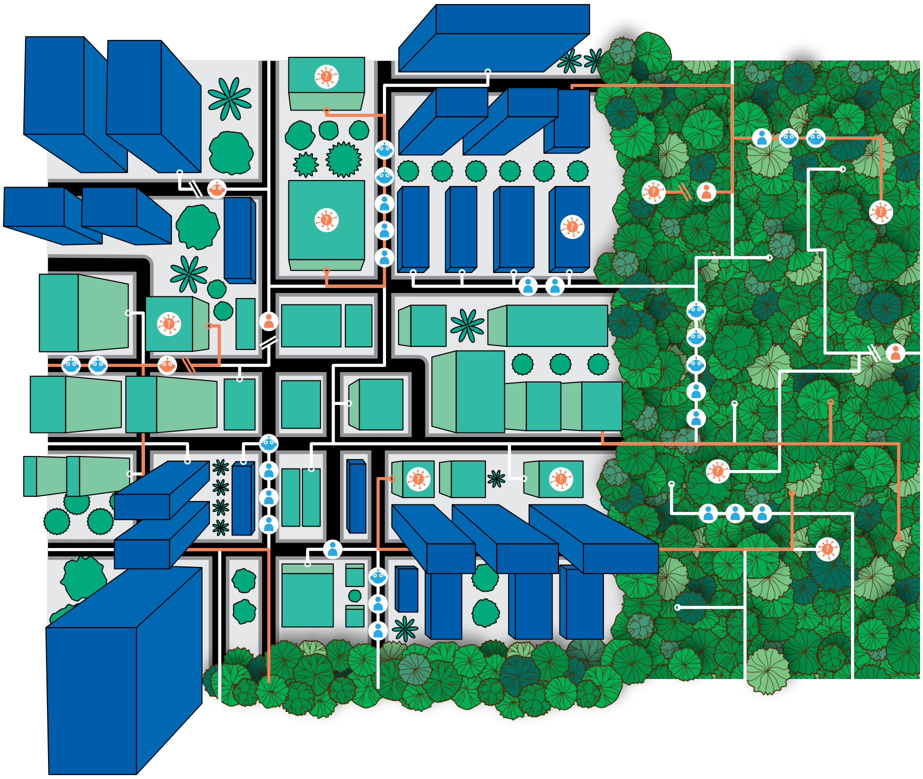
Check for updates

# A time-critical crowdsourced computational search for the origins of COVID-19

To the Editor — On 26 May, President Joe Biden called on the US Intelligence Community to redouble efforts to collect and analyse information on the origins of coronavirus disease 2019 (COVID-19), and report back to him in 90 days. After more than 18 months of intensive global investigations, the deadline will put pressure on a task that is proving to be substantially more challenging to solve than the origins of other threats[1].

I want to offer some technical considerations on the vast — but not insurmountable — complexity of the task ahead. My advice builds on a decade of experience leading teams that participated in[2], designed[3] and analysed[4] challenges involving the time-critical search for hard-to-find information entities. I also borrow insights from my field of study, network science, which has tackled the theoretical[5] and empirical[6,7] aspects of searching for rare information spreading on a network. A successful investigation will, I believe, benefit from implementing five principles: incentive structures, transparency, unbiased search, crowdsourcing and human–machine computational sense-making.



Credit: DTP+Graphics, Max Planck Institute for Human Development

## Incentive structures

It is improbable that one or a few agencies alone can uncover enough evidence to determine which origin scenario is most likely. I expect this effort to require a cooperative and coordinated collective search for evidence by a diverse range of people: national security personnel, scientists, medical practitioners, journalists and, crucially, citizen scientists. Such a collaboration requires the design and deployment of multi-level symmetrical incentive mechanisms[8] that connect and coordinate the entities performing the search, while efficiently allocating rewards for finding and sharing information. These incentives should proportionally promote the search for information and the recruitment of others that could help the investigators in new unexplored domains, balancing exploration (increasing geographical coverage) and exploitation (delving into a particular source of information). Such rewards should be carefully designed and could take the form of monetary and reputational incentives[9], including prestige emblems, medals and other certifications for valued work that governments can offer. Without these incentives, the human power available for the search will be notably smaller and confined to only a narrow set of experts with potentially conflicting motivations. The absence of carefully designed incentives can also lead to dishonesty, such as hoarding information, providing counterfeit data, and even data theft and cybernetic sabotage.

