Université Constantine 2
جامعة قسنطينة 2

Big Data Processing and Analysis

Reminder: Big data

Dr. Rostom Mennour
Faculty of New Technologies
rostom.mennour@univ-constantine2.dz

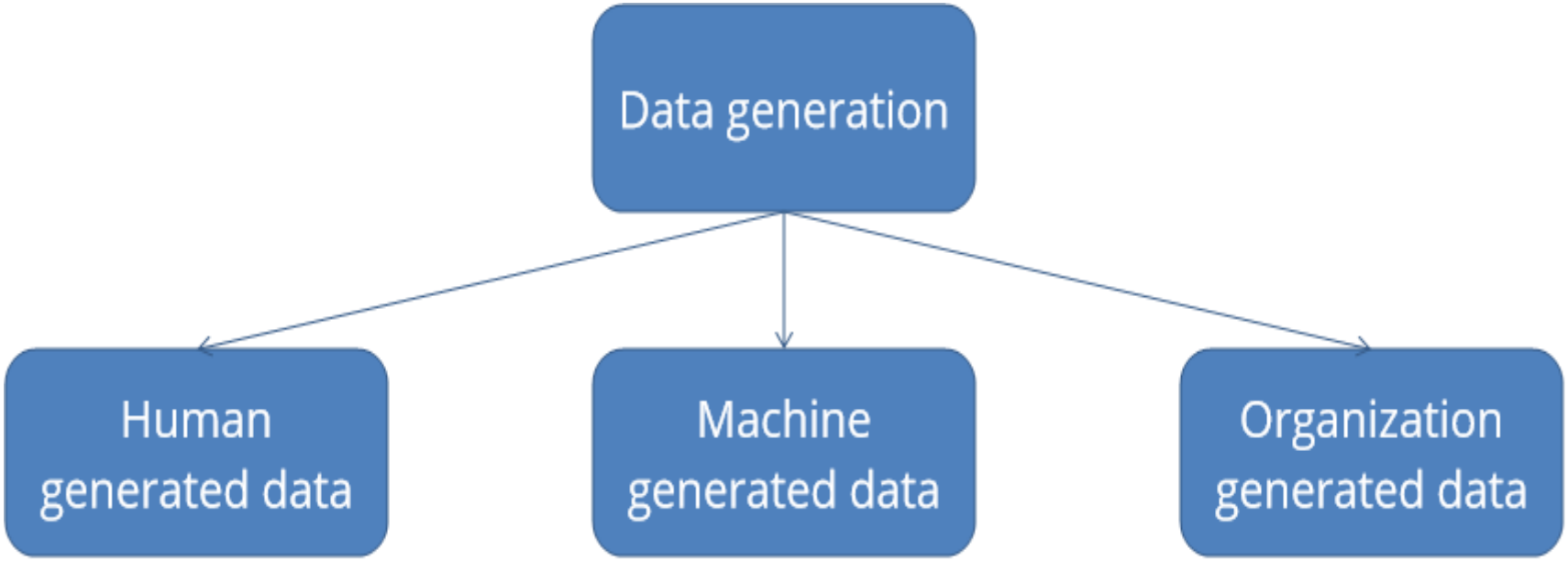
Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master 2	SDIA

What have you done today ?

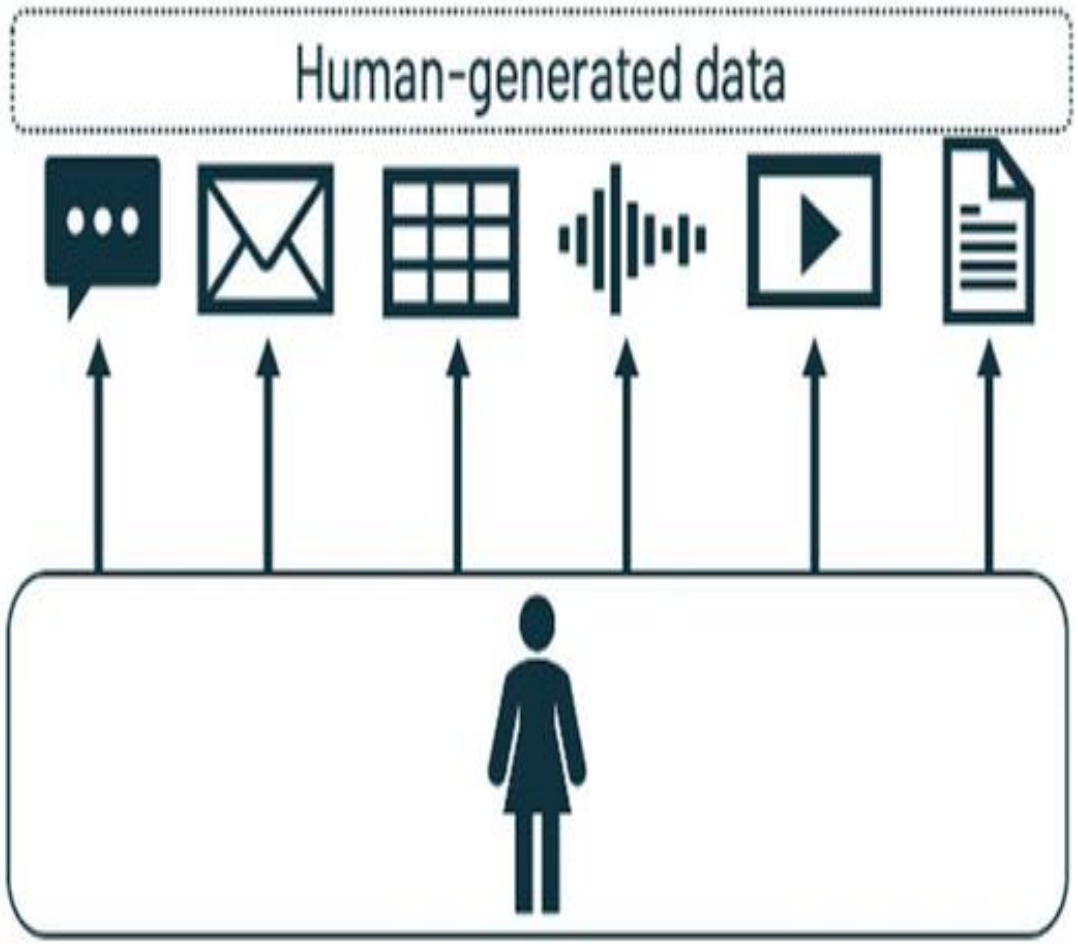
- Did you text someone or post something on a social media ?
- Did you exercise and log your workout using a fitness tracker ?
- Did you log into the e learning platform to take a course ?
- Did you play online games ?
- What else ?

