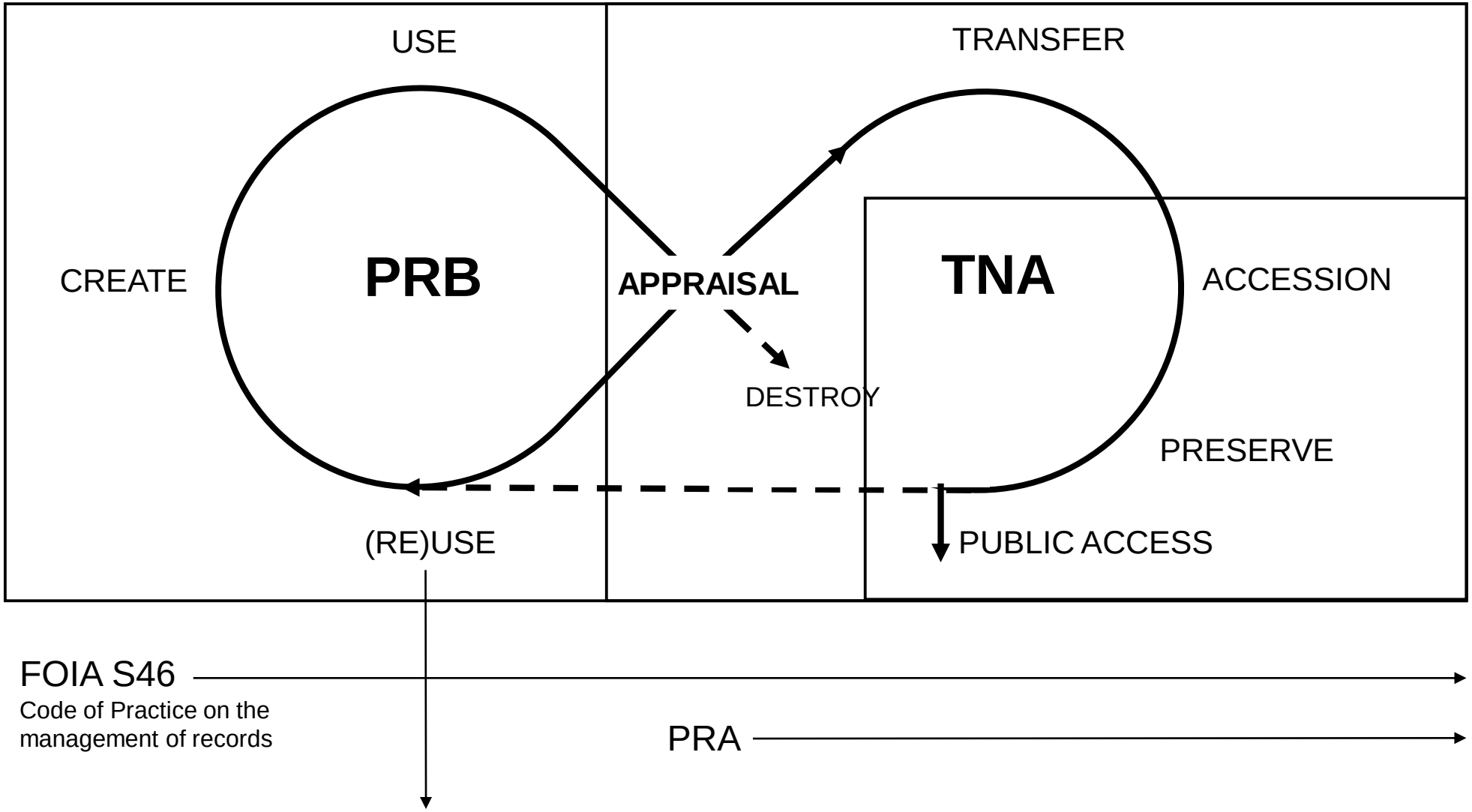
THE NATIONAL ARCHIVES

LLM for Records Appraisal

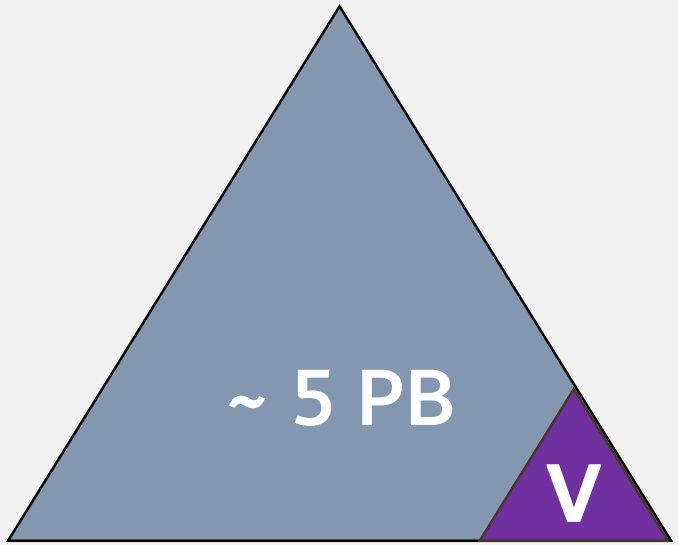
Balint Csollei & Chris Royds
Cross Government Engagement


USE
TRANSFER
CREATE
PRB
APPRAISAL
TNA
ACCESSION
DESTROY
PRESERVE
(RE)USE
PUBLIC ACCESS
FOIA S46
Code of Practice on the management of records
PRA
National Data Strategy
Mission 3: Transforming government's use of data to drive efficiency and improve public services

Volume of digital material in Government

- ~16 billion emails
- ~3 billion documents
- Growing by ~8 billion emails / 1 billion documents in a year
- Estimated volume was 5 PB in 2018
- **It would take 990 people 100 years** to review by using an analogous approach to that used for paper

Better Information for Better Government (2018)



This growing volume of information is poorly understood and presents risks to departments, including:

- **Non-compliance with the Public Record Act (PRA) and other statutory requirements like DPA and FOIA.**
- **Increasing liability and costs due to a lack of understanding of the content and context of information.**
- **Inability to identify and extract value from vast amounts of legacy records.**

TNA Research - Previous experimentations

Applying traditional approaches to the massive volume of born-digital records is not viable;
