
# COMP30810
# Intro to Text Analytics

Dr. Binh Thanh Le

thanhbinh.le@ucd.ie

Insight Centre for Data Analytics

School of Computer Science

University College Dublin

# Today Goals:

- Where do Data come from? Big Data?
- Related concepts in Data Science?
- CRIPS-DM?
- Text analytics?

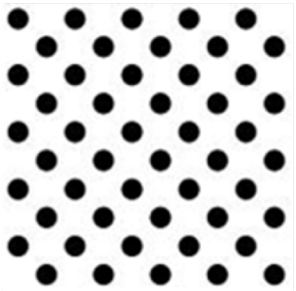
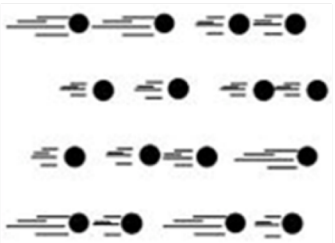
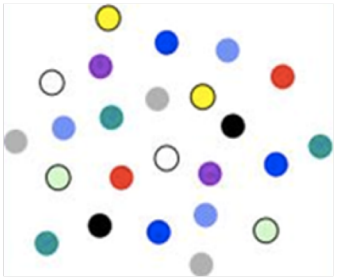
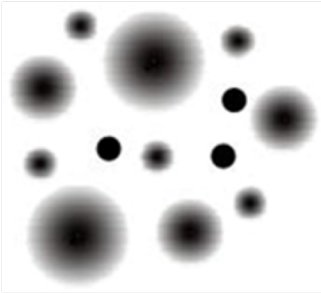
# WHERE DO DATA COME FROM?

- LinkedIn
- Instagram
- Snap chat
- Tweet
- Google engine
- Facebook
- YouTube
- WhatsApp
- Etc.


# What is Big Data?

# What is Big Data?

**Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. (from sas)**

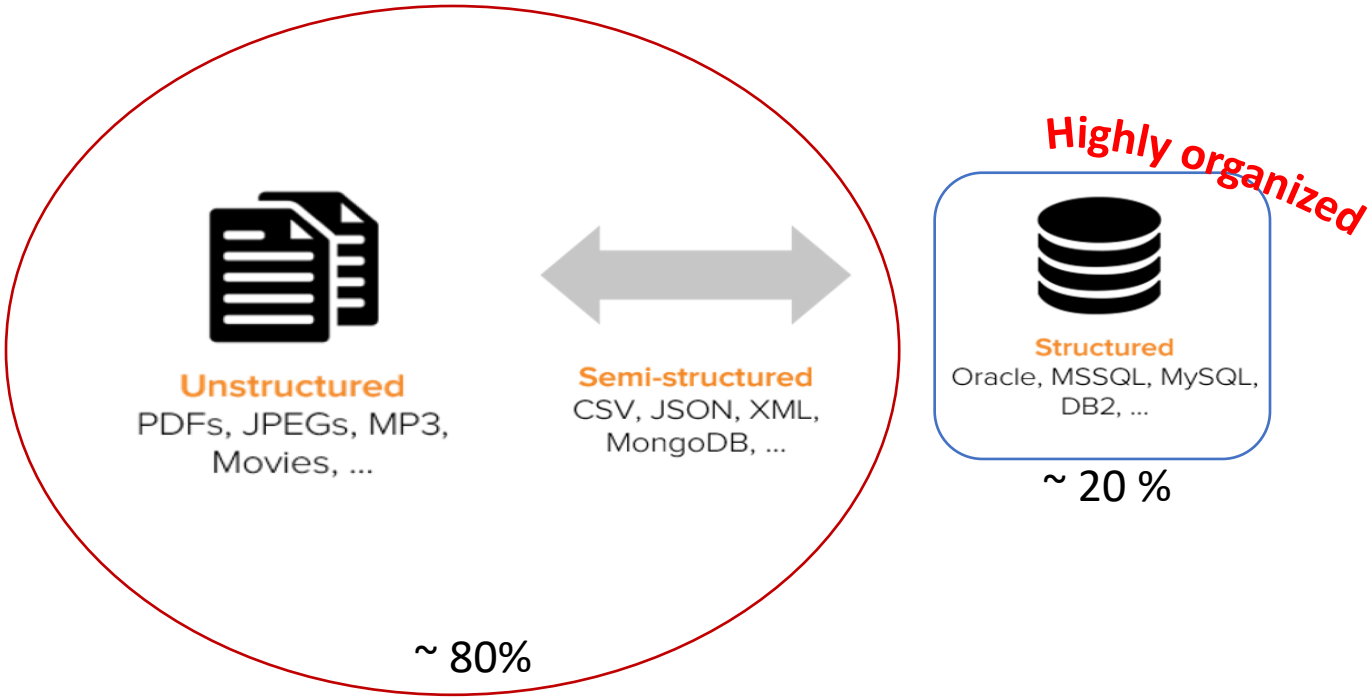
| Volume | Velocity | Variety | Veracity | Value |
|---|---|---|---|---|
|  |  |  |  |  |
| **Data at Rest** | **Data in Motion** | **Data in Many Forms** | **Data in Doubt** | **Data into Money** |
| Terabytes to Exabytes of existing data to process | Streaming data, requiring milliseconds to seconds to respond | Structured, unstructured, text, multimedia,… | Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations | Business models can be associated to the data |

Adapted by a post of Michael Walker on 28 November 2012

*http://informationcatalyst.com/vision-experience/big-data-value/*

# Types of Data used in Analytics

- Data are organized in three categories: Structured, Semi-structured, and Unstructured Data.



# Types of Data used in Analytics

- However, the data types can be transferred

Example:


