



# AI IN EVERYDAY LIFE

## Unit 5 – Natural Language Processing



UNIVERSITÀ DEGLI STUDI  
DI TRENTO  
Dipartimento di Ingegneria  
e Scienza dell'Informazione



**DataScientia**  
Unitas per Varietatem



OPEN  
UNIVERSITY OF  
CYPRUS

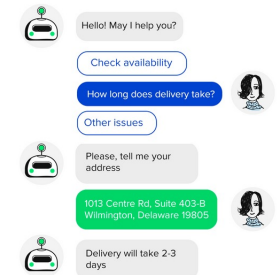
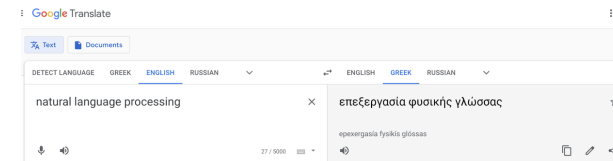


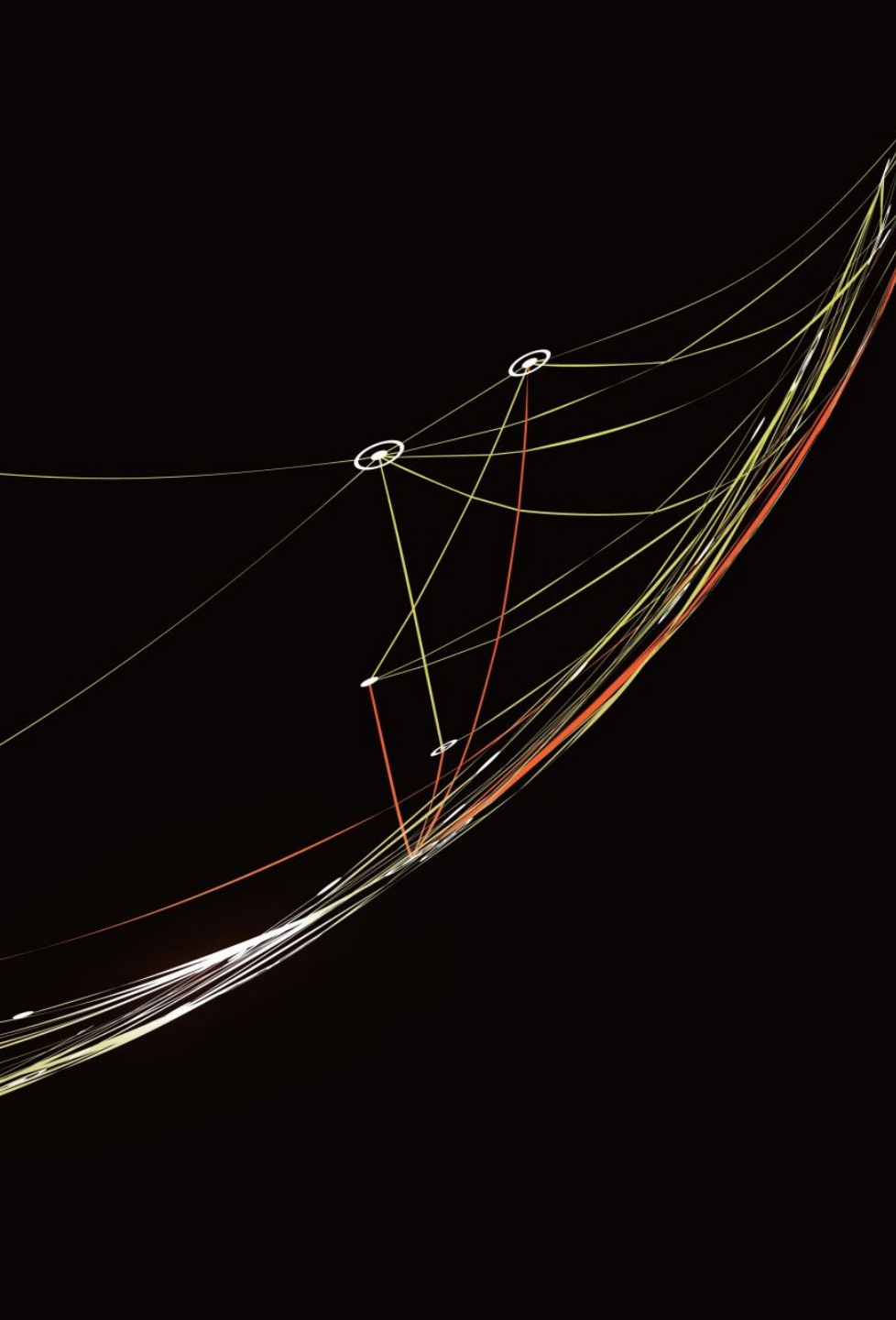
cy. center for  
algorithmic  
transparency



# OUTLINE

- What are the general goals of NLP?  
Some common tasks?
- How do everyday applications use NLP?
- What are some of the benefits and some possible drawbacks?





