A culture for doing good things with data

The why, what, how, and who of a progressive culture for data in public health

An essay by Chad Heilig

2023-10-22

Contents

Preface: On doing good things with data ***2***

1 Introduction: Why, what, how, and who ***3***

2 Why ***4***

3 What ***5***

3.1* What data science *is **5**

3.1.1 Data science is what data scientists do 5

3.1.2 Learning from data 6

3.1.3 Core activities and critical reflection 8

3.1.4 Commitments: life cycle, centered on analysis, subject to norms 10

3.2* What data science *is not **11**

3.2.1 Data science is not statistics 11

3.2.2 Data science is not data analysis (not even machine learning) 12

3.2.3 Data science is not informatics 13

3.2.4 Data science is not just good science 13

4 How ***14***

4.1 Foster a progressive culture **14**

4.2 Foster technical skills **15**

4.3 Foster nontechnical skills **17**

4.3.1 Intellectual character 17

4.3.2 Ethics and values 20

4.4 Foster community and leadership **21**

5 Who ***22***

5.1 Who gets to do data science? **22**

5.2 *Learning* in a progressive culture for data **23**

5.2.1 Relational learning 24

5.2.2 Formal curricula and structured programs 25

5.3 *Doing* in a progressive culture for data ... 26
5.4 *Staffing* in a progressive culture for data ... 27
5.4.1 Data science staff should be cultivated, hired, *and* outsourced ... 27
5.4.2 Data science staff should be organized to do data science ... 31
5.5 *Leading* in a progressive culture for data ... 36
6 Redux: Who, how, what, and why ... 37
7 I get to do data science ... 38
7.1 How I think about data science ... 38
7.2 My personal history with data science ... 39
7.3 Why a progressive culture? ... 40
8 Machine learning and artificial intelligence ... 41
8.1 Context: what is familiar or known ... 42
8.2 Context: methodology ... 43
8.2.1 Machine learning and statistics ... 43
8.2.2 On "predictive analytics", ML, and AI ... 44
8.3 Context: history ... 45
8.4 Context: organizational culture ... 46
Declaration: a progressive culture for data in public health ... 48
References ... 50

Preface: On doing good things with data

Everyone who wants to do good things with data should have the intellectual support to do so, as long as they proceed with rigor and stand behind their work. This is my *credo*—my fundamental, animating belief in a culture for doing good things with data in public health.

This essay presents views that I have developed throughout my career, especially during my 8-year tenure as the Associate Director for Data Science in CDC's Center for Surveillance, Epidemiology, and Laboratory Services from January 2015 through CSELS's dissolution in January 2023. I drafted this essay primarily in early 2022 to bring together as a coherent whole several related ideas on how CDC should think, talk about, and support a work culture oriented to doing good things with data. In my view, this is CDC's single greatest area for gains from doing good things with data: connecting technical excellence and analytic rigor to doing science better, getting better at learning things about the world, and getting better at doing things with what has been learned. In my experience, CDC has tended to overemphasize technology and underemphasize critical reflection and practical wisdom in doing things with data.

In the next 4 sections, I expand on *why* we should care about doing good things with data, *what* data science is and is not, *how* to construct and support a culture for doing good things with data, and *who*

plays various roles and carries out the functions of a culture for doing good things with data. In the penultimate section, I add my personal history with data science. Then I extend my discussion on machine learning and artificial intelligence as a salient, contemporary set of issues for doing good things with data. Finally, I cap this essay with an aspirational declaration for creating and fostering a progressive culture for data in public health.

This essay is a snapshot in time. It reaches back to extensive reading and study that I undertook from 2014 through 2016, but it stops around 2021. The field continues to change rapidly. But as much as data science is about keeping up with fast-moving methods, tools, and technology, the schema for data science is itself stable. So, for example, where I describe how CDC should think about machine learning and artificial intelligence, I don't specifically address recent large language ("chat") models.

I dedicate this essay to the dozens of folks whom I have mentored since I joined CDC as a federal employee in February 2000. I believe in you. You're the reason that I wholeheartedly believe that a progressive culture for data centers on learners—because learners believe.

1 Introduction: Why, what, how, and who

Data science acts on the belief that if you approach data in just the right way, you can discover and unlock its meanings. As data become more varied and complex, data science helps in removing impediments to data's meanings so that no data are off limits, no data have to go unlearned. Sometimes I approach data gently, as a data-whisperer intent on codiscovering with the data its own potential to reveal things about the world and to inform action in the world. Sometimes I wade in gingerly; sometimes I dive in; and sometimes I catch and ride the waves as the story within the data comes to the surface.

It's easy to be skeptical of the concept of data science, especially when it seems like it means many things but not much of anything. "Data science is what data scientists do," wrote Davenport and Patil (2012). Does the phrase convey anything substantive? Does it offer anything new compared, say, to the data-oriented fields of statistics and informatics? Let's open with the why, what, how, and who of data science and then unpack these themes.

> **Why**: Foremost, data science is about learning from data. Its purpose is broadly to bring together, in a rigorous way, all that goes into doing good things with data. Data science promotes principled use of the full breadth of methods, from the familiar to the unfamiliar, along with the norms to ensure that methods and results stand up to scrutiny. Data science helps us to keep up with evolving methods, tools, and technology for learning from data of all structures, sizes, shapes, and speeds in a way that other disciplines do not. Dynamic and complex technologies and data motivate but do not define data science.
>
> **What**: Data science studies how to learn from data—especially complex or nontraditional data. It combines analytic, computational, and subject-matter methods to connect the whole life cycle of data: Frame what you want to figure out. Obtain and prepare data to engage the question. Preserve and share what was learned, how it was learned, and how that learning fits in with what is already known and with other choices that could have been made.
>
> **How**: At the individual level, data science calls for technical and nontechnical skills. At the collective level, it calls for a forward-looking but grounded culture that supports putting those skills to

use for doing good things with data. Technical skills cover analytic methods, such as statistics, machine learning, or causal inference, and computational methods, such as data wrangling and implementing and scaling algorithms. Nontechnical skills support good science generally and good data science specifically, such as the ability to approach a problem with curiosity, attentiveness, perseverance, open-mindedness, and creativity.

Who: Everyone who wants to do good things with data should get to make the effort, as long as they are rigorous and accountable. A rich culture for data science includes expert and nonexpert doers, learners, mentors, supporters, and advocates, organized to operate and keep up with fast-moving methods, tools, and technology for doing good things with data effectively and sustainably.

I unpack these 4 circumstances—the why, what, how, and who—in the next 4 sections.

2 Why

The purpose of data science is broadly to bring together, in a rigorous way, all that goes into doing good things with data—for learning from data and for building things with data to put those learnings to use, for example, in safeguarding public health. CDC consumes a lot of data to support its public health mission. Traditional sources include surveillance, vital records, surveys, program evaluation, studies of health services, and clinical trials. More recent sources include billing and claims data, electronic health records, social media, and sensor data. From small, structured data to high-volume, unstructured data, over time the scope and scale of those data expand and become more complex.

Why should we focus on data science? Because we need to keep up with rapidly changing methods, tools, and technology for extracting meaning from data.

Data science makes available tools for classic problems, such as working with data through the life cycle from problem formulation to collection to data management, through analysis, interpretation, and presentation. Contemporary issues in data science arise from movements to be open and to expand the scale of data and sophistication of analytic methods. Making data available for wide audiences, while ensuring adequate protections of individual privacy. Describing practices to make analyses reproducible, or at least traceable. Working with high-volume data, such as genomics, and high-velocity, real-time data, as found in syndromic surveillance and claims data feeds. And making sense out of unstructured text, images, and other nontraditional data types. Most contemporary problems differ from classic problems in scale rather than kind. For example, whether administrative data come from paper-based registers in resource-constrained settings or from massive stores of insurance claims data, they pose the same problems for inferring causes. Data science promotes principled use of the full breadth of methods, from the familiar to the unfamiliar, along with the norms to ensure that methods and results stand up to scrutiny. Data science crosses disciplines.

There's another reason to focus on data science: Data science affords a measure of autonomy for practicing and honing data skills, because of ready access to open methods, tools, and technology. Some technical skills require special equipment (like growth media or microscopes in a microbiology laboratory) or access to humans (like clinical medicine or behavioral counseling). In contrast, to learn about

data and from data, it is often enough to have data in hand, widely available software, and the persistence to jump into a problem and break it open. Software is now often freely available, with growing contributions by the very active R and Python user communities. So data science can be practiced with a great deal of self-determination. With that autonomy comes the latitude to own and direct one's learning. Thus, the enterprising scientist can capitalize on that autonomy in order to keep up with fast-moving methods, tools, and technology, in part by continuing to learn how to learn from data. This autonomy presents a paradox that we will try to resolve later (in section 5.4.2): Data science is necessarily interdisciplinary, and not every practitioner needs to cross all the disciplines. So how can one be autonomous and team-oriented at the same time?

In summary, we focus on data science because we want to learn from data, learn about data, and learn with data.

Learning *from* data: Data have value because data help us learn things about the world. What we learn helps us to make informed choices about how we interact with the world, for example, through public health interventions.

Learning *about* data: Data come in many structures, sizes, shapes, and speeds, from small, flat data tables to massive, unstructured data streams. Data conform to a variety of standards, or no standards at all. The varied characteristics of data both enrich and constrain the ways that data reveal characteristics of the world.

Learning *with* data through its full life cycle: Analytic knowledge and skills allow us to pose rich questions about the world, amenable to rich methods; guide how we generate, transmit, obtain, and prepare data; probe data to answer questions about the world; place answers from data in context, mindful of assumptions and alternatives; present data-driven answers to audiences clearly and correctly; and preserve those answers and ensure that the entire life cycle is transparent, accessible, traceable and, to the extent possible, reproducible.

The field of data science addresses a wide variety of problems (what), and the practice of data science straddles autonomous and collaborative styles (how). Thus, we also focus on data science so that we can build and sustain a culture (who) for doing good things with data, for continuously learning things about the world, and for empowering choices informed by those learnings, and for being ever ready to learn from and act on data.

3 What

3.1 What data science *is*

We have no shortage of definitions of "data science", because different definitions serve different purposes. I will start with a tautological description that almost amounts to an operational definition.

3.1.1 Data science is what data scientists do

Writing in 2012 for the *Harvard Business Review*, Tom Davenport and DJ Patil (the US Chief Data Scientist during the Obama Administration) wrote the following:

> [W]hat data scientists do is make discoveries while swimming in data. … At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. [D]ata scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data. Data scientists realize that they face technical limitations, but they don't allow that to bog down their search for novel solutions. [T]he dominant trait among data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested. (Davenport and Patil 2012)

3.1.2 Learning from data

For pith, we can turn to Donoho (2017): "Data science [is] the science of learning from data, with all that this entails," to which he adds, "it studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner."

For plainness, we can turn to my working definition: "a set of core activities to ask good scientific questions and to line up the tools to answer them rigorously using data." I developed that formulation in 2016, drawing on *The Art of Data Science* (Peng and Matsui 2015). My operating definition of data science links rather than separate the tasks of stating and solving problems, mediated through data. Peng and Matsui enumerated 5 core activities, to which I add a sixth. I explicitly link each of these 6 core activities to analysis, since data analysis is a central commitment of data science, as I explain further below.

Pose good questions. The set of potential questions is enriched through awareness of the kinds of learning that various analytic methods support.

Prepare data to address those questions. With the purposes of analysis in mind, the practitioner can obtain, manage, and explore data to ensure the data's fitness to address the analytic purposes.

Probe the data. Conduct rigorous analysis to address questions, which includes developing and critically assessing one or more analytic models. The value of analysis itself comes from the ability to answer the question and to convey what is learned from data.

Place analytic results in context. Interpretation binds the question to the method, binds the method to the result, and puts them all into the context of assumptions about the data, technical assumptions in the analytic models, existing domain knowledge, and alternative analytic approaches that could have been considered. Understanding what specific data can't tell you, or what phenomena those data rule out, is as valuable as interpreting what the data show.

Present methods and results. Communication shares what has been learned from data and how it was learned, and it also subjects the life cycle to scrutiny and transparency.

Preserve the entire life cycle. Ensure that the life cycle is traceable, accessible, reproducible, and enduring to the extent possible. In addition to communicating methods and results, transparency ensures that the data and analytic code are as available as possible, subject to privileges of access where necessary. This transparency in turn supports fundamental scientific norms.

Peng and Matsui took care to explain that their core activities do not need to, and often do not, occur in sequence. Rather, with the execution of each core activity, careful reflection could lead the practitioner to repeat, jump back, or jump forward. I expand on this idea later in this section.

Blei and Smyth (2017) wrote, "Although each of [statistical, computational, and human perspectives] is a critical component of data science, we argue that the effective combination of all three components is the essence of what data science is about. ... The practice of data science is not just a single step of analyzing a dataset. Rather, it cycles between data preprocessing, exploration, selection, transformation, analysis, interpretation, and communication. One of the main priorities for data science is to develop the tools and methods that facilitate this cycle."

The Data Science Association focuses on meaning: Data science is "the scientific study of the creation, validation and transformation of data to create meaning" , and a data scientist is "a professional who uses scientific methods to liberate and create meaning from raw data." (Data Science Association)

The National Institutes of Health *Strategic Plan for Data Science* defined it as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data." (National Institutes of Health 2018)

The National Academies of Sciences, Engineering, and Medicine described data science along with its relationship to other fields, its primary tasks, and its primary purposes:

> [Data science centers on] multidisciplinary and interdisciplinary approaches to extracting knowledge or insights from data for use in a broad range of applications. It is the field of science that relies on processes and systems (mathematical, computational, and social) to derive information or insights from data. It is about synthesizing the most relevant parts of the foundational disciplines to solve particular classes of problems or applications while also creating novel techniques to address the 'cracks' between those disciplines where no approaches may yet exist … because the volume and variety of data available are expanding swiftly, data are more available immediately, and decisions based on data are increasingly automated and in real time. (National Academies of Sciences, Engineering, and Medicine 2018)

Finally, the National Institute of Standards and Technology (NIST) in 2015 defined data science as the "Extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing." Per NIST, a data scientist is "a practitioner with sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle … The end-to-end role ensures that everything is performed correctly to explore the data, create and validate hypotheses." (NIST Big Data Public Working Group 2015) The data life cycle is "a set of processes in an application that transform raw data into actionable knowledge":

- **Collection**: Gather and store raw data.
- **Preparation**: Convert raw data into cleansed, organized information.
- **Analysis**: Synthesize knowledge from organized information.
- **Action**: Use synthesized knowledge to generate value for the enterprise.

NIST's definitions appear in the context of discussing "big data" that cannot be accommodated by traditional architectures. Data volume (size), variety (sources, domains, and types), velocity (rate of flow), and variability (changing characteristics) "drive the shift to … architectures for data-intensive applications." (NIST Big Data Public Working Group 2015) I prefer plain-language versions of these concepts:

Table 1. Attributes of "big data"

Attribute	V	S
Number of data elements	Volume	Size
Multiple repositories, domains, or types	Variety	Scope, Sources
Rate of flow (records per unit time)	Velocity	Speed
Richness, complexity	Variability	Shape, Structure

These varied definitions of data science explicitly invoke learning, answering questions, creating meaning, and extracting knowledge—epistemic tasks that take us into the nature, sources, structure, and limits of empirically derived knowledge. Learning defines data science, which in turn centers data science on the role of analysis. These definitions vary in whether and how they appeal to processes (such as a data life cycle and adaptive problem-solving), means (complex data), and methods (quantitative and analytical approaches and technology). These definitions also suggest but do not enumerate the skills that are needed to do data science.

I agree with the NIH and NASEM definitions in that current and evolving complexity drive a need for adapted methods. As data and technologies become more complex, the drive for adaptively learning from data also intensifies. The risk I see is that we might overinvest in narrow (but potentially useful) skills while giving short shrift to broadly applicable but underserved skills: We should respect the fundamentals and avoid an unfortunate tendency to overemphasize the exotic or the complex at the expense of those fundamentals. Indeed, Donoho (2017) makes a similar case. We cannot lose sight of the need for learning from conventional or familiar data, using conventional or familiar methods. In addition to basic skills for managing and analyzing data, we need always to esteem skills for critical reflection and reasoning in creative but disciplined ways. In this sense, complexity and sophistication *motivate* data science, but they *do not define* data science, and it is a mistake to overidentify data science with those drivers. It would also be a mistake to focus on the technology rather than the science: Learning from data is a scientific act, enabled by evolving methods and tools. The value of learning from data needs to be judged by scientific rather than technological norms. Have we posed questions of social value and scientific validity? (See also Freedman 1987.) Have we prepared the data to answer those questions? Have we adequately probed the data and placed findings in context? Are proposed conclusions traceable and defensible? Is the reasoning coherent? We will return to this point when we talk about *how* data science is done.

