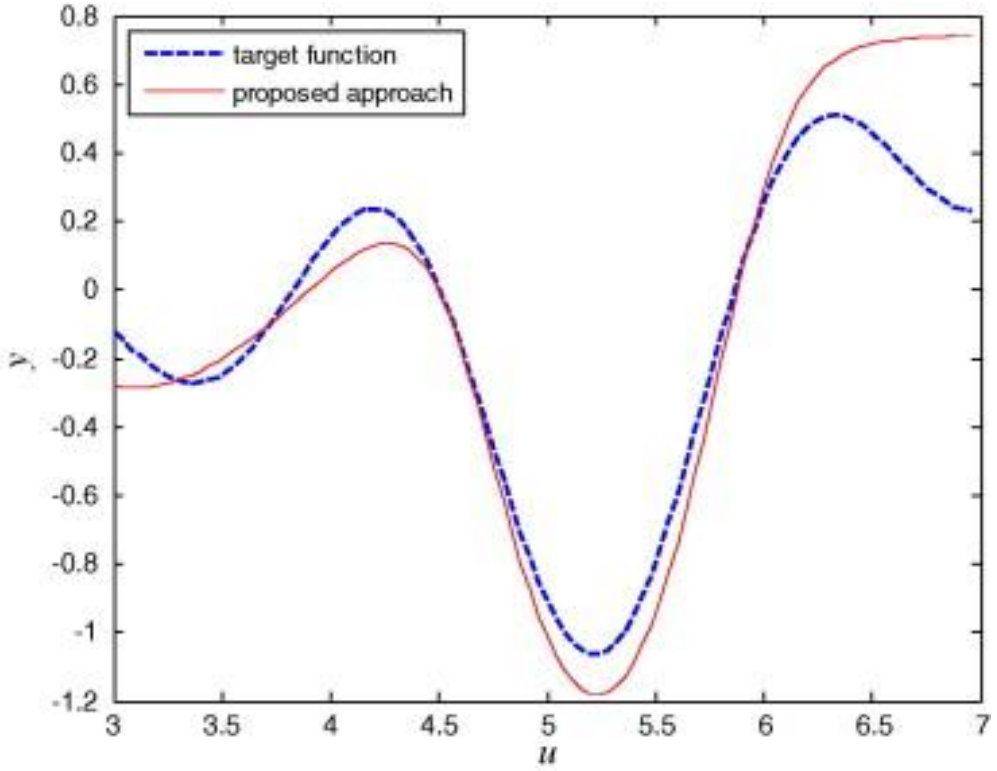
All in all...

Where are you?  **Who are you?**

Big Data Processing - Downside

All in all...

Where are you? ➡ **Who are you?**



Big Data Processing - Downside

But this is not all...

Big Data Processing - Downside

But this is not all...

all the above examples were based on the exterior, in actual actions performed by individuals...

Big Data Processing - Downside

...but what if we change slightly the perspective and start looking at the interior...

