Big data, development and governance

LIRNE*asia*

http://lirneasia.net/projects/bd4d/

Carnegie India, July 19, 2017

This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK . Also supported by New Venture Fund.

If “development” includes . . .

- Human settlements and infrastructure services that serve people better than they do now
- More efficient delivery of government services

If ideal governance* is about . . .

- Public discourse about public goods and about how positive/negative externalities are dealt with
  - Based on evidence
  - Using modalities that do not unduly privilege/disadvantage groups

* In reality, the ideal could be, to greater or lesser extent, a rhetorical cover for power-laden actions

- In which case, understanding the ideal and the ways in which actual practice deviates from it would be useful

Big data analytics is . . .

- Definitely needed

- Low-impact data collection
- More accurate than present methods, when correctly used
- Cheaper than present methods
- Can support policy experimentation

But . . .

- Data-information-knowledge are integral to control
- More the state knows, more it can control, for good or ill

- To solve governance problems using big data, we have to address problems of governance of big data

Some development problems that may be addressed using big data . .

- Worldwide, more people live in cities than in rural areas since 2008
  - How can we make cities more livable?
  - Is there a role of ICTs, not just more roads, transit, etc.?
- Infectious diseases are posing threats
  - Can we make better decisions re allocating scarce resources?
- Governments are flying blind, with ineffective National Statistical Organizations unable to give timely data needed to better target expenditures, assess programs or achieve development goals such as SDGs
  - Are there ways to remedy this?

Comprehensive coverage of population needed for most public-policy problems. Sources of data?

- Administrative data
  - E.g., digitized medical records, insurance records, tax records
- Commercial transactions (transaction-generated data)
  - E.g., Stock exchange data, bank transactions, credit card records, supermarket transactions connected by loyalty card number
- Online activities/ social media
  - E.g., online search activity, online page views, blogs/ FB/ twitter posts
- Sensors and tracking devices
  - E.g., road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones communicating with base stations, satellite/ GPS devices

Mobile Network Big Data is only option for some problems at this time

Country	Mobile Subscriptions/100	Internet Users/100	Facebook Users/100
	2016	**2016**	**2017**
Pakistan	71.4	15.5	15.8
Bangladesh	77.9	18.3	15.8
India	87.0	29.6	15.9
Myanmar	89.3	25.1	29.2
Philippines	109.2	55.5	59.7
Sri Lanka	118.5	32.1	25.0
Indonesia	149.1	25.4	44.4
Thailand	172.7	47.5	70.3

LIRNEasia
Pro-poor. Pro-market

Sources: http://www.itu.int/net4/itu-d/icteye/AdvancedDataSearch.aspx; http://datatopics.worldbank.org/hnp/popestimates; facebook advertising portal;

