Organizations in many different industries are in the early stages of developing cognitive applications. From healthcare to manufacturing to governments, decision makers need to quickly make sense of large volumes and varieties of data. Problem solving often requires the aggregation of a multitude of disconnected data sources including a combination of internal and external data. In addition, it is increasingly likely that the data required to answer problems or deliver new insights is unstructured—such as text, videos, images, sound, or sensor data. Valuable insights may remain hidden because the volume, variety, and velocity (speed) of the data are so hard to manage. Organizations are now beginning to recognize the potential benefits of using cognitive applications to find the patterns in data that can help to improve outcomes.

Chapters 11–13 provide examples of emerging cognitive computing applications across multiple industries. Although the domains and applications described in these chapters differ, certain common attributes of each situation make them a good fit for cognitive applications. Organizations that are implementing cognitive applications typically face similar challenges regarding data and the decision making process such as:

- Large volumes of unstructured data that must be analyzed to make good decisions.
- Decisions must be based on constantly changing data, new sources, and forms of data.

- A significant amount of knowledge about the domain is transferred from senior experts to trainees through a mentoring and training process.
- Decision making requires the analysis of a variety of options and solutions to a problem. Individuals often have to quickly weigh the relative risks and benefits of each alternative and may have to decide based on confidence rather than certainty.

This chapter examines the seven key steps involved in designing a typical cognitive application:

1. Defining the objective
2. Defining the domain
3. Understanding the intended users and their attributes
4. Defining questions and exploring insights
5. Acquiring the relevant data sources
6. Creating and refining the corpora
7. Training and testing

The Emerging Cognitive Platform

The majority of early cognitive applications have been built from scratch by vendors in collaboration with their customers. The vendors and customers were experimenting and learning together. As the number of cognitive applications under development and deployment grows, vendors are using their experience to codify packaged services, APIs and delivery models that can help customers build cognitive applications more independently and quickly. Most new cognitive applications are developed on a cloud-based cognitive engine that provides the ability to scale processing, storage, and memory. In addition, customers will access a set of well-defined foundational services to speed development of cognitive applications. These foundational services may include a corpus service, analytics service, a data engine such as a graph database, training services, presentation and visualization services, and others. The expectation is that moving forward, cognitive applications will be built on an engine and well-defined APIs that provide some or all these foundational services.

In many ways, vendors are collaborating with partners on these early cognitive applications in a role similar to a systems integrator. Vendors are responsible for the development of the cognitive engine, but much of the development of associated tools and services are created jointly with partners based on their requirements. The partners begin by focusing on defining the domain for their cognitive application, collecting and curating the data sources, and understanding

the types of questions and information that their users will be interested in. Typically, the development of the model, including the training and testing of the system, is completed in collaboration with the vendors providing the cognitive platform.

Each phase of developing a cognitive application can be time-intensive and requires the input of domain experts and end users. Many initiatives have required a significant amount of manual intervention in areas such as building and refining the corpus and training and testing the system. If cognitive applications are going to become more accepted and deliver value across many industries, vendors need to provide packages and tools that enable customers to get new applications up and running quickly. One of the most time-consuming aspects of building a cognitive application is selecting, accessing, acquiring, and preparing data for the corpus. Therefore, vendors are beginning to offer corpus services that include industry-specific, pre-ingested, and curated data. For example, in the healthcare industry, these sources might include a healthcare-specific semantic taxonomy and ontology of disease codes and symptoms. Training is critical to the success of the applications and can also be time-intensive. Vendors could provide a pretrained data set for the application in a particular domain or problem area. There will also be extensive use of APIs that abstract some of the challenging aspects of developing and maintaining the application. For example, APIs can simplify the process of importing data for visualization rendering or extracting relationships from data.

Defining the Objective

Creating a cognitive application has much in common with developing any other enterprise application. You need to understand what the objectives are for your application and how you will achieve those objectives. Therefore, the first step in developing a cognitive application is to understand the types of problems your cognitive application is going to solve. Your objective needs to consider the types of users you will be appealing to and if there will be multiple constituencies in your user base. What issues will your users be interested in, and what do they need to know? One of the unique differences between a cognitive application and traditional applications is that users should expect more than answers to queries. A cognitive application should provide answers to questions but also go deeper and explore context related to how and why something happened.

Building traditional applications often begins with business process. In contrast, in a cognitive application you need to develop an objective based on knowledge and data. Therefore, in the design you need to set some parameters around the type of knowledge that is pivotal to your corpora. In other words, your objective should probably focus on a specific segment of an industry rather than attempt

to solve all problems for a particular industry. Several examples of objectives for cognitive healthcare applications follow:

- Provide personalized information and social support to help individuals optimize their health and wellness.
- Help health consumers take a more active role in managing their own health and the health of people they help care for.
- Help determine if the treatment plan selected for the patient is the best and most cost-effective.
- Provide additional knowledge to medical students to support what they learn on their subspecialty rotations.

Cognitive applications are also a good solution in situations in which you would like to provide assistance to a customer service representative or salesperson. Consider a retail organization with a large and diverse group of sellers. This company has a small number of sellers with many years of experience and deep knowledge of the products the company sells. If a customer has specific requests or needs help comparing alternatives, these knowledgeable sellers can answer all their questions and make the sale. However, the company also has high turnover, and many sellers lack the knowledge to provide the right level of support to customers. The company decides to introduce a cognitive application designed to make all the sellers as smart as the most knowledgeable and experienced seller in the company.

Defining the Domain

The next step is to specify the domain or subject area for your cognitive application. Defining your domain is a prerequisite to identify and assess which data sources you will need for your applications. In addition, your domain definition will be useful in determining the subject matter experts that will be helpful in training the system. Table 10-1 provides a few examples of cognitive application domains and a sample of the data sources and subject matter experts that can create the knowledge base for that domain. As described in the previous section, your stated objective is likely to help narrow the domain focus. For example, a cognitive application designed to train medical students would require medicine as the domain, whereas a cognitive application designed to help clinicians select the right treatment plan for breast cancer patients would require breast oncology as the domain. The medical domain will require comprehensive and broad-based medical taxonomies, ontologies, and catalogues, whereas the breast oncology domain will require segments of the medical ontology as well as additional data specific to the field.

Table 10-1: Examples of Cognitive Application Domains

Domain	Selected Data Requirements	Subject Matter Experts
Medicine	International Classification of Diseases (ICD) codes, Electronic Medical Records (EMR), and research journals	Senior physicians in medicine and key specialties
Airplane Manufacturing and Maintenance	Complete parts list, maintenance records per airplane, and spare parts inventory	Mechanics who know how to anticipate failure and create effective repairs and experienced pilots
Retail	Customer and product data	Experienced sales associates

Although the domain helps to define the data sources you might need, you may also include data sources that are not typically associated with solving problems in that domain. The inclusion of non-typical data sources is needed because cognitive systems support problem solving in a different way from traditional systems. A cognitive application is good at helping users to assimilate knowledge quickly and efficiently. Although some of this knowledge might be found in specific data sources, it may also incorporate information that is typically learned by experience. One of the advantages of a cognitive application is that it can provide every user with proven business practices and industry-specific knowledge that is well known to your most experienced domain experts. A cognitive system's greatest value comes from its ability to combine information from industry data sources with testing and refinement based on interactions with highly experienced experts. For example, when faced with an unusual problem, an airplane mechanic with 30 years of experience might remember similar situations that occurred in the past and recommend something like, "the problem is likely to be A or B and we should take these five steps to get the best result."

Understanding the Intended Users and Defining their Attributes

You need to understand the types of users who will be accessing your cognitive application. Expectations for user and system interactions will have an impact on the development of the corpus, the design of the user interface, and how the system is trained. The level of accuracy required in a cognitive application will depend on the intended use case. For example, a scientist requires a much more precise level of accuracy than a customer service representative answering questions about replacement parts. However, it is unnecessary and unwise to attempt

to anticipate all the questions your users will ask and all the different ways your cognitive application will be used. A cognitive application assumes that the data will grow and change as new data sources are discovered and added. In addition, the machine learning algorithms will refine the way questions are analyzed and answered. You need to build in flexibility so that your application can change as user requirements change. The learning process for a cognitive system is continuous, and as a result, your application will get smarter and deliver greater value to end users the more it is used.

The following best practices can help to ensure that your cognitive application has the flexibility it needs to provide the right level of support for its users:

- **Understand your users' level of understanding of the domain.** Is your cognitive application intended for consumers or domain experts? Will your users understand the meaning of industry-specific terms? Will your cognitive application be used to help train users in a particular domain?
- **Plan for variations in types of questions and analysis required.** Will your application have users with varied backgrounds and levels of expertise? For example, if you are planning for consumers and domain experts, they are likely to ask questions using different words and language styles. Although these users may be looking for insights into a similar topic, they may have widely differing expectations for the level of insight required. The consumer may look for a definition, whereas the domain expert wants to compare alternative solutions to a complex problem.
- **Keep the scope of your application broad enough to support different types of users.** If you are too specific or narrow in your definition of the domain, there may be subject areas that are not covered adequately in the corpus. It is better to err on the side of a little more coverage of your domain than less, as the learning process will successively refine the corpus toward the "right size" with increased usage.

Defining Questions and Exploring Insights

As discussed in Chapter 1, "The Foundation of Cognitive Computing," a cognitive system delivers insight relative to a domain, a topic, a person, or an issue, based on training and observations from all varieties, volumes, and velocities of data. A cognitive system creates models to represent the domain and generates and scores hypotheses to answer questions or provide insight. To ensure your cognitive application delivers the insights your users are looking for, you need to begin by mapping out the types of questions they may ask. Users of a well-defined and trained cognitive application can benefit in many ways. One significant benefit is the capability to receive alternative answers to questions along with associated confidence levels. These benefits can be achieved only if

the right set of data for the domain and the system is ingested into the corpus and then is properly trained and tested. However, before you even begin to train the system, you need to consider the types of questions your users will ask and the types of insight that your users will be looking for.

Many of the early cognitive applications are of two main types: customer engagement, or discovery and exploration. Customer or user engagement applications typically leverage advanced Question-Answer Systems designed to answer questions as part of an ongoing dialogue with the user. Answers to questions may be provided as a set of alternatives with associated confidence levels. Discovery and exploration applications begin with data analysis rather than by asking questions. You may not know what to expect or exactly what questions to ask. Discovery applications are used in situations such as genomic exploration, security analysis, or threat prevention. Typically, in these situations, your cognitive application will begin by looking for patterns and anomalies in the data.

Because the Question-Analysis approach to cognitive applications requires a more rigid structure to understand potential user questions, this process is described next. All questions need to be suitable for evidenced-based analysis; however, the questions do not all need to be initiated by the user. Actually, one of the defining features of a cognitive application is that users can engage in a dialogue with the system. In anticipatory systems, the application is designed to analyze the data and make suggestions or recommendations for the user without the user needing to ask a specific question. As a result, the user can move through paths of analysis not previously anticipated and develop new insights based on the user/application interaction. The cognitive system can make associations between questions, answers, and content to help the user understand the subject matter at a deeper level. The questions users will ask can be placed in two general categories:

- **Question-Answer pairs**—The answers to these questions can be found in a data source. There may be conflicting answers within the data sources, and the cognitive system will analyze the alternatives to provide multiple responses with associated confidence levels.
- **Anticipatory analytics**—The user engages in a dialogue with the cognitive application. The user may ask some questions but not all the questions. The cognitive application will use predictive models to anticipate the user's next question or series of questions.

Typical Question-Answer Pairs

Developers of question-answer cognitive applications have found that they need to begin with approximately 1,000–2,000 pairs of questions and answers. You have already defined the users of your application, and you need to keep them in mind when creating the question-answer pairs. How will your representative

group of users ask questions? Consider not only the content of the question, but also how it will be asked. The questions need to be in the voice of the end user. What style of language will they use? What technical terms are they likely to know? There are often many ways to ask the same question, and you need to consider the alternative styles of questioning when developing these initial questions. Although the answers need to use terms and a language style that will be understood by users, the content of the answer needs to be vetted by subject matter experts.

Table 10-2 provides an example of two questions that might be asked of a medical cognitive application, related to the use of morcellators: a health consumer asks one question and a gynecologist asks the other. The health consumer is looking for a definition, whereas the health professional is looking for more details on risks and benefits of a specific procedure. In a cognitive application, users of both types could engage in a dialogue that would provide more granular information on the topic.

table 10-2: Question-Answer Pairs for Different Types of Users

question	answer
Heath consumer: What is a morcellator?	A morcellator is a device with a spinning blade that is used to shred a fibroid through an incision on a woman's abdomen. The force and speed of the device may cause cellular particles from the fibroid to become dispersed in the abdomen.
Gynecologist: What are the risks and benefits of using a morcellator for surgical treatment of fibroids?	Risks include potential spread of an occult uterine sarcoma. Benefits include smaller incisions for the patient, less bleeding, and quicker healing and recovery.

You should define a sample set of questions prior to selecting the data sources needed to build the corpus. By choosing your information sources based on what is needed to answer a representative set of questions, your system can learn how to answer similar questions in the same domain. If you build the corpus first, you may make the mistake of tailoring your questions for training and testing to the information you already have at hand. When your cognitive application becomes operational, your users may have questions that cannot be answered by the system. It is expected that the corpus will need to be continuously updated during training and operation; however, you want to start out by including as many data sources as required to provide the right level of insight within your chosen domain.

Anticipatory Analytics

What if the user is not in a position to ask a specific question of the cognitive application? Anticipatory analytics can be used when there are many

unknown factors making it difficult for a user to know what questions to ask. For example, in military or security analytics, you may not know when or where a future event will occur or even what event will occur. You need to observe the data and look for patterns without knowing what you are looking for. The data you need to observe and analyze may be unclean or subject to inconsistent definitions and inconsistencies in metrics or measurements around time and place. However, when this data is used for a cognitive security application, unclean data may provide valuable clues to anticipate events or actions. The anomalies or outliers in the data are used to build the models and anticipate changes that can identify security threats or military events in time to take corrective action.

Anticipatory analytics is also used in cognitive applications that are designed to understand an individual's personal needs and help them to make good decisions. Because a user does not need to ask a question to be provided with a recommended action, the creators of the application need to focus on the different personal situations that might be best suited for assistance by a cognitive system. For example, a cognitive assistant could monitor a user's schedule and alert the user if there is a delay in a scheduled air flight or train. By monitoring personal medical devices and applications, a cognitive assistant could alert users they may be getting sick or help them keep on track with dietary goals. Users are increasingly sharing a lot of personal information on a variety of applications and devices—ranging from health monitoring devices to e-mail, travel, and calendar applications. A cognitive application can be trained to integrate this information to learn a lot about you. In addition, a cognitive application can be designed to be aware of what is happening in the world around you through geospatial, travel, health, and other applications. Therefore, a cognitive application that understands your location, your health and medical status, and the context of your questions can make personalized recommendations. An anticipatory cognitive application leverages data to make personal tasks easier and provide information you need before you ask for it.

Cognitive Commerce

Cognitive commerce refers to a cognitive application designed to anticipate user needs from a retail or commerce perspective. Organizations with mobile or Internet-based commerce sites are continuously trying to optimize their sites to increase sales. By making it easier for consumers to find what they want faster, these companies can reach their sales goals faster. For example, a company that provides streaming entertainment content could create a cognitive application to make it easier for customers to find the movie they want to watch and make it easier to watch on their mobile device.

Cognitive capabilities are built in to an existing commercial app or other environment. The user has previously provided permission to the commercial

Continues

(continued)

application to capture personal information (that is, health data, travel itinerary, and exercise tracking). As a result, the application can make suggestions or provide information to the user without the user needing to ask specific questions.

Builders of a cognitive application with commerce capabilities need to plan for the types of questions users will ask as well as the types of capabilities that will have a positive impact on sales. For example, you may expect that users will ask a question about ordering a specific item such as, "Do you have XBrand jeans available in dark wash size 29?" However, you may also want to plan for questions that are more open ended such as, "I saw the perfect silk dress on 'X' character in 'Y' on 'ABC' show. Can you find me something similar in size 4?" You may also want to submit a photographic image of a dress and ask the system to locate the item in a different color or size. A cognitive commerce application could accept complex user queries in natural language and make it much easier and faster for consumers to find the right item to purchase. In addition, by understanding your personal information in context, a cognitive commerce application can anticipate what you might like to buy next before you do.

Acquiring the Relevant Data Sources

When developing a corpus you should determine the most relevant data sources. This is challenging because you cannot know with certainty what type of insights users might require as their needs change over time. However, taking the time to evaluate data sources you currently own and those you may want to acquire also offers great opportunity. You may discover that you have internal data resources that can provide new insight when leveraged by a cognitive system. Additionally, you may want to include social media data or other external sources. Cognitive systems provide an opportunity to leverage data sources in new ways. To start building the corpus, you need to understand your requirements for a variety of internal and external data sources. As you move through the testing process and your application becomes operational, you need to be prepared to add new sources as they become available and the scope of the application expands.

The Importance of Leveraging Structured Data Sources

Much of the focus around cognitive computing has been data from unstructured data sources. However, cognitive solutions must gain insights based on the current state of customers or other constituents. Therefore, you need to know what internal data sources are going to be meaningful. For example, if the application is related to travel, the company needs internal data to relate to the details about customers or travel locations. A retail application needs data sources related to merchandise that has been ordered, what products have been sold, and who the customers are. A hospital-based healthcare application needs data on patient status, medical history, and hospital admissions. A manufacturing application

may need data that reports on sensor activity from the production floor. These data sources will most likely be stored as structured data in relational databases including customer data from a Customer Relationship Management system or patient data from an Electronic Medical Record for a healthcare application. Additionally, there could be streaming data sources that come from sensor networks.

Analyzing Dark Data

Dark data refers to data that has been stored over many years and sometimes decades. Much of this data has been stored but not previously analyzed. For example, dark data could be data about performance of a company's stock over a decade or data stored at the time of a security breach. With the cognitive system, the dark data can become the benchmark to analyze how things have changed over time. This data may provide new insights by using machine learning to look for patterns in data collected over many years. Given the advent of new analytics technologies, this dark data may now be an important internal data source depending on the domain.

Leveraging External Data

What external data sources will support users? External data sources may include everything from industry-specific technical journals that are focused on new research findings to industry taxonomies and ontologies. In medical research there are results from clinical trials that might provide insights into drug interactions. Most industries have a wealth of third-party databases with both structured and unstructured data. Increasingly, there are stores of videos, images, and sounds that are of particular interest to either a specific industry or a technical discipline.

Many industries have codified ontologies and taxonomies that are managed and updated by industry consortiums. These sources are critical in creating your corpus. However, you may find that you need to capture only a subset of the available data. These data sources often include the hierarchical classification of entities or concepts within a domain, which are important for determining context and meaning. Table 10-3 provides you with a sample of the types of ontologies and taxonomies available for certain specific industries.

You need to use caution when using these external data sources. For example, what is the origin of the data source? Who owns that data source and how and when was it created? More important, who is responsible for updating the data source on an ongoing basis? Equally important is the security and governance of the data sources. There are data sources that include private information that can be used under strict governance guidelines. If that data is misused, it can cause significant problems for an organization.

Table 10-3: Industry-Specific Taxonomies and Ontologies

Industry	Taxonomy/ Ontology	Publisher	Description
Healthcare	International Classification of Diseases (ICD)	World Health Organization	International codes for diseases, disease symptoms, and medical findings about diseases
Healthcare	Semantic taxonomy for the healthcare ecosystem	Developed by companies such as Healthline Corp.	Classifies healthcare information on the web and maps the relationship between consumer and clinical terminology
Construction	International Building Code (IBC)	International Conference of Building Officials	Standards and compliance regulations for international building codes
Finance	U.S. GAAP Financial Taxonomy	Financial Accounting Standards Board (FASB)	U.S.-based standards for financial accounting and reporting
Information Technology	NIST Cloud Computing Taxonomy	National Institute of Standards and Technology (NIST)	Companion to the NIST Cloud Computing Reference Architecture; goal to help communicate the offerings and components of cloud architecture

Creating and Refining the Corpora

Building a cognitive application requires extensive collaboration between the technology team and business experts. The initial steps in the development process include defining the objective and user expectations for the application. This stage requires substantial industry or domain expertise. The next series of steps in the application development process relies more heavily on the technology team. The actual creation of the corpus, model development, and training and testing of the system requires skills in areas such as software development, machine learning, and data mining.

The creation of the corpus is not a one-time process. There is an initial effort to build a quality corpus (or corpora) that includes the selected data sources. However, there needs to be continuous re-evaluation of the data sources to determine if new sources need to be added or if enhancements to existing sources are required to improve outcomes from the cognitive application. You need to understand the life cycle for each of the data sources because many of these sources need to be updated at regular intervals. Therefore, you need to set a process in place to ensure that updates to data sources are made on a timely basis.

Although a cognitive application leverages data from the corpora as its primary base of knowledge, not all the data sources used by the system need to be ingested into the corpus. Much of the data may be called as a cloud-based service and used by the application without being included in the corpus. A cognitive application may need to interact with a variety of data management systems including Hadoop, column store, graph, and other environments.

The process of creating the corpus includes preparing the data, ingesting the data, refining the data, and governing that data throughout its life cycle. These steps are described in the following sections.

