THE CENTRE FOR internet & society

Big Data in Governance in India: Case Studies

EDITORS **Elonnai Hickok, Sumandro Chattapadhyay, Sunil Abraham**

The Centre for Internet and Society, India

This work has been made possible by a grant from the John D. and Catherine T. MacArthur Foundation. The conclusions, opinions, or points of view expressed in the report are those of the authors and do not necessarily represent the views of the John D. and Catherine T. MacArthur Foundation.

Designed by **Saumyaa Naidu**

Shared under
Creative Commons Attribution 4.0 International license

Contents

Introduction

BY AMBER SINHA

This research seeks to understand the most effective way of researching Big Data in the Global South. Towards this goal, the research planned for the development of a Global South big data Research Network that identifies the potential opportunities and harms of big data in the Global South and possible policy solutions and interventions. The research was for a duration of 12 months and in form of an exploratory study which sought to understand the potential opportunity and harm of big data as well as to identify best practices and relevant policy recommendations. Each case study has been chosen based on the use of big data in the area and the opportunity that is present for policy recommendation and reform. Each case study will seek to answer a similar set of questions to allow for analysis across case studies.

What is Big Data

Big data has been ascribed a number of definitions and characteristics. Any study of big data must begin with first conceptualizing defining what big data is. Over the past few years, this term has been become a buzzword, used to refer to any number of characteristics of a dataset ranging from size to rate of accumulation to the technology in use.[1] Many commentators have critiqued the term big data as a misnomer and misleading in its emphasis on size. We have done a survey of various definitions and understandings of big data and we document the significant ones below.

Computational Challenges

The condition of data sets being large and taxing the capacities of main memory, local disk, and remote disk have been seen as problems that big data solves. While this understanding of big data focusses only on one of its features—size, other characteristics posing a computational challenge to existing technologies have also been examined. The (US) National Institute of Science and Technology has defined big data as data which "exceed(s) the capacity or capability of current or conventional methods and systems."[2] These challenges are not merely a function of its size. Thomas Davenport provides a cohesive definition of big data in this context. According to him, big data is "data that is too big to fit on a single server, too unstructured to fit into a row-and-column database, or too continuously flowing to fit into a static data warehouse."[3]

Data Characteristics

The most popular definition of big data was put forth in a report by Meta (now Gartner) in 2001, which looks at it in terms of the three 3V's—volume[4], velocity and variety. It is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.[5] Aside from volume, velocity and variety, other defining characteristics of big data articulated by different commentators are— exhaustiveness,[6] granularity (fine grained and uniquely indexical),[7] scalability,[8] veracity,[9] value[10] and variability.[11] It is highly unlikely that any data-sets satisfy all of the above characteristics. Therefore, it is important to determine what permutation and combination of these gamut of attributes lead us to classifying something as big data.

Qualitative Attributes

Prof. Rob Kitchin has argued that big data is qualitatively different from traditional, small data. Small data has used sampling techniques for collection of data and has been limited in scope, temporality and size, and are "inflexible in their administration and generation."[12] In this respect there are two qualitative attributes of big data which distinguish them from traditional data. First, the ability of big data technologies to accommodate unstructured and diverse datasets which hitherto were of no use to data processors is a defining feature. This allows the inclusion of many new forms of data from new and data heavy sources such as social media and digital footprints. The second attribute is the relationality of big data.[13] This relies on the presence of common fields across datasets which allow for conjoining of different databases. This attribute is usually a feature of not the size but the complexity of data enabling high degree of permutations and interactions within and across data sets.

Patterns and Inferences

Instead of focussing on the ontological attributes or computational challenges of big data, Kenneth Cukier and Viktor Mayer Schöenberger define big data in terms of what it can achieve.[14] They defined big data as the ability to harness information in novel ways to produce useful insights or goods and services of significant value. Building on this definition, Rohan Samarajiva has categorised big data into non-behavioral big data and behavioral big data. The latter leads to insights about human behavior.[15] Samarajiva believes that transaction-generated data (commercial as well as non-commercial) in a networked infrastructure is what constitutes behavioral big data.

Scope of Research

The initial scope arrived at for this case-study on role of big data in governance in India focussed on the UID Project, the Digital India Programme and the Smart Cities Mission. Digital India is a programme launched by the Government of India to ensure that Government services are made available to citizens electronically by improving online infrastructure and by increasing Internet connectivity or by making the country digitally empowered in the field of technology.[16] The Programme has nine components, two of which focus on e-governance schemes. Smart Cities Mission is an urban renewal and retrofitting program by the Government of India with a mission to develop 100 cities (the target has been revised to 109 cities) all over the country. To begin with, each of the schemes under these two wings were the subject of this case study. We looked at 67 e-governance schemes under these projects at different levels of implementation, and finally narrowed down to 5 case studies. In order to arrive at them, we looked at the stated objectives of each scheme, classified them under broad headings and focussed on schemes with a strong focus on three objectives. Integration and data consolidation, Interoperability and common standards, Data driven decision making.

While looking at what schemes could qualify as big data, we have considered schemes which are one of the following:

- Self Identified: Scheme policy documents describe the use of big data analytics and techniques.
- Publicly Identified: Described in publicly available third party sources as a scheme using big data or as big data being a critical component of the scheme.
- CIS Assessed: Schemes that indicate the use or generation of big data through aspects of the dataflow and that will enable a quantified society.

ENDNOTES

1. Thomas Davenport, Big Data at Work: Dispelling the Myths, Uncovering the opportunities, Harvard Business Review Press, Boston, 2014.
2. MIT Technology Review, The Big Data Conundrum: How to Define It?, available at https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/
3. Supra Note 1.
4. What constitutes as high volume remains an unresolved matter. Intel defined Big Data volumes are emerging in organizations generating a median of 300 terabytes of data a week.
5. http://www.gartner.com/it-glossary/big-data/
6. Viktor Mayer Schöenberger and Kenneth Cukier, Big Data: A Revolution that will transform how we live, work and think" John Murray, London, 2013.
7. Rob Kitchin, The Data Revolution: Big Data, Open Data, Data Infrastructures and their consequences, Sage, London, 2014.
8. Nathan Marz and James Warren, Big Data: Principles and best practices of scalable realtime data systems, Manning Publication, New York, 2015.
9. Bernard Marr, Big Data: the 5 Vs everyone should know, available at https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.
10. *Id.*
11. Eileen McNulty, Understanding Big Data: the 7 Vs, available at http://dataconomy.com/seven-vs-big-data/.
12. Supra Note 7
13. Danah Boyd and Kate Crawford, Critical questions for big data. Information, Communication and Society 15(5): 662–679, available at https://www.researchgate.net/publication/281748849_Critical_questions_for_big_data_Provocations_for_a_cultural_technological_and_scholarly_phenomenon
14. Supra Note 6.
15. Rohan Samarajiva, What is Big Data, available at http://lirneasia.net/2015/11/what-is-big-data/.
16. http://www.digitalindia.gov.in/content/about-programme.

Case Studies

Predictive Policing

BY AMBER SINHA

CCTNS

CCTNS (Crime and Criminal Network Tracking System) is an e-governance project under the Digital India mission which seeks to use ICT for better provision of citizen-centric services, connect about 14000 police stations across the country and facilitate investigation, detection and prevention of crime. Various states have decided to use predictive policing techniques. [1] The idea is to build on the already existing structured data based on geographic locations and the nature of crimes in locations, and databases of history sheeters and police reports with other alternative data. [2]

Predictive Policing

Predictive Policing is the application of big data analytical techniques to identify targets for prevention of crimes through police intervention or use of statistical predictions to solve crimes. [3] Predictive Policing usually works in the following four ways —

a) predicting places and times with an increased risk of crime,

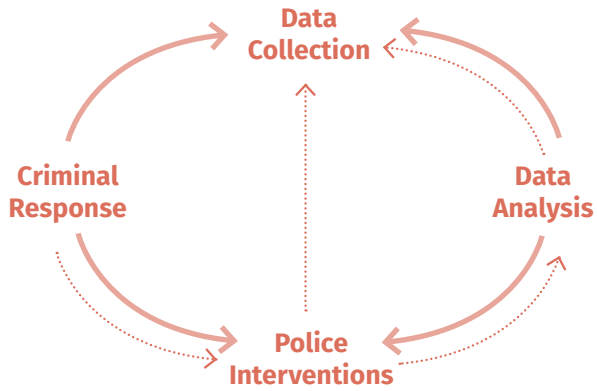
b) predicting potential future offenders,

c) creation of profiles for past crimes, and

d) predicting groups of individuals likely to be victims of crimes. [4]

Predictive policing draws from canonical theories of crime that focus on criminal events, crime-prone locations, and criminal opportunities. [5] The basic underlying assumption of predictive policing is that crime is not randomly distributed across people or places. Rather, patterns of crime are a "function of environmental factors that create vulnerabilities for victims and spaces at certain times." [6]

PREDICTIVE POLICING PROCESS



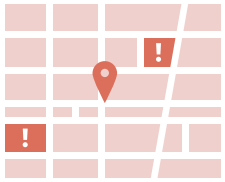
In the first step, data is collected from different sources — crime data from police station databases, environmental data including crime seasonal patterns, neighborhood composition, call data records and other mobile phone data. The next step is the Analysis, where the data collected is analysed based on a predictive method. Examples could include near-repeat theory, [7] social network analysis [8] and regression models using risk factors. The next step is making police interventions in response to the insights such as deploying more forces in a neighborhood. It is noteworthy that the very act of predictive policing also creates new data. The final step is to account for the criminal response to predictive policing and includes cases like displacement of crime to another area.

Inherent Biases

India has a history of tribes with a taint of inherent criminality which has led to racist and discriminatory practices by police.[9] This is widely reflected in their presence in history-sheets (database of people accused or suspected of a crime) and other police databases.

Pardhis (a de-notified "criminal" tribe) are routinely picked up by the police every time there is a crime in the area. A large number of people belonging to the community have records in police databases by virtue of belonging to the community.

1 A predictive policing model might take historical data about a particular type of crime, the location and time of that crime, and plot those past crimes in a way that would inform crime analysts about an unusual cluster of crimes.



2 The police looks at previous crime recorded to try to predict and prevent thefts in a locality. Towards this, the police performs Hot Spot analysis[10] to identify likely areas where crime will occur. Regression analysis[11] is also used, which looks at factors such as prior burglaries as well as counts for other types of crime, counts for vandalism and other types of disorder, numbers of homes in the area, numbers of unoccupied homes, the number of individuals with recent convictions for property crimes. This throws a list of areas which are crime prone. Further, Near Repeat Methods are used to narrow down streets which are most risk-prone in the immediate future.

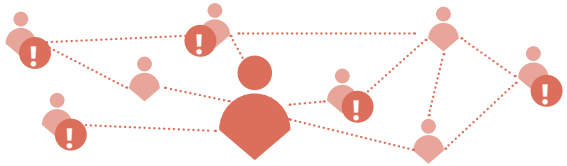
3 **HIRA**

Hira, a member of the Pardhi tribe lives on one of the streets identified.

Hira is a 'history sheeter' and it is possible that factors/data that lead to him being suspected are in itself biased data from police databases.

Incomplete and biased data

Data censoring, systematic bias and irrelevant data lead to incorrect patterns and inferences. Implicit biases in police data about criminal tribes can lead to amplified biases in the case of Hira. Discrimination may be an artifact of the data collection and analysis process itself. Even with the best intentions, algorithmic decision-making can lead to discriminatory practices and outcomes.[12]



4 Upon social network theory analysis, the police may find many connections to other criminals/suspects/history-sheeters, as he is from a community with a high number of history sheeters.

5 Given the data, the risk assessment for individual criminal behavior might suggest that Hira is at a high risk of committing crimes. This could lead to the police arresting him on reasonable suspicion and focussing the investigation solely on him

Data driven myopia bias/ Automation bias

In the presence of data and automated systems, law enforcements agencies have been reported to focus solely on these factors. For instance, in Washington DC, the officers had a tendency to only parol areas indicated in the heat maps and ignore other neighborhoods. Therefore, any use of predictive policing software must also be accompanied by a training of law enforcement officers to critically question its insights while applying them.

This perpetuates a vicious cycle of injustice against an already disenfranchised community.

BENEFITS

Better allocation of resources

In India, the police force is overburdened with work leading of health and social issues.[13] Cops usually work seven days a week and often have to attend to very long shifts. Therefore, any technology and system of policing which enables a more efficient allocation of resources is extremely desirable.

Preventive Policing

Predictive Policing offers the opportunity to the law enforcement agencies to preemptively act against predicted crimes by focussing on crime-prone areas and individuals at the risk of offending or being targeted. If crimes can be stopped before they are committed, it has great social and economic value not just for those at the risk of being victims of such crimes, but also for the offenders, as they can be stopped from making life altering mistakes.

More holistic analysis

Predictive Policing is a multi-disciplinary process which seeks to bring together insights from diverse fields such as actuarial science, statistics, criminology and a contextual understanding of local surroundings.[14] Therefore, it is hoped that this will lead to more comprehensive and holistic analyses of crime patterns.

HARMS

Lack of safeguards to prevent abuse

Predictive policing entails preemptively responding to the threat of crime. This automatically raises questions of preventive measures inconveniencing and infringing on the rights of innocent people. There are already existing provisions in the Code of Criminal Procedure in India which allows arrest upon suspicion.[15] Abuse of predictive policing could lead to warrantless arrests and detention without probable cause.

Discriminatory impact of predictive policing

Any data driven decisionmaking system runs the risk of amplifying existing inequities. Each intervention made under predictive policing also feeds into the data that informs decisions. Thus, if certain neighborhoods are identified as crime-prone, there will be greater policing in those areas, and consequently, lead to more focus on the same areas.

The form of discrimination in predictive technologies is institutional in that there are implicit biases in data. In India, there are documented problems with the police data which includes 'history sheets', 'rowdy sheets' and records of individuals classified as 'hooligans', 'goondas' and 'criminal tribes.'[16]

Further, the predictive policing algorithms are inscrutable to the courts in many circumstances. Therefore, despite constitutional protections preventing the discrimination, redressal mechanisms are not effective.

Data ideology

Use of predictive policing in other jurisdictions has raised questions of over-reliance on data and ignoring other factors. For instance, police patrols could have a tendency of only patrolling areas pointed out by heat maps.[17]

Surveillance without just cause

One of the key goals of predictive policing is to predict likely offenders. The likely interventions in response to such analysis would involve greater attention to the person in question in the form of surveillance. Such measures are problematic and raise new questions about how the standard of reasonable suspicion can be measured against algorithmic insights.

Opacity of predictive models

What data goes into the predictive models, what assumptions are being made by the algorithms, and what kind of contextual questions are does the algorithm ask are entirely opaque. There is a need for greater transparency and making information available in form which is accessible.

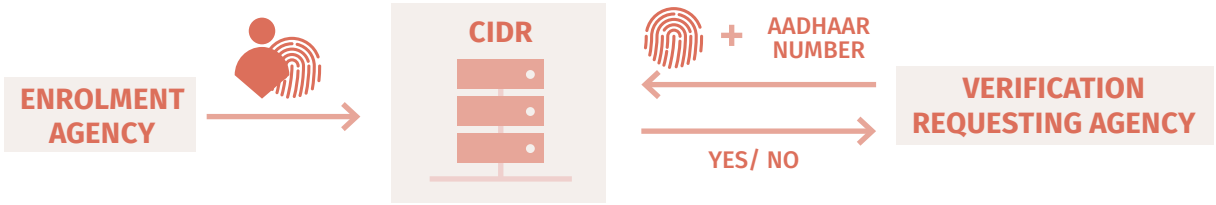
ENDNOTES

1. National Crime Records Bureau. "About Crime and Criminal Tracking Network & Systems - CCTNS." available at http://ncrb.gov.in/cctns.htm.
2. Bibhu Prasad Routray, Making a case for futuristic predictive policing in India, available at http://www.newindianexpress.com/magazine/voices/2012/sep/09/making-a-case-for-futuristic-predictive-policing-in-india-404220.html.
3. Sarah Brayne, Alex Rosenblat and Danah Boyd, Predictive Policing, DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE, available at http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf.
4. Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, available at http://www.rand.org/pubs/research_reports/RR233.html Andrew Guthrie Ferguson, Policing Predictive Policing, forthcoming in 94 WASH. U. L. REV. (2017).
5. Supra Note 3.
6. Andrew Ferguson. G. 2012. "Predictive Policing and Reasonable Suspicion." 62 Emory Law Journal 259.
7. Supra Note 4.
8. Id.
9. Rule 1054 (3) of the Karnataka Police Manual, which states: "History Sheets should be opened for those registered ex-notified tribe members...for whom the Superintendent or Sub-Divisional Officer thinks it is advisable to do so on account of their active criminality.
10. Hot spot methods predict areas of increased crime risk based on historical crime data. Hot spot methods seek to take advantage of the fact that crime is not uniformly dis- tributed, identifying areas with the highest crime volumes or rates. The underlying assumption—and prediction—is that crime will likely occur where crime has already occurred.
11. Regressions fit a mathematical relationship between the variable to be predicted and independent "explanatory" variables. In contrast with hot spot mapping, regressions project future crime risk based not just on past crimes but also on what can be a wide range of data.
12. Solon Barocas and A. Selbst, "Big Data's Disparate Impact." California Law Review. 104. SSRN Scholarly Paper.
13. "National Requirement of Manpower for 8-Hour Shifts in Police Stations", by the Bureau of Police Research & Development, Government of India, and Administrative Staff College of India, Hyderabad
14. Andrew Guthrie Ferguson, Policing Predictive Policing, forthcoming in 94 WASH. U. L. REV. (2017).
15. Section 41 of the Code of Criminal Procedure.
16. Mrinal Satish, Bad Characters, History Sheeters, Budding Goondas and Rowdies, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1703762
17. Supra Note 4.

The Unique Identity Project

BY AMBER SINHA

The Unique Identity Project in India, Aadhaar, which is reported to have over one billion registrants [5] is an online, digital and paperless identity system [6] which can act as a platform for a number of digital services, all of which produce enormous troves of data, precious to both the government and the private sector. Once residents have a uniform identification number and technology which is used across a range of services, it allows service providers to leverage this existing system, rather than engage in the verification process themselves. [7] Further, the presence of a digital identity allows for a host of services to move online and facilitates enormous transaction generated information which is useful behavioral big data. [8] We look at some key features of this project and ways in which we see Aadhaar manifesting itself as big data.



Digital Identity

In recent years, digital identity systems have been promoted as key to reducing fraud, facilitating financial inclusion, providing for efficient delivery of services, enabling political empowerment, and facilitating economic growth and security in developing countries. A Digital Identity system includes the following components: 1) identification, 2) authentication, the process of asserting an identity previously established during a process of identification, and 3) authorization, the process of determining what actions may be performed or services accessed on the basis of the asserted and authenticated identity. [1] Because of these features, a digital identity becomes critical for the use of online services such as e-governance or e-commerce platforms, and facilitates the generation of enormous amounts of transactional data - termed as transactional identity. [2] This transactional data results in behavioral big data. [3] A digital identity system can be both centralised and decentralised. [4]

CIDR

The Central Identities Data Repository (CIDR) is a centralised database containing all Aadhaar numbers, demographic information and biometric information. The presence of a central repository of all enrollment data has attracted severe criticism, [9] with commentators comparing it to a “honey-pot” with both the username (Aadhaar No.) and the password (biometric information). [10]

Authentication procedure

In order to authenticate an individual, the biometric information is matched with the corresponding Aadhaar’s number’s biometric information in the CIDR. [11] The false positive ratio (the probability that the identifiers of two persons will match) of this procedure has been estimated to be extremely high, and in a population size of 1.2 billion people, the expected proportion of resulting duplicands could be as high as 1/121. [12]

FALSE POSITIVE RATIO

1/121
Duplicands

AADHAAR

1.03 billion
People Enrolled

Seeding

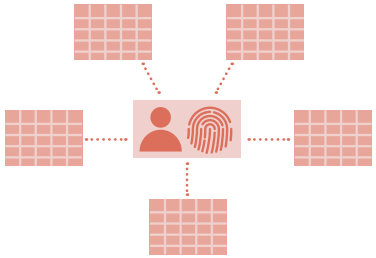
Seeding is a process by which the Aadhaar numbers of residents are included in the service delivery database of service providers for enabling de-duplication of database and Aadhaar based authentication during service delivery. Seeding allows different databases to be tagged with unique identifiers, thus establishing a relationality, a key feature of big data datasets. Once seeded, disparate datasets can come together and enable big data processing.

HOW DOES IT WORK?

The seeding process itself can be done through manual/ organic processes or algorithmic/in-organic processes.

Ginger Platform

Service providers that adopt the Aadhaar number must move their existing databases onto the Ginger platform, which then organizes the present and incoming data in the database by individual Aadhaar numbers. Once organized, automatically or manually, data can be queried by Aadhaar number by person's on the 'control' end of the Ginger platform.



MANUAL/ ORGANIC SEEDING

- BANKS
- DOOR-TO-DOOR COLLECTION
- COLLECTION AT POINT OF SERVICE DELIVERY
- CROWD SOURCING - IVRS, SMS, DROPBOXES
- 3rd PARTY SERVICE PROVIDERS
- COMMON SERVICE CENTER (CSC)

ALGORITHMIC/ IN-ORGANIC SEEDING

- USAGE OF AADHAAR SEEDING APP ON SMART PHONE / POS DEVICE / TABLET, ETC.
- MATCHING THROUGH USE OF STATE RESIDENT DATA HUB (SRDH)
  - Database of Beneficiaries
  - KYR + DATA from Enrolment Agencies
  - EID-UID Database from UIDAI

OPEN QUESTIONS

Lack of data protection regulation

The lack of a comprehensive data protection regulation and privacy law in India translates into an absence of rules on how government agencies may share data with each other. This could lead to issues such as ambiguity over who has data ownership, indiscriminate sharing of data across databases, and mission creep where data collected for one purpose is used for other secondary purposes.

Convergence

Anyone having access to the 'control' end of the Ginger platform can access all data associated to an Aadhaar number. That convergence can now easily be initiated with databases on the Ginger platform and the profiling of individuals can take place through the linking of data points via this platform. Thus, UIDAI effectively get access to all "seeded" data.

Technological failure

Poor technology with low resolution cameras often result in problem in enrollment and authentication. Such technological failures have very high costs in India where people may rely on the technology for essential services like rations, benefits and subsidies, and credit opportunities. [13]

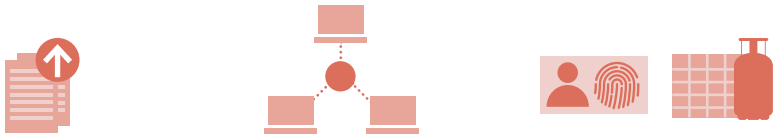
A Sample of what the database looks like
BEFORE AADHAAR SEEDING

Village Name	Applicant No.	Scheme Identifier

A Sample of what the database looks like
AFTER AADHAAR SEEDING

Village Name	Applicant No.	Scheme Identifier	Aadhaar Number

DATA EXTRACTION → CONSOLIDATION → NORMALIZATION → MATCHING



AADHAAR DATA DIGITIZATION

Data digitization may be done using existing Department Data (through use of physical Aadhaar letters/ e-Aadhaar print outs/ e-KYC).

DATA CENTRALIZATION

It does not necessarily mean collating all data at one physical location.

Software application users with authorized access should be able to access data online in a seamless fashion while providing service benefit to residents.

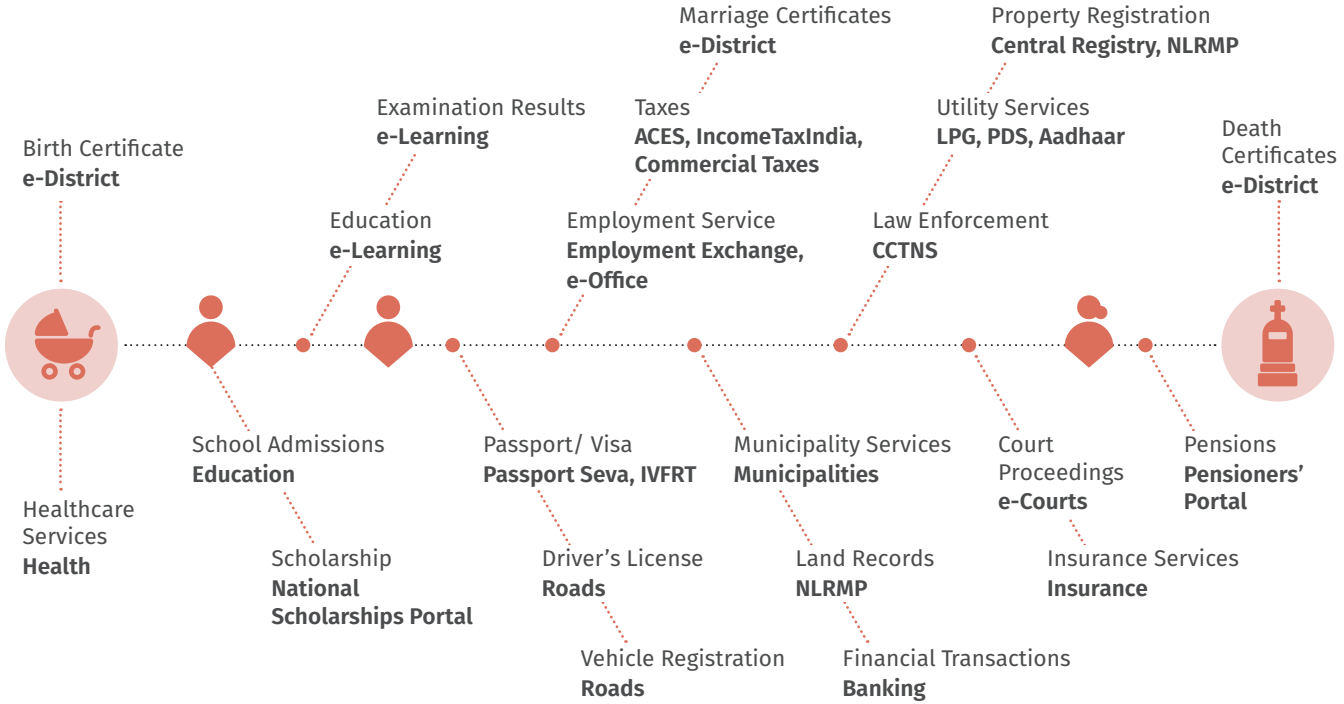
AADHAAR SEEDING

Aadhaar Seeding and Data centralization do not follow any particular order, and any activity may precede another, or both activities may run in parallel.

Cradle to Grave

DIGITAL INDIA

The original press release by the Central Government reporting the approval by the Cabinet of Ministers of the Digital India programme speaks as '**cradle to grave**' digital identity as one its vision areas. The characteristic of this identity are mentioned as "**unique, lifelong, online and authenticable.**"[14] We have done a survey of the e-governance schemes under **Digital India**, and tried to map those which seek to service individuals and create a welfare state.[15] In order to provide services effectively, these schemes shall collect and store information through the lifecycle of an individual. The result, as we can see, is building databases on individuals, which when combined, as enabled by the **Seeding** process, will provide **a 360 degree identification**. The schemes under Digital India conceptualise a digital ecosystem,[16] with the Aadhaar functioning, if not as the singular, the primary form of identity that would be used to enter this ecosystem.[17] This diagram represents a snapshot of how the entire lifecycle of an individual may be documented.



BENEFITS

Efficient service delivery

A consolidated view of the entire data would facilitate the social welfare department of the state to improve the service delivery in their programs, while also being able to ensure that the same person is not availing double benefits from two different districts

Convenience for the citizen

Convergence of data will mean that citizens are not asked for the same data many times and there could be a potential one stop shopping.[13] Therefore, updates like changes in address need not be done across multiple agencies and databases.

Better fraud management

Convergence of databases in a digital format from different government departments will make available more information in order to aid detection of cases such as tax fraud and money laundering.

Better information dissemination and training

One of the benefits of a holistic approach to government modernization and innovation is the ability to communicate information to citizens effectively about more than one program. Disparate projects require communications targeted to that individual program, and the citizen has only so much bandwidth to spend on figuring out how to do things and get service.[18]

HARMS

Profiling

Disparate pieces of data collected by different agencies shall come together to create a 360 degree profile of the citizens. This enables greater surveillance and impeded any efforts towards anonymity. This could lead to chilling effects on the citizens' right to free speech and expression.

