Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing

Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, Bill Howe
{megyoung,billhowe,rodriglr,efkeller,fs377,boyangsa,janwhit,billhowe}@uw.edu
University of Washington

ABSTRACT

Data too sensitive to be "open" for analysis and re-purposing typically remains "closed" as proprietary information. This dichotomy undermines efforts to make algorithmic systems more fair, transparent, and accountable. Access to proprietary data in particular is needed by government agencies to enforce policy, researchers to evaluate methods, and the public to hold agencies accountable; all of these needs must be met while preserving individual privacy and firm competitiveness. In this paper, we describe an integrated legal-technical approach provided by a third-party public-private data trust designed to balance these competing interests. Basic membership allows firms and agencies to enable low-risk access to data for compliance reporting and core methods research, while modular data sharing agreements support a wide array of projects and use cases. Unless specifically stated otherwise in an agreement, all data access is initially provided to end users through customized synthetic datasets that offer a) strong privacy guarantees, b) removal of signals that could expose competitive advantage, and c) removal of biases that could reinforce discriminatory policies, all while maintaining fidelity to the original data. We find that using synthetic data in conjunction with strong legal protections over raw data strikes a balance between transparency, proprietorship, privacy, and research objectives. This legal-technical framework can form the basis for data trusts in a variety of contexts.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**;
• **Applied computing** → *IT governance*;

KEYWORDS

data ethics; data governance; privacy; algorithmic bias; data sharing

ACM Reference Format:
Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, Bill Howe. 2019. Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287577

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287577

1 INTRODUCTION

The mechanisms by which algorithms can be made more fair, accountable, and transparent require broad access to sensitive data, data that is typically "closed" due to concerns over proprietorship and privacy. Data ownership models in competitive markets foreclose the possibility of inter-organizational data sharing and collaborative analysis between researchers, firms, and the public sector. Data deemed too sensitive to release through open data efforts typically remain unavailable, leading to convenience sampling effects where researchers, startups, and the general public put disproportionate attention on already opened data, whether or not it is suitable for their purposes. To combat this perceived dichotomy between open and closed data, we emphasize the release of semi-synthetic datasets that control bias and account for privacy, organized through a suite of data governance policies to facilitate responsible data sharing in public-private partnerships.

The transportation sector helps to motivate and illustrate our approach. Private firms hold an increasing share of information about urban transportation provision; including widely adopted services like car share[31], ride share [42], bike share [67], prediction apps for public transportation [24], and routing apps [13]. Like the taxi and limousine services that preceded them, city agencies increasingly require these new services to share data in order to enforce permit requirements, enable integrative models of demand and ridership, and analyze their policy implications. To date, existing data sharing paradigms have failed to deliver granular access to firm data; when it is shared, it is often encumbered with contractual obligations that preclude linking data across competing firms. As corporate data is zealously guarded to protect competitive advantage, releasing data via open data portals or detailed APIs is untenable in many situations. Notably, in the absence of access to firm information, researchers spend considerable resources simulating it, as evidenced by work modeling the supply and distribution of car share vehicles [31]. Researchers have also used data mining [29], social experiment [27], and intercept surveys at "hot spot" [51] to collect TNC related data due to the absence of direct access to firm information. However, each data source has its own biases, which may lead to distortions in research findings. For example, disparities in wait times based on passenger race were not present in data mined from Uber's API [29] but were present in data collected via 1,500 rides in a controlled field study [27].

This paper describes dilemmas attributable to the current data sharing paradigm for (i) privacy, (ii) fairness, and (iii) accountability in the urban transportation data domain. In each case, we examine how purely technical approaches are incomplete. In turn, we provide evidence in favor of co-designed legal and technical tools to address these gaps. In the discussion, we describe the design of a legal-technical infrastructure called the Transportation Data

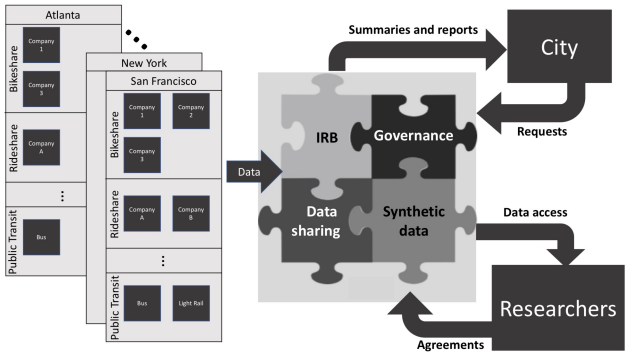

Figure 1: Overview of the Transportation Data Collaborative

Collaborative (TDC) that offers an alternative to "open" or "closed" dichotomy. The TDC emphasizes the release of customized synthetic datasets for most use cases, along with structured data use agreements to govern access to high fidelity data. These mechanisms provide flexible access that balance the competing interests between individuals, firms, governments, and researchers. Finally, we report from initial experiences operating the TDC in Seattle, WA, where it is being used by government agencies, private firms, and university researchers to enable granular data access, integration across competing firms, and the removal of bias that can propagate inequity. We find it can enable research access to data that would otherwise remain unavailable while protecting privacy and enforcing compliance.

2 RELATED WORK

Researchers focused on "wicked" urban problems [52] such as housing discrimination, transportation management, and crime reduction have been limited to working on a small set of canonical open datasets and published algorithms. While this approach affords the opportunity to refine and validate findings, it necessarily limits their representativeness. For example, a preponderance of predictive policing scholarship [41] [23] examines the algorithm underlying one vendor's software, PredPol™, which the company published in an academic paper [46]. Other salient predictive policing algorithms are not available for in-depth or comparative analysis. A similar problem is evident in recidivism and racial bias research, which has been limited to COMPAS datasets [1]. Despite the scale and granularity of its data about criminal defendants, scholars have found significant flaws in its use for recidivism analysis; it reflects and reproduces racial bias even without using race as an attribute[7], and lacks predictive power over both simpler models and untrained humans in a lab experiment [20]. These examples indicate that dedicating resources to a convenience sample of available datasets and algorithms falls short in addressing real-world problems that necessitate a more representative array of sources.

In practice, the availability of sources as open government data also guides the direction of scholarship. An increasing number of academic researchers use open government data as a primary data source, or to validate their findings [63]. To the extent that such initiatives serve dual aims of government transparency and public collaboration, it is difficult to fully satisfy either [11, 58]. Open data is both labor intensive to produce and 'self-selected' by nature [19]. As a result, open government advocates argue that it is unlikely to disclose certain types of information [60, 65]. Finally, open data quality issues reduce its usefulness for research purposes. Challenges including lack of sufficient metadata, curation, findability, interpretability, completeness, interoperability, and granularity [50, 59, 64] have been documented in open data programs across various stages of maturity.

