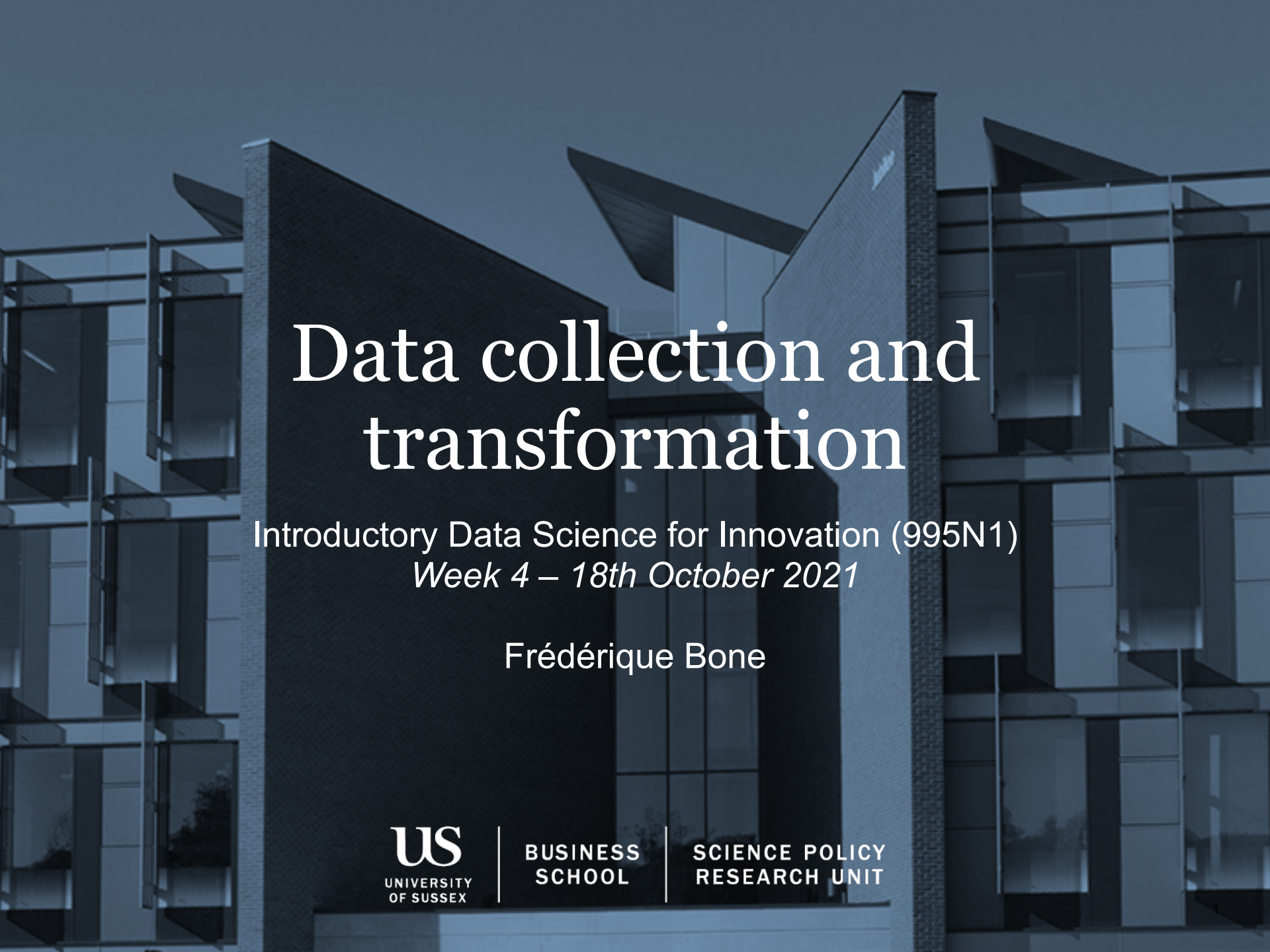


Data collection and transformation

Introductory Data Science for Innovation (995N1)
Week 4 – 18th October 2021

Frédérique Bone

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL | SCIENCE POLICY RESEARCH UNIT

Lecture objectives

Bases of data science:

This lecture will present you with an overview of **data sources** which are relevant for innovation studies. A few of these sources will be explored in depth in the next two lectures.
Secondly the lecture also aims to give a brief introduction of **the format the data** can take. How to treat these will be explored further in the seminars.
Finally, this lecture will give you a broad overview of **good practice when using data** collected from the web.

