Challenges in Digitising Kampala's Vehicle Collision Data

Michael Thomas Smith
Department of Computer Science
University of Sheffield
m.t.smith@sheffield.ac.uk

Jimmy Owa Konyonyi
Department of Computer Science
Makerere University, Kampala
jymgonza@live.com

Abstract

Road traffic collisions are becoming a public health crisis, in particular in developing countries. However, very little data is available regarding the details of when and where the collisions happen. This data is essential for both planning interventions and monitoring their efficacy. We describe crashmap.org, a website developed to allow the police traffic records to be crowd-sourced into a digital form. We consider a few basic analyses of this data; looking at the effect of rainfall on the statistics, and at the effect time and day have on the results. We find an unexpected trend towards drier days having more collisions and a peak in collisions around 8pm or 9pm, later than the peak in other parts of the world. We summarise some of the caveats around the current dataset, and the ways in which it will be utilised in the near future, in particular for the placement of vehicles in the city's new ambulance service.

1 Introduction

Over a million people are killed in road traffic collisions around the world annually, and it is the leading cause of death among 15 to 29 year olds (WHO, 2013). These injuries and fatalities occur disproportionately in developing countries. Despite increasing international efforts to reduce road traffic fatalities (exemplified by the UN's Decade of Action for Road Safety), there is a dearth of data from developing countries on the statistics of these events. This data is essential for both planning interventions and monitoring their efficacy.

In Kampala, the collision data is stored in handwritten log books. In this paper I summarise the current work of collecting and transcribing this data, and the intended processing, disseminating and utilisation that we are starting to perform. This article is deliberately informal, with the intention of including some of the aspects of coincidence and chance that are often lost in formal submissions. It is also an incomplete project and requires further data collection, analysis and dissemination.

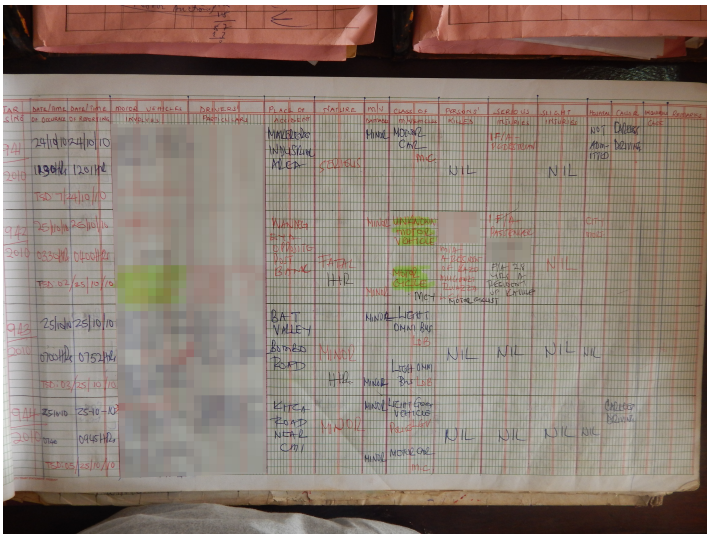
I finally, briefly look at the potential biases, and connections with future work aimed at mitigating some of these and summarise some of the more immediate uses for the data.

2 Data Collection and Transcription

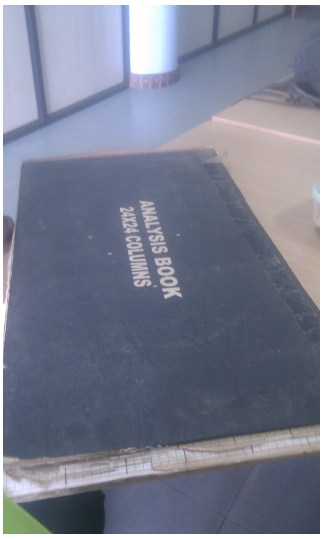
2.1 Finding the data

The project has been largely opportunistic, and depended on chance meetings and informal connections with data holders and users. I'll briefly describe the events which led to this dataset.

