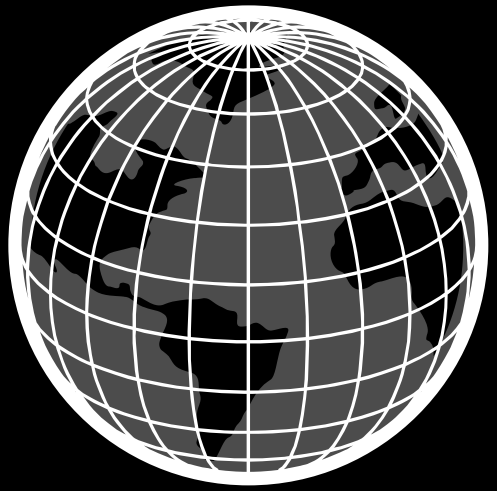
Natural Language Processing

In Big Data

How Much Data Do We Create Everyday



2.5 quintillion bytes

Amounts of data in…

Internet

3.7 billion humans use the internet

Twitter

456,000 tweets

Linkedin

More than 120 professionals

Communication

16 million text messages

Skype

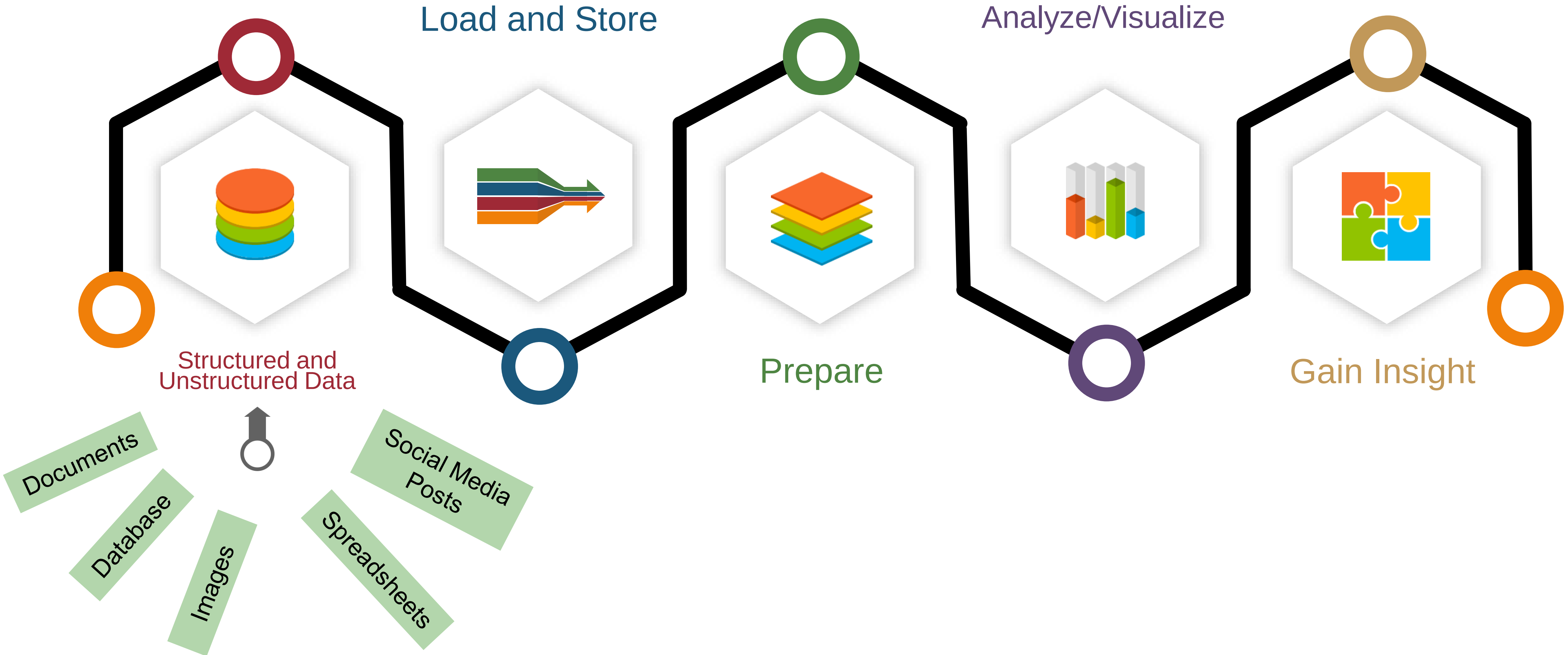
154,200 calls

Instagram

46,740 photos

Sources: Bernard Marr, **"How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read"**, May 21, 2018
Domo's Data Never Sleeps 5.0 report

