Big Data uses in Crowd Based Systems

Cristian Chilipirea, Andreea-Cristina Petre, Ciprian Dobre
University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers,
Department of Computer Science, Bucharest, Romania
{cristian.chilipirea; ciprian.dobre}@cs.pub.ro
andreea.petre@cti.pub.ro

Abstract. There are currently many trends in computer science, like Smart Cities, Internet of Things, and Wireless Sensor Networks. Many of these systems require or could dramatically benefit from having information about crowds. First of all, many of the systems are built to improve the life of people, and they require information about them to be able to know when to activate their functionality in order to help them. Secondly people represent a dynamic component of the entire systems, which is unpredictable. Measuring crowd dynamics is not an easy task. Each city consists of millions of individuals and their location needs to be known at all times. Furthermore, the other systems need to be able to extract the needed information for them to be able to function correctly while maintaining every individual's privacy. With crowd dynamic understood we open the way to the opportunity that is given by crowd sensing systems. Systems where data is gathered by sensors carried by individuals.
Even more, crowd dynamic information can be supported by context, context that can be gathered from multiple sources, mostly available free on the Internet. With the vast amount of data on crowd dynamics and the context that surrounds them, the only option to extract information from these systems is given by Big Data. This is where Big Data meets crowd sensing. By having accurate, correct analysis of the crowd data and its context, the information extracted can be used by all other systems in order to be able to take smart decisions.

Keywords: crowd management, crowd sensing, crowd dynamics, big data.

1 Introduction

Crowd Sensing represents a new paradigm whose potential and limitations still require a lot of research. Classically if someone wanted to gather information about an area she would deploy a large enough number of sensors that provided the required data. This solution has some severe limitations. First of all, the number of sensors can be extremely high, depending on the application. Take for instance, measuring the noise level inside a city. This would require microphone-like sensors deployed on every street. These sensors would need to communicate in order to send the data to a central location. This in turn means deploying a wire or wireless infrastructure. Furthermore, the sensors require energy. They can either use high power expensive batteries, that require maintenance or a connection to an electrical infrastructure, which is not easily available. All this problems: the sensors, the communication, the power requirements

impose an extremely high price for the simple application of noise monitoring. And the problem doesn't stop with the price. Sensors need to be replaced when they are broken, they need to be protected so that the quality of the measurements is not disturbed and because they are static they are limited in their view of the noise in parts of the city where they are not deployed.

Here comes the idea of using a crowd. Instead of deploying an expensive, difficult to maintain infrastructure, one could use the power of the crowd in order to obtain the same, or even better measurements. The power of the crowd is well known. The first example of this was made by Wikipedia [1], where a collection of articles, created and edited by individuals across the Internet, built the largest encyclopedia. This work was not managed or coordinated in any way. All participants contributed with as much or as little as they felt.

The idea of using a crowd is further explored by applications such as Amazons Mechanical Turk [2]. This is a marketplace application that proposes the use of crowds in order to solve small tasks in exchange for small amount of money. Many similar specialized marketplaces have appeared over time.

All these systems are based on crowd sourcing. Having a large number of individuals solve one small problem. In most cases crowd sourcing requires the direct involvement of the individual in order for the problem to be solved.

Crowd sensing is a subset of crowd sourcing. If crowd sourcing proposes the use of crowds in order to solve a problem, crowd sensing requires that the problem be of a "sensing" one by nature. The previously presented noise problem is a sensing one.

To take advantage of crowd sensing in order to solve the noise monitoring problem one would require only a simple application, installed on the phones of many users. Smartphones are an ideal platform: they are now ubiquitous; they all have one of three popular operating systems, Android [3], iOS [4] or Windows Phone [5]; they are very powerful complex pieces of hardware with a processor, internet access and many sensors; and finally they are very mobile. Particular to the noise application, they also have microphones already included in them.

With an application that takes microphone data, process it to determine a noise level, and sends it to a central location one can use the power of crowds to monitor large cities. This is all done with minimal costs or resources. Because of human mobility, many different areas are constantly being monitored. The only thing that is required is the good will of the crowds, and this has been shown time and time again through platforms such as Wikipedia. Furthermore, the goodwill can always be replaced with incentives, monetary or of different nature, such as gamification.

Multiple crowd sensing tasks can be run at the same time. For instance, one task can measure the noise level, while another measures the pollution level or even pedestrian density. With more tasks, more data is being generated and more difficult it is to process. But with more data more information can be extracted and specialists can make more sense of what is going on.

There are multiple factors that introduce an extremely high scale to the crowd sensing problem: Large number of users that generate the data; Multiple concurrent monitoring tasks; Constant monitoring, all devices are periodically gathering data. This creates an extremely large volume of data.

Volume is the first of the "three V's of Big Data" as defined by Gartner [6]. The V's are: Volume, Velocity and Variety. Crowd sensing data has all these features. The data is constantly being generated at a very high speed by many sensors, this represents Velocity. The data can be generated by different, complex sensors, representing Variety. The volume is given by the scale of this data set.

IBM added a 4th V as a Feature of Big Data [7]. This V represents Veracity, the fact that the data is uncertain, or unclear. This is especially true in crowd sensing, where there can be only little trust in the data being generated.

Crowd sensing raises interesting problems that never existed in the case of any other sensing system. Crowd sensing is extremely dependent on crowd dynamics. Sensors are where people are. This does mean they can cover a large area, but this also means they are not always available. For instance, take a task of sensing pollution levels in a city. During the night people are mostly inside their homes and very few are outside offering readings. In the day, most people are mobile and so are the sensors they are carrying, covering large areas and delivering large amounts of relevant data. This makes it very important to firstly understand crowd dynamics so that crowd sensing data can be correctly evaluated and used.

Fortunately, there is currently a large amount of research that tries to make sense of crowd dynamics. Crowd tracking applications make use from anything from GPS, WiFi signals to inertial sensors. The data from these applications have very similar properties to the crowd sensing data. There is a high volume of data, given by the large number of individuals, with a high velocity given by the constant location updates as well as high variety and veracity given by the different methods in which this data can be acquired.

In the next section we offer related work and a motivation for this research. This is followed by a detailed analysis on Crowd Sensing systems. Next we present different methods of measuring Crowd dynamics, information relevant to the crowd sensing systems. We continue by presenting different types of Context that can be used in conjunction with Crowd tracking and crowd sensed data. Finally, we discuss how Big Data can be used to extract information from this data sources and finish with our Conclusions.

2 Related Work

Crowd based systems are raising more interest. More and more research projects aim the construction of such a system. In this chapter we are interested in two categories of crowd based systems, in systems that use the crowd for the purpose of sensing and in systems that try to monitor a crowd.

Sensing data using a crowd, or simply called crowd sensing represents an interesting promise, of cheap, scalable, powerful data gathering system. These systems make use of the mobility of crowds in order to extract sensed from multiple locations. On overview on crowd sensing and a listing of many possible applications is available in [8].

