Data Science for Bridging the Digital Divide and Beyond

Gareth Tyson, Ignacio Castro, Jon Crowcroft, Mahesh Marina & Mark Graham.

This report covers the "*Data science for bridging the digital divide and beyond*" Workshop, arranged by the Alan Turing Institute on July 16th & 17th July, 2019. The "Digital Divide" refers to the gap between those who have access to ICT, and those who do not. The workshop aimed to bring together expertise across a number of disciplines, in an attempt to better understand the role that Data Science could play in bridging this digital divide (both domestically and internationally). As part of this, the workshop hosted a number of interactive sessions to expose participants to a number of differing views and ideas. The workshop was aimed to help bootstrap a multidisciplinary community, within the context of the Alan Turing Institute, which will tackle some of these issues. A further key outcome of the workshop has been a series of research ideas that the participants felt worthwhile exploring further. We created a Twitter hashtag, which people can search for to see some of the observations being raised: `#DigitalDivideTuring`.[1]

1. Overview of Sessions

On the first day we had three presentation sessions, covering complementary aspects of bridging the Digital Divide: (1) regulation and economics of infrastructure; (2) inequality, poverty and social factors; (3) the role of technological innovation. In this section, we offer a brief summary of each presentation. We had approximately 35 participants, led by several prominent academics and practitioners.

Figure 1: *Steve Song, Ignacio Castro and Andrew Button discussing data regulation.*

[1] https://twitter.com/hashtag/digitaldivideturing

Session 1: The politics of cables: regulation & economics of infrastructure

[Chair: Ignacio Castro]

This session focused on regulatory and policy issues within bridging the digital divide. It explored what challenges exist when trying to formulate good regulation, and how they could be overcome in the future.

Steve Song: Open Telecom Data: Why greater transparency in the telecom sector is essential to the healthy evolution of the internet. Steve argued that as access to digital communication becomes more strategically valuable, there is a corresponding need for more public data regarding the ownership, location and terms of use of telecommunications infrastructure. We discussed about how transparency is essential to understanding how to deliver affordable access into areas not deemed economically viable by incumbent operators. Steve further proposed that transparency is key to understanding how investment is reshaping global telecommunications infrastructure and by extension the internet traffic that relies on it.

Andrew Button: Addressing the problem: using geospatial science to unlock infrastructure intelligence. Andrew presented some of the ongoing work at Ofcom, the UK's regulatory and competition authority for the broadcasting, telecommunications and postal industries. He explained how Ofcom endeavours to achieve a street level view of digital services. He presented their work in balancing transparency with privacy, commercial with public interests, and showed how they are developing services to deliver fairness for all. We then moved on to discuss how data science and technology is enabling Ofcom to unlock information at a granular geographic level. Finally, he highlighted that there is still more to achieve, and lessons to learn, from pan-Government, academic and industry engagement. The discussion seemed to show that Ofcom's work in this area is a good exemplar for telecoms regulators more broadly.

Session 2: The human side: inequality, poverty and social factors

[Chair: Jon Crowcroft]

This session focussed on the human-side of bridging the digital divide, with a focus on inequality. It explored some of the ways we can integrate communities into the collection and use of data, as well as some of the key ethical questions that must be asked when starting such initiatives.

Reem Talhouk: Data Driven Technologies as Experienced by Syrian Refugee Communities. The session was opened by discussing some of the ethnography work carried out by Reem in Syrian Refugee communities. She described how emerging technological trends, such as distributed ledgers, are being integrated in to the humanitarian aid system. She explained that, as digital technologies start to play a bigger role in the way aid is distributed as well as generating data to inform humanitarian actors, the community needs to also consider how these technologies are experienced in the day to day lives of refugees. Reem presented some of her on-the-ground findings from co-designing research with Syrian refugees in Lebanon

and discussed how humanitarian technologies are experienced by refugee communities as well as the interplay between these technologies and refugee values and practices.

Enrico Calandro: After Access: Measuring digital inequalities in developing countries.
Enrico described much of the ongoing work performed at Research ICT Africa, with a particular focus on the large-scale pan-African surveys they have been carrying out. Enrico showed that, despite many of the UN's Sustainable Development Goals (e.g. gender equality, good health, quality education), we do not have accurate and good quality data to assess progress towards overcoming the 'digital divide'. For example, data collected nationally by regulatory agencies and internationally by the International Telecommunications Union (ITU) does not allow us to measure several basic socio-demographic indicators. Enrico explained how an effective way to measure progress towards achieving more equitable digital societies and economies, is to collect data on Internet usage through nationally representative demand-side surveys, and to combine them with financial performance indicators from operators and internet measurements, where available. He showed how RIA had been using these indicators to provide evidence for good policy and regulatory practices in the telecoms sector, as well as methods to assess their success.