Preparing the Data

All data ingested into the corpus must first be validated to ensure that it is readable, searchable, and comprehensible. As detailed in the previous section, structured, semistructured, and unstructured data are likely to be combined in various corpora included in a cognitive system. All data sources need to be evaluated to see if any transformations or enhancements are required prior to ingestion into a corpus. Are your text-based resources such as journal articles, textbooks, and research documents annotated with headings that provide queues to the cognitive system? Tagging should help the system identify and classify the content in specific articles. In addition, tagging can ensure that the cognitive system can quickly make appropriate associations between different data elements.

The requirements to transform the structure of the data can vary based on the cognitive platform you use. The corpora of some early cognitive systems were ingested with primarily unstructured text-based content. As a result, complex structured data sources needed to be transformed into unstructured content prior to ingestion into the corpus. Initially, this transformation was time-consuming. However, services have been developed to help speed up the process of transforming data structures. Vendors are continuing to improve data preparation services for cognitive systems, making it possible for structured data to be automatically transformed within the system. These transformation and other data preparation services can have a positive impact on adoption rates for cognitive applications. Data from structured data sources such as Customer Relationship Management Systems or other database applications needs to be easily and quickly ingested into the system if business users are going to begin using the applications at a greater rate. It is not necessary that the complete database be ingested as is. Actually, it is quite common that only a segment of the existing data source is required to meet the requirements for the domain.

Ingesting the Data

Managing the data ingestion process efficiently is critical to the success of a cognitive application. Data ingestion is not something that happens just once

during the development of the system. Existing data sources are subject to continuous updates and refinements to ensure they are accurate and up to date. The results of the training and testing of the models may indicate weak spots or limitations in the corpora that require the addition and revision of sources. In addition, changing user expectations are likely to result in new additions to the corpora. Delays in making the required updates to the corpora will decrease the effectiveness and accuracy of the system. Therefore, to maintain the viability of the cognitive system, data sources may need to be ingested in near real time. Typically, you will have access to a set of services that are designed to make the ingestion process fast, robust, and flexible. Although there may be some coding required, the ingestion services will include connectors and tools to make the process as seamless as possible.

As in traditional data management efforts, you need to have controls and supports in place to maintain governance and anticipate and correct for errors. For example, you must incorporate real-time traceability into the data ingestion process. If errors result in an unexpected halt in the ingestion process, you need to trace back to understand why the problem occurred and where you were in the ingestion process when it stopped. This is called *checkpointing*, and you can then use this information to restart the ingestion process in the right place. In addition, you may need to monitor the ingestion process to ensure that any records that are deleted or scrubbed to meet security requirements have been handled properly.

Refining and Expanding the Corpora

As mentioned in the previous section, a corpus needs to be continually refined to ensure that the cognitive application delivers accurate information and provides the right level of insight. Although you have completed extensive preparation for ingesting content needed to provide a good knowledge base for your cognitive application, it is hard to anticipate all data requirements at the outset.

Early in the training process, you may find that the accuracy of the answer to a certain question is below your accepted threshold. By increasing the coverage (adding more data) for certain topic areas in your domain, you should improve accuracy. Plan for multiple iterations of this process of training, observing results, and then adding to the corpus. You need to establish an ongoing process of updating data requirements and adding to the corpus as you proceed through the testing process and after your application becomes operational. You can use expansion algorithms to determine which additional information would do the best job of filling in gaps and adding nuance to the information sources in the corpus. There will be situations in which you need to enrich data by providing lookups to additional sources that might have detailed information about customers or definitions of technical data.

Governance of Data

The corpora in your cognitive application will include a wide range of data sources. There may be personal data that is subject to the same data privacy rules that apply to data used in other systems in your organization. Therefore, you will need to comply with the same data privacy and security requirements of any system. There will be data that will be ingested into the corpus that might have restrictions on use based on governance requirements. In some situations there might be copyrighted images or content that is part of your corpus. Therefore, you want to make sure that you have a license for use of that content. In healthcare there are patient privacy rules that require that personal information be anonymized. In a retail system it will be important not to expose customers' credit card data. If a corpus includes social media data, you must be sure that you are not violating the privacy of users of those sites. For example, users might decide that they no longer want to allow access to location data. In some countries there are restrictions on where customer data can be stored. A cognitive system may require the highest level of governance and security because over time it will include sensitive data about competitive best practices. Therefore, in designing and operating a cognitive system, governance and security cannot be an afterthought.

Training and Testing

It is through an iterative process of model development, analysis, training, and testing that the cognitive system begins to learn. Deploying a scalable training and testing strategy can ensure your application works as intended when it becomes operational. You need to measure responses to determine what is the minimal level of accuracy that is acceptable. After this is established through the testing process, you can begin to establish the *ground truth*—a set of data that is the gold standard for accuracy of a model. It may require you to try additional data sets so that the information used for testing is objective. Initially, you create a ground truth that establishes what the system knows and understands. In Question-Answer based cognitive applications, you have a set of question-answer pairs that establish your ground truth. The questions represent the types of questions your users will ask. The answers to those questions are accurate, having been approved by domain experts. These question-answer pairs are developed in clusters around a topic to help with the machine learning process. Algorithms help the system to understand context by looking for associations and patterns in the clusters of question-answer pairs. Your training and testing strategy needs to compare new analysis against the ground truth and add to the base level of truth when needed to improve the accuracy of the system.

This is often an iterative process; each time the data is trained, the accuracy of the application improves.

Cognitive systems are designed to learn from failure and improve through feedback. Your cognitive application may assign a high confidence level to answers that are obviously wrong. As part of the training process, you need to analyze why the system got the answer wrong. Although training the system should be an automated process, there are some aspects that involve manual intervention, particularly by subject matter experts. Figure 10-1 illustrates the steps that help to analyze the reason for the error and what corrective action to take to improve accuracy in the future. These errors are measured against key measures for monitoring cognitive system performance including recall, precision, and accuracy.

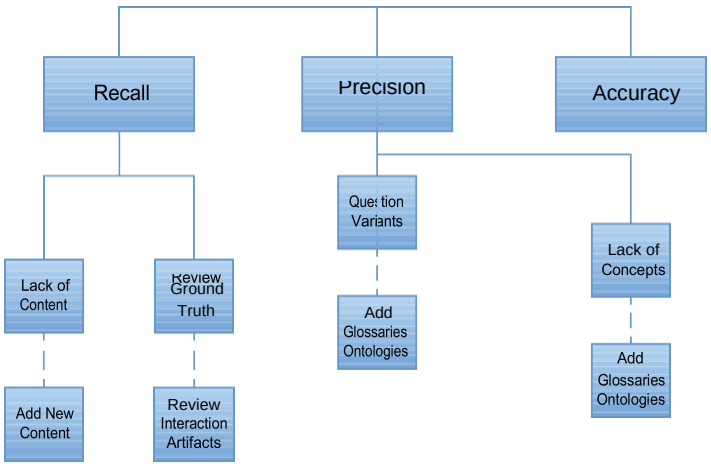

Figure 10-1: Improving accuracy of the models

The section "Creating and Refining the Corpora" earlier in this chapter detailed the importance of adding and updating data to ensure the corpus can support the cognitive application. However, lack of data is not the only reason why an application may provide incorrect answers. Subject matter experts may need to review the ground truth and make adjustments to the answers provided to the system. Other errors may occur because the models cannot capture some of the relationships and nuances between similar data sources. One approach to improve this is by adding glossaries and ontologies that provide the system with cues to learn more about key concepts.

Training and testing data can be one of the most time-consuming parts of the process of creating a cognitive system. The smaller the domain, the easier it will be to both create the corpus and find training data to ensure that the information can answer questions and learn over time. In this situation, you can select a sample data set that is representative of the type of questions and type

of problems you are addressing. If the domain is larger and more complex, it requires a larger set of sample data. In many situations, you can select sample data that is directly applicable to the problem. For example, the data regarding consumer questions about treatment of diabetes are well understood. However, you may have a situation in which there isn't a lot of certainty regarding outcomes. For example, if you want to understand data from traffic management in a large metropolitan area, you might need a vast amount of sensor data. You may not select the right set of data that is representative of the patterns you want to identify. As you can see, training and then testing results can be complicated by scope and scale issues.

The most important part of the training process is to have enough data so that you are in a position to test your hypothesis. Often the first pass at training provides mixed results. This means that you either might need to refine your hypothesis or provide more data. This process is not unlike learning any new discipline where you start with your assumptions based on incomplete knowledge. As you learn more, you can determine that you need more data from more sources. As you gain more insights from the data, your assumptions will change. At this point you are ready to test your understanding of the domain to see if you have the right amount of knowledge or if you are still required to collect more data and learn more. This is precisely what happens in an automated fashion when you design a cognitive system.

Summary

Implementing a cognitive solution is a multistep process that begins with understanding the goals and objectives of the project. These steps begin by establishing your objectives: the domain and key user attributes. You also have to define the type of questions you expect users to ask and what insight they may be looking for. You are also required to determine and find the relevant data sources both from internal and external sources. After these stages are complete, you create and refine the corpora. The final stage is the training and testing process. But keep in mind that this is not a serial process. Building a cognitive system is iterative because data continues to change, and the nature and attributes of users changes. A well-designed cognitive system can become a new model for gaining significant insights into business knowledge.

Chapter 11

Building a Cognitive healthcare application

The healthcare industry is a large and complex ecosystem that encompasses many different types of organizations that support patient wellness and care. The ecosystem is broad, with a number of well-defined roles including:

- Healthcare providers
- Healthcare payers
- Medical device manufacturers
- Pharmaceutical firms
- Independent research labs
- Health information providers
- Government regulatory agencies

Although there have been enormous technological advances that have enabled organizations to improve health outcomes for patients, the need for continued technical innovation is at a tipping point. Each segment of this ecosystem has typically managed healthcare information in a siloed way making it difficult to share patient and medical research data across the various stakeholders. The volume and variety of healthcare data that needs to be managed, analyzed, shared, and secured is growing at a fast pace. Even when participants are motivated to share information for mutual benefit, the required data is often inconsistent and disconnected, which can slow down progress in medical research and lead

to clinical errors that put people's lives at risk. Depending on the methodology used to measure medical mistakes, preventable harm leading to death is either the third leading cause of death in the United States behind heart disease and cancer or the sixth behind accidents and ahead of Alzheimer's disease.

This chapter looks at several healthcare organizations where experts are in the early stages of building cognitive applications that help them to solve well-known healthcare problems in new ways, and begin to solve what was previously intractable. These stakeholders in the healthcare ecosystem are beginning to use cognitive systems to help them find patterns and outliers in data that can help to fast track new treatments, improve efficiencies, and treat patients more effectively.

Foundations of Cognitive Computing for Healthcare

The healthcare ecosystem creates and manages a huge volume of data such as digital images from CT scans and MRIs, reports from medical devices, patient medical records, clinical trial results, and billing records. This data exists in many different formats ranging from manual paper records and spreadsheets to unstructured, structured, and streaming data managed in a variety of systems. Some of these systems are well integrated, but most are not. As a result, the vast amount of data generated and analyzed by the healthcare industry presents significant challenges. However, as organizations find new ways to manage and share this data, they are finding that there are amazing opportunities for improving health outcomes. For example, healthcare providers have implemented electronic medical record (EMR) systems to maintain integrated, consistent, and accurate patient records that can be shared by a medical team. Although the EMR is still a work in progress for many organizations, there are great benefits to having a complete, accurate, and up-to-date set of problems and treatment for each patient. Treatment decisions can be made more confidently and with greater speed if the medical information is available in a consistent and accurate form.

One of the persistent challenges for healthcare organizations is the need to find the patterns and outliers in both structured and unstructured data that can help them improve patient care. As shown in Figure 11-1, data management in the healthcare ecosystem is moving away from document-centric silos to well-integrated knowledge bases that include both structured and unstructured data.

The management of healthcare data will begin to follow a more standards-based approach to facilitate sharing of data where appropriate. Medical devices and sensors have the capability to generate valuable data about a patient's condition, but this data is not always captured effectively. There are great opportunities to improve screening of patients and anticipate changes in their medical condition by using predictive analytical models on the data streams. Cognitive systems

can capture and integrate this new generation of sensor-based data with the entire recorded history of medical research and clinical outcomes captured in natural language text to form a corpus. The system learns from experiences with this corpus, enabling significant improvements in outcomes.

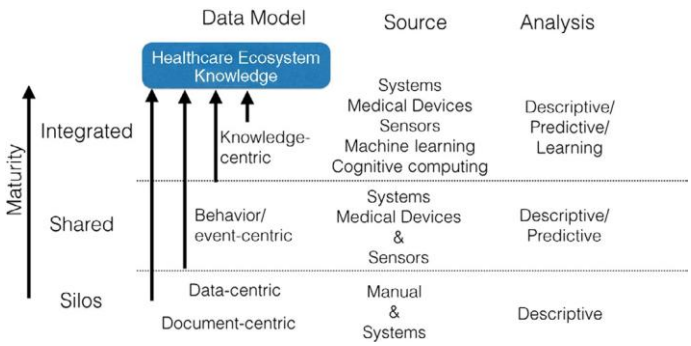

Figure 11-1: Foundations of cognitive computing for healthcare

For example, doctors in the neonatal department of a Toronto hospital developed analytical models that provide 24 hours of advance warning on which babies might develop a life-threatening infection. The infection, late-onset neonatal sepsis, is a blood infection that can occur in subsets of newborn babies. Prior to the analytics research done by Dr. Carolyn McGregor, Canada Research Chair in Health Informatics based at the University of Ontario Institute of Technology, the neonatal intensive care unit relied on monitors that collected data on infants' vital signs but stored only 24 hours of data at a time. By capturing the data from the monitors as an ongoing stream of data, the informatics team developed algorithms to analyze the data over time. The algorithm looks for patterns that occur before the infection becomes clinically apparent. With the new system doctors get a digital reading on respiratory rates, heart rates, blood pressure, and blood oxygen saturation, and can monitor infants' vital signs in real time and detect changes in their conditions.

Constituents in the Healthcare Ecosystem

The healthcare ecosystem has evolved to include a variety of organizations, each of which contributes to the development, financing or delivery of well-ness or treatment information, processes, or products. As shown in Figure 11-2, healthcare providers, payers, pharmaceutical companies, independent research groups, data service providers, and medical manufacturers all have access to different sources of relevant healthcare data. Government agencies and even the patients have a role in managing who sees which data. Some of this data is

shared, but much of it is controlled by regulations and security requirements. The relationships between the constituents in terms of data sharing are complex and in a state of flux. To move toward a more integrated approach to healthcare ecosystem knowledge that includes more predictive analysis and machine learning, there needs to be continued improvement in the consistency of data shared across the ecosystem.

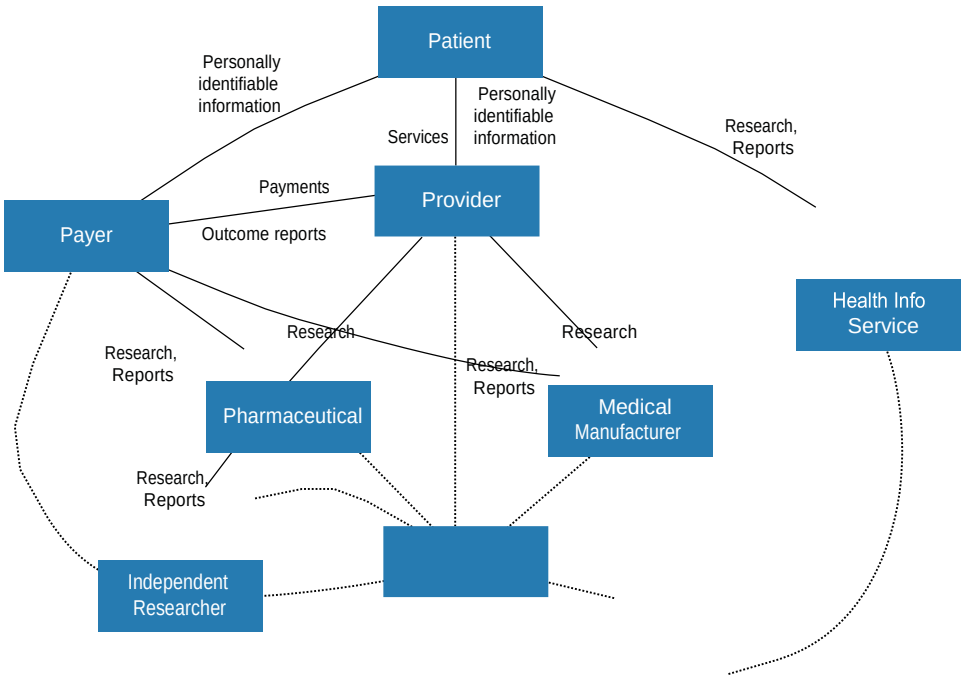

Figure 11-2: Healthcare ecosystems data sources

The data managed and leveraged by different constituents in the healthcare ecosystem includes:

- **Patients**—From family history and habits to test results, individuals participating in the healthcare ecosystem produce personally identifiable information, which may be aggregated anonymously, where permitted, to guide care for those with similar attributes.
- **Providers**—Data covers a broad range of unstructured and structured sources. Some examples include patient medical records (EMR, doctors' office notes, and lab data), data from sensors and medical devices, intake records from the hospital, medical text books, journal articles, clinical research studies, regulatory reports, billing data, and operational expense data.
- **Pharmaceutical companies**—Data to support pharmaceutical research, clinical trials, drug effectiveness, competitive data, and drug prescriptions by medical providers.

- **Payers**—Data includes billing data and utilization review data.
- **Government agencies**—Regulatory data.
- **Data service providers**—Prescription drug usage and effectiveness data, healthcare terminology taxonomies, and software solutions to analyze healthcare data.

Learning from Patterns in Healthcare Data

The benefit of cognitive computing is that healthcare professionals will more easily get the insights they need from all types of data and content to act with confidence and optimize their decision making. The risks of not finding the right relationships and patterns in the data are high in the healthcare industry. In this industry, if important pieces of information are overlooked or misunderstood, patients can suffer long-term harm or even death. By combining technologies such as machine learning, artificial intelligence, and natural language processing, cognitive computing can help healthcare professionals to learn from patterns and relationships discovered in data. The collaboration between human and machine that is inherent in a cognitive system supports a best practices approach that enables healthcare organizations to gain more value from data and solve complex problems.

Gaining more value from data is a multifaceted process that requires both technology and human knowledge. Getting the data right is paramount. The relevant data needs to be accurate, trusted, consistent, and available for access expeditiously. However, having accurate data is only the baseline for improving health outcomes for patients. Physicians need the skill and experience to make sense out of what is often a complex set of symptoms and diagnostic tests. They need to internalize best practices that enable them to ask the right questions and listen for answers from the patient. The solution to a patient's problem is not always obvious in the medical lab results and images. Best practices that focus on connecting all the disparate data points can help physicians, researchers, and others in the healthcare ecosystem to find the right solution.

Learning from patterns in data helps healthcare organizations to solve some of their most challenging problems. For example, The University of Iowa Hospitals and Clinics has identified patterns in a population of surgical patients that help to improve both quality and performance in surgery. The hospital has modeled data from hospital readmission, surgical site infections, and other hospital-acquired infections. The model enables physicians to predict which patients are most at risk for acquiring a surgical site infection while they are still in the operating room and corrective actions can be taken.

Other hospitals use predictive models to reduce costly and dangerous hospital readmission rates as well. The patterns identified from thousands

of hospital records are used to build a model that can analyze a patient's medical record to identify risk factors for problems that may occur after discharge from the hospital. Predictive analytics models look at a number of different factors to determine which ones have the greatest impact on hospital readmission rates. As shown in Table 11-1, these factors may be specific to the patient or the physician.

Table 11-1: Attributes to Consider for a Predictive Model on Hospital Readmissions

Patient Attributes	Smoker, drug abuse, alcohol abuse, lives alone, dietary noncompliance
Socio-economic Attributes	Educational status, financial status
Physician Factors	Incorrect medicines given, overlooked important information about patient

Understanding the risk factors can help hospitals to improve processes within the hospital and take corrective measures to decrease hospital readmission rates. The predictive model can help on a case-by case basis by indicating which patients may require more intensive follow up after they are discharged.

Building on a Foundation of Big Data Analytics

Although there is a great deal of interest and some exciting case examples of cognitive systems in healthcare, these implementations are at an early stage. However, healthcare organizations are not starting from scratch for cognitive computing and big data analytics. There are some high-profile examples of analyzing data and incorporating machine learning in medical environments. The next generation of these healthcare platforms are building on a strong foundation of big data analytics. As healthcare informatics capabilities mature to incorporate cognitive systems, the overall goals for the healthcare organization remain the same. There is a common focus on providing optimal high-quality care to patients and to continually improve healthcare options and outcomes in a cost-effective manner.

Much of the effort in healthcare IT has been focused on developing more integrated systems so that medical information can be safely stored and accessed as needed for research and patient care. For example, healthcare providers have implemented electronic medical records (EMR) to help provide a unified record of medical data for each patient. Much of the patient-related data is unstructured, and large volumes of this data come from digital images, lab tests, pathology reports, and physician reports. As described in the previous section, healthcare organizations are rapidly finding new ways to gain value from this data. In

addition to using EMR and other patient-specific data to make decisions for one patient, there is great value in leveraging data across large groups of patients to build predictive models that can improve outcomes for large populations of patients. These analytics efforts need to ensure that requirements for security and privacy are met by removing any personal identifying information from the data.

