# Applying Data Analytics to Big Data Obtained from Wearable Devices

Cindy Rodriguez, Andrew Barrow, Shalwak Dangore, Ujwal Pathak, and Joseph Talledo
Seidenberg School of CSIS, Pleasantville, New York

***Abstract*—In recent years, wearable sensors have progressed from niche use into the mainstream. These modern sensors are light, small in size and even wearable on user's wrists. Consumer products from vendors such as Apple with their Apple Watch and Fitbit with their line of fitness bands have become increasingly commonplace. With the introduction of these and similar devices, wearable sensors have become not only unobtrusive, but a regular part of daily life for many people. Despite their small size, they can record a large amount of varied data from heart rates to geolocation. As thousands of users utilize these sensors each day, these devices can store and transmit a wealth of information. This wealth of information has given rise to big data. Big data is a term to describe a large amount of data that can be processed and analyzed to derive insights, patterns and other trends. The emergence of big data has allowed many commercial corporations to gain insights on the consumer. In addition to this, there are a wealth of patterns and insights related to health that can be accessed from the data found within wearables. With it, one could perform analytics that could identify and predict health tendencies, overall fitness, and general health over time. These are examples of the different types of big data analytical methods explored.**

***Keywords—wearable computers, sensor systems, mobile security, data analysis, security***

## I. Introduction

The popularity of wearable technologies has raised tremendously in recent years. With the convenience of health and fitness monitoring and tracking, millions of users allow companies access to their personal data and sensitive information. This, in part, has led to the rise of big data analytics. While big data was in its infancy, many did not understand the benefits of analyzing data in order to recognize trends and patterns. The majority of the data was and still is unstructured. In other words, there is no organization to the data; it merely exists. With the millions of users dumping their data onto servers each and every day, the amount of data has only continued to grow. This growth has led many companies to begin to see the benefits in analyzing and structuring the data in order to gain insights into their consumers. It also allows many companies to inform consumers of their tendencies. An example of this would be FitBit tracking a user's heart rate over time in an effort to determine the overall health of the user. They then, in turn, alert the user of this information so that the user can either continue their current habits or change them.

Wearable sensors are prevalent in various forms. While common devices marketed to consumers such as the Fitbit, Apple Watch and similar sensors are wrist-based, other specialized forms of sensors have also found use. These include glasses, vests and full body suits. In totality, sensors in these various forms generate a large amount of data such as users' vital signs. As these sensors have moved out of controlled lab conditions and into the world at large, there are several lingering questions about the data they collect. One involves what data is being generated, shared, stored, and accessed.

The focus of this study is to examine modern wearable sensors through both review of existing literature and direct examination. The aim is to access the current state of wearable sensor technology and perform data analysis on the large amount of data collected. By analyzing a set of contemporary wearable sensors, the goal is to examine how the sensors work, the data that is collected, and how that data can be leveraged to create insights about its user. Finally, making informed observations about the way forward for wearable technology from a technical and consumer standpoint will contribute to wearable sensor research while offering sensible recommendations to ensure safe, reliable usage of wearables for the foreseeable future.

## II. Study Requirements

The primary idea of the project is to analyze the wearable sensor device and the data it generates. To implement the idea:

- The preliminary step is to extract the sensors data using an API model.
- Incorporation a back-end [Hadoop - Big Data Framework] to store huge amount of data received from the sensor device.
- The data received from the sensor devices will be unstructured and will require application of appropriate Data Integration techniques in order to make the data fit for analysis.
- Python /R [1] scripts will be written to pre-process the data and filter out any discrepancies.
- Spark [2] will be used to implement learning algorithms and derive insights, patterns, and trends.

The overall architecture technology requirements can be put forward as depicted in Figure 1.

