Thomas Setzer

Data-Driven Decisions in Service Engineering and Management

Today, the frontier for using data to make business decisions has shifted, and high-performing service companies are building their competitive strategies around data-driven insights that produce impressive business results. In principle, the ever-growing amount of available data would allow for deriving increasingly precise forecasts and optimised input for planning and decision models. However, the complexity resulting from considering large volumes of high-dimensional, fine-grained, and noisy data in mathematical models leads to the fact that dependencies and developments are not found, algorithms do not scale, and traditional statistics as well as data-mining techniques collapse because of the well-known curse of dimensionality. Hence, in order to make big data actionable, the intelligent reduction of vast amounts of data to problem-relevant features is necessary and advances are required at the intersection of economic theories, service management, dimensionality reduction, advanced analytics, robust prediction, and computational methods to solve managerial decisions and planning problems.

1 Introduction

Increasingly automated data capturing, the ubiquity of sensors, the spread of smart phones, and the penetration of life by social media leads to enormous and ever growing amounts of data. Novel technological advances in analytics and scalable data management promise to facilitate the capturing, storage, searching, sharing, analysing, and visualisation of relationships and trends hidden in large, high-dimensional data sets.

While, traditionally, scientists in areas such as meteorology, genomics, physic simulations, or environmental research were primarily faced with the challenges of exploring large, very high-dimensional data sets, today such challenges also affect areas like business informatics. In particular service design and management need to process data in order to spot business trends, determine and anticipate bottlenecks and quality of service, or prevent customer churn by identifying churn risk and triggering appropriate actions, to name only a few tasks. In general, enterprises that can use their data quickly and correctly can gain efficiency through data-driven decisions, anticipatory action and accelerated service support and delivery processes. As an example, those companies can utilise knowledge extracted from past customer behaviours to better understand customers in order to better convince them with smart, individualised offers and services.

1.1 Service Management

Traditionally, the aim of service management is to optimise service-intensive supply chains, which are typically much more complex than the supply chains of typical goods. Those require tighter integration with field service and third parties and must also accommodate inconsistent and uncertain demand by establishing more integrated and more robust information flows. In addition, most processes must be coordinated across numerous service locations. Interestingly, among typical manufacturers, after-sale services (support, repair, maintenance, etc.) comprise less than 20% of revenue. Among the most successful companies, those same activities on average generate more than 50 percent of their total profits

(Accenture 2006). This is one of many observations indicating that a profound understanding of customers and business partners and establishing high-quality service and information management is of crucial importance.

However, today enterprises provide an increasing number of services in an automated or semi-automated fashion by means of information technology (IT services), where customer behaviour and experience can only be 'observed' by tracking what a customer is doing, in particular how he uses one or more services over time. Providers even of IT-only services can no longer afford to focus on technology and their internal organisation, but need to consider the quality of the services they provide and focus on the relationship with customers. IT service management (ITSM) refers to the implementation and management of high quality IT services that meet the needs of customers. ITSM is performed by IT service providers through an appropriate mix of people, process and information technology (Office of Government Commerce (OGC) 2009).

Unfortunately, in particular with IT services, providers typically do not receive regular direct customer feedback that is required for marketing, further service improvements, and service innovation. However, there is an ever-growing amount of information how a customer uses a services (e.g., sensors of a rental car, log files of a Web-shop, browsing behaviour in on-line manuals, etc.), and these datasets can be analysed to get 'implicit' feedback as described for example in Choi and Ahn (2009).

1.2 Advanced Analytics

In fact, today's service enterprises have more data at hand about their markets, customers, and rivals than ever before. Analysing those vast amounts of historical and current data in an autonomic or semi-autonomic fashion allows for predicting service demand and usage, customer behaviour, and market dynamics. In addition, it allows for identifying novelty patterns in customer behaviour and improving short and long-term performance of enterprise business systems, which is vital for running a competitive service company.

In 'Competing on Analytics: The New Science of Winning', Davenport and Harris (2006) argue that the frontier for using data to make business decisions has shifted. Many high-performing companies are building their competitive strategies around data-driven insights that generate impressive business results. Those companies use advanced analytical procedures, sophisticated quantitative and statistical analysis and predictive modelling. Examples of analytics are the usage of novel tools to determine the most profitable customers and offer them the right price, to accelerate product innovation, to optimise and integrate supply chains, and to identify the major drivers of financial performance. Many examples from organisations such as Amazon, Barclay's, Capital One, Harrah's, and Procter & Gamble are presented, showing how to leverage analytics to drive business. However, various potential definitions for advanced analytics exist. Typically, the 'advanced' indicates quantitative, predictive or prescriptive models as described later in this paper.

