



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:

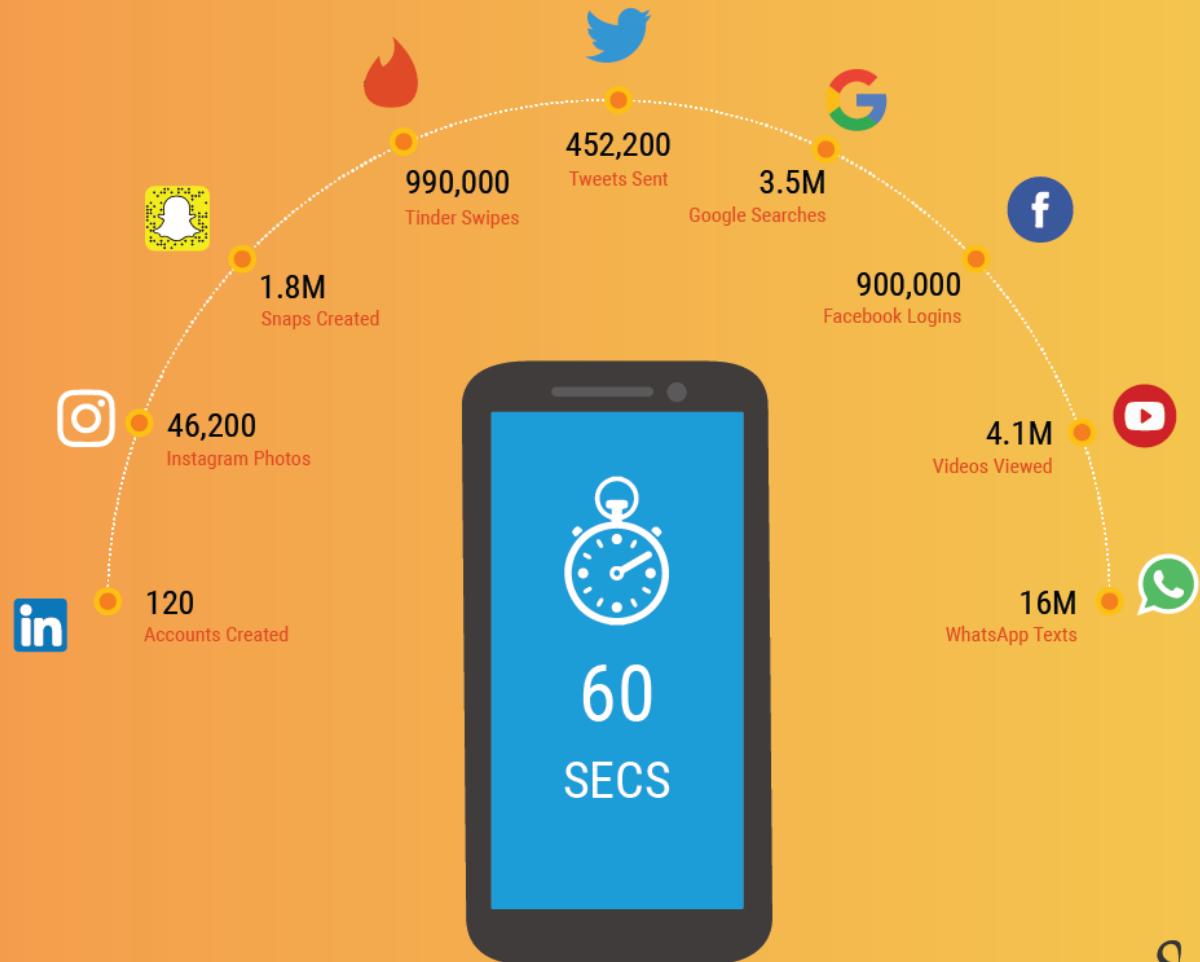
- Where do Data come from? Big Data?
- Related concepts in Data Science?
- CRIPS-DM?
- Text analytics?

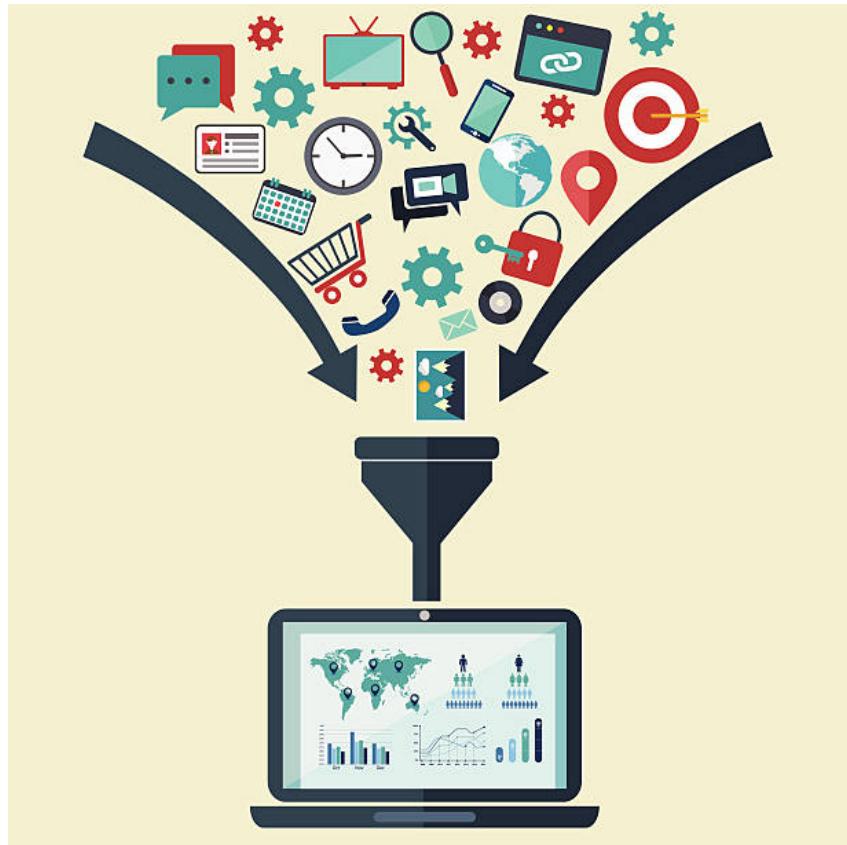
WHERE DO DATA COME FROM?

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

2017

What Happens in one internet minute?

