Par David Dubois, PhD Kellogg School of Management, INSEAD Associate Professor of Marketing & Cornelius Grupp Fellow in Digital Analytics for Consumer Behaviour (https://sg.linkedin.com/in/profdaviddubois) et Somya Gupta, Bachelor Student at Vellore Institute of Technology (https://in.linkedin.com/in/somya-gupta-309a46183).

In the race to fight Covid19, data is perhaps the most crucial asset. In the short-term, it can dramatically help governments and other organisations assess the progression of the pandemic and adjust their logistical responses (e.g., planning hospital beds, deciding on social distancing measures etc..). In the medium and long run, data may also help assess populations' general feelings and emotions. This data may, in turn, help identify and tackle potential negative byproducts of the pandemic (e.g., anxiety, depressions).

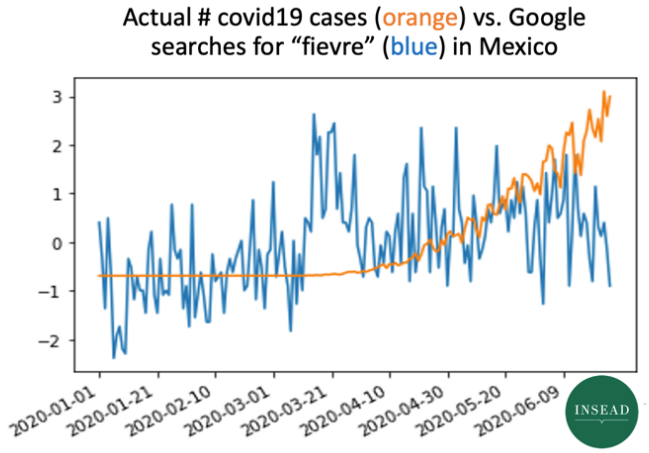
Since the SARS outbreak of 2003, global data capabilities have vastly improved. First, the increase in internet and mobile penetration has multiplied the available data. Organisations such as the World Health Organisation easily make this data one click away from researchers (e.g.,the WHO's dashboard). Second, improvements in data processing and machine learning have augmented our ability to process, visualise and analyse data. Governments across the world have started to take advantage of these possibilities through tracking apps enabling close contact detection – for instance in Taiwan, Korea or more recently Germany. In fact, some attribute Taiwan or Korea's successes in eradicating the pandemic to their effective use of mobile applications.

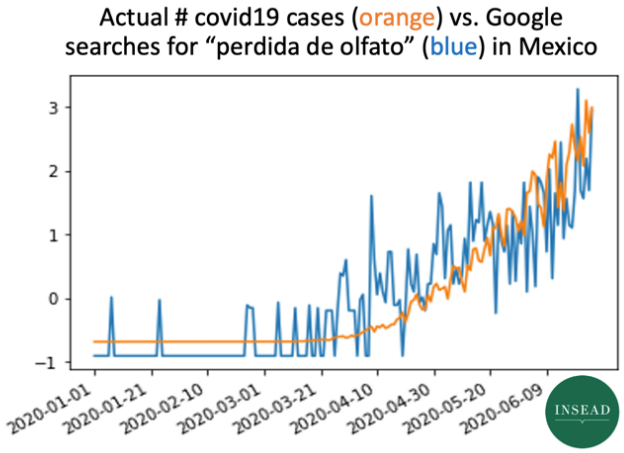
Despite these efforts, mapping the progression of the pandemic has proven a difficult task. India, Brazil, Mexico among others are still struggling to contain the epidemics while others are facing the possibility of a resurgence. To be most effective, scientists need to turn to data that (1) reflects weak signals (2) is live (3) reflects behaviours of entire populations and (4) can be easily combined with other types of data (e.g. actual progression of cases).

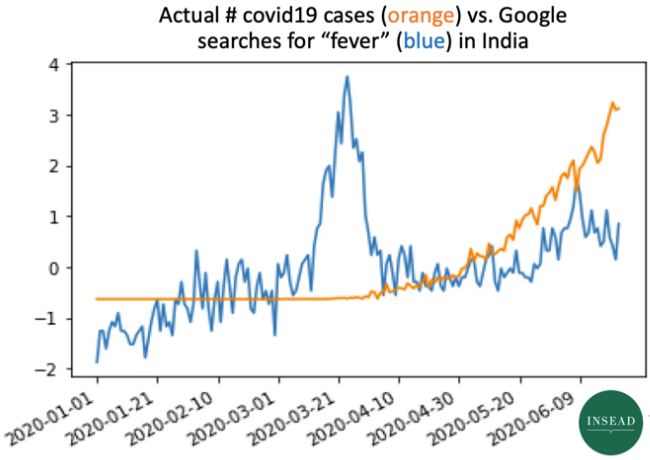
One such data source is search data. Google alone processes 3.5 billion searches per day – while other players fulfil similar functions in other countries (e.g. Baidu in China, Naver in South Korea etc.). Aggregate search behaviours open a door into what individuals think and desire at any given time – from their brand preferences to health concerns and inquiries about emerging symptoms. How can search data help tackle Covid19?