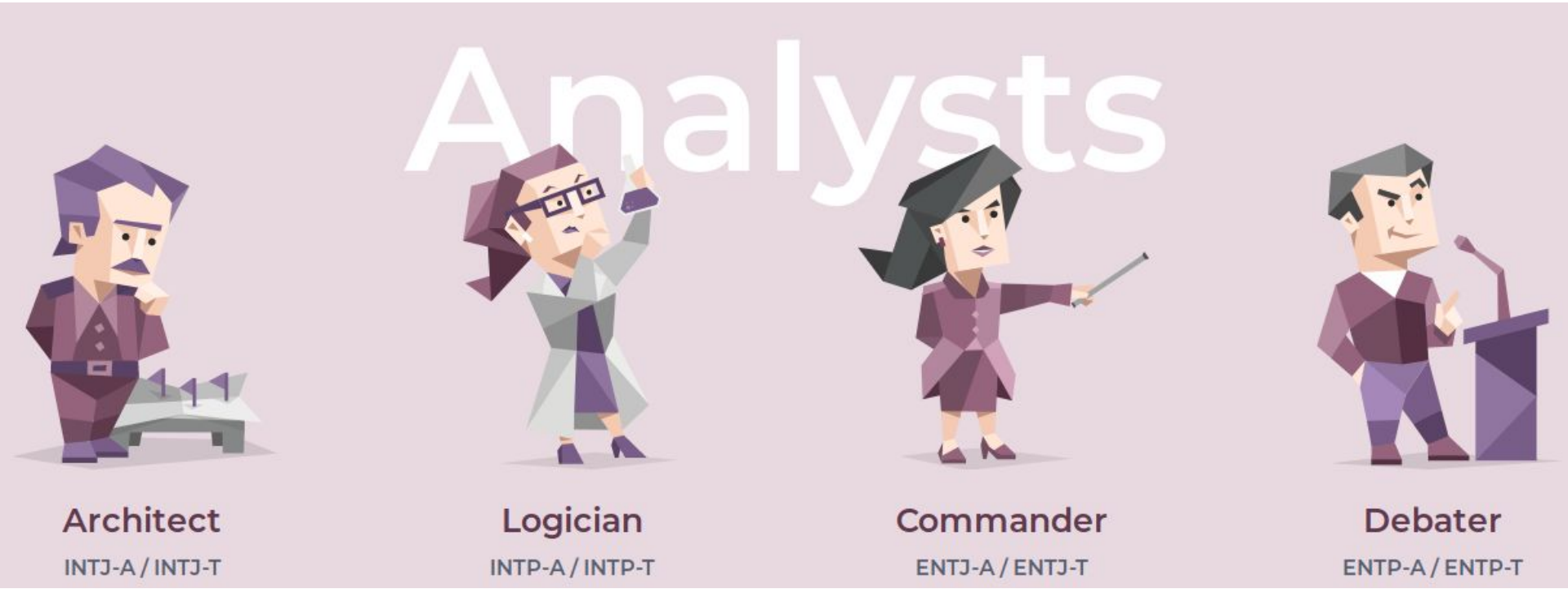
Big Data Processing - Downside

...but what if we change slightly the perspective and start looking at the interior...

at the thoughts, sensations and emotions of people (that is, at their inner feelings, at their personalities)

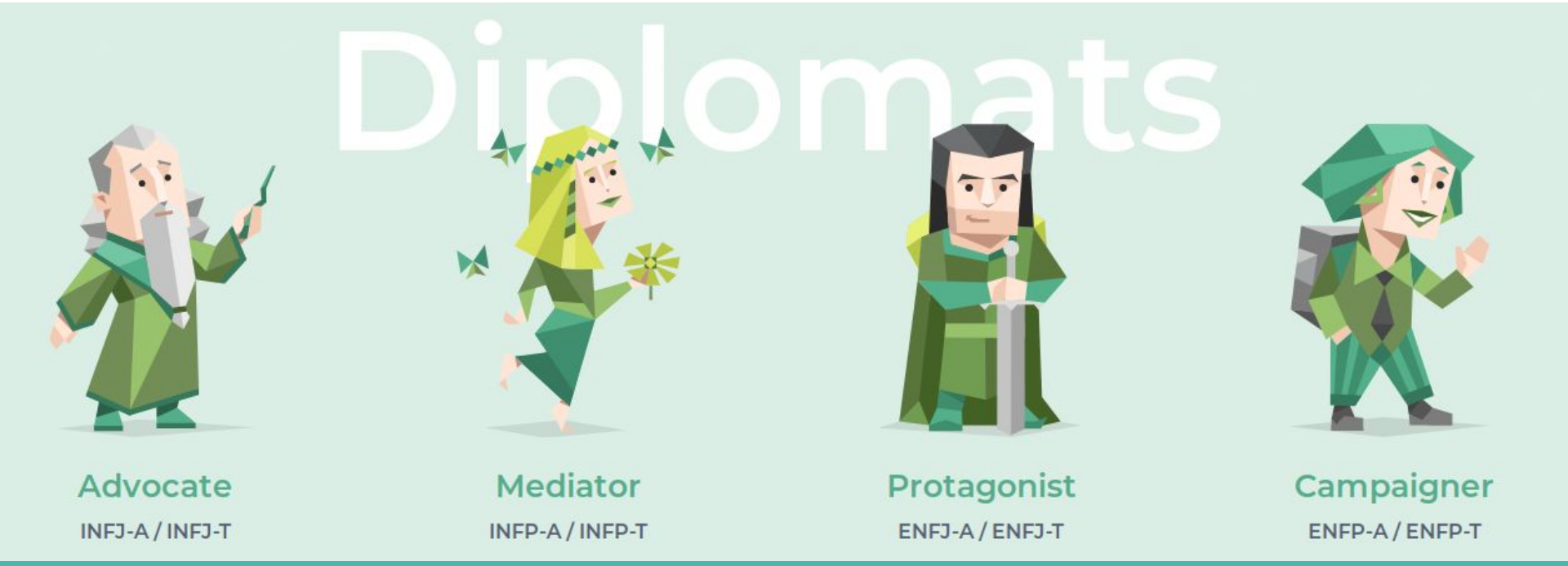
Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/



