# Practical Solutions to Protect Data Privacy

SQL Saturday San Diego

September 14, 2024

# We'll cover:

- Defining Data Privacy
- Why it's Important
- Disclosure Risk
- Privacy-enhancing techniques
  - Masking and Suppression
  - Differential Privacy
  - Synthetic Data: The New Frontier
- Comparisons

# Hi!


I'm Britton Gray

- Data professional for almost 20 years
  - My passions: advanced analytics, data engineering, data privacy
- Director of BI at Project Lead The Way

I'm also an actual scrum master!

Certifications:

present

past

# Defining Data Privacy

# Data Privacy (*n*):

Any information about an **identified** or <u>identifiable</u> natural person

## Data privacy rights:

**Who** has access to their information

**What** information can be accessed

**When:** Data kept only as long as it's legitimately needed

**Where:** Limitations on sharing and cross-border transfers

**Why:** the purpose(s) for storing/accessing that information

# Legitimate Purposes

Valid reasons for storing and processing data:

- ✅ Affirmative consent of the data subject
- ❤️ Vital interests of a person
- 📜 Fulfillment of a contract
- ⚖️ Legal obligation
- 🌍 In the public interest
- 🏢 ⇔ 🧍 Legitimate interests of processor – balanced by rights of the data subject

# Why is this important?

- Ethics!
- Legal sanctions
  - Fines
    - $2500-$7500 per person affected under CCPA/CPRA
    - GDPR: Max €20 million / 4% of turnover (whichever is greater)
  - Consent decrees
  - Private lawsuits
  - The "ban hammer“
  - Some could potentially be criminal

# Disclosure Risk

Two types

# Anonymized Medical Record

All identifiers have been removed

| | |
|---|---|
| **Hospital** | 162: Sacred Heart Medical Center in Providence |
| **Admit Type** | 1: Emergency |
| **Type of Stay** | 1: Inpatient |
| **Length of Stay** | 6 days |
| **Discharge Date** | Oct-2011 |
| **Discharge Status** | 6: Dsch/Trfn to home under the care of a health service organization |
| **Charges** | $71,708.47 |
| **Payers** | 1: Medicare |
| | 6: Commercial insurance |
| | 625: Other government-sponsored payers |
| **Emergency Codes** | E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl |

| | |
|---|---|
| **Diagnosis Codes** | 80843: closed fracture of other specified part of pelvis |
| | 51851: pulmonary insufficiency following trauma & surgery |
| | 86500: injury to spleen without mention of open wound into cavity |
| | 80705: closed fracture of rib(s); fracture five ribs-close |
| | 5849: acute renal failure; unspecified |
| | 8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury |
| | 2761: hyposmolality &/or hyponatremia |
| | 78057: tachycardia |
| | 2851: acute posthemorrhagic anemia |

| | |
|---|---|
| **Age in Years** | 60 |
| **Age in Months** | 725 |
| **Gender** | Male |
| **ZIP** | 98851 |
| **State Reside** | WA |
| **Race/Ethnicity** | White, Non-Hispanic |
| **Procedure Codes** | 5781: Suture bladder laceration |
| | 7939: 7919: Open/Closed reduction of fracture of other specified bone |
| **Physicians** | … |
| **…** | … |

# Newspaper Article

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash.

[News Review 10/18/2011]

*(name changed)*

# Reidentification

| | |
|---|---|
| Record | |
| Hospital | 162: Sacred Heart Medical Center in Providence |
| Admit Type | 1: Emergency |
| Type of Stay | |
| Length of Stay | 6 days |
| Discharge Date | Oct-2011 |
| Discharge Status | under the care of an health service organization |
| Charges | $71708.47 |
| Payers | 1: Medicare<br>6: Commercial insurance<br>625: Other government |
| Emergency Codes | E8162: motor vehicle traffic accident due t loss of control; loss control mv-nocycl |
| Diagnosis Codes | of other specified part of pelvis<br>51851: pulmonary insufficiency following trauma & surgery<br>276: hyposmolality &/or hyponatremia<br>78057: tachycardia<br>2851: acute orrhagic anemia |
| Age in Years | 60 |
| Gender | Male |
| ZIP | 98851 |
| State Reside | WA |
| | Non-Hispanic |

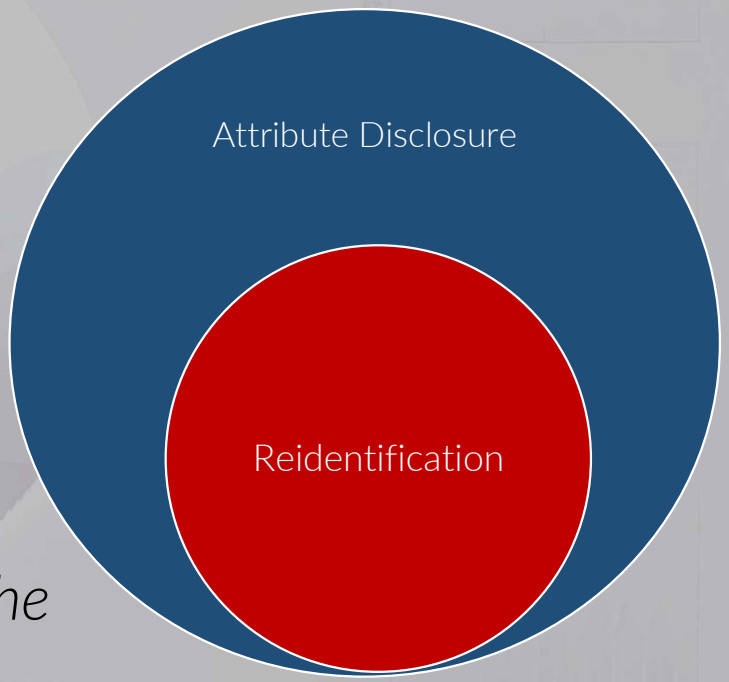
MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

