



# DATA ANALYSIS

Aim: getting the tweets of people talking about their symptoms.

## Preprocessing:

- Remove retweets (half of the dataset)
- Remove mentions (@mention) and url ([url])
- Dictionary of symptoms

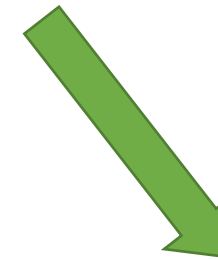
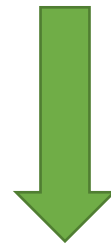
### Symptoms:

- Cough (toux, toussé)
- Fever (fièvre)
- Sore throat (maux de gorge, mal à la gorge)
- Headache (mal de tête, mal à la tête, mal de crâne)
- Breathing difficulties (difficultés à respirer, difficultés respiratoires, mal à respirer)
- Loss of taste and smell (perte du goût, perte de l'odorat)
- Symptom (symptôme)

**HANDS ON!** [Notebook 3 - Data analysis and visualization](#)

# CITIZEN SCIENCE

Issue: many **false positives** in our tweets mentioning symptoms



## Symptom before

« @mention J'ai eu beaucoup de symptômes.. fin Février debut Mars, fin Mars (essoufflement quoi que je fasse) je suis donc allée chez le medecin, "c'est le stress" 🤔 je me reconnais dans ce Tread..' »

## NOT a symptom

« Jcrois j'ai trop rigoler ojd j'ai mal à la gorge »

## General news

« En cas de fièvre, privilégiez le paracétamol. En effet, les anti-inflammatoires comme l'ibuprofène, de part leur mécanisme d'action, peuvent aggraver l'infection ! 🦠 En cas de doute, demandez conseil à un professionnel de #santé 🧑🏻‍⚕️📄🧑🏻‍⚕️📄 #COVID—19 [url] »

# CITIZEN SCIENCE

## Does this tweet contain self-reported COVID19 symptoms?

Symptoms can be about the person tweeting or someone they are referring to directly. They should be current and not about past events. If you are unsure about your annotation, hit the "skip" button."

@mention Diarrhée, perte d'odorat, orteils gonflés...  
les symptômes du Covid-19 se multiplient [url] via  
@mention

May 12, 2020

Yes

No

skip

## What is this about?

In this study, we aim to create **a model of how tweets about self-reported COVID19 symptoms can help predict upcoming pandemic waves**, and more generally the rise and fall of the disease. To that end, we crawled public tweets from the Paris region filtered by symptoms keywords, and plotted them in time (see the graph below).

However, this filtering is very crude, e.g people don't only tweet about symptoms when they are currently falling sick, but also about that one time a year ago when they fell sick, or when talking about the general news.

To filter out such false-positives we need **your** help! Which of these tweets are describing an acute symptom and which ones don't? Your contribution will make a direct impact!

If you want to learn more about the people behind this project you can [visit our About page](#).

Citizen science platform: <https://covid-twitter.thecommons.science/>

# CLASSIFICATION

***Based on the 7,300 annotations from the platform, can we predict which tweets are true positive?***

## **Steps:**

- Clean the labels: only consider tweets that have been labelled similarly several times
- Build the features from the textual data
- Train a classifier
- Define metrics to evaluate the results
- Refine the classifier with hyperparameters tuning and cross-validation
- Plot the corrected curve

**HANDS ON! [Notebook 4 - Classification](#)**

# THANK YOU!

Any questions, clarifications?

→ [naila.elhaouari@cri-paris.org](mailto:naila.elhaouari@cri-paris.org)