Text file

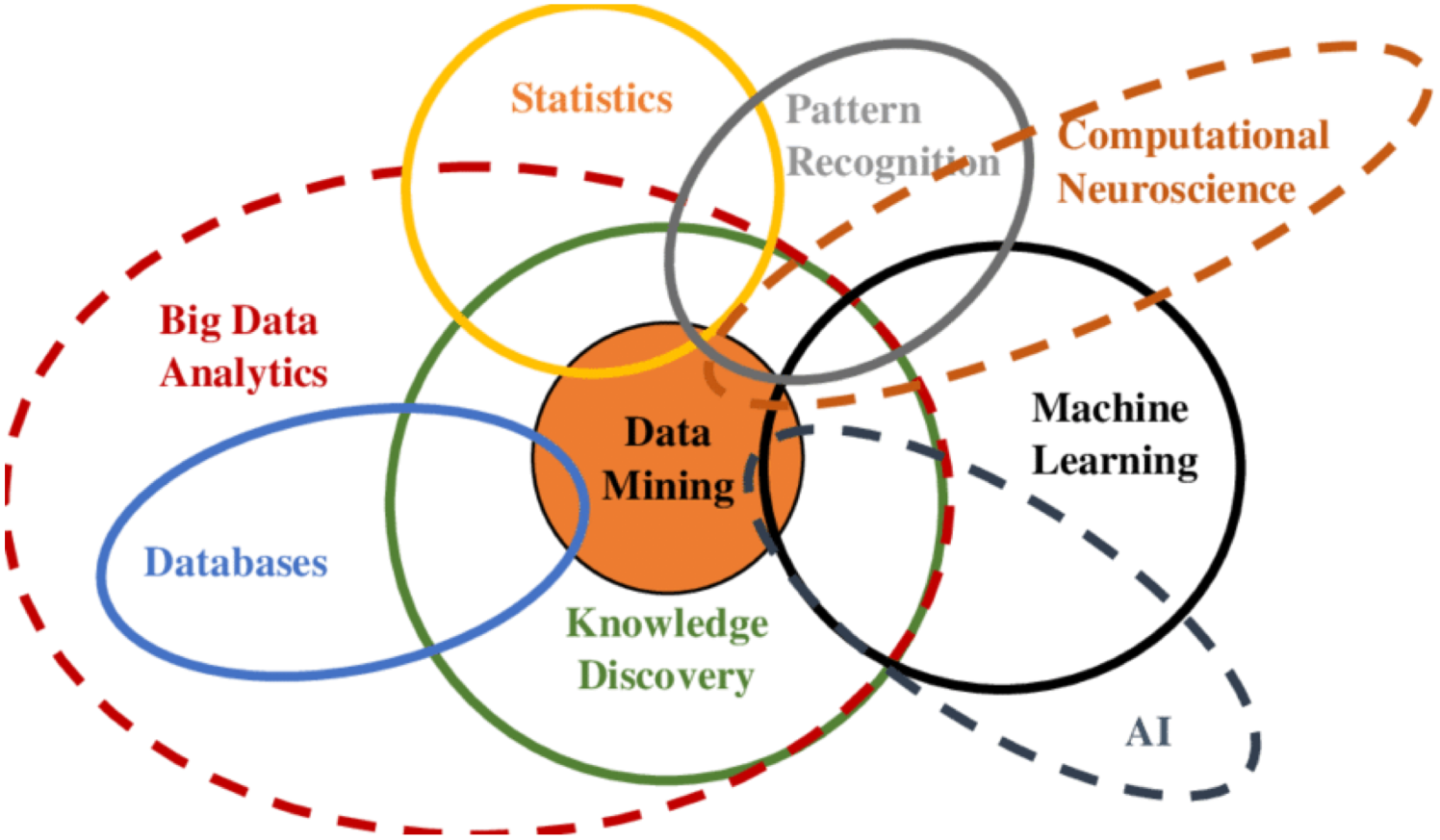
<Tag>

```
<?xml version="1.0" encoding="UTF-8" ?>
- <customer_order number="004985" date="2004-06-24">
  - <lines>
    - <line no="1">
        <item>Disc CD</item>
        <quantity>30</quantity>
        <price>0.95</price>
      </line>
    - <line no="2">
        <item>Disc CD-RW</item>
        <quantity>20</quantity>
        <price>2.95</price>
      </line>
    </lines>
  - <customer>
      <name>Technical University of Lublin</name>
      <street>Nadbystrzycka 38</street>
      <city>Lublin</city>
      <post_code>20-501</post_code>
    </customer>
  - <payment>
      <card_issuer>Master Card</card_issuer>
      <card_number>1234 567890 12345</card_number>
      <expiration_date month="10" year="2005" />
    </payment>
</customer_order>
```

organize

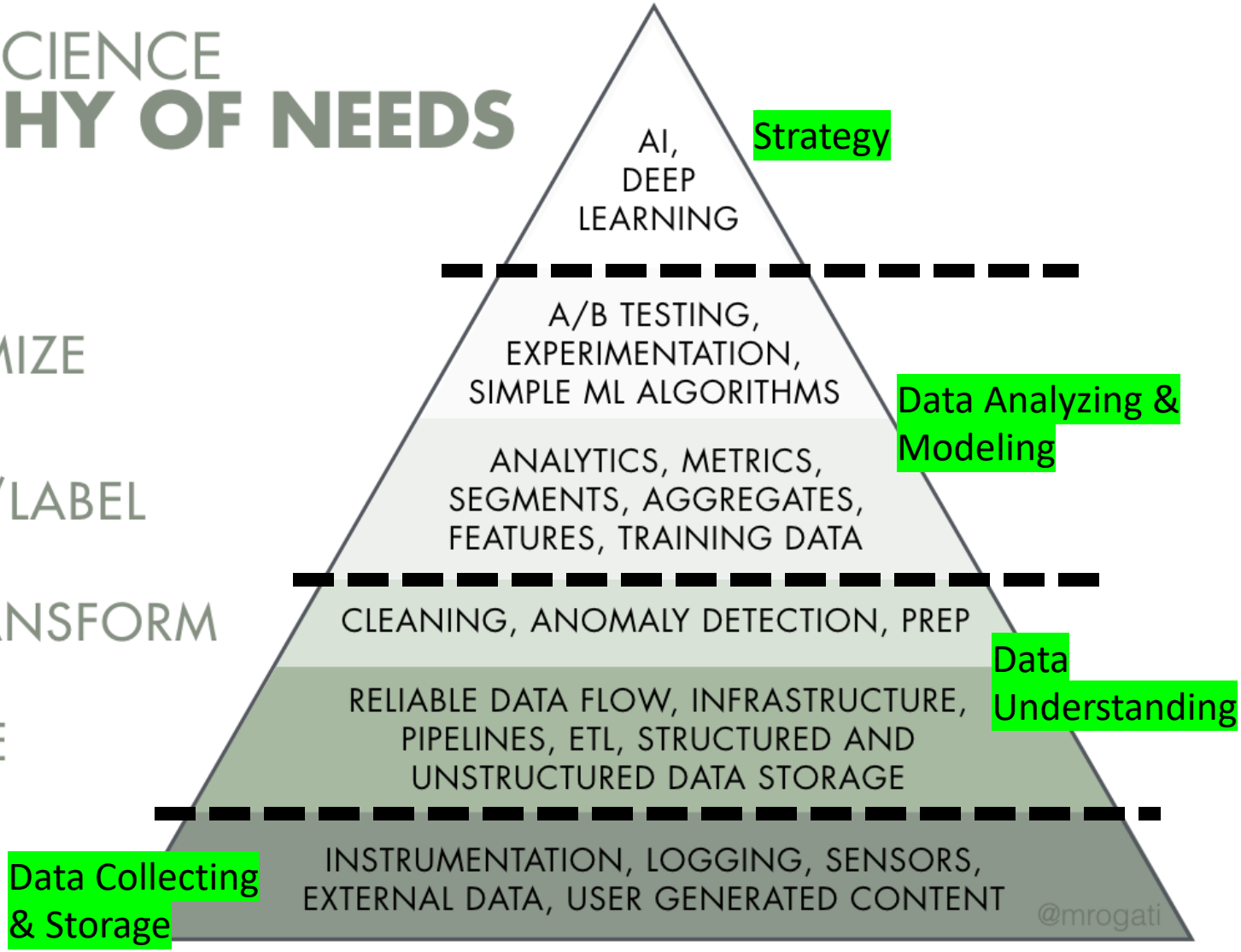
Relational Database

# Multidisciplinary nature of Big Data analytics




*[*] Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., ... & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced engineering informatics, 30(3), 500-521.*