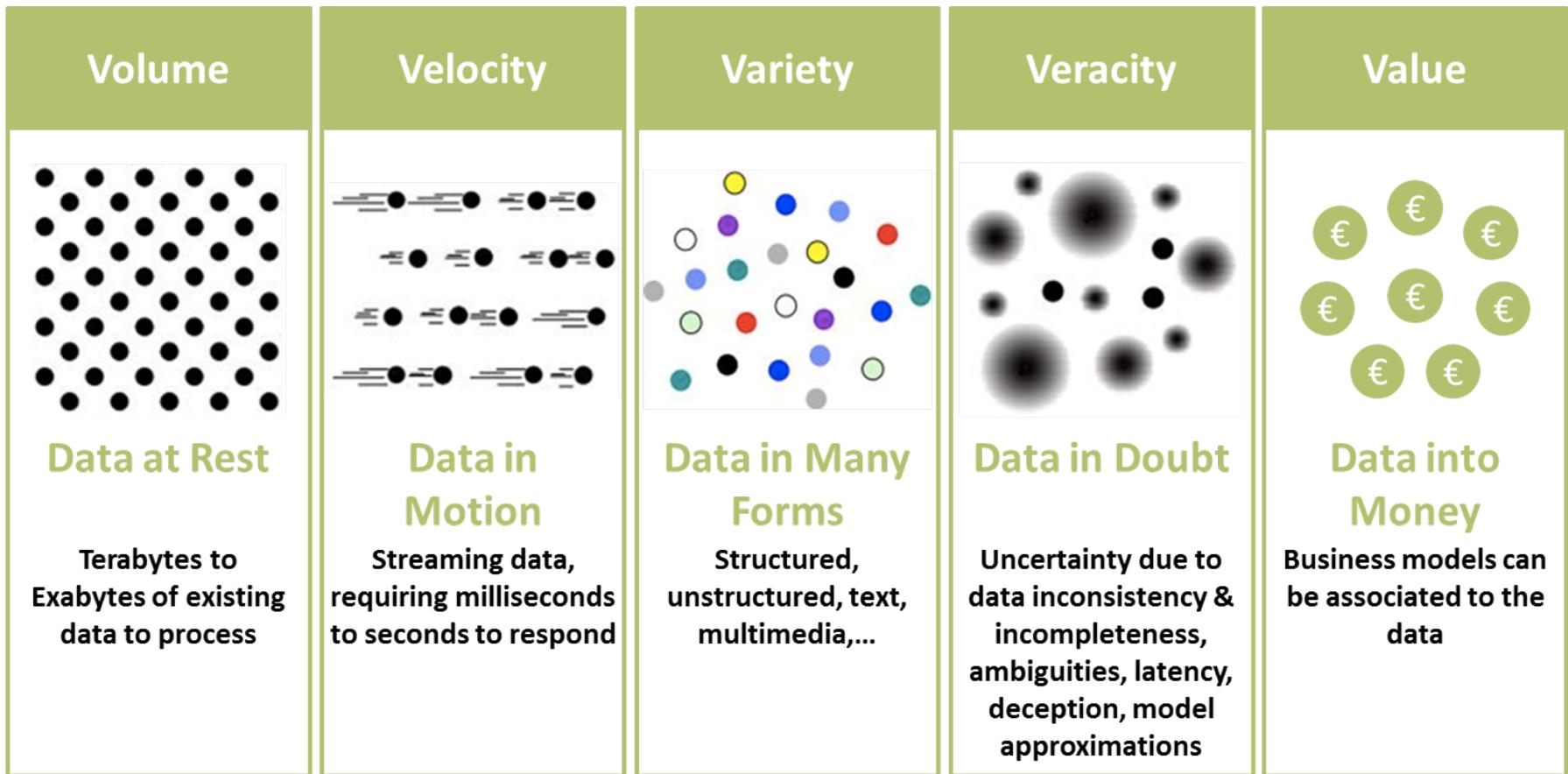




What is Big Data?

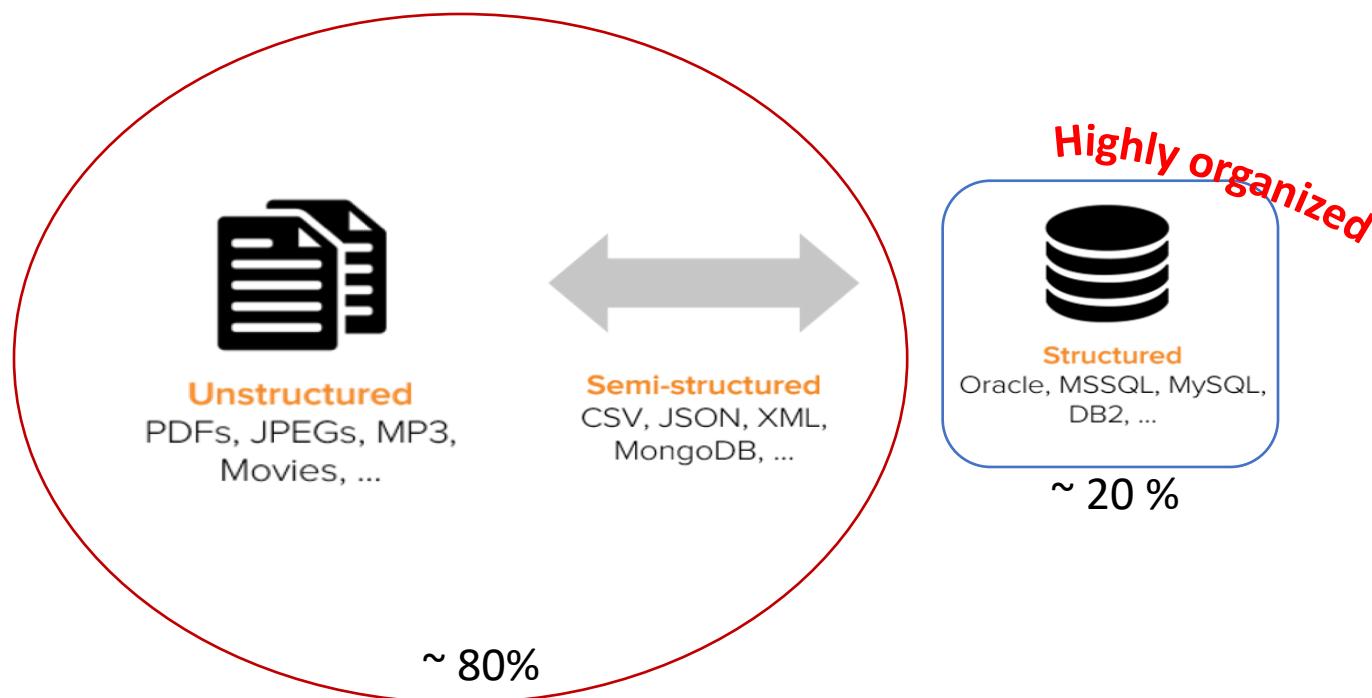
What is Big Data?

