Chapter 2
Open Data, Big Data, and Just Data

Abstract This chapter examines two cases in which data presents questions of justice. Many argue as a philosophical principle that data sources should be available as widely as possible, the principle at the heart of the open data movement. But as I argue in that chapter, open data can just as easily lead to injustice: Like programming, "Injustice in, injustice out" ought to be a principle of data. Social privilege can color the data that is opened and create serious inequalities in who can access and use ostensibly open data. Open data can also establish standards that exclude knowledge that is not part of the data system. In the second case, I consider what big data means for higher education. After discussing some recent examples, I identify two types of ethical challenges in the increasingly common use of predictive analytics at universities: challenges related to the direct consequences of the systems and those rooted in the ideology of scientism that inspire them. Both the open data and big data cases prove quite problematic if the aim is just data.

> "Technology is neutral—it's what you do with it that turns it into a public good." Condoleezza Rice, at ASU-GSV Summit. (@DeanOlian [Judy Olian] 2016)

> "So for example, pollution in China, environmental degradation is a hot political topic in China, and people can walk outside of their apartments, or wherever, and they know they can't breathe in Shanghai or Chengdu, or whatever. And for a long time the government was giving a pollution index number that clearly didn't bear any resemblance to reality. The U.S. Embassy started publishing a number or putting a number up, but there's also now apparently an app that you can buy that will measure the pollutants. So just the provision of information challenges the monopoly on information that an authoritarian government depends on for control and acquiescence." Condoleezza Rice (Freedman 2014)

With due respect to the former Secretary of State, technology cannot be both neutral and deterministic. If it is neutral, then authoritarian regimes will be able to use it to further their monopoly on information. If it is inherently democratic, then it is a public good regardless of what one does with it. Rice's contradiction, like many who make such arguments, is in part rooted in an equivocation on the meaning of "technology." In the former, technology is a pure object, often even an abstract concept: technology is neutral in that one can build specific technologies to further (more or less) any end. In the latter, technology has become embodied and purposed:

J.A. Johnson, *Toward Information Justice*, Public Administration and Information Technology 33, https://doi.org/10.1007/978-3-319-70894-2_2

a cell phone is used to provide information that "get[s] a little doubt in." Google's Jared Cohen argues:

> But the one silver lining in all of this is, the totalitarian societies—the true cults of personality—have literally been eliminated by the Internet in the same way that scientists were able to get rid of smallpox. Once North Korea changes, you'll never see a cult of personality again, because the ability to create a society without doubt will no longer be possible. (Freedman 2014)

And yet Donald Trump's campaign and presidency—not to mention cults of personality venerating Silicon Valley elites—certainly suggest the internet can do exactly that. Neither neutrality nor determinism seem to be effective in understanding the social effects of information technology.

This chapter takes a more complex view of the social effects of information technology (both in the abstract and in the form of specific technologies). Instead of assuming that information technology will be inherently good for society, I explore the way that questions of justice arise when information technologies are implemented in a society. I study two cases in which information technology has been held to be inherently and deterministically good: the open data movement and the use of big data and learning analytics in higher education. Each case explores injustice along different registers, open justice showing the mechanics of injustice, and big data demonstrating the conceptual levels at which injustice emerges. In both cases, the initial claims of, essentially, technological determinism founder on the myriad connections between technology and society, such as the values and assumptions built into the technologies, the complex of problems and applications in which the technologies are used, and the social structures within which the technologies operate. The argument is not that open data and learning analytics are inherently bad; such would be every bit as deterministic as their advocates argue. But Kranzberg's (1986) famed formulation that "Technology is neither good nor bad; nor is it neutral" applies well here. The fact that information technologies can have good or bad (more realistically, good *and* bad) outcomes does not mean that they are neutral either. The values and structures of technology in society ensure that any information technology will raise complex questions of justice.

2.1 Opening Government Data

With the proliferation of data in contemporary information societies comes an increasingly common call for that data to be publically accessible: an open data movement. This movement claims that open data will support democratic politics and individual liberty, unequivocally allowing individuals to use the wealth of data produced by governments and enterprises greater control over their lives and improving both their material and social conditions. "Free-as-in-speech" software and the aphorism that "Information wants to be free" as well as a distrust of political authority and consequent belief that "sunlight is the best disinfectant" have led many to argue as a philosophical principle that data sources should be available as widely as possible:

> The Internet is the public space of the modern world, and through it governments now have the opportunity to better understand the needs of their citizens and citizens may participate more fully in their government. Information becomes more valuable as it is shared, less valuable as it is hoarded. Open data promotes increased civil discourse, improved public welfare, and a more efficient use of public resources. (Open Data Working Group 2007)

The movement has come to be reflected in public policy. The U.S. government implemented an open data policy through the Office of Management and Budget's Open Government Directive, which called for agencies throughout the executive branch to take steps promoting transparency, participation, and collaboration in the publication and use of government data (Orszag 2009). Whether public or private, open data generally consists of a commitment to make data available publically in non-proprietary, machine-readable formats at the lowest level of granularity possible. As expressed by the U.S. National Science Foundation (2012), "The key principle being applied in executing the elements of the NSF Open Government Plan is: *Unless shown otherwise, the default position shall be to make NSF data and information available in an open machine-readable format.*" Similar programs range from international organizations such as the EU INSPIRE directive to local governments (Rich 2012).

This view of open data as inherently democratic is problematic, as we shall see, rooted in both a naïve view of technology and a simplistic view of politics. Open data has the quite real potential to exacerbate as much as alleviate injustices. So it comes as no surprise that open data's track record does not match its promises. The digitization of land records in the Karnataka region of India is a widely discussed case in point (Donovan 2012; Gurstein 2011; Raman 2012; Slee 2012). Three programs digitized the Record of Rights, Tenancy, and Crops (one type of land title record among others); the age, caste, and religion of owners and tenants; and spatial data. The former programs (called *Bhoomi* and *Nemmadi*, respectively) were created by the state government; the latter was part of the National Urban Information Systems program developed by the Government of India. A public-private partnership made the information accessible through internet kiosks deployed throughout the state. The promise was a system that would increase transparency and secure the rights of land tenants. The reality was a system that shifted power—and land—from local landholders to real estate developers. This is, unfortunately, rather typical of open data projects that simply approach openness as a technical condition of access. Such approaches to openness present challenges to justice in a number of ways.

2.1.1 Injustice In, Injustice Out

The constructed nature of data makes it quite possible for injustices to be embedded in the data itself. Whether by design or as unintended consequences, the process of constructing data builds social values and patterns of privilege into the data. Where those values and privileges are unjust, the injustice is then a characteristic of the data itself; no amount of openness can remedy such injustices, just as no amount of

statistical processing can undo inaccuracies in the original data. "Garbage in, garbage out" is a central concept in data ethics.

Data emerges often in the interaction of an individual with a bureaucratic organization such as the state or a business. But people and groups differ in their propensity to interact with such organizations. This difference provides an important point by which privilege can enter into data. Data over-represents some, and where those over-representations parallel existing structures of social privilege, it over-represents those already privileged and under-represents those less likely to be part of data producing interactions.

Interactions with the state are rife with disparities that reflect social privilege. One well-studied example is the undercount of the decennial U.S. census (Prewitt 2010). Since the problem of undercounting was first quantified in the mid-Twentieth Century, black and Hispanic households have been undercounted at higher rates than non-black households. The causes of this undercount are myriad:

> Households are not missed in the census because they are black or Hispanic. They are missed where the Census Bureau's address file has errors; where the household is made up of unrelated persons; where household members are seldom at home; where there is a low sense of civic responsibility and perhaps an active distrust of the government; where occupants have lived but a short time and will move again; where English is not spoken; where community ties are not strong. (Prewitt 2010, p. 245)

Two commonalities in these explanations are striking: the extent to which these causes are barriers to interaction with census takers and the extent to which they are correlated with racial and class privilege. The latter causes the undercount to disproportionately affect disadvantaged groups (hence, Prewitt argues, the focus on race in debates over census methodology between 1980 and 2000), while the former prevents those groups from being represented accurately in census data. Similar problems exist in collecting data on groups such as the homeless (Williams 2010).

