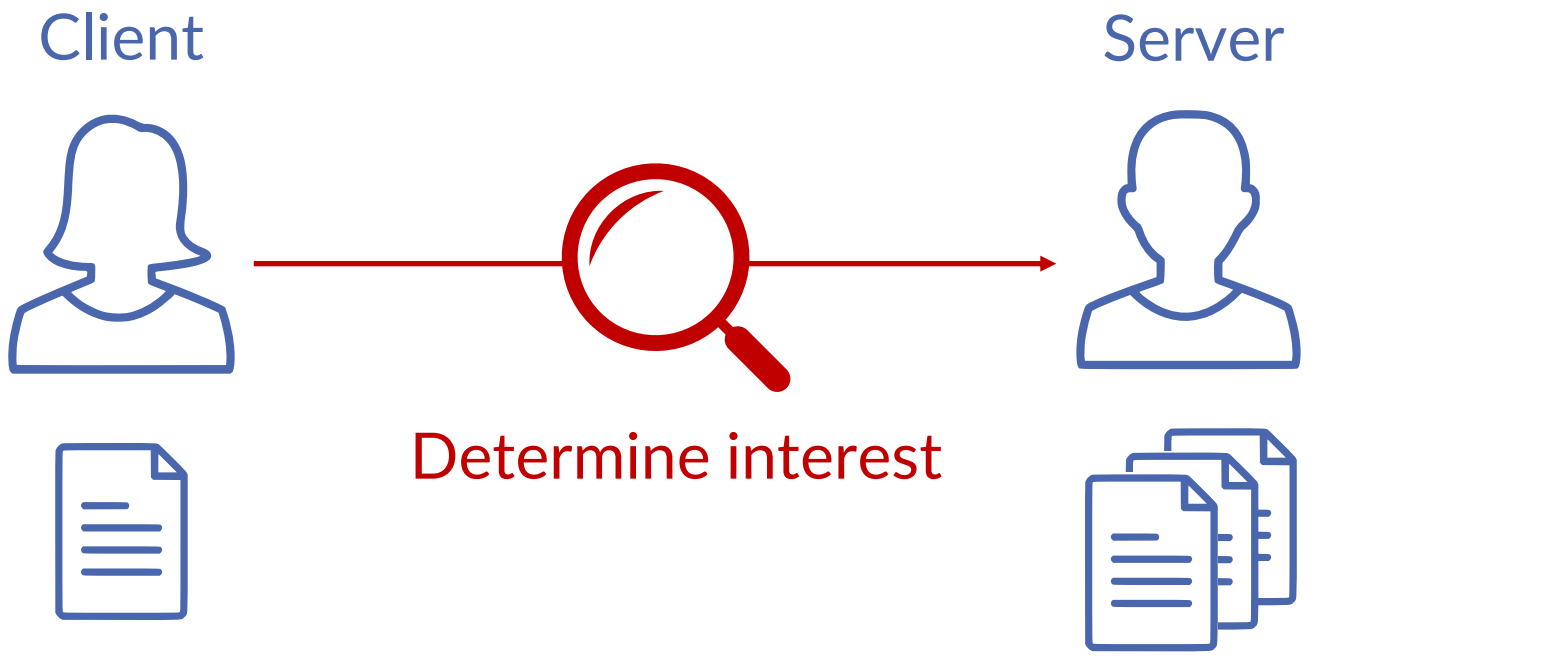
# PCM: Private collection matching protocols

Presented at NIST WPEC 2024 on September 24.

Kasra EdalatNejad (TU-Darmstadt/EPFL), Mathilde Raynal (EPFL), Wouter Lueks (CISPA), Carmela Troncoso (EPFL)

