Data Mining and Text Analytics

Postgraduate Programme in "AI for Business and Society "

A.Y. 2025-2026

Lecture_04

Prof. Alessandro Bruno

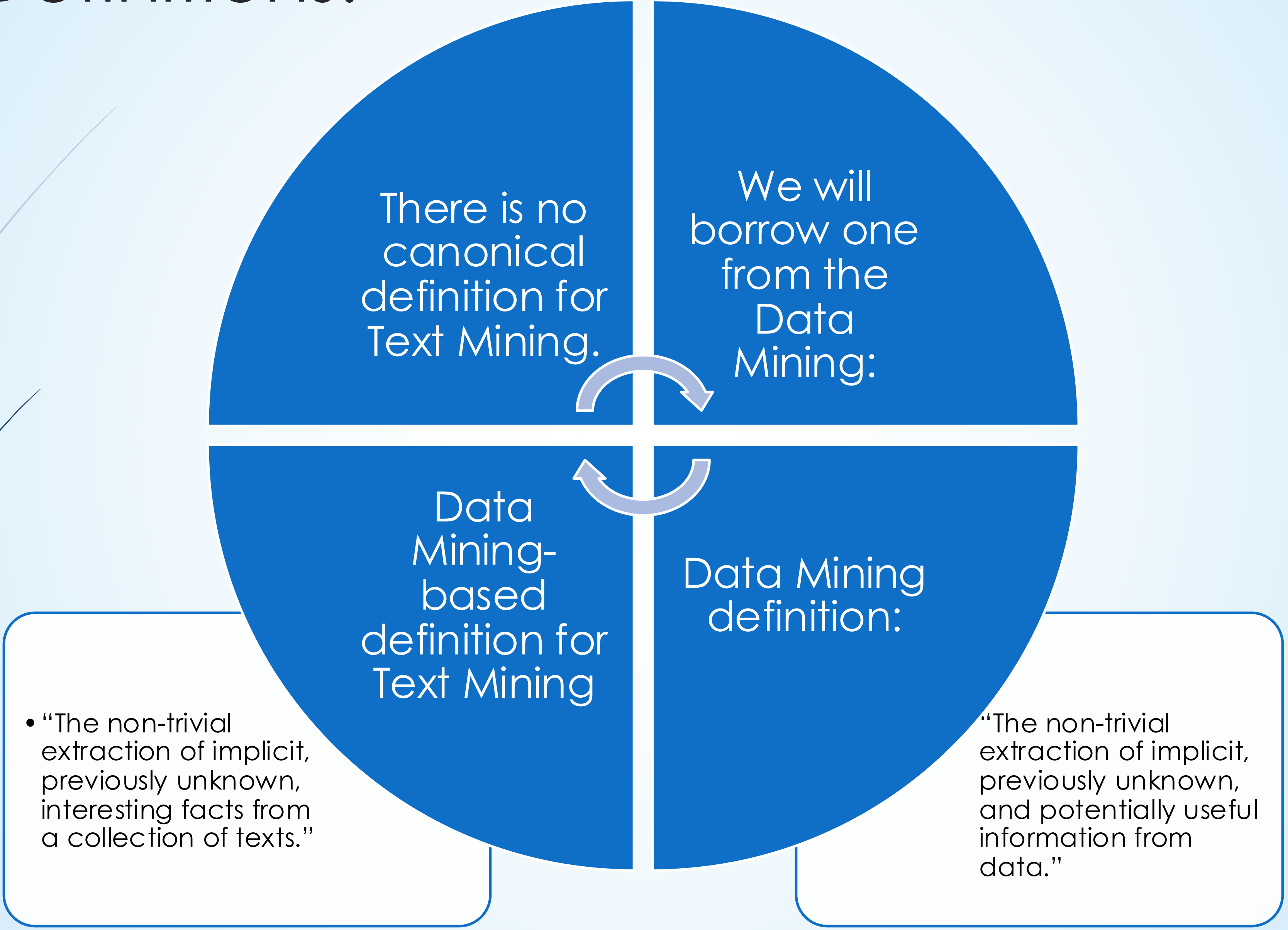
Dr Alessandro Bruno

Outline

- Definitions
- Basic Concepts
- Textual Content
- Text Mining and NLP
- Application domains
- Main Tasks
- LLMs (foundation models and GPT-based)
- Text Classification
- Text Clustering
- Topic Model
- Topic Detection and Tracking
- Information Extraction
- Text Summarisation
- Challenges
- Word-Vector Representation (Embeddings)

```
irror_mod = modifier_ob.
 mirror object to mirror
mirror_mod.mirror_object
operation == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
irror_mod.use_z = False
 _operation == "MIRROR_Y":
irror_mod.use_x = False
irror_mod.use_y = True
irror_mod.use_z = False
 _operation == "MIRROR_Z":
irror_mod.use_x = False
irror_mod.use_y = False
irror_mod.use_z = True

selection at the end -add
_ob.select= 1
ier_ob.select=1
ntext.scene.objects.active
("Selected" + str(modifier
irror_ob.select = 0
= bpy.context.selected_ob
ata.objects[one.name].se

int("please select exactl

- OPERATOR CLASSES ----

types.Operator):
 X mirror to the selected
ject.mirror_mirror_x"
rror X"

ontext):
ext.active_object is not
```

Definitions:



There is no canonical definition for Text Mining.

We will borrow one from the Data Mining:

Data Mining definition:

"The non-trivial extraction of implicit, previously unknown, and potentially useful information from data."

Data Mining-based definition for Text Mining

- "The non-trivial extraction of implicit, previously unknown, interesting facts from a collection of texts."

Credits go Rob Sanderson from Liverpool University to https://ucrel.lancs.ac.uk/events/htm06/RobSandersonHTM06.pdf

Definitions:

Gartner glossary provides a definition for Text Mining:

“The process of extracting information from collections of textual data and utilising it for business objectives.”

Source: https://www.gartner.com/en/information-technology/glossary/text-mining

Gartner, Inc. is an American research and advisory firm focusing on business and technology topics

Some Basic Concepts

Main text mining goals:

- Analysis & Modeling of unstructured natural language context

Textual content

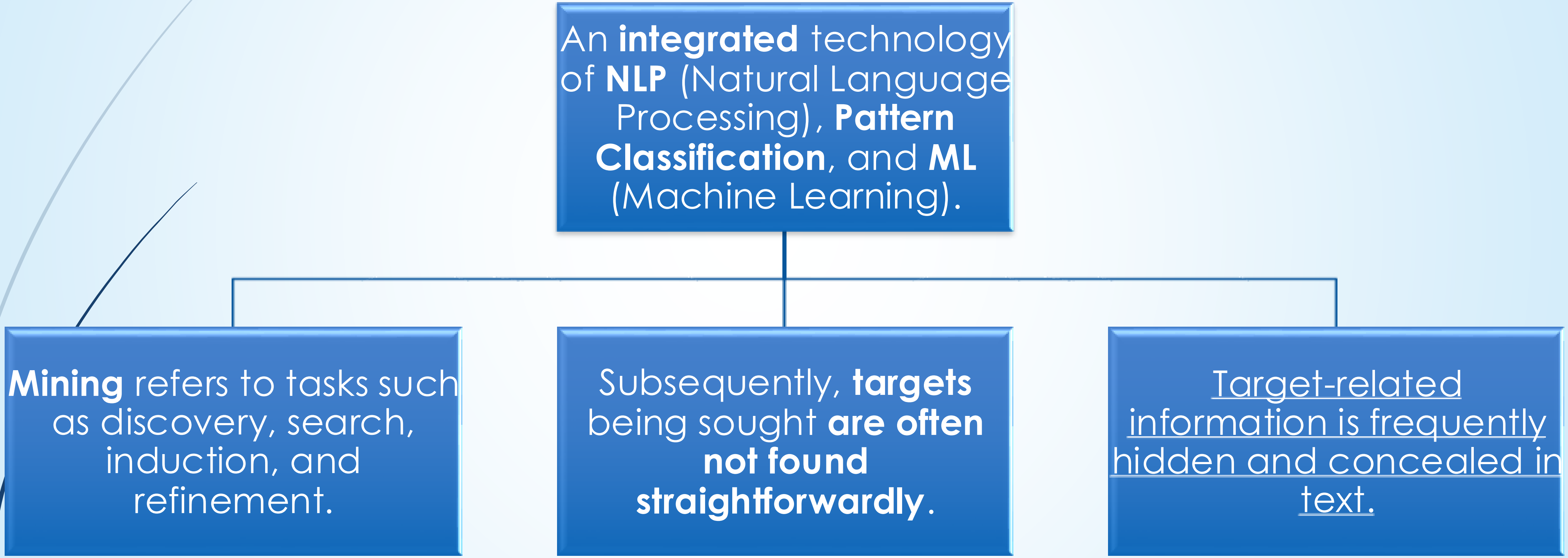
- almost always unstructured (unlike databases and data warehouse)
- described by natural language (ruling out graphics and images)

Structured vs Semi-Structure vs Unstructured Textual Content

- Structured Data: Data with high degree of organisation (in a spreadsheet-like manner).
- Semi-structured Data: Data with some degree of organisation.
- Unstructured Data: Data with no predefined organisation.
- Some examples:

Structured Data	Semi-structured Data	Unstructured Data
Excel spreadsheets; Comma-separated value file; Relational Database Tables.	HTML files; Json files; XML files.	PDF files; Word Files, Plain text files.
Around 20% of worldwide data is structured.	Characterised by hierarchical structure.	Most of data that is created today is unstructured. (Tweets, Facebook posts, social media comments).

