RECANATI-KAPLAN FELLOWS PROGRAM

# “If there was ever a time for data science, this is it.”

## How Data Science Can Help Us Fight COVID-19 Now and Prepare Us for Future Global Pandemics

John Wigle

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
MAY 2020

**Recanati-Kaplan Fellowship Program**
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

**www.belfercenter.org/fellowships**



Design and layout by Andrew Facini

RECANATI-KAPLAN FELLOWS PROGRAM

# “If there was ever a time for data science, this is it.”

## How Data Science Can Help Us Fight COVID-19 Now and Prepare Us for Future Global Pandemics

John Wigle

EDITED BY

Galen Hancock

Sandra Sandoval

Adele Weinstock

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
MAY 2020

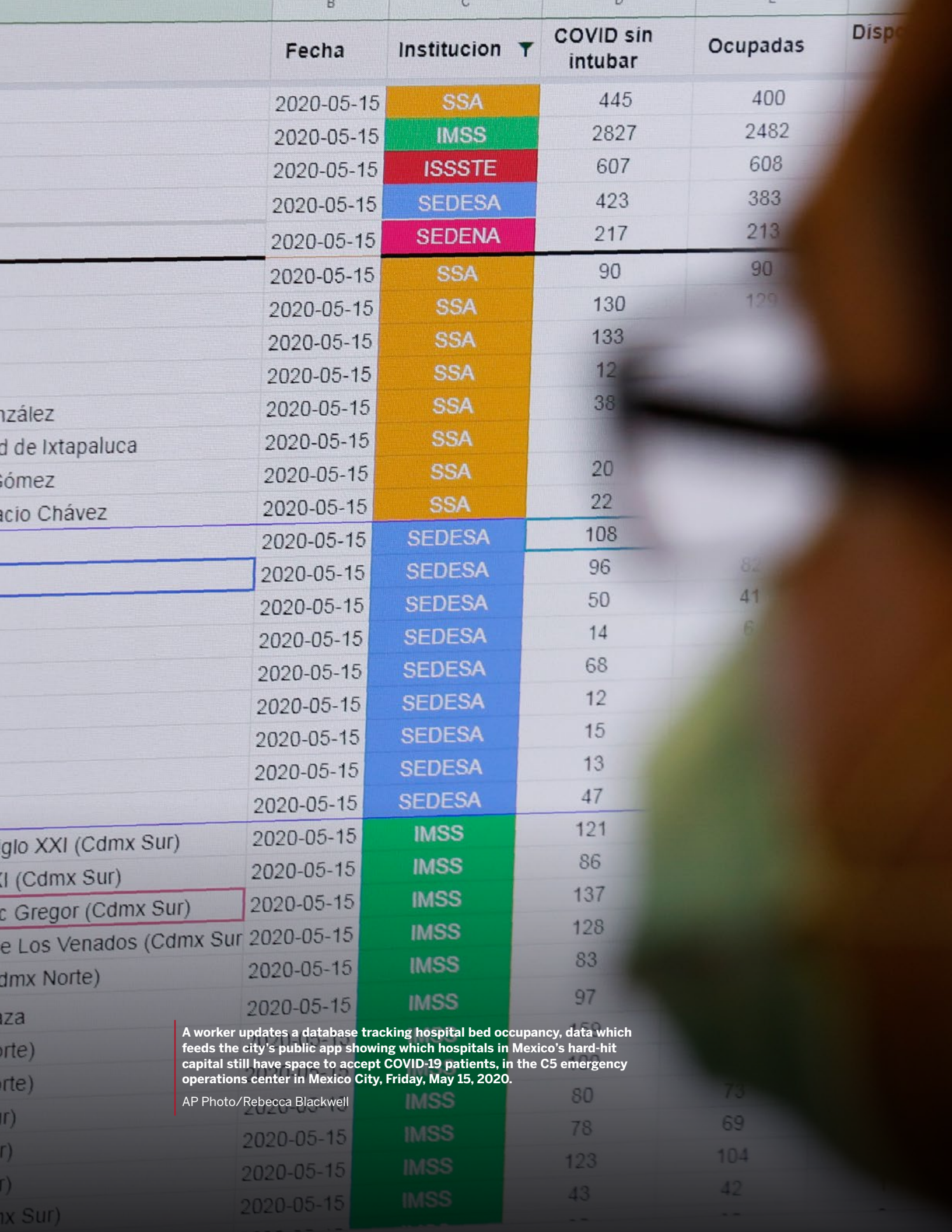
# Acknowledgments

Special thanks to Belfer Center for Science and International Affairs; the Recanati-Kaplan Fellowship; Paul Kolbe, the Director of the Intelligence Project; Caitlin Chase, Coordinator for the Intelligence Project; Calder Walton, Director of Research at the Intelligence Project; my fellow officers, data scientists, and professional colleagues; and my family.

# About the Author

**John Wigle** is a data scientist and a Recanati-Kaplan Fellow (AY19-20) at the Belfer Center for Science and International Affairs, Harvard Kennedy School of Government. He is a descendant of the Turtle Clan of the Tuscarora Nation, and previously was an adjunct lecturer at the Johns Hopkins University for over a decade before his selection as a Recanati-Kaplan fellow.

# Table of Contents




**A worker updates a database tracking hospital bed occupancy, data which feeds the city's public app showing which hospitals in Mexico's hard-hit capital still have space to accept COVID-19 patients, in the C5 emergency operations center in Mexico City, Friday, May 15, 2020.**

AP Photo/Rebecca Blackwell

# Introduction

I believe there is no better time for data science to serve a quintessential public good than during this time of national crisis. The fight against the COVID-19 pandemic has been described as a wartime footing by several heads of state.[1, 2] These times will require novel approaches like data science to manage this crisis.