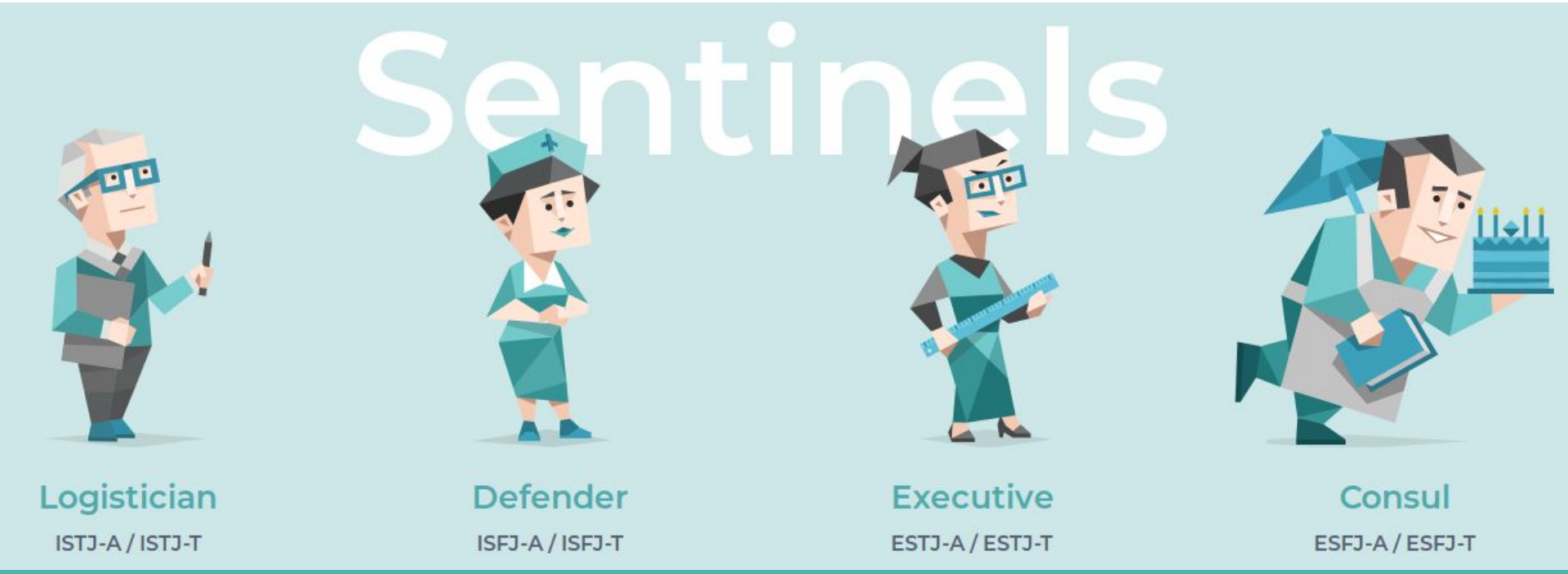
Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/



Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/



Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/

Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/

- Besides how accurate I found the stuff of this website, the two things blowing my mind the most are:

Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/

- Besides how accurate I found the stuff of this website, the two things blowing my mind the most are:

1. According to their model, people are not that much different.
   Indeed, just 4 traits and 16 personalities serve to represent the whole mass.

Big Data Processing - Downside

One of the websites I am most fascinated by is:
https://www.16personalities.com/

- Besides how accurate I found the stuff of this website, the two things blowing my mind the most are:

1. According to their model, people are not that much different.
   Indeed, just 4 traits and 16 personalities serve to represent the whole mass.
2. According to their model, it does not take much to match an individual to one of the 16 personalities; just quick 90 questions (that can be filled out in about 10 minutes) do the job.

Big Data Processing - Downside

And, needless to be said, knowing who everybody is, understanding each and everyone personalities, brings an incredible amount of opportunities...

Big Data Processing - Downside

And, needless to be said, knowing who everybody is, understanding each and everyone personalities, brings an incredible amount of opportunities...

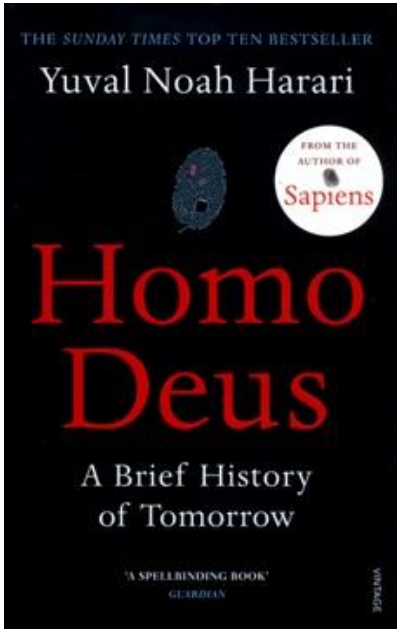
...together with an incredible amount of risks too!

Big Data Processing - Downside

- Recently, as a result of the scandal of Cambridge Analytica, the media has turned its attention to these risks:

Big Data Processing - Downside

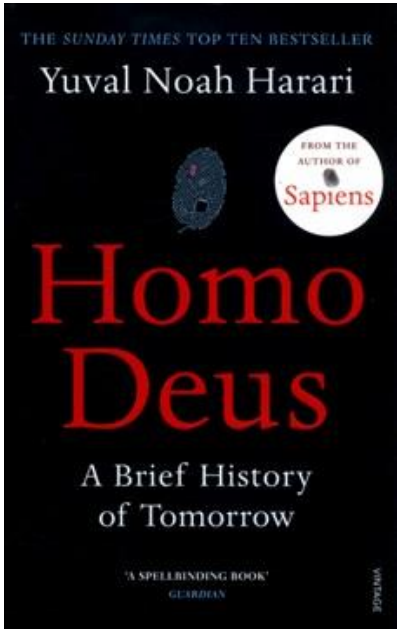
- But, in my humble opinion, the best way to reflect about the future that Data Analytics, Artificial Intelligence and Biotechnology will bring us to is by reading the book Homo Deus, from Professor Yuval Noah Harari.