Richard Heeks: Data Justice and the Next One Billion. Richard's presentation focussed on the growing datafication of international development. This included the use of new data sets and analytical techniques relating to the spread of mobile and internet connectivity. He explained the value of assessing this datafication from the emerging perspective of "data justice"; defined as the specification and pursuit of ethical standards for data-related resources, processes and structures. He explained how this enables us to evaluate next-billion data science along five dimensions of justice: procedural, rights-based, instrumental, structural and distributive. He presented the concept in terms of an assessment toolkit or, more loosely, as a set of ethical principles for next-billion initiatives.

Session 3: The role of innovation: technology for bridging the divide

[Chair: Nishanth Sastry]

This session explored some of the recent technological innovations that are being used in Bridging the Digital Divide. We started by discussing how data can be used to better support aid spending, before talking about the role data can play in improving Internet connectivity for underserved communities.

Tom Wilkinson: AIID - Establishing a common data and intelligence platform for the delivery of International Aid.
Tom discussed some of the work that has been taking place in the UK's Department for International Development (DFID), alongside the UK's Office for National Statistics. They have been striving to put DFID at the forefront of AI in International Development, leveraging the UK's advantages of world leading expertise in both Data Science and International Development. Tom presented some of the concepts and tooling they have been building to exploit data for better management and allocation of aid budgets. The initiative Tom described has been building on common data standards and open source tools for sharing data.

Josiah Chavula: Internet in Africa: from macroscopic to microscopic view of performance.
Josiah presented work looking at the performance of Internet infrastructure in Africa.
He quantified the performance issues with the Internet across various African countries, and explained some of the challenges faced (e.g. slow download speeds and high latency). He motivated the need to gather data on Internet performance, as well as the need to share such data to support operators in improving their infrastructures. The presentation also highlighted his efforts to characterise Internet performance from users' perspective, particularly in underserved communities, with a case-study of a township in South Africa.

Arjuna Sathiaseelan: GAIUS: Enabling a hyper local content ecosystem for emerging markets. The last presentation of the day described some of the problems with delivering Internet content to underserved communities, and the subsequent problem in supporting them with data-driven initiatives. He described (and demoed) GAIUS - a mobile app which allows users to create and share web content, without the need for full Internet connectivity. He also described some of the business models that can be built up in a bottom-up manner. He showed how GAIUS can function as an “Internet in a box” entirely independent from the rest of the Internet. He then discussed some of the forms of data that this might generate, as well as how it can be used, e.g. to optimise the delivery of content to people in underserved communities.

2. Overview of Breakouts and Mini Projects.

On the afternoon of the first day, we hosted three break-out discussions. The breakouts were themed around the broad topics explored during the expert presentations: (1) regulation and economics of infrastructure; (2) inequality, poverty and social factors; (3) the role of technological innovation. In this section, we offer brief summaries of each presentation. Within these break-outs, we also asked teams to think of ‘mini-projects’, highlighting critical topics worth further investigation.

Breakout 1: Evidence-based regulation - where is the shortfall?

This breakout, chaired by Ignacio Castro, discussed problems with data regulation and how these deficiencies might stunt expansion of low-cost telecommunications in developing regions. The breakout highlighted issues related to spectrum regulation for rural communities, and how regulators are not reactive even when facing evidence that counters policies. Despite this, the team agreed that increasing regulation transparency would be a good thing, particularly if it can be used to strengthen evidence-informed policy. As such, they brainstormed two key mini-projects. First, they felt it worthwhile performing more research into how they could better connect the logical view of the Internet (e.g. IP routing) with the underlying physical infrastructure. There are few available techniques to enable this, yet it is critical for understanding who regulators should contact to discuss problems. Second, and on a similar line, they proposed further research into developing methods to detect who owns infrastructure, and how this might impact the ability to build better global regulation of the telecoms sector.

Breakout 2: Is AI a risk to global equality?

