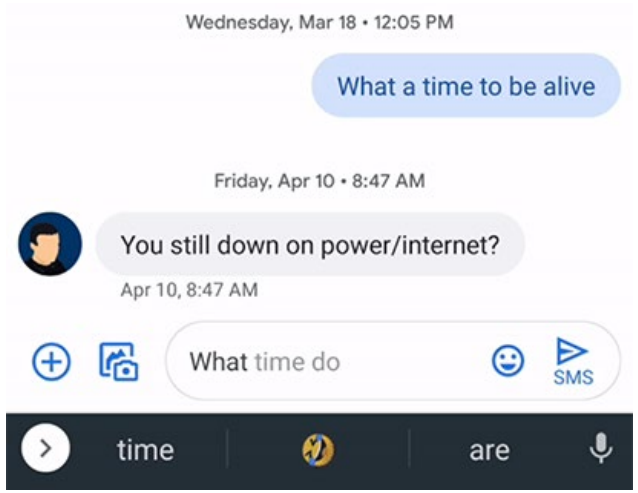
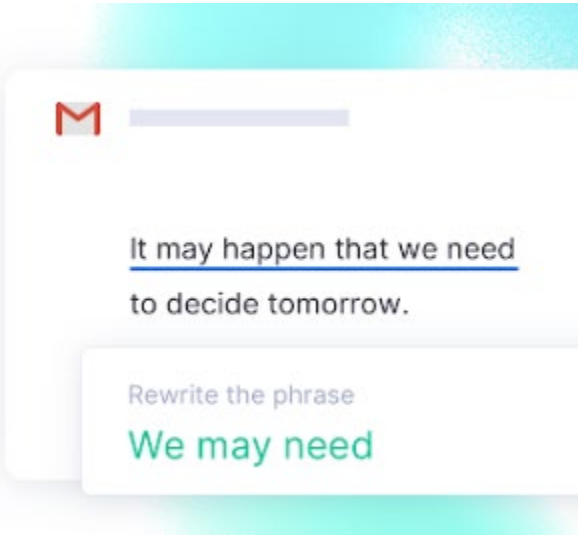
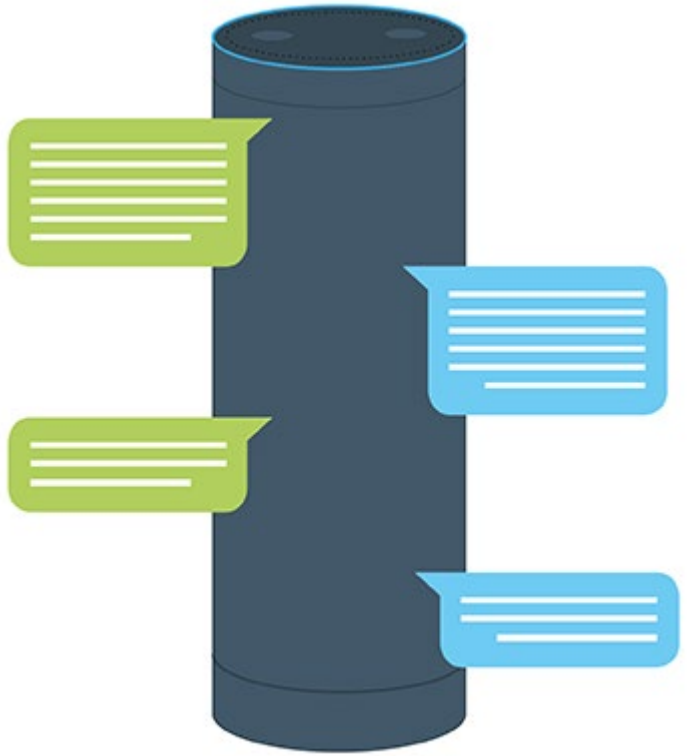
Fundamentals of Text Mining & NLP







Fundamentals of Text Mining

Why should we use text for analysis?

