

Andrew Marritt | CEO & Founder OrganizationView | People Analytics World 2023

Getting value from Text Analysis

Text Analytics in the age of GPT-4

Strategies for using Text Analytics in PA

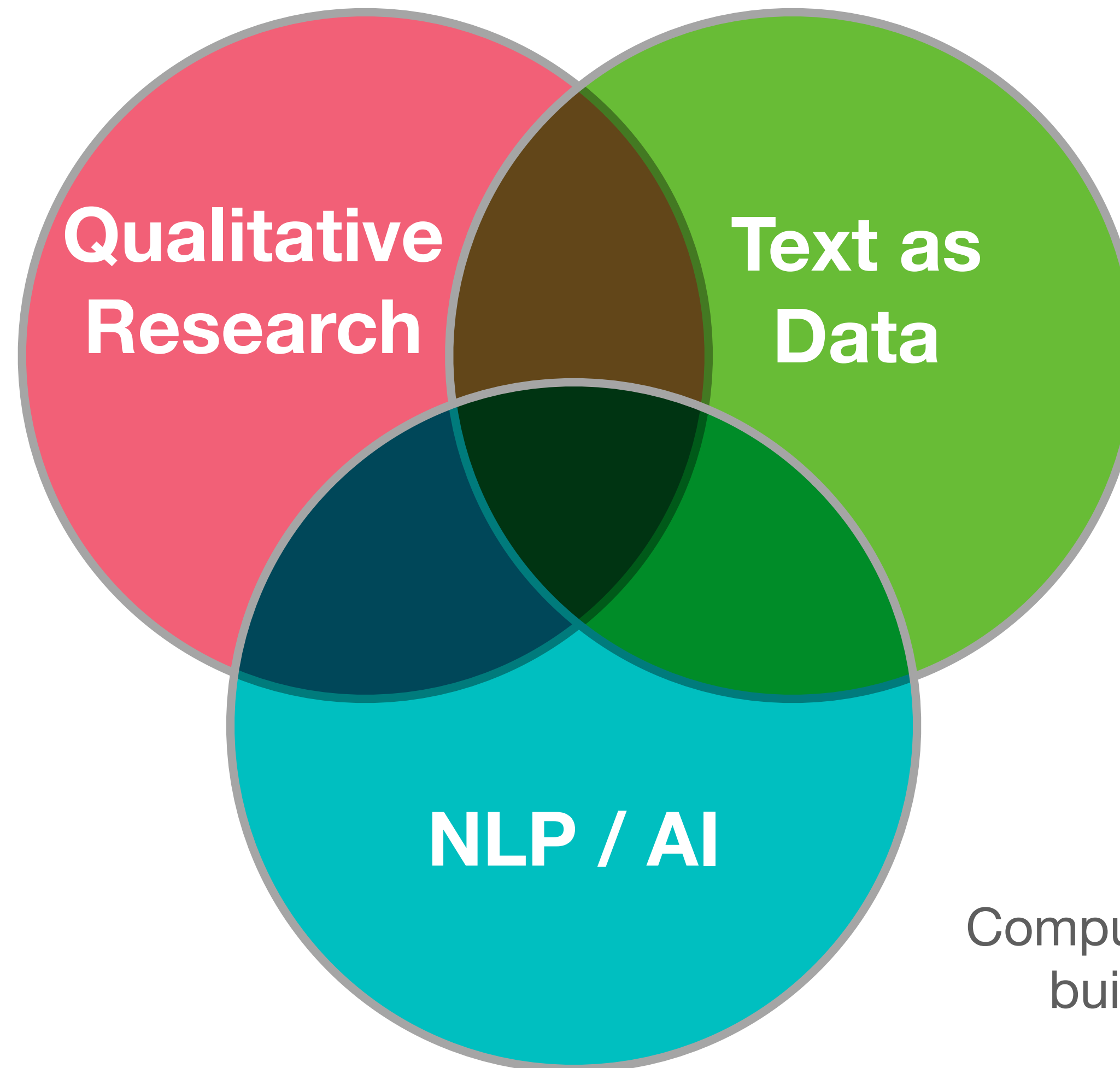
What we're going to cover

- Text Analytics - how to get the best quality results
- Realising value from your results
- BONUS: Using LLMs (GPT models) as part of the analysis process

The trilogy of successful text analysis

Don't reinvent the wheel or be attracted to shiny new toys

Your trying to do something
which has been done &
researched for a long time



Social scientists have
developed approaches to
answer similar questions
already

Computer scientists are rapidly
building better technical
approaches

What are you going to classify to?

Finding your categories

- Embedding models can help you find categories
 - **Themes** are mostly nouns
 - **Intents** are mostly verbs
- Old style models like LDA are easy to create, but hard to interpret
- Newer-style models like BERTTopic are rapidly outpacing traditional topic models
- There is really no harm just reading a selection!