Lecture objectives

We will see:

1) a variety of data sources which can be used to study science and innovation
2) how hard/easy it is to extract data from these data sources
3) considerations for the use of this data (data bias)
4) ethical considerations for data collection and use

Range of data on innovation

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL | SCIENCE POLICY RESEARCH UNIT

Finding data – existing databases

The usual suspects:

- Existing databases

Finding data – existing databases

- Existing databases

  The usual suspects:

eurostat

Europäisches Patentamt
European Patent Office
Office européen des brevets

WEB OF SCIENCE™

Scopus®

Dimensions

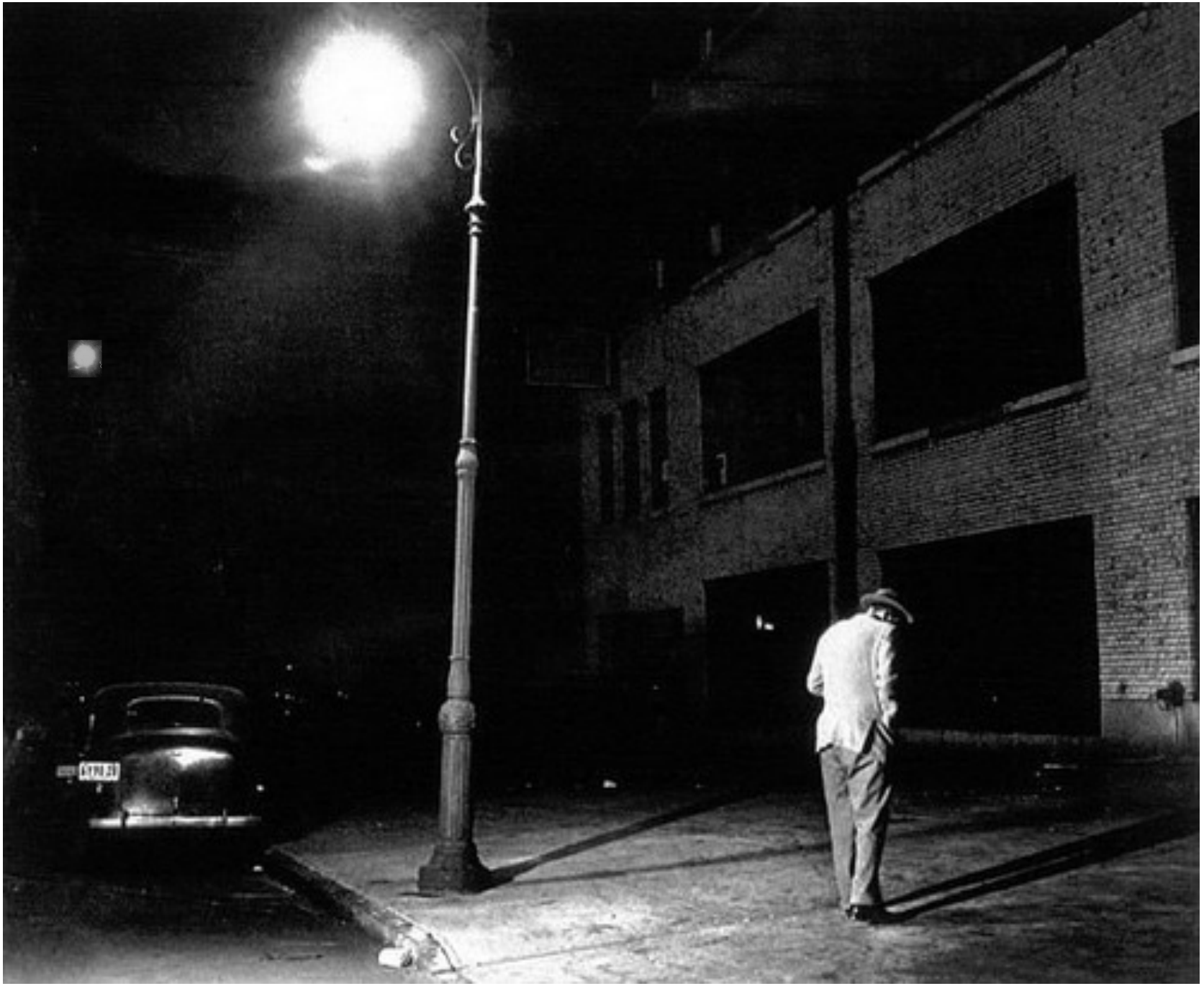
OECD

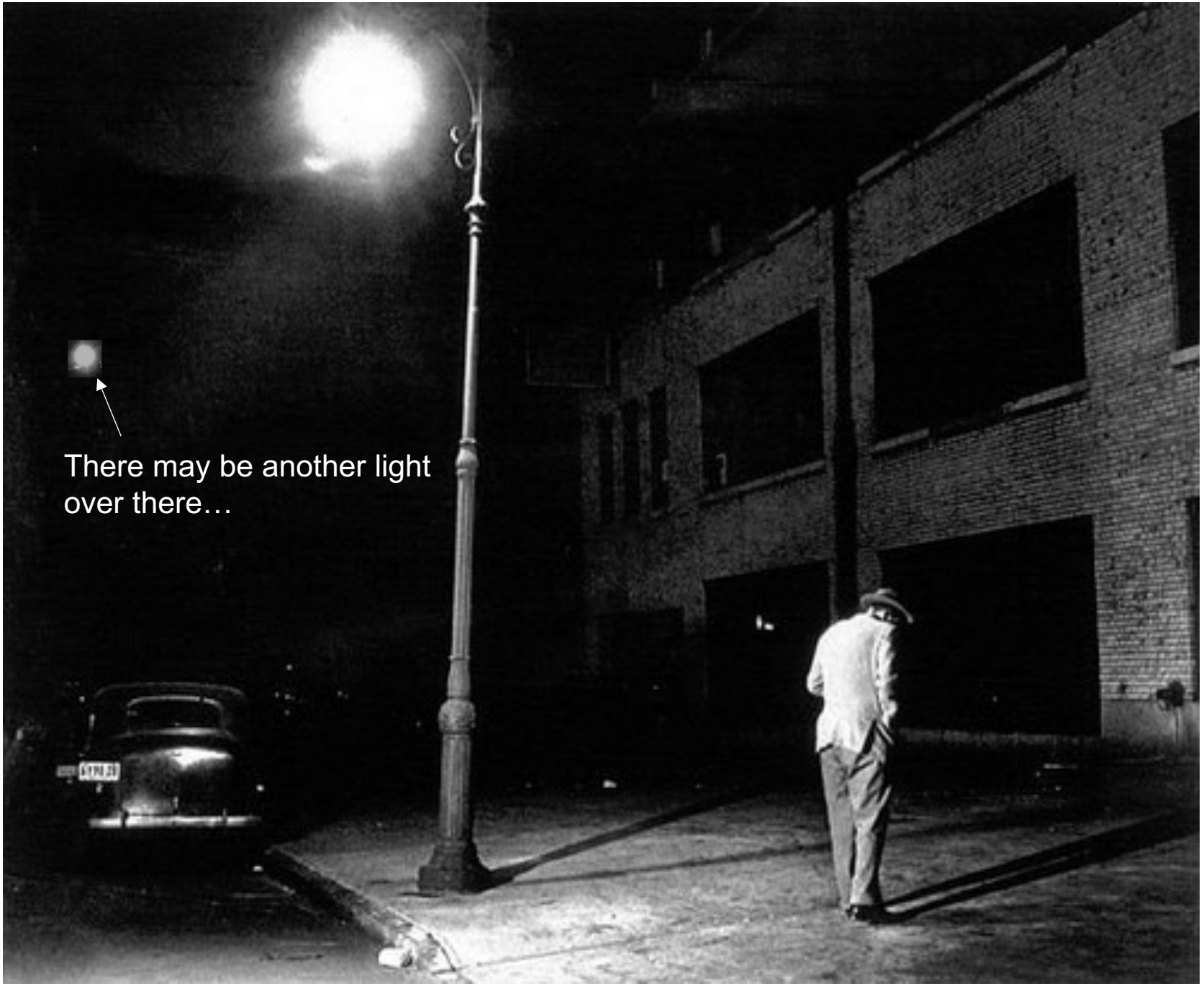
Office for National Statistics

orbis

PubMed.gov
US National Library of Medicine
National Institutes of Health

You may need to pay for access, or an institutional subscription.




There may be another light over there…

Finding data – existing databases

- Existing databases

  But other structured datasets can be used:

ORCID

Eurobarometer Data Service

gesis

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
data.gouv.fr

wellcome

data.gov.uk

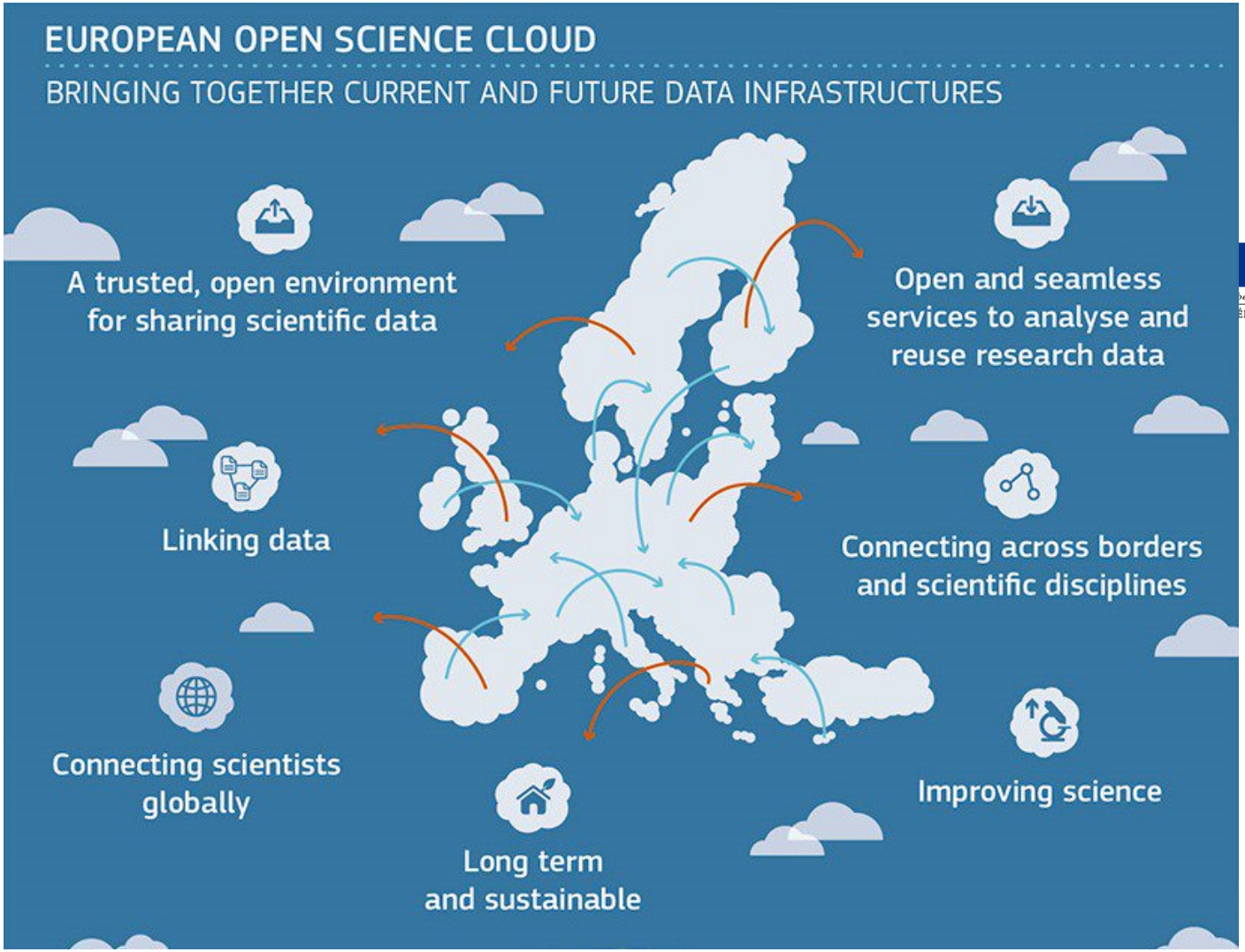
Open data

data.bl.uk

EUROPEAN
DATA PORTAL

EU Open Data Portal
Access to European Union open data

Finding data – existing databases



data.gouv.fr

Open data

EU Open Data Portal
Access to European Union open data


You can explore beyond here…

Finding data – through APIs

- Existing databases
- Data available through APIs

Finding data – through APIs

- Data available through APIs

  API (web application - Application Programme Interface)

  - Ask for a subset of data to a website (in an automated way)
  - The request is done through the URL

    <base-url> + <information about request>

  - The website returns a webpage with some structured text
    - You usually copy the content from the webpage
    - It is easily parsed due to its structure (XML, Json format)

  You could do that directly through a browser, or even using programming tools (without opening a browser).

Finding data – through APIs

- Data available through APIs

Easier to access a subset of the data (some databases are very large)