3.1.3 Core activities and critical reflection

As mentioned above, Peng and Matsui's schema for the core activities of the art of data science goes deeper than merely listing those activities: Each core activity calls for critical reflection, in which the practitioner reviews each core activity by framing, checking, and possibly revising or revisiting that activity.

First, set your expectations for the core activity. Then collect information and compare your expectations to your information. If they don't match, then take another look and either revise your expectations or your information. Maybe repeat the current core activity or return to a previous one. If they do match, then you're in a good place, but you should occasionally check again for good measure, lest you fall into a confirmation bias trap.

In a real sense, critical reflection puts the *science* in *data science*. We can associate each core activity with example prompts for critical reflection.

> **Pose good questions**. Is the question of interest? Valid and valuable? Consult with experts or the literature. If needed, revise the question.
>
> **Prepare data to address those questions**. Are the data suited to the question? Examine data early and often to learn about structure, content, and suitability. If needed, refine the question or obtain more or different data.
>
> **Probe the data**. Does the analytic model answer the question? Does it use data correctly and take a suitable form for explaining or predicting a phenomenon of interest? Challenge model assumptions and structure. If needed, revise model structure or inputs.
>
> **Place analytic results in context**. Does the analysis provide a meaningful answer that holds up to scrutiny and contributes to domain knowledge? Assess the totality of analyses—effect sizes, accuracy or bias, variability or uncertainty—in consideration of varying assumptions about the data, the model, and the subject-matter context for the question. If needed, revise the analysis to provide a specific, meaningful answer or conduct diagnostics or sensitivity analyses to assess limitations and assumptions.
>
> **Present methods and results**. Are the methods and results understood, complete, and meaningful to the audience? Assess content, style, and attitude and gauge audience feedback. If needed, revise the presentation to suit audience needs.
>
> **Preserve the entire life cycle**. Are the data and analysis as open and transparent as possible? Assess whether data and analytic code can be made available with or without alteration or access controls, for example, through public repositories or under user agreements. If needed, work toward the least restrictive means for sharing.

These critical reflections show why the core activities need not occur in sequence. Indeed, some or all core activities could occur more than once for a given undertaking.

Table 2. Core activities, critical reflection, and iteration

Core activity	**Critical reflection**		
	Set expectations	**Collect information**	**Resolve mismatch**
Pose: State a good question	Question is of interest, will advance public health	Consult experts, literature	Revise the question

Prepare: Obtain and explore data	Data are appropriate for the question	Examine data early, often to learn *about* the data and learn *from* the data	Refine the question or obtain other data
Probe: Build a formal model	Model answers question, to describe, explain, or predict	Challenge model assumptions and structure (e.g., sensitivity analysis)	Revise model structure or inputs
Place in context: Interpret results and implications	Analysis provides specific, meaningful answers	Totality of analyses—effect sizes, accuracy, uncertainty	Revise analysis to provide a meaningful answer
Present: Communicate methods, results, significance	Content, style, attitude meaningful to audience	Feedback from audience	Revise presentation
Preserve and **post**: Make life cycle transparent, enduring	Data, code, and methods available	Assess sensitivities to release and possible restrictions	Make as open as possible; document restrictions

Core activities and critical reflections are adapted, in part, from Peng and Matsui (2015).

3.1.4 Commitments: life cycle, centered on analysis, subject to norms

I have taken a broad view in surveying a variety of motivations and definitions for data science. In consideration of this breadth and variety, I contend that data science entails 3 main commitments:

1. **We learn from data in the context of an overall life cycle**: posing rich questions about the world, amenable to rich methods; guiding how we generate, transmit, obtain, and prepare data; probing data to answer questions about the world; placing answers from data in context, mindful of assumptions and alternatives; presenting data-driven answers to audiences clearly and correctly; and preserving those answers and ensure that the entire life cycle is transparent, accessible, traceable and, to the extent possible, reproducible.
2. **Analysis centrally connects the life cycle of data**. We pose questions, prepare data, place results in context, and present answers informed by the variety of available analytic approaches. If we have methods for analyzing images, then we can ask questions that only images can answer. To interpret and communicate a risk or a rate of change seen in a set of data, we infer meaning from the analytic method. As further discussed below, we have many analytic modes available to us beyond traditional statistical methods, such as causal inference and machine learning.
3. **We judge data science approaches and claims by scientific norms**. Since data science is about extracting knowledge through analyzing data, then it should be judged by the same criteria that apply to extracting knowledge from observations. This commitment is familiar within statistical practice; it needs to become familiar with other data-analytic modes, including machine learning.

These 3 commitments unpack what it means to learn from data, and they set some boundaries around that practice. They point to the practitioner's responsibility for respecting context, respecting analytic intent, and respecting quality and rigor. They also point us in the direction of needed investments in resources and learning. These commitments are not, however, meant to imply that each person who practices data science individually carries out each activity. (See also section 5.3.) Let's examine those boundaries and some disciplines that are related to, but distinct from, data science.

3.2 What data science *is not*

Data science overlaps other disciplines and scientific practices. Furthermore, as National Academies of Sciences, Engineering, and Medicine (2018) notes, the practice of data science necessarily crosses disciplines. How does data science relate to statistics and other modes of data analysis, to informatics, and to science in general? How should the practice of data science privilege science over technology and focus on meaning and rigor?

3.2.1 Data science is not statistics

Statistics is not the same as data science, though the field substantially overlaps with data science. Moreover, data science is not merely statistics dressed up with appealing marketing. I argue above that data science takes responsibility for the whole life cycle of data, connected centrally through analytic concerns. In this understanding, a statistician who limits their engagement solely to analysis and perhaps interpretation is not doing data science. A statistician who does analysis and engages the rest of the life cycle of data is doing data science—as is an epidemiologist, a sociologist, a microbiologist, or anyone else.

Donoho (2017) and Jones (2018) show that data science took shape as a discipline, in part, in reaction to the failure of academic statistics to focus sufficiently on pragmatic rather than theoretical concerns. This characterization cuts in 2 directions: While academic statistics might have shunned practical concerns, academic and applied statistics firmly root themselves in traditions of rigor and other scientific norms. On the other hand, machine learning and other analytic disciplines are not as firmly rooted. To be fair, academic machine learning—often located in computer science or information science departments—pays heed to rigorous mathematics, out-of-sample generalizability, and applied issues such as bias and fairness. But the traditions are not as deep, and the norms are not as strong.

Leo Breiman, the late UC Berkeley statistics faculty member and an early bridge-builder between statistical and machine-learning communities, said that he might advise a young person (in 2001), "Don't go into statistics." In the end, he would say, "Take statistics, but remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems." (Olshen 2001)

Andrew Gelman, Columbia faculty member and prolific blogger, wrote in 2013, "Statistics is the least important part of data science … Statistics is important—don't get me wrong … But it's not the most important part of data science, or even close." (Gelman 2013)

3.2.2 Data science is not data analysis (not even machine learning)

The field of statistics connects disciplines and practices for constructing and probing models grounded in probability theory and inference. This characterization holds for frequentist, Bayesian, and other approaches, whether the probability model is highly specified (as with parametric models) or loosely specified (as with nonparametric models). Many other approaches to data analysis might have a probability component that is not of primary concern or might have no formal probability component at all.

Machine learning has been described as the answer to the question, "How can computers learn to solve problems without being explicitly programmed?" In practice, computers "learn" to solve problems by looking for mathematically representable patterns in data (such as clusters or topic models) or by constructing mathematical tools to guess an output, given a set of inputs, modeled on examples that associate known inputs with known outputs. In these senses, machine learning is data analysis. As a field and collection of methods, machine learning overlaps substantially with statistics, distinguished by its emphasis on finding patterns and making predictions rather than constructing models that directly represent data—even when a machine learning model is explicitly probability-based or a model's performance is represented using concepts from probability.

There is no bright line between statistics and machine learning, and many methods inhere to both disciplines. For example, classical statistics has traditions of cluster analysis, dimensionality reduction, and regularization, and machine learning uses Bayes's theorem and logistic regression for binary classification tasks. Although machine learning is often associated with complex models based on vast amounts of data, statistical models can be complex, with many model parameters, or they can be based on large amounts of data. Conversely, machine learning models can be simple or based on small data. Since machine learning models tend to emphasize predictive performance (outputs given inputs) rather than internal model structure, however, the largest data-analytic models in practice tend to use machine learning methods. Some deep learning models have billions, or possibly trillions, of model parameters. I discuss machine learning, along with artificial intelligence, at greater length below in section 8.

Other modes of data analysis beyond statistics include causal inference, geospatial methods, econometric methods, and compartmental and agent-based modeling. Pearl (2009) explicitly characterizes causal inference as extrastatistical; without additional strong assumptions, no probability model can inherently represent causality. Structural equation models and inverse probability weighting can help to disentangle a causal signal from random noise, subject to those extrastatistical assumptions. Geospatial and econometric methods also often use probability components, for example, to accommodate correlations in space or time, but they wed those components to other concepts. Compartmental and agent-based modeling might or might not use empirical observations, but when they do, any probability components for solving or simulating systems also extend beyond strictly probability-based models.

From the perspective of data science, the life cycle of data can center on any or all of these analytic disciplines, not just statistics. This perspective covers 2 of my 3 commitments of data science: the life cycle and the central concern of data analysis. The third commitment, to scientific norms and rigor, obtains when the practitioner acknowledges and respects the norms inherent to the various models of analysis. For machine learning, for example, these norms include out-of-sample generalizability and model robustness and stability. Thus, not only is it wrong to characterize data science as an enhanced form of

statistics, such a characterization risks failing to hold other analytic modes to similar expectations of rigor and norms.

Much more could be written about whether machine learning or other modes of analysis reveal "meaning" in data, as some data science definitions seek to do. For now, I note that all such modes, including statistics, can be subjected to various methods for interpretation in terms of model structure and the relationship of models to input data. Furthermore, such interpretations and accompanying explanations warrant careful critical evaluation in view of a broad swath of scientific norms, not least because an apparent interpretation or explanation can itself be an illusion regardless of the method of analysis.

3.2.3 Data science is not informatics

Public health informatics is the systematic application of information and computer science and technology to public health practice, research, and learning. Informatics applies technology to obtain, store, and use information. Per Savel and Foldy (2012), it concerns "the how and why of technology and systems versus the common what and where of information technology … the *integration and proper application* of technology and systems to get data rather than just the technology and systems" (emphasis added) … "frequently the application of standards and structure that help with meaning before data science gets to it." My colleague Brian Lee has said (personal communication), "Informatics is all the work of understanding and making data available to determine meaning". Thus, we can see that informatics and data science overlap, especially regarding data wrangling, movement, accessibility, and scale, but the fields take different orientations: Data science seeks meaning from data, empowered by informatics to work with data. The disciplines in each field are important, and one can practice the collection of disciplines across those fields. It is important, however, not to conflate them nor to treat one field as a subset of the other.

3.2.4 Data science is not just good science

Much of scientific practice in public health uses, or purports to use, data that come from observations about individual health status and other aspects of the world. When public health science uses data, then it should conform to scientific norms and rigor, much as I have claimed data science should. Does it follow, then, that data science is just good science? The answer is *no*, both because data science inherits some commitments that do not apply broadly to the practice of science and because good science entails commitments that do not apply to data science.

By "good science" I mean, in brief, all those practices for building and organizing knowledge about the world through the methods and values of experiment and observation, neutrality, rigor, transparency, empiricism, reproducibility, minimizing subjective bias, and so on.

On the expectation that theoretical science need not use data at all, we can omit theoretical science and narrow our question to applied science. Even narrowed in this way, we can conclude that not all applied science uses data. While all applied science depends on observation and precedent, those contexts need not entail *data* in the sense of observations represented in a way that we can subject them to further analysis. For example, without implicating data, a scientist can classify an organism through observation or conduct a qualitative review of published literature to structure arguments and conclusions about the state of knowledge in a specified domain. Next, not all applied science that uses data, uses it well. While we could argue that applied science that uses data poorly is not "good science", we should also

acknowledge that applied science can consist of good and bad components in which the poor use of data does not undermine the entire project. Finally, and pivotally, not all applied science that uses data well also takes responsibility for the integrity of the life cycle of data and for connecting that data life cycle to how questions are posed, data obtained, analysis performed, results placed in context, methods and results presented, and the whole process preserved. Just as a statistician can conduct a rigorous data analysis without connecting that analysis to the life cycle of data, any other scientist can engage in portions of the scientific method, including portions of the life cycle of data, without having applied data science. Data science entails taking responsibility for the integrity of the life cycle of data—across the core activities of data science—in a way that does not apply to the full breadth of "good science". Without question, doing data science well overlaps with doing good science, but it is important not to conflate them.

The life cycle of data is consistent with, but not synonymous with, the scientific method. This distinction between "good science" and data science matters because the distinction informs how we do data science, which in turn differs in emphasis and kind from how we do good science. In particular, the technical and nontechnical skills that support the practice of data science, especially the skills for locating data analysis as a central focus in the life cycle of data, do not generalize or scale to the wholesale conduct of good science.

4 How

4.1 Foster a progressive culture

If we think of data as an asset, how does that asset produce value within our mission and our available resources? How can public health scientists who care about data keep up with fast-moving methods, tools, and technology for learning from data? A progressive culture intentionally orients itself proactively and not only reactively, toward advancement and not just tradition. While a progressive culture encourages innovation, more importantly this community continually expands the set of tools for doing good things with data and applies judgment for selecting among familiar or conventional options as well as unfamiliar or unconventional options. A progressive culture for data remains rooted in history, continues to learn from old data in new ways, anticipates the future, and handles evolving demands to keep up with fast-moving methods, tools, and technology.

Here I sketch a vision for fostering the practice of data science across disciplines and levels of experience by describing 3 components of a progressive culture for data:

1. developing know-how through data-savvy **technical skills** to bridge domain knowledge and methods for learning from data,
2. cultivating data-wise **nontechnical skills** to drive problem-solving with data (start inquiry, keep it on track, and deal with obstacles), and
3. participating in an empowering **community** of mentors and peers to enable self-learning and foster practical wisdom.

After describing technical and nontechnical skills, I map those skills to an expanded treatment of Peng and Matsui's core activities of data science. Then I sketch functions and roles in an empowering community. In the next section, on *who* does data science, I more fully articulate those functions and roles along with the level of technical and nontechnical skill needed for each.

4.2 Foster technical skills

What skills are required to practice data science rigorously? What about those who want or need to practice data science well but who don't need to be expert data scientists? The core activities of data science call for knowing how to pose a good question, how to compile and prepare the data to answer the question, and how to extract, interpret, and convey meaning from the data in answer to the question. Data science skills are often represented (for example, by NIST Big Data Public Working Group (2015)) as the cross-disciplinary intersection of 3 sets of technical skills that cover these core activities: domain-specific skills for posing a good question and interpreting and explaining results; computational skills for corralling, structuring, and applying algorithms to data; and data-analytic skills, including communication skills, for extracting, interpreting, and conveying meaning from data.

Domain-specific skills cover any subject about which one might want to use data to answer a question, including public health, epidemiology, medicine, microbiology, toxicology, and anthropology. In practice, different fields often call for different norms for rigor. Epidemiology establishes modes to reason about bias and causation. Medicine institutes norms for assessing preventive and therapeutic efficacy and effectiveness. Microbiology and toxicology work out how to establish and measure the presence of a pathogen or toxin for ascertaining individual cases.