Groups might also be disproportionately willing to participate in some interactions over others, such as differences in thresholds for reporting building code violations between the affluent and poor (Schönberger and Cukier 2013). This is an especially significant problem in the collection of public health data on minorities, where trust in government may be lagging. Migrant groups, especially indigenous groups, refugees, and undocumented workers, frequently fear that data collected by the state will be used to their disadvantage. In many cases, such communities maintain gatekeeper institutions through which outsiders must work in order to interact effectively with the community. These groups use such structures in part as protection from states and social actors that have histories of conflict with the group, or where the groups are accustomed to high-context institutions that provide a basis for trust. But the result is that even where such groups want the data being collected, the processes that generate trust in the data collectors exclude them from the datasets.[1] Since those groups tend to be those that lack privilege, this also embeds privilege in data.

[1] Evelyn Cruz, e-mail correspondence, March 29–31, 2013.

Such privileges are not confined to interactions with the state. Residential segregation especially is often tied to forms of institutional discrimination that would influence how often individuals interact with bureaucracies. Zenk et al. (2005) found that low-income, predominantly African-American neighborhoods in Detroit were, on average, 1.1 miles further from a supermarket than predominantly white neighborhoods with similar incomes, with consequently increased dependence on smaller food stores such as convenience stores or groceries. Similarly, Cohen-Cole (2011) argues that consumer credit discrimination based on the racial composition of applicants' neighborhoods is linked to increased use of payday loans. In both cases, the use of less bureaucratized businesses by groups already suffering from discrimination in the form of de facto residential segregation (either as the legacy of formal segregation or because of ongoing discrimination) results in data that is statistically biased against such populations and reinforces whites' privileged position. Businesses can analyze the needs of the (disproportionately white) customers with whom they interact and adapt accordingly; benefits thus accrue to the beneficiaries of social privilege.

Transforming information about a datized moment into data is equally problematic. Only some of the information about that moment will be datized. What information will be is not a natural consequence of the interaction but a design choice on the part of the data architects that reflects their purposes, resources, and values. An institutional survey director noted to me that survey data at the institution is subject to state open records laws and sometimes requested by the public and state legislators. As a result, the survey director encouraged the practice of not collecting data that the institution would not be comfortable making public.[2] In this case, the concern was privacy, but this reasoning is at least as likely when more self-interested motives are present. Regardless of the motivation, though, such decisions are value-laden; thus the data built on such decisions will embody those values and transmit them in the process of using the resulting data.

Less conscious assumptions such as those part of worldviews shaped by social privilege will also shape such decisions and likely be less amenable to challenge to the extent of their invisibility to lack of diversity among the data collectors. Higher education "net price calculators," which the U.S. government requires all institutions receiving Title IV aid to produce, are designed to help students and their families estimate the likely cost of attending an institution given the prevalence of "high-tuition, high-aid" business models. This assumes that the net price is what is important to students. But Sara Goldrick-Rab (2013) argues that the gap in applications to elite colleges between high-achieving, high-income and high-achieving, low-income students reported by Hoxby and Avery (2012) is rooted in "sticker shock" at the high gross price of such institutions among low-income families in spite of the institutions' often much lower net prices. Their disregard of net price is in part a lack of information, but more significantly a consequence of such families' lack of trust in institutions generally and substantially higher risk to such families if educational institutions fail to maintain the initial promises of aid, conditions that

[2] Jane Doe (pseudonym), personal communication, March 20, 2013.

make the net price of the institutions less credible: "Being told that a college is *likely* to give you aid is not the same thing as *getting the aid*, [emphasis in original]" Goldrick-Rab writes. Such students choose to apply at less expensive (and consequently less selective) institutions as they present less risk to themselves and their families.

If Goldrick-Rab is correct, the credibility that the middle class finds in state and social institutions that have generally protected their interests should be seen as underlying the decision to collect and report average aid amounts that do not vary by income: Middle class families can credibly take average aid as typical of people like them; low-income families cannot. One might expect the same to be true of first-generation students. With family members unfamiliar with the operations of universities, they will often be unaware of issues such as net price or even understand the financial aid process at all. Yet this background knowledge, like the credibility of a measure, is assumed in the selection of data to be collected. Those privileged with such knowledge find their privileges reinforced by this data; those who are not so privileged are further disadvantaged when they cannot see the data as meaningful.

Thus we find the outcome of the digitization of land records in India described in Sect. 2.1. The selection of the RTC as the definitive data form had consequences for the distribution of land ownership. Raman argues that the programs result in the exclusion of the claims of the Dalit caste (often referred to as "untouchables"), which are often not documented in the RTC records but are well supported in other sources. Adding to this the question of how that information is stored increases the complexity of the issue. Key features in the problematic *Bhoomi* experience with open data were not only the selection of only certain types of documentation for inclusion in the land title data but also the decision to store the resulting data in a relational database system (Raman 2012). These aspects of the system design effectively precluded informal and historical knowledge from being part of the open data system; such knowledge, which was the basis of the existing land claims of *Dalits*, cannot be queried by the systems used to store the data. The two features both inform and reinforce each other: excluding narratives and other unstructured data obviates the need for systems that can handle unstructured data such as those using text-analytics or Unstructured Information Management Architecture (UIMA), while the choice of a relational database precludes the use of narrative information.

The choice of the RTC and demographic data, and the decision to accord only the RTC legal status, is also a consequence of the programs' homes in the state department of revenue, as this data was already held by these departments and is needed by the department in the course of their responsibilities. But it also reflects a bureaucratic mindset:

> The architects of the *Bhoomi* and the *Nemmadi* projects viewed the prevalence of multiple records as a manifestation of "inefficient record keeping", "corruption of field bureaucrats" and the opacity of land records due to lack of modern systems of documentation …. They sought to resolve the conflicts by identifying a single owner to a single plot of land by according a legal status to the digital RTC. (Raman 2012)

This bureaucratic mindset builds data that reflects the bureaucratic values of efficiency and consistency, doing so at the cost of excluding data that cannot be accommodated to those values. Donovan (2012) cites this as an instance of Scott's (1998) "seeing like a state," in which the local government sought to simplify society by making it legible. The open data system incorporated this value in its choice of what to datize about the moment in which land was transferred. This incorporated a value structure into the data, one that is clearly not neutral in the competition for power.

Because of the myriad ways that social privilege can become embedded in datasets, open data cannot be expected to universally promote justice. It can just as easily marginalize groups that are not part of the data, people whose lack of privilege excludes them from the kinds of interactions that produce data and makes their viewpoints invisible to those who collect data. Opening datasets composed of such data simply propagates the injustices that came into the data as it was collected. Whatever steps are taken to promote fairness in using data that is at its root unjust, the result will almost inevitably be unjust as well. Data is very much a case of "Injustice in, injustice out."

2.1.2 Open to Whom?

Normatively "clean" data is a necessary starting point for the just use of data, but it is by no means sufficient to ensure just outcomes. While open data advocates assume that, once open, the use of data is entirely unproblematic, making data meaningful in fact requires turning raw information into "intelligence": conclusions that can inform actions or serve as the basis for evaluations. Data intelligence requires bringing many complementary structures to bear on the data itself, the absence of which can lead not to data equality but to "empowering the empowered" (Gurstein 2011). Gurstein posits a seven-layer model for promoting effective use of open data that identifies many of the most important complementary structures:

1. Sufficient internet access that data can be accessed by all users.
2. Computers and software that can read and analyze the data.
3. Computer skills sufficient to use them to read and analyze data.
4. Content and formatting that allows use at a variety of levels of computer skill and linguistic ability.
5. Interpretation and sense-making skills, including both data analysis knowledge and local knowledge that adds value and relevance.
6. Advocacy in order to translate knowledge into concrete benefits.
7. Governance that establishes a regime for the other characteristics.

In the absence of these conditions it is not likely that any open data will promote justice. Britz et al. (2012) argue that these conditions are required by Amartya Sen's capabilities approach to justice; in the absence of these conditions, diverse individuals are not able to use information to act on or become something that they value.

The *Bhoomi* program described in the previous section illustrates the problems that can arise in the absence of these conditions. Raman (2012) describes real estate developers as the main beneficiaries of the *Bhoomi* program. They are better positioned to gain access to and use the digital RTC records both because they have greater computational capabilities and interpretative skills in relation to the political and legal practices governing land tenure under the program. At the same time, they also have greater social and political power with which they can assert their interpretation of the data, increasing the probability that it will be the accepted interpretation. Open data under conditions of unequal capabilities—what Raman refers to as the "capture of information"—led to frequent mass evictions of residents of slums from "productive" (i.e., desirable to developers) parts of cities where previously their ability to present conflicting claims could at least stall such processes (Raman 2012).

