Chapter 3
The Construction of Data

Abstract In this chapter, I show that data is not an objective representation of reality but rather a constructed translation of observations into legible elements designed to support governance (be it by the state or by private actors). Both technical and social structures influence this translation; the technical aspects of database architecture are insufficient by themselves to define this translation regime. Such regimes can contain three characteristic translations: normalizing translations that separate the normal from the deviant, atomizing translations that separate complexity into individual elements, and unifying translations that group diverse characteristics into categories. At the same time, these data systems translate their subjects into "inforgs," representations that consist of bundled information rather than actually existing subjects. These acts of translation, I conclude, are significant exercises in political power.

Whether in business, government, or higher education, pressures toward "data-driven" or "evidence-based" decisions are ubiquitous, promising more insight, more efficiency, and better outcomes than was previously possible. Implicit in this view, however, is a scientifically realist view of data: Data can save us because it is an objective representation of observed reality that can thus transcend politics to bring organizations to the correct decision. But if data is a social construct requiring acts of choice and interpretation in its creation, then it becomes political, its power masked behind its false realism. The structures that shape these choices are thus central to understanding information justice.

This chapter establishes the translation regime as a mechanism by which the social construction of data takes place, and suggests that translation regimes should be viewed as political structures rather than technical ones. Data exists because organizations such as universities or states have a need to make the domains in which they act legible. Doing so, however, requires some process that narrows the many possible representations of a given state of the world to a single data state. This process is carried out within translation regimes: systems of technical rules and social practices that establish a one-to-one correspondence between a given state of the world and a data state. The technical structures of a relational database, such as tables, functions, business rules, and queries, translate states of the world into data states based on standards established by social structures such as cultures, states,

J.A. Johnson, *Toward Information Justice*, Public Administration and Information Technology 33, https://doi.org/10.1007/978-3-319-70894-2_3

and organizations. These regimes operate in a non-neutral fashion, carrying out a set of characteristic translations that favor certain groups over others. As such, information systems design is a political act, among other things shaping representation, asserting and protecting interests, and constructing normalized and deviant identities. Because these political acts are carried out through the technical structure of the translation regime, they appear as technical outcomes, making it more difficult to challenge them.

3.1 Data as Reality Made Legible

The ubiquity of data in contemporary society hides its peculiarity. Data is a very specific form of information, one in which the subject is broken down atomistically, measured precisely (in the sense of being measured to quite specific standards that may or may not involve a high level of quantitative precision), and represented consistently so that it can be compared to and aggregated with other cases. That this form of knowledge is more common in highly structured institutions and rose to ubiquity with the modern, bureaucratic state and the capitalist enterprise should surprise no one. Creating data should be regarded as a social process in which reality is made legible to the authorities of an institutional structure.

Scott (1998) argues that the driving force behind the creation of data is the need to make the subjects governed by an institution legible.

> Certain forms of knowledge and control require a narrowing of vision. The great advantage of such tunnel vision is that it brings into sharp focus certain limited aspects of an otherwise far more complex and unwieldy reality. This very simplification, in turn, makes the phenomenon at the center of the field of vision more legible and hence more susceptible to careful measurement and calculation. Combined with similar observations, an overall, aggregate, synoptic view of a selective reality is achieved, making possible a high degree of schematic knowledge, control, and manipulation. (Scott 1998, p. 11)

Legible knowledge transforms reality into standardized, aggregated, static facts that are capable of consistent documentation and limited to the matters in which there is official interest. Such facts emerge from a process in which common representations are created into which cases are classified and which can then be aggregated to create new facts on which the state will rely in making decisions (1998, pp. 80–81).

The importance of legibility for governance can be seen most clearly when Scott contrasts legible knowledge with local knowledge. The latter, with all of the specific practices, details, and dynamics of reality, is impossible to use for the kind of broad governance characteristic of the modern state; it lacks commonality with other localities and is not objective to outsiders. This obstructs governance in two ways, first by preventing synoptic understanding by authorities and then by denying the governing algorithms of the bureaucracy the standardized inputs they need to produce a standardized output. “A legible, bureaucratic formula which a new official can quickly grasp and administer from the documents in his office” (1998, p. 45) is a necessity for modern governance in both the state and the enterprise.

The need for legibility defines not only the form but also the substantive nature of data. It is common to regard data from a scientific realist perspective in which data is a technical artifact, a representation of information about some subject that is stored such that it can be related to other such representations. This is, for example, the approach used in the United Kingdom's Data Protection Act 1998. The act defines data as a qualified synonym for information: "Data is information which …" followed by a list of technical conditions relating to storage and processing; personal data is defined by the data's relation to an individual identifiable either in the data itself or in relation to other data, and sensitive personal data includes information about a specific list of personal characteristics (Information Commissioner's Office n.d., pp. 19–23).

This is a quite problematic view of data, however, as it suggests that the process of representing reality[1] is an automatic, even algorithmic process. Such a view is naïve, however; like virtually all technologies (Johnson 2006), data is a socio-technical construct in which human agency and social structure is central (Nissenbaum 2010, pp. 4–6) and the path from reality to data is contingent rather than determined (Seaver 2014). Rather than being an automatic process with a one-to-one relationship between reality and data, data states are underdetermined with a one-to-many relationship between reality and data: one state of the world can give rise to many possible data states, some of which are incommensurable with others. In order to make the world legible to human authorities and algorithmic bureaucracies, one data state must be chosen to represent a state of the world from among many possibilities. Reality constrains those possibilities but it does not, by itself, fully reduce the state of the world to a single data state.

Netflix provides an exceptionally valuable case, as it explicitly attempts to datize a cultural product and thus makes the socio-technical nature of the company's recommendation engine clear. Netflix's process is at its heart a form of structuralism, disassembling films into their smallest component parts and reassembling the "altgenres" (hyperspecific categories that Netflix users see organizing films that Netflix recommends, such as "visually appealing intellectual action-thrillers") that describe the common structures across films. Structuralism reveals the contingency of Netflix's purportedly objective recommendations:

> When you break an object down into its parts and put it back together again, you have not simply copied it—you've made something new. A movie's set of microtags, no matter how fine-grained, is not the same thing as the movie. It is, as Barthes writes, a "directed, *interested* simulacrum" of the movie, a re-creation made with particular goals in mind. If you had different goals—different ideas about what the significant parts of movies were, different imagined use-cases—you might decompose differently. There is more than one way to tear apart content. …Netflix's altgenres are in no way the final statement on the movies.

[1] For the purpose of this paper, I take "reality" to mean the physically existing world as interpreted by actors within it. Here I follow Charles Sanders Peirce in his seminal essay "The Fixation of Belief" in which there is an underlying reality that cannot itself be perceived but that can be asymptotically approached through repeated observation (Peirce 1992). This leaves open an interpretive space but does not deprive the concept of reality of meaning, allowing it to be bracketed as a distinct but related problem from that addressed in this paper.

> They are, rather, one statement among many—a cultural production in their own right, influenced by local assumptions about meaning, relevance, and taste. "Reverse engineering" seems a poor name for this creative practice, because it implies a singular right answer—a fact of the matter that merely needs to be retrieved from the insides of the movies. We might instead, more accurately, call this work "interpretation." (Seaver 2014)

Broadening Seaver's analysis to data generally, each possible data state can be regarded as a potential interpretation of the underlying state of the world to be datized. The contingency of the final forms of data requires some external source of stability in order for data to bring legibility to the world (Mitev 2005). What is needed is a process of translation from reality to data that constructs a single representation by serving as the external source of stability for representation. Such a process is inherently endogenous to the creation of the data as long as multiple interpretations are possible. In a realist view of data, all but one of these states must be regarded as errors or biases in the data, which can be corrected by validating the data against itself or the reality it purports to represent until a single data state that is fully consistent with reality remains. But the self-correcting process of scientific realism cannot do this; rules for interpretation are required in order to have data in the first place. All possible final data states will appear consistent with reality because they follow the rules of the specific interpretive process that leads to them. These processes have legitimized the data states resulting from them as the only acceptable representation of reality, all else—local knowledge in particular—being dismissed as anecdotal evidence.

Classifying individuals within a system of gender relations is a good paradigmatic case that demonstrates the operation of a constructivist understanding of data beyond the domain of cultural production. Simply within the binary gender system common in western cultures, people might be represented within a data system either by sex or by gender. These categories are not reducible to each other; the existence of transgendered and intersexed people is sufficient to make sex and gender incommensurable within such binary systems. Moreover, there is no inherent reason that a data system needs to be limited to a gender binary, even in predominantly Western contexts: Facebook recently introduction of more than 50 custom gender descriptions from which its members can choose (Facebook n.d.). The intellectual construct "gender" is thus insufficient to determine how data systems will represent a specific person; the reduction of gender realities to specific categories cannot be an objective, value-free, observational process. In spite of this, most data systems rely on the same binary coding frame, one in which gender is taken to have a one-to-one correspondence with biological sex. The representation of individuals' place in the system of gender relations is thus determined by neither reality nor by the technical requirements of the data system. It is a choice on the part of developers to reduce an exceptionally complex reality to a specific legible form.

