CHAPTER 5

Processing and Adding Vibrancy to Sensor Data

Intelligence is the ability to adapt to change.

—Stephen Hawking, Physicist

The integration of sensors into many aspects of daily life will generate enormous volumes of data, and that will only increase as progressively more sensor output start to feed into "big data." The term "big data," which has received significant attention in recent years, is used to describe the voluminous amounts of unstructured and semi-structured data companies, governments, institutions, and individuals generate each day using information and communication technologies (ICT). Sensors are expected to be one of the largest generators of data, especially as the Internet of Things (IOT) gains traction in our everyday lives. Big sensor data will leverage capabilities such as cloud infrastructures for data storage, processing, and visualization. Access to this data will become pervasive, especially through mobile devices. We will also be able to combine other sources of data with sensor data in innovative ways to reveal new insights.

Intelligent processing of data and context-based visualization are critical to delivering meaningful, actionable information. Presentation approaches should strive to engage users by adding vibrancy to the data and allowing users to interact with the data in collaborative ways. Figure 5-1 shows the essential elements of data literacy. Collectively, these elements play an important role in helping to develop the kind of understanding that enables us to effectively utilize sensor data. In this chapter, we will look at the various methods to process, interpret, and display sensor data to end-users.

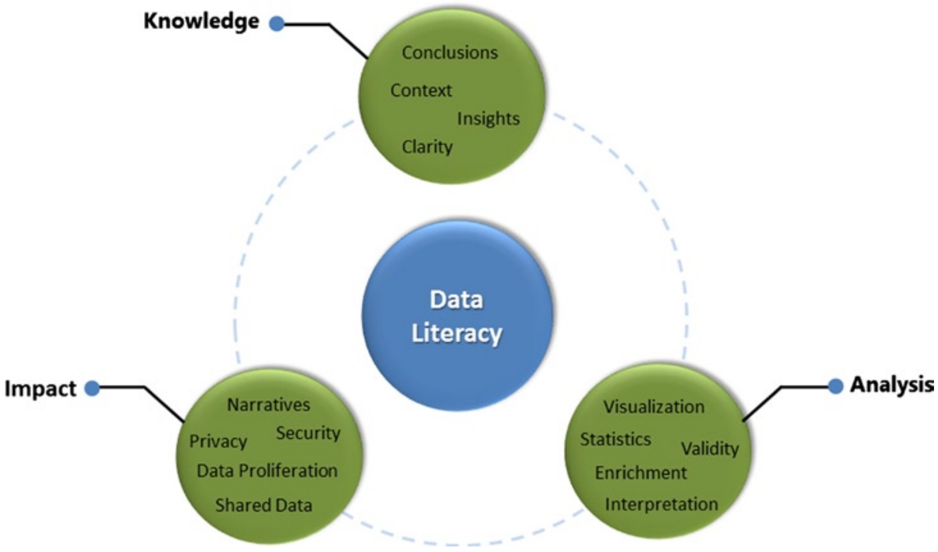


Figure 5-1. *The key elements in data literacy*

Data Literacy

Sensor data has value only if it allows us to do something useful with it, and we can't do that until we understand what the data is telling us. This process of understanding data is known as *data literacy* (Hart, 2011) and, from a knowledge perspective, it generally includes the ability to:

- correctly interpret visual representations of data, such as graphs
- accurately analyze the data and draw correct conclusions from the data
- utilize other datasets to add context
- know when the data is being interpreted in a misleading manner
- identify data that is inaccurate due to complete or intermittent sensor failure

The *New York Times* estimates that the United States will need 140,000 to 190,000 additional workers with strong data analytics expertise to deal with the abundance of data in areas ranging from science and public health to sports (Lohr, 2012). Data literacy needs to be the domain of more than just a few individuals; it must be embedded within an organization's culture to ensure that decision-makers understand data-driven insights and can act on them (Shelton, 2012).

Data literacy will become a more pervasive requirement, not only for specialists but also among the general population. Citizens will need to develop an understanding of mobile and ambient sensing and how this form can impact privacy, security, and risk. Casual technology users typically do not understand the security risks of data sharing. Therefore, it is increasingly important citizens become sufficiently data literate to grasp the implications of sensor-based data collection and the proliferation of shared data (Shilton et al., 2009).

Data literacy has three central themes, as shown in Figure 5-1: understanding the process of data analysis, understanding the impact of data, and gaining meaningful knowledge from the data. Clive Thompson, in his article *Why We Should Learn the Language of Data,* points out that many debates in the public domain, such as climate change or public health issues, often devolve into arguments over what the data means. In this context, he suggests that the new grammar is, in fact, statistics (Thompson, 2010).

Data literacy also enables individuals to utilize sensor data and other supporting data sources to provide contextualized observations and answers. Developing context may be as simple as knowing a sensor's expected range of measurement, which lets a person infer that measurements outside this expected range can indicate a malfunction. Alternatively, a clinician might be able to ignore a remotely acquired high blood-pressure reading if another data stream from a body-worn kinematic sensor shows that the patient engaged in physical activity before the measurement was taken, temporarily raising his blood pressure. Data structures, which provide contextual wrappers for healthcare-related sensor measurements, have also emerged in an effort to provide context for sensor measurements (Gonçalves et al., 2008). We can gain rich and meaningful insights into our health if we can ask the right questions from the data and understand the answers with a high degree of clarity. Without the clarity afforded by data literacy, we end up with "numerical gibberish, or data salad" (Bradshaw, 2012).

The Internet of Things

The IOT is a somewhat amorphous concept that continues to evolve along with the increased connectivity of a broad spectrum of technology devices. Early notions of the IOT focused on human-centric, Internet-enabled computing devices, such as smartphones, tablet devices, laptops, and so on. However, the IOT has grown to embrace a rich eco-system of devices including sensors, smart clothing, consumer electronics, utilities meters, cars, street lights, advertising signs, buildings, and more, all of which can now be connected to the Internet. One of the most salient definitions of the IOT comes from the US National Intelligence Council (Swan, 2012):

> *The "Internet of Things" is the general idea of things, especially everyday objects, that are readable, recognizable, locatable, addressable, and controllable via the Internet—whether via RFID, wireless LAN, wide-area network, or other means.*

The number of connected devices already exceeds the number of people on the planet, and Cisco estimated that the total number of connected devices will exceed 50 billion by 2020 (Cisco, 2011). A key driver of the IOT is sensors: discrete sensors, such as those for environmental monitoring; body-worn sensors, such as ECGs/EKGs; and sensors embedded into devices, such as accelerometers in smartphones. In fact, many devices feature multiple sensors. For example, a personal activity monitoring device may combine a kinematic sensor, such as an accelerometer; a physiological sensor to read heart rate; an ambient sensor to get the temperature; and a GPS to find the location. All these data streams can then be connected to the Internet via a smartphone.

Platforms like Xively (`https://xively.com/`), which allow hobbyists and companies to intuitively collect data from Internet-enabled devices, including sensors, are beginning to emerge. Xively supports secure sharing of data, builds collaborations around the data, and provides tools to visualize the data across multiple platforms. The platform has been applied in home environmental monitoring applications to gather and view sensor data streams, including barometric pressure, carbon monoxide, and temperature. New, innovative sensors for home environmental monitoring, such the Air Quality Egg (AirQualityEgg, 2013) are leveraging Xively to provide IOT capabilities. Popular platforms such as Arduino and Electric Imp (designboom, 2012) also provide the capabilities to connect sensors either directly to the Internet or via smartphones.

