UK Data Service

Web-scraping for Social Science Research: APIs as a Source of Data

Dr Diarmuid McDonnell
Research Associate

30 April 2020

New Forms of Data Training Series

Upcoming webinars:

- Being a Computational Social Scientist (12 May 2020)

Upcoming coding demonstrations:

- Introduction to Python for social scientists (06 May 2020)
- Collecting data I: web-scraping (13 May 2020)

Past webinars:

- Web-scraping for Social Science Research: Websites as a Source of Data
- Web-scraping for Social Science Research: A Case Study

UK Data Service

Table of Contents

What is an API?

An Application Programming Interface (API) is:

…a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service (Oxford English Dictionary).

In essence: an API acts as an **intermediary** between software applications.

What is an API?

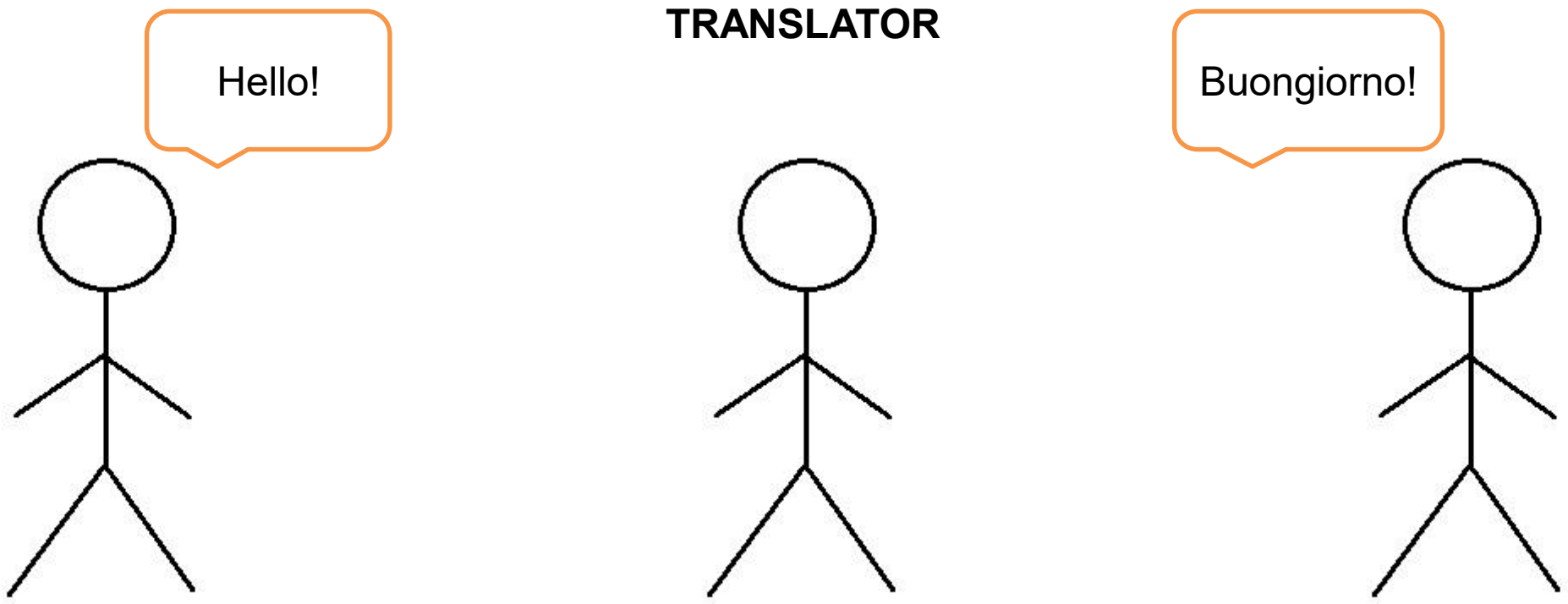

Image Source: https://uxfactor.wordpress.com/2012/12/20/usability/stick-figure-2/

UK Data Service

What is an API?