## III. Literature Review

It is undeniable that companies are finding new ways of using large quantities of data, known as big data, to gain insights into consumers. With the rise in the popularity of Internet of Things (IoT) devices, companies have become increasingly interested in big data analytics. Taking advantage of the data found on devices can help glean information of its users. Although there has been growing concerns over user privacy, as touched on in Arriba-Perez's paper, it hasn't stopped the depth of data collected from being immense. One

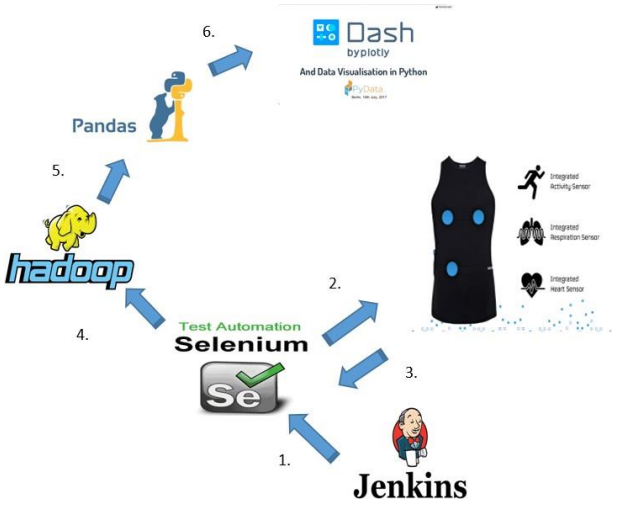


Fig. 1. Technology Work Flow

source of this big data has been wearable devices. Wearable devices are defined as any technical device that is worn by a consumer. With estimations that, by 2019, wearable devices will have reached 100 million units sold, it is unsurprising that the amount of data that is sourced from wearable devices alone is tremendous [3].

There are multiple insights that can be gained from the analysis of data from wearable devices. Common types of data collected by these devices include information related to fitness and overall health. One can learn a user's overall health, sleeping patterns, or even when they partake in certain actions, such as talking, eating or writing, from the data on their smart watches [3]. In addition to this, many wearable devices also include a location feature which uses the Global Positioning System (GPS). With the addition of GPS location sensors, a wearable is also able to keep data about a user's location as well.

Understandably so, many have been concerned with the potential use, or misuse, of the data collected. Much of the collected data is considered to be personal data. Local government officials define personal data as any information that can be used to identify a person's identity, whether directly or indirectly [4]. This includes a number of different identifiers including but not limited to name, address, location, or phone numbers. Ironically, consumers who purchase and use a wearable device own the device, but not the resulting data from that device [5]. Manufacturers who sell the device are the true owners of the data collected. In fact, manufacturers are able to sell this data to third party companies as long as the data is "anonymous". As pointed out by Piwek, however, the measures companies take to anonymize data are not sufficient. In some cases, it can lead to identity fraud as sophisticated algorithms have the ability to reveal a user's identity based off of the "anonymous" data.

Many companies claim they are dedicated to user privacy and yet, have had multiple breaches of security and information [6]. Unfortunately for consumers, a research study conducted by Hewlett-Packard revealed that 100 percent of the most popular devices contain critical flaws, easily exploitable by even a casual hacker [7]. Despite the sheer amount of breaches that have occurred over the past couple of years, numerous companies continue to advertise their abilities to protect the user, often making users think that they are secure. This false sense of security leads users to continue to use these services without realizing it can be bypassed and compromised. It is alarming the amount of data that many large companies have on their consumers and yet, there are not sufficient measures in place to protect that data.

To make matters worse, as Hamblen pointed out in his news article, the data collected by companies can potentially be later used against the user. An insurance company, for example, could deny a health claim based on data from the user's smart watch [8]. Even though the majority of users assume that the collected data is anonymous, it isn't. In fact, smart watches collect enough data to be able to create profiles of their users to predict future actions, habits, and preferences [8]. In Patrikakis' paper, he asserts that wearable devices are "capable of understanding [...] even the condition and mood of their owner [9]."