This problem is likely to be exacerbated by the emergence of "big" data. While the term has come to mean virtually all things to all people, four key threads emerge. The first is size: big data is often the result of device use or transactions, and so is much larger than an ordinary dataset. A common way of expressing the size is to say that "Your data might not fit easily on an Excel spreadsheet. Big Data doesn't fit on your laptop" (Charles 2013). Big data is frequently measured in petabytes, more than one million times larger than the gigabytes that measure memory in a desktop computer. But the role of size in big data is controversial; to a very important extent "big" data is Yodian: size matters not. Big data is as much about integrating multiple data sources, sources that lack common structure and in many cases lack structure at all (Craig and Ludloff 2011). The combination of size, multiple sources, and unstructured data then presents the problem of having sufficient computing power to process the data as well as the methodological skills needed to extract useful information from the data, advantages that played important roles in the re-election of Barack Obama in the 2012 U.S. presidential election campaign (Scherer 2012). Often these methods are rooted in artificial intelligence and machine learning, and the resulting output of big data analysis is more often not simply descriptive or even explanatory but in fact predictive (Baepler and Murdoch 2010).

The emergence of big data is driven largely by dramatic reductions in the cost of computing power and storage, which have made it possible for data administrators to produce data characterized by all three key values in data administration: velocity, volume, and variance.

> The advent of clouds, platforms like Hadoop, and the inexorable march of Moore's Law means that now, analyzing data is trivially inexpensive. And when things become so cheap that they're practically free, big changes happen — just look at the advent of steam power, or the copying of digital music, or the rise of home printing. Abundance replaces scarcity, and we invent new business models. (Croll 2012)

The temptation is thus to think that the intersection of big and open data, and especially of those with open-source software capable of managing and analyzing it such as Linux, MySQL, R, QGIS, and Hadoop, should minimize the capabilities differences that plagued the *Bhoomi* program.

But these tools also have capabilities requirements that often go far beyond those of ordinary citizens. Hadoop supports distributed computing and the management of unstructured data, but setting up and maintaining a Hadoop system is by no means an ordinary user skill. R and QGIS are free, but developing the skills needed to conduct advanced statistical or GIS analysis takes time and money. Petabytes of storage and teraFLOPS of processing power are "trivially inexpensive" to a large organization but not something readily available to the non-professional. In January 2014, the largest external hard drive available on Amazon.com was a mere 32 terabytes and cost $4,461.This likely explains why open data projects remain dominated by state and business users: Enterprises have the capacity to take advantage of big, open data, a capacity that citizens lack. A data store developed in Manchester, England, pooled content from ten local authorities but resulted in little citizen use beyond proofs of concept such as a bus timetable. Uses have emerged where compelling business cases can be made, and the state itself—police in particular—has proved to be an important user of open government data (Archer 2012).

The result is that big data is not, in practice, open to citizens. Opening data may allow citizens to analyze individual datasets, producing useful descriptive statistics. The empowering potential of even this should not be dismissed. But "citizen-open" pales in comparison to what might be called "enterprise-open" data. Enterprises will have the resources to get the most out of open data as they will be able to apply the full range of big data capabilities to it. They will be able to join multiple datasets together even where the data lacks structure using non-relational databases. They will be able to use proprietary business intelligence software to develop predictive models based on the data, and employ staff with the skills to both build such models and use their results. Such data is open in the sense that there are minimal restrictions on access. Insofar as it can be managed and analyzed using tools that are, to an enterprise, cheap, simple, and widely available, it is fully open to enterprises. But to the extent that such data requires capabilities that are beyond those of ordinary citizens, it cannot be understood as open to them.

2.1.3 The Normalizing Database

Injustice can emerge in systems of data as much as in any particular parts of such systems. Many of the systems of data collection to which open data advocates seek access can be usefully understood as disciplinary in nature (Adams 2013). As developed by Foucault (1995), disciplinary systems exist when individuals, regulated by a combination of detailed control and constant surveillance, self-discipline their behavior to reflect "normalizing judgment": an evaluation not of obedience to a command but of conformity to a standard of normalcy. This normativity can both impose itself on those who might wish to deviate from it and marginalize those who actually do so. Thus, to the extent that disciplinary systems take advantage of open data to impose unjust normalizing judgments or impose normalizing judgments unjustly, open data presents the possibility of undermining social justice.

This is astonishingly common in educational data, and usually deliberately and explicitly so. The U.S. Department of Education's Gainful Employment regulations required institutions to both disclose to potential students and report to the federal government information about program completion, employment of graduates, and student loan repayment. The regulations were a response to concerns about whether for-profit educational institutions were taking advantage of student aid programs to support programs that would not lead to "gainful employment" and thus expose students to excessive debt burdens and waste taxpayers' money. Preliminary data indicated that approximately 5% of programs covered under the regulations would not have met any of the benchmarks for employment and debt, jeopardizing their eligibility to offer aid. A Department of Education spokesperson stated that the regulations had led institutions "to think about what they were doing" and cut underperforming programs, a conclusion echoed by a spokesperson for Corinthian Colleges, a parent company for several for-profit colleges. The Gainful Employment regulations are a classic disciplinary program: hierarchical observation in the form of reporting requirements that are examined by an authority leads actors to adhere to an imposed norm on their own without direct coercion from the authority.

The Integrated Postsecondary Educational Data System (IPEDS) is the major postsecondary education data reporting process used in the United States. IPEDS requires educational institutions that offer Title IV financial aid to provide an extensive list of information about the institution to the National Center for Education Statistics (NCES), which then makes the data available publicly via the internet. Institutions that fail to comply risk losing their eligibility to award federal financial aid. While most of the data submitted is either demographic or input-driven (e.g., number of students enrolled or amount of state funding received), nearly all output measures IPEDS requires institutions to report concern retention and graduation. Institutions must report a first-year retention rate and graduation rates within specified percentages of normal program time. IPEDS does not require any measures of student performance, such as grade point averages, standardized test scores for postgraduate admissions, or licensing exam statistics.

These items establish the norm to which judgment is oriented: universities exist not in order to increase students' intellectual capabilities but in order to award degrees within the amount of time a normal person takes to get through the program. It must be stressed as well that "normal" most certainly does not mean "average." In practice, no disciplinary system can provide the kind of universal surveillance that Foucault describes, in which the universal possibility of observation is sufficient to ensure the self-discipline of the systems' objects. IPEDS limits the scope of surveillance by directing institutions to report graduation and retention rates on a specific subset of students, those who had first enrolled at the institution with no previous postsecondary education during a fall term intending to pursue the highest undergraduate degree offered by the institution on a full-time basis. This, too, is thus part of the norm: The "normal" student that postsecondary institutions exist to serve is the classic college student, going off to college immediately following high school graduation, studying full-time with minimal outside commitments.

IPEDS normalizes the 4-year residential university. Colleges and universities self-discipline themselves to conform to this normalizing judgment.

Educational institutions, in turn, are relying on big data techniques to create disciplinary systems that control their students. Austin Peay State University has developed an electronic advising system that suggests courses based on students' degree requirements, the extent to which courses can meet requirements for several degrees should students change their majors, and the likelihood of success in the course. Students must work through the system at registration, though they may disregard the recommendations after reviewing them. The system is a response to the problem of maintaining student aid and graduation rates:

> [Austin Peay Provost Tristan] Denley points to a spate of recent books by behavioral economists, all with a common theme: When presented with many options and little information, people find it difficult to make wise choices. The same goes for college students trying to construct a schedule, he says. They know they must take a social-science class, but they don't know the implications of taking political science versus psychology versus economics. They choose on the basis of course descriptions or to avoid having to wake up for an 8 a.m. class on Monday. Every year, students in Tennessee lose their state scholarships because they fall a hair short of the GPA cutoff, Mr. Denley says, a financial swing that 'massively changes their likelihood of graduating. … When students do indeed take the courses that are recommended to them, they actually do substantially better,' he says. (Parry 2012)

Certainly the institutional worldview that understands student success as simply completing a degree and its interest in maintaining financial aid should be apparent here. But this system, like similar systems at Arizona State University and Rio Salado College, goes a step further, using hierarchical observation and examination to promote student self-compliance with "wise choices" as the institution understands them. Here the tools of analysis and the construction of the data combine to create a data system that, open or closed, is about the institution imposing its values on students who may not share them; the data collected and analyzed is data that is relevant to a particular vision of education (credentialing) and of student success (completion). Opening the data (for instance, by allowing students to understand how the recommendations are made) does not change that in the slightest.