3.2 The Translation Regime

In order for the process of selecting a data state from among the many possible ones to be, in fact, legible, the process must be a rule-governed one. Creating data from reality is not simply an interpretation but a translation (or, more precisely, a series of translations) in which substantive content embedded in a set of technical rules determines how reality will be represented in the data system. For a relational database,[2] those rules are largely, but not entirely, contained within the data system itself, expressed as technical specifications within the database. The construction of data in relational databases consists mainly in the design and selection of rules such that they implement the demands of the content sources and only secondarily, when the rules and content sources are insufficiently precise, in the direct interpretation of reality by those entering data into the data system. Collectively, one might refer to these structures as the translation regime for a data system.

3.2.1 The Technical Structure of the Translation Regime

The most basic technical element of the relational database translation regime is the structure of individual data tables. The fields selected for inclusion in an individual table do much more than selecting which aspects of reality will be stored (though they most certainly do that as well). Those fields break down that reality into component parts. This is, of course, only a selection of the parts of the reality, and recombining these parts creates only an interpretation of reality rather than an objective and complete representation of it. A simple case is found in the fields representing the name of a person who is represented in a table row. The STUDENT table uses *FIRST_NAME*, *MIDDLE_NAME*, and *LAST_NAME*. These fields cannot be recombined to generate the formal name of everyone the data purports to represent. Truncating *MIDDLE_NAME* to a middle initial translates a student who goes by "G. Gordon Liddy" into "George G. Liddy"; a student from a country where family names precede given names named "Mao Zedong" is translated into "Zedong Mao."

A more complex example is seen in the information kept on a student's academic program. This data is kept in a hierarchy of fields within STUDENT: *DEGREE*,

[2] In a more general theory of data, the choice of database type would itself be understood as part of the translation regime. Raman (2012) shows that the choice of a relational database rather than one based on Unstructured Information Management Architecture (UIMA) to maintain land claims in India prevented the storage of knowledge held in narrative form, as was common among *Dalits* in the region. Narrative knowledge would have to be translated into an atomic structure in order to be stored in a relational database; in this case, such knowledge was simply excluded in favor of that contained in state-produced documents that could be stored in a relational system. Since all of the data currently used by UVU is contained in relational databases, the influence of database type must be investigated in another context.

COLLEGE, *DEPARTMENT*, *MAJOR*, and *CONCENTRATION*. This hierarchy standardizes the grouping of students by program in ways that may or may not reflect the actual operation of the program. The Behavioral Science major includes concentrations such as Psychology and Family Studies with such significant overlap in coursework, administration, and faculty that distinguishing students by concentration introduces an institutional separation of students to the data that is absent in reality. The major's Anthropology concentration, however, has much less overlap with the other concentrations. There is, in addition, a separate Social Work major that has stronger connections with the Psychology and Family Studies concentrations than the Anthropology concentration. The data fields, however, translate these varied conditions into a single, hierarchical set of student groups.

Moreover, each field in a table includes a definition restricting the type of data that can be entered into the field. These definitions define at the least the type of characters that can be put into the field and the maximum length of the field content; often field definitions might also include number formats, specialized formats such as times, or more precise tests of valid data. As such, they define what form the resulting data must take and proscribe the use of other forms. For a data element that represents a well-defined condition this is straightforward. But for conditions with more variability it is not at all so. Field definitions may thus permit or prohibit the entry of data that is valid in relation to reality but not in relation to the field definition. *TAX_ID*, for example, is defined as a variable character text field (to preserve the leading zero in some Social Security Numbers) of up to 63 characters. A more strictly defined field (for example, a fixed-width text field limited to nine characters) would prevent the entry of Federal Taxpayer Identification Numbers, which some students may have instead of a Social Security Number. The more flexible field definition of *TAX_ID* thus supports the translation of a wider range of conditions.

Commonly, some fields within a table will be indexed. Indexing a field stores information about the content of the field separately from the table itself, allowing the field to be searched rapidly. Typically, a table would index fields on which records would be selected, and then other data in those records could be returned promptly. In a small table, the difference in response time and server load between an indexed and non-indexed field may be minimal, but in a very large system might be the difference between practical and impossible searches. Indexing thus creates privileged translations of data, in extreme cases making fields that are conceptually equivalent incommensurable where one is indexed and the other is not: The indexed field is, effectively, the only field that can be used to represent the data in practice, and thus the only translation available for use. In COURSE, descriptive course information fields such as *SUBJECT*, *COURSE_NUMBER*, and *SECTION* are not indexed, but *COURSE_REFERENCE_NUMBER* is. This makes it quite practical to refer to courses by reference number and to identify descriptive course information given a reference number, but somewhat more difficult to do starting with the descriptive information, especially in the absence of other limits on the data needed.

Beyond the structure of individual tables, one might also look to the structures of a database that validate data across tables. Validation tables function in ways similar to field definitions. A validation table contains a list of values that are acceptable for

use in a field, used commonly in fields that contain categorical data with a limited number of possible values. The validation table for *COUNTY* contains a list of all counties in Utah, along with three residual values for all other cases. This prevents the entry of invalid county names. In the process, however, the validation table also determines the conceptual framework for the field itself. In this case, *COUNTY* becomes a characteristic held only by people from within that state. This is even clearer in the example of gender. The validation table for *GENDER* includes only the values "Male," "Female," and "Unspecified," imposing a binary gender schema on the people represented in the field. The "Unspecified" value as a residual is an especially strong reinforcement of the gender binary in this common validation frame: if one is not either male or female (whether because the translation regime insists on correspondence between sex and gender thus denying the existence of transgender identities, or because the person identifies as some form of non-binary identity), one is not even a residual "Other." One is presumed to, in reality, identify with one of the binary values and simply did not communicate that identification to those collecting data. These examples make clear the special importance of residual representation, often an afterthought, in validation tables' role in the translation regime.

A more complex validation structure is a business rule. Business rules place conditional requirements or constraints on the data in one or more fields based on the content of other fields, within or across tables. A common use might be to either require or proscribe certain external actions, for example, preventing a contract from being issued before a credit check has been performed by requiring that a row exists for a customer's credit check in a table of credit check data before a row can be created for that customer's contract in a table of contracts. Business rules can also be used to validate data across fields, preventing the entry of a state other than Utah in *STATE_ADMIT* (the state in which the student resided at admission) and a Utah county in *COUNTY_ADMIT*. In much the same way as validation tables, business rules impose a conceptual framework on the fields that they govern by limiting the data that can be entered to data that is consistent with the underlying concept. The central concept underlying a hierarchy of state and county of admission is the authority of a unitary state over its citizens at the local level. A business rule upholding that hierarchy would thus reinforce the structure of authority within state government in the United States. UVU's lack of such a rule has led to an exceptional amount of inconsistent data and thus inhibits the translation of a geographic location such as a street address into a political one such as a legislative district.

The relationships among data in different tables further shape the translation regime. In a relational database, data tables are structured so that tables can be joined to each other on common elements to allow cases in one table to be matched to cases in another. In the absence of appropriate commons field on which to join, however, data in different tables cannot be related to each other. The UVU reporting tables are designed expressly to facilitate this: COURSE and STUDENT_COURSE can be joined on the combination of *COURSE_REFERENCE_NUMBER* and *TERM*; STUDENT and STUDENT_COURSE can be joined on *STUDENT_ID* and *TERM*. Joining STUDENT to COURSE requires joining all three tables. As a result,

the translation of a particular characteristic of reality into an individual data field is also a translation of it into a context created by an extensive set of other data fields. A student is not simply a Botany major; joining STUDENT and STUDENT_COURSE makes the student a female Botany major who has not taken a course in the major in three semesters. This translation is much more interesting to those responsible for increasing retention of women in STEM degree programs.

All of the structures discussed above involve primary translation: the translation of a state of the world into data. But translation regimes include as well secondary translation processes, translating not reality into data but rather existing data into new data. Functions are a common structure that performs secondary translation. Fields can be defined with functions. Functions calculate a value for a field based on the content of other fields rather than being populated through direct entry of data. The function that calculates STUDENT.*INSTITUTIONAL_GPA* combines *CREDITS_ATTEMPTED* and *CREDITS_EARNED* across all rows in STUDENT_COURSE for a student to create a representation of that student's academic performance that does not exist in the absence of the function: an aggregate performance indicator. Functions thus widen dramatically the range of data contained in the data system and produced by the translation regime, illustrating the extent to which translation is not solely about creating equivalent representations of existing data but also about creating new data through the combination of existing data.

The data stored in a database is not necessarily the data that will be used in the final representation of reality. Data from relational databases is extracted through queries that specify precisely what data will be extracted, how it will be combined in new fields, and how it will be aggregated. A query will, at the least, specify which fields to retrieve for a record, and will usually specify which records to retrieve as well. The query thus selects, for example, whether students' academic performance will be represented by *INSTITUTIONAL_GPA* or *OVERALL_GPA*. But queries can also use the same set of functions that are used to define fields. A particularly common translation used at UVU when extracting demographic data for survey samples[3] is the creation of a binary ethnicity field. Given that minorities make up less than 20% of the UVU student body (Institutional Research & Information 2012a, p. 18), surveys are rarely large enough to provide reliable data when broken down by individual ethnic categories. Survey sample queries thus categorize student ethnicity using a function that parses *PRIMARY_ETHNICITY* and *RACE_COUNT* to identify students as either White or minority. This function is written into the standard queries that are used to generate samples, and often included in ad hoc queries for particular projects.