# WHAT IS NATURAL LANGUAGE PROCESSING?





# NLTK Tutorial: Introduction to Natural Language Processing

Steven Bird

Ewan Klein

Edward Loper

Revision 1.66, 7 Apr 2005

Copyright © 2005

**This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.**

The single and shortest definition of civilization may be the word *language*... Civilization, if it means something concrete, is the conscious but unprogrammed mechanism by which humans communicate. And through communication they live with each other, think, create, and act.

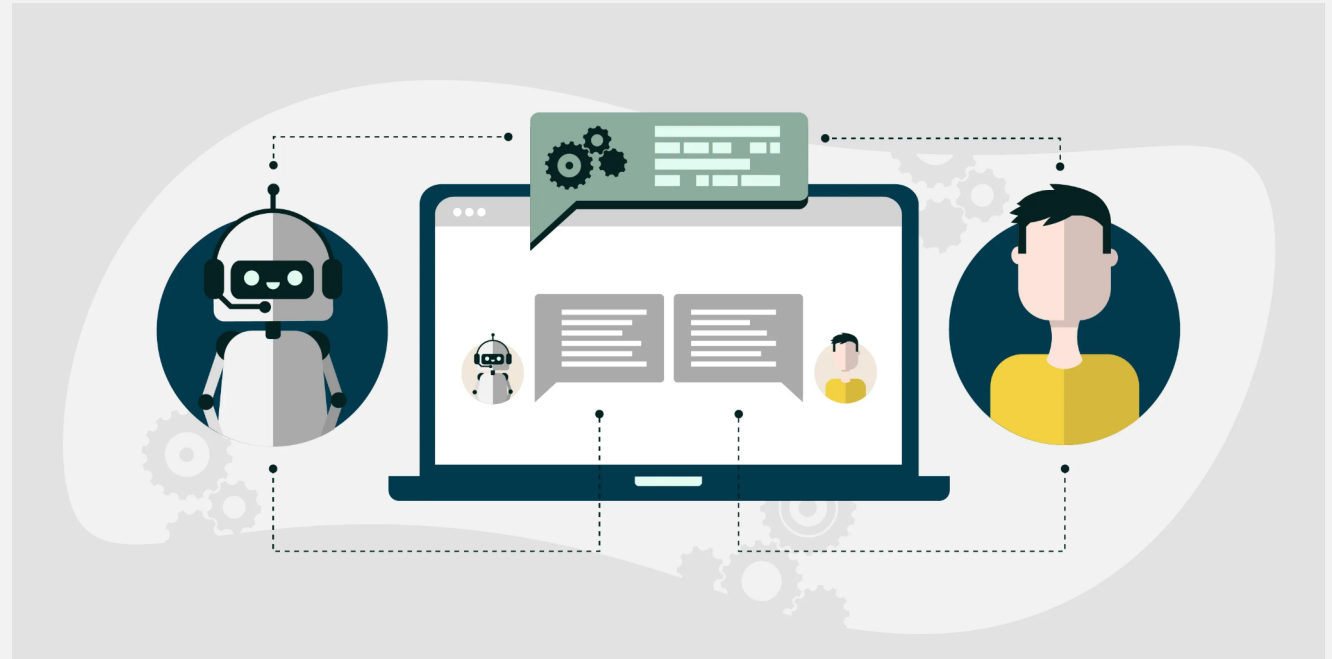
—John Ralston Saul

USEFUL  
RESOURCE



# WHAT IS NLP?

- "Natural language processing tries to build machines that **understand and respond to text or voice data** - and respond with their own text or speech - in the same way that humans do." - IBM



Source: <https://aliz.ai/natural-language-processing-a-short-introduction-to-get-you-started/>



## MAIN TASKS WITHIN NLP

- Text-to-speech
- Speech Recognition
- Machine Translation
- Information Retrieval,
- Extraction and Question
- Answering
- Sentiment analysis