Simple Big Data Journey



Load and Store

Analyze/Visualize

Structured and Unstructured Data

Prepare

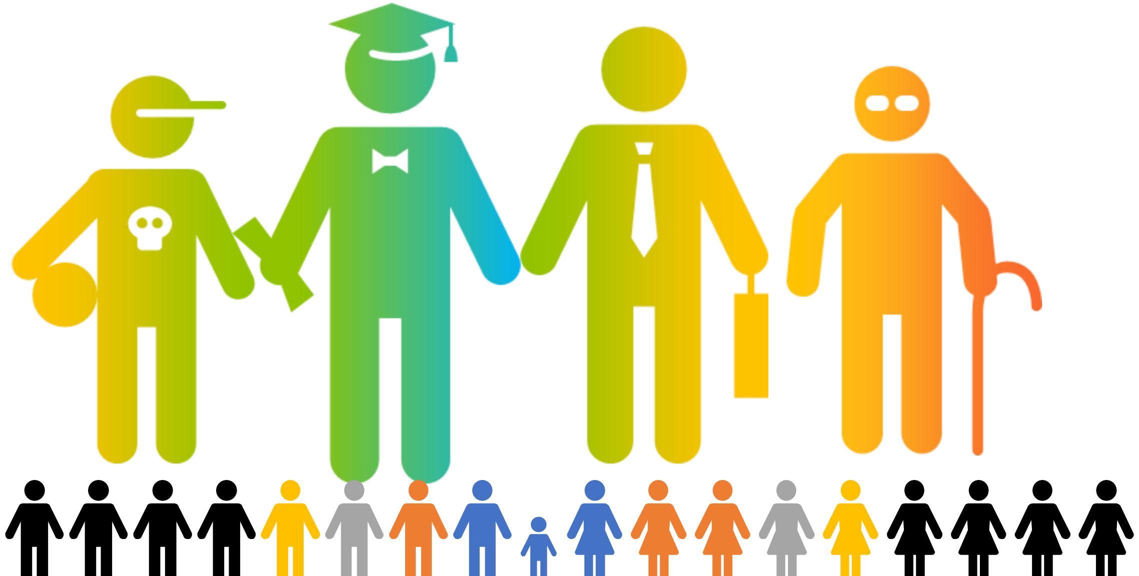
Gain Insight

What is Natural Language Processing?

Social Media Posts

Positive?

Negative?

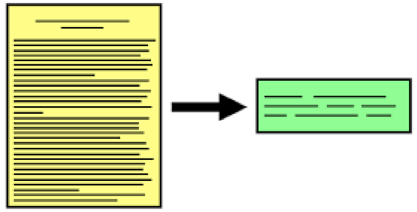


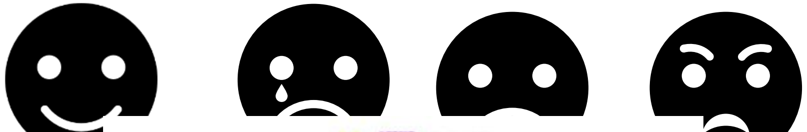
Natural Language Processing or NLP is the field of study that focuses on the interactions between human language and computers. It sits at the intersection of computer science, artificial intelligence, and computational linguistics. Wikipedia

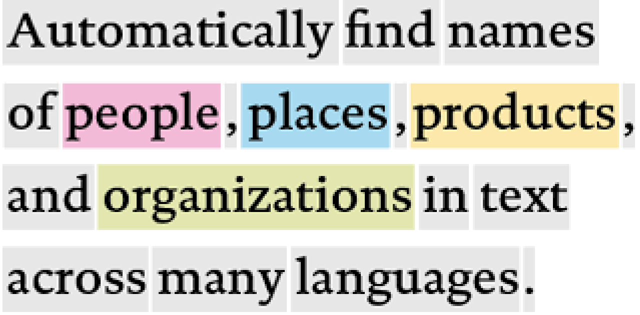
How is it being used?

NLP is used to analyse text, allowing machines to understand how human’s speak. It enables

- Automatic text summarization
- **Sentiment analysis**
- **Topic extraction**
- **Named entity recognition**







Automatically find names of people, places, products, and organizations in text across many languages.

How does it work?

Latent Dirichlet Allocation or LDA

-a generative probabilistic model of a collection of composites made up of parts. In terms of topic modelling, the composites are documents and the parts are words and/or phrases.

Purpose:

Learn the representation of a fixed number of topics, and given this number of topics, learn the topic distribution that each document in a collection of documents has.

Example

Sentence A: I spent a day at the beach tanning.
Sentence B: I ate sea foods and lechon.
Sentence C: I love tanning in beaches while eating sea foods.

LDA might say something like:

Sentence A is 100% about Topic 1
Sentence B is 100% Topic 2
Sentence C is 50% Topic 1, 50% Topic 2

Where LDA also discovers that:

Topic 1 represents things related to the beach
Topic 2 represents things related to food

How does LDA work?

An LDA model is defined by two parameters:

- **α**—A prior estimate on topic probability
- **β**—a collection of k topics where each topic is given a probability distribution over the vocabulary used in a document corpus, also called a "topic-word distribution."

LDA is a **"bag-of-words"** model

LDA is a generative model where each document is generated word-by-word by choosing a topic mixture θ ~ Dirichlet(α).

For each word in the document:

- Pick a topic z ~ Multinomial(θ)
- Pick the corresponding topic-word distribution β_z.
- Draw a word w ~ Multinomial(β_z).

Training the model:

The goal is to find parameters α and β which maximize the probability that the text corpus is generated by the model.