*Credit: "Only You, Your Doctor, and Many Others May Know" Journal of Technology Science*
Dr. Latanya Sweeney, 2015

# Attribute Disclosure

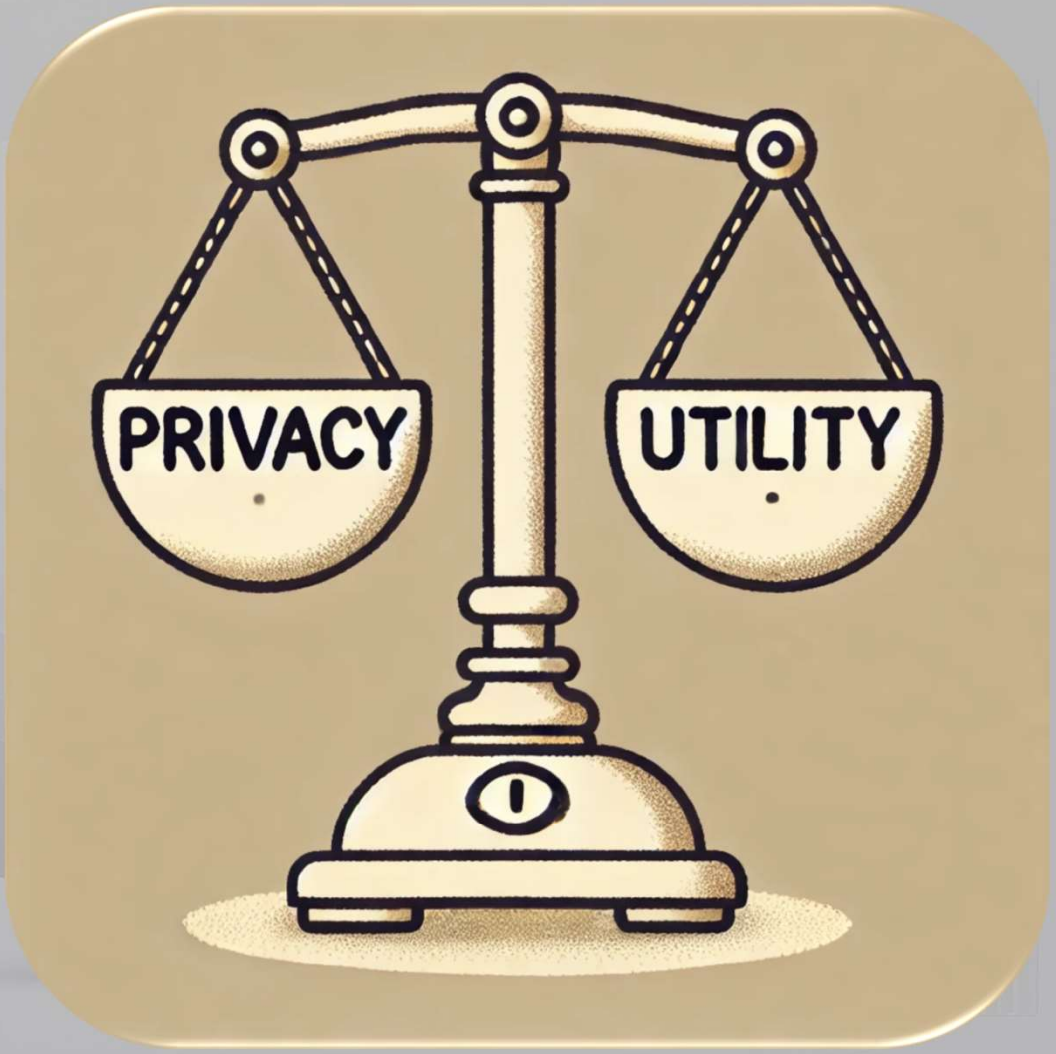
- One or more sensitive attributes or behaviors about a person become discoverable




*"100% of our 11th graders met standards on the state exam!"*

*- Actual school who sued a state because they wouldn't report it*

# Techniques



# Masking and Suppression

The simplest techniques

# Masking

- Useful for individual records and aggregations
- Thwarts re-identification to a degree
- Column / user / row-level security
- Two types of masking targets:
  - Would-be or could-be identifiers
  - Sensitive attributes

