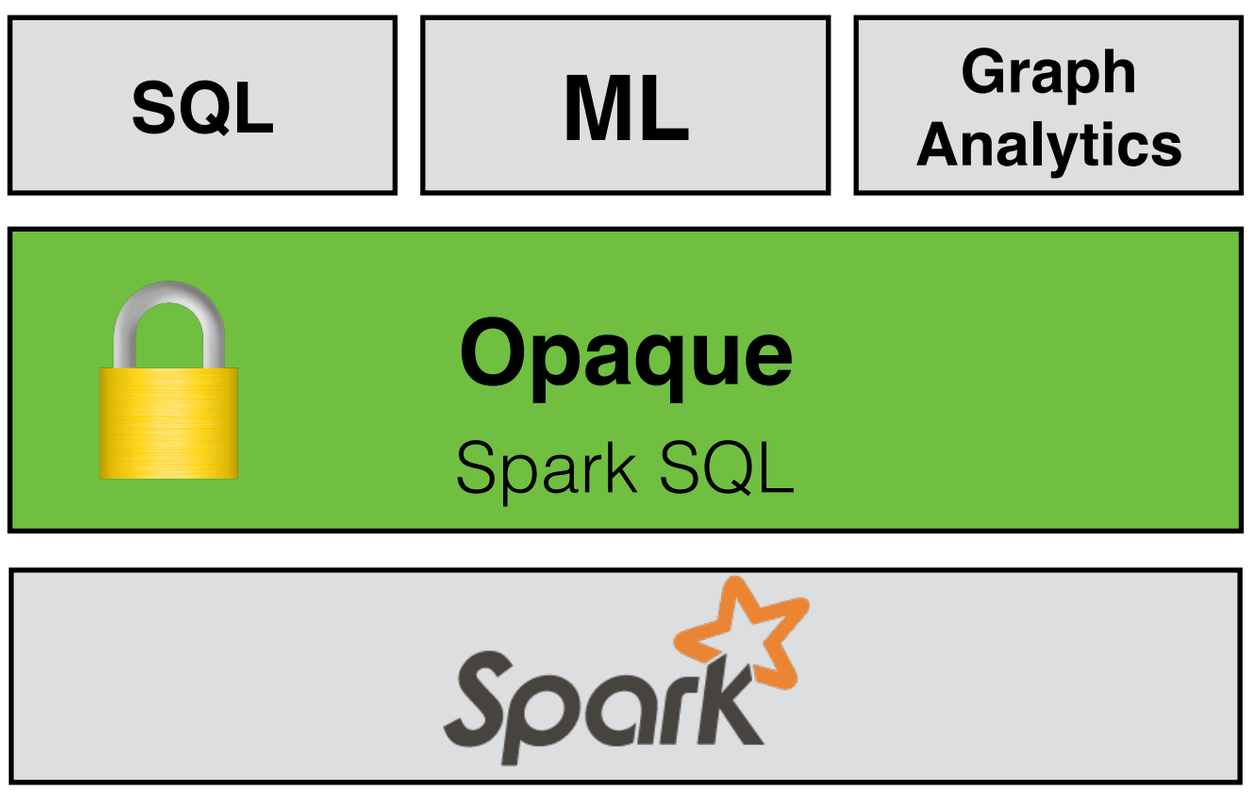
Opaque: An Oblivious and Encrypted Distributed Analytics Platform

[NSDI'17]

Wenting Zheng, Ankur Dave, Jethro Beekman, Raluca Ada Popa, Joseph Gonzalez, and Ion Stoica

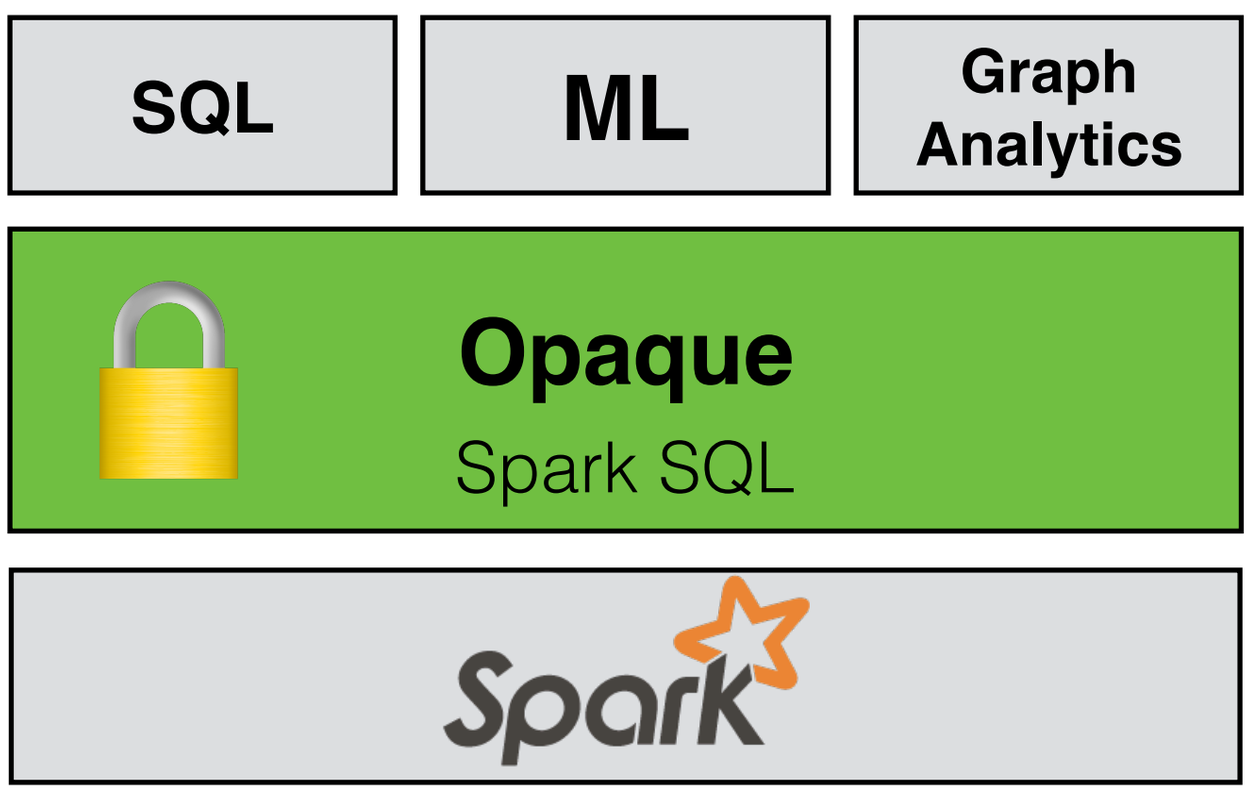
UC Berkeley

Opaque*: oblivious and encrypted distributed analytics platform



* Oblivious Platform for Analytic QUEries

Opaque*: oblivious and encrypted distributed analytics platform

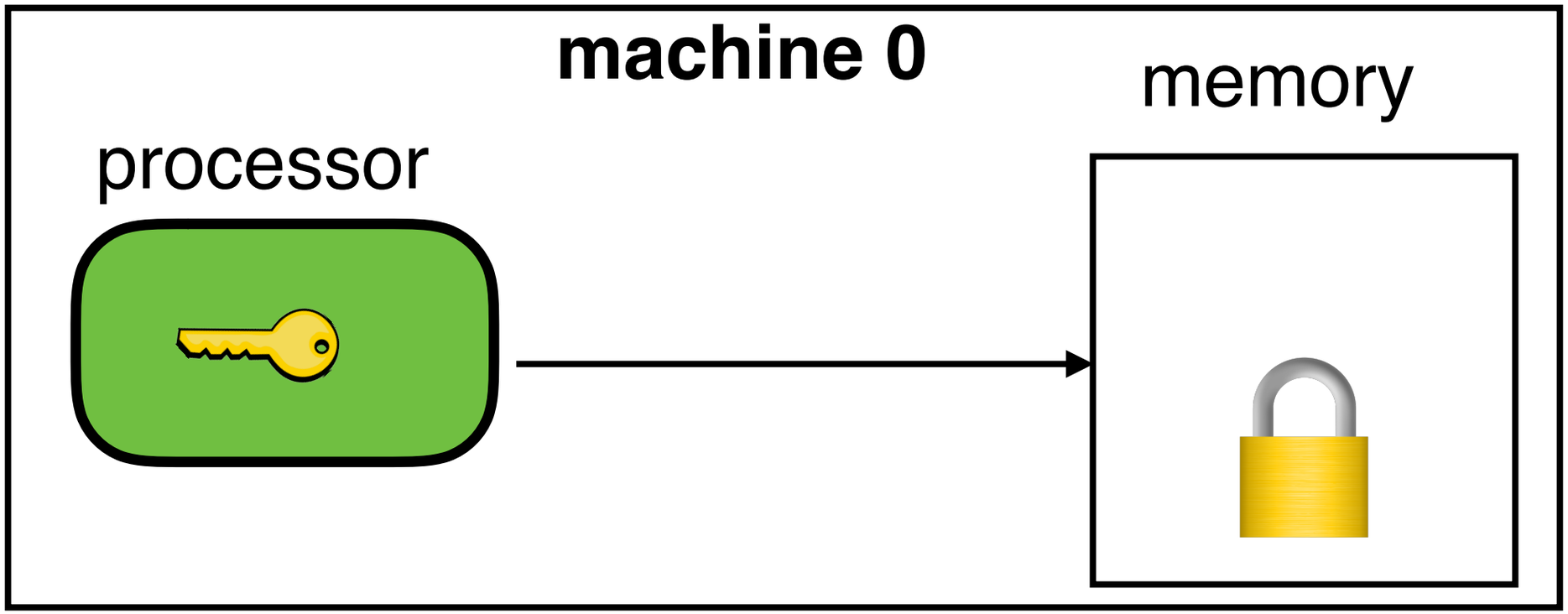


Leverages hardware enclaves (Intel SGX)

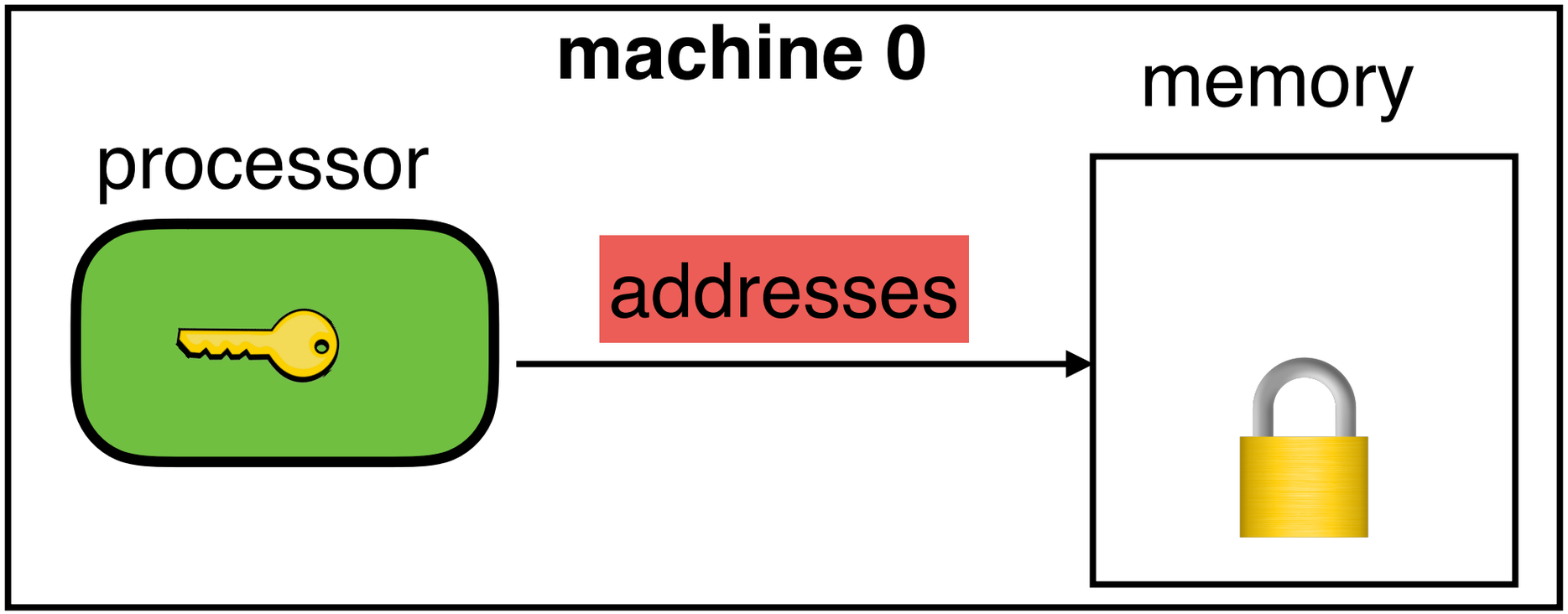
* Oblivious Platform for Analytic QUEries

Access patterns leakage

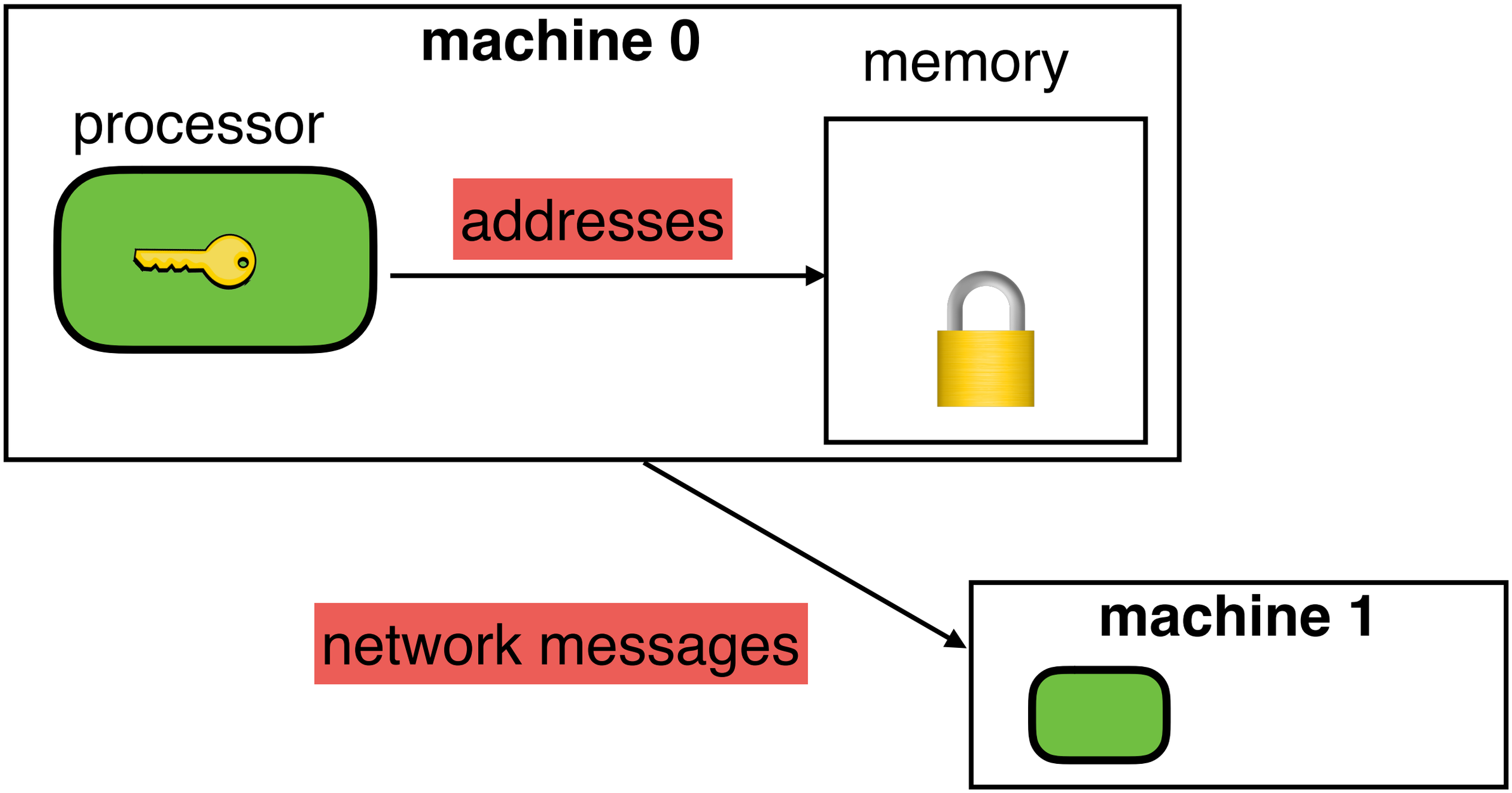
Access patterns



Access patterns



Access patterns



Example: network access pattern leakage

Example: network access pattern leakage

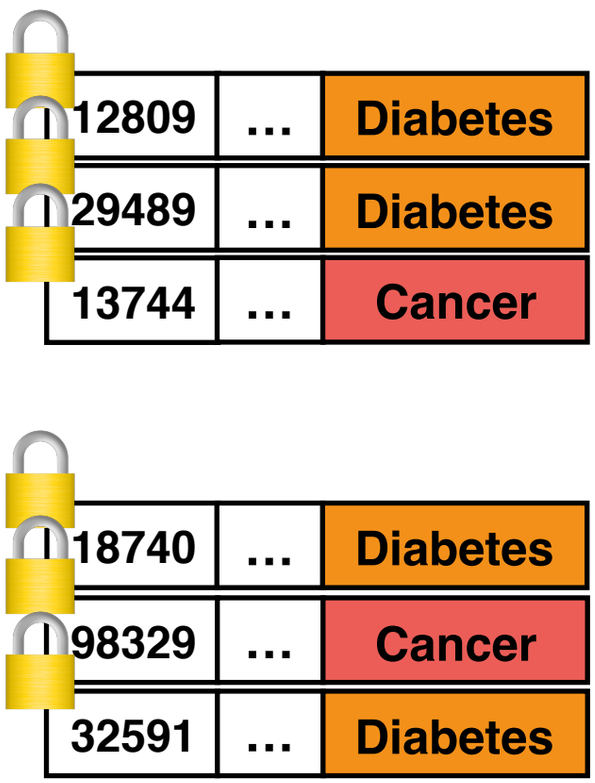
ID	Name	Age	Disease
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
98329	Ronald S. Ogden	53	Cancer
32591	Donna R. Bridges	26	Diabetes

Example: network access pattern leakage

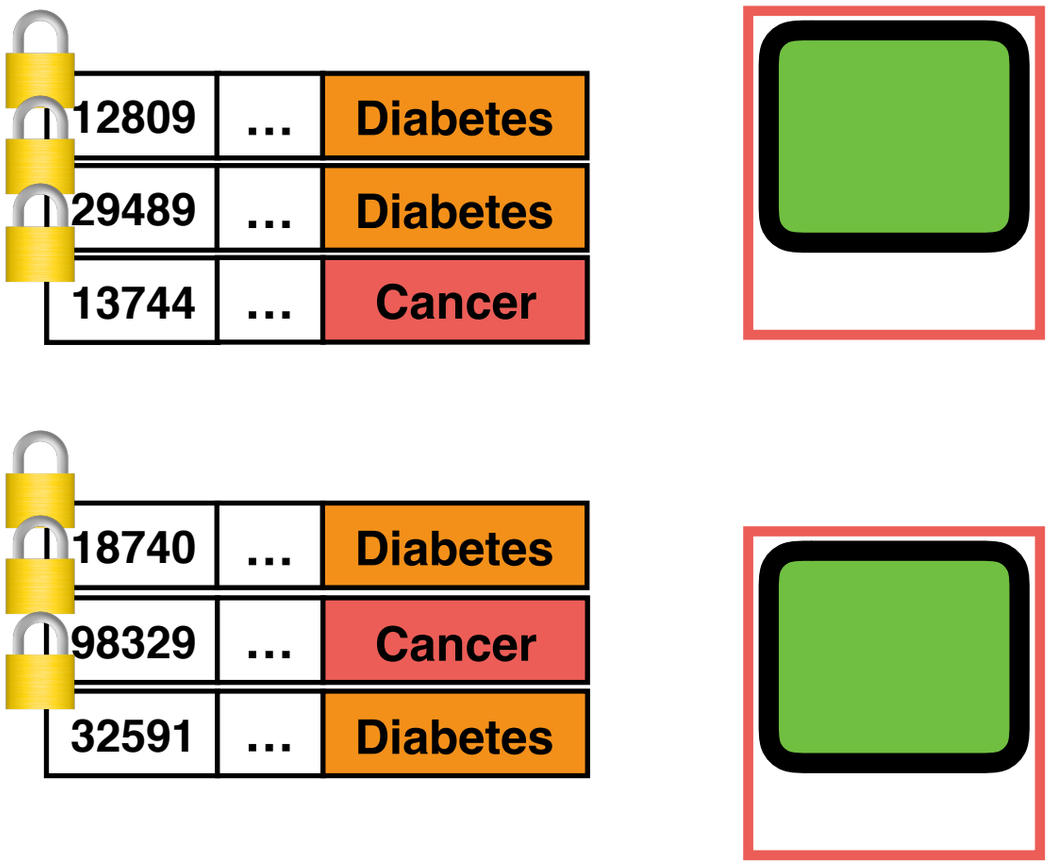
ID	Name	Age	Disease
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
98329	Ronald S. Ogden	53	Cancer
32591	Donna R. Bridges	26	Diabetes

SELECT count(*) FROM medical
GROUP BY disease

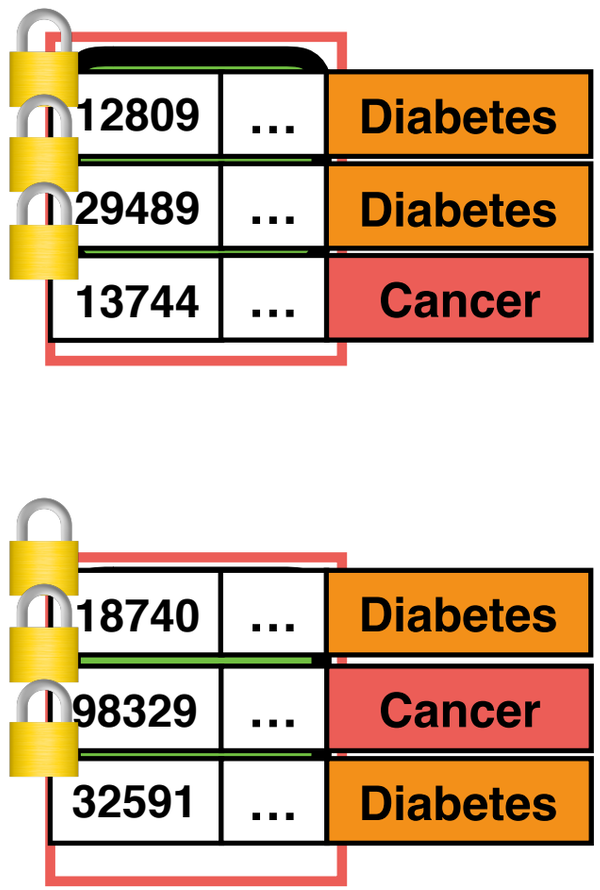
Example: network access pattern leakage



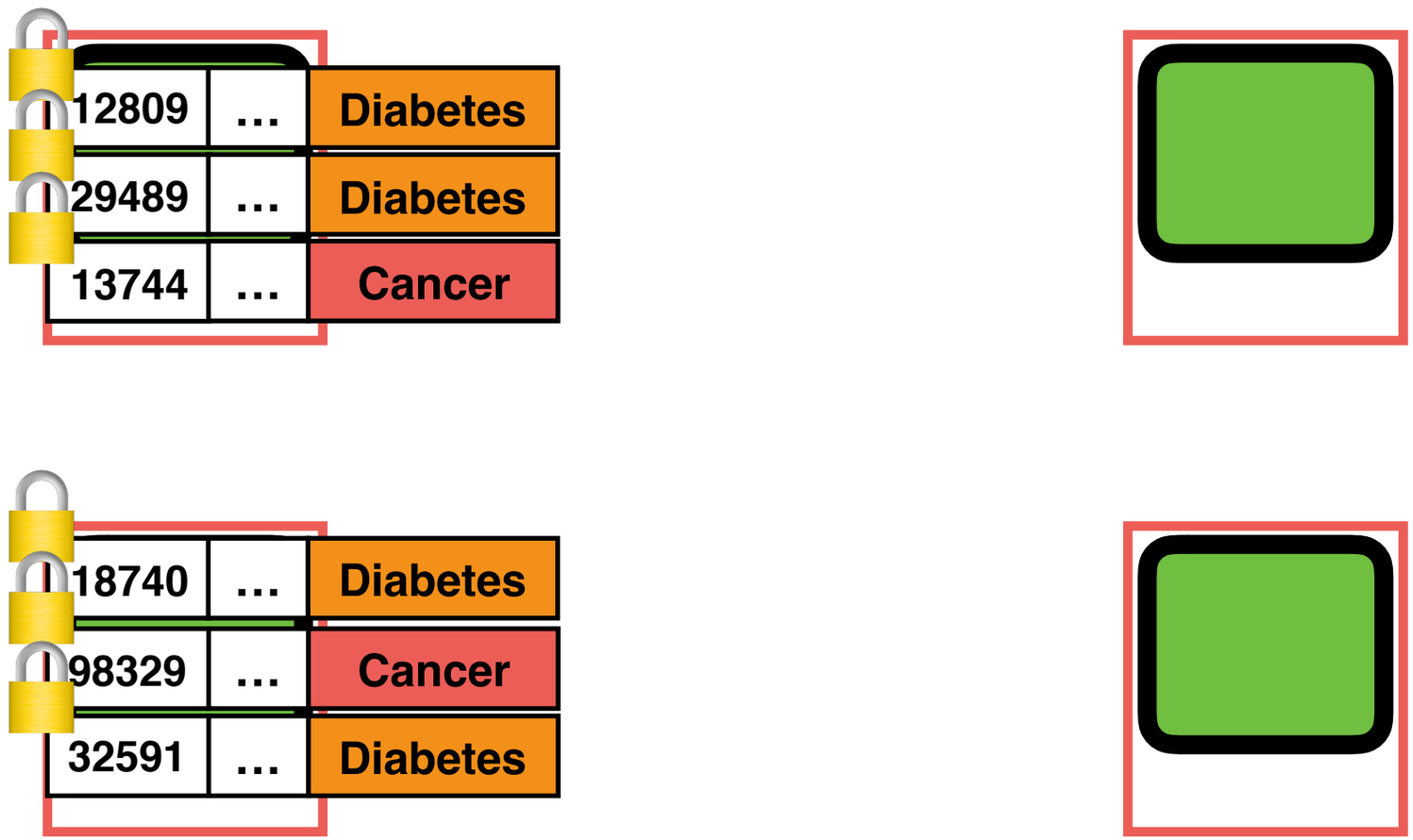
Example: network access pattern leakage



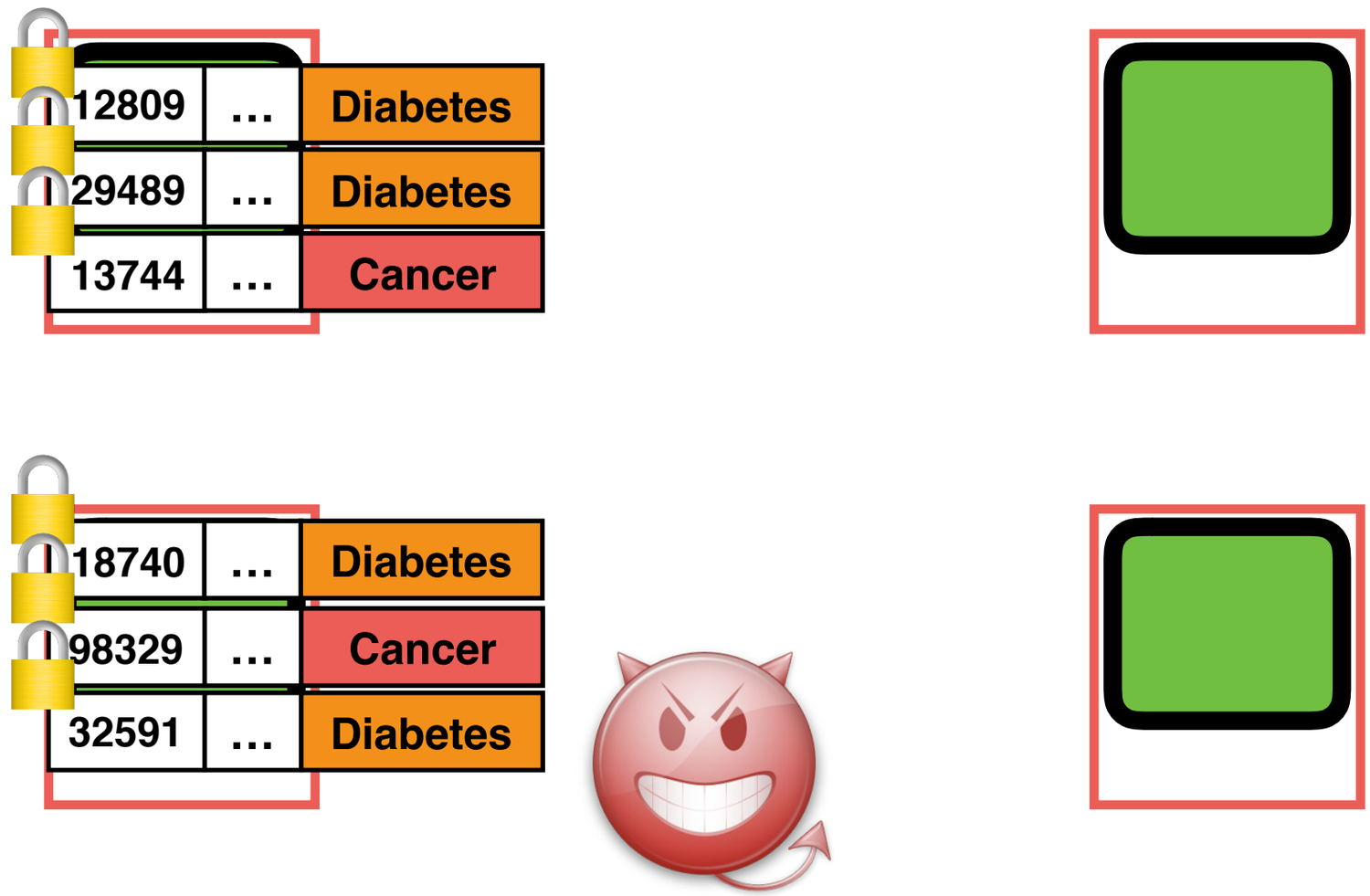
Example: network access pattern leakage



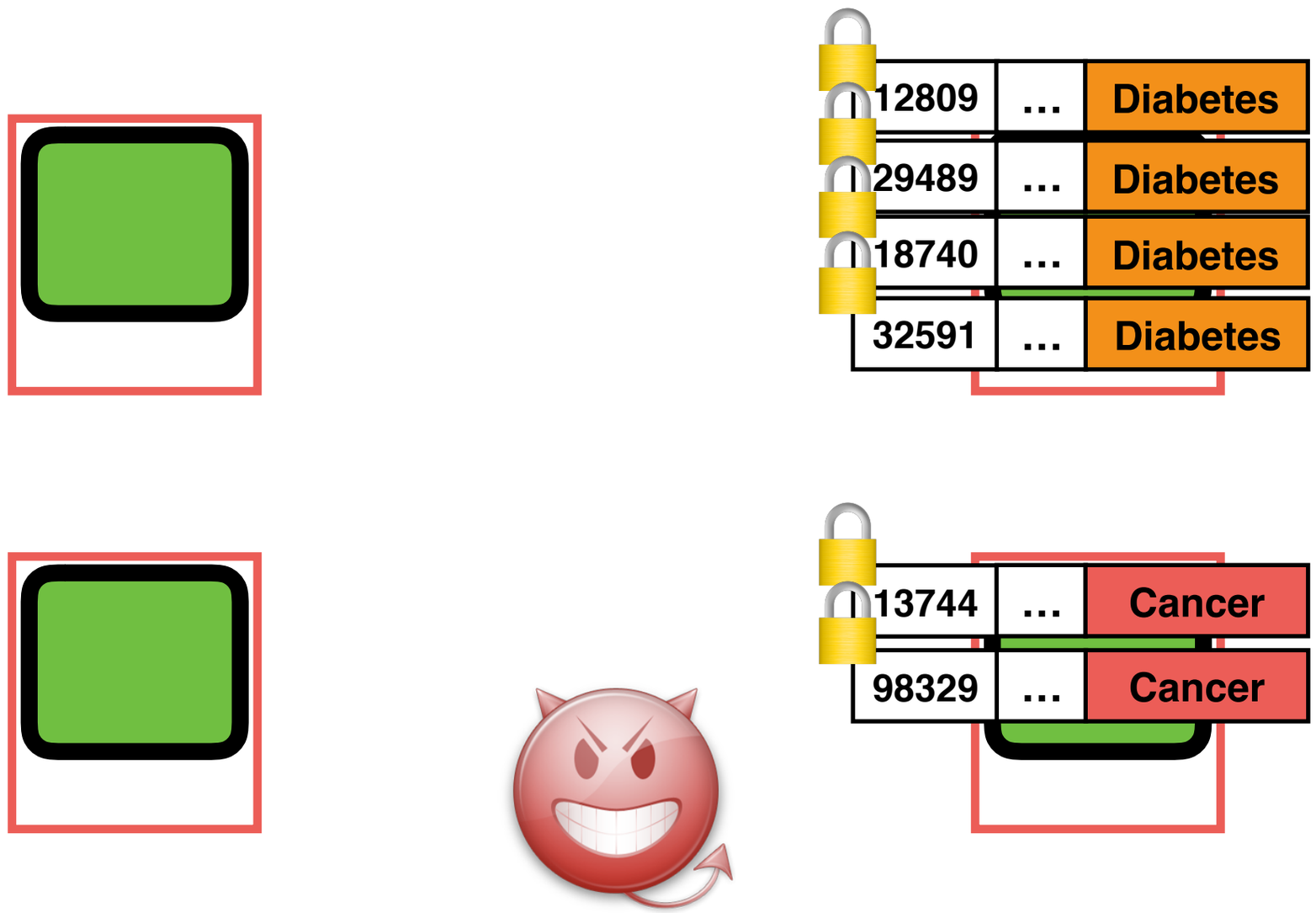
Example: network access pattern leakage



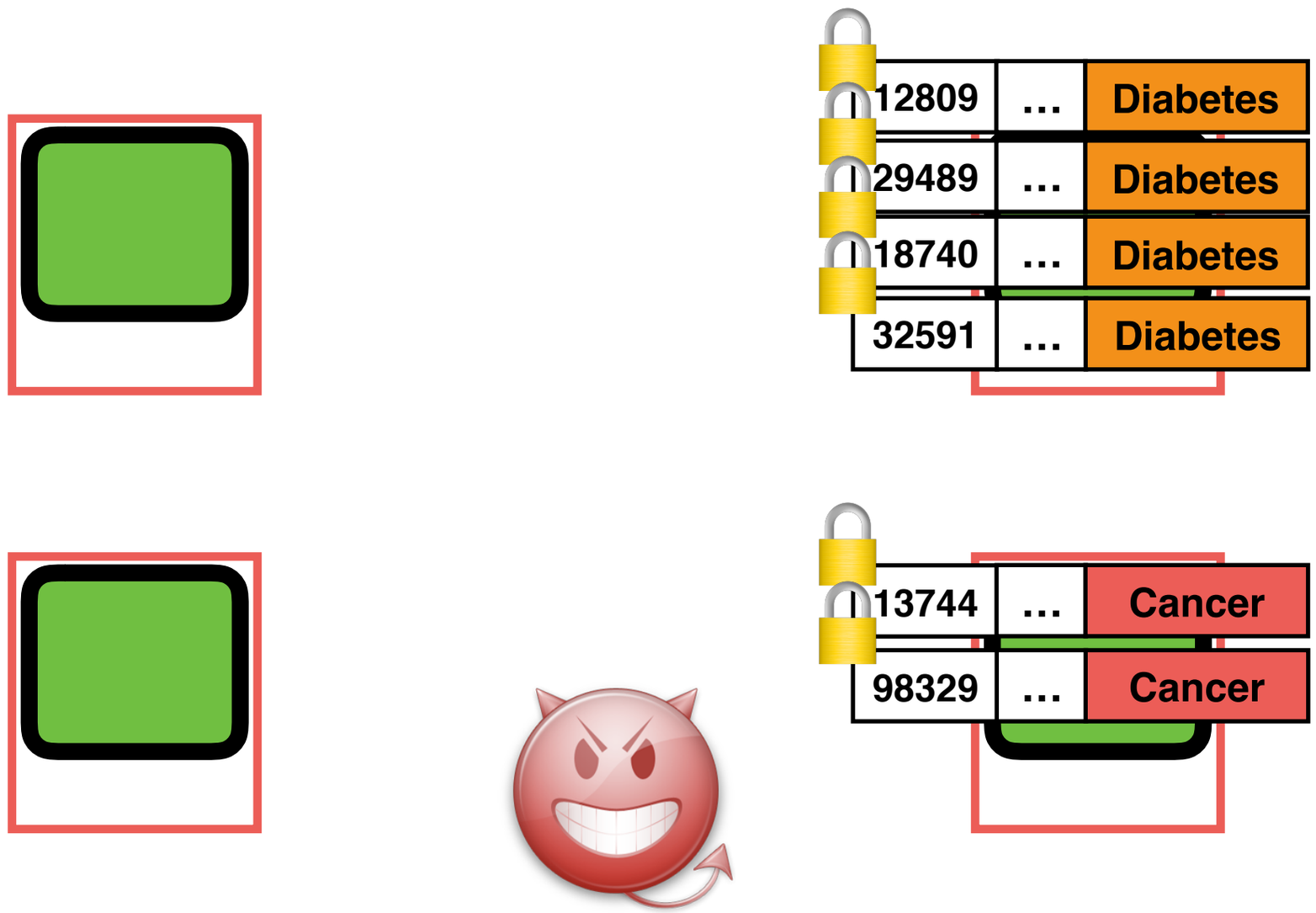
Example: network access pattern leakage



Example: network access pattern leakage

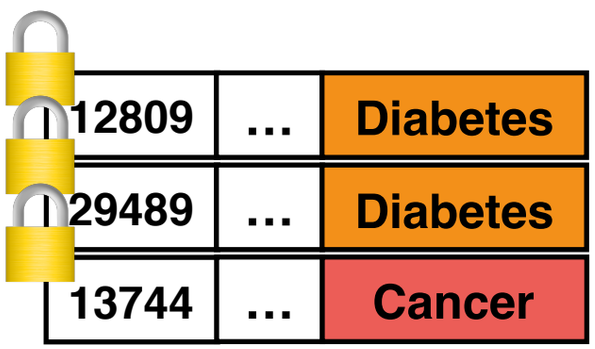


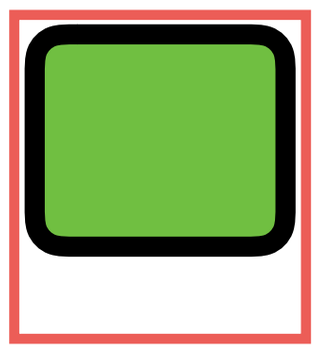
Example: network access pattern leakage

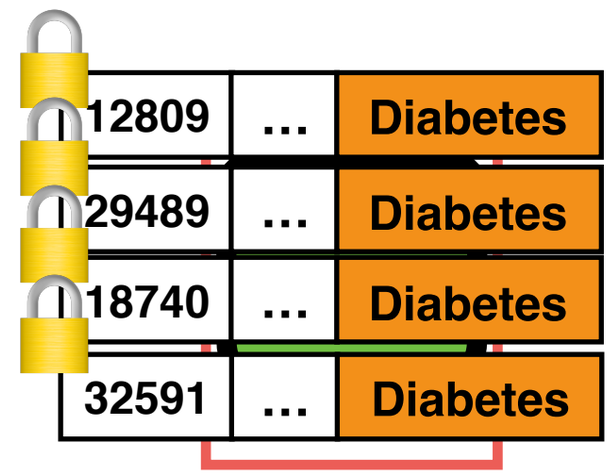


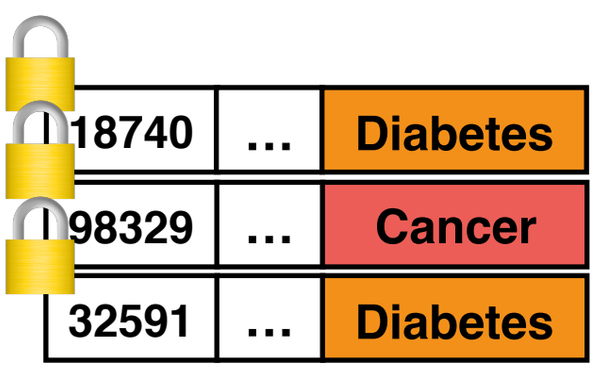
Public information:
Diabetes twice as common as cancer

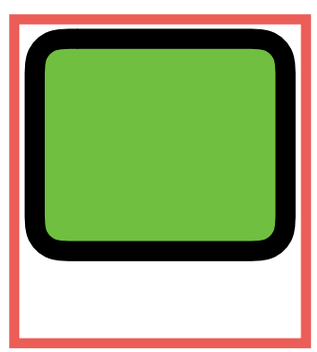
Example: network access pattern leakage

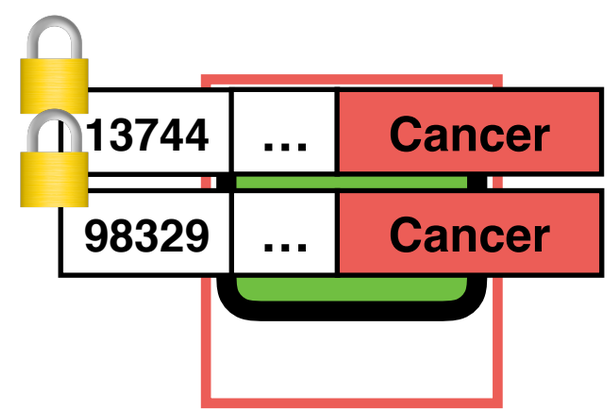






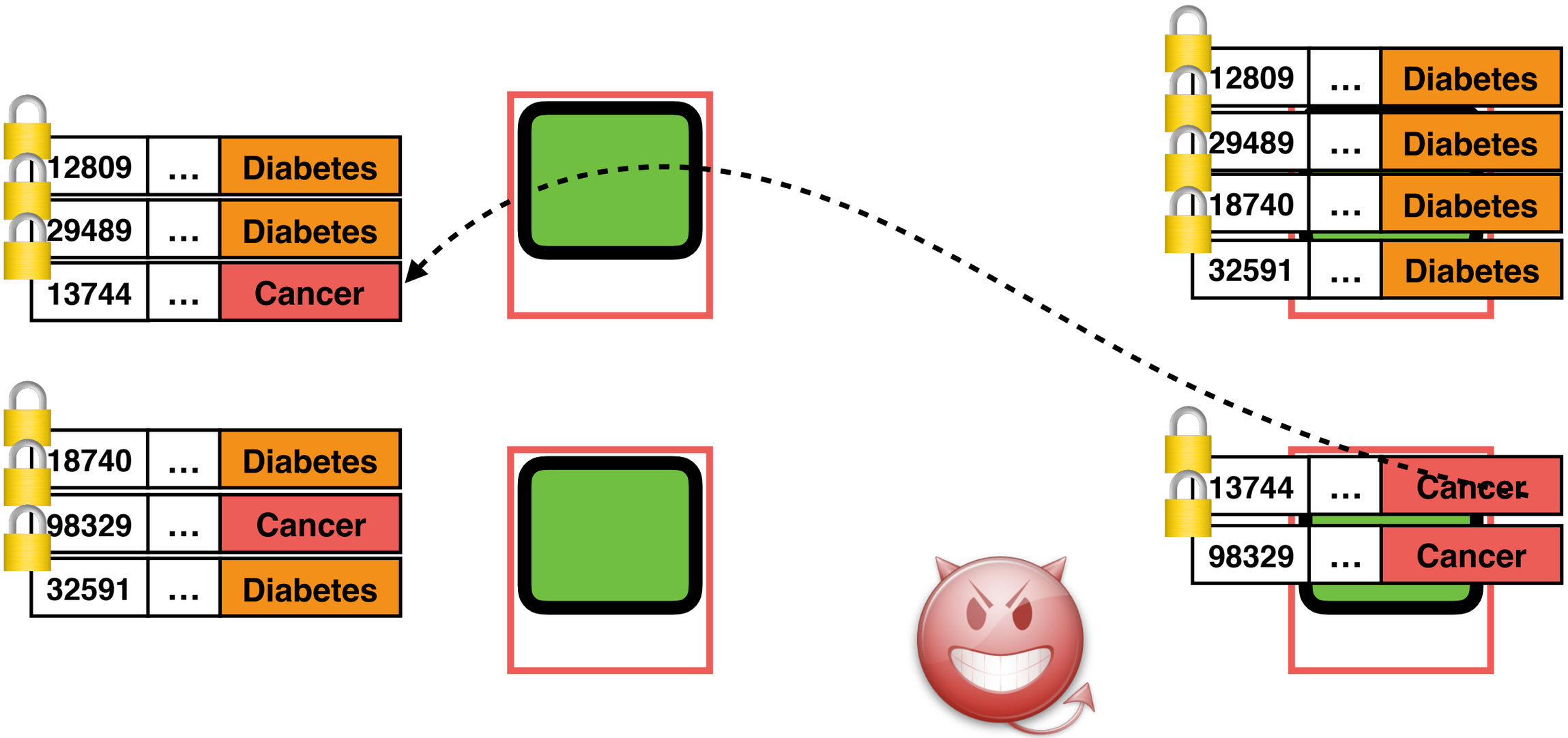






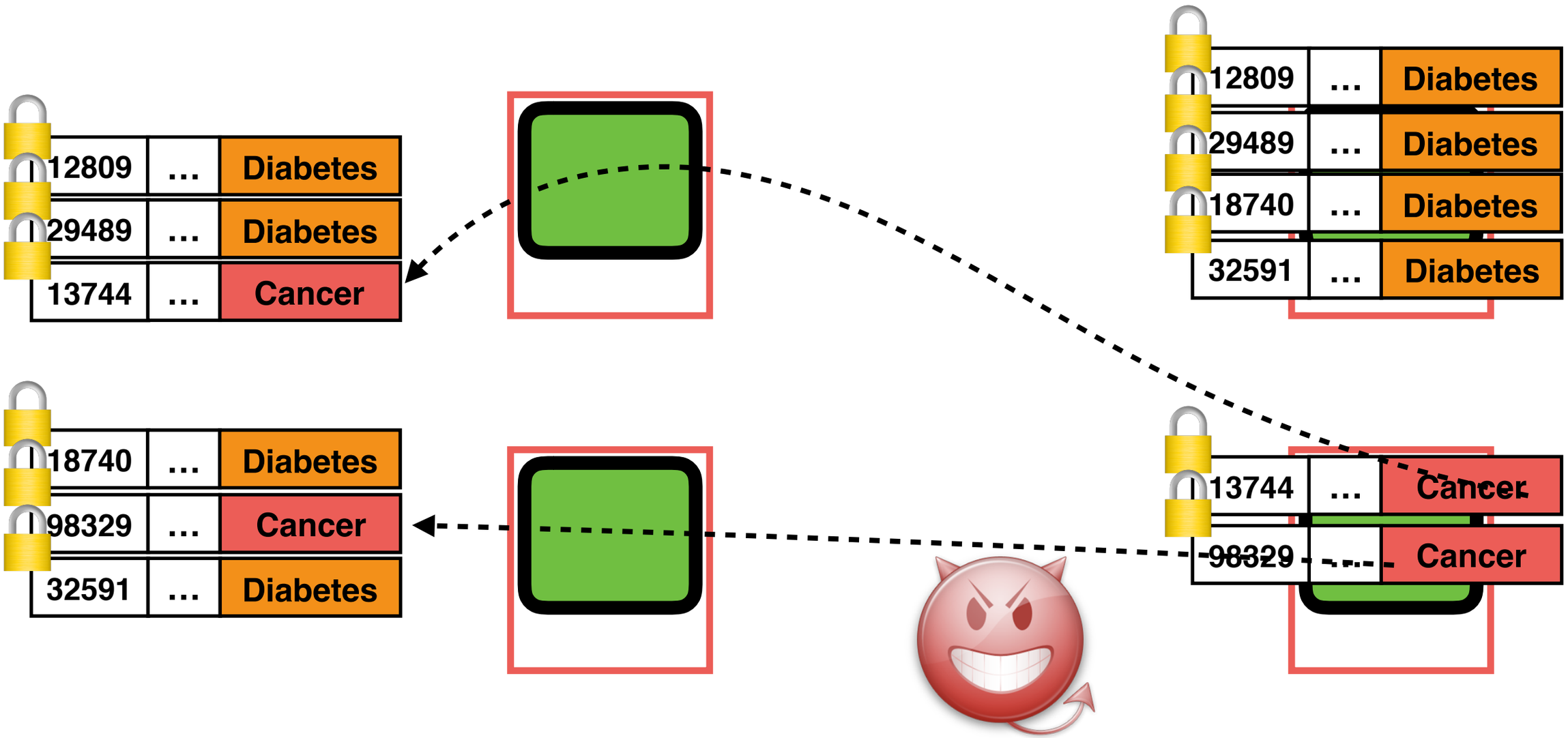
Public information:
Diabetes twice as common as cancer

Example: network access pattern leakage



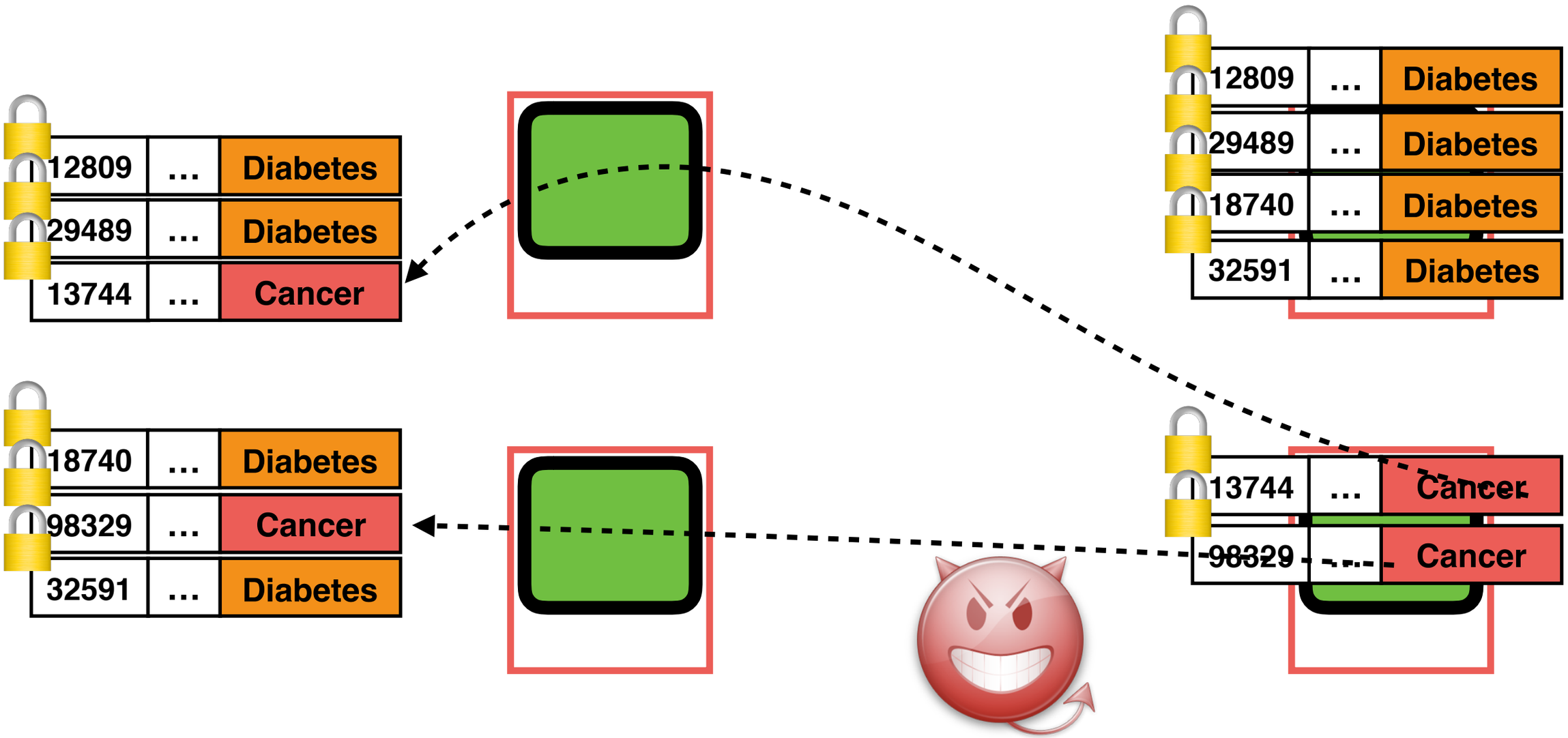
Public information:
Diabetes twice as common as cancer

Example: network access pattern leakage

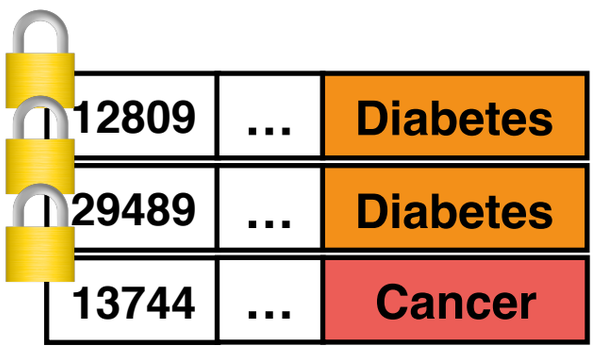


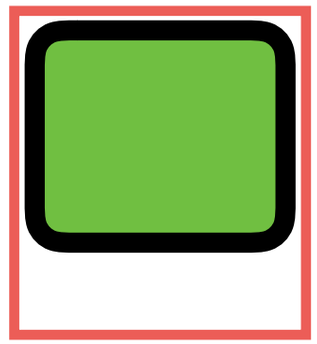
Public information:
Diabetes twice as common as cancer