# EVERYDAY AI USING NLP TASKS

## Voice-activated assistants

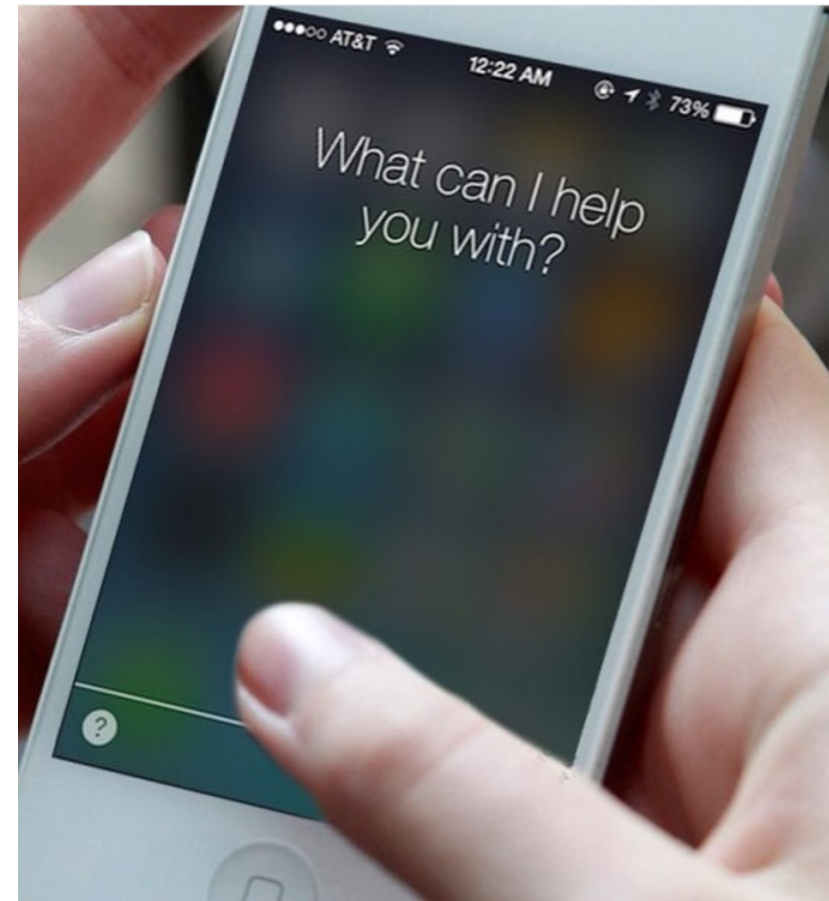
- Text-to-speech
- Speech-to-text

## Web search engines

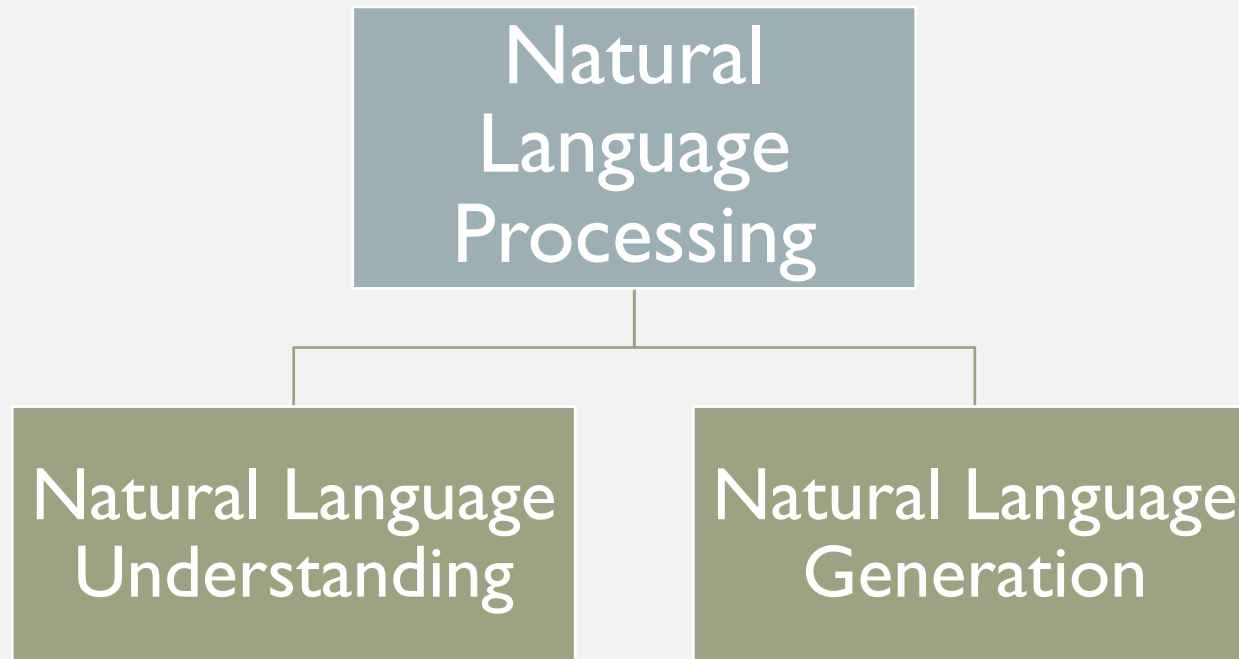
- Information extraction
- Information retrieval
- Question answering

