Key tools of Big Data for Transformation

Review

The challenges of big data can be captured succinctly as follows[1,2]:

- Volume; ever increasing volume which breaks down traditional data-holding capacity
- Variety; more and more heterogeneous data from many formats and types are bombarding the data environment
- Velocity; more and more data is time sensitive now; frequent updates are taking place instead of relying on historical old data and data in real time is being generated now by the internet of things, amongst others.
- Veracity; how valid and reliable is the data? Since now we have so much data, any point of view can be supported by selective adaption of data.

For volume, Map Reduce[3] works to harness the potential of billions of items of data. The first part is that the data is mapped down into key and value pairs; the reduce job combines the mapped data into smaller set of data by eliminating repetition and redundancy amongst others. Hadoop is open-source for handling big data, applying MapReduce and a variety of other distribution systems and clusters. and there are variants produced by many different vendors such as Cloudera, Hortonworks, MapR and Amazon. There also other products such HPCC and cloud-based services such as Google BigQuery.

For variety, NoSQL (Not-Only SQL)[4] is a new way of handling variety of data. Relational databases use rows and columns in handling data but NoSQL uses a number of other components such as giving unique key to every item in the data. Companies utilize NoSQL because it captures so many elements of supply chain that were previously only based on experience or hunches (Techterms, 2013). MongoDB – an open-source NoSQL database and other instances of NoSQL are Cassandra,

For velocity, Complex even processing or stream processing allows us to handle the velocity of time; real time generated by countless sensors involved in every bit and inch of the supply chain process can be automatically fed into stream processing which uses defined algorithms to analyze it almost instantly. Early alarm systems of supply chain would find this invaluable because red alert can be issued to company from the supply chain right when it occurs instead of giving the red alert after a sufficient time has elapsed which has given the disruption time to reckon havoc for the business. Apache spark has also been developed which is found to be around 100 times faster than hadoop

[1] Rozados, IV, Tjahjono, B, 2014; 6th International conference on operations and supply chain management, Bali.

[2] IBM [a], NA. "Infographics & Animations", available at:http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[3] IBM [b], NA;"What is MapReduce" available at: http://www01.ibm.com/software/data/infosphere/hadoop/mapreduce/

[4] Techterms, 2013. "NoSQL"

for MapReduce purposes. Storm is an open-source distributed computation system designed for processing multiple data streams in real time.

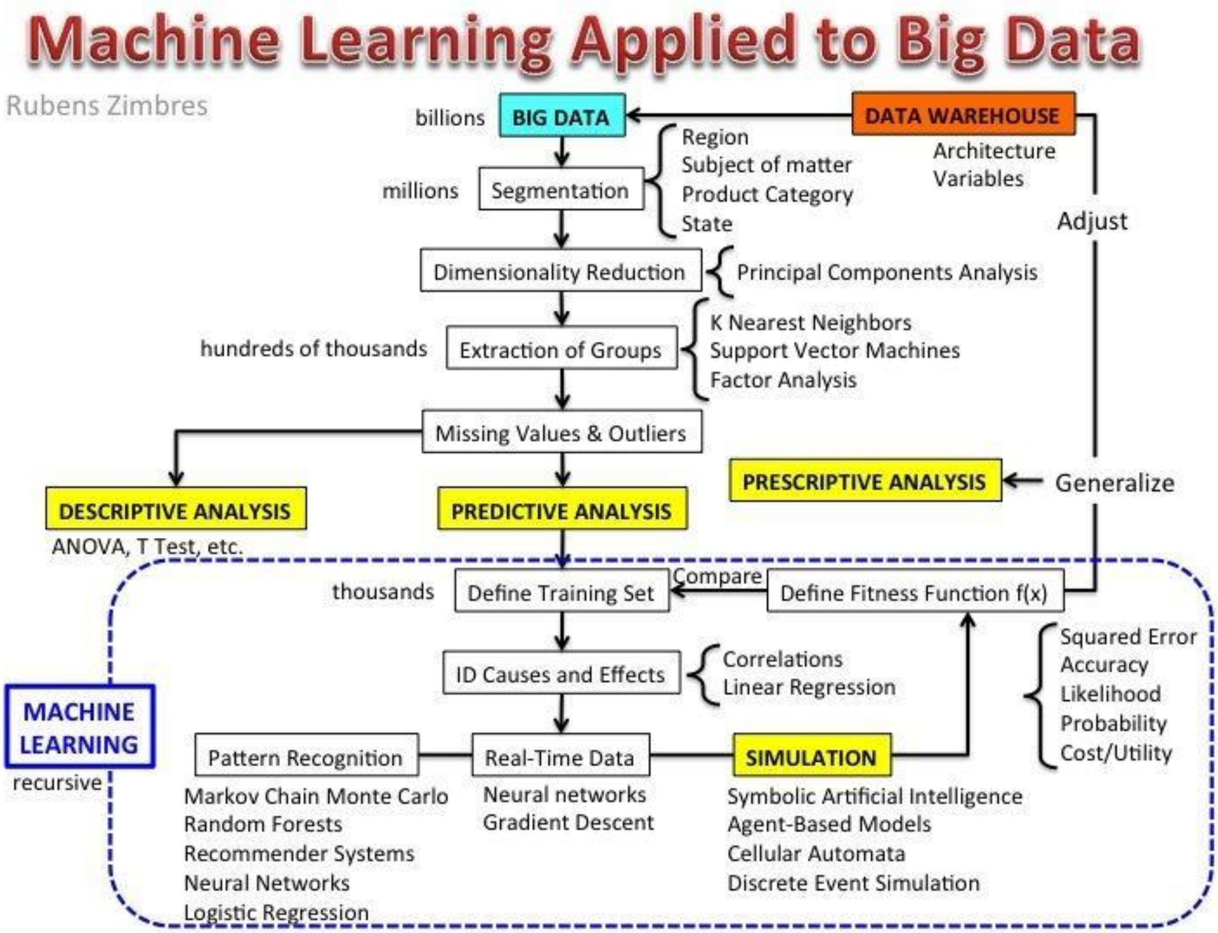
For veracity, Machine learning and data mining constitute a number of models taken from mathematics, statistics and artificial intelligence that make sense of the big data as well as ensure reliability of results associated with the data so that the skepticism brought by veracity for the management can be sufficiently diminished. Also for veracity, apache mahout, machine learning in R and python are particularly useful platforms for filtering of data as well as clustering and classification of data into model points which greatly reduces having to focus on redundant data. Powerful visualizations have also covered ground to make intuitive human sense of the plethora of data and results available to us. These types of visualizations are far more flexible, representative and diverse than the usual spreadsheet based visualization tools. Instances of powerful visualization can be made through tableau, data driven documents, google interactive maps , Ggplot, Shiny and R Markdown of R and so on.