Fig. 2. Opportunities with the insights gained from the analysis of big data

Although there are user privacy concerns, the potential actionable insights gained from the analysis of data from wearable devices also has potential benefits, as depicted in Figure 2 above. As mentioned previously, wearable devices are becoming increasingly popular in the market. Estimates by the CCS Insight's Wearables Forecast suggest that the number of wearable devices being shipped will reach 245 million per year by the end of 2019 [10]. Figure 3 breaks down the estimated sales by wearable device type, including devices such as eyewear, wristbands, watches, and others. Because there is a large pool of data being collected from millions of users, companies are able to leverage the data in order to create meaningful insights for users. This is arguably one of the greatest potential benefits to the user. By doing so, Nayar argues that the overall usefulness of wearable technology will be improved and user engagement will increase [10]. An example of this provided by Nayar is a company that can gain insights into the health of their employees in an attempt to reduce high levels of stress and create a happier, more effective workplace. The data, health of the employee in this case, can lead to an actionable insight and result, such as the company instilling a later starting time.

With the proliferation of wearable sensors, larger and larger amounts of data are being collected. This requires data

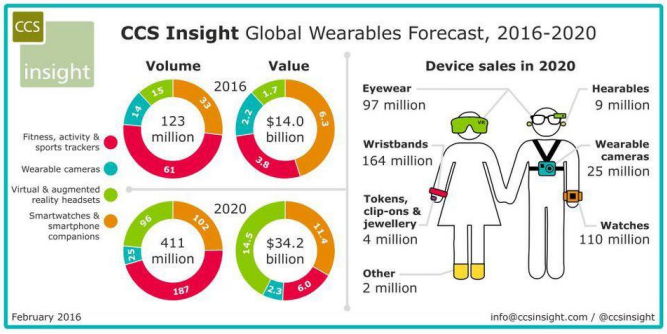


Fig. 3. CCS Insight Global Wearables Forecast

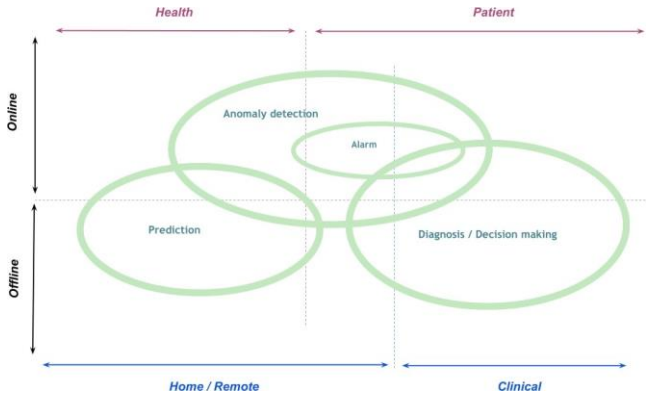


Fig. 4. A sensor data analysis method [12]

collection and analysis methods akin to the big data processing methods seen in various other fields, such as any number of online services.

While big data management with individual wearable sensors seems daunting at first, one relatively simple framework it described by Hussain, Kang and Lee. They distill big data handling with wearable sensors into several key steps: data acquisition, data transformation and data validation.[11]

Data acquisition entails capturing the raw data from the sensor(s) and passing it on for further processing. In some devices, this would involve connecting to the cloud. The next step is data transformation, or taking that raw data, categorizing it and turning it in a structured form such as CSV or text file. The final step, data validation checks for errors and inconsistencies. It can then be stored and further analyzed through a platform such as Hadoop.[11]

Data from wearable sensors, including within a healthcare context is not of much use unless that data is analyzed and turned into a useful form. With mainstream data collection from wearable sensors still largely in its infancy, there have yet to be a universally agreed upon set of standards and common processes to analyze sensor data.