Data generation

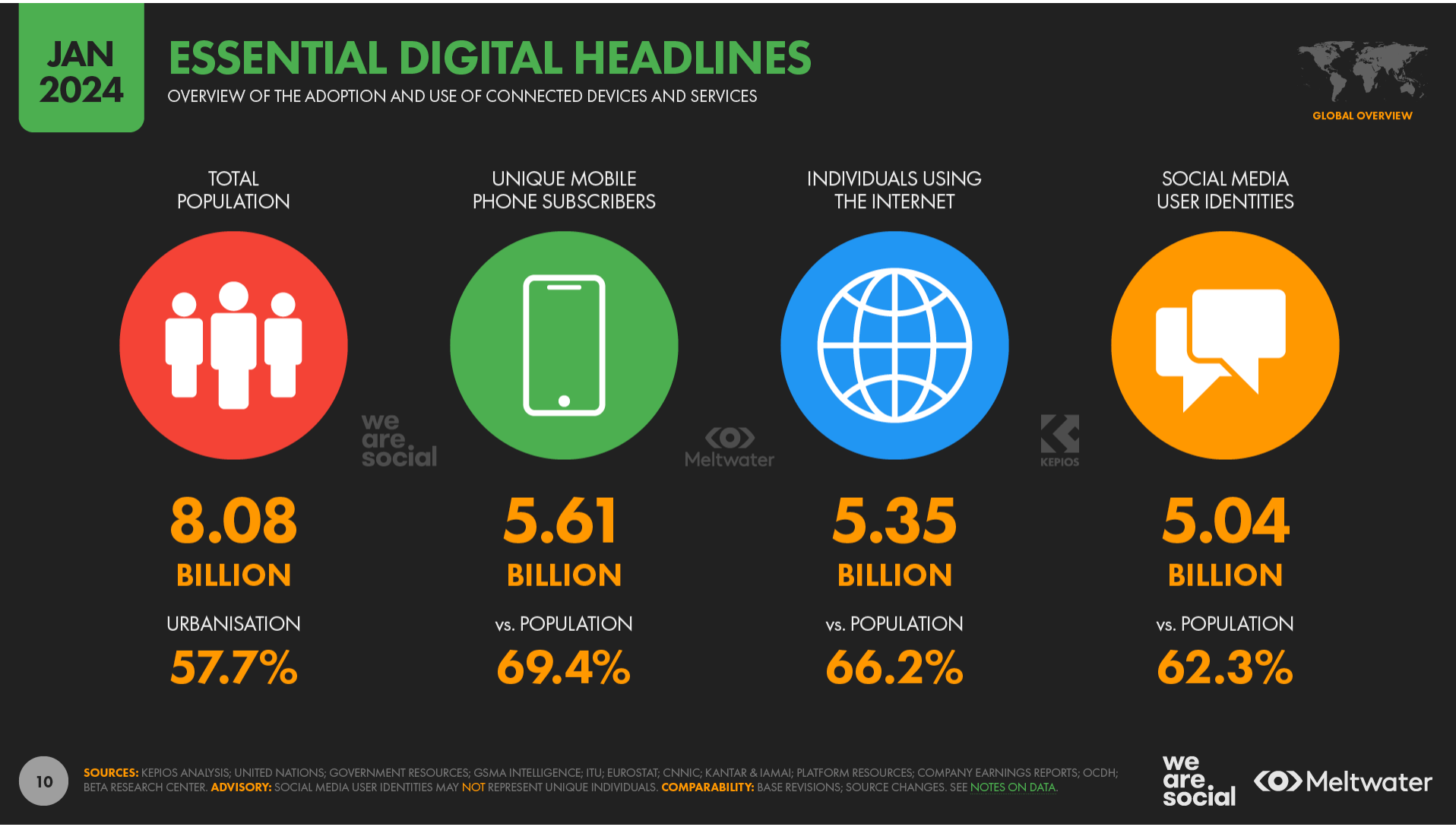


Human-generated Data

- Emails
- Social media posts
- Spreadsheets
- Presentations
- Audio files
- Video files

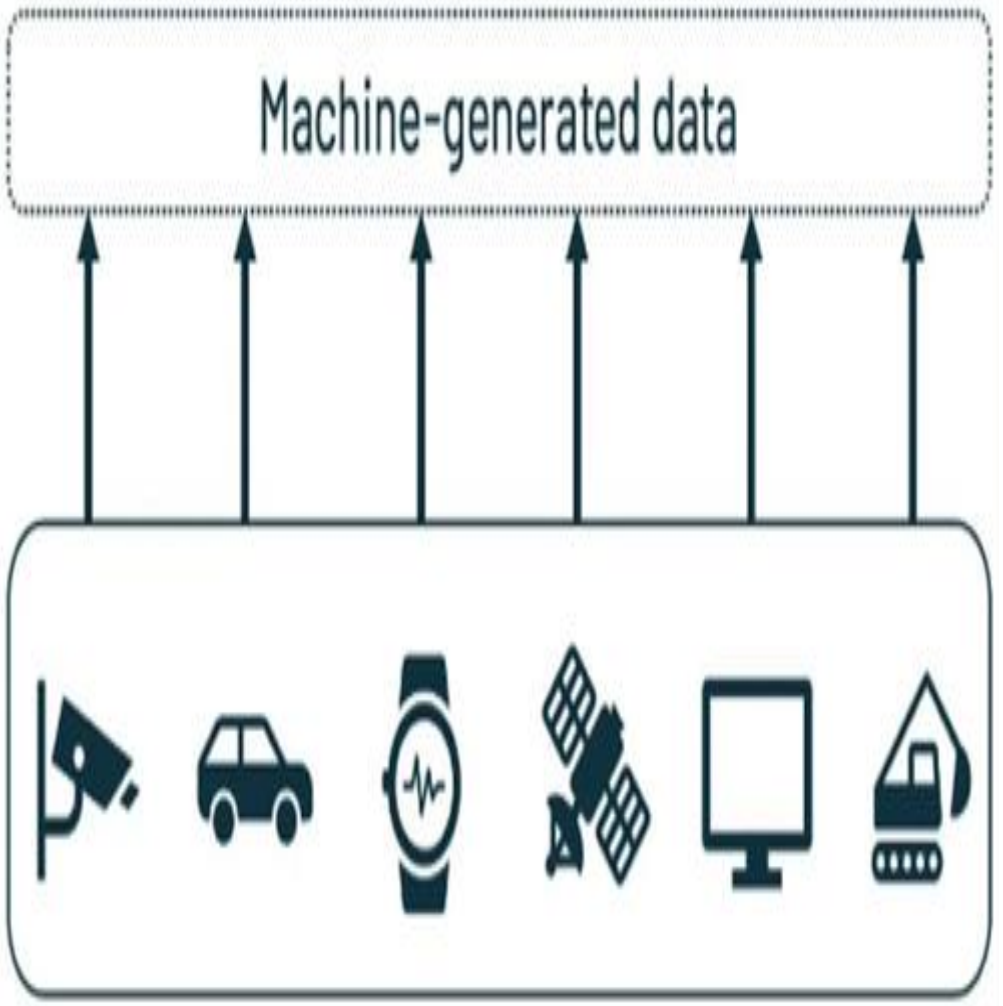


Human-generated data



Machine-generated data

- Sensors on vehicles, appliances and industrial machinery.
- Security cameras.
- Satellites.
- Medical devices.
- Personal tools, such as smartphones and fitness trackers



Machine-generated data

- 150 sensors (Mclaren f1)
- 4 data analysts per car
- One-lap = 2GB of data
- Single race = 3 terabytes of data

● **aUndersteer:** understeer, expressed in degrees (there is none here, 0.0°)

● **xSteerRack:** steering measurement (here the relative positioning on the steering rack: -0.68mm). It's often expressed as a steering angle.

● **rThrottlePedal:** acceleration measurement, expressed in percentage of the pressure applied by the driver on the throttle pedal. 0% means he does not step on it at all, 100% he is flat-out (here 52.4%)

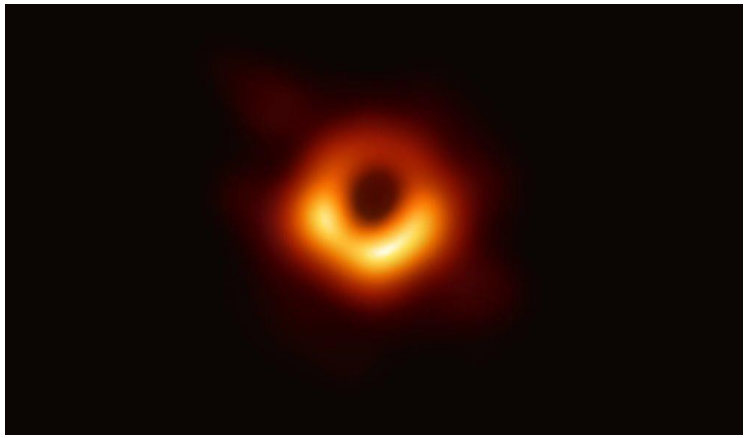
● **NGear:** Gear selected, from first to seventh (here: fourth)

http://www.f1i.com/magazine/73067-f1-telemetry-data-race.html/

A car race ?
Or
A data race?