Lack of trust

Sharing of data across government departments and agencies will mean that citizens may not know which organization ultimately holds and uses the data, and this will cause a lack of trust in the relationship between citizens and state.

Knowledge gap

A cradle to grave databases will lead to decisions being made using unrelated, inaccurate data without the knowledge of the citizen. People do not fully comprehend how powerful data mining techniques are, how much data analysis exists today and has been going on for decades, or who uses their personal data[19]

IndiaStack

IndiaStack is a set of APIs conceptualised by iSPIRT, a consortium of select IT companies from India, to be deployed and managed by several government agencies, including the National Payments Corporation of India (NPCI).[20] One of the key features of IndiaStack is supposed to be open-loop interoperability between providers, which will mean that different private and public parties can build over the APIs to create their own customized products and services.

Various talks about the IndiaStack speak of the following layers as integral to its structure: a) An identity layer supported by UID project, b) a presence layer which enables users and service providers to transact and communicate remotely, c) a paperless layer which leverages the existing functionalities built over UID such as e-KYC and e-Sign, d) a cashless layer which is an interoperable payments infrastructure, and e) a consent later based on the idea of OpenPDS conceptualised by a team at MIT Media Labs.[21] The cashless layer involves a payments switch which enables mobile account holders to transfer funds to and from any account in the country. With the implementation of the United Payments Interface (UPI), an architecture and set of APIs which will work on mobile,[22] users can transfer money between different banks, or even do an online transaction with far greater ease. As more and more transactions move from cash to noncash medium, the amount of data being generated increases dramatically. This is supposed to enable Anytime/Anywhere payments without a need to have access to cash.[23]

JANDHAN YOJANA

250 + million

Bank Accounts

MOBILE

1 billion +

Registered Mobile Phone Numbers

IDENTITY LAYER

An identity layer supported by the UID project.

PRESENCE-LESS LAYER

Digital technology that enables authentication and transactions remotely.

PAPERLESS LAYER

Functionalities like e-KYC, e-Sign and Digilocker built over Aadhaar to enable paperless registrations and verifications.

CASHLESS LAYER

A host of digital payments systems like Aadhaar Payments Bridge, Aadhaar Enabled Payments System, UPI and Bharat Bill Payment Service that can work with Aadhaar as identifier and on mobile phones

CONSENT LAYER

A XML artifact linked to the document/data being shared that specifies the provider, user, and purpose of use of that document/ data. Revocation of consent is only possible if the consent seeker allows it.

BENEFITS

Presence-less use

Once registered on Aadhaar, residents can remotely identify and authenticate themselves by inputting their fingerprint or iris information on a biometric reader. The Central Identity Data Repository (CIDR) which houses the Aadhaar enrolment data offers the basic validation service which takes as inputs the biometric and identity numbers and validates whether they match in the database. Residents can now sign up for services without having to be physically present.

Speedy and more efficient transactions

- A system that allows customers to use their Aadhaar numbers and biometric information to auto-populate information, converting what was a 30-day paper-based process into a short digital process
- A database that enables citizens to digitally store and retrieve loan documents, land titles, diplomas, training certificates and other documents in a single place.
- An API for digital signature in order to remotely and digitally sign and authenticate documents.

Reduce fraud

This data is valuable to governments (mitigate corruption and tax evasion, and making policy decisions), financial institutions (payments data to build credit scores), advertisers (purchasing data for targeted advertising), services providers (sales data to inform business plans).

One stop decentralized privacy control

Privacy and indiscriminate sharing of personal data has been an much spoken about problem in the digital world. The openPDS model intends to “allows users to collect, store, and give fine-grained access to their data all while protecting their privacy.”[24] Instead of sharing personal data and metadata directly, this model intends to employ SafeAnswers, which “allows services to ask questions whose answers are calculated against the metadata instead of trying to anonymize individuals’ metadata”[25]

HARMS

Complete loss of anonymity

Moving a cashless system as default mode of payment will decrease anonymity dramatically, and this could lead to greater surveillance and impacts on free expression and choice.[26] Data will be collected about purchases made, medical services sought and received, food and beverages consumed etc. This data can be used for targeted advertising and sharing of data across bodies can also lead to services providers making an assessment about recruitments, credits etc.

Potential denial of financial agency

Technological failures can be disastrous as it may deny financial agency. If services insist on payments through these modes only, it could lead to a completely inability of transact where there are technological failures. Similar cases of cashless mobile based payments in the Global South such as M-Pesa have faced problems of delays due to information having to pass through multiple platforms and connectivity issues.[27]

Predatory practices

Lack of regulation in the Fintech sector can enable predatory practices where right to remotely deny financial services can be granted to private sector companies. For instance, where all transactions are being handled through a mobile phone, inability to pay EMIs could lead to disconnection of mobile phone services.

Doubts over the consent layer

There is a lack of clarity about the whether there will be a blanket consent allowing service providers to authenticate the identity, and access all particulars of an individual, or the consent will be specific to each piece of information. Further, once consent has been provided, there are no effective opt-out mechanisms available in the Aadhaar Act or architecture. While the IndiaStack team has spoken of openPDS as the model for its consent layer, there is so far little clarity on how exactly it will work and how a model like SafeAnswers will be adapted.[28]

Regulation by code

An architecture like IndiaStack enables datafication of financial transactions in a way that enables linked and structured data that allows continued use of transaction data collected. This phenomenon is exacerbated due to the creation of a **state sanctioned monopoly** like IndiaStack as default mode for payments. This is a classic example of code behaving as law[29] and determining modes of behavior between citizens. Further, this is one of the first instances that a state sanctioned product has been developed by the private sector independently as opposed to under the RFP model, thus, circumventing regulatory oversights by the Comptroller and Auditor General and under the Right to Information Act.

ENDNOTES

1. Paul Makin, Steve Pannifer, Carly Nyst, Edgar Whitley, Digital Identity: Issue Analysis, available at http://www.chyp.com/wp-content/uploads/2016/07/PRJ.1578-Digital-Identity-Issue-Analysis-Report-v1_6-1.pdf
2. Decentralised Citizens ENgagement Technologies, Research on Identity Ecosystem, available at http://dcentproject.eu/wp-content/uploads/2015/08/D3.3-Research-on-Identity-Ecosystem_part1.pdf
3. Rohan Samarajiva, What is big data?, available at http://lirneasia.net/2015/11/what-is-big-data/.
4. On one side we have the view that we must adapt traditional ideas of identity, such as the ID card and ID number, to the digital age. This means that national governments should be the main providers of digital identities. On the other side are the proponents of using multiple identity sources and other more decentralised systems that do not rely on state assurance. These could provide a better balance between publicly controlled resources and personal data owned by citizens as commons.
5. http://www.ndtv.com/india-news/aadhaar-enrollment-crosses-1-billion-mark-ravi-shankar-prasad-1338621.
6. Jake Kendall and Stephen Deng, It's the ecosystem, stupid!, available at http://nextbillion.net/nexthought-monday-its-the-ecosystem-stupid-exploring-the-digital-poverty-stack-part-1/
7. *Id.*
8. Rohan Samarajiva, What is big data?, available at http://lirneasia.net/2015/11/what-is-big-data/.
9. Sunil Abraham, Surveillance Project, available at http://www.frontline.in/cover-story/surveillance-project/article8408866.ece.
10. Liliyan, UID Project in India - Some Possible Ramifications, http://cis-india.org/internet-governance/blog/uid-in-india
11. Section 2 (iii) and Section 7 of The Aadhaar (Targeted Delivery of Financial and Other Subsidies, Benefits and Services) Bill, 2016.
12. Hans Varghese Mathew, Flaws in the UIDAI Process, available at http://www.epw.in/journal/2016/9/special-articles/flaws-uidai-process.html.
13. Rajasthan presses on with Aadhaar after fingerprint readers fail: We'll buy iris scanners, available at http://scroll.in/article/806243/rajasthan-presses-on-with-aadhaar-after-fingerprint-readers-fail-well-buy-iris-scanners
14. http://digitalindia.gov.in/content/vision-and-vision-areas.
15. Sumandro Chattapadhyay, The Last Chance for a Welfare State Doesn't Rest in the Aadhaar System, available at http://thewire.in/30256/the-last-chance-for-a-welfare-state-doesnt-rest-in-the-aadhaar-system/.
16. Ravi Shukla. "Technologies in Aadhaar: A sociological view" in Ashish Rajadhyaksha ed., In the wake of Aadhaar: the Digital Ecosystem in India, available at http://www.jnu.ac.in/SSS/CSSP/RaviShukla.pdf.
17. Section 57 allows the Aadhaar number to be used in order to establish the identity of an individual by the government or private bodies. This provision runs counter to the claims that the Aadhaar number is for the purpose of identifying individuals for better targeted government subsidies.
18. https://www.oipc.ab.ca/media/389571/Report_Government_Information_Sharing_Jan2015.pdf
19. Peter Schaar, former data commissioner for Germany recently published a call to arms on data mining in a special edition of MIND http://www.collaboratory.de/images/1/11/PrivacyandInternetGovernanceMIND7.pdf.
20. http://www.indiastack.org/
21. Pramod Verma, Understanding the IndiaStack, available at https://www.youtube.com/watch?v=jwV0-NWmnbc.
22. http://www.thehindu.com/business/Economy/what-is-unified-payment-interface/article9037034.ece
23. Supra Note 15.
24. http://openpds.media.mit.edu/
25. Yves-Alexandre de Montjoye , Erez Shmueli, Samuel S. Wang, Alex Sandy Pentland, openPDS: Protecting the Privacy of Metadata through SafeAnswers, available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790
26. Dominic Frisby, Why we should fear a cashless world, available at https://www.theguardian.com/money/commentisfree/2016/mar/21/fear-cashless-world-contactless.
27. http://www.microsave.net/files/pdf/BN_94_Riding_M_Pesa_Rails_Advantages_Disadvantages.pdf
28. http://www.slideshare.net/ProductNation/india-stack-towards-presenceless-paperless-and-cashless-service-delivery-an-ispirt-initiative
29. Lawrence Lessig, Code v.2.0, available at http://codev2.cc/.

Big Data in Credit Scoring

BY AMBER SINHA

One of the key areas of focus in the **Digital India** mission is the use of Mobile and e-Banking technologies to enable greater financial inclusion.[1] Mobile and e-Banking technologies are apart of a larger shift that is happening in the banking sector - whereby companies are utilizing technology to provide financial services - often through new business models and platforms.

Fintech in India

The Fintech story in India, and in the Global South often is one of financial inclusion and empowerment, and relies on statistics demonstrating an untapped financial market comprising the poor.[2] The Indian fintech software market is forecasted to touch USD 2.4 billion by 2020 from a current USD 1.2 billion, as per NASSCOM.[3] Prominent business areas in the Fintech sector in India include credit-scoring, lending, payments, brokerage and insurance.[4] One of the fast emerging areas of opportunity is credit scoring due to the low penetration of financial services, and the emphasis on financial inclusion by the government.[5] Many of the FinTech business models focused on credit scoring by leveraging alternative credit data in combination with big data analytics to provide services.[6] Examples of emerging credit models include: small ticket unsecured loans, pre-paid plans for single medical procedures, instant point of sale credit, pay per day insurance, and micro-investment.[7]

Fintech Regulation

In this thread, the **Reserve Bank of India (RBI) has considered different aspects of FinTech and potential forms of regulating the same**. For example, the **RBI released a Master Circular** on Mobile Banking transactions in India in July, 2016, which recognized the importance of mobile phones in attaining financial inclusion. Earlier this year in April, 2016, the RBI had also released a Consultation Paper on P2P Lending as well as regulation of Account Aggregators. The RBI also provided an in-principles approval to a number of Fintech companies to set up payment banks.[8] We already see a discussion around the regulatory approaches that the government must adopt to these emerging sectors and what kinds of regulation may be desirable.[9]

EMERGING FINTECH, CREDIT SCORING, AND DIGITAL FINANCE SECTOR[10]

2 billion **INDIVIDUALS** & **200 million** micro, small, and midsize **BUSINESSES** in emerging economies today lack access to savings and credit

Widespread adoption and use of digital finance could increase the GDPs of all emerging economies by **6%** or a total of **$3.7 trillion** by 2025

Digital finance could provide access to **1.6 billion** unbanked people, **more than half of them are women**

An additional **$2.1 trillion** of loans to individuals and small businesses could be made sustainably

Governments could gain **$110 billion** per year by reducing leakage in public spending and tax collection

FINTECH AND CREDIT SCORING IN INDIA

Investment in India's fintech industry grew **282%** between **2013 and 2014**, and reached **US$ 450 million** in 2015[11]

In 2014, more than **400 million** people borrowed money

₹ But, fewer than 1 in 7 were approved for a formal loan[12]

2.4 billion USD market opportunity in India for companies able to take advantage of **big data** and **interoperable cashless transactions**[13]

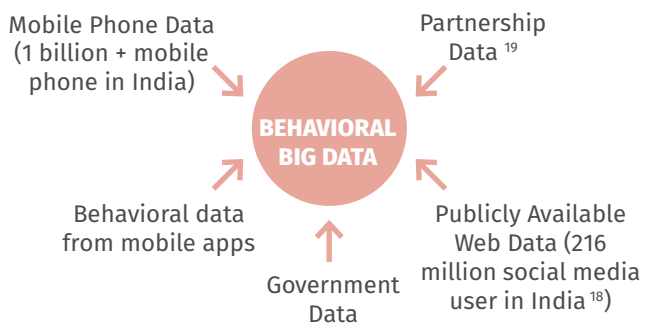
What is Credit Scoring?

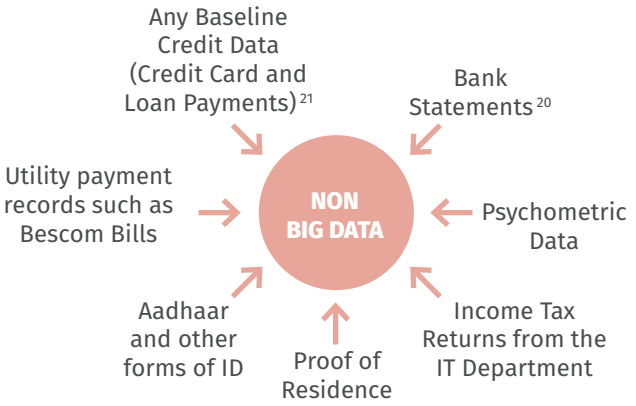
Credit scoring is a method of calculating the potential risk of credit applicants involving use of statistical techniques applied to historical data about the applicants.[14] The outcome of this exercise, the credit score is a summary of an applicant's creditworthiness that informs lending decisions.[15]

BIG DATA ENABLED CREDIT SCORING

With the increased use of information and communication technology, particularly through mobile phone penetration, everyday activities of people leaves behind a much **larger digital footprint which can serve as behavioral data**. This big data phenomenon has also impacted financial institutions and there is a **greater push to move beyond traditional sources of data** for credit scoring and underwriting, as well as use of big data technologies along with the conventional statistical techniques.[16] Big data proponents claim that big data will gives creditors a fuller picture of a consumer and therefore gives a more accurate prediction of the consumer's ability to repay. These practices involve analysis of numerous 'potential credit variables' in a manner that it provides insights about an applicant's creditworthiness. However, credit scoring using alternate data in a regulatory vacuum also poses the risk of bad lending decisions, discriminatory results and mission creep. So far, we have not seen too much movement on regulation of credit scoring companies in the Fintech sector. In 2014, the RBI has produced a report on Data Format for Furnishing of Credit Information[17] but no regulatory steps have followed it. The risks are exacerbated due to the lack of any fair credit underwriting and fair lending regulations in India. Further, unlike in other jurisdictions like the US, there is limited right to access information regarding the data used to generate credit reports.

SOURCES OF DATA





ANALYSIS

SOCIAL PROFILING

CALLS AND CONTACTS [22]

SOCIAL MEDIA DATA

WEB BROWSING DATA (social profile through search, websites visited)

EXAMPLES: An individual whose calls to others are returned may have stronger social connections that allow them to better follow through on entrepreneurial opportunities. Similarly, responsible borrowers may keep their phone topped up to a minimum threshold so they have credit in case of emergency.[23]

FINANCIAL PROFILING

SMS

- Financial profiling based on transactional information
- SMS data has been used a proxy for an individual's discretionary income [24]

BEHAVIORAL ANALYSIS

Behavioral signals from time taken to fill forms, answers revised and paused at, may provide insights which could lead to decisions about the individuals.[25]

Signals used to judge character, truthfulness, need, saving score to inform assessment of intent to repay.

LOCATION BASED PROFILING

Address and domicile information used to profile ethnic background, social strata, etc.

Location data from mobile phones and web browsing data

- Location based marketing,
- Profiles of customers based on their everyday routine through data on residence, workplace, places frequented, routes taken.

How BDCS may work?

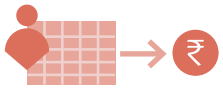
WHAT FINTECH PROMISES

HARI

Hari works at a general provisions store.

No credit opportunities

However, being a member of the unorganised sector, Hari is not a part of the credit ecosystem and has no credit opportunities available to him.



Credit from Fintech firms

The presence of Fintech firms providing small credit based on big data analysis, could potentially allow him access to credit for entrepreneurial opportunities or specific needs for him and his family such as education, housing etc.

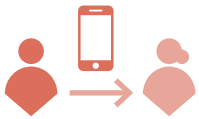
Credit history

Further, once Hari begins building a credit history with the BDSC firms, he could also have an opportunity to avail credit from the banking and financial companies in future.

GROUND REALITY

A Fintech small lending firm may analyse his recent bank account transactions, address, ID, other traditional forms of data, however they would rely on alternate data like mobile phone data to a large extent.

Location data is very useful to see patterns of movement and get an idea of the routine of the person. This will give an idea about whether the person has been consistently employed in the past, how regular he is at his job, whether he has a regular routine. The SMS data can give an idea of purchasing habits which could also speak to creditworthiness. Further his call records and contacts can be used to create his social profile.



However, Hari gave his old cellphone to his wife after using it for the few months. Hari purchase another phone and started using an old SIM lying at home. Often, people have multiple SIMs registered in their name, and used by different family member and friends.

When Hari's wife begins using his phone and SIM, her location data and purchasing habits would give a completely different picture

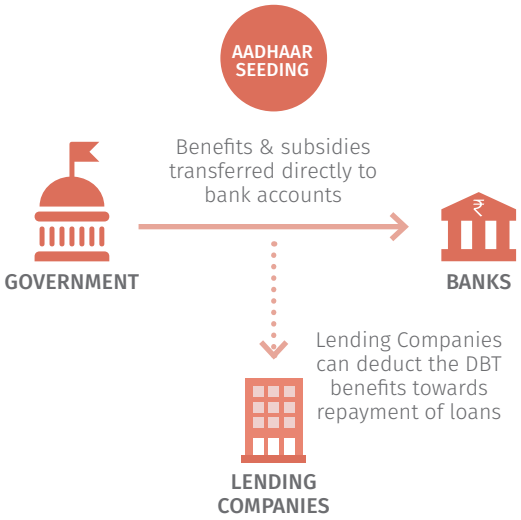
Hari also often takes extra night shifts at a local factory once a week to earn extra cash. Big data credit scoring could possibly correlate his routine travel at night as suspicious activity. Examples of analysis shared by the Fintech industry have shown that this anomaly in his routine could also lead to an assumption that he is travelling to meet to his second wife once every week, thus, adversely affecting his credit score.

No credit for Hari

The DBT Story

A review of big data enabled loan products by the National Consumer Law Centre in the US showed that they were very poor payday loan alternatives. Most of these products involved **annual percentage rates three times higher** than considered non-predatory. Most importantly, most products require **electronic access to the applicant's bank account** or some other arrangement of automatically deducting the the owed amount from the borrower's account. [26]

In India, the government has rolled out the direct benefits transfer scheme (DBT) in various states where **benefits and subsidies are transferred to the bank accounts** seeded with the Aadhaar numbers of the individuals directly in order to do away with the intermediaries involved in the flow of funds, thereby reducing leakages. [27] In the absence of regulations governing lending practices and credit reporting for small loans, lending companies could provide small loans to even those who may not have the capability to pay back loans, as long as **they can deduct the DBT benefits reaching the borrower's bank account towards repayment of loans**.



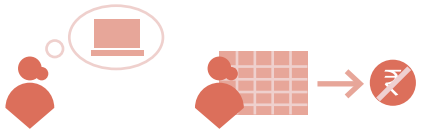
USE OF DBT FOR PAYDAY LOANS

During a **women empowerment campaign** to enable **greater use of information technology by women**, led by the government with the aid of a large IT hardware company and a host of Fintech companies providing small loans using alternate data, there are various **offers for women in rural areas to buy phones and data package**.

1

SEEMA

Seema works in the local post office in a village in Karnataka. She is impressed by the campaign and sees this as an opportunity to build new skill sets.



2

She applies for a loan to buy a laptop, however, an analysis by the Fintech firm does not suggest that she is creditworthy.

3

However, the firm notices that she is eligible to receive subsidies and benefits for rations, LPG, education, electricity and health, in the form of cash in her bank account every month through the DBT scheme. She is granted a loan, as long as the Fintech firm is allowed to deduct the DBT benefits directly from her bank account.

4

As suggested by the credit scoring, she is unable to repay the loans. and the benefits intended for essential services like food and housing are diverted towards consuming non-essential products.

Regulatory Vacuum

An argument often made is that for the Fintech firms to thrive and innovate, a regulatory sandbox for these firms is required so that they are not riddled with unnecessary overheads. Due to the new business models employed by Fintech companies, they often exist in regulatory vacuums where they fall outside the scope of regulatory frameworks governing traditional banking and financial institutions. There have also been calls for a regulatory sandbox for the Fintech sector in India. This would involve controlled limited-scale experiments of financial innovations in controlled environments., starting with a limited size and once there is data on their benefits and harms, formalise the regulatory leeway for the entire sector.[28]

However, these the following factors must be kept in mind:

1. **Opacity of big data**

   Big data enabled credit scoring poses the challenge of opaque algorithms using undisclosed and proprietary methodology which could be used to circumvent fair lending regulations.[29] Even in jurisdictions which provide right of access to citizens to check and verify the credit report,[30] use of credit scoring using big data will prevent them from examining how loan eligibility was determined.[31]

2. **Lack of non-discrimination regulations**

   The are is a lack of non-discrimination regulations in the credit scoring industry in India which prevent. Thus, there are no laws which prevent the firms from collecting data on religion, caste etc. which can be used toward disparate treatment. Even in other jurisdictions, there is a call for Fintech firms to be exempt from equal credit opportunity and fair credit regulations.[32] However, lack of regulations which prevent discriminatory practices are essential for any financial products introduced in the market.

3. **Greater protection required for the poor**

   People who lack the education, information, and other economic, cultural, and social capital that would allow them to take advantage of—and shield themselves against—the free market are most vulnerable and need greater protection

   The consequence of bad decisions are far more dire for those disadvantaged and lacking the resources—financial, psychological, social, and political—to compensate for their errors.

4. **Are de-risking strategies by Fintech firms legitimate?**

   The only way Fintech firms can drive financial inclusion is by 'de-risking' the those who otherwise be considered as risky borrowers. This can be done by collecting consumer data and using it to condition consumer behavior, for instance, through targeted advertising. If there is too much consumer spending on particular products and services, this data can be sold to companies providing these products and services.[33]

Keeping the above in mind, while it may be desirable that small Fintech firms are not saddled with the kind of regulatory overhead costs that traditional banks are subject to, the peculiar problems and abusive practices, such as contacting borrowers family and friends, dissemination of personal information, unauthorized transaction, high interest rates and predatory practices, that we have seen Fintech firms exhibit in their short life suggest a need for some regulation. These should include regulation ensuring fair lending practices and ensuring that the borrowers has the ability to pay back the loan, costs spread evenly across the lifetime of the loan, preventing overcharging of bank accounts and disallowing use of subsidies meant for essential services for repayment.[34]

BENEFITS

Financial inclusion

Use of alternate data can enable the provision of credit to a number of people ignored by banks that use a different set of data points for evaluating whether or not to give a loan, and bring them within the credit system. The employment of constellation of factors to price credit will lead to a reduction in the cost of credit, especially for low income borrowers. [35]

Substantial private sector opportunity

Big data enabled credit scoring presenta a 4-5 billion USD opportunity for the private sector. This will generate more jobs and revenue which will contribute to the GDP of the country. [36]

Data driven efficiency

Credit underwriting using big data has access to a far greater volume of data. This could lead to better underwriting and credit scoring. [37]

Ease of doing business

The possibility of getting loan approvals in a matter of hours or days, as opposed to weeks, and presence of various competing private sector companies offering the services, makes big data enabled credit scoring extremely attractive. [38] Lona applications in India typically require a number of documents such as Identity proof, residence proof, statement of accounts, salary slips, notarised copies of all documents, wet signatures, in-person verification, physical inspections of property etc. [39] A streamlined process that can cut through this paperwork is invaluable.

Personalised services

More granular data can enable firms to create more personalized portfolios for their clients which would be more beneficial for the lenders as well as the borrowers. [40]

HARMS

Use of proxy data for bias

Instead of using discriminating factors such as religion and caste which could be addressed by non-discrimination law, firms could use proxies of such factors such as neighbourhood of residence, purchasing habits. Because big data scores use closed and proprietary algorithm based technologies, it is impossible to analyze them for potential discriminatory impact. There is no law to prevent discrimination on the basis of the disparate impacts of data driven decisionmaking in India. [41]

Lack of data ethics

There is an absence of robust data protection framework and de-identification standards in India to protect the privacy rights of consumers. Fintech firms could clean data and sell inferences for additional revenue to advertising etc.

Imposition of homogenized metrics

Psychometric tests being used in the Global South based on Myer Briggs tests, or the SATs [42] designed in the western countries fails to take into account local contexts. Similarly, lack of grassroots understanding can lead to wrong inferences. For instance, mobile phone data being used to profile an applicant can be often misleading in India as mobile phone are often shared or handed down between members of a family.

Data ideology

A survey of management in companies in the Fintech space in India shows preponderance of those with technology and sales background and little banking/finance background. This suggests an over reliance on data, and a tendency to ignore other kinds of expertise which have been integral to the credit scoring industry.

Regulatory circumvention

Fintech firms engaged in lending practices are not subject to similar regulations as traditional financial institutions and seen as technology companies as opposed to financial companies. While this provides an opportunity to innovate without too many regulatory overheads, it also exempts such firms from regulations concerning confidentiality, data privacy, underwriting and ethical practices.

Mission creep

Practices such as checking credit scores during background verification for employment have been criticised for a long time. However, big data enabled credit scoring provides a far more granular profile dealing different behavioral aspects of a person and big data ecosystem provides more opportunities for credit data to be used for non-credit purposes.

