CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

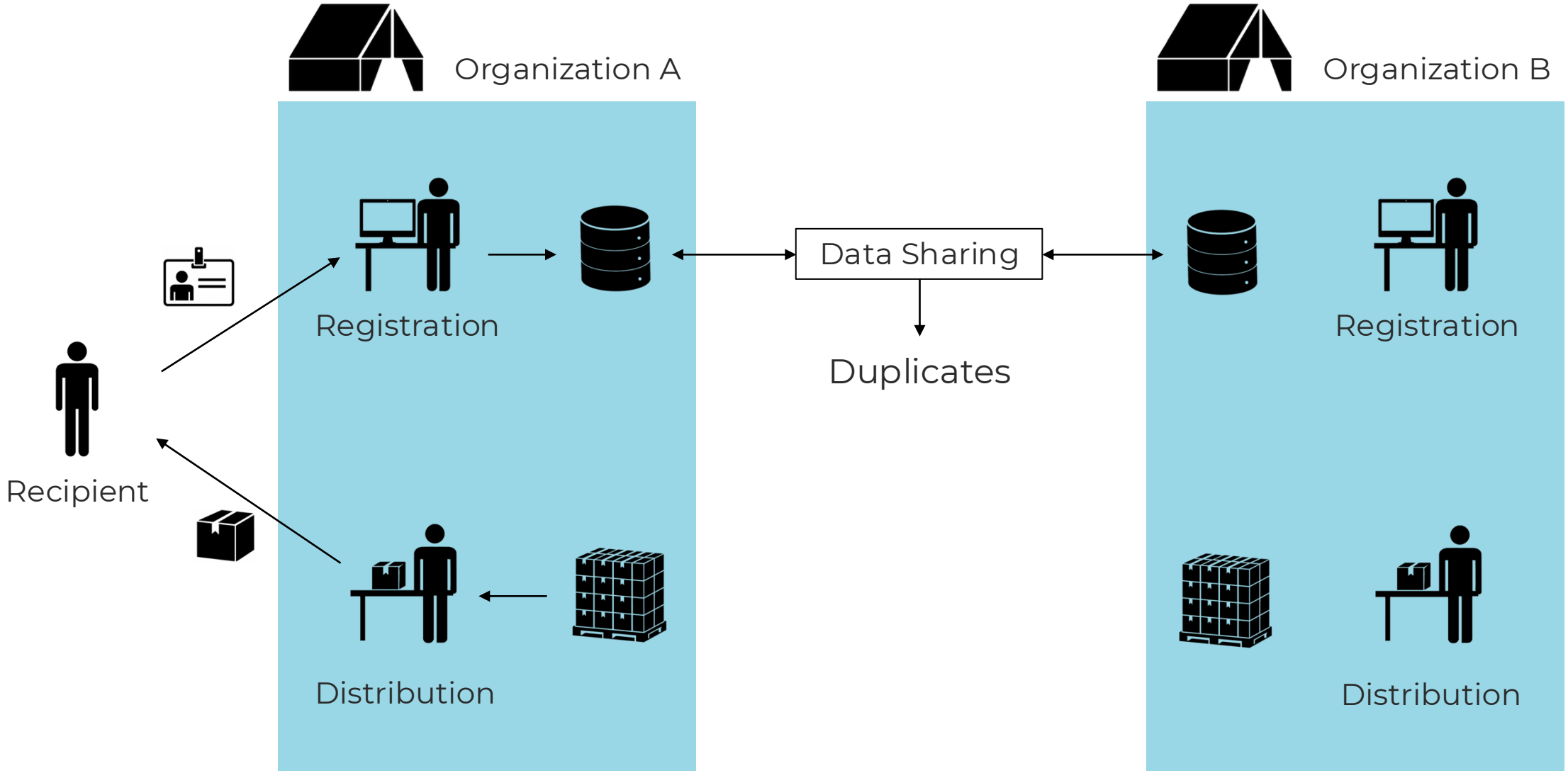
Fuzzy PSI meets Real-World Problem: Deduplication of Humanitarian Aid Recipients

Tim Rausch, *Sylvain Chatel*, *Wouter Lueks*

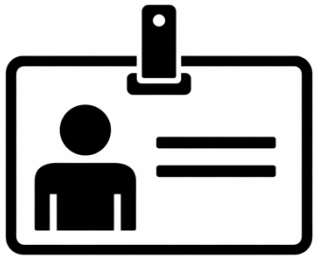
CISPA Helmholtz Center for Information Security

