Frontline Data Science: Lessons Learned From a Pandemic

Emily A. Beck¹ Hannah Tavalire² Jake Searcy¹

¹Presidential Initiative in Data Science, University of Oregon, Eugene, Oregon, United States of America,
²Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon, United States of America

Published on: Apr 27, 2023

DOI: https://doi.org/10.1162/99608f92.eea782bf

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
ABSTRACT
Disasters like the COVID-19 pandemic highlight the importance and difficulties of using data science to craft and implement effective disaster responses. While data science solutions are vital to the effectiveness and, importantly, the scaling of any frontline disaster response, lack of accessibility, training, and flexibility of these solutions can be extremely problematic. Here we outline our vision of Frontline Data Science and reflect on our successes and failures as frontline data scientists who pivoted from traditional research roles at a midsized nonmedical institution to build tools directly for the pandemic effort. We provide a summary of our experiences in building an interdisciplinary team founded on, and structured by, data pipelines and data scientists’ skills. We present ‘lessons learned’ from our experience building a COVID-19 Monitoring and Assessment Program (MAP), which provides SARS-CoV-2 information and testing to our university and large parts of our state. We further present ‘lessons learned’ from expanding disaster responses to traditionally unreached and disproportionately impacted groups. Finally, we provide a set of ‘principles for success’ that we hope will guide future efforts by data scientists wishing to immediately impact our society and provide information to organizations curious as to how data scientists can contribute to emergency response and disaster recovery.

Keywords: SARS-CoV-2, COVID-19 pandemic, accessibility, interoperability, public health, disaster recovery

Media Summary
As data scientists who pivoted from traditional academic research roles to leadership roles in building solutions to the COVID-19 pandemic, we provide a unique perspective on frontline data science—using data science in emergencies. We argue that data scientists, with training in computational techniques founded in interdisciplinary domains, are ideally suited to tackling large-scale emergencies regardless of and at times because of their domain expertise. Here we focus on specific strengths of our respective data science training and how those skills translated to helping a large team of health care workers scale SARS-CoV-2 testing at the height of the pandemic. We also provide a list of vital lessons that can be adopted by others when responding to future emergencies requiring immediate scaling of data curation and dissemination. We hope our experiences can inspire others to think outside their current roles in data science and about how their individual skillsets can be applied to help the broader community in times of crisis.

Visit the web version of this article to view interactive content.
1. Frontline Data Science in a Time of Crisis

In any disaster, curation and dissemination of data is vital. At face value, how to collect data from point-of-care locations and transfer it to the entities that need to make epidemiological or relief-based decisions appears to be a simple problem to solve, but even minor disruptions to data pipelines can cause massive delays and impact relief efforts. As large-scale data systems are adopted, it is therefore vital to address how they can fail during a disaster and who can best triage these failures. Many data system failures are not directly related to hardware or physical infrastructure problems but result from designs that do not scale to meet new demands, are incapable of integrating with larger disaster response efforts, or are not trusted by impacted communities.

When a crisis occurs and systems are not in place or are overwhelmed, triaging failures requires close collaboration between data creators, responders, and decision makers, as well as the skillsets to collect, collate, and disseminate data rapidly and accurately. We believe that careful foresight and investment in data systems, and in the cyber-infrastructure that supports them, are essential to prepare for the next crisis. We further contend that the data science community, with its broad focus on applied and collaborative work, is ideally suited to design and deploy emergency data infrastructure and solutions during crisis response; we refer to this work as Frontline Data Science.