With the increase of interest in crowd sensing systems, in recent years, many platforms have appeared that try to implement and solve this problem. Medusa [9] represents the most popular crowd sensing platform. It consists of a programming framework where people can define sensing tasks. These tasks are then spread to the users and they can gather the sensor data and forward it to the requester. In order to reward the users Medusa uses Amazon Mechanical Turk [2].

In comparison Metador [10], [11] represents a platform for crowd sensing that concentrates on context information. Tasks are created and delivered only to the individuals that are in the right context. For instance, information of a crowd sensing campaign requiring a picture of a building is sent only to individuals in proximity of the building.

Mosden [12] represents a platform for crowd sensed data whose main idea is to separate logic behind communication, storage and data acquisition. The platform is demonstrated using an application that uses crowd sensed data to determine noise pollution inside a city.

Another platform for crowd sensing is implemented on top of Cupus [13]. It uses a publish/subscribe interface based on the cloud. It is meant for Internet of Thing systems. The platform is used with an application for air quality monitoring. The air quality sensors are small dongles that can connect with Bluetooth to a smartphone. Using the Cupus publish/subscribe systems the data from multiple such devices is collected and processed in the cloud.

Probably the first platform that can be used for crowd sensing purposes is the mCrowd platform [14]. It integrates ChaCha [15] and Amazon Mechanical Turk [2] and offers tasks set on these systems on mobile platforms. The tasks take mostly the form of image or text responses and require human interaction to complete.

Platforms for mobile crowd sensing have appeared in order to make it simple to deploy crowd sensing campaigns or applications. There are a lot of problems that make it difficult to deploy crowd sensing applications. Manny of them are detailed in [16]. Among the problems identified we note: heterogeneity of mobile devices software and difficulty of installing applications, especially if each crowd sensing task requires a different application. There are also problems of scalability and resource utilization. Platforms for crowd sensing have the potential of solving most problems. With the large number of crowd sensed platforms there are also multiple applications for crowd sensing. For instance, in [17] the authors present MAP++ a system where semantic data is automatically added to maps using a crowd sensing system. Their system uses only sensors from inside smartphones, such as the accelerometer, and require no user interaction. Multiple road features can be detected such as bumps or roundabouts.

However, their system is calibrated to work only for smartphones carried inside vehicles.

A similar system that uses only sensors that already exist is smartphones is presented in [18]. Here the authors show how Bluetooth modules can be used in order to determine the density of people in different areas. Their application makes use of multiple smartphones that continuously record Bluetooth signals, making it a crowd sensing application. The authors prove that their system has an accuracy of 75%.

The Mahali project [19] represents probably the most daring of crowd sensing applications. It proposes the use of a dual frequency GPS receiver, in order to measure ionosphere features. This information can be used to get high resolution weather maps. Smart phones are likely to have dual frequency GPS receivers in the near future due to the need of precise localization. In the meantime, they are still useful to fill the gap in connecting remotely placed static sensors, without network access by being an intermediary node in connecting it to the internet. This method is called delay tolerant networking.

Crowd sensing can be used to further many science experiments that would otherwise require large costs and man power to complete. In [20] the authors propose a crowd sensing application intended for use by students. The application asks the users to go to specific locations and take pictures of different plants. These pictures are then used by environmental scientists in order to better understand the flora of the campus. The application has as basis the idea of gamification, the incentive of taking the pictures being a simple set of in-game points. The method of gamification is called floracahing and is similar to geocaching [21].

Not all applications that need sensing data require the deployment of a new crowd sensing campaign. For instance, in [22] the authors use Foursquare data in order to determine the growth of cities. Because of the scale of this data set popular locations in cities can be observed and the increase in density of these locations can be measured. Cities themselves have require many types of data if they are to become what is known as smart cities. The authors of [23] conduct an in depth look of multiple types of sensing that can be done in smart cities. Many of these can be applied to crowd sensing.

With all these platforms and application for mobile crowd sensing appearing in recent years it is clear that the potential given by these system is quite large. But, crowd sensing platforms and solutions all suffer from privacy issues. A few papers try to solve this problem. The work presented in [24] takes privacy as the most important consideration. In order to preserve the privacy of people generating the data they propose a cloud based approach with a system that gathers the data and obfuscates it before it is being sent to the data requester. Similarly, the authors of [25] create an application that provides anonymous authentication to their service. This application of crowd sensing permits park operators to determine important characteristics of the crowds that move around their park. For instance, the operators are able to determine queue lengths and even recommend better routes for incoming visitors. A survey on privacy of crowd sensing applications is available in [26].

Other works try to improve the reliability of crowd sensing systems. The authors of [27] tackle the problem of how to insure that crowd sensed data is accurate. For instance, when an application requires pictures from an event, how to confirm the pictures are actually valid. The solution of the authors is given by a Bluetooth communication between devices. Multiple devices use their GPS data and increase the trust that the image is indeed taken in that location. This is done without any human intervention. A similar solution is provided in [28] where trust in the data is treated. The requester needs to have trust in the data being sent by random users. In order to improve trust this paper proposes a way of discovering the "truth" in data by correlating multiple reports.

Finally, one has to determine the optimal number of users for a crowd sensing application. In order to determine this number a simulation environment is proposed [29]. This environment is specifically targeted at urban parking. The information extracted from the crowd sensing application being used in order to determine where a parking lot is available.

In order to completely understand crowd sensing and the data it offers, it is vital that data analysts have access to crowd dynamics information. There is a need to firstly understand the position of individuals that participate in the sensing process and then to understand the flows of crowds.

Measuring crowd dynamics is classically done using visual systems [30]. These system consists of a number of video cameras placed around the city and powerful software that can identify a face and track it across multiple cameras. Because facial recognition is still not a completely solved problem and the results are highly probabilistic, the results of these systems are not very accurate. Furthermore, implementing this type of systems requires huge investments in infrastructure, starting from cameras and the communication network for their data to be gathered at a central location, to large processing centers required to run the face recognition algorithms.

Other solutions have appeared in the literature. GPS [31] is the most known and most highly used positioning technology. It is present in most modern cars and in almost all smartphones, which are now ubiquitous. This solution has been implemented in the case of large crowd monitoring [32]. Taking advantage of the large number of smart phones the authors built an app that gathered GPS data from any device that had the app installed to a central location. They proved the feasibility of this method by deploying it during a three-day Swiss festival.

Requiring people to install an app in order to provide the required information is not a solution that scales very well. Individuals need to have a level a trust in the app and its use needs to be widespread, which in turn might require incentives. Methods that do not require the involvement of the individuals being tracked are preferred.

An increasingly popular alternative is given by the use of communication signals. Whenever a phone transmits data it uses one of several standardized protocols. By having scanners that can record these signals it is then possible, within a limited accuracy, to identify where a device is and how it is moving. Because protocols such

as WiFi includes the address of the device within most packets it is possible to track one device across multiple scanners.

