UnLynx: A Decentralized System for Privacy-Conscious Data Sharing

PETS 2017

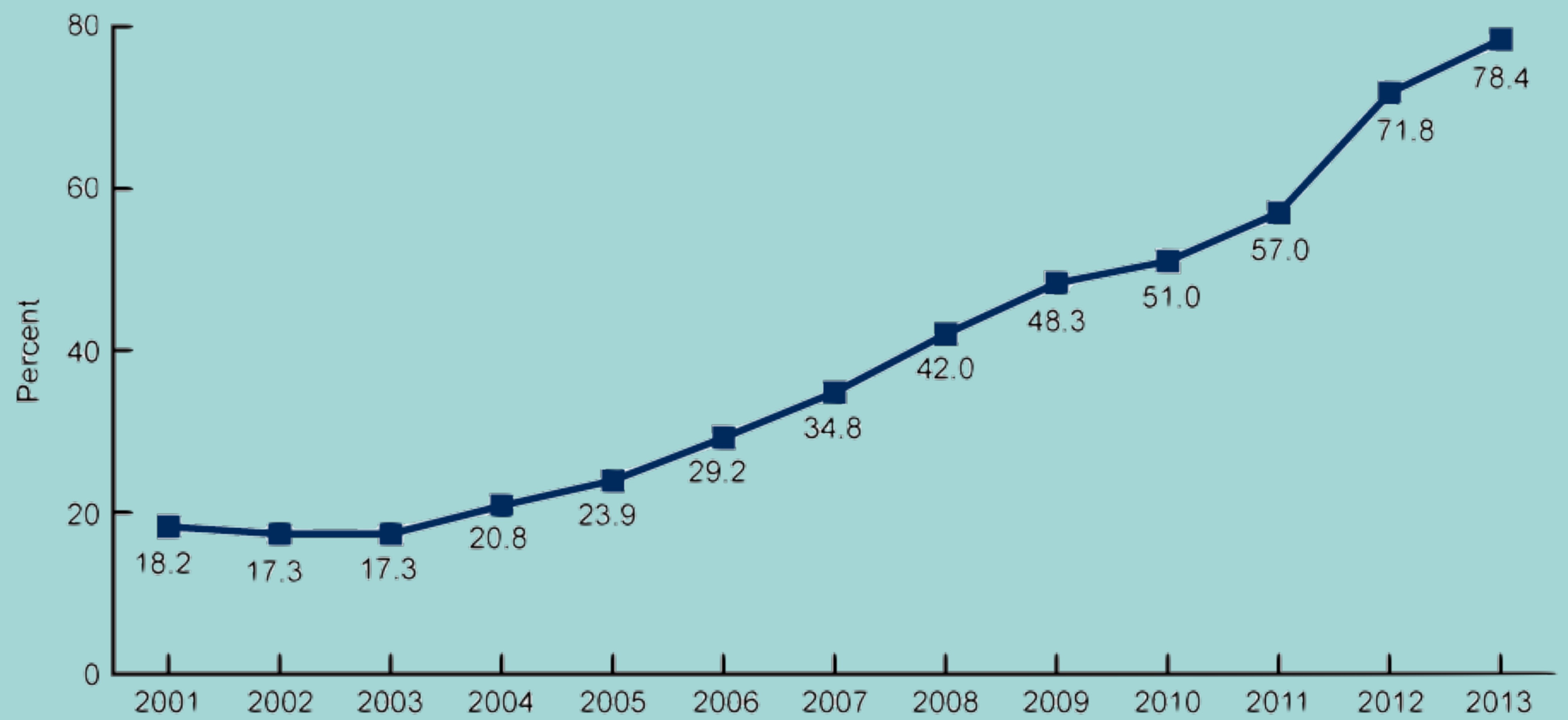
David Froelicher*#, Patricia Egger*#, Joao Sa Sousa*, Jean Louis Raisaro*,
Zhicong Huang*, Christian Mouchet*, Bryan Ford#, Jean-Pierre Hubaux*

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

*LCA1 laboratory
#DeDiS laboratory

More medical data are digitized

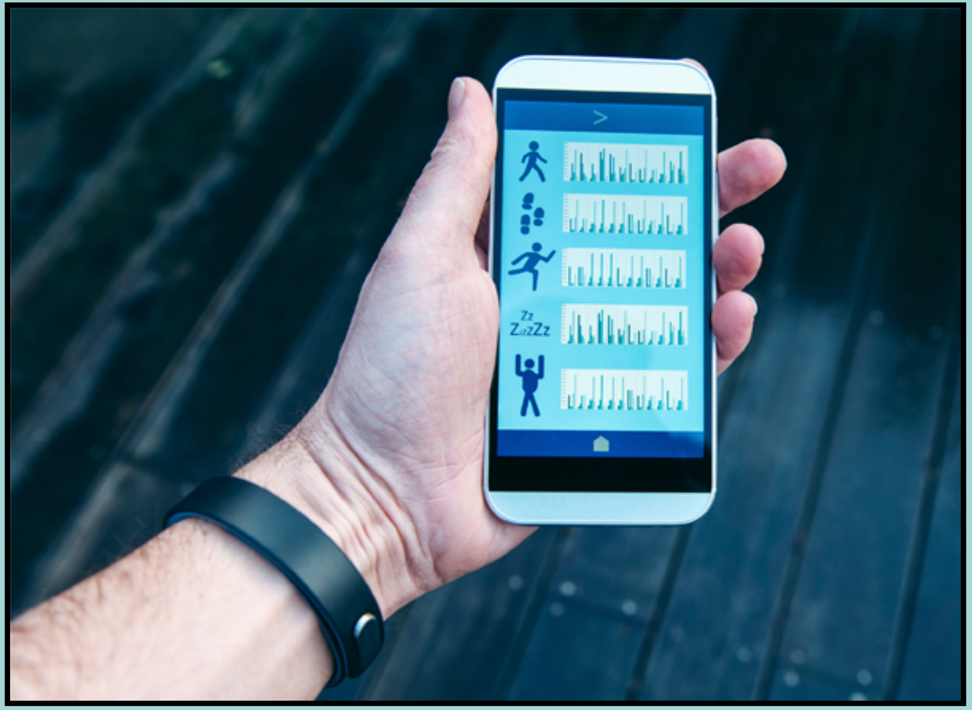
Percentage of office-based physicians with electronic medical records in U.S.A, 2001-2013



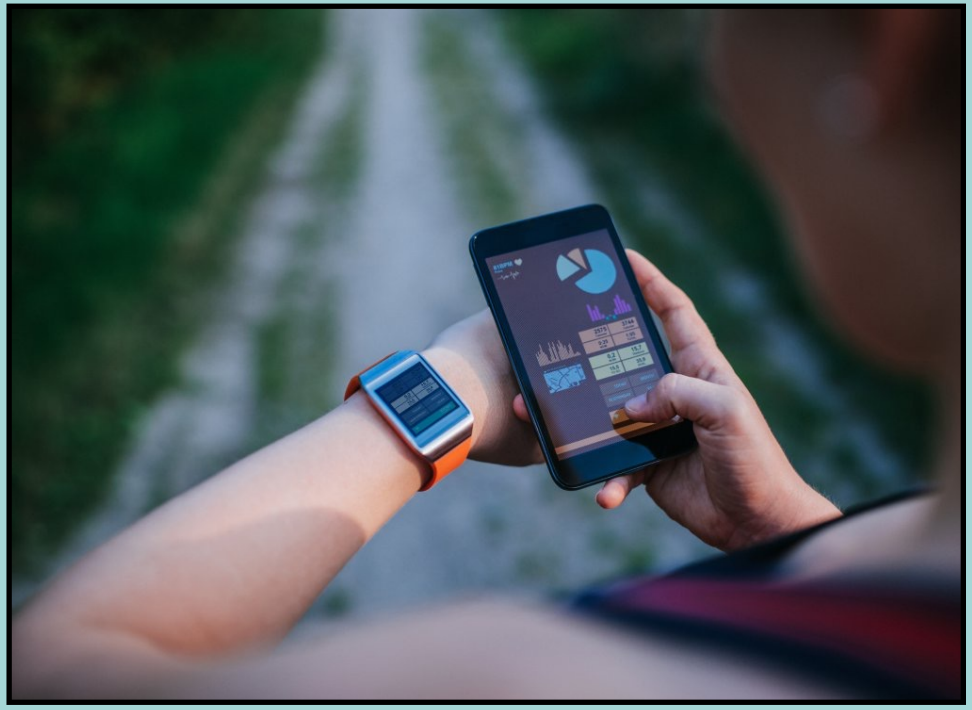
National Ambulatory Medical Care Survey (NAMCS)

http://howinforme.blogspot.ch/2015/09/electronic-health-record-how-to-use.html

More health data collected

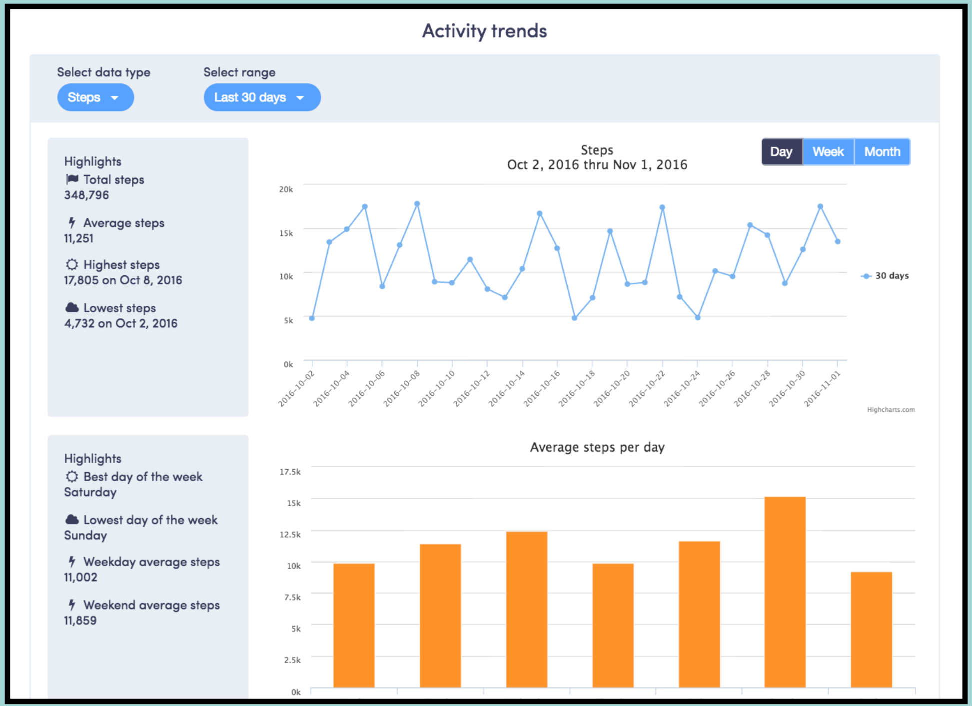


http://blog.stridekick.com/ultimate-guide-fitness-tracker-hacks-get-most-from-fitbit/



http://www.consumerreports.org/cro/news/2015/06/what-you-need-to-know-about-sharing-your-medical-data/index.htm

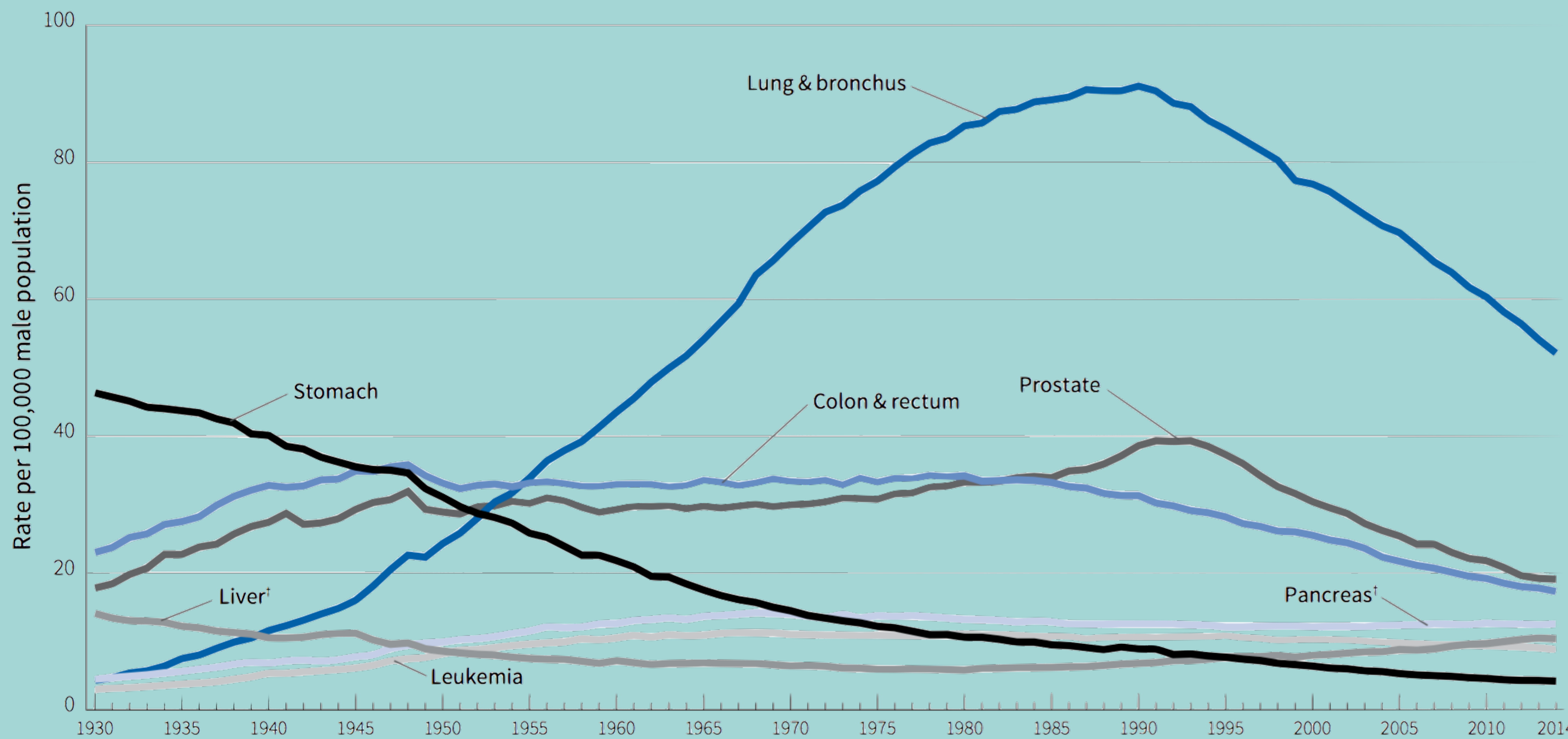
http://time.com/collection-post/3615161/sharing-health-data/



http://www.designindaba.com/articles/creative-work/smart-thermometer-crowdsources-info-real-time-health-tracking

More medical data = better treatments ?

Cancer death rates* among men, USA, 1930-2014



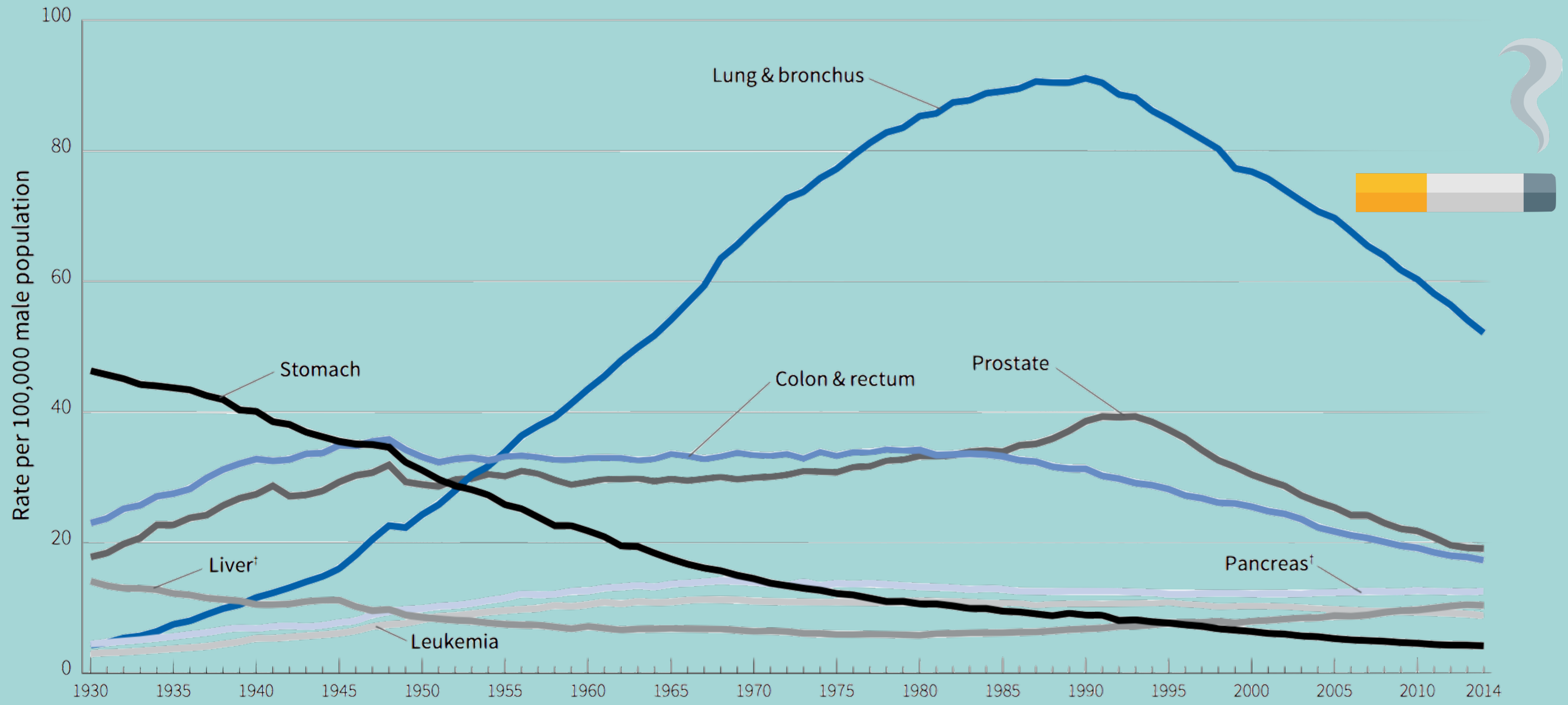
*Per 100,000, age adjusted to the 2000 US standard population. †Mortality rates for pancreatic and liver cancers are increasing.