# Masking Techniques

| Mask | Example | Use for |
|---|---|---|
| Redaction | Britton → NULL | No need to do any analysis on the field, NULL may have meaning |
| Constant Masking | Britton → **** | Same as above, but mask is obvious |
| Partial Masking | bgray@pltw.org → ***@pltw.org<br>8/6/2024 → 1/1/2024<br>12.34.56.78 → 12.34.56.xx | Part of a field may have meaning.<br>Decreases reidentification risk, but doesn't eliminate it |
| Substitution | Britton → Joseph<br>Pamela → Wendy | Making believable and sliceable data; makes mask less obvious.<br>Maintain mapping tables for consistency, but secure them |
| Hashing | Britton → 1e00fea58e20cc1 | Sliceable data that's obviously masked.<br>Salt the hash: ("AA1438-Britton") |
| Banding / Blurring | 33 → 25-34 (discrete)<br>33 → 30 (aggregable) | Numeric variables important to analysis but pose reidentification risk |
| Geo Blurring | ~~39.9155°N, 86.0650°W~~<br>39.92°N, 86.07°W | Treat precise geo data as identifying.<br>1/100 of a degree = about ½ mile |

# Masking: Data Platforms

Masking policies defined as CASE statements
- Apply policy to 1+ columns
- Can reference lookup tables

---

Masking functions written in SQL
- Applied as a part of column definition

---

Configurable pre-defined masking functions, applied to columns
- Implemented in Fabric's SQL analytics endpoint and Warehouse

---

Not native; ideally use OLS / RLS
Idea: Source view, UNION masked and unmasked, use RLS DAX

---

Same as Power BI
- Understand who can see what

# Complex Masking Behavior

## Based on User-to-State Table

### As a Policy

```
CREATE OR REPLACE MASKING POLICY NP_STAR_MASK AS
(VAL STRING, PERSON_STATE VARCHAR) RETURNS STRING ->
    CASE
    WHEN CURRENT_ROLE() in ('ACCOUNTADMIN', 'SECURITY_ADMIN') THEN VAL
    WHEN CURRENT_ROLE() IN ('DATA_ANALYST')
        AND IS_STATE_AUTHORIZED(PERSON_STATE) THEN VAL
    ELSE LEFT(COALESCE(VAL, '*'), 1) || '***' -- <--- Masking by default
    END;
```

| ANALYST_USERNAME | UNMASKED_STATE |
|---|---|
| GRANDPOOBAH | IN |
| GRANDPOOBAH | CA |
| SOMEONEELSE | TX |

### As a View

```
CREATE OR REPLACE VIEW V_MASKED_PERSON AS
SELECT DIM_PERSON_KEY,
       FIRST_NAME,
       CASE WHEN CURRENT_ROLE() in ('ACCOUNTADMIN')
            THEN LAST_NAME
       WHEN EXISTS (SELECT 1 FROM ANALYST_TO_STATE
                    WHERE ANALYST_USERNAME = CURRENT_USER()
                    AND DIM_PERSON.STATE = ANALYST_TO_STATE.UNMASKED_STATE)
            THEN LAST_NAME
       ELSE LEFT(LAST_NAME, 1) || '***' END AS LAST_NAME,
       STATE
FROM DIM_PERSON;
```

# Complex Masking Behavior

## Results

```sql
SELECT FIRST_NAME || ' ' || LAST_NAME, STATE
FROM DIM_PERSON
WHERE STATE IN ('IN','IL')
ORDER BY 1
LIMIT 5;
```

Results | Chart

| FIRST_NAME \|\|''\|\| LAST_NAME | STATE |
|---|---|
| 9th P*** | IL |
| A'ja W*** | IL |
| Aaron B*** | IL |
| Aaron Brink | IN |
| Aaron D*** | IL |

```sql
SELECT FIRST_NAME, LAST_NAME, STATE
FROM DIM_PERSON
WHERE LAST_NAME = 'Gray';
```

Results | Chart

| FIRST_NAME | LAST_NAME | STATE |
|---|---|---|
| Cyrus | Gray | CA |
| Darin | Gray | CA |
| Wesley | Gray | IN |
| Leonard | Gray | IN |
| Sylvester | Gray | CA |

```sql
SELECT COUNT(DISTINCT LAST_NAME)
FROM V_MASKED_PERSON
WHERE STATE = 'IL';
```

sults | Chart

| COUNT(DISTINCT LAST_NAME) |
|---|
| 31 |

Exact same query behavior: policy or view
(Policies have different metadata implications)

# Suppression

- Works on aggregated data
- Guards against attribute disclosure

The problem:

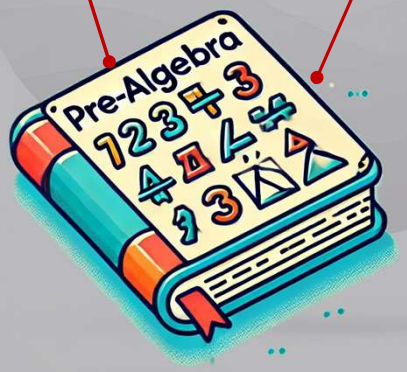
Everyone in CA makes more than $50K!