Text Mining



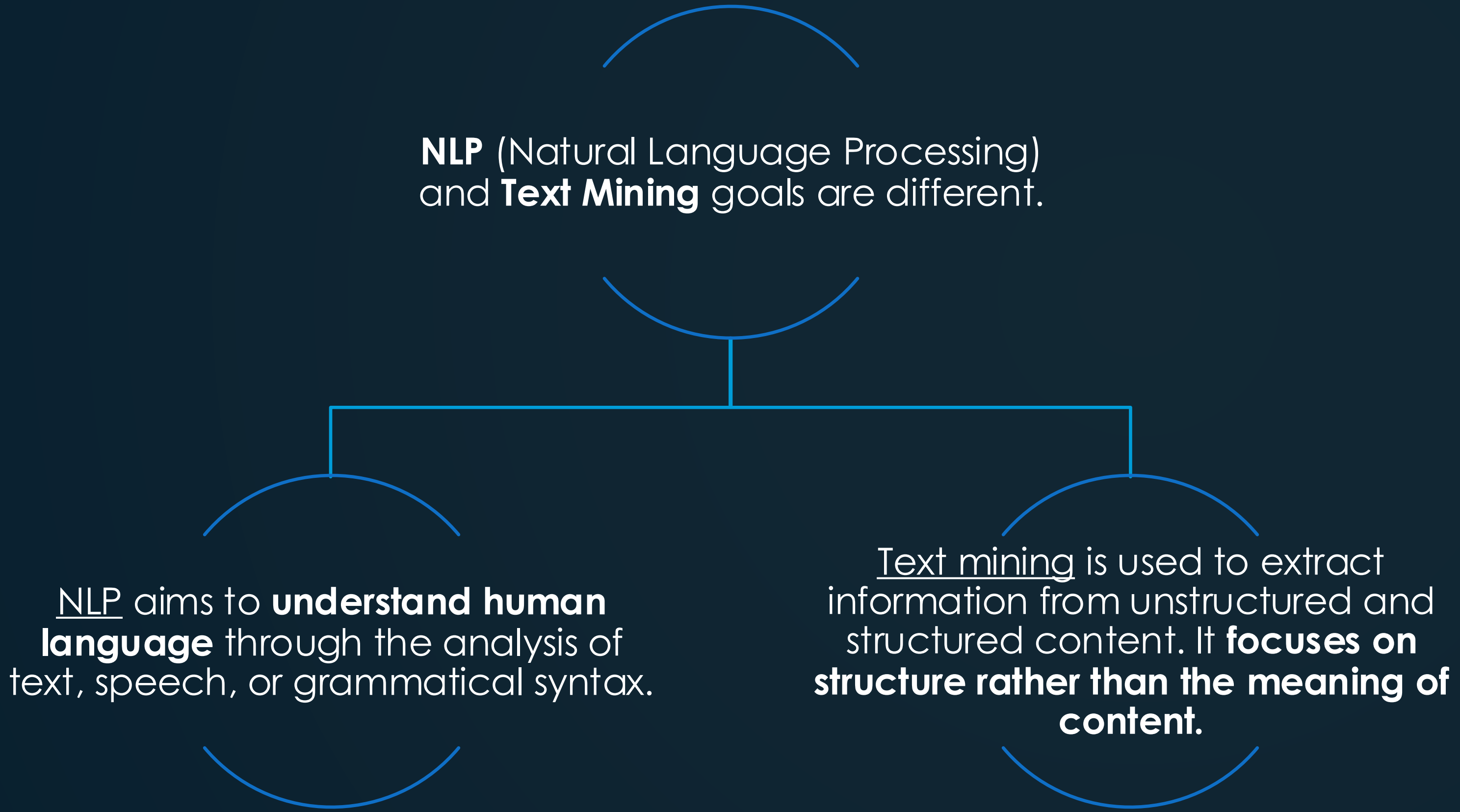
An **integrated** technology of **NLP** (Natural Language Processing), **Pattern Classification**, and **ML** (Machine Learning).

Mining refers to tasks such as discovery, search, induction, and refinement.

Subsequently, **targets** being sought **are often not found straightforwardly**.

Target-related information is frequently hidden and concealed in text.

Text Mining and NLP: Objectives

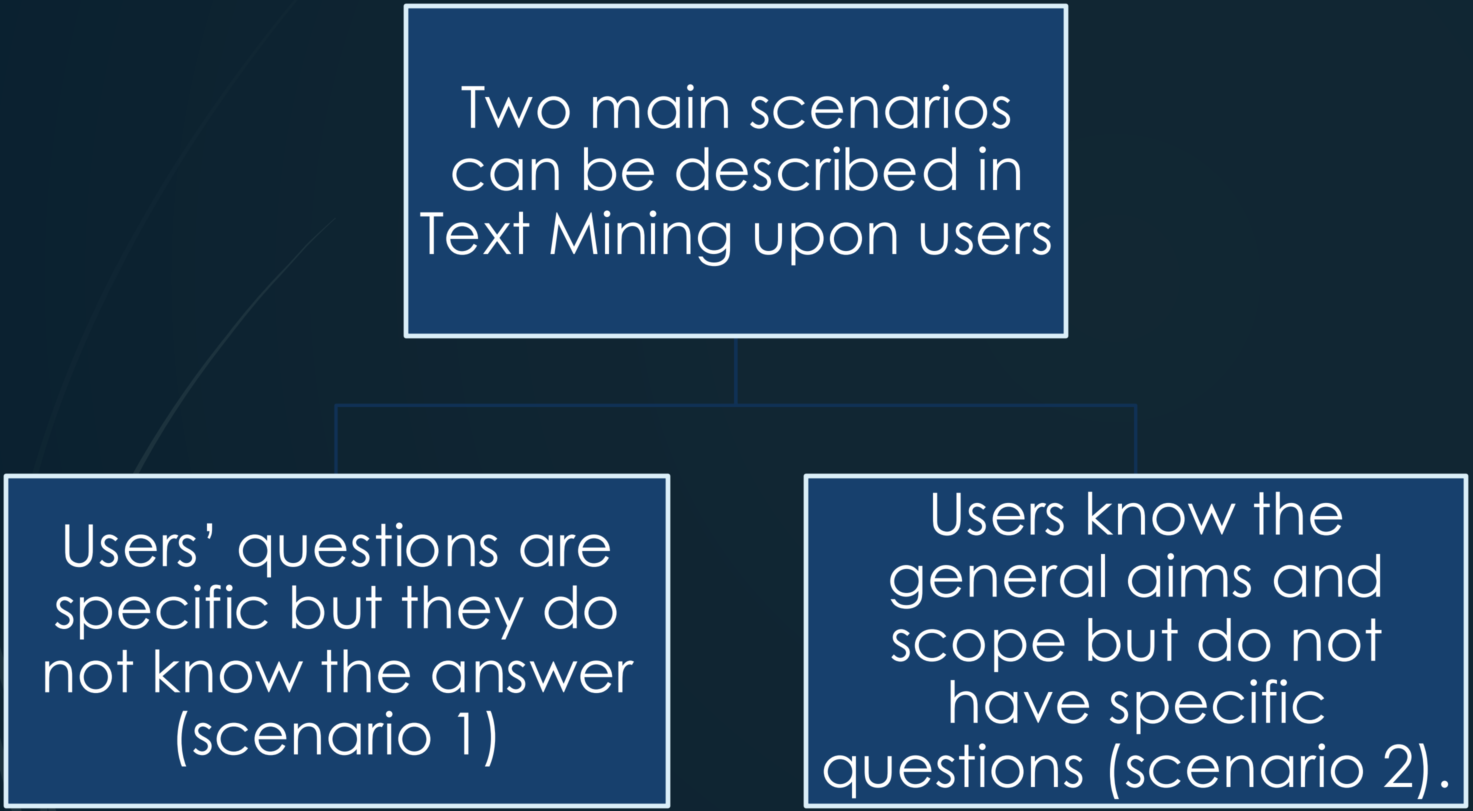


NLP (Natural Language Processing) and **Text Mining** goals are different.