The problem of scaling data curation and dissemination during a crisis was particularly apparent at the start of the COVID-19 pandemic. Holmgren et al. (2020, p. 1306), found “Public health agencies’ inability to receive electronic data is the most prominent hospital-reported barrier to effective syndromic surveillance.” In response Staes et al. (2020, p. 1822) identified “There is no question that inadequate resources have been a limiting factor for public health agencies to receive data from health systems. This problem is exacerbated by the many-to-one (hospitals-to-public health agency) nature of population health activities, the variable nature of hospital data contributions, and the resources required to onboard and manage interfaces with multiple health systems.” In addition, Houser (2022, p. 7) identified a “coherent data system for situational awareness,” as a common theme of failure during the COVID-19 response. Leonelli (2021, p. 4) highlighted a lack of prominence in two key areas of the data science community’s response to the pandemic: “Evaluation of Logistical Decisions” and “Identification of Social and Environmental Needs.” Similarly, calls for improved data management and analysis for situational awareness and crisis mitigation have been echoed in more immediate disaster scenarios such as earthquake recovery (Earthquake Engineering Research Institute, 2004) and wildfire response (Cuthbertson et al., 2021; Fitzgerald, 2021). Our goal in this panorama is to provide readers with concrete examples of our experiences as data scientists who pivoted from traditional research roles at a midsized nonmedical institution to build tools directly for the pandemic effort to inform data scientists interested in helping in similar situations and response managers who are considering their team’s composition.

2. Data Science About the Pandemic

It was not immediately clear how two data scientists with backgrounds in genetics (E.A.B) and particle physics (J.S.) would be useful to the pandemic effort. Neither of us had a background in health care or had worked with
electronic medical or health records systems previously. Despite this, we both had interdisciplinary training and extensive experience carefully handling large data sets, and admittedly we had the flexibility to contribute. Like many data scientists, we were no longer able to work at work and had no idea how long our regular jobs would be put on hold or be impacted by the pandemic. Fortunately, flexibility from campus leadership supported researchers pivoting from their typical jobs to help the pandemic effort, which proved to be a critical and fruitful decision for our community’s fight against COVID-19.

We, like many others, decided to draw on the skills we had as data scientists and spent the first months of the pandemic attempting to model the spread of SARS-CoV-2 infections. To this end, we built an agent-based model with the help of computational biologists and experts on the built environment to specifically model infections on the University of Oregon (UO) campus. Our initial goal was to help inform the UO pandemic response and provide guidance about potential impacts of masking policies, scheduling options for classes, and deciding to go to remote learning vs. in-person or hybrid learning. We were not alone in attempting to achieve this goal (Adiga et al., 2020).

The data science community has worked extensively with epidemiological data; modeling infection rates, predicting impacts of new viral variants, and tracking outbreaks. This work led to a boom in publications about the pandemic providing tools of varying quality to aid decision-making about pandemic responses. Unfortunately, we argue from experience that modeling, forecasting, and other popular research topics are ineffective without support from decision makers and risk contributing little actionable information once widespread political mandates are in place. For example, while local models may indicate that masking and isolation post-infection for longer periods may be the recommendation, Centers for Disease Control and Prevention guidelines may differ, leaving local entities unable to adapt to meet recommendations from local models. Our experience echoes the findings of Leonelli (2021) that policymakers should be involved in intervention and model development, and data scientists wishing to make immediate impacts should ask themselves if their results are capable of initiating change and who will initiate those changes before investing their effort. In our case, a critical limiting factor was that many modeling suggestions could not be implemented because of resource limitations. For example, in-person learning with less than five students per classroom is not a viable solution for a large university.

3. Data Science for the Pandemic

We do not wish to diminish the modeling work that was critical to understanding the effects of the pandemic. Instead, we wish to highlight the role data scientists can play in assisting teams that provide interventions and services to their communities. One finding that was consistent across our models—and the models of others—was that identifying positives and isolating them from the community was essential to slowing the spread. Unfortunately, a major limitation in most communities was the scalability of testing solutions. Most regions simply could not provide, process, and report tests at the rate needed to effectively prevent outbreaks. It was clear that to make an effective impact in our local community, we had to address the limitations in testing
capacity, not simply describe its impacts. We quickly found several roles in addressing testing limitations that we, as data scientists, were ideally suited for. We joined forces with another data scientist (H.F.T) with domain expertise in integrative biology and epidemiology to directly build solutions to scale testing in our community. The interdisciplinary nature of our data science team (genetics, physics, epidemiology) proved to be a strength of our team as we were able to provide unique perspectives with cohesive solutions. As a frontline data science team, we filled pivotal roles in the rapid development, deployment, and in-the-moment triaging of data systems to maximize impact of our university’s COVID-19 testing alongside a large team of public health experts and experts in prevention science.