| | < $50K | $50-100K | $100K+ | Total |
|---|---|---|---|---|
| California | 0 | 12 | 14 | 26 |
| Indiana | 6 | 14 | 2 | 27 |
| Total | 6 | 26 | 21 | 53 |

# Suppression Guidelines

- Suppress any aggregates if any population size (N) of less than 5-10

| | < $50K | $50-100K | $100K+ | Total |
|---|---|---|---|---|
| California | * | 12 | 14 | 26 |
| Indiana | 6 | 14 | * | 27 |
| Total | 6 | 26 | 21 | 53 |



# Suppression Guidelines

- Suppress any aggregates if any population size (N) of fewer than 5-10
- If you have <u>any</u> small *n*:
  - Keep suppressing until your total population is greater than threshold
  - Then suppress the next lowest value
- Avoid using subtotals if any subgroup is starred

| | < $50K | $50-100K | $100K+ | Total |
|---|---|---|---|---|
| California | * | 12 | 14 | * |
| Indiana | * | 14 | * | * |
| Total | * | 26 | * | 53 |

# Suppression: Data Platforms

Aggregation policies
- Disallows non-aggregate queries on a table
- Suppresses any aggregated results with fewer than *n* rows

---

Nothing built-in
Pre-aggregate results then apply masking logic

---

Use Python visual

---

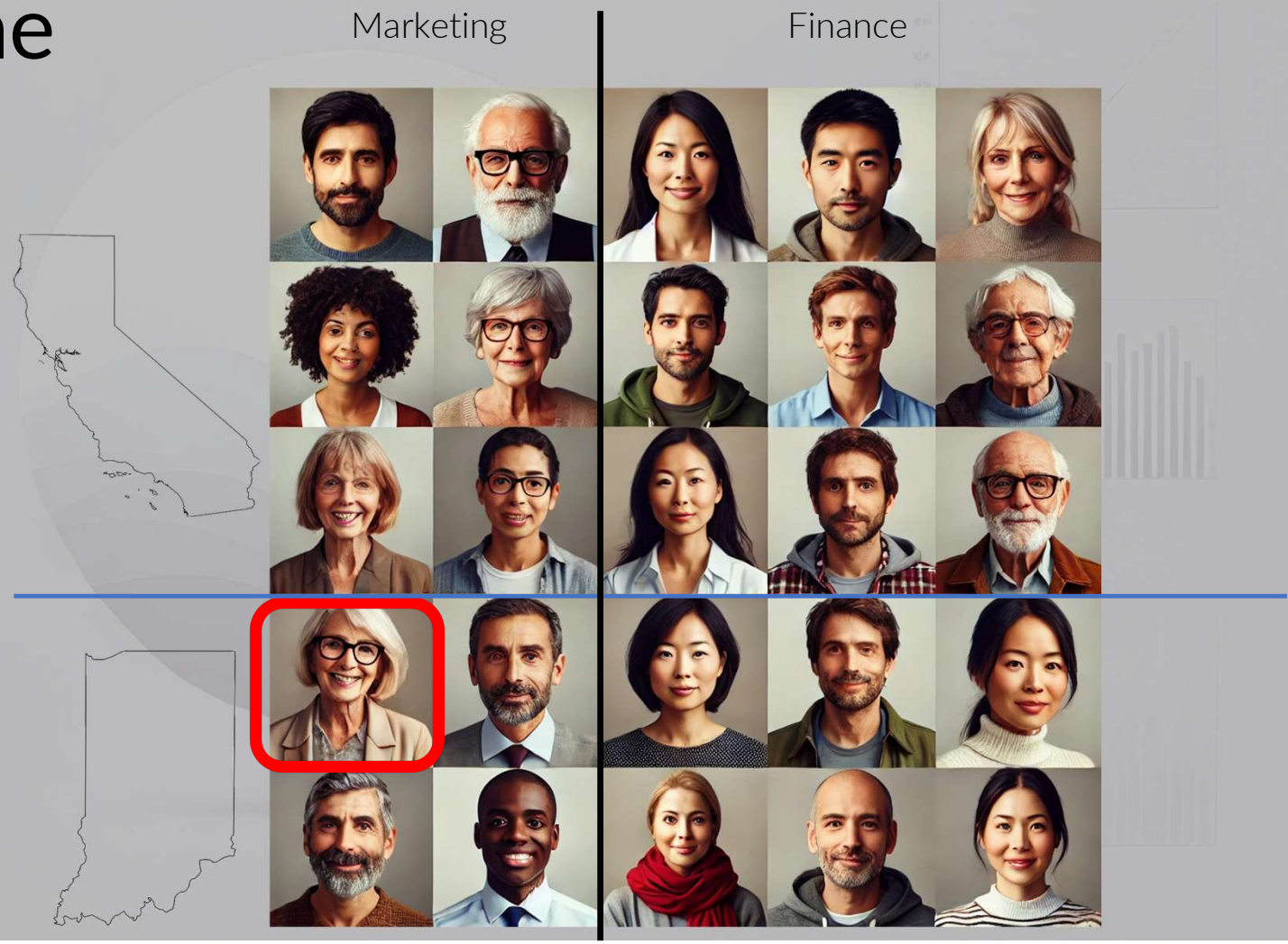
Possible via TabPy extension with some limitations
- Pass the data set in, get a suppressed one back

