ELEPHANT

IN THE LAB

OPINION

Science Without a Question

Title	Science Without a Question
Author	Fabian Stephany ¹
Author affiliation	¹ Oxford Internet Institute (OII)
Author bio	<p>Fabian is an Economist and Social Data Scientist researching the Oxford Internet Institute (OII), University of Oxford, and a Research Affiliate at the Humboldt Institute for Internet and Society in Berlin. With the iLabour project at the OII, he studies the global dynamics of Online Labour Markets. Fabian is interested in the application of computational social science in fields like digital skills, migration, innovation, and e-governance. He is a co-founder of the cross-country social data science research collaboration oxford.berlin. Fabian holds a PhD and degrees in Economics and Social Sciences from different European institutions, including Università Bocconi Milan and University of Cambridge. He has been working in the private sector and for various actors in the international policy landscape, such as the United Nations Development Programme, the World Bank or the OECD in Paris.</p>
Author social links	Fabian Stephany: Twitter
Date published	16 July 2020
DOI	10.5281/zenodo.3946747
Cite as (APA)	Stephany, F. (2020). Science Without a Question. <i>Elephant in the Lab</i> . DOI: 10.5281/zenodo.3946747

In the beginning, there was a question. This is how every science story seems to start. But what if it doesn't? Does research always require a question, or even a hypothesis? Is it possible to conduct science without a question? Can you "answer" a question that you haven't posed before?

My story started when social life had come to a halt. In late March 2020, the Coronavirus had paused public life and forced society into a global lockdown. Just as stores closed and travel plans were cancelled, our team, a group of Oxford and Berlin based researchers, had to pull the plug from a project we had been vigorously planning for the past months. Students from both university cities were supposed to meet for a data science competition on Sustainable Development Goals in Berlin. With datasets prepared, speakers invited, and rooms booked, everything had now been called off within a single day. Disappointed and frustrated, we arranged a team call to see how, nonetheless, we could turn slightly empty calendars and high-flying ambitions into something valuable. We asked ourselves, how could we, as social scientists and data geeks, contribute to making the corona crisis a little less messy?

As we observed recession-fighting governments around the world starting to prepare economic benefit packages of unprecedented volumes, we began to wonder: how do politicians actually know which industry is most in need of financial support? We explored options on how to measure the degree to which businesses and industries were impacted by the virus. It was clear from the beginning that our methods needed to be agile, since reaction time was limited, and that therefore real-time internet data might play a relevant role. In the end, we opted for a data-mining approach to analyse online texts from company risk statements [issued to the U.S. Securities and Exchange Commission \(SEC\)](#). Large firms, covering a third of all U.S. employees, are legally obliged to truthfully report their business outlook to the SEC. The magnitude and topical relationship in which companies expressed concerns related to the coronavirus were condensed in an index, which we made [interactively available to the public](#). The CoRisk-Index was born. It was first economic indicator measuring business risk assessments related to Covid-19. A later so-called "[indicator of business fear](#)".

Our first findings indicated that some industries were extremely worried about Covid-19, while others showed only little fear of the virus and its repercussions. Over the course of the crisis, we were able to see how business concerns switched from travel-related issues to uncertainties about supply chains and then to demand shocks. We received great feedback from researchers in macroeconomics who began using our data for their model calibrations on unemployment forecasting. Other colleagues and the media started to get interested in our work. [US](#) and [German](#) newspapers covered the story. We were invited to seminars and brown bag talks and asked to explain our research in [podcasts](#) and [interviews](#). Within a couple of weeks, our CoRisk-Index had itself gone viral. It was in these exciting weeks of feverish work, I had a somewhat mind-changing conversation with a senior colleague of mine. Visibly impressed by the speed and outreach of our project, he nonetheless started pressing where it hurt: "What is the

research question that you are answering with your index?”, he asked, “Where is your hypothesis?”. I had no clue. The cautious remark, all of the sudden, made me question the whole endeavour. We could see how the CoRisk “fever curve” looked, but did we know why? Or to put things in a bigger perspective: Did we really try to conduct research without a research question?

