VANDERBILT UNIVERSITY
Jean *and* Alexander Heard Libraries

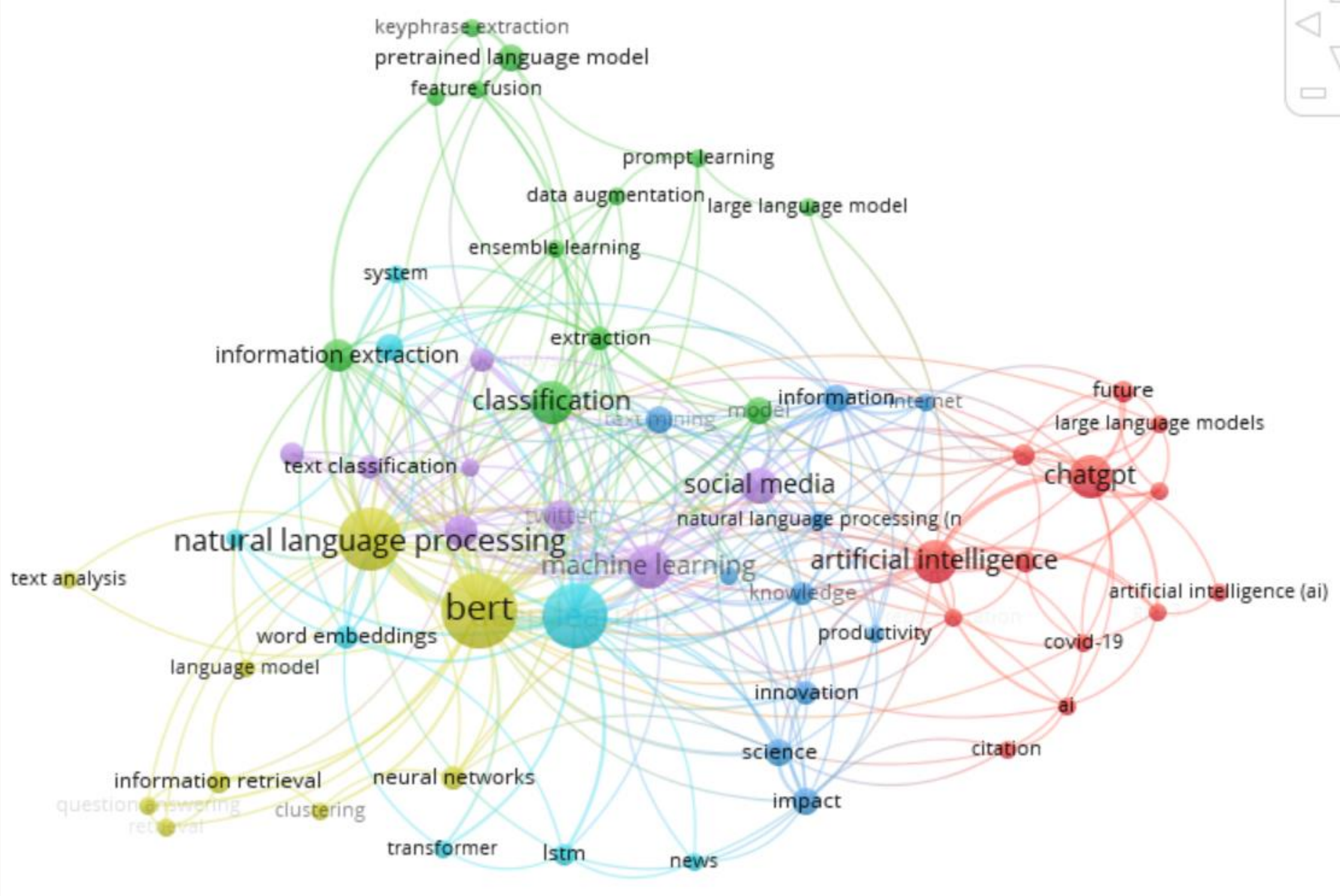
Topic Modeling: An Exploration

Text Analysis Community of Practice Presentation

Shenmeng Xu
Librarian for Scholarly Communications
Digital Lab
Vanderbilt Libraries

