CS 528 (Fall 2021)
Data Privacy & Security

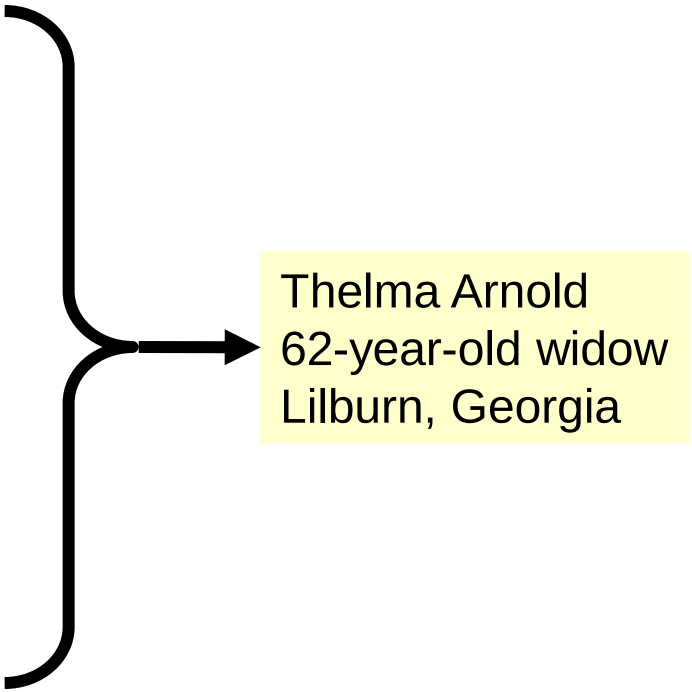
Yuan Hong
Department of Computer Science
Illinois Institute of Technology

Chapter 9-B
Private Information Retrieval

AOL SEARCH DATA SCANDAL (2006)

#4417749:

```
clothes for age 60
60 single men
best retirement city
jarrett arnold
jack t. arnold
jaylene and jarrett arnold
gwinnett county yellow pages
rescue of older dogs
movies for dogs
sinus infection
```



Thelma Arnold
62-year-old widow
Lilburn, Georgia

OBSERVATION

The owners of the database know a lot about the users!

This poses a risk to users’ privacy.

E.g., consider database with stock prices…

Really?

Can we do something about it?

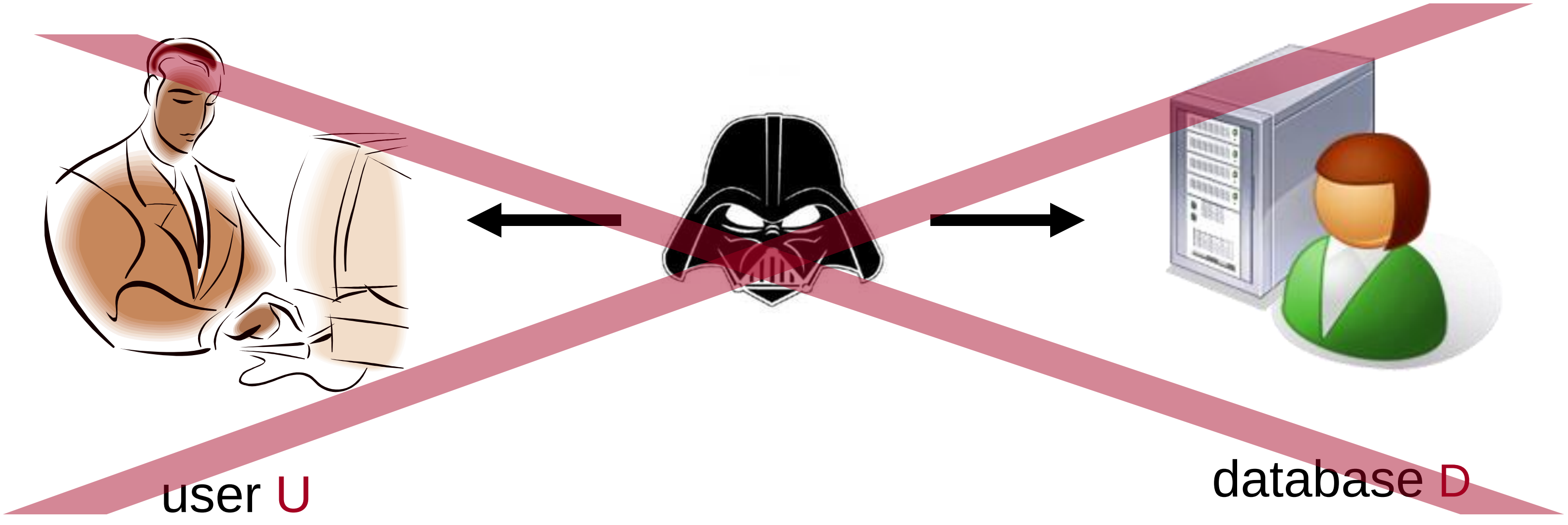
Yes, we can:

trust them that they will protect our secrecy,

or

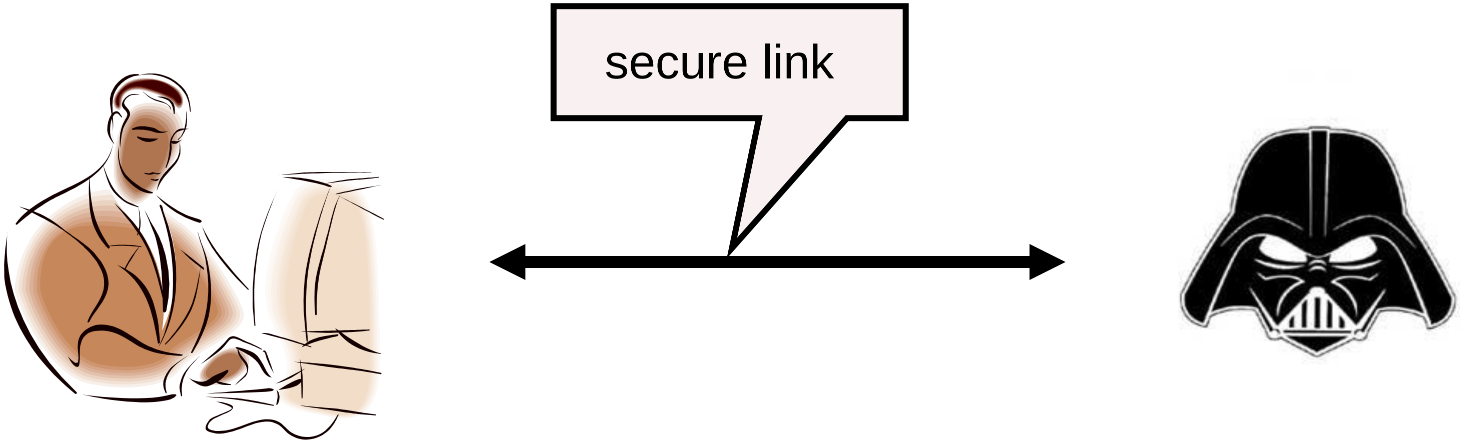
use cryptography!

HOW CAN CRYPTO HELP?



Note: this problem has nothing to do with side-channels, website fingerprinting, etc.

THREAT MODEL



secure link

user U

database D

A new primitive:

Private Information Retrieval (PIR)

PRIVATE INFORMATION RETRIEVAL (PIR) [CGKS95]

Goal: allow user to query database while hiding the identity of the data-items she is after.

Note: hides identity of data-items; not existence of interaction with the user.

Motivation: patient databases; stock quotes; web access; many more....

Paradox(?): imagine buying in a store without the seller knowing what you buy.

(Encrypting requests is useful against third parties; not against owner of data.)