appraisal, selection, and sensitivity review of digital records will only be feasible with machine assistance.

The lack of such assistance will lead to delays in the transfer of records to the archive, which negatively impacts our ability to support transparency and openness by providing public access to public records.

e-Discovery (2013)

Published a report on testing eDiscovery tools, originally developed for the legal industry, to locate information within large collections of documents.

>>> supports search for specific information but not suitable for high level appraisal

Machine Learning - AI for Selection (2019)

– market research and prototype to apply supervised learning to pre-labelled records.

>>> can be effective, but it requires training on a substantial volume of labelled data

LLM Sandbox (2024 -

Exploring the potential of transforming unstructured text into structured and meaningful forms, enabling labelling and text classification.

How efficient, effective, trusted this new intelligence and what are the use cases to assist KIM?

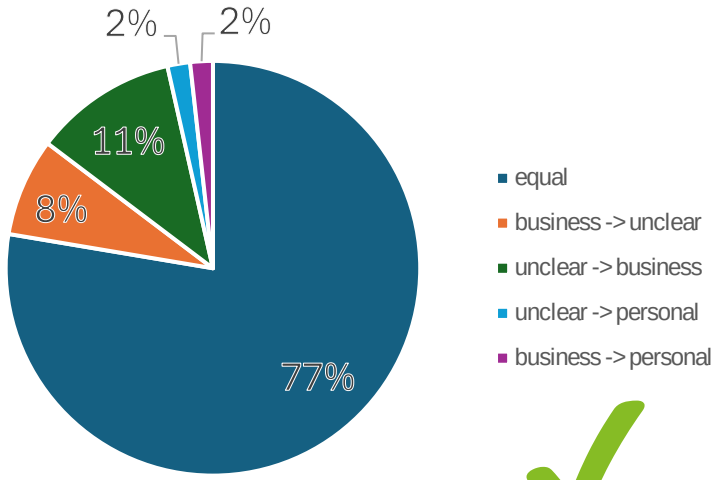
Classifying documents with LLMs

Being able to classify the contents of folders without investigating each file individually would be useful, and assist in high-level appraisal of potential records.

