```
diff --git a/.gitignore b/.gitignore
index a2f7741..6e7439c 100644
--- a/.gitignore
+++ b/.gitignore


@@ -26,0 +27 @@ _minted-thesis/
+viva_report
```

```
diff --git a/README.md b/README.md
index c735c93..d002bb7 100644
--- a/README.md
+++ b/README.md


@@ -3,3 +2,0 @@
- 1. Final PDF of the thesis is at https://github.com/sbmkvp/phd-
thesis/blob/master/outputs/full-manuscript.pdf
- 2. Word count is at https://github.com/sbmkvp/phd-thesis/blob/master/word-
count.md
- 3. Writing plan is at https://github.com/sbmkvp/phd-thesis/blob/master/writing-
plan.md
```

```
diff --git a/chapters/00_front.tex b/chapters/00_front.tex
index 177151c..7733b49 100644
--- a/chapters/00_front.tex
+++ b/chapters/00_front.tex


@@ -34,0 +35,16 @@
+
+%-------------------------------------------------------------------------------
+%      Dedication
+%-------------------------------------------------------------------------------
+
+\cleardoublepage
+~\vfill
+\begin{doublespace}
+\noindent\fontsize{12}{12}\selectfont\itshape \nohyphenation
+\hspace{4cm}Dedicated to my parents,
+
+\hspace{3.5cm}S. Kalavathy and K. Soundararaj.
+\end{doublespace}
+\vfill
+\vfill
+


@@ -40 +56 @@
-\setcounter{tocdepth}{0}
+\setcounter{tocdepth}{1}
```

```
@@ -44,0 +61,53 @@
+%--------------------------------------------------------------------------
+\chapter*{Glossary} \label{front:glossary}
+%--------------------------------------------------------------------------
+
+\begin{itemize}
+  \item \textbf{Active vs Passive Collection} - Active collection is where the
data collection process involves the active participation from the study subjects.
In passive data collection, no such participation is required. For example, a web
form based survey is an active data collection process while a script collecting
data on user's data on a website is passive. This shall not be confused with
primary and secondary data where the difference is mainly due to who collects the
data from the source.
+
+  \item \textbf{Anonymisation and Pseudonymisation } - The act of removing
personal or identifiable information from the data. For example, removing the
names and date of birth of people in a dataset. Anonymisation could be carried out
in various ways two most popular approaches are generalisation and perturbation.
Pseudonymisation is similar but the personally identifiable data is substituted
with artificial identifier. The difference between them is that in anonymisation
the personal information is permanently purged and there is no way of getting the
information back. De-anonymisation is the reverse process of getting personal
identifiable data from anonymised data.
+
+  \item \textbf{Big Data} - Generally defined as data which could not be handled
with regularly used tools and techniques. There are more nuanced definitions of
big data depending on the dimension, context and industry. These concepts are
discussed in detail in Section \ref{section:toolkit}.
+
+  \item \textbf{Bluetooth} - Short wavelength, low energy, short range wireless
technology used for transferring data between electronic devices. They are
generally used by mobile devices to connect to peripherals.
+
+  \item \textbf{CDRC} - The Consumer Data Research Center is an investment from
Economic Social Research Center, UK for working with consumer-related
organisations to open up their data resources to trusted researchers thus enabling
them to carry out important social and economic research.
+
+  \item \textbf{Cellular/Mobile Network} - Terrestrial, long range and wireless
network which provides connectivity to mobile devices embedded in them. Most
commonly used to denote the networks that provides telephony and internet services
to mobile devices using radio waves.
+
+  \item \textbf{Cryptographic Hashing} - This is the process of transforming a
variable set of characters or contents of a file into a fixed length string
(checksum). The process is usually one way and is not reversible without a rainbow
attack using a database of checksums of all possible values. This is generally
used for storing user passwords and to verify the integrity/ authenticity of data.
In this thesis hashing functions are used for the purpose of anonymisaiton
+
+  \item \textbf{Data Partner} - Local Data Company - the organisation which
developed the Smart Street Sensor project in conjunction with CDRC. The
organisation is solely responsible for the design, manufacture, installation and
maintenance of the Smart Street Sensors.
+
+  \item \textbf{Device Fingerprinting} - This is the process of identifying
devices through forensic analysis. This commonly used to identify users from data
collected by operational websites. This commonly done through identifying unique
configurations of the devices such as screen sizes, versions of software
installed, etc. In this thesis fingerprinting is used to denote such processes
where unique devices are identified from randomised data.
+
+  \item \textbf{Encryption} - This is the process of converting a information
into unintelligible format using an algorithm (cipher) to prevent unauthorised
access. The process is two way since the resulting cipher-text can be decrypted to
restore information. The most common methods used is a public-key based encryption
scheme known as 'Diffie Hellman key exchange'.
```

```
+
+  \item \textbf{Footfall} - Number of people at a given location at a give time.
In this thesis the term is used synonymous to footfall at a high street - which is
only the pedestrians walking along the particular sidewalk of the high street and
does not include people on the carriage way in vehicles.
+
+  \item \textbf{High Street} - The primary business street in a town or a local
area where most shops and commercial activity are located. This term is often used
to contrast and distinguish from 'Shopping centers' which are large designated
areas earmarked for retail activity exclusively.
+
+  \item \textbf{Localisation} - Though localisation could mean both GSM
localisation and indoor positioning where the location of a device is indirectly
derived from other information, in this thesis localisation is used mainly in the
context of indoor positioning of mobile devices using techniques other than GPS.
The localisation of objects is often relative to each other or to an object with
an established position.
+
+  \item \textbf{Location, Sensor and Install} - In this thesis and the Smart
Street Sensor project, 'location' refers to a physical or geographic location in
United Kingdom, 'sensor' denotes the electronic equipment used to collect data and
'install' refers to the time when a particular sensor was operating from a
particular location.
+
+  \item \textbf{Mobile Device} - A portable computing device such as a smartphone
or tablet computer. This also includes wearables and other devices which have
computing hardware and can operate independently of another device.
+
+  \item \textbf{Personally Identifiable Information} - Any data that could
potentially identify a specific individual. Any information that can be used to
distinguish one person from another and can be used for de-anonymizing anonymous
data can be considered personally identifiable information. In this thesis, the
Media Access Control address is considered as sensitive personally identifiable
information.
+
+  \item \textbf{Positioning} - Positioning is the measurement of the absolute
position (coordinates) of an object with reference to the earth. This is usually
achieved through the use of Global Positioning System or similar systems.
+
+  \item \textbf{Probe Request} - This is a special signal broadcast by Wi-Fi
enabled mobile devices to elicit a response (probe response) from Wi-Fi access
points which can receive them. The primary purpose of the probe request is to
enumerate the available Wi-Fi networks and there can also be secondary purposes
such as indoor positioning.
+
+  \item \textbf{Real-time} - This is highly subjective concept which could be
defined as the phenomena which occurs sufficiently immediately. In this thesis
real-time is used to describe data which is collected within an hour after the
event has occurred.
+
+  \item \textbf{Sensor Configuration} - The way the sensor is installed at a
particular location. This includes the this includes the position of the sensor in
terms of height and depth, the material of shopfront etc.
+
+  \item \textbf{Signal, Noise} - 'Signal' is used to denote the data received
from devices which are within study area and 'Noise' to denote the data from
devices outside the study area.
+
+  \item \textbf{Smart Street Sensor Project} - This project is a comprehensive
study of live footfall patterns across Great Britain using 1,000 Wi-Fi based
sensors located in high streets across 81 towns and cities across the country.
Smart Street Sensor is a Raspberry Pi based sensor designed and manufactured by
the Data partner which collects data for the project.
+
+  \item \textbf{Wi-Fi} - A family of medium range radio technologies based on the
IEEE 802.11 family of standards which are generally used for wireless local area
networking between devices.
+
+\end{itemize}
```

```
+
```

```
@@ -65 +134 @@
-\vspace{0.2cm} \par \noindent \textbf{Exhibition} - Real-time footfall in
Bloomsbury. (2017). \textit{Big Data Here - Big Data Week 2016}, London, United
Kingdom.
+\vspace{0.2cm} \par \noindent \textbf{Exhibition} - Real-time footfall in
Bloomsbury. (2017). \textit{big data Here - big data Week 2016}, London, United
Kingdom.
```

```
@@ -75,14 +143,0 @@
-%--------------------------------------------------------------------------------
-%      Dedication
-%--------------------------------------------------------------------------------
-
-\cleardoublepage
-~\vfill
-\begin{doublespace}
-\noindent\fontsize{12}{12}\selectfont\itshape \nohyphenation
-\hspace{4cm}Dedicated to my parents,
-
-\hspace{3.5cm}S. Kalavathy and K. Soundararaj.
-\end{doublespace}
-\vfill
-\vfill
```

```
@@ -102 +157 @@ In attempts to collect, analyse and link these widely available
datasets at a ma
-This research looks at one such data source - Wi-Fi probe requests gen- erated by
mobile devices - in detail, and processes it into granular, long-term footfall
information of the retail high streets of the United Kingdom (UK).
+This research looks at one such data source - Wi-Fi probe requests generated by
mobile devices - in detail, and processes it into granular, long-term information
on number of people on the retail high streets of the United Kingdom (UK).
```

```
@@ -105,2 +160 @@ This research starts with the design and implementation of
multiple experiments
-The thesis also details the uses of these datasets, along with the massive
dataset generated by the \textit{`Smart Street Sensor'} project, to devise novel
data cleaning and processing methodologies.
-These methodologies resulted in the generation of a high quality dataset which
describes the volume of footfall on UK retail high streets with a granularity of 5
minute intervals since August 2015.
+The thesis also details the uses of these datasets, along with the massive
dataset generated by the \textit{`Smart Street Sensor'} project, to devise novel
data cleaning and processing methodologies which result in the generation of a
high quality dataset which describes the volume of people on UK retail high
streets with a granularity of 5 minute intervals since August 2015 across 1000
locations (approx.) in 115 towns.
```

```
@@ -116 +170 @@ We live in the age of data deluge where data are generated at a
pace that far ex
-Putting these vast amounts of data to use within the constraints of available
resources and time, is one of the biggest challenges faced by researchers today.
+Putting these amounts of data to use within the constraints of available
resources and time, is one of the biggest challenges faced by researchers today.
```

```
@@ -118 +172 @@ The primary impact of this research is in solving this issue.
-This research utilised one such vast dataset - Wi-Fi signals generated by
millions of mobile phones all around the year and available to anyone with a Wi-Fi
receiver - then cleaned and processed them into highly granular and longitudinal
information on the volume of footfall at retail high streets across the UK.
+This research utilised one such dataset - Wi-Fi signals generated by millions of
```

```
mobile phones all around the year and available to anyone with a Wi-Fi receiver -
then cleaned and processed them into highly granular and longitudinal information
on the volume of footfall at retail high streets across the UK.


@@ -128 +182 @@ The work on  this `Medium-data toolkit' was presented at the
conference \textit{
-Moreover the research directly led to the creation and maintenance of the
\textit{aggregated footfall} data product disseminated by CDRC\sidenote[][-2cm]
{\textit{\url{https://data.cdrc.ac.uk/dataset/local.data.company.ucl.smartstreetsens
 and has served as the data source for multiple research projects within and
outside CDRC and UCL.
+Moreover the research directly led to the creation and maintenance of the
\textit{aggregated footfall} data product disseminated by CDRC\sidenote[][-2cm]
{\textit{Local Data Company \& UCL Smart Street Sensor Footfall Data: Research
Aggregated data - \url{https://bit.ly/2FNGmo0}}}, and has served as the data
source for multiple research projects within and outside CDRC and UCL.


@@ -132 +186 @@ From this information comes a variety of insights: retailers can
derive insights
-Finally, in the past 3 years, the outputs from this research have been
disseminated to the broader academic community and industry through a series of
paper presentations at conferences such as \textit{GIS Research UK} and
\textit{Conference of Complex Systems}, talks at \textit{Data natives},
\textit{Geo+Data London}, and \textit{Smart Urban Policy Futures Workshop},
industry events such as \textit{Oxford Retail Futures Conference}, and public
engagement events such as the \textit{Big Data Here} exhibition.
+Finally, in the past 3 years, the outputs from this research have been
disseminated to the broader academic community and industry through a series of
paper presentations at conferences such as \textit{GIS Research UK} and
\textit{Conference of Complex Systems}, talks at \textit{Data natives},
\textit{Geo+Data London}, and \textit{Smart Urban Policy Futures Workshop},
industry events such as \textit{Oxford Retail Futures Conference}, and public
engagement events such as the \textit{big data Here} exhibition.


@@ -144 +198 @@ I would like to thank \textit{Economic and Social Research
Council} for funding
-Last but not the least, I would like to thank all the contributors to the
hundreds of open source and free projects I have used in this research. I am
forever indebted to them and hope to repay them by contributing back to the
community for the rest of my career.
+Last but not the least, I would like to thank the contributors to all the open
source and free projects I have used in this research (appendix
\ref{appendix:software}). I am forever indebted to them and hope to repay them by
contributing back to the community for the rest of my career.
