

Why big data is a big deal

It's often said that insurance will be one of the many industries to benefit from the big data revolution. Peggy Brinkmann and Nancy Watkins break down the buzzwords and explain the techniques insurers can use to improve their risk analysis, detection of claims fraud and marketing

Big data and data science are ubiquitous phrases in business articles right now. Software and hardware companies new and old are pushing solutions for big data. Companies are trying to recruit data scientists. You may be wondering, what exactly is big data, and is it really a big deal?

WHAT IS BIG DATA?

The term big data is vague. Some have proposed that “big” implies any data where, because of its volume, its management and analysis is an issue given the current toolsets available. But as our hardware and software become more powerful, the threshold of big data changes, making this an unhelpful definition. “Bigger data,” as in “bigger than you can effectively deal with using traditional tools,” might be the more appropriate description.

One of the breakthroughs making big data valuable to businesses is the ability to quickly combine and evaluate both structured data (organised in a data model, such as rows and columns in a relational database) and unstructured data (not pre-organised, like text or images). Big data is often generated through automated or very inexpensive means, such as text, weblogs, transactions records, or sensors, including point-of-sale scanners and radio-frequency identification (RFID) tags. Rampant low-cost data generation means that data can grow to terabytes and petabytes (1,000 terabytes) very quickly.

WHAT IS DATA SCIENCE?

“Data science” and “data scientist” are other relatively new terms. The definition of data

science is even less clear than big data, but Drew Conway, a leading expert who writes and blogs about data science, has proposed the following definition: the intersection of hacking skills (meaning computer programming skills to manipulate files and implement algorithms, not the skills for breaking into secure systems!), statistics and machine learning knowledge, and subject matter expertise, shown in Figure 1.

Data scientists use both statistical techniques and machine learning algorithms for identifying patterns and structure in data.

The term big data is vague. Some have proposed that “big” implies any data where, because of its volume, its management and analysis is an issue given the current toolsets available.

In fact, their toolset is very similar to those of statisticians, data miners and predictive modellers. Here are some definitions of the various types of techniques used:

- If you make assumptions about a type of random distribution for your data in order to fit model parameters or make inferences, you are applying statistics. Regression models are a popular type of statistical technique that assumes data comes from a normal bell-curve distribution.
- If you are giving the computer rules and/or algorithms to automatically search through your data to “learn” how to recognise patterns and make complex decisions (such as identifying spam

emails), you are implementing machine learning.

- If you are giving the computer rules and/or algorithms to automatically search through your data to discover previously unknown and useful relationships in your data, you are data mining. Many of the algorithms used for data mining came from statistics and machine learning.

A predictive modeller may use any or all of these approaches to predict a value or likelihood of an outcome, given a number of input variables. A data scientist applies these same approaches on large data sets, writing code and using software adapted to work on big data.

In addition, data scientists have substantive expertise and deep domain knowledge. Drew Conway defines this as the ability to ask “motivating questions” (i.e. the important and valuable questions that can be answered with the data) and test hypotheses. An important aspect of a data scientist’s work is the ability to visualise information and communicate it effectively to the audience.

DEALING WITH BIG DATA

Delivering real value from big data requires approaches different from using data stored in a relational database. There are three fundamental approaches: massively parallel processing (MPP) databases, “Not Only SQL” (NoSQL) frameworks and columnar databases.

MPP and NoSQL both use cluster computing, in which a set of connected computers (called “nodes”) work together as a single system. In cluster computing, the data is divided up and stored on different computers; data processing and analysis

operations are run locally, or distributed, on each machine. Columnar databases are most effective with only special types of data (e.g. data fields that have only a few distinct values), which precludes their use for unstructured data.

MPP databases are relational databases specially designed to span the clusters. Examples of MPP databases are Teradata, Netezza and Greenplum. Oracle and IBM offer MPP “appliances” for their database products. Queries can be written in the familiar SQL language, and database administration is similar to that used for existing relational databases.

NoSQL technologies take a different approach to managing big data. These include frameworks such as Hadoop, Dremel and High Performance Computing Cluster (HPCC), with Hadoop being the most popular. Hadoop uses a simpler, more flexible data structure than a relational database and can achieve scale and performance on lower-cost commodity hardware. A key component of Hadoop is a programming model called MapReduce, which distributes processing across the nodes of the cluster. Data can also be accessed with a SQL-like language called Hive that builds MapReduce programs in the background. Still, using Hadoop requires new skills for system administrators, data analysts and predictive modellers.

All of these frameworks make it possible to query and summarise big data. This is very useful when selecting, sampling, and/or summarising data from a big dataset to a size manageable for further analysis and modelling using software available for a PC or server. But what if the analytics and modelling algorithms need to be directly run on the big data? There are two options: use software/packages with algorithms that have been written to take advantage of the distributed processing power of the cluster (either MPP or NoSQL), or write your own distributed algorithms.