NLP aims to **understand human language** through the analysis of text, speech, or grammatical syntax.

Text mining is used to extract information from unstructured and structured content. It **focuses on structure rather than the meaning of content.**

Text Mining



Two main scenarios can be described in Text Mining upon users

Users' questions are specific but they do not know the answer (scenario 1)

Users know the general aims and scope but do not have specific questions (scenario 2).

Text Mining Application Domains

Economy

Social Management

Information Services

Security

Risk Management

Customer Care Service

Fraud Detection

Business Intelligence

Social Media Analysis

Customer Churn Prediction

Q&A Systems

Marketing

Some historical hints

Traditional data mining relies on structured data such as database tables (relational model).

In the late 90s, researchers started using text as data, giving rise to text mining.

Early text mining applied data mining and machine learning algorithms on text data without using NLP techniques.

NLP (Natural Language Processing) has a longer history. It all started back in 1950s.

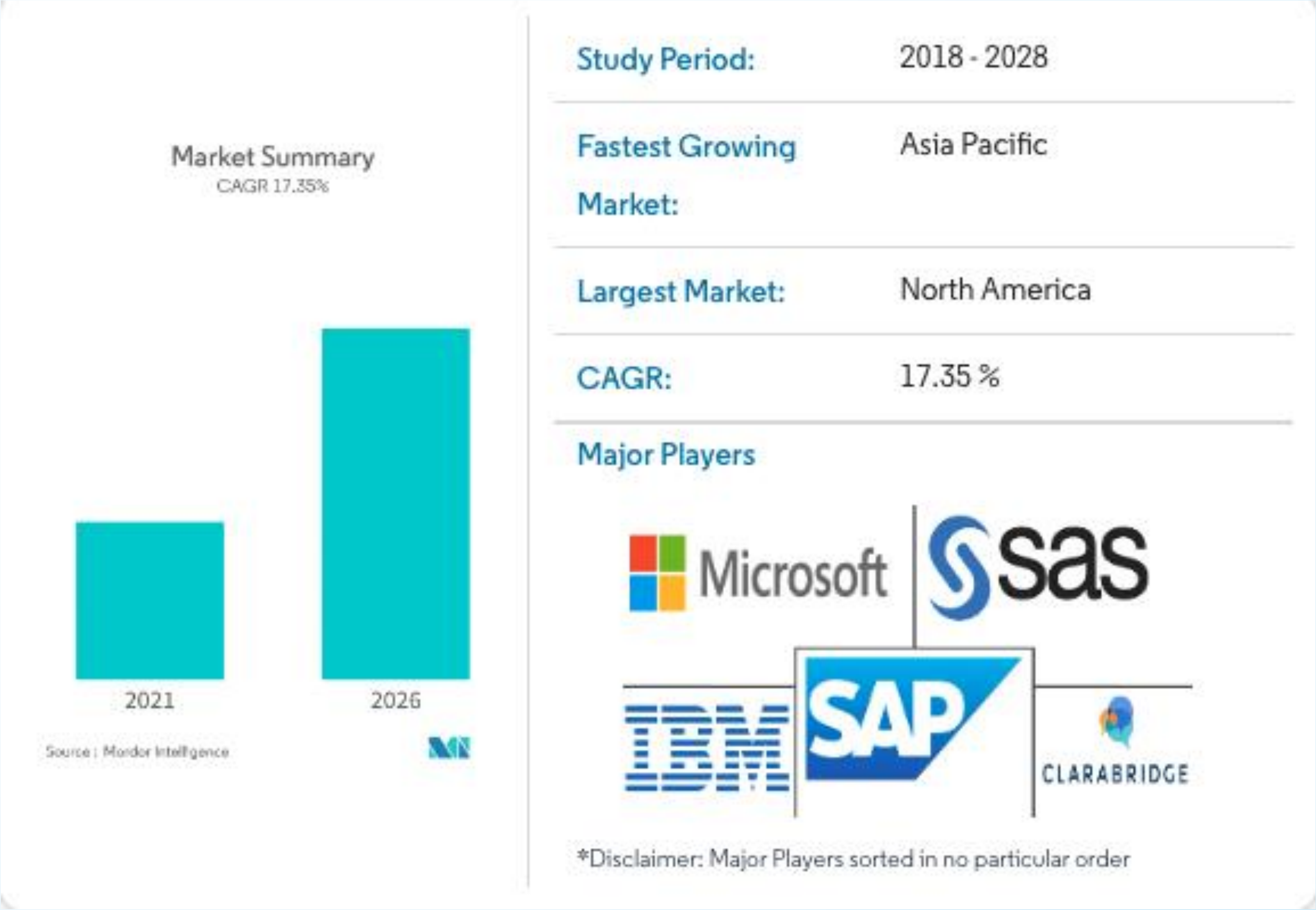
- GOAL: Make computers understand Human Language.

Later on, Text Mining started using NLP techniques. Subsequently, in the past 10 year, NLP and Text Mining have tackled some common points, albeit with different goals.

Professor Bing Liu's Interview from Sharon Dexter (Project, It Manager, Green Book)
Link to interview

Text Data Mining Figures

- CAGR in business and finance stands for Compound Annual Growth Rate
- Here is a diagram with figures on Text Analytics over the next few years.



Study Period:	2018 - 2028
Fastest Growing Market:	Asia Pacific
Largest Market:	North America
CAGR:	17.35 %

Major Players

Microsoft, sas, SAP, IBM, CLARABRIDGE

*Disclaimer: Major Players sorted in no particular order

URL: https://www.mordorintelligence.com/industry-reports/text-analytics-market

Main Tasks of Text Mining

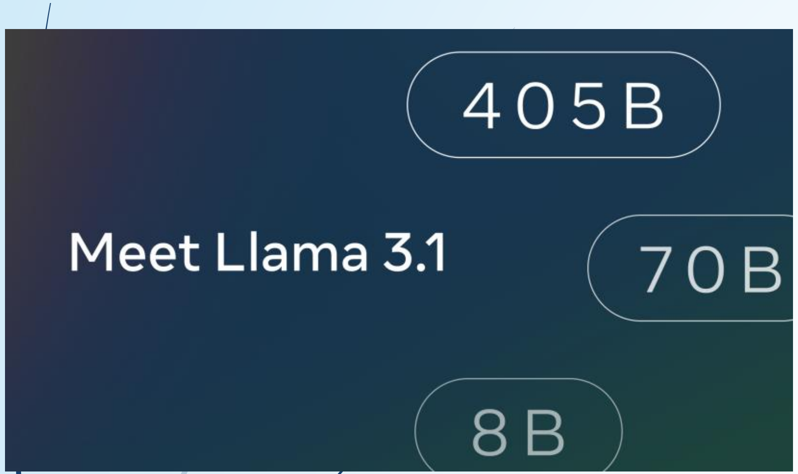
Text Mining is often hidden within several applications:

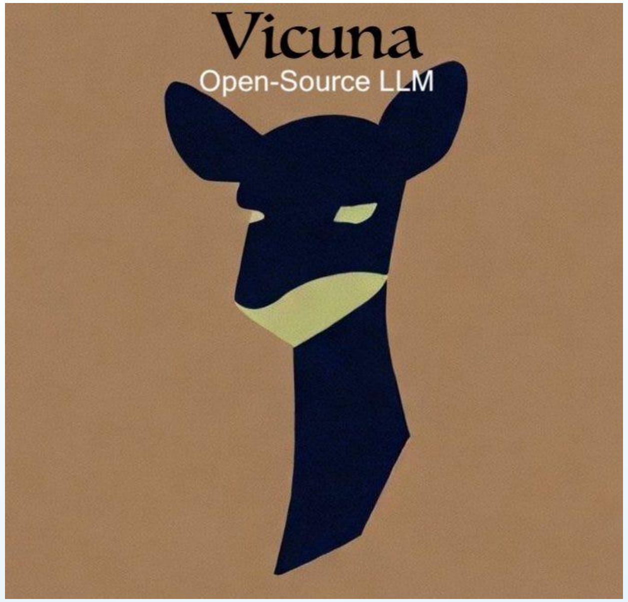
- For instance, Q&A (Question & Answering) systems have text data mining deal with tasks such as knowledge base search, inference and filtering of candidate answers, question parsing

In this unit, 7 main tasks are considered:

- Text Classification
- Text Clustering
- Topic Model
- Text Sentiment Analysis and Opinion Mining
- Topic Detection and Tracking
- Information extraction
- Automatic Text Summarisation

