An introduction to data management

Screen 1 – Title

Welcome to the first in a series of lectures exploring the topic of Data Management

Screen 2 – Learning outcomes

Explain what constitutes as data; recognise that it takes many forms

Define what data management is, and why it is required

Recognise what is good data and what is bad data, and how operations are impacted by both

Screen 3 – Introduction and outline

When we talk of data, what do we actually mean, and how are we supposed to manage it successfully? As you might expect, both subjects will be given great attention, not just in this lecture but throughout. Incidentally, this lecture sets the scene for the remaining lectures of the series, so don't worry if you are keen to discover more on a given topic presented here, I'll highlight the places where a greater amount of detail can be expected at the appropriate places.

What we will look at here is our ability to spot when we are analysing good data and the occasions when we are not which is of considerable importance when you need to rely on data to make informed operational decisions. Ultimately, this is our goal here. We need a new appreciation of the importance of data to our professional and perhaps personal lives. As we'll go on to look at, data should be considered as the most important commodity in your business. Very little can be achieved without it, but much can be lost should we not take the appropriate care in managing it.

Screen 4 – Data defined

Just as a point of note, you may not have known that the word data is plural, not singular. Watch out for this in scientific writing, where you will see sentences written such as 'the data **are** good' as opposed to 'the data **is** good'. When in conversation, it is commonplace to speak of data in the singular form, and as language evolves this is acceptable, but you will likely be corrected when writing about data if you do not use the plural form. And speaking of modern times, as you are no doubt aware, we have experienced something of an explosion in the growth of technology in recent times. Data surrounds almost every activity.

But what exactly are data? Well, you may have heard of the term bit and byte. A bit is the fundamental unit of data, and contains two pieces of information, for example yes and no. A byte consists of eight bits, so now we are talking about a total of 256 pieces of

information, as 2^8 equals 256. A short sentence or paragraph might equate to this; and you can see this for yourselves: perhaps type some text into a text editor and see how the size of the file changes with the amount of text you include within it. We'll look further at the other various terms used to describe data later in the course, but essentially you'll have encountered the doubtless-familiar prefixes of kilo- mega- giga- etc. before byte throughout your daily activities.

Having said this, we shouldn't forget that paper records are equally valuable, and just as valid in terms of any description of data. Many aspects of operational activities are still reliant and the humble piece of paper and pen, particularly when it comes to recording the context within which data are collected and why, which is also known as Metadata. In essence, the management of all manner of data is probably something we undertake as part of our daily lives, from personal banking to filling out forms to access local services. What we are trying to do here is to become more aware of the importance of data management, with the goal of maximising the efficiency and accuracy of the systems used to collect, store and analyse it.

Screen 5 – Good data/bad data

Data accuracy is something of a moveable feast, and can a compromise can often be made between quality and efficiency. This may seem a little counter-intuitive, as you might think that every effort should be made to make all data as accurate as possible. This is not the case for much instrumentation, for example instruments that record the level of the tide or measure currents can have variable settings that trade-off power demand against the number of measurements made for a given period of time.

The collection of data can also impact on quality. Instrumentation is one aspect, but when that data is recorded manually, problems can often occur. A typical example of this at sea can be observations of marine mammals. As skilled as the observers may be, they frequently have to endure long shifts and tiredness is often a problem. It is not uncommon under these conditions for periods of an individual watch to be removed or discounted due to unreliability.

Very much data are freely available and open source. As useful as this is, we cannot always be certain as to the reliability of this data, nor that it was collected appropriately. Due care must be taken, even if the Metadata appears to be rock-solid.

Ultimately, our ability to analyse the data efficiently may drive a further compromise over quality. As much as we would like all of our data to be of the highest quality all of the time, as long as we are transparent, that is honest about how it was collected and processed allowing any end-user to make an educated assessment of their own, then this will largely be sufficient. It really will depend on the specific circumstances of the situation with which you are faced.

Screen 6 – Using and processing data

Often the true picture of data quality will not materialise until all of what is termed the post-processing is complete, and you begin to analyse the signals of interest more closely. Perhaps only then will you determine that the collected data can help inform that operational decision. Careful inspection of the results of your efforts is required here; wisdom is required to ensure that results are not over-interpreted. For example, this image happens to be some current speed data for a location in the Western English Channel. Those of you with keen eyes will see that it is for a very specific period of time. What we cannot do is over-interpret this data and say that it applies to the whole of the English channel, all of the time.

Post-processing generally refers to the stage of data analysis that follows the treatment of raw data. It is the time when you get that clearer picture and truly understand what is going on for whatever situation you set out to study. Data such as these displayed here are insufficient to describe long-term activity at this location, but they still might be able to validate a model of current speed here, perhaps contributing to an enhanced ability to forecast conditions, or helping to determine some other important process locally. The key point is to use your judgement wisely, and question all aspects of what is being shown, rather than being distracted by what might otherwise appear to be solidly reliable.

Screen 7 – Managing the flow

There is a balance to be struck between the volume of data collected and the available resources to process and analyse it. Of course, this depends on what your particular goal is. Should you have automated systems that process the data without further effort, then perhaps more is better. For example, if I were to conduct an analysis of tidal currents in a particular location, perhaps in order to refine a model to predict future current speed, then more really is better. The increased level of detail or further back in history your measurements take you will always help in this situation. But sometimes, there is a need to interpret the results of the data in a specific way. Perhaps you are seeking to establish the most efficient way in which ships enter and leave port. It is unlikely that a fully automated system could manage this, without some specific guidance or input from a user that appreciates the context within which the computer arrives at a decision. In short, we need to think about managing the flow of information, which would hopefully be made clear in the planning process for the data collection itself.