One healthcare field where big data analytics is rapidly increasing the speed at which new research can be completed is the biopharmaceutical field. Revolutionary advances in DNA sequencing technology makes it possible to collect huge volumes of genomic information for analysis. To keep up the pace of the research, technology is used for sequencing data storage, processing, and downstream analytics. There are huge demands for new computational approaches to store and analyze genomic data. Advanced algorithms, methods, and tools enable scientists to effectively understand the data produced by genomic analyses, and to help answer important biological questions. Advanced modeling efforts are replacing many of the more manual efforts used in the past to analyze genomic data.

Cognitive Applications across the Healthcare Ecosystem

Many healthcare experts are building on what they have already achieved in big data and analytics initiatives to incorporate machine learning and cognitive computing. The goal is to continue to optimize results in healthcare research and clinical diagnosis and treatment. Gains in speed, innovation, and the quality of outcomes are dependent on how humans interact with the available technology and data. In addition, there is an exceedingly strong requirement within healthcare organizations for those humans with the most experience to put best practices in action and to share those best practices with the next generation of healthcare professionals. This transfer of knowledge takes place continuously through training programs for medical students and residents as well as assistant and mentoring programs in research labs. Introducing a cognitive system to support healthcare professionals as they learn can help drive this process of knowledge transfer forward. Right now the use of cognitive computing is at an early stage; however, the expectation is that over the next decade it will become well integrated into many healthcare processes.

Two Different Approaches to Emerging Cognitive Healthcare Applications

The implementations of cognitive healthcare applications are proceeding along two different paths: customer or user engagement applications and discovery applications. Customer engagement applications are designed to help find

personalized answers to questions. For example, several emerging companies have developed cognitive applications that provide consumers with answers to questions about managing their own health and wellness. Other cognitive systems provide support for healthcare payer customer service agents. With a corpus that contains more relevant information than people could possibly consume and retain, these systems answer relevant questions and provide new insight about their health. Discovery applications are used in situations such as drug discovery or to discover the optimal treatment for a patient. In both types of cases, the healthcare organization needs to begin by defining the end user of the system, the types of questions that will be asked, and the content that is required to build the knowledge base for the system. The cognitive system is used to understand relationships and discover patterns in data that may lead to improved healthcare outcomes.

You need to understand the types of users who access your cognitive healthcare application. What is the medical background and expertise of your users? For example, will the users be medical students, or medical clinicians with many years of experience? Or will your users be health and wellness consumers? Expectations for user/system interaction will have an impact on the development of the corpus, the design of the user interface, and how the system is trained. The user type also has an impact on the confidence levels required and level of accuracy the system needs to achieve. As user requirements and expectations change over time, these changes must be incorporated into the ongoing development of the cognitive system. The learning process for a cognitive system is continuous, and as a result the system gets smarter and delivers greater value to end users the more it is used.

The Role of Healthcare Ontologies in a Cognitive Application

Healthcare taxonomies and ontologies—a coding system or semantic network of medical terms and the relationships among these terms—are important to the development of a corpus for cognitive healthcare applications. These ontologies are used to map the relationships between terms with similar meanings. There are many ontologies that are already widely used in healthcare to organize terminology related to medical conditions, medical treatments, diagnostic tests, ingredients and dosing for clinical drugs, and drug complications. One example of a medical ontology is the International Classification of Diseases (ICD). The ICD-10 is the current version as endorsed by the World Health Organization. However, it has not yet become the standard in all countries. ICD-10 will become the standard in the United States some time after October 1, 2015. The ICD includes codes for diseases, disease symptoms, and medical findings about diseases. The ICD is only one of many different taxonomies and ontologies in use across the ecosystem. To build an efficient corpus for a healthcare application, you need to find a common language to

ensure that data from different sources can be integrated and shared. Without a taxonomy of terms, the cognitive system cannot learn as quickly, and the accuracy of results will be insufficient. Your systems will miss a lot of terms that have the same meaning.

Healthline Corporation has developed one of the largest semantic taxonomies for the healthcare ecosystem. It maps the relationships between consumer and clinical applications, which can help to support new consumer-focused cognitive health applications. Algorithms can reference the taxonomy to improve the semantic understanding of a query to the cognitive system. In addition, cognitive health applications can make more accurate associations between medical concepts by referencing a comprehensive and accurate ontology or taxonomy.

Starting with a Cognitive Application for Healthcare

Early stage examples of cognitive applications in healthcare are built on top of the cognitive engine or platform. To develop an application you need to begin by defining your target end user and then train the cognitive system to meet the needs of your user base. What is the general subject area for your cognitive application? What do you know about your users' level of knowl- edge in this area, and what are their expectations or requirements from the cognitive application?

A cognitive system needs to start with a base level of information from which it can begin to find the linkages and patterns that can help it to learn. Although the learning process begins with questions, a trained system can do much more than provide answers to a set of questions. The cognitive system can make associations between questions, answers, and content to help the user understand the subject matter at a deeper level. The basic steps required to build a cognitive application in healthcare follow.

Define the Questions Users will Ask

You want to begin by assembling the types of questions that will be asked by a representative group of users. After this step is completed, you can assemble the knowledge base required to answer the questions and train the system effectively. Although you may be tempted to begin by reviewing data sources so that you can build your knowledge base or corpus for your system, best practices indicate that you need to take a step back and define your overall application strategy. The risk to beginning with your corpus is that you are likely to target your questions to the sources you have assembled. If you begin with the corpus, you may find you cannot meet the needs of your end users when you move to an operational state.

These initial questions need to represent the various types of questions users will ask. What do users want to ask and how will they ask questions? Are you building a consumer-focused application that will be used by a general population of users, or are you developing a system that will be used by technical experts? Getting the questions right is critical to the future performance of the application. You need to seed the cognitive system with a sufficient number of question and answer pairs to start the machine learning process. Typically, 1000–2000 question/answer pairs seem to be the right number to get the process started. Although the questions need to be in the voice of the end user of the system, the answers need to be determined by subject matter experts.

Ingest Content to Create the Corpus

The corpus provides the base of knowledge used by the cognitive application to answer questions and provide responses to queries. All the documents the cognitive application needs to access will be included in the corpus. The question/answer pairs you have created help to drive the process of collecting the content. By beginning with the questions, you have a better idea of the content that will be required to build the corpus. What content do you need to answer the questions accurately? You need to identify the resources you have and which resources you may need to acquire to provide the right knowledge base. Examples of content include medical texts, background information on health subjects such as pharmaceutical research, clinical studies, and nutrition, medical journal articles, patient records, and ontologies and taxonomies.

The content you select needs to be validated to ensure that it is readable and comprehensible. Adding meta tags to your content can help with creating associations between documents. For example, you can use tagging to identify that an article pertains to a specific medical condition such as diabetes. In addition, content should have sections and headings to provide cues to the cognitive system. You may need to optimize the format of some of the source data to ensure that it can be properly identified and searched. For example, structured data sources such as a comprehensive nutrition table may need to be transformed into unstructured content prior to ingestion into the corpus. Simple tables can be read by the cognitive system, but more complex and nested tables should be transformed to unstructured text for clarity. The source transformation process is required to ensure that the corpus functions properly.

You need to understand the life cycle of documents you ingest to plan for appropriately scheduled updates. In addition, you may need to establish a process that will ensure you are notified of new and updated content. The corpus needs to be updated continuously throughout the life of the application to make sure it continues to be viable.

Training the Cognitive System

How does the training process begin? The cognitive system learns through analysis and training (refer to Chapter 1, "The Foundation of Cognitive Computing," for discussion on different types of machine learning and Chapter 3, "Natural Language Processing in Support of a Cognitive System," for details on natural language processing). Just think of how you might approach learning a new subject. Initially you may have a long list of questions. You do some reading and then your questions change in content and scope as you learn more about the subject. The more you read and understand, the fewer questions you have. A cognitive system is similar in that the more question/answer pairs that are analyzed, the more the system learns and understands.

Analyzing the question/answer pairs is a key part of the overall training process. Although it is important for representative users to generate the questions, experts need to generate the answers and finalize the question/answer pairs. The questions need to be consistent with the level of knowledge of the end user. However, the experts need to ensure that the answers are accurate and in line with the content in the corpus. As shown in Table 11-2, you are likely to have some overlapping questions or clusters of questions. These questions may ask about a similar topic using slightly different terms or from a different perspective. Or the questions may be basically the same except one version of the question abbreviates certain terms. The cognitive system learns from these clusters of questions.

Table 11-2: Questions Used to Train a Cognitive Application on Health and Wellness

Question 1	What is the difference between whole milk and skim milk?
Question 2	Is low fat milk different from whole milk?
Question 3	Is skim milk better than whole milk?

Question Enrichment and Adding to the Corpus

The training process for your cognitive application is used to ensure your application works as intended when it becomes operational. Initially, it needs to be repeated multiple times using training data, test data, and blind test data. As each of these tests is completed, you can add content to the corpus to cover areas in which there is inadequate information.

Plan to continually return to the training process after your application goes live so that you establish an ongoing process of updating question/answer pairs and adding to the corpus. Expansion algorithms are used to determine which additional information would do the best job of filling in gaps and adding nuance to the information sources in the corpus.

Using Cognitive Applications to Improve Health and Wellness

The patient (or healthcare consumer) is central to the healthcare ecosystem (refer to Figure 11-2). This complex ecosystem generates an enormous amount of data that describes the health and well-being of every individual in the system. Many organizations that manage a population of healthcare consumers have implemented various programs to help improve the group's overall health. The challenge is that these programs do not always provide the personalized responses and incentives that their members need to change behavior and optimize health outcomes. The payback of helping individuals to lose weight, increase exercise, eat a well-balanced diet, stop smoking, and make healthy choices overall is huge. Healthcare payers, governments, and organizations all benefit if communities as a whole are healthier and individuals do a better job of managing previously diagnosed conditions. The following list (developed by the Office of the Surgeon General (U.S.); Office of Disease Prevention and Health Promotion (U.S.); Centers for Disease Control and Prevention (U.S.); National Institutes of Health (U.S.); and the Rockville (MD, U.S.) Office of the Surgeon General (U.S.); 2001) shows the many different medical conditions and diseases increased weight is associated with. Even when faced with these facts, it is incredibly hard for many people to make the positive change they need. These conditions and diseases include:

- Premature death
- Type 2 diabetes
- Heart disease
- Stroke
- Hypertension
- Gallbladder disease
- Osteoarthritis
- Sleep apnea
- Asthma and other breathing problems
- Certain types of cancer
- High cholesterol

Finding ways to improve the connections and communication of individuals and the healthcare ecosystem is a priority for a number of emerging companies. Several of these companies are highlighted in the following sections.

Welltok

Welltok, based in Denver, CO, provides personalized information and social support to help individuals optimize their health with its CaféWell Health Optimization Platform. Welltok works with population health managers, such as health payers, to help decrease healthcare costs by providing a platform that gives people the support, education, and incentives (e.g., gift cards, premium reductions) they need to change their behavior and improve their health.

Overview of Welltok's Solution

Welltok's CaféWell Concierge is a platform designed to help individuals optimize their health by connecting them with the right resources and programs. It organizes the growing spectrum of health and condition management programs and resources, such as tracking devices, apps, and communities, and creates personalized, adaptive plans for each consumer.

Welltok partnered with IBM Watson to create the CaféWell Concierge app, which leverages cognitive technologies to dialogue with consumers and provide personalized guidance to optimize their health. Vast amounts of internal and external data sources are used to build corpora that form the knowledge base of the system. CaféWell Concierge uses natural language processing, machine learning, and analytics to provide personalized and accurate recommendations and answers to questions asked by individual health consumers.

As a mobile application, health consumers can engage with the CaféWell Concierge at a time and place that is convenient for them. Each individual receives an Intelligent Health Itinerary based on their health benefits, health status, preferences, interests, demographics, and other factors. The itinerary is a personalized action plan with resources, activities, health content, and condition management programs. For example, consumers with controllable health conditions like diabetes or asthma will receive an Intelligent Health Itinerary with educational information and guidelines tailored to them to help them make healthy choices on a daily basis.

Welltok's partners include health payers that make the app available for free to their members. The health payers typically offer incentives or rewards such as entry in a drawing for a gift card for completing a coaching session, or a reduction in health costs for improving your body mass index (BMI). CaféWell uses advanced analytics algorithms to align actions and behaviors with the right incentives and rewards to motivate consumers to get involved in their health. It also learns over time what individuals respond to and what type of incentives or value to offer for targeted behaviors.

Using natural language processing, consumers can dialogue with the application and ask questions related to health and wellness. Welltok followed the steps

in the previous section to build a cognitive application that can handle mass personalization, process large volumes of information, and answer open-ended questions in seconds. The architecture and data flow for the question-answer training process for CaféWell is illustrated in Figure 11-3.

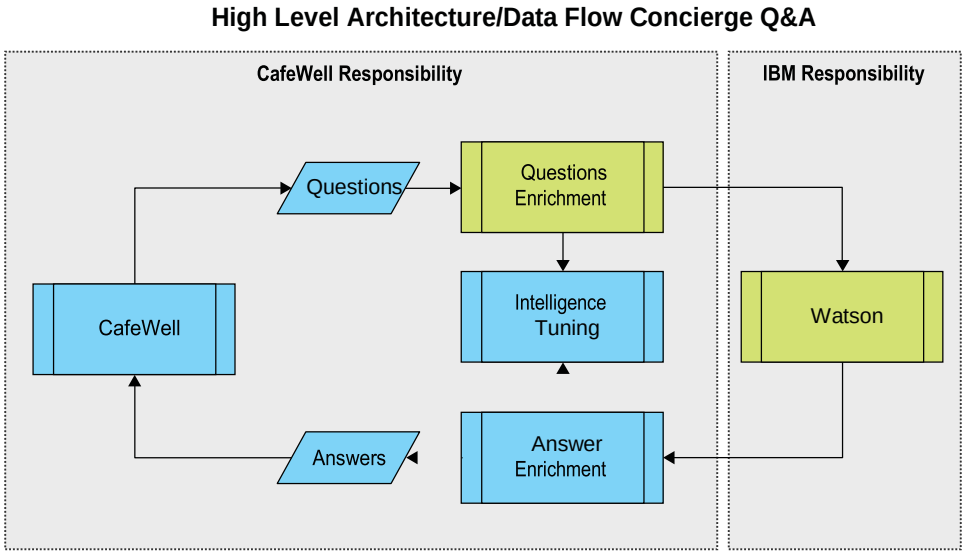

Figure 11-3: Welltok training architecture

To develop its question/answer pairs, Welltok collected input from consumers to create questions that would reflect their interests and used subject matter experts to answer the questions logically and accurately. Table 11-3 shows a sample of the thousands of question/answer pairs that Welltok created to begin the training process for CaféWell Concierge. After determining an initial set of question/answer pairs, Welltok developed the corpora (and ontologies) for the application to provide Watson with access to the information sources it needs. Welltok collected unstructured information from third-party healthcare sources to get all the information required for the corpora.

Welltok worked closely with IBM to train Watson for CaféWell Concierge. The iterative process of ingesting data to build the corpus, enrich content, and improve the intelligence of the cognitive system is illustrated in Figure 11-4. By leveraging Watson's cognitive capabilities, CaféWell can understand context and learn about a user's health concerns, goals, and preferences. Watson's machine learning capabilities enables CaféWell to continuously improve the quality of its responses and recommendations. Watson has dozens of different corpora covering many different aspects of health and wellness, including health insurance benefits, nutrition, and fitness. These corpora, in addition to

information about the individual, are used to support the advanced analytics algorithms that deliver the personalized recommendations and responses. The application goes beyond providing search results. It builds a relationship with the users—getting to know them, providing personalized recommendations and guiding them to optimal health.

Table 11-3: Sample of Welltok Question/Answer Pairs

What are some lifestyle changes that I should make if I have high blood pressure?	Lifestyle changes are just as important as taking medications. Reducing your weight by just 10 pounds may be enough to lower your blood pressure. Losing weight can help to enhance the effects of high blood pressure medication and may also reduce other risk factors, such as diabetes and high bad cholesterol.
How do you determine the calories burned by your body?	BMR is often calculated using the Harris–Benedict equation. This equation calculates basal metabolic rate based off of 3 variables: weight, height, and age. Using this approach, total energy expenditure can be calculated by multiplying BMR by an activity factor. **equation For Men:** BMR = 88.362 + (13.397 × weight in kg) + (4.799 × height in cm) – (5.677 × age in years)
Do my nutritional needs vary throughout life?	Nutritional needs vary throughout life. From infancy through adulthood, good nutrition is essential to growth and development, and to maintaining health in the later years.
Why should I read the food label on packaged foods?	Most packaged foods have a label listing nutrition facts and an ingredient list. In the U.S., the Food and Drug Administration (FDA) oversees the requirements and design of the Nutrition Facts label. The purpose of the label is to help consumers make quick, informed food choices that contribute to a healthy diet. Especially on a low-sodium diet, you need to look at the food label to limit sodium intake.
I have a grain allergy, what food should I avoid? What kinds of food are considered grains?	Any food made from wheat, rice, oats, cornmeal, barley, or another cereal grain is a grain product. Bread, pasta, oatmeal, breakfast cereals, tortillas, and grits are examples of grain products. There are whole grains, containing the grain kernel, and refined grains, which have been milled to remove bran and germ. There are many benefits to a diet rich in grains.

Using Watson's machine learning capabilities, CaféWell Concierge improves the quality of responses users receive with each interaction. And with its spatial and temporal capabilities, the application factors in time and location to provide highly relevant information. For example, it can recommend where and what to eat for lunch based on your location and your specific diet and nutritional requirements.

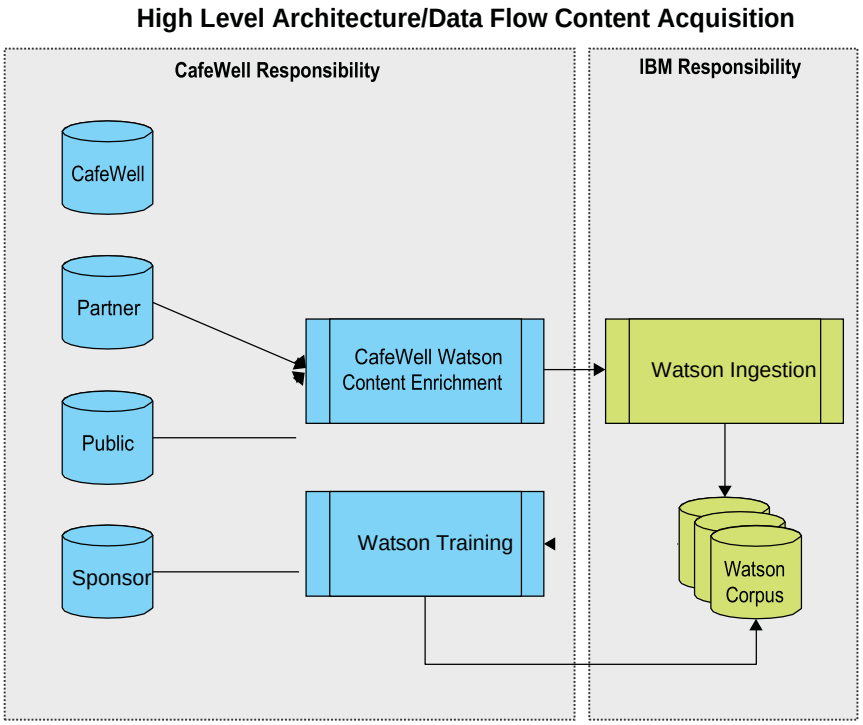

Figure 11-4: Welltok high-level architecture and data flow: data flow content acquisition

CaféWell Concierge in Action

CaféWell Concierge is intended to help individuals understand their health status and receive personalized guidance to help them achieve wanted health and wellness goals, and get rewarded for doing so. The following example shows how an individual with a new medical diagnosis could benefit from interacting with CafeWell Concierge.

Assume you just received a new diagnosis of pre-diabetes from your doctor. You saw your internist in his office last week and after examining you he took some tests. Today, you received a follow-up phone call with your diagnosis and a recommendation to change your diet, lose 20 pounds, and increase your exercise. However, you travel a lot for work leading you to eat a lot of meals at restaurants, and you never have enough time to get to the gym. What do you do next?

Without cognitive computing you might search the Internet for Type 2 diabetes and find information that leaves you confused and scared. Although there are many applications that allow you to search for nutritional information and monitor your weight and exercise activity, they only provide general information about pre-diabetes. A cognitive application can provide you with deeper

insight and more personalized high-quality support. CaféWell Concierge creates an Intelligent Health Itinerary for you including programs and resources such as a video coaching session on nutrition, food choices at local restaurants, a fitness tracking device with step goals to help reduce your BMI, and a social community for additional support.

Based on Watson's cognitive capabilities, CaféWell Concierge can integrate and analyze across multiple sources of information so that you can receive tailored and continuous support as you might from a personal concierge.

GenieMD

GenieMD is also focused on providing a cognitive health application for consumers. The company's mission is to help clients develop more meaningful conversations with their healthcare providers. The overall goal is to help health consumers take a more active role in managing their own health and the health of their loved ones. Users can ask questions in natural language and receive personalized responses and recommendations. These users can access GenieMD through a mobile application. The expectation is that patients will achieve improved health outcomes and that healthcare costs will be lowered. GenieMD aggregates medical information from a range of disconnected sources and makes that information actionable. GenieMD, which is powered by IBM's Watson, is following a development process that is similar to Welltok.

