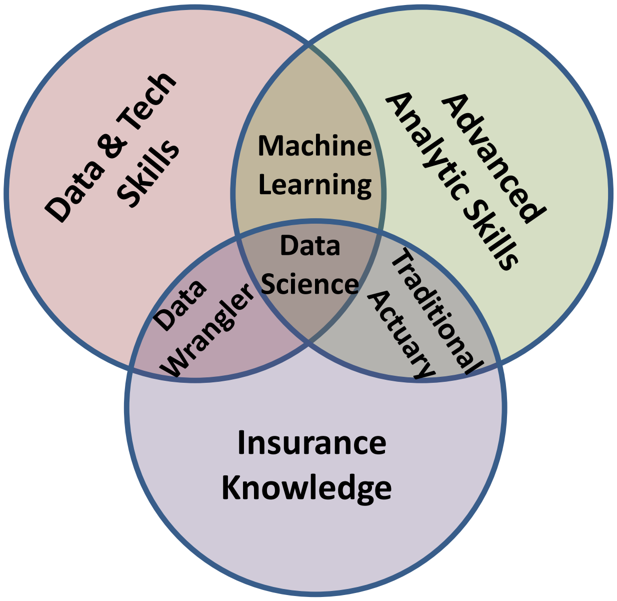
PREFACE

The evolving definition of Advanced Analytics and the emergence of the Data Scientist

In its infancy, Actuarial Science operated at the leading edge of contemporary analytic capabilities and could be easily said to be employing "advanced analytics." Over the past 50 years, however, relentless data and technology breakthroughs have created modern analytic capabilities that far outstrip many of our traditional actuarial pricing and reserving methodologies. The role of "data scientist" has emerged as the holistic practitioner in advanced analytics. The Casualty Actuarial Society (CAS) has begun to address the need to update our methodologies with recent predictive modeling additions to the syllabus, but to function as data scientists, we still need additional data and technology capabilities as well.



Working effectively with Information Technology is key to advancing the goals of the Insurance Industry

Similarly, during this revolution of data and analytics capabilities, information technology (IT) departments and vendors have embraced "data & analytics," "big data," and "data science" as the new frontier for informed decision-making. In the P&C insurance industry, the CAS actuary is uniquely well-positioned to partner with IT to advance the potential of these disciplines to benefit the industry. In order to be a participant in the conversation, however, the actuary must have knowledge of the language, practices, tools and techniques of the technology supporting this revolution.

Data and Technology Narratives

Pursuant to three aims, namely:

- to introduce actuaries to concepts critical to the pursuit of data science;
- to encourage actuaries to take leadership/sponsorship roles in data governance;
- to familiarize actuaries with technical concepts important for working with IT professionals to evolve data-driven decision making in the insurance industry,

this collection of papers aims to address concepts that will inform the actuary on key terms and concepts underlying the data and technology disciplines. The ultimate goal of these papers is to identify the knowledge and skills actuaries must possess in order to participate in the changes brought about by rapidly evolving technology supporting data and analytics. The narratives are designed to provide brief descriptions of the key terms and concepts and then point to recommended publications that the reader should reference for a greater appreciation of the subject along with practical applications.

In order to apply structure and scope to the material, the key concepts were aggregated into four major categories: data science, business intelligence (BI), data quality, and databases. Although it is somewhat subjective what topics were assigned to which category, the categories align closely to the current usage of terms found within the P&C insurance industry.

The topics included in the major categories are outlined as follows:

"Data science" includes:

- A common definition of "data science"
- Other related sub-disciplines associated with the term "data science"
- Discussion on "big data"
- Mathematical modelling techniques
- Definitions of terms and recommended readings

"Business intelligence" includes:

- Business intelligence solutions supporting the actuarial process
- Description of current BI software tools and their application in an insurance company
- An actuary's role in the design and delivery of a BI project
- Definitions of terms and recommended readings

"Data quality" includes:

- A common definition of "data quality"
- Data management, governance and roles
- The use of metadata in various settings to control quality
- An actuary's role in the application of data quality best practices
- Definitions of terms and recommended readings

“Databases” includes:

- Comparison and contrast between a database and a data warehouse
- Actuarial considerations in the use of SQL and data tables
- High-level schematics and diagrams of data architectures
- Discussion on other structures
- Definitions of terms and recommended readings

The more you know, the more you know you don’t know

Despite the different tone and structure of each paper, it is important to note that there is considerable overlap of perspective and terminology between the four topics. In fact, the interdependencies between these disciplines is what makes for compelling questions and unlimited opportunities for innovative solutions. For example, the reader should consider the following questions upon completing the readings:

- How do the databases that support data science differ from those that support actuarial process business intelligence deliverables?
- How does one build a business case for additional investment in data governance/management when competing with the forces that promote speed-to-market product development goals?
- What will the emergence of “self-service BI” mean for data warehousing strategy?
- How much of your actuarial analysis assumes your organization uses the same form of a reference data element (e.g., state code)? How will the lack of agreed-upon reference data format impact your database, data quality, business intelligence, and data science decisions?

With more formal education and research on data and technology topics, CAS actuaries will be better positioned to compete for data science roles and to partner with IT to use the combination of technology and analysis to develop innovative solutions to both long-standing and emerging challenges in the insurance industry and, likely, beyond.

Data Science and Analytics

Tom Davenport published an article in the October, 2012 Harvard Business Review (HBR) titled "Data Scientist: The Sexiest Job of the 21st Century." He described the data scientist as "a hybrid of data hacker, analyst, communicator and trusted advisor." What happened? Wasn't Actuary the best job in America not long ago? So what is data science, what makes the data scientist different and why aren't actuaries ranked at the top anymore[1]?

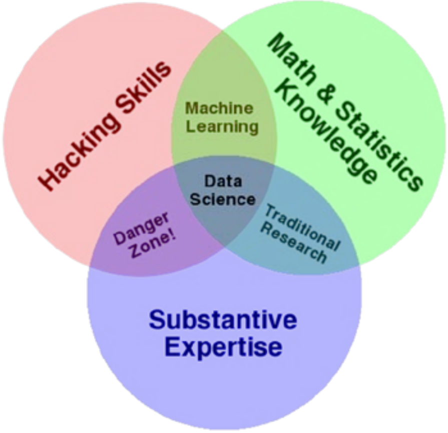
What is Data Science?

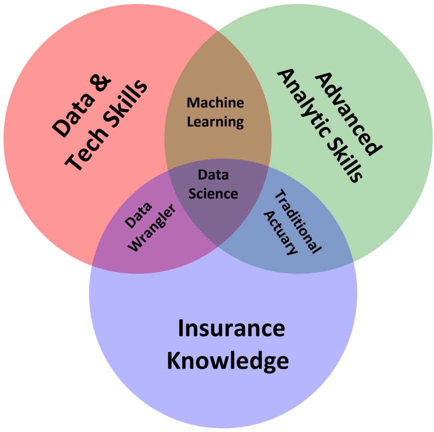
Let's start with a **definition** of data science. This is, unfortunately, not an easy task. There isn't an established professional or academic body to provide such a definition.

- Wikipedia defines it as: "an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases."
- TDWI asserts that data science "joins together contributions from several fields, including statistics, mathematics, operations research, computer science, data mining, machine learning (algorithms that can learn from data), software programming, and data visualization. It can cover the entire process of acquiring and cleaning data, methods for exploring the data and extracting value from it, and techniques for making insights actionable for humans and automated processes."
- Drew Conway provided[2] a popular view of the skills needed to be a data scientist using a Venn diagram, shown in the left panel of the figure below. For the purposes of this paper, we have created our own Venn diagram with labels that may be more familiar to the actuarial community, shown in the right panel of the following figure.

[1] In a recent careers survey, data scientist ranked first while actuary was at number ten: http://www.careercast.com/jobs-rated/jobs-rated-report-2016-ranking-200-jobs

[2] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram





Conway posited that like Actuarial Science, Data Science is an empirical science. However, the difference between a traditional actuary and a data scientist is the addition of what Conway called "hacking skills," namely "being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically" More generally in this context, "hacking" can refer to data acquisition and transformation at scale together with coding expertise required to implement production ready prototypes of the mathematical models.

In popular use "hacking" carries pejorative connotations but its intent here is to indicate a certain degree of fluency in dealing with programmable systems[3].

While "data science" has initially emerged as a label for analytics at web companies (Facebook and LinkedIn specifically), it is a reflection of deeper intellectual currents. A compelling account of the **history of data science** was recently given by a prominent academic statistician David Donoho [38], considering it in the context of the broader evolution of the practice of data analysis.

RELATION BETWEEN DATA SCIENCE AND OTHER ANALYTICS DISCIPLINES

It may at times seem difficult to differentiate between "data science" and more established fields of analytics. We believe that this is due to an increasing number of industries being affected by "digital transformation" – new business models facilitated by the ubiquitous availability of computing and telecommunication technologies. This "digital transformation" is to a large degree carried out by engineering / software-centric "web" companies following technology practices that have little overlap with traditional enterprise IT and adopting "data science" rather than traditional analytics.

[3] See the definition of "hacker" in the Jargon File: http://www.catb.org/jargon/html/H/hacker.html

Due to the widening front of "digital transformation," the scope of "data science" is also expanding. We are already quite close to it becoming an umbrella term for what historically have been largely disparate areas of applied mathematical modelling in the commercial setting. Some of these are listed below.

Online advertising and website optimization: Online advertising has grown into a massive ecosystem over the last two decades providing critical revenue for the majority of online services. The nature of the medium is eminently accommodating of tracking and analytics, resulting in one of the more dramatic applications of "data science." Most sophisticated solutions (e.g. AdWords) are deployed by inventory providers and aggregators, such as Google and Facebook. Live A/B testing is also prevalent among online businesses – something which is still a rarity in traditional enterprise. This is at present the biggest area of employment for "data scientists."

Manufacturing quality control, statistical process control, lean manufacturing, Six Sigma – this is an area of analytics supporting manufacturing activities and has been progressively developed since at least the 1930s. Among the main objectives is monitoring and elimination of variability in manufacturing processes (e.g., part dimensions), ensuring that defect rates are thereby controlled.

Operations Research, industrial engineering, revenue management, mathematical optimization, management science. Operations research began as a scientific study of military operations (e.g., convoy composition, bomber interception protocols, and logistics) during the Second World War and the principles have been exported to many other industries in the following years, in particular manufacturing, travel and transportation. Main tools include mathematical optimization and stochastic processes.

Statistics really requires no introduction; perhaps its main focus of interest in applications has been analysis of government data, polls and surveys and support for design of experiments and evaluation of experimental results in life sciences and medicine.

Applied finance, financial engineering, algorithmic trading, HFT, portfolio management. There are close parallels between data science and quantitative finance in the 1980s and 1990s. This is not surprising, because in-market execution is a key part of any model driven trading strategy, placing a premium on "hacking skills." At present it is perhaps reasonable to view the majority of "data scientists" as "quants" of digital advertising.

Engineering control, control theory, signal processing. Successful engineering applications of control and information theories span from fly-by-wire systems to cellular networks and synthetic aperture radar. Much less ambitious in scope than AI, these systems work reliably and are by now absolutely ubiquitous.

Econometrics, mechanism design, causal inference (from observational data) – due to the difficulty and costs of real world experiments in economics, econometricians have developed tools and conceptual frameworks for causal inference with observational data [39]. Furthermore mechanism design and the study of auctions have had significant impact on the design of online marketplaces.

Business intelligence, database / warehouse design, dashboards – business intelligence is primarily an IT led activity to support descriptive and diagnostic analytics. Business Intelligence will be given its own discussion in another section of this paper.

Machine learning, natural language processing, computer vision, data mining – machine learning is a branch of computer science that initially focused on more tractable aspects of **artificial intelligence**, primarily by constructing models from example data using statistical methods rather than designing them by hand from general principles. Two large application areas are computer vision and natural language processing, including machine translation. By now the differences between theoretical machine learning and statistics communities are largely superficial, amounting to little more than preferences for different styles of analysis of statistical procedures. Machine learning research has also provided many of the tools used in analytics for online advertising and algorithmic trading. **Data mining** has originated from the databases research community and has also mostly converged with machine learning in terms of both objectives and methodologies. Notably, there is a significant community of machine learning researchers working at technology companies who self-identify as such rather than "data scientists."

DATA SCIENCE AND "BIG DATA"

Data science has come to be associated with so-called "big data" – in this section we argue that "big data" projects that some insurance companies have undertaken are only tangentially related to the success of data science and instead the key lessons that insurers can derive from the experience of web companies lie in an integrated approach to product management and design and the adoption of live market testing.

"Big data" projects

The focus of "big data" projects in insurance and consumer finance to date has largely been on data processing 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 "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 materialized 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 utilizing clusters of relatively cheap commodity server hardware[4] 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 organization, even those who have not been the key focus in its design. (For example, general purpose data warehouses developed internally by insurers often turn out to have limitations that still require actuarial teams to operate their own independent processes to meet 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."

Another commonly cited benefit is the ability to construct a unified view of individual customer 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 behavior. This is one area where diversified market participants, providing consumer services outside insurance, are at a clear advantage relative to traditional carriers.

Virtually all "big data" technologies originated from the need to support analytics at web companies. However, it is important to note that these are purely enabling technologies and are not essential for data science itself except in situations where associated data processing tasks cannot be accomplished by other means. It is perhaps these common origins that have created an association between "big data" and "data science."

Analytics solutions at web scale usually need to address challenges around the so called "four Vs" of data (nomenclature predominantly adopted by IT vendors):

- velocity: data is gathered at an increasing speed;
- variety: data is gathered in a large number of forms and ways;
- volume: exponentially increasing volumes of data are being gathered;
- veracity: it becomes increasingly difficult to guarantee the quality of the data;

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

as well as satisfying certain operational properties:

- Automation: at scale, it becomes impossible to manually curate or even review models supporting operational decision. The entire pipeline needs to be fully automated. This is a challenging task if one wants results to be robust and credible.
- Speed of computation and algorithmic complexity of procedures involved come to the forefront with large volumes of data.
- Adaptability: special care needs to be taken in the design to allow adjustments to the analytics pipeline stemming from frequent changes to the front end systems.

It is important to remember that for most insurance companies with traditional product portfolios, issues relating to the scale of data are simply not present and data science solutions can be reasonably implemented on existing infrastructure, i.e., the link between "big data" and "data science" is very weak if it exists at all.

Limitations of "big data"

"Big data" technology is only a part of the solution to analytics-guided operational decision making - the standard operating practice of web companies. In this section we discuss another essential ingredient: live in -market testing.

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.
- Search engine campaigns directing traffic to the website are outsourced to a media agency.
- 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 administration system.

While this structure is readily understood in historical context, it is often 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.