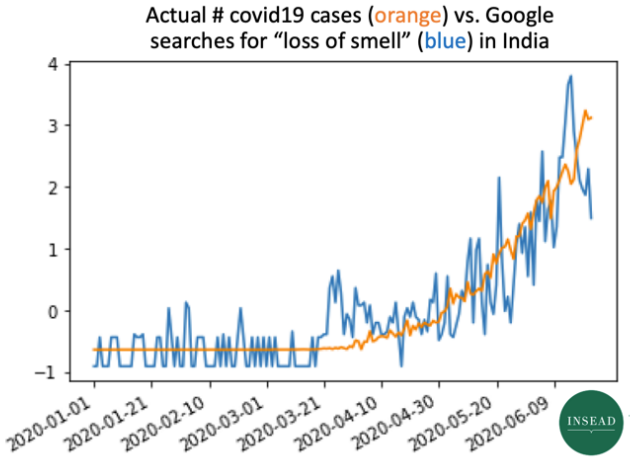
Classic Covid19 symptoms – common to similar viruses – include fever, dry cough and tiredness. Less common symptoms include aches and loss of taste or smell, among others. In this article, we decided for purposes of demonstration to focus on fever (a likely early symptom, associated with mild cases) and loss of smell (a seemingly more serious symptom associated with severe cases). We investigated the link between searches for fever, loss of

smell, and the actual number of cases. We predicted that searches for fever would act as an early predictor while searches for loss of smell would take place closer to the actual rise in number of cases. The data indeed supports this hypothesis (Figures below).









Together, these graphs and the underlying methodology -which can be easily applied for other countries or regions – is noteworthy on two fronts:

First, the lag between the peaks of searches for "loss of smell" and of COVID 19 cases overtime is smaller than that between the peaks of searches for "fever" and of COVID 19 cases overtime. This is likely due to loss of smell symptoms taking place later than fever symptoms. This difference shows how researchers and policy-makers may study different keywords along the patient journey to predict the rise in cases in the near future ("fever" searches) and monitor actual number of searches during the pandemic (i.e., "loss of smell" searches) – and potentially decide on second or local lockdowns.

Second, searches for loss of smell appear more sporadic among the public and may thus be driven by the outbursts of COVID 19 cases overtime. In fact, “loss of smell” may be a cleaner predictor because more tightly associated with the number of COVID 19 cases.

Moving forward, we encourage data scientists to complement survey data with other data such as online searches – reducing potential biases while increasing the transparency and timeliness of predictions. We believe this multi-method approach may complement survey data and provide both early and better monitoring of the pandemic. Practically speaking, using search data involves a three steps process: (1) keyword monitoring (2) analyses and graph production and (3) cross-checking and validation. Notably, several factors may alter the reliability of search data – for instance, differences in digital literacy across countries, as well as different keywords – over time or space. For instance, among English-speaking nations, while some may primarily search for the noun “loss of smell,” others may favor the verb use “can’t smell.” One also remember that Google’s 2008 initiative to predict flue epidemics – Google Flue – lacked reliability overtime due to year-on-year variations in symptoms, and thus of keywords searched.

We live in an increasingly complicated world in which establishing causality and certainty about when and why events such as pandemic occurare more and more complex. At the same time, powerful new tools analytics such as search represent goldmines for researchers willing to use them to uncover weak signals and combine them with other data such as search or existing datasets to uncover the dynamics of our complex global world.

- **Please cite:** Dubois, David and Somya Gupta 2020, “When and How Combining Data Analytics can Help Tackle the Pandemic,” Datacovid and INSEAD Covid-19 Response” (see also https://www.insead.edu/covid-19/expertise)
- **Supplementary material :** to access the codes for researchers who would like to build on/replicate the findings, click here : FINAL CODES WITH COMMENTS