THE DARKER SIDE OF BIG DATA

Nishan Weerasooriya
Vice President (IT Operations)
DFCC Bank

VK (formally known as VKontakte) is the largest social networking platform in Europe and is specially popular among Russian speaking users. As at June 2016, it had 369 million accounts and currently is ranked as the 2nd most visited website in Russia according to Alexa rankings.

In February of 2016, a new on-line service called "Findface" was launched in Russia which used face recognition artificial intelligence algorithms to match a random face with a profile in the Russia's most popular social media site, vk.com.

As an experiment, a Russian photographer called Yegor Tsvetkov started snapping random photos off the streets for six weeks and then tried to use Findface to match the faces to VK profiles and the result was astonishing. He matched 60-70% of the photos to correct VK profiles and published these findings under the project titled "Your Face is Big Data"[1].

This story does not end there. Just three days after the media reported on the "Your Face is Big Data" project, an underground site called "Dvach" launched a massive campaign to dox (search for and publish private or identifying information about a particular individual on the Internet, typically with a malicious intent) Russian porn actresses using Findface.

The average life span of an adult film actresses career is roughly six to eighteen months. Most actresses keep this profession hidden away from their personal life due to the social stigma associated with this profession.[2]

Dvach users obtained high quality facial shots off porn movies, submitted them to "Findface" which matched these faces against millions of faces on VK and identified the public profile of the porn actress. They shared and archived these VK profiles on their forums. After finding a match, hordes of Dvach users descended on the actress's VK profile, spamming friends and family of the actress exposing their discovery. They also targeted photos picked from another Russian website "Intimcity" where prostitution services are advertised on-line and similarly attacked their VK profiles[3].

Big Data analytics examines massive amounts of structured and unstructured data to uncover hidden patterns, correlations and other insights. With today's technology, it is possible to

get near real time answers from these massive dumps of data helping organisations to find business insights at a scale never thought 30 years ago.

One of the poster children of Big Data promoters was the Google Flu Trends (GFT). According to the World Health Organisation 3-5 million people fall victim to severe forms of influenza and 250,000 - 500,000 die every year [4].

U.S. Centre for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS) traditionally collected flu data from clinical visits of physicians. CDC published national and regional data on the influenza spread every week usually with a 2 week time delay.

In 2008, Google developed a model to predict flu trends around the world using Big Data analytics based on hundreds of billions of user search quarries spread over 5 years. Initially Google started off with 50 million most popular quarries and the algorithms finally came up with 45 most relevant query terms. The accuracy of the predictions of the final model Google published was high as 97% compared to CDC data.[5]

Once the GFT went live, the health workers could see the flu outbreaks near real time instead of waiting for 2 weeks. Combined with CDC data, GFT could even predict the infections one week in to the future. Limited medical resources could be now deployed more efficiently, effectively and on time to most critical areas before the flu turned in to an outbreak.

Google and its big data analytics were celebrated in all the major news media. CNN, New York Times, Wall Street Journal, and many more hailed the dawn of new era. Why not? Google managed to mine hundreds of billions of general user search queries to find strong correlation among less than 50 terms out of 50 million search terms and predict the flu patterns all over the world near real time. This is what big data could do to humanity.

Fast forward to 2013 and GFT starts to go terribly wrong, sometimes predicting almost twice the actual data[6]. Algorithms which worked so perfectly few years ago were now throwing predictions so wrong and was forced Google to shut down the service altogether.

The purpose of the above examples was not to discourage the adoption of big data but to demonstrate that just as we appreciate the advantages of using Big Data, we equally need to appreciate the limitations of Big Data analytics.

Gods, humans and algorithms

In the beginning humans believed that authority came from supernatural powers of divine beings such as, Devils and Deities. Then during last few hundred years philosophers preached and

popularised the idea of Humanism which emphasizes the value and agency of human beings, individually and collectively, which generally prefers critical thinking and evidence over superstition which shifted the authority to human beings from the supernatural powers.

Now the world is moving towards another stage of authority shift preached by Technology Gurus where the authority is being shifted from human beings to algorithms running on Big Data where all our actions are being datafied in an era of IoT and human beings are just data points.

