Connecticut Insurance Law Journal   School of Law

2014

Risk Classification's Big Data (R)evolution

Rick Swedloff

RISK CLASSIFICATION'S BIG DATA (R)EVOLUTION

RICK SWEDLOFF[†]

***

Insurers can no longer ignore the promise that the algorithms driving big data will offer greater predictive accuracy than traditional statistical analysis alone. Big data represents a natural evolutionary advancement of insurers trying to price their products to increase their profits, mitigate additional moral hazard, and better combat adverse selection. But these big data promises are not free. Using big data could lead to inefficient social and private investments, undermine important risk-spreading goals of insurance, and invade policyholder privacy. These dangers are present in any change to risk classification. Using algorithms to classify risk by parsing new and complex data sets raises two additional, unique problems.

First, this machine-driven classification may yield unexpected correlations with risk that unintentionally burden suspect or vulnerable groups with higher prices. The higher rates may not reinforce negative stereotypes and cause dignitary harms, because the algorithms obscure who is being charged more for coverage and for what reason. Nonetheless, there may be reasons to be concerned about which groups are burdened by having to pay more for coverage.

Second, big data raises novel privacy concerns. Insurers classifying risk with big data will harvest and use personal information indirectly, without asking the policyholders for permission. This may cause certain privacy invasions unanticipated by current regulatory regimes. Further, the predictive power of big data may allow insurers to determine personally identifiable information about policyholders without asking them directly.

Thus, while big data may be a natural next step in risk classification, it may require a revolutionary approach to regulation. Regulators are going to have to be more thoughtful about when price discrimination matters and what information can be kept private. The former, in particular, will require regulators to determine whether it will be acceptable to charge risky groups more for coverage regardless of the social context in which those risks materialize. Further, for both price discrimination and privacy issues, regulators will have to increase their capacity to analyze the data inputs, algorithms, and outputs of the classification schemes.

<div align="center">***</div>

I.   INTRODUCTION

Big data is at the insurance industry's door. It is frequently in the business, popular,[1] and academic[2] press. The predictive power of big data

---

[1] As proof of this claim, just consider the focus and coverage of three of the leading popular/business news sources. The *New York Times* has a blog dedicated just to technology issues: *Bits: The Business of Technology*, N.Y. TIMES, http://bits.blogs.nytimes.com/ (last visited Dec. 26, 2014). The blog is littered with entries about big data. *See, e.g.*, Steve Lohr, *Google Flu Trends: The Limits of Big Data*, N.Y. TIMES (Mar. 28, 2014), http://bits.blogs.nytimes.com/ 2014/03/28/google-flu-trends-the-limits-of-big-data/. Likewise, the Wall Street Journal has a technology blog, *Digits: Tech News & Analysis from the WSJ*, WALL ST. J., http://blogs.wsj.com/digits/ (last visited Dec. 26, 2014), and regularly has articles on big data. *See, e.g.*, Elizabeth Dwoskin*, 'Big Data' Doesn't Yield Better Loans*, WALL ST. J. (Mar. 17, 2014, 11:19 PM), http://online.wsj.com/news/articles/ SB10001424052702304732804579425631517880424; Ben DiPietro, *Survey Roundup: Money Laundering Moves Up Agenda, Big Data Potential*, WALL ST. J. (Mar. 28, 2014, 2:05 PM), http://blogs.wsj.com/riskandcompliance/2014/03/28/survey-roundup-money-laundering-moves-up-agenda-big-data-potential/. Forbes also has regular features and blog posts about the impact of big data in the financial sector. *See, e.g.*, Adam Tanner, *Different Customers, Different Prices, Thanks to Big Data*, FORBES, Apr. 14, 2014, *available at* http://www.forbes.com/sites/adamtanner/2014/03/ 26/different-customers-different-prices-thanks-to-big-data/.

[2] For a small smattering of just the legal academic articles about big data, consider the following titles: Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014);

analytics has been touted as game changing for goals as diverse as ending poverty,[3] stopping terrorism,[4] and transforming business practices.[5] Its evangelists see big data as the most important development since the advent of the Internet. However hyperbolic these claims, there is no doubt that this press has had some effect as a wide variety of businesses are using or considering how to use big data analytics.

Despite this, insurers have been slow to adopt big data analytics.[6] There are, however, few industries with as voracious an appetite for data, in any form, as the insurance industry. Carriers likely can no longer ignore the possibility that the algorithms driving big data will offer greater predictive accuracy than traditional statistical analysis alone. And, if realized, this additional accuracy could potentially benefit insurers in at least three ways.[7] First, by analyzing purchasing patterns, carriers could better target those individuals most likely to buy new coverage and retain those insureds most likely to switch to a different carrier.[8] Second, insurers may be able to use claims and settlement patterns to better distinguish

---

Jeffrey Alan Johnson, Senior Research Analyst, Inst'l Research & Info., Utah Valley Univ., Address at the Annual Conference of the Midwest Political Science Association: From Open Data to Information Justice (Apr. 13, 2013) (transcript available at http://ssrn.com/abstract=2241092); Frank Pasquale, *Restoring Transparency to Automated Justice*, 9 J. ON TELECOMM. & HIGH TECH. L. 235 (2011); Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013).

[3] WORLD ECON. FORUM, BIG DATA, BIG IMPACT: NEW POSSIBILITIES FOR ECONOMIC DEVELOPMENT (2012), *available at* http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development.

[4] John Arquilla, *Small Cells vs. Big Data: Can information Dominance Crush Terrorism?*, FOREIGN POL'Y (Apr. 22, 2013), http://www.foreignpolicy.com/articles/2013/04/22/small_cells_vs_big_data.

[5] Kenneth Neil Cukier & Viktor Mayer-Schoenberger, *The Rise of Big Data: How It's Changing the Way We Think About the World*, FOREIGN AFF., May/June 2013, at 28, *available at* www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data.

[6] *See* Eric Brat, Bos. Consulting Grp., Big Data and Paradigm of Insurance: Summary of BCG Recent Research, Presentation at the University of Connecticut School of Law Big Data and Insurance Symposium (Apr. 3, 2014).

[7] CHARLES NYCE, PREDICTIVE ANALYTICS WHITE PAPER 4 (Am. Inst. for CPCU/Ins. Inst. of Am. 2007), *available at* http://www.theinstitutes.org/doc/predictivemodelingwhitepaper.pdf.

[8] *Id.* at 4–5.

between real and fraudulent claims.[9] Third, again, to the extent greater predictive power is realized, carriers could use big data analytics to price their products more accurately.[10] This Article focuses on the implications of this third category. While big data analytics are a natural evolutionary step for insurers trying to price their products, the regulatory ramifications of this move are potentially revolutionary.

Insurers set prices by predicting the probability that any group of observationally identical individuals will suffer a loss and predicting the magnitude of that loss in the insurance period. Insurers individuate those prices by determining whether the particular observable characteristics of a particular insured correlate with particular harms.[11] For example, based on auto claim data, insurers believe that young men are more likely to be in auto accidents and cause more damage than other demographic groups.[12] Therefore, when a twenty-two year-old man purchases auto insurance, he pays more than a twenty-two year-old woman for the same coverage. Big data promises new opportunities to fine tune risk classification by using algorithms to mine new and complex sets of data to find new correlations and make predictions about behavior. Carriers can gather information about insureds from a variety of new sources, including phone records; the Internet; health records; sensors in cars and clothing, electrical grids, or communication devices.[13] In this way, carriers' use of big data may be a natural evolution in risk classification.[14]

Insurers are already doing some of this. For example, carriers have asked some drivers to equip their cars with electronic devices that monitor their driving patterns.[15] Carriers know that drivers who break harder, drive

---

[9] *Id.* at 5–7.

[10] *Id.* at 5.

[11] Kenneth S. Abraham, *Efficiency and Fairness in Insurance Risk Classification*, 71 VA. L. REV. 403, 408 (1985).

[12] *See* Jessica Bosari, *What Really Goes Into Determining Your Insurance Rates?*, FORBES (Jan. 1, 2013), http://www.forbes.com/sites/moneywisewomen /2013/01/08/what-really-goes-into-determining-your-insurance-rates/; *see also* JJ Arnett, *Developmental Sources of Crash Risk in Young Drivers*, 8 INJ. PREVENTION ii17, ii19 (2002).

[13] *See infra* notes 55–56 and accompanying text.

[14] NYCE, *supra* note 7, at 5.

