3

Frontiers of Big Data Business Analytics: Patterns and Cases in Online Marketing

Daqing Zhao

CONTENTS

INTRODUCTION

Computer technologies have changed our lives dramatically. The changes are still happening at an accelerating speed. Without a doubt, the digital information revolution will continue to change our society and culture.

As technologies advance, we have more and more ways to collect data. Using sensors, anything from our medical information to our Web surfing history, energy usage of our homes, and things that can be seen or heard or in some way measured now can be digitally recorded and stored. Digital data can be analyzed much better using computers and statistical tools than analog data. Computer technologies have the characteristic of increasing capability while lowering cost over time. Moore's law says that the number of transistors in an integrated circuit doubles every 18 months. Thanks to Moore's law, which has been true for decades, we get new computers with more processing power, larger storage, and wider network bandwidths at lower costs. As a result, we can collect more and more data, store and access them, as well as analyze them in more detail. Information which used to be too expensive to gather is now readily available. Data are being accumulated at an accelerating speed. With the abundance of data, more and more technical solutions for handling and utilizing the data are developed.

From the dawn of civilization to 2003, a total of five exabytes (one exabyte is one million terabytes) of information were collected; in 2010, collecting that amount took only two days (Siegler, 2010). New data sources include not only structured text and numerical data but also unstructured, free-format data, such as images, audio, and videos. Most data now are behavioral or sensor data in digital form, rather than insights and knowledge we are accustomed to seeing in print media. Data alone, without analysis, are not actionable. From sciences to government to companies, because of the limited number of people with data analytics expertise, more data are collected than can be analyzed. Most new data are stored and stay dormant. In time, this situation will only get worse. This is the big data era (Dumbill, 2012).

With the Internet and mobile technologies, people and devices are increasingly connected. A visitor can come from anywhere on earth to get information or do business, in the process leaving a trail of evidence of preferences and interests. Using a network, a large number of sensors can be connected and data aggregated into a single data set. Via the Internet, data can be shared and analyzed, and information can be consumed by a large number of people.

There are many examples of big data (Cukier, 2010). Now, collecting information about each and every visitor to a website is not only possible but necessary to optimize to achieve reasonable user experience and effectiveness. In astronomy, right now far more data about the universe are being collected than could be analyzed. In medicine, real-time information about a patient is available through small devices including smartphones. Together with lifestyle, behavioral data, and genomic information, doctors can use new information to improve patient's health significantly. Not only smartphones but smart TVs and smart homes all will collect more and more data about consumers. Every field has been or will be changed by the large amount of data available.

Before the advent of commodity storage and computing solutions, only the most important data were recorded in detail, such as financial data. Other data were collected only as samples and surveys. Web server log data were quickly purged without any detailed analysis. In the big data era, companies are collecting every page view, every click, every blog, and every tweet, as well as pictures and videos customers generate, in addition to transaction data, customer services data, and third-party data, to provide information about customers. A company may know more about its customers than not only families and friends know but also the customers themselves, which may be a scary thought. We may not remember all the websites and pages we visited during the last month, but web server logs never forget. We may not know many things about our friends, but information about them indirectly tells who we are.

Organizations and society are not yet ready to digest and to use information from the increasingly abundant data. Companies don't have enough data-savvy business managers to work with the data and turn them into business advantages. The bottleneck is not computing power but people, analysts and managers, operational processes, and culture.

BIG DATA ANALYTICS

Computers will not be able to outsmart humans in the foreseeable future. One reason is that the computing power of a single human brain is about the same as all the world's computers combined (Hilbert and López, 2011).

After millions of years of evolution and optimization, our brains have many features that are hardwired, but they are not yet adapted to handle a large amount of digital data. In processing data, computers have advantages in many ways, while humans have advantages in others. Computers are powerful tools to help people, and humans also need to learn to work with the technologies.

Things Computers Are Good At

Computers (including storage) have perfect memory, since they can record everything, every event of everyone. In the big data era, this is especially the case. Do you remember what you ordered for lunch for the last year? Or how much on average you spent on lunch? How about this kind of data for everyone in the country? Such information is readily available in the data customers left with their credit card processing companies. What did we say at some time in the past? Spoken words in a person's lifetime can now be easily stored in a thumb drive.

Computers are also very good at searching through a large amount of data to find a needle in the haystack, to identify fraud, to find evidence of criminal activities, to make the one-in-a-million perfect match, or to retrieve and send you the piece of information you are searching for. As the volume of data increases, the marginal value of additional data is lower. Using computers to handle more and more repeating tasks is the only scalable way to utilize big data efficiently.

Computers are very good at calculating tradeoffs among a large number of factors to come up with a conclusion. For example, let's say there is a potential customer, female, age 25–34, has a child less than 5 years old, Asian, earns $30K, rents a home, divorced, lives in zip code 90001, some college education, visited sites of Walmart, Coupons.com, Monster.com, drives a Toyota Camry, etc. Is she a buyer of product X? Computers can do much better than the best analyst, in milliseconds, remotely over the Internet. Credit scoring is another example. Even if our analysts are given all the information about customers, without the computer to do the calculations, we still won't be able to say how good their credit is. For a few customers, the analyst may have the advantage of meeting them to read more based on intuition, but in scale, the computer clearly wins. A model cannot tell whether an individual will have the behavior, but predicts how likely the behavior happens in a large number of people with similar profiles.