Big Data Processing - Downside

- But, in my humble opinion, the best way to reflect about the future that Data Analytics, Artificial Intelligence and Biotechnology will bring us to is by reading the book Homo Deus, from Professor Yuval Noah Harari.

Let's finish this section with the last paragraph of his book, which I found so inspiring:



Big Data Processing - Downside

Professor Yuval Noah Harari, Homo Deus.

- *If we think in term of months, we had probably focus on immediate problems such as the turmoil in the Middle East, the refugee crisis in Europe and the slowing of the Chinese economy.*
- *If we think in terms of decades, then global warming, growing inequality and the disruption of the job market loom large.*
- *Yet if we take the really grand view of life, all other problems and developments are overshadowed by three interlinked processes:*

1. *Science is converging on an all-encompassing dogma, which says that organisms are algorithms,and life is data processing.*
2. *Intelligence is decoupling from consciousness.*
3. *Non-conscious but highly intelligent algorithms may soon know us better than we know ourselves.*

Big Data Processing - Downside

Professor Yuval Noah Harari, Homo Deus.

➢ *These three processes raise three key questions, which I hope will stick in your mind long after you have finished this book:*

1. *Are organisms really just algorithms, and is life really just data processing?*
2. *What's more valuable – intelligence or consciousness?*
3. *What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?*

Outline

1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

The Technological Revolution Driven by Big Data

We can move now to a more technical discussion about Big Data.

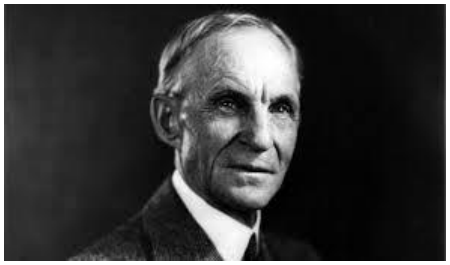
The Technological Revolution Driven by Big Data

When we talk about the technological race, we see it can be described by two kind of steps:

- Technological evolution (most of them).
- Technological revolution (most relevant ones).

The Technological Revolution Driven by Big Data

Example 1:



"If I asked the people what they wanted, they would have said faster horses"
– Henry Ford.

- Problem to be solved: Transport people.
- New problem? **No!**

- Inventing cars as an alternative to horses => ***Technological Revolution***.
- Create a faster car => *Technological Evolution*.

The Technological Revolution Driven by Big Data

Example 2:

Alan Turing was recruited to help to decode the messages produced by the Enigma machine in the second world war.

- Problem to be solved: Decode Messages.
- New problem? **No!**

- Develop a machine to defeat enigma => ***Technological Revolution***.
- Develop a new algorithm allowing the machine to double its decoding speed => *Technological Evolution*.

The Technological Revolution Driven by Big Data

"The good news is that Big Data is here. The bad news is that we are struggling to store and analyse it".
- Tom White, author of the book 'Hadoop: The definite guide'.

- Problem to be solved: Gather data and analyse it, so as to get some insights / hidden knowledge from it.
- New problem? **No!**

- Develop a more efficient hard drive, processor or data base so as to store and analyse data => *Technological Evolution*.
- Modify the hard drive, processor and data base architectures so as to store and analyse data => ***Technological Revolution***.

The Technological Revolution Driven by Big Data

What are the ingredients making Data Science and Big Data to become a technological revolution?

1. Data Scalability.
2. Processing Scalability.

The Technological Revolution Driven by Big Data

On the one hand...data scalability.

1. Historically, data was being generated and accumulated by workers. Employees of companies were entering data into computer systems.

The Technological Revolution Driven by Big Data

On the one hand...data scalability.

2. With internet, and specifically with the web 2.0, users can generate their own data. Think of facebook, all these users are signing in and entering the data themselves.

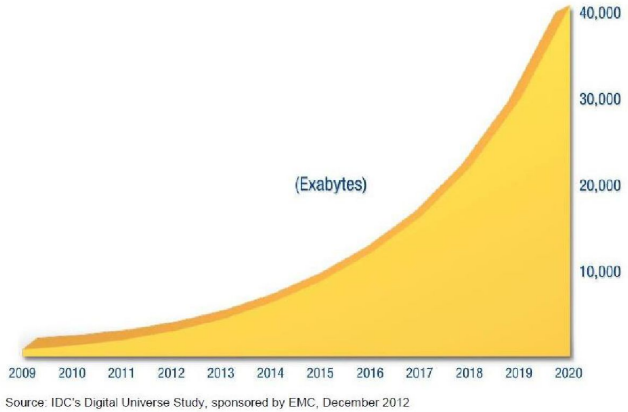
   This is larger amount of data (w.r.t. 1) by orders of magnitude!.