Queries are the final point in a relational database where translations take place. That does not mean, however, that the technical structures of the translation regime are limited to those processes that take place between data entry and data extraction. Applications, whether software systems or analytical processes that connect to a

[3] UVU commonly refers to the group invited to participate in the survey as a "sample" even when the invited group is in fact a census of a sub-population of students such as graduates in a term. I use "sample" in this sense here as well.

data system, can further translate the data extracted. UVU's "Stoplight" risk warning system translates 20 possible trigger conditions that the institution identified as characteristic of students at risk of failing courses into a color-coded risk rating that is shown to advisors and on class rosters (UVU Student Retention 2013). Stoplight operates as a new application built within the ODS, with a custom table carrying out this secondary translation and feeding data from it to advisors and instructors. UVU also maintains a website presenting data on mission fulfillment that is built, in part, on data that is extracted from the data system then aggregated and represented graphically using business intelligence software, translating individual data points into aggregated visual data. These applications are the point at which data finally meets a human who must act on the data, and thus mark the boundary of the technical structure of the translation regime.

3.2.2 The Social Sources of the Translation Regime

While the substantive content of the translations is inscribed in, and to an extent constrained by, the technical structure of the database, the bulk of the substance comes from sources external to the database itself. Culture, the state, the institution, and private sector actors all provide content for the translations that is then built into the technical structures.

As much as the language of conscious design and engineering permeates both the theory and practice of information systems, their conformity with their origin communities' cultural structures suggests that sociological institutions are at least as important. Like many organizational forms, data systems include "not just formal rules, procedures, or norms, but the symbol systems, cognitive scripts, and moral templates that provide the 'frames of meaning' guiding human action" as a mutually-constituting element of social action (Hall and Taylor 1996): Data systems are both composed of and instantiate cultural institutions. Some of these are relatively straightforward, such as *GENDER* including only "Male" and "Female" as valid values. A binary frame of meaning shared by most people in the institution defines which values are built into the validation table, for the most part without a conscious decision to do so.

A more interesting example is *STUDENT_CLASSIFICATION*. This field, in a technical sense, divides undergraduates into four classes based on the number of credits completed. Ostensibly, this indicates progress toward degree. But the values in *STUDENT_CLASSIFICATION* are more than categorizations or translations of a number. "Freshman," "Sophomore," "Junior," and "Senior" is a cultural script for understanding the social relations of a traditional residential institution. At an institution where many students are part-time, married, or returning adults and there is no campus housing, neither a 4-year academic career nor a distinct cohort are relevant concepts to most students.

STUDENT_CLASSIFICATION thus operates not as a reflection of student behavior or program structure but rather as a script for (mis)understanding and relating

students to each other and the institution. "Freshman" is a frame of meaning that assigns attributes to a student; the First Year Experience program acts toward such students as that frame says is appropriate, stressing that "College is different from high school," that independence is exciting but can be overwhelming, and that participation is a good way to make friends: messages appropriate to a traditional freshman on a residential campus but not to a recently retired Marine pursuing her bachelor's degree as a start to a second career. Similarly, the National Survey of Student Engagement samples all first-year students on the basis of institutional classification and asks a series of questions about differences between high school and college engagement experiences even though many freshmen at UVU are closer their children's graduation from high school than their own.

Political influences operate in a much more clearly conscious fashion, usually being deliberately designed into the data structures. The state shapes translations primarily by establishing formal data standards. Data standards define substantively and sometimes technically the content of a data field or record. UVU's main data standards, as is true for most public higher education institutions, are found in two sources: the federal IPEDS Glossary (National Center for Education Statistics n.d.) and a series of data dictionary files from USHE (Utah System of Higher Education 2013a, b, c, d). The USHE standard for *CITIZENSHIP_STATUS* in the student table, for example, translates the many categories of rights to presence in the United States under US law to five categories: US citizen, US national, resident alien (which includes all documented immigrants entitled to stay indefinitely in the United States), non-resident alien, and "non-immigrant undocumented students" (Utah System of Higher Education 2013c). This last category is particularly interesting, as it marks a quite significant departure from the typical discourse of undocumented *immigrants*, translating a person's intentions as well as their position within the immigration regime. The USHE data standard for *GENDER* became a stricter one with the inactivation of the "Unspecified" value in the USHE standards in 2012 (Utah System of Higher Education 2013c, p. S-13). This prohibited missing data in *GENDER*. As a result, gender nonconformity is no longer even translated as missing data; all students are translated into one of the binary gender categories. With such cases being typical examples of how data standards translate reality, it is clear that they should be viewed as substantive translations, not simply as technical coding procedures.

The translations created by data standards can be quite complex, especially when multiple data standards can apply to the same set of data. STUDENT supports three distinct data standards for ethnicity data to support competing, and in some cases conflicting, data standards. The USHE standard for *ETHNICITY* defines an eight-character field in which each character position represents an ethnicity with which a student might identify, with multiple identifications allowed, chosen among Hispanic or Latino, Asian, Black or African-American, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, White, Non Resident [sic] Alien, or Unspecified (Utah System of Higher Education 2013c, p. S-14). IPEDS currently used the Office of Management and Budget standards, in which students select all groups with which they identify among American Indians or Alaska Natives, Asians,

Blacks or African-American, Native Hawaiians or Other Pacific Islanders, or Whites and then identify whether or not they are Hispanic or Latino (National Center for Education Statistics n.d., p. R). UVU also supports older IPEDS standards that define students by a single ethnicity.

To do this, STUDENT includes one binary field for each possible ethnicity that it might report, a count of the total ethnicities selected by the student, and a primary ethnicity to be used with standards that do not support multiple ethnic identities. Reality having been translated into these data fields, a further translation of the data fields into the reporting identities takes place in querying and extracting data for reporting. This creates an extensive complex of translations that are not entirely consistent. The same student may be "White" in *PRIMARY_ETHNICITY,* multiracial in *IPEDS_ETHNICITY,* "Minority" in a query dividing students into "White" and "Minority," and "Non Resident Alien" in *USHE_ETHNICITY*. While inconsistent, none of these is fundamentally incorrect either as a translation of the ethnicity fields in STUDENT nor as translations of reality.

Political systems have more subtle means at their disposal to influence the translation regime as well. Especially for public institutions but, given the public mission of higher education generally, to some extent for all higher education institutions there exists a principal-agent relationship between the polity and those institutions similar to that between legislatures and bureaucracies. That relationship subjects the translation regime to many of the same oversight pressures as any regulatory regime. One of the most common responses to such pressure is bureaucratic anticipation: agencies, seeing signals from legislators about their desired outcomes, anticipate direction from the legislature and move to secure those outcomes without waiting for that direction to be made explicit (which, in many cases, never happens because the need for direction has been met) (Weingast and Moran 1983). This is not simply having the foresight to see a new formal requirement coming and implement it in advance; it is an act of anticipating the demands of political actors and meeting them as a means of satisfying those actors whether or not the demands are formalized.

Anticipation was a key factor in designing one of UVU's signature data applications, its Student Success and Retention dashboard (Institutional Research & Information 2012b). The dashboard was designed to assess efforts to improve the first-year retention rates and graduation rates reported to IPEDS. The appropriate federal data standard is thus the rates for the IPEDS cohorts: first-time, full-time, bachelor's degree-seeking students entering in the fall term. At the time this was being developed, however, the US Department of Education had begun public discussion of revised data standards to take effect in the 2014–2015 data collections, and constituencies and their legislators in both Utah and the federal government had raised significant concerns about whether higher education was meeting the needs of non-traditional students. Those involved in designing the dashboard recognized that significant political pressure was building to demand student success data for part-time and transfer students. The dashboard as completed in 2010, well before NCES made decisions on the new standards, was thus based on a fall new student cohort with both full-time and part-time students, and designed in a way that would facilitate the addition of transfer students by creating a transfer student cohort. UVU was

able to provide part-time data to the institutional administration, the community, and political actors well before it faced a formal requirement to do so. Neither standard was implemented by NCES until 2013, taking effect with the 2014–2015 IPEDS data collection, well after UVU had begun tracking the success of part-time students.

The private sector, both for-profit and non-profit, is an important source of content as well. Because UVU's data system is a customized version of a widely used commercial higher education data system, much of the translation regime's content comes from Ellucian, the makers of Banner. When UVU adopted Banner in 2005 and implemented the ODS in 2009, the institution started from a standard Banner database schema and then customized it to meet specific needs on the UVU campus (such as USHE reporting). This requires a notional higher education institution whose needs are representative of most institutions around which the out-of-the-box version of Banner can be designed; elements of the schema that were left unmodified thus reflect Ellucian's conception of what the content of fields should be based on that notional institution. UVU's class rosters, produced by an application within the Ellucian Luminis web services platform connecting to Banner data, provide students' formal names even with a preferred name field available. The University of California, Davis, has in fact implemented an option within Banner that allows students to use preferred names rather than formal names on many university documents including class rosters (Easley 2014), demonstrating that the standard form for class rosters in Banner is not a technical or legal constraint but an assumption on the part of the software designers.