# Many similar problems

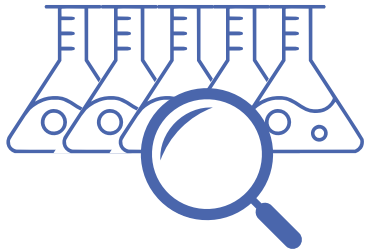



A single set

A collection of many sets

# Collection matching problems

Document search

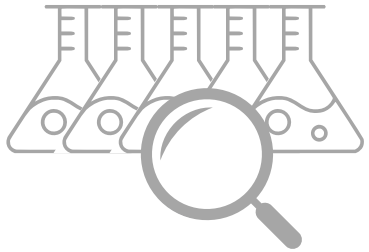


Chemical search

Mobile dating

# Collection matching problems

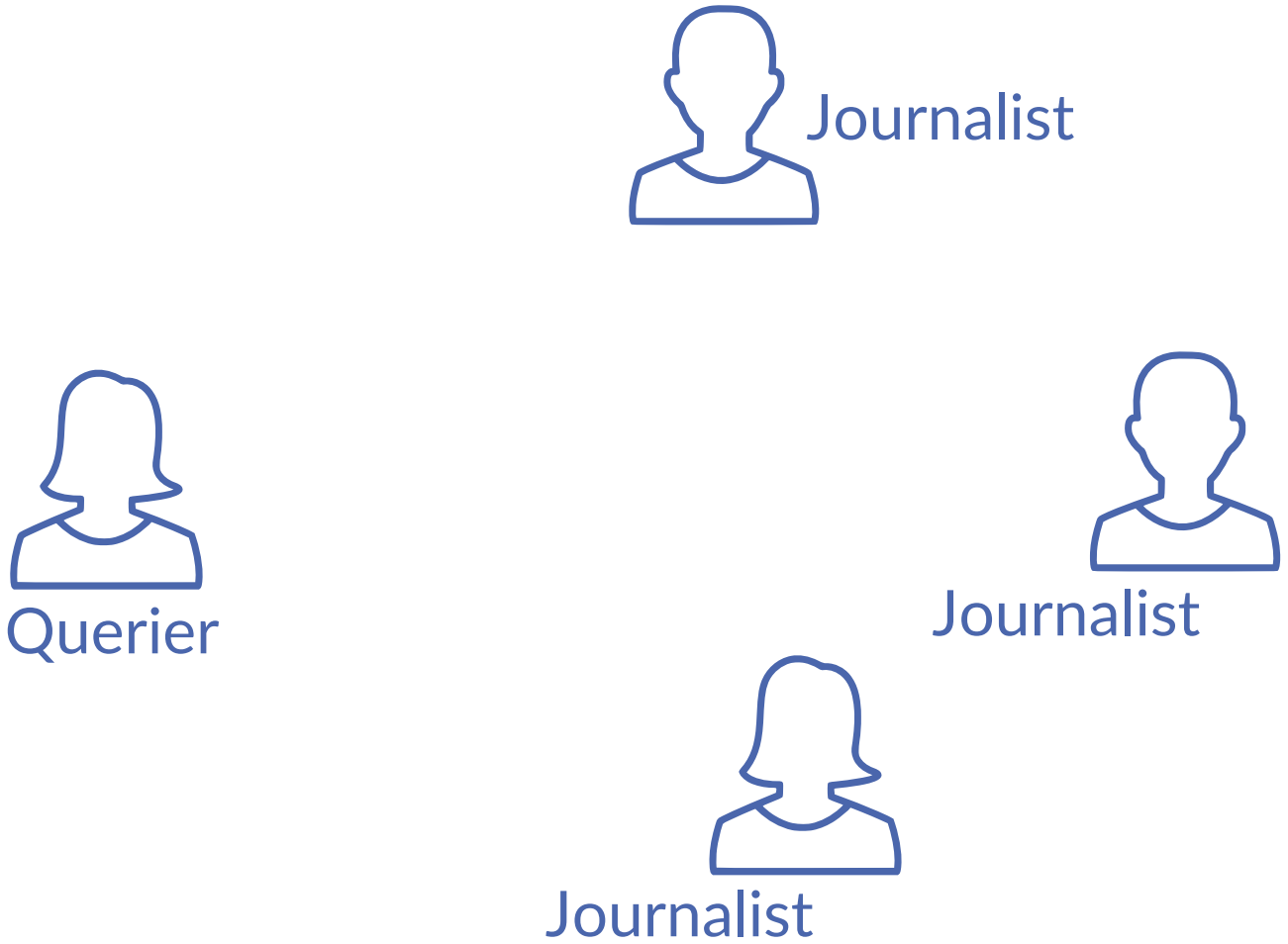
Document search



Chemical search

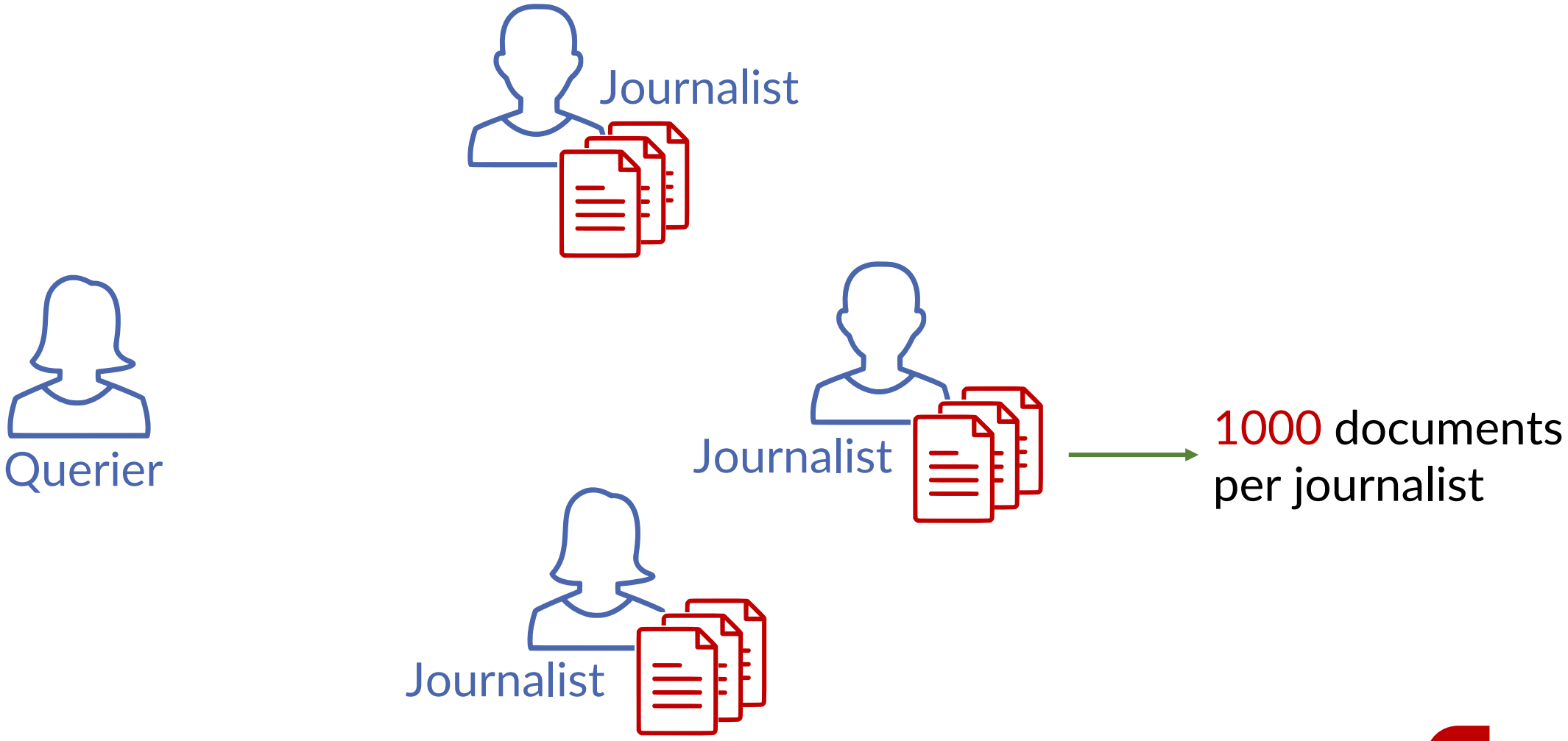
Mobile dating

# Document search for investigative journalists

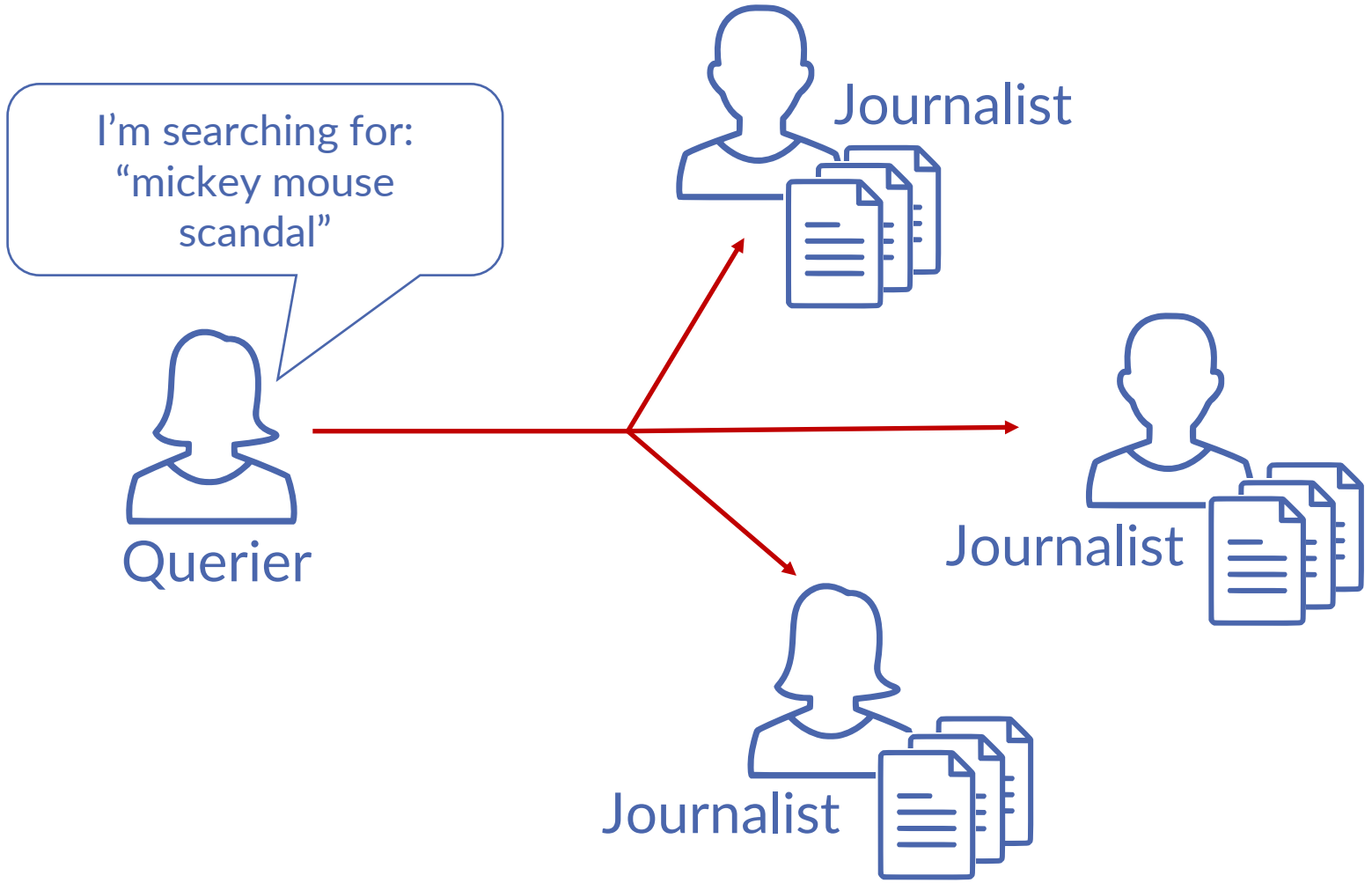



Kasra Edalatnejad, Wouter Lueks, Julien Pierre Martin, Soline Ledésert, Anne L'Hôte, Bruno Thomas, Laurent Girod, Carmela Troncoso: DatashareNetwork: A Decentralized Privacy-Preserving Search Engine for Investigative Journalists. USENIX Security Symposium 2020

