



# DATA SCIENCE CONSULTING

## Session 2

February 1st, 2021



# Agenda



BUSINESS4DATA



5 min

## 1. Data project kick-off

15 min

2. Hospitality industry & restaurants KPI definition

25 min

3. Deep dive into KPI

13 min

4. Hands-on & session closing

2 min

5. For next session

DATA4BUSINESS



# Data project presentation

*Ex: implement a Data Platform*



## IMPLEMENTATION TARGETS

- Identify & implement eligible use cases
- Build a roadmap with quick wins
- Reinforce data management and data protection
- Onboard key stakeholders to train the trainer

## SCOPE

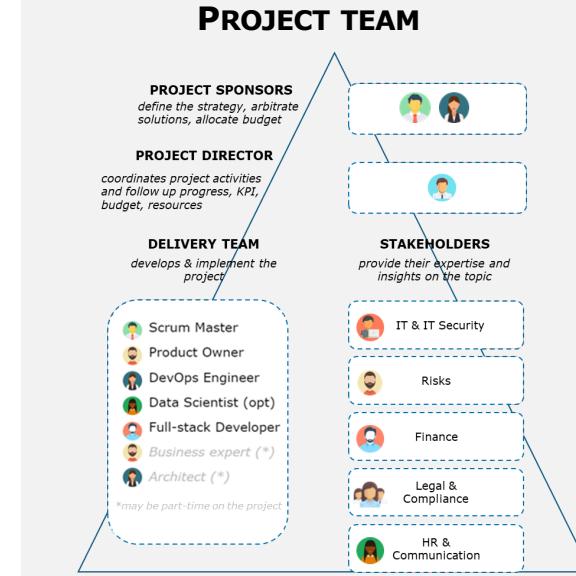
- Data to comply with regulatory reporting
- Data to comply with daily financial reporting
- Data on the customers' acquisition/retention
- Out of scope : Data on the employees' turnover

The **strategic objectives** of a data platform implementation are to:

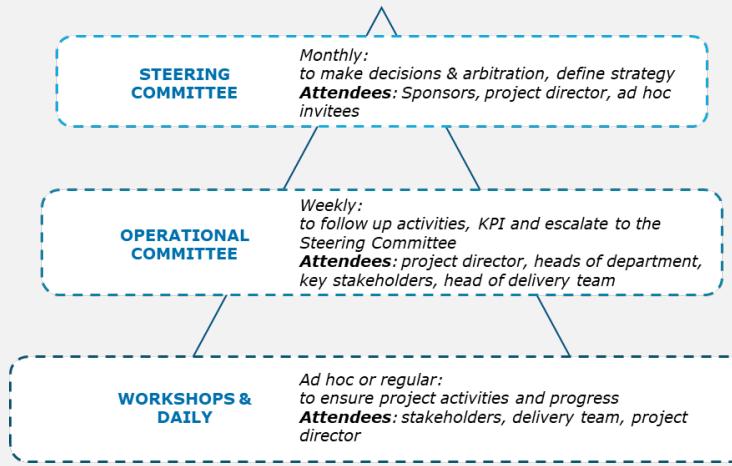
- Develop a data-driven approach to transform business processes
- Promote a data culture into the different parts of the company



## APPROACH



## COMITOLOGY



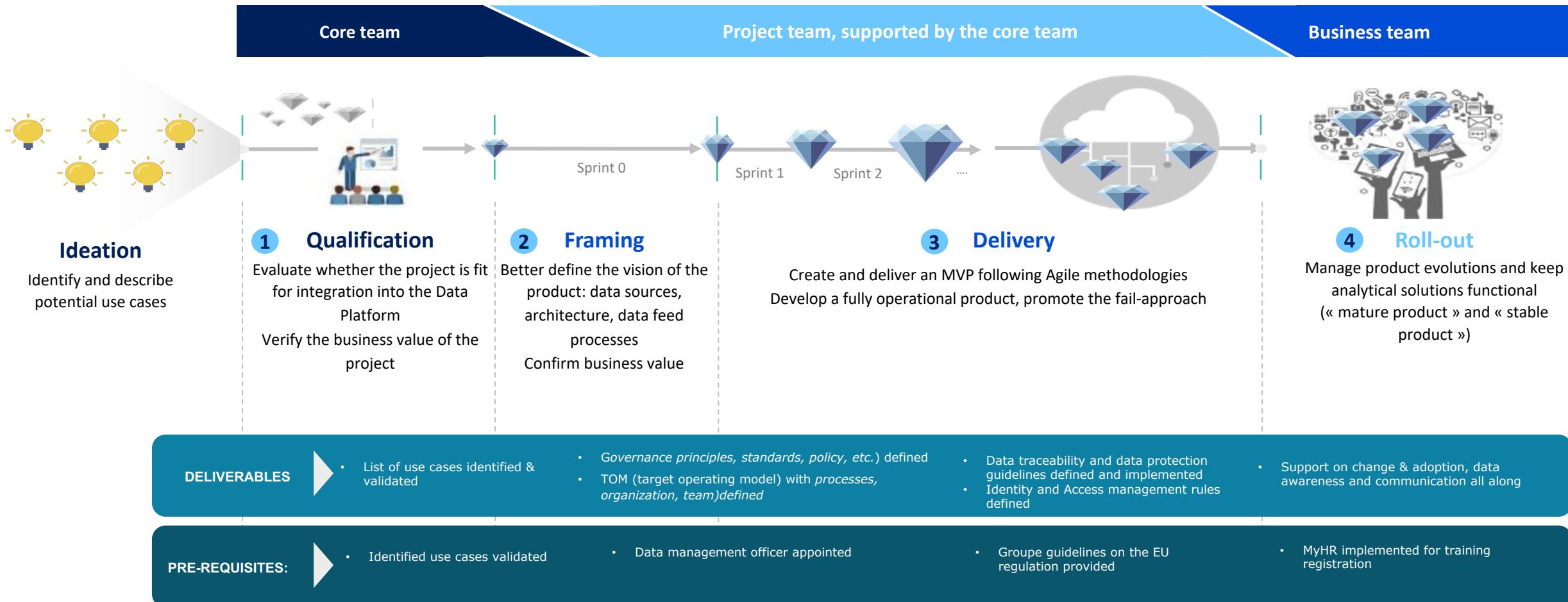
## KEY SUCCESS FACTORS

- Sponsor et stakeholder's availability
- Access to data
- Business alignment
- Data quality
- Clear client expectations



# Data project roadmap

*Ex: implement a Data Platform*





# Agenda



BUSINESS4DATA

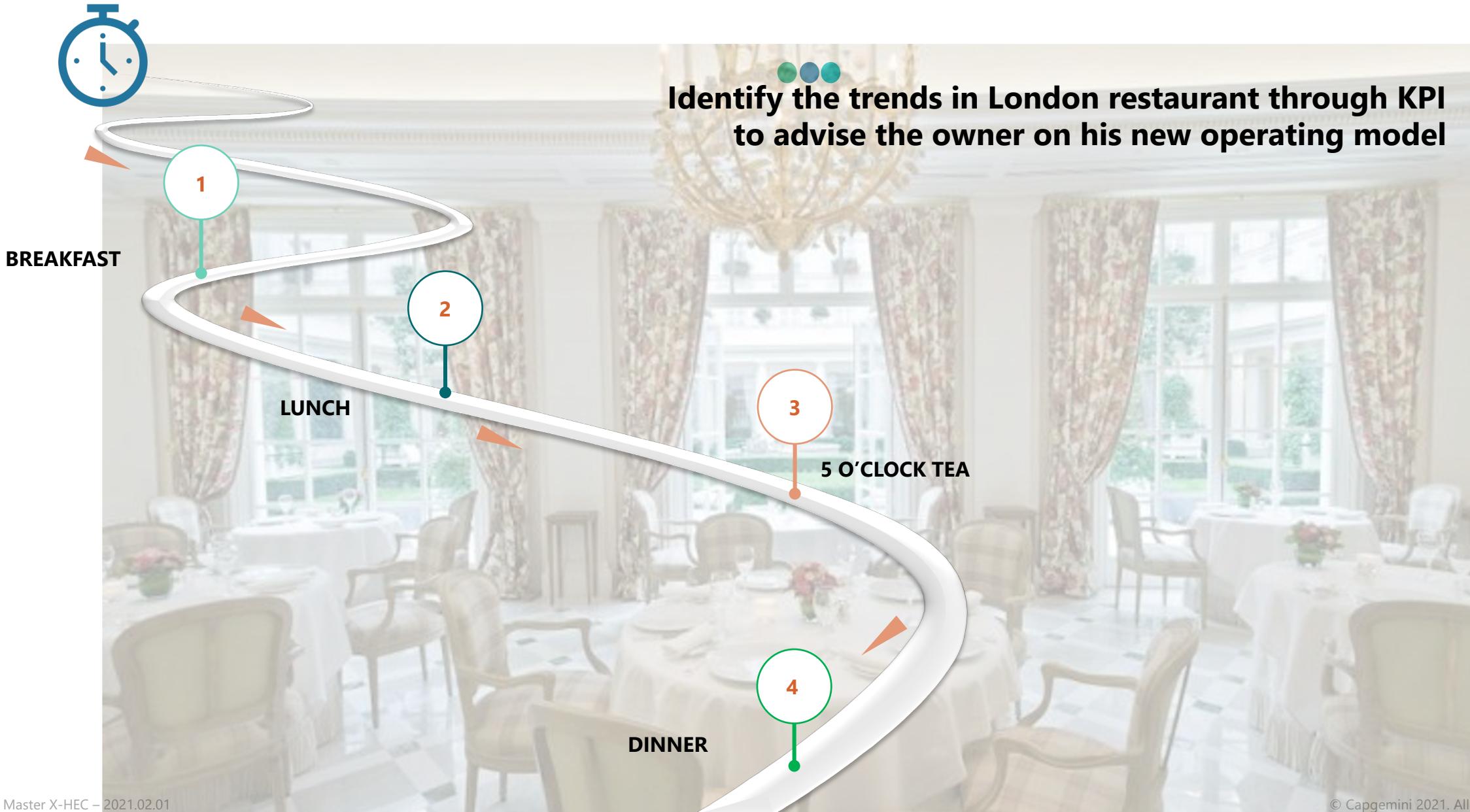


1. Data project kick-off 5 min
2. **Hospitality industry & restaurants KPI definition** 15 min
3. Deep dive into KPI 25 min
4. Hands-on & session closing 13 min
5. For next session 2 min