Given data, computers can help us build models to find repeatable patterns. Computers are very good at optimizing model parameters to predict how likely is it that some behavior will happen, using data of many similar people and their known behaviors. Using statistics and machine learning methodologies, computers are very good at finding out what insights or predictions we can get from the data, as well as what we cannot, and to what level of accuracy.

Events just don't happen in isolation. We may think of ourselves as individuals with our own freedom and judgment, but how we make decisions largely depends on who we are and what environment we are in. Our behaviors strongly correlate with those of our friends and neighbors. Before making a purchase, we inevitably have a sequence of activities, and we leave signals in our demographic profile, socioeconomic status, background, values, lifestyles, and preferences. When events happen, there is often some evidence left behind. If we collect a lot of data, we often find direct or circumstantial evidence of the event or behavior.

Once we have built models from the data to describe quantitatively how relevant a given set of variables and our concerned events are related, we can use the models to see what happens under some given scenario. This is computer simulation. Computers make extensive simulations possible. By selecting possible future scenarios, we can use computers to see how the concerned metrics change. This is just like flight simulators.

Computers can help us optimize using the models. Through generation of a large number of scenarios, including factors we can influence, we can evaluate which scenarios are most favorable or desirable. This is the most sophisticated use of computer modeling. We can try to get more of the good ones and fewer of the bad ones and to design strategies to best handle the situations. This is how we realize the value of data. The more data we have, the better model we have, and the better we can optimize. Most companies have managers look at the data at some level of aggregation and digestion and try to find value and opportunities to optimize using their heads. But as we discussed earlier, people are not good at estimating complicated tradeoffs among a larger number of factors.

Computers make scalable personalization solutions possible, offering the right information or product to the right people at the right time. Large-scale personalization is a great application of big data analytics. There is this narrative that the owner of a mom-and-pop store knows all of her good customers and builds personal relations, providing services tailored to their needs and preferences. As superstores come along, prices are lower because of scale, but at the expense of customer experience of personalized services. With computers and the Internet, companies now can know enough about the customers through the collection and analysis of a large amount of customer-level data. Large vendors now can provide personalized services at lower prices in scalable ways. The value of such personalization of services becomes more compelling as the cost of computers and storage continues to drop. Personalization solutions require not only customer data but also the computer power to do deep analysis on the data, as well as detailed data on products and services.

In addition to help improve services to customers, big data will allow companies to have better competitive intelligence (CI) as well. Companies can collect more detailed data about their own customers, products, and processes. Considering data as a valuable asset, they are very reluctant to share with competitors. It is usually more difficult to collect data about competitors. In order to gain insights of CI, companies often use syndicated data vendors, such as Nielsen and comScore, for services ranging from standard reports to custom data collections and analyses. In the big data era, individuals, organizations, and their relations are all more visible. Having easy access of customer sentiment and behaviors on the web, with a large amount of data from public sources as well as data vendors, inexpensive sensor data collections, and computer resources, companies will be able to have more comprehensive and accurate information about their competitors at lower costs. Data on the competitive environment should be part of the drivers for business decisions and optimization. At the same time, it is also more and more difficult to do business in scale and remain under the radar.

Computers Can't Do Everything

Even though computers can help a lot, they are only as good as the analyst who uses them. They follow the analyst's instructions.

Data, especially big data, are often disorganized and overwhelming like runoffs. Data may not have a taxonomy and context, and often there is no sufficient documentation. Some key data for some specific interest may not be collected at all. And then for sure no one, with whatever computer resources, would be able to make good predictions. Data are unreliable before they are thoroughly analyzed. Data collection is usually an engineering function. After building the data acquisition system, some data are collected and put into storage. Some quality assurance tests may be done on the software so that some numbers are there and some aggregate measures look reasonable. But this is no guarantee that the data are clean or even correct. Some subtle data issues may still be present. The more we analyze the data, from exploratory data analysis all the way to predictive modeling, the better we know the data and the better we identify issues. *Data are only as clean as the amount of effort used to analyze them.* This is similar to debugging a software product, which we all know is a long, laborious process. If we have not completely analyzed the data, they may not be correct. Without continued detailed analysis, additional issues may be introduced by new releases, and new usage exceptions may not be handled properly by an existing release.

Traditional Business Intelligence and Big Data

The traditional business intelligence (BI) is shaped like a pyramid (Dyche, 2007): from the standard report at the bottom to the multidimensional report, the segmentation/predictive modeling, and finally to knowledge discovery, which is at the top of the pyramid. Going from collecting a standard report to knowledge discovery, data maturity of the organization increases and there are fewer assumptions needed. This is similar to the capability maturity model in software development (Paulk, Curtis, et al. 1993).

The BI pyramid defines a sequence of efforts from simple to increasingly complex, as in crawl, walk, and run. Most organizations are somewhere in the middle in "maturity" level; they never go beyond the stage of multidimensional reporting or simple analysis. These companies may just have built a data collection infrastructure, or may not have the required analytic talents, or may not be ready due to organizational and cultural reasons to achieve a higher level on the pyramid. They never had a detailed analysis of the data; no predictive modeling was ever done. Again and again in our years of experience, we found data issues that are subtle enough to look normal without a detailed analysis. For example, a data warehouse may take many data feeds from different departments or regions, and only one of them has problems. The numbers are not missing, but they are not accurate or not correct.

If a company adopts a stepwise approach according to the traditional BI pyramid, business rules used to produce standard reporting will need to be decided beforehand. Before big data technology is available, because of the high cost of storage and computing power to process, most data are not collected or discarded. Only data deemed to be the most important are kept. Since analytic tools are built on databases, there is usually no easy way to analyze data in raw format. Therefore, assumptions have to be made about the data before we can look at them. We have to make decisions on data structure before loading raw data into a database. This can be a source of problems. Once the designs are implemented, they are difficult to change. Without the benefit of a thorough analysis, an initial design may hinder the optimal extraction of information and knowledge. This may not be optimal.