As in our modeling work, it was essential to identify teams or collaborators that could enact changes based on our work. To that end, we worked with a team of biologists, human research compliance experts, and information technology (IT) professionals to build and deploy the initial laboratory data system for the University of Oregon COVID-19 Monitoring and Assessment Program (MAP). MAP’s data systems—described below—depended on laboratory, contracting, and compliance work, leaving only a month for technical development, which required rapid development and deployment of flexible systems. In the beginning of the MAP development, we did not know how long our testing centers would be active or how permanent our solutions would need to be. We did, however, know that solutions were needed immediately. Speed to deployment was therefore our top priority. We spent three additional months refining lab and data processes, scoping collaborations, and identifying bottlenecks. After this phase, university IT began working with additional commercial contractors to build a more robust and scalable system. This commercial deployment was available six months later, allowing for continued scaling of our testing services, which now support K-12 school screening testing and other state initiatives. Flexible solutions crafted by data scientists enabled better and faster COVID-19 testing and created optimized processes that were codified into a long-term production system. The primary difference between a production quality system and our initial triage system was that the triage system required more human touch points (in this case data scientists) to identify and address problems. However, our initial rapid system deployment was ready in approximately one-sixth the time and resulted in approximately 75,000 tests being delivered before the full production quality system was launched.

4. Developing a COVID-19 Testing Pipeline

The primary need for expanding testing availability and increasing turnaround time for reporting lab results was the development and fine-tuning of a testing pipeline that included coordination between several teams (Figure 1). Typically, in a nonemergency situation, building such a large-scale tool would require working closely with a commercial provider to design all aspects of the data pipeline over an extended period to build a tailored, scalable, error-free system. While commercial systems are available, they are suboptimal for several reasons. First, these products often come with high initial and maintenance costs, which are easily shouldered by medical operations with insurance billing systems but serve as a barrier to free patient care programs.
Second, these systems are designed to offer wide-ranging services (types of testing, types of results, methods for accessing results), but on a smaller scale, or in cases of high-throughput testing centers, there are large teams dedicated to overseeing information processing, which was outside of the scope of our testing program. Given the limited time between inception and launch, our group had to pivot substantially to build a data system as testing sites were being developed. Under Emergency Use Authorization (EUA), data scientists built temporary flexible solutions to allow for immediate launch of testing at UO MAP. This required data skills that were in short supply among laboratory staff and a strong knowledge of laboratory processes not generally found in IT departments, but was heavily influenced by decision makers from public health, prevention science, and epidemiological backgrounds to ensure maximum impact.

Fortunately, infrastructure at UO, including access to Qualtrics, IT support, human subjects researchers, and open-source materials like Python and R were available and critical to our success. Furthermore, our interdisciplinary domain expertise was important for this role. In some cases, domain expertise, including familiarity with polymerase chain reaction (PCR) and genomics, was directly applicable to aiding the pandemic effort, while other skills, including statistical training and mathematical expertise, were more tangentially helpful. Still, despite our varied domain expertise, we found that core training in data science, including the ability to problem solve and work using real-world data with scientific rigor, was foundational to our effectiveness. Based on these experiences, we contend that the most successful team to tackle frontline data challenges is a highly inter- and trans-disciplinary team of data scientists with wide-ranging domain expertise. Differences in perspective come from differences in lived experience and training and are crucial to tackling large-scale problems requiring large teams. We do not propose to be exemplars but feel we have identified
important roles currently filled in an ad hoc manner in emergencies that should instead be considered key positions in any frontline team. We strongly contend that the data science community can learn from our experiences to allow flexibility in the roles data scientists play. We further urge administrators and hiring teams to support such flexibility in cases of emergency to allow data scientists to lend their skills to frontline teams to solve problems as they arise.

