Lecture 4:
Data Collection & Quality

Seminar ‘Foundations of Data Science’

Prof. Dr. Karsten Donnay, Assistant: Marcel Blum

Course Outline

– Part 1: Foundations

- *Day 1: Mon. 14.06.2021*: Information Coding & Data
- *Day 2: Tue. 15.06.2021*: Programming & Algorithms
- *Day 3: Wed. 16.06.2021*: Complexity & Efficiency

– Part 2: Applications

- ***Day 4: Thu. 17.06.2021*: Data Collection & Quality**
- *Day 5: Fri. 18.06.2021*: Research on Digital Media

Overview of this Session

- Data Collection
  - Digital Trace Data
  - Data Sources
  - Ethical Considerations
  - Regulatory Frameworks
- Data Quality
  - General Characteristics
  - Methodological Problems
  - Digital Trace Data & AI

University of Zurich[UZH]

Department of Political Science

Data Collection

Digital Trace Data – A Revolution?

[J]ust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact … . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin.

—Duncan Watts (2011, p. 266) “Everything is Obvious”

Digital Trace Data = Big Data?

- The terminology largely refers to the exact same thing
  - Digital Trace Data
    - Largely used in more academic settings
    - Definition refers closely to the sources and character of these data
      - Data from digital platforms (of all kinds)
      - Trace data in that sense that is is artifacts of techno-social systems
  - Big Data
    - Term coined in industry setting (business intelligence)
    - Concept more closely refers to one key characteristic
      - Sheer size of the data vs. manually collected ”small” data
      - Terminology is often used interchangeably for modern data science

What is Big Data?

"Big Data" refers to a combination of an approach to **informing decision making with analytical insight** derived from data, and a set of **enabling technologies** that enable that insight to be economically derived from at times **very large, diverse sources of data.**

John Akred
Founder and CTO, Silicon Valley Data Science

"Big Data" will ultimately describe any dataset large enough to necessitate **high-level programming skill** and statistically **defensible methodologies** in order to transform the data asset into something of value.

Reid Bryant
Data Scientist, Brooks Bell

Sources: "What is Big Data?" and "12 Big Data Definitions: What's yours?"

What is Big Data?

Big data is at the intersection of **collecting, organizing, storing**, and turning all of that **raw data into truly meaningful information**.

Prakash Nanduri
Co-Founder, CEO and President, Paxata, Inc

Big data refers to the approach to data of **"collect now, sort out later"**…meaning you capture and store data on a very large volume of actions and transactions of different types, on a continuous basis, in order to **make sense of it later**.

Rohan Deuskar
CEO and Co-Founder, Stylitics

Sources: "What is Big Data?" and "12 Big Data Definitions: What's yours?"