It’s raining LLMs





LLMs for text mining and NLP

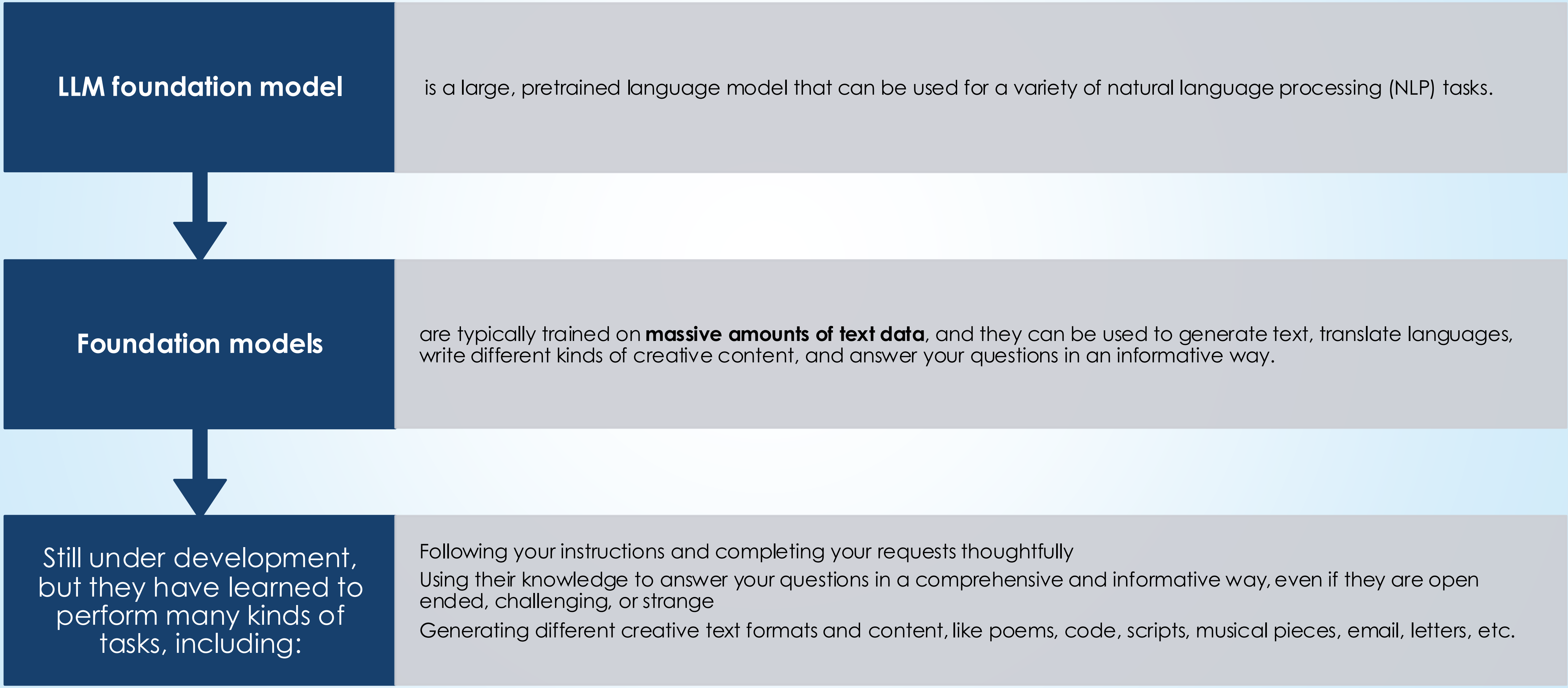
The seven steps listed in the previous slide refer to standard Text Mining and NLP methods that are still in use.

LLMs (Large Language Models) have made their way through Text Mining and NLP field.

In this unit, standard Text Mining and NLP as well as LLMs will be covered.

In this slide-deck some insights into LLMs will be given.

LLMs (Large Language Models)



LLM foundation model

is a large, pretrained language model that can be used for a variety of natural language processing (NLP) tasks.

Foundation models

are typically trained on **massive amounts of text data**, and they can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.

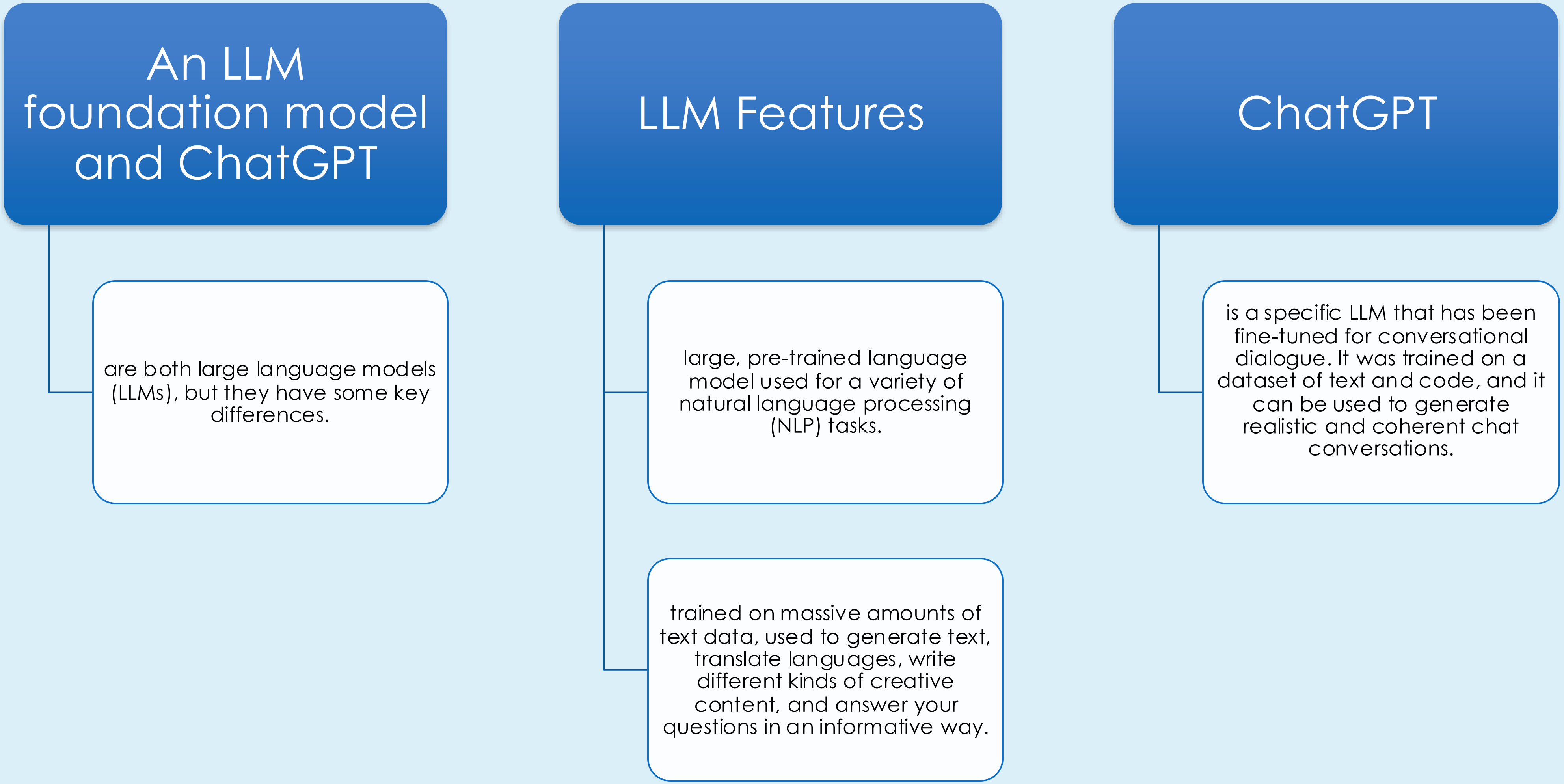
Still under development, but they have learned to perform many kinds of tasks, including:

Following your instructions and completing your requests thoughtfully

Using their knowledge to answer your questions in a comprehensive and informative way, even if they are open ended, challenging, or strange

Generating different creative text formats and content, like poems, code, scripts, musical pieces, email, letters, etc.

Foundation Models and ChatGPT



An LLM foundation model and ChatGPT

are both large language models (LLMs), but they have some key differences.

LLM Features

large, pre-trained language model used for a variety of natural language processing (NLP) tasks.

trained on massive amounts of text data, used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.

ChatGPT

is a specific LLM that has been fine-tuned for conversational dialogue. It was trained on a dataset of text and code, and it can be used to generate realistic and coherent chat conversations.

