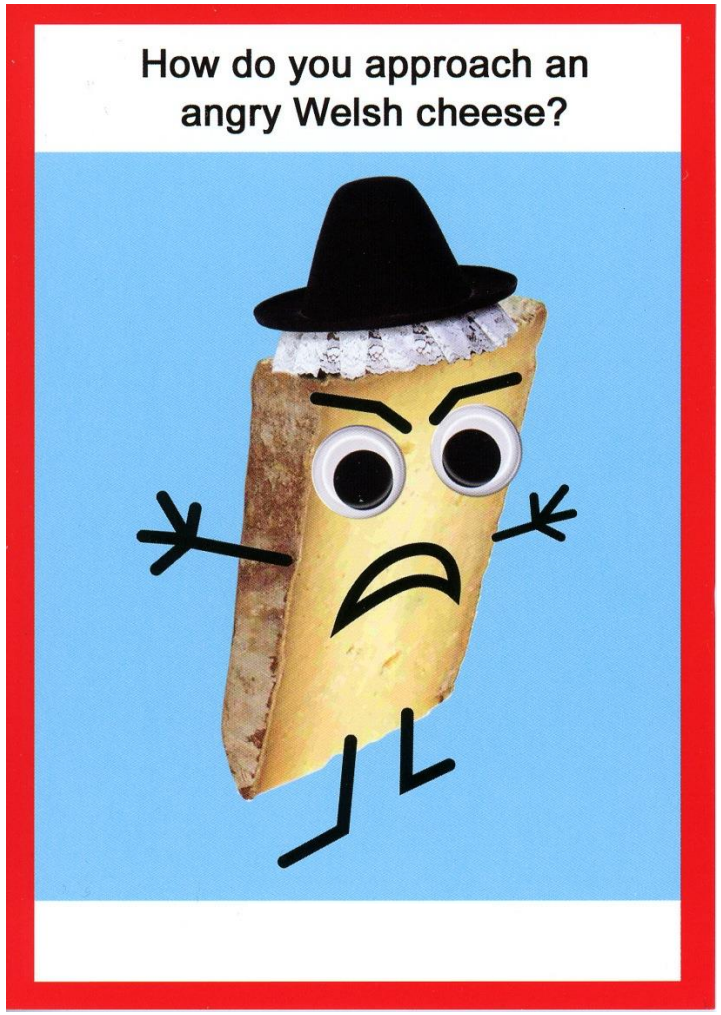
The ‘CDA Explorer’

Dave Camden, Director, Flare Solutions Limited

21st November 2016, The Geological Society, London

Interesting Challenge?



Can be done but tricky
Specific application
Needs training

But cannot ‘experience’ humour

Overview

The following presentation shows some of Flare's work to combine text analytics with existing oil and gas taxonomies to discover analogues based on user-selected criteria.

This approach is part of a journey in moving from a traditional database and search engine towards a semantically aware search based application (SBA).

Concepts

1. There is simply too much content in a typical E&P organisation for people to read and process. We need smart machines to assist in the analysis.

2. Words that occur together probably share some kind of semantic meaning.

3. Words that occur less frequently have more ‘information content’.

4. We are moving away from ‘Expert Systems’ with their hard-coded rules, lists and associations to one where we teach machines using text.

5. Combining text extraction (curation), natural language processing (meaning) and machine learning (prediction) allows us to create domain-specific search based applications/assistants.

6. These assistants can make reasoned suggestions, without cognitive bias, and exploit a company’s information assets to support creativity and innovation.

Text Analytics Applied to CDA Unstructured Information

The Text Source is approx. 25,000 unstructured CDA reports about UKCS wells

A prototype system was constructed that:

