General Insurance Seminar

Insuring Tomorrow

17-18 November
Hilton Sydney

Big data in insurance and its impact on the actuarial profession

Prepared by Dimitri Semenovich, Solai Valliappan

Presented to the Actuaries Institute
General Insurance Seminar
17 – 18 November 2014
Sydney

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9233 3466 **f** +61 (0) 2 9233 3446
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Big data in insurance and its impact on the actuarial profession

Dimitri Semenovich, Solai Valliappan

Abstract: In this paper we discuss some of the organisational practices developed by consumer technology (web) companies and examine their implications for the insurance industry. In particular, we argue that that the current industry view of "big data" is perhaps too narrow and suggest some further directions. The impact of the technological change on the actuarial profession is also considered.

Keywords: Big data, technology, analytics, experimentation, product design.

Acknowledgements: Authors thank David Whittle, James Patterson, Richie Haynes and Ben Lever for helpful discussions.

1 Introduction

While insurance can be viewed as a mechanism[1] for consumption smoothing via cross-sectional risk transfer within a group of economic agents, most concrete implementations amount to centralised information processing systems of varying degree of complexity. Insurance companies typically operate systems that capture details of exposure periods and facilitate claims processing with additional modules for pricing and risk management, accounting, payments and so on.

It is becoming increasingly clear that this infrastructure can be dramatically simplified if implemented on top of current web technologies (clusters of commodity PC hardware, running a combination of in-house and open source software). We argue that the current interest in "big data" can be seen as a reflection of this broader trend, and indeed "big data" movement itself can be usefully defined as the application of technological practices that have emerged at web companies to the more established areas of enterprise. It might be worthwhile to examine the implications for the insurance industry in several further aspects.

One underlying theme is viewing the organisation as more software centric, with sharp focus on programmatic "interfaces" for major services and the ability to uniformly query (subject to appropriate controls) all the data generated in the course of operations. There are many recent startups that have achieved global reach with no more resources than an engineering staff of less than 50 employees[2], strongly suggesting that dramatic gains in operational efficiencies across a broad range of information processing tasks can now be realised (expense ratios in the order of

[1] http://en.wikipedia.org/wiki/Mechanism_design

[2] E.g. at the time of USD 19 billion acquisition by Facebook earlier this year, WhatsApp employed 30 engineers, had over 400 million active users and processed 50 billion messages per day.

5% for a typical insurance operation are potentially within reach). Beyond cost savings, such a configuration can enable a new level of innovation in product design and distribution, at present to a large degree held up by challenges in implementation and legacy systems.

Another theme, and to some degree building on the operational capacity, is "analytics" — at the operational level it is the ability to effectively and objectively evaluate both pricing and product structures and competing approaches to customer experience. Active experimentation, particularly in the personal lines space, is a key tool to obtaining reliable insight into these questions, going beyond limited price testing and A/B testing variations in the design of the web frontend, that have been trialled in some markets.

Finally, the ability to leverage additional data sources (web, telecommunications, GPS, bank transactions etc), not directly available to traditional carriers, is quite likely to become essential. Census type datasets that are frequently updated and provide (probabilistic, deidentified) matching on many dimensions in addition to the residential address are likely to become the norm.

We challenge the incumbents' ability to quickly transform and adapt to the emerging environment while keeping to their acceptable risk envelope.

2 The industry view of "big data"

The current insurance industry focus of "big data" projects is largely on analytics infrastructure — information from production systems (web servers, policy and claims management, finance systems etc) is transferred in raw form into so called "data lakes" with the goal of subsequent "insight discovery".

In this sense much of the technology is a direct successor of the earlier generation of the "business intelligence" (BI) or "data warehousing" solutions, with the key difference being the abandonment of the fixed predetermined database schemas. Traditional BI architecture presupposed certain formats and relations to which all data was compiled, striving to present a "single source of truth" in one materialised data set.

Current big data tools (Hadoop, Spark etc.) replace this approach with computation - data views are not predesigned but are an output of a program run over the entire history of source system extracts. This approach is enabled by utilising clusters of relatively cheap commodity server hardware[3] and ideally ensures that no information is lost due to imposition of a schema and it is always possible to answer any (unanticipated) query addressable by historical data. This dramatically reduces both the upfront costs of data "ingestion" and transformation as well as making sure that the resulting system is potentially useful to many stakeholders in the organisation, even those who have not been the key focus in its design (it is a common experience, for example, that data warehouses developed by insurers have turned out, for various reasons required actuarial teams to continue operating some of their own independent processes to meet their pricing and valuation needs).

This approach has the potential to dramatically simplify many of the reconciliation, reporting and model building activities, as all of the enterprise data can be collected on a single "computational substrate".

[3] Cost reductions of two orders of magnitude per terabyte relative to vendor BI solutions are sometimes claimed.