We encountered several such problems while building our pipeline in a scalable, tailored way to be used by our teams (Figure 1). We needed to build custom pipelines for each stage of testing that could be utilized by team members with various computational backgrounds and identified many places where human error could impact testing efficacy (Figure 1). For example, we were able to write short, useable scripts to parse results to send to individual physicians. This single line of code reduced the amount of time the resulting team had to spend splitting result files into reports and was much more accurate than doing it by hand. We also used demographic data, including affiliation with the university to detect outbreaks early. For example, students who participated in club sports at the university indicated their sport affiliation upon registration, and we were able to detect if an outbreak was happening on a given team. We were also able to assist research projects and be respectful of their communities while remaining compliant with early electronic health record reporting standards. For example, the racial variable in the state system did not meet the needs of researchers because the categories were too broad. We built new systems to incorporate many new racial identities that could then be binned again when reports needed to be sent to the state. In most cases these were minor efforts from the data science team that had major impacts on efficiency and accuracy. Through these experiences we learned several lessons (described below) that we believe can be applied not only to COVID-19 testing but to other emergency situations that would benefit from a frontline data science response.

5. Lessons Learned

5.1. Prioritize Flexibility Over Automation During a Ramp-Up

One challenge arising early in the COVID-19 pandemic response effort was deciding where testing sites should be placed to maximize positive community impact. In addition to the university’s need to test its students, we wished to expand testing to the greater local community and eventually across the state of Oregon, focusing on reaching disproportionately impacted and historically marginalized groups. Several studies have shown consistent racial disparities in COVID-19 relief access across the United States, with Black and Hispanic groups being disproportionately impacted by the virus (i.e., higher positivity rates, higher hospitalization rates, lack of access to testing) (Escobar et al., 2021; Hu et al., 2020; Mullachery et al., 2022; Rubin-Miller et al., n.d.). One way we attempted to combat this disparity in resource availability was to develop new testing sites in regions previously unreached by COVID-19 relief efforts in Oregon. This work was focused specifically on meeting the needs of two particularly vulnerable and impacted groups, each with unique needs and various levels of accessibility. Because of the differing needs of our target underrepresented groups, we had to employ different flexible strategies to set up testing sites. As expected, we found that partnering with local government
and community-based organizations was vital to accomplishing this goal. However, it quickly became apparent that each new collaboration came with new data challenges.

We employed two alternate strategies to build effective testing sites outside our highly populated testing hub via UO MAP. The first target group(s) we attempted to reach were unhoused individuals and persons who inject drugs (PWIDs). In this case, we were able to tap into existing resources with outside agencies who had already built trust within these communities. We primarily partnered with the HIV Alliance (HIVA) needle exchange program offering free SARS-CoV-2 testing along with personal protective equipment to those participating in the needle exchange as well as anyone seeking resources. In this case, identifying target sites for testing and distribution was easy because we were able to use HIVA matriculation information to place sites at their most popular and already established needle exchange sites. While the opportunity to partner with other aid relief agencies is not always an option, this allowed us to immediately begin offering testing and was an excellent and effective option for quick action.

Our second target group included residents in rural communities, with particular focus on Latinx populations and undocumented individuals. In this case, no sites existed, and we had no infrastructure to lean on to provide effective interventions. To solve the problem of test site selection, data scientists were needed to identify optimal site locations capable of handling the needs of a new testing center. To this end, we utilized geographic information system and census data to target potential testing site locations to best serve these communities (Searcy et al., 2023). Importantly, to gain trust within these communities, we had to employ different approaches of community engagement targeted specifically to Latinx members, including offering multiple language options such as English, Spanish, and Mam and collaboration with existing community leaders. This meant alterations to every aspect of our data collection, testing, and result dissemination pipelines, which were completed quickly by our data science team. These additional efforts were instrumental in increasing accessibility and utility of our testing sites, with culturally tailored testing sites exhibiting greater utilization of target individuals (Degarmo et al., 2022).