The United States has invoked its Defense Production Act to redirect manufacturing to respirators and vaccine precursors,[3] and the pandemic's impact on the economy and unemployment has been compared to the great depression era just before World War II. The insights public health professionals and scientists share—from the numerous cases of infection with people who have stayed home, to the strange inflammation[4] called "COVID toes"—are a reminder that there is so much we don't know about this virus. Countries have applied various strategies from China's at-home detention of its citizens[5] to Sweden's bold experiment of herd immunity,[6] with no clear success story standing out. We are fighting this pandemic with limited resources, limited understanding, and under difficult economic constraints.

What are we to do? How can policy makers optimize our resources against this seemingly goliath micro-sized enemy to win and win quickly? What does a win look like? What are the acceptable losses? And is frequent hand washing, masks, and social distancing really enough to save us? With all these abounding existential questions and no clear answers, it is easy to see why many leaders have likened our current struggle with that of World War II.

---

1 World War II offers lessons—And warnings—For the coronavirus fight. (n.d.). Fortune. Retrieved May 11, 2020, from https://fortune.com/

2 Coronavirus and the language of war. (n.d.). Retrieved May 11, 2020, from https://www.newstatesman.com/

3 What World War II Can Teach Us About Fighting the Coronavirus. (n.d.). Wired. Retrieved May 11, 2020, from https://www.wired.com/

4 'COVID Toes': Mysterious Skin Condition Could be Linked to Coronavirus, Derms Say. (n.d.). NBC Chicago. Retrieved May 10, 2020, from https://www.nbcchicago.com/

5 Sealed in: Chinese trapped at home by coronavirus feel the strain. (2020, February 22). Reuters. https://www.reuters.com/

6 Orange, R. (n.d.). Will Sweden's herd immunity experiment pay off? Retrieved May 11, 2020, from https://www.prospectmagazine.co.uk/

Some have understandably chastised such an allusion citing the true horror of that war. But, growing up around a culture of oral history and a long-term community memory, I believe in the wisdom of crowds. As such, the allusion to the second World War may hold a nugget of wisdom that we should not forget. Looking back at how the allies handled World War II, we may care to remember the approach the allies took to optimize its resources against its enemies.[7] How did we decide what a win would look like? What were acceptable and unacceptable losses? And most importantly, how did we win quickly? The answer is a lot of heroic sacrifice of course, but it also included operational practices guided by field research—or operations research—a precursor to modern-day data science.