The Technological Revolution Driven by Big Data

On the one hand…data scalability.

3. With the Internet of Things (all these sensors being placed in our buildings, streets, satellites monitoring temperature, humidity, electricity) are generating data as well.

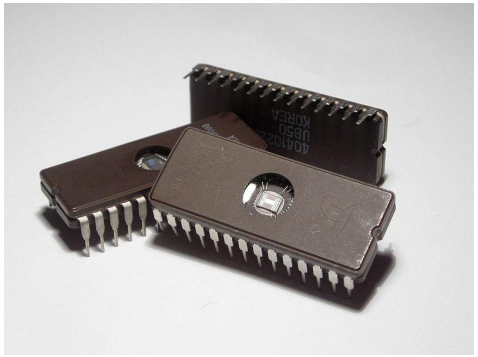
   This is larger amount of data (w.r.t. 2) by orders of magnitude!



The Technological Revolution Driven by Big Data

On the other hand… processing scalability.

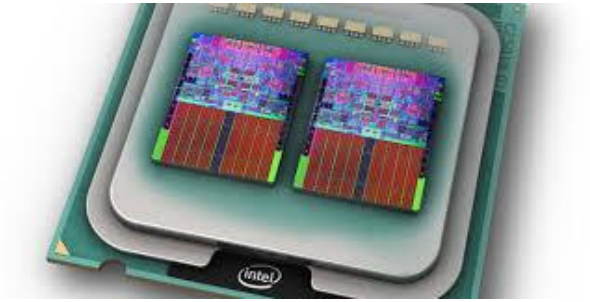
1. With the Integrated Circuit achievement 50 years ago the Moore's law stated that the number of transistors on each processor will be doubled each two years.



The Technological Revolution Driven by Big Data

On the other hand… processing scalability.

2. In 2002 it stopped, as the limit of transistors one can fit in a CPU without reaching electromigration was reached. Companies chose then to pursue the multi-core avenue.



The Technological Revolution Driven by Big Data

On the other hand... processing scalability.

3. However, even multi-core has a limit, with the refrigeration of the computer, its power consumption, its price, its maintenance...

The Technological Revolution Driven by Big Data

On the other hand… processing scalability.

4. Thus, pressure is making companies to leverage their own IT infrastructure and rely in own private clouds consolidated in external data centres.

The Technological Revolution Driven by Big Data

In summary.

- The problem of gathering data and analyse it is not new.

What makes the current context different is that companies are realizing that all the data collected by their business operations, and the one constantly being collected by web trends, consumer behaviour and social media can be combined in interesting and useful ways to gain competitive advantage or have better outcomes.

The Technological Revolution Driven by Big Data

The Technological Revolution Driven by Big Data

- However, to be able to do so, the older computer infrastructures and the algorithms operating on them are not applicable any more
- **Data Science Revolution has taken (and is taking) place.**

Outline

1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

Building a Distributed Storage and Compute Cloud

- Big Data refers to a huge volume of data that cannot be stored and processed with a traditional approach within a given time frame.

- Examples:
  - Process 10TB of images, in order to modify their brightness or resize them.
  - Monitor all information generated by all flights landing in a particular airport, and react to it in real time.
  - Store all messages generated by the Twitter users, and analyse them to decide what is a trending topic.

Building a Distributed Storage and Compute Cloud

- Big data is comprised of datasets that grow so large that they become cumbersome to manipulate using traditional database management tools or traditional data processing applications (store, retrieve, compute...)

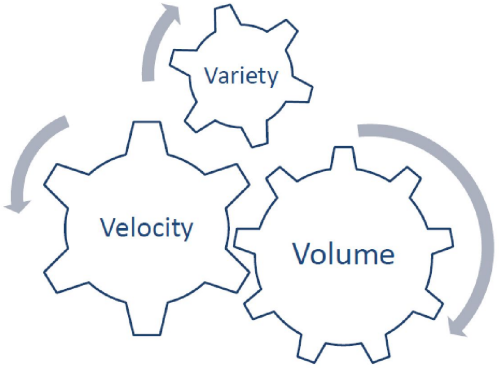
- These difficulties include:
  - Capture the data.
  - Store the data.
  - Search on the data.
  - Share the data.
  - Analyse the data.
  - Visualise the data.

Building a Distributed Storage and Compute Cloud

Big Data Classification.

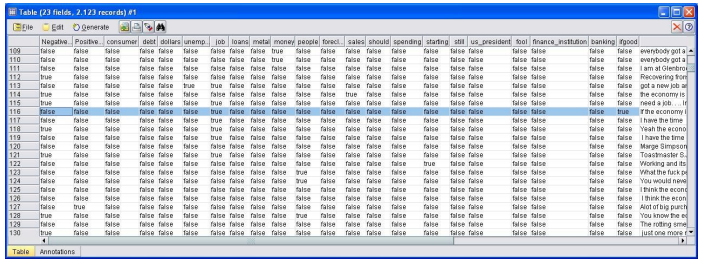
Big data is usually characterised by its three "V's":

1. Variety / Complexity.
    - o How diverse is the data being generated?
2. Velocity / Speed.
    - o How fast is the data coming in?
3. Volume.
    - o How much data is being generated?