Hence the opening of data can function as a tool of disciplinary power. Open data enhances the capacity of disciplinary systems to observe and evaluate institutions' and individuals' conformity to norms that become the core values and assumptions of the institutional system whether or not they reflect the circumstances of those institutions and individuals. Both individuals who deviate from these norms and the institutions that specialize in serving them are marginalized in policy debates; the surveillers evaluate all institutions according to the norm (and indeed data may only exist regarding it), and the institutions internalize the norms and orient their actions to them. With the norms reflecting the power structure of the society in which they developed, they reiterate the patterns of justice and injustice that open data set out to ameliorate.

2.2 Big Data in Higher Education

Data mining and predictive analytics are increasingly used in higher education to classify students and predict student behavior. Institutions of higher education, in some cases working with commercial providers, have begun to use these methods to recommend courses, monitor student progress, individualize curriculum, and even build personal networks among students. Institutional researcher E. Rob Stirton argues that data mining, as a major part of business intelligence, is part of a radically different future for higher education in general and institutional research in particular:

> Preparing predictive models through data mining changes the focus from trends and past performance to future-oriented projections, thereby allowing planning strategies to be based on leading indicators and scenarios, which further leverage our investment in people and computers. The story changes from describing what happened to foretelling what will likely occur. Providing statistically significant predictive analytics would alter every institution's approach to Strategic Enrollment Management. (Stirton 2012)

But while the potential benefits of such techniques are significant, realizing them presents a range of ethical and social challenges. Those who implement these techniques in higher education will thus be called on to not only build the technical processes but also to protect students, institutions, and society from their side effects.

One might consider two kinds of challenges that data mining poses for institutional researchers. The immediate challenge considers the extent to which data mining's outcomes are themselves ethical. Individually, those subject to data mining—primarily but by no means exclusively students—must be respected as human beings when data mining is used to understand their characteristics and guide their actions. This means protecting both their privacy and their individuality. Institutionally, data mining may undermine the purposes of higher education in a democratic society or the missions of individual institutions. A deeper challenge, one not readily apparent to institutional researchers or administrators, considers the implications of uncritical understanding of the scientific basis of data mining. Excessively scientistic views neglect the problems of acting on conclusions that are erroneously perceived to be scientifically justified and of the meanings, assumptions, and values that are embedded in data mining applications.

2.2.1 Data Mining and Predictive Analytics

Data mining and predictive analytics[3] encompass practices and methods that vary greatly in familiarity to those with quantitative backgrounds typical of researchers in education and the social sciences. Some techniques, such as various regression

[3] For the purpose of this chapter, I will use the terms *data mining* to refer to the general task of identifying relationships in large datasets without *a priori* theoretical bases and *predictive analyt-*

methods, are familiar but used in different ways. Other techniques, such as k-means clustering and decision tree algorithms, have been used extensively in business—the oft repeated examples of Netflix, Amazon, and Target are now clichés in data mining—but are only recently coming to the attention of institutional researchers and educational administrators.

Data mining presents different challenges to its users than do inferential research methods—often called "academic analytics" (Baepler and Murdoch 2010). At the outset of research, academic analytics, like inferential approaches to both social scientific research and business analytics, begin from a model developed *a priori* by the researcher. The purpose of data analysis is to test the hypothesized relationships predicted by the model. Data mining, however, eschews the hypothetico-deductive process, relying instead on a strictly inductive process in which the model is developed *a posteriori* from the data itself. The model does not need to be tested against the dataset from which it is derived, as the algorithm ensures an accurate fit to that data (Baepler and Murdoch 2010; Two Crows Corporation 2005).[4]

Operating without theory requires much different mathematical techniques than academic analytics. The inferential statistics used in academic analytics work from mathematical theory and include in most cases quite specific assumptions about the underlying data (e.g., that it is normally distributed or homoskedastic); techniques are designed to minimalize computational requirements and rely on detailed specification of model form at the outset. Predictive analytics relies heavily on machine learning and artificial intelligence approaches. These take advantage of vastly increased computing power to use brute-force methods to evaluate possible solutions. Detailed model specifications are not necessary at the outset, as the process is said to "learn" the best model form over multiple iterations of the algorithm (Two Crows Corporation 2005).

The results of the two approaches are also significantly different. Academic analytics produces models whose main goal is to characterize the general tendencies in a dataset. This is most clearly the case for descriptive statistics, but measures of association and hypothesis testing statistics also have the same goal of explaining, in a single value, the general relationship between variables or the degree to which distributions would be expected by typical random variation. Even regression models, which do in principle yield predicted values for individual cases, are most typi-

ics to refer to the mathematical and computational techniques used in the practice of data mining. Readers are advised, however, that this distinction is introduced in the paper for clarity and is not based on more broadly accepted convention in the field; the two terms are often used interchangeably in the broader literature.

[4] It is, of course, advisable that the model be tested against a new dataset, often a portion of the original dataset reserved for that purpose. With some predictive analytic techniques this is necessary, as it is possible for the model to over-fit the data. Neural nets, for example, will inevitably produce a model that exactly matches the dataset on which the net is trained if allowed sufficient iterations and hidden layers, but once the model begins to incorporate stochastic variation, it will show increasing error when applied to data on which the model was not trained (Two Crows Corporation 2005).

cally used to evaluate general relationships: β is interpreted as the effect on the dependent variable of a one standard-deviation change in an independent variable, r^2 is the proportion of variance explained, and p is the likelihood that the general relationship is attributable to random variation (King 1986). Predictive analytics, on the other hand, is designed to characterize specific cases, generating a predicted value or classification of each case without regard to the utility of the model for understanding the underlying structure of the data. Many predictive analytic techniques, in fact, do not yield models capable of generalized interpretation at all (Two Crows Corporation 2005).

The result of these three procedural differences is the key practical difference between academic analytics and data mining. Under the right circumstances and with appropriate limitations, the results of an inferential test are intended to be interpreted causally. Inferential research in retention can thus be said to aim at explaining why retention occurs, and relationships between variables that cannot be understood causally—ones displaying multicollinearity or that are likely to be spurious, for instance—are of no use (Pollack 2012). Data mining, however, aims strictly at identifying data relationships. Models such as Neural Nets or Classification and Regression Trees (CART) are difficult or impossible to interpret generally; the lack of theoretical guidance in the machine learning process makes even interpretable models such as Decision Trees or Multivariate Adaptive Regression Splines (MARS) as likely to include spurious as causal variables, especially when such variables display significant multicollinearity. Such variables are valuable in data mining because they may be more effective indicators of the response variable than an ultimately causal variable that is obscured by interactions.

This is the key—perhaps the sole—reason that a strictly inductive, non-hypothesis driven approach is of value: Data mining works for the quite different purposes for which it was designed, purposes which do not include ascribing causality (Baepler and Murdoch 2010). The aim of data mining is to identify relationships among variables that may not be immediately apparent using hypothesis-driven methods. Having identified those relationships it is possible to take action based on the fact that the relationships predict a given outcome. For example, retailer Target is able to identify pregnant customers based on changes in their habitual purchasing patterns. Target mined purchasing data from customers who had signed on to the company's baby registry and was able not only to identify pregnant customers who were not in the registry, but were able to determine their approximate due date. Using this data, Target was able to tailor advertising to those customers, with the aim of changing their overall shopping habits, an opportunity that coincides with major life changes (Duhigg 2012). Clearly nothing identified by Target's efforts to data mine purchases was causal. But for Target's purposes, cause was not relevant; the company simply sought cues that would predict when a customer would be inclined to purchase particular items. Data mining is the ideal tool for such situations.

2.2.2 Higher Education Applications of Data Mining

The use of data mining has attracted increasing attention in higher education over the past decade. Educational data mining aims at "making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in" (Baker 2010). As Delavari, Phon-Amnuaisuk, and Beizadeh argue:

> The hidden patterns, associations, and anomalies that are discovered by data mining techniques can help bridge this knowledge gap [between what those carrying out educational processes know and what they need to know] in higher learning institutions. The knowledge discovered by data mining techniques would enable the higher learning institutions in making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. (Delavari et al. 2008)

The growing interest in data mining is spurred, in part, by the increasing quantity of data available to institutional researchers from transactional databases, online operations, and data warehousing (Baepler and Murdoch 2010).

