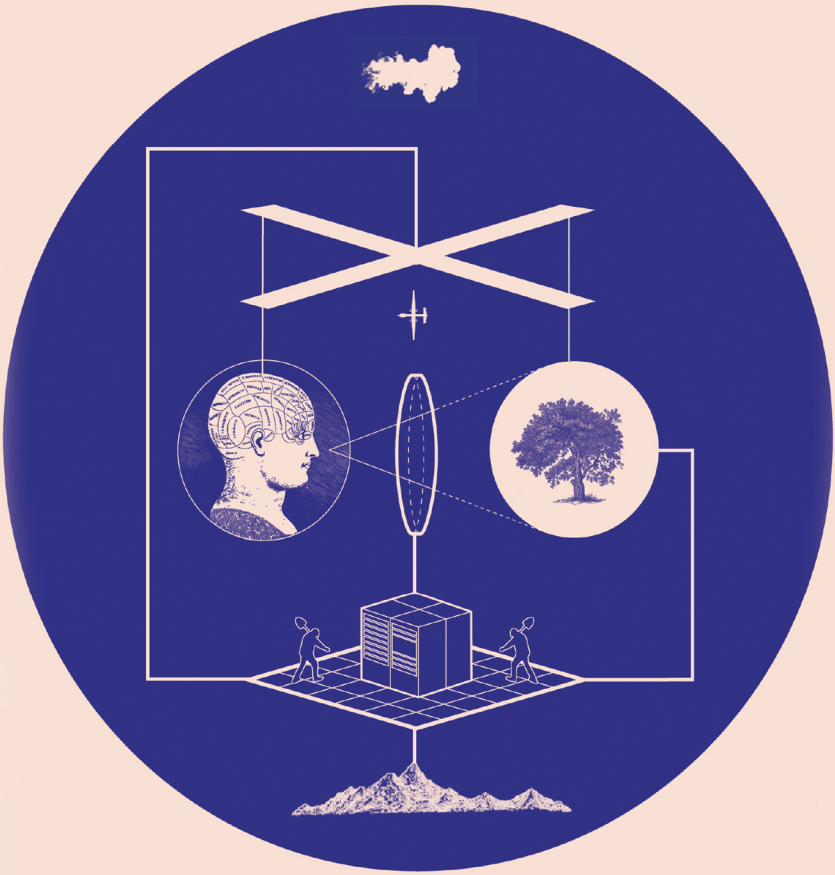
KATE CRAWFORD



ATLAS OF AI

Atlas of AI

This page intentionally left blank

Atlas of AI

Power, Politics, and the Planetary Costs of Artificial Intelligence

KATE CRAWFORD

Yale UNIVERSITY PRESS

New Haven and London

Yale University Press books may be purchased in quantity for educational, business, or promotional use. For information, please e-mail sales.press@yale.edu (U.S. office) or sales@yaleup.co.uk (U.K. office).

Cover design and chapter opening illustrations by Vladan Joler.
Set in Minion by Tseng Information Systems, Inc.
Printed in the United States of America.

Library of Congress Control Number: 2020947842
ISBN 978-0-300-20957-0 (hardcover : alk. paper)

A catalogue record for this book is available from the British Library.

This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

10 9 8 7 6 5 4 3 2 1

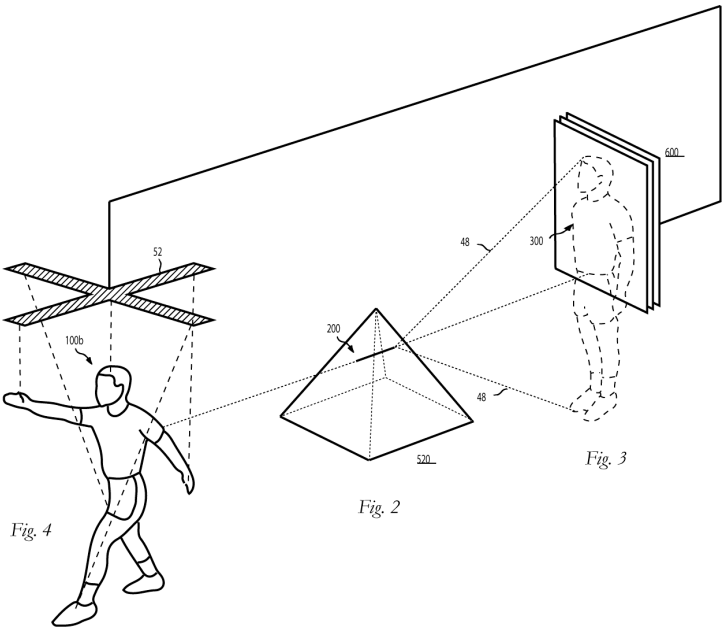


Fig. 4

Fig. 2

Fig. 3

3
Data

A young woman gazes upward, eyes focused on something outside the frame, as though she is refusing to acknowledge the camera. In the next photograph, her eyes are locked on the middle distance. Another image shows her with disheveled hair and a downcast expression. Over the sequence of photos we see her aging over time, and the lines around her mouth turn down and deepen. In the final frame she appears injured and dispirited. These are mug shots of a woman across multiple arrests over many years of her life. Her images are contained in a collection known as NIST Special Database 32–Multiple Encounter Dataset, which is shared on the internet for researchers who would like to test their facial recognition software.[1]

This dataset is one of several maintained by the National Institute of Standards and Technology (NIST), one of the oldest and most respected physical science laboratories in the United States and now part of the Department of Commerce. NIST was created in 1901 to bolster the nation's measurement infrastructure and to create standards that could compete with economic rivals in the industrialized world, such as Germany

Images from NIST Special Database 32—Multiple Encounter Dataset (MEDS). National Institute of Standards and Technology, U.S. Department of Commerce

and the United Kingdom. Everything from electronic health records to earthquake-resistant skyscrapers to atomic clocks is under the purview of NIST. It became the agency of measurement: of time, of communications protocols, of inorganic crystal structures, of nanotechnology.[2] NIST's purpose is to make systems interoperable through defining and supporting standards, and this now includes developing standards for artificial intelligence. One of the testing infrastructures it maintains is for biometric data.

I first discovered the mug shot databases in 2017 when I was researching NIST's data archives. Their biometric collections are extensive. For more than fifty years, NIST has collaborated with the Federal Bureau of Investigation on auto-

mated fingerprint recognition and has developed methods to assess the quality of fingerprint scanners and imaging systems.[3] After the terrorist attacks of September 11, 2001, NIST became part of the national response to create biometric standards to verify and track people entering the United States.[4] This was a turning point for research on facial recognition; it widened out from a focus on law enforcement to controlling people crossing national borders.[5]

The mug shot images themselves are devastating. Some people have visible wounds, bruising, and black eyes; some are distressed and crying. Others stare blankly back at the camera. Special Dataset 32 contains thousands of photographs of deceased people with multiple arrests, as they endured repeated encounters with the criminal justice system. The people in the mug shot datasets are presented as data points; there are no stories, contexts, or names. Because mug shots are taken at the time of arrest, it's not clear if these people were charged, acquitted, or imprisoned. They are all presented alike.

