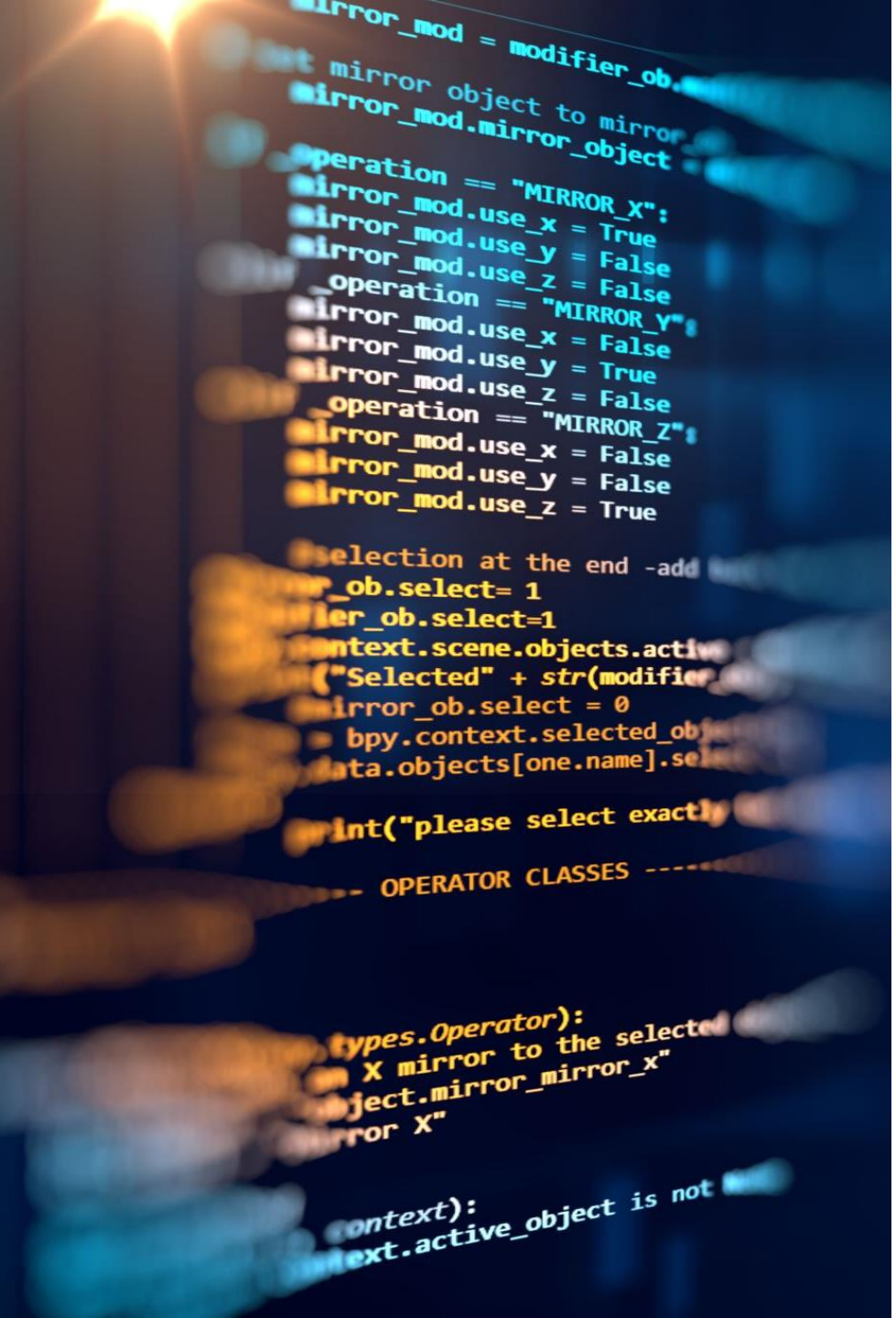
Salimah, M. (2022). Text Analytics [Powerpoint slides]

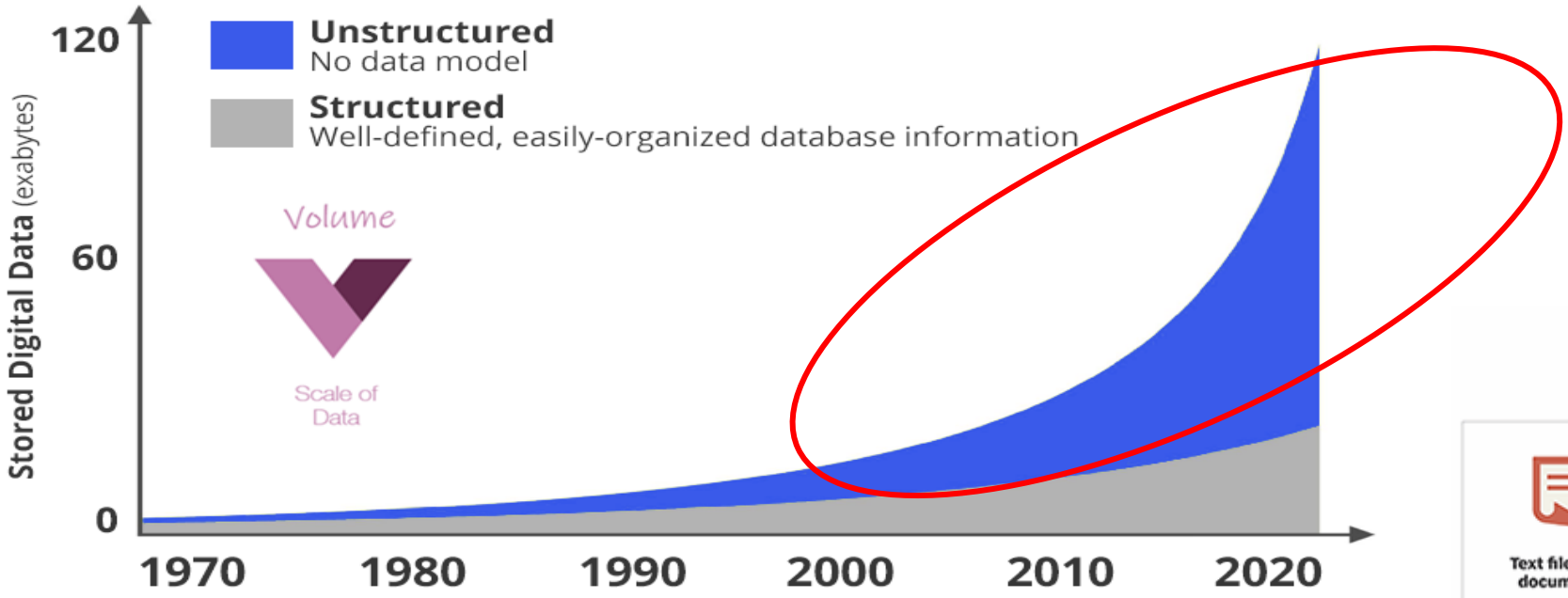


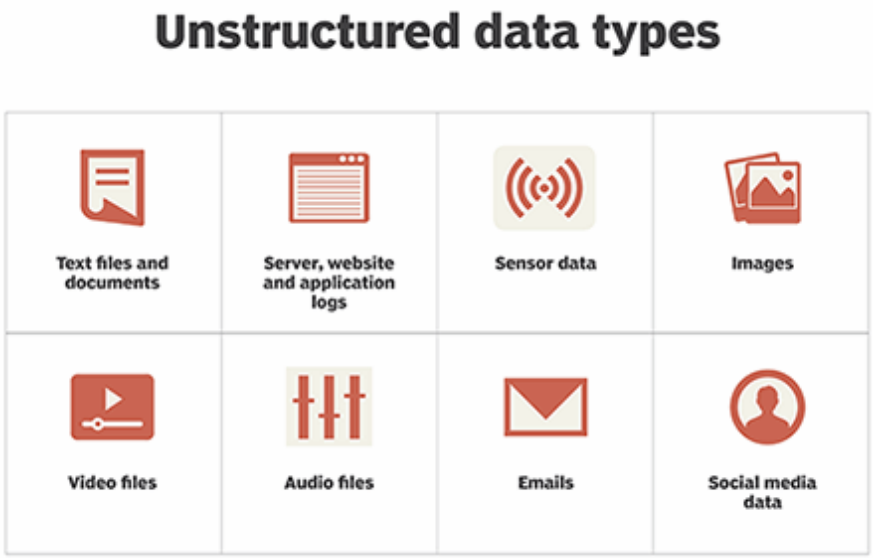
Learning Objectives

1. To describe unstructured data and its importance.
2. To revise NLP as a branch of AI.
3. To explain text analytics and related components IR, IE.
4. To differentiate between text mining and text analytics.
5. To review the process in doing text analytics.
6. To discuss about sentiment analysis.

Why Is Unstructured Data So Important?

Situation Today





People and **machines** are producing data at a very high rate than ever before.

The **volume** and **variety** of data being produced bring more challenges in **identifying useful information** from them.

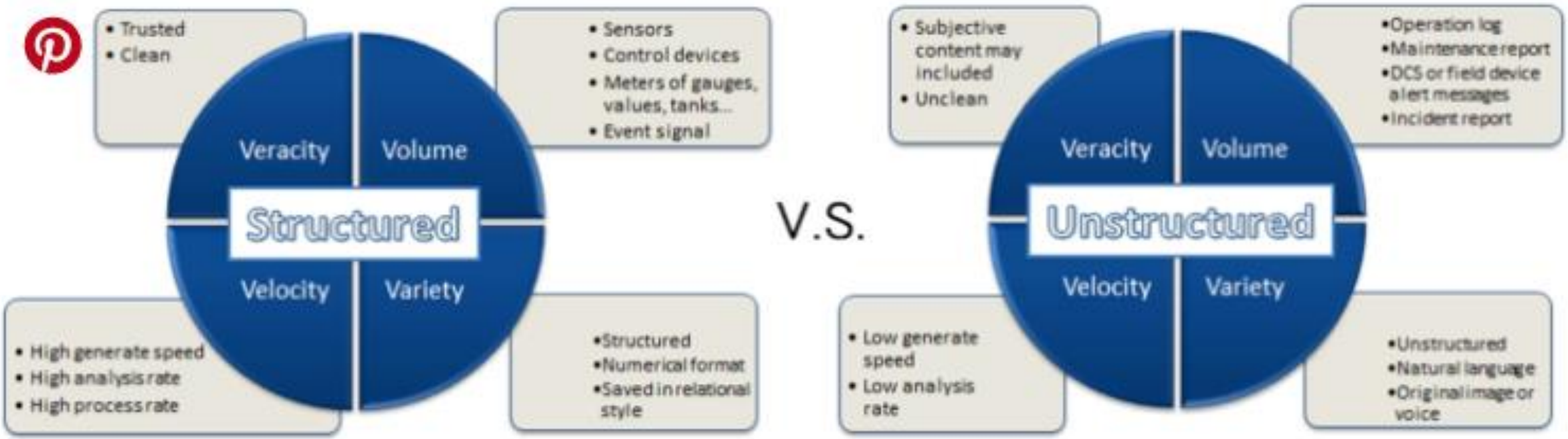
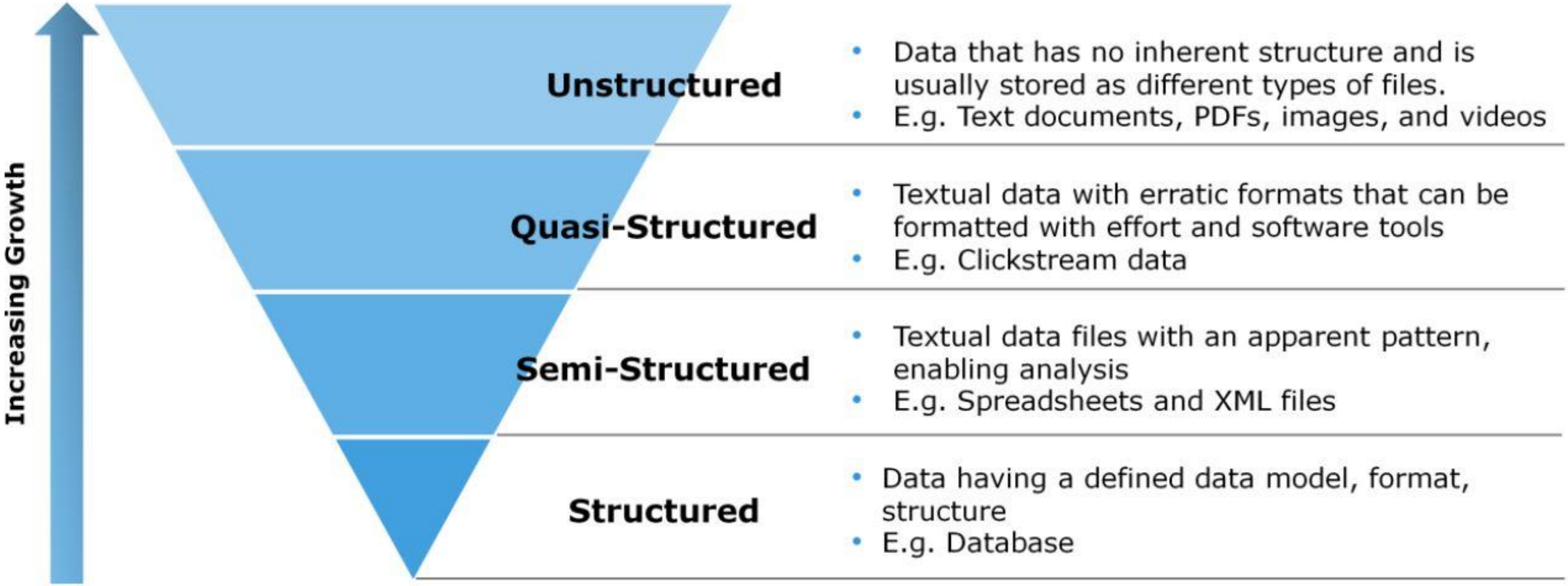


FIGURE 1. THE 4 CHARACTERISTICS OF BIG DATA IN A PLANT: STRUCTURED DATA VS UNSTRUCTURED DATA (SEMI-STRUCTURED DATA OMITTED)

https://blog.yokogawa.com/blog/find-value-in-unstructured-plant-data

Unstructured means that the datasets (typical large collections of files) that aren't stored in a structured database format.



Unstructured Data

Unstructured data has internal structure but is not structured via pre-defined data models or schema. It may be **textual** or **non-textual**, and **human-** or **machine-** generated. It may also be stored within a **non-relational database** like **NoSQL**.

Human-generated unstructured data :

- **Text files**: Word processing, spreadsheets, presentations, email, logs.
- **Email**: Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media**: Data from Facebook, Twitter, LinkedIn.
- **Website**: YouTube, Instagram, photo sharing sites.
- **Mobile data**: Text messages, locations.
- **Communication**s: Chat, IM, phone recordings, collaboration software.
- **Media**: MP3, digital photos, audio and video files.
- **Business applications**: MS Office documents, productivity applications.