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](#))



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

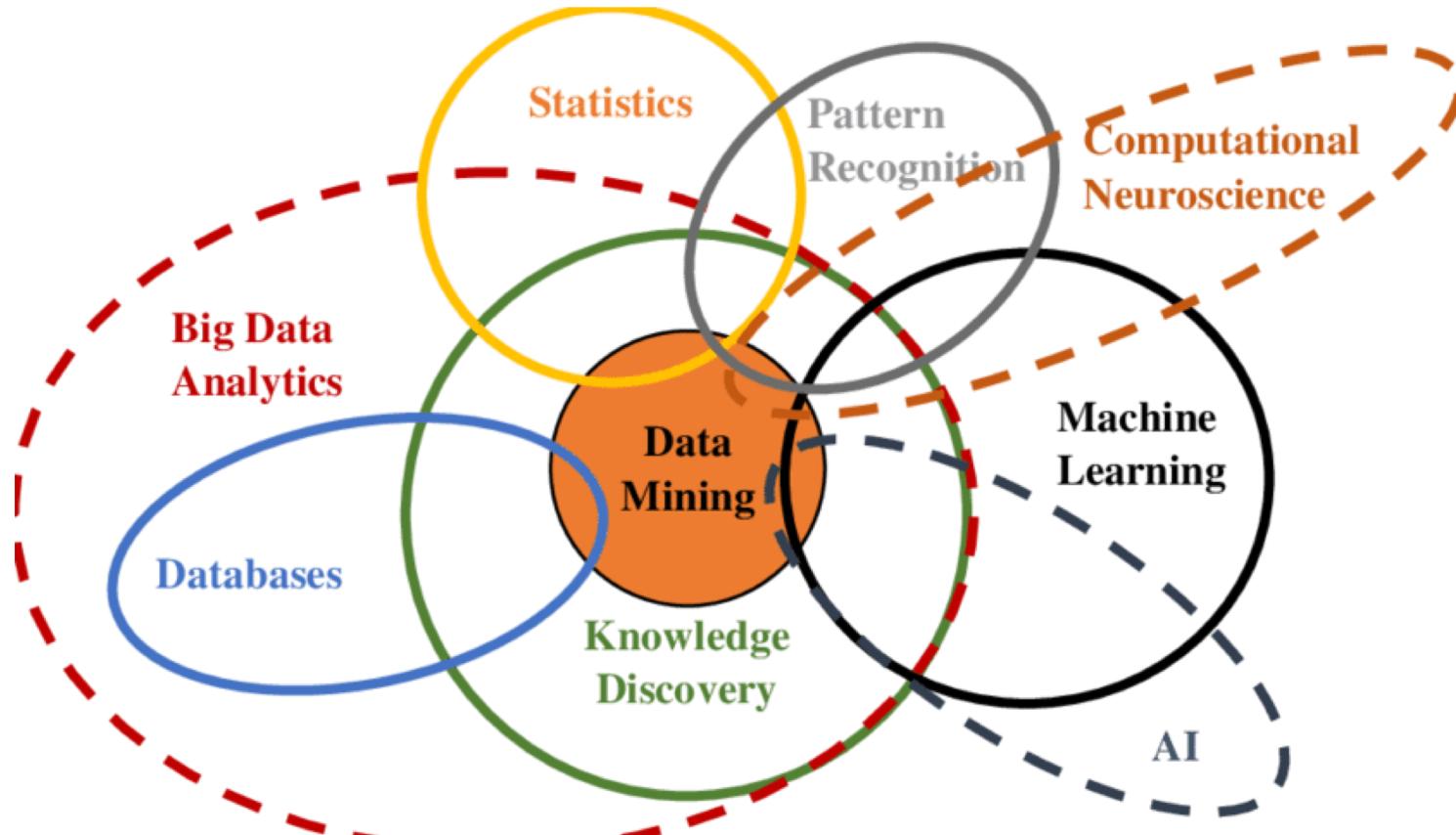


```
<?xml version="1.0" encoding="UTF-8" ?>
<customer_order number="004985" date="2004-06-24">
- <lines>
- <line no="1">
  <item>Disc CD</item>
  <quantity>30</quantity>
  <price>0.95</price>
</line>
- <line no="2">
  <item>Disc CD-RW</item>
  <quantity>20</quantity>
  <price>2.95</price>
</line>
</lines>
- <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>
```



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

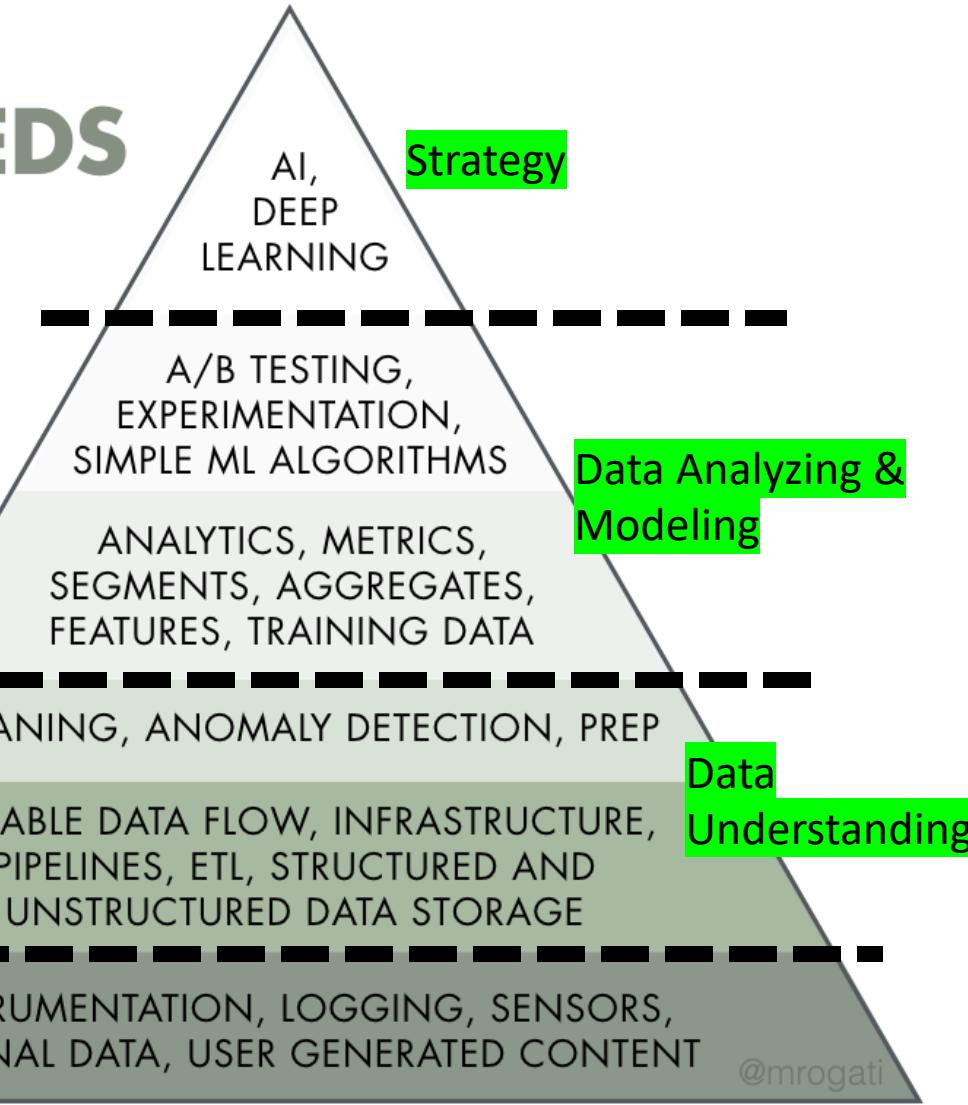
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Data Collecting & Storage

INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT



SMALL COMPANY
- STARTUP -

THE DATA SCIENCE HIERARCHY OF NEEDS

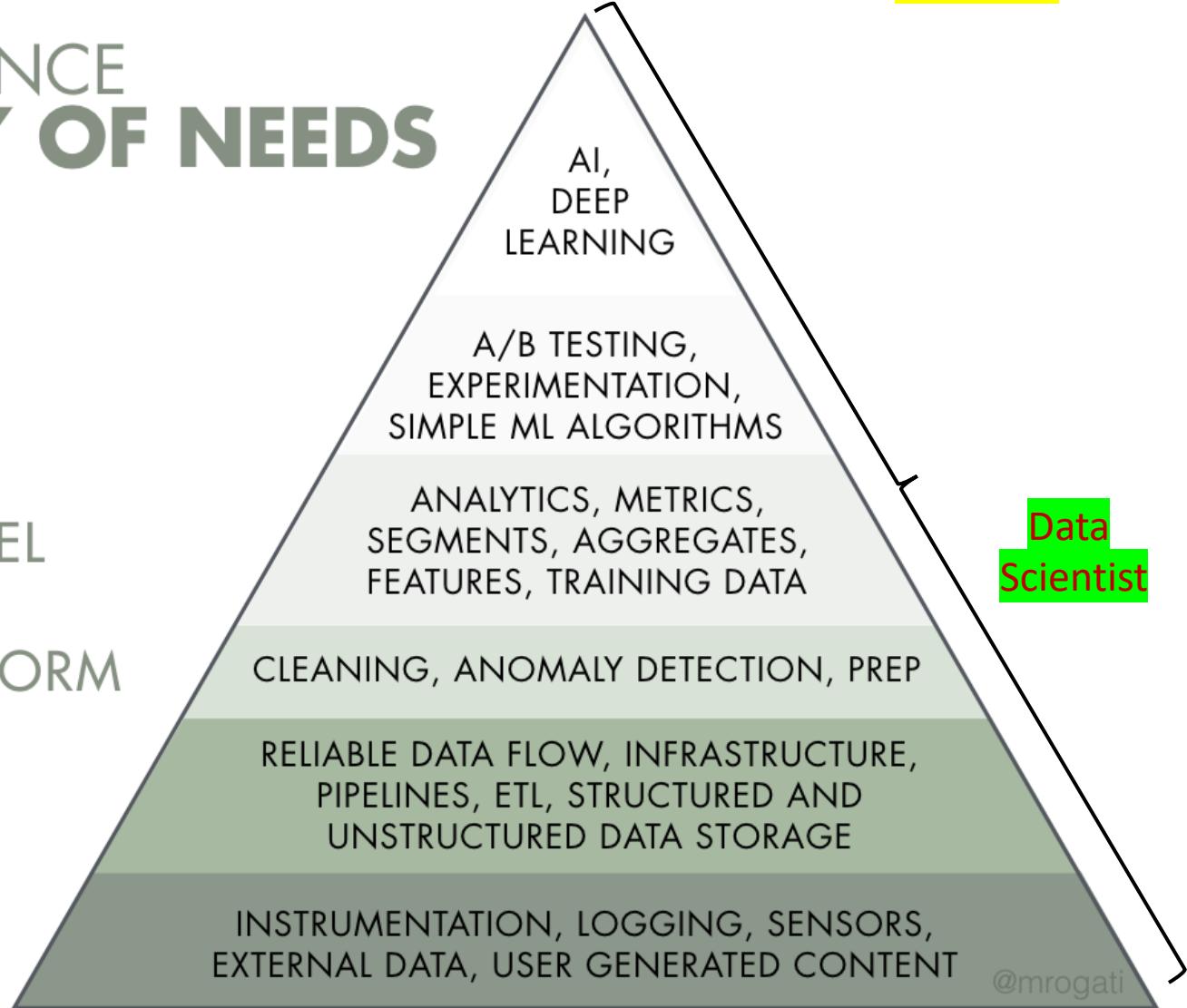
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



MEDIUM COMPANY

THE DATA SCIENCE HIERARCHY OF NEEDS

LEARN/OPTIMIZE

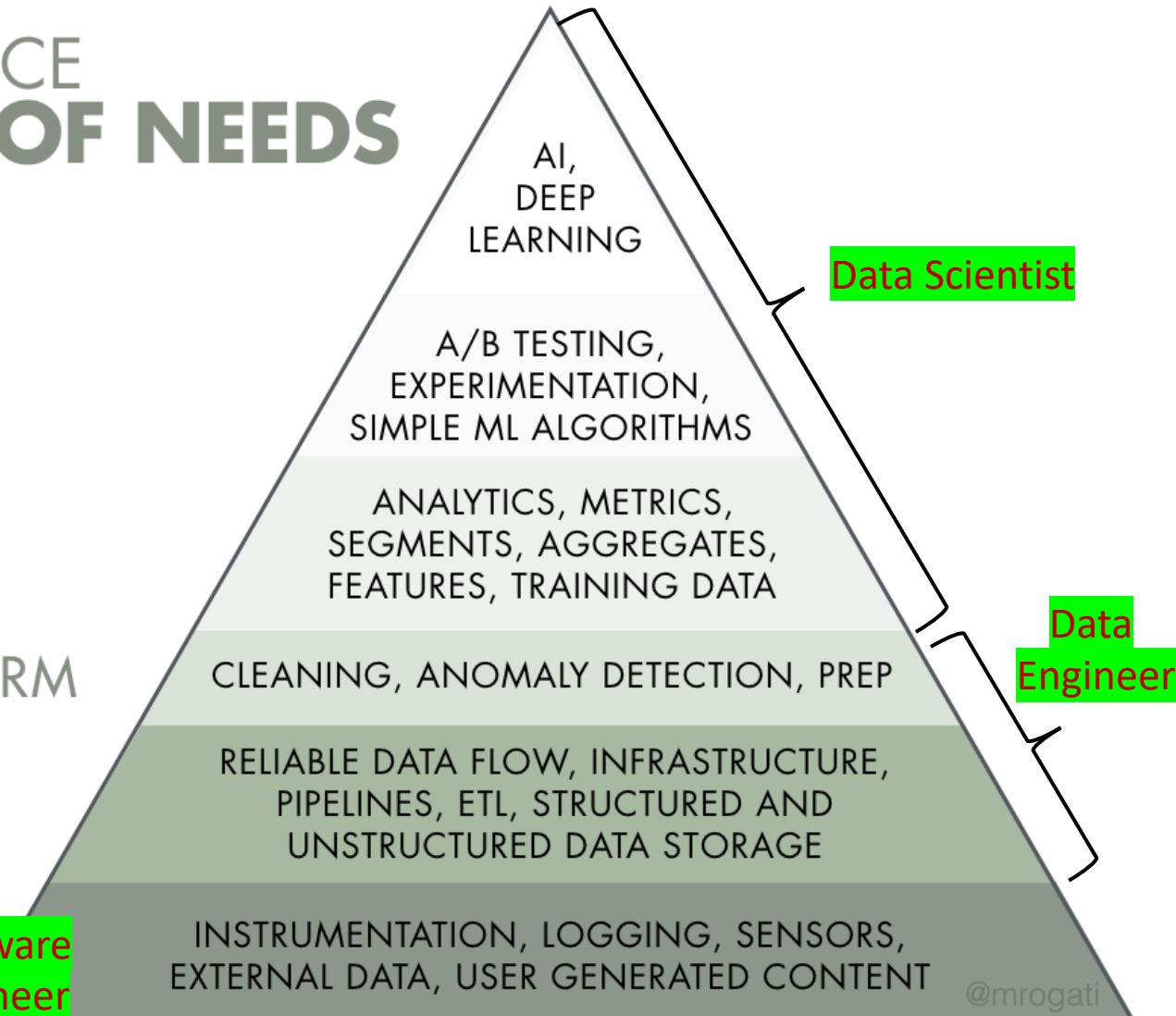
AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Software
Engineer



