



RESERVE BANK OF AUSTRALIA

Transforming Survey Analysis

Tools for Central Banks

Nicholas Gray and Dominic Jones
2025 BIS IFC Workshop on Data Science in Central Banking
February 2025

Surveys

- Surveys often include free-text open-ended questions.
 - They're great for eliciting what's top of mind – or getting at unstructured thoughts or things the survey had missed.
- If the survey sample is small enough, reading responses is the best option.
- But usually, the sample is huge! NLP techniques can help build a qualitative understanding using quantitative techniques.

Surveys at central banks

- Consumer payments surveys
- Trust and credibility
- Banknote design
- Inflation expectations
- Business surveys
- Lending surveys

Framework

- Gradual supervision
 - Making use of expert knowledge where we have it...
 - ...providing guides where we don't.
- Cluster exploration
 - Providing richer descriptions of topics to help with comprehension
- Iterative refinement
 - Subject-matter experts in the loop.
- Concept matrices
 - Are there any complex interactions worth thinking about?

Gradual supervision

- Usually, SMEs start knowing not much (but not zero) and gradually learn more about survey responses.
- Our old model – just run a topic model – didn't reflect this growing understanding of the data.
- Luckily, techniques now exist that cover the whole gamut of unsupervised to fully supervised – and allow us to incorporate growing expert knowledge!



Data scientists



Subject matter experts

Unsupervised

- Input: not much
- Easy
- Hard to interpret sometimes

Weakly supervised

- Input: anything you can think of
- Cheap
- Very good quality

Fully supervised

- Input: labelled examples
- Costly + high expert effort often required
- High quality



Unsupervised techniques

- Descriptive statistics: wordclouds, TF-IDF scores, phrasal segmentation algorithms.
- (Unsupervised) Topic modelling
 - BERTopic
- Information retrieval
 - BM25, pretrained sentence transformers.
 - Great for incorporating priors!

Weak supervision

- When SMEs start building a picture it helps to iterate fast.
 - Full supervision is expensive and slows down the loop.
 - Trying candidate models should be cheap.
- Guided topic models
 - BERTopic (again)
 - CatE
- Prompt-guided embeddings
 - Using natural language to shape the embedding space.
- Weakly-supervised classifiers
 - Fast to build!

Full supervision

- When SMEs have a strong understanding of the data – or want to be consistent over multiple surveys – it pays to train a fully supervised classifier.
- We often use **taxonomy classification** and train cross-encoders or text matching networks.
- It's useful to think about how to communicate these models and to track performance over time.

Recommendations

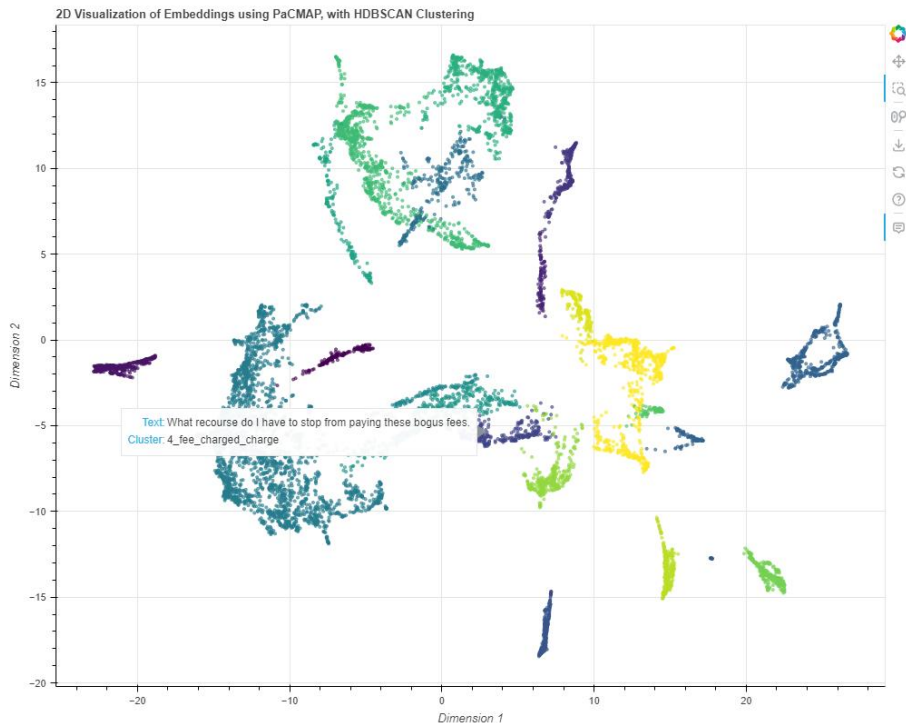
- Work closely with your SMEs.
 - Communicate the nature of what you're doing and be honest about limitations.
 - When eliciting guidance be clear about what format you want it in. Consider making a form for it!
- Reporting matters.
 - Do spend some time making reports easily digestible.
- When should we move up the chain? Questions to ask SMEs:
 - “Is this worth more investigation?”
 - “Should we cut this out of the analysis?”
 - “Are there any indicators of this sort of response?”
 - “Are there any sub-categories you're interested in?”
 - “Do you want to track this over time for the next survey?”