## Transparency and unbiased search

It may seem counterintuitive for an effort led by a national security agency to embrace the disclosure of information on the origins of the virus in real time, but not doing so will negatively impact the credibility of the investigation, rendering it useless at best or creating complex geopolitical puzzles at worst. It is thus essential to document each piece of evidence transparently, including how and who uncovered it, and what chain of evidence was followed. Notably, it is crucial to document negative evidence and negative results, including attempts to uncover specific evidence that proved false or inconclusive. Only by doing so, can the world be assured of an unbiased search.

## Crowdsourcing

All pieces of evidence should be available for crowdsourcing, from other entities and citizen investigators. Each item of evidence involved should thus allow participants to weigh in by labelling, annotating and enriching it, and it should be accessible and retrievable so that it can be verified independently. In return, the incentive structure should prompt citizen investigators to share back the observational or computational methods they used to perform verification and make their novel confirmatory datasets or code available for reproducibility. Not being able to openly verify a particular piece of evidence would not, of course, prove that this information is false or inaccurate. However, the accumulation of a range of independent evaluations — whether successful, unsuccessful or inconclusive — together with enriched clues generated along the way, will provide the necessary ingredients for the final step.

## Human–machine computational sense-making

The crowdsourced process will generate vast amounts of secondary information, considerably more than the direct, on-the-ground evidence itself. Thus, a collective[10] computational[11] sense-making approach becomes necessary to form coherent statistical models of the available data. These models can then be used to filter, to iteratively predict potential emergences and early evolution of the virus, and dynamically assign a likelihood to each possible hypothesis, which can be used to rank and prioritize the next steps in the investigation. Ideally, we should deploy an architecture that allows human–machine hybrid team predictions[12], yielding an evolving, scalable and ever more accurate set of estimations by experts, citizens and machine learning algorithms.

## Moving forward

We are still in the early days of the search for the origins of the virus. Quickly clarifying its origins is essential for scientific and geopolitical reasons, but the search is likely to require more than 90 days. By immediately implementing these five principles, some of which have recently been highlighted by others[13], we can though establish a solid foundation for the collective effort required. Delaying such planning, or ignoring altogether the complexities involved, is a risk we cannot afford. Investigating a twenty-first-century calamity requires a twenty-first-century technological response. ❐

Manuel Cebrian ✉
*Max Planck Institute for Human Development, Berlin, Germany.*
✉*e-mail: cebrian@mpib-berlin.mpg.de*

Published online: 12 July 2021
https://doi.org/10.1038/s41928-021-00619-4

### References

1. Cyranoski, D. *Nature* **552**, 15–16 (2017).
2. Pickard, G. et al. *Science* **334**, 509–512 (2011).
3. Pescetelli, N., Cebrian, M. & Rahwan, I. *Computer* **53**, 49–58 (2020).
4. Rutherford, A. et al. *Proc. Natl Acad. Sci. USA* **110**, 6281–6286 (2013).
5. Kleinberg, J. M. *Nature* **406**, 845 (2000).
6. Dodds, P. S., Muhamad, R. & Watts, D. J. *Science* **301**, 827–829 (2003).
7. Milgram, S. *Psychol. Today* **2**, 60–67 (1967).
8. Cebrian, M., Coviello, L., Vattani, A. & Voulgaris, P. *STOC '12: Proc. 44th Annu. ACM Symp. Theory of Computing* 775–788 (2012); https://doi.org/10.1145/2213977.2214047
9. Easley, D. & Ghosh, A. *ACM Trans. Econom. Computation* **4**, 16 (2016).
10. Tetlock, P. & Gardner, D. *Superforecasting: The Art and Science of Prediction* (Random House, 2015).
11. Malone, T. W. *Superminds: The Surprising Power of People and Computers Thinking Together* (Oneworld Publications, 2018).
12. Abeliuk, A., Benjamin, D. M., Morstatter, F. & Galstyan, A. *Sci. Rep.* **10**, 15940 (2020).
13. Bloom, J. D. et al. *Science* **372**, 694 (2021).

### Competing interests

The author declares no competing interests.