MODEL

Server: holds *n*-bit string *x*

** *n* should be thought of as very large**

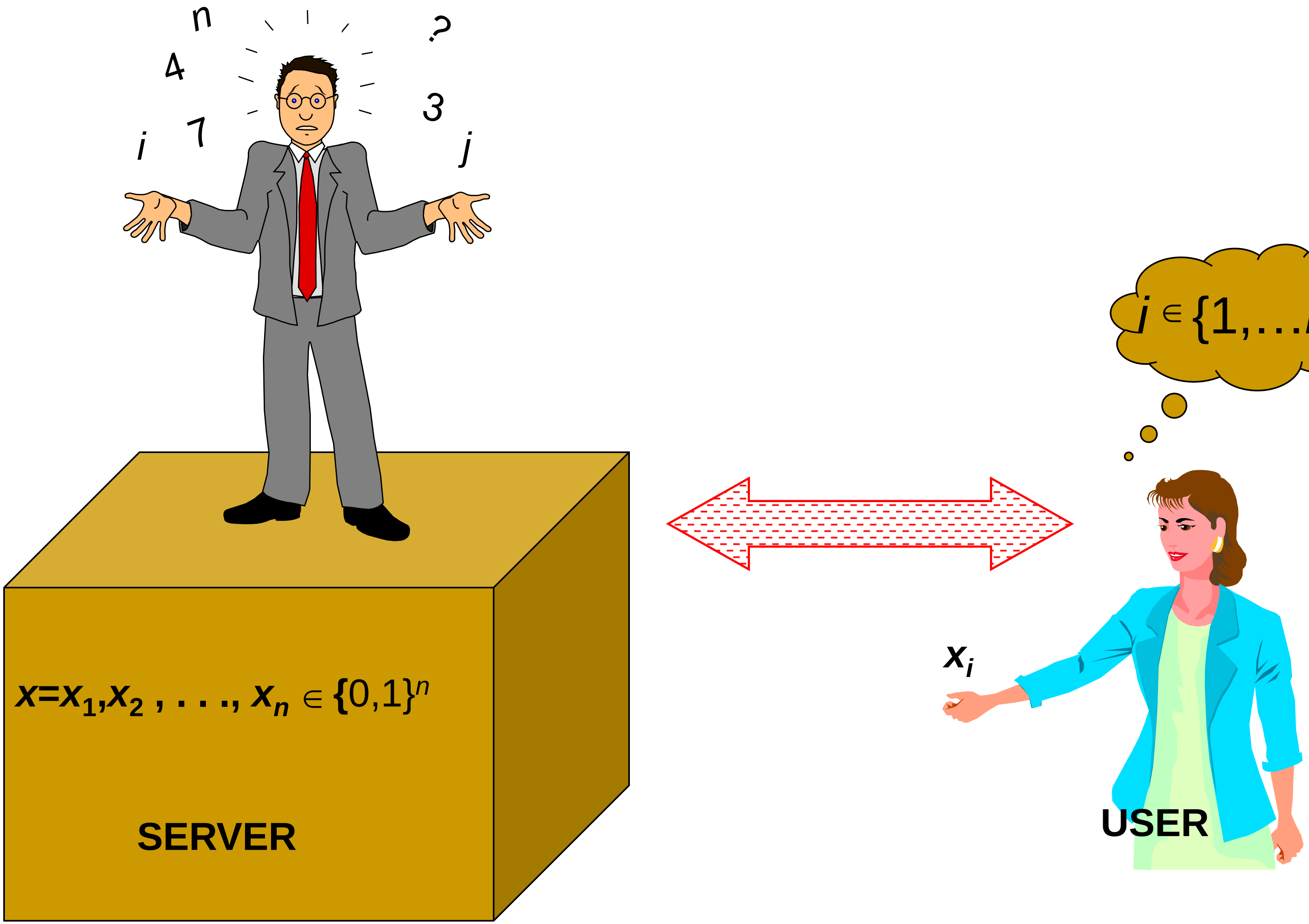
User: wishes

- to retrieve x_i

and

- to keep i private

PRIVATE INFORMATION RETRIEVAL (PIR)



$i \in \{1,\ldots n\}$

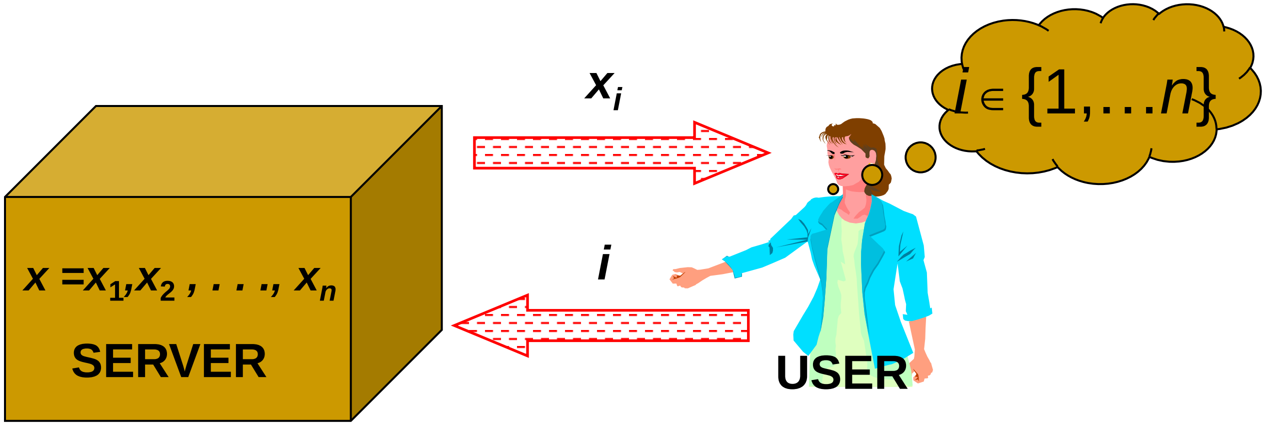
x_i

$x = x_1, x_2, \ldots, x_n \in \{0,1\}^n$

SERVER

USER

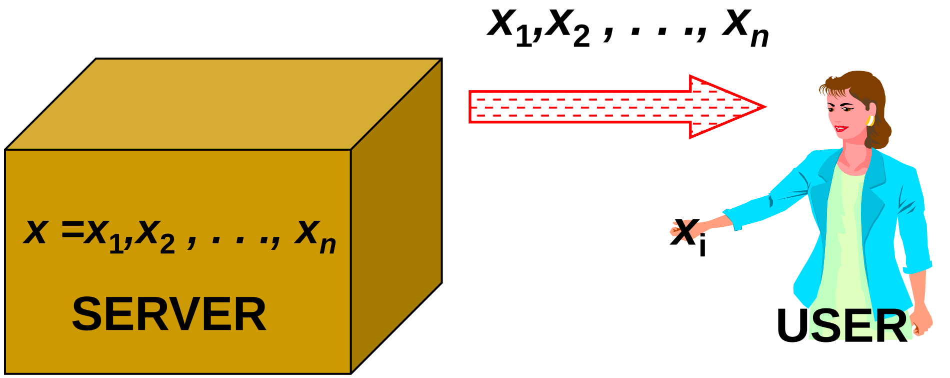
NON-PRIVATE PROTOCOL



NO privacy!!!