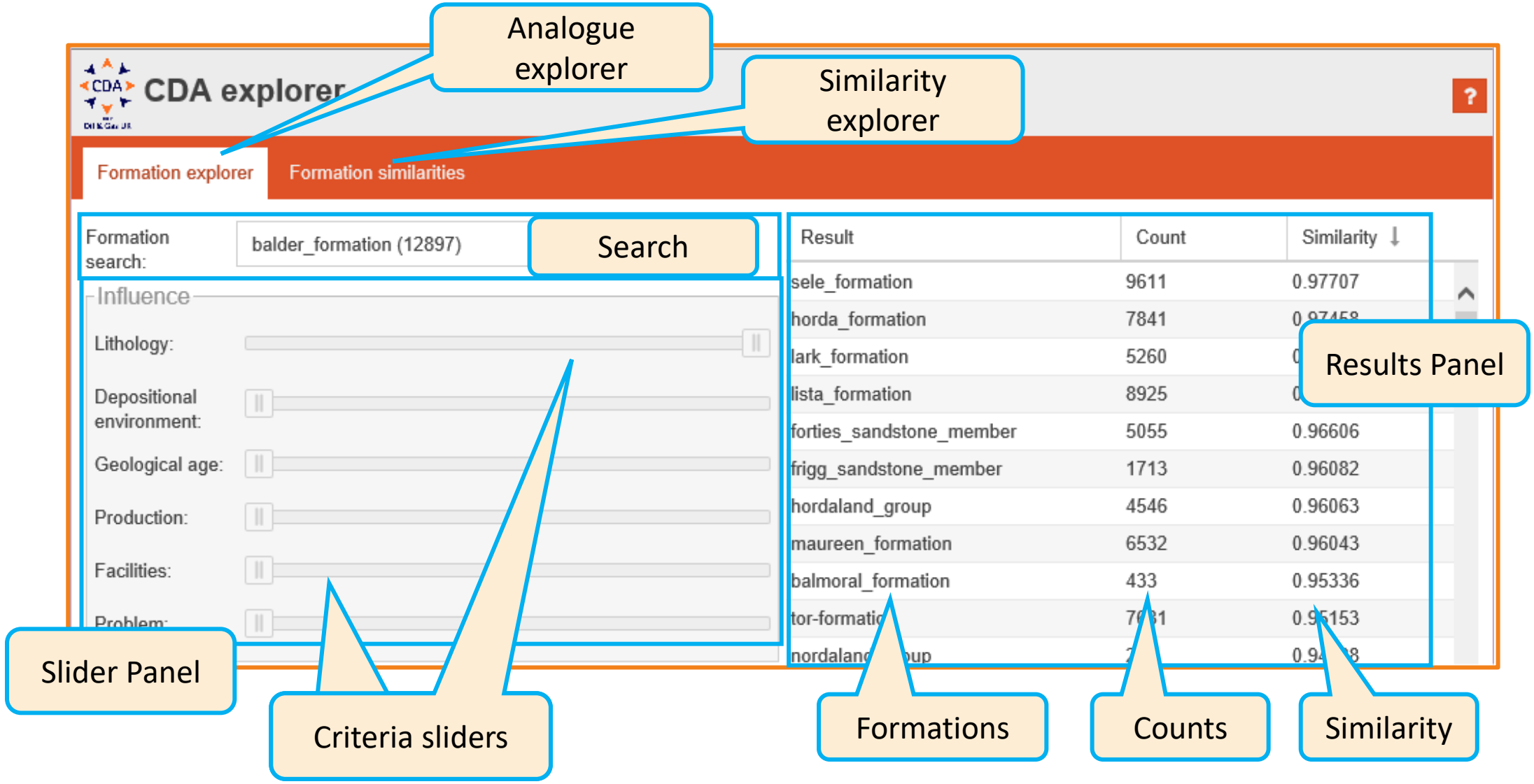
1. Finds ‘formation analogues’ based on a number of controllable criteria
    - Formation names extracted from the report text
    - Each formation is characterised by a number of taxonomical classes/terms, such as
        - Lithology (over 300 terms + aliases)
        - Geological ages (over 200 terms + aliases)
    - User enters a Formation Name and the results show matching Formations based on the selected criteria
    - Criteria can be changed in real-time to influence results

