Big Data, Big Business: Bridging the Gap

ABSTRACT
Business analytics, occupying the intersection of the worlds of management science, computer science and statistical science, is a potent force for innovation in both the private and public sectors. The successes of business analytics in strategy, process optimization and competitive advantage has led to data being increasingly recognized as a valuable asset in many organizations. In recent years, thanks to a dramatic increase in the volume, variety and velocity of data, the loosely defined concept of “Big Data” has emerged as a topic of discussion in its own right – with different viewpoints in both the business and technical worlds. From our perspective, it is important for discussions of “Big Data” to start from a well-defined business goal, and remain moored to fundamental principles of both cost/benefit analysis as well as core statistical science. This note discusses some business case considerations for analytics projects involving “Big Data”, and proposes key questions that businesses should ask. With practical lessons from Big Data deployments in business, we also pose a number of research challenges that may be addressed to enable the business analytics community bring best data analytic practices when confronted with massive data sets.

1. INTRODUCTION
The business world is undergoing a revolution driven by the use of data and analytics to guide decision-making. While many forces are at work, a major reason for the business analytics revolution is the rapid proliferation of the amount of data available to be analysed. In a report on enterprise server information, titled “How much information?” [10] the University of California estimated that enterprise servers processed 9.57 zettabytes of data globally in 2008, an amount equivalent to nearly six gigabytes of data daily for every person in the world. The same report suggested that if this information was shared out among all the world’s companies, each would have processed an average of 63 terabytes of data – enough to fill more than 13,000 DVDs.

Leading organizations increasingly recognize the importance of leveraging their data as a strategic asset. Some organizations undertake analytics initiatives to improve the quality of customer experience by measuring and acting on sentiments expressed by customers or by linking various divisions and operating units that customers tend to connect with. Others analyse data to predict a customer’s propensity to buy new products or services in order to proactively make recommendations for future purchases or offer discounts to encourage a longer-term relationship [3]. Often, data is analysed in order to fine-tune the enterprise itself, with analytical insights used to refine internal processes, promote safety, and pinpoint operational issues the resolution of which can drive up efficiency, profitability, and competitive positioning [4].

As professional services consultants, we are often approached by clients seeking help to navigate the worlds of business analytics and what has come to be called Big Data. However, while most clients are familiar with the term “Big Data”, many are still unsure about the costs and benefits of undertaking new Big Data projects. Part of the uncertainty is perhaps due to the fact that Big Data is a loosely defined, somewhat fuzzy concept. Furthermore, it is not immediately clear that “Big Data”, however defined, necessarily translates into commensurately valuable information and business insights that justify the requisite investments. So far, most practical Big Data success stories have emanated from internet giants, telecommunication companies, sensor networks and scientific organizations. The scale and throughput of operational data generated in these domains far exceeds that of most enterprises. Additionally, many clients are still seeking to optimize returns on their existing “ordinary data” analytics investments, including data warehouses and predictive analytics platforms, and are therefore reluctant to invest in additional Big Data technology. In short, the business case for investing in Big Data projects is domain-specific, and must turn on the informational content and marginal business value offered by “Big Data” over and above that of more traditional data analytic approaches.

It is therefore worth pondering over the key questions businesses should ask before undertaking Big Data projects, and how the data mining research community may help make Big Data more useful in business contexts? It is only by addressing these questions that businesses can maximize their investment in analytics, seize new opportunities and, ultimately, bridge the gap between their strategic objectives and Big Data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
BigMine’12, August 12, 2012, Beijing, China
Copyright 2012 ACM
2. BUSINESS QUESTIONS TO ASK AROUND BIG DATA MINING

We begin with a success story for Big Data analytics, from the forensics practices. Forensic investigators uncover irregularities in financial data, including errors, biases, duplicates and omissions. These problems share similarities with outlier detection and biased class distributions. The eventual goal of forensic analytics is to find out how and why – these irregularities exist and to find the source of the anomalies – especially when fraudulent activity is suspected.

Forensic analytics was applied on Big Data from the UN Oil-for-Food Program fraud investigation [6]. Under this program, the government of Iraq sold $64.2 billion in oil to 278 different international companies, and used $34.5 billion of that money to purchase humanitarian goods from 3,416 different international companies. As the program progressed, it was beset by allegations of bribes and kickbacks. In response, the UN set up an inquiry committee which directed an in-depth forensic analysis.

