siness Series

that help senior-level managers with

clude:

tutions: Driving Bottom-Line Results by

h Social Media and Mobility by Bernie

ert Laursen

s Intelligence beyond Reporting by Gert

m Approach to Maximizing Competitive
gam, and Stefanie Gerlach

ning Your Business in the Global Economy

de from the Experts by Tony C. Adkins

Information Technology, Second Edition

for Borrowers, Lenders, and Investors by

ting Intelligent Credit Scoring by Naeem

ion of the Truth, by Jill Dyche and Evan

ch to Forecasting by Charles Chase

chieving Strategic Objectives by Gregory

am, Jason Wahl, and Stuart Rose

cations for Credit Risk Management by

to Yen to Yuan: A Guide to Fundamental
wan

olution Model to Grow Your Business by

vity and Product Quality by Bobby Hull

ective Direct Marketing by Jeff LeSueur

to Make Knowledge Sharing Work by

ieces (to Close the Intelligence Gap) by

Execution, Methodologies, Risk, and

Cox

Carlos Andre Reis Pinheiro

ices and Providing Practical Solutions by

heir Data for Business Success by Tony

a Strategy: How Social Networks Are
homas and Mike Barlow

by Thornton May

th to Profitability by Evan Stubbs

an Cox, Marie A Gaudard, Philip J.

les, please visit **www.wiley.com.**

Taming the Big Data Tidal Wave

Finding Opportunities in Huge Data Streams with Advanced Analytics

Bill Franks

John Wiley & Sons, Inc.

CHAPTER 3

A Cross-Section of Big Data Sources and the Value They Hold

Wouldn't it be neat to receive an offer on your mobile phone for a discounted meal at a restaurant as you are driving past its parking lot? How happy would you be if a pit boss at a casino gave you the $20 that a dealer forgot to pay you? Imagine being able to quickly find people who match to your playing style in an online video game because the game can tell you who they are. Would you like to lower your car insurance rates? All of these are possible through big data.

In Chapter 2 we discussed web data. If web data isn't the original big data, it is probably the most widely used and recognized source of big data. But there are many other sources of big data as well, and they all have their own valuable uses. Some are fairly well-known and some are relatively obscure. In this chapter, we're going to take a look at nine more sources of big data and some ways to use them. Each will be covered at a high level. The purpose is to provide an introduction to what each data source is about and then review some of the applications and implications that each data source has for businesses.

Chapters 2 and 3 are not a top-10 list, because the claim is not that these are the most important sources of big data. The order in

which the sources are discussed doesn't imply any ranking either. The point is to provide a representative cross-section of big data sources so that the reader will understand the breadth and types of big data available, as well as the breadth of analysis that the data enables. Every reader should see at least a few to take a personal interest in.

One trend that becomes clear is how the same underlying technologies can lead to multiple big data sources in different industries. Also, different industries can leverage some of the same sources of big data. Big data is truly not a one-trick pony with narrow application. Its impacts will be far-reaching.

The big data sources we'll cover include:

- Auto insurance: The value of telematics data.
- Multiple industries: The value of text data.
- Multiple industries: The value of time and location data.
- Retail and manufacturing: The value of radio frequency identification (RFID) data.
- Utilities: The value of smart-grid data.
- Gaming: The value of casino chip tracking data.
- Industrial engines and equipment: The value of sensor data.
- Video games: The value of telemetry data.
- Telecommunications and other industries: The value of social network data.

AUTO INSURANCE: THE VALUE OF TELEMATICS DATA

Telematics has started to receive serious attention in the auto insurance industry. Telematics involves putting a sensor, or black box, into a car to capture information about what's happening with the car. This black box can measure any number of things depending on how it is configured. It can monitor speed, mileage driven, or if there has been any heavy braking. Telematics data helps insurance companies better understand customer risk levels and set insurance rates. If privacy concerns are ignored and it is taken to the extreme, a telematics device could keep track of everywhere a car went, when it was there, how fast it was going, and what features of the car were in use.

Telematics has the potential to lower rates for most drivers and increase profits for insurers. How can it both lower rates and increase profits? The answer is that insurers have to price insurance based on a risk estimate. Using traditional risk estimates, based on data such as age and demographics along with personal accident history, provides only a high-level profile. Especially for drivers with a clean driving record, there isn't much to differentiate them from the other people in a neighborhood.

Insurance companies have to plan for the worst. So they'll figure out what band they think people fall into on the scale of risk and then to be safe will assume their risk is at the higher end of that band. The more detail that an auto insurance company can get on people's specific habits and how risky they actually are, the narrower the risk bands will be and the less assuming the worst case within a band will increase their rate. That's how rates can lower and margins can go up at the same time. Insurers will have a much better feel for individual risk and they'll have less variability in the projected payouts that will have to be paid.

There are insurance companies pursuing telematics-based insurance in many countries throughout the globe, and the number is growing. Early programs focus on collecting minimal information from cars. They don't track everywhere a car has been, for example. What the early programs do track is how far a car is driven, what time of day it is driven, if speeding occurs, and if there is a lot of heavy braking. It is fairly basic information with limited privacy concerns, which is intentional. By avoiding the collection of highly sensitive information, a wider level of adoption will occur. The same principles are also being applied to commercial fleets. It is easier to set rates for a company's fleet of trucks if the insurer knows more specifically how the trucks are utilized.

Telematics data is taking hold initially as a tool to help consumers and companies get better, more effective auto and fleet insurance. Over time, telematics devices may end up being present in a large number of vehicles and uses for telematics data outside of insurance will emerge. There are already onboard computers managing systems within an automobile, but a telematics device can take it to an entirely new level. There are some very interesting uses for telematics data. Let's take a look at a few.

Using Telematics Data

There are some mind-blowing analytics possible if telematics truly takes off. Just imagine that a critical mass of millions or tens of millions of cars end up with telematics devices within your country. Let's also imagine that a third-party research firm arranges with consumers to collect very detailed telematics data from their cars in an anonymous fashion. As opposed to the limited data collected for insurance purposes, the data in this case has minute-by-minute or second-by-second updates on speed, location, direction, and other useful information.