Few, clear-cut categories

Prompt: Classify the given email as **“Business”**, **“Personal”**, or **“Unknown”** (if there is insufficient context to determine the correct category). Reply with ONLY ONE WORD. Do NOT provide any explanation of your choice. EMAIL: {email}

150 emails



Many, nuanced categories

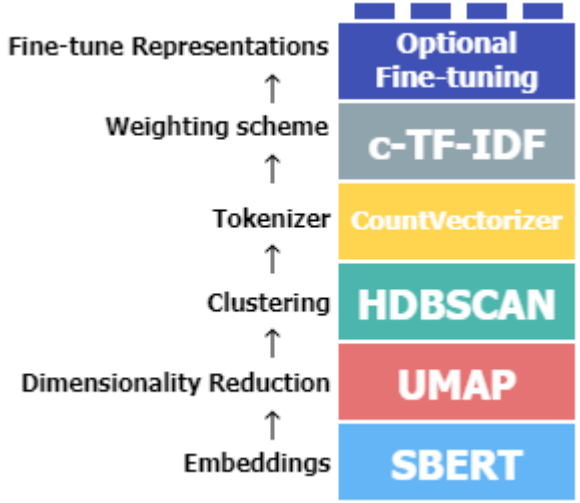
Prompt: Classify the given document using one of the following categories: {categories}, or “Unknown” (if there is insufficient context to determine the correct category). Reply with JUST THE CATEGORY. Do NOT provide any explanation of your choice. DOCUMENT: {document}

Range of Department for Education documents

- Education Policy
- Education Policy
- Education Policy
- Education Policy
- Education Policy

Clustering by Topics

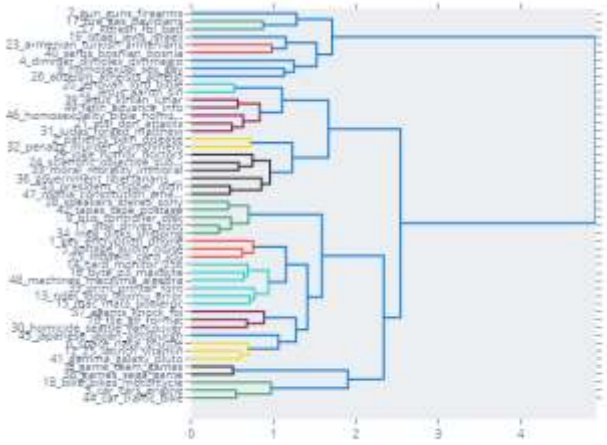
BERTopic is a modular tool which takes natural-language texts, produces vectorised embeddings, and then clusters the texts according to topics covered.

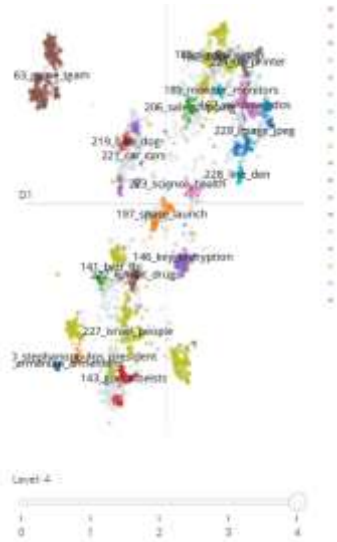


Our experiments focused on clustering ~7k DfE PDFs. The resulting clusters were difficult to interpret; by default, BERTopic reports clusters using their most common words, relative to the corpus as a whole – e.g. one of the clusters was reported as: ['cent', 'per', 'olds', 'eleven', '1997', 'reached', 'provisional', '2005', 'standard', 'up']

However, looking at the original URLs, and the documents themselves, the clusters were clearly still quite good!

