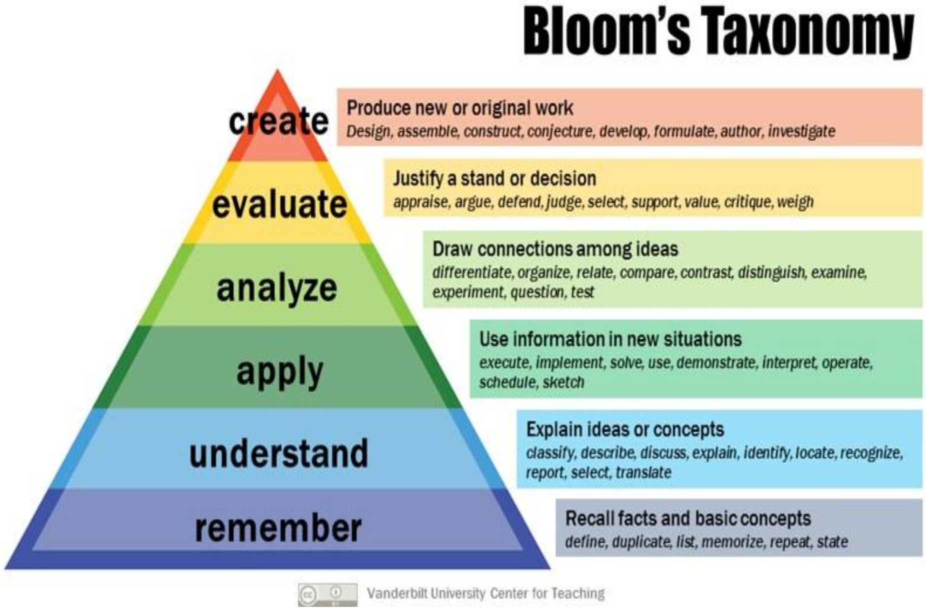
# Analysis of questions

## Post-final presentation

by JaMiMaKa group
Mikołaj Malec, Marceli Korbin, Kacper Grzymkowski, Jakub Fołtyn

# Experimental procedure

- Testing different approaches to topic clustering
  - LDA
  - Sentence embeddings
- Analysis of question complexity
  - LLM prompt engineering
  - DSI measure
  - Bloom's taxonomy
  - Question-words



Bloom's Taxonomy
create
Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate
evaluate
Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh
analyze
Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
apply
Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
understand
Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
remember
Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state
Vanderbilt University Center for Teaching

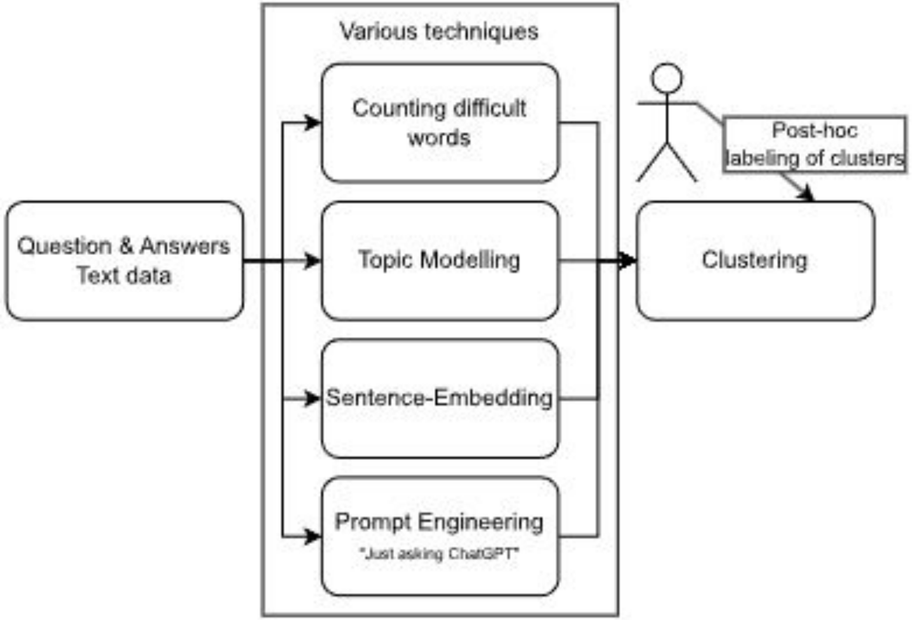
Image source: Vanderbilt University Center for Teaching