Feature	LLM foundation model	ChatGPT
Purpose	General-purpose NLP	Conversational dialogue
Training data	Massive amounts of text data	Dataset of text and code
Tasks	Generate text, translate languages, write different kinds of creative content, answer questions	Generate realistic and coherent chat conversations
Strengths	Versatility, accuracy	Fluency, realism
Weaknesses	Bias, computational cost	Limited scope

Foundation Models and ChatGPT

Quiz time!

- Scan the QR code or visit the link below to run a quiz
- https://forms.gle/z7mKKrptZNdLwXHP9

Text Classification

Goal: Divide test into **predefined text types**

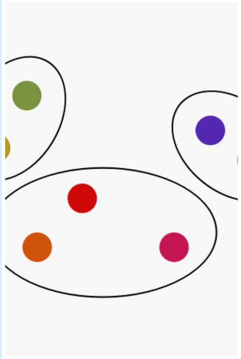
Type: It is a pattern classification technology

Example: Chinese Library Classification www.sina.com

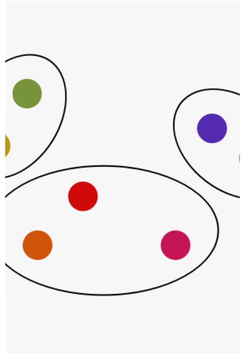
All books are grouped into 5 categories and 22 subcategories. The classification task is run on the content.

Link to Google Colab Project on Text Classification

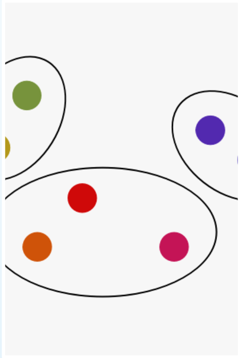
Text Clustering



Text Clustering divides a given text into different categories.



Clustering **does not include predefined categories**. The number and type of categories depend on some criteria, evaluation and indices.



Example: Some text can be clustered into news, entertainment, sports, finance. Based on users' viewpoint, a piece of text can be clustered into positive categories (positive and supportive attitudes) and negative categories (negative and passive attitudes).

Topic Model

- A topic can be generally expressed by words having strong conceptual and semantic relationships.
- Topic Model is a statistical approach assigning a topic probability value to each word.
- Each topic carries to a specific dictionary

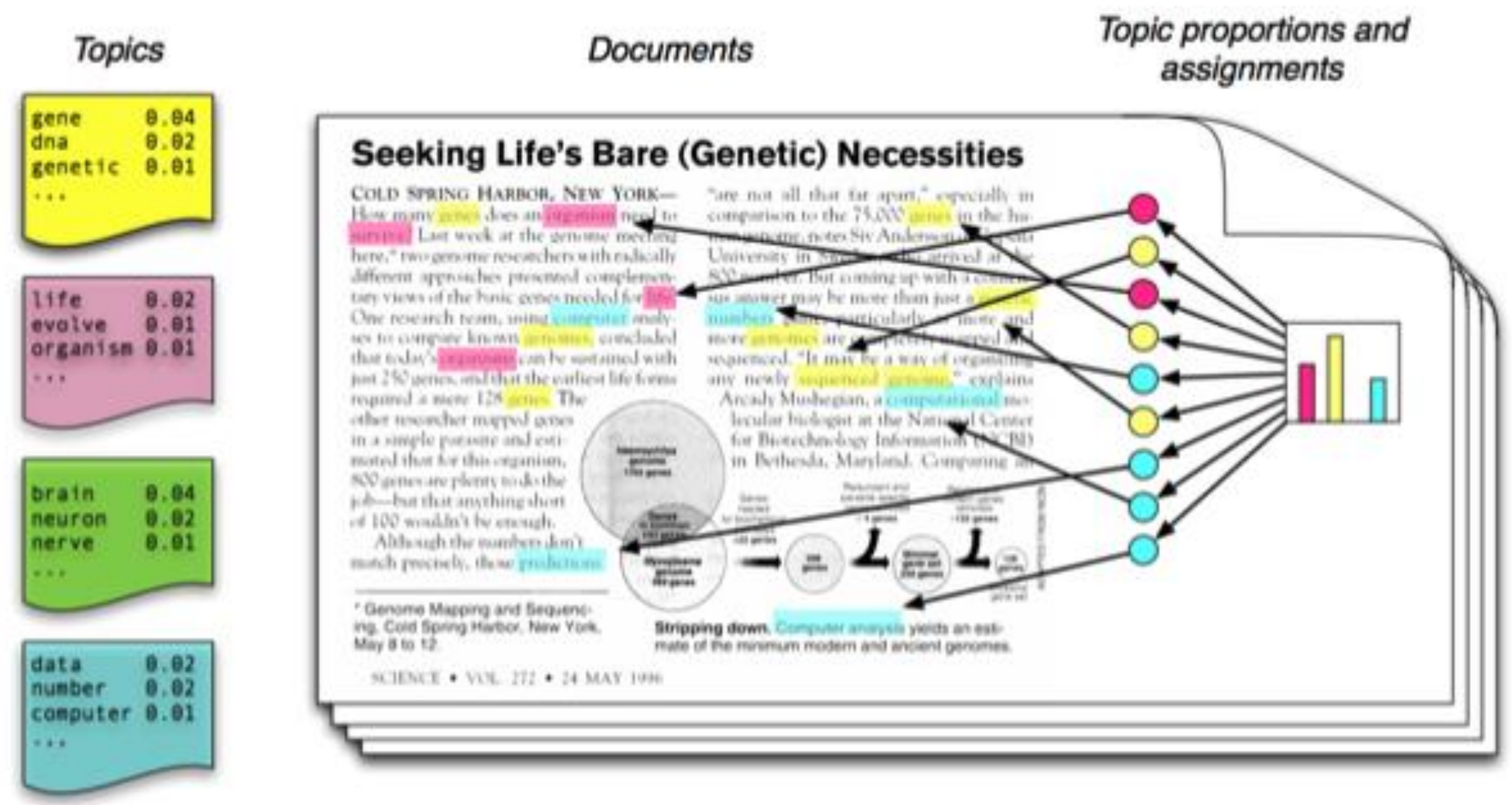
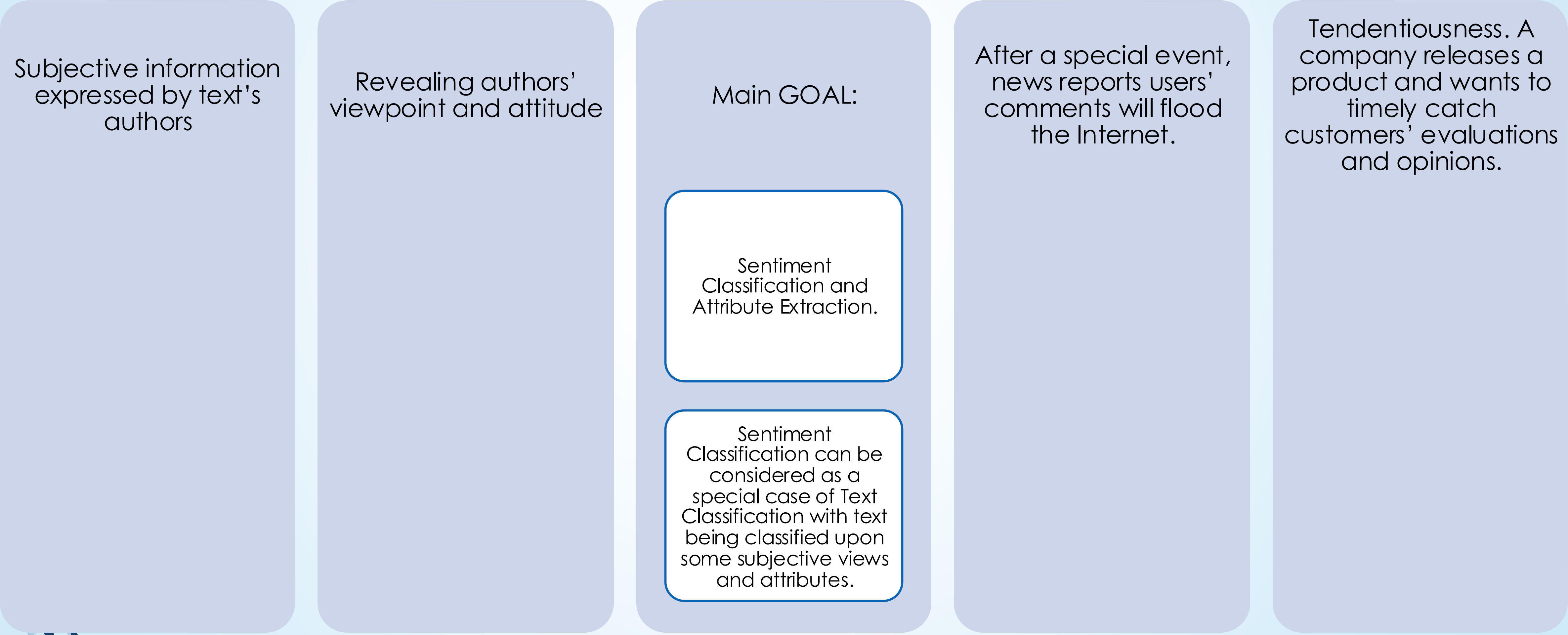


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Text Sentiment Analysis and Opinion Mining



Subjective information expressed by text's authors

Revealing authors' viewpoint and attitude

Main GOAL:

Sentiment Classification and Attribute Extraction.