In big data, data volume is so large so that raw data are stored as the persistent data, on a cluster of distributed computers with local storage. Also because of the size of the data, data access and analysis will need to be done on the same cluster of computers. A characteristic of big data analytics tools is that we can process data in raw format in a distributed way by using a large number of servers to manipulate data on their local storage. With big data analytic tools, we can and should do a more thorough analysis before generating standard reports. After analysis, the data are more reliable and we know better the basic patterns in the data, so we can better identify which variables are important and should be put in reports.

Some big data can be in a free format. Then relevant information has to be extracted before analysis can be conducted. Depending on the nature of raw data, there is usually no unique way or surely successful methodology to extract information from such data. Various strategies have their own perspectives and may yield different amounts of information with different levels of utility.

Therefore, *we need to conduct a detailed analysis before building standard reports*. This approach does imply that people who know how to analyze the data should be a part of the decision-making process on the data structure. We often say that knowledge is power. With big data, now we need to add that knowing how to discover knowledge is power.

Models Have to Be Designed by People

It is up to the analyst working with the stakeholder to define the question to be answered, to decide the model to be built, to select the dependent variable, which is the one we try to predict, and to choose all the independent predictors as well. For example, to improve services to our customers, we have to first decide how we measure quality of services. Our metric can be the number of clicks or conversions, transaction dollar amount, lifetime value, time spent, or visit frequency, and so forth. These measures are related but not identical, each with pros and cons and different emphasis or perspective. After we decide to choose, say, conversion, the metric is called the *target*, in the modeler's language. We then gather a set of variables to predict conversions, for example, day of week, time of day, geo, age, or gender, and these are the predictors. Again, we have to decide whether to include a particular variable. Computer algorithms may determine that a variable we include in the modeling is not predictive, but they cannot tell if a critical predictor is missing. It is up to the analyst to make these decisions.

Computers have no way of knowing whether there is a problem in a model. This can be very subtle. For example, during the model-building process, if a predictor data contains information about the event it is supposed to predict, the model produced will appear to be more accurate than it really is. In such a case, when we apply the model, its performance will be poor. This is called a leakage in predictive modeling. Only analysts know if these mistakes are present. Inexperienced analysts may solve correctly the wrong problems, and even experienced analysts may have a lapse of judgment.

Finally computers have no goals to achieve. It is not computers but people who decide on the purpose of the analysis and how knowledge will be used to take action. Computer models have to be designed and managed by people. Even after having built and deployed automated solutions to achieve scalability, we still need some analysts to assess and ensure their quality of performance, and to find new ways to improve and optimize.

Perfect data are all alike; every wrong data is wrong in its own way. In addition to some relevant data not being collected, it is also possible that some data feeds, but not all feeds to the warehouse, are incomplete. So when we query the table, data are there, but some rows or some values are missing. Without detailed knowledge, it may not be easy to realize that there is a problem. There can be multiple definitions of the same field, and each of them may be used for some period of time. There can be multiple business rules based on reasonable but different assumptions. For example, at an online university, if a new student took a single course and paid for it but dropped out after the first couple of classes, is he considered a student? One analyst may say that the person paid tuition and was a student for the classes. Another may say that he is just someone who took a single class and could hardly be considered a student. Both are reasonable, but they would result in not only different student enrollment counts but also metrics like average revenue per student.

Some data are incomplete due to business nature. For example, we have data that a customer has interest in some products, but we have no data on her interest in other products. The data are sparse, so it is difficult to tell whether there is a lack of behavior or it is an incomplete collection of data. One example is the separate log-in and log-out data for Internet portals. Due to privacy policies, the two sets of data cannot be analyzed together. Since people do not always log in, either data set is incomplete. Credit card purchase data reflect only a customer's partial behavior because of possible cash purchases. Data are never ideal. It is up to the analyst to decide if models should be built and if they are useful. This underscores the insight that detailed data issues need a thorough analysis to uncover.

Modeling Needs to Scale as Well

In traditional practice, predictive models take a long time to build. For example, it may take several months or even more than a year to build a model in property insurance. The training data sets for model building are quite small, and sample data are often relatively expensive to collect. Models can be built only for repeatable patterns over a long period of time.

Nowadays in the time of big data, data are cheap and abundant. We build more and more models; some of them may degrade in performance in weeks. With big data, the number of predictors or dimension of predictors can be very large. In addition, some variables may be categorical with a large number of values. In this new situation, human interactive model building is not scalable. We no longer have enough resources to build all the models with a lot of human interaction.

Reasons for interested events can be complex. Without some detailed analysis, it is often unclear which of a large number of variables drive the event. In traditional modeling, the number of predictors often is not more than a few dozen. Now, it is not uncommon to have thousands of variables. Increasingly, we need to rely on modeling methodologies which help build models somewhat automatically, using techniques like out-of-sample testing and off-the-shelf modeling.

Bigger Data and Better Models

Any model has two parts, the data and the analytic framework. For many complex questions, the ultimate determining factor to improve the quality of models is data. Not only will better data lead to higher-quality models, a larger data set will also generate more accurate results. Statistical analysis of really large data sets can often help us better answer difficult questions. One such example is "wisdom of the crowd," which says that for many questions aggregating responses from a large number of people will give better answers than asking an expert.

