International Journal of Communication 7 (2013), 556–583 1932–8036/20130005

# Data Mining Difference in the Age of Big Data: Communication and the Social Shaping of Genome Technologies from 1998 to 2007

PETER A. CHOW–WHITE
Simon Fraser University

SANDY E. GREEN, JR.
California State University, Northridge

If the 1990s was all about the information superhighway and the network society, then the first 10 years of the 21st century is perhaps best described as the decade of data. Actors in different enterprises worked feverishly to develop innovative database and data mining technologies for institutional goals such as marketing, social networking, and scientific discovery. These researchers and data entrepreneurs follow an emerging belief that gathering and mining massive amounts of digital data will give objective insight into human relations and provide authentic representations for decision-making. On the surface, the technologies used to mine big data have the appearance of value-free and neutral inquiry. However, as information entrepreneurs use database and data mining technologies to purposively organize the social world, this seeming neutrality obfuscates domain assumptions and leaves cultural values and practices of power unexamined. We investigate the role of communication and social shaping of database and data mining technologies in the institutional context of genome science to understand how various stakeholders (scientists, policy makers, social scientists, and advocates) articulate racialized meanings with biological, physical, and big data. We found a rise in the use of racial discourse that suggests race has a genetic foundation.

If the 1990s was all about the information highway and the network society, then the first 10 years of the 21st century is perhaps best described as the decade of data. Actors in different enterprises worked feverishly to develop innovative database and data mining technologies for institutional goals such as database marketing, social networking, and scientific discovery. They gathered disparate types of information from users and consumers—sometimes with the users' knowledge, sometimes without—and turned this information into analytical data points for measurement, sorting, and classification to achieve different organizational and institutional goals. For example, business analysts, scientists, and law-

Peter A. Chow-White: petercw@sfu.ca
Sandy E. Green, Jr.: sandy.green@csun.edu
Date submitted: 2011–11–02

enforcement agencies developed new forms of knowledge production around core database and data mining technologies (Gandy, 2009; Turow, 2006).

In a recent IBM ad campaign titled "Let's build a smarter planet," a collection of company technologists explains why database and data mining technologies matter. According to one of the IBM scientists, "Every day we are creating fifteen petabytes of new data. That's eight times as much data as there is in all of the libraries in the United States combined." Another IBM researcher in the commercial explains, "If we can analyze and mine this data, then we can understand it. If we can understand it, then we can understand trends about it. . . . The more data you have, the clearer you see." Data entrepreneurs such as these IBM researchers assume that gathering and mining massive amounts of data will give objective insight into human relations and concerns. However, what is the nature of these trends and new data-driven ways of seeing? The ad is an excellent example of technology boosterism in the Age of Big Data, where actors argue that data mining improves our understanding of social and organizational life. Yet , IBM fails to comment on how society shapes data mining technologies and how the use of these technologies may construct, perform, and categorize us in old and new ways. For example, recent research into genomics demonstrates an increased capacity to group people in new ways based upon their genetic characteristics (Rabinow 1992; Rose, 2007. The construction, management, and analysis of a database are more than simply technical exercises in data collection and processing (Bowker, 2010). Data-driven ways of seeing human relations purposefully organizes the social world via communicative acts that incorporate cultural values and practices of power. As scholars in science and technology studies (STS) argue, human actors' decisions, politics, and cultural values socially shape the direction and development of technology and innovation (Bowker & Star, 1999; MacKenzie & Wajcman, 1985). How will the social shaping of data mining technologies at the core of new data-intensive practices of seeing the world mediate social relations, identities, and practices?

We investigate these emerging data-driven ways of seeing the material, social, and digital world and their impact on knowledge production in the enterprise of genome science. In the early 21st century, innovations in communication technologies, data mining, and networks moved to the center of scientific practices (Castells, 2000; Chow-White & Garcia-Sancho, 2012). Molecular biologists and biotechnology entrepreneurs approach new technologies and DNA data in very similar ways to the IBM scientists. In the context of biology, scientists use data mining technologies to understand the molecular reality of race (Condit, 2008; Dubriwny, Bates, & Bevan, 2004; Kahn, 2006; Lynch, 2008; Parrott et al., 2004; Thacker, 2005). During the first decade of the 21st century, debates about the validity of race as a biological construct and the role of race in the new data mining enterprise of genomics ensued across the academy, public labs, policy circles, and biotechnology firms (Bamshad, Wooding, Salisbury, & Stephens, 2004; Duster, 2003; Fujimura & Rajagopalan, 2011 Reardon, 2005). Some scientists and scholars argued that data mining DNA would reveal the truth of race, like truths about gender differences between men and women, and essentialized notions of human capacities (Risch, Burchard, Ziv, & Tang, 2002). However, this perspective puts technology in a central role for mediating the relationship between nature and nurture, and anoints data as the best arbiter of truth for the many potential interpretations scientists might develop from the billion bases in the human genome.

Although genomic scientists conduct research in many various ways, several ethnographic studies (Bolnick, 2008; Fujimura & Rajagopalan, 2011; Fullwiley, 2007) suggest that scientists may articulate data, programming code, and discourses, in ways that shape the perceived reality of race as much if not more than any objective outcome of scientific data analysis. The data are only part of the story. Scientists collect human matter, extract and expand DNA strands, and represent them as sequences of genomic data, transforming mediated biological material into digital 1s and 0s (Thacker, 2005). They then use computers as a communication technology to decode, recode, and encode the data using data mining algorithms and statistical routines. Technologically driven communicative practices or discourses essentially constitute and shape these coding processes. Actors use them to make sense of data, and this sense making creates and promotes certain ways of seeing that tell particular stories or truths about race and human behavior (McCann-Mortimer, Augoustinos, & LeCouteur, 2004). We suggest that this development is a potential game changer for how the institution of science produces knowledge. We explore theoretically and empirically the new conditions of racial formation in the mediated space of communication technologies.

To illustrate our ideas, we examine the use of racial concepts in the context of genomic science as a case study for the role of communication in the social shaping of big data, digital databases, and data mining technologies. Although there are many potentially competing discourses within genomic science, we focus on two dominant discourses regarding the use of race in genome science: (1) racial realism and (2) social constructionism. Legal scholar Derrick Bell (1992) used the term *racial realism* to describe the conservative pushback against progressive social policies in the early 1990s. Conservatives coupled this ideology with discursive products such as the publication of *The Bell Curve* (Herrnstein & Murray, 1994) to reassert notions of innate human differences. In genome research, racial realists view race as a biological phenomenon and argue that the clustering of genomic variation reveals concordance with socially organized racial groups (see Risch et al., 2002). Since the post–World War II UNESCO Statements on Race from leading social scientists and natural scientists, scholars have advocated for the abandonment of biological notions of race in favor of political and social notions of race. Scholars argue that race is a social construct that derives from social conflicts over culture, politics, and values as opposed to a biological fact (e.g., Condit, 2008; Gould, 1996; Hall, 1998). Actors derive racial meaning and experience from the power network of social structures and discursive formations (Omi & Winant, 1994). In genome research, the term *social constructionist* refers to a stakeholder who argues that race is an invalid concept for genomics as there is no racial structure in genome variation. We acknowledge that focusing on these competing discourses may miss some nuances in what *race* means and how actors use it in biology and biomedical research (see Bliss, 2012; Condit, 2008). We employ these categories as heuristic devices in order to track and analyze macro trends in competing overarching logics (Green, 2004; Green, Li, & Nohria, 2009). We believe the use of these categories will help us show how new data mining technologies are discursively shaped over a specific and significant time-period. Moreover, many scholars suggest that genomics is just the latest episode of centuries-old debates and scientific research on the nature of identity and human bodies (Wailoo, Nelson, & Lee, 2012). Yet, the technological context mediating this trend is new. The data mining of DNA, genomics, *is* different from older debates about race and science. The larger trend in society toward a data-driven way of seeing intersects with specific institutional spaces to provide new life, challenges, practices, and technologies for old arguments.

To conduct our analysis, we first discuss the emerging relationship between data, information technologies, and race in the digital age and the role of data mining practices. Second, we describe briefly recent issues and controversies regarding the use of race in genome research. Third, we address the dearth of empirical evidence about the role of racial discourse in genome science (Lee, 2009; Sankar, Cho, & Mountain, 2007). To begin to fill this important gap, we measure the prevalence and variation of racial realist and social constructionist discourses in genomics with a longitudinal content analysis of top biomedical science journals from 1998 to 2007. We then use these measures to explore and describe how data mining as a communicative practice shapes the use and meaning of race in scientific research. We conclude our analysis with a discussion of the theoretical and practical implications of our findings.