Using WiFi in order to track crowds of people and understand human mobility has been done in for music festivals [33] or for campus monitoring [34]. More interesting applications try to use this type of data in order to inform on building facility planning [35] or even measure flows of people through airport checkpoints [36]. The technology has also been proved useful in measuring human queues [37].

Other works have searched ways of improving the use of WiFi scanning techniques. In [38] we try to show a number of filters that can be used in order to obtain a smaller data set with less noise. The solution for the RSS variance problems in this type of monitoring is solved presented in [39]. Improvements have also taken place in the form of building better tools for visualizing movement data [40].

In the search for accurate device positioning and tracking some works have tried using multiple technologies at the same time. This is apparent in [41] where multiple communications methods and magneto metric sensors are used at the same time, or in [42] where WiFi is used at the same time as inertial sensors.

3 Crowd Sensing

Crowd sourcing is becoming a very powerful tool for solving large tasks (even at a scale that makes them seem impossible) that can be broken down in a large number of simple tasks that require just a bit of human involvement. This idea is the basis of Wikipedia [1], a website where anyone can contribute to build one massive, open encyclopedia. This same website for instance offers an extensive list of cases where crowd sourcing is used in order to solve otherwise impossible tasks.

Crowd sensing is a part of crowd sourcing where the people inside the crowd are tasked with gathering sensor data. The authors of [43] define crowd sensing as a natural extension of participatory sensing. Where participatory sensing represents a task where the crowd is used to gather sensor data, and crowd sensing uses this data in conjunction with offline or previously gathered data. An even more important difference is given by the lack of need of participation that was implied by the participatory sensing systems. For instance, a campaign that tries to measure noise pollution requires that the mobile devices send data with regards to noise, but they don't require the mobile device owner to actively do anything. In contrast a campaign that requires images of a building or of an event, requires that the owner of the mobile device actively take these pictures.

To have a crowd sensing campaign, one requires that individuals inside an area carry a device that has the sensing capabilities required by the crowd sensing campaign. The data gathered by the devices then needs to be sent to the campaign manager, the one that requested the data, or to an individual to whom the data is useful.

Because of the popularity of smartphones, it is now possible to have large scale deployments of crowd sensing applications, where as many users that own smartphones

can participate. Smartphones are ubiquitous and with this the popularity of crowd sensing applications has grown. However, all crowd sensing applications that make use of smartphones are limited by the features that are found inside them. A solution is to extend the capabilities of smartphones with external hardware, like a Bluetooth dongle that contains only the needed sensor. An example of this is available in [44]. The smartphone is still an important part of the system because with its powerful CPU and communication capabilities it dramatically lowers the price of the scanner. With powerful sensors, such as pollution sensors, or sensors that can detect humidity or pressure, crowds can gather all types of data.

Campaigns that require external sensor are more difficult to deploy. They are more expensive because of the extra hardware and they require people to carry another device with them, which is inconvenient. Smartphones are now powerful computers with an already large variety of sensors. These sensors are the most popular for crowd sensing campaigns. The most popular sensors inside smartphones are:

- Accelerometer – The accelerometer in smart phones is commonly used as an inertial sensor, capable of inferring the speed with which the person is moving as well as offer support for many other applications as impressive as gesture recognition [45] or even authentication [46]. When a large number of these sensors are used applications such as earthquake detection and measuring [47] can be implemented.
- GPS – This sensor is used primarily for localization purposes. In conjunction with maps and interactive applications it can be used to assist with choosing the shortest path through a city. With many of these sensors crowds and congestions can be detected [48].
- Photo/Video Camera – Smartphones are now capable of taking high resolution photos and videos. People use them extensively and post photos on social networks and these can be used in crowd sensing campaigns [49]. The photos can also be used to enable citizen journalism [50].
- Microphone – Communication remains the main feature of a smartphone and microphones are the central sensor that enables it. With enough microphones crowd sensing campaigns can measure noise pollution [51].
- Magnetometer – Are sensors that detect magnetic fields, they are used in conjunction with accelerometers in order to improve positioning.
- Thermometer – Are mainly used to measure the temperature of the CPU to insure that it is working appropriately. However, these sensors can also be used to measure the temperature of the environment [52].
- WiFi – WiFi represents the module used to communicate data, mainly for internet usage. Because of the details of the 802.11 protocol this module can be used in order to detect other WiFi enabled devices through WiFi scanning [53]. Making a list of static WiFi devices is done through a technique called war-driving [54] and it can be used to improve localization.
- Bluetooth – Similar to WiFi, Bluetooth represents a data communication technology. It is mainly used for personal area networks and is meant to

connect smartphones to wearables or headphones. They can be used to facilitate communication with other devices and this means they can also be used to detect them and with them the size of crowds [18].

The data gathered by the sensors can be sent to a central authority. Alternatives are to store the data until an Internet connection is available or to send the data in a distributed manner. For instance, if the sensors measure the traffic on a street, the information can be useful to people located on adjacent streets.

In order for any crowd sensing campaign to be successful it requires people to participate in the crowd sensing process. This means people have to be incentivized in order for them to participate in the campaign. There are many articles that propose incentives for crowd sensing systems [55], [56], [57], [58] as this is the primary assumption that needs to be addressed in order to implement any such campaign.

With appropriate incentives crowd sensing systems can reach the required number of users. Because the sensors are usually small range, take for instance microphones which work for only few meters, this means lots of people need to take part in the crowd sensing system in order to be able to monitor large areas.

Crowd sensing itself is usually opportunistic. Sensed data is only gathered in locations where people are. If they don't go to different areas, then there will simply be no data from those areas. In order to improve this, active systems could be implemented in order to guide people to where sensed data needs to be gathered. This is the case for floracache [20] where people are told where to go to gather the data.

Outside of smartphones there are a few other classes of devices that can prove useful for crowd sensing initiatives. Cars now have many powerful computers and sensors as well as communication capabilities with central locations or even other cars [59]. The computers and sensors in these cars can be repurposed in order to participate in crowd sensing networks. Because of their large volume and carrying capabilities, cars can be equipped with larger more complex sensors. We ask the reader to consider if it is not fit for the cars that contributed to pollution to be the ones that are tasked to measure it.

Wearables [60] represent a large class of devices that people can wear. They are now popular in the form of smart watches and fitness bands, but they can take many forms. They have the advantage of being in extremely close proximity to the user. With increases in processing power and multiple sensors they will soon become another interesting alternative for crowd sensing.

To our knowledge there is no real open data set available that shows the results of a crowd sensing campaign.

With many alternatives for applications and many available platforms crowd sensing is likely to be an important part of our daily lives. But in order to understand the accuracy and the benefits of crowd sensed data it is vital that there is access to crowd dynamics information. Understanding crowd dynamics enables us to measure the quality of the sensing data, by understanding what areas are covered and which are not,

as well as be able to predict which areas will be covered in the future. This even enables campaign runners to deploy static sensors in areas that do not get enough traffic.

