Data Types

One of the key parts of the design process concerns understanding the different types of data (sometimes known as *levels of data* or *scales of measurement*). Defining the types of data will have a huge influence on so many aspects of this workflow, such as determining:

- the type of exploratory data analysis you can undertake;
- the editorial thinking you establish;
- the specific chart types you might use;
- the colour choices and layout decisions around composition.

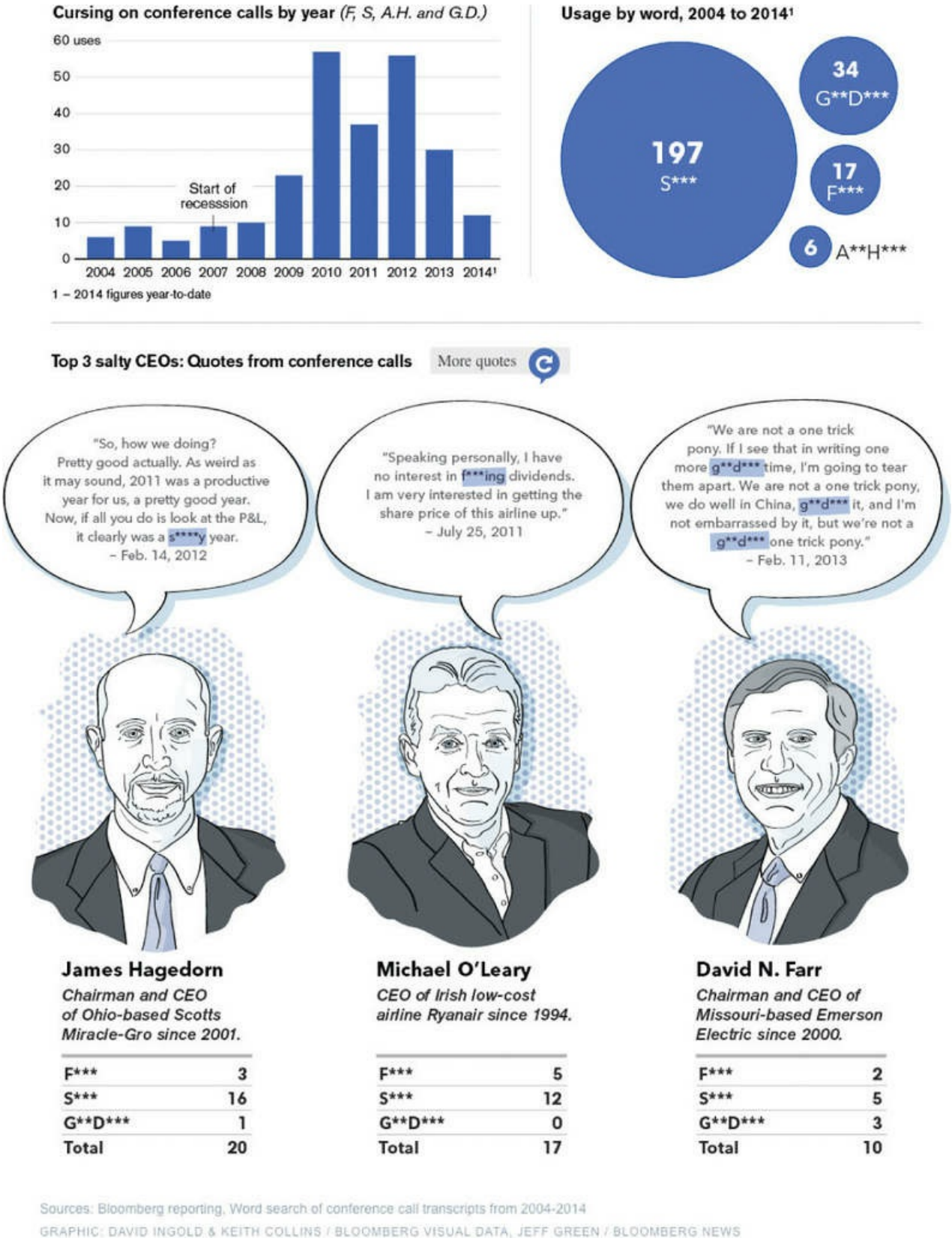
In the simplest sense, data types are distinguished by being either qualitative or quantitative in nature. Beneath this distinction there are several further separations that need to be understood. The most useful taxonomy I have found to describe these different types of data is based on an approach devised by the psychologist researcher Stanley Stevens. He developed the acronym NOIR as a mnemonic device to cover the different types of data you may come to work with, particularly in social research: Nominal, Ordinal, Interval, and Ratio. I have extended this, adding onto the front a 'T' – for Textual – which, admittedly, somewhat undermines the grace of the original acronym but better reflects the experiences of handling data today. It is important to describe, define and compare these different types of data.