Sentiment Classification can be considered as a special case of Text Classification with text being classified upon some subjective views and attributes.

After a special event, news reports users' comments will flood the Internet.

Tendentiousness. A company releases a product and wants to timely catch customers' evaluations and opinions.

Topic Detection and Tracking

Mining and Screening of text topics from news reports and comments

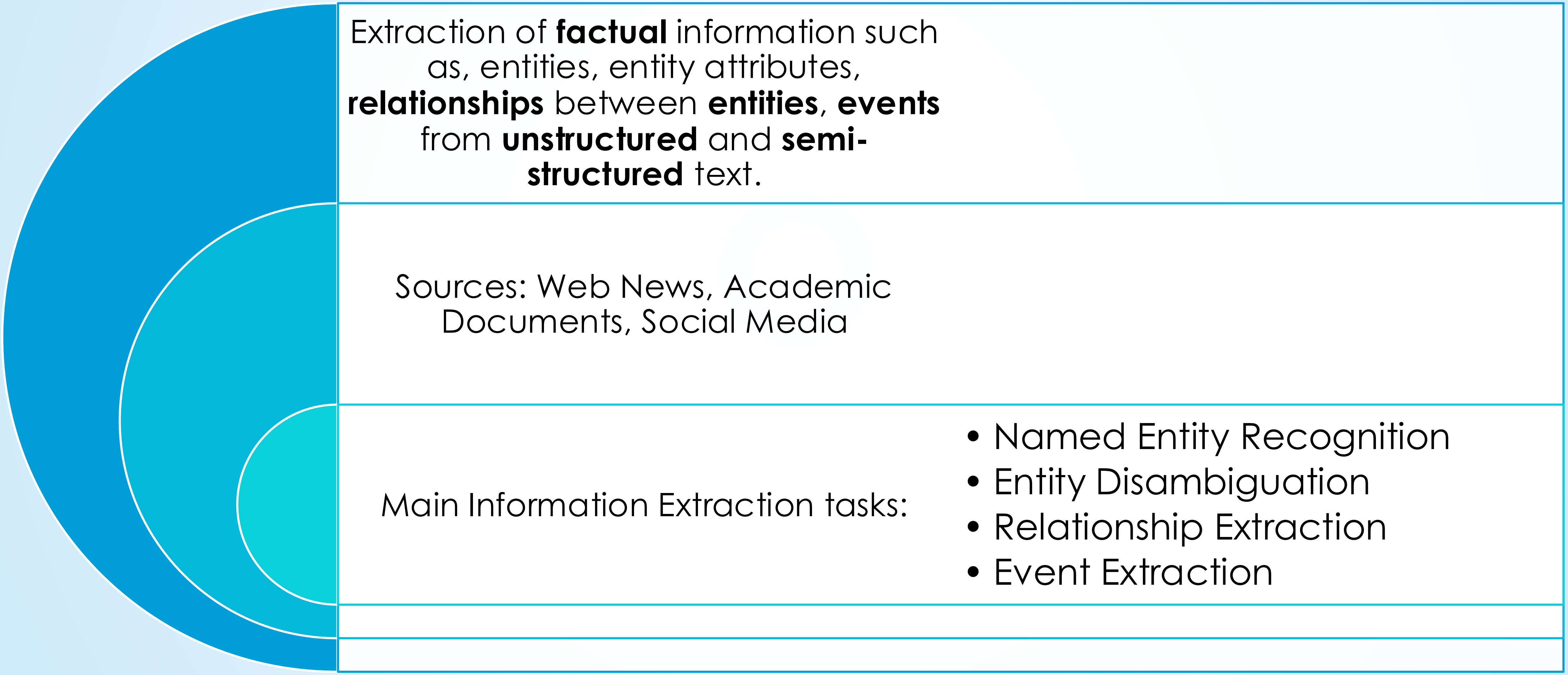
Topics most people care about (hot topics).

Hot topics detection, discovery and tracking are used in opinion analysis.

Example:

Hot Topics Today is a report on what is most attracting readers' attention from the news.

Information Extraction



Extraction of **factual** information such as, entities, entity attributes, **relationships** between **entities**, **events** from **unstructured** and **semi-structured** text.

Sources: Web News, Academic Documents, Social Media

Main Information Extraction tasks:

- Named Entity Recognition
- Entity Disambiguation
- Relationship Extraction
- Event Extraction

Information Extraction

Information Extraction has recently become popular across several application domains, such as, biomedical/medical text mining, financial field, social reputation.

It should be noted that the **relation in information extraction** usually refers to some semantic relation between two or more concepts.

In daily life, the way people describe events refers to aspects such as, when, where, and what happened.

In **event extraction**, the event usually refers to a specific state expressed by a certain predicate framework.

"*John meets Mary*"

In daily life, people think of a **story**

In **event extraction**, the "event" is "**triggered**" by a verb and is nothing more than a **state** or an **action**.

Automatic Text Summarisation

Technology focusing on automatically generating summaries using natural language processing (NLP) methods.

Visit the link https://www.textcompactor.com/

That is an online tool for text summarisation.

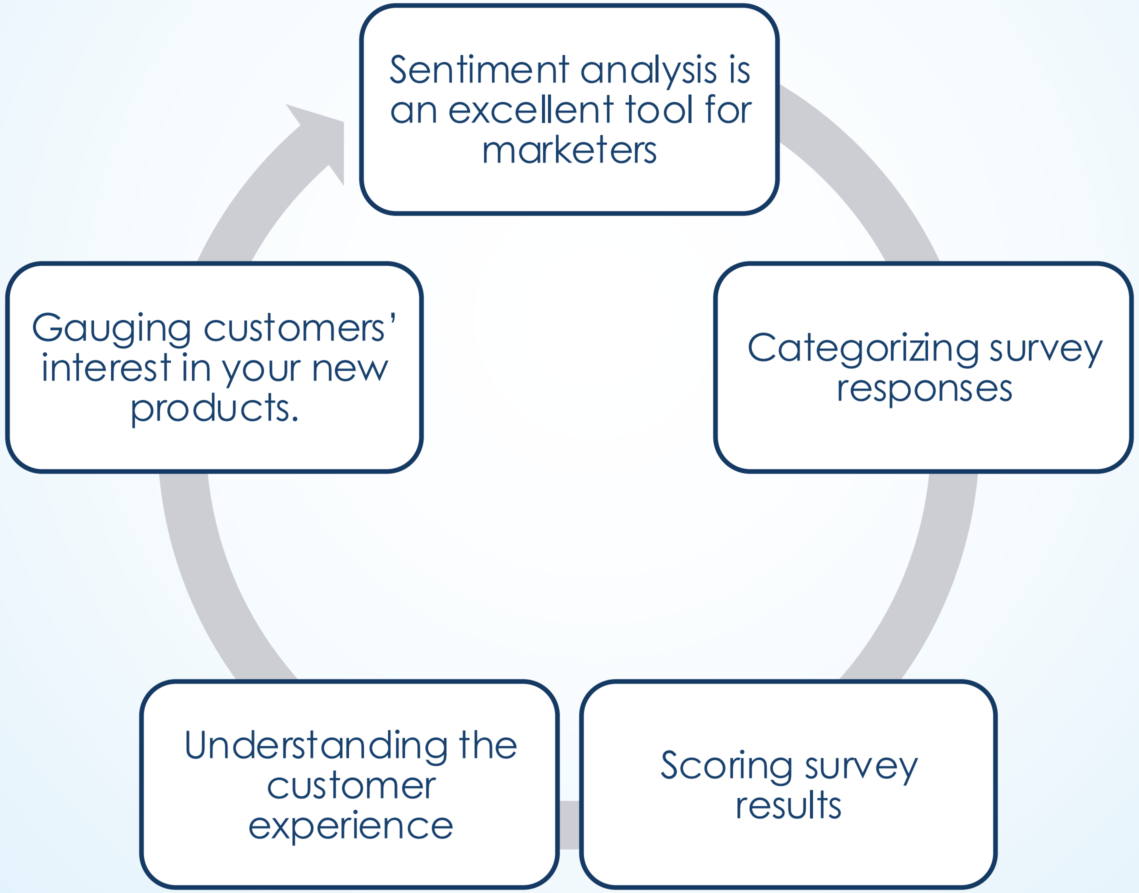
When information is saturated, several companies employs text summarisation to cut the chase and extract the most meaningful excerpts from text.