The non-profit sector's contributions to the content domain of higher education translation regimes should not be discounted. Institutions, in a bid to increase transparency (or at least the appearance thereof), are frequently participants in voluntary data sharing processes, each of which comes with their own data standards that may or may not be coordinated with others. The Voluntary System of Accountability (VSA) is one of the largest among public institutions, providing both input and outcome information with the aim of demonstrating the value of an institution's programs. UVU also participates in the Consortium for Student Retention Data Exchange, a program that facilitates peer benchmarking of multi-year retention and graduation rates. In both cases, these organizations' data standards exist alongside government standards. The tables supporting UVU's retention and graduation rate application, described above, translate student enrollment data into both IPEDS and CSRDE standards.

Despite these many external pressures, institutions themselves are important influences on the content of the translation regime. Data standards do not always offer precise operational definitions and logics to determine the data value; they often couple conceptual definitions with a set of valid end states, leaving institutions considerable leeway in the translation process itself. Institutions nearly always control the technical implementation of data standards. Under different alternatives, a particular state of the world can be translated into different values within a data standard depending on how the translation is performed. UVU can thus choose whether to use the state from an applicant's current or permanent addresses when selecting the value of *STATE_ORIGIN*, which USHE simply defines as "The state

code indicating the student's state of origin as described at the time of their first application to the institution, if one is available" (Utah System of Higher Education 2013c, p. S-11). That decision is embedded in functions and validation procedures, but the data standards do nothing to specify how the function should evaluate a student living in Twentynine Palms, California, who considers her permanent home to be Moab, Utah, so long as the function returns a valid value. That a function for doing so is provided in the base Banner package does not prevent the institution from changing that. The decision of how the function evaluates the primary data fields to create a secondary translation is a design choice rather than a predetermined outcome, one made and implemented at least in part by the institution.

The institution is also the data collection point, giving it the power to choose both what data to collect and what interactions to translate into data. This is a powerful tool in shaping data: interactions and characteristics that are not turned in to data are not simply missing; they are untranslatable and hence illegible. This prevents them from being considered in decisions. The standard Banner package includes a field for students' religious preferences. UVU does not collect that data from its students, however. Ostensibly, this is because of a concern that asking students to identify religious preferences would create the impression that UVU was supporting the dominant religion of its community, The Church of Jesus Christ of Latter Day Saints. This has not prevented UVU's Institutional Research & Information office from including that question on its student opinion surveys, the most recent of which that asked the religion question found that 77.2% of students identify with some form of the LDS faith (Institutional Research & Information 2013, p. 45). That data is not included in Banner, however; more than 75% of students' data records have a null value for *RELIGION*. As a result, the institution does not routinely consider religion in its decision-making, even though such a large number of students sharing a common worldview present many of the classic problems of in-group/out-group dynamics.

Religion is, to UVU, illegible. This is not at all to say that the institution is hostile to either LDS Church members or non-members; its President, Matthew S. Holland, is the son of one of the highest authorities in the LDS Church and an active church leader in his own right and yet has consistently promoted religious inclusivity toward those outside the LDS Church as an important element of UVU's Core Themes. But the decision not to collect data with which to populate *RELIGION* does leave religious preferences opaque to the institution. The institution cannot ask questions about the role of religion, either as a belief system or as a social institution, in the operation of educational programs. It cannot consider whether students who are not LDS Church members have lower retention rates, a possible sign that they feel excluded from the social life of the campus. It cannot consider whether LDS Church members are less likely to complete the FAFSA and thus to receive Pell Grants, a possible consequence of a strong ethos of self-sufficiency and financial conservatism within LDS theology and culture. UVU is quite effective addressing these questions within the limits of survey research methods, but a full canvas of students over time is impossible. This leaves UVU unable to "read" a characteristic that is central to many students' identities.

3.3 Characteristic Translations Within the Regime

The data translation regime is not substantively neutral; it favors certain types of outcomes over others. In a relational database such as that used at UVU, one can identify at least three characteristic types of translations in the data (as well as, of course, numerous translations that are relatively unique and not analyzed here). These characteristics translations describe how the ontological character and meaning of states of the world commonly change over the course of the translation process. The result is that translations are most often analytically incommensurable with the reality they purport to express: the words attached to the conditions may be similar, but they are embedded in an entirely new structure.

3.3.1 Normalizing Translations

One type of translation establishes certain states of the world as part of the realm of normally existing conditions, thus implicitly establishing all other states as deviations from normalcy in some sense. Such translations typically have the effect of reducing the states of the world to only those within the realm of the normal data states. Those represented in the database are thus represented only to the extent that they are capable of being represented within that normal realm; to the extent that they deviate from the normal world as it exists within the database they cease to exist analytically.

The simplest normalizing translation is from relevance to existence. Data is collected based on what the collectors find relevant to their interests: it may shed light on a question they need answered or a decision they may make, or it may be needed to comply with requirements of an external authority. Data is not collected, however, on matters that are not of interest to the institution, nor on matters for which the existence of data is counter to the institution's interests. One common objection to data collection and analysis within IRI was that UVU could be forced to make the data or subsequent analyses of it public under Utah's open records laws. Most frequently this objection was used with projects that might collect data that subjects might consider sensitive but that was not protected by privacy laws, a not unreasonable protection but nonetheless one that is driven by a specific interest on the part of the university. Those characteristics or states of the world were considered irrelevant to decisions, and thus not collected.

But when questions arise about such characteristics, irrelevance turns into nonexistence. The characteristics about which there is no data frequently function not as unknowns which need to be estimated or otherwise accounted for in analysis, but are rather ignored, treated analytically as if they do not exist or, at best, subsumed into platitudes about "context" that fade into the background when the data is available. This is more than just saying that nonexistent data does not exist: it is not data about a given characteristic that is translated into nonexistence but the characteristic

itself. Having determined that religion is irrelevant to decision-making and not collected information about it, UVU's students cease, analytically, to have religious preferences.

A similar process takes place with regard to the conditions that a characteristic might take on. Translation regimes transform the diversity of possible conditions of a characteristic into a set of acceptable data values. Those conditions that cannot be represented by a valid data state become represented not as themselves but as deviance: the data is missing; it is given a residual category value such as "other," "not applicable," or "not available"; it is forced into one of the valid data states even if that does not actually represent the state of the world. So the diversity of gender identifications are translated into categories of normalcy that are represented by the values "Male" and "Female," and invalid data that exists in a state of deviance from normalcy, first as "Unknown" and then, with the deprecation of that value in the USHE data standards, into a forced choice of a valid but untrue data state. Transgender identities are not simply statistically rare; they are abnormal. And as in the case of irrelevant characteristics of the world, deviant conditions of the world become analytically nonexistent, assumed to be trivial exceptions to a meaningful interpretation of reality.

It is important especially to understand what it means to say that states of the world *analytically* cease to exist. The qualifier is an important limitation. No one at UVU would deny that many students are religious; the lack of data does not preclude thinking about the characteristic. In a culture where decisions are legitimated in part based on the ability to support them with data analysis, a characteristic that is not datized cannot be analyzed, and so decisions about it cannot be legitimated and are unlikely to be built into policy. Nor can assumptions about the characteristic be questioned. This is perhaps the most pernicious aspect of the translation of relevance. While a characteristic for which data is unavailable may not exist analytically, it may be very prominent culturally, in many cases functioning as part of an ideal type representation and assumed to be true of all cases. The culture of the region carries with it a strong religious identity. The result is the assumption that any one student is a member of the LDS Church until they are known to be otherwise.

It is also important to recognize that analytical normalcy is different from social normalcy, by which I mean the existence of certain conditions as the normal or typical condition from which other conditions vary. Self/other distinctions are a form of social normalcy: whites or men represent the normal or typical, while people of color or women are an "other" defined in relation to the norm. The analytical normalcy that I posit here includes both the typical and other categories in normalcy; deviance constitutes existence outside of the set of recognized categories rather that existence within one of the atypical categories. Analytical normalcy does not imply social normalcy. "White" is, analytically, merely one category of *PRIMARY_ETHNICITY*, not different from other values within the translation regime despite being the only socially normal value. Nor, however, does analytical normalcy challenge social normalcy: the equation of "Male" with normal takes place outside of the translation regime, so that when the translation regime categorizes someone as male or female it does nothing to prevent the substitution of typical and atypical.