# THE DATA SCIENCE HIERARCHY OF NEEDS




LEARN/OPTIMIZE

AGGREGATE/LABEL

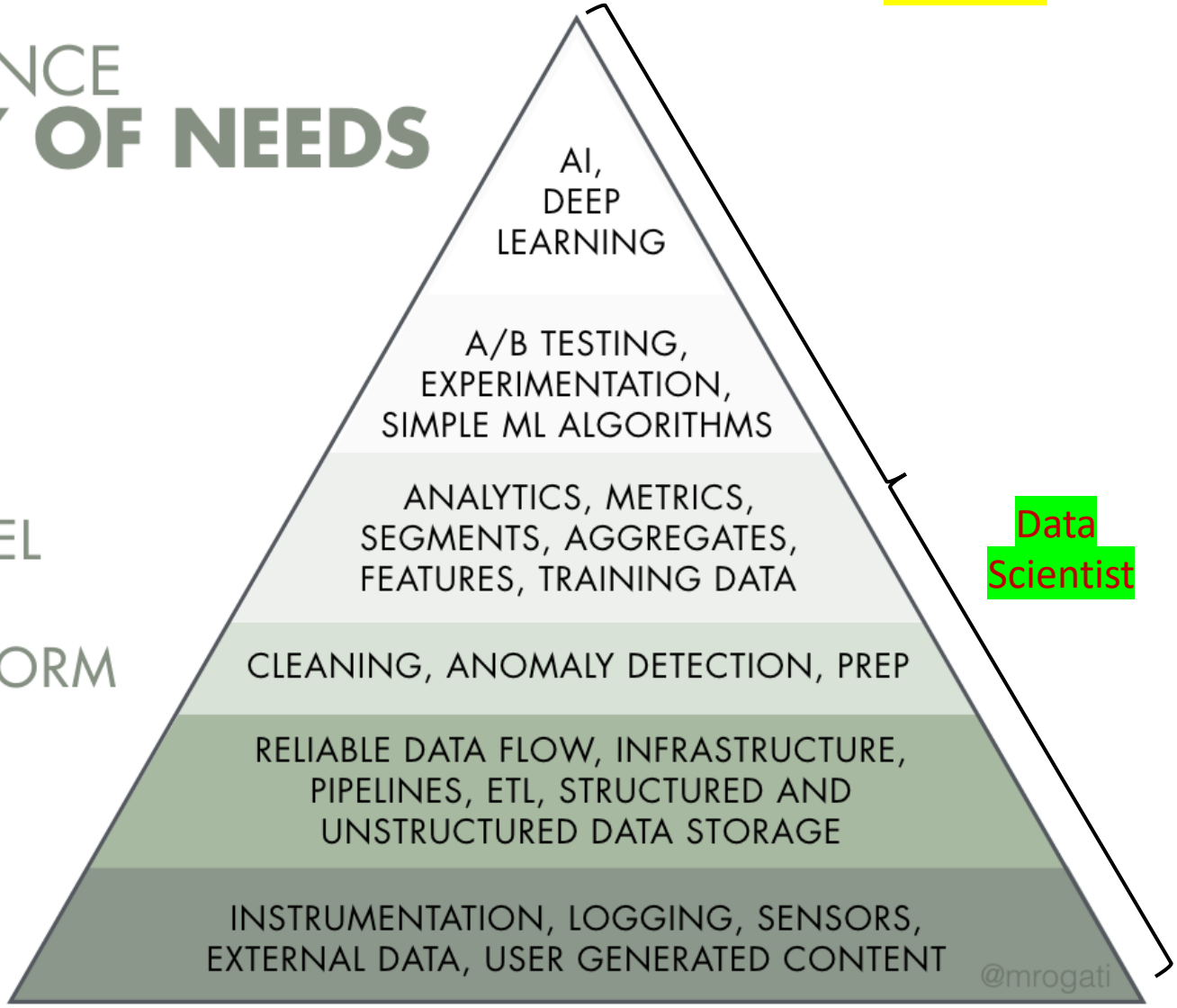
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