Cryptography Applications Workshop | May 10, 2026

Humanitarian Aid Distribution



Deduplication Options



Unique Identifiers

- Government-issued tax IDs
- May be unavailable

Biometrics

- Fingerprint or iris scan
- Highly privacy-invasive

Biographical Data

- Name, date of birth
- Needs to be verified

Biographical Deduplication

Organization A

Organization B

Tim	Rausch	Saarbrücken	=	Maria	Rossi	Rome
Jane	Doe	Washington		Tim	Rausch	Saarbrücken
Wouter	Lueks	Saarbrücken	≈	Max	Mustermann	Bonn
Erika	Musterfrau	Berlin		Walter	Lücks	Saarbrücken

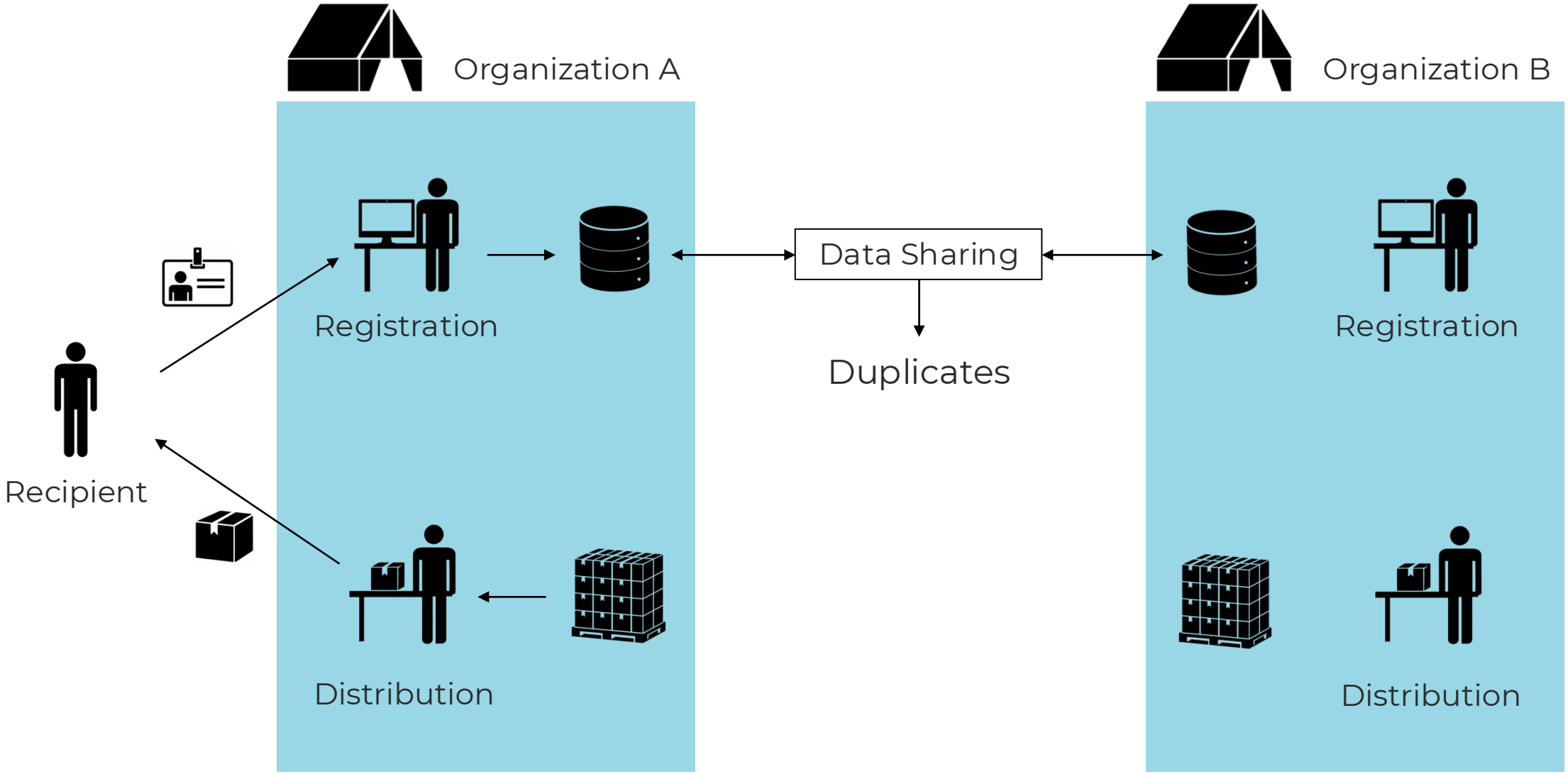
Functionality

Find duplicates by **fuzzily** comparing records

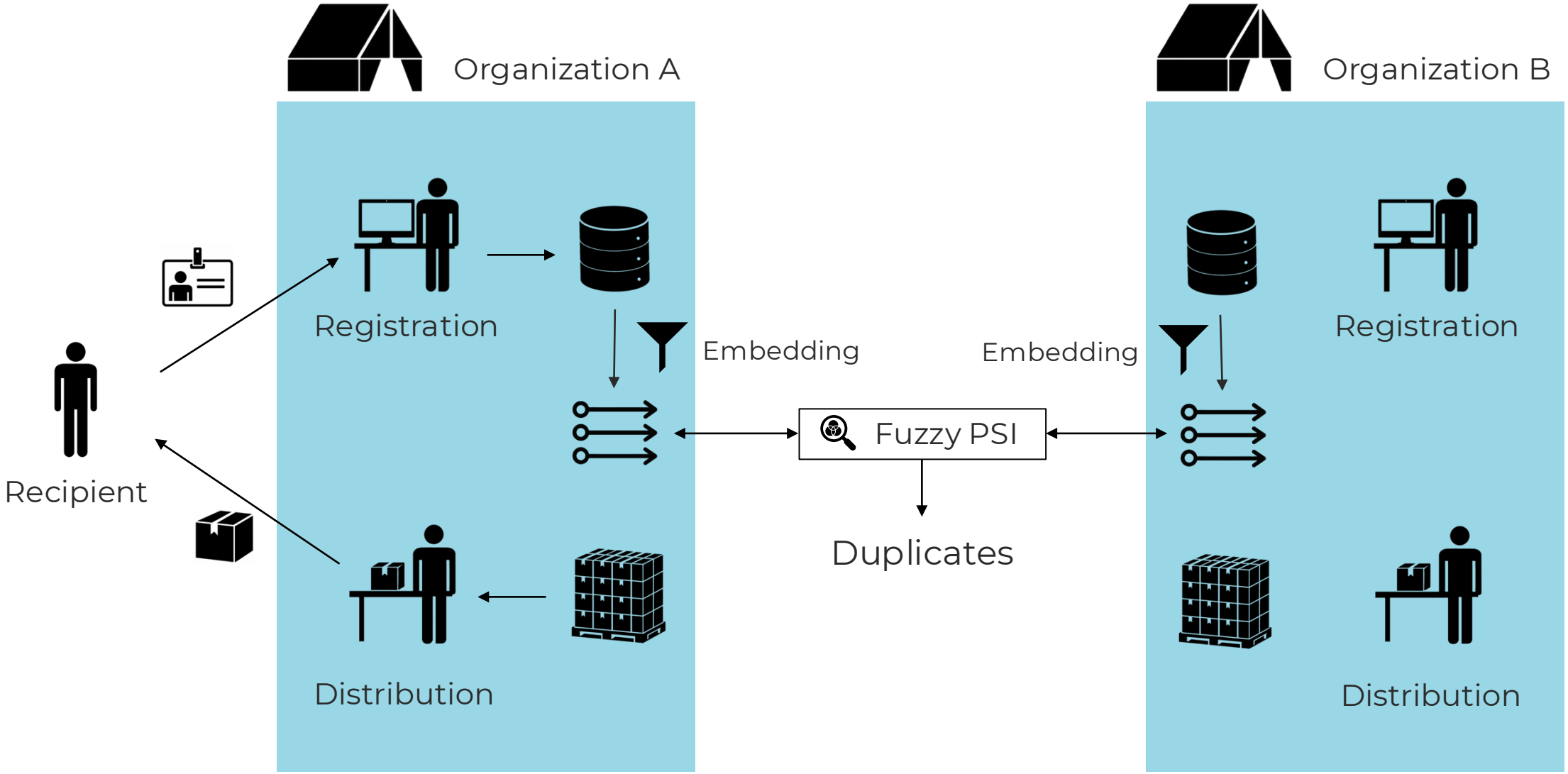
Privacy

No leakage to other organizations

System Design

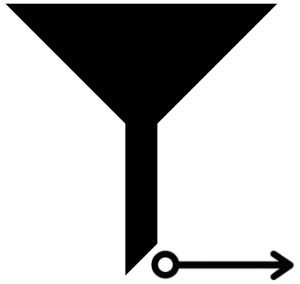


System Design



Deduplication from Fuzzy PSI

Jane Doe, Washington



Embedding

Transform registration records into points in a **metric space**

- Preserve similarity
- Context-**dependent**

Fuzzy Private Set Intersection (FPSI)

Privately identifies **close points** in metric space

- Close: $d(x, y) \leq \tau$
- Context-**independent**

Validating our Design