Creation of a financial bubble

Unlike the traditional credit scoring models, the efficacy of the new big data enabled credit scoring models are completely untested. Further, most big data enabled credit decisions are towards loans without any collaterals, thus, increasing the risks for banks and may translate into higher interest rates. This, bad lending decisions without collaterals could lead to a financial bubble similar to the one caused by subprime loans. [43]

ENDNOTES

1. www.narendramodi.in/shri-narendra-modi-shares-his-vision-for-digital-india-5944
2. ssir.org/articles/entry/romanticizing_the_poor
3. India emerging a hub for fintech start-ups, Business Standard website, http://www.business-standard.com/article/companies/india-emerging-a-hub-for-fintech-start-ups-116051700397_1.html
4. https://yourstory.com/2016/07/fintech-startups-2016/
5. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
6. Examples of Indian companies in FIntech credit scoring.
7. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
8. Reliance Industries, Aditya Birla Nuvo, Vodafone, Airtel, Department of Posts, Cholamandalam Distribution Services, Tech Mahindra, National Securities Depository Limited (NSDL), Fino PayTech, Sun Pharma and PayTM were granted in-principles approvals. www.thehindu.com/business/Industry/rbi-grants-inprinciple-nod-for-11-payments-banks/article7557908.ece.
9. www.rbi.org.in/scripts/bs_viewcontent.aspx?Id=3164#4
10. http://www.mckinsey.com/global-themes/employment-and-growth/how-digital-finance-could-boost-growth-in-emerging-economies
11. www.india-briefing.com/news/opportunities-india-emerging-fintech-sector-12477.html/
12. Omidyar Network, Big Data Small Credit, The Digital Revolution and Its Impact on Emerging Market Consumers, available at https://www.omidyar.com/sites/default/files/file_archive/insights/Big%20Data,%20Small%20Credit%20Report%202015/BDSC_Digital%20Final_RV.pdf.
13. https://assets.kpmg.com/content/dam/kpmg/pdf/2016/06/FinTech-new.pdf
14. Hussein A Abdou and John Pointon, Credit Scoring, Statistical Techniques and Evaluation Criteria: A Rreview of the literature, available at: onlinelibrary.wiley.com/doi/10.1002/isaf.325/abstract
15. Upturn Report www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf
16. Billie Anderson and Michael Hardin, Credit Scoring in the Age of Big Data, available at http://www.irma-international.org/viewtitle/107257/.
17. https://rbi.org.in/scripts/PublicationReportDetails.aspx?UrlPage=&ID=763.
18. https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/.
19. Riddhi Mittal, Finomena talk at The Fifth Elephant, 2016, available at https://www.youtube.com/watch?v=ZWlHKJZ3IqY.
20. Omidyar Network, Big Data Small Credit, The Digital Revolution and Its Impact on Emerging Market Consumers, available at https://www.omidyar.com/sites/default/files/file_archive/insights/Big%20Data,%20Small%20Credit%20Report%202015/BDSC_Digital%20Final_RV.pdf.
21. Robinson and Yu, Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace, available at https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf
22. Credit Suisse, India Financials Sector, available at https://www.credit-suisse.com/media/cc/docs/cn/india-digital-banking.pdf
23. Entrepreneurial Finance Lab at Harvard University, SETTLING THE SCORE: A PSYCHOMETRIC ASSESSMENT THAT UNLOCKS CREDIT TO ENTREPRENEURS IN THE MUSLIM WORLD, available at https://www.usaid.gov/div/portfolio/settling-score-psychometric.
24. Supra Note 12.
25. Supra Note 19.
26. http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf
27. http://dbtmis.planningcommission.nic.in/dbtentry/homepage.aspx
28. www.livemint.com/Opinion/dakF6QJNnl23xAVVyOF7qI/Creating-space-for-financial-services.html
29. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
30. The FCRA requires credit bureaus to permit individuals to dispute negative information on their credit reports and to give their own side of the story on disputed information in reports generated for potential creditors, insurers, and employers. It also requires that consumers receive some account of why their credit reports might have led to a denial of credit.
31. U.S. Department of the Treasury, Opportunities and Challenges in Online Marketplace Lending, May 10, 2016, p.20.
32. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
33. http://ftalphaville.ft.com/2015/07/31/2135943/when-financial-inclusion-stands-for-financial-intrusion/
34. The Pew Charitable Trust, Key Findings from Fraud and Abuse Online: Harmful Practices in Internet Payday Lending, October 2014.
35. http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf
36. Supra Note 10.
37. http://www.mckinsey.com/business-functions/risk/our-insights/new-credit-risk-models-for-the-unbanked
38. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
39. http://www.unionbankofindia.co.in/pdf/MudraLoanENGLISHRev180216.pdf.
40. Supra Note 12.
41. Nandan Nilekani interview, available at www.moneycontrol.com/news/economy/how-jam-trinity-will-transform-lending-propel-growth_7099701.html
42. Supra Note 23.
43. http://usf.vc/can-data-become-financial-inclusions-new-currency-and-not-just-in-fintech/

Smart Meters and Big Data

BY VIDUSHI MARDA

Smart Meters are systems that measure and analyse real time energy usage in short intervals and then communicate this information remotely to service providers. Unlike traditional meters used for the supply of electricity, these meters have the ability to collect and transmit data, monitor supply, and also communicate with other appliances. Smart meters empower consumers to understand their energy consumption and manage their usage more efficiently in real time by collecting granular data about consumption and usage activity. For service providers, smart meters enable analysis of this granular data and in turn reveal the best way for industry to optimize processes, redistribute services, and innovate further. As the Internet of Things (IoT) grows, the potential of data mining and collection from smart meters [1] increases correspondingly. [2]

THE SMART METER OPPORTUNITY

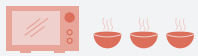
The worldwide installed base of smart electricity meters is projected to be 780 million by 2020 [3], out of which the Asia - Pacific region will account for around 65%. India intends to be "fully smart" by 2021 - 2022, with the Government looking to install a smart meter in every home. [4]

Towards this end, in 2010, the India Smart Grid Task Force (ISGTF) and the India Smart Grid Forum (ISGF) were set up to advise the Ministry of Power on policies, programmes and developments with respect to smart grids and also develop a smart grid roadmap for India. Following this, in August 2013, the Government of India notified its "Smart Grid Vision and Roadmap for India" which seeks to "*enable the development and deployment of an Indian Smart Grid model*", and also paved the way for the launch of the National Smart Grid Mission (NSGM) which would have "*its own resources, authority, functional & financial autonomy to plan and monitor implementation of the policies and programmes prescribed in the roadmap*". [5]

One of the most interesting initiatives so far has been the 14 Smart Grid pilot projects commissioned by the Ministry of Power in 2013, where the Government covers 50% of project cost. [6] The first trial was conducted in Puducherry, [7] where the state Government has recently signed a 46 crore deal with Chinese firm DongFan electronics for a smart meter project. [8] Ericsson, the Swedish telecom firm, also signed a deal with the government of Assam in early 2016 to install 15,000 smart meters over the next three years. [9] A model Smart Grid Regulation was developed by the Forum of Regulators in 2015, [10] and the Government also launched the Ujwal Discom Assurance Yojana (UDAY) to increase efficiency of electricity distribution, where one of the commitments is to make installation of smart meters compulsory for consumers using more than 200 units of electricity a month. [11]

These meters have the potential to increase energy saving, decrease energy spending and provide insight into energy consumption; i.e, they increase efficiency at all levels of the energy supply chain through the power of big data and analysis. However, the data that drives this efficiency also poses a number of big data risks and concerns because of the ability of these meters to construct a detailed map of individual patterns, that do not remain confined to just data on energy usage.

DATA COLLECTED IN REAL TIME



How often does an individual consume microwaved dinners?

How many times a week does an individual cook a meal?

How many hours of TV does an individual watch?

What kind of TV does the individual own?

When does the individual shower, and for how long?

How long does an individual take to leave the house after a shower?

How many hours does an individual sleep?

How often is the person home?

What TV program was being watched, and if a DVD movie being played had copyright-protected material

Worldwide installed base of smart electricity meters is projected to be **780 million** by 2020

Out of which the Asia-Pacific region will account for around **65%**

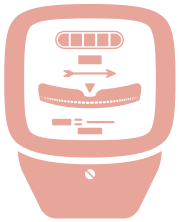
INDIA
Smart meter in every home by **2021-2022**

14
SMART GRID PROJECTS commissioned by the Ministry of Power in 2013

Unintended Consequences of Smart Meters

ANAND

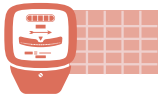
Anand installs a smart electricity meter in his home and looks forward to managing his energy consumption closely and lowering his costs.



After installation, Anand now has a better idea of how he can improve his usage. He finds that he is more careful with energy usage due to real time monitoring, and his bills are lower too.

After a few months, Anand approaches X Insurance to enrol for a health insurance plan.

Once the company has assessed his application, they inform him that they will need to charge him a higher premium than the standard rate.



This is because X Insurance bought consumer data from Anand's smart meter utility.

1.5 hours

Due to appliance load signatures, the smart meter data showed that Anand had at least 2 microwave meals a day and spent at least 1.5 hours watching TV.

According to X Insurance, this made him susceptible to health complications and thus liable to pay a higher premium.

Regulatory Response

Deregulation for innovation

Smart meters in India present an opportunity to harness the potential of big data in providing electricity, and in services like transportation. While the promise of efficiency and conservation is reassuring, these meters work with large amounts of data and information that could lead to an invasion of privacy without adequate oversight. Currently, there is no such oversight as the data collected from smart meters reaches units (namely, Data Concentrator Unit followed by the Meter Data Acquisition System and Meter Data Management System)[12] which are components of the Advanced Metering Infrastructure (AMI).[13] The ISGTF and ISGF under the Ministry of Power manage the AMI and therefore have access to the data collected by the smart meters. Section 43A of the IT Act applies to bodies which includes firms, companies or any association of individuals who carry on any commercial or professional activity.[14] The ISGTF and ISGF cannot be categorized into any of the bodies provided for by the Section.

Regulation for public interest

Granular data measurements are capable of revealing intimate information about consumers that choose to use them. This data is of great value not only for the effective running of these meters, but also has economic and societal repercussions. While the regulatory response should account for protection of individual's privacy, there is a school of thought that believes that excessive privacy protection would be a case of over-regulation; i.e would limit the benefits that can be untapped from smart meters.[15] To avoid over-regulation, a possible model would be to build a marketplace in which data analysts interested in obtaining access to data work with third party service providers who look to learn about consumer behavior and work with the data aggregator or the utility company which is the marketmaker. Providing consumers with information on potential data that can be derived from their activity, and also understanding the potential risk to privacy could determine a user's payoff in revealing this information.[16] Another issue to consider is where the control over such information lies - with the consumer, the utility or the service provider.[17]

Regulation to encourage technical measures that mitigate harm

A major concern and big data harm from smart meters is the ability to construct a detailed map of individual energy consumption by analysing application load signatures. It is possible to hide these appliance load signatures through rechargeable batteries,[18] power management models, power mixing algorithms etc.[19]

Regulation for interoperability

With variations in standards of communications protocols and meters, there is an issue of interoperability between devices. Regulation must ensure the standardization of this to facilitate better information exchange.This is already a concern amongst manufacturers,[20] and the regulatory response must account for this.

BENEFITS

Efficient Demand Side Management (DSM)

Data collected can help in using and providing electricity more efficiently. The aggregate technical & commercial losses are around 27-28% in India's power distribution segment, and smart meters have the potential to reduce these network losses.[21]

Accurate shaping of the market and industry

Smart meters provide data on how much excess energy will be available, whether the grid can transmit it and in turn, whether to sell it. Smart meter data can also help determine when and where equipment downtime and power failures are most likely to occur, helping in better load management and allocation of resources.

Security of supply

Smart meters enable usage of predictive models capable of applying knowledge acquired through the grid to improve daily performance. This means a more steady supply of electricity or more specifically, security of supply.[22]

HARMS

Unintended behavioral analysis tool

Smart meters will be able to recognise appliance load signatures to make a detailed map of an individual's energy usage. This combined with plug in hybrid vehicles with individuals taking this electricity usage outside their home, an electricity usage profile can soon become a behavioral analysis tool, that consumers do not intend to create by simply installing a meter.[23] Based on the data that can be collected by these meters, it is possible to determine a person's health (how many microwaved meals are consumed? How often is a home cooked meal made? How long is the person using the computer/television), work ethic (what time does she leave home? What time is she back? How much time is spent in front of the TV?), social life (how often is the person home on the weekend? Is it more or less than weekdays? In periods when the person is alone is there significantly more consumption of energy or less?). Hackers have also been able to identify "the number of PCs or LCD TVs in a home, what TV program was being watched, and if a DVD movie being played had copyright-protected material."[24] This information makes owners of smart meters vulnerable to theft, marketers, law enforcement, creditors, insurance companies etc.[25] This technological invention runs parallel to the functioning of the SCADA technology which collects data for the purposes of control and supervisory management.[26]

Social polarisation

In the UK, it was found that while a traditional meter costs 7-8 pounds plus installation costs, the various types of smart meters cost between 37-80 pounds plus data infrastructure, installation costs, customer display etc.[27] Affording initial investment in smart meters is a luxury that all households do not have. Some households will be able to signal their usage and consumption patterns to the market, while those households who cannot afford the initial outlay will not be considered. This discrimination harm will lead to an invisible, gradual polarisation of those households that are not yet consumers.

Social dumping

Serving marginal, low income and poor households consume small amounts of energy and are often associated with non-payment of bills, debt and disconnection, all of which raise significant costs for utilities.[28] Increasingly, with pre-payment option being adopted as default in newer meters, such households are presented with an option to self-disconnect due to inability to pay bills. This system transmutes the problem of disconnection (easily measurable by regulators and one of the parameters often used to judge the health of a market) to one of poor access to energy services. The number of disconnections will decrease, but these will not be representative of the increase in cases of switching off voluntarily by low income households,[29] making access to energy services an invisible issue.

ENDNOTES

1. http://www.ptc.com/product-lifecycle-report/iot-reduces-the-cost-and-complexity-of-smart-electric-meters
2. https://www.theguardian.com/sustainable-business/2015/mar/16/internet-of-things-energy-costs
3. http://tdworld.com/ami/smart-electricity-meters-total-780-million-2020-driven-china-s-roll-out
4. We should look at a smart meter in every home: Piyush Goyal, The Hindu BusinessLine, http://www.thehindubusinessline.com/specials/clean-tech/we-should-look-at-a-smart-meter-in-every-home-piyush-goyal/article8245384.ece, Feb 16, 2016.
5. http://www.cyantechnology.com/wp-content/uploads/2014/04/Smart-Grid-Vision-and-Roadmap-for-India.pdf, Page 3.
6. https://www.metering.com/14-smart-grid-pilot-projects-shortlisted-in-india/
7. https://www.metering.com/india-puducherry-smart-grid-pilot-to-be-expanded-to-87000-meters/
8. http://www.deccanchronicle.com/nation/politics/070716/puducherry-government-in-tie-up-with-dongfan-electronics-for-smart-grid-project.html
9. http://www.business-standard.com/article/companies/ericsson-to-deploy-15-000-smart-meters-in-assam-116041300896_1.html
10. http://www.indiasmartgrid.org/reports/FOR%20-%20Smart%20grid%20regulations.pdf
11. http://pib.nic.in/newsite/PrintRelease.aspx?relid=130261
12. http://www.isgw.in/wp-content/uploads/2016/03/ISGW-2016_Surendra-Acharya_Smart-Meter-to-Smart-Hubs---AMI-Rollout-Strategies-for-India_17-March-2016.pdf
13. http://www.indiasmartgrid.org/Advanced-Metering-Infrastructure.php
14. Section 43A, IT Act.
15. Smart Meters Big Data: Game Theoretic Model for Fair Data Sharing in Deregulated Smart Grids, IEEE Access Volume 3, Abdulsalam Yassine, Ali Asghar Nazari Shirehjini, Shervin Shirmohammadi, December 2015. http://ieeexplore.ieee.org/document/7342880/
16. Smart Meters Big Data: Game Theoretic Model for Fair Data Sharing in Deregulated Smart Grids, IEEE Access Volume 3, Abdulsalam Yassine, Ali Asghar Nazari Shirehjini, Shervin Shirmohammadi, December 2015. http://ieeexplore.ieee.org/document/7342880/
17. Smart Metering and Privacy: Existing Laws and Competing Policies, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1462285&download=yes.
18. Smart Meter Privacy Using a Rechargeable Battery: Minimizing the Rate of Information Leakage, 2011 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP), <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5916934> David Varodayan and Ashish Khisti, 12 July 2011, http://ieeexplore.ieee.org/document/5946886/
19. Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures, 2010 First IEEE International Conference on <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5619950> Smart Grid Communications (SmartGridComm), <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5619950> Georgios Kalogridis <http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Georgios%20Kalogridis.QT.&newsearch=true>et al, 4 November, 2011, http://ieeexplore.ieee.org/document/5622047/
20. http://www.vibrantgujarat.com/writereaddata/images/pdf/project-profiles/Smart-Meters-Smart-Cards-and-Smart-Lighting.pdf, Page 22.
21. http://www.vibrantgujarat.com/writereaddata/images/pdf/project-profiles/Smart-Meters-Smart-Cards-and-Smart-Lighting.pdf, Page 13.
22. https://www.ofgem.gov.uk/sites/default/files/docs/2006/03/13383-sustainability_first_response__0.pdf
23. http://www.santepublique-editions.fr/objects/smart-meter-problems-dec-2012-final-voir-p-28.pdf
24. http://www.networkworld.com/community/node/79486
25. http://www.santepublique-editions.fr/objects/smart-meter-problems-dec-2012-final-voir-p-28.pdf, Page 22.
26. Towards Privacy Protection in Smart Grid, https://www.nics.uma.es/pub/papers/1750.pdf
27. https://www.ofgem.gov.uk/sites/default/files/docs/2006/03/13383-sustainability_first_response__0.pdf, Page 23-25.
28. Pathways to "Smarter" Utility Meters: The Socio-Technical Shaping of New Metering Technologies, http://www.ncl.ac.uk/guru/assets/documents/ewp23.pdf
29. https://www.ofgem.gov.uk/sites/default/files/docs/2006/03/13383-sustainability_first_response__0.pdf

Intelligent Transport System

BY VANYA RAKESH

Intelligent Transport Systems (ITS) is a term for the integrated application of communications, control and information processing technologies to monitor and manage a transportation network,[1] which enables the gathering of data and intelligence, which is then analyzed and the derived insights are shared back to traffic managers and road-users.[2]

Such systems include stand-alone applications as well as cooperative ITS applications involving vehicle to infrastructure or vehicle to vehicle communications.[3]

The main objective of ITS is to improve commuter service as well as decision making, achieve traffic efficiency, enhance safety for commuters, drivers, etc., conserve energy and enhance overall performance and profitability of the service provider corporation.[4]

The components of ITS usually comprise of network of sensors, connected vehicles, GPS tracked public transportation, dynamic traffic lights, passenger information panels, automatic number plate readers, CCTV systems, navigation facilities, etc.[5]

In India, some of the ITS projects have been initiated in cities like Mumbai, Bangalore, Ahmedabad, Pune, Mysore[6] and states like Rajasthan, Maharashtra, Andhra Pradesh and Uttar Pradesh.

As per a World Bank study, by 2031, some **600 million** people are expected to live in India's cities

According to Census 2011, there are a total of **497 cities** in India, with around **90 cities** with a population over **5 lakh**[7]

However, only about **20 Indian cities** with populations over **500,000** have any kind of **organized public transport systems.**[8]

As of November 2015, the **current urban bus occupancy** in various Indian State Transport Units (STUs) is **<55%** and is reducing every year.

REASONS
Unavailability of buses
Long waiting time
Operational inefficiencies[9]

World Bank reports that the economic losses incurred on account of **congestion and poor roads** alone run as high as **$6 billion** in India[10]

Sources of Big Data in Transport

Vehicle Tracking System

Comprises of VTU and Voice-kit. The VTU shares real-time location information, vehicle and shift number (of the driver and conductor of the bus concerned). This Unit transmits data via mobile (GPRS) network, and can also be used by the driver to inform the ITS Control Room in case of an emergency.

Passenger Information System

This is the bus location information communication system, powered by the real-time data being generated by the VTS. It helps provide information to the passenger about buses going from a particular stop/location, the destination, the route via which the bus will go, the estimated time of arrival, etc. This information is generated by integrating information from the VTU installed in all buses.

Mobile Application

The VTU enables sharing of location of buses to show the estimated time of arrival (ETA) of a bus, along with bus routes, bus number, and destination of the bus.

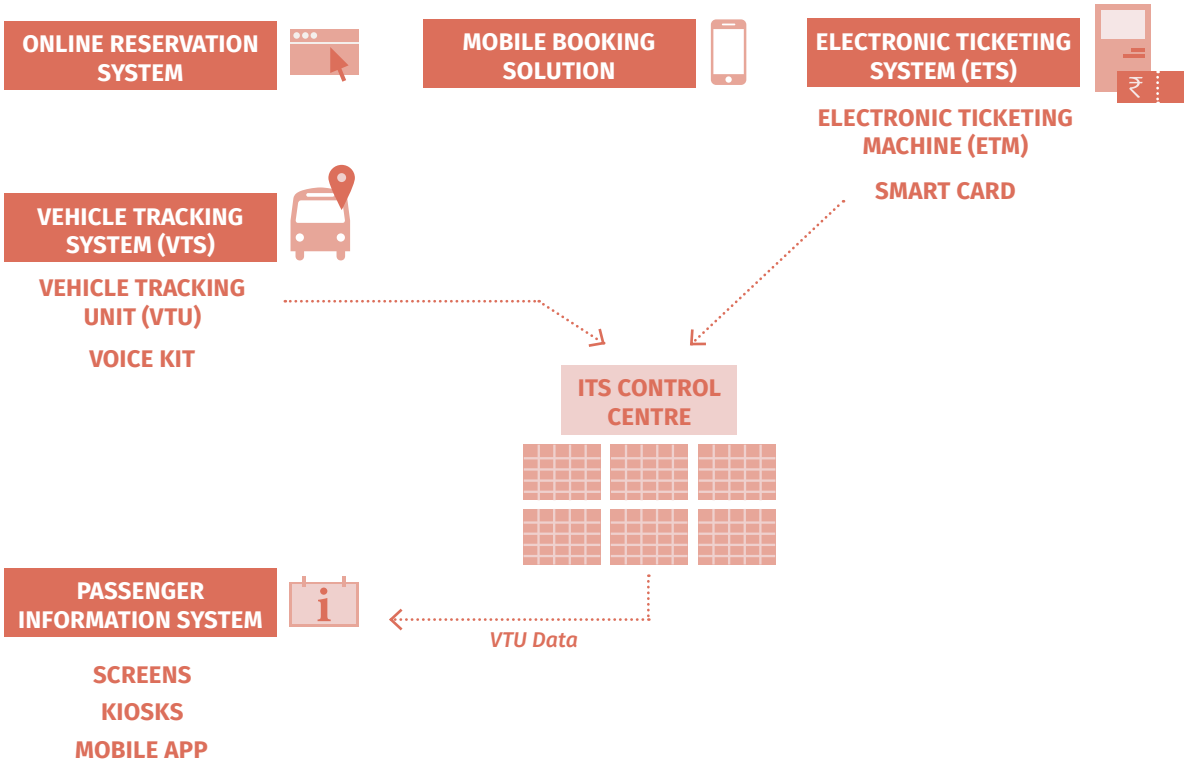
Electronic Ticketing System

Comprises of the Electronic Ticketing Machines (ETM) and Depot Application. The ETM transmits ticket data to ITS via mobile (GPRS) network, and includes information regarding the bus stop at which a particular ticket was issued, the ticket amount, details of the bus, timestamp, etc. Depot application is installed at all bus depots, which generates information on duty rota, the log sheet for the driver, kilometres performed and details about fuel usage, etc.

Call Data Records (CDR)

These records provide the location, time and duration of every call, generating huge databases about daily transactions performed by ordinary people. Every time a call is made on a mobile phone, the mobile tower associated with that call – usually the nearest mobile tower – is recorded, originally for billing purposes.

Overview of the ITS Infrastructure



DATA COLLECTION IN ITS

Karnataka (BMTC)

App collecting PI - location, media, photos, files, device ID and call information

Rajasthan (RSRTC)

App collecting PI - location, device ID and call information

Uttar Pradesh (UPSRTC)

App collecting PI - SMS, Camera, wi-fi connection information, device ID and call information

Maharashtra (MSRTC)

App collecting PI - SMS, Camera, wi-fi connection information, device ID and call information

Andhra Pradesh (APSRTC)

The mobile app says that for download, it does not require any special permissions. However, for tracking a bus a separate app needs to be downloaded, which requires access to an individual's PI.

Potential Users of Data

DATA →

- PASSENGERS
- URBAN PLANNERS
- TRAFFIC POLICE
- TRANSPORTATION COMPANIES
- TRIP-PLANNING WEB-SITES AND PORTALS
- URBAN LOCAL DEVELOPMENT BODIES

Comparison of Ongoing ITS Projects

Fleet size		Total Project Cost
9,500 Buses	UTTAR PRADESH STATE ROAD TRANSPORT CORPORATION (UPSRTC) [11]	Rs. 38.25 crores [12]
16,500 Buses	MAHARASHTRA STATE ROAD TRANSPORT CORPORATION (MSRTC) [13]	Rs. 25 crores
6,404 Buses	BANGALORE METROPOLITAN TRANSPORT CORPORATION (BMTC)	Rs. 79 crores [15]

In the year 2009, a sum of Rs 25 crore was invested in Trimax, which was utilised for MSRTC project. [14]

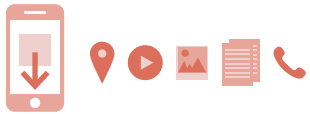
ITS Use Scenarios

PRIVACY

1

RAMA

Rama, a UG student in the city recently moved to Bangalore and started using BMTC bus services.



2

She tries to download the app to plan her trips in the city, which asks for access to personal information like PI - location, media, photos, files, device ID and call information.

No opt out

To download and use the ITS Mobile Apps for buses in India, one needs to allow access to information in one's mobile phone. Also, for making an online booking, an individual needs to mandatorily register and create an account for which various forms of PI are required to be submitted.

3

Being unaware of the consequences, she allows access.

Being a regular user of the BMTC App, the BMTC servers have details regarding her journey, location, her personal files from her phone, which makes it easier to track her and reveal personal details.

4

In absence of security measures/privacy policies for protection of such sensitive data, the servers might be hacked, revealing personal information online, rendering such data vulnerable to misuse.

Though 4 of these initiatives have a privacy policy available on the website, it does not provide for a comprehensive regulatory framework regarding collection, use, storage of data. Also, BMTC has no privacy policy. In case of Uttar Pradesh, Andhra Pradesh, Maharashtra and Rajasthan, the privacy policies do not clarify how they would collect, share or use the information collected through the mobile applications.

5

Also, the app only displays bus and route number, along with ETA of a bus (instead of the destination or the area via which that bus would go). Being new to the city, she is not yet familiar with bus numbers and their routes. As a result, she regularly spends extended periods of time at the stops to enquire about the correct buses, defeating the objective of PIS.

EXCLUSION

SHANTI

Shanti is a 65 year old woman working in the unorganised sector in Bangalore. She lives in Adugodi and works in K R Puram to which she takes a bus daily.

The fact that Shanti is an illiterate woman who does not use a smartphone and the BMTC Mobile App, renders her vulnerable to exclusion.

As data is increasingly being used to make decisions, passengers who leave less of a data trail may be under-represented in planning and decision making, which might impose a barrier for those with limited digital literacy.

INCIDENT MANAGEMENT

DEPOT MANAGER

With the introduction of ITS, the Depot Manager at Yelahanka bus depot in Bangalore has been able to track the location, speed and schedule of all buses that go from the depot.

One day, a bus carrying 60 passengers from Yelahanka to Koramangala broke down in Hebbal at 9pm in the night. The depot manager could instantly track the bus through the VTU and contacted the driver via the voice kit installed in the bus to know the exact location and the reason for the mishap. The information was instantly communicated to the ITS Control Room and without delay, BMTC managed to send another bus to the location and continue the trip to Koramangala.

BENEFITS FOR ORGANISATION

Reduce project costs

Big data presents opportunities to identify problems, analyse and reduce project costs. ITS intends to help in identification of pilferage, improve revenue management, and reduced fuel consumption.

Incident management

Tracking location of buses and passengers enables preventative maintenance and avert potential threat, whereas the Voice kit enables instant communication to inform about mishaps.

Promotes reliability on transports

Increased efficiency and improved forecasting promotes reliability on transports as reduced waiting time for passengers helps in increasing the passenger base.

Reduce traffic congestion

Providing actual data for route planning and route rationalization by way of VTUs can help in reducing traffic congestion on roads.

BENEFITS FOR CONSUMERS/SOCIETY

Improved user experience

Components like ETS help in demand prediction, which helps in improving user experience by increasing fleet of buses in areas where required.

Targeted services

Recognizing traffic patterns by investigating real time data can help transport authorities to understand commuters' behaviour, provide targeted information and identify policy interventions.

Reduction in traffic congestion in cities

Through big data, the smart city will be able to reduce traffic and accidents by opening new roads, enhancing the infrastructure based on congestion data, and collecting information on car parking and alternative roads.