SMALL COMPANY
- STARTUP -

# THE DATA SCIENCE HIERARCHY OF NEEDS




LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

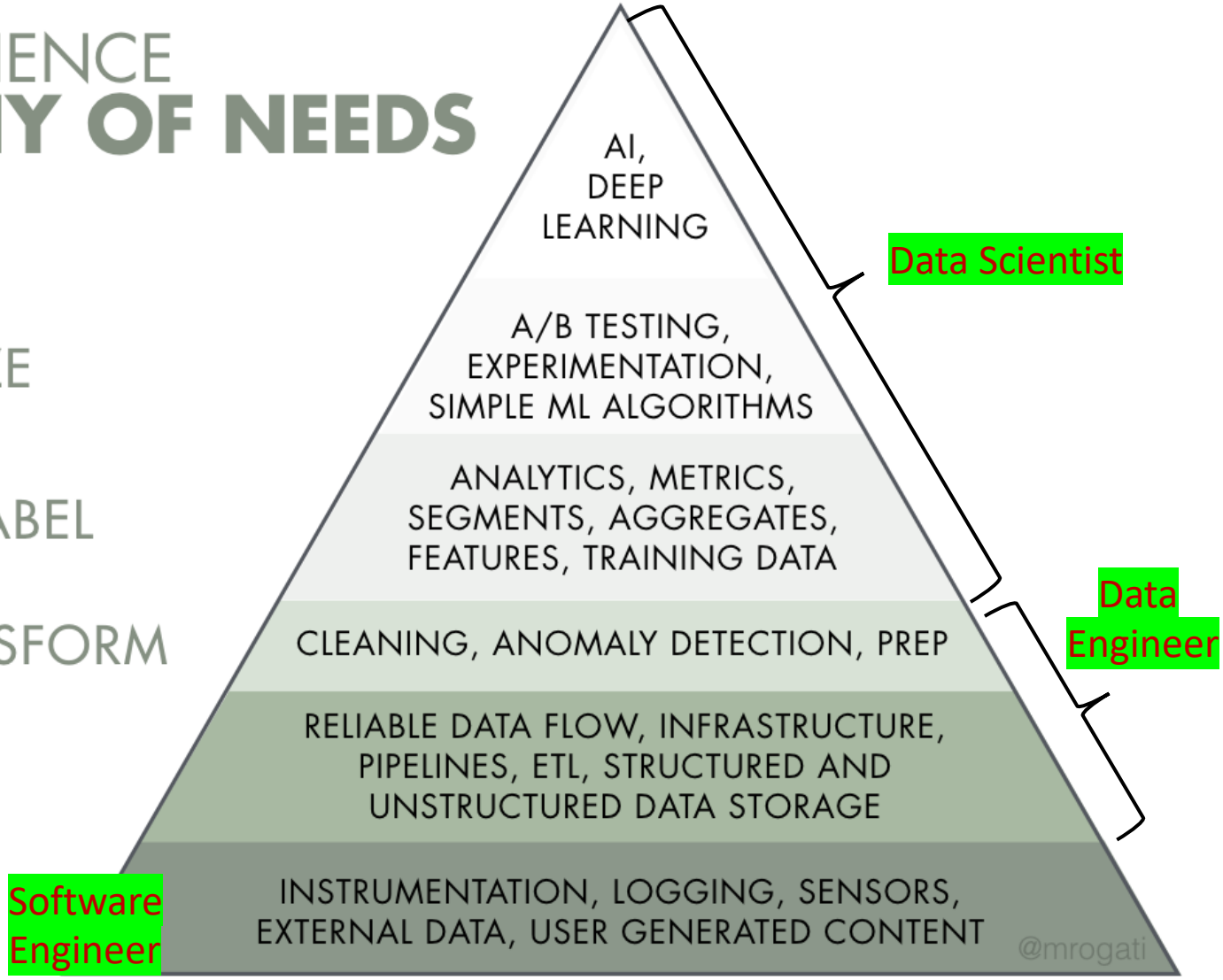
MOVE/STORE

COLLECT

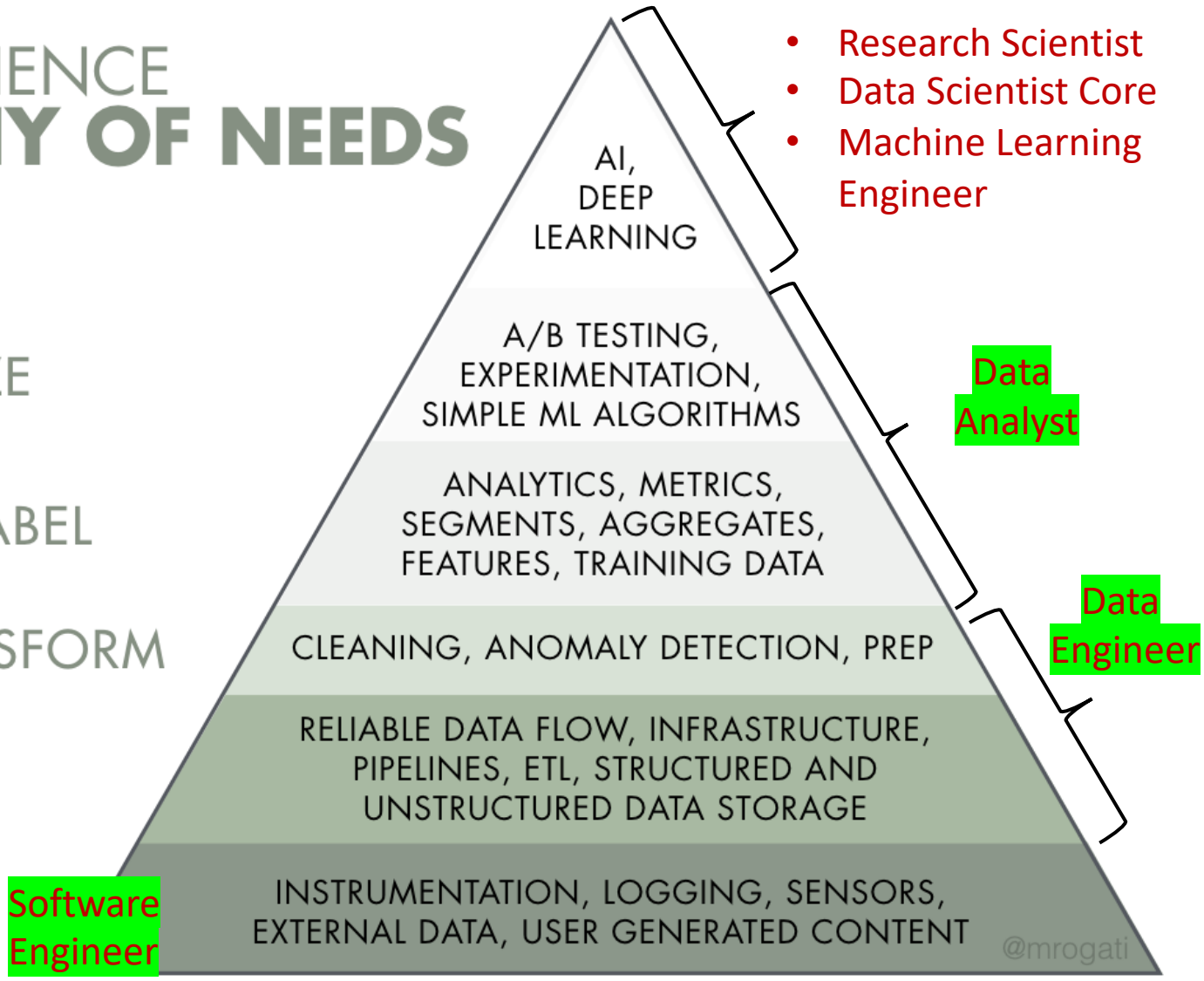
Data Scientist

MEDIUM COMPANY

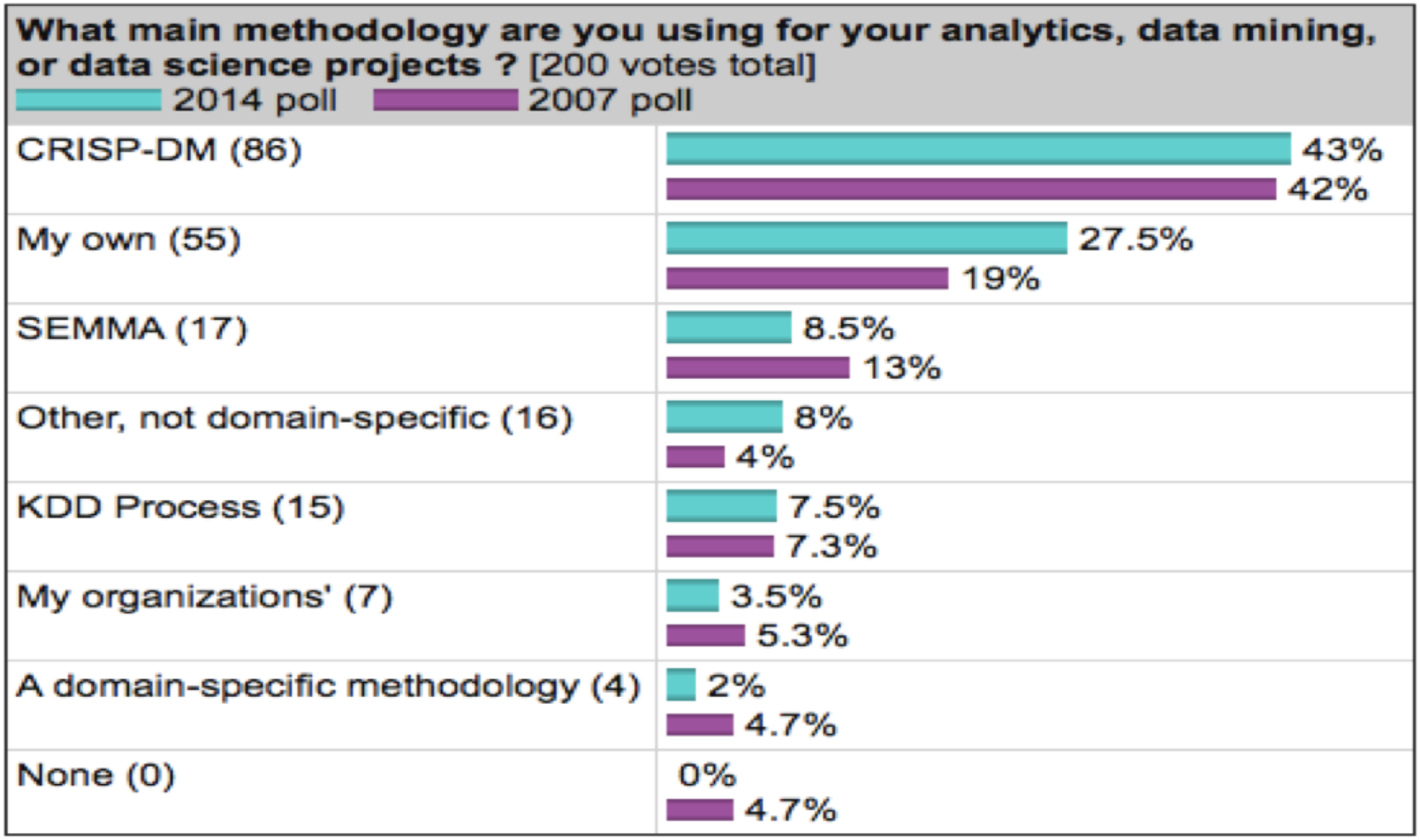
# THE DATA SCIENCE HIERARCHY OF NEEDS



# THE DATA SCIENCE HIERARCHY OF NEEDS



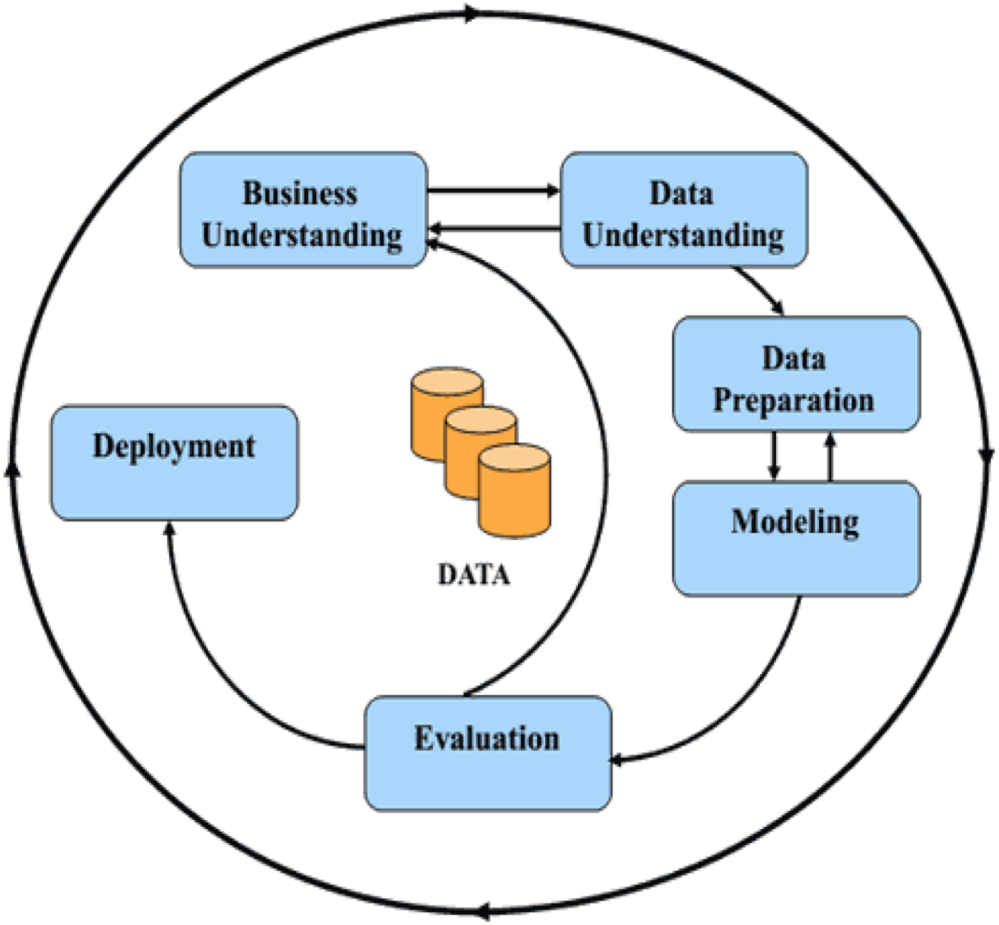
# Methodology for Data Analytics




*http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html*