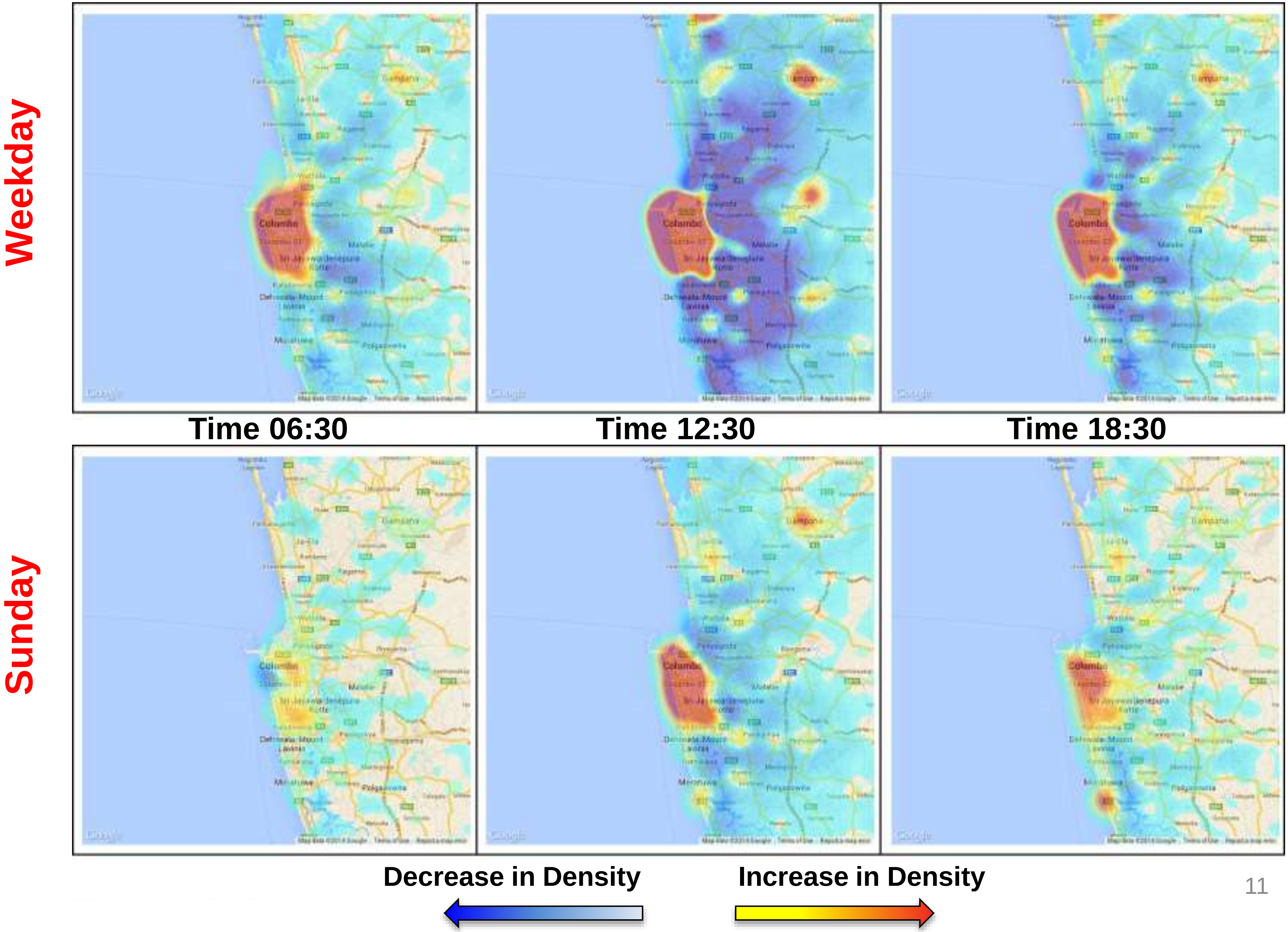
Data used in the research

- Multiple mobile operators in Sri Lanka provided four different types of meta-data
  - Call Detail Records (CDRs)
    - Records of calls
    - SMS
    - Internet access
  - Airtime recharge records
  - No Visitor Location Registry (VLR) data, because they are written over & not stored
- Data sets do not include any Personally Identifiable Information
  - All phone numbers are pseudonymized
  - LIRNE*asia* does not maintain any mappings of identifiers to original phone numbers
- Historical, not real time; therefore analyzed in batch mode in a hardware stack costing < USD 30k
- Cover 50-60% of users; very high coverage in Western (where Colombo the capital city in located) & Northern (most affected by civil conflict) Provinces, based on correlation with census data