The inclusion of these images in the NIST database has shifted their meaning from being used to identify individuals in systems of law enforcement to becoming the technical baseline to test commercial and academic AI systems for detecting faces. In his account of police photography, Allan Sekula has argued that mug shots are part of a tradition of technical realism that aimed to "provide a standard physiognomic gauge of the criminal."[6] There are two distinct approaches in the history of the police photograph, Sekula observes. Criminologists like Alphonse Bertillon, who invented the mug shot, saw it as a kind of biographical machine of identification, necessary to spot repeat offenders. On the other hand, Francis Galton, the statistician and founding figure of eugenics, used composite portraiture of prisoners as a way to detect a biologically determined "criminal type."[7] Galton was working within a physi-

ognomist paradigm in which the goal was to find a generalized look that could be used to identify deep character traits from external appearances. When mug shots are used as training data, they function no longer as tools of identification but rather to fine-tune an automated form of vision. We might think of this as Galtonian formalism. They are used to detect the basic mathematical components of faces, to "reduce nature to its geometrical essence."[8]

Mug shots form part of the archive that is used to test facial-recognition algorithms. The faces in the Multiple Encounter Dataset have become standardized images, a technical substrate for comparing algorithmic accuracy. NIST, in collaboration with the Intelligence Advanced Research Projects Activity (IARPA), has run competitions with these mug shots in which researchers compete to see whose algorithm is the fastest and most accurate. Teams strive to beat one another at tasks like verifying the identity of faces or retrieving a face from a frame of surveillance video.[9] The winners celebrate these victories; they can bring fame, job offers, and industry-wide recognition.[10]

Neither the people depicted in the photographs nor their families have any say about how these images are used and likely have no idea that they are part of the test beds of AI. The subjects of the mug shots are rarely considered, and few engineers will ever look at them closely. As the NIST document describes them, they exist purely to "refine tools, techniques, and procedures for face recognition as it supports Next Generation Identification (NGI), forensic comparison, training, analysis, and face image conformance and inter-agency exchange standards."[11] The Multiple Encounter Dataset description observes that many people show signs of enduring violence, such as scars, bruises, and bandages. But the document concludes that these signs are "difficult to interpret due to the

lack of ground truth for comparison with a 'clean' sample."[12] These people are not seen so much as individuals but as part of a shared technical resource—just another data component of the Facial Recognition Verification Testing program, the gold standard for the field.

I've looked at hundreds of datasets over years of research into how AI systems are built, but the NIST mug shot databases are particularly disturbing because they represent the model of what was to come. It's not just the overwhelming pathos of the images themselves. Nor is it solely the invasion of privacy they represent, since suspects and prisoners have no right to refuse being photographed. It's that the NIST databases foreshadow the emergence of a logic that has now thoroughly pervaded the tech sector: the unswerving belief that everything is data and is there for the taking. It doesn't matter where a photograph was taken or whether it reflects a moment of vulnerability or pain or if it represents a form of shaming the subject. It has become so normalized across the industry to take and use whatever is available that few stop to question the underlying politics.

Mug shots, in this sense, are the urtext of the current approach to making AI. The context—and exertion of power—that these images represent is considered irrelevant because they no longer exist as distinct things unto themselves. They are not seen to carry meanings or ethical weight as images of individual people or as representations of structural power in the carceral system. The personal, the social, and the political meanings are all imagined to be neutralized. I argue this represents a shift from *image* to *infrastructure,* where the meaning or care that might be given to the image of an individual person, or the context behind a scene, is presumed to be erased at the moment it becomes part of an aggregate mass that will drive a broader system. It is all treated as data to be run through functions, material to be ingested to improve techni-

cal performance. This is a core premise in the ideology of data extraction.

Machine learning systems are trained on images like these every day—images that were taken from the internet or from state institutions without context and without consent. They are anything but neutral. They represent personal histories, structural inequities, and all the injustices that have accompanied the legacies of policing and prison systems in the United States. But the presumption that somehow these images can serve as apolitical, inert material influences how and what a machine learning tool "sees." A computer vision system can detect a face or a building but not why a person was inside a police station or any of the social and historical context surrounding that moment. Ultimately, the specific instances of data—a picture of a face, for example—aren't considered to matter for training an AI model. All that matters is a sufficiently varied aggregate. Any individual image could easily be substituted for another and the system would work the same. According to this worldview, there is always more data to capture from the constantly growing and globally distributed treasure chest of the internet and social media platforms.

A person standing in front of a camera in an orange jumpsuit, then, is dehumanized as just more data. The history of these images, how they were acquired, and their institutional, personal, and political contexts are not considered relevant. The mug shot collections are used like any other practical resource of free, well-lit images of faces, a benchmark to make tools like facial recognition function. And like a tightening ratchet, the faces of deceased persons, suspects, and prisoners are harvested to sharpen the police and border surveillance facial recognition systems that are then used to monitor and detain more people.

The last decade has seen a dramatic capture of digital material for AI production. This data is the basis for sense-making in AI, not as classical representations of the world with individual meaning, but as a mass collection of data for machine abstractions and operations. This large-scale capture has become so fundamental to the AI field that it is unquestioned. So how did we get here? What ways of conceiving data have facilitated this stripping of context, meaning, and specificity? How is training data acquired, understood, and used in machine learning? In what ways does training data limit *what* and *how* AI interprets the world? What forms of power do these approaches enhance and enable?

In this chapter I show how data has become a driving force in the success of AI and its mythos and how everything that can be readily captured is being acquired. But the deeper implications of this standard approach are rarely addressed, even as it propels further asymmetries of power. The AI industry has fostered a kind of ruthless pragmatism, with minimal context, caution, or consent-driven data practices while promoting the idea that the mass harvesting of data is necessary and justified for creating systems of profitable computational "intelligence." This has resulted in a profound metamorphosis, where all forms of image, text, sound, and video are just raw data for AI systems and the ends are thought to justify the means. But we should ask: Who has benefited most from this transformation, and why have these dominant narratives of data persisted? And as we saw in the previous chapters, the logic of extraction that has shaped the relationship to the earth and to human labor is also a defining feature of how data is used and understood in AI. By looking closely at training data as a central example in the ensemble of machine learning, we can begin to see what is at stake in this transformation.

Training Machines to See

It's useful to consider why machine learning systems currently demand massive amounts of data. One example of the problem in action is computer vision, the subfield of artificial intelligence concerned with teaching machines to detect and interpret images. For reasons that are rarely acknowledged in the field of computer science, the project of interpreting images is a profoundly complex and relational endeavor. Images are remarkably slippery things, laden with multiple potential meanings, irresolvable questions, and contradictions. Yet now it's common practice for the first steps of creating a computer vision system to scrape thousands—or even millions—of images from the internet, create and order them into a series of classifications, and use this as a foundation for how the system will perceive observable reality. These vast collections are called training datasets, and they constitute what AI developers often refer to as "ground truth."[13] Truth, then, is less about a factual representation or an agreed-upon reality and more commonly about a jumble of images scraped from whatever various online sources were available.