The transformation of the value of data

When the value of large volumes of data was getting noticed in late 1990's the value was in the management decision making on the primary data. But over the years the value of primary use of data has diminished and it has moved to the secondary use. Not only the value has moved to secondary use, but the value of secondary use has exploded.

Secondary use in this context is the use of the data for reasons outside its primary reason (intended reason) for collecting the data in the first place. For example, Google collects the browsing habits of Internet users to optimise the advertisement targeting. Better the targeting, higher the click rate thus higher the revenue. The secondary use of this data is the use of the same data to predict flue trends.

In 2014, Facebook which had assets valued at 6.3 billion dollars floated its first IPO selling 421,000 shares at USD 38 raising its valuation to 104 billion overnight. The intangible asset value was almost 100 billion dollars and the primary intangible asset they had was data, customer data of 500 million active users. The IPO had placed a value of USD 200 for each user.

The Internet had also brought the concept of "Data Exhaust" along with the advancement of Big Data. This is the digital trail or information by-products resulting from all digital or on-line activities. These consist of storable choices, actions and preferences such as log files, cookies, temporary files and even information that is generated for every process or transaction done digitally. Every click you make on the browser is tracked by multiple sites and joined with an already existing digital profile of you. Facebook had access to the data exhaust of 500 million members including their friend networks, family members, their interests, browsing habits, hobbies, buying habits and a lot more information left as trails on the facebook. This is the value that the Facebook investors looked at, not the profit margins.

According to IBM, everyday humans create 2.5 quintillion bytes of data which comes from virtually everywhere; sensors used to gather environmental information, posts to social media sites, cell phone signals and more.

The value of this data exhaust was first used on a mass scale by Google upon which they built an entire business empire. The Google search was a free service which generated no income to Google but every time a user searched Google and selected a page, the data exhaust of that user was used to profile the user and most relevant advertisements were pushed through its ad serving network giving the advertisers the optimum coverage that no other provider could offer.

So today with the value of Big Data moving from its primary use to secondary use and the data exhaust left by all the internet activates of users are meticulously collected by websites is raising new questions of privacy in a new information age.[7]

Privacy in the world of Big Data

In 2006, AOL (America On Line) released 20 million search quarries to the public for research purposes but to protect anonymity, a unique number was assigned to the user. By anonymising the personal identification information of the users, no one could individually identify any user, so they thought. New York Times dug deep in the hundreds of quarries of the user number 4417749 and more and more information of the user was discovered. Cross referencing against a directory listing it did not take much time for the newspaper to zero in to Ms Thelma Arnold, a 62-year-old widow who lives in Lilburn, Georgia proving that personal data anonymisation in a world of big data has very little value[8]. AOL immediately detracted the data dump but not before it was copied to many other sites. The ensuing public outcry of this exposure led to the ouster of AOL's chief technology officer and two other employees.

But Netflix, the premier on-line movie rental site did not learn the lesson. Barely two months after the AOL incident it released 100 million movie rental records of half a million users and offered a bounty for any person who could develop a better recommendation algorithm predicting the movies customers would like. As earlier Netflix sanitized the records by anonymising sensitive customer information which could have been used to individually identify the customer. It was not surprising that just weeks after the contest began, two University of Texas researchers — Arvind Narayanan and Vitaly Shmatikov — identified several NetFlix users by comparing their "anonymous" reviews in the Netflix data to ones posted on the Internet Movie Database (IMDB) website.

Both AOL and Netflix examples show that the anonymisation carries no guarantee of privacy in the era of big data. In AOL case the identity was exposed using the same data set and and in the Netflix case the identity was exposed by combining other sources of big data with the existing data set.

State as the holder of Big Data

Fast forward to year 2054

> *Captain Anderton* : Mr. Marks, by mandate of the District of Columbia Precrime Division I'm placing you under arrest for the future murder of Sarah Marks and Donald Dubin that was to take place today, April 22nd, at 08.04.

Howard Marks : No! I didn't do anything. Oh, God. Don't put that halo on me! Sarah! I wasn't gonna do anything!

Woman : Officer Scott, I'm with the Precrime Trauma Response Unit. I want you to sit here a minute and listen to me. Your husband is being arrested by officers from Precrime.

Sarah : Oh, God, Howard, no! Howard, don't cry.