The available library of statistical and machine learning algorithms for evaluating and learning from big data is growing, but is not yet as comprehensive as the algorithms available for the non-distributed world.

An alternative to using analytic software is to write the algorithm in a MapReduce style of programming and run it on

Hadoop. There are a number of languages that support this, including R, an open source programming language for statistical computing and graphics. Even so, not all algorithms familiar to the statistician and data miner are easily converted to the distributed computing environment. The bottom line is that, while fitting models on big data has the potential benefit of greater predictive power, some of the costs are loss of flexibility in algorithm choices and/or extensive programming time.

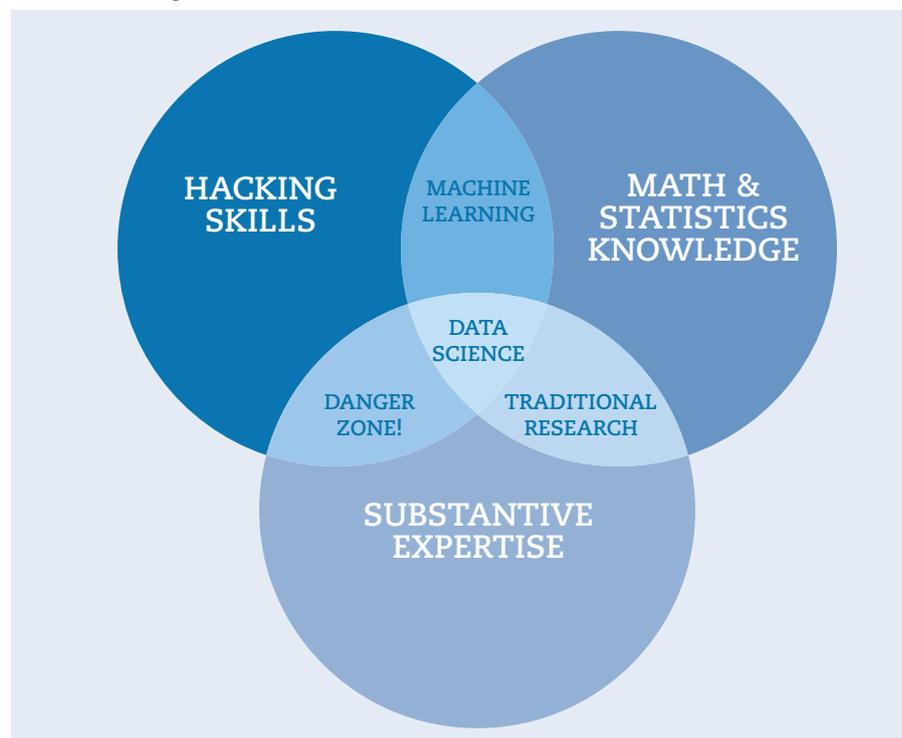
Algorithms written to run in an MPP environment are also called in-database algorithms. Some of the database vendors have their own software or R packages to run in-database algorithms. SAS sells a product that works with many types of MPP databases as well as Hadoop. Statistical and machine learning algorithms that run on Hadoop clusters are also available in Mahout, an open source algorithm library for Hadoop.

There are some other options for running analytics on “bigger” data outside of MPP/Hadoop, but each has trade-offs between algorithm flexibility, predictive power,

programming time, run time and cost. From the least expensive to the most expensive, some of those options include:

- Sampling the data. This may cost some predictive power, but shortens time and does not require investment in new software and/or hardware. If well thought out and coupled with the appropriate tools, sampling can work with minimal if any loss of predictive power.
- New software/packages that hold the data out of the memory, such as the bigGLM and bigfr packages in R and the RevoScale algorithms of Revolution R. These enable processing of larger datasets (and possibly better predictive power) but at the cost of more programming time and/or new software cost. Also, at present only a limited number of algorithms are available.
- New hardware with larger memory. This can be an expensive option for a very large amount of memory, but it enables processing of larger datasets with minimal additional programming time or software costs and no loss of algorithm choices. More processors can also help,

Figure 1: The Data Science Venn Diagram. Source: https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html