Current technical approaches. Data sharing and release has been studied from multiple perspectives, including causality-based reasoning for fairness and differentially private synthetic datasets. Recent reports on data-driven decision making underscore that fairness and equitable treatment of individuals and groups are difficult to achieve [6, 9, 45], and that transparency and accountability of algorithmic processes are indispensable but rarely enacted [12, 17, 57].

The importance of causality in reasoning about discrimination is recognized in recent work. Kusner articulated the link between counterfactual reasoning and fairness [36]. Datta et al. introduce quantitative input influence measures that incorporate causality for algorithmic transparency to address correlated attributes [16]. Galhotra et al. use a causal framework to develop a software testing framework for fairness [26]. Nabi and Shpitser use causal pathways and counterfactuals to reason about discrimination, use causality to generalize previous proposals for fair inference, and propose an optimization problem that recomputes the joint distribution to minimize KL-divergence under bounded constraints on discrimination [47]. The approach we use to remove bias as part of synthetic data generation builds on theoretical work relating causality to fairness [34].

Bindschaedler et al. consider plausible deniability for privacy [10], and Kifer and Machanavajjhala caution that any system looking to satisfy differential privacy must be both explicit about and careful of what it means to conceal participation of an individual in the data generating process, showing how this can lead to privacy breaches [33]. Our approach builds on prior work on publishing differentially private histograms, as summarized by Meng et al [43]. In particular, Xiao et al. propose a technique for using subcube histograms to improve accuracy in which the inputs are already binned into ranges [61]. Similarly, the concept of universal histograms helps Hay et al. [28] improve the accuracy over the original approach by Dwork et al. [22]. Building on the intuition that histograms depend heavily on bin choice, NoiseFirst and StructureFirst explicitly address both issues (Xu et al. [62]). All of these approaches allow for potential improvements to the accuracy of the released data under differential privacy, and are presented as purely technical contributions. In this paper, we consider how variants of differential privacy interact with legal constraints on data sharing.

3 WHY OPEN V. CLOSED FAILS

In this section, we describe examples of how data access and use under the dominant data sharing paradigm present obstacles to privacy, fairness, and accountability at present. For each topic, we

[1] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions, and refers to a software tool used in pretrial, parole, and sentencing decisions to assess the risk a criminal defendant will commit a crime.

describe how the coupling of legal infrastructure and technical approaches offers a potential solution.

3.1 Privacy

High-resolution mobility data holds inherent risks for user privacy. Any two individuals' location traces are unlikely to be similar, rendering them vulnerable to re-identification when trace records are linked with public information about home address or place of work. de Montjoye et al. [18] were able to re-identify 95% of the individuals in an anonymized mobile phone dataset with only four spatio-temporal points, demonstrating that even significant geographic aggregation is insufficient to protect privacy. Recent analysis demonstrates that as privacy preservation in large mobile phone datasets rises, its research utility declines[48]. In the case reported here, such work affirms the privacy concerns in de-identified data traces.

Legal context. Local laws also contribute to privacy concerns. Under freedom of information laws in the U.S., most government data is open on request [60]. In many states, records cannot be exempt from disclosure based on privacy concerns. For example, in the State of Washington, the state Public Records Act (RCW 42.56) defines privacy narrowly as information whose disclosure would be both "highly offensive to a reasonable person" and would "not [be] of legitimate concern to the public"– in practice, courts find few public records requests that rise to this standard (RCW 42.56.050). Rather, narrowly defined data attributes are exempt by legislative action, such as the residential addresses of public employees. In the absence of a legal privacy exemption, governments may collect and store high-resolution mobility data, but do not have the means to protect it. Public records laws further impede agency access to data, in that any industry information shared with regulators may be subject to requests that could unduly expose customer data to disclosure. These concerns reify a data sharing ecosystem in which high-resolution data is owned by transportation service providers, but not shared at a level of detail that would support analyses into multi-modal transit, management of public rights-of-way, or behavioral change in the transport sector.

Protecting privacy in shared mobility data. In the summer of 2017, the City of Seattle began a pilot program for 'dockless' bikes, issuing permits for three different companies: Lime, Ofo, and Spin. In exchange for a permit to operate in the city, Seattle required that each firm share granular data about its ridership, for the purpose of evaluating the services as they were being received by consumers throughout the city. Cities have long requested detailed travel information from service providers in exchange for a permit to operate on city streets, as evidenced by municipal taxi and limousine regulations. Such requests may include GPS traces of origins, destinations, and routes, with exact time stamps, as well as demographic details about the consumer. The aim of the city in requesting this data is complex; including the cost-effectiveness of services, public safety, the effects of services on social and economic inequality, the balancing of competing uses of public space, and plans for future investment in public infrastructure. These concerns change over time, creating the need for adaptive data-driven responses to questions as they are raised by community members and their political representatives. At the time, the city was seeking a way to conduct these analyses without violating the privacy of individual riders.

Differential privacy and synthetic datasets. Differential Privacy (DP), first proposed in 2006, has been applied to the problem of generating synthetic datasets for public release in various forms [21, 38, 66]. Differential privacy is an assertion about a specially designed query Q, called a *mechanism*: For any given result R, and any given individual i, we consider the probability that Q would return R if i was *included* in the dataset against the probability that Q would return R if i was *excluded* from the dataset. If these two probabilities are "close," we can conclude that the privacy for i has been protected, because it is difficult to infer whether or not i was included in the dataset. A parameter ϵ puts a bound on the ratio between these two probabilities for all individuals and all possible results of Q.

Most differentially private mechanisms achieve this bound by adding noise to the result of the query. For example, if we are interested in computing the average height of a set of people, we can add just enough noise to the answer to "hide" the presence of any individual. To use DP to generate a synthetic dataset, we design a special mechanism M that summarizes the original dataset (e.g., using a set of histograms), adds noise appropriately, and then samples these summaries to generate synthetic data. The goal is to produce a dataset that is 'statistically similar' to the original dataset (with respect to the summary selected) but protects the privacy of any particular individual.

Although attractive for its generality, differential privacy requires making a number of assumptions when deploying these techniques in practice.