Nov 2023

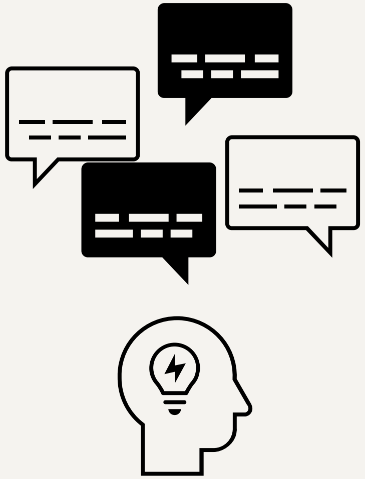
Understanding Topics…



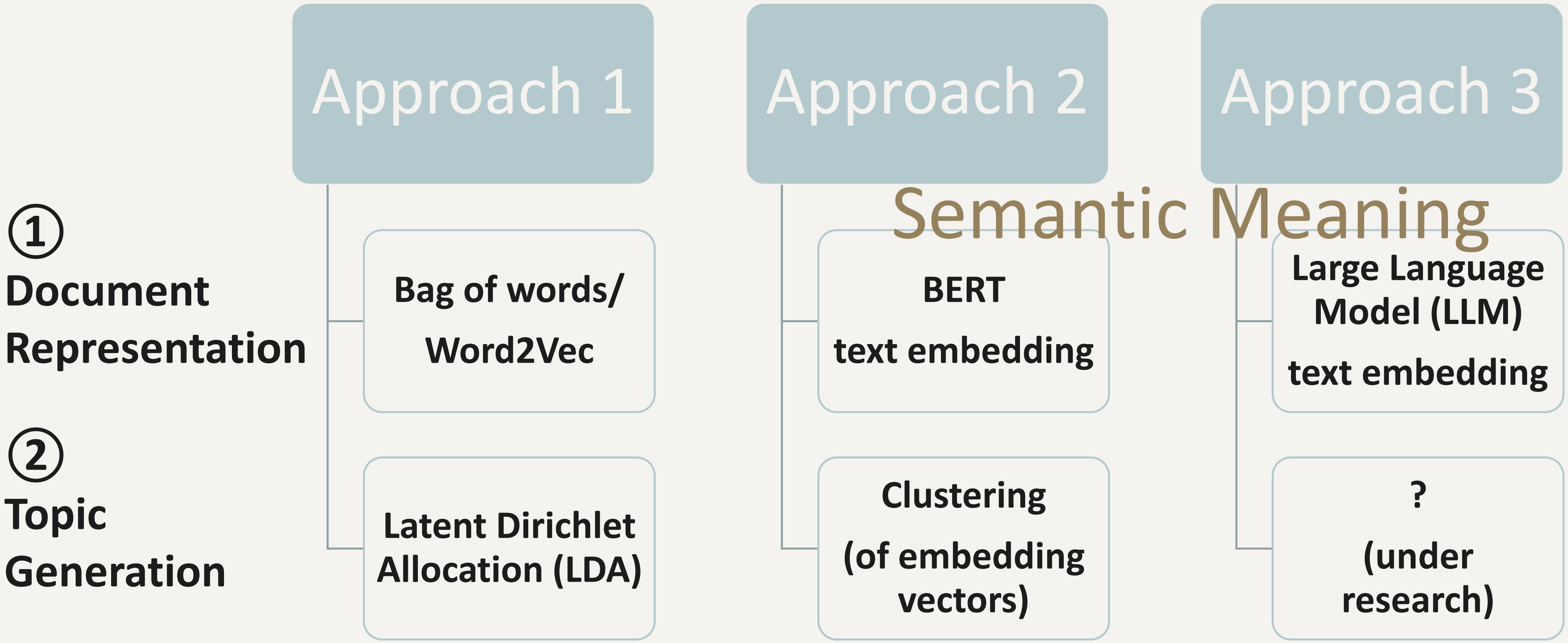
Keyword co-occurrence map created using VOSviewer, mapping research landscape in LLM research in the field of Information and Library Science based on publication data from Web of Science

Topic Modeling

- unsupervised machine learning
- uncovers hidden and abstract themes or topics within a collection of text documents
- makes sense of unstructured data