Initial research projects using data mining approaches studied several different types of outcomes such as student satisfaction (Thomas and Galambos 2004) and student assessment (Delavari et al. 2005). Based on this initial research, Delavari et al. (2008) suggested a wide range of potential applications, including predicting alumni contributions, predicting standardized test scores, creating learning outcome and institutional typologies, predicting outcomes and intervention success, predicting student performance and identifying at-risk students, and identifying appropriate academic programs for each student. Similarly, Baker (2010) suggests four areas of application: building student models to individualize instruction, mapping learning domains, evaluating the pedagogical support from learning management systems, and scientific discovery about learners. Kumar and Chadha (2011) suggest using data mining in organizing curriculum, predicting registration, predicting student performance, detecting cheating in online exams, and identifying abnormal or erroneous data. More recent applications have embraced such suggestions, exploring course recommendation systems (Vialardi et al. 2009), retention (Zhang et al. 2010), student performance (Ayesha et al. 2010; Baradwaj and Pal 2011), and assessment (Llorente and Morant 2011).

Unfortunately, these studies do not make for a promising foundation for the practice of educational data mining, because they suffer, on the whole, from major methodological flaws. None of the predictive efforts provide control data, for instance, commonly reporting a generic "accuracy rate" that is not even clearly described. For example, the course recommendation system designed by Vialardi and colleagues aimed to predict success in the recommended course and steer students away from courses in which they were likely to be unsuccessful. They reported a 73.9% accuracy rate with 80.2% of errors being false negatives. While this sounds impressive, the absence of any sort of proportional reduction in error measure of model accuracy prevents evaluation. If the pass rate for the course was 50%, the model would

be impressive indeed. But if the pass rate is 90%, the model is less accurate than simply predicting that all students would pass, and thus offers no improvement on existing methods. This problem is also present in the study by Zhang and colleagues. Llorente and Morant provide a crosstabulation of results with only column percentages and do not even provide the sizes of their treatment and control groups, making it impossible to determine the statistical significance of their findings. Barawaj and Pal provide no evidence at all that their decision tree in fact has any predictive value.

There is also an exceptionally casual attitude toward attributing causation. Delavari and colleagues, for example, report a strong relationship between instructors' performance and their marital status among those with weaker academic qualifications (2008). This relationship is almost certainly spurious, probably epiphenomenal with age and experience. Similarly, Thomas and Galambos hold, "studying a single student body begins to identify aspects of the college experience *that most affect* student satisfaction" (2004, p. 265, emphasis added), without any effort to describe a causal relationship between satisfaction and the variables identified by their CHAID method. Given that data mining was not designed to support causal inferences and provides no means for identifying potentially spurious relationships, such claims are not supportable. Both of these problems will prove problematic when considering the ethics of their use.

The cases of data mining reported in the scholarly literature above have been primarily pilot projects, limited to predictions for individual courses or academic programs. In spite of this and their methodological problems, however, data mining is gaining hold operationally at the institutional level. The most common applications are within courses. Rio Salado College has developed a system that predicts student success in online courses based on early performance in the course. The system provides information to instructors about predicted student performance so that instructors can intervene to promote success. The system's developer claims to be able to predict course success on the eighth day of class with 70% accuracy. Success with intervention, such as using welcome emails to encourage students to log in on the first day of class, has been mixed, however, according to descriptions in the media. A system in use at Arizona State University uses data mining to personalize content in online courses by adapting the course content to each student. The system, developed by educational software company Knewton, provides content in online and hybrid math courses based on student behavior and past performance, focusing students on the concepts they need help with, sequencing lessons based on individual needs, and presenting content in formats suited to their learning style. Data mining has filtered into traditional classrooms as well with systems such as Harvard's Learning Catalytics. The system matches students for in-class discussions based on answers to practice problems with the aim of stimulating discussion. Students with differing answers to the practice problem are matched together in real time to debate their answers (Parry 2011, 2012). A similar system is in place at the University of Texas (Deliso 2012).

The other major application of data mining has been in advising. Course recommendation systems are in place at several universities, including Arizona State University, the University of Florida, and Austin Peay State University. Austin Peay's

"robot adviser" is a response to the findings of behavioral economics that show the difficulty of making good choices when confronted with an overwhelming number of options and the often substantial consequences of marginal differences in performance. It uses recommendation algorithms similar to those used by Netflix to suggest courses based on major, degree requirements, student performance, and the past performance of similar students. Its grade predictions are accurate to one-half letter grade, administrators report, and they believe that students perform better when following the recommendations. ASU and Florida go a step further, monitoring student progress through their academic programs and sometimes intervening to force student action. ASU's eAdvising system requires students to choose a major and develops a plan for when to take courses. The plans front-load key courses so that students who aren't suited to the major are identified early. Students are marked "off-track" based on enrollment and performance, and may be forced to change majors after two such semesters. Austin Peay is implementing a similar system (Parry 2011, 2012).

Other applications of data mining are less common but may indicate where universities are taking data mining in the future. ASU mines campus identification card swipes at campus facilities to model campus social networks and student behavior, with an eye toward identifying lack of social integration or changes in behavior that suggest a student may withdraw. It can combine that data with other administrative data, for instance, requests for transcripts, to identify students to whom advisors should reach out (Parry 2012). ASU also mines Facebook data from students who have installed the university's Facebook app and recommends other students with similar interests (Deliso 2012). Admissions and recruiting are also growth areas for data mining. ConnectEDU, an online social network platform, operates as an eHarmony-like matching site for colleges. It matches students to colleges where they will fit well and allows colleges, indirectly, to contact students whose ConnectEDU profiles fit the institution's recruiting program (Parry 2011).

2.2.3 Consequentialism: The Immediate Challenge

Nearly from its inception, data mining has raised ethical concerns. Once implemented, a series of challenges for both the individuals who are the subjects of data mining and the institution that bases policy on it arise as consequences.[5] The most prominent of these are the related problems of privacy and individuality. The privacy of subjects in a data mining process is primarily a factor of information control: a subject's privacy has been violated to the extent that the opportunity for consent to

[5] In this section, I use "consequences" and related terms strictly in a non-technical sense, referring to moral conditions that arise consequent to the implementation of a data mining process. At this point, I take no position on the relative merits of formally consequentialist or deontological ethical theories in evaluating those circumstances, though it will become clear to readers familiar with the distinction through the examples that follow that I believe that both kinds of ethical theory at least raise questions that data miners should address.

collection or use of information is absent or in which personal information flows are used in ways that are incompatible with their social context (Nissenbaum 2010; van Wel and Royakkers 2004). The potential of data mining to violate personal privacy spans a range of applications. At its least intrusive, the data collection and storage capabilities that make data mining possible allow those with whom one interacts to develop a complete dossier about those interactions. This leaves one's privacy unprotected by the failures of human memory. Mining that data allows one to infer information about the data subject that some would not be comfortable divulging themselves (as in the Target example described above). At its worst, privacy violations allow for the manipulation of or discrimination against the subject, for example, by price discrimination and restrictive marketing (Danna and Gandy 2002).

These risks are very much present in higher education applications of data mining. Course recommendation or advising systems that consider student performance are a way of developing a comprehensive picture of student performance, in essence, an electronic reputation that the institution maintains and makes available to faculty and staff through dashboard and stoplight processes and administrative rules. Unlike a personal reputation among faculty in one's major, an electronic reputation seems more difficult to escape. It seems unreasonable to expect that Rio Salado College's students universally want a system to identify them as more likely to fail; even if the intent is to encourage faculty to reach out to those students, undoubtedly many students would feel stigmatized instead. Arizona State University's effort to identify students who intend to transfer is clearly not information that students would consistently want to divulge, as one ASU student reported (Parry 2012).

Privacy concerns can easily give way to challenges to individuality. To be sure, such challenges are not new; older techniques that describe central tendencies and typical relationships can easily be seen as contributing to a collectivization of subject, where all are treated identically based on the assumption that they are all "typical" students. Data mining can go far toward overcoming this because it recognizes and models diversity among subjects. Thomas and Galambos, for instance, used the CHAID decision tree method to find "a significant dimension of diversity among the undergraduates in a public research university … identifying different satisfaction predictors for different types of students" (2004, p. 259). To the extent that the model is reasonably comprehensive and causally supportable (necessarily by other means) and that the data mining technique does, in fact, aggregate characteristics to something that represents the whole person, this individualization is to be preferred over collectivization.

