

Hidden Errors in Big Data: The Case of Property Records

Evelyn Smith¹, Emma Harvey², Jacob S. Goldin³, Daniel E. Ho⁴

¹American Bar Foundation

²Cornell Tech

³University of Chicago, co-equal supervising author

⁴Stanford University, co-equal supervising author

esmith@abfn.org, evh29@cornell.edu, jsgoldin@uchicago.edu, deho@law.stanford.edu

Abstract

Big data are the foundation for an increasing share of academic research and AI models deployed in both the public and private sectors, prompting substantial growth over time in reliance on brokered datasets. Brokered property records, which are ubiquitous in studies of gentrification, inequality, and the property tax in the U.S. and serve as inputs to property valuation models, are one notable example. In this paper, we audit two prominent brokered property datasets, finding errors in these data which bias key measures of economic inequality. First, we document that for 1-2% of matched sales in Cook County, IL, from 2018-2021, broker-provided sale prices differ from ground truth sale prices by more than 5%. Moreover, missing data and conceptual differences in the reporting of deed and property characteristics lead to coverage errors ranging from 12 to 15% of transactions. Second, we show that misreporting is highly consistent between brokers: more often than not, brokers make identical reporting errors for the same transactions. Third, to illustrate the significance of these errors, we measure their impact on estimates of property tax regressivity, finding that they drive significant wedges between estimates depending on the data source. These findings generalize to two other large counties in the U.S., and highlight the crucial importance of open administrative data and transparency from brokers regarding data provenance and lineage.

1 Introduction

Big data now comprise a substantial share of the evidence base for disciplines ranging from medicine and genomics to climatology and the social sciences, while also serving as an important empirical basis for policymaking, law enforcement, and the machine learning systems that increasingly mediate government decisionmaking (Martin-Sanchez and Verspoor 2014; O’Driscoll, Daugelaite, and Sleator 2013; Mangal, Rajesh, and Misra 2020; Einav and Levin 2014; Burrows and Savage 2014; Lane 2016; Coulthart and Riccucci 2022; Hu 2015; Office of the Director of National Intelligence 2024; Wade 2023). While large, complex datasets have advanced our understanding of important topics, their scale and volume complicate their growing role in scientific research and public policy, making it difficult

for end users to validate these data for accuracy and representativeness. These challenges reinforce a broader concern that big data yield estimates which are simultaneously precise and misleading—a tension which Meng (2018) terms the “Big Data Paradox.” As governments adopt big data and machine learning tools for tasks ranging from law enforcement to property assessment, data quality becomes an essential determinant of system outcomes and thus an important policy concern; indeed, as Sambasivan et al. (2021) note, data errors cascade through AI systems, making these systems harder to maintain, harming the individuals these systems serve, and eroding institutional trust.

Accounting for big data’s provenance and lineage is a particular challenge (Wang et al. 2015; Chacko and Kumar 2017; Appelbaum 2016). This is especially true for records sourced from data brokers, who make crucial and often opaque decisions regarding how to source, clean, impute, and formalize their data. These decisions can introduce consequential errors if brokers disregard subtleties or key institutional details in their data sources (Zook and Spangler 2023; Donaldson et al. 2024). Moreover, data brokers frequently sell data to one another, creating opportunities for errors to propagate across networks of brokers (FTC 2014). These limitations are particularly alarming in light of the growing literature on data documentation for machine learning and pipeline-aware bias mitigation, which argue that downstream model behavior cannot be properly accounted for without careful scrutiny of upstream data practices (Gebu et al. 2021; Pushkarna, Zaldivar, and Kjartansson 2022; Black et al. 2023; Suresh and Guttat 2021).

In this paper, we audit the accuracy and coverage of two prominent property data brokers whose datasets are widely used in government and academic research: Cotality and ATTOM. Both brokers advertise comprehensive or near-comprehensive coverage of properties, mortgages, sales, and foreclosures across the U.S., and their data are ubiquitous in research on housing markets, as shown in Online Appendix Table A1.¹² Cotality and ATTOM also play a growing role in housing policy and property assessment.

¹ATTOM claims their data cover more than 158 million parcels; Cotality advertises 99.9% market coverage.

²Our online appendix can be found at <https://arxiv.org/abs/2607.28827>

ATTOM, for example, has contracted with the Department of Housing and Urban Development to provide foreclosure and sales data to inform agency policy, while both Cotality and ATTOM provide data and software, including ML-based valuation models, that counties can use to perform assessments (San Mateo County Board of Supervisors 2024; Cotality, Inc. 2026; U.S. Department of Housing and Urban Development 2024). These brokers are central to the administration of private-sector systems as well, such as mortgage financing and private appraisals. Cotality and ATTOM source their data in part from small units of local government, including county recorders and assessors, who number in the thousands and have diverse standards and practices for data reporting. To produce usable property and transaction data at the scale that they advertise, brokers must encode key concepts such as foreclosures and arms-length sales across tens of millions of inconsistent records—a laborious undertaking for any institution. However, as brokers generally do not publicize their data lineage, end users cannot account for the accuracy of these encodings or for errors that occur as concepts are harmonized across jurisdictions, even though these errors may have significant consequences for model validity (Jennifer, RES et al. 2021).

Our audit identifies substantial coverage and imputation errors in Cotality and ATTOM relative to administrative data for three jurisdictions and a sample period covering 2018-2021. After matching transactions of single-family homes between brokered and administrative data, we find that broker-reported sale prices substantially differ from county-reported prices for 1-2% of matched sales. These differences are on the order of hundreds of thousands of dollars per transaction, and are likely attributable to broker error in the process of imputing sale prices from transfer tax payments. Moreover, we identify coverage error ranging from 12-15% of ground truth transactions, largely owing to missing data and broker misreporting of filters used to subset to arms-length transactions of single-family homes. We find these errors significantly affect broker-derived measures of property tax assessment regressivity, a prominent metric in ongoing academic and policy debates.

We provide an overview of the data in Section 3. In Section 4, we identify the sources and extent of error separately for ATTOM and Cotality. Section 5 compares errors across brokers, examining to what extent these errors overlap, and Section 6 examines the implications of these errors for property tax assessment regressivity estimates.

2 Background and Related Work

Our work contributes to a landscape of responsible AI research that considers “data-centric factors” (Li, Goel, and Ash 2022) in AI development and evaluation, with a particular focus on how transparent data practices can promote accountability for data owners and data users (Chappidi et al. 2025). In line with prior work in this space, we interrogate large-scale, widely-used datasets (Agnew et al. 2025) in order to make the practices underlying data development clear (Rothschild et al. 2024; Zajac et al. 2023).

More specifically, our work contributes to a prior literature examining the impact of sampling, coverage,

and imputation errors on estimates of significant social importance derived from big data, including vaccine uptake (Bradley et al. 2021), influenza incidence (Lazer et al. 2014), electoral polling (Meng 2018), and the relationship between cardiovascular disease and COVID-19 mortality (Mehra et al. 2020). We also build on previous research auditing brokered datasets which has identified considerable discrepancies across brokers in the quality and coverage of their data, particularly in the context of mobility data (Hsu et al. 2024; Noi, Rudolph, and Dodge 2022). The most closely related body of research this paper examines the challenges of standardizing large-scale real estate microdata (Nolte et al. 2024; Brummet et al. 2016; Seeskin 2018). Consistent with the findings we present here, these prior papers identify substantial discrepancies between broker-provided and administrative records for features including home ownership data, mortgage data, and property tax liabilities, the scope and magnitude of which vary across geographies, and which are attributable to conceptual differences across jurisdictions and between brokers and governments in data collection and reporting. Our paper is the first to compare distinct brokered property databases to one another, identifying substantial similarities in errors made and measuring the impact of these errors on estimates of key statistics.