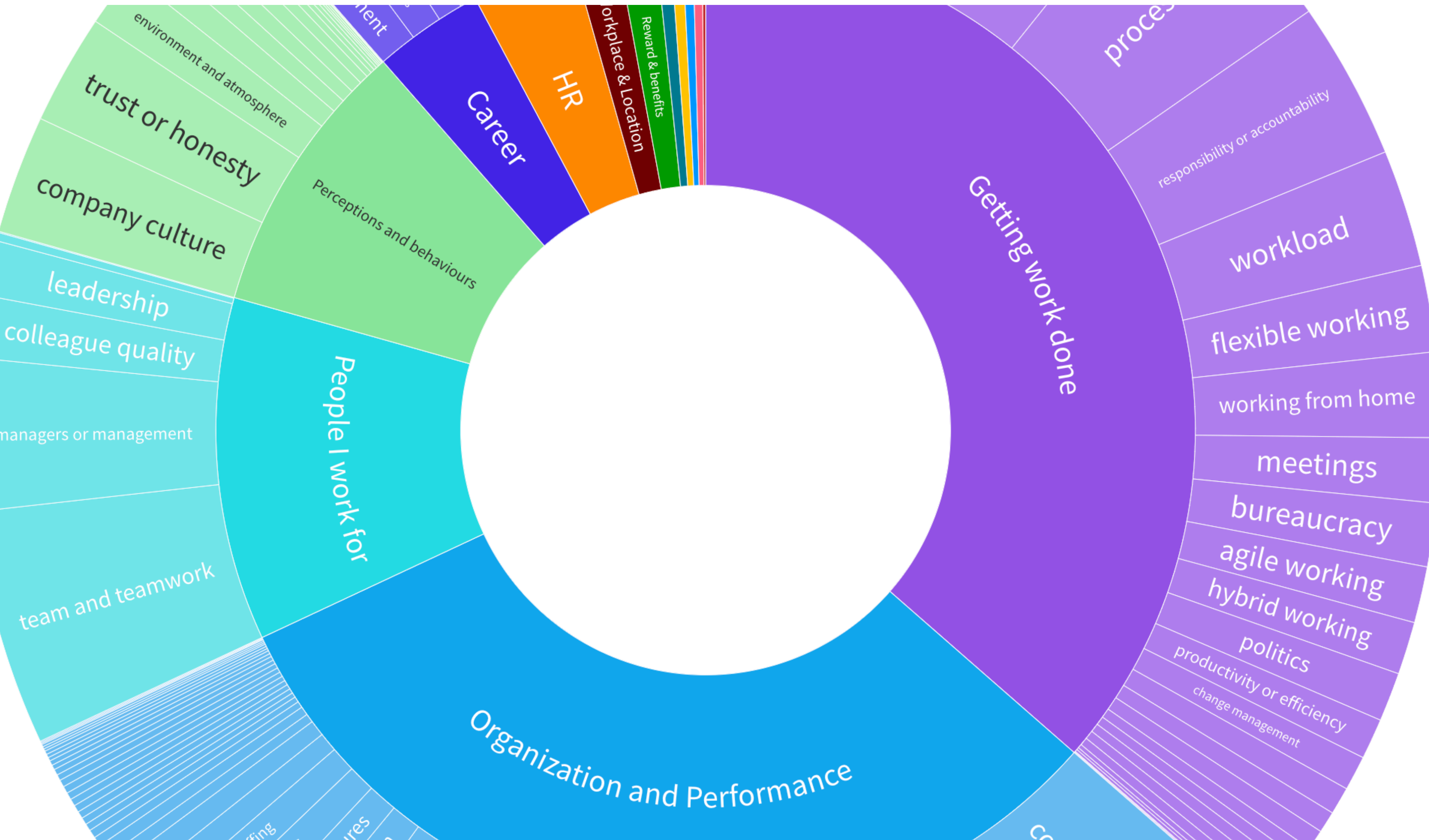
Classification models need examples

In text analysis the better examples you have the better your models are

- Need positive and negative examples
- Ideally your examples are close to how employees describe things
- Single words almost never work (ie you can't just search)
- You might need common spelling mistakes
- The objective is to get a large number of different examples (too many similar ones cause overfitting).

