ChatGPT

Integralność i wiarygodność Big Data: zasady naukowe, mechanizmy zniekształceń i kontrola jakości przy maskowaniu danych

Kontekst i formalizacja tezy o „pozornie niewinnej" podmianie wartości

Teza: w zbiorze zawierającym miliony rekordów globalna podmiana wartości (np. ciągu znaków „1.1.1.1" → „2.2.2.2") może wywołać istotne zniekształcenia wyników analitycznych; podczas gdy w małym zbiorze analogiczna operacja bywa postrzegana jako „nieszkodliwa", w warunkach Big Data potrafi podważyć poprawność całego wnioskowania.

Formalnie rozważmy zbiór danych jako relację $D = \{r_i\}_{i=1}^{N}$, gdzie rekord r_i zawiera atrybut kategoryczny A (np. identyfikator, adres IP, identyfikator klienta) oraz inne zmienne X i ewentualnie zmienną celu Y (np. etykietę zdarzenia, wartość finansową). Transformację „maskującą przez podmianę" można modelować jako funkcję

$$T(a) = \begin{cases} b & \text{gdy } a = a_0 \\ a & \text{w p.p.} \end{cases}$$

czyli deterministyczne przepisanie jednej wartości a_0 na inną wartość b, zastosowane do całego zbioru. Krytycznym przypadkiem, zawartym w tezie, jest sytuacja, gdy b **już występuje** w danych: wówczas transformacja nie jest różnowartościowa (nie jest iniekcją), lecz **scala** (kolizyjnie) dwie klasy rekordów $A = a_0$ i $A = b$. Tego typu kolizja jest fundamentalnie odmienna od bezkolizyjnego tokenizowania (1–1) lub permutacji po całej dziedzinie. W literaturze o prywatności i publikowaniu danych jest to klasyczny przykład konfliktu „privacy–utility tradeoff": techniki anonimizujące/maskujące redukują ryzyko ujawnienia, ale mogą degradować użyteczność analityczną, czasem w sposób trudny do przewidzenia bez formalnej analizy. [1]

W kontekście Big Data (duża skala, złożone potoki przetwarzania, liczne agregacje i segmentacje) problem jakości danych i jej oceny jest uznawany za warunek konieczny uzyskania wiarygodnej wartości informacyjnej. [2]

Fundamentalne zasady statystyczne istotne dla Big Data

W dyskusji „dlaczego błąd nie znika w dużych danych" kluczowe są trzy zasady, które można traktować jako „prawa" (w sensie stabilnych konsekwencji teorii statystyki i teorii informacji), szczególnie istotne dla maskowania i propagacji błędów.

Pierwsza zasada: duże N redukuje wariancję, ale nie usuwa obciążenia (bias). W klasycznym rozkładzie błędu średniokwadratowego estymatora (MSE) składnik wariancyjny zazwyczaj maleje wraz z N, natomiast obciążenie systematyczne nie ma powodu zanikać tylko dlatego, że rośnie liczba obserwacji. W rezultacie w Big Data „mikrobłąd" systematyczny może dominować nad losową niepewnością i

prowadzić do **precyzyjnie błędnych** wniosków (wysoka pewność, błędny wynik). Ten mechanizm jest spójny z analizami metodologicznymi wskazującymi, że w erze Big Data rośnie ryzyko nadinterpretacji istotności statystycznej przy ogromnych liczebnościach prób. [3]

Druga zasada: „Law of Large Populations" i „Big Data Paradox" (w ujęciu entity["people","Xiao-Li Meng","statystyk, harvard"]) pokazują, że dla wnioskowania o populacji sama liczebność danych nie jest miarą „informatywności", a błąd może rosnąć wraz z rozmiarem populacji, jeśli jakość (np. selekcyjność rejestracji, defekty pomiaru) nie jest kontrolowana. W pracy tej błąd estymacji średniej populacyjnej jest rozkładany na czynniki jakości danych, ilości danych i trudności problemu; wniosek: przy nawet drobnej korelacji defektu rejestracji z badaną cechą, pozornie „wielkie" dane mogą mieć dramatycznie małą efektywną liczebność. [4]