3 Data

Our data comprise the universe of arms-length, single-family home transactions in Cook County, IL listed in Cotality, ATTOM, and Cook County’s open data repository for a period covering 2018-2021.³ In each case, we exclude transactions with sale prices less than \$10K, as well as multi-parcel sales and properties which transacted more than once in a given year, which may include duplicate sales and recording errors.⁴ We also exclude transactions that are missing address, sale date, or sale amount information. Our filtered Cook County data contain 152,060 transactions from ATTOM, 147,731 transactions from Cotality, and 153,044 transactions from the ground-truth data for the whole of our sample period. Full details regarding the filters applied to the Cook County open data are provided in Online Appendix C.

Treating county data as ground truth. Throughout, we refer to county data as “ground truth,” as these data are derived directly from county administrative records and form the basis of brokered datasets. Even so, administrative records may contain errors, and discrepancies between brokered and administrative data may arise if brokers correct these errors while processing and standardizing records across jurisdictions. To support our interpretation of county data as ground truth, we manually verify broker and county prices against deed records published by the Cook

³Our analysis is focused on Cook County because most jurisdictions do not provide property data that is publicly available for download at scale. We use comparable data from New York City, NY, and Philadelphia, PA in robustness checks in Online Appendix B.

⁴A “multi-parcel sale” is a sale involving more than one property. These sales are excluded because the recorded price reflects the value of multiple properties.

County Clerk for a random 2% sample of transactions where the two sources disagree. For discrepancies with Cotality, the deed price matches Cook County's record 89.2% of the time, yielding an overall County accuracy of 99.8%.⁵ Similarly, for discrepancies with ATTOM, the deed price matches Cook County's record 97.2% of the time, yielding an overall County accuracy of 99.9%.⁶ Altogether, for the vast majority of disagreements, the county price aligns with the most recent underlying deed record, confirming that discrepancies between sources are predominantly attributable to broker processing rather than county-level errors. While a small number of cases do favor brokered data—suggesting that brokers may occasionally correct genuine administrative mistakes—such instances are rare enough that they are unlikely to materially affect our analysis.

Data appear comparable in aggregate. Table 1 and Figure 1 compare the distributions of price, assessed value, and assessment ratios across data sources for Cook County. In aggregate, the three data sources appear comparable: there are no visible differences between the densities, apart from a slight leftward skew in Cotality's sale price distribution and a corresponding rightward skew in its distribution of assessment ratios. However, as shown in subsequent sections, these summary distributions belie significant and consequential errors at the transaction level.

4 Coverage and Imputation Error in Brokered Data

We identify two types of consequential errors in brokered data. First, we find evidence of *imputation errors*. Imputation errors are issues of data accuracy, where records representing the same property transaction have different recorded sale prices in brokered data as compared to ground truth data, likely representing errors in how data brokers impute sale prices that are not directly available to them. Second, we find evidence of *coverage errors*. Coverage errors are issues of data representativeness, attributable to missing data or differences in how deed and property characteristics are reported, such that transactions recorded as arms-length sales of single-family homes in ground truth data are not recorded as such in brokered data.

4.1 ATTOM

We match transactions between ATTOM and ground-truth data to assess the coverage and accuracy of the ATTOM data. Our match process is described in Online Appendix D.

⁵This represents 33 of 37 cases. Of the remaining four cases, three match Cotality's data, and one involves two deeds issued several days apart—one consistent with the county's price and one with Cotality's.

⁶This represents 36 out of 37 cases. In the remaining case, the deed price matches ATTOM's data, while Cook County's price corresponds to the mortgage amount. In two instances, disagreement between ATTOM and county data arose due to corrections made after deeds were published. In these cases, ATTOM's price reflected the original price, while Cook County's data matched the corrected price, meaning that ATTOM data was more stale than County data and was likely not updated after initial ingestion.

As shown in Table 2, we are able to match 134,427 ATTOM transactions to our ground-truth data, comprising 87.8% of ground-truth sales. Among matched transactions, we find a total of 1,976 transactions that have discrepancies in sale price which exceed 5% of the ground-truth price. As shown in Table 3, the overwhelming majority of these discrepant transactions can be represented as simple ratios of broker-reported to ground-truth sale price. The most common ratios of ATTOM to ground-truth sale prices are $\frac{2}{3}$, $\frac{1}{3}$, and 2. While we also found instances of dropped, transposed, or otherwise mistranscribed digits,⁷ for the majority of cases, the discrepancies do not appear to be the result of data entry error.

Imputation error can potentially explain the pattern of discrepancies we observe. In discussions with one of the two brokers examined here, we confirmed that when sale price is not directly available from county-level records sourced through FOIA requests, it may be imputed using local transfer tax rates. These transfer taxes are typically structured as a fixed percentage tax on a property's purchase price with de minimus exemptions. In Cook County, there are 75 distinct municipalities that independently set transfer tax rates and can also divide responsibility for remitting the tax between the buyer and seller.⁸ In the event that a broker imputes a property's sale price using the incorrect municipality's transfer tax rate, or confuses the buyer's share of the tax for the seller's share, then the resulting imputed price will be an exact multiple of the actual price.

An additional 371 matched ATTOM transactions have discrepancies in assessed value that exceed 5% of ground-truth. These discrepancies do not follow a clear pattern, such as substitution between pre- and post-appeal assessed values. Just 6 instances are attributable to rounding error. Assessed value errors in ATTOM's data are far more limited in scope and magnitude than sale price errors.

We next turn to the portion of ground-truth transactions that we are unable to match to ATTOM's data, leading to coverage error between the two datasets. Ground-truth transactions generally fail to match to brokered data for one of three reasons. First, brokers may misreport sale and property characteristics, labeling an arms-length transaction as a foreclosure or a single-family home as a condominium. Since we filter each dataset to include only arms-length transactions of single-family homes, any transaction with a misreported filter will not match to its corresponding ground-truth sale. Next, a transaction may be completely missing or duplicated in the brokered data, in which case it will be dropped or missing prior to the match. Finally, transactions may not match due to the address and sale date similarity thresholds in our match procedure, which could exclude valid matches if too strict.

⁷Examples from the data: "\$7,000,000" instead of "\$700,000" (dropped digit), "\$200,000" instead of "\$220,000" (mistranscribed digit), and "\$271,000" instead of "\$217,000" (transposed digit).

⁸A full list of transfer taxes across municipalities in Cook County as of January 2025 is available at <https://www.cookcountyclerk.kil.gov/sites/default/files/pdfs/Transfer%20Tax%20By%20Municipality%2C%20January%202025a.pdf>.