PIS

PIS improves end-to-end customer experiences by enabling trip planning and saving time.

HARMS

Privacy and data security

Over-collection and unclear use of PII: Without clear purpose and over-collection of personal information about an individual for allowing them the use of mobile apps to help them get information about the location and route of a bus violates the principle of data minimization and poses a risk of identification.

No or inadequate Privacy Policies: In this case, no or inadequate privacy policies or established principles and guidelines to regulate ICT and big data usage reflects inadequate data security measures, which may lead to identification of individuals.

No opt-out: The compulsory provision of data as a prerequisite for the access and use of many key online services is making opting-out of data collection impossible, highlighting need for enacting meaningful consent from citizens and consumers. [16]

Unplanned use of data

Lack of clarity about the plans to utilise the big data potential of their project, even when the project has been implemented across the city and data collection has begun, like in case of BMTC, reflects planning and designing without clear consideration of the impact of big data in such projects.

Lack of accountability and transparency

In case the data is compromised accidently, lack of publically available information (like in case of BMTC) regarding data ownership and management may make it difficult to hold an entity in a PPP model accountable. This also reflects lack of transparency.

Data quality

Data captured by different systems and people and stored in distinctive databases can result in inaccurate data or poor data quality as collection and processing methods may vary, which may lead to unreliable analysis and decisions, further leading to exclusion and bias.

Exclusion

Though civic apps help may increase engagement of passengers and make the system more efficient, their capacity to bridge social divides remains unclear. The potential of Smartphone apps to address social problems of transport management should not be overstated. [17]

ENDNOTES

1. http://www.ectri.org/YRS07/Papiers/Session-15/Efimenko.pdf
2. http://meity.gov.in/content/intelligent-transportation-system-its
3. http://pearl.niua.org/sites/default/files/books/GP-IN1_UT.pdf
4. http://www.rakshakfoundation.org/wp-content/uploads/2015/01/2014_D141.pdf
5. http://unctad.org/meetings/en/SessionalDocuments/CSTD_2015_Issuespaper_Theme1_SmartCitiesandInfra_en.pdf
6. http://pearl.niua.org/sites/default/files/books/GP-IN1_UT.pdf
7. http://censusindia.gov.in/2011-prov-results/paper2/data_files/India2/Table_2_PR_Cities_1Lakh_and_Above.pdf
8. http://articles.economictimes.indiatimes.com/2015-09-02/news/66144023_1_smart-cities-public-transportation-bicycle-sharing-system/2
9. http://www.kpit.com/resources/downloads/whitepapers/realizing-smart-cities-technology-whitepaper.pdf
10. https://coeut.iitm.ac.in/ITS_synthesis.pdf
11. http://www.upsrtc.com/
12. http://telematicswire.net/an-overview-of-upsrtc-intelligent-transport-management-system-itms/#sf3vGS0lHlzrpi3x.99
13. https://www.msrtc.gov.in/
14. http://www.moneycontrol.com/news/business/trimax-may-go-public-to-raise-rs-250-300cr-this-fiscal_557585.html?utm_source=ref_article
15. http://bangalore.citizenmatters.in/articles/bmtc-new-year-gift-to-bengaluru-intelligent-transport-system
16. http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf
17. https://articles.extension.org/sites/default/files/Smart%20Phones%20and%20Immigrant%20Integration.pdf

Detailed Narratives of Case Studies

Predictive Policing

BY AMBER SINHA

CCTNS: e-Governance for Police

The Crime and Criminal Network Tracking System (CCTNS) is an e-governance project under the Digital India mission which seeks to use ICT for better provision of citizen-centric services, connect approximately 14000 police stations across the country and facilitate investigation, detection and prevention of crime. The Ministry of Home Affairs conceptualized the CCTNS project as a Mission Mode Project [1] under the National e-Governance Plan (NeGP), and like a number of other e-governance projects, it was later included in the Digital India mission. The primary goal of the scheme is to facilitate tracking of criminals by digitizing records of the police, prisons, courts, forensic labs, and prosecution records through the creation of a single database. Giving states access to the databases of other states will help in tracing an accused or missing person, lost vehicle, etc. It also aims at tracking and easing the process of traffic and resource management through the use of technology.

While the objective of the scheme is primarily to facilitate the functioning of the criminal system through online records, it is also proposed that these records be analyzed for the purpose of trend setting, crime analysis, disaster and traffic management, etc. Towards these purposes, various states have decided to use predictive policing techniques, [2] and plan on leveraging the idea is to build on the already existing structured data [3] collected from established legacy electronic data based on geographic locations and the nature of crimes in locations, and databases of history sheeters and police reports with other alternative data points. [4] While predictive crime analysis is a major objective of the Centre, so far most states have not reached this stage as data migration and data digitization is still in progress. The scheme seeks to link the databases across the country for identification and tracking of criminals, maintaining a record of suspected and missing persons, missing vehicles, etc. It also will facilitate the identification and police verification processes by linking of databases to the UID, NPR and Electoral Roll databases. [5] The linking of the databases under the CCTNS Scheme will also create a dynamic repository of information that could be used for big data analysis like predictive policing methods.

What is Predictive Policing?

Predictive Policing is the application of Big data analytical techniques to identify targets for prevention of crimes through police intervention or use of statistical predictions to solve crimes. [6] A RAND Report identifies the following fours ways in which Predictive Policing usually works —

a. predicting places and times with an increased risk of crime,

b. predicting potential future offenders,

c. creation of profiles for past crimes, and

d. predicting groups of individuals likely to be victims of crimes. [7]

Predictive policing draws from canonical theories of crime that focus on criminal events, crime-prone locations, and criminal opportunities. [8] The basic underlying assumption of predictive policing is that crime is not randomly distributed across people or places. [9] Rather, patterns of crime are a "function of environmental factors that create vulnerabilities for victims and spaces at certain times." [10] Therefore, instead of relying on a police officer's hunch, predictive policing seeks to rely on algorithmic insights derived using the power of 'big data'. Further, the use of algorithmic technologies also holds the ostensible promise of reducing the biases that exist in human decision-making. As a governance measure, predictive policing is reflective of a global trend towards data-driven decision-making, and its introduction in India sits well with recent efforts to make India a data-rich jurisdiction. [11]

How Predictive Policing Works

Data Collection

In the first step, data is collected from different sources — crime data from police station databases, environmental data including crime seasonal patterns, neighborhood composition, call data records and other mobile phone data. The CCTNS project intends to bring together data from various sources such as legacy electronic data, data available in the existing registers, reports, case files, (Physical Copies), Images and pictures within the case files, FIR, Charge Sheet, Seizure Memo, Conviction Memo, Arrest Memo, Crime (case/incident) data, criminals' data, the data from the police stations records rooms (from police registers), citizen complaint information. [12] An important source of information would be the Common Integrated Police Application [13] which was developed to be installed in police stations and to support the crime reporting, investigation and prosecution functions. [14]

Data Analysis

The next step is the Analysis, where the data collected is analysed based on a predictive method. Examples could include near-repeat theory,[15] social network analysis[16] and regression models using risk factors that may provide insights about crime in a particular location. These trends are used by software like PredPol to inform its models to identify hot-spots in an area.[17] For instance, if the analysis provides insights about the cause and patterns of crime such as seasonality and trends, then the algorithms identifying hot spots in an area can focus more on those causes and patterns. Place based theories focus on the vulnerabilities in the location as the factor leading to criminal activity.[18] For instance, PredPol factors in historical crime data from the police department and produces predictions on where and when a crime is most likely to occur, dividing the jurisdiction into 500 feet by 500 feet boxes in a map. On the other hand, event based theories focus like the routine activities theory suggest that crime is likely to occur where either offenders seem to converge or opportunities exist.[19] Similarly, risk assessment for individual criminal behavior try to assess the likelihood of short term term circumstances like losing one's employment, alcohol and drug abuse leading to criminal activity.[20] The CCTNS project also intends to leverage existing criminal information systems. For instance, the Crime and Criminal Information System (CCIS) facilitates statistical analysis of crime and criminals related information at the state level.[21]

Intervention and Response

The next step is making police interventions in response to the insights, such as deploying more forces in a neighborhood. Interventions can include deployment of more forces in a neighbourhood in anticipation of an increased risk, more surveillance of likely offenders or identification of people or location specific problems leading to crimes and measures to correct them.[22] It is noteworthy that the very act of predictive policing also creates new data.[23] The final step is to account for the criminal response to predictive policing and includes cases like displacement of crime to another area.

Status of the CCTNS Project

The CCTNS Project initiated by the Central Government had failed to meet its initial target in 2012. The current NDA Government has set a target for completion by March 2017 and another five year period till 2022 has been given for its operations and management.[24] As of November 2015, only 78% of the police stations are entering 100% FIRs through the CCTNS software.[25] NKN Connectivity for SDC-NDC has been implemented in 27 out of 35 States and Data Replication at NDC has started in 21 out of 25 States.[26]

There are several problems with the implementation as follows:[27]

1. Police, Law and Order is a state subject in India; and systems across States are not uniform.
2. Many of the states have delayed the selection of SIs. Two states—Bihar and Rajasthan are still in search mode after their flip-flop with the existing SIs.
3. There has been delay in establishment of the National Data Centre.
4. CCTNS is a centrally funded scheme that is to be implemented by States. Many States have stalled their implementation due to the lack of adequate funds.
5. The Core Application Software (CAS) provided by NCRB which has proved to be full of errors and crash-prone.
6. There is delay in data digitisation by SIs, hand holding resources are either of poor quality or yet to be deployed, payment for SIs and SPMUs are pending and there is serious deficiency in different version of CAS.

Predictive Policing in India - Contextual Factors

In light of the above factors, the move to opt for predictive policing solutions is odd. The status of digitisation of existing police records is far from complete and attempts to use such data for predictive analytics may lead to incorrect conclusions, causing not only inefficient use of state resources but also unfair discrimination against the rights of those targeted as a result of such error prone analysis. Further, the following implicit biases in the existing police data can prove to be highly problematic for any efforts towards data-driven decision-making in policing.

The Unfairness of 'History Sheets'

Since colonial times, the law enforcement agencies in India has had a practice of maintaining databases of 'history-sheeters", individuals accused, but not necessarily convicted of a crime, and put on police surveillance lists.[28] Mrinal Satish has shown the certain categories of persons continue to be under continual surveillance by the police. These include 'criminal tribes', a racist epithet for classes of people considered genetically criminal and habituated to a life of crime.[29] Aside from this, police in various states in India also maintains databases of individuals classified as "goondas" and "bad characters." These "history sheets" translate into constant supervisions or surveillance of individuals and often leads to selective application of laws against them.

Lack of Exclusionary Rules in India

Studies have demonstrated that courts very rarely refuse to include evidence that has been obtained in violation of provisions of the Code of Criminal Procedure in India. A perusal of judicial rulings in India demonstrates that illegally obtained evidence is admit regularly in Indian courts. In a landmark case, the Supreme Court of India has held that evidence obtained through illegal searches is not, by itself, vitiated but the court may be inclined to review it carefully.[30] Even more alarming is the legal position on warrantless arrests where the courts have held that police officers are not accountable for the discretion of arriving at the conclusion of reasonable suspicion while arresting a suspect.[31] The lack of these protections make it harder to hold police accountable for excessive or unlawful use of predictive policing methods. The promise of efficiency offered by predictive policing must be weighed against the competing interest of due process and rights of the accused. The courts in India have not historically given enough consideration to these competing interests, which magnifies the dangers of predictive policing.[32]

'Reasonable Grounds' and 'Reasonable Suspicion'

While the legal implications of predictive policing methods are far from clear, one aspect of criminal procedure law that it will definitely impact is the manner in which we construe 'reasonable suspicion'. Andrew Ferguson has analysed this in the context of areas where the police already uses 'prediction' while arriving at 'probable cause' or 'reasonable suspicion,' and see how predictive policing may reconfigure such concepts. In India, upon reasonable grounds, the police is allowed to conduct a warrantless search.[33] While this practice is common in other jurisdictions also, in India, for cognizable offences,[34] the police can arrest a person without warrant as well.[35]

The standard for 'reasonable suspicion' was articulated well by the Supreme Court when it held that there must be a "reason to believe that such an offence... has been committed and, therefore, an arrest or search was necessary." Further, the police must have "acted in a bona fide manner" as borne out by the facts and circumstances of the case,[36] and not merely subject to the subjective satisfaction of the officer.[37] Therefore, the requirement of reasonable suspicion demands acting in good faith. However, the Indian courts stop short of elaborating on this standard in terms of how factors such as informant's tips,[38] and profiling, which involve some level of prediction, measure against it. Therefore, use of predictive policing methods, not only give greater powers to the police to arrest people without warrants but may involve a judicial re-examination of the standard of 'reasonable suspicion.'

Benefits of Predictive Policing

Better Allocation of Resources

In India, the police force is overburdened with work leading of health and social issues.[39] Police officials usually work seven days a week and often have to attend to very long shifts. Therefore, any technology and system of policing which enables a more efficient allocation of resources is extremely desirable. Through hot-spots, risk terrain analysis and near-repeat theories, predictive policing promises analysis of data which is more easily available that could enable the law enforcement agencies to identify individuals and locations for targeted policing.[40]

Preventive Policing

Predictive Policing offers the opportunity to the law enforcement agencies to preemptively act against predicted crimes by focussing on crime-prone areas and individuals at the risk of offending or being targeted. If crimes can be stopped before they are committed, it has great social and economic value not just for those at the risk of being victims of such crimes, but also for the offenders, as they can be stopped from making life altering mistakes.

More Holistic Analysis

Predictive Policing is a multi-disciplinary process which seeks to bring together insights from diverse fields such as actuarial science, statistics, criminology and a contextual understanding of local surroundings.[41] Therefore, it is hoped that this will lead to more comprehensive and holistic analyses of crime patterns.

Harms

Discriminatory Impact of Predictive Policing

a) ***Amplification***

Any data driven decisionmaking system runs the risk of amplifying existing inequities. Each intervention made under predictive policing also feeds into the data that informs decisions. Thus, if certain neighborhoods are identified as crime-prone, there is will be greater policing in those areas, and consequently, lead to more focus on the same areas.

b) ***Institutional Bias***

The form of discrimination in predictive technologies is institutional in that there are implicit biases in data. In India, there are documented problems with the police data which includes 'history sheets', 'rowdy sheets' and records of individuals

classified as 'hooligans', 'goondas' and 'criminal tribes.'[42] A forthcoming paper in the Significance journal uses the publicly available version of the PredPol software to see how the data from 2010 used to predict the crime locations in 2011 compared against the actual crime locations in 2011.[43] Their conclusions raise doubts over the efficacy of PredPol and attract new questions whether it merely reinforces the implicit biases of race, locality, religion etc. that law enforcement personnel have.

c) ***Inscrutability of algorithms***

Further, the predictive policing algorithms are inscrutable to the courts in many circumstances. Therefore, despite constitutional protections preventing the discrimination, redressal mechanisms are not effective.

d) ***Excessive or illegal use of policing***

There is already existing literature which suggests selective application of criminal law in India targeting certain classes of individuals.[44] One of the key goals of predictive policing is to predict likely offenders. These problems are made more acute through use of the same data to predict offenders. The likely interventions in response to such analysis would involve greater attention to the person in question in the form of surveillance. Such measures are problematic and raise new questions about how the standard of reasonable suspicion can be measured against algorithmic insights.

ENDNOTES

1. A mission mode project (MMP) is an individual project within the National e-Governance Plan (NeGP) that focuses on one aspect of electronic governance, such as banking, land records or commercial taxes etc. Within Digital India, "mission mode" implies that projects have clearly defined objectives, scopes, and implementation timelines and milestones, as well as measurable outcomes and service levels.
2. National Crime Records Bureau. "About Crime and Criminal Tracking Network & Systems - CCTNS." available at http://ncrb.gov.in/cctns.htm.
3. Data collected widely includes: 1. Records of Crimes, suicides and accidental deaths at national level, 2. Prison statistics at national level, 3. Fingerprint records of convicted persons, 4. Details of FIRs filed, 5. Details of dead bodies, missing persons and vehicles, 6. Passport details, arms and hotel licenses and vehicle enquiries, 7. The Crime and criminal information repository contains the criminal images and fingerprints, 8. All crime related data – details of summons, bail, parole (Name, Details of crime, reason for parole, duration of parole, location of parole), details of accused persons in remand, details of convicted persons, case statuses, details of visitors of convict (name of convict and visitor, ID of visitor (DL/PAN/UID/Voter ID no), photograph, details of crime/conviction), 9. Details of police stations, districts, range, zone) and user hierarchy, 10. Online registration of Foreigners – Registration at Police Station with details of place of stay, duration, purpose etc., 11. The challan related information for all traffic violations; Traffic Officer/Operator 12. Details of the road certificate, 13. During an emergency call – call and caller details are recorded, 14. System should have details of Police personnel/equipment etc. required to be deploy for tackling emergency, 15. The detail of confiscated Vehicle/Individual/ Property with photograph to the concerned PS/Police Offices who raised the alert, 16. Employee information like Employee ID No., Name of the employee, Date of Birth, height, weight, Permanent Address, Name of the district, Caste, Education Details, Details of the family (Spouse, Children), Date of Joining, Due Retirement Date, any departmental enquiry, Character verification, joined as what designation, Service record of the employee, present deputation etc., 17. Employee includes leave details, transfer details, training, payroll, promotions, Annual Confidential Report (ACR), Service Book of the employee, 18. Details of Departmental enquiries against employees, 19. Post- retirement benefits like pension, PF, GPF, EPF details for employee, PPF Amount drawn by the employee, 20. demands of under various categories like Kits, Arms and Ammunitions, Motor Transport Vehicles, IT Goods, Miscellaneous goods & Items, 21. Stock holding of existing goods, 22. Stock holding of existing vehicles and demand for vehicles in several categories, 23. Case Property release and Personal Search release
4. Bibhu Prasad Routray, Making a case for futuristic predictive policing in India, available at http://www.newindianexpress.com/magazine/voices/article601920.ece?service=print.
5. Supra Note 2.
6. Sarah Brayne, Alex Rosenblat and Danah Boyd, Predictive Policing, DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE, available at http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf.
7. Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, available at http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf, Andrew Guthrie Ferguson, Policing Predictive Policing, forthcoming in 94 WASH. U. L. REV. (2017).
8. Supra Note 3.
9. Kate J. Bowers & Shane D. Johnson, Who Commits Near Repeats? A Test of the Boost Explanation, W. CRIMINOLOGY REV., Nov. 2004.
10. Andrew Ferguson. G. 2012. "Predictive Policing and Reasonable Suspicion." 62 Emory Law Journal 259.
11. DHNS, 'India to turn data-rich in 5yrs', available at http://www.deccanherald.com/content/499677/india-turn-data-rich-5yrs.html.
12. http://mha.nic.in/sites/upload_files/mha/files/CCTNSMHA_11012016.pdf.
13. Request for Proposal, CCTNS, Uttar Pradesh, Volume 1, available at http://www.ncrb.gov.in/BureauDivisions/CCTNS/All%20State%20RFP/Uttar%20Pradesh/UP_CCTNS_SI_RFP_Volume_I.pdf.
14. The NCRB's website described the current possibility of exchange of information between neighbouring police stations, districts or states as being "next to impossible." The aim of CCTNS is to address this gap and integrate and connect the segregated law enforcement arms of the state in India, which would be a foundational step in any initiatives to apply predictive methods.
15. Supra Note 4.
16. *Id.*
17. Request for Proposal, CCTNS, Uttar Pradesh, Volume 1, available at http://www.ncrb.gov.in/BureauDivisions/CCTNS/All%20State%20RFP/Uttar%20Pradesh/UP_CCTNS_SI_RFP_Volume_I.pdf
18. Charlie Beck & Colleen McCue, Predictive Policing: What Can We Learn from Wal-Mart and Amazon About Fighting Crime in a Recession?, POLICE CHIEF, Nov. 2009

19. Joel M. Caplan, Mapping the Spatial Influence of Crime Correlates: A Comparison of Operationalization Schemes and Implications for Crime Analysis and Criminal Justice Practice, 13 CITYSCAPE, no. 3, 2011
20. Peter K Manning, The Technology of Policing: Crime, Mapping, Information Technology and the Rationality of Crime Control, New York: New York University Press, 2008
21. Supra Note 17.
22. Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, available at http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf.
23. Supra Note 19.
24. http://www.rediff.com/news/report/the-ambitious- cctns-a- nation-wide- report-card/20121122.htm
25. MHA GO No. 22011/05/2010-SR/CCTNS, available at http://mha.nic.in/sites/upload_files/mha/files/AdvisoryCCTNS30Nov2015.pdf; Currently in 5 States Police Stations are now entering all the FIRs through CCTNS Software (100% of FIRs including both offline-online) including Karnataka, Tamil Nadu, Delhi, Tripura, Dadra and Nagar Haveli and Daman and Diu, while 4 States have not failed to implement the same in toto – Andaman and Nicobar Islands, Lakshdweep, Bihar and Rajasthan.
26. Status Report by MHA - http://mha.nic.in/sites/upload_files/mha/files/AllStates7thJan2016.pdf
27. http://computer.financialexpress.com/magazine/what-is-causing-the-cctns-delay/15536/#sthash.VfCSHr3r.dpuf
28. Mrinal Satish, Bad Characters, History Sheeters, Budding Goondas and Rowdies" : Police Surveillance Files and Intelligence Databases in India, (November 6, 2010). National Law School of India Review, Vol. 23, No. 1, p. 133, 2011, available at https://ssrn.com/abstract=1703762
29. Cesare Lombroso, Criminal Man (Translated by Mary Gibson and Nicole Hahn Rafter) (2006).
30. *Radhakrishnan v. State of Uttar Pradesh*, AIR 1963 SC 822.
31. *Gulab Chand Upadhyay v. State of Uttar Pradesh*, 2002 CriLJ 2907.
32. Rohan George, Predictive Policing: What is it, How it works, and its Legal Implications, Centre for Internet and Society, available at http://cis-india.org/internet-governance/blog/predictive-policing-what-is-it-how-it-works-and-it-legal-implications#_ftn118
33. Section 165 of the Code of Criminal Procedure states as follows: "*Whenever an officer in charge of a police station or a police officer making an investigation has reasonable grounds for believing that anything necessary for the purposes of an investigation into any offence which he is authorised to investigate may be found in any place with the limits of the police station of which he is in charge, or to which he is attached, and that such thing cannot in his opinion be otherwise obtained without undue delay, such officer may, after recording in writing the grounds of his belief and specifying in such writing, so far as possible, the thing for which search is to be made, search, or cause search to be made, for such thing in any place within the limits of such station*."
34. Cognizable offences have been defined in Criminal Procedure Code as follows— "'cognizable offence' means an offence for which, and `cognizable case' means a case in which, a police officer may, in accordance with the First Schedule or under any other law for the time being in force, arrest without warrant."
35. Section 41 (1) (a) of the Code of Criminal Procedure states as follows: "*Any police officer may without an order from a Magistrate and without a warrant, arrest any person who has been concerned in any cognizable offence, or against whom a reasonable complaint has been made, or credible information has been received, or a reasonable suspicion exists, of his having been so concerned*."
36. *State of Punjab v. Balbir Singh*, (1994) 3 SCC 299.
37. *Partap Singh v. Director of Enforcement, Foreign Exchange Regulation Act*., (1985) 3 SCC 72
38. The law on vercaity of tips in the United States was laid down in *Illionois v. Gates*, 462 U.S. 213 (1983). It was held that "an informant's 'veracity,' 'reliability,' and 'basis of knowledge'- remain 'highly relevant in determining the value'" of the said tip; "Anonymous tips need to be detailed, timely and individualised enough to justify reasonable suspicion. And when the informant is known to be reliable, then his prior reliability may justify reasonable suspicion despite lacking a basis in knowledge." Rohan George, Predictive Policing: What is it, How it works, and its Legal Implications, Centre for Internet and Society, available at http://cis-india.org/internet-governance/blog/predictive-policing-what-is-it-how-it-works-and-it-legal-implications#_ftn118
39. "National Requirement of Manpower for 8-Hour Shifts in Police Stations", by the Bureau of Police Research & Development, Government of India, and Administrative Staff College of India, Hyderabad
40. Predictive Policing in India based Crime Stats of 2013, available at http://www.datasciencedude.com/2015/07/31/predictive-policing-in-india-based-crime-stats-of-2013/.
41. Andrew Guthrie Ferguson, Policing Predictive Policing, forthcoming in 94 WASH. U. L. REV. (2017).
42. Mrinal Satish, Bad Characters, History Sheeters, Budding Goondas and Rowdies, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1703762
43. Jack Smith IV, Crime-prediction tool PredPol amplifies racially biased policing, study shows, Tech.Mic, available at https://mic.com/articles/156286/crime-prediction-tool-pred-pol-only-amplifies-racially-biased-policing-study-shows#.VGg36we9D.
44. Supra Note 28.

The Unique Identity Project

BY AMBER SINHA

Background

In recent years, digital identity systems have been promoted by international bodies such as the World Bank as key to reducing fraud, facilitating financial inclusion, providing for efficient delivery of services, enabling political empowerment, and facilitating economic growth and security in developing countries.[1] In this way, digital identity systems are being closely linked to narratives around development and security.

A Digital Identity system includes the following components: 1) identification, 2) authentication – the process of asserting an identity previously established during a process of identification, and 3) authorization – the process of determining what actions may be performed or services accessed on the basis of the asserted and authenticated identity.[2] Because of these features, a digital identity becomes critical for the use of online services such as e-governance or e-commerce platforms, and facilitates the generation of enormous amounts of transactional data - termed as transactional identity.[3] This transactional data results in behavioral big data.[4] A digital identity system can be both centralised and decentralised.[5] Centralised Identity architectural models include state issued electronic identity and monolithic identity providers (such as Google and Facebook logins). On the other hand, decentralised identity models include identity brokers (the Verify program in UK), personal identity provider models (Microsoft u–Prove) and non identity models (Blockchain identity providers).

The Unique Identity project in India (UID) is an example of a centralised identity architecture model. The Unique Identity Project in India, Aadhaar, which is reported to have over one billion registrants[6] is an online, digital and paperless identity system[7] which can act as a platform for a number of digital services, all of which produce enormous troves of data, precious to both the government and the private sector. The UIDAI was constituted as an attached office under the Planning Commission in 2009. The Unique Identity Project has its origin in the National e-Governance Plan (NeGP) adopted by the Government of India as one of the Mission Mode Projects. Prior to the Aadhaar number, there were at least eighteen documents that are recognized as acceptable proofs of identity.[8] The objective of the scheme has been to issue every resident in India a unique identification number based on an individual's biometrics by the Unique Identification Authority of India, which can be authenticated and verified online. The system was conceptualized and implemented as a platform to facilitate identification and avoid fake identity issues and delivery of government benefits based on the demographic and biometric data available with the Authority.[9]

The Aadhaar number has all the three features of a digital identity system—identification, authentication and authorisation. Once residents have a uniform identification number and service providers have adopted the authenticating technology, service providers can transact with residents via the UID system, rather than engage in the verification process themselves.[10] Various documents prepared by the UIDAI refers to the intended data sets as exponentially growing including "data coming in from residents, vendors and partners" and characterise this transaction data as "Big Data".[11]

How UID Works

Enrollment

The UIDAI has established a 'partnership model' for enrolling individuals into the UID scheme. This model consists of the UIDAI, registrars, and enrolling agencies. Registrars and enrolling agencies are the two bodies responsible for collecting information from individual. Registrars and enrolling agencies can be private or public bodies that have 'a keen business and social interest in ensuring the authentic identity of the people availing their services.[12] These entities are contracted and trained by the UIDAI. During enrollment the following is collected from the individual: name, gender, age, date of birth, address, email, mobile number, details of (father/mother/guardian/husband/wife), proof of identity/address (document based or introducer based) iris scans, photographs, ten fingerprints, and if the individual wishes to link their Aadhaar number to their bank account – their bank account details.[13] This is the baseline information that the UIDAI requires for issuing an Aadhaar number, but private services enrolling individuals are permitted to gather additional information from the individual at the time of enrollment with the consent of the individual. It is not clear if additional information collected by enrolling agencies is stored only by the enrolling agency, or if it is also shared and stored by the UIDAI.