Trzecia zasada: propagacja niepewności/błędu przez przekształcenia jest własnością modelu przetwarzania, nie tylko danych wejściowych. W metrologii i teorii pomiaru formalizuje się to jako „prawo propagacji niepewności": jeśli wyjście jest funkcją wejść, to wariancja/niepewność wyjścia zależy od (zlinearyzowanej) wrażliwości funkcji na wejścia (pochodne/Jakobian) lub – przy nieliniowości – od propagacji rozkładów (np. metodą Monte Carlo). Analogicznie, w potokach Big Data kolejne transformacje (czyszczenie, łączenia, agregacje, inżynieria cech) działają jak złożenie funkcji, które może wzmacniać błąd. [5]

W praktyce Big Data dochodzi jeszcze czwarty, „systemowy" komponent: złożoność potoków i automatyzacja. Narzędzia i prace systemowe na temat jakości danych podkreślają konieczność explicite testowania założeń i kontraktów danych (constraints, schematy, reguły), bo błędy wejściowe i transformacyjne są w stanie przejść niezauważone, a ich koszt rośnie w skali (liczbą downstream zastosowań i zależności). [6]

Mechanika zniekształceń: jak globalne podmiany propagują się przez agregacje i modele

Zniekształcenie rozkładu częstości i efekt „kolizji kategorii"

Niech n_a oznacza liczbę rekordów z wartością $A = a$, a $\hat{p}(a) = n_a/N$ – empiryczną częstość. Po transformacji $a_0 \to b$ otrzymujemy:

$$n'_{a_0} = 0, \qquad n'_b = n_b + n_{a_0}, \qquad n'_a = n_a \text{ dla } a \notin \{a_0, b\}.$$

To jest **sklejanie dwóch mas prawdopodobieństwa** w jedną kategorię. W prostym sensie odległości rozkładów w normie L_1:

$$\|\hat{p}' - \hat{p}\|_1 = \sum_a |\hat{p}'(a) - \hat{p}(a)| = 2\hat{p}(a_0),$$

czyli zmiana rozkładu jest proporcjonalna do częstości maskowanej wartości. W małym zbiorze $\hat{p}(a_0)$ często wynosi 0 (wartość nie występuje) lub $1/N$ (pojedynczy rekord), natomiast w Big Data $\hat{p}(a_0)$ stabilizuje się wokół częstości rzeczywistej, a liczba zmienionych rekordów wynosi w przybliżeniu $N \cdot p(a_0)$ (por. Wykres o skali Np). Sam fakt stabilizacji jest paradoksalnie niekorzystny: błąd staje się **powtarzalny i systematyczny**, zamiast „rozmywać się" losowo między eksperymentami. Z punktu widzenia wnioskowania statystycznego jest to przejście od błędu losowego do błędu systematycznego, z konsekwencją w postaci trwałego obciążenia. [7]

Dlaczego duże zbiory bywają bardziej wrażliwe niż małe

Na poziomie „samej matematyki rozkładu" relatywny błąd $\hat{p}(a_0)$ nie musi rosnąć z N. Teza o szczególnej szkodliwości w Big Data staje się jednak rygorystycznie prawdziwa, gdy uwzględni się typowe cechy analityki Big Data:

Po pierwsze, **w Big Data analizuje się na poziomie wysokiej granularności** (np. per użytkownik, per identyfikator, per adres), ponieważ liczebność pozwala. To powoduje, że atrybuty identyfikujące stają się osiami agregacji i cechami modelu. Wówczas sklejanie kategorii niszczy strukturę segmentacji: zmienia rozkład w każdym raporcie typu group-by, każdej metryce „top-k", w każdym wykresie trendu „per źródło". Ten efekt jest zgodny z ujęciami jakości danych w Big Data, gdzie „veracity" i semantyka atrybutów są traktowane jako krytyczne, a reguły integralności nie zawsze dają się zadeklarować a priori ze względu na różnorodność i dynamikę domeny. [8]