# A quick game

**Find the salary:**

```
SELECT
    AVG(SALARY),
    COUNT(*)
FROM EMPLOYEES
WHERE . . .
```

Do so with....
2 Queries, return 5+ rows

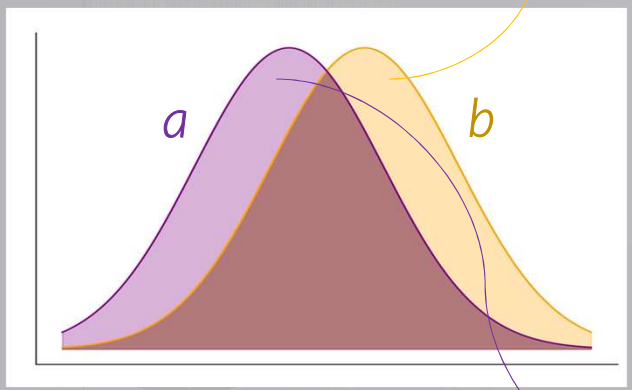



(specifically, a differencing attack)

# Differential Privacy

# Differential Privacy: Definition

- Adding statistical perturbation (noise) to aggregated query results




$a \approx b$

# The What & Why Differential Privacy

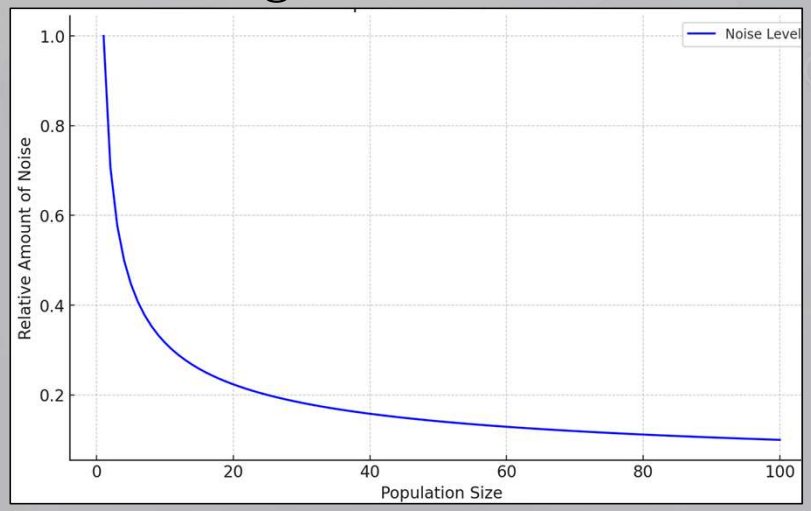
- Meant for data & statistical analysis cases
  - Used in aggregate queries only
- Guards against differencing attacks
  - Danger: combining the DP data set with other or external data
- Still possible to uncover protected attributes through brute force
  - Countermeasure: Privacy Budgets

**Jane Smith**
June 19, 2023

Happy 55th birthday, **John**! You're the best husband ever!

# Differential Privacy: Use

- Results phrased as x ± y or as an interval
  - Average salary: $90,538 ($88,517 - $92,559)
- More people = narrower range




- True differential privacy will infuse some “empty cells” with data

# Differential Privacy: Data Platforms

Snowflake: in public preview soon

Designed with active privacy attacks in mind, so it's rigid

Currently has several limitations; makes machine SQL difficult

---

Notebooks and clean rooms – third-party / open-source libraries

---

May be third party solutions

- Can use in-database Python with PyDP library

---

Can be approximated in both technologies

- More useful against accidental disclosure
- Will not be safe from a privacy attack