Authentication

Any private or public body can register with the UIDAI to become a 'requesting entity' and can adopt the 'Aadhaar' number as an authenticator for the delivery of services or completion of a transaction. An individual can authenticate with 'Aadhaar' by submitting their Aadhaar number and demographic details or Aadhaar number and biometric details.[14] Only online authentication is available for Aadhaar. Once submitted, the UIDAI will compare this information to that which

is stored in the CIDR and reply with a yes or no response, though potentially more information may be communicated and generated.

The UIDAI does not specify the type of 'metadata' that is collected about an individual during the authentication process, but at the least it could include: Aadhaar number/demographic/biometric information of the individual, time of authentication, location from where authentication took place, entity or organization requesting the authentication, and if the response was yes or no. Furthermore, according to section 8 of the Aadhaar (Targeted Delivery of Financial and other subsidies, benefits and services) Act, 2016 ("Aadhaar Act") requesting entities must inform the individual of users to which information received during authentication may be put to use. Section 29 further clarifies that identity information residing with the requesting entity can be used for purposes beyond those specified to the individual or disclosed further except with prior consent. Thus, it is clear that requesting entities will be storing and potentially sharing information associated with transactions.

As mentioned above, in order to authenticate an individual, the biometrics information is matched with the corresponding Aadhaar number's biometric information in the CIDR.[15] Any requesting entity will— (a) take consent from the individual before collecting his/her Aadhaar information; (b) use the information only for authentication with the CIDR. The entity requesting authentication will also inform the individual of the following— (a) what type of information will be shared for authentication; (b) what will the information be used for; and (c) whether there is any alternative to submitting the Aadhaar information to the requesting entity.[16] Some questions that the authentication process may address include the following:

1. Is this Aadhaar number a validly issued number?
2. Does this Aadhaar number share these biographical details such as name and address?
3. Is this Aadhaar number associated with these fingerprint biometrics?

The response to these questions is always a Yes or a No.[17]

UID as Big Data

Integrating Databases — Seeding

Government departments and institutions in India have always maintained databases on the Indian population, but due to institutional policy, or lack of technology, these databases do not "talk to each other." The digital technology and a 'logic of networks' seeks to create a system which enables interaction across databases.[18] In the UID scheme, data points within databases of service providers between departments, ministries and banks are organized via individual Aadhaar numbers through a process known as 'seeding.' It is a process through which the Aadhaar numbers of residents are included in the service delivery database of service providers for enabling Aadhaar based de-duplication and authentication during service delivery.[19] Though not directly stated, it is envisioned that the Aadhaar number will be seeded into the databases of service providers and banks to enable cash transfers of funds.[20]

The seeding process itself can be done through manual/organic processes or algorithmic/in-organic processes. In the algorithmic process the Aadhaar database is matched with the database of the service provider - namely the database of beneficiaries, KYR+ data from enrolment agencies, and the EID-UID database from the UIDAI. Once a match is found–for example between KYR fields in the service delivery database and KYR+ fields in the Aadhaar database–the Aadhaar number is seeded into the service delivery database. Organic seeding can be carried out via a number of methods, but the recommended method from the UIDAI is door to door collection of Aadhaar numbers from residents which are subsequently uploaded into the service delivery database either manually or through the use of a tablet or smart phone.[21] Seeding allows different databases to be tagged with unique identifiers, thus establishing a relationality, a key feature of big data datasets. Once seeded, disparate datasets can come together and enable big data analytics.

Cradle to Grave

The original press release by the Central Government reporting the approval by the Cabinet of Ministers of the Digital India programme speaks as '**cradle to grave**' digital identity as one its vision areas. The characteristic of this identity are mentioned as "**unique, lifelong, online and authenticable**."[22] We have done a survey of the e-governance schemes under **Digital India**, and tried to map those which seek to service individuals and create a welfare state.[23] In order to provide services effectively, these schemes shall collect and store information through the lifecycle of an individual. The result, as we can see, is building databases on individuals, which when combined, as enabled by the **Seeding** process, will provide **a 360 degree identification**. The schemes under Digital India conceptualise a digital ecosystem,[24] with the Aadhaar functioning if not as the singular, the primary form of identity that would be used to enter this ecosystem.[25] This diagram represents a snapshot of how the entire lifecycle of an individual may be documented.

The Aadhaar number has been recognized as core to the creation of such an identity by the government and the UIDAI. The architecture of the UID scheme and the provisions of the Aadhaar Act are crafted in such a way to enable such a 'cradle to grave' identity. From a legal perspective, the Act allows public and private bodies to adopt the Aadhaar number to establish identity, and for the delivery of services, the Bill allows the government to mandate the use of the Aadhaar number by offering individuals seeking services alternative forms of authentication only if it is not possible to issue them an Aadhaar number. With these provisions, any service can adopt the Aadhaar number, and the more services that adopt the number, the

more comprehensive an individual's profile will be. From a structural perspective, earlier documents from the UIDAI revealed that to enable authentication, the Aadhaar number would be seeded into the databases of service providers.

As part of this research, we have completed a survey of the e-governance schemes under Digital India. The schemes under Digital India conceptualise a digital ecosystem, with the Aadhaar functioning as, if not the singular, the primary form of identity that would be used to enter this ecosystem. We look at Digital India schemes that are Aadhaar-enabled or potentially Aadhaar-enabled and that are in various stages of implementation (implemented, partially implemented, and not yet implemented). In order to provide services effectively, the reviewed schemes require the collection and storage of information at different points of the lifecycle of an individual. This results in the creation of multiple databases containing different data points about an individual, that when seen in cohesion or when Aadhaar authentication requests from each service are compiled, could begin to create a "360 degree" profile of an individual.

a) ***Infancy***

Beginning at the first stage of infancy, use of ICT is intended to be leveraged to enhance the health system at a national level through the e-health initiative.[26] E-health projects include medical record and hospital automation, national tele-medicine infrastructure, integrated disease surveillance, a national medical college network, and a national digital medical library consortium.[27] These projects will entail collection, integration, and centralization of relevant health records from birth to death from multiple databases.[28] The National Knowledge Commission (NKC) has recommended the formation of the National Health Information Authority (NHIA) to support implementation on e-Health. The Ministry of Health & Family welfare and DEITY are developing infrastructure to facilitate the sharing of health information through a National Health Portal and the establishment of a national database containing records of citizens from birth to death.[29] This centralised electronic health record repository of all citizens will ensure that the health history and status of all patients would always be available to all health institutions. Aadhaar will play a key role in generating and tracking health documents and identifying health patients in the initiative and making the same available to public and private bodies.[30] For example, in States such as Haryana, the Aadhaar number is already being linked to the birth certificates of newborns to enable tracking of 'immunisations and status of school enrolment'.[31] Similarly, existing projects like MCTS (Mother and Child Tracking System), which is a system for improving delivery of health services to pregnant women and children up to 5 years of age through name based tracking, will become a part of the scheme. Such projects are in the process of adopting Aadhaar, for example the Health Ministry plans to integrate MCTS with Aadhaar for 'unique identification and generation of records.[32]

b) ***Childhood/ Adolescence***

In the next stage of the lifecycle, in student life, there is the involvement of various projects including National Scholarship Portal, e-Learning and the Education MMP. Optimizing the use of ICT in education is among the key agendas of the Ministry of Human Resources. When these projects are implemented, the details of admission, examination results, scholarships and streams of a student will be documented and centralized. For instance, the National Scholarship Portal,[33] which is being implemented, seeks the Aadhaar number at the time of registration.[34] Other projects such as eLearning and the Education MMP have not been implemented, but documents on the Digital Locker website indicate that these projects are intended to be issuers - and thus it can be presumed that they will also be Aadhaar enabled. A combination of these databases could provide a comprehensive view of the educational history of an individual. This example of streamlining the process for potential employers has been promoted as one of the advantages of Digital Locker. However, a system like this without strong consent requirements could override the choice of an individual before information about them is revealed.

c) ***Adulthood***

The next leg involves the process of issuing various identification and license documents including Passports, Driving License and Vehicle Registration being enhanced through ICT. Passport Seva under the Ministry of External Affairs leverages ICT for the process of issuing and re-issuing passports.[35] Indian passports are being linked to the Aadhaar number.[36] The Ministry of Road Transport and Highways is implementing the road and transport project Vahan (vehicle registration) and Sarathi[37] (drivers license)- software which digitizes the process of issuing drivers licenses and registering vehicles. State Registers (SR) & National Register (NR) have been established as a centralized database for Vahan & Sarathi related data which is meant to be readily available to other government departments.[38] As a note, drivers licenses and vehicle registrations are not backed by Aadhaar yet, though news items in 2013 reported plans to do so.[39] Such databases could enable the profiling of all international travel and to some degree the domestic travel of the individual.

Alongside, as the individuals enter adulthood, there are other services such as Employment Exchange through the National Employment Service, which is implemented and creates a database of potential employees to match against employer requirements;[40] e-Office which has yet to be implemented, but will have performance and attendance details in case of government employees;[41] and projects such as Automation of Central Excise and Service Tax,[42] Income Tax India and e-filing[43] and Commercial Taxes,[44] already partly or fully implemented which provide an e-platform for filing and declaring of taxes and expense. All of these services will be Aadhaar enabled as the Aadhaar number is being linked to the PAN number and according to the Income Tax Department's website, 4,307,343 PANs are Aadhaar-PAN linked,[45] while the Commercial Taxes Department collects Aadhaar

numbers at the time of user registration.[46] The e-District project will be responsible for issuing certificates for income, domicile, caste, birth, death etc.; issuing licenses such as arms licenses, issuing ration cards, disbursing pensions, assessing taxes, processing utility payments and linking to other relevant government projects.[47] Provision of marriage certificates and will collect personal data on previous marital status, religion along with personal detail about witnesses.[48] The e-district portal for a number of States uses the Aadhaar number at the login stage as a user's login ID.[49] All documents uploaded by the user will be available in his user ID (Aadhaar number).[50] Most of these initiatives may be aligned with the Digital Locker and documents such as PAN cards, TAN registration, Service tax registration, marriage certificates shall be uploaded on it.

Additionally, the CCTNS project is intended to connect all police stations across the country and provide them with tools for tracking and analysis. This will ensure that the records of any individual with a local police station shall be available nationally.[51] This will also mean that details of individuals involved in court proceedings in any way shall be available in the databases and potentially combined or shared with other databases such as CCTNS.

d) ***Late adulthood***

The final leg includes the Jeevan Pramaan scheme which entails a Pensioners Portal which shall function as the platform for retired government employees to register and seek their pensions. Jeevan Pramaan is an Aadhaar based Digital Life Certificate for pensioners. This will have details of their dependents, who may be eligible for the pension after they are deceased. Finally, the e-District platform shall be used to generate the death certificate upon the death of the individual.

The presence of an identifier like Aadhaar in e-governance schemes can potentially enable profiling, collation, and convergence of information in two ways:

a) Creating a 'common identifier' through disparate databases or disparate pieces of information - allowing them to easily combine, converge, and organize information as per the Aadhaar number. The use of Aadhaar in various e-governance schemes demonstrates that through application - Aadhaar is becoming a **versatile authenticator** as it is used in different ways by different platforms. For example, Aadhaar is linked to birth certificates, passports, and PAN cards; it is envisioned to be integrated into the MTSC project, it is a required field for registration for the Scholarship Portal, and it serves as the username for the e-District portal.

b) Creating comprehensive authentication trails as services require Aadhaar as a requisite for registration, login, or collection of a service. As per the Aadhaar Act, authentication records would be retained by the UIDAI, but can be requested with proper authorization by law enforcement.[52] Some of the services, like birth certificate, passport, PAN etc. issue identity documents to which Aadhaar is being linked. Yet, by linking Aadhaar to services such as Birth Certificate and passport, the UIDAI is 'tracking a tracker' and enabling centralized records of the use of such identity documents.

In the context of Digital India, the infrastructure that is being created and which Aadhaar is being integrated into can enable a 360 degree profile of an individual and its interaction with state services.

Aadhaar Act and Big Data

The Aadhaar Act passed recently by the Indian legislature also has provisions that retain the original intent of the UID project to enable big data. Section 3 (i) of the Act states that every resident of India is entitled to enrol and get an Aadhaar number. While the enrolment into the scheme is voluntary, in order to have access to certain benefits and subsidies, the Act allows for having an Aadhaar number as a precondition.[53] Thus, the size of population that needs to enrol into Aadhaar is effectively the entire population of the country.

Further, Section 57 allows the Aadhaar number to be used in order to establish the identity of an individual by the government or private bodies.[54] The Aadhaar (Authentication) Regulations notified a few months after the legislation lays down the entire framework for entities to use the Aadhaar Number as the identifier and authenticate it against the CIDR. These provision runs counter to the claims that the Aadhaar number is for the purpose of identifying individuals for better targeted government subsidies. Since the inception of the UID programme and particularly in wake of the various e-governance schemes under the National e-Governance Plan, and its successor, the Digital India programme, we see continued attempts to link Aadhaar with various programmes.[55] The use of UID as the primary identifier and seeding of databases with the Aadhaar numbers lays the groundwork for sharing of databases across government agencies. Data sharing usually includes data matching, joint access to repositories of data, file duplication, and methods of data access that enables more than one agency or organization to use personal data. Such data sharing results in personal data moving out of traditional data silos and being used in new ways, by different agencies, or for new or different purposes.[56] The Aadhaar (Authentication) Regulations mentions the process and duration for maintenance of data logs. While the regulations maintains that such records may not shared with any third parties without the consent of the data subjects,[57] broad privacy notices and blanket consents as a condition to use the services can easily enables a creation of a data marketplace around the authentication logs. The limitless scope of UID and the push to use it as the primary, if not the sole identifier for various government services leads to easy relational conjoining of databases, thus, enabling to big data.

Promises of UID as Big Data

a) ***Efficient service delivery***

A consolidated view of the entire data would facilitate the social welfare department of the state to improve the service delivery in their programs, while also being able to ensure that the same person is not availing double benefits from two different districts or through different identities.

b) ***Convenience for citizens***

Convergence of data will mean that citizens are not asked for the same data many times and there could be a potential one stop shopping. Therefore, updates like changes in address need not be done across multiple agencies and databases. It is noteworthy that so far the implementation of Aadhaar as an identifier has not reduced the number of identifications required, rather introduced an additional layer of identification in order to access services.[58] These issues will have to be addressed before this benefit is realised.

c) ***Better information dissemination and training***

One of the benefits of a holistic approach to government modernization and innovation is the ability to communicate information to citizens effectively about more than one program. Disparate projects require communications targeted to that individual program, and the citizen has only so much bandwidth to spend on figuring out how to do things and get service.[59]

Risks of UID as Big Data

a) ***Profiling***

Via Aadhaar - disparate pieces of data collected by different agencies can come together to create a 360 degree profile of the citizens. This enables greater surveillance and impedes any efforts towards anonymity. This could lead to chilling effects not only on privacy but also on the citizens' right to free speech and expression.

b) ***Function creep***

Increased collection of personal data also risks implicit or explicit "**function creep**," where data collected for one purpose is gradually used for others to which the individual has not consented. An additional risk in the case of UID is that the data ostensibly collected for more efficient public service delivery can be used in enabling datafication of human behavior by the private sector. In combination, **data aggregation** in which innocuous information come together to paint a picture of activities and individual attributes, greatly increasing an individual's vulnerability to dangers such as targeted discrimination (both public and private), intrusive surveillance, identity fraud, stalking or harassment.[60]

c) ***Change in the relationship between citizen and state***

Sharing of data across government departments and agencies will mean that citizens may not know which organization ultimately holds and uses the data, and this will cause a lack of trust in the relationship between citizens and state. Thus, we see greater use of Aadhaar as a digital identity, a push to move all services online and the emphasis on cashless transactions fundamentally altering the nature of the relationship between the state and citizens. The use of these technologies creates an additional layer of automated systems which the citizens must engage with without any understanding of how it works. The decisionmaking process is intended to be more data driven which also makes it an algorithmic black-box, limiting the ability of an individual to question it.[61]

d) ***Knowledge Gap***

Cradle to grave databases will lead to decisions being made using unrelated, data without the knowledge of the citizen. People do not fully comprehend how powerful data data mining techniques workare, how much data analysis is being done or who uses their personal data;[62] These kinds of black box systems where individuals are aware of some surveillance and analysis but not aware of how exactly it works often leads to greater paranoia and attempts to game the algorithms without truly understanding the rule of decision making.[63]

e) ***Lack of accountability***

In the absence of data governance regulations, there are no clear mandates on the different agencies getting access to data. For instance, UIDAI data is being used in data analytics in education in the state of Andhra Pradesh. However, the complete lack of regulatory oversight leads to no controls on how the data is used, who owns it and what security protocols are required for its protection. These risks gets further exacerbated with the involvement of the private sector through PPP projects (public Private Partnership) which is the norm in all e-governance projects in India.[64]

IndiaStack - The Future of Digital Identity System

What is IndiaStack?

IndiaStack is a set of APIs conceptualised by iSPIRT, a consortium of select IT companies from India, to be deployed and managed by several government agencies, including the National Payments Corporation of India (NPCI).[65] One of the key features of IndiaStack is the open-loop interoperability between providers which will mean that different private and public parties can build over the APIs to create their own customized products and services. The set of APIs bunched

under IndiaStack gives structured methods to public and private agencies to communicate with the CIDR operated by UIDAI, the UPI payment infrastructure operated by NPCI, and the DigiLocker system operated by the Ministry of Electronics and IT. This enables development of various independent client-facing services, mostly through mobile apps, that utilise the functions offered by CIDR, UPI, and DigiLocker, such as online verification of the identity and official documents of a client, and making payments from and to the bank account of the client. This allows these independent service providers to gain access to identity verification services and financial transaction data that were previously only available to government agencies and banks. [66]

This kind of digital identity and payment solution is emblematic of the move towards generative models. It is argued that attempts at financial inclusion are not met simply through greater investment into the microfinance sector and greater trickle down in the traditional financial sector. In order to reach low-income consumers, restraints to profitability for the private sector needs to be removed so that there is greater industry participation in creating business models that address the needs to those in the bottom of the pyramid. [67] The UIDAI has seen Aadhaar as an platform of not just the government and citizens, but also of vendors, developers and applications. The presumption has been that the identification, authentication and authorisation, the three key elements of a digital identity systems would become central to the economy, and Aadhaar offers the pivotal infrastructure to build these applications.

An important aspect of this ecosystem is the growing direct benefits schemes in India, the most important one being the Pradhan Mantri Jan Dhan Yojna, publicised widely in the JAM model of development, which comprises Jan Dhan (bank account), Aadhaar (digital identity) and Mobile (connectivity). The distribution of social welfare as physical goods has been the prime issue that welfare governments in the last twenty years have tried to address in the last few decades. While two decades back, there was a move towards governments procuring from the markets rather than producing the welfare benefits, the more recent narrative espoused by the proponents of UID and IndiaStack is to offer cash or coupons directly to beneficiaries instead of procurement by governments. [68]

Various talks about the IndiaStack speak of the following layers as integral to its structure:

a. An identity layer supported by UID project,

b. A presence layer which enables users and service providers to transact and communicate remotely,

c. A paperless layer which leverages the existing functionalities built over UID such as e-KYC and e-Sign,

d. A cashless layer which is an interoperable payments infrastructure, and

e. A consent later based on the idea of OpenPDS conceptualised by a team at MIT Media Labs. [69]

The cashless layer involves a payments switch which enables mobile account holders to transfer funds to and from any account in the country. With the implementation of the United Payments Interface (UPI), an architecture and set of APIs which will work on mobile, [70] users can transfer money between different banks, or even do an online transaction with far greater ease. Indian regulators such as RBI, SEBI (securities), IRDA (insurance) and TRAI (telecom) have broadly accepted Aadhaar-based eKYC. [71] As more and more transactions move from cash to noncash medium, the amount of data being generated increases dramatically. This shift will enable Anytime-Anywhere payments without a need to have access to cash. [72]

Promises of IndiaStack

a) ***Presence-less use***

Once registered on Aadhaar, residents can remotely identify and authenticate themselves by inputting his fingerprint or iris information on a biometric reader. The Central Identity Data Repository (CIDR) which houses the Aadhaar enrolment data offers the basic validation service which takes as inputs the biometric and identity numbers and validates whether they match in the database. Residents can now sign up for services without having to be physically present.

b) ***Speedy and more efficient transactions***

IndiaStack promises to enable the following:

1. a system that allows customers to use their Aadhaar numbers and biometric information to auto-populate information, converting what was a 30-day paper-based process into a short digital process. [73]
2. a database that enables citizens to digitally store and retrieve loan documents, land titles, diplomas, training certificates and other documents in a single place.
3. an API for digital signature in order to remotely and digitally sign and authenticate documents.

c) ***Fruits of a data rich economy***

The data generated through use of IndiaStack and application built over it is valuable to governments (mitigate corruption and tax evasion, and making policy decisions), [74] financial institutions (payments data to build credit scores), [75] advertisers (purchasing data for targeted advertising), services providers (sales data to inform business plans).

Risks of IndiaStack

a) ***Complete loss of anonymity***

Moving a cashless system as default mode of payment will decrease anonymity dramatically, and this could lead to greater surveillance and impacts on free expression and choice.[76] Data will be collected about purchases made, medical services sought and received, food and beverages consumed etc. This data can be used for targeted advertising and sharing of data across bodies can also lead to services providers making an assessment about recruitments, credits etc.

b) ***Potential denial of financial agency***

Technological failures can be disastrous as it may deny financial agency. If services insist on payments through these modes only, it could lead to a completely inability of transact where there are technological failures. Similar cases of cashless mobile based payments in the Global South such as M-Pesa have faced problems of delays due to information having to pass through multiple platforms and connectivity issues.[77]

c) ***Predatory practices***

Lack of regulation in the Fintech sector can enable predatory practices where right to remotely deny financial services can be granted to private sector companies. For instance, where all transactions are being handled through a mobile phone, inability to pay EMIs could lead to disconnection of mobile phone services.

d) ***Doubts over the Consent Layer***

There is a lack of clarity about the whether there will be a blanket consent allowing service providers to authenticate the identity, and access all particulars of an individual, or the consent will be specific to each piece of information. Further, once consent has been provided, there are no effective opt-out mechanisms available in the Aadhaar Act or architecture. Open PDS, which the consent layer is supposed to be based on, essentially acts as a bank for your personal information through interfaces that allow you to manage, query and selectively share data with others. SafeAnswers allows third-parties to inject code into your system and will safely aggregate certain allowed kinds of data and only then share safe summaries with third-parties.[78] There is so far little clarity on how exactly it will work and how a model like SafeAnswers will be adapted.[79]

e) ***Regulation by code***

An architecture like IndiaStack enables datafication of financial transactions in a way that enables linked and structured data that allows continued use of transaction data collected. This phenomenon is exacerbated due to the creation of a **state sanctioned monopoly** like IndiaStack as default mode for payments. This is a classic example of code behaving as law[80] and determining modes of behavior between citizens. Interestingly while OpenPDS is available to be installed on any server under the control of the individual (personal server, virtual machine, etc) or can be provided as a service (Saas model),[81] the option to install IndiaStack on an individual's server is not available, thereby limiting the level of control an individual has upon it and runs counter to the idea of decentralisation. Further, this is one of the first instances that a state sanctioned product has been developed by the private sector independently as opposed to under the RFP model, thus, circumventing regulatory oversights by the Comptroller and Auditor General and under the Right to Information Act.[82]

f) ***Identity marketplace as surveillance***

The collection and analysis of huge amounts of personal information is critical for most digital companies to establish a competitive advantage in the market. But even if the data is not strictly necessary for current processes, most companies feel compelled to collect it. The fact that data storage is much cheaper has led to data hoarding as the default state. Shoshana Zuboff, former Faculty Associate at the Berkman Center for Internet and Society at Harvard Law School, describes this as 'surveillance capitalism,' Zuboff refers to the commodification of human behavior as one of its key goals, which can be sold as information to different parties who may use it in their decisionmaking process.[83] The UID project enables this Identity Marketplace in the following two ways: — a) a unique identifier introduces relationality, the ability to conjoin different databases, and b) the introduction of IndiaStack, a set of APIs built upon the UID infrastructure could enable moving the datafication[84] of a majority of transaction by moving them from analog to digital, all of them also connected to a unique identifier.

ENDNOTES

1. http://pubdocs.worldbank.org/en/413731434485267151/BriefonDigitalIdentity.pdf
2. Paul Makin, Steve Pannifer, Carly Nyst, Edgar Whitley, Digital Identity: Issue Analysis, available at http://www.chyp.com/wp-content/uploads/2016/07/PRJ.1578-Digital-Identity-Issue-Analysis-Report-v1_6-1.pdf
3. Decentralised Citizens ENgagement Technologies, Research on Identity Ecosystem, available at http://dcentproject.eu/wp-content/uploads/2015/08/D3.3-Research-on-Identity-Ecosystem_part1.pdf
4. Rohan Samarajiva, What is big data?, available at http://lirneasia.net/2015/11/what-is-big-data/.
5. On one side we have the view that we must adapt traditional ideas of identity, such as the ID card and ID number, to the digital age. This means that national governments should be the main providers of digital identities. On the other side are the proponents of using multiple identity sources and other more decentralised systems that do not rely on state assurance. These could provide a better balance between publicly controlled resources and personal data owned by citizens as commons.
6. http://www.ndtv.com/india-news/aadhaar-enrollment-crosses-1-billion-mark-ravi-shankar-prasad-1338621.