@mrogati

THE DATA SCIENCE HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Software
Engineer

INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT

RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE

CLEANING, ANOMALY DETECTION, PREP
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA

A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS

AI,
DEEP
LEARNING

- Research Scientist
- Data Scientist Core
- Machine Learning Engineer

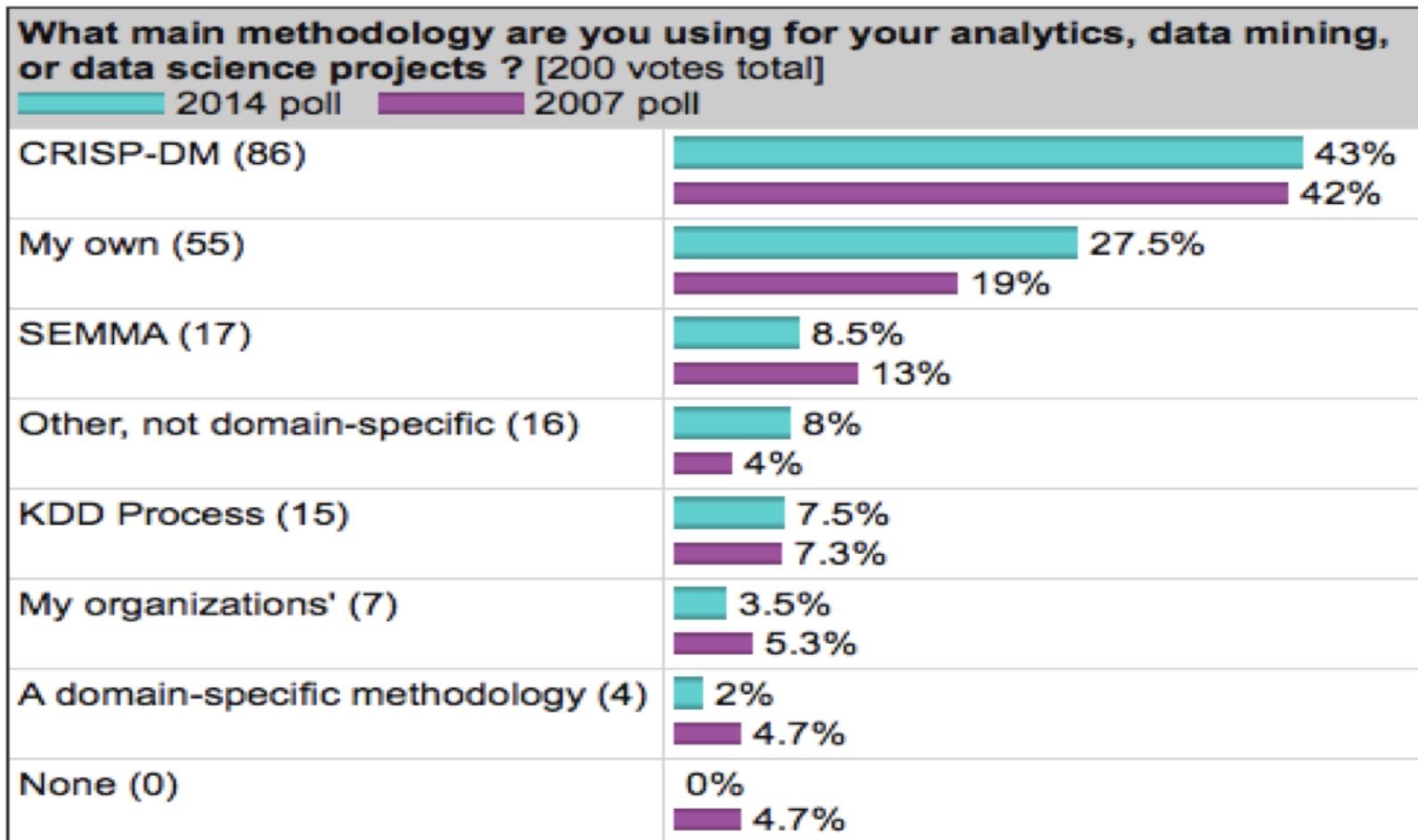
Data
Analyst

Data
Engineer

@mrogati

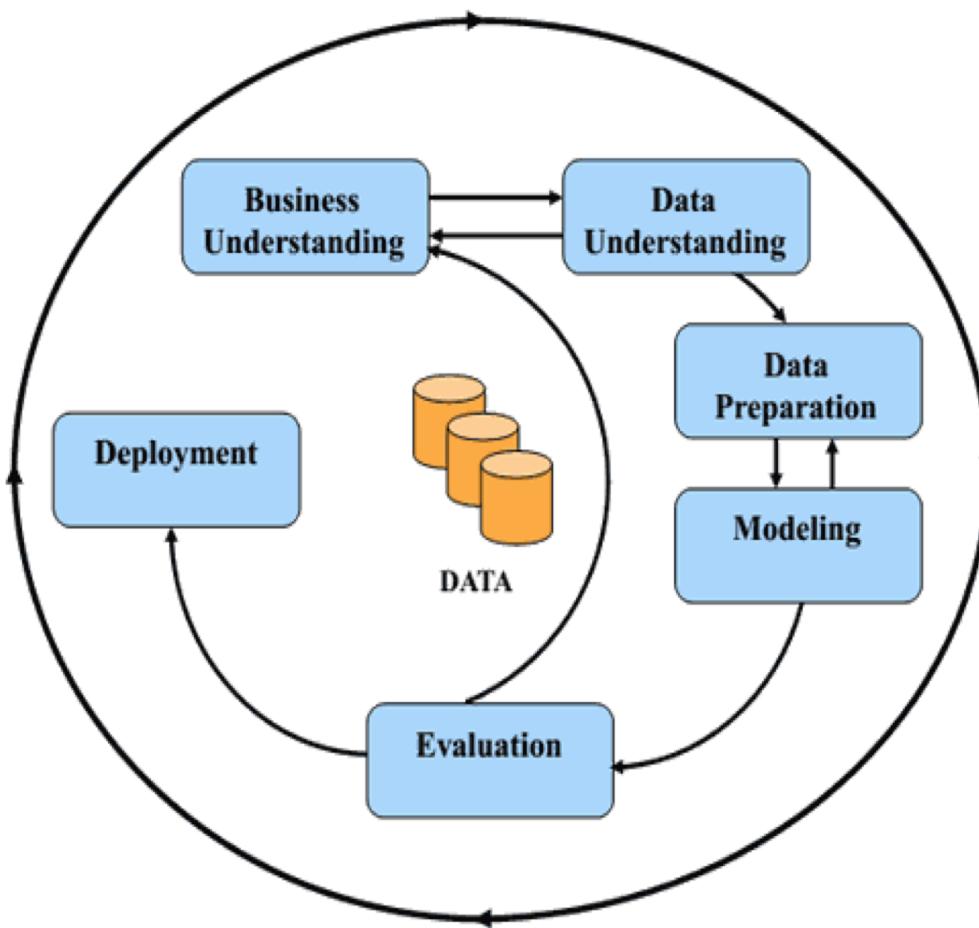
LARGE COMPANY

Methodology for Data Analytics



What is CRIPS-DM?