Access to complementary information

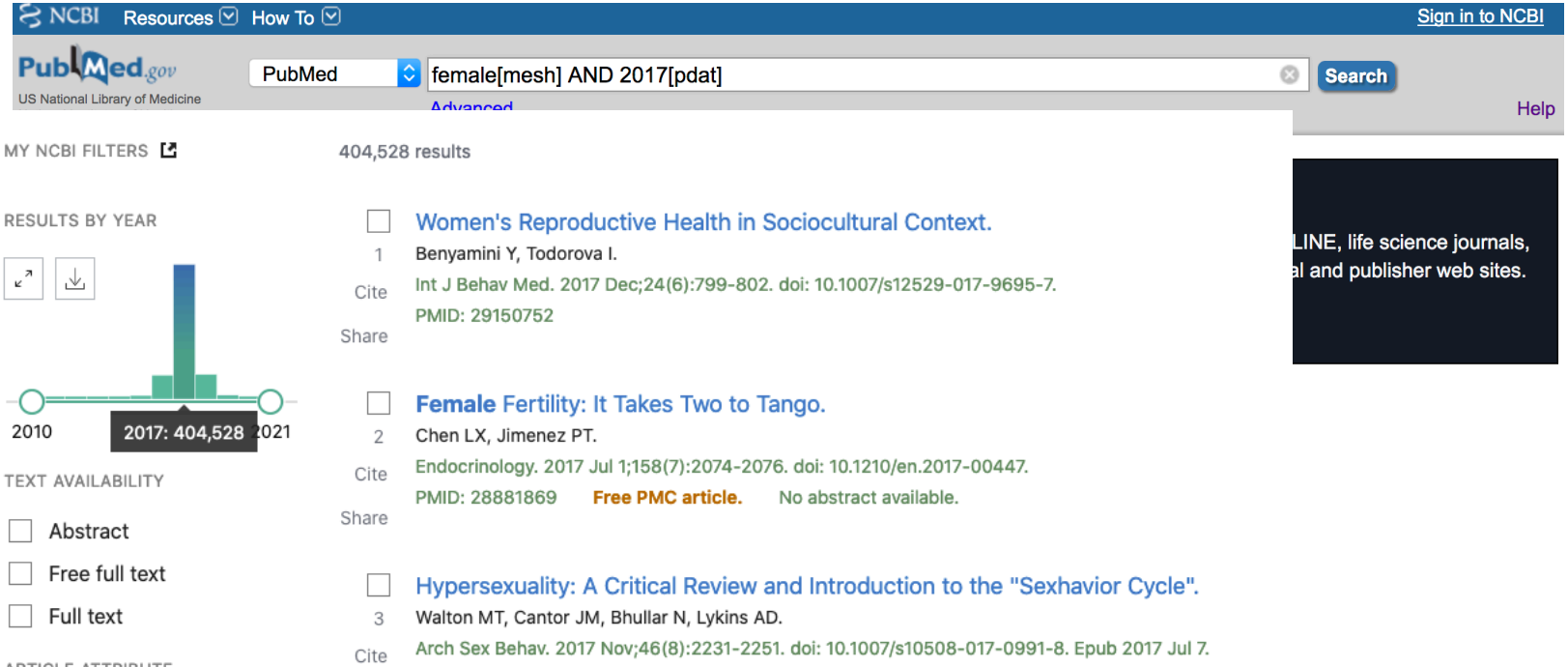
Finding data – through APIs

- Data available through APIs

Example from PubMed:

I want to find out how many publications there are when searching for a specific Meshterm “female” in the year 2017.

PubMed.gov
US National Library of Medicine
National Institutes of Health



Finding data – through APIs

- Data available through APIs

  Example from PubMed:

  I want to find out how many publications there are when searching for a specific Meshterm “female” in the year 2017.

  <base-url>+<information about request>

  URL:

  Base: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?

  Database: db=pubmed

  Search: term=female[mesh]+AND+2017[pdat]

  https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=female[mesh]+AND+2017[pdat]

```
−<eSearchResult>
    <Count>404707</Count>
    <RetMax>20</RetMax>
    <RetStart>0</RetStart>
  −<IdList>
      <Id>32768129</Id>
      <Id>32768128</Id>
      <Id>32768127</Id>
      <Id>32768126</Id>
      <Id>32430189</Id>
      <Id>32192603</Id>
      <Id>31747990</Id>
      <Id>31747988</Id>
      <Id>31679561</Id>
      <Id>31563388</Id>
      <Id>31439180</Id>
      <Id>31439179</Id>
      <Id>31439178</Id>
      <Id>31426919</Id>
      <Id>31426918</Id>
      <Id>31426916</Id>
      <Id>31409142</Id>
      <Id>31394985</Id>
      <Id>31343279</Id>
      <Id>31263379</Id>
    </IdList>
  −<TranslationSet>
    −<Translation>
        <From>female[mesh]</From>
        <To>"female"[MeSH Terms]</To>
      </Translation>
    </TranslationSet>
  −<TranslationStack>
    −<TermSet>
        <Term>"female"[MeSH Terms]</Term>
        <Field>MeSH Terms</Field>
        <Count>8916591</Count>
```

Finding data

- Available databases
- Data available through APIs
- Data you retrieve directly from digital sources

Finding data

- Data you retrieve directly from digital sources

  Any webpage is a structured document,

  and you can extract data from it.

  Let's say I want to study spin-offs from the University of Oxford:

  - Start by looking at the incubator website:

    https://www.sinc.co.uk/
  - Look for companies list
  - Can I export the list?
    - Developer tab (see the web page in a raw format)
    - Use of URL to iterate over the pages (use of loops)

Let’s Practice!

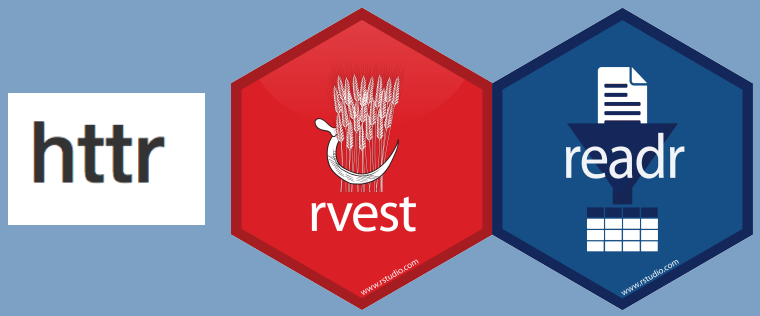
Let's start!