Consumer Health Data Platforms

Google, Apple, and Samsung are all developing consumer focused heath data platforms. These platforms are in the early stage, and the type and variety of data being collected is narrower than the applications discussed in the preceding sections. Google has a set of Google Fit APIs that it provides to developers to help them manage and combine different types of health data. At this point, the health data collected typically comes from wearable devices such as the FitBit, Nike Fuel Band, and other medical sensors that can detect biometric data. This data includes heart rate, steps taken, and blood sugar level. Nike FuelBand can publish user heath data that it collects to the Google Fit platform.

Using a Cognitive Application to Enhance the Electronic Medical Record

The electronic medical record (EMR) is a digital record of the medical and clinical data for each patient followed by a provider (independent physician or large medical centers with all physicians in one group). Typically, the EMR is

designed to store and retrieve data about a patient that can be used for diagnosis and treatment. It has some basic reporting capabilities such as flagging a lab test as low or high based on predetermined criteria. The EMR can be thought of as having three main functions: Think, Document, and Act. Today, the EMR documents information about a patient and supports the physician's ability to take actions on a patient's behalf. But, the EMR does not help with the "thinking" aspect of determining how to best deliver care to a patient. By incorporating machine learning, analytics, and cognitive capabilities into the EMR, physicians could be guided in understanding how a diagnosis was arrived at and the issues surrounding a treatment plan. Overall, healthcare organizations would like to gain more value from the EMR including finding ways to leverage the information in the EMR to improve coordination between different providers and providing more individualized high-quality care for patients.

Epic Systems, a healthcare software company providing EMR software, holds the medical records of approximately 50 percent of the patients in the United States. The company partnered with IBM to add a content analysis capability to the EMR. This enables physicians to use text-based information on a patient as part of the electronic medical record. IBM's natural language processing software, IBM Content Analytics, enables physicians to extract insights from the unstructured text in real time. The EMR can be used with a cognitive system that enables physicians to get answers to complex questions about patient diagnosis and treatment. The information stored in the EMR could be either incorporated into the corpus of the cognitive system or used as part of an analytics engine that is integrated with the cognitive system. Epic's approach provides for analysis of the physician's text-based notes on a patient and transforms these notes into a format that can be incorporated into the patient record. By automatically applying industry-standard diagnosis and treatment codes, significant improvements in accuracy and efficiency can be achieved.

Hitachi is working on a number of consulting projects with healthcare organizations related to enhancing the business value achieved from the EMR. In one project, Hitachi is working with a hospital and EMR vendor to help determine if the treatment plan selected for the patient is the best and most cost effective. Hitachi provides a clinical repository with an analytics engine and a database extractor tool. The focus is on gaining value from the unstructured content.

The Cleveland Clinic is working with IBM's Watson on a cognitive healthcare system focused on rethinking the capabilities of the EMR. How can the EMR be more accurate and used to help physicians learn about the thought process behind clinical decisions? Dr. Martin Harris, Cleveland Clinic, explained how important it is to create one unified and accurate problem list for all patients. One patient may see four specialists for different medical conditions, but for the benefit of the patient, there must be one problem list that includes all of that patient's medical information. Any omission from the EMR could lead to

incomplete knowledge of the patient's conditions, and the impact to the patient can be dangerous.

Although there is some risk that the EMR may be missing an important piece of information, there is typically a lot of information to review in each patient's record. Nothing is deleted in the EMR, and it can be hard to find the information you are looking for. If a patient has a complicated medical condition, the EMR can easily reach 200 pages or more. Given the volume of information in a patient's EMR, some physicians find it more cumbersome to use than the old paper records.

The Cleveland Clinic is building a comprehensive knowledge base using IBM's Watson that can be used to test for omissions and improve the accuracy of its EMR. The Cleveland Clinic ingested information from the EMR into the corpus of the cognitive system along with unstructured data including physician and hospital admission notes. When the unstructured data is compared to the problem list from the EMR, all too often omissions are identified. Using the cognitive system to ask questions would enable the hospital to make sure they are retrieving all information about a patient when needed for analysis. The goal of this project is to develop an EMR assistant that would provide a visual summary of a patient's condition. Users could type in keywords and receive visualizations that would help research a patient's medical history and improve decision making.

Using a Cognitive Application to Improve Clinical Teaching

The most senior and experienced physicians on the staff of a medical center are responsible for transferring knowledge about clinical diagnosis and treatment to medical students and residents. In addition, senior clinicians and researchers at large teaching medical centers want to share knowledge with staff at smaller community hospitals. Research in many areas such as cancer are advancing so quickly that experts at some of the largest medical teaching centers say that it can take years before the information on the newest treatments is translated to changes in treatment offered at community hospitals. In medicine, one is always a student. Each subspecialty has major conferences across the world where papers on the latest research are presented and medical knowledge is shared. In addition, physicians read journal articles to keep current with new research. One service used by many physicians is UpToDate, a clinical decision support resource that provides edited summaries of recent medical information in addition to evidence-based recommendations for treatment. Even with all these resources, it is extremely challenging to keep up with all the new discoveries in drugs and treatment options.

The training of the next generation of physicians is of the utmost importance to members of a senior medical team. Physician leaders at several top medical institutions are developing cognitive systems that may add new dimensions to the complex task of transferring knowledge about medical best practices and diagnostic skills. The expectation is that these new cognitive systems would be in addition to the traditional methods of personalized instruction followed in teaching medical centers. Physicians who train side by side with senior experts in the field learn lessons that they carry with them throughout their careers. A senior neurologist at a Boston teaching hospital described his role as one of "modeling the behavior that students need to follow in treating patients." A large team of medical students and residents accompany him as he makes rounds in the hospital. Students need to be exposed to a variety of diseases in each subspecialty and learn how to identify the differential diagnosis based on symptoms. However, his teaching goes much deeper than understanding the symptoms and treatments of diseases. He wants the students and residents to learn what questions to ask and how to ask them to get the information they need to deliver optimal care.

The Cleveland Clinic is working with IBM to develop a cognitive system called Watson Paths that will provide additional knowledge to students to support what they learn on their subspecialty rotations. Typically, students rotate through a series of subspecialty rotations, spending one month or more in each rotation. Students' clinical experience varies depending on the cases in the hospital during the time they are in the subspecialty unit in the hospital. Cognitive systems that have been well trained in how to treat a broad spectrum of diseases can change the way medical students are taught.

If a medical student is exposed to the cognitive system before the rotation, then the whole training process can become more powerful and can go deeper. If the student has a better understanding of the process of making a diagnosis on some of the most common conditions, the attending physician can focus more on the less obvious diagnosis. The focus needs to be on having as much information as possible to make a correct diagnosis. Considering that there are approximately 13,000 diagnosis codes in the ICD-9 and more than 68,000 diagnosis codes in the ICD-10, students have a lot to learn. A good internist may know approximately 600 medical diagnoses, whereas a subspecialist may have deep knowledge in 60 diagnoses. Fortunately, a cognitive system can ingest information on a huge scale. Cognitive systems (after training) can produce scenarios for the top 600 diagnoses and provide guidance to medical students to help them learn by showing the step-by-step approach in making a diagnosis. As cognitive systems can keep track of the evidence used to support their hypotheses and conclusions, they can justify the resulting confidence level that the diagnosis is right.

Students will be able to interact with Watson Paths to study different approaches to treating patients with a certain set of problems. The students can interact with

a system that offers reference graphs and probabilities of outcomes depending on the treatment approach the doctor and patient decide to follow. Watson Paths will focus on evidence-based learning: validating and calibrating the impact of treatment options that are selected. The system will annotate each decision—helping students to learn the impact of their decisions. As a result of its machine learning capabilities, the more people who interact with Watson paths, the better it gets in accuracy and understanding.

Memorial Sloan Kettering (MSK) is also working with IBM to develop a medical cognitive system powered by Watson. MSK is one of the top cancer research and treatment centers in the world, and its physician leaders are concerned about the length of time it takes for new research to reach the thousands of medical and surgical oncologists that are not based at one of the large cancer centers. MSK identifies the sharing of medical knowledge about cancer diagnosis and treatment as an important part of its mission. The medical center has more than 30 physicians working on the initiative to ingest data from a huge patient database and train Watson.

There is often more than one approach that will work to treat a particular cancer patient. MSK is helping to train Watson so that a physician can get help in assessing the potential outcomes of using one approach versus another. The expectation is that the oncology cognitive system will help to increase the speed at which new treatments can be disseminated. Watson will help with suggestions and support the physician who needs to make decisions about the best approach for his patient.

Summary

We stated previously in this chapter that it is early for cognitive healthcare applications. It is not easy to project how quickly these applications will evolve and become more integrated with operations across the healthcare ecosystem. However, the significant partnerships that are rapidly forming between healthcare experts and technology leaders in cognitive computing suggest a rapid increase in the pace of development. There are many reasons for the increasing investment of time and money in the development of cognitive healthcare applications. However, the most powerful driver for developing cognitive healthcare applications stems from the challenge of gaining insight from the large and rapidly growing volumes of structured and unstructured data managed by the healthcare ecosystem. There is an overabundance of data generated by the healthcare ecosystem that is not well integrated and not easily shared.

Many of the initial healthcare cognitive computing efforts are focused on how patients engage with their own data. Welltok's Café Well Concierge and GenieMD are excellent examples of this type of application. These applications focus on how the patient communicates with healthcare providers and gains

access to information about their medical conditions in a meaningful way. These are practical applications that can help a health consumer adjust his priorities for diet and exercise to improve his overall health. On the other end of the spectrum, there are some interesting applications focused on what can be learned from medical best practices. Clinicians and researchers make decisions that impact people's lives on a daily basis. All too often these decisions are made without the knowledge that comes from a comprehensive understanding of best practices. The goal of many of these emerging cognitive health applications is to ensure that all physicians have the opportunity to evaluate their clinical diagnosis and treatment options in collaboration with a well-trained cognitive system.

Chapter
12

Smarter Cities: Cognitive Computing in Government

One of the great challenges of the 21st century is how to leverage technology to solve a variety of problems that accompany the global trend toward urbanization. In cities everywhere, increasing population density strains physical systems and resources. Individual systems have been developed to collect and manage data for each of functional unit. When critical information cannot be shared across critical services, managers often cannot anticipate safety issues or opportunities to optimize services.

The promise of cognitive computing is to enable metropolitan areas to take advantage of data to evolve and become smarter, and deal with expected and unanticipated events effectively. The objective, therefore, is to learn from experience and patterns of data to improve the way cities function over time. This chapter reviews the problems confronting cities and demonstrates how cognitive computing has the potential to transform the way cities operate.

How Cities Have Operated

A city is more than the roads, buildings, bridges, parks, and even people found within its borders. Cities around the world have evolved in a similar way for centuries—agencies are created to provide services to the population in response to changing conditions and technologies. These agencies justify their existence based on their ability to collect the right data and manage that data to support their

constituents. For example, population density made the rapid spread of disease a public health issue and created a need for public health data tracking. New modes of transportation such as cars and planes required new transportation management departments that required the collection even more data.

Throughout history, these agencies produced paper records, which dominated the way cities managed processes. The problems with paper records are obvious: they are expensive to store, inefficient to retrieve, and subject to damage or loss from water, fire, or even rodents. Even as paper documents could be scanned via optical character recognition (OCR) to make searching text easier, it did not solve the problem. There was still no way to gain insights into the history, meaning, and context of these documents. The basic issue is that documents contain deep structure, which contains virtual knowledge that is not explicitly captured.

When all data was created manually and managed in the form of paper documents, it was difficult or impossible to recognize the relationships and dependencies between these systems. For example, relationships between education and hygiene, hygiene and disease, crime and poverty, are all obvious to us today. However, without a way to analyze data from different agencies or departments to see these patterns, they depended on insight and hypotheses that drove the quest for supporting data. As cities grew with data in departmental silos, it became increasingly difficult to look at data across departments in order to set budget priorities. There was no systematic way to put the pieces together and learn from experience. The real value in the information resided in the brains of these people.

Improved technology provided more efficient ways to collect, manage, and analyze data, the ability to understand and describe scenarios and predict outcomes improved dramatically. In the last few decades there have been significance advances in data-oriented city management. To support changing needs has required the movement from simply managing data more efficiently in databases to adopting analysis tools to make better decisions based on this data.

As shown in Figure 12-1, data management is on a path from document-centric silos to integrated repositories of standards-based structured and unstructured data. From manual systems to a sensor-based generation of relevant data, the progress has been dramatic and opens up new opportunities to develop systems that learn from their own data and experiences.

For example, a modern transportation department can use sensors or transponders, closed-circuit TV (CCTV) for video images, or perhaps mobile phone tower pings to more accurately determine not only how many vehicles pass a certain point at certain intervals. These systems can accurately track where data originates and where it goes. Knowing "who" is much more valuable for planning and operations than simply knowing "how many." When traffic counts were made by mechanical or manual traffic recorders, traffic workers knew only how many cars had passed. With more details about "who" (even anonymized), "where to," and "where from," they can build models that begin

to predict flows much more accurately than simulations, and with machine learning algorithms, even begin to control flow by adjusting traffic signals based on actual usage.

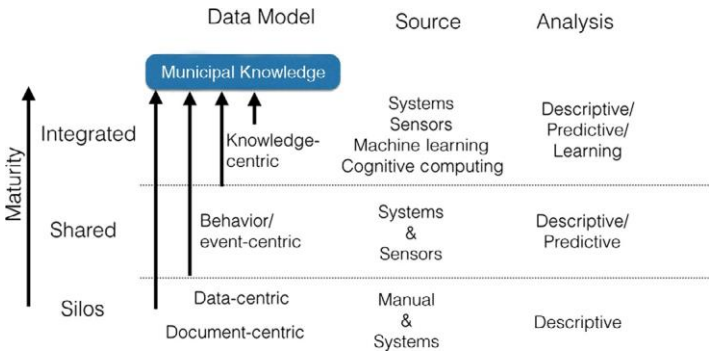

Figure 12-1: Foundations of cognitive computing for smarter cities

As similar advances were made in the data collection opportunities for other agencies, from health to safety to areas such as education, new and improved opportunities for analyses have emerged. As data from all these sources—across departments, structured and unstructured—is captured and made available in standard formats that encourage interagency sharing, cities become an ideal, data-rich environment for developing smarter applications.

The Characteristics of a Smart City

As noted, a city may best be understood as a combination of complex systems that have to work in collaboration with each other—sometimes called a *system of systems*. This is what makes cities and metropolitan areas so difficult to manage. Take a typical large city such as New York or Tokyo. These types of cities include roads and bridges, commercial and residential buildings, public transportation systems, private transportation, water systems, schools, and the public safety infrastructure. Although each of these elements is a universe within itself, they are all interdependent. Operationally, cities rely on smart managers to figure out best practices for managing and improving the way cities work best. But as cities have grown, it is simply impossible for smart managers to approach data driven problems systematically.

A city can become "smarter" if enough data is collected, analyzed, and managed so that critical improvements can be made. What does it mean for a city to be smart? It means that those managing a city collect the right information from a variety of sources and create a unified corpus of data that defines the components that make up a city infrastructure.

Figure 12-2 shows a typical set of government agencies, ranging from the basics of emergency services through utilities, public health, transportation, and human capital management. As individual citizens increasingly have persistent or mobile Internet access and become accustomed to interacting with their government in an ad hoc manner, you can also view community engagement as a function and potential differentiator for cities and other geopolitical units. The following sections break down these functions by task.

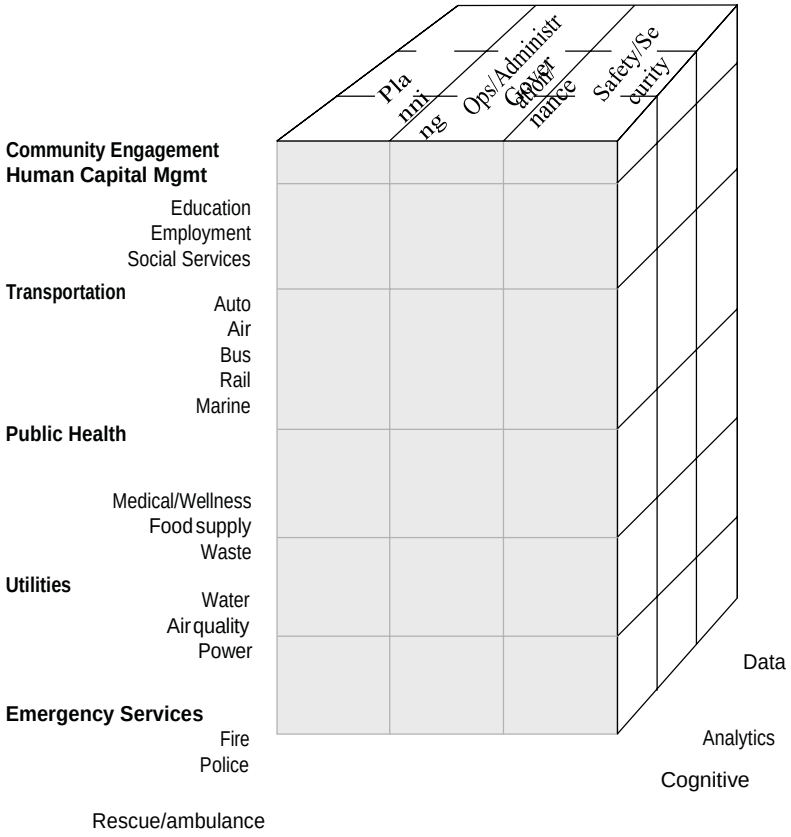

Figure 12-2: Data/knowledge management for cities

Collecting Data for Planning

Each agency in a city collects data for planning, operations, and safety and security, all of which are ongoing activities. The movement toward smarter cities includes better collection and analysis techniques at each stage.

City planning requires that management focus on a broad range of activities in order to promote growth. This growth has to be coordinated with the right scale of operations that allow that growth to be sustainable. At the same time, growth has to be planned so that it can improve and protect the quality of life for citizens. Planners evaluate options and codify them in policies. Planning requires managers to use their intuition and available

data to make decisions about the future. The best decision makers are those managers with enough experience to understand what will work and what actions will likely cause problems. However, even the most knowledgeable managers are helped when they get access to better analytical data. What are the trends for extreme weather? What events are happening in the region that might lead to civil unrest, agricultural failures, or drastic population shifts? What are the revenue shifts in industry? What is the outlook for wages? How do all these factors impact the way a metropolitan area can be managed efficiently and effectively? If these professionals managing cities are armed with analytics that can determine patterns and anomalies, they are better prepared for changes.

Cognitive computing applications are well suited to become dynamic planning assistants. Key technologies for cognitive planning systems would include hypotheses generation and evaluation, machine learning, and predictive analytics. The capability of some systems to read vast quantities of unstructured natural language text and analyze it for relevant events and trends will make planners everywhere more effective over time. The cognitive system does not simply analyze data from past events but collects data across anything that impacts the way a city operates. The system looks at the context and relationships between data elements and learns from the data that is ingested and managed over time.

Managing Operations

When policies are in place, various agencies are required to manage daily operations. The use of data and simple analytics has, of course, changed the way most departments operate. Creating repositories of "open data" by public agencies increases visibility into the processes of government and can improve community engagement. At the same time, this open process creates opportunities for commercial ventures to add value through analysis of the data, or even simply improving the way it is presented to the public.

Internally, this new data becomes even more valuable to civil servants because it makes systems based on prescriptive analytics more effective. City managers and planners are constantly trying to anticipate equipment and infrastructure failures and trying to determine how to provide better services without increasing costs. Cognitive computing approach solutions will soon become a mainstay of city operations centers because these systems can learn the rules of the policies while recognizing the realities of the terrain to "think outside the box." Today, manufacturers of complex and expensive machines like airplanes are already using machine learning to predict mechanical failures and ensure that the replacement parts are in the right vicinity. Tomorrow, no modern city will be without systems to do the smart repair and maintenance management from trucks to supplies, such as salt for icy roads.

Managing Security and Threats

Virtually all cities assume some responsibility for public safety and security. Beyond emergency services, protection from natural and man-made threats is an ongoing concern. Identifying potential threats that may occur with little or no warning (from gas line leaks or explosions to hurricane forces), internal and external threats must be monitored, assessed, and mitigated or met with managed responses. Cognitive systems are designed to identify trends and events from a variety of unstructured document sources combined with data from sensors, social media, and community sites. Leveraging all of this data is needed to make safety and security initiatives more efficient and targeted.

Managing Citizen-produced Documentation and Data

For each of the major departments, such as transportation, public health, and emergency services, there is a role for the foundation technologies and cognitive workloads (shown in Chapter 2, "Cognitive Computing Defined," Figure 2-1). Citizens, in these interactions with government, generate a considerable amount of data. All generate or capture data (refer to Figure 12-1), so the foundation structured and unstructured data management workloads are the requirement for advanced systems. At a higher level, senior manage- ment in each department has a role in planning and managing operations. By taking advantage of a cognitive assistant that helps to generate and evaluate hypotheses, a manager can gain insights and take corrective actions faster. The wisdom that comes from experience, once held tightly as a job require- ment, can be codified in a departmental corpus and used to make workers at all levels more effective.

Any system that gets input from the public or provides assistance to its residents can benefit from natural language processing (NLP) interfaces. However, the most value is created by systems that actually learn from experience as they build a corpus from use by residents and city employees, and increasingly from sensors, as previously discussed. For example, a transportation system that can accept updates about road conditions and traffic flow in real time from sensors, ad hoc natural language input from citizens, and is connected to the maintenance scheduling system for transportation department machines, will be more effective in dispatching the right equipment at the right time to make repairs, start preventative maintenance, or even schedule system upgrades, than a system that is missing any one of these components. Learning from experience—the defining characteristic for all cognitive computing solutions—is the key technology to improve this performance.