	Ground Truth	ATTOM	Cotality
N	153,044	152,060	147,731
Mean Sale Price	\$350,405	\$356,661	\$367,500
Sale Price P1	\$27,342	\$25,000	\$45,000
Sale Price P25	\$181,000	\$186,000	\$195,000
Sale Price P50	\$270,000	\$275,000	\$280,000
Sale Price P75	\$400,000	\$405,000	\$412,000
Sale Price P99	\$1,650,000	\$1,680,000	\$1,720,000
Mean Assessed Value	\$30,059	\$29,718	\$30,213
Assessed Value P1	\$4,549	\$4,499	\$4,499
Assessed Value P25	\$15,243	\$15,359	\$15,669
Assessed Value P50	\$22,774	\$22,799	\$23,000
Assessed Value P75	\$34,079	\$33,917	\$34,091
Assessed Value P99	\$141,517	\$137,475	\$139,278
Mean Ratio	0.10	0.10	0.09
Ratio P1	0.03	0.03	0.03
Ratio P25	0.07	0.07	0.07
Ratio P50	0.09	0.08	0.08
Ratio P75	0.10	0.10	0.10
Ratio P99	0.32	0.34	0.25

Table 1: **Comparison of Sale Prices, Assessed Values, and Assessment Ratios Between Ground-Truth and Broker-Provided Data for Cook County, 2018–2021.** The table displays summary statistics (counts, means, and percentile values) for each of the Cook County (ground truth), ATTOM, and Cotality datasets for arms-length, single-family home transactions from 2018 to 2021. “Ratio” indicates the ratio between the recorded assessed value and property sale price. “P#” indicate percentiles.

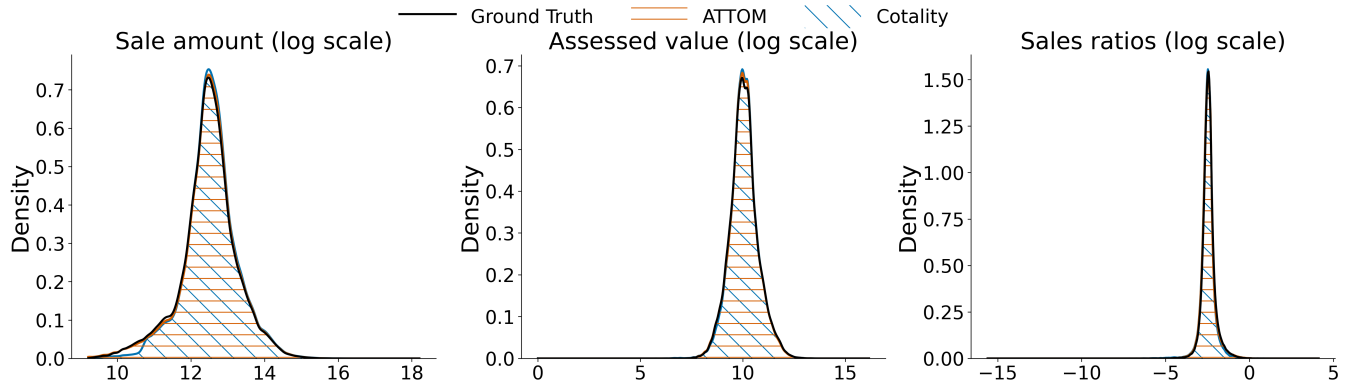


Figure 1: **Distribution of Sale Price, Assessed Value, and Assessment Ratios by Data Source, 2018-2021.** The figure shows the distributions of sale price, assessed value, and assessment ratios by data source regressivity measures by year and data source for a pooled sample of single-family home sales in Cook County covering 2018-2021. “Ground truth” values are derived from data published by the Cook County Assessor’s office.

Table 4 provides a breakdown of the causes of match failure for ATTOM. About half of match failures occur due to broker misreporting of sale and property characteristics. Another third are accounted for by missing records, while an additional 9% are attributable to duplicate sales. The remaining 5% of failed matches are largely due to the thresholds for address and sale month similarity in our matching process. We examine the robustness of our results to different choices of matching thresholds in Online Appendix D1.

4.2 Cotality

Using the same match procedure applied above and described in Online Appendix D, we are able to match

129,606 transactions between Cotality and Cook County, comprising 84.7% of all ground-truth transactions. Among these transactions, we find 1,919 transactions where the sale price listed in Cotality differs from the ground-truth price by more than 5%. Table 5 provides a summary of these discrepancies by year of sale.

Similar to ATTOM, we find that for the overwhelming majority of transactions with sale price discrepancies, the sale prices recorded by Cotality are exact multiples of ground-truth sale prices. And as with ATTOM, the most common ratios of Cotality to ground-truth sale prices are $\frac{2}{3}$, $\frac{1}{3}$, and 2 as shown in Table 6, indicating that Cotality’s data also plausibly suffers from imputation error.

	2018	2019	2020	2021
N matched	32,656	31,935	34,315	35,521
Matched as % of ground-truth sales	86.08%	86.91%	87.36%	90.90%
Share with sale price error >5%	1.27%	0.74%	3.70%	0.16%
Conditional mean error (%)	260.80%	79.33%	63.99%	78.97%
	(137.63% - 383.97%)	(65.29% - 93.37%)	(59.06% - 68.92%)	(44.28% - 113.66%)
Conditional mean error (\$)	\$398,934	\$166,999	\$200,411	\$162,269
	(\$245,412 - \$552,457)	(\$138,619 - \$195,380)	(\$187,999 - \$212,823)	(\$125,376 - \$199,162)
Share with assessed value error >5%	0.02%	0.02%	0.06%	0.95%
Conditional mean error (%)	58.13%	35.54%	53.65%	13.57%
	(-26.69% - 142.94%)	(9.77% - 61.30%)	(-14.88% - 122.19%)	(12.73% - 14.42%)
Conditional mean error (\$)	\$18,383	\$13,300	\$33,008	\$4,292
	(\$4,273 - \$32,494)	(\$-3,588 - \$30,187)	(\$-19,923 - \$85,940)	(\$4,010 - \$4,573)

Table 2: Discrepancies between ATTOM and Ground Truth Sale Prices and Assessments in Cook County, 2018-2021. The table displays summary statistics of the discrepancies between ATTOM and ground-truth transaction data for sales of single-family homes in Cook County from 2018–2021. “Error” is calculated as the absolute difference between the data broker and ground truth transaction price or assessed value as a fraction of the ground truth price or value. Conditional means are computed using only matched transactions with error >5%. Records were matched using property address, latitude, longitude, and sale date. 95% confidence intervals are displayed in parentheses.

Ratio	Count	Share of Discrepancies (%)
0.667 (2/3)	998	50.51
2.000	152	7.69
0.333 (1/3)	116	5.87
0.666 (2/3)	95	4.81
0.833 (5/6)	28	1.42
0.665 ($\approx$ 2/3)	18	0.91
0.250 (1/4)	15	0.76
10.000	13	0.66
1.999 ($\approx$ 2)	11	0.56
1.133 (17/15)	10	0.51
Other (random)	432	21.86
Other (simple ratio)	88	4.45
Total	1,976	100.00%

Table 3: Distribution of Ratios of ATTOM to Ground Truth Sale Prices for Cook County Discrepant Sales, 2018-2021. The table shows the distribution of ratios of ATTOM-recorded sale prices to ground truth prices for matched discrepant sales in Cook County covering 2018-2021. A “discrepant sale” is any sale where the absolute difference between the broker and ground-truth sale price exceeds 5% of the ground-truth price.

An additional 352 matched Cotality transactions have discrepancies in assessed value that exceed 5% of ground-truth. Similar to ATTOM’s assessed value discrepancies, these discrepancies do not follow a clear pattern, such as substituting between pre- and post-appeal assessed values. Just 6 instances are attributable to rounding error. As with ATTOM, assessed value errors in Cotality’s data are far more limited in scope and magnitude than sale price errors.