Communication: 1

TRIVIAL PRIVATE PROTOCOL



Server sends entire database *x* to User.

Information theoretic privacy.

Communication: *n*

Not optimal !

OTHER SOLUTIONS

User asks for additional random indices.

Drawback: leaks information, reduces communication efficiency

Employ general crypto protocols to compute x_i privately.

Drawback: highly inefficient (polynomial in n).

Anonymity (e.g., via Anonymizers).

Note: different concern: hides identity of user; not the fact that x_i is retrieved.

TWO APPROACHES FOR PIR

Information-Theoretic PIR [CGKS95,Amb97,...]

Replicate database among *k* servers.

User queries all the servers

Computational PIR [CG97,KO97,CMS99,...]

Computational privacy, based on cryptographic assumptions.

KNOWN COMM. UPPER BOUNDS

Multiple servers, information-theoretic PIR:

2 servers, comm. $n^{1/3}$ [CGKS95]

k servers, comm. $n^{1/\Omega(k)}$ [CGKS95, Amb96,…,BIKR02]

log n servers, comm. Poly(log(n)) [BF90, CGKS95]

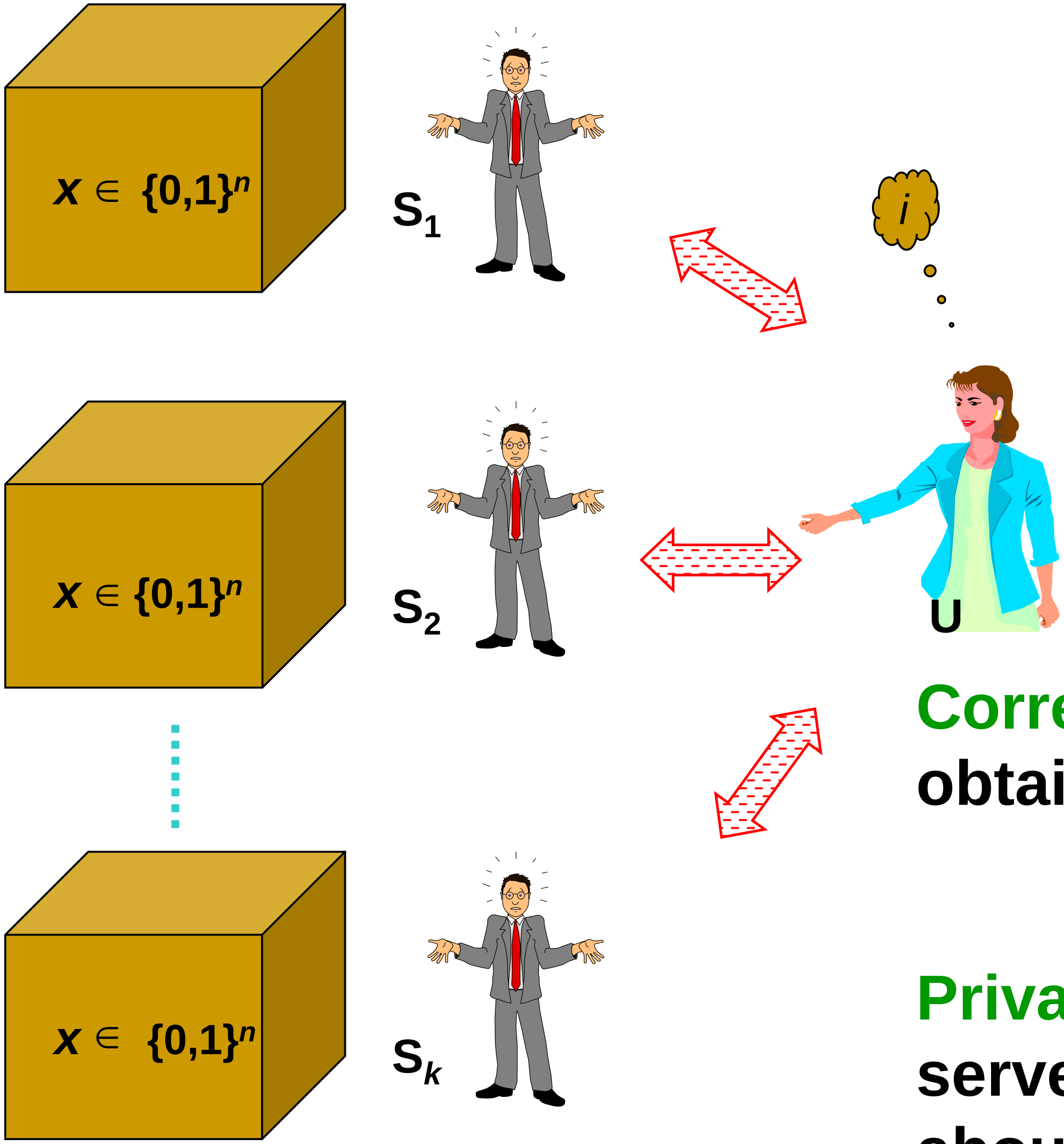
Single server, computational PIR:

Comm. Poly(log(n))

Under appropriate computational assumptions [KO97,CMS99]

Sub-linear with n

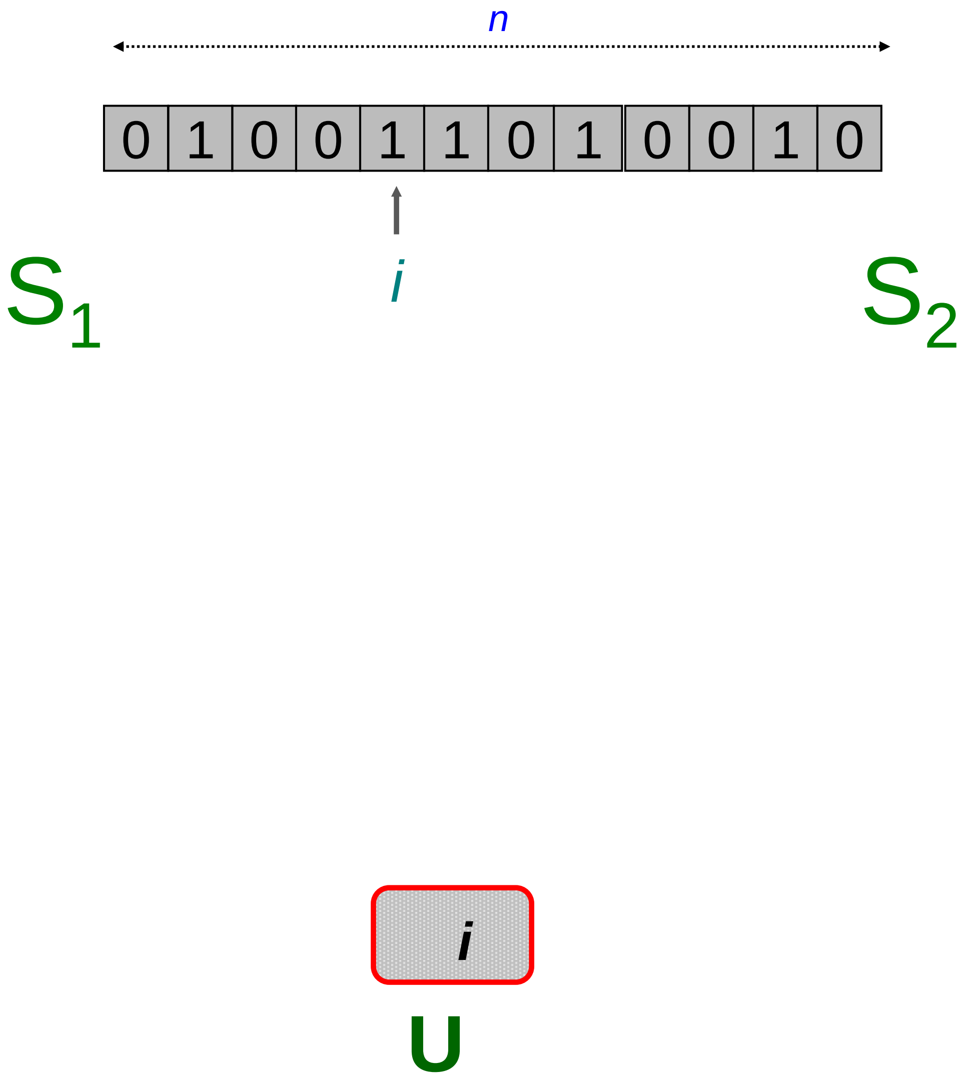
APPROACH I: *K*-SERVER PIR



Correctness: User obtains x_i

Privacy: No *single* server gets information about i

A 2-SERVER INFORMATION THEORETICAL PIR