## Data Mining and the Social Sorting of Racial Difference in the Digital Age

In the 1990s and early 2000s, scholars, policy makers, activists, and entrepreneurs commonly used the word *information* to talk about the social and organizational changes brought on by the expansion of digital networks into all walks of everyday and institutional life. As the Internet grew in the first decade of the 21st century, along with innovations in communication technologies such as the smartphone, gaming technologies, and Web-based apps, the conversation shifted from information to data and database technologies. A number of scholars pointed out the central role of database and data mining technologies in the information economy for managing data, networking information, and constructing institutional knowledge (Elmer, 2004; Gandy, 2009; Manovich, 2001). Data mining is the nontrivial process of using algorithmic techniques to discover (faster than is humanly possible) hidden patterns and unknown relationships among many variables in masses of observed data to produce understandable, meaningful, and potentially useful information for knowledge building and decision-making. In the information age, data mining technologies have diffused across many different institutional settings and represent a shift away from hands-on knowledge-making practices toward digital and algorithmic ways of seeing and mediating the social world.

Two key features of the social life of data mining technologies are the proliferation of the technology across organizational and institutional settings and the interaction of algorithmic practices of seeing meaningful patterns in databases with already existing, institutionally or organizationally based practices of knowledge building and decision-making. Models that represent the data mining process tend to start with data sets and neglect to show how the data got there in the first place. Simply put, the models are silent about the discourses, decision-making, and politics involved in constructing the data and designing data mining algorithms.

Thacker (2005; see also Chow-White, 2008, 2009) examines the role of communication in the social shaping of data mining in human genomics. Kuper and Szymanski (2009) use a data-driven approach to address everyday debates in the sport of soccer. In a more popular text, Lewis (2004) chronicles the uptake of a data-driven approach to building a professional baseball team, also known as Billy Beane baseball. Oscar Gandy (1993) calls the use of data mining by companies and governments the "panoptic sort" and suggests that data knowledge discovery in databases is the latest technology of surveillance. Surveillance has increasingly become dependent on digitized information infrastructures, "which simultaneously made them even less visible and even more powerful, and also produced some specific kinds of coding" (Lyon, 2002, p. 245).

Social media and Web 2.0 increases the number of people who can create and share their own content—what many refer to as user-generated content (UGC). Turow (2006) argues that people participate in these systems because of anxiety over what he calls "niche envy," which is the 21st-century digital version of "keeping up with the Joneses." That is, individuals fear they will miss retail or service opportunities if they are not included in the right database. To identify and target the best customers, companies collect a number of different types of information (demographics, geographic location, shopping history, etc.). These companies persuade people to give up their personal information freely or for an incentive such as a free e-mail or social media service membership. Companies relegate less-valuable customers to lesser services, such as when financial institutions score their clients and greet more-desirable clients with friendlier call-center scripts while excluding others from promotions or coupons (Turow, 2006).

This increase in data, however, raises questions for scientists and commercial interests who have been trying to figure out what to do with it and how to integrate, aggregate, and mine disparate sources and behaviors such as tweets, traffic patterns, consumer behavior, and DNA. The seemingly relentless increase of computer capacity, power, and speed—along with the connective capacity of Web 2.0 UGC applications—generate dramatically larger amounts of data. Amassing and analyzing large amounts of data is both easier and ubiquitous.

While every day people deal with information overload, policy makers, scholars, and entrepreneurs face a new problem: the deluge of data. The media regularly reports stories on the data deluge and the benefits of large-scale databases and data mining, such as fighting credit card fraud and database hacks (The data deluge, 2010). Social science and humanities funding bodies challenge scholars to address how the big data stored in continually expanding databases changes the nature of research (e.g., Digging into Data Challenge, 2011). Scientists largely credit the rapid success of the Human Genome Project (HGP) and advances in genomics and health to innovations in data mining technologies. For example, Leroy Hood, who developed one of the first DNA sequencers, refers to data mining in genomics as "discovery science," a type of science distinct from the conventional hypothesis- and theory-driven model (Hood, 2001). In the Celera Human Genome Project, Craig Venter pioneered a discovery process called the *shotgun method* that became widely used by scientists post-HGP. Almost a decade after the HGP, technology media claim that a data-driven approach makes traditional scientific hypothesis-driven approaches obsolete and point to the success of Google's business and mathematical models and the "end" of theory:

> Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising—it just assumed that better data, with better analytical tools, would win the day. And Google was right. (Anderson, 2008)

Observing the emerging mix of data, information, and analysts, Brown (2011) refers to the ability to mine massive amounts of data quickly for analytical guidance and economic value as *datanomics*. Entrepreneurs imagine data as the plastic of the new economy, easily shaped into knowledge for making

better decisions in business and creating innovative experiences for the customer and user.

When entrepreneurs following the big data model claim the end of theory, local expertise, and circumstance, scholars in the social sciences and humanities, especially communication studies, need to investigate because meaning, context, and action lay at the heart of human affairs. Since the 1940s when Claude Shannon proposed a limited, transmission model in his mathematical theory of the process of communication, scholars in communication, media, and cultural studies have been pursuing a more complex view of the communicative process, such as Stuart Hall's encoding/decoding model. The mediation of human relations by communication technologies lies at the core of the communication process, from language to print, radio, film, television, Internet, new media, and now algorithmic culture and data mining. Data collection, management, and analysis are communicative acts of decoding and encoding that transform and normalize complex activity and phenomena. We argue that we need this line of inquiry again in the age of big data.

*Data* differs from *information* insofar as entrepreneurs and scientists purposefully collect and organize objects into databases. Typical definitions focus on data as raw materials or objective facts in the form of signs or symbols about events, people, things, or activities (Khosrow-Pour, 2007). Actors can amass and assemble structured data through purposeful collection. Actors can also gather unstructured data, which may look like information, such as updates on Twitter, profile interests on Facebook, sports statistics, and human DNA sequences. In a critical examination, Hayes (1993) begins by referring to data as "recorded symbols" characterized as "primitive" (p. 2). Hayes borrows typical administrative definitions of data, but challenges common notions of data and facts being synonymous. Like critical information scholars Lyon (2005) and Terranova (2004), Hayes (1993) notes that sometimes "people will treat [data] as synonymous with facts" (p. 2). He argues that we need to make a distinction between data and the facts that data represent. If not, then uncritically equating data with facts, Hayes (1993) writes, "can produce innumerable perversions" (p. 2). The key point for our discussion is that *data are representations, cultural objects that stand in for stimuli and mediate relations*. In the haste to move massive amounts of information through new communication networks rapidly on a global scale, actors reduce information to bits and bytes and neglect to focus on the meaning-making and mediating role of the raw material of the information age. In the shift to big data, the same process is taking place. Data are not in the same state as natural raw materials. Data are no more pre-meaning or pre-communicative than information.

Some see data and data mining as ethically neutral because a computer chooses groups and clusters according to the affinity of data points and variables. Yet the reduction of social facts and human relations to data points hides the complex decision-making process of decoding and encoding human stimuli and naturalizes a set of assumptions and calculations that undergird and provide the rationale for classification systems. The digital code of databases hides politics, arguments, agreements, attitudes, and values (Bowker & Star, 1999). The processes of decoding and encoding an object or action and representing it as data do not simply reflect that object or action in the world; the very act of decoding and encoding also constitute the object and transforms complex human action and social relations. Decoding reduces stimuli to what appear to be basic units or sub-units of information in the form of objects in a database, numbers in a spreadsheet, and the 1s and 0s of digital code underneath. We argue that the transformation of stimuli to data (and the encoding of data as knowledge) is a political and

communicative act and we need to recover the complexity of this simplification process in order to understand the politics of the transformation and the social shaping that data mining technologies perform. Zarsky (2003), for example, argues that choice making in a data mining clustering approach is objective, in contrast to a human-based classification approach where a hypothesis drives the grouping of factors. The key difference between classification and data mining is that within classification human choices take place a priori and within data mining computer discovery emerges from the data. When new information technologies diffuse from the sphere of small groups of experts across institutions of origin, such as computing science, new organizational practices emerge. Genomics, computer-based algorithmic trading programs, search engines, and Billy Beane baseball all utilize data-driven methods of knowledge production to reach institutional goals.