2. Finds formations based on typed value(s)
    - User types in a term (or terms) and results show best match Formations

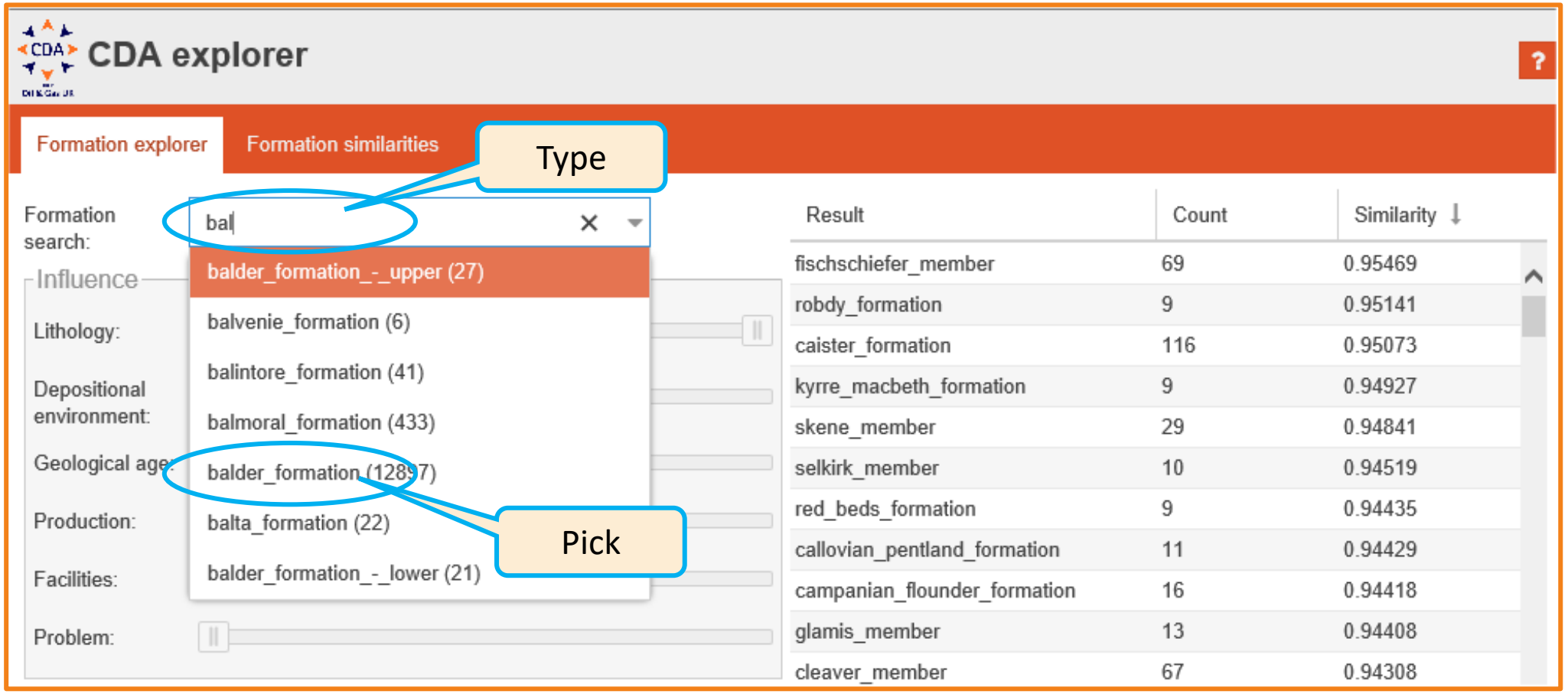
The Prototype