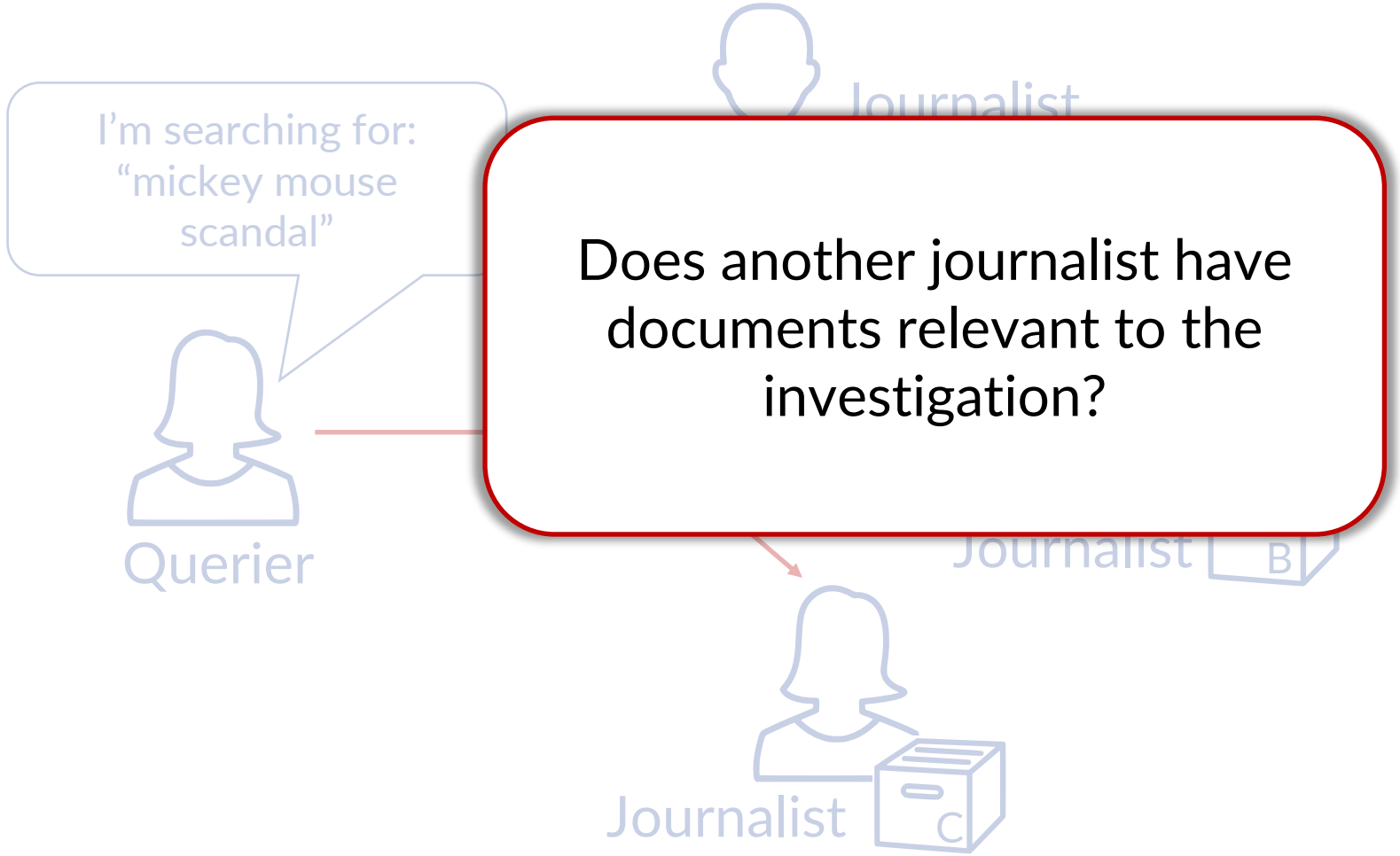
# Document search for journalists



# Document search for journalists

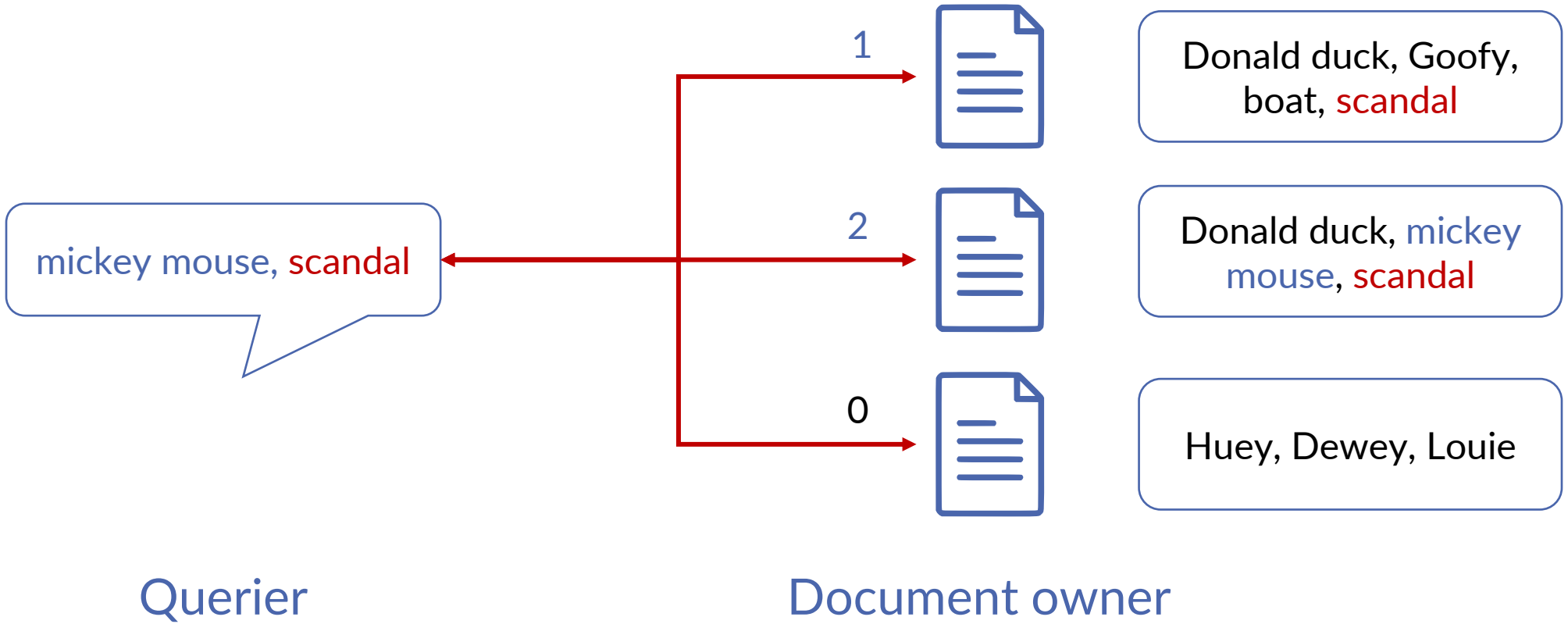


# Document search for journalists



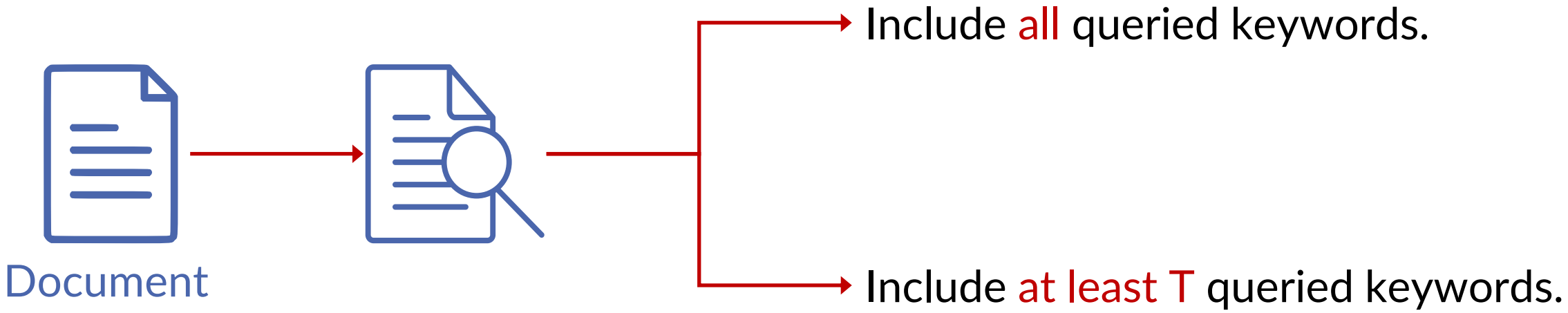
# Document search

# Private set intersection cardinality

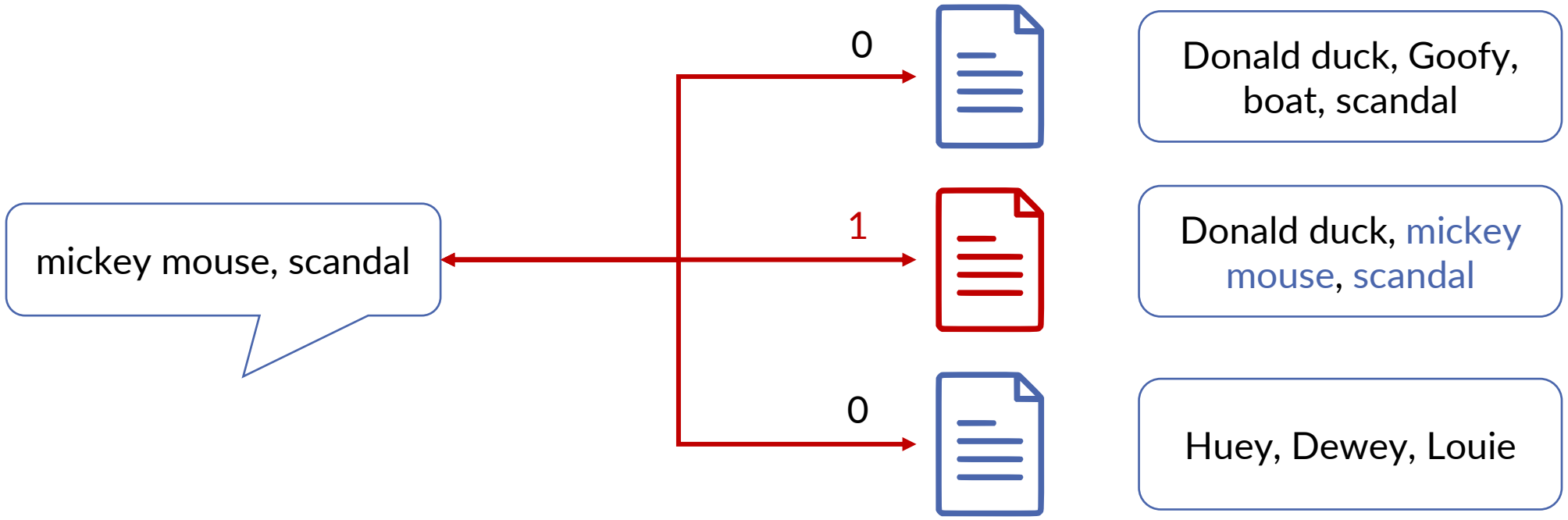


# Is revealing cardinality needed?

# Is revealing cardinality needed?



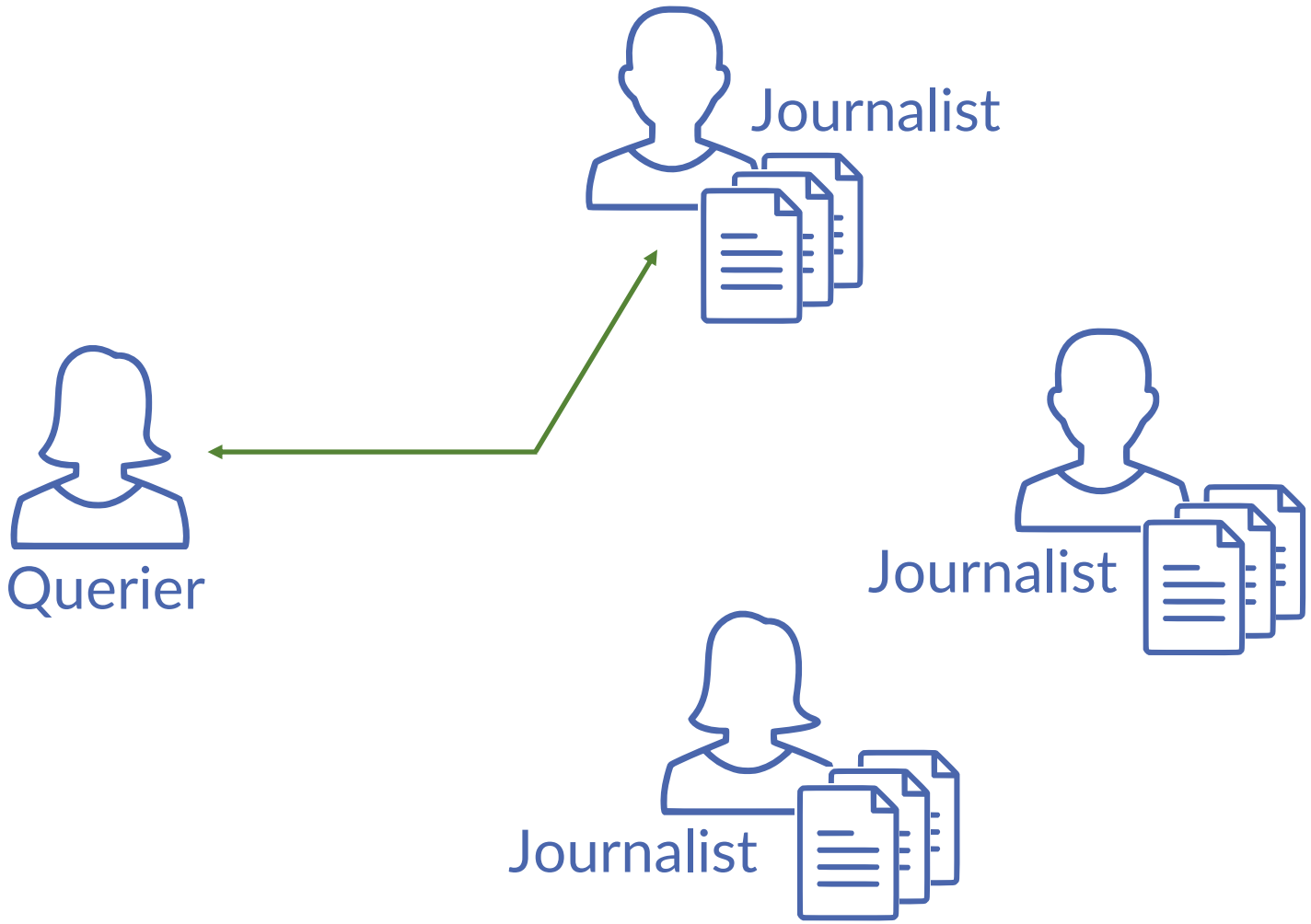
# Document relevance



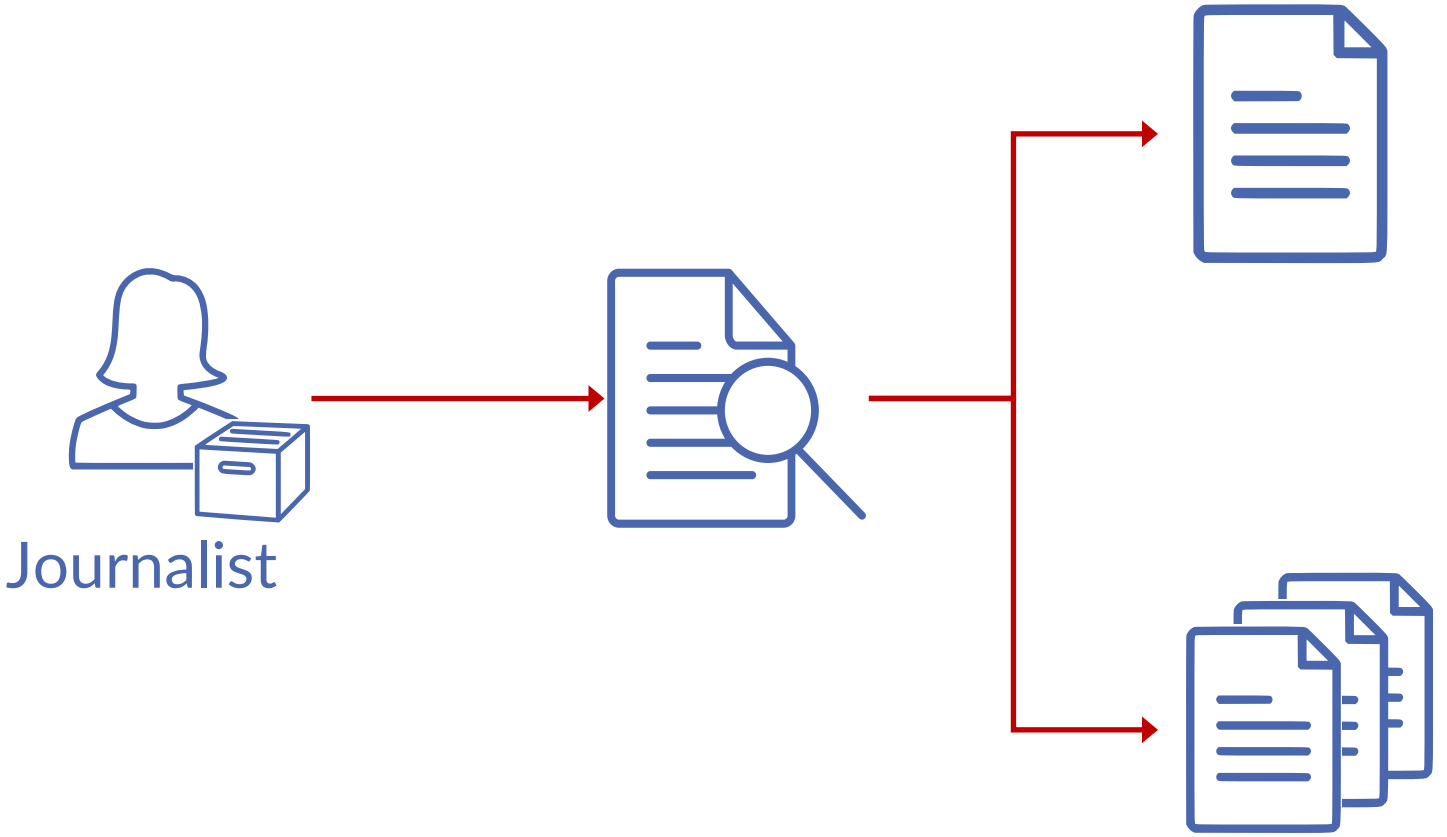
# After search

No retrieval

Screening talk



# Private collection search




Has at least 1 relevant document.

Has T relevant documents.

# A new class of problems

Common properties

- Clients want to compare their one set with all sets at the server.
- Clients do not need per-server set results, only an aggregated output.
- Clients and server want privacy.

# A new class of problems

## Common properties

- Clients want to compare their one set with all sets at the server.
- Clients do not need per-server set results, only an aggregated output.
- Clients and server want privacy.

## Differences

- (Matching) When a server set is of interest to the client?
- (Aggregation) How to combine individual set matching result?

# Requirements

# Requirements

1. Flexible matching criteria
   - Without revealing intermediate values such as intersections
   - Examples: all (or a threshold of) queried keywords are in a document

# Requirements

1. Flexible matching criteria
   - Without revealing intermediate values such as intersections
   - Examples: all (or a threshold of) queried keywords are in a document
2. Aggregate many-set response
   - Without leaking information about individual sets
   - Examples: at least 1 matching set exists, how many sets are of interest?

