

# Computational Social Science: An Introductory workshop

*Louise Capener*

*Research Associates at Cathie Marsh  
Institute and UK Data Service*

## You might be interested in other UKDS events

- Using QGIS to map 2021/22 census data from the UK Data Service (22/04)
- Introduction to anonymisation techniques for social sciences research data (07/05)
- Getting started with secondary analysis (11/05)
- **Health Studies User Conference (25/06)**
- **Family Finance Surveys User Conference (09/07)**

<https://ukdataservice.ac.uk/training-events/>

# Table of Contents for this workshop

What is up with 'computational social science'?

How do I become a computational social scientist?

8 steps of CSS - with discussion and project development!

Final thoughts, questions, etc.

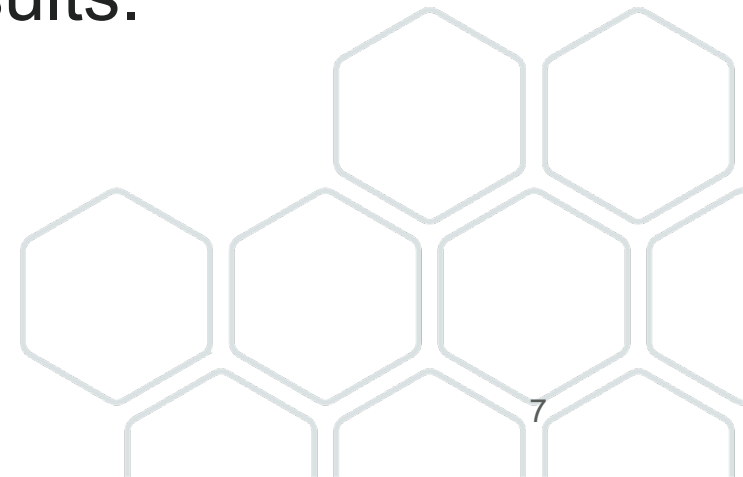


# Computational social science is...

The use of computational and empirical methods to address social science questions.

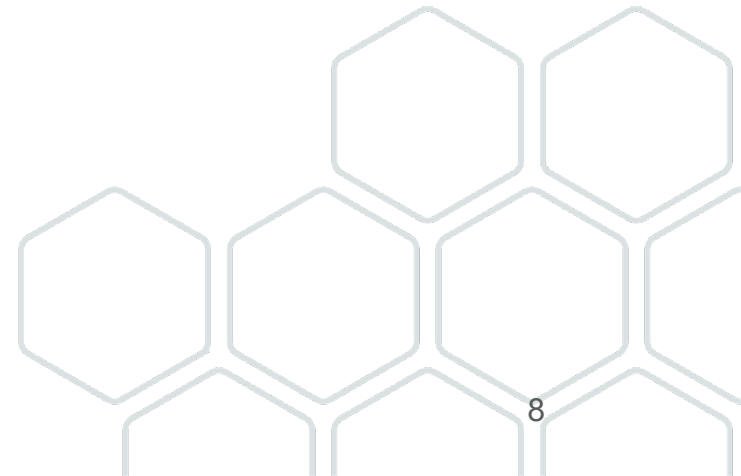
This requires:

- Human-thinking to identify important research questions,
- Computer-thinking to turn questions into computational/empirical methods,
- Human-thinking to effectively communicate the results.



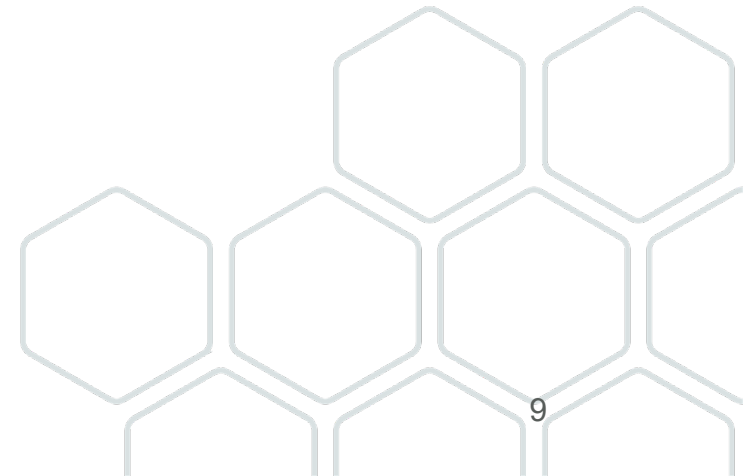
# Computational Social Science is NOT just...