The scope and complexity of the analysis was herculean and nearly unprecedented. Thousands of documents described communications, events, and transactions – both paper and electronic – in nearly a score of languages and just as many currencies. All pertinent documents had to be identified, collected, correlated, and analysed for probative facts that would confirm fraudulent activity. There was also quite a bit of traditional gum-shoe activity required and investigators travelled the world in search of evidence, witnesses, and accomplices. Ultimately, the investigation uncovered $1.8 billion US in kickbacks and bribes, and about $50 million in potentially fraudulent over-charges from the UN to the program.

Forensic analytic techniques were critical to the investigation, but there was also a good match of technique to problem, and attention paid to how the analytics results would be integrated with other types of detective work. We find that similar efforts are required even in the broader domain of financial crime analysis and investigation, which typically deals with large volumes of disparate data sources such as transaction records, geospatial signals, digital multimedia, social media, etc. [13]. We generalize from these examples and propose that organizations looking at potential applications of Big Data ask themselves three questions:

- **What is the business problem, or organizational goal?** The goal provides a strategic context for making cost-benefit decisions on whether to use Big Data techniques. In the Oil-for-Food investigation, the volume and variety of data justified the application of automated analytics techniques. An earlier project involved building a credit scoring algorithm for a major insurance company. A custom set of predictive variables based on historical transactional credit data was used rather than a vendor’s canned set of variables, for a variety of reasons: anticipated marginal segmentation power, control over the algorithm for risk management and regulatory compliance reasons. Now, suppose a continuously updated credit feed is available. Should the insurance company mothball its decade-old approach to credit scoring in favor of the real-time Big Data version? The answer really depends on the marginal predictive power of the Big Data version of credit over and above what they already have. If the business goal is merely to price a six-month insurance policy, then the company probably doesn’t need to monitor daily fluctuations and change their algorithm in response. However, if the business goal is to send real-time “nudges” to policy holders to ward off risky behavior, it’s a different conversation.

- **Given the goal, is the available data suitable?** Strictly on-line information was not sufficient for the Oil-For-Food investigation, which required merging document-derived evidence with observations from more traditional detective work to build a case. But the data suitability question should be asked even in more mundane circumstances. For instance, suppose we’re asked to do a sentiment analysis of Twitter feeds. How do we evaluate the degree to which this data is representative of the population about which we want to make a decision? A famous example from US political history applies here: in the 1930’s, a poll predicted that Franklin Roosevelt would lose by a landslide to his Republican opponent. But he didn’t. The poll data was biased because it was collected over the telephone, which was a luxury item disproportionately used by Republicans during the great depression.

- **What is the return of investment on Big Data?** The Netflix Prize attracted worldwide interest and thousands of entries, but Netflix decided not to implement the winning algorithm [1] for a number of practical reasons, including that the marginal accuracy improvement of the winner was not worth the complexity of the required implementation effort. This is not atypical in business, where algorithms are evaluated against multiple criteria beyond speed and accuracy. Before investing in Big Data deployment, businesses should also keep in mind other considerations such as the cost of collecting or licensing data of sufficient quality, the costs of transferring, cleansing, transforming, and integrating data across multiple sources, effort and skill sets needed to develop and maintain models, timing of data and result availability, cost of required infrastructure, and the salience, explainability, and marginal business value of the results.

In discussing these questions with a variety of organizations, we have learned that Big Data often requires different choices than are customary in more tightly controlled internal data contexts. Moving from “ordinary” to Big Data along the dimensions of volume, velocity and variety requires managing tradeoffs among data freshness, query response time, data quality, and sometimes, even result quality. Considering such tradeoffs is either implicit or just not done in the “ordinary data” world, but in Big Data we must manage them explicitly. The alternative, in some cases, would be missing the insights afforded by the signals present in Big Data, in other cases missing the operational efficiencies gained from keeping the infrastructure as simple as possible to achieve the organization’s goals.
3. RESEARCH TOPICS AROUND BIG DATA MINING

Business users must educate themselves on new data mining technology developments to ensure that the implementation context justifies Big Data complexity. But insights can flow in the other direction as well. Lessons learned from business can motivate new research that will widen the applicability and lower the overhead of Big Data solutions. Topics that could benefit from assistance from the research community include:

1. Supporting feature engineering and selection:

We have consistently found that we often get more mileage out of engineering better features/indicator variables than out of better algorithms. For instance, we developed a relatively accurate algorithm for assessing risks of bank failures from the call reports that US banks file with the federal government, containing about 2600 pieces of information. Doing this required filtering the 2600 down to 70+ relevant variables and working with subject matter experts to filter and recast these further, in some cases transforming numeric variables into categorical ones, and vice-versa.