Data Integration Across Government Departments

The importance of interaction between departments cannot be overstated. If water, electrical, or gas lines are to be replaced or maintained, that will often have a major impact on overall infrastructure in other areas. For example, replacing gas lines might mean that a city has to reroute traffic or even resurface major highways. The department responsible for managing utilities will have to share critical information with the transportation planning system so that there is minimal disruption and so that the public is prepared. Predicting major events and planning for alternative routes can help a city manage change. Again, systems that learn from their experiences can better share data among departments to make the overall system more effective. In a major city, no manager can have insights into every planned or unplanned activity. But an integrated cognitive computing system with a common corpus that aggregates data and knowledge from all departments can certainly capture it all and share what needs to be shared between departments.

Chapter 3, "Natural Language Processing in Support of a Cognitive System," (Figure 3-1) discussed the fundamentals of learning systems and the concept of fast thinking versus slow thinking. Fast thinking tasks require intuitive actions that a manager might do without difficult analysis—responding to a citizen complaint or alerting people of an impending weather event. In contrast, slow thinking requires deep thought, analyses, and judgment. In city systems, fast thinking tasks can be automated to make departments more efficient providing deterministic answers to city workers and the public. These systems either require on demand access via the popular 311 information systems in U.S. cities or based on events, such as dispatching alerts to people in a defined perimeter when it's necessary to evacuate for a gas leak, or to remain in place during a police emergency. These answers and notifications can be based on knowledge managed within the cognitive computing system because the system understands the context between elements such as events, people, systems, and the like. The cognitive system is designed to understand relationships and patterns.

For slow-thinking problems requiring consideration of multiple scenarios, or for situations for which there is no single right answer, a cognitive computing system can supply probabilistic responses. Again, knowledge about the user—employee or resident—can make the answers more relevant. For a public health manager trying to determine which course of action to take when facing the possibility of an infectious disease outbreak, confidence-weighted alternatives such as those discussed for healthcare in Chapter 11, "Building a Cognitive Healthcare Application," may help in ensuring that supplies of vaccines or treatments are adequate. Integrating this system with the education system may have a ripple effect on bus planning, which may, in turn, have a

ripple effect on transportation logistics. In a city, everything is connected and interdependent. Unified cognitive computing applications, sharing a city corpus and supported by open data, can make those interdependencies a strength rather than a weakness.

The Rise of the Open Data Movement Will Fuel Cognitive Cities

Nations have long shared data valuable to businesses and individuals when it didn't compromise their own interests. From nautical charts in the 19th century to GPS data in the 20th century, the trend toward openness as a default has grown. In the 21st century, this trend has accelerated among cities, aided by better communications systems, standards, and regulations affecting the distribution of publicly held data.

In March 2012, then Mayor Bloomberg of New York signed the New York City (NYC) Open Data Policy (Introductory Number 29-A) and tasked the Department of IT Telecommunications with developing and posting standards for all agencies to make public data available online.

A goal of the initiative was to provide access to all public data from all agencies through a single web portal by 2018. To date, more than 1,000 public data sets are available from NYC agencies, commissions, and other groups, and available for a variety of private and commercial uses.

NYC also has a program called BigApps to "help teams advance new or existing projects that aim to solve civic challenges and improve the lives of New Yorkers." Teams work with civic organizations to develop applications that use this open data, while vying for cash prizes (totaling more than $100,000 in 2014).

Making data available is the first step, but today the problem isn't a lack of data, or even access to data. The problem is the ability to understand what of value is actually *in* all that data. Using the foundational technologies that are part of the fabric of emerging cognitive computing solutions has the potential for revolutionizing how data can improve the way modern metropolitan areas are dynamically managed.

The Internet of Everything and Smarter Cities

Previous chapters explore the technologies that make cognitive computing possible. Here you can find answers to two critical questions for public sector cognitive computing: "Where does the data come from?" and "How is value created?"

Now attack the first question. In a modern smart city, or one with smarter aspirations, data comes from three primary sources: citizens, governments,

and businesses, through systems and sensors. For businesses, much of the information comes from smarter buildings, which are managed by systems that coordinate information from all the internal systems and sensors for heating, ventilation and air conditioning (HVAC), water, power, transportation (elevators and escalators), and security. The adoption of standards for enabling all sorts of devices to connect to the Internet has made it easy to envision a future where everything that can be connected will be connected. Assigning an Internet Protocol address (IP_address) to a device uniquely identifies it to the rest of the world and allows it to potentially share information with anything else on the Internet. Computers are no longer the only option for Internet communica- tion. Increasingly there are a range of devices that incorporate sensors. Today, devices including refrigerators to smart watches and clothing include the ability to enable machine to machine communication. Summary information is provided to the receiving system or the humans using these sensor-based systems on a need-to-know basis. This so-called "Internet of Things" (IOT) or "Internet of Everything" (soon to be known simply as the Internet) enables businesses and governments to unobtrusively collect or derive all sorts of data about people as they carry on their daily activities. And most of it is given willingly, or without much of a fight. From smart meters that monitor energy usage in the home to transponders that allow us to speed through toll booths without stopping to devices that monitor acceleration and location in our automobiles to closed circuit TV (CCTV) cameras that can monitor group and individual movement, there is no shortage of data being generated.

This city-centric big data, when used to power predictive analytics algorithms or to develop a corpus for a cognitive computing solution, can provide insights that would never be discovered in time to be useful if the data were kept in departmental silos and analyzed by those with no incentive to share with other departments. It is the integration of this data that enables cognitive computing applications for smarter cities of the next decade.

Understanding the Ownership and Value of Data

In the new cognitive computing era, the traditional questions of who owns such data, who benefits from it, who transforms it into knowledge, and who owns that knowledge takes on new importance as the interconnections create new opportunities to improve the quality of life. At the same time they threaten to take away all semblance of privacy and perhaps security.

Figure 12-3 shows some of the critical relationships between citizens, businesses, and government that enable a virtuous cycle of knowledge creation. With the advent of social media and greater citizen participation in explicit and implicit data reporting, cities get smarter every day. Explicit data reporting includes applications like Street Bump in Boston, a mobile phone app that collects

real-time data about city streets by monitoring a user's GPS and accelerometer to identify potential problems such as potholes. (Speed Bump "knows" where the speed bumps are located so it can avoid many false positives.)

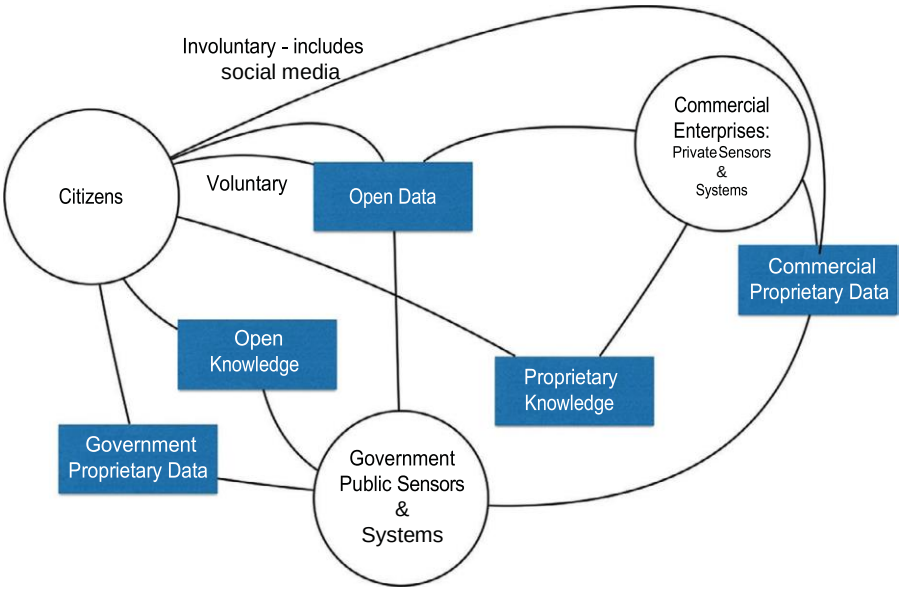

Figure 12-3: Modern city data sources and managers

Implicit data reporting includes activities such as making public comments about government activities on social media that are harvested by government agencies. It would also include systems or sensors that are commonly used with little thought given to the information gathering potential, from loyalty programs at retailers to facial recognition software used in public spaces on CCTV networks.

Cities Are Adopting Smarter Technology Today for Major Functions

This section presents a brief survey of projects in cities that are already working with leading vendors to implement cognitive and precognitive solutions (those based on foundation technologies like predictive analytics that pave the way for future cognitive computing efforts). As the field of cognitive computing is still in the early stages of maturity, most of these projects are ongoing and still being refined as the cities and vendors learn from their experiences. Using the typical urban organizational structure (refer to Figure 12-2), you

can focus on a few projects that may form the basis for integrated cognitive computing solutions in the future. In particular, you can consider opportunities to improve the quality of life through analytics and networks that learn from experience.

Managing Law Enforcement Issues Cognitively

It is not surprising that one of the greatest potential opportunities for smarter cities is in the area of law enforcement. This is an area in which there is a huge volume and variety of data that must be managed and analyzed. It is also an area in which patterns and anomalies play a huge role in solving and preventing crime. Leveraging a cognitive approach can benefit not only a single police department but also has the potential for providing repeatable best practices that can have wide use.

The Problem of Correlating Crime Data

The biggest problem facing police forces of metropolitan areas is the difficulty in correlating data from hundreds and even thousands of different data sources. Police departments often have access to large databases that capture historical data from local, regional, and national reports of arrests, crimes, and other recorded incidents. These data sources are often stored in relational databases. Other information sources are unstructured and stored in incident reports, paper files, interviews with witnesses, and the like. Still and video images and audio data must also be analyzed. In addition, there may be situations where a large volume of this data is produced on a daily basis from public and private surveillance cameras. Acoustic event detection monitors makes it impossible to manage manually. Recent advances in facial detection algorithms, however (based on recognition of facial parts and their spatial relationships, along with coloring), have helped to automate this task. The problem of analyzing real-time video as it is captured, however, remains a challenge. Big data provides new opportunities, but big cases still demand a lot of human personnel.

Departments, therefore, need enough time and enough skill to fit the pieces of data together to solve crimes. This is complicated because it is not always possible to correlate data manually. An experienced detective will know how to analyze data and will have mastered the process of connecting the dots. In addition, many crimes generate hundreds or thousands of new reports, pictures, bystander videos, and accounts that have to be considered, but which may be received as unstructured natural language voice recordings or text. The sheer volume and complexity of the data can be overwhelming. Even if the "answers" are in the data, finding them in time to prevent further crime can be impossible if a human has to find patterns between data sources.

The COPLink Project

COPLink is a law enforcement information management system originally developed by the Tucson, AZ police department and researchers at the University of Arizona. It has been commercialized by i2, an IBM company, and deployed in more than 4,500 law enforcement agencies in cities around the United States. This customized project is now being turned into a repeatable cognitive solution under the IBM Watson brand.

COPLink enables individuals, departments, and agencies to gather, share, and analyze historical and current crime information. Users can access data from local and national databases from virtually anywhere, using computers or mobile devices to improve the effectiveness of field personnel. COPLink supports general data standards XML and those commonly used in law enforcement, such as the Logical Entity exchange Specification Search and Retrieve (LEXS-SR), to simplify integration with other applications and to promote data sharing.

The Adaptive Analytic Architecture (A3) enables users to generate leads (hypotheses) for tactical follow-up without extracting data permanently. COPLink also offers agent-based alerts to let officers know when similar searching is made to help coordinate efforts, and can send notifications when new information becomes available based on saved searches.

COPLink uses analytics to actually create suspicious activity reports based on alerts and can share these with the relevant jurisdictions. Simplifying searches and providing this level of database integration makes personnel more efficient. Continuous monitoring of new data and events to improve communications makes them more effective.

There are a number of ways that this type of law enforcement application is being expanded with cognitive extensions, including:

- Providing NLP analyses of unstructured reports as they are created in the field.
- Leveraging hypothesis generation and evaluation technology from Watson to augment the case lead generation process that is currently handled manually by officers.
- Treating the locally generated databases as the foundation for a corpus that can be integrated with cognitive operational systems to plan and manage emergency staff and resources. (COPLink uses analytics to predict crime hotspots, but integrating its corpus—including new knowledge learned from experience with hypothesis testing—with planning tools would enable department managers to benefit from the actual experiences.)
- Integrating sensor data from smart transportation systems in (near) real time to support continuous monitoring of activities that may appear innocuous in isolation, but which may be part of patterns that would otherwise go unnoticed.

Smart Energy Management: From Visualization to Distribution

The importance of visual reports to help people make discoveries from complex data relationships has been discussed. Visual representations of the state of complex systems can also be used to ensure confidence in an operational system and to enable a human to participate in the management process when anomalies arise. For example, allocating energy production resources for a smart grid, or responding to expected changes in demand based on predictive analytics, may create situations that benefit from operator intervention. Visual abstractions make it easier for the operator to detect patterns than simply seeing relevant numbers scroll past on devices. As power systems integrate subsystems with different characteristics or energy sources, visual interfaces are the only practical way to present all the information as a continuous stream. This is much like the approach taken with an automobile dashboard that abstracts some data into simple color maps—red for danger, yellow for caution, and green for normal operations—but provides more detail for other functions, such as a numerical estimate of the distance that may be traveled before refueling. In a single system that balances power generation from water for hydroelectric generation, wind, nuclear generation, and solar and storage in the form of batteries or supercapacitors, visualization is the key to helping the user understand what requires attention as soon as the system can detect it.

The Problem of Integrating Regional Utilities Management

A new smart city is being developed in Kashiwa-no-ha, Japan, through a collaboration between public agencies and private sector enterprises with three stated goals: environmental symbiosis, support for health and longevity, and creation of new industries. Creating a new city by design presents an opportunity to build a utilities-based infrastructure that is based on state of the art technologies with minimal constraints from aging systems.

The Area Energy Management Solutions Project

The Area Energy Management Solutions (AEMS) project in Kashiwa-no-ha, Japan, uses advanced analytics to provide an integrated, comprehensive solution for energy production, provisioning, and optimization. AEMS design and development is led by Hitachi Consulting. Hitachi, the $90BUSD global technology and services provider, has emerged as a leader in the smarter cities movement. Its 2014 Annual Report, titled "Social Innovation: It's Our Future," lays out a vision based on building and leveraging technology to address social challenges, including three that resonate with smart city planners: securing water resources, energy and food; replacement of aging infrastructure systems; and improving transportation systems.

The AEMS project is focused on using analytics to manage energy (including electricity, water, gas, and any other production technology that is ultimately adopted by Kashiwa-no-ha) by forecasting demand and dynamic provisioning based on continuous reporting signals from sensors throughout the city. Creating a solution that incorporates renewable energy sources (solar and wind) with more conventional sources and storing excess capacity through storage batteries allows the system to schedule production when it is convenient and cost effective (daylight for solar and medium-to-high wind for turbines) and puts it into the grid on-demand directly or from batteries that store excess capacity from a period of over-production.

AEMS uses analytics to predict peak loads and evaluate alternative approaches to distribution (from interbuilding sharing of resources, similar to the cloud-based model for sharing physical resources).

By planning for an integrated set of energy systems rather than developing hydroelectric, solar, and other data management systems in isolation, the developers created a design that leverages data from each subsystem and also uses analytics to efficiently load balance the overall system to predict demand, provision resources, and distribute more efficiently. AEMS shares information with building energy management systems in commercial, government, and residential facilities.

The Cognitive Computing Opportunity

From the published plans and reported progress about Kashiwa-no-ha, it is clear that the planners want to take advantage of all the smarter cities' products and practices that can be applied to the operation of their new home. Kashiwa-no-ha is expected to become a model, or test bed, for emerging cognitive technologies. Further integration, to include systems for each of the major constituencies discussed in this chapter, will be a cornerstone of system design and provide opportunities for collaboration. The Hitachi AEMS is new but expected to be integrated with other smart systems in Kashiwa-no-ha over time. For example, integration with its transportation management system, and even weather forecasting and monitoring, would improve the performance of all the systems by sharing new patterns discovered by machine learning algorithms. Consider a day when the forecast called for mild weather, but the actual weather that day is unseasonably wet and cold. An integrated system that has learned from its shared experiences might see a correlation with changing traffic patterns, resulting in changing power demands from public transportation systems. In this scenario AEMS might predict an increase in the ratio of people working from home, and prepare to reallocate power based on the behavior of residents, not a preset model of consumption. Integration of algorithms and data between these systems could use real-time or just-in-time sensor data about the discrepancy between forecast and actual weather to "know" that power demands would change in time to adjust production or distribution patterns.

Protecting the Power Grid with Machine Learning

In the United States, power grids are included in the energy sector of critical infrastructure that must be secured against cyber attacks under provisions of the National Cybersecurity and Critical Infrastructure Protection Act of 2014. That requires the Secretary of Homeland Security to conduct activities to "protect from, prevent, mitigate, respond to, and recover from cyber incidents." Energy companies are turning to machine learning algorithms to keep up with the volume of data that must be continuously monitored to mitigate risk and maintain compliance with increasingly stringent regulations. One emerging software company called Spark Cognition in Austin, Texas has developed a cognitive system intended to help secure Internet of Things environments such as power grids.

The Problem of Identifying Threats from New Patterns

Electrical grids are attractive targets for a variety of physical threats, from vandalism to terrorist attacks. More and more, however, the threats are from cyber attacks that attempt to disrupt service. The scale of the physical infrastructure and complexity of connections and dependencies make automation of threat and breach detection critical and potential threat detection a requirement. As new patterns for threats and vulnerabilities emerge, utilities must prepare responses before they materialize.

The Grid Cybersecurity Analytics Project

C3 Energy (C3) is an energy information management firm founded by the founders of Siebel Systems with a mission to make energy systems more efficient and secure through analytics. C3, working with researchers at the University of California and a National Science Foundation Cybersecurity Center called TRUST, developed Grid Cybersecurity Analytics (GCA)—a smart grid analytics application that uses machine learning algorithms to identify and detect potential threats. GCA was designed to understand the characteristics of normal operations (communications traffic levels, asset activity, and such) and identify potentially threatening activities or anomalies. The machine learning algorithms enable GCA to adapt and identify emerging threats as they evolve or mutate over time by reading up to 6.5 billion records/hour and providing petabyte-scale analysis. That level of performance is required for large grids serving cities.

The Cognitive Computing Opportunity

The Grid Cybersecurity Analytics system already leverages machine learning algorithms and the latest research in threat assessment and response. Natural extensions to such a system beyond its application in grid security would center on leveraging data and lessons learned as the system gains experience with threats that have commonalities with potential attacks on other critical infrastructure

assets, such as bridges, tunnels, and information networks. A scenario is also envisioned in which systems such as GCA share data with systems such as COPLink and sensor-based smart asset-monitoring systems to help prevent or mitigate attacks by sharing data on individual or groups whose behavior raises a threat alert. On the mundane side, data from the underlying analytics platform could also be shared with a corpus for a resource management system that could be tied to incentives for better energy management and to planning systems to recommend incentives for electric vehicles.

Improving Public Health with Cognitive Community Services

Public health in cities is concerned with wellness and medical care. Wellness includes access to information and preventative care, feedback on behavior that impacts health, and the availability of a full range of care when prevention is not enough. In some jurisdictions it may also include monitoring or managing risks in the food supply and waste management that can improve or diminish overall health. Some of the earliest adopters of cognitive computing technology have been commercial health management firms that have offered personalized recommendations on activities and nutrition, and sometimes incentives for good behavior. In cities, however, many efforts—such as restricting food sodium content or limiting soft drink sizes—have taken a one-size-fits-all-citizens approach. Cognitive computing technology, which can tailor findings and recommendations for diagnoses and individual wellness plans, is the next logical step.

Smarter Approaches to Preventative Healthcare

In many established cities, access to preventative care is seen as a social service for economically disadvantaged people or offered as a perk through private or public employers' insurance policies. As the cost of personal health monitoring drops, through the availability of free or modestly priced online educational resources and access to sensor-based feedback devices, local governments can begin to justify community-oriented health services that offer personalized care driven by analytics.

The Town Health Station Project

Kashiwa-no-ha, Japan, has worked with academia (The University of Tokyo) and businesses such as Mitsui Fudosan to develop the latest in community healthcare supported by cognitive computing solutions to aid in preventative care. Its Town Health Station is a model for public health that is expected to be emulated around the world. The project was designed from the outset to

embrace the Internet of things and take advantage of new, continuous sources of health information such as the output of personal devices, from mobile phones to exercise bracelets to workplace and residential sensors. The health station is the center of a partnership between local government, the University of Tokyo Institute of Gerontology, the Center for Preventive Medical Science at Chiba University, and residents, to promote long-term health.

The Cognitive Computing Opportunity

The Town Health Station project (and ancillary programs such as community exercise activities) are already creating a corpus of individual and community data from sensors and professionals that can be shared with other communities to improve the quality of care locally and remotely. As a planned smart city, Kashiwa-no-ha uses analytics to manage traffic and even plan for shared transportation resources in the spirit of the emerging global sharing economy. Integrating these systems in a common corpus that learns from the experiences and behavior of its residents is a natural evolution for the town, and one that would keep it at the vanguard of smarter cities. Similar to the integration envisioned for the Hitachi AEMS solution, the learning systems that share new knowledge between the town, academia, and the medical community is also expected to share the anonymized unstructured health and wellness data with every system that can use it to improve performance.