For supervised machine learning, human engineers supply labeled training data to a computer. Two distinct types of algorithms then come into play: *learners* and *classifiers*. The learner is the algorithm that is trained on these labeled data examples; it then informs the classifier how best to analyze the relation between the new inputs and the desired target output (or prediction). It might be predicting whether a face is contained in an image or whether an email is spam. The more examples of correctly labeled data there are, the better the algorithm will be at producing accurate predictions. There are many kinds of machine learning models, including neural networks, logistic regression, and decision trees. Engineers will

choose a model based on what they are building—be it a facial recognition system or a means of detecting sentiment on social media—and fit it to their computational resources.

Consider the task of building a machine learning system that can detect the difference between pictures of apples and oranges. First, a developer has to collect, label, and train a neural network on thousands of labeled images of apples and oranges. On the software side, the algorithms conduct a statistical survey of the images and develop a model to recognize the difference between the two classes. If all goes according to plan, the trained model will be able to distinguish the difference between images of apples and oranges that it has never encountered before.

But if, in our example, all of the training images of apples are red and none are green, then a machine learning system might deduce that "all apples are red." This is what is known as an *inductive inference,* an open hypothesis based on available data, rather than a *deductive inference,* which follows logically from a premise.[14] Given how this system was trained, a green apple wouldn't be recognized as an apple at all. Training datasets, then, are at the core of how most machine learning systems make inferences. They serve as the primary source material that AI systems use to form the basis of their predictions.

Training data also defines more than just the features of machine learning algorithms. It is used to assess how they perform over time. Like prized thoroughbreds, machine learning algorithms are constantly raced against one another in competitions all over the world to see which ones perform the best with a given dataset. These benchmark datasets become the alphabet on which a *lingua franca* is based, with many labs from multiple countries converging around canonical sets to try to outperform one another. One of the best-known competitions is the ImageNet Challenge, where researchers com-

pete to see whose methods can most accurately classify and detect objects and scenes.[15]

Once training sets have been established as useful benchmarks, they are commonly adapted, built upon, and expanded. As we will see in the next chapter, a type of genealogy of training sets emerges—they inherit learned logic from earlier examples and then give rise to subsequent ones. For example, ImageNet draws on the taxonomy of words inherited from the influential 1980s lexical database known as WordNet; and WordNet inherits from many sources, including the Brown Corpus of one million words, published in 1961. Training datasets stand on the shoulders of older classifications and collections. Like an expanding encyclopedia, the older forms remain and new items are added over decades.

Training data, then, is the foundation on which contemporary machine learning systems are built.[16] These datasets shape the epistemic boundaries governing how AI operates and, in that sense, create the limits of how AI can "see" the world. But training data is a brittle form of ground truth—and even the largest troves of data cannot escape the fundamental slippages that occur when an infinitely complex world is simplified and sliced into categories.

A Brief History of the Demand for Data

"The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it." So said Vannevar Bush, the inventor and administrator who oversaw the Manhattan Project as director of the Office of Scientific Research and Development and later was integral to the creation of the National Science Foundation. It was July 1945; the bombs were yet to drop on Hiroshima and Nagasaki, and Bush had a theory about a new kind of data-connecting system that

was yet to be born. He envisaged the "advanced arithmetical machines of the future" that would perform at extremely fast speed and "select their own data and manipulate it in accordance with the instructions." But the machines would need monumental amounts of data: "Such machines will have enormous appetites. One of them will take instructions and data from a whole roomful of girls armed with simple key board punches, and will deliver sheets of computed results every few minutes. There will always be plenty of things to compute in the detailed affairs of millions of people doing complicated things."[17]

The "roomful of girls" Bush referred to were the keypunch operators doing the day-to-day work of computation. As historians Jennifer Light and Mar Hicks have shown, these women were often dismissed as input devices for intelligible data records. In fact, their role was just as important to crafting data and making systems work as that of the engineers who designed the wartime-era digital computers.[18] But the relationship between data and processing machinery was already being imagined as one of endless consumption. The machines would be data-hungry, and there would surely be a wide horizon of material to extract from millions of people.

In the 1970s, artificial intelligence researchers were mainly exploring what's called an expert systems approach: rules-based programming that aims to reduce the field of possible actions by articulating forms of logical reasoning. But it quickly became evident that this approach was fragile and impractical in real-world settings, where a rule set was rarely able to handle uncertainty and complexity.[19] New approaches were needed. By the mid-1980s, research labs were turning toward probabilistic or brute force approaches. In short, they were using lots of computing cycles to calculate as many options as possible to find the optimal result.

One significant example was the speech recognition group at IBM Research. The problem of speech recognition had primarily been dealt with using linguistic methods, but then information theorists Fred Jelinek and Lalit Bahl formed a new group, which included Peter Brown and Robert Mercer (long before Mercer became a billionaire, associated with funding Cambridge Analytica, Breitbart News, and Donald Trump's 2016 presidential campaign). They tried something different. Their techniques ultimately produced precursors for the speech recognition systems underlying Siri and Dragon Dictate, as well as machine translation systems like Google Translate and Microsoft Translator.

They started using statistical methods that focused more on how often words appeared in relation to one another, rather than trying to teach computers a rules-based approach using grammatical principles or linguistic features. Making this statistical approach work required an enormous amount of real speech and text data, or training data. The result, as media scholar Xiaochang Li writes, was that it required "a radical reduction of speech to merely data, which could be modeled and interpreted in the absence of linguistic knowledge or understanding. Speech *as such* ceased to matter." This shift was incredibly significant, and it would become a pattern repeated for decades: the reduction from context to data, from meaning to statistical pattern recognition. Li explains:

> The reliance on data over linguistic principles, however, presented a new set of challenges, for it meant that the statistical models were necessarily determined by the characteristics of training data. As a result, the size of the dataset became a central concern. . . . Larger datasets of observed out-

> comes not only improved the probability estimates for a random process, but also increased the chance that the data would capture more rarely-occurring outcomes. Training data size, in fact, was so central to IBM's approach that in 1985, Robert Mercer explained the group's outlook by simply proclaiming, "There's no data like more data."[20]

For several decades, that data was remarkably hard to come by. As Lalit Bahl describes in an interview with Li, "Back in those days . . . you couldn't even find a million words in computer-readable text very easily. And we looked all over the place for text."[21] They tried IBM technical manuals, children's novels, patents of laser technology, books for the blind, and even the typed correspondence of IBM Fellow Dick Garwin, who created the first hydrogen bomb design.[22] Their method strangely echoed a short story by the science fiction author Stanislaw Lem, in which a man called Trurl decides to build a machine that would write poetry. He starts with "eight hundred and twenty tons of books on cybernetics and twelve thousand tons of the finest poetry."[23] But Trurl realizes that to program an autonomous poetry machine, one needs "to repeat the entire Universe from the beginning—or at least a good piece of it."[24]

Ultimately, the IBM Continuous Speech Recognition group found their "good piece" of the universe from an unlikely source. A major federal antitrust lawsuit was filed against IBM in 1969; the proceedings lasted for thirteen years, and almost a thousand witnesses were called. IBM employed a large staff just to digitize all of the deposition transcripts onto Hollerith punch cards. This ended up creating a corpus of a hundred million words by the mid-1980s. The notoriously antigovern-

ment Mercer called this a "case of utility accidentally created by the government in spite of itself."[25]

IBM wasn't the only group starting to gather words by the ton. From 1989 to 1992, a team of linguists and computer scientists at the University of Pennsylvania worked on the Penn Treebank Project, an annotated database of text. They collected four and a half million words of American English for the purpose of training natural language processing systems. Their sources included Department of Energy abstracts, Dow Jones newswire articles, and Federal News Service reports of "terrorist activity" in South America.[26] The emerging text collections borrowed from earlier collections and then contributed new sources. Genealogies of data collections began to emerge, each building on the last—and often importing the same peculiarities, issues, or omissions wholesale.