## Chatbots

- Sentiment analysis
- Information extraction



# A SIMPLE TAXONOMY



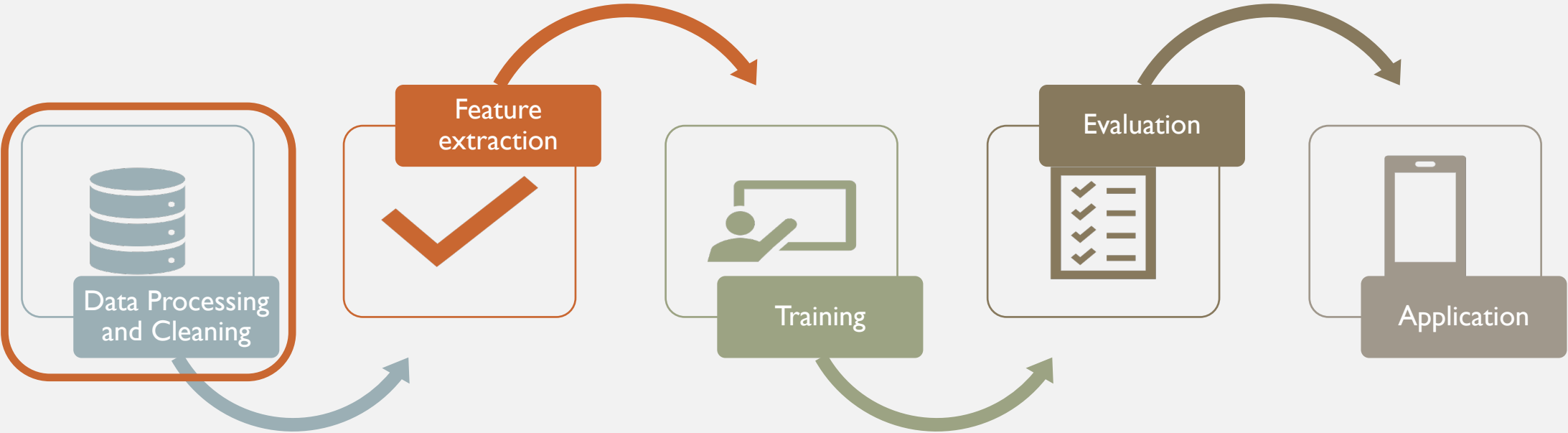




# HOW DOES IT WORK?



# TYPICAL NLP CYCLE



# DATA PREPROCESSING

## Tokenization

- Segment text into sentences or words. Punctuation, numbers and symbols are also removed. Convert capitals to small.

## Stop words removal

- Intentions, links, articles are removed. For example, “and” “the” “a”.

## Stemming

- Process of reducing words by converting them to their root form.

## Word embedding

- Word vectors representing words as numbers. Synonyms have a similar representation.

## TF-IDF

- Frequency (relative) of occurrence of word(s) in a document.