7. Jake Kendall and Stephen Deng, It's the ecosystem, stupid!, available at http://nextbillion.net/nexthought-monday-its-the-ecosystem-stupid-exploring-the-digital-poverty-stack-part-1/
8. These include passport, PAN card, ration/ PDS photo card, voter id, driving license, government photo id cards, NREGS job card, photo id issued by a recognized educational institution, arms license, photo bank ATM card, photo credit card, pensioner photo card, freedom fighter photo card, kissan photo passbook, CGHS / ECHS photo card, address card having name and photo issued by department of posts and certificate of identify having photo issued by group a gazetted officer on letterhead, disability ID Card/handicapped medical certificate issued by the respective State/Union Territories.
9. UID FAQ: Aadhaar Features, Eligibility. Available at: https://resident.uidai.net.in/faqs
10. Id.
11. ANALYTICS - EMPOWERING OPERATIONS: THE UIDAI EXPERIENCE available at https://uidai.gov.in/images/FrontPageUpdates/uid_doc_30012012.pdf; AADHAAR TECHNOLOGY & ARCHITECTURE, available at https://uidai.gov.in/images/AadhaarTechnologyArchitecture_March2014.pdf.
12. https://aadhaar.maharashtra.gov.in/upload/EnglishFAQAADHAAR.pdf
13. https://www.karnataka.gov.in/nammaaadhaar/documents/Application%20form%20-%20English.pdf
14. https://aadhaar.maharashtra.gov.in/upload/EnglishFAQAADHAAR.pdf
15. Section 2 (iii) and Section 7 of The Aadhaar (Targeted Delivery of Financial and Other Subsidies, Benefits and Services) Bill, 2016.
16. Section 8 of the Aadhaar Act, 2016.
17. However, the Aadhaar Act also provides that the CIDR may, in addition to a Yes/No response, also provide other 'appropriate' information to the requesting entity. This has raised questions about sharing of collected data.
18. Swagato Sarkar, The UID Project and the new 'Bureaucratic Moment' in India, available at http://www3.qeh.ox.ac.uk/pdf/qehwp/qehwps194.pdf.
19. UIDAI. Approach Document for Aadhaar Seeding in Service Delivery Databases. Version 1.0. Available at: https://authportal.uidai.gov.in/static/aadhaar_seeding_v_10_280312.pdf.
20. Id.
21. Elonnai Hickok, Supreme Court Order is a Good Start, but is Seeding Necessary?, Centre for Internet and Society, available at http://cis-india.org/internet-governance/blog/supreme-court-order-is-a-good-start-but-is-seeding-necessary#fr3.
22. http://digitalindia.gov.in/content/vision-and-vision-areas.
23. Sumandro Chattapadhyay, The Last Chance for a Welfare State Doesn't Rest in the Aadhaar System, available at http://thewire.in/30256/the-last-chance-for-a-welfare-state-doesnt-rest-in-the-aadhaar-system/.
24. Ravi Shukla. "Technologies in Aadhaar: A sociological view" in Ashish Rajadhyaksha ed., In the wake of Aadhaar: the Digital Ecosystem in India, available at http://www.jnu.ac.in/SSS/CSSP/RaviShukla.pdf
25. Section 57 allows the Aadhaar number to be used in order to establish the identity of an individual by the government or private bodies. This provision runs counter to the claims that the Aadhaar number is for the purpose of identifying individuals for better targeted government subsidies.
26. http://economictimes.indiatimes.com/industry/healthcare/biotech/pharmaceuticals/why-ehealth-is-a-key-pillar-of-digital-india/articleshow/48069561.cms
27. http://telemedindia.org/india/e-Health%20Initiatives%20in%20India.pdf
28. http://indianexpress.com/article/india/india-news-india/digital-india-scheme-to-focus-on-health-sector-arvind-gupta/
29. http://www.ijser.org/paper/Authenticating-Indian-E-Health-System-Through-Aadhaar-A-Unique-Identification.html
30. http://indianexpress.com/article/india/india-news-india/digital-india-scheme-to-focus-on-health-sector-arvind-gupta/
31. http://indianexpress.com/article/cities/chandigarh/digital-india-digital-india-initiative-software-technology-park-of-india-software-technology-park-in-panchkula-panchkula-news-under-the-digital-india-initiative-software-technology-park-of-india/
32. http://pib.nic.in/newsite/mbErel.aspx?relid=114861
33. https://scholarships.gov.in/main.do
34. http://registrations.scholarships.gov.in/newReg.do
35. http://passportindia.gov.in/AppOnlineProject/online/faqServicesAvailable
36. http://passportindia.gov.in/AppOnlineProject/online/downloadEFormStatic
37. https://sarathi.nic.in:8443/nrportal/sarathi/HomePage.jsp
38. http://www.cips.org.in/road-transport-mission-mode-project
39. http://timesofindia.indiatimes.com/city/mumbai/Aadhar-card-may-be-linked-to-driving-licence/articleshow/26011418.cms
40. http://www.employmentservice.nic.in/
41. http://www.deity.gov.in/content/e-office
42. https://www.aces.gov.in/
43. http://incometaxindiaefiling.gov.in/
44. https://www.apct.gov.in/apportal/
45. http://incometaxindiaefiling.gov.in/
46. https://www.apct.gov.in/apportal/DLRServices/Dlremail_login.aspx
47. http://www.deity.gov.in/content/e-District
48. http://nisg.org/files/documents/A05040001.pdf
49. http://timesofindia.indiatimes.com/city/delhi/SC-Aadhaar-order-poses-hurdle-for-e-district-project/articleshow/47994912.cms and https://edistrict.kerala.gov.in/registerPortalUser.do?token=144368111091717689873277741 73
50. http://edistrict.delhigovt.nic.in/in/en/Public/TermsAndConditions.html
51. http://ncrb.nic.in/All%20State%20RFP/Uttar%20Pradesh/UP_CCTNS_SI_RFP_Volume_I.pdf
52. Sections 29 and 33 of the Aadhaar Act.
53. Section 7 of the Aadhaar Act.
54. Section 57 of the Aadhaar Act reads as follows: Nothing contained in this Act shall prevent the use of Aadhaar number for establishing the identity of an individual for any purpose, whether by the State or any body corporate or person, pursuant to any law, for the time being in force, or any contract to this effect: Provided that the use of Aadhaar number under this section shall be subject to the procedure and obligations under section 8 and Chapter VI.
55. Software Freedom Law Centre, India's Surveillance State, available at http://sflc.in/wp-content/uploads/2014/09/SFLC-FINAL-SURVEILLANCE-REPORT.pdf
56. Stephanie Perrin, Jennifer Barrigar and Robert Gellman, Government Information Sharing: Is Data Going Out of the Silos, Into the Mines?, available at https://www.oipc.ab.ca/media/389571/Report_Government_Information_Sharing_Jan2015.pdf.

57. Regulation 18 of Aadhaar (Authentication) Regulations, 2016.
58. Anumeha Yadav, Rajasthan presses on with Aadhaar after fingerprint readers fail: We'll buy iris scanners, available at http://scroll.in/article/806243/rajasthan-presses-on-with-aadhaar-after-fingerprint-readers-fail-well-buy-iris-scanners.
59. https://www.oipc.ab.ca/media/389571/Report_Government_Information_Sharing_Jan2015.pdf
60. Megan Brewer, Nicholas Menzies and Jared Schott, Identification Systems don't always serve the bottom 40%, available at http://documents.worldbank.org/curated/en/774661467986277535/Making-identification-systems-work-for-the-bottom-40-percent.
61. Daniel Solove, Privacy self-management and consent dilemma, 2013 available at http://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2093&context=faculty_publications
62. Peter Schaar, former data commissioner for Germany recently published a call to arms on data mining in a special edition of MIND http://www.collaboratory.de/images/1/11/PrivacyandInternetGovernanceMIND7.pdf.
63. Frank Pasquale, The Algorithmic Self. The Hedgehog Review, available at http://www.iasc-culture.org/THR/THR_article_2015_Spring_Pasquale.php
64. Jandhyala BG Tilak, INDIA: State seeks greater role for private sector, available at http://www.universityworldnews.com/article.php?story=2011111121430132.
65. http://www.indiastack.org/
66. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
67. Jake Kendall and Stephen Deng, It's the Ecosystem, Stupid – Exploring the 'Digital Poverty Stack,' Part 1, Next Billion, available at http://nextbillion.net/nexthought-monday-its-the-ecosystem-stupid-exploring-the-digital-poverty-stack-part-1/.
68. Swagato Sarkar, The UID Project and the new 'Bureaucratic Moment' in India, available at http://www3.qeh.ox.ac.uk/pdf/qehwp/qehwps194.pdf.
69. Pramod Verma, Understanding the IndiaStack, available at https://www.youtube.com/watch?v=jwV0-NWmnbc.
70. http://www.thehindu.com/business/Economy/what-is-unified-payment-interface/article9037034.ece
71. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
72. Supra Note 15.
73. Supra Note 70.
74. Id.
75. Id.
76. Dominic Frisby, Why we should fear a cashless world, available at https://www.theguardian.com/money/commentisfree/2016/mar/21/fear-cashless-world-contactless.
77. http://www.microsave.net/files/pdf/BN_94_Riding_M_Pesa_Rails_Advantages_Disadvantages.pdf
78. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790
79. http://www.slideshare.net/ProductNation/india-stack-towards-presenceless-paperless-and-cashless-service-delivery-an-ispirt-initiative
80. Lawrence Lessig, Code v.2.0, available at http://codev2.cc/.
81. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790
82. Pramod Verma et al., Open Ecosystem Questions(OEQ) Hangout to introduce the IndiaStack, available at https://www.youtube.com/watch?v=x4_1Fv9Dc2M.
83. Shoshana Zuboff, Big Other: Surveillance Capitalism and the Prospects of an Information Civilization (April 4, 2015). Journal of Information Technology (2015) 30, 75–89, available at https://ssrn.com/abstract=2594754.
84. Datafication has been defined as the transformation of social action into online quantified data, enhancing real-time tracking and predictive analysis in the hope that this will boost profits and create new sources of revenue.

Big Data in Credit Scoring

BY AMBER SINHA

The Fintech Story

The Fintech story in India, and in the Global South, often is one of financial inclusion and empowerment, and relies on statistics demonstrating an untapped financial market comprising the poor.[1] The Indian fintech software market is forecasted to touch USD 2.4 billion by 2020 from a current USD 1.2 billion, as per NASSCOM.[2] Prominent business areas in the Fintech sector in India include credit-scoring, lending, payments, brokerage and insurance.[3] One of the fast emerging areas of opportunity is credit scoring due to the low penetration of financial services, and the emphasis on financial inclusion by the government.[4] Many of the FinTech business models focused on credit scoring by leveraging alternative credit data in combination with big data analytics to provide services.[5] Examples of emerging credit models include: small ticket unsecured loans, pre-paid plans for single medical procedures, instant point of sale credit, pay per day insurance,and micro-investment.[6]

One of the key areas of focus in the **Digital India** mission is the use of Mobile and e-Banking technologies to enable greater financial inclusion.[7] Mobile and e-Banking technologies are part of a larger shift that is happening in the banking sector, whereby companies are utilizing technology to provide financial services - often through new business models and platforms.

Big Data in Credit Scoring

What is Credit Scoring?

Credit scoring is a method of calculating the potential risk of credit applicants involving use of statistical techniques applied to historical data about the applicants.[8] The outcome of this exercise, the credit score is a summary of an applicant's creditworthiness that informs lending decisions.[9] These credit-scoring approaches typically assess three characteristics: a) the consumer's identity, in order to minimise fraud, b) the ability of a consumer to repay, and c) the willingness of a consumer to repay.[10]

Big Data Enabled Credit Scoring

With the increased use of information and communication technology, particularly through mobile phone penetration, everyday activities of people leave behind a much **larger digital footprint which can serve as behavioral data**. This big data phenomenon has also impacted financial institutions and there is a **greater push to move beyond traditional sources of data** for credit scoring and underwriting, as well as use of big data technologies along with the conventional statistical techniques.[11] Big data proponents claim that big data will gives creditors a fuller picture of a consumer and therefore gives a more accurate prediction of the consumer's ability to repay. These practices involve analysis of numerous 'potential credit variables' in a manner that it provides insights about an applicant's creditworthiness.

Credit scoring using alternate data in a regulatory vacuum also poses the risk of bad lending decisions, discriminatory results and mission creep. So far, we have not seen too much movement on regulation of credit scoring companies in the Fintech sector. In 2014, the RBI has produced a report on Data Format for Furnishing of Credit Information[12] but no regulatory steps have followed it. The risks are exacerbated due to the lack of any fair credit underwriting and fair lending regulations in India. Further, unlike in other jurisdictions like the US, there is limited right to access information regarding the data used to generate credit reports.

How Alternate Data Informs Credit Scoring?

Big data enabled credit scoring works by collecting, identifying and analysing data that can be used a proxies, as mentioned above, for the the three key questions in any credit scoring model—a)identity, b) ability to replay and c) willingness to repay. With the advent of big data and greater digitization and datafication of information, new data sources such as telecom data, Utilities data, Retailers and wholesale data and government data, are available. Below are some of the examples of alternate data and how they are used to calculate creditworthiness.

Telecom data

The prepaid data and recharge patterns is said to provide insights about a person's cash flows. The daily call patterns and location data can indicate whether a person is working a steady job or not. One of the key sources of proxy data about income and spending is the texts about payments, and the credit and debit texts received on the consumer's mobile phone. Payment of bill, purchases made, regular remittances and made or received are all deemed very useful in predicting a consumer's ability and intent to repay.

Utilities data

The digitisation of records and the use of digital payment mechanism to pay utilities bills make this data available for analysis. This data not only shows the

consumption patterns of an individual, but also how timely the person is in making payments. The payments cycles for utilities bills are usually monthly like monthly repayment cycles and therefore considered highly indicative of how the person handles their monthly financial obligations.

Retailers' data

With the greater use of loyalty cards, the retailers maintain the records of purchases made by its consumers. This data can be used to evaluate the individual's expendable income, their family structure, other relevant characteristics, for instance, purchase of certain goods can suggest health consciousness while others may indicate risk taking abilities.

Mobile app data

The metadata collected by the mobile apps used by small lending firms are analysed to derive insights about the consumer. The mobile apps typically seek various permissions to access other data on the person's mobile phone and their logon identities like Facebook and Google. Further, psychometric analysis of the manner in which the consumer fills the online form on the app, such as time taken on each question, the number of times an answer is changed etc. are also seen as indicative of the individual's character.

Regulating Fintech

The Reserve Bank of India (RBI) has considered different aspects of FinTech and potential forms of regulating the same. For example, the **RBI released a Master Circular** on Mobile Banking transactions in India in July, 2016, which recognized the importance of mobile phones in attaining financial inclusion. Earlier this year in April, 2016, the RBI had also released a Consultation Paper on P2P Lending as well as regulation of Account Aggregators. The RBI also provided an in-principles approval to a number of Fintech companies to set up payment banks.[13] We already see a discussion around the regulatory approaches that the government must adopt to these emerging sectors and what kinds of regulation may be desirable.[14] In general, money lending laws are concerned with protecting the interests of borrowers by imposing a ceiling on interest rates, mandatory licensing requirement for money lenders, and making further and better provision for the control of money-lenders and for the regulation and control of money-lending. As evidenced by the refusal in a 2011 decision by a High Court in India to recognise NBFCs within the definition of a 'moneylender' under a state legislation,[15] new Fintech companies who may be in the position of a lending company are not governed by existing money-lending laws.

An argument often made is that for the Fintech firms to thrive and innovate, a regulatory sandbox for these firms is required so that they are not riddled with unnecessary overheads. Due to the new business models employed by Fintech companies, they often exist in regulatory vacuums where they fall outside the scope of regulatory frameworks governing traditional banking and financial institutions. There have also been calls for a regulatory sandbox for the Fintech sector in India. This would involve controlled limited-scale experiments of financial innovations in controlled environments, starting with a limited size. Once there is data on their benefits and harms, the regulatory leeway is formalised for the entire sector.[16] However, these the following factors must be kept in mind:

Opacity of Big Data

Big data enabled credit scoring poses the challenge of opaque algorithms using undisclosed and proprietary methodology which could be used to circumvent fair lending regulations.[17] Even in jurisdictions which provide right of access to citizens to check and verify the credit report,[18] use of credit scoring using big data will prevent them from examining how loan eligibility was determined.[19]

Lack of Non-discrimination Regulations

There is a lack of non-discrimination regulations in the credit scoring industry in India which prevent intentional discriminatory use of data or obligations to safeguard against unintentional disparate impacts of data-driven decisionmaking. Thus, there are no laws which prevent the firms from collecting data on religion, caste etc. which can be used toward disparate treatment. Even in other jurisdictions, there is a call for Fintech firms to be exempt from equal credit opportunity and fair credit regulations.[20] However, regulations which would prevent discriminatory practices are essential for any financial products introduced in the market.

Greater Protection Required for the Poor

People who lack the education, information, and other economic, cultural, and social capital that would allow them to take advantage of—and shield themselves against—the free market are most vulnerable and need greater protection. The consequence of bad decisions are far more dire for those disadvantaged and lacking the resources—financial, psychological, social, and political—to compensate for their errors. A review of big data enabled loan products by the National Consumer Law Centre in the US showed that they were very poor payday loan alternatives. Most of these products involved annual percentage rates three times higher than considered non-predatory. Most importantly, most products require electronic access to the applicant's bank account or some other arrangement of automatically deducting the the owed amount from the borrower's account.[21]

In India, the government has rolled out the direct benefits transfer scheme (DBT) in various states where benefits and subsidies are transferred to the bank accounts seeded with the Aadhaar numbers of the individuals directly in order to bypass the intermediaries involved in the flow of funds, thereby reducing leakages.[22] In the

absence of regulations governing lending practices and credit reporting for small loans, lending companies could provide small loans to even those who may not have the capability to pay back loans, as long as they can deduct the DBT benefits reaching the borrower's bank account towards repayment of loans.

Are De-risking Strategies by Fintech Firms Legitimate?

The only way Fintech firms can drive financial inclusion is by 'de-risking' those who would otherwise be considered as risky borrowers. This can be done by collecting consumer data and using it to condition consumer behavior, for instance, through targeted advertising. If there is too much consumer spending on particular products and services, this data can be sold to companies providing these products and services. [23]

Keeping the above in mind, while it may be desirable that small Fintech firms are not saddled with the kind of regulatory overhead costs that traditional banks are subject to, the peculiar problems and abusive practices, such as contacting borrowers family and friends, dissemination of personal information, unauthorized transaction, high interest rates and predatory practices, that we have seen Fintech firms exhibit in their short life suggest a need for some regulation. These should include regulation ensuring fair lending practices and ensuring that borrowers have the ability to pay back the loan, costs spread evenly across the lifetime of the loan, preventing overcharging of bank accounts and disallowing use of subsidies meant for essential services for repayment. [24]

Potential Benefits

A. Financial Inclusion

Use of alternate data for credit scoring offers the promise of increased financial inclusion. According to a study done by Omidyar Foundation, "in China, Brazil, India, Mexico, Indonesia, and Turkey—alternative credit scoring using big data can enable anywhere between 325 million and 580 million people to get access to credit." This could potentially provide millions of people with access to the formal economy and its advantages such as savings and insurance services, and consequently greater choices and opportunities which will lead to more dignity. [25]

B. Substantive Market Opportunity

Market experts have been extremely optimistic about the potential for private sector opportunities available through the Fintech sector in general, and a larger credit markets in particular. A report by McKinsey estimates a private sector opportunity in the region of USD 4-5 billion in India in the next few years. This could offer numerous job opportunities and contribute significantly to the economy.

C. Data Driven Efficiency

The argument for big data enabled credit scoring also points to the limited scope of traditional underwriting and lending practices. As the scope of data collection under big data is much more comprehensive, it is argued that it will lead to better decisionmaking about the creditworthiness of a consumer. On the other hand, critics of big data have warned against the untested correlations that big data relies on and its efficacy in predicting creditworthiness.

D. Ease of Doing Business

The possibility of getting loan approvals in a matter of hours or days, as opposed to weeks, and presence of various competing private sector companies offering the services, makes big data enabled credit scoring extremely attractive. [26] Loan applications in India typically require a number of documents such as Identity proof, residence proof, statement of accounts, salary slips, notarised copies of all documents, wet signatures, in-person verification, physical inspections of property etc. [27] A streamlined process that can provide a respite from the paperwork could be invaluable.

Harms

A. Use of Data as Proxy for Discrimination

Instead of using discriminating factors such as religion and caste which could potentially be addressed by non-discrimination laws, firms could use proxies of such factors such as neighbourhood of residence, purchasing habits. Because big data scores use closed and proprietary algorithm based technologies, it is impossible to analyze them for potential discriminatory impact. There are no regulations that may be used to address discrimination on the basis of the disparate impacts of data driven decisionmaking in India.

B. Imposition of Homogenised Standards

Psychometric tests and behavioral analysis being used in the Global South based on Myer Briggs tests, or the SATs [28] designed in the western countries fails to take into account local contexts. Similarly, lack of grassroots understanding can lead to wrong inferences. For instance, mobile phone data being used to profile an applicant can be often misleading in India as mobile phone are often shared or handed down between members of a family. [29]

C. Data Ideology and Lack of Domain Expertise

A preliminary survey of the emerging companies in the Fintech sector in India and the profiles of their management teams shows a preponderance of those with

technology and sales background and a lack of individuals trained in banking and finance. This suggests an over reliance on data, and a tendency to ignore other kinds of expertise which have been integral to the credit scoring industry. This is reflective of the narrative that data is exhaustive and comprehensive enough to provide inferences that negate the need for domain expertise, theoretical models and interpretivism. However, this assumption has been greatly critiqued and various authors have pointed out the perils of the over-reliance on data.[30]

D. Regulatory Circumvention

Fintech firms engaged in lending practices are not subject to similar regulations as traditional financial institutions and seen as technology companies as opposed to financial companies. While this provides an opportunity to innovate without too many regulatory overheads, it also exempts such firms from regulations concerning confidentiality, data privacy, underwriting and ethical practices.

E. Mission Creep

Practices such as checking credit scores during background verification for employment have been criticised for a long time. However, big data enabled credit scoring provides a far more granular profile dealing different behavioral aspects of a person and big data ecosystem provides more opportunities for credit data to be used for non-credit purposes. In light of the lack of regulation in the Fintech sector, there is risk of such practices emerging as a business model to generate additional revenue by the companies.

F. Creation of Financial Bubble

Unlike the traditional credit scoring models, the efficacy of the new big data enabled credit scoring models are completely untested. Further, most big data enabled credit decisions are towards loans without any collaterals, thus, increasing the risks for banks and may translate into higher interest rates. Thus, bad lending decisions without collaterals could lead to financial bubbles which leaves financial institutions with non-performing assets in the form of unrecoverable loans.

ENDNOTES

1. http://ssir.org/articles/entry/romanticizing_the_poor
2. India emerging a hub for fintech start-ups, Business Standard website, http://www.businessstandard.com/article/companies/india-emerging-a-hub-for-fintech-start-ups-116051700397_1.html
3. https://yourstory.com/2016/07/fintech-startups-2016/
4. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
5. Fintech companies engaged in credit scoring on the basis on alternative data include Gyan-Dhyan, ftcash and Finomena.
6. https://yourstory.com/2016/08/the-parallel-universe-of-indian-fintech/
7. http://www.narendramodi.in/shri-narendra-modi-shares-his-vision-for-digital-india-5944
8. Hussein A Abdou and John Pointon, Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the literature, available at: http://onlinelibrary.wiley.com/doi/10.1002/isaf.325/abstract.
9. Upturn Report https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf
10. http://www.mckinsey.com/business-functions/risk/our-insights/new-credit-risk-models-for-the-unbanked
11. Billie Anderson and Michael Hardin, Credit Scoring in the Age of Big Data, available at http://www.irma-international.org/viewtitle/107257/.
12. https://rbi.org.in/scripts/PublicationReportDetails.aspx?UrlPage=&ID=763.
13. Reliance Industries, Aditya Birla Nuvo, Vodafone, Airtel, Department of Posts, Cholamandalam Distribution Services, Tech Mahindra, National Securities Depository Limited (NSDL), Fino PayTech, Sun Pharma and PayTM were granted in-principles approvals. http://www.thehindu.com/business/Industry/rbi-grants-inprinciple-nod-for-11-payments-banks/article7557908.ece.
14. https://www.rbi.org.in/scripts/bs_viewcontent.aspx?Id=3164#4
15. Radhe Estate Developers Vs. Versus Mehta Integrated Finance Co. Ltd. and Ors, available at https://indiankanoon.org/doc/1357712/
16. http://www.livemint.com/Opinion/dakF6QJNnl23xAVVyOF7qI/Creating-space-for-financial-services.html
17. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
18. The FCRA requires credit bureaus to permit individuals to dispute negative information on their credit reports and to give their own side of the story on disputed information in reports generated for potential creditors, insurers, and employers. It also requires that consumers receive some account of why their credit reports might have led to a denial of credit.
19. U.S. Department of the Treasury, Opportunities and Challenges in Online Marketplace Lending, May 10, 2016, p.20.
20. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
21. http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf
22. http://dbtmis.planningcommission.nic.in/dbtentry/homepage.aspx
23. http://ftalphaville.ft.com/2015/07/31/2135943/when-financial-inclusion-stands-for-financial-intrusion/
24. The Pew Charitable Trust, Key Findings from Fraud and Abuse Online: Harmful Practices in Internet Payday Lending, October 2014.
25. Omidyar Foundation, Big Data Small Credit
26. http://financialservices.house.gov/uploadedfiles/hhrg-114-ba15-wstate-glevi-20160712.pdf
27. http://www.unionbankofindia.co.in/pdf/MudraLoanENGLISHRev180216.pdf.
28. Entrepreneurial Finance Lab at Harvard University, SETTLING THE SCORE: A PSYCHOMETRIC ASSESSMENT THAT UNLOCKS CREDIT TO ENTREPRENEURS IN THE MUSLIM WORLD, available at https://www.usaid.gov/div/portfolio/settling-score-psychometric.
29. LInnet Taylor and Schroder
30. Scott Mason, Nature of Knowledge, Cenre for Internet and Society, available at http://cis-india.org/internet-governance/blog/nature-of-knowledge.

Smart Meters and Big Data

BY VIDUSHI MARDA

Background

Smart Meters are systems that measure and analyse real time energy usage in short intervals and then communicate this information remotely to service providers. Unlike traditional meters, they possess power quality monitoring, real time sensors and power outage notifications. Smart meters analyse power consumption and transmit the data to the service provider who in turn monitor usage and redistribute energy accordingly.[1] Smart meters are not only capable of analysing the energy usage of an electronic appliance but also how long and at what time of the day the appliance is being used.[2] Service providers have been vouching for the installation of such meters in every household by replacing the traditional meters, as this will result in the production of substantial amounts of profits for them.

Smart Meter Schemes and Initiatives

The Government of India has indicated its intention of installing a smart meter in every home, with the goal of being "fully smart" by 2021-2022.[3] Towards this end, government initiatives, public-private partnerships and schemes have arisen.

India Smart Grid Task Force (ISGTF) and India Smart Grid Forum (ISGF)

ISGTF and ISGF were set up by the Ministry of Power in order to facilitate and accelerate the development of programmes/policies related to smart grid technologies in the Indian power sector. These bodies were set up in 2010 for the coordination and integration of smart grid technologies in India, collaboration on the interoperability framework and for making recommendations to the Government regarding policies on smart grid technologies, based on rigorous research.[4]

Smart Grid Vision and Roadmap for India

This followed ISGTF and was rolled out in August 2013 with the vision to "Transform the Indian power sector into a secure, adaptive, sustainable and digitally enabled ecosystem that provides reliable and quality energy for all with active participation of stakeholders."[5] This mission also aimed at developing an indigenous affordable smart meter and to ensure their installation with every consumer by 2027. This vision and roadmap was laid out with a broad aim of furthering and achieving the goals of ISGTF and SGF.[6]

National Smart Grid Mission

The National Smart Grid Mission (NSGM) also works for the advancements of the smart grid projects in the country. This mission undertakes activities including cost-benefit analysis, selecting of technology, pre-feasibility studies among other things which facilitate the formulation of projects. The mission along with other financing agencies and State Discoms also ensures the funding of the formulated projects. It is also largely involved in promoting consumer awareness initiatives. The primary methodology followed for implementation of the projects is to ensure the installation of smart meters in cities in order to establish and further the smart grid project.[7]

Pilot Projects by the Ministry of Power

The Ministry of Power has taken up a pilot project in order to assist the establishment of smart grids and has shortlisted fourteen projects across distribution companies in various states. This step was taken based on the recommendations made by the ISGTF. The Government aims to fund these projects and the amount of funding to be received by each project will be demarcated in the 12th plan which will be brought out in 2017.[8] As a part of this, Chinese major Dong Fong has been selected among the bidders by the Puducherry State Government for the implementation of the smart grid project. The cost of the project was estimated to be Rs. 46.1 Crores with the central government funding 50 per cent of the project.[9] Similarly, Swedish telecom provider, Ericsson has entered into a contract with a public sector power company for the installation of 15,000 smart meters over the next 3 years in Guwahati, Assam.[10]

With a steep increase in the implementation of smart grid projects, the need for certain regulations was felt and in the year 2015, the Forum of Regulators formulated the Model Smart Grid Regulations. This model laid down the rules and regulations to be followed in the process of implementing the smart grid projects. It also provides for the constitution of a smart grid cell and the roles and responsibilities of the same.[11] The model also provides the process to be followed while evaluating and monitoring the smart grid projects in order to ensure efficiency and cost management.

Ujwal Distribution Company (DISCOM) Assurance Yojana

The Government of India launched the Ujwal DISCOM Assurance Yojana in order to provide financial assistance to the power distributors across the country.

Apart from reducing costs of power, improving efficiency and enforcing financial discipline on DISCOMS, this scheme made the installation of smart meters compulsory for consumers using more than 200 units of electricity in one month in order to effectively and efficiently distribute electricity to all consumers and also avoid wastage.[12]

Applying Section 43A of the IT Act to Data Collection through Smart Meters

The data that is being mapped by smart meters reaches the Data Concentrator Unit followed by the Meter Data Acquisition System and Meter Data Management System which collect, analyse, authenticate and manage the data thus received.[13] All these units are components of the Advanced Metering Infrastructure (AMI) which has been used by the Indian Government for implementing the smart grid project.[14] The ISGTF and ISGF under the Ministry of Power manage the AMI and therefore have access to the data collected by the smart meters. Section 43A of the IT Act applies to bodies which includes firms, companies or any association of individuals who carry on any commercial or professional activity.[15] However, the ISGTF and ISGF have been set up under the Ministry of Power for aiding the smart grid project and hence cannot be categorized into any of the bodies provided for by the Section. Their activities do not fall under commercial or professional activities as it is a government initiative and scheme for the establishment of a smart grid through installing smart meters in place of the traditional meters. The Section is worded in such a fashion so as exclude government bodies and their subsidiaries as they are not corporate bodies and do not engage in commercial or professional activities.