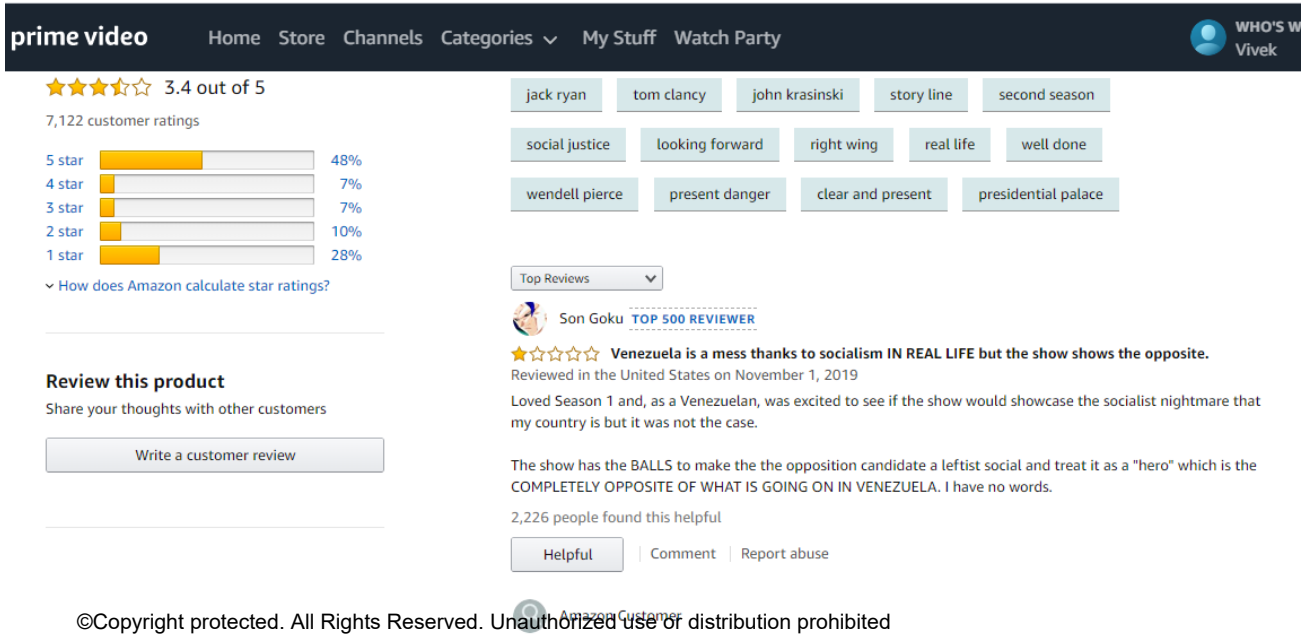
Quantitative vs qualitative insights

Data Sources for text mining & NLP

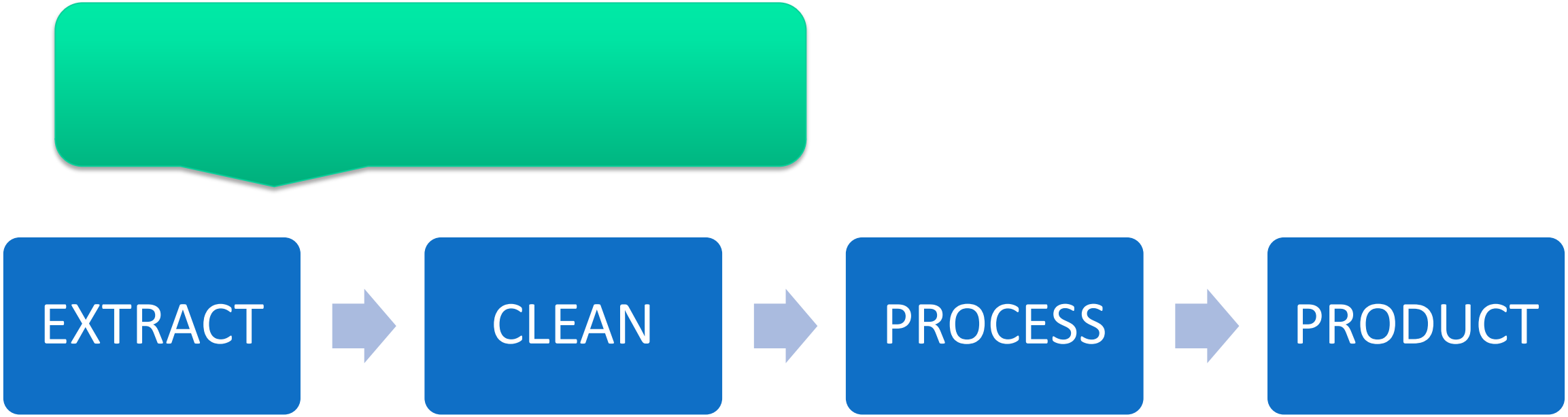
Google

work from home furniture

work from home furniture **chennai**
work from home furniture
work from home furniture **online**
work from home furniture **india**
work from home furniture **wipro**



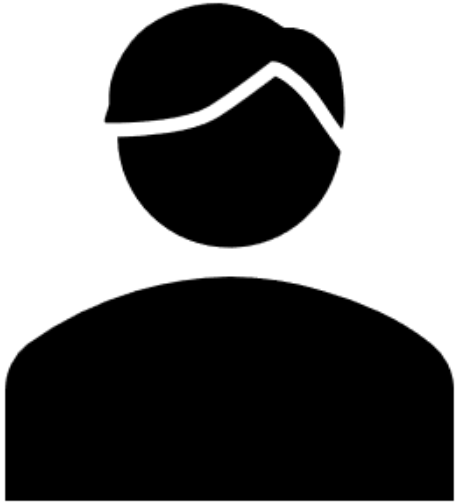
What is Text Mining?



What is Text Mining?

What is NLP?