## STEP I: TOKENIZATION

my

dog

loves

to

eat

meat

all

day

the

cat

drinks

milk

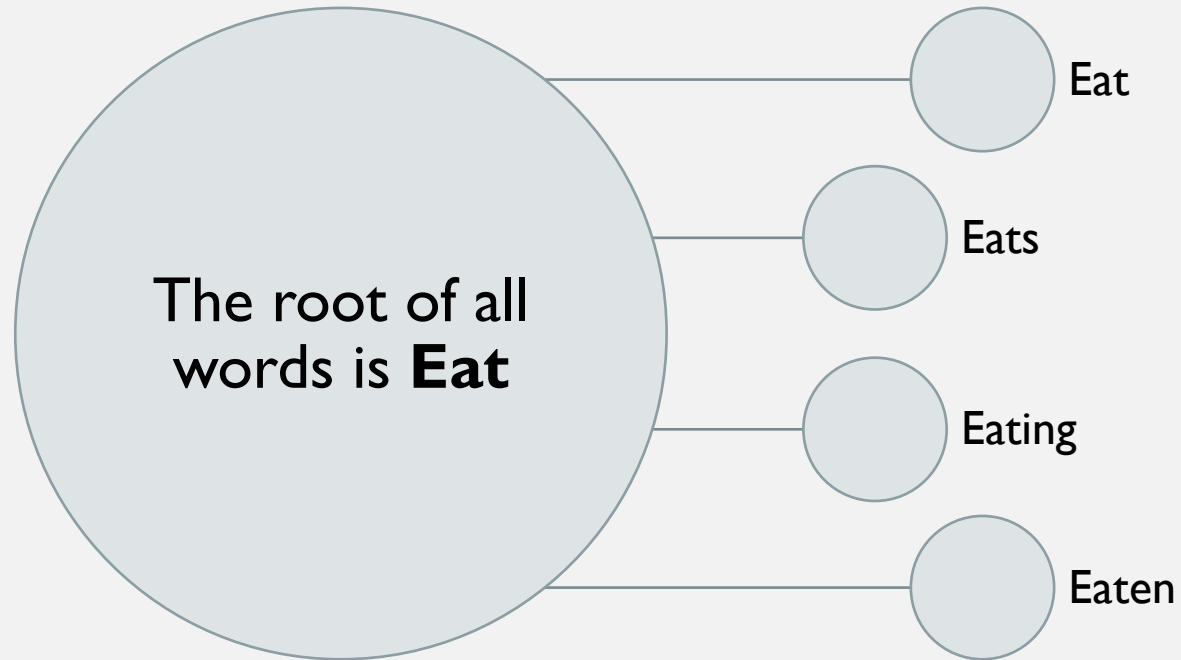


## STEP 2: STOP WORDS REMOVAL

my dog loves to eat meat all day  
the cat drinks milk



# STEP 3: STEMMING



## STEP 4: WORD EMBEDDING

The numbers in the table below show how many times 2 words appear together in the 3 sentences.

- I love Data Science.
- I love coding.
- I should learn NLP.

	I	love	data	science	coding	should	learn	NLP
I	0	2	1	1	1	1	1	1
love	2	0	1	1	1	0	0	0
data	1	1	0	1	0	0	0	0
science	1	1	1	0	0	0	0	0
coding	1	1	0	0	0	0	0	0
should	1	0	0	0	0	0	1	1
learn	1	0	0	0	0	1	0	1
NLP	1	0	0	0	0	1	1	0



## STEP 5: FREQUENCY OF WORDS IN A DOCUMENT (TF-IDF)

We use search engine algorithms to calculate how relevant a document is to keywords.

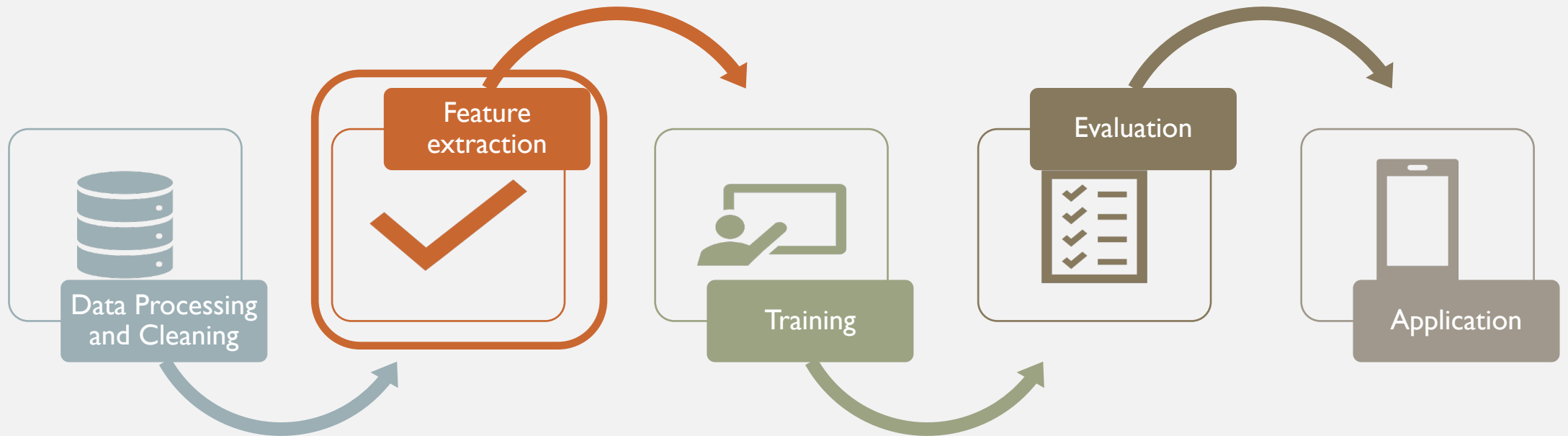
$$TF-IDF = TF * IDF$$

- Term Frequency (**TF**): Calculate frequency of a word/phrase in the document
- Inverse Document Frequency (**IDF**): Calculation of the importance of the specific word/phrase
  - e.g., the words “is”, “are” have no special significance in the text.





