Session 2

- Data & data sources
- Data from Twitter streams
- More Spark
- Exercises 1 & 2
- INFO319:
  - practical information
  - background / expectations
  - programming projects
  - essay

Data
(Kitchin's chapter 1, also a little from 2-3)

What are data?

- Not general agreement!
- Usually agreed on properties:
  - material (matter or energy at bottom)
  - this material basis can vary (lack of uniformity)
  - the variations (or lack of v.) represent something
- Representation:
  - direct correspondence:
    - “the property/state of the data corresponds to some property/state of something else” (natural/intentional)
  - symbolic correspondence:
    - “the data contain symbolic language that describes something else” (intentional)

What are data?

- “Non-data”: material variations (or lack of v.) that do not represent anything
- “Natural” data, data in the wild: material variations (or lack of v.) in nature that represent something else
- Human-made, artificial data: ← *We are (mostly) here!*
  - material variations (or lack of v.) that represent something else by human action
  - direct, “hand-made” artificial data
  - indirect, machine-generated artificial data
  - non-recorded, recorded ← *We are (mostly) here!*
  - non-rivalrous, non-excludable, marginally free
- Data do not only represent, they also constitute reality

Luciano Floridi: Information – A Very Short Introduction

Data are not information!

- A common distinction:
  - data may accommodate multiple interpretations
  - information = data + meaning
    - ...the interpretation has become (more) fixed
- Information is carried by (constituted by) data,
  - but is not bound to particular data:
  - a letter can be scanned into a PDF file.
  - when the letter is shredded, the data are lost.
  - but the information is still there in the PDF
- Data are in themselves, but the same information can be carried by different data in different forms at the same time (or different times)

(Some) types of data

- Analog and digital
- Qualitative and quantitative
  - nominal, ordinal, interval, ratio for quantitative
- Structured, semi-structured and unstructured
- Primary (main purpose) and exhaust (side effect)
  - secondary, tertiary
- Metadata:
  - about content: syntax, semantics
  - about dataset: descriptive, structural, administrative
- Indexical and attributive
- Small and big!

Shapers of data

- Collected data are not neutral, but shaped by:
  - prevailing power structures
  - background and interests of collectors
  - data generation context
  - field of view / sampling frame
  - technology and platform used
  - data model / ontology
  - regulatory environment:
    - e.g., privacy, data protection, security
- Big data tend to be opportunistic / convenient
  - small data tend to be purposeful / targetted

Big and small data

	Small data	**Big data**
Volume	Limited to large	Large to very large
Exhaustivity	Samples	Entire populations
Resolution and identification	Coarse and weak to tight	Tight and strong
Relationality	Weak to strong	Strong
Velocity	Slow, freeze-framed, bundled	Fast, continuous
Variety	Limited to wide	Wide
Flexible and scalable	Low to middling	High
Origin	Targetted, purposeful	Convenient, opportunistic
Purpose	Specific	Generic

Qualities of open data sets and sites

- Clean, high-quality, validated, interoperable
- Comply with data standards
- Associated metadata and documentation
- Preservation, backup and auditing policies
- Reuse, privacy and ethics policies
- Administrative arrangements, management organisation, governance mechanism, financial stability
- Long-term plan for development and sustainability

Open data

- Example definition:
  - “knowledge is open if anyone is free to access, use, modify, and share it
  - subject, at most, to measures that preserve provenance and openness”
- Requirements:
  - *technically open:* open, standard format, physical availability, no-DRM or similar constraints (DRM: Digital Rights Management software)
  - *legally open:* no legal restrictions, explicit open licences
- Examples:
  - OpenDefinition of Knowledge: http://opendefinition.org/od/2.1/en/
  - Open Government Data, 8+7 principles: https://opengovdata.org/
- From product to service thinking?

Why open data?

- Long tradition in some countries
  - ...other are opening up
- Drivers:
  - measure success of (public) organisations, decision making, transparency, accountability, value for money
  - active and informed citizenship: choosing schools and hospitals, political involvement, participative democracy, social innovation
  - evidence-based monitoring and decision making, improved operational efficiency, competence and productivity, using information across departments, broader ("holistic") views of organisations, more eyes
  - low economic value → high commercial value, e.g. map data
  - brand enrichment, customer contact, trust and reputation

Why open data?

- Obstacles:
  - first-time preparation has a cost
  - requires repurposing
  - curation (anonymity, aggregation)
  - developing new systems / services
  - partly market-financed state agencies
  - legal limitations:
  - public / private sector competition
  - lobbying from third-party resellers

Funding open data

- Arguments for direct government backing:
  - increased societal costs are offset by reduced company costs
  - free additional labour, improved data quality, crowd innovation
  - simpler, better, more efficient customer-handling
  - diverse consumer surplus value
  - new innovations and markets (GPS!), corporate revenue, corporate tax

Funding open data

- Funding models for open data:
  - premium version of free product/service
  - freemium product/service (graded options)
  - open source
  - free trial (razor), then paid (blades)
  - value-added services (i.e., semantics)
  - product/service store
  - advertising
  - customisation
- ...resembles the funding models for open software!