n

0	1	0	0	1	1	0	1	0	0	1	0

i

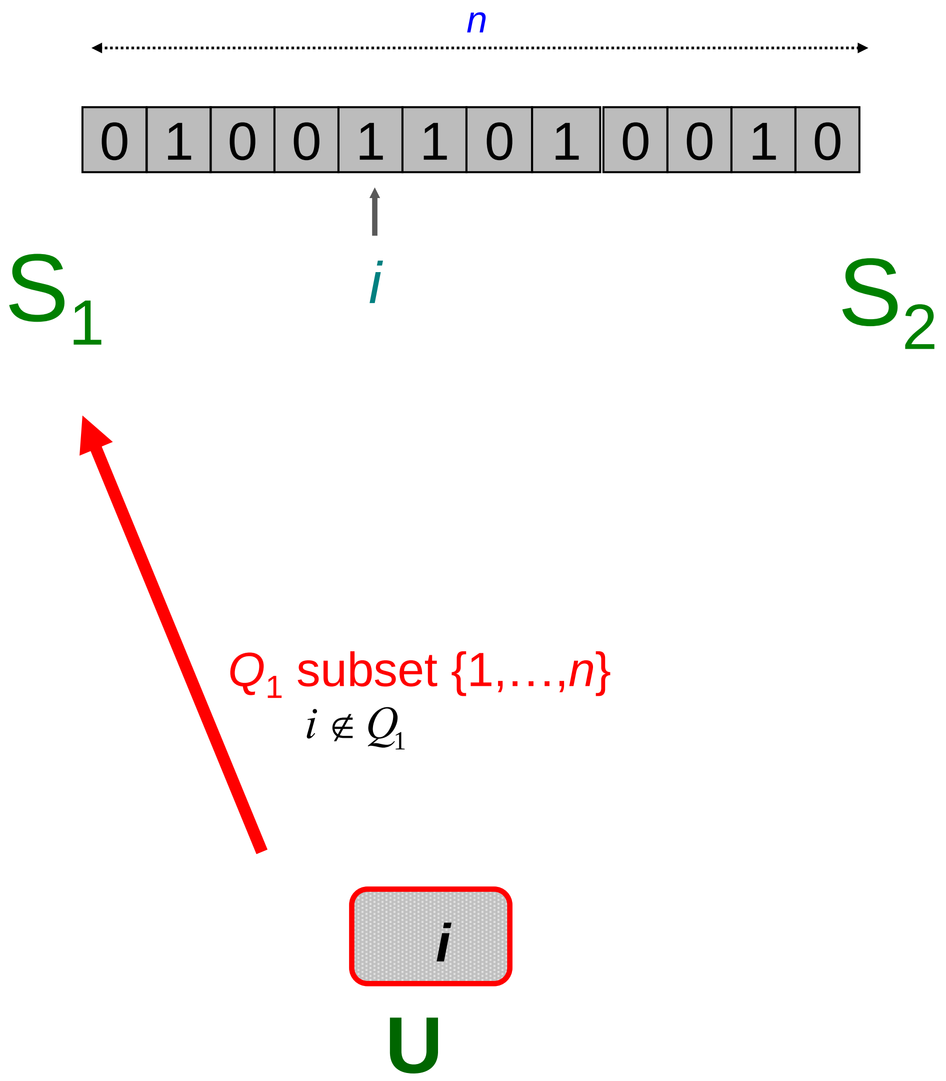
S_1

S_2

i

U

A 2-SERVER INFORMATION THEORETICAL PIR



n

0	1	0	0	1	1	0	1	0	0	1	0

i

S_1

S_2

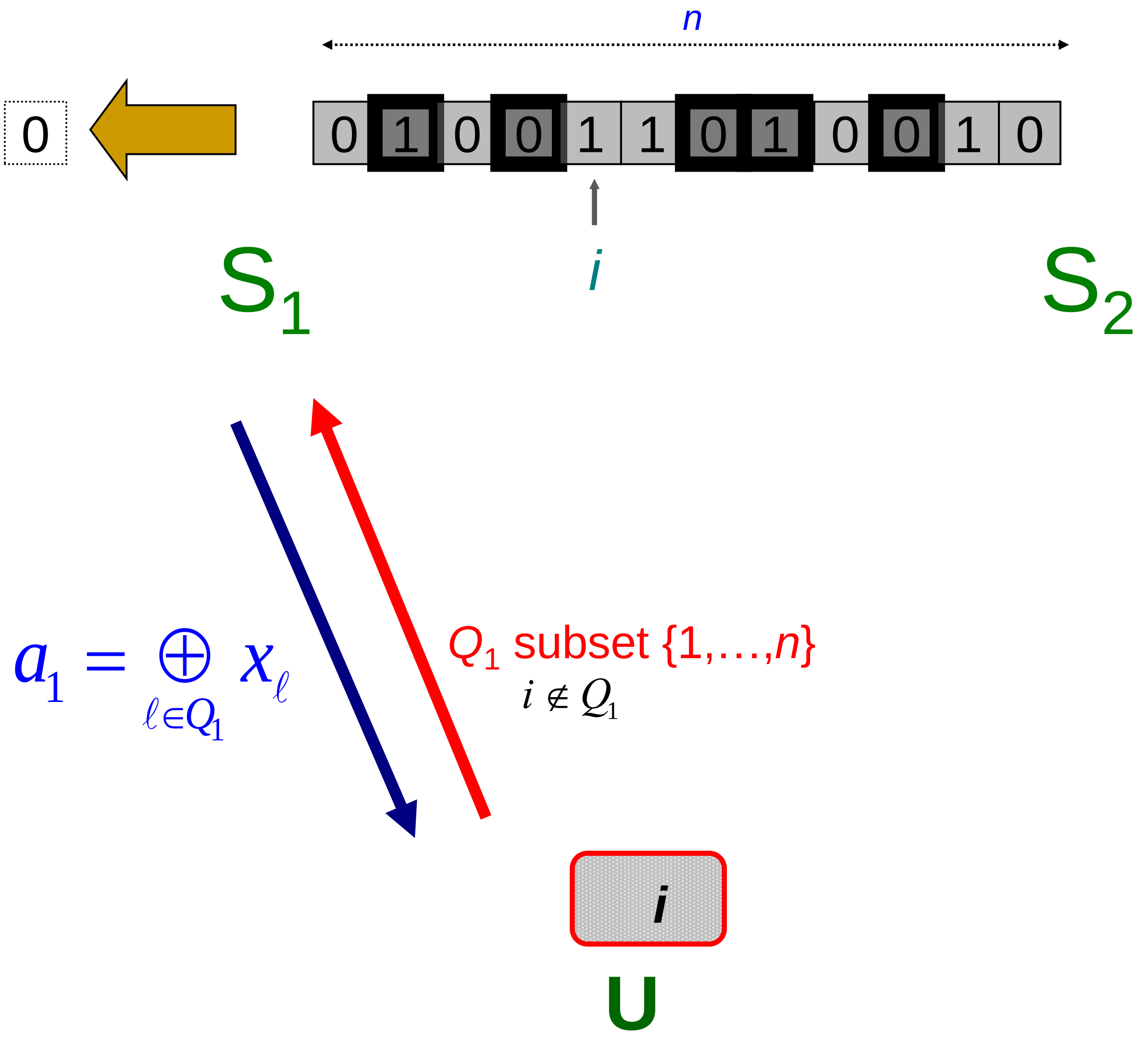
Q_1 subset $\{1,\ldots,n\}$

$i \notin Q_1$