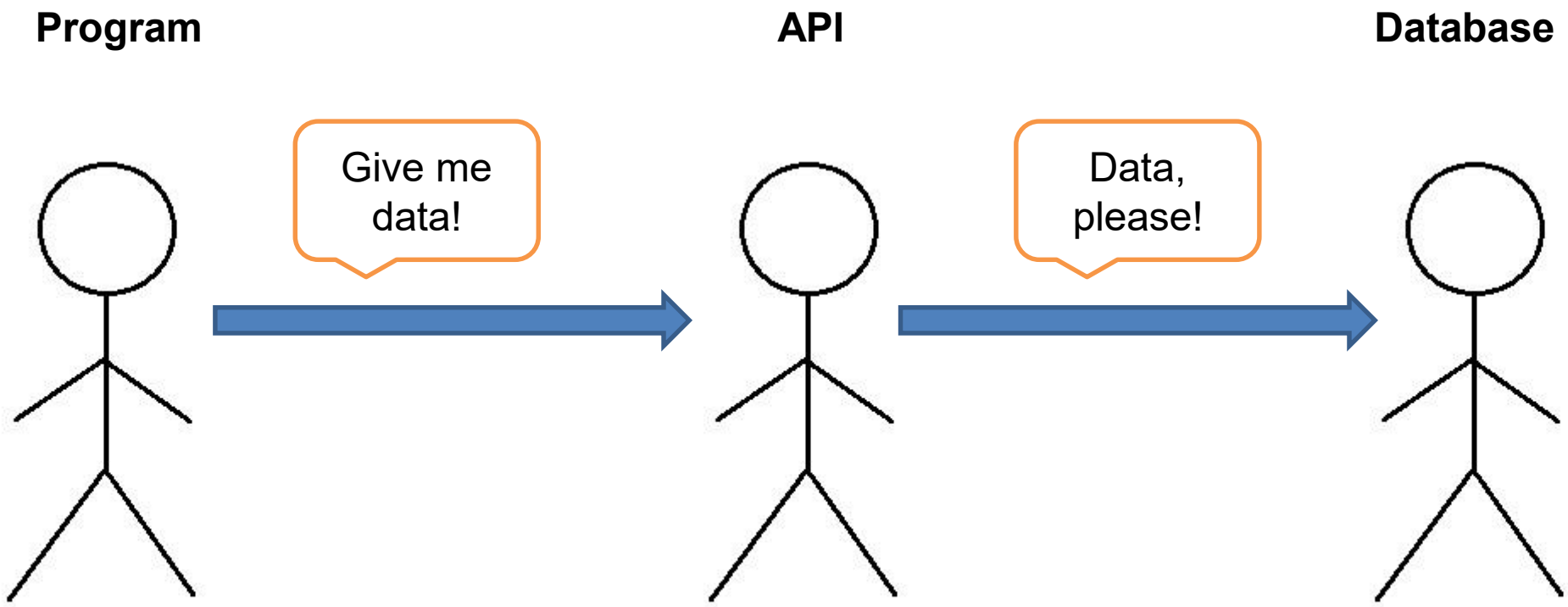

Image Source: https://uxfactor.wordpress.com/2012/12/20/usability/stick-figure-2/

UK Data Service

What is an API?

