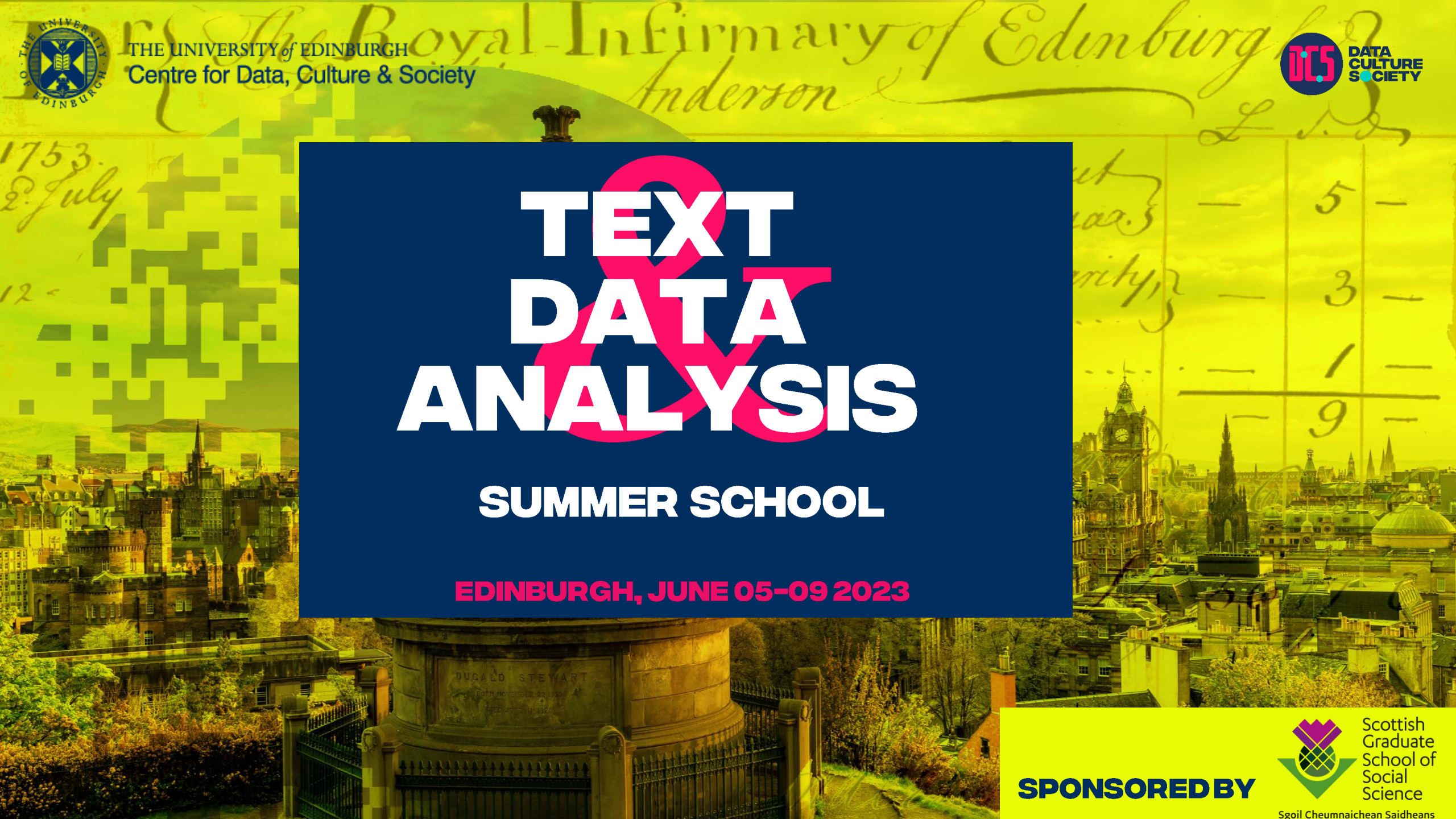
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

TEXT & DATA ANALYSIS

SUMMER SCHOOL

EDINBURGH, JUNE 05–09 2023



SPONSORED BY

Scottish Graduate School of Social Science

Sgoil Cheumnaichean Saidheans

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

HOUSE KEEPING

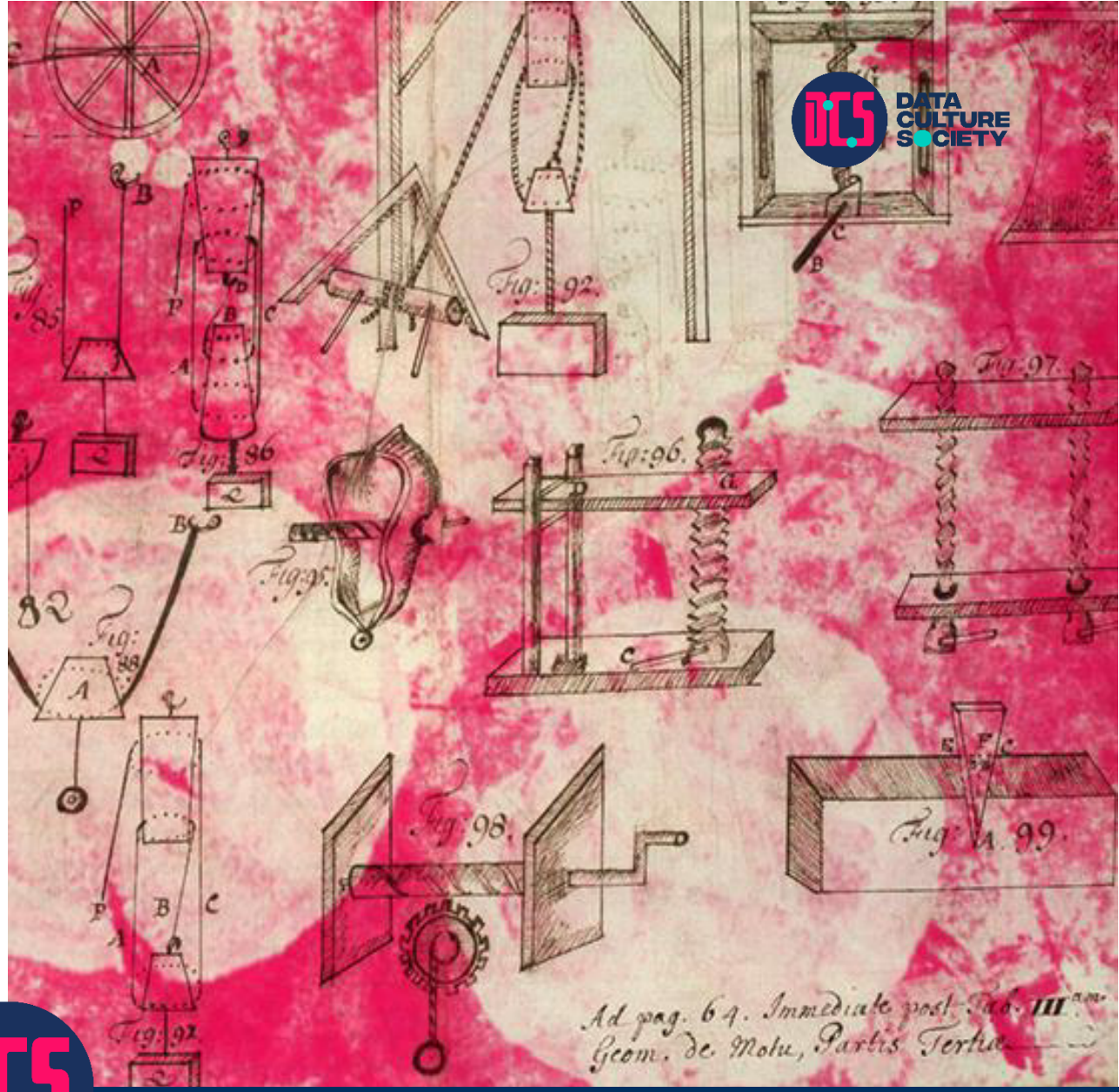
DATA CULTURE SOCIETY

- Toilets
- Food Consumption
- Water Fountains
- Fire Alarm
- Code of Conduct


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TODAY'S SCHEDULE
Seminar: Affective Partisan Sorting in the UK: Turbulent Times
Hands-on session 1: Sentiment Analysis
Hands-on session 2: Data Wrangling
BYOD Session: 3