# TYPICAL NLP CYCLE



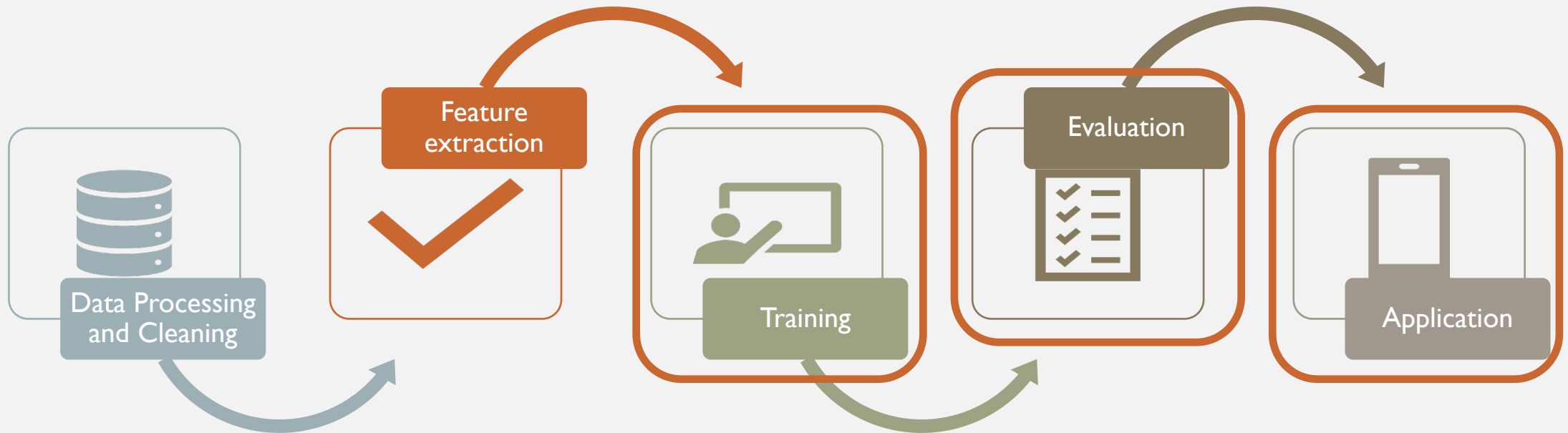


# FEATURE EXTRACTION FOR A TASK

- Topic Modelling
- Sentiment Analysis
- Part-of speech tagging
- Named-entity recognition