Evaluation Dataset

Generate **synthetic** evaluation dataset

- Based on voter registration DB
- Synthetically introduce duplicates
- Models common errors:
  - E.g., typos and missing fields

Representative of a deduplication task

Embedding

Embed records into **Hamming space**

- Using standard record linkage techniques
- Parameterization:
  - Embedding dimension $l = 512$
  - **Threshold $\tau \approx 128 = l/4$**
  - FPR of 0.1% (deduplicating records against a DB of 131k existing registrations)
  - Recall of 99.4%

Missing: FPSI Protocol for **Hamming Space**

Why Existing FPSI does not Work

FLPSI
[Uzun et al., USENIX '21]

DA-PSI
[Chakraborti et al., USENIX '23]

Approx-PSI
[Chongchitmate et al., 2024]

Fmap-FPSI
[Gao et al., Asiacrypt '24]

PE-FPSI
[Blass and Noubir, USENIX '26]

Why Existing FPSI does not Work

	Threshold Dependent	Insufficient Precision
FLPSI [Uzun et al., USENIX '21]	τ	$\approx$
DA-PSI [Chakraborti et al., USENIX '23]		
Approx-PSI [Chongchitmate et al., 2024]		
Fmap-FPSI [Gao et al., Asiacrypt '24]		
PE-FPSI [Blass and Noubir, USENIX '26]		

FLPSI

Approximates Hamming Distance by **Subsampling**

