Data & Civil Rights: Technology Primer

by Solon Barocas, Alex Rosenblat, danah boyd, Seeta Peña Gangadharan, and Corrine Yu

Produced for Data & Civil Rights Conference / October 30, 2014

Data have assumed a significant role in routine decisions about access, eligibility, and opportunity across a variety of domains. These are precisely the kinds of decisions that have long been the focus of civil rights campaigns. The results have been mixed. Companies draw on data in choosing how to focus their attention or distribute their resources, finding reason to cater to some of its customers while ignoring others. Governments use data to enhance service delivery and increase transparency, but also to decide whom to subject to special scrutiny, sanction, or punishment. The technologies that enable these applications are sometimes designed with a particular practice in mind, but more often are designed more abstractly, such that technologists are often unaware of and not testing for the ways in which they might benefit some and hurt others.

The technologies and practices that are driving these shifts are often described under the banner of "big data." This concept is both vague and controversial, particularly to those engaged in the collection, cleaning, manipulation, use, and analysis of data. More often than not, the specific technical mechanisms that are being invoked fit under a different technical banner: "data mining."

Data mining has a long history in many industries, including marketing and advertising, banking and finance, and insurance.[1] As the technologies have become more affordable and the availability of data has increased, both public and private sectors—as well as civil society—are envisioning new ways of using these techniques to wrest actionable insights from once intractable datasets. The discussion of these practices has prompted fear and anxiety as well as hopes and dreams. There is a significant and increasing gap in understanding between those who are and are not technically fluent, making conversations about what's happening with data challenging. That said, it's important to understand that transparency and technical fluency is not always enough. For example, those who lack technical understanding are often frustrated because they are unable to provide oversight or determine the accuracy of what is produced while those who build these systems realize that even they cannot meaningfully assess the product of many algorithms.

This primer provides a basic overview to some of the core concepts underpinning the "big data" phenomenon and the practice of data mining. The purpose of this primer is to enable those who are unfamiliar with the relevant practices and technical tools to at least have an appreciation for different aspects of what's involved.

Key Definitions

Terms like "big data" and data mining obscure as much as they reveal. They are both often used to describe nearly any attempt to find something of value in a dataset. Although "big data" is primarily a buzzword, data mining does have a more precise technical meaning that warrants careful consideration. It is also worth explaining how data mining relates to other key concepts.

Data: Data are the most basic unit of enumeration. They index, represent, or describe some entity or phenomenon by translating these into abstract inscriptions like numbers, letters, and other symbols. In a narrower conception, data are ways to classify and quantify observations, frequently through measurement techniques that involve special instruments. Most familiar are data in the form of so-called "categorical" or "continuous" variables - recordings that have parsed observations in terms of fixed categories (e.g., that's a cat) or incremental numerical ranges (e.g., that cat weighs 10.3 pounds). There is, invariably, a politics to categories and counting. Contrary to what the etymology of the word might suggest, data, from the Latin *datum*, the givens, are never simply there for the taking. They are artifacts of human intervention, not facts imparted by nature itself. And any discussion of the data and civil rights should attend to the conditions and nature of data's production.

Metadata: Metadata is data *about* data and it is plays a crucial role in the practical functioning of data analytics. Think of the top row of a spreadsheet, the row that often has a title for each column. Absent this information (or some outside knowledge), a table with many long columns of numbers is meaningless. Metadata provides this meaning. It furnishes a definition for the cells and fields that data occupy. Metadata has become a more familiar term in the wake of the Snowden revelations, which suggested that the state relied on a distinction between the content of phone calls (i.e., the actual sounds communicated by phone) and metadata for those calls (e.g., the participants in the call) to explain the noninvasiveness of its surveillance activities. Simply knowing who is talking – the metadata – is extremely revealing.

Big Data: This term, firmly established in the popular discourse, most often to refers to data that is either too large or too complex for an institution to process and interpret with legacy data management tools (see entry on database below). There are no precise technical definitions, but big data is often discussed in terms of volume (i.e., the size of the dataset), variety (i.e., the diversity of data types), and velocity (i.e., the rate of accumulation). It is also used to reference a broader contemporary phenomenon in which data have been invested with hype that reaches mythical proportions; the technology is not as advanced as many imagine it to be.[2] In the context of this event, we use "big data" to reference the cultural dimensions of data analytics, technological development, and organizational shifts that are currently underway.