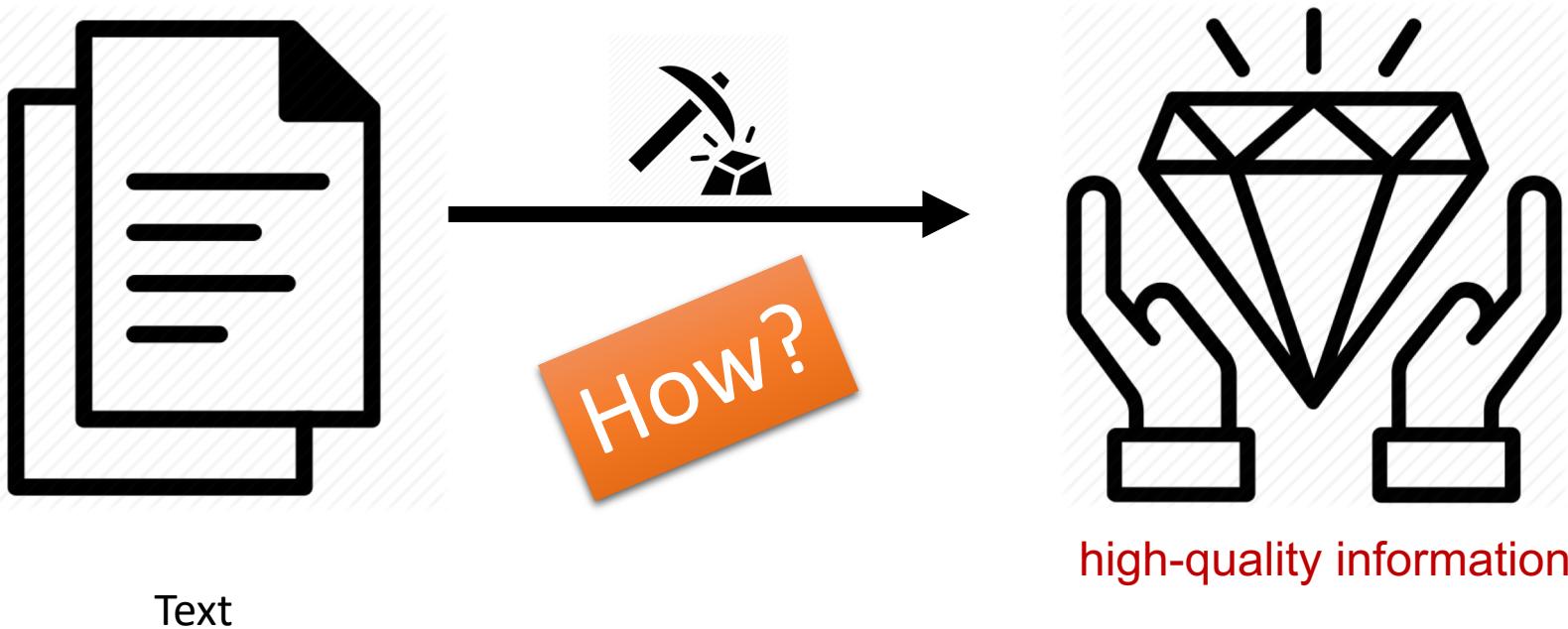
CRISP-DM: CRoss-Industry Standard Process for Data Mining (1996)



- 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**.



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

English ▾ Graphical ▾

I really enjoyed using the Canon Ixus in Madrid on March 4. The Panasonic Lumix is a bit disappointing, but the Canon camera is not bad at all. All I want when taking photos is point it and then just press the button. For only 200 dollars, a really fair price, this camera is perfect for me. Besides, I have had a good customer service experience.

*(characters: 347 / 400)

Analyze Text ►

Results Legend

I **① really enjoyed using** the **① Canon Ixus** in Madrid on March 4. The **② Panasonic Lumix** **② is a bit disappointing**, but the **③ Canon camera** is **③ not bad** at all. All I want when taking photos is point it and then just press the button. For only 200 dollars, a **④ really fair** **④ price**, this **⑤ camera** is **⑤ perfect** for me. Besides, I have had a **⑥ good** **⑥ customer service experience**.

Positive

Negative

Text mining around us

- Document summarization

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

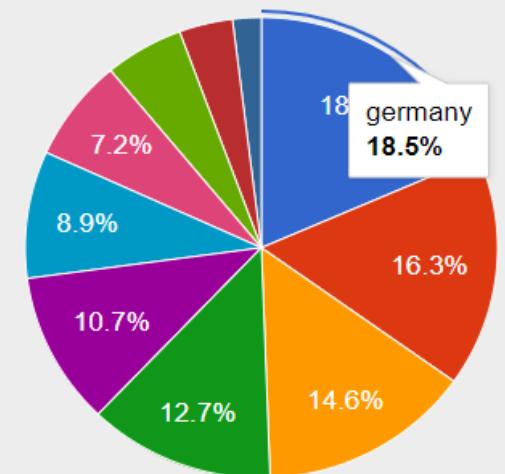
[View in Browser](#) [Learn More](#)

 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)

Informative Keywords

- germany
- carbon emissions
- country
- carbon country
- carbon releasing
- united kingdom
- correspondingly values
- lowest measures
- largest emitter
- biggest emitter



Text mining around us

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

Based on your Profile and Career interests

3 job titles · County Dublin, Ireland · Any industry · 200 to 10,000+ employees ... [Update Career interests](#)

The image shows three LinkedIn search results cards. The first card on the left is for a 'Machine Learning Engineer' at 'Xilinx' via 'Saggart'. It includes a LinkedIn logo, the job title, company name, and contact information, along with a '1 alum' badge and a timestamp of '2 days ago'. The middle card is for a 'Junior Data Analyst' at 'National Treasury Management Agency' in 'Dublin 2, Dublin, Ireland'. It shows a green triangle icon, the job title, company name, location, a '4 company alumni' badge, and a timestamp of '2 days ago'. The third card on the right is for another 'Machine Learning Engineer' at 'Xilinx' via 'Csonant'. It follows a similar layout with the LinkedIn logo, job title, company name, and a timestamp of '2 days ago'.

Google News

The image shows a Google News sidebar with a dark background and white text. At the top, it says 'Recommended'. Below are several news items with small thumbnail images, titles, and brief descriptions. The first article is about a manhunt for Las Vegas strip shooting suspects. The second is about John Avlon's 'anti-Rove' Nazi ad. The third is about tracking State of the Union proposals. The fourth is about Ronda Rousey's UFC bout. The fifth is about Cornell bioengineers printing human ears. The sixth is about DOJ moves against Armstrong.

Recommended

Manhunt underway for suspects in fatal Las Vegas strip shooting
Reuters - 7 hours ago

Editor's note: John Avlon Is a CNN contributor and senior political columnist for Newsweek and The Daily Beast. He is co-editor of the book "Deadline Artists: America's Greatest Newspaper Columns."