i

U

PROTOCOL I: 2-SERVER PIR

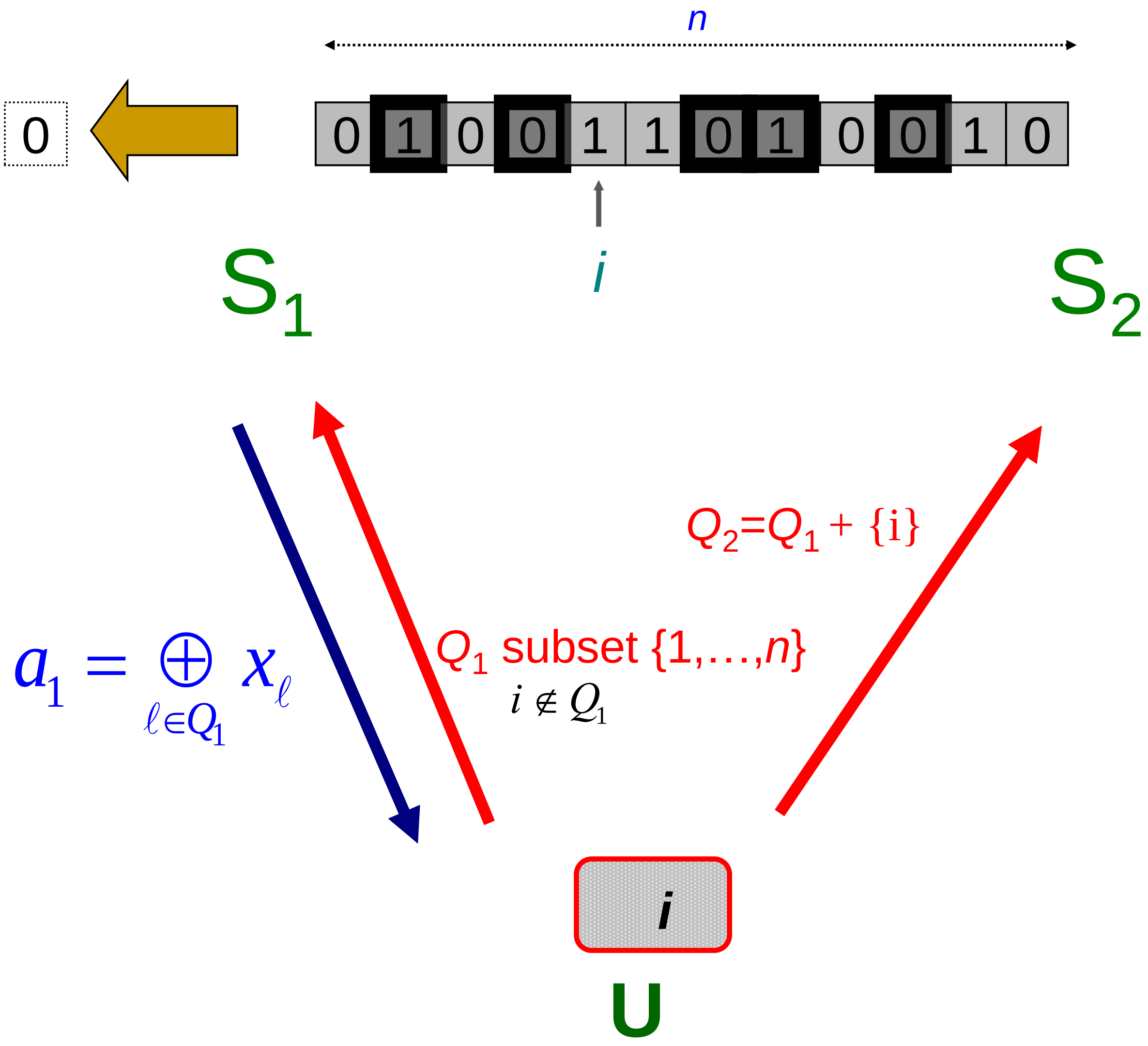


$$a_1 = \bigoplus_{\ell \in Q_1} x_\ell$$

Q_1 subset $\{1,\ldots,n\}$

$i \notin Q_1$

PROTOCOL I: 2-SERVER PIR



n

0

0 1 0 0 1 1 0 1 0 0 1 0

i

S_1

S_2

$Q_2=Q_1 + \{i\}$

Q_1 subset {1,…,*n*}