Database: Data that have been organized in some way are commonly described as databases. Generally speaking, databases impose structure on the data by defining the specific fields into which different kinds of data should slot and the entities to which these should relate. They provide a logical schema for parsing data and for maintaining associations between different pieces of data and the entities that they describe. To take a simple example, an address book is a kind of database because it parses contact information into its constituent fields (i.e., name, address, phone number, etc.) and associates this information with a specific person. In much the same way, computerized databases provide a means to formally represent the relationship between different fields, but they also allow operators to redefine and create new relationships on the fly. For instance, they can allow for the creation of new databases that pull the names and addresses from an address book and match these criminal records. This modularity is characteristic of all so-called "relational databases". Emerging approaches to the management of data do away with the kinds of "tabular" representations that are common to relational databases. So-called "post-relational" databases are structureless in the sense that they do not require that fields be specified in advance. Any piece of data can be 'tagged' after the fact. You may hear terms like NoSQL (a genre of database), Hadoop (a database product), and MapReduce (a programming model) associated with this emergent type of database.

Aggregation**:** Aggregation refers to the assembly of data from multiple sources. This can take many forms: joining together datasets form different institution that each contains the same kind of information (e.g., pooling the medical records for all patients with diabetes from different hospitals), matching records from different institutions that each refers to the same person (e.g., piecing together John Doe's medical and financial records), or a combination of the two. The second type of aggregation often employs a 'fuzzy' matching strategy—that is, formal rules that attempt a best match between records based on similar—but not perfectly overlapping—identifiers (e.g., matching John Doe at 123 Main Street with John J. Doe at 123 Main St.).[3] These are probabilistic determinations, and occasionally result in mismatches that saddle one John Doe with the unfavorable records of a different John Doe.

Aggregation poses other risks, too. Separate bits of information, readily shared in discrete moments and distinct contexts, may serve as the basis for discoveries that are more than the simple sum of their parts. Taken together, these facts can support deductive inferences that could wrest what once seemed like 'private' facts from 'public' information. Consider a contrived but instructive example: one set of records indicates that John Doe always visits a specific location on Tuesday afternoon; another set of records shows that the only event that takes place in this location at that time is an Alcoholics Anonymous meeting; combined, these records reveal that John Doe attends Alcoholics Anonymous meetings.[4] In such cases, clever—rather than especially keen—observers can, deducing from more easily observable behavior (where John Doe travels in his car) and more easily accessible information (what events take place at a certain location) that something otherwise more difficult to discern must be true as well (that John Doe attends Alcoholics Anonymous meetings).

The capacity to draw deductive inferences from aggregated observations of everyday activities has been the subject of recent legal analyses (considered by the Supreme Court), culminating in what some now call a 'mosaic theory' of privacy.[5] This work builds on the notion that, pieced together, a sequence of observations may reveal rather sensitive details about an individual's habits, associations, and beliefs, throwing into stark relief the problem with the public-private distinction upon which Fourth Amendment doctrine has long relied to judge what constitutes an unreasonable search.

Algorithms*:* Algorithms refer to specified sequences of logical operations designed to accomplish a particular task. They are step-by-step instructions for acting on some kind of input to achieve a desired result. This could be as simple as a set of instructions for adding the value of two input variables together or as complicated as instructions for rank-ordering the websites that are most relevant to an inputted keyword query. In trying to determine how to solve a problem computationally, developers devise an algorithm. They try to figure out how to break a problem into a series of questions that a particular sequence of logical operations can then answer—and which a computer can then execute automatically. In this sense, algorithms are a kind of abstract strategy for problem-solving; a computer program implements these strategies more concretely as a set of logical operations expressed in formal code.

This more generic definition of algorithm encompasses any kind of decision-making that passes data through a fixed and formal decision procedure, whether or not it results in a computer-automated decision. In practice, algorithms tend to be deployed in three different ways: (1) as fully automated systems that render and apply decisions without any human intervention; (2) as semi-automated systems that integrate human operators into part of the decision-making process; or (3) as decision-support systems that return results directly to a human as information or advice.[6]

Machine Learning: Machine learning is a sub-field within computer science that grew out of artificial intelligence. Unlike traditional artificial intelligence, which attempted to hand code the logical operations involved in human cognition, machine learning 'trains' computers how to reason. Specifically, machine learning is a kind of learning by example, one in which an algorithm is exposed to a large set of examples from which it has been instructed to draw general lessons. "Learning" occurs when the algorithm extracts logical rules that are not simply a recapitulation of the specific properties of the examples. To get a better handle on this, consider cats. Humans routinely recognize that a cat is indeed a cat, even if they have never encountered the particular cat before. We extract a general concept of cat from the necessarily limited set of cats to which we have been exposed, and this concept allows us to recognize other, future cats. These general concepts involve a complex set of characteristics that allow humans to discern that the animal before them is indeed a cat (i.e., four legs, fur, cuddly, etc…). This kind of learning is so foundational to human cognition that it can be hard to appreciate the challenge it poses for computers. The core task of machine learning is to generalize from examples—to be able to learn general rules that allow a computer to make sense of cases that are not simply those to which it has been exposed.