Traditional organizational structure is a major obstacle faced by established insurers seeking to adopt a "data science" approach to product management as it generally hinders rapid in-market testing of different variations of customer experience. "Big data" solutions are of little help in this environment as data captured from various systems will be by its nature observational - generated in the course of normal business operations - but only limited insight can be systemically obtained from observational data. For example, it is generally straightforward 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 expected business volume. The latter requires a model of demand elasticity, which is not identified[5] without active intervention or external shocks (i.e. known 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, ultimately requiring in-market testing. We will revisit this in later sections.

Data quality considerations

Actuaries are quite familiar with data quality considerations when it comes to rate filing or reserving exercises and traditional data quality principles are discussed in detail in another section of this paper.

Different criteria, however, will apply when devising rules for operational decision making, e.g., choosing a particular version of an online quote form. While for a valuation missing data for 10% of policies would clearly be unacceptable, data missing (at random) for 10% of customers would have no significant effect on performance.

In the big data space, suitable determinations have to be made for each individual use case and it is at times necessary to significantly relax standards actuaries might be accustomed to. One issue specifically worth mentioning is the situation when the dataset used for analysis contains information not available at the time the decision needs to be made (e.g. due to a pre-processing step incorrectly incorporating knowledge of future transactions). This type of error has the potential to undetectably undermine the model validation protocol described in the following sections – underperformance only revealed once the model has been deployed in production.

OBJECTIVES OF DATA SCIENCE

At least one of the goals of data science is to bring rigor to optimizing operational decision making through integration of analytical and technological expertise. Additionally it seeks to incorporate rich new data sources such as text, audio, images and video into both analysis and decision making – this

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

latter objective is made possible by advances both in the costs of hardware and progress made over the last two decades on the associated pattern recognition tasks (e.g. [41]).

Some questions around operational decisions can be answered effectively by constructing models (see next section) of observational data gathered in the course of normal business operation, sometimes referred to as "predictive analytics."

Situations where "predictive analytics" are directly applicable are not universal – insurance premium rating happens to be one such case, fraud detection in settings where notifications are reliably received from injured parties is another.

A great discussion of the limitations of "predictive analytics" approach in the context of evaluating effectiveness of advertising is given in [42]. The paper shows that the critically important questions of causality (in their case sales uplift from a particular advertising campaign) cannot be answered reliably from observational data alone without randomized intervention to fully remove confounding. Randomized interventions on the production systems are in many cases the only known way to reliably estimate and therefore optimize the effects of operational decisions. Sometimes this is called "prescriptive analytics" although the term is often also used in engineering applications where system dynamics can be reliably estimated from general theory and do not require ad hoc experimentation.

An example of "prescriptive analytics" in the insurance context would be so-called premium optimization. Demand based premium adjustments, however, is just one form of intervention and the exact same framework can be applied to evaluating the color and position in which the quoted premium is displayed vs. any loading applied to the amount itself.

Indeed it is in the design, execution and analysis of live market tests of this type that technical expertise of a data scientist is often crucial, 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."

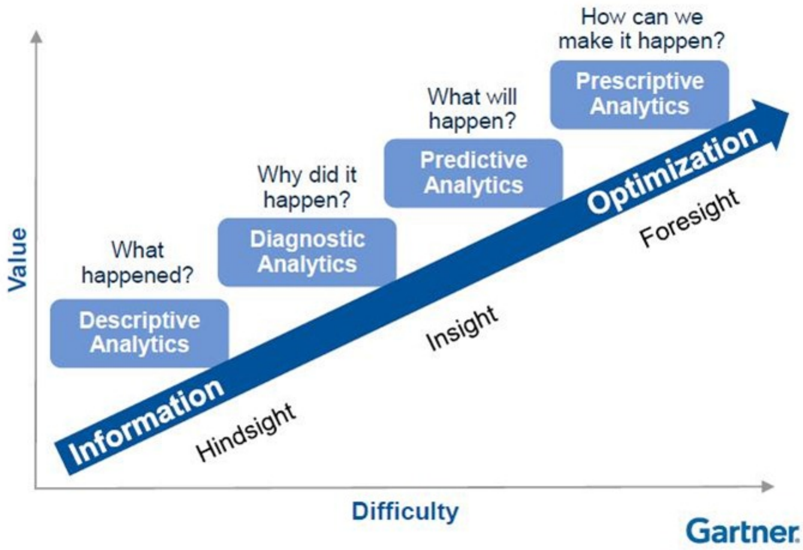
In practice this means that some formal metric needs to be defined that can be estimated in a relatively short period of time - it could be conversions, click through rates, retention, net promoter scoring and so on – as well as the size of the test or an appropriate stopping rule[6].

Design of large scale sequential experiments and analysis of resulting data is an active area of research in the machine learning community, with [43] offering the most accessible introduction to date. Pervasive testing is likely to prove the key analytics lesson to be adopted from the consumer tech

[6] https://en.wikipedia.org/wiki/Sequential_analysis

companies - Google, for instance, runs hundreds[7] of parallel experiments on its search product alone, as does Microsoft[8].

Business analytics is also sometimes said to follow a maturity model. While there are many sources with slight variations around the same theme, we have included the Gartner model in this paper.



Here we also see the progression from “predictive” to “prescriptive” analytics. While actuarial analysis is generally “predictive” there is considerable room for advancement when it comes to model validation, and live testing in insurance remains exceedingly rare.

METHODOLOGY: SOME SPECIFIC TECHNIQUES

Mathematical modeling techniques

There is dramatic variation in mathematical tools used in different areas of “analytics.” It is not infrequently observed that different groups of practitioners develop solutions that are operationally very similar, while diverging significantly in ideology and mathematical apparatus. Despite significant overlaps it can still be useful to consider the major approaches as they form major components of respective intellectual traditions.

Summary tables are pervasive in business reporting, and while this aspect is usually ignored, one must make implicit assumptions about the underlying data generating process in order to make inferences from such information.

[7] http://research.google.com/pubs/pub36500.html
[8] http://www.exp-platform.com/Documents/2014\%20experimentersRulesOfThumb.pdf

Basic probability and statistics are quite familiar to actuaries, perhaps the biggest gap being hypothesis testing, should experimental methodology become more widely adopted in the insurance industry. In particular, the general confusion between Fisherian, Neyman-Pearson and Bayesian points of view[9] in introductory texts make it difficult to acquire fluency in the correct application of standard methods.

Parametric conditional models – these include traditional regression tools, like generalized linear models, quantile regression etc. Multiple useful extensions have been developed, including regularization, random effects and additive models, all of which are closely linked to actuarial credibility. It is also sometimes possible to "predict" more complex objects than just a single dependent variable, such as part of speech labelling for an entire sentence (e.g. so called conditional random fields).

Dynamical systems – dynamical system models seek to incorporate the temporal aspects of the phenomenon or control process under consideration. These are particularly effective when the evolution of the system can be somewhat reliably predicted (e.g. on the basis of physical laws). Time series models are without exception special cases of dynamical systems.

Mathematical optimization – many inference problems, from testing to regression and beyond rely on solving optimization problems (e.g. maximum likelihood). Mathematical optimization studies both properties of such problems as well as computational procedures that can be used to find or approximate solutions.

"Model free control" – in many situations it is not possible to construct a reasonable model of the system to be optimized from general principles; this includes the majority of applications of data science in analysis and control of live experiments. Such settings necessitate joint estimation and control. For example, in online advertising, before a click through rate for a new ad can be estimated it has to be displayed a certain number of times in different contexts. Investigation of methods for doing this efficiently while simultaneously optimizing for an overall objective, such as revenue, is a central problem in "reinforcement learning," a subfield of machine learning.

Bayesian modelling – it is possible to consider most of the above settings from the Bayesian point of view. Notoriously difficult computationally, algorithmic advances (geometric MCMC, variational methods) and the availability of open source software make this approach tractable for a growing range of practical problems. Some aspects of Bayesian analysis are known to actuaries as credibility theory.

[9] R. Christensen, Testing Fisher, Neyman Pearson and Bayes, http://www.stat.ualberta.ca/~wiens/stat665/TAS%20-%20testing.pdf

"Non-parametric" conditional models – gradient boosting machines, support vector machines, much of "deep learning" or neural networks – these types of methods are most commonly associated with machine learning or data science. They extend the usual regression models by introducing non-linear dependence of output on the input variables while still maintaining the ability to control overall model complexity.

With the proliferation of specific analytical methods and implementations, it is important for actuaries to be able to place specific methods into a theoretical framework to evaluate their relative merits and specific applicability. For example it would be valuable to understand connections between actuarial credibility and penalized regression and ensemble methods developed in machine learning and computational statistics literatures [46].

There exists a multitude of such frameworks at various levels of abstraction. One particularly useful viewpoint is that of optimization [19] – it is generally very difficult to understand whether two statistical procedures are related, especially if they are presented in the form of algorithms. Understanding what objective function is minimized or maximized by a given procedure allows us to readily appreciate similarities between methods. As an example, it turns out that the maximum likelihood estimator for logistic regression is almost identical to the optimization problem solved by "support vector machines" popular in machine learning. Many examples of optimization models in premium rating are given in [44].

Algorithmic thinking

Increasing volumes of data brings to the forefront computational issues around mathematical modelling and data processing. Beyond certain problem sizes, algorithms with second or higher degree polynomial complexity simply stop working (i.e., they do not terminate in any reasonable time). Actuaries must be aware of this possibility and some common workarounds where they exist.

Finally, we should point out that all popular environments for cluster computing (Hadoop, Spark, etc.) impose significant limitations on the user in terms of how the computation needs to be structured relative to using a single computer. Understanding these limitations and how they can impact common tasks require both familiarity with distributed system architectures as well as the underlying algorithms.

Visualization and exploratory data analysis

Sanity checks on the available data and trying to understand how recorded observations relate to the generating process are the core activity in "data science" and indeed among actuaries. John Tuckey has referred to this "Exploratory Data Analysis" in his influential book [45]. This type of investigation is made particularly important when working with heterogeneous data originating from rapidly evolving systems.

Advances in theory [47] and computer software (e.g. ggplot2) have made advanced visualization [48] readily available.

Model validation

Data driven model validation has emerged as one of the central themes in "data science." These methods have seen relatively limited use in ratemaking to date due to the labor-intensive nature of model construction (model validation requires fitting multiple models to different subsets of data).

Basic model validation involves dividing the available modelling data set into three disjoint subsets:

1. Training – this is a (random) subset of the data used to construct successive iterations of the model.
2. Test – this (random) subset of the data is not used in model fitting but only to evaluate model performance. If a model iteration performs well (relative to all other model iterations as well as some known baseline) on the test set according to some formal metric, such as AUC or RMSE, that iteration is declared the winner. It is usual to consider dozens if not hundreds of iterations in a course of a modelling project[10].
3. Validation – this set of the data is withheld for a final validation of the model that passes the testing process. Often most reliable validation is, in fact, not a random subset of the data at all, but an "out of time" dataset that more accurately approximates live deployment.

This approach can then be naturally extended to multiple participants and has enabled steady progress in a number of applications of machine learning, particularly computer vision and natural language processing, with the more general framework as follows:

1. A dataset is made available (perhaps publicly) containing for each observation a value to be predicted (these can be numeric, categorical, or more complex structures altogether).
2. An objective function which the prediction rule or model is to optimize is communicated to the participants.
3. A referee who is able to evaluate models on a separate dataset whose objective values are not visible to the participants and report back the scores.

The goal of the participants is the construction of model which minimizes deviations from the objective values as reported by the referee. Beyond academic research, this is the mechanism that is used by Kaggle, a company that provides "crowd sourced" modelling solutions to companies willing to share their data publicly.

In the view of the authors, the key to success in applying "hard" data science to business problems is the creation of appropriate evaluation frameworks that can rigorously evaluate the quality of decision rules – sometimes historical observational data alone is sufficient (e.g., for models of claim costs) and sometimes live market testing may be required.

[10] To get more accurate estimate of out of sample performance when limited data is available, it is common to repeat the process over multiple training/test splits, e.g. so called "cross-validation."

DATA SCIENCE RESOURCES

In what follows we list some particularly noteworthy graduate and undergraduate courses that could help develop a broad fundamental understanding of computing and mathematical modelling. These could be argued to be core "data science" skills for addressing future business problems, with increasing number of processes and low-level operational decisions subject to automation. In compiling these resources we have intentionally stayed away from "flavor of the month" or introductory offerings, focusing instead on fundamentals.

Analytics at web companies

To get an impression of what the future of insurance analytics might look like, it is worthwhile to review some of the courses offered by people with experience implementing analytics solutions for the leading web companies. Examples include CS281B "Scalable Machine Learning" at UC Berkeley [25] by Alex Smola (formerly of Yahoo) and "Big Data, Large Scale Machine Learning" at NYU [26] by Yan LeCunn (currently at Facebook). In particular the first course offers an interesting insight into the importance of understanding systems, numerical methods and statistics to develop analytics solutions at web scale.

Prerequisites for this material include linear algebra, basic probability and statistics and, ideally, convex optimization and an introduction to machine learning, as discussed next.

Mathematical background and numerical computing

Numerical linear algebra is the most essential tool in applied mathematics. The majority of computational procedures for solving mathematical models ultimately reduce to iteratively solving systems of linear equations.

An excellent introductory treatment of linear algebra is given by Gilbert Strang in MIT 18.06 [2]. The material is further developed in MIT 18.085 [3] and 18.086 [4], demonstrating a very broad range of applications across engineering subfields. The observation that the differential operator can be discretized as e.g. a tri-diagonal matrix (the so called "finite differences" method) is the key connection between linear algebra, traditional calculus (in the form of integral and differential equations) and computing.

Another take on the material is given in Stanford EE263 taught by Stephen Boyd - in addition to basic linear algebra, the course gives a highly intuitive exposition to least squares regression, regularization, singular value decomposition and linear dynamical systems (which can be viewed as a generalization of a wide class of time-series models in the CAS syllabus). The material above should provide sufficient background to appreciate some of the technology behind modern robotics platforms, such as those formerly developed at Boston Dynamics, now part of Google (MIT 6.832

Underactuated Robotics [12]).

Finally, the Fourier transform is one of the most famous special cases of a linear operation– an intuitive introduction to the subject and its multitude of applications, including the Central Limit Theorem, is given in Stanford EE261 [17].

Optimization

Beyond differential equations, one of the main applications of linear algebra is in mathematical optimization or "mathematical programming." Optimization based models are pervasive in analytics, whether it be maximum likelihood estimation, "empirical risk minimization," Neyman-Pearson hypothesis testing, optimal control, Markowitz portfolio theory or option pricing.

Prof. Stephen Boyd's course EE364A Convex Optimization [19,20] not only gives a solid grounding in the theory but also considers many of the above-mentioned examples. Convex optimization is widely seen as the foundation of modern statistics, machine learning and signal processing. Familiarity with theory and algorithms will enable the practitioner to identify and implement solutions to a very wide range of problems across industries.