This data feed will provide information on thousands of cars in any given traffic jam on any given day. Researchers will know how fast each car was moving along the way. They will know where traffic started, where it ended, and how long it lasted. This is an amazingly detailed view of the reality of traffic flow. Imagine the impact on the study of traffic jams and the planning of road systems!

LOOK BEYOND THE INTENDED USE

The wealth of possibilities for telematics data is a terrific example of putting big data to use in a way that wasn't initially foreseen. Often, the most powerful uses for a given data source will be something entirely different from why it was created. Be sure to consider alternative uses for every big data stream encountered.

Once researchers have access to thousands of cars in every rush hour, every day, in every city, they will have the ability to diagnose traffic causes and effects in immense detail. They'll be able to pinpoint the answers to questions such as:

- How does a tire in the road impact traffic?
- What happens if a left lane gets blocked?
- When a traffic light gets out of sync, what are the effects?
- Which traffic intersections are poorly timed, even if they're acting the way they were intended to act?
- How fast does a backup in one lane spread to other lanes?

It's almost impossible to effectively study such questions today, outside of very focused and expensive testing. It is possible to physically send people out to monitor a given stretch of road and record information. Or, to put down sensors to count cars that go by. Or, to install a video camera. But those options are very limited in practice due to costs.

It's a traffic engineer's dream to have the telematics information outlined here. If telematics devices do become common, any location populated enough to have traffic can be studied. The changes that are made to roads and traffic management systems, as well as the plans for how roads are built in the first place, will provide huge benefits to all of us. Telematics got its start as a mechanism to assist in insurance pricing. But it may well revolutionize how we manage our highway systems and improve our lives by reducing the stress and frustration we experience when sitting in traffic.

MULTIPLE INDUSTRIES: THE VALUE OF TEXT DATA

Text is one of the biggest and most common sources of big data. Just imagine how much text is out there. There are e-mails, text messages, tweets, social media postings, instant messages, real-time chats, and audio recordings that have been translated into text. Text data is one of the least structured and largest sources of big data in existence today. Luckily, a lot of work has been done already to tame text data and utilize it to make better business decisions.

Text analytics typically starts by parsing text and assigning meaning to the various words, phrases, and components that comprise it. This can be done by simple frequency counts or more sophisticated methods. There's an entire discipline called natural language processing that comes into play heavily in such analytics. We won't get into that here. Text mining tools are available as part of major analytical tool suites, and there are also standalone text mining packages available. Some of these text analysis tools focus on a rules-based approach where users have to tune the software to identify the patterns that they're interested in. Others use machine learning and other algorithms that will help to find patterns within the data automatically. Each approach has advantages and disadvantages, but that discussion is out of scope

for our purposes. We'll focus here on how to use the results, not produce them.

Once the parsing and classification phases are done, the results of those processes can be analyzed. The output of a text mining exercise is often an input to other analytic processes. For example, once the sentiment of a customer's e-mail is identified, it is possible to generate a variable that tags the customer's sentiment as negative or positive. That tag is now a piece of structured data that can be fed into an analytics process. Creating structured data out of unstructured text is often called information extraction.

For another example, assume that we've identified which specific products a customer commented about in his or her communications with our company. We can then generate a set of variables that identify the products discussed by the customer. Those variables are again metrics that are structured and can be used for analysis purposes. These examples illustrate how it is possible to capture pieces of unstructured data and create relevant and structured data from it.

CREATE STRUCTURE WHERE NONE EXISTS

Text analysis is a terrific example of taking purely unstructured data, processing it, and creating structured data that can be used by traditional analytics and reporting processes. One major part of taming big data is getting creative in the ways that unstructured and semi-structured data is made usable in this way.

Interpreting text data is actually quite difficult. The words we say change meaning based on which words we emphasize and also the context in which we state them. When looking at pure text, you won't know where the emphasis was placed, and you often won't know the full context. This means that you'll have to make some assumptions. We'll discuss this issue more in Chapter 6.

Text analysis is both an art and a science and it will always contain a level of uncertainty. When doing text analysis, there will be issues with misclassification as well as issues with ambiguity. That's okay. If a pattern can be found within a set of text that enables a better decision to be made, then it should be used. The goal of text analysis is improvement of the decisions being made, not perfection. Text data can easily cross the bar of improving decisions and providing better

information than was present without it. This is true even given the noise and ambiguity that it contains.

Using Text Data

One popular use of text analysis today is what's known as sentiment analysis. Sentiment analysis looks at the general direction of opinion across a large number of people to provide information on what the market is saying, thinking, and feeling about an organization. It often uses data from social media sites. Examples include:

- What's the "buzz" around a company or product?
- Which corporate initiatives are people talking about?
- Are people saying good or bad things about an organization and the products and services it offers?

As discussed earlier, one tough part of text analysis is that words can be good or bad depending on the context. It will be necessary to take that into account, but across a lot of individuals the direction of sentiment should become clear. Getting a read on the trends of what people are saying across social media outlets or within customer service interactions can be immensely valuable in planning what to do next.

If an organization captures sentiment information at an individual customer level, it will provide a view into customers' intent and attitudes. Similar to how it is possible to use web data to infer intent, knowing whether a customer's general sentiment about a product is positive or negative is valuable information. This is particularly true if the customer hasn't yet purchased that product. Sentiment analysis will provide information on how easy it is going to be to convince that customer to purchase that product.

Another use for text data is pattern recognition. By sorting through complaints, repair notes, and other comments made by customers, an organization will be more quickly able to identify and fix problems before they become bigger issues. As a product is first released and complaints start to come in, text analysis can identify the specific areas where customers are having problems. It may even be possible to identify a brewing issue in advance of a wave of customer service calls coming in. This will enable a much faster, more proactive reaction. The corporate response will be better both in terms of putting in place

a fix to address the problem in future products, and also in what can be done to reach out to customers and mitigate the issues that they're experiencing today.

Fraud detection is also a major application for text data. Within health insurance or disability insurance claims, for example, it's possible to use text analysis to parse out the comments and justifications that have been submitted. Then, patterns can be identified that are associated with fraud so that claims can be flagged as high or low risk. Claims with higher risk patterns can be checked much more carefully. On the flip side, it's possible to do some claims automation. If there are patterns, terms, and phrases that are associated with clean, valid claims, those claims can be identified as low-risk and can be expedited through the system while resources are focused on the claims that have a higher risk.