Realizing the potential of sensors to become the "*finger of the Internet*" will take a number of years. Several key factors still remain to be addressed in order to fully achieve this goal. These factors include reducing the cost of sensors; increasing battery life; developing reliable and powerful energy-harvesting capabilities; building more robust wireless data transmission capabilities; making wireless backhaul coverage ubiquitous (3G and 4G, for example); building data analysis and visualization tools capable of dealing with large scale, high-frequency sensor data streams; and, finally, understanding how to convert data streams into meaningful, real-time, personalized recommendations with appropriate context.

Beyond human-centric sensing utilities, machine-to-machine (M2M) applications, commonly known as ubiquitous or pervasive computing, will be a key driver of Internet-enabled sensing. It is estimated that by 2020 there will be 12.5 billion M2M devices globally, up from 1.3 billion devices in 2012. They will deliver a wide variety of applications for, among others, environmental monitoring and control (water management, smart cities, weather event monitoring, and more). Over time, M2M connectivity will change our daily experiences; altering how we interact with the world around us, from our home environments to the places in which we conduct our daily lives.

Sensors and the Cloud

Cloud computing has become one of the most active areas in information technology; it is already starting to transform how businesses manage and use their computing and storage infrastructure. The cloud-based model affords flexibility and scalability for computing, storage, and application resources, optimal utilization of infrastructure, and reduced costs. Cloud computing has the potential to provide storage, processing, and visualization capabilities for sensors and sensor networks, as shown in Figure 5-2. The sensors can be discrete or part of a geographically distributed network, and feature highly dynamic data throughputs. This cloud-based integration is able to accommodate dynamic loads and sharing of sensor resources by different users and applications, in flexible usage scenarios.

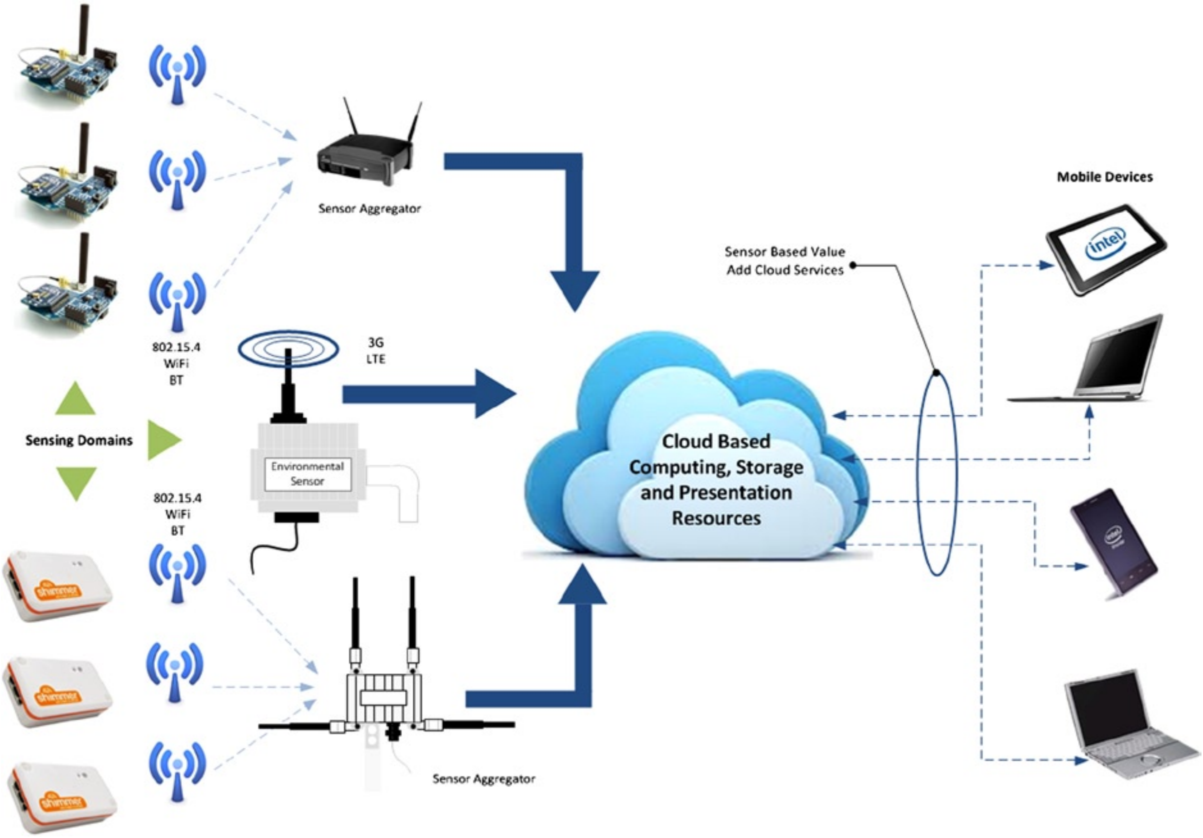


Figure 5-2. *A sensor cloud architecture*

Smart Cities is an application domain for which the use of sensor clouds has been proposed (Mitton et al., 2012). For example, a citywide environmental monitoring system would require significant computational and storage resources during an exceptional weather event, but would return to standard requirements at the end of that event. A cloud-based approach has the potential to reduce overall cost in a sensor deployment, as it can easily support elastic consumption of resources. Clouds are typically based on usage models, allowing application developers to optimize frequency and resolution of data against cost. Commercial software as a service (SaaS) solutions for sensor data are already starting to emerge, including the SensorCloud system from MicroStrain (MicroStrain, 2013) and sensorcloud from temperature@lert (temperature@lert, 2012).

Vast amounts of sensor data can be processed, analyzed, and stored using computational and storage resources of the cloud. Once the sensor data is stored in the cloud, novel applications based on aggregated data sets can be created from different organizational resources or from crowd sourcing. Essentially, the cloud can be used to create virtualized sensors that can be accessed by any user or application. This breaks down the siloing that currently exists with many sensor applications. There are number of efforts to realize this goal, including Sensor-Cloud from IBM (Yuriyama et al., 2010), Cloud@Home from ANR (Recherche, 2011), and SensorCloud from Aachen University (Hummen et al., 2012).

Users and applications do not need to be concerned about the physical location of sensing resources when using cloud-based functions, because they are essentially virtualized. End users can dynamically provision and group virtual sensors based on need, and terminate them easily when no longer required. New physical sensors can usually be added to the cloud through a registration process. This process includes a mechanism to register sensor characteristics, in a format similar to TEDs as described in Chapter 3, as well as a way to describe the sensor's data type in an XML format, such as SensorML.

Data Quality

The success of any sensor application depends on the quality of the data. Without trust in that quality, the value of the data is significantly undermined, and it's observational, diagnostic, or actionable value is limited. It is important, therefore, to ensure that data quality is an integral part of any sensor application development process. A variety of issues can affect the data quality during the application lifecycle, in all of its phases:

- Sensor system design, development, and validation
- Deployment
- Protocol design
- Data processing and visualization

Some issues can be mediated through careful design of the sensor system. A tightly controlled deployment process or active management strategy can proactively identify issues that affect data quality. Key factors to be considered include data consistency, measurement accuracy, and reliability during data collection, processing, storage, and transmission. The primary objective should be to minimize, or ideally eliminate, both the general and application-specific data deficiencies. A risk matrix can be useful in prioritizing data-quality impacts. In this matrix, priority is normally allocated to both high-impact and high-frequency risks. The rationale for this prioritization is based on propensity for significant impact. These are then followed in order by lesser influences, such as outlier detection. The key factors affecting data quality are outlined in Table 5-1 (Puentes et al., 2013).