It is said that operations research was a decisive factor contributing to the allied victory in World War II. General Doolittle expressed his appreciation for operations analysts, the data scientists of his time, saying they made "substantial contributions toward the success of the Eighth Air Force."[8] General Carl Spaatz expressed his appreciation for his data scientists during the war, describing them as essential, and prophetically stated, "[w]e all hope that no similar national crisis will arise in the future... [i]f that time ever comes we shall call upon you again as we called on you before."[9] I believe that time of national crisis has arrived.

**WHAT IS DATA SCIENCE?**

Data science is field research that uses data and the application of appropriate analytic and statistical methods usually at a large-scale to glean insight that can inform policy makers of the current environment, its anomalies, trends and opportunities. It also applies algorithms, artificial intelligence, machine learning, hypothesis tests, and computer models to improve operational practices that can achieve a policy maker's desired outcome.

7 Gill Bennett, a British historian, coined the phrase "may care to remember" regarding our forgetfulness of history's lessons.

8 McArthur, Charles. (1990) Operations Analysis in the U.S. Army Eighth Air Force in World War II. Providence, RI: American Mathematical Society, Pg. 324.

9 Ibid.

# What Data Science Can Do For You

All public health officials are under pressure. They are faced with overwhelmed hospitals, equipment shortages, and operating with few insights to meet this challenge.[10] We have to move forward wisely and swiftly. We have to get data science into the field where operational practices will decide our collective fate. We can deploy data science teams to provide some answers to public health officials, improve existing processes, and map out a direction forward, just as operations research teams did during World War II to navigate our way to victory.[11]

Public health officials could deploy data science teams to hospitals, testing sites, nursing homes and other hot spots to better assess and understand the challenges they are facing. The teams would use their tradecraft and methods to sort out authoritative and questionable reporting, improve the analysis of contact tracing, build apps to automate data collection like we have seen in Asia,[12] discover the rates of community transmission of the virus, verify the effectiveness of testing, recommend how to optimally deploy resources to combat this pandemic, spot clusters of infection, identify vulnerable communities, and uncover surprising vectors of infection… just to name of few of things they could do.

We have limited resources that must stay aligned with our needs to save lives. Data science teams can gather and analyze data quickly and efficiently to keep our resources on target.  The teams can work with doctors, nurses, paramedics, emergency medical technicians, police and others who are the team's subject matter experts. These experts know their communities and are best positioned to observe important variations and anomalies and tip these leads. The data science team can then collect data and provide the analysis.  Those on the front line can in-turn adjust their operational practices based on the new insights gleaned from the data. Those practices can then be tested in real-time for their efficacy to improve public health results. So how do we organize such an effort?

10 Grimm, C. A. (n.d.). Hospital Experiences Responding to the COVID-19 Pandemic: Results of a National Pulse Survey March 23–27, 2020. April 2020, 41.

11 Operations research—History. (n.d.). Encyclopedia Britannica. Retrieved May 11, 2020, from https://www.britannica.com/topic/operations-research

12 Thompson, D. (2020, April 7). The Technology That Could Free America From Quarantine. The Atlantic. https://www.theatlantic.com/

# The Data Science Team

It is difficult to find a single individual—or full performance data scientist—who possesses the skills of a data engineer, a statistician, a computer scientist, subject matter expert, and a skilled communicator. A past lesson from World War II suggests that having a small team of people encompassing multiple skill sets will be best right now. I agree. The best data science teams are small three or four person teams. These teams usually have the following people (or skills):

- a data engineer,
- a mathematician or statistician,
- an analytic methodologist,
- a programmer or computer scientist, and
- a subject matter expert tailored to the research question.

