Big Data and Knowledge

Fiona Allen

Australian Parliamentary Library

The physical and the conceptual

- Endless growth
- Structured
  - Logs, RFID data, financial transactions, information from wearable devices, geolocation details, medical tests, records from sensor-equipped infrastructure, other types of automated record keeping
- Unstructured—as much as 95%
  - email messages, social media posts, photographs, content of internet forums, telephone transcripts
- Meaning can be imprecise and uncertain
- Requires specialised infrastructure—physical and conceptual

PARLIAMENT OF AUSTRALIA
DEPARTMENT OF PARLIAMENTARY SERVICES

PARLIAMENTARY
LIBRARY
INFORMATION | ANALYSIS | ADVICE

Do we risk creating knowledge but no understanding?

PARLIAMENT OF AUSTRALIA
DEPARTMENT OF PARLIAMENTARY SERVICES

PARLIAMENTARY
LIBRARY
INFORMATION | ANALYSIS | ADVICE

The applications

- Public policy and service delivery
  - patterns of crime, public health trends, traffic congestion, planning of public infrastructure, and provision of medical, educational, and welfare services
- Research
  - predict and contest elections, medical research, climate and weather analysis, earthquake risk analysis
- Commercial
  - examine economic behaviour, track communication dynamics, operation of financial markets, online and traditional commerce, and manage and control machines, factories, and other infrastructure
- Security
  - understand conflict and violence, combat terrorism, intelligence gathering

The questions

- Collect—store—manage—disseminate?
  - storage—physically in servers, computationally in aggregated and relational databases
  - curation—taxonomies, metadata standards, retrieval methods
  - preservation—forward and backward compatibility
  - dissemination—access and usability
- Analyse?
  - algorithms
  - teams of computer scientists, statisticians, economics, social scientists, anthropologists, psychologists, and more
- Use the studies produced by others?

The interpretations …

- Raw data and objectivity
  - Generated, norms and classifications
  - inclusion and exclusion
  - scrubbing and cleaning
  - prior organisation
- Size and representation
- Errors
- Commensurability
- Aggregation and meaning
  - combining unstructured data—social media posts, photographs, conversation streams, geolocation—so that it can be meaningfully evaluated
  - interpretation in unstructured data—quantification of meaning

… pile up …

- Analytical tools
    - algorithms—complex mathematical equations that instruct a computer to perform certain actions
    - reflect biases of programmers and datasets, contain embedded judgements, generally subjective, dependent on limitations of data
    - human beings as 'inputs' not complex beings capable of an array of thoughts, feelings, desires, hopes, perspectives

… and keep piling up …

- Predictive policing
  - collective term—variety of analytical tools and subsequent law enforcement practices
  - identify where crime will occur and direct deployment of law enforcement officers
- Data collection phase assumptions
  - data is accurate
  - systemic problem of gap between crimes committed and reported—type of crime, characteristics of the victim, bias of law enforcement officers
  - categorisation and classification, decision to record as a crime
  - feedback loop—the more police are present, the more crimes will be reported
  - algorithm becomes implicated in the process of prediction—subsequent actions of law enforcement officers affects that which is being measured (crime)
  - impossible to measure predictive accuracy and effectiveness concurrently

… endlessly

- Data analysis phase
  - based upon particular model of crime—crime rates in different areas at different times assumed to be constant
  - variables assumed relevant depend on information collected about crime—depends on how much is known about a particular crime; some variables omitted because expensive or difficult legally to collect
  - identification of correlations—difficult to understand causes
- Data can't speak for itself

The correlations

- Identify correlations not causes
- Attributes of previously unrelated datasets are comparable and can reveal something of significance about a particular issue
- Skill of data analyst and familiarity with data set important and source of variability
  - find different patterns each time a data set is analysed
  - risk in every data set of finding false correlations that appear statistically significant
  - the larger the dataset, the greater the number of correlations that can be found
  - no guarantee correlations are accurate
- Big data sets can identify correlations that would be otherwise missed, cannot always determine if they are meaningful

The questions not asked

- Arab Spring with twitter
  - using twitter keyword analysis, we can predict dissent and regime change
  - are we looking for answers in places where none exist?
  - can we reduce the dynamic of events in Egypt to measuring how often certain words were used on twitter?
- Questions that can't be asked
  - conditions of possibility—what was happening on the ground, what were the complex factors feeding into the discontent, what was the diversity of opinion, what were the myriad factors motivating each individual
  - importance of contextual factors—government policies and actions, state of the economy, weather, support and agenda of broader societal groupings
- The influence of the medium
  - limit characters, use hashtags, gather followers, and truncate language and emotion

PARLIAMENT OF AUSTRALIA
DEPARTMENT OF PARLIAMENTARY SERVICES

PARLIAMENTARY
LIBRARY
INFORMATION | ANALYSIS | ADVICE

The presuppositions

- We can truly and accurately understand phenomena
- Radicalisation
  - data on characteristics of people who are radicalised—age, socio-economic status, residence, friends, religion, views, opinions, behaviours, frustrations
  - assumptions of interpretation and meaning—require an entire system of meanings for a proposition to be sensible, this system is not fixed

The way to where?

- Promising potential

- Colonising potential

- To evaluate the quality of research, we must understand the substance of the assumptions and methods by which it was produced

- All research is marked by interpretations, blind spots, and questions not asked or capable of being asked