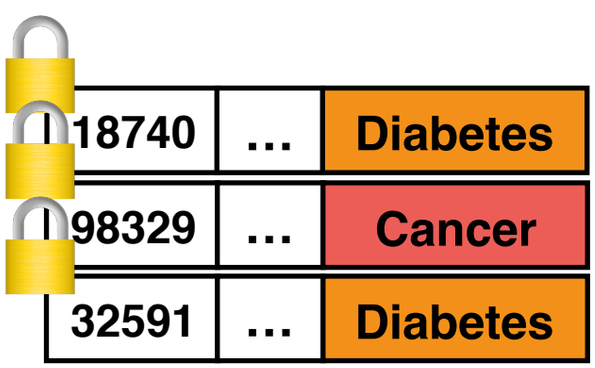
Example: network access pattern leakage

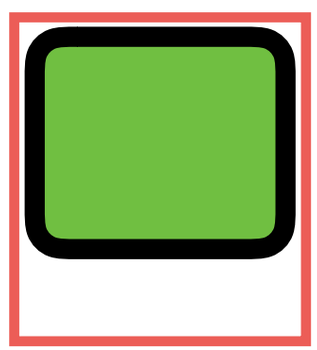


Example: network access pattern leakage

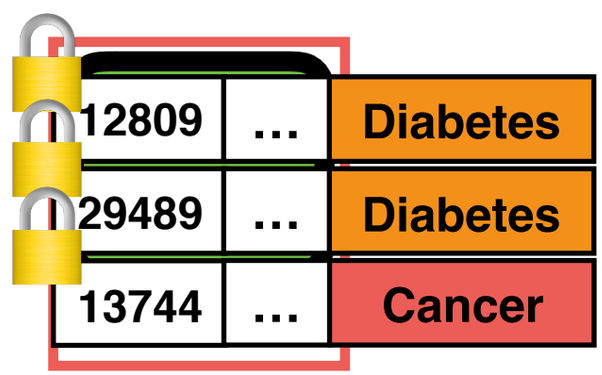


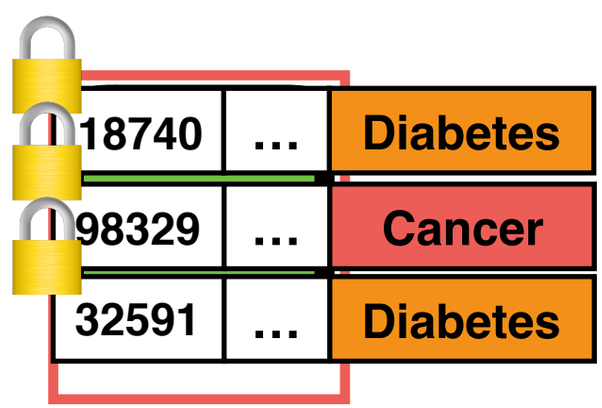




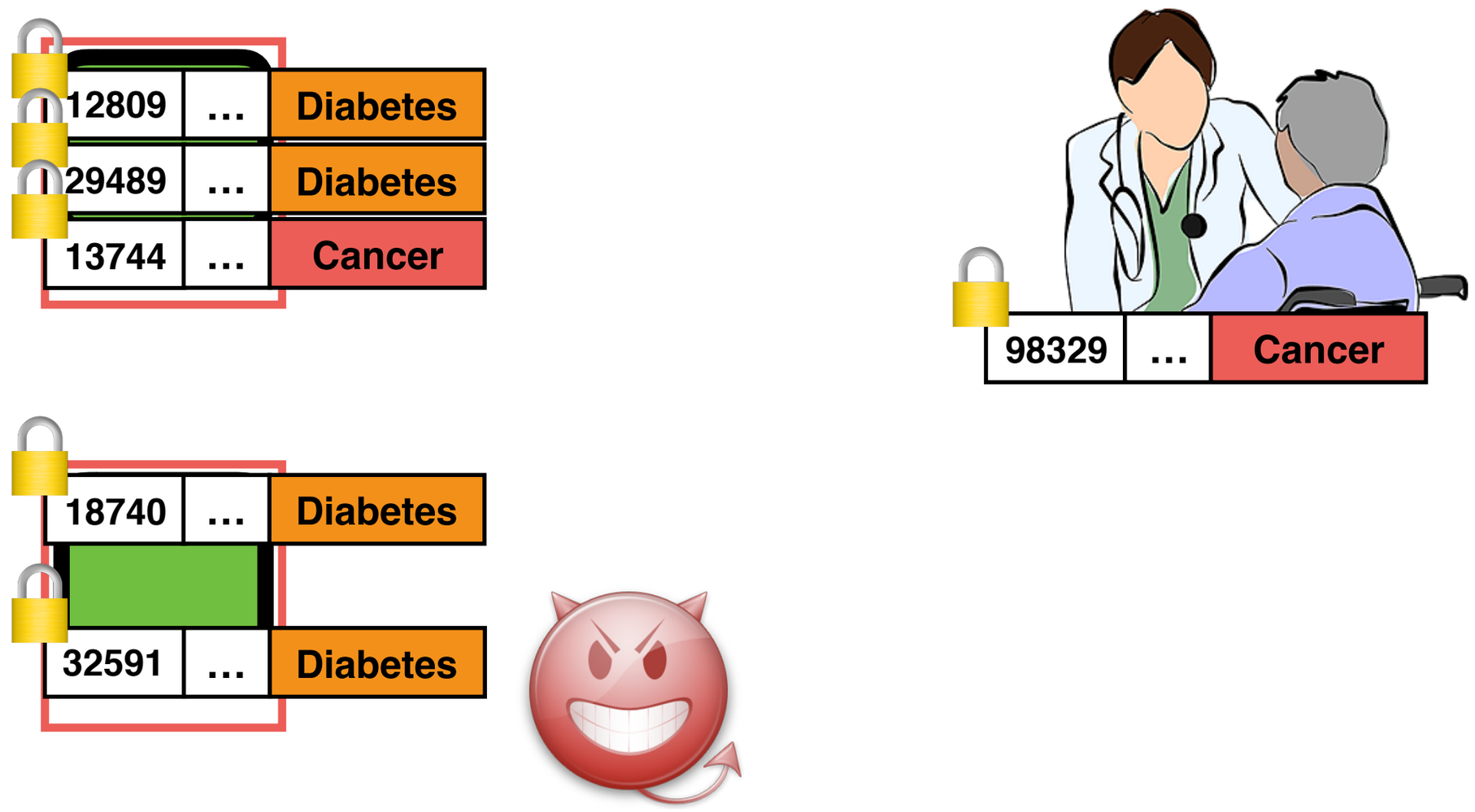


Example: network access pattern leakage

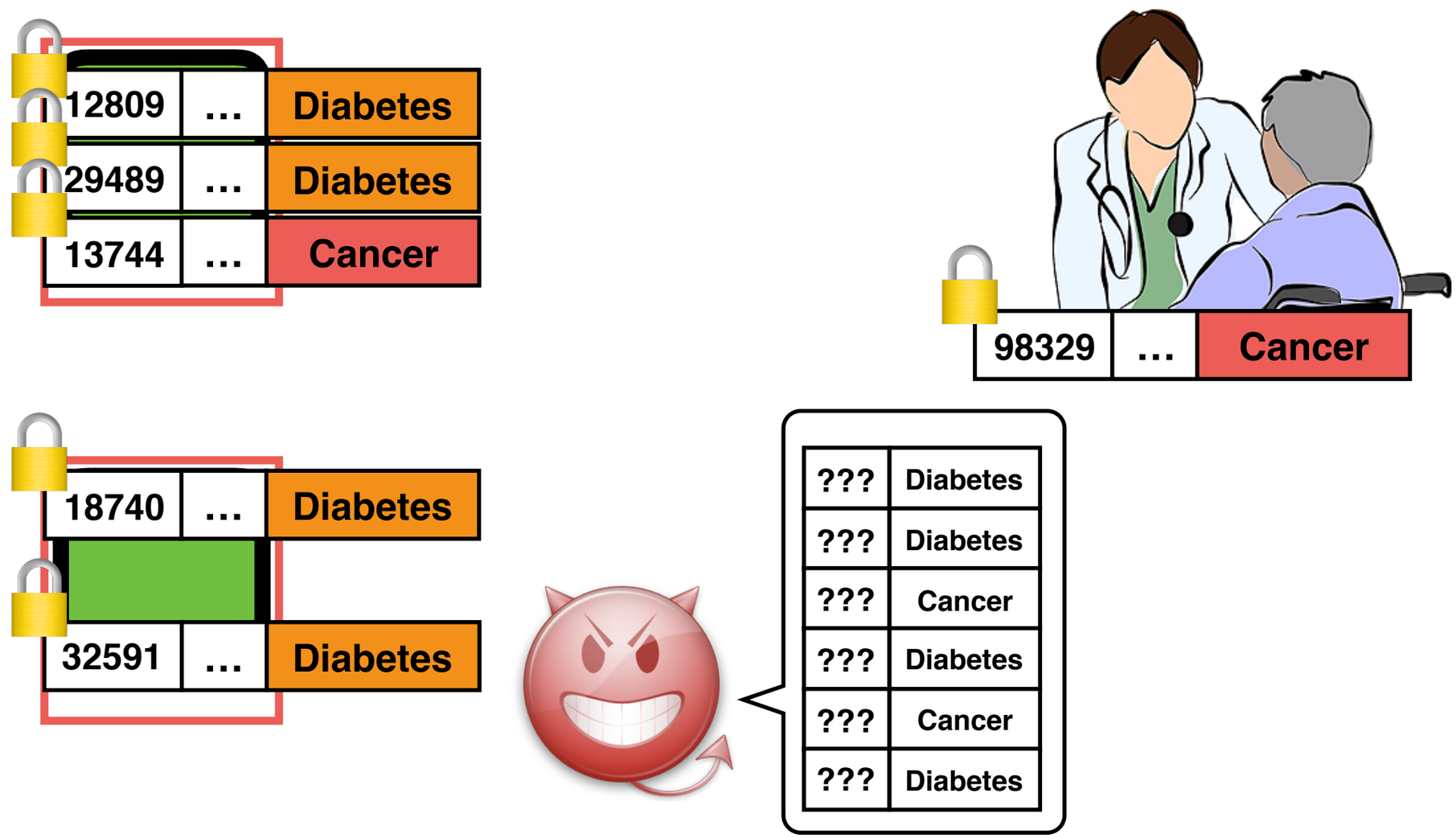




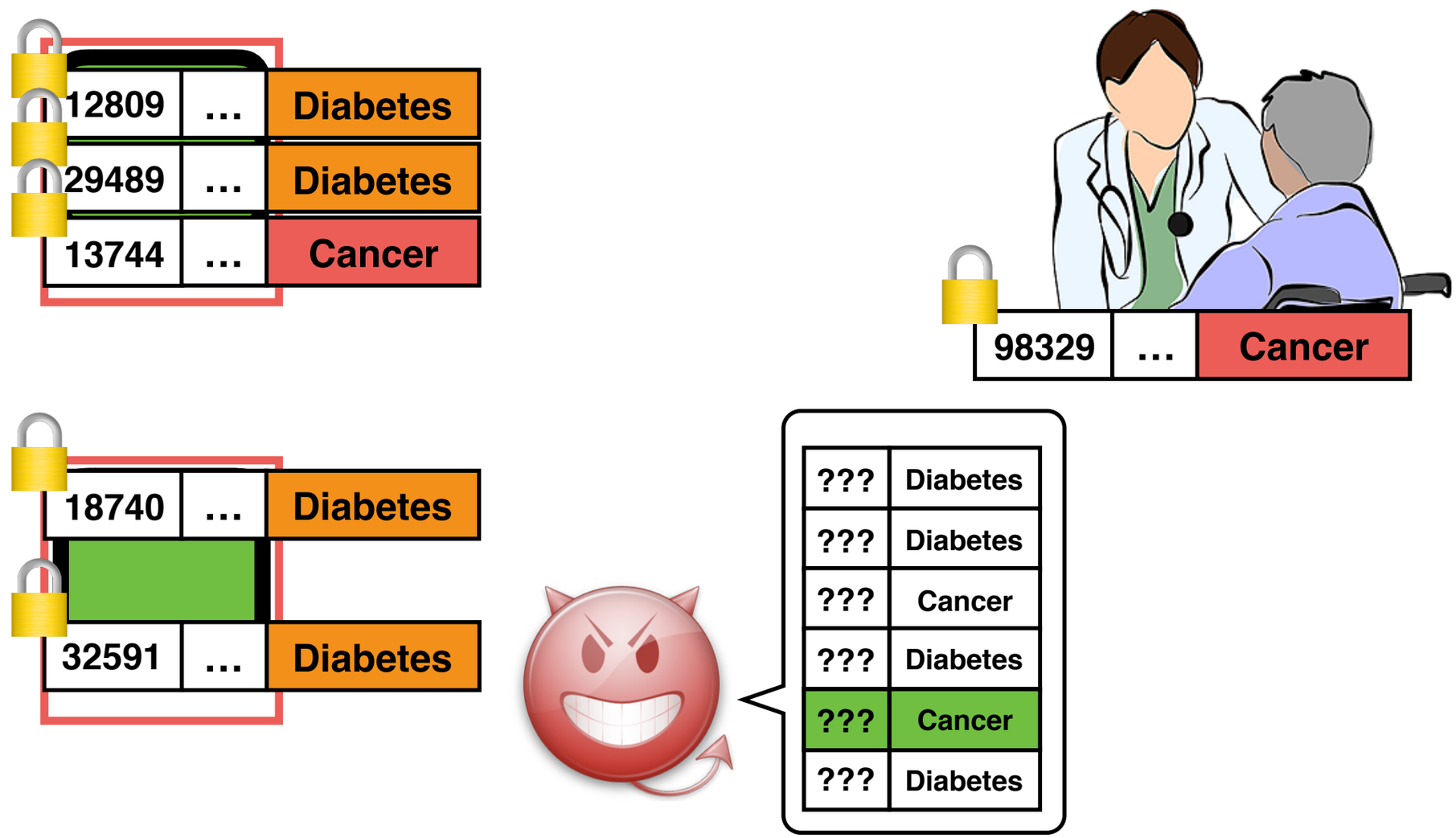
Example: network access pattern leakage



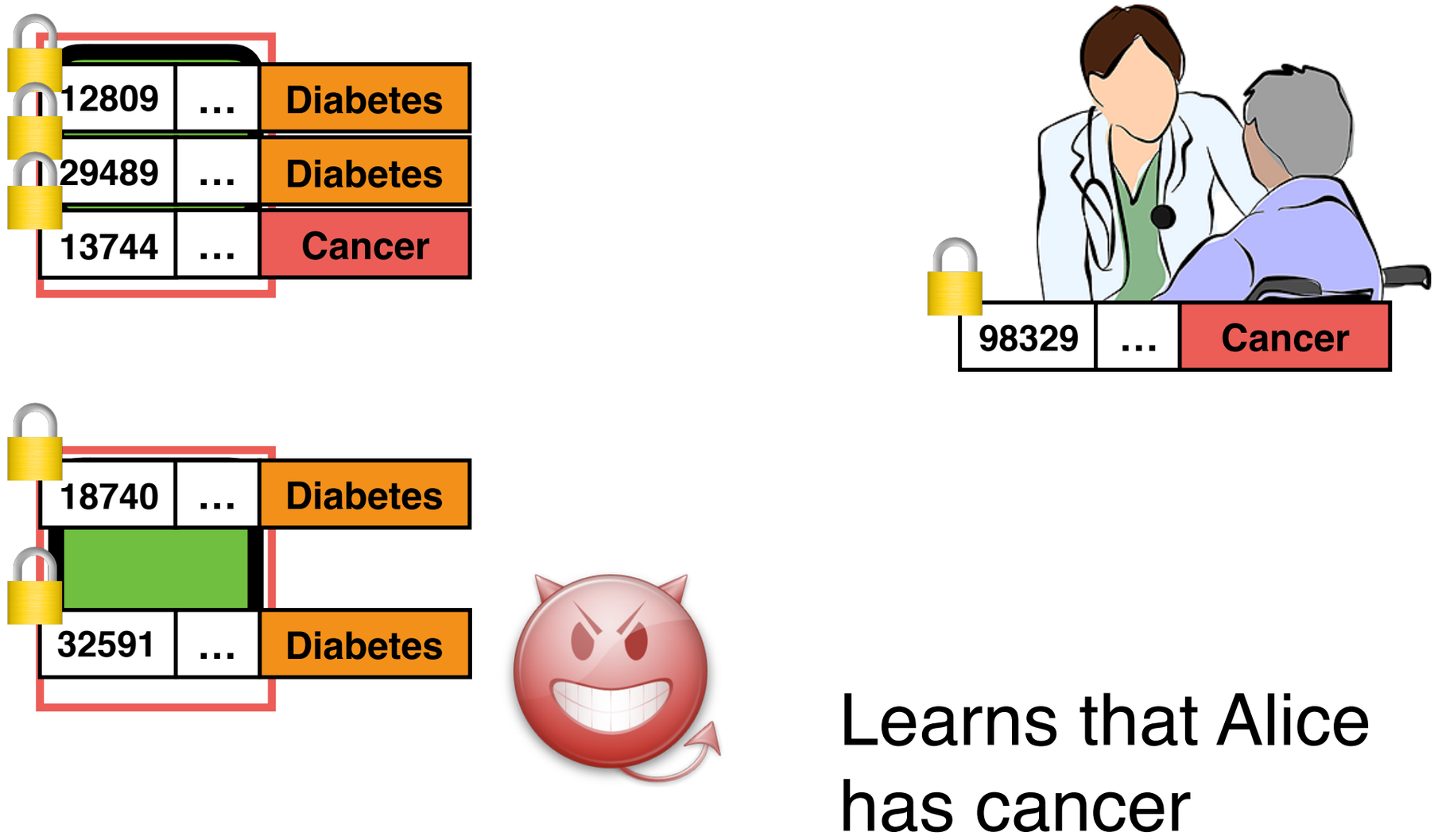
Example: network access pattern leakage



Example: network access pattern leakage



Example: network access pattern leakage

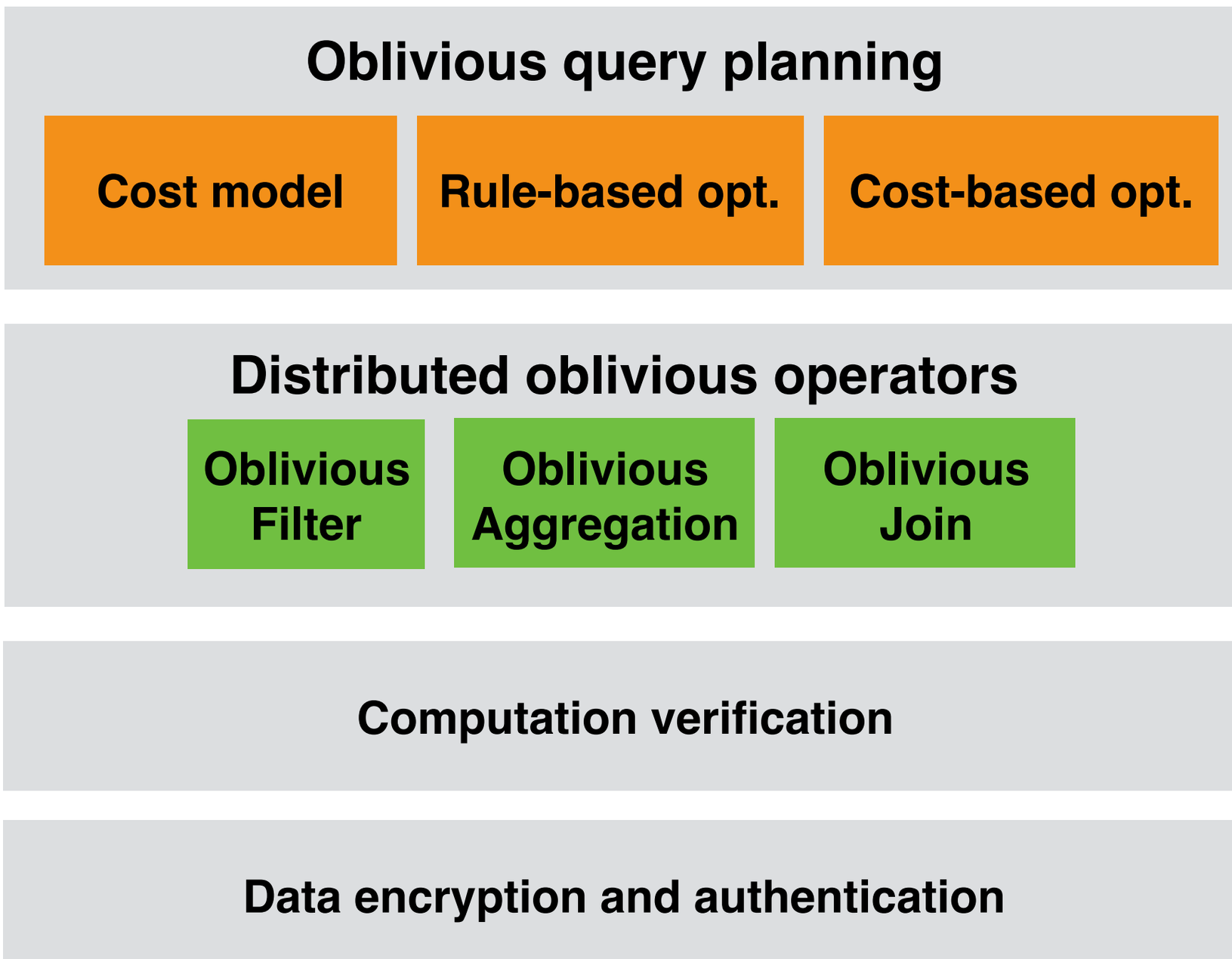


Goal: oblivious distributed analytics

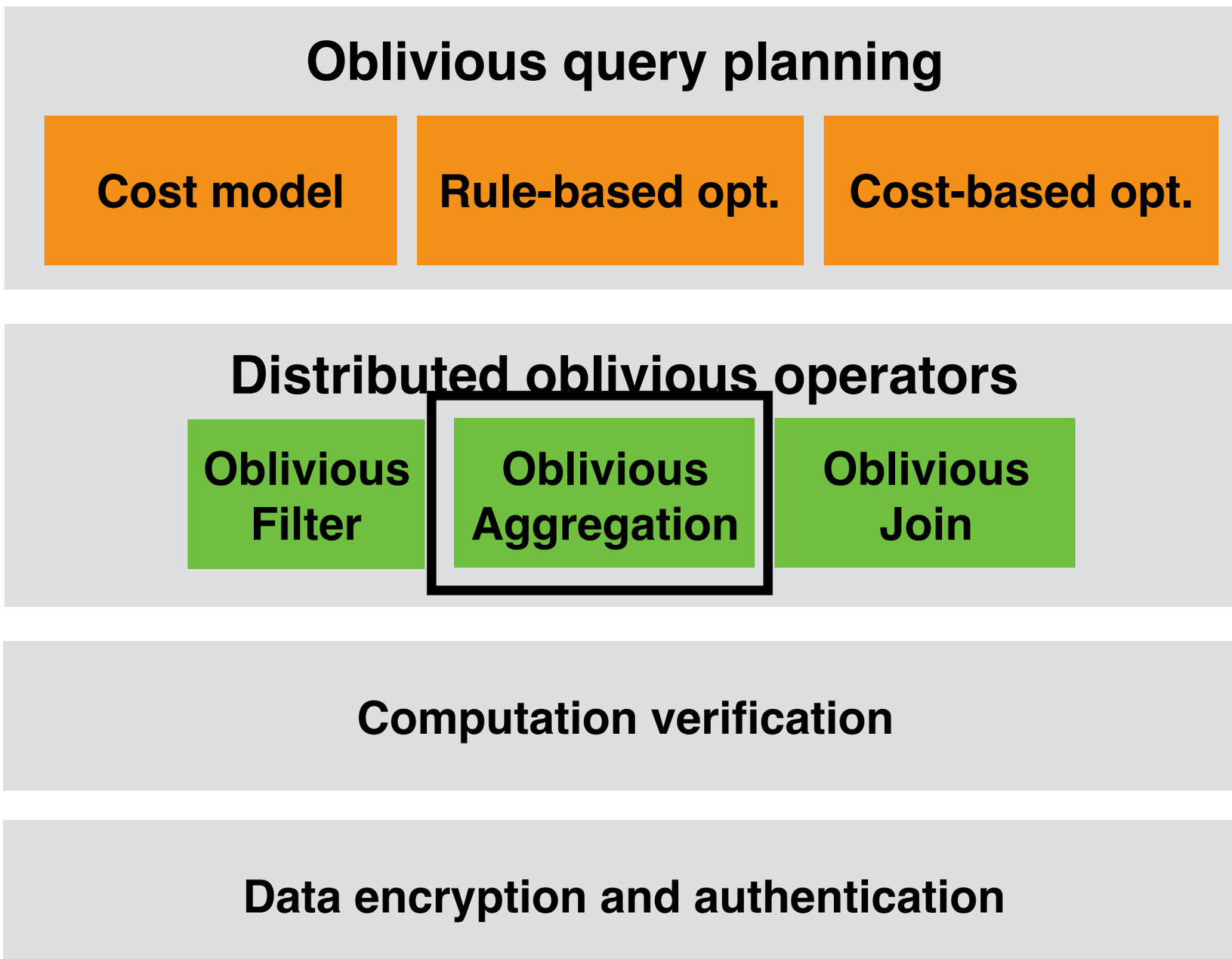
access patterns are independent of data content

Goal: oblivious distributed analytics

Opaque components

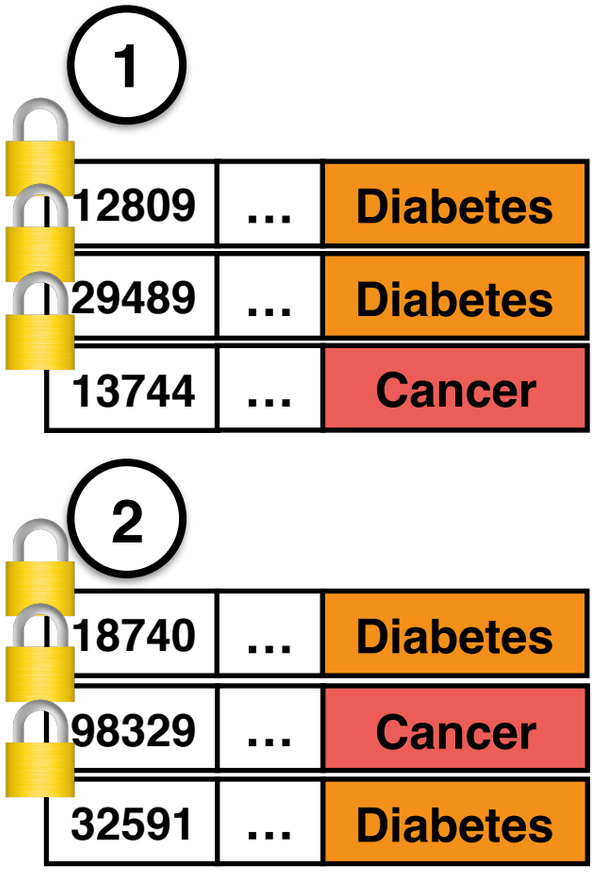


Opaque components



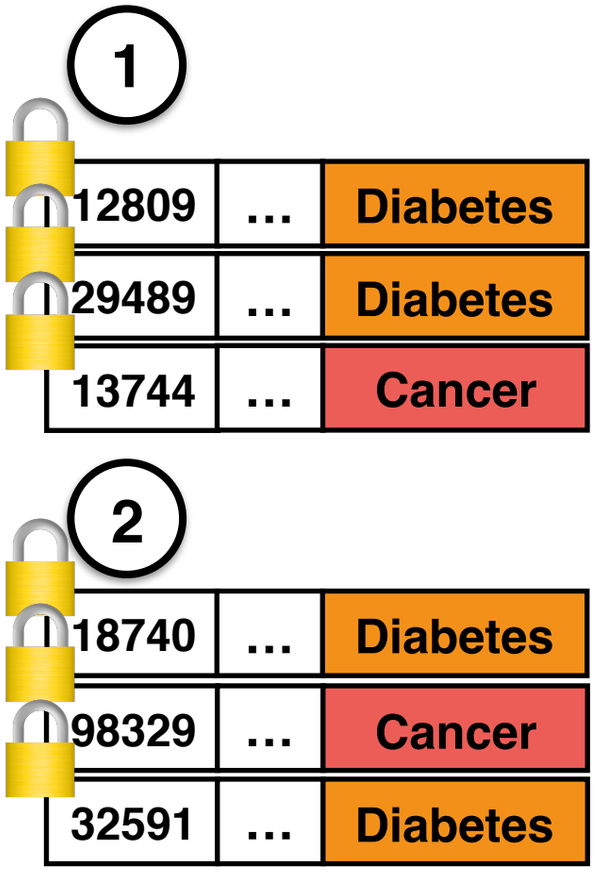
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



Oblivious aggregation

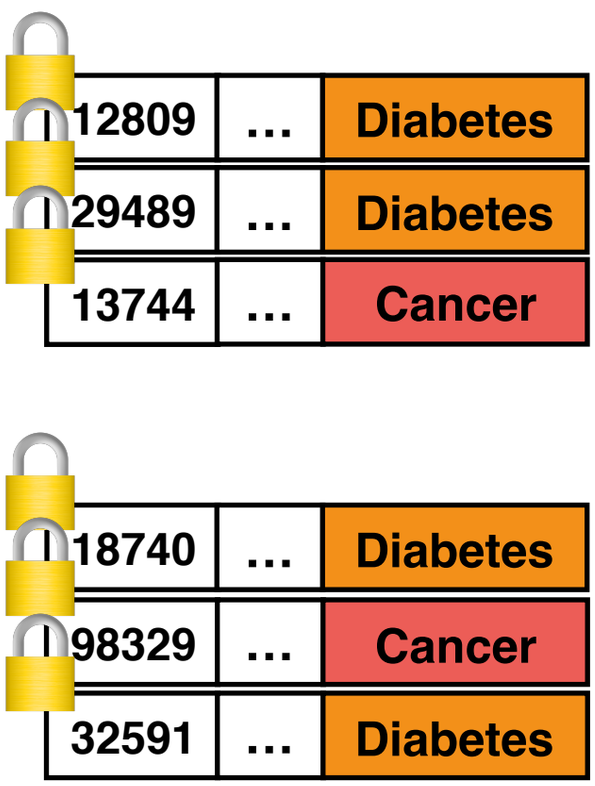
SELECT count(*) FROM medical GROUP BY disease

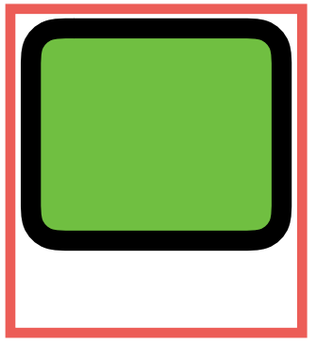


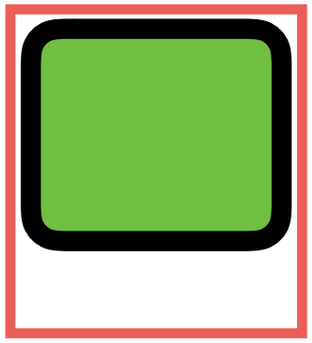
There can be many partitions

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



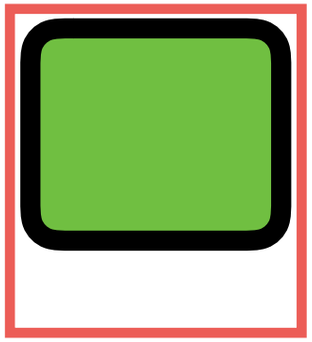


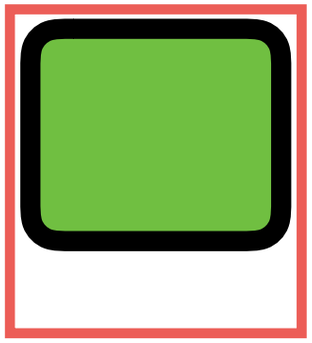


Map

Oblivious sort
[CLRS, Leighton '85]

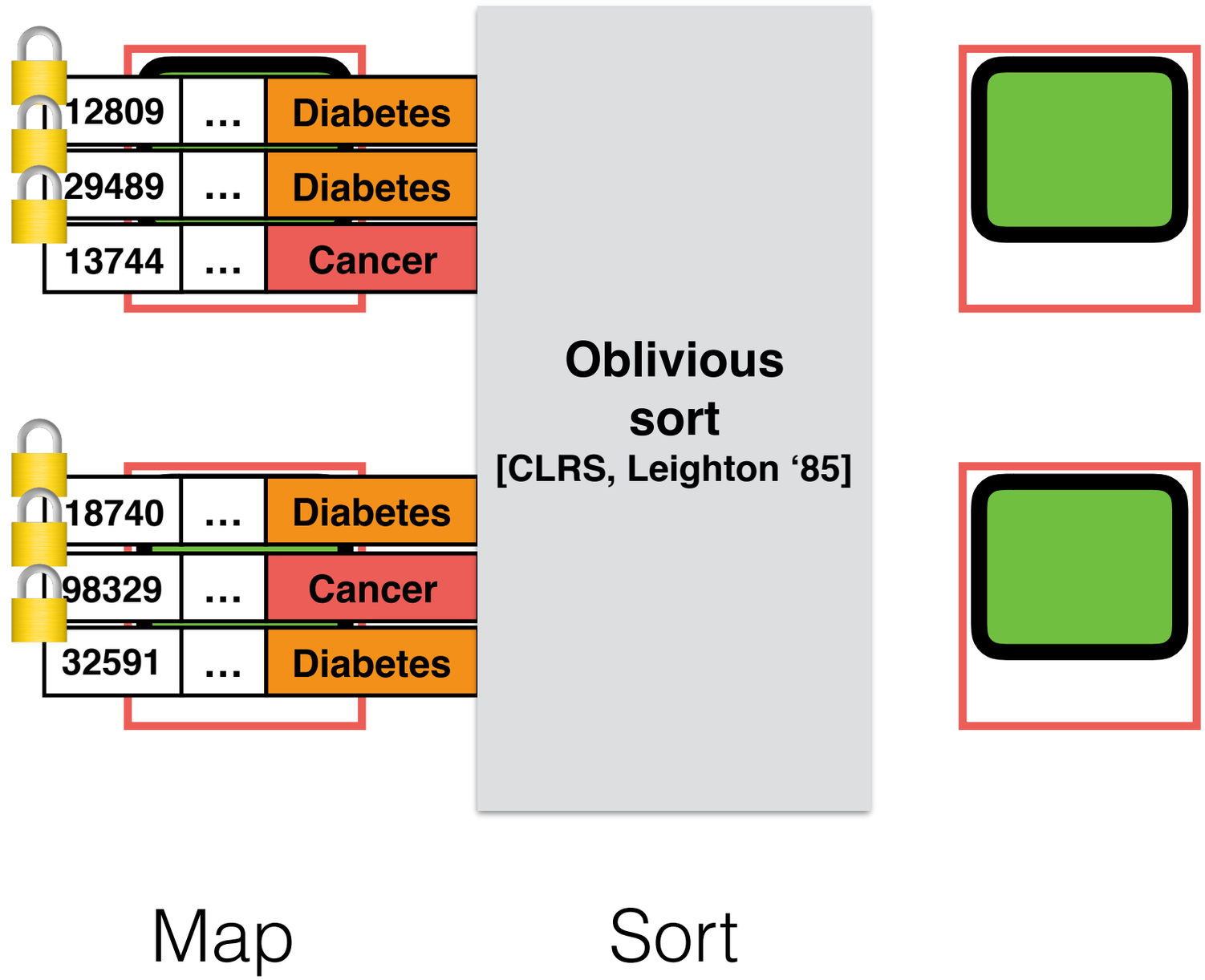
Sort





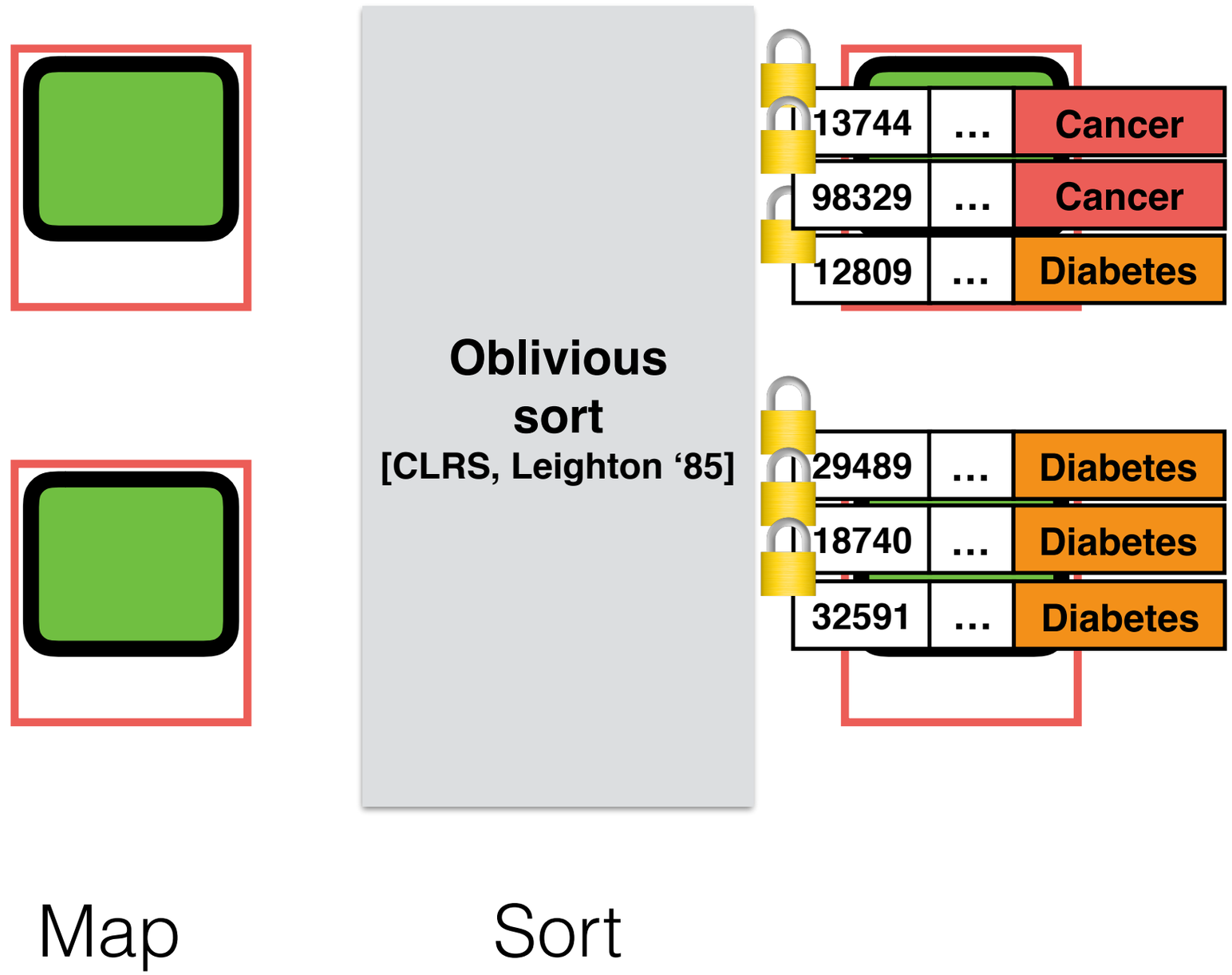
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



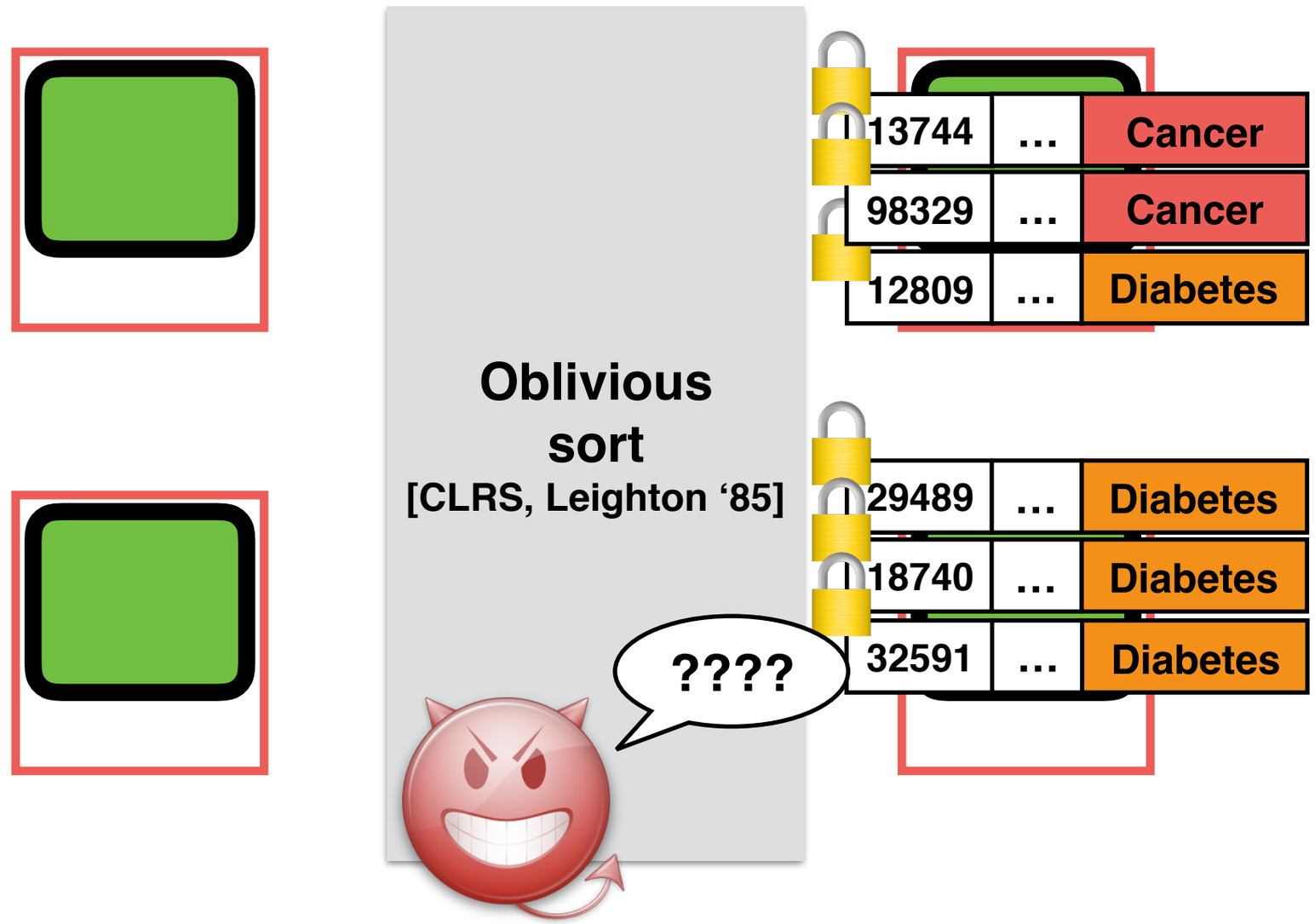
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



Oblivious aggregation

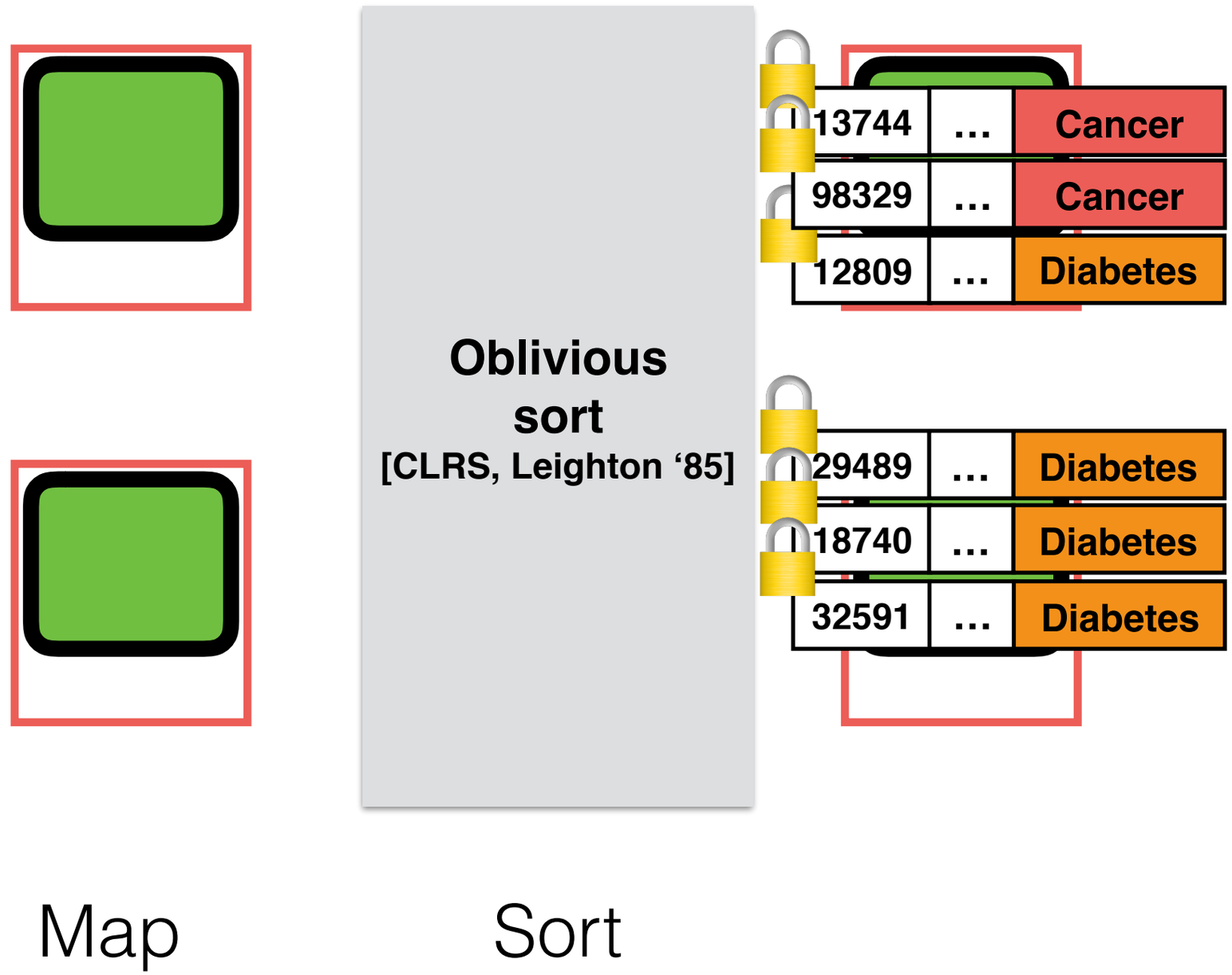
SELECT count(*) FROM medical GROUP BY disease



Map Sort

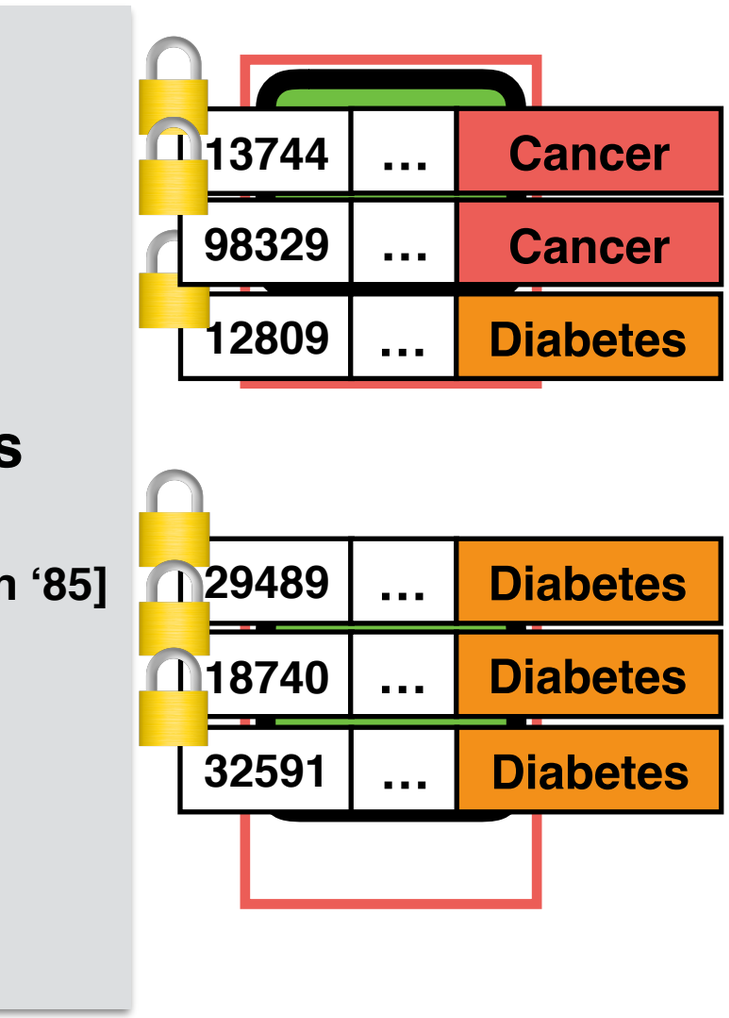
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



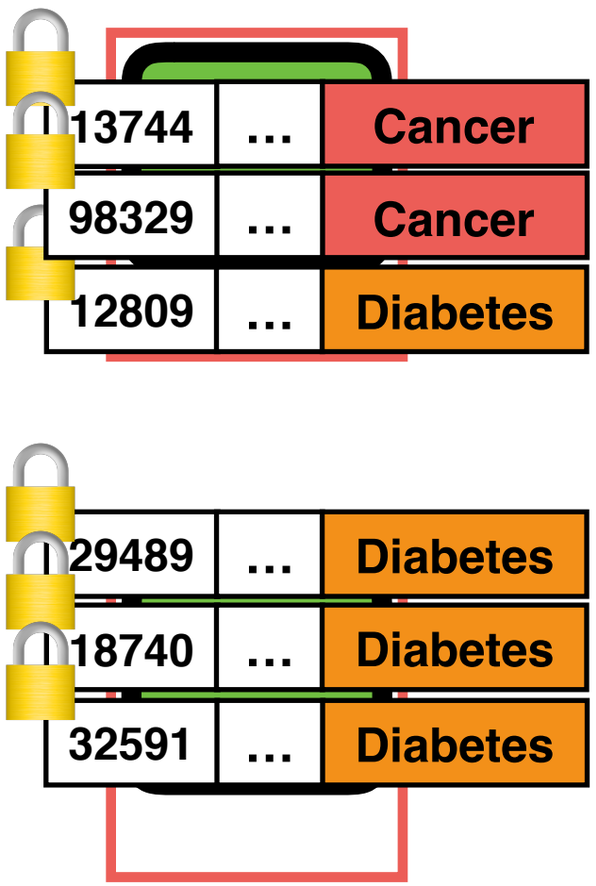
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



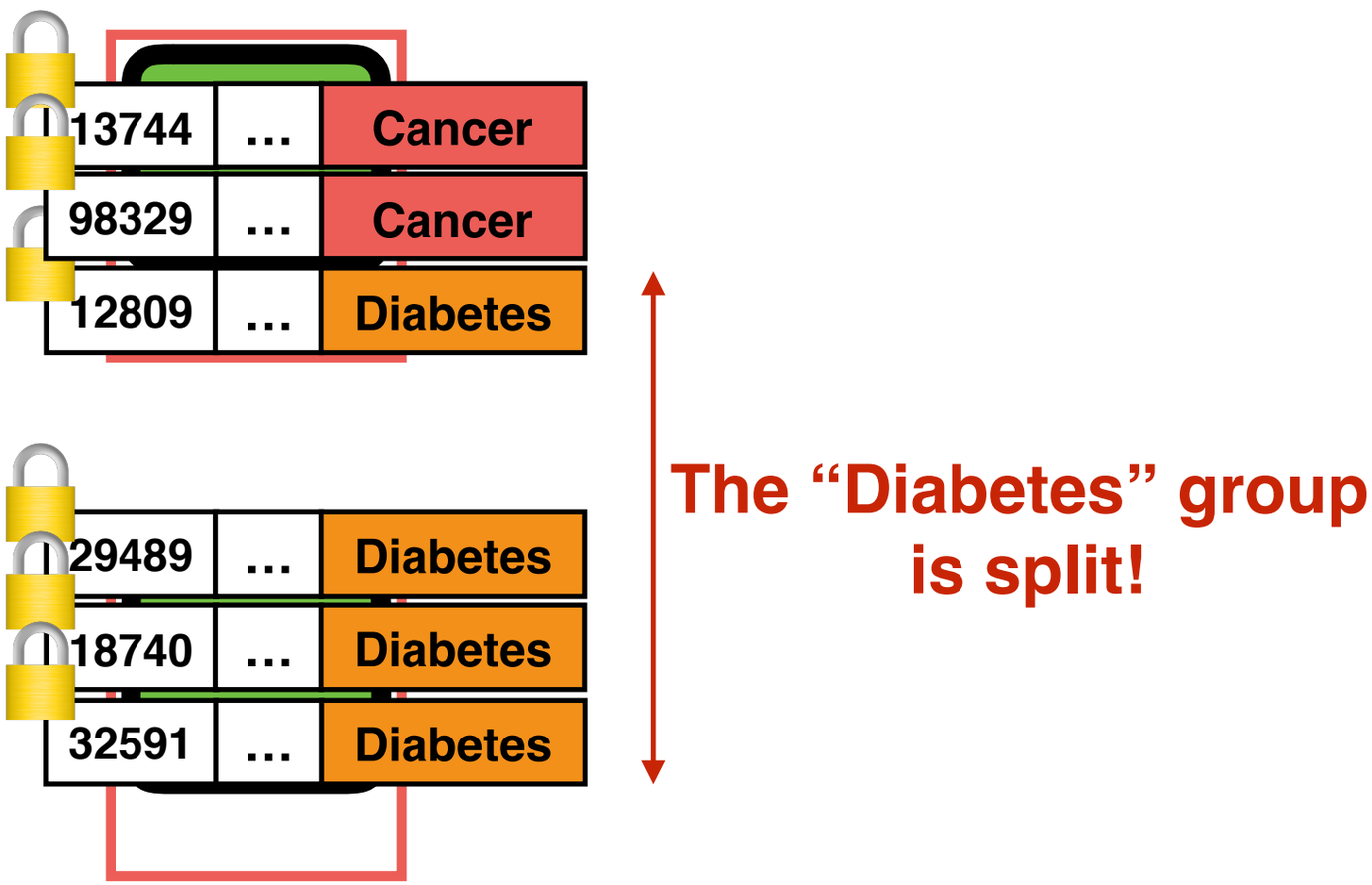
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



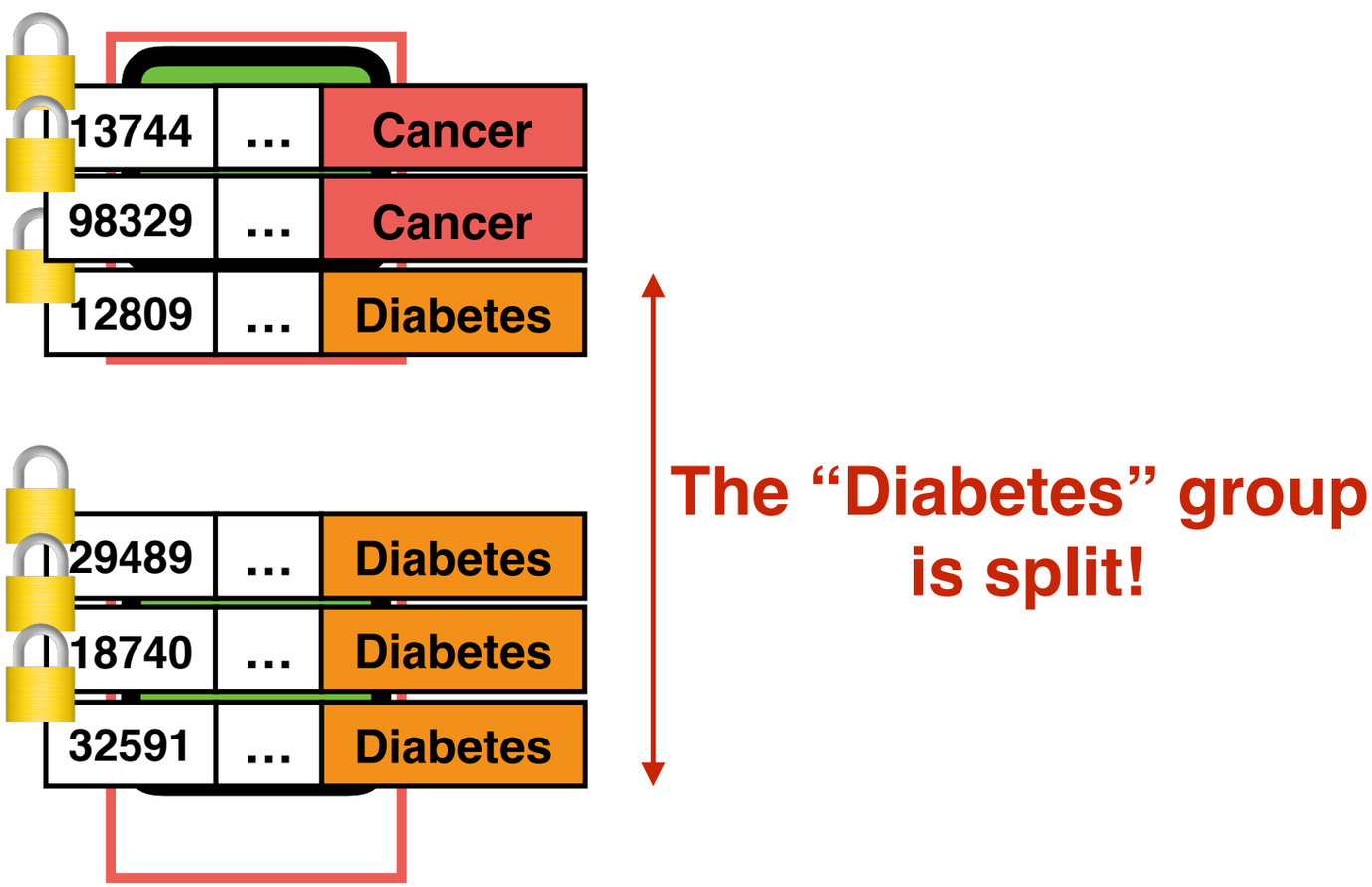
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



Oblivious aggregation

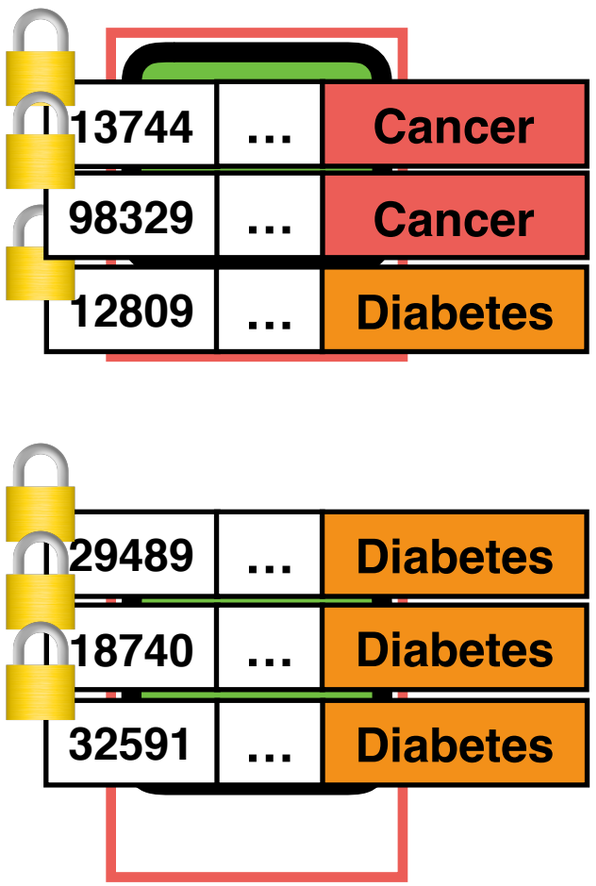
SELECT count(*) FROM medical GROUP BY disease



How to aggregate *obliviously* and in *parallel*?