Machine-generated unstructured data:

- **Satellite imagery**: Weather data, landforms, military movements.
- **Scientific data**: Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance**: Surveillance photos and video.
- **Sensor data**: Traffic, weather, oceanographic sensors.

https://www.datamation.com/big-data/structured-vs-unstructured-data.html

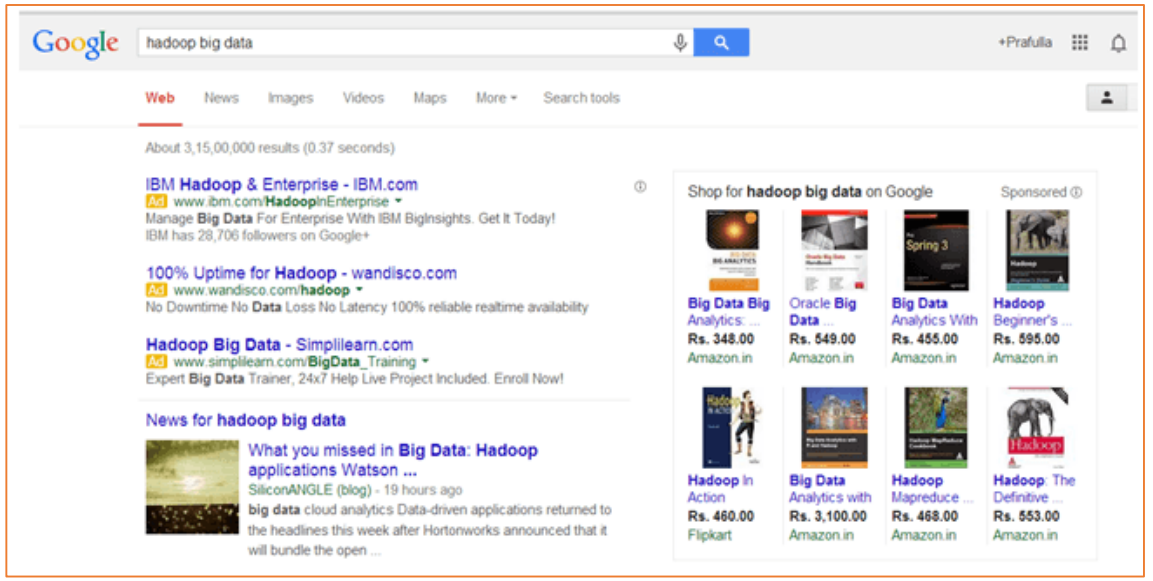
https://www.youtube.com/watch?v=T5ibveutnnU&feature=emb_logo

Structured Data

Employee_ID	Employee_Name	Gender	Department
2365	Rajesh Kulkarni	Male	Finance
3398	Pratibha Joshi	Female	Admin
7465	Shushil Roy	Male	Admin
7500	Shubhojit Das	Male	Finance
7699	Priya Sane	Female	Finance

An Employee table in a database

Unstructured Data

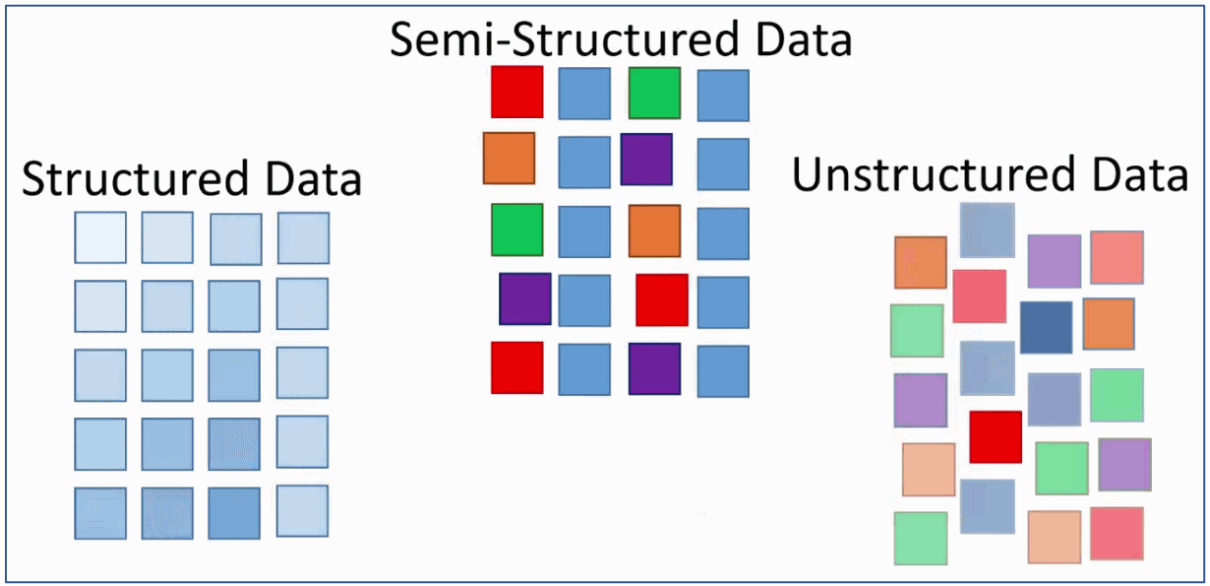


The output returned by Google Search

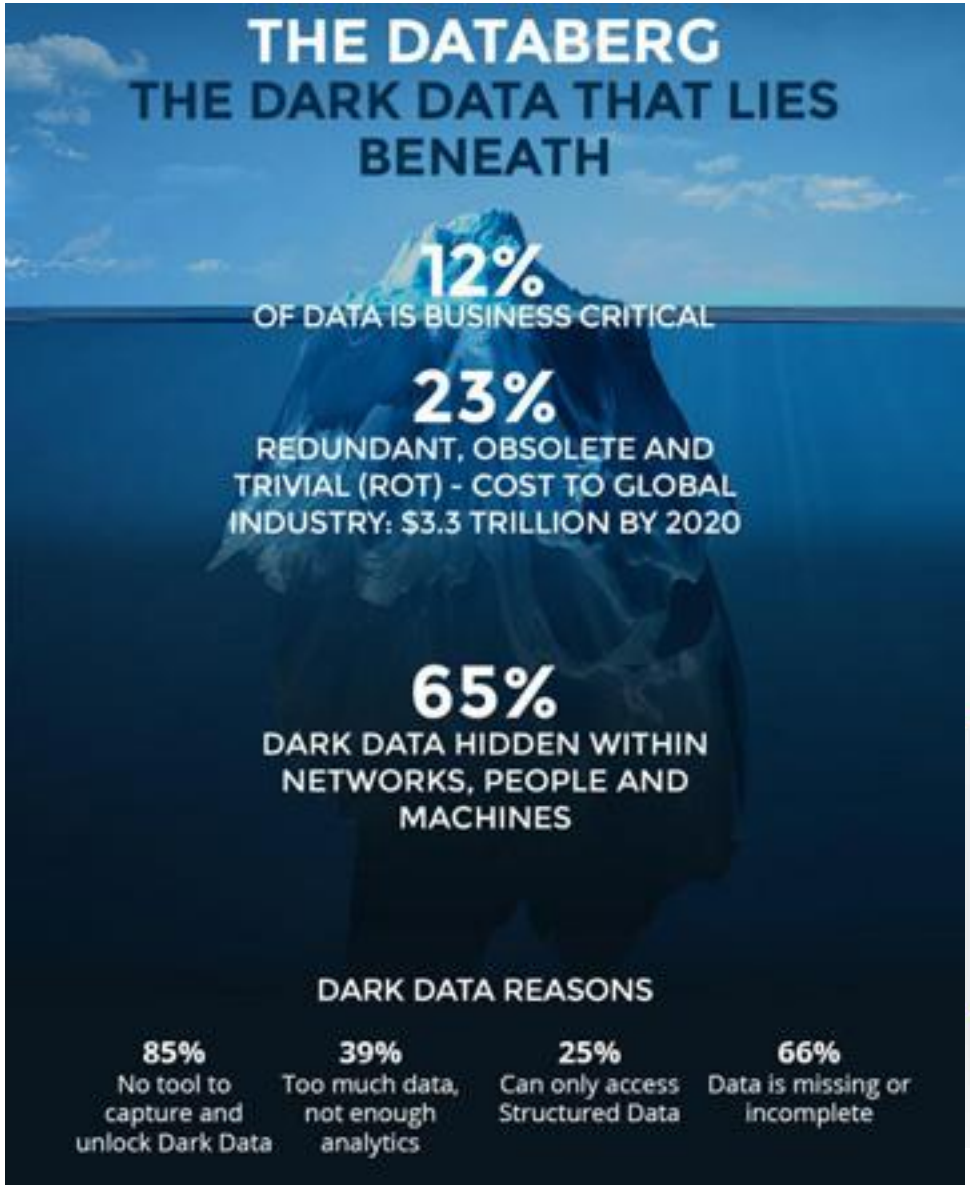
Semi-Structured Data

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Personal data stored in an XML file



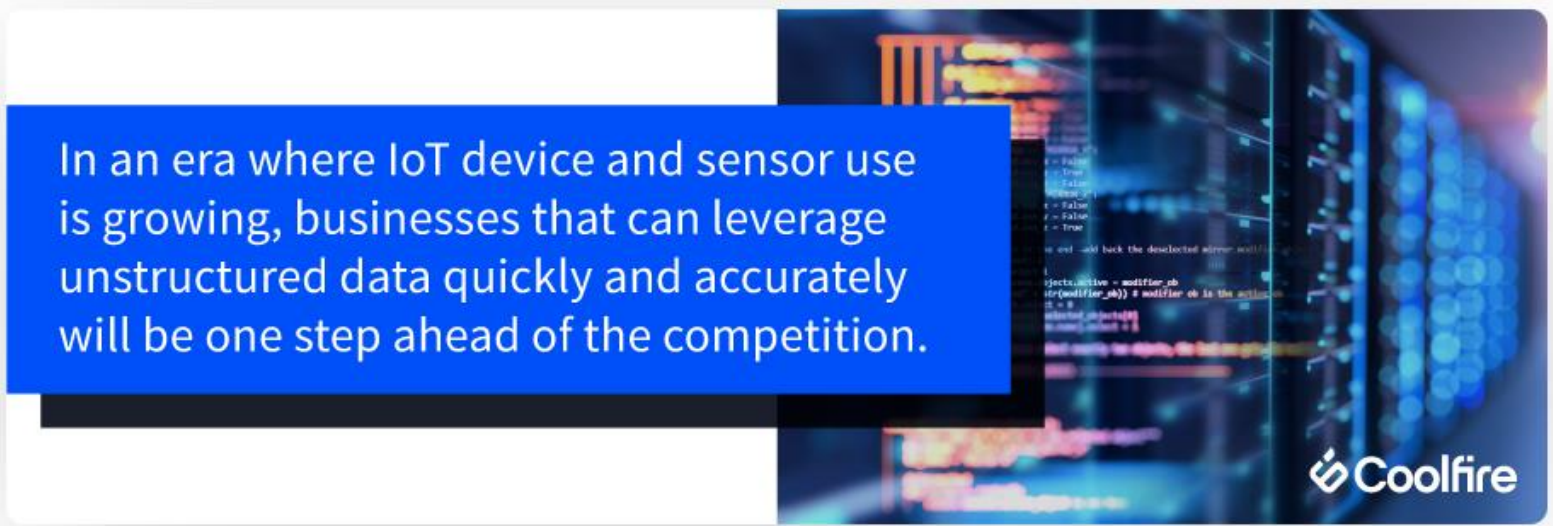
Importance!



expert.ai

Unstructured Data: The Data Too Important to Ignore

Jay Selig - 12 October 2020



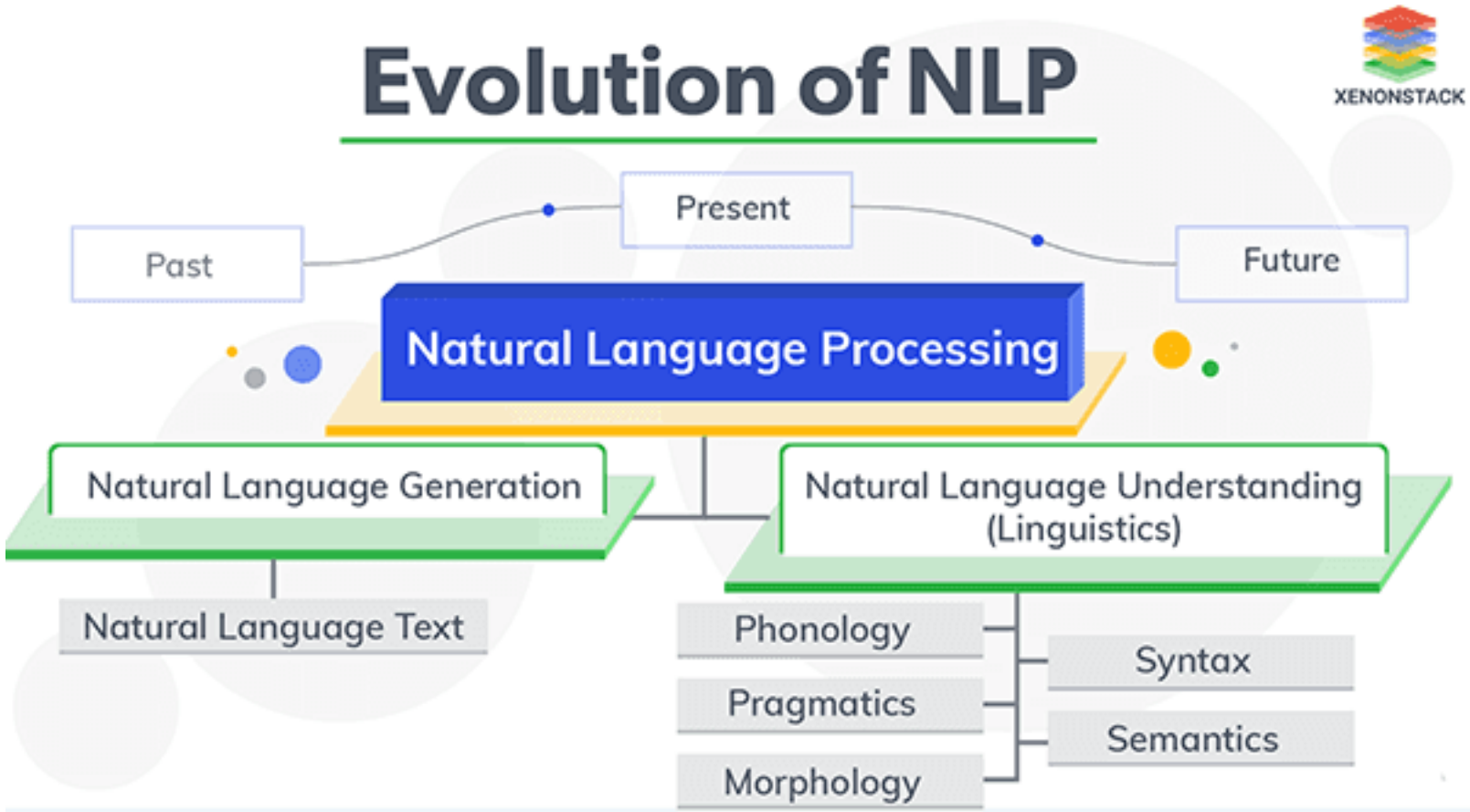
Natural Language Processing (NLP)

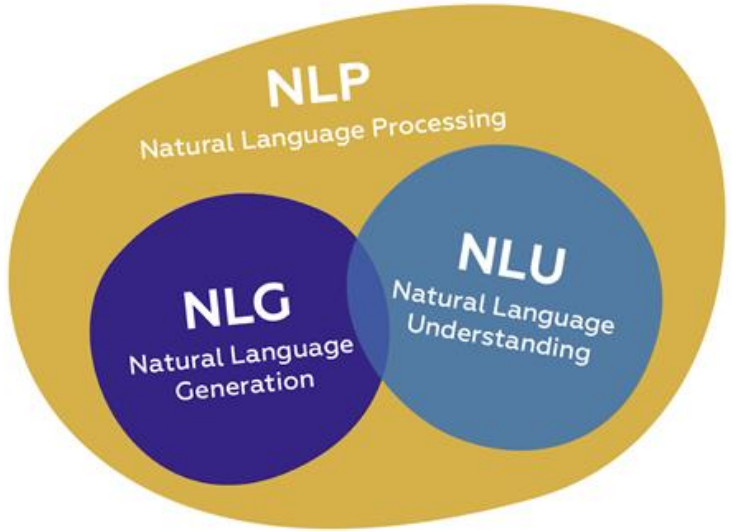
NLP is a branch of **Artificial Intelligence** that deals with the interaction between **computers** and **humans** using the **natural language.**

- The ultimate objective of NLP is to read, decipher, understand, and make sense of the **human languages** in a manner that is valuable.
- Most **challenging problems** for computers to understand natural languages as humans do.
- NLP research pursues the **vague question** of how we understand the meaning of a sentence or a document. What are the indications we use to understand who did what to whom?
- The role of NLP in text mining is to deliver the system in the information extraction phase as an input.