Table 5-1. *Factors That Affect Sensor Data Quality*

Factor	Impact
Sensor limitations	*Operational limitations*—such as over-sensitivity to environmental influences. *System-design limitations*—such as data-throughput constraints, processing bottlenecks, reliability of communications.
Calibration error; drift	Incorrect calibration for the required operational range, or frequent recalibration needed to maintain accuracy. Accuracy deteriorates over time.

(*continued*)

Table 5-1. (*continued*)

Factor	Impact
Environmental influences	Performance variation due to temperature, humidity, ingress of moisture, and the like. Degrading of sensor materials. Malicious damage to the sensor or the measurement environment. Damage from wildlife. Damage from vehicles.
Malfunctioning sensor	Sensor ceases to function correctly, resulting in erroneous output.
Incorrect Values	Incorrect sensor measurements can arise due to external influences such as noise.
Unsuitable Protocol	Measurement protocol cannot be utilized correctly due to its complexity. Unclear how to use the sensors correctly. Protocol does not acquire data at required periodicity or range.
Human Influences	Skewed results due to location of humans in the measurement environment. Incorrect use of the sensor: for example, incorrectly attached electrodes a body-worn application.
Pertinence	Data collected is irrelevant or has no utility.
Incorrect installation	Inaccurate sampling.
Trust and repudiation	Inability to guarantee origin of data limits value of data, particularly for diagnostic purposes. Data traceability and robust security need to be established.

Addressing Data Quality Issues

Identifying and addressing quality issues in sensor applications are important. A wide variety of potential influences and impacts can affect the accuracy of sensors readings, but checking manually is extremely tedious and time-consuming and does not scale to accommodate larger sensor deployments. Automated methods represent the most pragmatic approach to monitoring data. Statistical modeling, machine learning, and other data mining techniques can be utilized to identify anomalous or outlier data in real time or during post-processing. It may even be possible to eliminate the outlier or anomaly and replace it with an appropriate calculated value. Some of the common approaches to monitoring data quality are discussed in the next section.

Outlier Detection

Outlier detection, also known as anomaly or deviation detection, is commonly used to monitor data quality and improve robustness in wireless sensor networks. This approach identifies malfunctioning sensors or malicious network attacks from external parties by detecting measurements outside the expected range. A number of methods exist to deal with outliers. Visual inspection can be used to spot erroneous data, which can be removed manually.

Another approach involves plotting data in histograms or scatter plots to help identify outliers. However, these approaches are time-consuming and not very scalable. For greater efficiency, model-based approaches are commonly utilized. Here are five useful approaches (Zhang et al., 2010):

- *Statistical modeling*: This method is based on the normal probability distribution of the data under standard operational conditions. Actual data is evaluated against this model to determine whether it fits correctly. Data identified as low probability by the model is classified as outliers.
- *Nearest neighbor*: This method compares the data value with respect to its nearest neighbors. The distance (such as Euclidean distance or a similarity measure) between the new sensor measurement and previous measurements is calculated. Too much distance from neighbor measurements results in the measurement being identified as an outlier.
- Clustering: In this approach, data instances are grouped into clusters that have similar behavior or characteristics. Instances of data are identified as outliers if they do not fit into a predefined cluster or if they generate a cluster that is significantly smaller than other clusters. Euclidean distance is often used as the dissimilarity measure between two data instances.
- *Classification*: This approach is based on the development of a classification model that is then trained using representative data sets. Once the model has been taught the expected data distributions, it can classify unseen data into either a normal or outlier class. As new data is collected or the operational parameters of the sensor or sensor system change, the classifier can be updated to reflect new instances of normal data. Popular approaches for developing the classifier include support vector machines (SVM) and Bayesian networks.
- *Spectral decomposition*: Principal component analysis (PCA) is used to reduce the number of dimensions within a dataset to those, such as variability, that capture the expected behavior of the data. Data whose components fall outside these structures are considered to be outliers.

When dealing with outlier data, it is vital to have an informed and objective decision-making process regardless of the techniques used. Without such a process, there is always the risk of introducing unintentional bias into the outlier removal process. It is also important to ensure that the integrity of any processed data is maintained by clearly identifying that outliers have been removed along with information on the process utilized. And it is essential to maintain the raw data (at least for a defined period of time) should access to it be required for clarification purposes.

Automated Data Cleaning

Automated data cleaning builds on outlier detection by using techniques like machine learning, artificial neural networks (ANN), and clustering techniques such as k-nearest neighbors (KNN). These techniques predict the value of a sensor reading based on the readings from a set of related sensors. The sensor measurement can simply be rejected or substituted with another value, based on the predicted value (Ramirez et al., 2011). The obvious downside to this approach, particularly where the true accuracy of sensed value is of critical importance, is the rejection of a true outlier sensor measurement. There is also a significant computational overhead associated with this approach, which makes it more suitable for post-processing than real-time implementation.

Plausibility Testing

A plausibility framework can test whether a received sensor reading is credible based on predefined validity checks. The expected ranges or thresholds for the characteristics utilized are defined using statistical models designed to capture the natural variations under normal operational circumstances. The test data are generally updated dynamically to account for expected temporal variations. Characteristics such as range, persistence, and stochasticity in the measurement data are interrogated for agreement with expected limits under a given set of conditions. Each test produces a simple binary pass/fail output. Plausibility testing has found application in the environmental monitoring domain (Taylor et al., 2012).

Fusing Sensor Data

Single sensors can't always be relied on to produce the measurement of interest. Measurements often contain noise, are incomplete, or lack context. In some cases, it may not be possible to measure the data of interest directly, and other approaches will be required. Sensor fusion and virtual sensors are two such approaches, which are widely used to improve the informational value of sensor data.

In the sensor fusion process, the sensor data or the derived sensor data is combined with data from other sensors or sources. The resulting information is superior to what could be achieved using the sensor data or other resources in isolation. In some applications, multiple sensors may be required to either fully quantify the measurement of interest or provide context sensitivity or situational awareness for the measurement. For example, non-contact measurement of gait velocity requires multiple sensors at fixed distances in order to calculate the velocity of people as they move past. Sensor fusion can also provide context for remote physiological measurements. In this case, kinematic sensors can identify whether a person was active prior to a physiological measurement such as ECG/EKG or blood pressure (Klein, 2004).

The way sensor data streams are fused depends on the application's requirements, sensor resolution, and available processing resources. The fusion process can occur at the sensor-system level if the microcontroller (MCU) has sufficient computational capabilities. This is particularly useful if real-time measurements are necessary. Alternatively, if real-time measurements are not necessary, the fusion process can occur on the data aggregator or on the backend IT infrastructure during post-processing of the data.

