

What is data?

Clément Levallois

2017-31-07

Table of Contents

1. Definition of data	1
2. Examples!	1
3. Three take aways from the examples	1
a. Think about data in a broad sense	1
b. metadata is data, too	2
c. zoom in, zoom out	2
4. Some essential vocabulary to discuss data	3
a. Formats, types, encoding	3
b. Data presented as a table	4
c. Data according to who owns it	4
d. Data: "sociodemo" or "behavior"?	4
5. Finally: data and size	5
The end	5



1. 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 entities ("data points"). We can speak also of datasets instead of data (so a dataset is a collection of data points).

2. Examples!

A date	A color	A grade
A relation of friendship	A sound	A heartbeat
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. Three take aways from the examples

a. 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 business situations and think: "where is the data?"

b. 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, as shown with the NSA scandal: [3: <http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata>]

NEWS DESK

WHAT'S THE MATTER WITH METADATA?



By Jane Mayer June 6, 2013

c. zoom in, zoom out

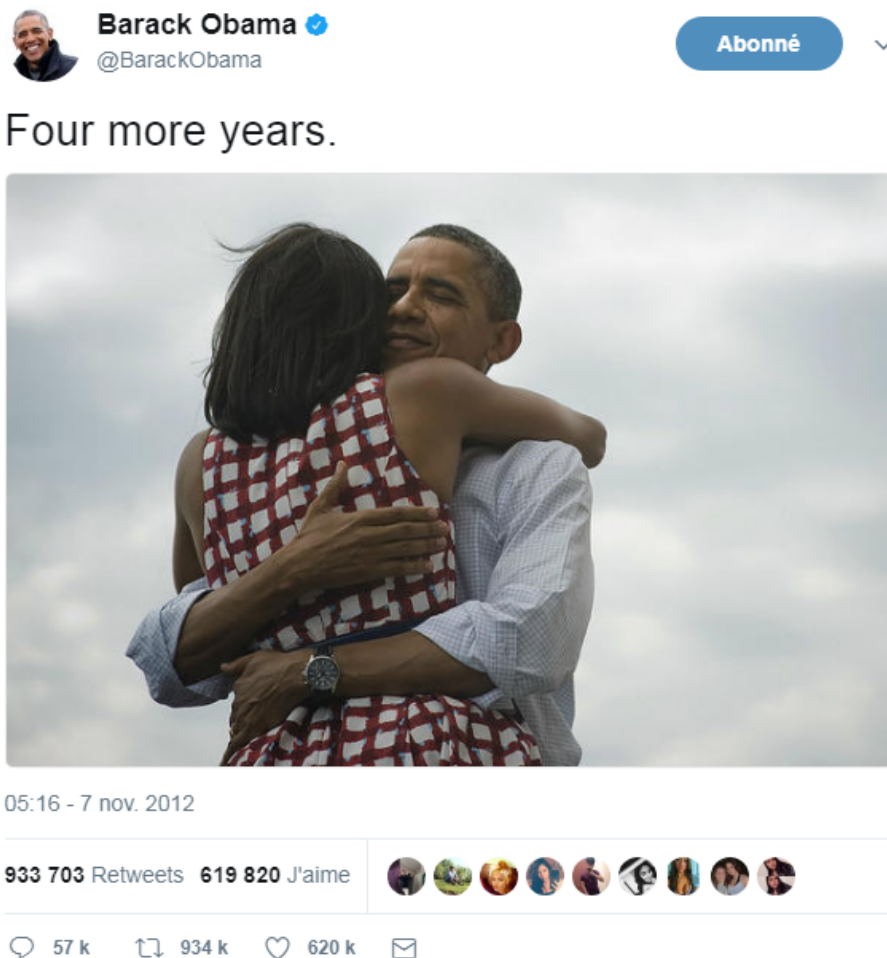
We should remember considering that a data point can be itself a collection of data points:

- 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)

4. Some essential vocabulary to discuss data

a. Formats, types, encoding



- 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).

b. Data presented as a table

A spreadsheet, or a **table**. This is still the most common way to represent a dataset.

Columns. Each represents an **attribute** of the data.

Header: these are the names of the attributes.

Rows, or lines. Each represents a data point

A value. (can be empty).

	A	B	C	D	E	F	G
1	id	civilete	particule	first name	name	maiden name	year of birth
2	10997	M		William	Pruitt		unknown
3	10998	F		Marian	Oconnor		unknown
4	10999	M		Sammie	Robertson		unknown
5	22529	M		Efren	Smith		1970
6	22528	M		Nigel	Simon		unknown
7	22527	M		Bruce	Bowers		unknown
8	22526	M		Chester	Hicks		1987
9	22525	M		Bernardo	Lott		unknown
10	22524	F		Elisabeth	Nash		unknown
11	22523	M		Kristopher	Stanton		unknown
12	10990	M		Dennis	Sparks		1989
13	22522	M		Sean	Ewing		1950
14	10991	M		Cedrick	Hoffman		1983

c. Data according to who owns it

- First party data: the data generated through the activities of your own organization. Your organization own it, which does not mean that consent from users is not required, when it comes to personal data.
- Second party data: the data accessed through partnerships. Without being the generator nor the owner of this data, partners make it available to you through an agreement.
- Third party data: the data acquired via purchase This data is acquired through a market transaction. Its uses still comes with conditions, especially for personal data.

d. Data: "sociodemo" or "behavior"?

- Sociodemographic or "sociodemo" data refers to information about individuals, describing fundamental attributes of their social identity: age, gender, place of residence, occupation, marital status and number of kids.
- Behavior data refers to any digital trace left by the individual in the course of it life: clicks on web pages, likes on Facebook, purchase transactions, comments posted on Tripadvisor...

Sociodemo data is typically well structured or easy to structure. It has a long history of collection and analysis, basically since census exists.

Behavior data allows to go further than sociodemo data: each individual can be characterized by its acts and tastes, well beyond what an age or marital status could define.

But behavior data is typically not well structured and harder to collect.

5. 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](#).



This course is made by Clement Levallois.

Discover my other courses in data / tech for business: <http://www.clementlevallois.net>

Or get in touch via Twitter: [@seinecle](#)