Useful Applications in the NLP field

- Search, spell checking, keyword search, finding synonyms, complex questions answering.
- Extracting information from websites such as: products, price, dates, locations, people or names.
- Machine translation (i.e. Google translate), speech recognition, personal assistants (think about Amazon Alexa, Apple Siri, Facebook M, Google Assistant or Microsoft Cortana).
- Chat bots/dialog agents for customer support, controlling devices, ordering goods.
- Matching online advertisements, sentiment analysis for marketing or finance/trading.
- Identifying financials risks or fraud.





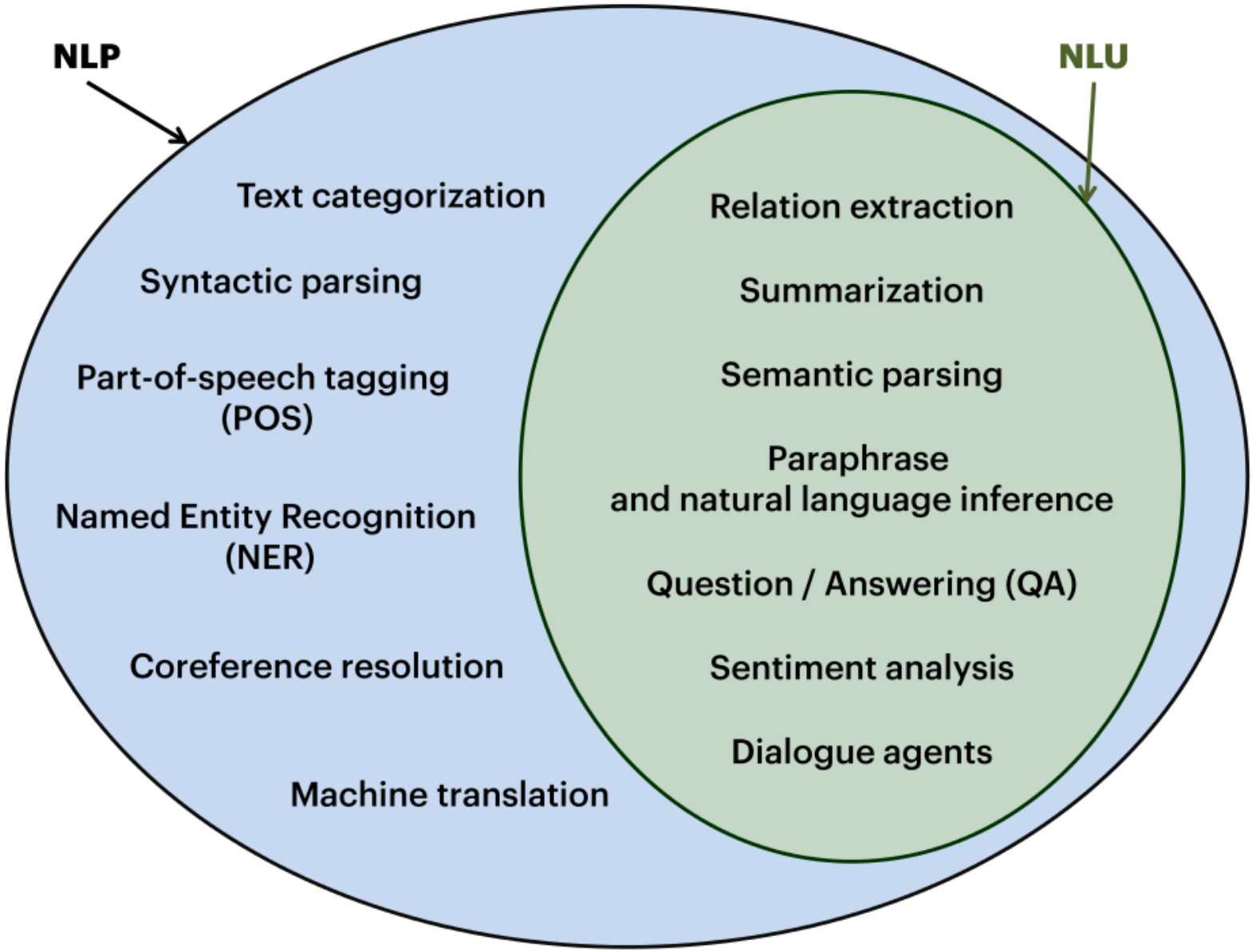
Phonology – understanding of **sound** and **speeches**.

Pragmatics – understanding the **meaning** in language in a particular **context**.

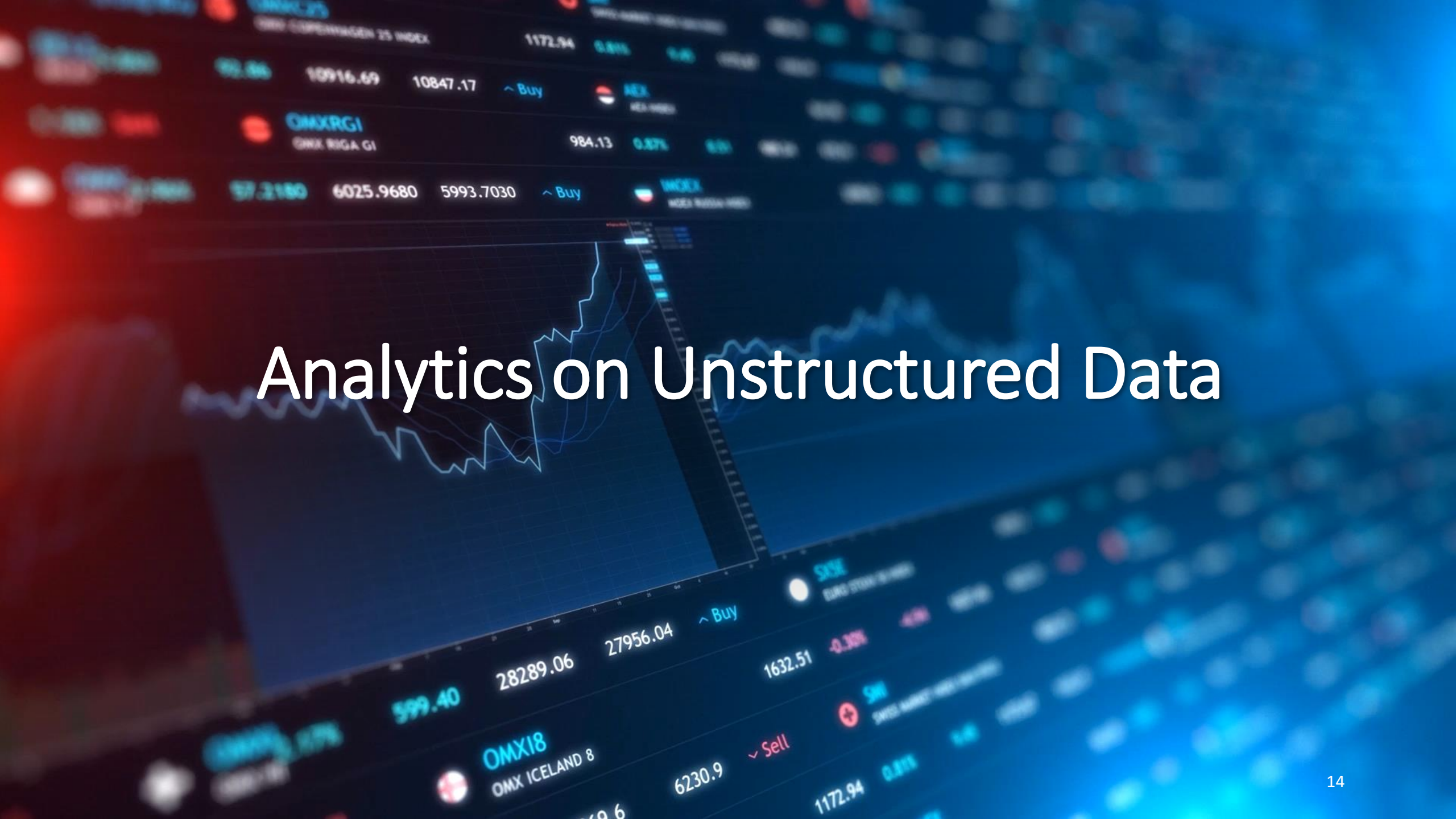
Morphology – understand the **structure** and **construction** of words.

Syntax – understanding the **grammar** of the text.

Semantics – understanding the **literal meaning** of the text.



https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf


Analytics on Unstructured Data

What is Text Analytics?

Text analytics is the **automated process**

(combines a set of **machine learning**, **statistical** and **linguistic** techniques)

of translating large volumes of **unstructured text**

(such as tweets, articles, reviews and comments) into

quantitative data

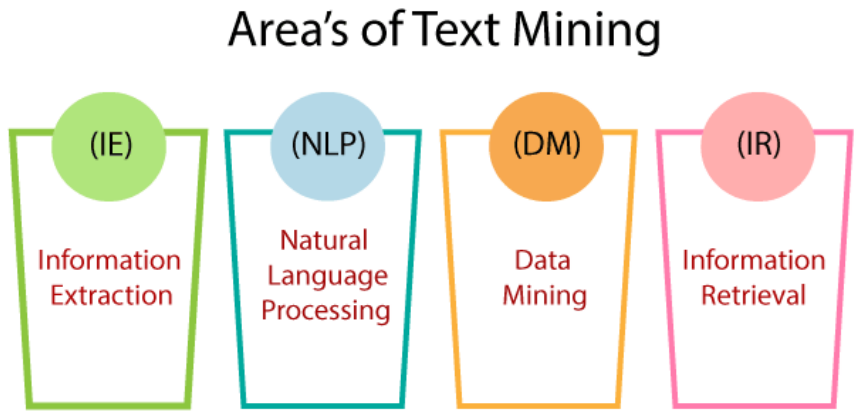
(transforming unstructured text documents into usable, structured data – structuring the unstructured!)

to uncover **insights**, **trends**, and **patterns**.

What is Text Analytics?

This process involves **3 major tasks**:

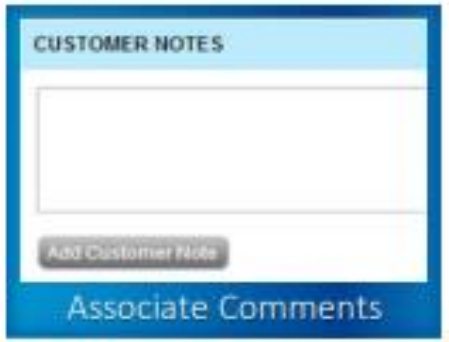
- **information retrieval** (gathering the relevant documents),
- **information extraction** (finding information of interest from these documents),
- **data mining** (discovering new associations among the extracted pieces of information).



Data mining uses **structured data**, found in most business databases.
Text mining uses **unstructured** or **semi-structured** data from a variety of sources, including media, the web, and other electronic data sources.

Common Textual Data Sources


Call Center Notes



Associate Comments


Survey Feedback

Research & Publications

Claims & Case Notes

Live Chat


Factory/Tech'n Notes

Emails


Medical/Health Records

Contracts & Applications

Online Forums

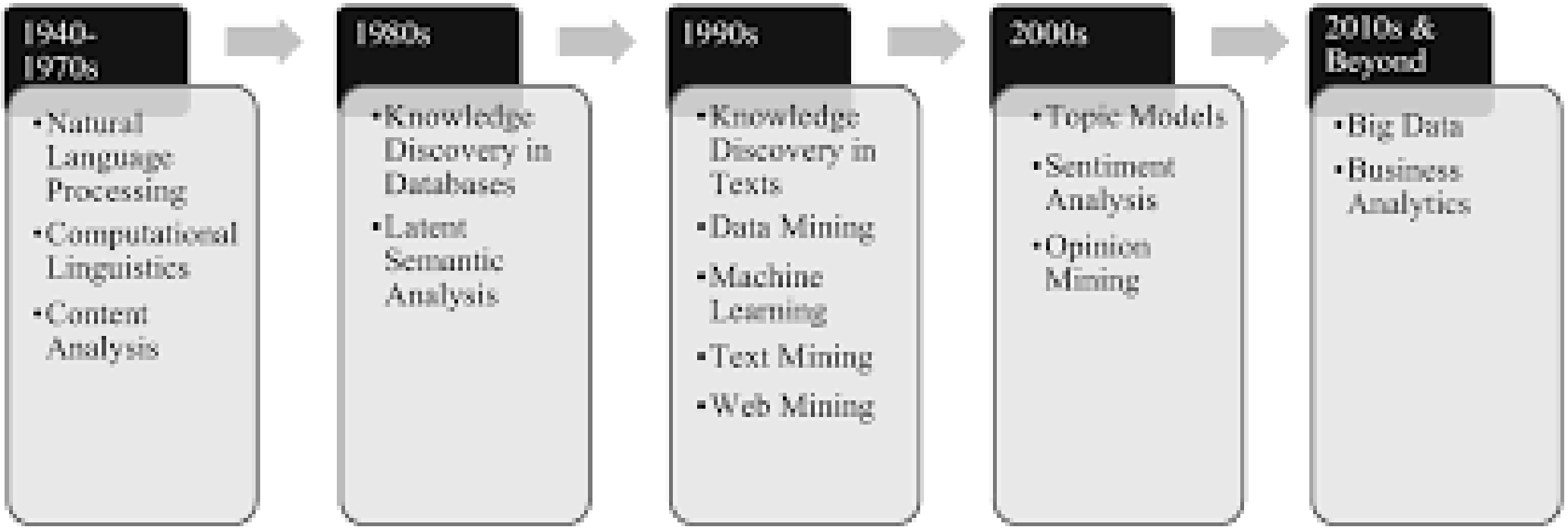
Blogs

Consumer Reviews

Online News

Social Networks

Text Analytics Timeline



Text Analytics Use Cases

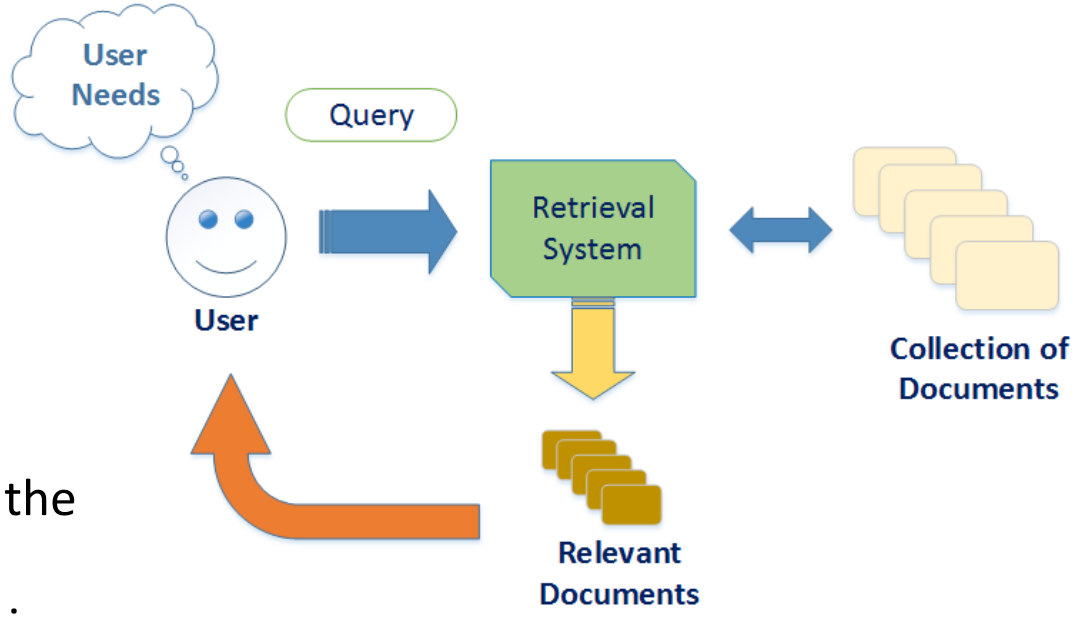
Information Retrieval (IR)

- **Information retrieval** (IR) is finding material (usually documents) of an **unstructured** nature (usually text) that **satisfies** an **information need** from within large collections (usually stored on computers).