An important application of sensor fusion is motion analysis. 3D-accelerometer, 3D-gyroscope, and 3D-magnetometers have been utilized for motion-related applications, such as falls detection. When used in isolation, these sensors have some limitations that can impact accuracy and sensitivity. For example, accelerometers are sensitive to on-body position or might generate a signal when a subject is at rest. To compensate for individual sensor limitations, a sensor fusion approach combines the 3D-accelerometer, 3D-gyroscope, and 3D-magnetometer signals to deliver a 9-DoF (degrees of freedom) motion-capture solution. Such systems can provide accurate motion analysis capabilities, which are significantly cheaper and more flexible than standard optical systems (See Chapter 9). Sensor fusion is also used in compass applications, enhanced navigation, and 3D-games (Ristic, 2012). The growth in sensor fusion applications is likely to continue. Devices such as smartphones, tablets, and ultrabooks with dedicated sensor hubs that can support a wide range of sensing capabilities will enable delivery of new and exciting applications.

As discussed in Chapter 4, virtual sensors are software-defined sensors, as opposed to physical sensors. The function of a virtual sensor is to provide an indirect measurement of an abstract quantity or measurand that can't be measured directly. This measurement can be consumed by an application or user without reference to or knowledge of the contributory sensor streams (Kabaday, 2008). Virtual sensors and sensor fusion are related, as the sensor fusion process is required to create a virtual sensor. However, virtual sensors fuse data from real sensor data streams only. In smartphones, the device orientation is determined using the virtual sensor output, generated by fusing accelerometer, magnetometer, and gyroscope data. As outlined earlier in this chapter, the growth in cloud computing will result in the proliferation of virtual sensors, creating rich data sets that can be virtualized to generate novel observations. These virtual sensors will lead to new commercial and new public domain applications, created by both hacktivists or interested citizens.

Data Mining

Extracting useful and actionable knowledge from raw sensor data is a nontrivial task. Data mining is an important tool that can be applied to the knowledge-discovery process on sensor data. Effective for modeling relationships among sensor measurements, data mining is used to reveal non-obvious patterns among the data and to determine data quality issues, such as outliers. Data mining leverages a wide range of techniques, from traditional statistical analysis and nearest-neighbor clustering to more modern techniques, such as ANN, decision trees, and Bayesian networks.

Before data mining can be applied, the data typically requires some form of preprocessing to address issues such as noise, outliers, missing data, or data from malfunctioning sensors. In some applications, particularly those that require real-time or near real-time performance, data reduction may also be required. The high volumes of data generated by some sensor applications make it extremely challenging to maintain the entire data set, which is required to achieve optimal algorithm performance, in memory. This is also an issue for applications that utilize in-memory databases to achieve improved application performance (Tan, 2006). Preprocessing of sensor data typically incorporates one or more of the following activities:

- Data cleaning or filtering, such as noise reduction.
- Outlier detection and removal or replacement of outlier data.
- Reducing the data set size by removing redundant values using techniques such as statistical sampling.
- Reducing the number of dimensions within the data using methods such as principle components analysis (PCA).
- Feature extraction such as event detection; for example, identification of R-wave maxima (QRS point) in an EKG/ECG signal.

Preprocessing of data can occur in a distributed manner at the sensor node or at an aggregator node, or in a centralized manner on the backend aggregator or sensor cloud. If the sensor node has sufficient computational capabilities and energy budget, initial data processing should occur there to reduce the size and frequency of transmissions. This also helps to improve the performance of the data-mining algorithm, by reducing the data set, which is particularly important for real-time processing.

Although a wide variety of mining techniques can be applied to sensor data, they can be rationalized into four broad categories (Duda et al., 2001):

- *Classification* is based on a machine-learning approach and uses methods such as decision trees and neural networks. The basic principle is to classify a measurement into one of a predefined set of classes, based on a feature vector. Classification performs well on sensor measurements that do not contain significant feature variability.
- *Clustering* combines groups of similar sensor values together based on a set of common characteristics. Two main types of measures are used to estimate the relationship to a group: either distance (such as Euclidean distance) or similarity measures. Clustering defines the classes into which a measurement should be added as opposed to classification which adds measurements to predefined classes. Common clustering methods include: *k*-Means, fuzzy clustering and single linkage (nearest neighbor).
- *Regression* identifies a functional description of the data that can be used to predict further values for measurements. The most commonly used implementation of regression is linear regression, where the function is linear with respect to the input variables.
- *Rule induction* identifies all possible patterns in a systematic manner. Accuracy and significance are added to indicate how robust the pattern is and the probability that it will occur again. This is probably the most common form of knowledge discovery in unsupervised learning systems.

The ultimate goal of these methods is to provide a model that can be used to interpret existing data and, if necessary to predict future sensor values in an automated manner. Due to the variety of available methods, a key step is determining the most appropriate approach for modeling a given data set. With practice, experience, and some expert guidance, this selection process should become somewhat intuitive. For most applications, an iterative process will be required to optimize the particular technique being utilized. In some cases, more than one method may be needed to achieve the desired result, especially where the output of one model forms the input to another.

Standard visualization techniques, such as typical 2D or 3D bar charts and line graphs, may not be successful due to screen size constraints on mobile devices, human visual limitations, and restrictions on available computational resources. Visual data-mining approaches address these limitations by applying techniques like geometrically transformed displays (such as scatter plot matrices), dense pixel displays, stacked displays, or icon displays to the inspection and interpretation of very large data sets. Visual data exploration serves three general purposes:

- Presenting data in a graphical form that allows users to gain insight into the data.
- Confirmatory analysis that lets users confirm or reject hypotheses based on insights gained through direct interaction with the data.
- Exploratory analysis, resulting in the development of new hypotheses.

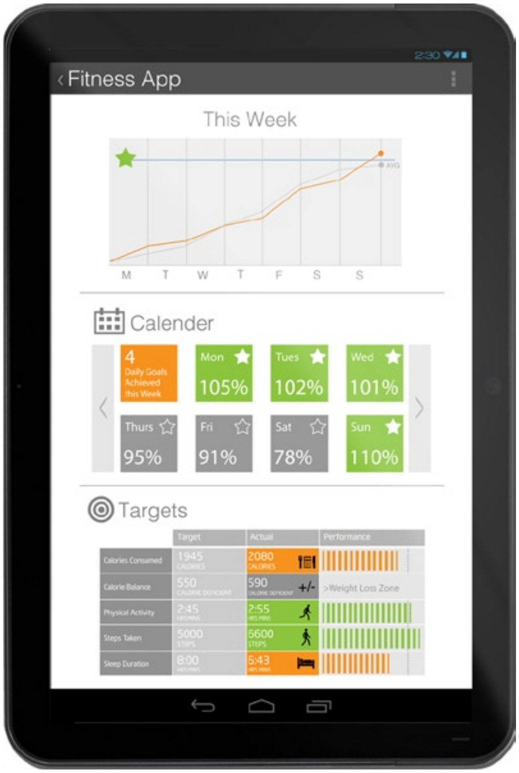
Presenting the data under scrutiny in an interactive, graphical form can facilitate new insights into a data set. It can provide a deeper understanding that is not readily discernible using standard data-mining techniques. This approach has been used in a variety of domains, including the oil industry and IT forensics, and is being applied with large-scale sensor data (Rodriguez et al., 2012). The key requirement is a presentation tool to generate initial views, navigate data sets with complicated structures, and to deliver the results of analysis. Many analytical methods don't involve visualization or have limited visualization capabilities. As the application of visual data mining continues to mature, it will evolve beyond current limitations into a highly functional and flexible tool.