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

AFFECTIVE PARTISAN SORTING IN THE UK: TURBULENT TIMES



Dr Ugur Ozdemir,

Lecturer in Quantitative Political Science at the University of Edinburgh

AFFECTIVE PARTISAN SORTING IN THE UK: TURBULENT TIMES

Ugur Ozdemir - PIR

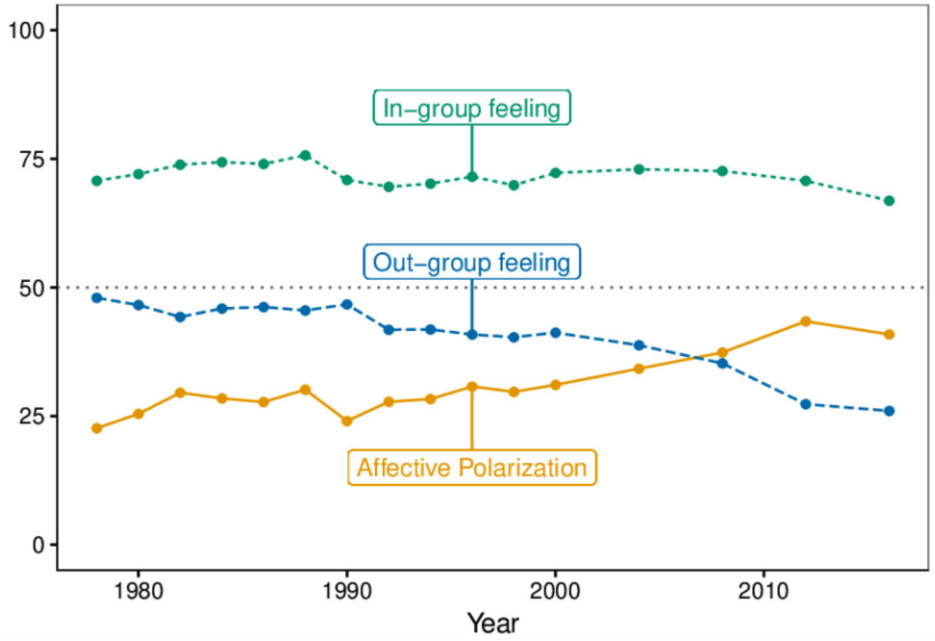
CDCS Summer School

7 June 2023

THE UNIVERSITY *of* EDINBURGH

AFFECTIVE PARTISAN POLARIZATION (APP)

- **Definition:** Strong emotional attachments towards co-partisans and hostility towards opposing partisans.
- Theoretical background: Social Identity Theory.
- More important than issue polarization?
- Mainly in relation to American politics (Iyengar, Sood and Lelkes 2012; Iyengar and Westwood 2015; Mason 2015; Mason 2018)
- Some recent work focus on outside the US:
  - Hobolt, Leeper, and Tilley (2020): UK
  - Reiljan (2020): Europe
- Measured through 'thermometer questions' in the surveys – 'how much do you like-dislike the political parties'.



APP – MEASUREMENT IN MULTIPARTY SYSTEMS

- For two party systems, median distances are usually used.
- Not much until recently.
- Eiljan (2020):

$$\text{API} = \sum_{n=1}^{N} \left[\sum_{\substack{m=1 \\ m \neq n}}^{N} \left((\text{Like}_n - \text{Like}_m) \times \left(\frac{\text{Vote share}_m}{1 - \text{Vote share}_n} \right) \right) \times \text{Vote share}_n \right]$$

  - Not an individual level measure – works with party averages. Hence it does not take the distribution into account.

IDEA OF THIS PAPER

- Affective partisanship as a 'latent variable' - use all the like-dislike information to construct a unidimensional 'affective partisanship space'.
- Method: Item Response Theory
  - IRT: Family of statistical models for measurement of latent traits from observable indicators.
  - Graded response model (Samejima, 1969): The model used when you have ordinal indicators, like the Likert scale we are using.
- Using the characteristics of that space, examine the sorting/polarization dynamics of the UK electorate.

METHOD

- Every voter is a K sized vector where K is the number of parties and each element is how much the voter likes that party.
- If we use an IRT model to reduce the dimensionality of that NxK matrix to a NX1 matrix, what would be the underlying latent variable?
- Can we use this variable to measure polarization?
- What about the «loadings»?
- What about the proportion of the variance explained by the first dimension?

DATA

- British Election Study Internet Panel Waves 1-23
- Time span: February 2014 – May 2022
- Sample sizes are around 30.000.

partyLikeGrid1

partyLikeGrid1 (Like/PTV randomised W1-W6)	GRID	All waves

How much do you like or dislike each of the following parties?

ROWS

likeCon	Conservatives	
likeLab	Labour	
likeLD	Liberal Democrats	
likeSNP- Show if country==2	SNP	All countries in second half of wave 5 and wave 6
likePC- Show if country==3	Plaid Cymru	

COLUMNS

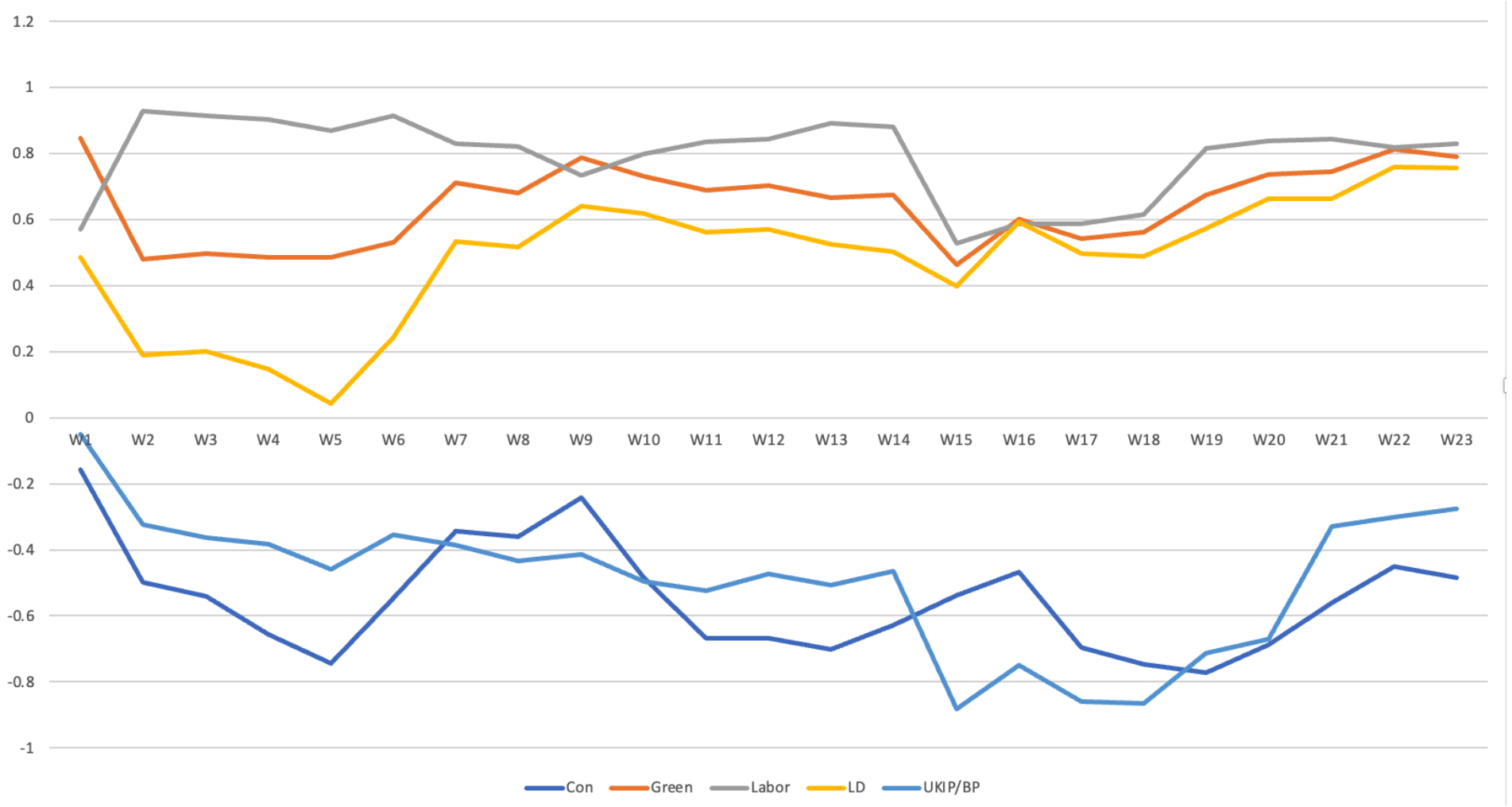
0	0	Strongly dislike
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	Strongly like
9999	Don't know	