```

```
diff --git a/chapters/10_introduction.tex b/chapters/10_introduction.tex
index 7e784a1..bb6766a 100644
--- a/chapters/10_introduction.tex
+++ b/chapters/10_introduction.tex


@@ -2 +2 @@
-\chapter{Introduction}
+\chapter{Introduction} \label{chapter:introduction}


@@ -9 +9 @@ The field was further broadened in the later part of the twentieth
```

```
century to in
-Moreover, with the dawn of the information age around the turn of the millennium,
the built environment can now be viewed as the tangible result of information
exchange; where cities can be seen as high density clusters of information
exchange, rather than as places with a concentration of physical infrastructure
such as buildings and roads.
+Moreover, with the dawn of the information age around the turn of the millennium,
the built environment can now be viewed as the tangible result of information
exchange; where cities can be seen as high density clusters of information
exchange, in addition to as places with a concentration of physical infrastructure
such as buildings and roads.


@@ -21 +21 @@ Most of the data collection activity has also become `passive', i.e.
collected w
-This has vastly increased the capacity of the data collection process, which has
led to the emergence of `Big Data' and consequentially, to the need for advanced
and automated data-mining techniques to extract value from these vast datasets.
+This has vastly increased the capacity of the data collection process, which has
led to the emergence of `big data' and consequentially, to the need for advanced
and automated data-mining techniques to extract value from these vast datasets.


@@ -24,0 +25,4 @@ For example, functional regions of a country has been derived
from call detail r
+%-------------------------------------------------------------------------------%
+\section{Challenges}
+%-------------------------------------------------------------------------------%
+


@@ -40,3 +44,3 @@ In addition to privacy concerns, this deluge of data introduces
significant tech
-Both academia and industry have produced extensive `Big Data' research which
develops the theory, methods and tools to tackle the challenges posed by such
large assemblages of data, in order to derive meaningful insights from them.
-This `Big Data' research promises to solve a lot of the technological and
logistical challenges incurred in many disciplines, but not without significant
additional overheads in terms of cost and resources.
-In the case of research projects, blindly jumping on the `Big Data' bandwagon
without consideration, has the potential to cause more problems than advantages.
+Both academia and industry have produced extensive `big data' research which
develops the theory, methods and tools to tackle the challenges posed by such
large assemblages of data, in order to derive meaningful insights from them.
+This `big data' research promises to solve a lot of the technological and
logistical challenges incurred in many disciplines, but not without significant
additional overheads in terms of cost and resources.
+In the case of research projects, blindly adopting the `big data' methods without
consideration, has the potential to cause more problems than advantages.


@@ -44 +48 @@ The discipline of Geography, especially geographic information
systems and scien
-Hence, it is extremely important to be mindful while adopting the contributions
from `Big Data' discourse for research so that the solutions are implemented where
the actual problems are located.
+Hence, it is extremely important to be mindful while adopting the contributions
from `big data' discourse for research so that the solutions are implemented where
the actual problems are located.


@@ -48 +52,27 @@ Moreover, there needs to be an inquiry into a framework for how
these considerat
-In this context, the research described in this thesis works on the opportunities
and pitfalls presented above in the following ways: by first describing the
collection of large volumes of passively generated data, then by solving the
uncertainties in the data which arises due to their high variability and the
mechanisms designed to protect the privacy of the users, and finally, there is an
analysis of the data to produce useful information regarding the distribution and
```

```
dynamics of footfall in the country.
+%-------------------------------------------------------------------------------%
+\section{Research Question \& Methodology}
+%-------------------------------------------------------------------------------%
+
+The motivation for the research began with the collection of the Wi-Fi probe
requests at a national level within the 'Smart Street Sensors' (SSS) project.
+The primary objective of the project was to develop a business venture providing
quantitative data on footfall to retailers across the country.
+This research was initially designed to supplement the above project by exploring
the possibility of validating the data collection methodology and understanding
the uncertainties and biases present in them.
+It was also designed to explore methods and analyses which provide insights and
value to the retailers from the base footfall volumes.
+As discussed in the previous section, the preliminary analysis of the data
collected revealed two major problems areas.
+First is the need to improve the accuracy of the footfall estimation by
circumventing the MAC randomisation process and the second is the need to protect
the privacy of user by developing methods that neither reverse engineer nor
collect any personal data.
+With these two challenges in mind, the primary question posed for the research is
as follows,
+
+\begin{displayquote}
+  "Can dynamics of footfall inferred from passively collected big dataset without
putting the privacy of users at risk?"
+\end{displayquote}
+
+In this context, this thesis works on answering the question while exploiting the
opportunities presented above in the following ways: by first describing the
collection of large volumes of passively generated data, then by solving the
uncertainties in the data which arises due to their high variability and the
mechanisms designed to protect the privacy of the users, and finally, by analysing
the data to produce useful information regarding the distribution and dynamics of
footfall in the country.
+
+Contrary to regular methodology, where the research starts from the question,
moves to a literature search, data collection, analysis and finally discussion and
conclusions, this research starts from the availability of large comprehensive
national level dataset.
+This research starts from this dataset, studies both the data and literature
surrounding it in detail, finds gaps, problems and unanswered questions in the
field, the tries to solve or answer them.
+In this pursuit of understanding the nature of the data that were available
without using personal data of users, the research also devises and conducts
series of controlled experiments which provides valuable insight into improving
the method that could be used for improving the estimation of people or footfall
around the sensors.
+The effectiveness of the methodology was also tested using various sets of
manually collected data on footfall information at sample locations.
+
+%-------------------------------------------------------------------------------%
+\section{Outline}
+%-------------------------------------------------------------------------------%
+


@@ -51,0 +82 @@ The literature review resulted in the identification of the best
possible data s
+


@@ -55,2 +86,3 @@ Three sets of initial experiments were conducted and results
from the experiment
-The chapter also introduces the `Smart Street Sensors'(SSS) project - a national
project which collects Wi-Fi data at a large number of retail locations.
-The chapter concludes with a detailed evaluation of all the data collected from
the experiments and the SSS project, in terms of the bias, noise and uncertainties
present in them.
```

```
+The chapter also introduces the `Smart Street Sensors' project - a national
project which collects Wi-Fi data at a large number of retail locations.
+The chapter concludes with a detailed evaluation of all the data collected from
the experiments and the Smart Street Sensor project, in terms of the bias, noise
and uncertainties present in them.
+


@@ -61,0 +94 @@ In section \ref{section:pipeline}, both the `data toolkit' and
methods are combi
+


@@ -63,0 +97,3 @@ Chapter \ref{chapter:applications} details a variety of
applications of the rese
+%-------------------------------------------------------------------------------%
+\section{Impacts \& Applications}
+%-------------------------------------------------------------------------------%


@@ -69,0 +106 @@ Availability of such datasets can revolutionise academic research
in fields such
+
```

```
diff --git a/chapters/20_literature.tex b/chapters/20_literature.tex
index 39817fc..cb501b1 100644
--- a/chapters/20_literature.tex
+++ b/chapters/20_literature.tex


@@ -10 +10 @@ Alternatively, smaller datasets such as sample surveys and traffic
counts are co
-This pursuit for identifying a data source which has the best features of both
type of datasets started as an inquiry into methods to estimate and interpolate
highly granular data from existing regional level aggregate data.
+This pursuit for identifying a data source which has the best features of both
type of datasets started as an inquiry into methods to estimate and interpolate
finer data from existing regional level aggregate data.


@@ -20 +20 @@ The aim of this survey was to evaluate the stage at which the
research is curren
-A comprehensive survey of over 300 publications which discuss this area of
research was undertaken covering the major themes and trends in the last 70 years.
+A comprehensive survey of over 300 publications which discuss this area of
research was undertaken covering the major themes and trends in the last 40 years.
```

```
diff --git a/chapters/21__research_themes.tex b/chapters/21__research_themes.tex
index 35d5e71..6e469f1 100644
--- a/chapters/21__research_themes.tex
+++ b/chapters/21__research_themes.tex
```

```
@@ -34 +34 @@ The focus of the research around this time was primarily on
interpolation from t
-There have been numerous fairly successful attempts with methodologies where a
broad dataset such  as regional level population summaries and modelling or
interpolating more granular data from them by augmenting with other sources of
data such as street networks \citep{reibel2005}, remote sensing \citep{sutton1997,
yuan1997, chen2002} etc.
+There have been numerous fairly successful attempts with methodologies where a
broad dataset such  as regional level population summaries and modelling or
interpolating finer data from them by augmenting with other sources of data such
as street networks \citep{reibel2005}, remote sensing \citep{sutton1997, yuan1997,
chen2002} etc.


@@ -42 +42 @@ There were efforts in using device level sensors such as global
positioning syst
-There have been studies on looking at people distribution as granular as queue
lengths as discussed by \citep{wang2013} to city level dynamic population mapping
where the limitations of traditional datasets generated through censuses and
surveys \cite{deville2014}.
+In terms of scale, there have been studies on looking at distribution of people
at a highly granular level such as queue lengths \cite[-5.5cm]{wang2013} as well
as broader level such as cities \cite{deville2014}.


@@ -46 +46 @@ Social media data such as twitter \citep{lansley2016a} and other
consumer data s
-Recently, with increased concerns and legislation on privacy, there have been
studies which go back to the effort of interpolating granular data from broader
datasets but using more data and processor intensive technologies such as agent
based modelling, deep learning, small area estimation \citep{crols2019,
shibata2019, rao2015} etc..
+Recently, with increased concerns and legislation on privacy, there have been
studies which go back to the effort of interpolating finer data from broader
datasets but using more data and processor intensive technologies such as agent
based modelling, deep learning, small area estimation \citep{crols2019,
shibata2019, rao2015} etc..


@@ -53 +53 @@ Though there have been a lot of work done in most of the directions
in this rese
-Study of movement of people is one of the major areas of research which have
significantly benefited from the decentralised collection of data at a granular
level \cite{castells2000}.
+Study of movement of people is one of the major areas of research which have
significantly benefited from the decentralised collection of data at a granular
level \cite[-1cm]{castells2000}.


@@ -94 +94 @@ This is currently an active field of research and there is immense
opportunity f
-Visualising the temporal dynamics of data collected on human activities through
decentralised processes poses significant challenges when approached with
traditional cartographic concepts \cite{maceachren2001, hallisey2005}.
+Visualising the temporal dynamics of data collected on human activities through
decentralised processes poses significant challenges when approached with
traditional cartographic concepts \cite[-4cm]{maceachren2001, hallisey2005}.


@@ -111 +111 @@ With the population getting more and more connected
\citep{castells2010}, the na
-This flood of hard data \cite{nature2008} was accompanied not only by optimism in
its potential \citep{thomas2001} but also by the questions raised on the
challenges in handling the diverse, large scale, non standardised data it produces
and the usefulness or representativeness of the resulting analysis
\citep{miller2010, arribas-bel2014a}.
+This flood of hard data \cite[-1cm]{nature2008} was accompanied not only by
```

```
optimism in its potential \citep{thomas2001} but also by the questions raised on
the challenges in handling the diverse, large scale, non standardised data it
produces and the usefulness or representativeness of the resulting analysis
\citep{miller2010, arribas-bel2014a}.


@@ -121 +121 @@ There is immense value in uniquely identifying and profiling
information on peop
-Strictly protecting personal information while ensuring the information is usable
for research by maintaining the uniqueness in the data is the major concern which
was addressed by devising frameworks for secure practices in confidentially
collecting and using the location data \citep{duckham2006, tang2006, lane2014}.
Some efforts sought to accomplish this task through cryptographic hashing
algorithms (Pang, 2007) while others aimed to thwart identification and tracking
at the device level by techniques such as MAC randomisation \citep{gruteser2005,
greenstein2008}.
+Strictly protecting personal information while ensuring the information is usable
for research by maintaining the uniqueness in the data is the major concern which
was addressed by devising frameworks for secure practices in confidentially
collecting and using the location data \citep{duckham2006, tang2006, lane2014}.
Some efforts sought to accomplish this task through cryptographic hashing
algorithms \citep{pang2007} while others aimed to thwart identification and
tracking at the device level by techniques such as MAC randomisation
\citep{gruteser2005, greenstein2008}.


@@ -132 +132 @@ There is opportunity in this area for research in applying the
cryptographic sol
-  \caption{Outline of the `Medium data toolkit' devised to collect, process,
visualise and manage the Wi-Fi probe requests data}
+  \caption{The evolution of research since 1980 categorised based on their major
theme.}


@@ -137 +137 @@ Figure \ref{figure:literature:themes:timeline} shows the volume of
research done
-We can observe that there are distinct trends in the research over time, which
evolved around the development of technology in the last two decades.
+We can observe that following distinct trends exist in the research, which
evolved around the development of technology in the last two decades.