- Run time depends on subsampling parameters τ
- **Most** registrations are false duplicates $\approx$

Why Existing FPSI does not Work

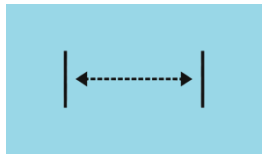
	Threshold Dependent	Insufficient Precision	Prohibitively Expensive
FLPSI [Uzun et al., USENIX '21]	τ	≈	
DA-PSI [Chakraborti et al., USENIX '23]	τ	≈	🐌
Approx-PSI [Chongchitmate et al., 2024]			
Fmap-FPSI [Gao et al., Asiacrypt '24]			
PE-FPSI [Blass and Noubir, USENIX '26]			

DA-PSI

Approximates the Hamming Distance

- **Quadratic** in τ τ
- Slower than Garbled Circuits 🐌
- **Every** record is false duplicate ≈

Why Existing FPSI does not Work

	Threshold Dependent	Insufficient Precision	Prohibitively Expensive	Unfulfillable Assumption
FLPSI [Uzun et al., USENIX '21]	τ	$\approx$		
DA-PSI [Chakraborti et al., USENIX '23]	τ	$\approx$	🐌	
Approx-PSI [Chongchitmate et al., 2024]				
Fmap-FPSI [Gao et al., Asiacrypt '24]				
PE-FPSI [Blass and Noubir, USENIX '26]				

Approx-PSI

Subquadratic through input assumption:

Points are

- Close: $d(x, y) \leq \tau$, or
- Far apart: $d(x, y) \geq 3\tau$

With $\boldsymbol{\tau \approx l/4}$, points must be $d(x, y) \geq 3/4 \cdot l$ apart