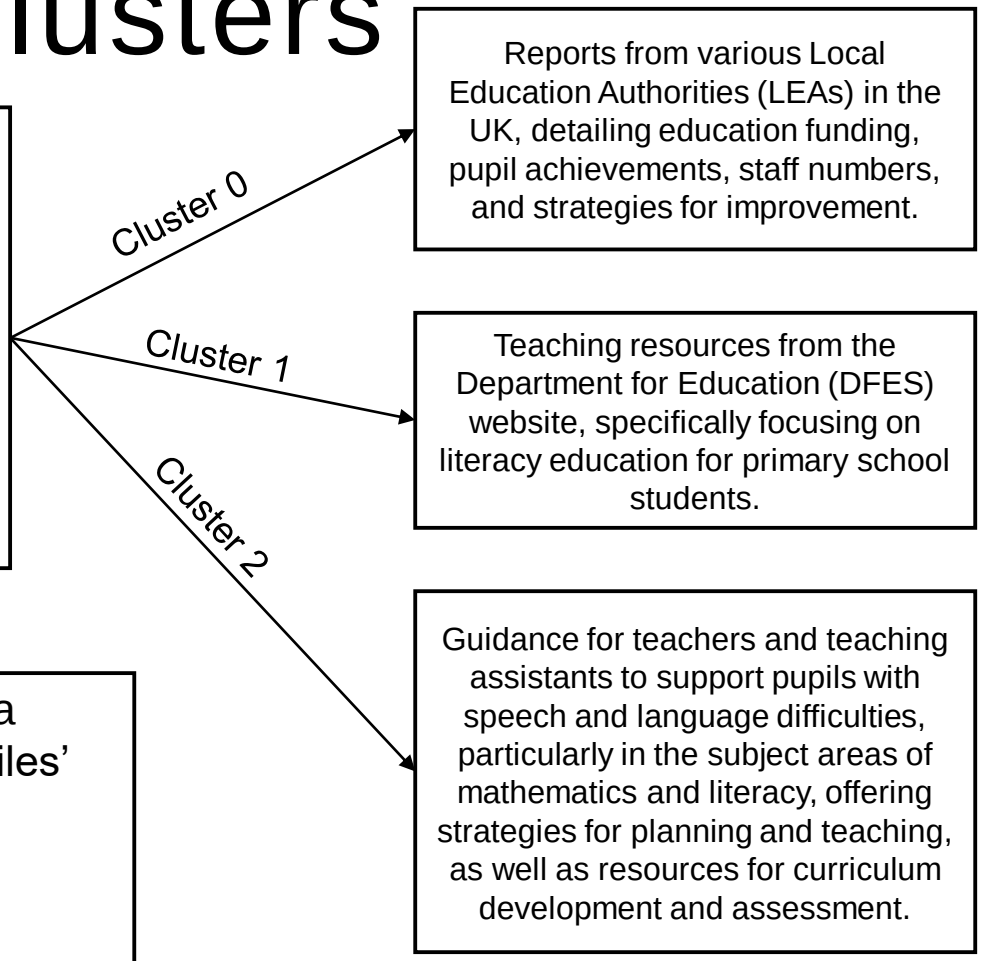
>>> Demo in Power BI





Summarisation of clusters

Prompt: Here is a sample of three documents from Set {cluster_number} - a set of documents from the DfE; words which appear commonly in the set include {list_of_keywords}: SAMPLE 1: {sample_1}; SAMPLE 2: {sample_2}; SAMPLE 3: {sample_3} END OF SAMPLES. Produce a 1-sentence summary of this set of documents, using the following template: Set 0: Reports on performance of teachers in low-income neighbourhoods.