Both healthcare and environmental sensor applications are areas where data mining can have meaningful utility. Sensors used to monitor a patient's condition, both in the hospital environment and outside of it, have the potential to generate significant volumes of data. These data sets will likely continue to grow unabated. However, such data sets are vastly under-utilized, despite their potential to deliver insights into the future well-being of patients, particularly for time-critical scenarios (Sow et al., 2013). And various sensor types are already being used to track the environment. Again, tremendous amounts of sensor data are being generated through air, water, climate, soil, and ecology monitoring. These data sets, if harnessed correctly through the careful application of data mining, have the potential to determine both short- and long-term trends in our environment and climate. By detecting events and revealing cause-and-effect relationships, such data will enable us to be more responsive and proactive in situations where the environment, and consequently our health and well-being, is under threat (Karpatne et al., 2013). But, as outlined in Chapter 11, significant challenges remain before this vision can be fully realized. For example, the Argo project has deployed a global array of 3660 floats containing temperature and salinity sensors in the oceans. The purpose of the project is to provide real-time data for use in climate, weather, oceanographic, and fisheries research (Argo, 2013). The challenge is how to expand the available sensing, including other measurements of interest such as pH, oxygen, and nitrate levels. Developing sensors that can operate autonomously, reliably, and accurately in this harsh environment is technically challenging and costly (West, 2011).

Mining of sensor data can involve significant overhead costs, including IT infrastructure, software tools and licenses, and networks and staff to maintain and grow the infrastructure over time. Therefore, it is important to continually question whether the correct data is being collected at the correct sampling rate. There is little point in collecting and mining data if it can't be used to drive meaningful actions. The question of whether the data mining and associated costs are delivering a return on investment must also be continually asked. Vast quantities of sensor data mined in a variety of sophisticated ways may have limited impact if the resulting information has no real predictive value. The output of the data-mining process should also be used to validate any subsequent actions taken on the basis of the analysis. It should also help determine the utility of these actions on a continual basis. Collecting more sensor data is useful only if it supplements and strengthens the quality of the analysis process. It becomes counter-productive if used as a substitute for informed analysis.

Data Visualization

Generally, people prefer graphics and visuals to pages of numeric values or text. Visual representations help address issues such as information overload or *data glut*. They allow people to more easily see the patterns and connections that are meaningful and important. If used appropriately, data visualization is a key component in the value-add chain of sensor data processing because it adds vibrancy to the data. It supports pattern recognition and can act as a primary catalyst in changing the manner in which sensor data is acted upon, either at an organizational level or by affecting the behaviors of individuals (for example, to initiate and maintain a sufficient level of physical activity). As the visualization process enables us to bring various information sources together, including non-sensor sources, context can be added that informs the interpretive process. Ultimately, visualization lets us create designs that tell a story about the data (McCandless, 2013). Good design, particularly in the health and wellness domains, often utilizes relative data, which connects the sensor data to peer data sets or to values that generate a fully rounded and qualified picture, rather than one based on absolutes, which could be misleading. The ability to visualize sensor data in a compelling manner allows individuals to evaluate the utility of a decision or to identify personal benefit to their health and wellbeing (as shown in Figure 5-3).



Calender

4 Daily Goals Achieved this Week	Mon 105%	Tues 102%	Wed 101%
Thurs 95%	Fri 91%	Sat 78%	Sun 110%

Targets

	Target	Actual	Performance
Calories Consumed	1945 CALORIES	2080 CALORIES	
Calorie Balance	550 CALORIC DEFICIENT	590 CALORIC DEFICIENT +/-	>Weight Loss Zone
Physical Activity	2:45 HRS MINS	2:55 HRS MINS	
Steps Taken	5000 STEPS	5600 STEPS	
Sleep Duration	8:00 HRS MINS	6:43 HRS MINS	

Figure 5-3. *Mobile app visualization of a personal activity record on a temporal basis with supporting lifestyle targets*

The kinds of visualizations that are effective differ according to the context. For example, 3D visualizations typically require higher resolution, which may impose certain restrictions on their use in small form factor displays. It is therefore important to match the visualization wishes of the user—for example, differing views depending on specified context (the viewing device)—while addressing a meaningful real-world problem (Richter, 2009). There are seven potential classes of visualizations that can be used depending on the composition of the underlying dataset (Shneiderman, 1996), as shown in Figure 5-4.

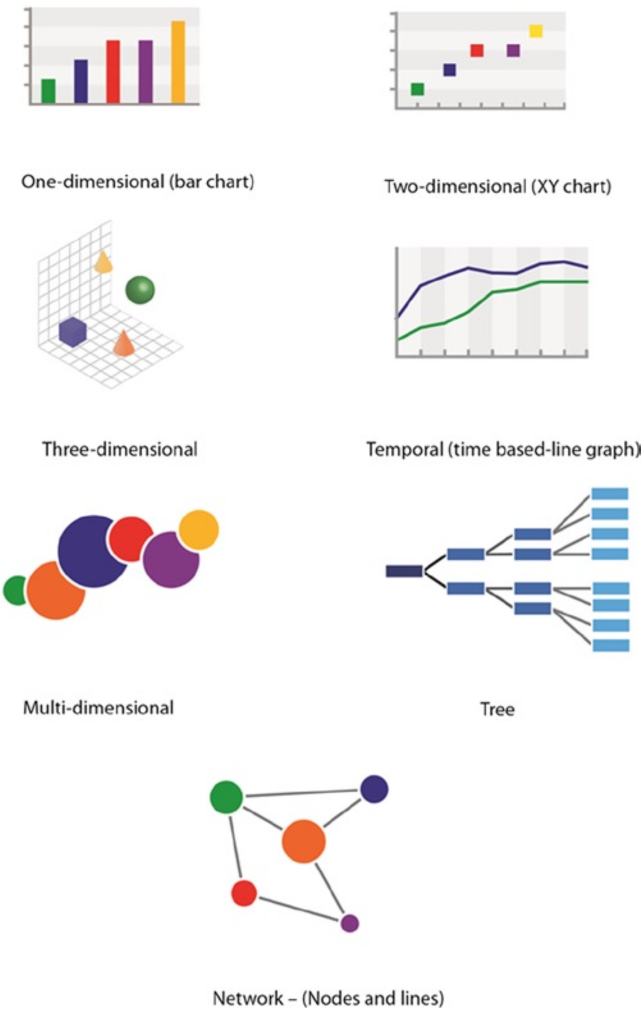

One-dimensional (bar chart)
Two-dimensional (XY chart)
Three-dimensional
Temporal (time based-line graph)
Multi-dimensional
Tree
Network – (Nodes and lines)