Benefits of Smart Meters

1. Efficient Demand Side Management- The data collected enables the suppliers to monitor power consumption and regulate it, thus minimizing the costs of distributing energy and firing up of reserve generators. It has been estimated that the setting up of these meters in India will reduce Average Technical and Commercial losses from 22 per cent to 15 per cent and eliminate the growing gap between Average Revenue Realized and Average Cost of Supply by 2019.[16] Smart meters help consumers understand the impact of their energy usage, thus incentivizing them to minimize the use of appliances whenever possible which will ultimately result in lower electricity bills.[17]

2. Accurate shaping of the market and industry- Smart meters facilitate the functioning of smart grids and provide data as to the amount of extra energy available along with whether the grid has to capacity to transfer such energy. These meters also have the capacity to determine the time and place of power failures. This helps the suppliers to better allocate the available resources in order to arrive at the most efficient, economically viable solution.[18] Such an ascertainment helps energy suppliers with better load management, thus resulting in a regulated supply of power.[19]

3. Security of Supply- Smart meters act as predictive tools and have the ability to determine shortage of power, power failure, excess usage of power by consumers etc. This helps the suppliers to provide and distribute energy accordingly. This will ensure a systematic and sustained supply of energy to the consumers and avoid energy overflows. These meters will also enable the suppliers to redirect energy to places where there is a shortage and prevent excess supply to areas which have sufficient supply.

Harms of Smart Meters

1. Unintended Behavioral Analysis Tool- Due to smart meters, energy suppliers will be in a position where they have an easy access to the daily life patterns of an individual.[20] The smart meters can analyse how many times an individual uses the microwave, the number of hours for which she watches the television, the kind of television owned by her, the time taken by her to take a shower, the time at which she leaves the house and the time at which she comes back among various other things that can be deduced from the patterns of energy consumption.[21] These meters eventually end up extracting data which is not confined to energy usage but extends to the behavioural patterns of an individual citizen.[22] Large scale installation of smart meters provides energy suppliers with vast amounts of data about the habits and routines of all the consumers in the state or country.[23] These suppliers will be in possession of data which might not have been consented to be provided by the consumers, thus placing them in a risky position where the suppliers might use such data to the detriment of the consumers. This technological invention runs parallel to the functioning of the SCADA technology which acquires data for the purposes of control and supervisory management.[24] This technology is also being used by the Ministry of Power in its smart cities project which primarily involves the setting up of smart meters across households in order to control and manage the consumption of energy. This has been planned for the twenty smart cities that the MoP has identified for the first phase of the smart cities project.[25]

2. Social Polarization- The Department of Energy and Climate Change estimated the cost of installing a smart meter at £ 214.50 per household.[26] This gets passed onto the consumers along with other network costs.[27] This initial investment in smart meters is not affordable to all classes of people. Further, the returns that can be expected from the installation of such meters has been estimated to be disproportionately low compared to the costs incurred for setting up the same.[28] Therefore, people who are not so affluent will avoid the installation of smart meters in their households and hence will be polarized. The people who can afford the smart meters will be within the

circuit and will be able to communicate their patterns of energy usage to the suppliers, thus enhancing power supply and ensuring a safe supply of energy as per their requirements. Although the smart grid programmes do not inherently discriminate between various sections of the society, the cost of this technological innovation is resulting in an invisible polarization of the have-nots in terms of efficient energy usage and supply.

3. Social Dumping- The marginally poor and low-income groups are often identified with indebtedness, low energy consumption, non-payment of electricity bills which eventually leads to disconnection from supply. [29] All of these characteristics and aspects raise serious problems and costs of utilities. Initially, the solution to this was the coin operated prepayment meter, but additional costs for utility were created due to problems of theft, fraud and collection. The smart meters which are now being installed in the households work on the prepayment model and this provides the citizens with the choice to disconnect at the beginning of the month. [30] Therefore, the lower income groups who cannot afford to pay the bills at the beginning of the month have to choose to disconnect themselves before they get disconnected due to nonpayment. [31]

4. Smart Meters as Big Data- As extensively discussed before, smart meters are capable of transmitting data that the consumer might not be willing to otherwise make public. The technology with which the smart meters have been built enables them to determine the number of hours for which an appliance is being used along with the amount of energy consumed by it. It is also capable of ascertaining which appliance is being used at what point of time in the day. This indirectly helps in mapping the behavioural patterns of an individual and how she or he functions on a daily basis. [32] This is an intrusion of an individual's privacy as most of them are unaware of the kind of data that can be mapped from the smart meters. [33] Further, there is a lot of ambiguity about who receives this data or will be in possession of it, once it has been transmitted by the smart meter. Who ever is on the receiving end of this data will be in possession of voluminous amounts of details about a person's daily life routines and patterns without their knowledge. [34] Such possession of data can be used by the possessor to the detriment of the consumers or can be sold to a third party without the consent of the consumer. [35] Therefore, in the absence of data protection laws, the idea of smart meters comes off as a tool which invades the privacy of an individual as it is not only collecting data about energy consumption, but such data is further revealing other activities and patterns of the person's life. [36]

Conclusion

While the project of implementing smart meters across India is a step further to harness the potential of big data in supplying energy and the project is lucrative as it assures efficiency in the consumption and distribution of energy, these smart meters can deliver to their full capacity only when allowed to collect significant amounts of data from the consumers which essentially means large scale invasion of privacy without any adequate data protection laws. [37] Smart meters are able to draw such maps based on consumption by analysing the appliance load signatures. However, some models to hide these application load signatures have been proposed: they include using rechargeable batteries which can be used to hide the appliance load signatures. [38] Power mixing algorithms and power management models can also be used to minimize the data that can be extracted by a smart meter if installed. Another issue that needs to be addressed is the control over the information that has been received through the smart meters.

ENDNOTES

1. http://www.dailymail.co.uk/news/article-2315863/Big-brother-switch-fridge-Power-giants-make-millions--pay-sinister-technology.html
2. http://news.nationalgeographic.com/news/energy/2012/12/121212-smart-meter-privacy/
3. We should look at a smart meter in every home,: Piyush Goyal, The Hindu BusinessLine, http://www.thehindubusinessline.com/specials/clean-tech/we-should-look-at-a-smart-meter-in-every-home-piyush-goyal/article8245384.ece, Feb 16, 2016.
4. http://www.indiasmartgrid.org/about-us.php
5. http://www.cyantechnology.com/wp-content/uploads/2014/04/Smart-Grid-Vision-and-Roadmap-for-India.pdf, page 8.
6. http://www.cyantechnology.com/wp-content/uploads/2014/04/Smart-Grid-Vision-and-Roadmap-for-India.pdf
7. http://www.nsgm.gov.in/content/nsgm.php
8. https://www.metering.com/14-smart-grid-pilot-projects-shortlisted-in-india/
9. http://www.thehindu.com/news/cities/puducherry/smart-grid-project-back-on-track/article8036407.ece
10. https://www.metering.com/news/ericsson-bags-smart-meter-deal-in-india/
11. http://www.indiasmartgrid.org/reports/FOR%20-%20Smart%20grid%20regulations.pdf
12. http://pib.nic.in/newsite/PrintRelease.aspx?relid=130261
13. http://www.isgw.in/wp-content/uploads/2016/03/ISGW-2016_Surendra-Acharya_Smart-Meter-to-Smart-Hubs-–-AMI-Rollout-Strategies-for-India_17-March-2016.pdf
14. http://www.indiasmartgrid.org/Advanced-Metering-Infrastructure.php
15. Section 43A, IT Act.
16. http://pib.nic.in/newsite/PrintRelease.aspx?relid=130261
17. https://www.theguardian.com/british-gas-smart-meter-challenge/2014/sep/25/new-smart-meters-offer-a-host-of-benefits

18. http://www.cpuc.ca.gov/General.aspx?id=4853
19. http://www.cpuc.ca.gov/General.aspx?id=4853
20. CHERYL DANCEY BALOUGH, Privacy Implications of Smart Meters, 86 CHI.-KENT L. REV. 161, 166 (2011).
21. BRANDON J. MURRILL ET AL., CONG. RESEARCH SERV., R42338, SMART METER DATA: PRIVACY AND CYBERSECURITY 3 (2012), available at http://www.fas.org/sgp/crs/misc/ R42338.pdf.
22. http://www.forbes.com/sites/federicoguerrini/2014/06/01/smart-meters-friends-or-foes-between-economic-benefits-and-privacy-concerns/#341c103d51a9
23. http://news.nationalgeographic.com/news/energy/2012/12/121212-smart-meter-privacy/
24. Towards Privacy Protection in Smart Grid, https://www.nics.uma.es/pub/papers/1750.pdf
25. http://www.indiasmartgrid.org/smartcityproject.php
26. http://www.telegraph.co.uk/finance/personalfinance/energy-bills/11975065/Smart-meters-will-cost-11bn-but-youll-be-lucky-if-yours-saves-you-30.html
27. http://www.telegraph.co.uk/finance/personalfinance/energy-bills/11975065/Smart-meters-will-cost-11bn-but-youll-be-lucky-if-yours-saves-you-30.html
28. http://www.telegraph.co.uk/finance/personalfinance/energy-bills/11975065/Smart-meters-will-cost-11bn-but-youll-be-lucky-if-yours-saves-you-30.html
29. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.6693&rep=rep1&type=pdf
30. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.6693&rep=rep1&type=pdf
31. https://www.ofgem.gov.uk/sites/default/files/docs/2006/03/13383-sustainability_first_response__0.pdf
32. http://news.nationalgeographic.com/news/energy/2012/12/121212-smart-meter-privacy/
33. http://www.politico.com/story/2015/01/energy-electricity-data-use-113901
34. http://www.bbc.co.uk/news/technology-18407340
35. http://www.forbes.com/sites/federicoguerrini/2014/06/01/smart-meters-friends-or-foes-between-economic-benefits-and-privacy-concerns/#341c103d51a9
36. Samuel J. Harvey, Smart Meters, Smarter Regulation: Balancing Privacy and Innovation in the Electric Grid, http://www.uclalawreview.org/pdf/61-6-10.pdf
37. Id
38. Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures, http://ieeexplore.ieee.org/document/5622047/

Intelligent Transport System

BY VANYA RAKESH

Introduction

With efficient urban mobility and public transport stated as an element of a Smart city and Intelligent transport management system listed as a viable solution for transport sector in the 100 Smart Cities Mission guidelines, [1] this case study aims to study 5 key ITS projects in Indian cities, in light of the benefits and harms such a project encompasses, with large amounts of data being collected and aggregated.

As per a World Bank study, by 2031, some 600 million people are expected to live in India's cities. However, only about 20 Indian cities with populations over 500,000 have any kind of organized public transport systems. [2] As of November 2015, The current urban bus occupancy in various Indian State Transport Units (STUs) is <55% and is reducing every year. Reasons: unavailability of buses, long waiting time, operational inefficiencies. [3] World bank reports that the economic losses incurred on account of congestion and poor roads alone run as high as $6 billion in India. [4] According to Census 2011, there are a total of 497 cities in India, with around 90 cities with a population over 5 lakh. [5] These statistics highlight the need for an organised and efficient solution for transportation in Indian cities.

The Intelligent Transport Systems (ITS) is a term for the integrated application of communications, control and information processing technologies to monitor and manage a transportation network, [6] which enables the gathering of data and intelligence, which is then analyzed and the derived insights are shared back to traffic managers and road-users. [7] Such systems include stand-alone applications as well as cooperative ITS applications involving vehicle to infrastructure or vehicle to vehicle communications. [8] The main objective of ITS is to improve commuter service as well as decision making, achieve traffic efficiency, enhance safety for commuters, drivers, etc., conserve energy and enhance overall performance and profitability of the service provider corporation. [9] The components of ITS usually comprise of network of sensors, connected vehicles, GPS tracked public transportation, dynamic traffic lights, passenger information panels, automatic number plate readers, CCTV systems, navigation facilities,etc. [10] In India, some of the ITS projects have been initiated in cities like Mumbai, Bangalore, Ahmedabad, Pune, Mysore [11] and states like Rajasthan, Maharashtra, Andhra Pradesh and Uttar Pradesh.

Sources of big data in transport

1. **Vehicle Tracking System (VTS)**: A VTS Usually comprises of the Vehicle Tracking Unit (VTU) and Voice-kit. The VTU shares real-time location information , along with vehicle number and shift number (of the driver and conductor of the bus concerned). This Unit transmits the data via mobile (GPRS) network, and can also be used by the driver to inform the ITS Control Room, that is the control room of the ITS, in case of an emergency.

2. **Electronic Ticketing System (ETS)**: Comprises of the Electronic Ticketing Machines (ETM) and the Depot Application. The ETM transmits ticket data to the ITS via mobile (GPRS) network, and includes information regarding the bus stop at which a particular ticket was issued, the ticket amount (proxy for destination bus stop), details of the bus, timestamp, etc. Depot application is installed at all bus depots, which generates information on duty rota, the log sheet for the driver, kilometres performed and details about fuel usage, etc. This information is sent to the desktop application at the ITS Control Room, which generates schedules and timetables on the basis of that. This data is also fed in the ETM, which is used by the bus conductors and is issued on the duty start date.

3. **Passenger Information System (PIS)**: This is the bus location information communication system, powered by the real-time data being generated by the VTS. The passenger is given information about buses going from that stop/ location, the destination, the route via which the bus will go, the estimated time of arrival, etc. This information is generated by integrating information from the vehicle tracking unit installed in all buses.

4. **Mobile Application:** The VTU enables sharing of location of buses to show the estimated time of arrival (ETA) of a bus, along with bus routes, bus number, destination of the bus.

5. **Call Data Records (CDR)**: These records provide the location, time and duration of every call, generating huge databases about daily transactions performed by ordinary people. Every time a call is made on a mobile phone, the mobile tower associated with that call – usually the nearest mobile tower – is recorded, originally for billing purposes. Studies [12] suggest that CDRs could become a valuable tool for transport planners and traffic management by monitoring average number of daily passengers, congregations of people, peak and off-peak travel patterns, communities, and traffic. [13]

Stakeholders with whom such data is/can be shared include : passengers, urban planners, traffic police, Transportation companies, Trip-planning web-sites and portals, urban local development bodies.

ITS for smart solution in bus services in India

In India, some of the ITS projects have been initiated in cities like Mumbai, Bangalore, Ahmedabad, Pune, Mysore and states like Rajasthan, Maharashtra, Andhra Pradesh and Uttar Pradesh. This case study maps the key features/ solutions, scale of operations and cost of project for 5 states and cities across the country.

STATE/CITY PROJECT	KEY FEATURES AND SOLUTIONS	SCALE OF OPERATIONS
Uttar Pradesh State Road Transport Corporation (UPSRTC) [14]	• Integrated Ticketing System (ITS) • Vehicle Tracking System (VTS) • Passenger Information System (PIS) • MIS analytics & Decision Support System [15] • eTicketing Solution • Smart Card RFID Pass system • Public Online Reservation system – Through Self Service Portal • Online Reservation Solution – For UPSRTC counters and Agents [16]	• It has a fleet size of over 9500 buses plying over 32 lakh kilometers per day and carrying over 15 lakh passengers. • the capital cost of the project is Rs. 38.25 crores of which 50% (Rs. 19.125 crores) is being funded as a grant under ACA from Government of India (GoI). [17]
Rajasthan State Road Transport Corporation (RSRTC) [18]	• Electronic ticketing issuance system • Online reservation system.	• RSRTC has a fleet of over 4700 buses and has deployed 6000 conductors. • More than 10 lakh passengers (on an average) use the RSRTC services daily. [19]
Maharashtra State Road Transport Corporation (MSRTC) [20]	• E-ticketing services. • Smart Card RFID Pass system • Mobile Booking Solution • Public Online Reservation system – Through Self Service Portal • Online Reservation Solution – For MSRTC counters and Agents [21]	• 16,500 buses which ferry 7 million passengers daily on 18,700 routes. • Cost : in the year 2009,) a sum of Rs 25 crore was invested in Trimax, which was utilised for MSRTC project. [22]
Andhra Pradesh State Road Transport Corporation	• VT & PIS will cover core systems such as Vehicle Tracking System, Real Time Passenger Information System and Central Control Centers. [23]	• The project is planned to cover about 3,502 buses • The Corporation's buses cover 83.19 lakhs KMs. and carry 136.13 lakhs people to their destinations every day. They connect 23,388 villages to all major towns and cities in A.P which constitutes 95% of Road Transport. [24]

STATE/CITY PROJECT	KEY FEATURES AND SOLUTIONS	SCALE OF OPERATIONS
Bengaluru Metropolitan Transport Corporation (BMTC)	The Intelligent Transport System (ITS) comprises of • Vehicle Tracking System (VTS), • Passenger Information System (PIS), and • Electronic Ticketing System (ETS), including a mobile application that provides passengers with information about buses available from a particular stop/location, the route of the buses, the estimated time of arrival, and allows for trip planning. [25]	• Scale of BMTC's operations, which includes 6,404 buses and 6,216 schedules, 5,200,000 daily passengers, 2,400 routes and over 75,000 trips, and 1,290,000 KMs of daily service. • Cost of project: The total cost of this Build-Operate-Transfer (BOT) project is Rs 79 crore. [26] • The ITS project in Bangalore saw integrated implementation of over 10,000 internet-enabled electronic ticketing machines, and over 6,400 vehicle tracking units.

Global Concerns and Initiatives

The International Transport Forum at the OECD is an intergovernmental organisation which acts as a think tank for transport policy that improve people's lives. In its report [27] titled "Big Data and Transport- Understanding and assessing options", several concerns that have been listed include : difficulty in anonymising location and trajectory data, lagging data protection policies in comparison to new modes of data collection and uses, the need to design more effective protection of location data upfront into technologies, algorithms and processes, introduction of new models of public-private partnership involving data-sharing to leverage all the benefits of big data and the challenge of enacting meaningful consent from citizens and consumers.

Also, the National Transport Commission (NTC) in Australia released a discussion paper [28] to examine the policy implications of Cooperative Intelligent Transport Systems (C-ITS). According to the paper, the key issues that affect regulatory policy include : human factors (for example: Humans can often behave in complex and unpredictable ways, whereas technology behaves as programmed, so the interaction of the two can lead to unforeseen results. Also, if users are not aware of a system's capabilities and limitations they can overestimate that capability), mix of old and new technology in the transport system and mixing different road users (example : C-ITS enabled vehicles will need to be tolerant of other road users that may be not have C-ITS, in particular cyclists and pedestrians), data accuracy, security and anonymity, liability, whether C-ITS applications operate as warning systems or trigger automated interventions.

The European Union has also launched major initiatives [29] to overcome the slow and fragmented uptake and deployment of ITS in road transport. The European Commission's ITS Action Plan and — in the form of the ITS Directive — dedicated EU legislation on ITS together constitute a concerted policy framework to boost ITS across Europe. [30] Some of the challenges that have been identified are: defining the roles of the public and private sectors as well as rules for cooperation on data exchange, content and service provision, Rules in EU countries on the collection of road and traffic-regulation data have been uneven and often completely lacking, to ensure easy access to the digital road databases maintained by thousands of European road authorities in a standardised, non-discriminatory and transparent way, to ensure the integrity, confidentiality and secure handling of data, including personal and financial details, and show that citizens' rights are fully protected, to list a few. [31]

Benefits

For Organisation

1. **Reduce project costs**: Big data presents opportunities to identify problems, analyse and reduce project costs, leading to better transportation asset utilization and making the most of transportation structure. Also, ITS is intended to help in identification of pilferage, improve revenue management, and reduced fuel consumption.
2. **Incident management**: Tracking location of buses and passengers enables preventative maintenance and avert potential threat. Voice kit enables instant communication to inform about mishaps. Pinpointing locations of accidents or vehicle breakdown is important to handle the emergency situations.
3. **Promotes reliability on transports**: Increased efficiency and improved forecasting promotes reliability on transports. Reduced waiting time for passengers helps in increasing the passenger base.
4. **Reduce traffic congestion by optimization of routes**: providing actual data for route planning and route rationalization by way of VTUs can help in reducing traffic congestion on roads.

For Consumers/Society

1. **Improved user experience**: Components like ETS help in demand prediction, which helps in improving user experience by increasing fleet of buses in areas where required.
2. **Targeted services**: Recognizing traffic patterns by investigating real time data can help transport authorities to understand commuters' behaviour, provide targeted information and identify policy interventions.
3. **Reduction in traffic congestion in cities**: Through big data, the smart city will be able to reduce traffic and accidents by opening new roads, enhancing the infrastructure based on congestion data, and collecting information on car parking and alternative roads.
4. **PIS**: PIS enables passengers plan their trips, select routes and save time. Hence, improved end-to-end customer experiences. Information about arrival of public transport helps in choice of travel mode and reduces wait delays.

Harms

1. **Decision making and transparency**: There can be a lack of transparency about decision making based on big data and analytics that could lead to unintended consequences.
2. **Privacy and data security**:
   - Over-collection of PII, uncertain and unclear use of the data : The purpose of collection of personal information about an individual for allowing them the use of the mobile apps to help them get information about the location and route of a bus on one hand, and to enable the organisation improve decision-making on the other hand, is unclear. The over-collection of information violates the principle of data minimization and poses a risk of identification of the individuals, with the challenge of anonymising data.
   - No or inadequate Privacy Policies: In this case, no or inadequate privacy policies or established principles and guidelines to regulate ICT and big data usage reflects inadequate data security measures, which may lead to identification of individuals from the database of the companies.
   - No opt-out: the compulsory provision of data as a prerequisite for the access and use of many key online services is making opting-out of data collection impossible, highlighting the need for designing measures in light of new technologies with the challenge of enacting meaningful consent from citizens and consumers. [32]
3. **Unplanned use of data**: like in case of BMTC, the officials are not clear about the plans to utilise the big data potential of their project, even when the project has been implemented across the city and data collection has begun. This reflects planning and designing without clear consideration of the impact of big data in such projects.
4. **Lack of accountability and transparency**: Since the tenders for ITS projects are awarded to private companies for developing the system, lack of information regarding data ownership and management available publically may make it difficult to hold an entity in a PPP model accountable or responsible in case the huge amount of data so stored and collected is compromised. For example, in case of BMTC, the officials revealed that the CFP for awarding the project was available on the website only for a specific duration and was not available for anyone to access it. This also reflects lack of transparency.
5. **Data Quality**: Data captured by different systems and people and stored in distinctive databases can result in inaccurate data or poor data quality as collection and processing methods may vary. Further, the data is rarely stored in standard and interoperable formats. Entering inaccurate or low-quality data from unreliable data sources in the system may lead to revenue loss and process inefficiency, where the analysis and decisions based on such data may be unreliable. This may also lead to exclusion and bias.
6. **Exclusion**: Though civic apps do help passengers inform about location and time of arrival of buses may increase engagement and make the system more efficient, their capacity to bridge social divides remains unclear. The potential of smartphone apps to address social problems of transport management should not be overstated. [33]

Recommendations

1. Need for greater transparency about role of private players and accountability. This requires defining the roles of the public and private sectors as well as rules for cooperation on data exchange, content and service provision.
2. Need to design more effective protection of data in comparison to new technologies- comprehensive data privacy policies and adoption of suitable technical standards ensures interoperability of technologies and ensures that services are safe, reliable and of good quality. For example-ISO/TC 204 - Intelligent transport systems. [34]
3. Need for a clear and wider dialogue and framework on smart solutions to analyse benefits and harms of the project.
4. Need for increased co-operation and coordination between government departments to facilitate decision-making.
5. Involved entities must establish guiding principles of openness, transparency, participation, and collaboration to keep the exchange and flow of big data

under control. Consideration of the impact of big data and its consequences to formulate comprehensive policies and regulations to control and regulate collection, sharing, usage of data by IT solution providers for smart solutions in transportation.

6. As big data smart city applications involve large scale heterogeneous systems and data, it is advantageous to follow an open standard for designing and implementing such solutions. This will add flexibility for upgrading, maintaining, and adding more application features for smart cities.

7. Citizens must be aware of how to use ICT solutions for smart city correctly and safely. Their active participation in providing information related to the different issues they may encounter with smart city applications will help in enhancing the quality of collected data and the performance of the applications.

ENDNOTES

1. Mission Statement and Guidelines, http://smartcities.gov.in/writereaddata/smartcityguidelines.pdf
2. http://articles.economictimes.indiatimes.com/2015-09-02/news/66144023_1_smart-cities-public-transportation-bicycle-sharing-system/2
3. http://www.kpit.com/resources/downloads/whitepapers/realizing-smart-cities-technology-whitepaper.pdf
4. https://coeut.iitm.ac.in/ITS_synthesis.pdf
5. Cities having population 1 lakh and above, Census 2011, http://censusindia.gov.in/2011-prov-results/paper2/data_files/India2/Table_2_PR_Cities_1Lakh_and_Above.pdf
6. http://www.ectri.org/YRS07/Papiers/Session-15/Efimenko.pdf
7. http://meity.gov.in/content/intelligent-transportation-system-its
8. http://pearl.niua.org/sites/default/files/books/GP-IN1_UT.pdf
9. http://www.rakshakfoundation.org/wp-content/uploads/2015/01/2014_D141.pdf
10. http://unctad.org/meetings/en/SessionalDocuments/CSTD_2015_Issuespaper_Theme1_SmartCitiesandInfra_en.pdf
11. http://pearl.niua.org/sites/default/files/books/GP-IN1_UT.pdf
12. http://blogs.worldbank.org/transport/files/2014_trb_cdr_boston_rio.pdf
13. http://www.scidev.net/south-asia/governance/analysis-blog/big-data-can-make-south-asian-cities-smarter.html
14. http://www.upsrtc.com/
15. http://www.rakshakfoundation.org/wp-content/uploads/2015/01/2014_D141.pdf
16. http://trimax.in/resource-center/success-stories/transport/
17. http://telematicswire.net/an-overview-of-upsrtc-intelligent-transport-management-system-itms/#sf3vGS0lHlzrpi3x.99
18. http://transport.rajasthan.gov.in/rsrtc/
19. http://trimax.in/doc/RSRTC.pdf
20. https://www.msrtc.gov.in/
21. http://trimax.in/resource-center/success-stories/transport/
22. http://www.moneycontrol.com/news/business/trimax-may-go-public-to-raise-rs-250-300cr-this-fiscal_557585.html?utm_source=ref_article
23. http://telematicswire.net/vehicle-tracking-passenger-information-system-vt-pis-in-apsrtc-a-case-study/#wM1Z2RM2DVvm7dHl.99
24. http://telematicswire.net/vehicle-tracking-passenger-information-system-vt-pis-in-apsrtc-a-case-study/#3DBzEhxeSEdFZqgl.99
25. http://bangalore.citizenmatters.in/articles/bmtc-new-year-gift-to-bengaluru-intelligent-transport-system
26. http://bangalore.citizenmatters.in/articles/bmtc-new-year-gift-to-bengaluru-intelligent-transport-system
27. http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf
28. https://www.ntc.gov.au/Media/Reports/(BB07ED22-C3D7-3B62-8455-C8A1E31D6374).pdf
29. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52014DC0642
30. More documents available at : http://ec.europa.eu/transport/themes/its/road/action_plan/its_reports_en.htm
31. http://www.gppq.pt/h2020/_docs/brochuras/transportes/intelligent%20transport%20systems%20in%20action.pdf
32. http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf
33. https://articles.extension.org/sites/default/files/Smart%20Phones%20and%20Immigrant%20Integration.pdf
34. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=54706

Big Data Regulation & De-regulation Agenda

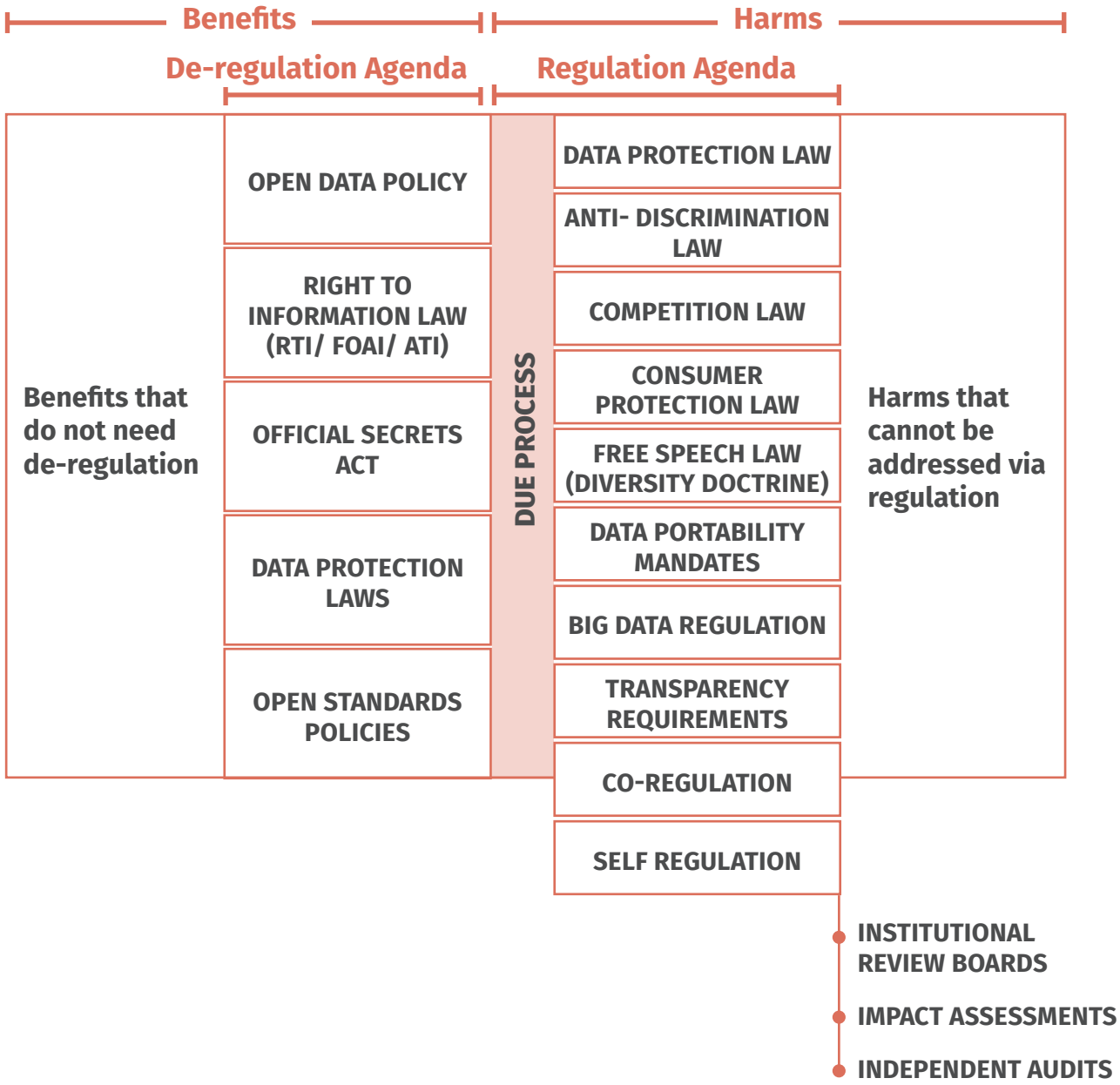
BY SUNIL ABRAHAM

While personal information, including personally identifiable information and sensitive personal information is governed by data protection law, personal information in the public domain is not. A potential research question is to understand how two competing areas of policy, i.e data protection and open government data policy govern them.