Building a Smarter Transportation Infrastructure

Intracity transportation management is, in many ways, more difficult than regional and national transportation management due to population and structure density. As cities build up, it is increasingly important to manage transportation and traffic flow through better use of information. When it becomes prohibitively expensive and disruptive to add infrastructure such as additional lanes or levels to roads and rails, getting smarter about who can go where, and when, will drive cities to cognitive computing solutions.

Managing Traffic in Growing Cities

In cities everywhere, traffic congestion leads to frustrated drivers; excess energy consumption, inefficient commerce as people spend more time driving and parking when they could be working and shopping; delayed emergency responses; and higher pollution. It also requires assets to manage peak loads that may be underutilized in general. Almost any change to flow involves trade-offs that inconvenience people and interrupt commerce, while potentially slowing emergency responses. In 2014, lane closures on a heavily traveled bridge between

two metropolitan areas in the United States nearly became a political scandal when paramedics took twice as long as usual to respond to an accident with multiple injuries.

The Adaptive Traffic Signals Controller Project

The city of Toronto, Canada, adopted the Multi-Agent Reinforcement Learning Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC), a system for smarter traffic management, with impressive results. Toronto reports that downtown delays have already been reduced by 40 percent on an average day by simply managing traffic signals more effectively. Xerox, the $23 billion information and document management company, worked with the University of Toronto's Intelligent Transportation Centre to develop and deploy the system, which incorporates camera images and machine learning chips to enable real-time communication between traffic lights to detect patterns and dynamically adjust their timing. Helsinki has a similar traffic project underway with Xerox, with comparable results and similar opportunities to expand into a cognitive computing environment. As Xerox continues its focus on business transformation and migrates away from a dependence on manufacturing devices, it is building up its credentials in smarter city professional services and integration with projects such as MARLIN-ATSC.

The Cognitive Computing Opportunity

MARLIN-ATSC is already an example of machine learning and adaptive, integrated devices. In the next several years, many opportunities may be available to integrate systems such as this with other cognitive computing municipal management systems, and to extend its capabilities by collecting data from external systems such as the cars themselves.

For example, the U.S. Transportation Department is working on a "connected vehicle" initiative that would promote the use of car-mounted wireless devices that communicate with each other and with traffic signals to increase the effectiveness of systems such as MARLIN-ATSC. In addition to the obvious issues of individual privacy and security, such a program faces years of testing and changes in legislation before the government would require the use of these devices in new cars. It is, however, technically feasible now.

Transportation management is one of the most developed domains for analytics and cognitive solutions. In part, this is due to the ready availability of sensors and systems to collect and share data. Virtually every component of a transportation system is amenable to measurement, from cars and buses to planes, trains, and boats, plus the highways and ports that support them. Combining data from this type of signal management with data from CCTV or transponders—even anonymously—could further improve the capability

of cities to reduce congestion while providing input for security and planning. Integration with security systems such as COPLink to identify patterns after a criminal event could help law enforcement's capability to predict future events. Integration with systems such as the Grid Cybersecurity Analytics may have limited threat improvement opportunities, but offer real potential to improve the use of electric vehicles by better understanding their usage patterns to optimally locate charging stations and introduce variable pricing to encourage usage patterns that minimize infrastructure requirements.

This type of integration may also help speed the way for the adoption of autonomous vehicles (self-driving cars, buses, and so on) by promoting communication between driverless and human-piloted vehicles to make a smoother transition.

Using Analytics to Close the Workforce Skills Gap

Human capital management in cities requires management of significant interdependencies among employment, education, and social services. Cities also need to find ways to try to lower unemployment, because it can increase crime and lower the tax revenue at a time when assistance is needed the most. Neighborhoods with high concentrations of unemployment face additional problems as lower income disproportionately impacts small businesses and real estate values, too. As businesses tighten their skills and experience requirements for hiring, it is important that new workforce entrants and the unemployed have the right skills when openings appear.

Identifying Emerging Skills Requirements and Just-in-Time Training

A representative from Boeing, one of the largest aerospace firms noted, "We don't have necessarily a labor challenge, we have a skills challenge." Even with thousands of job openings and thousands of applicants, many jobs at this firm still go unfilled. Preparing residents for opportunities by identifying and training them on appropriate skills before the need arises requires an understanding of emerging skills requirements supported by analytics. Matching skills training to aptitude in an applicant pool is a perfect opportunity for cognitive computing applications.

The Digital On-Ramps (DOR) Project

Leaders in the city of Philadelphia, PA, with a population of approximately 1.5 million, recognized that their residents were falling behind the digital literacy and technical skills demands of industry and recognized that significant action had to be taken. In 2011, it was estimated that by 2030, it could have 600,000

citizens who were unprepared to compete effectively for new jobs. At that point, only 41 percent of homes in Philadelphia were connected to the Internet, but increasing mobile access was not being leveraged effectively to improve training or education. Working with IBM under a grant from its Smarter Cities Challenge program and support from the Clinton Global Initiative, Philadelphia developed Digital On-Ramps (DOR), a learning delivery system that includes in-person and digital education, tailored to anticipated industry needs and individual aptitudes and learning styles. Similar to the electronic health records (EHR) discussed in Chapter 12, DOR creates a universal ID and "digital record depository" for each learner that tracks school, training, and work experiences and accomplishments. Captured as structured and unstructured data in a personal portfolio, it is used to guide the learner toward a job goal with the advice of a counselor. Today, the counseling is done by human practitioners, but the system is capturing data that is useful for evaluating progress using descriptive analytics. This data should soon be useful as input to predictive analytics when the last two stages in the DORS process—building individual skills and matching a skill set to jobs and networks—have generated sufficient data to build a corpus for a cognitive computing solution.

Early results from DOR indicate that creating individualized programs for future industry needs while leveraging a variety of disparate resources, from the free library to local schools and corporate philanthropic programs, can be a successful endeavor. Human will reinforced by analytics has already enabled Philadelphia to secure a MacArthur Foundation grant to develop a certification process to help guide employers when hiring graduates of the DOR programs.

The Cognitive Computing Opportunity

A strong workforce attracts businesses, which strengthens the workforce, creating a virtuous cycle. DOR was launched in Philadelphia, PA, but from the outset it was considered to be a model for other cities struggling with the mismatch between skills and jobs in industry. Integrating cognitive computing functionality into DOR will make it more powerful for those in the program, but also for ongoing city planning and management. For example, adding NLP capabilities for interactions between individuals and the system will lower barriers to participation and provide more tailored responses.

As participants become comfortable with sharing information about personal aspirations and experiences with the system (in natural language). It could learn from experience what works and what doesn't (self-training), and begin to make better recommendations and suggestions for midcourse corrections if it detects changing employment conditions. From reading and interpreting labor statistics, job ads, and editorials in unstructured text from around the country and around the world, such a system could provide ongoing personalized guidance to a population that would otherwise be left to its own devices.

Finally, an integrated cognitive computing environment that shares data across all major city systems could use DOR data to improve transportation planning and operations, utilities demand forecasting, and public health as the impact of new jobs propagates throughout the city and region.

Creating a Cognitive Community Infrastructure

As you have seen, cognitive computing solutions benefit from retaining knowledge created during operation as hypotheses are tested and refined over time. When natural communities—groups of people with similar interests, experiences, or simply geographic proximity—communicate, they raise the collective intelligence of the community. Professional communities that communicate with or through a cognitive computing solution can amplify the learning, as you have seen through medical diagnostics. Physical communities, such as city residents, can likewise benefit from collaboration via cognitive computing.

The Smart + Connected Communities Initiative

In New Songdo City, South Korea, a partnership between the Korean government and private industry is building a new, green, smart city designed from the outset to make extensive use of sensors and analytics. The $35 billion (U.S. dollars) project is one of the largest new city ventures in the world. Cisco, the $48 billion U.S. dollars global networking and communications firm, has wired every home and business in New Songdo City with video screens and systems to facilitate collaboration and a sense of digital community within a physical community expected to have a population of 65,000 by 2016. This Smart + Connected Communities initiative is the first in what Cisco and the development organization hope is dozens of similar projects throughout Asia. Cisco is a leader in developing devices for the IOT and was an early pioneer of the connected world concept. Cisco is promoting New Songdo City as a showcase for the benefits of networking and remote control of residential and commercial energy management and security. These systems will be integrated from the outset, and machine learning tools such Cisco's Cognitive Threat Analytics—which learns and adapts as new threats are identified—will be deployed as New Songdo City nears completion.

The city, set to be operational in 2015 and home of the largest skyscraper in South Korea by 2016, includes an international school for children of visit- ing business executives, which is connected at all levels with a sister-school in California. Extensive use of video for collaboration and communication is planned to develop a community of learners within the general population. As video analysis techniques mature, that technology will be integrated to continue the automation of learning support.

The extensive use of telepresence and sensor data to deliver more personalized services to every residential and commercial location should foster communication while discouraging energy-intensive travel. This approach fits with New Songdo City's goal of world-class sustainability.

The Cognitive Computing Opportunity

New Songdo City is starting with many advantages over existing cities—before the people arrived, the initial wiring was in place to encourage communication and capture sensor-based, video-based, and system-based information. Beyond community sharing, these systems integrate with transportation, energy, and water management systems, providing an unprecedented opportunity for a city whose population is supported by and bound together by cognitive computing technologies.

The Next Phase of Cognitive Cities

As more of the global population moves into urban areas, the importance of city-oriented cognitive computing will increase. Making better use of the big data generated by the daily activities of individuals, public sector agencies, and businesses will support differentiation for cities as they compete to lure valuable talent and businesses and increase the effectiveness of collaborative efforts for alliances.

For each area, however, you can expect to see an ongoing conflict over private versus public ownership of the data and knowledge, as competition to market intelligent solutions to the densely populated urban residents increases. This is not an issue for cognitive computing per se; rather it reflects the great value that will be created as a result of smarter processing of dumb data at scale.

Most of the examples in this chapter have come from the introduction of analytic and cognitive computing technologies to address straining physical and information infrastructures in established cities. That will continue to be an issue for the foreseeable future, but the desire for urbanization is also driving the creation of new cities to meet the demand. Although planned communities of the last century, from Reston, Virginia, to Celebration, Florida, were built around architectural and open space patterns harvested from the study of centuries of organic city growth, the next generation of planning will be based on communication and collaboration needs, much better suited to cognitive computing analyses than to nostalgic looks at successful town squares.

Kashiwa-No-Ha, Japan, and New Songdo City, Korea, are essentially new cities unencumbered by aging infrastructure and preconceptions. As these two cities mature, and others like them spring up in the next two decades, they will demonstrate the value of integrated analytics and cognitive computing to

improve municipal life, and even the most conservative cities will have to adapt or see their relevance erode in the age of cognitive computing.

Summary

Smarter cities have moved beyond collecting data such as vehicle and animal registrations and tax-based data on homes and income. Now, city managers are collecting data through collaboration between departments with optimization in mind. Some leaders have appointed chief data officers to look for opportunities to share data whenever new systems are created. A new public health record database, for example, may be designed to capture elements that can be shared with an education database, or a security database. Most cities are actively making as much data as possible open for free use by the public and by innovative entrepreneurs who can build new services on this data. Every shared use or new commercial use can potentially produce value for residents.

As smart city managers, including elected officials and civil servants, begin to understand the power of data to transform city life, they are embracing advanced analytics and cognitive computing as a source of added value. Smarter cities make better use of all their resources, and having good data is at the heart of these better decisions. Today, the best systems for creating this type of data are the class of emerging cognitive computing applications that learn from their experiences and can guide their users to better decisions and outcomes. In the future we may take for granted the cognitive computing applications that make cities safer and more efficient, while anticipating our needs as residents. Today, the journey is just beginning but the benefits are already clear.

CHAPTER

13

Emerging Cognitive Computing Areas

Cognitive computing is beginning to have an impact on a number of different industries. Initially, cognitive computing is starting to transform the healthcare industry, as discussed in Chapter 11, "Building a Cognitive Healthcare Application." Cognitive computing applications are already:

- Making and enabling new discoveries about patterns of symptoms that indicate specific diseases based on the power to read more research and case files than any individual physician
- Explaining the results of findings to medical students to help them refine their own diagnostic abilities
- Democratizing specialized knowledge and making it available to practitioners, who would rarely, if ever, see certain symptoms or conditions

All these capabilities that help drive the adoption of cognitive computing in healthcare will have the same potential in other industries. For example, finding patterns for disease diagnosis and making treatment recommendations are special cases of the general problem of identifying faults and remedies in complex systems. This capability to diagnose problems has applications in industries such as manufacturing—machine maintenance and repair from household appliances to oil rig apparatus, and even call centers for problem resolution. Evidence-based explanations that help to train new professionals can be used in any field where a large or complex body of knowledge is codified. Finally, almost every profession

has the potential to thrive by aggregating and analyzing specialized knowledge and packaging and selling that data as a service. Professionals ranging from lawyers to accountants and stockbrokers can be democratized by giving newly-minted professionals and inexperienced users the power to leverage this knowledge with the assistance of a cognitive computing guide.

This chapter identifies attributes of problem domains that make them well suited to similar transformations in other industries, and shows how some functional areas across all industries can also be transformed.

Characteristics of Ideal Markets for Cognitive Computing

At the heart of every cognitive system is a continuous learning engine that improves with experience and that can return probabilistic results when the data supports multiple candidate answers. Advanced systems make use of natural language processing (NLP) to capture meaning and nuances in text from publications. In addition, these systems can make use of images, gestures, and sound. The systems can increasingly capture streaming data from Internet-enabled sensors. Mapping this functionality to typical business workloads—datacentric tasks found within and across industries—reveals a set of attributes that make some problem domains well suited to cognitive computing applications.

Ideal candidates for cognitive computing include:

- Industries with rapidly increasing or changing volumes of domain-specific knowledge, which make presales advice valuable to buyers but costly to sellers. This would include areas such as retail, where products and offers change constantly, and travel, where options and opportunities change constantly.
- Industries where post-sales support/diagnostics—professional services—are cost centers or revenue opportunities due to complex products. This is especially important when staff turnover is high and changes in product data requires constant training. This includes virtually every industry currently using call centers, from retail to enterprise software support.
- Industries characterized by a lot of specialty knowledge or experience that is highly concentrated in a small group of experts. This includes many professions and complex product presales where configuration decisions are critical.
- Industries that traditionally follow a modern apprentice/intern model for training and certification. This would include most data rich professions such as law, medicine, and financial services.

- Industries in which best practices are known or knowable, and there is a wide variance between the best and least-effective practitioners in the field. (The top experts are differentiated from average performers by their ability to draw on personal experiences or recognize patterns that are unknown or too complex for inexperienced workers.) This would also include most professions.
- Industries in which the sudden availability of sensor data creates opportunities that cannot be exploited by conventional means. This includes industries from transportation to healthcare; wherever sensors can capture meaningful data where real-time analysis by experts has value, cognitive computing applications can be adopted.
- Industries in which success depends on discovering patterns in a large volume of data, particularly unstructured natural language text. Any field that produces a large stream of new research, beyond the ability of practitioners to absorb on their own, is a candidate. This includes most of the natural sciences, pharmaceuticals, and the professions (new case law around the world, changing regulatory requirements, and so on).

As a general rule, the introduction of cognitive computing systems can make an organization smarter by making the top performers more efficient. In fact, even the most sophisticated professional will not be aware of all the new findings in a field. A cognitive system prevents biases from driving decision making. It is quite common for the most experienced professionals to select an approach based on their own best practices without taking into account that there is new data that they have not been exposed to. Less experienced professionals can perform at a much higher level because they can benefit from the shared knowledge.

The cognitive system is not static. It is continually ingesting new information from data and from both successes and failures. The dynamic nature of this environment has the potential to raise the level of expertise of the entire organization.

Vertical Markets and Industries

This section looks at a few market segments that are already using advanced analytics to improve their performance, and have leaders exploring cognitive computing solutions. The intent is to show how the common characteristics are driving this adoption and acknowledges that there are many other fields with similar attributes and many proof-of-concept projects underway that will change the public perception of cognitive computing and create demand for smarter systems in almost every field.

Retail

Retail is a notoriously competitive industry. To survive and thrive, retailers have to anticipate what products to purchase based on forecasting trends in advance. They have to understand the impact of changing economic, social, and demographic factors. Retailers also have to make sure that their employees do a good job at both representing the company and the products sold. External factors can also wreak havoc with plans and forecasts. Unseasonable weather, changing gas prices, fluctuating employment rates, and even political unrest can impact buying behavior. Larger firms have used predictive analytics and scenario planning tools for decades. These firms have optimized supply chains to reduce the lag between ordering an item and delivery (lowering the risk of receiving out of fashion or unwanted goods). However, too often retailers miss subtle changes in buying preferences and do not anticipate opportunities to gain advantages over competitors. Cognitive solutions have the potential to help retailers leverage knowledge in inventive ways. For example, a typical large retailer relies heavily on its supply chain automation to deal with customer problems. These systems are weak when unanticipated problems arise. Being able to provide creative ways to deal with problems so that customers remain satisfied and loyal.

Cognitive Computing Opportunities

Many retailers use predictive analytics tools to detect interesting correlations to discover insights based on loyalty-card data. For example, analytics help retailers to recognize changing habits, buying preferences, and changes in life circumstances (marriages, pregnancy, and such). It is possible to differentiate between an anomaly and a true change in buying preferences. For example, a product may be suddenly popular because of a single event. (A sudden storm causes consumers to purchase snow shovels in a region that gets little snow.) With cognitive computing, these changes may be detected earlier, enabling the retailers to implement innovative new practices and approaches that can change the customer experience. Retailers will be more interactive with customers through natural language dialoging with high-value customers. These retailers will have the ability to learn from the information gleaned from social media and customer interaction with call centers. They will bring together all this data so that it learns from practices of the most successful sales personnel and from the most successful campaigns. Armed with this type of dynamic environment, a retailer could discover new approaches to retail that can change the relationship with customers. Even increasing sales from repeat customers by a small percentage can result in huge profits.

Personalized Customer Service

In retail sales, customers are constantly faced with new choices, features, or fashion and must decide which product to buy, when to buy, and from

whom to buy. Although recommendations from friends on brands and sellers are important, the customer experience in a store or on a website can make the difference between casual interest and an actual purchase. With that in mind, retailers struggle with the issue of how much personalized attention to provide for each shopper—sometimes sacrificing volume for higher margins. Buyers usually have to choose between stores that offer knowledgeable sales staff and those that don't, whether the purchase is made in a physical store or online. For higher-priced items, from luxury goods to complex home electronics, this creates a tension when good advice in a store often leads to an online purchase based on price. The promise of cognitive computing in this case is to democratize good advice. This requires that retailers offer personalized attention to a buyer's wants and needs by using extensive knowledge about the products. It also requires that retailers gain insights about those customers so that they can make recommendations that are meaningful. Taking advantage of a system that continuously learns and adapts based on collecting and analyzing massive amounts of data can make a huge difference between success and failure.

An early entrant into this market is Fluid, a 15-year-old firm specializing in building online customer experience tools for leading retailers. The company partnered with IBM (who announced an equity investment in Fluid) to build a Watson-based platform that enables customers to communicate with online retail sites to provide customized product recommendations. The goal of this service is to lead the buyer through a dialog that mimics the customer engagement of a personal shopper. Fluid's first customer is NorthFace, a retailer that offers hundreds of products for outdoor activities and caters to "explorers" who are willing to pay a premium for quality.

The Watson-based platform will enable customers to enter a natural language description of their requirements from "I need a sleeping bag for a trip to Argentina in May" to "What should I take on a camping trip with three young children?" By engaging in a conversation with the buyer, the system can leverage information in its corpus about each item in inventory, and each type of activity in its ontology (for example, from camping to hiking). Through this dialoging process, the system narrows the scope of suggestions based on a match between what is available and the knowledge it acquires about the customer from this conversation. The system also stores all the background information from previous customer interactions and queries. The system can also look for similar queries and outcomes from other customers. This is not the same as a simple recommendation engine that provides a series of potential options. In the Watson system, the user can generate questions based on the assumption that the system contains depth of knowledge. As the system collects more information about consumers over time, the level of confidence in those recommendations increases. Retaining information between sessions allows the system to learn from each interaction and provide better advice in subsequent sessions with the same or different users.

Retail Staff Training and Support

It is critical that in-store personnel provide consistently good advice based on product knowledge and good customer interaction skills. However, retailers typically have high turnover so that average salespeople lack deep knowledge about the products they sell. UK retail technology firm Red Ant commissioned a study of 1,000 retail workers aged 18–55 by an online polling firm. The results were revealing:

- 50 percent of respondents reported feeling embarrassed by their own lack of product knowledge
- 43 percent said they lie to customers every week due to product knowledge deficiencies
- 73 percent said they send customers to another store
- 57 percent said they were given less than 2 hours of training before being sent to help customers

The results clearly indicate that there is a critical need for training. Employees need to understand the products they sell and need access to best practices. Employees often quit or are fired under less than ideal working conditions. It was clear to Red Ant that there was an opportunity to use a cognitive approach to improve the performance of retail workers.

Red Ant specializes in helping retailers improve their processes through analyzing customer and retailer behavior. Therefore, Red Ant is developing a Watson-based retail sales-training application to market. The goal of the product offering is to help sales associates analyze customer demographics and purchasing histories. The application will provide access to product information and market feedback from sentiment analysis to help retail workers make better recommendations to customers. Bringing together customer information with product information while the customer is in the store will enable a more engaging dialog using NLP from Watson and will create a more personalized shopping experience for the consumer. Every interaction with that customer is recorded and can be compared with interactions with similar digital histories. This corpus of data is intended to help predict what will be most effective and record the results. The sales associate may get recommendations via custom on-screen prompts, which could be shared with the customer or via text-to-speech messages through an earpiece.