1.3 Big Data Analytics

Over the last two years, the term Big data is propagated by major companies offering information management software such as Intel[1], SAP[2], or IBM[3], and has become more and more a synonym for data analysis and advanced analytics. For many SMEs and also for larger companies, this is in some sense counter-productive as nowadays enterprises collect massive amounts

[1] http://www.intel.de/content/www/de/de/big-data/big-data-analytics-turning-big-data-into-intelligence.html

[2] http://www54.sap.com/pc/tech/in-memory-computing/hana/software/analytics/big-data.html

[3] http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

of various metrics, such as historical sensor, monitoring, and customer usage data, hoping that the data will turn out to be useful one day for prediction and optimisation.

Accordingly, as Big data analytics is now a popular topic for management, many information management companies offer tools and solutions to extract and project relationships between a vast amount of high-dimensional data vectors (structured, semi-structured, or unstructured ones), and to process, reduce, correlate and interpret data in a much more flexible fashion compared to traditional database management and business intelligence systems.

Over the last years, enterprises such as Software AG, Oracle, IBM, Microsoft, SAP, EMC, and HP have spent more than $15 billion on software firms only specialising in data management and analytics. Since the last three years, this industry was worth more than 100 billion US-dollars and was growing at around 10 percent a year: about twice as fast as the software business in general (The Economist 2010).

1.4 The Curse of Dimensionality

While in principle the vast and ever-growing sets of available data would allow for deriving increasingly precise predictions and optimised planning and decision models, the complexity resulting from the consideration of large volumes of multivariate, fine-grained, often noisy and incomplete data leads to the fact that relationships within the data are not found, algorithms do not scale, and traditional statistics as well as data-mining techniques collapse because of the well-known curse of dimensionality (nowadays also called the curse of big data) (Bellman 1961; Lee and Verleysen 2007).

Despite these dimensionality-intrinsic problems, biases in how data are collected, a lack of context, gaps in what's gathered, artefacts of how data are processed and the overall cognitive biases that lead even experienced researchers to determine non-existing patterns (and vice versa) shows that even if a company has Big Data, making use of such data typically not only requires appropriate tools but also data scientists with expertise and know-how, hacking-skills, domain knowledge, and deep mathematical and data management skills; unfortunately, as of yet data scientists of that sort are still a very scarce human resource (Davenport and Patil 2012).

The result is that – in practice – data are often collected and then ignored or aggregated in a problem-agnostic fashion, and finally for most problems rather simple and conservative solution heuristics are applied by rules of thumb or using coarsened data. The authors of this article are not aware of many companies besides the financial institutions and telecommunications companies that make excessive use of their collected data; however, most enterprises spend an increasing amount of money and effort in monitoring systems and data collection. That is also the outcome of numerous studies and expert interviews conducted and summarised by Ross et al. (2013).

Interestingly, already today leading data scientists are telling us that Big Data can and must be reduced intelligently to small data, so that finally for most decision problems one does not need Big Data at all.[4,5]

1.5 Collecting the Right (Amount) of Data

Large, global companies already recognise that there is a need to stop collecting more data and start a focused collection of *the right* data required to make decisions and to run a business successfully (Nokia Siemens Networks 2013).

Suppose a company is gathering the right data: attributes and dimensions really relevant for planning and decision-making. There is still the question whether the return on adding more data

[4] Big Data: Maybe You Don't Need It : http://www.datacenterjournal.com/it/big-data-dont/

[5] Most data isn't big, and businesses are wasting money pretending it is:
http://qz.com/81661/most-data-isnt-big-and-businesses-are-wasting-money-pretending-it-is/

points diminishes after passing a certain volume of data collection, or certain data granularities (such as monitoring intervals), and if – in a particular situation – gathering additional data will cost more than it will actually yield.

Cleary, an answer to that question depends on the concrete enterprise planning and decision problem, the importance of the problem, the scalability of engines/algorithms processing the data, the tolerance of the algorithms regarding artifacts and noise, the skills of the managers processing and interpreting the data, and many more factors.