Computer Language vs Natural Language

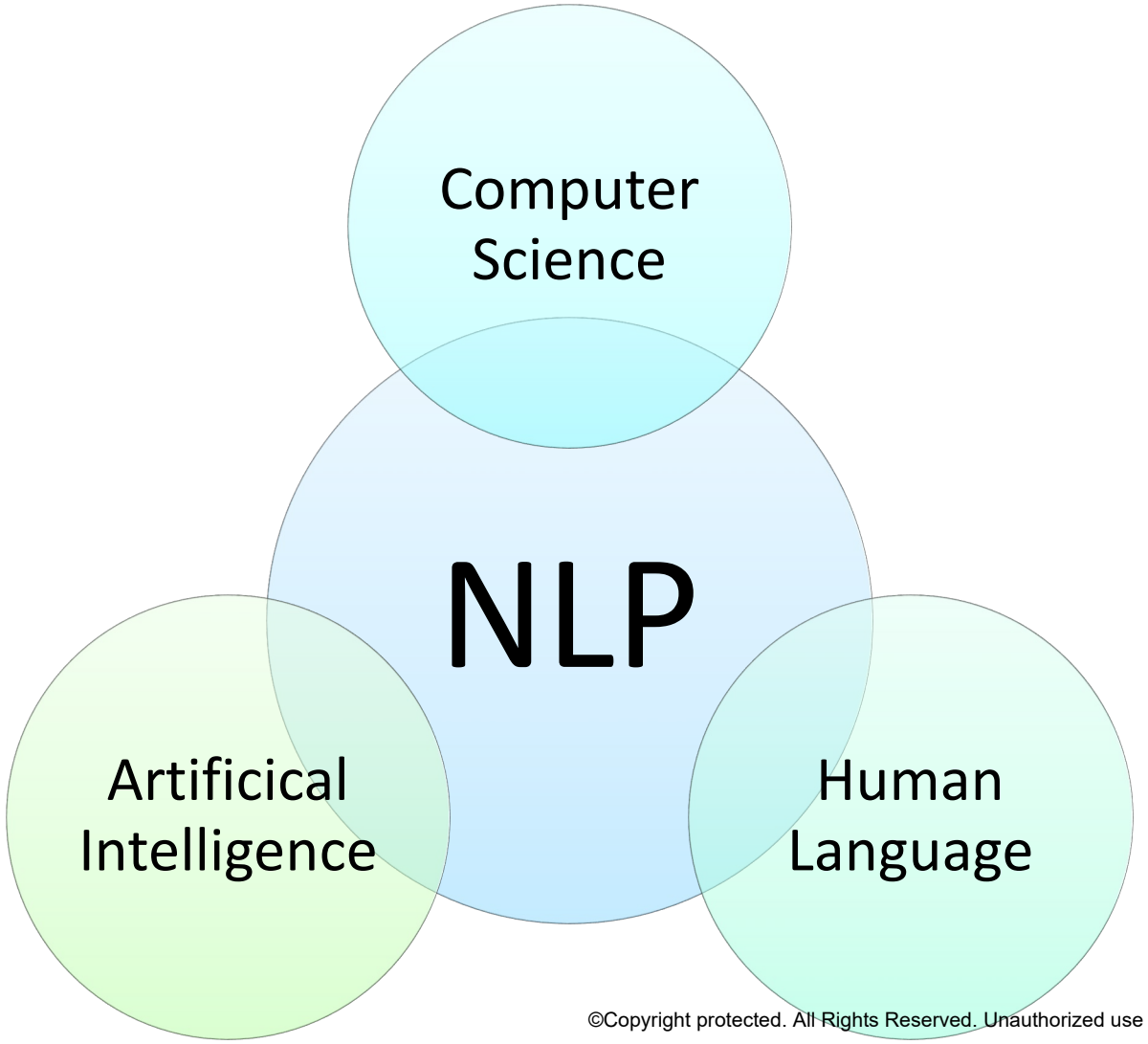


What is NLP?

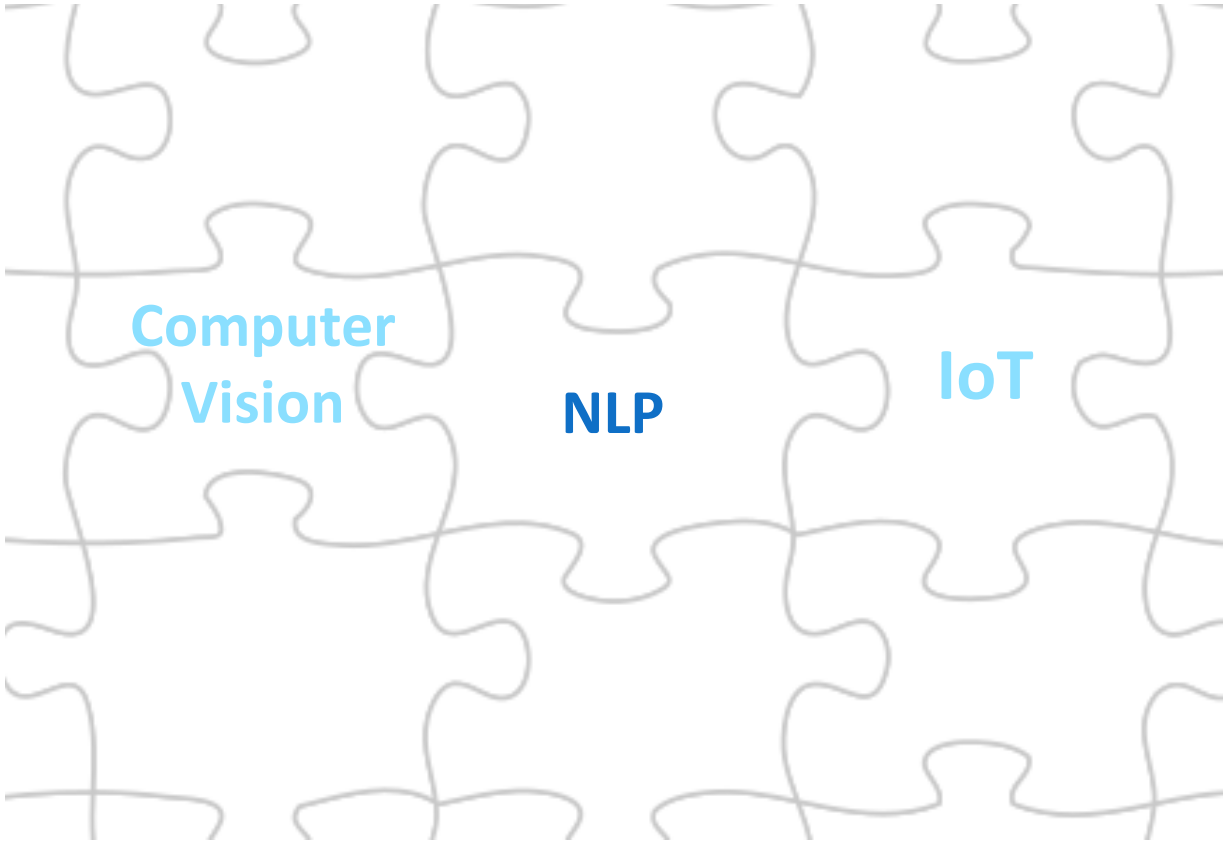
Natural Language Processing

Goal of NLP

What is NLP?