Cluster exploration

- Clusters, groups, classes don't need to be set in stone.
- Especially when understandings change!
- Agnostic of method, geometry of embedding space can be utilised.
- Providing interactivity and multiple perspectives important for reducing potential bias.

Cluster exploration

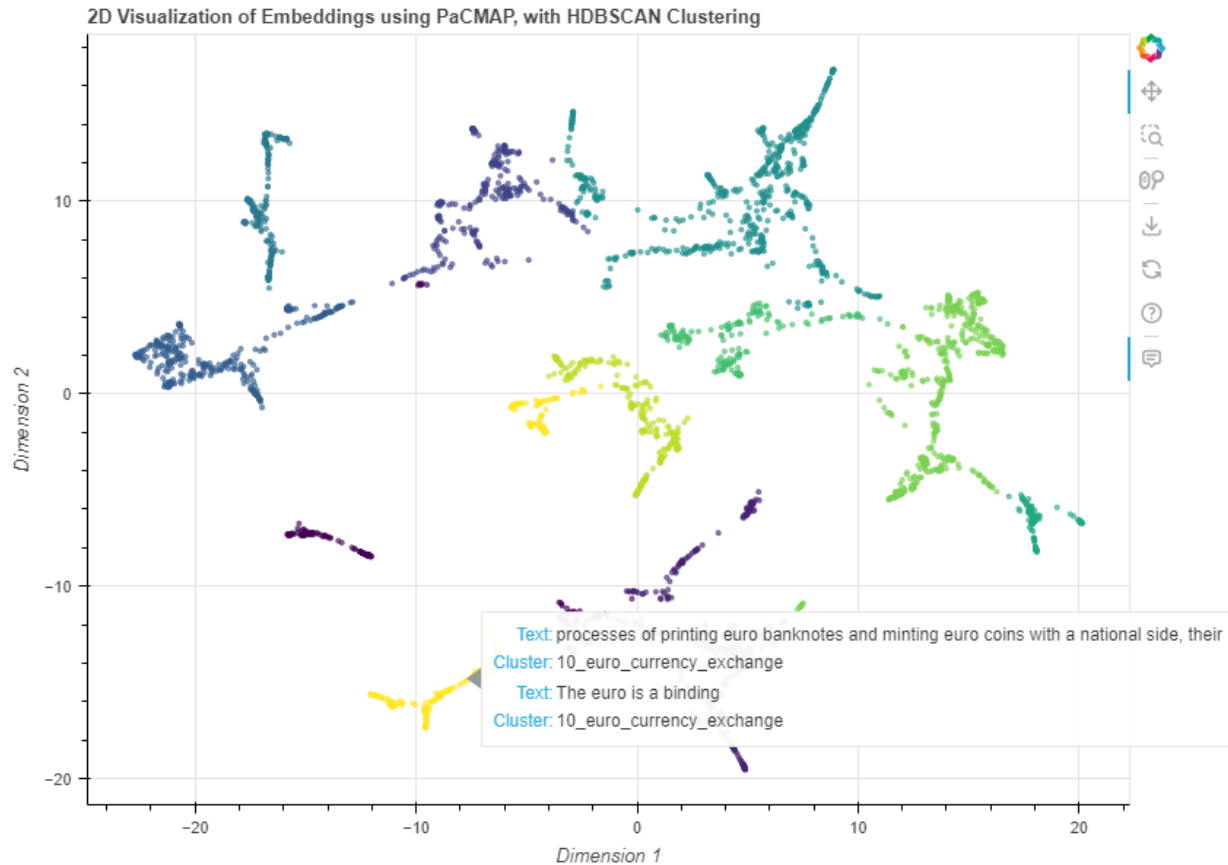
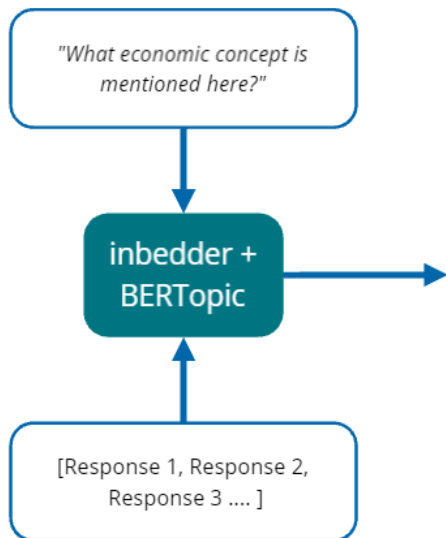
Interactivity:



