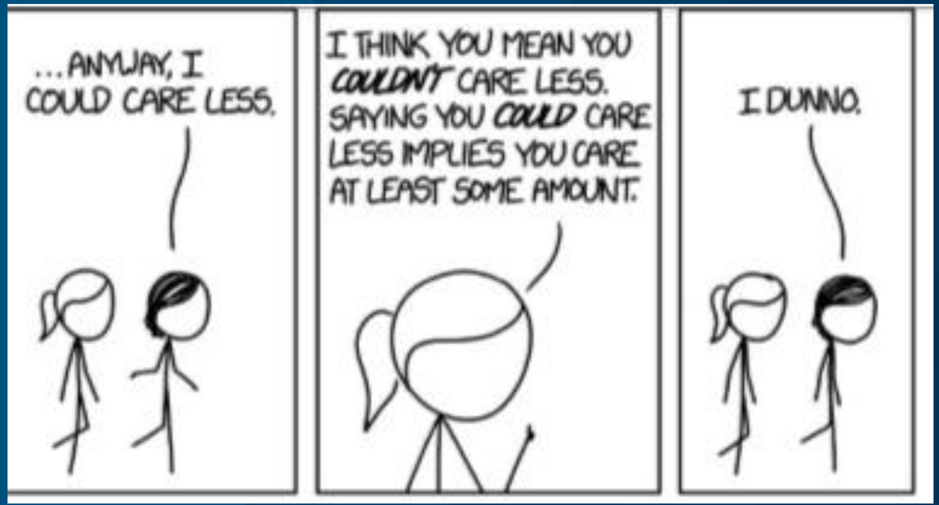
NLP Workshop – APPAM 2024

Ruhan Circi, Principal Data Scientist, AIR

Bhashithe Abeysinghe, Researcher/Computer Scientist, AIR



Source: xkcd.com

Circi, R., Abeysinghe, B., “Natural Language Processing and Artificial Intelligence for Text Analysis” *Association for Public Policy Analysis and Management* 2024

Introductions

RUHAN CIRCI

BHASHITHE ABEYSINGHE

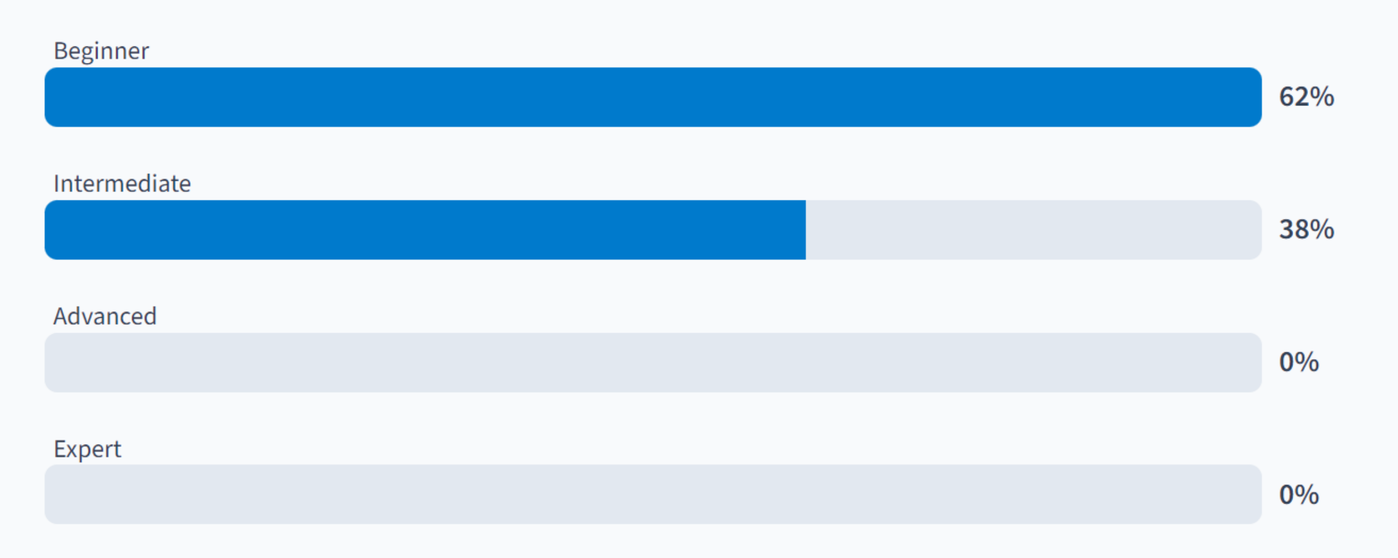
Your Turn

What is your motivation for this training?

Your Turn

What is your familiarity with

Natural Language Processing?



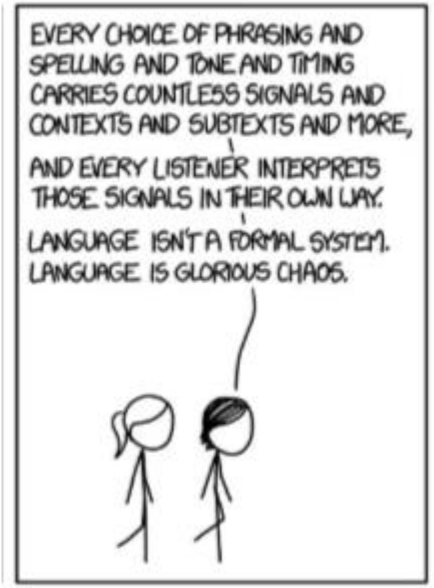
Agenda

Schedule (.am)	Activity
8.30-8.45	Introductions, agenda and goals for the day
8.45-8.55	**Topic:** Text as Data and NLP Use Cases [Presentation]
8.55-9.05	Breakout one ▪ Thought experiment: How analyzing text data could be valuable in your field of expertise or area of interest
9.05-9.15	Regroup and Q&A/Audience discussion
9.15-9.30	**Topic:** NLP as a Field [Presentation]
9.30-10.00	**Topic:** NLP Pipeline I [Presentation] ▪ Analysis of environmental science & policy research abstracts
10.00-10.15	Required break
10:15-10:30	Welcome new attendees & recap
10.30-11.15	**Topic:** NLP Pipeline II [Presentation] ▪ Analysis of environmental science & policy research abstracts cont.
11.15-11.30	Breakout two ▪ Thought experiment cont.: Data analysis, policy implications, and fairness concerns
11.30-11.40	**Topic:** Bias and Risks [Presentation]
11.40-11.45	Wrap up, next steps

Natural Language

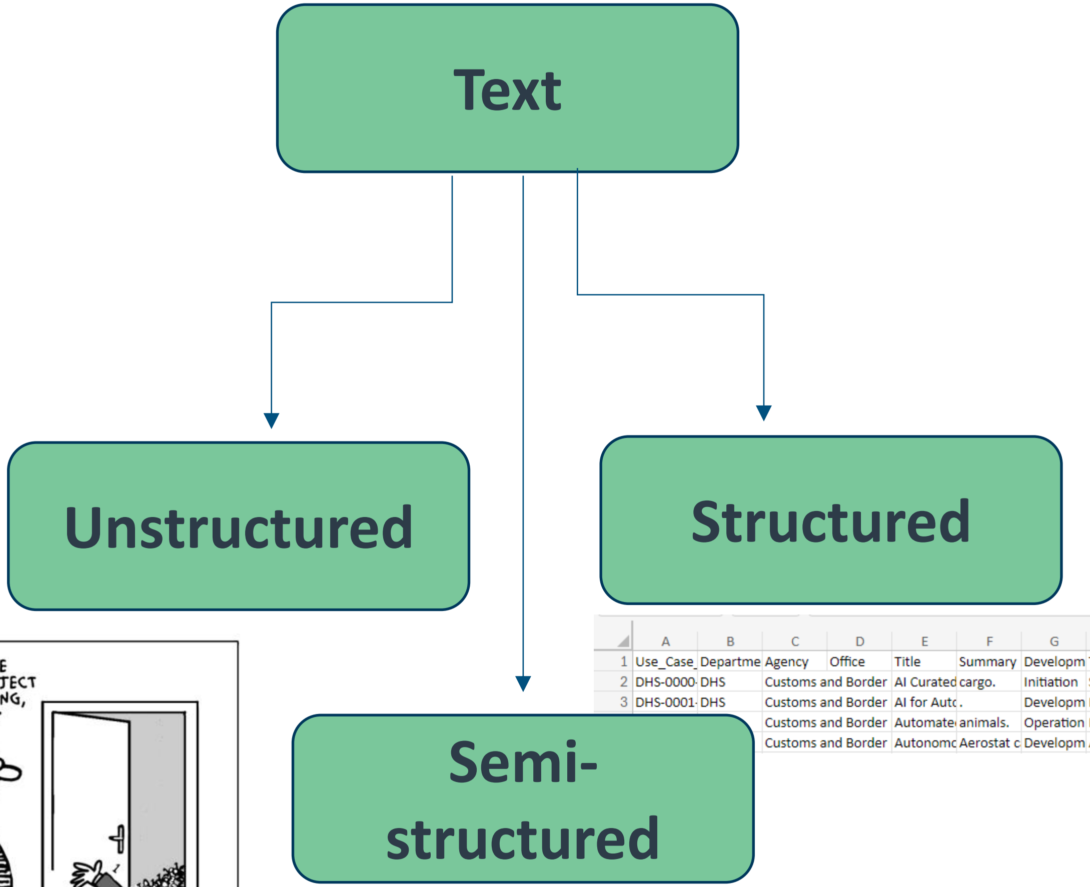
- Language used for everyday communication
  - English
  - 中文
  - Italiano
  - Español
- Any output we produce to communicate
  - Speech
  - Process Dialog
  - Retrieve information
  - Write



Source: xkcd.com

- Reports (long and short)
- Peer reviewed articles
- User comments
- Online documents
- Emails
- Videos
- Speech



	A	B	C	D	E	F	G	H	I	J	K	L
1	Use_Case_	Departme	Agency	Office	Title	Summary	Developm	Technique	Source_Cc	Department		
2	DHS-0000-	DHS	Customs and Border		AI Curated	cargo.	Initiation	Synthetic Image Gene		Department of Homeland Security		
3	DHS-0001-	DHS	Customs and Border		AI for Autc	.	Developm	Machine Vision		Department of Homeland Security		
			Customs and Border		Automate	animals.	Operation	Machine Learning		Department of Homeland Security		
			Customs and Border		Autonomc	Aerostat c	Developm	Automation & Roboti		Department of Homeland Security		


```
{
    "glossary": {
        "title": "example glossary",
                "GlossDiv": {
            "title": "S",
                    "GlossList": {
                "GlossEntry": {
                    "ID": "SGML",
                                "SortAs": "SGML",
                                "GlossTerm": "Standard Generalized Markup Language",
                                "Acronym": "SGML",
                                "Abbrev": "ISO 8879:1986",
                                "GlossDef": {
                        "para": "A meta-markup language, used to create markup languages such as DocBook.",
                                        "GlossSeeAlso": ["GML", "XML"]
                    },
                                "GlossSee": "markup"
                }
            }
        }
    }
```


Image source: Cartoon by David Fletcher

Unstructured Text: For Whom?