Thus, if we want to know the price of an item, we should look it up in eBay auctions; if we are looking for the value of a keyword on Google paid search, we should place bids on the auction engine to find out; if we wonder how good a book is, let's look at its reviews on Amazon.com; if we want to compare which of the two web page layouts has better conversion rates, let's do an A/B test for a large number of site visitors to decide, and so forth.

Other examples are Google's spell checking in search and the Translation product, which are based on big data–driven models. Research shows that model results continue to improve as the amount of data becomes larger and larger (Norvig, 2011).

Big Data and Hadoop

There are some characteristics in big data analytics. In big data, often raw data are stored and appended but not updated. There are no aggregations for the purpose of saving storage. This is mainly because the volume of raw data is too large for normal database technologies to handle. When data sizes are larger than several hundred gigabytes, a single server will not be able to process the data in a reasonable amount of time. For example, it may take a day for a server just to scan one terabyte of data from a storage disk.

To get results in a reasonable amount of time at a reasonable cost, a technique now often used is MapReduce, a distributed computing paradigm developed at Google (Dean and Ghemawat, 2004). The basic idea is the following: We use a cluster of commodity servers with local storage to work as a single computer. We read and process intermediate results in parallel using many servers on local data, which is called the Map step. And then we aggregate at the end, which is the Reduce step. We may need to repeatedly execute Map and Reduce steps to complete a task. In order to address the issue of slow speed of disk read and write, we bring computing closer to the data. A cloud of servers using MapReduce often scales linearly as the number of servers increases, but not always. As data get larger and larger, a cloud of commodity servers is the only way to scale.

MapReduce is a data processing strategy that can be implemented on different platforms. Google has its own implementation. Ask.com built an

SAS cloud using the MapReduce paradigm for an online educational institution, which was discussed in an invited talk at SAS Global Forum (Zhao, 2009). The setup can process billions of ad impressions and clicks at the individual customer level in a scalable way. One advantage of using SAS to implement MapReduce is the availability of a large portfolio of statistics procedures already in SAS to process and analyze data. This is an especially good solution for organizations with SAS site licenses. Hadoop is an open-source implementation of MapReduce used widely on commodity servers and storage. Many major companies, such as Yahoo!, Facebook, and Ask.com, have large Hadoop clouds consisting of thousands of servers. Using these clouds, we can search the data to find a needle in a haystack in milliseconds; model computations usually would take years to compute, but now can be completed in minutes. Using cloud computing, we can build models in scale. In 2010, Google was using 260 million watts of electricity, enough to power 200,000 homes (Glanz, 2011). This implies that the total number of servers is on the order of several hundred thousand or more. At one location near the Columbia River at The Dalles, Oregon, where electricity is less expensive, Google has two football-field-sized data centers. Facebook, Yahoo!, and other Internet companies have similar large data centers.

ONLINE MARKETING CASE STUDIES

Wine.com One-to-One e-Mails

During the dot-com era, Digital Impact was an e-mail marketing company committed to the vision of "the right message to the right customer at the right time." It was one of the main intermediary players between customers and vendors. Now e-mail marketing is still a widely used and effective channel to engage customers.

In 1999, I led the analytics project to help the e-commerce site wine.com develop a one-to-one e-mail program. Armed with wine.com's house opt-in e-mail list, and permissions to send marketing e-mails, wine.com sent weekly newsletters, with each customer having a different set of six or eight recommended wines. Before using the one-to-one e-mail solution, weekly e-mails contained static wine offers, with every customer getting the same recommendation, selected by wine.com's merchandiser, along with some news articles on wine and related information. Wine.com had an inventory of more than 20,000 wines. Due to state-level alcohol regulations, there are distribution constraints for various states.

As one of the early pure e-commerce sites, wine.com had relatively clean data. We were able to get purchase and product data, as well as e-mail behavioral data. For each purchase, we obtained time of purchase, products, spend, and associated campaign. Wine product profiles were also quite complete, with product-level data on price, color, variety, vintage, country of production, the producer, and a description of the wine. Wine.com also gave us a set of taste profiles of the wine, including oak, sweetness, acidity, body, complexity, intensity, and tannin in a scale of 1 to 7. We also had e-mail response click streams linked to each wine, and we collected self-reported preferences and demographic data, such as age, gender, zip code, and others, as well as preference for types of wines, and optionally, drinking frequency, purpose of purchase, level of knowledge about wine, and so forth. There were no explicit customer ratings of products. Most customers had only one or two data points, while a small percentage of customers had a lot of purchases and e-mail clicks.

The goal of the one-to-one e-mail program was to lift purchase revenue. We achieved this by optimizing the selection of a subset of wines that a customer is more likely to buy. The efficacy of the program was measured by A/B testing against weekly static selections by merchandisers. Our challenge was to produce consistent lift over a long period of time and many e-mail campaigns.

We designed an algorithm called *preference matching*. Instead of building elaborate logistic regression or decision tree models to predict interest category, we put our focus on the most important predictor—customer behavior profile—which was built using the detailed wine product profiles. We built both implicit profiles from purchases and e-mail clicks data and explicit preference profiles. More active customers had more behavioral data points, so that they have more refined profiles. We also considered the overall popularity and seasonality factors included, for example, champagne wines are more popular near the new year.

We then decomposed purchases into values in product attributes. Even if a customer had only a single click, we still could generate a profile. We augmented the profiles by adding association rules such as "Customer who bought these also bought …" An advantage of such an approach is that when the specific wine goes out of stock, its profile information is still very much usable. New releases have no purchase history, but as long as we know the product attributes, they can be immediately mapped to existing profiles. For new customers, we augmented their profile with nearest neighbors who had more purchases as "mentors."