Success = a really strong training dataset

Size isn't everything



- You need to aim for a minimum of 100 examples of each category.
- Variety is probably more important than frequency
- Data imbalance can cause significant issues

Ambiguity is the enemy

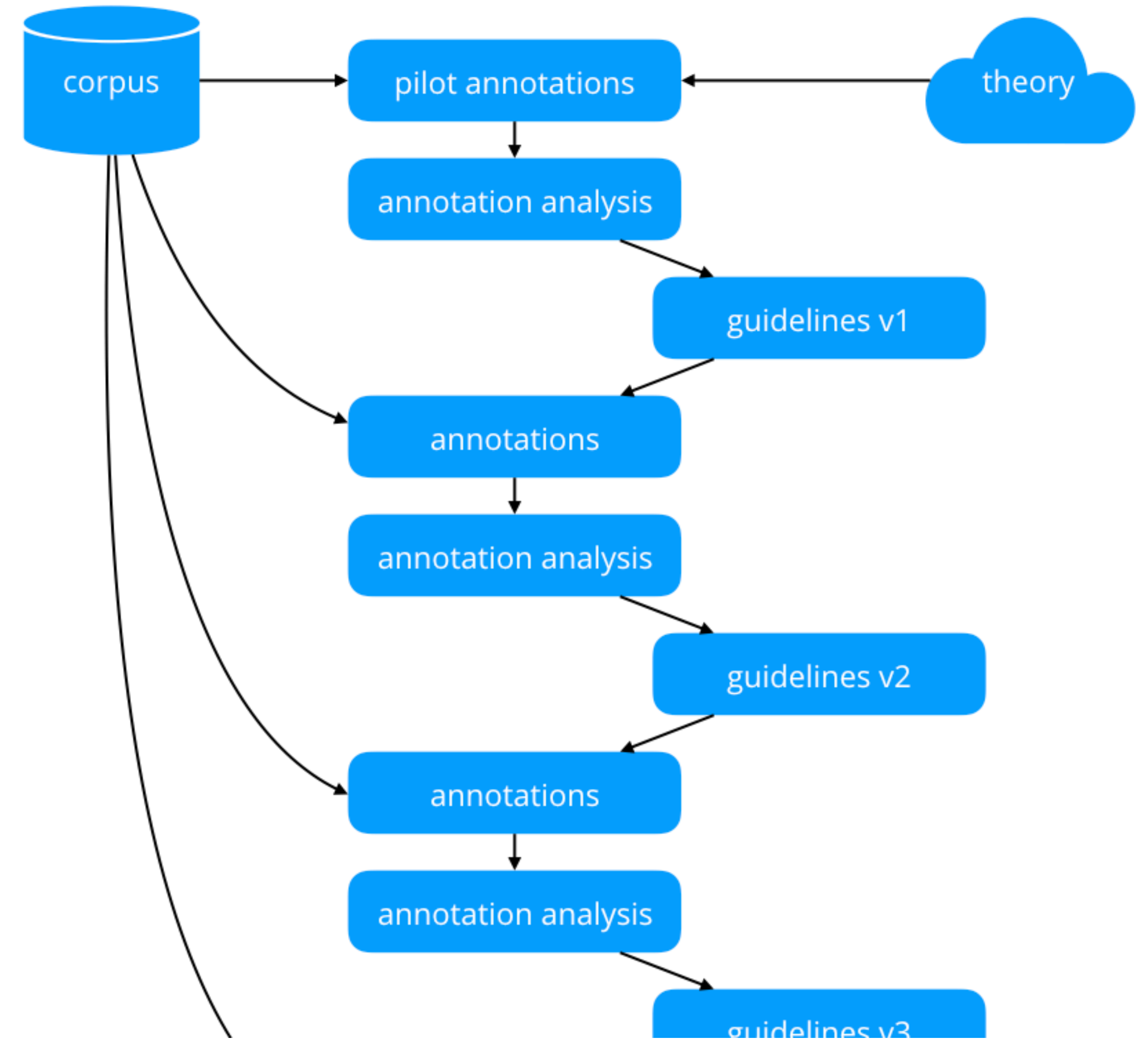
Most cases you learn by seeing them

- The key issue driving misclassification in an otherwise performant model tends to be ambiguity
 - Ambiguity in the training examples
 - Ambiguity in the answers to classify
- The context of the question / organization can resolve a lot of ambiguity