Textual (Qualitative)

Textual data is qualitative data and generally exists as unstructured streams of words. Examples of textual data might include:

- 'Any other comments?' data submitted in a survey.
- Descriptive details of a weather forecast for a given city.
- The full title of an academic research project.
- The description of a product on Amazon.
- The URL of an image of Usain Bolt's victory in the 100m at the 2012 Olympics.

Figure 4.3 Graphic Language: The Curse of the CEO



In its native form, textual data is likely to offer rich potential but it can prove quite demanding to unlock this. To work with textual data in an analysis and visualisation context will generally require certain natural language processing techniques to derive or extract classifications, sentiments, quantitative properties and relational characteristics.

An example of how you can use textual data is seen in the graphic of CEO swear word usage shown in Figure 4.3. This analysis provides a breakdown of the profanities used by CEOs from a review of recorded conference calls over a period of 10 years. This work shows the two ways of utilising textual data in visualisation. Firstly, you can derive categorical classifications and quantitative measurements to count the use of certain words compared to others and track their usage over time. Secondly, the original form of the textual data can be of direct value for annotation purposes, without the need for any analytical treatment, to include as captions.

Working with textual data will always involve a judgement of reward vs effort: how much effort will I need to expend in order to extract usable, valuable content from the text? There are an increasing array of tools and algorithmic techniques to help with this transformational approach but whether you conduct it manually or with some degree of automation it can be quite a significant undertaking. However, the value of the insights you are able to extract may entirely justify the commitment. As ever, your judgment of the aims of your work, the nature of your subject and the interests of your audience will influence your decision.

Nominal (Qualitative)

Nominal data is the next form of qualitative data in the list of distinct data types. This type of data exists in categorical form, offering a means of distinguishing, labelling and organising values. Examples of nominal data might include:

- The 'gender' selected by a survey participant.
- The regional identifier (location name) shown in a weather forecast.
- The university department of an academic member of staff.
- The language of a book on Amazon.
- An athletic event at the Olympics.

Often a dataset will hold multiple nominal variables, maybe offering different organising and naming perspectives, for example the gender, eye colour and hair colour of a class of school kids.

Additionally, there might be a hierarchical relationship existing between two or more nominal variables, representing major and sub-categorical values: for example, a major category holding details of 'Country' and a sub-category holding 'Airport'; or a major category holding details of 'Industry' and a sub-category holding details of 'Company Names'. Recognising this type of relationship will become important when considering the options for which angles of analysis you might decide to focus on and how you may portray them visually using certain chart types.

Nominal data does not necessarily mean text-based data; nominal values can be numeric. For example, a student ID number is a categorical device used uniquely to identify all students. The shirt number of a footballer is a way of helping teammates, spectators and officials to recognise each player. It is important to be aware of occasions when any categorical values are shown as numbers in your data, especially in order to understand that these cannot have (meaningful) arithmetic operations applied to them. You might find logic statements like TRUE or FALSE stated as a 1 and a 0, or data captured about gender may exist as a 1 (male), 2 (female) and 3 (other), but these numeric values should not be considered quantitative values – adding '1' to '2' does not equal '3' (other) for gender.

Ordinal (Qualitative)

Ordinal data is still categorical and qualitative in nature but, instead of there being an arbitrary relationship between the categorical values, there are now characteristics of order. Examples of nominal data might include:

- The response to a survey question: based on a scale of 1 (unhappy) to 5 (very happy).
- The general weather forecast: expressed as Very Hot, Hot, Mild, Cold, Freezing.
- The academic rank of a member of staff.
- The delivery options for an Amazon order: Express, Next Day, Super Saver.
- The medal category for an athletic event: Gold, Silver, Bronze.

Whereas nominal data is a categorical device to help distinguish values, ordinal data is also a means of classifying values, usually in some kind of ranking. The hierarchical order of some ordinal values goes through a single ascending/descending rank from high or good values to low or bad values. Other ordinal values have a natural 'pivot' where the direction changes around a recognisable mid-point, such as the happiness scale which might pivot about 'no feeling' or weather forecast data that pivots about 'Mild'. Awareness of these different approaches to 'order' will become relevant when you reach the design stages involving the classifying of data through colour scales.