Po drugie, w Big Data typowe są operacje o dużej „wrażliwości funkcji" (w sensie teorii prywatności różnicowej): liczniki unikatów, rankingi, maksimum/minimum, progi alertowe, detektory anomalii, metryki grafowe. W prywatności różnicowej wrażliwość (sensitivity) formalizuje maksymalną zmianę wyniku zapytania przy zmianie pojedynczego rekordu; istotne jest, że zmiana K rekordów może w przybliżeniu skalować wpływ na wynik jak K razy wrażliwość jednostkowa. Podmiana wartości stosowana „hurtowo" zmienia nie jeden rekord, lecz wszystkie o $A = a_0$ – czyli $K = n_{a_0}$, które w Big Data może być duże. [9]

Po trzecie, Big Data zwiększa **moc testów i „łatwość wykrycia" minimalnych różnic**, co w praktyce prowadzi do sytuacji, w której nawet mikroskopijne przesunięcia (np. wywołane maskowaniem) stają się „statystycznie istotne", a tym samym nadają się do wygenerowania pozornie solidnych, lecz błędnych narracji wnioskowania. W literaturze polskiej jest to dyskutowane m.in. w kontekście roli p-value i jakości danych w erze big data. [10]

Propagacja przez łączenia (JOIN), deduplikację i inżynierię cech

Najbardziej destrukcyjny wariant tezy dotyczy sytuacji, gdy A pełni rolę *klucza* (jawnie lub niejawnie).

Jeśli w potoku istnieje etap łączenia dwóch tabel po A (np. logi sesji ↔ zdarzenia bezpieczeństwa; transakcje ↔ profile), zamiana $a_0 \to b$ może wywołać dwa typy szkód równocześnie: (i) utratę prawidłowych dopasowań dla a_0 (false negatives), (ii) pojawienie się fałszywych dopasowań do b (false positives). W najgorszym przypadku liczba błędnych par po JOIN rośnie jak iloczyn liczności klas po obu stronach:

$$\text{błędne pary} \approx n_{a_0}^{(L)} \cdot n_b^{(R)},$$

gdzie $n^{(L)}$ i $n^{(R)}$ to częstości po lewej i prawej stronie łączenia. To jest mechanizm „wzmacniacza": pojedyncza decyzja maskowania może spowodować multiplikację rekordów w wyniku join i radykalnie przestawić agregaty downstream (sumy, średnie, liczniki). W praktyce systemów Big Data powiązanych z bezpieczeństwem sieciowym i analizą przepływów jest to jeden z powodów, dla których anonimizacja identyfikatorów (np. IP) musi być projektowana pod konkretne zadania analityczne. [11]

W inżynierii cech bardzo częsty jest mechanizm „kodowania informacji per kategoria" (np. target encoding, liczniki częstości, wskaźniki ryzyka per identyfikator). Jeżeli $m(a) = \mathbb{E}[Y \mid A = a]$ jest efektem, który chcemy oszacować z danych, to po sklejeniu a_0 z b estymator dla b staje się średnią ważoną:

$$\widehat{m}'(b) = \frac{n_b \widehat{m}(b) + n_{a_0} \widehat{m}(a_0)}{n_b + n_{a_0}}.$$

To oznacza, że podmiana pojedynczego identyfikatora może zmienić **nie tylko rekordy pierwotnie z** a_0, lecz także wszystkie rekordy z b (bo ich cecha grupowa została przeliczona). Ten efekt jest szczególnie groźny w Big Data, gdzie (i) liczności n_b bywają ogromne, a więc rozmiar populacji dotkniętej zmianą jest ogromny, oraz (ii) modele są trenowane i walidowane na wielkich próbach o małych błędach standardowych, więc nawet małe przesunięcie cechy może realnie zmienić parametr modelu i predykcje. Te zależności są klasycznym tematem teorii błędu pomiaru i błędnej klasyfikacji (misclassification) w modelach statystycznych. [12]

Związek z „kolizjami" w mapowaniach (hashing) – analogia teoretyczna

Zamiana $a_0 \to b$ jest deterministyczną kolizją dwóch symboli. W uczeniu maszynowym analogicznym mechanizmem jest „feature hashing" (hashing trick), w którym wiele symboli mapuje się do ograniczonej liczby kubełków, dopuszczając kolizje; praca o własnościach geometrycznych takiego haszowania dostarcza granic błędu i pokazuje, że kolizje są kluczowym źródłem degradacji, zależnym od liczności i częstości kolidujących cech. Jest to formalne potwierdzenie, że „niewinne mapowanie" kategorii do wspólnej reprezentacji może być problemem strukturalnym, a nie jedynie „szumem". [13]

Studia przypadków z literatury: gdy anonimizacja/maskowanie zniekształca analitykę

Pierwsza klasa studiów przypadków pochodzi z analizy danych sieciowych i logów, gdzie adresy IP i identyfikatory hostów są centralne dla agregacji, detekcji anomalii i atrybucji zdarzeń.