One set of primary data analysis methods identified by Banaee, Ahmed and Loutfi includes several tasks, which have some degree of overlap. These tasks are anomaly detection, prediction, and diagnosis/decision making. Each of these tasks can be performed to varying degrees both online and offline. [12]

Anomaly detection involves detecting variations from the normal, expected behavior from the data. Statistically this can involve the identification of outliers in the data. Identifying anomalies can assist clinical personnel in determining necessary treatment modalities. Detection of anomalies can go hand in hand with an alarm system alerting relevant parties such as healthcare personnel as to any deviations from the norm.[12]

Prediction, as the name implies, involves identifying future events. This is a growing area of sensor data analysis as it can potentially predict future health conditions and allow providers to enact proper preventive care measures.[12]

The diagnosis/decision making process involves assessing condition (e.g. patient status) based in part on sensor data. The decision making process entails what action is taken as a result. Part of this intersects with anomaly detection in vital signs, as anomalies may be indicative of particular medical issues. However other data also plays a factor in this process, such as a medical history and other factors.[12]

These tasks, performed in conjunction with the usual data cleaning and filtering processes outline a particular method for analyzing sensor data in a healthcare context.[12] The classification of data can prove to be a challenge for sensors that for example detect different types of activity. This issue becomes more important as data sets become larger and larger. Efforts to accurately classify sensor data have involved advanced data processing and filtering and have resulted in more efficient usage of sensor data to classify varying activities.[13]

## IV. Methodology

This study was conducted to examine wearable sensor devices, the data they collect, and to investigate any emerging patterns. The wearable devices used in our research were Apple Watch, Fitbit, and Hexoskin body vest. To gather the necessary data, descriptive study method were employed, investigated using the qualitative and the quantitative methods.

### *A. Data Collection*

Extraction of the data was attempted from the Hexoskin API Dashboard using Selenium scripts to ensure the data integrity and by extension its validity.

- Data from the Apple Watch was retrieved via an unencrypted backup of the iPhone paired with the Watch.
- Data from the Fitbit was retrieved via the Fibit web API on a Windows 10 machine.
- Data from the Hexoskin body vest was retrieved using the Representational State Transfer API.

### *B. Data Analysis*

The Raw data retrieved from our sensors which were time, breathing rate, heart rate, activity and cadence (steps per minute). Once the data was gathered from the Apple, Fitbit, and Rest API in csv format it was stored and processed using the Hadoop big data warehouse. We compared to the heart rate data retrieved from the API to the normal heart rate and

illustrated in the multiple graphs comparing heartrate over time and how the changes could indicate periods of intense activity. [14].

## V. Preliminary Findings/Results

Hexoskin provides an online dashboard to view the data synced by our wearable sensors through its HexoService app which also stores data locally. It uses REST API that stands for Representational State Transfer which provides interoperability between computers on the internet. When the data from the Hexoskin device is synced to the Hexoskin API servers it gets reflected on the online dashboard as a new record. To download the data from the online dashboard in real-time which was implemented with the help of a Selenium script in java which is triggered periodically by a continuous integration server (Jenkins) that completes our data gathering process. Selenium is an automation framework, which provides modules/libraries to automate testing process. It provides flexibility to incorporate the libraries in any programming languages.The entire data gathering process can be illustrated as follows.

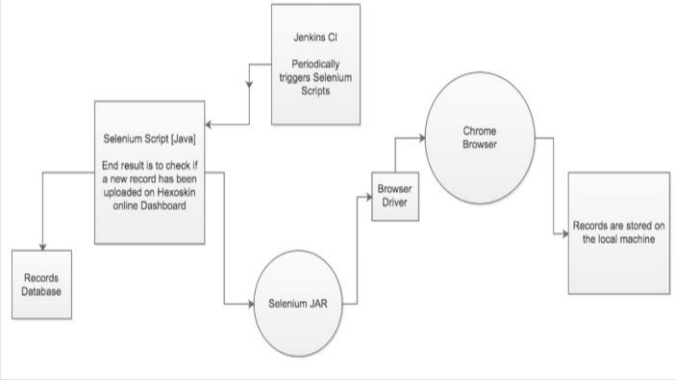


Fig. 5. Real-Time Data Gathering

The data being uploaded over a timeline on the Hexoskin servers consists of 100,000+ records per file. This sparked a need of a big data warehouse, and thus a Hadoop one node cluster was configured to store the streaming data over a timeline. Hadoop is an open source Big Data framework which incorporates distributed data processing across multiple nodes of a cluster which helps faster processing of big data sets.