Machine-generated data

- The amount of data came in at about 5 petabytes (PB)
- 55 million light years
- 5 petabytes to under a 5 megabyte image.

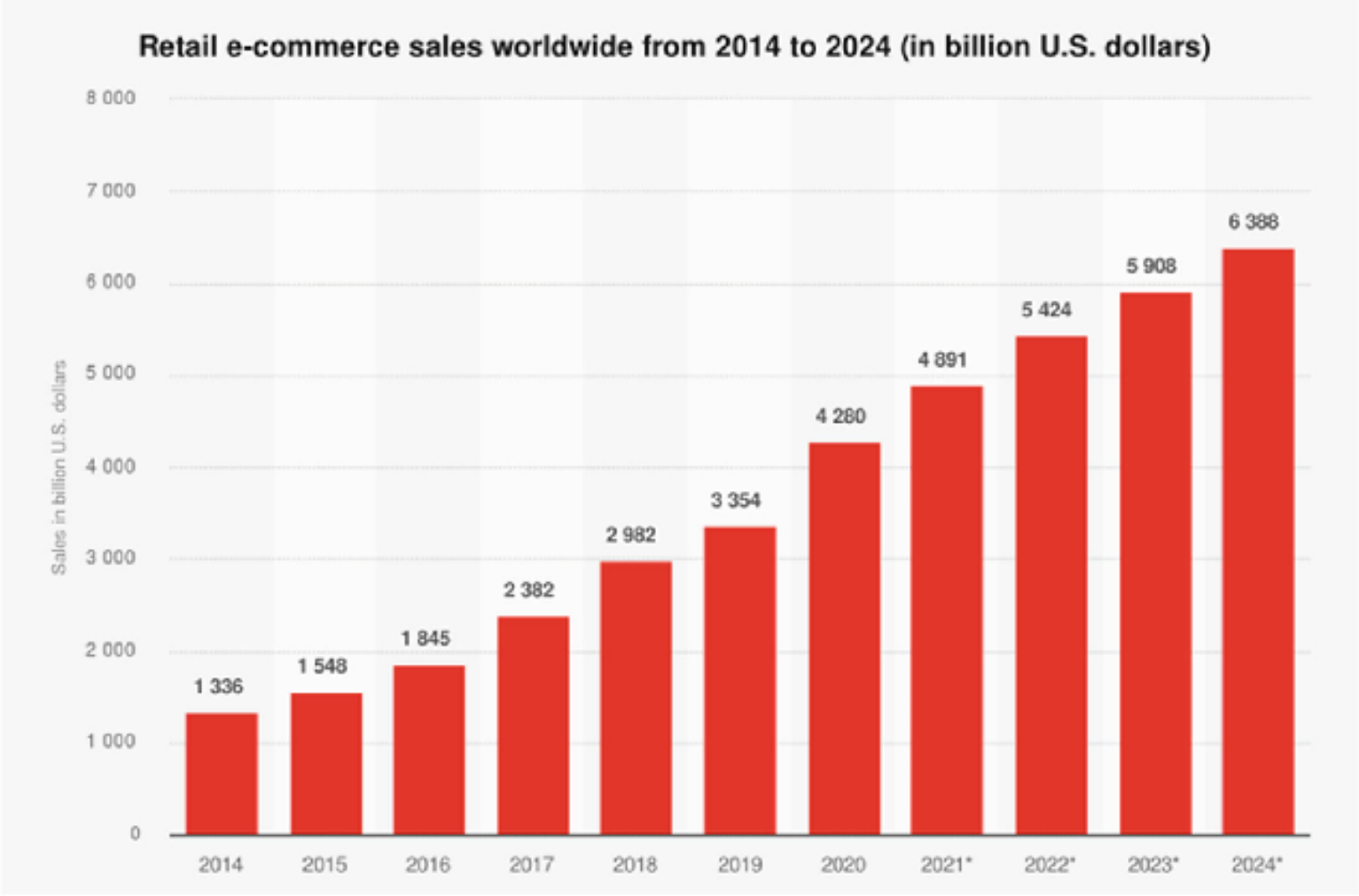


Organization-generated data

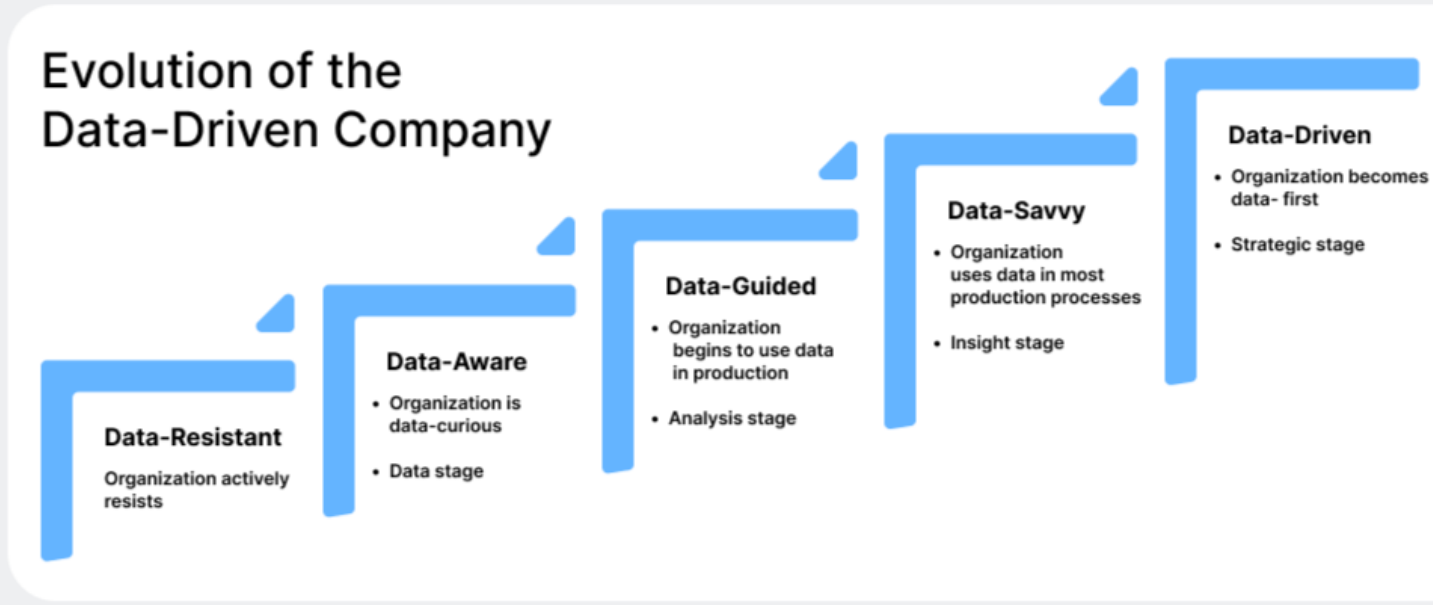
Records generated every time you make a purchase at an online or physical store. Things like unique customer numbers, the items you purchased, the date and time you purchased items and how many of each item you purchased

CustNumer	Item	Date	Time	Quantity
A23498273	Eraser	15/5/2020	4:25:05 AM	2
B34572934	Pencil	24/5/2020	8:17:43 AM	5
B98798172	Pen	12/5/2020	11:39:08 AM	12
C71298748	Marker	23/5/2020	11:55:32 AM	8