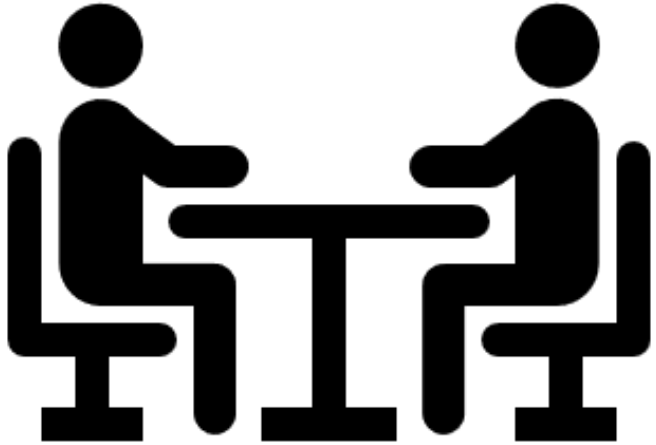


NLP in the AI landscape



NLP Verticals

Big questions of NLP



What is NLP based on

Text mining vs NLP

Text mining vs NLP

Topic	Text Mining	NLP
Goal	Extract important information from unstructured data	process sequential data to understand human language
Tools	Statistical models Machine Learning	Deep neural network
Data source	•documented collections	•Resources can be form of any natural human communication methods like text, speech, signboard, etc
Scope	•Extracting the important features for natural language documents	•extract the semantic meaning and grammatical structure from the input
Outcome	1.Frequency of words 2.Patterns of words 3.Correlation within words	1. Analysing sentiment 2.Language translation 3.Grammatical structure

Challenges associated with NLP

Syntactic vs Semantic Analysis

Syntax

Syntactic analysis

Semantic analysis

Syntax vs semantics

A happy person barks well

Context matters!

Tim likes to have coffee when we wakes up

INTRODUCTION TO LANGUAGE MODELS

Distributional semantics

- when we do not know a word, we can guess what it means by knowing the *context* where it's used
- words that are used and occur in the same contexts tend to purport similar meanings

What is a language model?

Why is language modelling a challenge?

Types of language models

Predict the next word

- I want a _____
- I want a glass of _____ wine juice
- I want a cup of _____ coffee tea
- I want a pinch of _____ salt

Context
2 words to the left

Add more context

- I want a glass of _____ juice
- I want a glass of ____milk____ to go with my cereal
- I want a glass of ____juice____ on a hot summer afternoon

Context – few words to the left and few words to the right

Context – Nearest word on either side

Example of language model – Machine translation

- P(tall man) P(intelligent man)

Example of language model – Spell check program

Please accept my apologies

Please except my apologies

Example of language model – Speech recognition

Hey Alexa, place an order for 2kg of wheat flour

Hey Alexa, place an order for 2kg of wheat flower

NLP METHODS OVERVIEW

Common nlp methods

- Word frequency
- Collocation
- N-grams
- Entity Extraction
- Content classification
- Document comparison
- Topic modelling
- Pattern matching and text similarity

What is word frequency?

What is Topic Modelling?

What is Collocations?

Example of Collocations

Collocations	
Correct	**Incorrect**
• High temperature • Have an experience • Heavy rain	• Tall temperature • Make an experience • Thick rain

What is N-grams?