4 Measuring Crowd Dynamics

Crowd dynamics represent all movements and stationary actions that crowds make. They can apply to a small area such as an indoor facility or be as large as an entire city. Crowds display flows and they can merge or split throughout their movement. They are affected by the density of people that make out a crowd.

Ideally a crowd tracking system would be able to offer a precise positioning for every person at any time. No known system is capable of such a task, but instead they all offer approximations.

The most popular method measuring crowds and tracking them is by use of video feeds. This is mostly made apparent by CCTV systems [61]. These systems however require expensive infrastructure and large processing centers aimed at analyzing the video feeds. Because face recognition and individual tracking is still difficult to implement with computers, these systems still have a large room for improvement.

Newer systems make use of smartphones which are considered ubiquitous. This can be done by installing an application of the devices of many citizens or by scanning for the communication signals they send. The application can use GPS positioning to determine the location of the device. Unfortunately, GPS only works outdoors and has large errors, in the order of meters. Scanning for WiFi hotspots and matching them with locations obtained from war-driving represents a possible alternative which has the advantage of being energy efficient. However, the accuracy is even lower than GPS and the exact location of hotspots is not always known.

Installing software at a large scale is difficult. People need incentives and they are worried about the security of their device as well as their own personal privacy. Because of this, methods that simply scan for communication signals are simpler to implement and deploy.

It is important to understand that all communication signals suffer from a high variety of noise sources. First of all, not all devices are built the same. Then, the weather or surroundings can affect the way in which the signal propagates. This phenomenon is made even worse when we consider mobile features of the surroundings such as cars or other pedestrians.

Not all people own smartphones and from those that do own them not everyone uses the WiFi module. Other people stop the WiFi module when they are not using it due to energy consumption constraints and the fact that they wish to preserve battery. This means that only a portion of the population is being tracked using these systems.

Another important factor that needs to be considered when deploying crowd monitoring applications is the privacy of the people being monitored. The data needs

to be obfuscated in a way in which it remains useful but hides the identities of all the individuals being tracked.

If the crowd dynamics information is obtained by deploying applications on the user devices this makes it a crowd sensing application that is able to reveal for instance density. There are many applications that can result from this type of data, for instance we were able to infer the map of Roma [62] given only GPS localization data from several taxis.

5 Crowd data for crowd dynamics

As we mentioned previously crowd sensing data is not openly available. We did manage to identify an important source for open crowd tracking, and crowd related data. This source is called crawdad [63] and it hosts a large number of scientific data sets related to wireless data.

In order to have a better picture of what crowd tracking data is and what it can be used for we make a small analysis on five data sets from the crawdad website. All these data sets contain localization data.

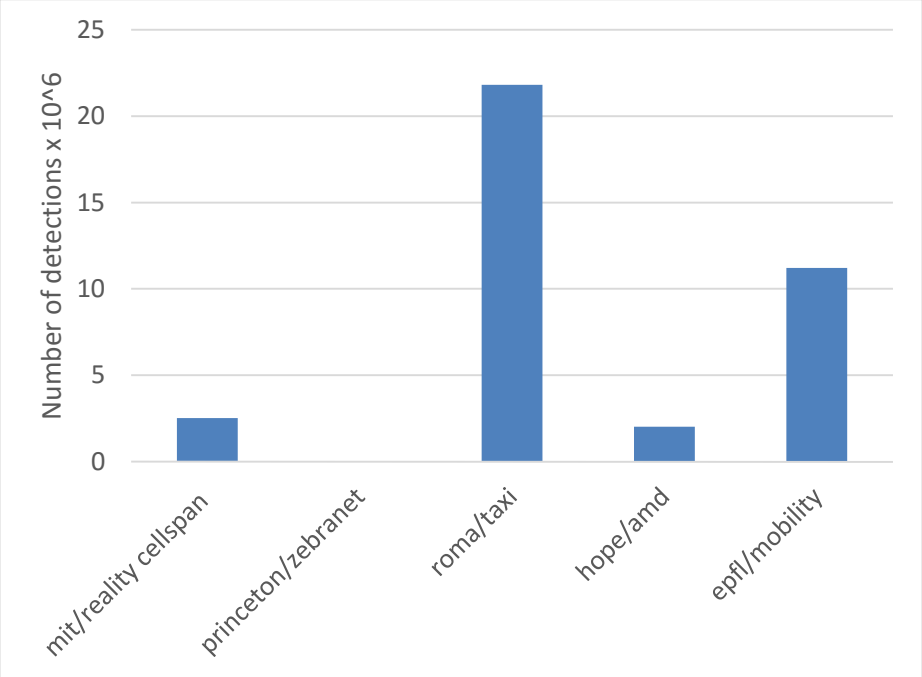

Figure 1 # of detections

The five data sets we chose are:

- Reality [64] – Represents a data set gathered from 100 subjects from MIT during an academic year. The people, consisting mainly of students carrying mobile phones capable of recording the cells of GSM towers.
- Zebranet [65]- Applied on the Sweetwaters Game Reserve near Nanyuki, Kenya. Zebras were fitted with collars that contained a GPS receiver, a CPU as well as storage and wireless transceiver. Their system uses opportunistic communication in order to gather the data set.
- Roma Taxis [66] – In the city of Rome taxis were fit with special devices and software that recorded their movements.
- Amd [67] – At the hackers on planet earth conference guests were given RFID tags. These tags were detected by static sensors placed in different rooms of the conference.
- Epfl [68] – Similar to the Roma data set, the data is gathered from taxis. This data set is gathered in the city of San Francisco.

The statistics on the five data sets are available in Figure 1 where the number of recorded data elements is displayed. In Figure 2 we compare the number of devices in each of the data sets.