However, independent of particular problems and individual factors as aforementioned, the answer also depends on purely statistical or mathematical criteria regarding redundancy and noise within the datasets. That is because such criteria can determine if another piece of data can bring novel information at all, or whether it can be fully or approximately derived from data already available (for example by means of collaborative mechanisms such as regression or causal reasoning).

Furthermore, for reasons of robustness and scalability it is disadvantageous to parametrise prediction models and mathematical decision programs with correlated or even collinear data vectors. In fact, efficient decision mechanisms should be rather elastic and adaptable to the anatomy and the information contained in the input data, while today typically the signatures and internal algorithms of enterprise decision modules are of rather static nature.

Consider a resource allocation mechanism for enterprise services in a data centre. If demand forecasts were expected to be highly precise for certain indicators over a defined period of time, a rather aggressive allocation mechanism operating with deterministic demand curves would be appropriate. Once the demand prediction tool downgrades its confidence levels and shrinks the horizon of the look-ahead period considered as reliable, more conservative allocation mechanisms might be appropriate.

If the forecasting horizon approaches zero time intervals, conservative online mechanism should be applied that allow for handling unexpected demand phases immediately, as sophisticated offline-planning would not beneficial in such situations: plans would be invalid shortly after their computation.

This paper reviews theory and practice of data reduction in service management with regard to the various targets addressed with the different data reduction techniques. First, we argue that a really efficient and intelligent data reduction requires the prior definition of business problems and algorithms how to address these problems with reduced data. Second, we argue that mathematical programs and algorithms for planning and decision-making should not be applied in a data-agnostic fashion. In contrast, programs and algorithms should be sensitive and adjustable to available data and the amount of dependencies, reliability, and stochastics within data, which typically vary over time, use-case, domain, and planning horizon.

2 Data Understanding and Reduction

The first and most important step in analytics is a proper understanding of the available data, the involved variables and how these are measured. Data quality, appropriate data cleaning and handling missing values as well as detecting outliers and errors must be performed prior to any data analysis. Knowing that data preprocessing is arguable the most complex and time-consuming step in analytics, for now we assume these tasks have been already performed.

We will now characterise various techniques to reduce data to relevant features, structures, and developments. In order to separate approaches aimed at descriptive, predictive, and prescriptive analytics, we will group the techniques accordingly. Descriptive analytics will be further differentiated in simple aggregations (Sect. 2.1),

and approaches that exploit statistical dependencies in and between data objects and variables (Sect. 2.2). In Sect. 2.3, we focus on data mining approaches aimed at gaining knowledge from the data to reduce uncertainty regarding the realisation of a particular variable (or label). A typical task would be the determination of the probability of a positive response of a customer, and the determination of data (features) necessary to learn this probability. In Sect. 2.4 we then summarise approaches to predict whole vectors or time series. Finally, in Sect. 2.5, we focus on prescriptive data reduction techniques that differ from prescriptive techniques as data selection and reduction needs to be aligned with a particular, potentially combinatorial and computational very complex mathematical optimisation problem. In the latter case, the goal is not only to gain insights and reduce uncertainty of future values of data, but to select and transform data in a way that is beneficial for solving a particular planning and decision problem

2.1 Data Aggregation for Descriptive Service Analytics

The purpose of aggregating data for descriptive service analytics is to summarise what happened in the past. For example, in Web analytics metrics are considered such as *number of page views*, *conversion rates*, *check-ins*, *churns*, etc. There are literally thousands of such metrics, on their own typically simple event counters. Other aggregations for descriptive service analytics might be the results of simple arithmetic operations, such as *share of voice*, *average throughput*, *average number of positive responds to a campaign*, etc. Most of what the industry called analytics is nothing but applying filters on the data before computing the descriptive statistics, sometimes combined with a linear statistical forecast. For example, by applying a geo-filter first, a company can get metrics such as average revenue per week from USA vs. average revenue per week from Europe. Structuring aggregated data to reports derives the well established and broadly used reporting functions based on information stored in data warehouses. Management dashboards usually provide the means of presenting such aggregated data to managers to support their business decisions.

2.2 Data Compression and Approximation

The most generic way to reduce (and not just aggregate) data is to exploit dependencies in and between data vectors – in a problem-agnostic way – by multivariate statistics and matrix approximation techniques, mostly based on linear algebra. Examples are variance-preserving approximation techniques such as *Empirical Orthogonal Defactorisation* derived by *Eigen-approaches* such as *Truncated Singular Value Decomposition* or *compact Principal Component Analysis* (*PCA*). More and more, techniques such as *Independent Component Analysis* (*ICA*) are applied to derive more meaningful features (in contrast to solely reducing data). By exploiting communalities, such techniques are very useful to reduce data to the maximum amount of variation (as a proxy for information) in the data sets and are often shown to derive the best low-dimensional approximation of data in very useful mathematical senses such as the L_2 norm.