Load the Script of this week:

W3-webdata_exercise.Rmd

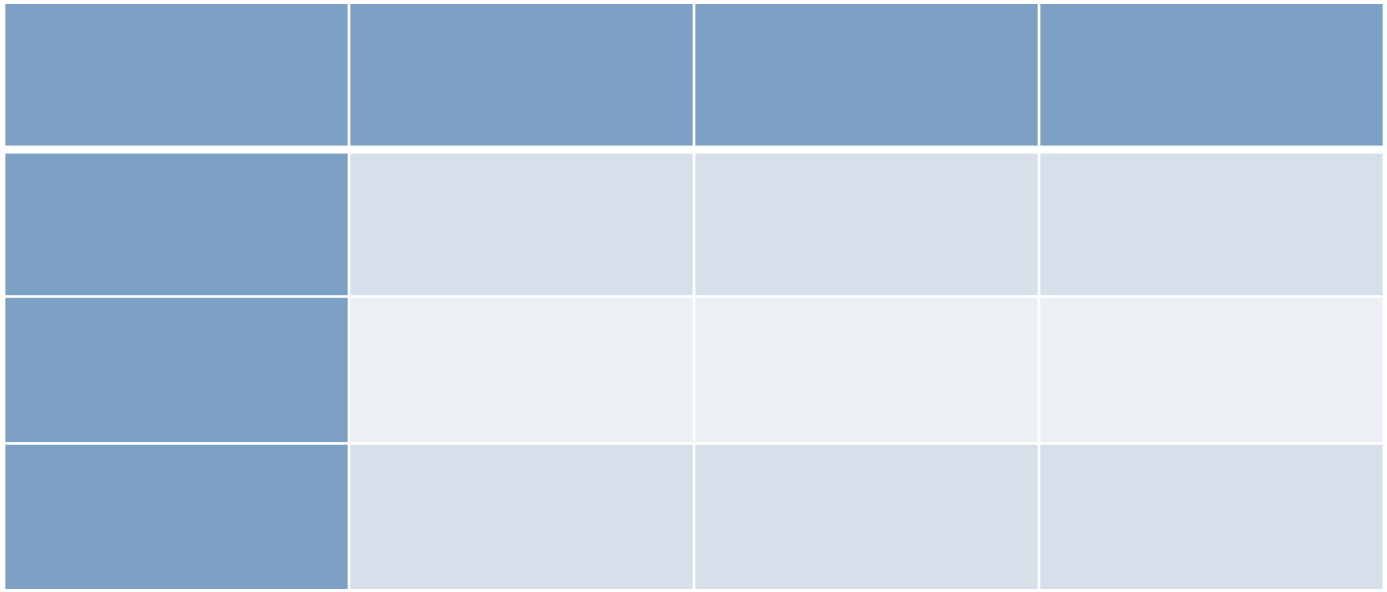
We are going to build a dataset from the previous website to build a incubator start-up dataset using Rvest

In which form will I get the data?



Data Format

1) Tables - two dimensional data:



Data Format

1) Tables - two dimensional data:

- Excel spreadsheet (multiple sheets in a file)
- .csv files which are usually comma delimited but can also have other types of delimiters (tabs generally).
- Proprietary formats .sav (SPSS), .dta (Stata)

Data Format

Example of a csv file:

```
rule_id,rule,keyword,freq.occ
2,Make parts interchangeable,gauges,6456
2,Make parts interchangeable,interchangeability,2023
2,Make parts interchangeable,reproducibility,7
2,Make parts interchangeable,specification,10958
2,Make parts interchangeable,standardization,406
2,Make parts interchangeable,tolerance,1059
```

Data Format

2) Nested /Hierarchical data:

httr

- .json
- XML files which are usually comma delimited

→ **Data gathered through APIs are usually coming in this format.**

NYT API


```
response:
  docs:
    0:
      abstract:        "  The heat wave this month will raise overall electricity consumption in the region above last year's levels, according to utilities officials. Demand for electricity fell last year, after a rise the year before, a year that also had unusually hot summer. Three-fifths of that decline came in residential use, which is chiefly affected by use of air conditioners. Generally, the troubled economy has moderated residential usage, which rose through the 1980's as consumers bought larger homes, many of them with central air conditioning, and acquired personal computers, microwaves, color TVs and other consumer appliances."
      web_url:         "https://www.nytimes.com/1993/07/19/nyregion/pulse-electricity-consumption.html"
      snippet:         ""
      lead_paragraph:  "The heat wave this month will raise overall electricity consumption in the region above last year's levels, according to utilities officials. Demand for electricity fell last year, after a rise the year before, a year that also had unusually hot summer. Three-fifths of that decline came in residential use, which is chiefly affected by use of air conditioners. Generally, the troubled economy has moderated residential usage, which rose through the 1980's as consumers bought larger homes, many of them with central air conditioning, and acquired personal computers, microwaves, color TVs and other consumer appliances."
      print_section:   "B"
      print_page:      "1"
      source:          "The New York Times"
      multimedia:      []
      headline:
        main:          "Electricity Consumption"
        kicker:        "PULSE"
        content_kicker: null
        print_headline: "PULSE; Electricity Consumption"
        name:          null
        seo:           null
        sub:           null
      keywords:
        0:
          name:        "glocations"
          value:       "Long Island (NY)"
          rank:        1
          major:       "N"
        1:
          name:        "glocations"
```