Computational skills cover how to use theory, hardware, and software to represent and work with data of various structures, sizes, shapes, and speeds, to enable transmission and exchange of data and other information among systems and among users, to implement algorithms for working with and analyzing data, and to manage the efficiency of all of these undertakings. How should textual information, audiovisual information, and other types of information be represented for further computational access and use? Having obtained and stored various types of information, how should they be processed and arranged in preparation for analysis? How can algorithms for working with data make the best use of available computational resources, such as memory and processing time? How can algorithms be implemented to work with increasing volumes, speed, and complexity of data while ensuring that computational results are available in an acceptable amount of time and other limitations? Computational skills cover or overlap programming, data-wrangling, software engineering, statistical computing, and methods for breaking up high-volume, high-velocity, or otherwise intensive data problems into smaller pieces, processing them, and reassembling the output.

Data-analytic skills, as discussed above, encompass statistical methods, machine learning, and other modes of data analysis. Statistical modeling typically refers to using probability to think about how data might have been generated and then using data to figure out how we might separate a representation of something about the world (signal) from variability or uncertainty about that representation (noise). Statistical methods include simple summaries like means and medians and more complicated summaries like tables, regression models, and time-to-event models. Machine learning typically refers to asking whether we can find patterns within a set of data, like clusters of similar counties or patients or topics in a set of documents, or patterns that relate inputs to outputs based on examples, such as for predicting a

patient's disease status or prognosis from available insurance claims and billing information. Other data-analytic approaches include causal, geospatial, and econometric methods. These methods often overlap, and they often incorporate but don't always center on probability components.

I include **communication skills** primarily with data-analytic skills, because an analyst often has primary responsibility for interpreting, representing, and conveying methods and results. These skills include the ability to use verbal narrative, tables, and graphics to explore, develop, and tell a story that translates results into stories, decisions, and actions.

While data science is often represented at the 3-way intersection of domain-specific, computational, and data-analytic skills, it is also instructive to review their pairwise overlap. The combination of domain-specific and computational skills could encompass domain-specific software development, as with medical or laboratory applications. Domain-specific and data-analytic skills entail applied research, as in epidemiological applications. And computational and data-analytic skills overlap in statistical computing, machine learning, and other applications that implement mathematical algorithms and optimization.

We can roughly associate each core activity in data science with technical skill areas.

> **Pose good questions**. Domain knowledge is needed to state and refine a good question. An awareness and understanding of a broad and rich variety of data-analytic methods can also enhance the kinds of questions that one could pose.
>
> **Prepare data to address those questions**. Domain knowledge informs what to measure or assess, and computational skills inform how to obtain, organize, store, transmit, extract, and transform data. Data-analytic skills support assessments of whether the data can answer the question.
>
> **Probe the data** through rigorous analysis. Building a formal model depends primarily on data-analytic skills, supported by strong computational skills for implementing the analysis, especially when working with complex data or methods.
>
> **Place analytic results in context**. Interpreting models depends on the data-analytic skills to construct them and to critique model-related assumptions, as well as domain knowledge to place the results in context of what is already known or perceived about the domain subject.
>
> **Present methods and results**. Communication draws on data-analytic skills for correctly describing methods and formal results, as well as domain knowledge for correctly describing and relating to subject-matter.
>
> **Preserve the entire life cycle**. Predominantly, computational skills support procedures for openness and traceability, including preparation of data and code for restricted or unrestricted sharing.

Of course, it is very likely that every core activity will draw on all 3 types of technical skills.

4.3 Foster nontechnical skills

To practice data science well and to keep up with constant change, it is not enough to focus on technical skills and knowledge. Technical skills cover the know-how for answering good scientific questions rigorously using data, but technical skills have limits and can become obsolete as methods, tools, and technology advance.

Nontechnical skills (sometimes called "soft skills") are personality traits, goals, motivations, and preferences that are valued in an applied domain. For example, collaboration and communication call for interpersonal skills. In addition, many sources (such as Davenport and Patil (2012)) emphasize that those who practice data science should be passionate, curious problem-solvers. Here I pay special attention to traits that support, and even empower, learning from data through its life cycle, centered on analysis and subject to scientific norms. In other words, I describe and unpack the traits that flow from a love of knowledge and learning, followed by traits that support the ethical conduct of data science.

4.3.1 Intellectual character

In a progressive culture for data, fostering intellectual character can cultivate responsible learners and inquirers who are better able to keep up with fast-moving methods, tools, and technology. Intellectual virtues flow from a love of knowledge and learning, aiming at "cognitive goods", like truth and understanding (King 2014; see also Costa and Kallick 2008). In data science, the practitioner seeks understanding mediated through data and the life cycle of data. Intellectual virtues animate scientific practice in general and data science in particular. This subsection draws heavily on the work of Jason Baehr, especially Baehr (Baehr 2013a; Baehr 2013b; Baehr 2015).

Baehr (2015) describes 3 dimensions of an intellectual virtue: First, an **ability or skill** specific to a virtue and leading to action. For the trait of curiosity, this skill is asking good questions. Second, the **motivation or commitment** to apply the virtue. With curiosity, the motivation is to ask good questions because of a love of knowledge or learning. Third, the **judgment or sensitivity** to know when and how to exercise virtuous abilities or skills. With curiosity, the sensitivity concerns when to start, continue, pause, or stop inquiry. In addition, each virtue can be seen as the mean between vices—too little of a good thing and too much of a good thing. Too little curiosity is the vice of indifference, while too much curiosity is the vice of obsession or fixation.

We can identify several intellectual virtues by examining the dispositions associated with stages of inquiry when approaching an objective: starting to learn and heading in the right direction, keeping the inquiry on track, and dealing with obstacles. For each stage of inquiry described below, I list stage-related virtues and use the pipe character ("|") to delimit each virtue's corresponding ability or skill, motivation or commitment, judgment or sensitivity, and vices representing too much or too little of the virtue.

Start learning and head in the right direction. A few intellectual virtues relate to how to start learning or start an inquiry and ensure that it heads in the right direction: In addition to curiosity, intellectual autonomy is the ability to think for oneself, and intellectual humility is the ability to admit one's limitations—to know what you don't know.

> **Curiosity**: Ask good questions | to learn | discerning when to start, continue, pause, or stop the inquiry | mediating between indifference and fixation.
>
> **Intellectual autonomy**: Think for oneself | to achieve independent thought or self-assuredness | discerning when to yield to others or differentiate from others | mediating between conformity and radicalism.
>
> **Intellectual humility**: Admit one's limitations | to recognize what one is able or unable to do or to locate oneself in the context of others' interests | discerning when to assert oneself or to stand back | mediating between arrogance and self-deprecation.

Keep the learning process on track. After starting an inquiry, a few intellectual virtues assist the learner in keeping on track: Attentiveness is the ability to engage, to look and listen, and to notice details. Carefulness is the ability to spot and avoid errors. Thoroughness is the ability to go deep in order to gain understanding and to explain.

> **Intellectual attentiveness**: Look and listen | to remain alert to details | discerning when to tune out or to focus more intently | mediating between distractedness and preoccupation.
>
> **Intellectual carefulness**: Avoid errors | to assure or control the quality of one's output | discerning when to ease up or to double down on quality control | mediating between sloppiness and perfectionism.
>
> **Intellectual thoroughness**: Go deep to understand | to ensure sufficiently complete coverage or treatment | discerning when to fill gaps or let well enough alone | mediating between superficiality and meticulousness.

Deal with obstacles. Even on track to learning, one is likely to encounter obstacles. A learner benefits from intellectual virtues that help work through or around obstacles: Open-mindedness helps to think outside the box when confronted with a challenge to solve. Courage helps to be bold and to take intellectual risks. Flexibility helps to adapt as needed. Tenacity or perseverance helps to embrace struggle while working through a challenge.

> **Open-mindedness**: Think outside the box | to consider new or unfamiliar ideas and seek diversity and inclusion | discerning which ideas to dismiss or to entertain an idea | mediating between narrow-mindedness and gullibility.
>
> **Intellectual courage**: Take intellectual risks | to allow for bold action despite potential for failure | discerning when to tolerate more or less potential for failure | mediating between cowardice and foolhardiness.
>
> **Intellectual flexibility**: Adapt as needed | to allow for change, especially for improving outcomes | discerning when and how much to stand firm or alter activity | mediating between intransigence and suggestibility.
>
> **Intellectual tenacity**: Carry on | to continue toward learning objective, even when challenged | discerning when to persist and when to stop trying | mediating between fickleness and stubbornness.

Intellectual virtues can conflict with each other. For example, courage can conflict with humility when the drive to take an intellectual risk runs counter to the limitations of one's abilities (when one's reach

exceeds one's grasp). To navigate these conflicts, the good learner or thinker is aided by the mediating virtue of **practical wisdom**. This trait allows the inquirer (phronimos, per Baehr (2013a)) to grasp which intellectual activity is most valuable for attaining one's goals. Recall that Baehr (2015) identifies one dimension of an intellectual virtue as judgment or sensitivity about when and when not to exercise that virtue. Practical wisdom undergirds this dimension, and it allows the good learner or thinker to take suitable action when intellectual virtues conflict. (See Turri et al. (2021) and Baehr (2013a).)

Even if, as a good learner or thinker, you are motivated to apply intellectual virtues, you still need to develop the abilities and judgment that connect your motivation to right action. Intellectual virtues are developed by practicing them and by critical reflection on your own actions and dispositions. You get better at courage by practicing courage—by taking intellectual risks and learning from the consequences. You get better at humility by practicing humility—by owning your limitations and not shying away from them. You also develop or cultivate practical wisdom—to avoid vice and to mediate conflicting virtues—through practice and guidance and seeing them modeled by others. Curricula and other resources, including literature, computing resources, and mentors, can help intentionally and systematically cultivate intellectual virtues. I return to these ideas in the section on learning data science in community.

Just as we associated each core activity with technical skills, we can also associate the critical reflection process with nontechnical skills. Setting expectations corresponds to starting learning and heading in the right direction, which calls for curiosity, autonomy, and humility. Collecting information and comparing expectations with that information corresponds to keeping the learning process on track: attentiveness, carefulness, and thoroughness. And dealing with matched or mismatched expectations and information corresponds to dealing with obstacles: open-mindedness, courage, flexibility, and tenacity.

Table 3. Intellectual virtues, by stage of inquiry

	What: skill	*Why: drive*	*How: practical wisdom*	
Virtue, by stage	Skill or activity	Motivation or commitment	Judgment or sensitivity	Mediating between too little and too much
Start learning				
Curiosity	Ask good questions	learn about the world	when to start, continue, pause, or stop the inquiry	indifference / fixation
Intellectual autonomy	Think for oneself	achieve independent thought or self-assuredness	when to yield to others or differentiate from others	conformity / radicalism
Intellectual humility	Admit one's limitations	recognize what one is able or unable to do	when to assert oneself or to stand back	arrogance / self-deprecation

Keep learning on track				
Intellectual attentiveness	Look and listen	remain alert to details	when to tune out or to focus more intently	distractedness / preoccupation
Intellectual carefulness	Avoid errors	assure or control the quality of one's output	when to ease up or to double down on quality control	sloppiness / perfectionism
Intellectual thoroughness	Go deep to understand	ensure sufficient coverage or treatment	when to fill gaps or let well enough alone	superficiality / meticulousness
Deal with obstacles				
Open-mindedness	Think outside the box	consider new or unfamiliar ideas	which ideas to dismiss or to entertain an idea	narrow-mindedness / gullibility
Intellectual courage	Take intellectual risks	allow for bold action despite potential for failure	how much potential for failure to tolerate	cowardice / foolhardiness
Intellectual tenacity	Carry on	continue toward objective, even when challenged	when to persist and when to stop trying	fickleness / stubbornness

4.3.2 Ethics and values

Where intellectual virtues connect a love of knowledge and learning to the practice of asking and answering questions, ethics and values promote behaviors to achieve other goods, including trust, equity, and fairness.

We seek to protect personal privacy, and to balance privacy and utility, with specific behaviors throughout the life cycle of data: posing questions that do not raise undue risk to respondents; obtaining, using, communicating about, and sharing data in ways to limit risks to privacy and confidentiality; interpreting and communicating findings in ways that respect other rights and the welfare of the subjects of analysis; and promoting openness, transparency, and other aspects of data utility to make the overall process, methods, and final products available for scrutiny.

Further considerations concerning ethics and values in the practice of data science stem from the conduct of research involving human subjects, the conduct of public health surveillance, scientific integrity, and public service. Many of these considerations pertain to data, data systems and informatics, and data analysis. They go beyond privacy and confidentiality to justifications for gathering information; for balancing benefits and harms, burden and utility, access and security; self-determination and substantive engagement; justice; duties to limit collections and to use what is collected, and responsibility to avoid fabricating, falsifying, and plagiarizing. These duties are covered extensively elsewhere and are often implemented through regulation, policy, checklists, and other forms of guidance. (See, for example, CDC's Office of Public Health Ethics and Regulations and Privacy Program.)

In a progressive culture for data, we *value* data because data help us to learn things about the world and to make informed choices about how we interact with the world. We *value* innovation and technology insofar as they help us to continue expanding the means for doing good things with data, but we do not

seek innovation or technology as ends in themselves. An extensive set of tools gives us the broadest options for doing good things, so we remain open both to the unfamiliar or unconventional and the familiar or conventional. Based on these values, we practice pragmatic, principled pluralism by exploring and using wisely and well all methods that can help achieve technical excellence to learn from, about, and with data. Principled pluralism allows honest disagreement about methods, results, and interpretation.

4.4 Foster community and leadership

Community and leadership form the essence of *culture* in a progressive culture for data—so essential that I defer the full discussion to the next section (5). In this section, I sketch *how* community and leadership enable the practice of data science along the following dimensions:

> **Learning**. Community supports learning about data, and learning how to do data science, by centering on learners. Learning can follow formal curricula and be encouraged in structured programs, but substantial portions of learning occur in informal settings. Learners benefit from interactions with peers and mentors. Mentors benefit from meta-mentors. Advocates influence, guide, and support the learning-oriented community.
>
> **Doing**. Community supports the practice and profession of data science by giving everyone who wants to do good things with data the resources to do so. Data science learner-practitioners with basic or intermediate data skills come from any discipline to do good things with data. Expert practitioners go deep on data science methods and guide practitioners to proceed with rigor and stand behind their work. Managers supervise practitioners and experts, to ensure that they have the resources and direction that they need to achieve good things with data. Lay advocates, as persons literate in the value of data, work in community with practitioners, experts, and managers and help ensure supportive resources to enable the practice of data science.
>
> **Staffing**. Community creates and ensures the capacity for data science through staffing and career development by all available means to recruit, retain, organize, and develop learners, doers, and supporters. This includes identifying and building on the data science potential among existing staff, finding and using mechanisms to bring on learners as well as other federal and nonfederal staff, and organizing formal and informal structures for staff to learn, do, manage, and support data science effectively.
>
> **Leading**. In a progressive culture for data, leadership aims toward and flows from practical wisdom. Leadership is part of the practice of data science, and not separate from it. Leaders include practitioners, experts, managers, and laypersons, regardless of their career stage, job title or series, credential, or location in the hierarchy (subject to some structural constraints in the federal system).

Within and across these dimensions, an individual can carry out more than one role or function. For example, in the *doing* dimension, a practitioner can serve as both an expert and a manager. The same person can also serve as meta-mentor and advocate in the *learning* dimension. In the next section, I expand on the roles and functions that align with these dimensions.

5 Who

5.1 Who gets to do data science?

Everyone who wants to do good things with data should have the intellectual support to do so, as long as they proceed with rigor and stand behind their work. I first formulated this credo in 2016. I asked, "Who gets to do data science?" to express both empowerment (who has the ability) and privilege (who has the authority). I was proclaiming that one need not be an expert in statistics or computer science to perform well when working with data. Indeed, nonstatisticians can, and often do, perform great work with data.

I believe this credo because of my own experience mentoring, coaching, and advising learners through a variety of CDC or CDC-adjacent programs, including fellowship and student internship programs, with undergraduates, masters and doctoral students, and postgraduate learners. For the first few years in my CSELS ADDS role, I emphasized the learning-oriented, empowering component of data science: "*every-one* who wants to do good things with data". In my mentoring experience, a learner's specific analytic or technical background has been a poor predictor of how well they would do, especially in programs that don't recruit specifically for previous analytic or technical education. For example, I have worked with several physicians who had no specific statistics background, who went on to execute superb analyses, some even winning awards. In each case, they proceeded with rigor and stood behind their work. My role was merely as mentor; they took up the challenges of doing data science. Conversely, some learners with apparent analytic background either shunned rigorous analysis or fumbled badly. CDC programs *can* select for prior technical or analytic experience, but I don't believe that CDC programs *need* to do so.