Moreover, behavioral finance has highlighted a number of ground breaking cognitive biases that human falls into; such as recency effect of seeing only the recent efforts and generalizing them to the long term anchoring which is relying on expert opinion or the primacy effect which is giving undue importance to first impressions. Anchoring has a special context in organizations which is relying unduly on the opinion of highly paid and powerful leaders in the company. Big data can clear up the clutter of such cognitive biases by introducing data-driven decision making into the company.

Even with robust sensitive checks in place to counter over-fitting, there is still a challenge around the concepts Game Theory and the Butterfly Effect of Chaos theory and their application in insurance analytics. Modelers need to at least partly take into account these reactions within strategies to maximize the benefit and minimize any potential damages. This includes more frequently updating analysis than before to assess if there are any discrepancies between model anticipation and reality. Assessment ideally needs to incorporate horizon scanning techniques and emerging risk assessment, which can be in-built into assumptions for stochastic analysis or the robustness of sensitivity analysis[5].

[5] ASTIN Big Data/Data Analytics Working Party – Phase 1 Paper- April 2015

An excellent illustration by Rubens Zimbres captures the synergistic interaction between machine learning and big data and is shown as a mindmap as follows:



Key thing to note here is how big data goes from billions to less and less data which is more and more refined and useful. Machine learning is then applied onto the data to arrive at actionable insights. Key models of simulation mentioned here belong to the complexity science domain detailed before. As such, this mindmap presents a powerful consolidation of many algorithms by showing relationships and links between them, as these on their own, can seem confusing and dispersed.

Along with the server clusters provisioned in company data centers or leased from virtual cloud environments, the software to deploy and manage Big Data application environments is a crucial element. The complexity of deploying "Big Infrastructure" clusters has been somewhat lessened by a new generation of open-source software frameworks. The leading solution is Apache Hadoop, which has gained great popularity due to its maturity, ease of scaling, affordability as a nonproprietary data platform, ability to handle both structured and unstructured data, and many connector products[6].