Organization-generated data

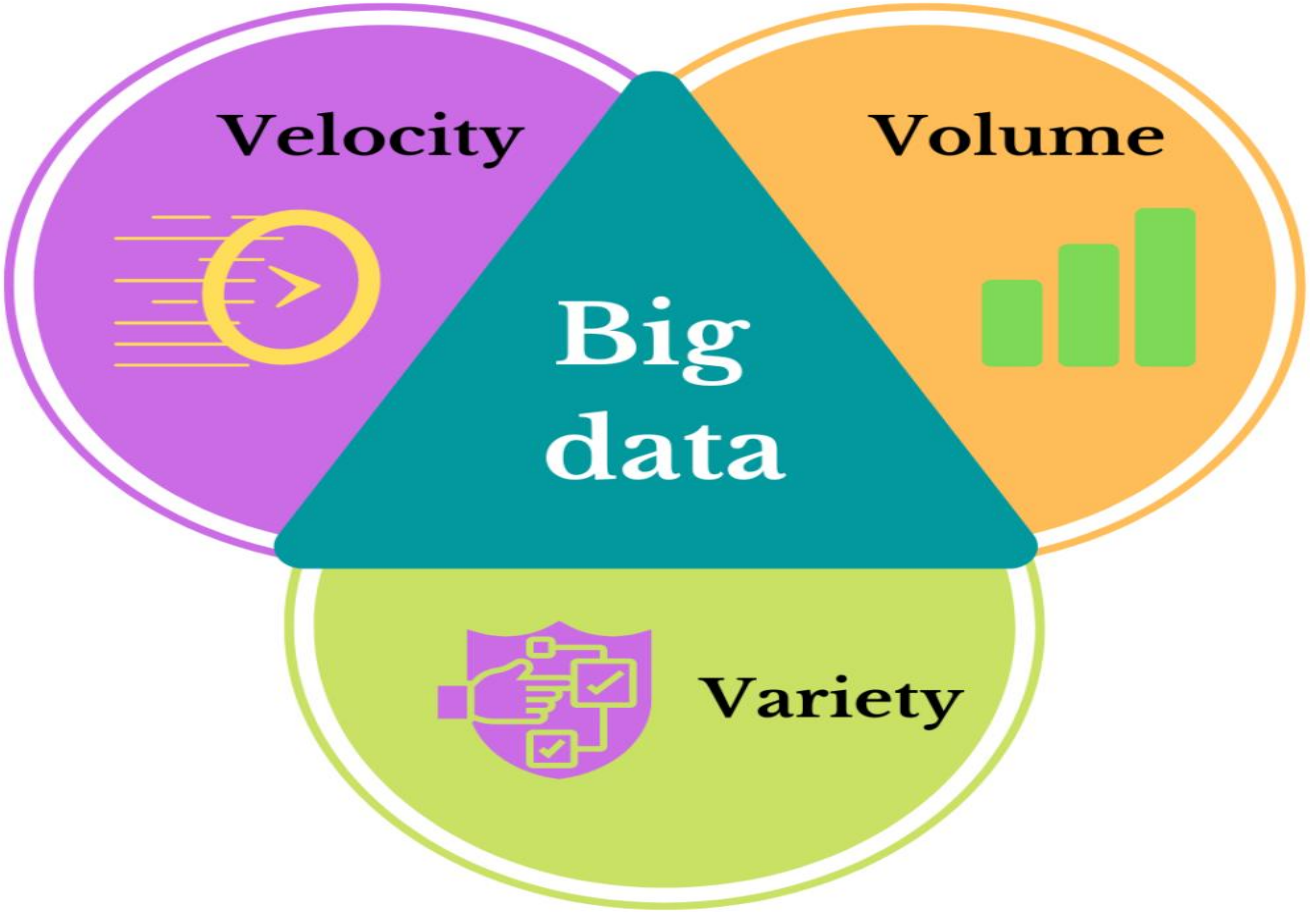


Organization-generated data



What makes big data « Big » ?

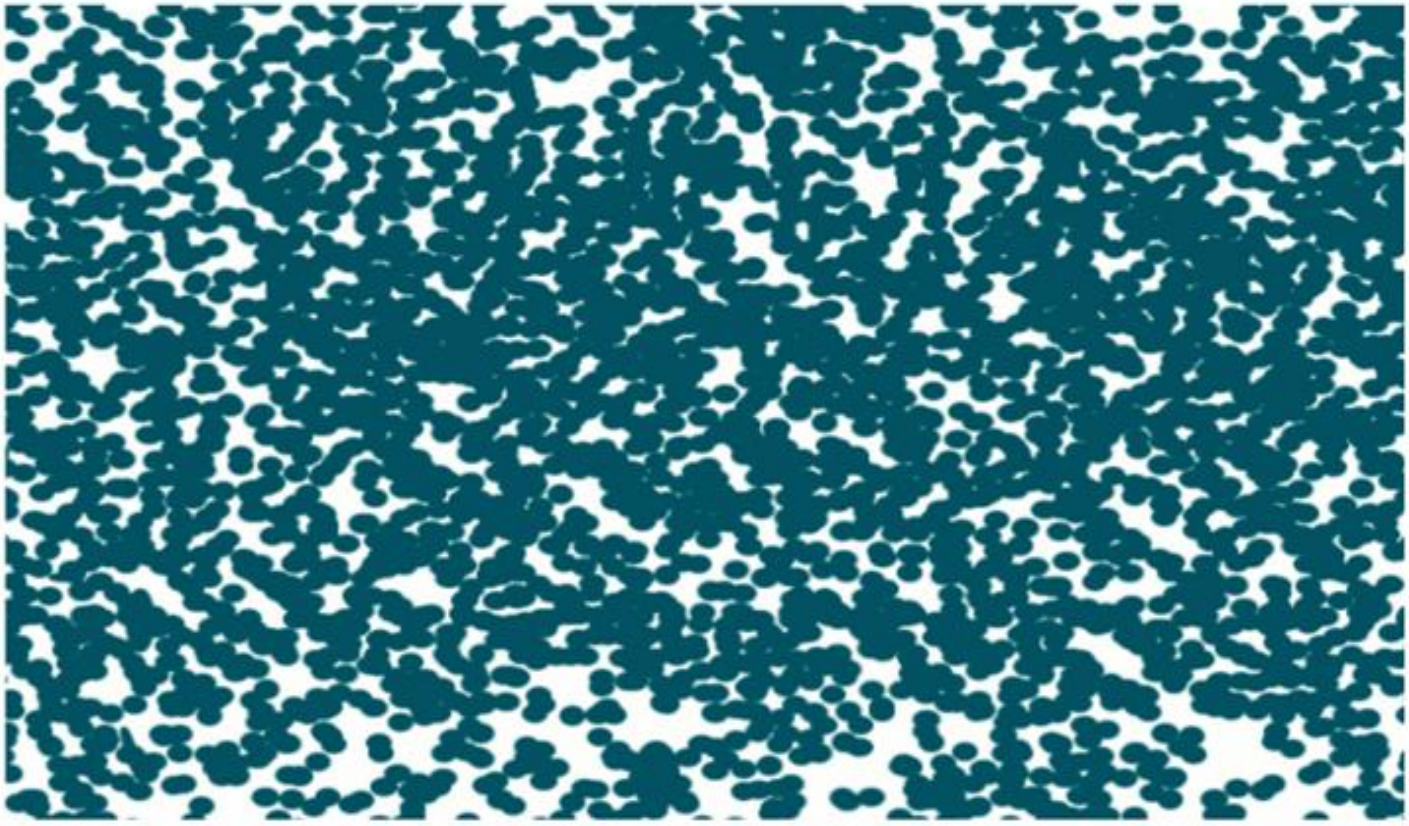
The Three Vs model



Volume

Volume

The amount of data being generated



Where does big data start from?