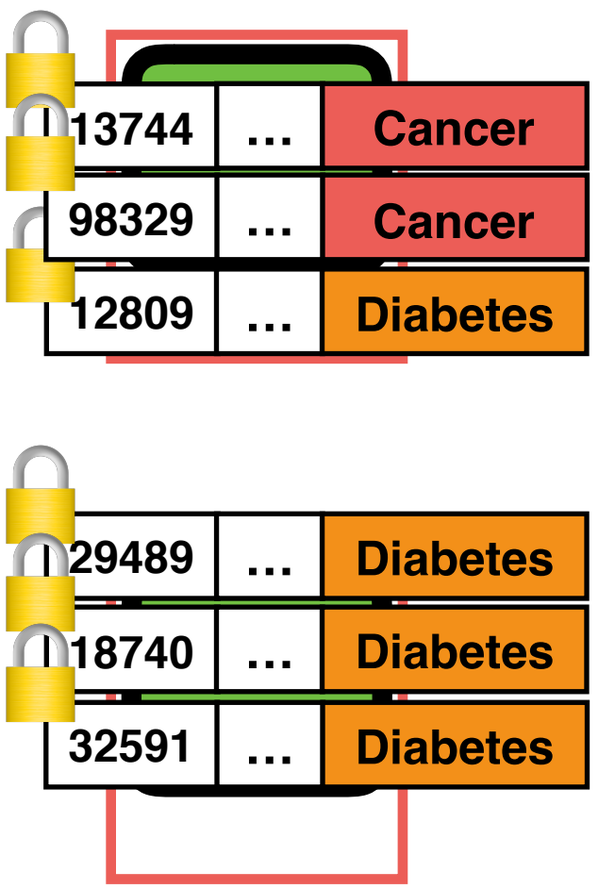
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



Oblivious aggregation

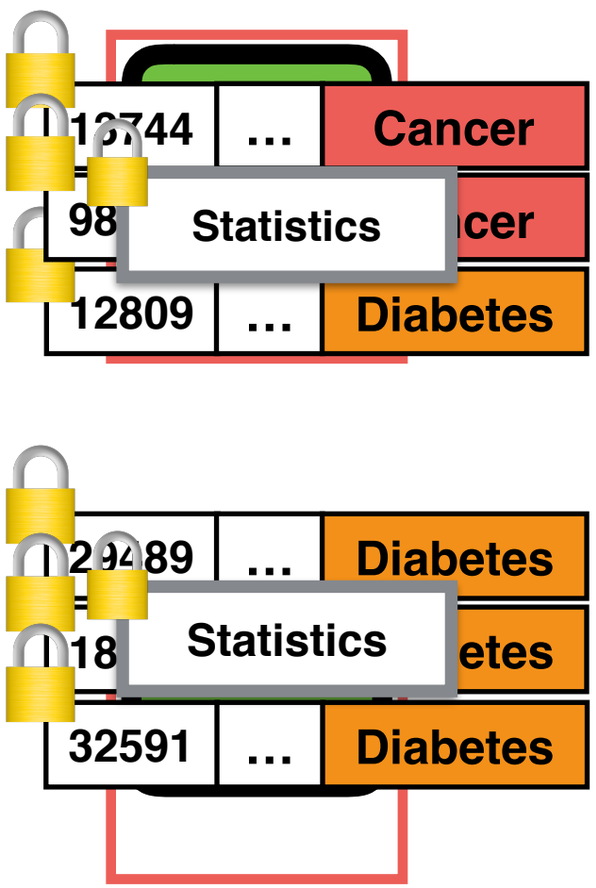
SELECT count(*) FROM medical GROUP BY disease



Scan

Oblivious aggregation

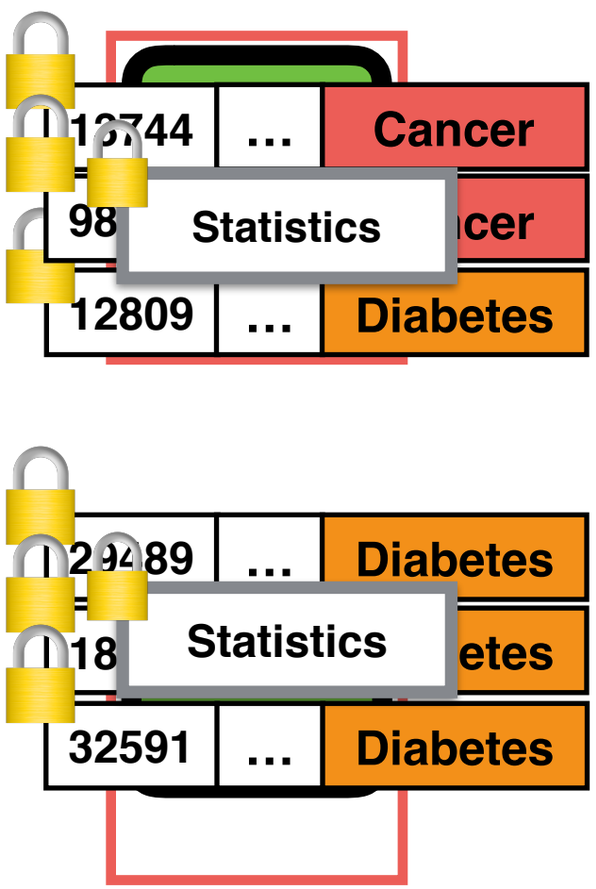
SELECT count(*) FROM medical GROUP BY disease

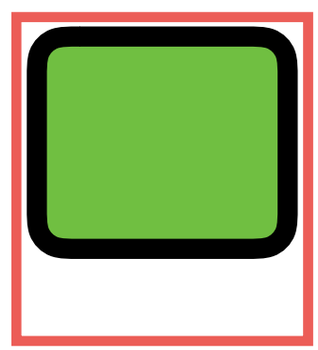


Scan

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



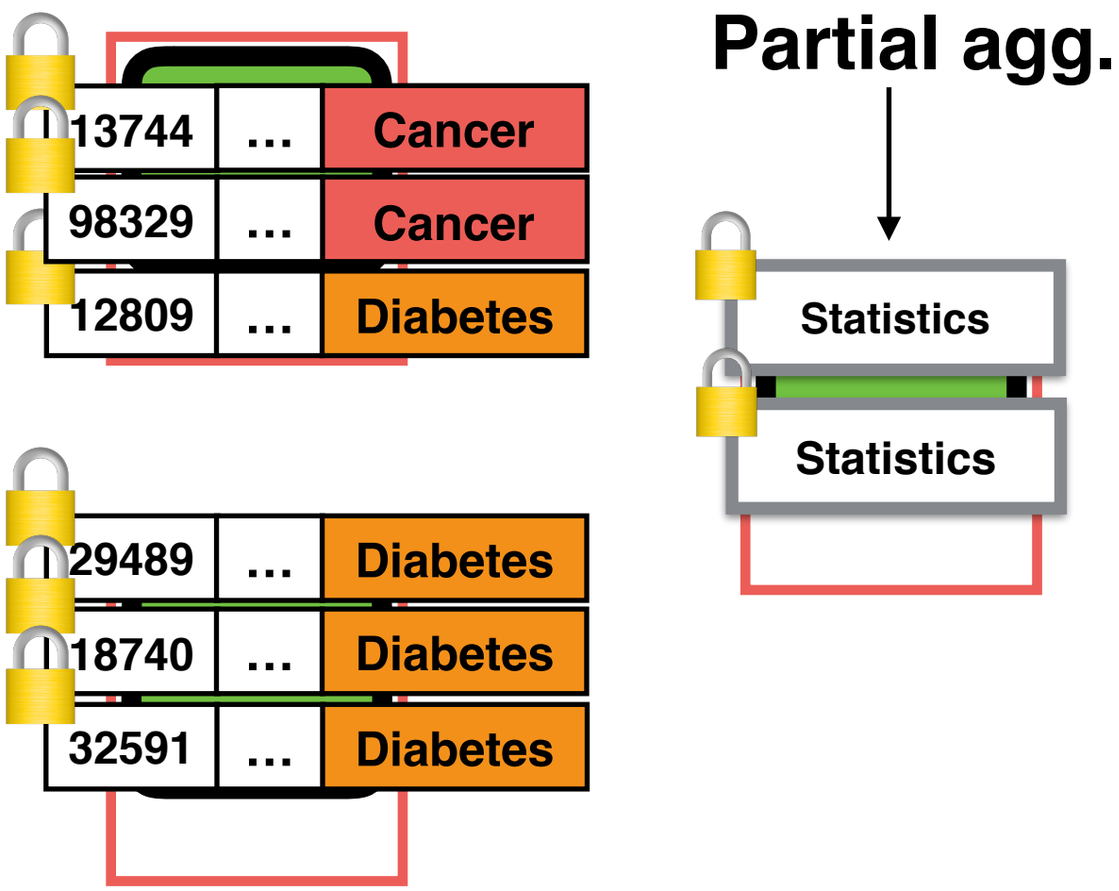


Scan

Boundary processing

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

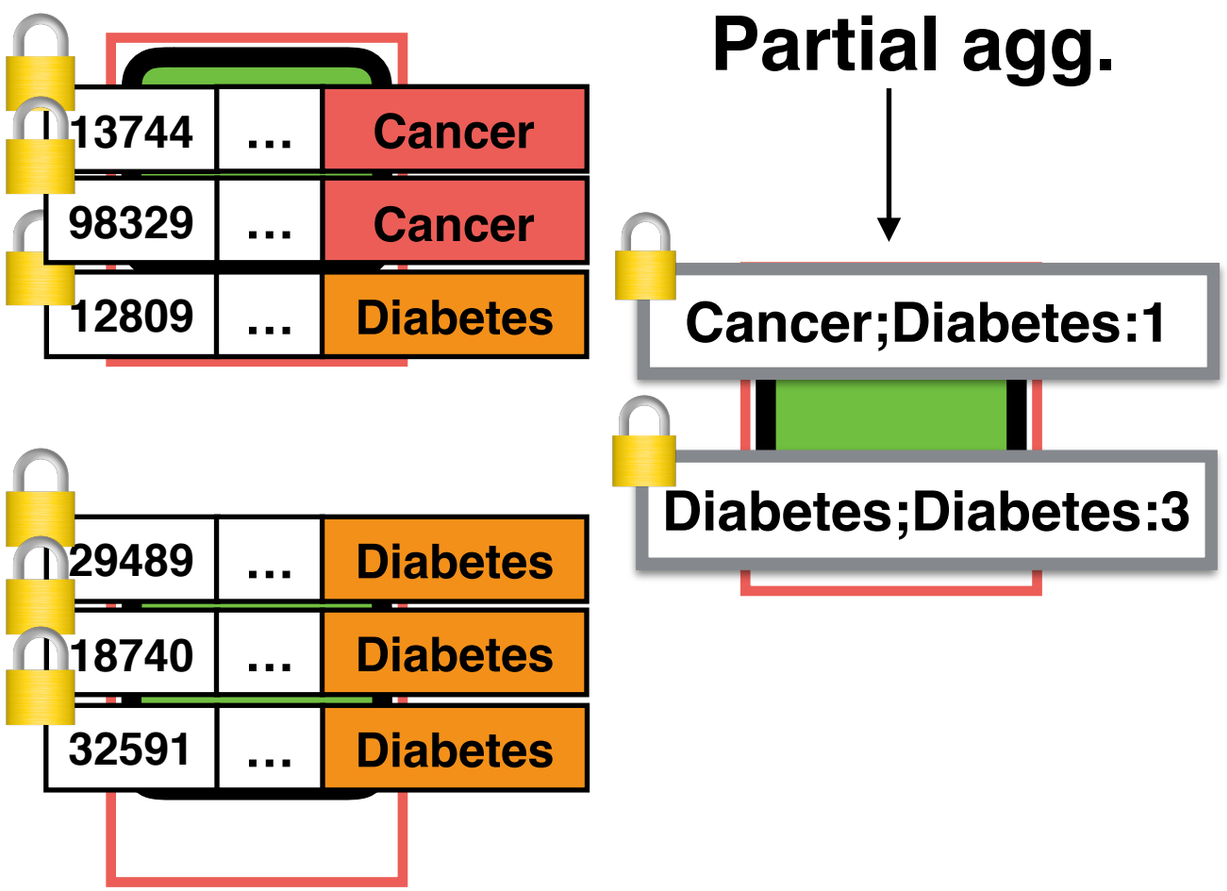


Scan

Boundary processing

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

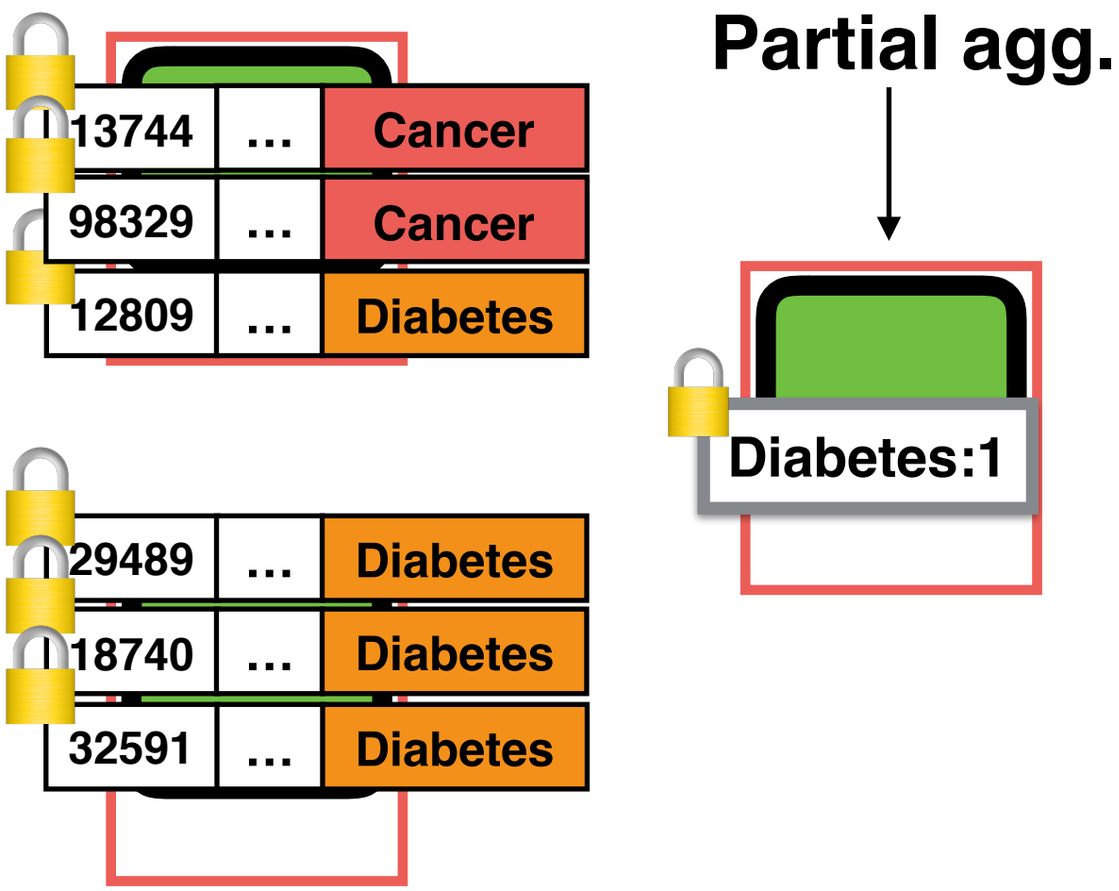


Scan

Boundary processing

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

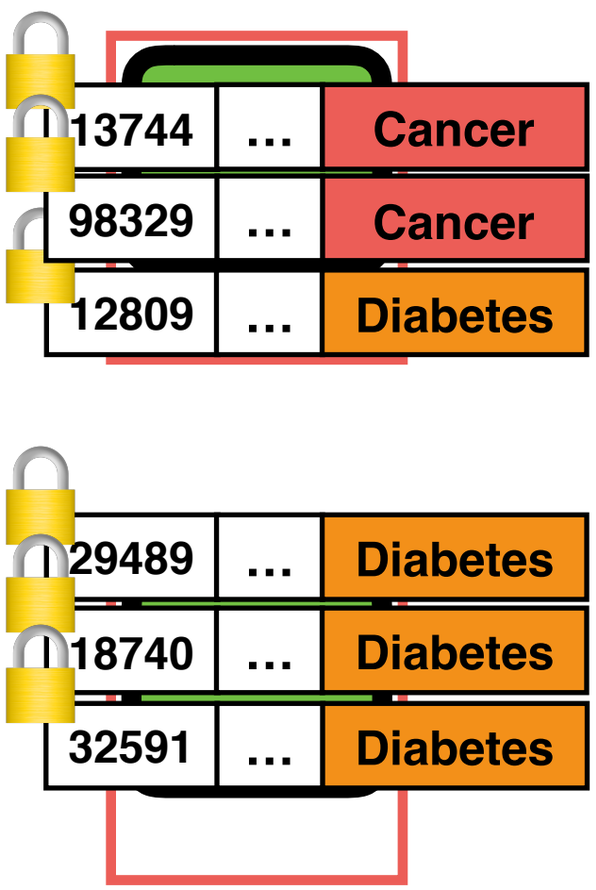


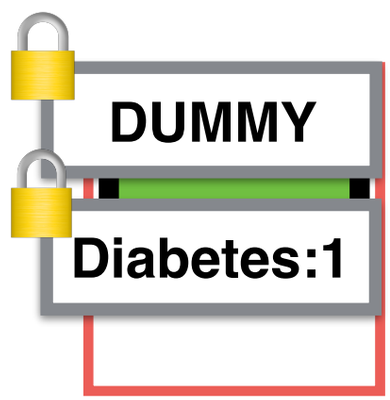
Scan

Boundary processing

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



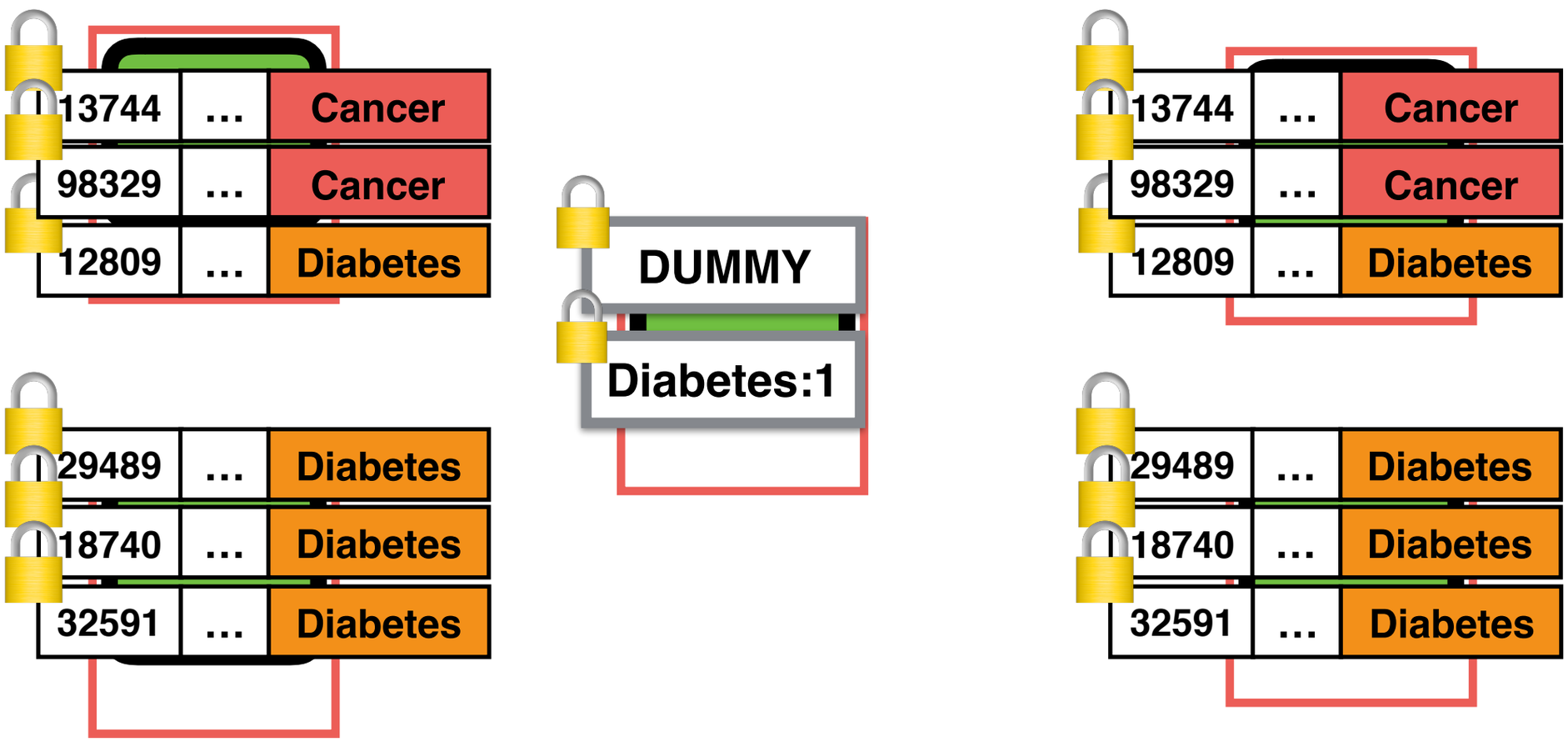


Scan

Boundary processing

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

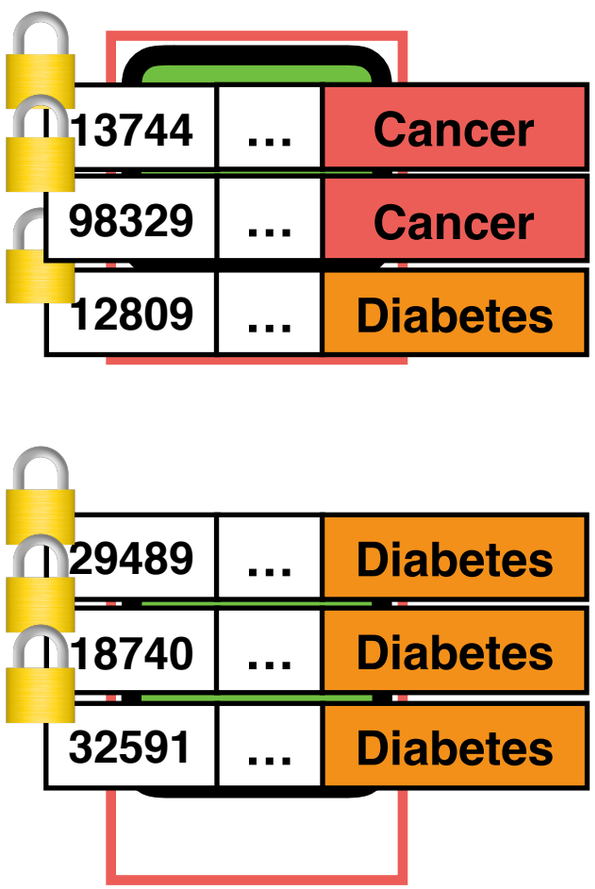


Scan

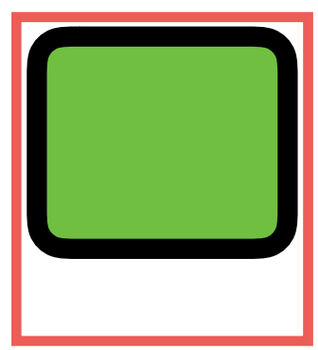
Boundary processing

Oblivious aggregation

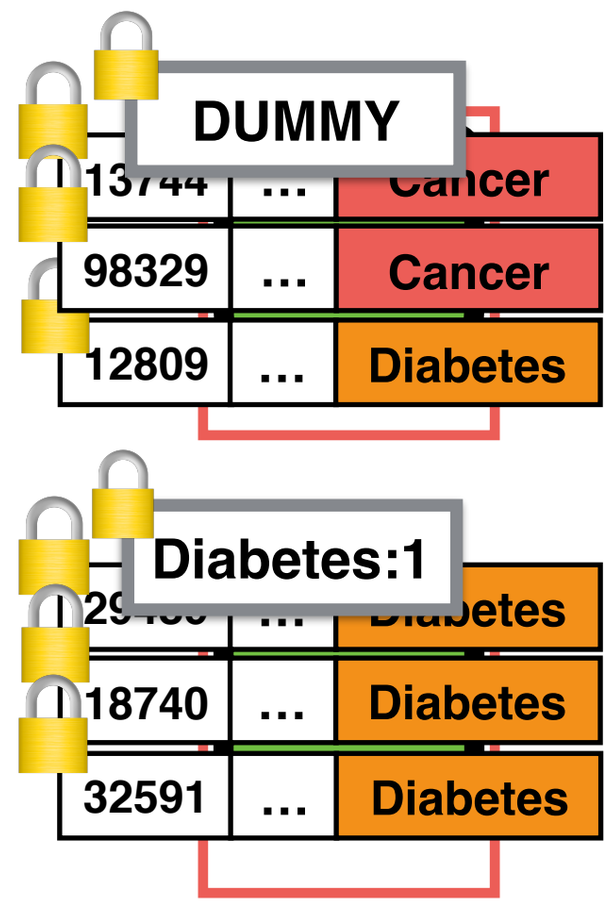
SELECT count(*) FROM medical GROUP BY disease



Scan

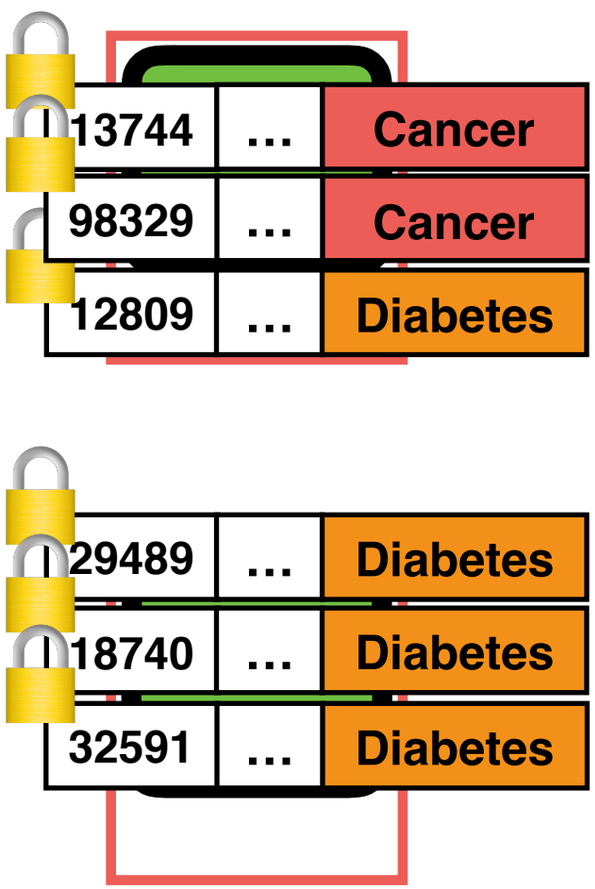


Boundary processing

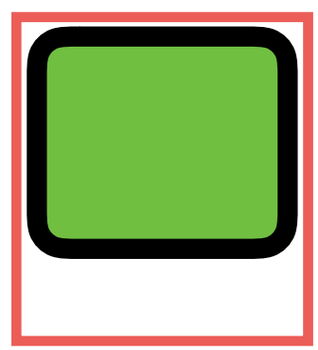


Oblivious aggregation

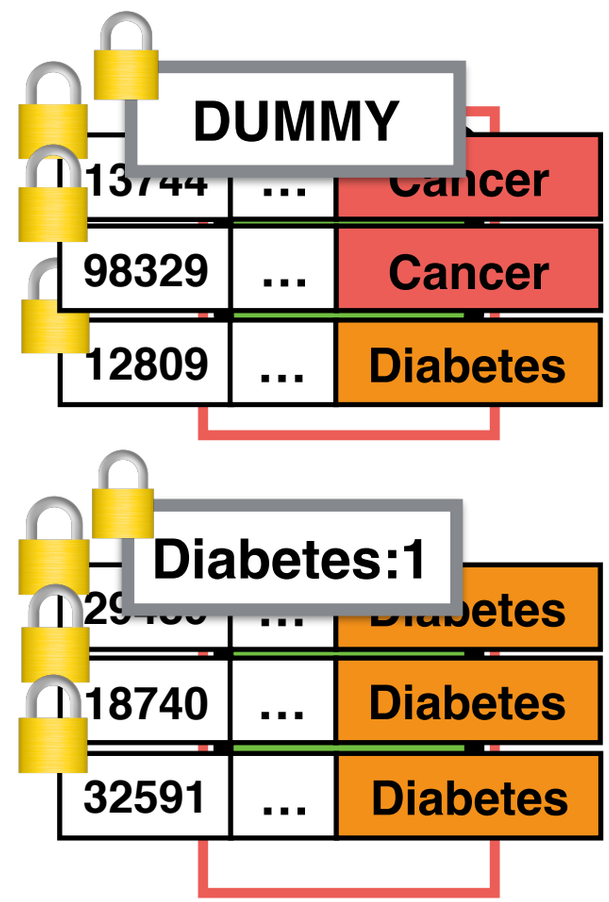
SELECT count(*) FROM medical GROUP BY disease



Scan



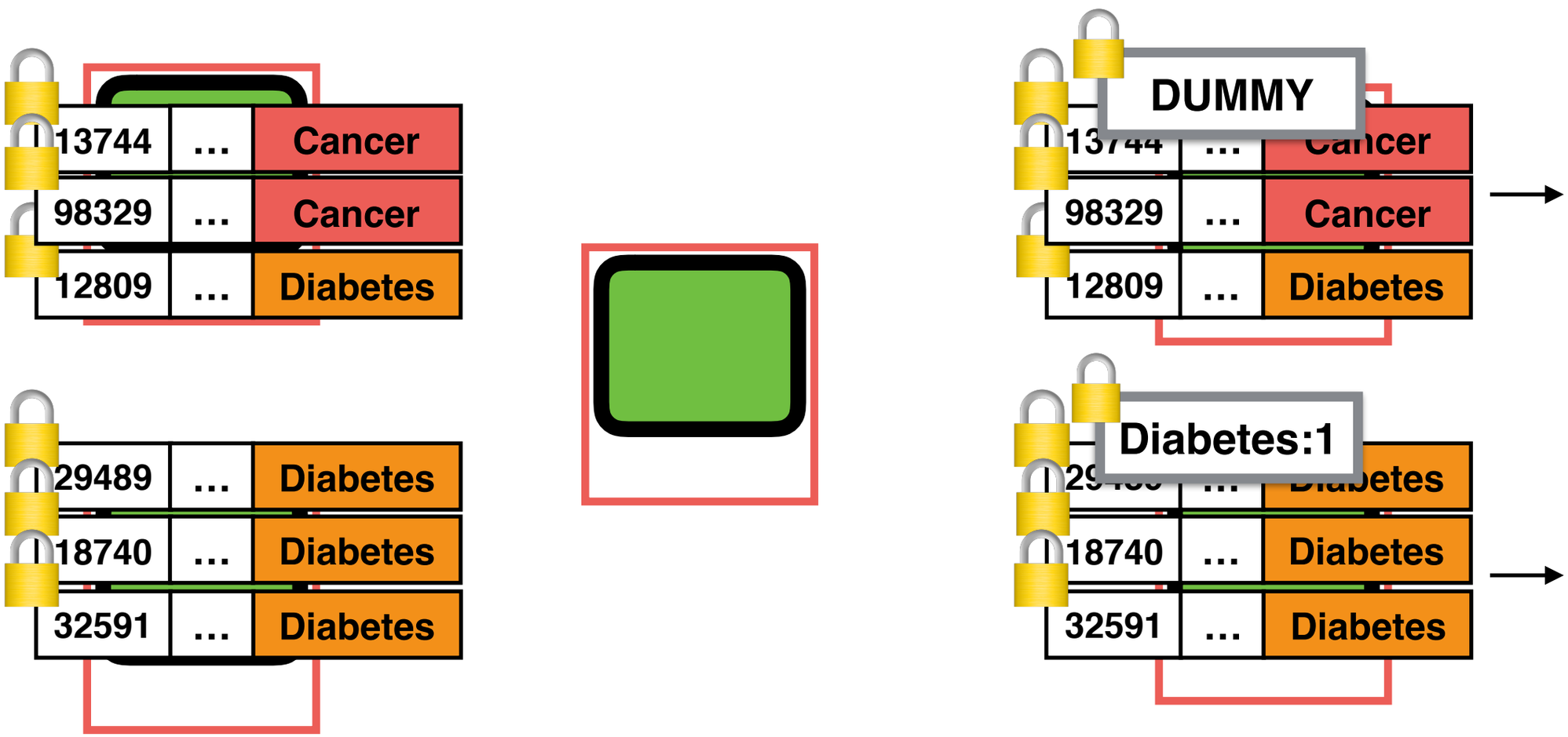
Boundary processing



Scan

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



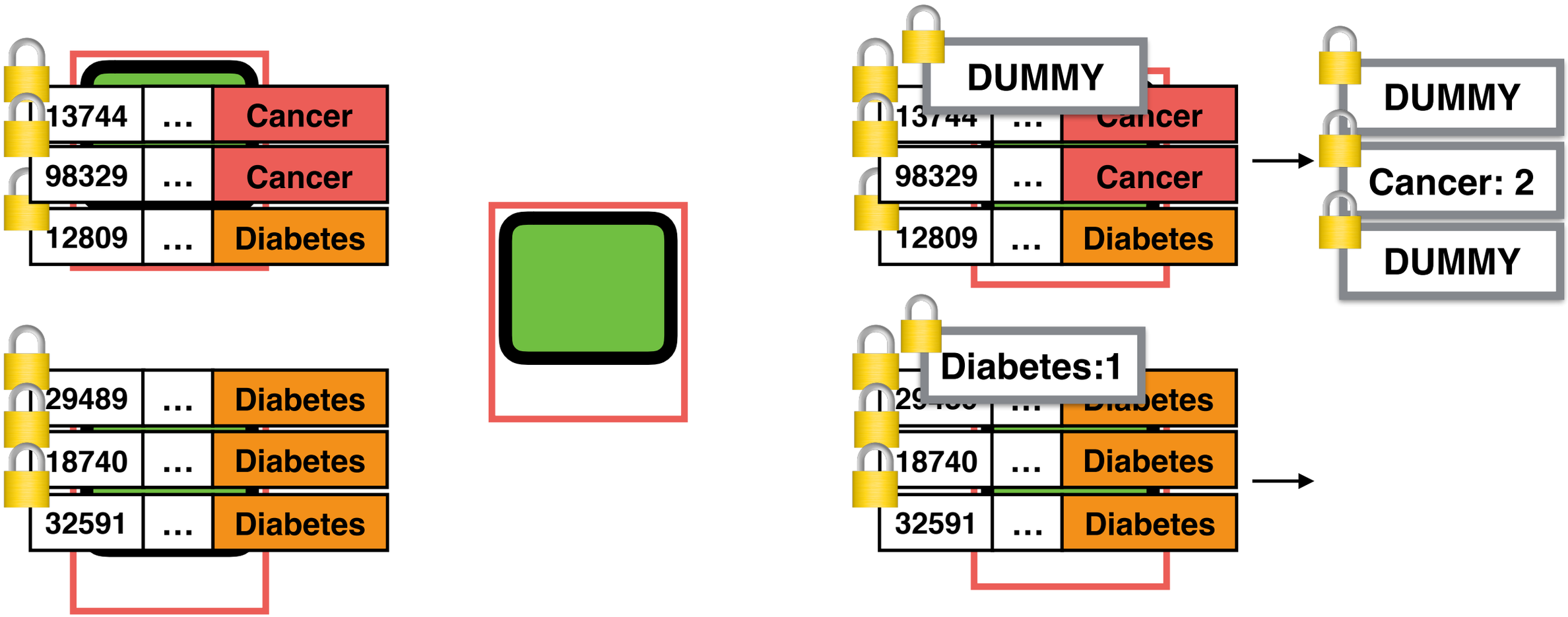
Scan

Boundary processing

Scan

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



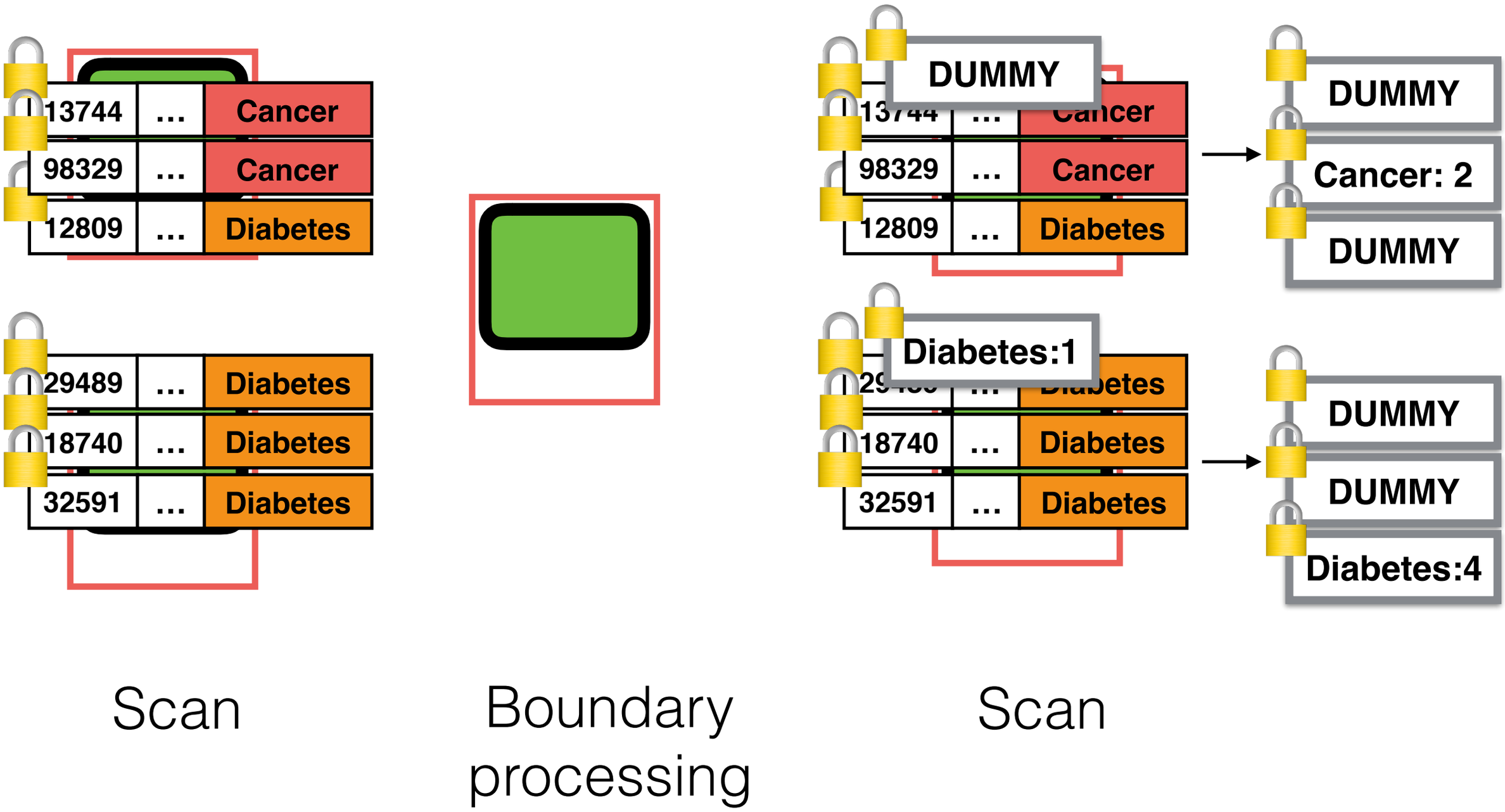
Scan

Boundary processing

Scan

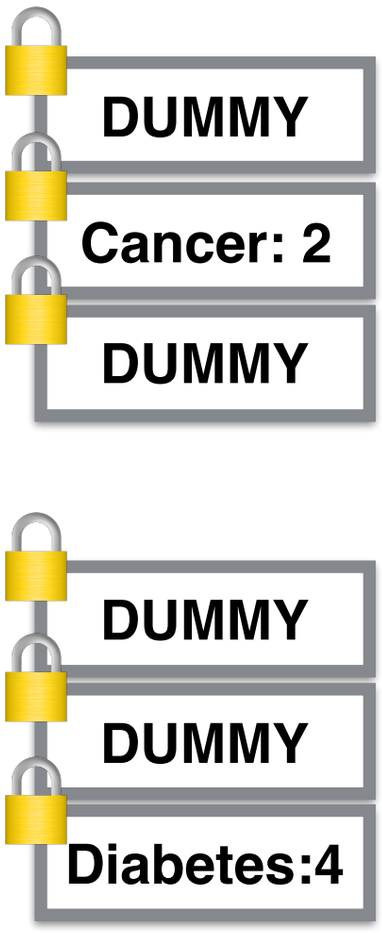
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



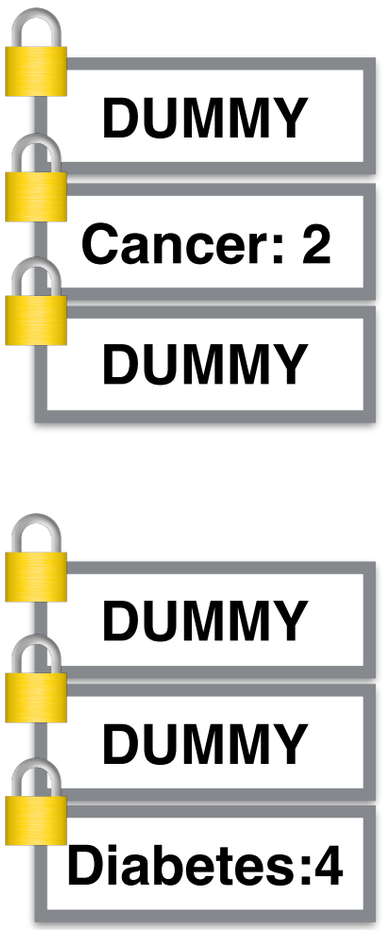
Oblivious aggregation

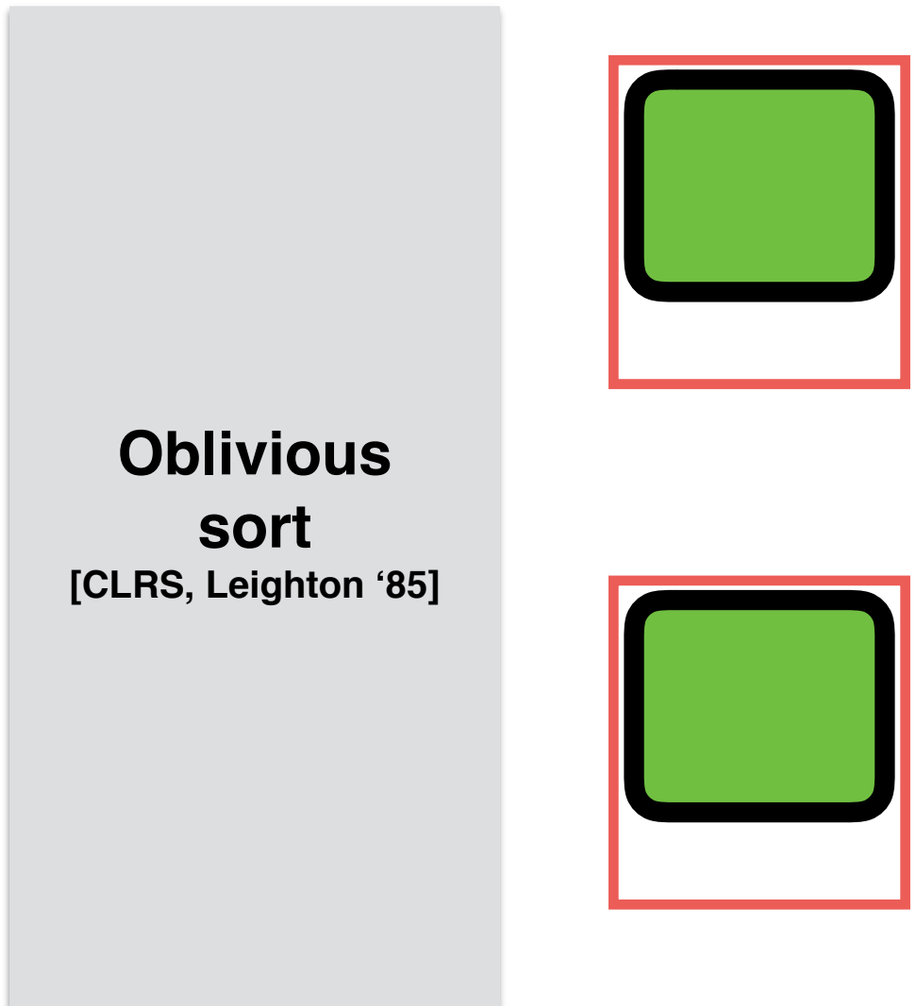
SELECT count(*) FROM medical GROUP BY disease



Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

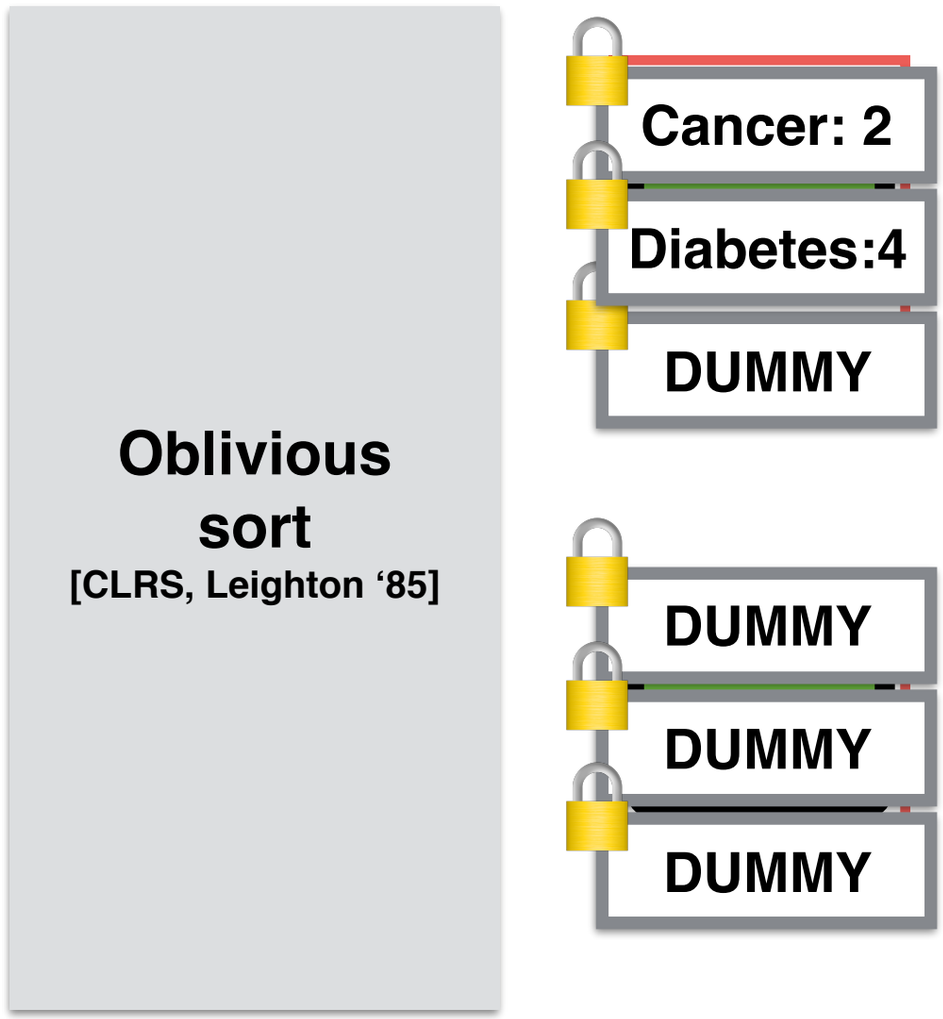




Sort

Oblivious aggregation

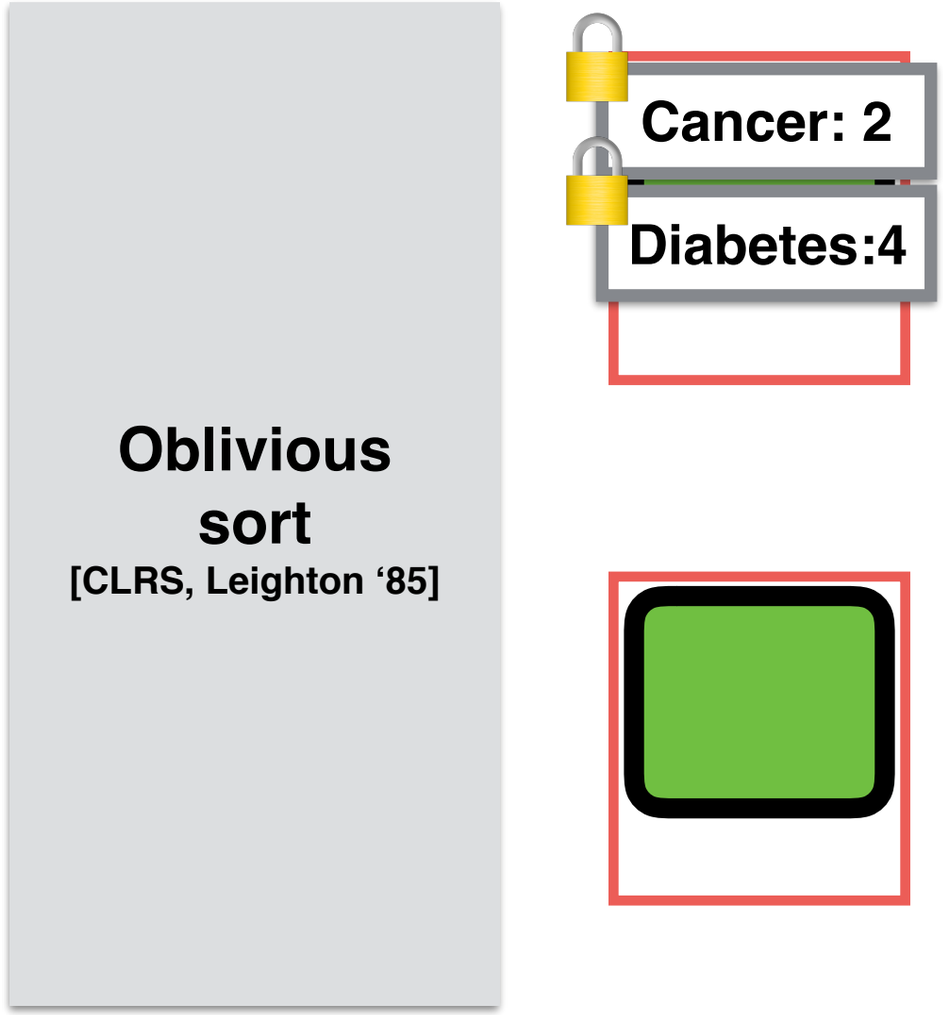
SELECT count(*) FROM medical GROUP BY disease



Sort

Oblivious aggregation

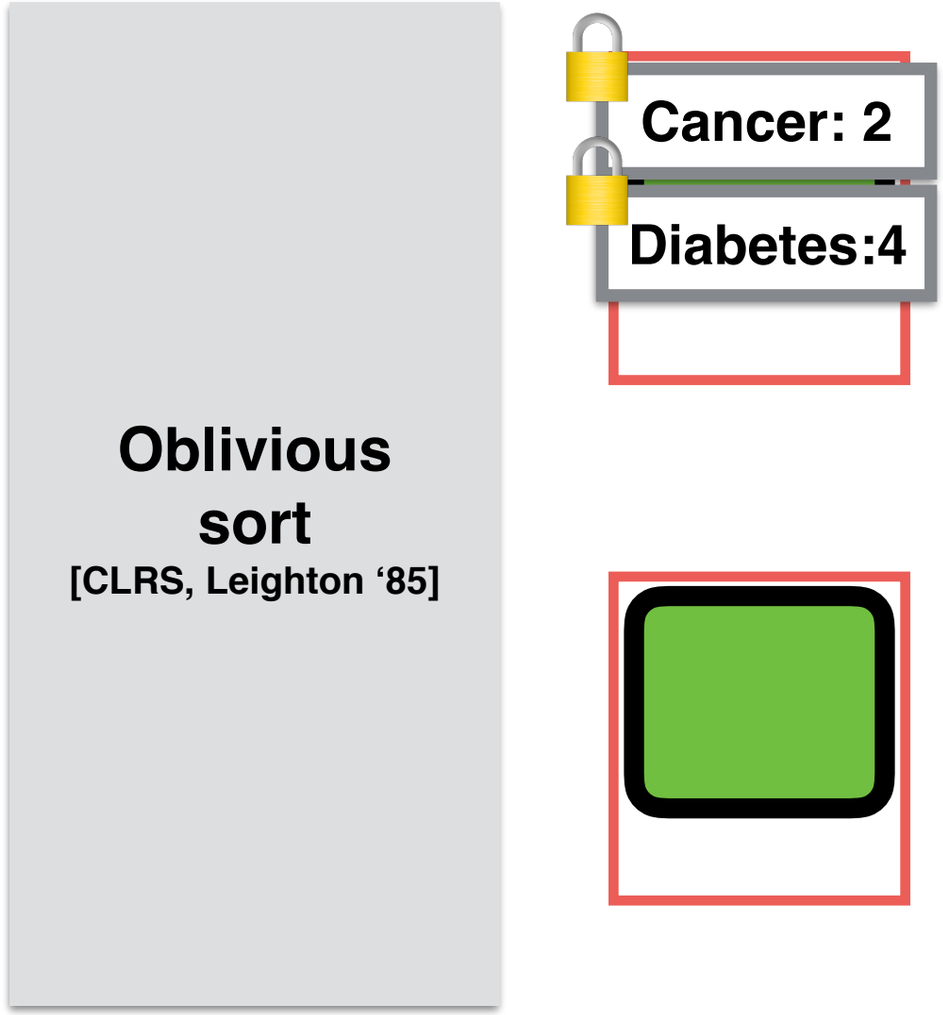
SELECT count(*) FROM medical GROUP BY disease