The Big Data Paradigm

One of the first papers that had cited our work was [a meta-level analysis](#) collecting all sorts of research articles that had investigated the repercussion of the coronavirus. Many of the pieces were medical studies that seized the opportunity of what is arguably the largest health experiment in the 21st century. Some of the papers, like ours, were situated in social science. All studies made use of data science methods by scanning news articles, analysing X-ray images or clustering Twitter mentions. The possibility to harvest information about human behaviour on a large scale with methods of computational statistics, had enabled dozens of studies that were rather more interested in revealing patterns than answering questions, it seemed. The “theorylessness” of these studies, frequently presented as works of [nowcasting](#), sparked the strange feeling of a déjà-vu.

It took me a bit of mind-wandering to find the origin of my odd sensation; in a [2008 Wired article](#) by Chris Anderson. At that time editor in chief of the magazine, Anderson wrote a provocative piece titled “The End of Theory”. He was referring to the new paradigm of Big Data in which methods of machine learning could analyse data in a way that the emerging patterns would tell us more about the world than theory-guided experiments, carefully crafted by domain experts, ever could. Today, this notion, notably established by tech giants like Google and Facebook, has not only stirred popular imagination but also initiated a new epistemological paradigm. The new paradigm tells us that machine intelligence replaces theories and hypotheses; it reveals unknown patterns and trends to us without requiring any prior knowledge. Or as Anderson concludes, “Petabytes (of data) allow us to say: ‘Correlation is enough.’”

While in the realm of techy advertisement, correlation might suffice, for doing science, correlation alone does not suffice. On the [Ladder of Causation](#), a conceptual reflection on cause and effect introduced by the statistician, Professor Judea Pearl, correlation alone does not allow us to exceed recognising simple associations between things. As much as correlation tells us that A happens frequently together with B, it reveals nothing about WHY this might be the case. Correlation - to put it with the words of the philosopher [Byung-Chul Han](#) - is in fact “the most primitive form of knowledge”. It is wisdomless, as it tells us nothing about cause and effect. Nothing is understood with plain correlation. Which left me with nothing but the nagging question in my head: Was our research meaningless, after all?

Reason Upon Uncertainty

In the midst of sleep-depriving self-doubt, I remembered an anecdote that a friend and colleague of mine was once told in a seminar by his PhD supervisor, [Karl Milford](#). In the footsteps of his own doctoral supervisor, Karl Popper, Milford used a trivial example to show the students the power of implicit theory. He asked one student to step at the auditorium window and count all red cars that would pass by within the next minute. When later asked to judge how “scientific” this task was, the seminar turned oddly quiet. While the counting car experiment seems mundane at first, as the professor explained, the observer requires a whole set of theoretical concepts. What is a car? Does a motorbike qualify as a car? How about a truck? What does the colour red mean? When does orange stop being yellow and starts being red? What if a car stops right in front of the building? And so on. The punch-line was that no empirical science, even the seemingly most trivial type, can exist without theory and questions. These questions might be invisible at first, but it does not mean that they do not exist. They are implicitly woven into the fabric of the research. Do risk statements contain information about business concerns? Do industries with higher risks have more unemployment? Can the findings of our work be generalised to other economic crises? As I remembered this anecdote on the hidden nature of implicit theory, the foggy mist of unconsciousness lifted and all the questions of our own work appeared to me, one by one.

In the end, science is not about finding patterns, it is about explaining them. However, finding patterns is certainly part of the scientific process based on implicit theory. The current fast-changing crisis has shown researchers how fragile this implicit knowledge is. Scientific certainties can change from day to day. The process of falsification is accelerated. But we could also welcome these uncertain times, as my colleague [Benedikt Fecher](#) suggests, for they invite researchers to sketch a utopian state of how we believe the world might be. We are then asked to falsify this utopia via “prospective falsificationism” enabling society to arrive at a better state. We, as researchers, should not be afraid of drawing this utopia that could require adjustment at any time in light of new information. The current pandemic and future crisis to come will certainly question the way we conduct research. But an open-minded, forward-looking, and data-driven science can help us to reason upon uncertainty.