First, multiple definitions of differential privacy have emerged, partly in response to the difficulty of retaining utility under the original definition. Different definitions pertain to different notions of exactly what disclosure an information publisher wants to protect against. Using the example of a survey, this could include protecting any individual's answer to any single given question on the survey (attribute-level DP), obscuring whether or not an individual even participated in the survey (individual DP), or protecting a particular class of people who were surveyed (group DP). See Kifer and Machanavajjhala [33] for more detail on what it means to conceal attributes, participation, or groups.

Additionally, all definitions of differential privacy include one or more parameters controlling the *privacy budget* (the bound on the probabilities ϵ in the discussion above can be viewed as a "budget" that must be "spent" to protect against different kinds of attacks.) These budget parameters are related to the accuracy of the result: a high budget ϵ implies high fidelity to the original data, but a weaker privacy guarantee.

These parameters are difficult to interpret, and must be chosen by the data publisher with only a few guidelines in the form of community best practices to follow [37]. Furthermore, if the same individual is represented in multiple datasets (as they are likely to be when examining multiple transportation modalities), these multiple synthetic datasets taken together constitute a weaker privacy guarantee than any one would individually. If the first dataset is generated with a privacy budget of ϵ_1 and the second with ϵ_2, then

this individual has the privacy guarantees of each individual data source violated as the resulting information disclosure is equivalent to using a privacy budget of $\epsilon_1 + \epsilon_2$.

Advantages of a data trust for privacy. A data trust such as the Transportation Data Collaborative (TDC) provides a mechanism by which these assumptions can be standardized and made explicit. Indeed, technical instantiation of privacy protections are highly dependent on assumptions about the structure and distribution of underlying data holding true. Moreover, even when separate firms are required to deliver data in a mutually agreed-upon format, there is often significant variation with respect to data structure and provenance. Governments and successive research teams must also re-purpose datasets over time, often in new contexts with new requirements and assumptions. This heterogeneity requires technical expertise and labor to bring disparate pieces into alignment, a process known in the computer-supported collaborative work (CSCW) field as articulation work[15]. A data trust reduces such frictions by: (i) using legal agreements to reach shared understanding with data contributors as to their contents and allowed uses, (ii) connecting rich datasets to the expertise required to create DP mechanisms appropriate for each use case, (iii) ensuring that data is used responsibly by researchers, and (iv) communicating the assumptions on which DP mechanisms are based.

Centralization of data would also allow the publishing of synthetic data that is more representative of real world conditions, while simultaneously strictly meeting the mandated privacy budget. Such limits can be supplemented with contractual legal guarantees as to the level of privacy protection that the data trust personnel agree to provide. Data trust personnel leverage their own technical expertise to assume responsibility in the event that parameters are set too loosely, and re-identification attacks occur. Robust technical and legal approaches for formally enforcing privacy protections create a framework where sensitive, granular, and proprietary data is more likely to be shared.

3.2 Fairness and Bias

Remediating structural inequalities in shared mobility data. Despite unresolved privacy concerns, public agencies must use data to ensure the equitable distribution of resources. The public interest mission of agencies in the transportation sector in general — and in bike share programs, specifically — inclines them to focus on particular features of shared-mobility options, such as (i) accessibility for low-income, elderly, and differently-abled persons; (ii) access to remote areas or those not currently served by transit; (iii) reliability across modes by providing transportation alternatives for the 'last mile' between riders and their homes or workplaces. For instance, combining bike sharing origin-destination (OD) pairs that indicate a high volume of trips can indicate that a particular corridor is not well-served by existing transit networks — emphasizing the need for increased bus service on that route. However, programmatic decisions based on usage may be prone to reproducing certain kinds of structural bias reflected in the data.

In such cases, bike share companies may have targeted their service provision to particular market segments based on ability to pay; use patterns reflected in the data should be understood as skewed with income. Studies suggest that the majority of bicycle trips captured in surveys or volume counts are made by cyclists who are Caucasian, male, and well-educated [25] and take place in highly bicycle-accessible areas [1]. Bicycle facility planning prioritized by the volume data alone fails to serve the neighborhoods most disadvantaged in terms of accessibility [32]. Early deployments of transportation services may favor wealthy neighborhoods, inadvertently discriminating along racial lines due to the historical influence of segregation [3]. Releasing data "as is" would complicate efforts to develop fair and accurate models of rider demand.

A related example of bias in data comes from the NYC 311 reporting system. Kontokosta et al. [35] analyzed the usage of the NYC 311 reporting system, a centralized platform of services and information requests and non-emergency reports, by residential area. Neighborhoods using the system disproportionately less often had a higher minority population, higher unemployment rate, and more non-native English speakers. On the contrary, neighborhoods that tend to over-report are more likely to have higher rents and incomes, and higher educational attainment. This example illustrates that service provision based on non-remediated data sources would reflect and potentially reinforce structural inequities.

Synthetic data generation. To enable responsible data use in these sensitive situations, we advocate releasing "algorithmically adjusted" datasets that *destroy causal relationships between certain sensitive variables* while *preserving relationships in all other cases.*

There are several potential *sensitive causal dependencies* in an urban transportation setting:

- Company A may be marketing to male riders through magazine ads, leading to a male bias in ridership that could be misinterpreted as demand.
- Ride hailing and taxi services allow passengers to rate and tip the drivers; gender or racial patterns in tips or ratings may encourage discrimination by drivers and should be eliminated before attempting to develop economic models of tip revenue.

To remove these sensitive patterns, the data publisher specifies a causal relationship between two attributes X and Y that they wish to eliminate in the adjusted dataset, conditioned on another set of attributes Z. Then the causal repair problem is to set the mutual information between X and Y to zero, conditioned on Z. This approach is composable with differential privacy techniques that add noise to prevent re-identification. We will describe this approach in more detail in Section 6.1.

Advantages of a data trust for fairness. Governing data under a data trust model spreads risk of identification across multiple firms. More importantly, the collection of trip and rebalancing details from the full contingent of firms in a data trust affords researchers access to the full geographic distribution of attributes needed in order to inform adjustment strategies. When a trust houses the full collection of firms participating in the market — as in the case of Seattle bike-share — it becomes possible to decouple the firm from the origin and destination of trips. This transformation better protects their competitive advantage while maintaining similar global properties about the demand to support compliance audits and analysis. We describe the method we adopt for generating synthetic datasets to achieve these properties in Section 6.1.