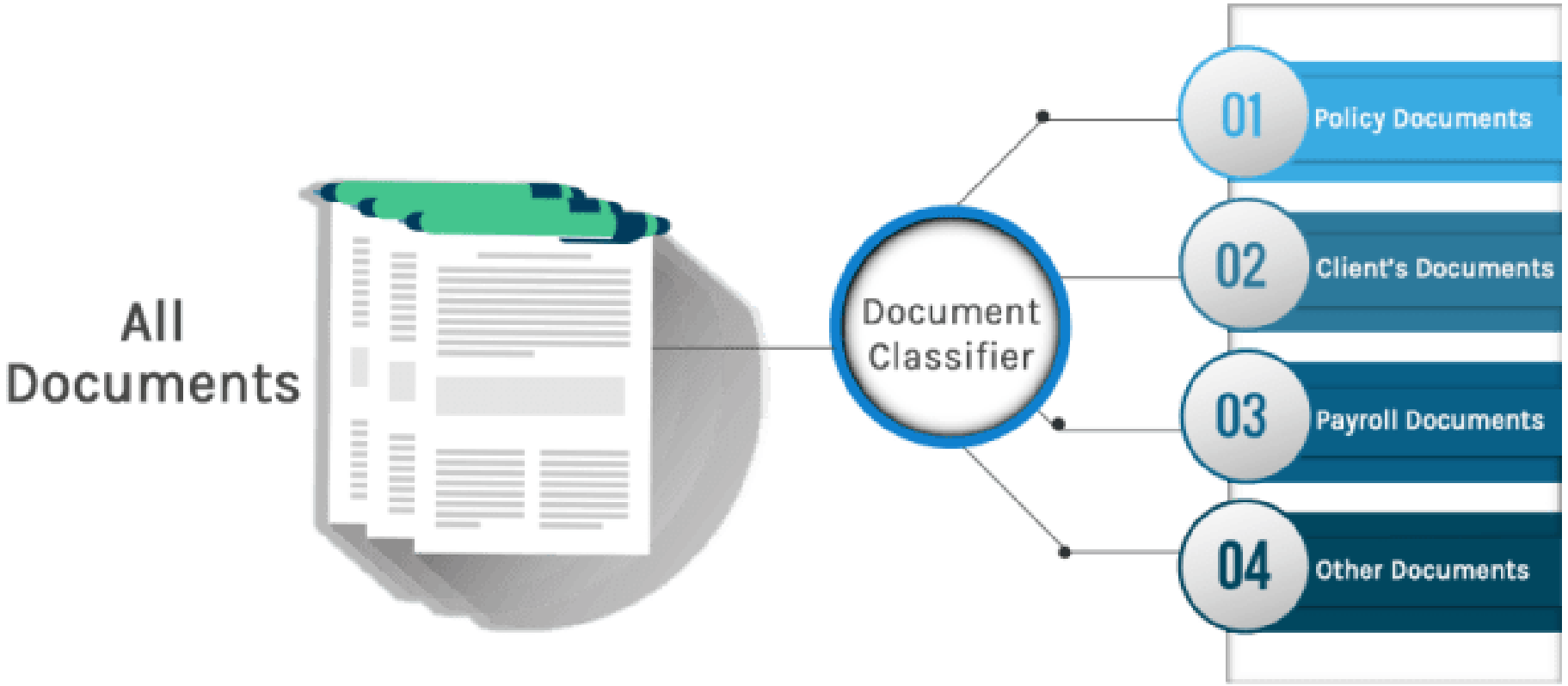
This is a sentence

What is Entity extraction?

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

What is Content Classification?