Homework assignment No. 1

- “Obviously, language is not only about computation. The most difficult thing to do is to create a tool that is “smart”, i.e. it can understand not only grammar but also the semantics of a text.”

- Above is a quote from the article by Marco Belmondo (Chief Marketing Officer at Datrix group) titled “**Why use an automatic text summarization tool for digital content?**”
- Read the whole article and answer the question below:
- What are the benefits of Text Summarisation for Businesses? (Max 100 words).

Text Mining applications in Marketing



Sentiment analysis is an excellent tool for marketers

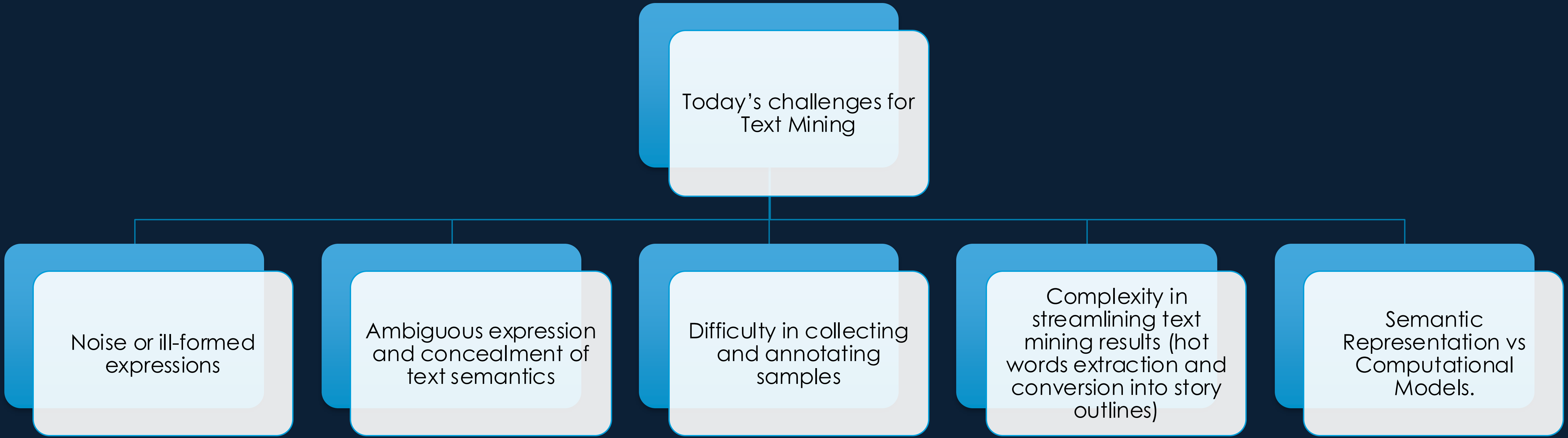
Categorizing survey responses

Scoring survey results

Understanding the customer experience

Gauging customers' interest in your new products.

Main challenges in Text Mining



Today’s challenges for Text Mining

Noise or ill-formed expressions

Ambiguous expression and concealment of text semantics

Difficulty in collecting and annotating samples

Complexity in streamlining text mining results (hot words extraction and conversion into story outlines)

Semantic Representation vs Computational Models.

Main Challenge in Text Mining

- **Noise or ill-formed expression:**
  - Literary works, academics publications, political article, news article from TV channels and other media channels stick to some standard semantic rules.
  - If we compare the above-mentioned scenarios with online text, things change dramatically
  - Several ill-formed expressions are in online websites
  - Ill-formed expressions make it harder to NLP and Text Mining tasks to be accurate.
  - CWS (Chinese Word Segmentation) achieves accuracy rates over than 95% when fed by People's Daily
  - CWS (Chinese Word Segmentation) accuracy rates drop on online text (below 90%).

Main Challenge in Text Mining

Ambiguous expression and concealment of text semantics

- "Bank" may refer to a financial bank or a river bank
- "Apple" may refer to the fruit or to a product such as Apple Iphone, Apple Macbook, etc.
- "I saw a boy with a telescope"
- The sentence above is ambiguous: it may refer to a boy holding a telescope being seen by me; it may refer to a boy being seen by me through a telescope.
- The correct parsing of statements like this is challenging for NLP systems.
- There are no effective methods to sort out this challenge.

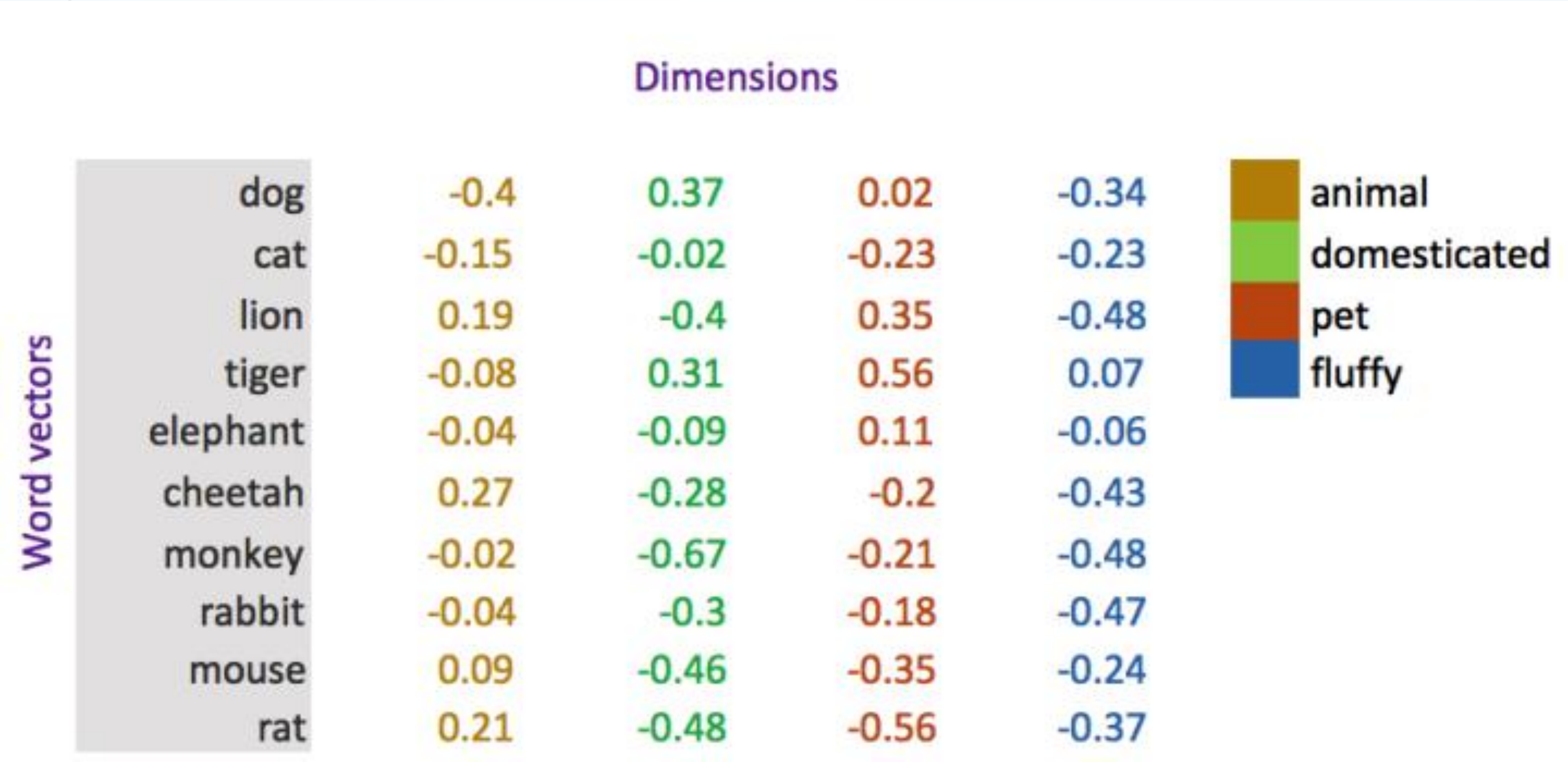
Main Challenges in Text Mining

Data Collection and Annotation

- The mainstream text mining methods are machine learning-based
  - Traditional statistical-based machine learning
  - Deep Learning
- Large collections of annotated data are necessary
- Obtaining large amount of data from online websites might be difficult due to copyright protected contents.
- Even when there is no copyright related issues, online content is generally unwell formatted and needs several processing steps
- On top of that, if contents pertains to a specific area, experts are needed to manually annotate the content itself.