Legal endeavors also benefit from text analysis. In a legal case, it is routine that e-mail or other messaging histories are subpoenaed. The messages are then examined in bulk to identify statements that may contain information tied to the case at hand. For example, which e-mails have potential insider information in them? Which people made fraudulent statements as they interacted with others? What is the specific nature of threats that were made?

Applying such analytics in a legal setting is often called eDiscovery. All of the preceding analytics can lead to successful prosecutions. Without text analysis, it would be almost impossible to manually scan all the documents required. Even if an effort was made to manually scan them, there would be a good chance of missing key information due to the monotonous nature of the task.

Text data has the potential to impact every industry. It will be one of the most widely used forms of big data. Learning how to capture, parse, and analyze text is critical for organizations. Text is one big data source that must be tamed.

MULTIPLE INDUSTRIES: THE VALUE OF TIME AND LOCATION DATA

With the advent of global positioning systems (GPS), personal GPS devices, and cellular phones, time and location information is a

growing source of data. A wide variety of services and applications from foursquare, to Google Places, to Facebook Places are centered on registering where a person is at a given point in time. Cell phone applications can record your location and movement on your behalf. Cell phones can even provide a fairly accurate location using cell tower signals if a phone is not formally GPS-enabled.

There are some very novel ways that consumer applications use this information, which leads to individuals allowing it to be captured. For example, there are applications that allow you to track the exact routes you travel when you exercise, how long the routes are, and how long it takes you to complete the routes. The fact is, if you carry a cell phone, you can keep a record of everywhere you've been. You can also open up that data to others if you choose. As more individuals open up their time and location data more publicly, very interesting possibilities start to emerge.

Many organizations are starting to realize the power of knowing "when" their customers are "where" and are attempting to get permission to collect such information from their customers. Of course, this should always be done on an opt-in basis, and very clear privacy policies should be developed and adhered to rigorously. Today, organizations are coming up with compelling value propositions to convince customers to release time and location information to them.

Time and location data isn't just about consumers, however. The owner of a fleet of trucks is going to want to know where each is at any point in time. A pizza restaurant will want to know where each delivery person is at any given moment. Pet owners want to be able to locate pets if they get out of the house. A large banquet facility wants to know how efficiently servers are moving around and covering patrons in all areas of the facility.

As an organization collects time and location data on individual people and assets, it starts to get into the realm of big data quickly. This is especially true if frequent updates to that information are made. It's one thing to know where every truck ends up at the start and end of every day. It's another thing to know where every truck is every second of every day. Time and location data is going to continue to grow in adoption, application, and impact.

Using Time and Location Data

Time and location data is one of the most privacy-sensitive types of big data. There are serious questions that deal not just with privacy, but even with ethical and moral issues. Should chips be placed in children's arms so that they can be tracked down if they go missing? What about elderly people with dementia who are known to walk away from their house or care facility? Certainly the potential for the misuse of time and location data is high. But the upside when it is used appropriately is also high. Let's look at some examples.

Soon, people may be able to register with local police and fire agencies and provide information on where they typically travel. This way, if there's an event like a major accident, flood, fire, or closed road, people can receive an alert from the fire or police department telling them that one of the spots along their typical route has trouble and that they may want to go a different way. This could mitigate traffic disruption as people proactively avoid a problem area and save a lot of time individually by preventing them from getting stuck. Eventually, agencies may even be able to receive real time information on your location if you allow it.

One application of this data that's only beginning to be leveraged is the development of time- and location-sensitive offers. This is going to be huge in the future of marketing. It's no longer just about what offer to develop for a customer today or this week, but it's about what offer is best for that customer based on when they will be where. Today, this is typically achieved by having customers check in and report where they are so they can receive an offer. Eventually, organizations may track the whereabouts of customers continually and react as necessary.

For example, perhaps a customer communicates that he's going to be commuting home from the office at 5:30 and he's going to drive by Exit 5 sometime between 5:45 and 6:00. He's looking for dinner and wants to know what you have to offer if he stops by your store or restaurant. You need to provide him something that matches his need at that point in time in that place. Giving him an offer tomorrow morning via e-mail will be too late. You want to give him an offer that's good at only the location he'll pass and only in the brief time around when he'll pass it.

PROVIDE OFFERS FOR THE "HERE" AND "NOW"

An emerging trend in marketing is the generation of offers for customers that are only good for a specific time period and a specific location. Such offers can be far more powerful and targeted than offers for a broad range of time and locations. Early adopters of these approaches have seen eye-opening results.

Of course, the complexity of managing offers goes up a few notches because now it isn't just about keeping track of what offers each person is eligible for this week. Instead, it is necessary to worry about where each person is at any point in time and what offers they are eligible for as a result. Time- and location-based offers do add complexity and will be more difficult to manage. Over time, however, the success rate of such offers should greatly surpass traditional personalized offers if they are done well. History has repeatedly shown that the more targeted and specific an offer, the better the response will be.

Another application of time and location data deals with enhancing social network analysis. In addition to a wireless carrier being able to identify relationships based on voice or text interactions, time and location data allows identification of what people were at the same place at the same time. For example, who attended a given concert or movie? Who went to a specific sporting event? Who was dining at a specific restaurant at the same time?

By identifying who ends up in similar locations at similar times repeatedly, it is possible to identify people who may not know each other or be part of the same social network today, but who have a lot of common interests. Imagine a dating service with this information to help you find your match! It could be worth encouraging people to get to know each other or giving them offers for products that are relevant to the type of people and communities that they appear to be associated with.

Time and location data not only helps understand customers' historical patterns, but also allows accurately predicting where customers will be in the future. This is especially true for customers who stick to a regular schedule. If you know where a given person is and where they're heading, you can predict where they might be in 10 minutes or an hour based on that information. By looking at where customers

were going historically when on the same route, you can make an even more educated prediction as to where they are going now. At minimum, you can greatly narrow down the list of possibilities. This enables better targeting.