Note: Due to changes in ICD coding, numerator information has changed over time. Rates for cancers of the liver, lung and bronchus, uterus, and colon and rectum are affected by these coding changes.

Source: US Mortality Volumes 1930 to 1959 and US Mortality Data 1960 to 2014, National Center for Health Statistics, Centers for Disease Control and Prevention.

More medical data = better treatments ?

cancer death rates* among men, USA, 1930-2014



*Per 100,000, age adjusted to the 2000 US standard population. †Mortality rates for pancreatic and liver cancers are increasing.

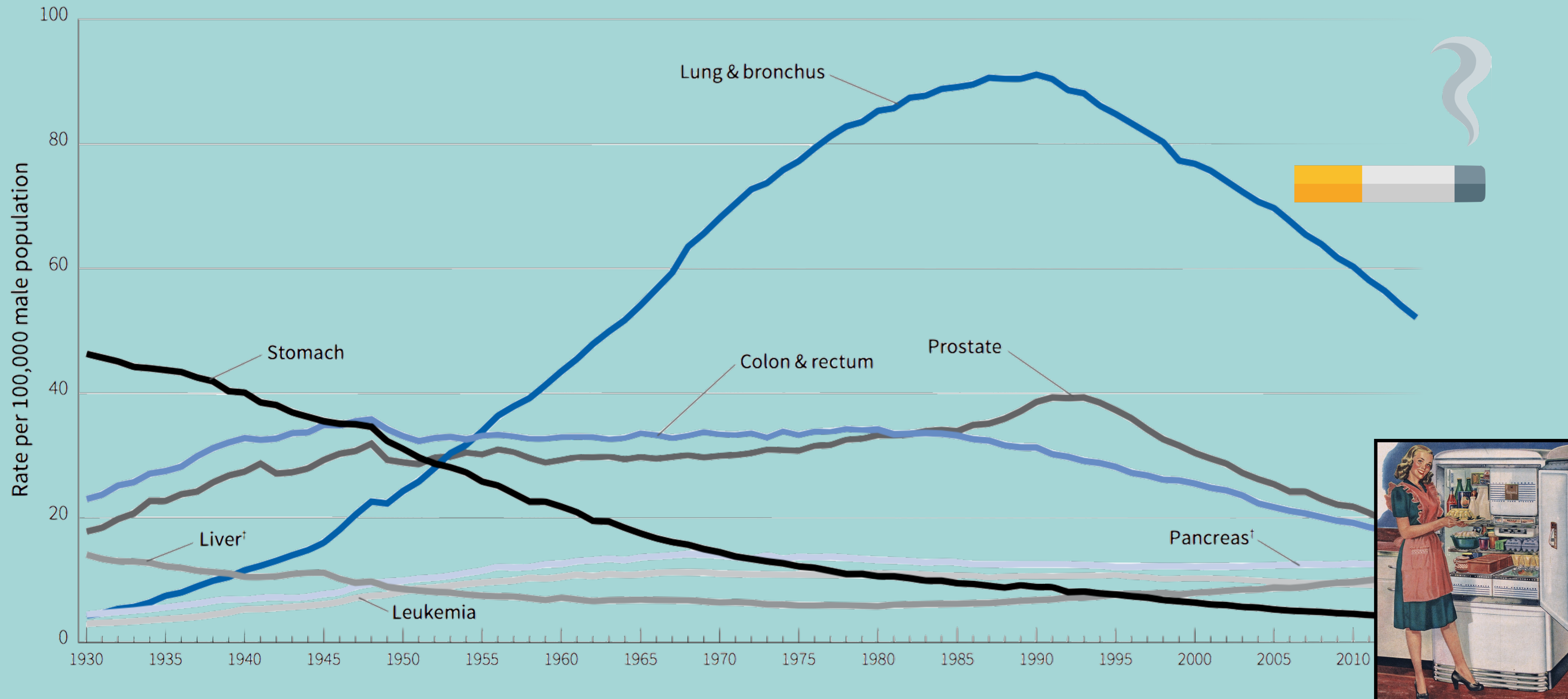
Note: Due to changes in ICD coding, numerator information has changed over time. Rates for cancers of the liver, lung and bronchus, uterus, and colon and rectum are affected by these coding changes.

Source: US Mortality Volumes 1930 to 1959 and US Mortality Data 1960 to 2014, National Center for Health Statistics, Centers for Disease Control and Prevention.

https://www.iconexperience.com/g_collection/icons/?icon=cigarette

More medical data = better treatments ?

cancer death rates* among men, USA, 1930-2014



*Per 100,000, age adjusted to the 2000 US standard population. †Mortality rates for pancreatic and liver cancers are increasing.

Note: Due to changes in ICD coding, numerator information has changed over time. Rates for cancers of the liver, lung and bronchus, uterus, and colon and rectum are affected by these coding changes.

Source: US Mortality Volumes 1930 to 1959 and US Mortality Data 1960 to 2014, National Center for Health Statistics, Centers for Disease Control and Prevention.

https://www.iconexperience.com/g_collection/icons/?icon=cigarette
https://www.zazzle.co.uk/vintage+refrigerator+advertisement+gifts

Sensitive-data sharing is difficult

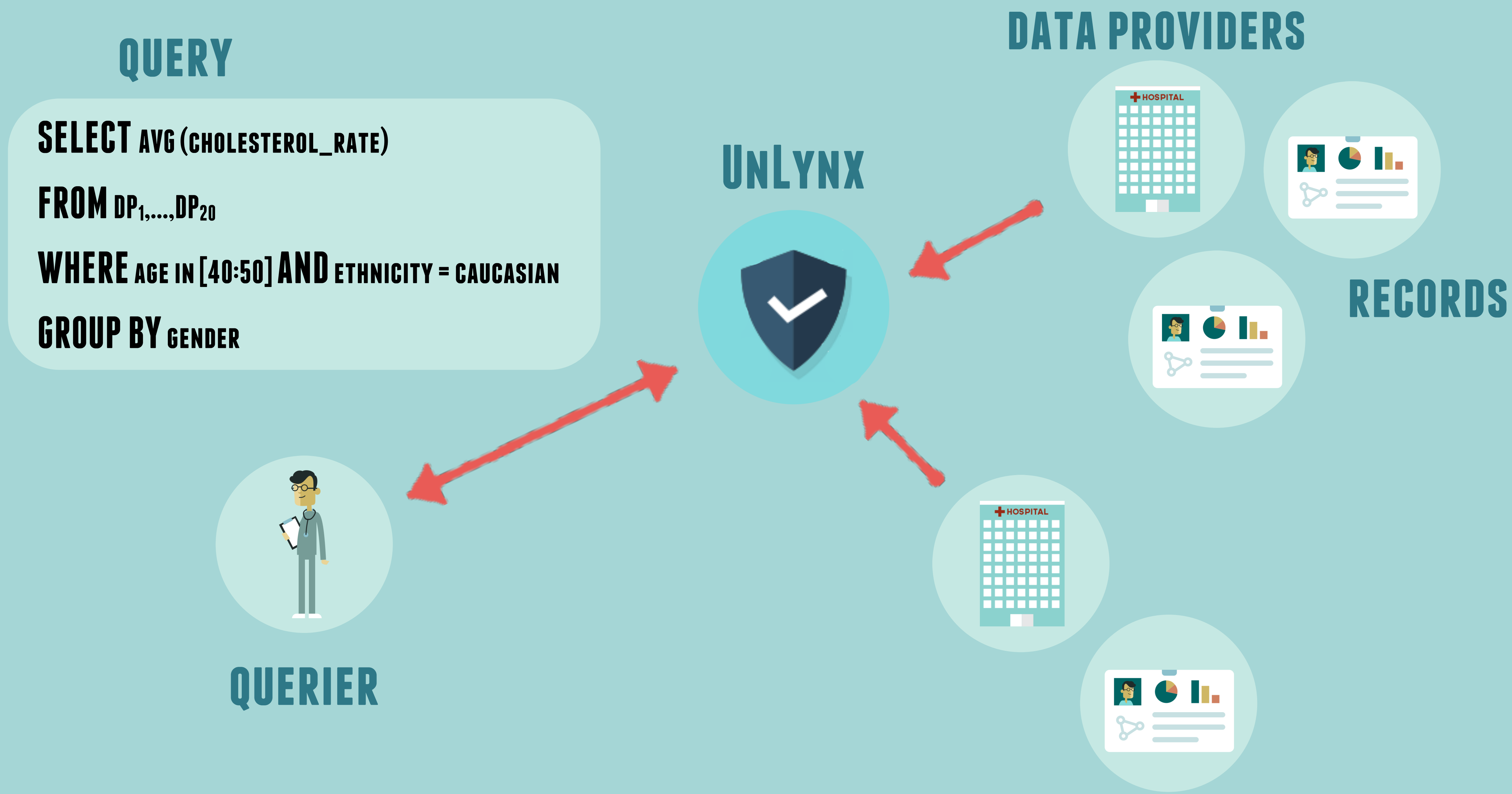
http://clubpenguin.wikia.com/wiki/Money_Bag

http://blog.lpinnovations.com/

Sensitive-data sharing is difficult

http://www.gmmill.net/proje-Grain-Storage-Silos

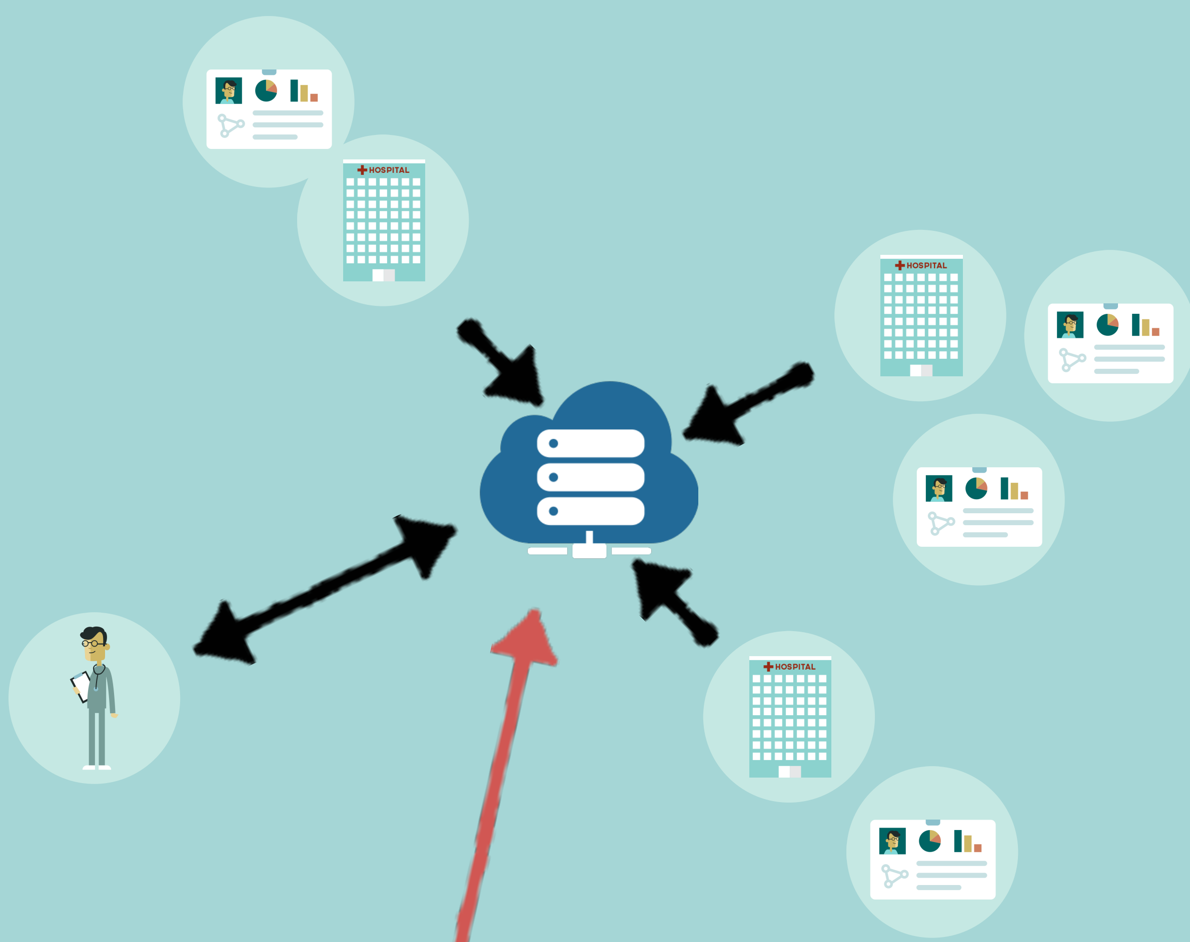
UnLynx



Allow statistical queries on multiple independent databases while ensuring privacy and confidentiality for data providers.