[15] Allstate explains how this works on its website: "A telematics device is generally a system that you install in your car that records information about your driving habits, such as the number of miles you drive, your speed, and how quickly you brake. These systems sometimes analyze the time of day when you drive, as

faster, or drive during particular times of day are more likely to get into accidents and/or have more severe accidents. Using the data gathered from the devices, carriers can price auto insurance to better reflect the risks posed by the drivers. In the future, carriers could gather this same information from other sources, including communications devices, E-ZPass records, or sensors in the road. It is not a stretch to imagine harnessing more and different information to price different types of policies. For example, carriers could determine whether people who use cell phones at certain times of day, post revealing pictures on social media, or have certain search habits on the Internet are more likely to have liability claims, live shorter lives, or suffer more unemployment claims.

But the potential benefits of big data (to the extent carriers can recognize them) will not be free. Like any improvement in risk classification, additional expenditures on big data analytics could be socially wasteful and privately inefficient. Further, like all risk classification refinements, to the extent that the promised gains in predictive accuracy materialize, classifying risks with big data analytics may undermine important risk spreading goals of insurance. Lastly, mining individual data to build the data sets or to identify whether a potential insured falls into a particular risk category could invade policyholder privacy.

Algorithmic parsing of new and complex data sets may also raise problems unseen in the past. First, machine driven risk classification could yield unexpected correlations with risk. For example, it may be that people from a particular racial or ethnic group have certain Internet search patterns: for example, Jews may search for the time of sundown more often than other groups. Insurers may find those search results yield correlations to particular risks (like Tay Sachs). Carriers focusing on strange algorithmic correlations, like Internet searches to risk of disease, may inadvertently burden these groups with higher prices.[16] Second, insurers classifying risk with big data will harvest and use personal information

---

well. If you use a telematics device from your insurer, you agree to allow the device to send this information to your insurance company." *Tools & Resources: What Is a Telematics Device?*, ALLSTATE, http://www.allstate.com/tools-and-resources/car-insurance/telematics-device.aspx (last visited Dec. 26, 2014).

[16] Of course, it is likely that insurers can already identify individuals by race, gender, ethnic group, etc. without asking these questions. What is different about big data is that the algorithms identifying the correlations may mask the fact that particular groups are being charged higher prices.

indirectly, without asking the policyholders for permission. This may cause certain privacy invasions unanticipated by current regulatory regimes. Further, the predictive power of big data may allow insurers to determine personally identifiable information about policyholders without asking them directly. This means that insurers could be invading new zones of privacy or finding ways to invade zones of privacy once thought protected.

Thus, while it may be a natural evolution for carriers to use big data to classify risk, there may be significant financial and social costs to doing so. These costs may require a revolutionary approach to regulating risk classification. Regulators can no longer rely, to the extent they ever could, on discriminatory intent to protect certain groups from higher prices. To the contrary, regulators must recognize that big data may make it even more likely that certain groups will be burdened with higher prices without any evidence of intentional discrimination. Whether this matters depends on whether a jurisdiction views a particular line of insurance as a means to spread risk generally across society or whether the jurisdiction is comfortable charging risky groups more for coverage regardless of the social context in which those risks materialize. Thus, as will be discussed, big data requires a move from regulating based on discriminatory intent to disparate impact. Further, regulators must determine what information, if any, policyholders may keep private. To protect those privacy matters, regulators will have to increase their computing capacity to analyze the data inputs, algorithms, and outputs of insurers' classification schemes.

This Article looks at the impact of the opaque proxies created by big data and offers some regulatory suggestions to control the risk that individuals or groups will be unfairly burdened by the classification scheme and minimize the risk that insurers will invade individual privacy in new or more nuanced ways.

A.   THE RISK CLASSIFICATION FRAMEWORK

Insurers classify risks by trying to predict the probability that a potential insured will suffer a loss and the magnitude of that loss should it come to pass. To make that prediction, underwriters have traditionally looked at the features and the experience of a potential insured to determine whether and how those features and experiences correlate to insurable losses.[17] Feature rating bases prices on the observable traits of an insured.

---

[17] Abraham, *supra* note 11, at 413–14.

These traits could be inherent, like age, race, gender, or national origin. Feature rating could also look to certain systems that insureds have in place to prevent loss, like smoke detectors, risk management protocols, or whether the insured has taken a particular kind of risk management class (e.g., drivers education).[18] Some of these characteristics are malleable; others are not. That is, an insurer can only control some of these features. In contrast to feature rating, experience rating prices risk based on the loss history of the individual policyholder.[19]

Some individuals have a vector of characteristics that has a low probability of loss conditional on the observables. These individuals represent a low risk and are charged relatively low prices for their insurance. Others have characteristics that correlate more strongly with loss. These individuals represent a higher risk and are charged higher prices.

Insurers have a significant financial incentive to classify insureds properly on the basis of risk. Accurate risk classification can impact the company's bottom line in two ways. An insurer who offers lower prices for good risks could add low risk insureds into its risk pool and thus lower its own risk of paying out. And, if multiple insurers are in the market, accurate risk pricing could allow an insurer to skim good risks away from competitors, leaving the competitor with a comparatively worse risk pool, thus raising its competitor's risk of paying out.[20]

There are well-rehearsed benefits to and concerns with risk classification. On the positive side, accurate risk classification can help mitigate adverse selection and moral hazard. On the negative side, risk classification can be socially costly, may create unfair burdens on certain groups, and may implicate socially suspect categorizations such as race, national origin, or gender.

1.   Benefits of Risk Classification

In addition to the profit motives listed above, carriers may give three justifications for classifying and charging higher premiums on the

---

[18] *See* Ronen Avraham et al., *Understanding Insurance Antidiscrimination Laws*, 87 S. CAL. L. REV. 195, 202 (2014).

[19] *See id.*

[20] KENNETH S. ABRAHAM, DISTRIBUTING RISK: INSURANCE, LEGAL THEORY, AND PUBLIC POLICY 67 (1986); Tom Baker, *Containing the Promise of Insurance: Adverse Selection and Risk Classification*, 9 CONN. INS. L.J. 371, 377 (2003).

basis of perceived risk. These reasons are tied directly to the classic twin insurance dilemmas: adverse selection and moral hazard.

First, pricing based on risk allows insurers to combat adverse selection by marketing to low risks.[21] Potential insureds who are less likely to suffer harm may not want to pay a price that reflects the likely harm of the entire population, including high, medium, and low risks. Low risks (theoretically) may go without insurance rather than pay the premiums that reflect a mix of high and low risk insureds.[22] Thus, risk classification can help alleviate some of the consequences of adverse selection by allowing insurers to price products to entice low risks to enter the insurance pool.[23]

Second, and relatedly, pricing based on risk may be more fair to low risk insureds. All insurance pools are somewhat heterogeneous with low risks subsidizing higher risk policyholders. Risk classification can remove some of the heterogeneity by putting like risks together. The more refined the classification scheme, the more homogenous the resulting pools will be, which will then require less subsidization from low risks to high risks.

Third, risk classification is also a form of moral hazard mitigation.[24] Pricing based on risk provides a signal to insureds about their riskiness.[25] To the extent that insureds have control over the characteristic

---

[21] ABRAHAM, *supra* note 20, at 67.

[22] As Ken Abraham explained, "insurance is only one of a number of ways of satisfying the demand for protection against risk. With few exceptions, insurance need not be purchased; people can forgo it if insurance is too expensive." Abraham, *supra* note 11, at 407.

[23] ABRAHAM, *supra* note 20, at 67. The likelihood of this adverse selection is unclear. There is some evidence that low risk individuals are risk adverse and tend to buy insurance as well as take added precautions. *See, e.g.,* David Hemenway, *Propitious Selection*, 105 Q.J. ECON. 1063 (1990); *see also* Peter Siegelman, *Adverse Selection in Insurance Markets: An Exaggerated Threat*, 113 YALE L.J. 1223 (2004) (reviewing the literature on propitious selection). The impact of this propitious selection in various insurance markets is unclear and, even if there is some form of propitious selection, pricing based on risk remains a potential marketing opportunity to low risk groups that typically go without insurance.

[24] Tom Baker & Rick Swedloff, *Regulation by Liability Insurance: From Auto to Lawyers Professional Liability*, 60 UCLA L. REV. 1412, 1419 (2013).