Site selection was not the only challenge requiring data flexibility. Processes for collecting samples using a check-in station and a sample collection station could utilize shared databases in areas with good internet connectivity; the same system would fail in an area without that connectivity. Reporting positive results through a primary care physician worked well for our student population, who utilized student health center care, but largely did not apply to other community members. With expansion to rural communities, unhoused individuals, and less affluent groups came another list of challenges. Early iterations of our data collection, testing, and resulting pipelines relied on certain assumptions about those seeking tests. Primarily, that they had contact information (either a physical address or an email address to mail results to) and that they had internet access capable of registering them for a test either before arriving to be tested or at the testing site itself. While building our pipelines to help our testing team register patients and collect samples, we identified bottlenecks that needed data science solutions. Collecting patient demographic data required by the state was a time-
A key lesson in the above challenges is to not rush into automating all steps of your process and codifying them into production-grade systems. Flexibility and the ability to adapt quickly to new partners or new requirements should be a priority. A result of this priority is that many data processing and pipeline steps may have to be performed manually; for example, running scripts to check for quality errors may be triggered on demand, human checks can ensure compliance, and human intervention can track down missing or incorrectly entered data. Data scientists should focus on automating only the most time-consuming problems first and be ready to try new solutions regularly, even if these solutions will not scale in the long run. In this phase, a data scientist may have a clear operational role like any other member of the laboratory testing staff as opposed to just a systems maintenance and development one. This is significantly preferable to building a scalable system that is unusable due to a change in circumstance or one that prevents delivering services to new partners.

5.2. Data Scientists Can Reduce Data Burdens on Other Emergency Response Staff

Another key area where we believe data scientists are essential during an emergency response is reducing the burdens on existing staff. Data are vital for situational awareness and to provide information for direct intervention, but the burden of gathering this data often falls on laboratory and site staff or the participants themselves. Alleviating this burden can often be achieved through well-designed systems and the ability for data scientists to review and seek corrections for data gathered incorrectly. In a COVID-19 testing lab the primary point of interaction is with a laboratory information system (LIS), which is critical for the operation and compliance of a high-complexity clinical laboratory. A LIS is generally responsible for sample tracking, data quality control, and recording of medical results. Setting up a LIS can be a time-consuming process involving careful integration of each laboratory instrument into an integrated data pipeline. Many commercial LIS’s exist, however, each must be customized to fit the workflows and instruments available in a testing laboratory. This process can often take months of back and forth between a commercial development team and laboratory technicians. This can place difficult burdens on already overtaxed laboratory staff and limit time to deploy and test new capabilities. For example, single sample entry is often built into LIS solutions where a single sample is scanned and additional information entered by hand, however, to scale COVID-19 testing to thousands or more tests per day, bulk scanning of samples and the removal of human data entry was vital. This sample volume and upload challenged our commercial LIS’s database, and several improvements need to be rapidly deployed. Data scientists can play a key role in shifting the burden for developing and improving data processes from vital testing staff.

A key difference between data science and traditional IT expertise is data scientists must have a clear understanding of the science underlying test methodology in addition to the computational challenges in the data pipeline. This is another place where domain expertise in data scientists is important. Without both
experiences it is impossible to make informed decisions about investing effort into computational or laboratory process improvements. Data scientists must be the advocates for both lab process and staff time and have a reasonable understanding of the demands placed on IT teams when changes are requested. Data scientists need to own the process as much as the data systems. As an additional benefit to this, data scientists can advocate for process changes to speed up data handling during laboratory processes. In our case, we switched to asynchronous processing of laboratory samples and patient data entry, which significantly reduced reporting time and allowed for quicker interventions.