DATA4BUSINESS

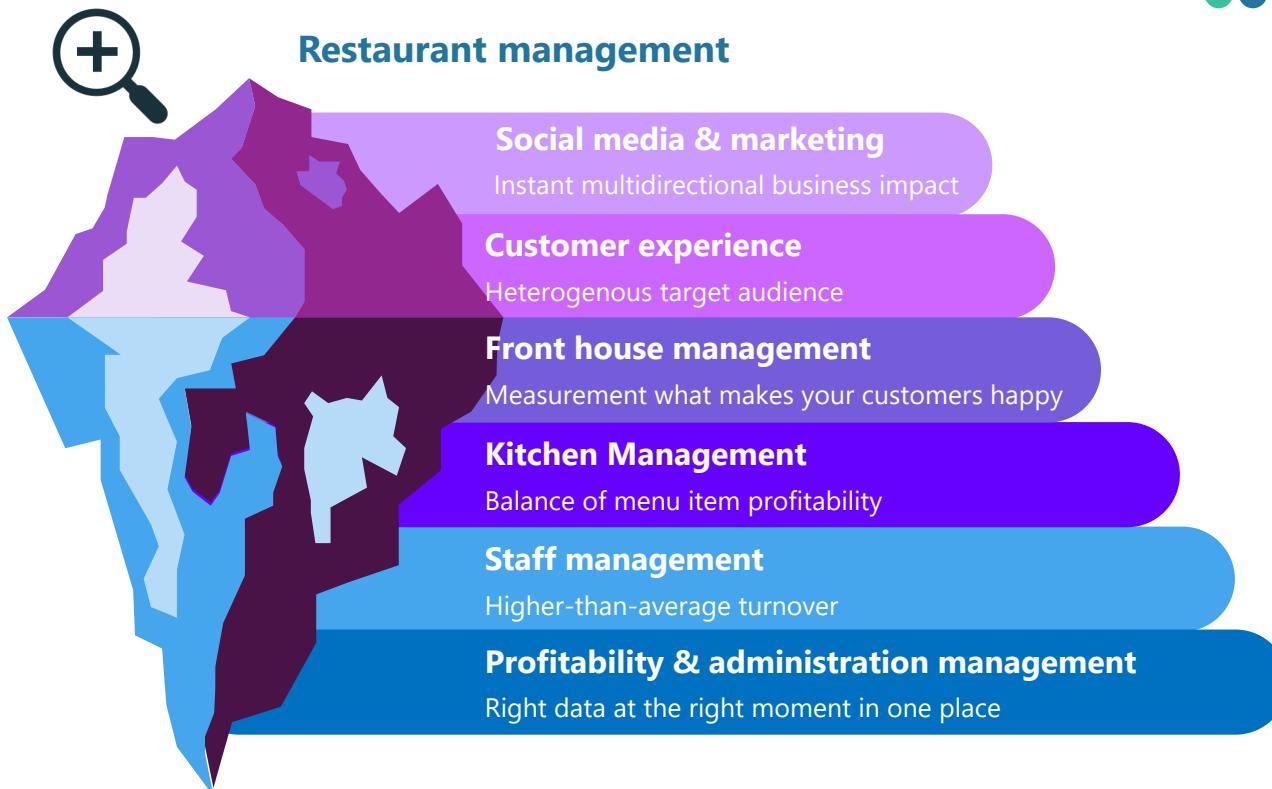


# Today's objectives:





# KPI: objectives, features & main domains



KPI monitoring **is not an end in itself**, but a **mean** to **achieve successful restaurant management**. KPI monitoring should be coherent and be integrated into global strategy.



## KPI: objectives, features

As a restaurateur, tracking your KPIs is a necessary part of daily business – and a key process when it comes to accelerating your success.

Essentially, a **KPI is a performance measurement** that is used to evaluate how effectively your company is achieving its key business objectives.

If you don't track (or haven't set!) KPIs, you can't evaluate how effectively you are operating, and eventually will be unable to pinpoint and **fix issues as they** – inevitably – **arise**.

### KPIs must be:

- Applicable
- Generated and measured with actual data
- Properly defined
- Communicated clearly
- Monitored regularly



# Agenda



BUSINESS4DATA



1. Data project kick-off 5 min
2. Hospitality industry & restaurants KPI definition 15 min
- 3. Deep dive into KPI** 25 min
4. Hands-on & session closing 13 min
5. For next session 2 min

DATA4BUSINESS



# Key social media & marketing KPI

In an age where a poor **website** and lack lustre **social media presence** can **make or break a business**, it's crucial to monitor these promotional channels effectively and to have a strategy that brings all of your marketing together.

For example, customers have a hard time finding what they're looking for, such as opening times, address or the menu, it won't take much for them to leave your website or social media profile and look for other options. **The less clicks it takes to find out about your restaurant, the better.**

## SOCIAL ENGAGEMENT METRICS

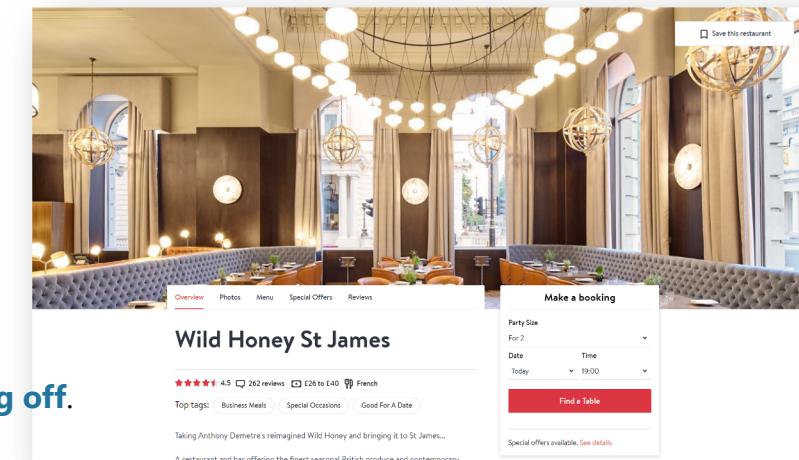
Social media can take considerable time and effort. If you're devoting time and money to amplifying your social media presence, you want to understand if it's worth it. All of the following metrics measures the level of interaction happening between your restaurant's social media profile and current as well as prospective customers:



- **Likes:** The number of clicks on your post "Like" buttons.
- **Comments:** The number of comments on your posts.
- **Shares:** The number of times a post has been shared on social media, retweeted on Twitter or repined on Pinterest.
- **Engagement rate:** This measures the level of interaction that happens between your social media account and audience.

By measuring these, you can **get an immediate sense of how your social media marketing efforts are paying off**.

A social media reporting tool like *Hootsuite* can quickly gather this data for you.



## WEBSITE TRAFFIC METRICS

Your website should be your number one promoter, tirelessly working for you 24/7 to attract new customers to your restaurant. These KPIs are surefire indicators of how well your website is doing and can easily be uncovered in your *Google Analytics* data.



- **Sessions by channel:** A useful way to look at your traffic is how it breaks down by referral source. Looking at the number of people who visit your site through organic search, referral, social, or paid search will give you a general idea of which channels are driving the most traffic to your site
- **Website engagement:** Bounce rate (the percentage of visitors who leave your site having only viewed one page) is another important metric. If getting users engaged with your content is a priority for your business, one another one worth monitoring is your average pages per session
- **Keyword rankings and search traffic:** By monitoring your keyword rankings and search traffic, you can gauge whether your restaurant is likely to be found by the right people
- **Conversions:** With the help of call tracking software and Google Analytics, you can track key conversions for your restaurant, such as bookings over the phone



# Customer experience KPIs

A **whopping 97% of customers read online reviews before** they buy, with almost half of those trusting 4-stars or better. Similarly, customer satisfaction should be at the top of any restaurant manager's list. **CUSTOMERS VOTE WITH THEIR WALLET** and to keep them coming back, a team of smiling and efficient staff is essential to encourage return visits.

- Online reviews :** The number of online reviews, as well as the average review scores on Google My Business, Tripadvisor and Yelp are also important metrics to monitor
- Bookings:** Track and measure your reservations throughout the year to identify patterns, and properly plan for restaurant usage during peak and low times. If you use a retail management system that offers integrated booking management, then this KPI will be really easy to track
- Customer retention rate :** Your restaurant's customer retention rate is the percentage of customers you keep, relative to the number you had at the start of a set period of time.

$$\text{CUSTOMER RETENTION RATE} = \frac{\text{THE NUMBER OF CUSTOMERS AT THE END OF A PERIOD} - \text{THE NUMBER OF NEW CUSTOMERS ACQUIRED DURING THAT PERIOD}}{\text{THE NUMBER OF CUSTOMERS AT THE START OF THAT PERIOD}} \times 100$$

This KPI holds immense value when you consider the cost of acquiring new customers. To keep growing your restaurant, you need to make up for every customer you lose on top of attracting new customers. Your customer retention rate also helps you gauge how loyal your current customers are and, by association, how good your customer service is.

A mere **5% increase in retention** can **increase company revenue by 25-95%**. 90% of happy consumers are more likely to purchase again, and 93% are more likely to be repeat customers at companies with excellent customer service.