3.3 Accountability and Transparency

Algorithmic accountability begins with data provenance [40], and data provenance requires application of norms across heterogeneous sources. The TDC achieves provenance as a side effect of auditability requirements imposed by IRB rules, research data retention laws, and our own agreements with firms. In this section, we consider accountability in a more traditional sense, where forms that operate within a jurisdiction must provide evidence that they are complying with relevant laws. The data sharing activities implied by compliance, as we will discuss, cause contention between the competing interests of cities, firms, and researchers.

Legal struggles in proprietary data sharing. Although open public records laws aim to improve accountability, they can paradoxically have the opposite effect by encouraging companies to proactively invoke legal mechanisms to withhold any and all data that may be shared with government. In Washington State, the Public Records Act is "liberally construed, and its exemptions narrowly construed" (RCW 42.56.030) so as to incline disclosure of any information public employees use, prepare, own, or retain. While the intent of the law is to promote government transparency, it also has consequences for individuals [60] and firms [2, 4]. At times, records requests implicate proprietary information (for example, a company's proposal for a contract bid, or information the firm provided in order to receive a permit to operate). The PRA provides for the government to notify the firm of such a request, giving the firm the opportunity to file an injunction with the court to prevent its records from becoming public. As a result, companies often mark the majority of the materials they share with public agencies as 'proprietary and confidential.' Nevertheless, some courts have decided that even information that a firm deems to be proprietary is of enough legitimate public interest to be disclosed (TODO find examples in legal precedent in WA). The potential for disclosure thus disposes firms to consider data sharing with the public sector as risky — further entrenching a siloed data sharing ecosystem.

Preserving competitive interests in integrated analysis. A recent court decision [4] highlights how concerns about fairness, accountability, transparency, and privacy become intertwined in conflicts about data sharing. In its 2016 round of permits for transportation network companies (TNCs, or rideshare companies), the City of Seattle mandated that each of the firms share statistics about the trips taken with their services. Each firm was required to share data aggregated at the zip code level. In January 2016, a man from Austin Texas interested in TNC service provision filed a public records request in the State of Washington for the most recent two quarters of data provided to the City of Seattle on:

- The total number of rides provided by each TNC.
- The percentage or number of rides picked up in each ZIP code.
- The pick-up and drop-off ZIP codes of each ride.
- The number of rides when an accessible vehicle was requested.

In the context of a public records request for this information, Uber and Lyft filed formal injunctions with the court to prevent the City from releasing their aggregated data, believing it to be key to their competitive interests. Of note is that in a market dominated by two players, aggregating marketplace data does not protect firm privacy. After escalating the case, the WA State Supreme court found zip code information to be of legitimate public interest & allowed for its public release. Two points are germane. First, under circumstances where Uber and Lyft were not compelled to offer aggregated data in order to receive a permit, the data would not have been made available for public use. Second, accountability of firms to government (and government in turn to private citizens) elided each firm's market interests.

Figure 2 shows a map of the ridership for the pilot program in Seattle and is indicative of the kind of data products the city requires to assess compliance with relevant policies, including equity.

In order to achieve accountability, the framing for data sharing and release must evolve beyond open v. closed and acknowledge the interests of all relevant stakeholders, including the companies. Geographic aggregation would appear to protect both individuals' privacy as well as firms interests, since it may not be obvious how to infer the contribution of an individual or a company from the sum. We discussed the limitations of aggregation in the context of privacy in Section 3.1, but aggregation also fails to protect companies. If there is a duopoly, as is essentially the case with Uber and Lyft, one company can immediately infer the data of the other company using the aggregated values. Even in less extreme cases, one can make inferences from aggregated data using exogenous information.

Our approach is to recognize that the "all or nothing" approach to protecting firms' information while maintaining accountability is neither possible nor necessary. The risk to competitive advantage or releasing information is much more nuanced. For example, data that is six months old is less valuable to a competitor than data that is current, and ridership data from busy routes is less sensitive than ridership data from neighborhoods where a firm is actively rebalancing bikes.

Protecting proprietary information in synthetic data generation. Our key observation is that we can use the same causality-based approach we outlined to remove bias for fairness reasons in Section 3.2 to also selectively remove certain signals from a dataset prior to computing aggregates. We can hide causal relationships that they consider too sensitive to reveal, just like we hide causal relationships that expose potential discrimination in the data.

Consider the bike share example: shared bikes are mainly used for medium to short distance travel and one-way trips. For example, in Seattle, bikes are commonly used to ride down large hills. Such usage patterns result in a spatially and temporally unbalanced distribution of bikes [14]. If the bikes are not redistributed, it will lead to a low efficiency system and a low quality of service [49]. To avoid customer dissatisfaction, firms have hired personnel to "rebalance" bike distribution across the city [39]. The particular rebalancing strategy of one company could be easily copied at the cost of competitive advantage; it is not in the best interest of the firm to share this data directly with public agencies. How can a public agency know when the rebalancing strategy of a firm is working, or needs adjustment? The same methods noted above to mitigate bias can be applied here, where the causal relationship in question is between the bike company, the timing and location of trips, and the timing and location of rebalancing bike movements.

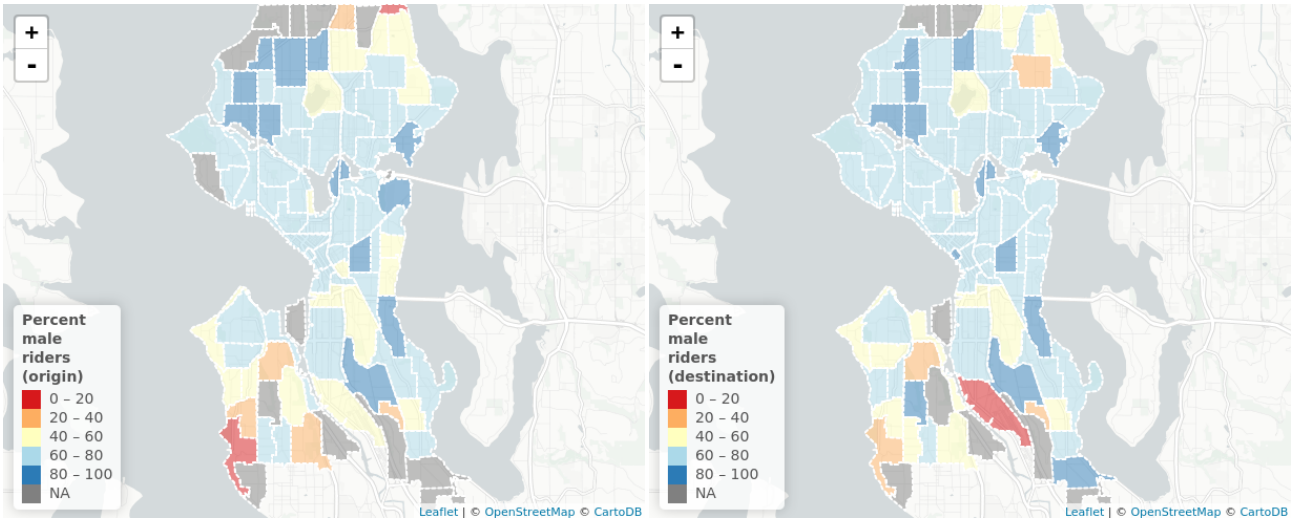

Figure 2: Percentage of bikeshare trips in Seattle with male riders by origin and destination neighborhoods