[15]

The raw data collected from the sensor has following parameters

1) Time
2) Breathing (Rate per minute)
3) Heart Rate(Beats per minute)
4) Activity
5) Cadence

The data is pre-processed i.e. empty rows are deleted, garbage and redundant values are removed. The time format in the raw data is in Trans-Norm which was converted into UTC format using appropriate DateTime libraries in Python. To get appropriate insights from the data, ideal values of the parameters were researched. For example, "A normal resting heart rate for adults ranges from 60 to 100 beats a minute.

Generally, a lower heart rate at rest implies more efficient heart function and better cardiovascular fitness. A well-trained athlete might have a normal resting heart rate closer to 40 beats a minute".[16] Sample heart rate was studied and compared to the ideal heart rate range. The results were visualized using matplotlib library in Python. The link[14] to the Github can be found in the list of references. For the preliminary analysis, variation of heart beats were visualized.

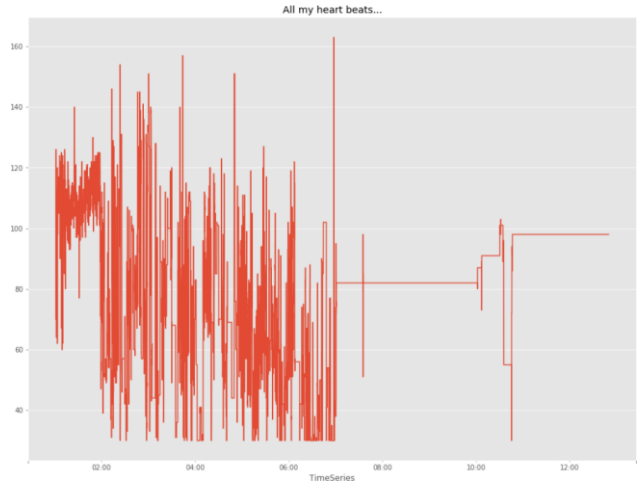


Fig. 6. Variation of Hearbeat per minute over a timeline

To take a new look at the data, activity parameter was studied and categorized into various sections i.e. intense, moderate, resting activities. The segregation was implemented by calculating the average body movements over a timeline. The result gives an insight to the user about his daily activities.

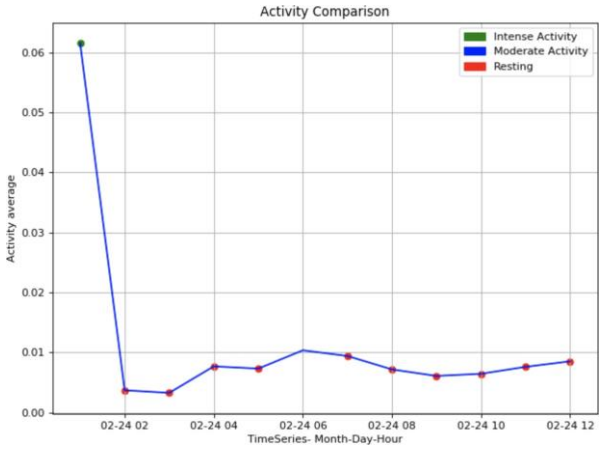


Fig. 7. Activity Collation

To dig deeper into the activities of the user, cadence [steps per minutes] retrieved from the sensor were evaluated over a month. The variations in steps on daily basis were studied in order to derive any pattern in user activity at a particular period of the day. The frequency of steps per minute were visualized and marked with peak points to identify user activities and corresponding bpm [breathing rate per minute] at that particular point.