## TIPS & TRICKS

"Just take a few minutes to read posts. While preparing a trip, any advice is important. The references of [www.booking.com](http://www.booking.com) and [www.TripAdvisor.com](http://www.TripAdvisor.com) are very helpful.

There is one interesting point to mention. I used to compare tourists' comments on any hotel in European countries posted in different languages (English, French and Russian). Results are absolutely amazing! Feedback provided in different languages prioritizes different aspects of customer experience.

You may see that **English-speaking** tourists mostly pay attention to good location, friendly staff. Ready to continue? Comments in **French** contribute to understand whether the breakfast is delicious, the room is noisy. Just one more tip in **Russian**: you may check whether the hotel's attractive description complies with the reality, everything works properly in the room."

Source: <https://www.linkedin.com/pulse/preparing-your-multicultural-customer-journey-svetlana-olivier/>



# Front-house management KPI

Your customers are the most important part of your restaurant. Measuring what keeps them happy and returning to you is incredibly valuable when it comes to determining what to change, maintain or improve.

- **Food and beverage sales per guest:** get a handle on which menu items appeal most to guests, and whether time or day impact total spend. Tracking this metric will allow you to make menu adjustments – for example removing underperforming dishes that might negatively impact kitchen productivity while resulting in little profit or increasing prices on popular and hot-selling items to improve your margins. This metric can also help you determine which promotions to run – for example, happy hour to increase sales during slow periods.
- **RevPASH: Revenue per available seat per hour** is a daily/hourly calculation which takes time into account when calculating how effectively each seat in your restaurant is driving revenue. Developed by Dr. Sheryl Kimes at Cornell University, RevPASH is thought to be one of the best measures of success in the restaurant industry.

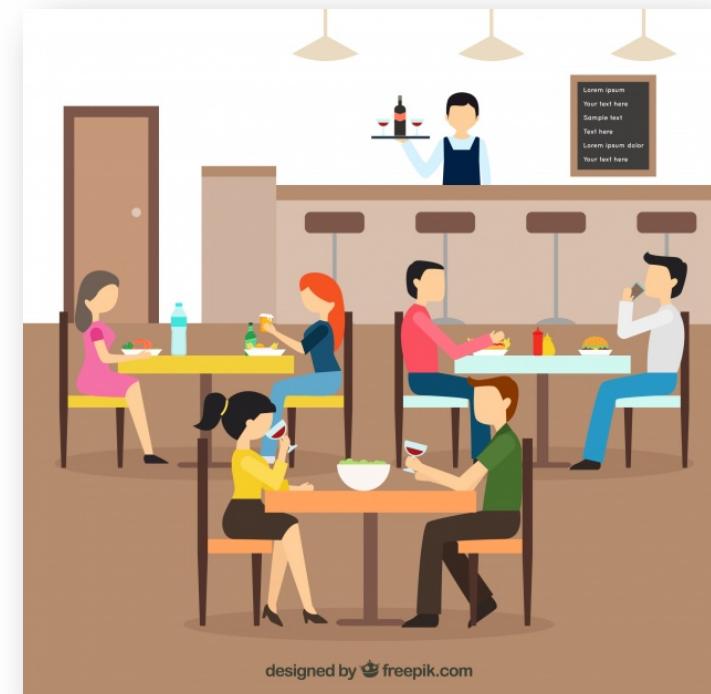
$$\text{RevPASH} = \frac{\text{OVERALL REVENUE}}{\text{SEATS AVAILABLE}} \times \text{OPEN HOURS}$$

- **Table turnover rate:** you'll want to ensure your table turnover is optimally timed to allow your customers to fully enjoy their dining experience and new parties aren't waiting around to be seated. A poor table turnover rate is detrimental to your restaurant's profit margins, which is why it's important to keep a close eye on this particular metric.

$$\text{TABLE TURNOVER RATE} = \frac{\text{PERIOD OF TIME}}{\text{NUMBER OF TABLES SERVED DURING THAT TIME PERIOD}}$$

- **Average table occupancy :** this metric tells you how many customers on average visited your restaurant during a particular period of time.

$$\text{AVERAGE TABLE OCCUPANCY} = \frac{\text{NUMBER OF OCCUPIED TABLES}}{\text{TOTAL NUMBER OF AVAILABLE TABLES}}$$





Social media &amp;

marketing

Customer experience

Front house management

Kitchen Management

Staff management

Administration &amp; profitability management

# Kitchen management KPI

The key to good kitchen management is finding a balance between providing great value to the customers, while keeping the cost down. This includes reducing waste, optimizing the menu and reducing waste to keep down food costs

Diversifying your restaurant menu for niche customers "Do you serve gluten-free options?" This question, along with other dietary restriction requests, have become common among restaurant-goers for many reasons.



**Menu item profit and popularity:** pricing items on your menu impacts profitability. A high-profit menu item is great – but not if it produces a statistically insignificant percentage of your sales. Items with a higher sales velocity but lower profit margin might be driving all your revenues. Get a true handle on the profits of each menu item to help drive menu engineering and promotion decisions

- **Production time per dish:** Knowing how long it takes to produce each dish is key, as it can help you determine the value of each single dish based on turnaround time and expectations. Track popular (or star) items as well dishes with low sales volume and consider ways to shorten production time and offer faster service. If a popular dish takes too long to prepare, it will negatively impact the customers experience as well as your kitchen's efficacy. And if items that are quick to produce are rarely ordered, consider swapping them out for more popular items and giving the kitchen more time to prepare the items customers want, faster.
- **Food wasted per food purchased:** food waste is a huge concern for restaurateurs worldwide. Monitoring food waste enables you to improve your demand forecasting, forces you to reconsider how and where you procure your foods, and leads to better management and storage your food stock. It can also help you determine if there are better methods to prepare your food, and if you need to reconsider your portion sizes or serving methods.

Other kitchen management KPIs to consider tracking are:

- Food costs per head
- Kitchen linens costs
- Value of stock
- Total food costs
- Best/worst selling items



## TIPS & TRICKS

Is there any correlation between napkins' choice & sales increase per guest?

In a high style restaurant, well known for its fantastic dishes, elegant design, great services, high prices, etc. the only stuff differs from everything: guest napkins are not of 100% linen, containing a few %tile of synthetic fabrics. It is not obvious while touching. However, napkins slide down sometimes. Why not to change these sliding napkins?

It is a matter of sales: every time a waiter assists the client to get the napkin back or to replace, it is an opportunity to replenish his ... glass of wine, water etc.



# Staff management KPI

Any restaurant is only as **good** as its people, **including both the management and staff**. Since labour costs tend to be among the highest costs in any establishment, it's important to have an idea about how productive your staff are and how much sales they bring in. Similarly, it's very important to ensure that **employees are engaged and happy at work**, as **high employee turnover** impacts productivity and the bottom line.

- **Sales per employee per hour:** looks at how much revenue each employee generates
- **Employee turnover:** this refers to the **percentage of employees who leave your restaurant during a certain period of time**. When you factor in the expenses associated with finding, hiring and training new employees, a higher-than-average employee turnover rate can prove to be an expensive problem for your restaurant.

$$\text{EMPLOYEE TURNOVER RATE} = \frac{\text{NUMBER OF EMPLOYEES WHO LEFT DURING THE TIME PERIOD}}{\text{AVERAGE NUMBER OF EMPLOYEES}} \times 100$$

Additional staff management KPIs to consider are:

- Wage cost percentage
- Total labor hours
- Average hourly pay

## TIPS & TRICKS

When thinking on equip the team with tools/tablets/apps, think twice. In a high-turnover environment, you may not invest too much time to train newcomers...





# Administration et profitability management KPI

Even if customers and food service make up the bulk of your daily life, no restaurateur can escape **the back-end management aspects** of the business. Knowing how your sales and marketing efforts affect **your revenues and being aware of your cash flow position** is necessary for you to thrive.

- **Cash flow** : It takes many things to start and run a successful restaurant – a passion for good food, determination and the right staff, but the best-laid plans will go nowhere without **one key ingredient**. And **that's cash**.

One of the most important things to track as a restaurant is **the amount of money that's going in and out** of your restaurant and how much cash you have on hand. If you're not assessing your cash flow on a regular basis, you could be setting your restaurant up for failure.

$$\text{CASH FLOW} = \text{BEGINNING CASH} - \text{ENDING CASH}$$

- **Cost of goods sold** (or COGS) : helps you measure the amount of money that goes into buying supplies and good ingredients for your menu items. Before you can even think about calculating your restaurant's profit, you need to know your cost of goods sold. If you track only one thing in your kitchen, make it this one. Cost of goods sold is **likely the biggest expense for any restaurant**; knowing where you can reduce the costs is a key factor in increasing profitability.

$$\text{COGS} = \text{BEGINNING INVENTORY} + \text{PURCHASES DURING THE PERIOD} - \text{ENDING INVENTORY}$$

- **Labour cost percentage**: this is critical restaurant KPI, **it shows how much is spent on labour as a percentage of revenue**. Just knowing your staff's hourly or annual rates isn't enough. Consider wages alongside all the other labor expenses a staff member incurs, including sick and vacation time, staff discounts, benefits and insurance, as well as taxes and other local wage-related laws. This value should be expressed as a percentage of your total sales.

$$\text{LABOUR COST PERCENTAGE} = \frac{\text{AMOUNT SPENT}}{\text{TOTAL SALES}} \times 100$$

- **Return on investment (ROI)** : As any business owner knows, this metric is essential to determining profit.

$$\text{ROI} = \frac{\text{NET VALUE OF CASH FLOWS}}{\text{COST OF INVESTMENT OR CAPITAL EMPLOYED}} \times 100$$

## TIPS & TRICKS:

The proper way to calculate a return is using the "cash flow method", it should meet **at least 15% ROI minimum in your first year**, and you are in a good business if you could **reach 20 to 25% annual profit vs capital**. Good restaurant business requires sustainability over 5 years. In fact, **only 1 of 100 restaurants could reach over 25% profit vs capital**.