When actors collect and organize stimuli in a database, they compress and instrumentalize the stimuli for specific institutional and organizational goals. Geof Bowker points out that in scientific work the database is not merely a means to an end in the process of knowledge production, "the database is the object, an end in itself" (Bowker, 2006, p. 199; see also Hine, 2006). Although one can see the explicit social and political processes involved in making a database, in terms of who owns it and what data is included and excluded, we want to highlight the implicit and tacit processes behind digital code. Decoding and encoding processes are often hidden from public view and sometimes even from the view of coders themselves. Yet actors make decisions about what constitutes different types of data, create algorithms for analyzing it, and attach discourses and interpretations to both the process design and the results. We only see the output of the decoding and encoding of data at the interface of computer monitors, social media sites, or in the pages of scientific research publications. In this context, actors analyze, interpret, and structure the data using words, symbols, metaphors, and programming code. This communicative process is inherently shaped by the cultural and social context, and thus potentially rife with both politics and bias. What happens when actors compile, manage, and analyze data in digital databases and recode as knowledge? Who did the coding? Why did they do it? How and where did the coding take place? What role does data mining play in the new economy, scientific innovation, knowledge production, and decision-making?

These research questions are broad, and we do not intend to answer all of them in the limited space of a journal article. Consequently, we begin our exploration by examining data mining in the context of genomics, which is rapidly becoming a very important area of algorithmic innovation as well as what Nikolas Rose (2007) refers to as a key site of the politics of life itself. The birth of genome science is largely due to the convergence of data mining technologies and biology (Chow-White & Garcia-Sancho, 2012). The rising use and importance of data mining technologies in genomics raises important issues and concerns with regard to race, class, privacy, and the future of health care.

## Data Mining Difference in Genome Research: The Persistence of the Use of Race in Science

Human genomics is an important enterprise in the information age. Scientists, policy makers, and entrepreneurs make enormous promises about the health and social benefits of genome science and related DNA technologies (see Collins, Green, Guttmacher, & Guyer, 2003). The rewards of this field range

from the evolutionary human histories contained in our DNA to the genetic origins of complex diseases to personalized health profiles. Other scientists, social scientists, and community advocacy groups argue there are significant risks as well, such as ignoring the environmental causes of disease, health disparities, and reproducing racist biological notions of human difference.

For some scholars, the initial completion of the Human Genome Project in 2001 heralded the end of bio-race as a legitimate concept when project leaders declared humanity similar at the molecular level (Gilroy, 2000). Many hoped the HGP would provide conclusive evidence to refute the "race as biology" thesis and thus support the social constructionist position put forth in the UNESCO Statements on Race 50 years earlier. The lead scientists on both sides of the HGP public and private teams, Francis Collins and Craig Venter, as well as President Bill Clinton, cited a new mantra at a historic press conference in 2000 that was drawn from the data, code, and digital sequences: "We are all 99.9% the same at the DNA level."

The HGP and genome-sequencing technology did not close the door on notions of bio-race and the validity of race as a variable in scientific research, as some predicted (e.g., Gilroy, 2000). Instead, racial population difference became a key variable in studying the existence and meaning of difference and variation at the genetic level. Perhaps scholars should have expected this considering science's long entanglement with racial classification. However, the persistent use of race did not take place without controversy. Many scholars across the academy argued vigorously to abandon race in genomic and genetic research (e.g., Bliss, 2012; Schwartz, 2001). For example, when organizers of the next large genomic sequencing project, the International HapMap Project, chose sample groups from Asian, African, and European origin in 2001 planning meetings, social scientists and natural scientists together debated different ways to name the groups to avoid the obvious connections to racial classifications (Chow-White, 2008). They attempted to color blind the database by choosing names that indicated geographic origins and opted for the term *populations* to refer to the groups. Major scientific publications such as *Science* reported on the new project and referred to the groups as populations (Couzin, 2002). However, by 2007, two years after the completion of the first phase of the HapMap project, the publication *Science* returned to referring to the HapMap samples as "races" (Couzin, 2007). The earlier racial labels and discourses those scientists had attempted to "bury alive" (Duster, 2003) post-HGP continued to proliferate. *Why and how did these notions or discourses about racial realism grow and persist so vigorously after the HGP?* As opposed to the end of race as a biological concept in the sciences, the completion of the HGP now appears to mark the beginning of a new phase of biopolitics for racialized bodies. Although the debates of race and biology have a long history in modern science, the rise of information and digital technologies in general, and use of databases and data mining in particular, has brought both renewed promise and problems to these issues.

## Data and Methods

In order to understand how discourses about race and genomics coevolved with the use and meaning of genome technology and data mining during the 1990s and 2000s, we conducted a content analysis of scientific discourse on race in genomics in major scientific journals. To collect discourse about genomics in science journals, we gathered articles on genomics and population studies from a robust and

representative set of scientific journals. We selected these journals because genomic experts we interviewed identified these journals as important to the field. In addition, these journals also ranked among the highest for Journal Citations Reports impact factors. We collected 464 general science, genetic, and biomedical journal articles (see Table 1) from 1998–2007. We chose this time period because the interviewees indicated that it encompassed important and relevant events (e.g., the lead up to and completion of the HGP) within genomic science. We built a Web-based search engine and crawler to identify and download the electronic articles directly from the journal websites using a Boolean search string that included the words *genom* AND human AND (population OR ancestr* OR continent OR geograph*) OR (ethnic* OR Africa OR Asia OR Europe* OR Hispanic)*[i]. We cleaned the dataset by hand in order to include relevant articles and remove irrelevant articles returned from the search. For example, our cleaning process removed articles focusing on animals or plants, gene patenting, gender, and some technical sequencing articles. In order to capture the breadth and totality of discourse across the journals, we included a broad and diverse set of articles, including research reports, commentary, letters, editorials, and news reporting.

***Table 1. Sampled Journals by Impact Factor.***

| Sample sets (N=8) | Journals sampled (Abbreviation; 2007 JCR impact factor) | Proportion of total sample |
|---|---|---|
| General science | *Nature* (Nat; 28.751) | 7% |
| | *Science* (Sci; 26.372) | 9% |
| Genetic | *Nature Genetics* (NG; 25.556) | 32% |
| | *Genome Research* (GR; 11.224) | 13% |
| | *Genome Biology* (GB; 6.589) | 6% |
| Clinical | *New England Journal of Medicine* (NEJM; 52.589) | 20% |
| | *Journal of the American Medical Association* (JAMA; 25.547) | 5% |
| | *British Medical Journal* (BMJ; 9.723) | 8% |

In our discussion of the discourse in the journals, we supplemented the content analysis data with interview findings from a genome-research project that included 114 stakeholders from across the academy, biotechnology companies, NGOs, and policy circles in various settings such as offices, conferences, meetings, and workshops from 2005 to 2009. Our interviewees were evenly distributed across the natural and social sciences. The transcripts from these interviews provided insights into the way actors used discourse to manage the risk of using or not using race in genomic science. The interviews led us to the content categories as well as directed us to investigate the trends in racial discourse in the sciences. For example, a number of interviewees claimed to notice an increase of racial-

comparison studies in the bio-medical and scientific journals during the 2000s, such as this geneticist and director of a bioethics department:

> When I look and do searches on different types of studies by going through PubMed or something there seems to be this overwhelming, tri-group [i.e., African, Asian, and European] comparison that happens over and over again. (Interview 1022)

A molecular biologist and director of a major genomics research center observed,

> [I]f you look at the construct of the literature on population studies and, especially, of disease in populations you'll find black/white studies all over the place. (Interview 1006)

We include some interview findings to support and aid in illustrating the core concepts and trends in the racial discourse. The selected interview data also shows a concordance between the highly structured journal discourse and the more informal reporting of experiences, values, and opinions of stakeholders. In short, the interviews helped us validate and check our measures and interpretation of our findings. Ideally, we would prefer to include an extended analysis of the interview data and comparison with the journal data. However, we will have to pursue this on another occasion due to the space constraints of a journal article and the primary goals of our study.