Concern: neoliberal and market interests

- Open data are not neutral
- Example claims:
  - driven by commercial forces
  - exploiting public goods for private benefit
    - in turn weakens public data resources
    - must perhaps buy back from private sector
  - public accountability drives neoliberal, NPM reorganisation
  - transparency talk is just rhetorical
    - business interests in disguise
    - not similar support for whistleblowing, IP liberalisation, DRM-restrictions etc.

Concern: benefits the already empowered

- Open data is not all
  - also: which data, and how they can change society
- Example claims:
  - open public data enhance value of privately held data
  - data are not neutral:
    - which data to collect, generate and make open about who and what is highly political
    - which interests are included, which excluded?
    - leveraging open data is labour and skill intensive: technological, contextual, argumentative...
  - two types of public data: operational and citizen

Concern: sustainability and utility

- Supply focus
  - many open data sites / sets are low-hanging fruit
  - more interesting data sets may require curation
    - e.g., repurposing, privacy concerns, regulation
  - created by volunteers, short-term projects
  - less focus on maintenance over time
  - danger of vicious cycles
  - shift needed:
    - holdings → archives → infrastructures

Concern: sustainability and utility

- Supply focus
  - many open data sites / sets are low-hanging fruit
  - more interesting data sets may require curation
    - e.g., repurposing, privacy concerns, regulation
  - created by volunteers, short-term projects
  - less focus on maintenance over time
  - danger of vicious cycles
  - shift needed:
    - holdings → archives → infrastructures

Enablers of big data

Enablers of big data

- Computation
  - “Moore's law” of transistor numbers (1965 –)
  - from *more powerful* cores to *more cores* (20005 –)
- Networking
  - “Gilder's law” of network bandwidth (2000 –):
  - global *bandwith* doubles every 6 months
- Storage (cloud, *aaS, NOSQL)
- Pervasive and ubiquitous computing
  - sensors and actuators
  - from dumb to smart things (cars)
  - exhaustive data collection
  - *“ambient computing”, “the age of everyware”*
- Standardised identifiers *(e.g., the URIs again)*

Pervasive versus ubiquitous computing

- Pervasive computing:
  - computing “in everything”
  - make them interactive and smart
  - divergent: more and more things become smart
  - needs situational awareness
- Ubiquitous computing:
  - computing “in every place”
  - moves with the person
  - convergent: smart things we carry do more and more
  - needs context and location awareness

Overlapping concepts / ideas

- Internet of Things (IoT):
  - sensors, actuators, other devices are on the internet
  - TCP / IP → IP4 / 6 addresses of their own
    - ...connected through gateways
- Web of Things (WoT):
  - web APIs, HTTP, HTML / CSS / JS dashboards / UIs
  - sensoring, actuating etc. *-as-a-Service*
- Cloud of Things (ClouT):
  - gateways and services can be *hosted*
  - *digital twins* with *histories, extrapolations, simulations*...
  - the *cloud*, the *edge* and the *fog*
- ...a bit hyped as usual, but at least a quantitative change going on

Enablers of big data

- Indexicality
  - growth of unique identifiers
  - people: user names/handles, personal numbers / SSNs, passports, driver's licences, health cards, biometry, IMSIs
  - things (and information): product type codes (bar code, QR code), RFID for individual products, auto passes, MAC addresses (medium access control), IMEIs, URIs, including ISBNs, ISSNs, DOIs, etc.
  - places: post codes, addresses, geo coordinates
- Machine-readable identification
  - more and more are becoming digital
    - ...and remotely readable

Sources of big data

- Three types:
  - directed
  - automated
  - volunteered
- Directed data collection
  - organised and structured surveillance
  - personal or through technological lens
  - census, government forms, inspections, CCTV cams
  - surveillance technology is becoming *digital, smarter, directable, internetworked…*
- Big data too are a *representation* and a *sample*
  - there are no “raw”, only “cooked” big data

Automated data collection

- Automated surveillance
  - e.g., smart electricity meters, electronic transportation tickets, passenger counting systems, car tolls, radar/lidar speed guns, ANPR
- Digital devices
  - smart phones/tablets + lots of others
  - actively produce data
  - primary: cameras, videos, GPS units, medical devices
  - exhaust: mobile phones (also primary), cable boxes
  - logjects = objects that log their (+ their users') history
  - object logs can also be re-combined
    - e.g., GPS-car data combination

Automated data collection

- Interaction data
  - all ICT-based transactions leave traces
  - using a web shop, net bank, ATM
  - sending an email
  - accessing the internet from home or a mobile device
- Scan data
  - machine-readable identification codes
  - barcodes, QR (“Quick Response”) codes
  - magnetic cards, chip card/smart card/ICC
- Sensor (sensed) data
  - inexpensive sensors generate continuous data streams
  - smart cities gauging noise, temperature, light, CO2...