# Requirements

1. Flexible matching criteria
   - Without revealing intermediate values such as intersections
   - Examples: all (or a threshold of) queried keywords are in a document
2. Aggregate many-set response
   - Without leaking information about individual sets
   - Examples: at least 1 matching set exists, how many sets are of interest?
3. Extreme imbalance
   - Clients have limited computation and communication power
   - The size of the servers input can be 1,000,000 larger than client's input
   - Example: searching a database of 1 million compounds

# Why existing PSI does not work?

| |
|---|
| Comparison-based (OT) |
| Oblivious pseudorandom function |
| Oblivious polynomial evaluation |

# Why existing PSI does not work?

| | Privacy |
|---|---|
| Comparison-based (OT) | ✗ |
| Oblivious pseudorandom function | ✗ |
| Oblivious polynomial evaluation | ✗ |

## Privacy

Do not reveal intersection or its cardinality

Do not reveal per-set result

# Why existing PSI does not work?

| | Privacy |
|---|---|
| Comparison-based (OT) | ✗ |
| Oblivious pseudorandom function | ✗ |
| Oblivious polynomial evaluation | ✗ |
| Circuit-PSI | ✓ |
| Generic SMC | ✓ |

## Privacy

Do not reveal intersection or its cardinality

Do not reveal per-set result

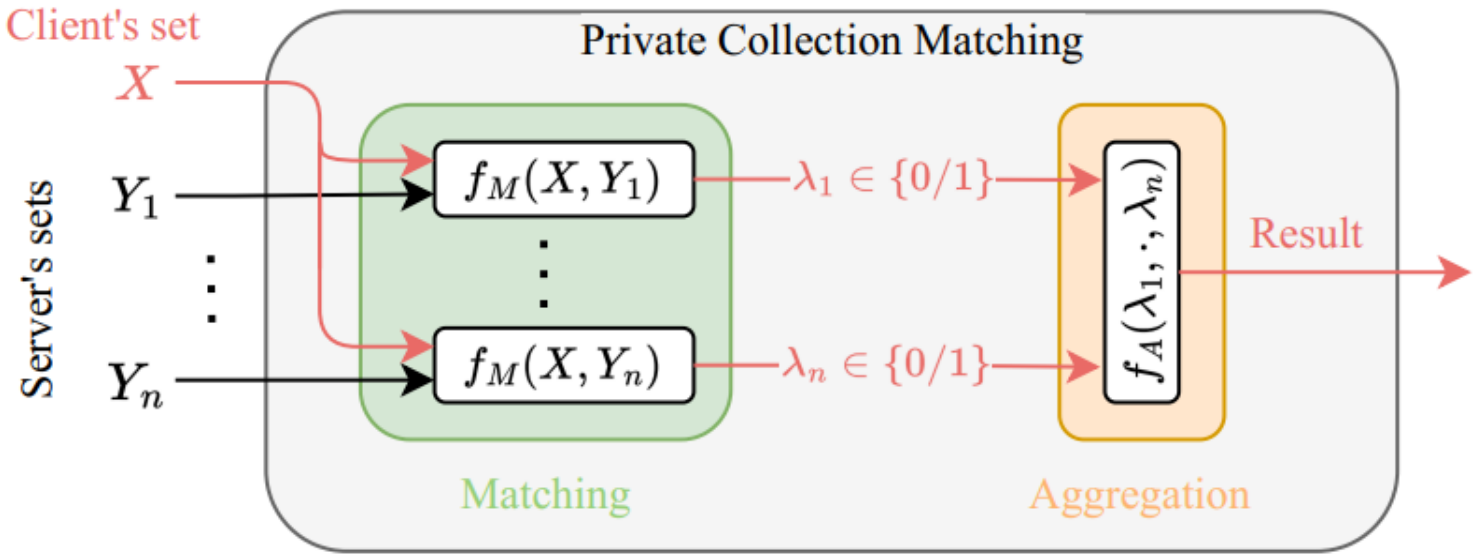
# Why existing PSI does not work?

| | Privacy | Client efficiency |
|---|---|---|
| Comparison-based (OT) | ✗ | ✗ |
| Oblivious pseudorandom function | ✗ | ✓ |
| Oblivious polynomial evaluation | ✗ | ✓ |
| Circuit-PSI | ✓ | ✗ |
| Generic SMC | ✓ | ✗ |

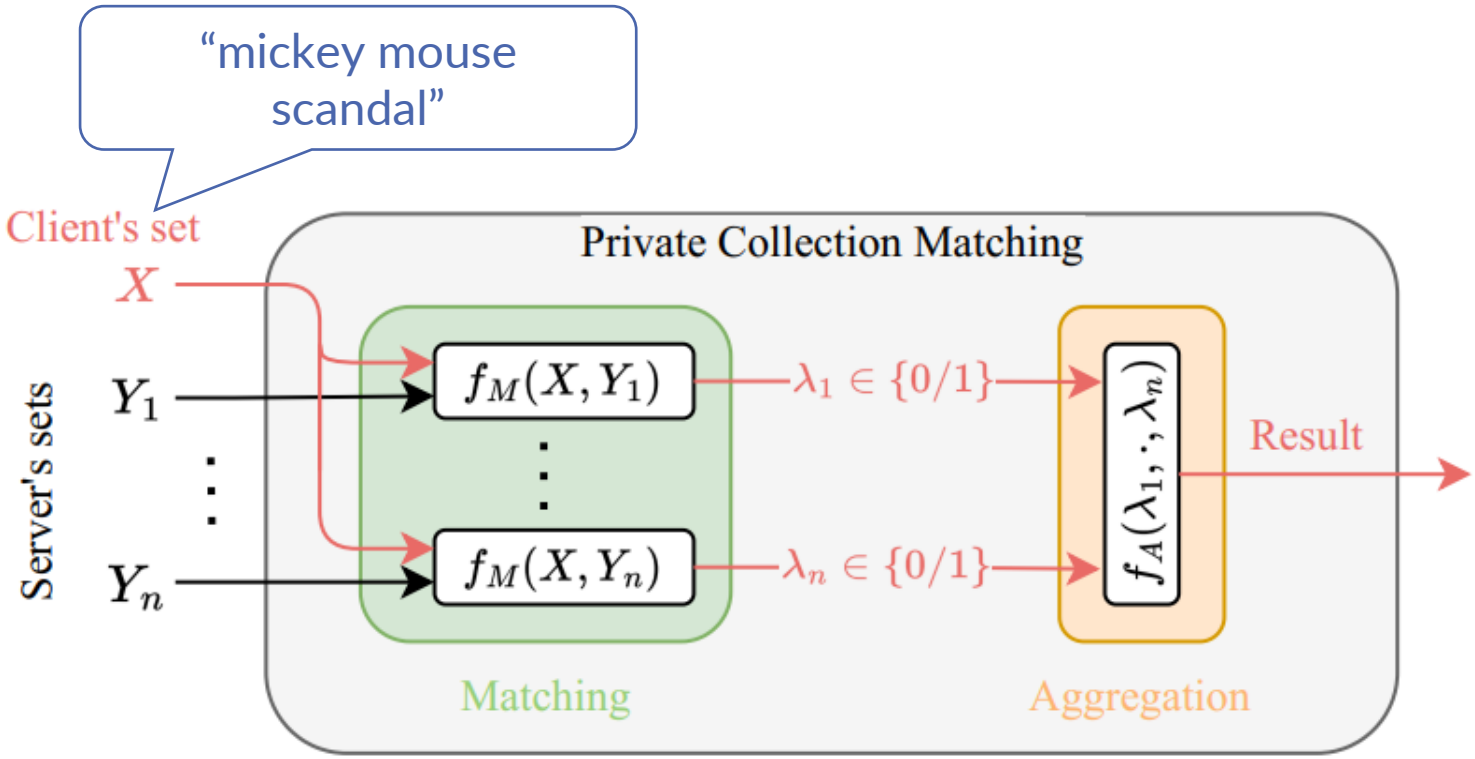
## Client efficiency

Client computation and communication costs should be independent of the server input size