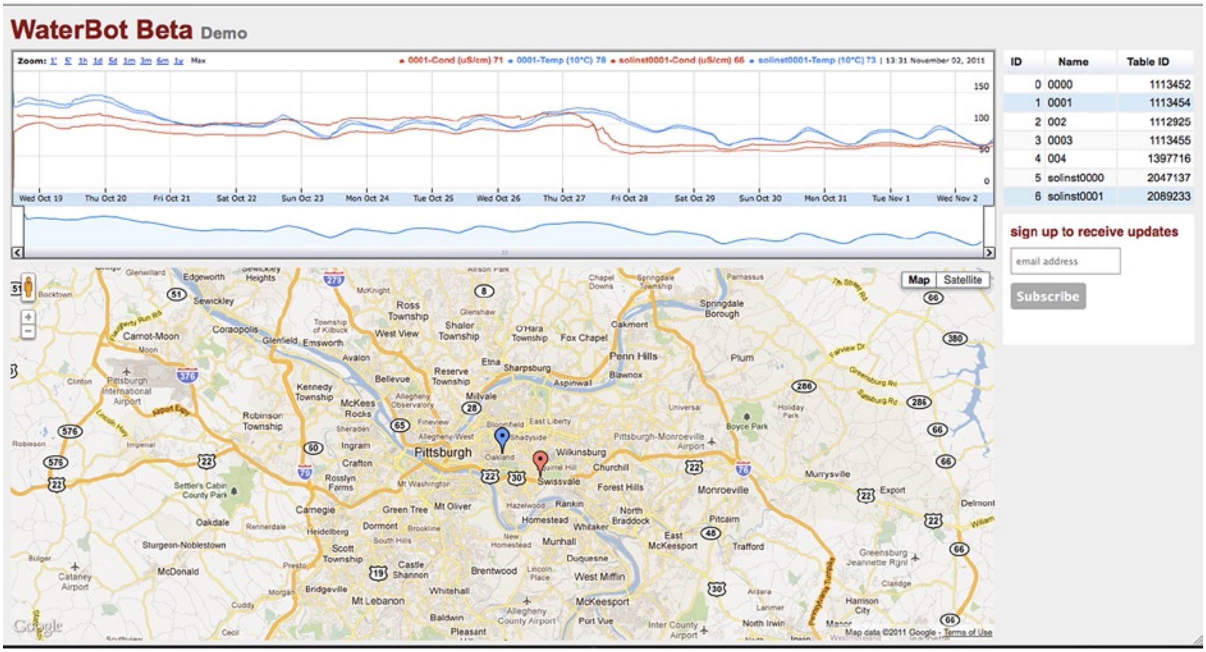
Figure 5-4. *Seven classes of visualization based on underlying data type*

1-, 2-, and 3-dimensionalrepresentations are common ways of organizing sensor data sets based on combinations of characteristics, such as temporal or spatial components along with properties such as frequency, amplitude, and so forth. Visualization of multidimensional data sets may require modifying the dimensions in some manner, either reducing the number of dimensions for display or separating dimensions into different components for display. This approach is particularly advantageous when dealing with high-dimensional data sets that must be transposed into 2 or 3 dimensions for visualization (Rodrigues et al., 2010). Temporal representations are one of the most familiar methods of visualization because the data contains definite start and end times and can be represented by a timeline.

When the dimensionality of data can't be reduced without loss of information, a multidimensional visualization approach is required. This often involves the creation of virtual sensors or other information sources to add context for sensor measurements. These forms of visualization are often customized for a particular application. For example, visualizations could fuse environmental sensor data with geospatial and cartography data in a hierarchical structure in the form of a tree. These are very useful for demonstrating the whole evolution of an aggregated value or virtual sensor reading or the relationships between groupings of similar sensors. Networks have similarities to trees in how they connect sensor measurements, but do so in a non-hierarchical manner.

As sensor deployments grow in scale, the need to visualize distributed sensor networks, such as wireless sensor networks (WSNs), will increase. A number of efforts to address this requirement have been reported. HiperSense is designed to provide scalable sensor data visualization with the ability to handle up to 6200 independent stream of data (Chou et al., 2009). In the environmental sensing domain, Teris et al. discusses an environmental monitoring system based on MicaZ motes that was used to monitor soil temperature and moisture. The resulting data was visualized using Microsoft Research SensorMap, which enabled the geographic coordinates of the motes to be accessed via a web service interface. This allowed users to find specific sensor locations and to drill down into that location for both current and historical measurements, providing both a micro and macro view of the data with geographical context (Terzis et al., 2010).

Another popular tool for visualizing sensor data is Google's Fusion tables. Fusion is a web-based application that allows users to gather, visualize, and share large data tables (Bradley et al., 2011, Fakoor et al., 2012). Once the data has been collected, the user can apply filters and create summaries across rows, up to hundreds of thousands. The user can visualize the data using charts, maps, or custom layouts and embed the visualization in a web page to share it with others. The tool enables citizens to upload a variety of environmental data sets, such as air and water quality data and contextual metrological data. It has also been used by researchers, such as those at the Waterbot program at the Create Lab in Carnegie Mellon University (CMU), to visualize data from water-quality sensors (temperature and conductivity sensing) deployed at the Nine Mile Run watershed in Pittsburgh, as shown in Figure 5-5. The researchers used Google Fusion tables to present their sensor data against a reference sensor (Solinst) with geographical sample mapping. They have been able to identify conductivity spikes that corresponded to sewage overflow events from heavy rainfall. In a blog post on the Climate Code, Professor Illah Nourbakhsh at CMU described Google Fusion tables as a key enabler of citizen science. He outlines how access to web-based data collection and visualization enables democratization of data, and in doing so empowers citizens to make informed decisions about their environment (Nourbakhsh, 2012).


WaterBot Beta Demo
Zoom: 1' 5' 1h 1d 5d 1m 3m 6m 1y Max
● 0001-Cond (uS/cm) 71 ● 0001-Temp (10°C) 78 ● solinst0001-Cond (uS/cm) 66 ● solinst0001-Temp (10°C) 73 | 13:31 November 02, 2011
150
100
50
0
Wed Oct 19
Thu Oct 20
Fri Oct 21
Sat Oct 22
Sun Oct 23
Mon Oct 24
Tue Oct 25
Wed Oct 26
Thu Oct 27
Fri Oct 28
Sat Oct 29
Sun Oct 30
Mon Oct 31
Tue Nov 1
Wed Nov 2
sign up to receive updates
email address
Subscribe
Map
Satellite
Pittsburgh
Map data ©2011 Google - Terms of Use

ID	Name	Table ID
0	0000	1113452
1	0001	1113454
2	002	1112925
3	0003	1113455
4	004	1397716
5	solinst0000	2047137
6	solinst0001	2089233

Figure 5-5. *Visualization of water quality sensor measurements using Google Fusion tables (with permission from the Create Lab, Carnegie Mellon University)*

In the health and wellness domain, the ability to mine and visualize longitudinal data sets, such as behavior patterns at home, can provide early indications of abnormal behaviors and health-related issues (Lotfi et al., 2012). Visualization designs for this domain should accommodate both individual- and peer group-level norms as appropriate. Keep in mind that we are all individuals with our own unique biology, so adopted designs must be carefully selected in order to prevent "loaded" representations that yield narrow and restricted representations and do not reflect individual complexity.

Visualization can provide the tools for driving awareness in domains of interest. But, ultimately, its usefulness is determined by the ability to build accurate and relevant models that produce correct answers. Visualization is a powerful and effective tool for adding texture to sensor data sets and providing a mechanism to compress significant amounts of knowledge into an intuitive visual metaphor. However, visualization can only provide answers and insights to properly constructed questions. It is not a magic tool that can make sense from data chaos. It requires careful guidance to ensure data is correctly interpreted. In the future, data visualization will move beyond the current static modes of interactions to more interactive modalities, which will enable individuals to interact with data. Augmented reality solutions, such as Google Glass, will enable real-time delivery of visualized data that enhances our perception of the world and its impact on us. Ultimately, visualization is about adding vibrancy to sensor data so we can make connections to it and to tell meaningful stories about it.