Are the retrieved documents about the target subject **satisfies** an information need?

- up-to-date?
- from a trusted source?
- satisfying the user's needs?

An **information need** is the topic about which the user desires to know more about.



How should we rank documents in terms of these factors?

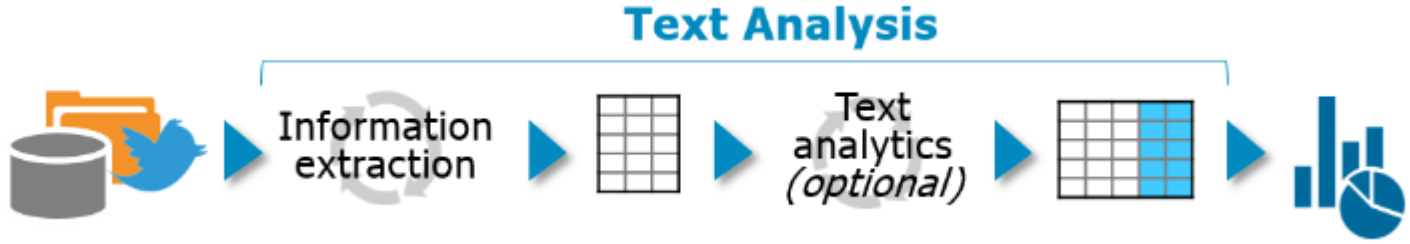
The **effectiveness** of an IR system (i.e., the quality of its search results) is determined by two key statistics about the system's returned results for a query:

Precision: What fraction of the returned results are relevant to the information need?

Recall: What fraction of the relevant documents in the collection were returned by the system?

Information Extraction (IE)

- Information Extraction is the task of automatically extracting **structured** information from **unstructured**. In most of the cases, this activity includes processing human language texts by means of **NLP**.



The first step when analyzing text is to recognize what's useful and what's not. The typical tasks are:

- ❑ Detect terms based on data types, specific words and linguistic context
- ❑ Classify and organize terms, often leveraging dictionaries or ontologies
- ❑ Describe these terms, ranging from simple rating systems to complex statistical methods

Consider this sentence: "Software revenue of $25,932 million increased 1.9 percent as reported and 3 percent adjusted for currency in 2013 compared to 2012."

[Software revenue] of [$25,932] [million] [increased] [1.9 percent] as reported and

- [Software revenue]: Revenue Type
- [$25,932]: Revenue Amount
- [million]: Revenue Units
- [increased]: Growth Rating = Modest
- [1.9 percent]: YoY Growth

3 percent adjusted for currency in [2013] compared to 2012.

- [2013]: Revenue Year

What Can We Learn From Text?

So I am prefacing this with saying that I am a huge fan of both iOS and Android phones. I got to play around with the iPhone X today and I was underwhelmed. First off, bad on Apple for branding it the iPhone 10. It should have been branded the iPhone X because that is what most people are identifying it as.
Second, this is the best display ever... on an iPhone. Compare this to the Note 8 and its not even close. I know Apple wanted to go with natural colors but natural colors look flat compared to the pop of Samsung phones.
Next, I noticed a lot of wasted space in apps. You can see it when you have the keyboard up in Safari or Messages. That is surprising to me that Apple would do that. Don't get me wrong, the iPhone X is an excellent phone, but it is not 1000 dollars excellent.

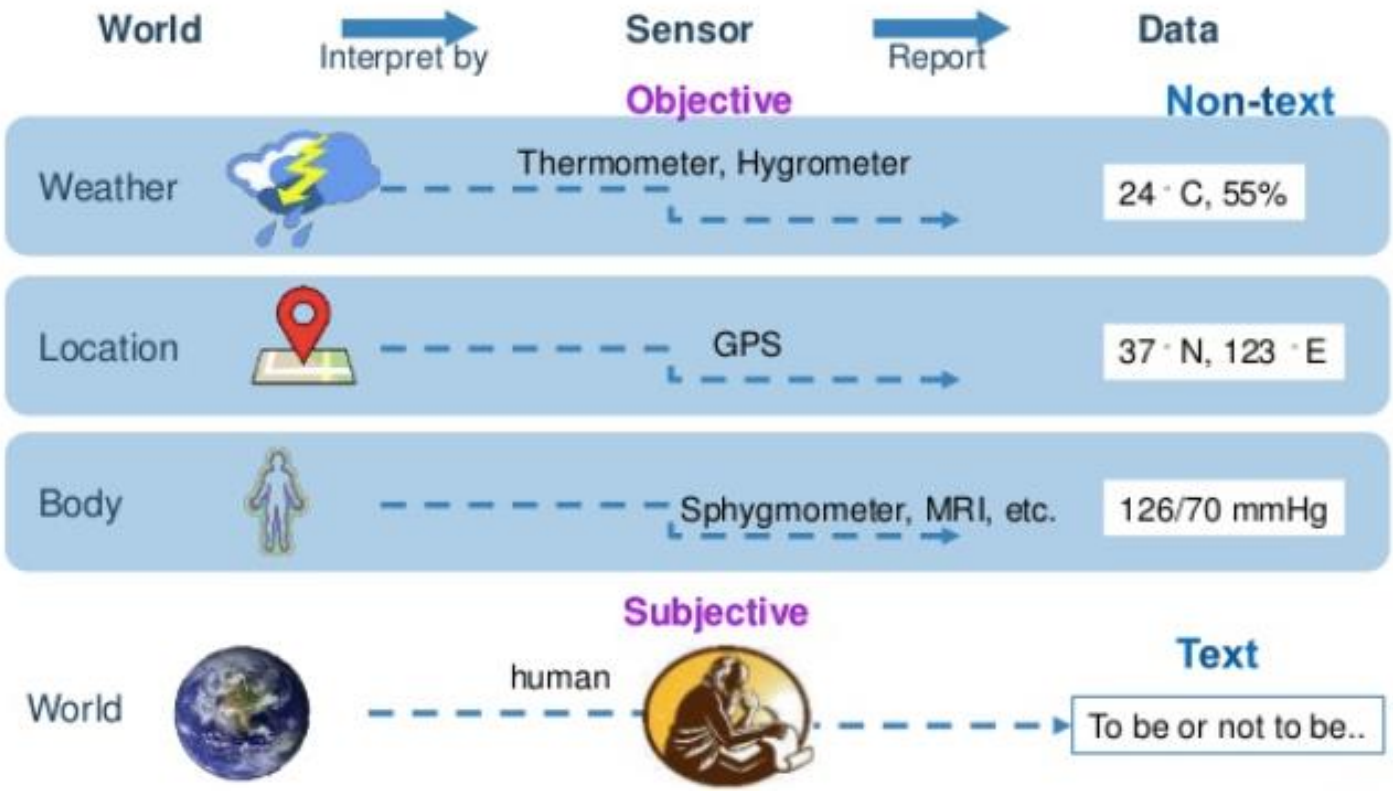
Share ›

What we can learn?

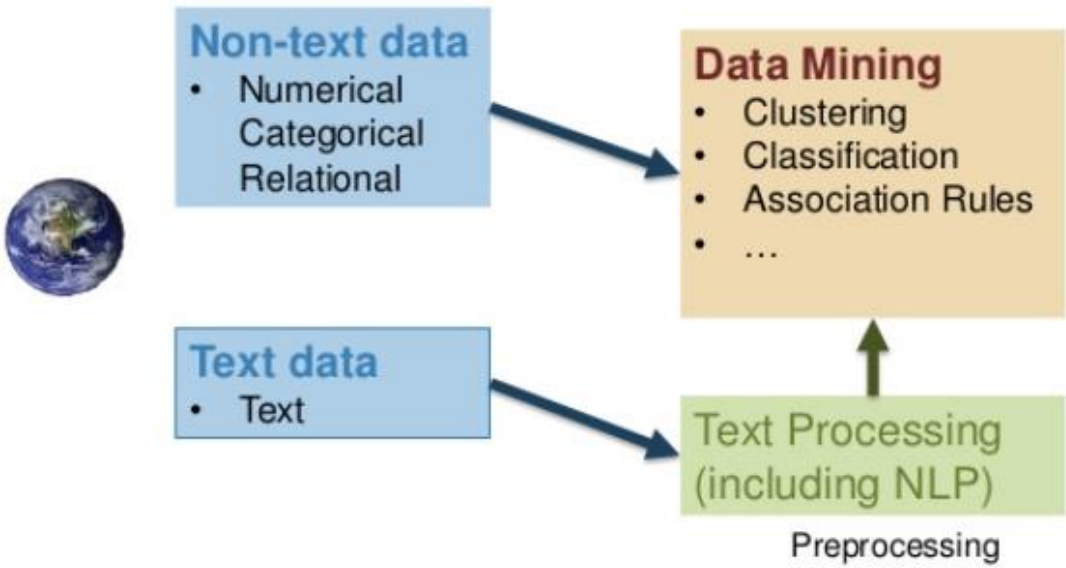
- Reviewer's Attitude
- Reviewer's Opinion on the Features (Display, Branding, Design, Price)
- Degree of their Opinion
- Competitor Information

Now, add these to your existing knowledge...

Data (Text vs. Non-Text)



Data Mining vs. Text Mining



Text Mining

Meaning
- is basically cleaning up of data to be available for text analytics.

Concept
- a tool that helps in getting the data cleaned up.

Framework
- similar to ETL (Extract Transform Load), which means to be able to insert data into database these steps are carried out.

Language
- Python and R are the most famous text mining tools out there for text mining.

Examples:
- text categorization
- text clustering
- concept/entity extraction
- sentiment analysis
- document summarization
- production of granular taxonomies
- Entity relation modeling

Text Analytics

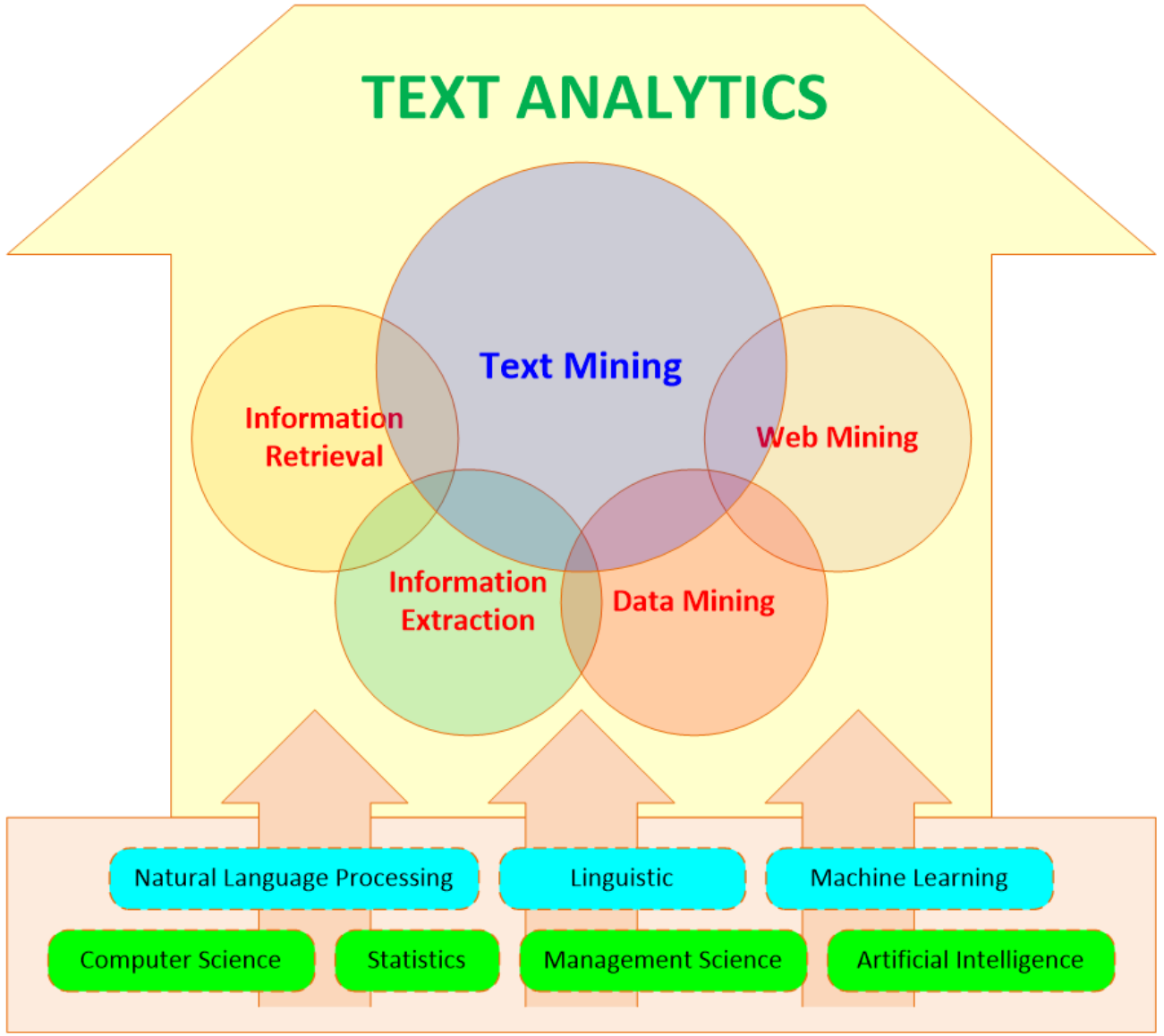
- is applying of statistical and machine learning techniques to be able to predict /prescribe or infer any information from the text-mined data.
- is the process of applying the algorithms.
- this data is used to add values to the business, example creating word clouds, bi-grams frequency charts, N-grams in some cases.
- python and R, including others e.g., Power BI, Azure, KNIME, etc.