# Top tips for 2021

## BE SENSITIVE

In a challenging climate it's easy to be too reserved and stagnate, but at the same time don't get caught up in shiny or distracting technology that can't be easily implemented into your business. Review business, customers' and your needs first.

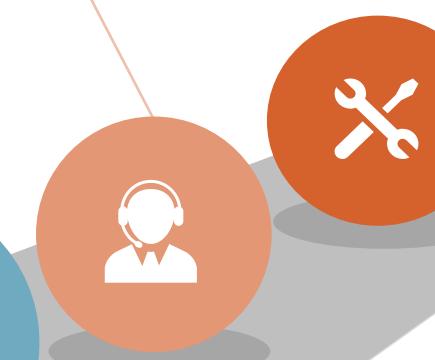
## KNOW YOUR CUSTOMER

Know who are you targeting. There is no point using technology for the sake of it. Your decision should be well considered and depend entirely on the needs of your customer base.



## FOCUS ON PROCESSES

Improve processes and create a frictionless experience. A better experience brings back customers.



## KNOW YOUR NEEDS AND BUDGET

Have a clear understanding of your needs and your budget before you start your research

## SELECT EASY-TO-USE TECHNOLOGIES

Prioritize intuitive technologies to benefit employees and your business as a whole. In a high employee turnover environment, make sure staff can be trained quickly and use it effectively to get real ROI

WHAT DO YOU THINK ABOUT COVID-19 IMPACT ON THESE 5 SUCCESS INGREDIENTS?



# Agenda



BUSINESS4DATA



1. Data project kick-off 5 min
2. Hospitality industry & restaurants KPI definition 15 min
3. Deep dive into KPI 25 min
- 4. Hands-on & session closing** 13 min
5. For next session 2 min

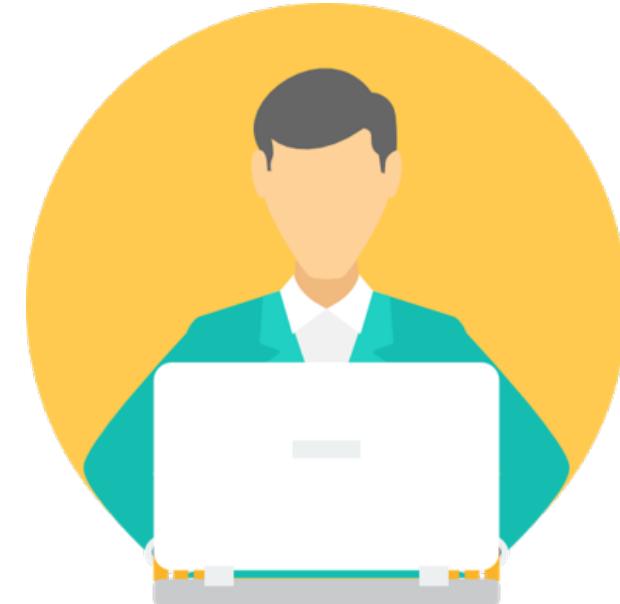
DATA4BUSINESS



# Hands-on & session closing

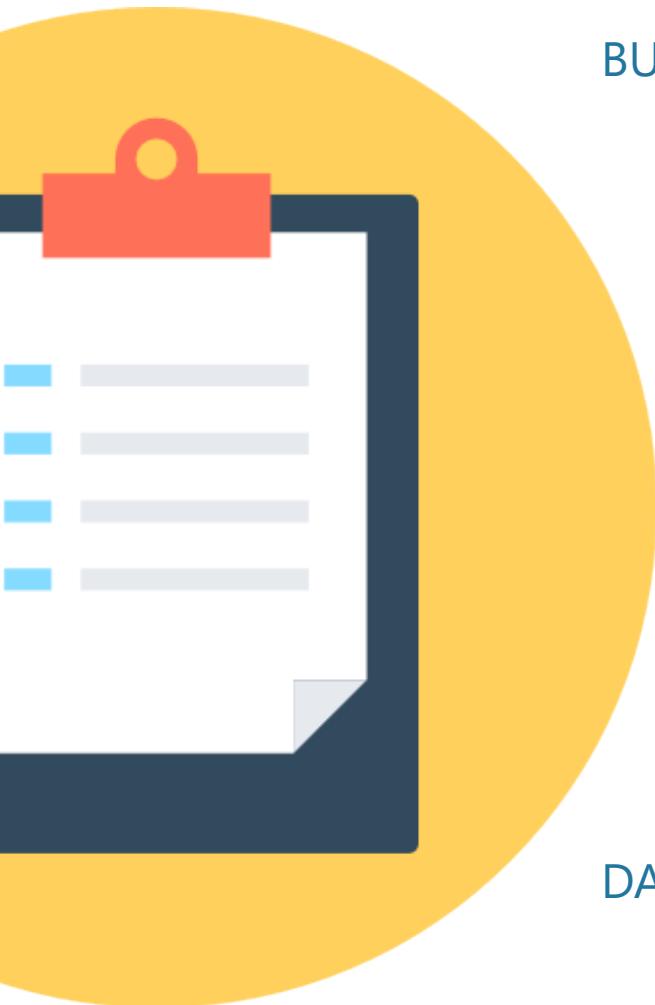
**Go to [www.menti.com](https://www.menti.com) and use the code 87 45 19 9**

*SHARE YOUR FINDINGS!*





# Agenda



BUSINESS4DATA



5 min

1. Data project kick-off

15 min

2. Hospitality industry & restaurants KPI definition

25 min

3. Deep dive into KPI

13 min

4. Hands-on & session closing

2 min

**5. For next session**

DATA4BUSINESS

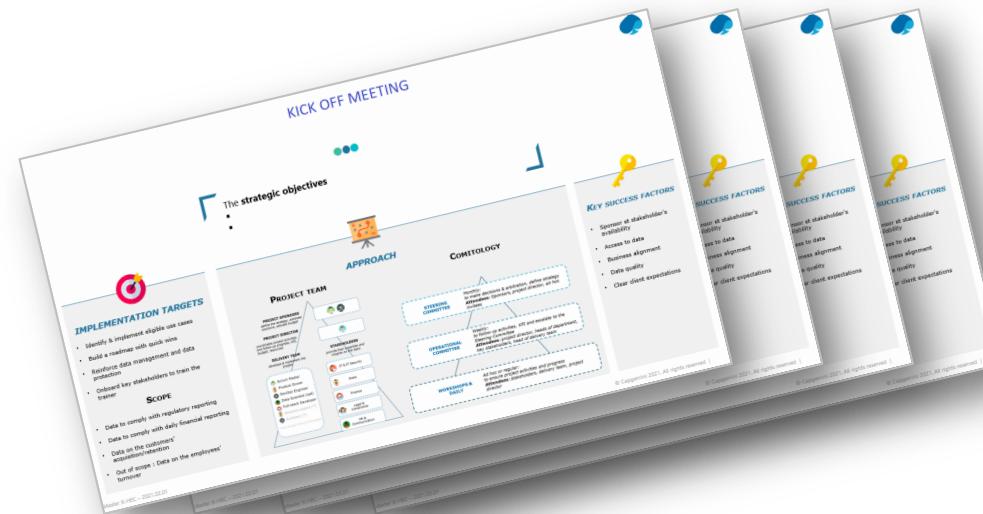


# For the next session



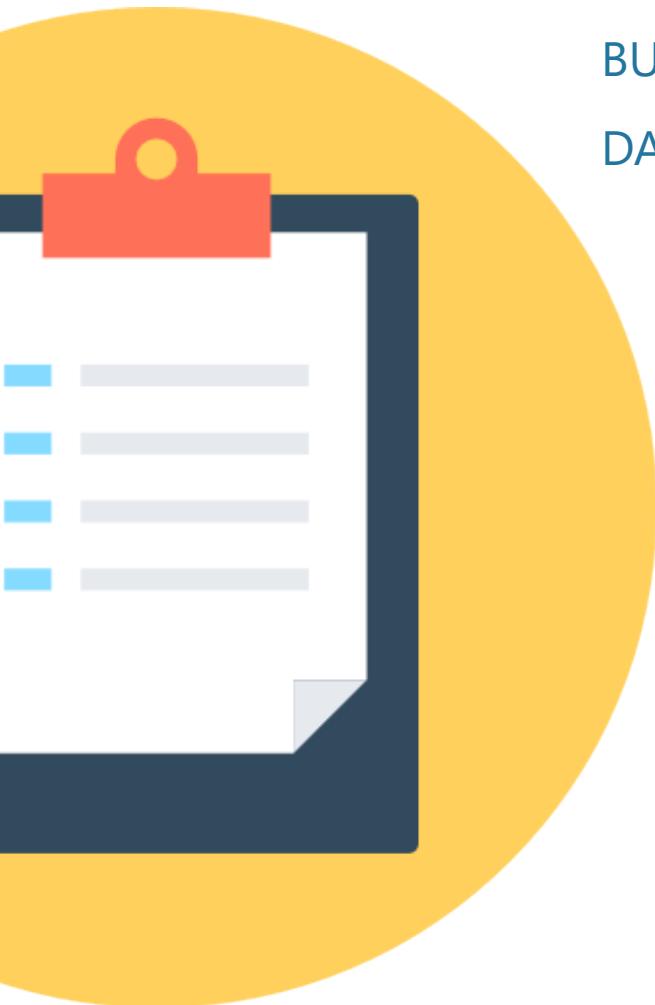
## INSTRUCTIONS

- Prepare the kick-off of your project using tools as set on slide 3
- Create sets of KPI (both evident and shadow) you may collect from available open sources
- Share your presentation with Capgemini team





# Agenda



BUSINESS4DATA

DATA4BUSINESS

1. **Data pipeline - Scrapping restitution**
2. Data Cleaning : introduction & basic processing
3. Stemming & Lemmatization
4. Text representation
5. Summary of the session