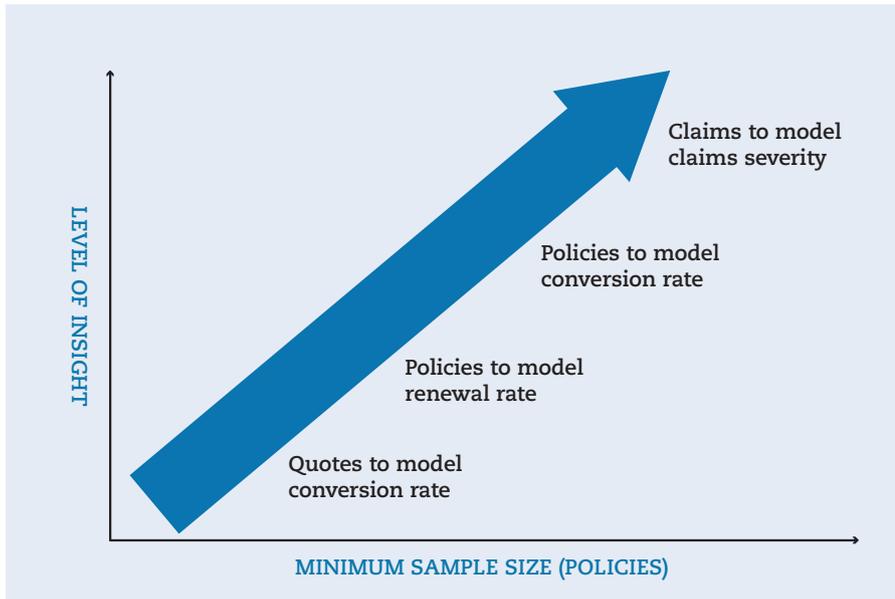


Figure 2: Opportunities for Data Analysis

if the analysis software has been written to take advantage of them (this is called multi-threading).

BIG DATA AND INSURANCE

Currently only the largest insurers have traditional policy rating and claims data that would be considered “big” (i.e. potentially unmanageable with a relational database.) But insurance companies of all sizes are encountering big data from new sources such as their website traffic, vehicle telematics programmes and social media. Companies are already trying to use this data to market, price and underwrite their insurance policies, but this is just the beginning. In the near future, experts anticipate completely new data sources, and applications such as “home telematics” or “usage-based home insurance” – using sensors in structures to measure moisture and detect flooding in the house, track how many occupants are in the home at what times, and measure utility and appliance usage – to use as predictors of property claim risk.

Most companies have IT infrastructures that were not designed for the volume of data generated by their web traffic or telematics programmes. This big data is very expensive (if not impossible) to manage using their existing relational

databases. Companies often give up and discard it, keep only summaries or very short histories, or outsource the storage to a vendor that provides them only limited reporting and analysis capabilities. These choices leave a data scientist without the means to access, leverage and integrate the big data to find new insights and value.

For example, while social media monitoring is often outsourced, the next opportunities for insurers to leverage social media data are in areas such as distribution, underwriting and claim fraud detection, all of which requires integration with internal data sources. Telematics data can help identify marketing opportunities, streamline accident reconstruction or recover vehicle thefts. In order to benefit at all from big data, insurers must gather it, store it, and access it to discover these insights and make better decisions.

LEVERAGE THE LITTLE DATA

While the media is enamoured with the possibilities of big data, opportunities still exist to apply the tools of data science to our “little data,” such as:

- Leveraging unstructured data, including notes from customer relationship management systems or claim files
- Integrating data from multiple systems, and/or external sources such as vehicle

characteristics, geographic information systems, etc.

- Exploring machine learning techniques to see if there are previously unrecognised patterns in the data
- Using visualisation techniques and new software for data visualisation to explore data and explain insights to decision makers

Companies that have operational databases without a data warehouse to facilitate access for analysts and modellers have a tremendous asset that they cannot leverage. A data warehouse without good tools for business analysts to access it is a significant lost opportunity.

We have found that massive amounts of data are not required to discover valuable insights through predictive analytics. Even if there are not many policies, there are probably enough quotes to analyse. Figure 2 illustrates some of the types of analyses possible as a company grows in policy volume.

We have revamped homeowners insurance rating plans with only a few thousand policies. Useful models for claim severity have been built with as few as 300 claims. We have identified high- and low-profitability segments in a book of business with as few as 10,000 claims. Clients can use these models to track the expected profitability of their portfolios over time.

Of course, dataset size and the level of noise in the variables modelled will impact the power of the models and the types of analyses that are feasible, but it would be a mistake to underestimate the potential of “little data.”

NO GOING BACK

Big data has already arrived in the insurance industry, and there is no going back. Companies need to start creating an intentional path to manage and use it. Data science is the key to extracting value from big data, but the innovations of data science can be applied to “little data” as well to discover new insights and make better decisions. Big data is indeed a big deal, after all. ■

Peggy Brinkmann is an actuary and Nancy Watkins is a principal and consulting actuary for Milliman in San Francisco. Emails: peggy.brinkmann@milliman.com, nancy.watkins@milliman.com