Building a code book

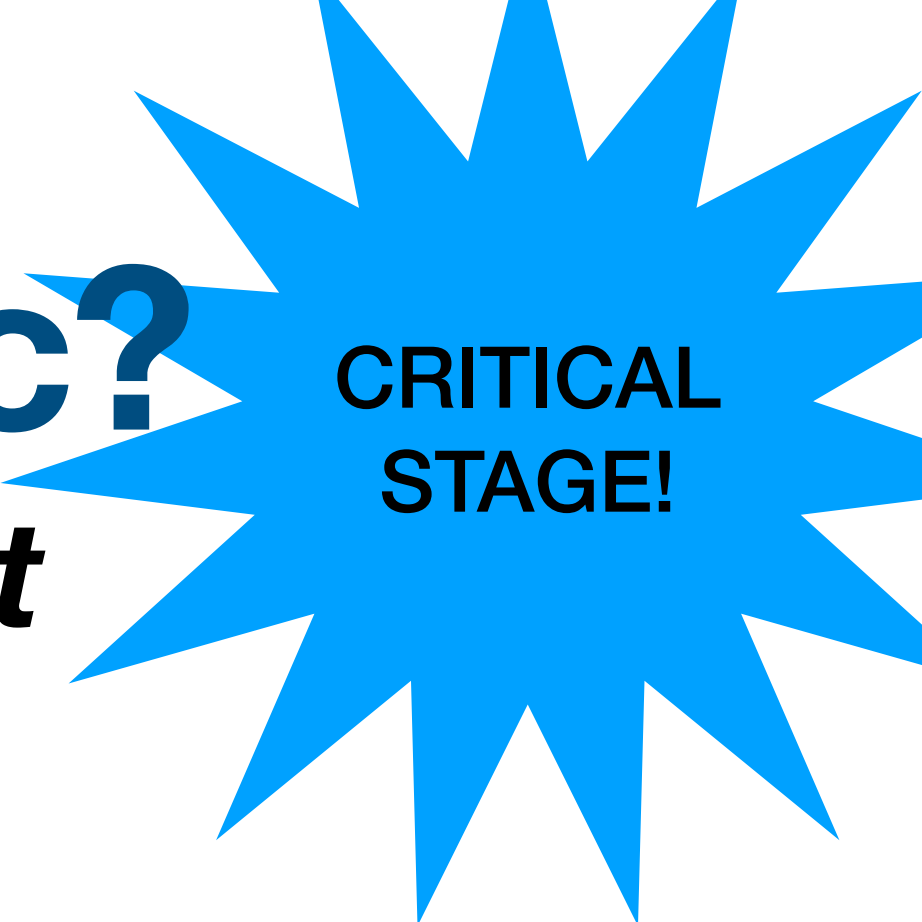
Consistency is all-important

- Annotation Guidelines are critical to ensure that your definitions are consistent
- Annotation Guidelines will likely be a living document
- You will need a process for updating / issue resolution
- Smart Annotation Guidelines can be used with GPT models - what was designed for human consumption can guide GPT models



How are people mentioning each topic?

To make decisions you need to understand *why* not *what*



CRITICAL
STAGE!

- In grounded approaches the analyst identifies hundreds of different expressions / ideas
- They then cluster together similar topics into themes
- In an ML approach we start with the topics (e.g. Training / Salary)
- We then identify different ways people mention each topic
- We can use various statistical techniques to group into themes

Topic frequency isn't that important

What is linked to things you really care about

- Techniques like Relative Importance, Feature Importance... can guide you to what is linked to the scale questions / other things you care about
- Text itself can indicate causal relations (*I feel X because Y happened to me*)
- Comparing importance across multiple questions / metrics can be important