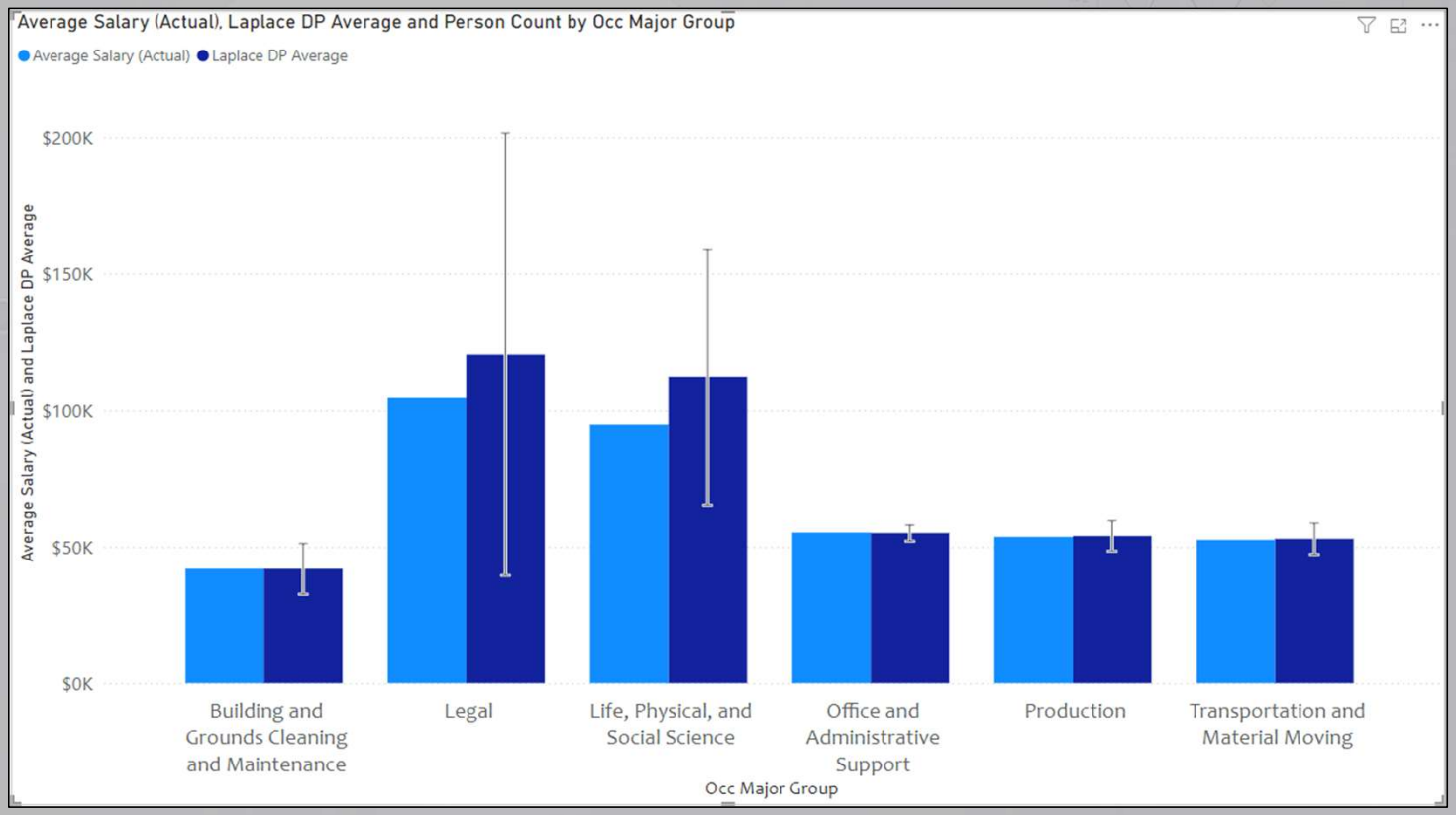
# Diff Privacy: Example

$\epsilon$ = 0.5 (moderate)

Legal:
8 people
- 95% CI: $162K

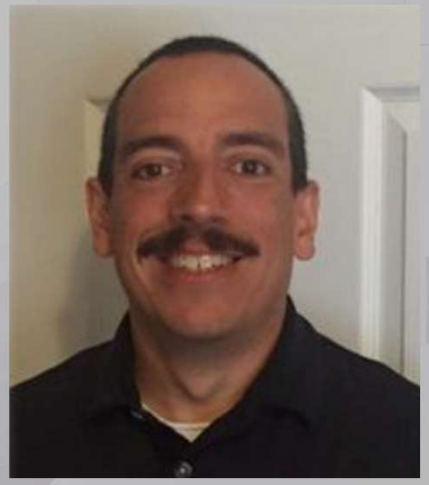
Building … Mtce.:
51 people
- 95% CI: $19K