Watch for an explosion in the use of time and location data in the coming years. Opt-in processes and incentives for consumers will begin to mature. For now, be very cautious and ensure your customers explicitly agree to let you use information in these ways before you do it. This will enable messaging to become even more targeted and personal than it is today. The idea of getting offers that aren't targeted to the here and now may well be considered old-fashioned in the not-too-distant future.

RETAIL AND MANUFACTURING: THE VALUE OF RADIO FREQUENCY IDENTIFICATION DATA

A radio frequency identification tag, or RFID tag, is a small tag placed on objects like shipping pallets or product packages. It is important to note that an RFID tag contains a unique serial number, as opposed to a generic product identifier like a UPC code. In other words, it doesn't just identify that a pallet contains some Model 123 computers. It identifies the pallet as being a specific, unique set of Model 123 computers.

When an RFID reader sends out a signal, the RFID tag responds by sending information back. It's possible to have many tags respond to one query if they're all within range of the reader. This makes accounting for a lot of items easy. Even when items are stacked on top of one another or behind a wall, as long as the signals can penetrate, it will be possible to get a response. RFID tags remove the need to manually log or inventory each item and allow a census to be taken much more rapidly.

Most RFID tags used outside of very high value applications are known as passive. This means that the tags do not have an embedded battery. The radio waves from a reader create a magnetic field that is used to provide just enough power to allow a tag to send out the information embedded within it. While RFID technology has been around for a long time, costs were prohibitive for most applications. Today a passive tag costs just a few cents and prices continue to drop.

As prices continue to drop, the feasible uses will continue to expand. There are some technical issues with today's RFID technology. One example is that liquids can block signals. As time progresses, these issues should be solved with updates to the technologies used.

There are uses of RFID today that most people will have come in contact with. One use is the automated toll tags that allow drivers to pass through a toll booth on a highway without stopping. The way it works is that the card provided by the toll authority has an RFID tag in it. There are also readers placed on the road. As a car drives through, the tag will transmit back the car's data so that the fact that you went through the toll can be registered.

Another major use of RFID data is asset tracking. For example, an organization might tag every single PC, desk, or television that it owns. Such tags enable robust inventory tracking. They also enable alerts if items are moved outside of approved areas. For example, readers might be placed by exits. If a corporate asset moves through the door without having been granted prior approval, an alarm can be sounded and security can be alerted. This is similar to how the item tags at retail stores sound an alarm if they haven't been deactivated.

One of the biggest uses for RFID today is item and pallet tracking in the manufacturing and retail spaces. Each pallet a manufacturer sends a retailer, for example, may have a tag. It makes it easy to take stock of what's in a given distribution center or store. Eventually, it is possible that every individual product in a store that is above a trivial price point will end up being tagged with an RFID chip, or an updated technology that serves the same purpose. Now that we've covered what RFID data is, let's look at some examples of how RFID data can improve businesses today.

Using Radio Frequency Identification Data

One application where RFID can add value is in identifying situations where an item has no units on the shelf in a retail environment. If a reader is constantly polling the shelves to identify how many of each item remains, it can provide an alert when restocking is needed. RFID enables much better tracking of shelf availability because there's a key difference between being out of stock and having shelf availability. It's

entirely possible that the shelf in a store has no product on it, yet simultaneously there are five cases in the storage room in the back.

In such a scenario, any traditional out-of-stock analysis is going to show that there is plenty of stock remaining and nothing to worry about. When sales start to drop, people are going to wonder why. If products have an RFID tag, it is possible to identify that there are five units in the back, yet no product on the shelf. As a result, the problem can be fixed by moving product from the back room to the shelf. There are some challenges in terms of cost and technology in this example today, but work is being done to overcome them.

RFID can also be a big help for tracking the impact of promotional displays. Often, during a promotion, product may be displayed in multiple locations throughout a store. From traditional point-of-sale data, all that will be known is that an item on promotion sold. It isn't possible to know which display it came from. Through RFID tags, it is possible to identify which products were pulled from which displays. That makes it possible to assess how different locations in the store impact performance.

As RFID is combined with other data, it gets even more powerful. If a company has been collecting temperature data within a distribution center, product spoilage can be traced for items that were present during a specific power outage or other extreme event. Perhaps the temperature of a section of a warehouse got up to 90 degrees for 90 minutes during a power outage. With RFID it is possible to know exactly which pallets were in that part of distribution center at exactly that point in time and appropriate action can be taken. The warehouse data can then be matched to shipment data. A targeted recall can be issued if the products were likely to be damaged or retailers can be alerted to double-check their product as it arrives.

IT'S THE COMBINATION THAT COUNTS

With RFID data, like many other big data sources, the power isn't just in what RFID data can tell you uniquely by itself. It is in what it can tell you when combined with other data. It can't be stressed enough that a big data strategy must aim to incorporate big data into the same processes as other data. Big data can't be a stand-alone effort.

There are operational applications as well. Some distribution centers may tend to be too rough with merchandise and cause a high level of breakage. Perhaps this is true only for specific work teams or even specific workers. A human resources (HR) system will report who was working at any point in time. By combining that data with RFID data that shows when product was moved, it is possible to identify employees who have an unusually high rate of breakage, shrinkage, and theft. The combination of data allows stronger, better quality action.

A very interesting future application of RFID is tracking store shopping in a similar fashion as web shopping. If RFID readers are placed in shopping carts, it is possible to know exactly what customers put into their carts and exactly what order that they added those items. Even if individual items aren't tagged, the cart's path can still be identified. Many of the advantages of web data discussed in Chapter 2 are possible in a store environment through such a use of RFID. These last two examples are again cases where privacy is an issue that must be considered. Perhaps customers won't want their shopping in the store tracked to them. In that case, "anonymous" shopping trips can be tracked where the person who generated the data is not identified.

One last application of RFID relates to how fraud can be reduced as criminals attempt to return stolen items. If an item has an RFID tag, the retailer can identify through the tag's unique identifier that an item being returned was part of a stolen batch of product and take appropriate action. In fact, it may come to the point where an RFID tag identifier is included as part of a receipt and required by the returns process. A retailer will know which specific RFID tag was on the item you purchased, not just that you purchased an item generically. When you come to the returns desk, you need to be returning that specific item with that specific tag. You can't pick up another of the same item from the shelf and fraudulently return it with your receipt. Using RFID in this manner will make it much harder to perpetrate fraud.