Representative responses:

Closest to centroid of <i>7_credit_score_report</i> cluster:	
Distance from centroid	Reponses
0.856	I didn't know this was being reported to the credit company until after it was done, I am trying to buy a house and this is causing problems.
0.827	They reported my account to the credit bureaus as 30 days late.
0.813	I contacted my bank in regards to this issue as they are reporting every month to the credit reporting agency 's with remarks which is affecting my credit score in a negative way.
0.805	I told them that was not fair because I cleared up my credit and it is in good standing.
0.804	This company has pulled my credit multiple times for no reason as well.

Prompts, too!



LLM summaries

- All running on local LLMs for privacy reasons.

Q: Provide summaries of the most unusual document in cluster*

Response:

Scam Involving Timeshare Conversion:

- Sale of a timeshare condo as a campground membership.
- Down payment of \$10,000 without delivery of promised timeshare product.
- Denial of refund despite a contract addendum.
- Accusation of credit card fraud protection guarantee non-honoring by Chase.
- Loss of significant funds without receiving promised item or refund.
- Seeking governmental and legal assistance due to fraud.

Unauthorized Charges and Identity Theft:

- Unauthorized changes to bank account details and subsequent unauthorized transactions.
- Denial of dispute by bank despite evidence of fraudulent activity.
- Credit bureaus listing unauthorized transactions as late payments and charge-offs.
- Loss of funds due to identity theft and fraudulent charges.
- Seeking assistance to recover stolen money and address unauthorized account changes.