# Presentation of the achievements so far



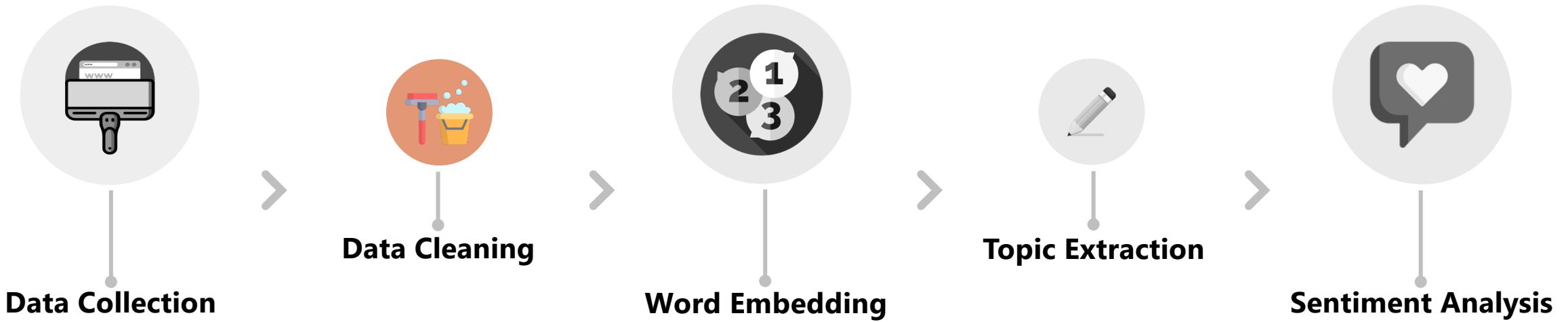
## Restitution of data scraped using Scrapy



- Pain Points
- Main takeaways
- How to improve ?

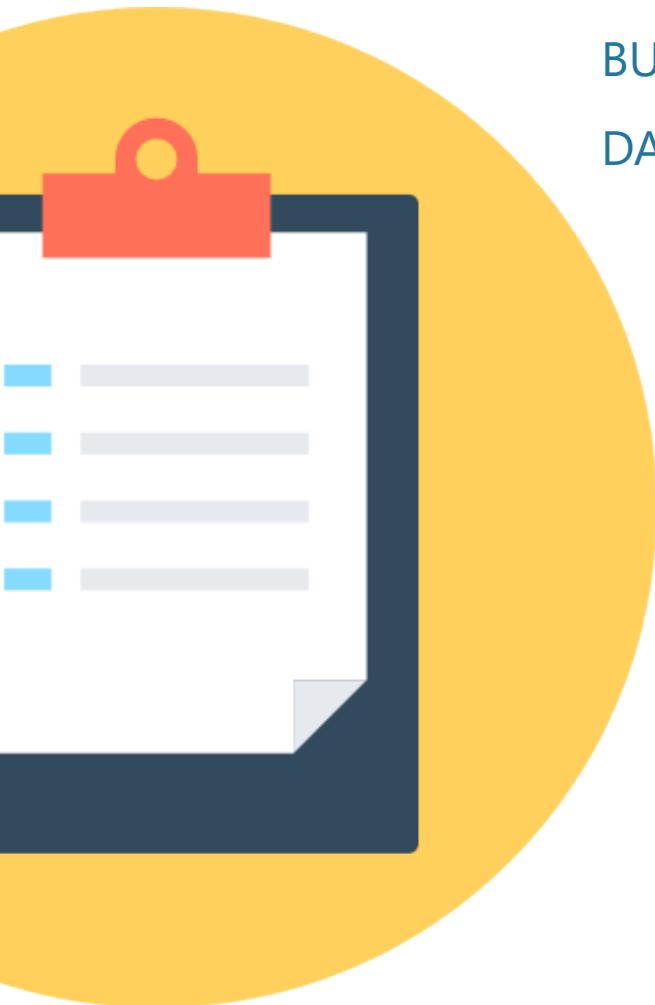


# Data pipeline





# Agenda



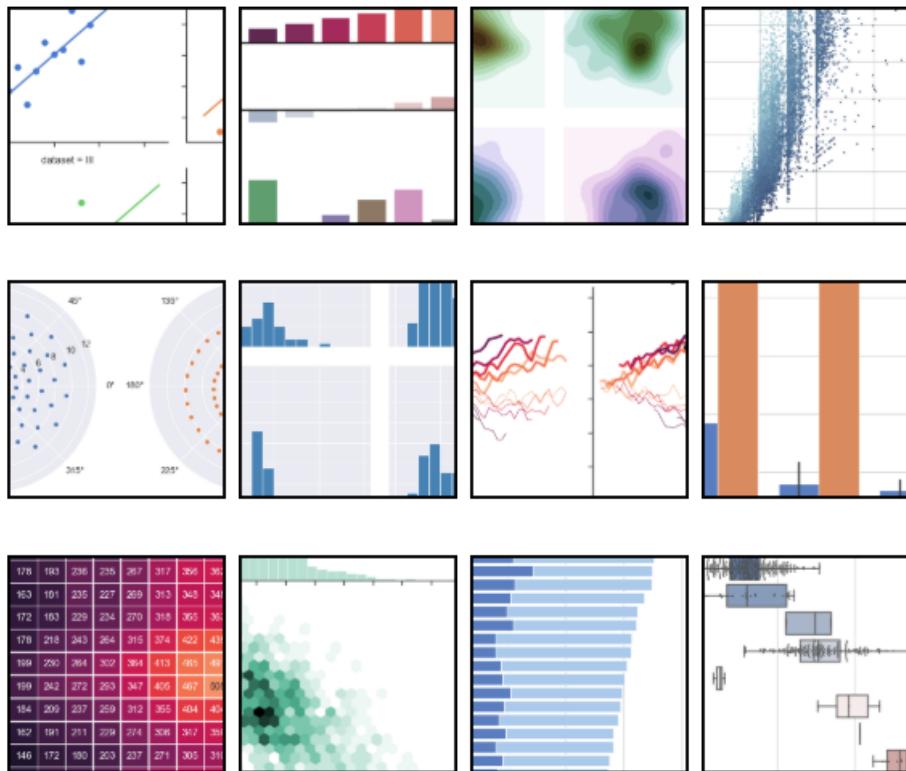
BUSINESS4DATA

DATA4BUSINESS

1. Data pipeline - Scrapping restitution
2. **Data Cleaning : introduction & basic processing**
3. Stemming & Lemmatization
4. Text representation
5. Summary of the session



# Exploratory Data Analysis



EDA is the process of using **statistical** tools and ideas to **examine** data in order to describe their main features

Exploring data begins with :

- **Examining each variable** by itself. Then move on to study the **relationships** among the variables.
  - Python methods like `df.describe()` and `df.info()`
  - Statistical tools
  - **Pandas-profiling**
- Making graphs **to visualize** and have a better comprehension of the dataset. There are some powerful libraries like :
  - Matplotlib
  - Seaborn



# EDA on Jupyter Notebook





# What is Natural Language Processing ?



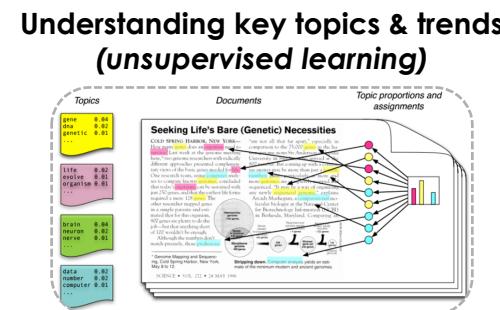
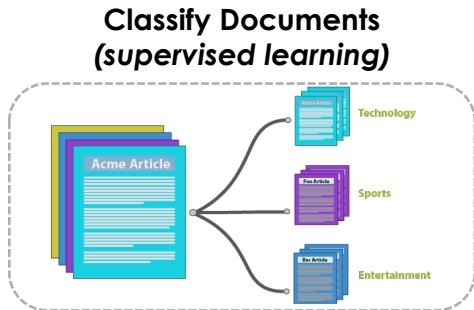
**Field of study focused on making sense of human written text**

It includes both linguistics and machine learning in order to make machines learn how to process large amount of natural language data

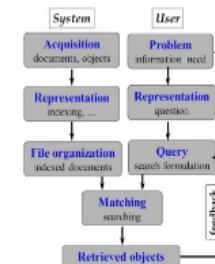
**You will learn the basics of NLP**

- Word Embedding
- Topic Extraction
- Sentiment Analysis

**NLP applications include**



**Find relevant information (information retrieval)**





# Pre-processing: Introduction



What processes should we apply on our data before anything  
to improve the quality of our analysis ?



```
[ '\n      Nous avons passé un excellent week-end, tout était bien entretenu\n      ',\n      "\n      Parc très agréable, difficile de s'y retrouver au début. Mais on s'y fait à la longue. Le cottage un\n      peu vieilli mais cela peut aller. Au niveau cuisine, il manque un peu de choses (une vrai poele, des couverts en plus\n      quand les premiers sont au lave vaisselle). Cadre très agréable dans la foret surtout avec les fortes chaleurs. La lo-\n      cation de la voiture électrique s'est bien passé et très vite (un peu cher sans doute pour la semaine !).",\n      '\n      Pas grand chose ne marche, ni l'internet, ni wifi (ce n'en seraient pas trop important s'il y avait du\n      réseau) Les cottage sont mal entretenus, j'ai trouvé au petit matin une puce et un tique dans les draps. ',\n      "\n      Moi je vais parler aujourd'hui du service commercial de center parc . J'ai fait une réservation pour\n      deux nuits sur le site de Center parcs au prix de 343 € quelques jours plus tard je vois sur vente-privee la même off-\n      re avec une nuit de plus pour 269€. Lors ce que j'appelle le service pour obtenir un geste commercial sachant que je\n      suis cliente fidèle il me répond: pas de chance pour vous et puis vous n'avez pas pris d'assurance annulation. Je le\n      urs explique que je veux pas annuler mais a titre commercial et car je suis cliente fidèle une compensation financière.\n      Je suis vraiment déçue\n      " ]
```