Advantages of a data trust for ensuring accountability. A data trust provides a marked improvement on traditional data sharing methods in that it respects proprietary and competitive interests firms assert in their datasets without precluding access to their granular, confidential form. Under current practice, governments and firms have entrenched adversarial positions with respect to the scope of data sharing in exchange for city permits. The proposed model refocuses governments on the specific analyses needed for policymaking. A data trust mediates and deduplicates such requests, benefitting from network effects as it scales across jurisdictions. Finally, technical and domain expertise available to university-housed data trusts allow synthetic data requests to be served while preserving salient causal links and removing those that implicate privacy, bias, or firm competitiveness.

A data trust also creates conditions that allow for chained and traceable data provenance. Transportation data from any number of sources are used to develop models, but encoding the data as trained weights in a model ends up "laundering" it — that is, it is no longer transparent to trace the source of the data through to the decisions reached by the model. A trust mitigates the lack of algorithmic accountability in part by emphasizing the use of synthetic datasets whenever possible during research and development; once a proof of concept is established, and access to the raw data is requested, ongoing data sharing relationships mediated through data sharing agreements present an opportunity to enforce provenance.

4 DESIDERATA FOR A DATA TRUST: BEYOND OPEN AND CLOSED

The previous case studies demonstrate the failures of conventional data sharing paradigms in real-world examples of FAT in algorithmic systems, and how integrated legal-technical approaches offer more comprehensive solutions for data access and remediation than technical techniques in isolation. To date, problems related to data sharing have been addressed primarily through (i) NDAs between firms and universities, (ii) privately owned data repositories created toward a shared business interest, or (iii) compelled disclosure of firm data by government.

Here we outline desiderata for an urban data trust, based on lessons learned from related data repositories and our own creation of a university-based repository for geolocation data called the Transportation Data Collaborative (TDC). Responding to the call for data trusts and intermediaries we provide a set of design implications in support of the following goals: (i) protecting individuals' and firms' privacy in granular data, (ii) providing access to researchers from multiple firms in a way that allows those sources to be combined for integrated analysis, (iii) facilitating government monitoring over firms for the purpose of ensuring accountability and equity in service delivery, and (iv) observing firms' proprietary interests in their data.

A third-party data trust works to 1) broker and facilitate query-specific granular data sharing across a public-private collaborative, and 2) research and develop new *integrative services* that provide value to participants without sharing identifiable information. Government agencies can benefit from the trust with evidence needed to enforce policy compliance without acting as stewards for sensitive data – data that they are often legally and technically ill-equipped to protect. For example, in the case of transportation data, a travel demand model trained on public transportation data, rideshare data, bikeshare data, and mobile phone traces can achieve better results than models based on individual data sources. Firm contributors realize value by satisfying requirements for reporting performance of their service to public agencies (i.e., as required for a permit to operate in the city, for example), while protecting the proprietary nature of their data, protecting the privacy of their customers, and by gaining access to new research and related services.

By providing protected access to data to researchers, these trusts centralize and proceduralize interactions between individual investigators and companies that are otherwise typically ad hoc. The resulting data sharing ecosystem provides streamlined and more

equitable access to sensitive data across transportation providers, social sciences, computer science, and statistics.

5 LEGAL INFRASTRUCTURE

To be effective in deriving datasets for these complex multi-party problems and associated interventions, the TDC requires a legal infrastructure by which multiple firms' and agencies' granular data sources could be collected and made interoperable. In its most basic form, the legal and technical infrastructure must support a range of users including private firms, government agencies, academic researchers, and third-party researchers. Each of these have different requirements, and the leadership governing the TDC plays a key role in mediating how the platform will be used.

Private firms collect a vast array of high-dimensional data attributes about urban residents and activities, and contractual or policy protections for firms' competitive and proprietary information are necessary preconditions for access to this data. Under the current paradigm, firms commonly share data with researchers via Non-Disclosure Agreement (NDA), a contract stating the purposes for which data is being shared, and limiting the uses that researchers are allowed to make of the information. NDAs can take many forms, but are generally not configured to allow for the analysis of data shared between multiple organizations. Borrowing from the health sector, we found data sharing and use agreements offer more flexibility for multiparty research than simple NDAs, and can be used in conjunction with NDAs to maintain confidential communications with firms.

Data sharing and use agreements with the TDC serve two distinct purposes. First, data sharing agreements specify the data to be shared by a firm with the trust, how it may be used by external researchers, and acknowledgment that the researchers hosting the data may carry out core activities (e.g., cleaning, linking, and FAT interventions such as bias remediation. Second, they are fully formed contracts that form the legal basis for determining whether the data delivered is of appropriate quality, and for arbitrating disputes in the event of an unauthorized disclosure.

When developing these agreements, firms express requests for bias adjustments, negotiating with TDC leadership as appropriate. Typically firms are willing to share data as long as it does not empower their competitors. In all cases, the TDC is transparent about the terms of the agreement — the goal is to protect sensitive information, not mislead stakeholders by releasing adjusted data as unadjusted data.

Data sharing agreements require substantial review by the counsel of each firm, and are considered to be somewhat permanent, modified as the technology or data itself changes. Data use agreements, in contrast, offer the opportunity for the data trust and the firm to identify and approve uses for the data that are not already ascribed to the data trust in their data sharing agreement. The possible uses of data are a moving target, because the accumulation of data in the trust expands the range of queries possible with the data and the parties interested in the data. Data use agreements are revisited on a regular basis by the TDC and its participating firms for this reason, but also because firms doing business with government agencies can benefit from the role of the data trust as an intermediary in that relationship.

The TDC is established as a "Corporate Affiliate Program" at the university. Approved projects using TDC data beyond its affiliate program are funded as traditional research projects. This model encourages long-term engagements with firms rather than one-off research engagements that have become the status quo.