This breakout, chaired by Jon Crowcroft and Richard Heeks, discussed equality issues with gathering and using data for bridging the Digital Divide. They discussed a number of issues with (unwisely) applying algorithms/datasets from the West in a developing world context. They brainstormed four potential mini-projects. First, they proposed a study into exploring algorithmic vulnerability in the Global South. This pertained the implications of using artificial intelligence in developing regions, including the reliance of algorithms trained using datasets that may not reflect the interests of people living in such regions. Second, they proposed to study cultural differences across data for development projects, and how cultural sensitivities could be better respected. This linked into earlier discussions about assumptions that people make when deploying technologies in developing regions (e.g. assuming that a country might already have postcodes that could be used within datasets). Third, they discussed the value of exploring the grey/black data economy, and how pushing technologies into new regions may impact the wider economy, e.g. exposing people to online scams. Fourth, they proposed further research into creating support for minority languages. They observed that people who are not fluent in English may struggle to fully engage with online services and data. Hence, they proposed using crowd-sourcing to create a semantic corpus for NLP in minority languages.

Breakout 3: Technology for good?

This break-out, chaired by Nishanth Sastry, explored some of the technical solutions that could be used to assist with exploiting data for bridging the digital divide. They highlighted the high barrier to entry for many people, particularly those not trained in Computer Science. They therefore proposed a mini-project, whereby a Data Science Platform could be built, and made available to the community. This would focus on easy-to-use services, such as the ability to upload data which can then be automatically visualised. They talked about running a data sharing and management services, which would make it easier for organisations to make data publicly available. They also touched upon the problem of powering Machine Learning, via manual annotations, i.e. Data Labour. A particular example was the need for people, often in developing countries, to work as content moderators (e.g. filtering offensive content for Facebook). The major social problems that this poses were discussed. Hence, a second mini-project idea was to research into the implications of this approach to moderation, and develop better technologies to replace manual work.

3. Future Directions

The second day focussed on future directions, striving to bring together some of the ideas discussed within the break-outs and presentations. Thus, we brainstormed several themes. Each theme consisted of a mix of three concepts: (a) Datasets that we think are vital to better understand and bridge the digital divide; (b) Research questions and lines of analysis that can be underpinned with such data; (c) Meta issues that must be considered while pursuing such questions, e.g. accuracy of data or the need for privacy.

The above was coordinated through a series of post-it note exercise, whereby participants were encouraged to write their ideas on coloured post-it notes (one colour for each of the three concepts above). We then rearranged the post-it notes into a set of disciplinary areas: (1) Social aspects of the Digital Divide; (2) Internet Access; (3) Infrastructure and regulation; (4) Green considerations and sustainability; (5) Human-Computer Interaction (HCI); (6) Economic aspects; and (7) Meta-issues. We present the post-it notes in Figure 2. Following on from the discussion, we transcribed the main points from the post-it notes and captured them below.

Figure 2: *Brainstorming Concepts, which includes (1) Datasets that we think are vital for better understand and bridging the digital divide; (2) Research questions and lines of analysis that can be underpinned with such data; (3) Meta issues that must be considered while pursuing such questions, e.g. accuracy of data or the need for privacy.*

Social aspects of the Digital Divide: Many of the participants at the workshop had ideas about exploring social aspects of the digital divide. These were mostly divided between ideas for new datasets (blue) and ideas of research questions (purple). The datasets focus on survey data that captures why people do not have access to the Internet, and how such data could be linked to Internet datasets (e.g. social media data) to understand social dynamics. Many of the meta issues (orange) focus on the ethics of such data collection. A full list of (informal) thoughts are:

- Research ICT Africa (RIA) survey data, which includes pricing information available: https://researchictafrica.net/ramp_indices_portal/
- RIA's After Access dataset, covering household and individual interviews on Internet usage:

https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/765/related_materials
- RIA's first level analysis of mobile internet access and use across 10 African countries: https://researchictafrica.net/wp/wp-content/uploads/2019/05/2019_After-Access_Africa-Comparative-report.pdf
- Further rich qualitative interview data on what means to individuals to be "connected" or "unconnected".
- Geospatial and social deprivation indexes.
- General information on the RIA survey methodology (aims, partners involved, data visualisation): https://afteraccess.net/
- Can we map and link social network data to measure social capital (as network transitivity)?
- Can we perform social science data analysis, that uses social network data from developing regions to observe evolving social patterns?
- Does infrastructure growth leads to more users and financial gain for the operators?
- Can we use maps of social connections (e.g. from Facebook) to model disease spread?
- Can we explore the role of Peer to Peer institutions (e.g. peer to peer lending, AirBnB, Deliveroo)?
- Can we explore social capital and network effects in small communities in developing regions?
- All analysis is also required to protect from the linking of personal information or personal identifiable information/privacy
- Data labour involved in the process needs to be straightforward, and we should explore how to make labelling interfaces more expressive. Data labour must also be ethical.
- Analysis should follow community-driven research data policies, e.g. how should data be collected?
- Analysis should include participation from the underrepresented in defining what are the data questions

Internet Access: The majority of post-it notes in this topic related to datasets. Many of the Internet measurement specialists were discussing various types of actively collected data (e.g. traceroutes), but there were a number of requests for traffic data provided from telecoms operators. A full list of (informal) thoughts are:

- DNS manipulation data (e.g. from Princeton).
- Traceroute data to identify number of AS hops within developing regions.
- Operational network data, e.g. traffic traces.
- Dataset containing web complaints related to content accessed in developing regions (e.g. related to illegal content).
- Cell data (e.g. device mobility).
- Call records.