There is also an interesting connection between mathematical optimization and classical algorithms studied in undergraduate computer science courses (e.g. [7]) - many of the problems such as sorting, shortest path, max flow, etc. turn out to be special cases of linear programming (itself a special case of convex optimization).

The follow up course EE364B [21] provides more detailed background on scalable and distributed optimization as well as the clearest introduction to the General Equilibrium theory of microeconomics you are likely to find. The background for these courses is limited to linear algebra [2,18] and basics of multivariable calculus (gradient, Hessian) [1].

Probability, statistics, machine learning, information theory.

There are few unequivocally great introductory probability and statistics courses publicly available, at least at the moment. MIT 6.041 [9] is a useful probability refresher. A worthwhile follow up is MIT 6.262 [10] "Discrete Stochastic Processes."

When it comes to statistics, or at least a take on the topic that is more attuned to analytics applications, Stanford Statistical Learning [21] is a solid introduction from the authors of the well-known book. A closely related subject area is machine learning, with the introductory course by Andrew Ng [23] and a much more in depth treatment by Alex Smola [24]. So called "deep networks" are a recent "hot" topic in machine learning, providing state of the art performance for many recognition tasks. This material is covered in [27].

Information theory provides perhaps one of the most successful and widely used applications of probability. There are also important connections to statistics and machine learning (as efficient compression requires effective conditional probability estimation). MIT 6.450 "Principles of Digital Communications I" [11] is an excellent course by the pioneer of digital communications Rob Gallager, who invented one of the most effective known coding schemes and was a founding engineer at Qualcomm where he designed the first 9600 baud modem. Information theory is an essential foundation of all digital information processing technology.

Another excellent discussion of information theory is given in the course taught by David MacKay at Cambridge [36], bringing together topics from coding theory, statistics and machine learning.

Convex optimization provides a very helpful background for the courses in this section even if it is not explicitly alluded to.

Programming

There exists a very wide range of high quality introductory programming courses. Perhaps the Stanford sequence deserves a particular mention [15,16]. Alternatives include the introductory courses at MIT [5,8].

MIT 6.001 [6] (now superseded) is the most celebrated introductory programming course of all, with the textbook "Structure and Interpretation of Computer Programs" used in dozens of top universities. While Scheme, the language that it uses for teaching programming concepts, has for long time been considered less than practical, over the recent years there has been a dramatic resurgence of popularity of the related body of ideas called "functional programming," underpinning many of the latest "big data" technologies.

Beyond the introductory courses, "Programming Paradigms" [17] gives a useful overview of design choices behind a variety of programming languages and [31] offered on Coursera by the University of Washington, provides a more advanced grounding in the functional programming paradigm.

An introduction to Scala, an increasingly popular compatible replacement of Java, is available from its creator on Coursera [32].

No such list would be complete without an algorithms class [7]. Conceptual links with optimization or "mathematical programming" offer a connection back to the material in the earlier sections.

Finance, economics and social science

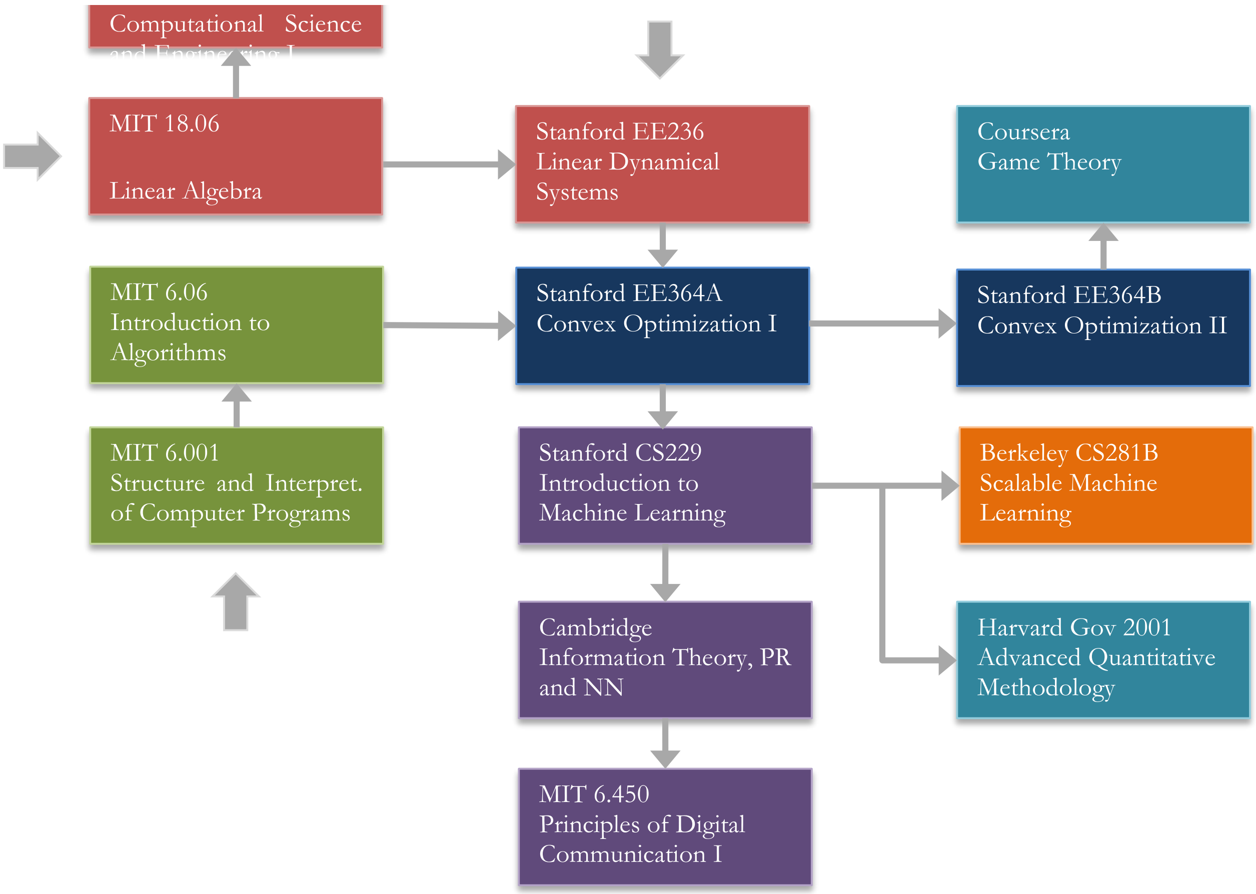
While the exact relation between actuarial pricing and financial economics is not clearly set out in the actuarial curriculum, it has been understood in the academic literature for some time as the so called "incomplete markets" setting. An introductory discussion of the modern theory of finance

(CAPM, option pricing, etc.) from this more advanced point of view is given in John Cochrane's (University of Chicago) class "Asset Pricing" on Coursera [30].

A useful generalization of the concept of an optimization problem (see e.g. Stanford EE364A [19,20]) is offered by game theory. Instead of considering a "central planning" problem where all the decisions are taken by a single agent, game theory looks at situations where there are multiple self-interested parties involved. Coursera classes [28,29] provide an introduction to a range of topics, including auctions and mechanism design. Applications of game theoretic methods to the study of social insurance, optimal taxation and related ideas are given in the Harvard course "Public Economics" [34].

Problems addressed by "business analytics" are not dissimilar to those found in the social sciences, especially when it comes to identifying what is sometimes called "actionable insights" - a social scientist may instead talk about "policy targets." While causal attribution is oftentimes not necessary, it is important to be aware of limitations of analyses carried out purely on observational data. One example in social science where large-scale experiments have been possible is "development economics." The MIT course 14.73 [35] offers an in depth discussion of considerations that go into designing a convincing experimental study. A broad introduction to the design of quantitative methods that are directly applicable to the question being studied is given in Gary King's excellent methodology course at Harvard [33].

Possible curriculum


Computational Science and Engineering I
MIT 18.06
Linear Algebra
Stanford EE236
Linear Dynamical Systems
Coursera
Game Theory
MIT 6.06
Introduction to Algorithms
Stanford EE364A
Convex Optimization I
Stanford EE364B
Convex Optimization II
MIT 6.001
Structure and Interpret. of Computer Programs
Stanford CS229
Introduction to Machine Learning
Berkeley CS281B
Scalable Machine Learning
Cambridge
Information Theory, PR and NN
Harvard Gov 2001
Advanced Quantitative Methodology
MIT 6.450
Principles of Digital Communication I

REFERENCES

[1.] MIT 18.02 Multivariate Calculus
http://ocw.mit.edu/courses/mathematics/18-02sc-multivariable-calculus-fall-2010/

[2.] MIT 18.06 Linear Algebra
http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/

[3.] MIT 18.085 Computational Science and Engineering I
http://ocw.mit.edu/courses/mathematics/18-085-computational-science-and-engineering-i-fall-2008/

[4.] MIT 18.086 Mathematical Methods for Engineers II
http://ocw.mit.edu/courses/mathematics/18-086-mathematical-methods-for-engineers-ii-spring-2006/

[5.] MIT 6.00 Introduction to Computer Science and Programming
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-00sc-introduction-to-computer-science-and-programming-spring-2011/

[6.] MIT 6.001 Structure and Interpretation of Computer Programs
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-001-structure-and-interpretation-of-computer-programs-spring-2005/

[7.] MIT 6.06 Introduction to Algorithms
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-006-introduction-to-algorithms-fall-2011/

[8.] MIT 6.01 Introduction to Electrical Engineering and Computer Science
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-01sc-introduction-to-electrical-engineering-and-computer-science-i-spring-2011/

[9.] MIT 6.041 Probabilistic Systems Analysis and Applied Probability
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-041sc-probabilistic-systems-analysis-and-applied-probability-fall-2013/

[10.] MIT 6.262 Discrete Stochastic Processes
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/

[11.] MIT 6.450 Principles of Digital Communications I
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-450-principles-of-digital-communications-i-fall-2006/

[12.] MIT 6.832 Underactuated Robotics
http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-832-underactuated-robotics-spring-2009/

[13.] UNSW COMP1917 Higher Computing
http://www.youtube.com/playlist?list=PL6B940F08B9773B9F

[14.] Stanford CS106A Programming Methodology
http://see.stanford.edu/see/courseinfo.aspx?coll=824a47e1-135f-4508-a5aa-866adcae1111

[15.] Stanford CS106B Programming Abstractions
http://see.stanford.edu/see/courseinfo.aspx?coll=11f4f422-5670-4b4c-889c-008262e09e4e

[16.] Stanford CS107 Programming Paradigms
http://see.stanford.edu/see/courseinfo.aspx?coll=2d712634-2bf1-4b55-9a3a-ca9d470755ee

[17.] Stanford EE261 Fourier Transform and its Applications
http://see.stanford.edu/see/courseinfo.aspx?coll=84d174c2-d74f-493d-92ae-c3f45c0ee091

[18.] Stanford EE263 Introduction to Linear Dynamical Systems
http://see.stanford.edu/see/courseinfo.aspx?coll=17005383-19c6-49ed-9497-2ba8bfcfe5f6

[19.] Stanford EE364A Convex Optimization
http://see.stanford.edu/see/courseinfo.aspx?coll=2db7ced4-39d1-4fdb-90e8-364129597c87

[20.] Stanford CVX101 Convex Optimization
https://class.stanford.edu/courses/Engineering/CVX101/Winter2014/about

[21.] Stanford Statistical Learning
https://class.stanford.edu/courses/HumanitiesScience/StatLearning/Winter2014/about

[22.] Stanford EE364B Convex Optimization II
http://see.stanford.edu/see/courseinfo.aspx?coll=523bbab2-dcc1-4b5a-b78f-4c9dc8c7cf7a

[23.] Stanford CS229 Machine Learning
http://see.stanford.edu/see/courseinfo.aspx?coll=348ca38a-3a6d-4052-937d-cb017338d7b1

[24.] CMU 10-701 Introduction to Machine Learning
http://alex.smola.org/teaching/cmu2013-10-701/

[25.] UC Berkeley CS281B Scalable Machine Learning
Slides - http://alex.smola.org/teaching/berkeley2012/
Videos - http://www.youtube.com/playlist?list=PLOxR6w3fIHWzljtDh7jKSx_cuSxEOCayP

[26.] NYU Big Data, Large Scale Machine Learning
http://cilvr.cs.nyu.edu/doku.php?id=courses:bigdata:start

[27.] NYU Deep Learning
http://cilvr.cs.nyu.edu/doku.php?id=courses:deeplearning:start

[28.] Coursera - Stanford/UBC - Game Theory
https://www.coursera.org/course/gametheory

[29.] Coursera - Stanford/UBC - Game Theory II: Advanced Applications
https://www.coursera.org/course/gametheory2

[30.] Coursera - University of Chicago - Asset Pricing
http://www.coursera.org/course/assetpricing

[31.] Coursera - University of Washington - Programming Languages
https://www.coursera.org/course/proglang

[32.] Coursera - EPFL -Principles of Functional Programming in Scala
https://www.coursera.org/course/progfun

[33.] Harvard Gov 2001 Quantitative Research Methodology
http://projects.iq.harvard.edu/gov2001/home

[34.] Harvard Econ 2450a
http://obs.rc.fas.harvard.edu/chetty/public_lecs.html

[35.] MIT 14.73 The Challenge of World Poverty
http://ocw.mit.edu/courses/economics/14-73-the-challenge-of-world-poverty-spring-2011/

[36.] Cambridge Information Theory, Pattern Recognition and Neural Networks
http://www.inference.phy.cam.ac.uk/itprnn/Videos.shtml

[37.] https://www.soa.org/library/research/transactions-of-society-of-actuaries/1990-95/1995/january/tsa95v475.pdf

[38.] D. Donoho, 50 Years of Data Science
http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

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

[40.] Gill Press, Forbes Contributor, A Very Short History of Data Science, 5/28/2013

[41.] Stanford CS231n Convolutional Neural Networks for Visual Recognition
https://www.youtube.com/watch?v=NfnWJUyUJYU&feature=youtu.be

[42.] B. Gordon et al., A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook
http://www.kellogg.northwestern.edu/faculty/gordon_b/files/kellogg_fb_whitepaper.pdf

[43.] Agrawal et al., Multi-World Testing: A System for Experimentation, Learning and Decision-Making
http://research.microsoft.com/en-US/projects/mwt/mwt-intro.pdf

[44.] D. Semenovich, Applications of Convex Optimization in premium rating, CAS E-Forum 2013
https://www.casact.org/pubs/forum/13spforum/Semenovich.pdf

[45.] J. Tukey, Exploratory Data Analysis, 1977

[46.] H. Miller and P. Mulquiney. Credibility, penalized regression and boosting; let's call the whole thing off, 2011
https://www.casact.org/education/infocus/2011/handouts/AM2-Fry.pdf

[47.] L. Wilkinson, The Grammar of Graphics, 1999

[48.] E. Tufte, The Visual Display of Quantitative Information, 2001

Business Intelligence Technology and Tools: A Primer for Actuaries

BUSINESS INTELLIGENCE DEFINED

While the term "Business Intelligence" is widely used, it doesn't have a single, widely accepted definition. However, several authoritative sources have each published definitions close enough conceptually to provide a good starting point for any discussion of the subject.