Sort

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

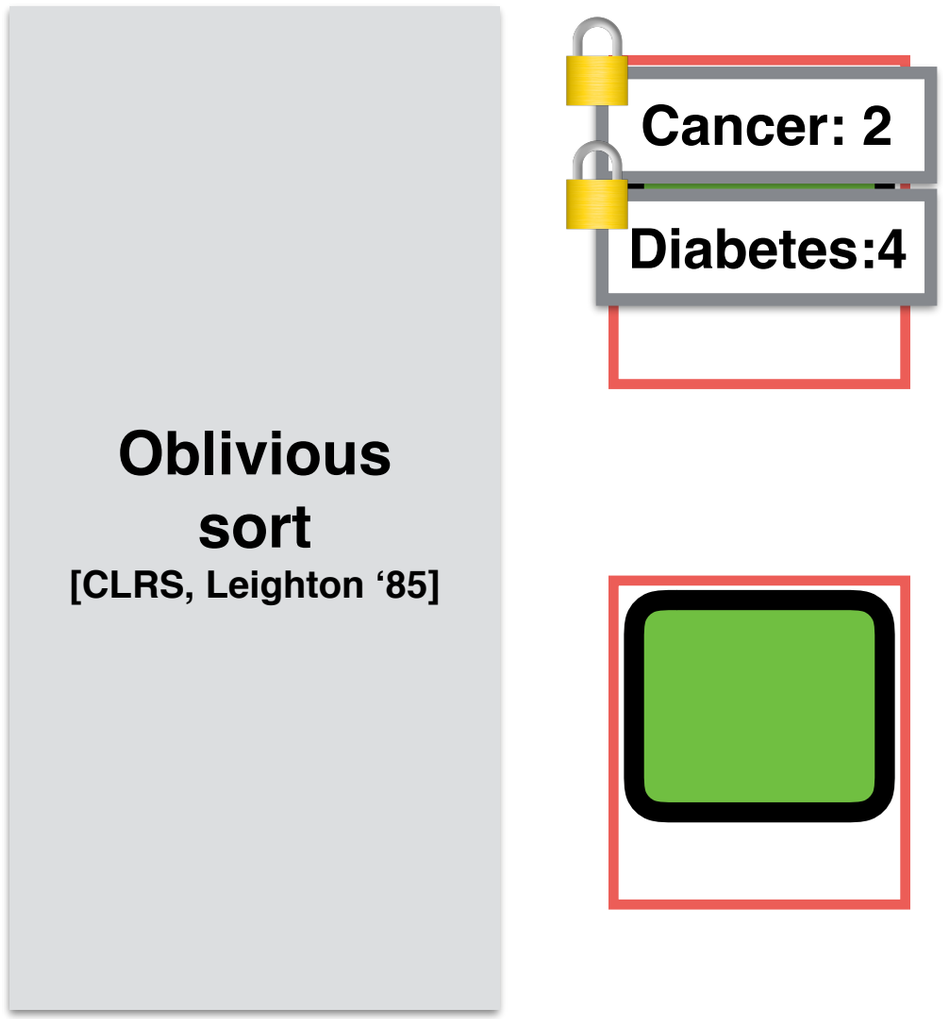


Sort

Final result

Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease



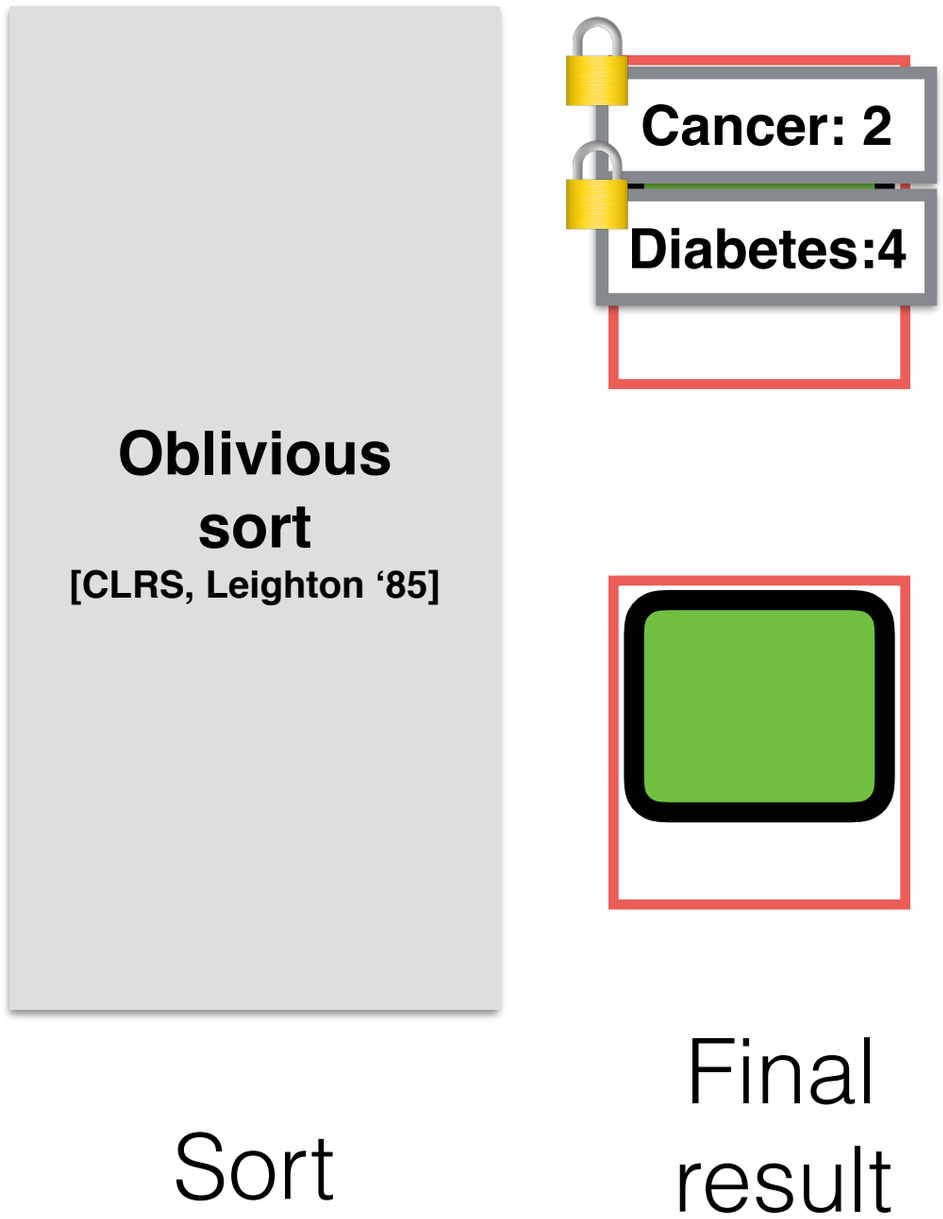
Sort

Final result

Aggregation has <u>two</u> sorts…

Oblivious aggregation

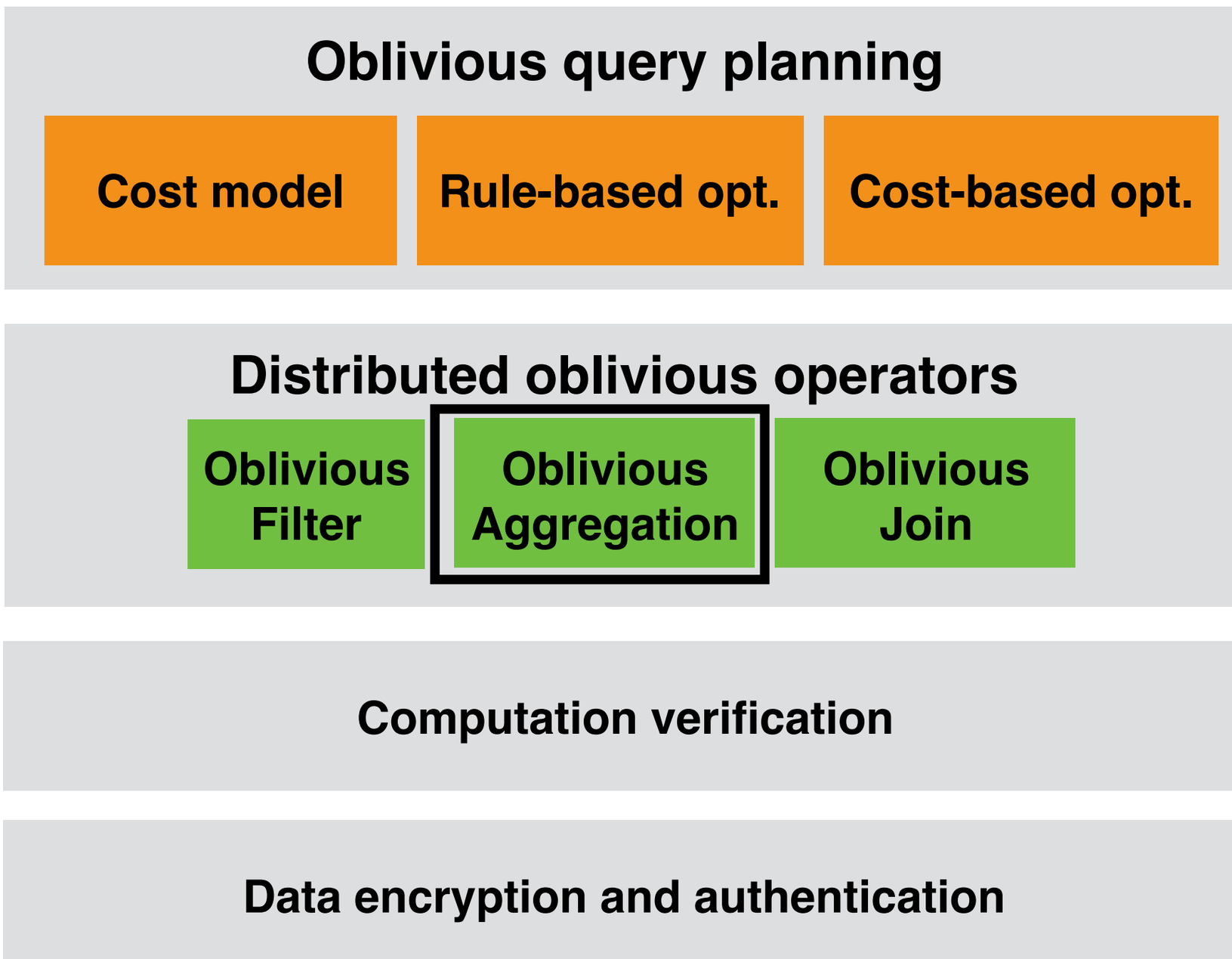
SELECT count(*) FROM medical GROUP BY disease



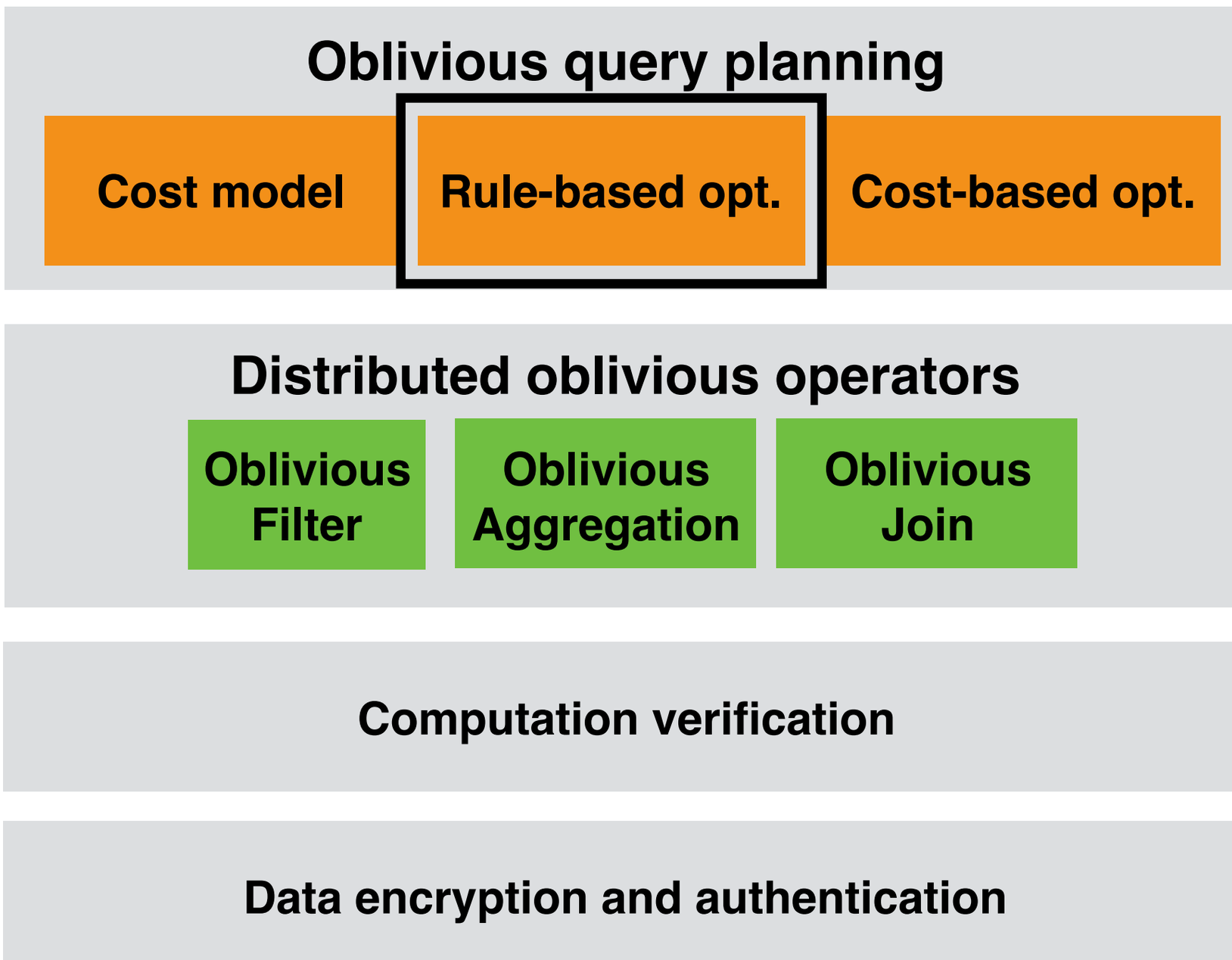
Aggregation has <u>two</u> sorts…

Can we do better?

Opaque components



Opaque components



Rule-based optimization

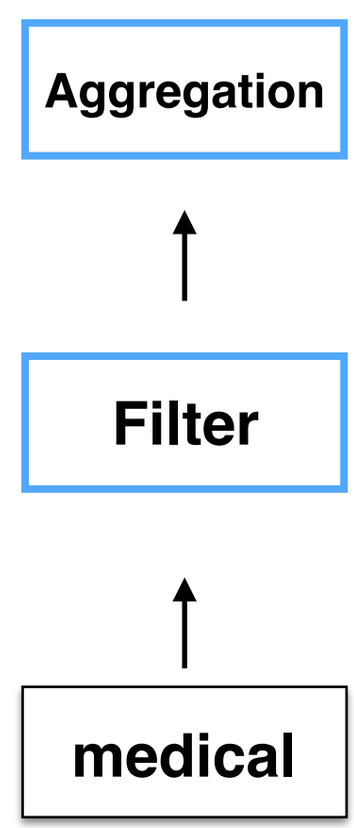
Rule-based optimization

```
SELECT count(*)
FROM medical
WHERE age > 30
GROUP BY disease
```

Rule-based optimization

Logical op.

```
SELECT count(*)
FROM medical
WHERE age > 30
GROUP BY disease
```



Insight 1

Insight 1

1. Split each logical operator into smaller Opaque operators

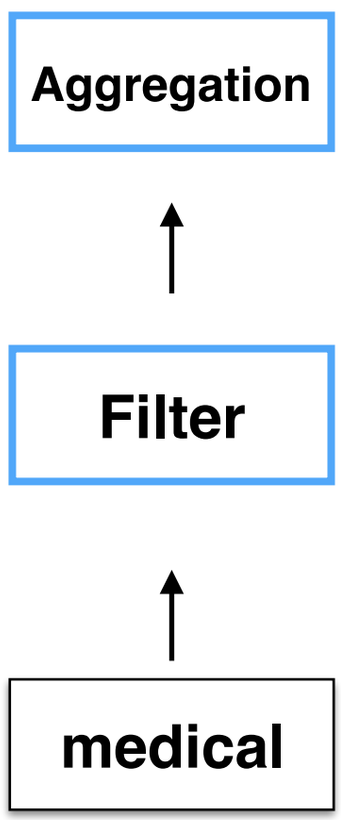
Insight 1

1. Split each logical operator into smaller Opaque operators

2. Take a global view across the plan to remove some Opaque operators

Rule-based optimization

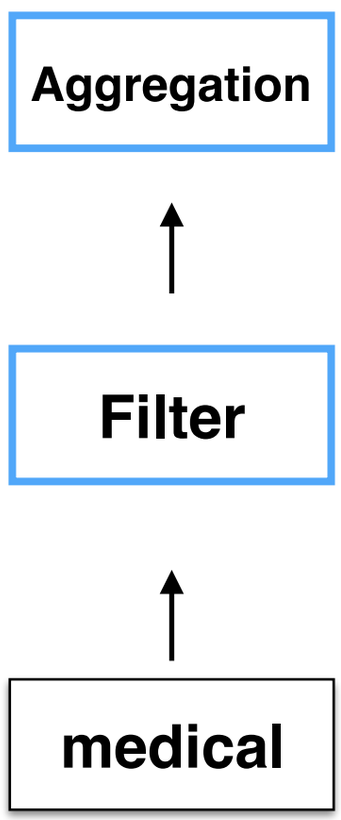
Logical op.



Rule-based optimization

Opaque op.

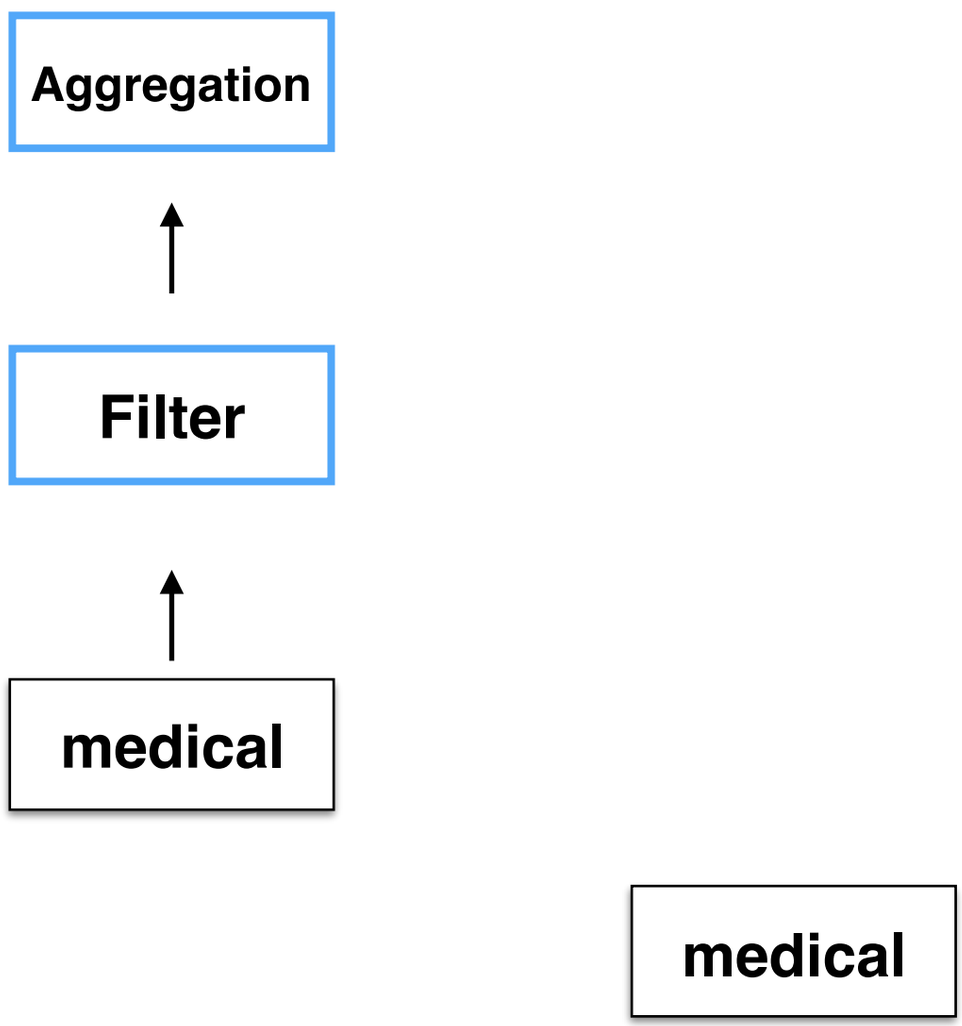
Logical op.



Rule-based optimization

Opaque op.

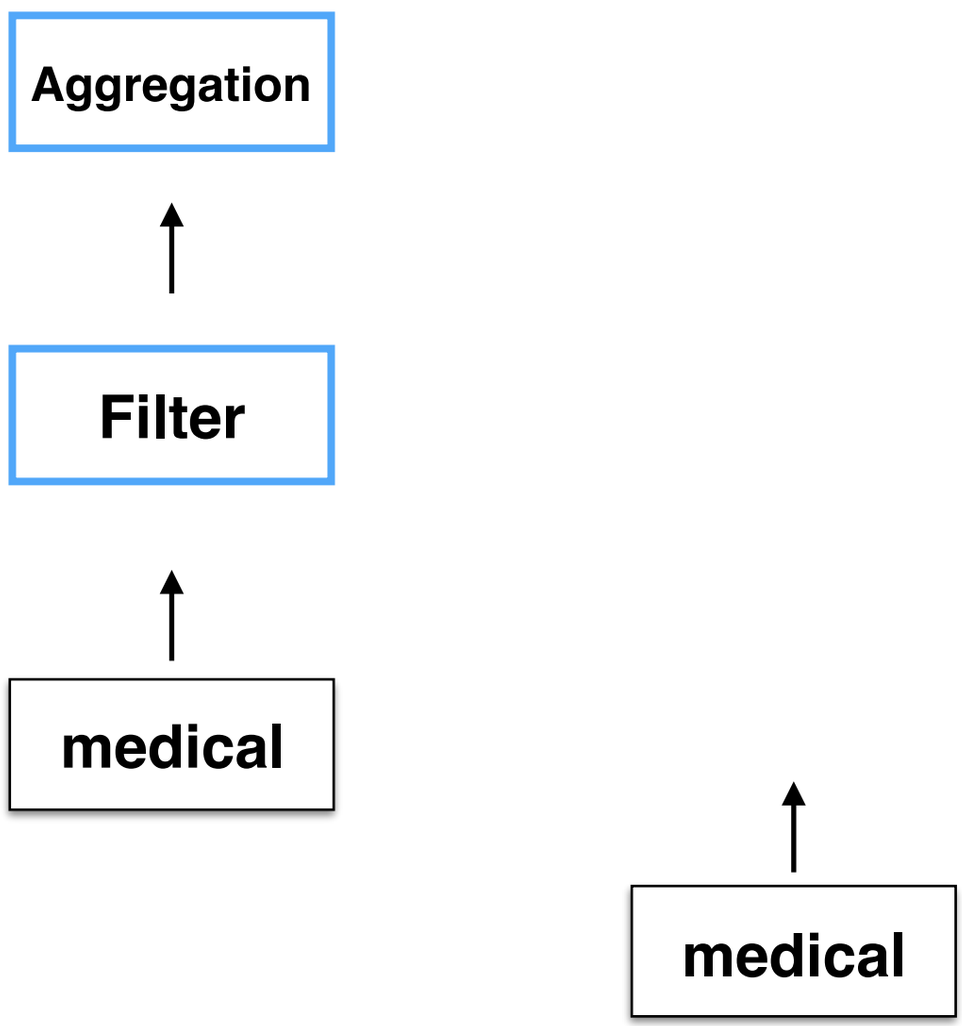
Logical op.



Rule-based optimization

Opaque op.

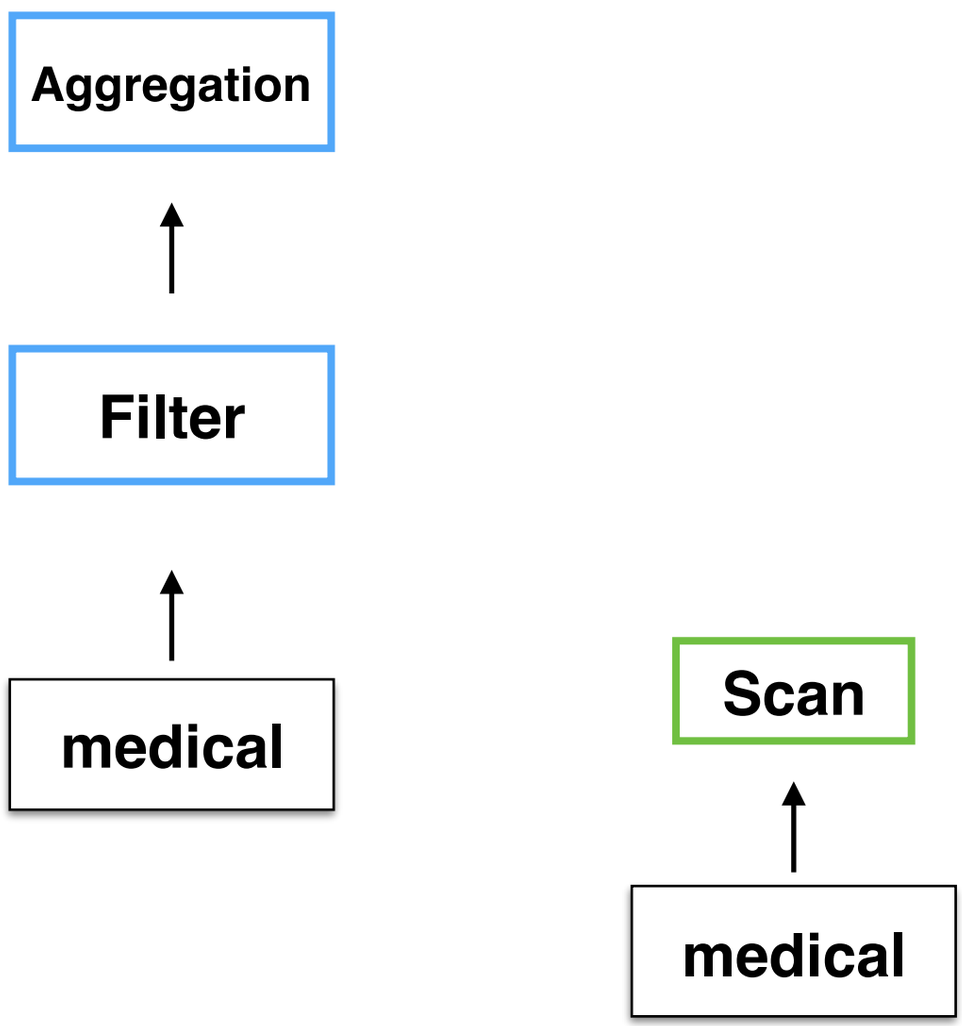
Logical op.



Rule-based optimization

Opaque op.

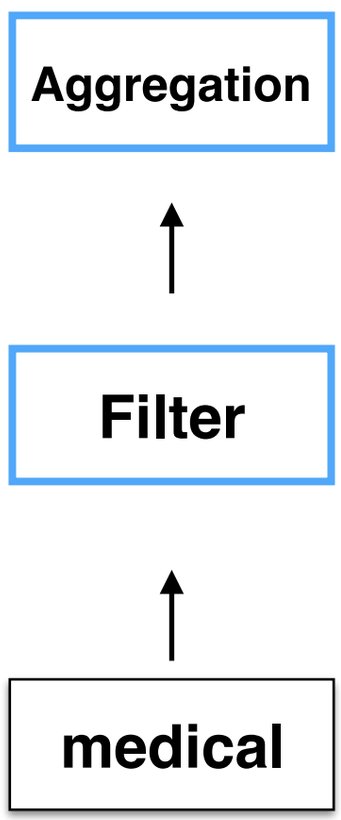
Logical op.

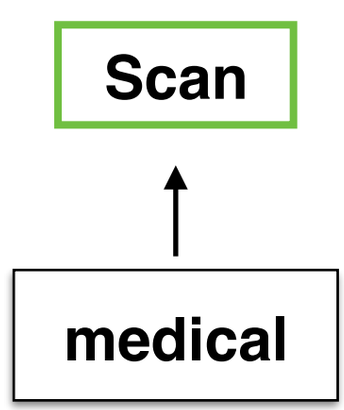


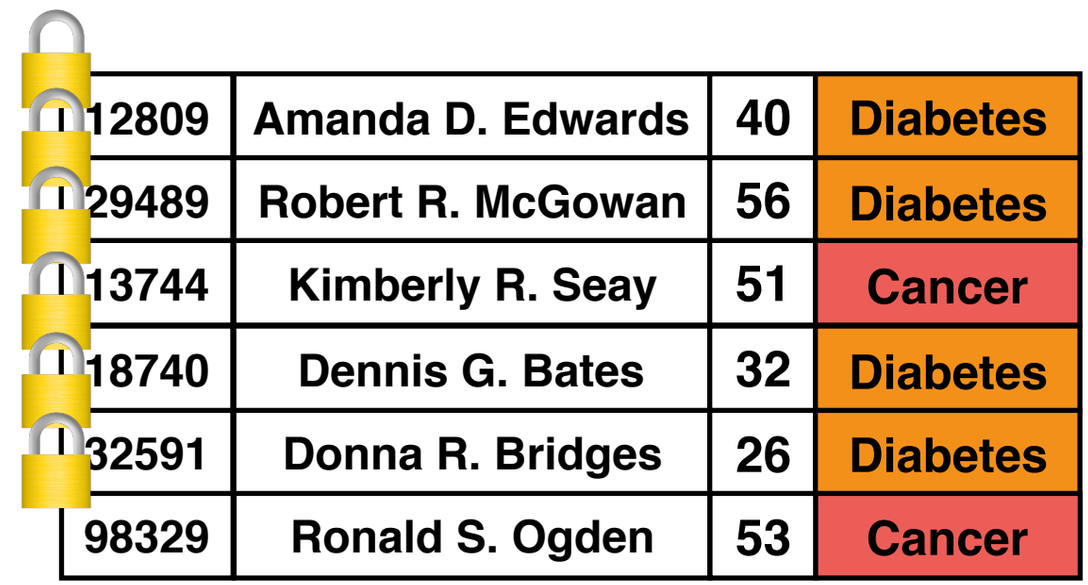
Rule-based optimization

Opaque op.

Logical op.





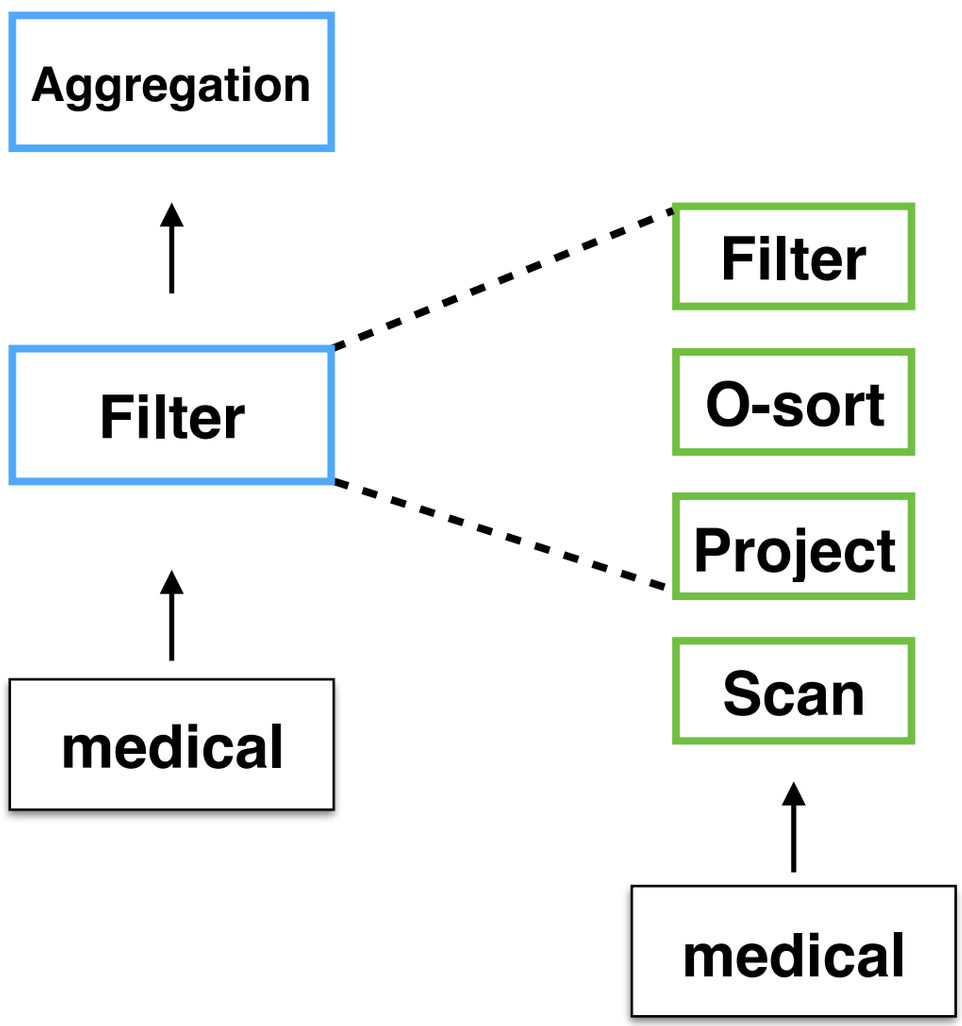


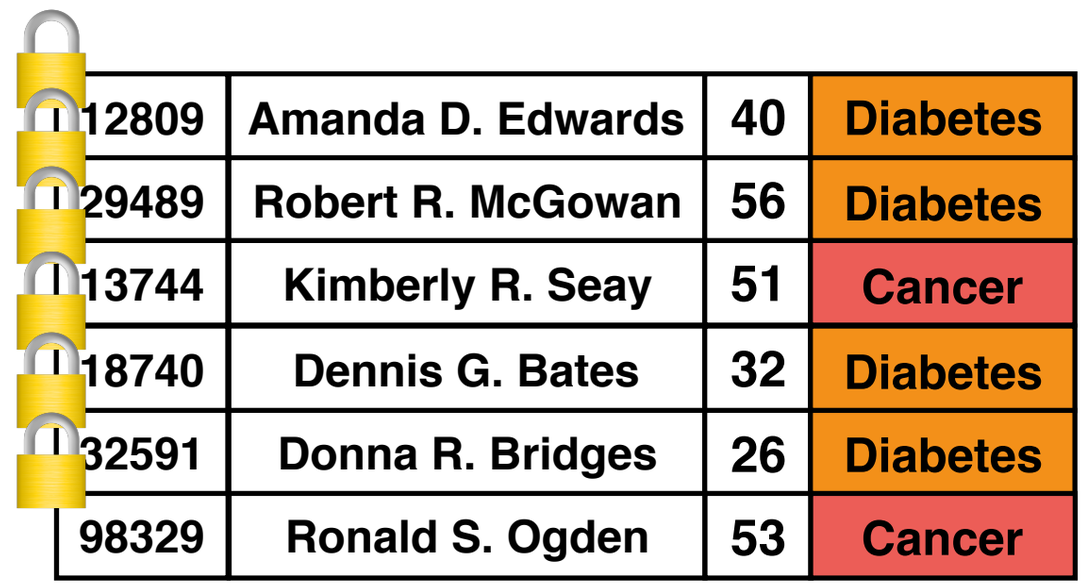
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

Rule-based optimization

Opaque op.

Logical op.



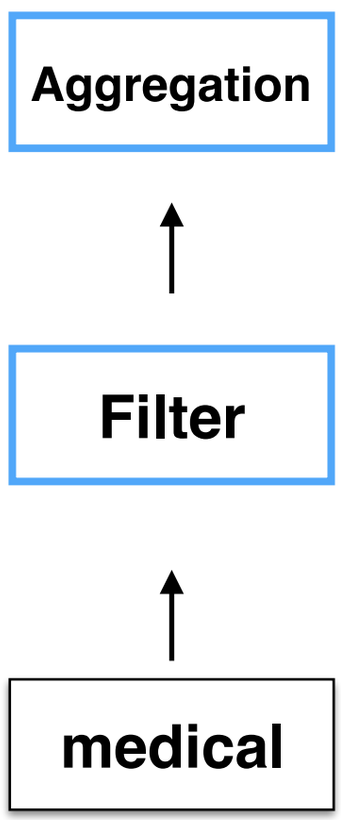


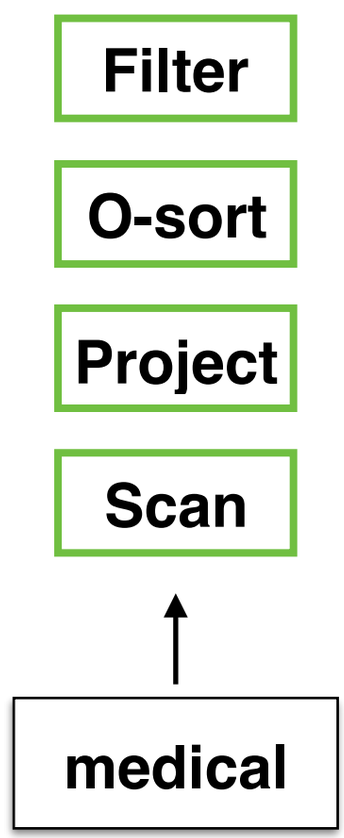
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

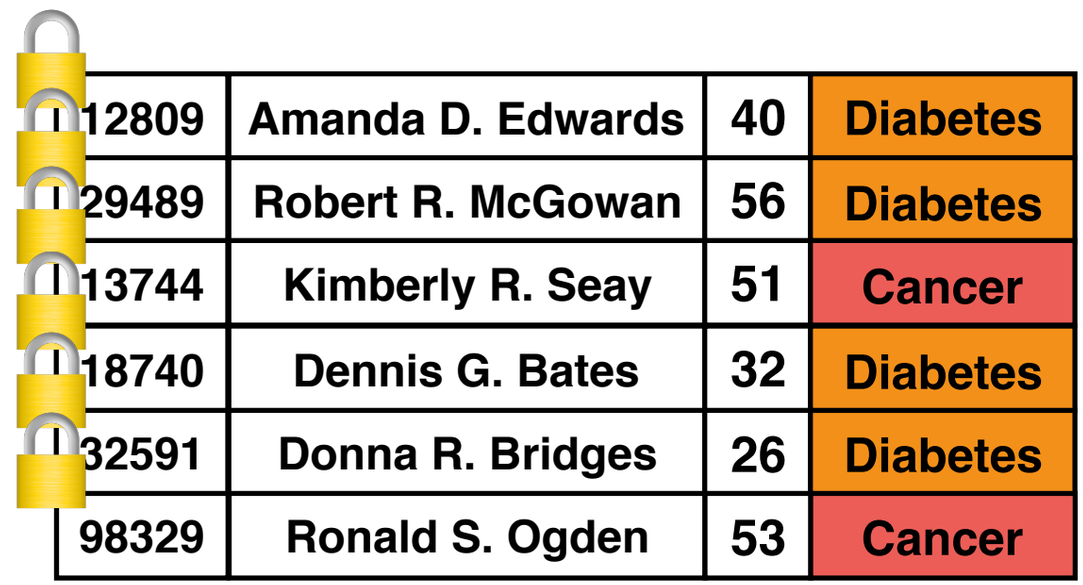
Rule-based optimization

Opaque op.

Logical op.





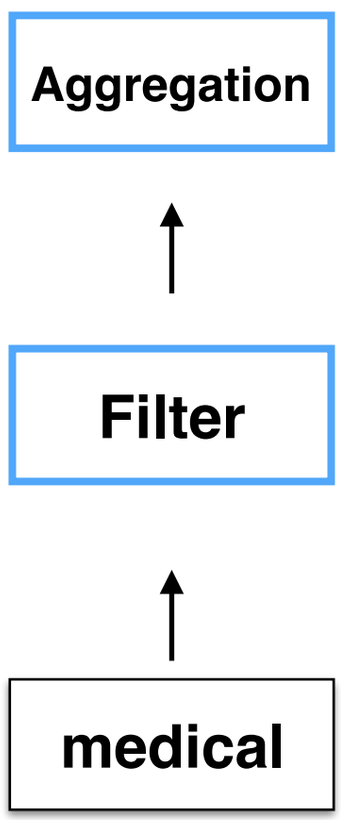


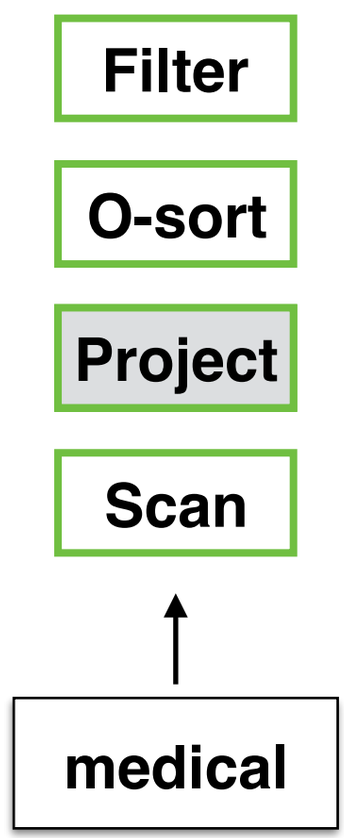
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

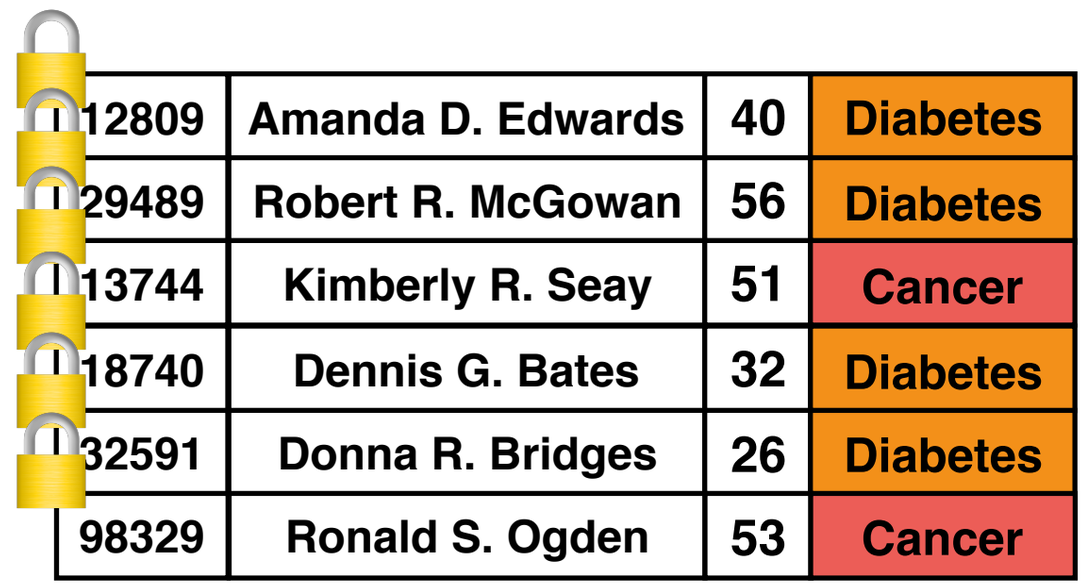
Rule-based optimization

Opaque op.

Logical op.





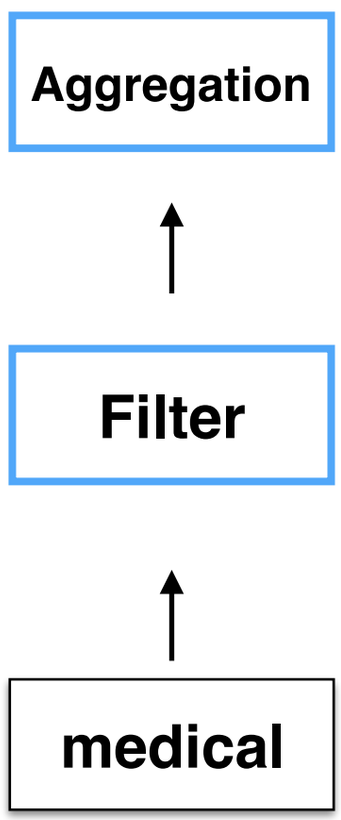


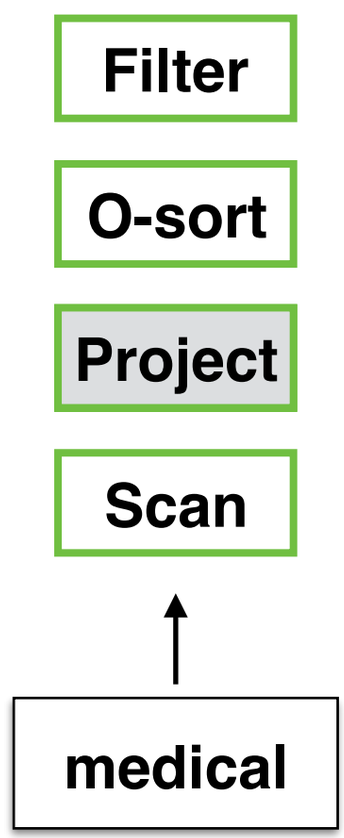
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

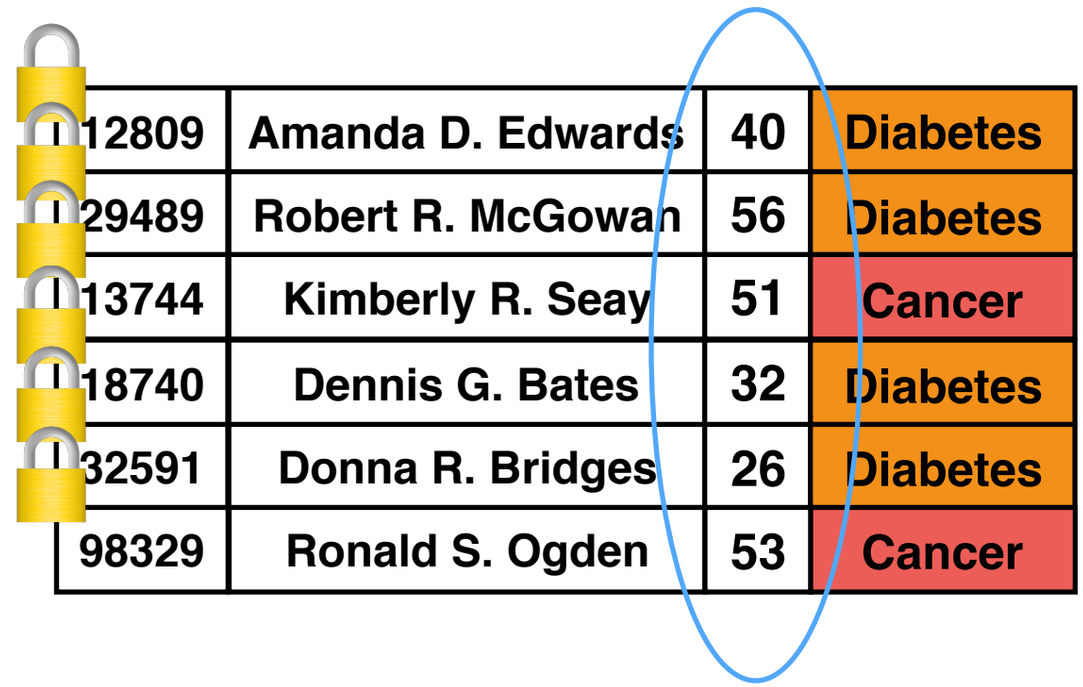
Rule-based optimization

Opaque op.

Logical op.





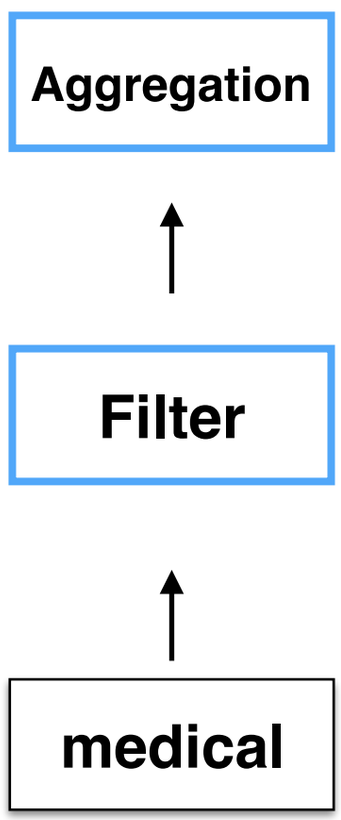


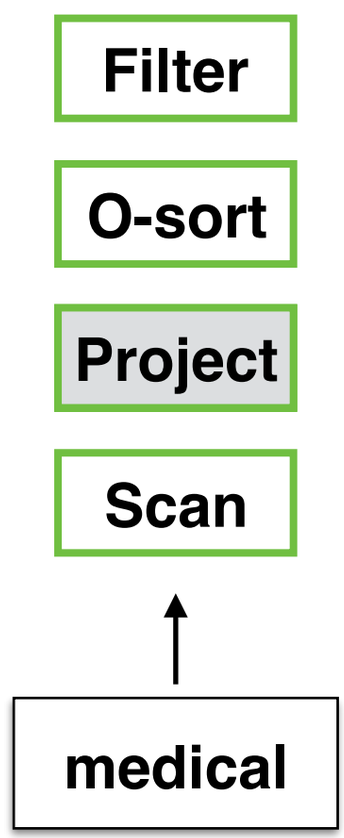
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

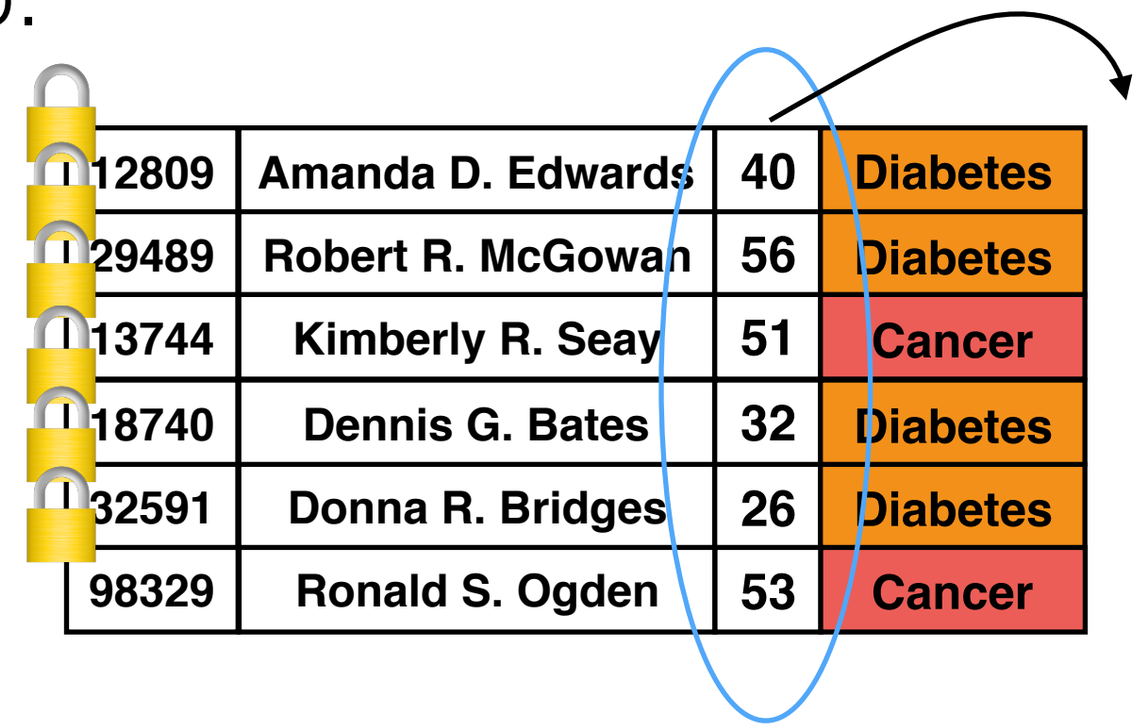
Rule-based optimization

Opaque op.

Logical op.





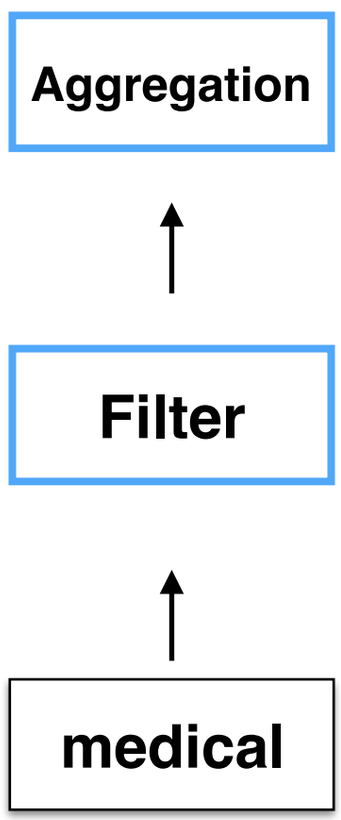


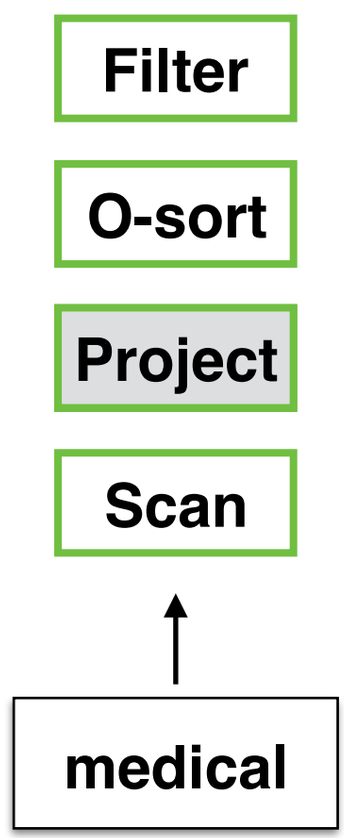
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

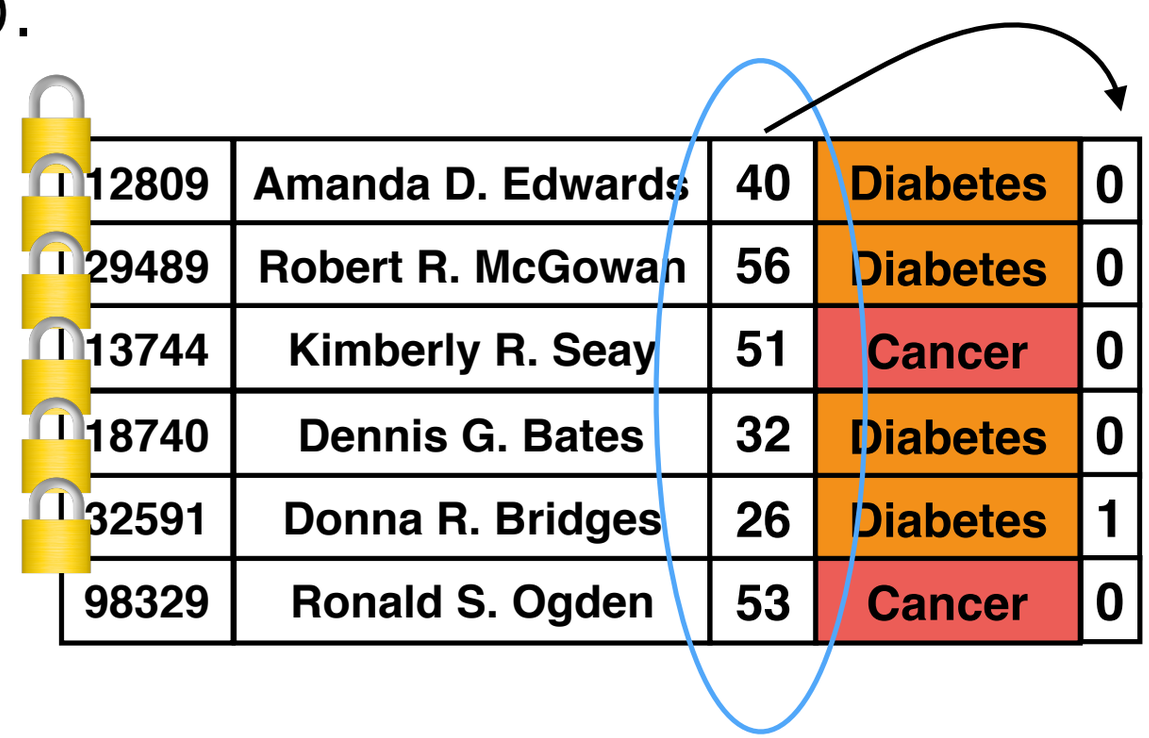
Rule-based optimization

Opaque op.