- using computers within a social science research project,
- using digital versions of purely traditional social science methods,  
or
- using digital but purely non-empirical methods.



## Let's have some examples of CSS projects:

- Collect, process, and analyse millions of online news articles to show changing political attitudes
- Use real-time weather and traffic data to show how travellers react
- Combine data from novel wearables/apps to establish correlation between social media activity and heart rate
- Import, process and format centuries of parish records to map family names over time



# Key factors in CSS:

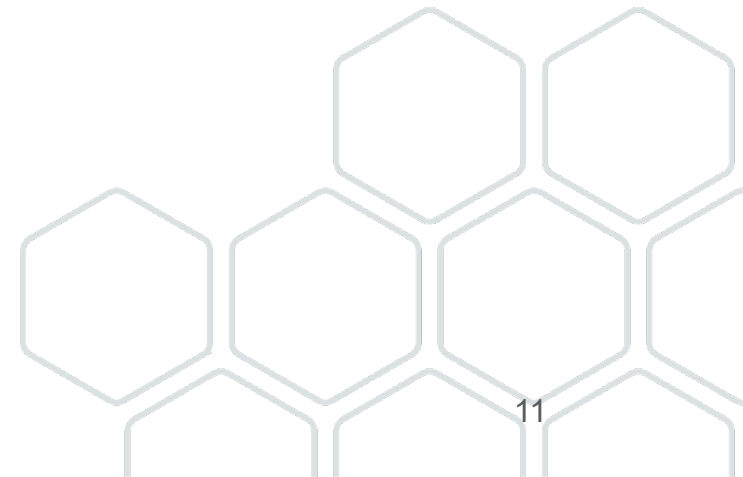
- Data volume, complexity, speed, difficulty or novelty is more important than exact data source/type.
- Data must pertain to people, actions, behaviours, choices, statements, etc.
- Exact research question is not important BUT must be a social science question.



Image from Pexels [pexels-photo-163064.jpg(5472x3648)]

## In essence, CSS is:

“an opportunity to do socially valuable research that would not be possible without computational methods and tools”  
(Halford & Savage 2017)



# Social Scientists ...



Image from Pexels [pexels-photo-6914014.jpg(5771x3847)]

- Social scientists “think like people”
- Study people, interactions, behaviours, etc.
- Thinking skills = abstraction, inference, fuzzy categories, background knowledge, etc.
- Data skills = response categorisation/coding, quality evaluation, pattern detection, etc.
- Use computers, but do not usually write computer code

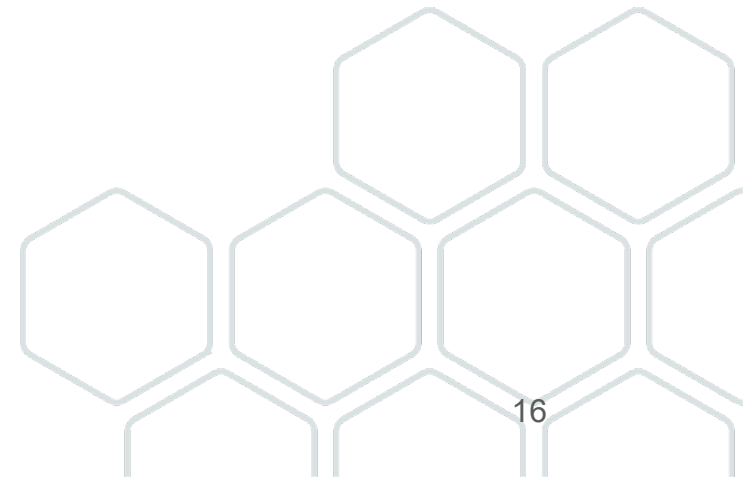
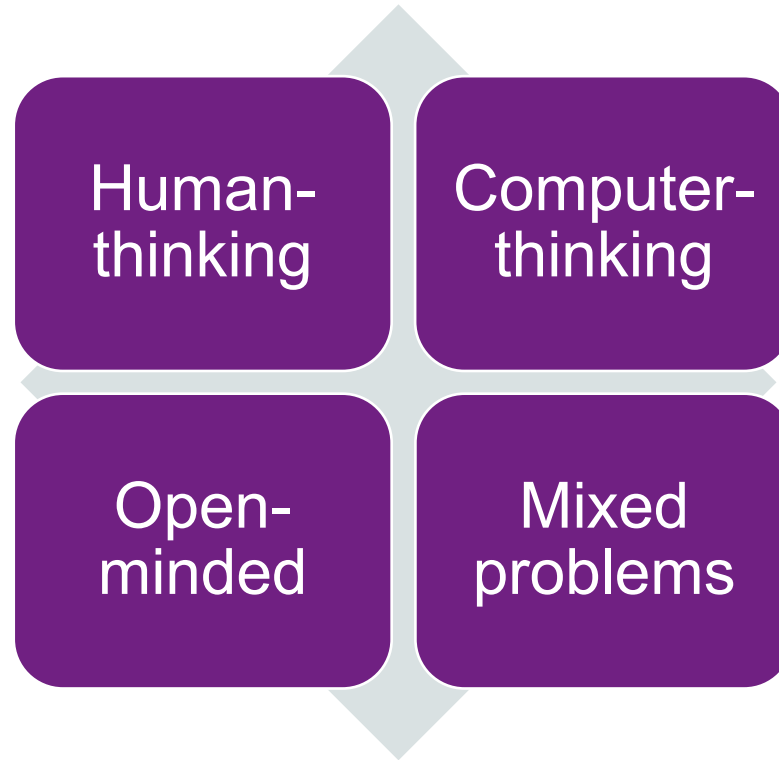
## ... and Computer Scientists