The algorithm used cosine distances to measure similarity in taste profile by color, and we also used price range, as well as text attributes on producer name, region, and country of production, to recommend similar wines. In successive campaigns, we shuffled among higher-scored wines. This way, repeated campaigns took care of prediction errors. We also deduped recent recommendations and purchases so that we didn't repeat what customers obviously were familiar with. We used decaying memory functions to put more weight on recent profiles and factored in seasonality. We always use simulations to ensure recommendation quality and user experience. Through reinforced learning, which is repeated test and optimization, we find algorithms and weights that give the highest lifts.

The one-to-one e-mails using these algorithms increased revenue up to 300 percent relative to the control cell. The program performed by 40 percent over more than a two-year period. We found that lifts in revenue were more significant than those in click-through rates. This finding underscores the importance of selecting the right metric of customer service. We found that purchasing data were the most important in recommending wines that customers are more likely to buy again. E-mail response data were also predictive. This says that the customer puts money where his mouth is. Self-reported preferences tend to be broader in range than the purchased sets. It is "talk the talk" versus "walk the walk." Aggregated web behavioral segments were least useful, and it is likely that this had to do with the way in which the early dot-com web analytics vendor processed and aggregated the data.

We built similar programs for other vendors, for example, Intel Channel Marketing to recommend, in biweekly newsletters, time-sensitive news on product releases, price drops, white papers, marketing collaterals, and training, based on purchases and e-mail response behavior, achieving the goal of sending the right information to the right customer at the right time. The general strategy of these programs is to improve relevance, to help customers search information, and to engage the customers.

Yahoo! Network Segmentation Analysis

In 2003, Yahoo! was the web portal on the Internet with 200 million users. Yahoo had more than 100 properties or websites, such as Mail, Search, Messenger, Personals, Sports, News, Finance, Music (Launch), Shopping, Health, and others, with many properties being ranked as top sites at the time in their respective categories. Yahoo!'s privacy policies forbade explicit user-level analysis using combined login data and logout data. So we did the analysis using only login data. We separately did a sample analysis on combined login and logout data, which was encrypted to comply with privacy policies, and found similar results.

We asked, Who are Yahoo!'s users and how do they use Yahoo!'s properties? The intention was to use monthly page views in different properties to build a monthly profile for each user, and use clustering algorithms to group users into a finite number of segments. Each user belongs to one and only one segment. The benefit of this approach is that we can target individual customers based on the segments.

Potentially every customer can be different, which would result in 200 million segments. For 100 properties, if we use 1 for users and 0 for nonusers, we would get 2^{100} possibilities, which is an astronomical number. In reality, people's behaviors had a limited number of usage patterns. We expected the number of segments to be a much smaller number, say only around 100.

Each property has its own typical usage levels. For example, Mail had an average of several hundreds of page views per user per month, while News had an average of a few dozen page views, and Shopping may only have a few page views. Some of the differences were due to the various stages of adoption of the products and others to just the nature of the product. We would expect that a user generates fewer Shopping page views than e-mail page views or Sports ones. We did some normalization so that even though Mail was the most heavily used property, there weren't too many people in the Mail segments. Shopping page views are low, but user values are high. We don't want to see Shopping page views getting swamped by those from Mail or Sports.

After some optimization on the cluster analysis, we got 100 segments. Not surprisingly, Mail was still the largest segment, with a third of all users. The Search segment was the second largest. Shopping was around a few percent. Eighty-five percent of customers were in the top 15 segments.

After the clustering, we did some analysis profiling the segments. Since we used only login data, we were able to append gender, age, and other information. We found that some properties were gender neutral, such as Mail and Search, but interestingly some segments were highly selective for gender and age groups. For example, News and Finance were used mainly by male and older users, Music by young females, and Sports predominantly by young males and healthy older females. Not surprisingly, Search users had high user values, while Music and Sports had very low user values.

One of the obvious strategies to increase customer value is to integrate the more engaging properties such as Mail or Sports, with better monetized ones, such as Search and Shopping. Implementing features of Mail Search together with Web Search is an obvious integration tactic, so that we can have more Mail customers use Search more often.

Yahoo! e-Mail Retention

Mail was the stickiest service of Yahoo! If customers become Y!Mail users, the likelihood of their coming back is much higher. Users of other web properties, such as Search and Shopping, are more fickle. Therefore, increasing Mail customers is good for Yahoo!'s overall retention.

At the time, 40 percent of new Y!Mail users never came back after their initial signup. An analysis indicated that for customers who had e-mail activity immediately after signup, the retention rate would become normal. A more detailed analysis found that frequent page views in certain sections, such as Help and Junk folders in Mail, were predictive for mail retention. We tried to find actionable retention drivers and strategies, such as sending welcome e-mails, to improve customer service, user experience, to reduce Mail churn, and so forth.

There are many ways to analyze the retention problem. One approach is to look at profiles and activities of a cohort of Mail users in one quarter and see if they come back the next quarter. Some analysts are more comfortable with this formulation due to its simplicity. One of the problems of this approach is that retention depends strongly on tenure. For newborns, when we plot infant mortality rates versus time, we find that the rates were high immediately after birth but they decrease and stabilize after a couple of weeks. Similarly, new e-mail customers tend to have high attrition rates initially, and the rates stabilize after some period of time. If we choose a time interval that is too large, we would lose information about this feature.