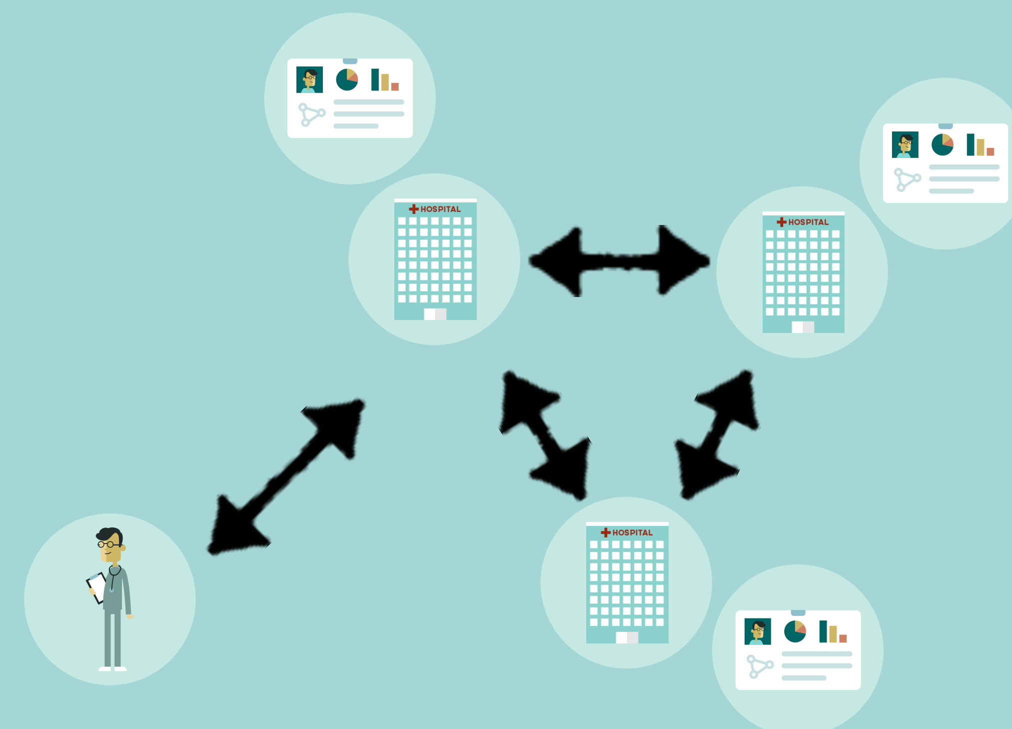
Existing data sharing solutions

Centralized Solutions



Single point of failure

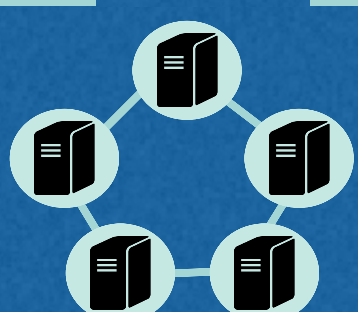
Decentralized Solutions



Limited number of data providers/computation entities in an adversarial model

Requirements

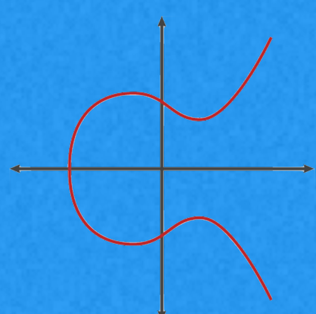
Building Blocks



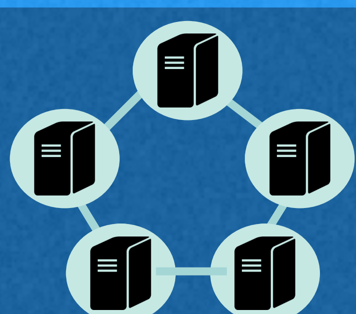
Collective Authority

DECENTRALIZED
TRUST

Building Blocks

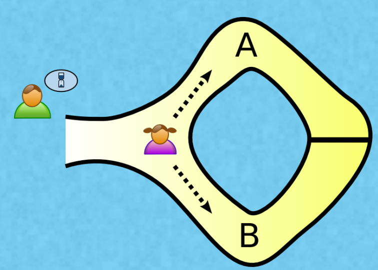


Additively-homomorphic ElGamal crypto scheme



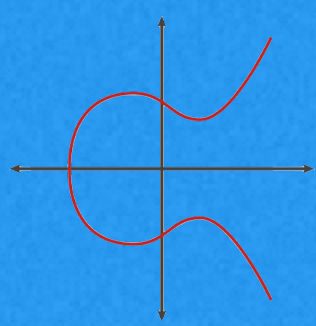
Collective Authority

Building Blocks

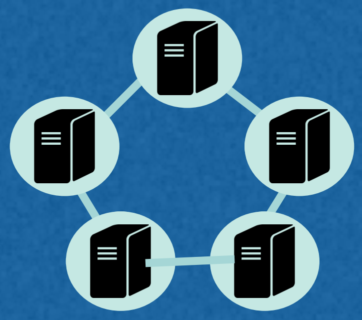


https://simple.wikipedia.org/wiki/Zero-knowledge_proof

Zero-Knowledge Proofs of Correctness



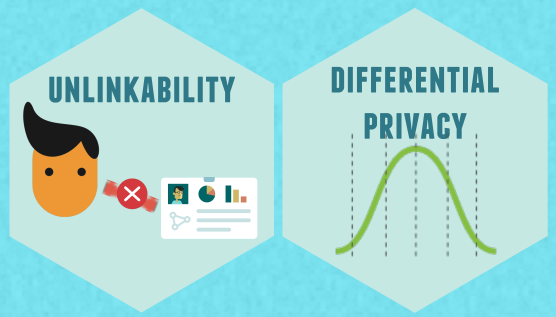
Additively-homomorphic ElGamal crypto scheme

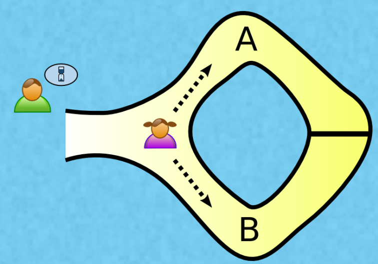


Collective Authority

Building Blocks

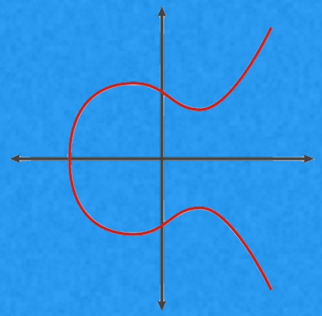
Verifiable Shuffle



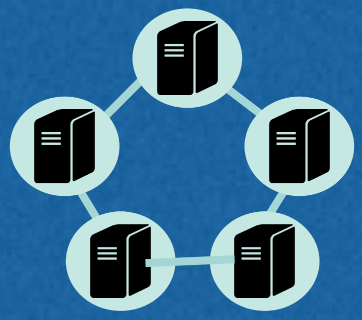


https://simple.wikipedia.org/wiki/Zero-knowledge_proof

Zero-Knowledge Proofs of Correctness

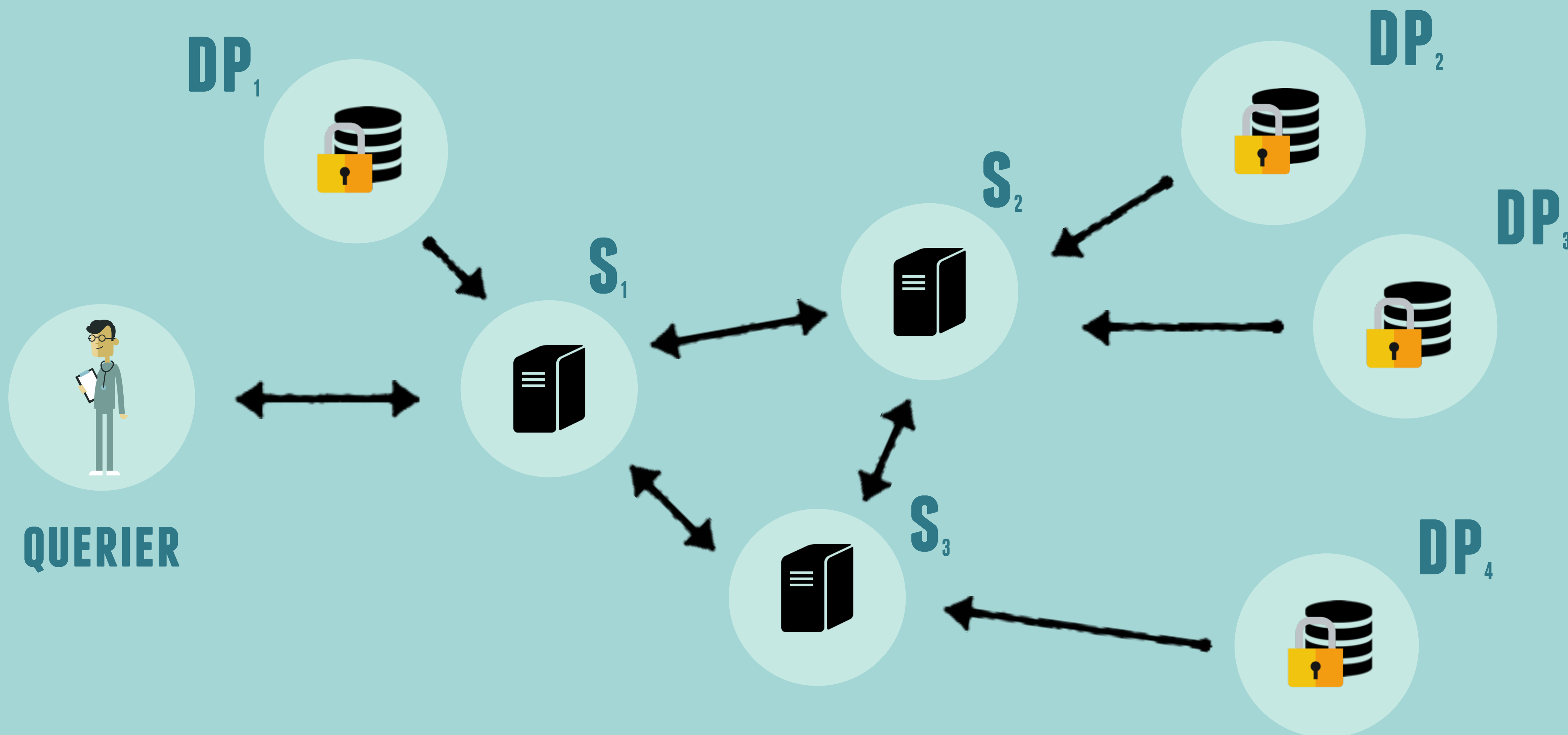


Additively-homomorphic ElGamal crypto scheme



Collective Authority

System Model

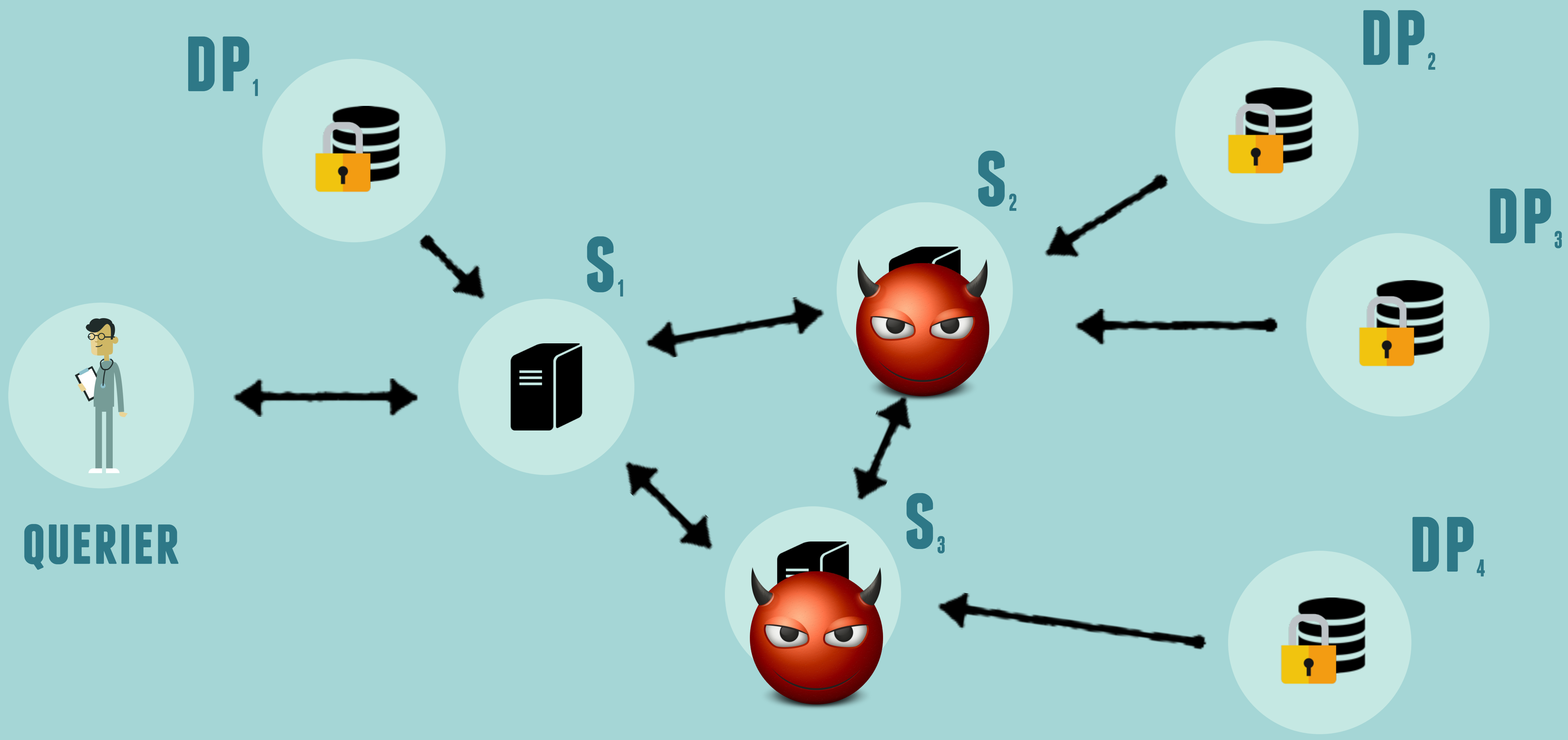


- Collective authority of *m* servers *S*
- *n* Data Providers *DPs*
- Clients *Q* querying the system

DP = Data Provider

S = Server

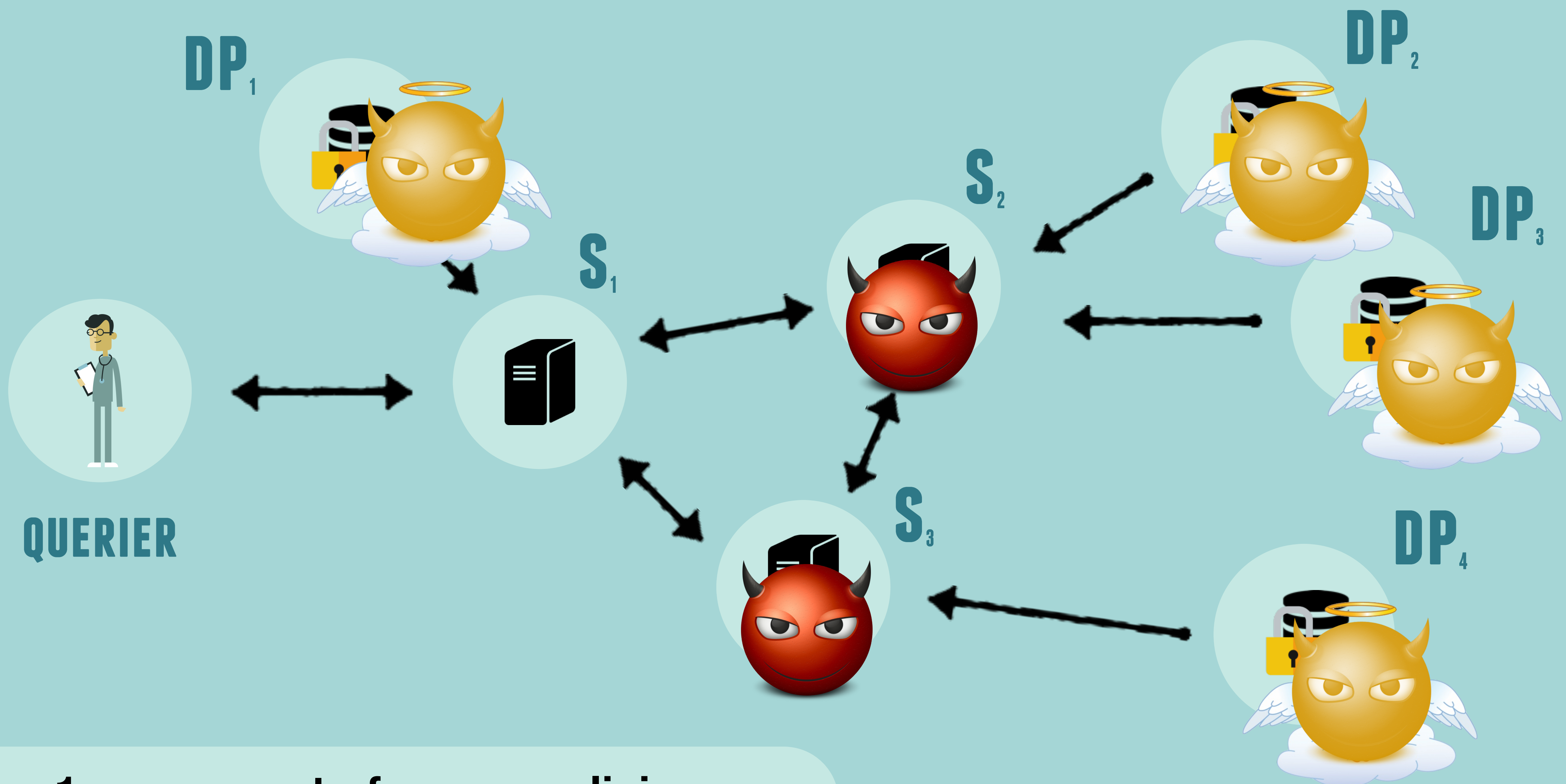
Threat Model



- m-1 servers out of m are malicious (Anytrust Model)

