What is data?

Clément Levallois

2017-31-17

Table of Contents

last modified: 2017-07-31

Definition of data

The English term "data" (1654) originates from “datum”, a Latin word for "a given". [1: http://www.etymonline.com/index.php?term=data] "Data" is a single factual, a single entity, a single point of matter.

Using the word "data" to mean "transmittable and storable computer information" was first done in 1946. The expression "data processing" was first used in 1954. [2: http://www.etymonline.com/index.php?term=data]

> Thoughts: the etymology suggests that data is "a given". Can you question this?

Data represents either a single entity, or a collection of such entitities ("data points"). We can speak also of datasets instead of data (so a dataset is a collection of data points).

Examples!

A date	A color	A grade
A relation of friendship	A sound	A hearbeat
A user input	A duration	A curriculum vitae
A picture	A longitude and latitude	A price
A number of friends	A temperature	A list of favorite movies
etc...	etc...	etc...

3 take aways from the examples

1. Think about data in a broad sense

Data is not just text and figures. You should train in thinking about data in a broader sense:

- pictures are data
- language is data (including slang, lip movements, etc.)
- relations are data (you know individual A, you know individual B, but the relationship between A and B is data as well)
- preferences, emotional states… are data
- etc. There is no definitive list, you should train yourself looking at buisness situations and think: "where is the data?"

2. metadata is data, too

Metadata: this is some data describing some other data.

Example:

```
The bibliographical reference ①
describing
a book ②
```

① the metadata

② the data

- Data without metadata can be worthless (imagine a library without a library catalogue)
- Metadata can be informative in its own right [3: http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata]

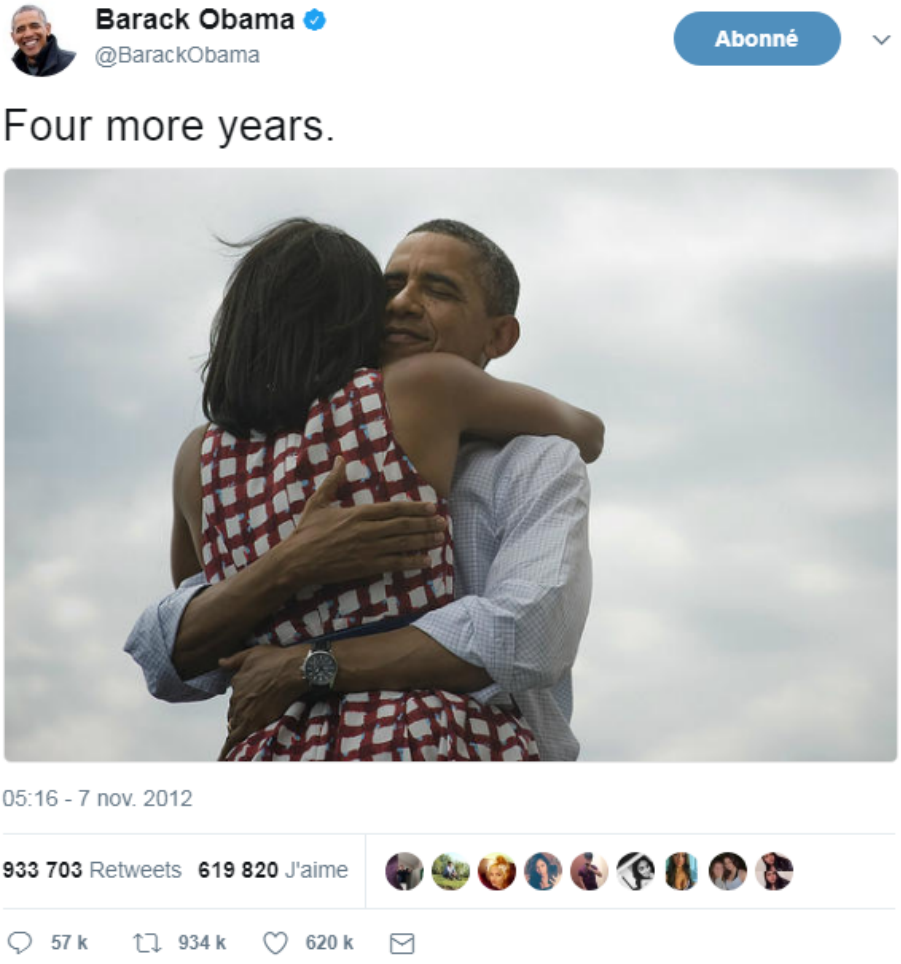
3. zoom in, zoom out

We should remember considering that a data point can be itself a collection of data points:

- a person walking into a building is a data point.
- however this person is itself a collection of data points: location data + network relations + subscriber status to services + etc.

So it is a good habit to wonder whether a data point can in fact be "unbundled" (spread into smaller data points / measurements)

Some essential vocabulary to discuss data



- This is a digital **medium** (because it's on screen as opposed to analogic, if we had printed the pic on paper)
- The **type** of the data is textual + image
- The text is **formatted** in plain text (meaning, no special formatting), as opposed to more structured data-interchange formats (check json or xml).
- The **encoding** of the text is UTF-8. Encoding has to do with the issue: how to represent alphabets and signs from different languages in text? (not even mentioning emojis?). UTF-8 is an encoding which is one of the most universal.
- The tweet is part of a list of tweets. The list represents the **data structure** of my dataset, it is the way my data is organized. There are many alternative data structures: arrays, sets, dics, maps...
- The tweet is stored as a picture (png file) on my hard disk. "png" is the **file format**. The data is **persisted** as a file on disk (could have been stored in a database instead).

Data presented as a table

[table]

Finally: data and size

1 bit		can store a binary value (yes / no, true / false…)
8 bits	1 byte (or octet)	can store a single character
~ 1,000 bytes	1 kilobyte (kb)	Can store a paragraph of text
~ 1 million bytes	1 megabyte (Mb)	Can store a low res picture.
~ 1 billion bytes	1 gigabyte (Gb)	Can store a movie
~ 1 trillion bytes	1 terabyte (Tb)	Can store 1,000 movies. Size of commercial hard drives in 2017 is 2 Tb.
~ 1,000 trillion bytes	1 petabyte (Pb)	20 Pb = Google Maps in 2013 == The end Find references for this lesson, and other lessons, here.