I lectured at Makerere, Kampala, for the majority of 2014. During my time there I was allocated several MSc students to supervise. One of the students, Jimmy Owa Konyonyi (JK), was also a

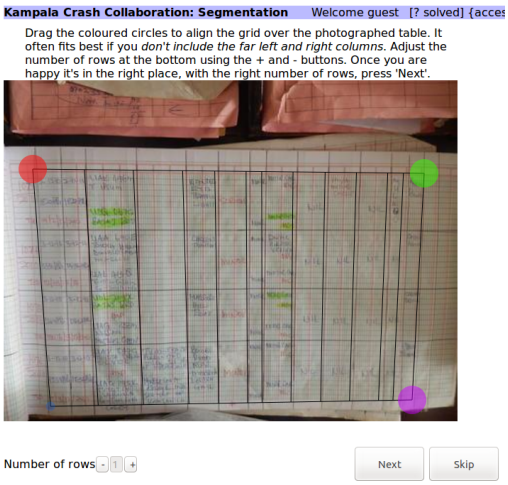


(a) Example page from a traffic collision log book (with sensitive items redacted).



(b) One of the books.

Figure 1: The data for this project was collected from police records.



(a) Segmentation of the page into cells.

(b) Transcription of the location.

Figure 2: Stages in crowd sourced transcription: (a) The pages of the book are down-scaled to make the text illegible, then a volunteer drags the grid over the page, adjusting the number of rows to match. (b) Each cell is then transcribed separately. For location cells the volunteer clicks on a google map to mark the location.

police officer (teaching digital forensics) in the Ugandan Police Force. After some discussion about possible projects, it was found that he could receive permission to use the road collision data held by the Kampala traffic police.

2.2 Collection

Kampala is divided into several areas (e.g. Katwe, Kiira Road, Old Kampala, Wakiso, Kakiri, Wandegeya, etc) by the traffic police. Each has an area office, where reports are kept on collisions. These reports are stored as hand-written rows in large log books (see figure 1). So that the books didn't need to leave the police offices, JK photographed each page of the books on site.

2.3 Transcription

We considered several options for the text transcription. Most OCR researchers agreed that automatic segmentation and transcription would be problematic. We looked at mechanical turk from Amazon and discussed the issue with Captricity, a business involved in form-transcription. The latter's workflow however depended on significant automated transcription and carefully aligned form cells. We also spoke with people from Zooniverse, but it was felt that maybe this project did not fall sufficiently within their scientific remit.

Providing the full pages for crowd-source transcription would have been the most simple transcription method. However, the pages contain a combination of sensitive data (such as names and number plates) and less-sensitive data (although see section 4.1 for details). In summary each cell of the data tables fell into one of two categories; it either contained private information that couldn't be released to even a single transcriber (e.g. someone's name and address), or it contained information that alone did not pose a privacy violation (such as a date), but when combined with other cells could provide enough information to compromise a data subject's privacy. JC, through his police training had access to students with sufficient privileges to see the sensitive cells in the dataset. However there were far more pages than a small population of students could transcribe. To resolve this, we built our transcription system to process the data in two stages.

Stage one; The whole crowd can take part in segmenting the images. Each page is resized to an image small enough that individual cell text elements are illegible. The crowd then decide where the boundaries are of each cell, by dragging a grid over the page (figure 2a).

Stage two; These images are then sliced (segmented) into individual cells. Each cell is then transcribed by the crowd. Cells which describe those injured or killed sometimes contained sensitive information and so were only visible to users given privileged accounts (for example the police training staff). The cell describing location was text-transcribed, but also the crowd-member clicked on a google map to indicate the collision location (figure 2b).