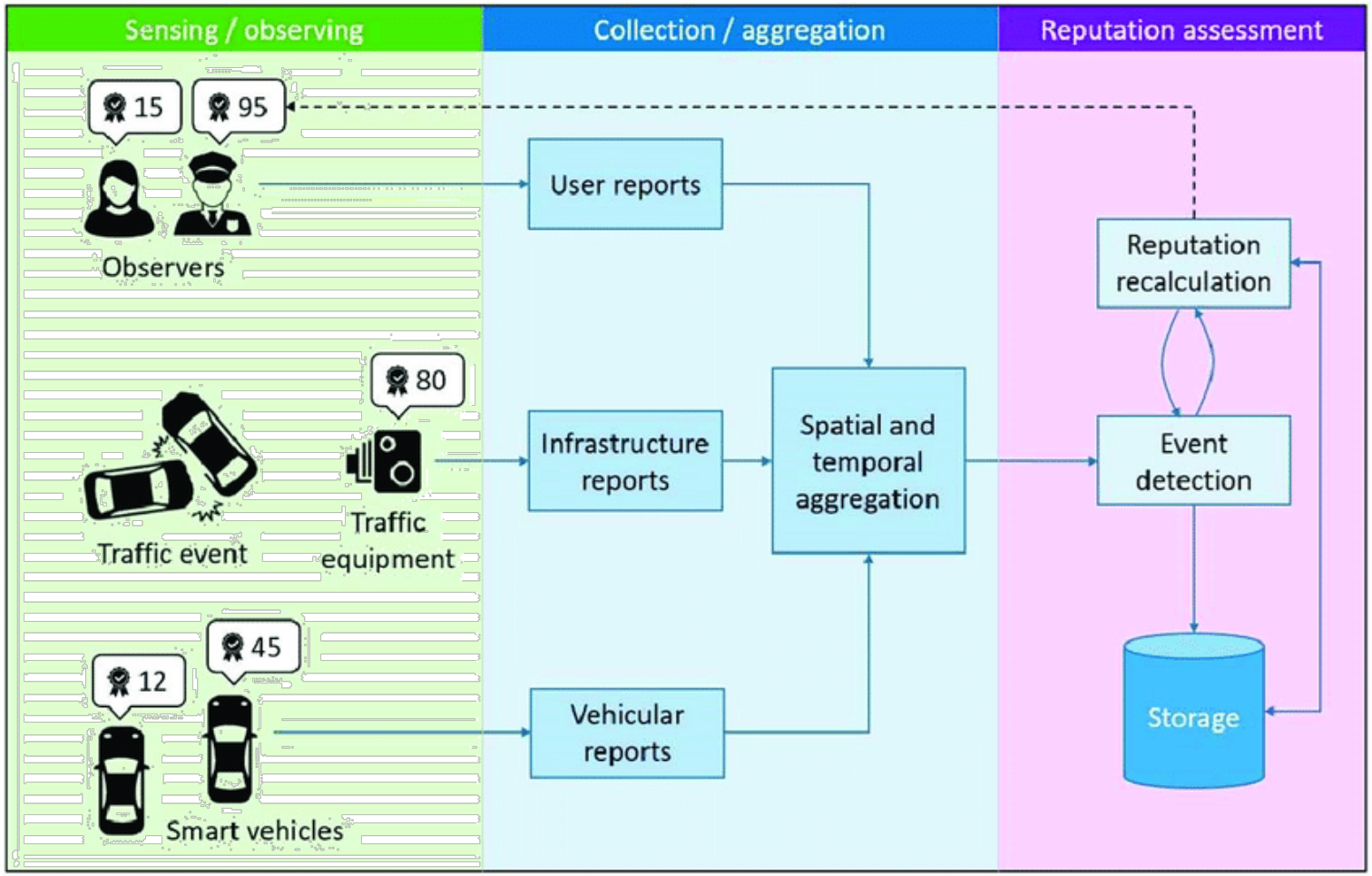
NLP APPLICATIONS

Information Extraction

Spam mail filter

Pattern Matching

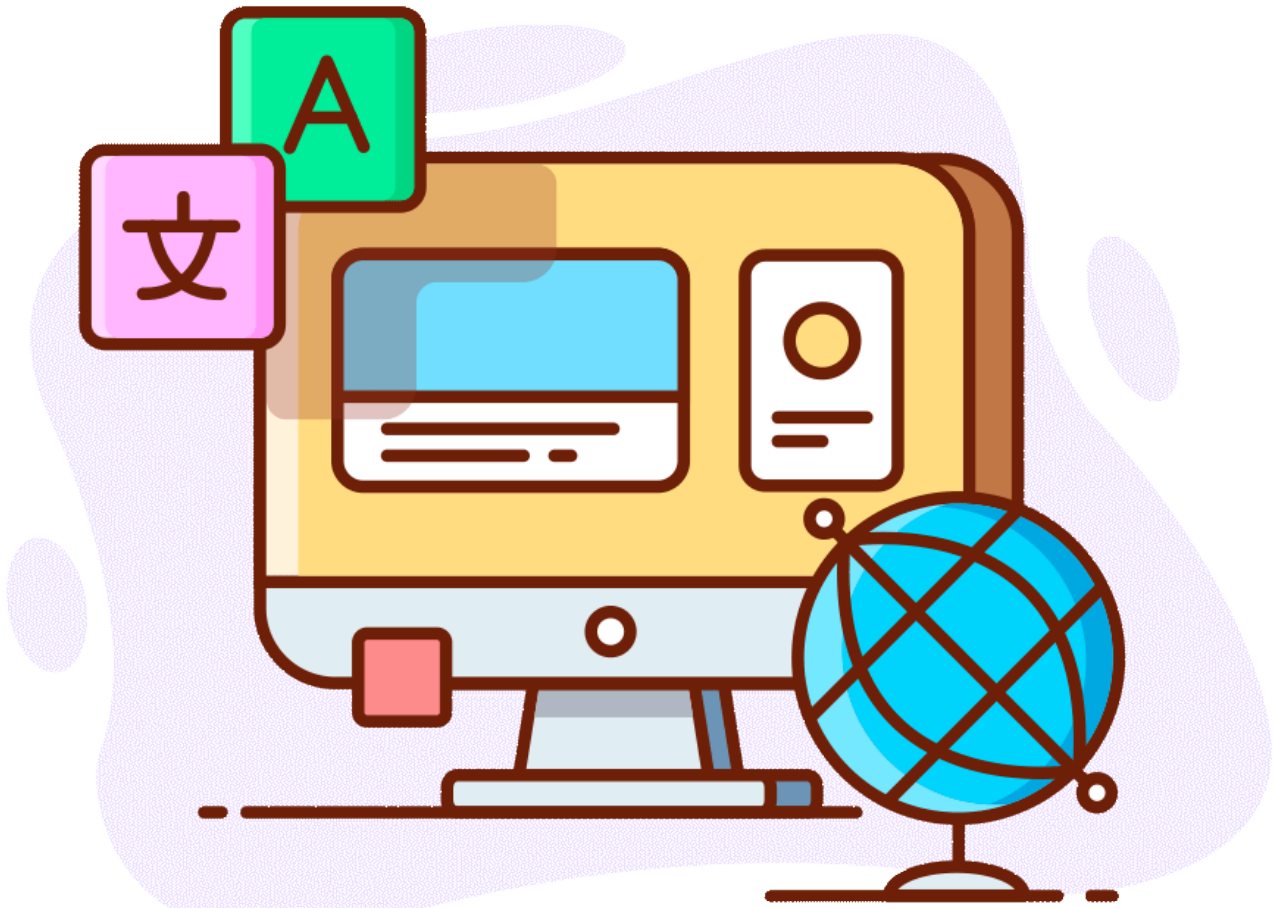
Event detection



Social Media monitoring

Sentiment Analysis

Machine Translation



Speech Recognition

Chabot

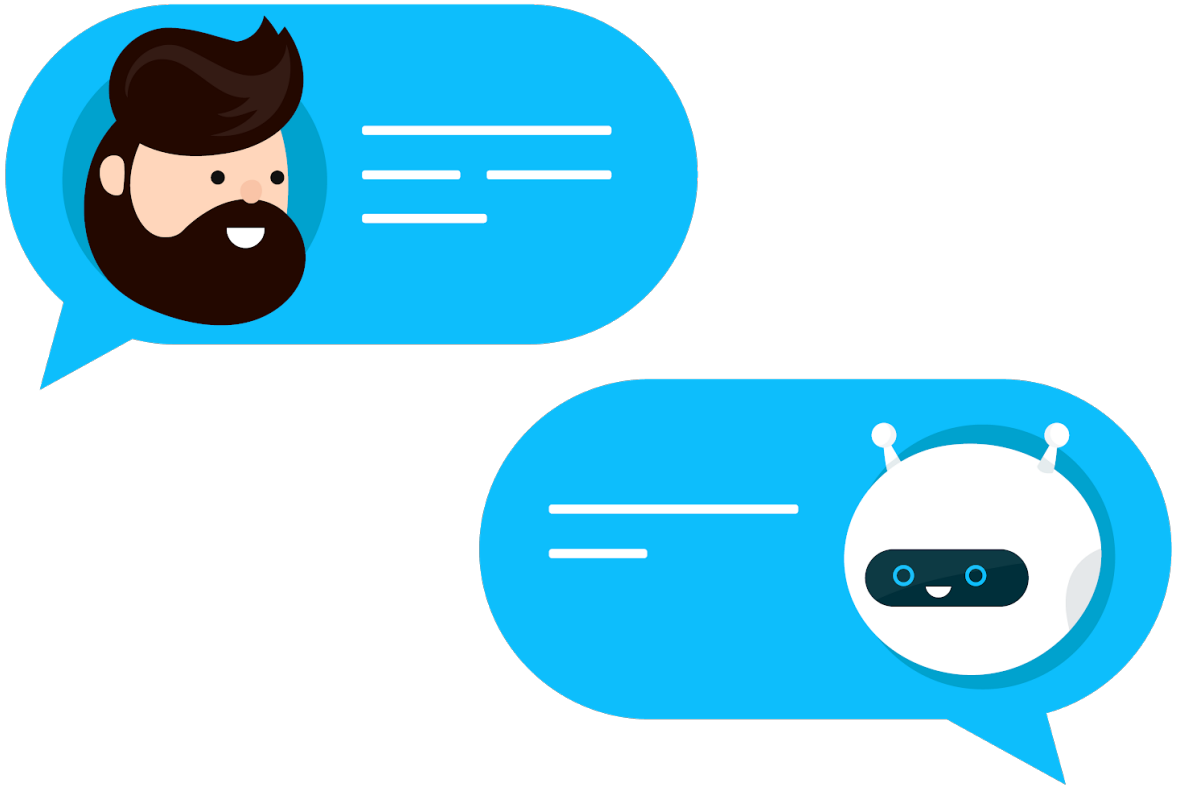


Image caption generator

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

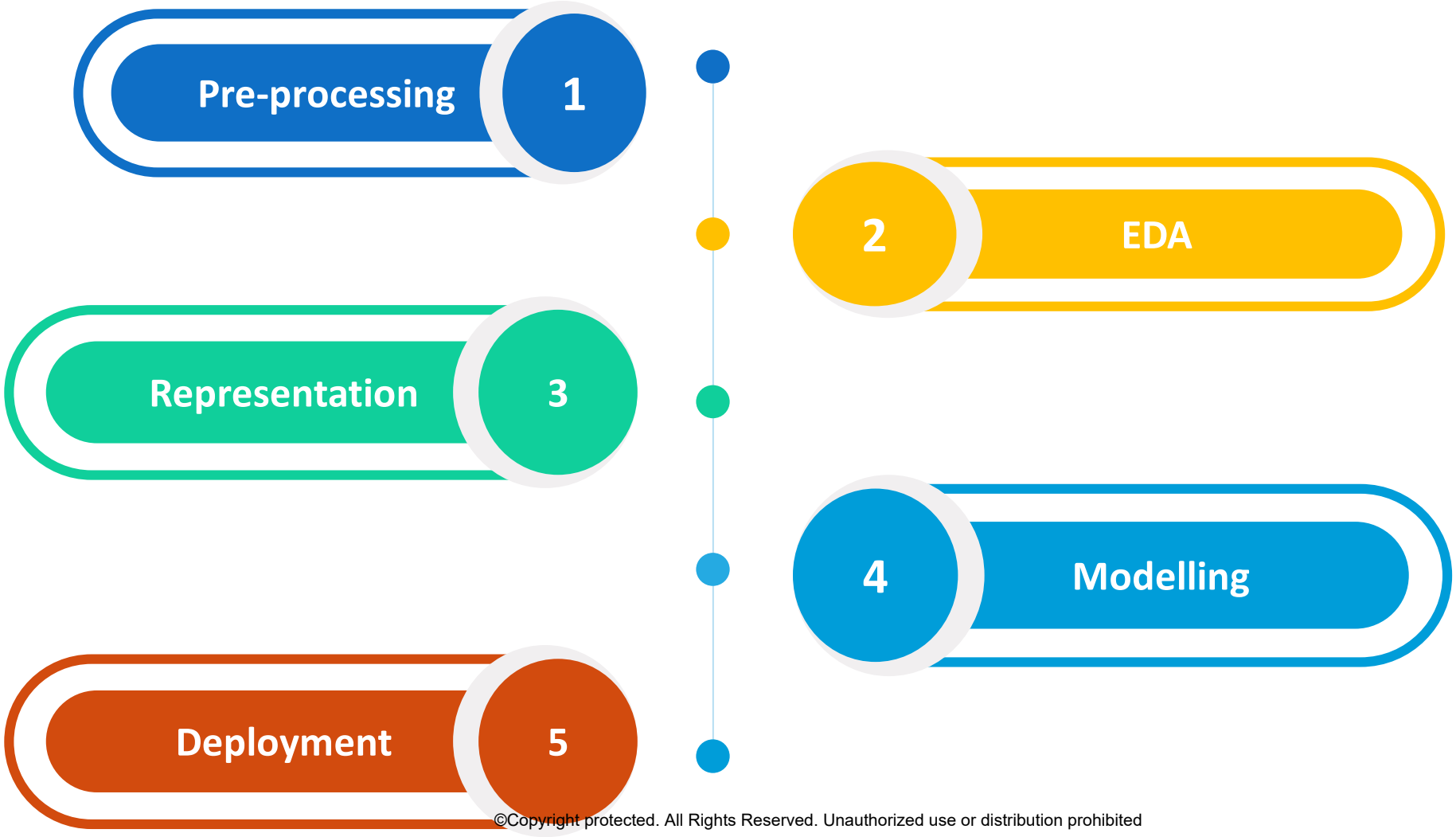
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

NLP Workflow

Broad outline

Workflow of NLP



Workflow of NLP: Pre-processing

Pre-processing 1

Text processing

Workflow of NLP: Exploratory Data Anaıysıs

2 EDA

EDA (Exploratory Data Analysis)

Workflow of NLP: Representation

Representation 3

Representation

Tokenization

Text to sequence

Padding sequence

Workflow of NLP: Modelling

4 Modelling

Modelling

Recurrent Neural network (RNN)