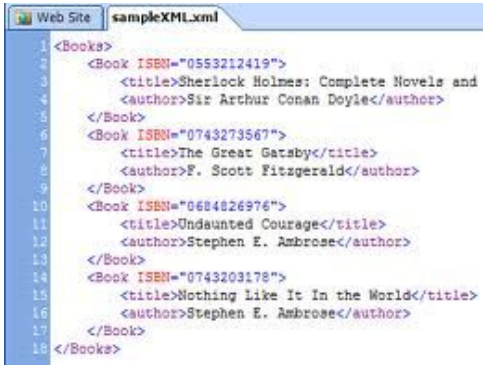
Building a Distributed Storage and Compute Cloud

1. Variety / Complexity



Big data 3 categories:

- Structured data: Data that has a proper format associated to it.
  - Excel, RDBMS table…
- Semi-structured data: It has a format, but it is flexible.
  - JSON, HTML…
- Unstructured data: It does not have any format associated to it.
  - Audio, video…

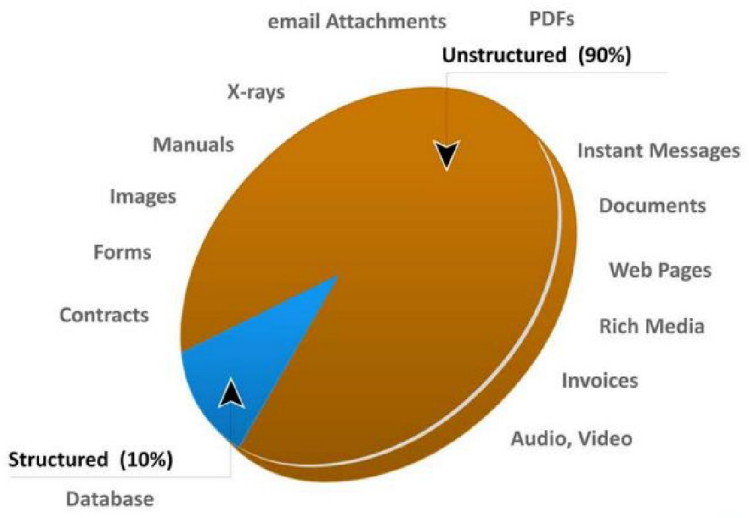


Building a Distributed Storage and Compute Cloud

1. Variety / Complexity

Data is no longer just about text and numbers. The information is now linked, and consists of multiple data types.

Applying typical algorithms for search, storage and categorisation in this data is now much more complex and inefficient.



Building a Distributed Storage and Compute Cloud

2. **Velocity / Speed.**

- Streaming media over the internet, slow motion video for surveillance or high-definition video have very high ingestion rates.
- Business have to keep up with the data flow to make the information useful.

Building a Distributed Storage and Compute Cloud

3. **Volume.**

- All collected data must be stored in a secure location which also makes it always available.
- With such high volumes of data, IT teams have to decide when it is “too much data”.
- For example, they might flush all data each week and start over the following week. However, for some companies this is not an option, which has its impact in the infrastructure responsible of storing it and make it accessible.

Building a Distributed Storage and Compute Cloud

Big Data Analysis:

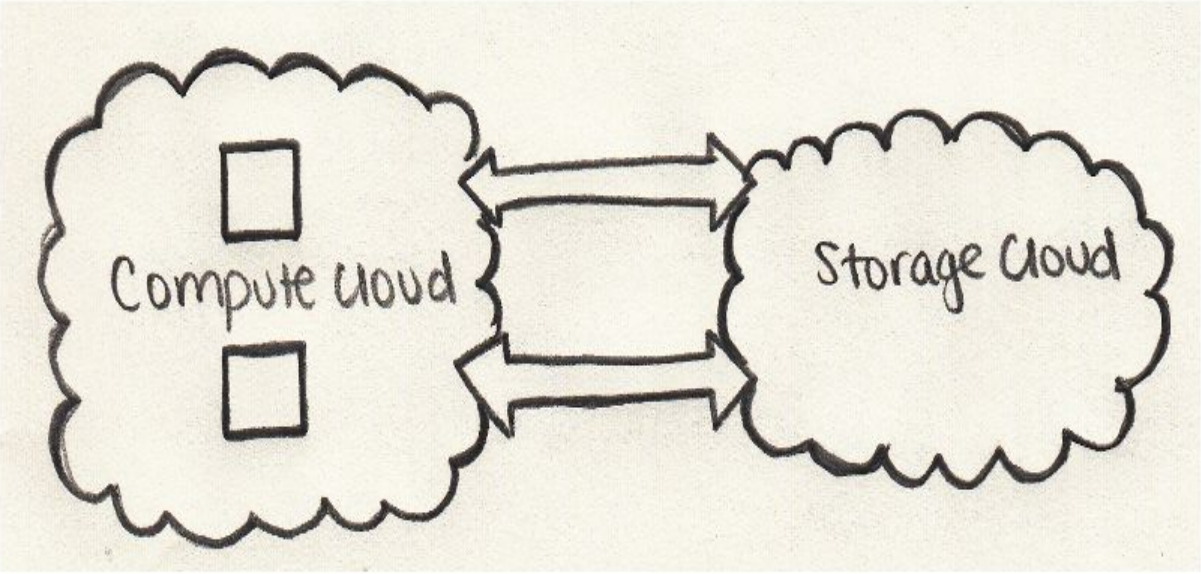
- Analytics for extreme large data sets focus on providing efficient analysis. It is all about turning the data into high-quality information, providing deeper insight about the business to enable better decisions.

