

Self fixing data



Talking to various subsurface specialists around the world in various oil companies you come across many different attitudes to data. The most common view is concern that the current data's poor quality is causing some risky business decisions to be taken, usually this is rooted in worry that checking, keeping and publishing the data are being given too low a priority. However, just occasionally some data users have, what is, to me at least, a much more surprising attitude. The users in this set claim that data quality is a non-issue, their view is "if there is anything wrong with the data, I'll spot the mistake as soon as I try to use it". This view is rare but I've come across it all over the world, it's certainly not a

cultural thing. My observation is that it tends to be inexperienced engineers that have this outlook, but I've also heard the same observation from even the most senior oil company staff.

In his famous 1948 paper¹ Claude Shannon provided the mathematical basis for investigating information handling systems. This includes a direct trade-off between redundancy and the ability to detect or correct errors. Take the fly-by-wire system in an aircraft, for example, manufacturers know that any design of sensor, control system or actuator will have certain situations where it doesn't do what is expected, so they add extra redundancy, with maybe three independently designed sensors each linking to three independently designed control systems. This ensures that each component "checks up" on the others and any error in any one element is quickly spotted and doesn't cause the whole aircraft to fall out of the sky. This is expensive and 99 times out of 100 not needed, but for aeroplanes one time in a hundred is not good enough and the extra cost can be easily justified.

Now, of course, there are some data mistakes that anyone can spot, surely we've all experienced that moment when you spot that the well is showing up 400 miles South of Accra (in Ghana) and realisation dawns that the location has not been set. But once any data is corrupted it can only be properly corrected by finding an alternative corroborating source. In other words by employing redundant copies to identify errors and fix them. My personal suspicion is that the way budgets in most oil companies are designed is to drive to increasing efficiency, that is to ensure that every piece of data is only generated when it absolutely has to be. Of course we can always fill in a value that seems to make sense, but that's as liable to end up destroying a drilling rig as finding oil.

It is well known that stock traders consistently overestimate how effective they are, whenever their personal impression of how well they did last year is compared with the actual results it turns out that their own assessment is wildly optimistic. I can only assume this is the type of dynamic that is behind the myth of self-fixing data, because I can't believe that anyone really believes entropy can be forced to go backwards.

¹ "A Mathematical Theory of Communication" originally published in The Bell System Technical Journal and available from <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>