Besides, though many statistical and data mining methodologies can be used to perform dimension reduction, it is often challenging to extract features from huge data over a large number of dimensions. For example, consider sensing and collecting automotive data, including speed, acceleration and breaking patterns for the purpose of understanding risk profiles of individuals. The velocity and variety of the data generated from each source makes it difficult to extract relevant features and events automatically in coordination with other information about the individual. How can existing data mining tools give explicit support to make feature engineering and selection more systematic and intuitive?

2. Learning from partially labeled examples:

Classical supervised learning techniques often assume access to a large repository of examples labeled as being either positive or negative instances of the target class. In business applications, such a labeled repository is often not available or is too expensive to obtain. For instance in fraud detection, it may be possible to get positive labeled instances (examples of known fraud), but hard to get negative examples with confidence (examples known to be legitimate, non-fraud, instances). This is especially true when the underlying examples are each complex enough with hard to verify substructure, for instance in company reports summarizing millions of transactions. In medical diagnosis, there may not be certainty that a patient does not have a disease.

Partially labeled data can be considered as a type of noise/inaccuracy in the labeling target class, and it is often feasible to filter out or cleanse traditional “ordinary data” using exploratory data analysis. However, automated data cleansing approaches that break data into subsets to automatically learn rules for noise identification [14], would be difficult to implement in large volume and velocity of data streams that occur in many Big Data scenarios. How can algorithms that learn from positive and unlabeled instances, or semi-supervised or active learning algorithms that interact with users to obtain feedback on examples of interest be applied in realistic situations?

3. Managing missing data across heterogeneous data streams:

Missing data poses a multitude of challenges in the Big Data world. Traditional approaches that analyse missing data involve applying appropriate imputation techniques. For example, data that is often not Missing Completely at Random (MCAR) requires a systematic strategy for imputation and handling missing data to avoid introducing bias in the analysis. However, such techniques may not be feasible when the volume of data is too large. Even in simpler cases where data is Missing at Random (MAR), popular approaches such as Multiple Imputation are impractical in Big Data, as they require creating series of imputed data sets and carrying out the analysis model on each data set.

Nevertheless, we find that coping with the variety of data sources is often a greater challenge for many of our clients than the sheer volume of data. A common scenario is the joining of data from structured, generally internal, transaction data with unstructured, often external, data from the Web, especially clickstream behavior and social media. The joining of these data sources, whether around geography, customer name, channel, or other segments is generally very lossy, with some reporting success rates as low as 40 or 50%, meaning that over half the source instances is often missing in the merged results.

The granularity of the data among sources also varies widely, with some data available at the individual level, and other data only at the aggregate level. The temporal granularity or sampling frequency for different data sources often varies. Social media is a particular challenge as contributors use multiple channels and sometimes multiple identities. For privacy reasons, many individuals will opt-out and elect not to share their data, a trend which may increase as people understand the capabilities of new data mining methods. All of these factors mean a merged data stream may be very sparse. Different portions of the results may have different levels of confidence associated with them. Are there principled ways for assessing and improving the reliability of analytic conclusions drawn from data with such known gaps?

4. Managing online and offline learning:

Traditionally, most advanced analytics have worked with offline data, while much simpler models are implemented on streaming data. Offline models allow more complicated models and provide flexibility, but they introduce a time lag. In fraud detection systems, for example, simple rule-based approaches to flagging suspect claims or transactions in real-time are combined with the efforts of analysts working offline seeking to identify new patterns – new rules must be proposed, tested and reviewed before being put into production.
Though recent advances in stream-based outlier detection allow for online detection of new anomalies [2], how can the offline and stream-based approaches combine to learn? What kind of analytics infrastructure is needed to support such combined learning?