- Thus, provided the vast amount of information, the main challenges become:
  - How to efficiently access to each piece of data?
  - How to efficiently perform computations across all data?

Building a Distributed Storage and Compute Cloud

Big Data Analysis:

- The traditional data processing model has data stored in a 'storage cluster', which is copied over to a 'compute cluster' for processing, and the results are written back to the ‘storage cluster’.



Building a Distributed Storage and Compute Cloud

Big Data Analysis:

- This model however doesn't work well for Big Data, because copying so much data out to a compute cluster might be too time consuming or impossible.
- Get the data to the processor can become a bottleneck:
  - Typical disk data transfer rate: 75MB/sec
  - Time taken to transfer 100GB of data to the processor: approx 22 minutes!
  - Assuming sustained reads, actual time will be worse, since most servers have less than 100GB of RAM available

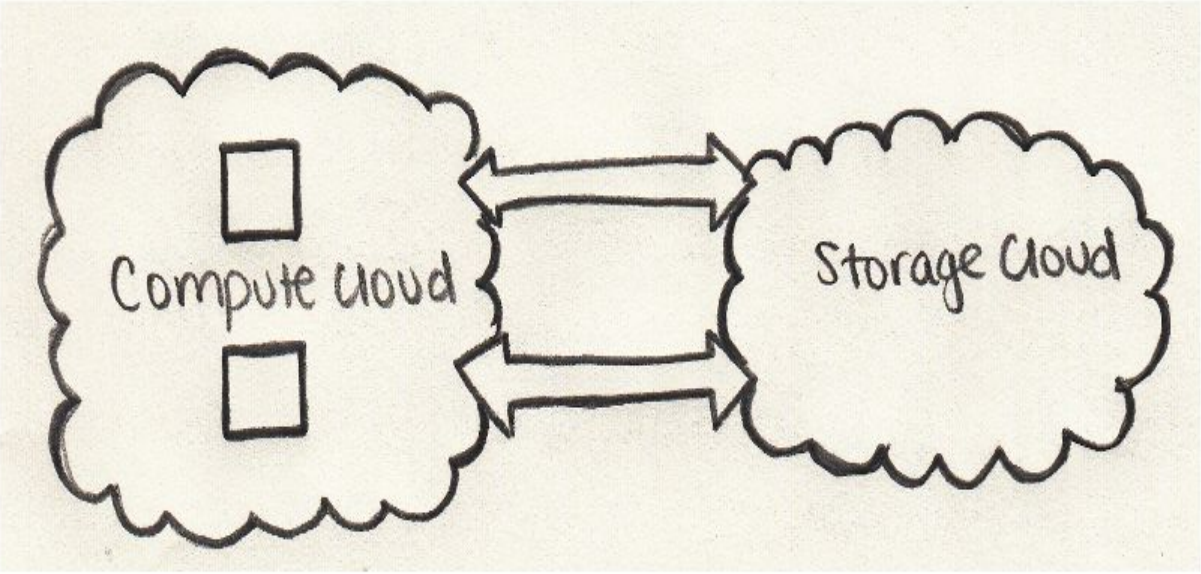
A different approach is needed!

Building a Distributed Storage and Compute Cloud

Big Data Analysis:

New processing model:

If data cannot go to the processor…bring the processor to the data!