Other examples are topology-preserving techniques such as *Local-Linear-Embedding* (*LLE*) (Roweis and Saul 2000) or *isoMap* (Tenenbaum et al. 2000), where the objective of data reduction is not to capture maximum variance of the data sets with fewer dimensions, but to preserve the topology of the data objects, i.e., their distance relationships.

Likewise, multivariate techniques such as vector quantisation and linear and non-linear regression techniques fall into this category of data reduction according to pre-defined mathematical objectives.

2.3 Data Reduction by Information Gain and other Criteria

Unlike the approaches described in Sect. 2.1 and Sect. 2.2, the analysis step of discovering knowledge in databases is aimed at discovering patterns in sets of data involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal is to extract pattern in a data set and transform it into structural dependencies for further use. Aside from the raw analysis step, it involves database and data management aspects, inference considerations, interestingness metrics, complexity considerations, post-processing of identified structures, visualisation, and on-line updating mechanisms.

Typical goals are the automatic or semi-automatic analyses of large quantities of data to extract previously unknown patterns such as groups of data records (segmentation analysis), unusual records (anomaly detection) and dependencies via association rules, decision trees, or other methods. For instance, data mining techniques might identify multiple groups in the data, which can then be used to obtain more accurate prediction results and more focused marketing campaigns by a decision support system. Here, data is reduced to gain information about the general structure of the data (clustering), or the class prediction of records with an unknown label due to similarities with other records where labels are already known.

As discussed in Sect. 1.4, clustering and classification do not perform well with high-dimensional data because of the curse of dimensionality. Beyer et al. (1999) and Aggarwal et al. (2001), amongst others, have shown that standard measures for proximity or distance that are used for standard *k-means clustering*, are becoming more and more meaningless with growing dimensionality. To circumvent this problem, approaches as proposed in Aggarwal et al. (2001) introduce novel distance calculations that are still meaningfull even in high-dimensional data space of 15 dimensions and more. Alternative streams of research (see Tsymbal et al. 2002 as an example) propose approaches that do not work (cluster) on original data but on reduced data as a result of compression steps as described in Sect. 2.2.

2.4 Data Reduction for Predictive Service Analytics

Predictive analytics is based on information extracted by the three previous data understanding and reduction steps; it uses all of the gained insights to make robust prediction of developments of important indicators, metrics, and variables (Stewart et al. 2012).

An intuitive way to understand predictive analytics is to apply it to the time domain. The most familiar predictive analytic tool is a time series model (or any temporal model) that summarises past trajectories found in the data, and use either auto- or (lagged) cross-correlations and regression to extrapolate time series to a future time where data is not yet existing. This extrapolation in the time domain is what scientists refer to as forecasting or prediction.

Although predicting the future is a common use case of predictive analytics, predictive models are not limited to predictions in temporal dimensions. Such models can theoretically predict anything and, hence, predictive analytics are somewhat overlapping with data mining and knowledge extraction as described in Sect. 2.3. The predictive power of a model needs to be properly validated by criteria addressing the robustness of the prediction such as using pre-whitened predictors, perpendicularity of predictors, by using information criteria such as BIC or AIC, and finally out-of-sample testing using consecutive samples. The essence of predictive analytics, in general, is that we use existing data to build a model. Then we use the model to predict data that doesn't (yet) exist.

However, only with concrete use cases in terms of business problems in mind, one can decide which pieces of information in the data set are ultimately relevant for a company, and which

pieces are not. This brings one directly to data reduction for prescriptive analytics that will be described in the next subsection.

2.5 Data Reduction for Prescriptive Service Analytics

Prescriptive analytics not only predicts a possible future, it predicts multiple futures based on the decision maker's actions. Therefore a prescriptive model is, by definition, also predictive and significant effort must be undertaken to guarantee internal and external model validity. As it is seen today, a prescriptive model is actually a combination of multiple predictive models running in parallel, one for each possible input. Since a prescriptive model is able to predict the possible consequences based on different choices of action, it can also recommend the best course of action for any pre-specified outcome, given the data set used to predict the future (together with its confidence or uncertainty). The goal of most prescriptive analytics is to guide the decision maker towards decisions that will ultimately lead to an (near) optimal and robust business outcome.