But while academic analytics tends to collectivize the students by treating them all identically to the central tendency case, data mining has a tendency to disaggregate the whole individual into nothing more than the sum of a specified set of characteristics. Data mining can create group profiles that become the persons represented:

> Profiling through web-data mining can, however, lead to de-individualisation, which can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits. … In non-distributive group profiles, personal data are framed in terms of probabilities, averages and so on. (van Wel and Royakkers 2004, p. 133)

These profiles treat the subject as simply a collection of attributes rather than a whole individual, and interfere with treating the subject as more than a final predictive value or category. Course recommendation systems are just such a case; students are encouraged to do what students like them have done before. Austin Peay's system does not consider student interests, while Arizona State's eAdvising system is built specifically to identify students whose "ambitions bear no relation to their skills" (Parry 2012). This suggests that the students, far from being understood as individuals, are simply bundles of skills that need to be matched to an outcome.

At its extreme, data mining can undermine individuals' autonomy. Broadly speaking, autonomy can be understood as the ability to critically reflect on and act so as to realize or modify one's preferences, particularly preferences among conceptions of the good. This raises the questions of whether coercion and paternalism are ever justified, questions that are often addressed on the basis of a principle of preventing harm to others, furthering ends that the objects of the paternalism values themselves, or addressing a limited capacity for autonomy on the part of the object (Dworkin 1995). An especially complicated form of interference is the creation of disciplinary systems, wherein the control of minutiae and constant surveillance lead subjects to choose the institutionally preferred action rather than their own preference, a system that generally disregards autonomy (Foucault 1995).

Data mining can easily be coercive, paternalist, or disciplinary. ASU's system of compelling students making insufficient academic progress to change their major is very much coercive. The sense of promoting "wise" choices in Austin Peay's course recommendation system is a classic example of paternalism. Classifying students and communicating the classification to the professor used at Rio Salado College is virtually identical to Foucault's example of the Nineteenth Century classroom (1995, pp. 146–149) and could be expected to have similar effects: encouraging conformity to a set of behaviors that the institution has conceived of as successful. One might justify these interferences with student autonomy as preventing waste of taxpayers' money (a harm to the taxpayer, arguably), as furthering the educational ends that students presumably have when they enroll, or as guidance for those who are still not fully mature or lacking information about the consequences of a decision. But it remains necessary to provide such a justification in each case, as violations of the principle of autonomy are generally justified only as exceptions to the broad aim of allowing each person the maximum autonomy consistent with all others also having such autonomy.

It is not only the individuals whose data is mined, however, whose moral status comes into question when institutions use data mining. Many of the applications of data mining discussed above present moral concerns regarding the institution not as an actor but as one affected by the action. These chiefly concern the relation of data mining to the purpose of higher education, especially in a liberal democratic society. There are, of course, many such purposes. Peters argues that education is a process that leads "to the development of an educated man in the full sense of a man whose knowledge and understanding is not confined to one form of thought or awareness" (2010, p. 14), a perspective that one might call critical education. Flathman (1996) goes further, arguing that education ought to enable the individual to make critically

informed choices among conceptions of the good life, which he sees as the essence of liberal education. University of Pennsylvania president Amy Gutmann argues that democratic education ought to prepare students to participate in the processes of public deliberation over policy that guide representative government; higher education, especially, has an important place as a refuge for unpopular ideas, promoting values for professions that are not promoted by market forces such as professional virtue, and promoting communities that share intellectual and educational values (1999, pp. 172–193).

At the same time, higher education also has more practical purposes. It smacks of elitism to deny that students should pursue higher education for vocational purposes. It is as naïve to disregard higher education's role in establishing and maintaining social classes as it is cynical to disregard its role in promoting class mobility. Moreover, discussion of the purpose of "higher education" in general ignores the fact that each university may have its own specific purposes as well, deriving from its history, community, and governance. The practical and unique purposes are as important to a university's moral circumstances as are general views of what higher education should be.

Data mining can both contribute to and undermine these purposes. Mining data to find courses and majors in which students will be successful, like Arizona State, Florida, and Austin Peay do, may contribute to the vocational goals that many students have when they enroll in higher education. Students who find fields in which they will be academically successful are, it stands to reason, more likely to be professionally successful as well. But at the same time, those may be courses and majors in which students are successful because they are not challenged; likewise, personalized curriculum may provide the easiest path to course completion but not the surest path to learning. Where they are not challenged academically, they may not ever be critically educated in Peters' meaning. Where they are not challenged by divergent ideas, they may not ever be liberally educated in Flathman's sense or able to deliberate rationally as Gutmann would have them. ASU's social data mining is especially problematic for both democratic education and the status ambitions of many students in that it will almost certainly tend to reinforce the class relationships that students have when they enroll, preventing them from deliberating with a view toward the perspectives of others and from forming networks with others that would aid their social mobility.

2.2.4 Scientism: The Deep Challenge

The consequential challenges of data mining are the most prominent ones, but they are not the only ones. In fact, the most difficult challenges may be ones of which institutional researchers are least aware. In the process of designing a data mining process, institutional researchers build both empirical and normative assumptions, meanings, and values into the data mining process. These choices are often obscured by a strong tendency toward scientism among data scientists. For philosophers of

science and technology, the term refers (almost always critically) either to the claim that the natural sciences present both epistemologically and substantively the only legitimate way of understanding reality or to instances of scientific claims being extended beyond the disciplinary bounds in which the claim can be supported (Peterson 2003).[6] Such perspectives introduce the temptation to uncritically accept claims that purport to have scientific backing. This was a recurring theme in Twentieth Century political philosophy, one reflected in Dewey's (1954) critique of expertise, Arendt's (1973) analysis of Hitler's racial theories, and Habermas' (1990) communicative ethics. Given the mathematical precision and rigor of data mining, the temptation to accept the results as scientifically established and thus an unequivocal representation of reality is strong.

Scientism has a long tradition in the social sciences, and especially in the study of education (Hyslop-Margison and Naseem 2007). Critics of scientism in education see a fetishization of the scientific method, which manifests itself in contemporary policies such as *No Child Left Behind* and mandates "scientific" evidence of effectiveness as an authoritative practice of politics (Baez 2009). The preponderance of such methods in education research—and especially in the kinds of studies produced by institutional research offices—point to the assumption that traditional scientific methods are the ideal approach to understanding contemporary higher education. Indeed, one need look no further than the AIR standards for designation of a presentation as a "scholarly" paper: "Scholarly papers must include research questions, methodologies, literature reviews, and findings" (Association for Institutional Research 2012). Surely one would not dismiss disciplines such as philosophy or literature as non-scholarly for not being organized as AIR suggests; that is not the appropriate organization for scholarly work in those disciplines. The AIR standards confuse "scholarly" with "empirical," a confusion rooted in the positivist dismissal of the non-observable as unknowable "metaphysics."

Scientism is a trap that, if not avoided, can do substantial harm to students. But unfortunately, current examples of data mining in higher education have embraced, rather than rejected, scientism. The lack of attention paid to the major methodological flaws described in the previous section is a good example of scientism at work. A non-scientistic perspective critically evaluates methods and evidence before taking action upon it. But the casual attitudes toward causality and the ignorance of even statistical uncertainty in the studies of data mining in higher education suggest that the authors have taken an uncritical attitude toward the underlying science of data mining. Assuming that the relationships uncovered by data mining are inherently causal and reasonably certain can lead to ineffective actions and actions that reinforce rather than interdict causal mechanisms. Similar problems can occur when uses of data mining are insufficiently appreciative of the uncertainty present in the models; especially among users who only see the predictions and are unfamiliar with the model itself, predictions of a marginal change in likelihood can easily be implemented as predestined certainty.

[6] See, for example, Haack's (1993) critique of Quine's naturalism for a technical treatment. A useful non-technical perspective on scientism can be found in Kitcher (2012).

Rio Salado College's lack of success with intervention is telling. The welcome emails assumed that the relationship between first-day login and course success was causal; encouraging students to log in on the first day would thus increase their likelihood of success. But if both course success and first-day login are caused by students' self-motivation, a single email is unlikely to affect course success even if it does result in a first-day login; a sustained effort rather than a one-time intervention is needed. While this intervention is unlikely to harm, at the least an opportunity has been missed to make an effective intervention. The same cannot be said of potential actions stemming from the findings about lecturer marital status by Delavari et al. (2008). If the relationship between lecturer marital status and student performance is epiphenomenal to that between lecturer experience and student performance but is nonetheless used in hiring practices,[7] the university will certainly have harmed the subject of the model.