Computers discover the telltale signs of a kind of 'cat-ness' by testing to see which specific set of details from the examples happen to distinguish cats from other entities. This is more challenging that it might seem. Moving from the specificity of the limited set of examples of cats to a more general concept of cat requires separating out those features that are relevant in recognizing *all* possible cats from those features that just happen to distinguish cats in the set of examples under consideration. Learning is really a matter of abstracting a general concept of cat from the infinite specificity of the examples of cats. A concept that is too closely wedded to the peculiarities of the examples will fail to generalize to the unknown and slightly different cats that the machine will encounter in the future. This is known as the problem of "overfitting" and it is one of the central preoccupations of machine learning as a field.[7] For machine learning to successful teach a computer to recognize cats, it must privilege but—crucially—also discount certain details from the examples to arrive at a more general concept of cat that will allow the computer to recognize any future cats it will encounter. An overly narrow concept of cats might expect all cats to have the same color patterns in their coats as those in the examples. Underfitting is no less of a problem, and can result in a concept of cat that is so general that it encompasses other animals that happen to be furry and four-legged.

As these examples suggest, over- and underfitting can result in two types of misclassification: an individual can be incorrectly categorized into a group to which they do not actually belong; this is a type I error—a false positive. Or an individual can be mistakenly excluded from a category to which they actually do belong; this is a type II error—a false negative. Setting an acceptable false positive and false negative rate depends on the context-specific consequences of those errors. Further complicating this decision is the fact that models with the *same* error rates may distribute those errors differently across a population. Dealing with these errors can be especially troublesome because they tend to occur at inverse rates: as you might expect, a decrease in the false negative rate comes at the cost of an increase in the false positive rate. This trade-off is yet another motivating problem within machine learning. Although a number of tools exist to help express and therefore evaluate these trade-offs, finding right balance between the true positive rate and false positive rate is a matter of necessarily subjective judgment. What might be acceptable for marketing might be intolerable for medicine. Still, as Anthony Danna and Oscar Gandy note, in many contexts these errors are often accepted because "the benefits derived from the small percentage of predictions that are 'right' outweigh the costs of not having made any predictions at all."[8] Unless specifically structured to do so, common evaluation methods will be agnostic on the objectionableness of *distribution* of errors; they will not take into consideration that the 5% error rate actually reflects that the lessons learned apply very effective for the majority of the population, but happen to be ineffective for a specific minority population.[9]

Note, too, that helping a machine to learn a concept of cat in this way is radically different than hand-coding a set of explicit logical statements that describe the defining features of a cat. In effect, machine learning furnishes the computer with a set of tools and large set of examples to teach itself the defining features of a cat. For all its strengths, this approach involves some serious hazards. First, to develop a sufficiently nuanced concept of a cat, machine learning requires an enormously large set of examples. Indeed, the number of examples tends to be the single most important factor in successful applications of machine learning, explaining much of the excitement about big data among practitioners.[10] It also needs a diverse set of examples. Stated simply, the computer will not be able to learn from examples to which it has not been exposed. If, for whatever reason (and there are many),[11] the set of examples fails to represent the extent of variability in cat appearance, machine learning will fail to impart the lessons necessary for a computer to recognize the types of cats omitted from the examples. Third and relatedly, the lessons learned may be rather fragile because cat breeds may evolve over time, developing new appearances and qualities that differ from those in the examples. In effect, the concept of cat may become outdated. Fourth, a computer can learn the wrong lessons if the assembled set of examples happens to include other animals besides cats. If, by mistake, images of dogs have been labeled examples of cats, these examples may confuse the machine into thinking that certain properties of dogs are also good indicators of whether something is a cat. The danger here is that machine learning must accept the labels assigned to the examples as ground truth, even where there is the possibility that certain examples may have been mislabeled. Finally, nuance often comes at the cost of significant complexity. This means that humans may sometimes struggle to understand what, exactly, the machine has learned, even when it has learned to do a good job of recognizing cats. The specific set and weighing of features that the machine has identified as the telltale signs of a cat may be so complex that they are impossible for a human to interpret. In fact, the most successful applications of machine learning in many domains are frequently those that are most inscrutable. Practitioners have long accepted that, in certain settings, "predictive accuracy is paramount; the need to use the extracted knowledge to explain a recommended action is less important."[12] To put it simply: successful learning may come at the expense of understanding.

All that said, this approach has been enormously successful across a wide variety of domains. It is the driving force behind so-called "computer vision" (i.e., automating the identification of entities in images (e.g., cats)), natural language processing (i.e., automating the interpretation of a text), speech recognition (i.e., automating the recognition of words), and high frequency trading (i.e., automating the detection of lucrative trading opportunities), to name just a few. As this final example suggests, machine learning can not only automate, but also far exceed the kinds of interpretative work performed by humans. It can learn to do things that humans cannot. It can teach a computer to recognize signals to which humans are not attuned or detect subtle patterns that would escape human notice.