```

```
diff --git a/chapters/22__technologies.tex b/chapters/22__technologies.tex
index e38a823..2a66af4 100644
--- a/chapters/22__technologies.tex
+++ b/chapters/22__technologies.tex


@@ -4,0 +5,8 @@
+\begin{marginfigure}
+  \forcerectofloat
+  \includegraphics[trim={1.1cm 1cm 1cm 1cm},clip]{images/literature-
technology.png}
+  \caption{Distribution of research across various techniques and technologies}
+  \label{figure:literature:timeline}
+\end{marginfigure}
+\marginnote{\noindent\fontsize{7}{7}\textit{Measured in the number of papers
published}}
+
```

```
@@ -6 +14 @@ When we look at the literature from the technology perspective, we
observe that
-Figure \ref{figure:literature:tech:timeline} shows the distribution of the
research in terms of the main technique/ technology used over the past 50 years.
+Figure \ref{figure:literature:timeline} shows the distribution of the research in
terms of the main technique/ technology used over the past 40 years.


@@ -13,6 +21,5 @@ A detailed account of these studies is given below,
-\begin{marginfigure}[1cm]
-  \includegraphics[trim={1.1cm 1cm 1cm 1cm},clip]{images/literature-
technology.png}
-  \caption{Distribution of research across various techniques and technologies}
-  \label{figure:literature:timeline}
-\end{marginfigure}
-\marginnote{\noindent\fontsize{7}{7}\textit{Measured in the number of papers
published}}
+\begin{figure*}
+  \includegraphics{images/literature-tech-timeline.png}
+  \caption{The evolution of research since 1980 in terms of the the technology
used in the research.}
+  \label{figure:literature:tech:timeline}
+\end{figure*}


@@ -43,7 +49,0 @@ This gives us the need to identify and use techniques which are
more general in
-\begin{figure*}
-  \includegraphics{images/literature-tech-timeline.png}
-  \caption{Outline of the `Medium data toolkit' devised to collect, process,
visualise and manage the Wi-Fi probe requests data}
-  \label{figure:literature:tech:timeline}
-\end{figure*}
-
-


@@ -54 +54 @@ This gives us the need to identify and use techniques which are more
general in
-The rise of mobile phones as ubiquitous personal devices for the broader
population has provided us with a viable alternative for collecting data with
greater granularity at large scales.
+The rise of mobile phones as ubiquitous personal devices for the broader
population has provided us with a viable alternative for collecting data with
finer granularity at large scales.


@@ -64,2 +64,2 @@ Even though this approach has been acknowledged to have inherent
biases such as
-Visual exploration of use of such data using interactive interfaces to evaluate
quality of service and scenario testing has been tested for the optimisation of
public transport \citep{sbodio2014}.
-Such network data with the active and passive information collected from them can
be used to create trajectories of people \cite{schlaich2010}, detect their daily
routine \citep{sevtsuk2010} and classify those routes in terms of function
\citep{becker2011a}.
+Visual exploration of use of such data using interactive interfaces to evaluate
quality of service and scenario testing has been tested for the optimisation of
public transport \cite[-4cm]{sbodio2014}.
+Such network data with the active and passive information collected from them can
be used to create trajectories of people \cite{schlaich2010}, detect their daily
routine \cite[2cm]{sevtsuk2010} and classify those routes in terms of function
\citep{becker2011a}.


@@ -73 +73 @@ This social networks and their spatio-temporal structure can also be
used for cl
-The data collected from the cellular network when examined at granular levels
such as inter-personal communication and economic activity can be used to create
```

```
estimations of micro area-level population density \citep{pulselli2008, ng2017}
and also the characteristics \citep{girardin2009} and the nature of the activity
\citep{phithakkitnukoon2010}.
+The data collected from the cellular network when examined at granular levels
such as inter-personal communication and economic activity can be used to create
estimations of micro area-level population density \cite[-4cm]{pulselli2008,
ng2017} and also the characteristics \cite[1cm]{girardin2009} and the nature of
the activity \citep{phithakkitnukoon2010}.


@@ -76 +76 @@ The spatial patterns thus uncovered can then be applied to urban
planning \citep
-For example, population influxes measured from changes in mobile network usage
can be used to model spread of diseases \citep{buckee2015}.
+For example, population influxes measured from changes in mobile network usage
can be used to model spread of diseases \cite{buckee2015}.


@@ -100 +100 @@ Apart from GPS and Wi-Fi, there have been efforts in exploring
other possibiliti
-There have also been successful implementations of frameworks to predict movement
of people by combining Wi-Fi and Bluetooth \citep{vu2011}.
+There have also been successful implementations of frameworks to predict movement
of people by combining Wi-Fi and Bluetooth \cite{vu2011}.


@@ -118 +118 @@ There have been successful efforts in extracting behaviours and
patterns from su
-In traffic and transportation, GPS trajectory from mobile devices is used to
estimate \citep{calabrese2011} and expand \citep{jing2011} OD matrices, detect the
mode of travel \citep{gong2012, rossi2015} and calibrate existing spatial
interaction models \citep{yue2012}
+In traffic and transportation, GPS trajectory from mobile devices is used to
estimate \citep{calabrese2011} and expand \citep{jing2011} origin-destination
matrices, detect the mode of travel \citep{gong2012, rossi2015} and calibrate
existing spatial interaction models \citep{yue2012}


@@ -123 +123 @@ Patterns \citep{jiang2012} and structures in routines
\citep{eagle2009} can be e
-It can also utilised in classification of the population at a particular location
at a given time \citep{pappalardo2015}.
+It can also utilised in classification of the population at a particular location
at a given time \cite[-1cm]{pappalardo2015}.


@@ -139 +139 @@ It is a distributed server-client based system where the client
connects to acce
-Every device in the network has a unique hardware specific MAC address, which is
transmitted between the device and AP before the connection is made.
+Every mobile device in the network has a unique hardware specific MAC address,
which is transmitted between the device and AP before the connection is made.


@@ -144 +144 @@ With this background we can see that Wi-Fi provides a fair middle
ground between
-Since the network infrastructure is distributed and deployed for Internet it
offers near complete coverage, is very resilient,  and can encapsulate and
reinforce civic space in cities \cite{torrens2008}.
+Since the network infrastructure is distributed and deployed for Internet it
offers more coverage than most of the technologies discussed except or cellular
network. It is also very resilient and can encapsulate and reinforce civic space
in cities \cite[-2cm]{torrens2008}.


@@ -146 +146 @@ Since the network infrastructure is distributed and deployed for
Internet it off
-Though Wi-Fi is a location less technology, there are reliable methods to
triangulate the location of the device by the signal strength and the known
```

```
locations of APs \citep{he2003, moore2004, lamarca2005, dinesh2017, lin2018}.
+Though Wi-Fi is a location less technology, there are reliable methods to
trilaterate the location of the device by the signal strength and the locations of
APs known through either targeted surveys or crowdsourced volunteer effort
\citep{he2003, moore2004, lamarca2005, dinesh2017, lin2018}.


@@ -148 +148 @@ This can overcome the usual shortcoming of GPS, which struggles
for precision an
-Utilising this, we can easily and quickly estimate trajectories of the mobile
devices just using the Wi-Fi communication the device has with multiple known APs
\citep{xu2013}.
+Utilising this, we can easily and quickly estimate trajectories of the mobile
devices just using the Wi-Fi communication the device has with multiple known APs
\cite[-4cm]{xu2013}.


@@ -152 +152 @@ It can also be used in transportation planning and management to
estimate travel
-Being a general network protocol designed to be used by mobile devices, Wi-Fi
devices relay a range of public signals known as probe request frames on regular
intervals throughout its operation, for the purpose of connecting and maintaining
a reliable and secure connection for the mobile device \citep{freud2015}.
+Being a general network protocol designed to be used by mobile devices, Wi-Fi
devices relay a range of public signals known as probe request frames on regular
intervals throughout its operation, for the purpose of connecting and maintaining
a reliable and secure connection for the mobile device \cite{freud2015}.


@@ -156,2 +156,2 @@ These device/user fingerprinting techniques are demonstrated
by \citep{franklin2
-Using the semantic information present in these probe requests it is possible to
understand the nature of these users at a large scale \citep{di2016}.
-Using the received signal strengths from pre placed devices we can monitor the
presence and movement of entities that are not even carrying a Wi-Fi enabled
device \citep{elgohary2013}.
+Using the semantic information present in these probe requests, such as names of
previously connected APs, it is possible to understand the nature of these users
at a large scale \citep{di2016}.
+Using the received signal strengths from pre placed devices we can monitor the
presence and movement of entities that are not even carrying a Wi-Fi enabled
device \cite{elgohary2013}.


@@ -162 +162 @@ Since MAC randomisation is not a perfect solution
\citep{mathieucunche2016} ther
-The methods used are decomposition of OUIs where detailed device model
information is estimated by analysing an already known dataset of OUIs
\citep{martin2016}; Scrambler attack where a small part of the physical layer
specification for Wi-Fi is used \citep{bloessl2015}; and finally, the timing
attack where the packet sequence information present in the probe request frame is
used \citep{matte2016, cheng2016}.
+The methods used are decomposition of OUIs where detailed device model
information is estimated by analysing an already known dataset of OUIs
\cite{martin2016}; Scrambler attack where a small part of the physical layer
specification for Wi-Fi is used \citep{bloessl2015}; and finally, the timing
attack where the packet sequence information present in the probe request frame is
used \citep{matte2016, cheng2016}.


@@ -167 +166,0 @@ Similar to GPS data this can be used as an additional control
layer for interpol
-A comparison of various approaches was done through experiments on a telecom
operator dataset, where it was showed that using network-driven mobile phone
location data is more advantageous compared to the widely used event-driven
ones\cite[-3cm]{pinelli2015}.
```

```
diff --git a/chapters/23__research_gaps.tex b/chapters/23__research_gaps.tex
index 3d7063d..98aa65c 100644
--- a/chapters/23__research_gaps.tex
+++ b/chapters/23__research_gaps.tex


@@ -13 +13 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Coverage & Local & City & All & Local & All\\
+        Coverage* & Local & City & All & Local & All\\


@@ -15 +15 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Certainty & Very Low & High & Medium & High & Medium \\
+        Certainty* & Very Low & High & Medium & High & Medium \\


@@ -19 +19 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Intrusiveness & Low & Medium & High & High & Medium \\
+        Intrusiveness* & Low & Medium & High & High & Medium \\


@@ -21 +21 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Granularity & Very Low & Very High & Medium & High & High \\
+        Granularity* & Very Low & Very High & Medium & High & High \\


@@ -23 +23 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Ease of Collection & Medium & Low & Medium & Low & High \\
+        Ease of Collection* & Medium & Low & Medium & Low & High \\


@@ -25,3 +25 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-        Scalability & Medium & Low & High & Medium & High \\
-        \addlinespace[0.2cm]
-        Privacy Risk & Low & Medium & High & High & Medium \\
+        Scalability* & Medium & Low & High & Medium & High \\


@@ -34 +32,3 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-\marginnote[-3.75cm]{\textit{*independence from secondary data collected by a
third party.}}
+\marginnote[-3.75cm]{\textit{* coverage - the density and extent of the current
infrastructure. Certainty - the lack of uncertainty in the data. Independence -How
much the technique depends on secondary data. Intrusiveness - the potential for
infringement of users' privacy. Granularity - the smallest spatial and temporal at
which data could be collected. Ease of Collection - how efficient it is collect
data in terms of time and resources. Scalability - the potential for the
technology to improve coverage.}}
+
+We can observer that Wi-Fi offers the best possible technology in terms of
flexibility and scalability for data collection through mobile devices at an
individual level while posing some risk to privacy of participants and involves
uncertainty regarding the field of measurement.
```

```
@@ -36 +36,10 @@ Table \ref{table:literature:technologies} summarises the above
discussion to eva
-We can observer that Wi-Fi offers the best possible technology in terms of
flexibility and scalability for data collection through mobile devices at an
individual level while posing some risk to privacy of participants and involves
uncertainty regarding the field of measurement.
+\citep{pinelli2015} looks at a comparison of various approaches of collecting and
analysing mobile phone location data.
+The research identifies two major approaches in collecting device location data -
Event-driven and Network-driven.
+The event-driven approach is centered around the mobile devices generating data
+through their day to day activities.
+The major sources of event-driven data are Call Detail Records(CDR) and internet
use.
+Network-driven approach is centered around the service provision infrastructure
such as cellphone towers, Wi-Fi base stations etc.
+The methods used to collect network-driven data are periodic update - where the
device sends an update stating the base station it is connected to, handover -
where the device information is recorded as they are moving between base stations
and location update - where the location of the device is recorded based on the
base stations it is connected to.
+The research used a set of anonymised mobile phone location data from a Belgian
telecom operator for the city of Mons from which various event-driven and network
driven scenarios were simulated.
+The authors compared these simulated scenarios for application-independent and
application-dependent cases such as spatial dispersal, classes of users, count
estimation and flow estimation to understand their relative advantages and
disadvantages.
+Through these comparisons it was shown that using network-driven mobile phone
location data is more advantageous compared to the widely used event-driven ones.


@@ -39 +48 @@ From the literature search we can summarise that there is a
considerable opportu
-The potential for research gaps broadly fall under the following areas,
+The potential for research gaps are discussed in detail in the following
sections.