Another commonly cited benefit is the ability to construct a unified view of individual customers' interactions with the company, records of which may be split across multiple systems. The data can then be used to both improve risk models and in some cases derive insights around other aspects of customer behaviour. This is one area where diversified market participants, providing consumer services outside insurance, are at a clear advantage relative to traditional carriers.

2.1 Limitations of the industry view

While the above is a compelling story, it ultimately does not capture the standard operating practices of web companies, where much of big data technology has originated.

Consider the typical online quoting process for a personal motor policy — the only interaction the customer has with the insurance company in this case consists of being presented a sequence of web forms. Who within the company is responsible for the overall customer experience? For some insurance companies the responsibilities may be separated as follows:

- A product team is responsible for policy options and associated wordings - in the online world this translates into available check boxes and sliders on the quote screen.
- Pricing function is responsible for the actual quote amount displayed for a particular product configuration.
- Design, form layout and flow may be handled by a dedicated "channel" team.
- Banners or cross sell offers may be managed by the marketing function
- Underwriting may have input into what information is collected as well as business rules for generating referrals for manual processing.
- Finally, IT function would be responsible for the integration of the web front end and the "core" policy admin system.

While this structure is readily understood in historical context, it is also unclear who is ultimately accountable for the customer experience and any substantial change typically involves interdepartmental coordination which can further complicate or delay the process. In a modern web company, all of these responsibilities would be handled by a single "product" team, where "product" is not a particular policy wording but rather the software artifact that generates the customer experience with product options, wordings and prices all integral parts of the whole.

This potential lack of focus and agility leads us to the key challenge faced by any established insurer seeking to adopt a "big data approach": data captured from various systems will be by its nature observational — generated in the course of normal business operations. Only limited insight can be obtained from observational data — for example, it is generally straight forward to estimate risk premium for a new cohort of business based on the history for a comparable book, but much more difficult to answer more pertinent questions around the impact of a proposed rate change on the loss ratio. The latter requires a model of demand elasticity, which is not identified[4] without active intervention or external shocks (i.e. changes in competitors' prices). The same applies to many business questions around the product offering and marketing strategies — few of them can be answered with any degree of credibility by analytics on historical data alone, with limited live market testing often the only real alternative.

[4] http://en.wikipedia.org/wiki/Parameter_identification_problem

This is indeed the key lesson to be adopted from the consumer tech companies - Google, for instance, runs hundreds of parallel experiments on its search product alone[5]. Analytics is vastly more effective in identifying the most effective intervention for a given customer according to a prespecified metric or suite of metrics, based on a market test, rather than uncovering "customer insights". It appears likely that to fully take advantage of analytics and big data technology insurers will need both an increased rate of product innovation (supported by changes to organisational structure) and a much higher degree of flexibility in front and back end software systems, enabling parallel data driven evaluation of a large range of product options and customer experiences.

3 Taking full advantage of big data

We have earlier proposed to define "big data", perhaps somewhat vaguely, as applying tools and operating principles developed in the consumer technology industry to the traditional insurance operations. In this section we explore some of the possible implications of following through with this premise.

3.1 Software focus as a foundation for the business

As already stated, insurance companies operate systems that capture details of exposure periods and claims processing with additional modules facilitating pricing and risk management, accounting, payments and so on. From the technology point of view it is by no means inconceivable to have new product options rolled out on a weekly release schedule or various forms of pricing, valuation or risk accumulation analyses carried out effectively instantaneously, updated for every bound or expired risk. Similarly it should be possible to carry out market testing for hundreds of narrowly targeted offers (subject to controls and relevant regulatory frameworks) at the same time, each potentially altering what are today seen as immutable aspects of product design.

Whether the above set of capabilities is essential to the the future success of the insurance enterprise is of course up for debate, but it will likely form the core advantage available to any emerging technology centric competitors.

Achieving sustained product innovation and effective utilisation of digital channels (it is likely that in the medium term there may remain no other significant channels, at least in personal lines) will require the elimination of existing divide between business and IT in favour of the model almost universally adopted by consumer technology firms, where the main division is between "platform" and "applications", both software centric.

Platform teams are usually responsible for both hardware and software that constitute computational "fabric" for applications; applications are then what essentially replaces "products", "channels" and analytics functions, with domain experts embedded within relevant teams. Majority of applications would be developed internally, predominantly assembled from open source[6] components. Collaboration between completely unrelated organisations enabled by the open source

[5] http://research.google.com/pubs/pub36500.html

[6] http://en.wikipedia.org/wiki/Open_source

movement brings a large class of custom infrastructure projects within reach of companies who are not primarily technology vendors.