Logical op.





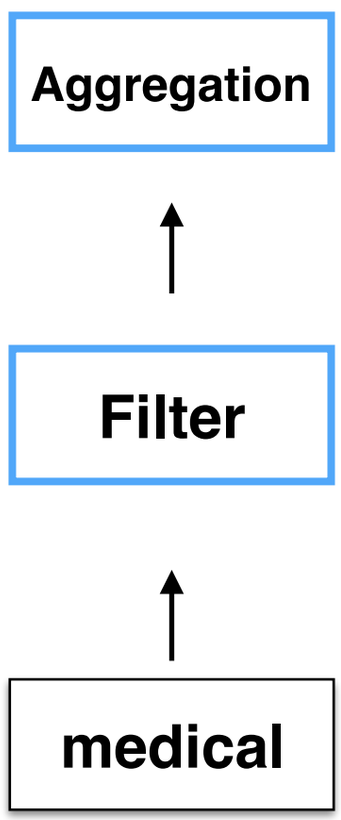


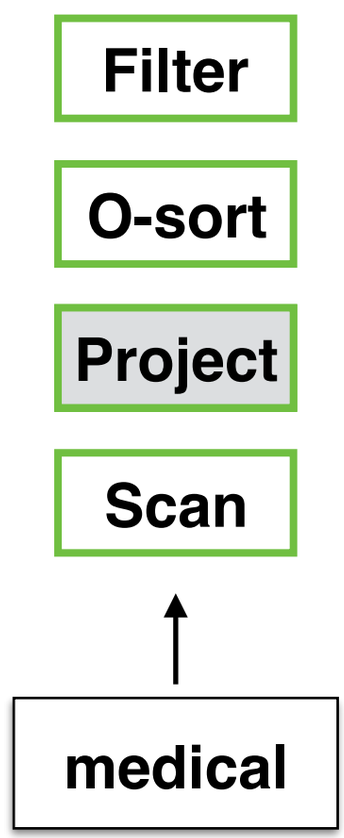
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

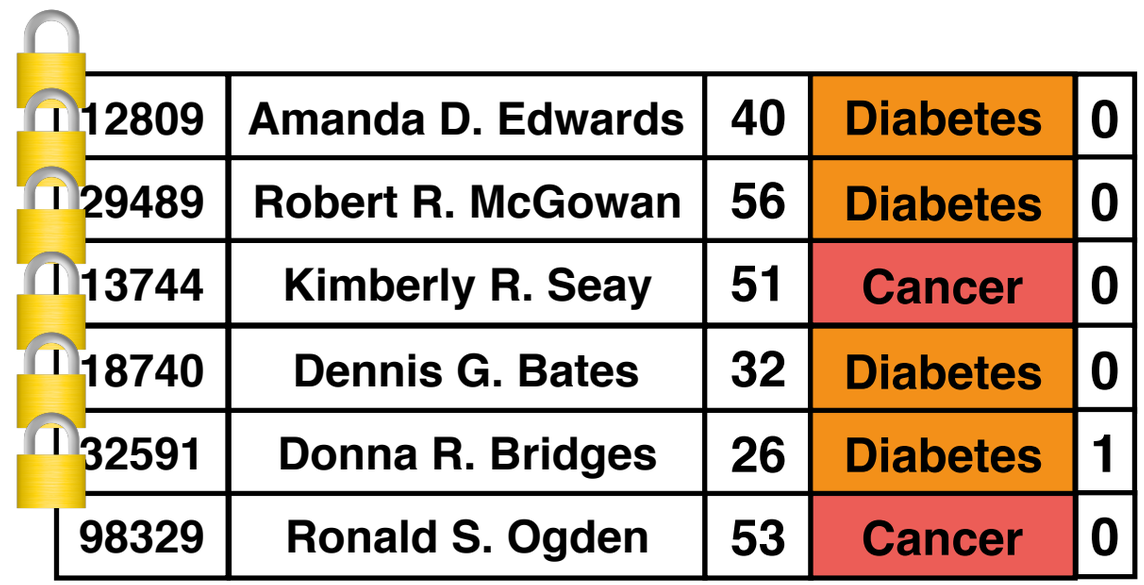
Rule-based optimization

Opaque op.

Logical op.





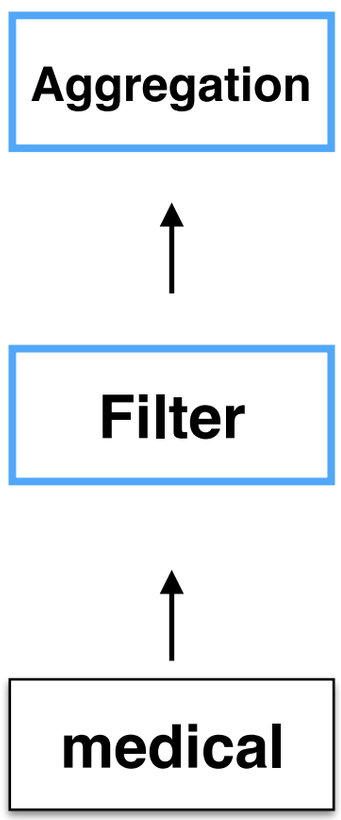


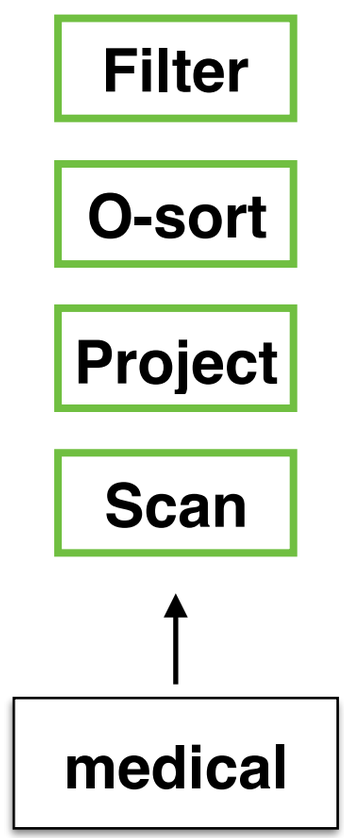
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

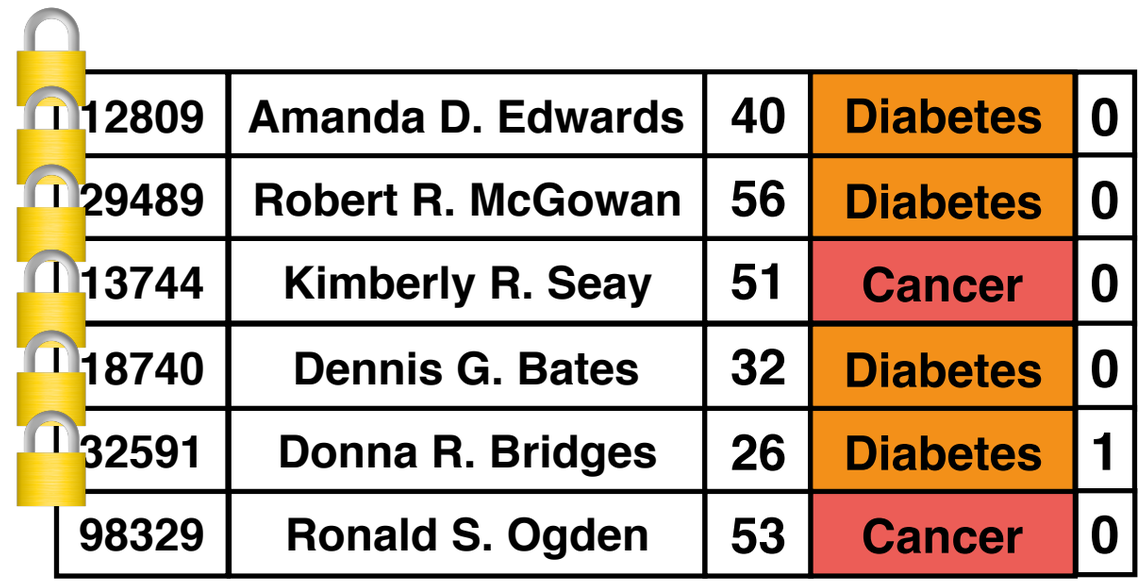
Rule-based optimization

Opaque op.

Logical op.





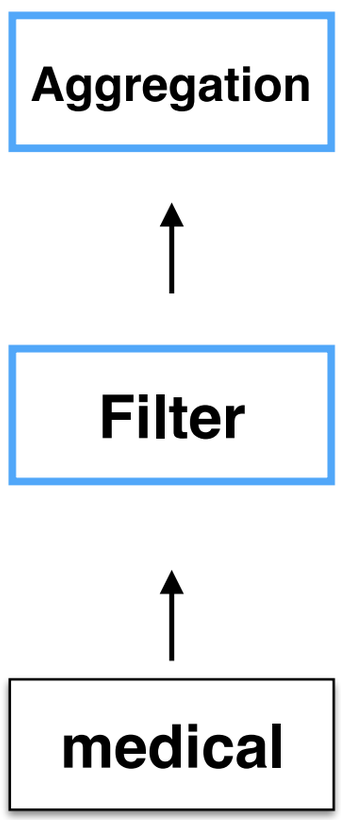


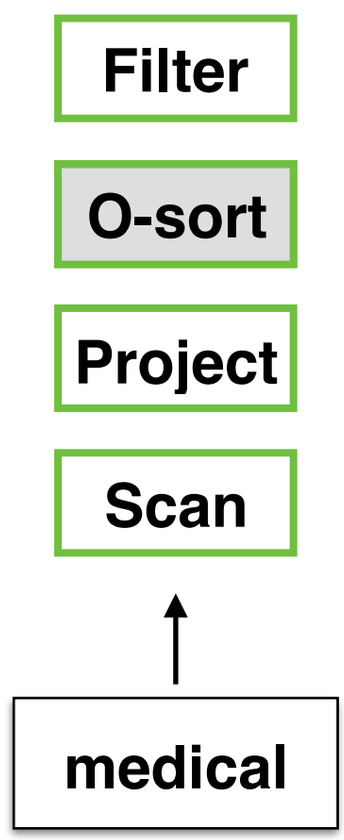
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

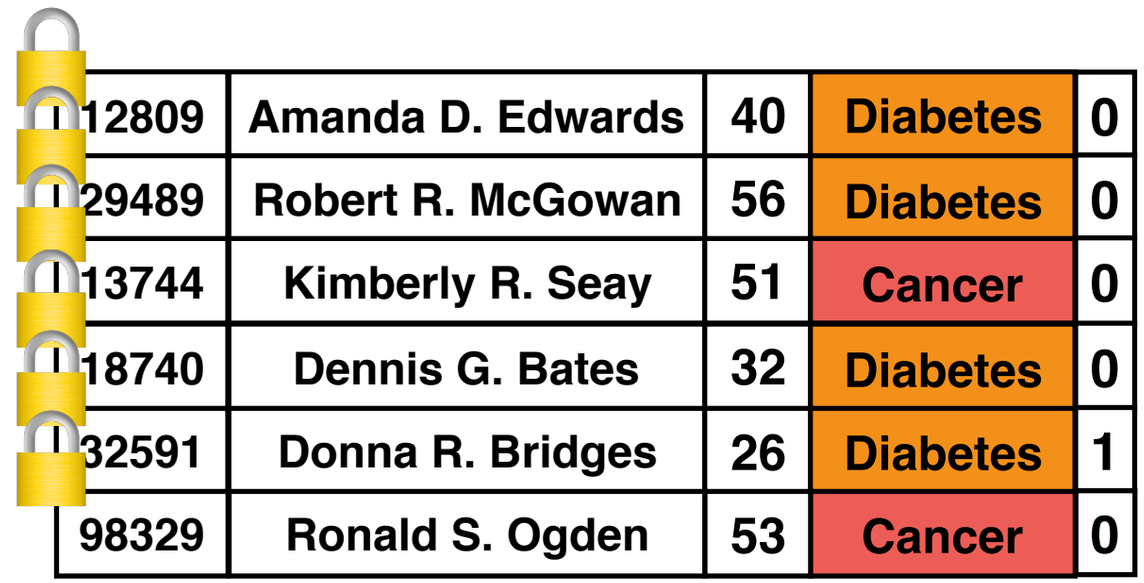
Rule-based optimization

Opaque op.

Logical op.





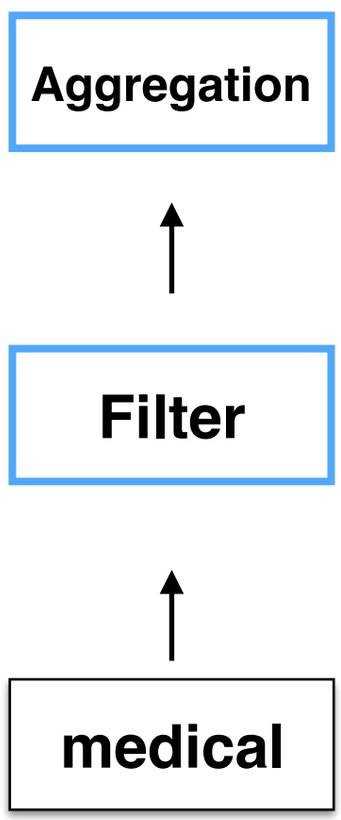


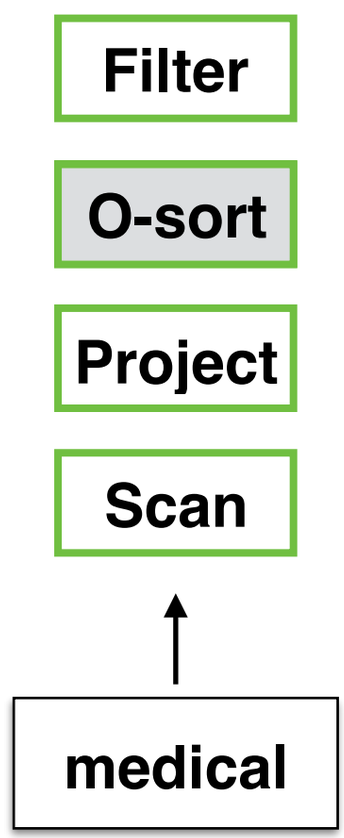
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

Rule-based optimization

Opaque op.

Logical op.



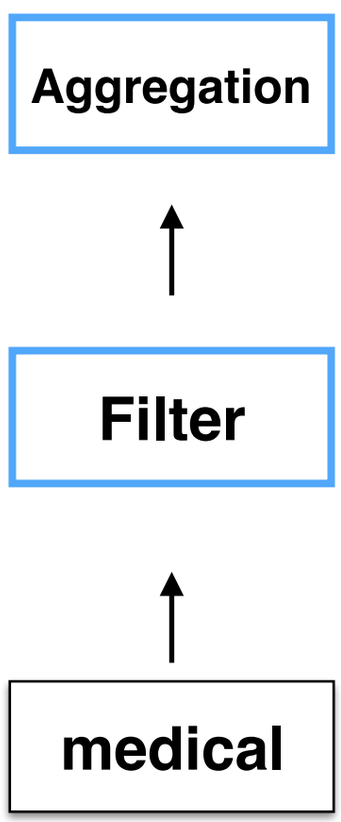


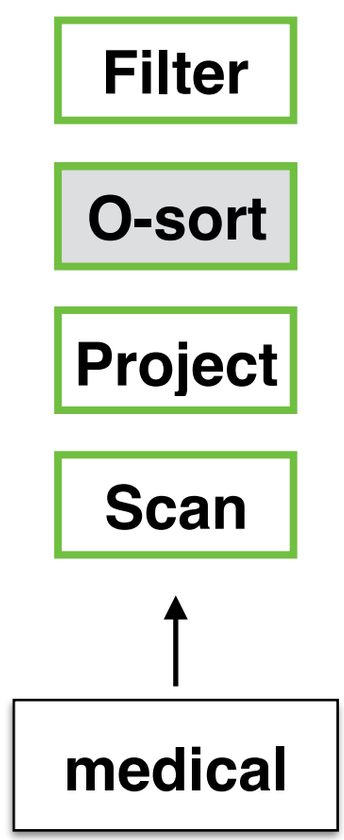
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

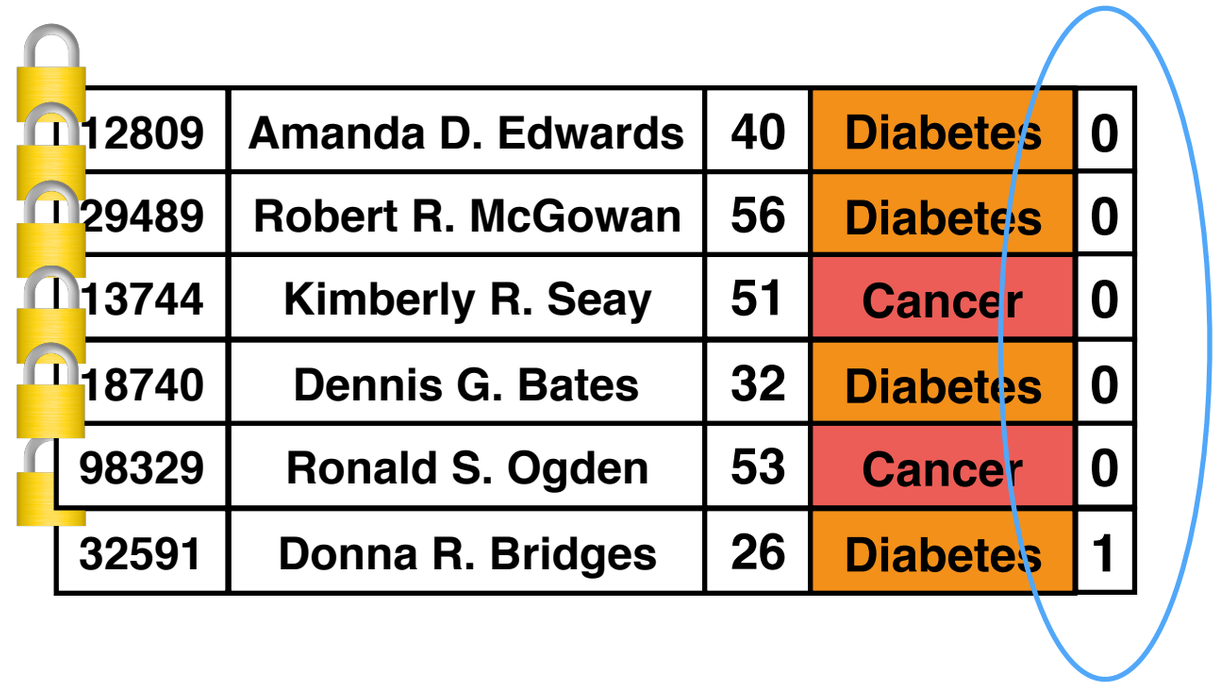
Rule-based optimization

Opaque op.

Logical op.





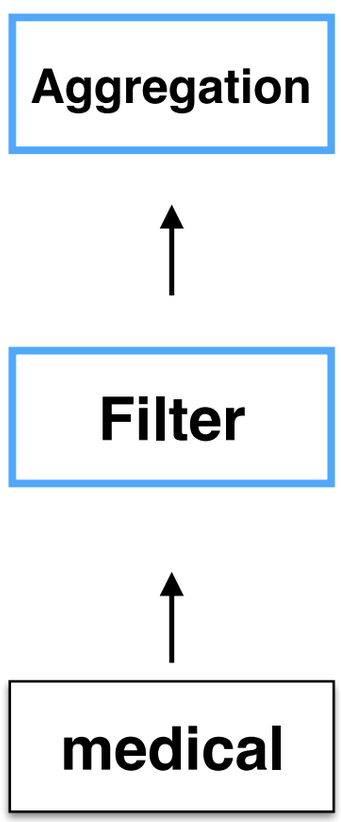


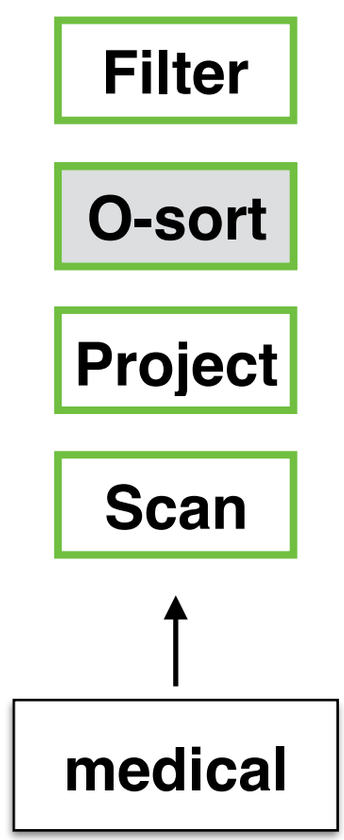
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0
32591	Donna R. Bridges	26	Diabetes	1

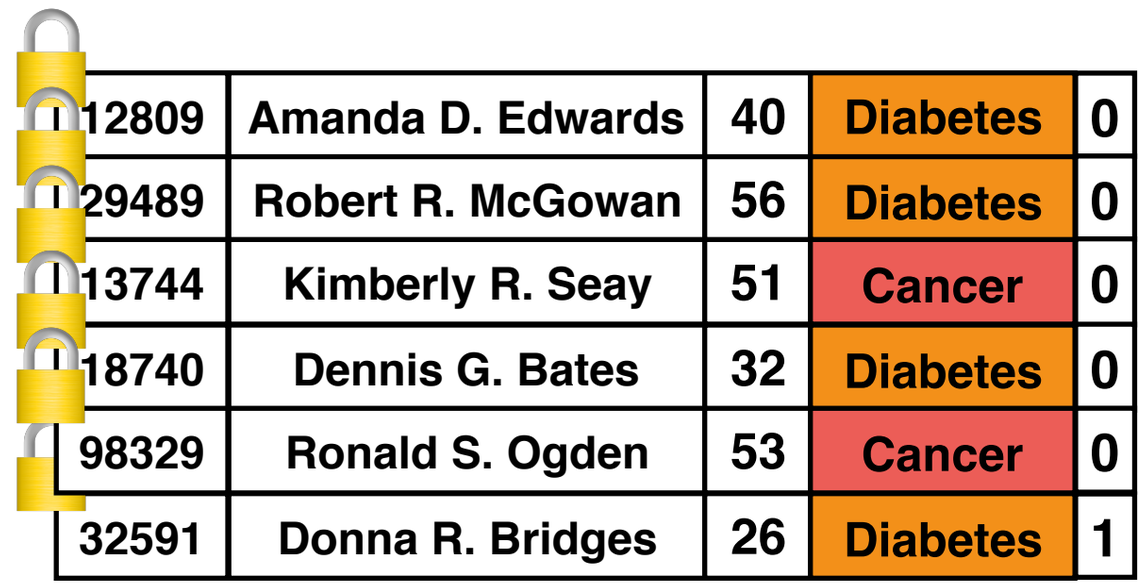
Rule-based optimization

Opaque op.

Logical op.





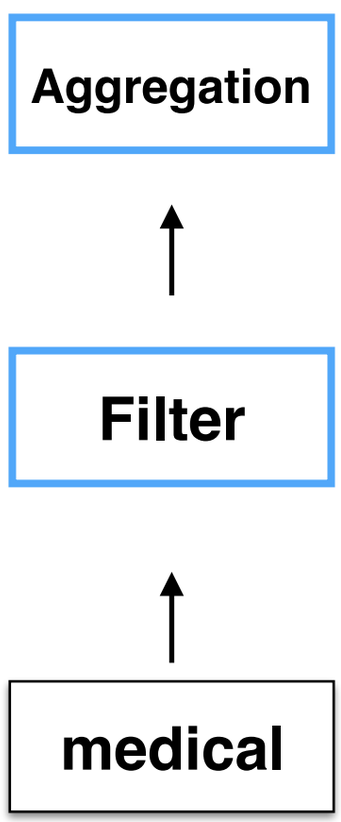


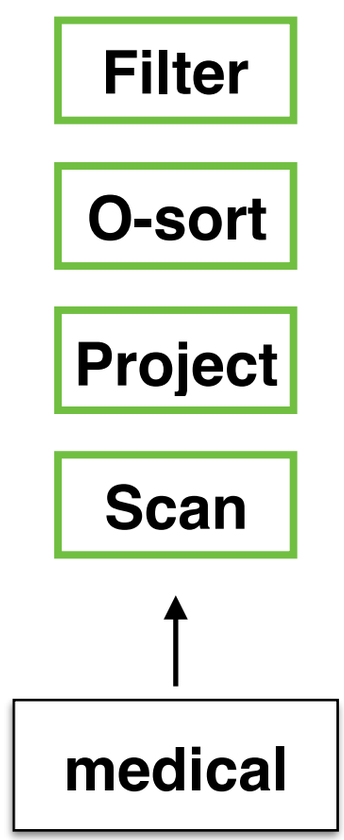
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0
32591	Donna R. Bridges	26	Diabetes	1

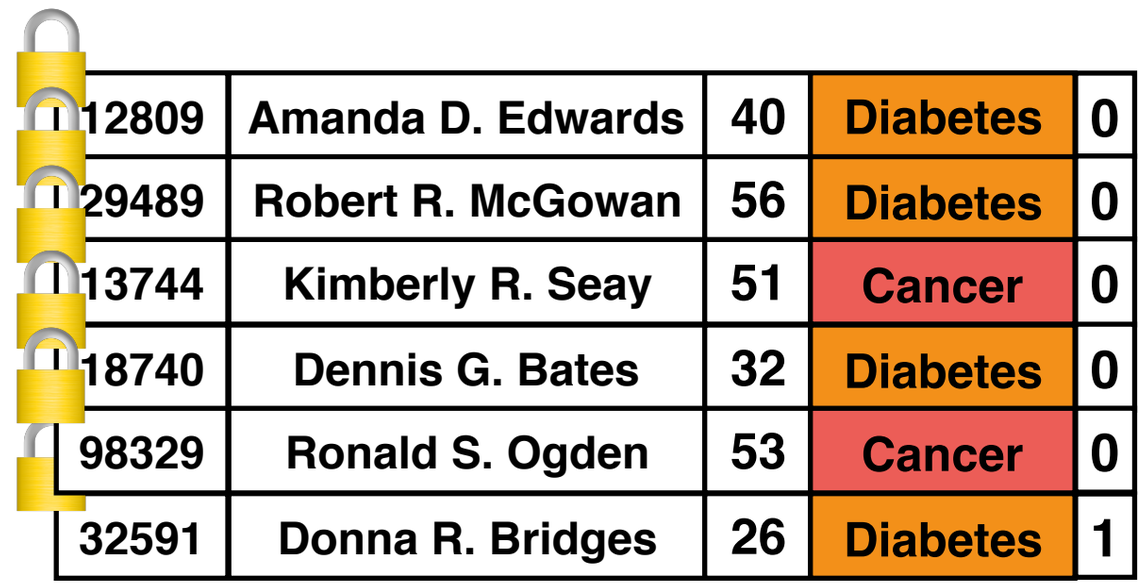
Rule-based optimization

Opaque op.

Logical op.





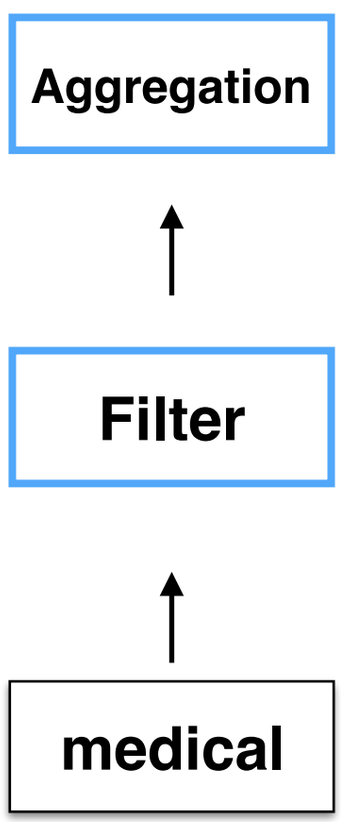


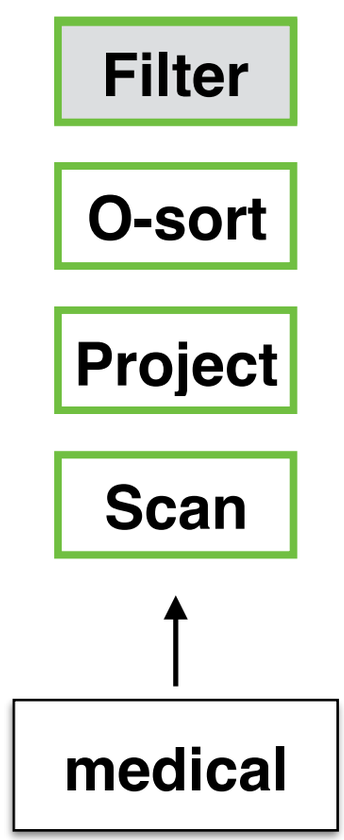
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0
32591	Donna R. Bridges	26	Diabetes	1

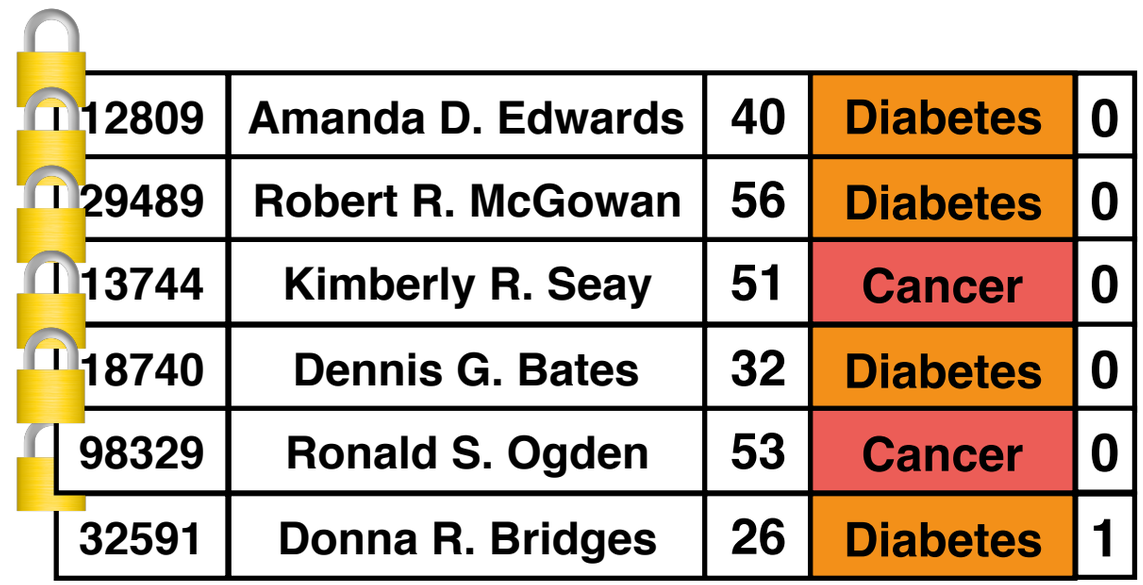
Rule-based optimization

Opaque op.

Logical op.





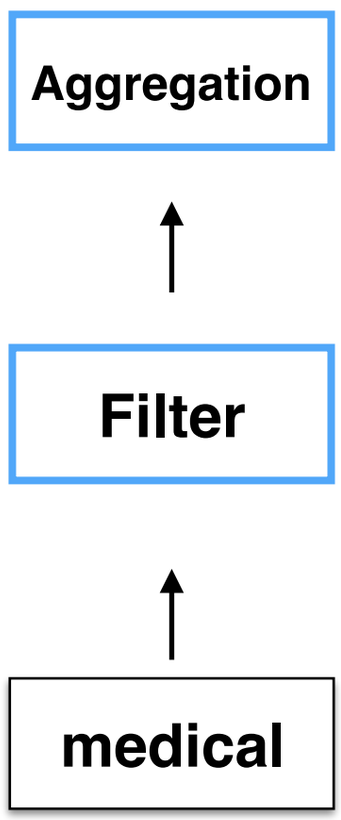


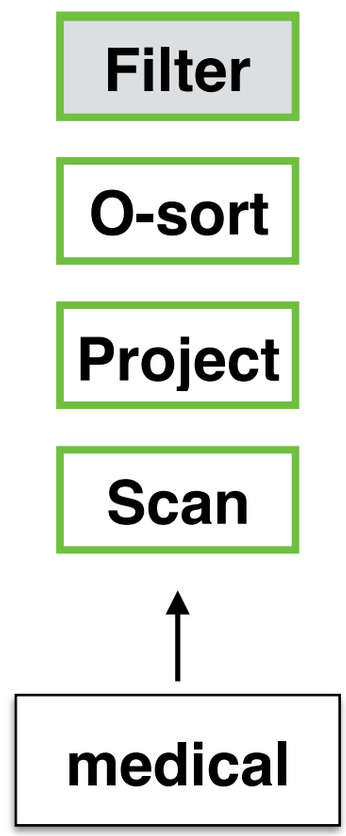
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0
32591	Donna R. Bridges	26	Diabetes	1

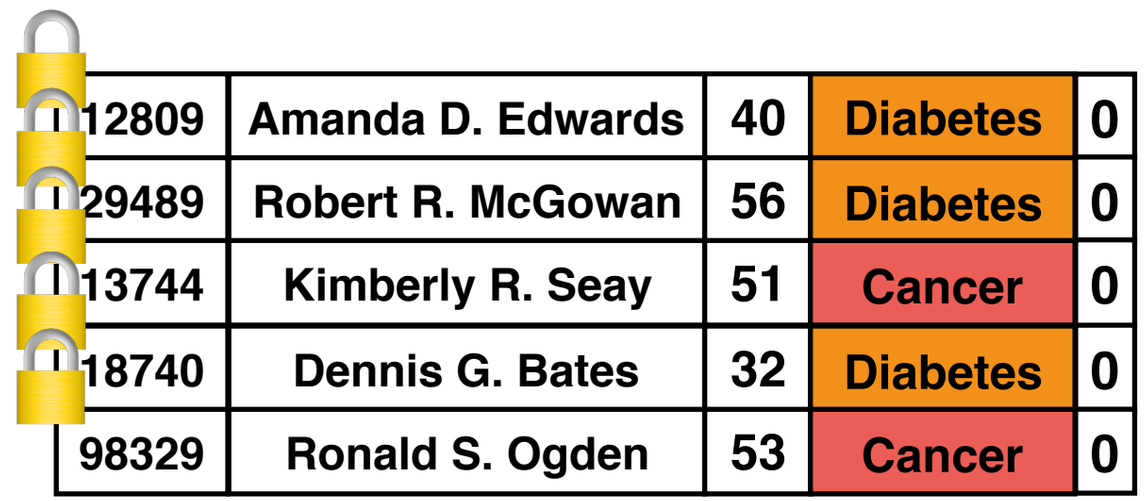
Rule-based optimization

Opaque op.

Logical op.





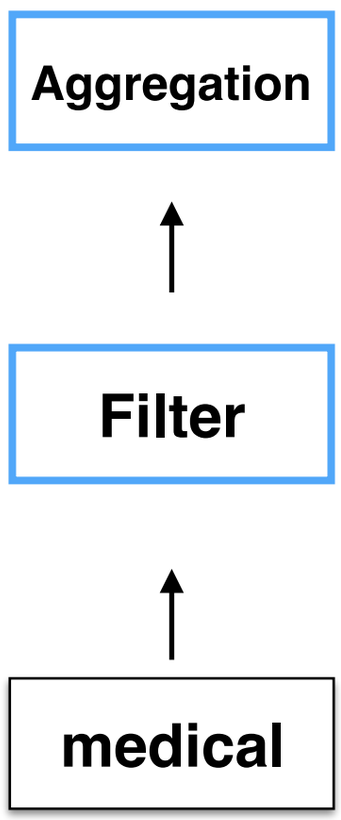


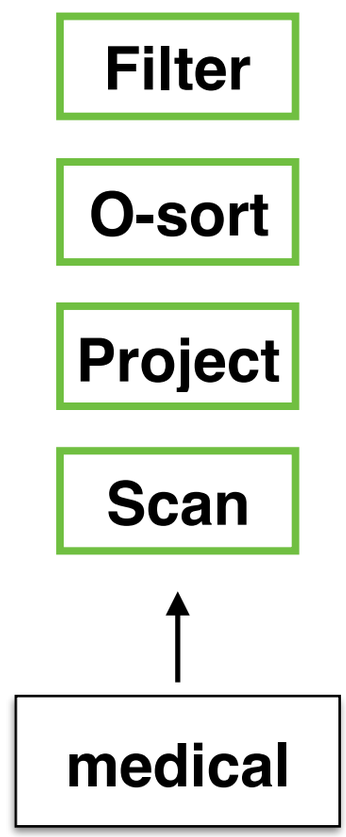
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0

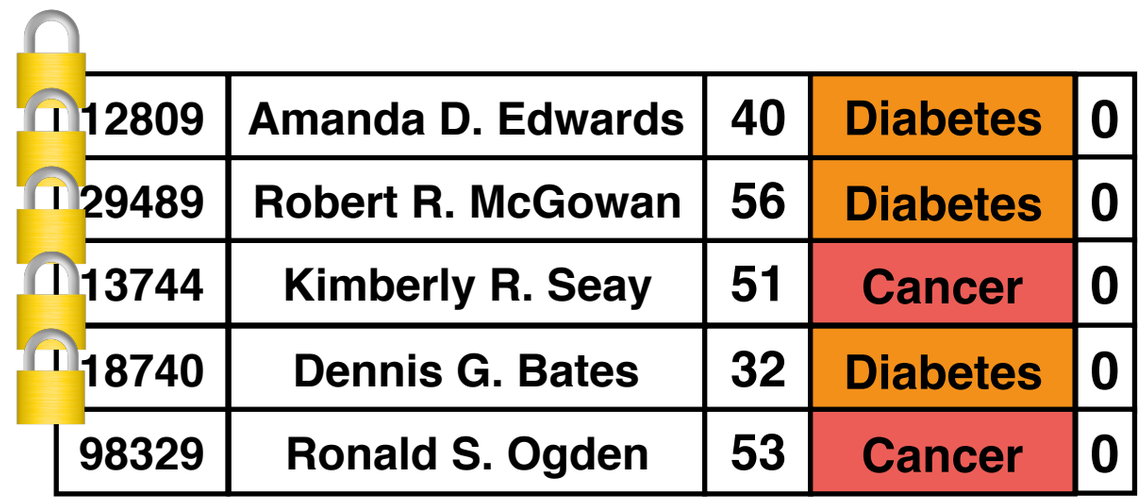
Rule-based optimization

Opaque op.

Logical op.





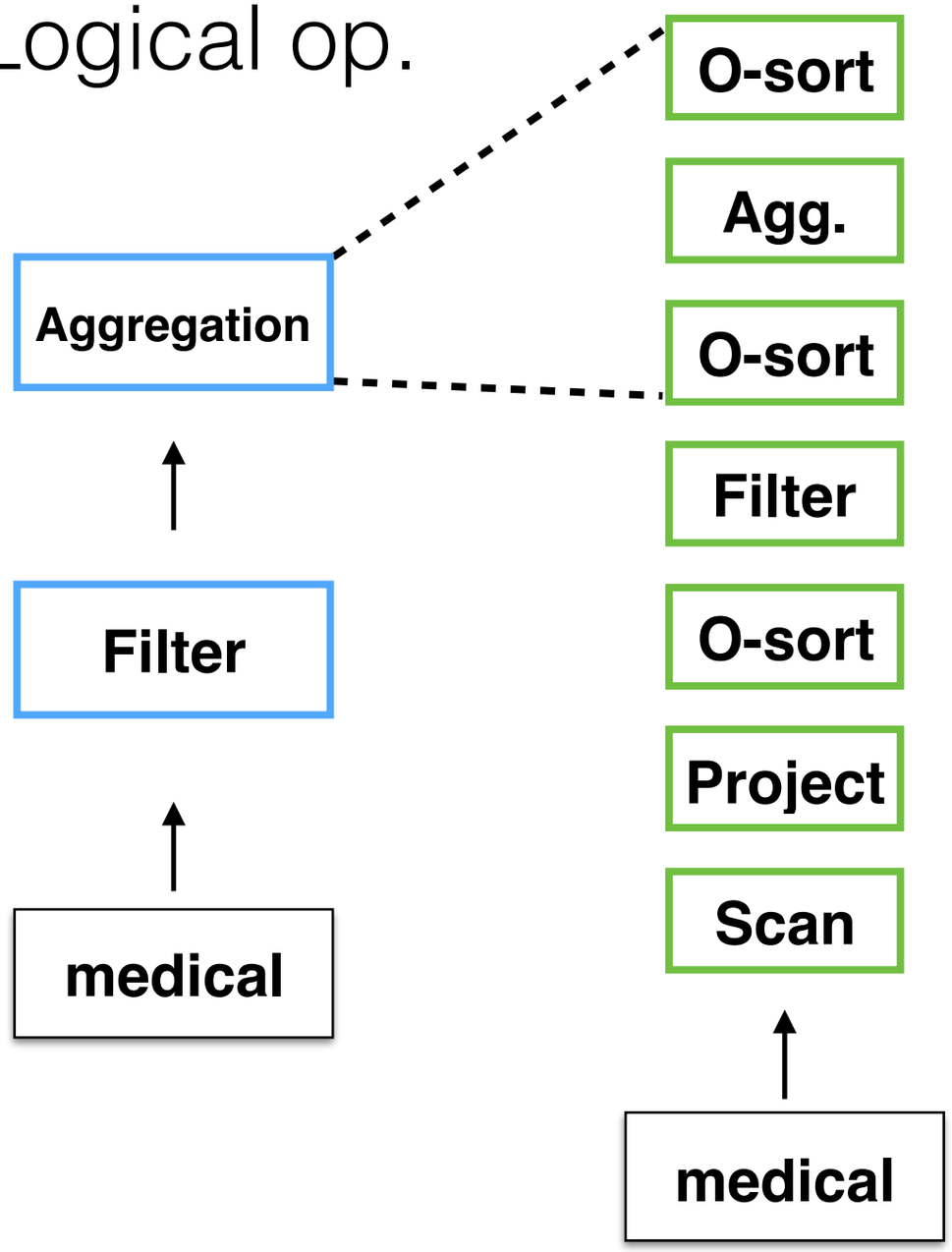


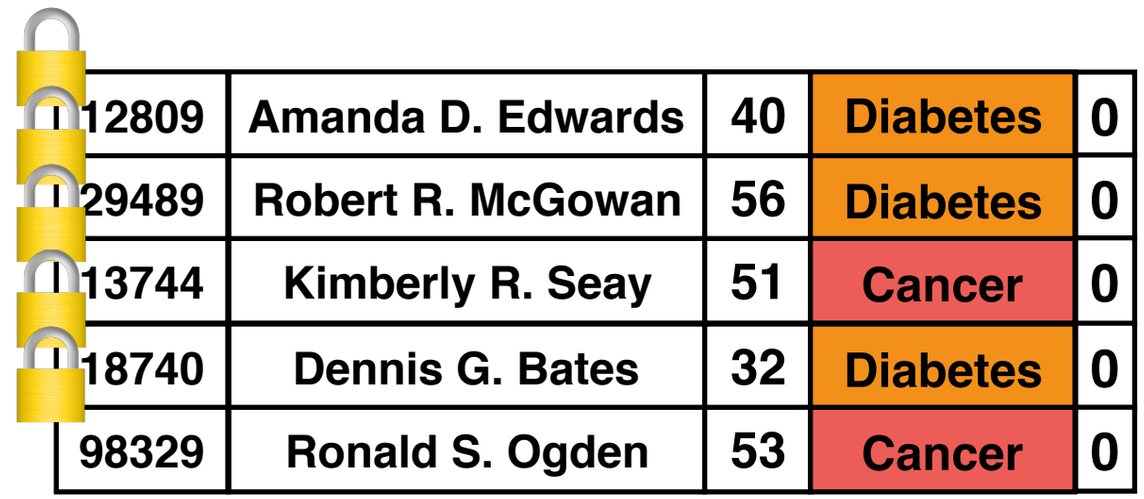
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0

Rule-based optimization

Opaque op.

Logical op.

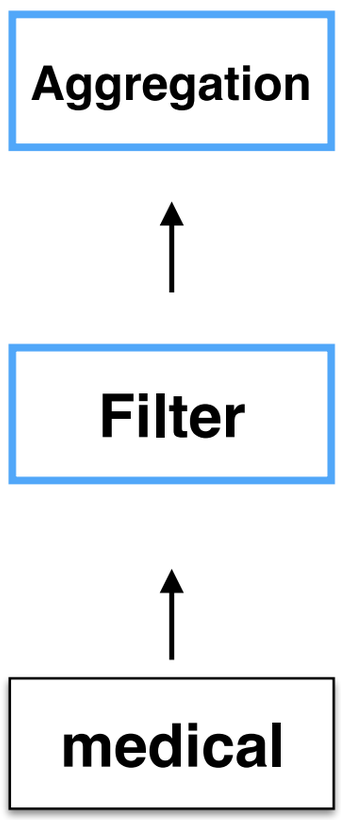




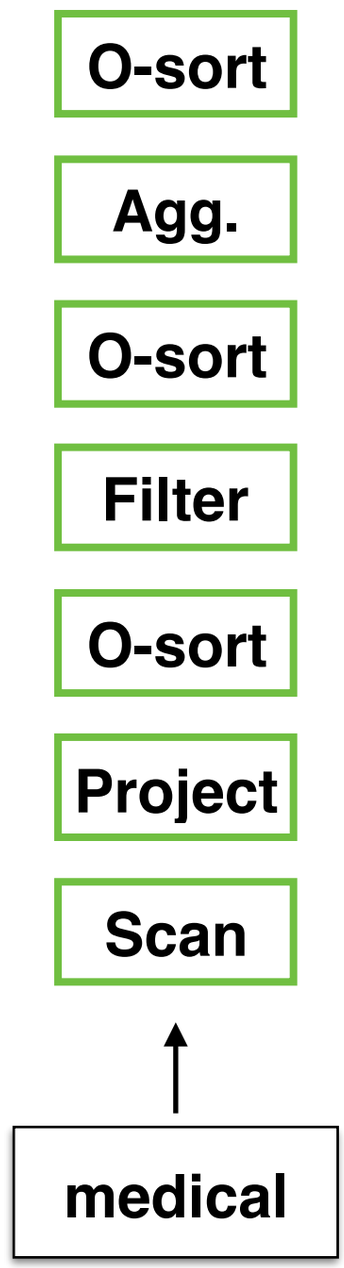
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0

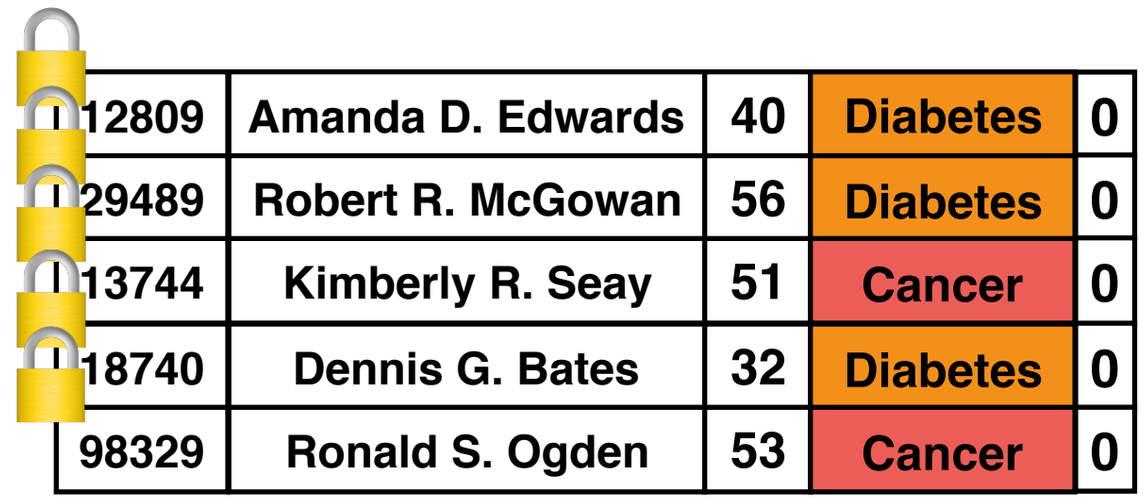
Rule-based optimization

Logical op.



Opaque op.

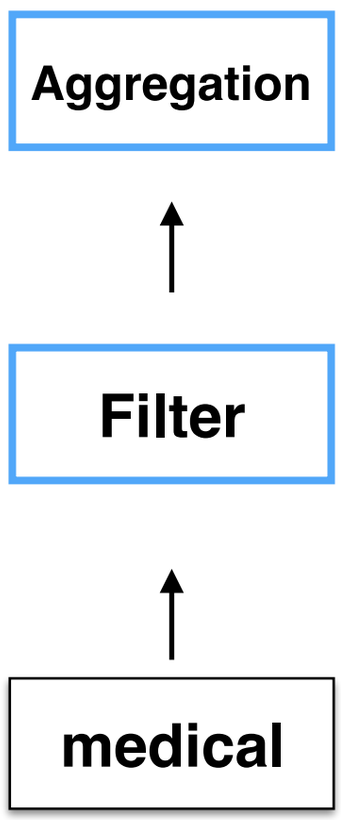




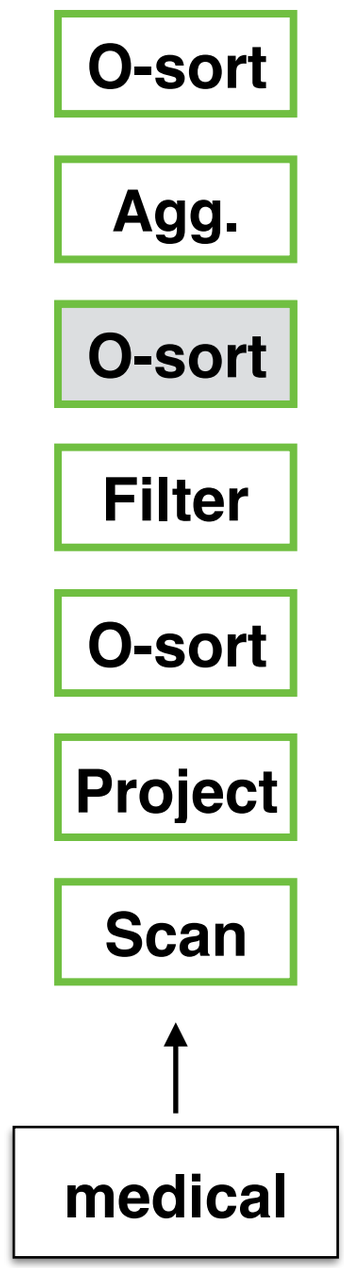
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0

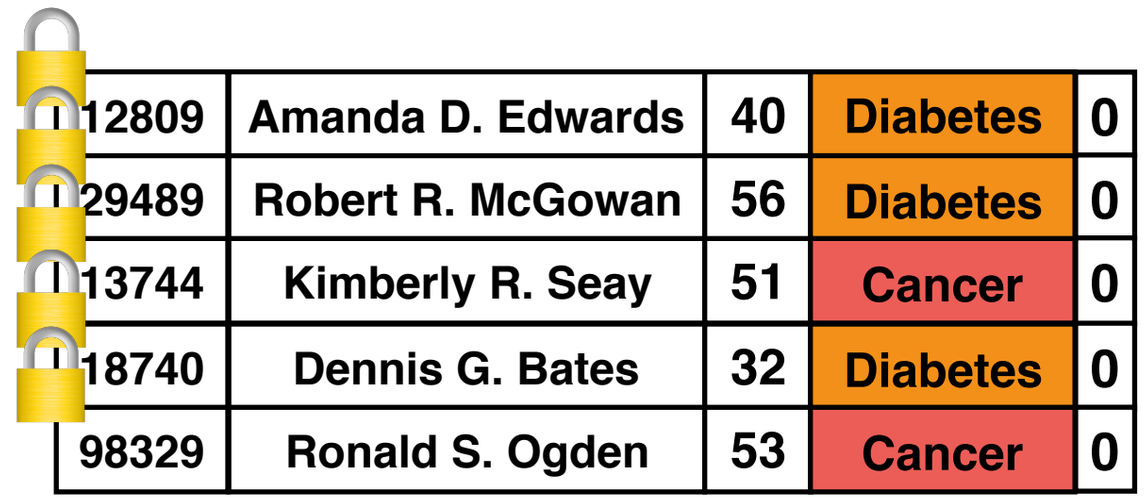
Rule-based optimization

Logical op.



Opaque op.