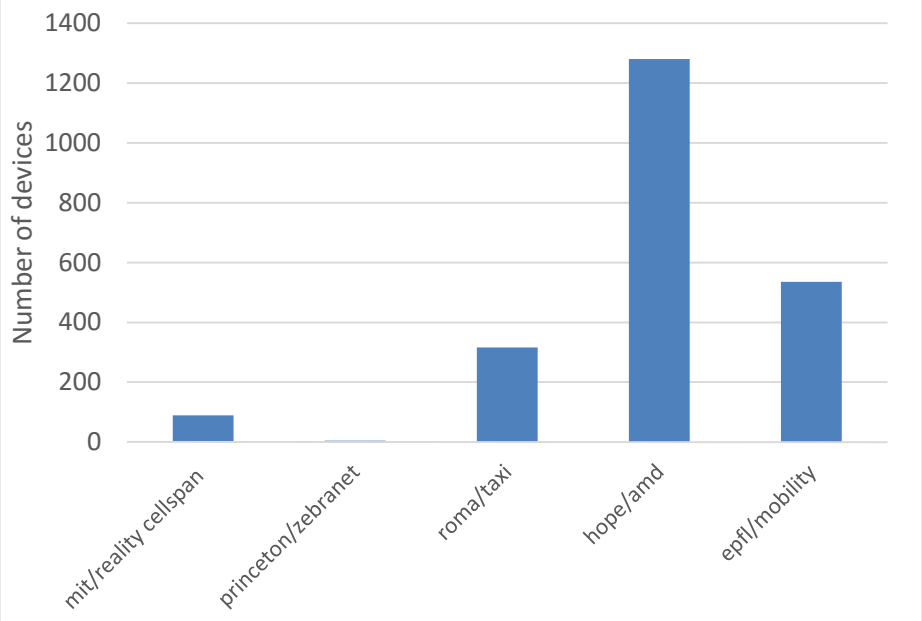

Figure 2 # of Devices

Figure 3 shows the duration of each of the data sets. These first 3 figures show how diverse the data sets can be. They can contain data recorded over any number of days with a large variation in the number of elements in the data set or the number of devices. The large variety in the data sets permits anyone to do analysis on very different scenarios and use cases.

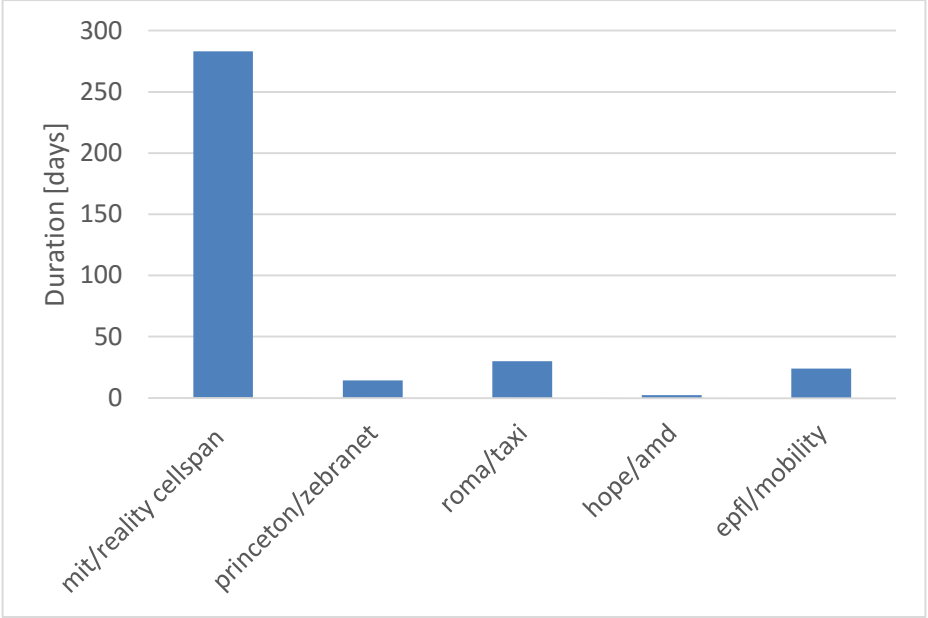

Figure 3 Duration

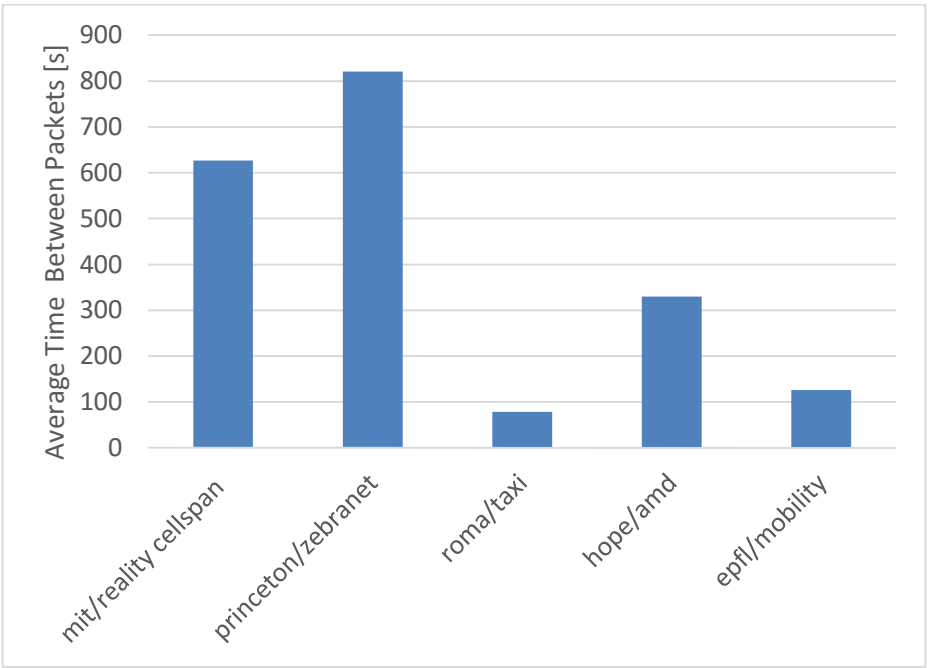

Figure 4 Average Time Between Consecutive Packets

It is not only the scenario that differs but the technology used to gather the data set. These technologies span from GPS recordings to WiFi detection to even RFID tags. Because of this difference there is also a difference in the average time between two consecutive detections of the same device. This is best made apparent in Figure 4. This difference is made not only by the technologies but by the settings and the behaviors of the participants.

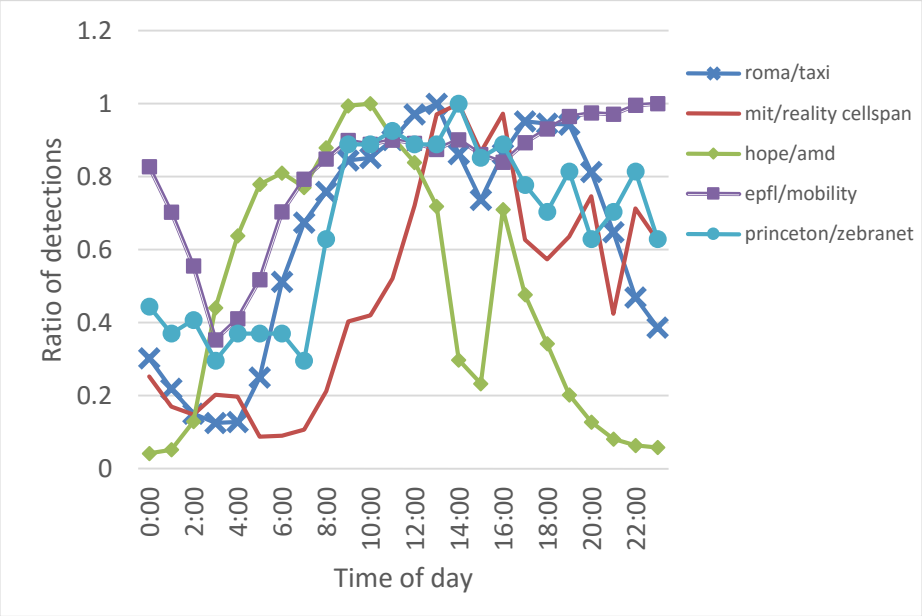

Figure 5 # of Detections Over Time

Finally, we tried to identify the similarities between the data sets. Given our experience with various crowd tracking data sets we learned to expect a day night pattern in any data set that involves people. We managed to identified this feature in all of the data sets. In Figure 5 we display the number of detections or data set elements as they change during the day. The data is normalized by having a ratio to the maximum number of detections for each data set. The day was chosen randomly from the data sets. It is clear that in all of these the number of packets raises during the day and drops during the night.

