DATA ANALYTICS

DATA, OBJECTS, ATTRIBUTES & MODES

WHAT IS DATA?

It is difficult to give a clear-cut definition of **data** (is it singular or plural?).

Linguistically, a *datum* is “a piece of information”; **data** means “pieces of information,” or a **collection** of “pieces of information”.

Data represents the whole (greater than the sum of its parts?) or simply the idealized concept.

Is that clear?

WHAT IS DATA?

Is the following data?

4,529 red 25.782 Y

Why? Why not? What, if anything is missing?

The Stewart approach: “we know it when we see it.”

Pragmatically, we think of data as a collection of facts about **objects** and their **attributes**.

OBJECTS AND ATTRIBUTES

Object: *apple*

- **Shape:** spherical
- **Colour:** red
- **Function:** food
- **Location:** fridge
- **Owner:** Jen

Object: *sandwich*

- **Shape:** rectangle
- **Colour:** brown
- **Function:** food
- **Location:** office
- **Owner:** Pat

Remember: an object is not simply **the sum of its attributes**.

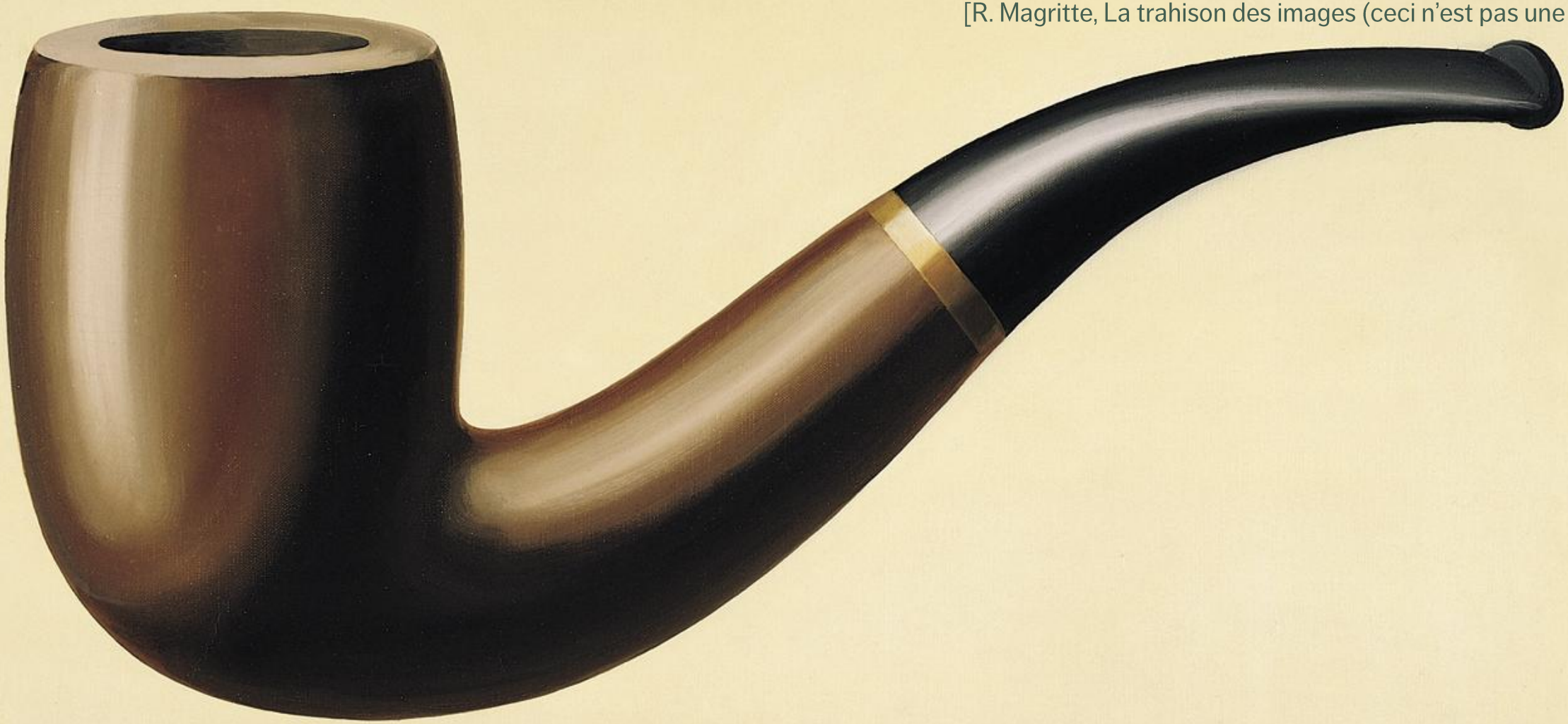
OBJECTS AND ATTRIBUTES

Ambiguities when it comes to **measuring** (and **recording**) the attributes:

- apple picture is a 2-dimensional representation of a 3-dimensional object
- overall shape of the sandwich is vaguely rectangular, it is not exact (**measurement error**?)
- insignificant for most, but not necessarily all, analytical purposes
- apple's shape = volume, sandwich's shape = area (**incompatible measurements)**
- a number of potential attributes are not mentioned: size, weight, time, etc.
- are there other issues?

Measurement errors and incomplete lists are always part of the picture; is this collection of attributes providing a reasonable **description** of the objects?

[R. Magritte, La trahison des images (ceci n'est pas une pipe)]



Ceci n'est pas une pipe.

FROM OBJECTS AND ATTRIBUTES TO DATASETS

Raw data may exist in any format.

A **dataset** represents a collection of data that could conceivably be fed into algorithms for analytical purposes.