The following slides show how the user interacts with the system

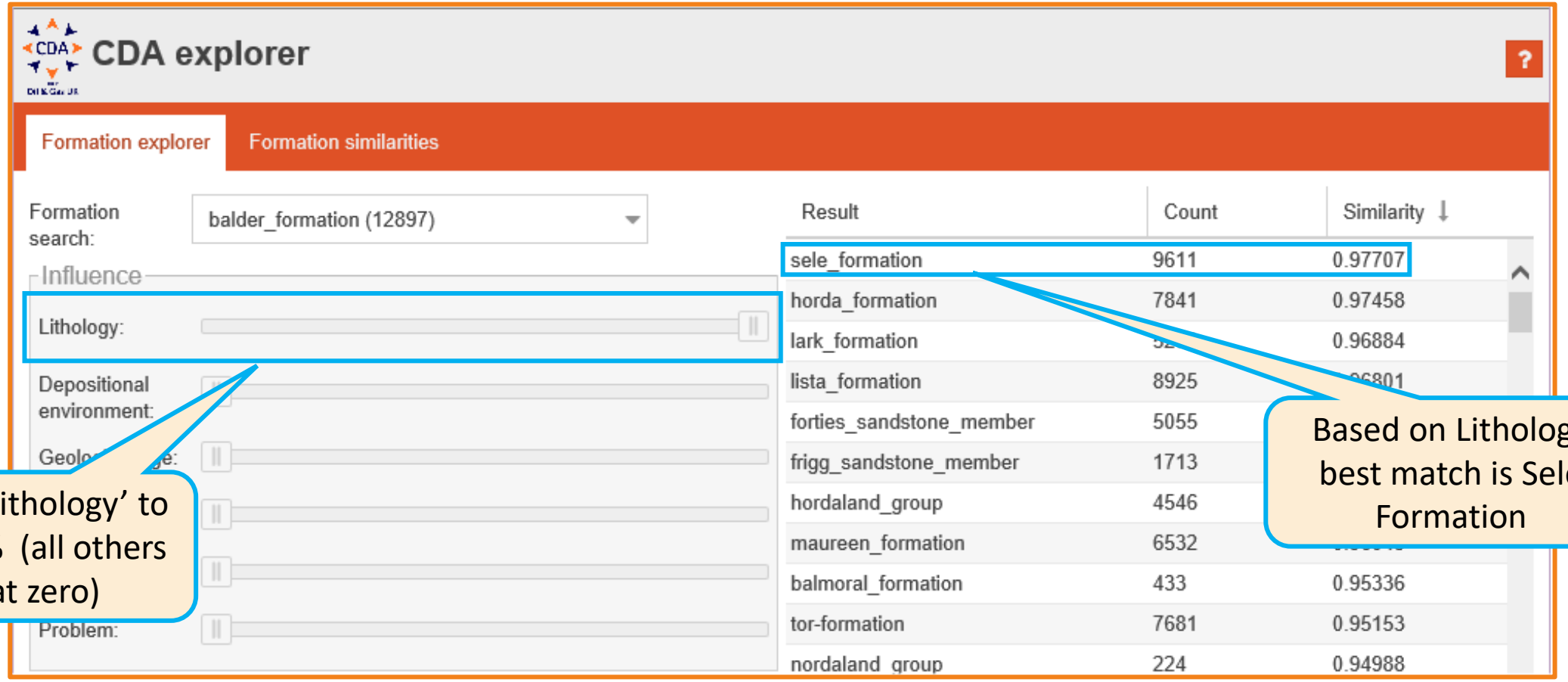
Prototype Interface



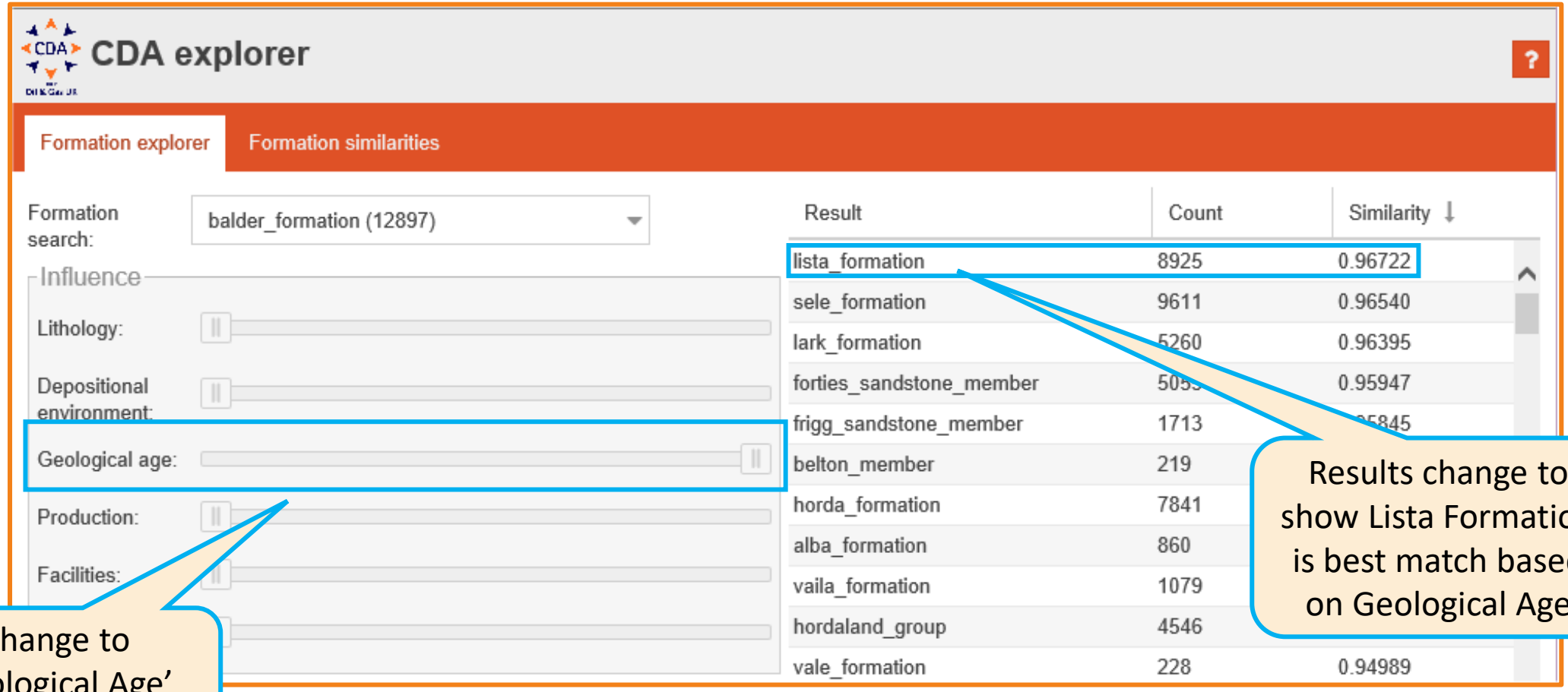
1. Find Formation Analogue – Enter Formation Name