In prescriptive analytics, one also builds a predictive data model. However, the model must have two more added components in order to be prescriptive. A company not only needs a rigorously validated predictive model, the model must be actionable, i.e., managers must be able to take actions supported by the model. In addition, the prescriptive model must have a feedback system that collects feedback data for each type of action, which will additionally increase data volume by some orders of magnitude. Therefore, prescriptive analytics is very challenging even with scalable data infrastructures and the talent/expertise to make sense of the feedback data (e.g., sensitivity analysis, causal inference, or risk models).

That makes prior data reduction even more important and requires a focus on the pieces of input data really relevant for decision-making and optimisation.

3 Information Gain versus Optimisation Gain

Each department of a service provider has a set of typical tasks to perform on an operational, tactical, or strategic level. Taking for instance the Customer Relationship Management (CRM) department. CRM is aimed at the optimisation of a company's interactions with current and future customers. Objectives of CRM are the reduction of overall churn by adequate customer service and support, or by identifying and rewarding customers that have been loyal over a period of time but now show certain behaviours that increase churn probability (reduced call frequency, churns of neighbor nodes in the telecommunication network, etc.) Another objective might be the identification of customer segments for particular campaigns such as cross-selling offers based on score-values of customers. Scores are derived by data analytics and reflect the probability of a certain customer to respond positively depending on a customer's profile and past behaviour. Such procedures are aimed at gaining information from datasets regarding the probability of an unknown label in data records (for instance, class predictions such as churn: yes/no, upselling: yes/no, etc.) and are in the primary focus of business intelligence solutions.

However, usually strict business rules exist that complicate the selection of target customers. As a simple example, consider the case where one single customer is not allowed to be contacted more than twice a year (a common rule-type in telecommunications companies' campaign management). This in fact leads to predictive and finally to prescriptive analytics, as combinatorial decision problems based on expected behavioural developments of customers are required (beyond the calculation of current scores). Besides a customer's score-value for a planned campaign, knowledge of future campaigns are of importance as well as on future developments of customers in order to predict their responses. In addition, it has been shown by Goel and Goldstein (2013), amongst others, that the structure

of the communication or social network and the prediction of future behaviour of a customer's neighbors play important roles, which brings a decision maker to network models, multivariate forecasting models and collaborative prediction.

While there is a huge body of knowledge of broadly used methods and sophisticated tools exist to perform individual tasks such as classification, time series prediction, or mathematical optimisation, the integration of these tasks to derive efficient and robust overall solutions is still left to the expertise and preferences of individual decision makers, typically based on trial-and-error procedures or rules of thumb.

For each task, different data reduction techniques and feature-combination might be adequate, while the interplay of these tasks might lead to the fact that certain data considered as highly relevant in one task might not or only slightly impact the overall solution (and vice versa). For instance, it might turn out that the prediction of features relevant to compute current scores are too difficult to predict for future campaigns and the forecast cannot be considered as reliable. Formulating a stochastic optimisation model might reveal that the solution is highly sensitive to even small planning errors or rather insensible to larger ones, which makes the predictability of a feature either less or more important. Hence, each type of problem requires individual data and model selection procedures if the goal is to make optimal decisions.

This leads to a novel concept in prescriptive analytics that we will refer to as *optimisation gain* of data. Optimisation gain differs from information gain (or derivatives such as information gain rations, GINI, etc.) or matrix approximation quality norms of a residual matrix. Those metrics are aimed at quantifying the quality of a data prediction or approximation without contextual knowledge on how information is used in subsequent optimisation steps.

By optimisation gain we mean the dependency of a solution (the solution quality) derived by a mathematical model or algorithm to additional data, which might be more fine-grained data, more data in terms of a longer reliable planning horizon, or simply an additional attribute or dimension under consideration.

Optimisation gain also differs from concepts such as sensitivity, robustness, or stability of a solution. With optimisation gain we address the different and more general problem of quantifying, if (and how much) the optimality or robustness of a solution would benefit for example from the consideration of a novel data feature in a particular planning or decision problem. Addressing such questions is challenging as this typically requires the re-formulation of the mathematical program formulation for numerous input-data combinations and transformations. The intuition of optimisation gain is the quantification of the solution quality expected with different input data for a particular type of optimisation problem analytically, without expensive and time-consuming (and potentially infeasible) trial-and-error-procedures. The vision is a new generation of criteria by integrating data and model selection and configuration.

