Big data - wozzat?

So what is this big data thingy?

Big data has become a hot topic lately. The people who deal with it ("data scientists") have become much in demand by companies wanting to find important business insight in amongst their sales data, twitter mentions and blogposts.

Which confuses three different concepts.

    - Big Data is defined as the processing of data that's larger than your computer system can store and process at once. It doesn't matter so much where the data is from – it's more important that it's too huge to handle with 'normal' processing methods.

- Social media mining looks for patterns in the posts, tweets, feeds, questions, comments et al that people leave all over the Internet. It's a logical consequence of Web 2.0, that idea that we could not only read what people put on their websites, but contribute our thoughts etc to it too.

- Data analysis looks for patterns in any data. It doesn't have to be big data (though it might be), and it doesn't have to come from Internet use (although it might be that too). It's just data, and the tools that have been used to understand meaning and find insights in data still apply. The data scientists have a saying "everything old is new again", and it's lovely to see a whole new generation discover Bayesian analysis and graphs as though they were shiny new super-exciting concepts.

So what are the data scientists trying to do here, and why is it special?

Well first, a lot of data scientists are working on Internet data. Which is why big data and Internet data often get confused: a collection of blogs, tweets etc can be seriously big – especially if you're trying to collect and analyse all of them.

And they're analyzing the data and using the results to help drive better business decisions. Yes, some people do this for fun or to help save the world, but mainly it's popular because better decisions are worth money and those analysis results are big business differentiators.

Which is great until you realize that up 'til now most of those decisions were made on data from inside the company. Nice, structured, controlled, and often quite clean (as in not too many mistakes) data, often stored in nice structured databases. Which is not how the Internet rolls. What data scientists often end up with is a mix of conventional structured data and data with high structural variance: data that looks kinda structured from a distance (time, tweeter, tweet for

example) but has all sorts of unstructured stuff going on inside it. Sent from a mixture of conventional systems and devices. That companies often ask to be analysed in the same way they're already analysing their structured data.nnSo, alongside the usual corporate data, we now have 3 new types of data that we can access and process: structured data stored in warehouses, unstructured internet-style data (blogs, tweets, sms) and streams of information.

Lets back up just a little. To do analysis, you need a question, some data and some tools (or techniques, if you will). It also helps to have someone who cares about the results, and it's even better if they care enough to explain what's important to them and why.

The Question

First, the question. Asking the right question is difficult, and often an art. Sometimes it'll be obvious, sometimes it'll come from staring at a subset of the data, sometimes the question will be given to you and you'll have to hunt for the data to match. We'll talk about the question later.

Handling the Data

So we have a question and some data. And if the data is big, this is where the Big Data part of the story comes in. If you suddenly find yourself with data that you can't analyse (or possibly even read in) using the computing resources you have, then it's big data and your choices (thank you Joseph Adler) are:

    - Use less data. Do you really need all the data points that you have to make those business decisions (try reducing down to a statistically significant number of points, or reducing down to mostly the points that are important to you)? Do you really need all the variables you've collected (do a sensitivity analysis)? Are there repeats (e.g. twitter retweets) in your dataset (tidy it up)?

- Use a bigger computer. You'll need to both store and process the data. "The cloud" is a generic term for storage that's outside your home or office that you can still access from wherever you want (e.g. over the internet). Amazon Web Services is a prime example of this; other cloud storage includes Microsoft Azure (sql datastore), Cassandra (bigtable datastore), Buzz Data, Pachube (primarily storage for sensor outputs, a.k.a. the Internet of Things), Hive (data warehouse for Hadoop) and sharded databases.

- Use parallel processing across multiple computers. A popular process for this is map/reduce, which splits data into chunks that are each processed by a different machine. Places where map/reduce is available include Hadoop, which also has a higher-level language, Pig, that reduces down to map/reduce instructions.

- Get smart. Get lateral about the problem that you're trying to solve (see any good statistics

textbook for ideas).

The processing

And then we have the techniques part of the equation (sorry – couldn't resist the pun). Again a post for later – there are many tools, packages and add-ons out there that make this part of the process easier.

Explaining the results

If you're doing big data analysis, you're doing it for a reason (or you really like fiddly complex tasks). And the reason is often to increase the knowledge or insight available to an end user. For this, we often use visualisations. Which is another post for later.