Table 7 provides a breakdown of the causes of match failure for Cotality. In contrast to ATTOM, misreported sale and property characteristics explain a smaller share of match failures (32% versus 50%), while missing sale records explain comparatively larger share (56% versus 37%). Dupli-

N ground truth sales	153,044
N sales not matched to ATTOM	18,617
Reason for match failure:	
Record missing	36.8%
Matches duplicate sale	8.6%
Broker filter misreported	49.8%
Arms-length sale	38.0%
Not multiparcel	2.6%
Sale price > \$10k	13.5%
Single-family home	13.5%
Does not meet fuzzy match criteria	4.0%
Other	0.9%
Total	100.0%

Table 4: Sources of coverage error: ATTOM. The table displays the sources of coverage error between ground truth and ATTOM data for arms-length sales of single family homes in Cook County from 2018-2021.

cate sales in brokered data explain about 9% of match failures, and our choice of matching thresholds account for most of the remaining 3% of unmatched transactions.

5 Shared Errors Across Brokers

One of the most common approaches to assessing robustness to data errors is a sensitivity analysis to different data sources (Hsu et al. 2024; Desai et al. 2019; J’Neka et al. 2019). We therefore investigate whether reporting errors are correlated across sources at the transaction level, which would limit the diagnostic value of cross-source comparison, and find substantial dependence across data sources in errors for specific transactions. One reason for this dependence is that data brokers frequently sell and share data with one another, meaning errors can quickly propagate across networks of brokers (FTC 2014). In the context of property data, antitrust concerns specifically resulted in an FTC consent degree that required Cotality (then-CoreLogic) to share

	2018	2019	2020	2021
N matched	30,763	31,116	34,212	33,515
Matched as % of ground-truth sales	81.09%	84.68%	87.10%	85.76%
Share with sale price error >5%	1.04%	0.60%	3.71%	0.43%
Conditional mean error (%)	94.40%	89.90%	67.78%	133.24%
	(78.32% - 110.47%)	(72.26% - 107.54%)	(62.11% - 73.45%)	(93.07% - 173.41%)
Conditional mean error (\$)	\$215,846	\$156,821	\$203,642	\$431,771
	(\$168,451 - \$263,242)	(\$127,478 - \$186,164)	(\$189,252 - \$218,032)	(\$286,907 - \$576,634)
Share with assessed value error >5%	0.02%	0.00%	0.06%	0.97%
Conditional mean error (%)	67.20%	0.00%	54.48%	15.61%
	(-23.46% - 157.87%)	(0.00% - 0.00%)	(-13.88% - 122.85%)	(13.32% - 17.90%)
Conditional mean error (\$)	\$18,979	\$0	\$30,792	\$4,869
	(\$-7,826 - \$45,784)	(\$0 - \$0)	(\$-22,166 - \$83,749)	(\$4,268 - \$5,471)

Table 5: Discrepancies between Cotality and Ground Truth Sale Prices and Assessments in Cook County, 2018-2021. The table displays summary statistics of the discrepancies between Cotality and ground-truth transaction data for sales of single-family homes in Cook County from 2018–2021. “Error” is calculated as the absolute difference between the data broker and ground truth transaction price or assessed value as a fraction of the ground truth price or value. Conditional means are computed using only matched transactions with error >5%. Records were matched using property address, latitude, longitude, and sale date. 95% confidence intervals are displayed in parentheses. Conditional means are not reported for cells with zero observations.

Ratio	Count	Share of Discrepancies (%)
0.667 (2/3)	1,093	56.96
0.333 (1/3)	138	7.19
0.666 (2/3)	100	5.21
2.000	44	2.29
0.833 (5/6)	28	1.46
0.250 (1/4)	17	0.89
0.095	14	0.73
0.948	11	0.57
0.665 ($\approx$ 2/3)	10	0.52
1.133 (17/15)	9	0.47
Other (random)	359	18.71
Other (simple ratio)	96	5.00
Total	1,919	100.00%

Table 6: Distribution of Ratios of Cotality to Ground Truth Sale Prices for Cook County Discrepant Sales, 2018-2021. The table shows the distribution of ratios of Cotality-recorded sale prices to ground truth prices for matched discrepant sales in Cook County covering 2018-2021. A “discrepant sale” is any sale where the absolute difference between the broker and ground-truth sale price exceeds 5% of the ground-truth price.

data with ATTOM (then-RealtyTrac) (Lane 2018), and as such an analysis of shared errors is particularly appropriate.

First, we consider transactions where broker and county-reported sale prices differ. Out of 1,976 discrepant ATTOM transactions and 1,919 discrepant Cotality transactions, we find that 1,512 transactions overlap—meaning brokers share roughly 3 out of every 4 discrepant transactions in common. Moreover, among these 1,512 shared discrepant transactions, the broker-reported sale prices are identical in virtually all cases (99.8%).

We next turn to coverage error, considering overlap between brokers in the ground truth transactions we are unable to match to their data, as well as the overlap in underlying reasons why such transactions fail to match. The

N ground truth sales	153,044
N unmatched sales	23,438
Share of unmatched:	
Record missing	56.0%
Matches duplicate sale	8.9%
Broker filter misreported	32.0%
Arms-length sale	28.4%
Not multiparcel	3.1%
Sale price > \$10k	9.0%
Single-family home	9.0%
Does not meet fuzzy match criteria	2.3%
Other	0.7%
Total	100.0%

Table 7: Sources of coverage error: Cotality. The table displays the sources of coverage error between ground truth and Cotality data for arms-length sales of single family homes in Cook County from 2018-2021.

results of this analysis are shown in Table 8. ATTOM and Cotality share 11,094 unmatched transactions in common, representing roughly 50 to 60% of each broker’s total unmatched transactions. Among these transactions, the reasons for match failure are largely consistent across brokers. For example, one reason a transaction may fail to match is because it is missing in the broker’s data. We find that, among the subset of unmatched transactions these brokers have in common, if a transaction is missing in Cotality’s data, there is a 97% chance it is also missing in ATTOM’s data (5,432/5,615). We find similar agreement along the other dimensions we consider.

Overall, we find considerable overlap in the errors present in ATTOM and Cotality’s data. The most straightforward explanation for the similarities we observe in this particular case is the data sharing agreement between Cotality and ATTOM as mandated by the FTC’s 2018 consent decree. However, this is not the only mechanism through which errors can propagate, and we cannot rule out other potential

		ATTOM error type					Total
		Missing	Duplicate sale	Filter misreported	No fuzzy match	Other	
Cotality error type	Missing	5,432	150	625	40	4	6,251
	Duplicate sale	24	810	290	137	1	1,262
	Filter misreported	149	93	2,762	29	7	3,040
	No fuzzy match	8	1	38	377	3	427
	Other	2	1	8	2	101	114
	Total	5,615	1,055	3,723	585	116	11,094

Table 8: **Match failure concordance matrix: Cotality vs. ATTOM.** The table shows the concordance in reasons for match failure between ATTOM and Cotality among unmatched ground-truth transactions the brokers have in common for Cook County from 2018-2021. “Missing” indicates that the unmatched record is completely missing in the brokered dataset. “Duplicate sale” indicates that record is duplicated in the brokered data and therefore dropped prior to matching. “Filter misreported” indicates that one or more data filters (single-family home, arms-length transaction, sale price > \$10K, or single-parcel sale) is misreported in the brokered data. “No fuzzy match” indicates that there is a potential match for the sale in the brokered data that does not meet our fuzzy match criteria for address and sale month similarity, as described in Online Appendix D. Bolded values along the diagonal indicate cases where the reasons for match failure are the same across brokers.

explanations. For example, these brokers may rely on similar imputation and processing methods. As previously noted, it is also common for brokers to sell data to one another (FTC 2014); such transactions are rarely disclosed to end users and are subject to limited regulatory oversight (Sherman 2023). These practices pose several challenges for researchers. First, researchers cannot use one broker’s data to validate findings or correct errors in another broker’s data if both datasets suffer from the same shortcomings. Second, combining or pooling brokered datasets to improve coverage or precision may yield smaller gains than expected if coverage gaps are largely shared. Finally, seemingly independent replication across studies using different brokers may provide a false sense of robustness, since convergent findings could reflect common errors rather than unbiased agreement.

