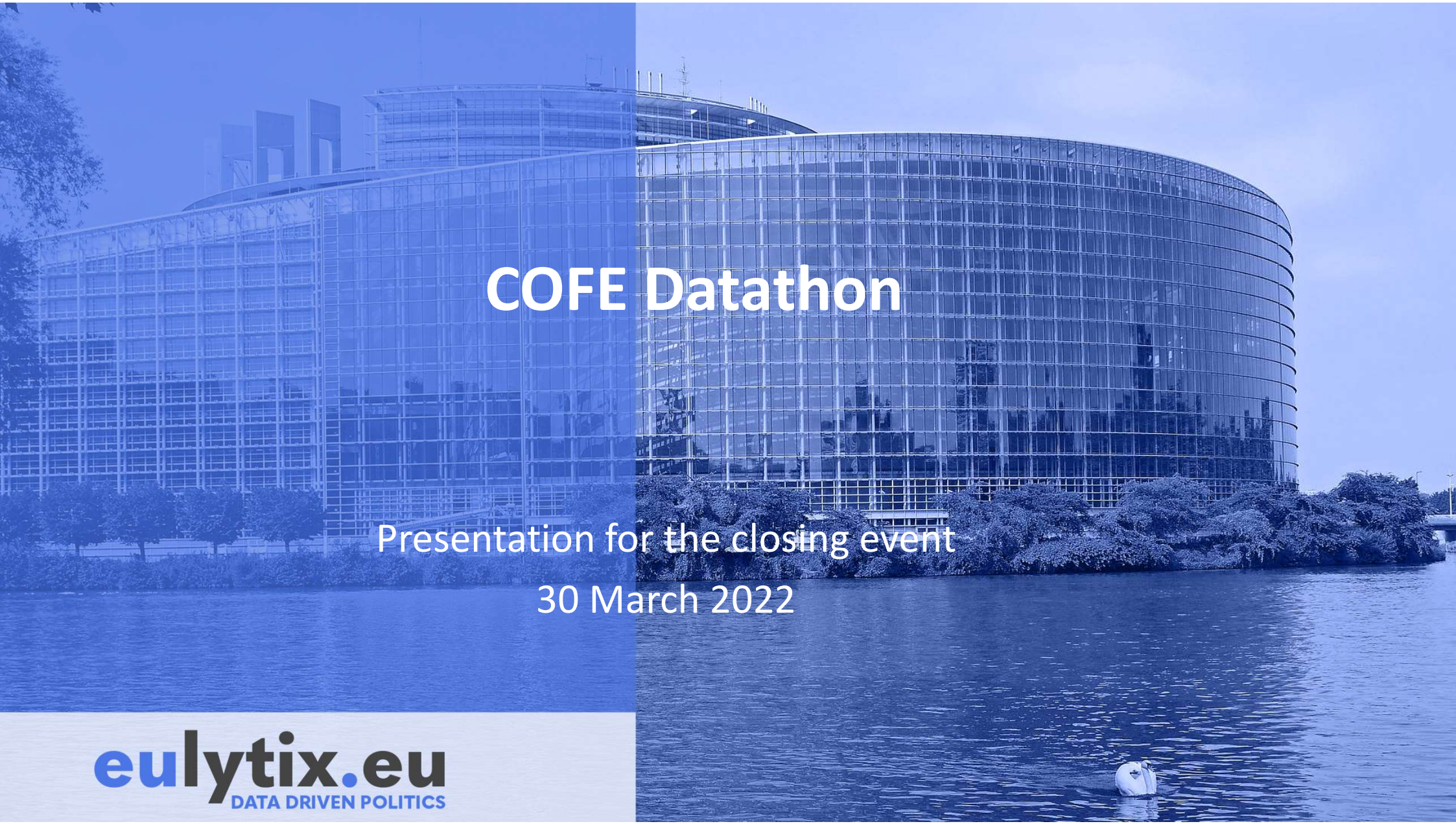

COFE Datathon

Presentation for the closing event

30 March 2022

eulytix.eu

DATA DRIVEN POLITICS

Challenge 2 – Proposal Clustering

- Goal: to identify topics based on the semantic similarity of the texts
- Method: Top2Vec
- Steps:
  - A joint embedding of words and proposals with Doc2Vec
  - Mapping the sematic space to a lower dimension with UMAP
  - Identification of clusters with HDBSCAN

Solution:

131 topics were identified and they were aggregated to 50, 20 and 10 topics

Figure 1: Illustration of the UMAP-reduced document vectors of the CoFE dataset and the identified clusters. Different colours indicate different clusters. Note that different clusters may have the same colour.

Solution:

Topic0	Topic1	Topic2	Topic3	Topic4
teachers	asylum	army	teu	patients
educational	seekers	nato	cjeu	medical
teaching	reception	military	court	patient
curriculum	migrants	defence	rule	medicines
education	refugees	armies	article	doctors
students	refugee	armed	law	medicine
erasmus	migrant	afghanistan	bravo	healthcare
school	migration	capabilities	fundamental	health
pupils	immigration	defense	judicial	chronic
teacher	immigrants	allies	judges	treatments

Table 1: Keywords of the top 5 topics

topic_id	topic_frequency
0	772
1	600
2	526
3	487
4	466
5	445
6	403
7	394
8	388
9	386
10	377
11	360
12	342
13	318
14	273
15	270
16	268
17	250
18	249
19	231
20	228
21	219
22	216
23	210
24	209

Table 2: Top 25 topics

Thank you for your attention!

Questions? Comments?

research@eulytix.eu