# What is CRIPS-DM?

CRISP-DM: **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (1996)




**1. Business Understanding:** What is the business **problem**?

**2. Data Understanding:** What is the **data** required to solve the business problem?

**3. Data Preparation:** Where is the **data**, how should it be collected, transformed and stored?

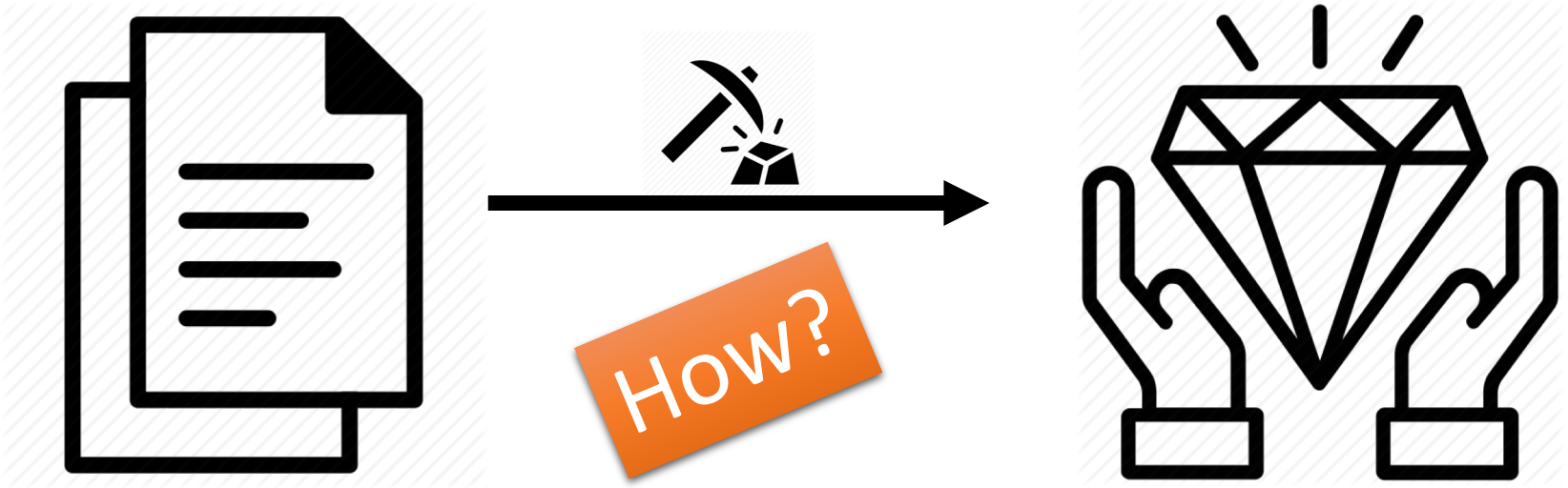
**4. Modeling:** What data analytics **algorithms** should be used?