Office & Admin Support:
240 people
- 95% CI: $6K



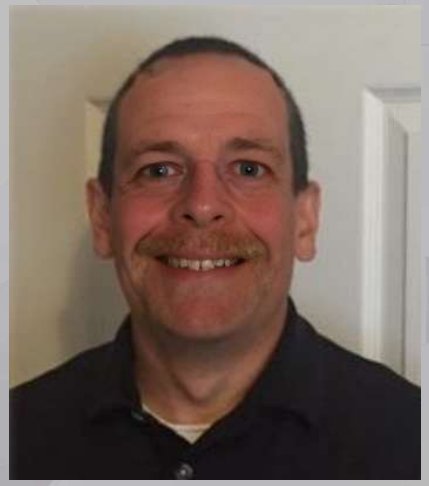
# Synthetic Data

Britton



Warren

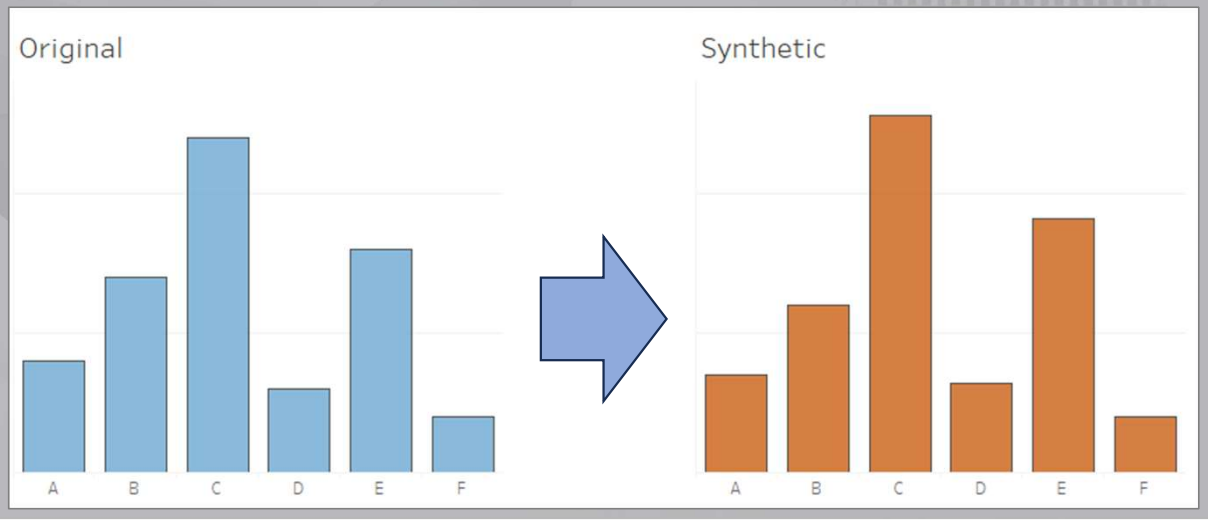
Britren



Warton

# What is it?

- A new data set generated using real data
- Statistically consistent with actual data
- Retains no actual source data (or some if desired)
- Use synthetic data exactly as you would use original data



# How is it created?

Most modern tools use Generative Adversarial Networks

# Tools and Platforms

- Snowflake: In development
- Microsoft may have one in the works
- Paid tools *(some offer free limited-user plans)*

gretel hazy MDCLONE MOSTLY·AI

- Open-source tools

gretel YData datacebo CTGAN datacebo SDV The Synthetic Data Vault

- Some require flat data sets, some allow tabular sets

# Example

MOSTLY·AI

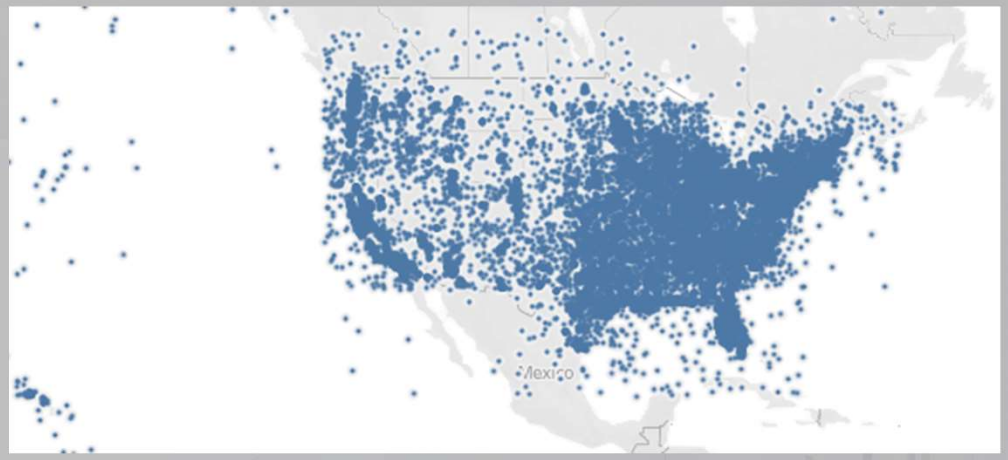
| Table | Primary key | Foreign keys |
|---|---|---|
| DIM_PERSON | DIM_PERSON_KEY | - |

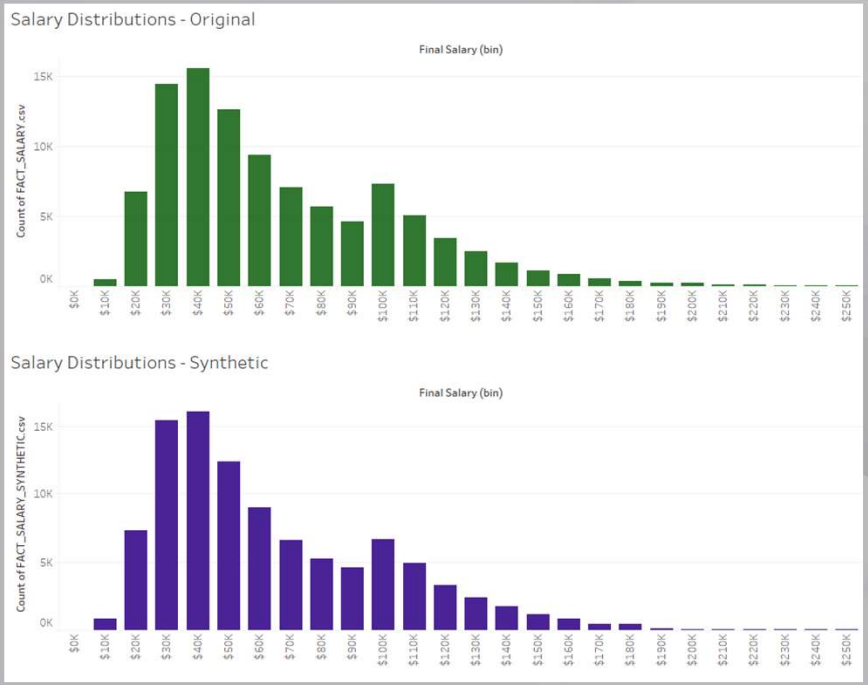
| Include | Name | Encoding type |
|---|---|---|
| ◯ | DIM_PERSON_KEY | Primary key |
| ● | FIRST_NAME | Categorical |
| ● | LAST_NAME | Categorical |
| ● | TYPE | Categorical |
| ● | GENDER | Categorical |
| ● | APPROX_AGE | Numeric: Auto |
| ◯ | POP_SCORE | |
| ● | ADDRESS | Character |
| ● | CITY | Categorical |
| ● | STATE | Categorical |
| ● | ZIP | Categorical |
| ● | LATITUDE | Latitude, Longitude |
| ● | LONGITUDE | Latitude, Longitude |