    - African topology data and performance measurements, e.g. what is the page load time of websites?
    - What are the average end-to-end network delay in small geographical areas?

Infrastructure & Regulation:
There was a strong interest in better understanding the availability and presence of physical telecoms infrastructure. A number of datasets were listed:

- Open map of fibre optic infrastructure in Africa: https://afterfibre.nsrc.org/
- Examples of Good Practice in Open Telecom Data: https://wiki.opentelecomdata.org/good-practice/transparency
- GSMA Mobile Coverage Maps for Africa: https://www.mobilecoveragemaps.com/
- Open Spectrum Assignments for African Countries: https://opentelecomdata.org/spectrum-chart/
- Steve Song also maintains a regularly updated list of connectivity statistics, maps, and reports: https://github.com/stevesong/awesome-connectivity-info/blob/master/README.md

Green and Sustainability: A set of participants were highlighting sustainability problems in the developing world, and were suggesting that social surveys should also cover environmental issues. They were also keen to explore climate change data, and to look at how these could be used to improve resilience to things like flooding. A full list of (informal) thoughts are:

    - RIA Survey data covering energy consumption. This should be nationally representative (ZA, Nigeria, Kenya, Mozambique, Rwanda, hana, Uganda, Tanzania, Seneal, Lesoto).
    - Datasets covering energy consumption of networking hardware.
    - Data and climate change: how new datasets and new analytics can build resilience to flooding, storms, sanitisation, etc?

Human Computer Interaction: A particularly novel idea here was developing programming languages that can assist with data science, particularly targeted at people in developing regions. Another focus was on how people should be integrated in the Data Science process in a bottom-up fashion, i.e. the people using the technology should also be co-designing it. A full list of (informal) thoughts are:

    - Can we build programming languages for data science, targeted as use in developing regions?
    - Datasets should be compiled via co-creation and co-ownership of datasets
    - Can we use things like story-telling and narrative to put people in the center of answering these questions. For example, can we create “data stories” - narratives of how data is actually used, e.g. eVouchers.

Economic aspects: Naturally, economics played a prominent role in the discussions. People

were requesting datasets that explain Internet pricing patterns across developing regions. Another request was for wholesale Internet pricing (between networks), which is quite difficult to obtain. An interesting observation was also the commercial sensitivity of some of the datasets discussed; hence it was proposed that there is a need to offer better economic incentives for companies to share datasets. A full list of (informal) thoughts are:

- Internet pricing data across Africa, covering consumer prices across providers and packages.
- International wholesale Internet prices dataset.
- Mobile money transaction data, explaining how and when people spend.
- Can we use computational sources to predict/estimate the distribution of wealth measured more quickly and cheaply?
- Can we devise non-GDP measures of wealth?
- How many IoT operators are using cellular services to connect over the Internet? Can we propose Value Added Services for IoT operators?
- As the datasets are often commercial sensitivity of data, how can companies be incentivised to share?
- How can all the questions be linked into valuable business problems?

Meta-issues: Finally, a number of people put 'meta-issues' up that applied across all of the disciplinary domains discussed above. These touched up ideas such as the difficulty of compiling longitudinal data, and the challenges of collecting such data in a privacy-preserving fashion. A full list of (informal) thoughts are:

- We need to explore the capture of longitudinal data, and how we can ensure that comparisons are consistent and robust.
- We need to explore and understand the trade-off between privacy vs. data disclosure.
- We need data aggregation, and this might also link into anonymisation requirements.
- We should start to consider the costs of releasing datasets (e.g. time to prepare them), and thinking of ways to streamline this process.
- Before collecting or releasing datasets, we should think about how the data might be used and whether it could have a negative impact?
- We need to think about how we can measure impact, both negative and positive, whenever trying new initiatives.

4. Conclusions

This report has described some of the key discussion points in the "*Data science for bridging the digital divide and beyond*" workshop. Much of the workshop was dedicated to brainstorming future research directions and potential 'mini-projects' which could bootstrap new collaborations. It is hoped that this can offer a foundation to build further work at the Alan Turing Institute on this topic.