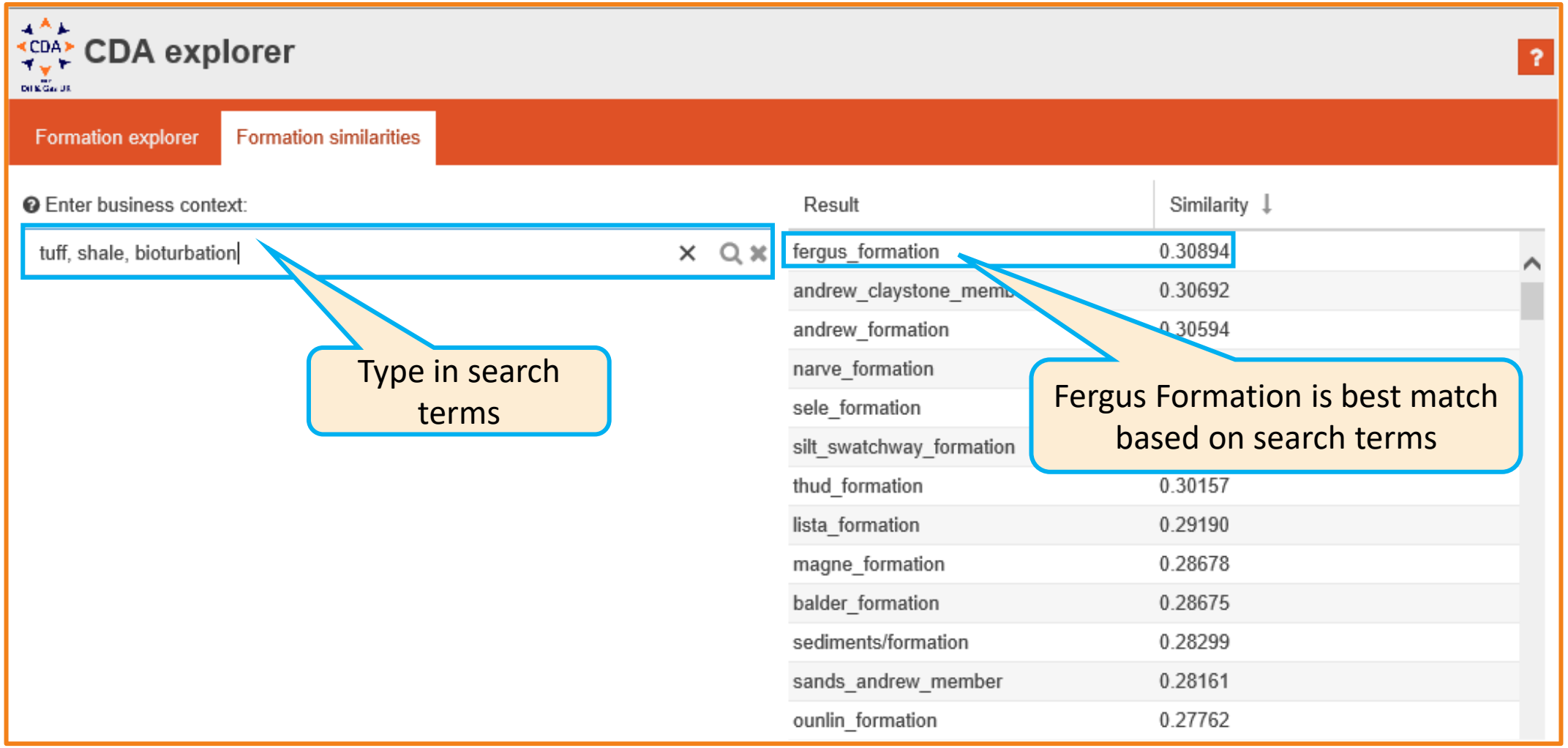
1. Find Formation Analogue – Choose ‘Lithology’



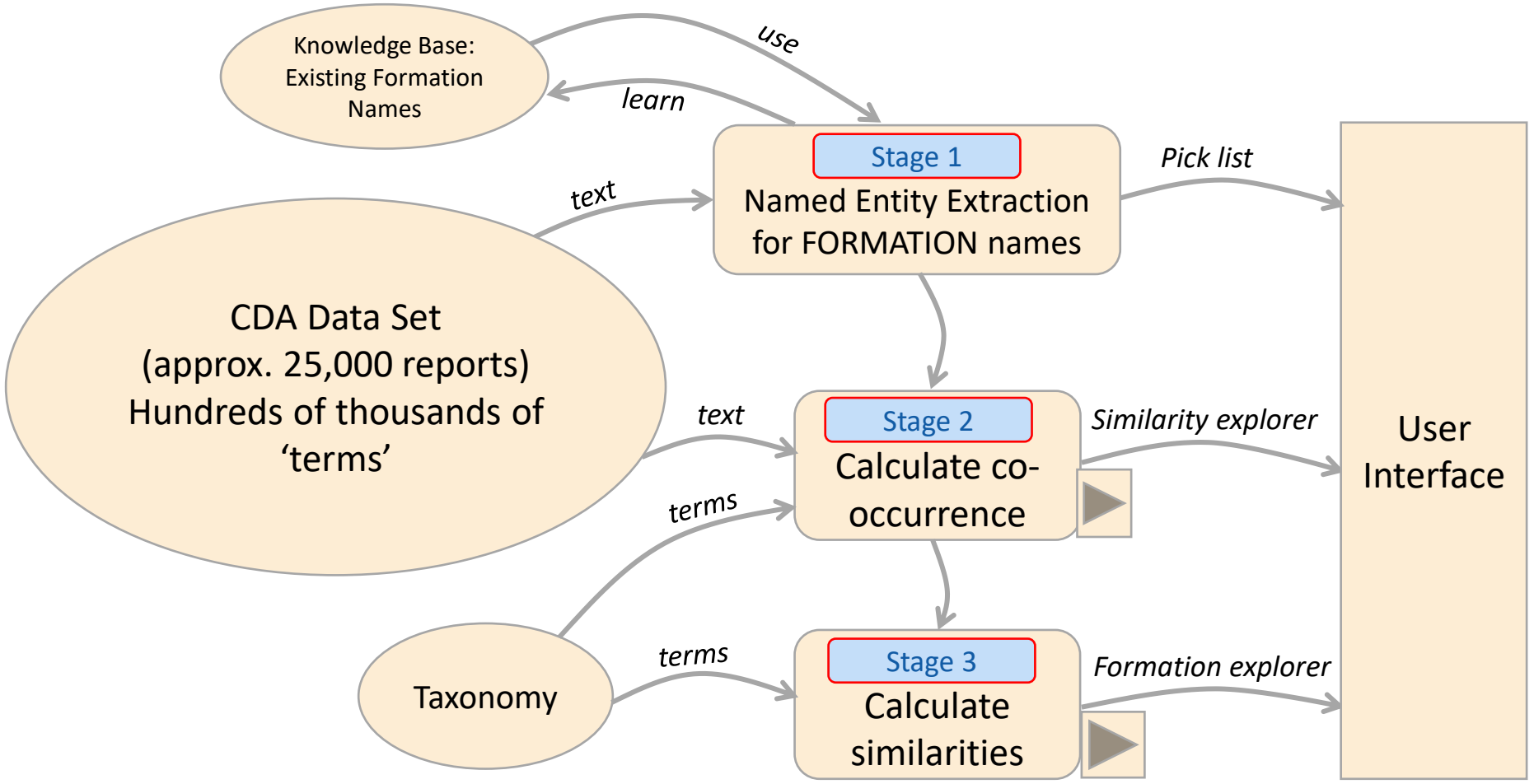
1. Find Formation Analogue – Choose ‘Geological Age’