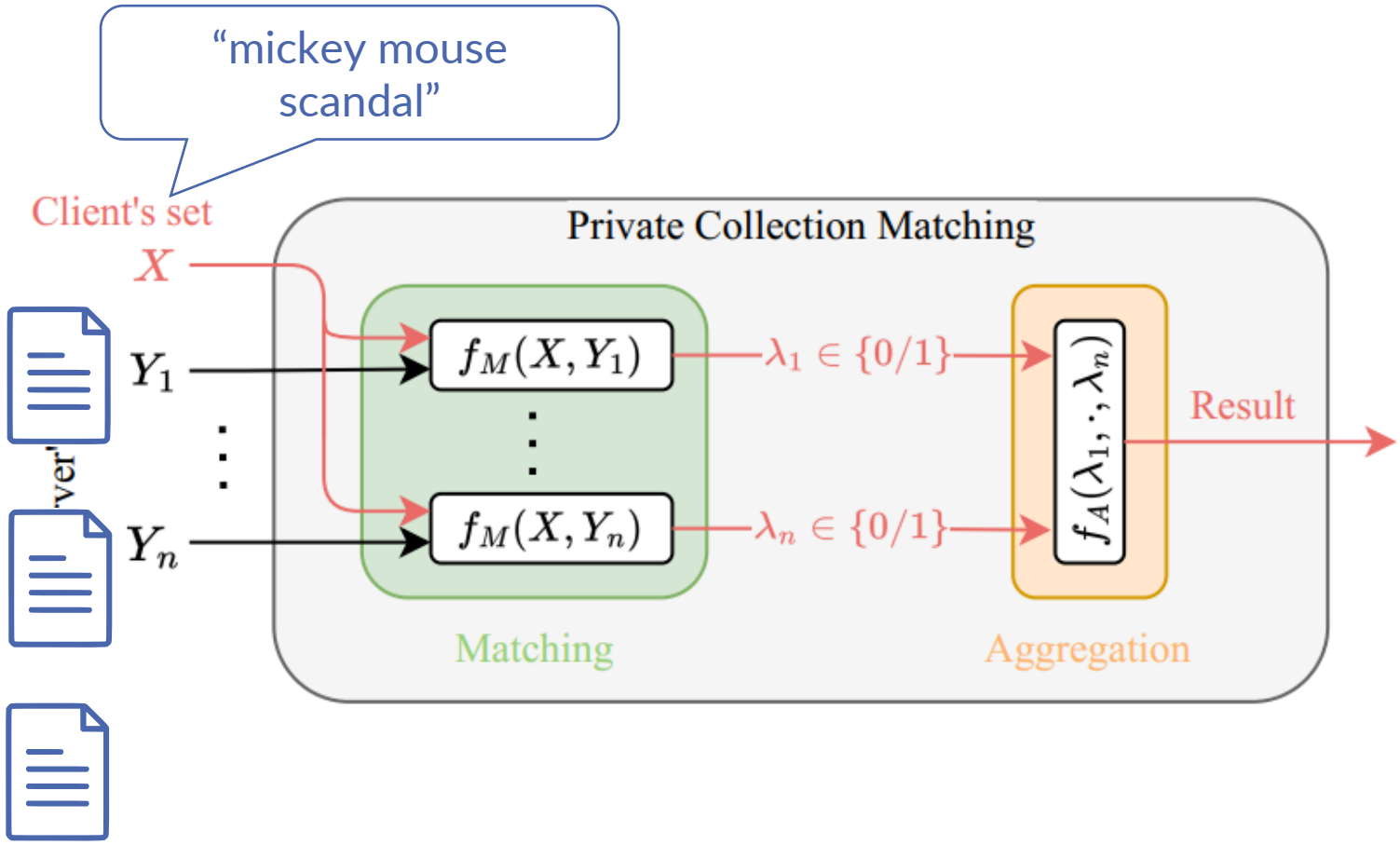
# Private collection matching



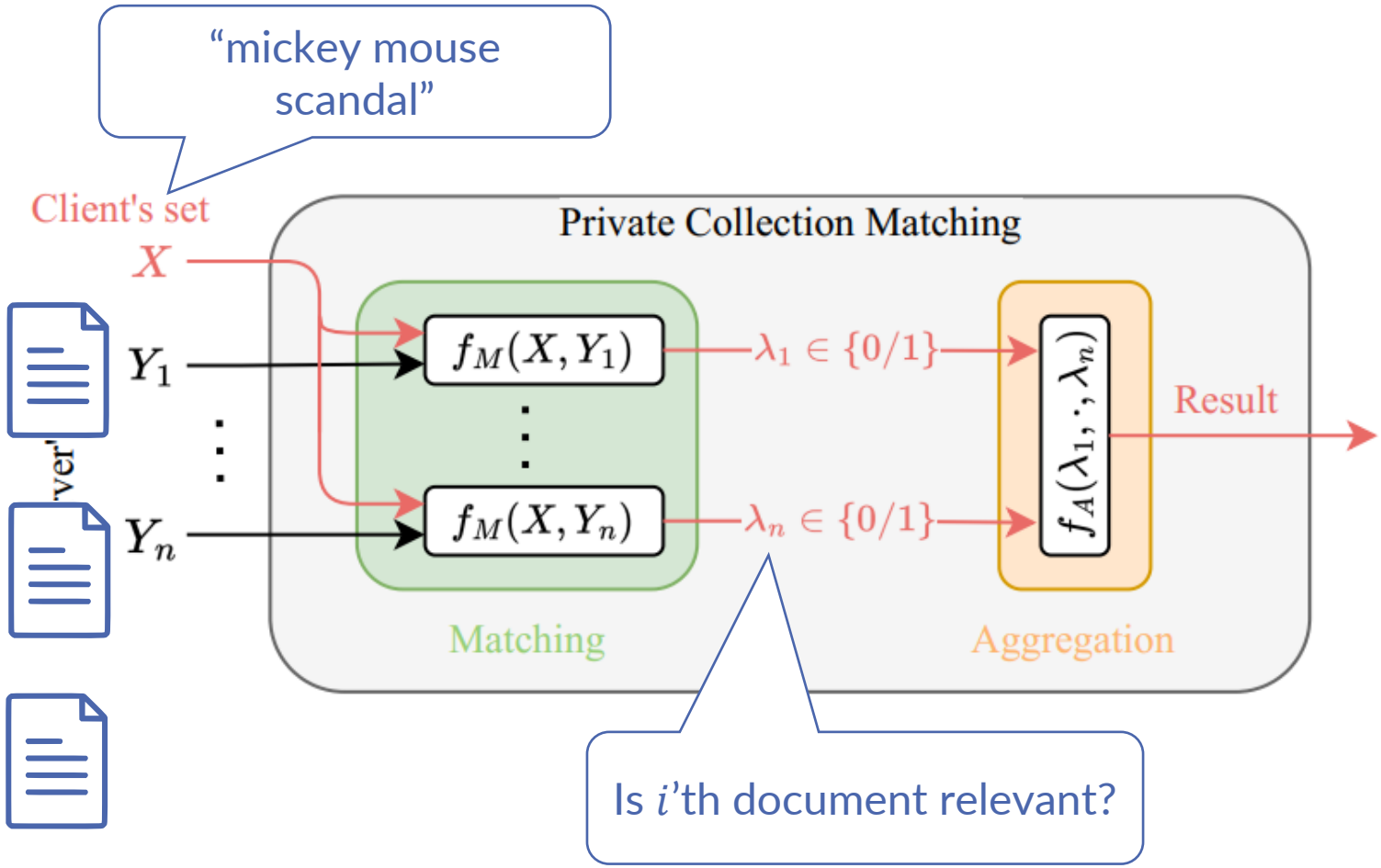
# Private collection matching



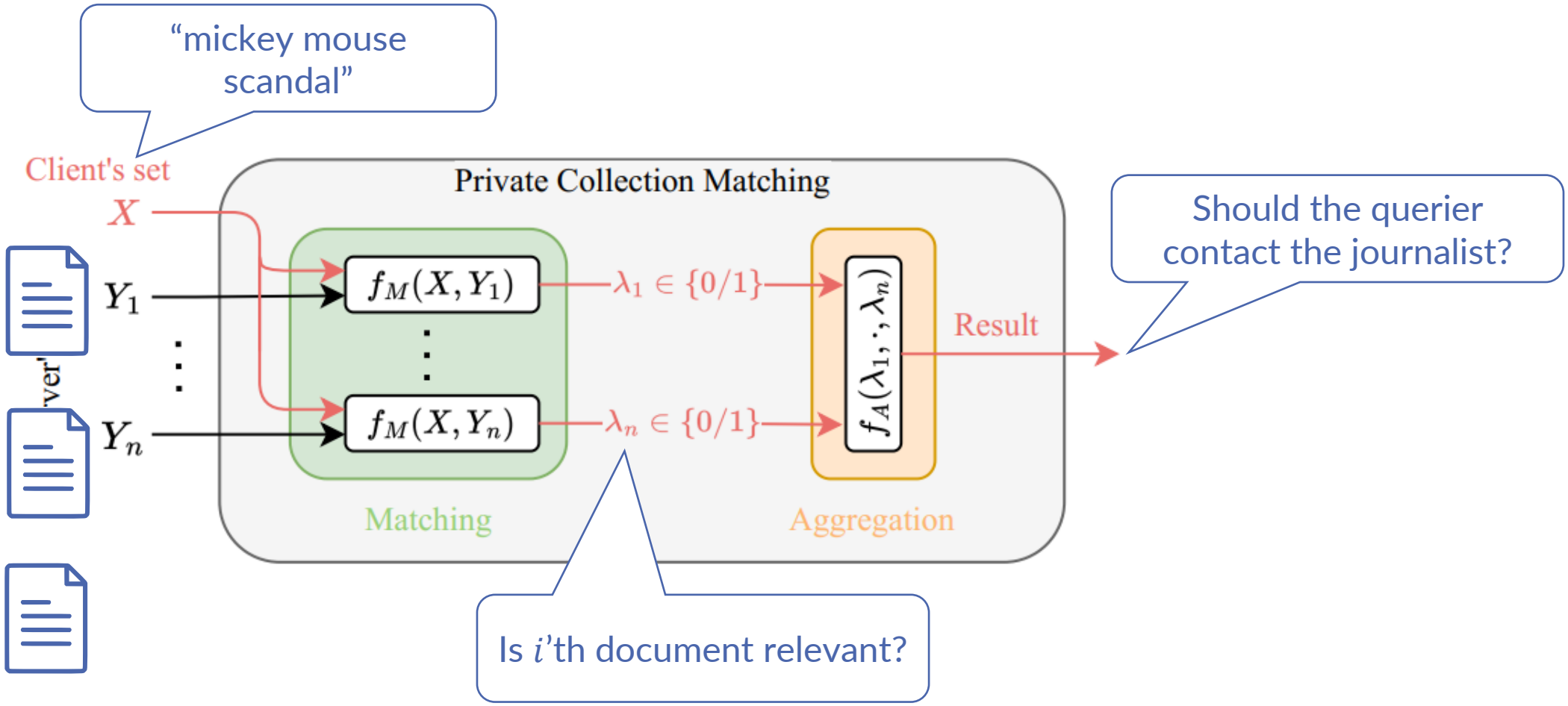
# Private collection matching



# Private collection matching



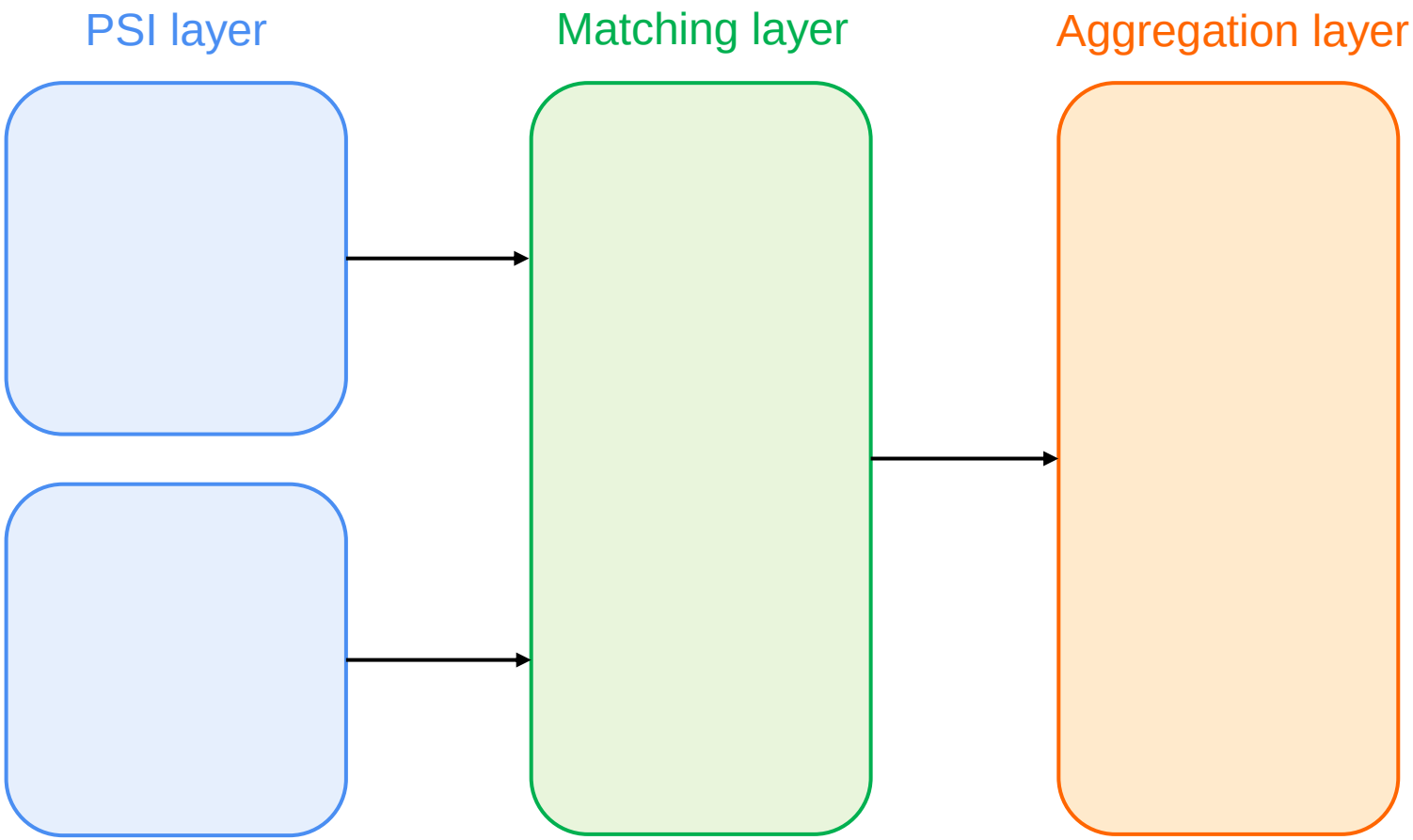
# Private collection matching



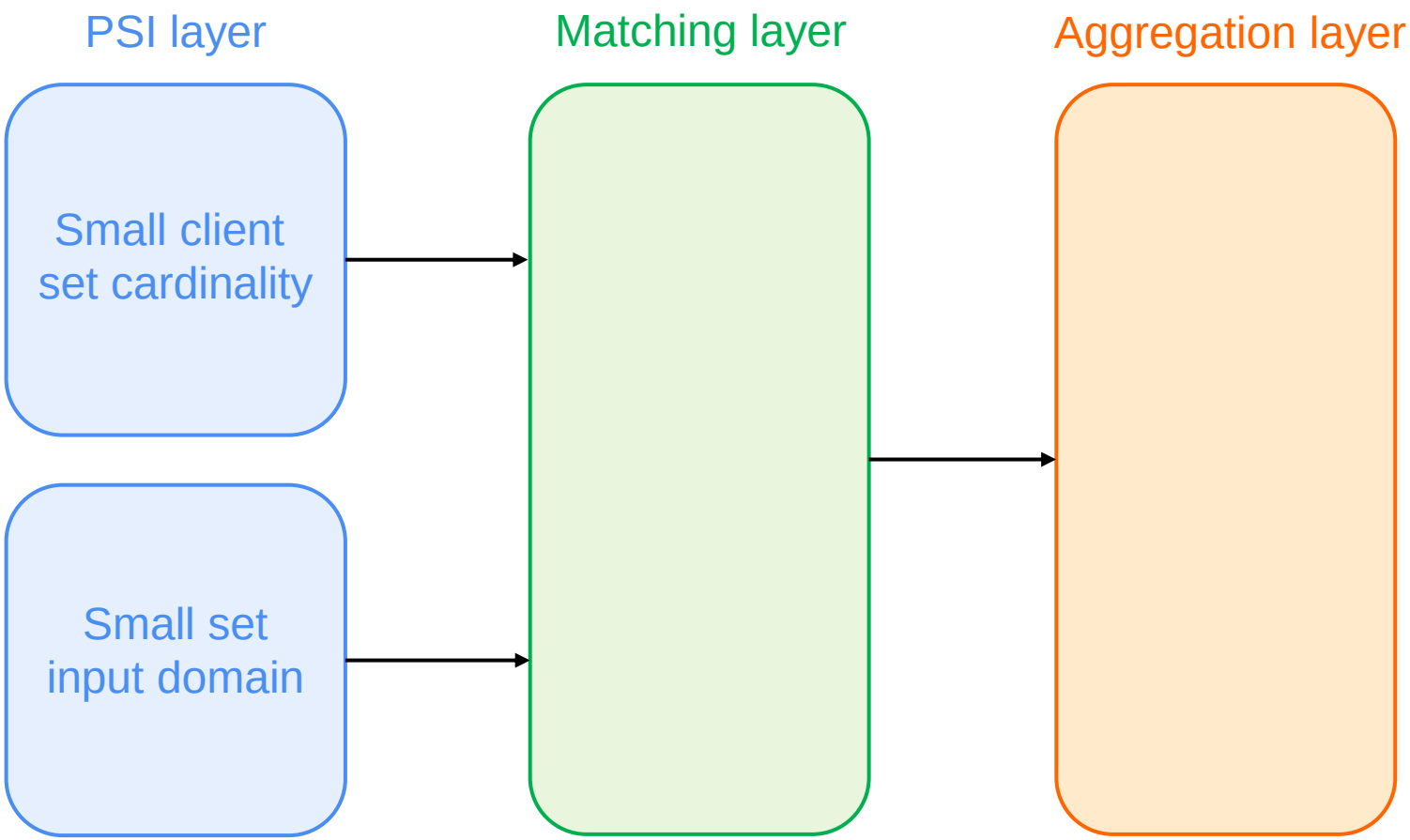
# Properties

- Correctness
  - The computed answer is correct with overwhelming probability
- Client privacy
  - The server learns no information about $X$ beyond its size
- Server privacy
  - The client learns no information about the server input beyond the size of $Y$ and the intended output of the protocol.