Government agencies have complex and changing needs for transportation data. As an intermediary between public agencies and firms, the trust has the domain expertise necessary to identify, in cooperation with both parties, a mutually acceptable scope of inquiry, and to manage changes to that scope over time. For the data trust in relation to the firm, the data use agreement captures this changing scope over time. For the trust in relation to the public agency, a separate contractual relationship ensures that the needs of the public agency in reporting to the general public, and the development of a basis in evidence for transportation and related policy can be met. The public sector also owns, contracts for, and operates data intensive systems from which it is able to contribute data to the TDC.

5.1 Human subjects protections

In any research context, access to data that can be used to re-identify individuals begins with consideration of the human subjects represented in the data. In the U.S., this requirement is governed via regulatory proxy in the university setting by an Institutional Review Board (IRBs). IRBs emerged in the U.S. in 1974 to provide ethical oversight and protection of human subjects in research that utilizes federal funding. In recent years, questions have emerged about whether IRB protocols intended for a traditional model of research are applicable or sufficient for addressing the ethical dimensions of new, dynamic forms of 'big data' analysis, in which data representing scores of individuals are re-purposed without an opportunity to use conventional means of informed consent [8, 54]. Data science analysis using publicly available datasets is generally considered exempt from human subjects protections, or subject to minimal review, on the basis that it poses low risks to individuals. However, research based on observational or secondary data collection (in which no direct intervention has occurred) challenges traditional notions of human harm associated with physical or psychological results from active participation in a research experiment [44].

Although university IRB approval was pivotal to establishing the TDC, the use of geospatial trace data from public and private sources for research purposes requires privacy protections for individuals and firms that expand beyond the model provided. A new paradigm in ethics protections would rethink privacy protections for high-dimensional data, which carry the risk of re-identification even when following all available policy prescriptions for anonymization. The TDC makes this leap with a new repository-based approach to the protection of human subjects for research.

The first step in creating this approach was to learn lessons from data repositories in health and medicine[2] that have established procedures to comply with the Health Insurance Portability and

[2]Healthcare repositories have led the way on innovative sociotechnical applications to support stakeholder interests. For instance, the Data Query, Extraction, Standardization, Translation (Data QUEST) and Cross-Institutional Clinical Translational Research (CICTR) projects utilize a federated model in which local partners store their own data and approve each data extraction. These health data platforms provide a socio-technical approach to providing data sharing infrastructure for demographic and medical visit data [55].

Accountability Act (HIPAA), institutional requirements and sector-based ethical concerns. Mechanisms include the development and enforcement of rules in the form of contractual obligations for vendors and researchers and constraints on data use [56], implemented through encryption-based keys for separating or joining datasets in ways that achieve privacy protection, restrictions on downloading data from cloud infrastructure, and keystroke-based logs with manual audits of the patient identifiers accessed and analytics conducted by each data user [5]. Similarly, data security requirements in the health sector demand strict adherence to ethical standards of privacy protection and risk mitigation in sensitive data.

Though researchers may be accustomed to viewing IRB as a source of constraints on research, the IRB approval for the TDC establishes a shared understanding in the research community of the sensitivity of the data – and therefore the nature of the protections called for in research – through unified requirements for protecting the human subjects represented in the data.

6 TECHNICAL INFRASTRUCTURE

In this section we describe two technical components in more detail: the method for generating synthetic data using causal analysis, and the data management architecture designed to afford ingest and processing of heterogeneous data from a variety of sectors, companies, and cities with minimal administration overhead.

6.1 Synthetic data generation

We generate semi-synthetic datasets to balance competing interests: individuals want to retain privacy, firms want to protect competitive advantage, researchers want to study mobility services, their effects on society, and relevant methods. Our approach is to compose established techniques of differential privacy with the causal approach of Rodriguez et al. 2018 [53] to selectively remove sensitive relationships in the dataset.

This process requires that the data owner specify a causal relationship between two attributes that they wish to have removed in the synthetic dataset. Participating firms express these requirements as part of the data sharing agreements. Specifically, they identify a causal relationship between a variable X and a variable Y, where the relationship is conditional on another set of specified attributes Z. Given these attributes, the goal is to force X and Y to be independent with respect to Z by adjusting the data such that the mutual information between X and Y to zero conditioned on Z while preserving all other relationships.

Once these attributes have been specified, we use them to count the co-occurrence of X and Y conditional on different values of Z. We interpret these counts as a probability density function for each of the grouped combinations of attributes, and then update these counts according to a factorization of the chain rule probability for the attribute set. By adjusting the counts in this way, we ensure independence and create a version of the dataset that we can easily sample, yielding a synthetic dataset where X and Y are mutually independent.

Consider an example. Mobility data can be aggregated to *Origin-Destination* (OD) pairs: a set of location pairs representing city blocks or neighborhoods, along with the traffic flow between the pair of locations. We can extend the OD histogram by adding attributes besides origin and destination. For example, bikeshare data includes an attribute *gender* with domain $(male, female, other)$, a binary attribute *helmetuser*, and an attribute *company* with domain (A, B, C) in addition to *origin* and *destination* attributes, each with a domain of 90 neighborhoods. A released dataset then might include the tuple $(female, A, Downtown, Ballard, 245)$ indicating that there were 245 trips taken by female riders on bikes owned by company B from Downtown to Ballard during the time period covered by the dataset. This histogram of ridership represents the joint probability distribution of mobility. By adjusting the ridership counts, we can adjust the joint probability distribution to force independence between company and gender while retaining the other relationships in the data. We may suppress relationships to avoid propagating discrimination into downstream application (Section 3.2), or we may suppress relationships to protect proprietary company information (Section 3.3).

We can add calibrated noise to the counts after the bias repair but before sampling to satisfy differential privacy. We currently use a simple direct application of the Laplace mechanism [22], but this could easily be extended to any of the more sophisticated methods summarized by Meng et al. [43].

As shown by Rodriguez et al [53], this method does not tend to change the underlying distribution more than would a bootstrap sample. This result suggests that we can treat the resulting semi-synthetic dataset as if it were sampled from the same underlying population as the original dataset, but with the conditional mutual information between the specified variables removed.

6.2 Data management architecture

The data management infrastructure to support the trust requires scalable ingest and storage, robust security, and flexible queries. Though these security and access requirements would have been difficult and expensive for a data trust to meet as recently as ten years ago, the boom in cloud computing resources allows a small team to build and administer a platform comparable with those operated at the enterprise level.