6 Consequences of Data Errors

Finally, we examine whether the coverage and imputation errors we identify above significantly affect estimates derived from brokered data using measures of property tax regressivity, a key economic statistic. Property taxes comprise a substantial share of revenues for local governments throughout the U.S. These taxes are allocated to households via property tax assessments, which are typically conducted at the county level and involve mapping property characteristics to market values using valuation tables or predictive models (Dornfest et al. 2019). While property taxes are meant to be proportional to each property’s underlying market value, several studies have demonstrated that these taxes and corresponding assessments are in fact regressive: higher-valued properties tend to be under-assessed and have lower effective tax rates than lower-valued properties (Berry 2021; Amornsiripanitch 2024; McMillen and Singh 2020; Smith et al. 2026).

Studies of assessment regressivity typically compare assessed values to sale prices among homes that sell in a given year, under the assumption that these sale prices reflect the true underlying value of the home. These studies generally rely on data either obtained from counties through FOIA requests or compiled by data brokers such as Cotality and ATTOM (Berry 2021; Amornsiripanitch

2024; Avenancio-León and Howard 2022). We estimate and compare standard regressivity measures from this literature using both broker-reported and ground-truth data. We consider the price-related differential (PRD) (IAAO 2017), the log coefficient (IAAO and Almy 1978), and the Suits Index (Suits 1977). Descriptions of each measure are provided in Online Appendix E.

Our results, presented in Figure 2, show that regressivity estimates derived from broker-reported transaction data significantly differ from ground-truth values for at least one broker in all four sample years. In particular, Cotality data tend to underestimate the level of regressivity relative to ground truth data for the log coefficient and PRD measures, while ATTOM-derived estimates generally overstate regressivity for all three metrics. Differences between broker-reported and ground-truth estimates narrow by the end of our sample period.

Both coverage and imputation error contribute to the gap between broker-reported and ground-truth measures of regressivity. To determine their relative contributions, we compare three estimates:

- a **ground-truth estimate**, computed from the full population of ground-truth data;
- an **imputation-error estimate**, computed from the same ground-truth population but replacing ground-truth prices and assessments with broker-reported prices and assessments for matched transactions; and
- a **broker estimate**, computed from the full population of broker-reported data.

The difference between the imputation-error and ground-truth estimates approximates the contribution of imputation error, while the difference between the broker and imputation-error estimates approximates the contribution of coverage error. The results of this analysis are provided in Tables 9 and 10.

We observe key differences in the relative contributions of coverage and imputation error across brokers. For ATTOM, coverage error dominates in 2018 and 2019, when differences between its estimates and ground-truth estimates are statistically significant and the rate of imputation error

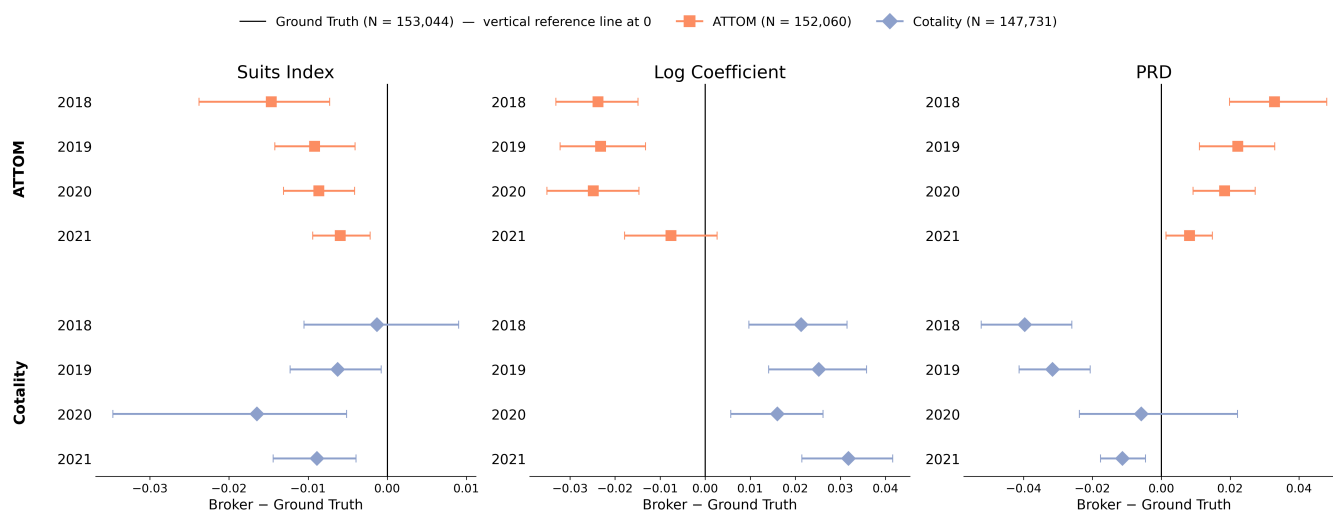


Figure 2: Difference between Ground Truth and Broker-Derived Regressivity Estimates by Year and Data Source, 2018-2021. The figure shows the differences between estimates of regressivity derived from ground-truth administrative data and brokered data by year and data source for single-family home sales in Cook County covering 2018-2021. Ground truth values are derived from data published by the Cook County Assessor’s office. 95% confidence intervals are derived from a nonparametric percentile bootstrap (1,000 iterations), with each dataset resampled independently. For the Suits Index, lower values are more regressive. For the Log Coefficient, lower values are more regressive. For the PRD, higher values are more regressive.

in the data is relatively low (~1%). In 2020, the share of transactions suffering from imputation error jumps to 3.7%, and imputation error accounts for a larger portion of the overall difference in estimates. Finally, in 2021, the differences between ATTOM and ground-truth estimates more or less disappear, and the relative shares of imputation and coverage error reflect division of near-zero differences.

Cotality’s data, by contrast, is susceptible to a higher degree of coverage error, as discussed in Section 4.2. Correspondingly, coverage error predominates in virtually all cases where the differences between ground truth and Cotality estimates are statistically significant. We observe a slight increase in the share of error explained by imputation in 2020, where Cotality’s data—similar to ATTOM’s data—exhibits a temporary spike in the overall share of transactions suffering from imputation error (from 1% to 3.7%). But overall, it is the difference in the population of arms-length transactions rather than difference in reported prices that drives most of the differences between Cotality and ground-truth estimates.