Beyond workflow and data system design, it is inevitable during a crisis that things will go wrong, and both may need to change rapidly. Supplies such as testing reagents or pipette tips can be in short supply, necessitating rapid changes to instruments and processes. Data scientists can quickly code fixes to normalize instrument outputs preventing large-scale changes in database structures or process design, or code fixes to keep operations working while longer term improvements can be made. A final important word of caution is that while flux is inevitable during rapidly evolving situations, it must be managed with rigorous adherence to ethical and legal standards. Regular testing for errors after changes is a must. A key difference for a data scientist to consider in an emergency scenario is that human verification and validation can often be used for checking results even if these checks would not be necessary when using a mature electronic system in a static environment. It is important for clear communication with technical and medical directors to ensure accurate compliant procedures are always maintained. It is increasingly common for first responders and laboratory techs to bear the burden of data systems and data entry, a role that they often have little desire or interest in performing, and frontline data scientists can help alleviate this burden in rapidly changing situations.

5.3. Data Scientists Should Be Involved at Every Step of the Pipeline

The most important lesson learned from working with each of these teams was that the data science team needed to be involved in the design of every aspect of testing for ensuring accuracy and quality assurance of data pipelines. As we continued to scale our testing system, we encountered an array of human errors that are unavoidable with our initial system when working on a large scale. While human error cannot be eliminated, data scientists can rapidly craft quality assurance and compliance solutions to catch and fix errors. In our case, errors ranged from typos on patient registration data, empty tubes being sent for testing, incomplete registration information being submitted, and so on. While our production system could detect and prevent many of these entry errors, our initial processes focused on identifying such errors through downstream systems or human checks. Data scientists were able to build error-catching systems in the registration forms and in team hand-off documentation that allowed us to process more samples faster. We found data hand-offs to be particularly problematic during development. We encountered a not-my-problem effect where once a team handed off samples or data, there were no checks in place to make sure the next team was notified or that the data arrived at all. Also, due to the large number of partners and collaborations, data models of one group would often be changed without notifying downstream users. This was exemplified when a regular check-in with our county
A health organization revealed that they were not receiving our results through the state system. We had reported our results through their portal, but due to a glitch in the system the results were not making it to the state’s data collection team and hence were not propagated to the county health department. Once identified, the problem was quickly remedied, and all data were restored to the state. Problems like this were only solvable in a timely manner because of the integrated role data scientists played in every aspect of the data chain. The data science team was the only team that understood the data chain from collection to utilization, which was critical for debugging and working around problems. With access to all data, across all stages of testing, we could identify problems, craft fixes, or alert collaborators to potential problems.

There were also key jobs performed by each team that required data science solutions. Because of this, it was essential that the data science team be trained in all aspects of the testing pipeline. Building custom solutions for our LIS and developing scripts to generate reporting files were only possible because of the intimate knowledge of the testing pipelines and the important domain expertise of our data scientists. Ensuring rapid patient resulting and reporting was a top priority for our entire testing team. Because our goal was to identify and notify positives as quickly as possible, we needed high-throughput data solutions to connect testing results from deidentified samples back to patients. Unfortunately, contacting patients is not a simple task that requires a single solution. Different legal restrictions on the way(s) results can be delivered and who can deliver them can often slow the process. We encountered several problems using low-tech solutions such as email or slower processes like USPS mail. Despite efforts to speed up the reporting process to patients using email, many patients did not have email addresses or provided email addresses with typos, leading to a backlog of results that failed to reach the intended patients. This meant that even if a positive result was obtained by the resulting team within 12 hours of testing, a patient may not have been notified for 3–5 days. This was an extremely important problem that needed immediate attention as positive results were being delivered days too late. To solve this problem, our data science team worked with commercial partners to build an online portal for patients to log in and receive their results. Data scientists, with their knowledge of the entire pipeline and bottlenecks encountered by the resulting and reporting teams, were able to build a high-throughput, easily accessible, accurate reporting system to alleviate this pain point.