# Dataset

- The Stanford Question Answering Dataset
- Not the best choice, but:
  - Available
  - Decently sized
  - Decently clean
- Time constraints
  - We wanted to get something out of the gate

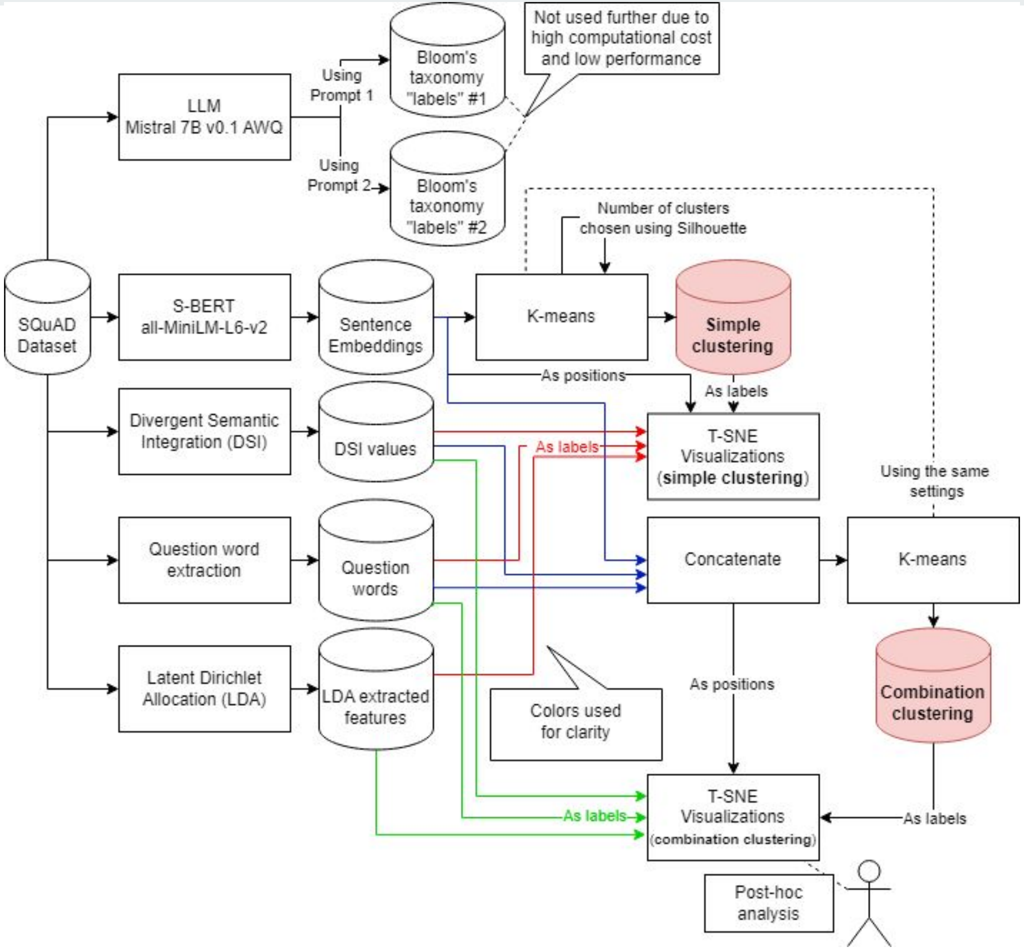
# Overview

- *OK, but what did we actually do?*
- This diagram got a bit more complicated
- While working on the project, it was clear to us…
- But not for anyone else
  - ○ The point of reviews