CONTEXT

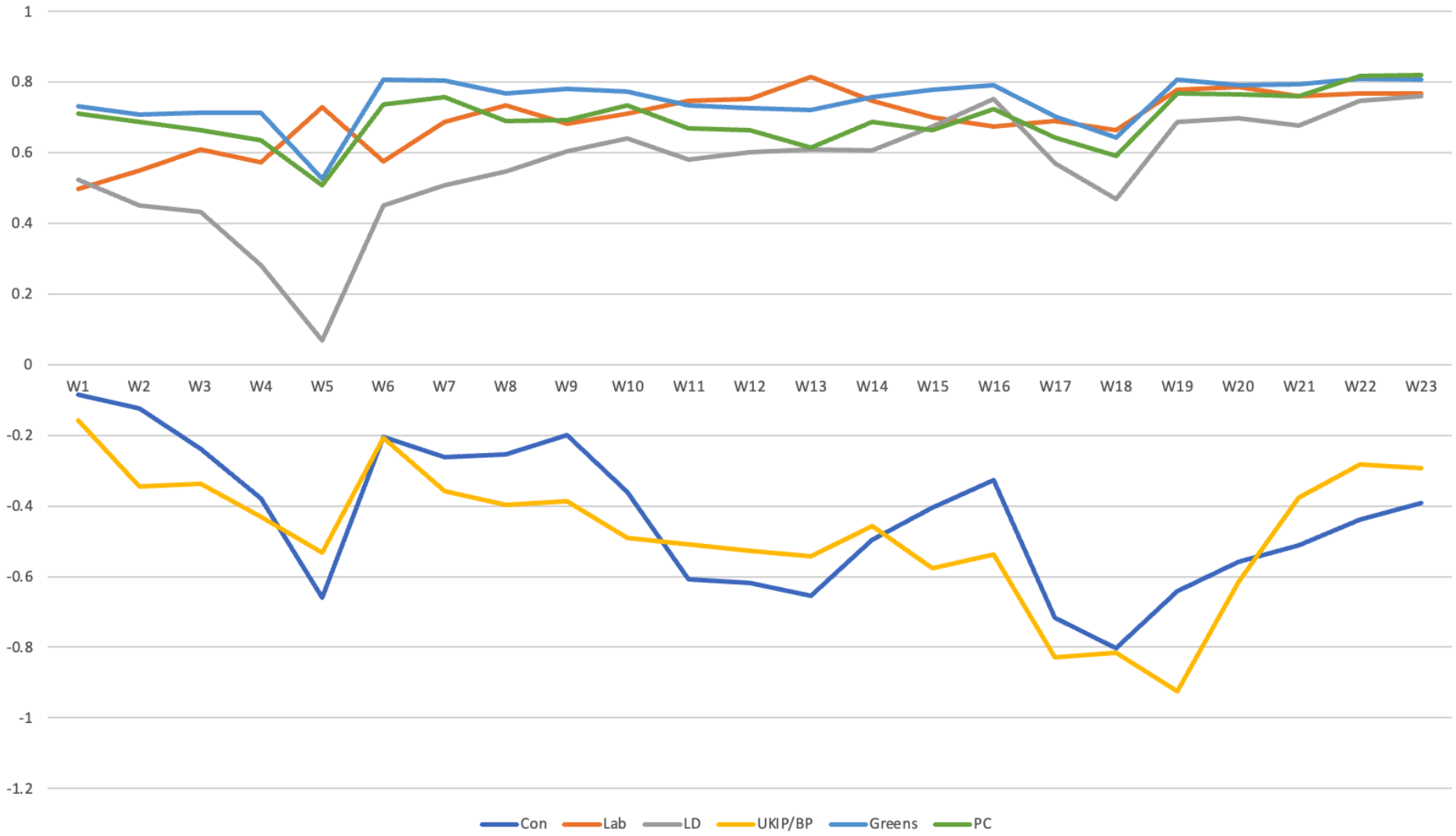
- Panel data covers the 'turbulent times' – 6 months before the independence referendum to aftermath of the 2021 elections.
- Gives us a unique opportunity to test how issue polarization and affective polarization are related.
- Multiple states and 'interstate conflicts' make the case more interesting.

	Event	Begin	End
2	**Wave 1**	Thursday, 20 February 2014	Sunday, 9 March 2014
3	**EP Elections**	Thursday, 22 May 2014	
4	**Wave 2**	Thursday, 22 May 2014	Wednesday, 25 June 2014
5	**Independence**	Thursday, 18 September 2014	
6	**Wave 3**	Friday, 19 September 2014	Friday, 17 October 2014
7	**Wave 4**	Wednesday, 4 March 2015	Monday, 30 March 2015
8	**Wave 5**	Tuesday, 31 March 2015	Wednesday, 6 May 2015
9	**2015 GE**	Thursday, 7 May 2015	
10	**Wave 6**	Friday, 8 May 2015	Tuesday, 26 May 2015
11	**Wave 7**	Thursday, 14 April 2016	Wednesday, 4 May 2016
12	**Wave 8**	06 May 2016	22 June 2016
13	**Brexit**	23 June 2016	
14	**Wave 9**	24 June 2016	04 July 2016
15	**Theresa May**	13 July 2016	
16	**Wave 10**	24 November 2016	12 December 2016
17	**Wave 11**	24 April 2017	03 May 2017
18	**Wave 12**	05 May 2017	07 June 2017
19	**2017 GE**	08 June 2017	
20	**Wave 13**	09 June 2017	23 June 2017
21	**Wave 14**	04 May 2018	21 May 2018
22	**Gov Defeats**	15 Jan - 12 March	
23	**Wave 15**	11 March 2019	29 March 2019
24	**Art. 50 Extensions**		
25	**Wave 16**	24 May 2019	18 June 2019
26	**Wave 17**	01 November 2019	12 November 2019
27	**Wave 18**	13 November 2019	11 December 2019
28	**2019 GE**	12 December 2019	
29	**Wave 19**	13 December 2019	23 December 2019
30	**EU Withdrawal Ratified**		
31	**COVID**		
32	**Starmer Elected**		
33	**Wave 20**	03 June 2020	21 June 2020
34	**Elections - UK Local / Scottish Parliament /Senedd**		06 May 2023
35	**Wave 21**	07 May 2021	25 May 2021
36	**Wave 22**	26 November 2021	15 December 2021
37	**Wave 23**	06 May 2022	26 May 2022