Travel

The travel industry has seen tremendous upheaval in the past two decades as information about rates and schedules of transportation, lodging, and leisure activities has been made freely available online. Self-service booking sites have enabled individuals to make their own reservations after searching multiple sites for descriptions, prices, and even reviews, without paying a fee, and other sites have emerged to harvest results from multiple sites with a single search.

That has eliminated the personal touch of an experienced travel agent who could get to know the customer and ask qualifying questions to ensure a good fit between personal goals and desires, and available inventory of transportation, lodging, and experience options.

Although much progress has been made in terms of making information visible, optimizing yield on flights and cruises based on predictive analytics and customer history, what is lost is a store of personal information about the *why* aspects of trip planning, and experience-based recommendations that understand the client's objectives. An individual may have different preferences for pleasure and business travel, and different preferences based on duration, location, and who is paying for the trip or even who is accompanying them, but no single site captures all this information today.

Cognitive Computing Opportunities for the Travel Industry

It was once common to work with a travel agent who could get to know the preferences of an individual and be on the lookout for new options and new opportunities. Today, the traveler typically has to provide a profile of standard options for each site they use. However, none of these sites provide inferences are made based on observable behavior. This leaves a big opportunity for a cognitive computing travel application that captures information explicitly by capturing patterns of travelers' behavior. There is also the potential for implicitly understanding travelers by monitoring social media streams. There is also an opportunity to allow travelers to interact via an NLP interface with the system.

An example of what you can expect is a company founded by Terry Jones, the founding CEO of Travelocity and an early chairman of Kayak. WayBlazer is a startup founded by Jones with the intention of leveraging cognitive computing services from IBM Watson to add evidence-based advice to its product. WayBlazer is also built on top of a cloud solution built by Cognitive Scale, a company that provides a cognitive-insight-as-a-service platform. The company is collaborating with the Austin, Texas Convention and Visitor's Bureau to create an application that will provide customized recommendations for individuals. Over time, the company intends to expand to provide concierge services to hotels and airlines to improve the overall user experience and provide additional revenue opportunities to these ecosystem partners. WayBlazer is using the NLP and hypothesis generation/evaluation capabilities from Watson to evaluate a corpus initially populated with data from the destination and transportation vendors (suppliers), but which will be augmented by knowledge it acquires by monitoring traveler requests and outcomes. In addition to earning the customary fees from transactions, the system can learn about individuals and group behavior. WayBlazer will collect valuable data that it will be able to sell to the suppliers. The travel industry is fertile territory for a cognitive approach and there will be many competitive services that will provide evidence-based advice to travelers.

Transportation and Logistics

Transportation and logistics companies face stiff competition, regulatory pressures, and danger from man-made and natural causes, from terrorism to tornados. Keeping the infrastructure safe is an ongoing concern. In addition, there is a need to identify patterns of customer behavior that may foreshadow new revenue opportunities. Logistics firms were among the first to optimize route times with such tricks as minimizing left turns in cities and improving yield with highly optimized hub and spoke terminals. The rise of sensor technology and GPS tools has uncovered further efficiencies, but we are now on the cusp of a new, smarter industry as cognitive computing technologies are applied across the board.

Cognitive Computing Opportunities for Transportation and Logistics

Many changes in technology are transforming the transportation and logistics industries that will be helped by processing and managing complex data. The first change is the rise in the use of sensor data that needs to be interpreted in near real time to identify opportunities to improve efficiency and safety. The second change is the capability of cognitive computing models to provide diagnostic and preventative maintenance recommendations. These recommendations will help make fleet operations and maintenance more effective. For example, these recommendations will help managers schedule preventative maintenance while minimizing disruption.

CSX, a 185-year-old transportation and logistics firm based in Jacksonville, Florida, has implemented this type of system. With more than 21,000 miles of railroad track, connecting virtually all the population and manufacturing hubs in the United States, CSX links more than 240 short line railroads and 70 water ports. The company replaced a manual-intensive track inspection system that required 600 road masters and track inspectors to record track condition information on paper. This information was then manually input into a system for analysis and reporting. The replacement system, called an Integrated Track Inspection System (ITIS), was developed by CSX leveraging analytics technology from SAP. ITIS replaced the manual system with more functionality and mobile access to recording and predictive analytics tools.

CSX and SAP are also developing a complementary planning system that uses natural language processing and sentiment analysis of unstructured customer feedback. Combining data from these systems into a single corpus with data about traffic patterns and sales data will enable CSX to identify new revenue opportunities in a continuous learning environment. As CSX increases its use of sensors to provide real-time data, integration of these systems will create more opportunities to make the lines safer and able to continuously learn from results. This learning will result in new best practices to improve the efficiency of CSX's operations.

Telecommunications

Telecommunications providers live and die by performance metrics that may be easy to measure but difficult to manage. Customers of these providers are often large companies that resell services to large companies of managed service providers. To be successful they have to provide a predictable level of service. They are often required to provide service-level agreements (SLAs), which specify required performance levels for service delivery. If the telecommunications vendor cannot deliver the service, there are often financial penalties. It is incumbent upon the provider to constantly monitor and manage performance and to demonstrate compliance with the SLA or quickly identify and rectify any subpar performance. Telecommunications providers have matured from providing voice communications channels and basic access to relatively static data, to streaming video to consumers on home and mobile devices, and responding to consumer demand for data on the latest mobile application. As the variety of services offered by telecommunications providers increases, so does variability in demand.

The requirement for continuous performance monitoring, which may have subsecond response time mandates, has led to the deployment of sensors and probes at the edge of the network that can give a real-time view of the actual service level available to the customer. Demand may change due to a variety of events ranging from routine maintenance to a sudden surge in demand because of natural disasters.

Cognitive Computing Opportunities for Telecommunications

Collecting all this data, even in real time, is the easy part. The hard part is to identify conditions that indicate an impending change in demand in time to reconfigure or reallocate services to ensure ongoing SLA compliance based on weak signals—patterns that have traditionally escaped the notice of even well-trained and experienced network engineers. This is where the benefits of cognitive computing come into play. The problem for telecommunications companies is to evaluate enough historical data to discover and understand patterns and causality, while evaluating signals from the environment about impending events that may trigger demand change.

Hitachi Data Systems has developed a solution aimed at helping telecommunication managed service providers monitor and manage real-time data using a combination of machine learning algorithms, proprietary and open source intellectual property, and third-party offerings integrated through APIs. Hitachi uses a component library built from historical customer data including spatial-temporal event detection, complex event processing, event extraction from unstructured data, and root cause analysis to augment the machine learning algorithms. The system continuously monitors current performance and compares it with historical

performance. The system also analyzes unstructured data such as social media streams (weather emergencies or a popular television event that is about to happen) that can impact network performance. By combining real-time data analytics with unstructured data analytics, the system can anticipate changes in demand based on patterns. With continuous learning at its core, such a cognitive computing solution could alert network engineers of impending demand, or even dynamically adjust capacity proactively to prevent a crisis.

Security and Threat Detection

Commercial network security is a concern for business continuity and general risk management in virtually every industry today. Networks, websites, and applications in the cloud are all attractive targets. Cyber-terrorism attacks made for commercial gain or simply to exploit vulnerabilities to demonstrate the proficiency of the attacker are on the rise and show no signs of abating. Even constant vigilance with conventional technology is not enough to keep up as attackers employ more and more sophisticated approaches to theft and disruption.

Cognitive Computing Opportunities for Security and Threat Detection

Following are three big drivers for adopting cognitive computing for threat detection:

- The speed at which new threats are developed
- The speed with which damage can be done before an attack can be controlled
- The complexity of networks that are getting beyond the capabilities of conventional systems and network managers to protect

In the past, as new threats were detected, new rules were rolled out to network administrators or individuals with subscriptions to security and antivirus packages. The delay between detection and updates could be hours, days, or weeks.

Fortunately, machine-learning solutions can monitor network access points continuously and compare current activity to historical activity while looking for anomalies—without being told what to look for. Instead of waiting for an update, the system can highlight unexpected activity patterns and even take actions such as quarantining data and network segments while an operator evaluates the situation. For a false positive (an anomaly that simply represents a new but safe activity pattern) the system can learn that the new pattern is benign and update its own knowledge so that future occurrences will be recognized as the new normal and not register as a threat.

The Cognitive Threat Analytics solution from Cisco is an early entrant in this market. It uses machine-learning algorithms to analyze traffic through a secure

gateway and look for symptoms—anomalous behavior—without concern for the method of attack. This eliminates the old loop that required threat identification as a first step. Cisco can build a corpus of normal behavior patterns by analyzing the activities of individual users and larger groups of similar users. When an unexpected pattern is found to represent a new, benign activity, the updated corpus can be made available immediately to all users of this cloud-based service. Looking at behavior within the network without the bias of rules that were constructed based on previous threats allows the system to learn based only on relevant evidence.

Other Areas That Are Impacted by a Cognitive Approach

Although several industries that can be helped by cognitive computing have been mentioned, many others are good candidates. In some of these areas projects are already underway. Other markets will adapt continuous learning solutions in the coming years. This next section indicates which areas will be impacted by a cognitive approach.

Call Centers

The call center is a cross-industry function that is critical to the reputation and management of an organization. The call center staff is required to have deep knowledge of products and customer issues. However, call centers have notoriously high rates of personnel turnover. When highly skilled staff members leave, their knowledge and best practices leave with them. There is enormous pressure to know intricate details about products and services and to provide the "next best action" to retain customers and sell them other products and services. In addition, call center agents must understand and comply with compliance requirements for their industries.

Cognitive Computing Opportunities

A considerable amount of data can be applied to creating a cognitive computing solution for call centers. Structured data exists in customer support databases. A considerable amount of data is available in notes and documents related to customer interaction and recommendations that can be added to a corpus for a call center application. Over time, the machine learning process can provide guidance for best practices for addressing customer issues. An NLP interface enables a customer support agent to determine next best actions. In addition, customers can interact directly with an online system to determine solutions without long waits on call center phone lines. Eventually, many inbound tasks—getting input from the customer—should be automated using NLP and hypothesis generation. The refined query can be handed off to a human for action or

may be handled directly by a cognitive call center application. (The system may determine whether the caller would prefer a human response by asking or by evidence from previous experiences with the same or similar customers.)

Solutions in Other Areas

A number of other areas exist in which work has started on creating cognitive computing solutions. These are all areas in which a large amount of both structured and unstructured data exists. Some promising areas include:

- **Financial services**—In a data rich environment such as financial services, it will be possible to gain an understanding of an individual's requirements and the best product offerings. The data will be put in context with a vast volume and variety of data from multiple customers. The cognitive system can learn based on patterns of success to provide best next actions.
- **Legal applications**—The legal industry is heavily focused on unstructured documents that include the details for discovery and compliance. This data comes from records ranging from e-mails to tweets to clinical trial results, which must be kept for years and made available upon demand. These legal activities may be carried out by in-house counsel or outsourced, but they all require electronic discovery. This often requires significant resources to scour relevant documents and filings that were created in unstructured natural language. Advanced NLP systems and pattern recognition algorithms found in continuous learning systems are ideal for these applications. Today, a common practice is to use the Electronic Discovery Reference Model (EDRM, from a coalition of attorneys, IT managers, and other interested parties). In the future, using a cognitive computing system could simplify the process, and people could also be trained to alert the business to new opportunities (investments that fit a profile, for example) or risks (scenarios that foretell legal action) by combining information from the company corpus with updated sentiment analysis of social media data and news feeds that indicate impending litigation before it is filed.
- **Marketing applications**—Most of the applications for marketing analyze results of existing campaigns or use predictive analytics to anticipate future customer requirements. The opportunity exists to actively monitor the information related to customer and prospect interactions. On the outbound side, message and pricing can be structured as hypotheses that can be tested against a corpus of data about the industry, firm, current clients and prospects, and competitors. A well-trained continuous learning system could evaluate alternatives and help marketers refine messaging and pricing by asking the right questions early in the process. Constant monitoring and updating of the corpus with relevant social media and

news items—using NLP—would add significant value to that process and to the monitoring of public perceptions of the brand. Sentiment analysis is already used in this context; the cognitive advantage would come from intelligent hypotheses and questioning from the continuous learning component.

Summary

Early cognitive computing successes in areas such as medical diagnosis, manufacturing fault prediction, and healthcare research have convincingly demonstrated the potential for continuous learning systems to change the way we think about entire industries. In the next decade, these learning systems will likely be applied in every industry or business functional area that is characterized by a rapidly increasing or changing volumes of domain-specific knowledge, a high concentration of specialty knowledge in a small group of experts that has great value to a broader audience, or ones that are undergoing great transformations where uncertainty—that is, there is not one single right answer in most situations—is the rule.

As the cost to deploy these systems becomes an operational rather than capital expense (for example, ecosystem revenue sharing models for cognition as service offerings) the barrier for entry will be lowered even further and adoption should increase rapidly. The trend to offer all sorts of functions as services has already transformed the SMB (Small/Medium Business) market in areas ranging from expense management to productivity suites to customer relationship management applications. The next wave of enterprise-level cognitive computing applications is on the horizon, and the cognition as a service wave for functional areas is not far behind.

Chapter

14

Future applications for Cognitive Computing

The development of cognitive computing is at the early stages; however, the building blocks to create this new generation of systems are in place. Over the coming decade there will be many advances in both hardware and software that will impact the future of this important technology. So, the future of cognitive computing will be a combination of evolution and revolution. The evolutionary aspects of cognitive computing are foundational technolo- gies such as security, data visualization, machine learning, natural language processing, data cleaning, management, and governance. There will be revolutions in the capability of systems to improve human-to-machine interactions. In addition, some of the biggest revolutions will come in the areas of hardware innovation.

For decades, advances in chip technology were based on increasing levels of component density and systems integration. Although conventional architectures will continue to improve along these lines, fundamentally different architectures are emerging that will have a bigger impact on cogni- tive computing performance. Neuromorphic architectures, which are "brain inspired" and use processing elements modeled after neurons, will have a profound impact on speed and portability. In particular, neuromorphic hardware will bring a new level of performance for scale up and will allow data to be processed closer to the source, including direct processing on mobile devices. Quantum computing architectures, based on properties of quantum mechanics, offer great promise for fast processing of large data sets that are

often found in cognitive computing applications. This new generation of chips and systems will enable demand for context-aware computing to be met. This chapter looks forward to the coming decade and what is coming and what will be possible.

Requirements for the Next Generation

The need to share knowledge has always been a top requirement for large and small organizations. Myriad attempts have been made over the decades to try to create learning systems that could codify knowledge in a way that does not require years of coding and software development. Emerging technologies that speed the capability to manage and interpret data to gain insights are emerging. A number of important innovations will change the way organizations can translate data into knowledge that is dynamic, shar- able, and predictable.

Leveraging Cognitive Computing to Improve Predictability

Advanced analytics is going to be integrated with cognitive solutions. As cognitive computing matures, companies will find more automated methods of capturing and ingesting massive amounts of data to create solutions. As the corpora of data expand with more experience, it will be possible to incorporate advanced analytics algorithms to a corpus or subset of available data for analysis to determine next best actions or to correlate data to find hidden patterns. This will require a set of tools that can also automate the process of vetting data sources to ensure that data quality is at the level it needs to be. After analysis has been completed, the results can be moved into the cognitive system to update the machine learning models. This will be part of the process of ensuring that a cognitive system can take advantage of the wealth of knowledge and expertise to make better decisions.

The New Life Cycle for Knowledge Management

In a sense, there will be a new life cycle of knowledge management. You begin by creating a hypothesis for the problem you want to solve; you then ingest all the data that is relevant to that problem area; and then vet the data sources, cleanse them, and verify those sources. You train the data, apply natural language processing (NLP) and visualization, and refine the corpus. After the system is put into use, the data is continuously analyzed with predictive analytic algorithms to understand what is changing. Then the process starts all over again. This life cycle from hypothesis through Big Data analytics creates a sophisticated and dynamic learning environment (see Figure 14-1).

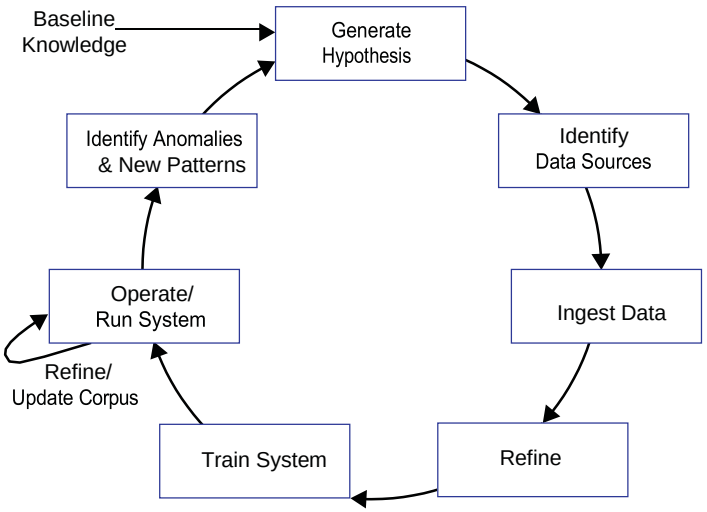

Figure 14-1: The life cycle of knowledge management

Creating Intuitive Human-to-Machine Interfaces

The most sophisticated applications in the first generation of cognitive systems rely heavily on a natural language interface. NLP will continue to be the foundation of how we interact with cognitive systems. However, there will be additional interfaces available for use depending on the nature of the task. For example, there are times when the interface needs to provide visualization so that the researcher can determine where a pattern exists that requires additional exploration. If a biotech researcher is trying to determine the affinity between a disease molecule and a potential therapy, visually detecting patterns will speed the development of a potentially powerful new drug. Other interfaces are also beginning to emerge. For example, improvements in voice recognition technology that can detect emotions such as fear through detection of hesitance could be useful in guiding a user and system through a complex process. When the voice indicates that the instructions are unclear, the system will react with a new explanation. Over time, the system could begin to create new sets of directions that are clearly for a majority of users. A voice recognition system could be helpful in working with the elderly. If the system can detect panic or evidence of a stroke through slurred speech and other cues, it could send help to the elderly individual living at home.

One of the most intriguing experiments with visual interfaces is from an experiment called BabyX developed at the University of Auckland at the Laboratory for Animate Technologies. It is creating "live computational

models of the face and brain by combining Bioengineering, Computational and Theoretical Neuroscience, Artificial Intelligence and Interactive Computer Graphics Research," according to the University's website (http://www.abi. auckland.ac.nz/en/about/our-research/animate-technologies.html). The University explains the BabyXproject:

> *BabyX is an interactive animated virtual infant prototype. BabyX is a computer generated psychobiological simulation under development in the Laboratory of Animate Technologies and is an experimental vehicle incorporating computational models of basic neural systems involved in interactive behaviour and learning.*
>
> *These models are embodied through advanced 3D computer graphics models of the face and upper body of an infant. The system can analyse video and audio inputs in real time to react to the caregiver's or peer's behaviour using behavioural models.*
>
> *BabyX embodies many of the technologies we work on in the Laboratory and is under continuous development, in its neural models, sensing systems and also the realism of its real time computer graphics.*

The laboratory researchers developed a visual modeling technique that enabled programmers to build, visually model, and animate neural sys- tems. The language being developed is called Brain Language (BL). Armed with this language, researchers can interactively work with the simula- tions and model new behavior. To gain some insights into the potential for this type of interface, you may want to view some of the videos of BabyX (http://vimeo.com/97186687).

Requirements to Increase the Packaging of Best Practices

Most cognitive computing applications are based on custom projects in collaboration with subject matter experts. As with any emerging technology sector, pioneers often blaze their own trail. Over time as there are more and more implementations, it will be possible for these results to be codified into patterns that can be used with other projects looking to solve similar problems. Initially, there will be a set of foundational services that developers can use. However, over time there will be a set of packaged services that has been proven through multiple uses by organizations in similar industries. In a sense there is a corollary with what is now thought of as a packaged application. The difference is that a traditional packaged application is a black box. The user can change data, add rules and business process, but the application itself is sealed from the user.

In a packaged cognitive system, there is a level of transparency. First, it will be critical to understand the assumptions and hypotheses that are built into models as well as the source of the data in the package. In this way, a user could use a subset of the package if the use case is different. There will also be packages that are ubiquitous best practices that will become industry standards. This will

have many varied uses for these packaged cognitive applications from training new professionals in a complex field to creating new cognitive applications within a few months rather than a year or more.

Technical Advancements That Will Change the Future of Cognitive Computing

We have made it clear through this book that we are at the early stages of the evolution and maturation of cognitive computing. Many of the foundational technologies are already in place. However, we still require the evolution of other technologies to get to the predictability and repeatability needed to make systems easy to create and manage. Speed of learning is perhaps the area most in need of innovation. Real-time processing is at the heart of fast learning. On the software side, data has to be analyzed in real time especially to process information in data-rich environments such as video, images, voice, and signals from sensors. These systems will require better clarity, and faster identification of the meaning of these signals. The key to success is improving the time to meaning, not just data acquisition. For example, getting to the point at which the system recognizes and understands the actions of a specific individual within a video fast enough to respond in a threat situation enables more meaningful outcomes. Identifying and then processing the relationships between data in real time can help establish context.