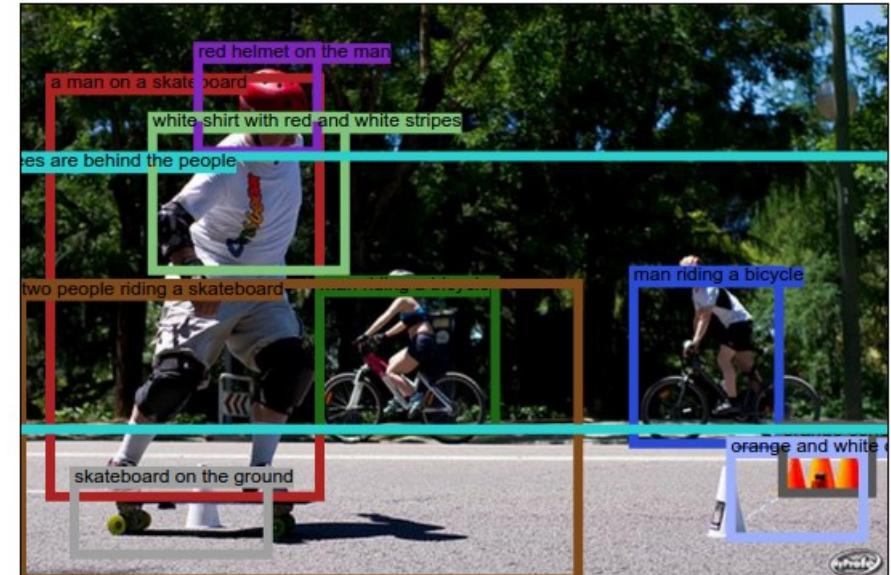
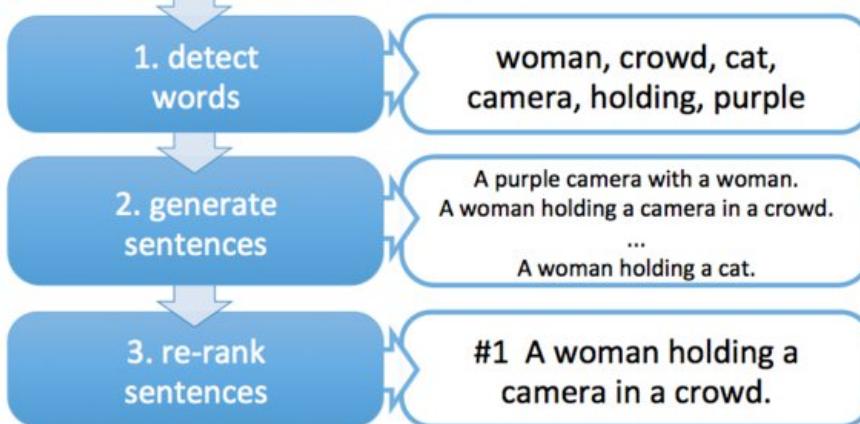
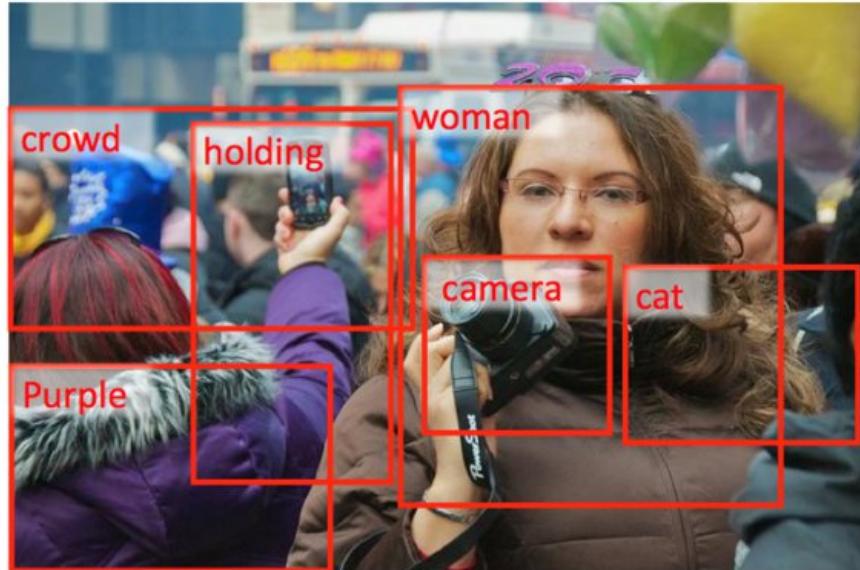
Tracking 66 State of the Union proposals
USA TODAY - Feb 20, 2013

Ronda Rousey's maverick ways lead to landmark UFC bout
Los Angeles Times - 3 hours ago

Cornell bioengineers print human ears
Cornell News - 10 hours ago

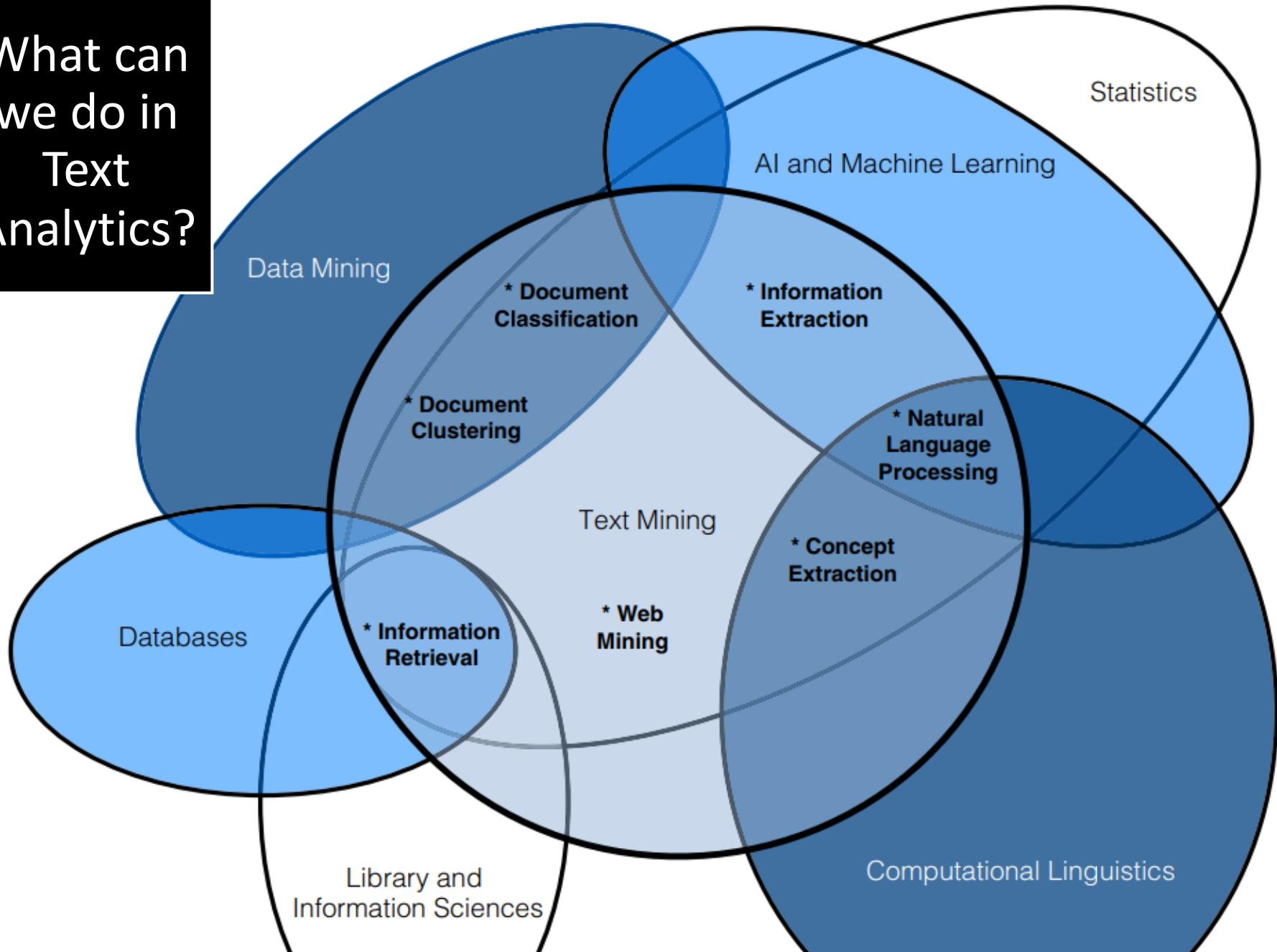
DOJ move a bad sign for Armstrong
USA TODAY - 14 hours ago