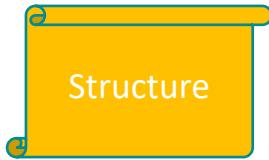
"[ Alice ' s Adventures in Wonderland by Lewis Carroll 1865 ] CHAPTER I . Down the Rabbit - Hole Alice was beginning  
to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped i  
nto the book her sister was reading , but it had no pictures or conversations in it , ' and what is the use of a book  
' thought Alice ' without pictures or conversation ? ' So she was considering in her own mind ( as well as she could  
, for the hot day made her feel very sleepy and stupid ), whether the pleasure of making a daisy - chain would be wor  
th the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her .  
There was nothing so VERY remarkable in that ; nor did Alice think it so VERY much out of the way to hear the Rabbit  
say to itself , ' Oh dear ! Oh dear ! I shall be late ! ' ( when she thought it over afterwards , it occurred to her t  
hat she ought to have wondered at this , but at the time it all "



# What are the preprocessing steps to get clean text data ?



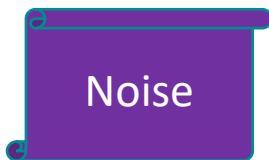
## Today's next steps



- **Transforming the text into a “corpus”**
- **Tokenizing**



- **Removing punctuation**
- Removing/replacing specific characters (highly recommended)
- Replacing accents (depending on language)



- **Removing stop words**
- **Lemmatization**
- **Stemming (Optional)**



All of these is part of what we call **Natural Language Processing**, which leads to **Natural Language Understanding**, which we will focus on during the next session.

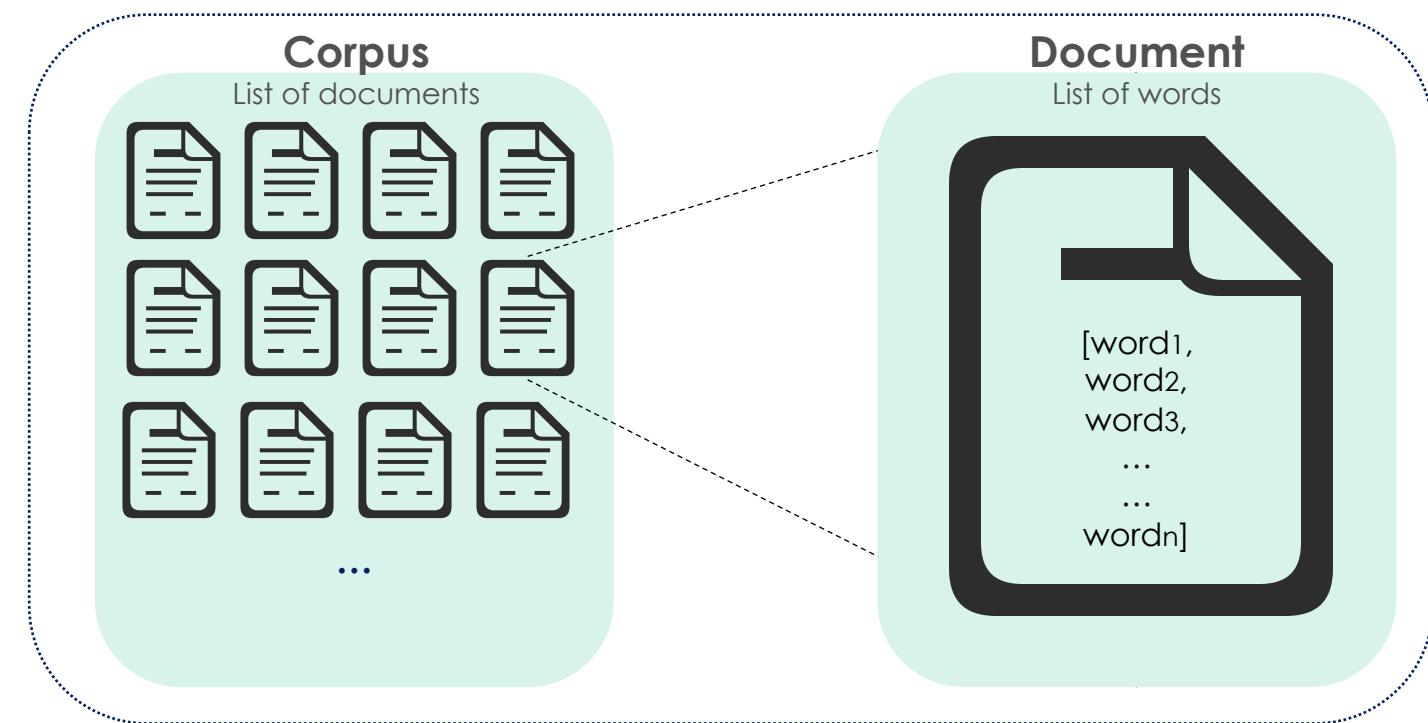


# Corpus of texts



## Before getting started, let's introduce a basic concept: The Corpus

- A corpus or text corpus is a **large and structured set of texts** from which we perform our analysis
- Within each corpus we will have separate documents, which might be articles, stories, or book volumes, each document is treated as a separate entity or record





# Tokenization



This method is used to **tokenize** the text, tokenization is the process of **breaking a stream of text** into words, phrases, symbols, or other meaningful elements called tokens

## 1st level: considering a sentence as a token

- Related to the structure of the text
- In a novel, dialogs should not be considered as the rest of the text

**Input:** ["How is it to be a Data Scientist? Olivier Auliard answered: It is super cool to be a Data Scientist"]

**Output:**

- ["How is it to be a Data Scientist",
- "Olivier Auliard answered",
- "It is super cool to be a Data Scientist"]

## 2nd level: considering the word as a token

- We split the document word by word

**Input:** " it is super cool to be a Data Scientist"

**Output:**

- ["It", "is", "super", "cool", "to", "be", "a", "Data", "Scientist"]

## 3rd level: N-grams

- We consider few words together
- Unigrams are tokens of one word, bi-grams are tokens of two, etc.

**Input (Bi-grams):** " it is super cool to be a Data Scientist"

**Output:**

- ["It is", "is super", "super cool", "cool to", "to be", "be a", "a Data", "Data Scientist"]



# Cleaning on Jupyter Notebook





# Hands-on 1: Pre-processing your data



**Load, explore and clean your dataset**





# Break

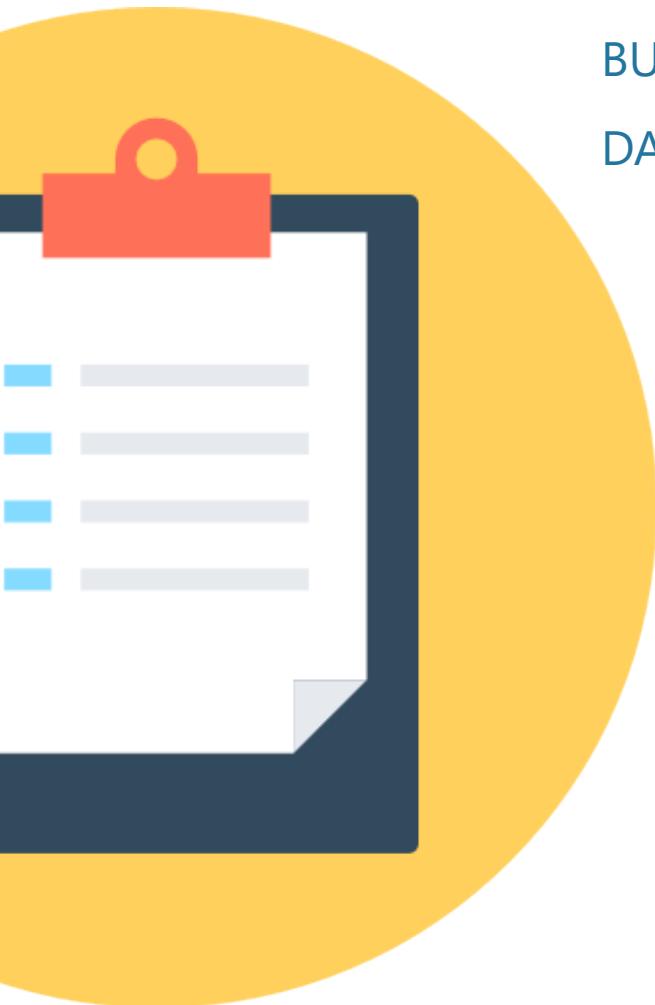


**Feel free to help yourself !**





# Agenda



BUSINESS4DATA

DATA4BUSINESS

1. Data pipeline - Scrapping restitution
2. Data Cleaning : introduction & basic processing
- 3. Stemming & Lemmatization**
4. Text representation
5. Summary of the session



# One possibility to clean text : Orthographic correction



Detecting words with low frequency of apparition and removing them



Making a comparaison with a dictionnary

Using specific algorithms like Levenshtein Distance Spelling Correction



Those methods are often very time and memory consuming





# Inflected language



In grammar, inflection is **the modification of a word to express different grammatical categories** such as tense, case, voice, aspect, person, number, and gender. An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change.