The trust has a mandate to accept data feeds from a variety of companies, agencies, and researchers. These data providers have varying degrees of technical maturity and cannot all be assumed to follow best practices in data quality or format. However, the trust cannot reject data that does not conform to desired standards, as the mission is to provide secure and policy-aware access to all data, since even storing the raw data under our legal framework offers some value.

To accommodate the heterogeneity in data and provider, the architecture has loosely coupled tiers: A triage tier for unstructured data, a scalable lake tier for semi-structured data, and a warehouse tier for structured data. All tiers are implemented as thin administrative layers on top of existing cloud services.

The Triage tier supports ingest of raw files that may exhibit unfamiliar format, structure, content, or quality. This tier provides no analysis capabilities, and essentially offers only secure and policy-protected storage. The Triage tier is implemented as a policy, authentication, and auditing layer over Azure Blob storage, using

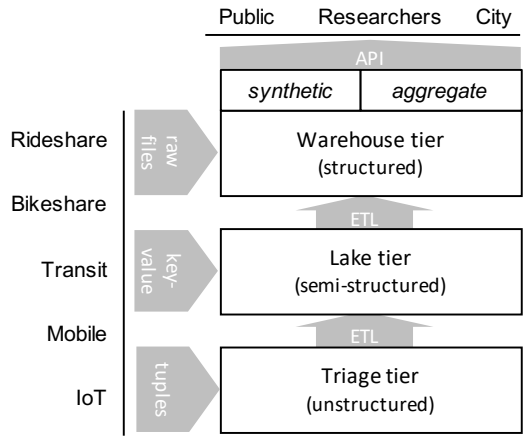

Figure 3: A three-tier architecture to capture heterogeneous data sources: raw files, semi-structured data, and structured relations.

shared access signatures (SAS) to mediate access to Azure and provide implementation transparency.

The Lake tier offers scalable semi-structured query for quality control, analysis, restructuring, and integration. Data is ingested either directly from providers (in "easy" cases where providers can conform to our requirements) or from Extract Transform Load (ETL) processes from the Triage tier. Data is assumed to be in semi-structured json format, following current best practices. The Lake is implemented using the Azure service CosmoDB.

The Warehouse tier supports structured data management and enforces integrity constraints. Data may be ingested directly into this tier via managed APIs, or may be produced from ETL processes from the Lake tier. Most reports and data products, including those based on synthetic datasets, will be derived from data in the Warehouse tier to ensure that quality assumptions are met. Synthetic datasets will typically be managed in the warehouse tier. The warehouse tier will be implemented using a relational database, building on existing capabilities [30].

Access to data can be customized and controlled for each stakeholder, allowing free access to their own data, restricted access to certain synthetic datasets in the Warehouse, and even access to original data from other stakeholders given the proper data sharing agreements. This access structure also allows for the creation of an auditable log of data interactions for full transparency.

7 CONCLUSION

Whereas computational approaches tend to define fairness, accountability, and transparency in technical terms, we argue for the need for these approaches to be closely coupled with data governance frameworks. This sociotechnical approach to the problem embraces law, policy, and practice as instrumental to promoting privacy protection, transparency, and accountability in high-dimensional data.

Fairness, accountability, and transparency issues typically arise from data management or mismanagement. We use terms like "algorithmic transparency," yet the algorithm is rarely the source of opacity. The open v. closed paradigm of public data sharing leads to convenience sampling effects that undermine a broader agenda of fairness, accountability, and transparency. Our response is to design a sociotechnical system that incentivizes and regulates sharing certain sensitive data that cannot be fully opened due to privacy and proprietary interests, but must still be made available to policymakers and researchers to assess equity, optimize the delivery of public services, and enforce the law.

Our approach has been to assume "synthetic by default," using new methods for removing unwanted bias and proprietary information from datasets prior to release, and combining these methods with differential privacy techniques. When these synthetic datasets are insufficient for the analysis, we invoke structured data use agreements backed by strong governance. With synthetic data, we can engage those researchers reluctant to sign legal agreements during the pilot phase of a project, allowing them to test their methods before committing to stronger governance policies. Meanwhile, we limit the risk surface for privacy and discrimination violations by controlling who has access to the real data, and for what purpose.

The data trust is a first step. Its limitations include balancing research with pre-approved uses of participating firms, the need to communicate how synthetic dataset products can be used appropriately, and unresolved challenges in adapting differential privacy for real-world use. Overall, we find this legal-technical approach can be used in a variety of sectors to facilitate public-private partnerships around algorithmic decision-making over sensitive data.

REFERENCES

[1] Promoting public bike-sharing: A lesson from the unsuccessful Pronto system - ScienceDirect.
[2] Progressive animal welfare society, respondent, v. the university of washington, 125 wn.2d 243, paws v. uw. http://courts.mrsc.org/supreme/125wn2d/125wn2d0243.htm, 1994. (Accessed on 08/23/2018).
[3] Amazon doesn't consider the race of its customers. should it? *Bloomberg*, 2016.
[4] Lyft, inc. v. city of seattle. http://www.courts.wa.gov/opinions/pdf/940266.pdf, 2018. (Accessed on 08/23/2018).
[5] N. Anderson, A. Abend, A. Mandel, E. Geraghty, D. Gabriel, R. Wynden, M. Kamerick, K. Anderson, J. Rainwater, and P. Tarczy-Hornoch. Implementation of a de-identified federated data network for population-based cohort discovery. *J. Am. Med. Inform. Assoc.*, 26, 2011.
[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, May 23, 2016.
[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica, May*, 23, 2016.
[8] S. Barocas and H. Nissenbaum. Big data's end run around procedural privacy protections. *Communications of the ACM*, 57(11):31–33, 2014.
[9] S. Barocas and A. Selbst. Big data's disparate impact. *California Law Review*, 2016.
[10] V. Bindschaedler, R. Shokri, and C. A. Gunter. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment*, 10(5):481–492, 2017.
[11] A. Blok, C. Marquet, A. Courmont, K. Minor, M. Young, R. Hoyng, and C. Nold. Data Platforms and Cities . *Tecnoscienza. Italian Journal of Science & Technology Studies*, 2017.
[12] R. Brauneis and E. P. Goodman. Algorithmic transparency for the smart city. *Yale Journal of Law & Technology*, forthcoming.
[13] A. M. Brock, J. E. Froehlich, J. Guerreiro, B. Tannert, A. Caspi, J. Schöning, and S. Landau. Sig: Making maps accessible and putting accessibility in maps. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page SIG03. ACM, 2018.
[14] L. Caggiani, R. Camporeale, M. Ottomanelli, and W. Y. Szeto. A modeling framework for the dynamic management of free-floating bike-sharing systems. *Transportation Research Part C: Emerging Technologies*, 87:159–182, Feb. 2018.
[15] J. M. Corbin and A. Strauss. *Unending work and care: Managing chronic illness at home.* Jossey-Bass, 1988.
[16] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE SP*, pages 598–617, 2016.