@@ -47 +56 @@ The potential for research gaps broadly fall under the following
areas,
-Previous research in this area of study has been limited to either national/
regional level studies using centrally collected night time population data such
as censuses or to area level studies conducted with mobile devices based
technologies.
+Previous research in this area of study has been limited to either national/
regional level studies using centrally collected residential population data such
as censuses or to area level studies conducted with mobile devices based
technologies.


@@ -49 +58,4 @@ Though there were some efforts in collecting and using mobile
phone data at nati
-The explosion of consumer data both publicly available and privately held
presents previously unseen opportunity and also limited by the privacy concerns
that arise with them.
+
+For example, \citep{qin2013} demonstrate that it is possible to detect and
quantify human presence at locations using probe requests with a detection rate of
86\%.
+Along with the evaluation of the various algorithms for channel switching the
research also successfully classifies these detected human presence into distinct
activities in a non-intrusive way.
+Though this work predates both the MAC address randomisation and wide spread use
of mobile experienced these days, the explosion of consumer data both publicly
available and privately held presents previously unseen opportunity and also
limited by the privacy concerns that arise with them.
```

```
@@ -74,2 +86,5 @@ As we see later, even the estimation of ambient population is
limited by these d
-Though there are prior research in this area, most are conducted from security
perspective evaluating the robustness of the randomisation/obfuscation procedure.
-These research focus around de-anonymising the obfuscated data to recreate the
personal information from them thus demonstrating vulnerabilities and risks for
the users.
+
+\citep{vanhoef2016} presents several novel methods of abusing the features of the
Wi-Fi standards to track mobile devices even when the MAC addresses were
randomised. This research shows the possibility of using the information elements
present in the probe requests along with the sequence numbers to fingerprint the
mobile device which sent the request with an accuracy of the 50\% within a 20
minute interval with a possibility of improvement with known scrambler 'seeds' -
the randomisation factor used by popular commercial devices. Though this sounds
promising for short intervals, since this research, manufacturers have stopped
including non-mandatory information elements which can affect the accuracy
significantly. The research also features two other methods to reverse engineer
the original MAC addresses from the randomised ones - first where known hotspots
were spoofed to trick the mobile devices in revealing their real addresses and the
second where a different protocol requests were used. Both these methods cannot be
used extensively since the former is not ethically sound and the latter is not
widely used by all mobile devices.
+
+Since the above study and the following ones were conducted from security
perspective - evaluating the robustness of the randomisation/ obfuscation
procedure,
+they focus on de-anonymising the obfuscated data to recreate the personal
information from them while demonstrating vulnerabilities in the standard and
associated risks for the users.


@@ -93 +108,3 @@ Thought research have been done on this area using social media
data, a longitud
-The near real-time aspect of such research also provides us opportunities in
fields of disaster management, smart cities etc.
+\citep{kontokosta2016} discuss the use of Wi-Fi data for a 'real-time' census of
the city with a case study of New York City's Lower Manhattan neighborhood.
+The research collects around 20 million Wi-Fi data points during 2015 and
presents a model to create real-time, on-the-fly population estimates with fine
granularity.
+The research demonstrates the feasibility of the pursuit along with the potential
significance of such localised population estimates for use within the domains of
city operations and policy, strategic long-term planning processes, emergency
response etc.


@@ -111,0 +129,5 @@ Even when the problem of the identifying unique fingerprints
of users in the dat
+
+\citep{musa2011} use the Wi-Fi probe requests collected in a 12-hour trial on a
busy road to describe a passive tracking system for mobile devices.
+The research proposes a trajectory estimation method based on Viterbi's algorithm
which estimates the most-likely spatial path taken from the information on when
and where they have been detected.
+Although the research extends this trial into a 9-month deployment and
demonstrates trajectory estimates with high accuracy, the problem still remains
where we need to extract trajectories of users without actually being able to
identify them.
+


@@ -129 +151 @@ Though there are extensive research in using ambient population
and people's mov
-With access to granular and long-term data sets, we can hope to look into the how
the places have changed over time and how the external factors such as policy and
economy has affected them.
+With access to highly granular and long-term data sets, we can hope to look into
the how the places have changed over time and how the external factors such as
```

```
policy and economy has affected them.


@@ -138 +160,7 @@ There are opportunities to ask questions such as,
-\marginnote{\textit{\textbf{Opportunity 5:} Using long term data to detect the
nature and change of form and function of a place. } }
+\marginnote[-3cm]{\textit{\textbf{Opportunity 5:} Using long term data to detect
the nature and change of form and function of a place. } }
+
+
+
+
+
+
```

diff --git a/chapters/30_data_collection.tex b/chapters/30_data_collection.tex
index f89e871..5127127 100644
--- a/chapters/30_data_collection.tex
+++ b/chapters/30_data_collection.tex

```
@@ -5 +5 @@
-From the literature review in Chapter \ref{chapter:literature}, we observed that
of all the technologies discussed, Wi-Fi seems to be the most promising one.
+From the literature review in Chapter \ref{chapter:literature}, we observed that
of all the technologies discussed, Wi-Fi seems to be the most promising one for
our purposes.


@@ -25 +25 @@ We also introduce the `Smart Street Sensor' project - a national
scale effort fo
-Finally we summarise the data collection procedure with a detailed look at the
each set of uncertainty in these datasets and draw conclusions for further lines
of research into alleviating the uncertainty and noise so that the datasets can be
used to estimate human activity with confidence.
+Finally we summarise the data collection procedure with a detailed look at the
uncertainties in these datasets and draw conclusions for further lines of research
into alleviating the uncertainty and noise so that the datasets can be used to
estimate human activity with confidence.


@@ -39 +39 @@ Though the end goal of internet connectivity is the same, Wi-Fi
greatly differs
-The first difference is the range of the network: unlike mobile infrastructure
where a single tower can serve mobile phones for miles, Wi-Fi is designed to be
and extension of the wired networking, thus creating short range network with a
range of 20 meters.
+The first difference is the range of the network: unlike mobile infrastructure
where a single tower can serve mobile phones for miles, Wi-Fi is designed to be an
extension of the wired networking, thus creating short range network with a range
of 20 meters.


@@ -52 +52 @@ There have been numerous iterations and versions of the IEEE 802.11
standards bu
-These frames have the information that is being exchanged along with the meta
data about information and information on the device that is sending them.
+These frames have the information that is being exchanged along with the meta
data and information on the device that is sending them.
```

```
@@ -57,0 +58,7 @@ The structure of a probe request is shown in Figure
\ref{figure:collection:probe
+\begin{figure*}
+  \includegraphics[width=0.9\textwidth,trim={0 -30 0 -10},clip]{images/probe-
request-structure.png}
+  \caption{Structure of a probe request frame. }
+  \label{figure:collection:proberequest}
+\end{figure*}
+\marginnote[0.75cm]{\textit{Source: IEEE 802.11 specification}. DA - Destination
Address. SA - Sender Address. BSSID - Broadcast or multicast address. FH -
Frequency hopping. OS - Optional. CF - Contention free. DS - Direct Sequnce. }
+


@@ -60 +67 @@ There are additional meta data such as the sequence number of the
frame, and con
-There is also a dearth of variable information which can be used to transfer
data.
+There are also a number of variable information which can be used to transfer
data.


@@ -64,0 +72,4 @@ The significant information present in a probe request is
detailed in Table \ref
+Essentially the above information is sent over and over by the mobile device
which expects a reply from nearby APs so that it can keep a list of networks it
can connect to.
+This process is usually carried out even when the Wi-Fi is switched off in the
operating system so that the connection times are faster once it is switched on.
+Moreover operating systems use the replies they get for these probe requests and
triangulate the device location with respect to the APs with location information
on AP's collected through surveys or crowdsourcing, thus acting as a quick and
easy localisation solution which along with the above makes this probing process
almost non-stop.
+


@@ -74 +84,0 @@ The significant information present in a probe request is detailed
in Table \ref
-        \addlinespace[0.2cm]


@@ -76,3 +86 @@ The significant information present in a probe request is detailed
in Table \ref
-        \addlinespace[0.2cm]
-        Signal Strength (RSSI) & The strength of the received signal\\
-        \addlinespace[0.2cm]
+        Received Signal Strength (dBm) & The strength of the received signal\\


@@ -80 +87,0 @@ The significant information present in a probe request is detailed
in Table \ref
-        \addlinespace[0.2cm]


@@ -82 +88,0 @@ The significant information present in a probe request is detailed
in Table \ref
-        \addlinespace[0.2cm]


@@ -84,3 +90 @@ The significant information present in a probe request is detailed
in Table \ref
-        \addlinespace[0.2cm]
-        Known SSIDs & Networks that are already known to the device\\
-        \addlinespace[0.1cm]
+        Known Networks & Name of networks that are already known to the device\\
```

```
@@ -94,10 +97,0 @@ The significant information present in a probe request is
detailed in Table \ref
-Essentially the above information is sent over and over by the mobile device
which expects a reply from nearby APs so that it can keep a list of networks it
can connect to.
-This process is usually carried out even when the Wi-Fi is switched off in the
operating system so that the connection times are faster once it is switched on.
-Moreover operating systems use the replies they get for these probe requests and
triangulate the device location with respect to the APs with known location
information thus acting as a quick and easy localisation solution which along with
the above makes this probing process almost non-stop.
-
-\begin{figure*}
-  \includegraphics[width=0.9\textwidth,trim={0 -30 0 -10},clip]{images/probe-
request-structure.png}
-  \caption{Structure of a probe request frame. }
-  \label{figure:collection:proberequest}
-\end{figure*}
-\marginnote[0.75cm]{\textit{Source: IEEE 802.11 specification}}


@@ -108 +102 @@ Moreover operating systems use the replies they get for these
probe requests and
-Media Access Control (MAC) address is a 12bit unique identifier assigned to a
device on a network.
+Media Access Control (MAC) address is a 6 byte unique identifier assigned to a
device on a network.


@@ -111 +105 @@ Although the IP address of a mobile device might change regularly,
the MAC addre
-The MAC address has two parts: the first 6bits are known as the Organisationally
Unique Identifier (OUI) and gives us information about the manufacturer of the
network card.
+The MAC address has two parts: the first 2 bytes are known as the
Organisationally Unique Identifier (OUI) and gives us information about the
manufacturer of the network card.


@@ -113 +107 @@ Organisations need to register with IEEE to be assigned an OUI
which they can us
-the second 6bits are unique to device itself.
+the second 2 bytes are unique to device itself.
```

```
diff --git a/chapters/31__initial_experiments.tex
b/chapters/31__initial_experiments.tex
index 9047a8f..5155b9d 100644
--- a/chapters/31__initial_experiments.tex
+++ b/chapters/31__initial_experiments.tex
```

```
@@ -26 +26 @@ First, we put the Wi-Fi module of the laptop in `Monitor mode' where
it behaves
-Then we invoke the command line interface of the Wireshark programme tshark to
collect the Wi-Fi probe requests received by the laptop in Character Separated
File (CSV) format.
+Then we invoke the command line interface of the Wireshark programme tshark to
collect the Wi-Fi probe requests received by the laptop in Character Separated
File (CSV) format in the file system.
```

```
@@ -62 +62 @@ The fields marked with \textit{-e} are the ones which were collected
and they co
-The name of the manufacturer/ vendor of the Wi-Fi module  is extracted from the
\textit{wlan.sa\_resolved} field into a separate column and the original field is
hashed using the SHA256 algorithm implemented in R.
+The name of the manufacturer/ vendor of the Wi-Fi module  is extracted from the
\textit{wlan.sa\_resolved} field into a separate column and the original field is
hashed using the SHA256 algorithm \cite{gueron2011} implemented in R.


@@ -103 +103 @@ This is where the signal strength shows good promise.
-Looking at the Table \ref{table:collection:proberequests} we can see that two of
these vendors show significantly low average signal strength - Google and Fn-
LinkT, which can easily correspond to the two devices present in the room.
+Looking at the Table \ref{table:collection:proberequests} we can see that two of
these vendors show significantly high average signal strength - Google and Fn-
LinkT, which can easily correspond to the two devices present in the room.


@@ -262 +262 @@ Figure \ref{figure:collection:ucl:signal} also shows the results
of k-means clus
-\marginnote[-1.75cm]{\textit{Class intervals calculated using k-means clustering
with the number of clusters defined as 4.} }
+\marginnote[-1.75cm]{\textit{x-axis is measure in dBm as a proxy for distance.
The class intervals calculated using k-means clustering with the number of
clusters defined as 4.} }
```

diff --git a/chapters/32__pilot_study.tex b/chapters/32__pilot_study.tex
index 444d076..53839e6 100644
--- a/chapters/32__pilot_study.tex
+++ b/chapters/32__pilot_study.tex

```
@@ -50 +50 @@ The information collected from each probe request at these locations
are,
-The manual counting at these locations were done using a custom application
\citet{bala2018}. The application was built for recording pedestrian footfall with
precision and accuracy which was not possible when counted manually. The app
records the precise time stamp of every footfall with the precision of micro
seconds which can be aggregated later at different time intervals. The code for
the app is detailed in Section \ref{appendix:clicker}.