# TYPICAL NLP CYCLE

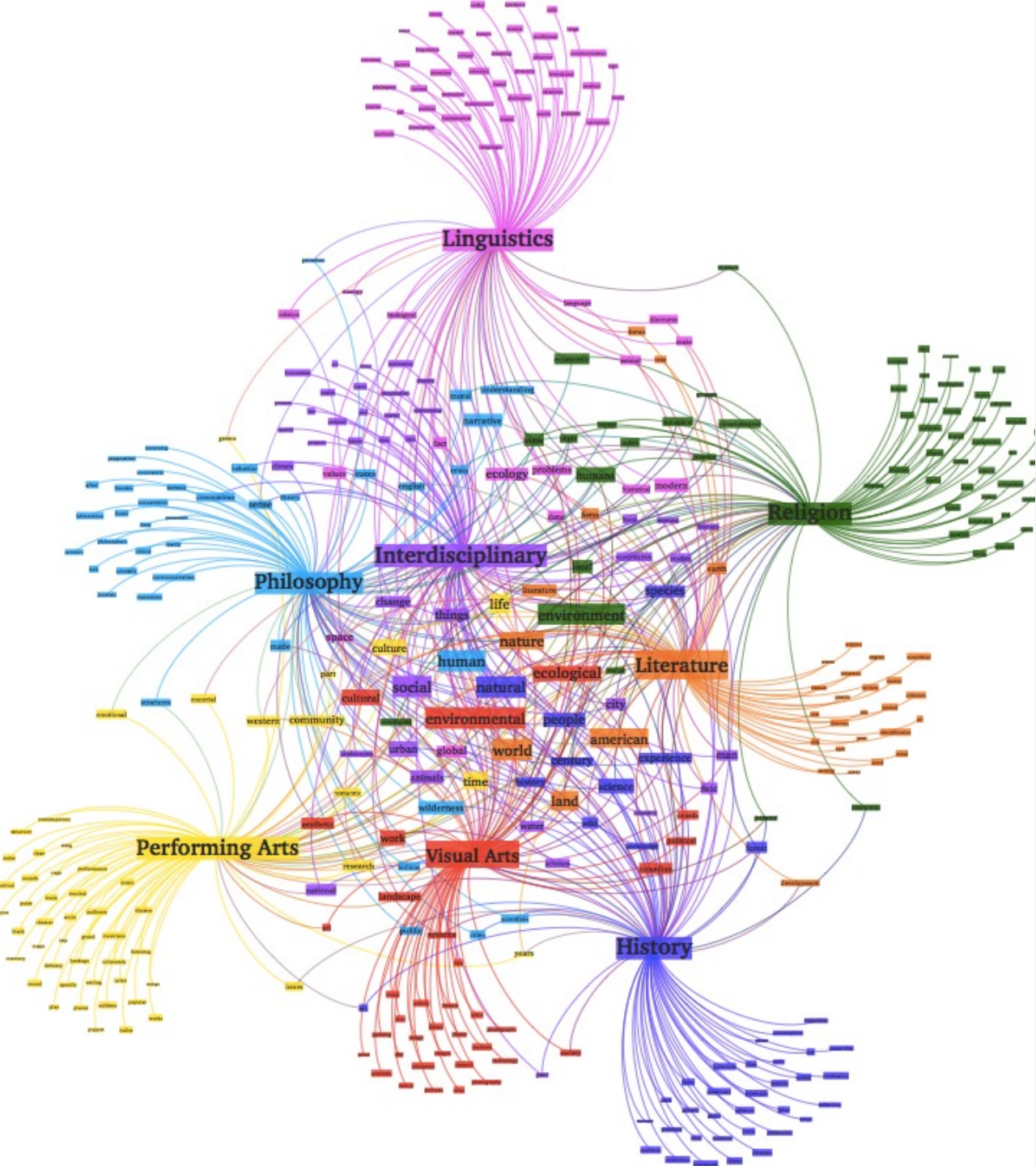


WHERE IS NLP USED?





# TOPIC MODELING




- Extract the main topics from a text or set of texts.
- Each text document is modeled as a statistical distribution of topics and each topic is modeled as a distribution of words.
- Creating features that are useful for training machine learning models for classification.
- Topic modeling is valuable for Hate Speech Detection, as certain topics are more likely to elicit sensitive and/or abusive comments.



# SENTIMENT ANALYSIS

General feeling (polarity) arising from a text (usually an opinion), e.g., positive, negative, neutral.

Emotion detection, such as anger, sadness, and happiness.