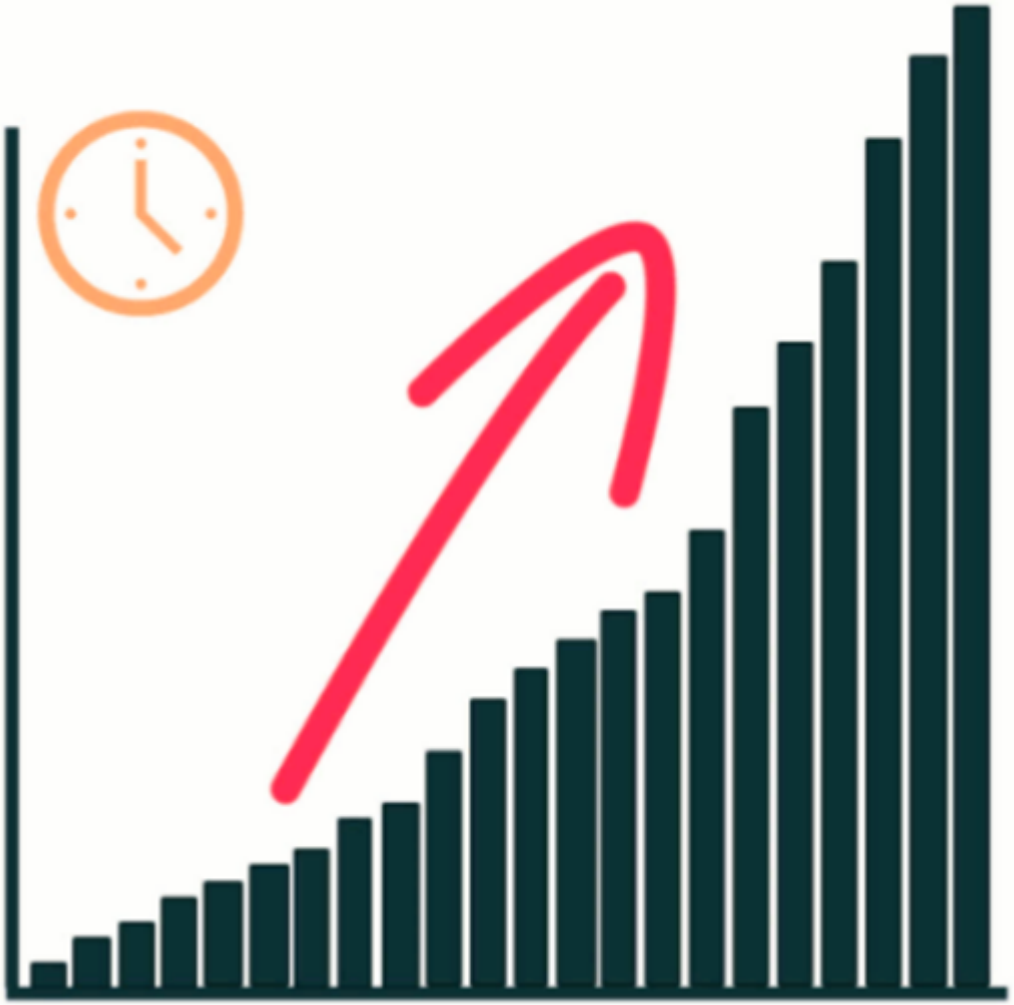
Name	Symbol	Value
Kilobyte	KB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^12
Petabyte	PB	10^15
Exabyte	EB	10^18
Zettabyte	ZB	10^21
Yottabyte	YB	10^24

How big is big data?

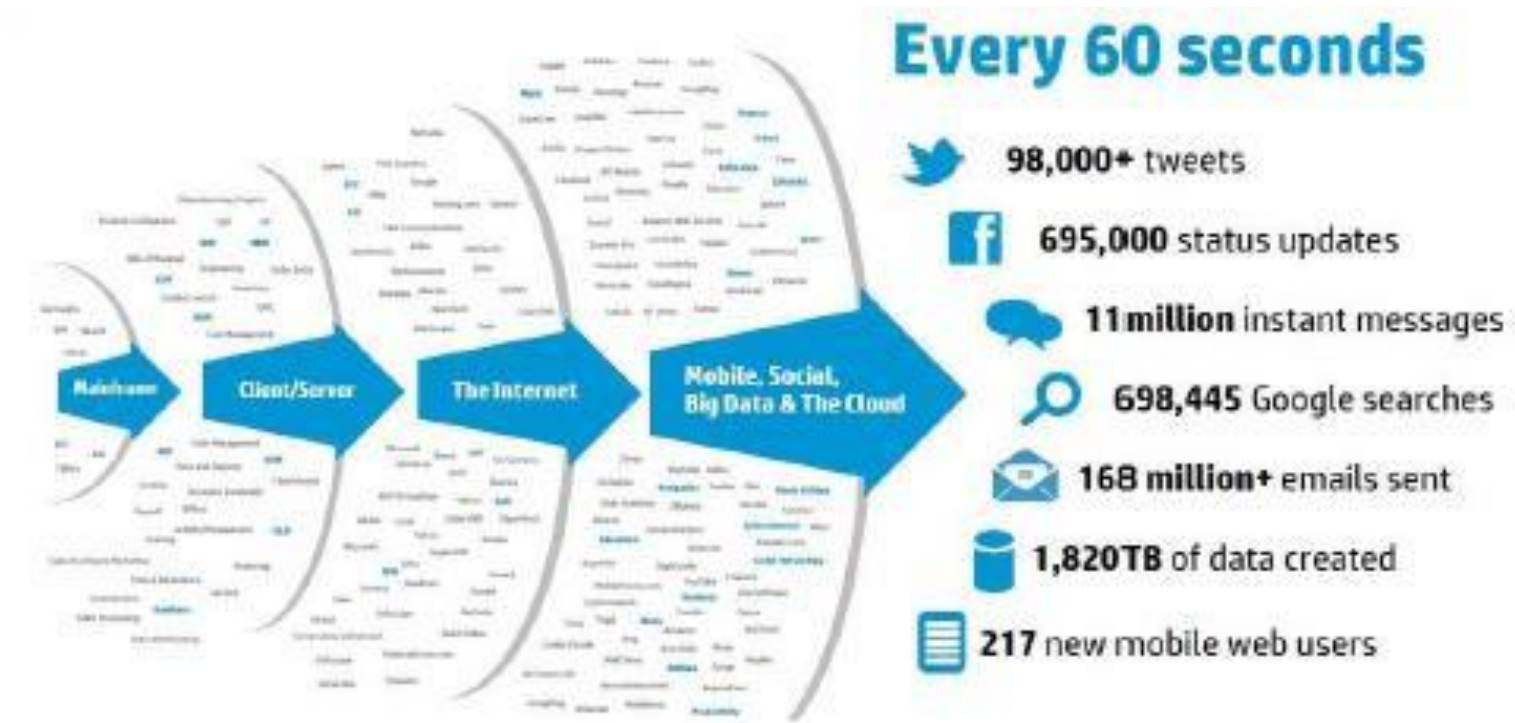
Velocity

Velocity

The speed at which data is generated



What happens every 1 minute



Variety

Variety

The different types of data generated

What are the types of big data ?