3 Analysis

The transcription process is still ongoing (as of 21st June, 2016; 20,073 items have been transcribed, including 1,661 collision locations identified). As the dataset is currently very incomplete the analysis is preliminary and experimental.

3.1 Time and Place

Figure 3 illustrates the locations of all the collisions currently transcribed. Only three divisions have been transcribed completely, with another three still in progress. Many more books are being added to the dataset. In spite of this, the data so far does start to give an idea of the distribution, with concentrations around junctions and along main roads.

3.2 Effect of Rainfall

It was hypothesised, based on observation of Kampala traffic, that there would be more accidents on days that it rained (defined for the data-analysis as those days in which the $0.25^\circ \times 0.25^\circ$ tile containing Kampala experienced more than 1mm of rain. Data from Funk et al. (2014) - CHIRPS). Unexpectedly we found a (non-significant) trend in the other direction, with fewer collisions reported with more rainfall, with the number of collisions reported on rain-free days 12% greater than on rainy days ($t = 1.52$, $p = 0.13$). However, looking at particular time periods (for example, during the evening) show significant differences. A multiple-comparison corrected analysis will be conducted once data collection is complete. As a quick example, we note that there are significantly fewer collisions between 6pm and 9pm on days with rain ($t = 2.50$, $p = 0.013$).

4 Dissemination and Utilisation

With the permission of the police department we aim to provide the dataset online for unrestricted use. However, there are several issues that need addressing,