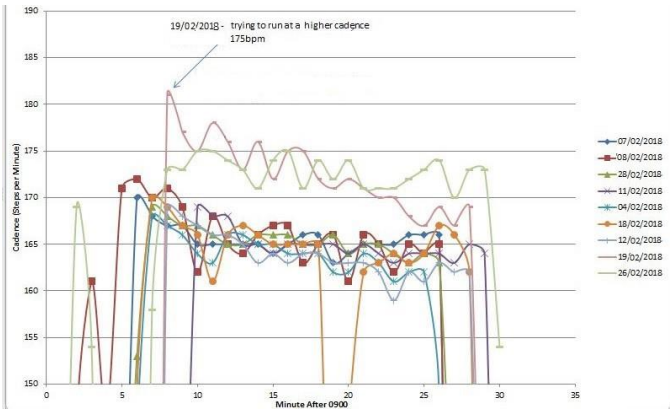


Fig. 8. Cadence Insights

The figure above illustrates the peak point of user activity [Sprint /Run] on 19th Feb and the corresponding bpm 175 from 09:00 to 09:35. The visualization provides cognition to the user about his activity variations via which he can improve on his workout.

From a user's perspective, hiding the background complexity is a key point for building any system. Instead, a user should have access to a user interface, or UI, with which they can interact. In this case, a dashboard was developed in order to allow a user to create and view any number of charts as they please. This dashboard is a concise, effectively decipherable, graphical show of key performance indicators. It gives an overview of the sensor data collected and provides an option to the user to see all of the charts with respect to particular parameters. The user can also filter out values as desired. This online dashboard was built using Dash by plotly. The initial analysis was implemented and examined using python.

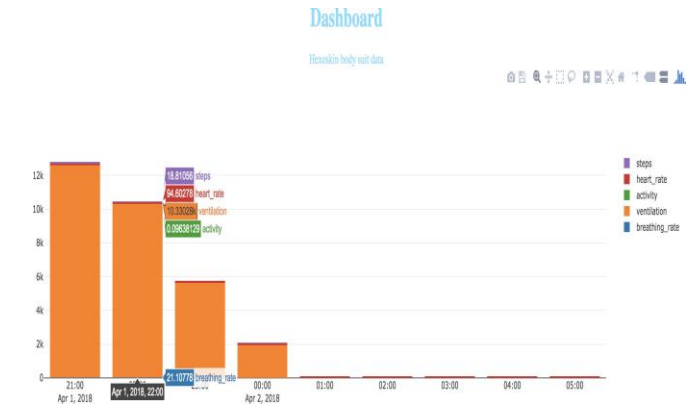


Fig. 9. Dashboard

The examination of the Hexoskin suit only represented a single subject. In actuality, there are thousands of individuals utilizing wearable sensors every day. Various data collection efforts have monitored different subjects and their use of wearable sensors. One of the largest is Insight for Wear, a lifelogging app with a continuously updated database. Access is closed to the general public, but can be requested by data scientists. The app has over 1,000 installs and millions of data points pertaining to smartwatches. [17] A similar initiative from the University of California, San Diego is the ExtraSensory app. This app monitored sensors from mobile devices, including wrist-based devices from 60 users.[18]

These datasets are anonymized, as can be expected for such personal information. Thus a natural line of inquiry is whether one can distinguish individual subjects from each other just by looking at the data. Superficially this is a data privacy issue, in that someone can be individually identified from their sensor data. But wearable sensor readings, if unique enough can presumably be a signature particular to an individual. This has possible implications on fields such as authentication, wherein users may be able to verify their identities based on their distinct sensor readings. This also has implications on studies on crowd behavior modeling and automated recognition of gestures and body language.

A cursory examination on the big data implications of wearable sensor data possibly being unique to an individual is possible. Looking at the ExtraSensory data set, we can see a microcosm of how this would work.

Activity detection was chosen as the method of examination as it was common between the Hexoskin suit (which is included along with the ExtraSensory data). The data from the Hexoskin suit and the ExtraSensory data calculate the magnitude of 3-axis acceleration as a function of time, using the unit of standard gravity. For the ExtraSensory data, the watch accelerometer data was isolated. The sensor data from between the 1-3Hz frequency band was extracted as it matched up with the frequency from the Hexoskin suit. The combined dataset represented over 300,000 data points.