The teams work independently but are goal oriented with a research question like, "find out which vector is the biggest transmitter of the virus" or as explained later in this paper, "find out how to keep our bombers on target."[13] Once the team has its marching orders, they will collect the data, interview people, and look at the problem from a variety of angles. They will use their nuanced understanding of the data, math and computer models to determine the best approach that is effective, simple, reliable, and has the greatest impact.

These teams can uncover answers quickly because they are in the field getting their hands dirty in the data. They intimately understand the strengths and weaknesses of what they are investigating. As such, they evaluate results and discern meaning, what is questionable, and the amount of error that may be present in their analysis and conclusions. This is the secret sauce of data science teams: their rich understanding of the subtle distinctions in data, its impact on analysis, its ability to explain the variation in results, how much it accurately reflects what they see in the field, and

13 McArthur, C. W. (1993). An Eighth Air Force Bombardier Looks Back at Operations Analysis. Interfaces, 23(5), 56–61. JSTOR.

most importantly the ability to see a possible solution even if its messy and unconventional. It's a combination of the science lab meets the real world.

These teams can help us explain, for example, why there are large numbers of hospitalizations by people who have stayed home.[14] Data is tricky, its dirty, and necessarily messy. Data science teams can untangle this riddle and give Governor Cuomo an answer. Are the numbers from poor data collection, a flawed survey, or just bad math? Or does the team need to visit these homes to observe conditions that may have contributed to the spread of the virus? Did all these patients come from tall high-rise buildings with one or more elevators out of service? Such conditions would require people to share cramped spaces and trapped air more frequently. Or does the building have a central HVAC unit with people sharing the same air? These teams can tease out these answers quickly, and in less obvious cases develop answers using math or other sciences.

## We Must Get Data Science in the Field during a Pandemic

First, an example from World War II. The allied bombing raids in World War II were very costly. Bombers took months to build at great expense. When a bomber was shot down, it was not only the loss of life, but the expense of a new plane, the lost time placing a new one back in service, and a lost opportunity to pursue our enemies. Teams of operation analysts studied the problem of aircraft survivability.[15] They determined if a plane survived long enough to return to an allied airfield then it could be saved or at least rebuilt quicker and at a lower cost than a total loss.[16] Even if a bomber crashed close to an airfield, the parts could be salvaged to keep other bombers flying. These simple insights uncovered by operations

14 Breuninger, N. H.-D., Kevin. (2020, May 6). Cuomo says it's "shocking" most new coronavirus hospitalizations are people who had been staying home. CNBC. https://www.cnbc.com/

15 Mangel, M., & Samaniego, F. J. (1984). Abraham Wald's Work on Aircraft Survivability. Journal of the American Statistical Association, 79(386), 259–267. JSTOR. https://doi.org/10.2307/2288257

16 Wald, A. (1980). A reprint of: A Method of Estimating Plane Vulnerability Based on Damage of Survivors. (CRC-432). CENTER FOR NAVAL ANALYSES ALEXANDRIA VA OPERATIONS EVALUATION GROUP. https://apps.dtic.mil/

analysts enabled the allies to keep a limited resource—the bombers—literally on target throughout the war.

> *"Many would say, 'what a great way to manage the spread of the virus,' but the data scientist sees a better way to save even more lives."*

But how would we apply data science during a pandemic? Some would argue that we already do, but I don't see Health and Human Services deploying data science teams to the field. Some may point to the data models developed by Justin Lessler, an associate professor of epidemiology at Johns Hopkins Bloomberg School of Public Health and the model developed at the University of Washington's Institute for Health Metrics and Evaluation (IHME) as data science during the pandemic.[17] These useful, excellent models do employ a facet of data science—modeling—but when data science enters the lab and stays there, I call it, well, um, "science." Data science is focused on daily operational practices and the messiness of humanity that does not fit so neatly into the ones and zeros of data, a petri dish, or a lab flask. It's not an arm chair sport watching from a distance, but an engaged field research activity that is often messy and imperfect, but demands the scientific method and its rigor. It is precisely this field element that gives data science an insight that complements the important science taking place back in the lab.