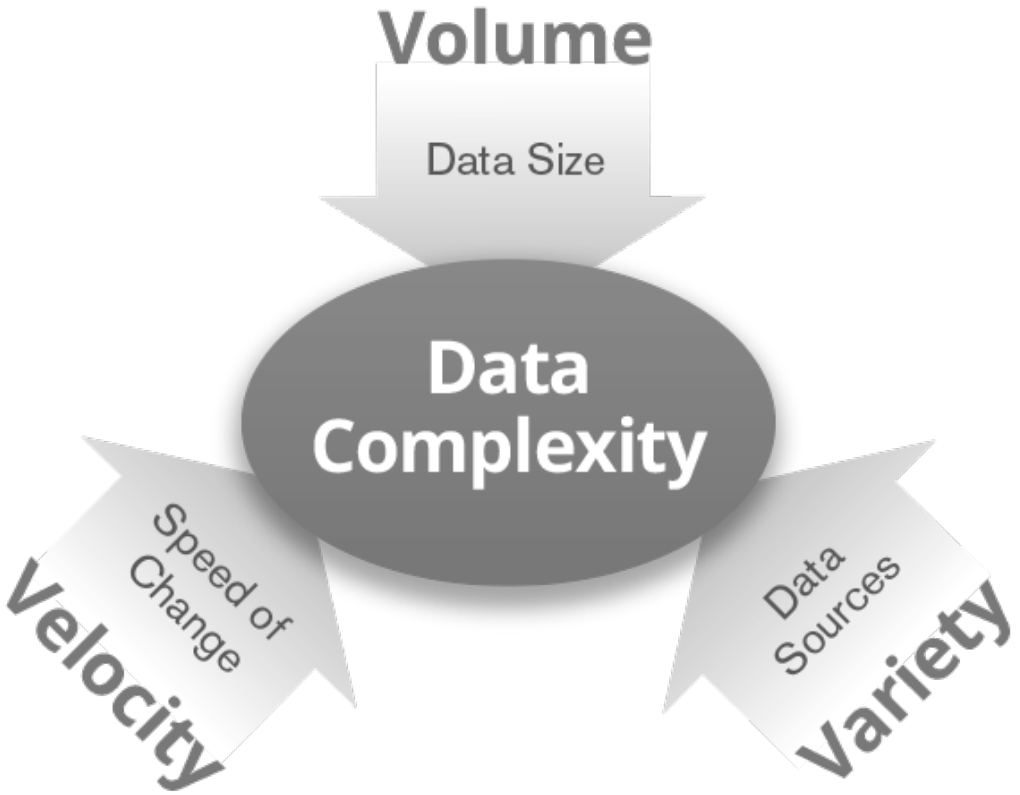
Definitions

- There is not one clear definition of what constitutes Big Data/Digital Trace Data but there are distinct features that characterize it
  - **Technical:**
    Machine-readable data that has been automatically collect and is too large to be processed without the help of specialized processes and software for data handling and processing.
  - **Conceptual:**
    Usually massive, very complex and messy data on all kinds of human interactions that is collected without a specific research question in mind.

Defining Characteristics

- Not research data:
  - What is collected from whom and when is in most cases outside of our control because of post-hoc collection
    - Thus also referred to as “found data” or (trace data)
  - Data is characterized only by either
    - Original purpose for collecting it
    - Parameters of the specific process it arises from
  - Often no standard format or coding:
    - How data is recorded and information is coded is outside of or control
    - Implies uncertainty about standards and stability of coding rules
    - Often most research-relevant pieces of information are not recorded

Defining Characteristics



Source: www.datameer.com

Data Sources

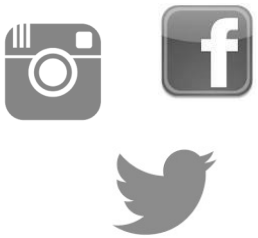
- Two main conceptual ways in which Big Data is generated:
  - **Systematic** and deliberate (automatized) collection:
    - Traffic drones, customer tracking, government databases, monitoring systems for financial transactions, immigration etc., Twitter user behavior
  - **Artifacts** of human interactions with/on techno-social systems:
    - Mobile metadata, WiFi or Bluetooth usage, software logs, Tweets, Reddit, blog posts, websites

Data Sources

website data

device data

social media data



administrative data

proprietary company data

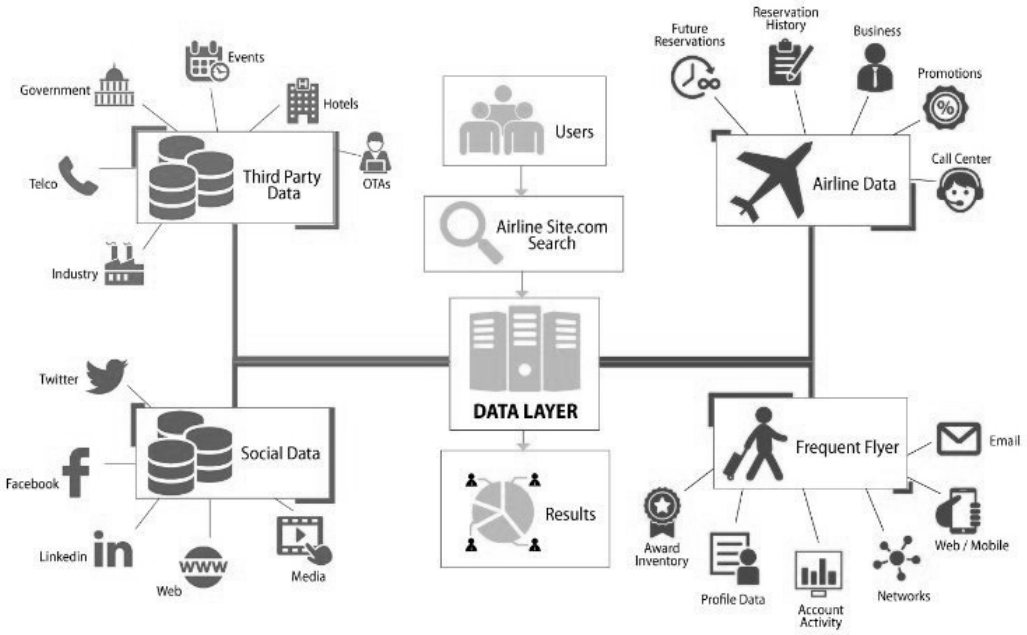
blogs etc.