NLP helps us

“Climate change is causing rising temperatures and extreme weather events”

Unstructured for Computer

Introduction – What is Natural Language Processing?

- An area of computer science that includes methods to analyze, model and understand human language (Vajjala et al., 2020)

- **Our definition**:

Make algorithm to operate with natural

language to do certain tasks that human can do

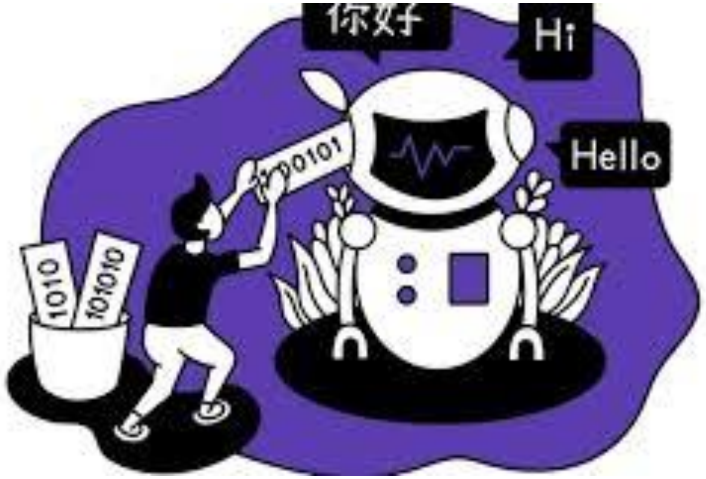
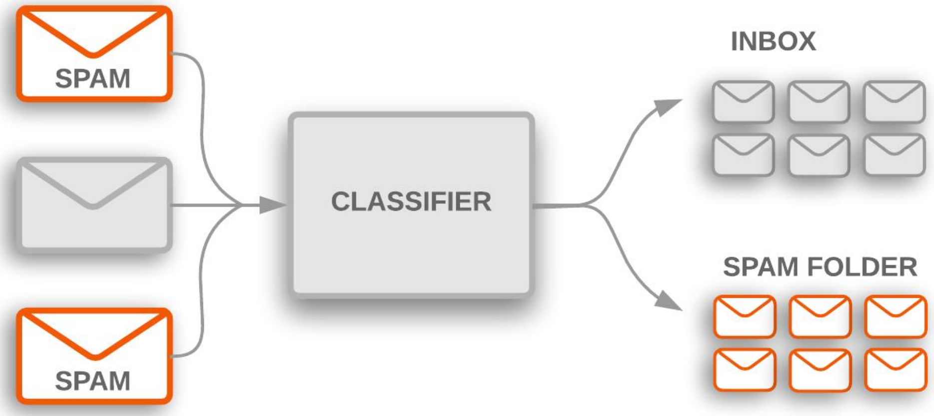


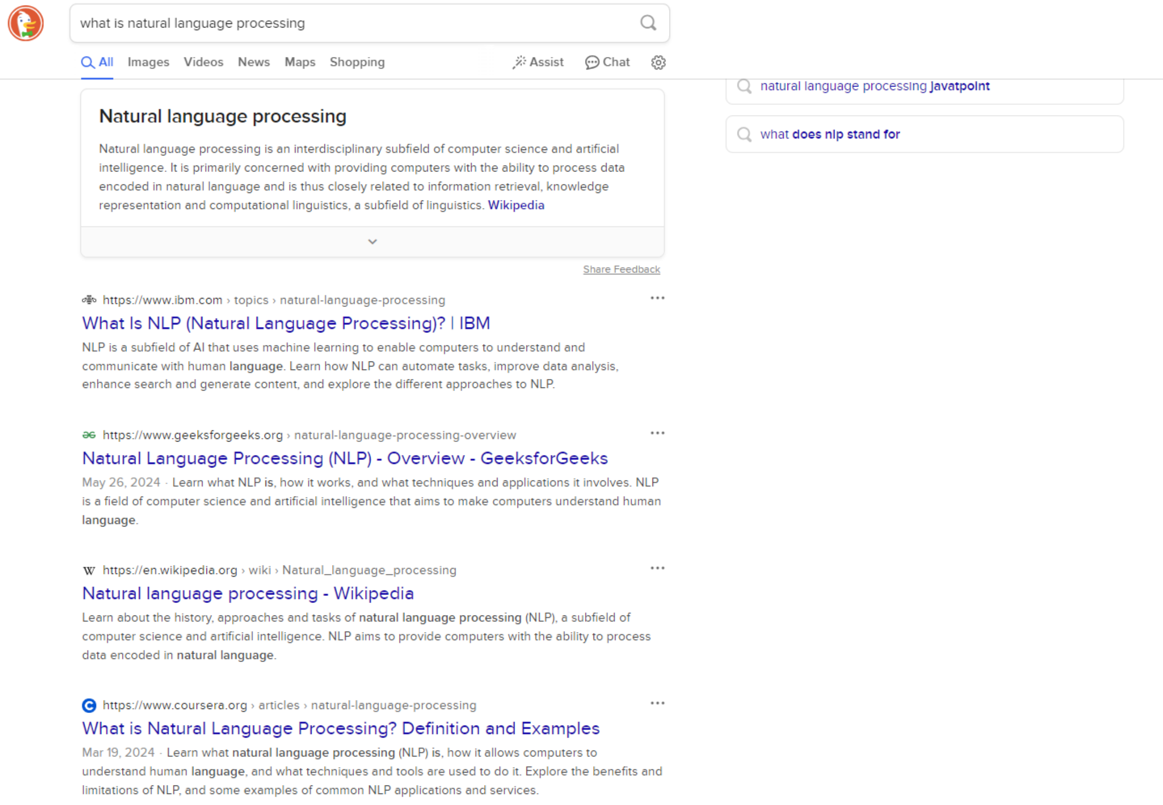
Image source: https://media.licdn.com/dms/image

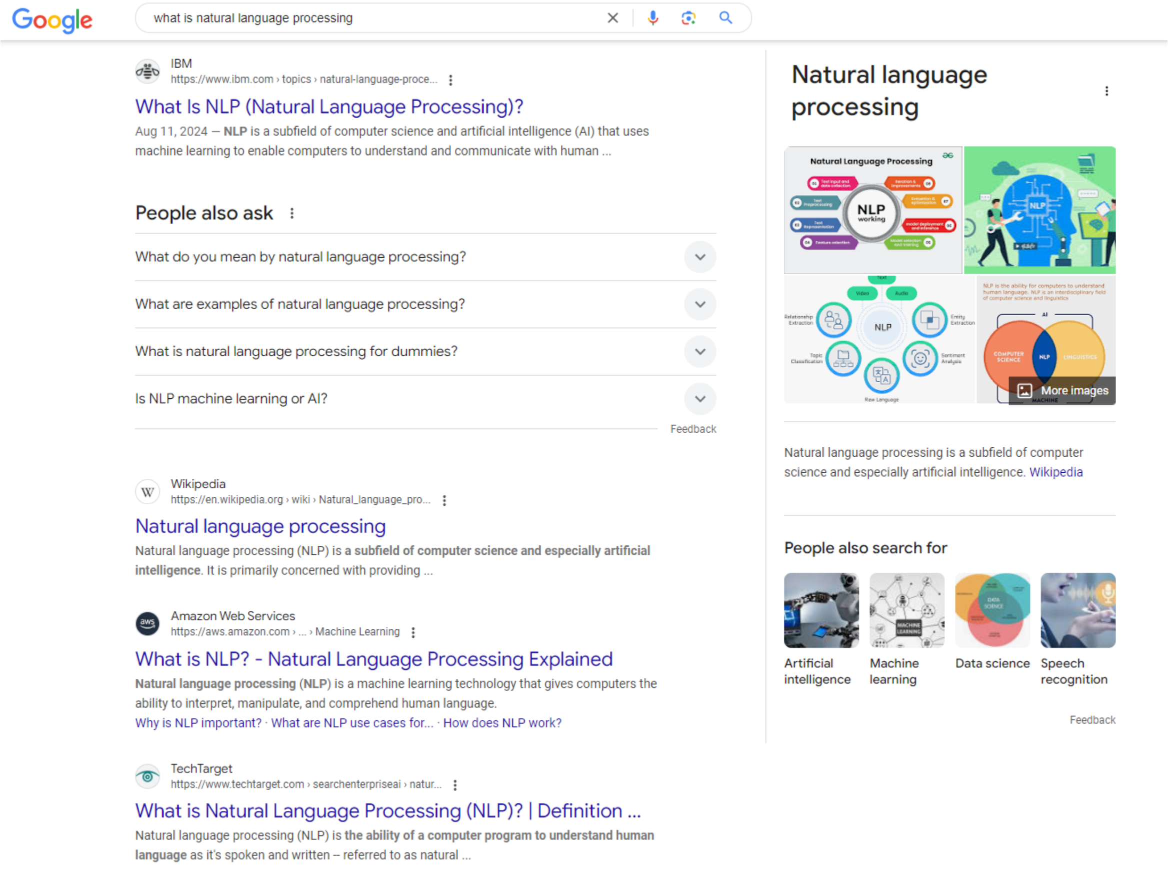
Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media.