The translation of irrelevance to nonexistence played a significant role in the creation of UVU's "15 to Finish" program (First Year Experience and Student Retention 2014), which encouraged students to take 15 credits per semester in order to graduate in 4 years. The assumption behind the program is that students who attend full-time are not only more likely to graduate on time; they are more likely to graduate at all. One of the core messages is of the program that it is better to reduce or eliminate outside work in order to attend full-time, even if that requires students to take out loans, because they will be more likely to finish, finish faster (especially within the limits of Pell Grant eligibility), and spend more years earning an income commensurate with their completed degrees. The analysis performed in support of the program did indeed show that this was the case. But it did not consider whether this was practical for all students. UVU does not collect effective data regarding the family status or family income of its students; the only systematic effort at data collection regarding the number of children students have or parents' income that is integrated into Banner is the FAFSA, but institutional privacy protections limit the transfer of FAFSA data outside of the financial aid office and low rates of FAFSA completion make such data unrepresentative in any case. Students with high family incomes might find it much easier to attend school without working, while those with families might find it especially difficult to reduce or eliminate outside work. Yet neither group exists at an analytical level. The program does have a strong ethos that 15 credit hours may not be appropriate for all students because of their family status or availability of parental support. But that is not implemented formally in the way that, for instance, the various triggers are built into the Stoplight program. "15 to Finish" is for all students, "with exceptions, of course." These characteristics are irrelevant to the institution's data collection efforts, become illegible because they are not collected, and ultimately cease to exist as part of the "normal" world in which administrators operate.

3.3.2 Atomizing Translations

One of the generally accepted best practices of relational database design is that data fields should be atomic, representing one and only one value for one and only one characteristic. To the extent that this is practiced (and it usually is), the result is that translation regimes will represent the world in atomistic terms, fragmenting characteristics that are defined as much by their relationship to other characteristics as by their specific conditions into distinct fields that are not connected to each other. These fields are then analyzed in isolation from each other rather than in the contexts that make them meaningful to the people represented in the database.

Individual identity is highly susceptible to atomization. Complex, intersectional identities frequently bring together different categories of identity into a coherent whole that does not exist within a database: "Jewyoricans" are fragmented into atomistic categories of religion, residence, and ethnicity without the relationships among them that are central to the identity of Jewish New York residents of Puerto

Rican descent. These categories reflect both the principle of atomicity—separate fields for separate characteristics—and the data standards to which the institution must conform. The USHE reporting standards for STUDENT maintain separate fields for ethnicity and state of origin (and, of course, do not collect information about religion) (Utah System of Higher Education 2013c). This makes representing complex identities that reflect not just one or another aspect of one's identity but the intersection of or relationships among multiple aspects of one's identity quite rare; data is often analyzed along ethnicity or gender, often sequentially but rarely both at once. There are people represented in UVU's data who are Black, and people who are female; there are some who are both. But there are no Black women in the data.

Atomizing translations can be especially complex when trying to translate a narrative into data. In such cases, it is necessary not only to separate characteristics but also to reduce complex conditions into nominal data states that conform to the validation rules and data standards. One might consider the case of students who transfer in large numbers of credits that reflect their prior educational experiences that are at best tangentially related to their current educational ambitions but don't meet the requirements of their current degree program. All of these characteristics are included in STUDENT: *PREVIOUS_EDUCATION* captures whether the student was enrolled at another university in the past, *TRANSFER_CREDITS* reflects the number of hours brought in, *TOTAL_CREDITS* identifies the number of credits earned at all institutions, and *STUDENT_CLASSIFICATION* performs a secondary translation that characterizes overall progress toward the degree. But the fields don't reflect the narrative of a student entering, leaving, and returning with different educational goals and having far more credits than are actually needed to graduate while not being anywhere near completing the current program. A student may be classified as a "Senior" but have perhaps 2 or 3 years of additional coursework to complete in order to graduate. The narrative that provides meaning to the value in *TOTAL_CREDITS* is lost; it is reduced to a name: 142.

In most cases, atomizing translations are driven by the content domain rather than the technical domain; the latter merely implements atomicities that are already practiced in other contexts. Technical limitations do not force atomicity on those using extracted data. The different characteristics can be quite easily brought together through simple concatenations of fields or crosstabulations of extracted data. The IPEDS data standards in fact do exactly this. Institutions are required to report enrollment by ethnicity separately for men and women, allowing the federal government to see the intersectionality of the two conditions. UVU's Student Success and Retention Dashboard allows analysis by two characteristics at once, making it possible to see the effects of a wide range of two-dimensional intersectionalities, and with some rather awkward technical gymnastics a very narrow set of three-dimensional ones, on graduation and retention rates. The multi-character *ETHNICITY* field in the USHE data standards shows how a secondary translation of atomic data can create a field that captures the complexity of multiracial identities.

Narratives, too, can be stored in data systems. Banner includes a data table in which comments can be stored. These can provide the narratives that are stripped

away by atomizing translations—if users actually use them. Extracting data from comments is notoriously difficult, requiring complex expressions, tedious analysis, and careful interpretation to make them legible to the institution. Suggesting that the best source for a particular data point is found in COMMENT is universally reviled with UVU's IRI office, but it can be, and sometimes is, done. But this is rarely the case, and even when it is the narrative structure of the comment is rarely used fully, the preference being to identify a nominal value that may be extracted from a comment rather than a more structured field. To be sure, many of these cases involve a re-translation of atomic data. But they do show that other possibilities exist and thus make clear the nature of translation as a choice rather than an inherent technical limitation.

3.3.3 Unifying Translations

In spite of the imperative toward atomicity, there is a counter-tendency toward unity in translation regimes as well. The detailed and diverse conditions of reality frequently exceed the capability of data systems to store them or the ability of analysts to manage them. A characteristic that can have thousands of potential values, especially when those values are expressed in a nominal level of measurement, does little to bring legibility to the state of the world. Diverse states of the world must often be translated into a small number of values that bring many different conditions together into a common data state.

One translation process that unifies disparate conditions is grouping a large number of possible conditions into a small number of data values. This creates a unified group that may not, in fact, exist in reality or that is at least far more complex than is expressed in a single value label. The USHE ethnicity categories are an example. The standard defines "Asian" as "A person having origins in any of the original peoples of the Far East, Southeast Asian, or the Indian subcontinent including for example Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam" (Utah System of Higher Education 2013c, p. S-14). The common label "Asian" hides a wide range of diversity within the definition; UVU had Asian students admitted from 27 countries other than the United States among its Fall 2013 students. It seems reasonable to expect that they would have considerable differences among them, and in many cases might find more in common with other racial groups. The borders Pakistan shares with Iran, Afghanistan, and Tajikistan defines Pakistanis as Asians and thus unifies them with students from Japan or Indonesia while separating them from citizens of the surrounding countries who are defined as "White," a category that includes those "having origins in any of the original peoples of Europe, the Middle East, or North Africa." Similar differences in racial identity exist between African and Black immigrants from Africa or the Caribbean and among immigrant groups themselves (Benson 2006), who are nonetheless unified into a single category of "Black or African-American."

Data systems may also unify characteristics temporally. Characteristics that vary over time become essential and fixed in data systems, stripping away the contingency that is often at work in them. Again, the USHE ethnicity standards are instructive here. The USHE standards make reference consistently to the origins of the student, suggesting that ethnic identity is a fixed part of a person's overall identity. As a result, it can be stored in the data systems and reported consistently over the course of a student's academic career. But there is considerable evidence that ethnic identity is not essential; rather it is a characteristic that is situated in particular circumstances and can change with them, such as when the student moves from a public to a private space or into and out of spaces dominated by heritage identities (Zhang and Noels 2013). One might expect this to be especially strong among students who identify with multiple ethnic or racial groups. This situational variability is not captured by the data system, however; the permanence of the data state implies a permanence to the state of the world it purports to represent that may be accurate on average but may not be so at any given moment.

3.4 Translating the Subjects of Data Systems

The technical structures of a relational database, such as tables, functions, business rules, and queries, translate states of the world into data states based on standards established by social structures such as cultures, states, and organizations. These regimes also translate the entities about which data is collected into "inforgs," entities that exist solely as bundles of information. Within many of the structures that guide data use and data-driven decision-making inforgs behave quite differently than people, fundamentally changing the power dynamics of representation in decision process. I explore two structures related to representation in this section. First, inforgs significantly complicate the way that data-driven decision processes can be considered representative of students. While a less data-driven process emphasizes a trustee model of representation in which the decisionmaker is seen as acting in the best interest of the student, a data-driven process that translates students as inforgs requires decisionmakers to create constructs that ultimately represent themselves rather than students. Standard approaches to protecting student privacy are also considerably more problematic in translated data processes. Approaches to privacy typically rely on restricting the flow of information. A traditional approach views this as a protection of an individual. But when the individuals exist solely as inforgs as in a data-driven decision process, restrictions on the flow of information destroy or at least degrade the inforg itself, excluding the associated person from the process.