- “person”
- “persons”
- “person’s”
- “persons”



person

- “processing”
- “processes”
- “processed”



process

- “university”
- “universe”



univers

- “am”
- “are”
- “is”



be



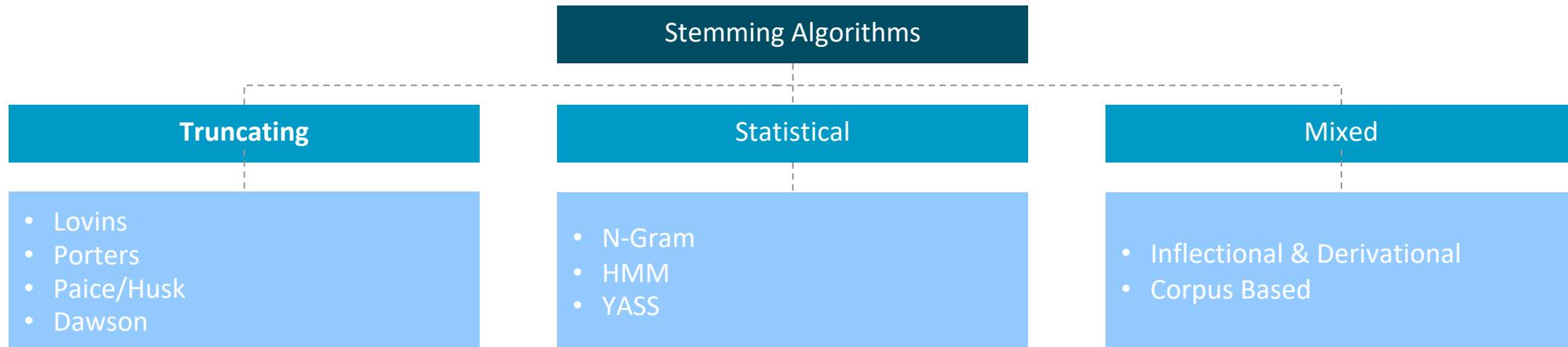
# Stemming



- **Stemming is the process of reducing inflection in words to their root** forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.
- By removing inflectional form of a word we focus on the **meaning**
- “Rude” process in the sense that it could remove the end of a word even if it is not an inflectional form



- Stemming algorithms can be classified into three groups: **truncating methods**, statistical methods, and mixed methods:



[https://www.researchgate.net/publication/284038938\\_A\\_Comparative\\_Study\\_of\\_Stemming\\_Algorithms](https://www.researchgate.net/publication/284038938_A_Comparative_Study_of_Stemming_Algorithms)



# Lemmatization



- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the **root word belongs to the language**.
- In Lemmatization root word is called **Lemma**.
- A trivial way to do **lemmatization** is by simple **dictionary lookup**. It is also the most common way.

## Example

« I have an important meeting today.  
Persons I'm meeting with always make the  
right decisions »

Lemmatizer

« I have an important meet today .  
Person I 'm meet with always make  
the right decision. »



Can you see some limits of lemmatization from this example?



# Trade-off between stemming & lemmatization



Stemming VS Lemmatization	
Stemming	Lemmatization
<ul style="list-style-type: none"><li>• produced by “stemmers”</li><li>• produces a word’s “stem”</li></ul>	<ul style="list-style-type: none"><li>• produced by “lemmatizers”</li><li>• produces a word’s “lemma”</li></ul>
<ul style="list-style-type: none"><li>• Better → Better/Bet</li><li>• am -&gt; am</li><li>• having -&gt; hav</li></ul>	<ul style="list-style-type: none"><li>• Better → Better/Good</li><li>• am → be</li><li>• having → have</li></ul>
<ul style="list-style-type: none"><li>✓ Fast</li><li>✓ Reduce vocabulary size</li></ul>	<ul style="list-style-type: none"><li>✓ Existing words</li><li>✓ Irregularities &amp; special cases</li></ul>
<ul style="list-style-type: none"><li>✗ Harsh method</li><li>✗ Non-existing word</li></ul>	<ul style="list-style-type: none"><li>✗ Slower</li><li>✗ May keep large vocabulary size</li></ul>

More about this: <http://stackoverflow.com/questions/17317418/stemmers-vs-lemmatizers>



# Let's do it with Python





# Hands-on 2: Stemming & Lemmatization



**Get ready for starting basic NLP : Stemming and  
lemmatization !**





# Agenda



BUSINESS4DATA

DATA4BUSINESS

1. Data pipeline - Scrapping restitution
2. Data Cleaning : introduction & basic processing
3. Stemming & Lemmatization
- 4. Text representation**
5. Summary of the session



# Bag-of-words (1/2)



**Challenge :** How to represent text with **numeric** values, in order to apply statistical methods on it ?

## Principle

- Consider a corpus of documents, each **document** is a text
- Each document contains a certain number of **words** for which we can do **tokenization**
- Each word, or group of words will be a **token** (monogram, bigram, ...)
- **Bag-of-word:** Representation of the document by a numerical vector containing the counts of each token



In this part of the course, we will use “**bag-of-words**” techniques, which means each word will be considered individually, **whatever its place in the sentence**.



# Bag-of-words (2/2)



## Description

- To get the most important topics in a collection of documents, a natural thing would be to look at the words and **compare their occurrences**.
- The **Document Term Matrix** synthetizes the words' occurrences in a collection of documents.
- Rows correspond to documents in the corpus
- Columns correspond to tokens present in the corpus (vocabulary)

## Example

- docA = 'I believe cats are better animals than dogs, I love cats !'
  - docB = 'I saw this movie named cats, it was quite bad'
  - docC = 'I went to the movies with catty last week'
  - docD = 'Catty has a gorgeous animal : a superb green parrot !'
- > [1,0,0,0,1,1,0,0 ...]

Document Term Matrix

	◆ believe ◆	◆ named ◆	◆ went ◆	◆ gorgeous ◆	◆ catty ◆	◆ love ◆	◆ dog ◆	◆ saw ◆	◆ week ◆	◆ quite ◆	◆ better ◆	◆ parrot ◆	◆ movie ◆	◆ superb ◆	◆ bad ◆	◆ animal ◆	...
docA	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	
docB	0	1	0	0	0	0	0	1	0	1	0	0	1	0	1	0	
docC	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	
docD	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1	



# TF-IDF



A score that better reflects how words are related to documents within a corpus

$$tfidf(t, d) = tf(t, d) \times idf(t)$$

## Term frequency

We have different possibilities to calculate the term frequency. The most frequent are raw count and **term frequency**.

## Inverse document frequency

A measure of how much information the word provides, i.e., if it's common or rare across all documents. It diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Formula of Inverse Document Frequency:

$$idf(t) = \log \left| \frac{n}{\{d \in D : t \in d\}} \right|$$

→ Can you comment this formula ? Why do we need a log ?

With  $n$  the total number of documents,

$D$  the set of all documents and  $d$  a given document in  $D$ ,

$t$  is a given term.

Thus,  $\{d \in D : t \in d\}$  is the number of documents in which the term  $t$  appears (at least once).

## Remarks

They are various formulas for tf and idf, see on [Wikipedia](#)

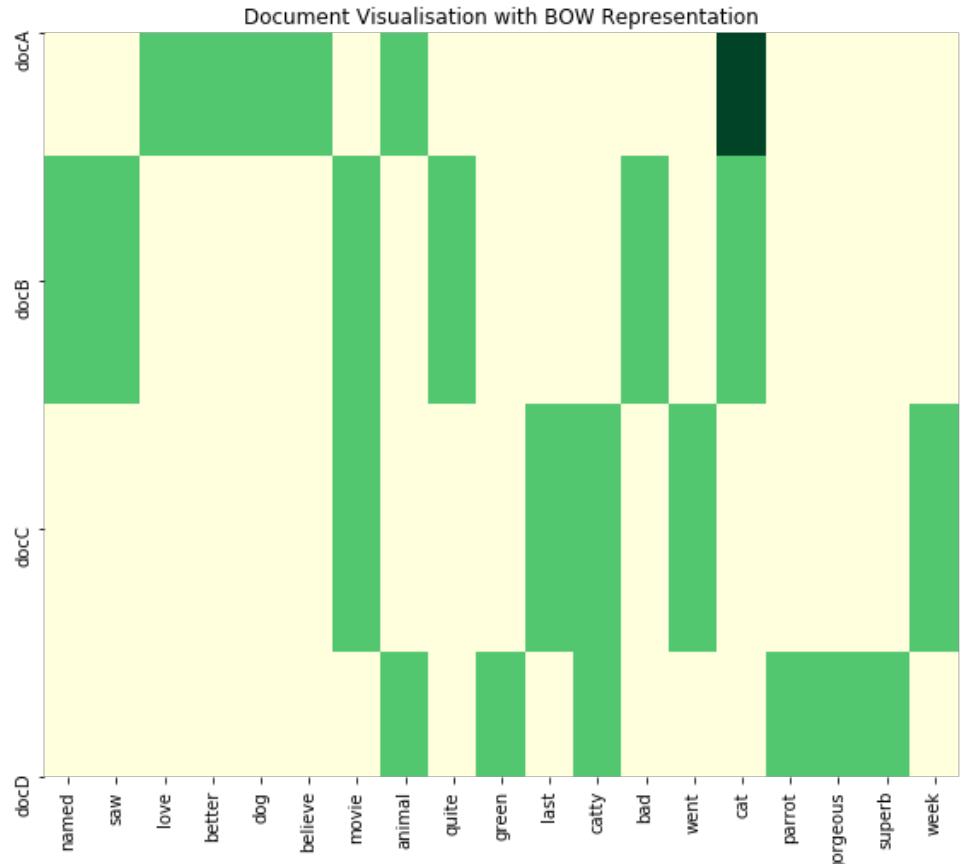
Choices for the value of tf ([Wikipédia](#))

weighting scheme	tf weight
binary	0, 1
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