In addition to notifying patients, our lab worked with several community partners and had several legal reporting requirements to individual health care providers and state agencies. Each of these entities had their own content and formatting requirements and required different file types to be compatible with their own unique data systems. A typical off-the-shelf commercial system for handling health care data could produce these distinct types of reports, but this was not a financially viable option. Again, this process relied on our data science team to create custom scripts to pull various subsets of data content from a metadata file that was then formatted to meet the needs of the various reporting agencies. We then ultimately worked with our commercial partners and IT team to build capabilities for generating each of these types of reports into our reporting system.
5.4. Interoperability Is More Important Than Robustness: An Ode to the Comma Separated Value (CSV) File

Another important lesson we learned was about interoperability of file types. In the health care system, Health Level 7 (HL7) files are standard for transferring health care data between systems. However, there still exist several versions in use, and purchasing software capable of handling these types of files is often expensive, which limits its use. This can be a major barrier to collaboration with partners who may not have electronic health records or use various standards. In our testing example, patient data would be received from collection sites and testing results would need to be sent to state systems and the ordering physicians, who mainly comprised county health officials but also included some private practices. This wide net of partnerships was vital for success, but despite a significant investment of time and effort, we found it largely untenable to require partners to conform to a set data format as any changes placed significant financial, time, and training burdens on already taxed teams. We found focusing instead on the simplest format possible provided the only way to manage these many different collaborations and hand offs. We found that simple CSV files were crucial to our success. While technically similar in format to HL7 messages, CSVs are readily producible with standard software, easily exportable by almost all systems and instruments, and require little training to understand or edit. The tradeoff, of course, is the lack of a consistent data model and increased opportunity for human error. This, however, again reflects our philosophy of making data production easy and shifting data burdens onto embedded data scientists who were responsible for normalizing data sets and identifying errors. Importantly, because CSV files have extensive existing support in almost all coding languages, we were able to craft in-the-moment solutions more quickly. In the end, we learned that while there might be more robust computational solutions to problems, prioritizing ease of collaboration and ability to shift the data burden away from partners and onto data scientists was a key requirement for initial collaborations.

6. Vision of Frontline Data Science

In the inaugural HDSR editorial, Meng (2019) proposed a broad ecosystem view of data science that includes practitioners spanning a wide variety of domains and technical skills. We believe a key feature of data scientists is their varying domain expertise, which is essential to crafting a diverse team capable of solving interdisciplinary problems. These domains are wide-ranging, from STEM fields to linguistics, law, and business. An important question is what are the skills and roles that define a frontline data scientist. In our vision, frontline data scientists have a charge to ease data burdens from first responders and emergency management, address unforeseen problems quickly, inform the development of more permanent solutions, and communicate across disciplines to ensure accurate and efficient transfer of information across several community partners. Technical skills include data manipulation with a careful eye to fidelity and testing to assure accurate data pipelines, the ability to quickly absorb the scientific and technical understanding of the domain in which they are working, and the ability to summarize and present information to others succinctly. Frontline data scientists must also be tightly integrated within any response team with the ability to inform and collaborate with a wide variety of roles that interact with or generate data, for example, compliance officers,
community engagement teams, laboratory technicians, information technology staff, and clinical personnel. In this role we advocate several principles for success for an aspiring frontline data scientist:

1. Providing services to the community should be a core driving principle even if conditions require immediate data collection without internet access or even without electronic formats.

Frontline data scientists require the ability to advocate for needed change, particularly when this change may cause burdens on other units. We found that frequent data gathering, particularly at the patient level, was complicated by electronic forms. Although these forms made data handling much easier and cleaner for data systems, a frontline data scientist’s goal of prioritizing the community in need may require additional data processing and cleaning steps even if they complicate existing data solutions.