The problem of scientism in data mining goes deeper than just poor methodology. Part of the scientist epistemology is the claim that science is objective, and thus it—and its products—is value-neutral. But one of the key recent findings in both the philosophy and the sociology of science is the value-ladenness of science and technology. This is more than just claims of biases in scientific inquiry that deviate from the norms of such inquiry; it is an inherent feature of science and technology that they embody and embed values as they are created within a complex web of technical and social interdependencies (Nissenbaum 2010, pp. 4–6). Contingent meanings are as important as evidence and function in their development, as scientists and technologists make choices among equally likely possibilities or equally useful practices. Design intent and assumptions about user behavior are especially significant sources of embedded values in technologies. As technologies are themselves embedded broader structures when implemented, the values embedded in the technologies become embedded in the social context in which the technologies are used. The iteration of the technology development cycle reinforces this relationship: social values are embedded in technologies, and technologies reinforce those values (Johnson 2006).

The connection between technological artifact and social purpose suggests that data mining applications in higher education are best understood as part of a problem-model-intervention nexus. In developing models data miners link their own meanings, values, and assumptions to similar ones taken from the problem and the intended intervention. Richard Clark points in this direction when he criticizes personalized learning for its assumption that students' performance is rooted in different learning styles (a pedagogical theory that has seen its support eroded by recent research) and for questionable interpretations of data points, such as what

[7] Delavari and colleagues do not identify the university in which their study is conducted or its location, thus whether there are legal constraints that would prevent such a policy is unknown. Even if there are such constraints, however, such constraints are external to the criticism being made here; the finding and the failure of the authors to address the question of its spuriousness suggest that such conclusions are likely in areas in which the law presents no such constraint to designing an intervention around a spurious relationship that would harm the subjects of the model.

differences in time spent on a topic or learning method indicate (Parry 2012). When used properly—that is, predictively rather than causally—these criticisms lose some of their effect; if students who spend more time with video than text in one lesson are more successful when presented with video in the next lesson, it does not matter whether the relationship is epiphenomenal to an underlying motivational effect, personal preference, or difficulty with the material. The students' past behavior is sufficient to predict the success of an intervention regardless of a causal relationship. Of course, susceptibility to scientism in this respect is also likely to make one susceptible to the previous respect as well; when (mis)interpreted causally, the embedded values and assumptions of interventions based on data mining can easily become self-fulfilling prophesies.

Even when properly used, the values embedded in a model nexus become part of the institutional context. Vialardi and colleagues note that predictive analytic models "are based on the idea that individuals with approximately the same profile generally select and/or prefer the same things" (2009, p. 191). This very behaviorist model of human nature is at the foundation of every data model. While it is generally reasonable, one should note that it directly contradicts the rational utility maximizer model of human nature used in microeconomics or the habitual perspective of behavioral economics, and has very different implications for interventions. This is especially problematic in that interventions often incentivize behavior, a prescription best suited for rational utility maximizers. Similar processes embed more specific values in specific models. Most models are developed with both a problem and an intervention in mind, as can be seen in Austin Peay Provost Tristan Denley's description of the university's course recommendation system:

> Denley points to a spate of recent books by behavioral economists, all with a common theme: When presented with many options and little information, people find it difficult to make wise choices. The same goes for college students trying to construct a schedule, he says. They know they must take a social-science class, but they don't know the implications of taking political science versus psychology versus economics. They choose on the basis of course descriptions or to avoid having to wake up for an 8 a.m. class on Monday. Every year, students in Tennessee lose their state scholarships because they fall a hair short of the GPA cutoff, Mr. Denley says, a financial swing that "massively changes their likelihood of graduating." (Parry 2012)

The wisdom of a student's choice and the difficulty of making such a choice under these circumstances is part of the model; what it is to predict is not just a choice that the student will like but one which will be, from the institution's perspective, wise. And the model is specific about what constitutes wisdom: conformity to a utility function that values high grades and rapid progress toward graduation.

The question that arises here, then, is threefold: are users aware of the assumptions, meanings, and values embedded in the data model; are they consistent throughout the problem-model-intervention nexus; and is the inclusion of them justifiable? This is a question that is specific to each application of data mining in higher education, because the question is not whether values should be included in the application *per se*. There will be values in any technology; ethical applications of data mining are not value-free. They can, however, be value-conscious, even

value-critical (Johnson 2006). They ask whether likelihood of success in a course is a good standard to use for recommending that a student take the course; perhaps a wise choice is one that gives opportunities to develop wisdom through struggle rather than to maintain the highest GPA possible. They ask whether a behavioral prediction regarding academic progress makes sense as the basis for a utilitarian intervention; perhaps habitual behavior needs more impetus for change than a changing utility function. Often, one hopes, the values included are entirely reasonable, and perhaps even necessary. But ethical data mining can't happen if the ethical and philosophical assumptions behind the models are not considered.

2.3 Conclusion

In Tom Lehrer's (1965) song "Wernher von Braun," the titular hypocritical/apolitical rocket scientist denies responsibility for his creations: "'Once the rockets go up / Who cares where they come down / That's not my department' / says Wernher von Braun." Data systems, similar to von Braun's rockets, are too often assumed to be value-neutral representations of fact that produce justice and social welfare as an inevitable by-product of efficiency and openness. Rarely are questions raised about how they affect the position of individuals and groups in society. But data systems both arbitrate among competing claims to material and moral goods and shape how much control one has over one's life. These are the two classic questions of philosophical justice, raising the question of information justice. Information presents questions of justice as data is created, as it is used, even by its mere existence in a data system. It presents immediate questions about the consequences of information and deeper questions about the ideology of information technology itself.

Data scientists cannot be content to say that the use of their systems is someone else's problem: where the rockets are meant to come down determines the design of the system. Understanding information as a social product requires that information scientists work with an eye toward the social, asking critical questions about the goals, assumptions, and values behind decisions that are too easily—but mistakenly—seen as merely technical. Information science requires an understanding of information justice, which requires an understanding of justice itself.

References

@DeanOlian [Judy Olian]. (2016, April 18). Technology is neutral – it's what you do with it that turns it into a public good. Condoleezza Rice, at ASU-GSV Summit. *Twitter*. https://twitter.com/DeanOlian/status/722251515629441024. Accessed 22 Apr 2016.

Adams, S. (2013). Post-panoptic surveillance through healthcare rating sites. *Information, Communication & Society, 16*(2), 215–235. https://doi.org/10.1080/1369118X.2012.701657.

Archer, P. (2012). *Report on using open data: Policy modeling, citizen empowerment, data journalism*. Brussels: W3C. http://www.w3.org/2012/06/pmod/report. Accessed 3 Mar 2013.