3.2 Active experimentation

Most questions about the effect of company's potential actions on customer experience can not be answered reliably purely through analytics on historical data, collected before the capability to execute these actions has been implemented in the production systems. Experiments or live market testing have long been recognised[7] as one of the few reliable approaches to gaging the effects of interventions — whether it be a rate change, introduction of a new product option, new direct marketing campaign, decrease in the page load delay on the main quote screen or changes to website design and layout[8]. While having all the data available in a centralised store would be helpful for the analysis, in all of these cases little can usually (but not always[9]) be done before some data has in fact been generated through a limited scale implementation of each proposed change.

Ability to define and rapidly implement complex changes to production systems to carry out market testing hinges on clear product ownership and in-house software expertise already alluded to — some companies still find challenging even parallel deployment of several rating structures, let alone any deeper modifications.

Furthermore, it is important to think of ways to measure effectiveness of each proposed change well before a test is carried out, quoting R. A. Fisher: *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of"*.

In particular, this means that some formal metric needs to be defined that can be estimated in a relatively short period of time and would affect the required size of the test — it could be conversions, click through rates, retention, net promoter scoring and so on. It is appealing to settle on a single organisation wide metric, such as the "customer life time value" (CLV), but extreme care needs to be taken how such a metric is operationally defined. CLV in particular is fraught with many difficulties, not the least, being dramatically affected even by minor variations of assumptions, all highly uncertain (in this and other regards it is not dissimilar to Margin on Services (MoS) accounting in Australian life insurance). Some blend of short term revenue and long term engagement metrics might prove much more preferable.

Discussion so far has presupposed that a single alternative needs to be chosen as an outcome of a market test — this is generally not the case, any variable or combination of variables can be used to define customer groups that react favourably to each of the options being tested, leading to customer level "customisation". A simple regression based approach to this problem is sometimes known as "uplift modelling"[10].

---

7 J. Agrist, J. Pischke, Mostly Harmless Econometrics: an empiricist's guide, PUP, 2009

8 http://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

9 http://en.wikipedia.org/wiki/Quasi-experiment

10 http://en.wikipedia.org/wiki/Uplift_modelling

3.2.1 Limitations of the experiment driven approach

While rapid prototyping and live market testing are almost universally practiced by successful web companies[11] and are often preferable to the standard practices of collecting subjective focus group feedback and slow moving pilot programs, these methods are not without their share of limitations:

- There are attribution problems with "above the line" advertising (e.g. billboards).
- It can be often hard to agree on the metric to be optimised and to deal with the time delays in its measurement (e.g. claim development).
- The main focus is on a large number of small incremental improvements, unlikely to lead to breakthrough innovation.
- Some ideas can be altogether too expensive to market test, requiring more manageable proxies.

3.3 External Application Programming Interfaces (APIs)

Technology also enables to "outsource" large parts of operations, with majority of communications between handled through APIs (application programming interfaces). In particular, aspects of distribution and product design and can be modularised in this way by exposing programmatic interfaces for policy creation and management. It is perhaps worth mentioning that the current mainstream adoption of the "cloud computing" paradigm is in no small part due to Amazon's early commitment[12] to modularise its application infrastructure and then expose the internal interfaces to the public, gaining the ability to sell excess capacity.

Similar idea in insurance has a long history in the guise of underwriting agencies, where insurers provide a recognised brand name and deal with risk management, regulatory compliance and back end systems, while product design and distribution are handled by more nimble external partners. Communication through APIs can both empower these partners to innovate and mitigate the risks for insurers.

Availability of APIs can also allow for substantially increased flexibility in product design and help to lift the level of engagement with the customers. One example might be the ability to turn on or off certain product options on demand via a mobile app eg. a customer going overseas may purchase an extension covering foreign medical expenses and at the same temporarily remove own damage cover from their car.

3.4 New approaches to enterprise systems architecture

The Conway's law[13], frequently cited with respect to large scale software projects states that *"organisations which design systems ... are constrained to produce designs which are copies of the communication structures of these organisations"*. The reverse is also potentially true — the design of (software) systems has the potential to affect organisational structure.

[11] http://www.exp-platform.com/Documents/2014%20experimentersRulesOfThumb.pdf

[12] http://apievangelist.com/2012/01/12/the-secret-to-amazons-success-internal-apis/

[13] http://en.wikipedia.org/wiki/Conway%27s_law

Existing "big data" solutions seek to gather information from a multitude of production systems for later analysis. They do not directly help to address the challenges of legacy systems integration or making rapid changes to interdependent production systems, the latter being the critical component enabling rapid and cost effective market testing.

We can speculate that the evolution of enterprise software is in the direction where the "big data" approach is gradually extended to incorporate majority of the production systems. Instead of multiple application specific databases and peer to peer communications between systems, all enterprise software could be developed against a single distributed transaction log and a central data store. The data store will be "immutable", which means that the data can not be overwritten but only invalidated, enabling reconstruction of the entire system state at any point in the past, a key requirement for both analytics and "systems of record". Furthermore immutability readily enables "branching", the ability to run multiple applications versions in parallel without directly affecting each other, dramatically simplifying rapid market testing of new features.