Who gets to do data science? I can restate the question echoing my credo as follows: Who wants to do good things with data, proceed with rigor, and stand behind their work? I can restate the question again, echoing my working definition of data science: Who will line up tools to ask and answer good questions rigorously using data? I have the same answer for all 3 versions of the question. **Self-learning problem-solvers get to do data science**: people who connect a love of knowledge to self-learning and solving problems, people who ask thoughtful questions, pay close attention to details, honestly acknowledge what they don't know, probe for deeper meaning, and persist in the face of obstacles. (See also Baehr 2013b.)

In 2016, I felt energized to tout such an empowering message focused on learning rather than specific disciplines. I slowly realized that this message was incomplete. While I situated self-learning problem-solvers in learning-oriented communities with mentors and advocates, I needed to say more about who learners, mentors, and advocates are, where specific technical and nontechnical skills fit in, how that community operates beyond learning. So from mid-2016 through mid-2019, my primary formulation transmuted from *Who gets to do data science* to the more expansive and inclusive *Who participates in a progressive culture for data*. I began acknowledging that nonexperts who do good things with data often need guidance from experts to empower those achievements. Furthermore, nonexperts and experts alike need support and other resources from other members of the progressive culture for data.

I also believe that CDC has a substantial, untapped well of potential among existing staff for doing good, and better, things with data. In other words, CDC could achieve, or make great strides toward, a progressive culture for data with the right attention and direction regarding learning, doing, staffing, and leading. I have not seen CDC as a whole make those moves, though pockets here and there show promise.

Who gets to do data science? Echoing NIST, the Office of Personnel Management (Reinhold 2019) says, "Practitioners with sufficient knowledge in the areas of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle. Overlapping skills including data analysis, analytical applications, big data engineering, algorithms, domain expertise, statistics and machine learning. [They] use expertise in one or more of these domains to solve complex data problems." This answer lacks poetry but specifies a few details in terms of knowledge, skills, and work activities. These details are useful for creating staffing strategies, occupational qualifications, and position descriptions.

Finally, who leads in a progressive culture for data? Everyone should get to.

5.2 *Learning* in a progressive culture for data

In 2018, I added the following to my growing collection of personal mottos: "Less training, more learning." By this I meant that our culture should explicitly recognize and support personal initiative and self-direction as ways—in my view the most important ways—of gaining knowledge and experience for solving real problems, especially after entering the workforce. Community supports learning about data, and learning how to do data science, by centering on learners. Learning can follow formal curricula and be encouraged in structured programs, but substantial portions of learning occur in informal settings, as through reading, self-guided learning, and interaction with peers. In my experience, a culture that overemphasizes training risks undervaluing the full gamut of the ways that learners learn.

A focus on learning respects the agency and responsibility of the learner, who must take an active role not only in receiving instruction but in practicing and honing what is learned. A focus on learning further opens the way for various models and modes of learning, including self-teaching through reading, crafty on-line searching, independent tutorials, and experimentation. Self-guided and experiential learners need guidance from others, mentoring, and help identifying or filtering through material that can assist in learning. Learning concerns not only individual development but also better serving the shared mission, for example, by asking better questions, working better with data sources or structures, and communicating rigorously and clearly to a variety of audiences. Learning should not focus only on technical skills but also on the means to exercise data acumen, good sense, and judgment.

Beyond developing skills among current staff, we need an agency culture that provides intellectual support to everyone who wants to do good things with data, whether they already have the skills or not. For those who already have skills, it's a matter of supporting good practice, supporting continuing development, and encouraging that they support others. For those who lack skills, it's a matter of providing support that is more oriented to learning new skills or to adapting skills from other areas.

5.2.1 Relational learning

I start here with the relationship between learners—typically fellows, students, or early-career scientists—and mentors, since this relationship formed the initial impetus for my entire conception of a progressive culture for data.

Mentors create supportive conditions to guide other scientists to learn about data and to learn from data. A mentor is responsible for guiding technical skills to encourage a personalized direction for self-learning, sometimes as specific as skills for managing relational databases, creating graphics that smooth binary outcomes, modeling seasonality, or exploring categorical data using mosaic plots.

In my experience, *experience* has been more important for learning than specific subject-matter knowledge. I typically guide the learner through clear thinking and critical reflection more than through particular methods. Mentors also model intellectual virtues. They show by example as they openly practice curiosity, courage, humility. Mentors also create regular opportunities for the learner to practice intellectual virtues by stimulating curiosity, rewarding courage, and fostering humility. In some ways, these nontechnical skills are more fundamental to scientific practice than technical skills are: learning how to learn and, more than that, how to exercise judgment regarding level of effort, intensity of exploration, extent of experimentation, making interesting mistakes. Practical wisdom matures in part through guidance and in part through reflecting on one's own lapses in intellectual virtue. A mentor can help the learner with both.

In the context of education, Baehr (2013b) explained that educating for intellectual character growth is **personal**, because it involves thinking of learners as persons whose basic beliefs, attitudes, and feelings about knowledge and learning also matter critically to the quality of their education; it is necessarily **social** or **relational**, because personal change and growth occur most readily in the context of trusting and caring relationships; and, it is **reflective**, because it involves reflecting on and discussing with learners the value of thinking and learning, regularly pausing to identify or reflect on the significance of what is being learned. The mentoring relationship responds to and nurtures a love of and interest in thinking, learning, answering good questions. "Intellectual virtues flow, not from a desire for praise or approval, but out of a genuine interest in thinking and learning." (Baehr, 2015)

As a mentor, I often explain my thought process to the learner, as messy and nonsensical as that thought process might be. Thinking out loud with a learner, usually when I have no idea whether I'm making sense, is real and honest. It is also an exercise in vulnerability. In turn, the mentor must earn the trust and respect of the learner, so that the learner is able to take risks to make interesting mistakes in their own thought processes.

CDC needs to do a better job of finding and supporting mentors in informal and formal ways. In the discussion below on staffing and capacity for data science, I propose some approaches through articulating competencies and accounting for performance.

In a robust culture supported by mentoring, the mentoring relationships go in several directions. I started with the relationship of the mentor to the learner. In communities of peers, learners can mentor each other in the same personal, relational, and reflective ways as mentors do with learners. In addition, mentors could benefit from guidance and wisdom from experienced mentors, or what I call "meta-mentoring". Finally, advocates influence, guide, and support the learning-oriented community.

Learners ask questions, solve problems, reflect critically on process, and improve their skills. They take responsibility for self-learning, seek mentorship, and practice technical skills and intellectual virtues. Learners take risks and make interesting mistakes, from which they learn how to exercise judgment. Learners use data responsibility to improve the world. Learners participate in community through fellowships and other learning programs, communities of practice, scientific workgroups, interest groups, and user groups. The Statistical and Machine Learning Community of Practice was a community for learning from data, an example grass-roots network of learners that brought together members of scientific workgroups, user groups, and other groups.

Advocates influence the practice and profession of data science, promote and reward commitments to data and to self-learning, remove needless barriers (e.g., technology-related procedures), support instructive failures and interesting mistakes, encourage those who practice data science, and uphold those who profess data science. Advocates include managers, decision-makers, associate directors, directors, and others.

5.2.2 Formal curricula and structured programs

A progressive culture for data can promote more systematic or standardized learning through formal curricula for technical skills and for nontechnical skills. Numerous vendors now offer courses on a wide variety of computational and data-analytic skills. CDC makes many of these available through OCIO, CDC University, programs like Advanced Molecular Detection, and sessions organized by CDC workgroups and user groups. In addition to the endless options for courses on technical skills, curricula are also available for nontechnical skills (e.g., Educating for Intellectual Virtues), though they are not as readily available.

CDC has a rich, long, and successful history of structured, experiential learning programs, including some for professionals outside of CDC. Until recently, these programs have not intentionally and explicitly addressed data science. The EIS program has addressed analytic skills in limited ways. Informatics and prevention effectiveness programs have specific, narrow technical areas of focus. Participants in these programs regularly benefit from communities of mentors and peers, but those have also not purposefully addressed data science.

Some recent developments demonstrate incremental shifts: The EIS program has piloted and continues to provide mentoring to EIS fellows for advanced analytic projects. The fellowship programs have also piloted efforts to create teams of fellows from different programs for intentionally interdisciplinary work. Recent modernization initiatives have sponsored a few fellows to focus on data science. The Data Science Upskilling (DSU) program launched in 2019 with about a dozen teams of incumbent federal staff and fellows, each focusing on a primary project, on-demand, online courses, and cross-team activities on 5 components of data science: statistics, machine learning, computing, visualization, and ethics.

DSU allows federal staff and other learners to set aside time to go deep on data and on methods that are new or unfamiliar to them, as well as time for trial and error. The program is predicated on explicit organizational support, including from supervisors, as well as structured leadership and access to experts, learning resources, technical tools, and fellow learners. Participants refine existing skills or learn new skills in analytic methods, software such as R, Python, and Power BI—importantly, not limited by their prior or primary occupational series or disciplines. They generally learn how to establish clearer

boundaries on their motivating data science projects, develop workable (if preliminary) solutions, and establish a community of practice. Most of the methods and tools used in DSU are not new, though they might be unfamiliar or uncommon within CDC's general culture. The program brings both an overarching purpose and specific value to learners and their programs by focusing on specific, mission-oriented problems. They thus establish, expand, and apply methods, tools, and technology available for CDC to use rigorously. Furthermore, they enrich their own and their teams' ability to adapt to fast-changing contexts: newer questions, less familiar data sources, and less familiar methods and technology, all of which comes close to my vision for the motivation and disposition of a progressive culture for data.

5.3 *Doing* in a progressive culture for data

A progressive culture for data centers on learning in order to empower the practice and profession of data science. Community supports the practice and profession of data science by ensuring that everyone who wants to do good things with data has the resources to do so. On this account, I see 4 primary roles in that culture: learner-practitioners (which I also call "learners and doers"), expert practitioners, managers, and lay advocates. These roles can change over time and overlap with each other and with the roles that I articulated above as learning-oriented roles (mentor, learner, and advocate). The roles capture the essential distinctions for the primary practical needs of that culture for doing good things with data.

Learner-practitioners with basic or intermediate data skills come from any discipline, not just computer science or statistics, to do good things with data, mindful of the full life cycle of data.

Description	seek to do good things with data come from any discipline, not just computer science or statistics
Data-oriented skills	basic or intermediate literate in data fundamentals, such as the design of data collection methods, data quality assurance, conventional flat, tabular and multidimensional, relational data, and common analytic methods interpret, communicate, and memorialize learning from data
Goals and approach	achieve, or work toward, data proficiency, building on fundamentals to work rigorously with more complex data or methods learn continuously and show how modern tools and methods solve modern problems mindful of the full life cycle of data

Expert practitioners achieve data mastery and go deep on data science methods and provide the intellectual foundation for good practice.

Description	provide the intellectual foundation for doing good things with data, aiming for scientific quality and analytic rigor master complex data structures or methods
Data-oriented skills	literate in advanced, contemporary methods for complex data structures or methods, such as high-volume or high-velocity data; analysis of patterns and predictions as well as inferences; visual and other methods interpret, communicate, and memorialize learning from data
Goals and approach	practice personal proficiency ensure that everyone who wants to do good things with data, can set norms for data-oriented practice and for learning from, about, and with data

	enable, guide, correct, and empower practitioners to proceed with rigor and stand behind their work mindful of the full life cycle of data

Managers supervise learner-practitioners and experts to ensure that they have the resources and direction that they need to achieve good things with data, now and in the future.

Description	give learner-practitioners and experts resources and direction to do good things with data
Data-oriented skills	data fluency, acumen, or proficiency how to assess scientific quality and analytic rigor of data-oriented solutions how to allocate investments in data-oriented learning and technology how to allocate data-oriented assignments
Goals and approach	foster and reward curiosity, invest in learning (not just training), encourage creativity and interesting mistakes hold practitioners and experts to account for producing knowledge learned from, about, and with data advocate for the means to enable practitioners and experts to continue increasing their capability, efficiency, and effectiveness

Lay advocates work in community with practitioners, experts, and managers as persons literate in the value of data to help learn things about the world.

Description	support doing good things with data, in community with practitioners, experts, and managers
Data-oriented skills	data fluency or acumen how to assess basic quality of data-oriented solutions how to allocate investments in data-oriented learning and technology
Goals and approach	help ensure supportive resources to enable learning and achievement

5.4 *Staffing* in a progressive culture for data

How does a progressive culture build and sustain the capacity to keep up with fast-moving methods, tools, and technology? How are people brought in, organized, and kept around?

Harvard Business Review headlined data scientist as "the sexiest job of the 21st Century" (Davenport and Patil 2012). Fast Company has called it one of the best 25 jobs in America (Dishman 2016).

5.4.1 Data science staff should be cultivated, hired, *and* outsourced

Amidst the mixture of excitement and marketing hype about data scientists, there's a recurring question about whether data scientists are recruited and hired from the outside or cultivated from the inside.

> Data scientists are hard to find and attract. … Data scientists are rare commodities. … What data scientists do—curate data, ask the right questions, build explanatory analytical models, implement the models into various applications—is simply not scaling at the pace of demand. (Millis, 2015)
>
> A prominent data scientist in Silicon Valley ... doesn't hire on the basis of statistical or analytical capabilities. … [He] seeks both a skill set—a solid foundation in math, statistics, probability, and computer science—and certain habits of mind. He wants people with a feel for business issues

> and empathy for customers. Then, he says, he builds on all that with on-the-job training and an occasional course in a particular technology. (Davenport and Patil, 2012)
>
> I believe you indeed learn data science on the job. It is true that data scientists should know [some specific technical skills] … And self-learners can catch up quickly …. But focusing only on people who call themselves data scientists is a mistake. (Van Cauwenberge, 2015)

In these 3 quotations, we sense that data scientists are hard to come by. Furthermore, since they need to keep up with fast-moving methods, tools, and technology, they need a firm foundation in technical and nontechnical skills as well as a disposition and self-sufficiency for continuous learning.

Federal workforce flexibilities afford a rich variety of staffing mechanisms and organizational options for achieving and sustaining an effective mix: career development among federal staff and other learners, recruiting new federal staff and learners, adding collaborators from academia and other partners, and acquiring data science services through contracts. This section provides a brief, opinionated summary narrowly focused on a few considerations. I organize the discussion around 6 broad, mutually exclusive segments:

1. Federal employees already on staff, including civil service and uniformed staff
2. To-be-recruited federal employees
3. Federal and nonfederal staff in learning programs, glossing over some nontrivial nuances distinguishing federal learners (e.g., some fellows hired under Title 42) from nonfederal learners (e.g., ORISE research participants)
4. Collaborators from academia or funded under a grant or cooperative agreement
5. Research and development contractors from federally funded research and development centers, university-affiliated research centers, and national laboratories
6. Commercial vendors

For ease of presentation in this section, I will sidestep some details that cannot be ignored in practice. For example, by learning programs, I mean staffing mechanisms such as fellowships, not coursework or programs like Data Science Upskilling. In addition, I include academic collaborators under the Intergovernmental Personnel Act or as Special Government Employees along with grantees, even though IPA funding is executed like a contract (acquisition) rather than a grant (assistance) and SGEs are technically civil federal employees.

Some of the material in this section corresponds to similar, more expansive discussions of the HHS Data Council's Data-Oriented Workforce Subcommittee (Gehrke et al. 2021; Wagner 2022). The subcommittee's reports present rich, thoughtful, comprehensive detail on staffing and organizing for data science in the federal workforce. While I provided some critical input to the subcommittee, the views that I present here are my own.

5.4.1.1 Federal employees

The federal government has been expanding options for classifying and developing federal employees to do data science. Historically, occupational series in science, technology, engineering, and mathematics (STEM) have represented narrow but workable disciplines, including engineering (0801), operations research (1515), mathematics (1520), statistics (1529 and 1530), computer science (1550), and to some

extent information technology specialist (2210); some of these have been combined into interdisciplinary positions, such as 0601/1530. Other scientific or technical series in social and behavioral sciences (0101), microbiology (0400 group), and health sciences (0600 group) have been used for positions that focus on research or analysis. Around 2016, I wrote CDC's first position description (in series 1530) that explicitly included machine learning.