This chilling dialogue is from the 2002 movie 'Minority Report' directed by Steven Spielberg and is loosely based on a short story of the same name by Philip K. Dick. The story is built around a revolutionary programme in Washington D.C. Where a specialized police department PreCrime stops murderers before they kill, reducing the murder rate to zero. Murders are predicted using three psychics, called "Precogs", who "pre-visualize" crimes by receiving visions of the future. Would-be murderers are imprisoned in their own happy virtual reality. The programme is so successful that the Federal government is on the verge of adopting the controversial program to spread it country wide.

Although this was science fiction back in 2002, we are already on track to such a future which could realise long before 2054.

The large corporates however, are governed by the laws and regulations and can be held responsible for their actions. But both the gains and risk to the individual are the greatest when the state holds and use big data analytics.

Project Blue CRUSH (short for Crime Reduction Using Statistical History) is a proactive and predictive policing approach based on statistical history is a well-known Memphis success. While Blue CRUSH had revolutionized police tactics in Memphis is one of the first success stories on predictive policing. It was one of the first police programs which began in around 2005.

Today various state police departments in USA and UK already use massive collections of records of past crimes to predict the crime hotspots to plan their resource allocations. Los Angeles and Santa Cruz Police Departments have reported that there have been a 33% reduction in burglaries, 21% reduction in violent crimes and 12% reduction in property crime in the areas where

predictive software is being used[9]. Outside US, London police has been using data driven policing to combat gang violence since 2014.

The Department of Homeland Security in USA announced its project called FAST (short for Future Attribute Screening Technology) in 2008. The system would identify you as a future terrorist by scanning your pulse rate, skin temperature, breathing, facial expressions, body movements, pupil dilation and other psycho physiological/ behavioural patterns. The technology would mostly be used at airports, borders, and special events. Fox News reported that the mobile units transmit data to analysts, who use a system to recognize, define and measure seven primary emotions and emotional cues that are reflected in contractions of facial muscles. The results are transmitted back to screened.

Some researches and scientists have questioned the rate of false positives such a system could raise and the ability of the system to 'read people's thoughts', it is potential violation of privacy laws of the US Constitution.

But going back to basics, FAST increasingly looks like a digitised version of a polygraph, popularly referred to as a lie detector which was invented way back in 1921. The polygraph too measures and records several physiological indices such as blood pressure, pulse, respiration, and skin conductivity while the subject is asked and answers a series of questions.

Another entrance to the predictive policing is a system developed in Rutgers University. Risk Terrain Modelling, or RTM, is an approach to spatial risk analysis and is used to identify risks that come from features of a landscape and model how they co-locate to create unique behaviour settings for crime. With a diagnosis of the attractors of criminal behaviour or other hazardous outcomes, the law enforcement agencies can make forecasts on potential crimes.

Other superpowers such as China are not too far behind using big data for predictive analysis of people. China maintained a document known as Dang'an (meaning 'record' in Chinese) on each and every citizen of mainland China since the times of Mao Zedong. The Dang'an includes information possibly found in a CV plus many other documents that would be considered private in some other countries. This dossier is said to include appraisals by supervisors and peers, academic reports from primary school to university, professional credentials, any criminal convictions or administrative penalties, club/society memberships, employment records and political history (such as Youth League and party membership and assessments). Some of the material is composed by the subject. Even the death certificate and eulogy may be placed in the file. The Dang'an together with the Hukou (the government system of household registration), has been an important part of the Chinese government's efforts to maintain control of its people.

In March 2016, Bloomberg reported that the the Communist Party of China has contracted its largest government owned defence contractor China Electronics Technology Group, to develop software to collate data on jobs, hobbies, consumption habits, and other behaviour of ordinary citizens to predict terrorist acts before they occur. "It's very crucial to examine the cause after an

act of terror," Wu Manqing, the chief engineer for the military contractor, told reporters at a conference in December. "But what is more important is to predict the upcoming activities."[10]

Much of the project is shrouded in secrecy. Only Wu, the engineer at China Electronics Technology, would speak on the record. This system would be able to draw portraits of suspects by cross-referencing information from bank accounts, jobs, hobbies, consumption patterns, and footage from surveillance cameras.