Conclusion: Clustering and summarising gives a useful, high-level view of an unstructured set of files' contents

Drawbacks:

- Very reliant on drawing the right samples
- Does take time (~5 minutes per cluster)
- ~15% of documents weren't clustered at all

High Level Appraisal will need data driven approaches and new interfaces to view, filter and interact with records at scale

source: https://cambridge-intelligence.com/

Image: AI Generated by using DaVinci – AI app. Keywords were virtual head set, record manager woman, selection records, in a virtual sphere

Developing a **proof-of-concept** tool that provides **insights** and **new ways to interact** with **records at scale**, supporting **record managers' high-level appraisal decision-making** for content stored on **Shared Drives**.

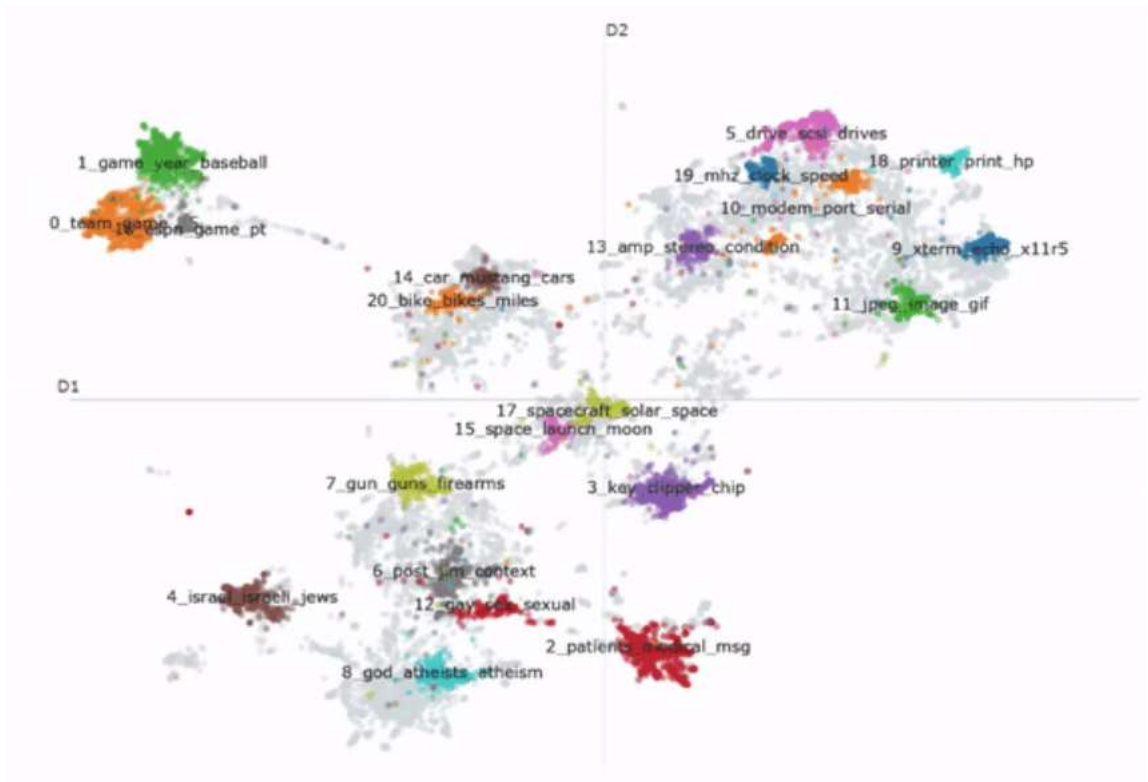
Assumptions:

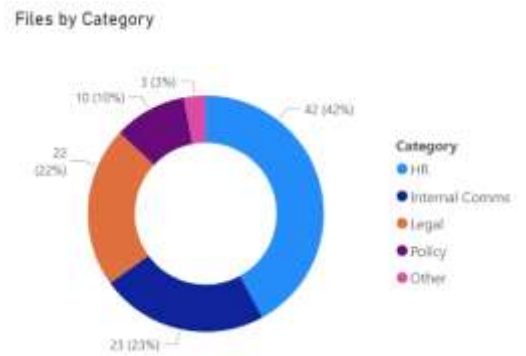
1. AI, particularly Large Language Models (LLMs), will play a significant role in the analysis alongside and in conjunction with traditional data-driven methods.
2. Data visualization can offer a more efficient method to interact with records at scale.