"Business Intelligence" has been defined in the following ways:

> "…a set of concepts and methodologies to improve decision making in business through the use of facts and fact-based systems"[1]
>
> "BI is neither a product nor a system. It is an architecture and a collection of integrated operational, as well as decision-support, applications and databases that provide the business community easy access to business data."[2]
>
> "Business intelligence encompasses data warehousing, business analytic tools and content knowledge management." [3]
>
> "…the ability to transform data into useable, actionable information for business purposes. BI requires:
>
> - Collections of quality data and metadata important to the business
> - The application of analytic tools, techniques and processes
> - The knowledge and skills to use business analysis to identify/create business information
> - The organizational skills and motivation to develop a BI program and apply the results back to the business"[4]
>
> *"…* an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies that allows actuaries: (1) to have interactive access (sometimes in real time) to data, (2) to manipulate data, and (3) to conduct appropriate analyses. The process of BI is based on the transformation of data to information, then to decisions, and finally to actions."5

INTRODUCTION

The purpose of this narrative is to introduce our actuarial audience to business intelligence terms and tools. The technologies involved are not all new, but are evolving as insurers use more and more computations and analytics to drive their business.

[1] Howard Dresner, Gartner Group

[2] Larissa T. Moss and Shaku Arte, Business Intelligence Roadmap, Pearson Edition, 2003

[3] David Loshin, Business Intelligence: The Savvy Manager's Guide, Addison Wesley, 2003

[4] TDWI Course titled, "Business Intelligence Fundamentals…"

[5] Sharda, Ramesh; Delen, Dursun; Turban, Efraim; Aronson, Janine; Liang, Ting Peng (2014-01-14). Business Intelligence and Analytics: Systems for Decision Support (Page 14). Pearson Education. Kindle Edition

Casualty actuaries have been concerned with data used in decision making since the beginning of the Casualty Actuarial Society (originally the Casualty Actuarial and Statistical Society of America) or CAS. Actuaries used computers as personal support tools to complete their work, but today business intelligence systems are coupled with data mining and automated processes to greatly expand an insurer's understanding of market trends.

When looking to improve their processes with increased use of the business intelligence landscape, the actuary needs to have a good working relationship with the IT[6] department who often maintains the data and tools used for the actuarial work. The nature of the relationship relies on the actuary having a clear idea of the business goal. Together they will determine what data and tools the actuaries will provide and what data and tools the IT department will provide as inputs to the actuarial work. The relationship and how they communicate together often determines if these tools will be successful.

Sources for further information

Source	Author	Publisher
Business Intelligence Roadmap	Larissa T. Moss and Shaku Arte	Pearson
Business Intelligence: The Savvy Manager's Guide	David Loshin	Addison Wesley
Business Intelligence Fundamentals	TDWI Course	TDWI
Business Intelligence and Analytics: Systems for Decision Support	Sharda, Ramesh; Delen, Dursun; Turban, Efraim; Aronson, Janine; Liang, Ting Peng	Pearson Education
Decision support and Business Intelligence Systems	Turban, Sharda & Delen	Prentice Hall
Understanding Actuarial Management: the Actuarial Control Cycle, 2nd Ed.	Bellis, Lyon, Klugman and Shepard. Ed.	Institutes of Actuaries of Australia and the Society of Actuaries

[6] This paper takes the perspective that the actuary is supported by an IT department within his or her organization. For those actuaries who do not have that benefit, "IT" can be also thought of as an external software or hardware vendor.

Source	Author	Publisher
Understanding Actuarial Practice	Stuart Klugman, Ed	Society of Actuaries
Deloitte – "How to Build a Successful BI Strategy"	Prashant Pant	Deloitte
Applied Insurance Analytics	Patricia Saporito	Pearson

BUSINESS INTELLIGENCE SOLUTIONS IN THE CONTEXT OF ACTUARIAL PROCESSES

Business Intelligence solutions are subject to data governance processes which provide for "an enterprise-wide data governance body, a policy, a set of processes, standards, controls, and an execution plan for managing the data."[7] While data governance processes control the flow and allocation of data as a resource, it is up to the actuary to determine how these resources are turned into actuarial work products. In order to ensure a smooth process of data-to-information and to determine the proper use of actuarial information, the actuary must have an active working relationship with the process, people and technology of their Business Intelligence landscape. This includes familiarity with the BI development/delivery lifecycle as well as the various capabilities of the larger organization's accepted BI toolset. Embracing with academic curiosity this decidedly non-actuarial discipline and developing these relationships can often be the critical factors in the success or failure of an actuarial department.

Actuarial resources, products and workflow must be managed like any other business activity. This activity can be described as the Actuarial process. This process can be broken into distinct components.

- Define the Problem
- Design the Solution
- Monitor the Results

Each of these components involves large and small tasks. Large tasks, like measuring and reporting annual statement loss reserves; and small tasks, like answering a regulator's inquiry about a rate filing, all require historical information. The source of required data can be internal to the corporation, such as that arising from policyholders and claimants, or external, such as economic and political data.

[7] Deloitte – "How to Build a Successful BI Strategy"

Business Intelligence Architects design and direct the flow of data into the Actuarial Process

BI solution architects and managers are charged with managing the corporate data as a resource. They are responsible for delivering:

- Faster computations
- Improved communications and collaboration
- Increased productivity of supported groups
- Improved data management
- Better management of actuarial data reservoirs
- Improved quality of decisions
- More agile support
- Increased cognitive limits
- Using the web
- Anytime, anywhere support

Optimally, the IT architects and the actuaries work together to decide on the best tools and data for supporting the actuarial process.

Sources for further information

Source	Author	Publisher
Decision support and Business Intelligence Systems	Turban, Sharda & Delen	Prentice Hall
Understanding Actuarial Management: the Actuarial Control Cycle, 2nd Ed.	Bellis, Lyon, Klugman and Shepard. Ed.	Institutes of Actuaries of Australia and the Society of Actuaries
Understanding Actuarial Practice	Stuart Klugman, Ed	Society of Actuaries
Loss Models: From Data to Decisions	Klugman, Panjer & Willmott	Wiley & SOA
Introduction to Scientific Computation and Programming	Daniel Kaplan	Thompson
Intelligence and Other Computational Techniques in Insurance: Theory and Applications	Shapiro & Jain, ed.	World Scientific

THE TECHNOLOGY OF ACTUARIAL PROCESSES AND BUSINESS INTELLIGENCE SOLUTIONS

Actuaries, underwriters, claims examiners and all others within an insurance organization are part of a complex data-driven system aimed at making accurate assessments to advance the goals of their management. Technology is applied at both divisional and enterprise levels to achieve these goals. Each organization balances this technology distribution between division and enterprise differently depending on the goal targeted and/or their appetite for standardization.

The Technology of Actuarial Processes

As professionals, actuaries work through an actuarial process that constantly looks at loss experience from the past, adjusts that experience to current conditions and then reports the findings for actionable decisions. Problems are defined, solutions are designed and reported, decisions and actions are monitored, all in a cycle of activity. The actuary leverages multiple technologies for this monitoring, adjusting, and analysing past experience.

Actuarial Technology Characteristics and Capabilities are frequently characterized by the following:

- They can be standalone.
- They are used for multiple levels of management.
- They are adaptable and flexible.
- They are interactive and easy to use.
- Actuaries control the process.

Actuarial Systems

Certain actuarial processes can rely heavily on technology to maintain standardization. Reserving and capital allocation processes are typical examples where heavy investment in the build of a structured Actuarial System is not uncommon. Best practices of Actuarial Systems suggest they include a database management subsystem, a model management subsystem, and a user interface subsystem.

The database management subsystem consists of a decision support database with a DBMS, a Data dictionary, and a Query facility. The data in the decision support database (primarily premium, exposure and loss data) are non-volatile, cleaned, in a standard format and not used in a transaction processing environment. IT manages the directory, data quality, query facility, data integration, data scalability and data security.

The second component of a structured Actuarial System is a model management subsystem which contains strategic planning models, capital allocation models, pricing and reserving modes and predictive models.

The final component of structured Actuarial System is a user interface system which provides for clear communication of the results of analysis. Of growing importance in user interface systems is data visualization. Data visualization capabilities are a growing expectation for both structured Actuarial Systems as well as smaller actuarial deployments of technology.

Modeling Languages – Many languages can be used in Actuarial Processes. Earlier languages were FORTRAN, Basic, and APL. SQL has replaced many of the processes previously driven by these languages. Today, statistical languages like R and SAS are becoming more common.

Despite the breadth of modeling languages available, it is not uncommon for many actuarial process to be supported exclusively through Microsoft Office applications like Access and Excel spreadsheets including VBA and Excel Add-ins like those from Palisades Corporation.

The Technology of Business Intelligence Solutions

As discussed, actuarial processes may or may not depend on the technology commonly associated with Business Intelligence solutions. With the desire to include more data into the actuarial process, it has become important to consider the use of more sophisticated and scalable Business Intelligence solutions into actuarial processes. In this section, specific tools are mentioned that have been known to create successes within financial institutions like insurance companies. It is important, however, to recognize that new tools are introduced to the market every year which may or may not improve on the capabilities of existing products.

Business Intelligence & Analytics Software Tools

This term refers to the software that supports the business processes, methodologies, and metrics, used by the insurance enterprise to measure, monitor, interpret and forecast business performance. From a cost perspective it is desired to have a single software platform support these processes. In practice, however, the need to balance the varied metrics, timing and detailed appetite of different business support personnel generally requires that different software tools be implemented throughout the enterprise. IBM Cognos and SAP Business Objects are seen as leaders in the space of standard reporting delivered across an enterprise. Their highly structured design, however, creates challenges to meeting new expectations in the marketplace. Frequently they add capabilities through acquisition of innovative competitors. Despite the perceived lack of flexibility of reporting, these giant technology companies benefit from their software deployments by early adopters of the enterprise BI movement in the early 2000's. They are likely the mainstay of many financial organizations where consistency and a "single source of truth" are critical to forming and delivering on expectations. As the desire for "data discovery" continues to widen to support functions that seek trend and correlation insight over precise measurement, tools that specialize in "visualization" have moved to the forefront. Tableau and Qlik, once the newcomers are now becoming established providers looking to maintain market share against steady upstart competition. Microsoft Excel, easily considered the current tool of choice of financial analysts and actuaries, continues to evolve to compete with these mid to large scale BI tools. Depending on the controls required for the business process, it may still be the optimal choice.

Data Warehouses

Data warehouses can but frequently do not have an operational role within an organization. That is, traditionally the role of a data warehouse is to provide control and efficient storage for timely reporting of financial values recognized widely throughout the organization. Data warehousing tools are designed to store and refresh large volumes of structured data (discrete values with a limited number of bytes). With the onset of "big data," the concept of the data warehouse has become less *en vogue*. Despite its current departure from popularity, data warehouses (in one form or another) are the foundational data structures of many organizations. This is particularly true for financial organizations that require stringent internal and regulatory controls on financial information, both historical and prospective. Database tools such as SQL Server, Oracle, IBM DB2 are strongly associated with traditional data warehouses.

Data Mining Applications

These are technology that use statistical, mathematical and artificial intelligence techniques to extract and identify useful information and patterns obtained from large sets of data.

Any tool that can parse text (which includes Excel and SQL), can technically be considered a data mining tool. That notwithstanding, an unstructured dataset can easily exceed a terabyte which generally calls for a tool that can comfortably accommodate such volume. Both SAS and Hadoop are seen as successful tools for data mining projects. A look at Gartner's 2015 *Magic Quadrant for Business Intelligence and Analytics Platforms* drives home the rapid expansion of BI tools that can manage large databases although there are few tools that appear to be a "one-size-fits-all" solution. In fact, many tools in this space are configured to work with other tools recognizing the varied goals of data mining projects.

Sources for further information

Source	Author	Publisher
Decision support and Business Intelligence Systems	Turban, Sharda & Delen	Prentice Hall
Excel 2010 Data Analysis and Business Modeling	Wayne Winston	Microsoft Press
VBA for Modelers: Developing Decision Support Systems with Microsoft Office Excel, 4th Ed.	s. Christian Albright	South-Western Centage Learning
@Risk: Advanced Risk Analysis for Spread Sheets		Palisade Corporation
Computational Actuarial Science with R	Arthur Charpentier, ed.	CRC Press
Session 57 L: Business Intelligence for Actuaries	Rigby & Levine	Society of Actuaries
Practical Management Science, 4th ed.	Winston & Albright	South-Western Centage Learning
Magic Quadrant for Business Intelligence and Analytics Platforms	Gartner	Gartner
Wikipedia: “Database” “Business intelligence tools”	Various	NA
Making Successful Presentations: A Self-Teaching Guide	Terry Smith	Wiley Press
Practical Data Science with R	Zumel & Mount	Manning
Applied Insurance Analytics	Patricia Saporito	Pearson

THE ROLE OF THE ACTUARY IN BUSINESS INTELLIGENCE PROJECTS

Business Intelligence Project Challenges

Despite the millions of dollars spent on BI projects, many BI projects fail or, at the very least, fail to meet their expected potential. There is no shortage of publications lamenting this observation. Even in the financial industry (which includes insurance), where BI projects are most attempted, success stories are limited. A post-mortem of a project's diminished success frequently boils down to two critical gaps:

- Sufficient and sustainable support at the sponsorship level
- Concrete recognition and consensus on the return on investment

Of course, these two causes are frequently related. Sometimes BI projects are embraced with only a fuzzy understanding of the potential or, sadly, because "everyone else is doing it." Many are justified by the assumption that current processes are so fractured and inefficient that a BI project can only improve the organization.

BI projects are generally costly (in the millions) and lengthy, frequently scheduled over years to sufficiently spread the costs. Unless the BI project's business objectives are clear and core to the long term success of the organization, the risk is great that the originally envisioned expectations of the project will fall victim to changing short term priorities or unexpected forces.

Even when sufficiently clear goals and an expected ROI are embraced at the onset, technological advances unveiled during the course of a project can easily sway an impatient sponsor (the CFO or CEO, for example) into changing directions that appear to be faster and less costly. Similarly, a change in the "c-suite" can derail a BI project mid-stream as the new chief officer may want to recast the future BI landscape to their liking and not to that of their predecessor's.