The magnitude of estimated assessment regressivity differs significantly across brokered data sources, and the direction of bias is broker-specific. Decomposing the discrepancies into their underlying sources indicates that imputation error introduces a consistent, if modest, attenuation bias: both brokers’ data initially understate the true degree of regressivity. Coverage error, by contrast, varies in sign and magnitude across brokers and years, and it dominates the total discrepancy for Cotality throughout the sample period and for ATTOM in earlier years. While ATTOM and Cotality share roughly half of their unmatched transactions, the remaining portion is large and idiosyn-

cratic enough to drive substantial wedges between their regressivity estimates. The substantial overlap in errors documented in Section 5 therefore does not translate into agreement at the level of regressivity estimates.

7 Discussion

Brokered data now comprise a substantial share of the evidence base for academic and government research and the training data for machine learning models, including models deployed in the public sector. The credibility of these endeavors crucially depends on the brokers who source, process, and publish this data, and who generally treat their data lineage as proprietary and confidential. This places users of brokered data in a difficult position. On the one hand, rising expectations—for the scope and coverage of contemporary empirical research, or the technical capacity of government—encourage reliance on brokered data. On the other hand, users cannot account for how these data were sourced and processed, and they often have no practical way to validate that the data are as accurate or comprehensive as brokers claim.

In this paper, we audit two central brokered property datasets, documenting significant shortcomings in their representativeness and accuracy. We find similar patterns of errors in each dataset and a high rates of concordance between brokers in the data reporting errors they make for specific transactions. While errors in sale price and assessed value tend to attenuate regressivity estimates, errors in data coverage bias estimates in unpredictable ways, and are responsible for a much larger portion of the bias we observe.

The types of errors our audit identifies—nonrandom label noise and nonrandom sample selection—pose particular

Panel A: Differences in Levels					Panel B: Share of Total Difference (Absolute Contributions)				
	Metric	Total Difference	Imputation Error	Coverage Error		Metric	Imputation Error	Coverage Error	
2018	Log Coefficient	0.020*	0.008	0.012	2018	Log Coefficient	39.93%	60.07%	
	PRD	-0.026*	-0.012	-0.015		PRD	43.5%	56.5%	
	Suits Index	0.009*	0.008	0.001		Suits Index	88.28%	11.72%	
2019	Log Coefficient	0.017*	0.002	0.015	2019	Log Coefficient	11.96%	88.04%	
	PRD	-0.015*	-0.002	-0.013		PRD	11.76%	88.24%	
	Suits Index	0.003*	0.001	0.003		Suits Index	23.41%	76.59%	
2020	Log Coefficient	0.022*	0.013	0.009	2020	Log Coefficient	60.09%	39.91%	
	PRD	-0.015*	-0.010	-0.005		PRD	65.54%	34.46%	
	Suits Index	0.006*	0.006	-0.000		Suits Index	99.48%	0.52%	
2021	Log Coefficient	0.003	0.003	-0.001	2021	Log Coefficient	85.2%	14.8%	
	PRD	-0.004*	-0.003	-0.001		PRD	83.66%	16.34%	
	Suits Index	0.002*	0.000	0.001		Suits Index	26.97%	73.03%	

Table 9: **Decomposition of Differences Between Ground Truth and Broker Estimates: ATTOM.** Panels A and B decompose the relative contribution of imputation and coverage error to the observed differences between ATTOM and ground-truth regressivity estimates. We derive the values in each panel using the ground-truth, imputation error, and broker estimates of regressivity derived in Section 6. Panel A shows the overall difference between ground truth and ATTOM estimates for each metric. Asterisks indicate statistical significance at the 5% level. Panel B shows the absolute percentage contribution of Imputation and Coverage error to the total differences displayed in Panel A. “Absolute percentage contribution” is measured as $|error_type|/(|imputation_error| + |coverage_error|)$.

Panel A: Differences in Levels					Panel B: Share of Total Difference (Absolute Contributions)				
	Metric	Total Difference	Imputation Error	Coverage Error		Metric	Imputation Error	Coverage Error	
2018	Log Coefficient	-0.025*	0.003	-0.028	2018	Log Coefficient	8.89%	91.11%	
	PRD	0.046*	-0.003	0.049		PRD	5.33%	94.67%	
	Suits Index	-0.004	0.002	-0.006		Suits Index	24.3%	75.7%	
2019	Log Coefficient	-0.031*	0.002	-0.033	2019	Log Coefficient	4.91%	95.09%	
	PRD	0.038*	-0.001	0.040		PRD	3.0%	97.0%	
	Suits Index	0.001*	0.000	0.000		Suits Index	85.43%	14.57%	
2020	Log Coefficient	-0.019*	0.011	-0.030	2020	Log Coefficient	26.67%	73.33%	
	PRD	0.009	-0.008	0.017		PRD	31.58%	68.42%	
	Suits Index	0.014*	0.006	0.008		Suits Index	42.64%	57.36%	
2021	Log Coefficient	-0.037*	0.005	-0.042	2021	Log Coefficient	10.74%	89.26%	
	PRD	0.016*	-0.003	0.019		PRD	13.62%	86.38%	
	Suits Index	0.005*	0.003	0.002		Suits Index	59.12%	40.88%	

Table 10: **Decomposition of Differences Between Ground Truth and Broker Estimates: Cotality.** Panels A and B decompose the relative contribution of imputation and coverage error to the observed differences between Cotality and ground-truth regressivity estimates. We derive the values in each panel using the ground-truth, imputation error, and broker estimates of regressivity derived in Section 6. Panel A shows the overall difference between ground truth and Cotality estimates for each metric. Asterisks indicate statistical significance at the 5% level. Panel B shows the absolute percentage contribution of Imputation and Coverage error to the total differences displayed in Panel A. “Absolute percentage contribution” is measured as $|error_type|/(|imputation_error| + |coverage_error|)$.

challenges for machine learning practitioners. First, noisy labels can degrade model accuracy, increase model complexity, and increase the number of training samples required to achieve reliable performance (Frénay and Verleysen 2013). Straightforward approaches to filtering errors, such as censoring or winsorizing extreme values, are unlikely to address the data issues we identify, as the vast majority of discrepant sales in both Cotality and ATTOM’s datasets are not outliers with respect to sale price, assessed value, or assessment ratios. Furthermore, unrepresentative samples can bias model predictions (Zadrozny 2004). Practitioners can of course re-weight brokered data to better reflect the population of interest; however, doing so requires access to population-level characteristics, and researchers may not think to take this step if they believe their data are already comprehensive.

Our findings underscore the importance of accessible open data from federal, state, and local governments. Federal governments, in particular, are the primary alternative to data brokers in supplying large-scale data on topics in-

cluding population dynamics, economic activity, healthcare, and the environment. In contrast to data brokers, governments do not face competitive pressure to keep their data lineage closely held, and are in fact generally required to publicly disclose how data are sampled and processed (Eurostat (European Commission) 2018; Office of Management and Budget, Executive Office of the President 2014, 2008). Researchers equipped with these disclosures can more easily identify and account for imperfections in government data as compared to brokered data, whose provenance is opaque to them. Even smaller-scale open data, like the county-level data we consider here, provide researchers the means to validate the robustness and accuracy of their findings, and in certain settings can enable researchers to adjust brokered data for measurement and sampling error using sampling weights, instrumental variables, or other statistical tools. Brokers, of course, could also disclose their data lineage so that researchers have a better understanding of how their evidence is sourced and processed prior to analysis.