[25] *See* Abraham, *supra* note 11, at 413 ("Risk classifications should reflect differences in expected losses between classes of insureds; ideally, they should also create loss prevention incentives based on variables within each insured's control.").

upon which they are being classified, the signal of a higher price may encourage potential insureds to change their behavior—either to take more precaution or to reduce the frequency of the risk creating activity. To provide policyholders such incentives to change extant behaviors, insurers must reevaluate and reclassify policyholders periodically.[26]

2. Dangers of Risk Classification

Even assuming that the classification accurately predicts risk, properly mitigates adverse selection and moral hazard, and allows insurers to increase their profits, allowing insurers to make these kind of distinctions among potential insureds raises three distinct types of concerns: efficiency, fairness, and privacy. [27]

a. Efficiency

Risk classification may be inefficient in several ways. First, it may be socially wasteful. Risk classification is socially beneficial to the extent that insurers succeed in bringing new, low risk entities or individuals into the overall risk pool. To the extent that insurers only succeed in moving low risks from one carrier to another, the money spent on risk classification is socially wasteful.[28] This is especially problematic when it is particularly costly for the insurer to acquire the information it needs to segregate risk classes.

Second, risk classification may be inefficient if the higher prices inhibit high-risk, but socially beneficial behaviors.[29] For example, if high medical malpractice insurance premiums for obstetricians drive physicians out of that field and into others, risk classification may create inefficiencies.[30]

---

[26] *Id.*

[27] Ronen Avraham et al., *supra* note 18, at 204-20; Abraham, *supra* note 11, at 419–420.

[28] Avraham et al., *supra* note 18, at 208–09.

[29] *Id*. at 205.

[30] If it is inefficient to classify on the basis of risk in this type of situation, then there are still questions about who should subsidize the behavior. For example, should the entire insurance pool (in the example above, all physicians) pay a higher premium so as not to disincentivize the behavior? Or should the public at large subsidize the behavior through tax subsidies or caps on damages?

Third, risk classification may be inefficient because it may inhibit private acquisition of socially useful information. If risk classification is based, in part, on the knowledge of the insured (e.g., in the case of genetic diseases known only through testing), insureds may choose not to obtain that information.[31]

b. Unfair burdens

Beyond these concerns, risk classification may unfairly burden particular groups. Some view insurance as a means of spreading risks throughout an entire population.[32] Risk classification undermines these risk spreading ideals. If all of society is (or all policyholders are) included in the pool, each individual can use insurance to maintain the status quo. But if insurers classify on the basis of risk, or deny insurance based on the amount of risk a potential insured presents, some individuals may be significantly burdened or even locked out of the safety net provided by insurance.[33] Said differently, if insurance is a means to promote social solidarity, the economic costs of risk factors should be distributed evenly across society. Classifying on individual characteristics "undermine[s] this feature of insurance by 'fragmenting communities into ever-smaller, more homogenous groups.'"[34]

This ideal is particularly undermined if insurers classify risk based on a suspect category.[35] Obviously, some groups are more likely to incur certain types of expenses than others. Women are more likely to incur medical costs associated with pregnancy and breast cancer. Men may have

---

[31] *Cf.* Alexander Tabarrok, *Genetic Testing: An Economic and Contractarian Analysis*, 13 J. HEALTH ECON. 75, 80 (1994) (explaining why people may choose not to get genetic testing even if there is a possibility that the information gained could help minimize the risk of future harm).

[32] *See* Baker, *supra* note 20, at 392–96 (arguing that those who believe risk classification is a fair mutual aid fail to see that the fairness justification for classification lacks the moral force its proponents believe it has).

[33] *See id.* at 392 (explaining how in the late 1980s insurance companies tried to exclude battered women from the insurance pool).

[34] Avraham et al., *supra* note 18, at 215.

[35] Indeed, the most obvious classifications will be based on just such distinctions. Age, sex, race, etc. have been traditional underwriting criteria. *See, e.g.*, Regina Austin, *The Insurance Classification Controversy*, 131 U. PA. L. REV. 517, 517 (1983).

lower life expectancies than women.[36] African Americans are more likely to have medical costs associated with sickle cell anemia; Jews are more likely to have medical costs associated with Tay-Sachs. The elderly are more likely to die than the young. The young are more likely to get in car accidents than the middle aged. The list could go on and on. And, in some sense it may make sense to charge members of these groups more for different types of insurance because of their higher risk status. But there may be significant social and other reasons to ignore the additional risk factors.

First, to the extent that these groups are constitutionally protected based on race, religion, or national origin, there might be concerns that the classification system "reinforces or perpetuates broader social inequalities or . . . causes some sort of expressive harm by acknowledging and legitimating that prior unfair treatment."[37] Said differently, even if it is true that a particular group is more likely to suffer a particular kind of loss, one might be concerned that by being charged more, the extra charge reinforces negative stereotypes, the group suffers certain dignitary harms, and/or the group is unfairly burdened.

Even if the classification is not based on a constitutionally protected class, risk classification may still be viewed as unfair if the rate is based upon a characteristic that is undeserved or when the potential insured does not have control over the characteristic.[38] For example, even if it is true that women who have suffered domestic abuse tend to require additional health care services over the course of their lives, it may be unfair to charge these women higher premiums, because the victims do not deserve their high-risk status.[39] It is, of course, difficult to determine

---

[36] *World Health Statistics 2014*, WORLD HEALTH ORG. (May 15, 2014), http://www.who.int/mediacentre/news/releases/2014/world-health-statistics-2014/en/.

[37] Avraham et al., *supra* note 18, at 217.

[38] *See* Deborah S. Hellman, *Is Actuarially Fair Insurance Pricing Actually Fair?: A Case Study in Insuring Battered Women*, 32 HARV. C.R.-C.L. L. REV. 355, 384–85 (1997); Austin, *supra* note 35, at 559–60.

[39] *See* Hellman, *supra* note 38, at 356–57, 369, 384. This intuition is doubly true when the characteristic is both undeserved and uncontrollable. As Alexander Tabarrok notes in the context of pricing based on health risks:

> First, the intuition that those with higher risks should bear the costs seems less justifiable when the higher risk is not a matter of choice. Is it right that someone with the Huntington's gene should have to pay potentially staggering insurance bills or

whether an insured deserves or can control a characteristic,[40] but these are debates into which I need not wade for purposes of this Article. It is merely important to note that even if certain characteristics predict risk accurately there may be reasons that insurers should not classify their insured on that basis. Likewise it may be unfair to burden individuals with higher rates for socially valuable activity. For example, women in their twenties and thirties are more likely to incur medical expenses related to child birth. But it may not be fair to charge them higher premiums based on those expected medical expenses.

Further, and relatedly, some may view as unfair risk classifications based on characteristics that do not seem to cause the particular harm predicted. Of course, all risk classifications are based on correlation, not causation. But some correlations have a strong causal backbone. For example, the link between Huntington's disease and death is more than just correlative, and, for example, insurers can tell strong causal stories about the links between obesity and health. Other correlations may seem random, or even discriminatory, but actually have certain causal links. For example, facially there does not appear to be a link between credit scores and automobile accidents. There may, however, be common psychological and biological roots to financial risk-taking and risky driving.[41] But, where the

---

> even [be] denied health insurance altogether? Second, charging higher premiums will not reduce the number of people with Huntington's. Thus, in this case, there is no efficiency gain from charging high risk elements larger premiums (only a wealth transfer).

Tabarrok, *supra* note 31, at 80.

[40] Take health status, for example: in some respects, insureds can control their risk factors: they can stay fit, eat right, and abstain from smoking or drinking too much. But, of course, fit people can get sick, many obese people live until old age, and smokers may not get cancer. So what does it mean to control one's health status? *See* Avraham et al., *supra* note 18, at 215.

[41] *See* Patrick L. Brockett & Linda L. Golden, *Biological and Psychobehavioral Correlates of Credit Scores and Automobile Insurance Losses: Toward an Explication of Why Credit Scoring Works*, 74 J. OF RISK & INS. 23, 26 (2007). Further, to the extent that bad credit scores significantly correlate with suspect or vulnerable characteristics, there may be statistical methods to isolate and eliminate these proxy effects while maintaining the predictive accuracy of the variables. *See generally* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J.: ECON. POL'Y 206 (2011).

characteristic is causally remote from the predicted loss and is thus perceived to be non-causal (perhaps in a but-for sense of the word), the use of the characteristic may be challenged on the ground that the classification is unfair.[42] For example, there is near perfect correlation between the per capita consumption of cheese and the number of people who die by becoming entangled in their own bed sheets.[43] But there is little argument that the amount of cheese consumed in the United States says anything interesting about death by entanglement. If there is no causal connection, it is unclear that it is reasonable for insurers to base rates on spurious correlations.