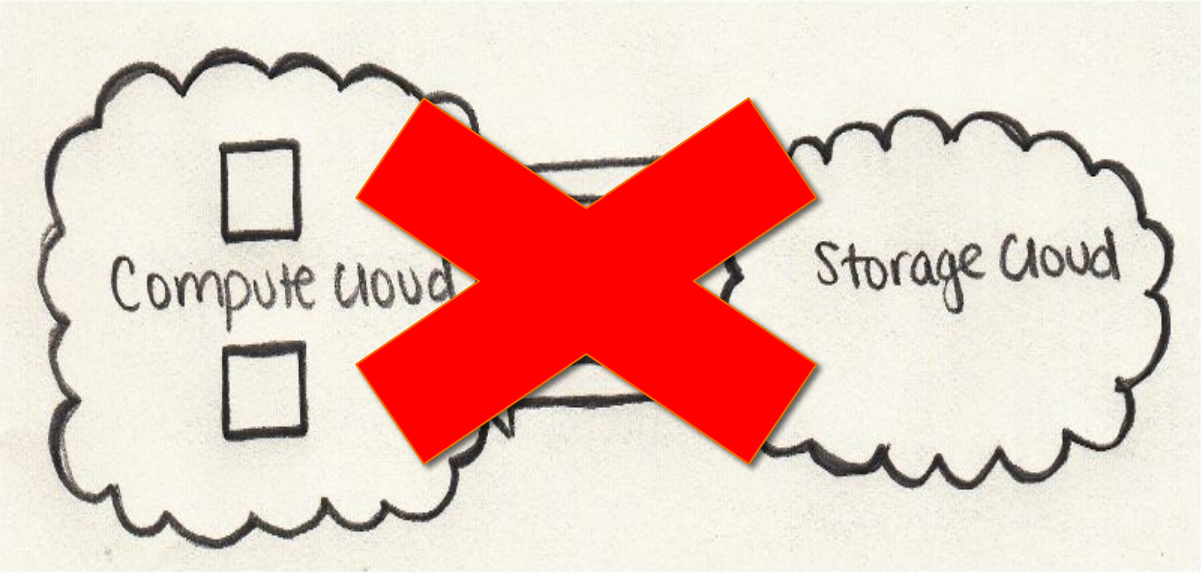


Building a Distributed Storage and Compute Cloud

Big Data Analysis:

New processing model:

If data cannot go to the processor…bring the processor to the data!

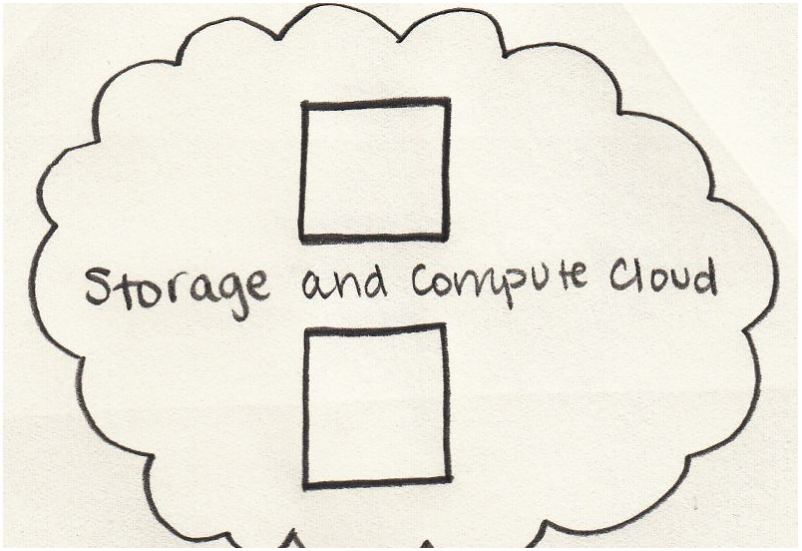


Building a Distributed Storage and Compute Cloud

Big Data Analysis:

New processing model:

If data cannot go to the processor…bring the processor to the data!



Building a Distributed Storage and Compute Cloud

Big Data Analysis:

- Core concept: distribute the data as it is initially stored in the system.
  - Computation happens where the data is stored, wherever possible.
  - No data transfer over the network is required for initial processing.
  - Data is replicated multiple times on the system for increased availability and reliability.

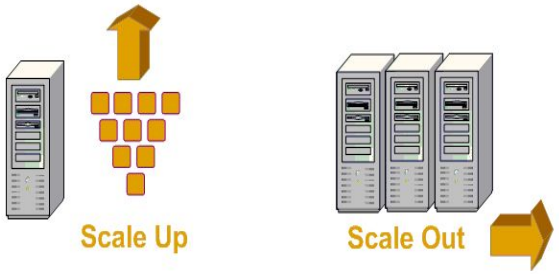
Building a Distributed Storage and Compute Cloud

Big Data Analysis: Distributed Storage and Processing.

But, if the data is split among different nodes in the cluster, and each piece of data is processed on the node it is hosted, then...
...our processing model has suddenly become a distributed one!

- Scale up vs scale out

"In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers."



Building a Distributed Storage and Compute Cloud

Big Data Analysis: Distributed Storage and Processing.

- Pros and cons of distributed systems:

+ Performance.
+ Scalability.
+ Fault Tolerance.

- Management.
- Synchronization.
- Interconnection.

Outline

1. Big Data Processing - Upside.
2. Big Data Processing - Downside.
3. The Technological Revolution Driven by Big Data.
4. Building a Distributed Storage and Compute Cloud.

Thank you for your attention!