Big Data and Data Sharing

David J. Hand

Imperial College, London

and

Winton Capital Management

June 2017

(As you all know) the world of data is changing

Not something ***after which*** we settle into the new world

Rather ***change is the only constant***

- *"The future you have tomorrow won't be the same future you had yesterday."* Chuck Palahniuk
- *"In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists."* Eric Hoffer
- *tomorrow's data environment will be different from today's*

Corporate change

Corporate change

For example:

Bebo: Overtook Myspace in UK

2008 sold to AOL for $850m

2013 sold to Michael and Xochi Birch for $1m

Corporate change

For example:

Bebo: Overtook Myspace in UK

2008 sold to AOL for $850m

2013 sold to Michael and Xochi Birch for $1m

Myspace: Overtook Google in US

2005 sold to News Corp for $580m

2011 sold for $35m

Technical change

- *size* *of data sets*: “**big data**”

- ***sharing*** of data
  - ***speed*** of acquisition of data: “streaming data”
  - ***diversity*** of data
  - ***source*** of data: automatic acquisition of data
  - ***societal*** aspects of data sharing

Size: Large data sets

administrative data, register-based data

- some countries (e.g. Scandinavian) ahead of the field

transaction data

- social media, Google searches, twitter messages, email transaction logs, phone logs, transport logs, ...

social media data

geospatial data

image data

text data

Data sets with billions of data points are common

And they arise as a consequence of data sharing

Sharing

Two kinds of sharing:

1) individual sharing “their own” data with larger database

Contrast data which ***needs to be retained***

e.g. hospital records

with data which ***can be discarded*** after processing

e.g. travel cards

Statistics as a study of aggregate phenomena

vs

Statistics as a study of the individual:

by *sharing* data and linking *datasets*:

Statistics as a study of aggregate phenomena

vs

Statistics as a study of the individual:

by *sharing* data and linking *datasets*:

e.g. medical treatment: combine data describing your symptoms and diagnoses with data from clinical trials and big epidemiological data which showed which treatment was most effective

e.g. credit scoring: combine data describing you and your circumstances with big data summarised in a credit scorecard

2) linking, merging, combining data sets

Sharing of data sets by public or private bodies

e.g. police forces

e.g. government departments

Challenge of combining data of diverse and heterogeneous types:

- interesting theoretical challenges

Speed: realtime data collection – and analysis

Several major implications,

e.g. 1: timeliness
e.g. 2: analytic tools and methods

1: Timeliness

Balance timeliness against accuracy

Example: UK GDP

- 1st estimate: 44% of the data available by 25 days,
- 2nd estimate: 88% by 55 days,
- 3rd estimate: 85 days

Example: inflation rate

Elaborate procedure to collect sample data

Contrast with direct recording from transactions

And from web-scraped prices

2: Analytic tools and methods

“Streaming data”:

the data keep on coming, like water from a hose

Permanently executing analytic tools, processing the data as it accumulates

- anomalies
- changes
- summaries (trends, averages, variability, maxima, ...)

Realtime → *automatic* analysis

Contrast:

(a) the familiar fixed database

(b) unable to store the data after processing

In case (b) we need to know what questions we will ask as we collect the data

We cannot later ask arbitrary questions, but only those that can be answered from our summary statistics

Summarising a stream

Subsetting a stream: sampling, but requires different approaches from classical survey sampling

Filtering a stream: accept only those cases which meet some criterion

Diversity of data

Survey, census, administrative, transaction, experimental, ...

Numerical tables, image, text, signal, networks, ...

Different kinds of data have different properties

e.g. survey data: answers to the questions you choose but slow and expensive to collect, response bias?

e.g. transaction data: fine granularity, both spatial and temporal, immediate, but may not address the question you want

→ ***an opportunity***:

Data of different kinds can be combined synergistically, to overcome the problems of each individual kind

Stitching different kinds of data together

- *Linking*
- *Matching*
- *Merging*
- ***Sharing***

Technical challenges have begun to be addressed in different fields

- e.g. medical combination of information from scans with traditional numeric, text, and image data
- e.g. administrative and survey data

“survey and census data is what people *say*: administrative and transaction data is what people *do*”

“survey and census data is what people *say*: administrative and transaction data is what people *do*”

New forms of data are closer to social reality ?