Lastly, there are fairness concerns based on the fact that risk classification is expensive and imperfect.[44] Despite the benefits of risk classification, carriers do not have an incentive to make risk classes completely homogenous (nor could they necessarily do so). Risk classification is expensive, and at some point the marginal increase in homogeneity may cost more than the marginal benefit to the insurer.[45] Thus, some members of the group will always be a higher risk than other members of the same risk class. To the extent that the burden of the imperfections and inaccuracies in the classification scheme falls disproportionately on one group over another, risk classification may implicate additional fairness concerns.[46]

c. Privacy

Risk classification raises a number of privacy concerns. To classify risks, insurers may have to ask about or otherwise discern particularly intimate information about an insured, such as credit score, HIV status, genetic information, or sexual orientation.[47] Insurers could also

---

[42] *See* Avraham et al., *supra* note 18, at 218–20.

[43] *See* Tyler Vigen, *Per Capita Consumption of Cheese (US) Correlates with Number of People Who Died by Becoming Tangled in Their Bedsheets*, SPURIOUS CORRELATIONS, http://www.tylervigen.com/view_correlation?id=7 (last visited Dec. 26, 2014).

[44] Ken Abraham refers to this problem as differential inaccuracy. *See* ABRAHAM, *supra* note 20, at 84–89; Abraham, *supra* note 11, at 429–36.

[45] Avraham et al., *supra* note 18, at 217 ("Efficient insurance regimes will only invest in improving classification to the extent that the resulting benefits are larger than [the cost of doing so].").

[46] Abraham, *supra* note 11, at 429–36.

[47] Avraham et al., *supra* note 18, at 220.

ask questions about drug and alcohol use, lifestyle, exercise, etc. Of course, these are not just idle questions. Failure to answer or answer truthfully could have significant ramifications. Potential insureds refusing to answer could be denied coverage.[48] And policyholders who respond inaccurately could be denied coverage after suffering a loss.[49] These privacy concerns are redoubled when one considers that insurance is a de facto requirement for a number of important life activities like driving a car and owning a home.[50] Thus, many may be forced to divulge particularly intimate information about themselves to obtain insurance.

Which areas are off limits and which questions delve too deeply into private spheres depends on the product line and one's prior assumptions about the strength and meaning of privacy. For example, one's use of alcohol, tobacco, or illegal drugs might be relevant to life expectancy and thus some may not view questions about these topics on a life insurance applications as invasions of privacy. Others, however, may view those questions as intrusive of a personal sphere of privacy regardless of the relevance of the information to the line of insurance, because they represent inquiry into a particular type of personal activity. As with the issues related to control over a particular risk factor,[51] it is not necessary to settle debates about which questions are appropriate in which policy lines and which questions invade a particularly private sphere. It is enough to note that risk classification may implicate privacy concerns even in the absence of the big data concerns to be raised later in this essay.

---

[48] *See* Philip R. Reilly, *Public Policy and Legal Issues Raised by Advances in Genetic Screening and Testing*, 27 SUFFOLK U. L. REV. 1327, 1356 (1993) (refusing to answer questions about genetic background can lead to denial of coverage); James Monroe Smith, *When Knowing the Law is Not Enough: Confronting Denial and Considering Sociocultural Issues Affecting HIV Positive People*, 17 HAMLINE J. PUB. L. & POL'Y 1, 34-35 (1995) (failing to disclose HIV diagnosis is a material misrepresentation justifying denial of coverage).

[49] Avraham et al., *supra* note 18, at 210.

[50] *Id.* at 220.

[51] *See id.* at 215.

II.   THINKING ABOUT BIG DATA

     A.   BIG DATA AND THE DATA DRIVING IT

        Big data derives its name from the mountain of information created
by daily activities and gathered by all types of commercial and
governmental entities. The data includes such sources such as Internet
"transactions, email, video, images, clickstream, logs, search queries,
health records, and social networking interactions."[52] These online sources
could include both the primary record (e.g., a tweet or Facebook post) and
the metadata of the record (e.g., the time and date of posting, the type of
media used in the post, the number of retweets, etc.). But big data is not
limited just to information from the Internet. Big data can also include
traditional data sets and it increasingly includes "sensors deployed in
infrastructure such as communications networks, electric grids, global
positioning satellites, roads and bridges, as well as in homes, clothing, and
mobile phones."[53]
        Given the vast reach and the variety of types of data, there is a
tendency, especially among commercial entities, to define big data in terms
of the amount of this information and the ability to manage that data. For
instance, McKinsey Global Institute, an offshoot of McKinsey &
Company, defines big data as "datasets whose size is beyond the ability of
typical database software tools to capture, store, manage, and analyze."[54]
These quantity definitions often refer to the rapidly increasing amount of
data created every year.[55] Other definitions point out that it is not just the
amount, but also the type of data being gathered that matters. For example,
Forbes, writing for a corporate clientele, defined big data as "a collection of
data from traditional and digital sources inside and outside your company

---

[52] Tene & Polonetsky, *supra* note 2, at 240.

[53] *Id.*

[54] JAMES MANYIKA ET AL., BIG DATA: THE NEXT FRONTIER FOR INNOVATION,
COMPETITION, AND PRODUCTIVITY 1, (McKinsey Global Inst. ed., 2011), *available
at*   http://www.mckinsey.com/insights/business_technology/big_data   _the_
next_frontier_for_innovation (click download full report).

[55] Kenneth Cukier, *Data, Data Everywhere*, ECONOMIST (Feb. 25, 2010),
http://www.economist.com/node/15557443   ("[T]he   world   contains   an
unimaginably vast amount of digital information which is getting ever vaster ever
more rapidly.").

that represents a source for ongoing discovery and analysis." [56] This collection of data, according to Forbes, includes structured and unstructured data. The former refers to data points that are easily placed into databases. The latter refers to inherently more messy data like text in tweets, video uploads, pictures, etc.[57]

Simple size-and-kind definitions, however, tend to be driven by companies selling analytic products, marketers selling big data services, insurers trying to optimize offerings, and Wall Street traders interpreting and predicting the market.[58] These definitions overstate the importance of the amount of data and understate the way the data is analyzed and the sociological meaning of the term. "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets."[59] Big data analytics do not necessarily rely on large data sets—in fact, the set of data may be smaller than traditional (non big) data sets.[60] Rather than think of big data as different because it relies on *big* data sets, it is better to think of big data analytics as different because big data uses complex algorithms to mine messy and diverse data sets.[61] What is unique about big data is that the algorithms driving the analytics are not like

---

[56] Lisa Arthur, *What is Big Data?*, FORBES (Aug. 15, 2013), http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/.

[57] *Id.*

[58] Nicole Wong, Twitter's former legal director and the Obama administration's Deputy U.S. Chief Technology Officer, tweeted, "Tweeps, can you point me to the best available definition of 'big data'? A lot of marketing-speak out there, low on precision." Nicole Wong, Twitter (Jan. 25 2014, 12:56 PM), https://twitter.com/nicolewong/status/426413033200812033. *See also* Tim Harford, *Big Data: Are We Making a Big Mistake?*, FIN. TIMES (March 28, 2014), http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html# axzz30tH6hAOd ("As with so many buzzwords, 'big data' is a vague term, often thrown around by people with something to sell."); Danah Boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662, 663 (2012).

[59] *See* Boyd & Crawford, *supra* note 58, at 663.

[60] Size-based definitions are limiting in two respects. First, they are limited temporally given the ever-expanding computational power of computers. What once required so-called super computers can now be done on simple desktop machines. Second, the definition is over-inclusive. Some of the data "encompassed by Big Data (e.g. all Twitter messages about a particular topic) is not nearly as large as earlier data sets that were not considered Big data (e.g. census data)." *Id.*

[61] *See* Crawford & Schultz, *supra* note 2, at 96.

traditional statistical techniques, and allow data scientists to look at data that was once thought unusable. These new techniques have also given rise to the sociological meaning of big data. As Crawford and Schultz argue, there is a growing and pervasive "belief that large data sets generate results with greater truth, objectivity, and accuracy."[62]

Despite this belief in the perfection of big data, there may be serious concerns about the data and the outputs. First, there are a number of errors that may exist in the data. As Boyd and Crawford explain, "[l]arge data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together."[63] Moreover, some have expressed concerns about which data are collected and used. For example, "in case of social media data, there is a 'data cleaning' process: making decisions about what attributes and variables will be counted, and which will be ignored. This process is inherently subjective."[64] Even choosing to use certain data can be misleading. Not everyone is on Twitter or Facebook, and those who are aren't created equally. Some users post far more often than others.[65] And the data sets themselves are far from pure. Twitter, for example, doesn't make available all tweets and any sampling will likely over-represent the present.[66] Moreover, even if the data were clean and unbiased, there is a problem of over fitting. Given the enormous number of data points considered, there is a risk that the algorithms will find correlations with statistical significance even if there is no meaningful connection between the variables.[67]

That said, private actors have every incentive to find meaningful correlations and data analysts are well aware of the problems listed above. Thus, it is unsurprising that these concerns have not dampened either the demand for big data analytics or the belief in the power of the correlative and predictive outputs.[68] This demand has created a business of collecting

---

[62] *Id*.