Interval (Quantitative)

Interval data is the less common form of quantitative data, but it is still important to be aware of and to understand its unique characteristics. An interval variable is a quantitative and numeric measurement defined by difference on a scale but *not* by relative scale. This means the difference between two values is meaningful but an arithmetic operation such as multiplication is not.

The most common example is the measure for temperature in a weather forecast, presented in units of Celsius. The absolute difference between 15°C and 20°C is the same difference as between 5°C and 10°C. However, the relative difference between 5°C and 10°C is not the same as the difference between 10°C and 20°C (where in both cases you multiply by two or increase by 100%). This is because a zero value is arbitrary and often means very little or indeed is impossible. A temperature reading of 0°C does not mean there is no temperature, it is a quantitative scale for measuring relative temperature. You cannot have a shoe size or Body Mass Index of zero.

Ratio (Quantitative)

Ratio data is the most common quantitative variable you are likely to come across. It comprises numeric measurements that have properties of difference *and* scale. Examples of nominal data might include:

- The age of a survey participant in years.
- The forecasted amount of rainfall in millimetres.
- The estimated budget for a research grant proposal in GBP (£).
- The number of sales of a book on Amazon.
- The distance of the winning long jump at the 2012 Olympics in metres.

Unlike interval data, for ratio data variables zero means something. The absolute difference in age between a 10 and 20 year old is the same as the difference between a 40 and 50 year old. The relative difference between a 10 and a 20 year old is the same as the difference between a 40 and an 80 year old ('twice as old').

Whereas most of the quantitative measurements you will deal with are based on a linear scale, there are exceptions. Variables about the strength of sound (decibels) and magnitude of earthquakes (Richter) are actually based on a logarithmic scale. An earthquake with a magnitude of 4.0 on the Richter scale is 1000 times stronger based on the amount of energy released than an earthquake of magnitude 2.0. Some consider these as types of data that are different from ratio variables. Most still define them as ratio variables but separate them as non-linear scaled variables.

> If temperature values were measured in kelvin, where there is an absolute zero, this would be considered a ratio scale, not an interval one.

Temporal Data

Time-based data is worth mentioning separately because it can be a frustrating type of data to deal with, especially in attempting to define its place within the TNOIR classification. The reason for this is that different components of time can be positioned against almost all data types, depending simply on what form your time data takes:

Textual: 'Four o'clock in the afternoon on Monday, 12 March 2016'

Ordinal: 'PM', 'Afternoon', 'March', 'Q1'

Interval: '12', '12/03/2016', '2016'

Ratio: '16:00'

Note that time-based data is separate in concern to duration data, which, while often formatted in structures such as hh:mm:ss, should be seen as a ratio measure. To work with duration data it is often useful to transform it into single units of time, such as total seconds or minutes.

Discrete vs Continuous

Another important distinction to make about your data, and something that cuts across the TNOIR classification, is whether the data is discrete or continuous. This distinction is influential in how you might analyse it statistically and visually.

The relatively simple explanation is that discrete data is associated with all classifying variables that have no 'in-between' state. This applies to all qualitative data types and any quantitative values for which only a whole is possible. Examples might be:

- Heads or tails for a coin toss.
- Days of the week.
- The size of shoes.
- Numbers of seats in a theatre.

In contrast, continuous variables can hold the value of an in-between state and, in theory, could take on any value between the natural upper and lower limits if it was possible to take measurements in fine degrees of

detail, such as:

- Height and weight.
- Temperature.
- Time.

One of the classifications that is hard to nail down involves data that could, on the TNOIR scale, arguably fall under both ordinal and ratio definitions based on its usage. This makes it hard to determine if it should be considered discrete or continuous. An example would be the star system used for rating a movie or the happiness rating. When a star rating value is originally captured, the likelihood is that the input data was discrete in nature. However, for analysis purposes, the statistical operations applied to data that is based on different star ratings could reasonably be treated either as discrete classifications or, feasibly, as continuous numeric values. For both star review ratings or happiness ratings decimal averages could be calculated as a way of formulating average score. (The median and mode would still be discrete.) The suitability of this approach will depend on whether the absolute difference between classifying values can be considered equal.