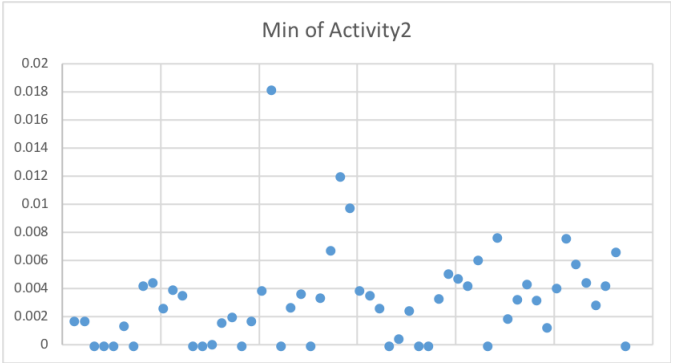


Fig. 10. Scatter plot of min values from ExtraSensory/Hexoskin data

As can be seen in the graphs, the minimum and maximum values of activity are largely distinct in this sample, but are closely clustered. The maximum values in particular are even more clustered. As the number of subjects grows exponentially larger, it conceivable that it would be increasingly difficult to discern individual records from one another. The values of activity average are more promising, with data points more spaced out. As the number of subjects grows larger, activity average seems a far more likely way to distinguish individual subjects from one another.

What is evident however is that a multitude of measurements from a wearable sensor would be needed to determine an individual signature for a particular user. Activity level for

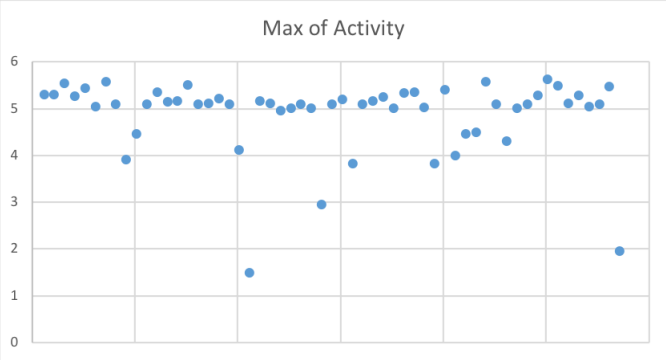


Fig. 11. Scatter plot of max values from ExtraSensory/Hexoskin data

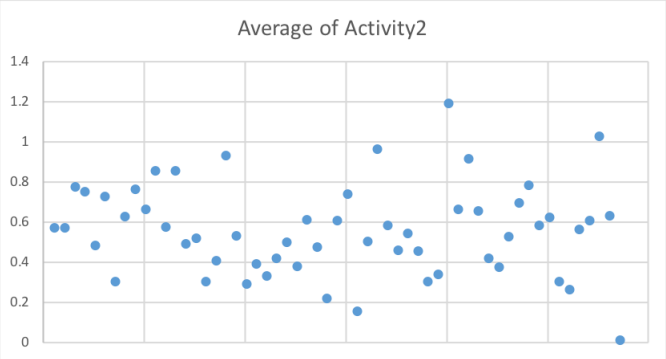


Fig. 12. Scatter plot of average values from ExtraSensory/Hexoskin data

example, paired with corresponding heart rate could prove to be useful in distinguish one particular wearable sensor user from another.

## VI. Conclusion

These findings regarding wearable sensors were very enlightening. What was readily evident at the beginning of the study was that the research field of wearable sensors as well as the consumer market for wearable sensors is largely untapped territory that is in a period of defining what part these sensors will play in the future. In addition, sensors have evolved to the point that consumer acceptance of their use in daily life has resulted in marketable consumer products. This initial success has meant that understanding how they work and analyzing the data that these sensors has produced becomes increasingly important.

Another attribute of these wearables was they have varied methods of storing and accessing their data. Due to the varied ways in which these devices deal with their data, work had to be done with different APIs and other methods to extract and analyze the sensor. In devices such as Hexoskin suit and Fitbit, it is possible to extract data via their corresponding APIs and download it from their respective web sources onto local machines.