Why Existing FPSI does not Work

	Threshold Dependent	Insufficient Precision	Prohibitively Expensive	Unfulfillable Assumption
FLPSI [Uzun et al., USENIX '21]	τ	≈		
DA-PSI [Chakraborti et al., USENIX '23]	τ	≈	🐌	
Approx-PSI [Chongchitmate et al., 2024]				⟷
Fmap-FPSI [Gao et al., Asiacrypt '24]	τ			⟷
PE-FPSI [Blass and Noubir, USENIX '26]				

Fmap-FPSI

Linear through input assumption: Each vector has $\tau + 1$ **unique compontents**

- Bit vectors have one of **two** values in each compontent ⟷
- **Quadratic** in τ τ

Why Existing FPSI does not Work

	Threshold Dependent	Insufficient Precision	Prohibitively Expensive	Unfulfillable Assumption
FLPSI [Uzun et al., USENIX '21]	τ	$\approx$		
DA-PSI [Chakraborti et al., USENIX '23]	τ	$\approx$	🐌	
Approx-PSI [Chongchitmate et al., 2024]				⟷
Fmap-FPSI [Gao et al., Asiacrypt '24]	τ			⟷
PE-FPSI [Blass and Noubir, USENIX '26]	τ		🐌	

PE-FPSI

Linear communication through predicate encryption.

- **Linear** in τ τ
- **Inefficient:** Processing 2k new registrations takes months 🐌