2. Find Similar Formations based on Search Terms



Methodology



Methodology

- Stage 1:
  - Parsing of words from CDA Reports
    - Identify FORMATION names
    - Identify matches to known Reference terms (only partial for this demo)
    - Some STOP words (again, partial)
    - Some manual QC to remove errors and resolve 'aliases'
- Stage 2:
  - Run 2-layer neural net text co-occurrence analysis with dimension=300, word window=8
  - Output is matrix (vector):
    - Rows are terms (words)
    - Columns are dimensions (abstract) that represent a 300D descriptor, or 'co-occurrence fingerprint' for each term. Terms include the formation names and Taxonomy terms.
    - Similarity can be calculated by comparing 'closeness' of co-occurrence descriptors between terms (Cosine similarity)
  - For the Similarity explorer
    - Uses the raw vector analysis outputs to match the typed term(s) to the 'most similar' based on the 300D vector (Cosine) similarities
- Stage 3:
  - For the Formation analogues:
    - Calculate, for each formation, the similarity between the formation and EACH term in each slider group (e.g. all Lithology terms).
    - Result is a set of matrices where rows are Formation names and columns are terms appropriate to each slider (so there are 6 matrices in all). So, each Formation has 6 'similarity fingerprints', one for each slider set.
    - When sliders are moved the cell similarity values are scaled appropriately to calculate the overall similarity between the search Formation and all other Formations based on the slider terms.

Future Developments

- This is early experimental work to test some ideas. The eventual implementation of this, or similar systems, will be based around graph Db technology (which this prototype is not) allowing far more flexibility
- The parsing based on ‘known’ values from the Taxonomy, will be improved
- Further leveraging of the hierarchical nature of the Taxonomies to influence similarity weights (teach the machine)
- Many other parsing concepts are not yet included (Part of speech, spelling etc.)
- Add natural language parsing to understand concepts like ‘not permeable’, ‘capped by salt’, ‘later than the Triassic’ etc.
- Add ‘learning’ functions – influence similarity weights and adding terms to the Taxonomy
- Incorporate numerical data

Summary

Looking for analogues is about searching for ‘similar things’. It has wide application across a range of analytical and search domains:

- Other types of more complex analogue matching like plays, prospects and fields.
- Finding similar patterns in operational environments such as equipment failure modes.
- Finding ‘non-similar’ or unusual related things – relevant information the user didn’t expect.
- Improved auto-classification of content by characterising the ‘container’ (document etc.) and matching to ‘reference’ containers (the Taxonomy) – to recognise key documents like well reports, regulatory approvals/notifications, field development plans etc.
- Machine learning – enhancement of knowledge bases by recognising extensions to existing knowledge. For both Assets (names of wells, fields, formations, facilities…) and contexts plus relationships.
- Linking Information Management to Information Exploitation.

END

Questions?

Dave Camden

Flare Solutions

PETEX, 17-November, 2016

SPARES and Linked Slides

- These slides provide further technical information for further discussions

Term Vectors (300-Dimensions)

Technology is a 2-layer neural network, where inputs are the ‘terms’ in the analysed text.

Output, for each term, is a characterisation by a 300d parameter set - or ‘vector fingerprint’.