Datasets appear in a **table** format, with rows and columns; attributes are the **fields** (or columns, variables); objects are **instances** (or cases, rows, records).

Objects are described by their **feature vector** (observation's signature) – the collection of attributes associated with value(s) of interest.

FROM OBJECTS AND ATTRIBUTES TO DATASETS

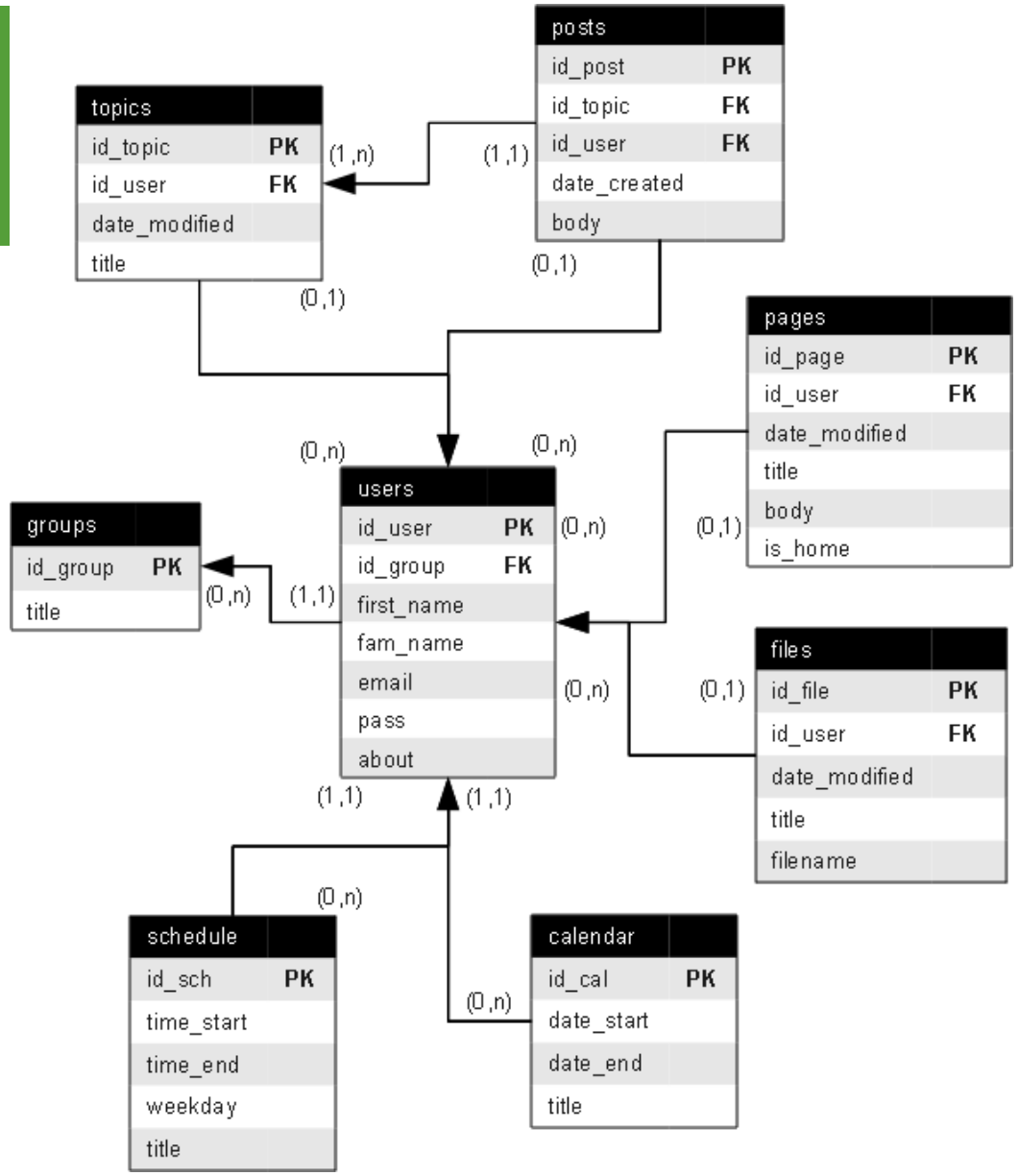
The dataset of physical objects could start with:

ID	shape	colour	function	location	owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	school
...	...	...	...	...	...

[Wooptoo, "Entity - relationship model." Wikimedia]

FROM OBJECTS AND ATTRIBUTES TO DATA

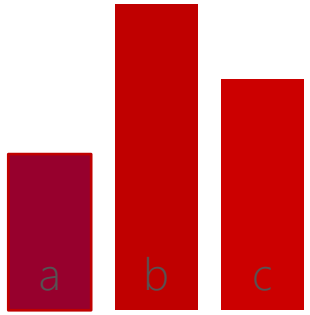
In practice, more complex **databases** are used, for a variety of reasons that we briefly discuss at a later stage.



ANALYTICS MODES

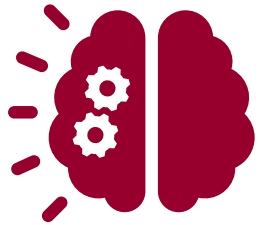
Looking Backwards

Descriptive



Show **what** happened

Diagnostic



Explain **why** something happened

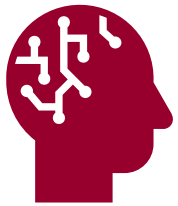
Easy

Difficult

Predictive

Guess **what will** happen

Prescriptive



Suggest **what should** happen

Looking Ahead

DATA ANALYTICS

DATA, OBJECTS, ATTRIBUTES & MODES