Big Sensor Data

Big data is a catchphrase that has gained considerable traction in the business analytics world. Big data is said to be the "new oil." Of course, this is something of a misnomer. Data is not a finite resource like oil. In fact the opposite is true, with new reserves of data being created on a daily basis (Thorp, 2012). It is hard to avoid the fact that we are generating ever increasing amounts of personal and systems-based data each day from connected devices. According to IBM, mankind and its supporting infrastructure generate a massive 2.5 quintillion bytes of data daily (IBM, 2013). Whether "big data" is truly a new phenomenon or just a part of the continuous evolution of a technological society is likely to remain a point of contention among the business intelligence analytics vendors and industry veterans and academics (Few, 2012).

The proliferation of sensors into all aspects of our lives—including smartphones, body-worn sensors, homes, and monitoring future smart cities—is often used to illustrate how sensors are and will be a key contributor to big data growth. It is important to realize that collecting data from new sensors may not involve collecting new data types. Instead, it will simply add to the overall cumulative effect of existing sensor data. Sensors will slowly add incremental data sources only as new technologies emerge. They will, though, add increasing volumes of the same measurement types and become more pervasively used. Most big data sources typically generate data in an automated manner without human intervention. Automated sensing does occur in the health and wellness domain (such as home-based activity monitoring). At the same time, a significant human interaction element remains in the sensing process, for example, physiological sensing like blood pressure or blood chemistry monitoring (for diabetes control, for example). Despite the large "manual" element in health- and wellness-related sensing, the volumes of data generated continue to grow. Another characteristic of data in these domains is the high levels of interaction and participation it engenders. People often feel a strong sense of personal ownership of the data and can be highly motivated to interact with it and compare their "numbers" with either peer groups or other sources of aggregated data.

The distinct category of "big sensor data" has started to emerge with the advent of sensor-enabled smartphones and tablets, coupled with cloud-based services, which enables persuasive user feedback (Lane et al., 2010). In the healthcare domain, large-scale data sets, such as those that can be generated by smartphone sensors, do present challenges—particularly with respect to the manner in which the data is mined. Population-level data can often have the effect of degrading the differences among people, which is particularly problematic for classification-based systems. This issue is commonly called the "classification diversity problem." As mentioned in the preceding section, the uniqueness of an individual's biological and biomechanical composition makes creating highly generalizable models extremely difficult. Generalized models typically afford only indicative value, as opposed to diagnostic value, which require highly granular models tailored to small population groups with common epidemiology (Campbell et al., 2012).

The value of big sensor data is only as good as the information you can extract from it. The ability to interpret the data correctly is also critical and dependent on the data literacy of citizens as discussed earlier in this chapter. A lack of awareness can result in confusion among the public and lead to misinformed debate. New methods of analysis may not necessarily be required. Enhancements to current tools will be necessary to handle and analyze the large sensor data volumes in order to deliver the performance muscle needed, especially if real-time analytics are required. Existing analytics solutions can process temporal sensor measurements but have difficulty in correlating data sets with other data sources quickly. The new generation of big data analytics tools, including NoSQL databases, Hadoop, and MapReduce, are designed to address these requirements. Hadoop's open-source stack uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The Hadoop stack includes utilities, a distributed file system, analytics and data storage capabilities, and an application layer. These features support distributed processing, parallel computation, and workflow and configuration management. MapReduce, the software programming framework in the Hadoop stack, simplifies processing of large data sets by giving programmers a common method for defining and orchestrating complex processing tasks across clusters of computers.

Big data analytics remains an area of active and growing research. For example, in Ireland the national research body Science Foundation Ireland recently announced the Insight Centre at a cost of over 75 million euros. The focus of this center is to develop a new generation of data analytics technologies focusing on key application areas, such as healthcare and the sensor web(Insight, 2013). It is the biggest ICT R&D investment in the history of the country, demonstrating the strategic importance of big data analytics to the national economy. Within the research domain, active topics include algorithmic techniques for modeling high-dimensional data, knowledge discovery for large volume dynamic data sets, and methods for automated hypothesis generation. In data storage research, the focus is on data representation, data storage, data retrieval, and new parallel data architectures, including clouds.

The value of big sensor data can only be systematically realized by a bottom-up approach. That approach starts with basic tenets: What do we want to measure and how do we measure it? Scaling data collection to large sample sizes will then enable us to see the patterns and connections that are important. This process will move us into data-driven discovery, which in turn will lead to an accelerating pace of innovation.

Summary

In this chapter we looked at the importance of data literacy and the steps required to extract new knowledge about our environment and our health and well-being. The process of turning sensor data into actionable information includes: data preparation, mining, and visualization. But to realize the value of this process, people need be sufficiently data literate to understand the information sensors can provide. We also saw how the visualization process is a key element in adding vibrancy to data. That process can help turn data into stories that enable individuals, groups, and organizations to make a connection to a data set. We have seen that the amount of sensor data is growing through the emergence of a connected world and the Internet of Things, leading to the phenomenon of 'big sensor data.' The value of these large scale data sets is dictated by its quality and whether it is measuring something of interest in an accurate and contextualized manner.

References

Hart, Robert V. "*Data Information Literacy?*," Last Update: July 26th, 2011, http://esciencecommunity.umassmed.edu/2011/07/26/data-information-literacy/

Lohr, Steve. *The Age of Big Data*, The New York Times, New York, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=1&_r=1, 2012.

Shelton, Steve. "*Building a Big Data workforce: how can we get started?*" [Blog]. Last Update: 17th August, 2012, https://www.baesystemsdetica.com/news/blogs/building-a-big-data-workforce-how-can-we-get-started/

Shilton, Katie*, et al.*, "Designing the Personal Data Stream: Enabling Participatory Privacy in Mobile Personal Sensing," presented at the 37th Research Conference on Communication, Information, and Internet Policy (TPRC 2009), Arlington, Virginia, 2009.

Thompson, Clive, "Why We Should Learn the Language of Data". *Wired Magazine,* (May)*,* 2010

Gonçalves, Bernardo, José G. Pereira Filho, and Giancarlo Guizzardi, "A Service Architecture for Sensor Data Provisioning for Context-Aware Mobile Applications," presented at the ACM Symposium on Applied Computing (SAC '08), Fortaleza, Ceará, Brazil, 2008.

Bradshaw, Leslie. "*Data Dreaming*," Last Update: December 14th, 2012, https://medium.com/american-dreamers/4ee4351f8aab

Swan, Melanie, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," *Journal of Sensor and Actuator Networks,* vol. 1 pp. 217-253, 2012.

Cisco. "*The Internet of Things*," Last Update: 2011, http://share.cisco.com/internet-of-things.html

AirQualityEgg. "*AirQualityEgg*," Last Update: 2013, http://airqualityegg.wikispaces.com/AirQualityEgg

designboom. "*Electric Imp for the Internet of Things*," Last Update: May 17th 2012, http://www.designboom.com/technology/electric-imp-for-the-internet-of-things/

Mitton, Nathalie, Symeon Papavassiliou, Antonio Puliafito, and Kishor S. Trivedi, "Combining Cloud and Sensors in a Smart City Environment," *EURASIP Journal on Wireless Communications and Networking,* vol. 2012 (247), 2012.