What means do individuals and institutions have to correct errors in brokered data and redress harms caused by the application of such data in private and public sector systems? In California, individuals can compel brokers to fix errors and delete data upon request (California Legislature 2018). However, for an individual to exercise these rights, they must know that brokers have collected data about them or their property, know that this data is incorrect, and know that they have the right to correct errors or request deletion. While individuals may request that brokers share data they have collected about them and their properties under California and other state laws, these individuals cannot fully understand the impact of flaws in these brokers' data on their property's appraisal or tax bill without far more substantial visibility over brokered datasets and how appraisers and assessors use such data. Furthermore, for such laws to be effective, brokers must act on requests for data to be modified or deleted. In a study of data brokers registered in the state of California, van Kempen et al. (2025) find that roughly half of brokers do not respond at all to such requests (Gueorguieva et al. 2026).

Courts and administrative agencies can provide relief for egregious harms. For example, in 2025, the Federal Trade Commission issued a consent agreement addressing General Motors' collection and sale of drivers' precise geolocation data to consumer reporting agencies without drivers' express consent. The agreement ordered GM and related parties to delete the data and to stop collecting similar data in the future without clear, affirmative consent from drivers. This consent agreement had positive spillovers beyond GM in protecting consumers' right to privacy. However, such agreements are exceptional. They occur after harm has been done, and they cannot substitute for effective disclosure requirements regarding brokered data provenance and other ex ante consumer protections.

Overall, our findings underscore the importance of broker transparency regarding the conceptual and statistical methods they use for data gathering and imputation. In the absence of such transparency, end users are left in the dark regarding potential shortcomings of their data, and brokers miss out on opportunities to improve data quality. Based on our findings, we recommend that researchers and practitioners evaluate the robustness of their estimates and model performance using open, ground-truth administrative data where available. We also encourage clear standards of transparency for data brokers regarding data provenance.

Limitations. Our analysis focuses primarily on Cook County, IL. This is because Cook County is an outlier in terms of the amount and quality of property data that it makes available for researchers to download and use. To address this, we provide robustness checks for two other major U.S. cities with similar, albeit slightly less comprehensive, open data in Online Appendix B, showing that coverage and imputation errors also arise in those cities. Second, although we treat county-provided data as "ground truth", we note that, in a very small minority of cases, county-provided data contain errors that brokered data do not (Section 3). Although we believe that these instances are rare enough that they do not materially affect our analysis, we

note that future audits of brokered data should continue to validate county-provided data against administrative records directly. Finally, without access to data brokers' documentation of their data provenance and processing steps, we cannot definitively verify our hypotheses about the potential causes of the coverage and imputation error that we identify.

8 Conclusion

Through an audit of two prominent brokered property datasets against county-provided "ground truth" data on sales of single-family homes from 2018 to 2021 in Cook County, IL, we find evidence of coverage and imputation errors which bias key measures of economic inequality. Coverage error, which affects 12 to 15% of transactions, occurs when missing data and conceptual differences in the reporting of deed and property characteristics cause transactions recorded as arms-length sales of single-family homes in ground truth data are not recorded as such in brokered data. Imputation error, which affects 1-2% of matched transactions, occurs when records representing the same property transaction have different recorded sale prices in brokered data as compared to ground truth data, likely representing errors in how data brokers impute sale prices that are not directly available to them. We show that misreporting is highly consistent between brokers, suggesting that researchers cannot use one broker's data to validate findings or correct errors in another broker's data. Our findings highlight the crucial importance of open administrative data and transparency from brokers regarding data provenance and lineage.

Ethical Considerations

Although all of our data are based on public records, brokered data do include personally identifiable information, including the names and contact information of individuals involved in property transactions. In order to respect the privacy of those individuals, we did not use those fields in our analysis; further, we accessed brokered data only via a secure server.

Adverse Impacts

The primary potential adverse impact of this work is that it runs the risk of providing "free labor" for data brokers by identifying issues in their data on their behalf so that they can be corrected. While we do acknowledge this as a risk, our hope is that this work instead draws attention to errors in brokered data, encouraging further audits, more transparency from brokers, and the wider release of administrative data.

References

- 2014. Data Brokers: A Call for Transparency and Accountability. Technical report, Federal Trade Commission.
- Agnew, W.; Barnett, J.; Chu, A.; Hong, R.; Feffer, M.; Net-zorg, R.; Jiang, H. H.; Awumey, E.; and Das, S. 2025. Sound Check: Auditing Recent Audio Dataset Practices. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 26–40.

- Amornsiripanitch, N. 2024. Why are residential property tax rates regressive? Available at SSRN 3729072.
- Appelbaum, D. 2016. Securing big data provenance for auditors: The big data provenance black box as reliable evidence. *Journal of emerging technologies in accounting*, 13(1): 17–36.
- Avenancio-León, C. F.; and Howard, T. 2022. The assessment gap: Racial inequalities in property taxation. *The Quarterly Journal of Economics*, 137(3): 1383–1434.
- Berry, C. R. 2021. Reassessing the property tax. Available at SSRN 3800536.
- Black, E.; Naidu, R.; Ghani, R.; Rodolfa, K.; Ho, D.; and Heidari, H. 2023. Toward operationalizing pipeline-aware ML fairness: A research agenda for developing practical guidelines and tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11.
- Bradley, V. C.; Kuriwaki, S.; Isakov, M.; Sejdinovic, D.; Meng, X.-L.; and Flaxman, S. 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890): 695–700.
- Brummet, Q. O.; Cronkite, D.; Flanagan-Doyle, D.; and Rinz, K. 2016. Analysis Results: A Comparison of Consumer Expenditure Survey Data with CoreLogic Property Tax and Deeds Data. Technical report, Center for Administrative Records Research and Applications, U.S. Census Bureau.
- Burrows, R.; and Savage, M. 2014. After the crisis? Big Data and the methodological challenges of empirical sociology. *Big data & society*, 1(1): 2053951714540280.
- California Legislature. 2018. California Consumer Privacy Act. Cal. Civ. Code tit. 1.81.5. https://leginfo.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.
- Chacko, A.; and Kumar, S. M. 2017. Big data provenance research directions. In *TENCON 2017-2017 IEEE Region 10 Conference*, 651–656. IEEE.
- Chappidi, S.; Cobbe, J.; Norval, C.; Mazumder, A.; and Singh, J. 2025. Accountability Capture: How Record-Keeping to Support AI Transparency and Accountability (Re)shapes Algorithmic Oversight. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 554–566.
- Cotality, Inc. 2026. Marshall & Swift Products. <https://www.cotality.com/products/marshall-swift>. Accessed: 2026-02-27.
- Coulthart, S.; and Riccucci, R. 2022. Putting big data to work in government: the case of the United States border patrol. *Public Administration Review*, 82(2): 280–289.
- Desai, R. J.; Sarpatwari, A.; Dejene, S.; Khan, N. F.; Lii, J.; Rogers, J. R.; Dutcher, S. K.; Raofi, S.; Bohn, J.; Connolly, J. G.; et al. 2019. Comparative effectiveness of generic and brand-name medication use: A database study of US health insurance claims. *PLoS medicine*, 16(3): e1002763.
- Donaldson, D. R.; Schleyer, T. K. L.; Riordan, G. G.; and Lian, J. 2024. The Transparency of an Honest Data Broker in Providing Electronic Health Record Data Sufficient for Reuse. *International Journal of Digital Curation*, 18(1).
- Dornfest, A. S.; Rearich, J.; Brydon III, T. D.; and Almy, R. 2019. State and Provincial Property Tax Policies and Administrative Practices (PTAPP): 2017 Findings and Report. *Journal of Property Tax Assessment & Administration*, 16(1).
- Einav, L.; and Levin, J. 2014. Economics in the age of big data. *Science*, 346(6210): 1243089.
- Eurostat (European Commission). 2018. European Statistics Code of Practice: For the National Statistical Authorities and Eurostat (EU Statistical Authority). Technical Report KS-02-18-142-EN-N, Publications Office of the European Union, Luxembourg.
- Frénay, B.; and Verleysen, M. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5): 845–869.
- Geburu, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Gueorguieva, A.; Panidapu, A.; Ho, D.; and King, J. 2026. Privacy Without Remedy: An Assessment of Data Broker Compliance with California Privacy Law. In *Proceedings of the 2026 ACM conference on fairness, accountability, and transparency*.
- Hsu, C.-W.; Liu, C.; Nguyen, K. M.; Chien, Y.-H.; and Mostafavi, A. 2024. Do human mobility network analyses produced from different location-based data sources yield similar results across scales? *Computers, Environment and Urban Systems*, 107: 102052.
- Hu, M. 2015. Big data blacklisting. *Fla. L. Rev.*, 67: 1735.
- IAAO. 2017. Stanard on Mass Appraisal of Real Property.
- IAAO; and Almy, R. R. 1978. *Improving real property assessment: A reference manual*. International Association of Assessing Officers.
- Jennifer, M.; RES, R.; et al. 2021. Garbage in, garbage out: Implications of data quality for valuation models. *Journal of Property Tax Assessment & Administration*, 18(1): 1.
- J’Neka, S. C.; Lutsey, P. L.; MacLehose, R. F.; Chen, L. Y.; Lewis, T. T.; and Alonso, A. 2019. Geographic disparities in the incidence of stroke among patients with atrial fibrillation in the United States. *Journal of Stroke and Cerebrovascular Diseases*, 28(4): 890–899.
- Lane, B. 2018. FTC sanctions CoreLogic for failing to comply with DataQuick acquisition agreement.
- Lane, J. 2016. Big data for public policy: The quadruple helix. *Journal of Policy Analysis and Management*, 35(3): 708–715.
- Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of Google Flu: traps in big data analysis. *science*, 343(6176): 1203–1205.
- Li, N.; Goel, N.; and Ash, E. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’22, 396–410. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