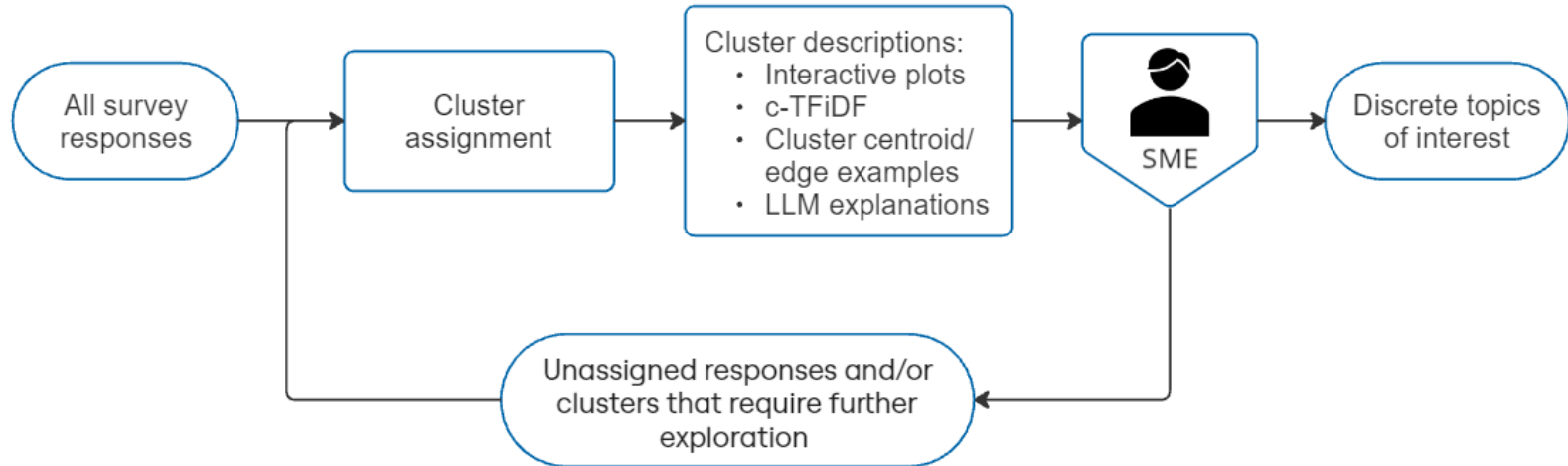
Fraudulent Transactions and Disputes:

- Merchant fraudulently disputing charges and misrepresenting to credit bureaus.
- Credit bureaus listing unauthorized transactions as legitimate.
- Credit card companies failing to resolve disputes and continuing to charge customers.
- Scenario involving a guest with a Capitol One card charged twice for a non-existent hotel stay.
- White Dial Stainless Steel watch purchase fraudulently disputed due to account access by an unauthorized third party.

* More prompt engineering behind the scenes

Iterative refinement

- Cluster assignments – no matter what algorithm you use – can be a mixed bag.
- More problematic: sometimes they're hard to explain!
- Having SMEs “in the loop” helps here.



Iterative refinement

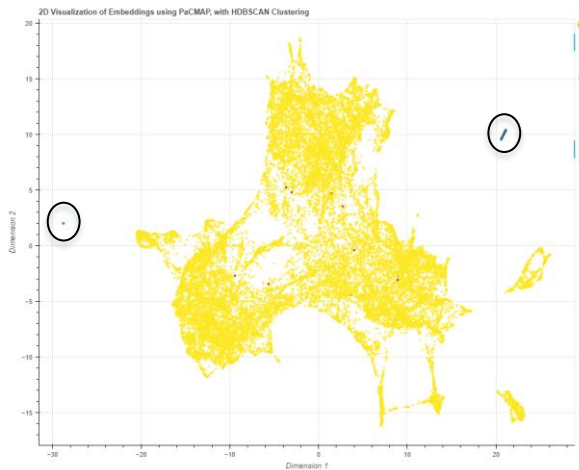
- Cache discrete clusters (circled)
- Rerun clustering on large central cluster



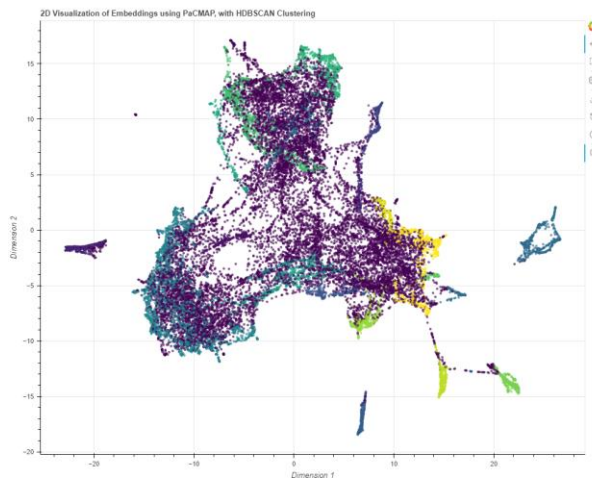
- Remove noise cluster (for downstream re-clustering)
- Focus in on clustered response