- Computer scientists “think like computers”
- Solve information/processing problems
- Thinking skills = concrete definitions, absolutes, strict hierarchies/categories, clearly defined and scoped variables/rules, etc.
- Data skills = Collect/analyse/manipulate data through code/tech/computational methods
- Not usually taught to identify/motivate research projects with societal impact/value



Image from Pexels [pexels-photo-8879107.jpg(6485x4323)]

# How to do CSS?



# Human-thinking

Skills like:

- Identifying important problems or knowledge gaps,
- Considering possible solutions,
- Connecting problems to relevant theories or perspectives, and
- Collecting relevant information and research to frame approach.

Easy(ish) for social scientists trained in abstraction, communication, subtle context, and shared societal knowledge.

Harder for computer/data scientists not trained in ill-defined, overlapping, context-dependent concepts or using assumptions/background knowledge for interpretation

# Computer-thinking

Skills like:

- accessing, organising, processing and vast and/or complex data,
- writing (collaborative) code, and
- documenting workflows.

Easy(ish) for computer/data scientists trained in computational methods, strict rules, exclusive definitions, and extremely formal and structured processes (Jewett and Kling 1991)

Harder for social scientists, but they can build on training to code responses, format surveys, and draw statistical analyses from complex data (among others)

# Open-minded (and eager to learn)

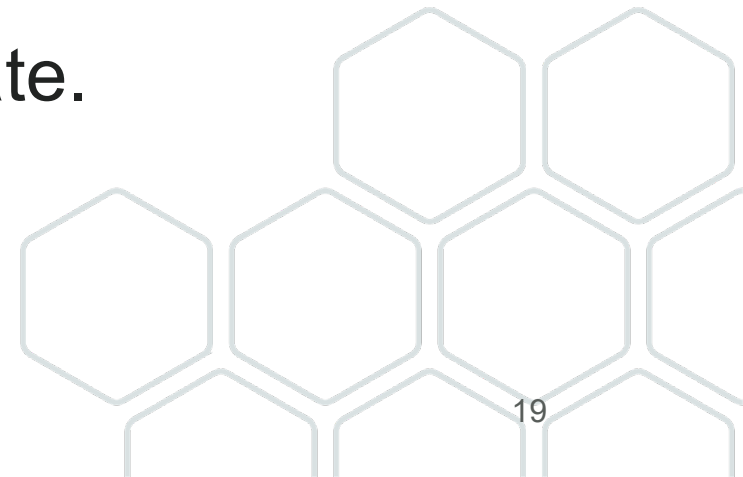
No one starts out with all of the skills they need.

No one knows all the skills they might need to acquire

Approach with an open mind, curiosity, and a willingness to learn.

Some skills will be easier to pick up or use than others.

You can't do it all yourself - be prepared to collaborate.