DP = Data Provider
S = Server

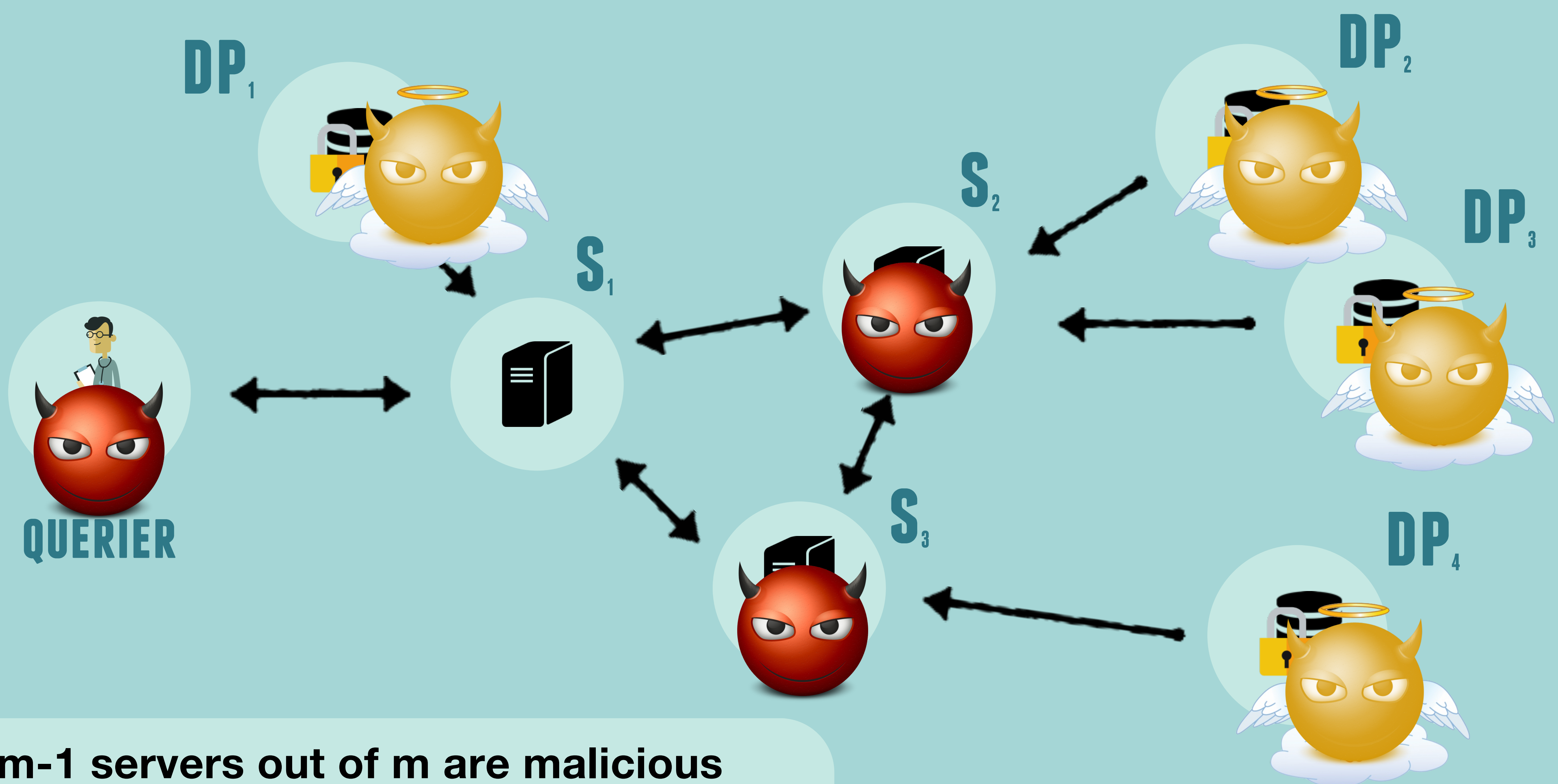
Threat Model



- m-1 servers out of m are malicious (Anytrust Model)
- Data Providers are honest-but-curious

DP = Data Provider
S = Server

Threat Model

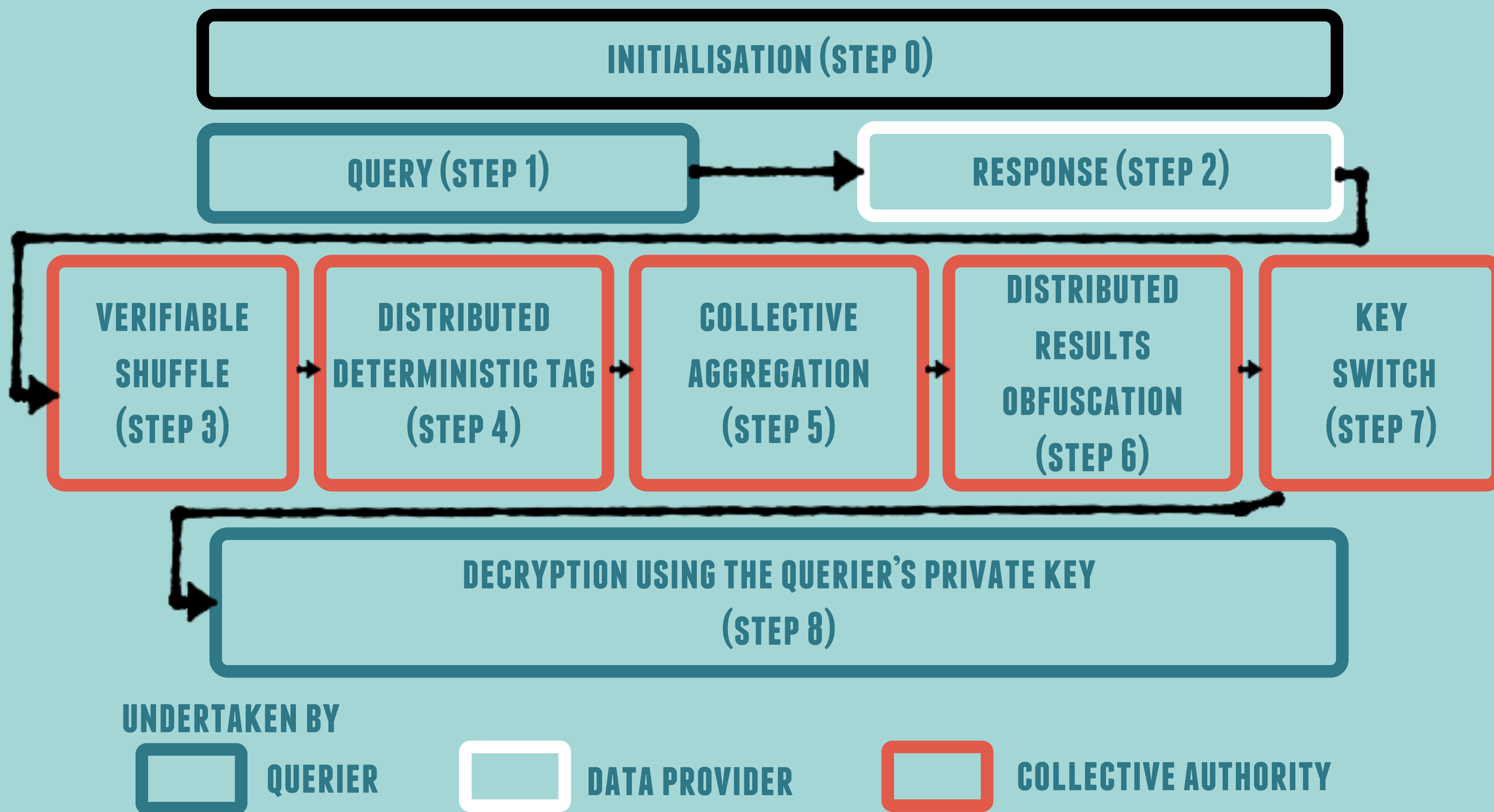


- m-1 servers out of m are malicious (Anytrust Model)
- Data Providers are honest-but-curious
- Queriers are malicious

DP = Data Provider

S = Server

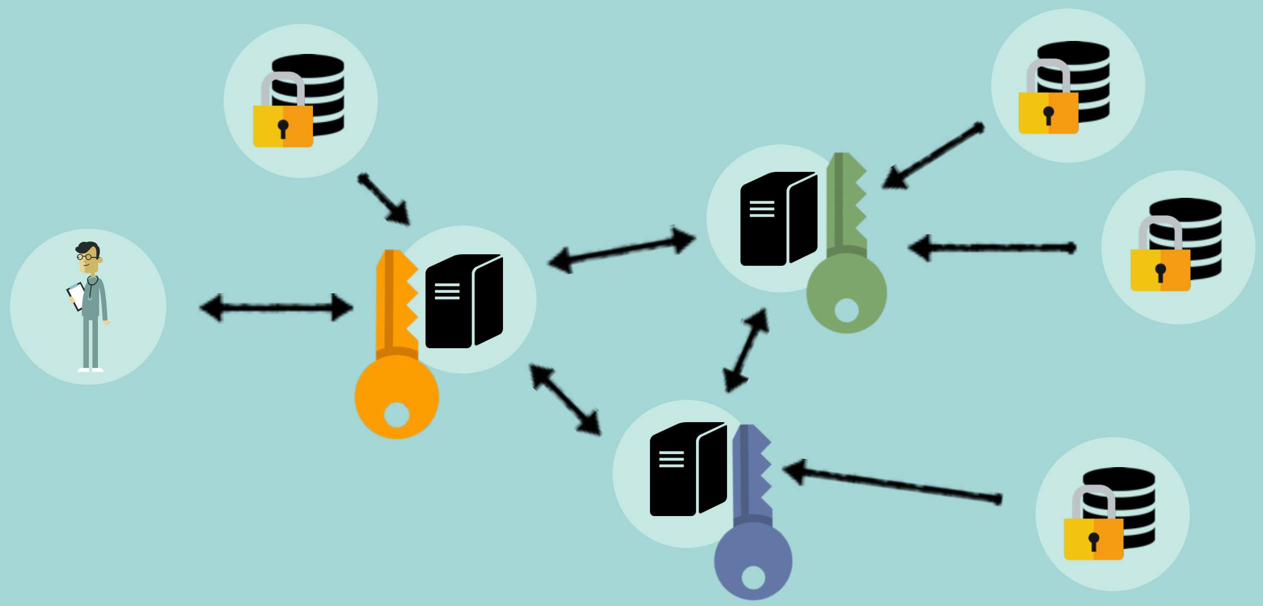
Query processing Workflow



Workflow - Initialisation (Step 0)

INITIALISATION (STEP 0)

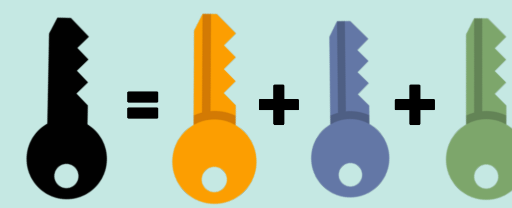
Each server constructs his public-private ElGamal Key pair.

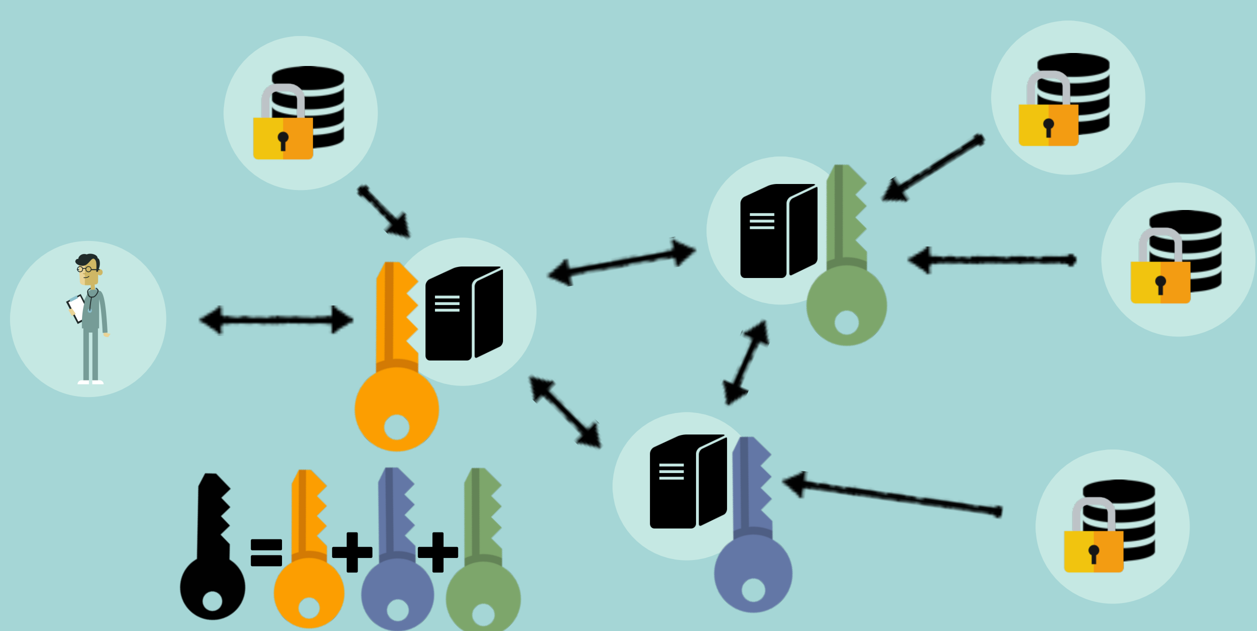


Workflow - Initialisation (Step 0)

INITIALISATION (STEP 0)

Each server constructs his public-private ElGamal Key pair.

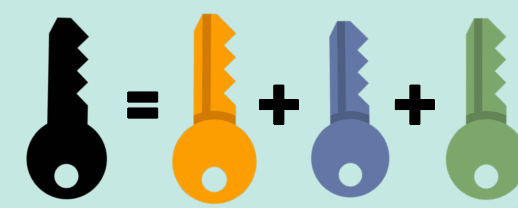
Collective Key: 



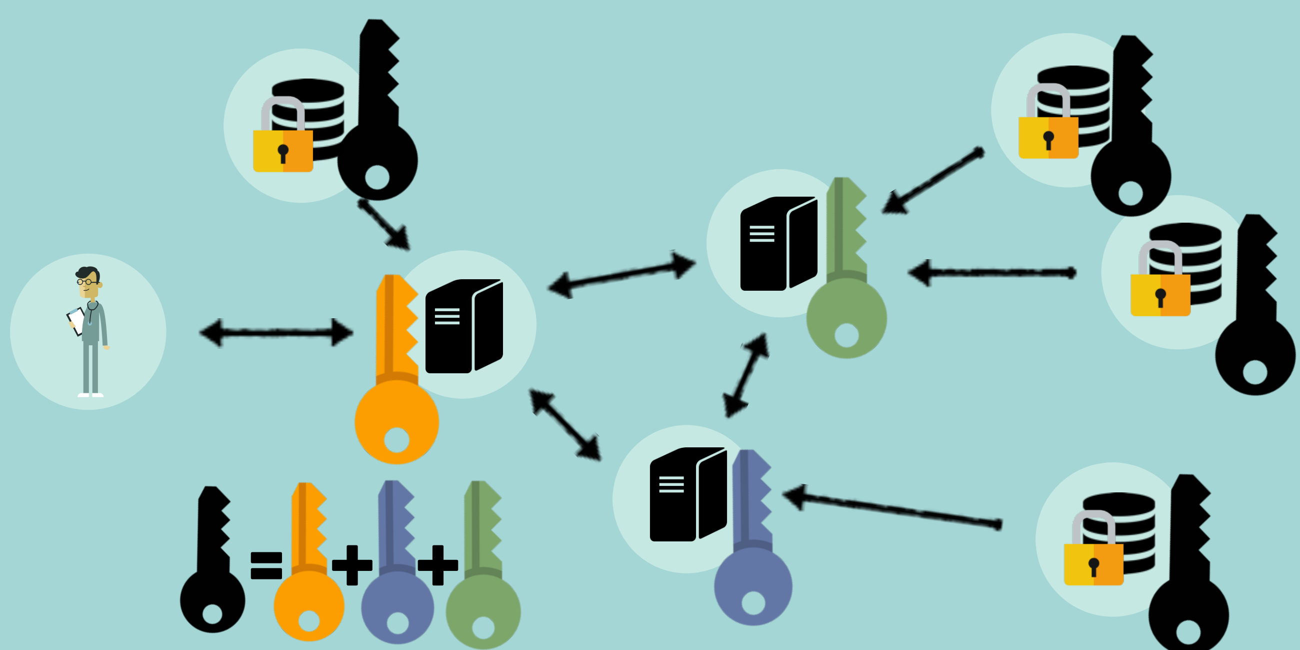
Workflow - Initialisation (Step 0)

INITIALISATION (STEP 0)

Each server constructs his public-private ElGamal Key pair.