**5. Evaluation:** How well do the **algorithms** work?

**6. Deployment:** How can the analytics results be integrated into the work process (specific to the organization)?

# What is Text Analytics?

**Text mining**, also referred to as **text data mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from **text**.




Text

high-quality information

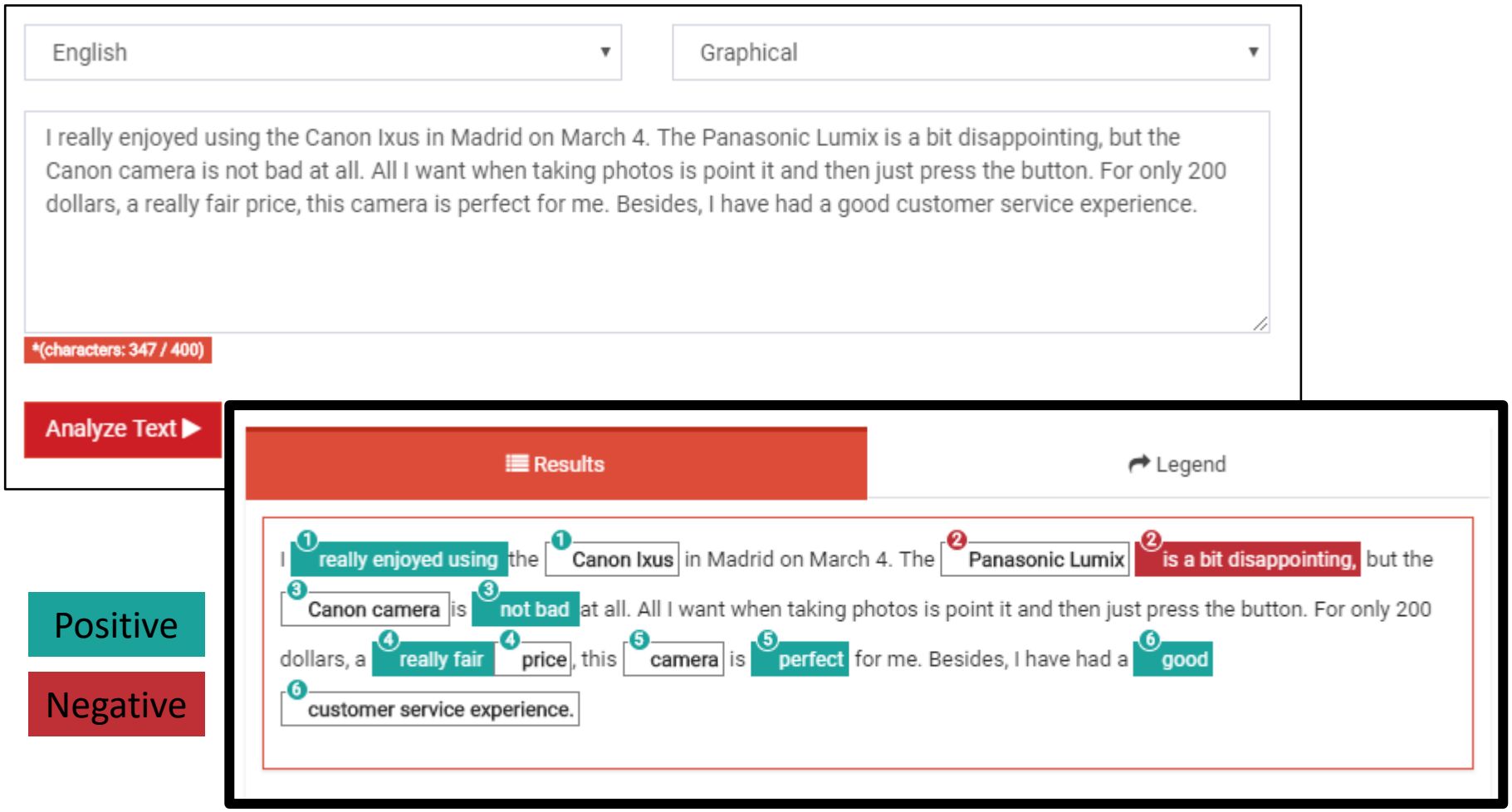
# What is Text Analytics?

**Text mining**, also referred to as **text** data **mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from **text**.

Example

# Text mining around us

- Sentiment analysis



# Text mining around us

- Document summarization

exampletext.txt

The bar chart compares the amount of carbon emissions in various countries from 1975 to 2005.
At the first glance, the biggest emitter of carbon country was the USA, and followed by the China. In 2005, there were the significant rise in China's carbon emission, leading to the similar value to the USA's.
From 1975 to 2005, the USA and China were the two largest emitter of carbon countries with the correspondingly values from 1,200,000 to 1,600,000 thousand metric tonnes (in the USA), and from 300,000 to nearly 1,600,000 thousand metric tonnes (in the China). On the other hand, the United Kingdom and the Canada had the two lowest measures of carbon releasing with around 180,000 thousand metric tonnes, and around 150,000 thousand metric tonnes respectively. (20) The average of carbon emissions by the Germany and the India during 30 years were mostly equivalent with around 210,000 thousand metric tonnes. (22)
From 1990 to 2005, the China had a critical increasing in carbon emissions measurement when it was raised from 650,000 to 1,500,000 thousand metric tonnes. By contrast, the Germany was the only country which had a distinct decreasing of carbon emissions from 250,000 to 200,000 thousand metric tonnes during 30 years. (27)

Ok. Let's summarize your attachment!

Here is your summary! Please, see full summary at the link provided below.

