# Examine different types of data

4 minutes

Data is just another word for collected information. There are lots of kinds of information out there, whether that be information about people on the Titanic, or your hairdresser's favorite color.

There are also lots of ways we can categorize data. To work in machine learning we need to understand the kind of data, and how it's being stored digitally.

## Continuous, ordinal, and categorical data

When we work with data, sometimes we need to be aware of what kind of thing it represents. This awareness can help us to pick the right machine learning model or organize our data in particular ways.

*Continuous data* are numbers that can be increased or decreased by any amount. For example, you can add 1 mm to 1 m resulting in 1.001 m.

*Categorical data* are data that don't fall on a spectrum. In our scenario, people on the Titanic were categorized as "staff" or "passengers". Categorical data can't be stored as numbers in an obvious way.

*Ordinal data* are categorical data that have an order, and so can be stored as numbers. For example, big, medium, small are ordinal data because they can be ranked like so: big > medium > small. By contrast, 'apple', 'orange' and 'coconut' are categorical because they can't be ranked. Ordinal data can also refer to numbers that can be increased or decreased, but only by set amounts. For instance, the number of people boarding a boat is guaranteed to be a whole number: no-one can half board.

*IDs* are a special kind of categorical data where each sample has its own ID. For example, in our dataset, each person on the Titanic has their own ID, even if they have the same name as someone else. Identities are useful for helping us find our way around a dataset, but they aren't data we analyze per-se.

## Datatypes

All data we use for machine learning must be able to be stored and processed by a computer. While we can put almost any data we want on a piece of paper with a pencil, computers store

information as series of 0's and 1's. This means how we use information is much more restricted.

Datatype, refers to the type of data that is stored on a computer. The common basic datatypes are:

- integers: counting numbers, like 2
- floating-point numbers: numbers with decimal places, like 2.43
- strings: letters and words
- booleans: true and false
- None, void, or null: not data, but rather the absence of it

The exact terms and implementation for these concepts varies from language to language, but the basic way they work is much the same.

It’s notable that in some circumstances two different datatypes offer equivalent functionality. For example, true/false values can often be encoded as Booleans (true or false), strings (‘y’, or ‘n’), integers (0 or 1) or even floating-point numbers (0.0 or 1.0).

# Derived datatypes

By now you’ve seen there are datatypes that are more impressive than these! Computers can store dates, images, 3D models, and so on. These are called derived datatypes and are constructed by one or more primitive types.

Often, in machine learning, it's helpful to convert derived types into simpler representations. For example, a date—1st January 2017—can be stored as an integer or floating point number: 20170101. This makes the mathematics underneath our models easier to calculate.

# Too many choices?

Knowing the kind of data you have can help you pick the right datatype.

The correct datatype can depend on the package you're using to run your models, though generally packages are quite permissive. Generally:

- If you're working with continuous data, floating point numbers are best.
- Ordinal data are typically encoded as integers.
- Categorical data for only two categories can usually be encoded as Boolean or integer data. Working with three or more categories can be slightly more complicated. Not to worry – this is something we'll cover in the next lesson!

In the next exercise, we'll practice visualizing our data in order to understand it well. When we do, take special note of the datatypes that are being processed, and try to identify which of these are continuous, ordinal, or categorical.

## Next unit: Exercise - Work with data to predict missing values

Continue >