Structured data

Structured

Conforms to a schema

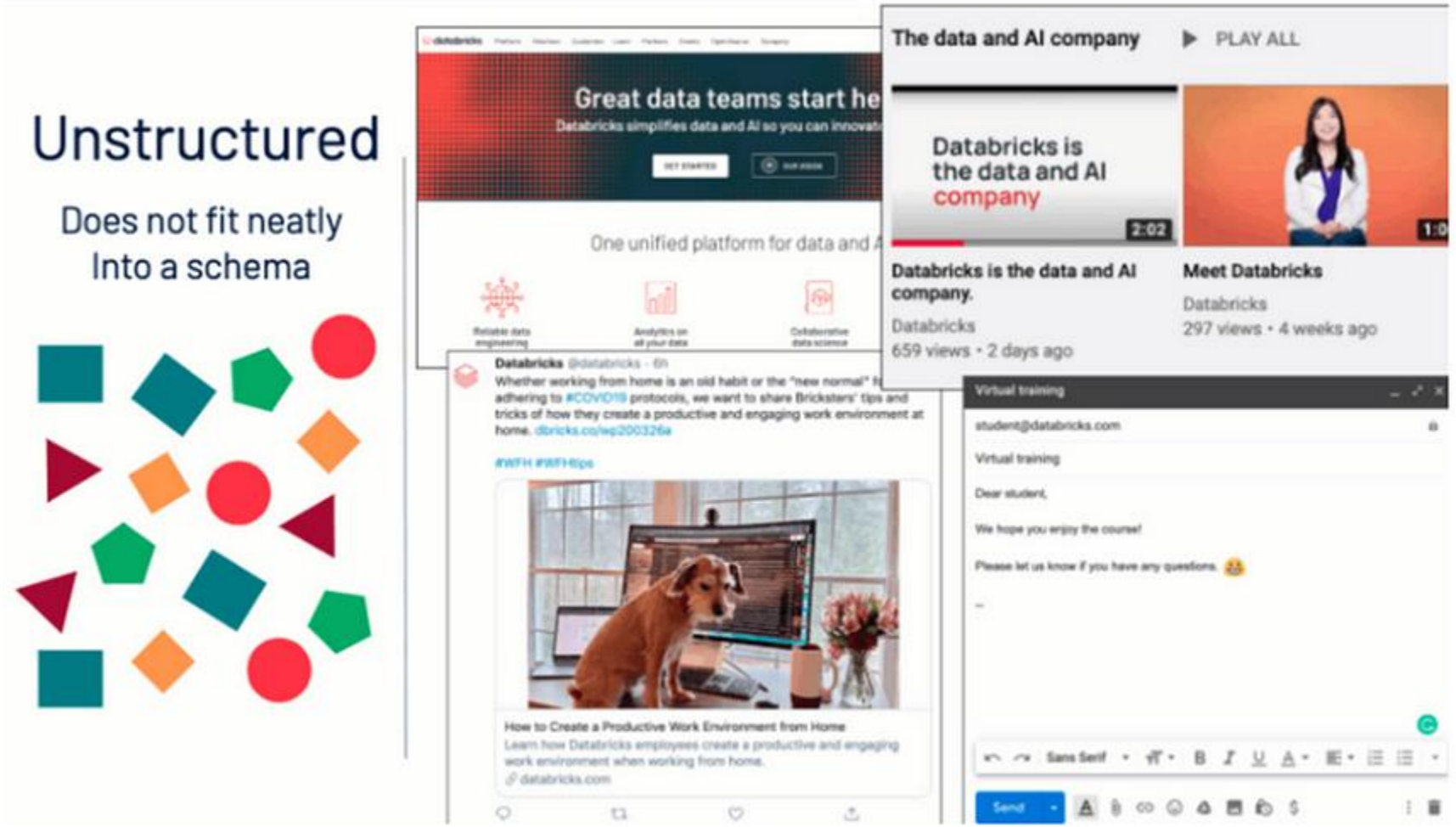
Column

Defined formats

Order	CustID	Month	Item	Color	Price
101	20051	Dec	Pen	Red	2.99
102	20045	Mar	Pencil	Blue|Yellow|Red	3.99
103	29584	May	Eraser	Blue	1.25
104	29584	May	Pen	White	2.25
105	29584	May	Pencil	Blue|Yellow|Red	2.99
106	27485	Jan	Eraser	Blue|Yellow	2.75
107	29574	Jan	Marker	Green	1.75
108	24447	Feb	Marker	Yellow|Blue	7.25
109	26466	Jul	Pen	Black|Red	5.25
110	27467	Jun	Pencil	Black	2.95

Row

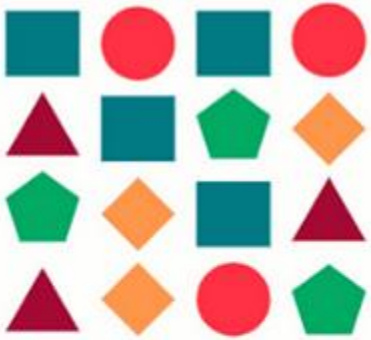
Unstructured data



Semi-structured data

Semi-structured

Some level of organization



Defined way to express data

```
▼<div class="new-main-menu">
  ▼<div class="header-desktop-block">
    ▼<div class="container new-menu">
      ▶<a class="main-logo" rel="home" href="https://
      databricks.com/_" title="Databricks">…</a>
      ▼<div id="new-m" class="menu-bar">
        ▼<div id="mega-menu-wrap-headerNew" class="mega-menu-
        wrap">
          ▶<div class="mega-menu-toggle">…</div>
          ▼<ul id="mega-menu-headerNew" class="mega-menu max-
          mega-menu mega-menu-horizontal" data-event=
          "hover_intent" data-effect="fade_up" data-effect-speed=
          "200" data-effect-mobile="disabled" data-effect-speed-
          mobile="0" data-panel-width="body" data-panel-inner-
          width="#new-m" data-mobile-force-width="false" data-
          second-click="close" data-document-click="collapse"
          data-vertical-behaviour="standard" data-breakpoint=
          "1199" data-unbind="true">
            ▶<li class="mega-main-bar-li mega-menu-item mega-
            menu-item-type-custom mega-menu-item-object-custom
            mega-menu-item-has-children mega-menu-megamenu mega-
```

What is Big data?

Definition of Big Data

Any type of data source that has three or more shared characteristics:

✓ Extremely large data volumes (**Bioinformatics & social network**)

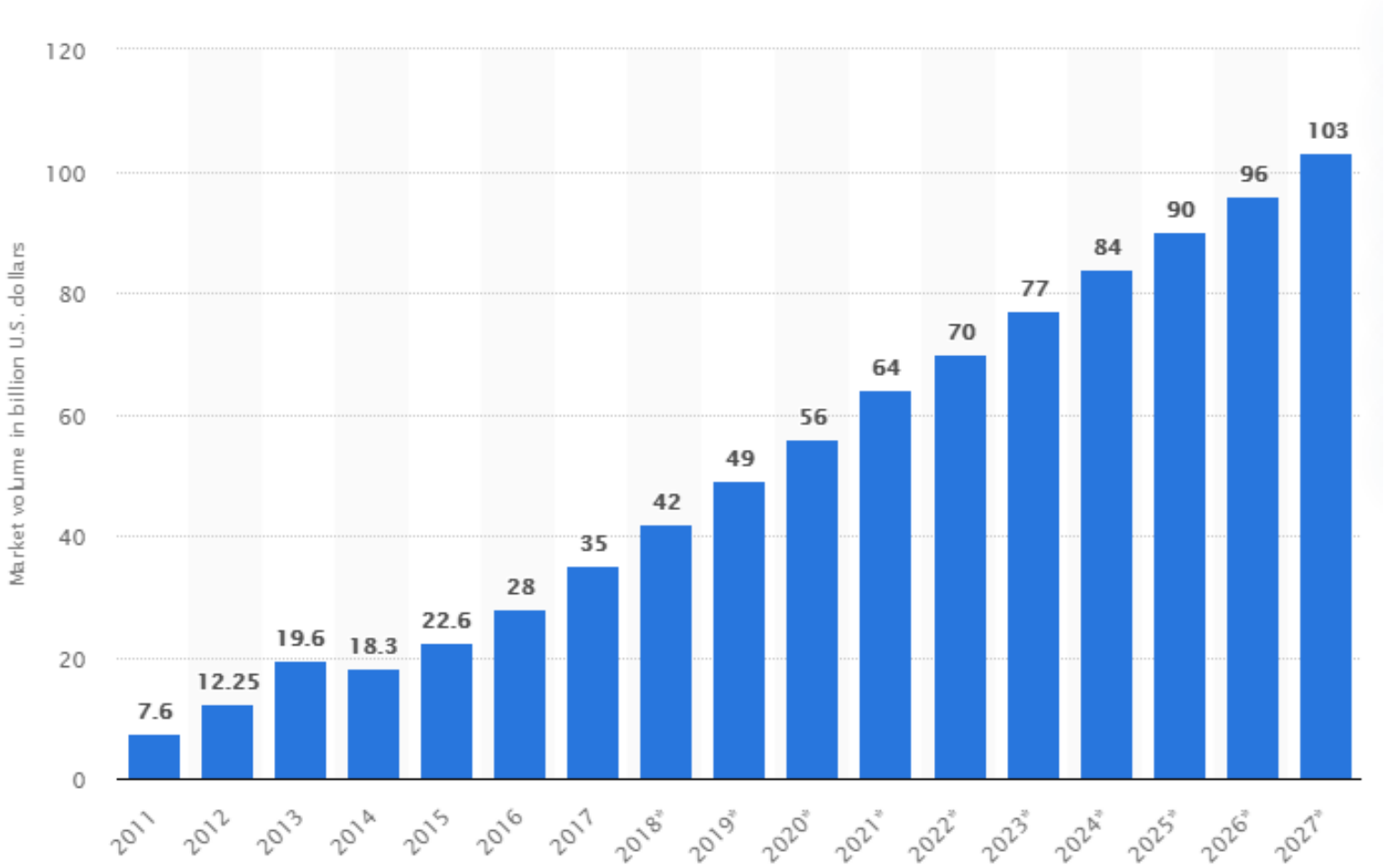
✓ Extremely high data speed (**Telemetry**)