Volunteered data collection

- Social media, collective projects (online)
  - production + consumption = prosumption
- Transactions
  - voluntary registration, clickstreams, review data
- Some of the automated collection was volunteered:
  - actively produced data
  - primary: cameras, videos, GPS units, medical devices
  - some logjects (objects that log history)
- Sousveillance
  - *(fr.) sur*-: above, *sous*-: below
  - self-monitoring, e.g.,
    wearable fitness equipment, dieting apps, ...

Volunteered data collection

- Social media
- Crowdsourcing
  - to create one new product
  - to create many new products/concepts/ideas
  - to assess many existing products/concepts/ideas
- Citizen science

Some
big data sources
for news

Example big data sources

- Social media (pre and post news):
  - Twitter's open API (https://developer.twitter.com/en)
  - Meta's CrowdTangle API (https://www.crowdtangle.com/)
  - YouTube Search API (https://developers.google.com/youtube/v3/docs/search/list)
  - Reddit (https://www.reddit.com/dev/api)
  - TikTok API (forthoming)
- Published news (post news – ...from raw to processed…):
  - NewsAPI (https://newsapi.org/)
  - EU Media Monitor (https://emm.newsbrief.eu/overview.html)
  - DataMinR (https://www.dataminr.com/)
  - GDELT (https://www.gdeltproject.org/)

GDELT

The GDELT project (→ from INFO216)

- Global Database of Events, Language, and Tone (GDELT)
  - free open platform
  - monitors the world’s broadcast, print, and web news
  - focus on crises, but much broader in practice
  - globally in over 100 languages
  - identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images, events
  - *“can we map happiness and conflict, provide insight to vulnerable populations and even potentially forecast global conflict in ways that allow us as a society to come together to deescalate tensions, counter extremism, and break down cultural barriers?”*

https://www.gdeltproject.org/

The GDELT project

- Archives back to 1979 (expanding back to 1800)
- Increasingly integrating social media
- Translations from 65 languages into English
- Supported by Google
  - runs in the Google Cloud
- Almost a knowledge graph, but
  - not native RDF
  - not fully linked
  - no ontology

https://www.gdeltproject.org/

The GDELT project: data streams

- Downloadable CSV files (every 15 minutes)
  - http://data.gdeltproject.org/gdeltv2/lastupdate.txt
  - *Events* (...export.CSV, ~400k)
    - low-level actor - event type – actor triples
  - *Mentions* (...mentions.CSV, ~600k)
    - where in and which source is each event mentioned?
  - *Global Knowledge Graph* (...gkg.CSV, ~50M)
    - which people, locations, organizations, themes, sources, emotions, counts, quotes, images, events are mentioned where and in which source?
  - Also available as Google BigQuery tables
- Lots of other datasets and streams, raw and analysed, native language or translated to English

https://www.gdeltproject.org/

The GDELT project: data streams

- Other data streams:
  - *Visual GKG*
    - codifying the world's news images in real time
    - random sampling, Google’s Vision API
  - *Global Entity Graph*
    - experimental, random sampling of news articles
    - deep learning, Google’s Natural Language API
    - provides Wikidata links for entities
  - *Global Relationship Graph*
    - experimental, related to the global entity graph
    - extracts verbs and the words in their context
    - groups new articles with similar verbs-in-context

https://www.gdeltproject.org/

The GDELT project: Events 2.0

- For each event:
  - global event id and datetime
  - actor 1 and 2:
    - name (person, organisation, location, ethnicity, religion, type) and CAMEO code
  - event:
    - CAMEO code and importance of event type
    - numbers of mentions and sources, tone
  - geography
- Codebooks
  - http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf
  - http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

https://www.gdeltproject.org/

The GDELT project: Mentions 2.0

- For each event
  - global event id and datetime
  - mention type and datetime
  - source name and identifier (e.g., a URL)
  - sentence number
  - actor 1 and 2 mentions (character indices)
  - confidence
  - source length and tone
- Codebook
  - http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf

https://www.gdeltproject.org/

The GDELT project: GKG 2.0

- For each document:
  - record id and datetime
  - source and document identifier (e.g., a URL)
  - keywords/themes (taxonomies of 50k keywords)
  - person and organisation names and types
  - locations, they types, names, geo-coordinates
  - counts, their types and counted objects
  - average tone, positive/negative score, polarity
  - ...and lots of other stuff
- Codebook
  - http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf

https://www.gdeltproject.org/