NLP Use Cases: Text Classification

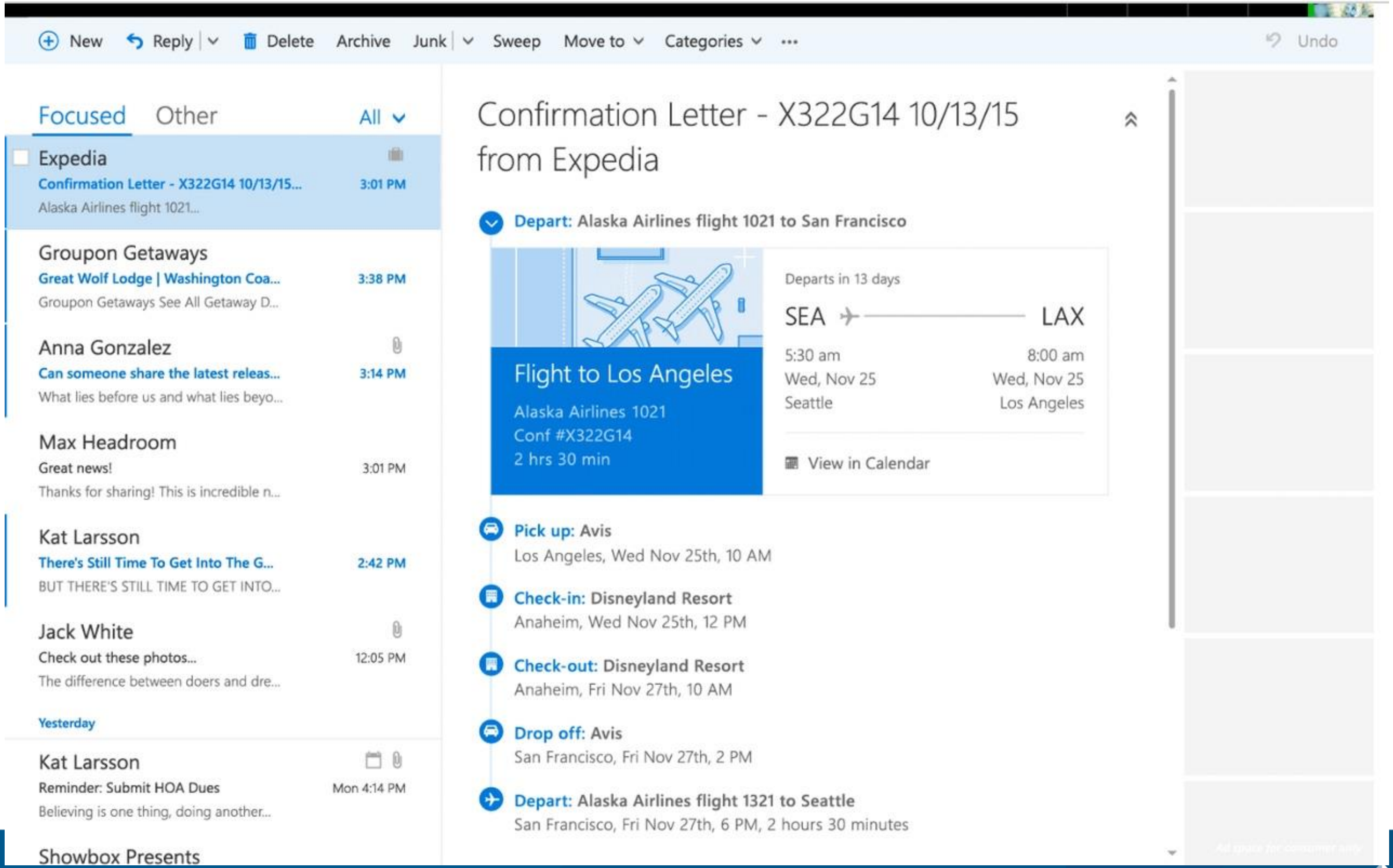


NLP Use Cases: Information Retrieval





NLP Use Cases: Information Extraction

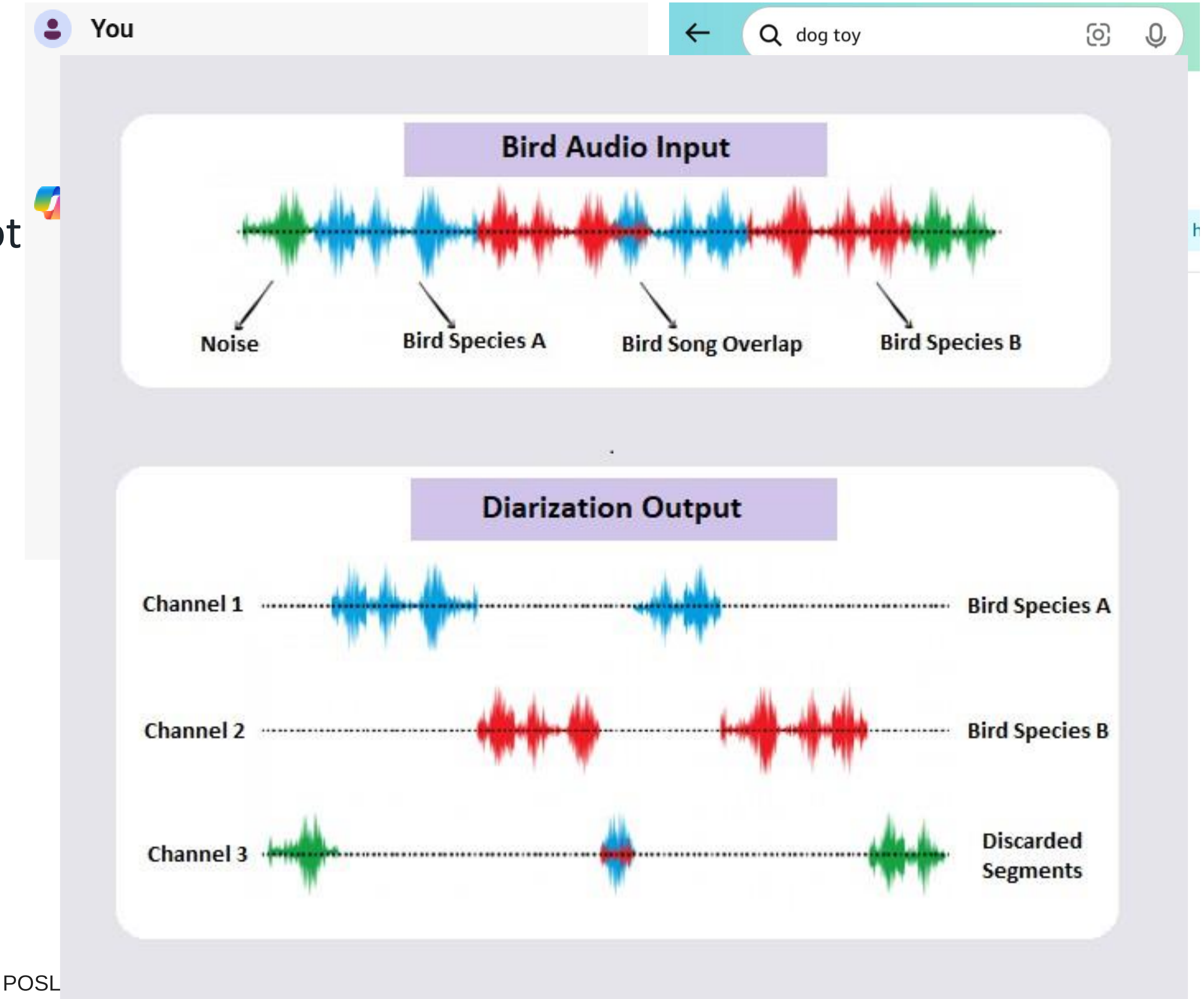


AIR

NLP Use Cases: Machine Translation

NLP Use Cases More

- Have you used Virtual Assistant/Chatbot
  - Question Answering
- Amazon reviews
  - Sentiment analysis
  - Aspect based sentiment analysis
- Disentanglement/Diarization



Abeysinghe, B., Shah, D., Freas, C., Harrison, R., & Sunderraman, R. (2022, April). POSL
ACM/SIGAPP Symposium on Applied Computing (pp. 1756-1763).

Breakout I : Activity

Think about a project or problem in your field that analyzing text (e.g., reports, feedback, articles) played or could play a key role.

Are there any text data in your field that you would like to use? What will be the source?

What kind of questions would you like to answer from this text data?

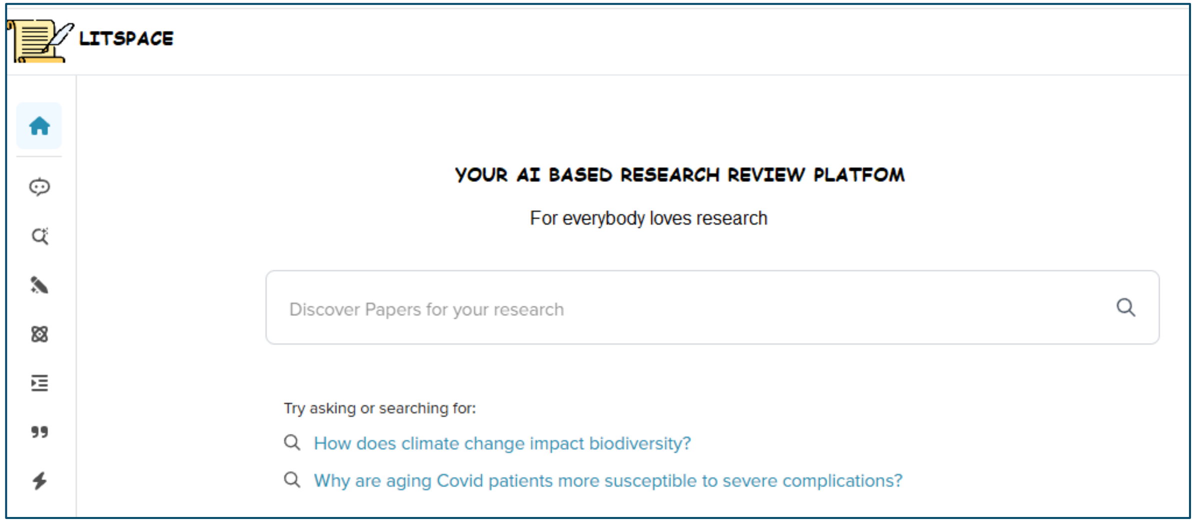
Why Do We Need NLP?

- Extract valuable **insights from large volumes** of unstructured text data
- **Automate** tasks related to text processing, such as **document summarization**, content categorization, and sentiment analysis
- Power **virtual personal assistants** and **chatbots**
- Improve search engines' ability to **understand** user queries and yield more relevant results
- **Generate human-like text**, aiding in content creation for various purposes and more…

Why Do We Need NLP? An Example Case

For example, reading 1000s of research articles to include them in a systematic literature study

- » May take hours for a human
- » Algorithm may complete this in minutes



How to Teach Language to Computers

- How do you instruct computers to do tasks?
  - Programming languages
    These get parsed into binary instructions [0/1]
  - If you don't know programming languages, then getting computers to do your tasks may not be trivial
- **NLP is the connector between language and computer**

```python
print('Hello world')
```

✓ 0.0s

```
Hello world
```

```
SSSTSSTSSSL:Push_+1001000=72='H'_onto_the_stack
TL
SS:Output_'H';_SSSTTSSTSTL:Push_+1100101=101='e'_onto_the_stack
TL
SS:Output_'e';_SSSTTSTTSSL:+1101100=108='l'
TL
SSSSSTTSTTSSL:+1101100=108='l'
TL
SSSSSTTSTTTTL:+1101111=111='o'
TL
SSSSSTSTTSSL:+101100=44=','
TL
SSSSSTSSSSSL:+100000=32=Space
TL
SSSSSTTTSTTTL:+1110111=119='w'
TL
SSSSSTTSTTTTL:+1101111=111='o'
TL
SSSSSTTTSSTSL:+1110010=114='r'
TL
SSSSSTTSTTSSL:+1101100=108='l'
TL
SSSSSTTSSTSSL=+1100100=100='d'
TL
SSSSSTSSSSTL:+100001=33='!'
TL
SS:Output_'!';_L
L
L:End_the_program
```

Whitespace (a programming language)

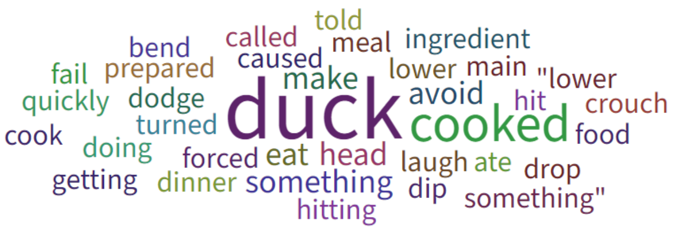
How to Teach Language to Computers?