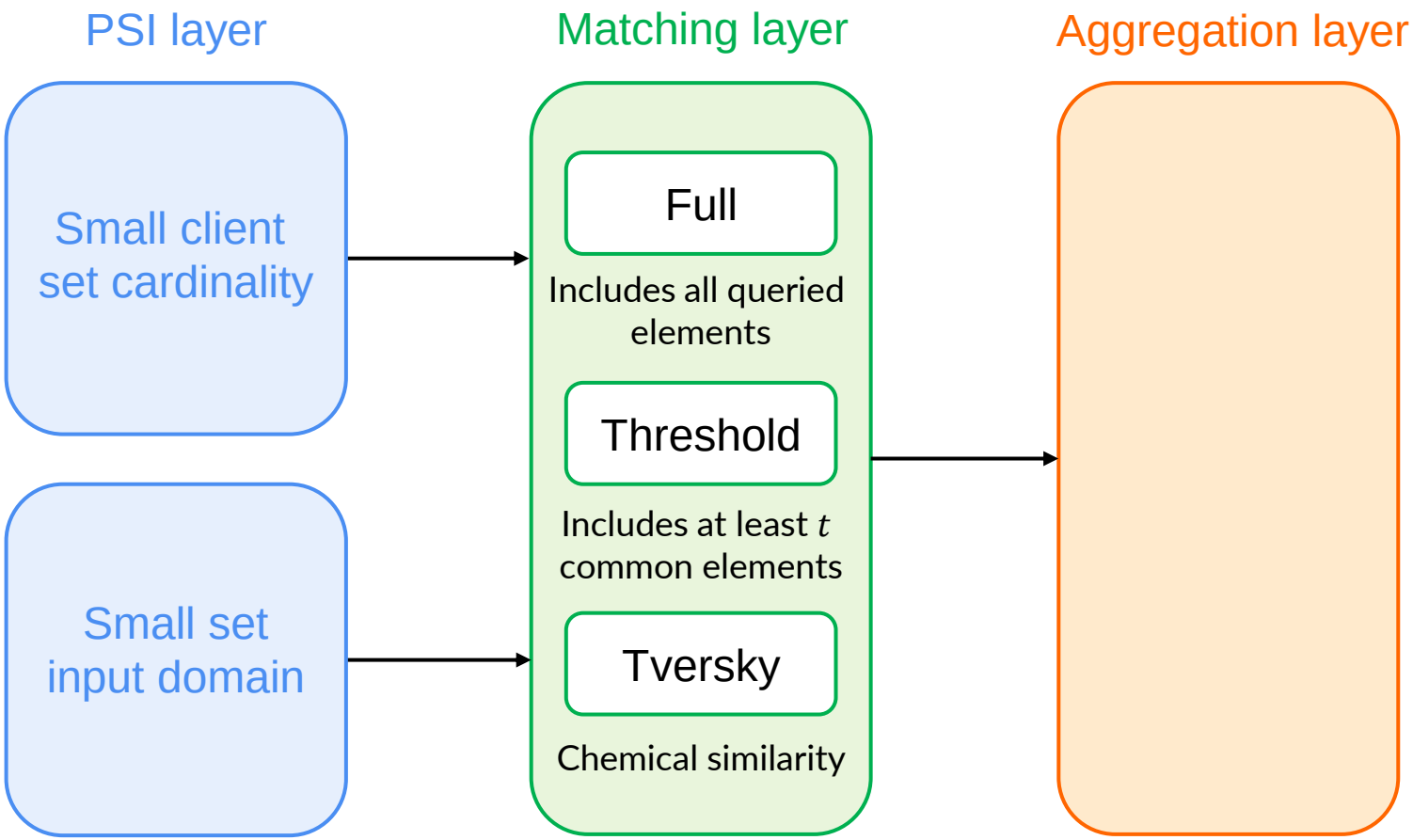
# A modular framework for PCM problems



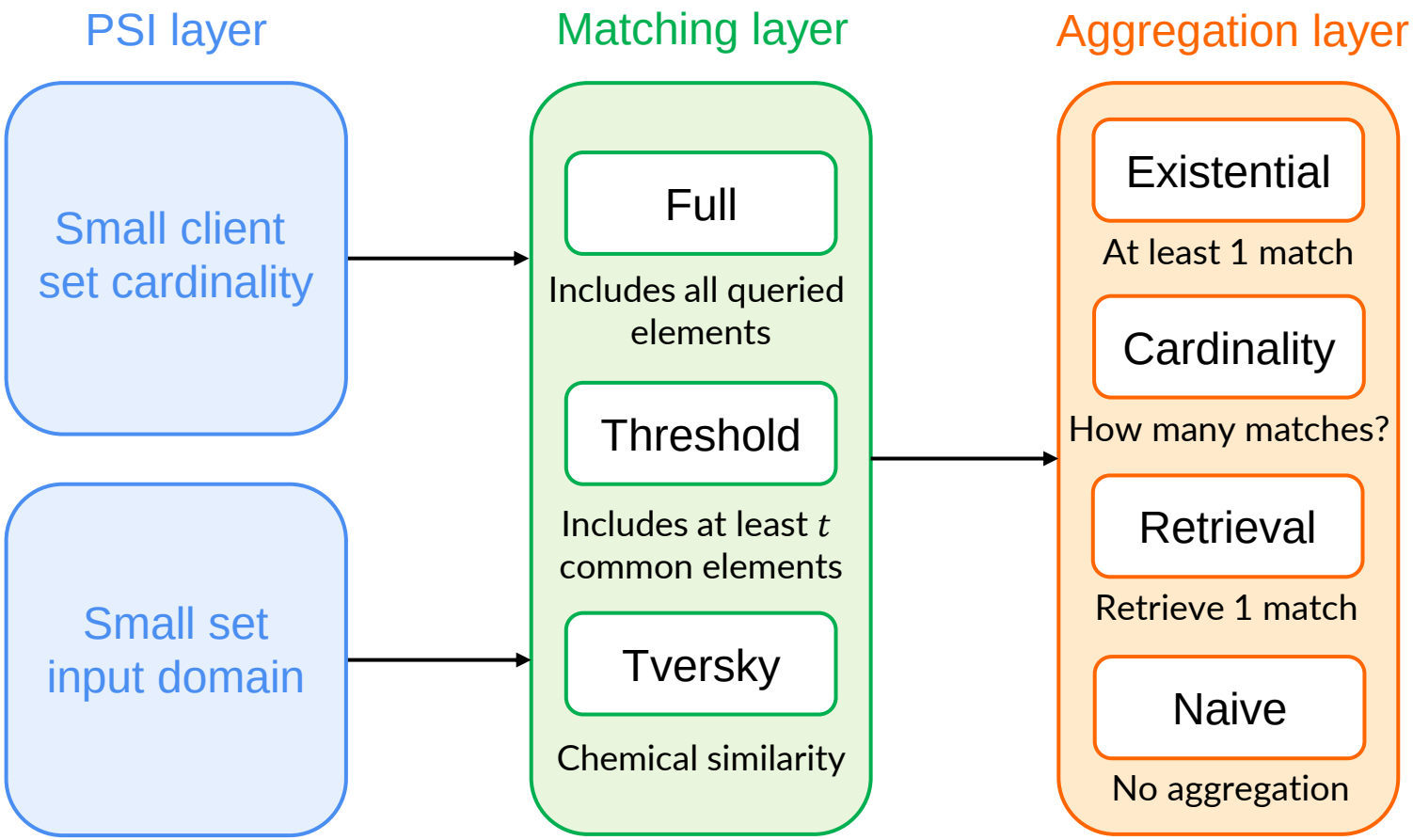
# PSI layer


PSI layer
Matching layer
Aggregation layer
Small client set cardinality
Small set input domain

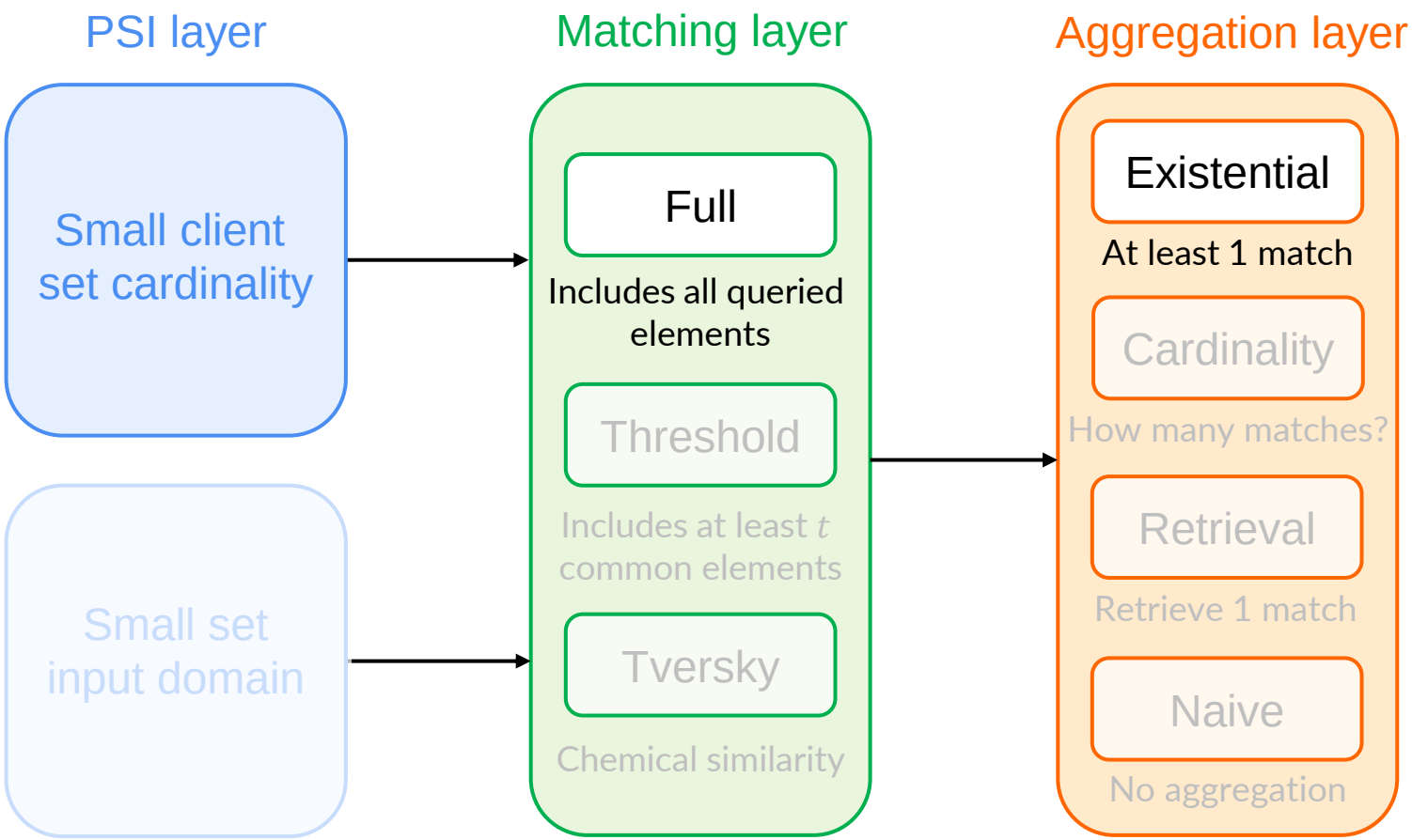
# Matching layer



# Aggregation layer



# Document search



# Somewhat homomorphic encryption

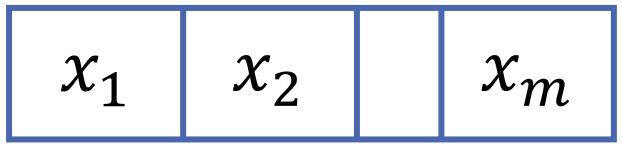
- KeyGen
- Encryption
- Decryption

$$pk, sk \leftarrow \mathsf{KeyGen}(\mathsf{param})$$
$$[\![x]\!] \leftarrow \mathsf{Enc}(pk, x)$$
$$x \leftarrow \mathsf{Dec}(sk, [\![x]\!])$$