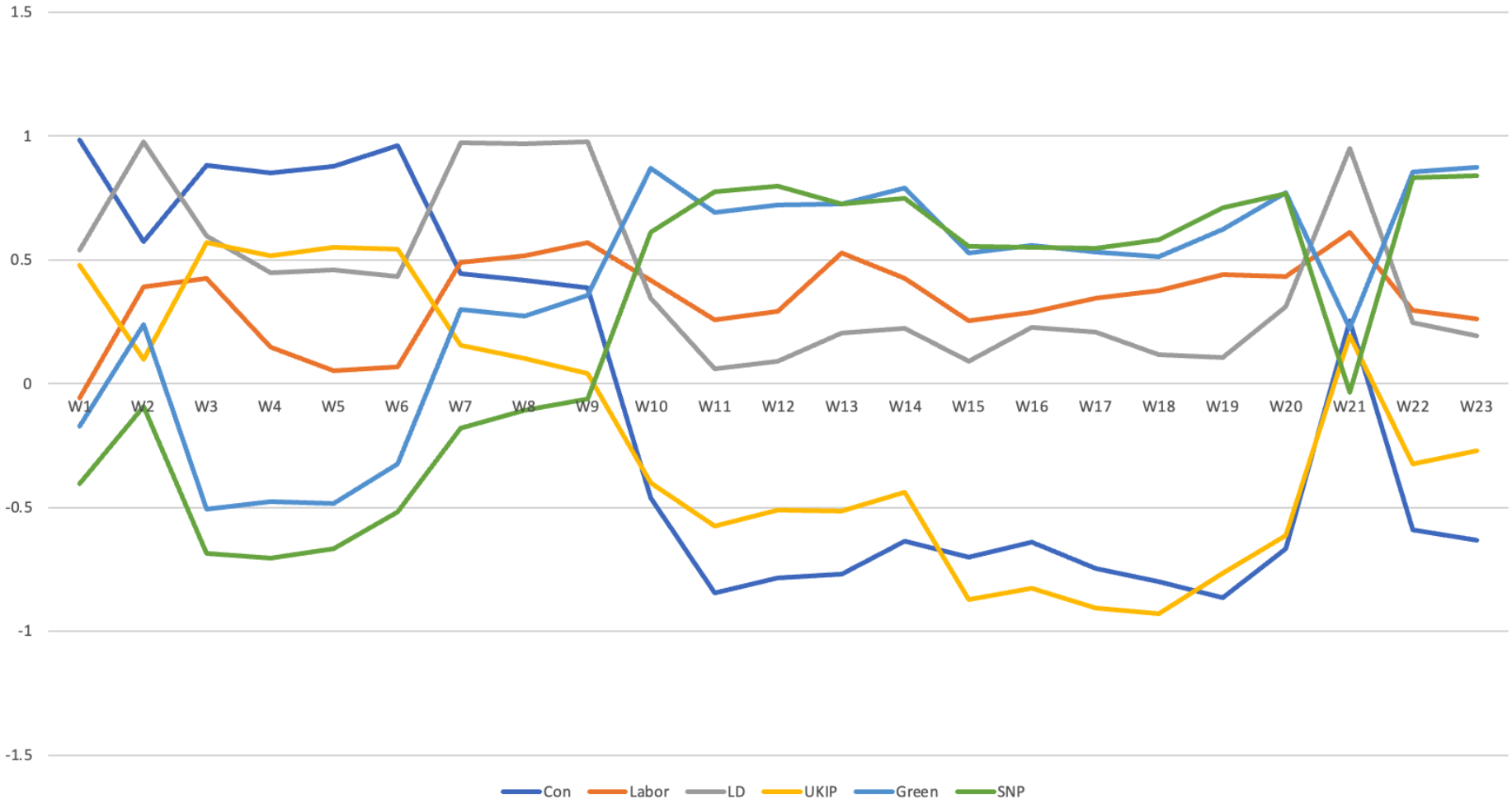
LOADINGS - ENGLAND



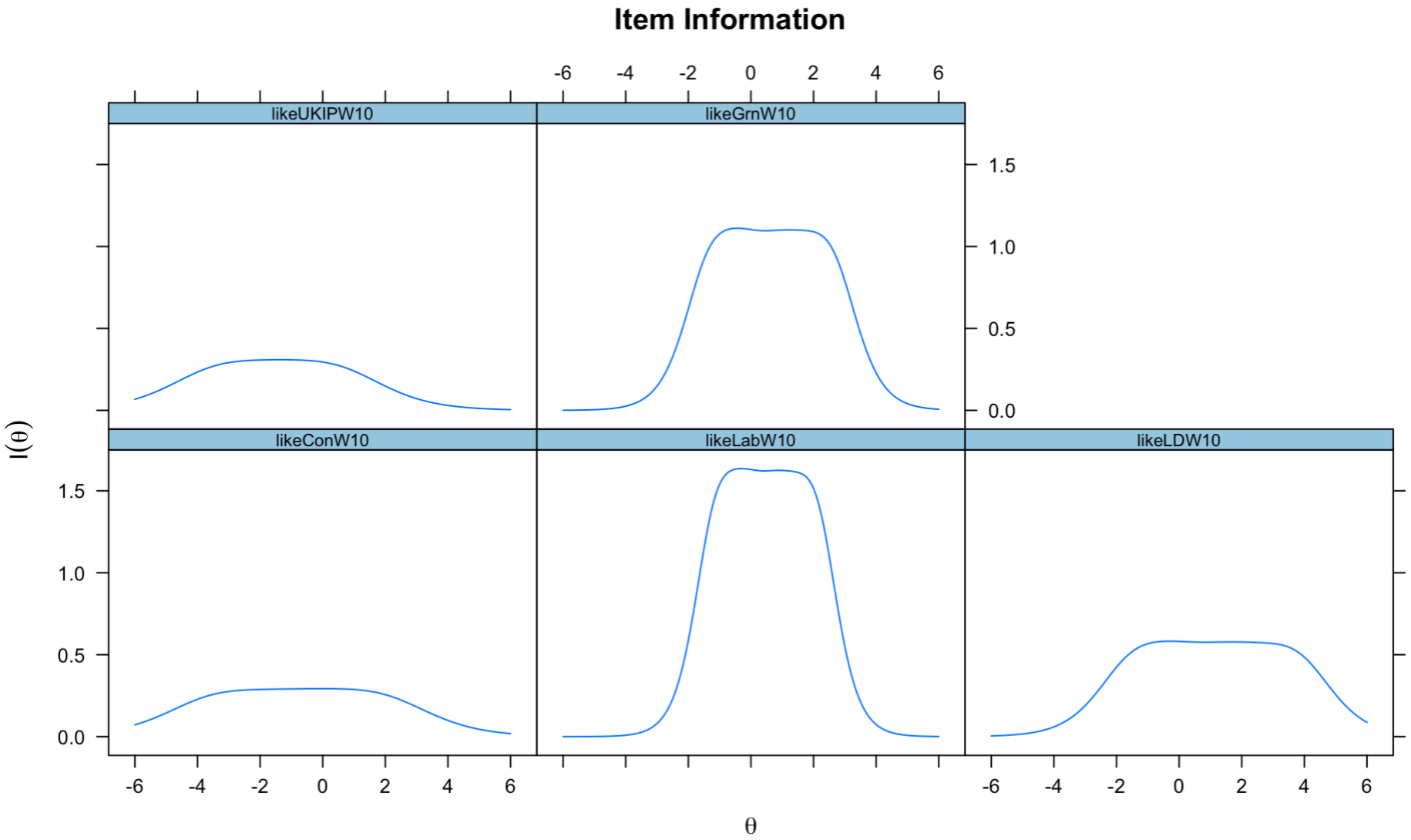
LOADINGS – WALES



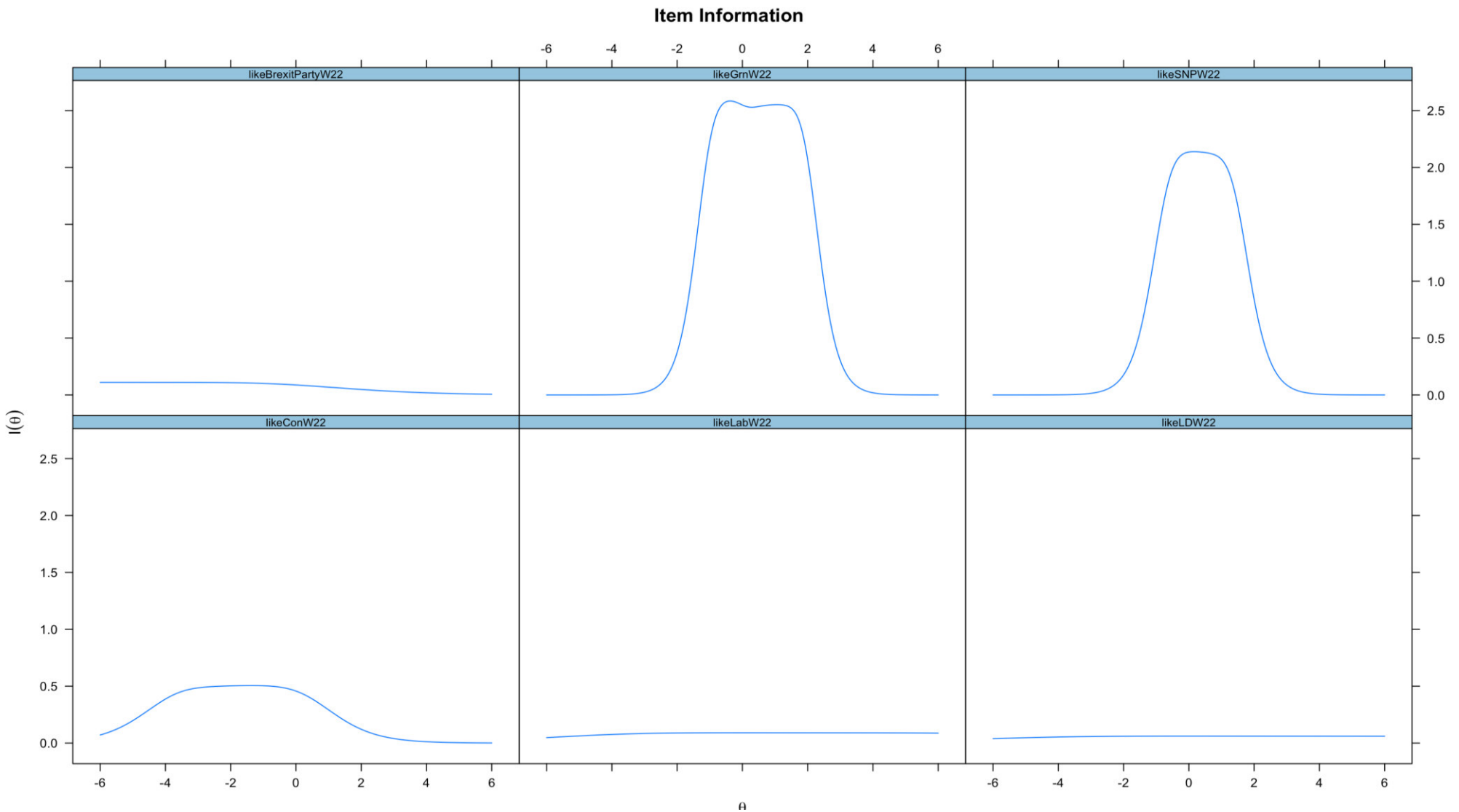
LOADINGS – SCOTLAND



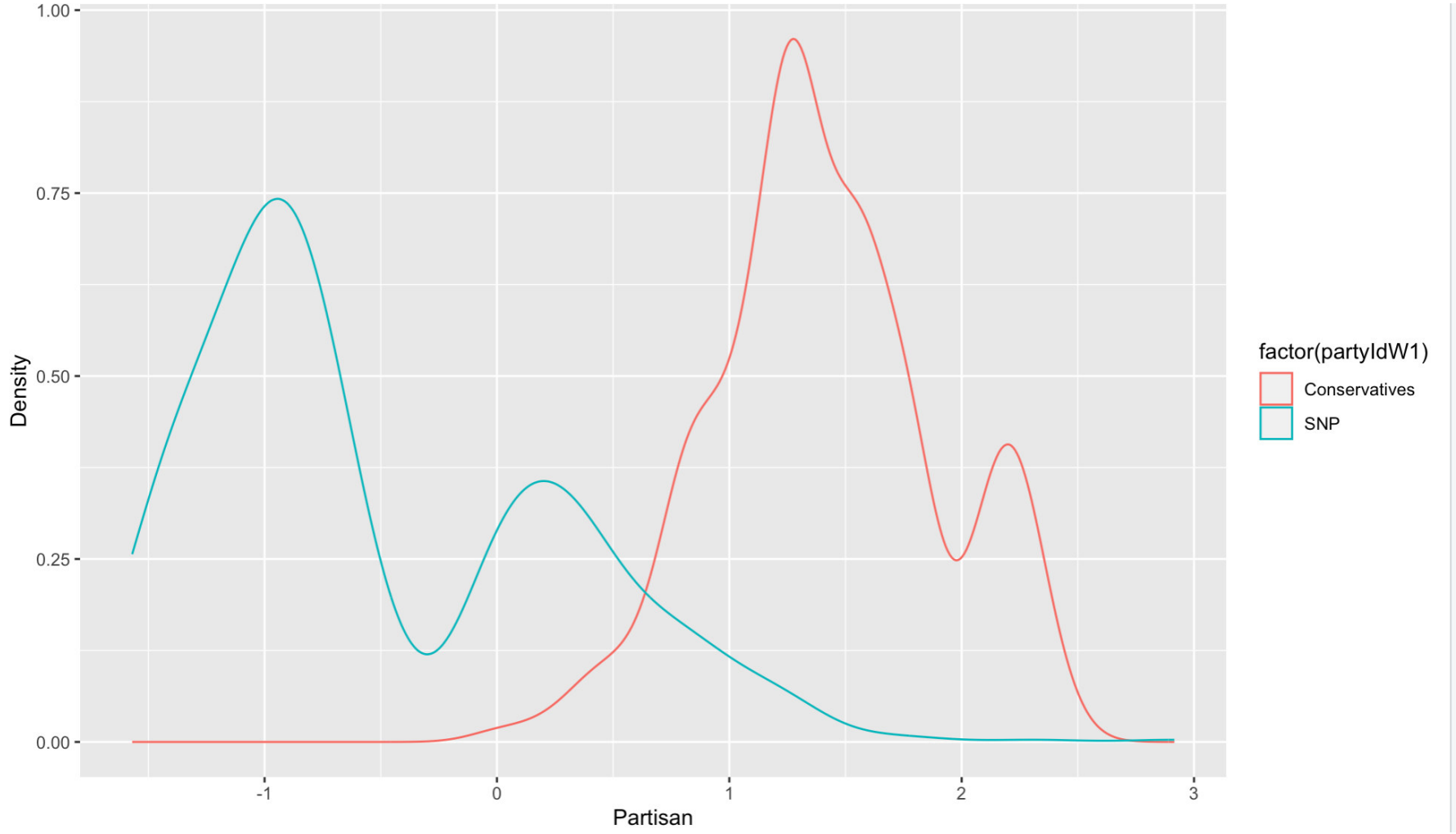
ITEM INFORMATION – ENGLAND EXAMPLE



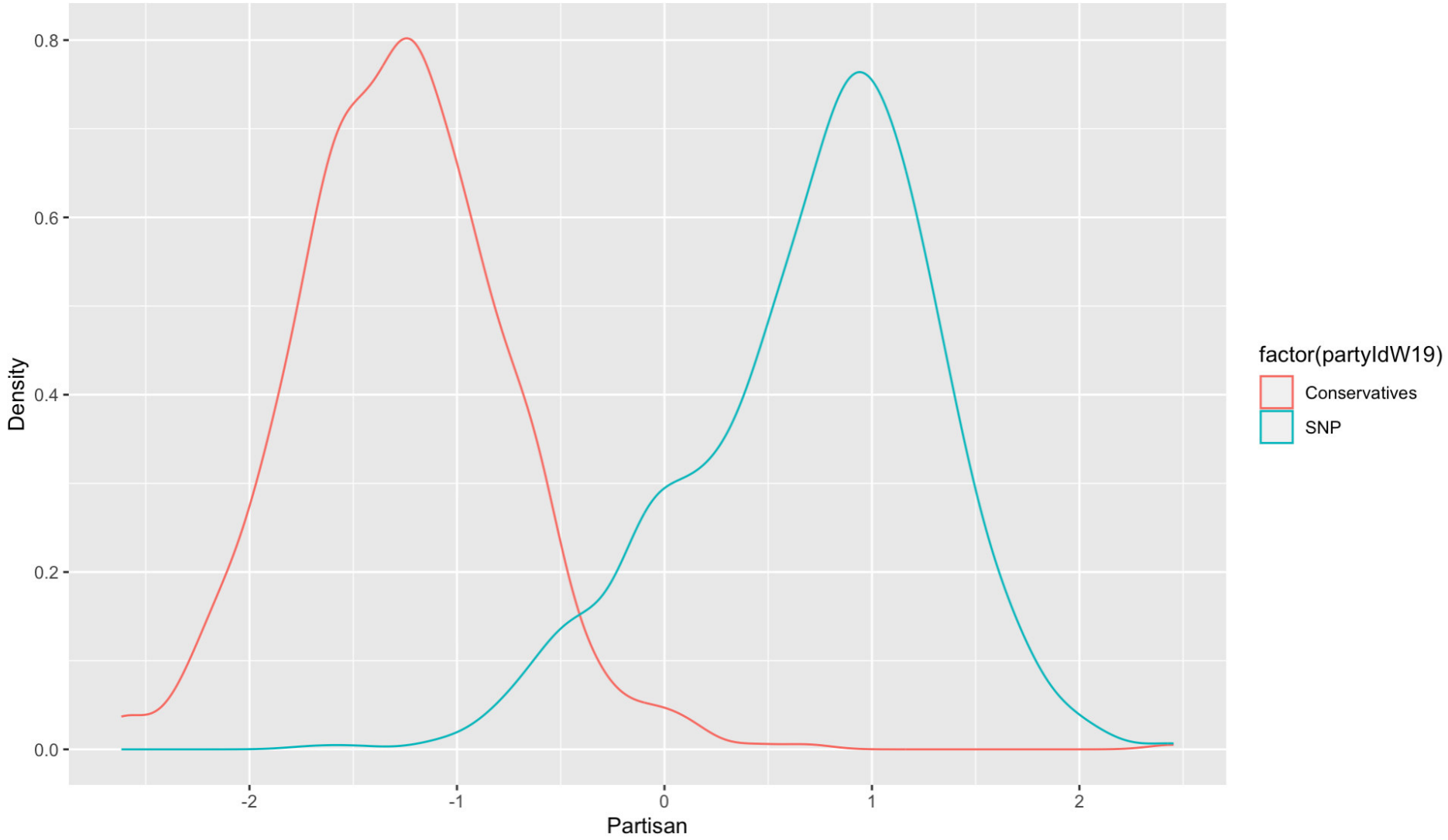
ITEM INFORMATION – SCOTLAND EXAMPLE



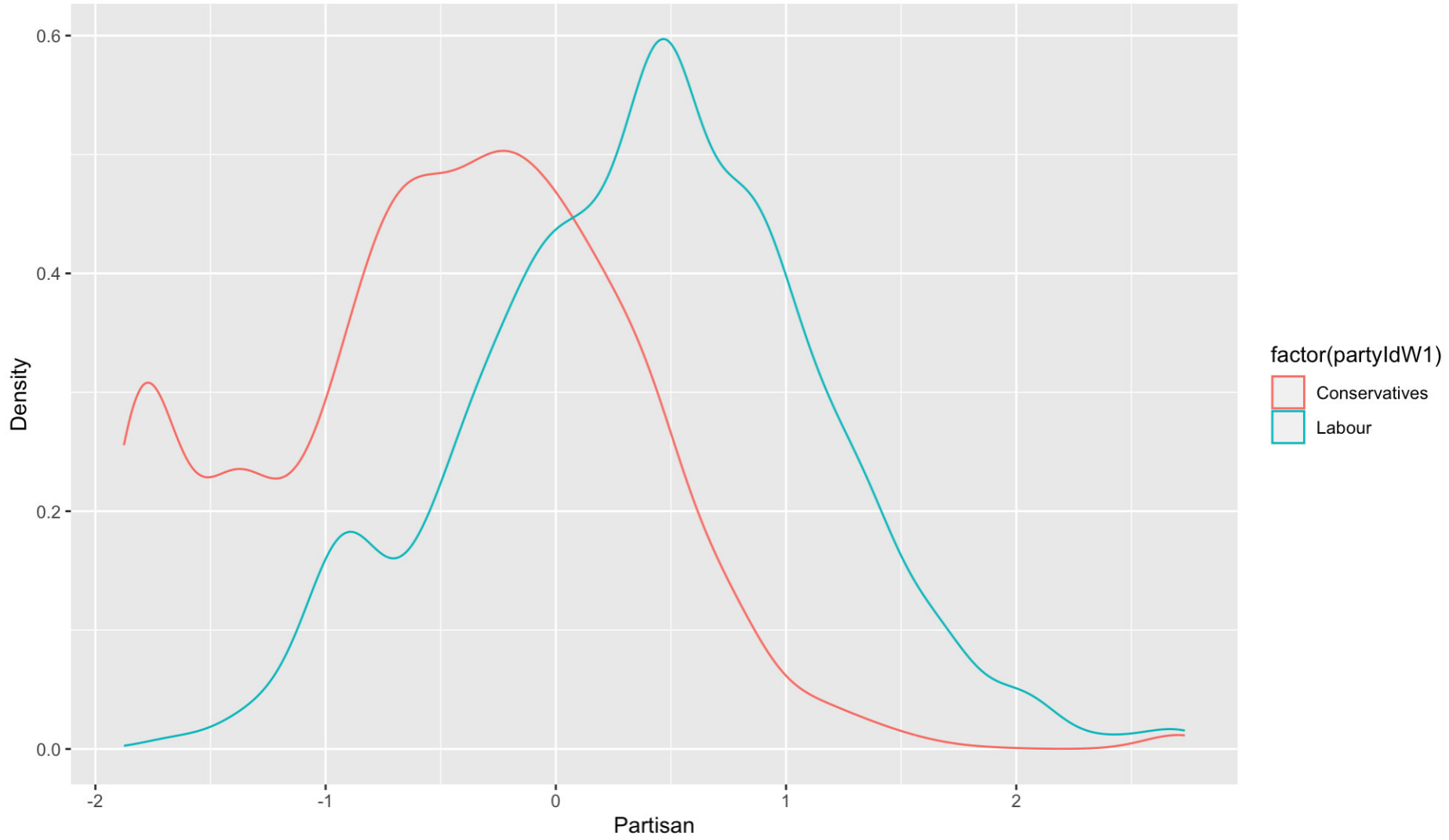
«IDEAL POINTS» – SCOTLAND W1



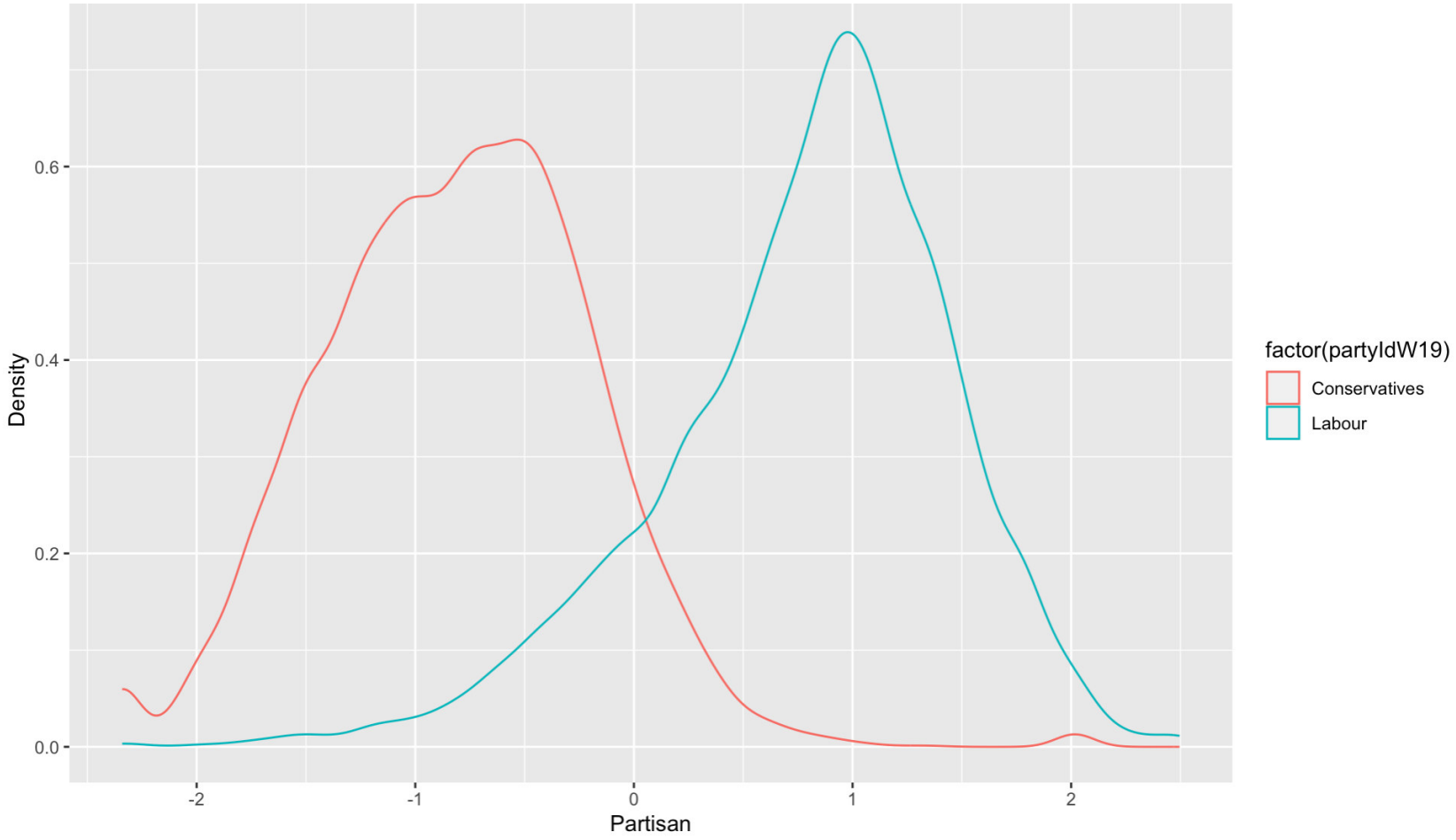
«IDEAL POINTS» – SCOTLAND W19



«IDEAL POINTS» – ENGLAND W1



«IDEAL POINTS» – ENGLAND W19

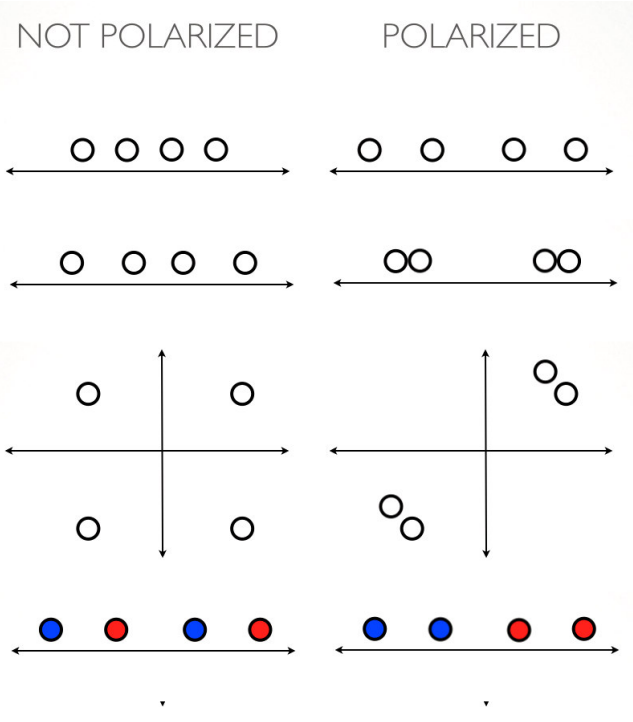


AFFECTIVE PARTISAN SORTING

- The interpretation seems to have 'face value'.
- Issue polarization is leading to affective partisan sorting:
  - Jump in Scotland around the indep. referendum but not sticky.
  - Huge jump after the Brexit vote.
  - Elite level gridlocks and polarization leading to affective partisan sorting more than the electionsand referenda.
  - The whole Brexit process seems to sharpen the affective partisan divisions.
  - There is a significant overall increase during the turbulent times.

WRAP-UP

- Issue polarization is leading to **affective partisan sorting** in the UK.
- Sorting is in fact a mechanism for polarization, hence, opinion based group conflicts are fueling affective partisan polarization.
- Promising methodology for measuring affective partisan identities and polarization in multiparty settings.
- What is next?
  - How do different subgroups look like? Leavers vs. Remainers, YES vs. NO, youngs vs. olds etc.
  - Clustering the affective partisan scores?
  - Use standard polarization measures like Duclos-Esteban-Ray using the individual scores.
  - Apply the methodoloy to other countries for which there is survey data for more than one wave. (Already started the analysis for Turkey using the CSES data.)
  - This can go beyond partisan affective polarization.