In 2018, the Office of Personnel Management issued direct-hiring authority for STEM positions in economics, biology, engineering, physical sciences, and math fields. Then in 2019, OPM released guidance for adding parenthetical "(Data Scientist)" titling to several of these series (Reinhold 2019). Managers in the National Center for Injury Prevention and Control developed a set of standard "(Data Scientist)" position descriptions in several series and grades.

In 2019, CDC hosted a sequence of Future of Work (FoW) workshops to develop data science profiles. I appreciated the focused attention, but I perceived that the approach did not provide much latitude for existing federal staff who are experts in data science to influence the shape and direction of the effort. Data-oriented experts already in the workforce would have the direct experience to inform what is needed for doing good things with data—like existing supports and motivators (such as interesting problems and supervisory support) as well as persistent challenges (like barriers to nimbly using no-cost data science software). FoW's contract support staff could say something about the ways that industry improves its use of data, but they lacked awareness from within CDC's own culture of working with data. I also perceived that the approach risked conflating informatics with data science rather than clarifying the distinctions between them. On the benefit side, FoW fleshed out the concept of data fluency as a minimum competency for much of the federal workforce. In my schema above for *doing* in a progressive culture, managers and laypersons would best support the culture by achieving at least data fluency.

Finally, in late 2021, OPM issued the new data science occupational series 1560. The accompanying flysheet substantially emphasizes the defining importance of the life cycle of data, but it covers job activities that are diffuse or ill-defined enough that it will take special care to use the series effectively. I would have preferred improving the way that federal agencies use existing series, including flexibility with titling and combining series, but the development deserves to be taken seriously. Thus, CDC is actively working to develop qualifications, competencies, position descriptions, and other resources for recruiting and hiring data scientists.

Based on my experience with learners, user groups, and other early-career professionals at CDC, I believe that unrecognized and untapped potential already exists among incumbent federal staff and that CDC has so far failed to see and characterize this potential. To realize this latent capacity, we need to shift our thinking from traditional assessments of existing skills and traditional emphasis on training, to assessments of aptitudes and habits of mind and a radically different take on on-the-job learning that rewards self-learning and nurturing networks with peers and mentors. At least as important, we should be finding out from employees and learners with these experiences or interests what they need and want in order to do good things with data, rather than a narrow top-down focus on what only managers perceive—especially managers unfamiliar with the motivations, commitments, and prospects of data science. It makes little sense to me to talk about recruitment and retention without examining what makes prospective or practicing data science practitioners *want* in order to join the workforce and stay in it.

Tapping this potential also calls for a culture shift among staff themselves who do or can do data science. While it can be important, for example, for a statistician to maintain the professional identity of their discipline, statisticians (and computer scientists and others) need to see themselves as part of, rather than separate from, intentional cross-disciplinary engagement.

Turning to hiring, CDC faces well known challenges competing with other sectors. Aware of limited flexibility to enhance incentives for prospective hires, what nonfinancial incentives can CDC offer? Foremost, CDC's unique mission and public service already draws employees from many disciplines; that is, CDC appeals to many recruits' personal values. Second, if CDC cultivates a truly progressive culture for data—one that rewards a drive to learn as well as a drive to contribute—then CDC becomes that much more attractive to exactly the kind of people who can sustain and enrich that progressive culture. But the culture must be genuine, or else its attractiveness will fade.

Stepping back from the fine details of series and grade, whether federal data science staff are cultivated from within or hired from outside, the most important *operational* considerations pertain to competencies and performance. CDC needs practitioners who are able to do data science, whether as part or as all of their duties. As a side benefit of CDC efforts to flesh out series 1560, human resources staff have been developing a richly varied set of competencies, work activities, and proficiencies. Those supportive resources can and should shape other series beyond the new 1560. A 1530 statistician could adopt the more expansive data-analytic competency or the enriched competency for machine learning and artificial intelligence. And those same human resource concepts can and should be adapted into performance elements and statements, so that everyone who does data science can be accountable and rewarded for doing so. In addition to competencies and performance elements that arise from series 1560, CDC should also develop competencies for skills associated with intellectual virtues. Skills and competencies oriented to learning, practical judgment, and mentoring could also help to differentiate proficiency and grade within series and could (and should) apply to other scientific series.

5.4.1.2 Fellows and other learners

CDC manages or partners on dozens of structured learning programs on dozens of topics, open to persons with a variety of educational backgrounds. Fellows stimulate, and demonstrate CDC's commitment to, a vital culture of learning. CDC sometimes hires fellows as federal staff, often as an intentional career path. Although CDC fellows contribute to CDC's product, their primary purpose is to learn, not to augment staff.

Some CDC fellows focus largely on doing good things with data. More fellows get to do good things with data, whether they focus on data or not.

I believe that CDC should commit to helping CDC programs develop data science capacity through a focus on fellows, with the follow-on intention that other members of a learner's program unit can also develop their own data literacy or competency. Early drafts of the 2018 Public Health Data Strategy called for a ready response unit of data scientists who would work with CDC programs as needed. If such a unit were to be created, I recommended having it focus on working through fellows, such that the requesting CDC program would develop the capacity to address the data-oriented need rather than relying on outside staff to take care of it and move on. (As a side note, the Center for Forecasting and Outbreak Analytics largely goes the opposite direction from my recommendation, investing substantial data science resources in that center and rather than distributing them among other CDC programs.)

All these fellows need support from peers and mentors. To that end, CDC should foster mentoring as a supported competency, with accountability and reward through performance appraisal and other incentives. CDC should not, however, overinstitutionalize mentoring, because the role itself needs latitude and flexibility for fostering both technical and nontechnical skills.

5.4.1.3 Nonfederal collaborators

In addition to federal employees and learners, nonfederal collaborators serve some of CDC's data science needs, through joint research or other projects with academic or public health partners, through research and development organizations, and through commercial vendors. CDC often engages with academic and public health collaborators through grants and cooperative agreements or through a so-called mobility agreement under the Intergovernmental Personnel Act (which is administered more like a contract than a grant). Research and development organizations include federally funded research and development centers (FFRDCs, such as those operated by the MITRE Corporation or the RAND Corporation), university-affiliated research centers (UARCs, such as the Georgia Tech Research Institute and the Applied Physics Laboratory at Johns Hopkins University), and national laboratories (such as Oak Ridge National Laboratory and Sandia National Laboratories). Finally, commercial vendors include a vast collection of entities that bid to sell proprietary services to CDC under the Federal Acquisition Regulation.

These outside contributors can especially help by filling in gaps in CDC's own capacity for data science activities varying in discipline, skill, or scale that CDC can't address on its own. It's important for CDC, through advocates and managers, to strive toward building capacity among CDC's federal staff and to avoid assuming that only outside collaborators can do a particular thing (such as some forms of text or image analysis). As I've argued elsewhere in this essay, CDC's federal staff and learners likely have substantial, unrecognized capacity for extending CDC's data science capabilities into unrealized directions. It would be a mistake to outsource based on a faulty assumption.

Many needs *do* exceed CDC's current capacity. When it is necessary to turn to nonfederal collaborators, it becomes especially important to have enough expertise among CDC's federal staff (or at a bare minimum among trusted nonfederal partners) to ensure that contributions from nonfederal collaborators meet the intended need. How do we know if we're getting something useful, or what we need, from these collaborators? I have observed more than one project in which a nonfederal collaborator—sometimes academic, sometimes commercial—supplied a deliverable that the home CDC program was unable to evaluate. In those instances, greater data science expertise within the CDC program, or through another service or community within CDC, could have helped to ensure that the proposed deliverables would be worth the investment and that the actual deliverables met the need.

5.4.2 Data science staff should be organized to do data science

5.4.2.1 Organizing data science capacity

As described in previous sections, a progressive culture for data needs data science learner-practitioners (from a variety of disciplines), expert practitioners (specifically data science disciplines), managers, and lay advocates. A discussion that focuses only on experts is incomplete and short-sighted. Not everyone needs to be a data scientist to be empowered to do good or to be held to a high standard. And not everyone needs to be held to a high standard.

Should analysts or data scientists be integrated with staff from other disciplines or set apart? This question and the reality cut both ways: statisticians and other data staff are often set apart, and they often prefer it that way. In a largely remote, post-Covid workplace configuration, the organizational question comes down to 2 main characteristics that we can think of as *within* and *between*: Should data staff be placed in units that are homogenous or mixed with collaborating staff of other disciplines? How connected should data staff be across distinct units? I've seen some version of each configuration. The idea of grouping data scientists together seems like a wise way to manage limited resources, but in my experience, it fosters the notion that data scientists *ought* to be separate. During CDC's Futures Initiative in the early aughts, there was talk of putting all statisticians together in one center. It would have both made it harder to work with analysts and constrained the development of statisticians. I think that the most effective all-around configuration is to mix data scientists with other professionals so that there are other data scientists nearby and all data science practitioners in a division, say, regularly interact with each other to work through problems together and to learn.

5.4.2.2 Assessing data science capacity

CDC's ability as an agency to do data science depends on all the cultural components that I have listed above: intentional cultivation of learners as well as constructive support and direction for data science practitioners and experts that not only respects but also appeals to their know-how and their drive—both their technical skill and their nontechnical skills. An assessment of data science capacity needs to include, and go beyond, characterizing the aggregate set of those technical and nontechnical skills. It is important also to discern from people who do good things with data what they need and what they want in order both to continue and to improve. Let's break those ideas down by focusing on people who do data science (practitioners and experts) and people who directly empower, enable, or support them (managers). Because we are in the federal system, we will also need to distinguish federal staff, (nonfederal) learners, and other nonfederal staff. Finally, we want to characterize individual data science competencies as well as unit-level competencies at increasing levels of aggregation, such as teams, branches (collections of teams), and so on.

For staff who do data science, we want to know their proficiency and aptitude with technical skills in data analysis and computation as they apply across the life cycle of data. In my experience, the most effective way to discern technical and nontechnical skills and aptitude is for experts to see the skills in action, either prospectively or retrospectively: How well can the practitioner frame a problem? Work out what kind of data address the problem? Arrange, explore, and analyze the data using suitable tools? Correctly describe and critique the analysis? Demonstrate critical reflection throughout? Keep the activity directed toward the ultimate goal and deal with obstacles by acting on traits such as curiosity, attentiveness, perseverance, open-mindedness, and creativity? In addition to this broad set of competencies, we also want to know about particular strengths, for example, with programming in Python or deep learning or time series, as well as areas that warrant learning in order to address intended data science tasks. No one staff person needs to master all the technical skills, but they should have sufficient acumen to discern where their skills apply and where they do not.

For staff who supervise data science practitioners or lead projects that apply data science, we need to assess and edify their data fluency, sufficient to guide and empower practitioners and experts. Data fluency includes the ability to understand the components of the life cycle of data, how those components

relate to each other, the skills that each core activity calls for, and the intellectual traits and practices that support critical reflection and adaptation throughout the life cycle. Managers could be, but do not need to be, data science practitioners or experts. Where a manager lacks expertise, they will need the humility and wisdom to turn to experts. Furthermore, managers should demonstrate the skills needed to foster both learning and mentoring.

Some staff enable data science but do not practice it or manage those who practice it, such as information technologists or cloud engineers. For these staff, we also need to assess and edify their data fluency and their understanding of the life cycle of data, centered on analysis.

Data science is interdisciplinary. To ensure domain knowledge in addition to computational and data-analytic skills, we need to account for the combined set of skills and knowledge as groups of staff roll up into teams and other aggregated units. *And* we need to consider additional nontechnical skills for collaboration and negotiation. When considering a collection of staff and their joint mission, what are the specific strengths, weaknesses, and gaps in their collective ability to prepare, conduct, and communicate analysis? Do they have special strength or notable weakness in areas that could affect their ability to meet their mission, such as detailed knowledge of longitudinal claims data or time series methods? Assessment of larger units could especially call for evaluative expertise from outside the unit, as practitioners and managers might not be able to identify their own gaps.

A capacity assessment extends beyond individual and collective skills and traits, however. What do incumbent staff think that they *need* in order to do data science well and to keep doing it better? What do incumbent staff *want* in order to do data science well and to keep doing it better? Taking staff interests seriously can nurture morale and foster staff retention, but it also recognizes that staff are often the experts on supporting and bolstering their own capacity. Just as modernization should pay due heed to early-career professionals (the epitome of modern), and world-class analytics should pay due heed to data science practitioners and experts (the epitome of data-savvy), an assessment of data science capacity should pay due heed to the staff who actually do things with data. And yet these staff are often overlooked when they should be intentionally and directly engaged. Enlightened organizations often conduct exit interviews with departing staff, in part as an after-action analysis of the counterfactual: Now that you're leaving, under what conditions might you have stayed? In a progressive culture for data, practicing staff are continuously seen as partners, or even experts, in knowing how their unit can best function to keep up with fast-moving methods, tools, and technology. The culture, through supervisors and other governance, should continually engage with data-oriented practitioners proactively throughout their tenure, to empower them, facilitate their ongoing achievement, assure forward-looking resources, direct their efforts, and hold them to account.

5.4.2.3 Shaping and developing data science capacity

Assessment lets us know where we are and a little bit about how prepared we are to move in the directions that we want to go. But how do we shape and develop that capacity to do good things with data? This section outlines a way to structure the mission and focus of data science practice using 3 organizing rubrics predicated on concepts presented earlier in this essay. Those rubrics then translate into organizing principles, which lead to specific practices.

The 3 rubrics encompass (1) the core activities of the practice of data science, (2) the prepositional calculus of learning through data, and (3) a primary but fluid commitment to specific topics and services within the unit's mission.

> **Rubric 1: Core activities of the practice of data science**. Data science intentionally connects all core data science activities across the life cycle of data, as explained above, together with critical reflection at each core activity.
>
> **Rubric 2: Modes of learning through data**. We use data to learn about the world in at least 3 ways:
>
> 1. Learn about data, to understand the kinds of questions they might be used to answer.
> 2. Learn from data, in support of answering questions put to the data.
> 3. Learn with data, by using data to develop, explore, or evaluate methods.
>
> This prepositional calculus distinguishes assessing quality and utility from making inferences, which are in turn distinguished from a focus on methods themselves for learning about or from data.
>
> **Rubric 3: Topical goods and services**. The third rubric distinguishes the goods or services delivered as a result of engaging with the life cycle of data. Under this rubric, technical assistance to collaborating partners is an essential service, as are developing methods for ensuring data validity, evaluating case definitions, and collaborative analysis of population health.

We translate these 3 rubrics into organizing principles by linking them to data science activities.

> **Rubric 1 → Principle 1**. Link all data science activities to 1 or more of the core activities of the practice of data science, in the context and awareness of the other core activities. A team's primary skills and products should be organized around performing these core activities, with explicit notice of the scientific or business question of interest, the source(s) and transformation of data, and so on.
>
> **Rubric 2 → Principle 2**. Link all data science activities to 1 or more of the learning prepositions. Is the purpose of a given activity to understand the structure and attributes of some data source (learn about), to make claims about the world such as trends in asthma (learn from), or to get better at carrying out one or both of those purposes (learn with)?
>
> **Rubric 3 → Principle 3**. Identify data science activities as subject-matter inquiry, service, or both. Establish the value and priority of each of these purposes.

Bringing it all together, data science capacity, and the skills to support and expand that capacity, should be linked directly to priority tasks and interests, which in turn are tied to core activities, "prepositional calculus", and inquiry vs service.

Finally, identify put these principles in practice, as with the following examples:

Consider a team that focuses primarily on applying data science to the practice of syndromic surveillance. The team carries out tasks primarily related to data engineering (learning about data) and to supporting routine analysis of emergency department data (learning from data) for surveilling opioid overdose, hurricane-related morbidity, heat-related illness, Covid-19, and other conditions of public

health interest. Although the team's work covers all the core activities of data science, they focus primarily on the activities concerning obtaining, exploring, and analyzing data.