Concept	Example
Phonetics and Phonology	"route" as /ruːt/ vs. /raʊt/ in GPS navigation
Morphology	"unhappiness" as "un-" (negation) + "happy."
Syntax	"How to bake a cake" vs. "Cake how to bake.“ on search engine
Semantics	"This product is cool" (positive sentiment, not temperature).
Contextuality	"Do you have a minute?" meaning "Can we talk?"
Discourse	I forgot to bring my umbrella. Thankfully, it didn’t rain.

Your Turn

- NLP is challenging
  - Ambiguity (example from Vajjala et al. (2020))
    - » "I made her duck"

"I made her duck" what does this mean for you?



NLP is Closely Related to Following

- Algorithm
- Artificial Intelligence
- Machine Learning
- Deep learning

"No, hashtags are not a part of speech."

Clear the Jargon : Algorithm

- An algorithm is a set of instructions to accomplishing a task, step by step

Task: Count the cupcakes on the tray

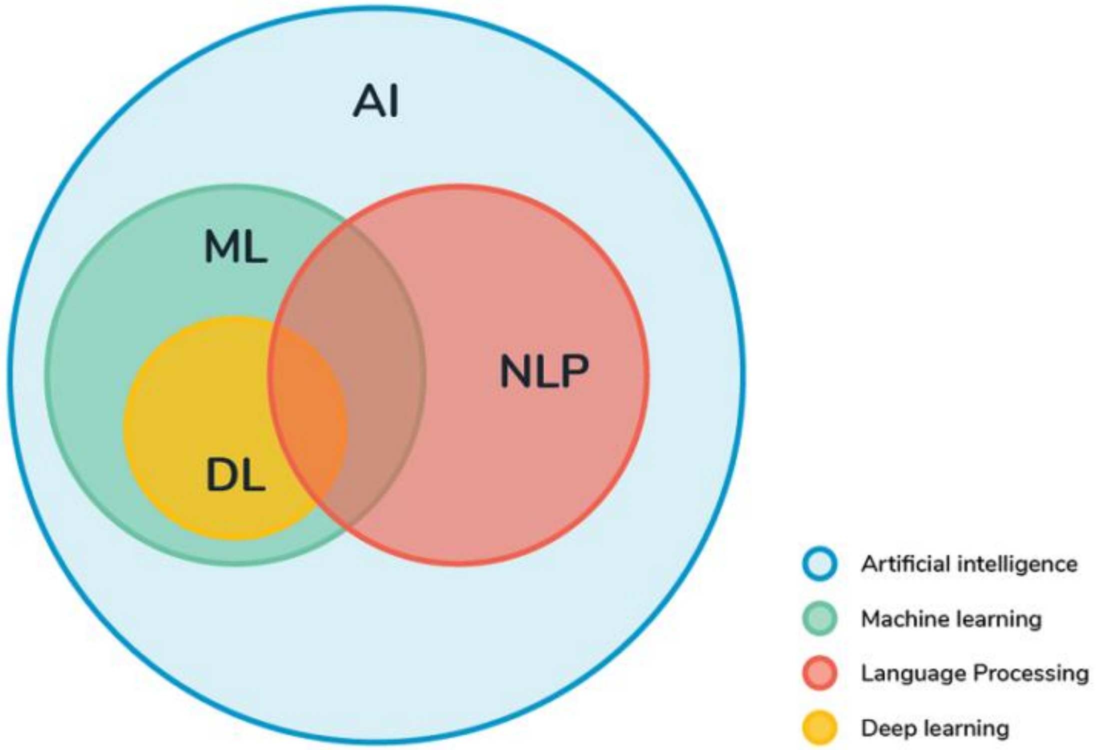
Let **Ncake = 0**

for each cupcake on the tray

Set **Ncake = Ncake +1**

Clear the Jargon: Relationships

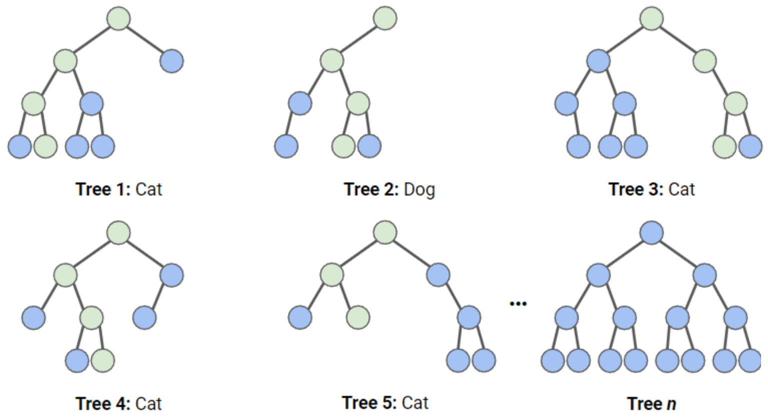
- NLP is a subdomain in the world of AI
  - Many NLP tasks are achieved through applying ML or DL
- ML and DL are algorithmic processes to discovering patterns in data
  - ML has a lot of algorithms
  - DL refers to using various neural network architectures

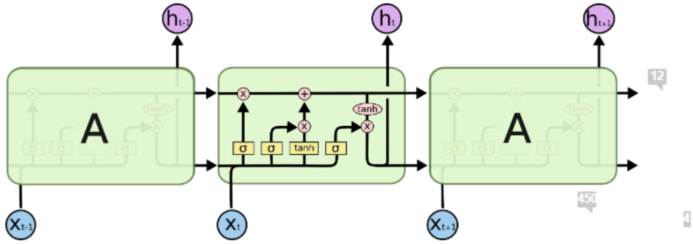


https://www.nomidl.com/natural-language-processing/difference-between-deep-learning-and-natural-language-processing/

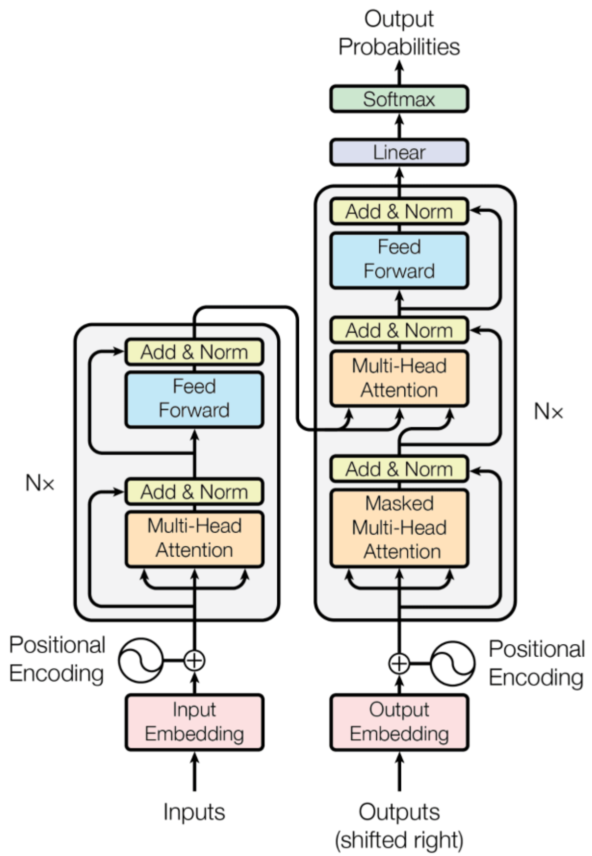
Clear the Jargon: Methods

- Popular ML methods
  - Naïve Bayes
  - Support Vector Machines
  - Hidden Markov Model
  - Conditional Random Fields
  - Random Forests, Decision trees
  - XGBoost
- Popular Deep Learning
  - Recurrent NN, LSTM
  - Convolutional NN
  - Transformers





The repeating module in an LSTM contains four interacting layers.



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

NLP PIPELINE

NLP Pipeline

Data Acquisition

Documents (text, image)

Data Preprocessing

Data Cleaning

Data Transformation
- Tokenization
- Stop words
- Stemming

Linguistic Aspects
- POS tagging
- Parsing

Feature Engineering (text to numeric)

Doc	Token1	Token2
1	0.25	1
2	0.66	0

Techniques

Bag of Words
N-grams
TF-IDF

Word embeddings

Modeling

Rule-based

Machine learning
- Supervised
- Unsupervised
- Semi-supervised

Deep learning

GenAI - LLMs

Model Evaluation

Confusion matrix
Quality against reference text

Data acquisition	Preprocessing	Feature engineering	Modeling	Evaluation

Our Experimental Goal Today

Goal: Examine the changes in common topics in environmental science & policy research over time

Research Question: What are the main topics in studies involved in environmental science & policy research?

Data Acquisition

- Can be challenging
- You can obtain data through
  - Surveys
  - Public repositories
  - Scraping
  - Product logs
  - (Data augmentation) last resort

```
def get_citations(doi):
    request = urllib.request.Request(f'https://api.crossref.org/works/{doi}')
    page = urllib.request.urlopen(request).read().decode('utf-8')

    #return json.loads(page)[0]['count']
    return json.loads(page)['message']['is-referenced-by-count']
```

Our Dataset

Data source

Article data via Elsevier [2000 (Jan) -2024 (Nov)]

```
request = urllib.request.Request(f'https://api.elsevier.com/content/article/pii/{pii}?APIKey=
page = urllib.request.urlopen(request).read().decode('utf-8')
```

Data provided via API

Our Dataset for Today's Session

- Collection of meta-data and abstracts of articles

Evolving paradigms for landscape-scale renewable resource management in the United States

Scott A. Mullner [a], Wayne A. Hubert [a], Thomas A. Wesche [b]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/S1462-9011(00)00095-2 Get rights and content

Abstract

Patterns of selection and evolution of renewable resource management paradigms appear when strategies are considered across a temporal scale. The roles of renewable resource managers were established during the early twentieth century and have since evolved. Autocratic natural-science-based management (ANM) of renewable resources was institutionalized in the early twentieth century following the principle of management based on science with administrative decisions by professional agency employees. A more recent form of management paradigm is interactive natural-science-based management (INM) which provides for limited stakeholder involvement in the decision-making process. These historic paradigms often inadequately addressed social and political aspects of renewable resource management leading managers to adopt new management paradigms involving communications and negotiations among stakeholders, not just science and administrative decisions. The inability of either ANM or INM paradigms to win uncontested agency and public acceptance, coupled with demands to increase spatial scales of management and public (stakeholder) involvement, is providing impetus for emergence of a new paradigm. The evolving paradigm can be defined as collaborative natural- and social-science-based management (CNSM) and provides a framework for approaching and finding solutions to landscape-scale problems. Successful evolution of this paradigm will require removing barriers to societal involvement in management decision-making institutionalized over the past century.