3.4.1 *Inforgs in Data-Driven Decision Processes*

In recent decades, higher education in the United States has seen dramatically increasing corporatization, bureaucratization, and rationalization of higher education derived from the for-profit sector but increasingly common in the public and private non-profit sectors as well. A central feature of this has been the emergence of accountability regimes, in which

> a politics of surveillance, control, and market management disguise[es] itself as the value-neutral and scientific administration of individuals and organizations (Tuchman 2009). Related to strategic planning, this accountability regime supposedly minimizes risks for an organization (or corporation) by imposing rules about how work will be done and evaluated. (McMillan Cottom and Tuchman 2015, p. 8)

The scope of such regimes goes far beyond traditional notions of legal and financial risk, reaching into the realm of operational control through data-driven decision-making processes. Accrediting bodies demand that mission fulfillment and student learning be demonstrated through "meaningful, assessable, and verifiable data—quantitative and/or qualitative, as appropriate to its indicators of achievement" (Northwest Commission on Colleges and Universities 2010, Sect. 4.A.1) and that institutions practice "regular, systematic, participatory, self-reflective, and evidence-based assessment of its accomplishments" (Northwest Commission on Colleges and Universities 2010, Sect. 5.A.1). The results of these data-driven analyses are "used for improvement by informing planning, decision-making, and allocation of resources and capacity" (Northwest Commission on Colleges and Universities 2010, Sect. 4.B.1). Institutions that fail to use appropriate data-driven processes to evaluate mission fulfillment and student learning risk punitive actions by accreditors. For example, in June 2013, the Middle States Commission on Higher Education, the largest of the regional accrediting bodies in the US higher education system, issued warnings that the accreditation of ten schools was in jeopardy; nine of these institutions had failed to demonstrate compliance with standards relating to planning, effectiveness, and learning assessment (Middle States Commission on Higher Education 2014).

The reliance on data in assessment, evaluation, and planning—arguably the most important decision processes in a university—is a paradigmatic case of the broader model of data-driven decision-making. Mandated at the primary and secondary levels in the United States by the now superseded No Child Left Behind Act of 2001, data-driven decision-making compels institutions to use data "to stimulate and inform continuous improvement, providing a foundation for educators to examine multiple sources of data and align appropriate instructional strategies with the needs of individual students" (Mandinach 2012, p. 72). The model is based on business management theories (especially those derived from manufacturing), including Total Quality Management and Continuous Improvement. The model organizes and interprets multiple types of data into information that is meaningful to the users. This then becomes actionable knowledge when users evaluate and synthesize the available information, ultimately using the information to either inform discussion or to choose actions. This process is cyclical and takes place within a range of vary-

ing organizational contexts (Marsh et al. 2006). The result is held to be a more rigorous and informed decision process that allows educators to teach more effectively and administrators to operate more efficiently and reliably (Mandinach 2012).

Unexamined in this model is the nature of the data that is driving decision-making. Data is, from the perspective of data-driven decisions, seen as an objective representation of a real world. This realist view is fundamentally flawed, however. In order to understand what a data point means, it must be understood as a representation of something within a nexus of problems, models, and interventions rather than as an abstracted object. The process of making reality legible reflects a fundamental problem: the relationship between that which is to be represented and the data state ultimately representing it is one-to-many; therefore data systems must select a single data state from among the many possible in order to produce legible knowledge. Hence the second key element: that data is itself constructed by social processes. I have elsewhere (Johnson 2015) called this process the *translation regime*, which one might define as the set of implicit or explicit principles, norms, rules, and decision-making procedures through which single, commensurable data states are selected to represent states of the world[4] that provides an external source of stability for the data system and allows it to bring legibility to the represented conditions (Mitev 2005). One could look to gender as a paradigmatic case of translation, with myriad possible gender expressions reduced to a small number of values, most commonly "male" or "female," by data standards and validation tables that reflect social norms, in particular those at work in the accountability regime of the institution.

From this perspective, data-driven decision-making takes place within an abstracted model world that resembles any reality external to it in one of many possible ways selected by the translation regime. In a data table, data exists in columns where the data has a common framework, but it also exists in rows that relate data points in different columns to each other through association with some sort of entity: data is information *about some things*, students and courses in the case of UVU's core institutional research data systems. These things in the database can have no more objective existence than the characteristics that the database attributes to them. The translation regime does not simply translate the characteristics of objectively existing entities into the columns of a database; those entities that make up the rows are also translations, whose existence is defined strictly by the information with which they can be associated.

These data entities are best described as what philosopher of information Luciano Floridi terms "inforgs":

> In many respects we are not stand-alone entities but rather interconnected informational organisms or *inforgs*, sharing with biological agents and engineered artefacts a global environment ultimately made of information, the infosphere. This is the informational environment constituted by all informational processes, services, and entities thus including informational agents as well as their properties, interactions, and mutual relations. (Floridi 2010, p. 9, emphasis in original)

[4] This definition follows that of Krasner's (1982, p. 186), used to define regimes in international relations.

An inforg is characterized as an entity that is de-physicalized, typified (represented as an instance of a class of identical objects), perfectly clonable, and existing only through its interactions with other inforgs. While the extent to which this ontology, which Floridi calls "informational structural realism," is an adequate description of being more broadly remains controversial, the sense of inforgs inhabiting an infosphere captures well the ontology of the model world in which a data-driven decision process takes place. In such a model world, data consists of signifiers of states that attach to inforgs. In a star schema, for instance, data is divided into fact tables that describe entities and dimension tables that describe conditions that those entities can take on. Each row in the fact table represents one entity, named by the data table's primary key, and that entity has no characteristics other than the facts stored in the row, that can be joined to the row, or that are stored in the related dimension tables. These inforgs are thus the only kind of entity that can exist within a data-driven decision process.

3.4.2 Informational Representation

Decisions in higher education are political decisions in the most basic sense: they are decisions made to govern a collective entity, in this case a postsecondary educational institution. As such, those that are affected by this decision, as in all political decisions, have a legitimate claim that they ought to have meaningful input into it in some fashion. This is the origin of the problem of representation, a problem not challenged by the fact that the decision takes place in a bureaucratic rather than legislative institution. Presumably, then, decisionmakers in higher education intend for their decisions to represent, in some form and among other considerations, the students about whom they are making decisions.

One might analyze different modes of representation along two dimensions. The first concerns the level of participation. Participatory models involve all those who have a claim to input in the process of making the decision; representative models vest that power in a relatively small group of individuals who act for the group as a whole. A second dimension considers the relationship between the decisionmakers and the group. Promissory models view the decisionmaker as an agent who acts on behalf of those they represent as principals, while autonomous models allow the decisionmakers the freedom to act on their own. The most common models fall into either the autonomous/participatory or the promissory/representative quadrants. Direct democracy, in which all members of the polity participate directly in policy-making, is the standard case of the former; the trustee-delegate dichotomy, in which representatives act respectively in the best interests of the represented or as the represented themselves would, is the basis of the latter.

This is not to say that the only coherent models of representation fit into one of these two quadrants. Frameworks of representation in the two other quadrants are less commonly observed but nonetheless important. In descriptive representation, representatives act without any moral obligation toward the positions of the repre-

sented but "in their own backgrounds mirror some of the more frequent experiences and outward manifestations of belonging to the group" (Mansbridge 1999, p. 628). This correspondence of backgrounds acts as a mechanism to ensure correspondence between the interests of the representative and the represented so that a representative acting in their own self-interest is coincidentally acting in that of the represented as well rather than acting out of an obligation to do so. Descriptive representation is an important case of representation that is both autonomous and representative used especially to study representation in bureaucracies (see, for example, Wilkins and Keiser 2004). Jean-Jacques Rousseau's *On the Social Contract* proposes a system in which citizens participate directly in government but represent not their particular individual wills but the "will that one has as a citizen," which he terms "the general will," thus directly participating in government but as an agent of the collective body of citizens that serves as principal. However, neither of these models is of practical value in higher education decision processes. In the case of descriptive representation, decisions are made by actors who cannot resemble the key characteristic of those they might be taken to represent: administrators are not students. Concepts related to the general will have never been shown to be sufficiently clear in any applied context to be of use in making a specific decision. Analysis of representation will thus focus on the direct and promissory models of representation.

In a personalized decision-making context, which we might define in contrast to a data-driven process as one in which either single or multiple decisionmakers use their personal judgment to make what they consider the best decision given the available information under some degree of uncertainty, higher education tends toward a trusteeship model of representation. Even at the smallest of institutions, direct participation in all decisions is impractical because of the number of students and of decisions involved in governing the institution. But there is also a strong strain of paternalism in decision-making at colleges and universities. Students, it is frequently held, cannot be counted on to do what is best for them. Consider, for instance, Austin Peay State University's use of predictive analytics in student advising:

> [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. (Parry 2012)

Such students would, if they chose themselves, make choices that run counter to their true interests (presumably, in receiving a generic college degree at minimum cost); decisionmakers must therefore choose not what the students *would* choose but what they *should* choose. Such a model of representation is defensible only to the extent that the decisionmakers do, in fact, have an adequate view of that interest.

This model of representation breaks down when students are translated into inforgs. Initially, one is tempted to see the translation of students (or of anyone with a claim to voice in a political process) as a gain for direct participation. The promissory models both break down when applied to inforgs. The trustee and delegate approaches both require a unifying concept that acts as the wholeness of the represented (interest or will, respectively) that guides how the agent acts on behalf of the principal, one that is lacking when the principal is no more than a bundle of information: which piece of information defines that unifying concept? But while a personalized process of direct participation requires some complex structure that allows universal participation in the process of developing policy alternatives, manages extensive deliberation among those alternatives, and aggregates preferences into a decision, a data-driven process can bring the participants in as inforgs and then aggregate their informational characteristics. The capacity for participation in data-driven decision-making is apparently limited only by the power to collect and process the information that constitutes the inforgs.