A more appropriate method to analyze customer retention is the survival analysis, a statistical method for analysis of patient survival data under medical treatments. If some treatment yields a higher survival rate than the placebo, it is said to have a certain efficacy. In consumer behavioral analysis, customer "survival" means customer retention as indicated by continued visits.

Customer Lead Scoring

In 2009, an online university was one of the largest online marketers, and it worked with a large number of lead-generation vendors. A lead is a customer name, contact information, and some basic profile of the area of study, high school degree, possible association with the military, and other fields, as well as the permission to contact.

For those of us familiar with online marking, customer life cycle is usually from an impression to a conversion. But for a lead, the experience from an impression to signup is just one third of the life cycle. After the university receives the lead, its call center and enrollment counselors will discuss with the candidate the topic of enrolling at the university. After months of effort, only a few percent of leads will enroll as students. Students can stop taking classes anytime, and those who are easy to enroll in the university may also be quick to drop out, with only a small percentage of them ever graduating many years later.

Lead vendors have their own media strategies, reaching various segments within the population to collect candidates with different levels of interest in college education. Being at different locations in the conversion funnel, some leads are ready to enroll immediately while others may be just looking around. Therefore, leads from the vendors often have very different enrollment rates. Because of the long enrollment process, it may take many months before we know the quality of a cohort of leads from a vendor. The university paid the vendors every month and had to agree with each vendor on price per lead and volume without the benefit of any direct information about the leads.

To assess the quality of leads, we need student data over a long period of time including not only enrollment information but also class completions. Ideally we should use lifetime values and brand values tied to the leads to determine media allocation and to buy a number of the best leads at the lowest cost while enhancing the brand.

One way to estimate quality is lead scoring. Analogous to credit scoring, the model uses given information at time of lead submission to score leads on the propensity for enrollment. This is similar to a car dealer running a credit report before deciding if we qualify for financing when buying a car. Using this approach, we can also build models on, say, completion of first one or three courses.

A lead may have been marketed multiple times from various channels. To build a good lead-scoring model, we need to track lead-level data in search, display, landing page, home site, call center, enrollment, courses completion data, and other factors. Ideally, we need to have a 360-degree view of a lead's signup and conversion process, as well as student life cycle. Lead quality may also depend on major, credits finished, demo, socioeconomic status, first-generation students, lead source, lead form entries, and so forth. Some degrees and majors have different desirable student profiles and may require different scoring models.

Vendors also have different levels of fluctuation in enrollment rates from month to month. When we buy leads, we take risk in the value of leads relative to the cost of leads, just as when we buy stocks we take risk in the company's prospects. Using financial theory of efficient frontiers, we can calculate a larger price discount if the vendor has a higher variability in enrollment rates, and we can construct a portfolio of lead vendors with a lower risk than that of an individual vendor.

Customer Lifetime Value

Let's consider the case at online universities, although similar arguments can be made for customers of other vendors. Online universities often face the question of student retention, sometimes called *persistence*. If a student drops out, it is a loss to both the student, who has to pay tuition, and the university, which has to spend resources on recruiting and educational services. What are the overall costs and returns of a student during time at the university?

Student attrition is not just absence for a period of time. A student who takes off for a period of time before assuming study is still retained. Some assumptions have to be made about the point in time of a student's attrition, for example, by defining a churn as someone who has taken a break longer than a certain period of time. We then analyze events up to that time and find their correlation with risk factors, such as if the student had a baby, failed some courses, had a family member who became sick, etc., to estimate the probability of attrition. By definition, the retention curve is nonincreasing in time, while cumulative attrition is nondecreasing. Starting at 100 percent initially, a retention curve eventually goes to zero. This is because in time, a cohort of students will decrease in number as more and more students either drop out or graduate.

With retention curves, we can consider lifetime revenue generated by a student. Like financial assets, we pay acquisition and service costs and receive revenue when the student takes a sequence of classes, considering the duration of the degree program. Since a student may or may not take the next course, the lifetime value is the average of revenue minus cost, weighted by the probability of retention.

To calculate lifetime values, we assume that student acquisition costs, marketing costs, and enrollment costs are shared by all new students, but not by returning students. Course instructional costs and salaries of faculty and academic counselors are proportional to the number of courses the student has taken. Campus and online students have different service costs, fixed or variable.

Longer programs have higher student lifetime values. In traditional four-year universities, student attrition rates may be very low. In community colleges and online universities, attrition rates are quite high initially and then stabilize after a few courses. This is because these universities serve primarily adult and part-time students, who have more retention risk factors. Many students receive credit for their past college courses or work experience. Because of the varying number of transfer credits, each student needs to take a different number of courses to reach graduation. This also affects the lifetime value in a degree program.

We built retention curves by degree and program and other variables and calculate lifetime values for each segment. Retention rates may depend on some other variables, such as age and gender, lead source, geographic location, modality and socioeconomic factors, and others.

We can attribute expected value of a student to a lead source, a search keyword, or a display ad impression, and we then can use the information to optimize media spend.

Ad Performance Optimization

Tribal Fusion (part of Exponential Interactive) was one of the pioneers of the display ad network. Aggregating a large number of reasonably large high-quality web publishers, Tribal Fusion serves display ads for premium advertisers, using a revenue-sharing model. By 2005 it became one of the top three display ad markets, reaching around 70 percent of the U.S. unique users, with billions of impressions per month. One of the efforts at Tribal Fusion was ad performance optimization. We used information about the publishers, channels, customer geo information, past behavior, demographic data, data append, session depth, and other factors to score each impression.

