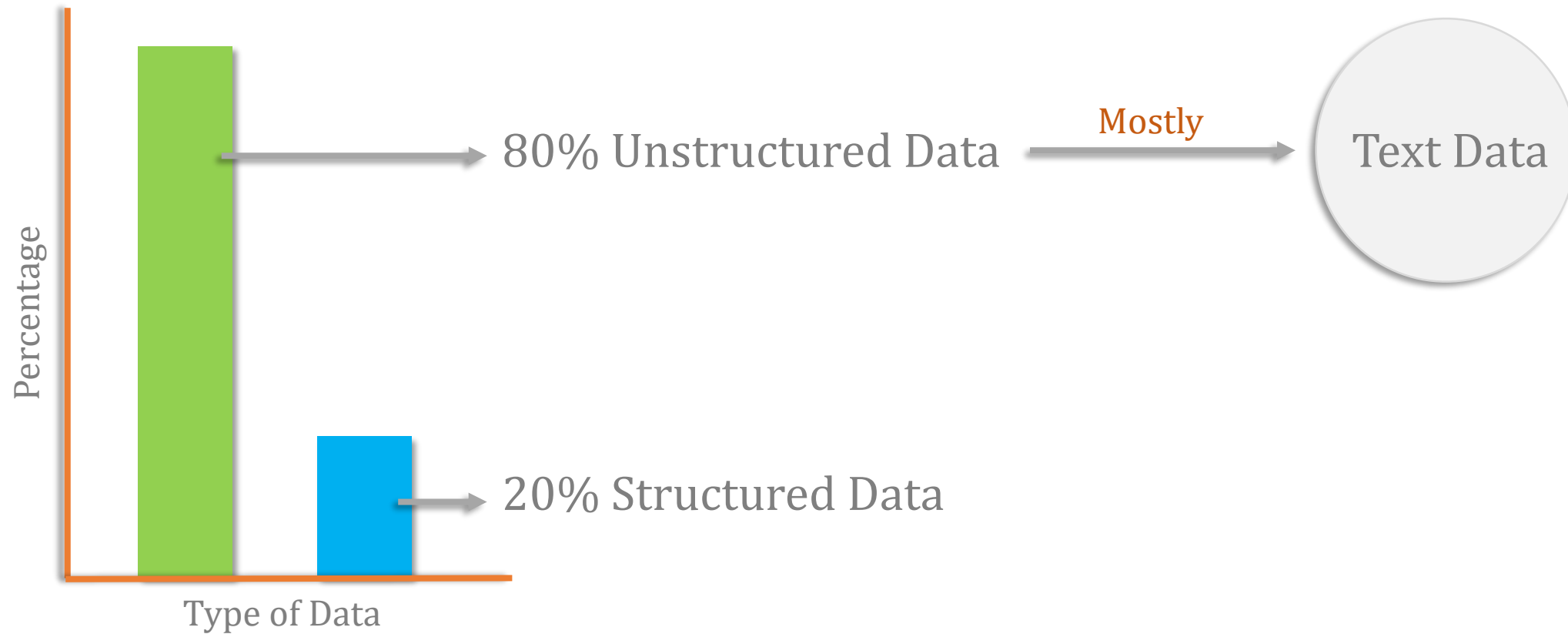


Natural Language Processing (NLP)



Do You Know ?



Textual form is: **Highly Unstructured In Nature**

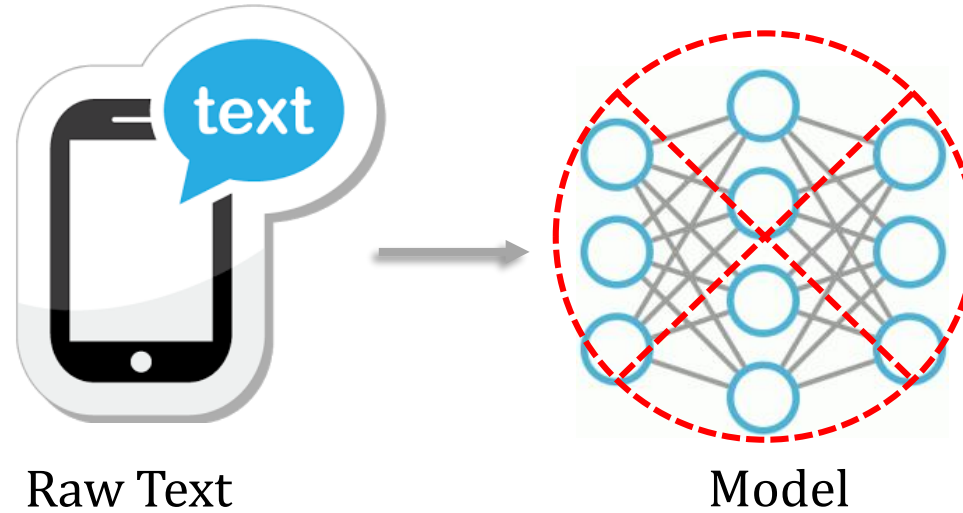
What is NLP?