Another classic corpus of text came from the fraud investigations of Enron Corporation after it declared the largest bankruptcy in American history. The Federal Energy Regulatory Commission seized the emails of 158 employees for the purposes of legal discovery.[27] It also decided to release these emails online because "the public's right to disclosure outweighs the individual's right to privacy."[28] This became an extraordinary collection. Over half a million exchanges in everyday speech could now be used as a linguistic mine: one that nonetheless represented the gender, race, and professional skews of those 158 workers. The Enron corpus has been cited in thousands of academic papers. Despite its popularity, it is rarely looked at closely: the *New Yorker* described it as "a canonic research text that no one has actually read."[29] This construction of and reliance on training data anticipated a new way of doing things. It transformed the field of natural language processing and laid the foundations of what would become normal practice in machine learning.

The seeds of later problems were planted here. Text archives were seen as neutral collections of language, as though there was a general equivalence between the words in a technical manual and how people write to colleagues via email. All text was repurposable and swappable, so long as there was enough of it that it could train a language model to predict with high levels of success what word might follow another. Like images, text corpuses work on the assumption that all training data is interchangeable. But language isn't an inert substance that works the same way regardless of where it is found. Sentences taken from Reddit will be different from those composed by executives at Enron. Skews, gaps, and biases in the collected text are built into the bigger system, and if a language model is based on the kinds of words that are clustered together, it matters where those words come from. There is no neutral ground for language, and all text collections are also accounts of time, place, culture, and politics. Further, languages that have less available data are not served by these approaches and so are often left behind.[30]

Clearly there are many histories and contexts that combine within IBM's training data, the Enron archive, or the Penn Treebank. How do we unpack what is and is not meaningful to understand these datasets? How does one communicate warnings like, "This dataset likely reflects skews related to its reliance on news stories about South American terrorists in the 1980s"? The origins of the underlying data in a system can be incredibly significant, and yet there are still, thirty years later, no standardized practices to note where all this data came from or how it was acquired—let alone what biases or classificatory politics these datasets contain that will influence all the systems that come to rely on them.[31]

Capturing the Face

While computer-readable text was becoming highly valued for speech recognition, the human face was the core concern for building systems of facial recognition. One central example emerged in the last decade of the twentieth century, funded by the Department of Defense CounterDrug Technology Development Program Office. It sponsored the Face Recognition Technology (FERET) program to develop automatic face recognition for intelligence and law enforcement. Before FERET, little training data of human faces was available, only a few collections of fifty or so faces here and there—not enough to do facial recognition at scale. The U.S. Army Research Laboratory led the technical project of creating a training set of portraits of more than a thousand people, in multiple poses, to make a grand total of 14,126 images. Like NIST's mug shot collections, FERET became a standard benchmark—a shared measuring tool to compare approaches for detecting faces.

The tasks that the FERET infrastructure was created to support included, once again, automated searching of mug shots, as well as monitoring airports and border crossings and searching driver's license databases for "fraud detection" (multiple welfare claims was a particular example mentioned in FERET research papers).[32] But there were two primary testing scenarios. In the first, an electronic mug book of known individuals would be presented to an algorithm, which then had to locate the closest matches from a large gallery. The second scenario focused on border and airport control: identifying a known individual—"smugglers, terrorists, or other criminals"—from a large population of unknown people.

These photographs are machine-readable by design, and not meant for human eyes, yet they make for remarkable viewing. The images are surprisingly beautiful—high-resolution

photographs captured in the style of formal portraiture. Taken with 35 mm cameras at George Mason University, the tightly framed headshots depict a wide range of people, some of whom seem to have dressed for the occasion with carefully styled hair, jewelry, and makeup. The first set of photographs, taken between 1993 and 1994, are like a time capsule of early nineties haircuts and fashion. The subjects were asked to turn their heads to multiple positions; flicking through the images, you can see profile shots, frontal images, varying levels of illumination, and sometimes different outfits. Some subjects were photographed over several years, in order to begin to study how to track people as they age. Each subject was briefed about the project and signed a release form that had been approved by the university's ethics review board. Subjects knew what they were participating in and gave full consent.[33] This level of consent would become a rarity in later years.

FERET was the high-water mark of a formal style of "making data," before the internet began offering mass extraction without any permissions or careful camera work. Even at this early stage, though, there were problems with the lack of diversity of the faces collected. The FERET research paper from 1996 admits that "some questions were raised about the age, racial, and sexual distribution of the database" but that "at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals."[34] Indeed, FERET was extraordinarily useful for this. As the interest in terrorist detection intensified and funding for facial recognition dramatically increased after 9/11, FERET became the most commonly used benchmark. From that point onward, biometric tracking and automated vision systems would rapidly expand in scale and ambition.

From the Internet to ImageNet

The internet, in so many ways, changed everything; it came to be seen in the AI research field as something akin to a natural resource, there for the taking. As more people began to upload their images to websites, to photo-sharing services, and ultimately to social media platforms, the pillaging began in earnest. Suddenly, training sets could reach a size that scientists in the 1980s could never have imagined. Gone was the need to stage photo shoots using multiple lighting conditions, controlled parameters, and devices to position the face. Now there were millions of selfies in every possible lighting condition, position, and depth of field. People began to share their baby photos, family snaps, and images of how they looked a decade ago, an ideal resource for tracking genetic similarity and face aging. Trillions of lines of text, containing both formal and informal forms of speech, were published every day. It was all grist for the mills of machine learning. And it was vast. As an example, on an average day in 2019, approximately 350 million photographs were uploaded to Facebook and 500 million tweets were sent.[35] And that's just two platforms based in the United States. Anything and everything online was primed to become a training set for AI.

The tech industry titans were now in a powerful position: they had a pipeline of endlessly refreshing images and text, and the more people shared their content, the more the tech industry's power grew. People would happily label their photographs with names and locations, free of charge, and that unpaid labor resulted in having more accurate, labeled data for machine vision and language models. Within the industry, these collections are highly valuable. They are proprietary troves that are rarely shared, given both the privacy issues and the competitive advantage they represent. But those outside

the industry, such as the leading computer science labs in academia, wanted the same advantages. How could they afford to harvest people's data and have it hand-labeled by willing human participants? That's when new ideas began to emerge: combining images and text extracted from the internet with the labor of low-paid crowdworkers.