RFID has the potential to have huge implications within the manufacturing and retail industries in the years to come. It has had a slower adoption rate than many hoped. But as tag prices continue to

drop and the quality of the tags and readers continues to improve, it will make financial sense to pursue wider adoption.

UTILITIES: THE VALUE OF SMART-GRID DATA

Smart grids are the next generation of electrical power infrastructure. A smart grid is much more advanced and robust than the traditional transmission lines all around us. A smart grid has highly sophisticated monitoring, communications, and generation systems that enable more consistent service and better recovery from outages or other problems. Various sensors and monitors keep tabs on many aspects of the power grid itself and the electricity flowing through it.

One aspect of a smart grid is what's known as a smart meter. A smart meter is an electric meter that replaces traditional meters. On the surface, a smart meter won't look much different than the meters we've always had; but a smart meter is much more functional than a traditional meter. Instead of a human meter reader having to physically visit a property and manually record consumption every few weeks or months, a smart meter automatically collects data on a regular basis, typically every 15 minutes to every hour. As a result, it's possible to have a much more robust view of power usage both for every household or business individually, as well as across a neighborhood or even the entire grid.

While we'll focus on smart meters here, the sensors placed throughout a smart grid deserve a mention. The data that utilities capture from unseen sensors placed throughout the smart grid dwarfs smart meter data in size. Synchrophasors that take 60 readings per second across the power system, and home area networks recording each appliance cycling on or off are just two examples. The average person won't have any idea that most of these other sensors exist, but they will be crucial for the utilities. Such sensors will capture a full range of data on the flow of power and the state of equipment throughout the power grid. The data generated will be very, very big.

Smart-grid technology is already in place in some parts of Europe and the Americas. Virtually every power grid in the world will be replaced with smart-grid technology over time. The amount of data on electricity usage that utility companies will have available to them

as a result of smart grids is going to grow exponentially. How might such data be used? Let's take a look.

Using Smart-Grid Data

From a power management perspective, the data from smart meters will help people to better understand demand levels from customers all the way up the chain. But the data can also benefit consumers. An individual homeowner, for example, will be able to explicitly test how much power various appliances use by simply turning them on while holding other things constant and then monitoring the detailed power usage statistics that flow from their smart meter.

Utilities around the world are already aggressively moving to pricing models that vary by time of day or demand, and the smart grid will only accelerate that. One of the primary goals of the utility firms is to utilize new pricing programs to influence customer behavior and reduce the demand during peak times. It is the peaks that require additional generation to be built, which drives significant costs and environmental impacts. If the cost of power can be flexibly applied by time of day and measured by the meter, customers can be incented to change their behavior. Lower peaks and more even demand equate to fewer new infrastructure requirements and lower costs.

The power company, of course, will be able to identify all sorts of additional trends through the data provided by smart meters. Which locations are drawing power on off-peak cycles? What customers have a similar daily or weekly cycle of power needs? A utility can segment customers based on usage patterns and develop products and programs that target specific segments. The data will also enable identification of specific locations that appear to have very unusual patterns. That might point to problems needing correction.

In effect, power companies will have the ability to do all of the customer analytics other industries have been able to do for years. Imagine a phone company knowing your month-end total bill, but none of your calls. Consider a retailer knowing only your total sales, but none of your purchase details. Think about a financial institution knowing only your month-end balances, but none of the movements of money and funds throughout the month. In many ways, power

companies have been dealing with data that is equally poor for understanding their customers. They had a simple month-end total usage, and even that month-end figure was often an estimated, not an actual, usage amount.

BIG DATA CAN TRANSFORM AN INDUSTRY

In some cases, big data will literally transform an industry and allow it to take the use of analytics to a whole new level. Smart-grid data in the utility industry is an example. No longer limited by monthly meter readings, information on usage will be available at intervals measured in seconds or minutes. Add to that the sophisticated sensors throughout the grid, and it is a whole different world from a data perspective. The analysis of this data will lead to innovation in rate plans, power management, and more.

With smart meter data, a whole variety of new analytics will be enabled that will benefit all. Consumers will have customized rate plans based on their individual usage patterns, similar to how telematics enables individualized rates for auto insurance. A customer who's using power during peak periods is going to be charged more than non-peak users. It will encourage us all to shift our usage patterns once we are able to see the incentives to do so. Perhaps we'll run the dishwasher late in the afternoon instead of right after lunch, for example.

Utilities will have much better forecasts of demand, as they are able to identify where demand is coming from in more detail. They'll know what types of customers are demanding power at what points in time. The utility can look for ways to drive different behaviors to even out demand and lower the frequency of unusual spikes in demand. All of this will limit the need for expensive new generation facilities.

Each household or business will gain the power through smart meter data to better track and proactively manage its energy usage. This will not only save energy and make the world more green, but it's going to help everyone save money. After all, if you are able to identify where you're spending more than you intended, you will adjust as needed. With only a monthly bill, it is impossible to identify such opportunities. Smart-meter data makes it simple.

GAMING: THE VALUE OF CASINO CHIP TRACKING DATA

Earlier, we discussed RFID technology as it is applied to the retail and manufacturing industries. However, RFID technology has a wide range of uses, many of which also lead to big data. One other use for RFID tags is to place them within the chips at a casino. Each chip, particularly high-value chips, can have its own embedded tag so that it can be uniquely identified via the tag's serial number.

Within a casino environment, slot machine play has been tracked for many years. Once you slide your frequent player card or credit card in the machine, every time you pull the handle or push a button it is tracked. Of course, so are the amounts of your wagers and any payouts that you receive. Robust analysis of slot patterns has been possible for years. Casinos did not have the capacity to capture such details from table games. By embedding tags within gaming chips, they're now evolving to have it.

Traditionally, a casino primarily tracked chips via robust security camera networks and people on the ground tasked with ensuring that chips are moving around appropriately. A pit boss watched frequent players and estimated their average bets and lengths of play in order to award frequent-player benefits. While pit bosses are very good at this and also leverage the help of other staff, it is still possible to end up giving too much or too little credit for play. This happens if a player is watched when he happens to bet more or less than usual. In fact, some players even try to game the system by increasing their bets when they think they are being watched.