Given this bleak track record, you might wonder why organizations continue to pursue large BI projects instead of operating locally with the power of spreadsheet technology. The answer remains the same as it always has been: the decision-making confidence gained by centralized and integrated data exceeds the pain and cost of a BI project. In fact, as the need to integrate data from external sources and stored descriptive (non-financial) data proliferates, BI projects are more critical than ever if an organization is to remain competitive.

Envisioning improvements to the Actuarial Process within a Business Intelligence Project Scope

Similar to the standards employed in developing an actuarial opinion, creating and delivering a BI report to be used by an organization has an accepted set of appropriate steps or "best practices."

Without these best practices, the delivered report is likely to not meet the expectations of the users in all of the familiar disappointing ways: the report will not be available as scheduled, it will not address all of the expected needs of the user (aka the actuary), nor will it be as flexible to use as desired. Given the investment of time and efforts of the multiple individuals involved (often far more than the creation of an actuarial opinion), it only makes sense that actuaries embrace these best practices as they would uphold the standards of their own work product.

The BI Development/Delivery Life Cycle can be arranged in varying ways within an organization. Some methodologies require formal, detailed meetings and documents with sign off responsibilities for each invested party. Others are less formal and may rely on multiple iterations of prototypes in order to build out the final product. Even when the planning and management is driven more by the business than IT, it is important to align resources, timelines and the efforts of others not directly involved in the project in order to make the best use of resources. It would do well for the actuary to inquire and adopt the expected approach and terminology used by those planning the project.

Despite the variations of approach (waterfall or agile[8]), BI solution development has a commonly understood set of required activities (although the terms used to describe the activities may vary), all of which are critical in order to achieve the goals of the product. It is unlikely that an actuary will be 100% dedicated to a BI project targeted to meet actuarial needs, but the actuary has a defined role in each of these activities.

Planning

"Planning" can be thought of as a constant drive to innovate and/or improve processes. Throughout their career, actuaries should seek out opportunities to brainstorm with co-workers in IT and other areas (finance, operations) on system or reporting improvements that would benefit all parties. That way, when there is an appetite to invest in technology, there is a better chance that the actuary's needs are known across the various support functions. If, on the off chance, the ensuing BI project targets the actuarial department's needs directly, that actuary is likely to have influence on the ultimate design.

Analysis

The use of the term "analysis" in this context generally refers to "Business Requirements Analysis" which is the gathering and documenting of the BI needs and expectations of the targeted users. The actuary will be well served to take the term quite literally from an actuarial science perspective and

[8] In the phases of a traditional waterfall development arc, you move to the next phase only when the previous one is complete.
However, [in an agile development model] instead of tackling all the steps for all of your product features at once, you break the project into *iterations* (smaller segments of the overall project), called *sprints*.
Mark C. Layton, *Agile Project Management for Dummies*

craft out any statistics to support business impacts that would assist in prioritizing needs, either initially or later during the project.

It can be a challenge to sort through the needs and wishes of everyone to determine core deliverables that would enable the users to experience desired results. One recommended approach is to ask "why?" repeatedly until solutions that meet the common needs of all engaged users are uncovered. It is not uncommon for any technical analyst to balk at this process. It can seem patronizing and a waste of time to discuss with non-actuaries what an actuary does and to what end. It is important during this process for the actuary to remember the values of methodical exploration, numerical or verbal, and to recognize that their organization is investing significant resources to the project and resistance to an accepted methodology is the true waste of time.

Design

This includes the finalization of the technical design and specifications. Although it may be discussed during Analysis activities, during Design the technology (software package and infrastructure) will be determined. As discussed earlier, it would be optimal for the actuary to inquire extensively on the software options available to the project. If IT is leading the project, they might assume that, like other support areas, an actuary would have little interest in the varying options of one software package or another. An actuarial department, however, is likely to rely heavily on Excel and custom queries in the existing process and should provide considerable input into the required capabilities of the new solution.

Development

During Development it is tempting for an actuary to "get back to their real job" and wait until they are beckoned again by the project leads. It is during the development process, however, that the hard decisions and compromises are made. Those closer to the project during Development are likely to have the most influence on the initial output. Additionally, knowing where the initial project design had to be "tweaked" due to unforeseen development, will provide the actuary insight later on during the testing phase. It is widely accepted that modifications during this phase are considerably less effort (aka cheaper) than changes made near the end of the project.

Test

Unfortunately, Testing is frequently considered the "last call" for any changes to the BI deliverable. In fact, testing should be performed throughout the project to constantly validate the requirements. As such, the actuarial area should strive to develop test cases that will not only challenge the speed of performance but should also seek to test the unusual but valid request. Some test failures will be more critical than others. It is important to classify the degree of failure as some fixes are likely to be postponed to later releases. The decomposition of these failures is likely to lead to the discovery of a poorly understood business requirement or software capability not captured earlier in the project.

Implementation

Assuming testing was thorough, this process should be relatively painless for the actuary. If, however, the testing process was abbreviated, it is likely to be the most painful. Even if implementation goes smoothly, the actuary will likely need to advocate for the new solution and demonstrate to others its contributions to the Actuarial Control Process.

Sources for further information

Source	Author	Publisher
Business-Driven Business Intelligence and Analytics: Achieving Value through Collaborative Business/IT Leadership	David Stodder	TDWI
Seven Strategies for Creating High-Performance BI Teams	Wayne Eckerson	TDWI
Wikipedia "Systems development life cycle"	Various	NA
Agile Project Management for Dummies	Mark C. Layton	Wiley

CONCLUSION

Providing management data and information for supporting insurance financial systems is the main task for professional actuaries. The data, the models, the communications and professional advice are all resources that actuaries use to support decisions.

For all actuaries currently working, from the actuarial student to those in the final years of their professional life, one constant is and has been the rapidly evolving technology that supports the collection and analysis of data and the proliferation of data sources available. Actuaries in all practice areas and certainly CAS actuaries should stay vigilant to the opportunities and problems brought about by this rapid evolution in technology.

Data Quality Overview
Actuarial Concepts in Data Quality

DATA QUALITY PRINCIPLES

In its simplest terms, Data Quality can be defined as data "fit for its intended use." In other words, Data Quality is measured in terms of how well it fits the data consumers' expectations. The categories in which Data Quality is evaluated include:

- Validity - is the information captured in correct formats, with codes or values that are appropriate for the business? For example: Is Zip Code 10019 is valid for the state of New York?
- Accuracy - does the information captured truly reflect the business information? Continuing with the above example, although valid, the data is not accurate if it is for a risk in Upstate NY while the intended zip code was meant to describe New York City.
- Completeness is used to measure the breadth of the data. Is all of the data that is supposed to be in the file or analysis included? What may have been excluded or duplicated?
- Timeliness is associated with the 'freshness' or time-lag of the data. If we need to support near real time customer service calls, data that is a month old may not meet the quality expectations of the consumer.
- Reasonability refers to the consistency or materiality of the data, given the business conditions. For example, a significant shift in the distribution or profile of a company's book of business may be reasonable, if the company has entered into a new territory or market.
- Data Lineage is a newer category of Data Quality. It includes ability to transparently trace the data path from creation to reporting, including data transformations. This path provides information about the reliability of the data.

Managing Data Quality

Data Quality is typically managed through the development and monitoring of metrics. These metrics must be measurable and should be quantifiable within a discrete range. Note, however, that while there are many things that can be measured, not all translate into useful metrics, implying the need for business relevance. Therefore, every data quality metric should demonstrate how meeting its acceptability threshold correlates with business expectations. The above data quality dimensions should frame the business requirements for data quality. Quantifying how quality is measured along the identified dimension provides hard evidence of data quality levels. The determination of whether the quality of data meets business expectations can be based on specified acceptability thresholds; if the score is equal to or exceeds the acceptability threshold, the quality of the data meets business expectations. Any measurable characteristic of information that is suitable as a metric should reflect some controllable aspect of the business. In other words, the assessment of the data quality metric's value within an undesirable range should trigger some action to improve the data being measured.

If the score is below the acceptability threshold, the appropriate data steward must be notified, and some action must be taken. These data quality metrics can be organized by the data quality dimensions noted above. For example:

- Completeness: Metric to monitor whether total dollars on claim records (e.g., a loss run) balances to a total on a control report;
- Timeliness: Is the claims upload delivery in the agreed upon time range agreed upon among stakeholders completed?
- Validity: What percentage of zip codes in the data are actual valid US zip codes?
- Integrity: What percentage of policyholder records contain a missing or null field?

Quantifiable metrics enable an organization to measure data quality performance improvement over time. Tracking helps in monitoring activities within the scope of data quality service level agreements and demonstrates the effectiveness of improvement activities. Once an information process is presumed to be stable, tracking enables the institution of statistical control processes to ensure predictability with respect to continuous data quality.[1]

Sources for further information

For more information on the overall principles of data quality, please see:

Source	Author	Publisher	Link (if applicable)
Data Quality - The Field Guide	T. Redman, Ph.D.	Digital Press	
Data Quality Assessment	A. Maydanchik	Technics Publications	
Risk Management and Insurance Operations - CAS Course 1 For Preparation For Exam CA1 - Assignment 16 - Actuarial Data Management	Casualty Actuarial Society	The Institutes	

Outside of the basic principles of data quality, the concepts which many actuaries would benefit being aware of fall under an umbrella of two broad topics, joined by their aims to promote a higher standard of information quality within the organization. As discussed in detail in the following sections, these are:

- Data Governance Concepts
- Data Documentation Concepts

[1] IDMA Tools for Managing Data Effectively One Day Class 2014 Insurance Data Management Association

DATA GOVERNANCE CONCEPTS

Quick Glossary of Important Terms

Actuaries Should Know:	
Term	**Definition**
Data Governance	Data governance is "the exercise of authority, control, and shared decision making (planning, monitoring, and enforcement) over the management of data assets."[2] Similarly: "Formalized behavior associated with data. Includes execution and enforcement of authority over the management of data and data-related assets/processes."[3]
Data Stewardship	"Formalized accountability over the definition, production, and use of data and data-related assets/processes."[4]
Data Governance Committee	A data governance council or committee (DGC) is a cross-functional group with members from both IT and the organization's operational side. Members of the DGC generally include the Chief Information Officer (CIO), Chief Data Officer (CDO), the Data Management (DM) leader, and a business executive who acts as Chief Data Steward. It is not uncommon for this group to include executives representing other functions, such as actuarial, underwriting, and claims. The DGC makes high level, strategic decisions about data governance as an integrated function within the organization.

[2] [1] The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK) First Edition, Mark Mosley Editor, Technics Publications LLC, New Jersey, copyright 2009 DAMA International, p., p. 37

[3] Robert S. Seiner, TDAN.com (The Data Administration Newsletter)

[4] Robert S. Seiner, TDAN.com (The Data Administration Newsletter)

Data Governance: Overview

As described above, the primary goal of any corporate data governance initiative is to manage data for the purpose of delivering accurate, valid, timely, and complete data which can be used to inform decisions across the company.

Data governance, therefore, may encompass elements extending beyond simply data. For example, there are the classic people, policy, process, and technology dimensions to data governance. Each must be individually defined with goals in order to successfully support an organization's data governance strategy. Let's examine the elements of each pillar:

- Process – Data governance processes which provide for "an enterprise-wide data governance body, a policy, a set of processes, standards, controls, and an execution plan for managing the data."[5]
- People – Clearly defining roles and responsibilities across the data managers or influencers in the organization. These may include:
    - Business Analysts – Those who actively utilize the data. Those who utilize the data are often the best to provide feedback on the data required for robust analysis. Actuaries, for example, may be considered business analysts.
    - IT Architects – Those who design and direct the flow of data within the organization.
- Technology – As will be discussed in more detail in a subsequent section of this paper, this involves deploying the best suited tools and data infrastructure needed to support the objectives of the organization. For example, if real-time data is necessary, data architects must design technology appropriately suited to deliver the information to the analysts in real-time.

Under the umbrella of these pillars, "Data governance focuses on the delivery of trustworthy, secure information to support informed business decisions, efficient business processes, and optimal stakeholder interactions. It is therefore not an end in itself, but merely the means: data governance supports your most critical business objectives."[6]

While the paper is written from the perspective of the need for data quality, there are business considerations which must be considered in this pursuit. Data quality initiatives involve significant commitments of time, resources, IT architecture, and capital. While actuaries may strive for the absolute best data quality, this must be balanced with an analysis of the marginal returns of greater and greater quality initiatives. For example, does spending $1m for a marginal increase in data quality warrant the business benefits? Actuaries must be prepared to discuss with management the cost implications for the data quality initiatives discussed in this paper.

[5] Deloitte – "How to Build a Successful BI Strategy"

[6] https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/metadata-management-data-governance_white-paper_2163.pdf

Data Governance Committees [*People and Process*]

Overview

"The Data Governance Council's (DGC) primary duty is to ensure responsibility, accountability and sustainability of data practices. The framework for effective data governance planning contemplates the personnel, technology and policies and procedures necessary to ensure the preservation, availability, security, confidentiality and usability of the company's data. Furthermore, a DGC encourages strategic thinking and the creation of opportunities surrounding the appropriate use of data within the organization."[7]

In simpler terms, the DGC manages all projects related to and impacting data within the firm. The advantages of establishing a DGC are numerous and include:

- **Enhanced Deployment of Resources**. Various stakeholder groups may have similar data request needs which are completed or attempted in parallel tracks, resulting in overlap or inefficient deployment of resources. The result may be a patch work of data flows rather than a logical, consolidated flow. A DGC coordinates IT efforts from a focal point of authority.

- **Singular Management of Data.** Data has typically been collected and managed at the business unit or stakeholder level based on individual needs. With increasing needs to integrate or exchange data, organizations need coordinated management to effectively:

  - Migrate data from legacy platforms to current, more advanced solutions;
  - Integrate various systems which may speak a different language (e.g., essentially contain the same information such as dates but in different formats);
  - Determine the accepted definitions of the data for reporting and analysis needs;
  - Report data consistently and ensure its fit for use.

A DGC which centrally manages the data assets of an organization ultimately improves the quality of data and information across the organization. Via consistent naming standards, definitions, formal metrics and calculations, there is an improved understanding of data. This facilitates better communication and understanding of the data, which in term aids in the ability to share or re-use data.

- **Representation of Stakeholders**. The DGC has cross-functional representation and works to understand the needs of corporate stakeholders from a data perspective. Understanding the needs of stakeholders is a key component in creating a synergistic data strategy.

Membership and Actuarial Roles in Data Governance

As insurance companies begin to establish formal Data Governance Committees, representatives

[7] http://www.insidecounsel.com/2014/01/27/inside-establishing-a-data-governance-committee-as

from both the business and technical data stakeholder groups are often included:

- Senior Executives (e.g., Chief Information Officer, Chief Data Officer, and/or Chief Technology Officer)
- Business Stakeholders from:
  - Financial Reporting;
  - Underwriting;
  - Claims;
  - Actuarial
- Technical stakeholders:
  - Data architects
  - Business analysts
  - Project managers

Actuaries usually have a seat or two at the DGC table. Many organizations choose an actuary to be the Chief Data Steward if there is not a formal chief data officer.