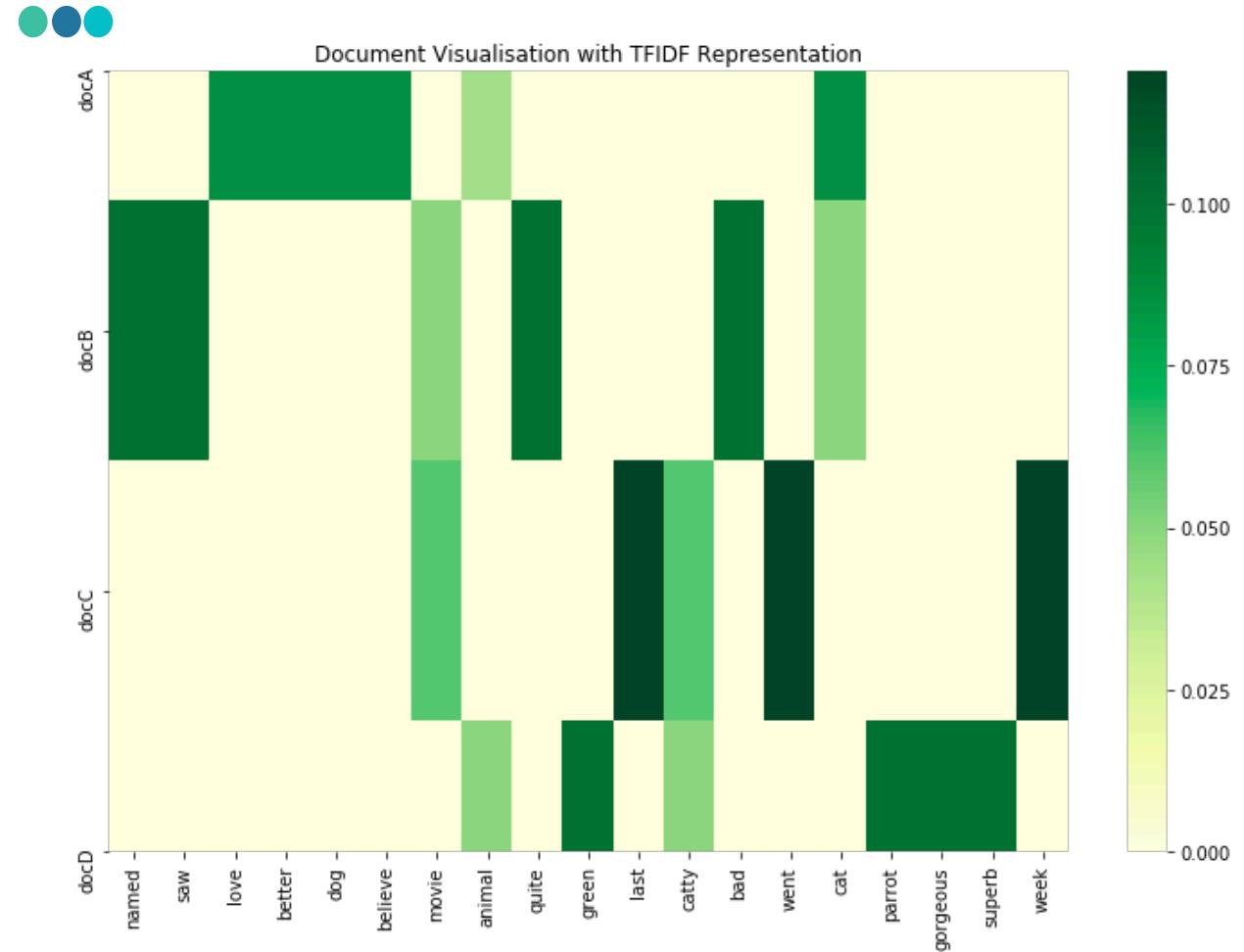


# BOW vs TF-IDF

**BOW**



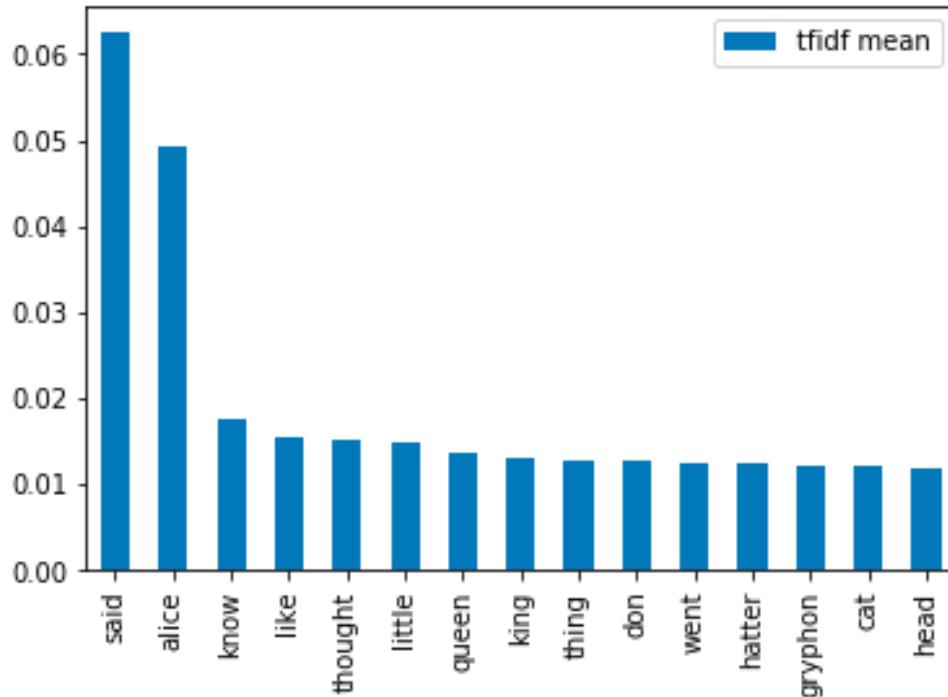
**TFIDF**



What are the pro's and con's of each representation ?



# Interpret the TF-IDF matrix



## How can we improve our analysis ?



# Let's do it with Python !





## Hands-on 3: BOW - TFIDF



**Use the notebook provided to build your own TF-IDF matrix from scratch.**



10'

**Then use a Python library (TfidfVectorizer).**

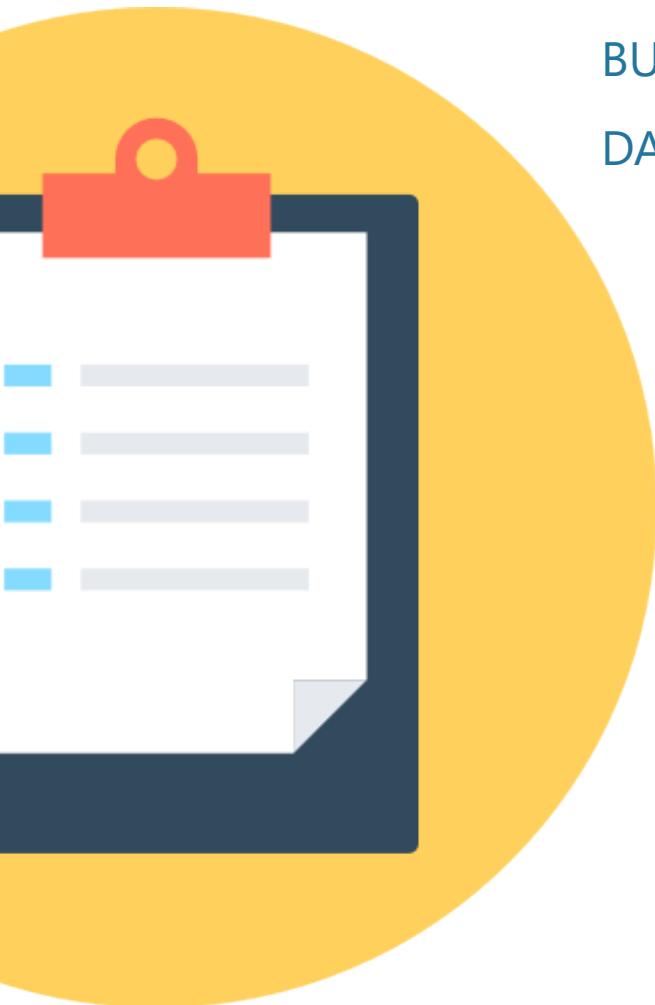
**What are the most relevant terms in the comments of your scrapped reviews**

?





# Agenda



BUSINESS4DATA

DATA4BUSINESS

1. Data pipeline - Scrapping restitution
2. Data Cleaning : introduction & basic processing
3. Stemming & Lemmatization
4. Text representation
- 5. Summary of the session**



# Today we learnt



## Cleaning steps

- Convert documents into a corpus
- Adjust characters:
  - Convert to lowercase
  - Remove special characters, punctuation, accents, double spaces
- CAREFUL: Some punctuation sets can be useful to keep (ex. : “;”)
- Retreat some words: spell-checking, stemming / lemmatizing, remove stop-words & tokenization



## Approach & organization of a data science consulting project

- Represent the corpus as a Document Term Matrix
- Identify and select terms that are important
- Tools: bag-of-words, TF-IDF, wordcloud



# Work for next week



## Instructions for Sunday 23<sup>rd</sup> !

- Be sure that you have applied every steps of text processing to your reviews
- Create a TF-IDF Matrix with all the reviews you scrapped on the web, and find the best way to represent it (WordCloud ?)
- Bonus reward : build 2 functions :
  - One that takes a corpus of raw text and creates a new corpus with cleaned and lemmatize text.
  - One that takes a corpus (or a dataframe text column) and creates a Wordcloud from it.
- Create sets of KPI (both evident and shadow) you may collect from available open information
- The idea is to anticipate what you'll present on the fourth session for the client meeting simulation

Our email addresses :

- [mariam.ikken@capgemini.com](mailto:mariam.ikken@capgemini.com)
- [svetlana.ollivier@capgemini.com](mailto:svetlana.ollivier@capgemini.com)
- [francois.lemeille@capgemini.com](mailto:francois.lemeille@capgemini.com)
- [ismail.mebsout@capgemini.com](mailto:ismail.mebsout@capgemini.com)

Feel free to contact us by email or Slack !





# Course evaluation



**Did you like that second course ? It's time to share your  
feedbacks !**





**THANK YOU!**