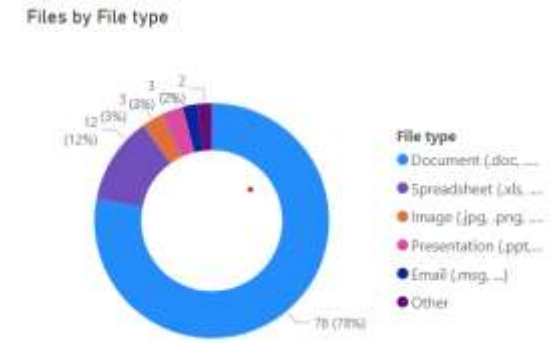
User research	Technical exploration	Prototype development	User Testing and Evaluation

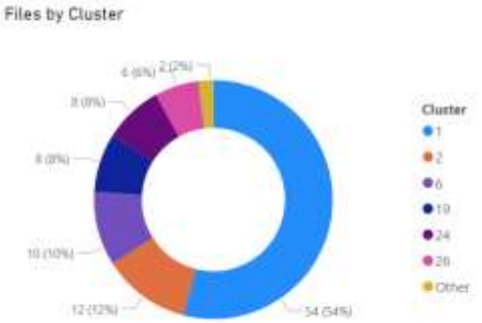
Record Appraisal Tool – Proof of Concept mockup