Example 1 (about). Among the list of proposed and actual activities directed to assessing or assuring data quality and utility, establish relative priorities and contingencies. These activities include at least the following:

- Develop, automate, routinize, and integrate measures of the health of data feeds, which in turn include characteristics such as completeness, timeliness, conformance to standards, and fitness for purpose; products include regular reports, on-demand reports, and summary dashboards.
- Assess value and limitations of data content, such as demographic data as received or as imputed.
- Assess mechanics and quality of auxiliary data sources, including laboratory and vital records.
- Assess mechanics and quality of using more than 1 data source for ecological analysis, such as merging at ZIP Code or county level, then aggregating post-fusion analytic results.

Example 2 (from). Among the list of proposed and actual activities directed to addressing specific, descriptive public health inquiry, establish relative priorities and contingencies. These activities include at least the following:

- Measure coverage and representativeness with methods and results that pass peer review.
- Characterize persons included in data sources, by demographic and (inferred) clinical factors.
- Develop and evaluate methods for monitoring specific conditions, integrating external knowledge of the epidemiology of those conditions, to detect temporal anomalies in a way that balances the utility of automated signals with the effort to attend to those signals.

Example 3 (with).

- Document methods for processing and learning about data sufficient to motivate independent re-implementation, in the interest of transparency, reproducibility, and intellectual credit.
- Advance the ability to process and use data from multiple sources.
- Advance the ability to develop data queries with a focus on conditions of interest, going beyond matching substrings by including machine-assisted record retrieval and semantic analysis.
- Advance the ability to incorporate temporal and spatial information for detecting anomalies, focused on specific conditions and jurisdictions of interest.

This framework for shaping and developing data science capacity does not independently invoke learning, because learning is an essential defining characteristic of a progressive culture for data. Rather, this framework orients learning toward the rubrics, principles, and practices of doing data science, both to prepare to do data science and to do it. In a truly progressive culture, learning that is not specifically oriented to a product or service can still serve an essential good, because it prepares the mind to see possibility and, one hopes, to keep up with it. Louis Pasteur said, "In the field of observation, chance favors only the prepared mind." (Pasteur and Vallery-Radot)

5.5 *Leading* in a progressive culture for data

Data science practitioners and experts—learners and doers—should lead CDC and ATSDR into the modern era through learning and advocacy. In a progressive culture for data, leadership is part of the practice of data science, and not separate from it. Leaders include practitioners, experts, managers, and laypersons, regardless of their career stage, job title or series, credential, or location in the hierarchy. Practitioners and experts must play a prominent, visible role in creating and leading a progressive culture for data in public health. As professionals who engage directly with data, we should ensure that the agency adopts, masters, and promotes an appropriately diverse set of tools and mindsets for using data to solve problems by showing how to learn from data and with data and by empowering others to do the same. Leadership continually shapes and sustains the culture of good data practice. As a community, we need to advocate to ensure that our interests and needs are folded into modernization initiatives as the agency becomes better tuned to meeting a modern mission. If learners and doers see leadership as separate from the practice of data science, then we risk leaving ourselves out.

We should invest in and take pride in personal technical excellence in doing good things with data—to construct, analyze, and interpret models of public health or administrative outcomes. To lead, though, we need to go further than technical excellence.

We should emphasize learning from data—unlocking meaning through analysis. We need to be as practical and solutions-oriented as public health is. And we need to be rigorous, to ensure that all data-analytic practices hold up to scrutiny, even when there's honest disagreement about methods or conclusions.

We should be principled pluralists on methodology. We see misapprehension about imputation methods ("making up data"), Bayesian methods ("too subjective"), and machine learning ("black box", "data dredging"). But all these methods and more can help us learn from data, if those tools are used wisely and well. This is largely what Leo Breiman was saying in 2001 (Breiman 2001).

And we should promote and praise good data-analytic practice, regardless of job title, credentials, or occupational series. Everyone who wants to do good things with data should have the intellectual support to do so, as long as they proceed with rigor and stand behind their work. This is as true for sociologists and microbiologists as it is for epidemiologists and statisticians.

We should provide leadership on how to integrate data science into interdisciplinary efforts and put data science on equal footing with other specialties. We need to be able to serve as an integral part of a team with collaborators from other backgrounds or disciplines, to apply and translate rigorous data science concepts for the benefit of collaborating scientists, and to explore and respect the rigorous application of concepts from other domains as part of collaborative undertakings. We must learn and practice methods for interpreting complex concepts for nonspecialists, without unduly sacrificing rigor.

That said, experts in data science are the foundation for good practice by practitioners, helping them to use data-analytic tools wisely and well. In a progressive culture for data, leadership aims toward and flows from practical wisdom.

We should hold fast to solid norms in how we learn from data as a basis for high-consequence decisions. The Covid-19 pandemic has been a time of high pressure, fast movement, substantial uncertainty, intense collaboration, and rapid turnover. It can be tempting, under these circumstances, to cut corners

on rigor—to try to get it done faster but to make concessions on quality. The opposite is needed: During times like this, and Ebola, and EVALI, and other high-consequence events, integrity is as important as ever. Data science practitioners with varied expertise have shown that we can achieve both high speed and high quality.

We should lead from every level. Front-line analysts lead by showing how modern tools and methods help solve modern problems. Team leaders and branch chiefs lead by fostering and rewarding curiosity, investing in learning (and not just training), and encouraging the interesting mistakes that come with innovation. Division and center leaders and associate directors help ensure that our infrastructure—our people, processes, and technology—can support modern and future tools and methods. In all of this, those who specialize in methodology and analysis lead from wherever they are, so that everyone who wants to do good things with data, can.

6 Redux: Who, how, what, and why

I have written about some ways that CDC and other organizations could better support a culture for doing good things with data, especially in view of fast-moving methods, tools, and technology. At CDC, we have a shared mission, a commitment to public service, and an intense, pragmatic need to draw on expertise across disciplines in multifaceted teams. We should surely hire great talent. But we also need to tend to staff who are already on board.

For our data generation, we must foster technical skills and a mastery of technique that allows scientists to extract information from data; we must foster intellectual virtues, including practical wisdom, that guide both inquiry and self-learning, and that enable scientists to ask good questions and to line up tools to answer those questions rigorously with data; and, we must foster a culture of mentoring, peer support, and advocacy in a community of practice that empowers data science learners and doers to keep up with fast-moving methods, tools, and technology.

> **Who**: Everyone who wants to do good things with data should get to make the effort, as long as they are rigorous and accountable.
>
> **How**: At the individual level, data science calls for technical skills in computation and data analysis and nontechnical skills to keep inquiry directed toward learning from data and to deal with obstacles. At the collective level, it calls for a progressive culture that supports putting those skills to use for doing good things with data.
>
> **What**: Data science studies how to learn from data by combining analytic, computational, and subject-matter methods to connect the whole life cycle of data, subject to norms of scientific quality and analytic rigor.
>
> **Why**: Foremost, data science is about learning from data. Data science helps us to keep up with fast-moving methods, tools, and technology for learning from data of all structures, sizes, shapes, and speeds.

A progressive culture remains rooted in history and continues to learn from old data in new ways, and it anticipates the future and handles evolving demands. Cultivating a progressive data culture in the present will best position the field of public health as ever ready to learn from and act on data.

7 I get to do data science

Who gets to do data science? I do!

7.1 How I think about data science

In January 2015, I started in CSELS as CDC's first, and (for at least 7 years) only, Associate Director for Data Science (ADDS). NIH had an ADDS by that time, and other CDC centers have had informatics or statistics leads, and some now have data science leads. But the title ADDS remained unique within CDC. I stepped into the role about 15 years after becoming a CDC employee and about 30 years after I first started working with data, statistics, and computing. In the role of CSELS ADDS, I tried to make sense and nonsense of the term "data science", thinking through what data science is and is not, why it matters for CDC, and most importantly how CDC can do public health better by doing data better. Early on, my favorite framing became not the definitional "*What is* data science?" but the cultural "*Who gets to do* data science?"

My personal values motivated me to pursue technical excellence and to offer those skills in public service. I entered public health because I wanted to use rigorous methods to help better the human condition. I am a methodologist, trained in math, statistics, some philosophy, and a smattering of other disciplines. I like to figure out how one can measure and count what is observed and quantify uncertainty about what remains beyond direct observation. Our culture perpetuates the notion that, with the inevitable march of dispassionate science, humanity will take the upper hand against what threatens or saddens us. How does a scientist resolve the apparent discordance between values and the cultural myth of dispassionate objectivity? We start by acknowledging the tools and power of scientific inquiry, and we respect their role in how we develop knowledge about the world. Values motivate and shape scientific endeavors, and passion itself can fuel scientific pursuits. None of us should shrink from or apologize for our commitment to the mission of public health. Our stories from data necessarily express perspectives and values; we have to commit to portraying and defending those worldviews.

My mentoring experience is the single greatest influence on how I think and talk about the values that can motivate and shape data science, as well as the skills that undergird data science practice. Since early 2000, I've had the pleasure of mentoring dozens of early-career scientists—post-doctoral fellows in the Epidemic Intelligence Service, Prevention Effectiveness, Public Health Informatics, and the Oak Ridge Institute for Science and Education (ORISE); undergraduate, master's, and doctoral students; all budding scholars and professionals in public health, medicine, philosophy, physics, mathematics. Into each of these relationships, I have poured a bit of myself and my respect for managing data, for coaxing meaning from data, for delighting in discovery from data, and for sharing stories from data with colleagues. Every one of these mentoring relationships has changed me as a data scientist and reinforced my belief that learners believe.

I came to believe 2 things about the practice and profession of data science in my time at CDC: (1) Everyone who practices data science should have the intellectual support to do so rigorously, whether a statistician, epidemiologist, philosopher, or some other brand of scholar. Rigorous practice entails standing behind your methods and conclusions, which can be an intimidating duty when reaching beyond your expertise. (2) Everyone who commits to doing data science as a profession should accept the intellectual

responsibility to contribute their expertise, both collaborating with and leading scientists from other disciplines. We have to commit ourselves to communicating clearly with those who share our specialty and with those who don't but who respect the ways that our specialty bolsters and advances public health research and practice.

7.2 My personal history with data science

I've been practicing or professing data science one way or another for a long time, typically under the title statistician or mathematical statistician or methodologist. When I was 5 years old in first grade, I thought that I might want to be a mathematician (or an artist or a basketball player). In sixth grade, I got to play with an Apple II, with its BASIC programming and 5.25-inch floppy diskettes. As an undergraduate, I helped teach the obscure but powerful programming language APL (literally "A Programming Language") to high school students. I earned a Bachelor of Arts degree in mathematics in 1991 and a Doctor of Philosophy degree in statistics in 1997.

In late 1997, about 2 months after filing my dissertation, I started with CDC as a contractor in the Division of Reproductive Health (DRH). I became a federal employee in early 2000, continuing to work in DRH, mostly on cohorts and clinical trials, until late 2004. Then I spent about 3 years overseeing CDC's institutional review boards and thinking about the connection between how we justify research risk and how we learn from data.

I served from mid-2007 through early 2015 in the Division of Tuberculosis Elimination (DTBE), working largely on clinical trials and creative but rigorous ways to get better at finding TB to stop TB. While in DTBE, I became a self-appointed evangelist for the R statistical computing environment. In 2012, I articulated a vision for leadership in mathematical sciences, which included skills specifically in technical excellence and clear communication—the seeds of my belief in leadership in a progressive culture for data. In March 2014, I spoke on "A Scrappy Little Division That Cares a Lot About Data: A Vision for Data Sciences in DTBE". That presentation included my first use and definition of the phrase "data science[s]", with particular attention to "data science tasks: end to end", later called the data life cycle.

In August 2015, shortly after I started as CSELS ADDS, I brainstormed dozens of potential topics for an internal CDC blog to explain and promote data science, called "expression(data, science)". I wrote the blog title as if it were in a fictitious programming language, monospace font and all: "`expression(data, science)`". The blog never really happened. This essay revives many of the topics that I had brainstormed.

In November and December 2015, I presented "The Art of Data Science: The Intense Pragmatism of Data in the Service of Public Health" at the EIS fall course. I told the story of data science through real-life experiences of 9 EIS fellows whom I had mentored. Although none of those fellows had had a background specifically focused on data analysis, many of them achieved great things with data, and others made interesting mistakes worth learning from.

In May 2016, 14 months into my tenure as CSELS ADDS, I delivered a CSELS science seminar entitled "Data Science and Data Wisdom: Cultivating a Data Generation" (a pun on "generating data"; maybe I should let up on the puns), in which I laid out the cultural components to support the practice of data

science. That presentation has evolved over the past few years to become “Who Gets to do Data Science? A Progressive Culture for Data in Public Health”, placing data science in the context of related but distinct disciplines and emphasizing who does data science more than what is data science.

In my latter days in CSELS, I turned my attention largely to machine learning (ML) and artificial intelligence (AI). We should regard ML as extending the set of data-analytic tools available to us, and we should use those tools where they help us learn things about the world—including potentially better ways to do public health surveillance and to adapt flexible and powerful ways to find disease and improve health. ML, and its scalable applications through AI, should not be mysterious or intimidating, and these tools should not be regarded as any more magical than familiar methods. Moreover, they should be subject to the same critical reflection and norms as other methods for scientific inquiry. Current agency discussion about the potential for ML and AI risks focusing too narrowly on technology and not enough on learning from data in ways that aim for scientific quality and rigor, as discussed at length in the sections 3.1 and 3.2 above: from posing questions of social value and scientific validity through ensuring that conclusions are traceable and defensible, that reasoning is coherent, and that the whole process is neutral, subject to minimal subjective bias, rigorous, transparent, reproducible, and so on. I expand further on ML and AI in section 8.

7.3 Why a progressive culture?

By the end of 2016, when I had been the CSELS ADDS for almost 2 years, CDC's Surveillance Strategy had successfully led to demonstrable improvements in technology for mortality, case-based surveillance, syndromic surveillance, and laboratory-based surveillance. Early formative efforts for a nascent Public Health Data Strategy in 2018 tapped dozens of midcareer and senior leaders to shape next-phase modernization. From those early days, I lodged 2 substantive concerns: (1) Regarding data, staff who work directly with public health data should join in co-leading the nascent data strategy because they know first-hand the challenges that they have in getting things done with data. (2) Regarding modernization, early-career staff should also join in co-leading the modernization effort, because they are more likely to have an essentially modern take on data and progress than mid- and late-career leaders alone. In August 2018, I nominated a “data science breakfast club”—an interdisciplinary collection of a dozen early-career data science practitioners—to discuss their experiences doing data science at CDC and how CDC could effectively develop a data science-savvy workforce. By early 2019, neither concern about data leadership and modernization leadership had gained appreciable traction, and the breakfast club never convened. I was told that the “movement" was open to early-career and data-involved staff, but it became clear that the movement would not engage them intentionally on their terms, for their co-leadership.

The strategy also struggled to describe *why* it was important to do data better. The emerging federal data strategy acknowledged the importance of data as an asset. And both the federal level and agency level connected that asset to informed decision-making and action. But neither the federal level nor the agency level explicitly articulated in what sense data were an asset and in what way data could inform decisions and action. I developed and shared a metaphor that we were conceiving of data as a treasure, and we were coming to acknowledge that we were largely hoarding that treasure, as if in a cave. Like the treasure in a cave, data are an asset because they have value, but we realize that value only when we *use* or *spend* rather than store the data. **Data have value because we use data to learn things about**

the world and to do things with what we learn, including but not limited to making decisions and taking action. Data have value because we use them to build things, like artificial intelligence tools, that promise to help us interact with the world more efficiently and effectively.

I became disheartened by what I perceived as regressive blinders. Despite my pessimism, I still believed that the intentions of the emerging strategy were largely sound. So I challenged myself to invert my pessimism and asked, "What would it take for CDC to be *progressive*?" About 3 months later, I had drafted a manifesto for a progressive culture for data in public health, appended to this essay (page 48).