+The manual counting at these locations were done using a custom application
\citet{bala2018}. The application was built for recording pedestrian footfall with
precision and accuracy which was not possible when counted without the
application. The app records the precise time stamp of every footfall with the
precision of micro seconds which can be aggregated later at different time
intervals. The code for the app is detailed in Section \ref{appendix:clicker}.


@@ -100 +100 @@ Id & Location & Type & Installation notes &
Probes\textsuperscript{*} & Footfall
-\marginnote[-1.5cm]{\textit{* Total probe requests in \(\times10^{6}\)(per
minute)  ** Total footfall (per minute)}}
+\marginnote[-1.5cm]{\textit{* Total probe requests in \(\times10^{6}\)(per
minute)  ** Total footfall counted manually (per minute)}}
```

```
diff --git a/chapters/33__smart_street_sensor.tex
b/chapters/33__smart_street_sensor.tex
index 8c967ce..614d52f 100644
--- a/chapters/33__smart_street_sensor.tex
+++ b/chapters/33__smart_street_sensor.tex


@@ -63 +63 @@ We must also note that the locations are retail and any insight from
the data ne
-A custom footfall counting technology using Wi-Fi sensors (Figure
\ref{figure:collection:sss:hardware}) was also designed and developed by LDC, and
the sensors were installed the identified locations.
+A custom footfall counting technology using Wi-Fi sensors (Figure
\ref{figure:collection:sss:hardware}) was also designed and developed by LDC, and
the sensors were installed at the identified locations.


@@ -73 +73 @@ There are also a small percentage (3\%) of the devices which are
installed withi
-Each device collects data independently and uploads the collected data to a
central Azure container at regular intervals of 5 minutes through a dedicated 3G
mobile data connection.
+Each device collects data independently and uploads the collected data to a
central microsoft cloud facility (Azure) container at regular intervals of 5
minutes through a dedicated 3G mobile data connection.


@@ -84 +84 @@ We have collected around 2TB of data comprising of around 73 billion
probe reque
-Due to the scale and the commercial nature of the project, the sensors collect
fewer data than the previous experiments.
+Due to the scale and the commercial nature of the project, the sensors collect
fewer data per probe than the previous experiments.









diff --git a/chapters/34__discussion.tex b/chapters/34__discussion.tex
index 973f0ec..b6fc31a 100644
--- a/chapters/34__discussion.tex
+++ b/chapters/34__discussion.tex


@@ -7 +7 @@ The pilot study extended this further by collecting data for a longer
time in re
-The Smart Street Sensor project is the most comprehensive study which collects
very small focussed set of data at a national level for very long periods of time.
+The Smart Street Sensor project is the most comprehensive study which collects
very small focussed set of data in a probe requestat a national level for very
long periods of time.


@@ -127 +127 @@ Though we can simply aggregate these probe requests based on the
unique informat
-We to consider this uncertainty in detail while making any simple assumptions on
the relationship between number of probe requests and the number of mobile devices
that generated them.
```

```
+We need to consider this uncertainty in detail while making any simple
assumptions on the relationship between number of probe requests and the number of
mobile devices that generated them.
```

```
diff --git a/chapters/40_data_processing.tex b/chapters/40_data_processing.tex
index 77048bf..c995971 100644
--- a/chapters/40_data_processing.tex
+++ b/chapters/40_data_processing.tex

@@ -12,2 +12,2 @@ A brief review was conducted on the topic of 'big data and big
data tools' which
-Using this framework, the datasets were evaluated in each of their 5 dimensions
to understand their nature and the challenges posed in these dimensions.
-We find that the Wi-Fi based datasets are 'medium data' which can benefit from
customised toolkits which increases the efficiency.
+Using this framework, the datasets were evaluated in each of their dimensions to
understand their nature and the challenges posed in these dimensions.
+We find that the Wi-Fi based datasets are 'medium data' (as opposed to 'big
data') which can benefit from customised toolkits which increases the efficiency.
```

```
diff --git a/chapters/41__data_toolkit.tex b/chapters/41__data_toolkit.tex
index bc1084a..f7f7a9c 100644
--- a/chapters/41__data_toolkit.tex
+++ b/chapters/41__data_toolkit.tex

@@ -20 +20 @@ Though we discussed the methods conceptually, we left out the
rationale behind c
-We start by discussing the concept of `Big Data' in general and look at previous
literature to understand its definition, nature and the challenges they pose.
+We start by discussing the concept of `big data' in general and look at previous
literature to understand its definition, nature and the challenges they pose.

@@ -29 +29 @@ Finally we put together a toolkit to suit our datasets built from
simple small U
-\subsection{What is `Big Data'?}
+\subsection{What is `big data'?}

@@ -52,3 +52,3 @@ This approach of looking at 'Big data' is too subjective and
falls short of givi
-One of the most quoted definitions pertains to the scale of the data in the
dimension of volume - size of the data, velocity - speed of the data and variety
- the complexity of the data \cite{laney2001}.
-This has also been extended to include more dimensions such as, veracity - the
reliability or truthfulness of the data, visualisation - the complexity in visual
interpretation and presentation of the data, and others such as visibility
validity, variability, volatility and value.
-There have also been other alternative dimensions proposed such as Cardinality,
```

```
continuity and complexity \cite{suthaharan2014}.
+One of the most quoted definitions pertains to the scale of the data in the
dimension of volume - size of the data, velocity - speed of the data and variety
- the complexity of the data \cite[-4cm]{laney2001}.
+This has also been extended to include more dimensions such as, veracity - the
reliability or truthfulness of the data, visualisation \cite[-2.5cm]{li2016}- the
complexity in visual interpretation and presentation of the data, and others such
as visibility validity, variability, volatility and value.
+There have also been other alternative dimensions proposed such as cardinality,
continuity and complexity \cite{suthaharan2014}.


@@ -62 +62 @@ Challenges in each one of these processing activity arises due to
the data being
-The data being big in volume, velocity and variety poses challenges in data
acquisition, aggregation, cleaning and analysis \cite{li2016}.
+The data being big in volume, velocity and variety poses challenges in data
acquisition, aggregation, cleaning and analysis.


@@ -127 +127,2 @@ It is important to note that this can be reduced many folds by
using compression
-      Study & Maximum* & Minimum* & Average* & Total** \\
+      Study & Maximum & Minimum & Average & Total* \\
+            & (GB per year) & (GB per year) & (GB per year) & (TB per year) \\


@@ -139 +140 @@ It is important to note that this can be reduced many folds by
using compression
-\marginnote{\textit{* Measured/ Estimated for each location in gigabytes per
year. ** Measured/ Estimated for 920 locations in terabytes per year} }
+\marginnote{\textit{* Estimated for 920 locations} }


@@ -184 +185,2 @@ Table \ref{table:toolkit:velocity} compares the datasets we
collected on Wi-Fi p
-      Study & Maximum* & Minimum* & Average* & Total** \\
+      Study & Maximum & Minimum & Average & Total* \\
+            & (per 5min) & (per 5min) & (per 5min) & (Mn per 5min)\\


@@ -196 +198 @@ Table \ref{table:toolkit:velocity} compares the datasets we
collected on Wi-Fi p
-\marginnote{\textit{* Measured/ Estimated for each location in number of requests
per 5 minutes. ** Measured/ Estimated for 920 locations in Millions of requests
per 5 minutes} }
+\marginnote{\textit{* Estimated for 920 locations} }


@@ -218 +220 @@ The first set of variety present in the Wi-Fi probes data set
arises from the 'i
-The structure of a probe request is discussed in detail in the data collection
chapter and is summarised in Figure ?.
+The structure of a probe request is discussed in detail in the data collection
chapter and is summarised in Figure \ref{figure:collection:proberequest}.


@@ -250 +252 @@ Since the sensors are part of retail establishments they can be
switched on and
-Figure \ref{figure:toolkit:veracity:gaps} demonstrates the veracity of the data
in terms of missing data for a sample of locations in London.
+Figure \ref{figure:toolkit:veracity:gaps} demonstrates the veracity of the data
in terms of missing data for a sample of 5 locations in London.


@@ -257 +259 @@ It is often not clear if the change in the data is due to actual
changes at the
-For example, Opening of a mobile shop next door to the sensor can increase the
estimated footfall without any change in actual footfall at the location.
```

```
+For example, opening of a mobile shop next door to the sensor can increase the
estimated footfall without any change in actual footfall at the location.


@@ -359 +361 @@ Designing and implementing of such system is time consuming,
cumbersome and usua
-On the other end of this spectrum, we have end-to-end solutions such as Blix,
Walkbase, Ecuclid, Retail Next, Pygmalios, etc. where the data is collected
through multiple sensors and sources and syndicated into a clean footfall
information by a third party service provider.
+On the other end of this spectrum, we have end-to-end solutions such as Blix,
Walkbase, Euclid, Retail Next, Pygmalios, etc. where the data is collected through
multiple sensors and sources and syndicated into a clean footfall information by a
third party service provider.


@@ -399 +401 @@ The relational databases are optimised to deal with relational
data and usually
-In general they can handle large number of rows and are designed to scale
vertically.
+In general they can handle large number of rows and are designed to scale
vertically through addition of more resources to the DBMS such as CPU and Memory.


@@ -402,2 +404,2 @@ The document based databases are optimised to deal with
unstructured data and ca
-In general they can handle large number of columns and are designed to be
distributed and scaled horizontally.
-Being distributed, most document based databases try to pick a focus and
compromise on others as specified in CAP theorem \sidenote[][-1cm]{Brewer's
theorem or CAP theorem states that it is impossible to simultaneously guarantee
consistency, availability and partition tolerance in a distributed data store.}.
+In general they can handle large number of columns and are designed to be
distributed and scaled horizontally by adding more instances of the databases
which balance the load and redundancy between them.
+Being distributed, most document based databases try to pick a focus and
compromise on others as specified in CAP theorem \sidenote[][-2cm]{Brewer's
theorem or CAP theorem states that it is impossible to simultaneously guarantee
consistency, availability and partition tolerance in a distributed data store.}.


@@ -405,0 +408,11 @@ There a numerous databases systems which prioritise different
things and the rig
+Since the publication of the paper on `Google file system' by Google \cite[-2cm]
{sanjay2003}.
+There have been significant effort in designing and building 'big data' file
storage systems which can large data in the range of petabytes.
+These systems are designed to be distributed and optimised for high throughput
for queries on them.
+Hadoop Distributed File System (HDFS) is one such file system which is also the
most widely adopted.
+There are numerous cloud based, third-party solutions built with these file
systems making them easy to use.
+There are also numerous tools, libraries and frameworks which emulate the
features of database systems on these distributed file systems making them easier
to use further.
+The primary advantage of these systems is the sheer scalability they provide when
it comes to data volume.
+The primary disadvantage is the associated overheads in terms of cost and time
incurred in learning, designing and implementing them.
+Unless the project is sufficiently large, the advantages gained usually do not
justify the overheads introduced.
+Table \ref{table:toolkit:storage} summarises the above discussion along with
relevant examples.
+


@@ -449,10 +461,0 @@ There a numerous databases systems which prioritise different
things and the rig
```

```
-Since the publication of the paper on `Google file system' by Google \cite[-2cm]
{sanjay2003}.
-There have been significant effort in designing and building 'big data' file
storage systems which can large data in the range of petabytes.
-These systems are designed to be distributed and optimised for high throughput
for queries on them.
-Hadoop Distributed File System (HDFS) is one such file system which is also the
most widely adopted.
-There are numerous cloud based, third-party solutions built with these file
systems making them easy to use.
-There are also numerous tools, libraries and frameworks which emulate the
features of database systems on these distributed file systems making them easier
to use further.
-The primary advantage of these systems is the sheer scalability they provide when
it comes to data volume.
-The primary disadvantage is the associated overheads in terms of cost and time
incurred in learning, designing and implementing them.
-Unless the project is sufficiently large, the advantages gained usually do not
justify the overheads introduced.
-Table \ref{table:toolkit:storage} summarises the above discussion along with
relevant examples.


@@ -471 +474 @@ We chose PostgreSQL because of the PostGIS extension which gives
us flexibility
-\begin{marginfigure}
+\begin{marginfigure}[-5cm]


@@ -481 +484 @@ Hence we require tools which are capable of dealing with the
veracity of the dat
-The traditional approach to deal with such dataset is to load it into a general
purpose analysis tool such as R or a GIS packages and process it.
+The traditional approach to deal with such dataset is to load it into a general
purpose analysis tool such as R or a GIS package and process it.


@@ -492 +495 @@ On the other end we have big data analysis tools which are built
for dealing wit
-Since the publication of the paper on MapReduce, there have been immense
developments in the Big data analysis landscape.
+Since the publication of the paper on MapReduce \cite{dean2008}, there have been
immense developments in the Big data analysis landscape.


@@ -522 +525 @@ This is where we come across the possibility of using standard
Unix tools along
-In some cases, a data processing pipeline made using command line Unix tools have
been demonstrated to be 230 times faster than using big data toolkits
\cite{adam2014}.