THE SAME TECHNOLOGY CAN DRIVE MULTIPLE BIG DATA STREAMS

Retailers and manufacturers leverage RFID technology. So do casinos. How they use RFID is totally different in many ways, and has similarities in others. The interesting part is that a single technology can be used by different industries to create their own distinct sources of big data.

Just as the example of casino chip tracking is a unique, but additional, application of RFID, there will be others. This example

illustrates that some of the same underlying technologies will enable different big data streams that are similar in nature but completely different in scope and application. What's exciting is how one fundamental technology can have different and completely distinct uses that generate multiple forms of big data in multiple industries.

Using Casino Chip Tracking Data

One obvious benefit of casino chip tags is the ability to precisely track each player's wagers. This ensures a player gets full credit in a frequent-player program, but no more or less. This benefits players and the casino. For the casino, resources will be allocated more precisely to the correct players. Over-rewarding the wrong players and under-rewarding the right players both lead to suboptimal allocation of limited marketing resources. Players, of course, always want their credits to be accurate.

Wager data collected across players will allow casinos to better segment players and understand betting patterns. Who typically bets $5 at a time, yet every once in a while jumps up to a $100 bet? Who bets $10 every time? Players can be segmented based on these patterns. The betting patterns can also point to those who are card counting in blackjack as there are certain patterns of wagers that will become clear when a player is using card counting techniques.

With chip tracking, it is also a lot harder to purposely defraud a casino or even for a dealer to make a mistake. Since bets and payouts can be tracked to the chip, it's easy to go back and compare video of the results of a blackjack hand and the payouts that were made. Even if arms or heads obscure what chips were put down or picked up, the RFID readings will provide the details. It will allow casinos to identify errors or fraud that took place. One example is when a player puts down extra chips after the fact when the dealer is looking the other way.

Analysis over time can identify dealers or players involved in an unusual number of mistakes. This will lead to either addressing fraudulent activity or providing additional training for a dealer who just happens to make a lot of innocent mistakes. Errors will also be lowered in the counting of chips in the casino cage. Counting large stacks of

chips of different denomination is monotonous, and people can make mistakes. RFID allows faster, more accurate tallying.

Taking the preceding example further, the tracking of individual chips is a terrific deterrent for thieves. If a stack of chips is stolen, the RFID identifiers for those chips can be flagged as stolen. When someone comes in to cash the chips, or even sits at a table with the chips, the system can realize it and alert security. If thieves remove or alter the chips so that they can't be read, that will be a flag. The casino will know exactly what chip IDs exist and will expect each chip to report a valid ID. When a chip doesn't report an ID, or when the ID reported isn't valid, they can take action.

As with any business, the more a casino can stop fraud and ensure that appropriate payouts are being made, the less risk it has. This will lead to the ability to provide both better service and better odds to players since there will be fewer expenses to cover. It can be a win for both casinos and their players.

INDUSTRIAL ENGINES AND EQUIPMENT: THE VALUE OF SENSOR DATA

There are a lot of complex machines and engines in the world. These include aircraft, trains, military vehicles, construction equipment, drilling equipment, and others. Keeping such equipment running smoothly is absolutely critical given how much it costs. In recent years, embedded sensors have begun to be utilized in everything from aircraft engines to tanks in order to monitor the second-by-second or millisecond-by-millisecond status of the equipment.

Monitoring may be done in immense detail, particularly during testing and development. For example, as a new engine is developed, it's worth capturing as much detailed data as possible to identify if it's working as expected. It's quite expensive to replace a flawed component once an engine is released, so it is necessary to analyze performance carefully up front. Monitoring is also an ongoing endeavor. Perhaps not every detail is captured for every millisecond on an ongoing basis, but a large amount of detail is captured in order to assess equipment lifecycles and identify recurring issues.

Consider an engine. A sensor can capture everything from temperature, to revolutions per minute, to fuel intake rate, to oil pressure

level. This data can be captured as frequently as desired. All of this data gets massive quickly as the frequency of readings, number of metrics being read, and number of items being monitored in such a fashion increases. Why should we care? Let's look at a few examples.

Using Sensor Data

Engines are very complex. They contain many moving parts, must operate at high temperatures, and experience a wide range of operating conditions. Because of their cost they are expected to last many years. Stable and predictable performance is crucial, and lives often quite literally depend on it. For example, taking an aircraft out of service for maintenance will cost an airline or a country's air force a lot of money, but it must be done if a safety issue is identified. It's imperative to minimize the time that aircraft and aircraft engines, as well as other equipment, are out of service.

Strategies for minimizing down time include holding spare parts or engines that can be quickly swapped for the asset requiring maintenance, creating diagnostics to quickly identify the parts that must be replaced, and investing in more reliable versions of problem parts. All three of these strategies depend on data for effective implementation. Data is used to create diagnostic algorithms and as input to the algorithms to diagnose a specific problem. Engineering organizations can use sensor data to pinpoint the underlying causes of failure and design new safeguards for longer, more dependable operation. These considerations apply whether the engine is in an aircraft, watercraft, or ground-based equipment.

By capturing and analyzing detailed data on engine operations, it's possible to pinpoint specific patterns that lead to imminent failures. Patterns over time that lead to lower engine life and/or more frequent repair can also be identified. The number of permutations of the various readings, especially over time, makes analysis of this data a challenge. Not only does the process involve big data, but the analysis that must be developed is complex and difficult. Some examples of the types of questions that can be studied:

- Does a sudden drop in pressure indicate imminent failure with near certainty?

- Does a steady decrease in temperature over a period of a few hours point to other problems?
- What do unusual vibration levels imply?
- Does heavy engine revving upon start-up seriously degrade certain components and increase the frequency of maintenance required?
- Does a slightly low fuel pressure over a period of months lead to damage to some of the engine's components?

LACK OF STRUCTURE WITHIN STRUCTURED DATA

Sensor data offers a difficult challenge. While the data collected is structured and its individual data elements are well understood, the relationships and patterns between the elements over time may not be understood at all. Time delays and unmeasured external factors can add to the problem's difficulty. The process of identifying the long-term interactions of various readings is extremely difficult given all the information to consider. Having structured data doesn't guarantee a highly structured and standardized approach to analyzing it.