LSTM

Bi-LSTM

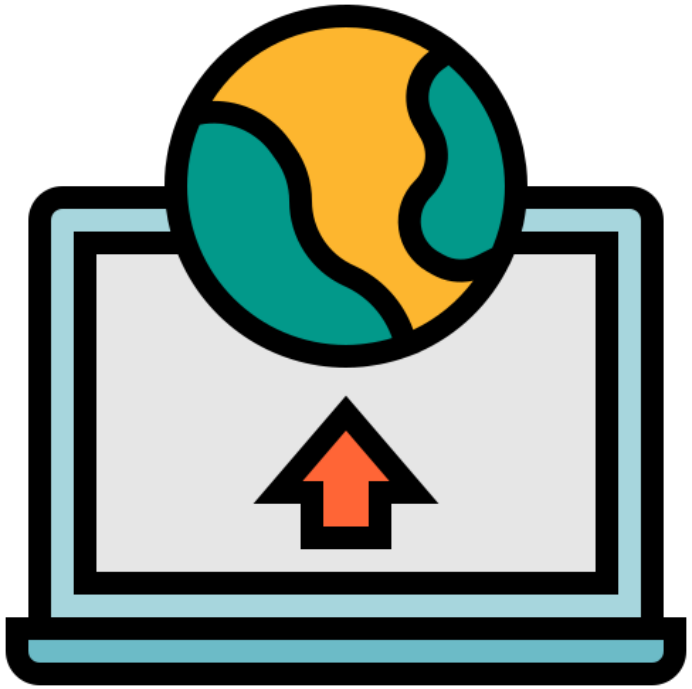
GRU

Deployment 5

Workflow of NLP: Deployment

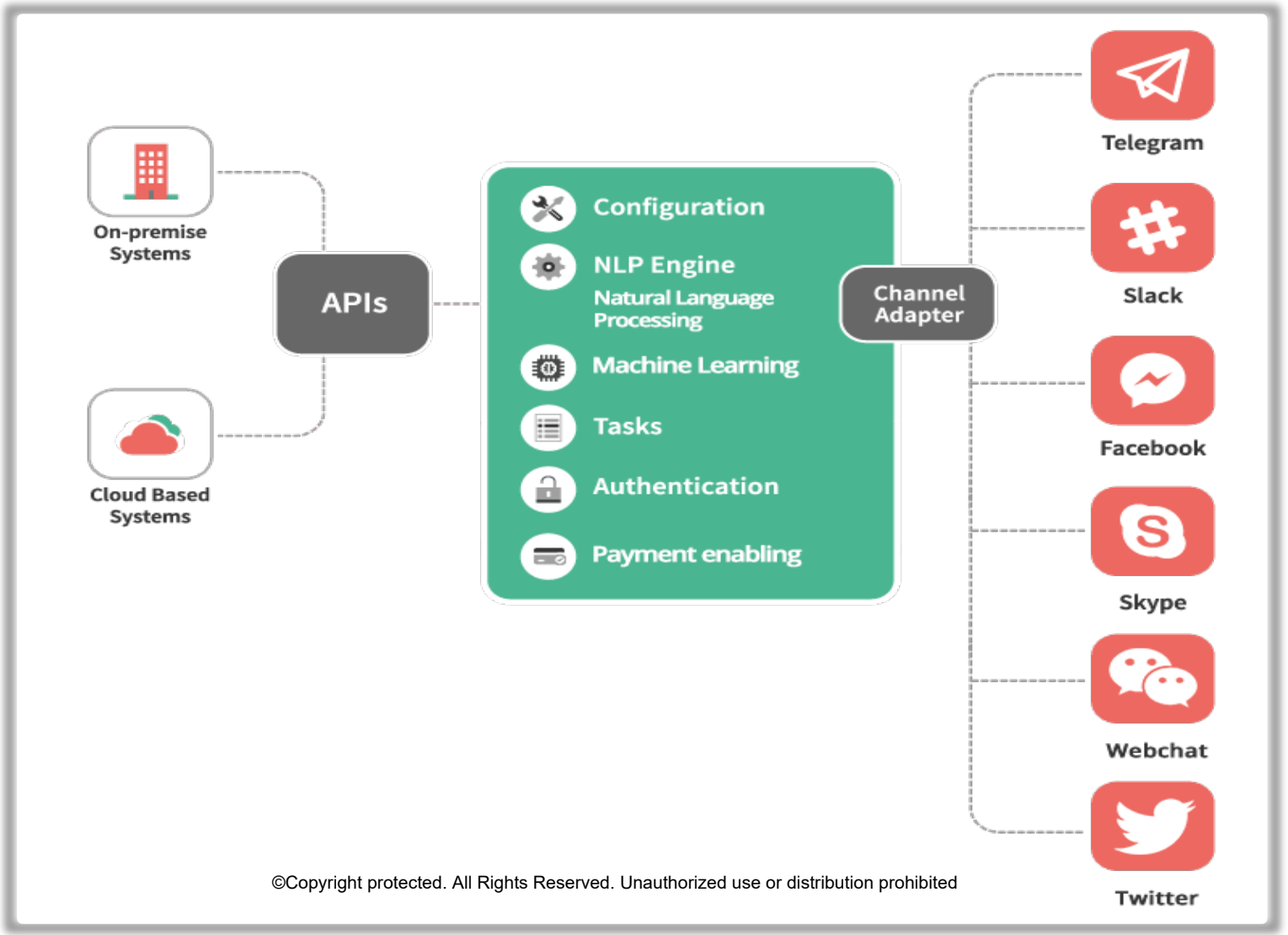
Deployment 5

Deployment



Prediction and model evaluation

Real Time application of NLP



Mapping of NLP methods vs applications

Domain specific NLP use cases – Marketing & Retail

Domain specific NLP use cases – Banking & FInance

Domain specific NLP use cases - Healthcare

Domain specific NLP use cases - Sports

COMMON NLP TERMINOLOGY

Corpus

Vocabulary

My friend is afraid of spiders.
But he is not afraid of lizards.

Tokenisation

My friend is afraid of spiders.

Normalisation

Stemming

Lemmatisation

Stop words

My friend is afraid of spiders.

Bag of words

My friend is afraid of spiders.
But he is not afraid of lizards.

Part-of-speech tagging

My friend is afraid of spiders.

Named Entity Recognition

New Delhi: A new Rs 75,000-crore fund has been announced by Google to help accelerate India's digital economy, its Chief Executive Officer Sundar Pichai said today

Word Sense Disambiguation

NLP & Deep Learning

Thank You