+In some cases, a data processing pipeline made using command line Unix tools have
been demonstrated to be 230 times faster than using big data toolkits
\cite[-1.5cm]{adam2014}.


@@ -525,10 +528 @@ They are developed in line with the Unix philosophy which
focuses on modular and
-The core tenants of the Unix philosophy has been summarised by Doug McIlroy as
below,\cite{mcilroy1978},
-
-\begin{enumerate}[rightmargin=1cm]
-  \setlength{\itemindent}{2em}
-  \itemsep-0.25em
-       \item Make each program do one thing well. To do a new job, build afresh
rather than complicate old programs by adding new "features".
-       \item Expect the output of every program to become the input to another,
as yet unknown, program. Don't clutter output with extraneous information. Avoid
stringently columnar or binary input formats. Don't insist on interactive input.
-       \item Design and build software, even operating systems, to be tried
```

```
early, ideally within weeks. Don't hesitate to throw away the clumsy parts and
rebuild them.
-       \item Use tools in preference to unskilled help to lighten a programming
task, even if you have to detour to build the tools and expect to throw some of
them out after you've finished using them.
-\end{enumerate}
+The core tenants of the Unix philosophy has been summarised by Doug McIlroy as
below,\cite[-2cm]{mcilroy1978},


@@ -563,0 +558,10 @@ The core tenants of the Unix philosophy has been summarised
by Doug McIlroy as b
+
+\begin{enumerate}[rightmargin=1cm]
+  \setlength{\itemindent}{2em}
+  \itemsep-0.25em
+       \item Make each program do one thing well. To do a new job, build afresh
rather than complicate old programs by adding new "features".
+       \item Expect the output of every program to become the input to another,
as yet unknown, program. Don't clutter output with extraneous information. Avoid
stringently columnar or binary input formats. Don't insist on interactive input.
+       \item Design and build software, even operating systems, to be tried
early, ideally within weeks. Don't hesitate to throw away the clumsy parts and
rebuild them.
+       \item Use tools in preference to unskilled help to lighten a programming
task, even if you have to detour to build the tools and expect to throw some of
them out after you've finished using them.
+\end{enumerate}
+


@@ -586 +590 @@ Finally the modular structure also gives us the advantage of using
the best tool
-All of the gives us an extremely minimal and efficient toolkit to process the raw
Wi-Fi probes data into counts in a scalable way.
+All of this gives us an extremely minimal and efficient toolkit to process the
raw Wi-Fi probes data into counts in a scalable way.


@@ -625,0 +630,4 @@ Every dimension of big data poses unique set of challenges and
we need make righ
+We then closely examined the Wi-Fi probes data we collected with this framework
and found that the data, though posed significant challenges with traditional data
processing techniques, do not exhibit `big data' properties in all its dimensions.
+Only veracity of the data was found to have any meaningful big data properties,
while volume and velocity was found to be 'medium' at best.
+The datasets lacked any variety and posed minimal challenge in the visualisation
dimension because of it high temporal granularity.
+Thus we arrived at the requirements for a bespoke `medium data toolkit' which is
able to deal with these challenges.


@@ -632,5 +639,0 @@ Every dimension of big data poses unique set of challenges and
we need make righ
-We then closely examined the Wi-Fi probes data we collected with this framework
and found that the data, though posed significant challenges with traditional data
processing techniques, do not exhibit `big data' properties in all its dimensions.
-Only veracity of the data was found to have any meaningful big data properties,
while volume and velocity was found to be 'medium' at best.
-The datasets lacked any variety and posed minimal challenge in the visualisation
dimension because of it high temporal granularity.
-Thus we arrived at the requirements for a bespoke `medium data toolkit' which is
able to deal with these challenges.
-


@@ -641 +644 @@ To process the raw data we chose to devise a processing pipeline
using an assort
-We also demonstrated that this processing pipeline can be 400 times faster than
```

```
the using a monolithic pipeline even with a small sample of locations.
+We also demonstrated that this processing pipeline can be 400 times faster (20
minutes to 3seconds) than the using a monolithic pipeline even with a small sample
of locations.
```

```
diff --git a/chapters/42__data_cleaning.tex b/chapters/42__data_cleaning.tex
index 1d45d3a..e3e3a51 100644
--- a/chapters/42__data_cleaning.tex
+++ b/chapters/42__data_cleaning.tex


@@ -12,6 +12,6 @@ The next step was to explore the extent of the noise generated
by looking at bot
-\begin{marginfigure}[-7cm]
-  \includegraphics[trim={10 10 10 10},clip]{images/processing-error-signal.png}
-  \caption{The share of noise present in the data from outside the field of
measurement in the initial experiment.}
-  \label{figure:processing:error:signal}
-\end{marginfigure}
-
+% \begin{marginfigure}[-7cm]
+%   \includegraphics[trim={10 10 10 10},clip]{images/processing-error-signal.png}
+%   \caption{The share of noise present in the data from outside the field of
measurement in the initial experiment.}
+%   \label{figure:processing:error:signal}
+% \end{marginfigure}
+%


@@ -28 +28 @@ It was observed that about 53\% of the total probe requests
collected were from
-As such the errors could be enormously reduced by simply filtering out the noise
as shown in Figure \ref{figure:processing:error:signal}.
+As such the errors could be enormously reduced by simply filtering out the noise
%as shown in Figure \ref{figure:processing:error:signal}.


@@ -83 +83 @@ That is, in configurations where a specific source of background
noise was at a
-For example, imagine a sensor in the middle of room such as in the initial
experiment in this thesis, with devices in and outside the room.
+For example, imagine a sensor in the middle of a room such as in the initial
experiment in this thesis, with devices in and outside the room.


@@ -87 +87 @@ Both of these changes can be identified by the ‘breaks’ in the
distribution
-Identification of these breaks in the data can be carried out using traditional
one-dimensional clustering algorithms such as ‘jenks natural breaks’, ‘k-means’,
‘quantile’ and ‘hierarchical clustering’, etc.
+Identification of these breaks in the data should be carried out using
traditional one-dimensional clustering algorithms such as ‘jenks natural breaks’,
‘k-means’, ‘quantile’ and ‘hierarchical clustering’, etc.


@@ -173 +173 @@ On a project with a large scope, such as the Smart Street Sensor
project, since
-The three methods — signal strength filtering, sequence number clustering, and
manual calibration - together provide a unified methodology for converting the Wi-
Fi probe requests into footfall.
```

```
+The three methods — signal strength filtering, sequence number clustering, and
manual calibration - together provide a complete methodology for converting the
Wi-Fi probe requests into footfall.


@@ -204,6 +204,6 @@ It also demonstrated the need for filtering the data before
aggregating them int
-     Quantile                                          & 0.002         &  27 \%
\\
-     K-Means                                           & 0.007         & -23 \%
\\
-     Hierarchical Clustering   & 172.520       &  -9 \% \\
-     Bagged Clustering                             & 0.135         & -30 \% \\
-     Fisher                                            & 3.034         & -30 \%
\\
-     Jenks Natural Break               & 556.279       & -30 \% \\
+     Quantile                                          & 0.002         & 27 \% \\
+     K-Means                                           & 0.007         & 23 \% \\
+     Hierarchical Clustering   & 172.520       &  9 \% \\
+     Bagged Clustering                             & 0.135         & 30 \% \\
+     Fisher                                            & 3.034         & 30 \% \\
+     Jenks Natural Break               & 556.279       & 30 \% \\


@@ -223,2 +223,2 @@ The results are shown in Table
\ref{table:processing:oxst:classification}.
-It was found that out of all algorithms, hierarchical clustering and jenks
natural break provided the least amount of errors.
-However, these two algorithms were designed to identify class intervals in much
smaller datasets and were extremely resource intensive for practical use with a
larger dataset.
+It was found that out of all algorithms, hierarchical clustering provided the
least amount of errors.
+However, this and jenks natural break were designed to identify class intervals
in much smaller datasets and were extremely resource intensive for practical use
with a larger dataset.


@@ -234 +234 @@ These concerns needed to be addressed with a larger survey with
multiple locatio
-  \caption{Finding the optimum time threshold $\alpha$ and sequence threshold
$\alpha$ through trial and error.}
+  \caption{Finding the optimum time threshold (s) $\alpha$ and sequence threshold
$\beta$ through trial and error.}


@@ -263,0 +264,3 @@ A comparison of minute by minute counts resulting from
different filtering proce
+To summarise, the data from the initial experiments suggest that both filtering
using signal strength and the clustering using sequence numbers worked well on
complex, real world data and resulted in fairly accurate pedestrian counts with a
MAPE of 20\%.
+It was also found that ‘k-means’ and ‘quantile’ are the best algorithms for
clustering signal strengths, and the optimum thresholds for time and sequence
numbers for the clustering algorithm were around 16 and 60 respectively.
+


@@ -269,3 +271,0 @@ A comparison of minute by minute counts resulting from
different filtering proce
-
-To summarise, the data from the initial experiments suggest that both filtering
using signal strength and the clustering using sequence numbers worked well on
complex, real world data and resulted in fairly accurate pedestrian counts with a
MAPE of 20\%.
-It was also found that ‘k-means’ and ‘quantile’ are the best algorithms for
clustering signal strengths, and the optimum thresholds for time and sequence
numbers for the clustering algorithm were around 16 and 60 respectively.
```

```
diff --git a/chapters/43__data_architecture.tex
b/chapters/43__data_architecture.tex
index b36a07c..9b150d5 100644
--- a/chapters/43__data_architecture.tex
+++ b/chapters/43__data_architecture.tex


@@ -42 +42 @@ First the probe requests were separated into global and local based
on the OUI p
-The global probe requests were then aggregated into ambient population estimation
employing the following steps,
+The non-randomised, global probe requests were then aggregated into ambient
population estimation employing the following steps,
```

```
diff --git a/chapters/50_applications.tex b/chapters/50_applications.tex
index 8a3a958..e810685 100644
--- a/chapters/50_applications.tex
+++ b/chapters/50_applications.tex


@@ -3,0 +4,2 @@
+
+\vspace{-0.5cm}


@@ -11 +13 @@ Finally the chapter briefly describes a way to calculate the flow of
pedestrians
-\section{United Kingdom Footfall Index}
+\section{Footfall Indices}


@@ -12,0 +15,2 @@ Finally the chapter briefly describes a way to calculate the
flow of pedestrians
+\subsection{UK Footfall Index}
+


@@ -29 +33 @@ In addition to showing larger trends, this footfall index also
showed sudden sho
-One such as example was the storm in February 2018 which corresponded with some
of the lowest footfall experience across the UK.
+One such as example was the storm in February 2018 named - "Beast from the East",
which resulted in a week of unusually low footfall experience across the UK.


@@ -41,3 +45 @@ It can be observed that April 2019 has been slower than last year
in most towns
-This kind of granular insights into trends in footfall could be valuable for
local authorities who can measure and monitor the health of their retail areas
closely.
-
```

```
-The difference in even smaller intra-day patterns in cities could be derived from
footfall data which could show the nature of their economies
+This kind of granular insights into trends in footfall could be valuable for
local authorities who can measure and monitor the health of their retail areas
closely. The difference in even smaller intra-day patterns in cities could be
derived from footfall data which could show the nature of their economies


@@ -46,0 +49 @@ Figure \ref{figure:applications:cities:profiles} shows an average
daily footfall
+


@@ -52 +55 @@ It can be observed that all three locations have completely
different patterns o
-Leicester Square was mostly a night time destination where the footfall peaks
around evening while Regent street is a mostly office location with three distinct
peaks corresponding to morning commute, evening commute and lunch.
+Leicester Square was mostly a evening destination where the footfall peaks around
evening while Regent street is a mostly office location with three distinct peaks
corresponding to morning commute, evening commute and lunch.


@@ -121 +124 @@ This coincided with the cold wave in UK named ‘Beast from the
East’, which b
-The other identifiable events are bank holiday weekends which result in higher
than normal footfall, and the holiday shopping season when footfall is at its
highest.
+The other identifiable events are bank holiday weekends which result in higher
than normal footfall, and the Christmas shopping season when footfall is at its
highest.


@@ -141 +144 @@ This not only shows the usefulness of the data in understanding
the effect event
-  \includegraphics[trim={0 0 0 0},clip]{images/applications-football-sample.png}
+  \includegraphics[trim={0 24 0 70},clip]{images/applications-football-
sample.png}


@@ -157,2 +160,2 @@ This research specifically removes any personally identifiable
information becau
-However, this problem can be solved by examining the movement of people in the
Smart Street Sensors network at a fine spatial and temporal resolution using a
novel methodology in the field of Big Data which uses mathematical models from
information theory: Transfer Entropy (TE).
-Using an area in central London, this section serves as a case study to
demonstrate the usefulness of TE as a measure of the flow of
pedestrians\sidenote{Work undertaken was in collaboration with Roberto Murcio and
Karlo Lugomer.