Source: Modern data capture technologies

Automatic data collection:

- electronic measurements: point of sale credit card terminals, petrol pumps, contactless travel cards, phone records, emails, GPS, CCTV cameras, ...

“Properties” of automatic data collection:

- immediate
- complete ???
- untouched by human hands ???

Internet of things

Social media data – data directly from the web

Administrative data

Data not primarily collected for research purposes

e.g. supermarket purchases, credit card transactions, tax records, education records, health records, transport movements, ...

Administrative data research is ***secondary*** analysis, so

- may not be ideal for the research purpose
- issues of consent may arise
- changes to the collection procedures may change nature of data
- quality issues different from those of surveys
- selection distortion – who’s in the database?

The *Administrative Data Research Network*

Aim: "*to facilitate access to and linkage of de-identified administrative data routinely collected by government departments and other public sector organisations*"

Four centres: England, NI, Scotland, Wales + ADS

Partnerships with Nat Stats Institutes

UK-wide governance

Safe and secure data access

Accredited researchers

Public engagement

Societal aspects of data sharing

Confidentiality

Often unclear what should be regarded as confidential, or indeed what it's feasible to regard as so.

Is the fact that we are here at this meeting confidential?

Ipsos MORI 2014 survey of public attitudes to the use and sharing of their data:

> Revealing an intrinsic suspicion of potential data sharing, but coupled with an increased enthusiasm for shared data when the advantages were spelt out

Trust

People readily give data to supermarkets, travel companies, phone companies, credit card companies,..

Concern about government misuse, targeting subgroups

The importance of formal separation of statistical offices from government

Privacy

Amongst the main conclusions as to what people *think* about data privacy were:

- Losing data is one of the worst things a company can do;
- Selling *anonymous* data is not far behind;
- A sense that data sharing is inevitable in modern world;
- Very few think either government or companies have their best interests at heart when using data;
- Both government and internet companies are a threat to privacy – but especially internet companies.

In preparation for the UK's 2021 census, which is to use administrative data as well as data collected by more conventional means, the ONS also carried out a programme of work exploring public attitudes (ONS, 2014):

- there is generally a very low level of public understanding about data, how it is collected and used;
- the public generally does not understand the difference between operational and statistical uses of personal data;
- nearly half of the public assume that government already routinely links data about the population from multiple sources in a central data store;
- around three quarters of people do not object to data held by other government departments being shared with ONS;

- the public are supportive of data sharing when personal or public benefit can be demonstrated and these are communicated effectively;
- data linking and storage is more acceptable if the personal data are anonymised;
- any objections to the use of personal data are largely related to security and privacy concerns;
- the public is generally positive towards the decennial census as a means of gathering information about the population; and
- when provided with reassurance with regard to security and privacy, the public broadly support ONS re-using administrative data to produce statistics.

Is it sensible to speak of public attitude?

- diverse
- volatile (depends on media reports)

When presented with cogent and balanced arguments the public are sympathetic to data sharing

Monopolies can be broken by *requiring* them to share their data with other organisations.

- analogy to journals requiring researchers to publish their data

Downside

- if data took individuals years to collect
- if the IP is within the data

Upside

- more eyes, more chance of important discoveries
- quality improvement
- false claims recognised

Conclusion

Consent may be difficult to require or grant

- Data being used for multiple purposes
- Down the line
- Individuals irrelevant

- Distinguish operational from aggregate

Conclusion

Consent may be difficult to require or grant

- Data being used for multiple purposes
- Down the line
- Individuals irrelevant

- Distinguish operational from aggregate

Is data ownership a meaningful concept

- Who owns the fact that you are here at this meeting?

thank you