✓ Vast variety of data (**Bioinformatics**)

Big Data is collection of data which you cannot store or process using the traditional database system within the given time frame.

Big data revenues around the world:

Big data market size revenue forecast worldwide from 2011 to 2027 (billion U.S dollars)



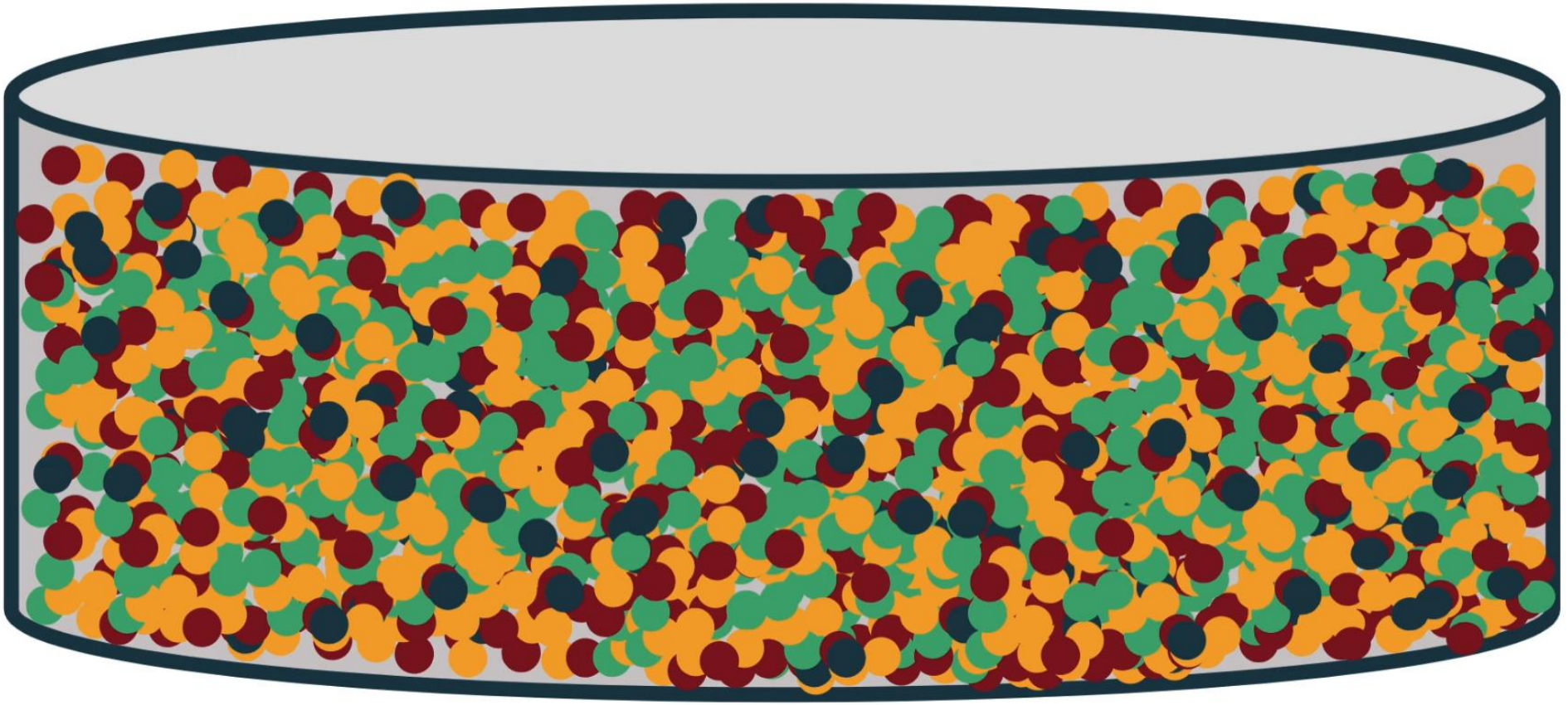
https://www.statista.com/statistics/254266/global-big-data-market-forecast/

Most in-demand skill?

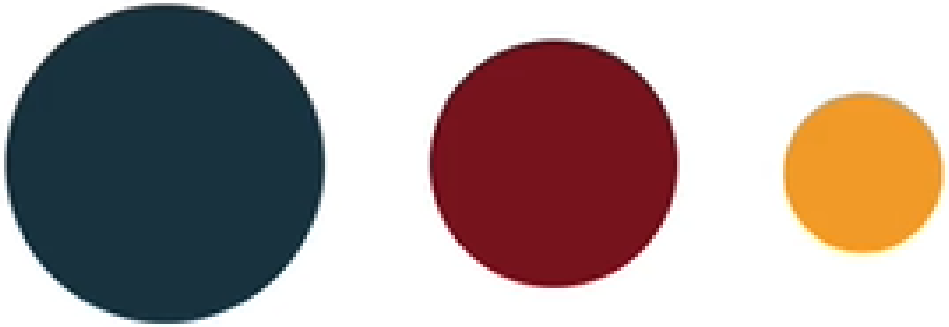
- An estimated 2.7 million data analytics and science jobs are expected in the United States by the end of 2020.