W pracy o kompromisie „risk–utility" dla ucinania (truncation) adresów IP wykazano empirycznie, że anonimizacja przez truncation może szybko degradować użyteczność danych dla detekcji anomalii, a tempo degradacji zależy od metryki (np. liczniki unikatów vs metryki entropijne). Raportowano m.in., że użyteczność liczników adresów wewnętrznych może zostać praktycznie utracona przy nawet niewielkiej truncacji (rzędu kilku bitów), podczas gdy pewne metryki oparte o entropię są odporniejsze. [14]

W badaniach i raportach o anonimizacji śladów sieciowych podkreśla się, że „brutalne" usuwanie lub prymitywne maskowanie pól może unieważniać całe klasy analiz (np. badania opcji TCP, analizę zależności czasowych), a jednocześnie nie gwarantuje bezpieczeństwa (możliwe są ataki reidentyfikacyjne wykorzystujące cechy uboczne). To jest klasyczny dowód na to, że maskowanie „dla bezpieczeństwa" bez modelu użyteczności i bez przeglądu ryzyka może prowadzić do równoczesnej utraty prywatności i utraty jakości analitycznej. [15]

Druga klasa studiów przypadków dotyczy Big Data w sensie predykcji i inferencji na danych pośrednich (proxy). W głośnym przypadku prognozowania zachorowań na grypę na podstawie zapytań internetowych autorzy pokazali pułapki „big data hubris": duże, pasywne dane mogą sprawiać wrażenie, że zastępują klasyczne pomiary, ale zmiany w mechanizmie generowania danych (np. algorytmy wyszukiwarki, zachowanie użytkowników) prowadzą do błędów predykcji, które nie są kompensowane samą skalą danych. Jest to przykład, w którym problemem jest dynamika i drift, a nie „niedobór danych", i w którym jakość/wiarygodność sygnału jest kluczowa dla całego wnioskowania. [16]

Trzecia klasa studiów przypadków, bliższa dokładnie rozważanej tezie, to literatura o „maskowaniu z zachowaniem wnioskowania" (inferential integrity). W pracy o model-targeted masking i wielokrotnej imputacji proponuje się ramę, w której maskowanie jest projektowane tak, aby minimalizować zmianę

wniosków inferencyjnych (np. estymacji i przedziałów ufności) przy równoczesnej kontroli ryzyka ujawnienia. Jest to formalnie istotne, bo pokazuje, że ad hoc podmiana wartości (często spotykana w praktyce) nie jest neutralna; aby zachować integralność wnioskowania, maskowanie musi być sprzężone z modelem oraz procedurą oceny zniekształceń. [17]

Metody wykrywania, ograniczania i korekcji zniekształceń w Big Data

Metody naprawy problemu „podmiana niszczy analizę" należy rozdzielić na (i) projekt maskowania, (ii) detekcję zniekształceń, (iii) redukcję wpływu błędów na wnioskowanie.

Projekt maskowania: minimalizacja kolizji i zachowanie własności analitycznych

Najbardziej podstawowa zasada inżynieryjno-statystyczna brzmi: jeśli atrybut jest kluczem (łączenia, deduplikacji, agregacji), maskowanie powinno zachować relację równoważności „ten sam identyfikator" bez sklejania z innym identyfikatorem. Oznacza to preferencję dla tokenizacji/permutacji 1–1 (w idealnym przypadku bijekcji na zbiorze identyfikatorów) zamiast mapowania do istniejącej wartości. W obszarze danych sieciowych rozwinięto techniki anonimizacji IP zachowujące strukturę prefiksów (prefix-preserving), aby utrzymać użyteczność dla analiz zależnych od topologii i hierarchii adresacji. [18]

W obszarze danych osobowych i usług chmurowych tokenizacja i format-preserving encryption są wskazywane jako techniki pozwalające zachować format danych i kontrolować odwracalność/ nieodwracalność, co bywa kluczowe dla zachowania spójności systemów i testów. [19]

Z perspektywy naukowej należy podkreślić różnicę między „maskowaniem dla ukrycia wartości" a „maskowaniem dla zachowania relacji". Teza o destrukcyjności „1.1.1.1 → 2.2.2.2" dotyczy dokładnie tego, że relacje (segmentacja, zliczenia, łączenia) są niszczone przez kolizję.