Please notice that optimisation gain can become negative as too many parameters can lead to an explosion of the search spaces and increased complexity, where optimal solutions are much harder to find. For instance, node-sets of branch & cut solvers might increase dramatically, and the quality of solutions that can be found in pre-defined periods of time might decline sharply with the number of features and constraints under consideration. Furthermore, models operating with too many data dimensions are more likely subject to over-fitting as artefacts and collinear configurations of (stochastic) variables used as model-input worsen the quality of decision-making. From a business perspective, the marginal gain of considering more data might further decline as collecting and managing data comes at additional costs for data scientists that need to analyse the data, as well as costs for monitoring, IT infrastructures, storage, and licenses.

We argue that the role of optimisation gain of data is a highly relevant concept in prescriptive analytics, and key to reducing Big Data efficiently to a manageable and actionable set of features Also, INFORMS, the leading scientific and professional organisation for OR professionals, decided to stake its claim on the analytics movement. The organisation recognised that the trend toward data-driven and analytical decision-making presents tremendous opportunities and challenges for OR professionals (Libertore and Luo 2011). Since 2009, INFORMS organises an own conference at the intersection of analytics and OR named *Business Analytics and Operations Research*, with a focus on how to apply data science to 'the art of' business optimisation. It features presentations on real-world applications of analytic solutions, presented by industry and university leaders.

Optimisation gain can provide a means of significantly reduce the effort spent for monitoring, collecting and managing data, as ideally only data is collected that is indeed supposed to improve decisions. Unnecessary frequent measurements are also avoided as the collection of correlated data that is (statistically) already captured by other variables. These ideas are closely related to visions such as smart measurement and collaborative monitoring systems, but with an additional focus on the impact on the business relevance of gathered data. We will further detail on this in Sect. 4.

4 Feature-based Optimisation and Model-Data-Integration

As aforementioned, certain units in enterprises have specific tasks to perform, usually composed by structured or at least semi-structures processes. For instance, in IT service management, the role of capacity management is to ensure sufficient capacity to provide high-quality services to customers efficiently, i.e., at reasonable (low) costs to the business. In capacity management, it is important to have a clear picture of the expected service demand and the corresponding resource demand that needs to be supplied in future points of time. Considering the case of private clouds, with the potential of hosting services in virtual machines (VM) in a flexible manner, e.g., by co-hosting VMs temporarily on the same physical server, sharing and multiplexing a servers capacity for resources such as CPU, memory, or I/O. In such an environment, IT service managers try to minimise the number of servers by assigning enterprise services in virtual machines efficiently to physical servers, but at the same time provide sufficient computing resources at each point in time. It is worthwhile to notice that running servers are (independent of their utilisation levels) the main energy drivers in data centers, where energy costs already account for 50% or even more of total operational costs (Filani et al. 2008).

Without going into too much detail, the resulting VM allocation problem can be reduced to a stochastic multi-dimensional bin-packing problem, a well-known NP-hard problem. As it is the case with every bin-packing problem, the goal is to fill-up the available spaces (resource capacities) of bins (servers) as much as possible, and, hence, come out with fewer servers while not exceeding the capacity of servers, as this would result in overload and SLA violations.

Theoretically, historical workload data would allow for accurate workload demand forecasting (for more than 80% of typical operational business services) and optimal allocation of enterprise applications to servers. In various experiments and studies with smaller VM sets it has been shown that such approaches lead to a reduction of required server by around 30% (Speitkamp and Bichler 2010). Unfortunately the volume of data and the large number of resulting capacity constraints in a mathematical problem formulation renders this task impossible for any but small instances and is of little use for IT service providers with server parks of hundreds or thousands of VMs to be consolidated.

Looking at the core of each packing problem, in particular at bin-packing problems, the challenge is to find complementarity in objects to be packed (in our case, the demand profiles of VMs for various resources over time) to achieve high average server utilisation levels. It makes sense to co-host VMs with peak loads in the morning hours and VM having their peak loads later during a day. Similarly it makes sense to combine a VM with high CPU and low memory demand with one having lower CPU but high memory demand.

When we consider relevant features of workload profiles for the packing problem as aforementioned, features describing the complementarities between VM profiles could be of great value, besides features describing the absolute resource demand curves of VMs.