Why Existing FPSI does not Work

Insufficient Precision

FLSPI, DA-PSI

Unfulfillable Assumption

Approx-PSI, Fmap-FPSI

Sub-quadratic complexities come at a **cost**

Threshold-Dependent Complexity

FLSPI, DA-PSI,
Fmap-FPSI, PE-FPSI

Prohibitively Expensive

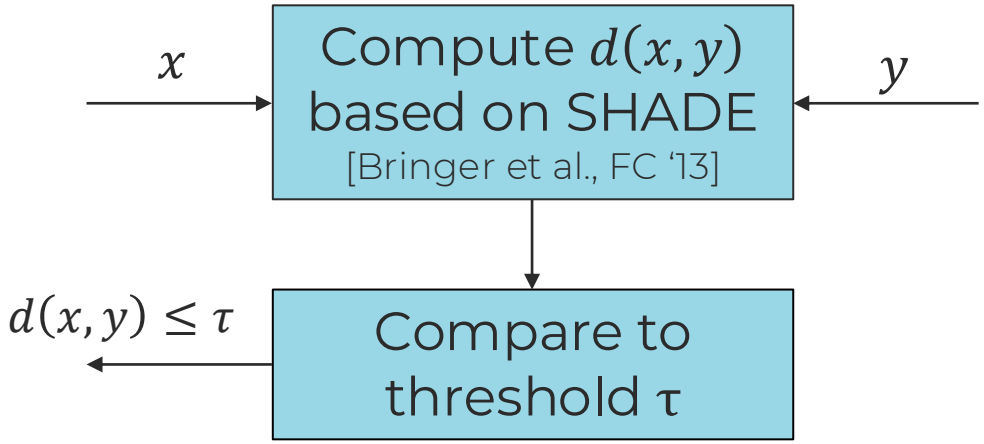
DA-PSI, PE-FPSI,
arguably others

Asymptotics do not protect against **concrete** inefficiecy

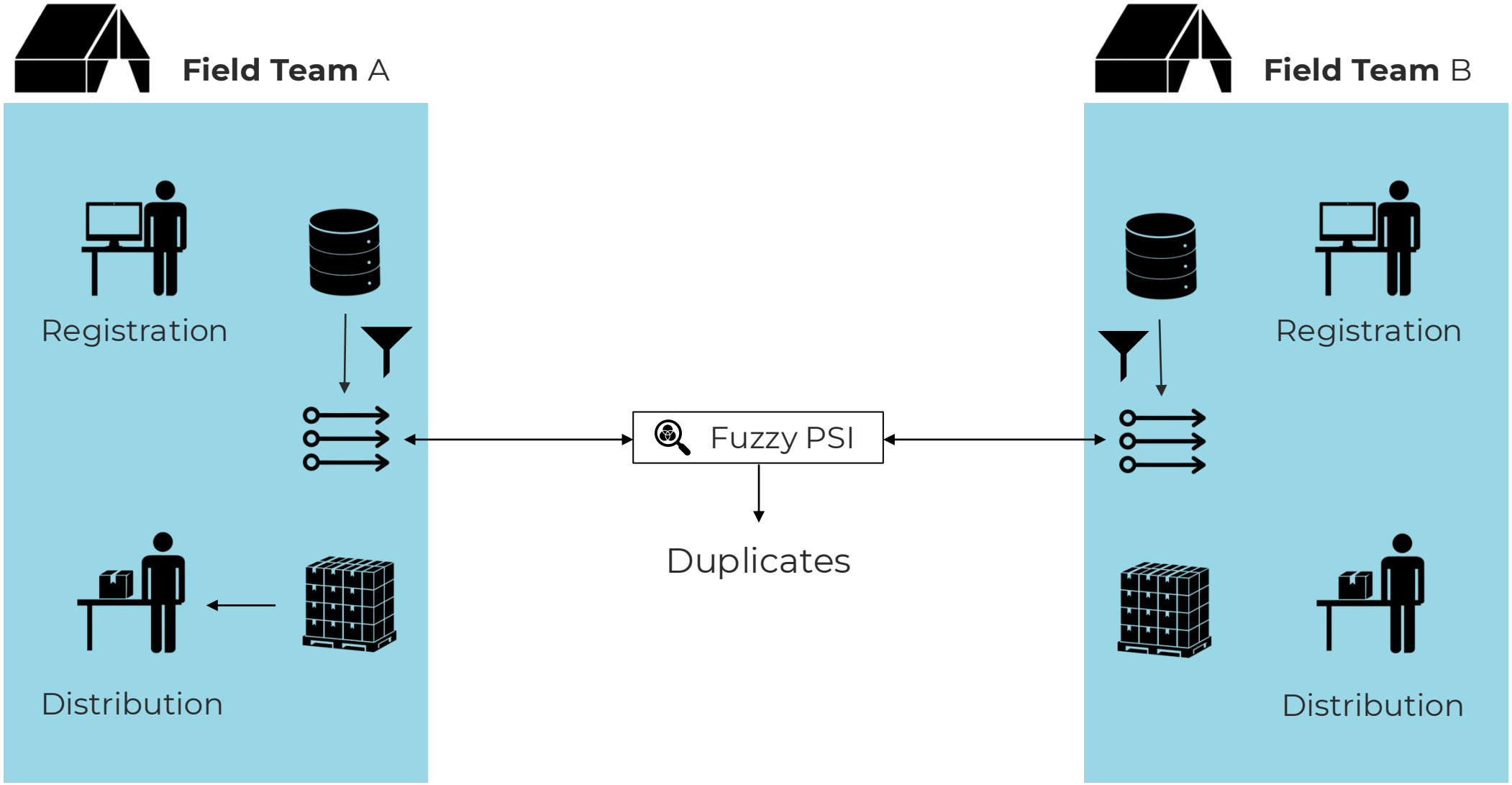
Our otFPSI

- Simple protocol, based only on **Oblivious Transfer** (OT)
- Accept complexity **quadratic** in set size – ensure **concrete** efficiency
  - No assumptions on inputs necessary