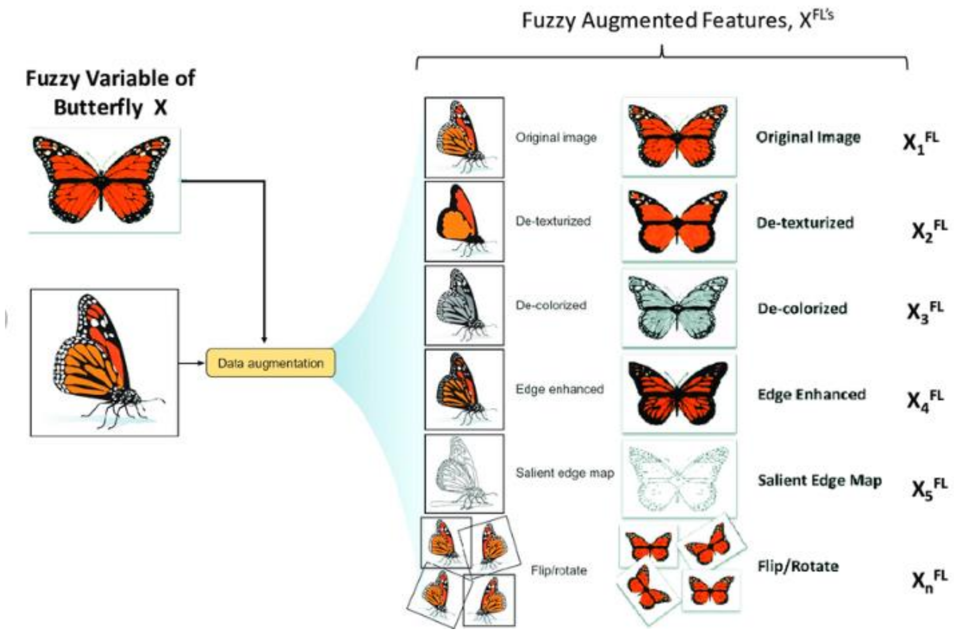
Sample Size

- Smallest data set for statistical analysis ?
  - Commonly cited number is 30
- Small dataset for NLP task ?
  - What is small?
    - » Small data may lead to poor generalization
    - » Large data may increase computational burden

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, *11*(3), 269. https://doi.org/10.14778/3157794.3157797

Larger Data Set?

- Augmentations
  - Synonym replacement
  - Back translation
  - Entity replacement (replace "Georgia" with "Atlanta")
  - Adding noise to data
    - » Replace words with other words that are closer in spelling (Levenshtein distance)
  - Snorkel (Ratner et al. 2017)



"Climate change impacts ecosystems." → "Ecosystems are affected by climate change."

Buah, E., Linnanen, L., Wu, H., & Kesse, M. A. (2020). Can artificial intelligence assist project developers in long-term management of energy projects? The case of CO2 capture and storage. *Energies*, *13*(23), 6259.

Data Cleaning and Parsing

- Cleaning is generally data wrangling – no real NLP methods here
  - Parsing from HTML files, PDFs etc.
  - Scanned documents (Tesseract, OCR)

<html>
<author:…>
<yearValue:…>

Author	Year	Abstract
…	…	…
…	…	…

Let's Discuss

- Some abstracts are missing
  - Issues with crawling/scraping the website
- How should we handle missing data?

abstract

Missing:	0 (0%)
Distinct:	3029 (75%)

-1:	25%
Hydropower is very important for electricity su...	<1%
This paper brings together institutional theorie...	<1%
Other:	75%

The United States is reprioritizing domestic extraction anc
To properly address the polycrisis we need to tackle the u
Much of the research on forestry innovation is based on r
While the Anthropocene has seen the dissolution of natu
-1
-1
The Nanjing University of Information Science and Technc
The urgent need for climate change adaptation is becomi
Western wildfires present a complex sustainability challen

Preprocessing Steps

- Example: `'Attending to the unattended: Why and how do local governments plan for access and functional needs in climate risk reduction? '`
  - Sentence segmentation, word tokenization

```
['Attending', 'to', 'the', 'unattended', ':', 'Why', 'and', 'how', 'do', 'local', 'governments', 'plan', 'for', 'access', 'and', 'functional', 'needs', 'in', 'climate', 'risk', 'reduction', '?']
```

  - Stop-words removal

```
['Attending', 'unattended', ':', 'Why', 'local', 'governments', 'plan', 'access', 'functional', 'needs', 'climate', 'risk', 'reduction', '?']
```

  - Lemmatizing (was -> be, better -> good) and Stemming (running -> run, functional -> function)

```
['attend', 'unattend', ':', 'whi', 'local', 'govern', 'plan', 'access', 'function', 'need', 'climat', 'risk', 'reduct', '?']
```

  - Special characters (numbers, Unicode)

Our Dataset Before Preprocessing

```
df=pd.read_feather('ESP.feather')
df.assign(authors=df.authors.str.replace('<#>', ';'))[['authors', 'title', 'pub_date', 'abstract', 'cite_count', 'doi', 'pii', 'openaccess']]
```

✓ 0.0s Python

	authors	title	pub_date	abstract	cite_count	doi	pii	openaccess
0	Zhang, Fengxiu;Xiang, Tianyi	Attending to the unattended: Why and how do lo...	2024-12-31	Research and practice in climate risk reductio...	0	10.1016/j.envsci.2024.103892	S1462901124002260	false
1	Yoshida, Yuki;Sitas, Nadia;Mannetti, Lelani;O'...	Beyond Academia: A case for reviews of gray li...	2024-12-31	Gray literature is increasingly considered to ...	0	10.1016/j.envsci.2024.103882	S1462901124002168	true
2	Pietrzyk-Kaszyńska, Agata;Olszańska, Agnieszka	Of heroes and villains – How coalitions shape ...	2024-12-31	Policy narrative analyses provide important in...	0	10.1016/j.envsci.2024.103899	S1462901124002338	false
3	Zurba, Melanie;Suchet-Pearson, Sandie;Bullock,...	Enhancing meaningful Indigenous leadership and...	2024-12-31	This is the first global empirical study that ...	0	10.1016/j.envsci.2024.103864	S1462901124001989	true
4	Lemke, Leonard Kwhang-Gil;Beier, Julia;Hanger-...	Exploring procedural justice in stakeholder id...	2024-12-31	In the face of complex societal challenges, st...	0	10.1016/j.envsci.2024.103900	S146290112400234X	true

Our Dataset After Preprocessing

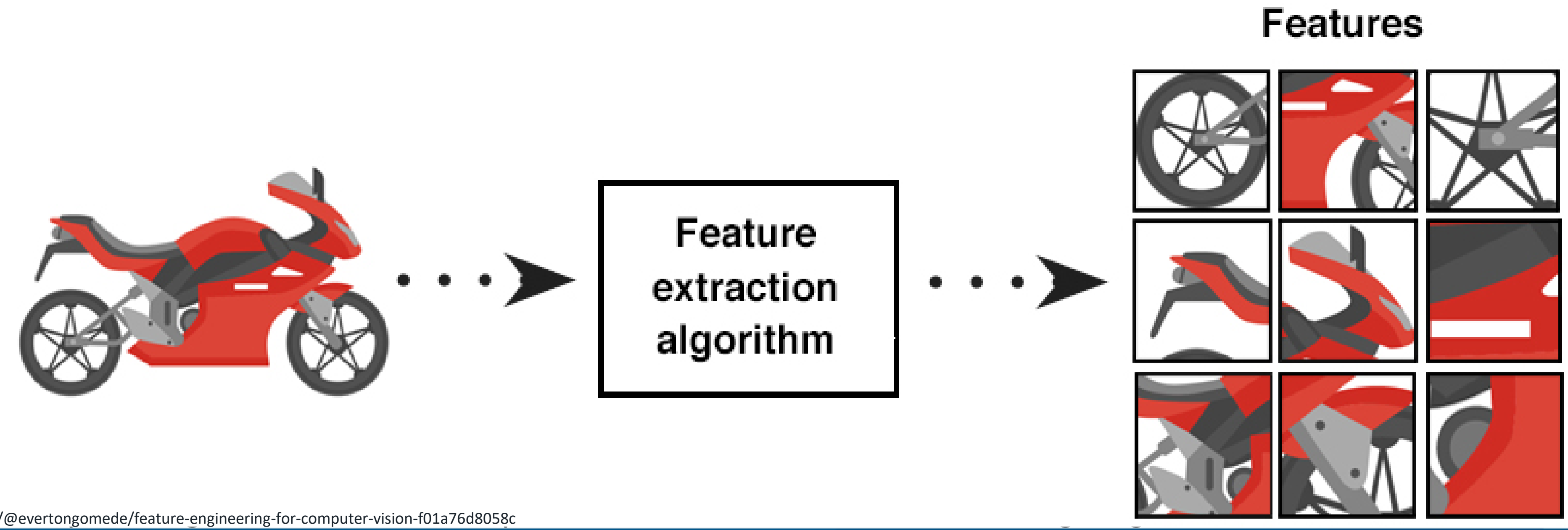
```
remove_special_chars().lower_case().tokenize().remove_stopwords().lemmatize().get_text()
```

- No missing
- Removed columns which are not necessary

pub_date	preprocessed
Missing: 0 (0%) Distinct: 201 (7%) 201 Distinct values	Missing: 0 (0%) Distinct: 3028 (>99%) 3028 Distinct values
2024-12-31	research practice climate risk reduction often view marginalized individual lens vulnerability however perspective lack specificity group need incor
2024-12-31	gray literature increasingly considered complement evidence knowledge peerreviewed literature sciencepolicy process applied research one hand
2024-12-31	policy narrative analysis provide important insight understand mechanism dynamic policy change also explore narrative shape bind coalition polic
2024-12-31	first global empirical study specifically explores perspective indigenous people people working indigenous people organisation ipo people workin
2024-12-31	face complex societal challenge stakeholder participationengagement knowledge coproduction become increasingly important sustainability scier
2024-12-31	smart specialization emerged vital strategy driving responsible research innovation across europe despite growing importance integration webbas
2024-12-31	world arguably existential crisis crisis manifesting nearly every facet existence education mental health culture democracy environment institution
2024-12-31	face growing pressure marine environment evidencebased decisionmaking realm marine conservation policy utmost importance boundary work c
2024-12-31	ambitious environmental policy regulation europe aim reduce pesticide use yet implementation face significant obstacle effective strategy gain su
2024-12-31	safe sustainable design ssbd concept integrates safety sustainability chemical material throughout entire life cycle minimizes environmental footpr
2024-12-31	article critically analyzes social political factor behind advancement technoscientific development modern brazilian agriculture second half 20th ce
2024-12-31	climate change pose significant threat ecosystem biodiversity conventional management strategy often fall short leading uncertainty addressing c
2024-12-31	biodiversity conservation increasingly recognized main challenge sustainability agenda human epicenter biodiversity crisis conserving nature requ
2024-12-31	climate change driving extreme weather heat flooding increasingly require evacuation recent study found inconclusive result determinant evacuat
2024-12-31	medium portrayal climate protester predominantly painted climate protester deviant antisocial protest paradigm leading negative reception publi

Feature Engineering :Vectorize the Text

- Words are just strings, not very helpful!
- We want to represent text numerically (**text → number**) via vectorization
- We want to create features and derive new features (e.g., count of syllables is not available in the text) for the analysis



https://medium.com/@evertongomede/feature-engineering-for-computer-vision-f01a76d8058c

Feature Extractions

- **Algorithmic extractions**
  - » Bag of Words (BoW)
  - » Term Frequency Inverse Document Frequency (TFIDF)
  - » N-grams
  - » Transformer embeddings – BERT, elmo, RoBERTa etc.