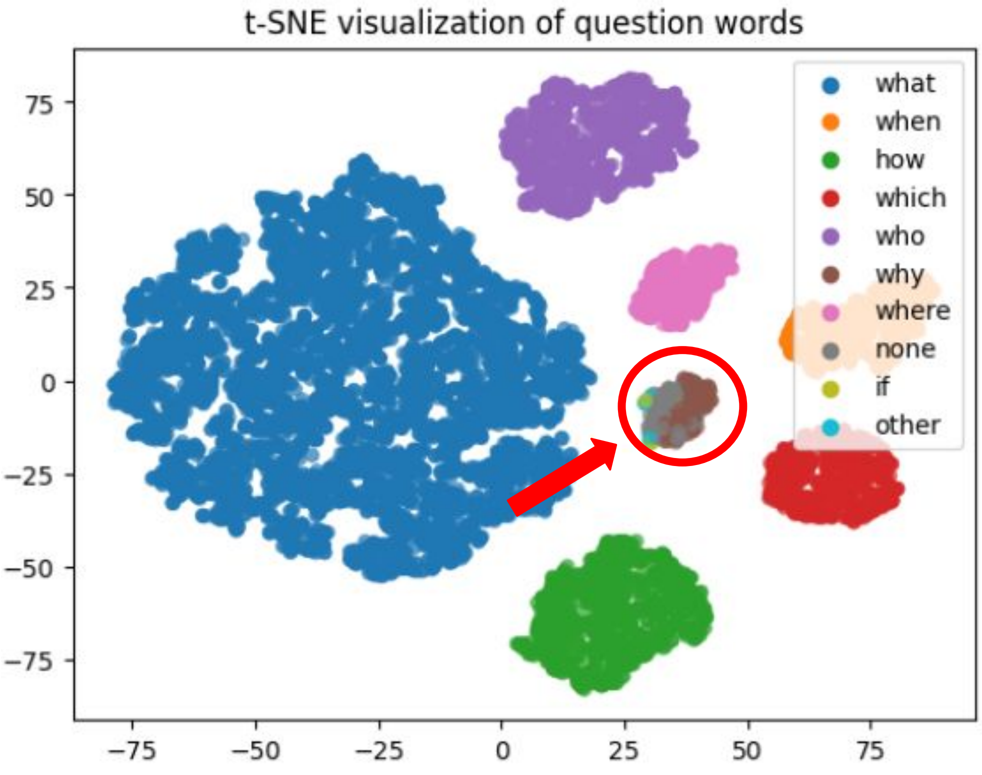
# New diagram

- Fair to say it was needed



# Our contribution

- Explored the data from the perspective of the problem
- We found what worked
- … and what didn't
- Found *something* interesting



# Lessons learned

- LLMs like Beyoncé

# Actual lessons learned

- LLMs aren’t always the best solution to a vague problem
  - Mostly a cost / effectiveness trade-off
  - Still worth exploring as an option
- Sentence embeddings roughly model the topic
- Modelling structure is much more difficult than modelling the topic
  - Especially when trying to model the structure based  on the topic…
- More careful selection and preparation of data
- Topic modeling is especially hard on short text data

# Technical lessons learned

- LLMs are large
  - Model quantized to 4 bits
  - Barely fit in memory
- The Python k-modes package doesn’t work on large data
- Preprocessing on question data is not as straightforward
  - Certain stop-words are significant for the meaning of the question

# Future works

- Better suited datasets
  - Mix of complex and simple  questions
  - Some labels – semi-supervised learning
- Method refinements
  - More statistically sound clustering
  - Analysis using multi-dimensional scaling instead of t-SNE
- More rigorously study structure / topic relationships

# Thank you for your attention

Any questions?