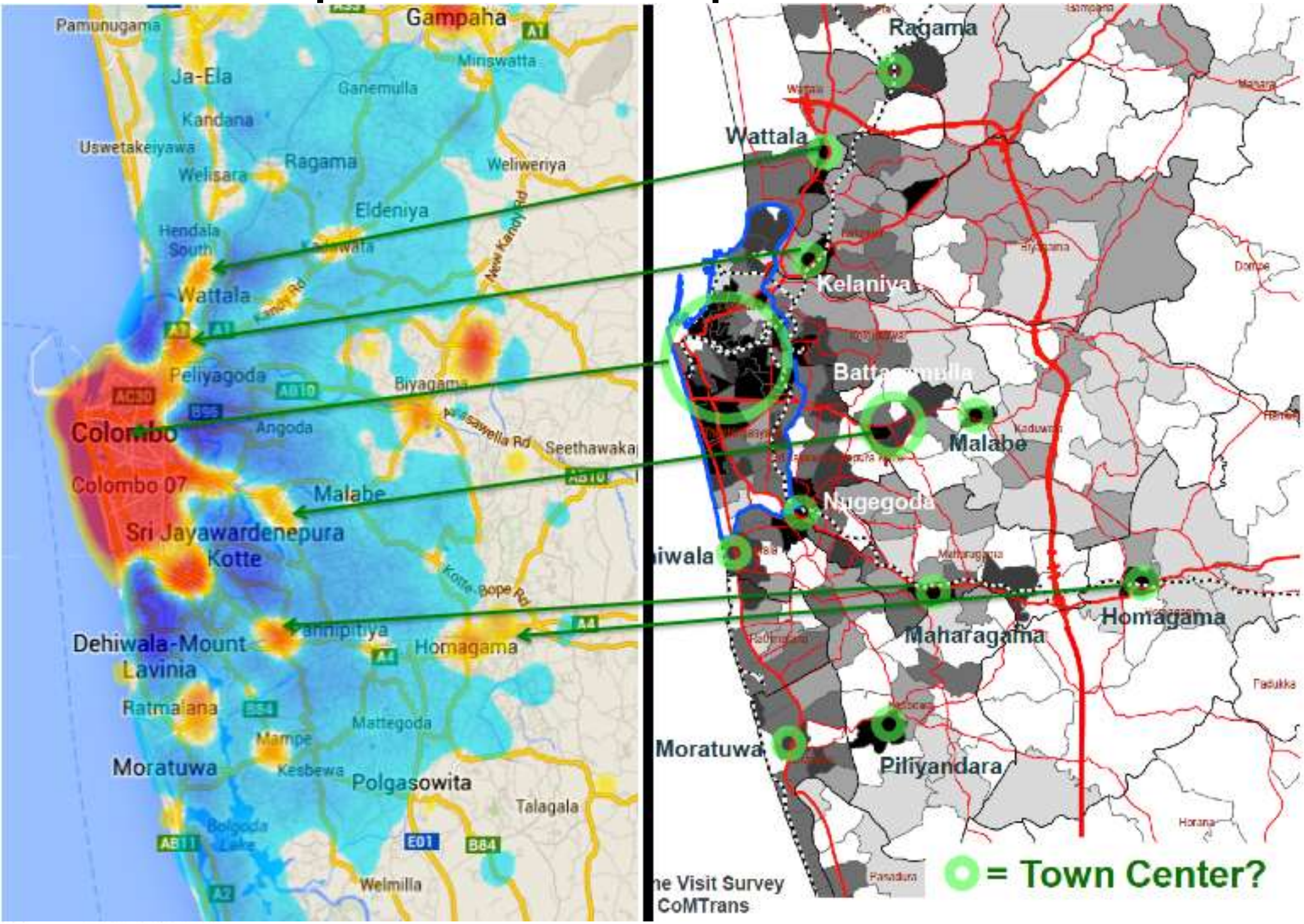
EXAMPLE: POPULATION DENSITY & MOVEMENT

Population density changes in Colombo region: weekday/ weekend

Pictures depict the change in population density at a particular time relative to midnight



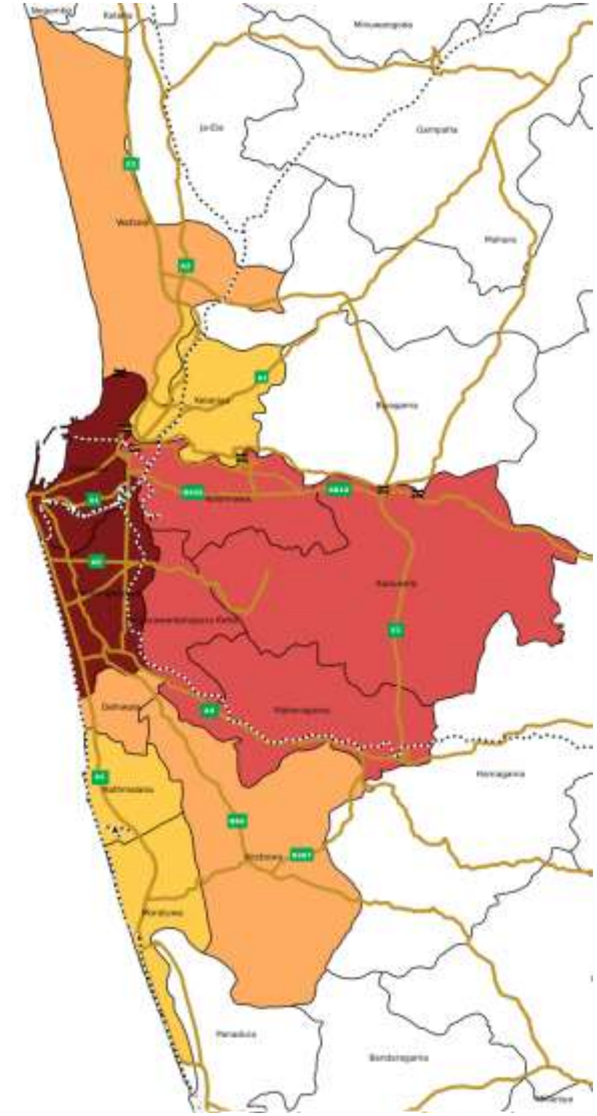
Our findings closely match results from expensive & infrequent transportation surveys; are cheaper & can be produced as needed



○ = Town Center?