Examples:
- Association analysis
- visualization
- predictive analytics
- information retrieval
- lexical analysis
- pattern recognition
- tagging/annotation

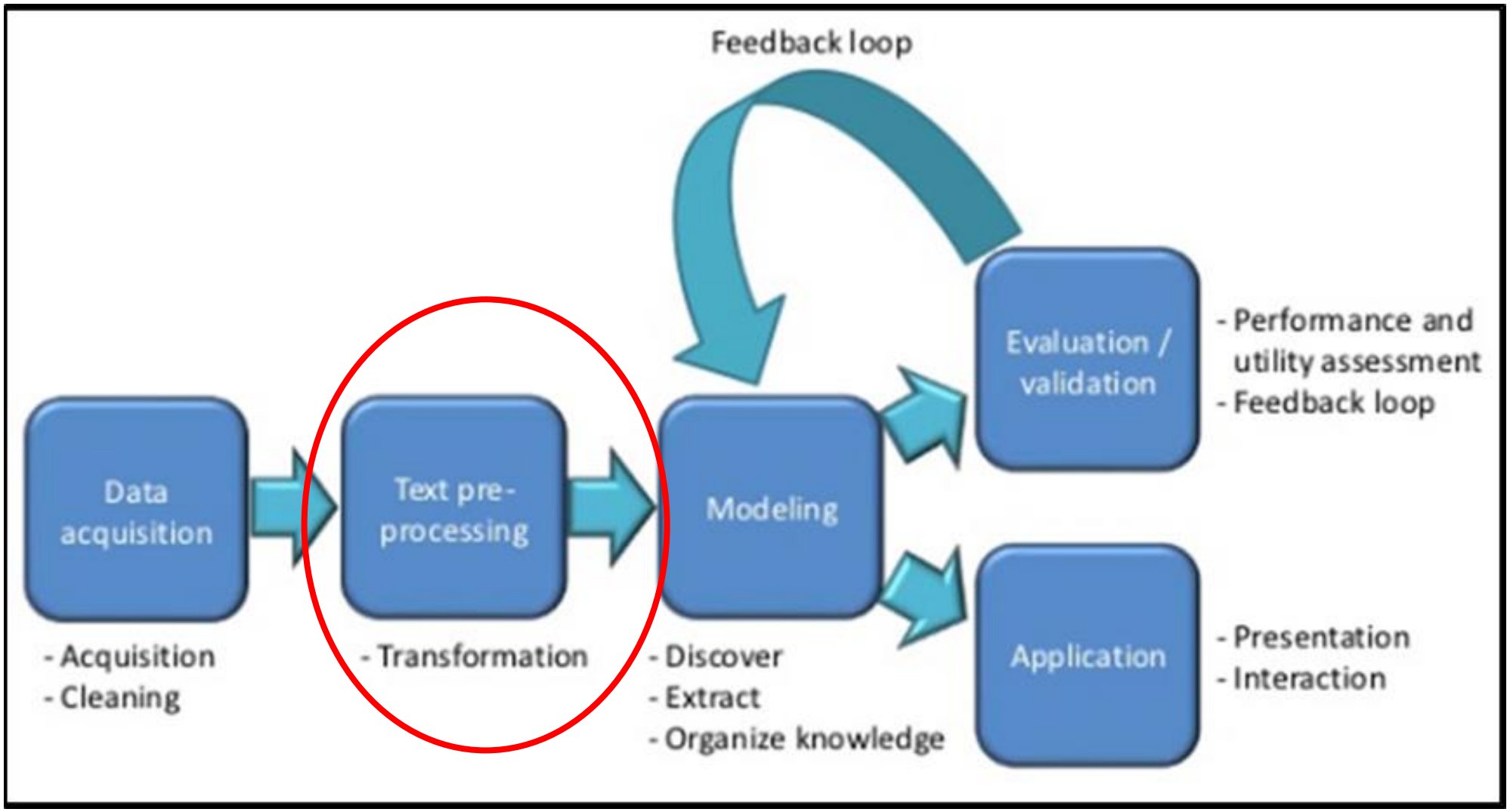
https://www.educba.com/text-mining-vs-text-analytics/

Text mining is pre-processed data for text analytics.

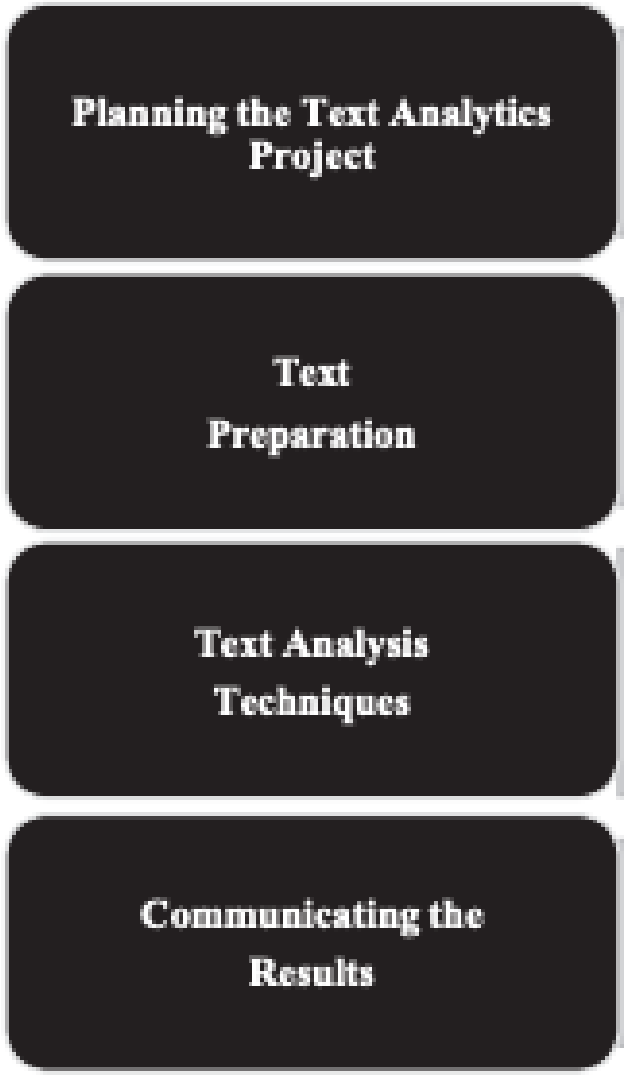
In **Text Analytics**, statistical and machine learning algorithm used to classify information.

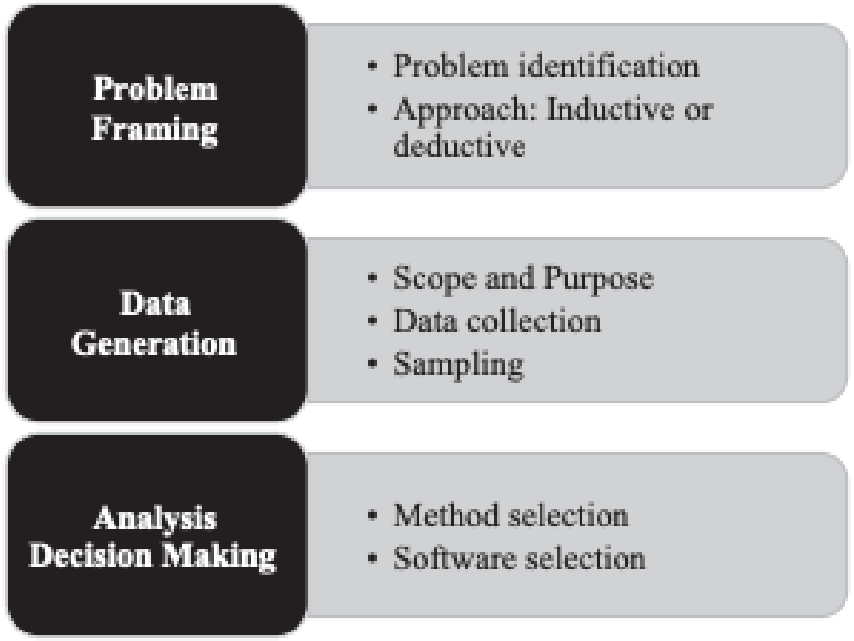


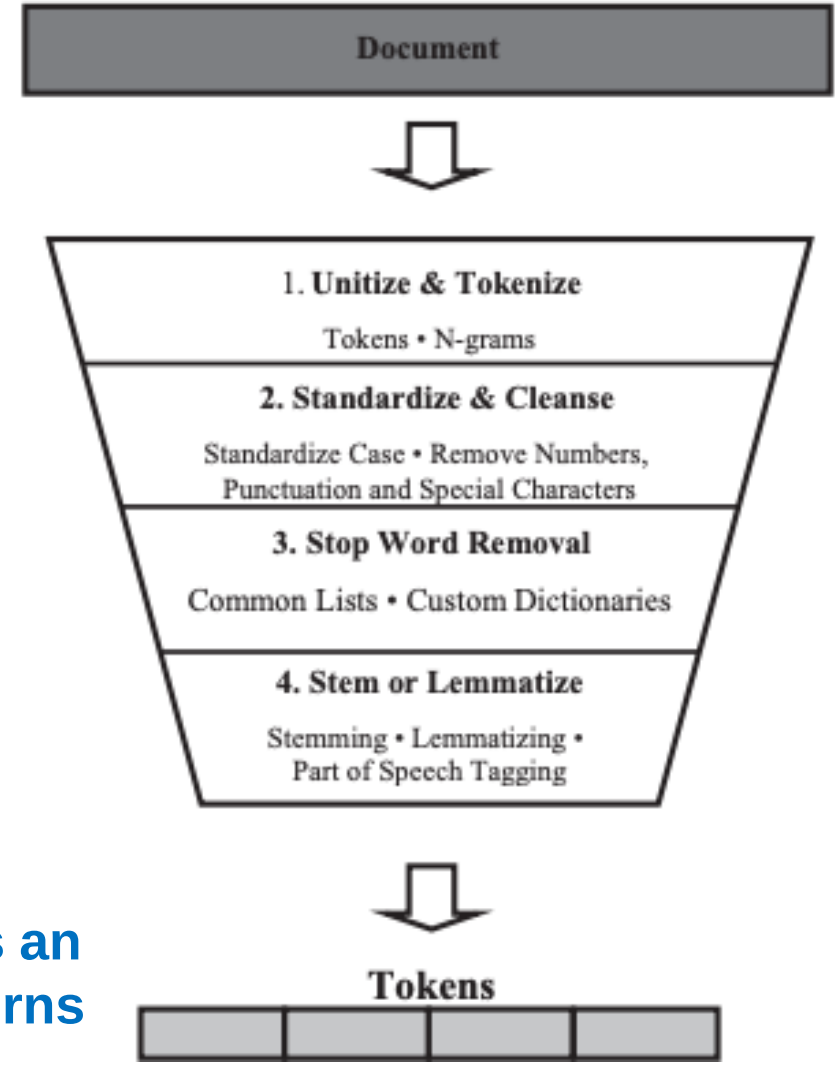
Typical Text Mining Process



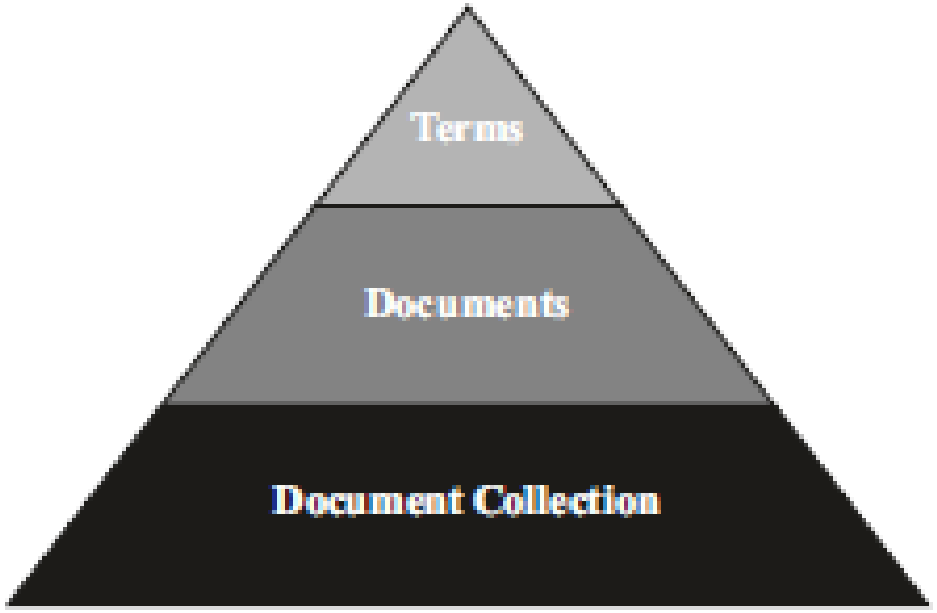
Text analytics process into four major steps: (1) planning, (2) preparing and preprocessing, (3) analysis, and (4) reporting.







Text preprocessing takes an input of raw text and returns cleansed tokens.



Hierarchy of terms and documents

A **document** is typically made up of many characters. The many documents make up a **document collection** or **corpus**.
Characters are combined to form words or **terms** in a given language. These words are the focus of our analysis, although groupings of terms can also be the chosen unit of analysis. The collection of terms is sometimes called the **vocabulary or dictionary**.

DOCUMENT 1
My favorite dog is fluffy and tan.
DOCUMENT 2
the dog is brown and cat is brown
DOCUMENT 3
My favorite hat is brown and coat is pink
DOCUMENT 4
My dog has a hat and leash. ♥
DOCUMENT 5
He has a fluffy coat and brown coats.
DOCUMENT 6
The dog is brown and fluffy & has a brown coat.
DOCUMENT 7
MY dog is white with brown spots.
DOCUMENT 8
The white dog has a Pink coat and the Brown dog is fluffy
DOCUMENT 9
The 3 fluffy dogs AND 2 brown hats are my favorites!
DOCUMENT 10
MY fluffy dog has a white coat and hat .

Fig. 4.2 Example document collection

Text Pre-processing

(1) **Tokenization**

Difficult for machines to understand the semantics and context of a document, a paragraph, or even a sentence as a whole.

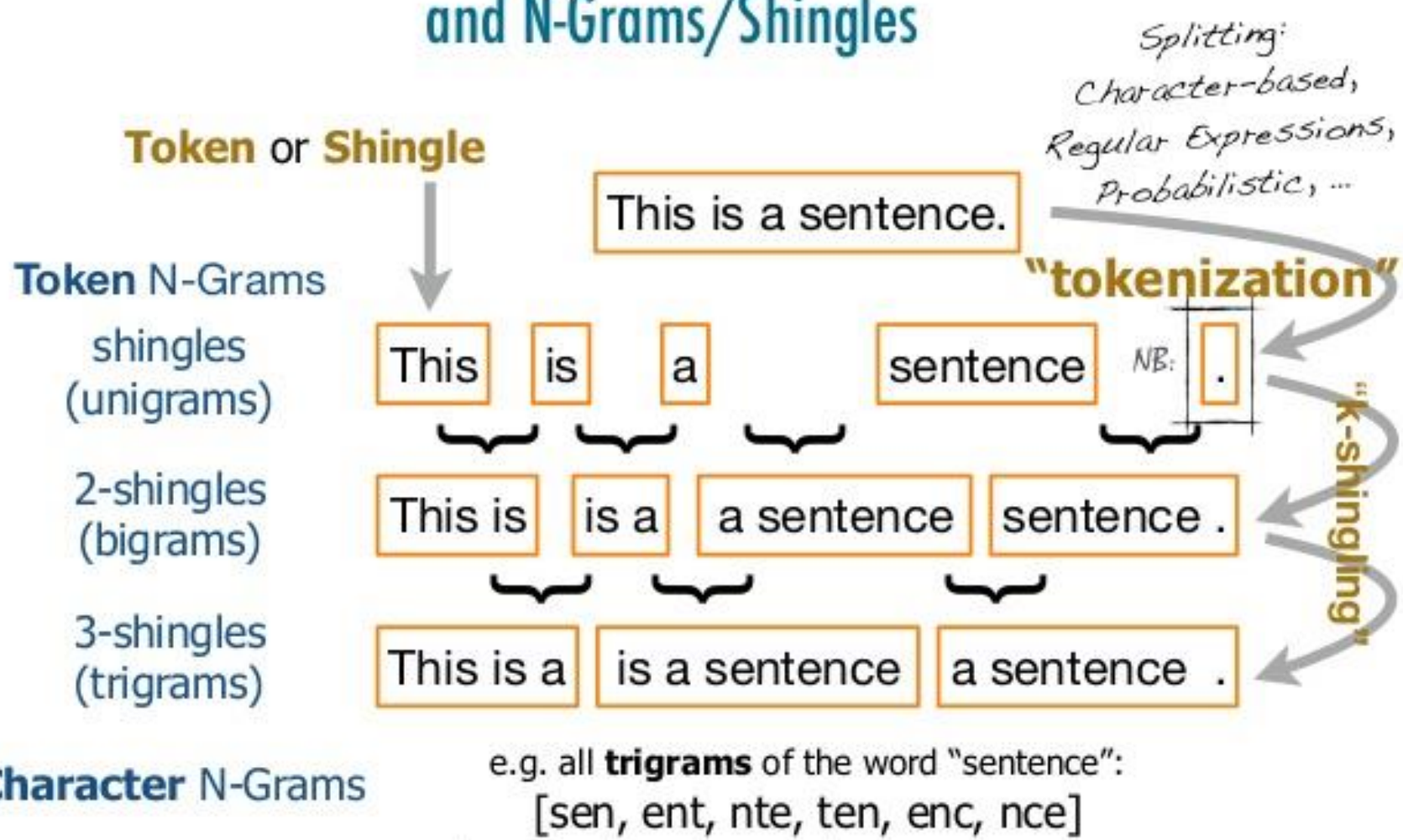
- So utilize the **tokenization process** to break them to down their smallest semantic units call ***token***.
- Eg. Given a sentence "*Dr. Salimah is an associate professor*", we can break them down into six tokens.