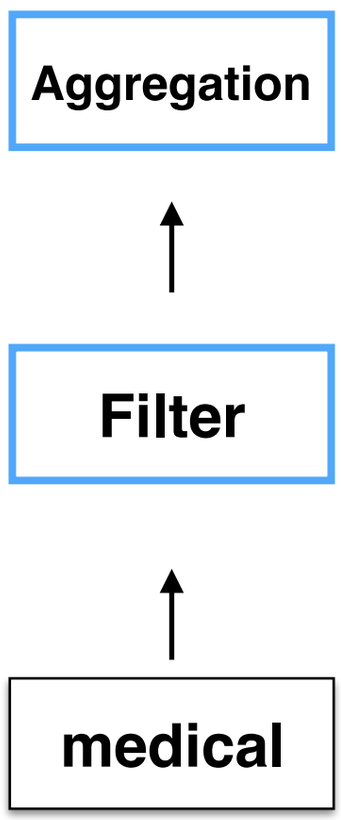




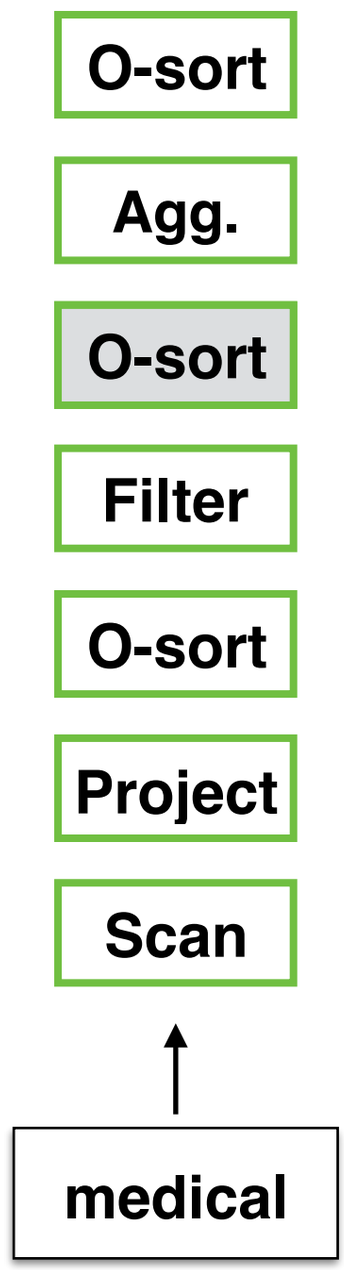
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
98329	Ronald S. Ogden	53	Cancer	0

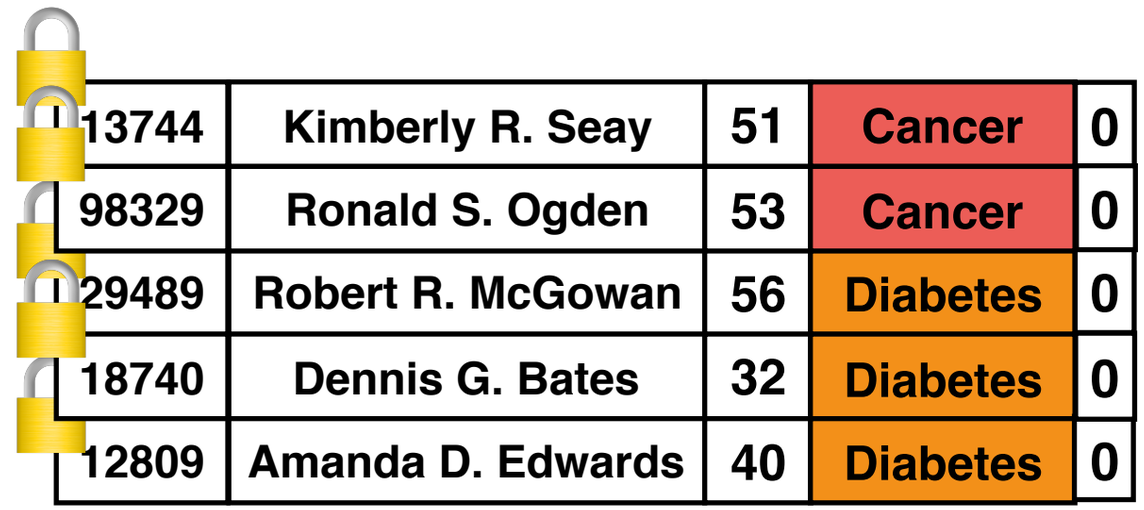
Rule-based optimization

Logical op.



Opaque op.

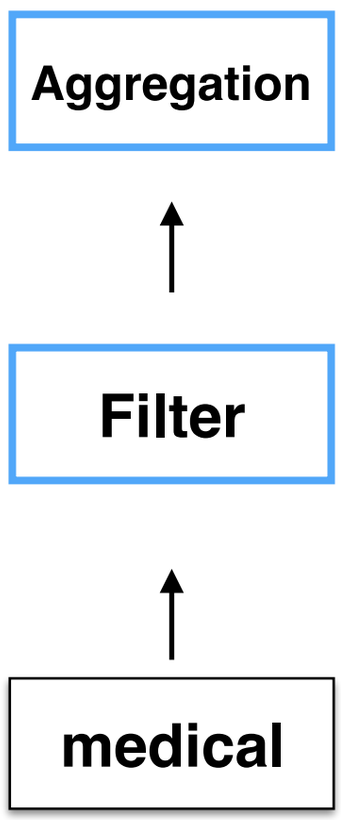




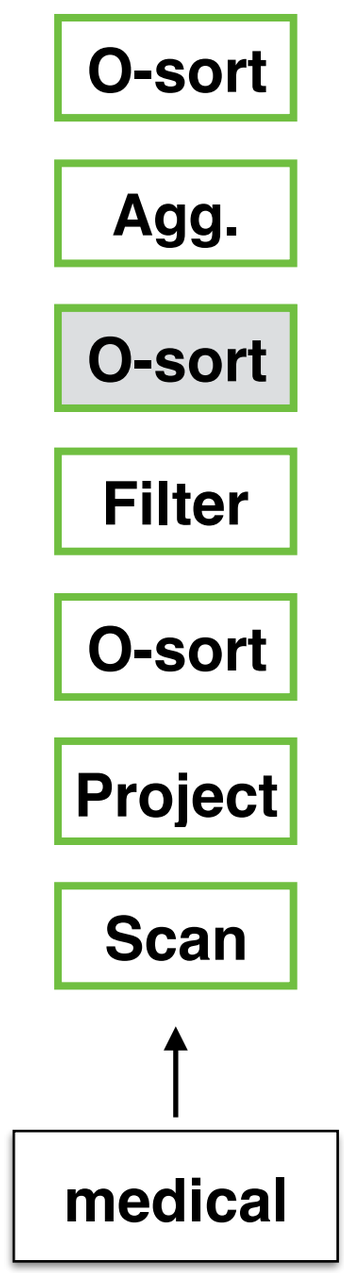
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
29489	Robert R. McGowan	56	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
12809	Amanda D. Edwards	40	Diabetes	0

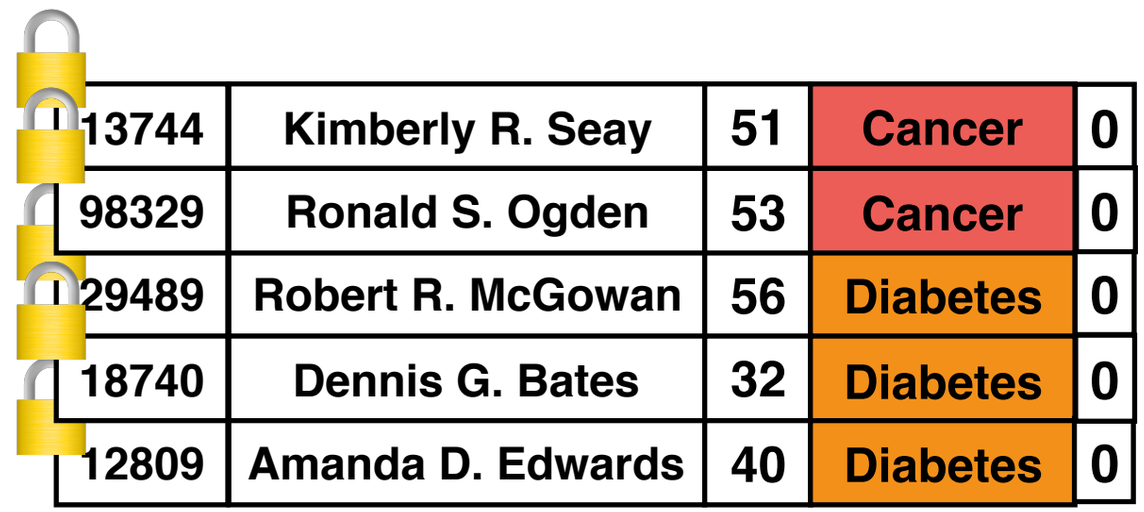
Rule-based optimization

Logical op.



Opaque op.



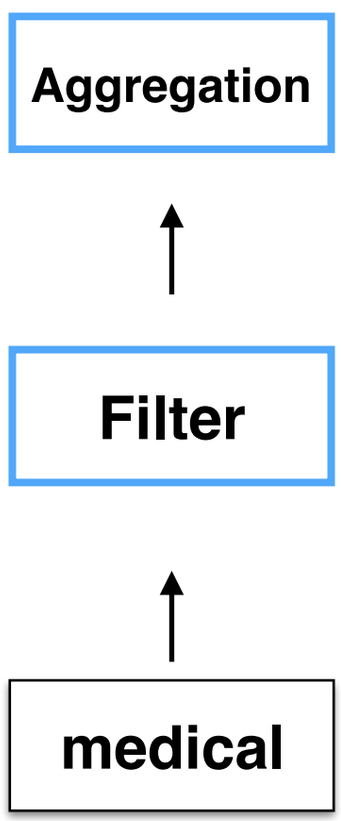


13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
29489	Robert R. McGowan	56	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
12809	Amanda D. Edwards	40	Diabetes	0

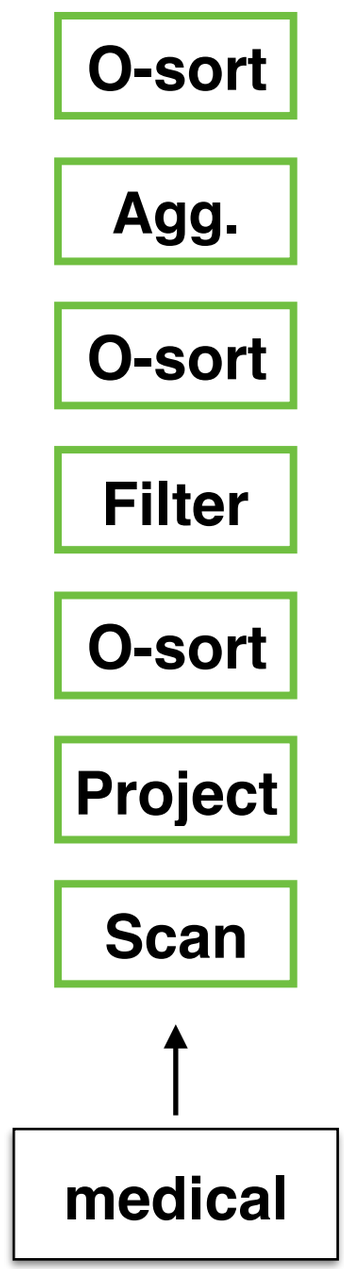
Can we remove any sort?

Rule-based optimization

Logical op.

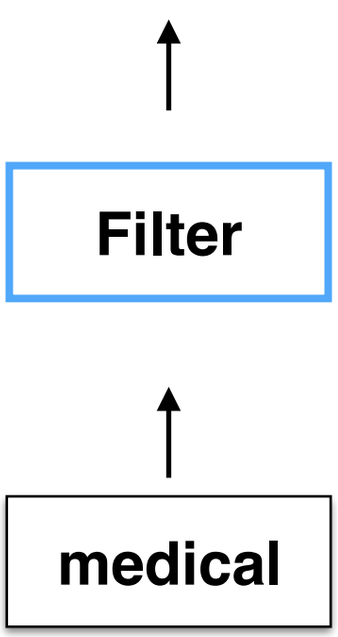


Opaque op.

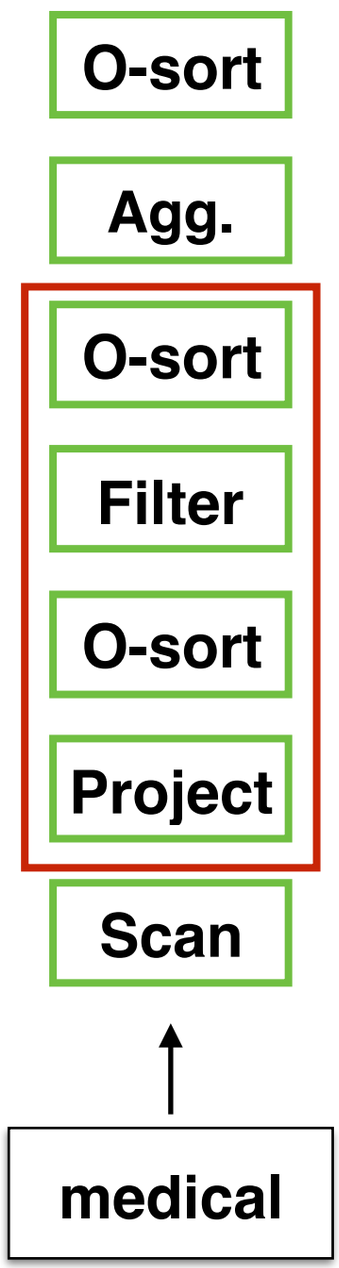


Rule-based optimization

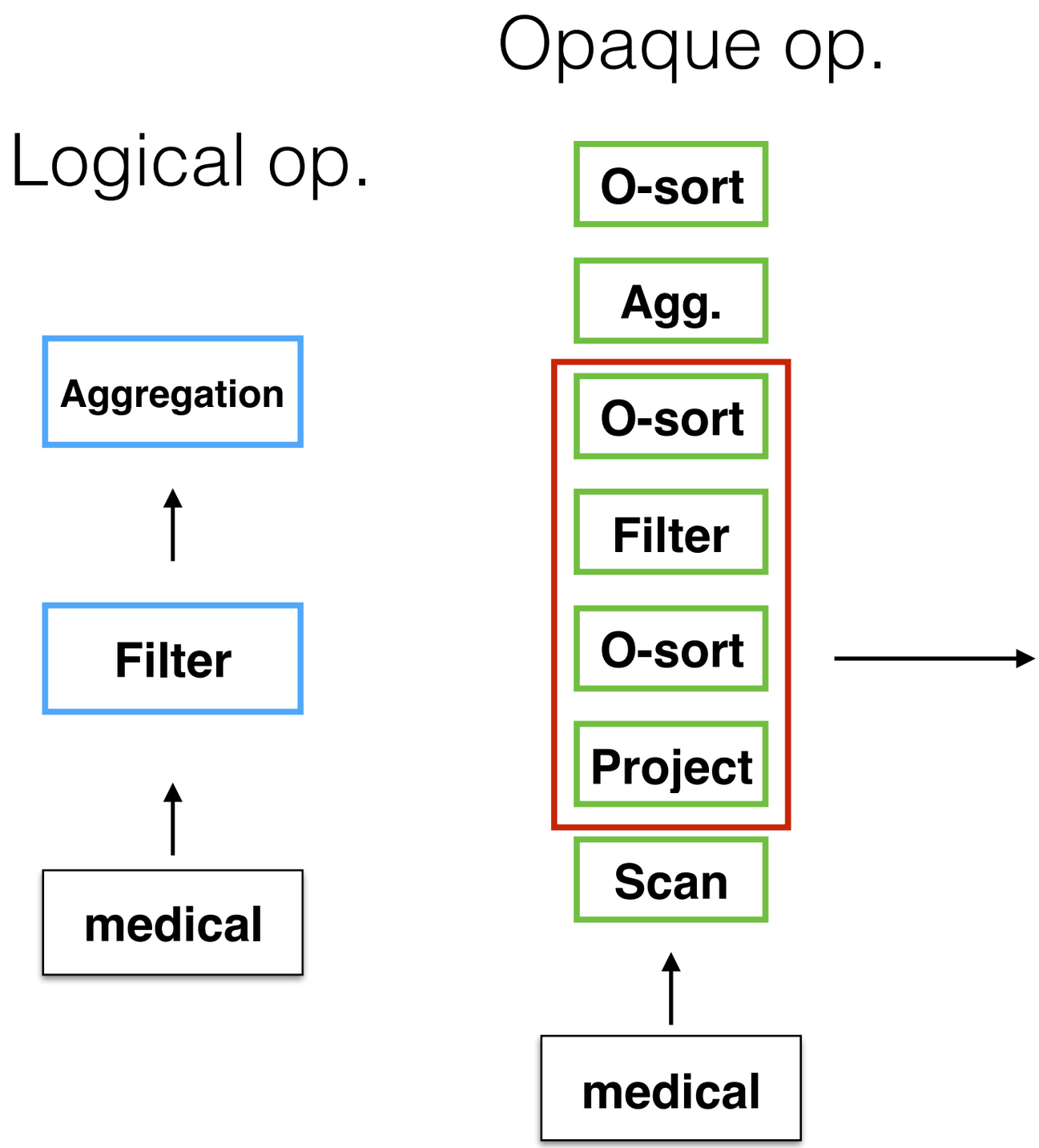
Logical op.



Opaque op.

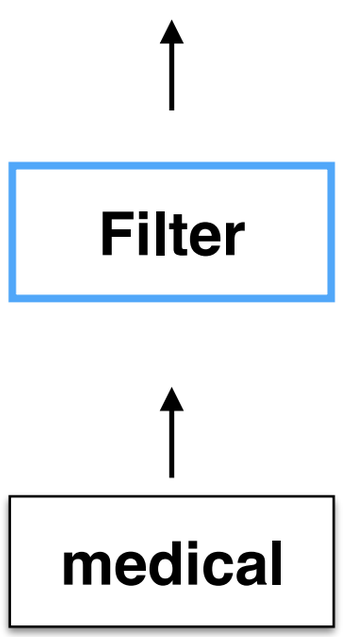


Rule-based optimization

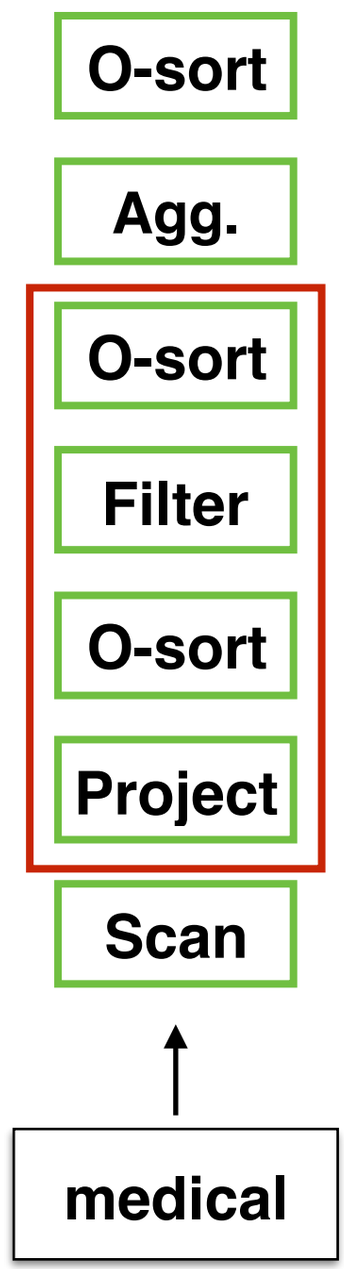


Rule-based optimization

Logical op.

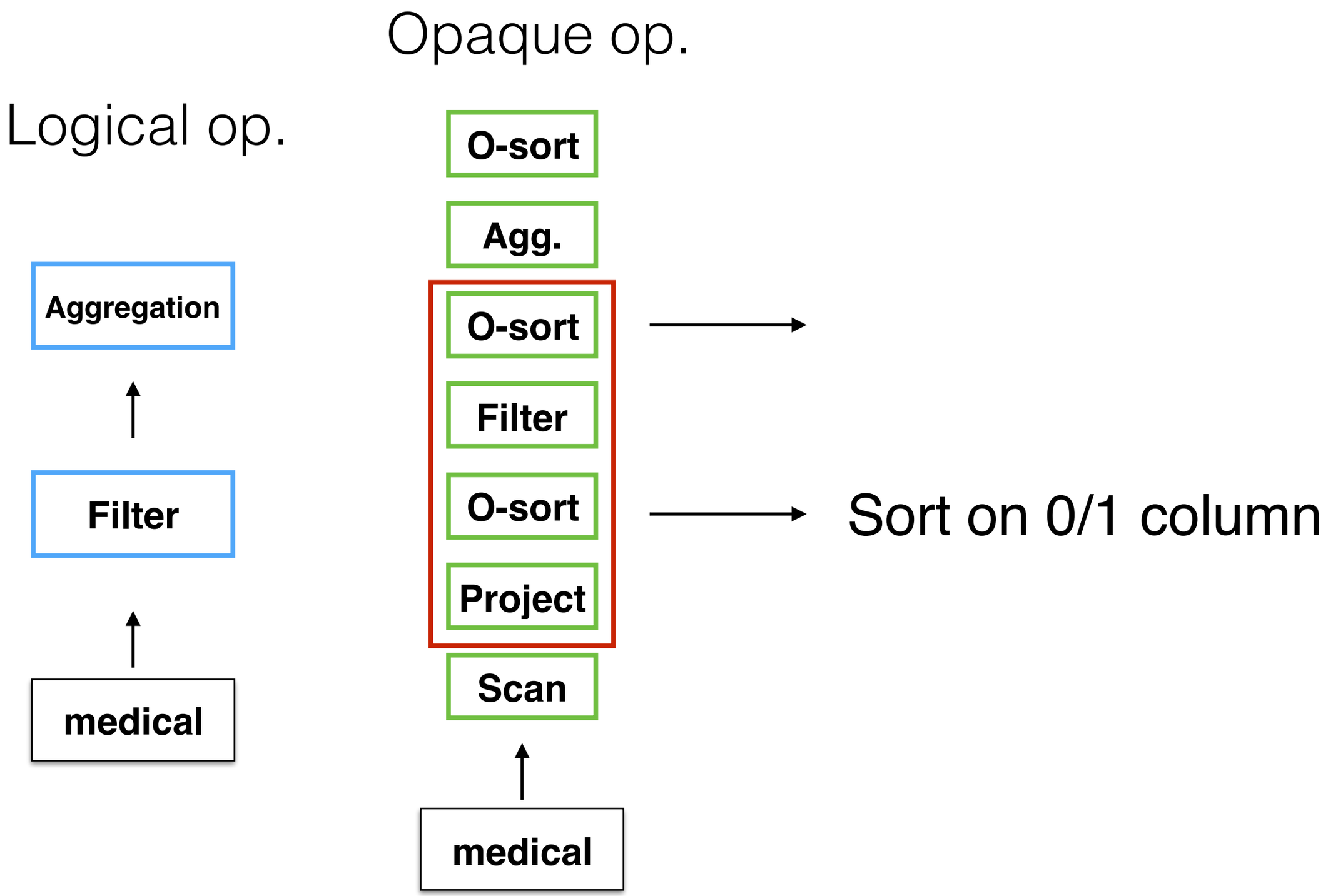


Opaque op.

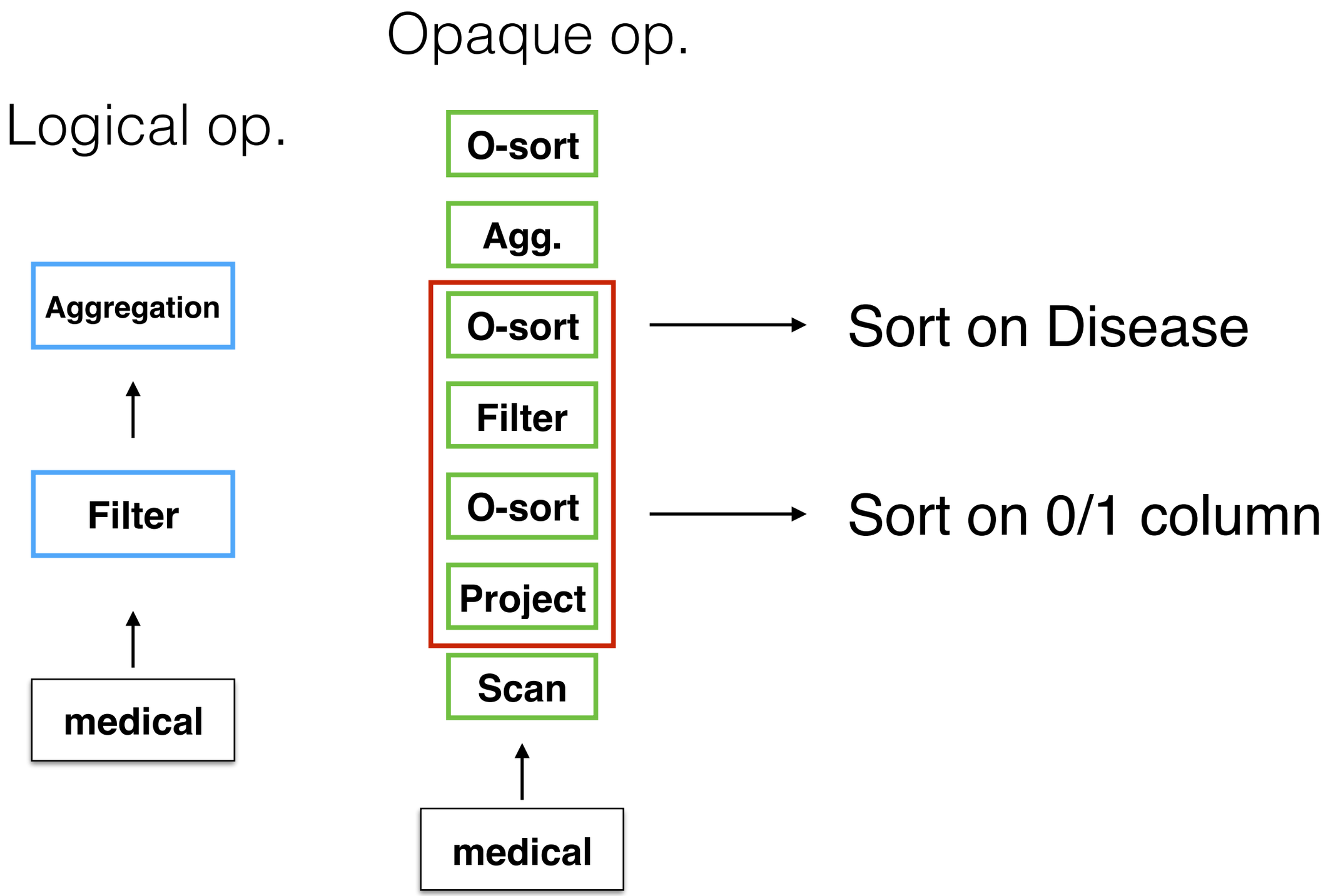


Sort on 0/1 column

Rule-based optimization

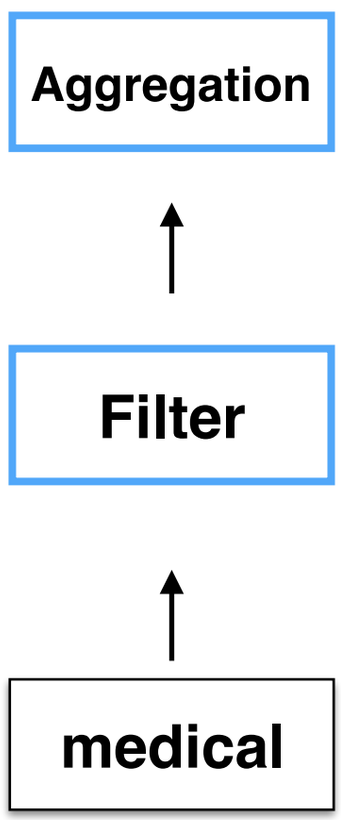


Rule-based optimization

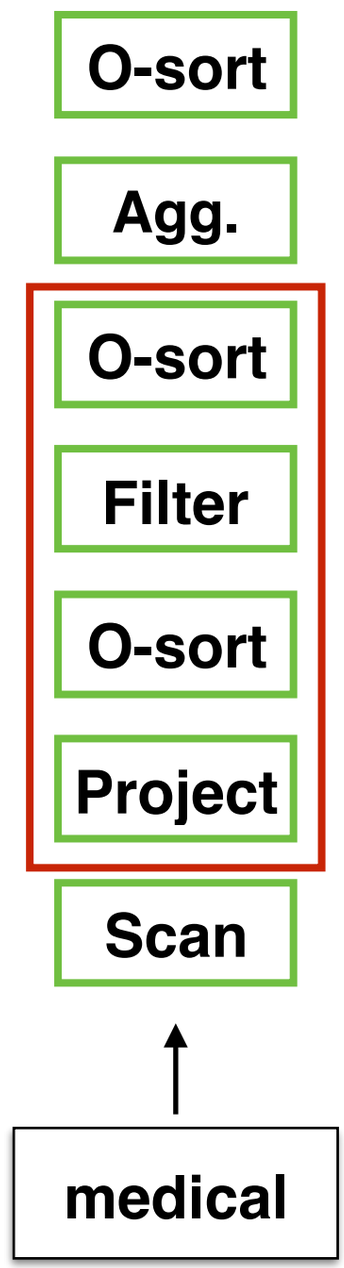


Rule-based optimization

Logical op.



Opaque op.

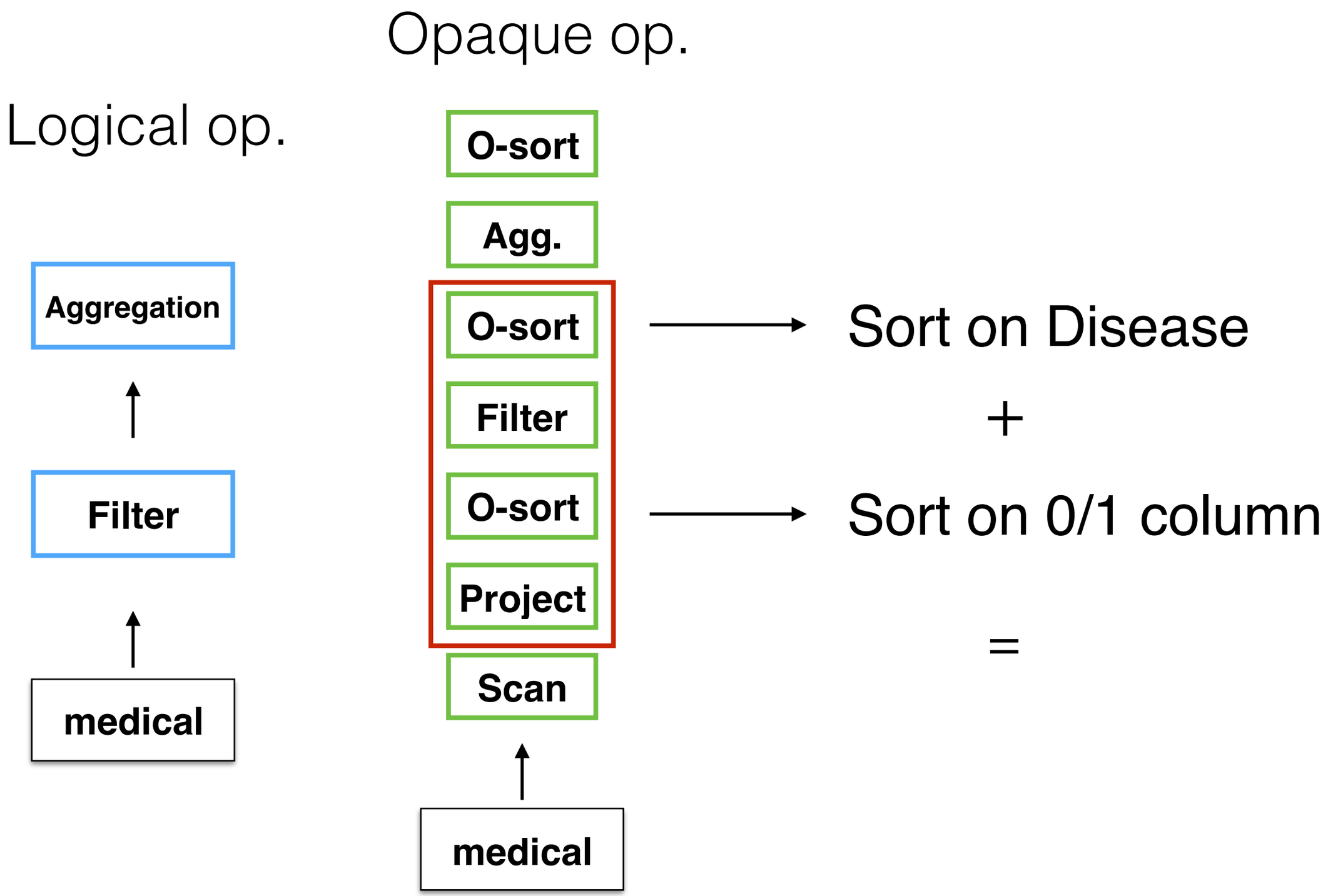


Sort on Disease

+

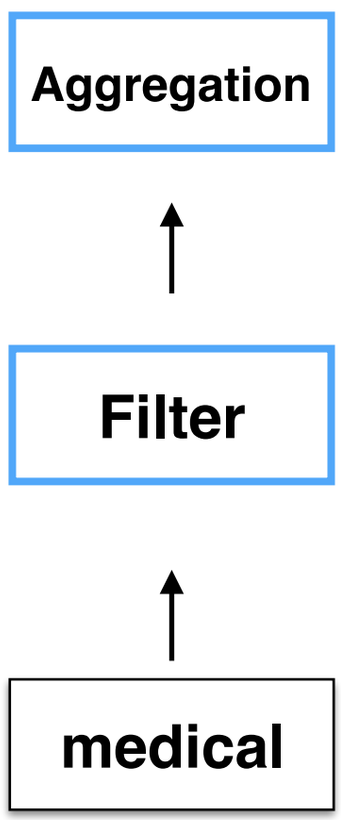
Sort on 0/1 column

Rule-based optimization

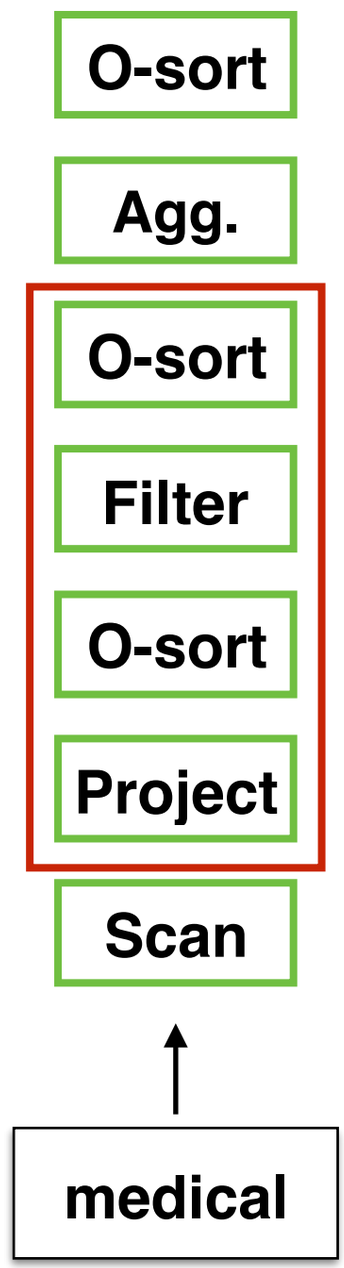


Rule-based optimization

Logical op.



Opaque op.



Sort on Disease

+

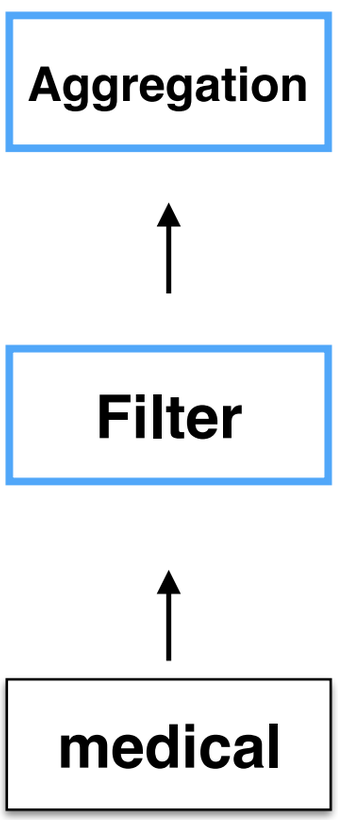
Sort on 0/1 column

=

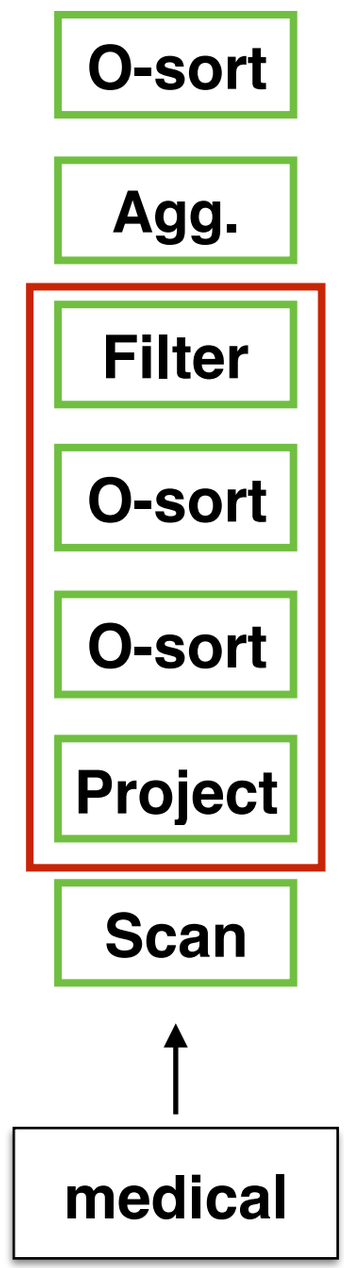
Sort on (0/1, Disease)

Rule-based optimization

Logical op.

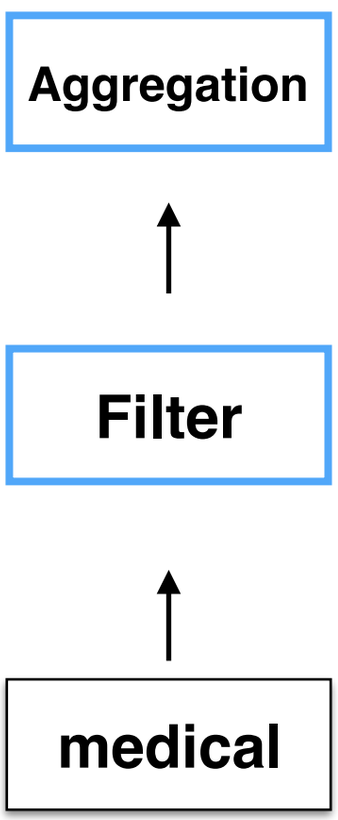


Opaque op.

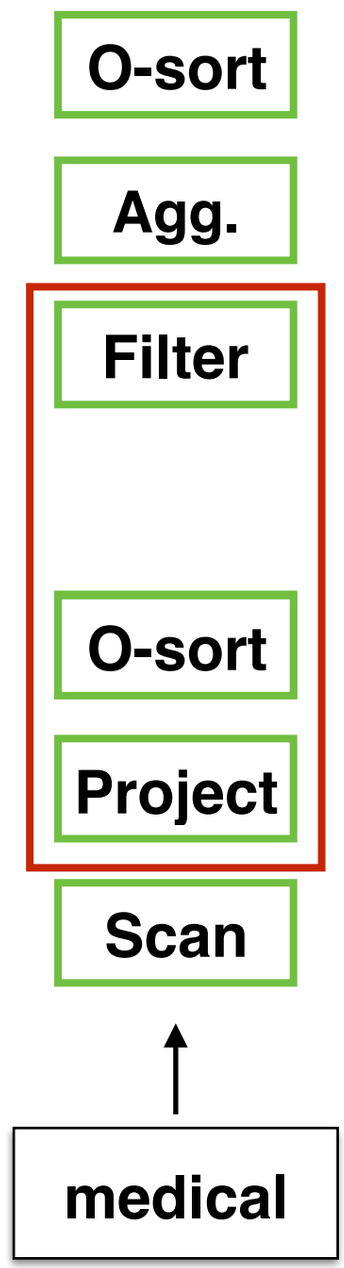


Rule-based optimization

Logical op.

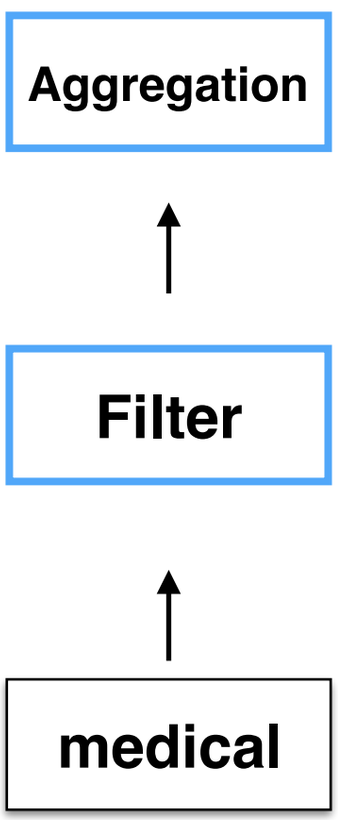


Opaque op.

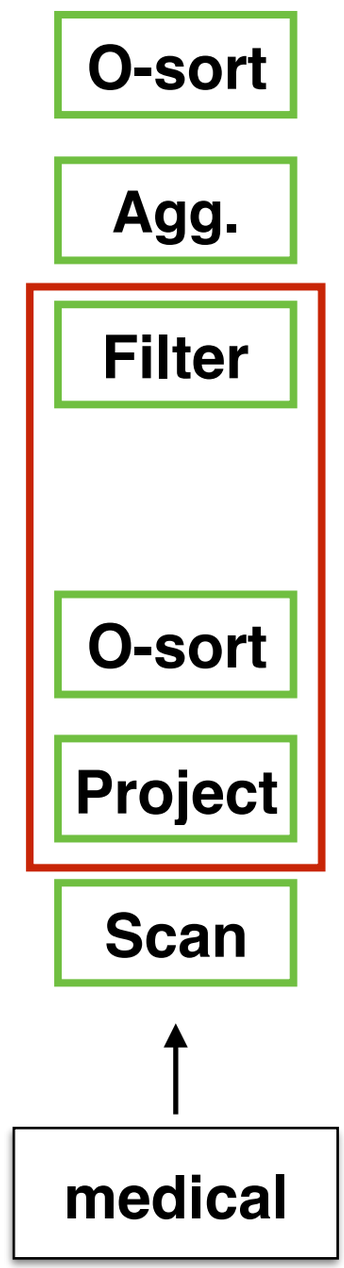


Rule-based optimization

Logical op.

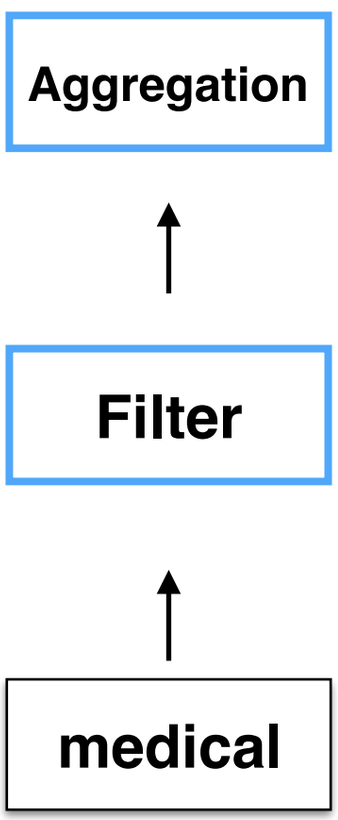


Opaque op.

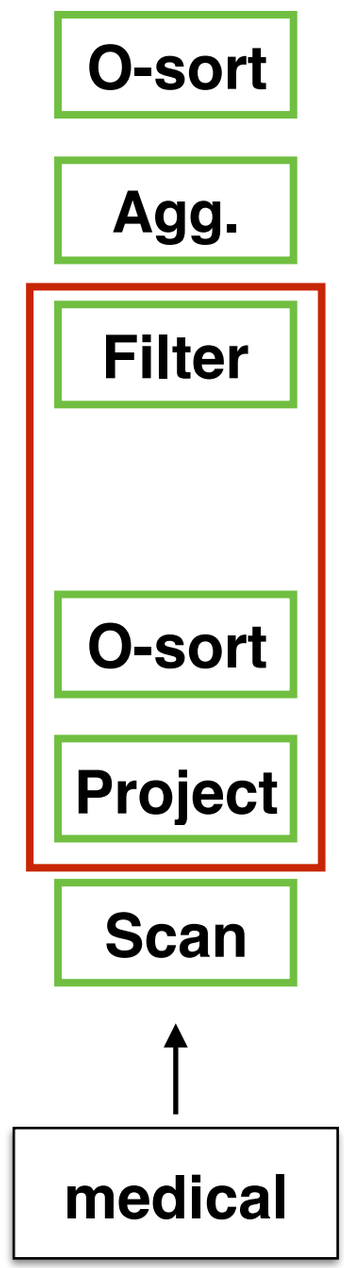


Rule-based optimization

Logical op.

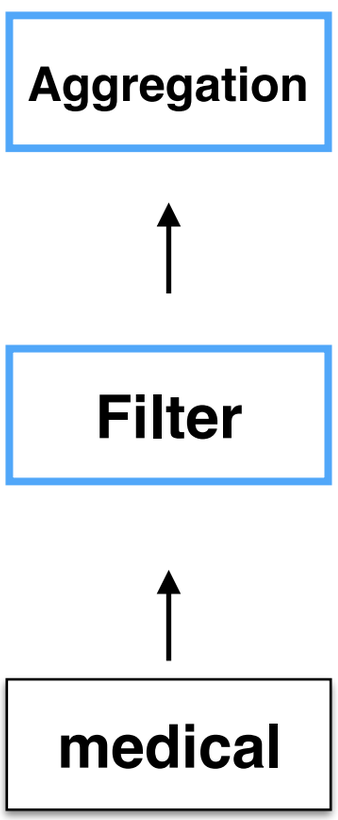


Opaque op.

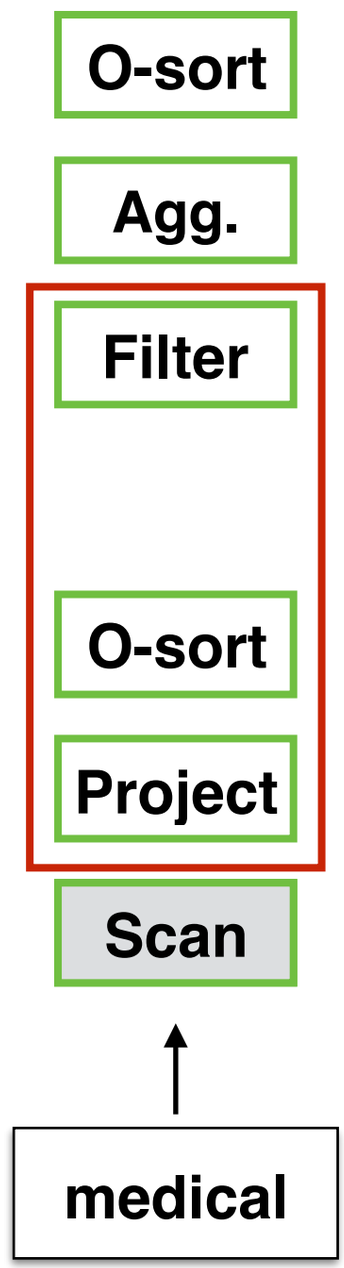


Rule-based optimization

Logical op.

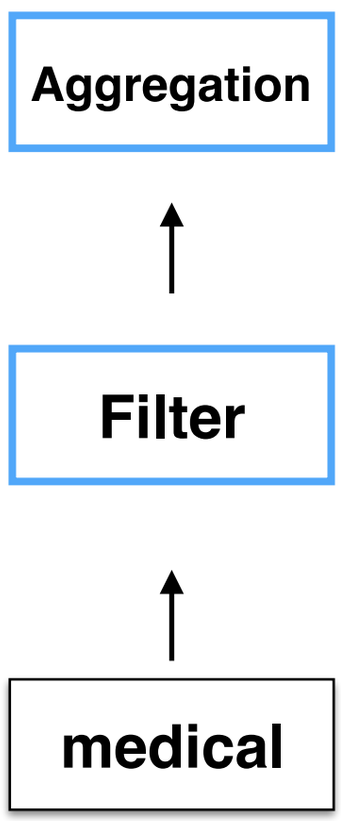


Opaque op.

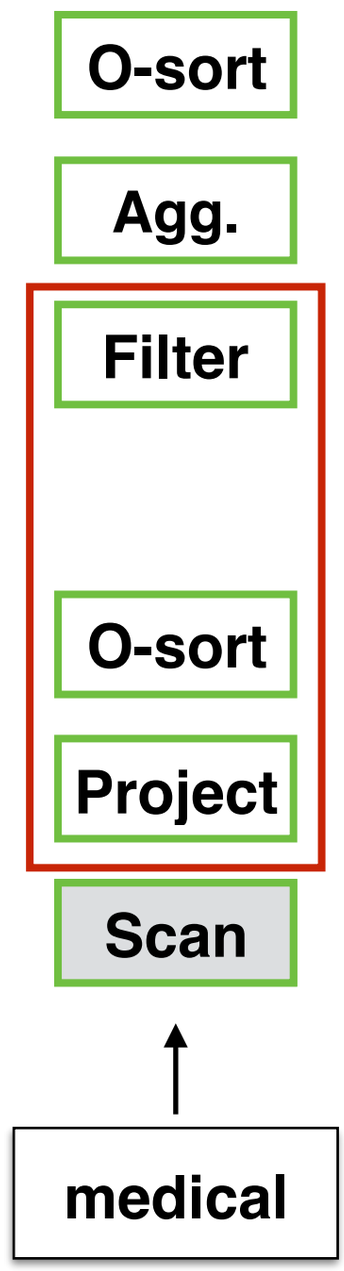


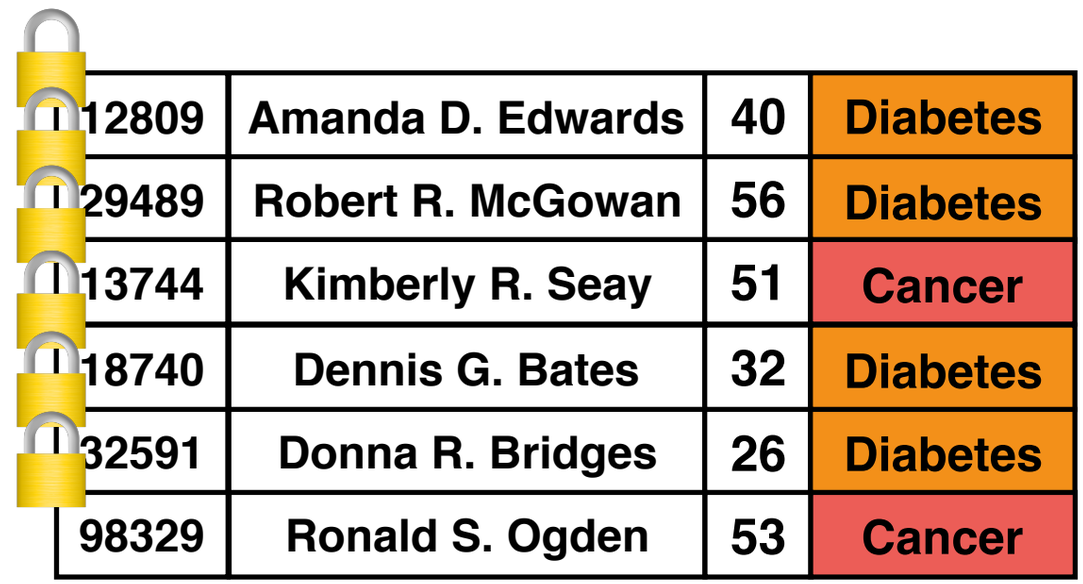
Rule-based optimization

Logical op.



Opaque op.

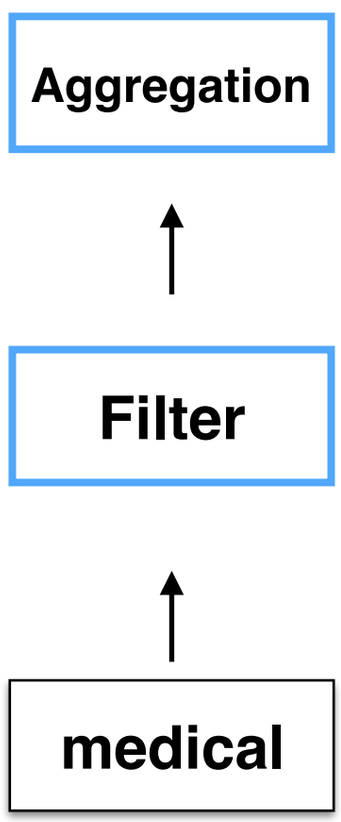




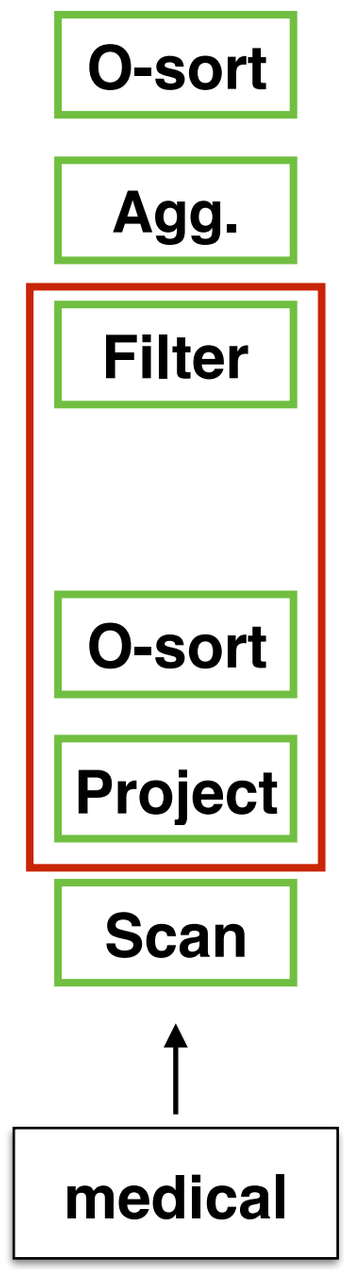
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

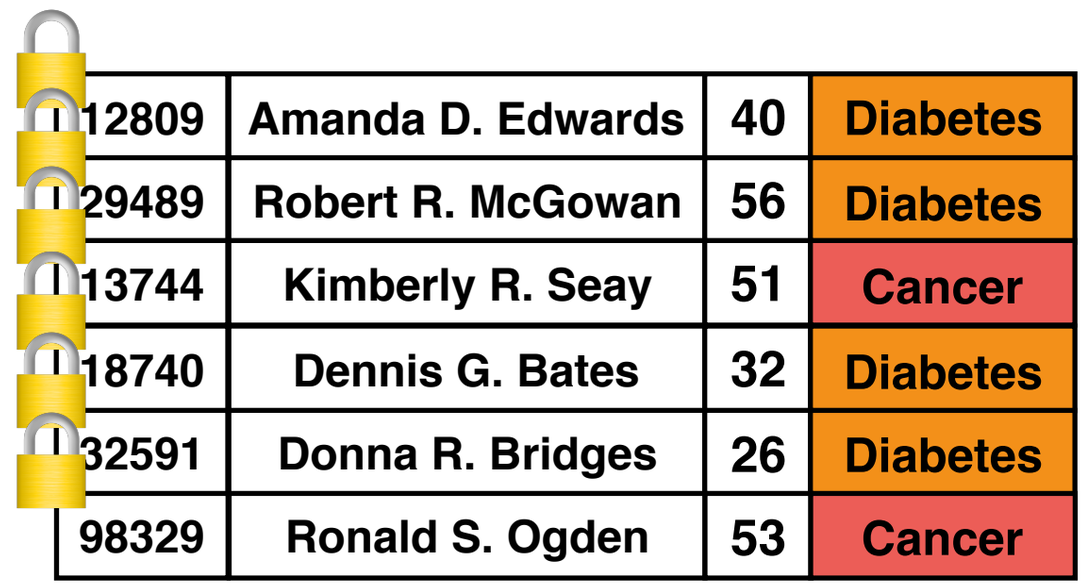
Rule-based optimization

Logical op.



Opaque op.