"Datomic"[14] is one recent database system with architecture following the above approach.

4 Additional sources of customer information

When it comes to data collection, the current (or near future) situation is well summarised by David Weinberg's dictum: *"Everything is a sensor for everything else"*. And indeed, much of "big data" discussions in insurance seem to concentrate on leveraging additional data sources for better risk pricing and selection. This is clearly an important topic and an efficient implementation can give an advantage to early adopters — whether the advantage persists depends largely on exclusivity and costs of access to the data source in question. This data is not necessarily "big"' as only a single variable[15] can at times substantially influence premium rating. It is also important to keep in mind that risk selection is not the only possible application — same information can also be used in the context of "uplift modelling" mentioned earlier.

One category of additional data is that collected through some automated process by insurance companies themselves, either through changes to product design (telematics most readily comes to mind) or utilising more fully the information being gathered as a byproduct of everyday operations (such as weblogs, including tracking via cookies or browser signatures).

Another category would be the information on consumers and SMEs collected through the provision of services not directly related to insurance: ISP logs, call records, supermarket point of sale data, credit card and bank transactions. These can be made available either through horizontal integration (bancassurers, financial services arms of supermarket brands) or through third parties providing anonymised matching services. It is not difficult to imagine arrangements in the latter case that appear compliant with the letter of the current privacy regulations.

All of these situations, benefit from a centralised "data lake" as described earlier, if not hosted by the insurer then by the third party providing "data enrichment" services.

How can this information be leveraged in practice? Additional data sources can be used to define a large assortment of customer "features" such as "count of sharp decellerations over last month", "number of times the cellphone was registered by a given cell tower ID in the preceding

[14] http://www.datomic.com/videos.html

[15] J. Monaghan, The impact of personal credit history on loss performance in personal lines, CAS Forum, 2000

6 months", "spending on white bread as proportion of total supermarket basket". These together with traditional risk factors can be in principle (the evidence is limited — credit scores are known to be effective, as well as obvious telematics derived driving style characteristics) used to build better risk cost or demand models as well as to augment analysis of any experimental data.

The inclusion of thousands if not hundreds of thousands of extra variables may seem daunting and intractable. However modern statistics and machine learning offer both theoretical and computational tools that can operate in this setting. For example free open source software[16] developed at Yahoo and Microsoft Research can deal with generalised linear model containing billions (!) of variables. While many challenges remain, it is a matter of only short time before the required expertise becomes relatively wide spread.

4.1 Markets in consumer data and integration with online advertising

The above picture, however, is far from complete — for one, it says little about online advertising and the associated ecosystem, powering all of the internet giants and indeed the source of the current wave of "big data" innovation. It is not difficult to realise that substantial friction is introduced by "third parties" (e.g. credit bureaus) when it comes to data sharing between "offline" companies in possession of data capturing important aspects of customer behaviour. The online advertising industry has left that model far behind, and is most reminiscent today of high frequency trading. So called real time bidding[17] infrastructure enables automated auctions to be carried out for every ad impression appearing on popular websites in the milliseconds the page takes to load. These auctions can in turn spawn side auctions for consumer data, such as browsing history gathered through tracking networks — one can readily imagine a future where data currently locked up in the "old world" companies is gradually brought online through the same mechanism. While there are clear economic pressures driving towards this model, its social desirability remains less than certain, at it can only be hoped that robust public debates will take place.

5 Impact on the actuarial profession

Historically, actuaries have been in a unique position in insurance, as a group with a strong appreciation for the links between technological, mathematical and consumer aspects of the business.

There is still an opportunity to capitalise on this, provided that there is a greater focus on modern technologies and approaches to analytics in education and CPD programs. Enhancements could include computing or mathematical optimisation and even more esoteric yet increasingly important topics like mechanism design. The difficulties of education reform and one possible approach have been outlined elsewhere[18] – any systematic change, however, is bound to take time and for the moment it is down to individual members to try to keep up with technological developments and where possible to take the lead. Broad perspective on technical and business issues could make actuaries well suited for leadership roles in the new regime.

[16] **Vowpal Wabbit** http://hunch.net/~vw/

[17] http://en.wikipedia.org/wiki/Real-time_bidding

[18] http://actuaries.asn.au/Library/AAArticles/2014/Actuaries191JULY2014p22t25.pdf

Failing this, the current trends could see the profession relegated into the role of fairly narrow compliance specialists. It is also important to keep in mind that regulatory protection is proving much less effective than originally anticipated by many industries (e.g. taxis, hotels) with "disruptive" technology centric entrants (Uber, AirBnB etc). While the situation is not directly comparable with insurance, and especially the Australian market, it provides some insight into possible scenarios globally.