Until now, cluster management products have been mainly focused on the upper layers of the cluster (e.g., Hadoop products, including the Hadoop Distributed File System [HDFS], MapReduce, Pig, Hive, HBase, and Zookeeper). The installation and maintenance of the underlying server cluster is handled by other solutions. Thus the overall Hadoop infrastructure is deployed and managed by a collection of disparate products, policies, and procedures, which can potentially lead to unpredictable and unreliable clusters.[7]

Insurance providers are looking beyond algorithmic ratemaking techniques that are claim-centric, to ones that are person–centric. These techniques focus on analysing policyholder behavior across claims, providers, and other sources of information (e.g. how many similar claims were submitted by the same individual, reported by the same individual), and extend to data sources beyond the firewall to analytics based on external information (e.g. cohort analysis - using a person's social graph to look for similar activities among connected individuals), and considering networks of people rather than just individuals.[8]

This person-centric approach requires integrating information across all providers involved in a claim, including counter-parties as well as partners (e.g. auto repair shops) requiring the schema-agnostic approach to data management mentioned earlier. Even when all the data lives within the firm, the agility provided by this approach makes it much more feasible to turn that data into useable information[9]. Telematics is the leading instance of ratemaking using such person-centric approach.

Big Data Application Case Study[10]

This section highlights big data application on handling liability catastrophes which does not let insurers have a good night's sleep.

Liability catastrophes are especially rare, high impact trends that are very difficult to know in advance. Even when the evidence starts becoming more and more corroborative and certain, the insurers may not collect adequate premium as actuarial modeling using historical claims data for ratemaking. This traditional way of ratemaking on historical data can mean that insurers are potentially underwriting their own graves and selling products features that will become their own gravediggers. Asbestos is the most famous liability catastrophe that has resulted in USD 85 billion in claims in US alone and bankruptcies in 73 insurers until 2004. Consequences of liability catastrophes cab include bodily injury, property damage or environmental damage Commercial general liability insurance covers such liability catastrophes usually.

[6] StackIQ white paper Capitalizing on Big Data Analytics for the Insurance Industry

[7] Ibid

[8] Bharal, P and Halfon, A. ACORD and MarkLogic (2013). Making Sense of Big Data in Insurance

[9] Ibid

[10] Lloyds and Praedicat, Innovation Series 2015. Emerging Liability Risks: Harnessing big data analytics.

To improve our chances of collecting adequate premium so that insurers do not go bankrupt when a new liability catastrophe arises, big data tools with machine learning algorithms focused around emerging risk approach is being utilized. Framework of emerging risk is more fruitful for liability catastrophes because the impact of new technologies can be much more complex than expected, and it can take many years before the broader consequences are brought onto the surface. When a liability catastrophe does occur, the products or business practices involved are usually discontinued and the companies usually bankrupt. Hence, each new liability catastrophe is likely to happen in a different industry and in a different way. Claims data therefore characteristically cannot be used to predict the next liability catastrophe, and this presents an open challenge for actuarial modeling.

Although the world has transformed since asbestos litigation emerged, the interplay between science, technology-driven-innovation and risk which can drive the accumulation of exposure has certainly not budged. Three prominent examples of emerging risks for current times are mobile telephones causing brain tumors, hydraulic fracturing causing earthquakes and nano-materials cause lung damage, range of plastics causing endocrinal damage causing autism and obesity and so on.

The main application is text mining in a big data environment. Text parsing, sentiment analysis, opinion mining and natural language processing using deep learning algorithm can collate, identify, catalogue and track the massive universe of resources available on the internet in a data driven manner to identify key emerging risks. While the risks are diverse and too many, the means through which scientists establish causation are common to all of them. Each new hypothesis published in the scientific literature sits somewhere along the road to establishing causation, and it is possible from the algorithms mentioned here to estimate whether causation will ever be established based on the current pace at which the research is progressing.

Moreover, the data mining as well as exploratory analysis with unsupervised machine learning algorithms of scientific literature can map these emerging risks with a particular insurer given its unique portfolio which can be deduced from public sources as well now. Some companies will be located on none of the emerging risks, and some will be located on many of them – and the map itself will vary according to what is on each insurer's list of emerging risks. Assembling companies into a portfolio and then adding up the risks across companies provides a possible method to measuring portfolio accumulation for a particular emerging risk.

These trends, mappings and probabilities can then be combined with quantitative estimates of mass litigations if these emerging risks were to occur. This show expected costs and these can vary across different industries, companies, regions and portfolios. This can serve as a vital guide for ratemaking for liability catastrophes and inform insurer on a number of key areas such as product specifications, list of exclusions, maximum sum insured, reinsurance arrangements, loading on premiums for liability catastrophes and so on. With historical analysis, insurers rely on scientific skeptics' approach where they are not convinced of liability catastrophes until they actually surface. This is a self-defeating approach and focusing instead on pragmatism and precautionary principle can serve the solvency interests of the insurers better. Big data tools with machine learning algorithms in an emerging risk framework can better aid risk-adjusted ratemaking of emerging liability catastrophes and apply the precautionary principle to work here.