This understanding of representation assumes that inforgs have an objective or realist ontological status, existing in their own right rather than being constituted by actors outside of themselves: the data row represents a physically existing student as they are in the "real" world rather than existing as an inforg that has been created by someone other than the represented. The analysis of the data structures above shows that this is not the case. Inforgs are themselves social constructs, and both their existence and their characteristics reflect the same social pressures and structures that data fields do. As such, the idea that inforgs are capable of being independently represented in a data-driven decision process is fundamentally unsound; what is represented is the constructive activity of those creating the inforgs. There is the appearance of direct participation, but the participants are not representations of students but actants created through the translation regime. What is represented is as much the constructors' understanding of students that is built into the data driving the decision process.

Data-driven decision processes thus present a fundamental contradiction. While they are instituted as objective processes, it is clear that no process of representing students can take place within them without the process of data creation also being a process of imposing external values and assumptions. The inforgs are created by those who create the data system, and decisions about them can only be made if decisionmakers supply their own concepts of interests of will to guide the application of promissory models of representation. This is, to be sure, true of personal decision models as well, but in those models there is a clear connection to individuals against which those assumptions can be checked. In a data-driven model, there is nothing to check against beyond the data; the students exist solely as data. The objectivity of the process, its supposed virtue, is thus a fiction needed to make the process work.

3.4.3 Destructive Privacy Among Infogs

Representing infogs becomes more seriously compromised when considered in relation to information privacy. In the United States, students are protected first and foremost by federal laws including but not limited to the Federal Education Rights and Privacy Act (FERPA), but also by a range of state laws, institutional policies, and data handling standards. All of this is meant to ensure that students are able to maintain a sphere of personal identity and activity safe from intrusion by others, including others' knowledge about the student. Most commonly this is protected by the twin principles of consent and anonymity: personal information may only be used or transferred with the consent of the subject; all other information must be stripped of personally identifying characteristics before use or transfer (van Wel and Royakkers 2004). Certainly these opt-in or opt-out procedures are the bedrock of most institutions' privacy policies, with the latter likely far more common than the former.

Increasing pressures on personal privacy have given rise to more complex perspectives on privacy. It is increasingly common to interpret privacy as a property right in information about one's self. Subjects hold initial ownership rights in information about them, and can exchange that information contractually in information markets, receiving appropriate compensation—or they can refuse to permit the use of such information in the absence of sufficient compensation to encourage the transaction (Solove 2004, pp. 76–81). This approach makes sense, for example, of the willingness of so many to give access to their email to Google: in exchange for an outstanding product, consumers are willing to allow Google to use the information captured to generate profit for itself. Alternatively, Helen Nissenbaum (2010) argues for a reliance on social context to protect privacy. As technosocial systems, the context of information flows is as much a defining feature of data exchange and use as the content of that information flow. The combination of situation, actors, information attributes, and practices of transmission for accepted information exchanges constitute an existing norm of practice that may be violated in the case of new uses of information, such as a data mining practice. Changes in this context that are not supported by its underlying norms are violations of the contextual integrity of the information flows, and in the absence of separate justification violate one's privacy rights. More recently, the European Court of Justice has embraced a "right to be forgotten" under which individuals are entitled to have information about them essentially destroyed, in the instant case by having Google remove links to information about them from search results (Costeja González 2014).

The common thread of each of these approaches to privacy is that they aim to restrict flows of information across parties, transactions, or both. This restriction is frequently considered the essence of data privacy. The centrality of collection (the flow of information from a subject to a data system) and dissemination (the flow of information across data systems or from a data system to subjects) in common definitions of information privacy makes restrictions on flow the *sine qua non* of data

privacy. Such a model of privacy is at least plausibly appropriate for the governance of subjects who are persons; preventing the transfer of information will, presumably, prevent those receiving information from using it to do harm to the subjects of that information. This meets the fundamental criteria of a wide range of ethical frameworks, such as Mill's harm principle, which permits the infringement of one's liberty in order to prevent harm to others, or the more recent proposal of a Hippocratic Oath making "do no harm" the first principle in the use of information and communication technology for development (Mill 2011, p. 17; Rodrik 2012).

Restricting the flow of information fundamentally fails, however, when the subjects are constructive inforgs. The flow of information is what translates subjects (in this case, students) into inforgs in the first place. To restrict that flow is to change the inforg itself. Such restrictions might, for instance, limit the data known about an inforg in absolute terms as privacy restrictions prevent the transfer of certain types of information (when, for example, the subject opts out of sharing of internet use information). Or it might do so in relative terms as it prevents the transfer of information from one source (when the subject installs a privacy plug-in in Chrome) but allows that same transfer from another source (when the subject doesn't bother reading the 31-page terms and conditions for the latest iOS update). Since an inforg is nothing more than a typified and clonable bundle of information, a difference in the information constituting the inforg violates the principles of typification (the difference resulting in inforgs that are instances of two different types) and clonability (the difference distinguishing two instances as different rather than as clones), and is thus the creation of a different inforg.

This becomes even more problematic when a subject opts out of a data system altogether. For a constructive inforg, a complete data opt-out is not simply a withholding of information; it is a complete destruction of itself as an inforg. Prohibition of data flows prevents the inforg from being constructed in the first place. It is perhaps only slightly overdramatic to characterize complete restriction of the flow of data as information suicide for a constructive inforg, as the inforg that protects its privacy ceases to exist in the model world of the data-driven decision process. The physical entity corresponding to the inforg (in this case, the actual student) is at best reduced to context—that there are some students who are excluded by privacy protections. But context, again, exists only in relation to data, which is to say in relation to inforgs. Students who opt to protect their privacy thus exist only as others to the inforgs' selves, defined not individually as entities in themselves but collectively as a typified characteristic of the inforgs (i.e., as a group of identical entities of which the inforgs are not members). Reduced to context that is meaningful only in relation to entities that have corresponding inforgs, those students cease to exist analytically and instead are subsumed as information into inforgs corresponding to other students.

That further complicates the problem of representation as well. Partial restrictions change how subjects are represented; complete prohibitions exclude subjects from being represented entirely. Students are faced with a difficult choice: they can be represented (with varying levels of adequacy given the process of constructing inforgs) in the data-driven decision processes that run the institution that shapes a

significant part of their lives both now and long into the future, or they can choose to minimize the extent to which that institution is allowed into the student's sphere of private activity and identity. To exactly the extent that students choose one good, they undermine the other. In personalized decision processes, the unifying concepts of principal-agent representation can moderate this, with decisionmakers taking into account expressions of students' best interests and wills regardless of—and perhaps taking into consideration—the privacy status of individual students, as these are not data-dependent. In data-driven decision processes, however, with those unifying concepts absent and decisions formally constrained by the available data, representation and privacy are fundamentally irreconcilable.

3.5 Conclusion

These transformations are political acts. The actors that design translation regimes are building structures that embed values and relationships within them that can advantage certain groups over others as the data rather than the actors it represents comes to play a defining role in decision processes. The translation regime begins by representing some groups and excluding other groups, representing some characteristics of individuals but not other characteristics of those same individuals, and representing the data subjects as the data system's designers would represent them rather than as the subjects would. In UVU's data systems, non-credit students and non-degree seeking students do not exist under most circumstances; nearly all queries are designed to filter such students out unless information about them is needed specifically. English as a Second Language students were until recently treated as non-degree seeking and thus left unrepresented in most data-driven decisions. Students' ethnicity is represented but their religion, the most commonly discussed aspect of identity in the student interviews, is not. White students are represented as an ethnicity rather than seeing themselves as ordinary people (who seem to lack ethnicity), as one White student described himself. These translations are not necessarily hostile to the students' representation, but they do quite clearly shape it.

Just as there are many characteristic translations, there are many political acts that take place through them. The creation of data systems is an assertion of self-interest on the part of the designers; the data system embeds their interests in the decision process but not those who have no influence on the design processes; the latter have no way to make themselves and their interests legible even to institutions that might want to take them into account in good faith, let alone those who might deliberately seek to exclude them. The categorization of characteristics creates and fragments groups that could assert their aims to the institution: Black women are forced to choose to work within the defined fields of *GENDER* and *ETHNICITY* to meet their needs and thus to accept racial inequality within the feminist movement or gender inequality within Black culture rather than identifying as Black women specifically and pursuing an intersectional strategy (Hill Collins 2009). Defining states of the world as valid or invalid (e.g., transgender identities) is at the least an imposition of a

normalizing judgment through a means other than surveillance, one that has the same kind of potential to construct individuals and groups as hate speech (Butler 1997).