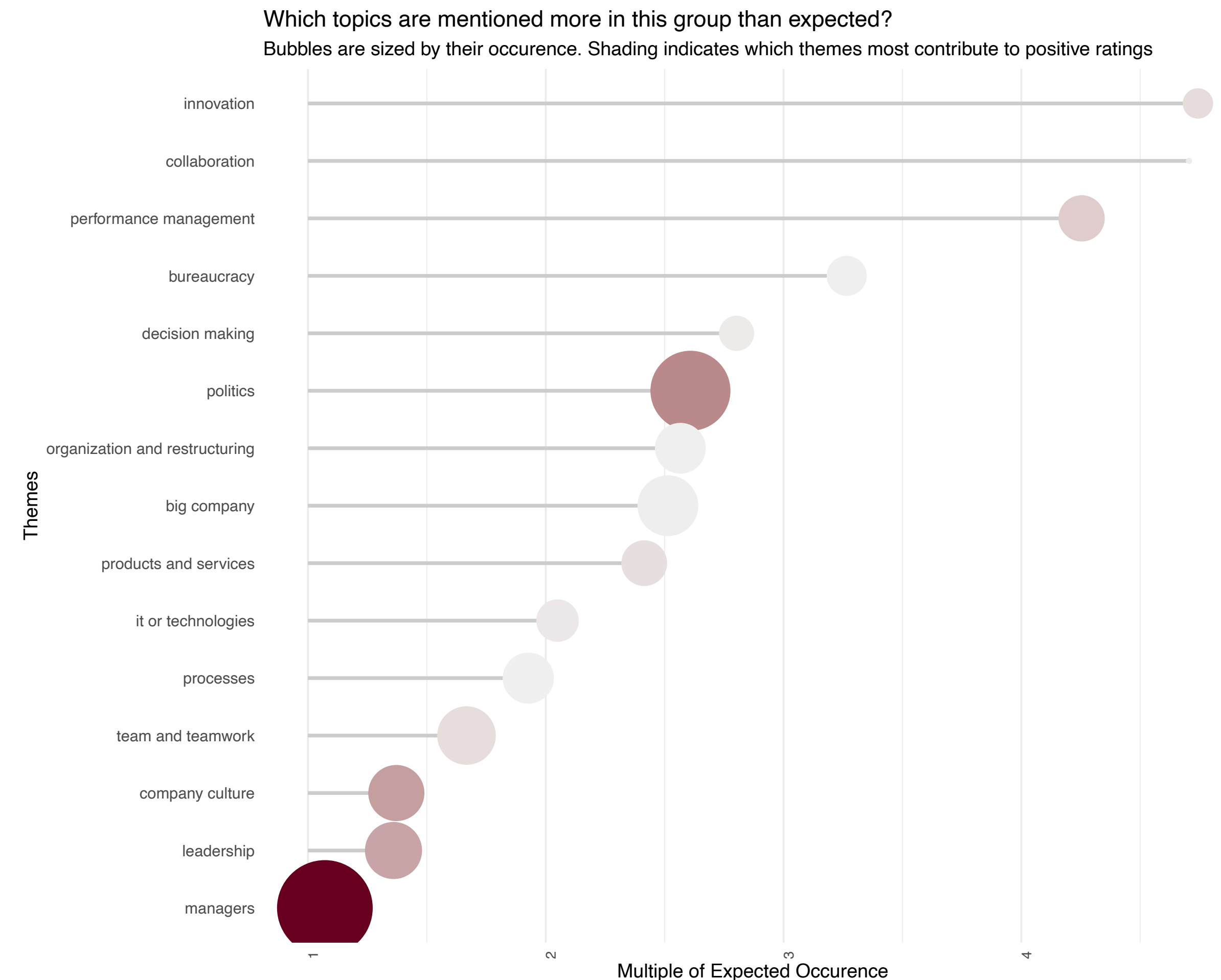
Qualitative Key Driver Analysis

Where are there pockets of similar topics?

Themes are rarely distributed equally

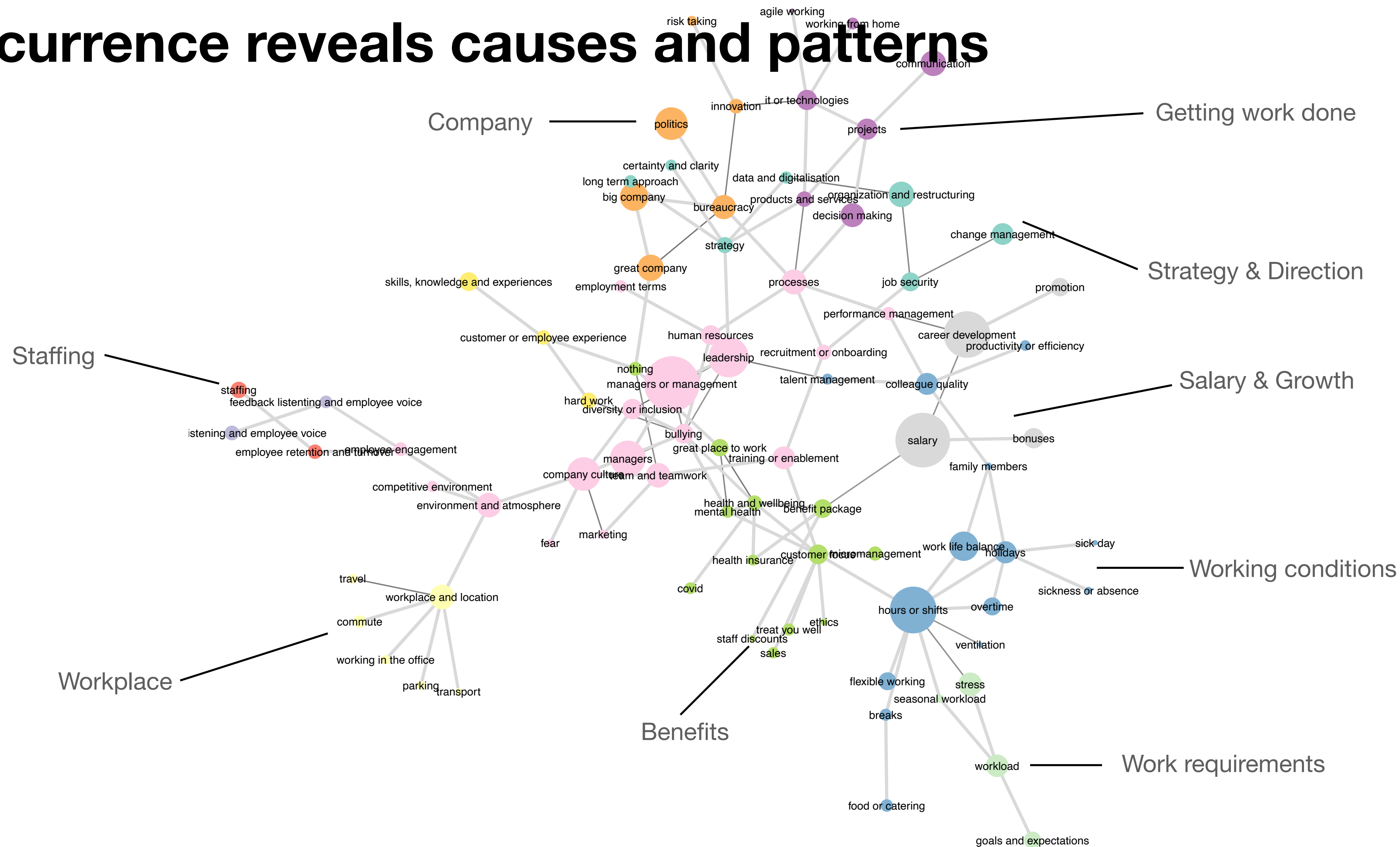
Look both at:

- Where themes are most likely to be used
- Which themes are more likely to be mentioned in a particular group
- Multiples of expected is a good measure



Which combinations of topics are used together?

Theme co-occurrence reveals causes and patterns

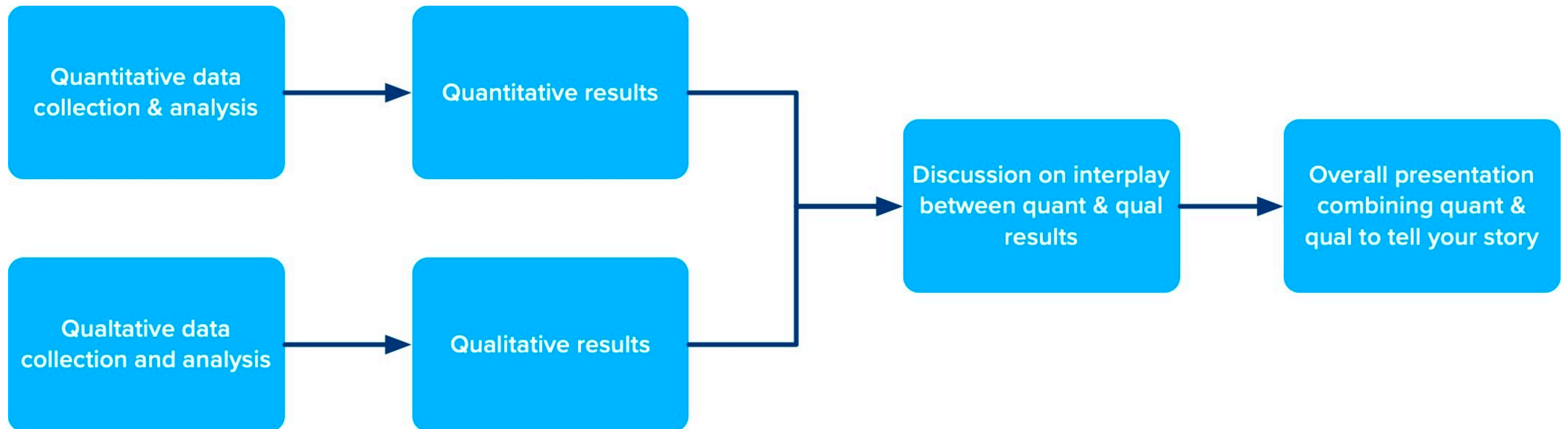


Mixed Methods Analysis

Making sense of very rich data

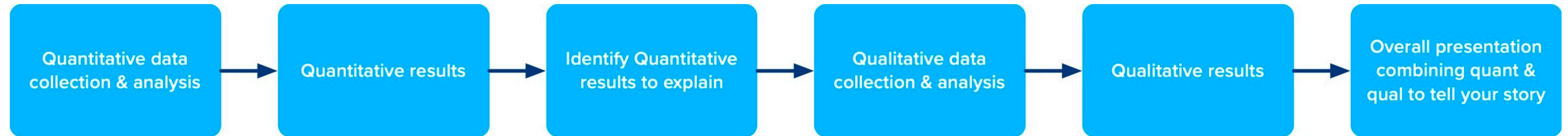
Method 1

Do both analyses in parallel



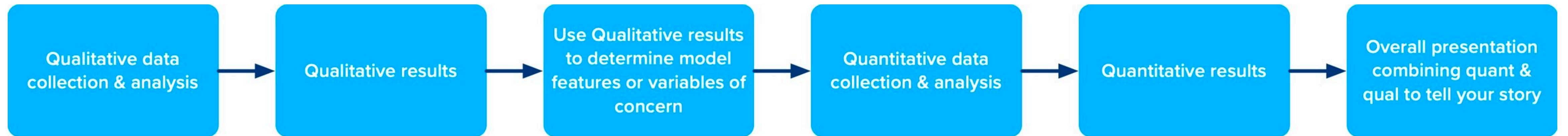
Method 2

Explanatory uses of text data



Method 3

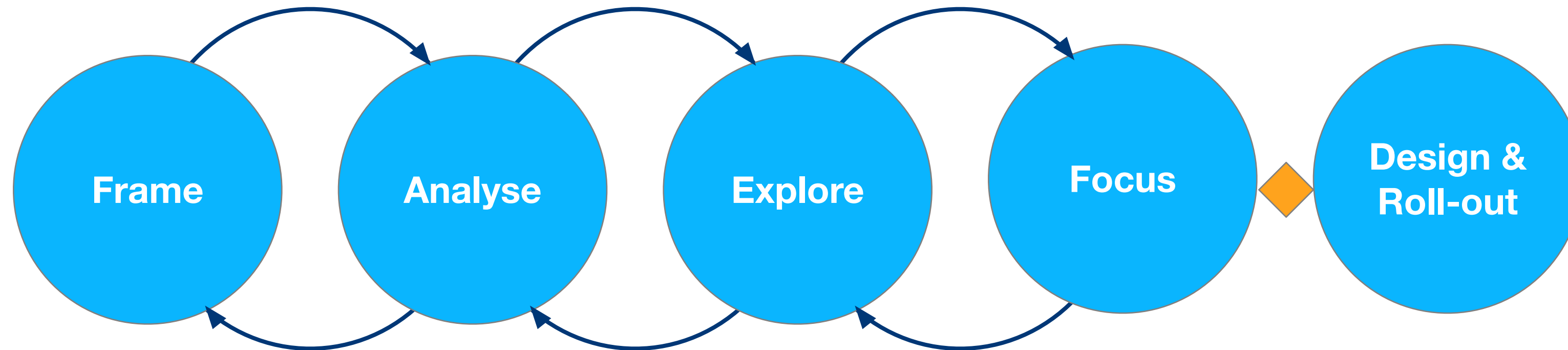
Exploratory use of text data



From data to action

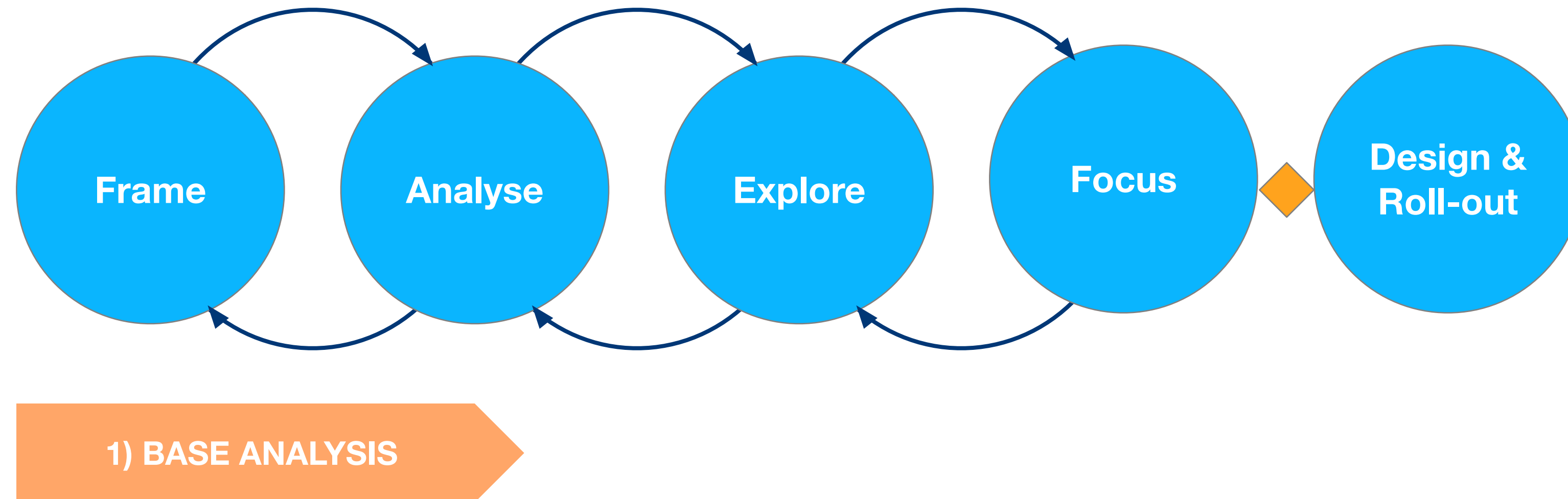
The hard work starts here

This is **TOTALLY** a domain-expertise effort



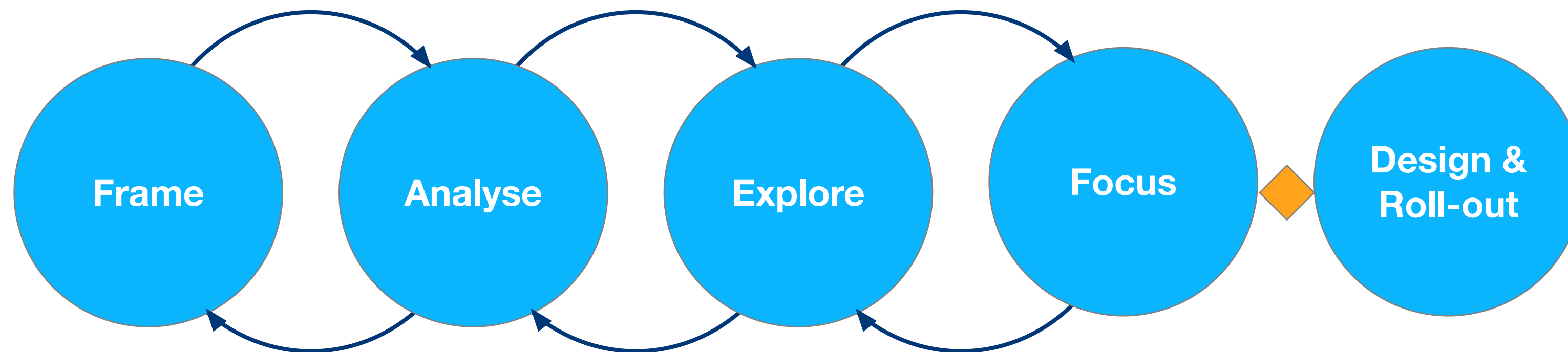
Base analysis

For our most mature clients



Making sense of the data

Typically between 1/2 and 1 day post analysis

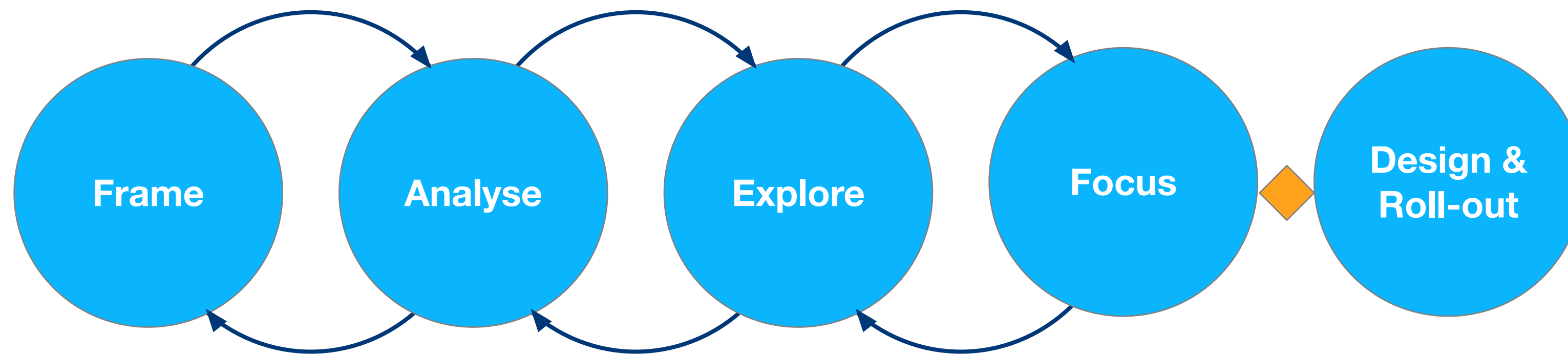


1) BASE ANALYSIS

2) BASE + GUIDED EXPLORATION