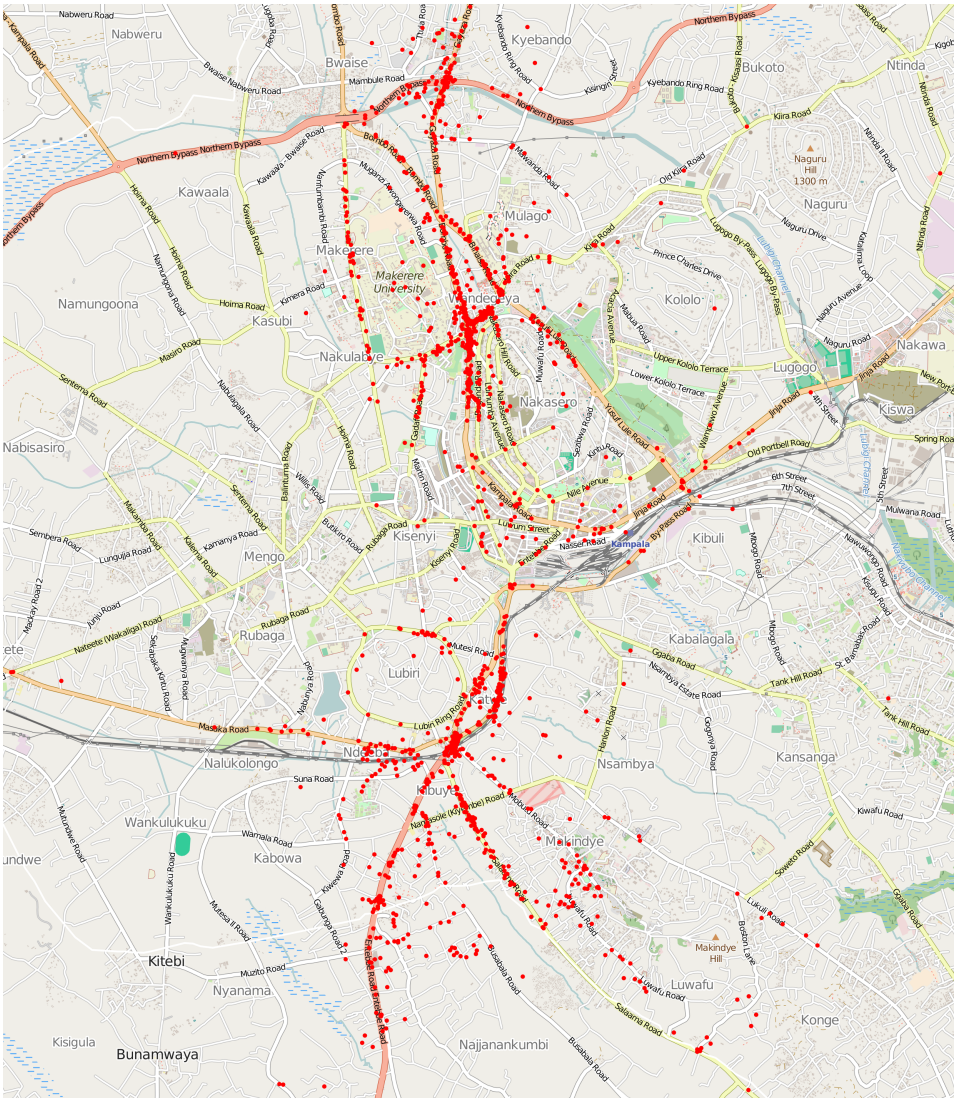


Figure 3: Map of transcribed collisions (as of 21-06-2016). Base layer provided by *www.openstreetmap.org*.

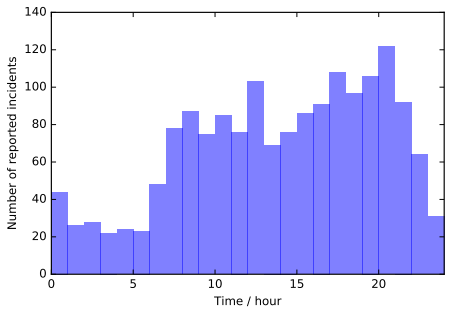
- The privacy of data subjects.
- Data quality.
- Data timeliness.

Below I briefly discuss these issues and potential mitigation strategies.

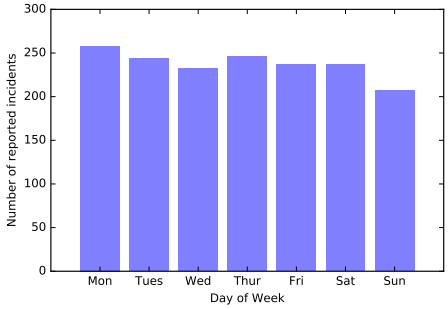
4.1 Privacy

Linkage attacks (where a subject's identity or attributes are revealed) must be protected against when releasing a dataset of this nature. Multiple examples exist in which another source of information is used in combination with an 'anonymous' dataset to cause a subject's membership or attributes to become revealed (Ohm, 2010; Barbaro et al., 2008; Narayanan & Shmatikov, 2008; Hern, 2014). An example of how this dataset could leak private data is as follows;

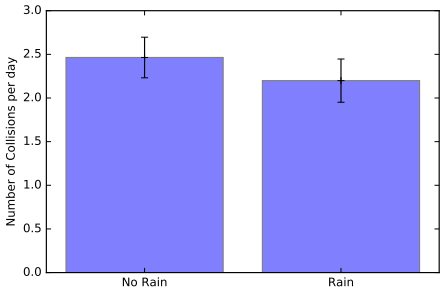
> Frances, a recovering caffeine addict, had a collision in her car, and when she got home explained to her partner, Beth, that she'd had a collision an hour earlier, outside work. A few weeks later, Beth was looking at the crashmap, filtering by



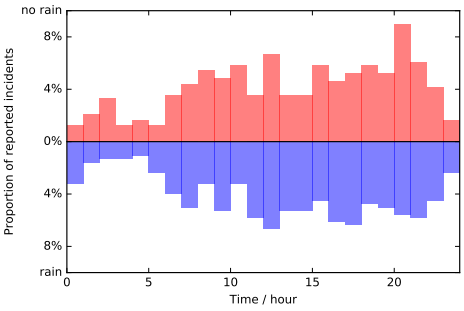
(a) Time of Day.