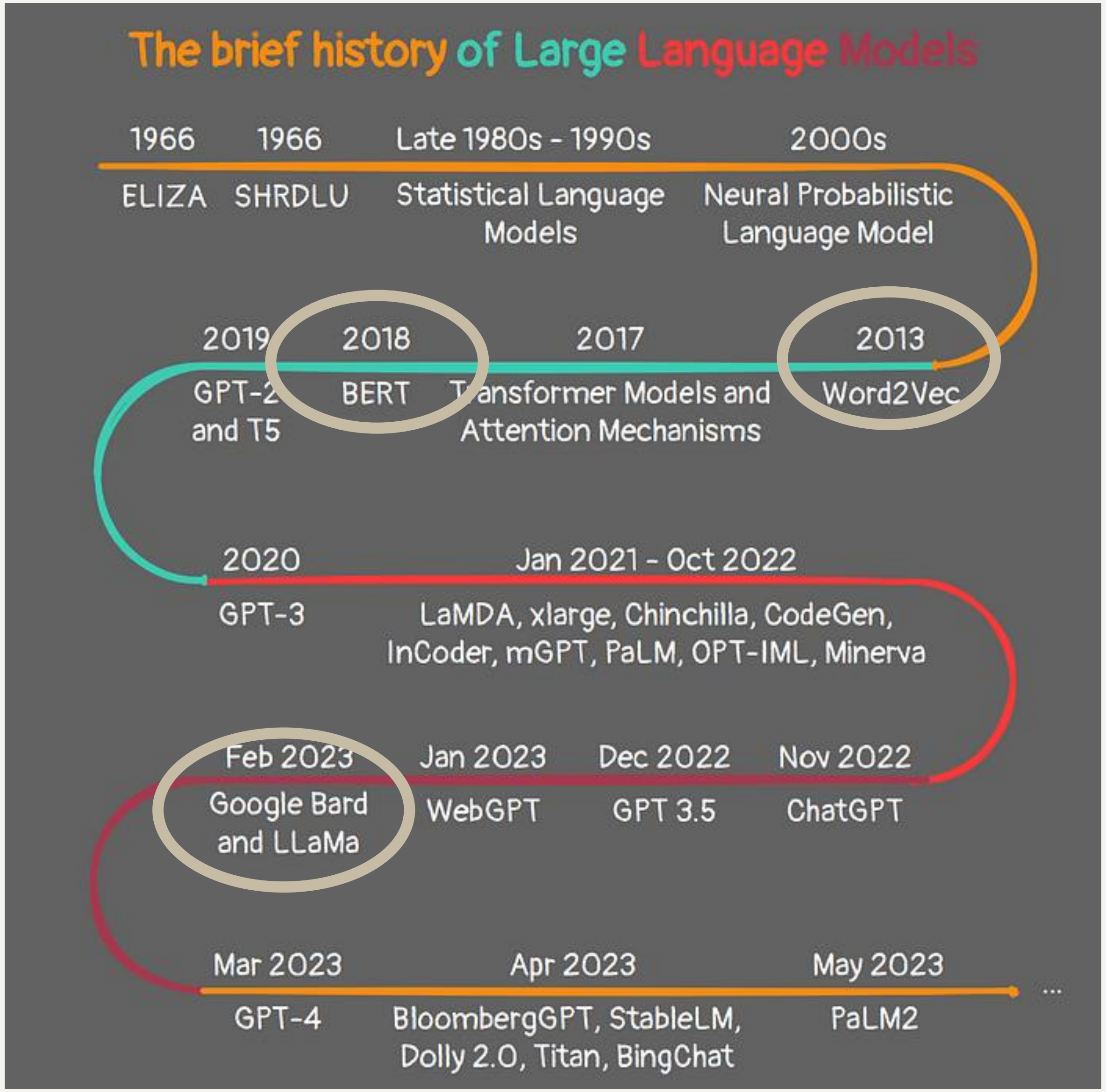
Experimenting with Different Approaches



	Approach 1	Approach 2	Approach 3
① Document Representation	Bag of words/ Word2Vec	BERT text embedding	Large Language Model (LLM) text embedding
② Topic Generation	Latent Dirichlet Allocation (LDA)	Clustering (of embedding vectors)	? (under research)

Semantic Meaning

Experimenting with Different Approaches



The brief history of Large Language Models

1966
ELIZA

1966
SHRDLU

Late 1980s - 1990s
Statistical Language Models

2000s
Neural Probabilistic Language Model

2013
Word2Vec

2017
Transformer Models and Attention Mechanisms

2018
BERT

2019
GPT-2 and T5

2020
GPT-3

Jan 2021 - Oct 2022
LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva

Nov 2022
ChatGPT

Dec 2022
GPT 3.5

Jan 2023
WebGPT

Feb 2023
Google Bard and LLaMa

Mar 2023
GPT-4

Apr 2023
BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat

May 2023
PaLM2

...

Graph by Armin Norouzi
Reference: The brief history of Large Language Models: A Journey from ELIZA to GPT-4 and Google Bard
https://levelup.gitconnected.com/the-brief-history-of-large-language-models-a-journey-from-eliza-to-gpt-4-and-google-bard-167c614af5af

Datasets

Dataset used in our experiments: newsgroups posts on 20 topics (split into a training set and a test set) https://huggingface.co/datasets/rungalileo/20_Newsgroups_Fixed

How to find other datasets:

- Hugging Face: https://huggingface.co/datasets
- Papers with code: https://paperswithcode.com/task/topic-models

Colab Notebook

https://colab.research.google.com/drive/18YM5HoLCi1_gb39OaiCgVnaRejAUyN4m?usp=sharing

Additional Resources

- Word Embeddings: https://ai.engin.umich.edu/2018/07/23/word-embeddings-and-how-they-vary/
- LDA: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- LDA Evaluation: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
- BERTopic: https://maartengr.github.io/BERTopic/index.html
- The Illustrated Transformer: https://jalammar.github.io/illustrated-transformer/

Thank you!