One of the most significant training sets in AI is ImageNet. It was first conceptualized in 2006, when Professor Fei-Fei Li decided to build an enormous dataset for object recognition. "We decided we wanted to do something that was completely historically unprecedented," Li said. "We're going to map out the entire world of objects."[36] The breakthrough research poster was published by the ImageNet team at a computer vision conference in 2009. It opened with this description:

> The digital era has brought with it an enormous explosion of data. The latest estimations put a number of more than 3 billion photos on Flickr, a similar number of video clips on YouTube and an even larger number for images in the Google Image Search database. More sophisticated and robust models and algorithms can be proposed by exploiting these images, resulting in better applications for users to index, retrieve, organize and interact with these data.[37]

From the outset, data was characterized as something voluminous, disorganized, impersonal, and ready to be exploited. According to the authors, "Exactly how such data can be utilized and organized is a problem yet to be solved." By extracting millions of images from the internet, primarily from search engines using the image-search option, the team produced a "large-scale ontology of images" that was meant to

serve as a resource for "providing critical training and benchmarking data" for object and image recognition algorithms. Using this approach, ImageNet grew enormous. The team mass-harvested more than fourteen million images from the internet to be organized into more than twenty thousand categories. Ethical concerns about taking people's data were not mentioned in any of the team's research papers, even though many thousands of the images were of a highly personal and compromising nature.

Once the images had been scraped from the internet, a major concern arose: Who would label them all and put them into intelligible categories? As Li describes it, the team's first plan was to hire undergraduate students for ten dollars an hour to find images manually and add them to the dataset.[38] But she realized that with their budget, it would take more than ninety years to complete the project. The answer came when a student told Li about a new service: Amazon Mechanical Turk. As we saw in chapter 2, this distributed platform meant that it was suddenly possible to access a distributed labor force to do online tasks, like labeling and sorting images, at scale and at low cost. "He showed me the website, and I can tell you literally that day I knew the ImageNet project was going to happen," Li said. "Suddenly we found a tool that could scale, that we could not possibly dream of by hiring Princeton undergrads."[39] Unsurprisingly, the undergraduates did not get the job.

Instead, ImageNet would become, for a time, the world's largest academic user of Amazon's Mechanical Turk, deploying an army of piecemeal workers to sort an average of fifty images a minute into thousands of categories.[40] There were categories for apples and airplanes, scuba divers and sumo wrestlers. But there were cruel, offensive, and racist labels, too: photographs of people were classified into categories like

"alcoholic," "ape-man," "crazy," "hooker," and "slant eye." All of these terms were imported from WordNet's lexical database and given to crowdworkers to pair with images. Over the course of a decade, ImageNet grew into a colossus of object recognition for machine learning and a powerfully important benchmark for the field. The approach of mass data extraction without consent and labeling by underpaid crowdworkers would become standard practice, and hundreds of new training datasets would follow ImageNet's lead. As we will see in the next chapter, these practices—and the labeled data they generated—eventually came back to haunt the project.

The End of Consent

The early years of the twenty-first century marked a shift away from consent-driven data collection. In addition to dispensing with the need for staged photo shoots, those responsible for assembling datasets presumed that the contents of the internet were theirs for the taking, beyond the need for agreements, signed releases, and ethics reviews. Now even more troubling practices of extraction began to emerge. For example, at the Colorado Springs campus of the University of Colorado, a professor installed a camera on the main walkway of the campus and secretly captured photos of more than seventeen hundred students and faculty—all to train a facial recognition system of his own.[41] A similar project at Duke University harvested footage of more than two thousand students without their knowledge as they went between their classes and then published the results on the internet. The dataset, called DukeMTMC (for multitarget, multicamera facial recognition), was funded by the U.S. Army Research Office and the National Science Foundation.[42]

The DukeMTMC project was roundly criticized after an investigative project by artists and researchers Adam Harvey and Jules LaPlace showed that the Chinese government was using the images to train systems for the surveillance of ethnic minorities. This spurred an investigation by Duke's institutional review board, which determined that this was a "significant deviation" from acceptable practices. The dataset was removed from the internet.[43]

But what happened at the University of Colorado and Duke were by no means isolated cases. At Stanford University, researchers commandeered a webcam from a popular café in San Francisco to extract almost twelve thousand images of "everyday life of a busy downtown café" without anyone's consent.[44] Over and over, data extracted without permission or consent would be uploaded for machine learning researchers, who would then use it as an infrastructure for automated imaging systems.

Another example is Microsoft's landmark training dataset MS-Celeb, which scraped approximately ten million photos of a hundred thousand celebrities from the internet in 2016. At the time, it was the largest public facial recognition dataset in the world, and the people included were not just famous actors and politicians but also journalists, activists, policymakers, academics, and artists.[45] Ironically, several of the people who had been included in the set without consent are known for their work critiquing surveillance and facial recognition itself, including documentary filmmaker Laura Poitras; digital rights activist Jillian York; critic Evgeny Morozov; and the author of *Surveillance Capitalism,* Shoshana Zuboff.[46]

Even when datasets are scrubbed of personal information and released with great caution, people have been reidentified or highly sensitive details about them have been revealed. In 2013, for example, the New York City Taxi and

Limousine Commission released a dataset of 173 million individual cab rides, and it included pickup and drop-off times, locations, fares, and tip amounts. The taxi drivers' medallion numbers were anonymized, but this was quickly undone, enabling researchers to infer sensitive information like annual incomes and home addresses.[47] Once combined with public information from sources like celebrity blogs, some actors and politicians were identified, and it was possible to deduce the addresses of people who visited strip clubs.[48] But beyond individual harms, such datasets also generate "predictive privacy harms" for whole groups or communities.[49] For instance, the same New York City taxi dataset was used to suggest which taxi drivers were devout Muslims by observing when they stopped at prayer times.[50]

From any seemingly innocuous and anonymized dataset can come many unexpected and highly personal forms of information, but this fact has not hampered the collection of images and text. As success in machine learning has come to rely on ever-larger datasets, more people are seeking to acquire them. But why does the wider AI field accept this practice, despite the ethical, political, and epistemological problems and potential harms? What beliefs, justifications, and economic incentives normalized this mass extraction and general equivalence of data?

Myths and Metaphors of Data

The oft-cited history of artificial intelligence written by AI professor Nils Nilsson outlines several of the founding myths about data in machine learning. He neatly illustrates how data is typically described in the technical disciplines: "The great volume of raw data calls for efficient 'data-mining' techniques for classifying, quantifying, and extracting useful information.

Machine learning methods are playing an increasingly important role in data analysis because they can deal with massive amounts of data. In fact, the more data the better."[51]

Echoing Robert Mercer from decades earlier, Nilsson perceived that data was everywhere for the taking, and all the better for mass classification by machine learning algorithms.[52] It was such a common belief as to have become axiomatic: data is there to be acquired, refined, and made valuable.

But vested interests carefully manufactured and supported this belief over time. As sociologists Marion Fourcade and Kieran Healy note, the injunction always to collect data came not only from the data professions but also from their institutions and the technologies they deploy:

> The institutional command coming from technology is the most potent of all: we do these things *because we can*. . . . Professionals recommend, the institutional environment demands, and technology enables organizations to sweep up as much individual data as possible. It does not matter that the amounts collected may vastly exceed a firm's imaginative reach or analytic grasp. The assumption is that it will eventually be useful, i.e. valuable. . . . Contemporary organizations are both culturally impelled by the data imperative and powerfully equipped with new tools to enact it.[53]

This produced a kind of moral imperative to collect data in order to make systems better, regardless of the negative impacts the data collection might cause at any future point. Behind the questionable belief that "more is better" is the idea that individuals can be completely knowable, once enough disparate pieces of data are collected.[54] But what counts as data?