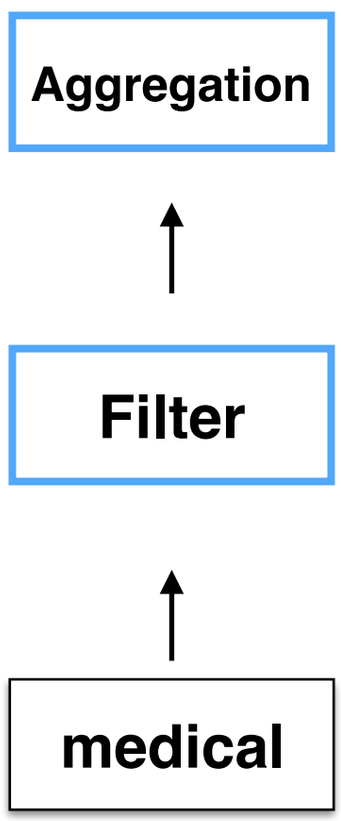




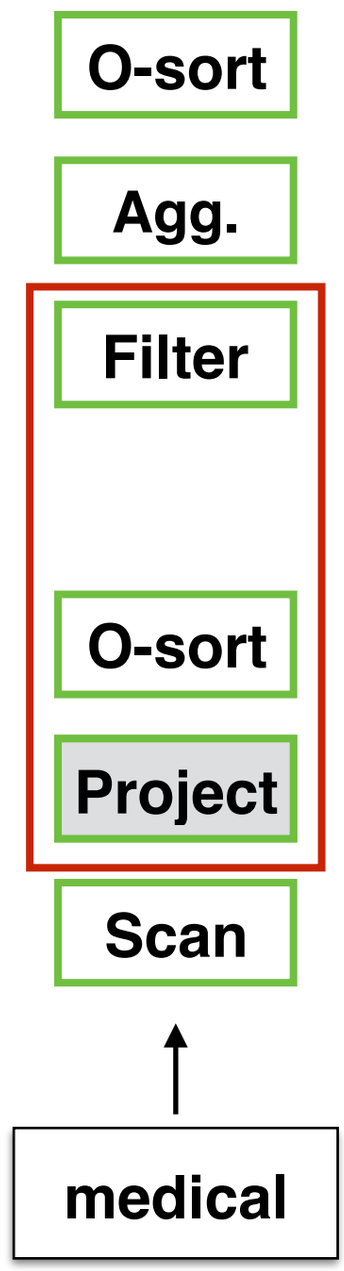
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

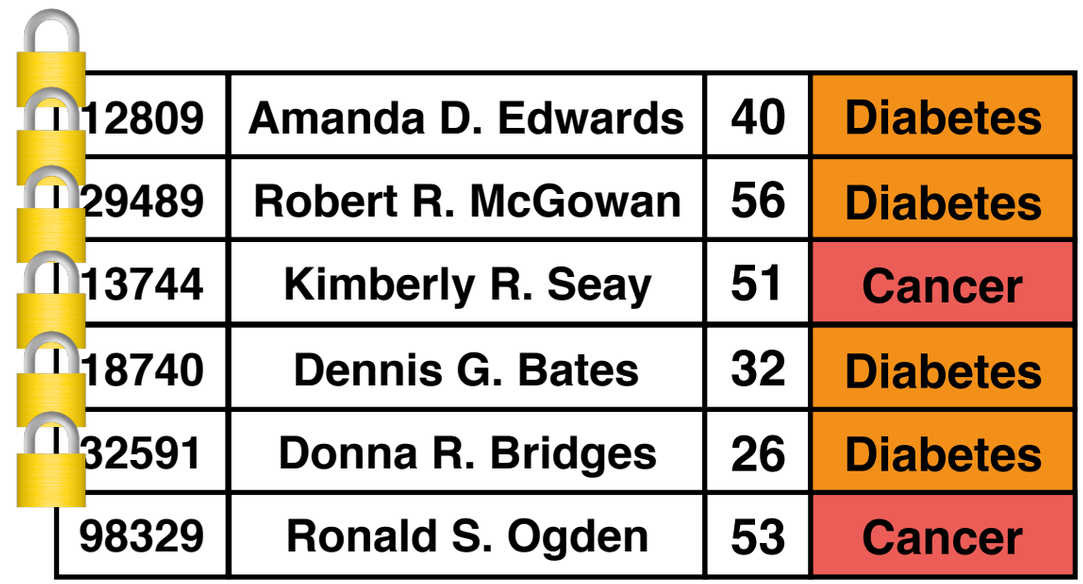
Rule-based optimization

Logical op.



Opaque op.

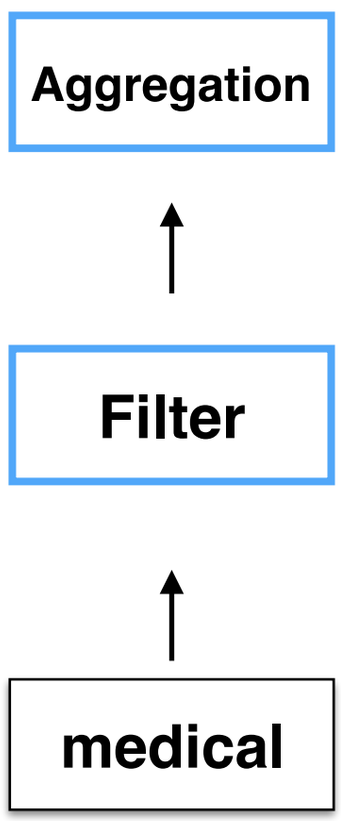




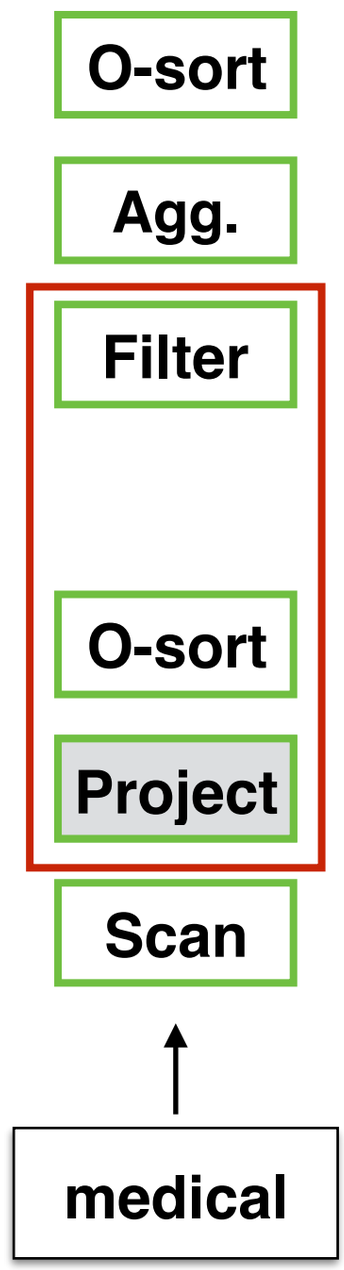
12809	Amanda D. Edwards	40	Diabetes
29489	Robert R. McGowan	56	Diabetes
13744	Kimberly R. Seay	51	Cancer
18740	Dennis G. Bates	32	Diabetes
32591	Donna R. Bridges	26	Diabetes
98329	Ronald S. Ogden	53	Cancer

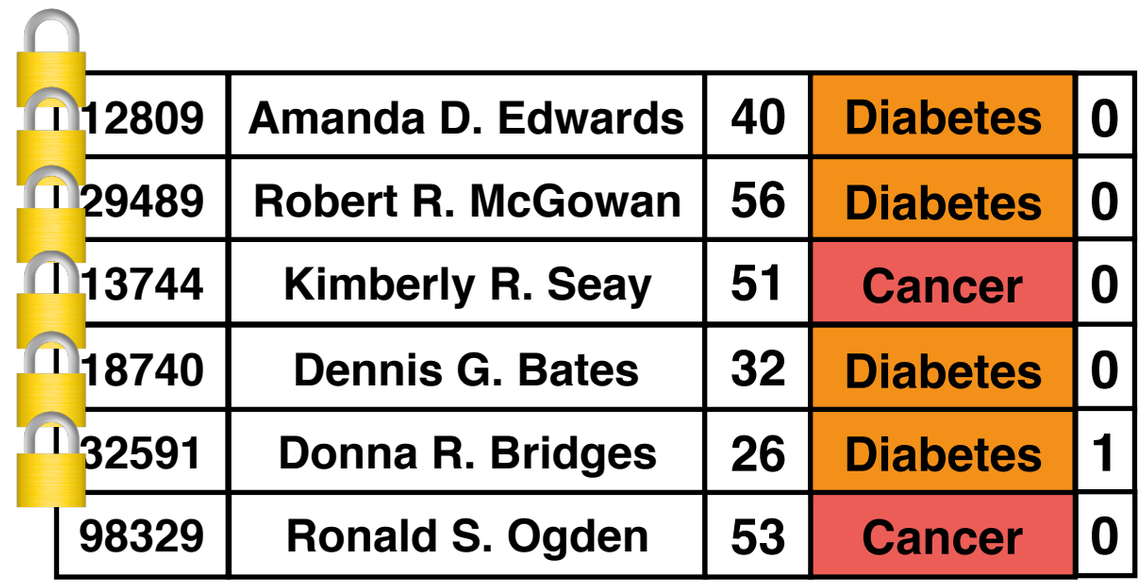
Rule-based optimization

Logical op.



Opaque op.

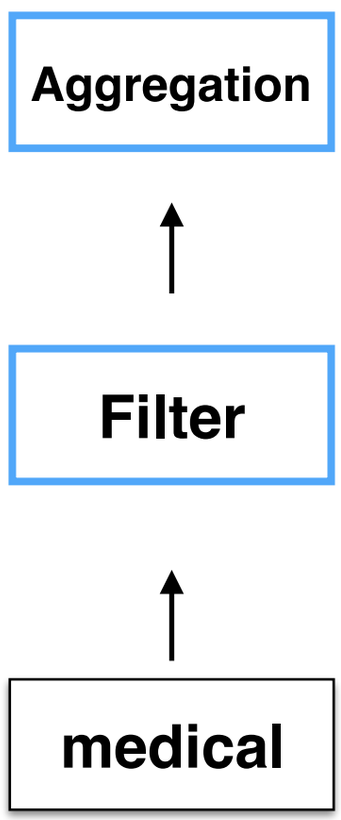




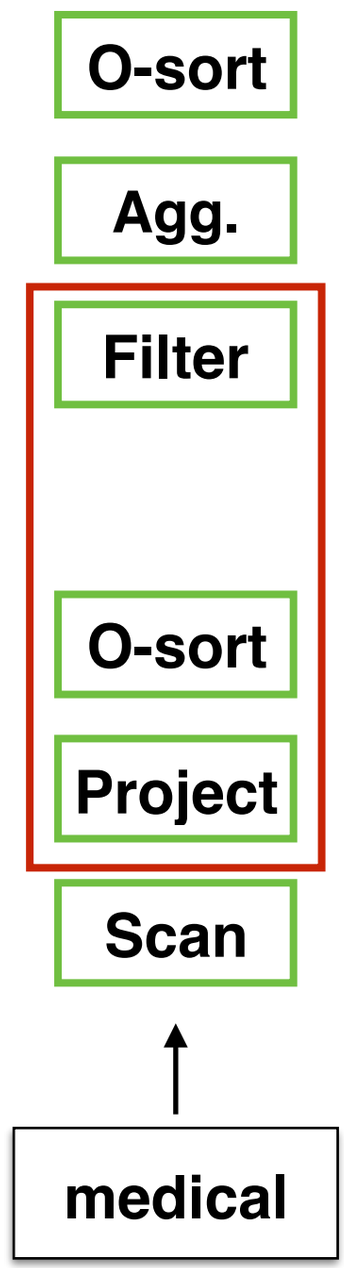
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

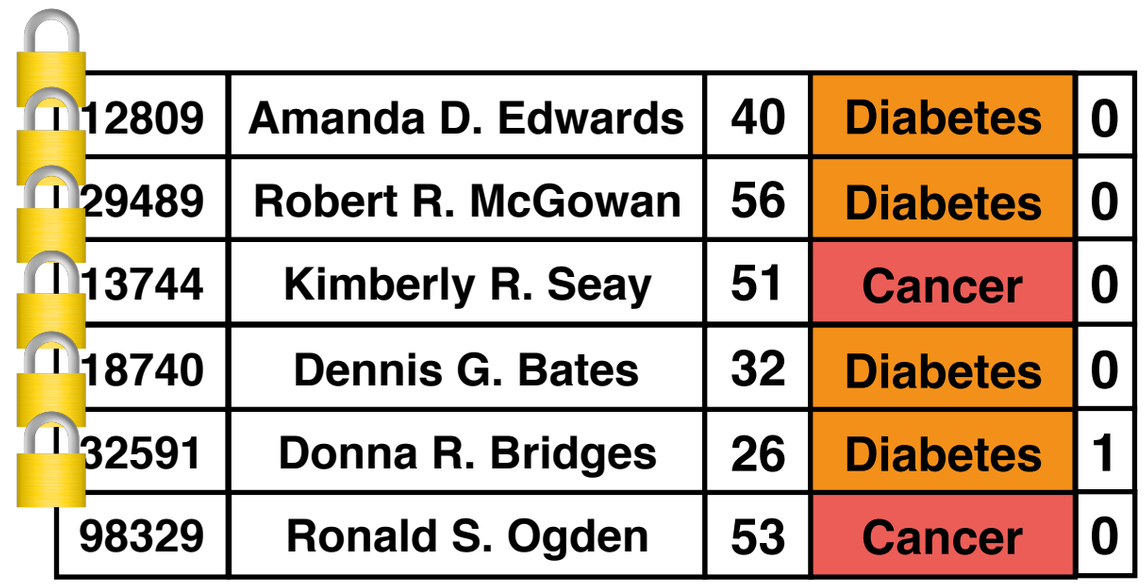
Rule-based optimization

Logical op.



Opaque op.

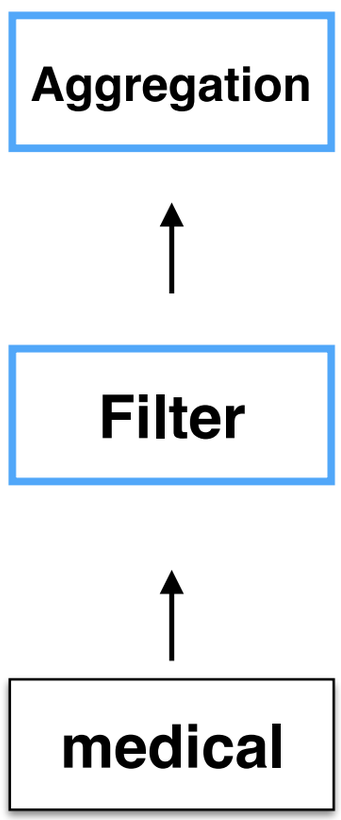




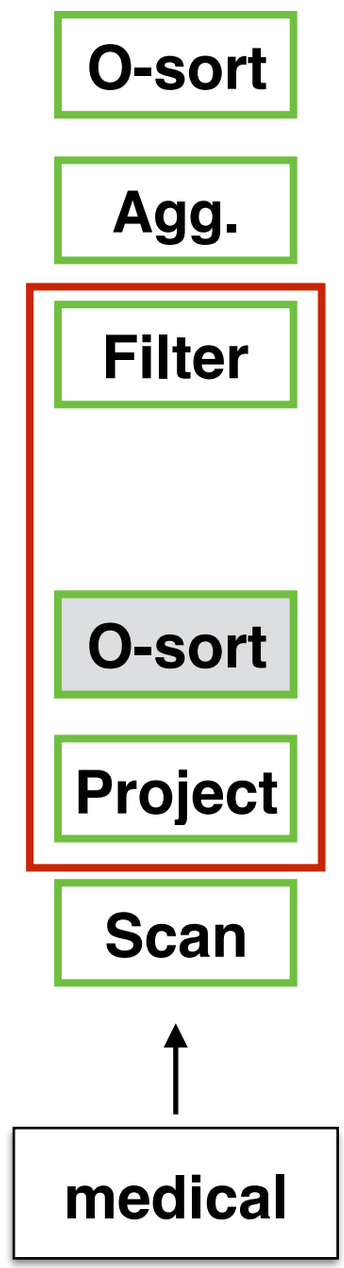
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

Rule-based optimization

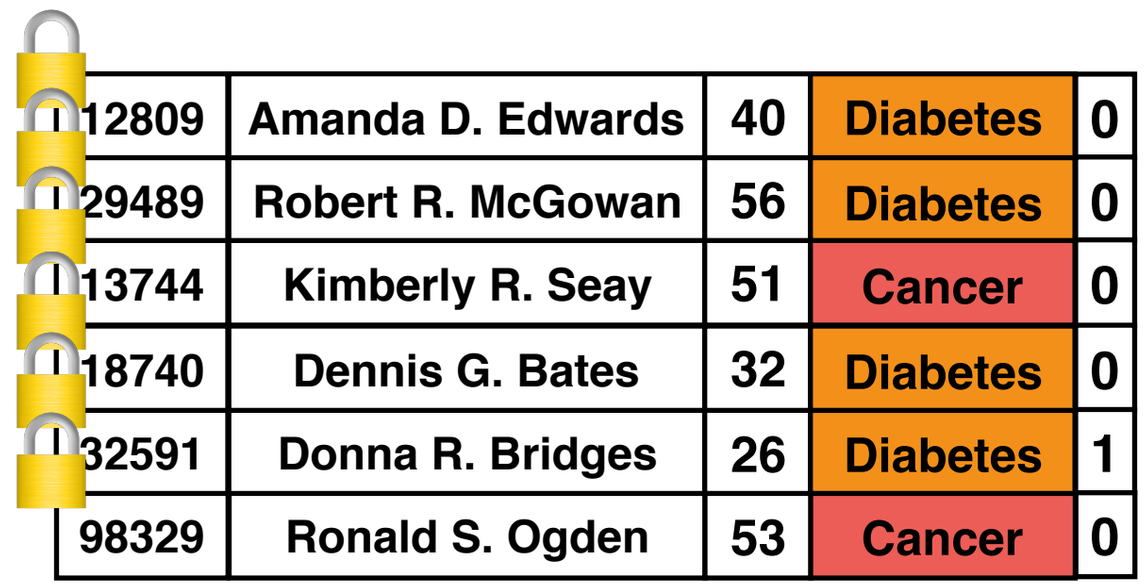
Logical op.



Opaque op.



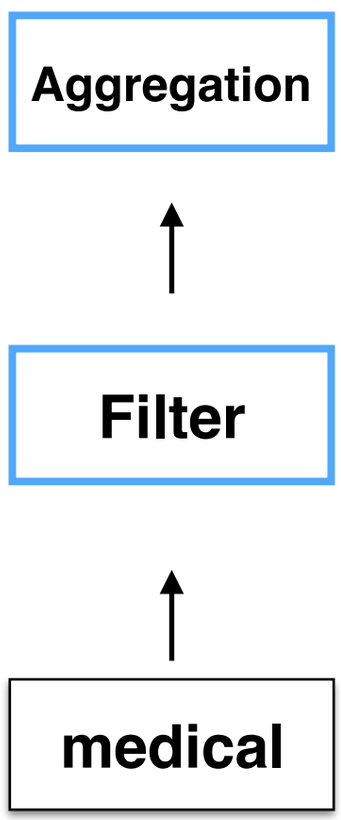
multi-column sort



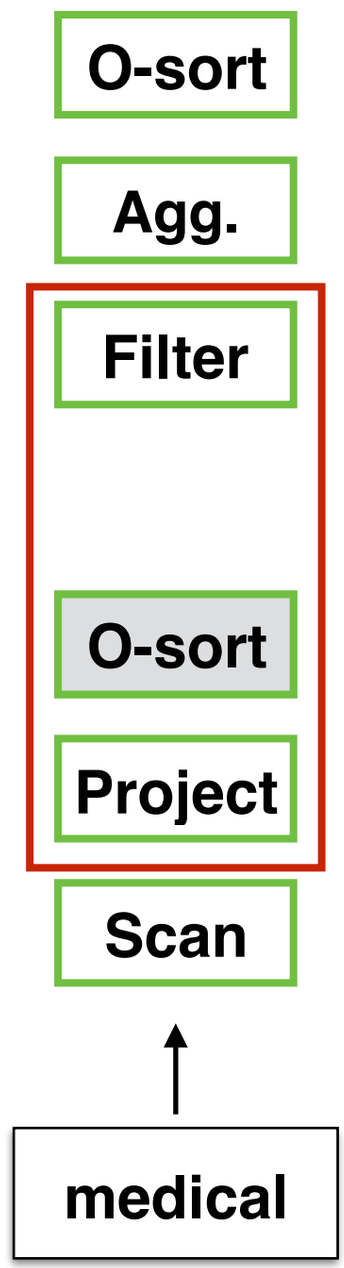
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

Rule-based optimization

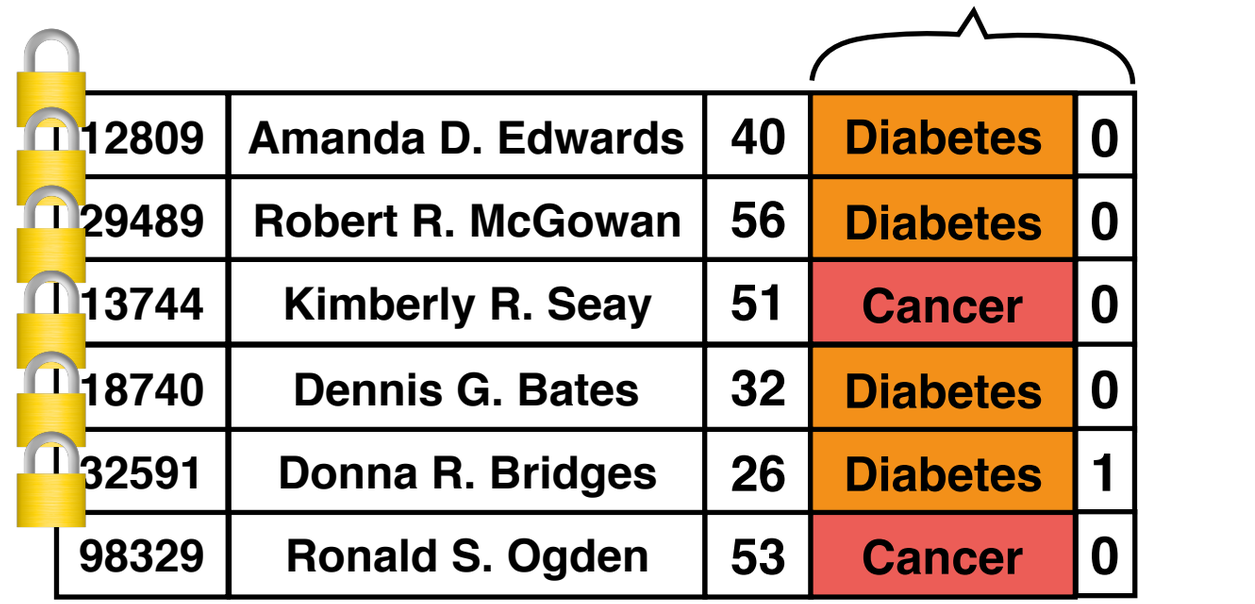
Logical op.



Opaque op.



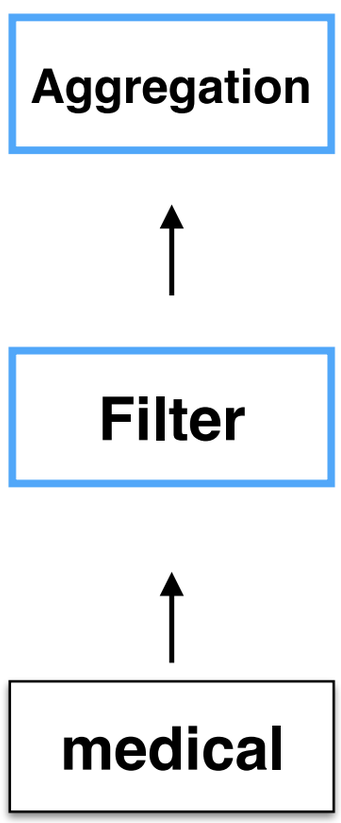
multi-column sort



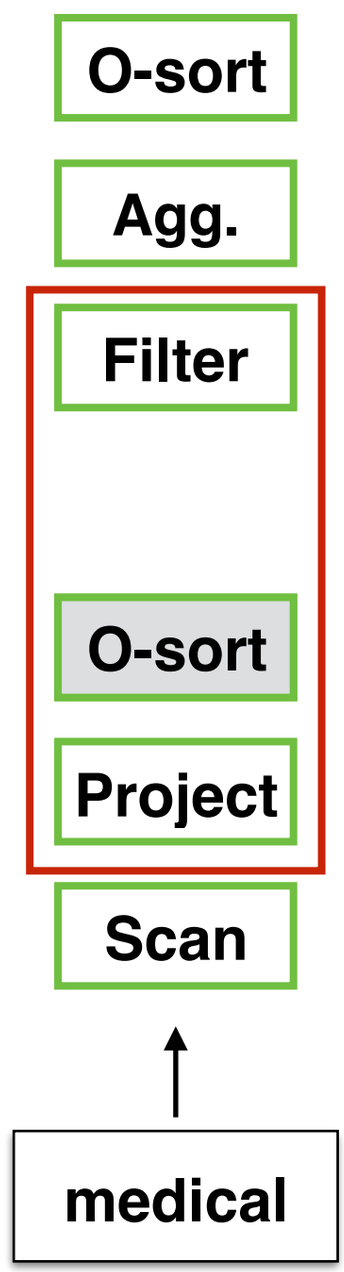
12809	Amanda D. Edwards	40	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
13744	Kimberly R. Seay	51	Cancer	0
18740	Dennis G. Bates	32	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1
98329	Ronald S. Ogden	53	Cancer	0

Rule-based optimization

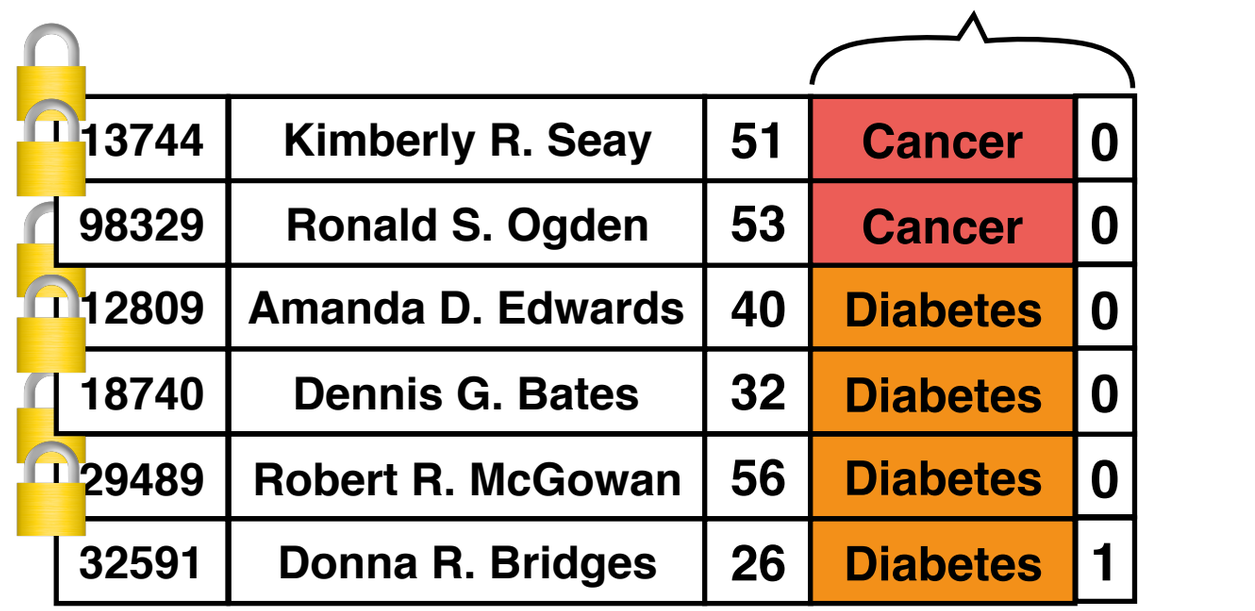
Logical op.



Opaque op.



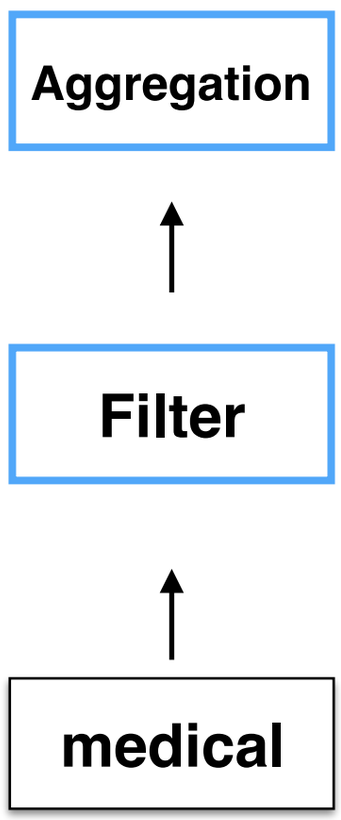
multi-column sort



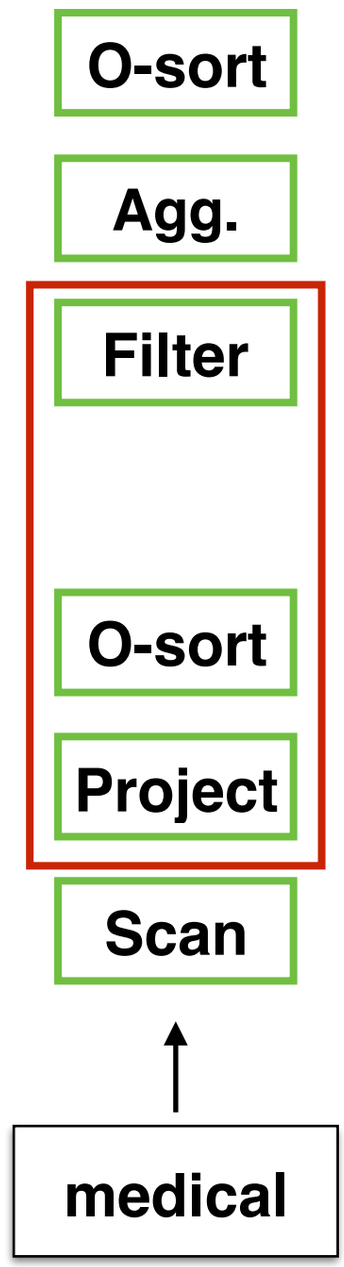
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1

Rule-based optimization

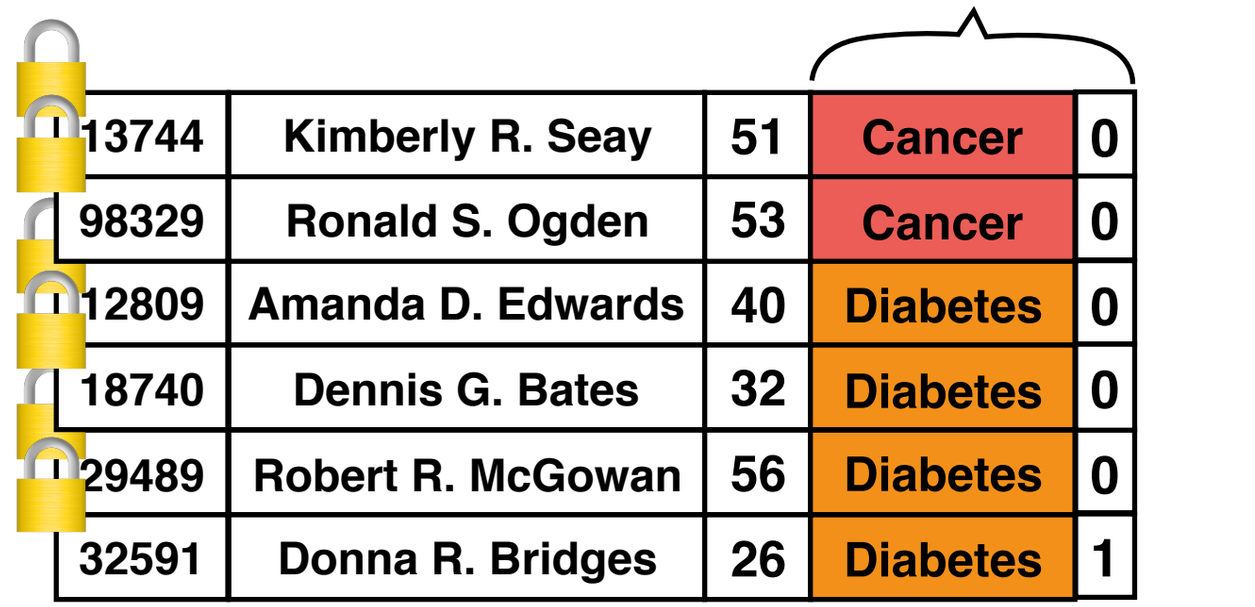
Logical op.



Opaque op.



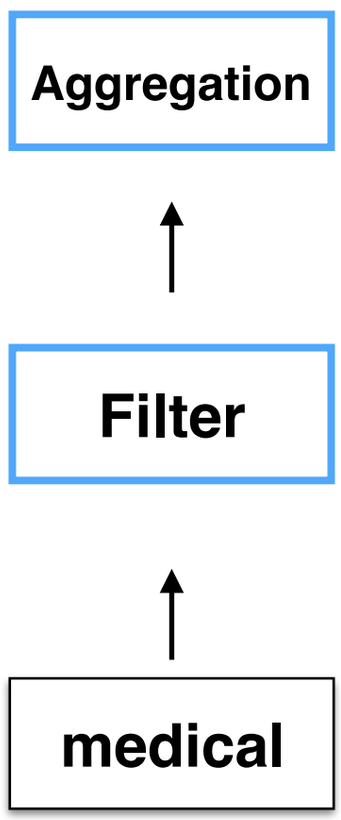
multi-column sort



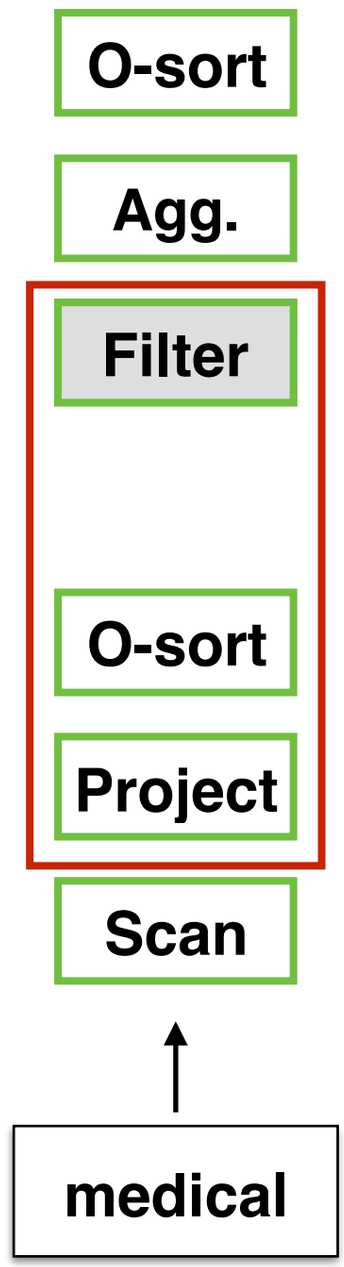
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1

Rule-based optimization

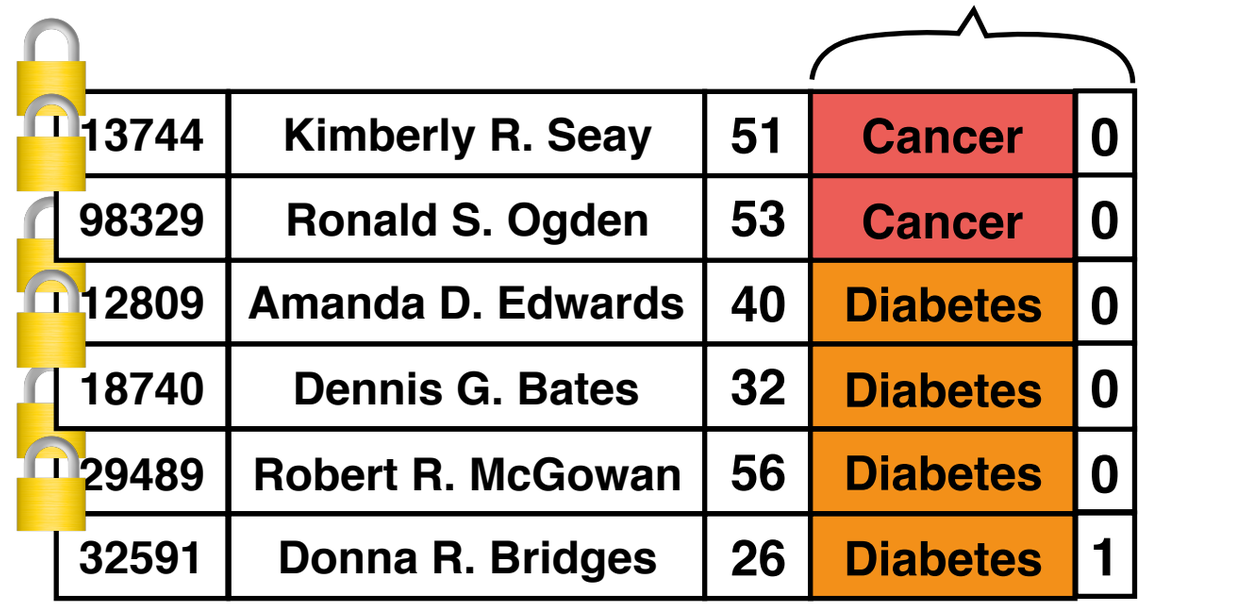
Logical op.



Opaque op.



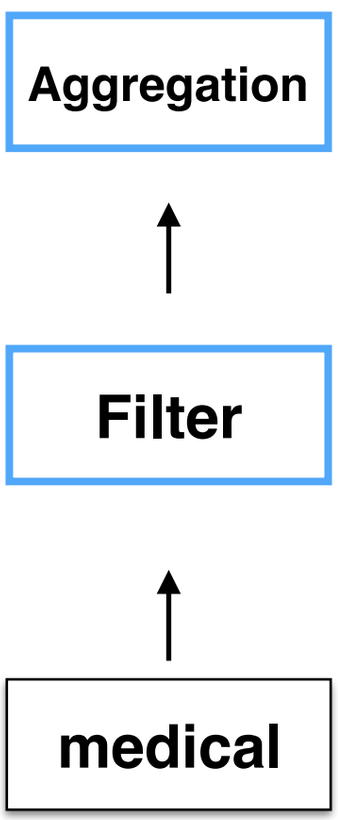
multi-column sort



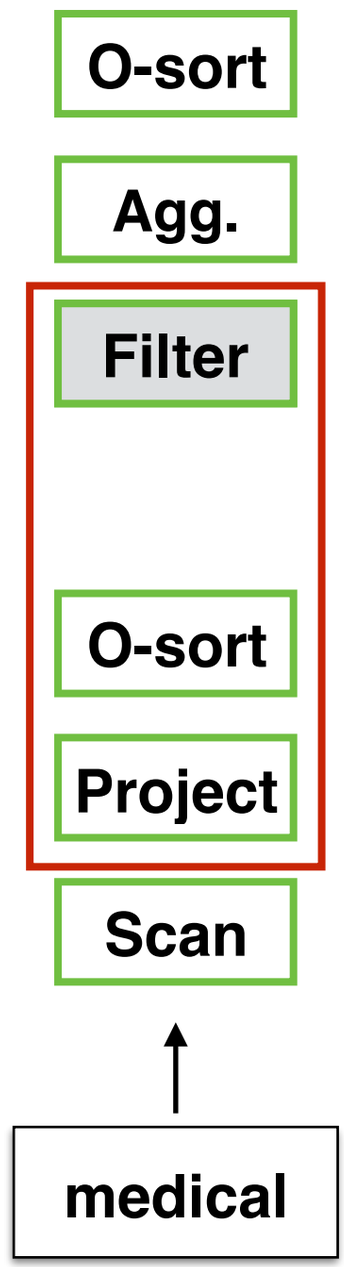
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0
32591	Donna R. Bridges	26	Diabetes	1

Rule-based optimization

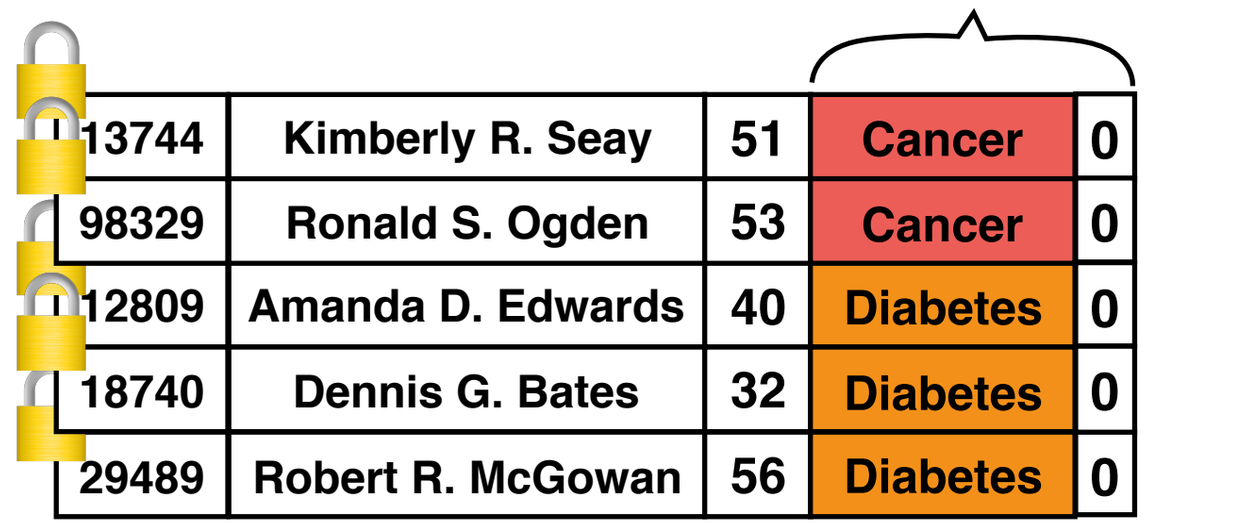
Logical op.



Opaque op.



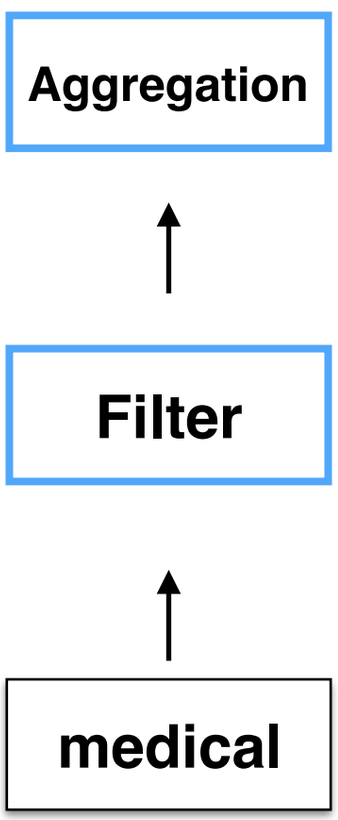
multi-column sort



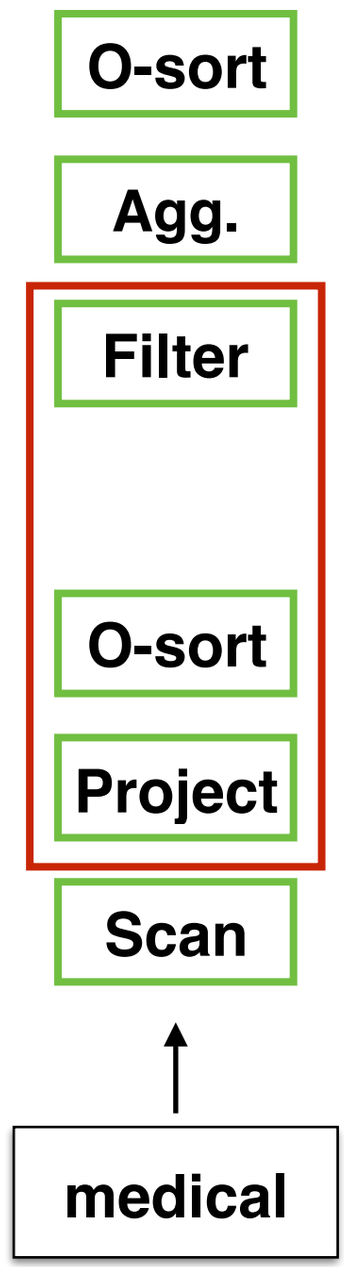
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0

Rule-based optimization

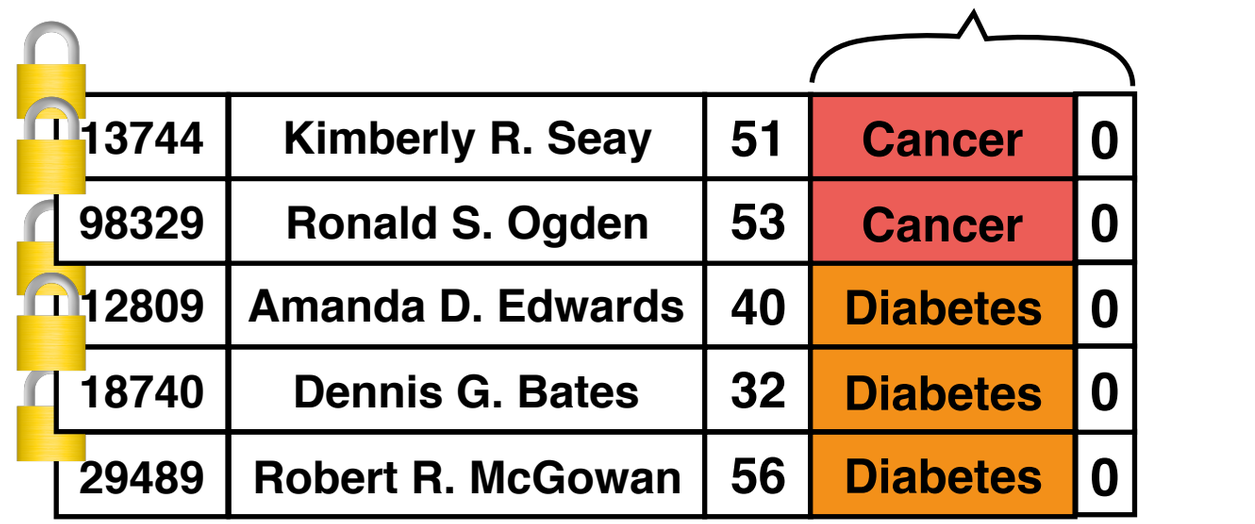
Logical op.



Opaque op.



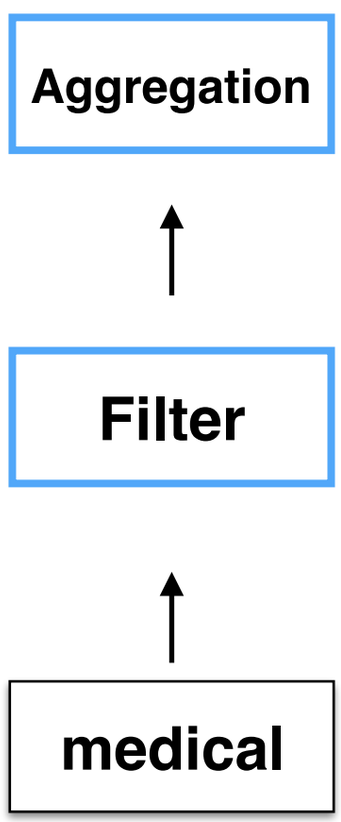
multi-column sort



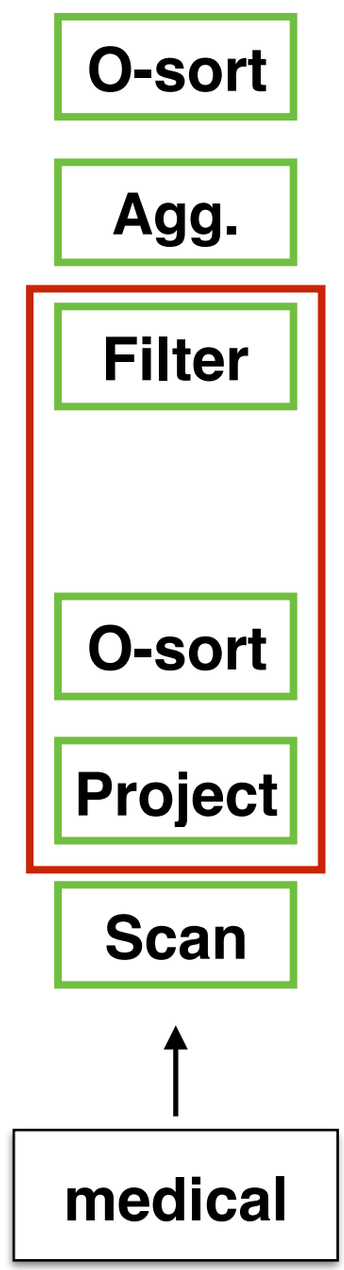
13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0

Rule-based optimization

Logical op.



Opaque op.



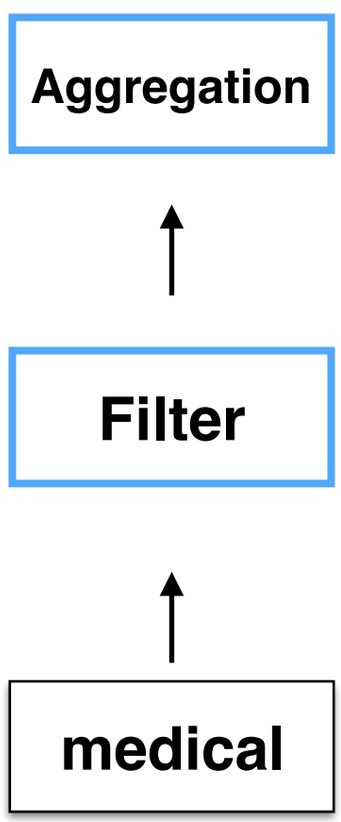
multi-column sort

13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0

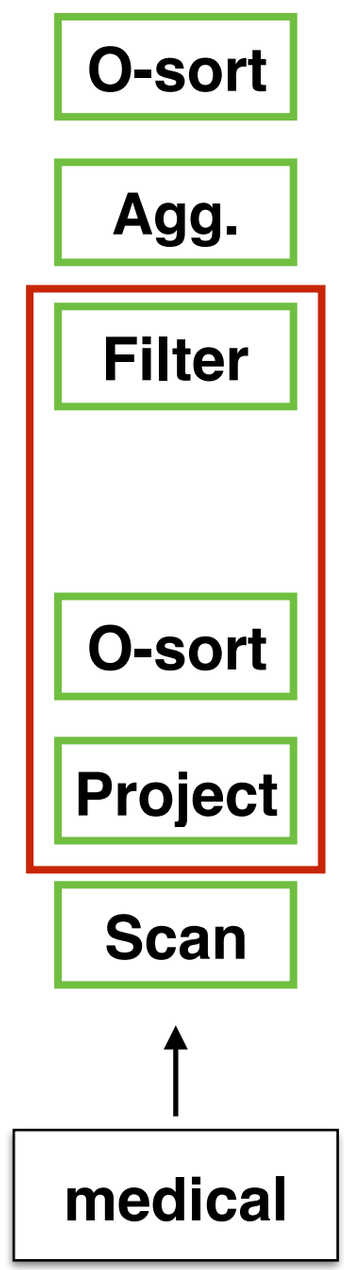
Eliminated one oblivious sort!

Rule-based optimization

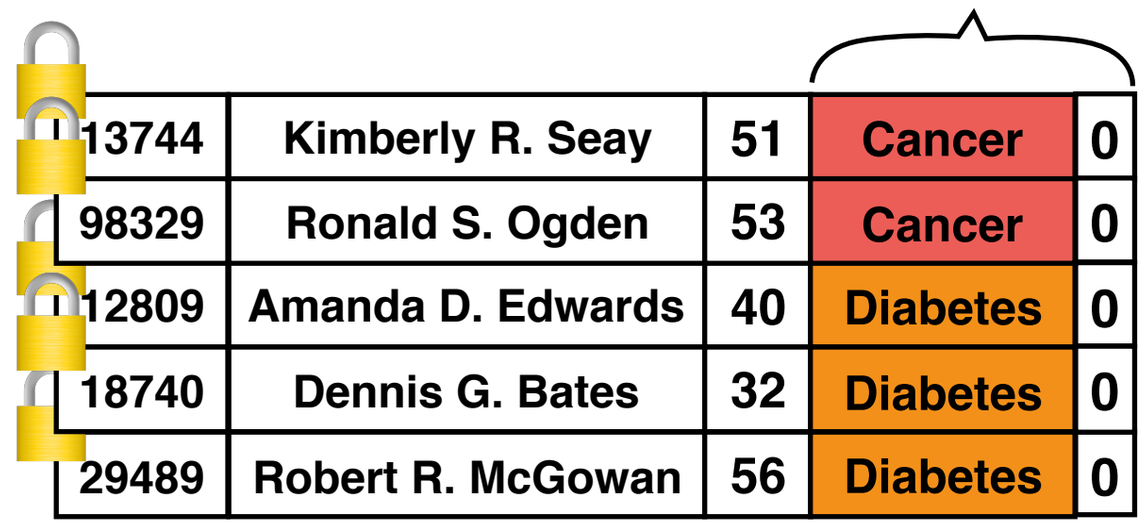
Logical op.



Opaque op.



multi-column sort

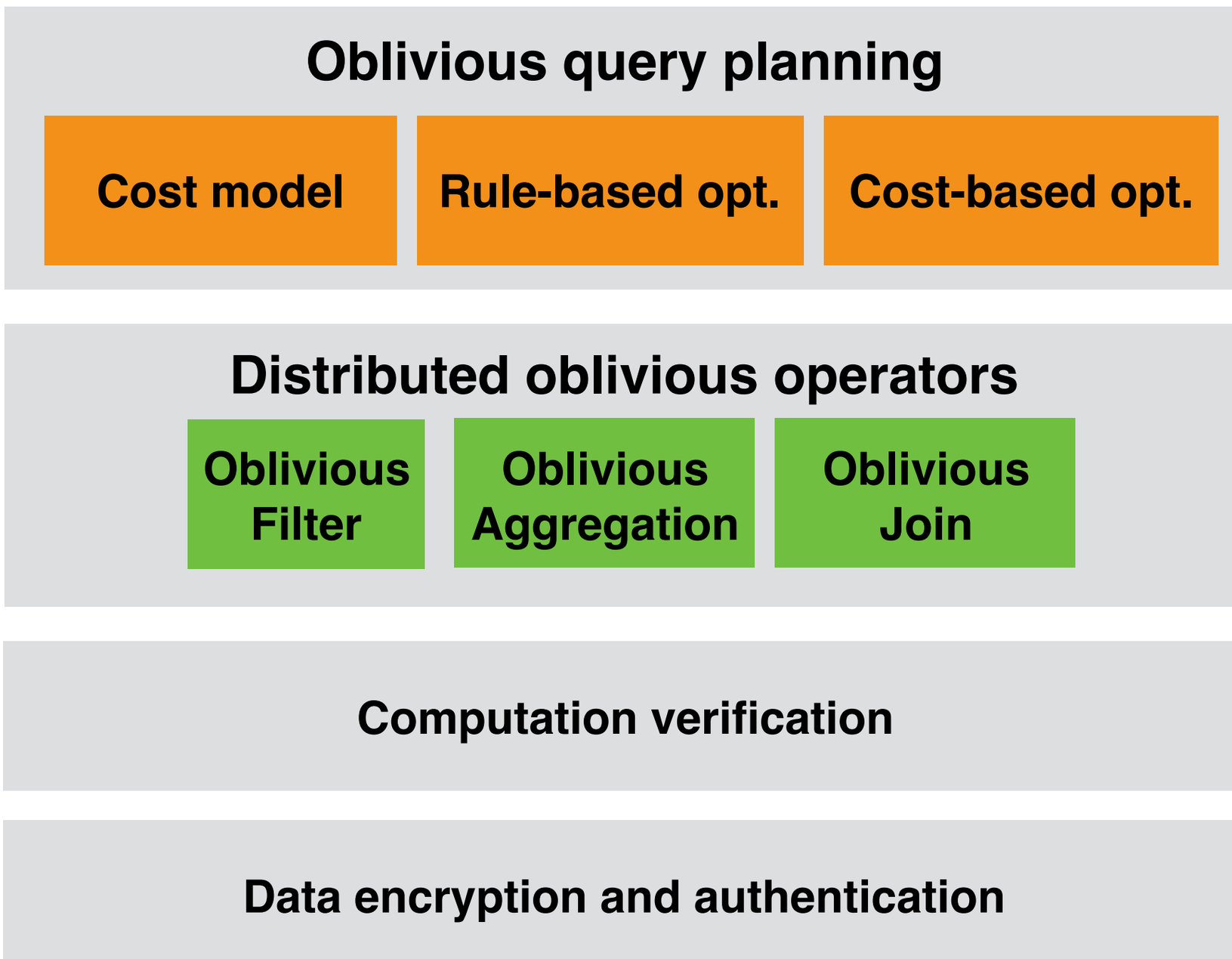


13744	Kimberly R. Seay	51	Cancer	0
98329	Ronald S. Ogden	53	Cancer	0
12809	Amanda D. Edwards	40	Diabetes	0
18740	Dennis G. Bates	32	Diabetes	0
29489	Robert R. McGowan	56	Diabetes	0

Eliminated one oblivious sort!