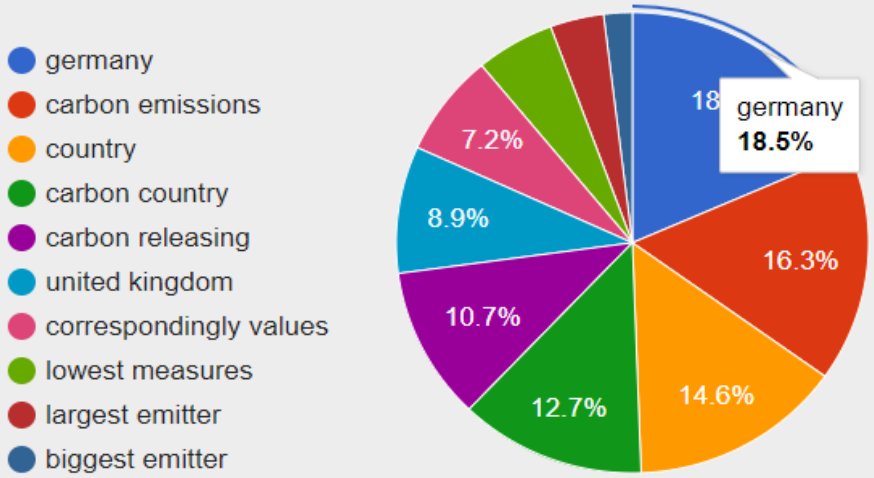
AI-Powered Summary

**Get to know more by reading less!**
View or Save your summary. Give us feedback!
www.summarizebot.com

View in Browser

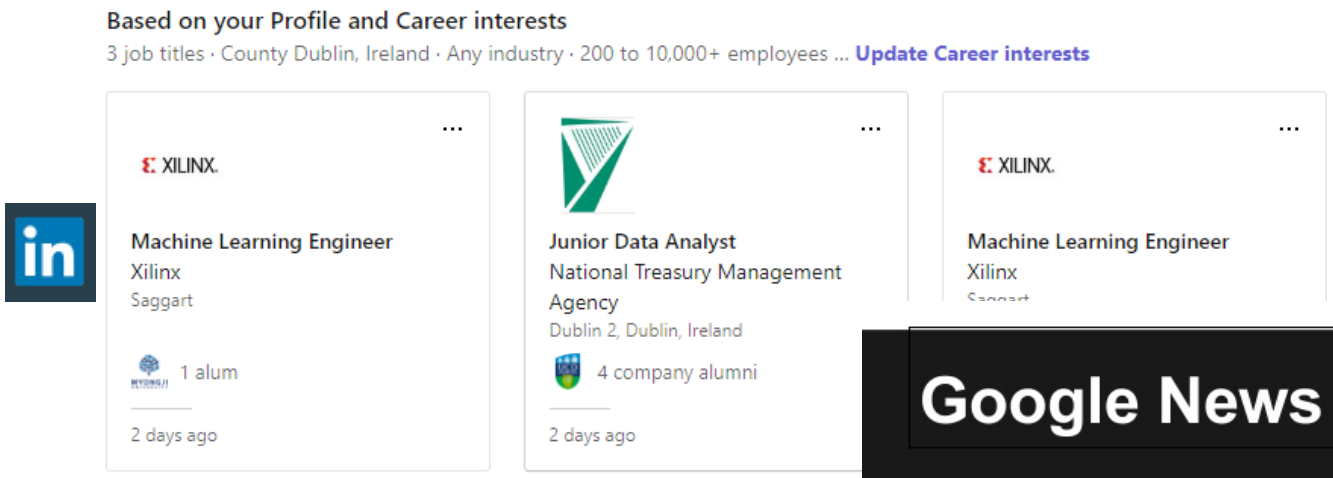
Learn More

**Informative Keywords**

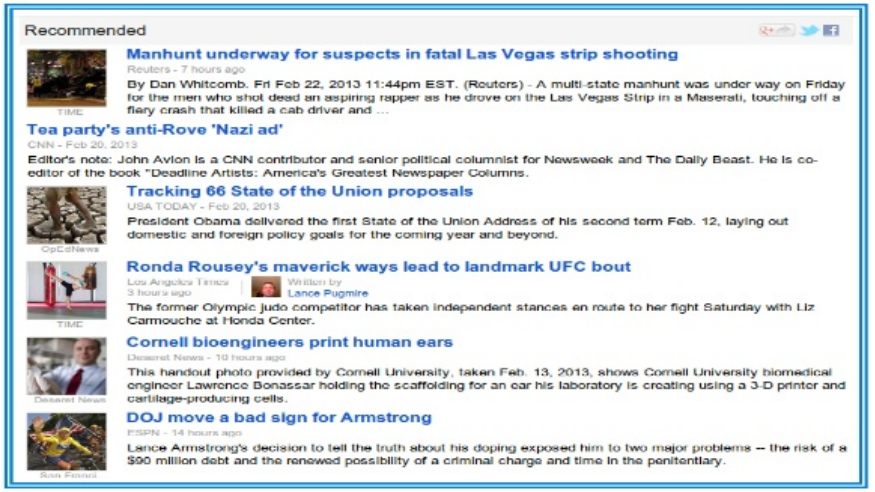


# Text mining around us

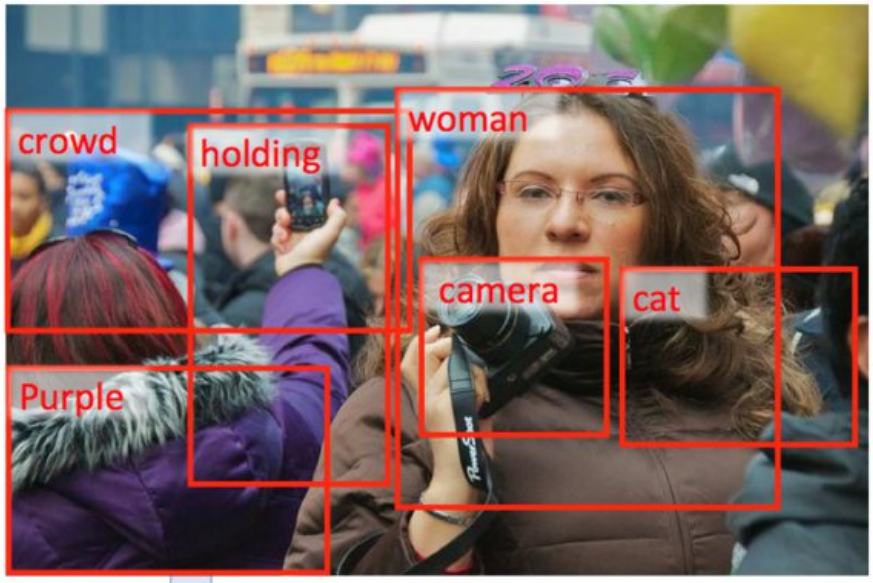
- Books, movies, news, hotel, career … recommendation







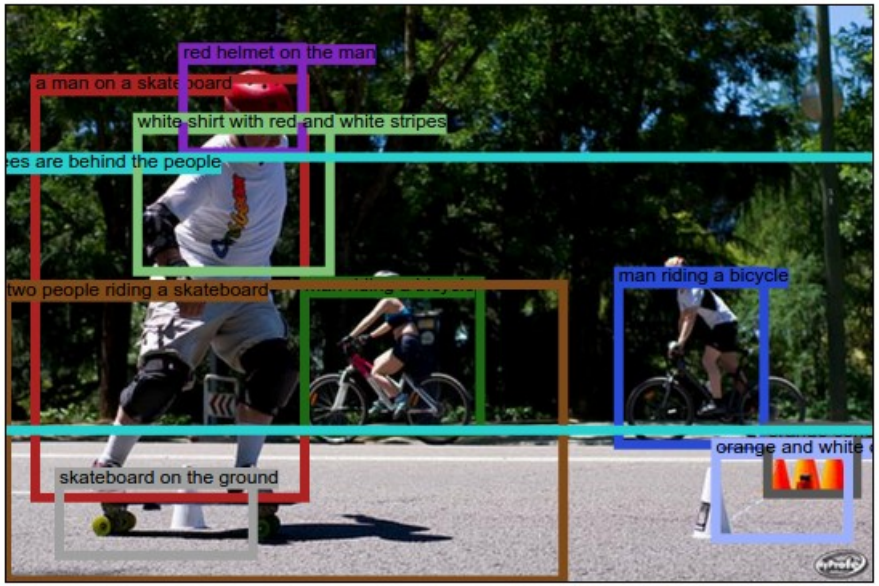
# Interdisciplinary Tasks: Image Captioning




1. detect words → woman, crowd, cat, camera, holding, purple

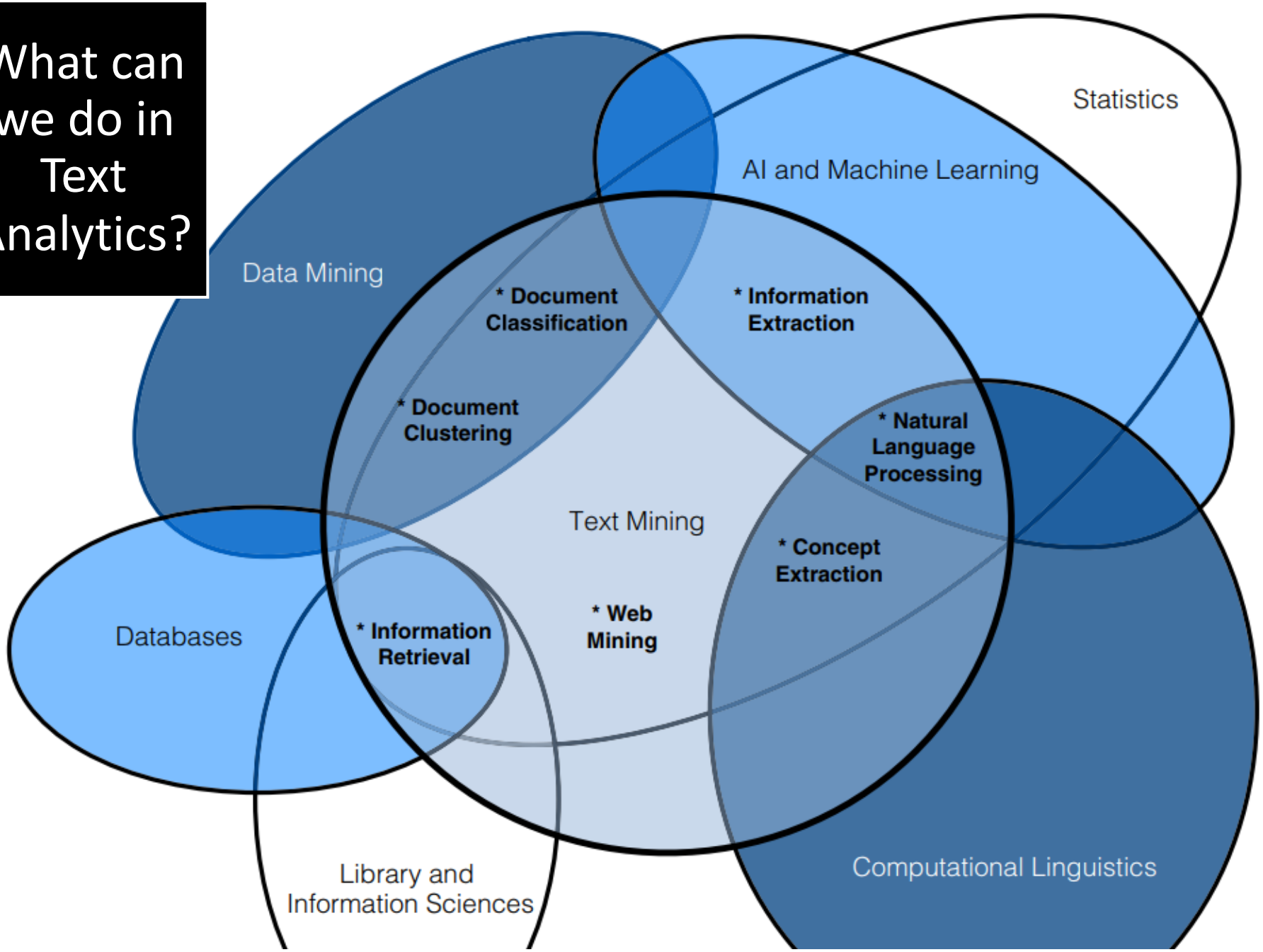
2. generate sentences → A purple camera with a woman. A woman holding a camera in a crowd. ... A woman holding a cat.