		Dimensions									
		d001	d002	d003	d004	d005	d006	d007	d008	.....	d300
Terms	Alness	0.0017	0.0043	0.21	0.51	0.093	0.112	0.00004	0.00311	.....	0.000311
	anhydride	0.0003	0.00092	0.038	0.092	0.091	0.0233	0.00345	0.0022	.....	0.004116
	anisotropy	0.0174	0.00411	0.4119	0.0233	0.0067	0.00191	0.002004	0.0902		0.00305
	annular	0.0022	0.00563	0.0322	0.0056	0.0611	0.1558	0.0204	0.03001	.....	0.01889
	anomalous	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	anthracene	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	anticline	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Aptian	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	aquifer	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	archie	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Asgard	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	average	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Beryl	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Berriasian	0.002	0.015	0.193	0.0717	0.0933	0.312	0.00492	0.01975		0.09226
	bioturbate	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Blue Lias	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	borehole_collapse	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	carrack	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	cretaceous	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	decline	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	decommission	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	sele	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	werraanhydrit	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc

Key

Asgard — Formation

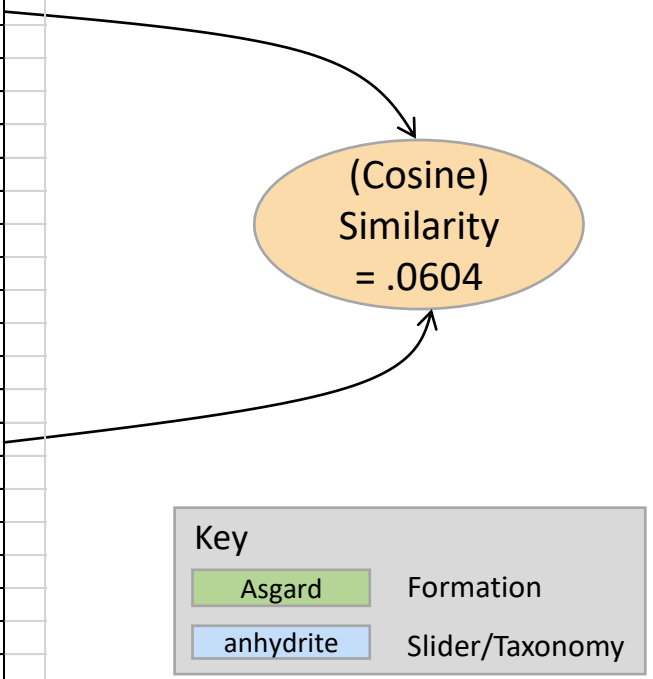
anhydrite — Slider/Taxonomy

Term Vectors (300-Dimensions)

For all of the ‘Formation’ terms, we calculate a similarity with all Taxonomy-based terms

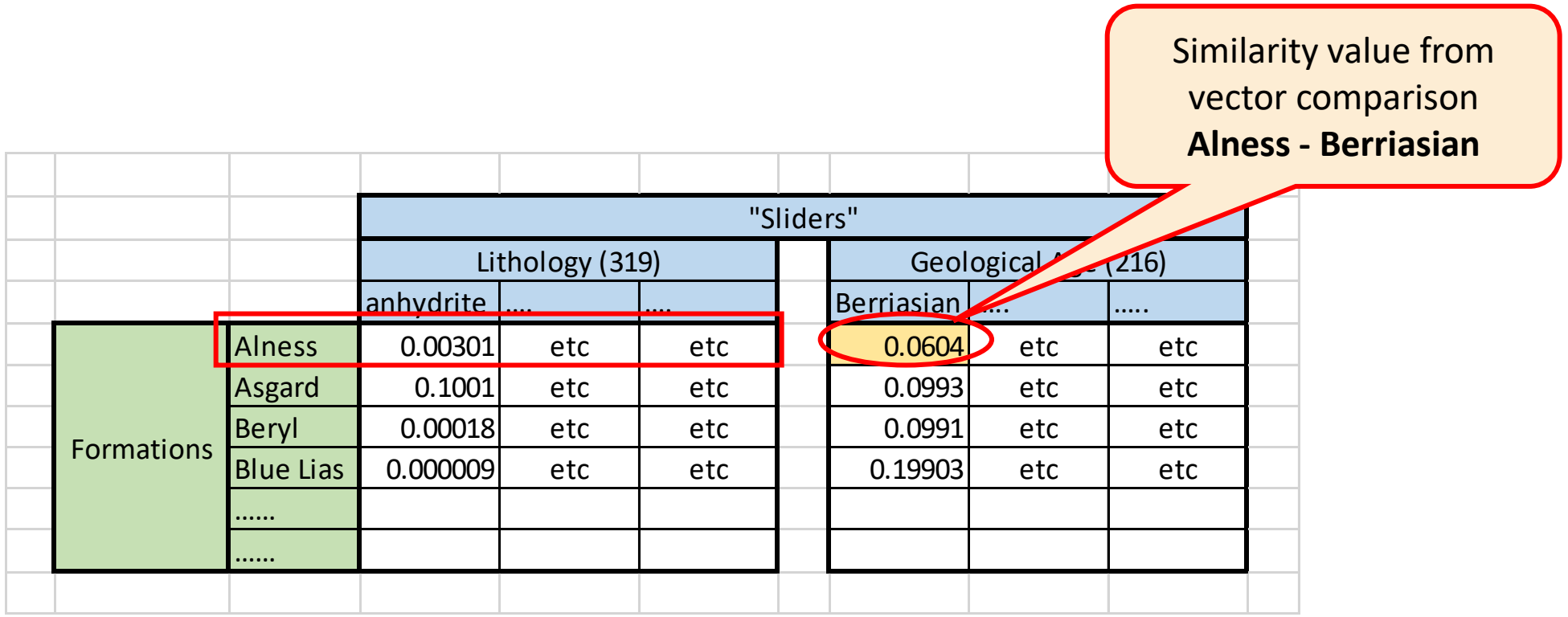
Terms that have similar vector fingerprints are ‘closer’ together (more similar) based on similarity of surrounding terms