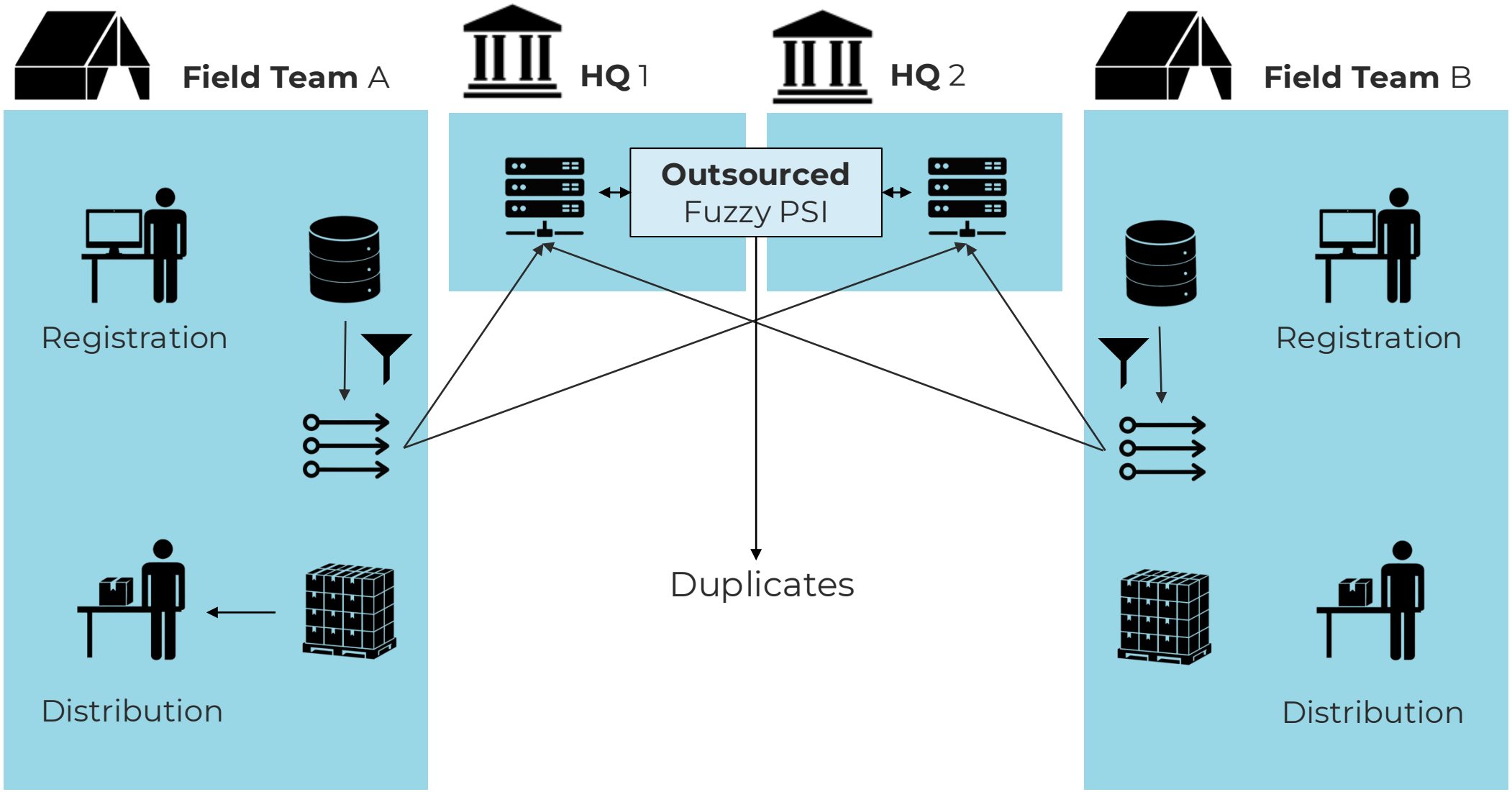
How it Works: Comparison protocol for two bit strings



System Design



System Design



otFPSI Evaluation

Overcomes limitations of previous protocols:

- **Exact Results**
- **Arbitrary Inputs**
- **Threshold Independent**
- **Outsourceable**
- **Concretely Efficient**

otFPSI Evaluation

otFPSI outperforms all existing Hamming FPSI protocols

FLPSI [Uzun et al., USENIX '21]
5x Faster than online computation
10x Faster than total computation

Fmap-FPSI [Gao et al., Asiacrypt '24]
≈ Similar online time
≥2 OoM Faster than total

DA-PSI [Chakraborti et al., USENIX '23]
3 OoM Faster for all set sizes

PE-FPSI [Blass and Noubir, USENIX '26]
3 OoM Faster for all set sizes

Approx-PSI [Chongchitmate et al., 2024]
8x Faster for large sets
3 OoM Faster for small sets

otFPSI Advantages

- **Exact Results**
- **Outsource-able**
- **Arbitrary Inputs**
- **Threshold Independent**

OoM: order of magnitude

Conclusion

- Deduplication seems like an ideal FPSI application – existing FPSI protocols fail
  - **Gap** between protocols and applications
- otFPSI overcomes limitations and outperforms

Exact Results	Outsource-able	Concretely Efficient
Arbitrary Inputs	Threshold Independent	

Read our paper!

https://arxiv.org/abs/2604.08019

To appear at IEEE S&P 2026

Contact

Tim Rausch
CISPA Helmholtz Center for Information Security
tim.rausch@cispa.de