## Content Analysis of Genome Discourse in Science Journals

In this section, we report on the content analysis of the aforementioned articles from scientific journals. In our content analysis, we focused on examining the role of racial discourse in shaping data mining technologies in genomic science. Building on previous scholarly research and insights contained within the interviews, we used the two main coding categories: racial realism (RR) and social constructionism (SC). We completed the coding in two steps. First, a team of six researchers, including the authors, defined an initial set of codes developed from critical race theory, interview findings, and relevant scholarly literature (see Table 2). We then used an iterative process to develop and refine the codes over a four-month period. During this time, we created, tested, revised, and refined the coding protocol by going back and forth between the data and the coding categories. Once we agreed that the codes effectively captured the features of the data, we created a codebook that outlined the coding procedures, rules, definitions, and examples (Capella et al., 2009; Krippendorff, 2004). Stakeholders who take an RR position tend to believe that genome variation can reveal racial differences. A typical RR statement in the journal discourse states, "Based on the average pigmentation difference between European-Americans and African-Americans of about 30 melanin units, our results suggest that SLC24A5 explains between 25 and 38% of the European-African difference in skin melanin index" (Lamason et al., 2005, pp. 1785–1786). In the interview data, a molecular biologist describes the validity of race as a category in genome research:

> I think it [race] is useful as a proxy for population in some cases. It is true particularly in some cases that the race reflects its phenomenal or genetic traits well . . . I basically agree that there is no biological basis for racial categories, but simultaneously still

> believe in some cases racial categories are useful . . . Racial groups also reflect some cultural or historical traits, not only genetic or phenomenal traits. (Interview 1018)

Stakeholders who take an SC position tend to believe that genome variation does not reveal racial differences. A typical SC statement in the journal discourse claims:

> Knowledge from the Human Genome Project and research on human genome variation increasingly challenges the applicability of the term "race" to human population groups, raising questions about the validity of inferences made about "race" in the biomedical and scientific literature. (Royal and Dunston, 2004, p. S5)

Stakeholders who argue for a social constructionist approach to genomics often refer to the findings of the HGP to support their position. One interviewee, a world-renowned geneticist at a major U.S. public research center, sums up the apparent conclusion of the HGP at the beginning of the 21st century, "[T]he Human Genome Project demonstrated that there is no such thing as race. I cringe when people say that the HGP did not do that [show that there are no racial differences]" (Interview 1013). Another geneticist suggests that race-based research is unnecessary:

> If you look at many manuscripts that have been published using race, quote un-quote, as a descriptor many of them don't even need to use race. They don't even need to look at specific populations. They are things that you can just look at, depending on what you are trying to find. And I think that many people just see that there is this need to divide and separate between these macro groups and not look at the diversity within some of those groups. And I think that is a problem that we have. There's a failure to recognize the diversity within the group as well as overlap among those groups. (Interview 1022)

Further, supporters of social constructionist approaches argue race has no validity as a category in scientific research.

Considering the nature of the data, we carefully trained the coders to develop a working understanding of the relevant scientific terminologies in the research articles in order to identify accurately the coding categories in the text. We spent more time on training than is normal for content analysis in order to make sure that coders were comfortable with this specialized terminology. We statistically sampled a subset of the 464 articles, and then tested coder inter-reliability on 30% of the articles. We found a rate of 91.4%. We sampled half of the 464 articles contained in the data set by selecting every other article beginning with the second article. We choose this starting point by randomly selecting between the first and second articles published in 1998, the first year of the time period. We then analyzed the resulting data set.

***Table 2. Discourse and the Social Shaping of Genome Technologies.***

| Racial Discourse in Genomics | Valence | Code |
|---|---|---|
| Racial Realism (RR)<br>Genome Variation = Race | Critique SC | cSC |
| | Support RR | sRR |
| Social Constructionism (SC)<br>Genome Variation ≠ Race | Critique RR | cRR |
| | Support SC | sSC |

After coding the 232 articles, we constructed a set of graphs to represent the trends in discourse over time. Our graphical analysis shows how RR and SC arguments changed over time in the scientific journal discourse. This coding supports and extends qualitative observations from our interview data. The results suggest important patterns and variations in how racial discourse formed and evolved. Figure 1 shows the total number of racial realist- and social constructionist-based articles for each year from 1998 to 2007. This allows comparison of each of the two types of racial discourse over time. Graphing the trends of these two types of racial discourse helped identify changes across racial framing in genomic discourse. Although RR discourse increased over the entire time period, RR discourse declined significantly in 2001 relative to SC discourse. In 2001, SC discourse increased relative to RR and became the dominant discourse. However, from 1998 to 2007 RR discourse increased relative to SC discourse from about 70% to about 90%. Perhaps the most striking and disconcerting trend is in the latter part of the time period where RR articles increase at an increasing rate and SC articles decline.

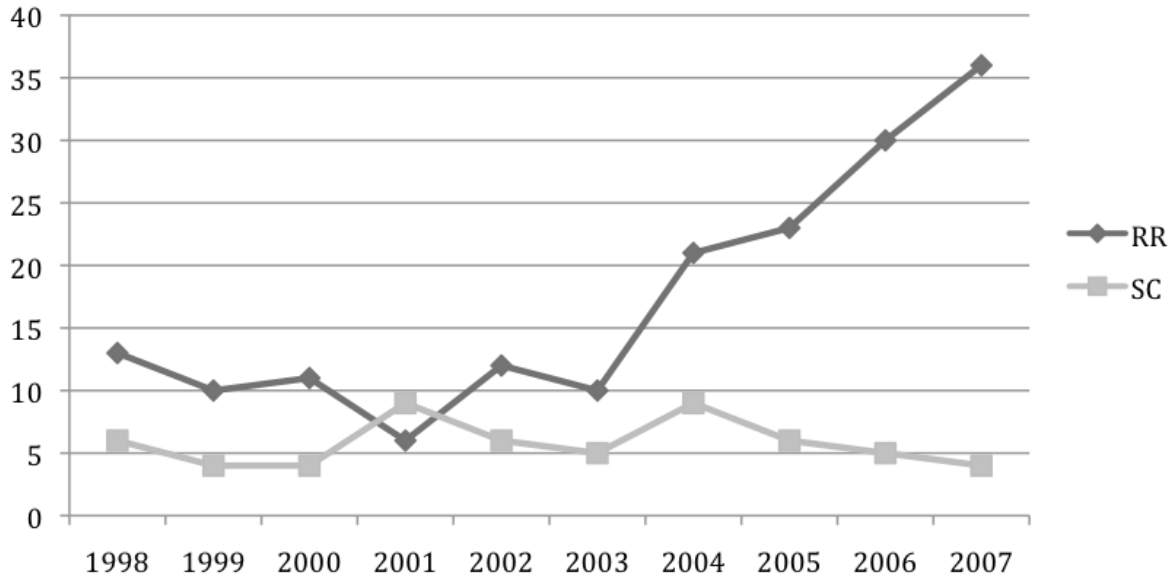


***Figure 1. RR and SC Articles per year, 1998–2007.***

Most years there are more RR articles relative to SC articles in the research than there are RR articles relative to SC articles in the commentaries. This is important because the overall number of commentaries decreased relative to the number of research articles over the time period. One could argue that the change in RR is a result of a change in the number of commentaries relative to research articles at the journals. However, we also found that the percentage of RR articles relative to SC articles increases over our study period in both the commentaries as well as the research articles.

To analyze further the variation of racial discourse in genomic research we also coded for the various arguments or instances of valence within each article. Actors using competing discourses within a social field may decrease the perceived risk of following their own logic by talking about the net benefits of their logic versus the net benefits of the competing logic (Green, Babb, & Alpaslan, 2008). We coded how arguments for RR and SC positions deployed racial discourse about the risks and rewards of believing RR versus believing SC. The communicative task for the RR approach is to decrease the costs and consequences of their approach by supporting the RR logic and extolling the benefits of this approach (i.e., coded as sRR) or increase the perceived costs and risks of an SC approach by critiquing SC (i.e., coded as cSC). Actors making sRR statements might suggest that humankind will benefit from an RR approach because this approach will increase our understanding of the genetic origins of a disease. For example, a typical sRR statement describing the benefit of using race as variable suggests that the use of racial variables *helps to guide the design and analysis of medical genetic studies* or *offers an important opportunity to explore and understand the evolutionary forces that shape variation in natural populations*.

Another example from the data uses geography as a proxy for racial groups in a global comparison of brain sizes:

> The gene Microcephalin (MCPH1) regulates brain size and has evolved under strong positive selection in the human evolutionary lineage. . . . We genotyped the diagnostic G37995C SNP in this panel to infer the frequency of haplogroup D chromosomes. Geographic variation was observed, with sub-Saharan populations generally having lower frequencies than others. (Evans et al., 2005, p. 1717)

An RR critique of SC (coded as cSC) might argue that SC is risky because ignoring racial structure is an attempt at political correctness and that this stifles genome research, the purity of science, and the search for truth. For example, the authors in this article from the *New England Journal of Medicine* argue against the SC position by deemphasizing the costs of using race as a variable:

> Although there are potential social costs associated with linking race or ethnic background with genetics, we believe that these potential costs are outweighed by the benefits in terms of diagnosis and research. Ignoring racial and ethnic differences in medicine and biomedical research will not make them disappear. Rather than ignoring these differences, scientists should continue to use them as starting points for further research. (Burchard et al., 2003, p. 1174)

One of the most common ways that scientists critique the social constructionist position is to claim that the racial realist logic is based on objective science and empirical data, whereas social constructionist logic is ideological and political in nature. In an interview, a biologist argues that

> [t]he data on difference between races is only beginning to emerge. It is starting to demonstrate that gene frequencies based on random genetic markers have a very strong correlation with geographic origin or what we call races or ethnicities. If we follow the ideology of political correctness, rather than an objective approach to science, we will miss the scientific reality of group differences. (Interview 1063)

Scientists such as this biologist argue there is a strong genetic component to race. While many interviewees agreed that race is largely social, some scientists critiqued the social construction of race as a political project and argue that empirical comparison studies of genomes would ultimately reveal the reality of race.