Because various advertisers had different conversion patterns, we used an array of predictive models, one for each advertiser, on conversion rate (or click-through rate) to work together with the auction engine in the ad server. We modeled using individual event-level information to predict a conversion rate for each impression.

We wanted to build a separate model for each of the hundreds of advertisers, but too many models were needed and there was too little time for them to be built by humans. Instead, the models were generated using an automated script that ran overnight.

Revenue Prediction

One of the tasks we were given for an online university client was to predict enrollment and revenue in the future within errors of a couple of percent, for the next month and in three months.

We were given all student transactional-level data for the online university from the finance department for three years as well as all data from the data warehouse, which had all the lead and student enrollment data and others. So in principle, we knew all the enrollment and all the associated revenue. Predicting future enrollment and revenue should be quite possible.

In reality the situation was far more complicated. The main problem was that there was more than one definition of revenue recognition and enrollment numbers by modality, campus or online, made by past business analysts, using reasonable business rules. Some rules were built into the BI reporting product, which the Financial Planning and Analysis team watched every month as only truth they know. We underestimated the difficulty of finding out explicitly the rules. It turned out that with IT/BI turnovers and rules changing over time, few people knew or knew how to articulate the rules. Without the rules, the enrollment and revenue numbers we calculated from the data were off by random errors of around 7 percent, larger than the prediction accuracy we wanted to achieve. After several meetings, we still had no correct rules that could reproduce the numbers from the reporting product. We also saw one-time data anomalies here and there. For some data problems, the finance team provided corrections, but for others, information was limited or absent.

Within the short time constraint, we found a way to get around these limitations. We modeled time series of reported data. This assumes that the relation between enrollment and revenue for campus and online modalities would be stable over time. In this way, one-time data errors were diluted, and rule changes long ago were also less weighted.

In the end, we were able to predict customer and revenue numbers for three to six months within a couple of percent. Time series models do have the assumption that some level, trend, and periodicity continue over the time window of prediction. Without the link between student-level information and revenue, it would be more difficult to use this approach to calculate the impact of student demographics and lead source information.

As we later found out, one of the issues was that some revenue from online enrollments was credited to campus, as an incentive to increase the use of online classes. These were campus students who also took some classes online.

Search Engine Marketing at Ask.com

Ask.com (formerly Ask Jeeves) was founded 16 years ago, and now it is part of InterActiveCorp, the IAC. Ask.com attracts 100 million global users and is one of the largest questions and answers (Q&A) sites on the web. Over the last two years, Ask.com has revamped its approach to Q&A with a product that combines search technology with answers from real people. Instead of 10 blue links, Ask.com delivers real answers to people's questions—both from already published data sources and from our growing community of users—on the web and across mobile.

Similar to other websites with original content, Ask.com uses multiple strategies of customer acquisition, with search engine marketing (SEM) being one of them. Using SEM, Ask.com places ads on major search engines to acquire customer traffic using the pay per click model.

One of the efforts is to identify keywords where Ask.com has an advantage. This is achieved by determining bids for each keyword using external data from the search engines, as well as internal data sources. If there were only a small number of keywords, it would be easy to let one or more analysts manage them; but Ask.com's keyword portfolio is very large, covering a wide range of topics and categories. To set bids for an extremely large number of keywords, data mining applications are developed. These applications run every day with new bids being automatically generated and pushed to major search engines. Through the use of reinforced learning, the algorithms are used to determine and optimize bids based on past performance data and to make further adjustment using new data.

We also propose and test hypotheses and optimize algorithms and their parameters via A/B tests.

In the bidding algorithms, we build models for revenue estimation at keyword and keyword group (cluster) level. This information, along with other information and business logic, is used to generate bids. Some of the variables we use are ad depth, which is the number of ads on the landing page; search engine click-through rates (CTRs); landing page click-through rates; quality score and minimum cost per click (CPC); effective CPC; keyword categories; natural language clusters; and search behavioral clusters.

One of the main assumptions is that similar keywords have similar performance, which tends to be the case, but not always. We found that contextual similarity to be more useful than similarity in performance metrics.

To group similar keywords together, we performed keyword clustering using text mining algorithms. We also clustered the keywords using behavioral associations, as well as metrics of keyword historic performance. We mapped out similarity metrics among keywords so that we can use information from similar keywords to help keyword management and expansion, and to leverage learning from keywords with more data.

One of the biggest challenges is to select profitable keywords at big data scale. Hadoop and Hive as well as machine learning suite Mahout are used to process and analyze the data, predicting keyword performance and bidding for the right keywords at the right price at the right time.

Although improving return on investment is important, our goal is to maximize profitable traffic volume. The algorithms generally increase click traffic for keywords of higher quality scores and higher click-through rates and reduce it for keywords of lower quality scores and lower click-through rates. We also optimize user experience through adjustment of the number of ads shown as well as the layout of the search result pages, not only to achieve profit goals but also to improve customer experience.

LESSONS FOR MODEL BUILDING

In predictive modeling, often there is leakage, which is the unintended mixing of information about the target in its predictors. For example, in building a lead scoring model, *lead source* was used to predict conversion. But some values of the field were populated only for converters that came from a different data source than nonconverters came from. Then the lead source becomes more predictive than it really is, contaminating the model. When being deployed, the model will have a lower predictive power.