- Most engineered features (**non-algorithmic**) are really useful in text classification
  – e.g., classifying documents that discuss about “deforestation” vs “pesticides”

NLP Pipeline

In our exploration scenario: *Examine the changes in common topics in environmental science & policy research over time*

- Algorithmic extractions are best
  - Because we don’t know exactly what topic we would see

Bag of words (BoW)

Description: Process of breaking text into individual word count statistics

a. The value associated with each word is its frequency in the document

b. Essentially converts a document/text into a numerical vector

c. Loss of sentence structure, dependency of words and grammar

```
docs = [
    '''Biodiversity is crucial for maintaining healthy ecosystems, but it is threatened by deforestation, which destroys habitats.''',

    '''Pollution from industrial activities contaminates air, water, and soil, further endangering biodiversity.''',

    '''Efforts to combat deforestation and reduce pollution are essential to protect the planet's biodiversity for future generations.'''
]
```

	activity	biodiversity	deforestation	industrial	maintaining	planet	pollution	protect	reduce	soil
Doc 1	0	1	1	0	1	0	0	0	0	0
Doc 2	1	1	0	1	0	0	1	0	0	1
Doc 3	0	1	1	0	0	1	1	1	1	0

https://ayselaydin.medium.com/4-bag-of-words-model-in-nlp-434cb38cdd1b

TF-IDF

Term frequency inverse document frequency (TFIDF)

a. Improves BoW by considering the importance of words
b. Term Frequency
   i. Measures how frequently a term appears in a document
c. Inverse Document Frequency (IDF)
   i. Measures how important a term is by considering how common or rare it is across all documents.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

	activity	biodiversity	deforestation	industrial	maintaining	planet	pollution	protect	reduce	soil
Doc 1	0.000000	0.425441	0.547832	0.000000	0.720333	0.00000	0.000000	0.00000	0.00000	0.000000
Doc 2	0.504611	0.298032	0.000000	0.504611	0.000000	0.00000	0.383770	0.00000	0.00000	0.504611
Doc 3	0.000000	0.278245	0.358291	0.000000	0.000000	0.47111	0.358291	0.47111	0.47111	0.000000

https://ted-mei.medium.com/demystify-tf-idf-in-indexing-and-ranking-5c3ae88c3fa0

BoW vs TF-IDF

BoW

	activity	biodiversity	deforestation	industrial	maintaining	planet	pollution	protect	reduce	soil
0	0	1	1	0	1	0	0	0	0	0
1	1	1	0	1	0	0	1	0	0	1
2	0	1	1	0	0	1	1	1	1	0

TFIDF

	activity	biodiversity	deforestation	industrial	maintaining	planet	pollution	protect	reduce	soil
0	0.000000	0.425441	0.547832	0.000000	0.720333	0.00000	0.000000	0.00000	0.00000	0.000000
1	0.504611	0.298032	0.000000	0.504611	0.000000	0.00000	0.383770	0.00000	0.00000	0.504611
2	0.000000	0.278245	0.358291	0.000000	0.000000	0.47111	0.358291	0.47111	0.47111	0.000000

- **BoW and TF-IDF are traditional approaches**
- **Mainly focus on frequencies within a limited context**

Coffee Break (10.00-10.15)

NLP Pipeline

Data Acquisition

Documents (text, image)

Data Preprocessing

Data Cleaning

Data Transformation
- Tokenization
- Stop words
- Stemming

Linguistic Aspects
- POS tagging
- Parsing

Feature Engineering (text to numeric)

Doc	Token1	Token2
1	0.25	1
2	0.66	0

Techniques

Bag of Words
TF-IDF
N-grams

Word embeddings

Modeling

Rule-based

Machine learning
- Supervised
- Unsupervised
- Semi-supervised

Deep learning

GenAI - LLMs

Model Evaluation

Confusion matrix
Quality against reference text

Word/Sentence/Document Embeddings

- Word embeddings: dense, continuous-valued representations of words that capture semantic relationships between words.

```
In [10]: nlp(u"pollution").vector

Out[10]: array([-1.5776649 , -0.16638929, -0.43895206, -0.04745665,  0.16214097,
               -0.03235428,  0.9215626 ,  0.3748377 ,  0.2827617 , -0.6613681 ,
                0.64951235, -0.97544885, -0.41579837,  0.7551622 ,  0.05494827,
                0.04006284, -0.81582147, -1.2004887 ,  0.9112468 ,  0.54774517,
                0.43714064,  0.42197418,  1.1573677 , -0.38139132,  0.28092423,
                0.38872135, -0.31069624,  0.643077  ,  0.00895375, -0.03980759,
                0.8954749 , -0.38518634,  0.68852556, -0.10997719, -0.11085591,
               -0.9712797 ,  0.62574136,  0.31639823,  1.0265087 ,  0.0928736 ,
               -0.25566372,  0.35664237, -0.2842762 ,  0.41854748,  0.21645036,
               -0.31171554, -1.0257293 ,  0.27165434,  0.6451773 , -0.77211404,
                0.7828064 , -0.06875579,  0.525453  , -0.85097736,  0.37000614,
               -0.6548801 , -0.2284891 , -0.72236097,  0.34280974, -0.71148694,
```

- Transformer embeddings
  - » Transformer embeddings are dense numerical representations of text generated by large, pretrained models like BERT and GPT
  - » They are usually trained on large datasets, so they can capture good contextual meaning and relationships

Transformer Embeddings

	abstract
0	Research and practice in climate risk reductio...
1	Gray literature is increasingly considered to ...
2	Policy narrative analyses provide important in...
3	This is the first global empirical study that ...
4	In the face of complex societal challenges, st...

Embeddings for whole document

Documents

	0	1	2	3	4	5	6	7	8	9	...
0	0.110548	0.007706	-0.061401	0.052673	0.086331	0.070158	0.000344	0.037999	-0.036692	0.011655	...
1	0.065996	0.066021	-0.008647	0.076451	0.014664	0.057418	-0.057534	0.022407	-0.026240	0.059444	...
2	0.049508	0.116063	-0.003712	0.062117	0.121245	0.042588	-0.041203	0.019516	0.041728	0.097406	...
3	0.050021	-0.000770	0.010225	0.047834	0.057153	0.014961	-0.062619	-0.007541	-0.012446	0.057593	...
4	0.068297	0.074023	0.005884	0.019898	0.090572	-0.003090	-0.011513	0.016520	0.024744	0.032146	...

Data for analysis

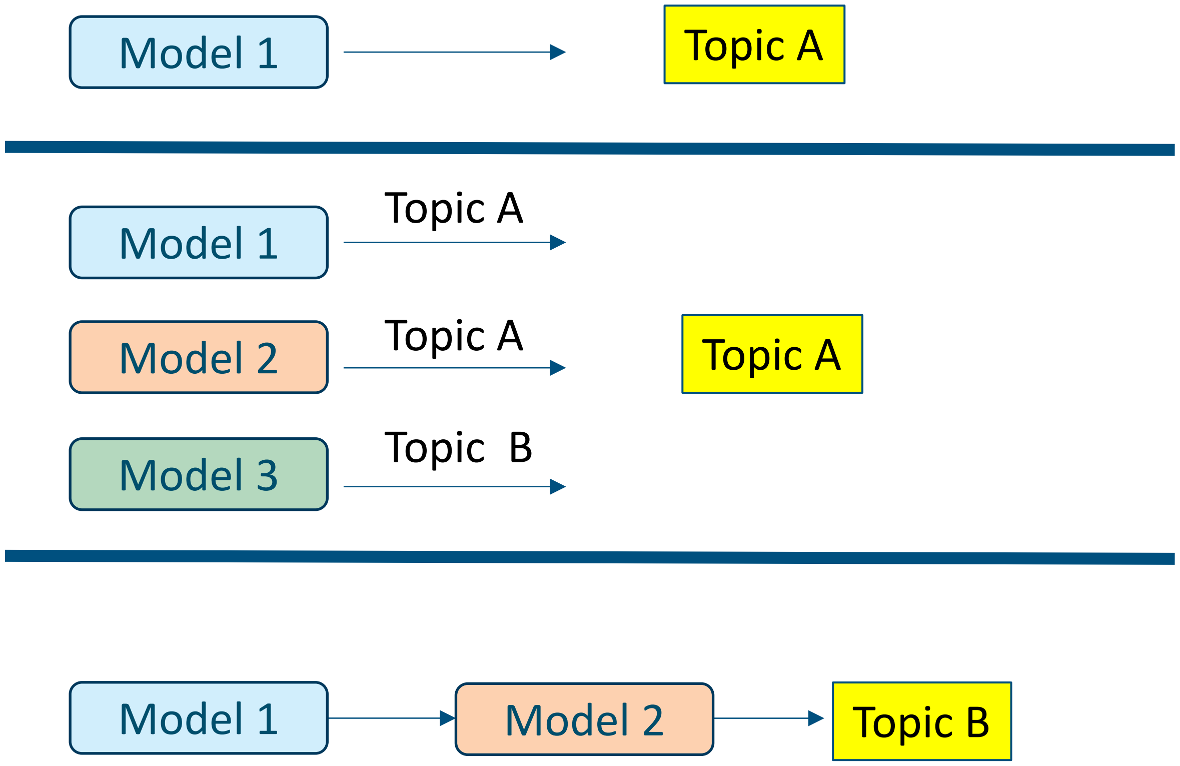
How to Pick a Model

- Depends on your purpose
  - Goal of our analysis in this section: *Identify common topics*
- Choose a Modeling Approach
  - Single vs Stacked/Ensemble Models
- Factors to Consider
  - Literature
  - Experience (e.g., LDA may struggle with short abstracts transformer based-model can be better)
  - Experimental testing (i.e., test multiple models)
    - » (Most reliable)

Single vs Stacked/Ensemble Models

1. **Single** model
    a. Simpler, faster, works well with straightforward tasks
2. Ensemble & Stacking
    a. Combine multiple models
        i. Combine predictions of multiple models and aggregate to find the final prediction – ***ensemble***

        i. Feed the outputs of a model into another model – ***stacking***



Our Modeling Approaches

Machine Learning	• Unsupervised • Clustering • Spectral Clustering : Chosen after experiments
BERTopic (End-to-End)	• BERT : Light transformer based model • Efficient and accessible for implementation • Ideal for topic modeling task with transparency and intuitiveness
GPT 3.5	• GPT: Advanced transformer based model • Superior performance across a range of tasks • Easy setup (no need to preprocess etc.) but opaque