Setzer and Bichler (2012) use techniques based on singular value decomposition (SVD) to extract significant features from a matrix of the expected (fine-grained) demand vectors of hundreds of VMs and provide a new geometric interpretation of these features as principal demand patterns, complementary between these patterns, and uncertainty. The extracted features allow for formulating a much smaller allocation model based on integer programming and allocating large sets of applications efficiently to physical servers. While SVD is typically applied for analytical purposes only such as time series decomposition, noise filtering, or clustering, here features are used to transform a high-dimensional allocation problem in a low-dimensional integer program with only the extracted features in a much smaller constraint matrix. The approach has been evaluated using workload data from a large IT service provider and results show that it leads to high solution quality. At the same time it allows for solving considerably larger problem instances than what would be possible without prescriptive analytics, intelligent data reduction and model transform. This work provides a first example of a highly integrated data reduction and optimisation approach.

The same authors argue that the overall approach can also be applied to other large packing problems. For instance, in Setzer (2013), the authors show that high-dimensional knapsack problems can also be intelligently reduced to smaller and computationally tractable ones, as long as there is a significant amount of shared variance amongst the dimensions to be considered. Please notice that, according to recent studies, knapsack-problems are amongst the top four problems to be solved in enterprises, although managers often do not know that their particular problems could be formulated as knapsack-problems.

Overall, we believe that there is a huge potential for solving particular decision problem with Big Data made small. However, to exploit these potentials, problems must be formalised before integrated data reduction and optimisation models can be developed.

Reconsidering the example of capacity management in private cloud infrastructures, we will now detail on the need for a decision model fabric that not only aligns the model to be used to changing environments by considering novel parameters. In contrast, completely different solution techniques are required depending on the (recent) structures and developments found in the data. Again, we will use private clouds for illustration.

Nowadays, live migration allows to move VMs to other servers reliably even during runtime and promises further efficiency gains (VMWare ESX, amongst others) (Nelson et al. 2005). Some platforms such as VMware or vSphere closely monitor the server infrastructure in order to detect resource bottlenecks by tracking threshold-violations. If such a bottleneck is detected they take actions to dissolve it by migrating VMs to different servers. For instance, if the CPU utilisation exceeds 80%, a VM is migrated away from that server to reduce total server load. On the other hand, if a controller detects phases of low overall workload, there is the possibility to concentrate workloads on fewer servers by vacating

servers and shutting down these source servers temporarily to further reduce energy consumption. We will refer to such techniques as dynamic resource allocation or dynamic control, as opposed to static VM allocation where allocations are computed and kept fixed for a longer period of time.

5 Towards Data-Elastic Decision-Making

On the one hand, dynamic control strategies are more flexible and should therefore lead to lower energy costs. On the other hand, migrations cause significant additional overheads and response-time peak, which are avoided with static allocation mechanisms. It has been shown that with well-predictable workloads of business applications, dynamic resource allocation during operational business hours does not lead to higher energy efficiency compared to static allocation even if future demand is known only to a certain extend (Wolke et al. 2013). However, if demand is completely unknown, dynamic control is the only reasonable option to avoid both: massive overprovisioning and service degradation. Depending on the share of stochastic developments in workload demand curves, hybrid models might be appropriate where basic allocations are computed for a given planning horizon in a more conservative fashion, considering the option of potential migrations to cope with uncertainty.

In summary, dynamic, data-based model selection is required that differs from parameter alignment, which simply would mean that for instance the alpha parameter in an exponential smoothing model is adjusted from time to time (which then leads to a different and hopefully better short term prediction), but where the same mathematical model is used for prediction.

In the example above, depending on the predictability of demand behaviour, which might be well predictable throughout certain periods but rather unpredictable in other periods of time, completely different allocation mechanisms are advised.

6 Conclusion and Vision

Analysing historical and current data in order to make better predictions is vital for running a competitive service company. Data-driven design and management of services demand interdisciplinary knowledge from the business domain, processes, data analytics, and mathematical optimisation. While in principle the ever-growing amounts of available data would allow for deriving increasingly precise forecasts and optimised input for planning and decision models, the complexity resulting from the consideration of large volumes of ever-growing volumes of multivariate, fine-grained data leads to the fact that dependencies and relationships within the data are not found, algorithms do not scale, and traditional statistics as well as data-mining techniques collapse because of the well-known curse of dimensionality. Hence, in order to make Big Data actionable, we are interested in the intelligent reduction of vast amounts of data to small sets of problem-relevant features. We argue that mathematical optimisation and planning models need to be transformed to be able to operate efficiently on highly reduced data. In addition, the selection of adequate planning and decision models must be adapted to (current) data and the reliability of relations and predictions extracted from that data, which requires time-dynamic and data-driven model selection and evaluation techniques.