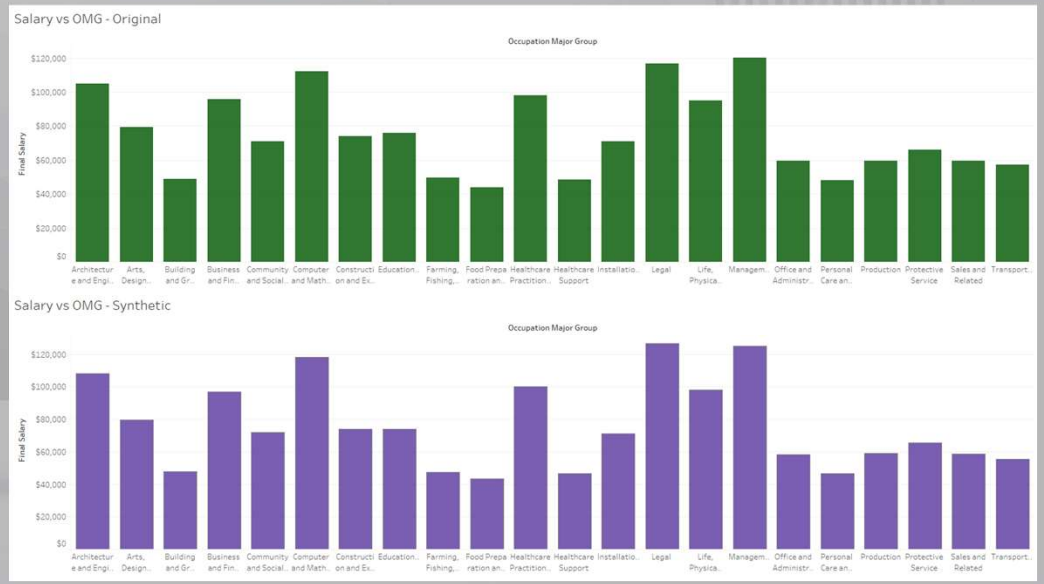
| | | |
|---|---|---|
| ↳FACT_SALARY | - | DIM_PERSON_KEY |

| Include | Name | Encoding type |
|---|---|---|
| ● | DIM_OCCUPATION_KEY | Categorical |
| ◯ | DIM_PERSON_KEY | Foreign key → DIM_PERSON |
| ● | FINAL_SALARY | Numeric: Auto |

| First Name | Last Name | Address | City | State | ZIP |
|---|---|---|---|---|---|
| Bobly | Piderser | 715 Main Street | Gonzales | LA | 70737 |
| Brandon | Brow | 3568 Broadford Street | Garner | NC | 50438 |
| Blaker | McDaug | 7737 Bayside Lanes Road | Valparaiso | IN | 46383 |
| Brianton | Hard | 1801 Main Street | Pottsville | HI | 32347 |
| Bryce | Tissell | 1632 Custer Street | Sumter | SC | 29150 |
| Brian | Morbien | 1513 Ellington Boulevard | Byron | GA | 31008 |
| Bee | Reiscinbr | 1102 Houston Street | Hutto | TX | 78634 |
| Bill | Allis | 1450 Cherry Lane | Raleigh | NC | 27604 |
| Brian | Stape | Null | Centerville | WA | 45459 |
| Bleta | Dunner | 2506 6th Street | Jacksonville | FL | 32224 |



# Comparisons

# Comparison

| | Masking | Suppression | Differential Privacy | Synthetic Data |
|---|---|---|---|---|
| Works at row level | ✅ | ❌ | ❌ | ✅ |
| Implementable in data platforms | ★★★★ | ★★ | ★★ | ❓ |
| Implementable in BI tools | ★★★ | ★ | ★ | Once created: ★★★★ |
| Availability with common BI skillsets | ★★★★ | ★★ | ★ | ★★★ |
| Efficacy against privacy attacks | ★ | ★★ | ★★★ | ★★★★ |

# Thank you!

Questions?

GitHub Repo: https://github.com/IDreamInSQL/data_privacy_resources

brittongray