5. Interactive and collaborative mining:
We recently participated in a modeling engagement where the definition of classes of interest, viz., retail customers likely to migrate from one value segment to another, or customers likely to become inactive, went through quite a bit of refinement. Initial findings from the data helped the client refine intuitions about customer behavior. Additionally, clients often have knowledge/beliefs about which indicator variables are controllable, and these need to be incorporated into model development. Practical applications of Big Data and business analytics in general, involve extended iteration and interaction among people who specialize in different areas of data extraction, modeling, and business users.

Collaborative environments where data scientists and domain experts work together to achieve a better outcome are quite invaluable [5]. Current tools for analytic interaction and collaboration offer relatively little support for even simple tasks such as annotating data sources, models and presentations with shareable comments and requests. Though research on areas such as automated class description [8] may support modeling and business experts to perform collaborative mining, they are infeasible in the face of large velocity and volume of data. “Crowdsourcing” of ideas and models from thousands of contributors adds another layer of complexity, including that of infrastructure to enable Big Data processing. How can the process of such “modeling-in-the-large” be improved?

6. Visualization of Big Data:
Because of the volume of Big Data as well as the complexity of analysis, business users especially appreciate interactive visualizations and it is important for visualization capabilities to keep track with advances in technology and analytics [11], [12]. Research from the University of Maryland’s Human Computer Interaction Lab shows that useful visualizations provide a capability to overview first, zoom-and-filter and provide details on demand [9], but heterogeneous data offers challenges to this framework.

Much of the data is in the form of graphical networks with multiple types of links, where traditional visualization techniques do not scale. Research efforts on large scale analytics and visualization for climate data [7] are promising, but complexity associated with high dimensionality and multiple data types/structures remain challenging. How can we help business users draw appropriate conclusions from visualizations when the underlying data may be unrepresentative, or known with varying degrees of accuracy?

7. Privacy and transparency:
The topic of privacy comes up in nearly every client conversation we have around Big Data and analytics. Individuals that are the source and subject of data have an increasing say in what is done with that data; in some cases, as a matter of law. Many privacy-preserving data mining standards (such as k-anonymity) are promising, but complexity associated with high dimensionality and multiple data types/structures remain challenging. How can we help business users draw appropriate conclusions from visualizations when the underlying data may be unrepresentative, or known with varying degrees of accuracy?

4. CONCLUSIONS

Big Data has become the new “mother of invention”, compelling organizations to take a fresh look at their data and ask themselves whether they are using it strategically. To maintain competitive advantage, organizations must focus on a well-defined business goal, and continually assess the business case for expanding their analytics activities to encompass Big Data. Business decisions should involve statistical considerations around the meaningfulness of the information contained in the data, business considerations around the marginal business value of this information, as well as economic considerations around the costs of storing, cleansing, visualizing, and analysing increasingly big data sets.

Research communities should be cognizant of these business imperatives, and make be encouraged to shape their Big Data research agenda to support business innovation. The data mining research community has manifold opportunities to provide the larger business analytics community with the tools and methodologies needed to approach “Big Data” in a scientific, practical, and economic fashion.

The authors welcome discussion around any of the thoughts above.
5. ABOUT DELOITTE ANALYTICS

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee, and its network of member firms, each of which is a legally separate and independent entity. Please see www.deloitte.com for a detailed description of the legal structure of Deloitte Touche Tohmatsu Limited and its member firms.

Deloitte provides audit, tax, consulting, and financial advisory services to public and private clients spanning multiple industries. With a globally connected network of member firms in more than 150 countries, Deloitte brings world-class capabilities and high-quality service to clients, delivering the insights they need to address their most complex business challenges. Deloitte’s approximately 195,000 professionals are committed to becoming the standard of excellence.

Deloitte has created the Deloitte Analytics Institute as a strategic investment in business analytics. The Institute, based out of centres in the US, UK and Asia, is dedicated to thought leadership arising from research and innovation. The Institute’s efforts span all of Deloitte’s functional businesses, addressing opportunities in business intelligence, data management, statistics, change management, technology, automation, risk, and governance. There is a particular focus on data-intensive sectors such as banking & insurance, telecom & media, energy & resources, life sciences, consumer goods, manufacturing, public sector, transportation, and healthcare.

Disclaimer This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the “Deloitte Network”) is, by means of this publication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

© 2012 Deloitte & Touche Financial Advisory Services Pte Ltd.

6. REFERENCES