Directory view | **Cluster view**









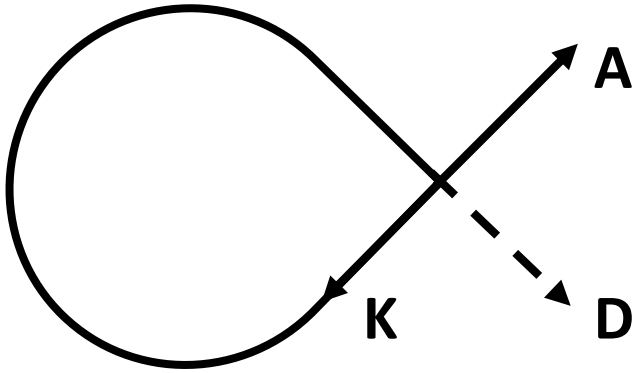
Filters

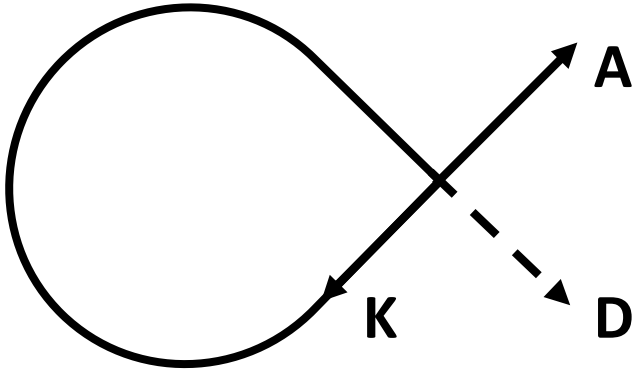
Mark selection for deletion

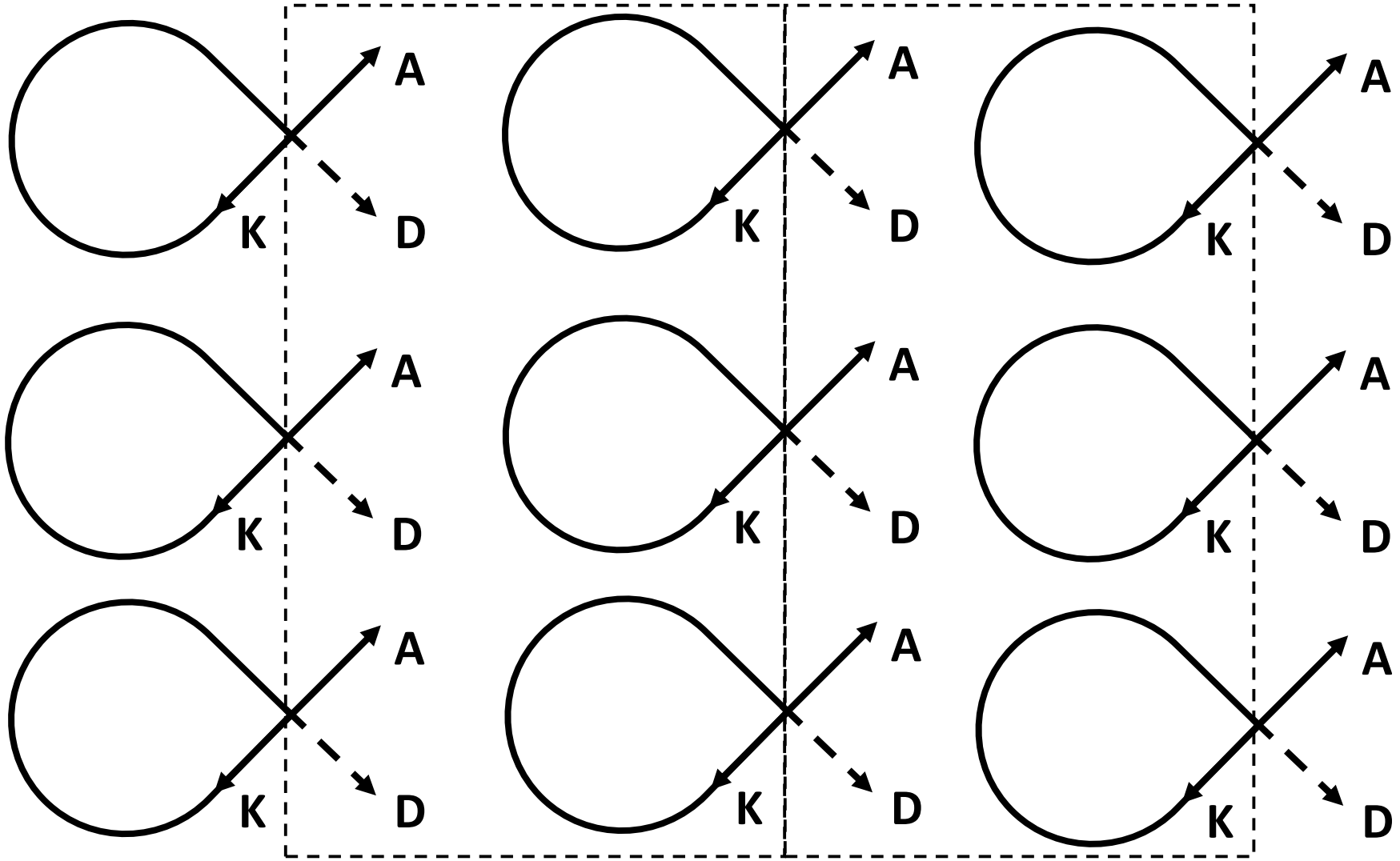
Mark for preservation

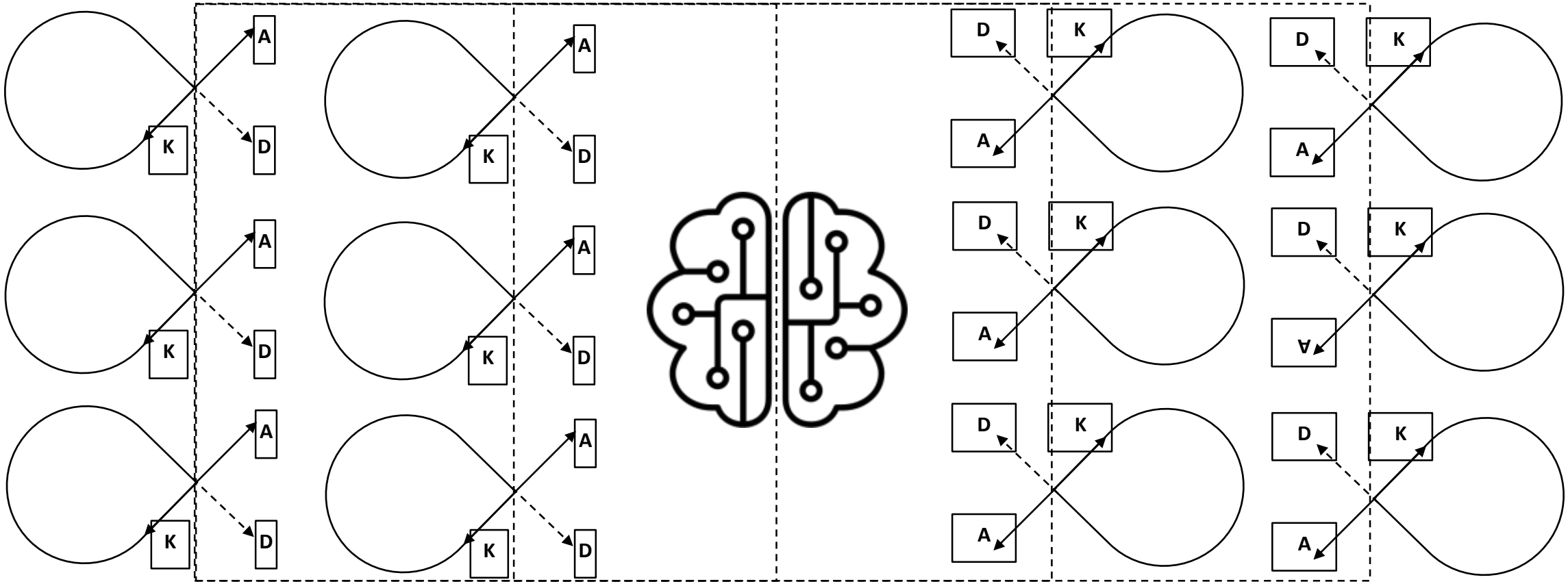
Generate decision report

Cluster	Cluster Summary
1	Documents relating to Project Zeus, instituted by politician X in 2009
19	Correspondence with minister about the cancellation of Project Zeus
2	Series of timesheets from 2001 to 2012
24	Spreadsheets providing analysis on finanical impact of Project Zeus
26	Documents primarily containing personal


A
K
D


A
K
D


A
K
D


A
K
D
A
K
D
A
K
D
A
K
D
A
K
D
A
K
D
D
K
A
D
K
A
D
K
A
D
K
A
D
K
A
D
K
A
D
K
A

Observations

- Any tools developed should focus not on automating deicisions, but on providing tools for Record Managers to help make those decisions
- Given concerns around scalability, sustainability, and environmental impact, as well as the need for data to remain in a private cloud, how do we focus our LLM capabilities where they're most valuable and effective?
- We may need to remove data completely from systems / models – in cases where takedowns would mean re-training a model from scratch (e.g. fine-tuning an LLM), what will be our approach?
- How can we evaluate the results of prompts? What metric can we use to determine if one summary is better / worse than another?
- How can we measure the consistency of responses? How can we demonstrate this effectiveness / consistency to users?
- To what extent are smaller LLMs good enough, balancing effectiveness against sustainability considerations? (In our experimentation, we have found that providing short samples to small LLMs is often more effective than providing full documents)