We offered only a small overview on some characteristics of crowd tracking data. Further analysis can reveal more interesting features such as patterns that people take each day in their movements. Obtaining these features and this information requires extensive data analysis usually done with the help of Big Data.

6 Context

Context represents all types of data set that can help in making more sense of the crowd data. Unlike the data sets we presented in the previous sections, context data represents unstructured data. It usually comes in the form of text and requires natural language processing systems [69]. Unlike the crowd sensed data and crowd tracking data, the origin of context data is not clear. There are various possible open internet sources as well as closed ones. The trust put in the data is dependent on these sources.

We have identified a few potential context data types:

- Weather – It directly affects the behavior of people. In case of bad weather people are more likely to remain indoors. This can explain a drop in the number of individuals performing crowd sensing outside.
- Schedules – There are many types of schedules, from the mostly static once of shops and school, to schedules of complex events. Using schedules one can better understand the reasoning behind crowd movements. This is made specifically apparent in day/night patterns observed in crowd data.
- Social network data – People post a high variety of information on social networks from picture to restaurant rating. Based on this data it I reasonable to expect certain behaviors in crowd movements. For instance, restaurants with bad reviews are avoided while the ones with good ones have a lot of traffic.
- News – Various events are not predictable. This is especially true for high impact disasters. News sources can be used in order to understand what is going on, beyond the normal schedule or social network information.

There are many sources of context data and many types of it. Making sense with it and matching it with the crowd sense data requires a lot of resources. This is where Big Data jumps in. It is specialized in processing large amounts of structured or unstructured data. Because of its variety context can be continuously extended as new sources of context data are identified and integrated with the existing ones.

7 Crowd data as part of Big Data

Big Data [70] represents a new paradigm in data processing. It goes beyond traditional database queries where extracting information was done in a straight forward manner. In contrast, instead of giving simple information such as the maximum or average of a data set, big data queries use machine learning techniques [71] to offer answers without the need for a question. They offer new information about the data without the analyzer even knowing what to look for. A good example of this is usually given in machine learning courses as the “beer and diapers” example. At a large supermarket chain Big Data analysis revealed that people who buy diapers also buy beer. The example sticks because it is clear how unlikely it is that someone would search through data sets to see if people buy diapers and beer at the same time, and the result has a simple application, place the 2 products next to each other. This type of

information can relevant for crowds. Take for instance groups of people that visit the same two shops in the same order. This would enable the prediction of crowd flows.

We showed in the previous sections that crowd data, be it crowd sensing or crowd tracking, exhibits the features of Big Data. Crowd data has all four Vs: Volume, Velocity, Variety and Veracity. In order to better fit with Big Data we propose the use of crowd sensing as well as crowd tracking. Crowd tracking provides important information as to where the crowd sensed data is gathered from. Context with its many forms bring only improvements to these two data sets. And all of them taken together pose an interesting challenge to Big Data and opens the door to many applications.

By applying Big Data techniques on crowd data we open the door to what is known as smart cities [72]. Places where real time data and information offers better living conditions, less traffic and overall increase in the quality of life. The information extracted permits automated systems that respond to citizen needs and better planning for the authorities.

But Big Data doesn't show only human behavior in order for decisions to be made based on past data, it can also help in making predictions on what will happen. For instance, when a large event is organized, having knowledge of possible crowd behaviors can help with many of the decisions.

Another important aspect is given by disaster scenarios. By having real life data as well as appropriate models of human movement, the rescue teams can respond faster, with more knowledge and more information. Even more, facilities can be built in such a way that evacuation is simple and it fits with appropriate models of human behavior.

Let's finish with a possible example of Big Data application with crowds. This is just one of many possible uses. Given a crowd tracking data set, which enables us to know where people are gathered and what their flows are, as well as a crowd sensing data set which enables us to map the noise level throughout the city. In the case of an unexpected event, we have more information on the behavior of crowds and how better to guide them. Adding context data can offer even more insight, it can determine if the event was a music event starting or if it was a dangerous event and emergency response units need to be deployed. The decision can even be taken before anyone even succeeds in reporting what happened.

Big Data opens the doors to a variety of applications and use cases that may be unimaginable, the more we explore it the more we can push its limits and be able to build truly smart cities and smart environments.

8 Conclusions

In this chapter we presented crowd sensing and crowd tracking applications. We discussed the high potential they have and their current limitations. We continued with a discussion on different types of context and how this context information can have a high impact in extracting information from crowd sensed and crowd tracking data.

Finally, we show how Big Data can be used to put everything together. We showed that crowd data is a type of Big Data and by using multiple crowd data sources in conjunction with context we can obtain various and even surprising new information.

Because the field is still young there is a lot that remains for future work. Multiple crowd sensing applications need to be deployed at a large scale and the data needs to be exchanged in an open manner. Privacy implications need to be considered in order to make this possible. Once the data is open multiple groups can identify different ways to best analyze and extract information from the given data sets. The multitude of crowd data sets and the high potential they have opens a lot of doors for future applications.

9 Acknowledgment

The research presented in this paper is supported by projects: MobiWay, Mobility beyond Individualism: An Integrated Platform for Intelligent Transportation Systems of Tomorrow - PN-II-PTPCCA-2013-4-0321; DataWay, Real-time Data Processing Platform for Smart Cities: Making sense of Big Data - PN-II-RUTE-2014-4-2731. We would like to thank the reviewers for their time and expertise, constructive comments and valuable insight.

10 References

[1] "Wikipedia," Wikimedia Foundation, Inc., 12 12 2015. [Online]. Available: https://en.wikipedia.org/wiki/Main_Page. [Accessed 12 12 2015].

[2] "Amazon Mechanical Turk," Amazon, 12 12 2015. [Online]. Available: https://www.mturk.com/mturk/welcome. [Accessed 12 12 2015].

[3] Google, "Android," Google, 12 12 2015. [Online]. Available: https://www.android.com/. [Accessed 12 12 2015].