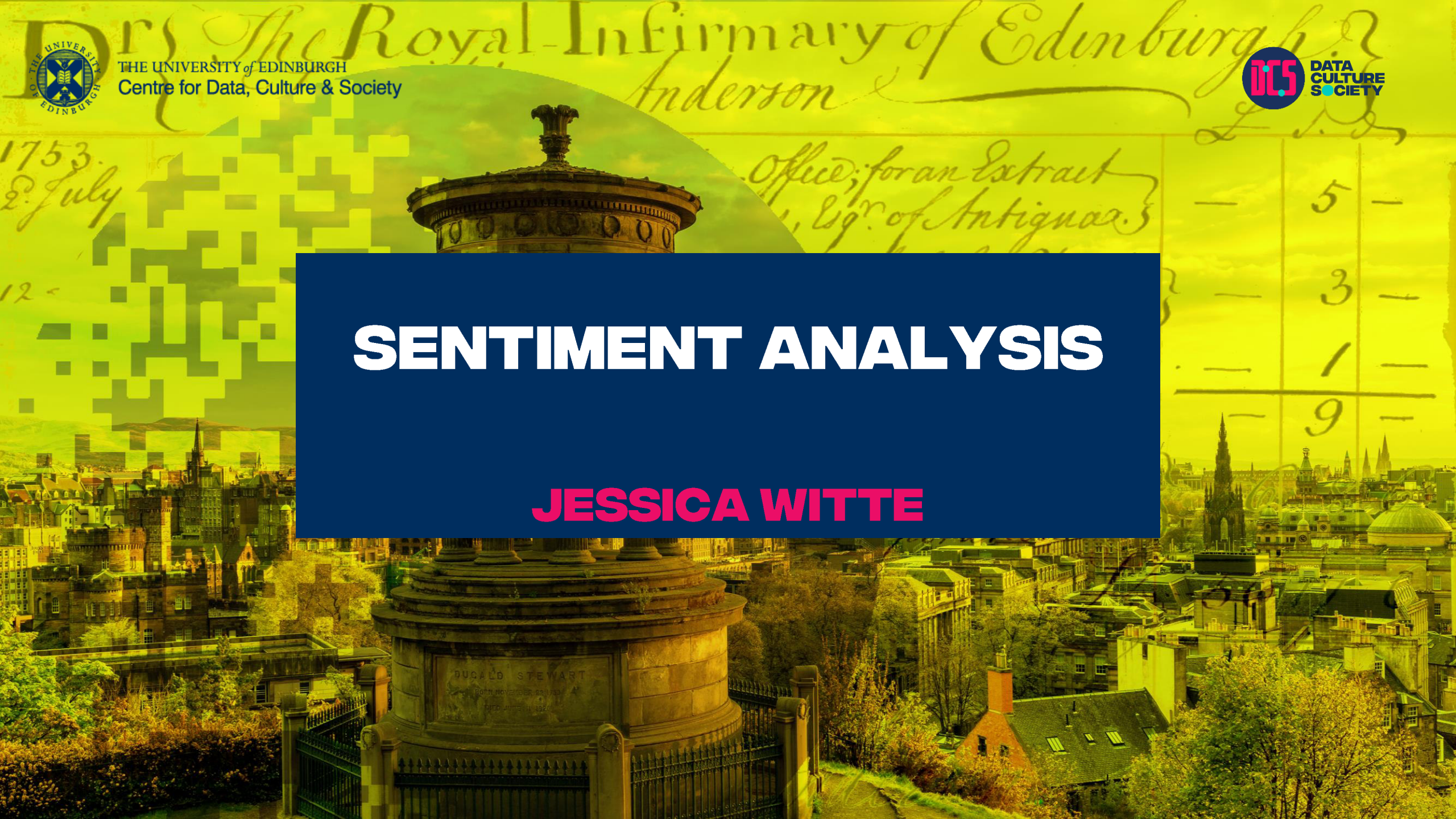


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

COFFEE BEAK

WE ARE GOING TO RESTART AT 11:00


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
SENTIMENT ANALYSIS
JESSICA WITTE

WHAT IS SENTIMENT ANALYSIS?

- A form of supervised learning that assigns a polarity score to textual data
- Existing tools classify text based on a lexicon or a set of rules; aspects, or features of the text; or though deep learning
- Can analyse text at various levels (e.g. word/aspect, sentence, paragraph, document)
- Designed to parse opinionated data
- For best results, training data should resemble data under analysis
- Some linguistic limitations include sarcasm, humour, slang, and connotation

“"I am so happy to report that I have been lucky enough to spend, on average, an extra 15 quid per week on groceries this year compared to 2022! I have never felt more financially secure."

Sentiment score: 0.868/1.0 (highly positive)

TABLE DISCUSSION

Can you think of a potential use case for sentiment analysis in your field (or specifically in your research)?

LESSON OVERVIEW

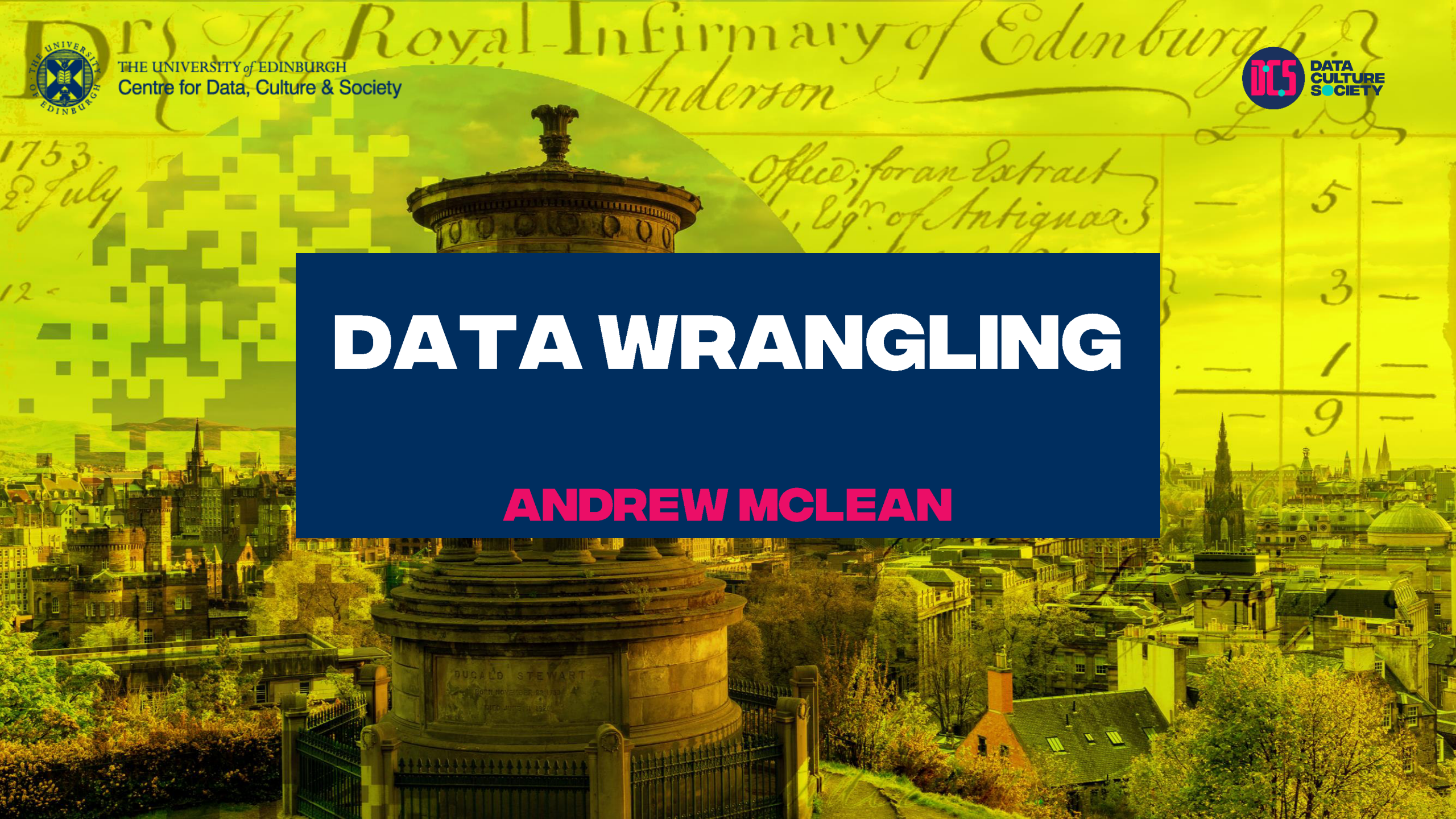
- **Datasets**: posts about the cost of living from 2022-23 scraped from **r/AskUK** and **r/Scotland**
- **Research questions**:

  1. What trends (if any) can we observe about sentiment about the CoL in each dataset?

  2. Based on one dataset only, is the CoL crisis getting better or worse in 2023? How do things differ in Scotland vs. the UK?
- **Topics covered:** data subsetting, sentiment analysis, data visualisation
- **New library**: vader (aka “Valence Aware Dictionary and sEntiment Reasoner”)


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TIME FOR R


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
LUNCH BEAK
WE ARE GOING TO RESTART AT 13:30



DATA WRANGLING

ANDREW MCLEAN

DATA WRANGLING

- Data is often mislabelled or unorganised
- Data wrangling and tidying is simply the process of organizing this data
- The aim is to make it more computer friendly, which isn't always more human friendly

A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

TIDY DATA

- Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In **tidy data**:

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.

A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

GOOD PRACTICE

- Have a separate project for each data analysis
- Keep your data, scripts and associated files somewhere in the working directory
- Save outputs to a separate folder in the working directory

```
> getwd()
[1] "C:/Users/amclea/OneDrive - University of Edinburgh"
> setwd("..")
```


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TIME FOR R


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 15:30


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
BYOD
SESSION 3

Effects of School Starting Age on Counterfactual Reasoning and Verbal Skills

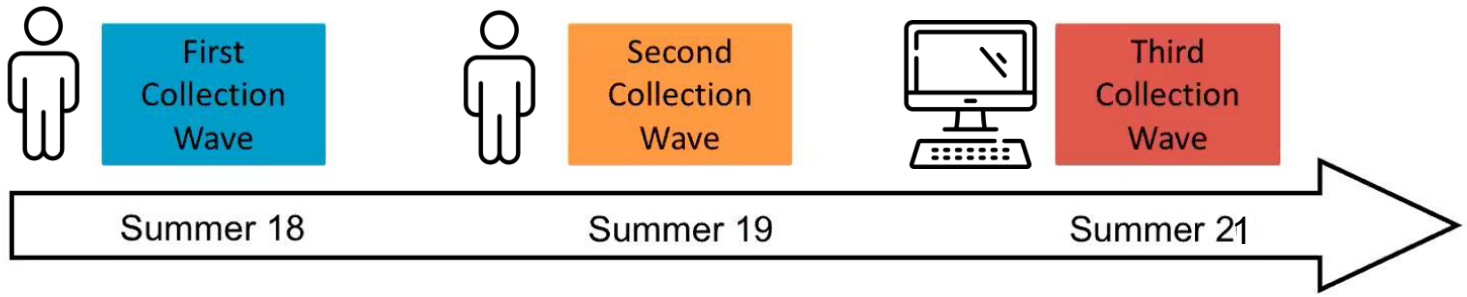
Scottish children born in: 2014

Jan/Feb ..... Dec

Entirely parents' decision

Starting School in:

- 2018 (4.5 years) "Early Schoolers"
- 2019 (5.5 years) "Late Schoolers"



Measurements:

Cognitive/Grades

- Working memory (fNIRS)
- Inhibitory control (fNIRS)
- Cognitive flexibility
- Episodic memory
- Counterfactual reasoning
- Numeracy Screener
- Vocabulary
- Academic Achievement

Noncognitive/Regulation

- Strength and Difficulties Questionnaire
- Parental Stress Index Child Domain
- Child Behavioral Questionnaire

Home & School Environment

- Home Conditions Scale
- HOME measure
- SES
- Parenting Daily Hassles
- Family Activities Scale
- Parent Teacher Involvement
- Parental Stress Index Parent Domain

Do late schoolers, due to being older when entering school, show larger schooling-specific cognitive changes in counterfactual reasoning and verbal skills compared to early schoolers?