[63] Boyd & Crawford, *supra* note 58, at 668.

[64] *Id.* at 667.

[65] *Id.* at 669.

[66] *Id.*

[67] Gary Marcus & Ernest Davis, *Eight (No, Nine!) Problems With Big Data*, N.Y. TIMES (Apr. 6, 2014), http://www.nyti.ms/1kgErs2; Harford, *supra* note 58 (detailing the downfall of Google Flu Trends as a "theory-free, data-rich model").

[68] In part this may be because data scientists managing big data analytics promise that they can massage the messy data and weed out correlations that have no real causal validity.

personal information either for use of the entity doing the collecting or for sale to third parties. The next section provides a brief taxonomy about how big data relates to personal information and the resulting privacy concerns.

B.   USES OF PERSONAL INFORMATION

There are serious concerns about the way entities collect and use personal information.[69] These privacy concerns have driven much of the debate about the use of big data. There are a number of different ways that data brokers and other entities could interact with an individual's personal information. Many of those ways could implicate a number of privacy concerns. Rather than catalog various privacy concerns and the debate surrounding them[70]—the contours of which are not directly relevant to this paper—what follows is a brief description of the ways in which big data could use personal information generally and a sense of the privacy implications.

First, personal information could be harvested to power the algorithms. As described above, companies obtain data from a diverse set of human activities, including online interactions such as ecommerce or social networking and other activities of daily living like using a cell phone or driving a car with E-ZPass.[71] Data brokers collect and categorize each of these data sources to identify correlations and predictions about individuals and their habits. Data brokers cull and sift reams of this personal data without the knowledge of those who generate the data. Generally speaking this data need not be identified with a particular person. Or, at least, in this context, the data are not used in a personally identifiable way. Rather, the data are grist for the algorithm mill. It is the raw material out of which the big data analytics create their correlations and predictions. From a privacy standpoint, one might be concerned that the data are being harvested without consent and often without the knowledge of the content

---

[69] Neil M. Richards & Jonathan H. King, *Big Data Ethics,* 49 WAKE FOREST L. REV. 393, 408 (2014).

[70] For an example of the types of concerns, see *id.* For a flavor of the debate, see Jules Polonetsky & Omer Tene, *Privacy and Big Data: Making Ends Meet*, 66 STAN. L. REV. 25 (2013).

[71] Tene & Polonetsky, *supra* note 2, at 240.

generators.[72] Further, while the data is not necessarily used to identify specific individuals, personal identity is also not scrubbed from the data.[73]

Second, companies run personally identifiable information through the algorithm. That is, companies use personal information from a particular individual to determine whether that individual's characteristics correlate to a particular set of outcomes. As above, there are significant concerns in this respect that individuals do not know what data is being harvested and used to determine correlations. For example, are banks using an individual's Facebook posts or pictures to modify his or her credit ratings? Or, in the context of this Article, are carriers gathering data about individual insureds to determine their riskiness? The data collected and used in this way are not anonymous, nor can they be. This raises, at a minimum, concerns about the access that corporations have to private data.

There may be second order concerns related to this algorithmic use of personal data. As Crawford and Schultz suggest, "[b]ig data processes can generate a model of what has a probability of being [personally identifiable information], essentially *imagining* your data for you."[74] For example, in 2012, Target used big data analytics to effectively predict which of its customers were pregnant and passed that information to its marketing arm.[75] That is, without asking any customers about their pregnancy status or harvesting that data in particular, Target was able to predict extremely sensitive and personal information about its customers.[76]

Third, and relatedly, companies harvest and use data without respect to who generates the data for marketing purposes. For example, companies typically gather all sorts of information from Internet searches to target marketing. While few express concerns about this targeted marketing, it is nonetheless another way that companies use private information (individual searches) without permission.

---

[72] Crawford & Schultz, *supra* note 2, at 94-95.

[73] Tene & Polonetsky, *supra* note 2, at 251-252.

[74] Crawford & Schultz, *supra* note 2, at 98.

[75] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[76] This so-called predictive privacy invasion may result in a number of harms. For example, marketers could attempt to avoid anti-discrimination statutes by simply directing on-line marketing to groups segregated by certain demographics, including race, gender, age, credit worthiness, etc. Crawford & Schultz, *supra* note 2, at 99-100. Crawford and Schultz also raise concerns about predictive policing and health care privacy.

III.   BIG DATA AND RISK CLASSIFICATION

When risk classification actually results in identifying better and worse risks and provides carriers the ability to price these differential risks correctly, the benefits of risk classification mostly redound to the insurer in the form of greater profits from a better risk pool and to low risks in the form of lower-cost insurance. The costs, on the other hand, manifest in the form of privacy invasions and higher prices on select groups. As such, it is easy to see why insurers would want to enhance their classification capabilities. Big data offers just such an opportunity.

There could, however, be a number of significant issues related to using big data to classify risk. This Article assumes away myriad potential problems with the data[77] by assuming that insurers only use good data—that is, data that represents a good statistical sample, has few biases in place, and no major errors. Further, this Article assumes that the data are providing correlations that represent actual differences between risk classes. That is, this Article assumes the data show that some set of people who have some set of characteristics is more risky than some other set. Even if all of this is true, there remain specific efficiency, fairness, and privacy concerns raised by insurer's use of big data to classify risks.

The social and private costs attached to using big data to classify risks may be significant and include inefficient investment of capital, unfair burdening of groups and individuals, and inappropriate invasions of personal privacy. These costs suggest potential regulatory responses. Whether and how regulators should respond, however, turns on a number of things including the incentives that private actors have in the marketplace to self-correct, the cost of any regulatory response, the costs created in the absence of a regulatory response, and views about the underlying purpose of insurance. Typically it is left to industry to fix problems stemming from inefficient investments. Carriers have significant incentives to determine for themselves whether investments in big data are profitable and adding new insureds to the pool. And it is not clear there is a role for regulators in solving whatever collective action problems might exist. On the other hand, regulators may have a reason to insert themselves into problems created by the disincentives created by big data, unfair burdens created by risk classification, and increased privacy concerns.

This part focuses on the costs created. Part IV addresses potential regulatory responses.

---

[77] *See* Boyd & Crawford, *supra* note 58, at 666-75.

A.   EFFICIENCY

Most carrier expenditures implicate a number of efficiency concerns. Carriers' expenditures on marketing, information technology, or even policy drafting could be unprofitable, socially wasteful or otherwise inefficient. Investments in big data to classify risk are no different. For example, it might be extraordinarily expensive to harness big data and generate more refined risk classifications. Each carrier might have to spend significant sums to make marginal improvements to their risk classification scheme. These costs could be exacerbated because carriers may feel a pressure to follow popular trends. Given the press coverage on the wonders of big data, firm leaders may spend exorbitantly even if the new classification scheme costs more than it generates in revenue for two reasons. First, carriers may hope that classifying based on big data now will reap profits in the future. Second, carriers may fear that if other insurers get better at classifying risks, they will lose low risk insureds, thus making their pool worse and forcing them to pay out more. These investments may be inefficient in two ways. First, it is unclear whether the refinements based on big data (to the extent they can be made) will bring in new, low risk policyholders into the insurance pool. If not, the expenditures on risk classification through big data will be socially wasteful, perhaps significantly so if the associated costs are particularly high. Second, the investment in big data may not be profitable. Given the collective action problem, firms may continue to invest so that a competitor that is using big data does not undercut their prices.

Whether, in fact, the expenditures to classify risk using big data are worth it for either the individual firm or for the industry as a whole is an empirical question. In thinking through this analysis, one must determine the following: is the use of big data profitable? Are new insureds being added? Is there a collective action problem spurring socially wasteful investments?