Interdisciplinary Tasks: Image Captioning



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?



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.

What makes Text Analytics so Hard?

Ambiguity



What makes Text Analytics so Hard?

Non-Standard Language



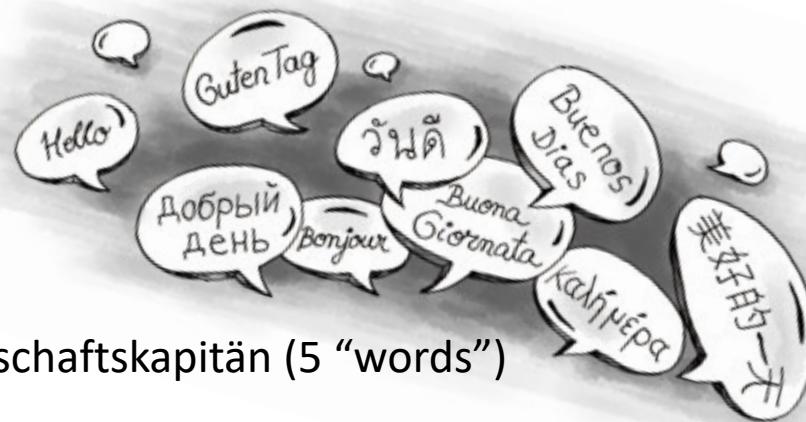
- LOL Lee O. Langdon
LOL laugh(ing) out loud



"Stop signing memos with your initials;
no one is taking them seriously."

What makes Text Analytics so Hard?

More Complex Languages Than English



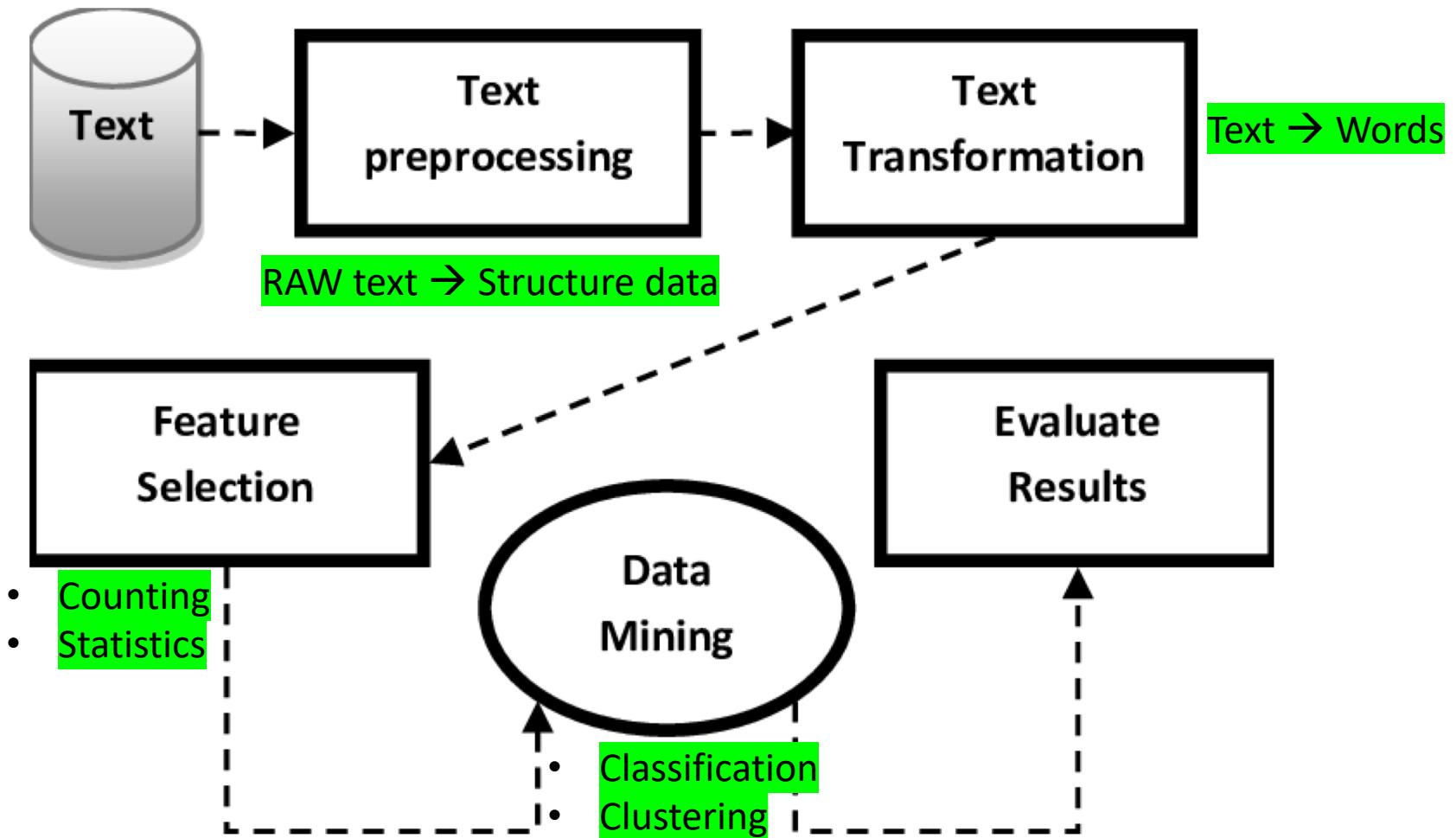
German: Donaudampfschiffahrtsgesellschaftskapitän (5 “words”)

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