MINING AND MEANING

Dear Linux Magazine Reader,

I'm looking at another crop of press releases and marveling at the ever-expanding market for products that fall under the ubiquitous heading of Big Data. Big Data is everywhere. (Or is it Big Data *are* everywhere – to reprise that classic quandary?) Gigantic file storage appliances from vendors like Oracle and HP and software tools like Apache Hadoop continually add new building blocks to the data infrastructure. And the goal of all this? Apart from more conventional uses for scientific modeling and analysis, one of the reasons for this huge interest in accumulating data is a phenomenon known as data mining, which Wikipedia defines as a process that "attempts to discover patterns in large data sets."

We're all aware that data mining technologies will have a big effect on advertising and retail, but I'm not sure the popular media is giving full attention to the radical change it brings to how we think about information and knowledge.

Computers grew up around academic science, and science grew up around a process known as the scientific method. The scientific method is roughly as follows:

- *Question* – Formulate a question about something in the natural world.
- *Hypothesis* – Restate that question in terms of a testable conjecture.
- *Prediction* – Devise a test that will prove or disprove the hypothesis, and make a prediction about the outcome of the test.
- *Test* – Perform an experiment that tests your prediction.
- *Analysis* – Analyze the results of the experiment, and determine whether your prediction was correct.

This process has a mythical significance to scientists, who have an aesthetic revulsion to sloppy procedures, in which the goals and the questions are not well defined, and the experimenter simply tinkers around like a child disassembling the toaster to see what will happen next.

Engineers also have an unwritten code of elegance, in which it is better to know what you are asking before you go looking for the answer. Social scientists, and even humanities specialists, also live by the creed of "no fishing expeditions," which means one shouldn't just troll around in the data trying to hook some meaning: better to have a vision in advance for what you want.

One of the stranger aspects of data mining is the potential for turning this paradigm on end. The whole point of data mining is to go inside a heap of data and look for patterns – in that sense, it really is a fishing expedition. The scary part is, it actually works; or at least, it is possible to use data mining to make predictions and optimize marketing practices, but you still need to keep a clear head about what you are actually getting and what you are actually "proving."

What is really happening is, the data collection phase is no longer part of the experiment – it becomes a precursor to the experimental process. Data collection captures a snapshot of reality, and the experiment is performed later on that snapshot, rather than directly on the natural world. An experiment carried off against the data store can be as scientific as any other experiment – one could easily formulate a question and hypothesis and descend down into this data to test it. But just remember you are testing a pre-compiled block of data, and the subtlety of your hypothesis is constrained by the fact that you didn't know what you were asking before you gathered the data.

Such experiments can lead to very remarkable findings, but one must be aware of the limitations on attributing *meaning* to this information. For instance, one could learn that men are 17.2% more likely to buy motor oil than women, and that might be enormously important information for a store manager, but it doesn't really explain anything meaningful about the differences between men and women in a way that a scientist would consider rigorous. The grail at the end of the process is not really a conclusion but is something that once would have been considered more like an observation. Yet, the brain leaps so effortlessly to meaning when it beholds a statistic. Such is the peril of data mining: Just as data collection shifts to the beginning of the process, so the potential for a fishing expedition shifts to the end.

The biggest danger is a future in which we start to ascribe too much importance to rudimentary information that might mean a little but doesn't mean much.

Joe Casad,
Editor in Chief