Another example for display ads is the conversion model. We may construct the data set by taking all converters and a random sample of nonconverters. We then predict conversion using user page view profiles. The problem, if we are not careful, is that in the sample of nonconverters there are customers who had no impressions of the display ad. Of course, one gets the trivial and useless prediction that those who never see the ads are less likely to convert. These errors can be subtle and can be overlooked even by expert modelers.

We worked with SBC Communications (now AT&T) to market digital subscriber line (DSL) services to consumers. DSL subscribers have one-year contracts. In a retention analysis, if churn events are measured for all customers in a month-to-month retention, we would find very high retention rates. This is because of the contracts with penalties if customers leave early. The analyst could declare that nothing needs to be done, but this approach would have omitted the renewal at the end of the contract. A better way is to model retention rates at the contract expiry, on only one-twelfth of the customers.

For a retention analysis, if we define retention rate as the fraction of customers who are acquired in one quarter and retained in the following quarter, we will find that those acquired early in the first quarter have a lower retention rate. This is because those customers have more time to churn. A correct way is to use survival analysis.

CONCLUSIONS

Big data analytics provide the most exciting opportunity in every field from science, government, and industry, affecting daily lives of everyone. Big data is a dream come true for data scientists, since we finally can have it all, to get exciting insights we could never have before.

Big data does not become big information and big knowledge without detailed analyses. Big data requires big and scalable storage solutions, as well as scalable analysis capabilities and applications. Analysis does not mean we can throw data at some machine-learning and statistics algorithms, such as neural networks, decision trees, support vector machines, and so forth and expect to have good results automatically.

The analyst should focus on the domain knowledge. Good modeling requires not only algorithms and procedures but also, more importantly, understanding of the business context, insights about the data, and how one may take actions based on results of the analysis. In modeling, it is most important to identify the key data. The analyst needs to understand how data are collected and know the context of data collection, as well as what data can and cannot be collected, and be able to balance the cost of collecting additional data and optimization of modeling. Identifying the smoking gun may make all the difference. Understanding of the business context and the data helps the modeler identify good data transformations. Using the link data in web pages, Google's search algorithm PageRank (Brin and Page, 1998) was a game-changing data transformation. In our wine.com case study, the wine similarity metric was also a key data transformation. Social graph is a key data transformation for fraud detection (Hardy, 2012). Using big data, it is especially important to identify the most import predictors and to come up with creative and useful ways to transform the data. Data are not reliable until after being seriously analyzed. Only detailed analysis can reveal subtle data issues. We have to do our due diligence on the data before we can be sure of their cleanliness and accuracy, as well as relevance.

Using the feedback loop to test hypotheses is a very effective way to gain better understanding of data insights as well as optimize models. To the extent possible, we should conduct simulations to see if changes are reasonable. Testing and optimizing in the real market can be crucial. We should always focus on customer experience, not model complexity or predictive accuracy.

Bigger data will support better models. The analyst's knowledge in natural sciences can be helpful in finding insights and building models in a given data set. Scientists are better at connecting the dots. We know Einstein's relativity was based on little data other than his "thought experiment," and now big data from space telescopes are providing support to his theory. Darwin wrote in *On the Origin of Species*, "Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed." His conclusion was based on his limited data from the Galapagos Islands. Now 150 years later, scientists use genomic big data to confirm the existence of a common universal ancestor (Steel and Penny, 2010).

In the case studies, we sampled some applications of customer segmentation, lead conversion, retention, lifetime values, targeted e-mails, predictions of trends, and seasonality of revenue, as well as keyword segmentation based on text and search behavior, based on our experience. One of the key features of these models and analyses is that they are built on individual customer and event level. The only way to scale these types of efforts, in the amount of data and in the number of customers, is through the use of big data.

To conclude, we use good advice from one of the greatest scientists ever:

> The best way to get good ideas is to have a lot of them.
>
> **—Linus Pauling**

REFERENCES

Brin, S., and L. Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30: 107–117.

Cukier, K. (2010). "Data, Data Everywhere." *The Economist*. Feb 25, 2010. Retrieved from http://www.economist.com.

Dean, J., and S. Ghemawat (2004). "MapReduce: Simplified Data Processing on Large Clusters." OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December.

Dumbill, E. (2012). "What Is Big Data? An Introduction to the Big Data Landscape." January 11, 2012. Retrieved from http://radar.oreilly.com.

Dyche, J. (2007). "BI Adoption Evolves." Retrieved from http://www.baseline-consulting.com.

Glanz, J. (2011). "Google Details, and Defends, Its Use of Electricity." *New York Times*. September 8, 2011. Retrieved from http://www.nytimes.com.

Hardy, Q. (2012). "Data Analytics Company Finds Fraud Is a Friend." *New York Times*. January 19, 2012. Retrieved from http://www.nytimes.com.

Hilbert, M., and P. López (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332: 60–65.

Norvig, P. (2011). "The Unreasonable Effectiveness of Data." October 11, 2011. Retrieved from http://www.youtube.com.

Paulk, M.C., B. Curtis, et al. (1993). "Capability Maturity Model for Software, Version 1.1." Retrieved from http://www.sei.cmu.edu.

Siegler, M.G. (2010). "Eric Schmidt: Every 2 Days We Create as Much Information as We Did Up to 2003." August 4, 2010. Retrieved from http://techcrunch.com. For more discussions on the topic, see Hilbert and López (2011).

Steel, M., and D. Penny (2010). "Origins of Life: Common Ancestry Put to the Test." *Nature* 465: 168–169.

Zhao, D. (2009). "The University of Phoenix Wins Big with SAS Grid Computing." SAS Global Forum 2009, Washington, DC. Retrieved from http://support.sas.com.