In 2019, the Public Health Data Strategy merged with the concurrently emerging Information Technology Modernization Strategy to become the Public Health Modernization Initiative and eventually the Data Modernization Initiative (DMI). In August 2019, as the Surveillance Data Platform was wrapping up its work, a presentation on the merged modernization initiative enumerated 5 pillars for a modernization strategy. Despite the strategy's stated intent to develop "world-class analytics", nothing in the pillars addressed the role of analysis. When I pointed this out, I was told that it was implicit in all 5 pillars. A value that is not explicitly stated risks getting ignored. With DMI came initial political success in the form of $50 million in appropriation to seed the effort, fleeting moments before the Covid-19 pandemic pushed modernization efforts and funding into overdrive. DMI and concurrent investments have prepared the public health sector to advance more rapidly in response to the pandemic than ever expected. Nonetheless, DMI remained slow to engage data practitioners and early-career professionals as leaders in this data revolution. Thus, the manifesto still complements DMI as a vision for realizing the value of data as an asset in a culture centered on the roles of data practitioners and experts, learners and doers.

That manifesto is now an organizing principle for this essay. In the manifesto, I state, "a progressive culture remains rooted in history and continues to learn from old data in new ways." This collection is itself rooted in a personal history.

8 Machine learning and artificial intelligence

CDC should think about machine learning (ML) as a collection of data-analytic methods (most of them decades old) akin to statistical methods. This collection of methods extends the set of tools that we have for extracting information from data and putting that extracted information to use, typically for finding patterns in data or guessing a likely output based on a set of inputs. Artificial intelligence (AI), in current practice, applies ML and other data-analytic methods to automate or assist with various tasks, especially repetitive tasks. Indeed, it is because AI follows largely from applications of ML that I write the pair as "ML/AI" rather than the opposite: ML leads and grounds AI.

Like other data-analytic methods, ML methods should be used with critical reflection: How well does a model perform on new data? How well does it hold up under different assumptions? Does it perform consistent with norms like fairness? As an application of those methods, AI should be subject to the same critical reflection and norms.

ML and AI can be simple or complex, but they don't have to be mysterious. CDC should use these wherever they help CDC achieve its mission better. CDC should not, however, use these tools just for the sake of it, just to satisfy a consultant's recommendation, or just to appear modern.

The following discussion proposes how to demystify ML and AI by establishing them in context: where ML/AI fit in with more familiar, related concepts; were ML/AI fit in with related data-oriented methods; where ML/AI fit in history; where ML/AI might fit in a data-supportive organization; and where ML/AI have already been practiced by CDC/ATSDR.

8.1 Context: what is familiar or known

People who have no direct experience with ML and AI especially conceive of ML and AI in many ways that don't relate very closely to what CDC might do with them. For example, I have heard it sincerely posited that ML is about robots—literally, machines learning. Some machine learning approaches do support robots, but that's not the most common or important meaning for CDC's purposes.

ML has been described as answering the question, "How can computers learn to solve problems without being explicitly programmed?" (Koza et al. 1996) I don't find that formulation especially useful for people who aren't already familiar with the idea. Let's rephrase this question as "How can computers look at examples and figure out patterns that can be applied to new data?" Those examples that computers look at are data, and "figure out patterns" means the use of algorithms to develop a model or representation for those patterns. In other words, **ML is a collection of data-analytic methods, typically used for finding patterns in data or guessing a likely output based on a set of inputs.**

- Finding patterns in data could involve putting counties into groups that are demographically similar or grouping tweets on a common topic. Pattern-finding tasks, called unsupervised learning, include methods such as cluster analysis or topic modeling.
- Guessing a likely output based on a set of inputs is also known as prediction; it could entail a best guess at whether a child meets the surveillance case definition of autism given just the words of their educational and psychological evaluations. Output-oriented tasks, called supervised learning, include methods such as regression and classification.

In publications that apply machine learning to public health issues, classification has appeared far more commonly than the other tasks, often as an application for separating cases from noncases.

An initial note on the word "prediction" in the previous paragraph: In the context described above, prediction focuses on relating an outcome—the predicted value—to corresponding given inputs, typically called "features". In contrast with its everyday use, the word "prediction" in this context might or might not have anything to do with the future. In the autism example above, a child's current case status is predicted from their existing evaluations. As another example, a model could be constructed to predict who will receive a Parkinson's disease diagnosis given their past claims history; this example includes a time component and a sense of the future, but even in this example, the model is developed from past data and continuously evaluated against future accumulating data.

Since ML focuses on using data for these tasks, it should be thought of principally as an analytic application, subject to scientific norms the same as or similar to the norms applied to other empirical, analytic approaches, like statistics or causal inference. (See section 4.2.)

In predominant current practice, **AI is the application of ML to automate or assist with recurring tasks**, especially scaling up repetitive tasks, such as assessing a patient's possible case status given their electronic health record. AI should be thought of principally as an application of technology, but the underlying ML should still be subject to scientific norms for data analysis.

In summary, we can think of ML (approximately) as data-analytic methods or practices and AI as the results of those data-analytic methods or practices deployed as applications to automate or assist with recurring tasks, especially when repeated at a large scale.

8.2 Context: methodology

8.2.1 Machine learning and statistics

As briefly mentioned above, ML should be put into the context of other data-analytic practices, including classical statistical analysis and causal inference, among others. This is true all the more because ML and statistical methods overlap. For example, logistic regression can be applied to a binary classification task (as an ML method) or to the task of estimating the probability of an outcome being present or absent given covariate values (as a common statistical method).

ML tends to focus on model performance, such as a measure of how well an output can be associated with inputs, especially inputs that the ML hasn't seen yet; this is called out-of-sample performance. In contrast, statistical applications tend to focus on the internal structure and goodness-of-fit of the analytic model, typically intending to assist with explanation. This contrast is sometimes described in terms of ML focusing on $\hat{y}$, denoting the estimated value of the response, and statistics focusing on $\hat{\beta}$, denoting estimated model parameters, such as regression coefficients.

ML tends to handle complex or nontraditional data structures better than common statistical methods do, including images, free text, and electronic health records, but statistical models can also be large or complex. Statistical models are typically based on explicitly constructed probability models; although such models can be large or complex, the size and complexity might be constrained to facilitate interpretability of the model. In contrast, since ML models tend to focus on performance rather than interpretability, complexity in and of itself is more acceptable when added model complexity improves model performance and avoids the disadvantages of overfitting.

ML tends to handle larger numbers of inputs, hence larger numbers of model parameters, better than common statistical methods. In traditional statistical practice, several heuristics might be used to constrain model size, including stepwise variable selection, best-of-all-subsets regression, and penalties that force a tradeoff between model fit and model size. ML methods deal with potentially large numbers of inputs in 2 main ways: feature engineering, which seeks to derive more performative inputs from old inputs (one example being principal components), and regularization, which trades off between model performance and model size by figuring out ways to downweight or upweight inputs for optimizing out-of-sample performance. To be sure, statistical models can and do use some of the techniques described here for ML models.

Table 4. Typical differences between machine learning and statistics

Machine learning tendencies	Statistics tendencies

Model	Model performance, especially for associating outputs $(\hat{y})$ with inputs (x)	Model structure and fit $(\hat{\beta})$; "interpretability"
Data structure	Complex or nontraditional data, such as free text	Highly structured data, especially tabular
Data breadth	Large number of inputs, complex models	Constraints on number of inputs or model complexity

I have listed 3 main contrasts between ML and statistics: (1) a focus on model performance vs model fit, (2) facility for complex or nontraditional data structures, and (3) facility for larger numbers of inputs. These are not sharp, exclusive distinctions, and they are not the only distinctions. Even with these differences in orientation and approach, ML, statistical, and other data-analytic models should be subjected to similar levels of scrutiny and rigor, as well as other norms such as accuracy (and its many variations), fairness, bias mitigation, interpretability, and explainability. Of note, interpretability and explainability can be distractions, as apparently interpretable models are not necessarily closer to being true than complex or obscure models.

Just as ML should be put into the context of other data-analytic methods and practices, AI should be put into the context of other data-analytic applications. Although AI is often implemented to automate and assist with tasks at scale, such as decision-making, other data-analytic methods similarly undergird practical applications. For example, the Framingham risk score, commonly used in medical practice, was derived from empirical data on thousands of participants. Many other algorithms were similarly empirically derived. Those applications and AI applications share some common concerns:

- Do the underlying data-analytic models conform to scientific norms?
- Are the models subject to undue bias or other characteristics that could affect or limit their applicability?
- Do those limitations breach ethical, legal, or social norms, for example, by imposing or leading to unfair conditions or outcomes?

Most of these concerns pertain to algorithmic decision-making in general rather than AI in particular. Some unique concerns, however, arise from the potential for AI applications to be especially complex or dynamic by continuing to learn from accruing data, such as challenges in identifying conditions or sets of input values under which the model performs especially poorly or identifying changes in model performance as training data accumulate over time.

8.2.2 On "predictive analytics", ML, and AI

In the early days of CDC's Public Health Data Strategy, it was asserted that "predictive analytic tools such as machine learning" hold the answers for modernization. The assertion seems to assume that perfect, high-speed data can inevitably support informed action to intervene in public health, especially outbreaks, if only the best methods are used. For example, one presentation said that "the reality" is "looking back: using data to see what has already happened" and "the opportunity" is "looking forward: using data to **predict and prevent threats**" (original emphasis). The same presentation asserted the goal "to transform CDC and our partners from a culture of primarily historical data analytics to predictive data science …".

No doubt better data should lead to better learning, but analytic methods must also adequately account for the limits of information inherent in the combination of data and methods; otherwise, we risk mismatching expectations with realistic possibility. An emphasis on "predictive analytics" does not acknowledge the real limits of even the best data and tips over into a (likely unintended) undervaluing of cumulatively understanding history. AI might or might not aid in making better decisions. Data-analytic workflows can improve our ability to forecast, anticipate, and preemptively intervene, but we should take care not to tip the balance too far. Even when we are able to attend as completely as possible to forward-looking workflows, we will still be **primarily** (my emphasis) looking back to see what has already happened. Getting smarter and more nimble about the future still requires us to remain rooted in history. I would like to see a responsible treatment of how to use all available tools—classical and conventional, statistics and machine learning, correlation and causation—and those yet to be available, to achieve public health practice that is less exclusively reactive and reactionary.

8.3 Context: history

The January 2022 report *Protecting the Integrity of Government Science* by the Scientific Integrity Fast-Track Action Committee states:

> New technology and new approaches to science—such as big data analytics, AI, and ML—have become central to many areas of science and Federal decision-making. While these technological advances provide opportunities to more deeply and efficiently learn about the world, they also present unique challenges and complexities for ensuring scientific integrity. … Additionally, scientific integrity policies can be extended to offices and work units not traditionally focused on research and that make use of the results of AI and ML-based analyses." (Scientific Integrity Fast-Track Action Committee 2022, p 27-28)

This passage includes a rare and important acknowledgment that data analyses, including those that are nonresearch, should come under policies for scientific integrity. Like the "predictive analytics" example above, however, it overstates the "new approaches" and "unique challenges and complexities" stemming from ML and AI. Although AI can introduce complex issues in assistance and automation technologies, the upstream issues that arise from data analysis are not especially unique to ML or AI.

People at CDC often talk about ML and AI as new methods or new technology. Some methods, especially those associated with deep learning, are relatively new and yet their potential is familiar because of their widespread use in search engines and smartphones. But other methods and uses for ML go back decades. For example, early neural networks became popular in the 1980s, classification and regression trees were publicized in 1984, support vector machines in the 1990s, and random forests in 2001. This isn't a quibble about history so much as encouragement to see these methods as perhaps unfamiliar rather than new, and to realize that all these methods have been subjected to vigorous, and often rigorous, analysis, testing, and scrutiny. Thus, they can be applied with normative confidence similar to more familiar methods of comparable complexity as well as subject to similar scrutiny.

As mentioned above, while ML and AI can be more complex than familiar statistical methods and their applications, ML and AI inherit longstanding issues common to other forms of data analysis and applications, including bias and privacy concerns. In that regard, all data and analytic efforts should take care to elucidate potential biases and to promote transparency. Where the data are complex or the methods

are complex, these efforts warrant special attention and perhaps special methods because of the *complexity*. Whether complex Bayesian methods, methods using rich electronic health records, multilevel surveys, data synthesized from sources of varying content and quality, all complex data and complex methods warrant critical scientific thinking and problem-solving, not because they use ML or AI. In contrast, if concerns arise from uncritical reliance on assistive or automating algorithms, then the unique criticism inheres more to that uncritical reliance than to the algorithms themselves.

Furthermore, many ML methods have already been applied to public health problems in hundreds of published, peer-reviewed manuscripts. Many dozens of those manuscripts have either included an author with CDC or ATSDR affiliation or have resulted from a project funded by CDC/ATSDR. See, for example, Goertzel et al. (2006), Holt et al. (2009), Menon et al. (2014), Gu et al. (2015), Petersen et al. (2015), Bertke et al. (2016), Ladd-Acosta et al. (2016), Maenner et al. (2016), Rubaiyat et al. (2016), Arnold et al. (2017), Goldstick et al. (2017), Kracalik et al. (2017), Bowen et al. (2018), Meyers et al. (2018), Yanamala et al. (2018), Lee, Levin, et al. (2019), Lee, Maenner, et al. (2019), and Wheeler (2019). These publications span applications to infectious and noninfectious conditions as well as cross-cutting areas like syndromic surveillance. They entail ML methods that include regularized regression, decision trees and tree-based ensembles (like random forests and gradient-boosting machines), support vector machines, other ensemble methods (like the super learner), and a variety of shallow and deep neural network architectures. Although most publications use supervised learning methods, especially classification, many use unsupervised methods, such as topic modeling.

In the current context, CDC is positioned to continue contributing rigorous work that employs ML methods, especially as the base of R and Python users grows within the agency to take advantage of high-quality, open-source tools. CDC's greater technical challenges at the moment entail incremental uptake of cloud-enabled technologies and supporting operations for deploying trained models, especially deep learning models that use graphics processing unit (GPU) hardware. Early efforts with proven models have been stymied by procedural glitches that prevent real implementation as AI. Nonetheless, because AI is seeing an ever-expanding collection of useful applications in clinical medicine, the prospects are strong for public health applications. For example, methods for using rich, possibly messy electronic health records hold promise for applications as varied as self-adapting triggers for electronic case reporting, enriching the use of emergency departments and other sources for syndromic surveillance, and forecasting the population prevalence of a wide variety of conditions, including autism spectrum disorder and Parkinson's disease. ML might or might not help with general forecasting and outbreak analysis, as other statistical methods could be suited for those purposes and warrant as much developmental attention as ML does.

8.4 Context: organizational culture

As we try to imagine the possible applications for ML and AI to CDC's mission, we should also conceptualize how ML and AI are normalized within the organizational structure. It has already happened, and will continue to happen, that every center at CDC uses ML in some way. Yet there is no central leadership on ML or AI.

Foremost, because ML undergirds AI, and because ML and other data-analytic approaches should be similarly subject to scientific norms, it follows that ML rather than AI should drive both growth and prac-

tice. If CDC promotes AI out of balance with ML, then we risk deploying technologies and purported solutions that do not hold up to scientific scrutiny, where out-of-sample performance, bias, and drift go underappreciated and undermine scientific integrity.

Furthermore, efforts to fortify workforce capacity should focus primarily on analytic literacy, including critical thinking and assessment. While CDC unquestionably needs ML engineers and other technology-adept skills, those roles and skills need to be carried out within the bounds of credible scientific practice. I draw here on the more general discussion in section 4 on how to foster a culture for doing good things with data by investing in technical skills, nontechnical skills, and community, tailored here to ML and AI. Many of these skills already exist in CDC's existing workforce, largely underrecognized and underappreciated. If CDC can come to recognize and appreciate existing technical and nontechnical skills among current federal employees, fellows and other learners, and nonfederal staff, then we can build on those skills faster. Moreover, as we work toward expanding capacity in ML and AI, we should include current data-analytic practitioners (including those who use ML) in *leading* those efforts. Finally, as we build capacity, we need to balance innovation with a respect for history. While we should continue to expand the set of tools available to us for learning from data and building things using data, we can't afford to lose sight of existing methods that also serve our purposes.