+However, this problem can be solved by examining the movement of people in the
Smart Street Sensors network at a fine spatial and temporal resolution using a
novel methodology in the field of big data which uses mathematical models from
information theory: Transfer Entropy (TE).
+Using an area in central London, this section serves as a case study to
demonstrate the usefulness of TE as a measure of the flow of
pedestrians\sidenote[][-3cm]{Work undertaken was in collaboration with Roberto
Murcio and Karlo Lugomer.


@@ -160,0 +164,7 @@ The methodology was formulated by Murcio; this author worked
on the implementati
+\begin{figure*}
+  \includegraphics[trim={0 0 0 0},clip]{images/applications-transfer-entropy.png}
+  \caption{Illustration of transfer entropies between set of locations along
Edgware Road, London.}
+  \label{figure:applications:transent}
+\end{figure*}
+
```

```
+


@@ -177,6 +186,0 @@ This measure is applied directly to our people's movement
problem and $X$ = loca
-\begin{figure*}
-  \includegraphics[trim={0 0 0 0},clip]{images/applications-transfer-entropy.png}
-  \caption{Illustration of transfer entropies between set of locations along
Edgware Road, London.}
-  \label{figure:applications:transent}
-\end{figure*}
-


@@ -187,2 +191,2 @@ The numbers at each line represent the $TE$ measured between
each pair of sensor
-The largest $TE$ value found was between 117 and 115.
-The asymmetry of the TE is clear here, as the value in the opposite direction
(115 to 117) is considerably lower.
+The largest $TE$ values found were from 114,115 and 117 to 116.
+The asymmetry of the TE is clear here, as the values in the opposite direction (
from 116 to 114,115 and 117) are considerably lower.
```

```
diff --git a/chapters/60_conclusions.tex b/chapters/60_conclusions.tex
index 5915aaa..42dd29e 100644
--- a/chapters/60_conclusions.tex
+++ b/chapters/60_conclusions.tex


@@ -10,0 +11,4 @@ This thesis identifies Wi-Fi probe requests as a source of data
from which infor
+%-------------------------------------------------------------------------------%
+\section{Summary of Findings}
+%-------------------------------------------------------------------------------%
+


@@ -39,0 +44 @@ Using all these methods for filtering, cleaning and adjusting Wi-
Fi data, this t
+Finally this research also provides a gallery of examples showing the possible
use of such granular and continuous data on footfall on a national level.


@@ -41,2 +46,7 @@ Using all these methods for filtering, cleaning and adjusting
Wi-Fi data, this t
-Finally this thesis provides a gallery of examples showing the possible use of
such granular and continuous data on footfall on a national level.
-The data were used to devise a 'footfall index' at various levels - national,
city, area and micro site locations showing how the retail related footfall have
been distributed in the UK and how this distribution has been changing over time
in high granularity both spatially and temporally.
+%-------------------------------------------------------------------------------%
+\section{Research Question}
+%-------------------------------------------------------------------------------%
+Looking back at our research question - "Can dynamics of footfall inferred from
passively collected big dataset without putting the privacy of users at risk?",
+we have demonstrated that the task is indeed feasible, using Wi-Fi probe
requests. Even when the identity of the devices were masked using randomisation
techniques we have demonstrated that aggregation and estimation could be done
without compromising the privacy of the users.
```

```
+In addition to this, we have also demonstrated the usefulness and application of
such footfall estimate with various examples.
+The footfall estimates derived from the method were used to devise a 'footfall
index' at various levels - national, city, area and micro site locations showing
how the retail related footfall have been distributed in the UK and how this
distribution has been changing over time in high granularity both spatially and
temporally.


@@ -46,0 +57,20 @@ Finally, it was also demonstrated that such detailed and
continuous footfall vol
+%--------------------------------------------------------------------------%
+\section{Further Work}
+%--------------------------------------------------------------------------%
+
+As we discussed in the literature review, the research on collecting and using
data on population distribution and dynamics have closely followed the advances
and changes in the consumer technology.
+Every new technology adopted for mainstream use spurred new wave of research in
using those technology.
+It is also noted that every new technology not only brought many advantages over
the previous ones but also introduced unique challenges.
+In this context, the larges opportunity in furthering the research exists in
identifying, evaluating and adopting new technologies.
+There is a significant opportunities in applying these new technologies for old
challenges and device methods to make them suitable to answer the questions raised
by research.
+Few such technologies are detailed below,
+
+\begin{itemize}
+  \item \textbf{5G} is the new generation of technology which aims to bring even
higher speeds of data transfer to mobile devices through cellular networks. This
may lead to the gradual decline and phasing out of Wi-Fi technology. Though this
cellular based technology doesn't provide the similar detail and flexibility
offered by Wi-Fi it has the potential to offer much more comprehensive picture of
the world if it gets  widely adopted.
+  \item \textbf{Bluetooth Low energy (BLE)} is the upcoming short-range, wireless
personal area network technology. With emphasis on being the technology used by
the Internet of Things (IOT) devices, this technology has the potential to
displace Wi-Fi as the choice of short-range communications. The explosion of
wearables and smart devices at home, the amount of data that could be available
form this technology could be staggering in the next decade.
+  \item \textbf{Ultra wide band radar} is another short-range technology which
has been developed for motion and object detection. Being primarily used to design
sensors for proximity and motion detection, this has the potential to become a
standard for vehicles. Moreover, with the recent uptick in self driving car
research and development, the cost of these devices has gone significantly down
thus providing amazing opportunities in creating comprehensive sensor networks
similar to Smart Street Sensor project.
+\end{itemize}
+
+In spite of being developed since 1980s, machine learning techniques have
received extraordinary interest in the last decade.
+This interest, along with advancements in the Big data tools and technologies has
set up the stage for research by applying supervised and unsupervised machine
learning techniques on large scale datasets collected through the above mentioned
technologies.
+There is a significant opportunity for applying unsupervised learning techniques
such as anomaly detection and neural networks in passively collected digital data
to improve data cleaning, interpolation, population estimations and time series
based predictions etc.


@@ -47,0 +78,5 @@ Finally, it was also demonstrated that such detailed and
continuous footfall vol
+Research ethics, safety and privacy are going to be the next big areas of concern
for advanced machine learning based techniques and big data analysis in the next
decade.
+The era of uninhibited large scale production, collection and consumption of
```

```
personal data through connected devices over internet without oversight is almost
over.
+People are increasingly concerned with protecting their privacy and are opposed
to the exploitation of their personal data.
+This concern has been addressed by legislation such as GDPR and technologies such
as cryptography and randomisation.
+All these developments provide us with various opportunities in further research.


@@ -48,0 +84,7 @@ Finally, it was also demonstrated that such detailed and
continuous footfall vol
+Firstly there is opportunity study the above mentioned technologies form a
privacy point of view to evaluate the advantages and risks presented by them and
advance the research in terms of both mitigation the risks while maintaining some
kind of usefulness.
+These inquiries can not only be done in terms of techniques but also on the lines
of legal compliance of such techniques.
+There is also opportunity for researching on the uncertainties and limits of
datasets when subject to robust privacy control methods.
+Secondly the immense research, innovation and advancements made in peer to peer
technologies in solving the various trust problems could be applied in the field
of sensor based population estimation or pedestrian flow detection.
+There is an opportunity for research into building a peer to peer network of
sensors where the data collected by the sensors never leave the device themselves
but the analyses are taken to the source of data.
+This act of "moving the analysis to data" can solve numerous problems of safety
of the personal data since there is not central point of failure and it can also
scale up indefinitely without overwhelming a central repository of data.
+Through these further research, we could take the field forward by not only
following the improvements in the technology of data collection but also push the
envelope in terms of developing more ethical and sage research environment while
handling large amounts of data.
```

```
diff --git a/chapters/70_appendix.tex b/chapters/70_appendix.tex
index 6249136..14e695b 100644
--- a/chapters/70_appendix.tex
+++ b/chapters/70_appendix.tex


@@ -2,0 +3,4 @@
+\vspace*{\fill}
+
+\textit{Note : All code used in this thesis has been made available online via
github - \url{https://github.com/sbmkvp/phd-thesis}.}
+


@@ -137 +140,0 @@ This bash script lists all the files in a given folder, parses
them into JSON da
-


@@ -144,0 +148,5 @@ This is a sample probe request captured using tshark and saved
in the JSON forma
+\pagebreak
+
+%=============================================================================%
+\section{Open-source Software Used}\label{appendix:software}
+%=============================================================================%
```

```
@@ -145,0 +154,83 @@ This is a sample probe request captured using tshark and
saved in the JSON forma
+This section provides a non-exhaustive list of the key open-source/free software
that have been used in this research.
+
+\begin{itemize}
+\item \textbf{R} - programming language for statistical computing
+    \begin{itemize}
+      \item \textbf{tidyverse} - An opinionated collection of R packages designed
for data science.
+      \item \textbf{imputeTS} - Package for imputation missing values in
univariate time series.
+      \item \textbf{tmap} - A flexible, layer-based, and easy to use package to
create thematic maps.
+      \item \textbf{lubridate} - Package for working with date-times and time-
spans
+      \item \textbf{ggplot2} - A system for declaratively creating graphics,
based on The Grammar of Graphics.
+      \item \textbf{classInt} -  Package for choosing univariate class intervals
for mapping or other graphics purposes.
+      \item \textbf{Cairo} - 2D graphics library with support for multiple output
devices.
+      \item \textbf{fmsb} - Package with methods and functions for demographic
analysis.
+      \item \textbf{digest} - Package for the creation of hash digests of
arbitrary R objects.
+      \item \textbf{ggrepel} - Package that provides geoms for ggplot2 to repel
overlapping text labels.
+      \item \textbf{ggridges} - Package for creating ridge plots.
+      \item \textbf{maptools} - Package for manipulating geographic data.
+      \item \textbf{tidyquant} - Package that brings financial analysis and
charting to tidyverse.
+      \item \textbf{treemapify} - Package for creating tree-maps.
+      \item \textbf{spatial features} - Package for manipulating geographic data
within tidyverse.
+      \item \textbf{RJSONIO} - Package for manipulating JSON objects.
+      \item \textbf{rgdal} - Wrapper for the Geospatial Data Abstraction Library.
+      \item \textbf{rgeos} - Wrapper for the Geometry Engine - Open Source.
+      \item \textbf{viridis} - Package providing pretty color scales for
visualisations.
+      \item \textbf{xtable} - Package for coercing data to LaTeX and HTML tables.
+      \item \textbf{scales} - Package for providing graphical scales mapping data
to aesthetics.
+      \item \textbf{showtext} - Package for managing fonts.
+      \item \textbf{reshape2} - Package for transforming data between long and
wide format.
+      \item \textbf{rmarkdown} - Package for integrating markdown with R
assisting reproducible research.
+    \end{itemize}
+  \item \textbf{Python} - An interpreted, high-level, general-purpose programming
language.
+  \item \textbf{PHP} - A general-purpose programming language originally designed
for server-side web development.
+  \item \textbf{JavaScript} - A high-level, interpreted scripting language for
client-side web development.
+    \begin{itemize}
+      \item \textbf{node.js} - JavaScript based runtime built on chrome's V8
engine.
+      \item \textbf{socket.io} - web sockets implementation for real-time,
bidirectional and event-based communication.
+      \item \textbf{moment.js} - JavaScript library for dealing with date-time
and time-spans.
+      \item \textbf{pm2} - Process management library for working with node.js
applications.
+      \item \textbf{express} - A fast, unopinionated, minimalist web framework
for Node.js
+      \item \textbf{Data Driven Documents} - A JavaScript library for visualizing
data with HTML, SVG, and CSS.
```

```
+      \item \textbf{jQuery} - A JavaScript library designed to simplify HTML DOM
manipulation.
+      \item \textbf{Bootstrap} - A framework for building responsive, mobile
first websites.
+      \item \textbf{highcharts} - A JavaScript library for drawing interactive
charts from data.
+    \end{itemize}
+  \item \textbf{GNU/Linux} - An operating system and an extensive collection of
open source and free computer software.
+    \begin{itemize}
+      \item \textbf{Arch Linux} - A lightweight and flexible Linux distribution.
+      \item \textbf{CentOS} - A community supported computing platform compatible
with Red Hat Enterprise Linux.
+      \item \textbf{Debian} - A Linux distribution focussing on stability.
+      \item \textbf{Ubuntu} - A Debian based Linux distribution focussing on ease
of use.
+      \item \textbf{Alpine Linux} - Ultra minimalistic Linux distribution
focussing on resource efficiency.
+    \end{itemize}
+  \item \textbf{git} - A simple distributed version control system.
+  \item \textbf{imagemagik} - Suite for displaying, converting and editing
images.
+  \item \textbf{ffmpeg} - Suite for converting and editing video files.
+  \item \textbf{fzf} - A general-purpose command-line fuzzy finder.
+  \item \textbf{ripgrep} - Rust based grep implementation for searching the
content of files.
+  \item \textbf{MySQL} - A relational database management system focussing on
speed and ease of use.
+  \item \textbf{Postgres} - A relational database management system emphasizing
extensibility and technical standards compliance.
+  \item \textbf{PostGIS} - Extension providing spatial objects for the PostgreSQL
database.