Future innovations in software and hardware will transform what today is complicated and time-consuming for data analytics. Today, to gain this level of expertise requires a lot of manual effort. In the future, machine learning will become more abstracted into the fabric of the development environment. It will be possible to interact with a system in real time as a pattern or connection is detected from the data. This evolution is required as we move from data to information to knowledge. The faster we can process knowledge and understand patterns and context, the sooner we can begin to make discoveries that change the pace of innovation and discoveries across markets and industries.

What the Future Will Look Like

What will a cognitive system look like in the future? The changes in technology that are needed will not happen all at once. Rather, there are two time horizons to consider: the first five years and then the long term, moving out to the next decade. There are three facets that will define the future of cognitive computing: software innovation, hardware transformation, and availability of refined and trusted data sources. All these are predicated on the development of standards. Before discussing the type of technology that will be at play in the future, take a quick look at what might be expected in five years and then in the more distant future.

The Next Five Years

There will be considerable change over the next five years. One of the most significant changes will be in the number of well-defined foundational and industry-specific components that are based on foundational and industry-specific elements. For example, there will be a service that can automatically build ontologies based on deep analysis of text in natural language within a domain. Today, the process requires a lot of manual intervention and consensus building. Although people are still expected to have the last word for the foreseeable future, their participation in the process will diminish as ontology building software learns from experience.

Within the travel market, there will be services that might automate the process of building correlations between destinations, predicted weather patterns, and social media data. As interfaces become standardized, it will be possible to automatically link these services together. These functional services will likely be packaged together into workloads using emerging container standards to enable cost-effective cloud deployment.

There will be a series of well-defined services for everything from ingesting specific types of data to analyzing that data in real time and providing visual interfaces that indicate where patterns exist and what they mean. Natural language interfaces will enable users to select the type of interfaces that are most appropriate for the type of analysis being done. One of the newer approaches gaining traction is to move from simply reporting or displaying data to delivering a "story" that explains the data using a narrative interface. Telling a story about how elements of data are related to each other based on wanted outcomes will bring the best clarity in certain situations. Today, many applications rely on graphs and charts to tell a story about the meaning of data.

In other situations when a customer asks a question on a retail selling site, they will be shown a set of products that best matches a keyword. However, what if that engine has a better understanding of the context of consumer's intent? The consumer looking for a sleeping bag might be ready to purchase sleeping bags for the whole family. The successful site would help build a story just for you, based on your needs, your aspirations, and even your financial constraints. Now you have moved from being shown a single item to being shown a story of your future engagement. The world of camping has many nuances, and there could be other products and services that a smart retailer could offer to the consumer. This new type of system will be as cognitive as you allow it to be. Trust and the ability to grant permission for one engagement or over the life of a relationship between a consumer and vendor will be key to the future of engagement.

Imagine a scenario in which the traveler is equipped with a cognitive trip system. The system knows your destination, your preferences for the way you drive, the gas stations along the way, the health of your car, your preferences in food, and the type of hotels you like to stay at. With the right level of input

and the right level of security, that system could make your reservations, alert you to alternative routes well in advance, and indicate that you should stop to get your car repaired—although a really smart system might advise you not to leave home on a trip if it can't be completed without predictable auto repairs or maintenance. The system could alert a store that has an item that you need and could even negotiate a price and alert you to the pickup time.

You can apply the same approach to how you deal with your insurance company. You may negotiate a deal with that company based on your habits that will be tracked by a device you wear (assuming that you have granted permission). The information you provide to your insurance company will be aggregated with hundreds of thousands of other insured customers to gain an understanding the level of risk. This could either drive costs down as insurance companies better understand actual risks or create a sharing economy pool of people with similar profiles. It could also lead to new government policies as cognitive systems for human capital management intervene to prevent catastrophic loss for the uninsurable, which could lead to unrest because those things can't be hidden in a transparent, connected society with access to communications and cognitive tools.

Looking at the Long Term

As the individual technologies that have been discussed continue to mature, we will see them built into the fabric of cognitive systems or platforms rather than assembled from discrete components. Learning will happen in real time and increasingly be influenced by gestures, facial expressions, and seemingly off-hand comments. These systems will, therefore, automatically understand context from events and data from yesterday or from 5 years ago. These systems will store and continuously analyze all social media history in a deep way. Armed with this level of analysis, the cognitive system will anticipate what you might do next and understand why.

Keep in mind, even in ten years permission-based interactions will be the rule. However, there will be more automated techniques that assume your permission level and then ask for confirmation. The system, in fact, is built by analyzing patterns across millions or perhaps billions of interactions. The level of permission the consumer allows will determine the interaction and level of security in this environment. The optimal system will act in the background, making suggestions or recommended actions when necessary but remaining silent most of the time. In essence, you will be dealing with an advanced automated agent that will allow you to create a persona for yourself that is comfortable for you. The agent software gets to know your preferences and personality over time based on the data that you provide directly and the learning system behind it that makes assumptions based on accumulated information. The system will be designed with a set of rules of etiquette based on how humans are comfortable interacting with machines.

In essence, this is the personal digital assistant for the new cognitive era. Rather than the physical device, it may take the form of a representation or personal agent in the cloud with the multiple types of context-appropriate interfaces available at the time of engagement. It could be you and your personal interactions; it could also be the interface to your washing machine. We have already begun to enter an era where the Internet is ubiquitous, but in the coming era, you will always be connected—unless you choose to disconnect. Depending on the situation, the interface is a natural language, a gesture, or a physical action. The cognitive system captures the nuances of your interactions and changes its interaction based on your changing needs and conditions. The system is constantly learning from your behavior and activities behind the scenes. It modifies its actions based on the learning over time. This technique will be widely applied to everything from traffic patterns in a city to security infrastructure.

As more devices with embedded sensors become ubiquitous, the level of data and actions will explode. Professional athletes will be equipped with sensors that know if they suffer a concussion even before a physician examines them. That same type of sensor-based device could warn a construction worker of an obstacle that he should avoid.

These types of cognitive systems will have potential to break down barriers for humans. A sensor-based device with a sophisticated interface can provide a different level of interaction with people who have trouble interacting in social situations. The system is nonjudgmental. Individuals on the autism spectrum could be helped by a system that learns the best ways to interact and has the potential to open lines of communications that have been blocked. The cognitive system adapts to the communication style most effective for different individuals with different disorders. It could be helpful for elderly suffering from Alzheimer's disease.

The most significant change in the coming decade is that cognitive computing will become part of the fabric of computing. Therefore, it will have a profound impact on many industries and many of the tasks that humans do. Machine learning and advanced analytics will be built into every application. Natural language interfaces will continue to be the foundation of how we interact with systems. Eventually, natural language processing will become a utility service rather than a separate market.

Emerging Innovations

What will it take to get from individual handcrafted systems to the state in which these technologies are deeply embedded in everything you use?

A number of existing technologies that are instrumental in cognitive computing are going to evolve over the next 5 years. That will improve the capability

of systems to be created faster with greater capability to solve complex issues. This section discusses the key technologies.

Deep QA and Hypothesis Generation

Today, Deep QA—which may require a system to generate a series of probing questions for a human to answer for the system to navigate multiple levels of meaning—is rare in practice. In IBM's Watson it is used interactively in a conversational mode with experts to refine their quest for possible answers in complex domains. For example, a doctor may describe a set of symptoms relevant to a patient, and Watson may ask questions that help it to narrow the range of possible answers or increase confidence in one or more diagnoses. It may ask if a particular test has been ordered or ask for more details about a family history. Deep QA requires the system to keep track of all the information that has been provided in previous answers for a session, and only ask further questions when the human answer can help it improve its own performance. It will evaluate the possible answers it may give and assign a confidence level in each, but look at what additional evidence could change that confidence to decide whether to ask for additional information.

If the learning experiences of a lot of systems that answer related questions are shared, that body of knowledge about the process could become a reusable pattern across a domain. In healthcare, for example, there may be enough deep QA analysis to discover the optimal treatment for a specific type of skin cancer because enough data exists— when aggregated—and enough analysis has been done on that data that has been vetted by the best experts in the world. Over time, some hypotheses will have been proven and accepted, so the same query asked at a later date may require less analysis and fewer generated hypotheses as the corpus matures. We may never run out of problems to solve, but for the most part we will see the process of problem solving in complex domains begin to coalesce around cognitive computing. Much like the scientific method guides discovery in the natural sciences, discovery through deep QA and hypothesis generation and testing is likely to become the default approach for many professional disciplines.

NLP

Advances in NLP have been dramatic in recent years as evidenced by the capability of IBM's Watson to derive meaning from unstructured text under conditions of intentional difficulty. (The QA format of Jeopardy! presents "answers" that may be ambiguous or require context or familiarity with idiomatic speech, and contestants must determine the meaning of the answer before identifying the most appropriate question as a response.) This format is challenging for many humans, but Watson had little difficulty finding the relevant meaning,

or alternatively, recognizing when it had low confidence in its answer. The Watson team prepared for the event by studying the way Jeopardy! writers used speech in the past. Those lessons will be valuable as IBM and others extend NLP technology to handle more general cases of slang, colloquialisms, regional dialogues, industry-specific jargon, and the like. A lot of the training is involved with understanding the context of language. NLP systems or services must understand state and conditions that may have been set previously.

Automating translation between natural languages that capture deep meaning remains a difficult problem for NLP. Vocabularies may be mapped from one language to another with reasonable precision (English to French, for example), but natural language communication involves strings or sentences built in to paragraphs and stories that may have explicit and implicit references to meaning expressed in other strings, paragraphs, or even historical references. A key NLP innovation—assuming some common constructs among languages that map to the same underlying deep structures—would be the identification and emulation of the manual process used by expert human translators to discover rules or heuristics they may be applying unconsciously. Analyzing different well-respected translations of books, for example, to identify commonalities and different interpretations, will provide insights into these rules. Today, even some shallow language analysis is so processor-intensive that mobile systems have to send the sentence or string from the device to a cloud-based service before responding. Enabling deep translation on the fly for more than simple statements on mobile devices will require these breakthroughs, or alternatively more powerful NLP chips on the devices themselves.

Cognitive Training Tools

It is tedious and time-consuming to build a corpus today by training a system based on ingested knowledge. A lot of trial and error and human judgment are involved for every new corpus. Much of the training work that is human-intensive today will become automated as we use current generation cognitive computing systems to examine the process to help build better tools. Similar to the way every generation of high-precision manufacturing tools were built with the previous generation of less sophisticated tools, cognitive computing technology will be used iteratively to discover ways to improve the process of building cognitive computing solutions.

Bias in training is one of the most important issues that will have to be addressed. With a lot of unstructured data and no standards to understand that data, experts make judgments based on their own experiences, which are biased because most have never seen the entire universe of possible interpretations. (Even in narrow medical specialties, for example, the most experienced practitioner has rarely seen every possible set of symptoms or treatment outcomes.) However, they aren't even aware of the bias they are bringing to the situation.

In the future, as cognitive tools become more powerful and apply more cognitive learning, it will be easier to determine the source of a bias and point that out to the expert.

Data Integration and Representation

Today, connectors, adapters, encapsulation, and interfaces are used to deal with complex data integration. Although this is sufficient if you have a good understanding of the data sources and they are well vetted, it is a different matter when you begin to bring thousands of data sources together. Data integration needs to be automated with a cognitive process so that the system begins to look for patterns across data sources and detect anomalies to see if they represent new, important relationships that were unknown before or problems with a data source being inconsistent.

You saw that ontologies can codify common understanding of complex relationships within a domain, but implementing an ontology is actually a crutch. In a perfect world, a cognitive computing system would not need an ontology because it could dynamically build its own model of the universe by understanding the relationships and context—but that works only if there is enough data and experience and it can process and understand fast enough. Today, we create ontologies so that performance is acceptable with current system constraints. If you could do that processing on the fly, you wouldn't have to predetermine what the ontology would be; you could discover an ontology rather than building one. With sufficient processing power, an ontology would actually be a system state during execution. It would be generated only on demand if it were required for auditing purposes, perhaps to understand why a decision or recommendation was made.

Emerging Hardware Architectures

Hardware innovations in both the short and long term will have a dramatic impact on the evolution of cognitive computing. Today, it is primarily traditional hardware systems that are used to build cognitive systems. Although parallel structures are used, these systems are still general purpose von Neumann architecture computers, in which all the actual processing takes place in registers within central processing units (CPUs) (or in adjunct processors such as graphical processing units [GPUs]). The real breakthroughs that are on the horizon over the next several years include major changes in chip architectures and programming models.

Complementary to the efforts in software and data architectures, we are seeing two different approaches to hardware architectures evolve. One is based on modeling neurosynaptic behavior (the relationship between neurons and synapses in the brain) directly in hardware. These neuromorphic chips feature many small processing elements that are most tightly interconnected to near

neighbors to communicate much like human brain neurons pass signals via chemical or electrical synapses.

The second promising approach is quantum computing, which is based on quantum mechanics (quantum physics), a branch of physics that explores physical properties at nano-scale. Unlike conventional computers whose fundamental unit of storage and processing is the bit (binary digit) which must be a 1 or 0 at any given time, quantum computers use the qubit (quantum bit), which may be in more than one state at any given time. The next two sections explore the prospects for these competing architectural approaches.

Neurosynaptic Architectures

Why should you look at this new generation of hardware architectures? Simply put, the complexity of identifying and managing relationships between data elements at the scale required for cognitive computing—Big Data—requires enormous computing resources with conventional architectures. Fundamentally, the challenge today is to partition the data effectively to funnel it into an architecture that processes 64 bits of data at a time.

The current basic Intel microarchitecture, for example, used in the Core i7 processor (found in many laptops) and the Xeon family of processors (used in Tianhe-2, currently the world's fastest supercomputer) processes data in increments of 64 bits. Over the past decades, computer scientists have developed elaborate workarounds to compensate for the limitations of hardware. For example, it is relatively easy to add processors to a cluster or system. The individual processors in Tianhe-2 are no faster than those in a modern laptop, but it links together 260,000 of them to harness 3,120,000 cores operating in parallel. The difficult part is to effectively distribute the workload across those similarly architected processors. Some cognitive computing techniques such as hypothesis generation are inherently parallel. Based on the data, it may be desirable to generate hundreds of hypotheses and then process them independently on different processors, cores, or threads.

Another task that would be valuable in a cognitive computing application is real-time image processing in a manner similar to human vision. That also requires mapping millions of bytes of information to look for patterns, which humans do in parallel rather than by breaking up the problem into sequential tasks. For still images, this can take thousands of processors. (The Google experiment mentioned in Chapter 2, "Cognitive Computing Defined," used 16,000 processors just to identify cats.) For video, the problem is much more difficult. A high-definition camcorder typically generates approximately 5 gigabytes of data per minute recording at 30 frames per second. If you want to analyze all the images, you need to analyze each frame and compare it to prior and subsequent frames to find patterns. For example, when evaluating video of a crime scene, detectives look for people whose behavior is not like the rest of the crowd. A

human can do that relatively easily with a single video stream, but when multiple streams are involved, it becomes a daunting task that could be automated with sufficient processing power. For most applications today, it is impractical to do large-scale hypothesis generation and evaluation or real-time video analysis.

Now contrast this first to the neurosynaptic hardware approach. The current large-scale leader in this field is IBM's TrueNorth (developed with funding from DARPA), a neurosynaptic chip with 1 million neuron-inspired processing units and 256 million synapses (connections between the processing units, similar to a computer bus but a lot more powerful and faster). Instead of improving performance by adding additional 64-bit register-limited machines, scaling up with a neuromorphic chip builds in the parallelism because while each neural processing unit executes a single function, it communicates with many others. Like neurons in the brain, they are so physically close and connected that they communicate virtually instantaneously. Test systems have already been constructed with multiple TrueNorth chips yielding a system with 16 M neurons and 4 B synapses.

The underlying principle that is modeled in neurosynaptic chips is Hebb's Rule, commonly simplified as "cells that fire together, wire together," —meaning that neurons in close proximity that fire together (actually in rapid sequence) reinforce learning. This was postulated in Donald O. Hebb's 1949 book, *The Organization of Behavior*, which formed the basis for much of the current understanding of associative learning and the development of parallelized pattern matching algorithms. Mapping the behavior of these human brain elements to fundamental constructs in the hardware architecture provides a natural bridge between the way we look at a problem and the way we solve it, which gives neuromorphic computing great appeal (as "brain-inspired" hardware). In the near future you can expect to see billions of processing units per neurosynaptic chip with trillions of synapses. When these chips are assembled into systems, the result will be a new standard for scalable parallelism that has practical applications for pattern matching and learning in cognitive computing systems.

Commercialization of this architecture will require a new programming model, a sophisticated software development environment and an ecosystem of professionals and companies to create a new industry around this model. Efforts to develop these tools and skills are already underway, but in the immediate future you may see hybrid solutions in which neuromorphic approaches will be combined with conventional computers. Similar to the way the average computer today often incorporates special processors for graphics and sound, neuromorphic chips integrated with a conventional system will enable you to take advantage of conventional programming models for much of the required preprocessing.

Why is this architectural approach so important? The emerging architecture enables you to populate each of the millions of neurons in parallel, rather than artificially constraining you to a 64-bit bandwidth for actual processing. When the data is loaded in these neurons, the chip or system can search for patterns in real time. Applications that now are impractical for conventional systems—for

example, massively parallel hypothesis processing in medicine and scientific exploration or human-like vision processing become feasible. Parallelism without partitioning is a huge advantage for neuromorphic architectures. The acts of partitioning and reassembling results take time and add complexity. Although there are multiple research efforts to build large-scale neurosynaptic chips, the same approach to mimicking neurosynaptic processing is already being commercialized in smaller scale special purpose chip sets for mobile devices. Qualcomm has a production chip set called Zeroth that is intended to capture patterns of human behavior based on the usage of the mobile device to provide context-aware services. This is planned to be put into production by 2015.

The architectures operate in parallel efficiently so that the total power consumption for a unit of work is lower than that of a register-based architecture. This makes these architectures appealing for mobile devices and at scale will reduce the power and space requirements for data centers. Scalability (up and down) and a simple architectural model will make the adoption of neuromorphic chips inevitable for some cognitive computing applications.

Quantum Architectures

The fundamental concept behind a quantum computer is to go beyond a binary, two-state (on/off; that is, 1s and 0s) atomic processing unit to a multistate unit called the qubit. A qubit can have multiple states as defined by the physics of quantum mechanics, including being in multiple states simultaneously (superposition). Conceptually, this will be extremely difficult to popularize because it is beyond the mathematical and scientific knowledge and experience of most of the world's population, but it is the most natural way to process quantum algorithms for learning and discovery. Quantum computers can be simulated using conventional computers by mapping each of the possible states to binary states, but, of course, the performance overhead is significant. For example, in a single conventional 64-bit register, you could represent 2^{64} values (ranging from a string of all 64 0s to 64 1s, or $1.8 * 10^{19}$). In a qubit with three possible values (0,1, or both) 64 qubits could represent 3^{64} values or $3.4 * 10^{30}$, that is, 200 billion times bigger than the binary solutions and impossible to process on a conventional system in anything approaching real time. And in theory, quantum computers can scale without the artificial register restriction, which makes them attractive for massively parallel computations and processing existing quantum algorithms. Like neuromorphic computing, quantum computing will require entirely different programming models, skills, and tools.

Perhaps the most significant barrier to quantum computing is that it requires physical materials to actually be in these superposition states, which requires the processing units to operate at a temperature near absolute zero. That precludes any mobile applications and modestly sized system installations, at least for the time being. Still, the performance potential is too great to ignore. Today, we

are seeing significant research and investment in quantum computing by IBM, Google, and DWave (which focuses exclusively on quantum computing). Google has set up a new effort to build its own quantum computer for AI research with academic researchers in the University of California system while continuing to support the independent efforts of DWave.

The energy, space, cooling, and mathematical skills requirements will keep quantum computing from becoming mainstream in the next decade. Although neuromorphic architectures are expected to grow in popularity quickly and be more pervasive at all levels than quantum computing, quantum architectures will continue to attract research funding because it is well understood that a few breakthroughs could lead to fundamentally faster supercomputers.

Alternative Models for Natural Cognitive Models

Although neuromorphic and quantum computing architectures are based on approaches to established science—neuroscience and quantum mechanics, respectively—that have active research communities in place, they are being challenged by a new approach pioneered by Jeff Hawkins. Hawkins, who changed the way we think about mobile devices when he introduced the Palm Pilot, has an alternative view of human learning. He founded the Redwood Center for Theoretical Neuroscience in 2002 to support research into a layered model of learning based on the functioning of the neocortex. His company Numenta is building applications and an infrastructure for cognitive computing based on his theory of the way the brain stores, processes, and retrieves information about events. His approach is based on the role of the neocortex in human memory as the central organizing principle for computer architecture rather than neurons and synapses. Although it is too early to evaluate the potential for this approach, its Grok for Analytics machine learning anomaly detection product has already demonstrated that it may be useful even if the theory behind it isn't ultimately adopted by the scientific community at large.

Summary

In the future, cognitive systems will be defined as an integrated environment, which means that software and hardware will work as though they are a single integrated system. This new architecture will scale up and down depending on the use case. For applications such as smarter cities and smarter healthcare, the high-end architectures will enable machine learning in near real time. With personal devices and sensor-based assistants, hardware embedded at the end points will provide processing at the source. This convergence between hardware, software, and connectivity will provide the platform for a huge flood of new use cases and applications for cognitive technologies.