		Dimensions									
		d001	d002	d003	d004	d005	d006	d007	d008	.....	d300
Terms	Alness	0.0017	0.0043	0.21	0.51	0.093	0.112	0.00004	0.00311	.....	0.000311
	anhydride	0.0003	0.00092	0.038	0.092	0.091	0.0233	0.00345	0.0022	.....	0.004116
	anisotropy	0.0174	0.00411	0.4119	0.0233	0.0067	0.00191	0.002004	0.0902		0.00305
	annular	0.0022	0.00563	0.0322	0.0056	0.0611	0.1558	0.0204	0.03001	.....	0.01889
	anomalous	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	anthracene	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	anticline	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Aptian	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	aquifer	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	archie	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Asgard	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	average	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Beryl	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Berriasian	0.002	0.015	0.193	0.0717	0.0933	0.312	0.00492	0.01975		0.09226
	bioturbate	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	Blue Lias	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	borehole_collapse	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	carrack	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	cretaceous	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	decline	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	decommission	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	sele	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc
	werraanhydrit	etc	etc	etc	etc	etc	etc	etc	etc	.....	etc

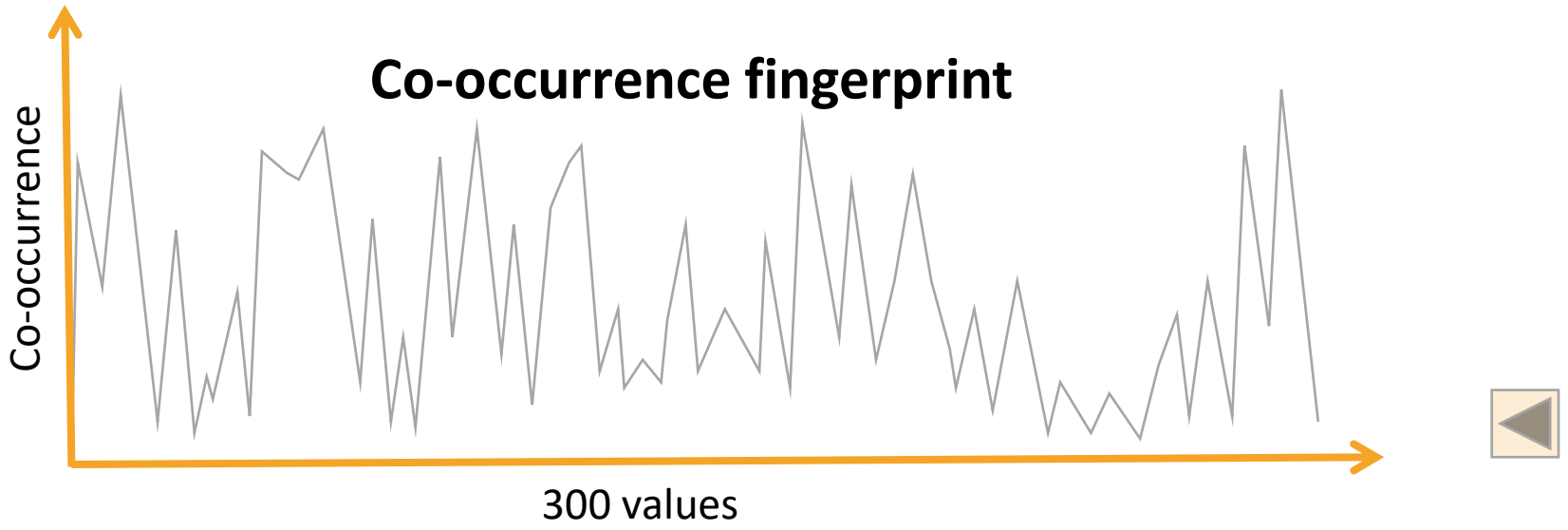


This is the Distributional Hypothesis – words that occur together share some kind of semantic meaning

Term Similarities



Every Term has a Co-occurrence Fingerprint



Similarity – Each Formation has 6 Similarity ‘Fingerprints’