[17] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1):92–112, 2015.

[18] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[19] J. Denis and S. Goëta. Exploration, extraction and 'rawification'. the shaping of transparency in the back rooms of open data. 2014.

[20] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[21] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284. Springer, 2006.

[23] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

[24] B. Ferris, K. Watkins, and A. Borning. Onebusaway: results from providing real-time arrival information for public transit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1807–1816. ACM, 2010.

[25] E. Fishman. Bikeshare: A Review of Recent Literature. *Transport Reviews*, 36(1):92–113, Jan. 2016.

[26] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, Paderborn, Germany, September 4-8, 2017*, pages 498–510, 2017.

[27] Y. Ge, C. R. Knittel, D. MacKenzie, and S. Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[28] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032, 2010.

[29] R. Hughes and D. MacKenzie. Transportation network company wait times in greater seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, 56:36–44, 2016.

[30] S. Jain, D. Moritz, D. Halperin, B. Howe, and E. Lazowska. Sqlshare: Results from a multi-year sql-as-a-service experiment. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 281–293. ACM, 2016.

[31] D. Jorge and G. Correia. Carsharing systems demand estimation and defined operations: a literature review. *European Journal of Transport and Infrastructure Research*, 13(3):201–220, 2013.

[32] M. Kent and A. Karner. Prioritizing low-stress and equitable bicycle networks using neighborhood-based accessibility measures. *International Journal of Sustainable Transportation*, 0(0):1–11, Mar. 2018.

[33] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[34] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[35] C. Kontokosta, B. Hong, and K. Korsberg. Equity in 311 Reporting: Understanding Socio-Spatial Differentials in the Propensity to Complain. *arXiv:1710.02452 [cs]*, Oct. 2017. arXiv: 1710.02452.

[36] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[37] J. Lee and C. Clifton. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[38] H. Li, L. Xiong, L. Zhang, and X. Jiang. Dpsynthesizer: differentially private data synthesizer for privacy preserving data sharing. *Proceedings of the VLDB Endowment*, 7(13):1677–1680, 2014.

[39] Y. Li, W. Y. Szeto, J. Long, and C. S. Shui. A multiple type bike repositioning problem. *Transportation Research Part B: Methodological*, 90:263–278, Aug. 2016.

[40] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 468–477. IEEE Press, 2017.

[41] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[42] S. Ma, Y. Zheng, and O. Wolfson. Real-time city-scale taxi ridesharing. 27:1782–1795, 07 2015.

[43] X. Meng, H. Li, and J. Cui. Different strategies for differentially private histogram publication. *Journal of Communications and Information Networks*, 2(3):68–77, 2017.

[44] J. Metcalf and K. Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):205395171665021l, 2016.

[45] MetroLab Network. First, do no harm: Ethical guidelines for applying predictive tools within human services. http://www.alleghenycountyanalytics.us/, 2018. [forthcoming].

[46] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[47] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[48] A. Noriega-Campero, A. Rutherford, O. Lederman, Y. A. de Montjoye, and A. Pentland. Mapping the privacy-utility tradeoff in mobile phone data for development. *arXiv preprint arXiv:1808.00160*, 2018.

[49] A. Pal and Y. Zhang. Free-floating bike sharing: Solving real-life large-scale static rebalancing problems. *Transportation Research Part C: Emerging Technologies*, 80:92–116, July 2017.

[50] H. d. S. Pinto, F. Bernardini, and J. Viterbo. How cities categorize datasets in their open data portals: an exploratory analysis. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, page 25. ACM, 2018.

[51] L. Rayle, D. Dai, N. Chan, R. Cervero, and S. Shaheen. Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco. *Transport Policy*, 45:168–178, 2016.

[52] H. W. Rittel and M. M. Webber. Wicked problems. *Man-made Futures*, 26(1):272–280, 1974.

[53] L. Rodriguez, B. Salimi, H. Ping, J. Stoyanovich, and B. Howe. MobilityMirror: Bias-Adjusted Transportation Datasets. *arXiv:1808.07151 [cs]*, Aug. 2018. arXiv: 1808.07151.

[54] E. Sedenberg and A. L. Hoffmann. Recovering the history of informed consent for data science and internet industry research ethics. *arXiv preprint arXiv:1609.03266*, 2016.

[55] K. A. Stephens, N. Anderson, C.-P. Lin, and H. Estiri. Implementing partnership-driven clinical federated electronic health record data sharing networks. *International journal of medical informatics*, 93:26–33, 2016.

[56] K. A. Stephens, N. Anderson, C.-P. Lin, and H. Estiri. Implementing partnership-driven clinical federated electronic health record data sharing networks. *International Journal of Medical Informatics*, 93:26–33, 2016.

[57] L. Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, 2013.

[58] N. Tkacz. From open source to open government: A critique of open politics. *Ephemera: Theory & politics in organization*, 12(4), 2012.

[59] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando. Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, 33(2):325–337, 2016.

[60] J. Whittington, R. Calo, M. Simon, J. Woo, M. Young, and P. Schmiedeskamp. Push, pull, and spill: A transdisciplinary case study in municipal open government. *Berkeley Tech. LJ*, 30:1899, 2015.

[61] Y. Xiao, L. Xiong, L. Fan, and S. Goryczka. Dpcube: differentially private histogram release through multidimensional partitioning. *arXiv preprint arXiv:1202.5358*, 2012.

[62] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett. Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822, 2013.

[63] A. Yan and N. Weber. Mining open government data used in scientific research. In *International Conference on Information*, pages 303–313. Springer, 2018.

[64] M. Young and A. Yan. Civic hackers' user experiences and expectations of seattle's open municipal data program. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[65] H. Yu and D. G. Robinson. The new ambiguity of open government. *UCLA L. Rev. Discourse*, 59:178, 2011.

[66] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):25, 2017.

[67] Y. Zhang, T. Thomas, M. Brussel, and M. van Maarseveen. Expanding bicycle-sharing systems: lessons learnt from an analysis of usage. *PLoS one*, 11(12):e0168604, 2016.