When major problems arise, it's immensely helpful to go back and examine what was taking place in the moments up until the problem manifested itself. In this case, sensors act similarly to how well-known airplane black boxes help diagnose the cause of an accident. The data from the sensors in an engine also provide data that can be leveraged for diagnostic and research purposes. The sensors being discussed here are conceptually a more sophisticated form of the telematics devices discussed in the auto insurance example. The use of data from sensors that are continually surveying their surroundings is a recurrent theme in the world of big data. While we have focused on engines here, there are countless other ways that sensors are also being used today, and the same principles discussed here apply to those uses.

If the sensor data capture process is repeated across a lot of engines and a long period of time, it leads to a wealth of data to analyze. Analyze it well, and it is possible to find glitches in equipment that can be proactively fixed. Weak points can also be identified. Then,

procedures can be developed to mitigate the problems that result from those findings. The benefits aren't just added safety but also lowered cost. As sensor data enables safer engines and equipment that remain in service a higher percentage of the time, it will enable both smoother operations and lower costs. That is a win all around.

VIDEO GAMES: THE VALUE OF TELEMETRY DATA

Telemetry is the term used in the video game industry to describe the capture of in-game activities. It has conceptual similarities to the web log data discussed in Chapter 2. This is because telemetry data captures the actions players take while navigating a game. Telemetry is most often captured for online games as opposed to game consoles.

In a hockey game, telemetry data will capture things such as where a player was when a shot on goal was taken, what type of shot it was, what speed the shot had, and if the shot was successful. In a war game, telemetry data will capture what weapon was fired, from where it was fired, what direction it was fired, and what damage the weapon did to various objects. Theoretically, any level of detail about the scene and actions can be captured.

This takes a video game producer well beyond simply knowing how many customers bought a game and perhaps how many hours games are played. Telemetry data makes it possible for game producers to know intimate details about how customers actually play and interact with the games they've created. The amount of data captured can be huge, and the video game industry is just starting to analyze this data in earnest. There are many areas where telemetry can have an impact. It will be easy to see the parallel between telemetry data and web data in terms of its advantages and uses. Let's take a look at a few.

Using Telemetry Data

Many games make money through subscriptions, so maintaining renewal rates is absolutely critical. By mining players' playing patterns, insights can be obtained into what types of game behavior are associated with renewals and which are not. For example, perhaps playing tournaments in a sports game while leveraging certain addi-

tional features leads to a large increase in renewal rates. The game manufacturer can then incent players to give tournaments a try while using those features if they haven't already done so.

TELEMETRY DATA WILL ONLY GET BIGGER

Currently, telemetry data largely captures controller or keyboard actions. As interactive games evolve that track player motions as opposed to depending on controllers, the amount of data will increase immensely. Knowing a player pressed a given button at a point in time is a lot less data than knowing where in space each of his many body parts was at the time and in what direction and with what speed each body part was moving.

Newer games often offer in-game purchases for a small fee. These are known as microtransactions. A special weapon might be available for 10 cents, for example. By analyzing game play, it's possible to identify specific areas in the game where such a microtransaction will have a high success rate. Perhaps there's a certain point in the game where a special weapon comes in very handy because many players struggle. A quick reminder on the screen that the weapon is available will lead to many players accepting that offer and making the purchase.

Customer satisfaction is as big an issue in the video game industry as in any other industry. What is unique is that video games have a very, very fine line to walk. A game has to offer a challenge to its players, but it can't be so challenging that it becomes frustrating and causes people to give up. If a game is either too easy or too hard, players will tire of it and move on.

By analyzing game play, it is possible to identify specific parts of a game that are easily passed by almost every player and specific parts that even top players have a very hard time completing. Such areas might be tweaked by adding or removing enemies, for example, to even out the difficulty level. Stabilizing the difficulty level across a game will provide a more consistent, more satisfying experience for the players. This will lead to higher renewal rates and additional purchases.

Through telemetry data, players can also be segmented by the style of play they utilize. Such information can be used to both design new

games and cross-sell other existing products. One player segment may go as fast as possible to beat a level without concern for anything but completing the level. Another segment may try to collect all the bonus items available before completing a level. Yet another segment might aim to explore every nook and cranny of a level as they work to the end. Based on those traits, a player can be educated about other games that benefit from his or her preferred method of play.

The level of knowledge about players that telemetry data can provide has the potential to totally change the video game industry. While the industry is just starting to use telemetry data, look for rapid advances in this area to occur in the near future. Also look for major changes to how games are made and marketed as a result of what the analysis of telemetry data unveils.

TELECOMMUNICATIONS AND OTHER INDUSTRIES: THE VALUE OF SOCIAL NETWORK DATA

Social network data qualifies as a big data source even though in many ways it's more of an analysis methodology against traditional data. The reason is that the process of executing social network analysis requires taking already-large data sets and using them in a way that effectively increases their size by orders of magnitude.

One could argue that the complete set of cell phone calls or text message records captured by a cellular carrier is big data in and of itself. Such data is routinely used for a variety of purposes. Social network analysis, however, is going to take it up a notch by looking into several degrees of association instead of just one. That's why social network analysis can turn traditional data sources into big data.

For a modern phone company, it is no longer sufficient to look at all calls and analyze them as individual entities. With social network analysis, it is necessary to look at who the calls were between and then extend that view deeper. You not only need to know who I called, but who the people I called in turn called, and who those people in turn called, and so on. To get a more complete picture of a social network it is possible to go as many layers deep as analysis systems can handle. The need to navigate from customer to customer and call to call for several layers makes the volume of data become

multiplied. It also increases the difficulty of analyzing it, particularly when it comes to traditional tools.

The same concepts apply to social networking sites. When analyzing any given member of a social network, it isn't that hard to identify how many connections the member has, how often she posts messages, how often she visits the site, and other standard metrics. However, knowing how wide a network a given member has when including friends, friends of friends, and friends of friends of friends, requires much more processing.