[4] Apple, "iOS," Apple, 12 12 2015. [Online]. Available: http://www.apple.com/ios/. [Accessed 12 12 2015].

[5] Microsoft, "Windows Phone," Microsoft, 12 12 2015. [Online]. Available: https://www.microsoft.com/en-us/windows/phones. [Accessed 12 12 2015].

[6] L. Doug, "3D data management: Controlling data volume, velocity and variety," META Group Research Note 6, 2001.

[7] IBM, "IBM Big Data & Analytics Hub," 12 12 2015. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data. [Accessed 12 12 2015].

[8] G. Bin, Z. Wang, Z. Yu, Y. Wang, N. Yen, R. Huang and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Computing Surveys (CSUR)*, vol. 1, no. 7, p. 48, 2015.

[9] R. Moo-Ryong, B. Liu, T. F. L. Porta and R. Govindan, "Medusa: A programming framework for crowd-sensing applications," *In Proceedings of the 10th international conference on Mobile systems, applications, and services*, pp. 337-350, 2012.

[10] C. Iacopo, D. Miorandi, A. Tamilin, E. R. Ssebaggala and N. Conci, "Crowd-sensing: Why context matters," *IEEE International Conference on In Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 368-371, 2013.

[11] C. Iacopo, D. Miorandi, A. Tamilin, E. R. Ssebaggala and N. Conci, "Matador: Mobile task detector for context-aware crowd-sensing campaigns," *IEEE International Conference on In Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 212-217, 2013.

[12] J. P. Prakash, C. Perera, D. Georgakopoulos and A. Zaslavsky, "Efficient opportunistic sensing using mobile collaborative platform mosden," *9th International Conference Conference In Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom)*, pp. 77-86, 2013.

[13] A. Aleksandar, M. Marjanović, K. Pripužić and I. P. Žarko, "A mobile crowd sensing ecosystem enabled by CUPUS: cloud-based publish/subscribe middleware for the internet of things," *Future Generation Computer Systems*, vol. 56, pp. 607-622, 2016.

[14] Y. Tingxin, M. Marzilli, R. Holmes, D. Ganesan and M. Corner, "mCrowd: a platform for mobile crowdsourcing," *In Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pp. 347-348, 2009.

[15] "Cha Cha," 12 12 2015. [Online]. Available: http://www.chacha.com/. [Accessed 12 12 2015].

[16] X. Yu, P. Simoens, P. Pillai, K. Ha and M. Satyanarayanan, "Lowering the barriers to large-scale mobile crowdsensing," *In Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, p. 9, 2013.

[17] A. Heba, A. Basalamah and M. Youssef, "Map++: A crowd-sensing system for automatic map semantics identification.," *Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 546-554, 2014.

[18] W. Jens and P. Lukowicz, "Bluetooth based collaborative crowd density estimation with mobile phones," *IEEE international conference on Pervasive computing and communications (PerCom)*, pp. 193-200, 2013.

[19] P. Victor, F. Lind, A. Coster, P. Erickson and J. Semeter, "Mobile crowd sensing in space weather monitoring: the mahali project.," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 22-28, 2014.

[20] H. Kyungsik, E. Graham, D. Vassallo and D. Estrin, "Enhancing motivation in a mobile participatory sensing project through gaming," *IEEE Third International Conference on Social Computing (SocialCom) and Privacy, Security, Risk and Trust (PASSAT)*, pp. 1443-1448, 2011.

[21] "Geocaching," 12 12 2015. [Online]. Available: https://www.geocaching.com. [Accessed 12 12 2015].

[22] D. Matthew, A. Noulas, B. Shaw and C. Mascolo, "Tracking Urban Activity Growth Globally with Big Location Data," *arXiv preprint arXiv:1512.05819*, 2015.

[23] H. Gerhard and J. G. Hancke, "The role of advanced sensing in smart cities," *Sensors*, vol. 13, no. 1, pp. 393-425, 2012.

[24] K. Ioannis and T. Dimitriou, "Privacy-respecting discovery of data providers in crowd-sensing applications," *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 249-257, 2013.

[25] K. D. Munirathnam, R. Deng, Y. Li, H. C. Lau and S. Fienberg, "Anonymous authentication of visitors for mobile crowd sensing at amusement parks," *In Information Security Practice and Experience*, pp. 174-188, 2013.

[26] P. Layla, L. Xiong, D. Garcia-Ulloa and V. Sunderam, "A survey on privacy in mobile crowd sensing task management," *Technical Report TR-2014-002, Department of Mathematics and Computer Science, Emory University*, 2014.

[27] T. Manoop, R. Curtmola and C. Borcea, "Improving location reliability in crowd sensed data with minimal efforts," *6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*, pp. 1-8, 2013.

[28] M. Chuishi, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding and Y. Cheng, "Truth discovery on crowd sensing of correlated entities," *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pp. 169-182, 2015.

[29] F. Károly and I. Lendák, "Simulation environment for investigating crowd-sensing based urban parking," *In Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 320-327, 2015.

[30] S. Nils and Maybank, "The advisor visual surveillance system," *ECCV 2004 workshop applications of computer vision (ACV)*, vol. 1, 2004.

[31] B. Rashmi, S. L. Ranaweera and D. Agrawal, "GPS: location-tracking technology," *Computer* , vol. 35, no. 4, pp. 92-94, 2002.

[32] B. Ulf, G. Troster, T. Franke and P. Lukowicz, "Capturing crowd dynamics at large scale events using participatory gps-localization," *IEEE Ninth Internation Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2014 .

[33] B. Bram, A. Barzan, P. Quax and W. Lamotte, "WiFiPi: Involuntary tracking of visitors at mass events," *IEEE 14th International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1-6, 2013 .

[34] K. Eftychia, Sileryte, M. Lam, K. Zhou, M. V. d. Ham, V. d. Spek and Verbree, "Passive WiFi monitoring of the rhythm of the campus," *In Proceedings of The 18th AGILE International Conference on Geographic Information Science; Geographics Information Science as an Enabler of Smarter Cities and Communities, Lisboa (Portugal)*, 2015.

[35] R.-R. Antonio, H. Blunck, T. Prentow, A. Stisen and M. Kjaergaard, "Analysis methods for extracting knowledge from large-scale WiFi monitoring to inform building facility planning," *IEEE International Conference on In Pervasive Computing and Communications (PerCom)*, pp. 130-138, 2014.

[36] S. Lorenz, M. Werner and P. Marcus, "Estimating crowd densities and pedestrian flows using wi-fi and bluetooth," *In Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 171-177, 2014.

[37] W. Yan, J. Yang, H. Liu, Y. Chen, M. Gruteser and R. Martin, "Measuring human queues using WiFi signals," *In Proceedings of the 19th annual international conference on Mobile computing & networking*, pp. 235-238, 2013.

[38] C. Chilipirea, A.-C. Petre, C. Dobre and M. v. Steen, "Filters for Wi-Fi Generated Crowd Movement Data," *10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 285-290, 285-290.