- Mangal, P.; Rajesh, A.; and Misra, R. 2020. Big data in climate change research: Opportunities and challenges. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 321–326. IEEE.
- Martin-Sanchez, F.; and Verspoor, K. 2014. Big data in medicine is driving big changes. *Yearbook of medical informatics*, 23(01): 14–20.
- McMillen, D.; and Singh, R. 2020. Assessment Regressivity and Property Taxation. *The Journal of Real Estate Finance and Economics*, 60(1-2): 155–169.
- Mehra, M. R.; Desai, S. S.; Kuy, S.; Henry, T. D.; and Patel, A. N. 2020. Retracted: Cardiovascular disease, drug therapy, and mortality in Covid-19. *New England Journal of Medicine*, 382(25): e102.
- Meng, X.-L. 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election.
- Noi, E.; Rudolph, A.; and Dodge, S. 2022. Assessing COVID-induced changes in spatiotemporal structure of mobility in the United States in 2020: a multi-source analytical framework. *International Journal of Geographical Information Science*, 36(3): 585–616.
- Nolte, C.; Boyle, K. J.; Chaudhry, A. M.; Clapp, C.; Guignet, D.; Hennighausen, H.; Kushner, I.; Liao, Y.; Mamun, S.; Pollack, A.; et al. 2024. Data practices for studying the impacts of environmental amenities and hazards with nationwide property data. *Land Economics*, 100(1): 200–221.
- Office of Management and Budget, Executive Office of the President. 2008. Statistical Policy Directive No. 4: Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies. Technical Report 46, Federal Register. 73 Fed. Reg. 12622 (Mar. 7, 2008).
- Office of Management and Budget, Executive Office of the President. 2014. Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units. Technical Report 231, Federal Register. 79 Fed. Reg. 71609 (Dec. 2, 2014).
- Office of the Director of National Intelligence. 2024. Intelligence Community Policy Framework for Commercially Available Information. Technical report, Office of the Director of National Intelligence. Unclassified policy framework, May 2024.
- O’Driscoll, A.; Daugelaite, J.; and Sleator, R. D. 2013. ‘Big data’, Hadoop and cloud computing in genomics. *Journal of biomedical informatics*, 46(5): 774–781.
- Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1776–1826.
- Rothschild, A.; Wang, D.; Jayakumar Vilvanathan, N.; Wilcox, L.; DiSalvo, C.; and DiSalvo, B. 2024. The Problems with Proxies: Making Data Work Visible through Requester Practices. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1255–1268.
- Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
- San Mateo County Board of Supervisors. 2024. 24-068 – 20240130.a.ACRE C3 AI Property Appraisal Valuation Final (Signed). <https://sanmateocounty.legistar.com/View.ashx?M=F&ID=12622213&GUID=EFFDB7B1-A24C-434B-9592-26421E46A1BD>. Legistar document, 51 pp.
- Seeskin, Z. H. 2018. Evaluating the utility of a commercial data source for estimating property tax amounts. *Statistical Journal of the IAOS*, 34(4): 543–551. Publisher: SAGE Publications.
- Sherman, J. 2023. Data Brokerage, the Sale of Individuals’ Data, and Risks to Americans’ Privacy, Personal Safety, and National Security. Written Testimony before the U.S. House Committee on Energy and Commerce, Subcommittee on Oversight and Investigations, Hearing on “Who is Selling Your Data: A Critical Examination of the Role of Data Brokers in the Digital Economy”. Duke University Sanford School of Public Policy.
- Smith, E.; Harvey, E.; Berry, C.; Goldin, J.; and Ho, D. E. 2026. Tradeoffs are Domain Dependent: Improving Accuracy and Fairness in Property Tax Assessments. arXiv:2605.15020.
- Suits, D. B. 1977. Measurement of tax progressivity. *The American Economic Review*, 67(4): 747–752.
- Suresh, H.; and Gutttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO, 1–9. – NY USA: ACM. ISBN 978-1-4503-8553-4.
- U.S. Department of Housing and Urban Development. 2024. Notice of Intent – Sole Source Contract: Foreclosure and Residential Sales Data Subscription (86614624Q00004 – Amendment 1). Notice of Intent to award a sole source contract to ATTOM Data Solutions LLC under FAR 13.501(a)(1)(ii).
- van Kempen, E.; Bagayatkar, I.; Frolikov, P.; Georgiou, C.; and Tsudik, G. 2025. Consumer Beware! Exploring Data Brokers’ CCPA Compliance. arXiv:2506.21914.
- Wade, A. 2023. The clocks are striking thirteen: Congress, not courts, must save us from government surveillance via data brokers. *Tex. L. Rev.*, 102: 1099.
- Wang, J.; Crawl, D.; Purawat, S.; Nguyen, M.; and Altintas, I. 2015. Big data provenance: Challenges, state of the art and opportunities. In *2015 IEEE international conference on big data (Big Data)*, 2509–2516. IEEE.
- Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114.
- Zajac, H. D.; Avlona, N. R.; Kensing, F.; Andersen, T. O.; and Shklovski, I. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, 351–362. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Zook, M.; and Spangler, I. 2023. A crisis of data? Transparency practices and infrastructures of value in data broker platforms. *Annals of the American Association of Geographers*, 113(1): 110–128.