(b) Day of Week.



(c) Effect of rainfall (error bars are 95% CIs).



(d) Effect of rainfall over the day. Top graph, frequency with no rainfall. Lower graph, frequency with rain.

Figure 4: (a) Histogram indicating when collisions are most common. Note: Nightfall is about 7pm. (b) Histogram showing how the day of the week affects collisions: There is not a great difference between weekdays and the weekend. (c) Effect of rainfall with no significant overall effect. (d) Some time periods do appear to have larger differences though.

> time, and noticed that the only reported crash involving a car, that had happened that hour had occurred outside The Coffee House. Had Frances started drinking coffee again?

The crashmap contains no personal information, but by using side-information (about when Frances was in the collision), Beth was able to infer Frances' location that evening.

4.1.1 Differential Privacy

Differential privacy is a formal method for reducing the likelihood of record release, while maximising the utility of the statistics released (Dwork et al., 2014). A recent paper by Hall et al. (2013) extended the differential privacy framework to functions and functionals. Our paper (Smith et al., 2016) applies this to Gaussian processes. We applied the techniques described in our paper to this dataset, to investigate how we might proceed with the anonymisation of the dataset. The differentially private result will have random noise added, but will also improve the privacy of the data-subjects. Figure 5 illustrates the Gaussian process density distribution with $\varepsilon = 1$ (a good level of privacy). The contours roughly encircle the locations with the greatest incident counts, suggesting that the large-scale structure of the underlying data is captured. Note however that they currently do not take into account road segments, and have a very long length-scale, causing small-scale details (such as information regarding the junctions associated with the greatest numbers of incidents) to become lost in the process.

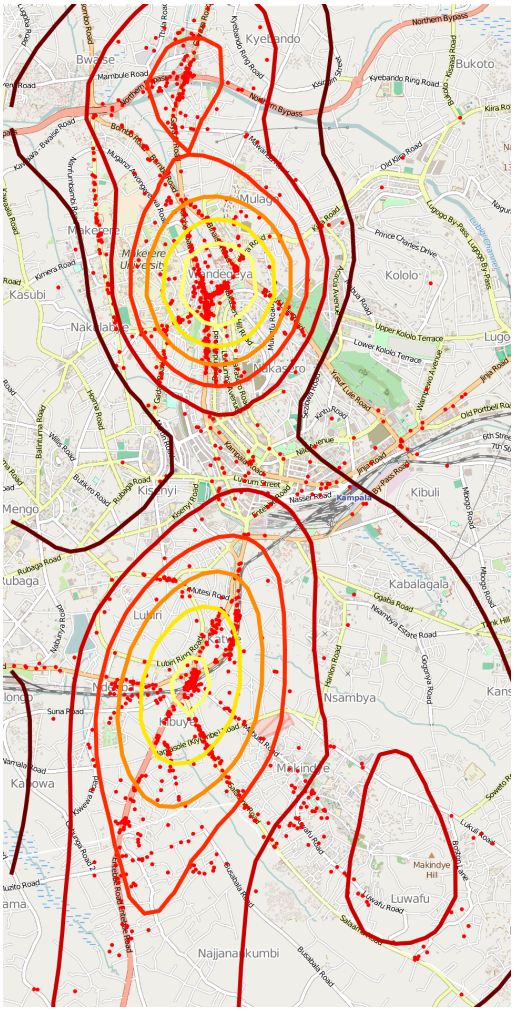


Figure 5: Differentially Private Density Map (using transcription data as of 21-06-2016)

4.1.2 Practical Privacy

There are other ways that data could be released, with, accidental or malicious releasing of data, being the most obvious. A specific issue to the crowd-source platform is the potential that an attacker could download and align the raw transcription image segments. This is somewhat mitigated by the small set of images available at any time, which usually do not have associated neighbouring cells to align to. The practical cost of having to separately transcribe these fragments also will be a deterrent. However, this is the biggest risk posed by this crowd-transcription platform methodology.