No change to the Spark SQL planner!

More in the paper..



Evaluation setup

Evaluation setup

- Single machine experiments:
  - Intel Xeon E3-1280 v5, 4 cores, 64 GB RAM
  - Intel SGX: 128 MB of enclave page cache (EPC)

Evaluation setup

- Single machine experiments:
  - Intel Xeon E3-1280 v5, 4 cores, 64 GB RAM
  - Intel SGX: 128 MB of enclave page cache (EPC)
- Distributed experiments
  - A cluster of 5 SGX machines

Evaluation

Evaluation

- **How does Opaque compare to Spark SQL?**

Evaluation

- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total

Evaluation

- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total
    - Queries 1, 2, 3: filter, aggregation, join

Evaluation

- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total
    - Queries 1, 2, 3: filter, aggregation, join
    - 1 million records

Evaluation

- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total
    - Queries 1, 2, 3: filter, aggregation, join
    - 1 million records
- **How does Opaque compare to state-of-the-art oblivious systems?**

Evaluation

- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total
    - Queries 1, 2, 3: filter, aggregation, join
    - 1 million records
- **How does Opaque compare to state-of-the-art oblivious systems?**
  - GraphSC (oblivious graph analytics)

Evaluation

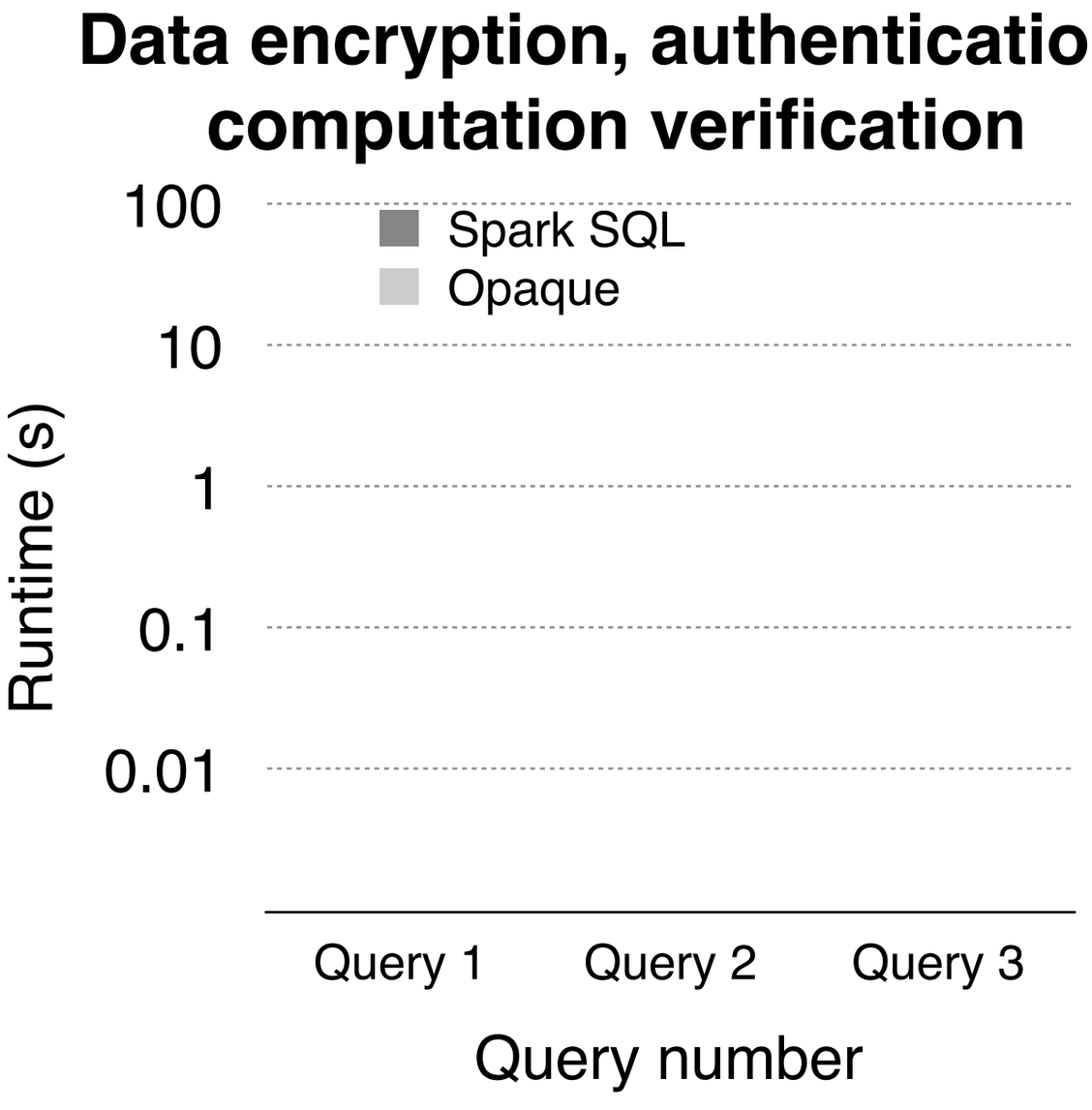
- **How does Opaque compare to Spark SQL?**
  - Big Data Benchmark (BDB); 4 queries total
    - Queries 1, 2, 3: filter, aggregation, join
    - 1 million records
- **How does Opaque compare to state-of-the-art oblivious systems?**
  - GraphSC (oblivious graph analytics)
    - PageRank

Big Data Benchmark (distributed)

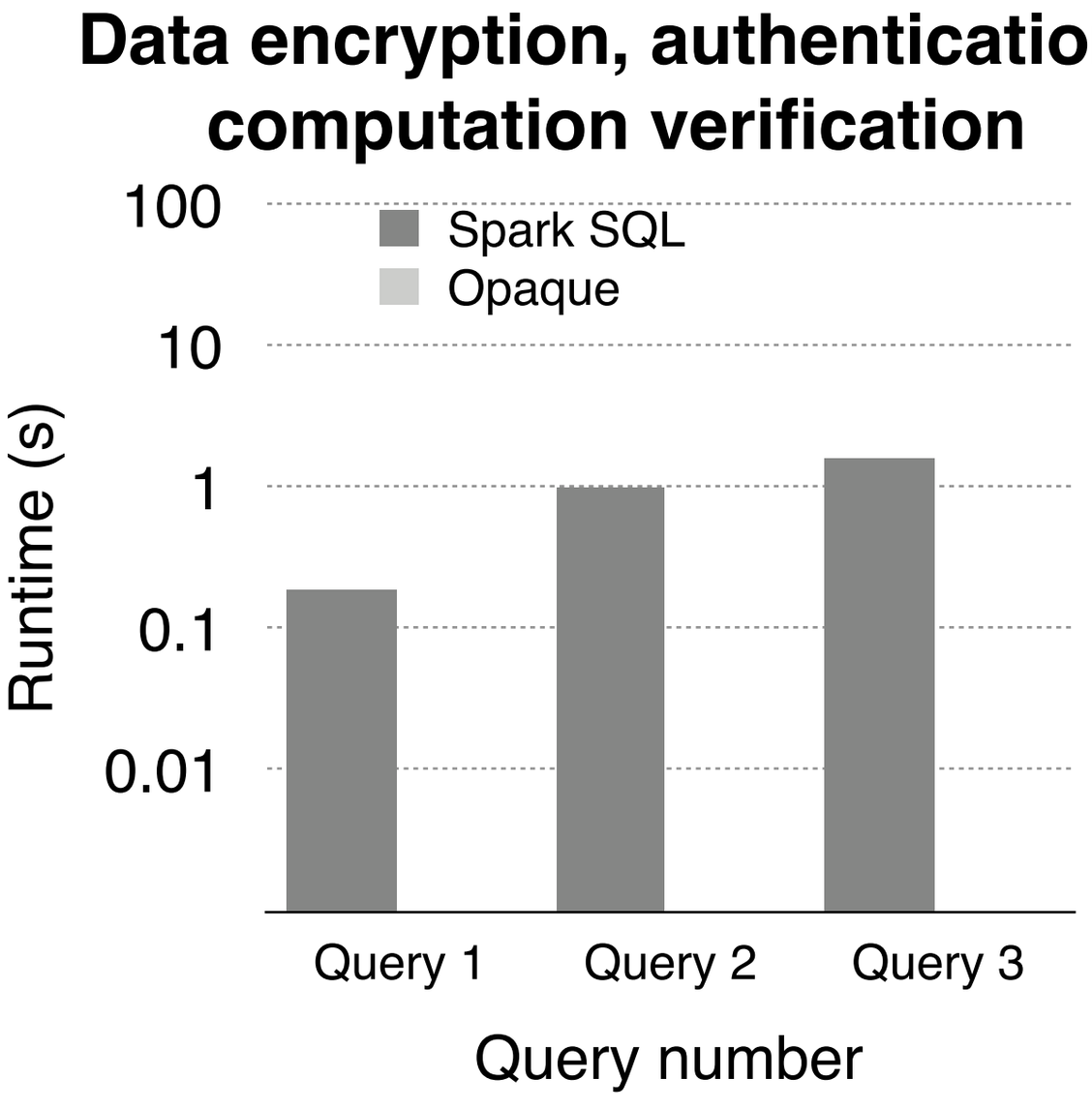
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

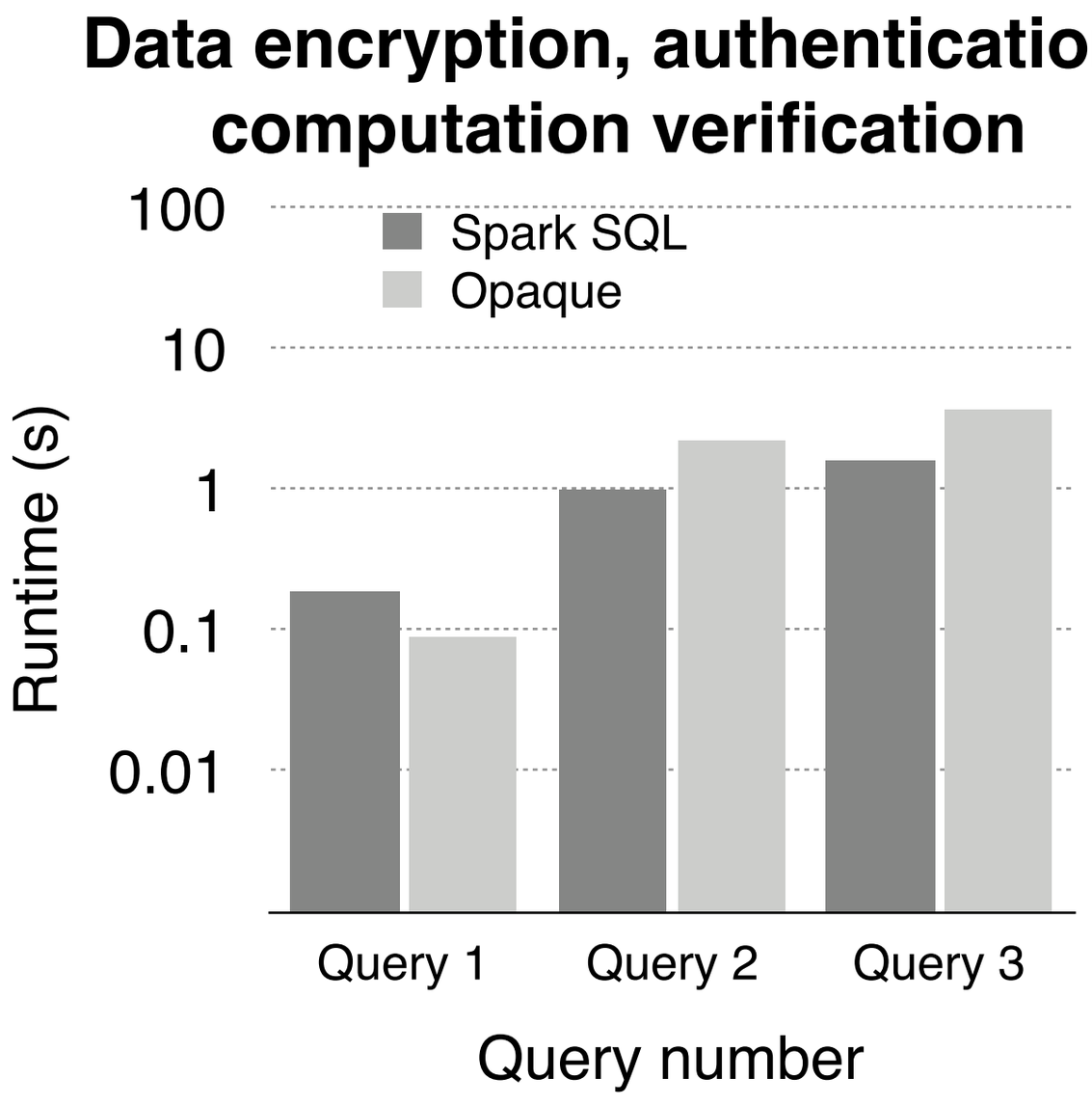
Big Data Benchmark (distributed)



Big Data Benchmark (distributed)

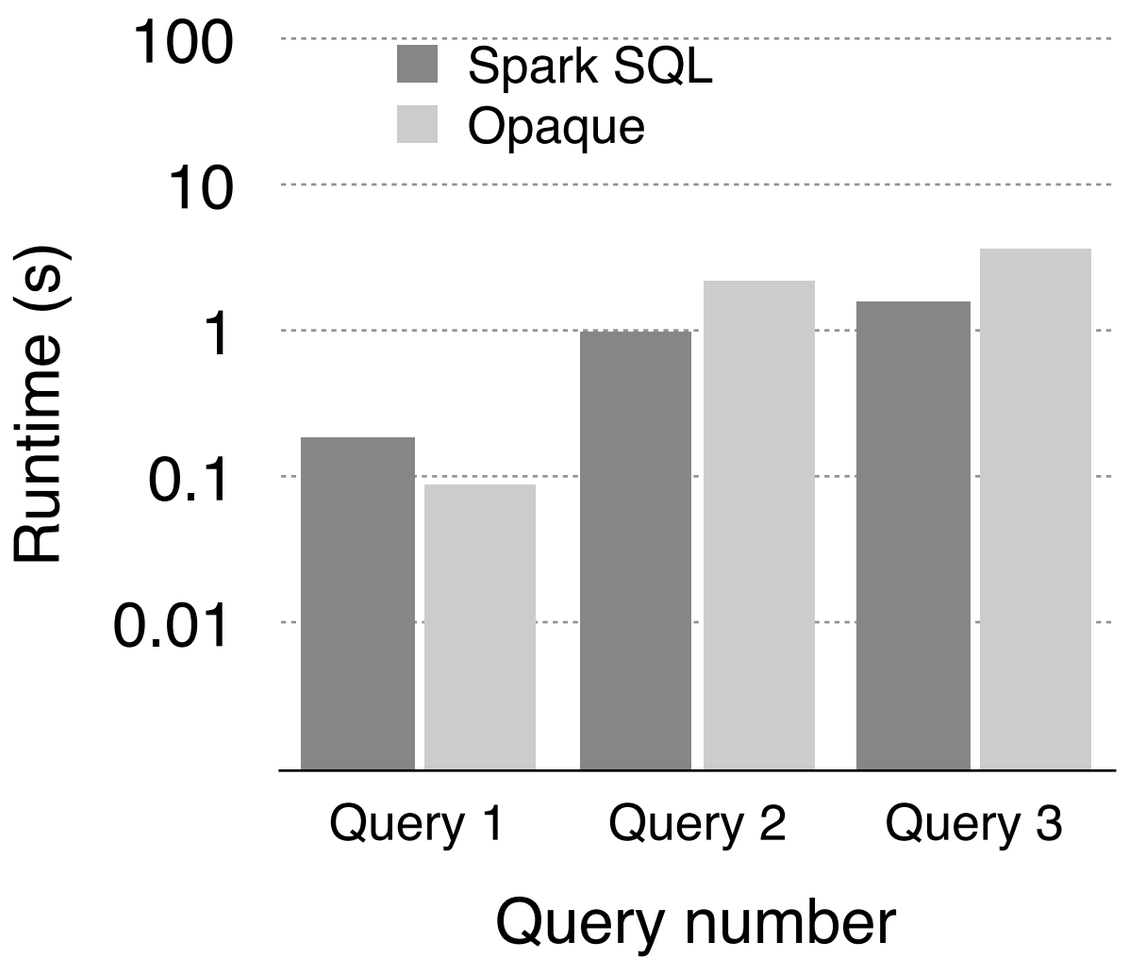


Big Data Benchmark (distributed)



Big Data Benchmark (distributed)

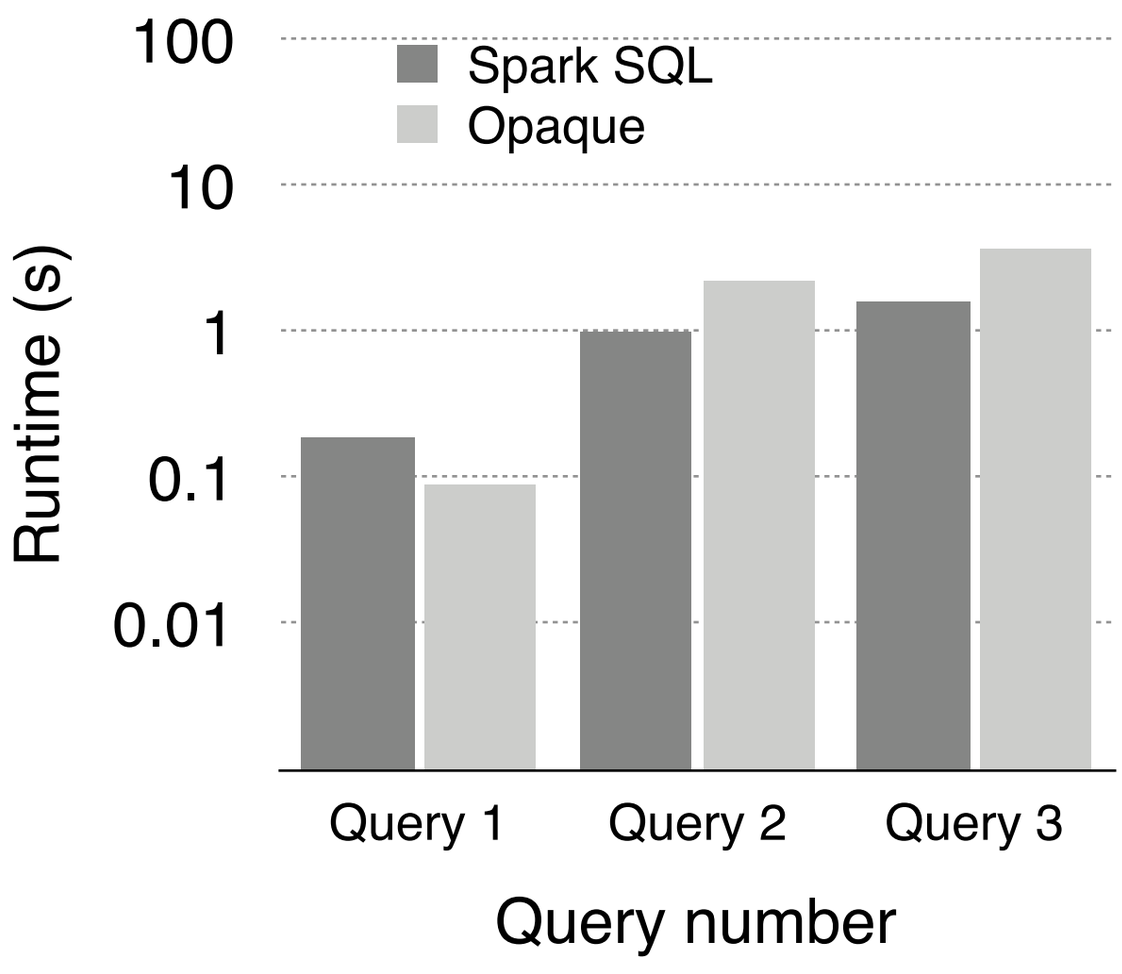
Data encryption, authentication, computation verification



Overhead: -0.47x to 2.3x

Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

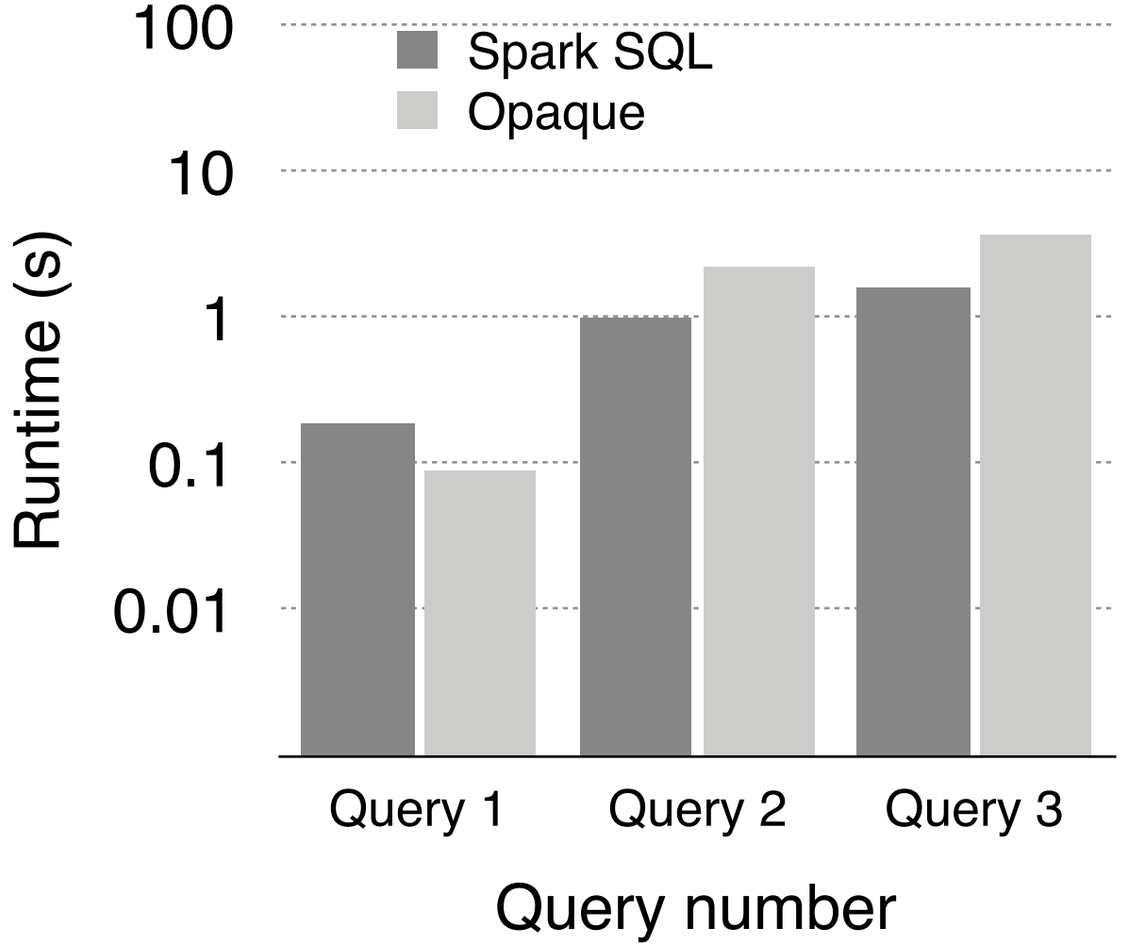


+ Obliviousness

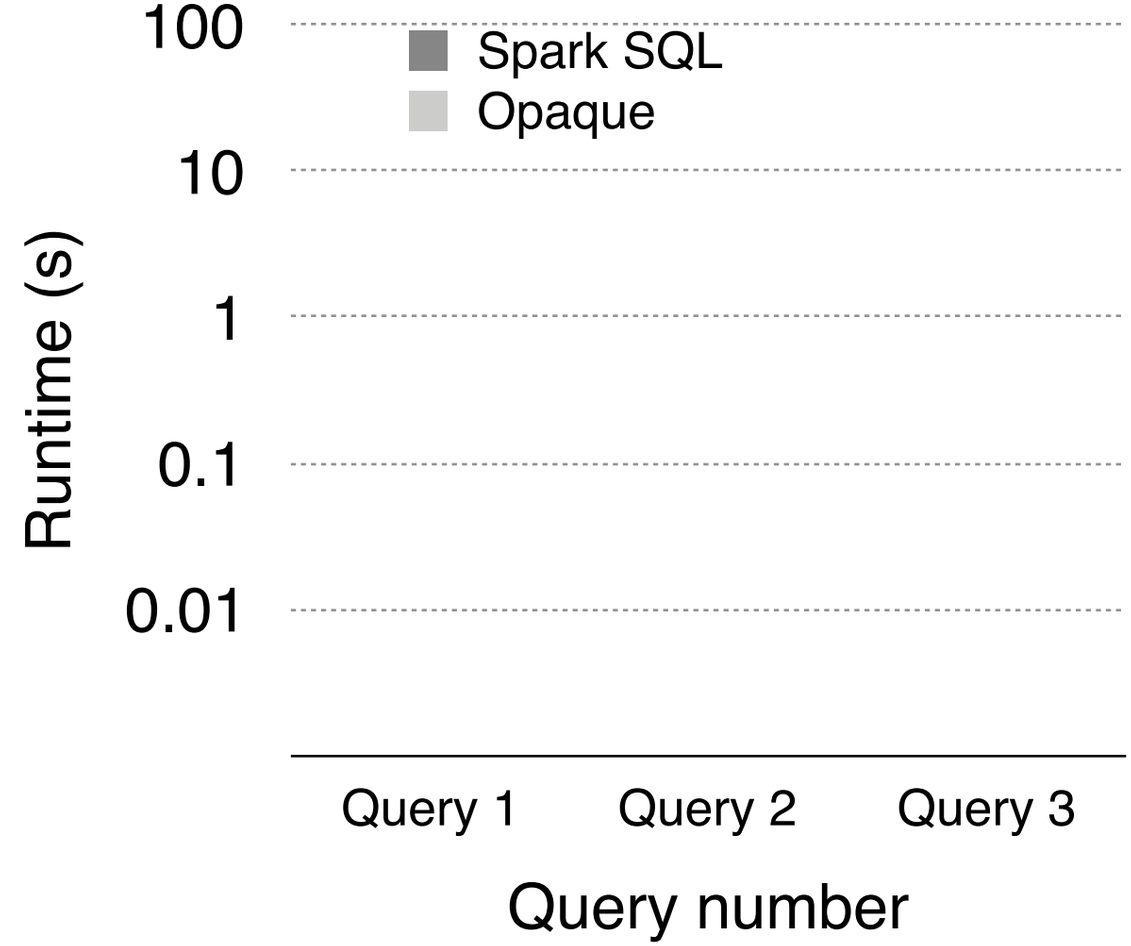
Overhead: -0.47x to 2.3x

Big Data Benchmark (distributed)

Data encryption, authentication, computation verification



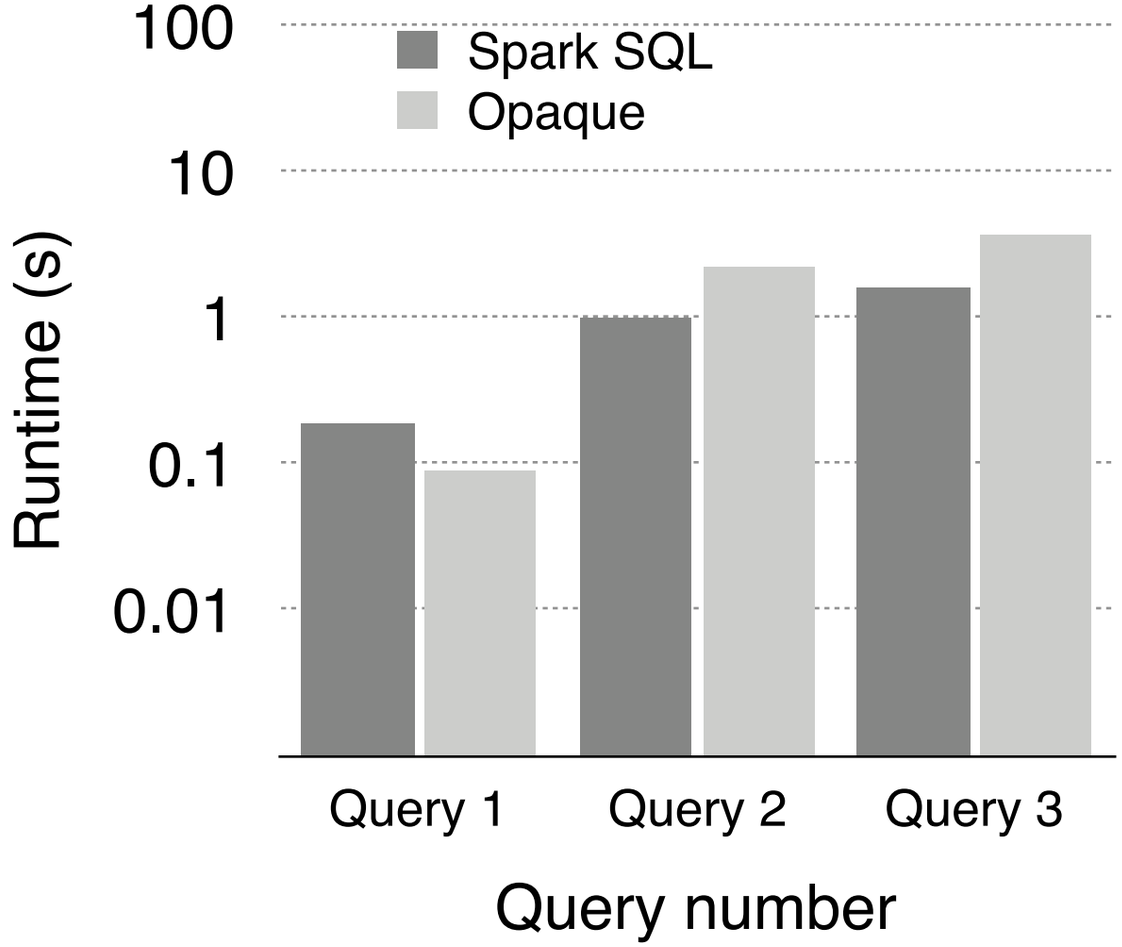
+ Obliviousness



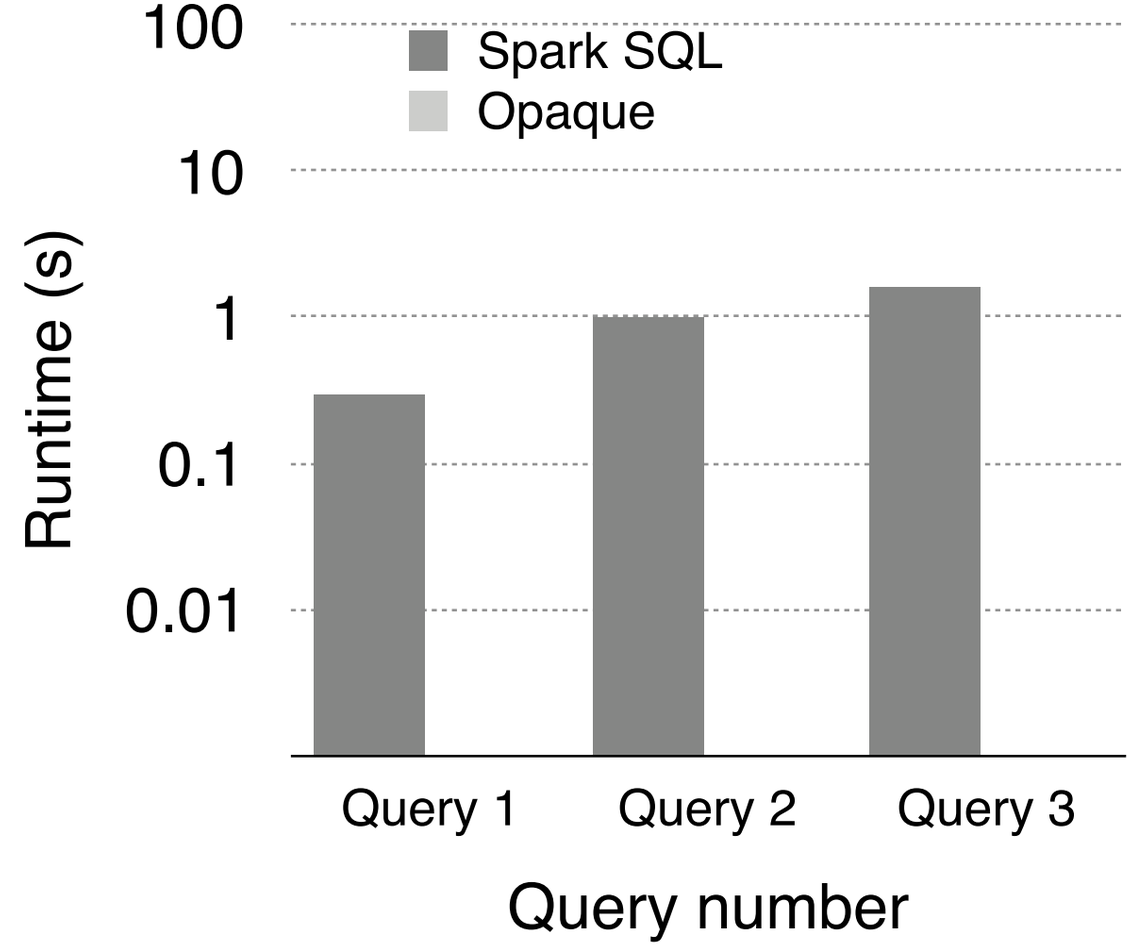
Overhead: -0.47x to 2.3x

Big Data Benchmark (distributed)

Data encryption, authentication, computation verification



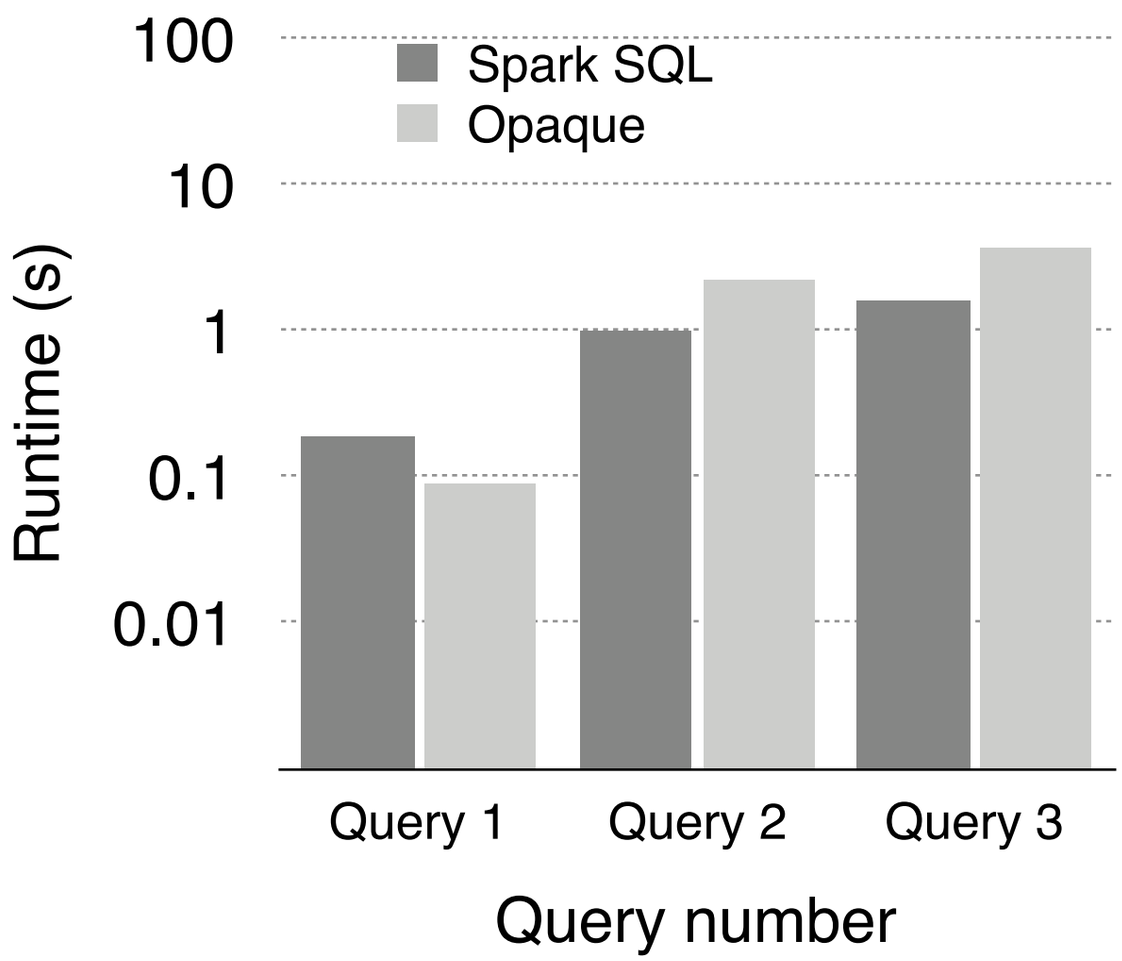
+ Obliviousness



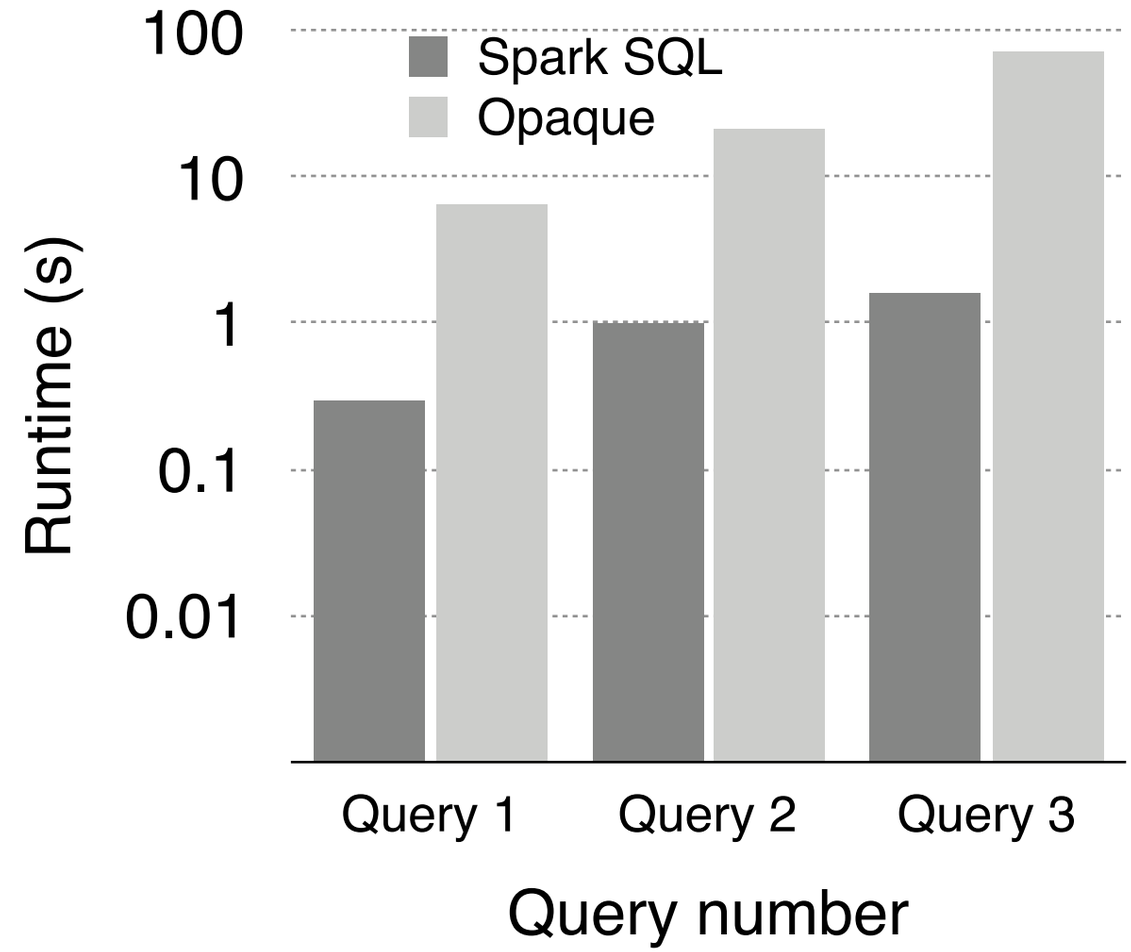
Overhead: -0.47x to 2.3x

Big Data Benchmark (distributed)

Data encryption, authentication, computation verification



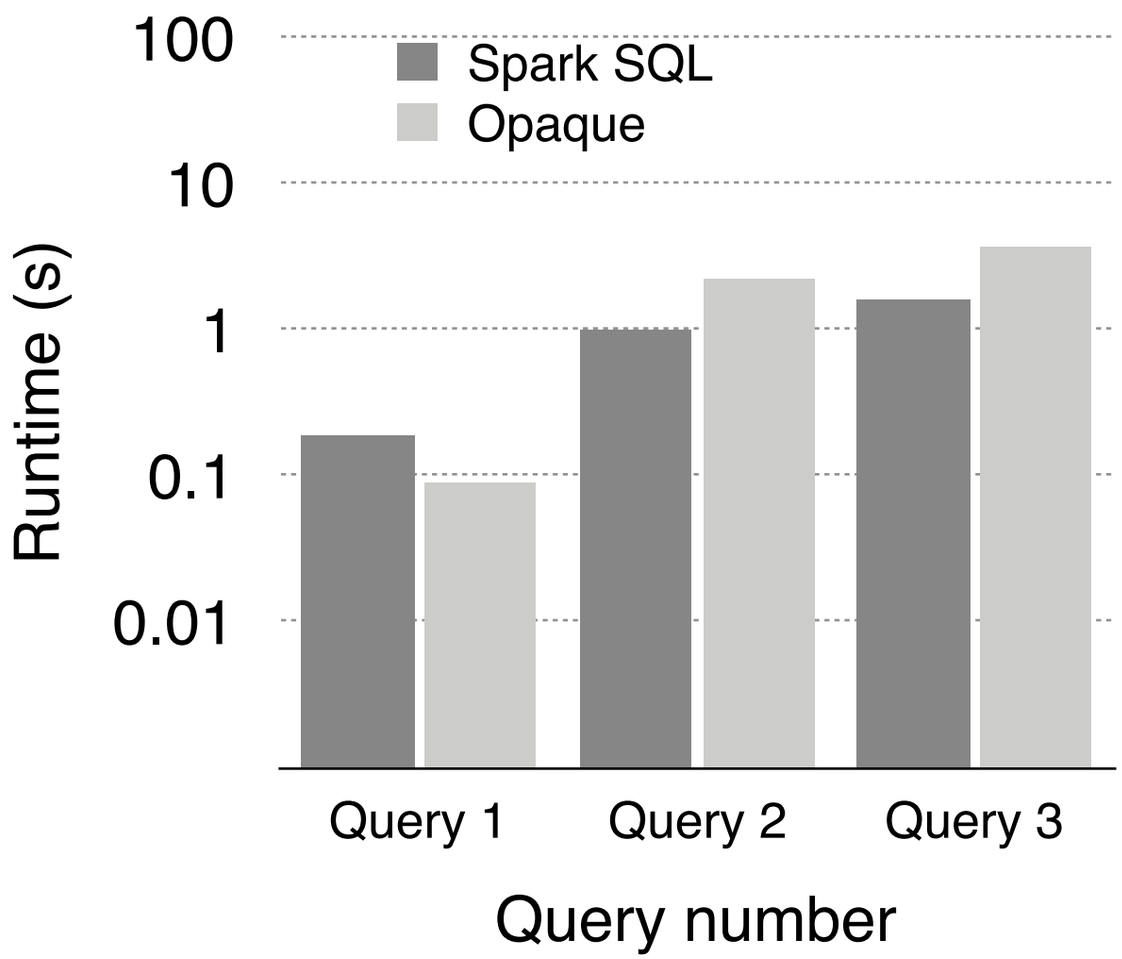
+ Obliviousness



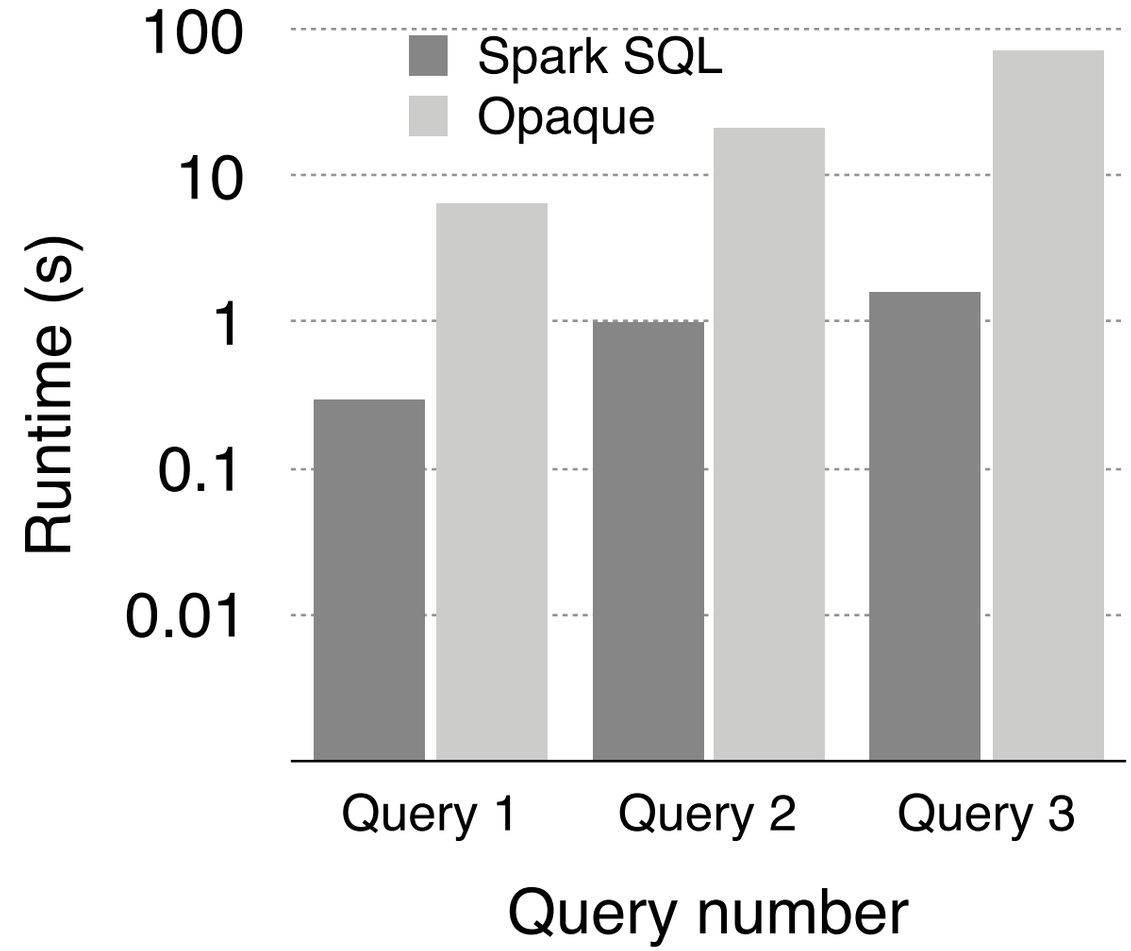
Overhead: -0.47x to 2.3x

Big Data Benchmark (distributed)

Data encryption, authentication, computation verification



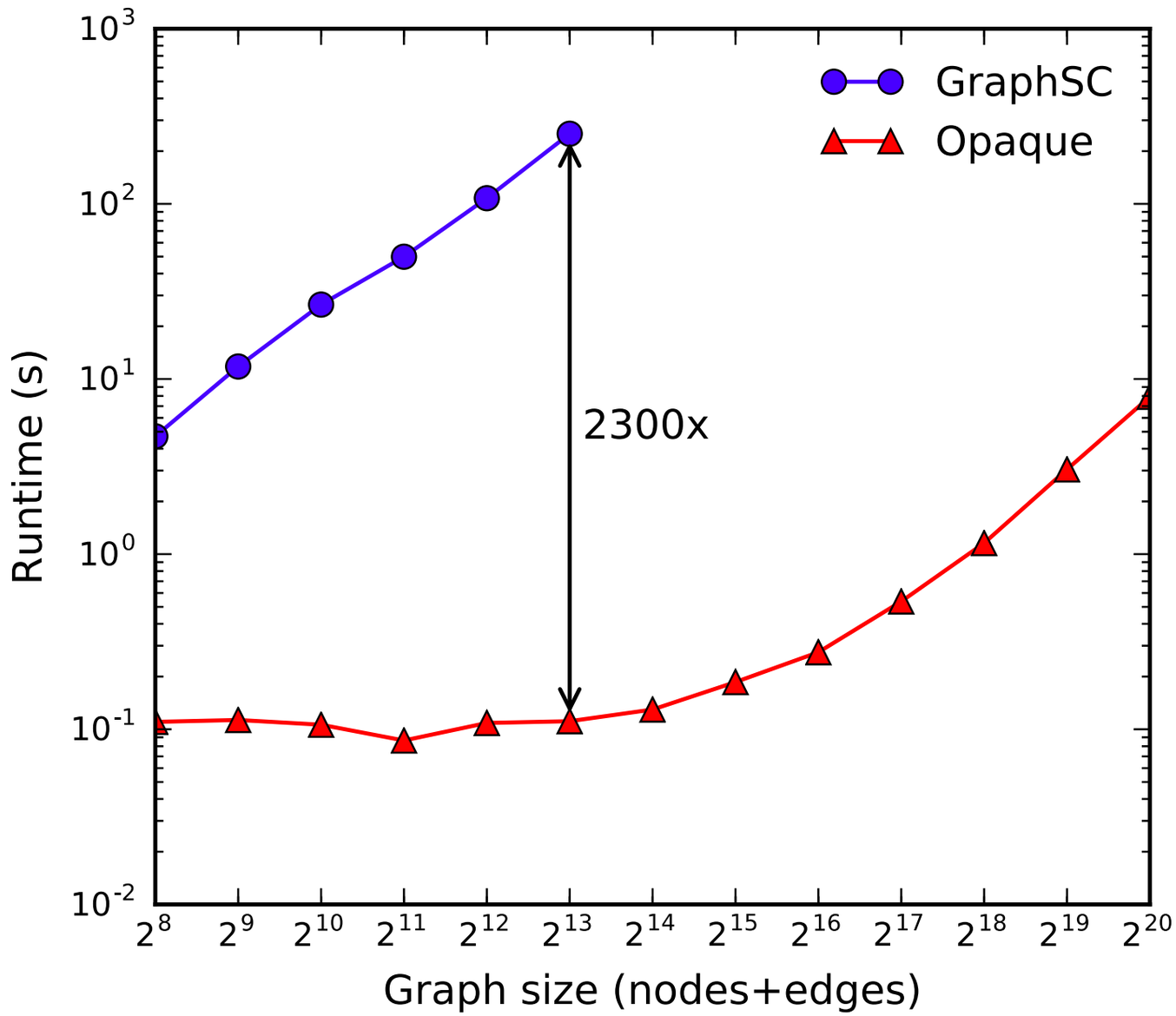
+ Obliviousness



Overhead: -0.47x to 2.3x

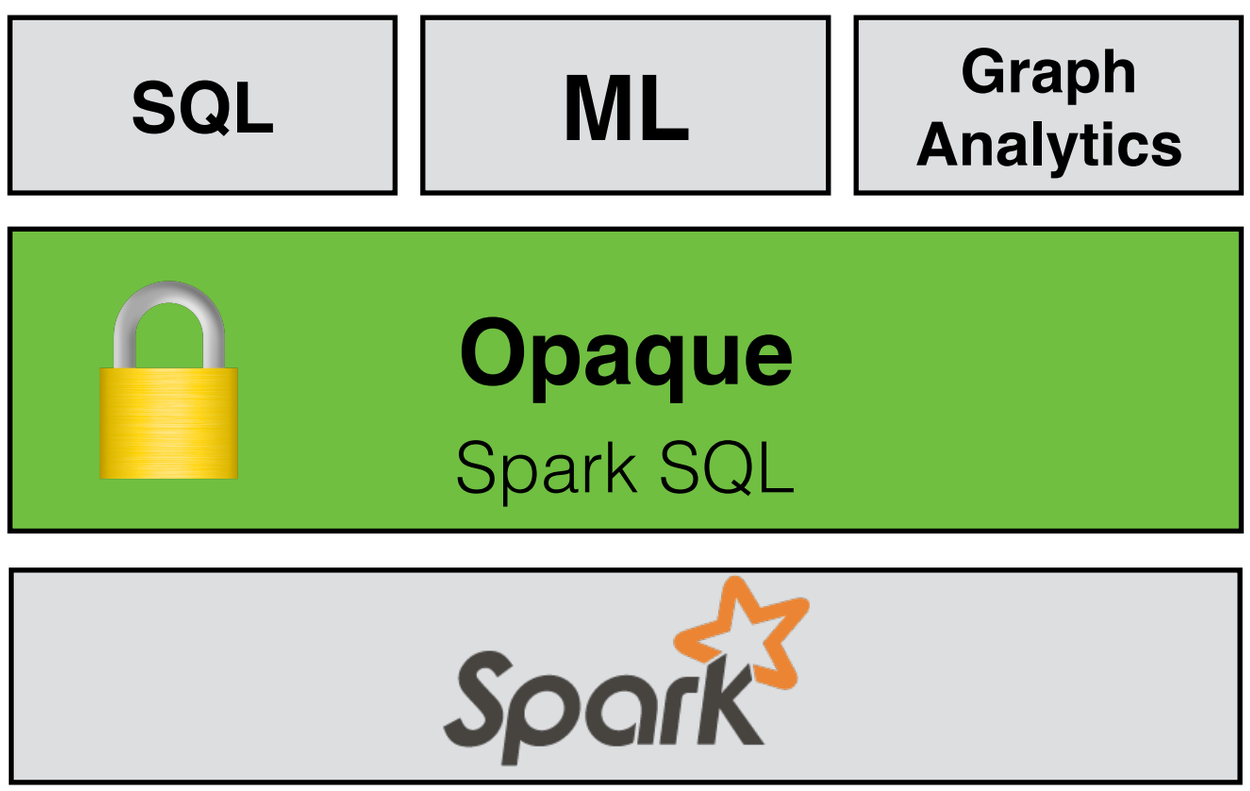
Overhead: 21x to 45x

PageRank: comparison with GraphSC (single machine)



Conclusion

Opaque is an oblivious and encrypted distributed analytics platform



WIP open source: github.com/ucbrise/opaque