...e Visit Survey
CoMTrans

46.9% of city's daytime population comes from outside. Potential configurations of a Metropolitan Corporation



Home DSD	Population	Percentage contribution to Colombo's daytime population
Colombo City (2 DSDs)	555,031	53.1
Maharagama	195,855	3.7
Total	555,031	53.1
Kolonnawa	190,817	3.5
Kaduwela	252,057	3.3
Sri J'pura Kotte	107,508	2.9
Dehiwala	87,834	2.6
Total	1,380,768	66.5
Kesbewa	244,062	2.5
Wattala	174,336	2.5
Kelaniya	[illegible]	2.1
Total	1,807,000	74.1
Ratmalana	95,162	2.0
Moratuwa	167,160	1.8
Total	2,204,015	79.9

GOVERNANCE OF DATA ANALYTICS

Marginalization

- City of Boston has an app called Street Bump for smartphones. Any citizen can activate the app at the beginning of a journey. The accelerometer of the smartphone collects data proven effective in identifying pot holes and speed bumps. At the end, the collected data including the GPS coordinates of starting and ending points are sent to City Hall. Algorithms differentiate between bumps that should be there and those that should not be. Roads with an excess of the latter get routed into the work order system for repairs.

- Should this be used at this time in
  - Patna?
  - New Delhi?

Privacy-associated harms

- As commonly understood, “is a sweeping concept, encompassing (among other things) freedom of thought, control over one’s body, solitude in one’s home, control over personal information, freedom from surveillance, protection of one’s reputation, and protection from searches and interrogations” (Solove, 2008, p. 1)
- Attempts to define privacy in terms of boundary control by individuals (e.g., Samarajiva, 1994: 90) are difficult to translate into practical policy

Our approach

- Focus on harms as identified by Solove's research that fall into four general types
  - Information collection;
  - Information processing;
  - Dissemination of information; and
  - Invasion
- Develop remedies to prevent harms
- Work with all stakeholders to operationalize safeguards

16 harms within umbrella meaning ➔ 9 of relevance to MNBD

1. Surveillance, interrogation (1/2)
2. Aggregation, identification, insecurity, secondary use, exclusion (5/5)
3. Breach of confidentiality, disclosure, exposure, increased accessibility, blackmail, appropriation, distortion (3/7)
4. Intrusion, decisional interference (0/2)

Most from information processing cluster; next from dissemination

LIRNEasia
Pro-poor. Pro-market

Examples of reasons for inclusion/exclusion

- Surveillance v ~~interrogation~~
  - Surveillance is obviously relevant
  - Interrogation is the pressuring of individuals to divulge information (physical coercion, not about information)
- Disclosure v ~~exposure~~
  - Both involve dissemination of true information, but exposure is limited to information about body and health
- ~~Decisional interference~~
  - "Right to privacy" is some countries encompasses a woman's decision whether or not to terminate pregnancy

Remedy	**Identified potential harm**	**Include in agreements transferring identifiable MNBD**	**Include in agreements transferring anonymized MNBD**
Mobile Network Operators (MNOs) will not engage in active surveillance of their customers, except as required by applicable law. MNOs will desist from collecting more data than are needed for the efficient operation of the networks and the supply of good service to customers. To the extent feasible, data collection practices will be transparent.	***Active surveillance***	No. Applying only to MNOs, this need not be included in agreements. However, active surveillance is a root cause of problems that could be manifested in other forms at the subsequent information processing and dissemination phases.	No. Applying only to MNOs, this need not be included in agreements. However, active surveillance is a root cause of problems that could be manifested in other forms at the subsequent information processing and dissemination phases.

Remedy	**Identified potential harm**	**Include in agreements transferring identifiable MNBD**	**Include in agreements transferring anonymized MNBD**
Any agreement transferring identifiable data to a third party will also transfer responsibility to maintain safeguards to ensure security of individually identifiable data.	***Insecurity***	Yes	No

Remedy	**Identified potential harm**	**Include in agreements transferring identifiable MNBD**	**Include in agreements transferring anonymized MNBD**
The agreement governing the transfer will include provisions to minimize risks posed by increased accessibility when data are released to third parties.	***Increased accessibility***	Yes	Yes

Group harms

Example: Socio-economic mapping

- Governments/IGOs wish to identify the poor so services may be efficiently delivered to them.
- Today, socio-economic mapping seeks to literally map or associate poverty on spatial representations. In future, may be extended beyond mapping in the literal sense. The analogy is to the zip-code-based voter mobilization efforts of past US elections versus the current precision-targeted get-out-the-vote exercises.