Further, the fear of big data may have inefficient impacts on policyholders and potential insureds. Individuals may refuse to invest in socially useful activities or fail to acquire important information for fear of being charged higher premiums or excluded from insurance altogether. For example, genetic testing could be both a socially useful activity and provide privately important information. It could both inform public understanding of genetic disorders and private decisions about health and welfare. But individuals may forgo genetic testing because insurers can

use the information discovered by those tests to set rates for or exclude individuals from life, disability, and long-term care insurance.[78]

B.   FAIRNESS

Policyholders may argue that using big data to classify risk unfairly burdens some groups. Of course, all risk classification burdens some groups more than others—that is the nature of differential pricing. Big data, however, has the potential to change old debates about risk spreading versus pricing based on risk. As discussed below, the algorithms driving big data analytics may find correlations between risk and suspect or vulnerable classes or based on non-causal factors without the insurer being aware that particular groups are being financially burdened.

Whether these higher prices should be thought of as unfair depends, in no small part, on one's belief about the underlying nature of insurance.[79]

1.   Proxies for Suspect and Vulnerable Classes

Insurers have long gathered data about policyholders' race, gender, age, and income level for many different lines of insurance. Insurers could easily use traditional statistical techniques to determine whether these or other suspect or vulnerable characteristics correlate strongly with loss. Even if characteristics that receive heightened constitutional protection (such as race, religion, and national origin),[80] characteristics that identify individuals as members of vulnerable groups (such as income), or characteristics that are otherwise undeserved (such as victims of domestic violence)[81] correlate more significantly with loss, there may be good policy reasons not to charge higher premiums on this basis alone. The cause of the higher risk rating may be bound tightly to sociological and historical

---

[78] *See* Kira Peikoff, *Fearing Punishment for Bad Genes*, N.Y. TIMES (Apr. 7, 2014), http://nyti.ms/1mVTzYS. The Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 100-233, 121 Stat. 881 (codified as amended in scattered sections of 42 U.S.C. and 29 U.S.C.), only covers the behavior of health insurers. It does not prohibit using information from genetic testing to price or exclude individuals from other insurance products.

[79] *See supra* notes 32–49 and accompanying text.

[80] Austin, *supra* note 35, at 517.

[81] *See* Baker, *supra* note 20, at 392.

conditions, making the higher risk status *undeserved.* Charging higher premiums "saddles people with all the consequences of their high risk status, whether deserved or not . . . [and] entitles other people to all the benefits of their low risk status, also whether deserved or not."[82]

These consequences could include making it more difficult to access insurance as a social safety net, reinforcing negative stereotypes, and causing dignitary harms. The first of these is obvious. Making insurance more expensive may make it impossible for some individuals to purchase the financial security that insurance provides. But charging more could have other negative effects. If it is known that members of a group pay higher premiums because they are members of the group (even if there are actuarial reasons for the higher premiums), it may reinforce a belief that the members of the group deserve their high-risk status or are burdens on society. For example, people may believe that Jews deserve Tay Sachs, that the poor actively choose not to take care of their health or property, or that victims of domestic abuse are responsible for their additional medical costs. This could serve to further reinforce negative stereotypes and thereby cause dignitary harms.

But insurers need not base the higher premiums directly on the characteristics listed above. There could be non-suspect individual characteristics that correlate with both a suspect or vulnerable characteristic and high-risk status. For example, property insurers could base higher property insurance rates on crime statistics. If people of color primarily live in areas with higher crime rates, the higher premiums would be based on a factor—crime rates—that correlates with race. Carriers could justify additional premiums based on the higher rate of loss in high crime areas. Outside of any current regulatory regime that prohibits disparate impact,[83] would it be normatively defensible to allow insurers to charge higher rates

---

[82] *Id*. at 394.

[83] The United States Department of Housing and Urban Development ("HUD") recently promulgated a rule establishing that a plaintiff can establish a Fair Housing Act claim on the basis of discriminatory effects. HUD's final rule declares that "[l]iability may be established under the Fair Housing Act based on a practice's discriminatory effect . . . even if the practice was not motivated by a discriminatory intent." 24 C.F.R. § 100.500 (2013). This regulation presumably prohibits charging higher rates for property insurance to people of color even if the rates are actuarially fair. That is, the rule would prohibit the disparate impact of the higher prices for property insurance. This rule is, of course, limited to those insurance types that lay within HUD's ambit. Neither this rule nor any other prohibits higher prices for life or auto insurance.

to people of color?   There are strong arguments on both sides of this debate.   On one hand, as with all risk classification, price differentiation allows a carrier to control adverse selection and moral hazard.   Further, some may view it as fairer to charge those who have less of a risk of loss less for their property insurance.[84] On the other hand, if one views insurance as a means of risk spreading, it may be unattractive to charge the high risk group higher premiums. This argument has additional weight in this example because there may be historical and sociological reasons for higher crime in particular areas. Further, insureds who live in high crime areas may not have the means to move.   Under this view, society as a whole bears some responsibility for the high-risk status of the insureds. And, importantly, because insureds cannot move, they likely cannot mitigate the risk of living in a high crime neighborhood. Further, to the extent that areas of high crime are predominately made up of people of color, there may be a risk that the higher premiums reinforce negative stereotypes, and thereby impose dignitary harms on those affected.

Big data has the potential to change some of this analysis, although it depends, in part, on the type of proxies that carriers find for high-risk status.   Insurers could find obvious correlations between non-suspect characteristics and both a suspect or vulnerable characteristic and high-risk status. It is easy to imagine the kind of data that may correlate more strongly with women than men; particular racial, religious, or ethnic groups; people from a particular country; or particularly vulnerable individuals. Women may "like" Oprah more often on Facebook, Jews may search more frequently for the precise timing of sundown on Google, people of Filipino decent may be more likely to follow @MannyPacquiao on Twitter, victims of domestic violence could search more frequently for women's shelters or about restraining orders, and the poor may be more likely to look up information about social services.

It is unlikely that carriers would make it known why policyholders fall into high-risk groups—for example, by explaining which behaviors correlate with higher risk. But if they were to do so, these obvious proxies raise a similar set of normative arguments as described above. Carriers could justify the higher rates on both the adverse selection argument and the argument that it may be fairer to the low risk group to pay less for coverage. The moral hazard mitigation argument, however, holds little water in this context. There is little argument that the correlatives to risk identified above are, in fact, causal. As such, there is little benefit to

---

[84] Baker, *supra* note 20, at 375.

encouraging, for example, fewer "likes" on Facebook or fewer Google searches for sundown. Further, to the extent that the proxies are obviously coextensive with a suspect or vulnerable group, there may be a risk that the higher premiums reinforce negative stereotypes and impose dignitary harms on those affected.

The far more likely scenario is that it will not be readily apparent to anyone why some individuals are charged more. The algorithms driving big data will simply spit out higher prices for some policyholders than others. Carriers will not directly explain nor will it be obvious to insureds or third parties why some individuals are charged higher premiums. Insurers may treat the information as proprietary and thus have an incentive to conceal the reason for the pricing from the policyholders (especially given that there is likely no moral hazard mitigation to be done). This may mean that the algorithms driving risk classification will identify groups of risky individuals without anyone intending or even knowing that many of the identified individuals are members of a suspect or vulnerable group.[85] As discussed below, this opacity changes the arguments for and against risk classification. Importantly, if it is not clear who is charged more for insurance or why, there is little argument that insurers are reinforcing stereotypes or that policyholders are suffering dignitary harms.

As with obvious proxies, carriers could argue that the risk classification helps mitigate adverse selection and is fairer to low risk groups. And, like obvious proxies, carriers cannot argue that the pricing helps mitigate moral hazard. There is no risk-related reason to encourage people not to buy certain types of paper towels or place cell phone calls at a particular time of day.

What is different is that the reasons against classifying risk look very different. Here, even if a proxy is coextensive with a suspect class, the reasons for the increased rates are obscured. The algorithms are simply spitting out high-risk groups. The carriers may not even know that many or most of those charged higher rates are members of suspect or vulnerable

---

[85] There could be another possibility: the carriers reveal the correlations with risk, but those correlations are not obviously linked to a particular suspect or vulnerable class. Hypothetically, imagine that individuals who buy a particular kind or amount of paper towel, who call particular area codes at particular times of day, or who use social media in a particular way are more susceptible to a particular type of risk *and* are more likely to be members of a suspect or vulnerable group. There is nothing obvious to link those behaviors to particular groups. In that case, the same arguments about opacity discussed below apply.

groups. And given this, it is unlikely that policyholders or the public know either. Thus, it is difficult to see how the higher rates reinforce stereotypes or cause particular groups to suffer dignitary harms.

The remaining argument against using characteristics that correlate with both a risk factor and a suspect class is that the group will be burdened unfairly. Whether this disparate impact matters depends in large part on whether one views insurance as a vehicle for social solidarity through risk spreading or not. As above, if a particular group has a propensity for higher risk, then one may consider it fair to charge that group more for coverage. If one views insurance as a mechanism for society-wide risk spreading, then risk classification is rarely acceptable.

