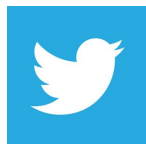# Data Science for Bridging the Digital Divide Workshop
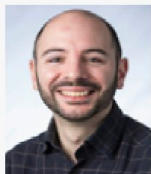
The
Alan Turing
Institute

#DigitalDivideTuring

Professor Jon Crowcroft

Ignacio Castro

Dr Mahesh Marina

Professor Mark Graham

Dr Gareth Tyson

Today's agenda

Day 2 (Wednesday 17th July)

10:00	Organisers	Summary from yesterday
10:15	Organisers	Data Gathering and Analysis for Bridging the Digital Divide
11:00	Coffee Break	
11:30	Organisers	Formulating Projects and a Research Roadmap of Bridging the Digital Divide
12:15	Organisers	Future Directions
12:30	Lunch	
13:00	Lunch	

Let's recap

The politics of cables: regulation, economics and infrastructure

- Steve: Open Telecom Data: Why greater transparency in the telecom sector is essential to the healthy evolution of the internet

- Andrew: Addressing the problem: using geospatial science to unlock infrastructure intelligence

The human side: inequality, poverty and social factors

- Reem: Data Driven Technologies as Experienced by Syrian Refugee Communities

- Enrico: After Access: Measuring digital inequalities in developing countries

- Richard: Data Justice and the Next One Billion

The role of innovation: technology for bridging the divide

- Tom: AIID - Establishing a common data and intelligence platform for the delivery of International Aid

- Josiah: Internet in Africa: from a macroscopic to microscopic view of performance

- Arjuna: GAIUS: Enabling a hyper local content ecosystem for emerging markets

Some common themes

- Data transparency vs. data usage
  - Some wanted data for transparency purposes, some required it for specific purpose
  - If it's just transparency - who is it for? Who benefits? How is it audited?
- Data availability
  - Who has data? Do they share it? How should it be shared? Can we force entities to share? Who are the winners and losers?
- Data accuracy
  - How do we know the data is correct? How do we incentivise it to be correct? What happens if it is not correct?
- Data abuse
  - What are unjust uses of data? Can it be used against citizens and are they aware of the implications? What are the privacy implications of data collection?

Some uncommon themes

- Diverse datasets been discussed
  - Coverage data, AID spending and delivery, economic information, Internet performance, usage information
- Diverse collection methodologies
  - Steve's reverse engineering, Enrico's surveys, Josiah's measurements, Tom's official collections, Arjuna's in-app etc.
- Diverse use cases & applications
  - Ranging from managing spectrum allocation to understanding AID budgets
- Role of the regulator
  - Pushing data transparency vs. pushing data protection

(Incorrect) assumptions?

- People want to share data, e.g. installing applications on their phone

- People won't game the system, e.g. exchanging eVouchers

- People won't always be truthful, e.g. lying on surveys

- We understand cultural factors, e.g. everybody has a postcode

- There is consistency across regions covered - little comparative data, e.g. using GDP per capita makes sense

- There are the underpinnings for this to be built on, e.g. availability of local material in language

- Data doesn't have a cost, e.g. manual annotations

Breakouts

Evidence-based regulation - where is the shortfall?

- Connecting the logical view of the internet with the physical infrastructure

- Physical infrastructure ownership study

Is AI a risk to global equality?

- Algorithmic vulnerability in global south
- Cultural difference for inclusion in data for development projects.
- Understanding the grey/black data economy, e.g. sharing pirate content, undertaking scams, capture workshops
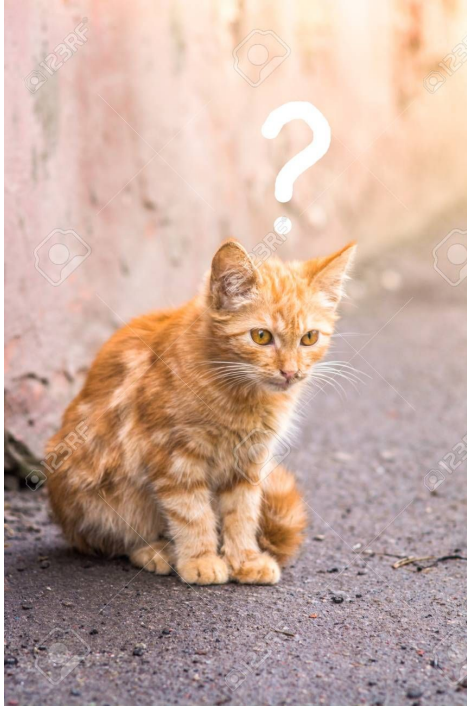- Crowd-sourcing to create semantic corpus for NLP in minority languages

Technology for good?

- Platform and model to help people share data, e.g. easy interface, provide visualisations etc.
- Understanding data labour, e.g. people moderating content

What datasets do we have?

- Telecoms: Fiber, tower, spectrum [Talk to Steve & Andrew]

- Economic: Pricing data [Talk to Enrico & Steve]

- Development data [Talk to Reem & Tom]

- Internet: traceroutes, latency measurements, web performance measurements [Talk to Josiah & Arjuna]

What datasets would you like?

What datasets should/could we collect?