Machine Learning Focus

- Our modelling process is **stacked**

Transformer embeddings → **Clustering** → **Topic Extraction**

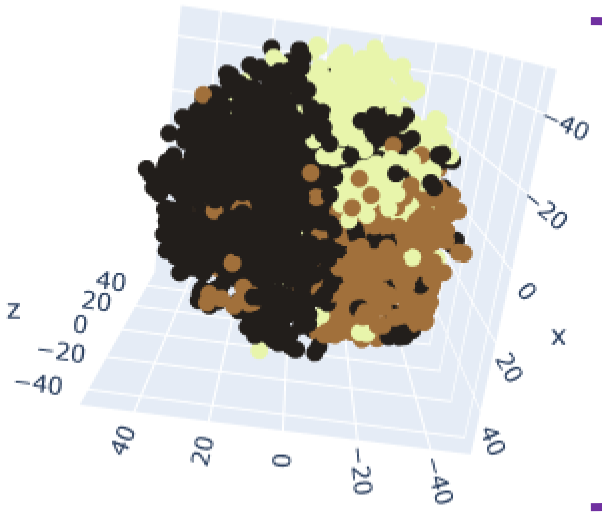
- Transformer models to choose from (all-Mini-LM)

- Clustering algorithms (DBSCAN SpectralClustering Birch)

- Class wise extraction (cTF-IDF)

Machine Learning Focus: Clustering

- Clustering with “Spectral Clustering”
  - **Definition: Clustering using a similarity matrix and its eigenvalues**



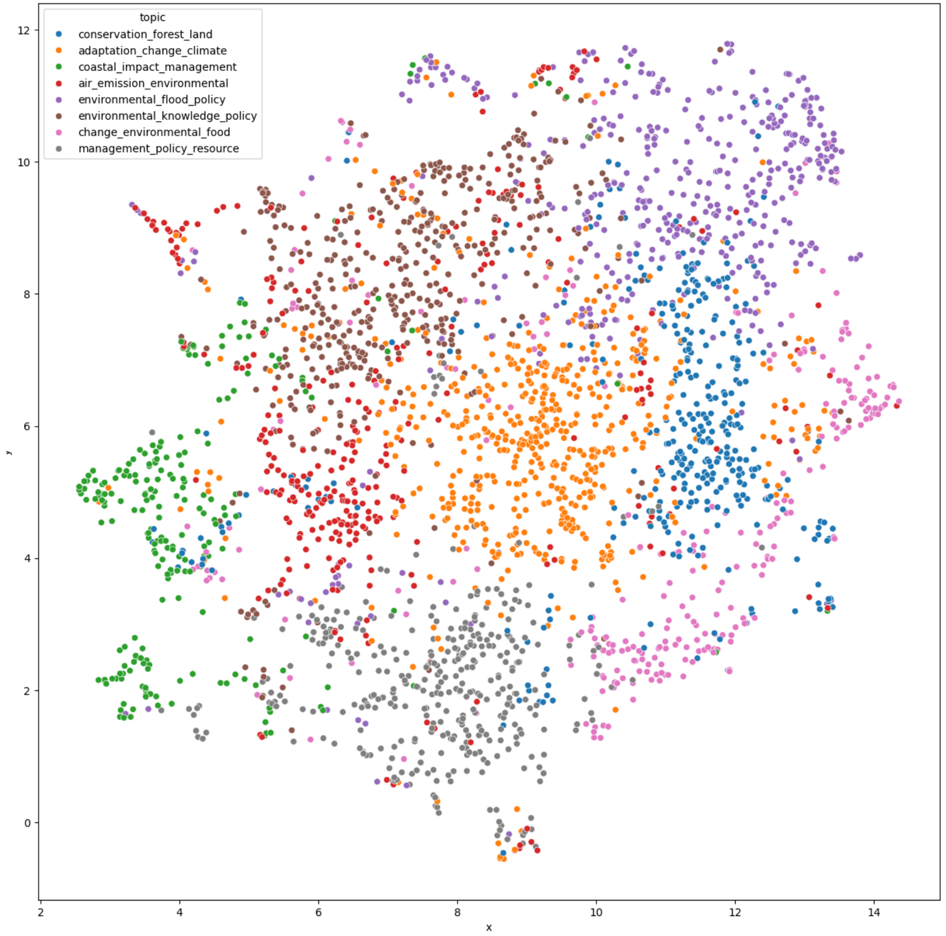
Clusters are not clearly separated

Machine Learning Focus: Optimization

- We compared clustering models
- Based Siluette scores, picked Spectral Clustering
- We used cTF-IDF to extract the topics
- Visualized the final topics (8 topics are extracted)

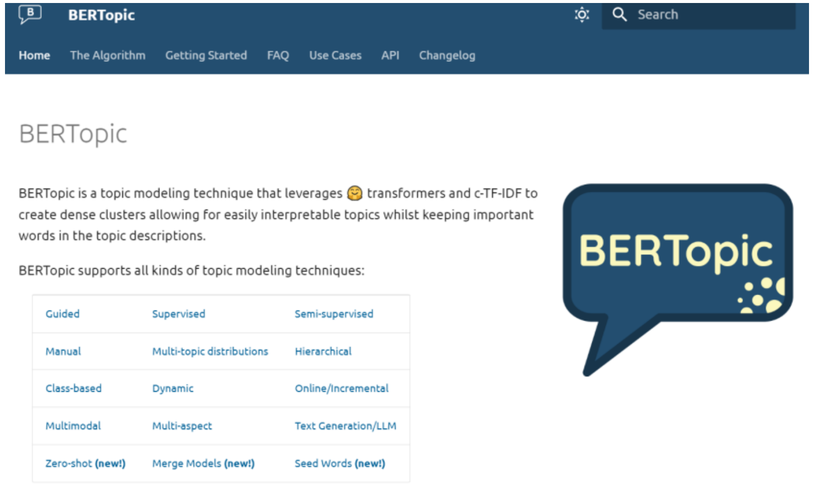
	model	sil_score
0	SpectralClustering()	0.292648
1	Birch()	0.292352
2	DBSCAN()	0.227754

```
['adaptation_change_climate',
 'environmental_knowledge_policy',
 'coastal_impact_management',
 'change_environmental_food',
 'air_emission_environmental',
 'conservation_forest_land',
 'environmental_flood_policy',
 'management_policy_resource']
```



BERTopic: End to End Process

- **End-to-End process**:
  - Input: the abstracts (no feature engineering)
  - BERTopic calculates embeddings
  - BERTopic extracts topics (provide topic labeling with words)



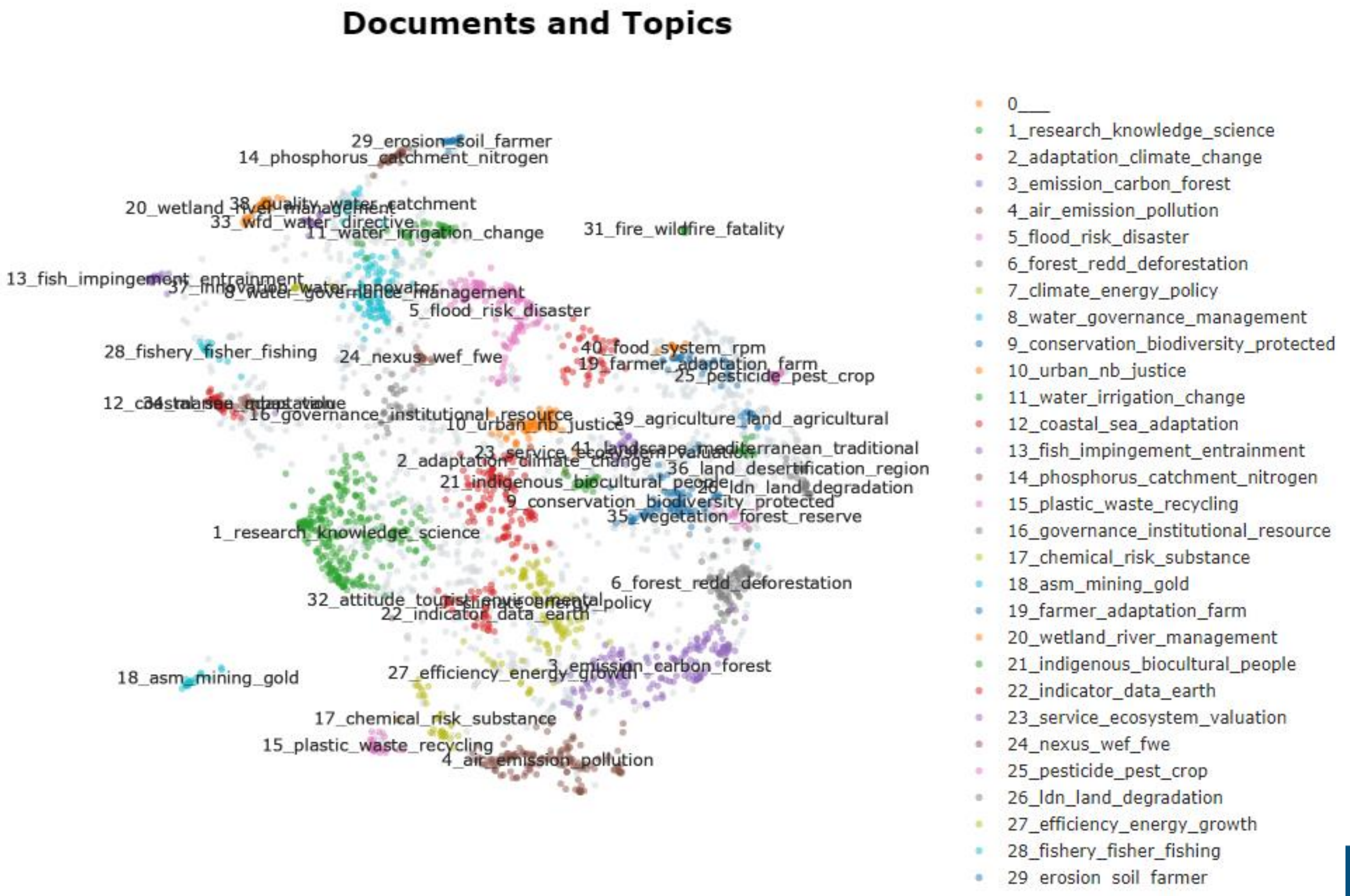

```
topic_model = BERTopic()
topics, probs = topic_model.fit_transform(pp.X.apply(lambda x: ' '.join(x)))
```

- **BERTopic** is an unsupervised topic modeling technique that leverages BERT (Bidirectional Encoder Representations from Transformers)

https://maartengr.github.io/BERTopic/index.html

BERTopic: End to End Process

```
topic_model.visualize_documents(pp.X.apply(lambda x: ' '.join(x)))
```