The table below summarizes these arguments. The three left columns represent the general arguments for risk classification. Where an "X" appears, carriers can reasonably make an argument in favor of classifying risk based on the type of proxy. As the chart makes clear, any time a characteristic correlates with risk—even if that characteristic also correlates with a suspect or vulnerable group—an insurer can argue that charging higher premiums helps fight adverse selection and is fairer to the low-risk group.[86] But, for most of these potential correlations, insurers have no reason to encourage their insureds to minimize the activity correlated with risk and thus do not mitigate moral hazard through pricing. Carriers can only mitigate moral hazard when the correlation to risk is known, is causal to the risk, and can be controlled by the policyholder.[87] For example, insurers can offer price breaks to install smoke detectors or take defensive driving classes. This helps mitigate the risk from materializing and controls moral hazard. On the other hand, if the price of auto insurance is based on age or sex, charging higher prices to young men does not encourage a different type of behavior. Policyholders are unlikely to be able to control most of the correlations found through big data. Even if the policyholder can control the characteristic upon which the carrier classified

---

[86] *See* Abraham, *supra* note 11. There are reasons to question the adverse selection story generally. It is, however, intuitively true that insurers can induce additional policyholders to pay for coverage by offering lower rates. In doing so, carriers may be mitigating some adverse selection, or at least enhancing their bottom line. This adverse selection argument is subject to a number of constraints. *See e.g.,* Ronen Avraham et al., *Towards a Universal Framework for Insurance Anti-Discrimination Laws*, 21 CONN. INS. L.J. (forthcoming 2014).

[87] *See generally* Baker & Swedloff, *supra* note 24 (discussing risk-based pricing as a means of mitigating moral hazard).

the risk (e.g., by defriending Oprah), it is unlikely that the changed behavior will actually result in fewer losses.

The arguments *against* classifying risk based on suspect classifications are far more equivocal. One could argue that charging more simply based on an underlying suspect or vulnerable characteristic reinforces structural inequality, reinforces stereotypes, and creates dignitary harms. It may be that where a carrier uses a proxy (whether through big data or not) that is obvious and fairly coextensive with a suspect class, the higher premiums will create the same harms. But, as the reasons for the higher premiums become less clear, as the algorithms obscure who is paying more and for what reason, the arguments change. With no obvious connection to a particular group, the extra premiums neither cause dignitary harms nor reinforce negative stereotypes. Thus, the only argument left against classifying risk in this way is that the high-risk group is unfairly burdened by the high premiums. This puts one's view of insurance front and center in the debate.

Arguments For/Against Using Proxies For Suspect Classes in Risk Classification

	Arguments for Risk Classification			Arguments Against Classifying Based on Suspect Class		
	Adverse Selection	Moral Hazard	Fairness to Low Risk	Reinforce Stereo-type	Dignitary Harms	Unfair burden
Suspect characteristic directly correlates to risk	X		X	X	X	X
Non-suspect characteristic correlates to suspect characteristic and risk	X	If insured can control the factor	X	X	X	X
Obvious big data correlation with suspect characteristic and risk	X		X	X	X	X
Nonobvious big data correlation with suspect characteristic and risk	X		X			X

2.   Non-Causal Correlations

Even if insurers identify correlations with risk that do not disparately impact suspect or vulnerable groups, there may other concerns with correlations identified by big data.   The algorithms may find correlations with risk for which carriers can tell no plausible story about the causal connection between the behavior and the loss. Big data is very good at finding subtle correlations, but these correlations may not be meaningful because the correlations are to activities that are unrelated to the underlying loss.[88] Of course, both traditional and big data risk classification are based on correlations.   As discussed above, some correlations, such as the connection between smoking and illness or early death, have a significant causal backbone. For other correlations, such as a link between age and driving, carriers can tell a plausible story: young men act rashly and do not have fully developed control over their rapidly changing emotions, and are therefore more erratic drivers.   But, there are certainly identifiable correlations with risk for which there is no plausible story—for example the link between consumption of cheese and death by entanglement in bed sheets.[89] If there is no causal connection, it is unclear that it is reasonable for insurers to base rates on spurious correlations.

Big data analytics exacerbate concerns that insurers will identify risks which have no causal relationship whatsoever to the insured loss. In part, this is due to the magic of big data. The Holy Grail for big data is finding subtle, yet undiscovered correlations. The problem, of course, is that finding such non-causally related correlations means that the policyholder cannot, and likely should not, try to minimize the activity, behavior, or characteristic. Imagine the following: using big data analytics, some carrier realizes that individuals who purchase vampire novels on

---

[88] "[A]lthough big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, it never tells us which correlations are meaningful. A big data analysis might reveal, for instance, that from 2006 to 2011 the United States murder rate was well correlated with the market share of Internet Explorer: Both went down sharply. But, it's hard to imagine there is any causal relationship between the two. Likewise, from 1998 to 2007 the number of new cases of autism diagnosed was extremely well correlated with sales of organic food (both went up sharply), but identifying the correlation will not by itself tell us whether diet has anything to do with autism." Marcus & Davis, *supra* note 67.

[89] *See* Vigen, *supra* note 43.

Amazon, "like" vampire related media on Facebook, or follow authors of vampire fiction on Twitter are more likely to engage in risky behavior. Reading vampire novels or being a fan of vampire fiction could be within the control of the policyholder, but should insurers be allowed to classify risks along these lines? There is likely little, if any, causal connection between being a fan of vampire fiction and an actual risk. Carriers have little reason to encourage the policyholder to be less of a fan of vampire fiction. So what is left to justify the higher prices? Carriers, of course, can still argue that prohibiting price discrimination—even for these non-causal characteristics—would create adverse selection problems and be unfair to low risk policy holders.[90]

Again, this pits two different kinds of fairness arguments against each other. Big data has laid bare the essential nature of insurance. Should individuals who are higher risks have to bear the burden of that status even when no one can tell a reasonable story about why they have that high-risk status? Should low risks subsidize high risks even if they do not have any reason for being in the low-risk group?

3. Opacity in Correlation

As noted above, big data is unlikely to provide simple, easily explainable reasons for higher premiums. Rather, carriers classifying risk in this way will likely just charge some group of policyholders higher premiums without explanation based a number of factors, each of which is obscured by the underlying algorithmic analysis. This lack of transparency raises a number of issues.

On the one hand, as noted above, opacity undermines fears that higher rates will create a particular stigma or dignitary harm to the high-risk group. Policyholders likely will not know whether they or others are paying more for insurance or whether any particular groups are being singled out for higher rates.   Thus, higher rates may not reinforce stereotypes, stigmatize a particular group, or create dignitary harms. But, the lack of transparency means that a policyholder may not be able to change his or her behavior even if he or she has characteristics that should and may be classified as high risk and can and should be controlled. In short, unless the carrier identifies which factors are leading to higher rates, there is little moral hazard mitigation to be done. All that is left to justify

---

[90] The likelihood of these claims depends in large part on the line of insurance. *See* Avraham et al., *supra* note 86.

the difference is a fear of adverse selection and a sense that it would be fairer to the low risks to charge them less.

Further, the opacity of the algorithm raises concerns about error. Imagine that the overall classification system works in that the insurer correctly identifies a certain set of characteristics that correlate with more risk, the carrier induces more insureds into the risk pool, and the classification system is otherwise efficient. There may still be individuals who are misclassified as high risk. The lack of transparency in the data collected and the algorithm deriving correlations means that these otherwise low risk individuals may not be able to determine why they were moved into the higher risk group or how to fix it.

C.   PRIVACY

Interestingly, both big data and risk classification raise significant privacy concerns. First, as noted above, privacy issues are raised any time a carrier classifies risks (with or without big data) on intimate, personal information, like HIV status, marital status, sexual orientation, or genetic information. [91] Likewise, privacy concerns are implicated any time a company obtains and uses personal information to augment its databases or any time a company feeds personal information through its big data algorithms for correlative or predictive purposes. Thus, it is natural that there would be significant privacy concerns when risk classification is combined with big data analytics. [92] There are two principal ways that big data raises new privacy concerns for risk classification.

First, insurers now may be able to collect information about current or potential policyholders from public sources that the carriers are prohibited from asking a policyholder about directly. For example, it is reasonably easy to imagine that carriers could access information that policyholders share via social media about themselves, including for example, sexual orientation. While policyholders may want to share that information with friends and family, they may not want a carrier to have it. If, to follow through on this example, carriers cannot ask about sexual

---

[91] *See* Avraham et al., *supra* note 18, at 220.