Dr. / Salimah / is / an / associate / professor

Dr.	Salimah	is	an	associate	professor

Tokenization splitting an English sentence (text) into a list of tokens using space

Words, Tokens, and N-Grams/Shingles



Character N-Grams

e.g. all **trigrams** of the word "sentence":

[sen, ent, nte, ten, enc, nce]

Snag: the terms "shingle", "token" and "n-gram" are not used consistently...
but "n-gram" and "token" are far more common!

Text Pre-processing

(2) Removing unnecessary **punctuation**, **tags**

```
Input : %welcome' to @geeksforgeek<s
Output : welcome to geeksforgeeks

Input : Hello!!!, he said ---and went.
Output : Hello he said and went
```

Dr. / Salimah / is / an / associate / professor.

(3) **Removing stop words** — frequent words such as, ”the”, ”is”, etc. that do not have specific semantic.

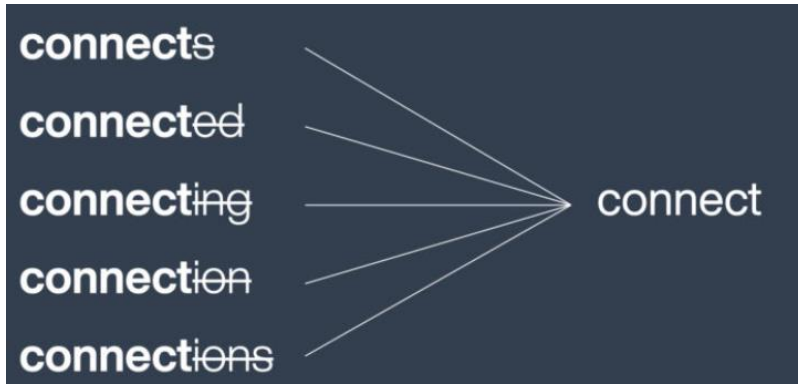
Stop words refer to terms that appear frequently in a natural language (e.g. is, am, a, in) but do not aid in understanding the meaning.
Removing stop words prevents high occurrence words from taking up space in our database and taking up the valuable processing time.

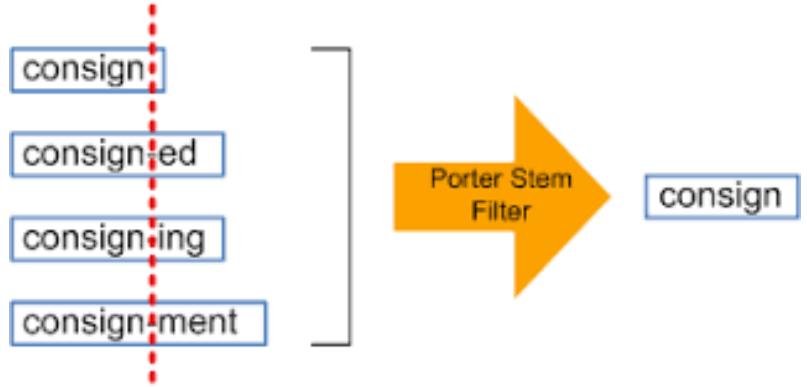
Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

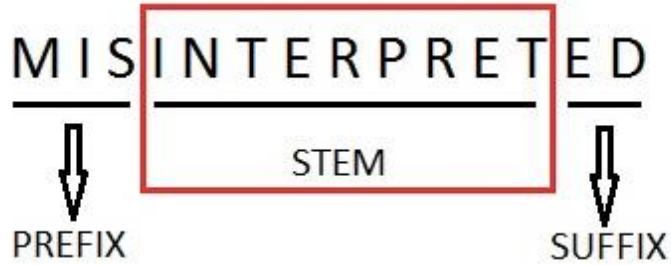
Text Pre-processing

(4) Stemming — converts a word into its **stem**(root form).

- Stemming is a process of obtaining the root word (**stem**) from a given word by eliminating its affixes (suffixes and sometimes prefixes).
- For example: Common suffix like: “es”, “ing”, “pre” etc.
  - A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “**chocolate**” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “**retrieve**”.







Common stemmer for English is **Porter Stemmer**.

Text Pre-processing

(5) Lemmatization —considers the context **and** converts the word to its meaningful base form, which is called Lemma(root form).

- It usually refers doing things properly with the use of a vocabulary and **morphological analysis** of words.
- It observes position and **Parts of speech (POS)** of a word before striping anything.

```
Examples of lemmatization:

-> rocks : rock
-> corpora : corpus
-> better : good
```

In linguistics, **morphology** is the study of words, how they are formed, and their relationship to other words in the same language.

Stemming

adjustable → adjust
formality → formaliti
formaliti → formal
airliner → airlin ⚠

Lemmatization

was → (to) be
better → good
meeting → meeting

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Text Pre-processing

(6) Part-of-speech (POS) Tagging

POS tagging is to identify the **role of a word** in a sentence.

This is based on the relationship and context with the surrounding words.

Part of speech includes nouns, verbs, adverbs, adjectives, pronouns, prepositions, and other sub-categories.

Part of speech (or POS) tagging marks the role of a word in a sentence

(7) Term frequency – Inverse Document Frequency (TF-IDF) is a numerical statistic that is intended to reflect **how important a word** is to a document in a collection or corpus.

Term frequency (TF) shows how frequently an expression (term, word) occurs in a document.

TF-IDF is not only counting the term frequency but also assigning different weights to each term according to the importance of the term to a document.

TF = (Number of time the term occurs in the document) / (Total number of terms in the document)
IDF = (Total number of documents) / Number of documents with term *t* in it)
TF-IDF = TF * IDF

Document A: The car is driven on the road.
Document B: The truck is driven on the highway.

From the table, we can see that TF-IDF of common words was zero, which shows they are not significant.
On the other hand, the TF-IDF of “car” , “truck”, “road”, and “highway” are non-zero. These words have more significance.

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

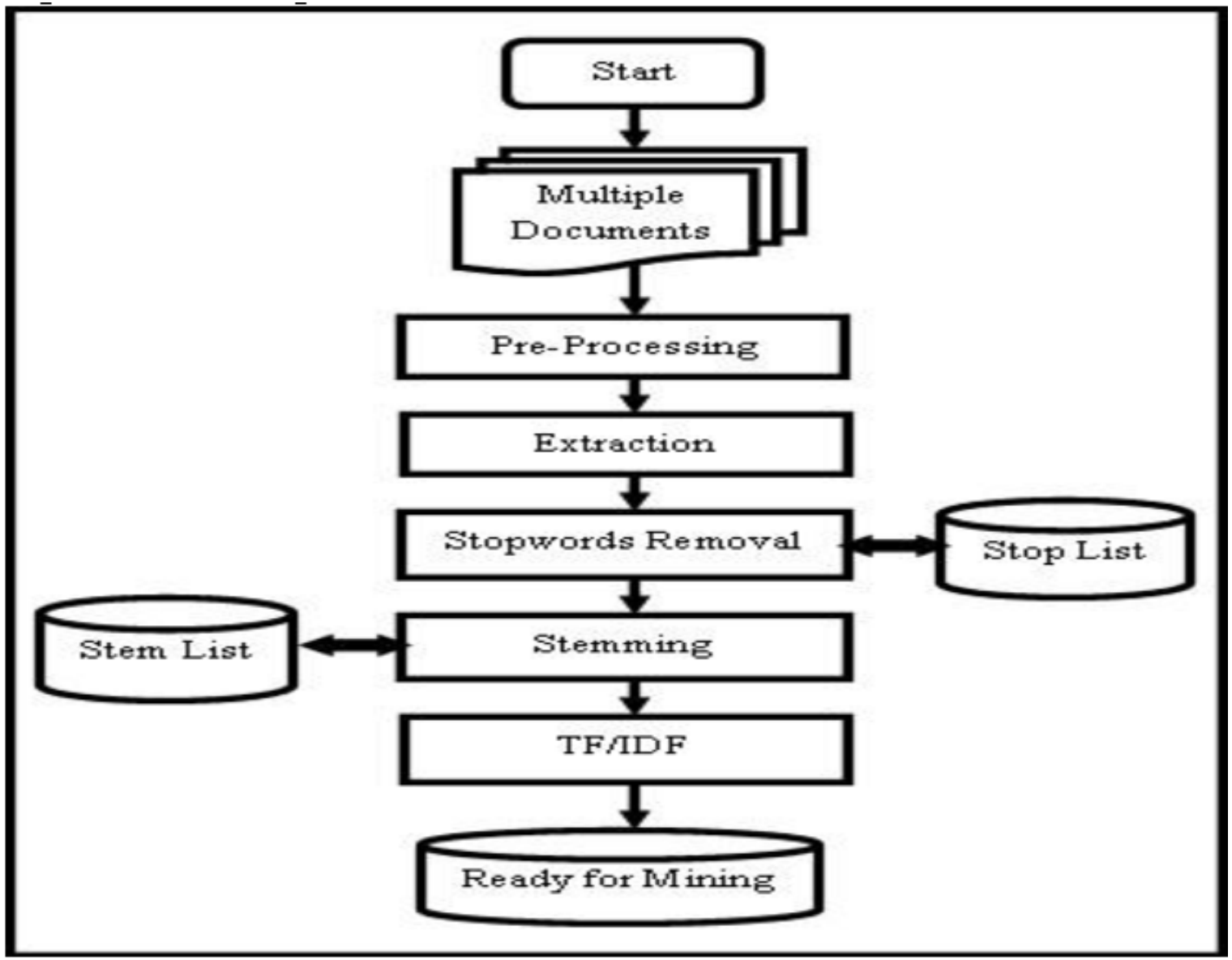


Figure 3. Text Mining Pre-Processing Techniques

Text Mining Techniques	Key Considerations
Organizing and Structuring Content	• Clustering • Categorization • Classification • Taxonomy
Text Processing	• Natural Language Processing (NLP) • Parsing • Tokenization • Stemming • Term Reduction • Parts-of-Speech (POS) Tagging
Statistical Analysis	• Term Frequency • Keyword Frequency • Distribution • Document Term Matrix (DTM) • Term Frequency – Inverse Document Frequency (TF-IDF) • Document Indexing
Machine Learning	• Clustering • Classification • Association Rules • Predictive Modeling
Classification Methods	• Naïve Bayes • Support Vector Machines • K-nearest neighbor
Model Evaluation	• Precision • Recall • Accuracy • Relevance

Text Analytics Software Solutions

Tools, servers, analytic algorithm-based applications, data mining and extraction tools for converting unstructured data into meaningful data for analysis.

Top Text Analytics Software

SAS Text Miner, Provalis Research Text Analytics Software, Google Cloud Prediction API, DiscoverText, IBM SPSS Text Analytics, Datumbox, Lexalytics Salience, Microsoft Azure Text Analytics API, Sysomos, IBM Watson Natural Language Understanding, Expert System, Twinword, UltiPro Perception, Google Cloud Natural Language API, Thematic, OpenText, Stratifyd, Semantria for Excel, Narrative Science Quill, indico, Clarabridge, MeaningCloud, Bitext, LingPipe, Medallia, Luminoso, STATISTICA Text Miner, Pingar DiscoveryOne, Brainspace, Keatext, Abzooba, OdinText, Semantria, NetOwl, TheySay, WordStat, Etuma, Smartlogic, Synapsify, Megaputer, AYLIEN, Text2data, Rosette Text Analytics, Oracle Endeca Information Discovery, Averbis, Angoos KnowledgeREADER, Buzzlogix, VOZIQ, Verint Systems, Aspect NLU, General Sentiment, Loop Cognitive Computing Appliance, Oracle Social Cloud, TextualETL, muText, ai-one, Ascribe, Language Computer, Open Calais, Semantria API, Rocket Search and Text Analytics, Intellexer, SYSTRAN, Relativity Analytics, SAP HANA Text Analytics, and AUTINDEX are some of the top software for Text Analysis, Text Mining, Text Analytics.

Top Free Text Analytics Software

General Architecture for Text Engineering – GATE, RapidMiner Text Mining Extension, KH Coder, Coding Analysis Toolkit, QDA Miner Lite, VisualText, TAMS, Datumbox, Carrot2, Natural Language Toolkit, Apache Mahout, Pattern, Textable, Twinword, Apache OpenNLP, Apache UIMA, Aika, tm – Text Mining Package, KNIME Text Processing, LingPipe, Gensim, Distributed Machine Learning Toolkit, LPU, Apache Stanbol, S-EM, LibShortText, and Coh-Metrix are some of the top Free Text Analysis, Text Mining, Text Analytics Software.

What You Want To Look For?
Fully Customizable, Balanced Approach to Text Analytics Problem

In order to gain business value from **unstructured text**, we need:

Structured + Unstructured

- To gain the lift in predictive accuracy from text

Supervised + Unsupervised

- To shorten the Time to Value, minimizing the manual effort while maintaining granularity and specificity

Linguistic Rules + Statistical Model

- To get desired level of granularity and customization ability

What Makes Text Analytics Hard? And Interesting …

Problem Specific

Domain Specific

- Out-Of-Box tool does not work
- Ability to customize is critical

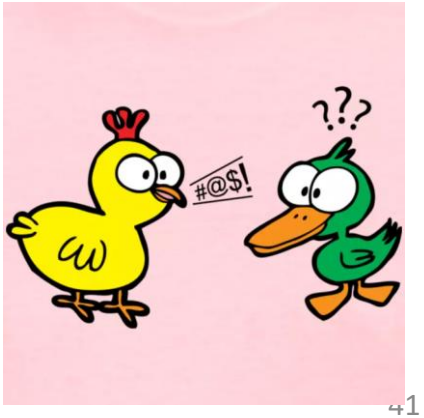
Same meaning, different words: Australians, Britons and Americans use the terms “daks,” “trousers” and “pants” respectively to refer to the same piece of clothing.

Human Subjectivity

- Same message conveyed in different ways
- Exactly the same statement in a different context may convey completely different meaning.

Language & Cultural Specific

- Requires deep knowledge about the language and the culture
- Social media as a new source of data



What Are Others Using Text Analytics For?