Developing recommendations & actions

Co-creating the action plan: 2-4 days post analysis



1) BASE ANALYSIS

2) BASE + GUIDED EXPLORATION

3) BASE + EXPLORE + FOCUS

Are GPT models the answer?
The shiny new toys

What are GPT Models?

There are *lots* of different models available

- “Next word generation” models - probabilistic approaches
- Typically you work with them via ‘Prompting’
- Size of the model is a big factor for performance
- ...Hence few can be run on a single machine
-So you’ll usually work with them via an API
-Which can be slow and expensive

Where are they useful for text analysis?

A key strength is zero- or few- shot learning

- “First Glance” results from the best GPT models are very good
- However, they are best when you condition (fine-tune) them
- OpenAI recommend 100 good examples per theme
- We find using your coding book in the prompt increases accuracy...
- ...but that obviously requires time creating the book
-and if you’ve done that you’ve probably labelled lots of data

Synthetic data

Data augmentation is a powerful use-case

- Creation of realistic examples of training data is a long-standing issue in ML
- GPT models are best when creating new text
- Accuracy isn't everything, so cheaper models can be useful
- Try and create examples in different tone-of-voice or even in slang
- Use your coding manual to guide the model to create example that are compliant with your definitions
- **HOWEVER: Using synthetic data means you import the GPT model biases**

Summarisation

GPT-4 seems to be the current king

- There are 2 types of summarisation
 - Extractive: finds sentences in the data that represent the data
 - Abstractive: rewrites the text to create a summary (GPT does this)
- GPT-4 significantly outperforms ChatGPT on summarisation and can summarise longer texts
- Summarisation of Summaries works well and is a good strategy to get shorter texts, or summarise texts longer than the token limit.