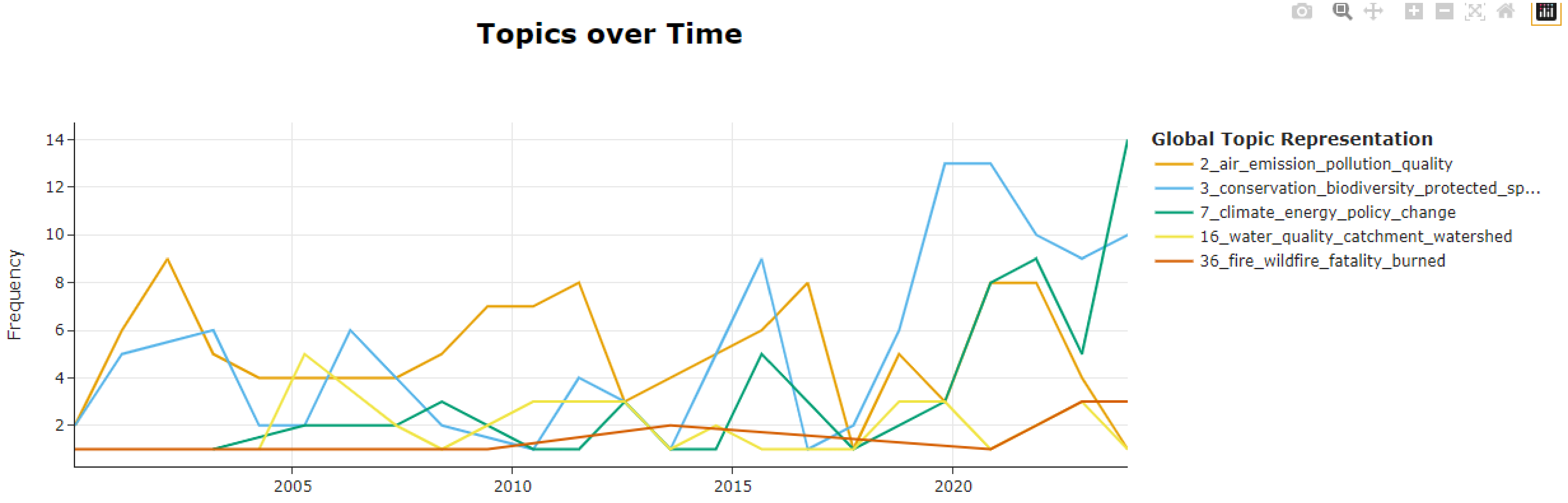


AIR

BERTopic: End to End Process

```
topics_over_time = topic_model.topics_over_time(pp.X.apply(lambda x: ' '.join(x)), data.pub_date, nr_bins=24)

topic_model.visualize_topics_over_time(topics_over_time, topics=[2, 36, 16, 3, 7, 24])
```



Generative Models to Identify Topics

- Generative models excel at both embedding contextual information into vector forms and generating language by predicting next word in a sequence.
- Both capabilities can be utilized for identifying topics in abstracts
  - Prompting : Guide model in generating relevant topics
  - Few shot learning : Starting with examples that demonstrate the task

Generative Models to Identify Topics

- Prompting for few shot learning

```
def build_prompt(abstract):
    return [{
            'role': 'system',
            'content': '''You will be given an abstract of a scientific article, you are to extract a topic for the article given its abstract.
                A topic will be few words, following is an example to how to extract topics.

                <abstract>:
                    Research and practice in climate risk reduction often view marginalized individuals through the lens of vulnerability.
                    However, this perspective lacks specificity of which groups and needs should be incorporated, features narrow wealth-based conceptualization and provides
                    insufficient operationalizable guidance for planning and implementation. This study highlights the theoretical and practical significance of a
                    functional-based approach. It transcends the apparent differences among social groups, instead identifying their shared activity limitations
                    and associated access and functional needs (AFNs) amid climate hazards. Those social groups generally include but not limited to people with
                    disabilities, limited language proficiency, restricted mobility and economic disadvantage, pregnant women as well as children and seniors.
                    We combine quantitative and qualitative analysis to investigate how and why local governments incorporate AFNs in their climate risk reduction.
                    Based on hazard mitigation and climate adaptation plans across local governments in California, our results show that AFN inclusion is consistently
                    predicted by AFN incorporation in higher-level plans, rather than the presence of AFN populations. Besides, plans embracing the functional-based approach
                    achieve greater comprehensiveness and depth of AFN inclusion. We further highlight the commonalities and differences between the two types of plans and
                    conclude with strategic and operational implications for risk reduction efforts.
                    Please only respond with the topic, nothing else
                <topic>: climate, adaptation, change
                '''
        },
        {
            'role': 'user',
            'content': f'''<abstract>: {abstract}
<topic>:
'''
        }]
```

Generative Models to Identify Topics

- May be slower than other processes depending on how busy the model's processing system is
  - We use AZURE OpenAI GPT3.5
  - 22mins for all 3000+ abstracts

Evaluation Approaches

- Each task requires a tailored evaluation approach

TASK	Example Evaluation Approach ML	Evaluation Approach Language Models
CLUSTERING	Silhouette Scores, Davies-Bouldin Index	Human Evaluation of Topic Coherence
TOPIC MODELING	Coherence Scores	Human Review
CLASSIFICATION	Accuracy, Recall and Precision, AUC of ROC	Accuracy, Recall and Precision, Few-Shot Prompt Performance

NLP Pipeline for Today's Example: Summary

Data Acquisition

Documents (text)

Data Preprocessing

Data Cleaning

Data Transformation
- Tokenization
- Stop words
- Stemming
- Lemmatization

Feature Engineering (text to numeric)

Doc	Token1	Token2
1	0.25	1
2	0.66	0

Techniques

Transformer embeddings

Modeling

ML based Clustering
BERTopic
GPT 3.5

Model Evaluation

ML Based Clustering
- Silhouette Score

Breakout II

What methods would you use to answer these questions? What did you learn in training?

What implications would your results can or cannot have?

Would you foresee any resistance to the results from this exercise?

Challenges and Pitfalls

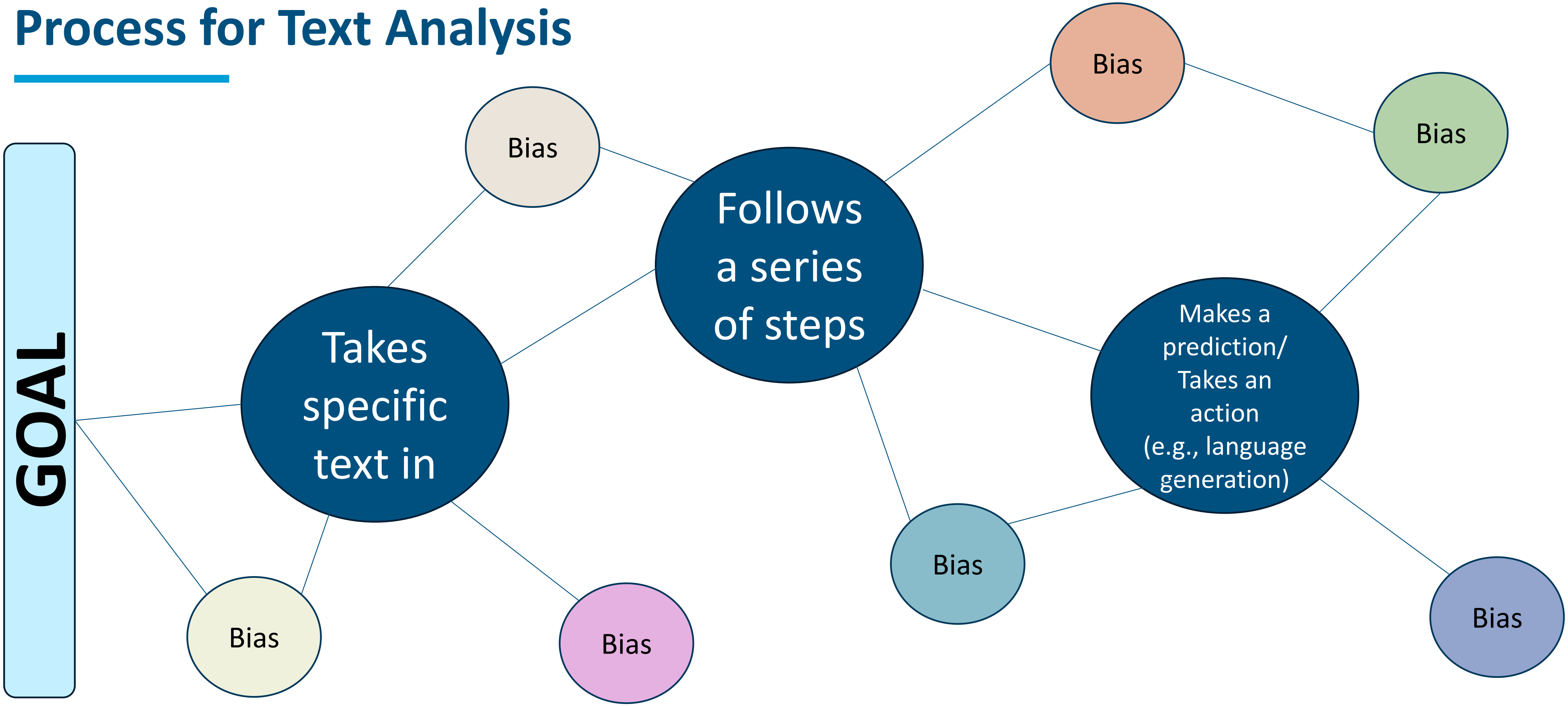
- List of positive applications of emerging technology is long
  - Economic development and poverty
  - Governance
  - Work and meaning
  - Education
  - Health and more
- Why to focus on risks?
  - They are only standing between our good intentions and good outputs

Bias and Risks

AI systems capable of understanding and generating human language by processing vast amounts of text data (definition by IBM)

AI systems capable of creating convincing text by processing a vast amount of the history of humanity, their judgments, their beliefs, and their priorities.

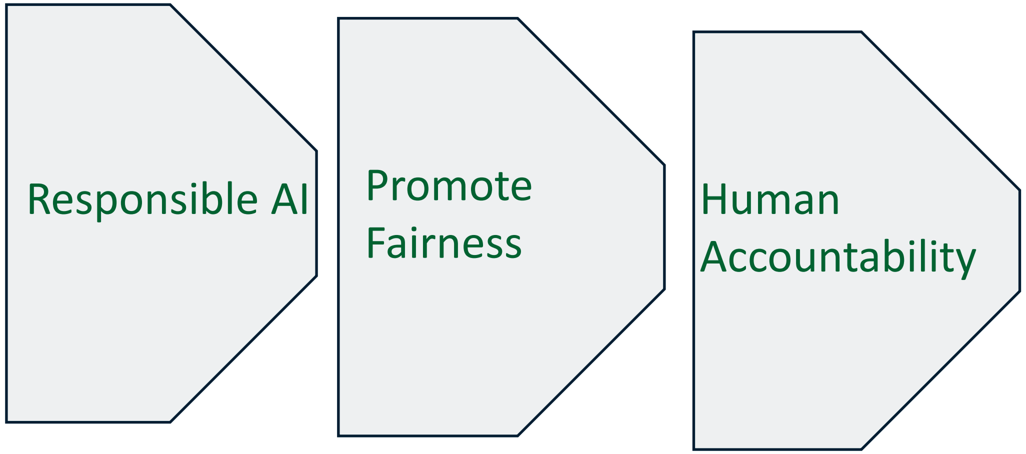
Process for Text Analysis



AIR

Highlighted Needs

- Available dataset (e.g., abstracts)
- Reliability, consistency, and replicability of large language model based results
- Wide gap in using language models
  - Domain specific fine-tuning
  - Bias mitigation



A Good AI Use vs. A Bad AI Use

Superman vs. Homelander

Wonder Woman vs. Harley Quinn

THANK YOU!

rcirci@air.org

babeysinghe@air.org