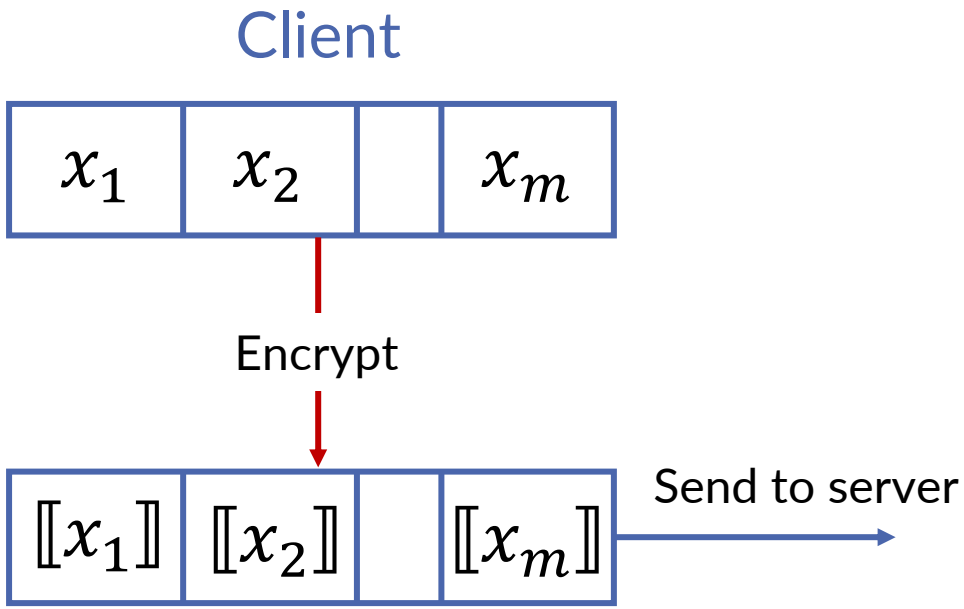
$$(ax + y) \mod q \leftarrow \mathsf{Dec}(sk, [\![a]\!] \cdot [\![x]\!] + [\![y]\!])$$

# PSI: small input

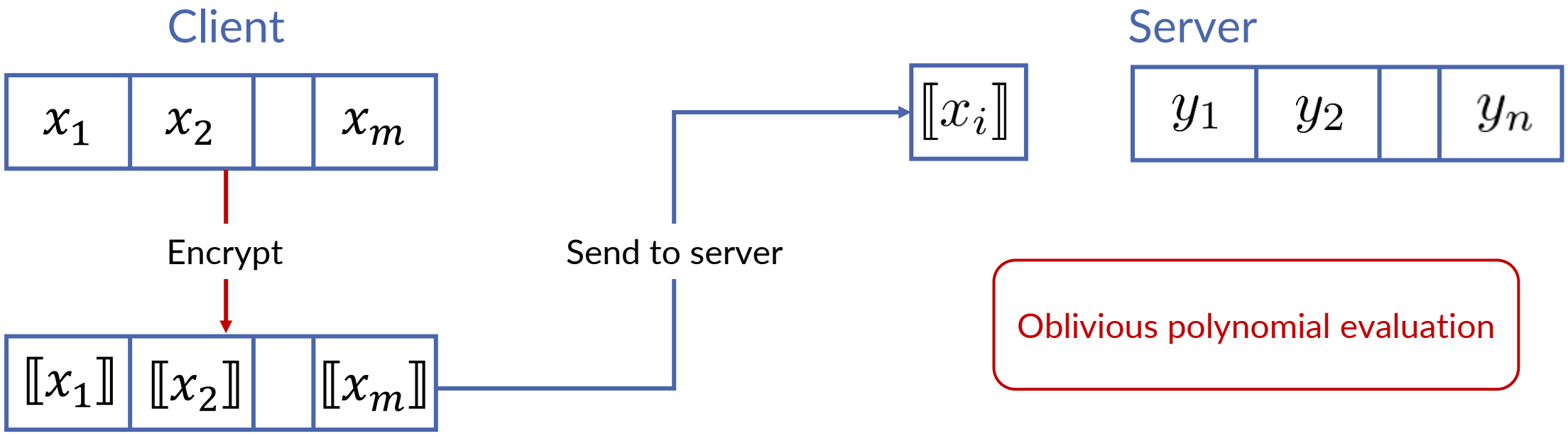
Client



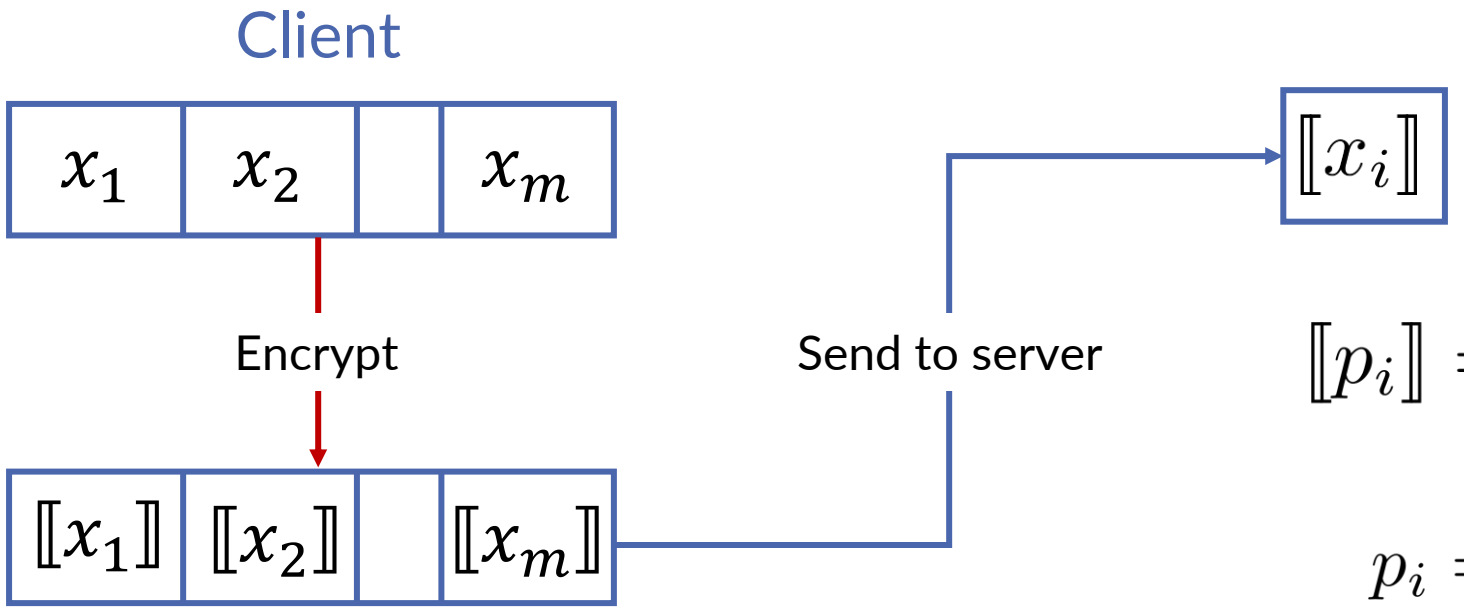
# PSI: small input

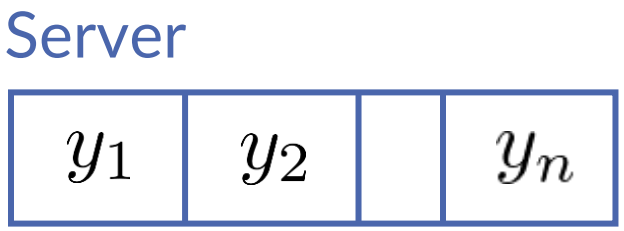


# PSI: small input



# PSI: small input



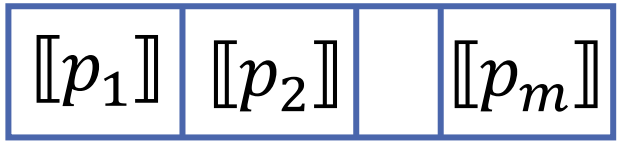




$$[\![p_i]\!] = r \cdot \prod_{y_j \in Y} ([\![x_i]\!] - y_j)$$

$$p_i = \begin{cases} 0 & : x_i \in Y \\ r \leftarrow^{\$} & : x_i \notin Y \end{cases}$$

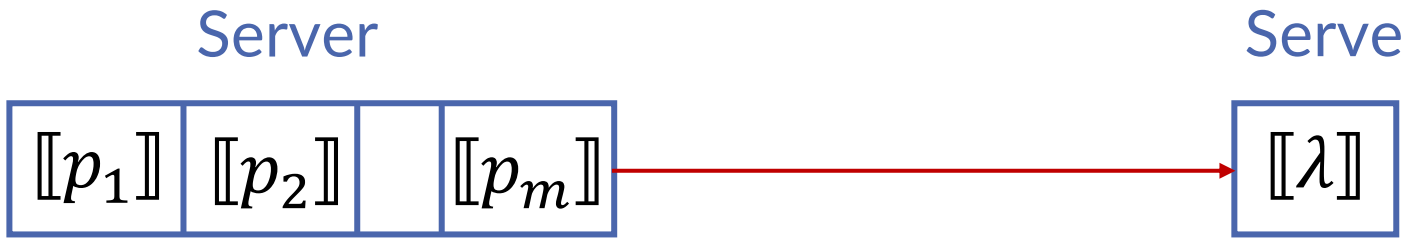
# Full matching

Server




$$p_i = \begin{cases} 0 & : x_i \in Y \\ r \leftarrow^{\$} & : x_i \notin Y \end{cases}$$

# Full matching

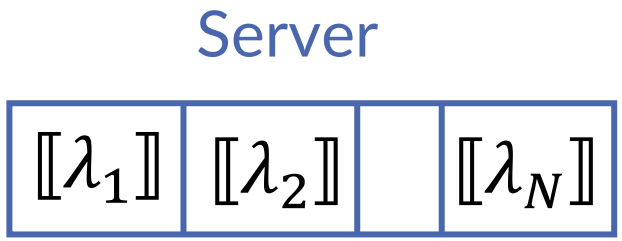



$$p_i = \begin{cases} 0 & : x_i \in Y \\ r \leftarrow^{\$} & : x_i \notin Y \end{cases}$$

$$[\![\lambda]\!] \leftarrow \sum_{i=1}^{m} [\![s_i]\!]$$

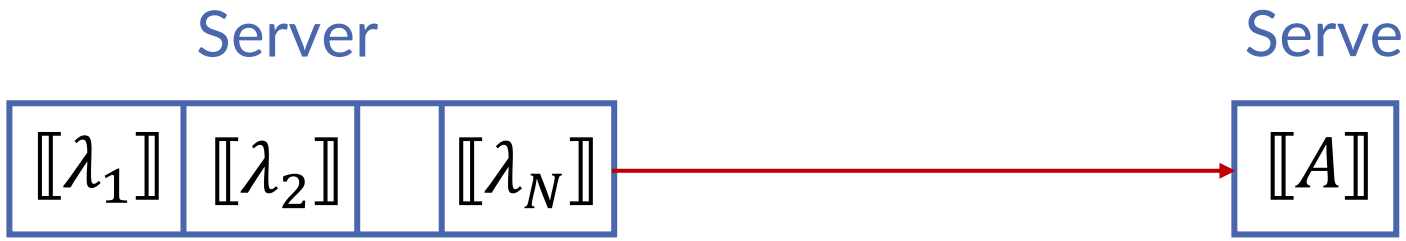
$$\lambda_i = \begin{cases} 0 & \text{if } Y_i \text{ is relevant} \\ r \leftarrow^{\$} & \text{otherwise} \end{cases}$$

# Existential aggregation




$$\lambda_i = \begin{cases} 0 & \text{if } Y_i \text{ is relevant} \\ r \leftarrow^{\$} & \text{otherwise} \end{cases}$$