Collective Key: 

Data Providers use the Collective Key to encrypt their data



Workflow - Query (Step 1)

Initialisation (Step 0)

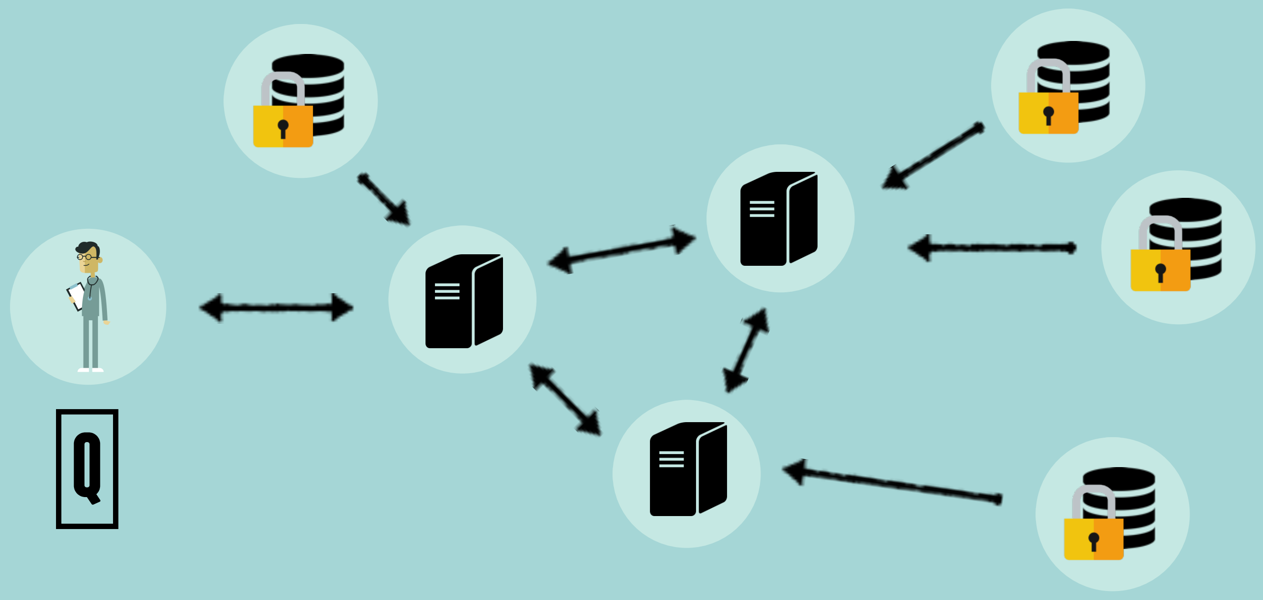
Query (Step 1)

SELECT SUM (cholesterol_rate) , COUNT(*)

FROM $DP_1,\dots,DP_{20}$

WHERE age IN [40:50] AND ethnicity = caucasian

GROUP BY gender



Workflow - Query (Step 1)

INITIALISATION (STEP 0)

QUERY (STEP 1)

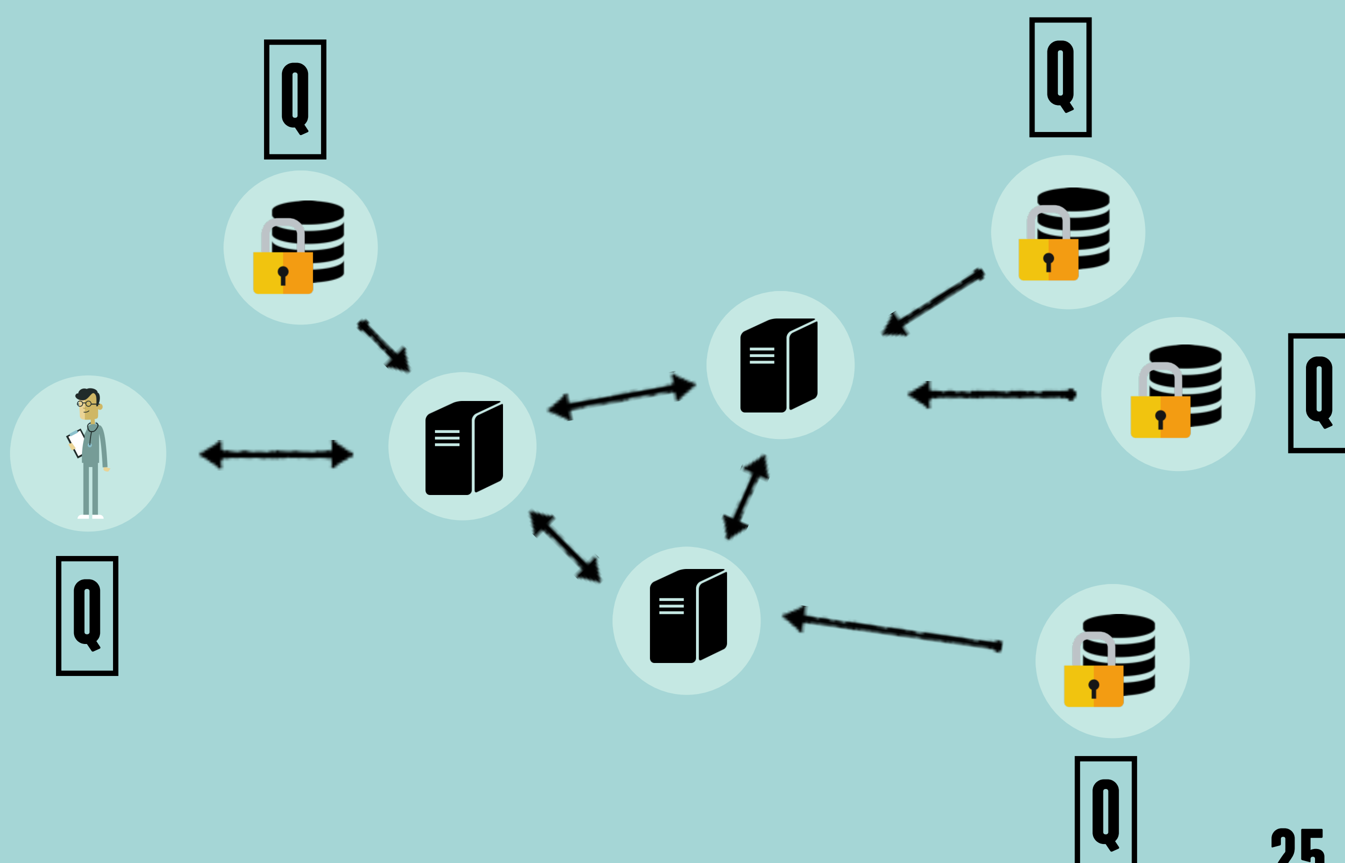
SELECT SUM (CHOLESTEROL_RATE) , COUNT(*)

FROM $DP_1,\ldots,DP_{20}$

WHERE AGE IN [40:50] AND ETHNICITY = CAUCASIAN

GROUP BY GENDER

Query broadcasted to Data Providers



Workflow - Response (Step 2)

INITIALISATION (STEP 0)

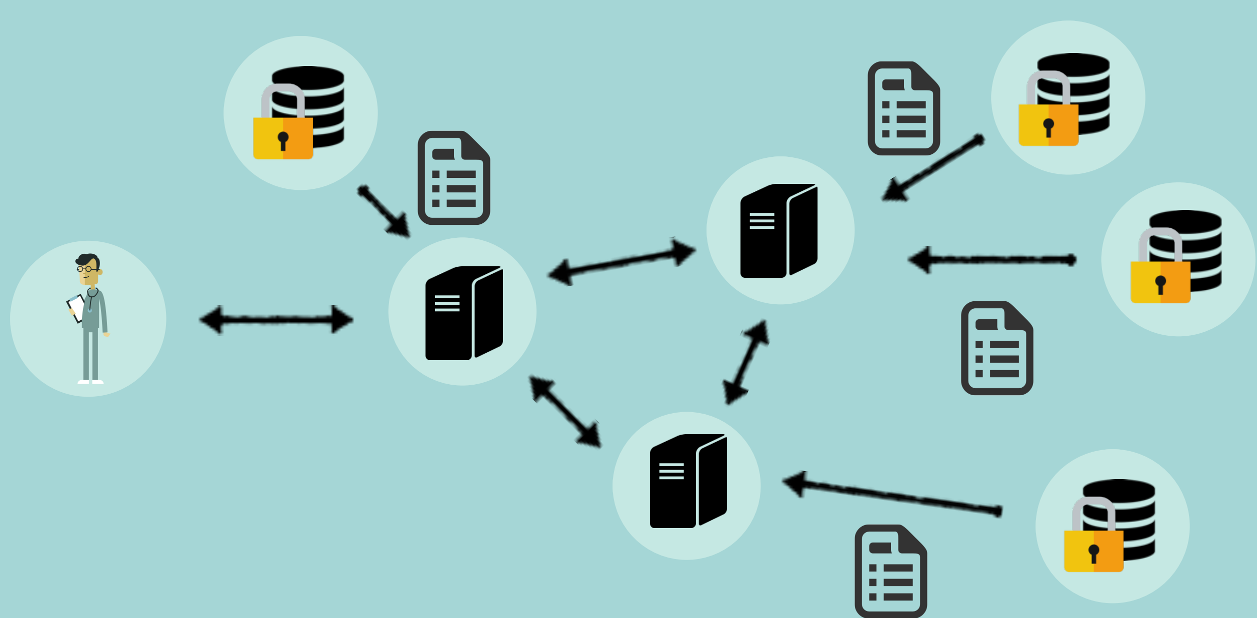
QUERY (STEP 1) → RESPONSE (STEP 2)

ID	Gender	Age	Ethnicity	flu	Cholesterol_rate	cancer
1	E(1)	E(40)	E(1)	E(1)	E(23)	E(0)
2	E(2)	E(40)	E(2)	E(0)	E(34)	E(0)
...	...	...	...	...	...	...

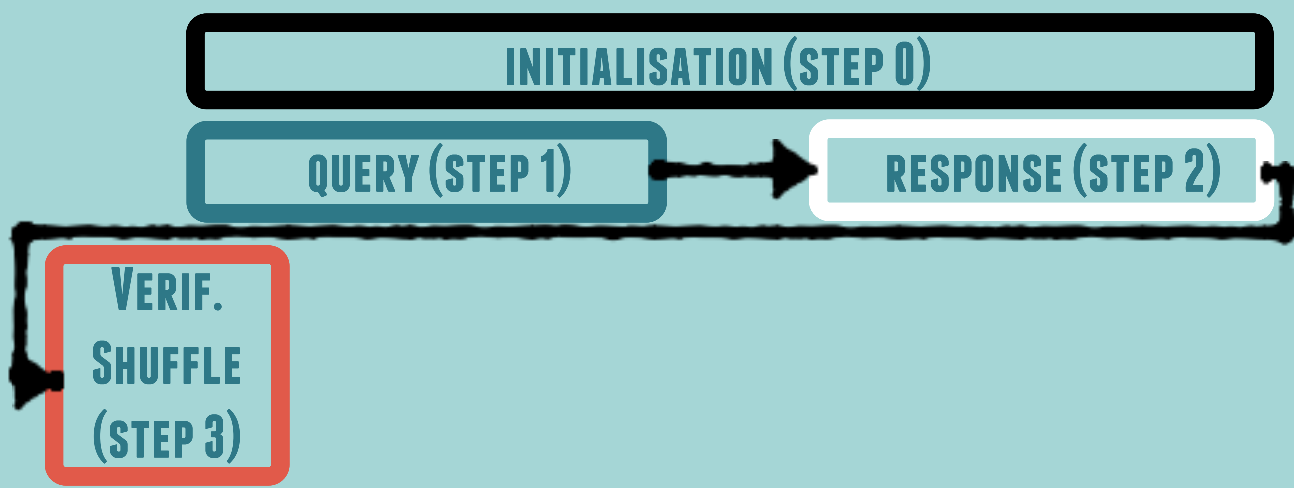
[[**group**. attr.], [**where**. attr.], [**aggr.** Attr.]]

[[E(1)], [E(40),E(1)], [E(23),E(0)]]

…



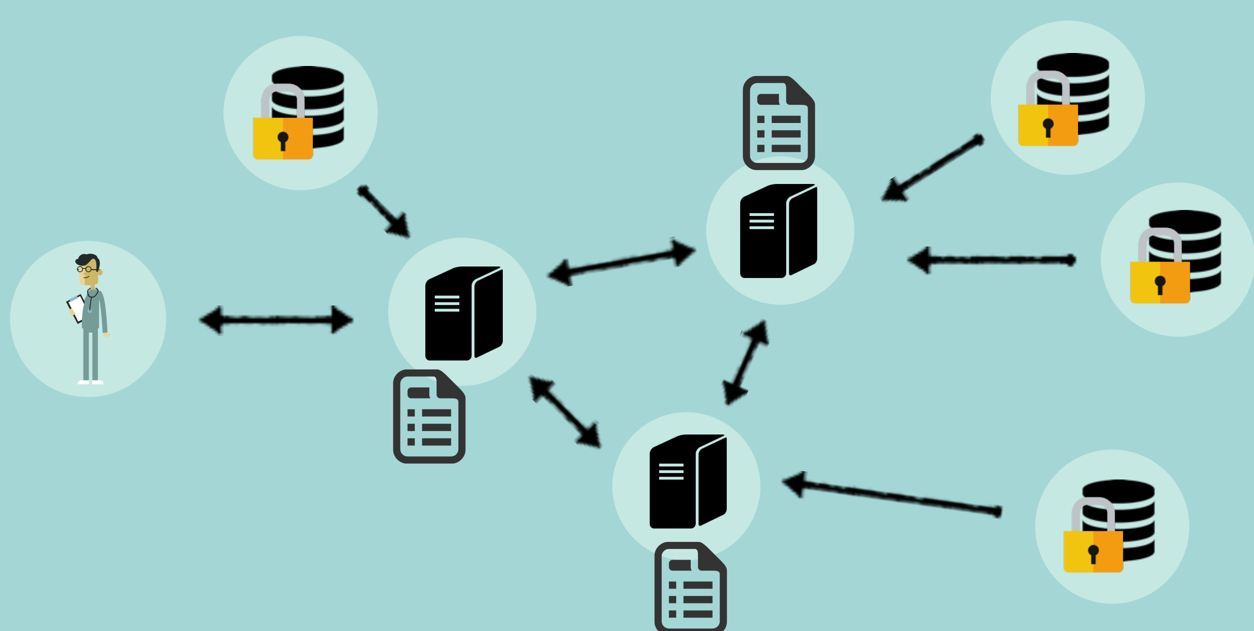
Workflow - Verif. Shuffle (Step 3)



Each server starts a **verifiable shuffle protocol**:

In this protocol each server sequentially:

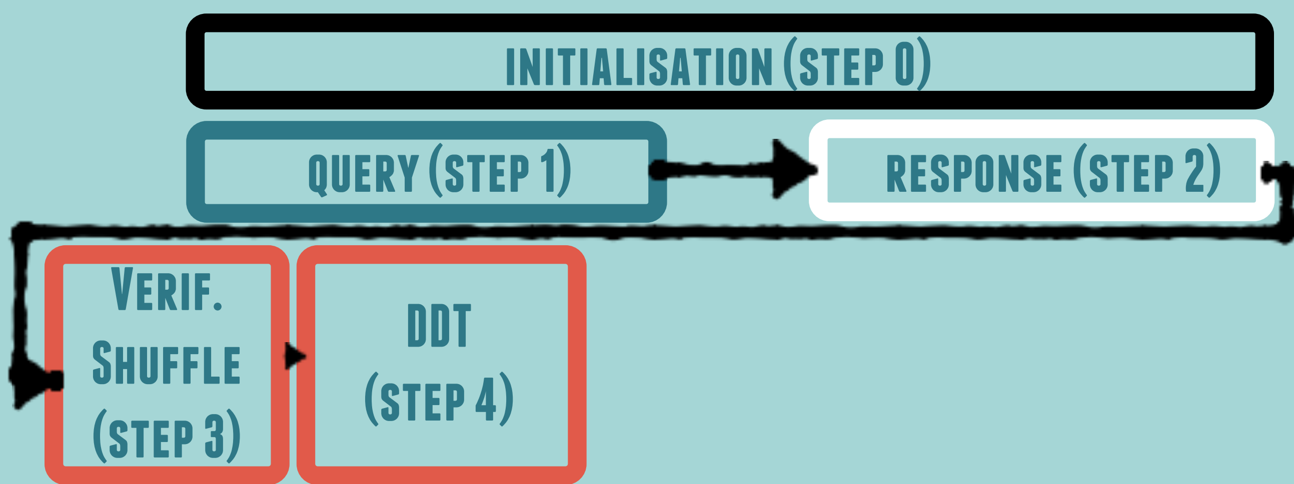
- **Shuffle** the list of responses
- **Rerandomize** (re-encryption) all the ciphertexts

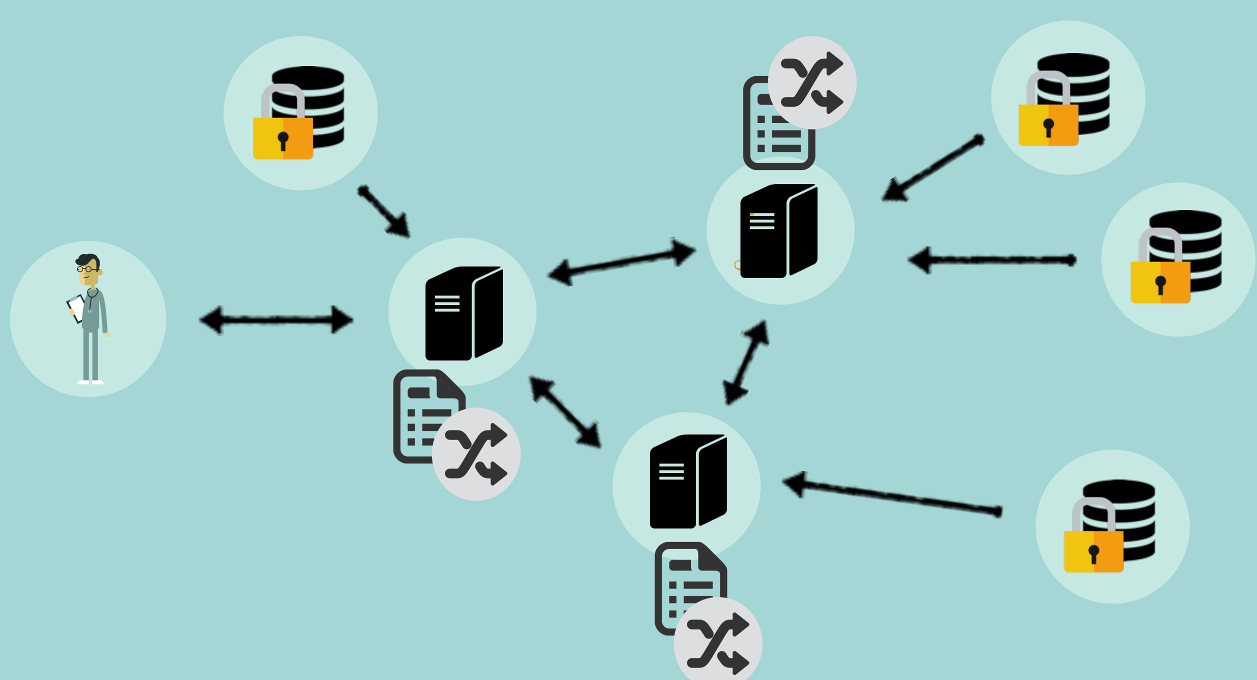


Using Neff Shuffle and the corresponding zero-knowledge proof [1]