Data Sources

- **Where** is data from?
  - Systematic collection
  - Artifacts of human interactions
- **What** information?
  - Specific
  - Generic
- **How** often updated?
  - Continuously
  - At varying intervals

Data Sources

– Example: Modeling Air Fares



Source: traveldatadaily.com

Ethical Considerations

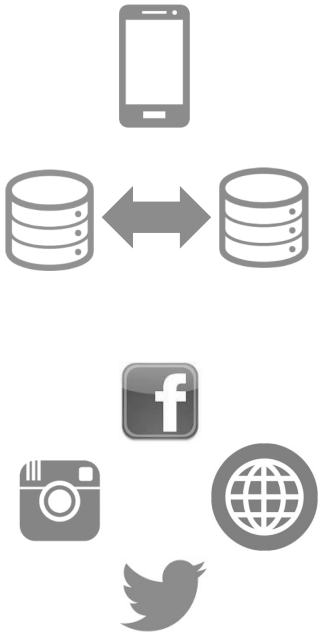
- **Trawler-net approach** to public data
  - Public private data
    - Individuals provide public information
    - Analytics generate link among data
    - Reveals unintended private information
    - e.g. public tweets or phone numbers; public work profiles
  - Anonymous data may not be anonymous
    - Data stripped of personal information
    - Reveal identity by combining with public data
    - e.g. anonymous mobile phone traces + public records

Ethical Considerations

- No real informed consent for internal data of online platforms
  - Unclear scope of analytics at time of collection
    - Difficult to give informed consent
    - No knowledge that behavior is tracked
  - Unclear secondary data use
    - Data sold to or shared with third parties
    - Private information enriched with public data
  - Legally covered under very generic blanket agreements

Ethical Considerations

- Privacy concerns
  - **Governments** systematically collect data
    - Blanket collection of phone records without cause
    - Linking of unrelated government databases
    - Monitoring of social media and online activity
    - Trade-off between invasion of privacy and (legitimate) security concerns



Ethical Considerations

- Privacy concerns
  - **Companies** systematically collect data
    - Personalized tracking of search or browsing activity, e.g. by Google, Facebook or Amazon
    - Advertisement tracking “off-site” without explicit consent of the users
    - Data and/or results of analytics are sold for profit to third parties

Ethical Considerations

- Implications for research
  - Large gray area in terms of secondary data use for research purposes
  - Example: Twitter
    - Public tweets can be obtained and analyzed
    - Users are aware that content it public but they do not know the scope of analytics
    - Can they truly give informed consent?
  - Public domain data more generally
    - Legally, research use is almost always allowed
    - Ethically, some things that are possible are quite questionable…

Regulatory Frameworks

– Largely uncurbed access for governments
  - No clear standards for privacy and individual data protection
  - Low regulatory oversight
  - Internet traffic is heavily monitored
  - Public opinion on surveillance is mixed

Regulatory Frameworks

- Recent progress in Europe with broader implications:
  - Key EU privacy legislation:
    - "Right to be Forgotten" (2014)
    - "General Data Protection Regulation" – GDPR (2019)
  - Curb private sector usage of personal information:
    - States continue to collect data
    - EU does not interfere with members
- Legal frameworks are generally still very underdeveloped
  - Not originally designed for internet communication

Regulatory Frameworks

- Implications of GDPR for research
  - Stricter norms and standards
    - Affects mostly primary data collection
    - Requires clearer declaration of user rights up front
    - Key user right is to deletion of information
    - Standards for storage of personal data and review of safety protocols etc.
  - Remaining uncertainty for research use
    - GDPR exempts research from some regulations but in practice a lot of details are unclear
    - Enforcement is potentially severe and institutions tend to err on the side of caution