Most consumer level wearable sensors operate in a similar fashion. They are largely dependent on traditional computing devices to which they sync, and store their data. These sensors normally transmit that data wirelessly usually through Bluetooth. The sensor's data is also sent to remote servers which facilitates pushing that data to other devices, and web access. The data that wearable sensors generate is often comprehensive. This data includes health data such as resting/active heart rate and other vitals as well other personal data such as geolocation.

Thus, going forward it is important that companies that produce wearable sensors take greater responsibility to ensure the integrity of user data. While a central standard for data processing and analysis may not be possible, it is conceivable that companies can make inroads towards more useful utilization of sensor data. It is also important for sensor manufacturers to continually update their data analysis processes according to the latest methods and communicate to their users how exactly their data is processed. This way both consumers and professionals can get the most out of the multitude of data points generated by wearable sensors. Further research in wearable sensors can examine how data collection and analysis has evolved, as well as best practices to ensure companies enact relevant, accurate data analysis procedures.

Wearable sensors will only become even more ingrained in the world. A future where healthcare is revolutionized by sensors that can not only detect vitals but potentially predict medical conditions is entirely conceivable. However that future will only happen if it is built on a solid foundation where wearable sensor data is accurate, valid and made useful for everyone.

## Acknowledgment

The authors would like to thank Prof Charles Tappert, PhD.

## References

[1] "R language,https://www.r-project.org/."

[2] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. Franklin, S. Shenker, and I. Stoica, "Fast and interactive analytics over hadoop data with spark," *Usenix Login*, vol. 37, no. 4, pp. 45–51, 2012.

[3] "Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios, https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5038811/."

[4] "Assessment of security vulnerabilities in wearable devices, http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1204&context=ism."

[5] "The rise of consumer health wearables: Promises and barriers, http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001953."

[6] "The effect of data security perception on wearable device acceptance: A technology acceptance model, http://csis.pace.edu/ ctappert/srd2017/2017pdf/d11.pdf."

[7] "Smartwatch security fails to impress: Top devices vulnerable to cyberattack, http://www.zdnet.com/article/smartwatch-security-fails-to-impress-top-devices-vulnerable-to-cyberattack/."

[8] "As smartwatches gain traction, personal data privacy worries mounts, https://www.computerworld.com/article/2925311/wearables/as-smartwatches-gain-traction-personal-data-privacy-worries-mount.html."

[9] "Wear it or share it? wearables and security., http://gala.gre.ac.uk/16310/7/16310

[10] "How analytics can take wearables to the next level, https://datafloq.com/read/how-analytics-can-take-wearables-to-the-next-level/2037."

[11] S. Hussain, B. H. Kang, and S. Lee, "A wearable device-based personalized big dataanalysis model,http://uclab.khu.ac.kr/resources/publication/c_304.pdf," 2014.

[12] H. Banaee, M. U. Ahmed, and A. Loutfi, "Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges,http://www.mdpi.com/1424-8220/13/12/17472," 2013.

[13] J. C. Davila, A.-M. Cretu, and M. Zarema, “Wearable sensor data classification for human activity recognition based on an iterative learning framework,https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5492798/,” 2017.

[14] Ujwal Pathak, Shalwak Dangore, *Python code we wrote to harvest data*. Available at https://github.com/ujwalp1994/Capstone_Project1.

[15] “Hexoskin api, https://api.hexoskin.com/docs/index.html#overview-and-technology.”

[16] “Ideal heart range, https://www.mayoclinic.org/healthy-lifestyle/fitness/expert.../heart-rate/faq-20057979.”

[17] R. Rawassizadeh, M. Tomitsch, M. Nourizadeh, E. Momeni, A. Peery, L. Ulanova, and M. Pazzani, “Energy-efficient integration of continuous context sensing and prediction into smartwatches.”

[18] Y. Vaizman, K. Ellis, and G. Lanckriet, “Recognizing detailed human context in-the-wild from smartphones and smartwatches,” 2017.