Historian Lisa Gitelman notes that every discipline and institution "has its own norms and standards for the imagination of data."[55] Data, in the twenty-first century, became whatever could be captured.

Terms like "data mining" and phrases like "data is the new oil" were part of a rhetorical move that shifted the notion of data away from something personal, intimate, or subject to individual ownership and control toward something more inert and nonhuman. Data began to be described as a resource to be consumed, a flow to be controlled, or an investment to be harnessed.[56] The expression "data as oil" became commonplace, and although it suggested a picture of data as a crude material for extraction, it was rarely used to emphasize the costs of the oil and mining industries: indentured labor, geopolitical conflicts, depletion of resources, and consequences stretching beyond human timescales.

Ultimately, "data" has become a bloodless word; it disguises both its material origins and its ends. And if data is seen as abstract and immaterial, then it more easily falls outside of traditional understandings and responsibilities of care, consent, or risk. As researchers Luke Stark and Anna Lauren Hoffman argue, metaphors of data as a "natural resource" just lying in wait to be discovered are a well-established rhetorical trick used for centuries by colonial powers.[57] Extraction is justified if it comes from a primitive and "unrefined" source.[58] If data is framed as oil, just waiting to be extracted, then machine learning has come to be seen as its necessary refinement process.

Data also started to be viewed as capital, in keeping with the broader neoliberal visions of markets as the primary forms of organizing value. Once human activities are expressed through digital traces and then tallied up and ranked within scoring metrics, they function as a way to extract value. As

Fourcade and Healy observe, those who have the right data signals gain advantages like discounted insurance and higher standing across markets.[59] High achievers in the mainstream economy tend to do well in a data-scoring economy, too, while those who are poorest become targets of the most harmful forms of data surveillance and extraction. When data is considered as a form of capital, then everything is justified if it means collecting more. The sociologist Jathan Sadowski similarly argues that data now operates as a form of capital. He suggests that once everything is understood as data, it justifies a cycle of ever-increasing data extraction: "Data collection is thus driven by the perpetual cycle of capital accumulation, which in turn drives capital to construct and rely upon a world in which everything is made of data. The supposed universality of data reframes everything as falling under the domain of data capitalism. All spaces must be subjected to datafication. If the universe is conceived of as a potentially infinite reserve of data, then that means the accumulation and circulation of data can be sustained forever."[60]

This drive to accumulate and circulate is the powerful underlying ideology of data. Mass data extraction is the "new frontier of accumulation and next step in capitalism," Sadowski suggests, and it is the foundational layer that makes AI function.[61] Thus, there are entire industries, institutions, and individuals who don't want this frontier—where data is there for the taking—to be questioned or destabilized.

Machine learning models require ongoing flows of data to become more accurate. But machines are asymptotic, never reaching full precision, which propels the justification for more extraction from as many people as possible to fuel the refineries of AI. This has created a shift away from ideas like "human subjects"—a concept that emerged from the ethics debates of the twentieth century—to the creation of "data

subjects," agglomerations of data points without subjectivity or context or clearly defined rights.

Ethics at Arm's Length

The great majority of university-based AI research is done without any ethical review process. But if machine learning techniques are being used to inform decisions in sensitive domains like education and health care, then why are they not subject to greater review? To understand that, we need to look at the precursor disciplines of artificial intelligence. Before the emergence of machine learning and data science, the fields of applied mathematics, statistics, and computer science had not historically been considered forms of research on human subjects.

In the early decades of AI, research using human data was usually seen to be a minimal risk.[62] Even though datasets in machine learning often come from and represent people and their lives, the research that used those datasets was seen more as a form of applied math with few consequences for human subjects. The infrastructures of ethics protections, like university-based institutional review boards (IRBs), had accepted this position for years.[63] This initially made sense; IRBs had been overwhelmingly focused on the methods common to biomedical and psychological experimentation in which interventions carry clear risks to individual subjects. Computer science was seen as far more abstract.

Once AI moved out of the laboratory contexts of the 1980s and 1990s and into real-world situations—such as attempting to predict which criminals will reoffend or who should receive welfare benefits—the potential harms expanded. Further, those harms affect entire communities as well as individuals. But there is still a strong presumption that publicly available data-

sets pose minimal risks and therefore should be exempt from ethics review.[64] This idea is the product of an earlier era, when it was harder to move data between locations and very expensive to store it for long periods. Those earlier assumptions are out of step with what is currently going on in machine learning. Now datasets are more easily connectable, indefinitely repurposable, continuously updatable, and frequently removed from the context of collection.

The risk profile of AI is rapidly changing as its tools become more invasive and as researchers are increasingly able to access data without interacting with their subjects. For example, a group of machine learning researchers published a paper in which they claimed to have developed an "automatic system for classifying crimes."[65] In particular, their focus was on whether a violent crime was gang-related, which they claimed their neural network could predict with only four pieces of information: the weapon, the number of suspects, the neighborhood, and the location. They did this using a crime dataset from the Los Angeles Police Department, which included thousands of crimes that had been labeled by police as gang-related.

Gang data is notoriously skewed and riddled with errors, yet researchers use this database and others like it as a definitive source for training predictive AI systems. The CalGang database, for example, which is widely used by police in California, has been shown to have major inaccuracies. The state auditor discovered that 23 percent of the hundreds of records it reviewed lacked adequate support for inclusion. The database also contained forty-two infants, twenty-eight of whom were listed for having "admitting to being gang members."[66] Most of the adults on the list had never been charged, but once they were included in the database, there was no way to have their name removed. Reasons for being included might be as simple

as chatting with a neighbor while wearing a red shirt; using these trifling justifications, Black and Latinx people have been disproportionately added to the list.[67]

When the researchers presented their gang-crime prediction project at a conference, some attendees were troubled. As reported by *Science,* questions from the audience included, "How could the team be sure the training data were not biased to begin with?" and "What happens when someone is mislabeled as a gang member?" Hau Chan, a computer scientist now at Harvard University who presented the work, responded that he couldn't know how the new tool would be used. "[These are the] sort of ethical questions that I don't know how to answer appropriately," he said, being just "a researcher." An audience member replied by quoting a lyric from Tom Lehrer's satiric song about the wartime rocket scientist Wernher von Braun: "Once the rockets are up, who cares where they come down?"[68]

This separation of ethical questions away from the technical reflects a wider problem in the field, where the responsibility for harm is either not recognized or seen as beyond the scope of the research. As Anna Lauren Hoffman writes: "The problem here isn't only one of biased datasets or unfair algorithms and of unintended consequences. It's also indicative of a more persistent problem of researchers actively reproducing ideas that damage vulnerable communities and reinforce current injustices. Even if the Harvard team's proposed system for identifying gang violence is never implemented, hasn't a kind of damage already been done? Wasn't their project an act of cultural violence in itself?"[69] Sidelining issues of ethics is harmful in itself, and it perpetuates the false idea that scientific research happens in a vacuum, with no responsibility for the ideas it propagates.