4.2 Data Quality

The dataset is far from accurate, with both bias and noise. Below a few issues are described, with some mitigation strategies suggested to manage them,

- Transcription errors: These are mainly random, but occasionally contain bias. For example the time of the report rather than time of incident is often wrongly transcribed by the crowd. This is dealt with by taking the earlier of any pair of times given. More insidious, is that some location descriptions given by the police are quite vague (e.g. a whole road). One might determine this uncertainty by the increased variance in the distribution of transcription locations, however, the crowd is likely to place the incident on the road label on the map. To mitigate this, modifications to the crowd-source platform are required to allow location uncertainty to be expressed. For example allowing a user to enter multiple locations,

or draw an ellipse or box over the map. Note that transcription errors can be reduced by combining the responses from several volunteers.

- Data availability bias: Currently we have only transcribed books from a short time period from a few police stations. This will improve as more books are entered into the system.
- Police reporting bias: The books only contain those collisions which have been reported. This is clearly affected by when and where the police are (collisions that occur immediately in front of a police officer will be more likely to be reported). For example, traffic police are stationed during rush-hour at the major intersections, directing traffic, which will mean the incident count at the intersections may be inflated. A possible mitigation will be to see the distribution of serious and fatal incidents, which may be more evenly reported.
- Collisions don't indicate dangerous roads, just busy roads: Depending on what a researcher is interested in, the raw number of collisions per km of road may not be of interest. For many applications it might be collisions per passenger-km travelled which is needed. We are currently in negotiations with ThinVoid and Tugende about using motor-bike taxi (locally 'boda-boda') GPS-recovery data to provide journey-density estimates for each road segment in Kampala. This will allow us to control for the number of journeys and potentially identify black-spots.

4.3 Data Timeliness

The dataset currently being transcribed is for incidents reported between 2010 and 2014. Senior personnel in both the police and ambulance service in Kampala have already asked about how to bring the dataset up to date, with the possibility of digitising the incident reporting system. It may be that demonstrating the possibilities offered by digitised data has motivated change within the system. It should be noted that such a transition is beyond the remit of this project, and is not without risk (e.g. the cost and skill demanded of a new system, and the long term integrity and security of the data). On-going transcription (possibly with sub-sampling) from record-books may be an acceptable interim solution. Although useful as a stand-alone dataset, its utility would be greatly increased by becoming a sustainable, maintained database of collision statistics, updated directly from police records.

4.4 Utilisation

Several agencies and departments approached me during this research, expressing an interest in the dataset.

- The police - Mainly interested in digitising their current system, they were also interested in using the data to improve officer deployment. These users will need access to individual rows through a form-style interface.
- Kampala Capital City Authority (KCCA) - Already installing speed-bumps and other traffic-calming measures, they could better target their interventions, and monitor the efficacy of an intervention with access to this detailed data.
- Ambulance Deployment - Joseph Kalanzi from the Department of Health was particularly interested in this dataset, as they are starting to deploy Kampala's first public ambulance service, and need to decide where to place the ten ambulances they have recently acquired. These users want an interactive visualisation to aid decision making.
- Researchers - We've had interest from researchers within academia, within business (e.g. GeoGecko) and charities (Tugende, ThinVoid) and within the UN Pulse Lab. This set of users would find simple (differentially private) machine readable data files the most useful.

5 Discussion

5.1 Time and Place

The current incompleteness of the dataset makes analysing the overall spatial extent of the collisions impossible. However, one can see significantly more collisions along busier roads and intersections,

as one might expect. We originally hypothesised the collisions would peak around 5pm or 6pm, following the pattern from other regions, such as Costa Rica (Flynn, 2013). However, the greatest rate of collisions appears to be after dark, at about 8pm or 9pm. My own observations confirm that the traffic is still very heavy at those times. I hypothesise that the 8:30pm peak is because, around this time, traffic is able to move more quickly, increasing the likelihood and severity of collisions. Combined with nightfall and the large number of pedestrians, an increase in collisions is likely. It will take further research to investigate whether this hypothesis is true, in particular by combining the collision data with the boda-boda GPS-tracking data to infer traffic flow.