Abstract	Web_url	…	headline			Keywords			
			main	kicker	…	name	value	rank	major

Data Format

2) Nested /Hierarchical data:
   - .json
   - XML files which are usually comma delimited

→ **Data gathered through APIs are usually coming in this format.**

→ **Difficulty in dealing with lists within lists**

→ **You may need to specify how to import the data in a data frame format**

Data Format

3) Free text format (with or without metadata)
   - Web pages
   - Scanned Text
   - Other .txt outputs

→ **You have to explore the underlying structure to find the best way to parse the data.**

Good practice:
Using data for research

How good is the data I get?

Data - quality

- Every dataset has a context and history which defines its content
  - Different purposes
  - Different coverage (time / country)
  - The dataset may have evolved depending on user needs
    - Fields are introduced over time (e.g. acknowledgement /funding fields in bibliometric databases)
    - Some categorisation evolves over time and can create data inconsistencies (see NYT)

Data - quality

- Example from Web of Science data:
  - Bias of representation of languages (Chavarro, 2016)
  - Bias of representation of disciplines (Ciarli & Rafols 2017)
  - Coverage limitation for funding data (Grassano et al. 2017)
- Example from the SPRU History project:
  - I built a database of SPRU staff to understand the scientific output over time.
    - Selection criteria of research staff
  - A colleague who wanted to re-use the data questioned its completeness.

Data - quality

- Existing databases
- Data available through API
- Data you retrieve directly from digital sources

The more bespoke your data, the more you will need to spend time on data cleaning…

No dataset is objective.

No dataset is objective.

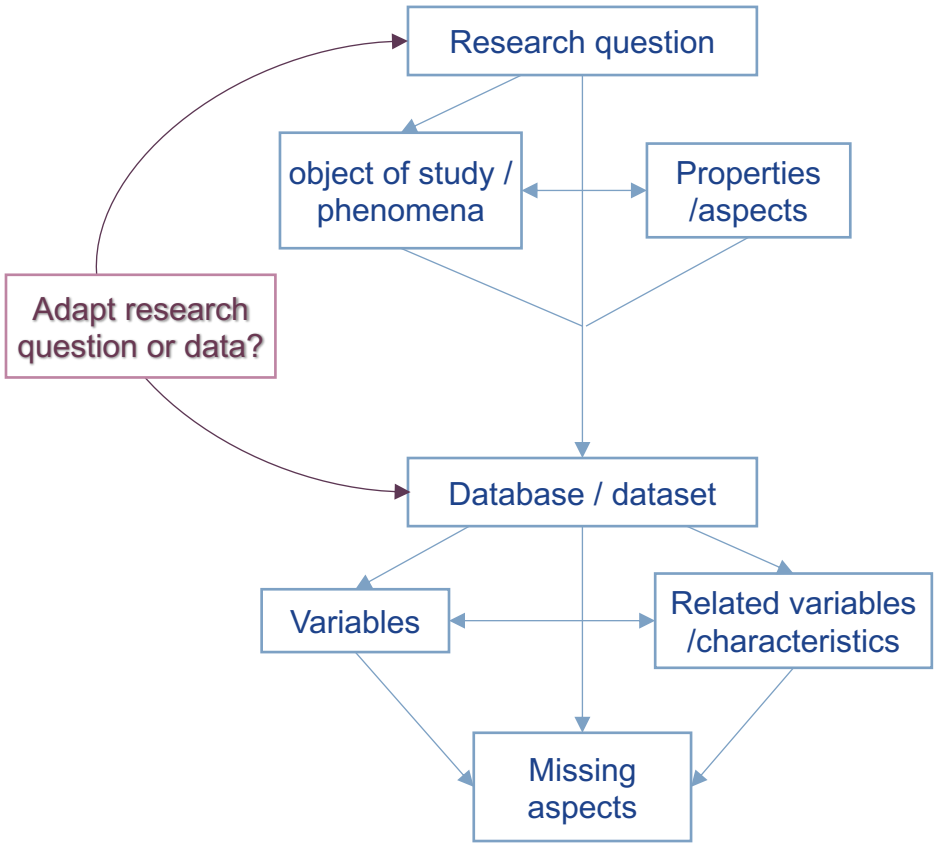
They have a context and history.