## Improving Contact Tracing

Data science belongs in the field to improve our operational practices that will decide our collective fate during this pandemic. One way we can apply it in the field is through contact tracing, but not in its present form. Contact tracing is very useful and often touted as one of reasons South Korea was so successful in getting control of the virus in its country.[18] It certainly is a useful tool: if you test positive for COVID-19, health officials ask who you may have exposed, they contact those people to get them

17 Yasemin Saplakoglu. (n.d.). Leaked White House document projects COVID-19 deaths will skyrocket. Livescience.Com. Retrieved May 10, 2020, from https://www.livescience.com/

18 McCurry, J. (2020, April 23). Test, trace, contain: How South Korea flattened its coronavirus curve. The Guardian. https://www.theguardian.com/

tested, and counsel them on self-isolating for 14 days. Many would say, "what a great way to manage the spread of the virus," but the data scientist sees a better way to save even more lives.

> *"True nobility is being superior to your former self."*
> *~ Ernest Hemingway*

Although there is a call for 100,000 volunteers to conduct contact tracing, they do not require the skills of a data science team.[19] This is a missed opportunity. Besides the plentiful discussion on automated contact tracing apps as seen in Asia, what if I told you that using methods from social network analysis, a mathematical discipline, could help us identify the most egregious vectors of infections? That it could highlight vulnerable populations, and could help us contain the spread quickly? And what if I told you this analysis requires minimal to zero changes in the current data collection practice of contact chaining? Why wouldn't we apply social network analysis, you might ask. The simple answer is officials don't know this is possible, but data science teams would intuitively apply it.

Here is how social network analysis (SNA) would work with contact tracing: the data is collected the same way, but it is connected together by a data science team showing the entire network of infection through a jurisdiction. The team then takes SNA measures of the network structure to determine where there are clusters—or hot spots—of infection, essentially triaging tens of thousands of infections to just a manageable handful of spots that warrant a closer examination. Data science teams can then deploy to these locations to investigate them. Going back to our New York City stay-at-home infections riddle, the teams may uncover these hot spots long before they hit the sixty percent margins on the governor's press charts.[20] If you can determine the attributes of transmission, then you can target the conditions with education, additional cleaning, and other changes in operational practices.

19 "Army" of contact tracers will be needed in coronavirus fight. Experts say that could cost billions. (n.d.). NBC News. Retrieved May 11, 2020, from https://www.nbcnews.com/

20 Breuninger, N. H.-D., Kevin. (2020, May 6). Cuomo says it's "shocking" most new coronavirus hospitalizations are people who had been staying home. CNBC. https://www.cnbc.com/

Conversely, a data science team could uncover locations that are similar to hot spots but showed significant reductions in the spread of the virus. Teams could deploy to these locations to learn if there are attributes that explain the reduced rate and if they can replicate these conditions elsewhere. Going back to our high-rise building example, if our hypothetical high-rise had shared airducts, but no spread of infection, how was this possible? The team may discover that the building installed far UV-C lights that are used to purify air in its elevators and air ducts years ago and it seems to stop the spread of the virus.[21] This information would be beneficial to the greater public good, and the team could get the word out quickly. The team could also consider how to operationalize far UV-C light in taxi cabs, Ubers, Lyft, schools and other close contact spaces.

## Improving an Aspect of Our Testing

There are many concerns regarding how the federal government has handled the current testing regime: particularly its availability and initially its accuracy. Although worthy of debate, I am setting those issues aside and will assume the tests are accurate and available for the moment to highlight something that irritates data scientists and statisticians about medical testing. This is particularly relevant with news of asymptomatic and unexplained infections.