[39] T. A. Wen, Y.-H. Chuang and H.-H. Chu, "Unsupervised learning for solving RSS hardware variance problem in WiFi localization," *Mobile Networks and Applications,* vol. 14, no. 5, pp. 677-691, 2009.

[40] A. Gennady, N. Andrienko and S. Wrobel, "Visual analytics tools for analysis of movement data," *ACM SIGKDD Explorations Newsletter ,* vol. 9, no. 2, pp. 38-46, 2007.

[41] M. Piotr, T. K. Ho, S. Yi and M. MacDonald, "Signalslam: simultaneous localization and mapping with mixed wifi, bluetooth, LTE and magnetic signals," *International Conference on Indoor Positioning and Indoor Navigation (IPIN),* pp. 1-10, 2013.

[42] E. Frédéric and F. Marx, "Advanced integration of WiFi and inertial navigation systems for indoor mobile positioning," *Eurasip journal on applied signal processing,* p. 164, 2006.

[43] G. Bin, Z. Yu, X. Zhou and D. Zhang, "From participatory sensing to mobile crowd sensing," *IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops),* pp. 593-598, 2014 .

[44] X. Chenren, S. Li, Y. Zhang, E. Miluzzo and Y.-F. Chen, "Crowdsensing the speaker count in the wild: implications and applications," *IEEE Communications Magazine,* vol. 52, no. 10, pp. 92-99, 2014.

[45] H. Bastian and N. Link, "Gesture recognition with inertial sensors and optimized DTW prototypes," *IEEE International Conference on Systems Man and Cybernetics (SMC),* pp. 2102-2109, 2010.

[46] M. Rene and H. Gellersen, "Shake well before use: Authentication based on accelerometer data," *In Pervasive computing,* pp. 144-161, 2007.

[47] F. Matthew, M. Olson, R. Chandy, J. Krause, M. Chandy and A. Krause, "The next big one: Detecting earthquakes and other rare events from community-based sensors," *10th International Conference on Information Processing in Sensor Networks (IPSN),* pp. 13-24, 2011.

[48] M. Gustavo and M. Roccetti, "Vehicular congestion detection and short-term forecasting: a new model with results," *IEEE Transactions on Vehicular Technology,* vol. 60, no. 7 , pp. 2936-2948, 2011.

[49] D. Murat, M. A. Bayir, C. G. Akcora, Y. S. Yilmaz and H. Ferhatosmanoglu, "Crowd-sourced sensing and collaboration using twitter," *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM),* pp. 1-9, 2010.

[50] Z. Arkady, P. P. Jayaraman and S. Krishnaswamy, "Sharelikescrowd: Mobile analytics for participatory sensing and crowd-sourcing applications," *IEEE 29th International Conference on Data Engineering Workshops (ICDEW),* pp. 128-135, 2013.

[51] M. Nicolas, M. Stevens and B. Ochab, "Participatory noise pollution monitoring using mobile phones," *Information Polity ,* vol. 15, no. 1, 2010.

[52] A. Siamak, A. Troiano and E. Pasero, "Environment sensing using smartphone," *IEEE Sensors Applications Symposium (SAS),* pp. 1-4, 2012.

[53] A. Arjun, C. Manikopoulos, Q. Jones and C. Borcea, "A quantitative analysis of power consumption for location-aware applications on smart phones," *IEEE International Symposium on Industrial Electronics,* pp. 1986-1991, 2007.

[54] C. Ionut, R. R. Choudhury and I. Rhee., "Towards mobile phone localization without war-driving," *IEEE Infocom,* pp. 1-9, 2010.

[55] L. Juong-Sik and B. Hoh, "Dynamic pricing incentive for participatory sensing," *Pervasive and Mobile Computing ,* vol. 6, no. 6, pp. 693-708, 2010.

[56] L. Tie, H.-P. Tan and L. Xia, "Profit-maximizing incentive for participatory sensing," *INFOCOM,* pp. 127-135, 2014 .

[57] L. Juong-Sik and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," *International Conference on Pervasive Computing and Communications (PerCom),* pp. 60-68, 2010.

[58] G. L. F. Cardone, P. Bellavista, A. Corradi, C. Borcea, M. Talasila and R. Curtmola, "Fostering participaction in smart cities: a geo-social crowdsensing platform," *Communications Magazine,* vol. 51, no. 6, pp. 112-112, 2013.

[59] E. Stephan, C. Schroth and J. Eberspächer, "Car-to-car communication," *In VDE-Kongress ,* 2006.

[60] Starner, "Human-powered wearable computing," *IBM systems Journal,* vol. 35, no. 3.4, pp. 618-629, 1996.

[61] G. Martin and A. Spriggs, "Assessing the impact of CCTV," *London: Home Office Research, Development and Statistics Directorate,* p. 2005.

[62] C. Chilipirea, A. Petre, C. Dobre, F. Pop and F. Xhafa, "Enabling Vehicular Data with Distributed Machine Learning," *Transactions on Computational Collective Intelligence,* vol. XIX, pp. 89-102, 2015.

[63] Y. Jihwang, D. Kotz and T. Henderson, "CRAWDAD: a community resource for archiving wireless data at Dartmouth," *ACM SIGCOMM Computer Communication Review,* vol. 36, no. 2 , pp. 21-22, 2006.

[64] N. Eagle and A. Pentland, "CRAWDAD dataset mit/reality (v. 2005-07-01)," *traceset: blueaware, downloaded from http://crawdad.org/mit/reality/20050701/blueaware, doi:10.15783/C71S31.*

[65] Y. Wang, P. Zhang, T. Liu, C. Sadler and M. Martonosi, "CRAWDAD dataset princeton/zebranet (v. 2007-02-14)," *traceset: movement, downloaded from http://crawdad.org/princeton/zebranet/20070214/movement, doi:10.15783/C77C78,.*

[66] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici and A. Rabuffi, "CRAWDAD dataset roma/taxi (v. 2014-07-17)," *downloaded from http://crawdad.org/roma/taxi/20140717, doi:10.15783/C7QC7M.*

[67] aestetix and C. Petro, "CRAWDAD dataset hope/amd (v. 2008-08-07)," *downloaded from http://crawdad.org/hope/amd/20080807, doi:10.15783/C7101B.*

[68] M. Piorkowski, N. Sarafijanovic-Djukic and M. Grossglauser, "CRAWDAD dataset epfl/mobility (v. 2009-02-24)," *downloaded from http://crawdad.org/epfl/mobility/20090224, doi:10.15783/C7J010.*

[69] P. Bernard, "Natural language processing," 2010.

[70] M. James, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[71] C. Jaime, R. Michalski and T. Mitchell, "An overview of machine learning," *Machine learning,* pp. 3-23, 1983.

[72] B. Michael, K. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis and Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics,* vol. 214, no. 1, pp. 481-518, 2012.