[1] Andrew Neff. Verifiable mixing (shuffling) of ElGamal pairs (2004)

Workflow - DDT (Step 4)





Each server starts a **distributed deterministic tagging protocol**:

Query:

WHERE *age* = $E_K(40)$ AND *ethnicity* = $E_K(2)$

↓

WHERE *age* = *DT*(40) AND *ethnicity* = *DT*(2)

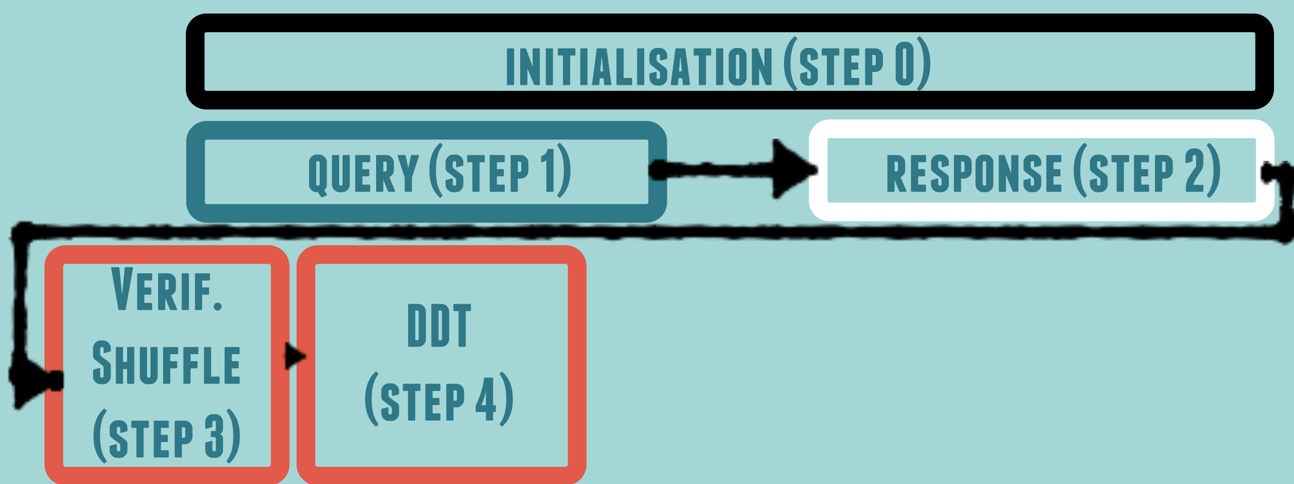
Data:

[[E_K(1)], [E_K(40),E_K(2)], [E_K(23),E_K(1)]

↓

[[DT(1)], [DT(40),DT(2)], [E_K(23),E_K(1)]

Workflow - DDT (Step 4)



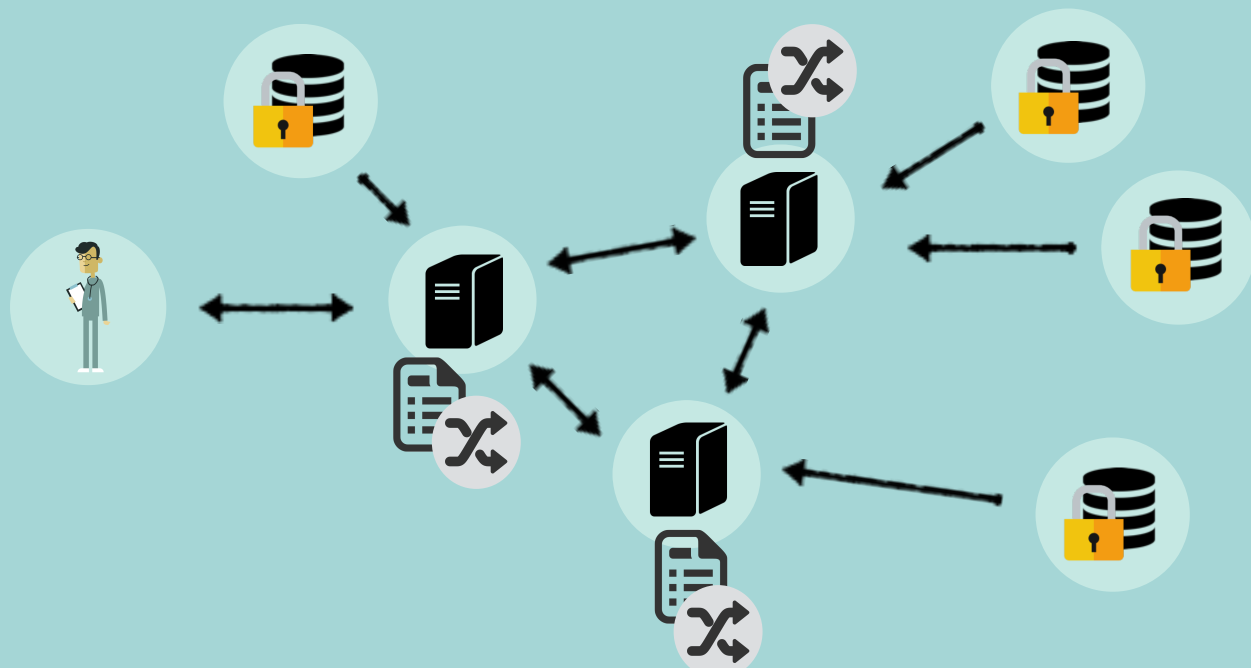
Each server starts a **distributed deterministic tagging protocol**:

In this protocol each server sequentially:

- **partially decrypt** the ciphertexts
- **Blinds** the message by multiplying the ciphertexts with a random **ephemeral secret key**

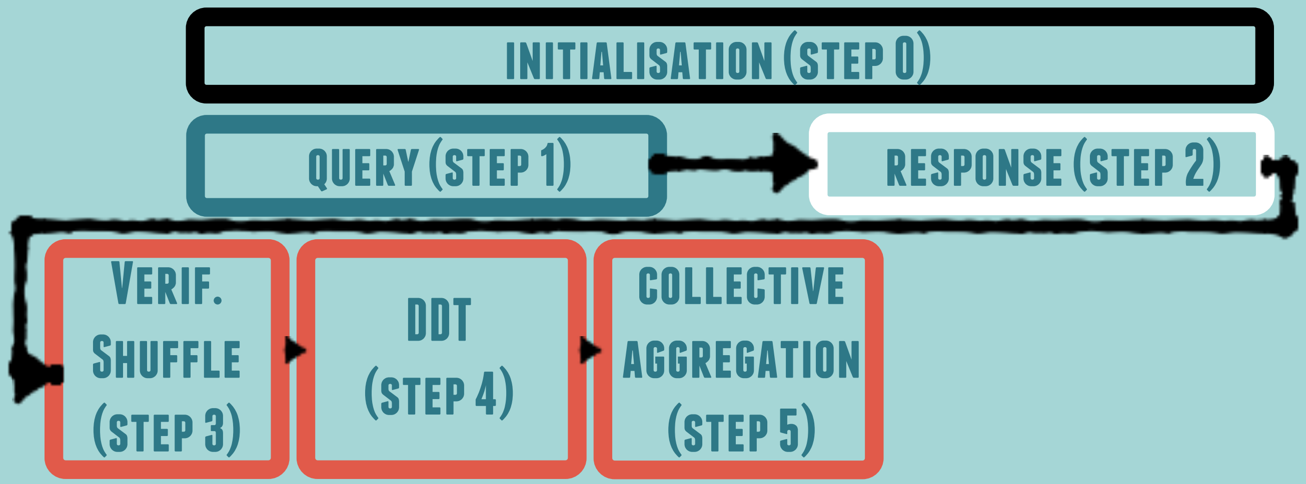
⟶ deterministic tag depending on the value of the encrypted message

All operations are done with zero-knowledge proofs from Camenisch et al.

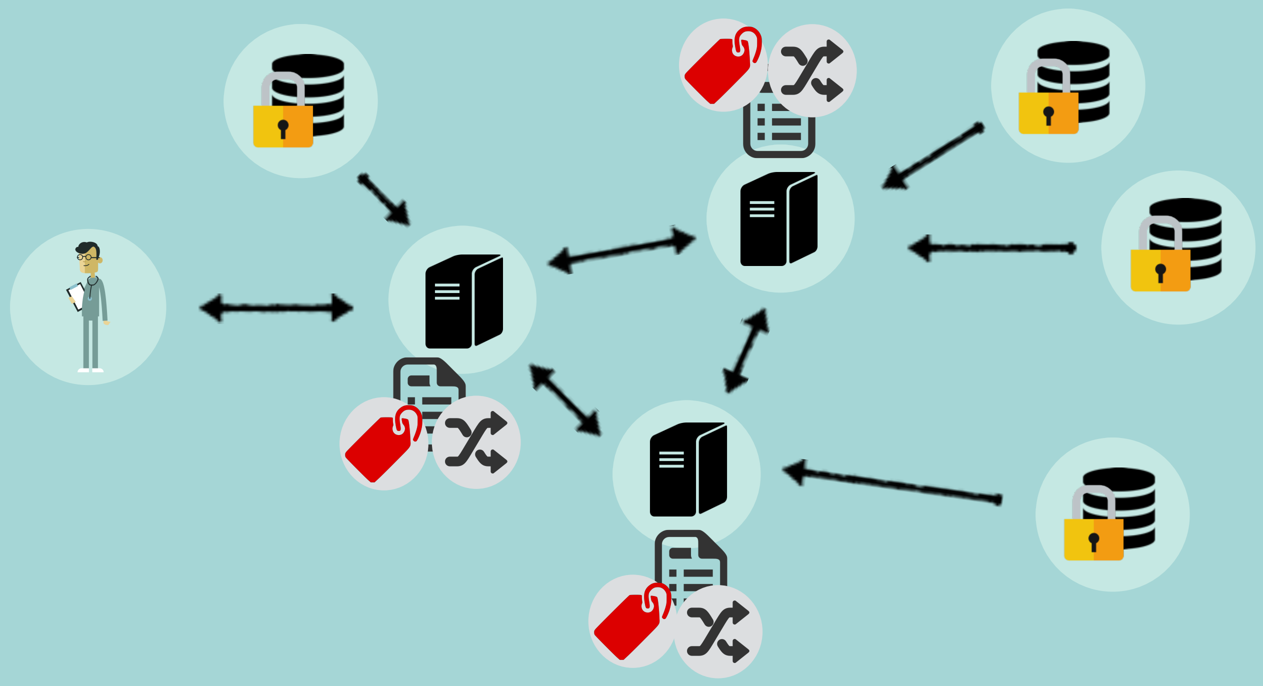


[1] Jan Camenish and Markus Stadler. Proof systems for general statements about discrete logarithms. (1997)

Workflow - Collective Aggr. (Step 5)

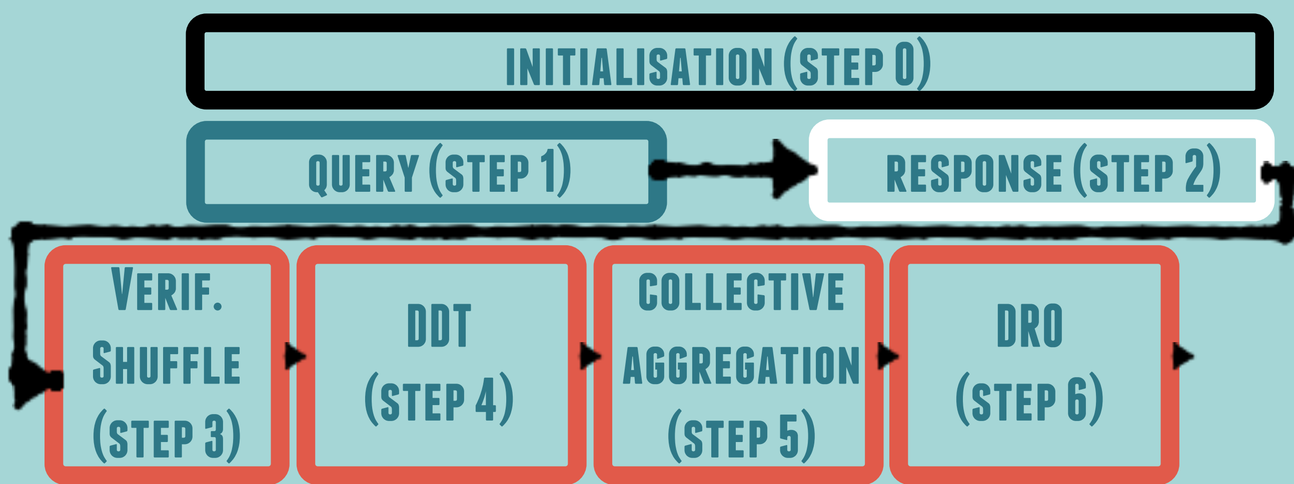


Servers **collectively aggregate** the responses by group.



Proofs consist in publishing the ciphertexts and the result

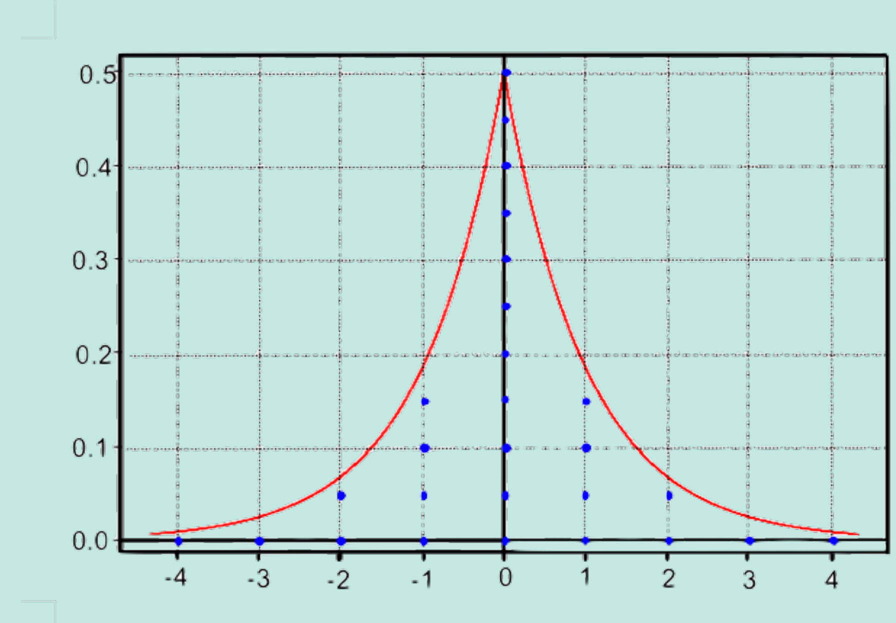
Workflow - DRO (Step 6)



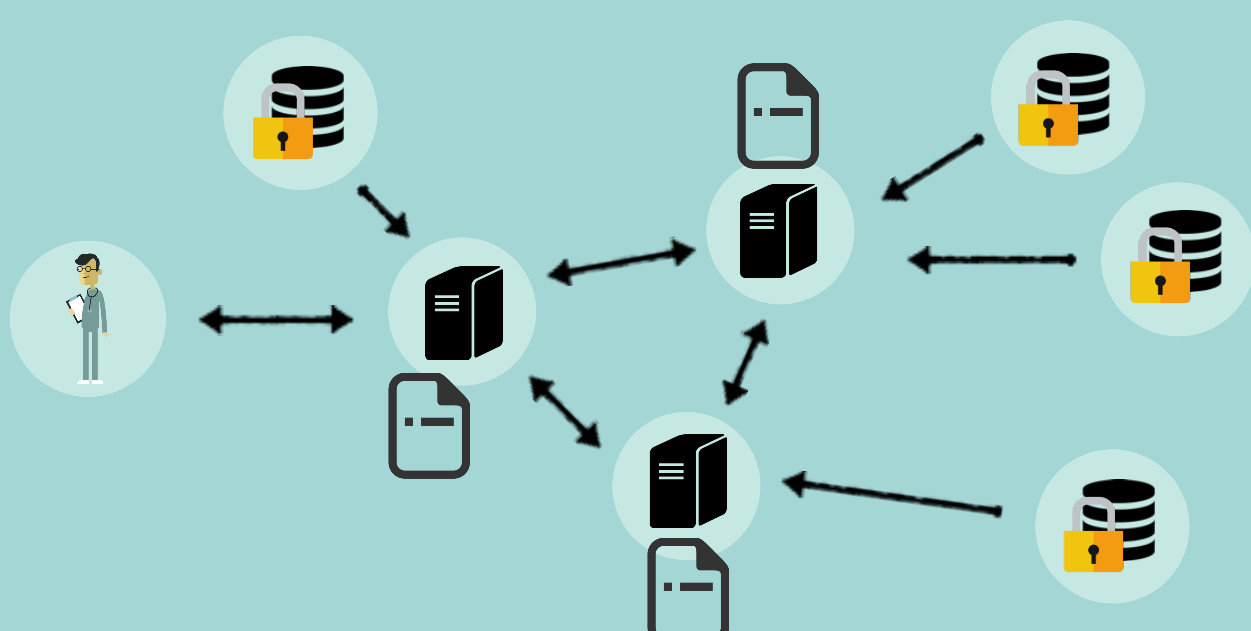
Distributed Results Obfuscation:

<u>Setup:</u>

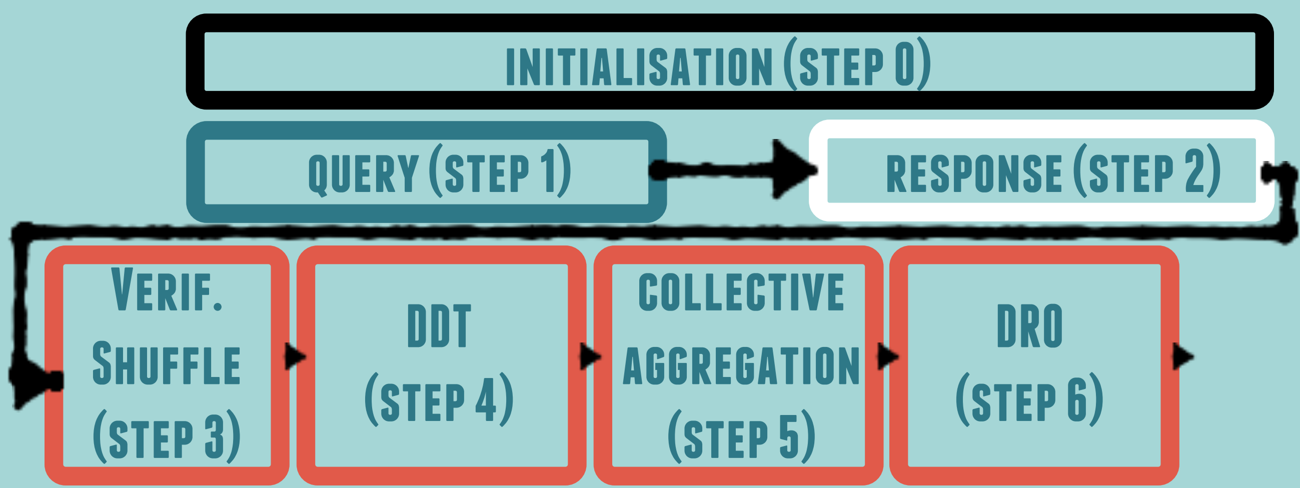
Servers agree on (ϵ,δ)-differential privacy parameters and produce:



→ [0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,...] = list of noise values satisfying (ϵ,δ)-differential privacy.



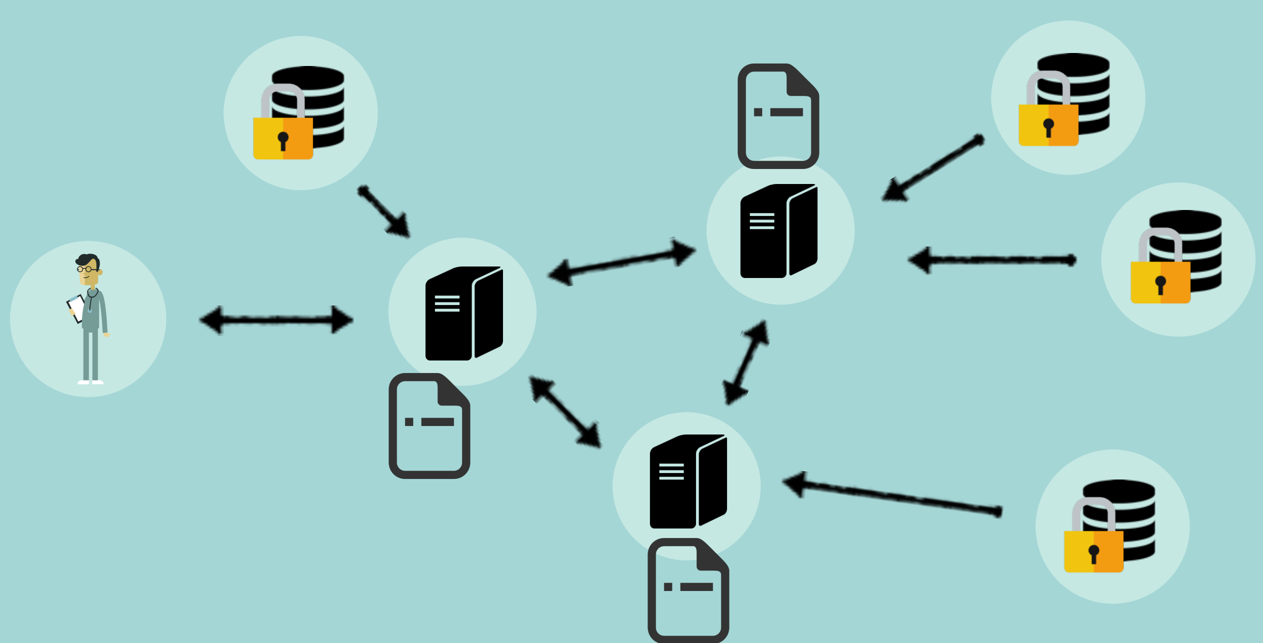
Workflow - DRO (Step 6)



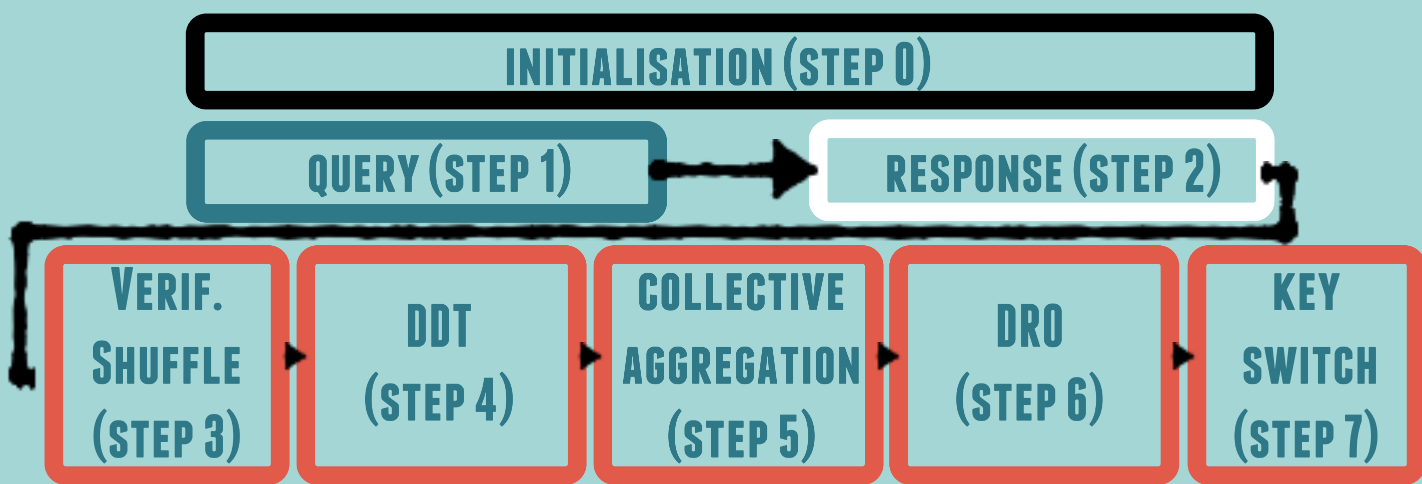
Distributed Results Obfuscation:
<u>Runtime:</u>

- A server starts a **collective shuffling** of the list of noise values
- **adds the first noise value** in the list to the query result.

Oblivious noise addition (shuffling **encrypts and shuffles** the list of noise values).



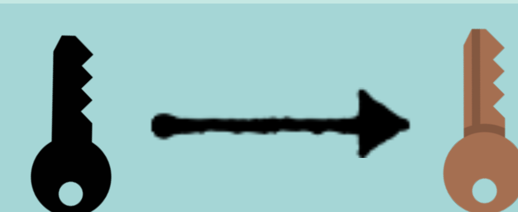
Workflow - Key Switch (Step 7)

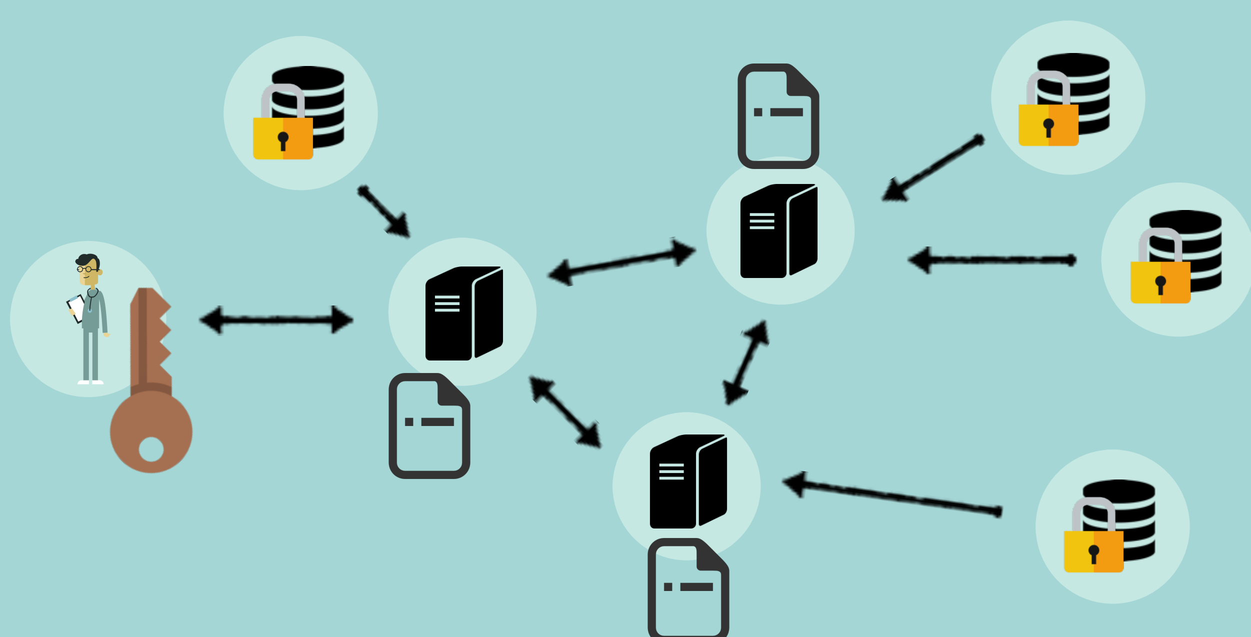


In the **key switch protocol** each server:

- **partially decrypt**
- **encrypt** with a new key all the ciphertexts.

⟶ **Encryption is switched** from the Collective Key to the querier's public key.

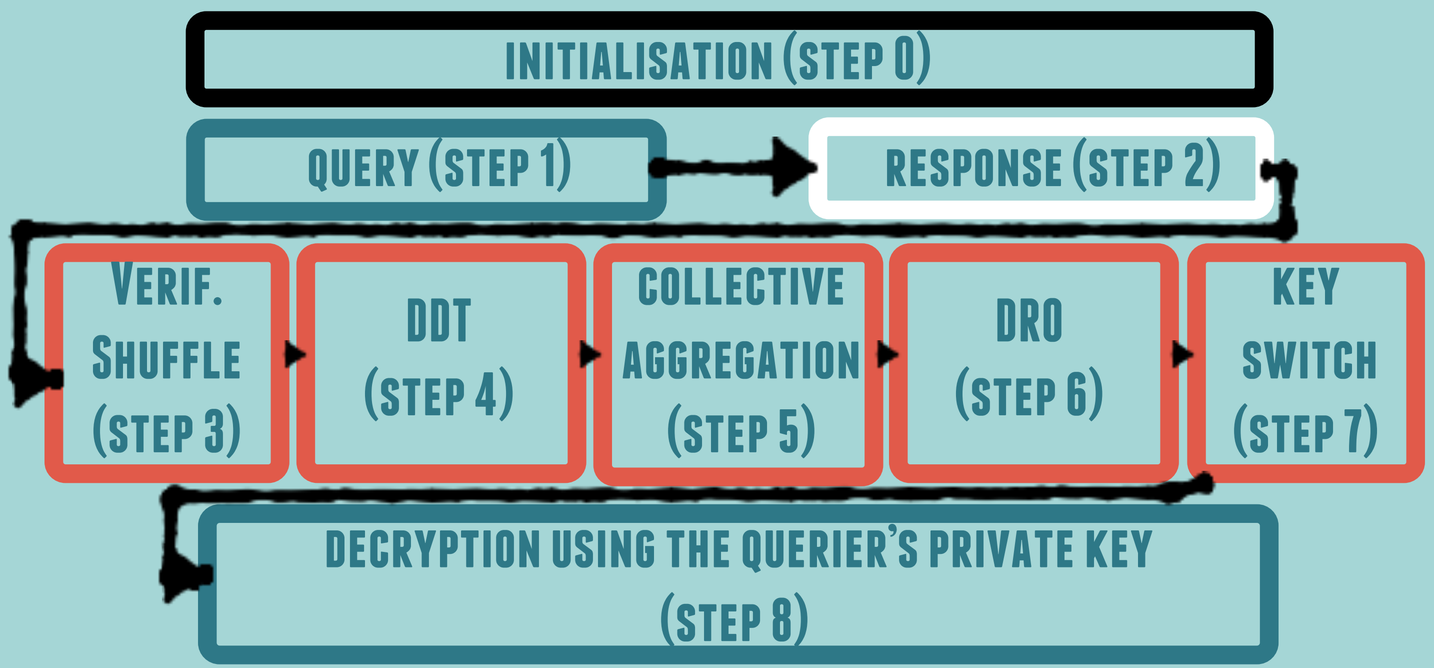




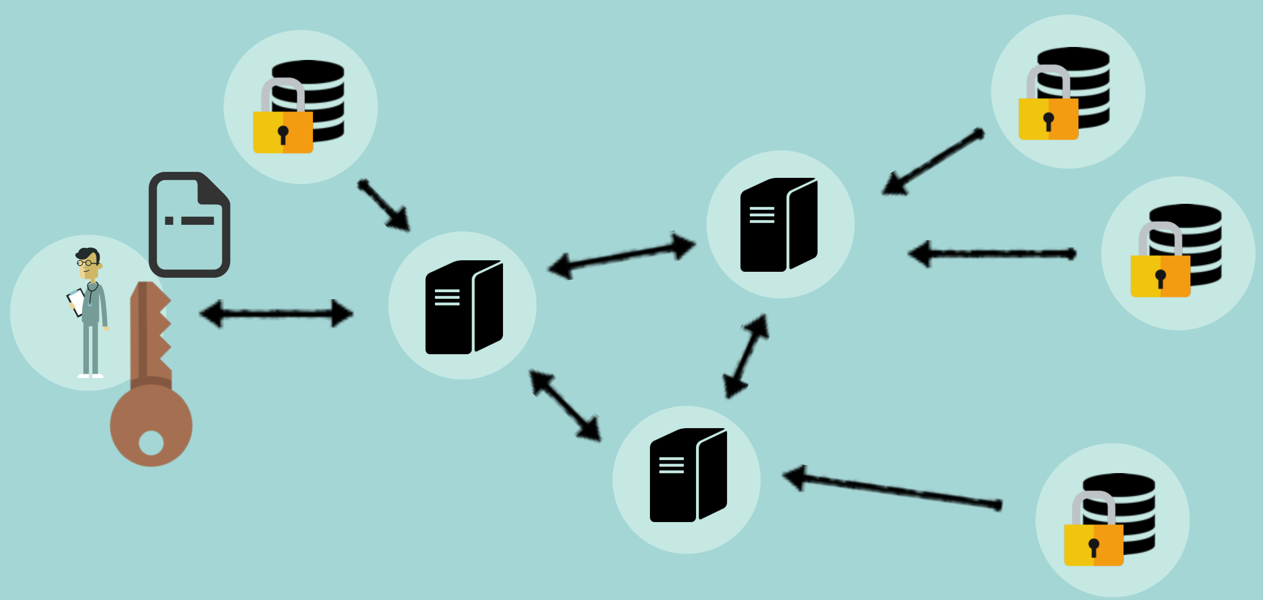
All operations are done with zero-knowledge proofs from Camenish et al.