Analytics	Deployment	Business Use Cases		
Topic Detection	**Sentiment Dashboard** Community View Product Line View Region View Business Line View	**Product** • Design Feature/Function • Competitive Landscape	**Operations** • Improve Call Center Agent-Customer Interactions • Improve Call/Support Agent Productivity and Workflow • 1:many / 1:1 Messaging via Social Channels	**Risk** • Detect regulatory and compliance violations • Detect common themes in cases of fraudulent charges, ID theft, scams, and phishing schemes • Enhance credit scoring & underwriting models
Sentiment Analysis				
Information Retrieval	**Event Stream Processing** Risk Alert Competitive Alert Customer Interactions	**Marketing** • Reduce Attrition; Identify At-Risk Clients • Enrich Customer segmentation models and "Path Analysis" • Root Cause Analysis on Customer Complaints		
Content Categorization	**Visualization & Reporting**			
Trend Analysis	**Enriched Model**			
Root Cause Analysis	**Case Management**			

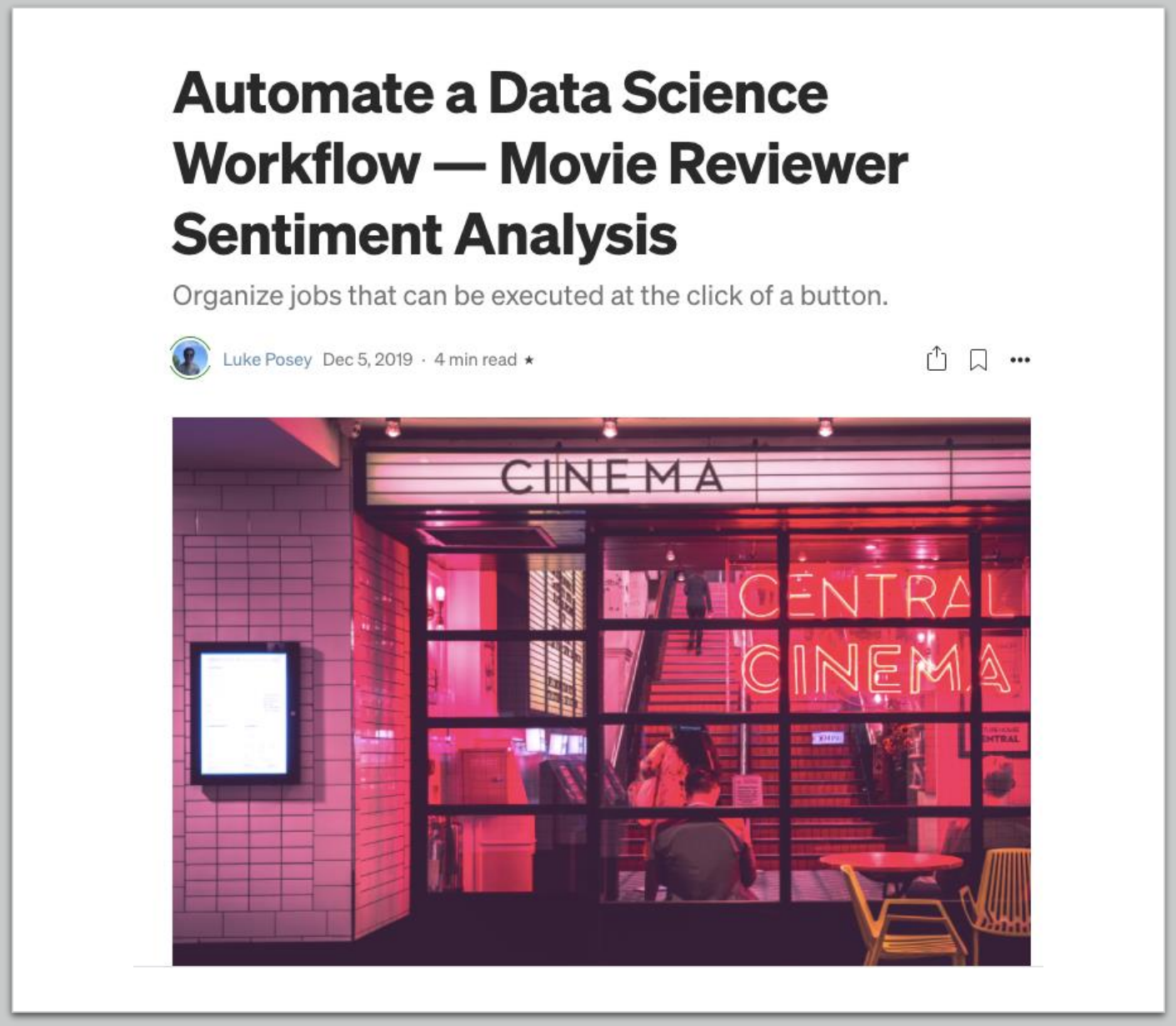
Text analytics is the process of analyzing unstructured text, extracting relevant information, and transforming it into useful business intelligence.
Sentiment analysis determines if an expression is positive, negative, or neutral, and to what degree.

Text analytics gives you the ***meaning***.
Sentiment analysis gives you ***insight*** into the *emotion* behind the words

SENTIMENT ANALYSIS

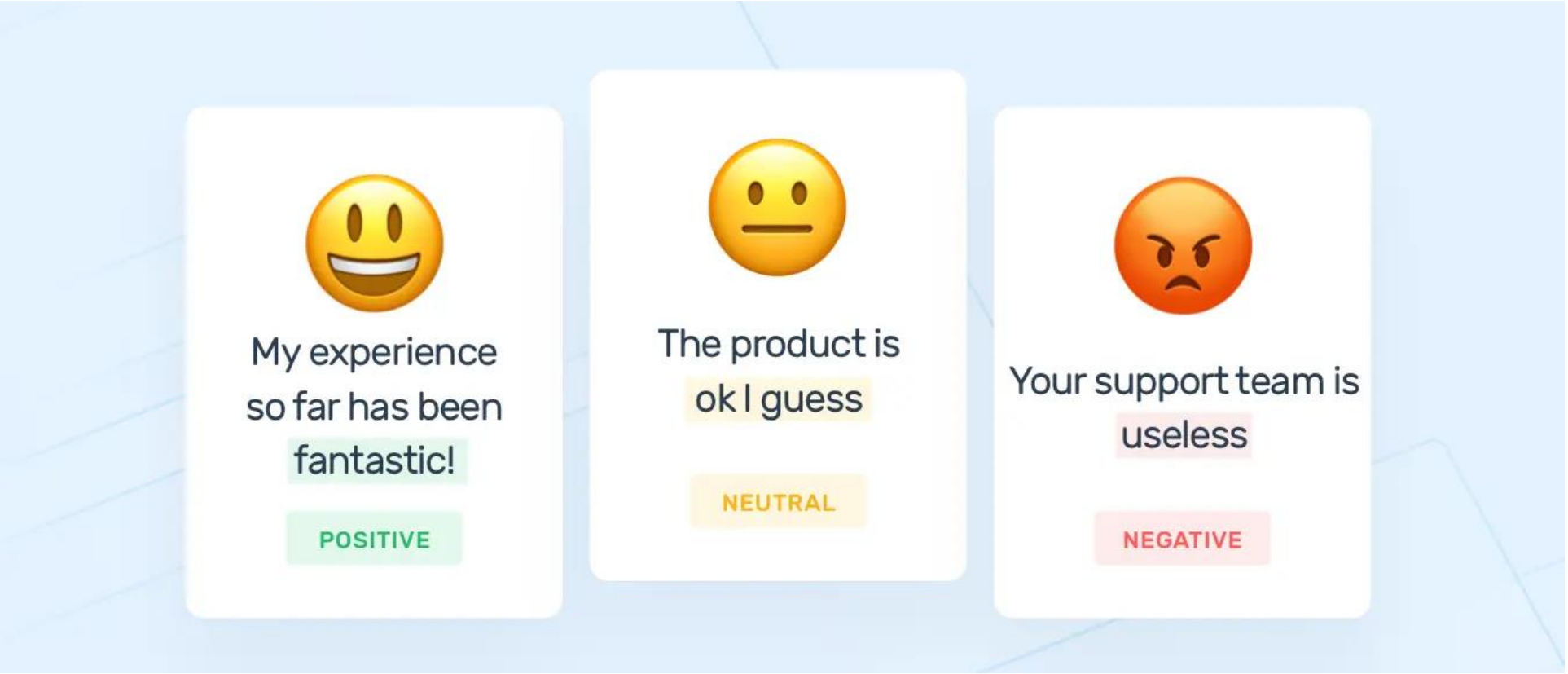
Sentiment Analysis (also known as **opinion mining** or **emotion AI**) is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news etc.

- Sentiment Analysis can help craft all this exponentially growing ***unstructured text*** into ***structured data*** using NLP and open-source tools.



Sentiment Analyzer

https://monkeylearn.com/sentiment-analysis-online/



https://monkeylearn.com/sentiment-analysis/

Sentiment Analysis

Sentiment → belief, view, opinion, conviction

- Aka opinion mining, opinion extraction, subjectivity analysis, and appraisal extraction.
- Detection of **attitude** towards objects / person.
- Study sentiment (**internal** or **external**) expressed in **text**.

The **goal** is to answer the question: *"What do people feel about a certain topic?"*

Usage or applications:

- Individual – make decision (e.g. on products)
- Business – benchmark products / services (market intelligence)
- Product rating
- Politics
- Public sentiment
- Opinion retrieval – provide general search for opinions.

Sentiment polarity

- Positive versus Negative
- … versus Neutral?

- Opinions from many people → opinion summarization → decision → action
- See Twitter Sentiment App (www.sentiment140.com)

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text.

Two Types of Opinions

- **Regular opinions**: Sentiment/opinion expressions on some target entities
  - Direct opinions:
  - "The touch screen is really cool."
- **Indirect opinions**:
  - "After taking the drug, my pain has gone."
  - Comparative opinions: Comparison of more than one entity.
  - E.g., "iPhone is better than Blackberry."
- We focus on **regular opinions**, and just call them opinions.

Free Sentiment Analysis Demo

https://text2data.com/Demo

Please enter your text in **english*** for analysis or leave default one.

I would like to express my greatest appreciation for the great learning experience and your dedication in conducting lectures.

Twitter-like content

SHARE THIS ANALYSIS

RUN ANALYSIS

I would like to express my greatest appreciation for the great learning experience and your dedication in conducting lectures.

great learning experience
greatest appreciation
greatest great

This document is: **positive (+1.00)**

*Magnitude: **0.50***

Subjectivity: **subjective**

Please enter your text in **english*** for analysis or leave default one.

You're the very most interactive and insightful educator and has strong desire to care deeply about the students and work hard to make sure no one is left behind.

Twitter-like content

SHARE THIS ANALYSIS

RUN ANALYSIS

You're the very most interactive and insightful educator and has strong desire to care deeply about the students and work hard to make sure no one is left behind.

care strong insightful

This document is: **positive (+0.98)** *Magnitude:* ***0.80***

negative neutral **positive**

Score Range -1 -0.25 +0.25 +1

Subjectivity: **subjective**

Please enter your text in **english*** for analysis or leave default one.

The COVID-19 pandemic has led to a rise in fear, anxiety, stress, and depression among the population.

Twitter-like content

SHARE THIS ANALYSIS

RUN ANALYSIS

The COVID-19 pandemic has led to a rise in fear, anxiety, stress, and depression among the population.

This document is: **negative (-0.93)** *Magnitude:* ***0.79***

Subjectivity: **objective**

Please enter your text in **english*** for analysis or leave default one.

I'm not sure if I like the new design

Twitter-like content

SHARE THIS ANALYSIS

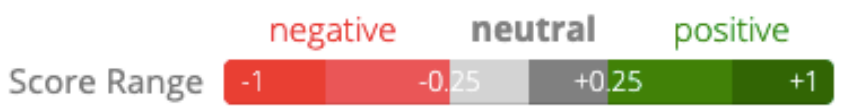
RUN ANALYSIS

I'm not sure if I like the new design

This document is: **neutral (+0.16)**

Magnitude: ***0.37***

Subjectivity: **subjective**



Types of Sentiment Analysis

These are the most common types of sentiment analysis:

- ❖ Standard sentiment analysis
- ❖ Fine-grained sentiment analysis
- ❖ Emotion detection
- ❖ Aspect-based sentiment analysis
- ❖ Intent detection
- ❖ Sarcasm detection

Source: https://monkeylearn.com/blog/sentiment-analysis-examples/

Standard Sentiment Analysis

It identifies the nuance of an opinion and classifies it as *Positive*, *Negative*, or *Neutral*. It's the most popular type of sentiment analysis. For example:

- *'I love how Zapier takes different apps and ties them together'* → `Positive`
- *"I still need to further test Zapier to say if its useful for me or not'* → `Neutral`
- *"Zapier is sooooo confusing to me'* → `Negative`

Fine-grained Sentiment Analysis

Also focused on polarity, this type of sentiment analysis adds a few more categories to obtain more granular results. Similar to 5-star ratings, it classifies opinions as:

- Very positive
- Positive
- Neutral
- Negative
- Very negative

For example, imagine having the following survey responses:

- *'The older interface was much simpler'* → `Negative`
- *'Awful experience. I would never buy this product again!'* → `Very Negative`
- *'I don't think there is anything I really dislike about the product'* → `Neutral`

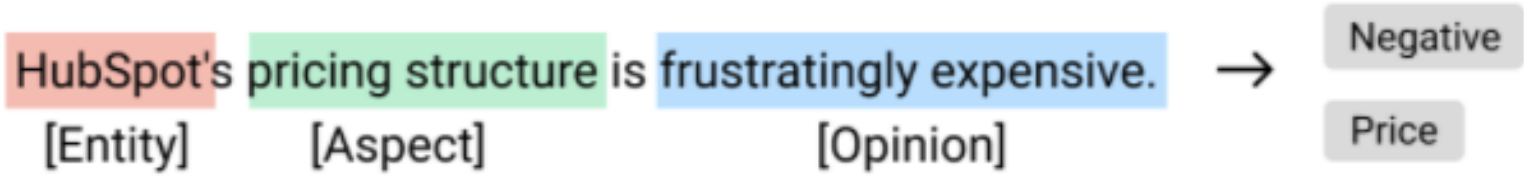
Emotion Detection

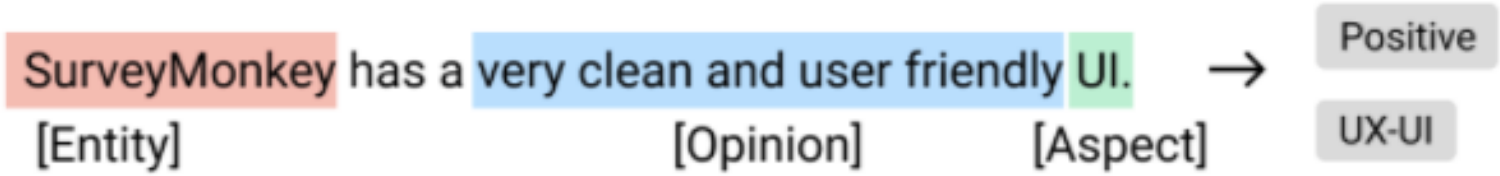
This sentiment analysis model detects the emotions that underlie a text. It makes associations between words and emotions like anger, happiness, frustration, etc. For example,

- *'Hubspot makes my day a lot easier :)'* → `Happiness`
- *'Your customer service is a nightmare! Totally useless!!'* → `Anger`

Aspect-based Sentiment Analysis

This type of sentiment analysis focuses on understanding the aspects or features that are being discussed in a given opinion. Product reviews, for example, are often composed of different opinions about different characteristics of a product, like `Price`, `UX-UI`, `Integrations`, `Mobile Version`, etc. Let's see some examples:





Intent Detection

This type of sentiment analysis tries to find an action behind a given opinion, something that the user wants to do. Identifying user intents allows you to detect valuable opportunities to help customers, such as solving an issue, making improvements on a product or deriving complaints to the correspondent areas:

- *"Very frustrated right now. Instagram keeps closing when I log in. Can you help?"* → `Request for Assistance`

Customers experiencing issues can be easily spotted thanks to sentiment analysis.

Sarcasm

- Greek: *sarkázein* (speak bitterly, use of irony to mock)
  French: *sarcasme*
- Nuanced form of language where often the speaker explicitly states the opposite of what she implies.
- Deliberately mean opposite of what is on the surface.

  "This talk looks like great fun ;)"

Most sentiment analysis system **fail to detect sarcasm** and thus **wrongly infer** the sentiment.

- Most sarcastic sentences show a shift in sentiment

  "I love the pain present in the breakups"

  → (shift in sentiment)
- There is a contradiction between sentiment of "love" and "pain of breakups". This is a hallmark of sarcasm.