References

Accenture (2006) Service Management – Enabling High Performance Through Supply Chain Management. Accenture.

Aggarwal C. C., Hinneburg A., Keim D. A. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Database Theory – ICDT 2001. LNCS. Springer, London, pp. 420–434

Bellman R. E. (1961) Adaptive Control Processes: A Guided Tour. Princeton University Press

Beyer K. S., Goldstein J., Ramakrishnan R., Shaft U. (1999) When is 'Nearest Neighbor' Meaningful. In: Proc. Int. Conf. Database Theory (ICDT '99). Springer, London, pp. 217–235

Choi D., Ahn B. S. (2009) Eliciting Customer Preferences for Products From Navigation Behavior on the Web: A Multicriteria Decision Approach With Implicit Feedback. In: IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 39(4), pp. 744–760

Davenport T. H., Harris J. G. (2006) Competing on Analytics: The New Science of Winning. Harvard Business Review Press

Davenport T. H., Patil D. J. (2012) Data Scientist: The Sexiest Job of the 21st Century. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Filani D., He J., Gao S., Rajappa M., Kumar A., Shah P., Nagappan R. (2008) Dynamic Data Center Power Management: Trends, Issues, and Solutions. In: Intel Technology Journal

Goel G., Goldstein D. G. (2013) Predicting Individual Behavior with Social Networks. In: Marketing Science

Lee J. A., Verleysen M. (2007) Nonlinear Dimensionality Reduction. Springer, Heidelberg

Libertore M., Luo W. (2011) INFORMS and the Analytics Movement: The View of the Membership. In: Journal Interfaces 41(6), pp. 578–589

Nelson M., Lim B.-H., Hutchins G. (2005) Fast Transparent Migration for Virtual Machines. In: Proc. USENIX. Berkeley, USA, pp. 25–35

Nokia Siemens Networks (2013) Exit Big Data. Enter Right Data. http://www.nokiasiemensnetworks.com/news-events/insight-newsletter/articles/exit-big-data-enter-right-data

Office of Government Commerce (OGC) (2009) Official ITIL Website. Last Access: [Online]. Available: http://www.best-management-practice.com/IT-Service-Management-ITIL

Ross J. W., Beath C. M., Quaadgras A. (2013) You May Not Need Big Data After All. In: Harvard Business Review 91(12)

Roweis S., Saul L. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. In: Science 290(5500), pp. 2323–2326

Setzer T. (2013) An Angle-Based Approach to Solving High-Dimensional Knapsack-Problems. KIT

Setzer T., Bichler M. (2012) Using Matrix Approximation for High-Dimensional Discrete Optimization Problems. In: European Journal of Operational Research 227, pp. 62–75

Speitkamp B., Bichler M. (2010) A Mathematical Programming Approach for Server Consolidation Problems in Virtualized Data Centers. In: IEEE TSC 3(4), pp. 266–278

Stewart T., Thomas R., McMillan C. (2012) Descriptive and Prescriptive Models for Judgment and Decision Making: Implications for Knowledge Engineering. In: Expert Judgment and Expert Systems – NATO AS1 Senes 35, pp. 314–318

Tenenbaum J. B., Silva V. d., Langford J. C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. In: Science 290(5500), pp. 2319–2323

The Economist (Feb. 2010) A Special Report on Managing Information: Data, Data Everywhere. In: The Economist

Tsymbal A., Pechenizkiy M., Baumgarten M., Patterson D. (2002) Eigenvector-Based Feature Extraction for Classification. In: Proc. Int. Artificial Intelligence Research Society Conference. Florida, USA

Wolke A., Bichler M., Setzer T. (2013) Energy efficient capacity management in virtualized data centers: optimization-based planning vs. real-time control. In: Proc. INFORMS Workshop on Information Systems Technologies. Milan, Italy

Thomas Setzer

Karlsruhe Institute of Technology
Englerstraße 14
76131 Karlsruhe
Germany
thomas.setzer@kit.edu