[1] Jan Camenish and Markus Stadler. Proof systems for general statements about discrete logarithms. (1997)

Workflow - Decryption (Step 8)



Querier **decrypts** the result with his secret key



Performance Evaluation

Servers configuration

- Memory: 256GB RAM
- Processor: Intel Xeon E5-2680 v3 (Haswell)
- **Cores: 24 (with 48 threads)**
- Frequency: 2.5GHz
- Bandwidth capacity: 1Gbps

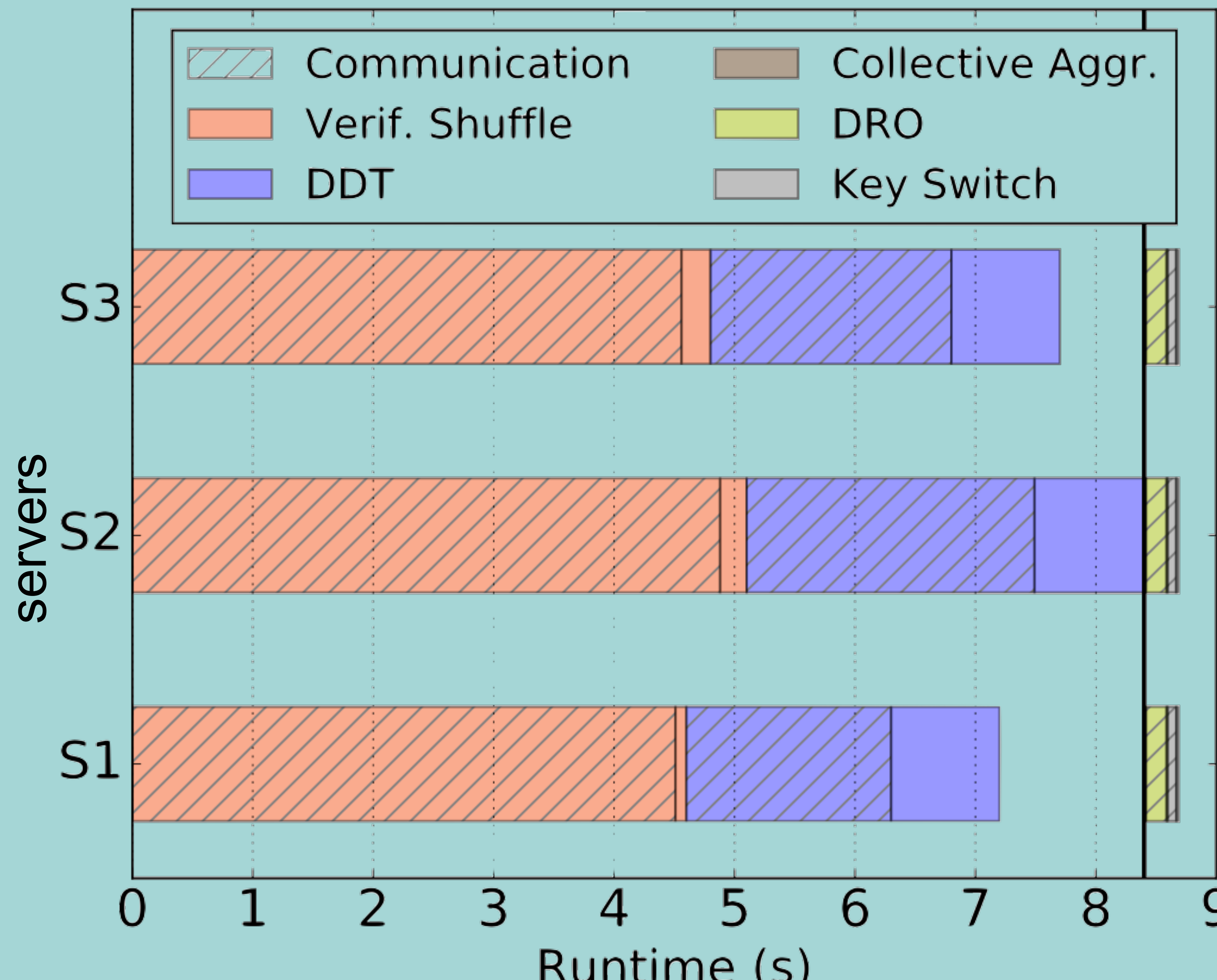
Network and Crypto

- Realistic virtual network emulation tool with 10ms delays btw. servers
- DeDiS' Onet library
- DeDiS' implementation of Ed25519 Elliptic Curve **(128-bit security)**

Default parameters

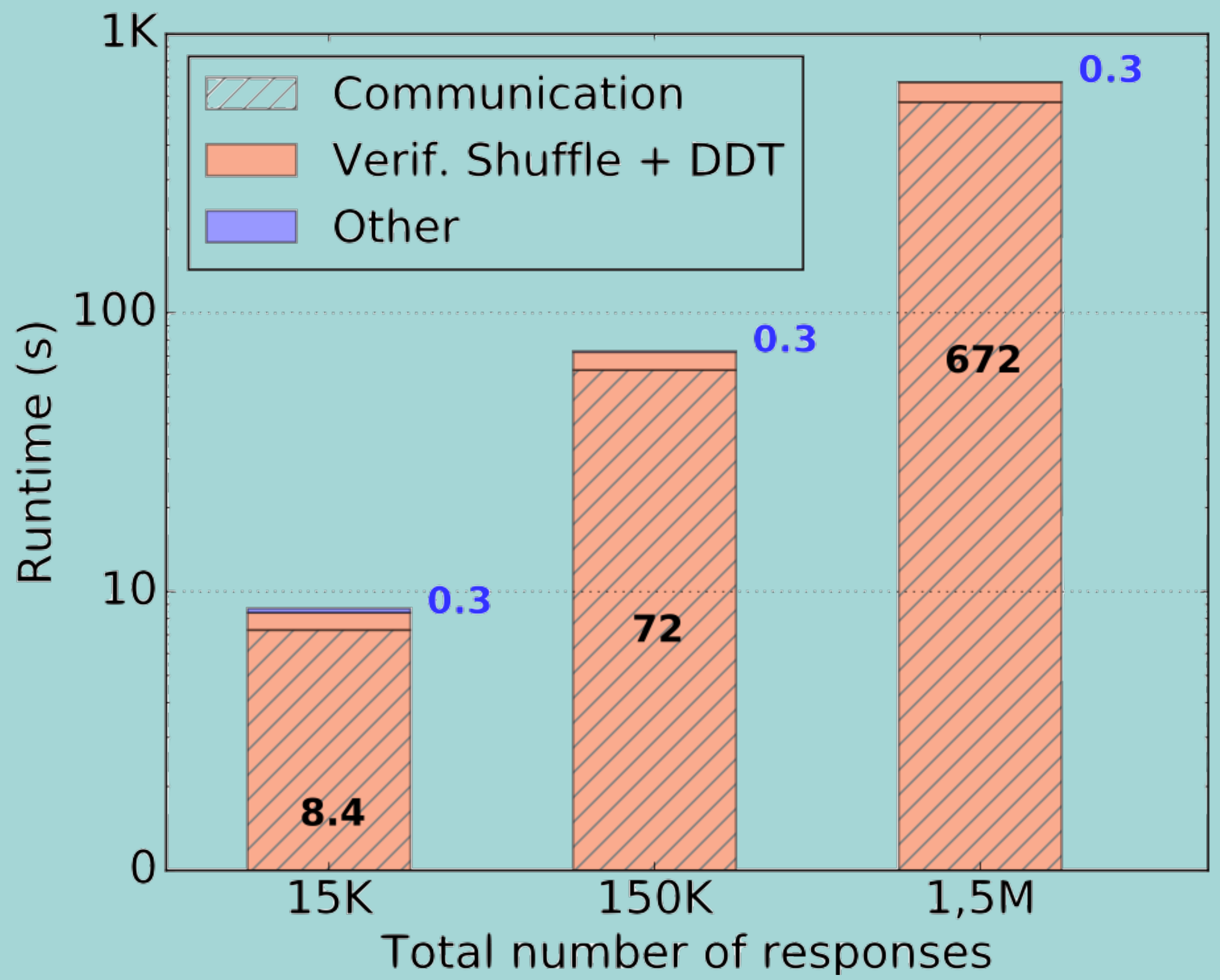
- **3 servers**
- **15,000 responses** in total (equally distributed in servers)
- **1 GROUP BY** attribute with 10 possible values , **1 WHERE** and **10 aggregating attributes**
- 1000 noise values

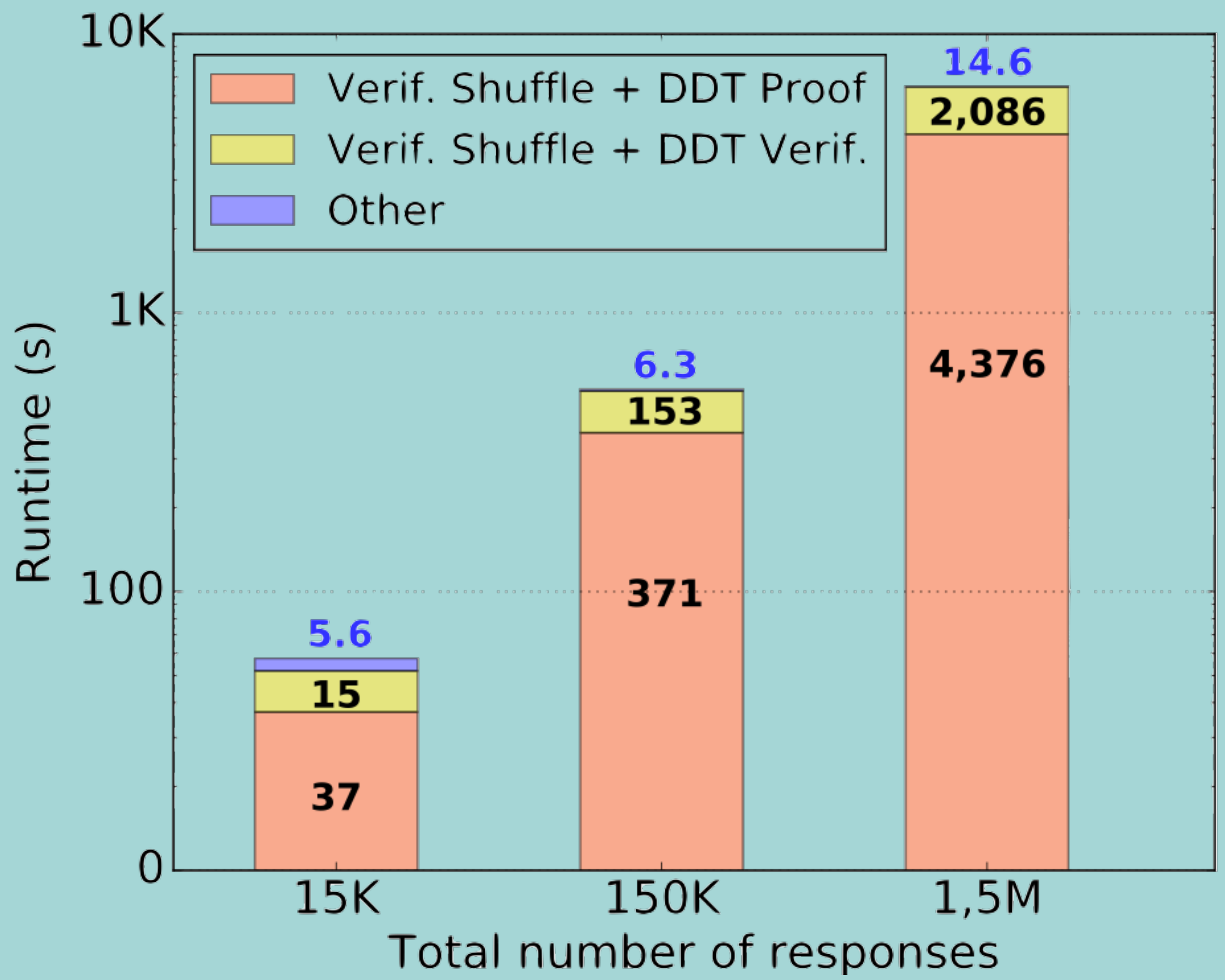
Servers collaboration



DDT = Distrib. Deterministic Tagging
DRO = Distrib. Results Obfuscation

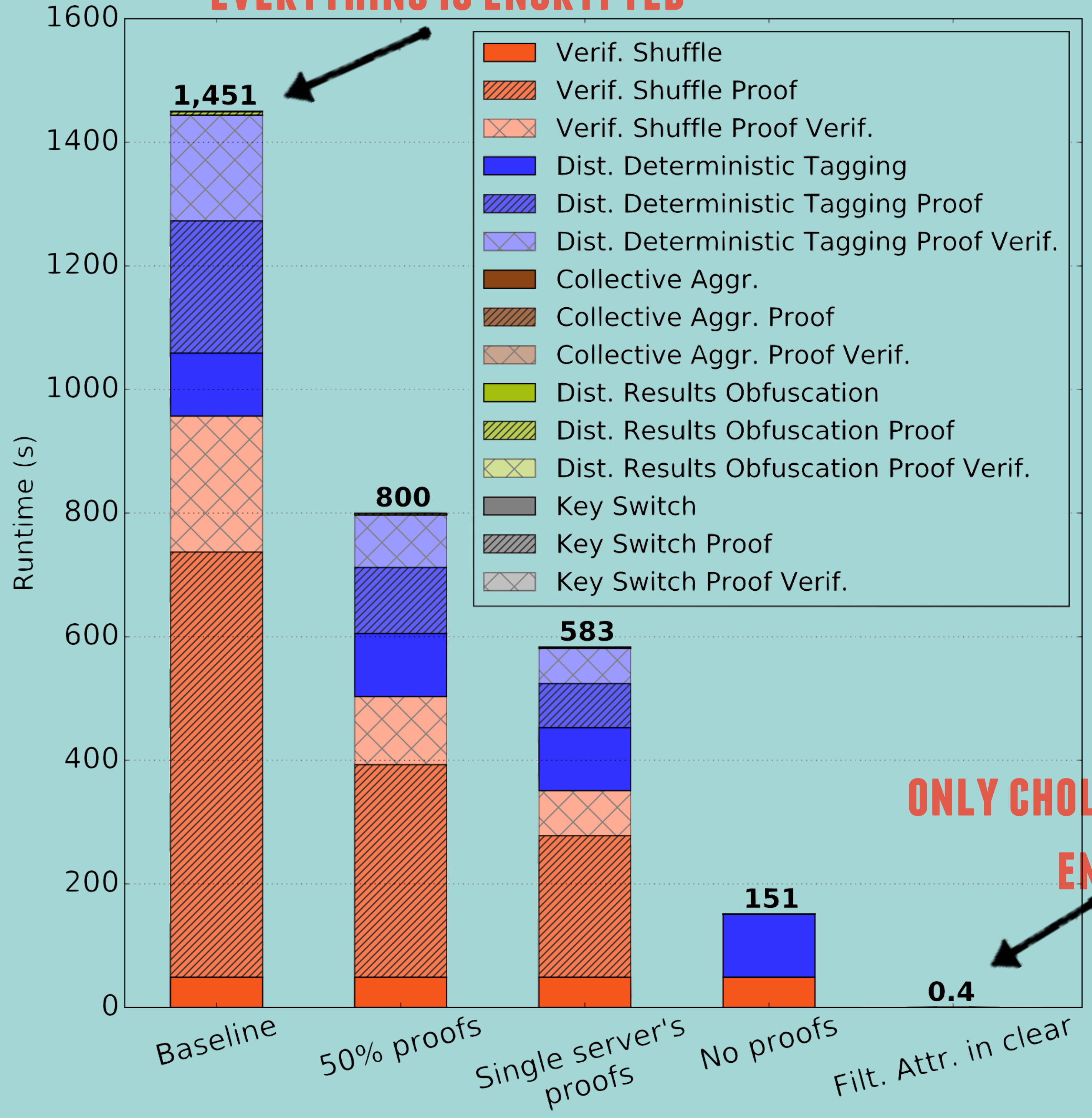
Runtime vs. nbr. of responses





Performance/Security Tradeoffs

EVERYTHING IS ENCRYPTED



ONLY CHOLESTEROL_RATE IS ENCRYPTED

SELECT SUM (cholesterol_rate) , COUNT(*)

FROM DP_1,...,DP_{20}

WHERE age IN [40:50] AND ethnicity = caucasian

GROUP BY gender

3 servers
400K responses with
- 1 GROUP BY attribute
- 2 WHERE attributes
- 2 aggregating attributes

Conclusion

A Decentralized System for Privacy-Conscious Data Sharing

- SQL statistical queries based on Boolean conditions
- Strongest-link security
- Data confidentiality
- Distributed differential privacy
- Distributed deterministic tagging of probabilistic ciphertexts
- Collective encryption key switching

- Runtime linear with the amount of data to process

github.com/lca1/unlynx

david.froelicher@epfl.ch