The program would flag unusual behaviour, such as a resident of a poor village who suddenly has a lot of money in his bank account or someone with no overseas relatives who makes frequent calls to foreigners. According to Wu, these could be indicators that a person is a terrorist. According to unnamed sources this anti-terrorism system would be first used on the citizens of the mountains of Xinjiang territory in China's northwest and the 3 million citizens in mountainous Tibet. Tibet has been occupied and ruled by China since 1951 and has caused ongoing tensions between the Chinese government and Tibetan people since then.

There are many who argue against the long term success of predictive policing. This is mainly due to the way predictive policing algorithms work. These systems rely on historic crime data to predict where crimes are likely to occur. Some systems also consume weather data, personal information of the family members, financial information of the citizens, property records and even social media posts of the citizens. But even with all this information, the systems can predict only events likely to occur similar to the past events.

The system will predict higher crime rates in areas which are known crime areas in the past. The police will focus more on these neighbourhoods increasing the number of arrests. These arrests will be fed back to the system reinforcing the algorithms predictions which will be used to predict more crimes in the neighbourhood. This feedback loop will work as a self fulfilling prophecy after some time ignoring new crime patterns which the system had not been fed or detected by the police who are depending on the system to carry out the policing work.

Another issue with such a system is the profiling of neighbourhoods and people. Bias against certain categories of people by police force is known all over the world. Although a system which profiles citizens and neighbourhoods purely based on algorithms without human intervention seems a magic bullet as against the police profiling, feeding the already biased crime data to the system will only enhance existing biases if not handled carefully.

The current systems can be even more granular focusing on individuals rather than communities.

The movie 'Minority Report' is based on a system which predict the crimes so accurately that the citizens are imprisoned for crimes that they are predicted to commit. The premise of punishing a would be perpetrator before committing a crime looks promising for a crime free society. No

one has to die of a bomb blast to punish a terrorist. No one has to die to punish a murderer and no one has to be raped to punish a rapist.

And we don't have to wait till 2054, that future is already here in 2016 creeping in to our lives undetected.

Big Data Analytics has already become a standard tool in the US. Various forms of automated "Risk Assessment" tools have now taken up permanent residency in the various judicial systems around the world.

In a paper submitted to National Conference of State Legislatures in USA in 2013, Alison Lawrence[11] states that a growing number of states have engaged in a "justice reinvestment" process that involves data collection and analysis of trends that drive prison populations and costs; and development and adoption of policies addressing those factors. At least 27 US states have enacted justice reinvestment reforms in recent years.

Setting bail has been an always difficult task for the judges. The judges has to consider whether the accused is likely to jump bail or commit another crime when setting out bail terms. In June 2015, New York Times reported that a USD 1.2 million "Risk Assessment" system would be rolled out to 21 jurisdictions in United States.

This system has found that only 10 key factors such as age, criminal record, past failures to appear in court and recent convictions will affect the defendant future decisions after crunching one half million past criminal cases. The algorithm would give defendants two scores — one for their likelihood of committing a crime and one for their risk of failing to appear in court — and flag those with an elevated risk of violence. These scores would be given to the judge before setting bail conditions.[12]

The intention of deploying such a system is noble. A growing body of evidence indicates that the USA's bail system keeps many low-risk defendants incarcerated before trial, while those who may pose a higher risk are released because they have the money to make bail.

One of the more controversial uses of big data analytics in the judicial system is the use of risk assessment tools for the sentencing of those charged with crime. The primary premise is to reduce recidivism by identifying high risk offenders and to reduce prison population by diverting low risk offenders away from prisons. The sentence is based on the Risk Profile Scores generated by different risk profiling systems used in each jurisdiction which not only consider the defendant's history, but also consider the demographics and profiles of parents which are beyond his immediate control.

In February 2013, the police of Wisconsin (USA) arrested Eric Loomis for driving a car that was used in a drive-by shooting. He had been arrested a dozen times before. Eric took a plea, and

the judge received an automatically generated risk score of Eric which indicated that he was likely to commit a violent crime in the future. He was sentenced to six years in prison plus five years of probation.