Data - quality

- Every dataset has a context and history which define its content
  - Different purposes
  - Different coverage (time / country)
  - The dataset may have evolved depending on user needs

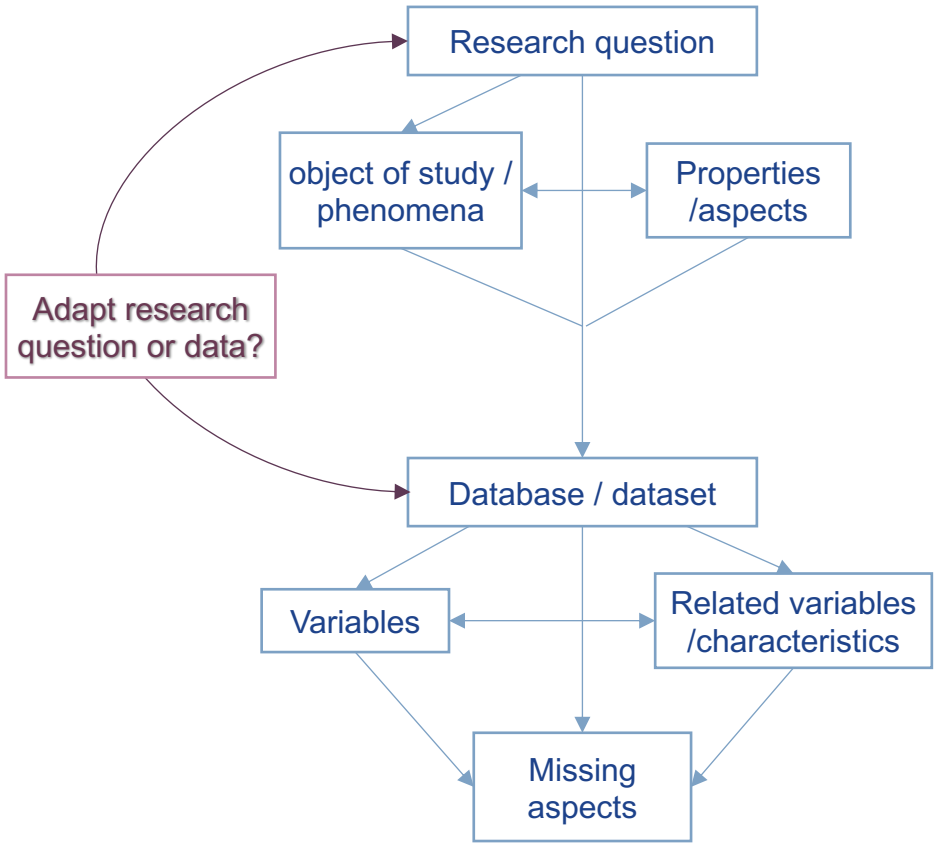
→ How do I choose the right data for my analysis?

Get the right data



- Start with your research question:
  - Object of study / phenomena
  - What aspects / properties are important?
- Is there an existing dataset that captures these:
  - Is it a good representation of the phenomena
  - Does it cover well the aspects which are important.
- How well does the data capture my research design?
  - Is there other data that I can use?
  - Is there complementary data?
  - Should I adapt my research question?

Get the right data



Trade-off:

- Fit of the proxy to the object of study
- Feasibility/time:
  - Searching for appropriate data
  - Cleaning, prepare the data for analysis
  - Conduct the appropriate analysis

Data - quality

- Usual concerns about getting data for your study:
  - Is the database clean enough for moving to the analysis
  - Does it include all the variables to do my 'fancy' analysis
  - Is the database large enough for me to do a large scale/ comparative analysis
- Concerns you should have when choosing data:
  - Is the database complete enough to generate sufficient insights for my analysis?

→ Meaning of the absence of data: non-existence vs. not captured

  - Is the proxy I use good enough to represent what I want to study?

No dataset is bad per se,

they vary in quality and coverage.

Do they fit your research question?

What's next?

Next lectures

The next weeks, we will deal with specific data sources:

- Next week : Funding data (by ResearchFish)
- Following week : Publication and patent data (by myself)
  - With a seminar practicing getting bibliometric data
  - Reading csv files, and use this type of data

Thank you.

Contact:

Dr Frédérique Bone
(f.bone@sussex.ac.uk)

Research Fellow at SPRU

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL | SCIENCE POLICY RESEARCH UNIT

Data - considerations

- https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf
- https://www.communia-association.org/2017/12/01/uk-government-report-right-read-right-mine/
- https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception
- https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf
- https://www.ouvrirlascience.fr/recommandations-sur-lanalyse-automatique-de-documents-acquisition-gestion-exploration/