Roles and Responsibilities

The roles and responsibilities of a DGC include, but are not limited to:

- Clearly establishing senior authority over data streams which cross organizational boundaries;
- Evaluating all internal data-related projects for coherence with the overall corporate data strategy, architecture, and overlapping work-streams to reduce inefficiencies, redundancies, and conflicting data streams;
- Designing the data related controls and processes for which data travels throughout the organization;
- Monitoring the compliance of data processes with controls mandated by various internal and external authorities (e.g., regulators, auditors, Sarbanes-Oxley);
- Responding to data process or compliance issues by prioritizing resources, approving remediation or strategic plans, and approving the data architecture utilized to support the data processes; and
- Conduct annual audits of strategic data processes.

Tools Available for Data Governance [Technology]

Data Governance Councils rely on the following tools or artifacts:

- **Policies and Procedures**: These are the 'rules' the organization has established for how data and information is processed and analyzed. They may include both internal and external Standards and Guidelines for how data and information is to be coded, processed or exchanged.

- **Enterprise Data Models:** These are the 'blueprints' for how data is organized in the various databases and systems. These are used to understand how the data is related to other data and processes and may be a source for data quality rules.

- **Collaboration Tools:** Most DGCs use a variety of collaboration tools to capture discussions, histories and revised policies and procedures. These aid in the recordkeeping and documentation of important decisions and actions.

Sources for further information

Source	Author	Publisher	Link (if applicable)
Stewardship Approach to Data Governance	Robert S. Seiner	The Data Administrators Newsletter (TDAN)	http://tdan.com/the-data-stewardship-approach-to-data-governance-chapter-1/5037
Deloitte – "How to Build a Successful BI Strategy"	Prashant Pant	Deloitte	http://www.loria.fr/~ssidhom/UE909R/1_BI_strategy.pdf
Establishing a Data Governance Committee as part of 2014 strategic priorities	David Katz	Inside Counsel	http://www.insidecounsel.com/2014/01/27/inside-establishing-a-data-governance-committee-as
Defining Organizational Structures	Gwen Thomas	The Data Governance Institute	http://www.datagovernance.com/defining-organizational-structures/
Risk Management and Insurance Operations - CAS Course 1 For Preparation For Exam CA1 - Assignment 16 - Actuarial Data Management	Casualty Actuarial Society	The Institutes	

DATA DOCUMENTATION CONCEPTS

Introduction

There are varying levels of data management and documentation, all of which an actuary can play an integral role in for an organization. The following glossary of terms are important in the subsequent documentation discussion. The interdependence of documentation and governance are concepts that will be explored further in this section.

Quick Glossary of Important Terms Actuaries Should Know

Term	Definition
Big Data	**Big Data** is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.[8]
Master Data	**Master Data** represents the business objects which are commonly agreed to and shared across the organization. Customer and / or Product IDs are examples of Master Data.[9]
Master Data Management	**Master Data Management (MDM)** is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file that provides a common point of reference. When properly done, MDM streamlines data sharing among personnel and departments. In addition, MDM can facilitate computing in multiple system architectures, platforms, and applications.[10]
Metadata	**Metadata** are business and technical information about an organization's data. They help put data in context, reveal their meaning and make them accessible.[11] Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.[12] Metadata summarizes information about data for the purpose of making that data easy to find and work with.

[8] www.gartner.com/it-glossary/**big-data**

[9] Wikipedia

[10] http://searchdatamanagement.techtarget.com/definition/master-data-management

[11] IDMA Curriculum Rewrite Task Force - Course 201 Assignment 6

[12] http://www.techrepublic.com/blog/it-security/is-metadata-collected-by-the-government-a-threat-to-your-privacy/

Term	Definition
Metadata Repository	A **Metadata Depository** is a database and software used to capture, manage and access metadata. It is where an organization collects, integrates, standardizes, consolidates, organizes, controls, and stores its metadata, and makes them available for shared general use.[13] A Metadata Repository is a special type of database containing information about another database, e.g., how the data in the other database was collected, transformed, and formatted, how frequently it is updated, and generally anything that can be useful to analysts that need to query data from that database.[14]
Data Flow Chart	A mapping of data flows from source systems (e.g., Policy Admin System, Enterprise Data Warehouse, etc.) to intermediate data stores (if any) and finally to end user.
Stakeholder Data Analysis	A comprehensive review of the data requirements of the enterprise stakeholders, both for analysis and reporting purposes. This often comprises the first step in formulating an overall enterprise data strategy.
Data Dictionary	"A data dictionary is a tool for displaying metadata to business and technical personnel. A data dictionary is important for expediting the transfer of knowledge regarding the meaning of data values stored in the data fields."[15]

[13] DMA Curriculum Rewrite Task Force - Course 201 Assignment 5

[14] http://www.casact.org/pubs/forum/05wforum/05wf274.pdf

[15] https://www.casact.org/pubs/forum/05wforum/05wf274.pdf

Data Documentation and the relationship to Data Governance

The key to superior data governance is the processes supporting the management and documentation of data. As will be discussed, there are varying levels of data management and documentation, all of which an actuary can play an integral role in for an organization.

"Maintenance of adequate documentation describing the data can help avoid problems associated with relying exclusively on people's memories of what is contained in the data. As actuaries we can help persuade our business and data management partners that system documentation is vital to the actuarial work product." Documentation is the cornerstone for well performing data governance – in that sense, data documentation and data governance are not mutually exclusive. Data won't govern itself.

Stakeholder Analysis

One of the first items of documentation a Data Governance Committee may seek to help formulate strategy, priorities, and objectives of the data governance function is a stakeholder analysis.

Stakeholders across different departments demand information presented in a way which aligns with their operational objectives – understanding what these stakeholder operational and reporting objectives are help define the data an organization needs to collect, store, and process.

As a simplistic example, personal auto coverage is often differentiated by actuaries into property damage and bodily injury components due to the differing development characteristics of these pieces. However, the profit center managers for personal auto do not view these components in isolation - they understand composite results and rates, and thus demand combined metrics. Furthermore, actuaries may look at the business pieces on a countrywide basis, while profit centers require a state by state breakdown of the results. As such, it's necessary for a data organization to understand what its stakeholders require from a data perspective – in this case, the actuaries would require losses and premiums on a state by state basis, split by property damage and bodily injury, for both analysis and allocation purposes.

Data and Process Flow Diagrams

Data and process flow diagrams are used to document the lifecycle that data goes through in the organization (again, one may hear the term data lineage when discussing the data lifecycle). It visually represents the various systems (input and outputs) that are involved the creation, consumption and transformation of data. It helps to ensure that all systems and processes are accounted for and is used to manage data lineage and data impacts.

Sources for further information

Source	Author	Publisher	Link (if applicable)
The Data Governance Institute			http://www.datagovernance.com/
The Data Administration Newsletter		Robert S. Seiner	http://tdan.com/
Risk Management and Insurance Operations - CAS Course 1 For Preparation For Exam CA1 - Assignment 16 - Actuarial Data Management	Casualty Actuarial Society	The Institutes	

Metadata: Overview and Technical Concepts

Metadata is one of the more topical and least understood data documentation concepts.

Metadata provides the context and descriptions of the data (the type, what it means, where it is located, how it used, etc.)[16] Metadata is important from a data lineage perspective – the ability to trace the data through its various stages and transformations is a key to ensuring data quality. In the broadest sense,

> Metadata can be considered the documentation of the contents of a database. In addition to the information about the data itself, metadata contains information about business rules and data processing.[17]

As defined in the CAS 2007 Winter Forum and in reference to Data Quality: the Accuracy Dimension by Jack Olson,

> Metadata is a term used by data management professionals for information about the data such as definitions, a description of permissible values and business relationships that define the data in a database. Comprehensive metadata is a prerequisite for good information quality.[18]

To that extent, Metadata can be classified into subtypes.

Types of Metadata

There are three types of metadata the actuary should be aware of:

1. Technical (also known as Structural)
2. Business (also known as Descriptive)

[16] http://dataqualitypro.com/data-quality-pro-blog/data-quality-through-metadata-strategy-anne-marie-smith)
[17] https://www.casact.org/pubs/forum/08wforum/actuarialIQ.pdf
[18] https://www.casact.org/pubs/forum/07wforum/07w279.pdf

3. Operational (also known as Administrative)

Technical Metadata

Technical metadata assists in understanding the format and definition of the data collected. For example, if one were sending a letter, the information contained on the envelope can be considered the data. The "data describing data" about that envelope may include a name and address and specify the format that "data about the data" is in.[19] This may include about the addressee:

- Surname – coded as 20 digits, alphabetic
- U.S. State – coded as two digit alphabetic code defined by US Postal Service
- Zip Code – coded as five digit numeric as defined by the US Postal Service

Technical metadata helps interpret the raw data – for example, we would know that NJ is a state abbreviation for New Jersey through structural metadata. Structural metadata is thus a key component in data quality – it gives context to the data (e.g., we know that NJ is a reference to the US State of New Jersey, how it should be coded – in this case with a 2 letter abbreviation, and its meaning).

Technical metadata includes the source table information so a user can know exactly where the information is sourced from. This is important in using metadata to generate data flow charts or maps.

Business Metadata

Business (or descriptive) metadata assists in describing a resource for purposes such as discovery and identification.[20] This type of metadata provides context around the data, including but not limited to the data field's name, definition of contents, related data, as well as the applicable business rules.

Continuing with our letter mailing example from above, business metadata may log characteristics of the transaction around sending the letter. This includes the addressee, sender, post-mark date, post offices handling, date of delivery, etc.[21]

Operational Metadata

Operational (or administrative) metadata provides administrative "data about the data", including date of last update, date of last access, user who last modified, movement from source to target, availability and usage. Operational metadata may also include a description of the types of data control or quality checks performed on the data, and where in the data processes these occur – this is a metric that allows for audit trails providing proof of compliance for data related controls.[22]

[19] http://www.riskandinsurance.com/the-whodunit-of-big-data/
[20] http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[21] http://www.riskandinsurance.com/the-whodunit-of-big-data/
[22] https://www.casact.org/pubs/forum/08wforum/actuarialIQ.pdf

Metadata as a Key to Successful Data Governance

Key to successful data governance is the management of metadata – the frame of reference giving data its context and meaning. Effectively governed metadata provides a view into the flow of data, the ability to perform an impact analysis, a common business vocabulary and accountability for its terms and definitions, and finally an audit trail for compliance.

High Level Metadata Example: Metadata in the News

Metadata's association with "data about data" is best seen through an example which is rather topical, albeit non-insurance related. In 2013, President Obama addressed the current data collection practices of the NSA with a reference to metadata:

> What the intelligence community is doing is looking at phone numbers and durations of calls. They are not looking at people's names, and they're not looking at content. But by sifting through this so-called **metadata**, they may identify potential leads with respect to folks who might engage in terrorism. [23]

Continuing with this example, suppose you make a phone call to a friend: The phone call's conversation (e.g., what was said) is raw data. Data without context may have little to no value – for example, a relatively mundane conversation may not generate any actionable information. However, the data about this data (i.e. the data about the conversation) may be classified as metadata and create the context needed to generate actionable information. For example, this metadata may include the date and time you called somebody, the duration of the phone call, the phone numbers involved, or the location of the participants.[24]

Using this information, as President Obama noted in his press conference, is how the NSA has generated links to those involved in terrorist activities.

High Level Metadata Example: Metadata in Health Care

An example of the power of metadata in cross-referencing data sources can be seen in the health care industry where vast amounts of patient information is collected, often by different users or systems. Metadata is a key component in tying this information together – "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information."[25]

This information may be tied together via a metadata repository which consolidates the metadata from various source systems, and from there integrate with the system through which end users query.

[23] http://blogs.wsj.com/washwire/2013/06/07/transcript-what-obama-said-on-nsa-controversy/

[24] http://www.theguardian.com/technology/interactive/2013/jun/12/what-is-metadata-nsa-surveillance#meta=1111111

[25] http://www.niso.org/publications/press/UnderstandingMetadata.pdf

In health care, it can be used to link patient information across sources. For instance, "If the physician prescribes the patient aspirin for a chronic headache, metadata could be used to retrieve other patient information, alerting the physician that the patient currently takes a blood thinner." [26]

Metadata Promotes Data Quality

Data quality intuitively can be measured by how well insurers can cross-reference, analyse, interpret and capitalize on the vast amounts of data collected. As seen above, robust metadata is a powerful tool in creating connections between data (and explicitly enhancing its quality and usability). Metadata is key to organizing the data collected, reducing confusion, and enhancing the usability and cross referencing ability of the data.

> "Without structural metadata, both descriptive metadata and, ultimately, the data content of the transaction, have no context."[27]
>
> "Good metadata management can lead to good data quality since having and relying on the metadata can identify poor data / incorrect data / missing data. Also, having good metadata shows an understanding of data management and shows that the organization is committed to good data – hence an improvement in data quality almost always follows." [28]

In the simplest sense, defining exactly how data is to be recorded within a database (format, character types, permissible values, etc.) is crucial to reducing the amount of time needed to scrub the data.

Metadata in the P&C Insurance Context

Let's expand the examples of metadata to an insurance context. One traditional compilation of metadata which we often take part in are the ISO and NCCI statistical plans. Statistical Plans are developed with a goal of providing a data base of homogeneous experience for comparable policies that fulfils both a regulatory need and a business need to correctly price the insurance product.

- For regulatory purposes, the statistical plans collect historical insurance company experience by state, by class, and by coverage. The minimum requirements for the regulatory needs are included in the National Association of Insurance Commissioners (NAIC) Statistical Handbook of Data Available to Insurance Regulators.
- For the business purpose of pricing the insurance product, Statistical Plans go beyond the regulatory mandated data elements (or standard data elements) and collect both additional detail within the standard data elements and additional or new data elements to perform research and development to better refine the rating or classification of insurance policies and to provide advisory prospective loss costs. By aggregating the data together from many insurers,

[26] http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_049357.hcsp?dDocName=bok1_049357

[27] http://www.riskandinsurance.com/the-whodunit-of-big-data/

[28] http://dataqualitypro.com/data-quality-pro-blog/data-quality-through-metadata-strategy-anne-marie-smith

> the resulting ISO data base provides a larger, more credible data base than any one insurer can do alone.[29]

The Statistical Plans are essentially the rules that described how the data is to be captured, including the format and codes or values that to be used. These descriptions and instructions help to form the metadata, definitions and business (and data quality) rules for the data.

Furthermore, metadata is crucial for insurers to capitalize on the advantages of big data. Robust metadata can be used to make connections between disparate data sources for use in analytics.