NLP is a way for computers to **analyse, understand, and derive meaning** from **human language**



Tasks in NLP

- Automatic Summarization
- Machine Translation
- Named Entity Recognition
- Relationship Extraction
- Sentiment Analysis
- Speech Recognition
- Image Caption Generation
- Language Modeling
- Topic Segmentation etc.



Straight from raw text to fitting a ML or DL model will never work

Steps

1. Text Cleaning
2. Prepare Data for Modelling or Feature extraction
3. Modelling

Steps are dependent on
your NLP task
(Refer Previous Slide)

Text Cleaning

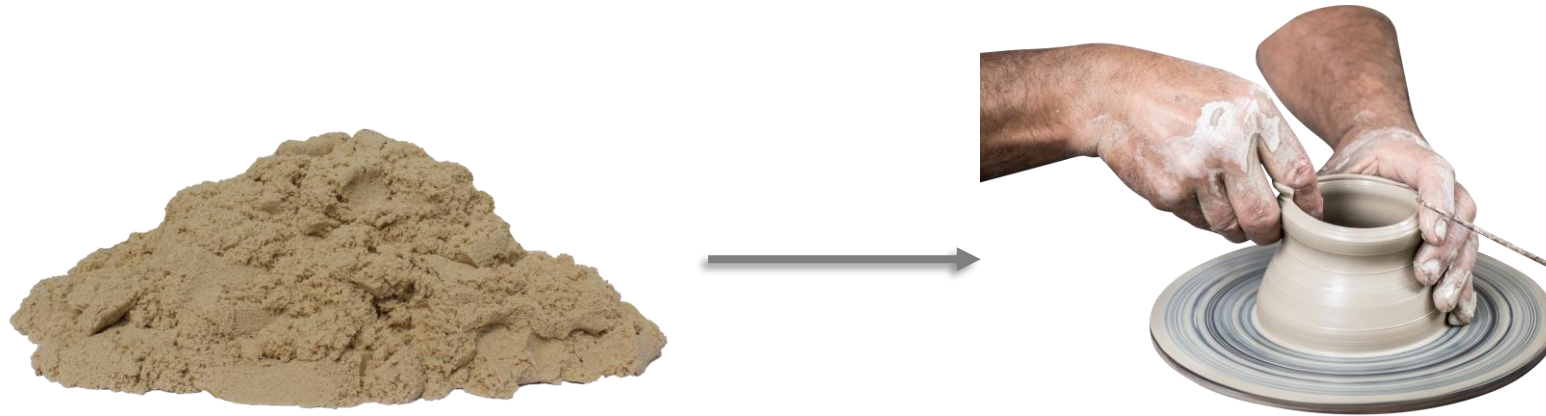
Converting raw text into a list of words or tokens



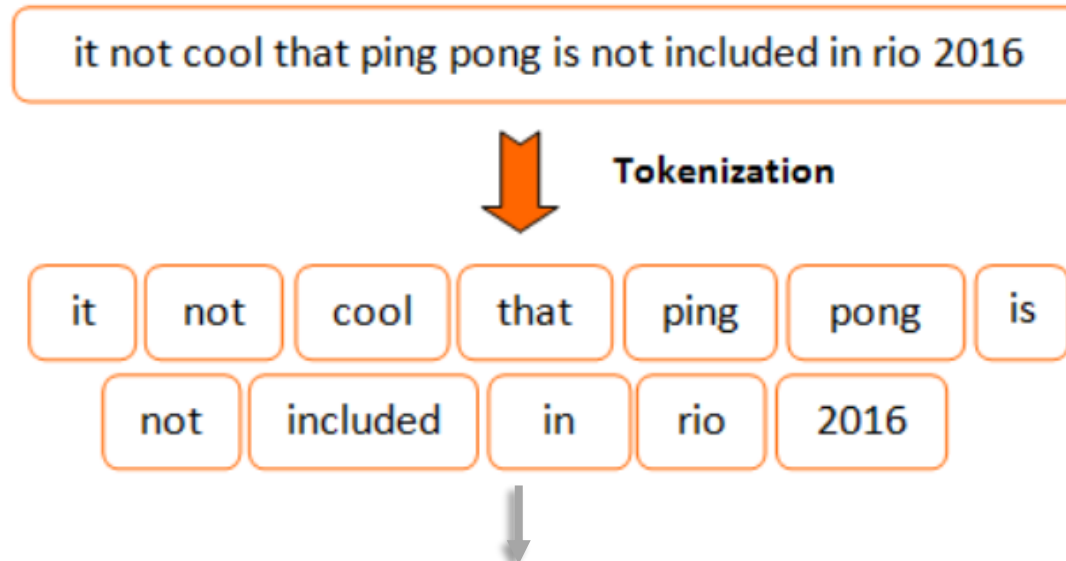
Different methods in Text Cleaning:

1. Split into Sentences
 2. Split into Words
 3. Filter Out Punctuation
 4. Filter out Stop Words
 5. Normalizing Case
 6. Stemming Words or Lemmatization
- Noise Removal

Prepare Data for Modelling or Feature extraction



Special preparation before using it for predictive modelling



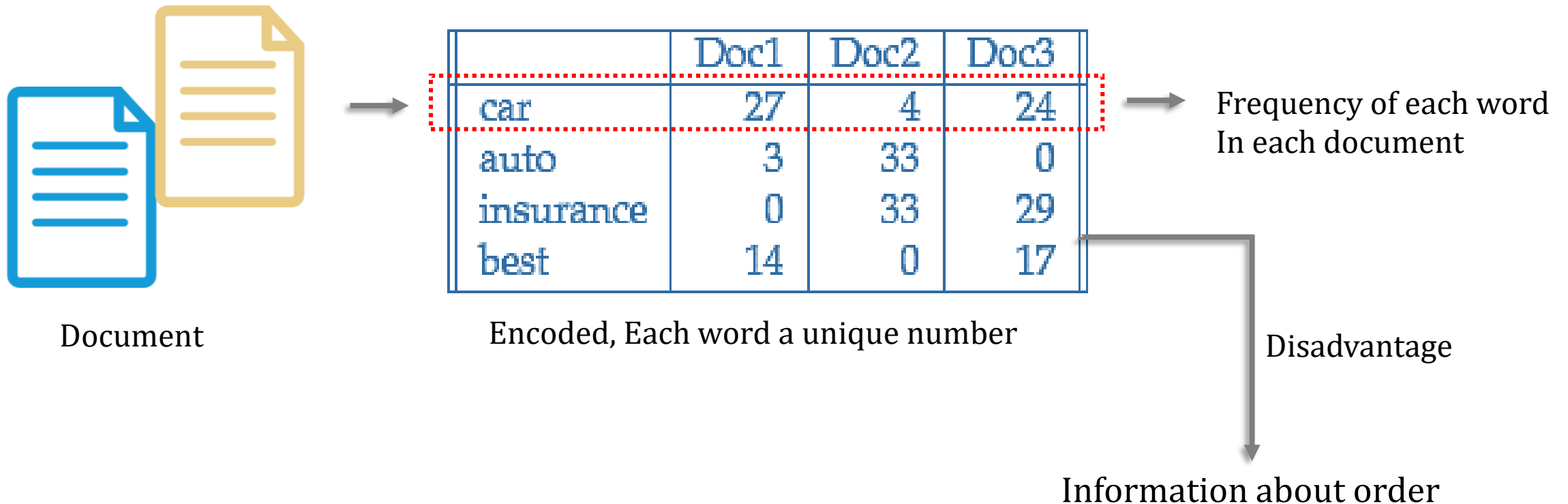
Words need to be encoded as integers

Bag of Words Model

We cannot work with text directly when using ML algorithms

We need to convert the text to numbers

1. Given document, assigns each word a unique number
2. Each position in the vector could be filled with a count or frequency of each word



Word Counts with Count Vectorizer

	!	.	brown	dog	fox	jump	lazy	over	quick	the
Doc1	0	1	1	0	1	0	0	0	1	1
Doc2	1	0	0	1	0	1	1	1	0	1

Contain a lot of zeros, called as “**Sparse**”

Number of times a token shows up in the document



Disadvantage

Simply counts and tell many times they appear

Word Frequencies with TfidfVectorizer

Term Frequency-Inverse Document Frequency

How often a given word appears
within a document

Down scales words that appear
a lot across documents

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	$\log(2/2) = 0$	0	0
Car	1/7	0	$\log(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log(2/2) = 0$	0	0
Driven	1/7	1/7	$\log(2/2) = 0$	0	0
On	1/7	1/7	$\log(2/2) = 0$	0	0
The	1/7	1/7	$\log(2/2) = 0$	0	0
Road	1/7	0	$\log(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log(2/1) = 0.3$	0	0.043

Looks for frequent in a document
but not across documents

TF-IDF are word frequency scores that try to highlight words that are more interesting

Hashing with HashingVectorizer

(For Larger Documents)

Hashing Trick

HashingVectorizer

"You better call Kenny Loggins"

tokenizer

['you', 'better', 'call', 'kenny', 'loggins']

hashing

[hash('you'), hash('better'), hash('call'), hash('kenny'), hash('loggins')]
= [832412, 223788, 366226, 81185, 835749]

Sparse matrix encoding

[0, ..., 0, 1, 0, ..., 0, 1, 0, ..., 0, 1, 0, ..., 0]