But Eric challenged the state's use of a risk score in his sentence, he cited many of the fundamental criticisms of the tools being used for risk profiling such as they are too mysterious to be used in court, that they punish people for the crimes of others, and that they hold your demographics against you. The Wisconsin Supreme Court ruled against Eric in July 2016, but the decision has validated some of his core claims which the legal experts consider as a jumping off point for future challenges questioning the validity of risk profiling systems within the judicial system.[13]

The third area in which the risk profiling is already being used in the judicial system is the parole process. Large number of Parole Boards in many states in USA already use predictions from Big Data analysis to decide whether to keep somebody in prison or to release him.

All the above examples starting from project FAST by the US state, predictive policing technologies such as project BLUE CRUSH to risk profiling algorithms used during various judicial processes have one thing in common. They analyse massive amounts of historical data and try to predict future crimes and pinpoint them to individual level.

Preventing crime before its happening is an alluring proposition. If the big data analytics can predict the future without error, algorithms would see our future with pin point accuracy. Free will and freedom of choice for the human kind well no longer be relevant. Every person on the planet will take the chosen path by algorithm. But as we know even the best algorithms cannot and will not be able to predict perfectly. The algorithms can make a prediction only based on a statistical probability limited by its data set. Even the accuracy of this probability figure will depend upon the validity of data set and the quality of the algorithm being used.

The fundamental issue with depending upon such a predictive system is that we impose a punishment before the crime is committed. Eric Loomis was given a 6 year prison term plus 5 years probation partly for the crimes he would commit in future. Similarly parole is refused for probable future crimes predicted by a system.

The would be perpetrator would never will be able to disprove his crime since it never happened. By punishing the accused before the crime, we would never know whether the crime would have been committed at all.

Such a system negates the very idea of presumption of innocence until proven guilty, the principle on which any legal system is built. When we impose punishments for the predicted crimes which may or may not happen in the future, the legal system denies the exsisitance of free

will and freedom of choice of its citizens. The denial of freedom of choice will be relevant only to the legal system but any action on which big data predictions are based.

Maybe one day the state will decide what education your child will have, which profession he will be entitled to enter based on your background, your parents background, race, culture, demographics based on the predictions of an algorithm. Though this idea seems to be far fetched, the technology is already here and being used in various spheres of which we are not aware. These predictive analytics are only expanding to new areas everyday unknown and unseen by the average person.

Changing context of privacy in an era of Big Data

Use of Big Data is at the infancy in the Sri Lankan context. The larger commercial entities are only now trying out the Big Data predictive analytics in selected niches. As Sri Lanka start this journey towards Big Data we need to be aware of the pitfalls of Big Data as owners of the data as well as the consumers of this data. In addition the protection of customers privacy will become a major challenge for the governments and organisations world over and new regulations are expected to change the way we protect and consume data.

The privacy laws of most countries are based on privacy principles of the Organisation for Economic Co-operation and Development (OECD) adopted in 1980 (http://oe.cd/1tC). These principles attempted to strike a balance between 'privacy of the individual and free flow of information'. At its core, processing of personal information must be lawful which in practice means that it is expressly permitted by law or the individual has consented to the use of the data after being informed the purpose for which the data is being used, at the time of data collection.

Once the major economies of the world encapsulated these principles in various privacy laws, the implantation of this regulation became a standard "Notice and Consent" which appeared on every website and on-line application. Users are provided with a lengthy and complicated privacy notice written by lawyers and given a binary option. "Agree" to proceed the site or "Disagree" and the user is pushed out of the site.

The validity and the usefulness of these notices has been much debated over the years but the in the era of the Big Data, a new challenge had arisen. Based on the "Purpose Specification Principle" of the OECD core privacy principles, the user had to be informed of the purposes for which personal data are collected not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes. In the era of Big Data the value or the purpose of the information collected is not apparent when the notice and consent is normally given. Since the value of Big Data has now completely moved from the primary use to the secondary use, the use of the information at the time of collection is never envisioned. Since the data collector would have to keep going back to the user each time the data set finds a new use, this becomes

prohibitively expensive to the collector and an irritating exercise to the individual who had already blindly agreed to a complicated and complex privacy statement. This situation has been further complicated by processing different data sets together to generate new insights. These complications present complex situations to most individuals who will fail to comprehend making the agreement useless[14].