# Mixed problems

- Mixed problems = need human-thinking AND computer-thinking



Image from Pexels [pexels-photo-7577763.jpg(3840x2160)]



# Mixed problems (and the need to pick 'em)

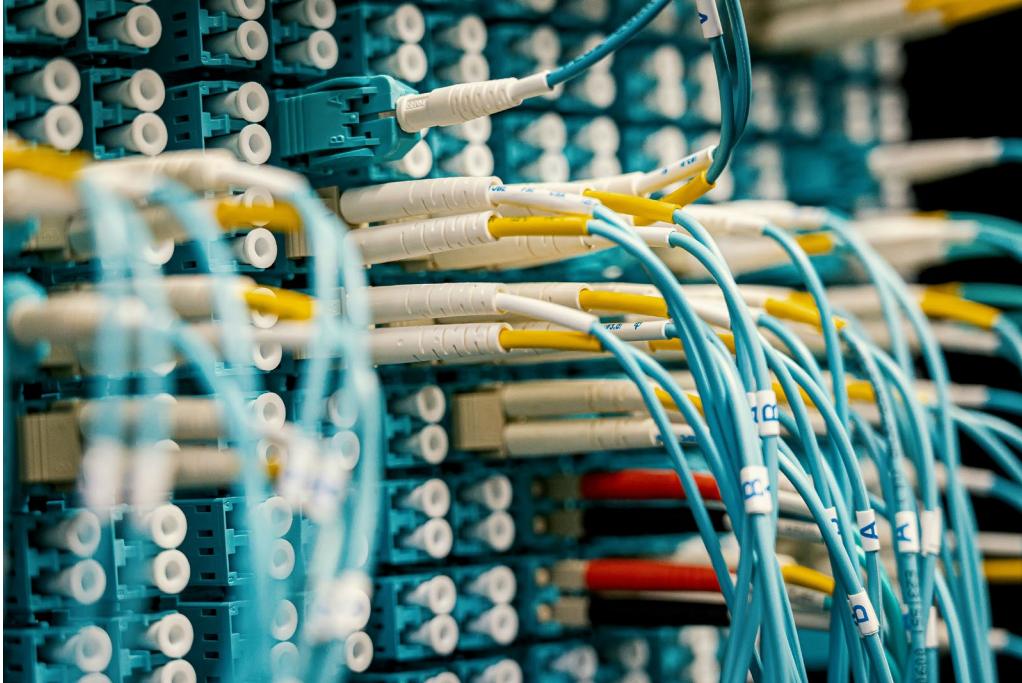
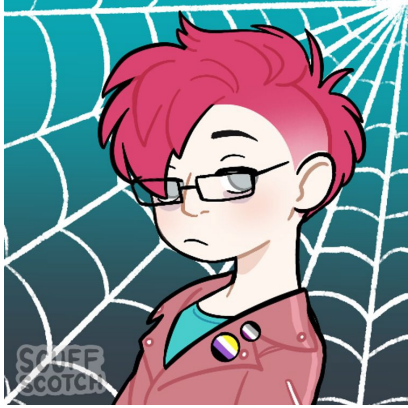


Image from Pexels [pexels-photo-4280702.jpg(5568x3712)]

- Mixed problems = need human-thinking AND computer-thinking
- Will become more important as:
- Resources are digitised
- Interactions, objects and processes become 'smart' or networked
- Large volumes of data are made available/are updated faster
- Other changes in the future

# Pathways into Computational Social Science

Dr. J. Kasmire



Linguistics (BA)

Evolution of Language and Cognition (MSc)

Transition Management in Complex Adaptive Systems (PhD)

Skills gained:

- text-mining and statistical analysis of human speech and writing.
- Advanced stats and agent-based modelling, data analysis.

Nadia Kennar

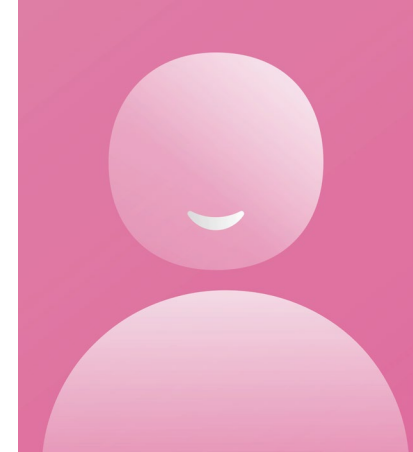


Criminology (BSc)

Criminology and Social Statistics (MRes)

Skills gained:

- Excel/STATA/SPSS and R
- EDA and regression analysis using census and crime data
- Advanced stats and geospatial stats
- Mapping



Politics (BA)

Data Science and Artificial Intelligence (MSc)

Skills gained:

- STATA/SPSS and Python
- Data mining techniques
- Key topics in AI, including ML and NLP

# What about coding?

## Python via Jupyter Notebook

The screenshot shows the RStudio environment. The source editor contains R code for loading and manipulating data. The console shows the output of the code, including package attachment messages and the creation of a data table.

```
44 # ...
45 # [r]
46 # Allows us to read-in csv files
47 library(readr)
48 # For data manipulation
49 library(dplyr)
50 # For regular-expression operations
51 library(stringr)
52 # library(shiny)
53 library(ggplot2)
54 # Used to create interactive visualisations
55 library(plotly)
56 # ...
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

```
filter, lag
```

The following objects are masked from 'package:base':

```
intersect, setdiff, setequal, union
```

Registered S3 method overwritten by 'data.table':

```
method from
print.data.table
```

R version 4.4.0 (2024-04-24) -- "Puppy Cup"  
Copyright (C) 2024 The R Foundation for Statistical Computing  
Platform: aarch64-apple-darwin20

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

```
>
```

## R via RStudio

The screenshot shows a Jupyter Notebook interface. The notebook contains Python code for loading a dataset and displaying its structure. The output shows a data table with columns for England and Wales Code, England and Wales, Gender identity (8 categories) Code, Gender identity (8 categories), and Observation.

```
4.5 Update_traces()
4.6 Update_layout()
5 Dataset 2
5.1 Data cleaning
5.2 Question
5.3 Data pre-proc
6 Interactive grouped
6.1 Interactive lege
6.2 Styling/adjust c
6.3 Further styling
6.4 Stacked bar ch
6.5 Or...
7 Dataset 3
7.1 Data Cleaning
7.2 Question
7.3 Data pre-proc
7.4 Calculate % of
7.5 Calculate Ethni
8 Merge both dataset
9 Interactive scatterp
9.1 Dropdown sele
10 Dataset 4
10.1 Data cleaning
10.2 Data pre-proc
10.3 Interactive line
10.4 Interactive leg
11 Sharing your inter
```

### 4 Dataset 1

The first dataset that we'll be focusing on is a really simple dataset which shows the total counts for 8 gender identity categories across England and Wales. We'll do a bit of data cleaning, remove unnecessary categories (such as 'Does not apply'), and then calculate the % of each gender identity category. Then, we'll create a simple interactive bar chart which displays the percentage by gender identity category, whilst enabling some interactivity when we hover over each bar.

```
In [157]: # Load in dataset
df = pd.read_csv('../Data/GI_det_EW.csv')
```

```
In [158]: # Brief glimpse of data structure
df.head(10)
```

```
Out[158]:
```

	England and Wales Code	England and Wales	Gender identity (8 categories) Code	Gender identity (8 categories)	Observation
0	K04000001	England and Wales	-8	Does not apply	0
1	K04000001	England and Wales	1	Gender identity the same as sex registered ...	45389635
2	K04000001	England and Wales	2	Gender identity different from sex registered ...	117775
3	K04000001	England and Wales	3	Trans woman	47572
4	K04000001	England and Wales	4	Trans man	48435
5	K04000001	England and Wales	5	Non-binary	30257
6	K04000001	England and Wales	6	All other gender identities	18074
7	K04000001	England and Wales	7	Not answered	2914625

```
In [159]: # Use .shape to return number of (rows, columns)
df.shape
```

```
Out[159]: (8, 5)
```

### 4.1 Data cleaning

- Clean column names
- Filter out unnecessary categories
- Clean gender identity category values - too wordy

### Code explanation: List comprehension

Okay, so we have our 'df.columns' variable, which is a list of our df's column names. We're using a 'list comprehension' (which is the for loop within the square brackets) to iterate through each column name and substitute any instance of text within parentheses (inc. the parentheses themselves) with an empty string. The result, i.e., the list of cleaned column names is then assigned back to 'df.columns'.

