Aperius Consulting

ESTABLISHING DATA CONFIDENCE

Summary:

- Assessing the validity of data is challenging.
- There are five core tests that should be applied to all data:
    - Is it plausible?
    - Is it from a reliable source?
    - Is it up-to-date?
    - Is it the cross-referenceable?
    - Is it complete?

Lies, Damn Lies and Big Data – Establishing Data Confidence

Roger Alderson, Aperius Consulting

A couple of years back, I came across an interesting new website, www.emergent.info. It's part of a research project at Columbia University that looks at how unverified information is reported in the media – effectively, a rumour tracker. It lists stories that are emerging and monitors how they're trending in traditional and social media.

Rumours gain acceptance because lots of people believe them. But, it doesn't make them facts.

I visited the site because it sounded interesting but afterwards I realised that it's a good example of how information morphs and flows. Rumours are often treated as facts simply because of the number of references to them and the volume of people that believe them. That's how 'common misconceptions' take root and grow. Things like: veins are blue because that's the colour of de-oxygenated blood, or glass is a slow-moving highly viscous liquid, or diamonds are formed from highly compressed coal. These artificial 'facts' evolve because they are repeated so many times by people who may appear to know what they're talking about.

And this happens in the business world too. Despite the talk of big data, small data and fast data, a lot of business decisions are based more on intuition and experience. People in authority, unencumbered with the burden of accurate data, speak with conviction (based on their years of experience) and those around them listen … and believe. The worst possible situation is when experienced business people espouse opinions based on intuition and experience and support their position by referencing inaccurate (or invalid) data.

But, how do you test the accuracy of information when it's presented to you? It isn't always immediately obvious to discern what data is relevant and appropriately accurate. I have an approach based on assessing five C's.

Credibility – Is the data plausible? This is the first test and one that most people apply. Who else believes this data? If somebody that you respect (like a teacher or a leader or a friend) believes it, then the perceived accuracy of the data increases. It's a valid test and shouldn't be ignored but nor should it be blindly accepted at face value.

Constancy – Is the source of the data trustworthy and reliable? This is the second test to apply to data. Where did it originally come from? Is the lineage/provenance of the data known and trusted? If the data comes from (for example) direct measurement using appropriately calibrated instruments, then the perceived accuracy of the data increases.

Currency – Is the data up-to-date? This is a test that is sometimes overlooked despite its importance, especially in today's fast moving world. What was accurate yesterday may not (and probably won't) be accurate tomorrow. Of course, currency is a relative term. When looking at trends in global climate change, currency is measured in millennia but when looking at trends in mobile advertising, currency is measured in months, weeks or days. Even when data is highly credible and trustworthy, if it is too old, then its accuracy could be excessively low.

Consistency – Is the data confirmed by multiple sources? Credible, trustworthy and up-to-date information has a good probability of accuracy. But there are many people that have been burned by not cross-referencing. It's always worth taking the time to seek out alternate data sources to see if they are in agreement with the primary source. Different sources can have slight variances (just as different people interpret the same situation differently), but check to see if there is reasonable correlation of the data.

Completeness – Is the data complete? The final test is to look at the breadth of the data and assess if it includes all relevant data-points. From a statistical perspective, this is where sample sizes are an important consideration and tie to confidence levels and intervals. But even on more general levels, the completeness of the data links to whether or not all appropriate elements are considered within the dataset. The omission of even one critical item of data can significantly skew results.

Assessing data objectively can help in building confidence it

Of course, it's rare that all five criteria line up and provide a strong indication of accuracy. Typically, more than one is less than conclusive and yet you still have to make a decision. I approach this by applying scores to my assessment of the data in light of each criterion – scoring each from -2 to +2 and then summing the results to get a final score. It really depends on the importance of the decision that you're facing, but the general rule of thumb that I apply is:

<0 – I have very little confidence in the data. At these levels, this really is just guesswork. Your HR director may value your "ability to make decisions without all the facts", but it's really just *guessing with panache*.

0 to 5 – I have some confidence in the data. At these levels, I feel that the data has merit and I'll start to look at it more closely. Although the data isn't 'perfect' it should provide a picture of the situation and is definitely worth assessing.

>5 – I have growing confidence in the data. At these levels I feel that the data will provide me with a good perspective and strongly steer my decision. These are the sorts of levels where data-driven decision-making begins to become a reality.

Summary

Don't think that I don't value intuition and experience. I do. I value it highly.
<u>I just don't value decisions that are made exclusively on it</u>.

CONTACT
M: +44 (0)7966-126882
E: roger@aperius.com
T: @RTAlderson
www.aperius.com