2. It is vital that electronic formats or systems do not create a barrier to collaboration.

Often, best practices and standards vary widely between data systems and products. Trying to connect native systems with multiple standards is difficult, expensive, and time-consuming. Focusing on simple common formats that are readily accepted by most software such as CSVs can provide a platform to build more robust systems.

3. Data scientists must be involved with all elements of a project’s data from beginning to end. If they do not integrate data systems into a coherent whole, no one else will.

In many cases where well-defined teams and handoffs exist throughout emergency management scenarios, it is important that a frontline data scientist be able to follow the data through each step of the process. Data can be lost in transit, and in a crisis, the receiving or transmitting unit will not always identify this. Regular checks that information has been delivered completely and accurately in a timely fashion can quickly yield huge benefits.

4. Emergency systems should prioritize speed of setup and the ability for humans to identify and correct data faults; streamlining and cleaning pipelines can come later.

Ideal electronic systems prevent data entry errors at the source with rigorous data standards that prevent messy or inaccurate data from being entered. However, these systems also take significant time investments to set up and failures in data entry can create bottlenecks that require testing staff to contact IT staff to properly understand and debug. Instead, systems that allow for error correction with follow-up contact or with human intervention should be prioritized.

5. Emergency systems should concurrently be replaced by longer term enterprise solutions that reduce or eliminate data errors and human interventions to the data process.
Following principles 1-4 in our case ultimately resulted in a quickly deployed and functioning system containing a patchwork of quick fixes. These systems are not scalable without continuous interventions by the data scientists who created them. While this may lead to short-term success, it should not be considered the end goal. Instead, these systems should be considered as experiments to inform the development of more robust systems.

In their description of competencies required for crisis management, van Wart and Kapucu (2011, p. 501) quote a participating manager: “In my opinion, [willingness to assume responsibility] is the competency where the differences are the greatest. The culture of the traditional response industry is to build strong protocols of mission/objective-oriented discipline. When there is a catastrophic event, the successful leaders are the ones who have the willingness to accept responsibilities outside their experience and training.” We believe data scientists should take this to heart. While it is true that crisis management is not normally a part of a data scientists’ skillset, we have found that there are numerous areas where data scientists can utilize their particular mix of careful data handling and applied computational skills to contribute in a time of crisis.

7. Vision for Frontline Data Science

The development and fine-tuning of our COVID-19 testing program was a massive learning experience. We were imperfect and there are certainly things we would do differently if we were starting from scratch. Primarily, we would recruit more data scientists from wide-ranging domains to help early and often. Still, we do see our efforts supporting MAP as a success. Most importantly, with MAP, the university stayed open. With regular testing, we were able to provide on-campus housing, in-person classes, and even student events like club sports. After codifying our pipelines, we were able to return to our research jobs at UO while our systems continued to be used, which have now made far-reaching impacts on our state. The scalability of our systems allowed for the expansion of testing to support local elementary, middle, and high schools, as well as communities across the state. Looking back, we feel we made a genuine impact on our community and we encourage data scientists to join the effort in our next crisis.

Acknowledgments

We would like to acknowledge the UO COVID MAP Team who allowed us the opportunity to play a role in their efforts to help our community. We would also like to thank the UO Presidential Initiative in Data Science for supporting the efforts of our data science team to the UO COVID MAP and supporting the development of this panorama.

Disclosure Statement

Emily A. Beck, Hannah F. Tavalire, and Jake Searcy have no financial or non-financial disclosures to share for this article.
References


Fitzgerald, A. M. (2021). A new wildfire watchdog: Alerts about forest fires shouldn’t depend on pets smelling smoke. We need smart infrastructure, and that needs zero-power sensors. *IEEE Spectrum, 58*(11), 38–43. [https://doi.org/10.1109/MSPEC.2021.9606510](https://doi.org/10.1109/MSPEC.2021.9606510)


©2023 Emily A. Beck, Hannah F. Tavalire, and Jake Searcy. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

**References**