5.2 Effect of rainfall

There are several hypotheses which might explain the reduced collision count during rainfall; 1. fewer people going into town for evening events, etc. 2. an increase in congestion due to rainfall, causing an associated reduction in traffic speed. 3. Fewer traffic police outside, observing incidents, and so fewer reports being generated.

With access to boda-boda transport data (provided by Tugende and ThinVoid) we'll be able to investigate the first two hypotheses. For the latter, we could restrict out analysis to the more serious incidents which we hypothesise will suffer from less under-reporting than the minor incidents (i.e. without injury).

5.3 Future Work

As discussed already, completing further data collection and transcription is the next step. It may be that to achieve the transcription rate required we will need to introduce a micro-payment system to compensate participants. We are also in discussions with Sinfa and Geogecko to scale up the transcription to the rest of the log books.

Regarding differentially private data release: It is likely that the Gaussian Process solution we applied was not a good fit to the dataset. Instead, we could use a more precise method within the differential privacy framework. For example, a modified Kernel Density Estimation distribution, based on the description in Hall et al. (2013) would provide a higher-resolution, with less noise added for the same privacy guarantee. It would also be interesting to treat the dataset as a Bayesian probabilistic modelling problem, in which the task would be to assign a probability to a collision occurring given a particular set of conditions.

Besides date, time and location, we are transcribing the type of vehicles and victims, and the severity of the collision. This is slowed somewhat by these columns being restricted to only those users with privileged access. Future analysis will be able to investigate those factors which affect particular types of victim (for example when are where are children most at risk?). By collecting more log books we'll be able to investigate long-term trends in the data.

Finally, integrating this dataset with the boda-boda GPS-tracking data will improve our estimates of a particular road's associated hazard. It may also allow the staging of ambulance locations to take account travel time (over the day).

5.4 Thoughts on Access to Data

From my experiences thus far in both developing and developed countries, it seems that initial access to data is often through these informal or chance meetings, rather than through formal, official lines of communication. It isn't clear whether this is changing in Uganda, although organisations in the region, like the Kenya Open Data portal (*opendata.go.ke*), may be moving us towards an ecosystem of more transparent and open data-access.

Sponsorship

crashmap.org is hosted using a donation from Amazon, for credit on Amazon Web Services.

References

Barbaro, M., Zeller, T., & Hansell, S. (2008). A face is exposed for AOL searcher no. 4417749. *New York Times*, *9 August*.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407.

Flynn, A. (2013). Maximizing resource efficiency in rural prehospital emergency medical services through call frequency analysis. *Research Journal of the Costa Rican Distance Education University*, *5*(2).

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., Verdin, J. P., Rowland, J. D., Romero, B. E., Husak, G. J., Michaelsen, J. C., Verdin, A. P., et al. (2014). A quasi-global precipitation time series for drought monitoring. *US Geological Survey Data Series*, *832*(4).

Hall, R., Rinaldo, A., & Wasserman, L. (2013). Differential privacy for functions and functional data. *Journal of Machine Learning Research*, *14*, 703–727.

Hern, A. (2014). New york taxi details can be extracted from anonymised data, researchers say. *The Guardian*, *27 June*.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, (pp. 111–125). IEEE.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA law review*, *57*, 1701.

Smith, M. T., Zwiessele, M., & Lawrence, N. D. (2016). Differentially private gaussian processes. *arXiv preprint arXiv:1606.00720 (submitted to NIPS 2016)*.

WHO (2013). *Global status report on road safety 2013: supporting a decade of action*. World Health Organization.