PI is the superset of personally identifiable information, sensitive personal data and any other data that can be associated with a person. Personally Identifiable Information (PII) includes any information that relates to an identifiable, living person which can be used to identify that person. This includes unique identifiers, direct identifiers and indirect identifiers.

Personal Information in the public domain refers to personal information voluntarily disclosed to the public by the data subject implying that an individual willingly forgoes his/her right to data protection and privacy. In most jurisdictions including India and the UK, such information is not regulated by data protection law. Personal information where the identifiers are replaced with pseudonyms or pseudo-identifiers is a way in which data controllers try to leverage PI without the associated harms to the data subject. Anonymised information is personal information where the identifiers have been eliminated.

Anonymous and pseudonymous information effectively exist in a regulatory vacuum between personal data policies and open data policies which leads to a reconsideration of the scope of data protection law. Some harms and some benefit may occur regardless of regulation, however, both regulation and deregulation are needed to solve the optimization problem of big data ie. unlocking benefits whilst mitigating harms through both substantial and procedural laws. Some regulatory options maybe developed ground up as self-regulatory or co-regulatory standards.



An Agenda for Big Data Regulation

BY AMBER SINHA

The ubiquitous nature of big data analytics and questions around its conspicuousness and transparency, its growing presence in disparate sectors like medicine, credit, healthcare, law enforcement, justice system and public policy, to name a few, and the unique questions it raises about existing legal systems (or lack thereof) presents a significant regulatory challenge.[1] So far most attempts to articulate a suitable regulatory response to big data are limited to an examination of data protections law, giving greater control to citizens over their data and holding companies engaged in big data more accountable.[2] However, most of these approaches seem to frame the question of big data issues solely as an impediment to privacy and thereby framing the solution also in terms of furthering the privacy as an information control argument and through giving individuals more rights over their data. This approach ignores the implicit manner in which data is harvested and inferred so much so that it may render individual control through informed consent meaningless, and also fails to take into account the various other implications of use of big data by framing the debate solely around privacy. In this section, we try to articulate a more multi-pronged approach to governing big data which takes into account its various implications, as well as the different means of regulating it. Given the pervasive and wide-ranging impacts of big data, we also advocate for a more multidisciplinary approach to regulate it which leverages social science research and technological solutions.

Law

A. Data Protection Law

A study conducted by Dean, Kalapesi and Rose in 2013 concludes that that for over 75 percent of consumers in a majority of countries, the privacy of personal data is the most important issue with regard to the phenomenon of big data. Various authors have pointed at the inefficacy of traditional data protection frameworks to address the privacy challenges of big data[3] and how data protection regulations need to be reconfigured.

The focus of most data protection frameworks globally is towards empowering individuals with information to enable them with greater choice in decisionmaking and disclosure of data. However, this framework has not been able to ensure rational choice by data subjects.

Therefore, there is a need for the focus of data protection regulation to move beyond the stage of collection to the actual use of data. There is a need for risk assessment frameworks which develop a methodology for organisations to apply, calibrate and implement abstract privacy obligations based on the actual risks and benefits of the proposed data processing. This would mean addressing the question whether "there is a significant likelihood that an identified threat could lead to a recognised harm with a significant degree of seriousness." This would require a matrix of threats and harms and an analysis of how threats can lead to tangible harms. Fred Cate has articulated the need to weigh benefits maximization against harms and the regulation must rest on the recognition that both information flows and individual privacy have value and are necessary in a democratic society and market economy.

B. Anti-discrimination Law

A growing area of research globally is the social consequences of big data with a particular focus on its tendency to replicate or amplify existing and structural inequalities. Problems such as data invisibility of certain excluded groups,[4] the myth of data objectivity and neutrality,[5] and data monopolization[6] contribute to the disparate impacts of big data. So far much of the research on this subject has not moved beyond the exploratory phase as is reflected in the reports released by the White House[7] and Federal Trade Commission[8] in the United States. The biggest challenge in addressing discriminatory and disparate impacts of big data is ascertaining "where value-added personalization and segmentation ends and where harmful discrimination begins."[9]

Some prominent cases where big data can have discriminatory impact are denial of loans based on attributes such as neighbourhood of residence as a proxies which can be used to circumvent anti-discrimination laws which prevent adverse determination on the grounds of race, religion, caste or gender, or adverse findings by predictive policing against persons who are unfavorably represented in the structurally biased datasets used by the law enforcement agencies. There is a dire need for disparate impact regulation in sectors where see the merging use of big data, and these regulations must address the following three issues:

a) Erroneous inferences which adversely impact individuals
This is usually a result of a poorly represented data. In India, there is huge challenge due to the issue of dark data. Only 16% individuals leave a digital footprint.[10] Large parts of the population is unrepresented in digital data. Further, the veracity and accuracy of the data available remains a big issue because of the complexities of local contexts makes it hard to devise a classification system which is accurate.

b) Adverse determinations which amplify existing inequalities
Even in cases where the inferences drawn are correct, they can lead to adverse determinations made against underprivileged classes. E-scores about individuals can inform the level of service offered to them, prices discrimination, insurance premium and access to credit. [11]

c) Use of data proxies to discriminate against protected classes
Anti-discrimination laws typically prevent adverse determinations on the basis of identified factors such as race, religion, caste and gender. Use of data points such as residential address, usage patterns and location data are often used as indicators of financial status, health, community etc. Adverse determinations such as denial of loans, price discriminations etc. on the basis of these indicators circumvent the anti-discrimination laws as these determinations are ostensibly not based on attributes protected by law.

C. Competition Law

The conversation on use of competition or antitrust laws to govern big data is still at ane early stage. However, the emergence of numerous data driven mergers or acquisitions such as Yahoo-Verizon, Microsoft-LinkedIn and Facebook-WhatsApp have made it difficult to ignore the potential role of competition law in the governance of data collection and processing practices. It is important to note that the impact of big data goes far beyond digital markets and the mergers of companies such as Bayer, Climate Corp and Monsanto shows that data driven business models can also lead to the convergence of companies from completely different sectors as well. So far, courts in Europe have looked at questions such as the impact of combination of databases on competition [12] and have held that in the context of merger control, data can be a relevant question if an undertaking achieves a dominant position through a merger, making it capable of gaining further market power through increased amounts of customer data. The evaluation of the market advantages of specific datasets has already been done in the past, and factors which have been deemed to be relevant have included whether the dataset could be replicated under reasonable conditions by competitors and whether the use of the dataset was likely to result in a significant competitive advantage. [13] However, there are limited circumstances in which big data meets the four traditional criteria for being a barrier to entry or a source of sustainable competitive advantage — inimitability, rarity, value, and non-substitutability. [14]

Any use of competition law to curb data-exclusionary or data-exploitative practices will first have to meet the threshold of establishing capacity for a firm to derive market power from its ability to sustain datasets unavailable to its competitors. In this context the peculiar ways in which network effects, multi-homing practices and how dynamic the digital markets are, are all relevant factors which could have both positive and negative impacts on competition. There is a need for greater discussion on data as a sources of market power in both digital and non-digital markets, and how this legal position can used to curb data monopolies, especially in light of government backed monopolies for identity verification and payments in India.

D. Consumer Protection Law

The Consumer Protection Bill, 2015, tabled in the Parliament towards the end of the monsoon session has introduced an expansive definition of the term “unfair trade practices.” The definition as per the Bill includes the disclosure “to any other person any personal information given in confidence by the consumer.” This clause excludes from the scope of unfair trade practices, disclosures under provisions of any law in force or in public interest. This provision could have significant impact on the personal data protection law in India. Alongside, there is also a need to ensure that principles such as safeguarding consumers personal information in order to ensure that the same is not used to their detriment are included within the definition of unfair trade practices. This would provide consumers an efficient and relatively speedy forum to contest adverse impacts on them of data driven decision-making.

E. Cross Border Agreements

There is growing trend of clauses in cross-border agreements that prevent signatory countries from mandating source code disclosure as a condition for import, distribution, sale or use of such software. Currently, Trans-Pacific Partnership, Trade in Services Agreement (TISA) and Regional Comprehensive Economic Partnership (R-CEP) are some of the agreements which have such clauses. While such measures have been criticised for promoting proprietary technologies, with increased algorithmic decision making, they can also prove to be an impediment in domestic regulators trying to ensure accountability of algorithm and need to be revisited.

Norms

A. Use of nudges for privacy enhanced choices

Daniel Solove has argued that insights from cognitive science, particularly using the theory of nudge would be an acceptable compromise between the inefficacy of privacy self-management and the dangers of paternalism. [15] His rationale is that while nudges influence choice, they are not overly paternalistic in that they still give the individual the option of making choices contrary to those sought by the choice architecture. This is an important distinction and it demonstrates that ‘nudging’ is less coercive than how we generally understand paternalistic policies.

B. Use of defaults in privacy notices

One of the nudging techniques which makes a lot of sense in the context of the data protection policies is the use of defaults. It relies on the oft-mentioned

status quo bias.[16] A number of data collectors have maximum disclosure as their default settings and effort in understanding and changing these settings is rarely employed by users. A rule which mandates that data collectors set optimal defaults that ensure that the most sensitive information is subjected to least degree of disclosure unless otherwise chosen by the user, will ensure greater privacy protection. Ryan Calo and Dr. Victoria Groom explored an alternative to the traditional notice and consent regime at the Centre of Internet and Society, Stanford University.[17] They conducted a two-phase experimental study. In the first phase, a standard privacy notice was compared with a control condition and a simplified notice to see if improving the readability impacted the response of users. In the second phase, the notice was compared with five notices strategies, out of which four were intended to enhance privacy protective behavior and one was intended to lower it. Shara Monteleone and her team used a similar approach but with a much larger sample size.[18] One of the primary behavioral insights used was that when we do repetitive activities including accepting online terms and conditions or privacy notices, we tend to use our automatic or fast thinking instead to reflective or slow thinking.[19] Changing them requires leveraging the automatic behavior of the individuals.

C. Usability Approach

Alessandro Acquisti, Professor of Information Technology and Public Policy at the Heinz College, Carnegie Mellon University, has studied the application of methodologies from behavioral economics to investigate privacy decision-making. He highlights a variety of factors that distort decision-making such as - "inconsistent preferences and frames of judgment; opposing or contradictory needs (such as the need for publicity combined with the need for privacy); incomplete information about risks, consequences, or solutions inherent to provisioning (or protecting) personal information; bounded cognitive abilities that limit our ability to consider or reflect on the consequences of privacy-relevant actions; and various systematic (and therefore predictable) deviations from the abstractly rational decision process." Acquisti advocates a usability approach that entails designing the system in way that is most intuitive and easy for users to decide whether to provide the information, along with a soft paternalistic approach which seeks to aid the decision-making by providing other information such as how many people would have access to the information, if provided, and set defaults such that the information is not visible to others unless explicitly set by the user. The last two approaches are typically cited as examples of nudging approaches to privacy.

D. Algorithmic Transparency

With more and more decisionmaking becoming automated and determined by algorithms, a growing concern has been the impact on the ability of the those affected by such decisions to review and question them. Frank Pasquale has advocated greater algorithmic transparency of big data systems. Disclosing the logic and flow of data driven decision-making can help individuals examine the conclusions drawn about them and their accuracy and fairness.[20]

Architecture

A. Greater 'Human-Data Interaction'

The idea of Human-Data Interaction has arisen from 'the need, both ethical and practical, to engage users to a much greater degree with the collection, analysis, and trade of their personal data, in addition to providing them with an intuitive feedback mechanism.'[21] One of the earliest proponents of such an approach in response to big data were Omar Tene and Jules Polenetsky who argued that "individuals must be offered meaningful rights to access their data in a usable, machine-readable format." Technological projects such as the Berkman Centre of Internet and Society's Project VRM (Vendor Relationship Management) which addresses issues such as vendor lock-in and tries to provide better and more equitable manners of engaging with vendors are examples of this approach. Potential examples of such measures could be greater control to individuals who use IndiaStack for digital payments on their data and metadata.

B. Algorithmic Affirmative Action

Cynthia Dwork and Deirdre Mulligan have argued that "Exposing the datasets and algorithms of big data analysis to scrutiny—transparency solutions—may improve individual comprehension, but given the independent (sometimes intended) complexity of algorithms, it is unreasonable to expect transparency alone to root out bias." Anupam Chander proposes affirmative action as the remedy against algorithmic discrimination, as they replicate or amplify real-world biases through their statistical methodologies. This would involve both design choices which address specific discriminatory impacts of an algorithm, as well as focus on data which the algorithms use without needing transparency in the design of the algorithm, and the data used to train an algorithm are evaluated for being embedded with implicit and institutional biases.[22]

C. Technological Due Process

An argument that is rapidly gaining emergence is to approach big data regulation from the point of view of procedural fairness of big data's analytical processes with regards to how they use personal data (or metadata derived from or associated with personal data) in any adjudicative process, including those whereby big data is being used to determine attributes or categories of an individual. Kate Crawford has pointed out that predictive harms cannot be foreseen or prevented because it is hard to determine at which point in the technical process the privacy violation occurs, and whether/how it will be used by companies and agencies that gain access to the PPI (personally identifiable information). For this reason,

the best solution would be to regulate the process through which individuals are adjudicated based on said PPI. For data due process, the affected party must receive an opportunity to present an argument, evidence and corrections to prejudice. The elements from traditional due process that are most suitable to data due process would be: "an unbiased tribunal", "the right to know the evidence against one", "the making of a record" and "a statement of reasons". The greater seriousness of the deprivation to the individual, the more protections should be in place.

D. Audits

Along with technological due process, another measure to address procedural fairness could be a requirement to explain in detail reliance on an automated system's decision, including any computer-generated facts or legal findings. Scoring systems should be subject to licensing and audit requirements when they enter critical settings like employment, insurance and health care. Automated administrative systems often fail to retain any audit record of how they made the decisions at issue or upon what data the decision was based. She hereby calls for an audit trail because it provides reassurance and increases accuracy. Access to these audit trails would also allow individuals to raise specific questions and objections to how and when their data is being used in various processes. Furthermore, she maintains that agencies should be required to regularly test their system's software for bias and other errors.

Markets

A. Adoption of best practices to address privacy and security concerns

While limitations of existing legal frameworks to regulate big data have come to the fore, various actors have also pointed to mechanisms which are self-regulatory or co-regulatory as possible solutions. In the United States, the White House and FTC have both published papers that look at industry codes of conduct and multi-stakeholder processes. Self-regulations which are typically common guidelines that economic operators formulate themselves while co-regulations are usually examples of societal goals arrived at through community consultations which economic operators are entrusted to achieve.

There have been growing privacy concerns regarding companies collecting behavioral data and then sell the results to other companies after analysis. These companies also attracted the attention of the Federal Trade Commission in United States which issued orders to nine companies in the data broker industry, requiring them to provide information on their consumer data collection and usage practices.[23] These issues are not limited merely to data brokers but also companies engaged in analysis of behavioral data in other sectors. Amy Dean, a data warehouse specialist with Emory University in Atlanta has recommended that varied and disparate data sources can be weighted or scored in terms of data quality to factor into the analytics, and that data sources be available for reference so as to allow for tracing back of data elements to their source.[24] Other self regulation practices include Institutional Review Boards in specific sectors which could govern operators in their domains and ensure fair use of sensitive datasets

B. Better Design Principles in Data Collection

An enduring criticism of the existing notice and consent framework has been that long, verbose and unintelligible privacy notices are not efficient in informing individuals and helping them make rational choices. While this problem predates big data, it has only become more pronounced in recent times, given the ubiquity of data collection and implicit ways in which data is being collected and harvested. Further, constrained interfaces on mobile devices, wearables, and smart home devices connected in an Internet of Things amplify the usability issues of the privacy notices. Some of the issues with privacy notices include Notice complexity, lack of real choices, notices decoupled from the system collecting data etc. An industry standard for a design approach to privacy notices which includes looking at factors such as the timing of the notice, the channels used for communicating the notices, the modality (written, audio, machine readable, visual) of the notice and whether the notice only provides information or also include choices within its framework, would be of great help. Further, use of privacy by design principles can be done not just at the level of privacy notices but at each step of the information flow, and the architecture of the system can be geared towards more privacy enhanced choices.

C. Standards as a means to address data issues

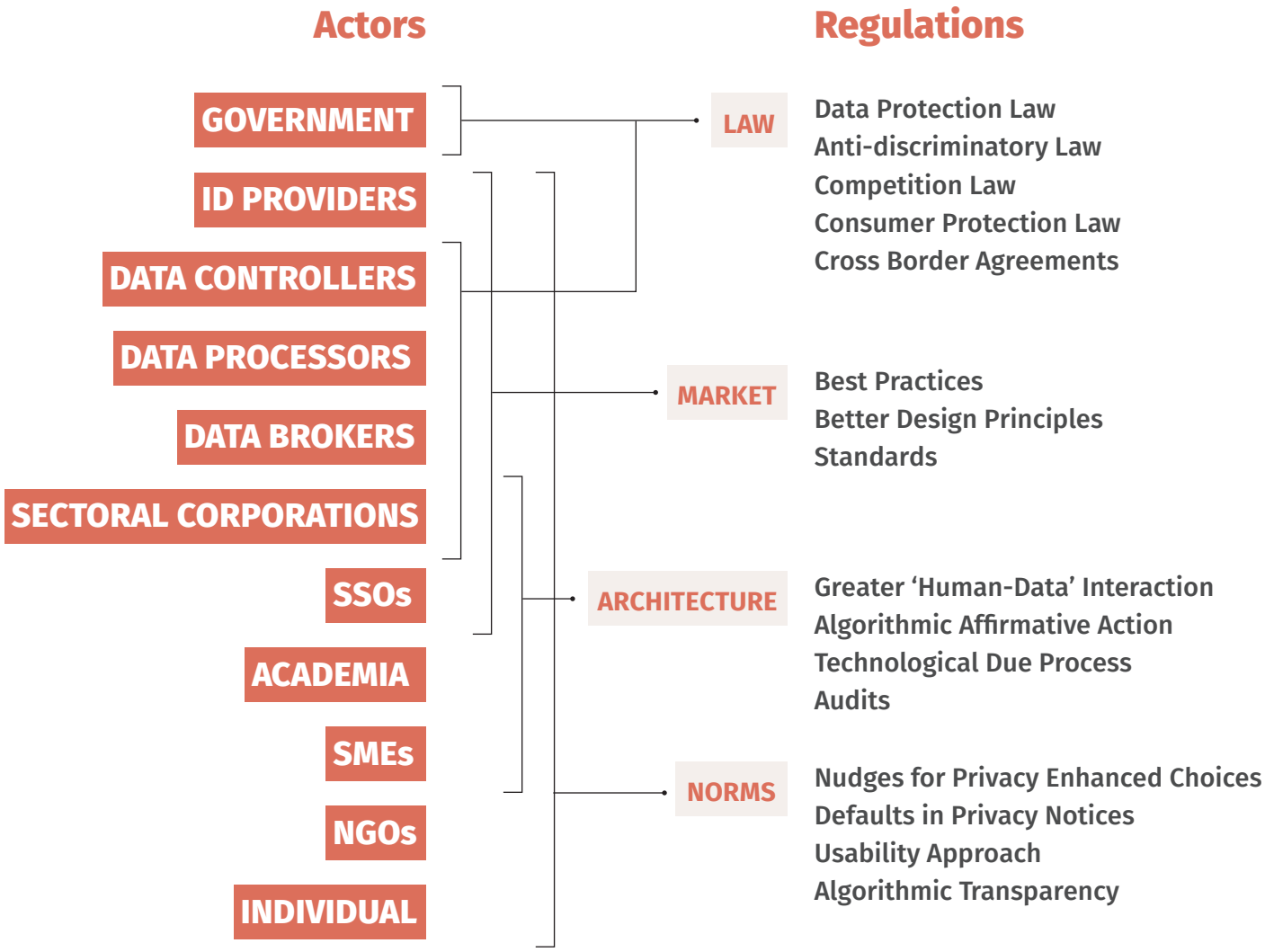
With digitisation of legacy records and the ability to capture more granular data digitally, one of the biggest challenges facing big data is a lack of standardised data and interoperability frameworks. This is particularly true in the healthcare and medicine sector where medical records do not follow a clear standard, which poses a challenge to their datafication and analysis. The presence of developed standards in data management and exchange, interoperable Distributed Application Platform and Services, Semantic related standards for markup, structure, query, semantics, Information access and exchange have been spoken of as essential to address the issues of lack of standards in big data.[25]

ENDNOTES

1. Maryanne Gaitho, How Applications of Big Data Drive Industries, available at https://www.simplilearn.com/big-data-applications-in-industries-article.
2. Quentin Baltus, Understanding big data key debates and controversies: a critical literature review, available at https://www.linkedin.com/pulse/understanding-big-data-key-debates-controversies-critical-baltus.

3. Fred Cate, Government Data Mining: The Need for a Legal Framework, available at http://www.law.harvard.edu/students/orgs/crcl/vol43_2/435-490_Cate.pdf.
4. J. Schradie, The Digital Production Gap: The Digital Divide and Web 2.0 Collide. Elsevier Poetics, 39 (1).
5. D Lazer, et al., The Parable of Google Flu: Traps in Big Data Analysis. Science. 343 (1).
6. Danah Boyd and Kate Crawford, Critical Questions for Big Data. Information, Communication & Society. 15 (5).
7. John Podesta, (2014) Big Data: Seizing Opportunities, Preserving Values, available at http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
8. E. Ramirez, (2014) FTC to Examine Effects of Big Data on Low Income and Underserved Consumers at September Workshop, available at http://www.ftc.gov/news-events/press-releases/2014/04/ftc-examine-effects-big-data-lowincome-underserved-consumers
9. M. Schrage, Big Data's Dangerous New Era of Discrimination, available at http://blogs.hbr.org/2014/01/bigdatas-dangerous-new-era-of-discrimination/.
10. Tara Thigarajan, India's challenge of 'dark data', Livemint, available at http://www.livemint.com/Opinion/JZ95vbh4jS5KI1ll4kjcIO/Indias-challenge-of-dark-data.html.
11. Charly Gordon, big data exclusions and disparate impact: investigating the exclusionary dynamics of the big data phenomenon, available at http://www.lse.ac.uk/media@lse/research/mediaWorkingPapers/MScDissertationSeries/2014/Charly-Gordon-MSc-Dissertation-Series-AF.pdf.
12. Case No COMP/M.4731 – Google/ DoubleClick available at http://ec.europa.eu/competition/mergers/cases/decisions/m4731_20080311_20682_en.pdf
13. French Competition Authority, Opinion n°10-A-13 of 1406.2010, http://www.autoritedelaconcurrence.fr/pdf/avis/10a13.pdf. That opinion of the Authority aimed at giving general guidance on that subject. It did not focus on any particular market or industry although it described a possible application of its analysis to the telecom industry.
14. http://www.analysisgroup.com/is-big-data-a-true-source-of-market-power/#sthash.5ZHmrD1m.dpuf
15. Daniel Solove, Privacy self-management and consent dilemma, 2013 available at http://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2093&context=faculty_publications.
16. Frederik Borgesius, Behavioral sciences and the regulation of privacy on the Internet, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2513771.
17. Ryan Calo and Dr. Victoria Groom, Reversing the Privacy Paradox: An experimental study, available at http://ssrn.com/abstract=1993125.
18. Shara Monteleon et al, Nudges to Privacy Behavior: Exploring an alternative approahc to privacy notices, available at http://publications.jrc.ec.europa.eu/repository/bitstream/JRC96695/jrc96695.pdf
19. Daniel Kahneman, Thinking, fast and slow, Farrar, Straus and Giroux, 2011.
20. Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, 2015.
21. Hamed Haddadi, Richard Mortier, Derek McAuley, Jon Crowcroft, Human-data interaction, available at https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-837.pdf
22. Michael Feldman et al., Certifying and Removing Disparate Impact, http://arxiv.org/abs/1412.3756 (July 16, 2015). Michael Feldman,. Certifying and Removing Disparate Impact. 2015.
23. http://searchsecurity.techtarget.com/feature/Managing-big-data-privacy-concerns-Tactics-for-proactive-enterprises.
24. http://searchsecurity.techtarget.com/tip/Data-sanitization-policy-How-to-ensure-thorough-data-scrubbing
25. http://www.iso.org/iso/big_data_report-jtc1.pdf

Regulation of Big Data



This document uses the following icons from **The Noun Project**

Fingerprint created by Daouna Jeong
Propane Tank created by Ambar Bhusari
Mobile Call created by Thays Malcher
Web created by Aneeque Ahmed
Cursor created by Arthur Shlain
Psychometric Analysis created by Chameleon Design
Factory created by Franc
Government created by Rflor
Bank created by Arthur Shlain
Electricity Meter created by Olivier Guin
Sleeping Bag created by Ed Harrison