Main Challenges in Text Mining

- One of the most well-know challenges regards the word-vector representation



Dimensions

Word vectors

dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

🟫	animal
🟩	domesticated
🟥	pet
🟦	fluffy

Word Vector Representation

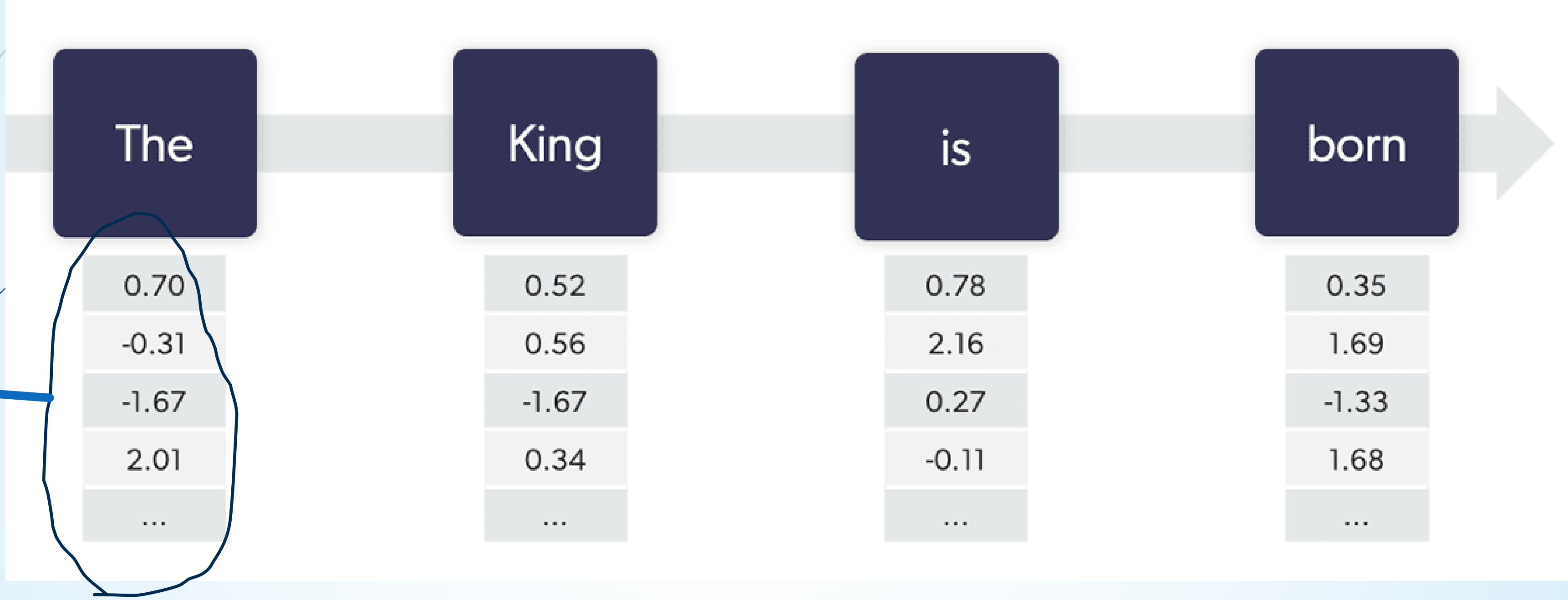
Word to vector representation, also known as **word embedding**, is a technique used in natural language processing (NLP) to represent **words** and **phrases** as **numerical vectors**.

There are many different methods for word embedding, but they all share the same basic goal of **capturing the semantic and syntactic relationships between words**.

One popular method is called **word2vec**, which was developed by **Google**.

Word2vec uses two architectures: continuous bag-of-words (**CBOW**) and continuous **skip-gram**. Both architectures work by predicting the surrounding words given a target word.

Word2Vec

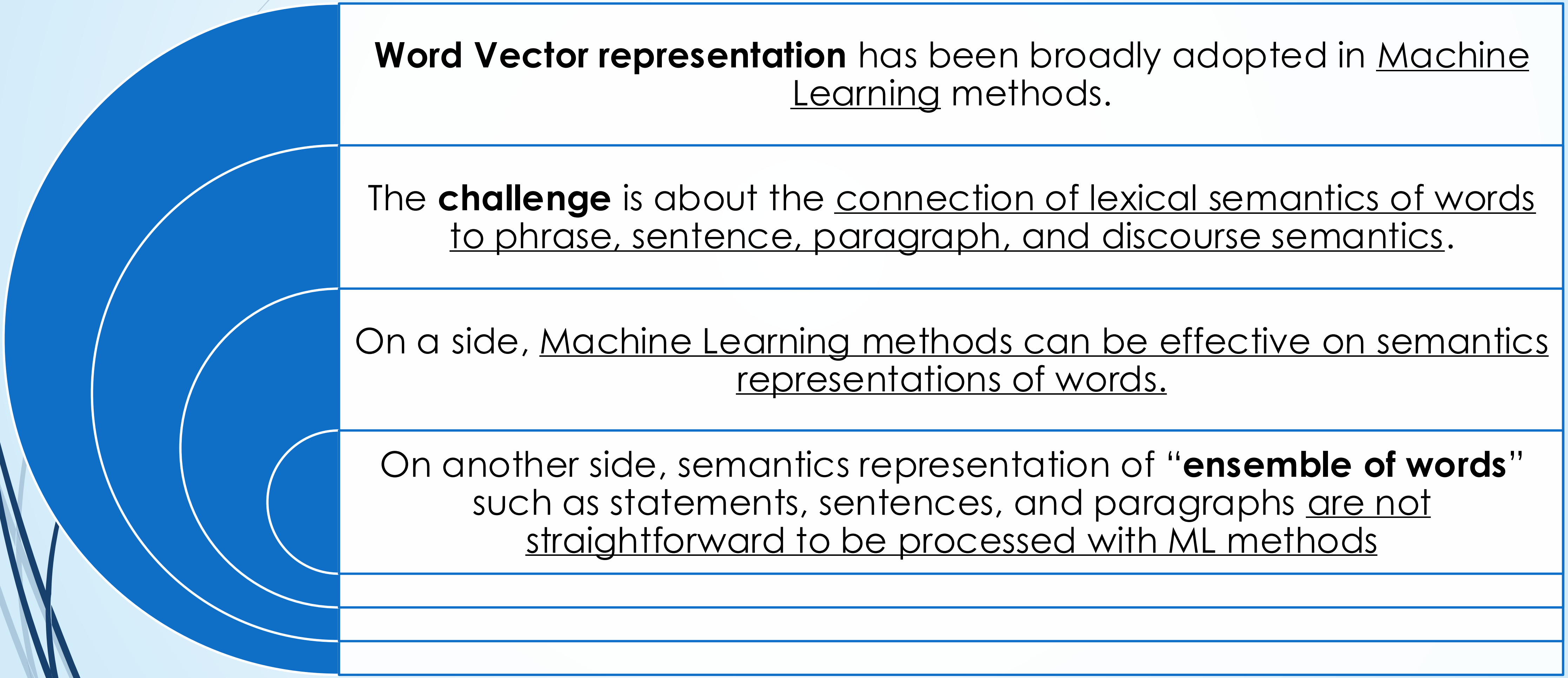


The	King	is	born
0.70	0.52	0.78	0.35
-0.31	0.56	2.16	1.69
-1.67	-1.67	0.27	-1.33
2.01	0.34	-0.11	1.68
...	...	...	...

Embedding

A phrase like "The King is born" is divided into words.
Each word is converted into numerical vectors (Embeddings).

Main Challenges in Text Mining



Word Vector representation has been broadly adopted in Machine Learning methods.

The **challenge** is about the connection of lexical semantics of words to phrase, sentence, paragraph, and discourse semantics.

On a side, Machine Learning methods can be effective on semantics representations of words.

On another side, semantics representation of “**ensemble of words**” such as statements, sentences, and paragraphs are not straightforward to be processed with ML methods

Takeaway points

- Text mining can be defined as the non-trivial extraction of implicit, previously unknown, interesting facts from a collection of texts
- Text content is almost always unstructured
- The following seven topics will be tackled in this unit: Text Classification; Text Clustering; Topic Model; Text Sentiment Analysis and Opinion Mining; Topic Detection and Tracking; Information extraction; Automatic Text Summarisation
- LLMs Foundation models and fine-tuned versions (ChatGPT) are also used to address NLP tasks