Data Mining**:** Commercial applications of machine learning are routinely described as data mining. Familiar applications known as data mining include spam or fraud detection, credit scoring, and insurance pricing. By exposing machine learning algorithms to previously identified instances of spam, fraud, default, and poor health, the algorithm learns which related attributes or activities happen to correlate with those qualities or outcomes of interest. This accumulated set of discovered relationships would be known as a "concept" in machine learning; in data mining, these are know as a "predictive model" or simply a "model". Adopting the model as a decision-aid allows institutions to treat spam, fraud, default, and poor health as a function of some other set of observed characteristics, and to automate the process of making decisions that turn on these inferences. Email providers can detect spam by looking at what the model has revealed as telltale signs.

As this suggests, real world applications of data mining actually involve two related, but very different forms of computer automation. In the first instance, data mining (like machine learning) automates the process of learning what distinguishes spam from legitimate email from a series of prior examples. That is, it automates the process of testing for useful statistical correlations between the properties of these example emails and whether they are or are not spam. But the results of this process—the discovered patterns and model—help to automate a second and very different set of operations: future inferences and decisions that involve similar types of data. This second order automation can resemble a kind of 'mining' because it will often routinize the process of evaluating the significance of incoming data and thus help identify cases of the relevant activity in the subsequent flow of information (e.g., whether an email is spam).

Early on, there was an attempt to distinguish data mining—the analytic step—from so-called Knowledge Discovery in Databases (KDD)—the entire process involved in undertaking a data mining project—the terms are now used interchangeably, if KDD is used at all.[13] Still, the original difference between the two term is significant and telling: while data mining referred exclusively to the application of machine learning algorithms, KDD referred to the overall process of transforming a question into a data mining problem, assembling and preparing the relevant data, subjecting the data to analysis, and interpreting and implementing the results.[14] Today's more fashionable terms—business analytics, predictive analytics, and data science—tend to share this more expansive definition.

[1] Heather Allen, Paul Gearan, and Karl Rexer, "2013 Data Miner Survey: Summary Report," (Winchester, MA: Rexer Analytics, 2013).

[2] boyd, danah and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication, & Society* 15, 5 (2012): 662-679.

[3] Data matching was a major concern in the seventies and eighties when government agencies first began to automate the selective matching of records across services (e.g., across the criminal, tax, and welfare systems). The 1988 Computer Matching and Privacy Protection Act, an amendment to the 1974 Privacy Act, responded to these concerns by legislating procedural uniformity, increased due process, and oversight in the matching programs. Although data matching has fallen out of favor as a topic of debate, now largely eclipsed by data mining, it remains an extremely common practice.

[4] Even though there are many possible reasons why he might attend such meetings (e.g., because he is a recovering alcoholic, a discussion facilitator, a caterer, etc.), the fact remains that he *attends* these meetings.

[5] See, for example, Orin S. Kerr, "The Mosaic Theory of the Fourth Amendment," *Michigan Law Review* 111, 3 (2012): 311–354; David C. Gray and Danielle Keats Citron, "A Shattered Looking Glass: the Pitfalls and Potential of the Mosaic Theory of Fourth Amendment Privacy," *North Carolina Journal of Law and Technology* 14, 2 (2013): 381–430.

[6] This is an adaptation of Citron's categorization; she refers to semi-automated systems as mixed systems. Danielle Keats Citron, "Technological Due Process," *Washington University Law Review* 85 (2007): 1249-1313.

[7] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge Discovery in Databases: an Overview," *AI Magazine* 13, 3 (1992): 57-70. http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929.

[8] Anthony Danna and Oscar H. Gandy, Jr., "All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining," *Journal of Business Ethics* 40, 4 (2002): 379.

[9] Moritz Hardt, "How big data is unfair: Understanding sources of unfairness in data driven decision making," *Medium*, September 27, 2014. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

[10] Alon Halevy, Peter Norvig, and Fernando Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24, 2 (2009): 8–12. *See also,* Brian Dalessandro, Claudia Perlich, and Troy Raeder, "Bigger Is Better, but at What Cost? Estimating the Economic Value of Incremental Data Assets," *Big Data* 2, 2 (2014): 87–96, accessed November 20, 2014. doi:10.1089/big.2014.0010.

[11] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *SSRN*, October 19, 2014. http://ssrn.com/abstract=2477899.

[12] Ronald Brachman, Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis, "Mining Business Databases," *Communications of the ACM* 39, 11 (1996): 42-8.

[13] Frawley, Piatetsky-Shapiro, and Matheus, "Knowledge Discovery in Databases,"; Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* 17, 3 (1996): 37-54.

[14] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge From Volumes of Data," *Communications of the ACM* 39, 11 (1996): 29.