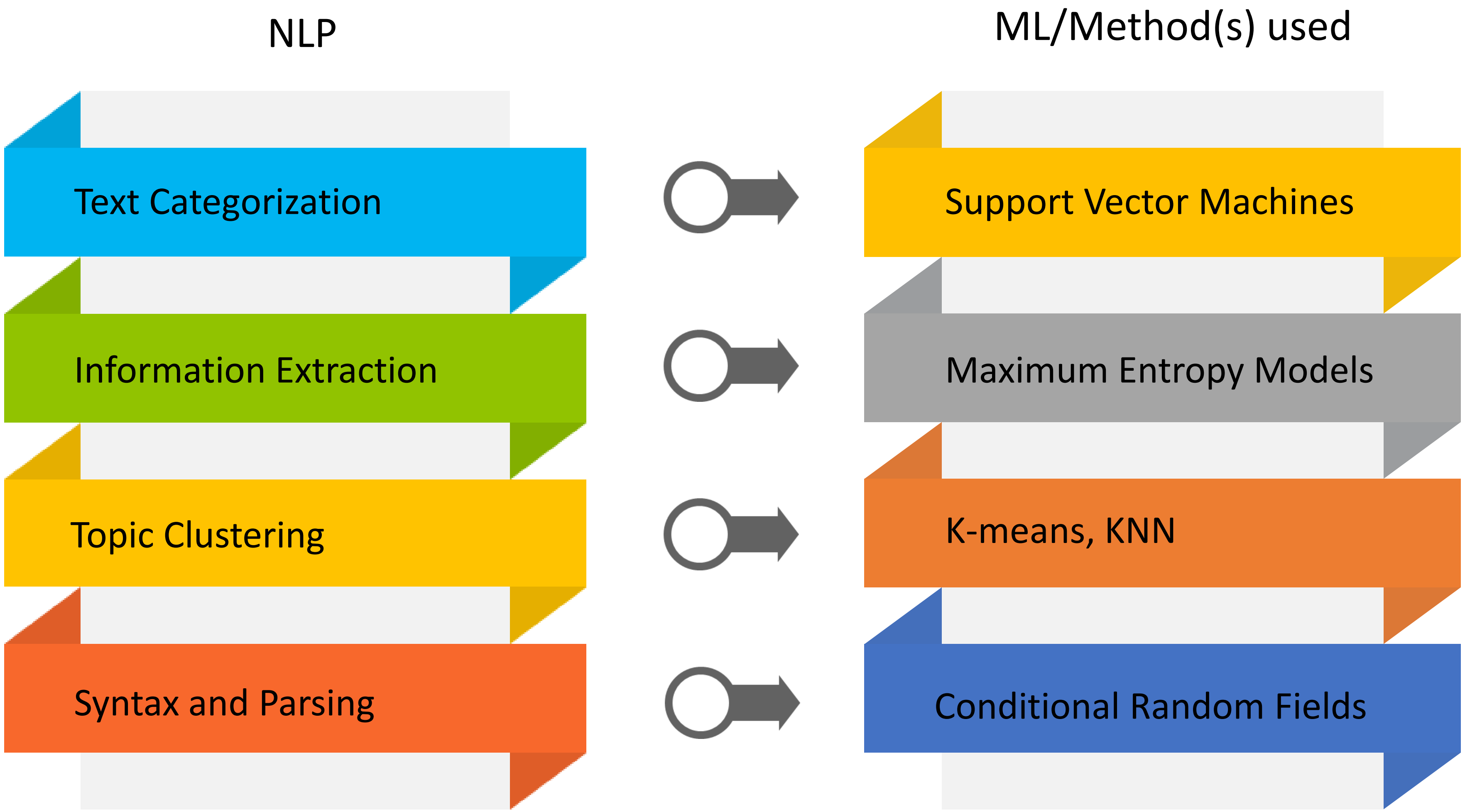
Methods for estimating the LDA model

Gibbs sampling

Expectation Maximization (EM)

Source: David M Blei, Andrew Y Ng, and Michael I Jordan. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.

Other Methods



NLP	ML/Method(s) used
Text Categorization	Support Vector Machines
Information Extraction	Maximum Entropy Models
Topic Clustering	K-means, KNN
Syntax and Parsing	Conditional Random Fields

Business Applications

"How can I keep my customers happy?"

"What are people saying about me?"

"What's happening with the competitors?"

"Who is interested with my product?"

"Is this applicant fit to the job opening?"

NLP for Filipino Language

English-Filipino machine translation system

e-Wika: Digitalization of Philippine Language
C. K. Cheng, R. E. O. Roxas, A. B. Borra, N. R. L. Lim, E. C. Ong and S. L. See College of Computer Studies, De La Salle University, Manila 2401 Taft Ave., Malate, Manila 1004, Philippines

Dito

Jusko

lit

D2

Juiceko

finsta

Goat

Thank you!

Ruby Jean L. Ocenar
Business Data Analyst
goFLuent,Inc
Personal e-mail : ruby_ocenar@yahoo.com
Work e-mail : rjocenar@gofluent.com
Mobile number : 09175616577