Similarly, data-driven decision-making becomes much more problematic when we recognize that data is made, not collected. As decision-making increasingly takes place within model worlds created by the process of collecting, managing, and analyzing data, it increasingly transforms people into inforgs and marginalizes considerations not rooted in data as mere context.[5] Data-driven decision-making is part of a larger ethos, one connecting managerialism, technocratic government, and neoliberal politics, that increasingly pervades higher education. The problems of representation and privacy, and especially the tension between the two, stem from the very core of this ethos.

Much of the politics that one would typically expect as groups compete is present in the translation regime. The politics of the translation regime is different, however, in that it is hidden behind a facade of technical specifications. The translations are, superficially, not exercises of power but simply functions and validation tables that store ostensibly objective information about reality. The scientistic ontology and ideology (Haack 1993; Peterson 2003) embedded in information systems creates the appearance of an apolitical process that is not open to contestation. It thus becomes quite difficult to engage from a political perspective. It cannot be challenged technically, as the translation regime is internally valid and self-legitimating. Any test against reality will confirm the validity of the regime so long as the rules are complied with, because the rules include what data can be considered. Data from within the regime will be correct, and there is no such thing as "data" from outside of the regime. The translation regime creates data; all else is anecdote and thus illegitimate. Challenges to the politics of the translation regime must, then, overcome the issue of legitimacy before the regime can be questioned.

The translation regime is thus a significant and problematic form of political power. Integrating both the technical and the social to render its subjects legible to the exercise of power, the characteristic translations that it produces also exercise power in their own right. As such, the fact that data is constructed through translation, among other processes, presents the need for a theory of information justice. Such a theory must rely on neither controlling the possession of information nor its use. If information is simply representational, these would be adequate safeguards. Privacy rights could protect transfer of information, and substantive regimes similar to human subjects protections might prevent against harmful uses. But the constructive nature of data makes these inadequate. Neither privacy use ethics addresses the content of information that is, within the internal framework of the translation regime, accurate. These approaches cannot address the questions that arise in build-

[5] To be sure, one might argue that the portrayal of data-driven decision-making presented here is something of a straw-man argument that neglects the subtleties of and importance of context in the models advocated in higher education. I would argue to the contrary that those models themselves only pay lip service to context; the more context can be used to override data and the more that conflicting data points are to be considered in the decision process, the less data-driven decision-making is distinct from personalized decision-making. If there is something distinct about data-driven decision-making, it is that data must take priority over context.

ing data systems in ways that their translations further rather than undermine the individuals represented in them. Instead, a theory of information justice should be oriented toward understanding data as a socio-technical system, promoting design practices that minimize their potential for domination and oppression.

References

Benson, J. E. (2006). Exploring the racial identities of black immigrants in the United States. *Sociological Forum, 21*(2), 219–247. https://doi.org/10.1007/s11206-006-9013-7.

Butler, J. (1997). *Excitable speech: A politics of the performative*. New York: Routledge.

Costeja González, M. (2014). Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (es), EU:C:2014:31.

Easley, J. A. (2014, March 18). If it's your preferred name, then we prefer it, too. *Dateline: News for Faculty and Staff*. http://dateline.ucdavis.edu/dl_detail.lasso?id=14756&dn=031814. Accessed 19 Mar 2014.

Facebook. (n.d.). *How do I edit basic info on my Timeline and choose who can see it?* https://www.facebook.com/help/276177272409629. Accessed 25 Feb 2014.

First Year Experience and Student Retention. (2014). *15 to finish*. https://www.uvu.edu/success/15tofinish/. Accessed 24 May 2017.

Floridi, L. (2010). *Information: A very short introduction*. Oxford: Oxford University Press.

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

Hall, P. A., & Taylor, R. C. R. (1996). Political science and the three new institutionalisms. *Political Studies, 44*(5), 936–957.

Hill Collins, P. (2009). *Black feminist thought: knowledge, consciousness, and the politics of empowerment* (2nd ed.). New York: Routledge.

Information Commissioner's Office. (n.d.). *The guide to data protection*. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.pdf. Accessed 25 Feb 2014.

Institutional Research & Information. (2012a). *Fact book 2012*. https://www.uvu.edu/iri/documents/additional_resources/factbook2012.pdf. Accessed 12 Mar 2014.

Institutional Research & Information. (2012b). *Student success/retention.* http://www.uvu.edu/iri/indicators/. Accessed 12 Mar 2014.

Institutional Research & Information. (2013). *Student omnibus survey – Fall 2012 results*. http://www.uvu.edu/iri/documents/surveys_and_studies/Omnibus%20Student%20Survey%20-%20Fall%202012%20Results.pdf. Accessed 20 Mar 2014.

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.

Johnson, J. A. (2015). Information systems and the translation of transgender. *TSQ: Transgender Studies Quarterly, 2*(1), 160–165. https://doi.org/10.1215/23289252-2848940.

Krasner, S. D. (1982). Structural causes and regime consequences: Regimes as intervening variables. *International Organization, 36*(2), 185–205.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064.

Mansbridge, J. (1999). Should blacks represent blacks and women represent women? A contingent "Yes". *The Journal of Politics, 61*(3), 628–657.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica: RAND Corporation.

http://www.rand.org/content/dam/rand/pubs/occasional_papers/2006/RAND_OP170.pdf. Accessed 16 Sept 2014.

McMillan Cottom, T., & Tuchman, G. (2015). Rationalization of higher education. In R. A. Scott, & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (pp. 1–17). http://onlinelibrary.wiley.com/book/10.1002/9781118900772. Accessed 22 Dec 2015.

Middle States Commission on Higher Education. (2014, June 26). *Summary of commission actions on institutions*. http://www.msche.org/institutions_recentactions_view.asp?dteStart=4/29/2014&dteEnd=6/26/2014&idCommitteeType=0&txtMeeting=Commission. Accessed 22 Sept 2014.

Mill, J. S. (2011). *On liberty (project Gutenberg eBook.)*. London: The Walter Scott Publishing, Ltd.. http://www.gutenberg.org/files/34901/34901-h/34901-h.htm. Accessed 30 Sept 2014.

Mitev, N. N. (2005). Are social constructivist approaches critical? The case of IS failure. In *Handbook of critical information systems research: Theory and application* (pp. 70–103). Northampton: E. Elgar Pub.

National Center for Education Statistics. (n.d.). *The integrated postsecondary education data system – Glossary*. http://nces.ed.gov/ipeds/glossary/. Accessed 11 Nov 2015.

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

Northwest Commission on Colleges and Universities. (2010). *Accreditation standards*. http://www.nwccu.org/Standards%20and%20Policies/Accreditation%20Standards/Accreditation%20Standards.htm. Accessed 22 Sept 2014.

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/.

Peirce, C. S. (1992). In N. Houser, C. J. W. Kloesel, & Peirce Edition Project (Eds.), *The essential Peirce: Selected philosophical writings* (Vol. 1–2., Vol. 1). Bloomington: Indiana University Press.

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.

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.

Rodrik, D. (2012). *A hippocratic oath for future development policy a hippocratic oath for future development policy*. http://www.policyinnovations.org/ideas/commentary/data/000244

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

Seaver, N. (2014, January 30). On reverse engineering: Looking for the cultural work of engineers. *Medium.com*. https://medium.com/anthropology-and-algorithms/d9f5bae87812. Accessed 28 Feb 2014.

Solove, D. J. (2004). *The digital person: Technology and privacy in the information age*. New York: New York University Press.

Utah System of Higher Education. (2013a, July 1). *Course data submission file, 2013–2014 submission year*. http://higheredutah.org/wp-content/uploads/2013/09/rd_2013DataDict_Course.pdf. Accessed 10 Mar 2014.

Utah System of Higher Education. (2013b, July 1). *Graduation data submission file, 2013–2014 submission year*. http://higheredutah.org/wp-content/uploads/2013/09/rd_2013DataDict_Graduation.pdf. Accessed 10 Mar 2014.

Utah System of Higher Education. (2013c, July 1). *Student data submission file, 2013–2014 submission year*. http://higheredutah.org/wp-content/uploads/2013/09/rd_2013DataDict_Students.pdf. Accessed 10 Mar 2014.

Utah System of Higher Education. (2013d, July 1). *Student_Course data submission file, 2013–2014 submission year*. http://higheredutah.org/wp-content/uploads/2013/09/rd_2013DataDict_Student_Courses.pdf. Accessed 10 Mar 2014.

UVU Student Retention. (2013). *Stoplight report*. http://www.uvu.edu/retention/advisors/stoplight.html. Accessed 13 Mar 2014.

Weingast, B. R., & Moran, M. J. (1983). Bureaucratic discretion or congressional control? Regulatory policymaking by the federal trade commission. *Journal of Political Economy, 91*(5), 765–800.

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.

Wilkins, V. M., & Keiser, L. R. (2004). Linking passive and active representation by gender: The case of child support agencies. *Journal of Public Administration Research and Theory, 16*(1), 87–102. https://doi.org/10.1093/jopart/mui023.

Zhang, R., & Noels, K. A. (2013). When ethnic identities vary: Cross-situation and within-situation variation, authenticity, and well-being. *Journal of Cross-Cultural Psychology, 44*(4), 552–573. https://doi.org/10.1177/0022022112463604.