3. re-rank sentences → #1 A woman holding a camera in a crowd.




a man on a skateboard. man riding a bicycle. orange cone on the ground. man riding a bicycle. two people riding a skateboard. red helmet on the man. skateboard on the ground. white shirt with red and white stripes. orange and white cone. trees are behind the people.

# What can we do in Text Analytics?




*http://www.textanalyticsworld.com/wp-content/uploads/2012/03/PracticalTextMining_Excerpt.pdf*

# What can we do in Text Analytics?

1. **Search and information retrieval (IR)**: Storage and retrieval of text documents, including search engines and keyword search.
2. **Document clustering**: Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
3. **Document classification**: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
4. **Web mining**: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.
5. **Information extraction (IE)**: Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text.
6. **Natural language processing (NLP)**: Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics.
7. **Concept extraction**: Grouping of words and phrases into semantically similar groups.

***http://www.textanalyticsworld.com/wp-content/uploads/2012/03/PracticalTextMining_Excerpt.pdf***

# What makes Text Analytics so Hard?

Ambiguity


*http://puterydelima.blogspot.com/2016/03/ambiguity.html*

# What makes Text Analytics so Hard?

Non-Standard Language


LOL Lee O. Langdon

LOL laugh(ing) out loud

# What makes Text Analytics so Hard?

More Complex Languages Than English

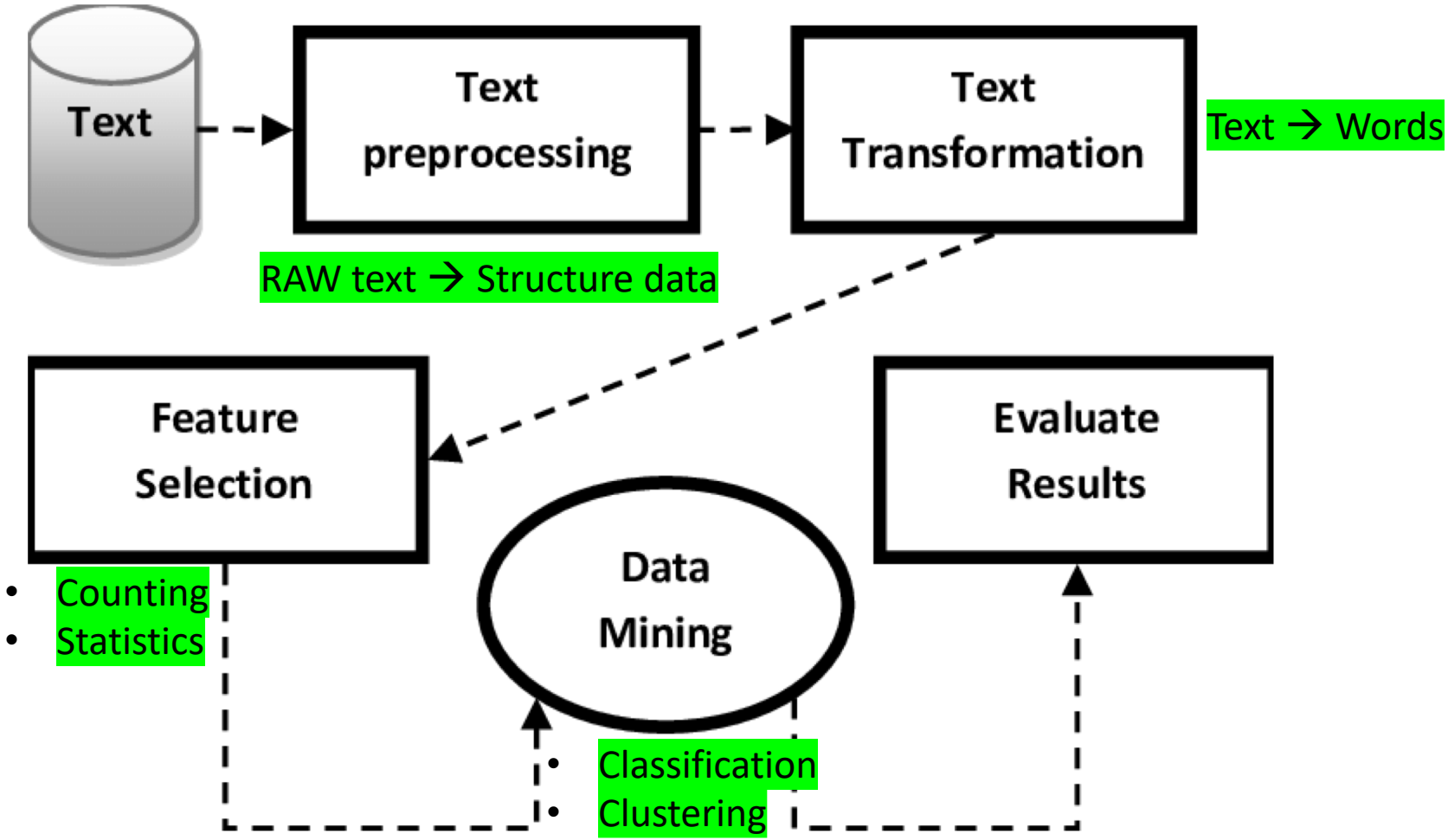

German: Donaudampfschiffahrtsgesellschaftskapitän (5 “words”)

Chinese: 50,000 different characters

Japanese: 3 writing systems

… … …

# Text Mining | Analytics Process



# Summary

- Understanding about
  - Big Data, Data Analytics and concepts
  - CRISP-DM
  - Data Science and branches
  - Text Analytics and roles in Text Mining
  - Text Analytics Process