Arendt, H. (1973). *The origins of totalitarianism*. New York: Harcourt Brace Jovanovich.
Association for Institutional Research. (2012). 2013 forum presenter information. *Association for Institutional Research Annual Forum*. http://forum.airweb.org/2013/Pages/Information/Presenting.aspx. Accessed 1 Apr 2013.
Ayesha, S., Mustafa, T., Sattar, A., & Khan, M. (2010). Data mining model for higher education system. *European Journal of Scientific Research, 43*(1), 24–29.
Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2), 17.
Baez, B. (2009). *The politics of inquiry: Education research and the "culture of science"*. Albany: State University of New York Press.
Baker, R. S. J. D. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (3rd ed.). Oxford: Elsevier.
Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63–69.
Britz, J., Hoffmann, A., Ponelis, S., Zimmer, M., & Lor, P. (2012). On considering the application of Amartya Sen's capability approach to an information-based rights framework. *Information Development*. https://doi.org/10.1177/0266666912454025.
Charles, N. (2013, March 4). Big data madness and my football prediction model. *Wallpapering Fog*. http://www.wallpaperingfog.co.uk/2013/03/big-data-madness-and-my-football.html. Accessed 24 May 2017.
Cohen-Cole, E. (2011). Credit card redlining. *Review of Economics and Statistics, 93*(2), 700–713. https://doi.org/10.1162/REST_a_00052.
Craig, T., & Ludloff, M. E. (2011). *Privacy and big data*. Sebastopol: O'Reilly. http://proquest.safaribooksonline.com/9781449314842.
Croll, A. (2012, August 2). Big data is our generation's civil rights issue, and we don't know it. *O'Reilly Radar*. http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html. Accessed 12 Mar 2013.
Danna, A., & Gandy, O. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics, 40*(4), 373–386.
Delavari, N., Beizadeh, M. R., & Phon-Amnuaisuk, S. (2005). Application of enhanced analysis model for data mining processes in higher educational system. In *2005 6th international conference on information technology based higher education and training*, F4B–1–F4B–6. doi:https://doi.org/10.1109/ITHET.2005.1560303.
Delavari, N., Phon-Amnuaisuk, S., & Beizadeh, M. R. (2008). Data mining application in higher learning institutions. *Informatics in Education, 7*(1), 31–54.
Deliso, M. (2012). How big data is changing the college experience. *OnlineDegrees.org*. http://www.onlinedegrees.org/how-big-data-is-changing-the-college-experience/. Accessed 12 Sept 2012.
Dewey, J. (1954). *The public and its problems*. Athens: Swallow Press.
Donovan, K. (2012). *Seeing like a slum: Towards open, deliberative development*, *SSRN Scholarly Paper No. ID 2045556*. Rochester: Social Science Research Network. http://papers.ssrn.com/abstract=2045556. Accessed 5 Mar 2013.
Duhigg, C. (2012). How companies learn your secrets. *The New York Times Magazine*. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=2. Accessed 16 Feb 2012.
Dworkin, G. (1995). Autonomy. In R. E. Goodin & P. Pettit, A. (Eds.), *Companion to contemporary political philosophy* (pp. 359–365). Cambridge, MA: Blackwell.
Flathman, R. E. (1996). Liberal versus civic, republican, democratic, and other vocational educations: Liberalism and institutionalized education. *Political Theory, 24*(1), 4–32.
Foucault, M. (1995). *Discipline and punish: The birth of the prison* (2nd ed.). New York: Vintage Books.
Freedman, M. (2014, March 26). What is the relationship between technology and democracy? *Insights by Stanford Business*. https://www.gsb.stanford.edu/insights/what-relationship-between-technology-democracy. Accessed 22 Apr 2016.

Goldrick-Rab, S. (2013, March 20). What have we done to the talented poor? *The EduOptimists*. http://theeduoptimists.com/2013/03/what-have-we-done-to-the-talented-poor.html. Accessed 24 May 2017.

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, *16*(2). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3316/2764. Accessed 5 Mar 2013.

Gutmann, A. (1999). *Democratic education*. Princeton: Princeton University Press.

Haack, S. (1993). *Evidence and inquiry: Towards reconstruction in epistemology*. Oxford: Blackwell.

Habermas, J. (1990). *The philosophical discourse of modernity*. Cambridge, MA: MIT Press.

Hoxby, C., & Avery, C. (2012). *The missing "One-Offs": The hidden supply of high-achieving, low income students*. (No. w18586. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w18586.

Hyslop-Margison, E. J., & Naseem, M. A. (2007). *Scientism and education empirical research as neo-liberal ideology*. Dordrecht: Springer. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=337528. Accessed 1 Apr 2013.

Johnson, J. A. (2006). Technology and pragmatism: From value neutrality to value criticality. In *Western political science association annual meeting*. Albuquerque. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2154654.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science, 30*(3), 666–687.

Kitcher, P. (2012, May 3). The trouble with scientism why history and the humanities are also a form of knowledge. *The New Republic*. http://www.tnr.com/article/books-and-arts/magazine/103086/scientism-humanities-knowledge-theory-everything-arts-science. Accessed 1 Jan 2013.

Kranzberg, M. (1986). Technology and history: "Kranzberg's Laws". *Technology and Culture, 27*(3), 544. https://doi.org/10.2307/3105385.

Kumar, V., & Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. *International Journal of Advanced Computer Science and Applications, 2*(3), 80–84.

Lehrer, T. (1965). In W. Von Braun (Ed.), *On That was the week that was*. Reprise/Warner Bros Records.

Llorente, R., & Morant, M. (2011). Data mining in higher education. In K. Funatsu (Ed.), *New fundamental technologies in data mining* (pp. 201–220). New York: InTech. http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/data-mining-in-higher-education.

National Science Foundation. (2012). *The national science foundation open government Plan 2.0*. http://www.nsf.gov/pubs/2012/nsf12066/nsf12066.pdf. Accessed 12 Mar 2013.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford: Stanford Law Books.

Open Data Working Group. (2007). 8 Principles of Open Government. https://public.resource.org/8_principles.html. Accessed 7 July 2015.

Orszag, P. R. (2009, December 8). Open government directive. *Office of Management and Budget*. https://obamawhitehouse.archives.gov/open/documents/open-government-directive. Accessed 24 May 2017.

Parry, M. (2011, December 11). Colleges mine data to tailor students' experience. *The Chronicle of Higher Education*. https://chronicle.com/article/A-Moneyball-Approach-to/130062/.

Parry, M. (2012, July 18). College degrees, designed by the numbers. *The Chronicle of Higher Education*. https://chronicle.com/article/College-Degrees-Designed-by/132945/

Peters, R. S. (2010). What is an educational process? In R. S. Peters (Ed.), *The concept of education* (pp. 1–16). Oxford: Routledge.

Peterson, G. R. (2003). Demarcation and the scientistic fallacy. *Zygon, 38*(4), 751–761. https://doi.org/10.1111/j.1467-9744.2003.00536.x.

Pollack, P. H. I. (2012). *The essentials of political analysis* (4th ed.). Washington, DC: CQ Press.

Prewitt, K. (2010). The U.S. decennial census: Politics and political science. *Annual Review of Political Science, 13*(1), 237–254. https://doi.org/10.1146/annurev.polisci.031108.095600.

Raman, B. (2012). The rhetoric of transparency and its reality: Transparent territories, opaque power and empowerment. *The Journal of Community Informatics*, *8*(2). http://ci-journal.net/index.php/ciej/article/view/866/909. Accessed 5 Mar 2013.

Rich, S. (2012, July 20). Palo Alto, Calif., to launch open data initiative. *Government Technology*. http://www.govtech.com/policy-management/Palo-Alto-Calif-Open-Data-Initiative.html. Accessed 12 Mar 2013.

Scherer, M. (2012, November 7). Obama wins: How Chicago's data-driven campaign triumphed. *Time Swampland*. http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/print/. Accessed 13 Mar 2013.

Schönberger, V., Cukier, K. (2013, March 6). Big data excerpt: How Mike flowers revolutionized New York's building inspections. *Slate Magazine*. http://www.slate.com/articles/technology/future_tense/2013/03/big_data_excerpt_how_mike_flowers_revolutionized_new_york_s_building_inspections.single.html. Accessed 8 Mar 2013.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Slee, T. (2012, June 25). Seeing like a geek. *Crooked Timber*. http://crookedtimber.org/2012/06/25/seeing-like-a-geek/. Accessed 5 Mar 2013.

Stirton, E. R. (2012). The future of institutional research – business intelligence. *eAIR*. https://www.airweb.org/eAIR/specialfeatures/Pages/default.aspx. Accessed 10 Sept 2012.

Thomas, E., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education, 45*(3), 251–269.

Two Crows Corporation. (2005). *Introduction to data mining and knowledge discovery* (3rd ed.). Potomac: Two Crows Corporation. http://www.twocrows.com/intro-dm.pdf.

Vialardi, C., Bravo, J., Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational data mining 2009: 2nd international conference on educational data mining, proceedings* (pp. 190–199). Cordoba: International Working Group on Educational Data Mining. http://www.educationaldatamining.org/EDM2009/uploads/proceedings/vialardi.pdf.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology, 6*(2), 129–140. https://doi.org/10.1023/B:ETIN.0000047476.05912.3d.

Williams, M. (2010). Can we measure homelessness? A critical evaluation of "Capture-Recapture". *Methodological Innovations Online, 5*(2), 49.1–4959. https://doi.org/10.4256/mio.2010.0018.

Zenk, S. N., Schulz, A. J., Israel, B. A., James, S. A., Bao, S., & Wilson, M. L. (2005). Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in metropolitan Detroit. *American Journal of Public Health, 95*(4), 660–667. https://doi.org/10.2105/AJPH.2004.042150.

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use data mining to improve student retention in higher education: A case study. In J. Filippe & J. Cordiero (Eds.), *Proceedings of the 12th international conference on enterprise information systems* (Vol. 1, pp. 190–197). Funchal: SciTePress.