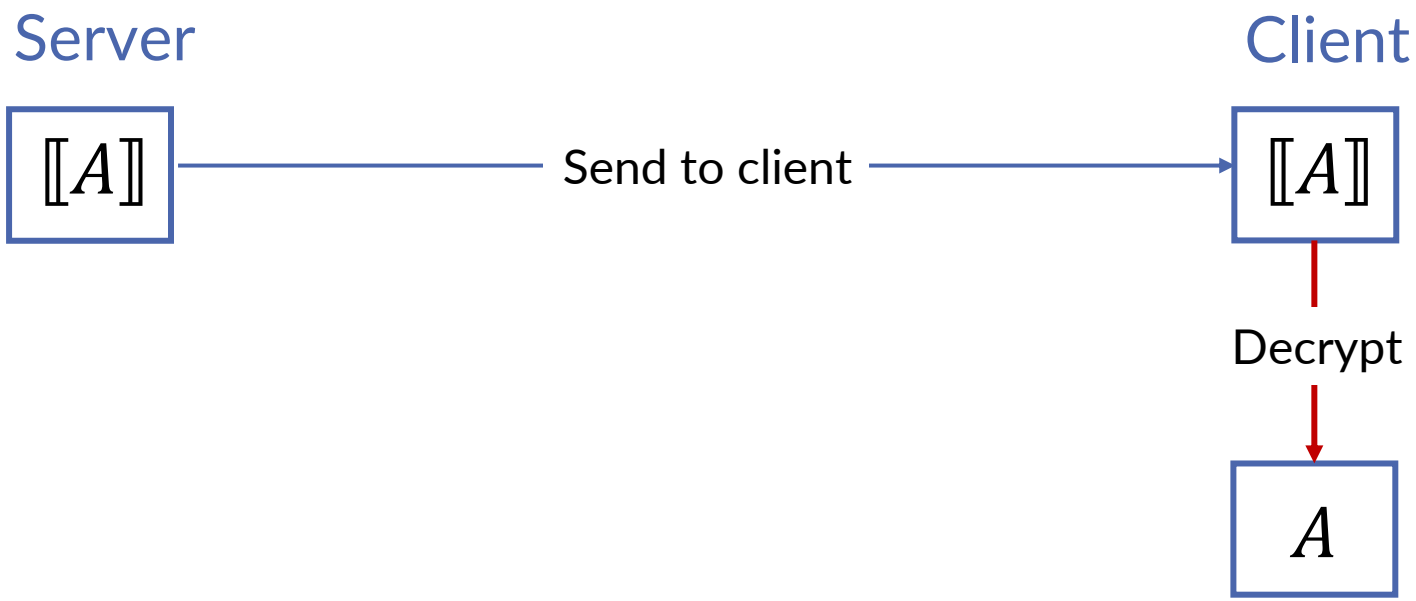
# Existential aggregation




$$\lambda_i = \begin{cases} 0 & \text{if } Y_i \text{ is relevant} \\ r \leftarrow^{\$} & \text{otherwise} \end{cases}$$

$$A \leftarrow \prod_{i=1}^{N} [\![\lambda_i]\!]$$

# Reveal



# Security and privacy

## Semi honest

- ✓Correctness
- ✓Client privacy
  - ✓Simulation proof
- ✓Server privacy
  - ✓Simulation proof

## Malicious

- ✓Client privacy
  - ✓ Reduction proof
- Server privacy
  - Limited protection

# Implementation

- We use BFV somewhat homomorphic encryption
- We use SIMD batching and replicate client query


- Repository: github.com/spring-epfl/private-collection-matching

# Disclaimer

- PCM works well in theory, but not all combinations are practical (reliance on an expensive zero detection). We optimize part of the framework to bypass zero detection.

# Document search

We implement two generic solutions with the same privacy

## Generic SMC

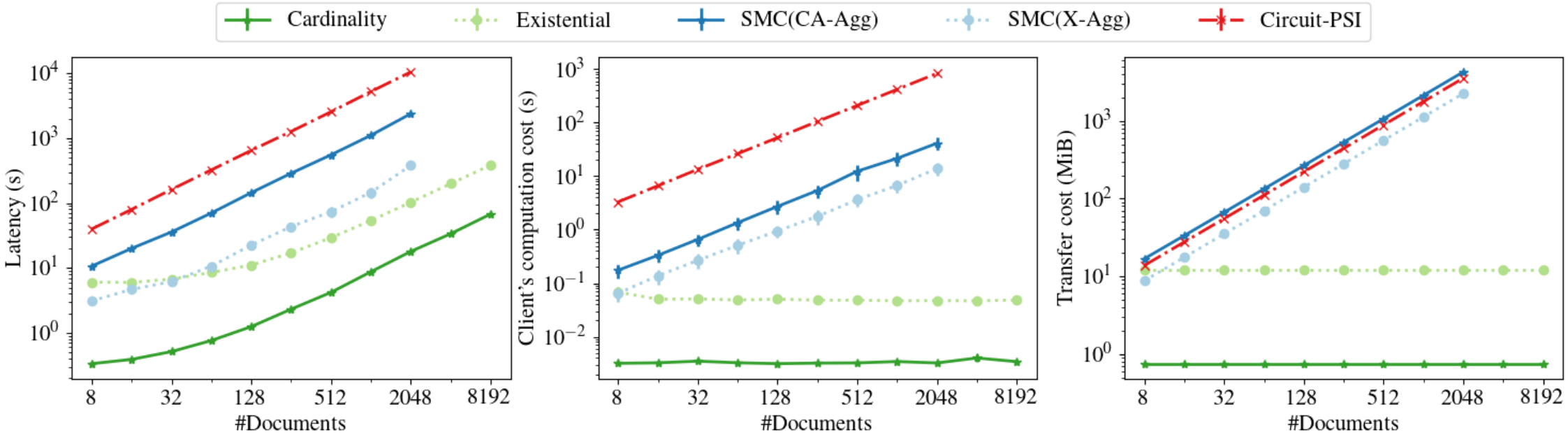
- We use  EMP-toolkit
- A semi-honest garbled circuit compiler

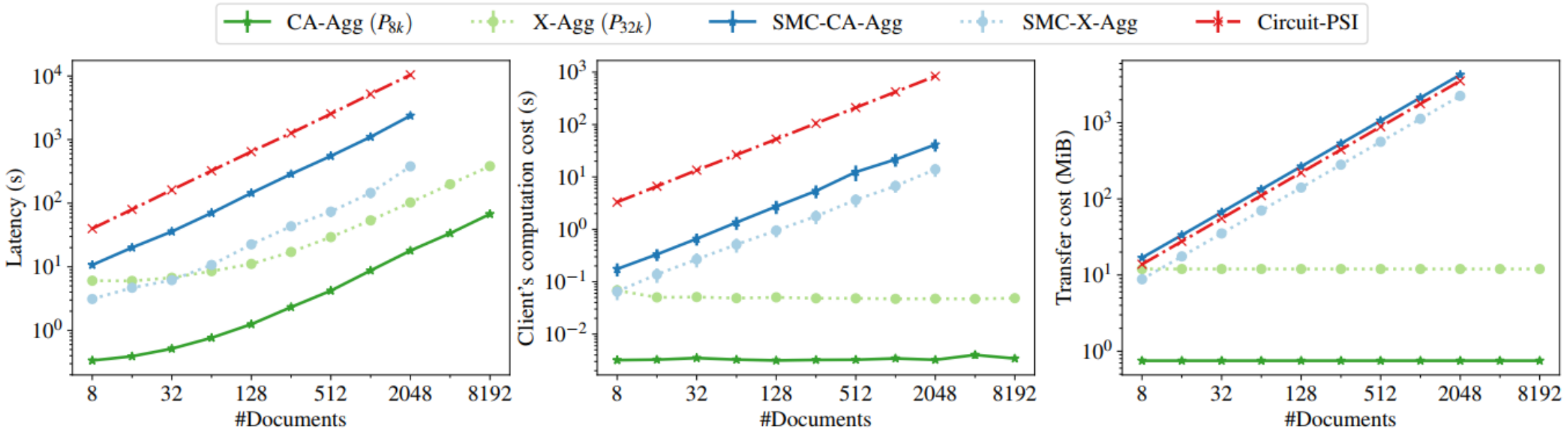
## Circuit-PSI

- Based on Chandran et al.
- Assumes equal client and server set sizes

Nishanth Chandran, Divya Gupta, and Akash Shah. Circuit-PSI With Linear Complexity via Relaxed Batch OPPRF. PoPETs 2022.

# Document search



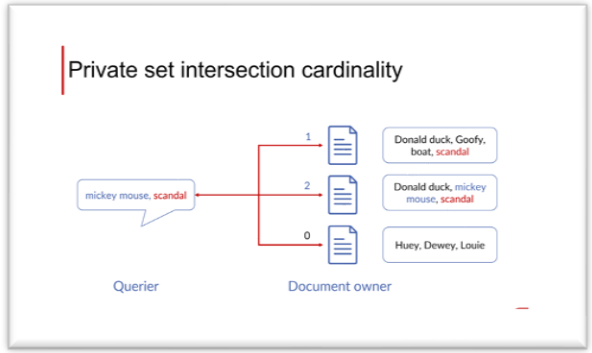


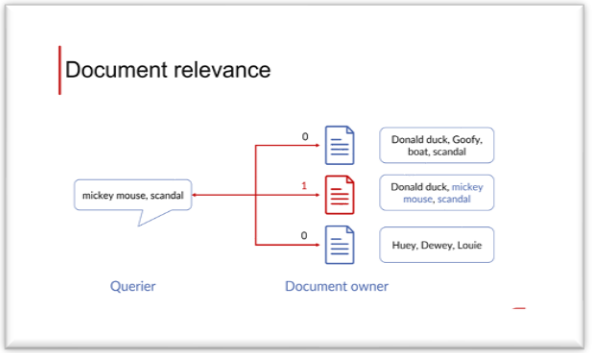

When searching 1000 document we improve client computation by up to 70,000x, latency by up to 96x, and transfer cost by up to 2,800x.

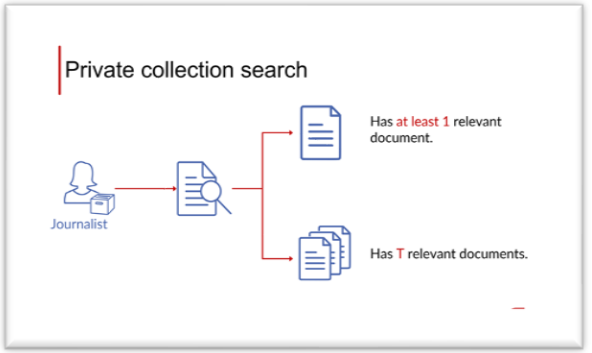
# Security proof

- Simulations do not prove that a system is secure
- Simulations show a system is as secure as the ideal-world

- Ideal-worlds may not be as secure as they sound.



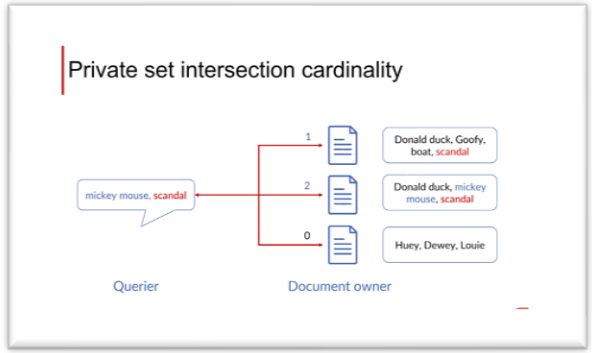



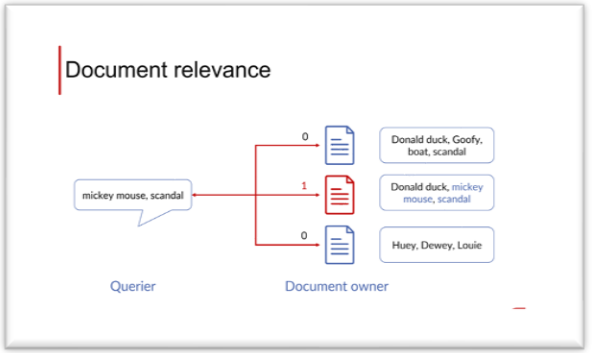



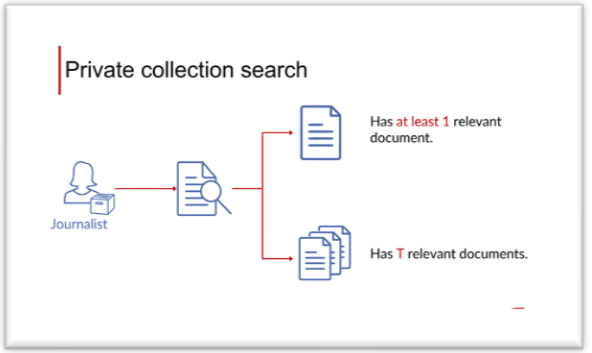
# Take away

- Simulations do not prove that a system is secure
- Simulations show a system is as secure as the ideal-world

- Ideal-worlds may not be as secure as they sound.



# References


- Kasra Edalatnejad, Mathilde Raynal, Wouter Lueks, Carmela Troncoso: Private Collection Matching Protocols. PoPETS 2023.
- Kasra Edalatnejad, Wouter Lueks, Julien Pierre Martin, Soline Ledésert, Anne L'Hôte, Bruno Thomas, Laurent Girod, Carmela Troncoso: DatashareNetwork: A Decentralized Privacy-Preserving Search Engine for Investigative Journalists. USENIX Security Symposium 2020.
- Nishanth Chandran, Divya Gupta, and Akash Shah. Circuit-PSI With Linear Complexity via Relaxed Batch OPPRF. PoPETs 2022.
- Icons made from https://www.onlinewebfonts.com/icon are licensed by CC BY 4.0.