Metadata in Predictive Analytics

"Metadata within data infrastructures enables us to locate and combine data, and to analyze its lifecycle and history. Consider, for instance, the addition of weather, geographical and social media data to the daily sales figures for a retail chain. It is easy to conceive that correlations with peaks and troughs in sales could be elicited: perhaps with good weather, word-of-mouth trends or road accessibility. With sufficient data, some of these events might even be found to be predictive of sales."[30]

Metadata to Detect Insurance Fraud

According to the Insurance Information Institutes, Property / Casualty insurance fraud amounts to about $32 Billion a year. From quote, policy issuance and even claims reporting, more and more insurance transactions are conducted online and through various mobile applications. Each internet-enabled device, which can include computers, tablets and cell phones, has metadata associated with it. The metadata can include which email accounts are associated with it, what the IP address is, etc. This metadata can be used to identify whether this electronic device is connected to an account (email) or other device with a known history of fraud. Technology is now available that can use this metadata, check for anomalies, attributes and activity levels of the device to determine a "Reputation" of the device. This device-based intelligence, gathered through metadata, is helping insurers identify fraud at the point of first contact.

Metadata in Property Risk Modeling

Without accurate address data, you can't property risk model without making assumptions about the risk characteristics of a particular property. Those assumptions usually take the form of using the average or modal value for a particular characteristic at the ZIP code level.

The Actuary's Role in Metadata

Actuaries are key individuals in developing the enterprise metadata and helping to promulgate its

[29] CAS STUDY NOTE: IS0 STATISTICAL PLANS, by Virginia R. Prevosto, FCAS, MAAA

[30] http://www.forbes.com/sites/edddumbill/2013/12/31/big-data-variety-means-that-metadata-matters/

usage and acceptance within the organization. Perhaps most importantly, actuaries play a role in the definition of the business metadata (e.g., providing clear definitions to the data to avoid confusion and misunderstanding across functional users or business units).

As an example, the definition of "loss" may significantly vary among business users depending on the context the term is used in. For a data field called "loss," it is important to exactly define what comprises these values. For instance, if we're talking in the context of Workers' Compensation, loss may include indemnity only loss, medical only loss, allocated loss adjustment expense, unallocated loss adjustment expense, etc. For a user in finance computing a "loss ratio" using the loss field as the numerator, potential confusion and erroneous indications may result without a clear context of what defines the loss field. Metadata is a key governance solution to avoid this type of confusion and provide business users a clear context for the data utilized in analysis or decision making.

Actuarial IQ has a well-defined starting list of questions an actuary can ask when helping IT create the enterprise metadata. The following is borrowed directly from the paper:[31]

- Are all the data elements listed?
- Has the source of each data element been provided?
- Is there a special value that is used to indicate missing data?
- Are there any transformations being applied to the data? (Note: data clean up such as filling in missing values should be considered data transformation).
- Have the contents and use of each data element been properly described?
- Have all the categorical values of each data element been properly described?
- In the case of numeric data, has the range of possible values for each data element been provided?
- Has the valuation date of all data been provided?
- Has a schedule of planned updates to the data been provided?
- Has the business process changed during the experience period?
- Have any of the data definitions changed during the experience period?

A good place to start is with our own actuarial work product. In many instances, we may produce or maintain databases that underlie our analyses. How well documented are these systems? How well understood are the sources that feed the actuarial systems? Once the actuarial systems are understood, one can start to drill back into the source systems. Along the way, missing metadata can be identified. The benefits and costs of producing the metadata can be weighed and ownership could be assigned.

[31] https://www.casact.org/pubs/forum/08wforum/actuarialIQ.pdf

Sources for further information

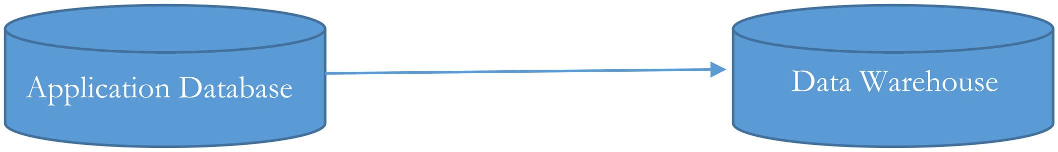
Source	Author	Publisher	Link (if applicable)
Data Quality: The Field Guide	Thomas Redman	Digital Press, 1st Edition, 2011	
Actuarial I.Q. (Information Quality)	CAS Data Management Educational Materials Working Party	CAS	https://www.casact.org/pubs/forum/08wforum/actuarialIQ.pdf
Metadata Management for Holistic Data Governance	Informatica Whitepaper	Informatica Whitepaper	https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/metadata-management-data-governance_white-paper_2163.pdf
Understanding Metadata	NISO		
Survey of Data Management and Data Quality Texts	CAS Data Management Educational Materials Working Party	CAS	https://www.casact.org/pubs/forum/07wforum/07w279.pdf
Actuarial Data Management In A High-Volume Transactional Processing Environment	Joseph Strube and Bryant Russell, Ph.D., ACAS, MAAA	CAS	https://www.casact.org/pubs/forum/05wforum/05wf274.pdf
Risk Management and Insurance Operations - CAS Course 1 For Preparation For Exam CA1 - Assignment 16 - Actuarial Data Management	Casualty Actuarial Society	The Institutes	
Actuarial Standard of Practice 23 - Data Quality	Actuarial Standards Board (ASB)	ASB	http://www.actuarialstandardsboard.org/wp-content/uploads/2014/02/asop023_141.pdf

Databases

Why an Actuary Needs to Know about Databases

Insurance organizations are increasingly becoming data driven. While most insurance companies have had Chief Actuaries for some time, the advent of the Chief Data Officer is a recent phenomenon, and increasingly common. In some instances, the Chief Actuary and Chief Data Officer are one-and-the- same. In other instances, Actuaries are leading the Data Management and Governance processes. Actuaries have traditionally been the prime users of data for analytical purposes at insurers, and as primary users of data, Actuaries should be at the forefront of the new data-driven culture. As this data culture grows, the discussions about data will become increasingly technical. To be a valuable and influential participant, it will be important that the actuary is reasonably fluent in data terminology, and knowledgeable about how data is stored. Databases come in many forms that vary by intended use; this paper will provide an overview of some of the more common forms.

APPLICATION DATABASES VS DATA WAREHOUSES



What's the difference between an application database (ADB) and a data warehouse (DW)? They are both databases, but they serve very different purposes.

As an actuary, you may rely on reports that will be as of yesterday, even though you know the data is already in "the system" today. Why can't your report include the data you know is there?

One reason is likely to be that your application data is in the ADB, but your report references the DW which may only be updated nightly. So why bother to have a DW when the data already exists in the ADB?

ADBs are designed to store transactional data efficiently and allow IT to efficiently add new transactions and update existing transaction data, while DWs are designed to serve as a source for reports and data analysis. Unfortunately, the underlying structures that optimize these functions are significantly different. For example, the following table illustrates a subset of data you might download in a report on current policyholders:

Insured ID	Insured Name	Premium	State Name	City Name
1	Aaron	90,000	Illinois	Chicago
2	Brian	30,000	Illinois	Chicago
3	Chris	30,000	Illinois	Chicago
4	David	40,000	Illinois	Springfield
5	Eddie	80,000	Wisconsin	Madison
6	Frank	10,000	Wisconsin	Madison
7	Gary	20,000	Wisconsin	Milwaukee
8	Henry	20,000	New York	New York
9	Isaac	40,000	New York	New York
10	John	30,000	New York	Albany
11	Kevin	90,000	New York	Albany

This data is easy to understand and once in Excel it is ready to serve as the source for a PivotTable or aggregating functions like SUMIFS(), as you might want to see the Premium at a state or city level. The format of this data is more likely to appear in DWs.

Data Table Normalization

The data above would likely come from an ADB, where it would probably be formatted as follows.

Insured ID	Insured Name	Premium	City ID
1	Aaron	90,000	1
2	Brian	30,000	1
3	Chris	30,000	1
4	David	40,000	2
5	Eddie	80,000	3
6	Frank	10,000	3
7	Gary	20,000	4
8	Henry	20,000	5
9	Isaac	40,000	5
10	John	30,000	6
11	Kevin	90,000	6

State ID	State Name
1	Illinois
2	Wisconsin
3	New York

City ID	State ID	City Name
1	1	Chicago
2	1	Springfield
3	2	Madison
4	2	Milwaukee
5	3	New York
6	3	Albany

The data has been "normalized" for use in an ADB. When data is normalized, it is reorganized so that it is as parsimonious as possible. As you can see, the data above was reorganized from the original insured table and has been split into an insured table, a state table, and a city table, and ID fields have been included. Since the City ID in the insured table maps the City Name, and the State ID in the city table maps the State Name, all the information of the original insured table is preserved in the normalized data tables.

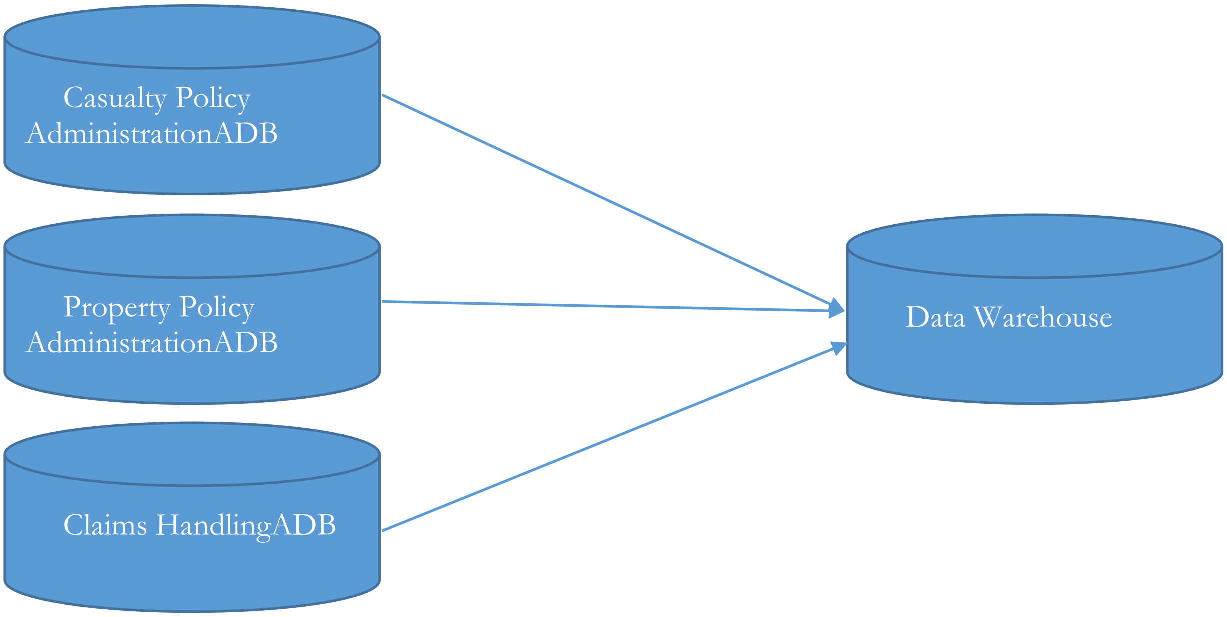
But why would normalized data tables be of use to IT in an ADB? One reason is to note that the addition of new data only requires a City ID instead of a State Name and City Name, since the City ID encapsulates both. Since new policyholders are added to the ADB more frequently than cities or states, less work is required to upload the same level of information.

Another reason to normalize data is to easily update entries across multiple tables using "primary keys," which have unique entries for every record in a table. For instance, let's say an insured's name was entered incorrectly and it must be updated in every table in the ADB. If the data tables are normalized, the insured's name can be corrected in the insured table, and any other table that references the insured's name does so via the Insured ID, so the update effectively takes place across all data tables. Otherwise, the insured's name would have to be tracked down across all the various data tables, which could be very time consuming—and imagine the problem when two insureds happen to have the same name! By using IDs as primary keys, IT is able to uniquely identify relevant data and update it efficiently.

In addition to making it easy for IT to make update data, tables with unique entries can be used to ensure data is entered consistently into the system. In order to prevent users from manually entering "New York City" instead of "New York," IT can reference the city table to validate the data before it goes into the system.

So if there's a good reason to normalize data tables in the ADB, why bother to de-normalize it in the DW? Why doesn't IT simply provide actuaries with reports directly from the ADB instead of reformatting the data in the DW?

Why bother with a Data Warehouse?


Casualty Policy AdministrationADB
Property Policy AdministrationADB
Claims HandlingADB
Data Warehouse

One reason is that a DW can get data from several ADBs associated with different business functions, like policy administration and claims handling. Furthermore, single business functions can rely on several ADBs that have evolved separately to handle specific lines of business or regulatory requirements. Extracting data from these ADBs with varying data formats and transferring it to a DW dramatically simplifies the development of reports for end users.

Structured Query Language (SQL)

Even if a DW represents data from a single ADB, the process of de-normalizing data makes it much easier for IT to develop reports. For instance, let's say you wanted to get the insured data presented in the first, de-normalized table. Since data in relational databases[1] is most often manipulated using Structured Query Language (SQL), the SQL statement written to query the table, "Denormalized," would look as follows:

```
SELECT [Insured Name], [Premium], [State Name], [City Name]

FROM Denormalized
```

Even if you're not familiar with SQL, the statement above is pretty straightforward. Now let's look at what the statement would look like to get the same data from three normalized tables, "Insureds," "States," and "Cities."

```
SELECT i.[Insured Name], i.[Premium], s.[State Name], c.[City Name]

FROM Insureds i
```

[1] Relational databases (like Microsoft Access and SQL Server) are based on tables ("relations") that are linked by their primary keys. While there are databases that make use of different formats, you can usually assume that when IT talks about a database, it will be a relational database unless they specifically qualify it.

```
LEFT JOIN Cities c ON i.[City ID] = c.[City ID]
LEFT JOIN State s ON c.[State ID] = s.[State ID]
```

By referencing normalized data tables, the query must explicitly state how the tables are related through JOIN statements, and also identify the tables that contain the various field names. And while the statement above uses LEFT JOIN statements, there are also RIGHT JOIN, INNER JOIN, OUTER JOIN, and FULL JOIN statements that might be appropriate depending on the nature of the data. You could think of the transformation of data to denormalized tables as a way of baking in the joins so that downstream queries don't need to deal with them. As the SQL statements increase in complexity, the value of referencing denormalized tables becomes clear[2].

Extract, Transform, and Load

The process of migrating data from ADBs to a DW is often referred to as ETL, which stands for Extract, Transform, and Load.

Extracting data from multiple sources can be challenging when those sources are not relational databases, and therefore require methods beyond traditional SQL statements. The extraction process can include a validation step that halts the process unless the data conforms to certain standards, preventing complications further downstream.

