Big Data and the Question of Objectivity

Federica Russo[a] & Jean-Christophe Plantin[b]

[a]University of Amsterdam | @federicarusso
[b]London School of Economics and Political Science | @JCPlantin

Overview

- The social sciences go *big*
  - *Quantitative* social science
- A big question: what conception(s) of objectivity in big-data social science practices?
  - What practices?
  - What objectivity?
- Why is this relevant?
  - Conceptual issues
  - Practical implications

The social sciences go *big*

Big and *quantitative*

The ‘first’ big data revolution in social science:

- Positivism and the birth of quantitative social science
  - Possibility of analysing more data, using the tools of statistics

Going quantitative helped the social science reach the ‘realm of the sciences’

- And yet questions related to the objectivity of the social sciences didn’t settle

Big and *problematic*

Current debates around big data and scholarship

Borgman (2015):

> Don’t conflate “ease of acquisition [of data] for ease of analysis”
>
> Need theoretical as well as methodological framework

A choir of old and new data-philosophers (e.g. Sellars, Floridi, Leonelli …)

> Data are not given
>
> Data, information, knowledge are not all the same
>
> Data are relational

Lots of questions already asked

How big / fast is ‘big’?

How much theory in big automated algorithms?

What kind of reasoning? Inductive?

What implications does the ‘big’ have at social / technical / scholarly level?

...

Our investigation

The question:

What exactly do data curators want to achieve with big-data practices?

A two-step answer:

1. Analysis of big-data practices in social science
2. Problematisation of 2 aspects of big-data practices:
    a. Making the data curator visible / invisible [(in)visibility]
    b. Standardisation of processes for data curation [standardisation]

In a nutshell: [a-b] force us to re-think the notion of *objectivity*

Big-data practices in social science

The manual processing ‘pipeline’

Stage	1. Deposit the dataset	2. Dispatch	3. Review and Process		4. Contact the PI (optional)	5. Metadata and Formatting		6. Verify	7. Publish
Action	The PI or acquisition department deposits a study for processing	The manager reviews and dispatches the study to a processor	The processor first reviews the data, identifies problems, and draws a processing plan	The processor then “fixes” the problems: “wild codes,” missing values, questions, labels, etc.	The processor, after contact with the manager, contacts the PI	The processo r writes the metadata for the study	The processor formats the datasets and the documents according to templates	The processor sends all the files to a manager and another processor for “Quality check”	Once reviewed, the manager approves the publication of the study on the website

“Taylorism” in the data archive

“We're more of an assembly line, and so it's production type of work” Paul, Archive Manager

Employment conditions that characterize “invisible technicians” in science (Shapin, 1989; Barley, Bechky, 1994; Star, Strauss, 1999)

- –Strict division of roles
- –Rhythm of work
- –No skills development
- –Short term employment and turn over
- –Highly standardized work, routine
- –Invisible contribution

Making data ‘pristine’

“We want [*the datasets*] to be right, and everything to read properly […] Trying to get that, so that the future users when they get [*the datasets*], they get everything in a pristine manner.” Paul, Archive Manager

Data processing and invisible labor

- **Complete invisibility outside the archive**
  - No critique allowed of the datasets: “Don’t get carried away”
  - Contacting the PI only as last resort
  - Strict formatting for standardized output

- **Complete visibility inside the archive**
  - Making all processing techniques explicit
  - Processing history file + Quality check
  - Homogenization of practices

Interrogating ‘pristineness’

- Cleaning data **twice**: traces of original context + traces of cleaning
- Reproduces erroneous conception of ‘raw data’ (Gitelman, 2013)
- Conceals contributions of data processors: protocol work (Downey, 2014) data packaging (Leonelli, 2016)

The question of objectivity

[(in)visibility] and [standardisation]:

Re-introduce old ideas about objectivity

Exemplify some more recent ideas about objectivity

But also: pull them in opposite directions

The data curator must be *invisible* from the *outside*

Data users don’t know / need to know about the process

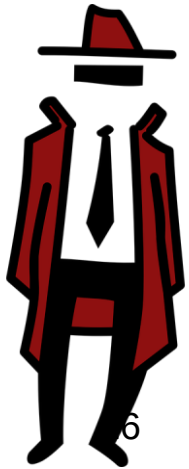
Focus on ‘end product’ (rather than process)

➔ Data are objectively clean, ready to (re)use

- No (interfering) curator behind data curation
- Objectivity is a property of data, not of the process

➔ An old ideal of objectivity: *objects’ objectivity*

  - Kitcher: the *Legend View of Neopositivism*



The data curator must be *visible* from the *inside*

At any time in the data curation process, who the data curator is and what s/he does must be visible, traceable, transparent

➔ The *process* is objective as long as procedures are respected

The curator is present at all times

Objectivity lies in the *procedure*

➔ A more recent idea of *procedural objectivity*

- Montuschi, Little, Cardano, … :
  - the social sciences can attain objectivity;
  - objectivity is in the process, not in the object of science

Procedural objectivity: pulling in opposite directions?

A good tool to have in the kit

- Liberate social sciences from inferiority complex
- Can value role of data curators
- Helps understand where the process can go wrong
- Increases objectivity of the ‘end product’ by self-reflectively work on process

...

A ‘procedural drift’ towards obsessive standardisation?

- Can / should we be flexible about procedure
- If so, do we lose or gain on objectivity?
- Is objectivity *just* a matter of procedure?
- What role is left to the data curator then?

...

What else does ‘obsessive procedural objectivity’ presuppose?

Strong procedures and data pristineness

Much of [(in)visibility] and of [standardisation] rest on

the myth of raw data and of clean data

Pristineness: data are cleaned twice (original PI and of traces of cleaning)

Here we sing with the choir of data-philosophers

No, data is not raw or clean

No, you can’t just assume their cleanness abstracting from curation procedures

No, maybe they shouldn’t be cleaned up so much after all

Yes, perhaps the social science need somewhat dirty data

To sum up and conclude

Social science practices go big

- The social sciences grew big already since Positivism
  - Introduction and development of quantitative methods
  - Demography and sociology; understanding and acting on social phenomena
- In the era of big data, they grow even bigger
  - More data: social media provide tons
  - More practices: data curation and automated data analyses

Big-data practices strive for objectivity

Two relevant aspects of these practices:

[(in)visbility] and [standardisation]

Two notions of objectivity at play:

[invisible] curators: the *objects* are objective

[visible] curators: the *procedures* are objective

[standardisation] of procedures: *procedures* are objectives … *too* objective?

Relevance of the discussion

- An interesting ‘philosophy of science in practice’ question
  - From the practice, bottom up crucial philosophical issues
  - Objectivity, an evergreen of phil sci. But what new is at stake with big data?
- Beyond scholarly questions
  - Open data and open science
    - Can we abstract from the alleged objectivity of these practices?
    - When are data objective enough to be safely re-used?
      - If [standardisation] doesn’t ensure it, what does?
      - Should we strive for *that* kind of objectivity?