Developing and Validating the Comprehensive Hierarchical Eustress Review (CHER)

Background. The CHER instrument aims to comprehensively assess positive stress (i.e., eustress). Based on the interdisciplinary literature, we created 47 items, each representing a unique feature of eustress.

Methods. We collected data from 262 UK adults on the CHER instrument, a distress measure (PSS), a eustress measure (ES), a personality inventory (HEXACO-60), and demographic information.

Data. Our dataset is in wide format, consisting of 133 columns (survey items) and 262 rows (participants).

Does the CHER instrument sufficiently cover eustress? What is the psychometric structure of eustress? Can we cluster participants into different eustress profiles?

Variables.

- **CHER.** 47 items divided into 3 subscales (18x goal-directed behaviours, 18x momentary experiences, 11x stable qualities of the agent)
  - 0 to 10 continuous response scale (one decimal place); ↑ rating = ↑ eustress feature
- **Perceived stress scale (PSS).** 10 items
  - 0 to 10 continuous response scale with one decimal place and higher ratings indicating higher levels of negative stress (i.e., distress)
- **Eustress scale (ES).** 10 items
  - 0 to 10 continuous response scale (one decimal place); ↑ rating = ↑ eustress
- **HEXACO-60.** 60 items divided into 6 subscales a 10 items (honesty, emotionality, extraversion, agreeableness, conscientiousness, and openness)
  - 1 to 5 continuous response scale (one decimal place); ↑ rating = ↑ personality trait
- **Demographics.** 5 items (age, gender, annual income, education, subjective social status)

Chat About Chat

A dataset of tweets about Large Language Models (LLM)

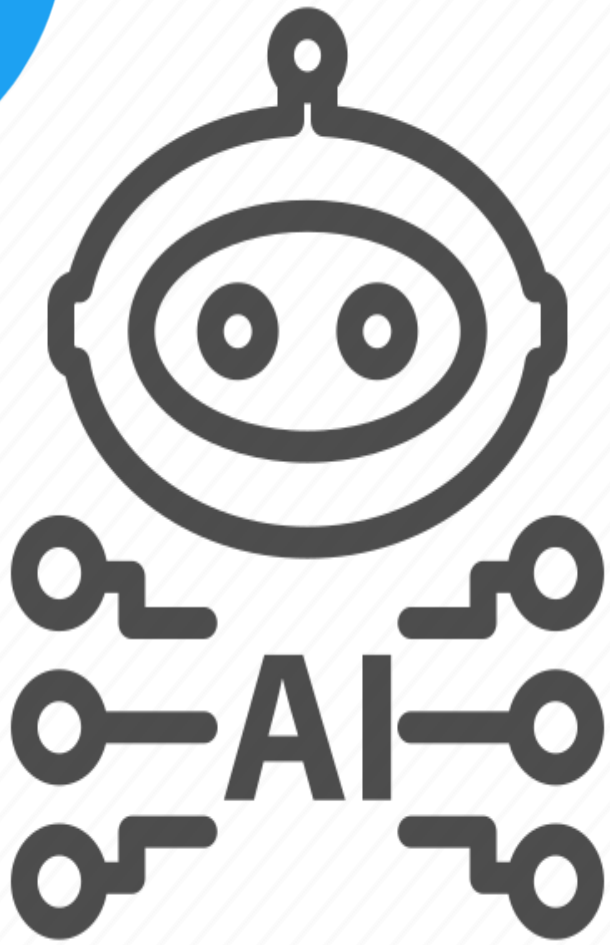
What? Tweets that mention or discuss large language models, such as ChatGPT.

Overview: Over 50,000 tweets from various sources, including researchers, practitioners, journalists and the general public.

Topics include development, use cases, performance, ethical considerations and impact on society.

Each tweet in the dataset includes information such as the tweet ID, timestamp, user ID, username, location, tweet text, hashtags and other metadata.

Uses: studying attitudes to and issues around large language models from a social media perspective, including text and sentiment analysis.



Dataset source: Konrad Banachewicz (Kaggle) https://www.kaggle.com/datasets/konradb/chatgpt-the-tweets

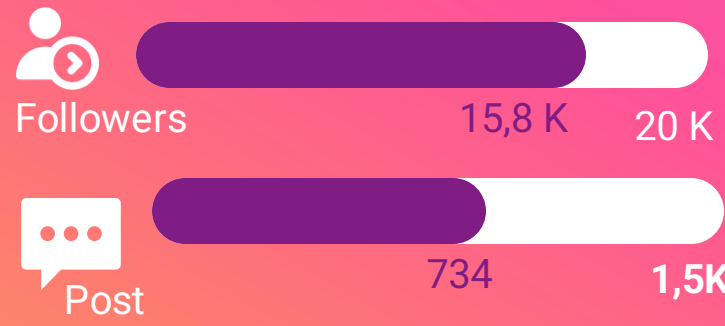
Millennium Cohort Study (MCS)

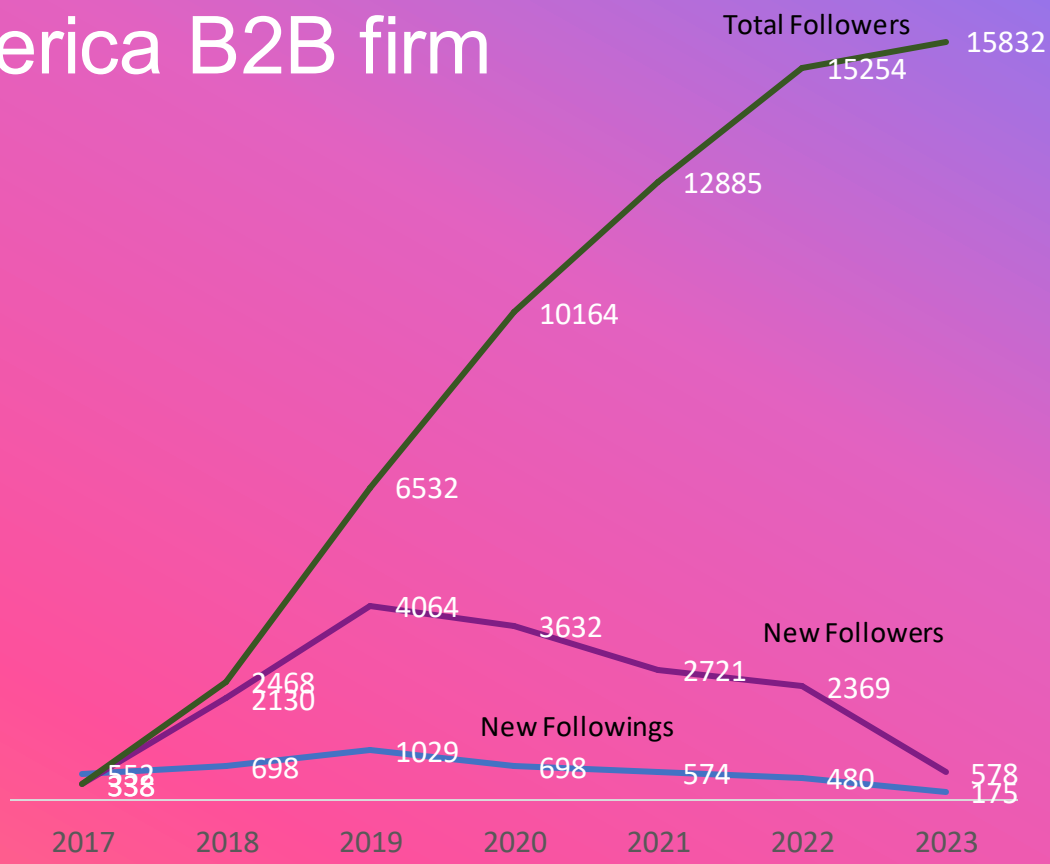
- First assessment in 2001 (families with children aged 9 months)
  - 18,818 children (from 18,552 families) from England, Scotland, Wales, Northern Ireland
  - Participants from deprived backgrounds and ethnic minorities over-represented
- Data available from seven time points
  - 2001 (age 9 months), 2004 (age 3 years), 2006 (age 5 years), 2008 (age 7 years), **2012 (age 11 years), 2015 (age 14 years), 2018 (age 17 years)**
- Relevant Datasets: young people aged 14 (N=11,859) and 17 (N=10,345), plus their parents at “baseline” (N=11,717) for demographic assessments

Instagram Account from a Latin America B2B firm

About Instagram Account

This company has the account from 2017. In 2020 was their best year and now comments are dropping.





What's NEXT?

Gloria Chacon
PhD Candidate – St Andrews University

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

CEILIDH

TEVIOT DEBATING HALL

13 Bristo Place, Edinburgh

EH8 9AJ

www.cdcs.ed.ac.uk