Transforming data can involve a number of adjustments, including the aggregation of data. Since the reports generated from the DW may not require the level of detail that exists in an ADB, performance could be enhanced by aggregating data during the transfer. For instance, a DW used by actuaries to evaluate experience by territory might only need data aggregate by ZIP code, county, or state, rather than the exact address of each insured. While street names and addresses would be lost in the course of aggregating the data, reports referencing this data in the DW would run faster. Another alternative would be for IT to preserve the detail for the data in the DW and implement an index, which keeps track of groups of records in a table, enabling more efficient retrieval of specific records. If indexes were implemented in an ADB, they would need to be maintained and updated with the addition of each new record, and if records are added more often than they are read, the cost in computing resources would probably outweigh the benefit. Other manipulations that may occur when moving data from an ADB to a DW include

- Adding accounting periods
- Calculating unearned premium reserves and earned premium
- Assigning loss development months for the creation of loss triangles
- Calculating reinsurance premium and loss from direct and assumed business
- Mapping premium and loss to lines of business defined by various regulatory regimes
- Mapping data to general ledger codes
- General scrubbing of data

The final step, Loading, is (hopefully) executed in such a way as to leave an audit trail, so that any data that looks odd can be traced back to the source ADB and verified. The loading process must also determine which of the preexisting data to overwrite, update, or leave alone. This step can be challenging, for what if an underwriter moves from California to New York, and a report aggregates

[2] For a useful reference on SQL statements, visit: http://www.w3schools.com/sql/default.asp

premium written by underwriter location? Should the underwriter's premium be re-classified as New York premium, so that the current reports won't be consistent with previous reports? One solution to this Slow Changing Dimension (SCD) issue is to duplicate the underwriter record and add fields that indicate the start date and end date for when that record is valid. This way the current reports are accurate but users can still generate reports as of earlier dates that tie to the original versions.

Levels of Normalization

The data formats discussed so far have been categorized as "normalized" or "denormalized," which is actually a convenient simplification. Data tables can go through increasing stages of normalization, with almost all of them in a state of at least "first normal form," or "1NF," and most being at "third normal form" or "3NF."[3] Meanwhile, denormalized tables used in DWs are referred to as "dimension tables" that contain various combinations of categories used to evaluate data held in "fact tables." DWs with more denormalized dimension tables can be described as having "star schemas" whereas less denormalized DWs with less denormalized dimensions tables can be described as having "snowflake schemas." Whatever the structure of your ADBs and DWs, the goal is to take data from an ADB that's optimized to be regularly updated with new data, and move it to a DW that's optimized to provide periodic reports. If you're interested in learning some of the more technical aspects of understanding and interacting with databases and data warehouses, the Microsoft Virtual Academy has tutorial videos intended to assist students preparing to take exams necessary for certification in SQL Server. You can access these videos here: https://mva.microsoft.com/colleges/mcsa-sql.

Development of a Data Warehouse / Business Information System (DW/BI)

For a DW / BI system to be successful, the business community must accept it. For them to accept it, information needs to be presented consistently, be easily accessible, timely, secure, authoritative and trustworthy.[4] A transaction schema focused on business processes and related measureable events at the most granular level allows for maximum flexibility when data is extracted for analysis. For insurance companies, core business processes could include policy issuance, premium collection and claim processing. A claim processing measurable event could include a claim payment or reserve change.

Facilitating a good system requires a team effort and should include Data Governance, IT, actuarial and other heavy users of analytic data. Actuaries need to be somewhat familiar with IT terminology to be effective members of the design team.

Data Architecture Terminology

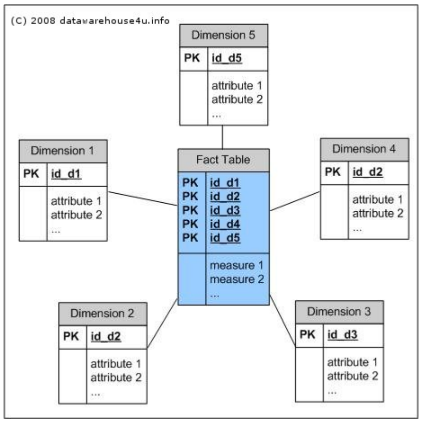
Design features described in previous sections are again described here with a few common data warehouse architecture terms. [5]

- **Star Schemas** – refers to the architecture of a dimensional model implemented in a relational database management system. It includes a fact table at the core surrounded by multiple dimension tables joined by keys in a star-like formation.

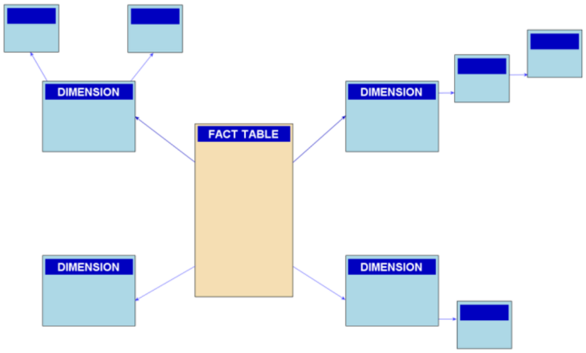
[3] There are higher order normal forms, but 3NF is usually sufficient for database administration purposes.

[4] [2] Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN): John Wiley & Sons. Chapter 1, Goals of Data Warehousing and Business Intelligence; p. 3-4.

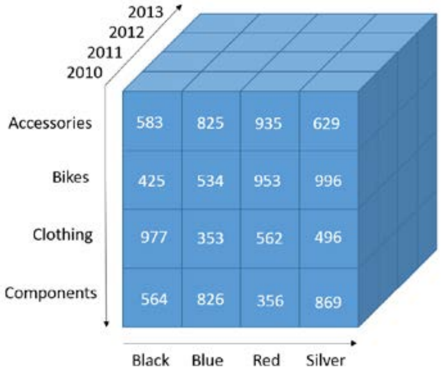
[5] [2] Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN): John Wiley & Sons. Chapter 1, Dimensional Modeling Introduction; p. 7-18.



- Snowflake Schema - represents a dimensional model which is also composed of a central fact table and a set of constituent dimension tables which are further normalized into sub-dimension tables. Because snowflake schemas are normalized, they are easier to maintain, but harder to query.



- **Online Analytical Processing (OLAP) cubes** – presentation area containing aggregations and precalculated summary tables. Data is loaded from the tables after additional processing.



- **Fact Table** – stores the performance measurements resulting from an organization's business process events. Most useful facts are numeric, additive and continuously valued. A non-additive fact would be rate per unit of exposure. Semi-additive facts like claim reserve balances cannot be summed across the time dimension. These tables consume the most storage. Potential facts might include premium in a policy transaction Fact Table; claim dollars in a claim transaction Fact Table.
- **Grain** – level of detail in each Fact Table row; three categories include:

  - **Transaction –** like reserve changes
  - **Periodic snapshot –** like triangles
  - **Accumulating snapshot –** like claim reserve balances
- **Dimension Table** – integral companion to a fact table; contains the textual context associated with a business process measurement event; often have many columns or attributes. Source of query constraints, groupings and report labels. Good practice is to minimize the use of codes in dimensional table for the sake of consistency and clarity across business processes. Potential Dimensions applicable might include policy effective date, policyholder, coverage, covered item, claimant, date of loss.
- **Keys** – join fact with dimension table. Generated when the fact table is created as sequential integers.
  - **Primary Key –** a **key** in a relational database that is unique for each record in a table
  - **Foreign Key –** a field in one table that uniquely identifies a row in another table
  - **Natural Key –** a key that uses its naturally occurring value as its unique identifier (e.g. Telephone Number)
  - **Surrogate Key –** a key that has to be created to uniquely identify a row in a table (e.g. Policy Number)

Steps in Dimensional Design Process for Insurance Company[6]

- Select the business process (e.g. ratemaking)
- Declare the grain – the lowest level of detail that will be stored (e.g. coverage)
- Identify the facts – the quantitative data that will be measured (e.g. premium)
- Identify the dimensions – the attributes of the facts in a dimensional database (e.g. state)

One of the collaboration tools commonly used in a Data Warehouse development process is a bus matrix.[7] The matrix is simply a table with the core business processes as rows and core dimensions as columns. It is a useful as a communication and documentation tool for DW / BI team participants. Below is an example:[8]

	Date	Policyholder	Covered Item	Coverage	Employee	Policy	Claim	Claimant
Policy Transactions	X	X	X	X	X	X		
Premium Snapshot	X	X	X	X	X	X		
Claim Transactions	X	X	X	X	X	X	X	X

[6] Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN): John Wiley & Sons.

[7] Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN): John Wiley & Sons. Chapter 4, Enterprise Data Warehouse Bus Architecture; p. 123-130.

[8] Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN): John Wiley & Sons. p. 389.

Conformed fact and dimension tables keep naming conventions and defined calculations consistent across all departments. DW/BI system should:

- Deliver data that is understandable to the business users.
- Deliver fast query performance.

OTHER TYPES OF DATABASES

Database types are dictated by their intended purpose. So far we have focused on application databases and Data Warehouses. Operational applications handle data input one transaction at a time; data warehouse business intelligence (DW/BI) systems maintain historical data for analysis by the business community. Other database types include:

Flat and Wide (F&W)

Flat and wide (or F&W) is a common data type given to actuaries from IT, usually in an Excel spreadsheet. It is denormalized and typically relatively small.

Operational Data Store (ODS)

An operational data store (or "ODS") is a database designed to integrate data from multiple sources for additional operations on the data. Unlike a master data store, the data is not passed back to operational systems. It may be passed for further operations and to the data warehouse for reporting.

Because the data originate from multiple sources, the integration often involves cleaning, resolving redundancy and checking against business rules for integrity. An ODS is usually designed to contain low-level or atomic (indivisible) data (such as transactions and prices) with limited history that is captured "real time" or "near real time" as opposed to the much greater volumes of data stored in the data warehouse generally on a less-frequent basis.

Columnar Databases

A column-oriented DBMS (or columnar database) is a database management system (DBMS) that stores data tables as columns rather than as rows. Practical use of a column store versus a row store differs little in the relational DBMS world. Both columnar and row databases use traditional database languages like SQL to load data and perform queries. Both row and columnar databases can become the backbone in a system to serve data for common ETL and data visualization tools. However, by storing data in columns rather than rows, the database can more precisely access the data it needs to answer a query rather than scanning and discarding unwanted data in rows. Query performance is often increased[9] as a result, particularly in very large data sets.

Another benefit of columnar storage is compression efficiency.[10] It is well known that a row of similar data, dates for example, can be compressed more efficiently than disparate data across rows. It's for this reason, columnar databases are well-known for minimizing storage and reducing the amount of I/O needed to read data and answer a query. Columnar databases most often are paired with Massively Parallel Processing (MPP) capability to allow for it to share the analytical workload across a cluster. They may also leverage Hadoop MPP capability for this purpose.

[9] Ventana; et al. (2011). "Ins and Outs of Columnar Databases."

[10] Ventana; et al. (2011). "Ins and Outs of Columnar Databases."

NoSQL Databases

A NoSQL (originally referring to "non SQL" or "non-relational")[11] database provides a mechanism for storage and retrieval of data which is modelled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century,[12] triggered by the needs of Web 2.0 companies such as Facebook[13], Google[14] and Amazon.com[15]. NoSQL databases are increasingly used in big data and real-time web applications.[16] NoSQL systems are also sometimes called "Not only SQL"[17] to emphasize that they may support SQL-like query languages.[18]

Graph Databases

In computing, a graph database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.

Most graph databases are NoSQL in nature and store their data in a key-value store or document-oriented database. In general terms, they can be considered to be key-value databases with the additional relationship concept added. Relationships allow the values in the store to be related to each other in a free form way, as opposed to traditional relational databases where the relationships are defined within the data itself. These relationships allow complex hierarchies to be quickly traversed, addressing one of the more common performance problems found in traditional key-value stores. Most graph databases also add the concept of tags or properties, which are essentially relationships lacking a pointer to another document.

[11] NoSQL DEFINITION: Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable. See: http://nosql-database.org/.

[12] Leavitt, Neal (2010). "Will NoSQL Databases Live Up to Their Promise?" (PDF). IEEE Computer.

[13] Mohan, C. (2013). History Repeats Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla (PDF). Proc. 16th Int'l Conf. on Extending Database Technology.

[14] "Dynamo Clones and Big Tables" http://www.eventbrite.com/e/nosql-meetup-tickets-341739151.

[15] Garling, Caleb (2012). "Amazon helped start the "NoSQL" movement." Wired Magazine.

[16] "RDBMS dominate the database market, but NoSQL systems are catching up". DB-Engines.com. 21 Nov 2013. Retrieved 24 Nov 2013.

[17] 'NoSQL (Not Only SQL)". NoSQL database, also called Not Only SQL

[18] Fowler, Martin. "NosqlDefinition." Many advocates of NoSQL say that it does not mean a "no" to SQL, rather it means Not Only SQL.

Unstructured Databases

To this point we have discussed the databases most familiar to actuaries. In the era of "Big Data", unstructured data is starting to play a substantial role in the world of data and analytics. Unstructured data includes any data whose structure is not compatible with the data warehousing structures discussed above. The incompatibility may be due to:

- Volume: databases too lard for data warehousing (e.g. weather data)
- Variety: data with formats incompatible with traditional warehousing (e.g. images)
- Velocity: data generated and delivered too frequently to be organized into a warehouse (e.g. telematics)

XML and geospatial data are often called "semi-structured" data as they contain their own inherent structure, but as their structure requires transformation before it can be combined with fully structured data.

Source Description	**Organization Name**	**Web Link**
Introduction to Databases, a free online course that covers database design and the use of database management systems for applications.	Platform by Coursera, class taught by Jennifer Widom of Stanford University	https://www.coursera.org/course/db
SQL Tutorial that teaches you how to use SQL to access and manipulate data in: MySQL, SQL Server, Access, Oracle, Sybase,ADB2, and other database systems.	W3Schools.com	http://www.w3schools.com/sql/default.asp
SQL Server Certification and Training videos for professionals working toward earning a Microsoft Certified Solutions Analyst (MCSA): SQL Server certification	Microsoft Virtual Academy	https://mva.microsoft.com/colleges/mcsa-sql
Kimball R, Ross M. 2013. The Data Warehouse Toolkit. Third Edition. Indianapolis (IN)	John Wiley & Sons.	http://www.kimballgroup.com/

Source Description	**Organization Name**	**Web Link**
Wikipedia	Various	• https://en.wikipedia.org/wiki/Operational_data_store • https://en.wikipedia.org/wiki/Column-oriented_DBMS • https://en.wikipedia.org/wiki/Graph_database • https://en.wikipedia.org/wiki/NoSQL • https://en.wikipedia.org/wiki/Unstructured_data