```
In [160]: # Finds all substrings which match the regular expression and replaces them with empty string
# string - just another word for text
```

# Practically, you can follow my 8-step process!

- Identify the problem
- Explore the problem
- Formalise the concepts
- Collect data, implement software, verify
- Experiment and analyse data
- Discussions and conclusions
- Communicate, publish, present
- Share, document and validate



Image from Pexels [pexels-photo-5383747jpg(3936x2624)]

# 1. Identify the problem

- Be as clear and specific as possible about the pattern, problem, lack of insight.
- Also identify who is involved, where it is, etc.



Image from Pexels [pexels-photo-28640155jpg(5101x3826)]

## 2. Explore the problem

- Gather information and perspectives in multiple ways (surveys, observations, secondary data analysis, app creation, web-scraping, API's, expert interviews, etc.).
- Spell out sub-problems, processes, relationships, simplicifations, assumptions, related issues, existing specialties, etc.



Image from Pexels [pexels-photo-36726329jpg(6000x4000)]

### 3. Formalise the concepts

- Make all the concepts and processes explicit, formal and both computer and human understandable.
- Often known as ‘pseudo-code’.
- Example:
  - “trust” is defined as a variable between 0 and 100.
  - “trust” between two parties increases following mutually beneficial interactions.
  - Existing levels of “trust” decrease to zero if an interaction is judged to be deceitful.
  - Etc.



Image from Pexels [pexels-photo-7964028jpg(6000x4000)]

## 4. Collect data, implement software, verify

- Select and implement one or more methods.
- The choice of method will be highly dependent on the research topic.
- Thoroughly check that the selected method has been implemented correctly – essentially answering the question “Did we do the thing right?”



Image from Pexels [pexels-photo-4659005jpg(6000x4000)]

## 5. Experiment and analyse data

- Run the experiments! Build the models! Analyse the data! Or otherwise use the methods selected in previous step!
- Identify and explain the results within the context of the experiments/model/method.



Image from Pexels [pexels-photo-8326459jpg(4709x3532)]

## 6. Discussions and conclusions

- Going beyond the experiment/model/method, draw some conclusions about what the results mean.
- Do you support policy recommendations?
- Who or what do these results affect? Why does it matter?
- What should change? Who benefits from that proposed change?



Image from Pexels [pexels-photo-7989092jpg(6016x4016)]

## 7. Communicate, publish, present

- All of the previous steps must be communicated to multiple audiences in multiple ways.
- Short term and long term engagement.
- Public, academic, political, students, etc.
- Consider conferences, journals, blogs, white papers, academic societies, workshops or university classes, etc.



Image from Pexels [pexels-photo-17796jpg(4608x3072)]

## 8. Share, document, validate

- Help make sure the ‘right thing was done’ by allowing your work to be studied, reproduced and/or modified as needed through openly available:
  - Workflows (methodologies/steps taken)
  - Code
  - Data
- As transparent, well documented and openly as possible (not always entirely possible)



Image from Pexels [pexels-photo-7346611.jpg(6514x4343)]

# Important to Note



Image from Pexels [pexels-photo-27098151jpg(6016x4000)]

- These steps are **NOT LINEAR!**
- Most (or all) will require many **ITERATIONS.**
- Documentation (step 8) actually applies **THROUGHOUT** all the other steps – don't wait until the end to start!

# References

- Arnold, B., L. Bowler, S. Gibson, P. Herterich, R. Higman, A. Krystalli, A. Morley, M. O'Reilly, K. Whitaker and others (2019). *The Turing Way: a handbook for reproducible data science* (Version v0.0.4).
- Brooker, P. D. (2019). *Programming with Python for Social Scientists*, SAGE Publications Limited.
- Halford, S. and M. Savage (2017). "Speaking Sociologically with Big Data: Symphonic. Social Science and the Future for Big Data Research." *Sociology* **51**(6): 1132--1148.
- Heiberger, R. H. and J. R. Riebling (2016). "Installing computational social science: Facing the challenges of new information and communication technologies in social science." *Methodological Innovations* **9**.
- Jewett, T. and R. Kling (1991). "The dynamics of computerization in a social science research team: A case study of infrastructure, strategies, and skills." *Social Science Computer Review*. **9**(2): 246—275.
- Kasmire, J. and D. McDonnell (2020). *Thinking like a computational social Scientist: Organising thoughts and organising data*. Big Surv '20 Big Data meets Survey Science.



# Thank you.

Louise Capener

Email to [louise.capener@manchester.ac.uk](mailto:louise.capener@manchester.ac.uk)

 @CapenerLouise on Twitter

UKDS

 @UKDataService on Twitter