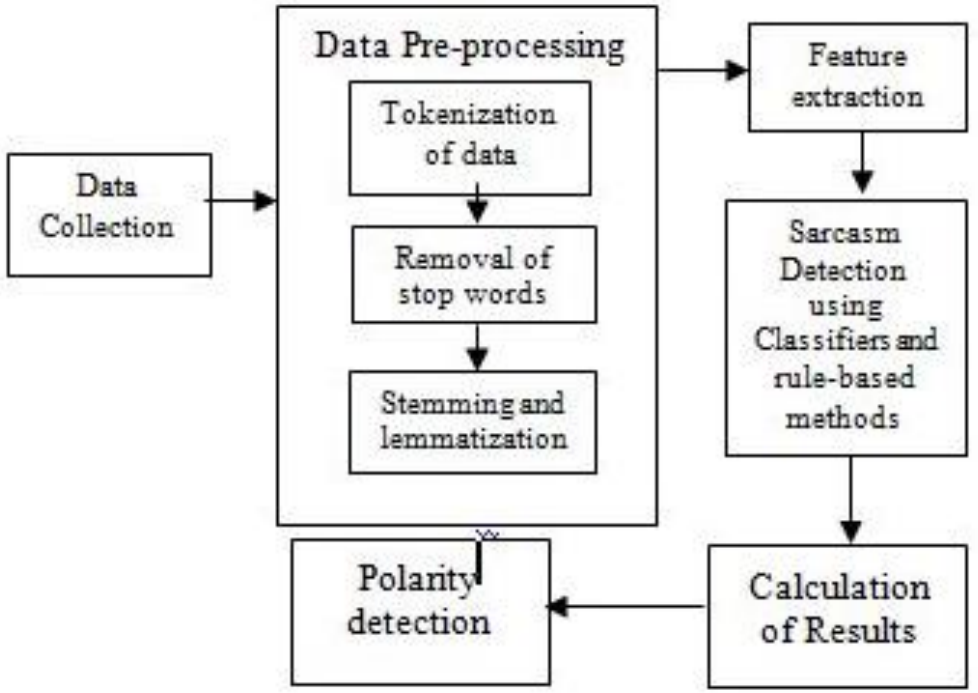
- Emotion: feelings such as happiness, anger, jealousy, grief, etc. One can have many emotions simultaneously. Subjective in nature.
- Sentiment: opinion or mental attitude produced by emotions about something. This is much more objective.
- Sarcastic sentences are rich in emotions.

Personality

- There is a body of work that argues that sarcasm is not just a linguistic phenomena but also a behavioral phenomena i.e. it not just about what is being said but also who is saying that is super important.
- i.e. sarcasm is user specific: some users have a stronger tendency to be sarcastic as compared to others*.

Can Sentiment Analysis Identify Sarcasm?

- Analysis is for **positive-negative sentiment** rather than **mood detection**.
- There is always **some degree of imprecision** in sentiment analysis, but the model is most useful when there is no hidden meaning or subtext to the content.
- Irony, sarcasm, humor, and similarly nuanced content rely on **cultural context** and norms to convey intent.
- This type of content is among the **most challenging** to analyze.



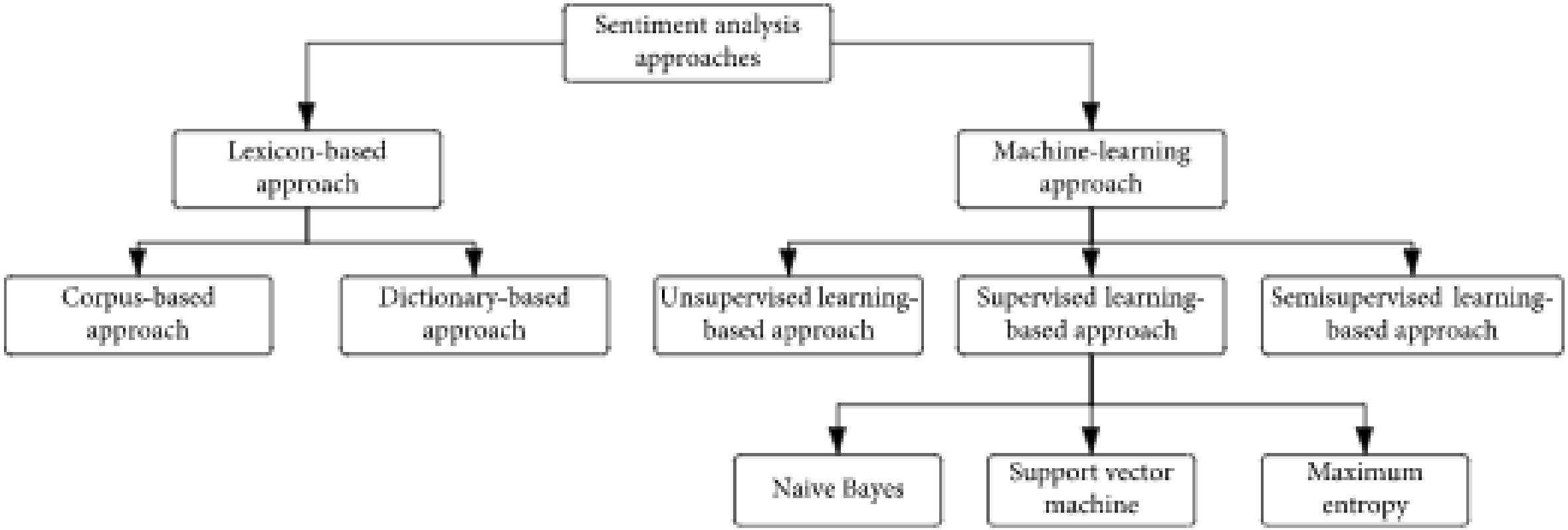
A COMPREHENSIVE STUDY ON SARCASM DETECTION TECHNIQUES IN SENTIMENT ANALYSIS

Researchgate.net

Approach

There are mainly two approaches for performing sentiment analysis.

1. **Lexicon-based**: count number of positive and negative words in given text and the larger count will be the sentiment of text.
2. **Machine learning** based approach: Develop a classification model, which is trained using the pre-labeled dataset of positive, negative, and neutral.


Sentiment analysis approaches
Lexicon-based approach
Machine-learning approach
Corpus-based approach
Dictionary-based approach
Unsupervised learning-based approach
Supervised learning-based approach
Semisupervised learning-based approach
Naive Bayes
Support vector machine
Maximum entropy

The different **lexicons**, the number of words in the lexicon, resolution, calculation method and classification method.

Lexicon	Number of words	Number of positive words	Number of negative words	Resolution	Calculation method to obtain score per year	Classification Method
AFINN	2,477	878	1,598	11	score individual words and sum	Manual
Bing	6,789	2,006	4,783	2	(number of positive words - number of negative words) / total words	Manual
Loughran	3,917	354	2,355	2	(number of positive words - number of negative words) / total words	Manual
NRC	5,555	2,312	3,324	2	(number of positive words - number of negative words) / total words	Amazon Mechanical Turk
SenticNet	23,626	11,774	11,852	continuous	score individual words and sum	Machine Learning
Sentiword	20,093	8,898	11,029	continuous	score individual words and sum	Machine Learning
Syuzhet	10,748	3,587	7,161	16	score individual words and sum	Manual
SOCAL	5,971	2,438	3,530	continuous	score individual words and sum	Amazon Mechanical Turk

Descriptions of Various Lexicons Used in Analysis

Source: https://bookdown.org/psonkin18/berkshire/sentiment.html

Using **VADER** to handle sentiment analysis with social media text

- Readily available pre-trained algorithms is called VADER *(**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner)* - a **lexicon** (dictionary of sentiments in this case) and a simple rule-based model for general sentiment analysis.
- a Python package
- Its algorithms are optimized ***to sentiments expressed*** *in social media like Twitter, online news*, movie/product reviews etc.
- It give a **Positivity** and **Negativity** score that can be standardized in a range of -1 to 1.

- **PHP Sentiment Analyzer** is a lexicon and rule-based sentiment analysis tool that is used to understand sentiments in a sentence using VADER.

4 Basic Components of Opinion

i. **Sentiment holder**: The holder of sentiment or opinion is a person or a group or an entity that expresses the sentiment.

ii. **Entity**: An object that can be a person, product, service, location, event, or text.

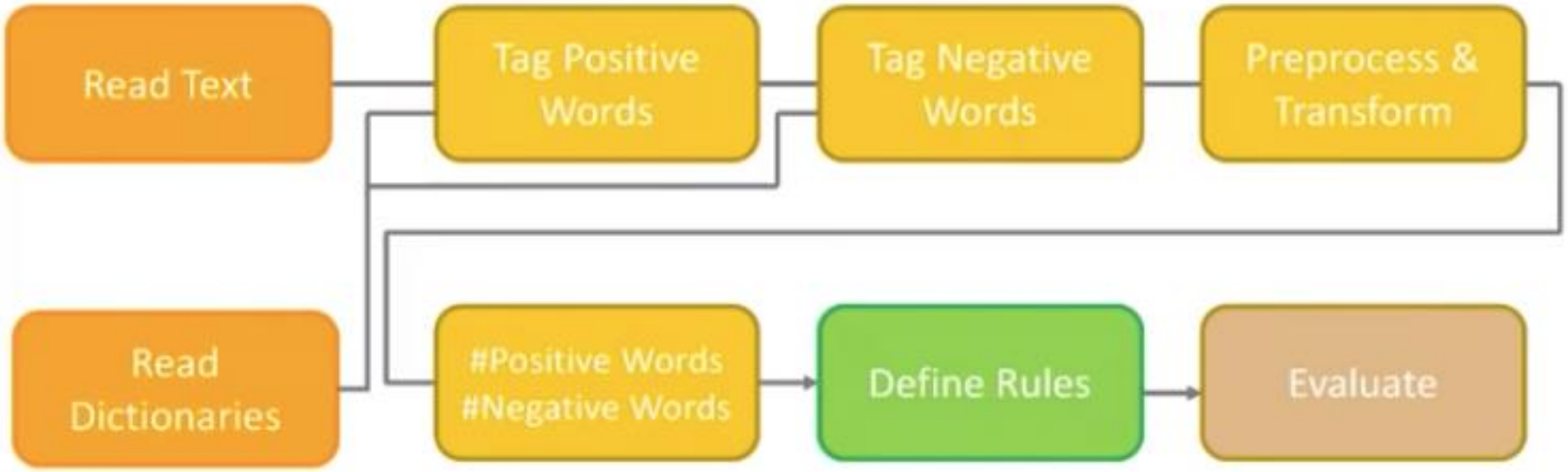
iii. **Feature**: A part (or an attribute) of the entity with respect to which evaluation is performed.

iv. **Sentiment orientation or polarity**: The orientation of an opinion on a feature represents whether the opinion is negative, positive or neutral.

Sentence	**"THE CAMERA PICTURE QUALITY IS WONDERFUL"**
Sentiment holder	Customer
Entity	Camera
Feature	Picture quality
Entity Opinion	Wonderful
Sentiment Orientation	Positive

Sentiment Analysis Workflow

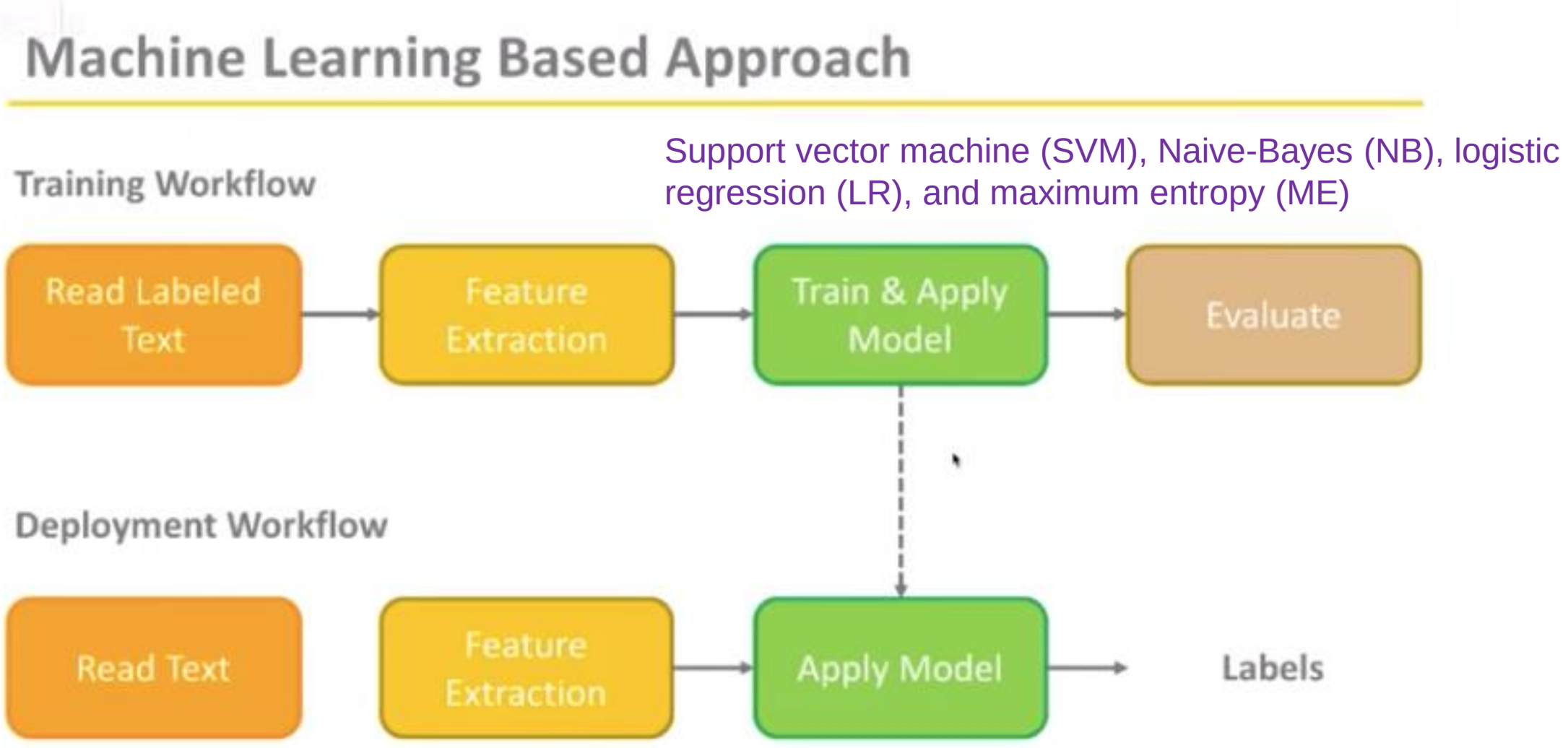
Lexicon Based / Rule Based Approach



sentiment score >0 => positive
sentiment score <= 0 => negative

$$, sentiment\ score = \frac{\#\ positive\ words - \#\ negative\ words}{\#total\ words}$$

Sentiment Analysis Workflow



Top Sentiment Analysis Software

Google Cloud Natural Language API, Lexalytics Salience, MeaningCloud, VisualText, Microsoft Azure Text Analytics API, IBM Watson Natural Language Understanding, Twinword, Rosette Text Analytics, NetOwl, Angoos KnowledgeREADER, Averbis, PrediCX, SAS Sentiment Analysis are some of the Top Sentiment Analysis Software.

Tutorial – Text Analytics in R

Text Analysis in R made easy with Udpipe

- https://towardsdatascience.com/easy-text-analysis-on-abc-news-headlines-b434e6e3b5b8

Learn how text analytics is done in R from the above link.

NO submission is required.

References

- Text Analytics in Action (Toronto Data Sciences Forum) by Cindy Zhong, Data Scientist, SAS Canada

https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/zhong_text_analytics.pdf

- What is Text Analytics? https://www.predictiveanalyticstoday.com/text-analytics/#content-anchor
- Why is analyzing text so hard? https://www.ibmbigdatahub.com/blog/why-analyzing-text-so-hard

https://m-clark.github.io/text-analysis-with-R/intro.html

https://tutorials.datasciencedojo.com/text-analytics/

https://monkeylearn.com/sentiment-analysis/