LORD MicroStrain, "SensorCloud," http://www.sensorcloud.com/, 2013.

temperature@lert. "*Temperature@lert Sensor Cloud Tour*," Last Update: 2012, http://www.temperaturealert.com/Remote-Temperature/Sensor-Cloud/Sensor-Cloud-Tour.aspx

Yuriyama, Madako and Takayuki Kushida, "Sensor-Cloud Infrastructure - Physical Sensor Management with Virtualized Sensors on Cloud Computing," in *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, 2010, pp. 1-8.

Agence Nationale De La Recherche, "Clouds@Home," http://clouds.gforge.inria.fr/pmwiki.php?n=Main.HomePage, 2011.

Hummen, René, Martin Henze, Daniel Catrein, and Klaus Wehrle, "A Cloud Design for User-controlled Storage and Processing of Sensor Data," presented at the IEEE CloudCom, Taipei, Taiwan, 2012.

Puentes, John, Julien Montagner, Laurent Lecornu, and Jaakko Lahteenmaki, "Quality Analysis of Sensors Data for Personal Health Records on Mobile Devices," in *Pervasive Health Knowledge Management*, Bali, Rajeev K., Indrit Troshani, and Steve Goldberg, Eds., New York, Springer, 2013, pp. 103-134.

Zhang, Yang, Nirvana Meratnia, and Paul Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials,* vol. 12 (2), pp. 159-170, 2010.

Ramirez, Gesuri, Olac Fuentes, and Craig E. Tweedie, "Assessing data quality in a sensor network for environmental monitoring," in *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, 2011, pp. 1-6.

Taylor, Jeff R. and Henry L. Loescher, "Automated Quality Control Methods for Sensor Data: A Novel Observatory Approach," *Biogeosciences,* vol. 9 (12), pp. 18175-18210, 2012.

Klein, Lawrence A., *Sensor and Data Fusion - A Tool for Information Assessment and Decision Making*. Bellingham, Washington: SPIE Press, 2004.

Ristic, Lj. "*Sensor fusion and MEMS for 10-DoF solutions*," Last Update: 3rd September, 2012, http://eetimes.com/design/medical-design/4395167/Sensor-fusion-and-MEMS-technology-for-10-DoF-solutions

Kabaday, Sanem, "Virtual Sensors: An Intutive Programming Abstraction," in *Enabling Programmable Ubiquitous Computing Environments: The DAIS Middleware*, Ann Arbour, Michigan, ProQuest LLC, 2008, pp. 36-57.

Tan, Pang-Ninh. "*Knowledge Discovery from Sensor Data*," Last Update: March 1st, 2006, http://www.sensorsmag.com/da-control/knowledge-discovery-sensor-data-753?page_id=1

Duda, Richard O., Peter E. Hart, and David G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2001.

Rodriguez, Claudia C. Gutiérrez and Anne-Marie Déry-Pinna, "Visualizing Sensor Data: Towards an Experiment and Validation Platform," in *Human-Centred Software Engineering*, Winckler, Marco, Peter Forbrig, and Regina Bernhaupt, Eds., Heidelberg, Springer-Verlag, 2012, pp. 352-359.

Sow, Daby, Deepak S. Turaga, and Michael Schmidt, "Mining of Sensor Data in Healthcare: A Survey," in *Managing and Mining Sensor Data*, Aggarwal, Charu C., Ed., New York, Springer US, 2013, pp. 459-504.

Karpatne, Anuj, *et al.*, "Earth Science Applications of Sensor Data," in *Managing and Mining Sensor Data*, Aggarwal, Charu C., Ed., New York, Springer US, 2013, pp. 505-530.

Argo. "*Argo - part of the integrated global observation strategy*," Last Update: 2013, http://www.argo.ucsd.edu/About_Argo.html

West, Amy E. "*Widespread floats provide pieces of the oceanic productivity puzzle*," Last Update: October 28th 2011, http://www.mbari.org/news/homepage/2011/johnson-floats/johnsonfloat.html

McCandless, David. "*Information is Beautiful*," Last Update: 2013, http://www.informationisbeautiful.net/tag/health/

Richter, Christian, "Visualizing Sensor Data - Media Informatics Advanced Seminar on Information Visualization," University of Munich, Munich, http://www.medien.ifi.lmu.de/lehre/ws0809/hs/docs/richter.pdf, 2009.

Shneiderman, B., "The eyes have it: a task by data type taxonomy for information visualizations," in *Visual Languages, Proceedings., IEEE Symposium on*, 1996, pp. 336-343.

Rodrigues, Pedro P. and João Gama, "A Simple Dense Pixel Visualisation for Mobile Sensor Data Mining," in *Knowledge Discovery from Sensor Data: Second International Workshop, Sensor-KDD 2008*, Vatsavai, Ranga Raju, Olufemi A. Omitaomu, João Gama, Nitesh V. Chawla, and Auroop R. Ganguly, Eds., Heidelberg, Springer, 2010, pp. 175-189.

Chou, Pai H., Chong-Jing Chen, Stephen F. Jenks, and Sung-Jin Kim, "HiperSense: An Integrated System for Dense Wireless Sensing and Massively Scalable Data Visualization," presented at the Proceedings of the 7th IFIP WG 10.2 International Workshop on Software Technologies for Embedded and Ubiquitous Systems, Newport Beach, CA, 2009.

Terzis, Andreas, *et al.*, "Wireless Sensor Networks for Soil Science," *International Journal of Sensor Networks,* vol. 7 (1), pp. 53-70, 2010.

Bradley, Eliza S., *et al.*, "Google Earth and Google Fusion Tables in support of time-critical collaboration: Mapping the deepwater horizon oil spill with the AVIRIS airborne spectrometer," *Earth Science Informatics,* vol. 4 (4), pp. 169-179, 2011.

Fakoor, Rasool, Mayank Raj, Azade Nazi, Mario Di Francesco, and Sajal K. Das, "An Integrated Cloud-based Framework for Mobile Phone Sensing," presented at the Proceedings of the first edition of the MCC Workshop on Mobile Cloud Computing, Helsinki, Finland, 2012.

Nourbakhsh, Illah, "*Citizen Science for Watershed Action: Big Data Meets Fusion Tables*," The Climate Code, Last Update: March 13th 2012, http://www.theclimatecode.com/2012/03/guest-post-citizen-science-for.html

Lotfi, Ahmad, Caroline Langensiepen, Sawsan M. Mahmoud, and M. J. Akhlaghinia, "Smart Homes for the Elderly Dementia Sufferers: identification and Prediction of Abnormal Behaviour," *Journal of Ambient Intelligence and Humanized Computing,* vol. 3 (3), pp. 205-218, 2012.

Thorp, Jer. "*Big Data Is Not the New Oil*," Last Update: 2012, http://blogs.hbr.org/cs/2012/11/data_humans_and_the_new_oil.html

IBM. "*What is big data*," Last Update: 2013, http://www-01.ibm.com/software/data/bigdata/

Few, Stephen, "Big Data, Big Ruse". *Visual Business Intelligence Newsletter,* vol. July/August/September, 2012, http://www.perceptualedge.com/articles/visual_business_intelligence/big_data_big_ruse.pdf

Lane, Nicholas D., *et al.*, "A survey of mobile phone sensing," *IEEE Communications Magazine,* vol. 48 (9), pp. 140-150, 2010.

Campbell, Andrew and Tanzeem Choudhury, "From Smart to Cognitive Phones," *Pervasive Computing*, vol. July-September, 7-11, 2012

Insight. "*The Insight Centre for Data Analyticsn*, Last Update: 2013, http://www.insight-centre.org/about/mission