Detekcja: testy kontraktów danych, profilowanie i wykrywanie driftu

W systemach Big Data coraz częściej przyjmuje się paradygmat „unit tests for data": użytkownik deklaruje założenia o danych (np. kompletność, unikalność, dopuszczalne zakresy, słowniki wartości kategorycznych), a system weryfikuje je skalowalnie na danych produkcyjnych. Platforma i biblioteka entity["company","Amazon","e-commerce company"] do walidacji jakości danych w potokach uczenia maszynowego jest opisywana jako mechanizm, który umożliwia jawne kodowanie założeń oraz ich automatyczne sprawdzanie na dużych zbiorach, ponieważ naruszenia założeń mogą prowadzić do awarii lub błędnych predykcji. [20]

Analogicznie, narzędzia walidacji danych w ekosystemach uczenia maszynowego opisują detekcję anomalii i driftu przez porównywanie statystyk danych ze schematem/kontraktem oraz przez porównywanie rozkładów między treningiem a serwowaniem. [21]

Ponieważ w Big Data testy statystyczne mają ogromną moc, detekcja zmian dystrybucji powinna łączyć miary odległości rozkładów i kryteria wielkości efektu (effect size), a nie opierać się wyłącznie na p-value; ten postulat jest zgodny z analizami metodologicznymi dotyczącymi istotności statystycznej w Big Data. [10]

Korekcja: reguły integralności, odkrywanie zależności i czyszczenie na skalę Big Data

Klasyczna szkoła jakości danych wskazuje na rolę wymiarów jakości (m.in. dokładność, kompletność, spójność, terminowość) oraz traktuje „fit for use" jako definicję nadrzędną. [22]

W praktyce formalne reguły integralności często przyjmują formę zależności danych (functional dependencies, conditional functional dependencies). W Big Data rozwijane są algorytmy odkrywania takich zależności w środowiskach rozproszonych (np. na bazie silników równoległych), aby wykrywać anomalie i wspierać czyszczenie danych. [23]

Istnieją systemy czyszczenia Big Data projektowane pod skalę i koszt obliczeniowy, które translują reguły jakości do serii transformacji wykonywalnych w rozproszonych frameworkach, adresując typowe bariery (enumeracja par, złączenia nierównościowe, funkcje użytkownika). [24]

Z kolei systemy zarządzania jakością danych w środowiskach rozproszonych dostarczają interfejsów zarówno do detekcji, jak i naprawy, integrując różne klasy metod (od detekcji wartości odstających po naprawę opartą o reguły). [25]

Redukcja wpływu na wnioskowanie: odporność statystyczna i prywatność „z gwarancją"

Jeżeli celem jest zachowanie wniosków statystycznych, a nie tylko „wyczyszczenie" danych, potrzebne są metody ściśle statystyczne: odporne estymatory oraz jawny model błędu (zanieczyszczenia). Model zanieczyszczenia (contamination) i estymacja odporna formalizują sytuację, w której obserwacje pochodzą z mieszanki „prawdziwego" rozkładu i zakłóceń; nawet mały ułamek zanieczyszczenia może wprowadzać znaczące obciążenie w estymatorach nieodpornych. [26]

W obszarze prywatności różnicowej gwarancje są formułowane przez ograniczenie wpływu pojedynczej obserwacji na wynik zapytania/statystyki przez dodanie losowego szumu skalibrowanego do wrażliwości funkcji. Jest to koncepcyjnie przeciwieństwo deterministycznej, kolizyjnej podmiany: zamiast strukturalnie sklejać kategorie, prywatność osiąga się przez kontrolowaną stochastyczną perturbację z parametryczną gwarancją. [27]