If the poor can be identified, so can the rich

- Will this result in prioritization of areas where the rich live in terms of service delivery?
- In competitive markets, suppliers are not expected to serve the entire market at the very outset or even at any point. Uncertainty about demand is normal. Therefore, suppliers enter in limited geographical areas or focus on particular market segments at the outset. It is only on the basis of feedback from these activities that the firm will scale up. Some firms will adopt niche strategies and never seek to serve the entire market.

Dangers of safeguarding against group harms, by creating new right of group privacy

- Rights are usually understood to belong to individuals, not to groups. The only group right recognized in international law is that of peoples having the right of self-determination
- Prejudice against actions based on group attributes would pretty much put an end to efforts to improve the functioning of society in systematic, evidence-based ways, e.g.,
  - Routine to associate various characteristics or behaviors with persons living in geographical areas (e.g., in poverty mapping), by age group and gender and so on
  - Desirable to “target” various policy measures to specific groups and indeed to improve the targeting by various means.
- Without group identification it will be impossible for modern societies to function

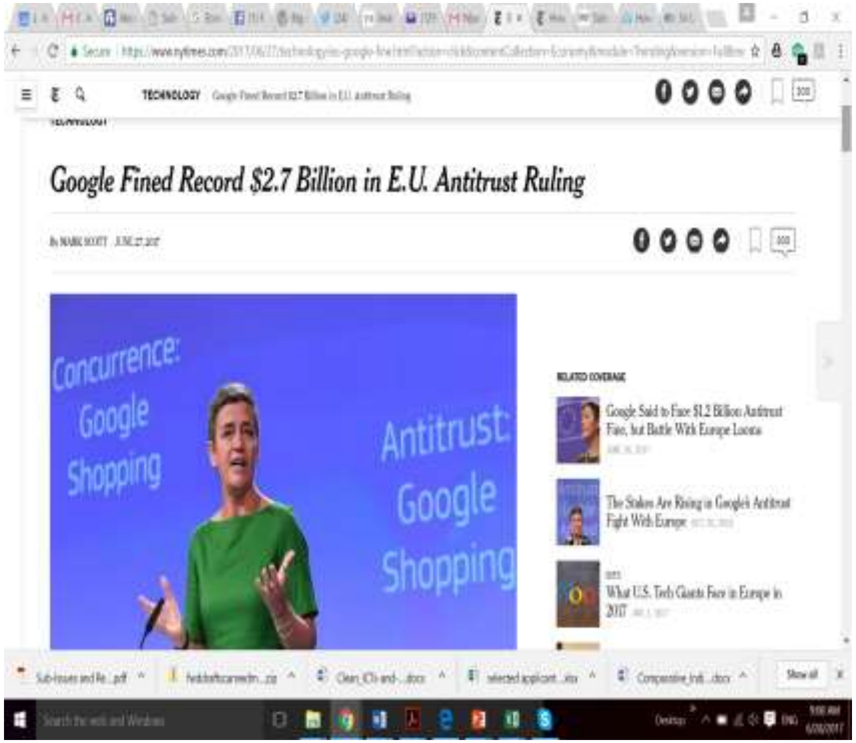
Competition harms

- If market power exists at one point of a value chain, public policy seeks to prevent its extension throughout the chain
- Privileged access to unique datasets = Market power?

Harms from algorithm-based decision making

- All decision making is “algorithm-based”
  - Rule of thumb is also an algorithm
- But in the past the rule of thumb (or better rule) could be explained, witnesses cross-examined, and biases corrected
- Now with more complex algorithms and artificial intelligence, difficulty in explaining how decisions are made
  - Is the solution algorithm regulation?

“Independent algorithm monitors”



TECHNOLOGY

Google Fined Record $2.7 Billion in E.U. Antitrust Ruling

By MARK SCOTT JUNE 27, 2017

RELATED COVERAGE

Google Said to Face $1.2 Billion Antitrust Fine, but Battle With Europe Looms

The Stakes Are Rising in Google's Antitrust Fight With Europe

What U.S. Tech Giants Face in Europe in 2017

- “Under European rules, the company — and not the regulator — must come up with proposals to guarantee that it treats competitors fairly when people make online search queries. The authorities can demand that Google make further changes if they are not satisfied with the initial proposals. **Analysts and many of Google’s competitors have called for an independent monitor to oversee the company’s digital services in Europe, which could include oversight of its search algorithms, some of Google’s most important intellectual property.** The company is likely to fiercely oppose such a move.”