Sentiment Analysis

My experience so far has been fantastic!  
POSITIVE

The product is ok I guess  
NEUTRAL

Your support team is useless  
NEGATIVE

MonkeyLearn



## EXAMPLE DICTIONARY - LIWC

ANXIETY

- Nervous, afraid, tense

ANGER

- Hate, kill, pissed

SADNESS

- Grief, cry, sad

POSITIVE EMOTIONS

- Happy, pretty, good

NEGATIVE EMOTIONS

- Hate, worthless, enemy





Filter by Review Date Input field filters

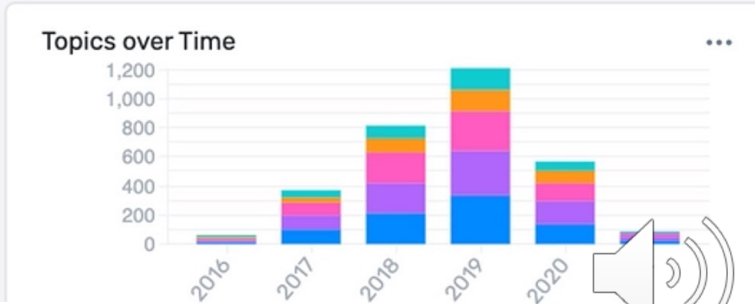
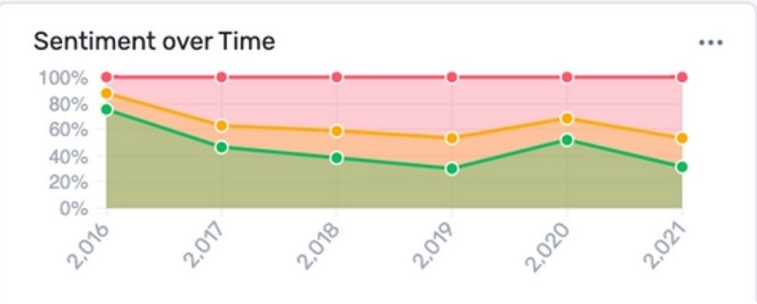
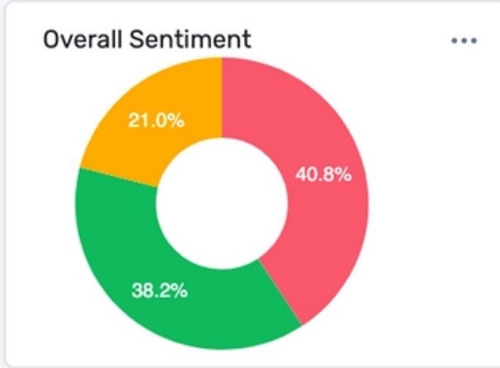
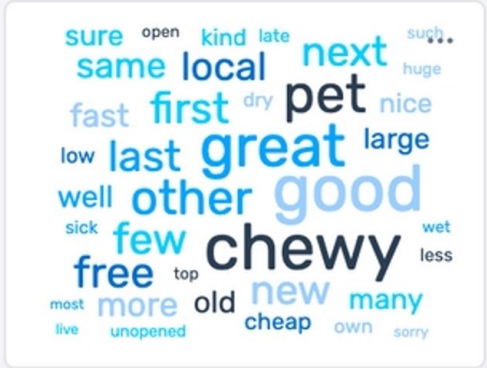
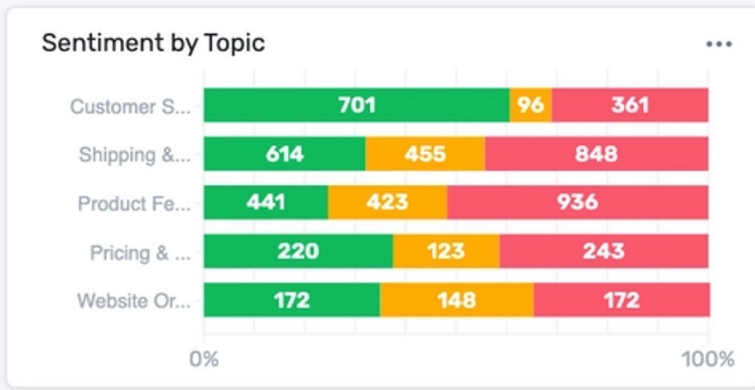
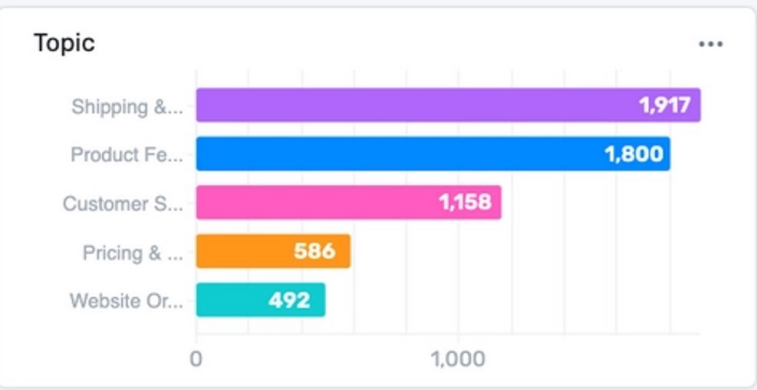
Topic Sentiment Intent Clear All

Share

Search... 5000 samples

rating	Text	Sentiment	Topic
5	and half empty. They replaced my litter for free. <b>Last November my beagle got really sick.</b> My vet said it was arthritis but it was kidney fa	Negative	Product Feedback
1	<b>I tried to see if I could do better buying my dog food through Chewy and the pricing did not even come close.</b> My wet dog food is .70 cheaper at the pet store n	Negative	Pricing & Fees Product Feedback
5	I have been using Chewy for about a year now. <b>I have my dog food shipped to me and have cat food shipped to my daughter.</b> She recently moved and gave me the wrong spelling	Neutral	Product Feedback
1	<b>I was extremely happy with Chewy's prior to the business being sold to PetSmart.</b> Customer Service was efficient and the delivery	Positive	Product Feedback

1-10/5000





# WHAT ARE THE BENEFITS AND DRAWBACKS?



## BENEFITS OF NLP

- Improving human-machine communication.
- Improved services
  - Useful for advertising companies, social networks. Customer support via chatbots.
- Enabling devices
- Speech recognition, useful for smart assistants, e.g., Alexa, Siri.
- Access to information
  - Information extraction and retrieval




## CHALLENGES IN NLP

- **Ambiguity:** The analysis of a word, phrase or sentence is amenable to more than one interpretation, leads to more than one solution.
  - E.g., I hit the thief with the axe. (Was the ax the weapon I used to hit the thief, or did I hit the thief who was holding the axe?)
- **Multiple senses:** big=large. Big sister (older)  $\neq$  large sister
- **Personality, different ways of expression**
  - “This topic is not important” “This topic is meaningless”
- **Emotions and style:** Some use irony and sarcasm to express themselves.




## DRAWBACKS IN NLP

DETECT LANGUAGE **ENGLISH** FRENCH SPANISH   **SWAHILI** HUNGARIAN ITALIAN

This rice is tasty. 

Mchele huu ni kitamu.

 'The rice is tasty' was wrongly translated to 'this uncooked rice is tasty' in Swahili. Photograph: Google Translate



## LINKS AND CONTACTS



<https://datascientiafoundation.github.io/datascientia-education-eai-2023-24-unitn>



<http://knowdive.disi.unitn.it/>



[@knowdive](#)



[matteo.busso@unitn.it](mailto:matteo.busso@unitn.it)

# THANK YOU!