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

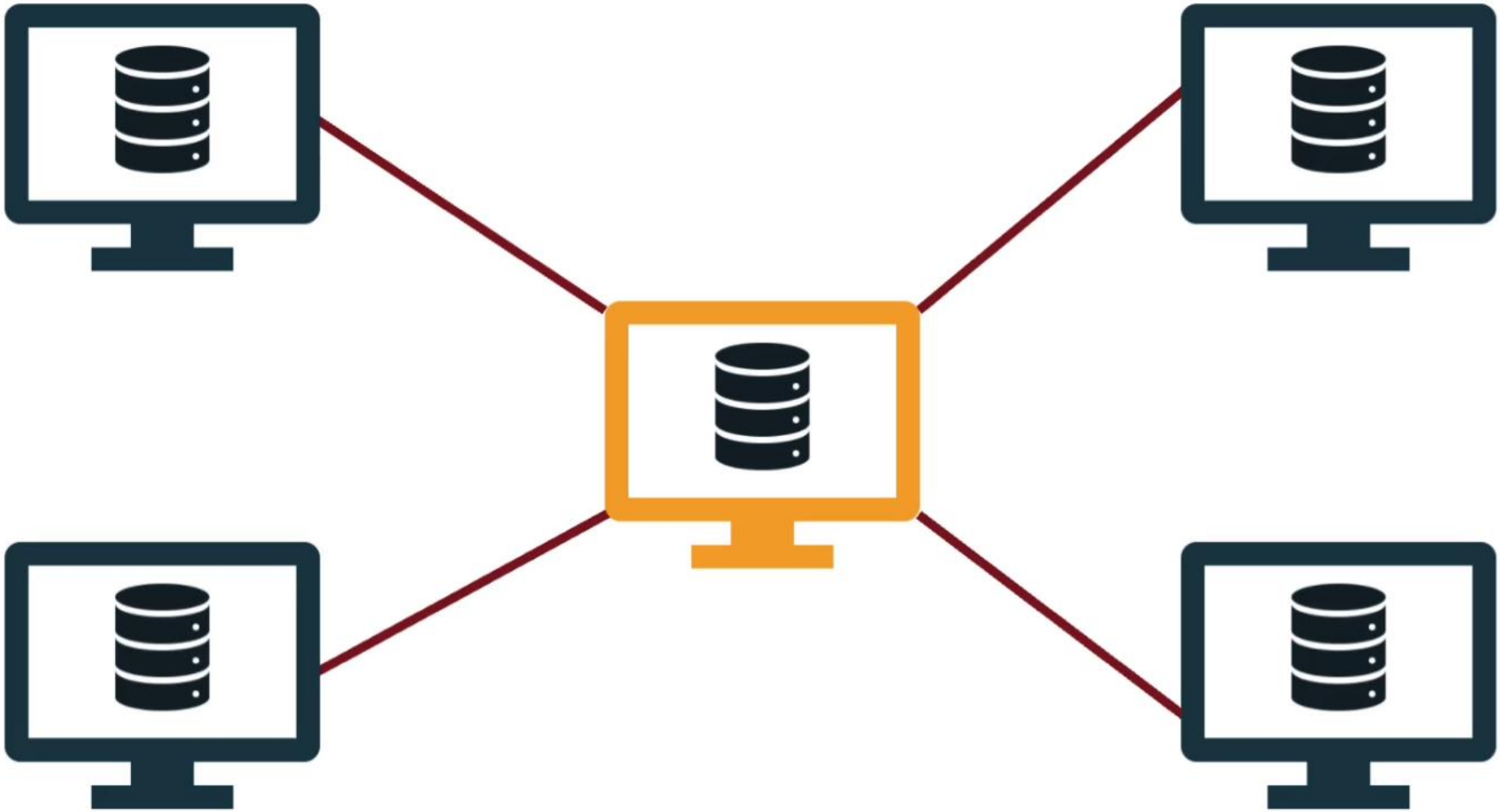
Distributed Computing



- Accurate count !
- Takes a really long time !



- You might not have an accurate count !
- Give you the count relatively quicklly



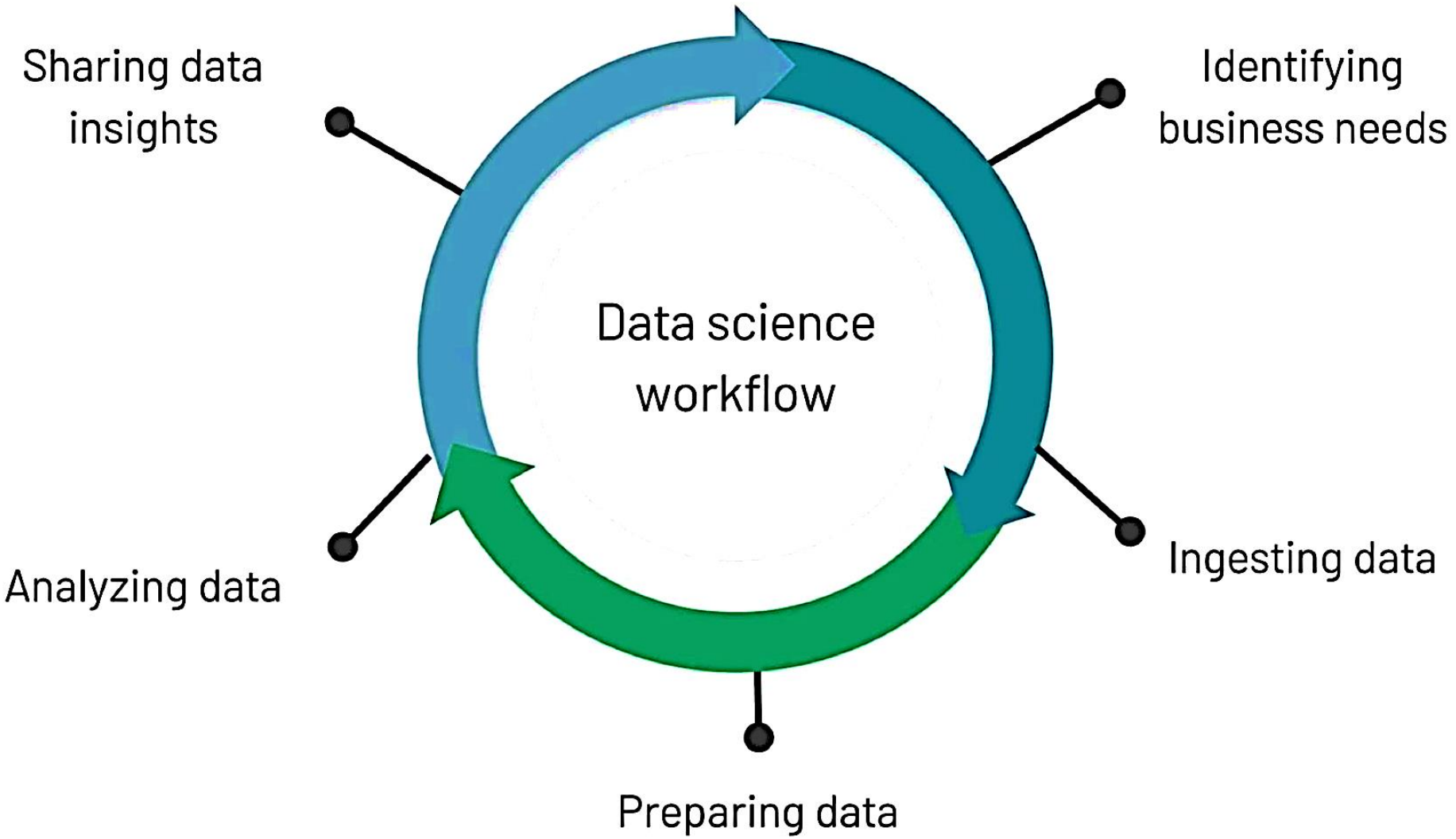
Batch and Streaming data

Batch data

Streaming data

The data science workflow

The data science workflow



Big data analysis for business

Big data analysis for business

Imagine that you work at a bank, and you accidentally transferred 200k DA into the wrong customer's account. All we know about this customer is that he:

- **Is male**
- **Is 20 years old**
- **Is single**
- **Currently resides in Constantine city**
- **Makes 150k da a mounth**

Big data analysis for business

Based on these information, what do you think our customer would do with the 20k DA ?

A. Give the mony back
B. Take the money and run

Big data analysis for business

- **Understand our customers better** – Who are our customers? What do they like? How do they use our products?
- **Improve our products** – Do we need to make changes to our products? What types of changes should we make? What do people like most about our product?
- **Protect our businesses** – Are we investing money in the correct things? Will the risks we take pay off?
- **Stay ahead of the competition** – Who are our biggest competitors? How will we do with upcoming trends in our industry?