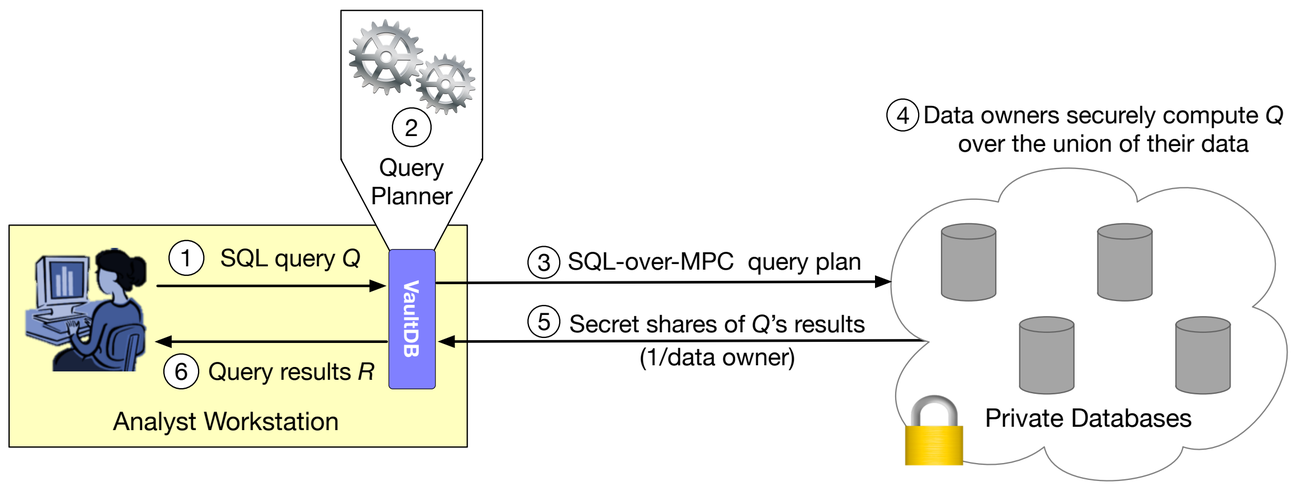
VaultDB: Efficient Query Processing for Private Data Federations

Jennie Rogers, jennie@northwestern.edu

Northwestern

VaultDB: A Private Data Federation

- Data federations let analysts query 2+ datastores as if they were a single engine.
- A *private data federation* queries the union of 2+ DBs s.t:
    - Analyst alone learns query results
    - Data owners learn nothing, store own data locally in the clear



Research Challenges

- How do we make ad-hoc SQL queries run efficiently over 2+ private datastores?
- Can we exploit SQL's declarative semantics to speed up oblivious query processing?
- Can we generalize relational query performance optimization to this setting?
- What trade-offs can we make to improve query runtimes? Accuracy? Privacy?

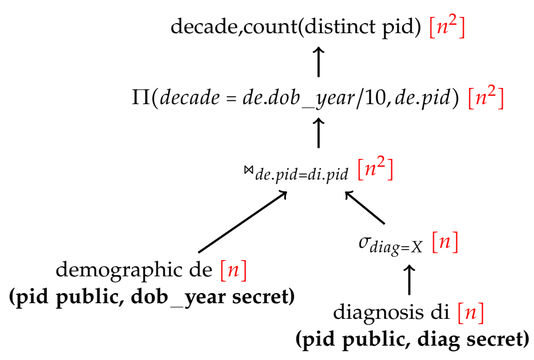
Scientific Impact

- Discover synergies between cryptography and DB research to make SQL-over-MPC queries scalable.
- Generalize principles of DB design to novel optimization opportunities for privacy-preserving analytics
- Investigates fundamental questions about the design of algorithms for efficient oblivious query processing

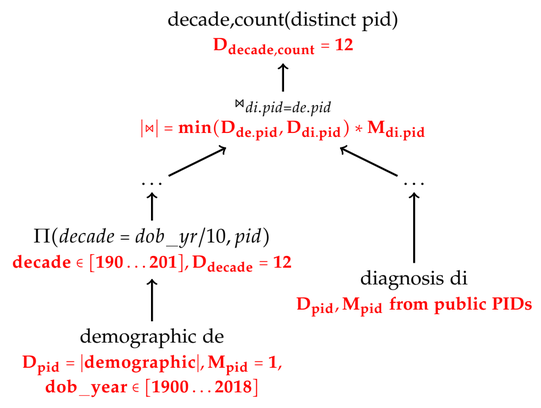
Technique: Intermediate Cardinality Bounds

```
SELECT de.dob_year/10 decade,
       COUNT(DISTINCT de.pid)
FROM demographic de, diagnosis di
WHERE di.diag = "X"
      AND de.pid = di.pid
GROUP BY decade;
```

Naïve Cardinalities



Algebraic Card. Bounds



Research Artifacts

- Open source software releases – enable SQL users to pick up privacy-preserving data analytics
- Preparing pilot with Chicago-area hospitals

CAPriCORN

Education and Outreach

- Formed DB+Security Reading Group (w/Xiao Wang)
- PI is faculty advisor for Grad Women in Computing @Northwestern
- Mentored two HS students for summer 2019

Broader Impact

- Research has immediate implications for:
    - Healthcare
    - Data markets
    - Finance
    - Public policy