Synteza porównawcza: mały zbiór danych a Big Data w obliczu maskowania przez podmianę

Poniższa tabela syntetyzuje różnice „mały zbiór vs Big Data" dla maskowania o typie „podmień wartość a_0 na istniejącą wartość b". W wielu punktach różnice wynikają z tego, że Big Data zachęca do granularnych analiz, ma ogromną moc testów oraz używa identyfikatorów jako osi łączeń i agregacji; to powoduje, że kolizje i obciążenia są bardziej destrukcyjne systemowo niż w małych próbach. [28]

Wymiar / metryka	Mały zbiór (np. $N \approx 10^3$)	Big Data (np. $N \approx 10^6$–10^9)
Odsetek zmienionych rekordów	$p = \hat{p}(a_0)$ (często 0 lub $1/N$)	p stabilne i niezerowe; błąd staje się systematyczny
Skala absolutna zmiany	$N \cdot p$ zwykle 0–kilka rekordów	$N \cdot p$ może oznaczać setki–miliony rekordów (patrz wykres skali Np)

Wymiar / metryka	Mały zbiór (np. $N \approx 10^3$)	Big Data (np. $N \approx 10^6$–10^9)
Odległość rozkładów kategorii	$\|\hat{p}' - \hat{p}\|_1 = 2\hat{p}(a_0)$, często bliska 0	$\|\hat{p}' - \hat{p}\|_1 \approx 2p(a_0)$ – trwałe zniekształcenie rozkładu
Wpływ na agregacje per kategoria	Często niewidoczny z powodu małych liczności klas	Zmienia rankingi, trendy i metryki w wielu przekrojach (duża granularność analiz)
Ryzyko zniszczenia łączeń (JOIN) po kluczu A	Mniejsze rozmiary, ale błędy mogą być jakościowo krytyczne (błędna atrybucja)	Potencjalna multiplikacja błędnych par $\sim n_{a_0}^{(L)} \cdot n_b^{(R)}$; stabilne „wstrzyknięcie" błędu do wielu downstream tabel
Wpływ na modelowanie (np. cechy per identyfikator)	Parametry niestabilne (duża wariancja), więc błąd może „zginąć" w szumie estymacji	Wariancja maleje, bias dominuje; model może być bardzo pewny, ale błędny („precyzyjnie błędne" wnioskowanie)
Interpretacja istotności statystycznej	Niska moc testów; małe efekty często niewykrywalne	Ogromna moc; minimalne przesunięcia stają się „istotne", co wymaga rygoru w ocenie efektu i jakości danych

W świetle tezy szczególnie niebezpieczny jest scenariusz, w którym atrybut podlegający maskowaniu jest (jawnie lub ukrycie) używany jako: (i) klucz łączenia, (ii) identyfikator do deduplikacji, (iii) oś agregacji, (iv) kategoria w modelu lub źródło cech zbiorczych. W takich warunkach podmiana $a_0 \to b$ nie jest „kosmetyczna"; jest to operacja o przewidywalnym, strukturalnym skutku: zlewanie klas i propagacja obciążenia przez cały potok przetwarzania. [29]

Bibliografia wybrana (źródła pierwotne i przeglądowe)

Kluczowe prace (wskazane w tekście cytowaniami) obejmują: teorię i praktykę jakości danych (w tym klasyczne ramy wymiarów jakości), teorię błędu i odporności estymacji, formalne modele prywatności i wrażliwości, oraz studia przypadków z anonimizacji danych sieciowych i pułapek Big Data w predykcji. Szczególnie istotne w kontekście tezy są: analiza „Big Data Paradox" entity["people","Xiao-Li Meng","statystyk, harvard"], studium „Google Flu" autorstwa m.in. entity["people","David Lazer","political scientist"] i entity["people","Gary King","political scientist"] opublikowane w entity["organization","Science","journal"], polska analiza znaczenia istotności statystycznej w Big Data (entity["people","Mirosław Szreder","statystyk, poland"]; entity["organization","Główny Urząd Statystyczny","warsaw, poland"]), oraz literatura o kompromisie użyteczność–prywatność w anonimizacji danych sieciowych i publikowaniu danych. [30]