[92] There is significant literature about whether these intrusions into personal space, or extrusions of personal information are privacy harms. It is beyond the scope of this Article to resolve any of these debates. Rather, at issue here is whether insurers using big data to classify risk implicate new or different privacy concerns.

orientation in classifying risk, they should not be able to use Facebook posts to identify the same information for classification purposes.

Second, and relatedly, carriers may be able to use predictive analytics to discern private information that they should not otherwise have or use as a basis for risk classification. For example, as described above, Target used shopping patterns to discern which of its customers were pregnant.[93] It is easy to imagine an insurer using the same or similar data to predict pregnancies or other personal information. Again, and without specifying where the boundaries are, if a carrier is prohibited from asking about the information in the first instance, the carrier should not be able to predict the same.

IV.   REGULATORY RESPONSE

The financial and social costs listed above suggest a regulatory responsibility to actively consider the ways that big data could change risk classification. Big data has the potential to strip away certain reasons for and against risk classification. Possibly gone are credible claims to the benefits of managing moral hazard and concerns about explicit harm from being singled out as different as a result of being a member of a suspect or vulnerable group. Left are old debates. Are low risks entitled to the benefits of their low-risk status? Or, should society subsidize high-risks because it is, for some reason, inappropriate to saddle high-risks with the burdens of their status?

Similarly, gone are old ways of protecting privacy. Insurers may not need to explicitly ask questions that invade particularly private spheres. Instead, carriers can base decisions on a set of correlations and predictions that may burden particular groups more than others or may invade particular zones of privacy.

Big data thus implies a move from conscious discrimination and explicit privacy invasions to unconscious proxies.[94] Whether and how regulators respond will depend on jurisdictional priorities. Is there a will to protect groups impacted by higher premiums or to protect certain intimate information? The answer to these questions may depend on the line of

---

[93] *See supra* notes 75–76 and accompanying text.

[94] *Cf.* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact* 20–21 (Oct. 19, 2014) (unpublished article), *available at* http://ssrn.com/abstract=2477899 (explaining that sorting using relevant criteria in making decisions also can serve as proxies for class).

coverage and the precise group burdened or information used. But, protecting these groups requires regulators to think actively about the harms and the remedies.

A.   REGULATION OF DISCRIMINATORY IMPACT

To the extent that there is any legislative or regulatory will to engage with these discrimination or privacy issues, big data changes the conversation.[95] To monitor, curb, control, and eliminate these concerns, legislators and regulators must look at the outputs of, rather than the inputs to, the classification system. That is, they can no longer—to the extent that they ever did—worry about whether carriers are directly grouping suspect classes or basing rates on other socially vulnerable characteristics. Instead, in the age of big data, regulators must look at how particular classes and individuals are being charged and then determine whether those charges constitute an impermissible burden.

Regulators must first determine whether insurers are charging higher premiums to particular groups or individuals with particular characteristics (such as characteristics that are non-causal to the potential loss or represent socially vulnerable groups). This will require some additional legwork on the part of carriers and regulators, because insurers will have to determine not just who is being charged more but whether there are any patterns to the classes of risk. Are, for example, African Americans being charged more for a particular line of coverage? Or, are people without children being charged more for other lines of coverage?

Legislators and regulators must then compare these groups and individuals against internal calculations about whether and how insurance should spread risks and in which forms. Even if risk and loss correlate with suspect classes, actuarial science should not necessarily govern insurance rates; higher rates of loss may reflect socioeconomic realities that should

---

[95] It is not at all clear that there is legislative will to engage with this, or for that matter, any discrimination. There is little federal oversight of discrimination within the insurance industry. *See* Avraham et al., *supra* note 18, at 198 (listing the limited number of federal laws and regulations on point). State regulation of discrimination in insurance is highly variable across jurisdictions and across lines of insurance. *Id.* at 268. For the most part, states have not even prohibited explicit discrimination based on race, religion, or national origin. *See id.* at 267 ("[L]aws often have little to say about the most important divisive types of discrimination: distinctions based on race, national origin, or religion.").

not burden one group over the population.[96] These calculations may differ across lines of insurance. There may be certain lines of insurance that require additional protection against discrimination. For example, given the semi-mandatory nature of homeowners insurance and the perceived importance of homeownership, there may be reasons to put more weight in the risk spreading rationale. This may be why federal regulators have instituted a very rare federal overlay of anti-discrimination regulation for homeowners' insurance.[97]

If the state chooses to make a commitment, legislators should prohibit carriers from placing any extra burden on suspect classes. This analysis highlights one clear fact: the regulatory response to big data in the risk classification sphere is going to turn on the underlying normative framework of the state. When a state believes that a particular line of insurance is designed more to spread risk, it must be on the lookout for disparate impacts.[98] When a state does not, it need not worry.

B.   REGULATION OF PRIVACY

The analysis for privacy intrusions is similar, but the prescriptions may be different. First, states must determine what, if anything, constitutes a privacy invasion in this context. Can carriers mine and use data anonymously? Can carriers use non-anonymous data about policyholders? Can carriers use predictive analytics to determine characteristic about the carrier that were otherwise private?

After determining what matters, regulators will face the same issues that others have flagged in a number of big data contexts: how to protect end consumers from privacy invasions and predictive analytics?[99] To resolve these issues, regulators need a two-pronged approach. First, regulators will have to audit insurers' classification systems looking at the "data sets mined" by the algorithms, as well as the "source codes and programmers' notes describing the variables, correlations, and inferences

---

[96] *See id.* at 267 ("Even when actuarial support can be found for these assumptions, that does not mean that they are not intimately tied up with socially suspect characteristics.").

[97] *See* 24 C.F.R. § 100.500 (2013); *see also supra* note 83.

[98] *Cf.* Citron & Pasquale, *supra* note 2, at 13–16 (describing how credit scores might have a disparate impact on racial minorities).

[99] *See, e.g.,* Citron, *supra* note 2; Crawford & Schultz, *supra* note 2, at 95; Richards & King, *supra* note 69, at 408.

embedded" in the algorithm.[100] These audits should focus on whether personal data is appropriately scrubbed from the data used to create the predictions, whether carriers are gathering inappropriate individual data, and whether the data are suggesting inappropriate correlative predictions. Second, regulators may want to institute a hearing procedure for individuals who believe that inappropriate data are being gathered or used.[101]

V.   CONCLUSION

Big data may be a natural evolution in risk classification. It makes sense for insurers to take advantage of new data sets and new algorithms to derive new correlations to risk. After all, insurers have a number of incentives to refine their pricing, including the possibility of higher profits and better management of adverse selection.[102] But, these new correlations may yield price discrimination that disparately impacts some suspect or vulnerable groups of people. Further, the algorithms may use or divine information that has otherwise been entitled to some privacy protection.

These two costs suggest a somewhat revolutionary approach to regulation. First, regulators will have to actively consider whether it is acceptable for each line of insurance to have prices that burden suspect or vulnerable groups. This will put in stark relief important choices about whether insurance is about risk assessment or risk spreading.[103] Regulators will have to consider whether to protect certain groups of people from higher insurance prices, even if there are sound business reasons for carriers to charge the affected policyholders more. To the extent that

---

[100] *See* Citron & Pasquale, *supra* note 2 at 23.

[101] *See* Crawford & Schultz, *supra* note 2, at 111; Richards & King, *supra* note 69, at 426.

[102] Unless big data (a) yields correlations that make transparent the policyholders' risky behavior and unless (b) that risk behavior is controllable and (c) has a causal relationship to the risk, there is no argument that the higher prices will control moral hazard.

[103] *See generally* ABRAHAM, *supra* note 20, at 65 ("In short, attitudes toward insurance always seem to be pulling in two directions—one that highlights the risk-assessment or efficiency promoting features of insurance classification and the other that stresses the risk-distributional function of insurance."); Baker, *supra* note 20, at 25 ("Thus, debates over the legitimacy of particular forms of risk classification invoke classic debates over the nature of distributive justice.").

regulators want to protect these groups, the regulatory regime will have to change from one based on prohibiting intentional discrimination to one based on prohibiting the disparate impact of business decisions.[104] Second, for both price discrimination and privacy issues, regulators will have to increase their capacity to analyze the data inputs, algorithms, and outputs of the classification schemes.

---

[104] As discussed above, HUD has already made that determination in the context of claims based on the Fair Housing Act. *See supra* note 83.