The reproduction of harmful ideas is particularly dangerous now that AI has moved from being an experimental discipline used only in laboratories to being tested at scale on

millions of people. Technical approaches can move rapidly from conference papers to being deployed in production systems, where harmful assumptions can become ingrained and hard to reverse.

Machine learning and data-science methods can create an abstract relationship between researchers and subjects, where work is being done at a distance, removed from the communities and individuals at risk of harm. This arm's-length relationship of AI researchers to the people whose lives are reflected in datasets is a long-established practice. Back in 1976, when AI scientist Joseph Weizenbaum wrote his scathing critique of the field, he observed that computer science was already seeking to circumvent all human contexts.[70] He argued that data systems allowed scientists during wartime to operate at a psychological distance from the people "who would be maimed and killed by the weapons systems that would result from the ideas they communicated."[71] The answer, in Weizenbaum's view, was to directly contend with what data actually represents: "The lesson, therefore, is that the scientist and technologist must, by acts of will and of the imagination, actively strive to reduce such psychological distances, to counter the forces that tend to remove him from the consequences of his actions. He must—it is as simple as this—think of what he is actually doing."[72]

Weizenbaum hoped that scientists and technologists would think more deeply about the consequences of their work—and of who might be at risk. But this would not become the standard of the AI field. Instead, data is more commonly seen as something to be taken at will, used without restriction, and interpreted without context. There is a rapacious international culture of data harvesting that can be exploitative and invasive and can produce lasting forms of harm.[73] And there are many industries, institutions, and individuals who are strongly incentivized to maintain this colonizing at-

titude—where data is there for the taking—and they do not want it questioned or regulated.

The Capture of the Commons

The current widespread culture of data extraction continues to grow despite concerns about privacy, ethics, and safety. By researching the thousands of datasets that are freely available for AI development, I got a glimpse into what technical systems are built to recognize, of how the world is rendered for computers in ways that humans rarely see. There are gigantic datasets full of people's selfies, tattoos, parents walking with their children, hand gestures, people driving their cars, people committing crimes on CCTV, and hundreds of everyday human actions like sitting down, waving, raising a glass, or crying. Every form of biodata—including forensic, biometric, sociometric, and psychometric—is being captured and logged into databases for AI systems to find patterns and make assessments.

Training sets raise complex questions from ethical, methodological, and epistemological perspectives. Many were made without people's knowledge or consent and were harvested from online sources like Flickr, Google image search, and YouTube or were donated by government agencies like the FBI. This data is now used to expand facial recognition systems, modulate health insurance rates, penalize distracted drivers, and fuel predictive policing tools. But the practices of data extraction are extending even deeper into areas of human life that were once off-limits or too expensive to reach. Tech companies have drawn on a range of approaches to gain new ground. Voice data is gathered from devices that sit on kitchen counters or bedroom nightstands; physical data comes from watches on wrists and phones in pockets; data about what books and newspapers

are read comes from tablets and laptops; gestures and facial expressions are compiled and assessed in workplaces and classrooms.

The collection of people's data to build AI systems raises clear privacy concerns. Take, for example, the deal that Britain's Royal Free National Health Service Foundation Trust made with Google's subsidiary DeepMind to share the patient data records of 1.6 million people. The National Health Service in Britain is a revered institution, entrusted to provide health care that is primarily free to all while keeping patient data secure. But when the agreement with DeepMind was investigated, the company was found to have violated data protection laws by not sufficiently informing patients.[74] In her findings, the information commissioner observed that "the price of innovation does not need to be the erosion of fundamental privacy rights."[75]

Yet there are other serious issues that receive less attention than privacy. The practices of data extraction and training dataset construction are premised on a commercialized capture of what was previously part of the commons. This particular form of erosion is a privatization by stealth, an extraction of knowledge value from public goods. A dataset may still be publicly available, but the metavalue of the data—the model created by it—is privately held. Certainly, many good things can be done with public data. But there has been a social and, to some degree, a technical expectation that the value of data shared via public institutions and public spaces online should come back to the public good in other forms of the commons. Instead, we see a handful of privately owned companies that now have enormous power to extract insights and profits from those sources. The new AI gold rush consists of enclosing different fields of human knowing, feeling, and action—every

type of available data—all caught in an expansionist logic of never-ending collection. It has become a pillaging of public space.

Fundamentally, the practices of data accumulation over many years have contributed to a powerful extractive logic, a logic that is now a core feature of how the AI field works. This logic has enriched the tech companies with the largest data pipelines, while the spaces free from data collection have dramatically diminished. As Vannevar Bush foresaw, machines have enormous appetites. But how and what they are fed has an enormous impact on how they will interpret the world, and the priorities of their masters will always shape how that vision is monetized. By looking at the layers of training data that shape and inform AI models and algorithms, we can see that gathering and labeling data about the world is a social and political intervention, even as it masquerades as a purely technical one.

The way data is understood, captured, classified, and named is fundamentally an act of world-making and containment. It has enormous ramifications for the way artificial intelligence works in the world and which communities are most affected. The myth of data collection as a benevolent practice in computer science has obscured its operations of power, protecting those who profit most while avoiding responsibility for its consequences.

69. Muse, "Organizing Tech."

70. Abdi Muse, personal conversation with the author, October 2, 2019.

71. Gurley, "60 Amazon Workers Walked Out."

72. Muse quoted in *Organizing Tech.*

73. Desai quoted in *Organizing Tech.*

74. Estreicher and Owens, "Labor Board Wrongly Rejects Employee Access to Company Email."

75. This observation comes from conversations with various labor organizers, tech workers, and researchers, including Astra Taylor, Dan Greene, Bo Daley, and Meredith Whittaker.

76. Kerr, "Tech Workers Protest in SF."

3
Data

1. National Institute of Standards and Technology (NIST), "Special Database 32—Multiple Encounter Dataset (MEDS)."

2. Russell, *Open Standards and the Digital Age.*

3. Researchers at NIST (then the National Bureau of Standards, NBS) began working on the first version of the FBI's Automated Fingerprint Identification System in the late 1960s. See Garris and Wilson, "NIST Biometrics Evaluations and Developments," 1.

4. Garris and Wilson, 1.

5. Garris and Wilson, 12.

6. Sekula, "Body and the Archive," 17.

7. Sekula, 18–19.

8. Sekula, 17.

9. See, e.g., Grother et al., "2017 IARPA Face Recognition Prize Challenge (FRPC)."

10. See, e.g., Ever AI, "Ever AI Leads All US Companies."

11. Founds et al., "NIST Special Database 32."

12. Curry et al., "NIST Special Database 32 Multiple Encounter Dataset I (MEDS-I)," 8.

13. See, e.g., Jaton, "We Get the Algorithms of Our Ground Truths."

14. Nilsson, *Quest for Artificial Intelligence,* 398.

15. "ImageNet Large Scale Visual Recognition Competition (ILSVRC)."