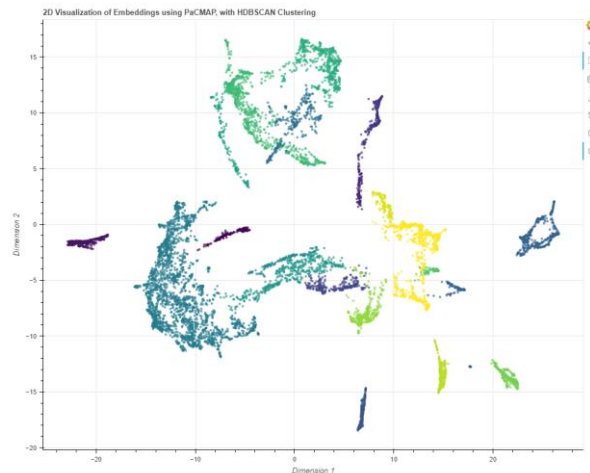
Run analysis on 20 clusters (plus 2 cached clusters from first iteration)



sentences = 23049
number of clusters = 3
noise cluster count = 12



sentences = 22882
number of clusters = 20
noise cluster count = 13818



sentences = 9064
number of clusters = 20
noise cluster count = 0

Recommendations

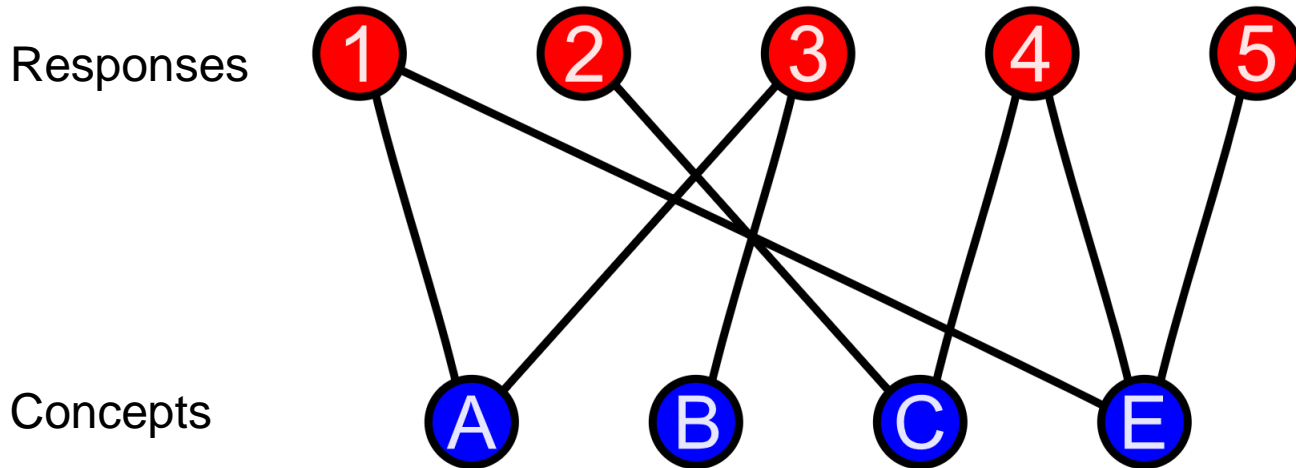
- Sit down with your SMEs and explain all your charts.
- Let them see the data underlying any LLM summaries – the chain of trust is important!
- Automate the process where possible – tight feedback loops are good.
- Feed the refined clusters back into the gradual supervision pipeline.
 - Usually, iterative refinement takes place somewhere between the un- and weakly-supervised steps.