Medical protocols require a person to exhibit more than mild symptoms before testing is even considered.[22] To understand the extent to which a population is infected, however, random testing of people with and without symptoms is necessary. Talk to any statistician about the criticality of random sampling. It allows us to approximate the rate of infection in the entire population without testing everyone. This allows us to use fewer tests to spot community transmission, approximate the rates of infection, and

21 Using the Power of Light: Preventing the Airborne Spread of Coronavirus and Influenza Virus. (2020, February 27). CRR. https://www.crr.columbia.edu/

22 CDC. (2020, February 11). Coronavirus Disease 2019 (COVID-19). Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html

decide if we have reached herd immunity numbers. But often the argument is made there are not enough tests to go around. As such, we don't know if there is community transmission of a virus. Then we are surprised by unexplained infections and are left guessing if community transmission is underway until it is too late.[23] Exacerbating the problem is the mounting evidence that there are asymptomatic people.[24] Asymptomatic people further support the case for random testing to know the true rate of community infection.

Data science teams will understand the wickedly profound value of random testing and use it to accurately inform public health officials about community transmission, the extent to which infection is propagating, and if we have reached herd immunity numbers. These are critical pieces of information to decide when to initiate social-distancing protocols, distribute resources, and eventually relax social-distancing and reopen communities.

Random testing also allows data science teams to identify and monitor unexpected blind spots. If random testing shows infections in places where it is truly unexpected, then the team can investigate and alert public health officials of its findings. For example, the current SARS-CoV2 virus is said to have originated in bats and passed to humans, possibly through a bite. If a random infection shows up in an unexpected place, the team can determine its possible origin. Hypothetically, the team may identify that the bat population in some areas are also spreading the virus, or bats nesting in airducts have contributed to the spread of the virus. These unexpected vectors of transmission would be caught and addressed.

Some efforts are underway to conduct tests on donated blood[25] to approximate a wider random testing regime. Local health department data are often unavailable for larger studies that would be helpful to public health officials. Data science teams in local health departments could organize efforts to facilitate a larger national study by organizing data for large scale analysis.

23 James, M. (n.d.). Second case of unexplained coronavirus appears in California, raising fears of community infection. USA TODAY. Retrieved May 10, 2020, from https://www.usatoday.com/

24 What We Know About The Silent Spreaders Of COVID-19. (n.d.). NPR.Org. Retrieved May 11, 2020, from https://www.npr.org/

25 Cohen, J. (2020, April 7). Unprecedented nationwide blood studies seek to track U.S. coronavirus spread. Science | AAAS. https://www.sciencemag.org/

# Data Science in Future Pandemics

I anticipate we will experience future viruses similar to this one for which there is no vaccine. We may care to remember in the future how we fought this virus today: knowing the approach we took to overcome it and the lessons learned from this episode. Armed with that knowledge, we could then add new data science tools and tradecraft to improve our future fights. Some of these data science tools and tradecraft include artificial intelligence, machine learning, and computer modeling using our latest know-how from the previous pandemic fight.

We could go beyond models of infection and mortality rates, and into the realm of predicting likely vectors of infection of future viruses using machine learning. We could use AI to identify probable symptoms based on genome sequencing.  We could model vaccines using strands of the virus RNA.

We could stand up data science teams to conduct health surveillance on bats and other animals that are prone to spread viruses to humans. Such health surveillance on animals could identify the next novel virus.  Data science can also identify the effective application of everyday low-tech protocols that can slow the spread of the virus, and determine the optimal deployment of high-tech or costly protocols that can efficiently stop the virus cold in its tracks.

Now is the time to apply data science to overcome this pandemic and prepare for our next fight. It will come again.

**Recanati-Kaplan Fellowship Program**
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

**www.belfercenter.org/fellowships**



Printed in the United States of America