16. In the late 1970s, Ryszard Michalski wrote an algorithm based on symbolic variables and logical rules. This language was popular in the 1980s and 1990s, but as the rules of decision-making and qualification became more complex, the language became less usable. At the same moment, the potential of using large training sets triggered a shift from this conceptual clus-

tering to contemporary machine learning approaches. Michalski, "Pattern Recognition as Rule-Guided Inductive Inference."

17. Bush, "As We May Think."

18. Light, "When Computers Were Women"; Hicks, *Programmed Inequality.*

19. As described in Russell and Norvig, *Artificial Intelligence,* 546.

20. Li, "Divination Engines," 143.

21. Li, 144.

22. Brown and Mercer, "Oh, Yes, Everything's Right on Schedule, Fred."

23. Lem, "First Sally (A), or Trurl's Electronic Bard," 199.

24. Lem, 199.

25. Brown and Mercer, "Oh, Yes, Everything's Right on Schedule, Fred."

26. Marcus, Marcinkiewicz, and Santorini, "Building a Large Annotated Corpus of English."

27. Klimt and Yang, "Enron Corpus."

28. Wood, Massey, and Brownell, "FERC Order Directing Release of Information," 12.

29. Heller, "What the Enron Emails Say about Us."

30. Baker et al., "Research Developments and Directions in Speech Recognition."

31. I have participated in early work to address this gap. See, e.g., Gebru et al., "Datasheets for Datasets." Other researchers have also sought to address this problem for AI models; see Mitchell et al., "Model Cards for Model Reporting"; Raji and Buolamwini, "Actionable Auditing."

32. Phillips, Rauss, and Der, "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results," 9.

33. Phillips, Rauss, and Der, 61.

34. Phillips, Rauss, and Der, 12.

35. See Aslam, "Facebook by the Numbers (2019)"; and "Advertising on Twitter."

36. Fei-Fei Li, as quoted in Gershgorn, "Data That Transformed AI Research."

37. Deng et al., "ImageNet."

38. Gershgorn, "Data That Transformed AI Research."

39. Gershgorn.

40. Markoff, "Seeking a Better Way to Find Web Images."

41. Hernandez, "CU Colorado Springs Students Secretly Photographed."

42. Zhang et al., "Multi-Target, Multi-Camera Tracking by Hierarchical Clustering."

43. Sheridan, "Duke Study Recorded Thousands of Students' Faces."

44. Harvey and LaPlace, "Brainwash Dataset."

45. Locker, "Microsoft, Duke, and Stanford Quietly Delete Databases."

46. Murgia and Harlow, "Who's Using Your Face?" When the *Financial*

Times exposed the contents of this dataset, Microsoft removed the set from the internet, and a spokesperson for Microsoft claimed simply that it was removed "because the research challenge is over." Locker, "Microsoft, Duke, and Stanford Quietly Delete Databases."

47. Franceschi-Bicchierai, "Redditor Cracks Anonymous Data Trove."

48. Tockar, "Riding with the Stars."

49. Crawford and Schultz, "Big Data and Due Process."

50. Franceschi-Bicchierai, "Redditor Cracks Anonymous Data Trove."

51. Nilsson, *Quest for Artificial Intelligence,* 495.

52. And, as Geoff Bowker famously reminds us, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care." Bowker, *Memory Practices in the Sciences,* 184–85.

53. Fourcade and Healy, "Seeing Like a Market," 13, emphasis added.

54. Meyer and Jepperson, "'Actors' of Modern Society."

55. Gitelman, *"Raw Data" Is an Oxymoron,* 3.

56. Many scholars have looked closely at the work these metaphors do. Media studies professors Cornelius Puschmann and Jean Burgess analyzed the common data metaphors and noted two widespread categories: data "as a natural force to be controlled and [data] as a resource to be consumed." Puschmann and Burgess, "Big Data, Big Questions," abstract. Researchers Tim Hwang and Karen Levy suggest that describing data as "the new oil" carries connotations of being costly to acquire but also suggests the possibility of "big payoffs for those with the means to extract it." Hwang and Levy, "'The Cloud' and Other Dangerous Metaphors."

57. Stark and Hoffmann, "Data Is the New What?"

58. Media scholars Nick Couldry and Ulises Mejías call this "data colonialism," which is steeped in the historical, predatory practices of colonialism but married to (and obscured by) contemporary computing methods. However, as other scholars have shown, this terminology is double-edged because it can occlude the real and ongoing harms of colonialism. Couldry and Mejías, "Data Colonialism"; Couldry and Mejías, *Costs of Connection;* Segura and Waisbord, "Between Data Capitalism and Data Citizenship."

59. They refer to this form of capital as "ubercapital." Fourcade and Healy, "Seeing Like a Market," 19.

60. Sadowski, "When Data Is Capital," 8.

61. Sadowski, 9.

62. Here I'm drawing from a history of human subjects review and large-scale data studies coauthored with Jake Metcalf. See Metcalf and Crawford, "Where Are Human Subjects in Big Data Research?"

63. "Federal Policy for the Protection of Human Subjects."

64. See Metcalf and Crawford, "Where Are Human Subjects in Big Data Research?"

65. Seo et al., "Partially Generative Neural Networks." Jeffrey Brantingham, one of the authors, is also a co-founder of the controversial predictive policing company PredPol. See Winston and Burrington, "A Pioneer in Predictive Policing."

66. "CalGang Criminal Intelligence System."

67. Libby, "Scathing Audit Bolsters Critics' Fears."

68. Hutson, "Artificial Intelligence Could Identify Gang Crimes."

69. Hoffmann, "Data Violence and How Bad Engineering Choices Can Damage Society."

70. Weizenbaum, *Computer Power and Human Reason,* 266.

71. Weizenbaum, 275–76.

72. Weizenbaum, 276.

73. For more on the history of extraction of data and insights from marginalized communities, see Costanza-Chock, *Design Justice;* and D'Ignazio and Klein, *Data Feminism.*

74. Revell, "Google DeepMind's NHS Data Deal 'Failed to Comply.'"

75. "Royal Free–Google DeepMind Trial Failed to Comply."

4
Classification

1. Fabian, *Skull Collectors.*

2. Gould, *Mismeasure of Man,* 83.

3. Kolbert, "There's No Scientific Basis for Race."

4. Keel, "Religion, Polygenism and the Early Science of Human Origins."

5. Thomas, *Skull Wars.*

6. Thomas, 85.

7. Kendi, "History of Race and Racism in America."

8. Gould, *Mismeasure of Man,* 88.

9. Mitchell, "Fault in His Seeds."

10. Horowitz, "Why Brain Size Doesn't Correlate with Intelligence."

11. Mitchell, "Fault in His Seeds."

12. Gould, *Mismeasure of Man,* 58.

13. West, "Genealogy of Modern Racism," 91.

14. Bouche and Rivard, "America's Hidden History."

15. Bowker and Star, *Sorting Things Out,* 319.

16. Bowker and Star, 319.

17. Nedlund, "Apple Card Is Accused of Gender Bias"; Angwin et al., "Machine Bias"; Angwin et al., "Dozens of Companies Are Using Facebook to Exclude."

18. Dougherty, "Google Photos Mistakenly Labels Black People 'Gorillas'"; Perez, "Microsoft Silences Its New A.I. Bot Tay"; McMillan, "It's Not