+  \item \textbf{QGIS} - Geographic Information System for creating, editing,
visualising, analysing and publishing geospatial information.
+  \item \textbf{gdal} - Geographic Data abstraction library.
+  \item \textbf{geos} - Geometry Engine Open Source.
+  \item \textbf{igraph} - R and Python library for dealing with networks /
Graphs.
+  \item \textbf{OpenStreetMap} - A collaborative project to create a free
editable map of the world.
+  \item \textbf{Leaflet} - A JavaScript library for mobile-friendly interactive
maps.
+  \item \textbf{Android } - Open source mobile operating system based on Linux.
+  \item \textbf{vim} - Vim is a highly configurable, modal text editor built with
focus on efficient.
+  \item \textbf{Latex} - A high quality professional typesetting system.
+  \item \textbf{jq} - A command line based JSON processor.
+  \item \textbf{Apache} - A feature rich web server.
+  \item \textbf{nginx} - A asynchronous, event-driven web server focussing on
resource efficiency.
+  \item \textbf{OpenJDK} - An open source implementation of the Java Platform.
+  \item \textbf{Wireshark} - A free and open-source packet analyzer
+  \item \textbf{OpenSSH} - A connectivity tool for remote login with the SSH
protocol.
+  \item \textbf{OpenSSL} - A full-featured toolkit for the Transport Layer
Security and Secure Sockets Layer protocols.
+  \item \textbf{GNUPG} - A complete and free implementation of the OpenPGP
standard.
+  \item \textbf{gnu-parallel} - A shell tool for executing jobs in parallel using
one or more computers.
+  \item \textbf{Libreoffice} - A free and open-source office suite built by The
Document Foundation.
+  \item \textbf{RaspberryPi} - A series of low-cost, flexible single-board
computers.
+  \item \textbf{Docker} - A platform for doing OS level virtualisation for
delivering software.
+  \item \textbf{termux} - A terminal emulator and Linux environment for Android.
+\end{itemize}
```

```
@@ -148 +239 @@ This is a sample probe request captured using tshark and saved in
the JSON forma
-\chapter{Research Article}
+% \chapter{Research Article}


@@ -150 +241 @@ This is a sample probe request captured using tshark and saved in
the JSON forma
-\includepdf[pages=1-15]{documents/ijgis-paper.pdf}
+% \includepdf[pages=1-15]{documents/ijgis-paper.pdf}
```

```
diff --git a/reading/abstracts.txt b/reading/abstracts.txt
new file mode 100644
index 0000000..db953b8
--- /dev/null
+++ b/reading/abstracts.txt


@@ -0,0 +1,36 @@
+Vanhoef
+
+We present several novel techniques to track (unassociated) mobile devices by
abusing features of the Wi-Fi standard. This shows that using random MAC
addresses, on its own, does not guarantee privacy.
+First, we show that information elements in probe requests can be used to
fingerprint devices. We then combine these fingerprints with incremental sequence
numbers, to create a tracking algorithm that does not rely on unique identi- fiers
such as MAC addresses. Based on real-world datasets, we demonstrate that our
algorithm can correctly track as much as 50% of devices for at least 20 minutes.
We also show that commodity Wi-Fi devices use predictable scram- bler seeds. These
can be used to improve the performance of our tracking algorithm. Finally, we
present two attacks that reveal the real MAC address of a device, even if MAC ad-
dress randomization is used. In the first one, we create fake hotspots to induce
clients to connect using their real MAC address. The second technique relies on
the new 802.11u standard, commonly referred to as Hotspot 2.0, where we show that
Linux and Windows send Access Network Query Protocol (ANQP) requests using their
real MAC address.
+
+
+Matte
+MAC address randomization is a common privacy protec- tion measure deployed in
major operating systems today. It is used to prevent user-tracking with probe
requests that are transmitted during IEEE 802.11 network scans. We present an
attack to defeat MAC address randomization through observation of the timings of
the network scans with an off- the-shelf Wi-Fi interface. This attack relies on a
signature based on inter-frame arrival times of probe requests, which is used to
group together frames coming from the same de- vice although they use distinct MAC
addresses. We propose several distance metrics based on timing and use them to-
gether with an incremental learning algorithm in order to group frames. We show
that these signatures are consistent over time and can be used as a pseudo-
identifier to track devices. Our framework is able to correctly group frames using
different MAC addresses but belonging to the same device in up to 75% of the
cases. These results show that the timing of 802.11 probe frames can be abused to
track individual devices and that address randomization alone is not always enough
to protect users against tracking.
+
+Martin
+Media Access Control (MAC) address ran- domization is a privacy technique whereby
mobile de- vices rotate through random hardware addresses in or- der to prevent
```

observers from singling out their traffic or physical location from other nearby devices. Adop- tion of this technology, however, has been sporadic and varied across device manufacturers. In this paper, we present the first wide-scale study of MAC address ran- domization in the wild, including a detailed breakdown of different randomization techniques by operating sys- tem, manufacturer, and model of device.
+We then identify multiple flaws in these implementa- tions which can be exploited to defeat randomization as performed by existing devices. First, we show that devices commonly make improper use of randomization by sending wireless frames with the true, global address when they should be using a randomized address. We move on to extend the passive identification techniques of Vanhoef et al. to effectively defeat randomization in ~96% of Android phones. Finally, we identify a previ- ously unknown flaw in the way wireless chipsets handle low-level control frames which applies to 100% of de- vices we tested. This flaw permits an active attack that can be used under certain circumstances to track any existing wireless device.
+
+Cunche
+Wi-Fi has imposed itself as one of the key radio technology in portable devices such as smartphones, tablets, and other wearable devices. Because they periodically scan for nearby access points, Wi-Fi devices act as portable radio beacons emitting short messages, called probe requests. The latter contain a unique identifier: the MAC address of the Wi-Fi interface. It can be used to passively track individuals. Owners of such devices are thus exposed to passive tracking in the physical world.
+MAC address randomization has been proposed as a way to prevent passive tracking, and is being progressively adopted by the industry. However, the MAC address is not the only piece of information that can be used for tracking. For instance, it has been recently shown that exploiting the content of frames as well as their timing could still lead to tracking despite MAC address randomization.
+
+
+
+Xu
+This paper presents a new simple mobile tracking system based on IEEE802.11 wireless signal detection, which can be used for analyzing the movement of pedestrian traffic. Wi-Fi packets emitted by Wi-Fi enabled smartphones are received at a monitoring station and these packets contain date, time, MAC address, and other information. The packets are received at a number of stations, distributed throughout the monitoring zone, which can measure the received signal strength. Based on the location of stations and data collected at the stations, the movement of pedestrian traffic can be analyzed. This information can be used to improve the services, such as better bus schedule time and better pavement design. In addition, this paper presents a signal strength based localisation method.
+
+
+Musa
+Smartphones with Wi-Fi enabled periodically transmit Wi-Fi messages, even when not associated to a network. In one 12-hour trial on a busy road (average daily traffic count 37,000 according to the state DOT), 7,000 unique devices were detected by a single road-side monitoring station, or about 1 device for every 5 vehicles.
+In this paper, we describe a system for passively tracking unmodified smartphones, based on such Wi-Fi detections. This system uses only common, off-the-shelf access point hardware to both collect and deliver detections. Thus, in ad- dition to high detection rates, it potentially offers very low equipment and installation cost.
+However, the long range and sparse nature of our oppor- tunistically collected Wi-Fi transmissions presents a signif- icant localization challenge. We propose a trajectory es- timation method based on Viterbi’s algorithm which takes second-by-second detections of a moving device as input, and produces the most likely spatio-temporal path taken. In addition, we present several methods that prompt passing de- vices to send additional messages, increasing detection rates an use signal-strength for improved accuracy.
+Based on our experimental evaluation from one 9-month deployment and several single-day deployments, passive Wi- Fi tracking detects a large fraction of passing smartphones, and produces high-accuracy trajectory estimates.
+

```
+Kontokosta
+New streams of data are being generated by a range of in-situ instrumentation,
mobile sensing, and social media that can be integrated and analyzed to better
understand urban activity and mobility patterns. While several studies have
focused on understanding flows of people throughout a city, these data can also be
used to create a more spatially and temporally granular picture of local
population, and to forecast localized population given some exogenous
environmental or physical conditions. Effectively modeling population dynamics at
high spatial and temporal resolutions would have significant implications for city
operations and policy, strategic long-term planning processes, emergency response
and management, and public health.
+This paper develops a real-time census of the city using Wi-Fi data to explore
urban phenology as a function of localized population dynamics. Using Wi-Fi probe
and connection data accounting for more than 20,000,000 data points for the year
2015 from New York City's Lower Manhattan neighborhood — combined with correlative
data from the U.S. Census American Community Survey, the Longitudinal Employer-
Household Dynamics survey, and New York City administrative records — we present a
model to create real-time population estimates classi- fied by residents, workers,
and visitors/tourists in a given neighborhood and localized to a block or
geolocation proximate to a Wi-Fi access point. The results indicate that the
approach has merit: we estimate intra-day, hourly worker and resident population
counts within 5% of survey validation data. Our building-level test case demon-
strates similar accuracy, estimating worker population to within 1% of the
reported building occupancy.
+
+
+Qin
+With the explosive growth and wide-spread use of smartphones with Wi-Fi enabled,
people are used to accessing the internet through Wi-Fi network interfaces of
smartphones. Smartphones periodically transmit Wi-Fi messages, even when not
connected to a network. In this paper, we describe the Mo-Fi system which monitors
and aggregates large numbers of continuous Wi-Fi message transmissions from nearby
smartphones in the area of interest using nonintrusive Wi-Fi sniffer sensors. In
this paper, we propose an optimized Wi-Fi channel detection and selection method
to switch the best channels automatically to aggregate the Wi-Fi messages based on
channel data transmission weights and human presence activity classification
method based on the features of human dwell duration sequences in order to
evaluate the user engagement index. By deploying in the real-world office
environment, we found that the performance of Wi-Fi messages aggregation of CAOCA
and CACFA algorithms is over 3.8 times higher than the worst channel of FCA
algorithms and about 76% of the best channel of FCA algorithms, and the human
presence detection rate reached 87.4%.
```

```
diff --git a/references.bib b/references.bib
index d29ae1d..8e5045f 100644
--- a/references.bib
+++ b/references.bib
```

```
@@ -2222 +2222 @@
-  author = {Gonz{\'{a}}lez, Marta C. and Hidalgo, Cesar A C{\'{e}}sar A. and
Barab{\'{a}}si, Albert-L{\'{a}}szl{\'{o}} and Gonzalez, Marta C and Hidalgo, Cesar
A C{\'{e}}sar A. and Barabasi, Albert-Laszlo},
+  author = {Gonz{\'{a}}lez, Marta C. and Hidalgo, C{\'{e}}sar A. and
Barab{\'{a}}si, Albert-L{\'{a}}szl{\'{o}}},
```

```
@@ -3200 +3199,0 @@
-  pages = {1-3534},
```

```
@@ -3753 +3752 @@
-  author={mikeazo and poncho},
+  author={Mikeazo and Poncho},


@@ -3806,0 +3806,18 @@
+
+@inproceedings{gueron2011,
+ author = {Gueron, Shay and Johnson, Simon and Walker, Jesse},
+ title = {SHA-512/256},
+ booktitle = {Proceedings of the 2011 Eighth International Conference on
Information Technology: New Generations},
+ series = {ITNG '11},
+ year = {2011},
+ isbn = {978-0-7695-4367-3},
+ pages = {354--358},
+ numpages = {5},
+ url = {http://dx.doi.org/10.1109/ITNG.2011.69},
+ doi = {10.1109/ITNG.2011.69},
+ acmid = {2006621},
+ publisher = {IEEE Computer Society},
+ address = {Washington, DC, USA},
+ keywords = {hash algorithms, SHA-512},
+}
+
```

```
diff --git a/template.tex b/template.tex
index 04b0a76..d6f9905 100644
--- a/template.tex
+++ b/template.tex


@@ -17,0 +18,2 @@
+\usepackage{glossaries}
+\usepackage{csquotes}
```

```
diff --git a/thesis.tex b/thesis.tex
index 53ed5ae..fa98e0e 100644
--- a/thesis.tex
+++ b/thesis.tex


@@ -7 +7 @@
-\title{Estimating Footfall From Passive WiFi Signals}
+\title{Estimating Footfall From Passive Wi-Fi Signals}
```

```
diff --git a/tufte-common.def b/tufte-common.def
index 52e0718..0470803 100644
--- a/tufte-common.def
+++ b/tufte-common.def
```

```
@@ -1600 +1600 @@
-    {\vspace{1.5\baselineskip}\begin{fullwidth}\LARGE\rmfamily\itshape} % above
(global formatting of entry)
+    {\vspace{1\baselineskip}\begin{fullwidth}\LARGE\rmfamily\itshape} % above
(global formatting of entry)


@@ -1608 +1608 @@
-    {\hspace*{2em}\contentslabel{2em}} % before w/label (label = ``2.6'')
+    {\hspace*{4em}\contentslabel{2em}} % before w/label (label = ``2.6'')
```