Therefore the current legal frameworks for protecting privacy of individuals are now hindering the advancements of the Big Data Analytics in addition to forcing data collectors to burden users with repeated consent requests for new uses of the same data.

With a growing need to change the way we collect and protect personal data, OECD revised the Privacy Principles in 2013 to reflect the changes in the use of personal data.

The new guidelines have put more focus on the data holders accountability of how the data is used and less on the individual consent. This adds to the burden on the data holder to protect the individual personal data. These will be the guiding principles for most countries including Sri Lanka, when the existing privacy laws are revised. Therefore, the corporate sector should expect to carry a bigger burden of ensuring the privacy of the consumer data which they collect and should be ready with the internal processes and technology to face the privacy challenges which are bound come in the near future.

(This article was inspired by *"Big Data: A Revolution That Will Transform How We Live, Work, and Think"* by Viktor Mayer-Schönberger and Kenneth Cukier)

Reference

1 Egor Tsvetkov. 2016. *Your Face is Big Data - Bird In Flight*. [ONLINE] Available at: https://birdinflight.com/ru/vdohnovenie/fotoproect/06042016-face-big-data.html. [Accessed 15 August 2016]

2 Live Science. 2016. *The Porn Myth: Uncovering the Truth about Sex Stars*. [ONLINE] Available at: http://www.livescience.com/27428-truth-about-porn-stars.html. [Accessed 27 August 2016].

3 Global Voices Advocacy. 2016. *Facial Recognition Service Becomes a Weapon Against Russian Porn Actresses - Global Voices Advocacy*. [ONLINE] Available at: https://advox.globalvoices.org/2016/04/22/facial-recognition-service-becomesa-weapon-against-russian-porn-actresses/. [Accessed 27 August 2016].

4 World Health Organization. 2016. *WHO | Influenza (Seasonal)*. [ONLINE] Available at: http://www.who.int/mediacentre/factsheets/fs211/en/. [Accessed 27 August 2016].

5 Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant. 2009. *Detecting influenza epidemics using search engine query data*. [ONLINE] Available at: http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html. [Accessed 27 August 2016].

6 Lazer, D, 2014. The Parable of Google Flu: Traps in Big Data Analysis. *SCIENCE*, 343, 1203-1205.

7 Wayne Williams. 2015. *Minority Report could one day be real, thanks to big data and predictive analytics*. [ONLINE] Available at: http://betanews.com/2015/09/07/minority-report-could-one-day-be-real-thanks-to-big-data-and-predictiveanalytics-qa/. [Accessed 27 August 2016].

8 New York Times. 2006. *A Face Is Exposed for AOL Searcher No. 4417749*. [ONLINE] Available at: http://query.nytimes.com/gst/abstract.html?res=9E0CE3DD1F3FF93AA3575BC0A9609C8B63. [Accessed 27 August 2016].

9 Mark van Rijmenam . 2014. *The Los Angeles Police Department Is Predicting and Fighting Crime With Big Data*. [ONLINE] Available at: https://floq.to/O0dUG. [Accessed 27 August 2016].

10 Bloomberg . 2016. *China Tries Its Hand at Pre-Crime*. [ONLINE] Available at: http://www.bloomberg.com/news/articles/2016-03-03/china-tries-its-hand-at-pre-crime. [Accessed 27 August 2016].

11 Lawrence A. 2013. *Trends in Sentencing and Corrections: State Legislation. Denver: National Conference of State Legislatures*. [ONLINE] Available at: http://www.ncsl.org/Documents/CJ/TrendsInSentencingAndCorrections.pdf. [Accessed 10 August 2016].

12 The New York Times. 2015. *Judges Replacing Conjecture With Formula for Bail*. [ONLINE] Available at: http://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html. [Accessed 10 August 2016].

13 *Bloomberg . 2016. This Guy Trains Computers to Find Future Criminals. [ONLINE] Available at: http://www.bloomberg.com/features/2016-richard-berk-future-crime/. [Accessed 10 August 2016].*

14 Fred H. Cate, Viktor Mayer-Schönberger. 2012. Notice and Consent in a World of Big Data. [ONLINE] Available at: http://idpl.oxfordjournals.org/content/3/2/67.abstract. [Accessed 10 August 2016].