The communicative tasks for an SC approach are similar. SC discourse attempts to decrease the perceived cost and consequences of a SC approach by lowering the risk of doubting (coded as sSC). An sSC statement might argue that humankind will benefit from an SC approach by emphasizing genetic similarities in order to avoid reproducing racist biological notions of race. For example, "Instead of using polymorphisms to seek racial distinctions, we can spark real progress in clinical research by using genetic variations to track down clinically relevant alleles and pathogenic mutations" (Schwartz, 2001, p.1392). In this example from the journal *Science*, the authors argue that scientists will find most genetic differences

within groups rather than between them:

> The average proportion of genetic differences between individuals from different human populations only slightly exceeds that between unrelated individuals from a single population. That is, the within-population component of genetic variation, estimated here as 93 to 95%, accounts for most of human genetic diversity. Perhaps as a result of differences in sampling schemes, our estimate is higher than previous estimates from studies of comparable geographic coverage, one of which also used microsatellite markers. This overall similarity of human populations is also evident in the geographically widespread nature of most alleles. (Rosenberg et al., 2002, p. 2381)

SC discourse also attempts to increase the perceived costs of an RR approach by raising the risks for scientists believing that they can use the genome to identify racial differences (i.e., coded as cRR). A cRR statement might argue that RR is risky because race is a spurious proxy for gross human groups or that racial classification reproduces racial ideology and racist practices. For example:

> Research to root out social injustice in medical practice needs continued support, but tax supported trolling of databases to find racial distinctions in human biology must end. . . . It will be difficult to abandon long-held preconceptions, but perhaps the first benefit of the Human Genome Project will be to lead us to the understanding that in medicine, there is only one race—the human race. (Schwartz, 2001, p. 1393)

In this example from the leading journal *Science*, the author reports on attempts to challenge the findings of the Evans study mentioned above. Scientist Bruce T. Lahn led the research that resulted in the Evans 2005 publication.

> And in October, a team led by geneticist David Reich of the Broad Institute in Cambridge, Massachusetts, reported at the meeting of the American Society of Human Genetics that it found no evidence for recent selection on *ASPM* when it used a method of analysis it considered superior to Lahn's. But Lahn, who is familiar with Reich's results, stands by his conclusions: "Their method has lower resolution . . . and is less reliable," he says. (Balter, 2006, p. 1872)

Many of the interviewees made statements about race being a social and political category rather than a biological one and argued that race has no place in genomics, such as this geneticist:

> Now that we have a genome with markers, where we can look at markers to find genes, we have a much more refined tool to try to track biology than the crude gross level of race. . . . And so the question becomes, in a particular individual, which track of which system has there been a change in that has thrown the system out of balance in a way that leads to disease . . . now the genomics in particular has pushed us to a point in biology where we need to dissociate equating or using race as a surrogate for biology. In other words there is no construct of the genome where you're going to find this

> construct or this gene in every person we call black, and not find it in people we call white. (Interview 1006)

However, a legal scholar and bioethics expert explains that using race in biomedical research is not the same as other types of color conscious programs such as affirmative action:

> Using race in medicine is more complicated than other discussions. There is more slippage into bio-determinism. Biology and race can get conflated much easier. Seems natural to use race as a category in genomic research as it pertains to health. But doctors are trained to look at biology. They don't look at race like social scientists, who understand race in much more complex ways. The doctors and scientists think genetics is difficult and race is easy. For them, race is obvious, not complicated. They say, I am just trying to save lives and the problems are the result of misinterpreted data. But this opens the door to characterizing blacks, for example, as different. (Interview 1112)

For this stakeholder, race enables a focus on health disparities and is potentially productive for genome research, but it also constrains alternative ways of thinking about human differences socially and biologically. Our coding protocol allows for the presence of any of the four valence codes in any single article. That is, an RR or SC article can contain arguments in support (sRR or sSC) or critique (cRR or cSC) of its own position.

Figure 2 shows the overall valence proportions over the time period and reveals how the discursive balance changed over time. Arguments for RR (sRR) are dominant over the time period. However, after the completion of the first draft of the HGP in 2000, SC arguments in support of SC (sSC) and against RR (cRR) increase and peak in 2001. They also shift in proportion from more cRR and less sSC to the reverse, more sSC and less cRR. At the same time, arguments against SC (cSC) appear. During the latter part of the period, especially following the publication of data from the International HapMap Project in 2005 that compares variation between populations from Africa, Asia, and Europe, the SC valence proportions shift back to more cRR and less sSC, and the overall proportion of SC valence diminishes. Arguments against SC are absent by the end of the time period, and RR discourse increases at an increasing rate and dominates by the end of the time period, peaking at almost 90%.

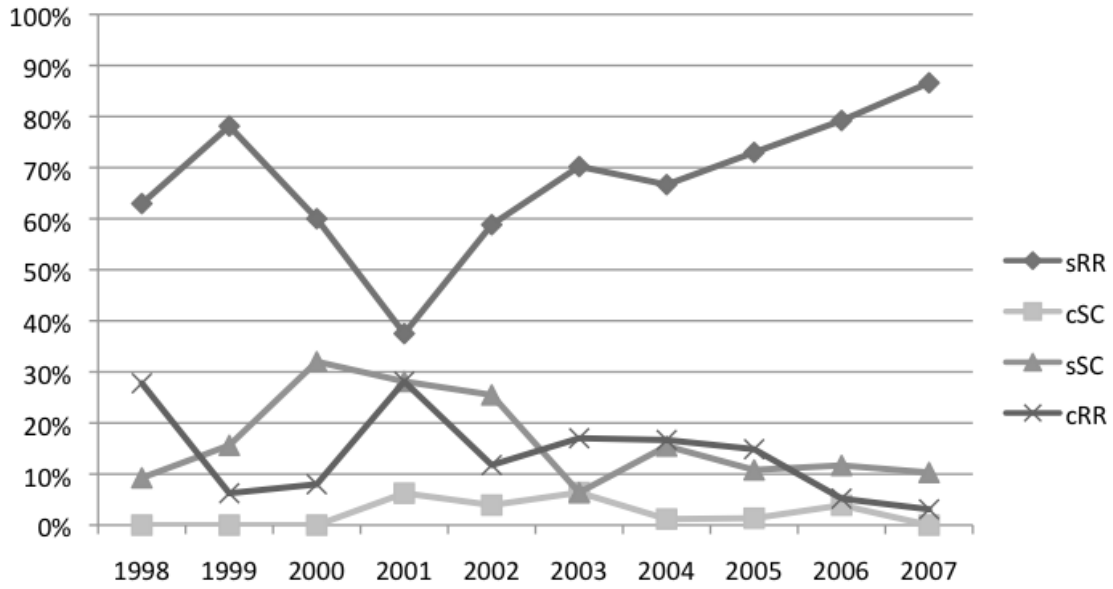


***Figure 2. Valence proportions by year.***

In order to understand how different actors used discourse to shape the use and meaning of genome technology, we created graphical representations of valence profiles between articles coded as RR and SC. We coded articles as *predominantly* RR or SC. Actors may support RR using one sided or dogmatic arguments. They can also admit that the other side raises important points worthy of consideration. Often, RR articles paid no attention to an SC perspective, and only sometimes RR articles were more nuanced or open in their discussion and argumentation. Authors supporting the SC perspective admitted merit to the other side more often than RR authors admitted merit to SC perspective. SC authors also sometimes mentioned possible shortcomings of their own approach. Figure 3 shows the valence profiles for RR articles and Figure 4 shows the valence profiles for SC articles. In Figure 3, we see proponents of RR logic focused on sRR or the rewards of an RR approach. The proportion of sRR arguments decreased from 1998 to 2001 and then increased from 2001 to 2007.