As you might expect, the management of large quantities of data can be a full-time occupation. From the perspective of the UK ocean scientist, this role falls to a publicly-funded body called the British Oceanographic Data Centre (BODC).

Screen 8 – Housekeeping and Metadata

We need to consider that any data we acquire may not be used immediately, nor even for the intended purpose for which is was collected, and a significant period of time might elapse between collection and use. This is where Metadata comes in, and strictly is the first and most important step in good data management. Metadata gives that all-important context to the circumstances within which the original data were collected. Without it, future users of the data are blind, and cannot apply any new thinking to old data if it missing this all-important component...

For the moment it is enough to say that small details surrounding the collection of the data can be incredibly important. As the image suggests, Metadata is truly a love note to the future users, as we cannot know in advance how others may want to use the data in the years that follow. Metadata can be anything related to the acquisition of the data, including simple things such as the date of collection, the positional coordinates, weather conditions, time of sampling, instrument settings and so on. To be on the safe side, report all of these things and more besides. The next user of the data will love you right back for it.

Screen 9 – Storage solutions

Assuming we have done a great job of collecting the data, recording the appropriate Metadata and made a start on the analysis, what do we do about data storage? It is indeed more complex than you might think, and whilst options are plentiful the solution must be appropriate for the needs of the sector within which you work. Can we afford to operate a large data warehouse? Do we have the resources required? Do we need to access the data instantly, from anywhere on the globe?

If so, then perhaps cloud computing might help with this and all other issues. But what if we work offshore, and require a stable connection with a lot of bandwidth? Clearly there is no one-size-fits-all solution, but there are plenty of options for us here so again we will return to this topic later in the series. One thing that we should be doing daily is backing-up our data. This includes personal and professional data. In my line of work as a research scientist, it is not uncommon to run several back-ups a day, as considerable effort can be expended finding solutions to problems that it would be inefficient in the extreme to repeat. Security of course is equally important, particularly if you are working in commercially sensitive environments, or are having to process data that is protected by legislation of any kind.

Screen 10 – Operational decision-making

The goal in all of this is to effectively manage our data to better inform our operational decision-making. If it sounds like there is a lot to think about here, it's because there is! We all work in complex sectors of what is a heavily data-influenced industry and if we hope to successfully negotiate complex scenarios, then this subject cannot be overlooked nor underestimated.

Ultimately, to better understand all aspects of the environment within which you work, successful data management is critical. There are other benefits associated with best practice, of course, in that those who are highly skilled in this area are in high demand. There is a global shortage of individuals with deep analytical data processing and management skills. A report from the Mckinsey Global Institute (2011) recently suggested that the global shortage of data-aware managers was in the region of 1.5 million. No excuse then to not give this aspect of your studies the fullest attention.

Screen 11 – Responsibility

Who is to take responsibility for ensuring that best practice in data management is followed? Frankly, it is up to all of us to ensure that these basic steps are followed. We cannot leave it to others and assume that all of our valuable data has been correctly stored, processed and all of the Metadata correctly written-up and attached to the raw data.

We must begin to appreciate that data is an important and valuable commodity that could influence whether we, and the organisations we represent, are successful. Whilst some of this may seem self-evident, I am sure we have all encountered more than one occasion when either we or someone we know has endured unnecessary data pain as a result of not following what may seem, at least superficially, to be good old-fashioned common-sense. I can certainly recall more than one student who has failed to submit their coursework on time because they had a problem relating to their project data! The buck should stop with us. Recognise the importance of good data management, implement some of its suggestions into the way in which you work with data and you will reap the benefits for the rest of your career.

Screen 12 – Learning outcomes reviewed

Explain what constitutes as data; recognise that it takes many forms

Define what data management is, and why it is required

Recognise what is good data and what is bad data, and how operations are impacted by both

Acknowledgements and references

McKinsey & Company (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, Cambridge.

Attributions

The content of this course is copyright to the Marine Learning Alliance (MLA) and you MAY NOT share it with anyone else, by any means, without prior permission from the MLA. Content provided in this module is provided for YOUR INDIVIDUAL USE ONLY. All pictorial images and diagrams used in this lecture have been created by MLA or are copyright free, unless otherwise stated. Where an image has not been created by MLA appropriate acknowledgement is given at the foot of this document.

Screen 3:

https://www.flickr.com/photos/malfet/6870465199

Screen 7:

https://upload.wikimedia.org/wikipedia/commons/5/53/USACE_Gavins_Point_Dam.jpg

By Robert Etzel, U.S. Army Corps of Engineers [Public domain], via Wikimedia Commons

Screen 8:

https://www.flickr.com/photos/centralasian/8071729256

Screen 9:

https://upload.wikimedia.org/wikipedia/commons/1/19/Interior_of_StorageTek_tape_library_at_NERSC_%282%29.jpg

By Derrick Coetzee from Berkeley, CA, USA [CC0], via Wikimedia Commons

Screen 11:

https://www.flickr.com/photos/smemon/6032417950/in/photolist-ac4HG3-au8uNg-q1mcsn-91iRws-9mq7H9-eNgdFU-nuUQM7-pkNGFp-pakM17-7YzBq9-kR9oZt-5sTuyk-dJvTjJ-5eLK2r-9fBHYt-nuURRo-8YV4K5-oeYpCH-kRawJq-kKPBSS-7FGCDx-9Vk8NS-qrEzjs-e7M5Qx-dDwovw-dBd8Rf-7cSb1j-cwujQN-fdgXiR-9RWUDD-5Tj7MG-8YUWaE-getGqG-kR9kG6-8YVJ4b-4CgAaf-nTGrou-bCbzWz-bVJFVV-5o2hnx-iNK4nX-getFvs-d4ouN3-pYfF75-9YoGVQ-7YzDyE-8YV91s-nuUSuA-pMdVmL-pembtL