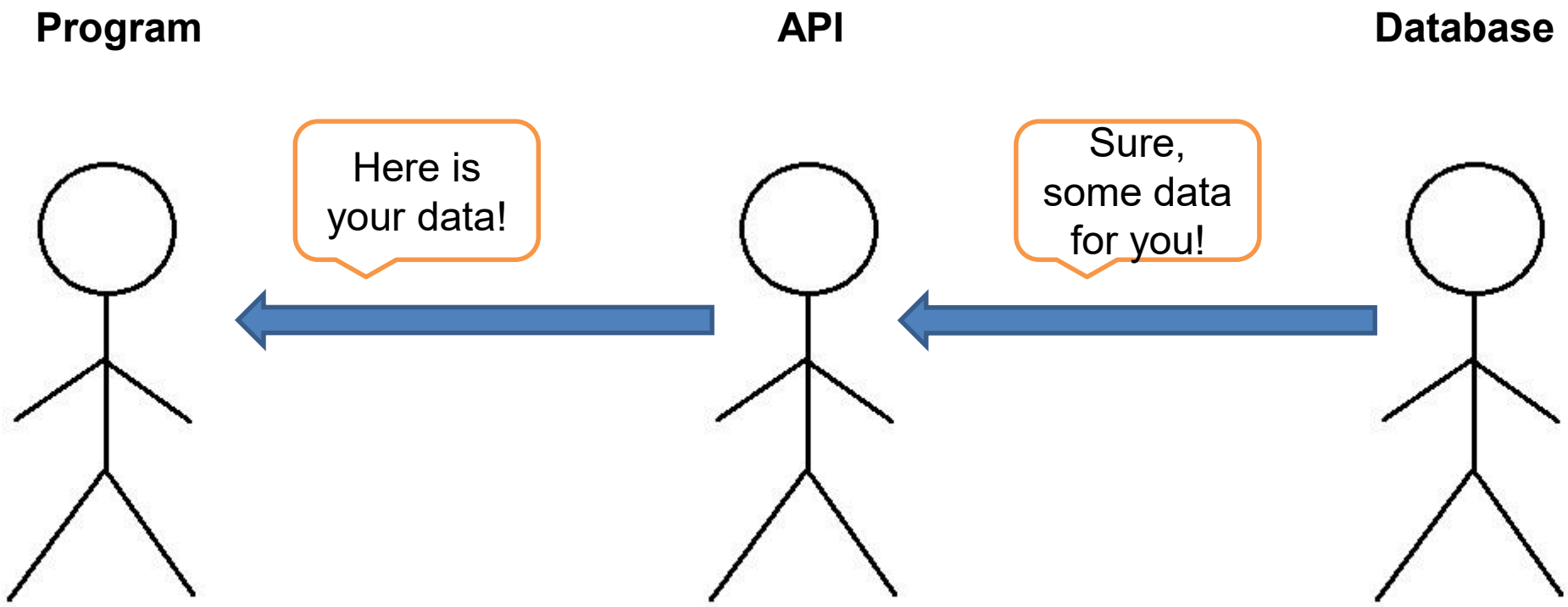

Image Source: https://uxfactor.wordpress.com/2012/12/20/usability/stick-figure-2/

UK Data Service

Why collect data from APIs?

APIs can be an important source of publicly available information on social phenomena of interest.

APIs allow *customised* access to data resources.

Once collected, data can be reshaped into a familiar format (tabular) and linked to other sources of social science data.

Logic of using an API

We need to **know** the following:

1. The location of the API (i.e., web address) through which the database can be accessed. For example, the UK Police API can be accessed via https://data.police.uk/api.
2. The terms of use associated with the API. For example, the UK Police API does not require you to provide authentication but restricts the number of requests for data you can make (15 per second) - the number of allowable requests is known as the *rate limit*.
3. The location of the data of interest on the API. For example, data on street-level crime from the UK Police API is available at: https://data.police.uk/api/crimes-street. The location of the data is known as its *endpoint*.

Logic of using an API

Then, we need to **do** the following:

4. Register your use of the API (if required).
5. Request data from the endpoint of interest, supplying authentication if required. This process is known as *making a call* to the API.
6. Write this data to a file for future use.

What is the value of APIs for social science research?

The process of interacting with an API is a common and mature computational method, with lots of established packages (e.g., `requests` in Python), examples and help available.

APIs provide access to data that is intended to be shared.

The richness of some of the information and data stored on APIs is a point worth repeating.

APIs provide flexible, customisable access to data.

The data you need might only be available through an API.

UK Data Service

What are the limitations of APIs for social science research?

APIs restrict the number of requests for data you can make.

The quality of an API's official documentation can vary wildly.

Data protection laws, such as the EU's General Data Protection Regulations (GDPR), impinge on the use of data you collect through APIs.

An API is a product and you must comply with the Terms of Service/Use associated with it.

APIs can be updated on a frequent basis, resulting in changes to the rate limit, authentication requirements, endpoints providing access to the data, cost of using the service etc.

What are the ethical implications of APIs for social science research?

First-and-foremost, the use of an API is a component of your research project, which itself must receive ethical approval from your institution.

Informed consent is a particularly relevant ethical issue (Lomborg & Bechmann 2014).

Let's take Twitter user data as an example:

- Can users reasonably be said to have given consent to participating in research using their data? Are you able to ask for consent?
- Are certain types of information available through the API too personal to analyse?
- Are there identification risks?
- What if you capture a user's personal data through the API, which at a later date the user deletes from their own profile: should you use this information in your research?

Questions

Dr Diarmuid McDonnell

diarmuid.mcdonnell@manchester.ac.uk

Further resources and help

Repository: https://github.com/UKDataServiceOpen/new-forms-of-data

Youtube: https://www.youtube.com/user/UKDATASERVICE

Help: ukdataservice.ac.uk/help/

Subscribe to UK Data Service news at https://www.jiscmail.ac.uk

@UKDataService

UKDataService

UK Data Service