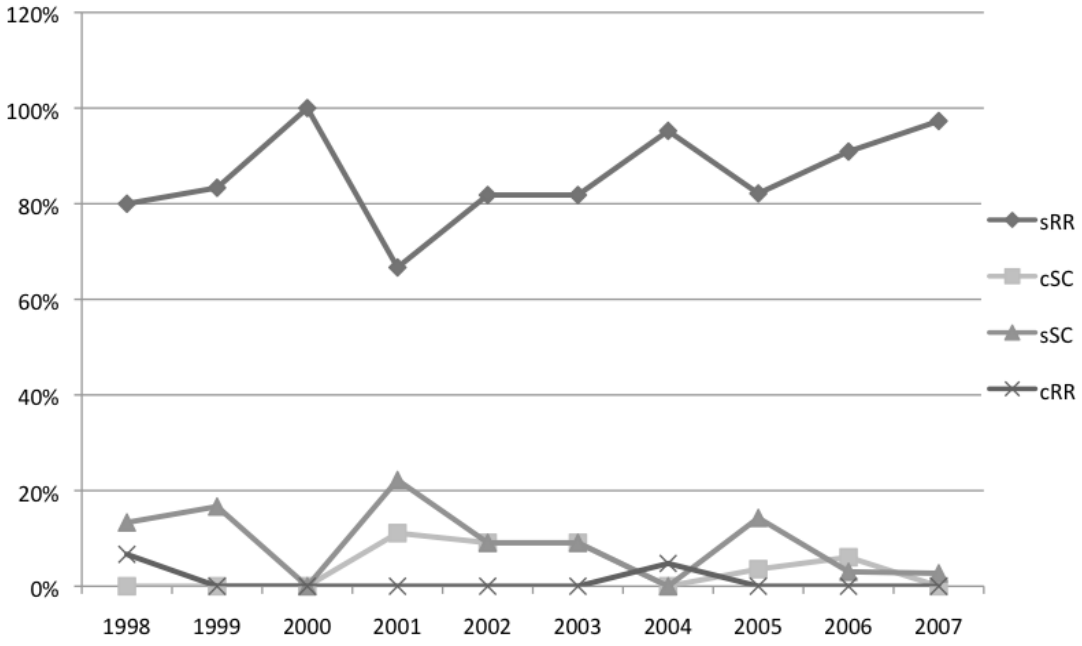


***Figure 3. RR valence profile by year (%).***

Although there were a few RR criticisms of SC from 2001 to 2006, in most years these criticisms were almost nonexistent. The sRR discourse or arguments for the rewards or value of an RR position dominated from 1998 to 2001. However, after the HGP, RR discourse stating the rewards of the RR perspective declined as racial discourse by actors in critique of SC increased. In sum, advocates of the RR logic seem to respond to the HGP by attacking SC with an increase in cSC. In contrast to RR discourse, we found a much more complex set of relations in the valence profile of SC discourse from 1998 to 2007 (see Figure 4). SC discourse appears to gain ground when cRR and sSC are high. We found that SC discourse declined when SC discourse shifted from arguing for the value or rewards of SC (sSC) to emphasizing the risks of RR (cRR). At the same time, SC articles increasingly acknowledged the value of the racial realist approach and even included criticism of its own approach.

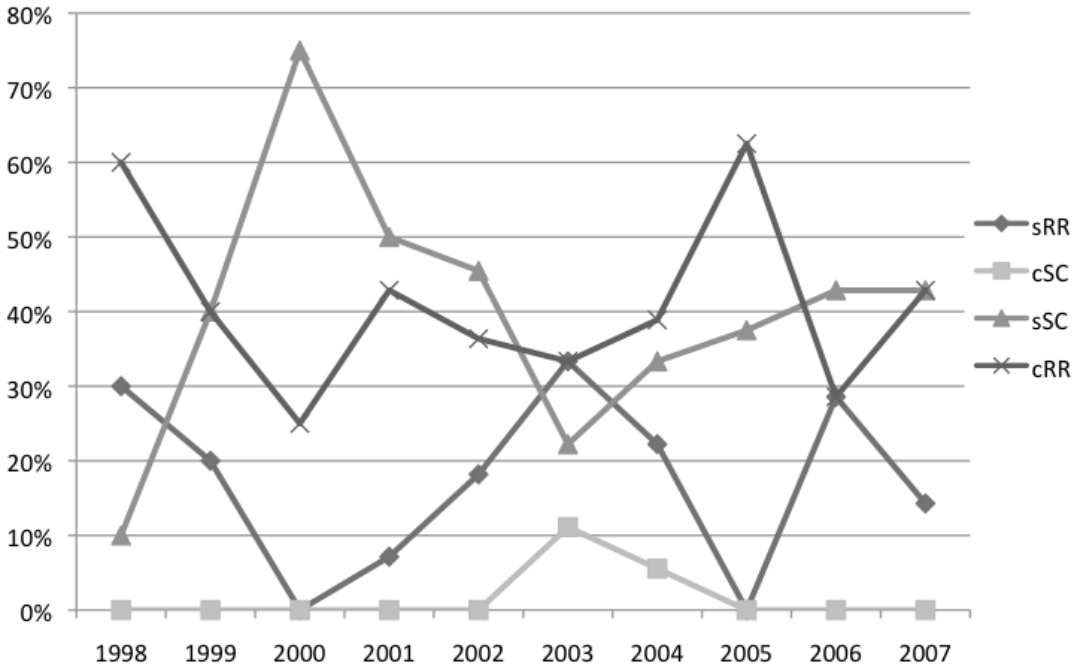


***Figure 4. SC valence profiles by year (%).***

The content analysis, supported by interview data, provides a longitudinal picture of racial formation in the development of genome technologies during the last decade. The interaction and tensions between racial realist perspectives and social constructionist perspectives shaped the development of data-driven genome population technologies in the 2000s and the encoding of scientific truths. Our detailed empirical data suggests that the assumptions and communicative practices of scientists shaped how they practiced genomic science with regard to race. Creating genetic samples linked to race is highly problematic.

In our discussions and interviews with scientists in the field about these problems, several respondents mentioned the theoretical and practical complexity of linking genetic samples to particular races. For example, one bioethics expert we interviewed described how scientists impose race on DNA in the labs:

> There is a racially specific imposition of race on DNA when it is labeled in the labs. Genome variants take on racialized identities even though most people share the variants. Bodies are parsed up, but they don't use race directly; they use ancestry. (Interview 1067)

Other interviewees told us how some researchers use direct observation or opinion experts to label biological samples from individuals as belonging to particular racial groups. Yet many researchers attested to the difficulty of actually predicting genetic patterns from commonsense notions of racial identity and visual perceptions of phenotype identifiers, such as skin color or shape of eyes, or symbolic representations, such as clothes or mannerisms. Alternatively, some respondents discussed how some biological samples that arrive at their labs pre-labeled by other labs or from an online public genome research database. This puts the researcher in a challenging situation that does not present an easy solution. Does she accept the existing labels even though she may find them inaccurate, or does she impose her own labels, which may misrepresent and alter the nature of the relationships of the people included in the data set?

Another option raised by our respondents describes how many researchers have the people giving the biological sample self-report their or their parents' race. The self-reporting of race is entangled with cultural notions of identity for all groups and especially so for those of mixed-race heritage. Techniques by the researcher to identify mixed samples by asking for reports of parents and grandparents cannot fully solve these problems. For example, imagine a self-report of someone who says they are black, yet their mother is white and their father is black. The researcher may throw this data out of his sample. However how does one identify a mixed-race sample at the level of grandparents, or great grandparents?

Finally, another option is to first find genetic markers and then link them to race. For example, some researchers link genetic markers to geography and then try to link geography to race. This requires researchers to make certain assumptions about migration patterns and spatial-demographic trends, such as colonialism and globalization. However, this technique is highly problematic because the researcher never establishes in any satisfactory a priori way the race of the person with the geographic sample or genetic marker. Conflating race with geography does not resolve these issues. All of these techniques for identifying racial identity leave the researcher theorizing why these genetic markers belong to some theoretical racial group, or identifying the racial group of a biological sample and describing the genetic markers from this sample.