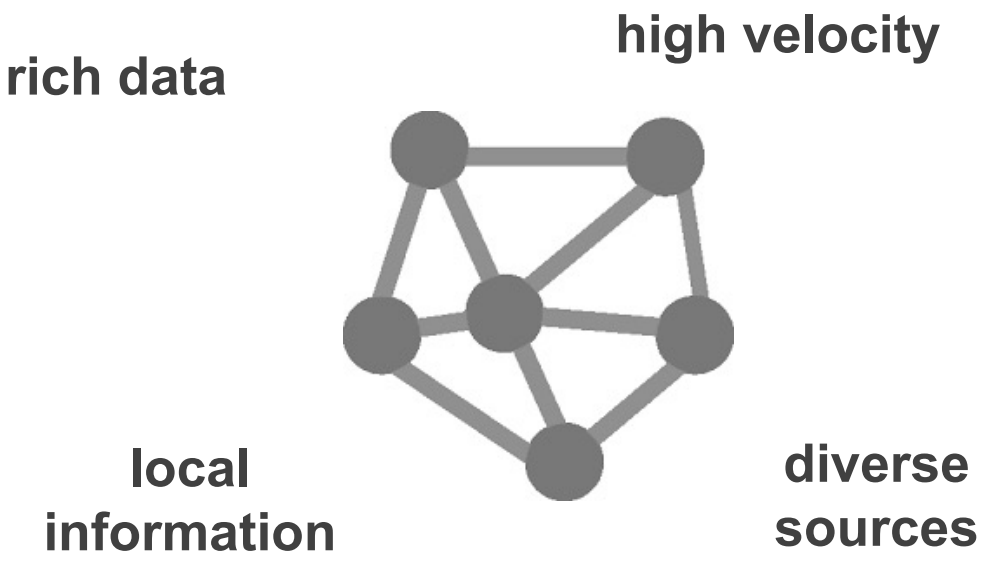
University of Zurich[UZH]

Department of Political Science

Data Quality

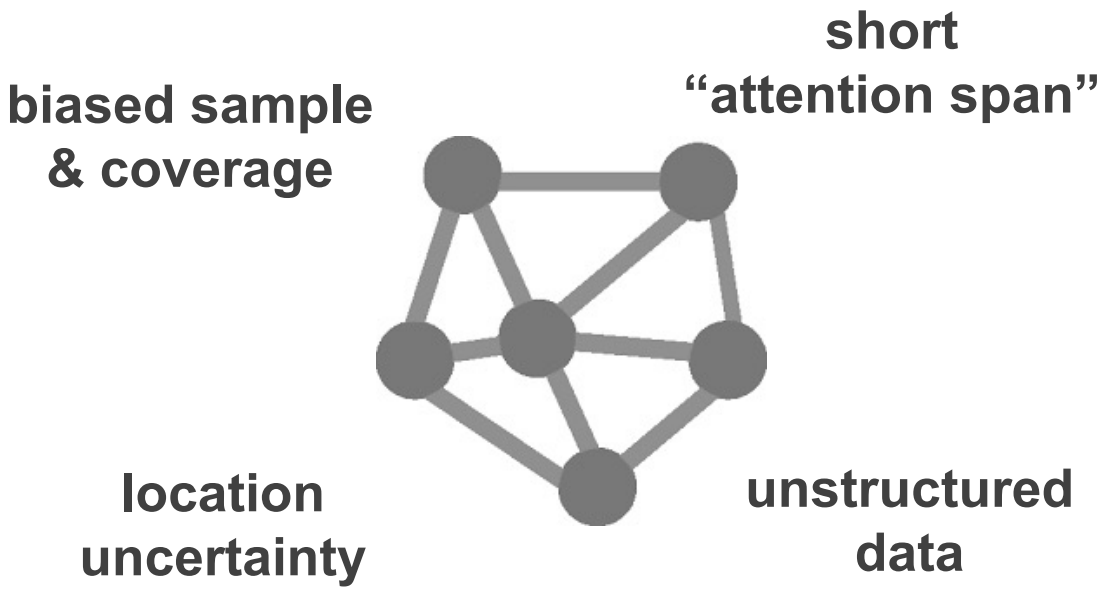
General Characteristics

– Huge **potential** as research data



General Characteristics

– Fundamental **challenges** for substantive analyses



Why Worry About Limitations?

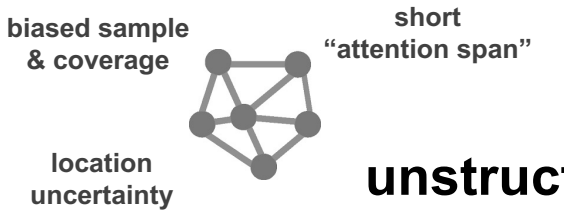
- "Big Data" means big problems?
  - Digital Trace data are not inherently better than any other data:
    - Suffers from all the same conceptual and methodological problems inherent to our "usual" social science data
    - Data collection is typically even more unsystematic than for other social science datasets
  - There are new problems arising from its sheer size:
    - Some of them are technical (beyond storage and handling)
    - Others are inferential, i.e.,
      - Is sampling precision still meaningful in a gigantic sample?
      - Correlation vs. causation
      - Multiple-comparison problem

Why Worry About Limitations?