---

[1] Privacy-preserving data publishing
https://dmas.lab.mcgill.ca/fung/pub/FWCY10csur.pdf?utm_source=chatgpt.com

[2] The Challenges of Data Quality and Data Quality Assessment ...
https://datascience.codata.org/articles/dsj-2015-002?utm_source=chatgpt.com

[3] [10] Istotność statystyczna w czasach big data
https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5982/7/62/1/ws_11_2019__05_istotnosc_statystyczna_w_czasach_big_data__miroslaw_szreder.pdf?utm_source=chatgpt.com

4 28 30 Statistical paradises and paradoxes in big data (I)
https://projecteuclid.org/journals/annals-of-applied-statistics/volume-12/issue-2/Statistical-paradises-and-paradoxes-in-big-data-I--Law/10.1214/18-AOAS1161SF.full?utm_source=chatgpt.com

5 JCGM 100:2008 (GUM 1995 with minor corrections
https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf?utm_source=chatgpt.com

6 Automating Large-Scale Data Quality Verification
https://www.vldb.org/pvldb/vol11/p1781-schelter.pdf?utm_source=chatgpt.com

7 12 Measurement Error in Nonlinear Models
https://ndl.ethernet.edu.et/bitstream/123456789/88802/1/CarrollRupertStefanskiCrainiceanu2006%20measurement%20error%20in%20nonlinear%20models.pdf?utm_source=chatgpt.com

8 Data Quality: the Other Face of Big Data
https://pdfs.semanticscholar.org/20e8/63fd7134ddc3467fca9ae212d3983b2c0426.pdf?utm_source=chatgpt.com

9 27 Calibrating Noise to Sensitivity in Private Data Analysis
https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf?utm_source=chatgpt.com

11 29 The Risk-Utility Tradeoff for IP Address Truncation
https://arxiv.org/abs/0903.4266?utm_source=chatgpt.com

13 Feature Hashing for Large Scale Multitask Learning
https://alex.smola.org/papers/2009/Weinbergeretal09.pdf?utm_source=chatgpt.com

14 The Risk-Utility Tradeoff for IP Address Truncation
https://arxiv.org/pdf/0903.4266?utm_source=chatgpt.com

15 devil.dvi
https://www.icir.org/enterprise-tracing/devil-ccr-jan06.pdf

16 The Parable of Google Flu: Traps in Big Data Analysis
https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf?utm_source=chatgpt.com

17 Balancing Inferential Integrity and Disclosure Risk via ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11466287/?utm_source=chatgpt.com

18 Prefix-Preserving IP Address Anonymization
https://www.csl.mtu.edu/cs6461/www/Reading/Xu02.pdf?utm_source=chatgpt.com

19 ITSP.50.108 Guidance on Using Tokenization for Cloud- ...
https://www.cyber.gc.ca/sites/default/files/cyber/2022-03/ITSP-50-108-Guidance-on-Using-Tokenization-for-Cloud-Based-ServicesV2-e.pdf?utm_source=chatgpt.com

20 Deequ - Data Quality Validation for Machine Learning Pipelines
https://learningsys.org/nips18/assets/papers/5CameraReadySubmissiondeequ.pdf?utm_source=chatgpt.com

21 TensorFlow Data Validation: Checking and analyzing your ...
https://www.tensorflow.org/tfx/guide/tfdv?utm_source=chatgpt.com

22 Data Quality Dimensions
https://web.mit.edu/tdqm/www/tdqmpub/WandWangCACMNov96.pdf?utm_source=chatgpt.com

23 DFD: Efficient Functional Dependency Discovery
https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2014_abedjan_dfd.pdf?utm_source=chatgpt.com

24 BigDansing: A System for Big Data Cleansing
https://cs.uwaterloo.ca/~ilyas/papers/ZuhairSIGMOD2015.pdf?utm_source=chatgpt.com

25 SparkDQ: Efficient generic big data quality management on ...
https://www.sciencedirect.com/science/article/abs/pii/S0743731521001246?utm_source=chatgpt.com

26 Robust Estimation of a Location Parameter
https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-35/issue-1/Robust-Estimation-of-a-Location-Parameter/10.1214/aoms/1177703732.full?utm_source=chatgpt.com