Taken as a whole, our interviews suggest that scientists' core assumptions about whether or not race is a useful proxy and research design for populations or human groups inherently drives the process of racial identification. Some scientists assumed a racial realist route where they viewed race as a useful proxy. Other scientists assumed a social constructionist position and viewed race as a poor and sometimes dangerous proxy. These core assumptions about race shaped how scientists went about collecting, measuring, analyzing, and interpreting genomic data in genomic databases. For example, one common step in the creation of race-specific genomic data often involve identifying individuals who fit a predetermined racial group and collecting biological samples via a swab of saliva or vile of blood and classifying the sample according to often strict U.S. census racial categories. Scientists working from a racial realist or constructionist view of race treat samples from racial categories not recognized by the U.S. census (.e.g., multiracial groups) differently. Scientists sequence DNA from the sample, turning human matter into 1s and 0s. The entry of the resulting computer code into a digital database compresses both complex biological stimuli and social relations into digital data. For example, in ethnographic studies of

two labs in the San Francisco Bay area, Fullwiley examined how scientists' biological notions of genetic differences between racial groups shaped their work. Fullwiley found that these lab scientists often applied commonsense notions of race to scientific discourses of populations, geography, and genetic variation (Fullwiley, 2007; see also Bliss, 2012; Bolnick, 2008; Fujimura and Rajagopalan, 2011).

In short, assumptions of racial realism and social constructionism appear to affect many aspects of the decoding and encoding process. For instance, scientists structure the identity of the data in the databases in the gathering-and-classifying stage of research. The decision to group racial data into three large samples of Asian, African, and European requires a complex social calculus for groups at the boundaries. It also requires making assumptions about where the centers of racial identities are actually located, as well as assumptions about why these groupings and not others are preferable. Finally, scientists data mine the digital DNA by running a series of statistical routines on millions or billions of data points that compare different sets of DNA and look for points of variation between the group data sets. They interpret the variation along the lines of ancestry, race, or whatever population variable they use, label it, and represent and recode the DNA with racial difference, thus encoding scientific facts. However, the assumptions of the scientist collecting, organizing, and mining the database often place the variation in the construction of the group itself.

## Conclusion

In the context of genome science, scientists in the 1990s and 2000s developed data mining as a new way of seeing the human body and the molecular DNA network. Our study provides insight into the role of communication in the shaping of data-driven genome technologies and the meaning of race. Despite discourses of colorblind racism and a postracial society in the public sphere, we found that the post-HGP promise of the end of biological notions of race did not pan out. Instead, we found a striking trend back toward racial realism in the social shaping of genome technologies. This upward trend seemed to coincide with the publication of the first round of global genome variation data comparing African, Asian, and European populations from the International HapMap Project, a multinational consortium.

We view this trend as more than simply a function of the legacy of race and science. We see this trend as an effect of the interaction between race, communication, data mining technologies, and knowledge production in the age of big data (see also Nakamura & Chow-White, 2011). Actors across institutions and enterprises utilize a data lens to see their operations and strategies more clearly. We believe this is happening across fields as the dissemination of data mining technologies quickens in pace. Early Internet technologies increased the amount of information available online, and new data mining technologies, such as genome technologies, aim to compress the social and biological world into data. The former produces more information and overwhelms us, and the latter is supposed to organize information, analyze it, and increase certainty and clarity in a value neutral way. There is a difference between overload and reducing the uncertainty of the overload. Scholars have theorized the former, but grounded theoretical work on the latter is still emerging.

As we noted at the beginning of this article, scientists who utilize data mining technologies believe that massive amounts of data can help them see the world more clearly: "The more data you

have, the clearer you see." For centuries, scientists, politicians, artists, and religious figures among others have commonly employed the metaphor of sight for knowledge. They are not talking about visual seeing, but about gaining a clearer understanding of a concept or phenomenon. In the age of big data, the IBM scientists argue, knowledge production through the practice of collecting and mining data creates a clearer view of the social, material, and digital world. On the face of it, this seems counterintuitive considering that a defining feature of the network society is information overload, where information and communication technology inundates us with massive amounts of data and information on a seemingly daily basis. However, in the scientists' estimation, the more data one can gather and analyze in real time, the better decisions actors can make within their institutional or organizational context.

Data mining enables actors to increase precision in answering questions. In a data-driven society, this claim to accuracy seems reasonable. Data mining may provide better and more valid answers and decrease the risk of false positive (Type I) and false negative (Type II) errors. However, it is less reasonable to assume that a data-driven approach decreases Type III errors: the probability of answering the wrong question correctly with the prevailing methods. Focusing the lens of our perception using data mining technologies may increase the clarity of our view of the world; however, this does not mean we are looking in the right direction or at the right things. While increasing data may include more points for analysis, the human decision-making behind the collection and organization of data frames what interpretations and knowledge is possible as we code from stimuli to data. We suggest that the technical wizardry and great promise offered to society by the construction of racial databases for genomic research may possibly increase the development of Type III errors. As our study points out, the construction of race in genomic data is highly problematic at best; however, we risk ignoring important potential biases in these databases as the siren call of big data entices us with breathtaking discoveries and practical knowledge in human biology and biomedicine. In such a world, Type III errors might proliferate if genomic scientists fail to question the meaning, validity, and reliability of genomic databases of racial data constructed with flawed questions and conceptions of how race and genetics cohere and relate. Deploying data mining without reflexively recognizing the inherent persuasiveness of precision and its corresponding increase in Type III errors is extremely dangerous because data mining practices have the power to rationalize and naturalize bias and domination under the guise of truth and knowledge. Like any tool, the technology of data mining is neither inherently good nor bad. In practice, however, the use of tools or technology is rarely, if ever, neutral. We have an ethical responsibility to make sure we are aware of the risks of Type III errors in the age of big data.

Recent scholarship echoes our concerns about the potential deleterious effects of data mining on society. For example, Gandy (2009) suggests that these new forms of knowledge production are producing *rational discrimination*. Rational discrimination refers to the forms of knowledge informed by statistical techniques of data analysis that facilitate identification, classification, and comparative assessment of groups generated analytically in terms of their expected value or risk. Gandy argues that we are already born into social positions that play a strong role in shaping our life chances such as race, class, gender, and sexuality. At the same time, Gandy argues that the data mining technologies used by companies contribute to what he calls *cumulative disadvantage* that "reinforces and reproduces disparities in the quality of life" (Gandy, 2009, p. 55) when companies use this data to control and coordinate access

to goods, services, and incentives. Cumulative disadvantage, rational discrimination, and the rise of racial realism in genomics provide excellent examples of the costs of making Type III errors.

Our key concern with these new data mining classification systems and technologies for producing knowledge is that they tend to obfuscate the myriad of decisions that create them. Users tend to see only the interface, the front end of a particular technology. Hidden away inside computers and software are attitudes, values, and politics that actors write into the code and "arguments, decisions, uncertainties and processual nature of decision-making" (Bowker & Star, 1999, p. 187). A number of ethnographic studies show that this process unfolds at the micro level in lab settings, data mining software development, and major global initiatives (e.g., Bliss, 2012; Bolnick, 2008; Fujimura & Rajagopalan, 2011; Fullwiley, 2007; Reardon, 2005). Data mining practices exist across many institutional contexts, and genomics is only one of them. We can assume that the zeitgeist to mine data as expressed by the IBM scientists in the introduction will continue as data mining costs decrease and the perceived benefits increase. We believe these issues are taking place in similar ways elsewhere and hope that communication scholars, following Gandy and Turow, will pursue them in different institutional settings.

**References**

Anderson, C. (2008, January 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*.

Balter, Michael. (2006). Bruce Lahn profile: Links between brain genes, evolution, and cognition challenged. *Science, 314*(5807), 1872.

Bamshad, M., Wooding, S., Salisbury, B. A., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. *Nature Genetics Reviews, 5*(8), 598–609.

Bell, D. (1992). Racial realism. *Connecticut Law Review, 24*, 363.

Bliss, C. (2012). *Race decoded: The genomic fight for social justice*. Palo Alto, CA: Stanford University Press.

Bolnick, D. A. (2008). Individual ancestry inference and the reification of race as a biological phenomenon. In B. Koenig, S. Lee, & S. Richardson (Eds.), *Revisiting race in a genomic age* (pp. 70–88). New Brunswick, NJ: Rutgers University Press.

Bowker, G. (2006). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, G. (2010.) The archive. *Communication and Critical/Cultural Studies, 7*(2), 212–214.

Bowker, G., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Brown, C. (2011, June). The rise and rise of datanomics. Retrieved from http://www.cnbcmagazine.com/story/the-rise-and-rise-of-datanomics/1394/1

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J. et al. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine, 348*(12), 1170–1175.

Capella, J., Mittermaier, D. J., Weiner, J., Humphreys, L., Falcone, T., & Giorno, M. (2009). Coding instructions: An example. In K. Krippendorff & M. A. Bock (Eds.), *The Content Analysis Reader*, (pp. 253–265). Thousand Oaks, CA: SAGE Publications.