The Department of Health and Human Services locates AI leadership within its Office of the Chief Information Officer. In my view, this office is predicated on some fundamental category errors that threaten to constrain or misdirect efforts to use and apply the full set of methods for learning from data and building things with data. The office's very definition of ML as "a type of artificial intelligence" (HHS OCIO 2021) obscures more than it reveals. This framing has precedent, as when MIT's management school presents ML as a subfield of artificial intelligence (Brown 2021). As I argued above, because ML methods "learn" from data, ML is about data analysis; I further argue that ML should be judged in ways similar to other empirical, specifically data-analytic, approaches. It is important to see ML as connected to the full range of data-analytic methods and tools, for at least 2 reasons: (1) Statistical and machine learning methods, and other data-analytic methods (such as causal inference), all have formal methods for characterizing performance and optimization, and those methods connect across fields. ML is not just a set of methods unto itself, but it emphasizes characteristics that differ from other domains. (2) By subsuming ML under AI, we lose the understanding that even conventional, classical statistical methods can drive AI, and we risk burdening ML practice more broadly. Without that grounding in science and related norms, ML and AI risk giving too much privilege to model performance. Indeed, the move to delimit "trustworthy" AI seeks out trustworthiness norms for this reason. While important concerns arise from implementing algorithms to assist or to automate, we can and should distinguish upstream issues, for example, that stem from input data or from model structure.

If CDC develops central leadership on ML and AI, it should *not* follow HHS's lead by aligning ML/AI primarily with technology. Instead, CDC should align ML/AI primarily with the practice of science, specifically data-intensive science, with all the norms that that entails. As I argued above, technology should take its lead from scientific interests. Technology can help to show what is possible, but it should neither push nor limit what is possible, within resource and security constraints. Technology should respond to and empower scientific advances.

Declaration: a progressive culture for data in public health

A progressive culture for data in public health keeps up with fast-moving methods, tools, and technology for continuously learning things about the world and empowering choices informed by those learnings. This declaration frames the major elements for cultivating and sustaining this progressive culture and for leading public health into the modern era. Public health scientists who care about data can flourish in a culture that fosters technical skills, inspires and rewards intellectual drive, and supports learners in community.

- Technical skills include math and statistics, programming and data structures, communications and visualization, and domain knowledge in public health and allied fields.
- Nontechnical skills put the traits of a good learner or knower into action, motivating and enabling those who care about data to learn and to practice wisely.
- Support for empowering and doing good things with data comes from a community of peers, mentors, and advocates centered on learning.

A progressive culture for data in public health manifests the following principles:

1. Data as object: A progressive culture for data is **dedicated to learning *from* data**. In a progressive culture, data have value because they mediate how we learn things about the world. Those learnings allow us to make informed choices about how we interact with the world, for example, through public health interventions.
2. Data as subject: A progressive culture for data is **dedicated to learning *about* data**, because data come in many structures, sizes, shapes, and speeds, from small, flat data tables to massive, unstructured data streams. Data conform to a variety of standards, or no standards at all. The varied characteristics and complexity of data both enrich and constrain the ways that data reveal characteristics of the world.
3. Data as mediator: A progressive culture for data is **dedicated to learning *with* data** through its full life cycle[1], primarily through knowing how analytic methods allow us
    - 3.1. to pose rich questions about the world, amenable to rich methods;
    - 3.2. to guide how we generate, transmit, obtain, and prepare data;
    - 3.3. to probe data to answer questions about the world;
    - 3.4. to place answers from data in context, mindful of assumptions and alternatives;
    - 3.5. to present data-driven answers to audiences clearly and correctly;
    - 3.6. to preserve those answers and ensure that the entire life cycle is transparent, accessible, traceable and, to the extent possible, reproducible.
4. Community supports doing good things with data through 4 primary roles:
    - 4.1. **Learner-practitioners** with basic or intermediate data skills come from any discipline to do good things with data, mindful of the full life cycle of data. In a progressive culture

[1] Items 3.1-3.5 are adapted from *The Art of Data Science*.

for data, everyone who wants to do good things with data has the intellectual support to do so, accepting that they must proceed with rigor and stand behind their work. Practitioners learn continuously and show how modern tools and methods solve modern problems.

4.2. **Expert practitioners** go deep on data science methods. Experts help ensure that everyone who wants to do good things with data, can. They set norms for the practice of data science and for learning from, about, and with data. They enable, guide, correct, and empower practitioners to proceed with rigor and stand behind their work.

4.3. **Managers** supervise practitioners and experts, to ensure that they have the resources and direction that they need to achieve good things with data, now and in the future. Managers foster and reward curiosity, invest in learning (not just training), and encourage innovation and interesting mistakes. They advocate for the means to enable practitioners and experts to continue increasing their capability, efficiency, and effectiveness. Managers hold practitioners and experts to account for producing knowledge learned from, about, and with data.

4.4. **Lay advocates** work in community with practitioners, experts, and managers as persons literate in the value of data to help learn things about the world. Laypersons know how to assess, use, and advocate for learning from data. They help ensure supportive resources to enable learning.

5. Members of a progressive culture for data **respect the fundamentals, value innovation, and practice pragmatic, principled pluralism**. Members apply wisely and well all methods that can help them achieve technical excellence to learn from, about, and with data. These methods include the classical, conventional, and innovative; statistics and machine learning; correlational, causal, and predictive inference; analysis, synthesis, and forecasting. Principled pluralism allows honest disagreement about methods, results, and interpretation.

6. A progressive culture for data in public health **fosters public trust**, motivated by public service to conduct itself ethically, **protect privacy**, and ensure that data and methods are **radically open and transparent**.

7. In a progressive culture for data, **leadership is radically and intentionally inclusive**, continually shaping and sustaining the culture of good data practice. Practitioners, experts, managers, and laypersons lead from every level, regardless of their career stage, job title, credential, or job series, so that everyone who wants to do good things with data, can. In a progressive culture for data, governance enables effective leadership, but governance does not substitute for leadership.

The public health sector faces constant challenges to stretch modest resources, to anticipate and respond to threats, and to promote population health. A progressive culture remains rooted in history and continues to learn from old data in new ways, and it anticipates the future and handles evolving demands and to keep up with fast-moving methods, tools, and technology. Cultivating a progressive data culture in the present will best position the field of public health as ever ready to learn from and act on data.

References

Arnold BF, Laan MJ van der, Hubbard AE, Steel C, Kubofcik J, Hamlin KL, Moss DM, Nutman TB, Priest JW, Lammie PJ. 2017. Measuring changes in transmission of neglected tropical diseases, malaria, and enteric pathogens from quantitative antibody levels. *PLoS Neglected Tropical Diseases*. 11(5):e0005616. https://doi.org/10.1371/journal.pntd.0005616

Baehr J. 2013b. Educating for intellectual virtues: From theory to practice. *Journal of Philosophy of Education*. 47(2):248–262. https://doi.org/10.1111/1467-9752.12023

Baehr J. 2013a. The cognitive demands of intellectual virtue. In: Henning T, Schweikard DP, editors. *Knowledge, Virtue, and Action: Putting Epistemic Virtues to Work*. 1st Edition. New York: Routledge; p. 99–118. https://doi.org/10.4324/9780203098486

Baehr J. 2015. *Cultivating Good Minds: A Philosophical & Practical Guide to Educating for Intellectual Virtues*. https://jasonbaehr.gumroad.com/l/IJxPL

Baehr JS. 2021. *Deep in thought: A practical guide to teaching for intellectual virtues*. Cambridge, Massachusetts: Harvard Education Press.

Bertke SJ, Meyers AR, Wurzelbacher SJ, Measure A, Lampl MP, Robins D. 2016. Comparison of methods for auto-coding causation of injury narratives. *Accident Analysis & Prevention*. 88:117–123. https://doi.org/10.1016/j.aap.2015.12.006

Blei DM, Smyth P. 2017. Science and data science. P*roceedings of the National Academy of Sciences*. 114(33):8689–8692. https://doi.org/10.1073/pnas.1702076114

Bowen DA, Mercer Kollar LM, Wu DT, Fraser DA, Flood CE, Moore JC, Mays EW, Sumner SA. 2018. Ability of crime, demographic and business data to forecast areas of increased violence. *International Journal of Injury Control and Safety Promotion*. 25(4):443–448. https://doi.org/10.1080/17457300.2018.1467461

Breiman L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*. 16(3):199–231. https://doi.org/10.1214/ss/1009213726

Brown S. 2021. Machine learning, explained. MIT Sloan School of Management. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Costa AL, Kallick B. 2008. Describing the habits of mind. In: *Learning and Leading with Habits of Mind: 16 Essential Characteristics for Success*. Association for Supervision; Curriculum Development; p. 15–41. https://www.ascd.org/books/learning-and-leading-with-habits-of-mind?chapter=learning-through-reflection-learning-and-leading-with-habits-of-mind

Data Science Association. Data Science Code of Professional Conduct. https://www.datascience-assn.org/code-of-conduct.html

Davenport TH, Patil DJ. 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dishman L. 2016. These are the top 25 jobs in the U.S. this year. *Fast Company*. http://www.fastcompany.com/3055629/the-future-of-work/these-are-the-top-25-jobs-in-the-us-this-year

Donoho D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics*. 26(4):745–766. https://doi.org/10.1080/10618600.2017.1384734

Freedman B. 1987. Scientific value and validity as ethical requirements for research: A proposed explication. *IRB: Ethics & Human Research*. 9(6):7–10. https://doi.org/10.2307/3563623

Gehrke A, Luo M, Takahata S. 2021. *Critical Factors for Building Successful Data Science Teams Final Summary Report*. Mathematica.

Gelman A. 2013. Statistics is the least important part of data science. https://statmodeling.stat.columbia.edu/2013/11/14/statistics-least-important-part-data-science/

Goertzel BN, Pennachin C, De Souza Coelho L, Gurbaxani B, Maloney EM, Jones JF. 2006. Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome. *Pharmacogenomics*. 7(3):475–483. https://doi.org/10.2217/14622416.7.3.475

Goldstick JE, Carter PM, Walton MA, Dahlberg LL, Sumner SA, Zimmerman MA, Cunningham RM. 2017. Development of the SaFETy score: A clinical screening tool for predicting future firearm violence risk. *Annals of Internal Medicine*. 166(10):707–714. https://doi.org/10.7326/M16-1927

Gu W, Vieira AR, Hoekstra RM, Griffin PM, Cole D. 2015. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. *Epidemiology and Infection*. 143(13):2786–2794. https://doi.org/10.1017/S095026881500014X

HHS OCIO. 2021. *Trustworthy AI Playbook*. https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

Holt AC, Salkeld DJ, Fritz CL, Tucker JR, Gong P. 2009. Spatial analysis of plague in California: Niche modeling predictions of the current distribution and potential response to climate change. *International Journal of Health Geographics*. 8(1):38. https://doi.org/10.1186/1476-072X-8-38

Jones ML. 2018. How we became instrumentalists (again): Data positivism since World War II. *Historical Studies in the Natural Sciences*. 48(5):673–684. https://doi.org/10.1525/hsns.2018.48.5.673

King N. 2014. What are intellectual virtues? Five key features of the intellectual virtues. https://cct.biola.edu/intellectual-virtues/

Koza JR, Bennett FH, Andre D, Keane MA. 1996. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: Gero JS, Sudweeks F, editors. *Artificial Intelligence in Design '96*. Dordrecht: Springer Netherlands; p. 151–170. https://doi.org/10.1007/978-94-009-0279-4_9

Kracalik IT, Kenu E, Ayamdooh EN, Allegye-Cudjoe E, Polkuu PN, Frimpong JA, Nyarko KM, Bower WA, Traxler R, Blackburn JK. 2017. Modeling the environmental suitability of anthrax in Ghana and estimating populations at risk: Implications for vaccination and control. *PLoS Neglected Tropical Diseases*. 11(10):e0005885. https://doi.org/10.1371/journal.pntd.0005885

Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, Schendel DE, Jones N, Daniels JL, Windham GC, et al. 2016. Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. *Environmental Research*. 144:139–148. https://doi.org/10.1016/j.envres.2015.11.014

Lee SH, Levin D, Finley PD, Heilig CM. 2019. Chief complaint classification with recurrent neural networks. *Journal of Biomedical Informatics*. 93:103158. https://doi.org/10.1016/j.jbi.2019.103158

Lee SH, Maenner MJ, Heilig CM. 2019. A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. *PLoS ONE*. 14(9):e0222907. https://doi.org/10.1371/journal.pone.0222907

Maenner MJ, Yeargin-Allsopp M, Van Naarden Braun K, Christensen DL, Schieve LA. 2016. Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PLoS ONE*. 11(12):e0168224. https://doi.org/10.1371/journal.pone.0168224

Menon R, Bhat G, Saade GR, Spratt H. 2014. Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth. *Acta Obstetricia et Gynecologica Scandinavica*. 93(4):382–391. https://doi.org/10.1111/aogs.12344

Meyers AR, Al-Tarawneh IS, Wurzelbacher SJ, Bushnell PT, Lampl MP, Bell JL, Bertke SJ, Robins DC, Tseng C-Y, Wei C, et al. 2018. Applying machine learning to workers' compensation data to identify industry-specific ergonomic and safety prevention priorities: Ohio, 2001 to 2011. *Journal of Occupational and Environmental Medicine*. 60(1):55–73. https://doi.org/10.1097/JOM.0000000000001162

Millis J. 2015. Data scientists don't scale. Machine intelligence does. http://blog.nutonian.com/data-scientists-dont-scale [site no longer available]

National Academies of Sciences, Engineering, and Medicine. 2018. *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24886

National Institutes of Health. 2018. *NIH Strategic Plan for Data Science*. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf. See also datascience.nih.gov/strategicplan.

NIST Big Data Public Working Group. 2015. *NIST Big Data Interoperability Framework: Volume 1, Definitions*. National Institute of Standards; Technology. https://doi.org/10.6028/NIST.SP.1500-1r2

Olshen R. 2001. A conversation with Leo Breiman. *Statistical Science*. 16(2):184–198. https://doi.org/10.1214/ss/1009213290

Pasteur L, Vallery-Radot LP. Discours prononcé à Douai le 7 décembre 1854 à l'occasion de l'installation solennelle de la faculté des lettres de Douai et de la faculté des sciences de Lille. In: *Réunies par Pasteur Vallery-Radot*. Vol. Tome 7. Masson (Paris).

Pearl J. 2009. *Causality*. 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Peng RD, Matsui E. 2015. *The Art of Data Science: A Guide for Anyone Who Works with Data*. Leanpub. http://leanpub.next/artofdatascience. Also available at bookdown.org/rdpeng/artofdatascience/

Petersen ML, LeDell E, Schwab J, Sarovar V, Gross R, Reynolds N, Haberer JE, Goggin K, Golin C, Arnsten J, et al. 2015. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 69(1):109–118. https://doi.org/10.1097/QAI.0000000000000548

Reinhold MD. 2019. Data Scientist Titling Guidance. https://www.chcoc.gov/content/data-scientist-titling-guidance

Rubaiyat AHM, Toma TT, Kalantari-Khandani M, Rahman SA, Chen L, Ye Y, Pan CS. 2016. Automatic detection of helmet uses for construction safety. In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*. Omaha, NE, USA: IEEE; p. 135–142. https://doi.org/10.1109/WIW.2016.045

Savel TG, Foldy S. 2012. The role of public health informatics in enhancing public health surveillance. *MMWR Supplements*. 61(3):20–24. https://www.cdc.gov/mmwr/preview/mmwrhtml/su6103a5.htm

Scientific Integrity Fast-Track Action Committee. 2022. *Protecting the Integrity of Government Science*. https://www.whitehouse.gov/wp-content/uploads/2022/01/01-22-Protecting_the_Integrity_of_Government_Science.pdf

Turri J, Alfano M, Greco J. 2021. Virtue epistemology. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy. Winter 2021*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue/

Van Cauwenberghe L. 2015. Can you learn data science on the job? https://www.gettingsmart.com/2015/12/03/can-you-learn-data-science-on-the-job/

Wagner RM. 2022. A*uthorities and Mechanisms for Hiring and Retaining Data Scientists at HHS*.

Wheeler MW. 2019. Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: An application to high-throughput toxicity testing. *Biometrics*. 75(1):193–201. https://doi.org/10.1111/biom.12942

Yanamala N, Orandle MS, Kodali VK, Bishop L, Zeidler-Erdely PC, Roberts JR, Castranova V, Erdely A. 2018. Sparse Supervised Classification Methods Predict and Characterize Nanomaterial Exposures: Independent Markers of MWCNT Exposures. *Toxicologic Pathology*. 46(1):14–27. https://doi.org/10.1177/0192623317730575