One thousand members or subscribers aren't hard to keep track of. But, they can have up to one million direct connections between them and up to one billion connections when "friends of friends" are considered. That is why social networking analysis is a big data problem. The analysis of such connections has a number of applications being pursued today.

Using Social Network Data

Social network data, and the analysis of it, has some high-impact applications. One important application is changing the way organizations value customers. Instead of solely looking at a customer's individual value, it is now possible to explore the value of his or her overall network. The example we're about to discuss could apply to a wide range of other industries where relationships between people or groups are known, but we'll focus on wireless phones since that is where the methods have been most widely adopted.

Assume a wireless carrier has a relatively low-value customer as a subscriber. The customer has only a basic call plan and doesn't generate any ancillary revenue. In fact, the customer is barely profitable, if the customer generates profit at all. A carrier would traditionally value this customer based on his or her individual account. Historically speaking, when such a customer calls to complain and threaten to leave, the company may have let the customer go. The customer just isn't worth the cost of saving.

With social network analysis it is possible identify that the few people our customer does call on that cheap calling plan are very heavy users who have very wide networks of friends. In other words,

the connections that customer has are very valuable to the organization. Studies have shown that once one member of a calling circle leaves, others are more likely to leave and follow the first. As more members of the circle leave, it can catch like a contagion and soon circle members are dropping fast. This is obviously a bad thing.

LOOK BEYOND INDIVIDUAL VALUE

A very compelling benefit of social network data is the ability to identify the total revenue that a customer influences rather than just the direct revenue that he or she provides. This can lead to drastically different decisions about how to invest in that customer. A highly influential customer needs to be coddled well beyond what his or her direct value indicates if maximizing a network's total profitability takes priority over maximizing each account's individual profitability.

Using social network analysis, it is possible to understand the total value that the customer in our example influences for the organization rather than only the revenue directly generated. This allows a completely different decision regarding how to handle that customer. The wireless carrier may overinvest in the customer to make sure that they protect the network the customer is a part of. A business case can be made to provide incentives that dwarf the customer's individual value if doing so will protect the wider circle of customers that he or she is a part of.

This is a terrific example of how the analysis of big data can help provide new contexts in which decisions that would have never happened in the past now make perfect sense. Without big data, the customer would have been allowed to cancel and the wireless carrier would not have seen the avalanche of losses that soon came as the customer's friends followed. The goal shifts from maximizing individual account profitability to maximizing the profitability of the customer's network.

Identification of highly connected customers can also pinpoint where to focus efforts to influence brand image. Highly connected customers can be provided free trials and their feedback can be solicited. Attempts can be made to get them active on corporate social networking sites through incentives to provide commentary and

opinion. Some organizations actively recruit influential customers and shower them with perks, advance trials, and other goodies. In return, those influential customers continue to wield their influence, and it should be even more positive in tone given the special treatment they are receiving.

Within social networking sites like LinkedIn or Facebook, social network analysis can yield insights into what advertisements might appeal to given users. This is done by considering not just interests customers have personally stated. Equally important, it is based on knowing what it is that their circle of friends or colleagues has an interest in. Members will never declare all of their interests on a social networking site, and every detail about them will never be known. However, if a good portion of a customer's friends have an interest in biking, for example, it can be inferred that the customer does as well, even though the customer never stated that directly.

Law enforcement and antiterrorism efforts can leverage social network analysis. Is it possible to identify people who are linked, even if indirectly, to known trouble groups or persons? Analysis of this type is often referred to as link analysis. It could be an individual, a group, or even a club or restaurant that is known to attract a bad element. If a person is found to be hanging out with a lot of these elements in a lot of these places, he or she might be targeted for a deeper look. While this is another analysis that raises privacy concerns, it is being used in real life situations today.

Online video gaming is another area where such analysis can be valuable. Who plays with whom? How does that pattern change across games? Social network analysis can augment the telemetry data we discussed earlier. It is possible to identify a player's preferred partners on a game-by-game basis. We earlier discussed how players can be segmented based on individual playing style. Do players of similar styles team up when playing together? Or do players seek a mix of styles? Knowing such information can be valuable to know if a game producer wants to suggest groups for players to team up with (for example, providing suggestions as to which group, out of the many options available in a list, a player might prefer to join in with when he or she logs in to play).

There have also been interesting studies on how organizations are connected. It starts by looking at the connections established through

e-mail, phone calls, and text messages within an organization. Are departments interacting as expected? Are some employees going outside typical channels to make things happen? Who has a wide internal influence and would be a good person to take part in a study on how to better improve communications within the organization? This type of analysis can help organizations better understand how their people communicate.

Social network analysis will continue to grow in prevalence and impact. One of the neat features it has is that it is always going to make a data source much bigger than it began due to the exponentially expanding nature of the analysis process. Perhaps the neatest feature is how it can provide insights into a customer's total influence and value, which can completely change how that customer is viewed by an organization.

WRAP-UP

The most important lessons to take away from this chapter are:

- While there are a wide range of big data sources across industries, common themes do exist. The same underlying technologies, such as RFID, can be utilized in multiple industries for different purposes.
- There are privacy implications with many big data sources, and privacy needs to be a serious consideration at all times.
- Telematics data is evolving to enable better pricing of auto insurance policies. However, the data collected also has the potential to revolutionize traffic management and planning.
- Text data is one of the biggest and most widely applicable types of big data. The focus is typically to extract key facts from the text and then use those facts as inputs to other analytic processes.
- Time and location data is a growing influence. Organizations will have to become much more sophisticated in order to target offers to customers at a given time and place.
- RFID data enables new analytics for retailers and manufacturers in areas from stock levels, to fraud, to employee performance.

- Smart grids will not only give utilities the ability to much better manage their power grids, but consumers will be enabled to better control their consumption.
- Tracking casino chips with RFID tags can help casinos more accurately track player activity, while also reducing payout errors and fraud.
- Sensor data provides powerful information on the performance of engines and machinery. It enables diagnosis of problems more easily and faster development of mitigation procedures.
- Telemetry data will enable video game producers to better target micro-transactions, make game flow improvements, and segment players by playing style.
- Social network data can lead to new ways of valuing customers. In the telecommunications industry, social network analysis has shifted focus from account profitability to network profitability.