Castells, M. (2000). *The rise of the network society* (2nd ed.). Oxford, UK: Blackwell.

Chow-White, P. A. (2008). The informationalization of race: Communication technologies and the human genome in the digital age. *International Journal of Communication, 2*, 1168–1194.

Chow-White, P. A. (2009). Data, code, and discourses of difference in genomics. *Communication Theory, 19*(3), 219–247.

Chow-White, P. A., & Garcia-Sancho, M. (2012). Bi-directional shaping and spaces of convergence: Interactions between biology and computing from the first DNA sequencers to global genome databases. *Science, Technology, & Human Values*, *37*(1), 124–164.

Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. (2003). A vision for the future of genomics research: A blueprint for the genomic era. *Nature, 422*(24), 835–847.

Condit, C. (2008). Race and genetics from a modal materialist perspective. *Quarterly Journal of Speech, 94*(4), 383–406.

Couzin, J. (2002). New mapping project splits the community. *Science, 296*(5572), 1391–1393.

Couzin, J. (2007). In Asians and Whites, gene expression varies by race. *Science, 315*(5809), 173–174.

Digging into Data Challenge. (2011). Retrieved from http://www.diggingintodata.org

Dubriwny, T. N., Bates, B. R., & Bevan, J. L. (2004). Lay understandings of race: Cultural and genetic definitions. *Community Genetics, 7*(4), 185–195.

Duster, T. (2003). Buried alive: The concept of race in science. In A. Goodman, D. Heath, & M. S. Lindee (Eds.), *Genetic nature/culture: Anthropology and science beyond the two-culture divide*. Berkeley, CA: University of California Press.

Elmer, G. (2004). *Profiling machines: Mapping the personal information economy*. Cambridge, MA: MIT Press.

Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, J. R., Vaez-Azizi, L. M. et al. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science, 309*(5741), 1717–20.

Fujimura, J. H., & Rajagopalan, R. (2011). Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research. *Social Studies of Science, 41*(1), 5–30.

Fullwiley, D. (2007). The molecularization of race: Institutionalizing human difference in pharmacogenetics practice. *Science as Culture, 16*(1): 1–30.

Gandy, Jr., O. H. (1993). *The panoptic sort: A political economy of personal information*. Boulder, CO: Westview Press.

Gandy, Jr., O. H. (2009). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Burlington: VT: Ashgate.

Gilroy, P. (2000). *Against race: Imagining political culture beyond the color line*. Cambridge, MA: Belknap Press of Harvard University Press.

Gould, S. J. (1996, 1981). *The mismeasure of man*. New York, NY: Norton.

Green Jr., S. E. (2004). A rhetorical theory of diffusion. *The Academy of Management Review, 29*(4), 653–669.

Green, Jr., S. E., Babb, M., & Alpaslan, C. M. (2008). Institutional field dynamics and the competition between institutional logics: The role of rhetoric in the evolving control of the modern corporation. *Management Communication Quarterly, 22*(1), 40–73.

Green, S., Li, Y., & Nohria, N. 2009. Suspended in self-spun webs of significance: A rhetorical model of institutionalization and institutionally embedded agency. *Academy of Management Journal*, *52*(1), 11–36.

Hall, S. (1998). Subjects in history: Marking diasporic identities. In W. Lubiano (Ed.), *The house that race built* (pp. 289–299). New York, NY: Vintage.

Hayes, R. M. (1993). Measurement of information. *Information Processing & Management, 29*(1), 1–11.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science, 36*(2), 269–298.

Hood, L. (2001, September). Under biology's hood. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/biomedicine/12575

Kahn, J. (2006). Genes, race, and population: Avoiding a collision of categories. *American Journal of Health Policy*, *96*(11), 1965–70.

Khosrow-Pour, M. (2007). *Dictionary of information science and technology*. Hershey, PA: Idea Group Reference.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Kuper, S., & Syzmanski, S. (2009). *Soccernomics: Why England loses, why Germany and Brazil win, and why the U.S., Japan, Australia, Turkey—and even Iraq—are destined to become the kings of the world's most popular s*port. Philadelphia, PA: Nation Books.

Lamason, R., Mohideen, M. P. K, Mest, J. R., Wong, A. C., Norton, H. L, Aros, M.C. et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, *310*(5755), 1782–1786.

Lee, C. (2009). ''Race'' and ''ethnicity'' in biomedical research: How do scientists construct and explain differences in health? *Social Science & Medicine, 68*(8), 1183–1190.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. New York, NY: Norton.

Lynch, J. A. (2008). Geography, genealogy and genetics: Dialectical substance in newspaper coverage of research on race and genetics. *Western Journal of Communication, 72*(3), 259–279.

Lyon, D. (2002). Everyday surveillance: Personal data and social classifications. *Information, Communication & Society, 5*(2), 242–257.

Lyon, D. (2005). The sociology of information. In C. Calhoun, C. Rojec & B. Turner (Eds.), *The SAGE Handbook of Sociology* (pp. 223–236). London, UK: SAGE Publications..

MacKenzie, D., & Wajcman, J. (Eds.). (1985). *The social shaping of technology: How the refrigerator got its hum*. Milton Keyes, UK: Open University Press.

Manovich, L. (2001). *The language of new media*. Cambridge, MA: MIT Press.

McCann-Mortimer, P., Augoustinos, M., & LeCouteur, A. (2004). 'Race' and the human genome project: Constructions of scientific legitimacy. *Discourse & Society*. *15*(4), 409–432.

Nakamura, L., & Chow-White, P. A. (2011). *Race after the Internet*. New York, NY: Routledge.

Omi, M., & Winant, H. (1994). *Racial formation in the United States: From the 1960s to the 1990s* (2nd ed.). New York, NY: Routledge.

Parrott, R., Silk, K., Weiner, J., Condit, C., Harris, T., & Bernhardt, J. (2004). Driving lay models of uncertainty about genes' role in illness causation to guide communication about human genetics. *Journal of Communication, 51*(1), 105–122.

Rabinow, P. (1992). Artificiality and enlightenment: From sociobiology to biosociality. In J. Crary & S. Kwinter, (Eds.), *Incorporations* (pp. 234-252). New York, NY: Zone Books.

Reardon, J. (2005). *Race to the finish: Identity and governance in the age of genomics*. Princeton, NJ: Princeton University Press.

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: Genes, race, and disease. *Genome Biology, 3*(7), comment2007.01–2007.12.

Rose, N. (2007). *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century*. Princeton, NJ: Princeton University Press.

Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., et al. (2002). Genetic structure of human populations. *Science* (298), 2381–2385.

Royal, C. D. M., & Dunston, G. (2004). Changing the paradigm from "race" to human genome variation. *Nature Genetics*, *36*(11), S5–S7.

Sankar, P., Cho, M. K., & Mountain, J. (2007). Race and ethnicity in genetic research. *American Journal of Medical Genetics Part A, 143A*(9), 961–970.

Schwartz, R. (2001). Racial profiling in medical research. *New England Journal of Medicine*, *344*(18), 1392–1393.

Terranova, T. (2004). *Network culture: Politics for the information age*. London, UK: Pluto Press.

Thacker, E. (2005). *The global genome: Biotechnology, politics, and culture*. Cambridge, MA: MIT Press.

The data deluge. (2010, February 25). *The Economist.* Retrieved from http://www.economist.com/node/15579717

Turow, J. (2006). *Niche envy: Marketing discrimination in the digital age*. Cam bridge, MA: MIT Press.

Wailoo, K., Nelson, A., & Lee, C. (2012). *Genetics and the unsettled past: The collision between DNA, race, and history.* New Brunswick, NJ: Rutgers University Press.

Zarsky, T. Z. (2003). "Mine your own business!": Making the case for the implications of the data mining of personal information in the forum of public opinion. *Yale Journal of Law & Technology*, *5*(4), 1–56.

***Appendix A. List of Interviews.***

| **Interview Code** | **Description** | **Interview Code** | **Description** |
|---|---|---|---|
| 1006 | Geneticist | 1063 | Biologist |
| 1013 | Geneticist | 1067 | Bioethics expert |
| 1018 | Biologist | 1112 | Legal scholar, bioethics expert |
| 1022 | Geneticist, bioethics expert | | |

[i] The articles we used in our sample utilize or discuss biological data that is either directly or indirectly data mined by genomic scientists. For example, when the human genome is sequenced from a blood sample, the sequence analysis is a type of string or pattern mining specific to biological strings. According to genome scientists we interviewed and the literature, although the phrase *data mining* is often missing in the sequence-analysis literature, its basic concepts are normally implied (e.g., Gaber, 2009, p. 207).