Concepts

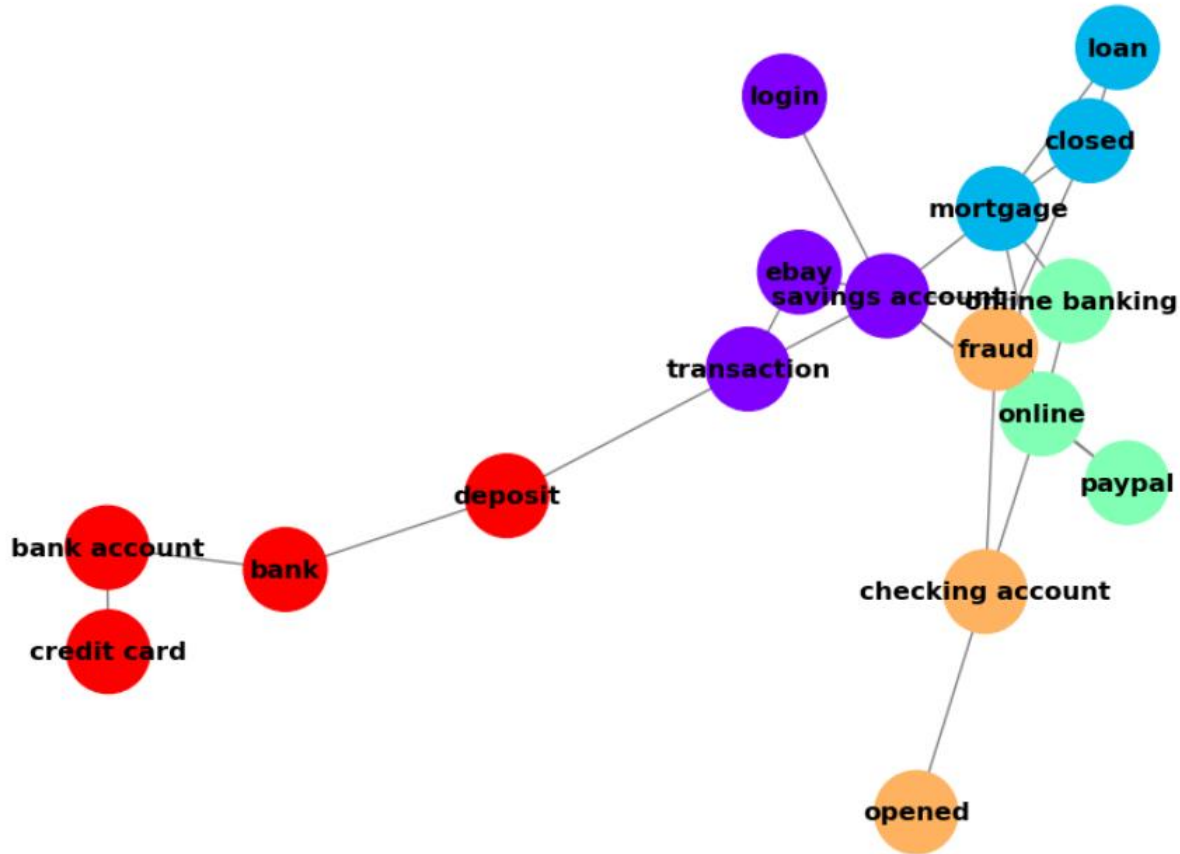
- When the output of gradual supervision is a set of classifiers, we call the classifier output *concepts*.
- Reporting on concepts is straightforward: bar charts, line graphs, etc.
- But what about more complex interactions?

Concept co-occurrence

- Each text response is associated with a set of concepts. This forms a “bipartite graph”. We can “project” these into concept-concept matrices.



Concept co-occurrence networks



Interpreting concept co-occurrence networks

- An edge means that two concepts are related (i.e. many responses are tagged with both concepts).
- Clusters on this graph can reveal more complex relationships that bar charts or group-bys wouldn't!
- If we see tight clusters (e.g. cliques) it's a sign that three separate concepts might be usefully thought of as a group.
 - In some cases!

Conclusions

- Guided supervision
 - Work closely with your SMEs
 - Use the whole gamut of techniques from un- to weakly- to fully-supervised.
- Cluster exploration
 - Important for reducing potential bias from basic topic summarization.
- Refinement
 - Build tight interactive loops with your SMEs to allow them to refine group assignments.
 - Use LLMs to reduce cognitive load on SMEs, but ensure they have access to the underlying data.
- Interactions
 - It's worth plotting a co-occurrence network if you've got a lot of concepts that might be interrelated.