- "Big Data" means big problems?
  - Temptation of purely data-driven analysis:
    - Often not more than a mere "fishing" expedition for systematic relationships in data
    - Fine line between "making sense" of patterns with theory and finding any pattern and arguing why it should be there
  - Interdisciplinarity is an asset but can also be a challenge:
    - Field of data science is currently dominated by technical applications from computer science
    - Too little attention to long-standing insights even for the systems that are now re-analyzed with new data
    - Lack of engagement with or knowledge of prior work significantly hampers substantive progress

Methodological Challenges

– Concrete technical and methodological **limitations**



extensive data cleaning

poor signal to noise ratio

algorithm dependent

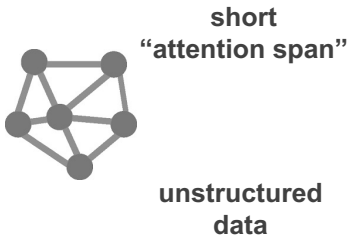
inherently subjective

NLP limitations

Methodological Challenges

– Concrete technical and methodological **limitations**

biased sample & coverage



changing sample properties

sampling error

sampling bias

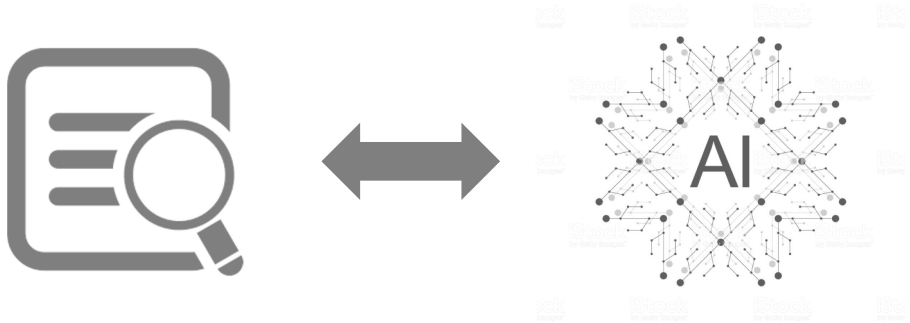
unknown sample population

Methodological Challenges

- Further methodological shortcomings
  - Inferential problems:
    - Correlation is not causation
    - Multiple-comparisons problem
    - Sampling precision vs. meaningful results
  - Design problems:
    - Data not collected for specific analysis
    - Miss decisive pieces of information
    - Unstable sample populations

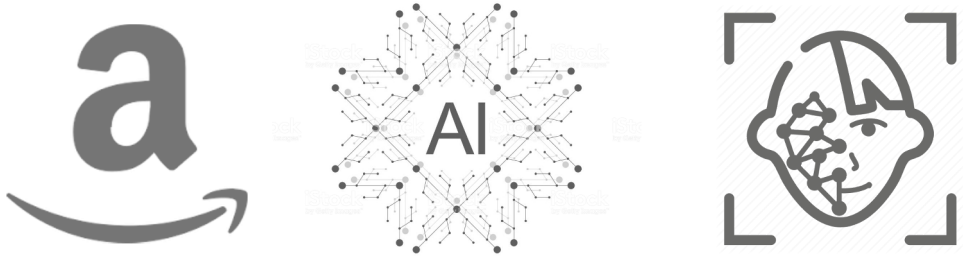
Digital Trace Data & AI

– Digital Trace Data as **fuel** for machine learning and AI
  - Rapid progress in development of machine learning and AI is directly linked to new abundance of data
  - Vast volumes of new data enabled these technologies and actually made them necessary
  - Powerful means of extracting “meaning” from massive collections of (often unstructured) data



Digital Trace Data & AI

– Risks of naïve machine learning and AI approaches
  - Inherent bias of big data sources is learned and directly translates to outcomes
  - Danger of putting predictive power and correlations over substantial insights
  - Often very weak out-of-sample performance, especially in complex settings



Digital Trace Data & AI

- Implications for research practice
  - Automatization from data science approaches is not everything
    - Necessary part of the processing pipeline for data retrieval, processing and analysis
    - Cannot replace ground-truthing and/or careful evaluation by researchers
  - Developing unbiased and carefully crafted “gold standard” datasets has become the new focus
    - Move away from just using “found” data and instead deliberately construct them from digital sources
    - Inserting human coders in the process for coding and validation boosts the quality of data science approaches
    - Increases interpretability of results, a key shortcoming of AI/ML

Up Next

- Exercise (this afternoon)
  - Scraping
  - Legal Issues
- Next lecture (tomorrow morning)
  - Research on Digital Media