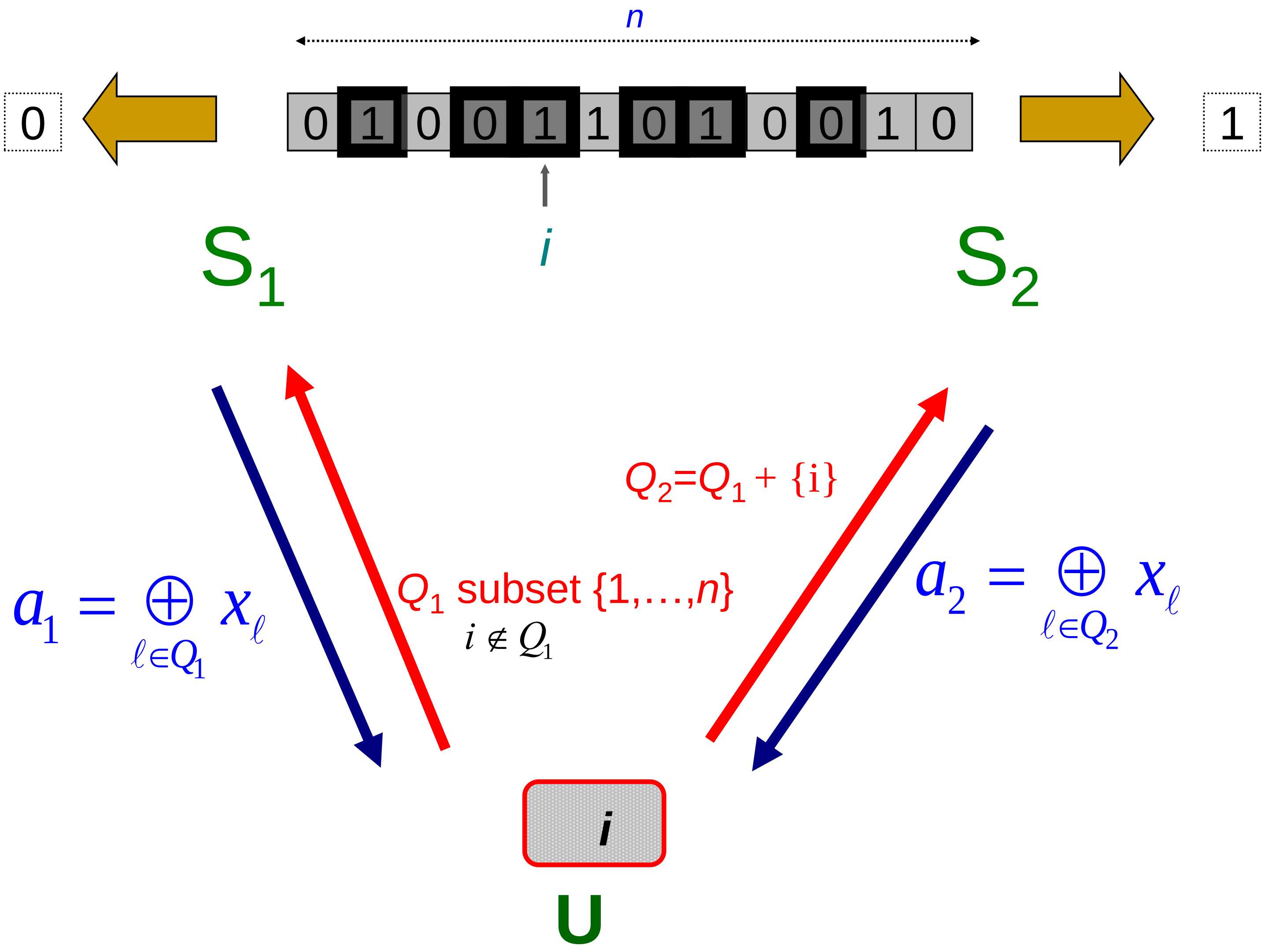
$i \notin Q_1$

$$a_1 = \bigoplus_{\ell \in Q_1} x_\ell$$

i

U

PROTOCOL I: 2-SERVER PIR



n

0 1 0 0 1 1 0 1 0 0 1 0

0

1

i

S_1

S_2

Q_2=Q_1 + {i}

Q_1 subset {1,…,*n*}

$i \notin Q_1$

$a_1 = \bigoplus_{\ell \in Q_1} x_\ell$

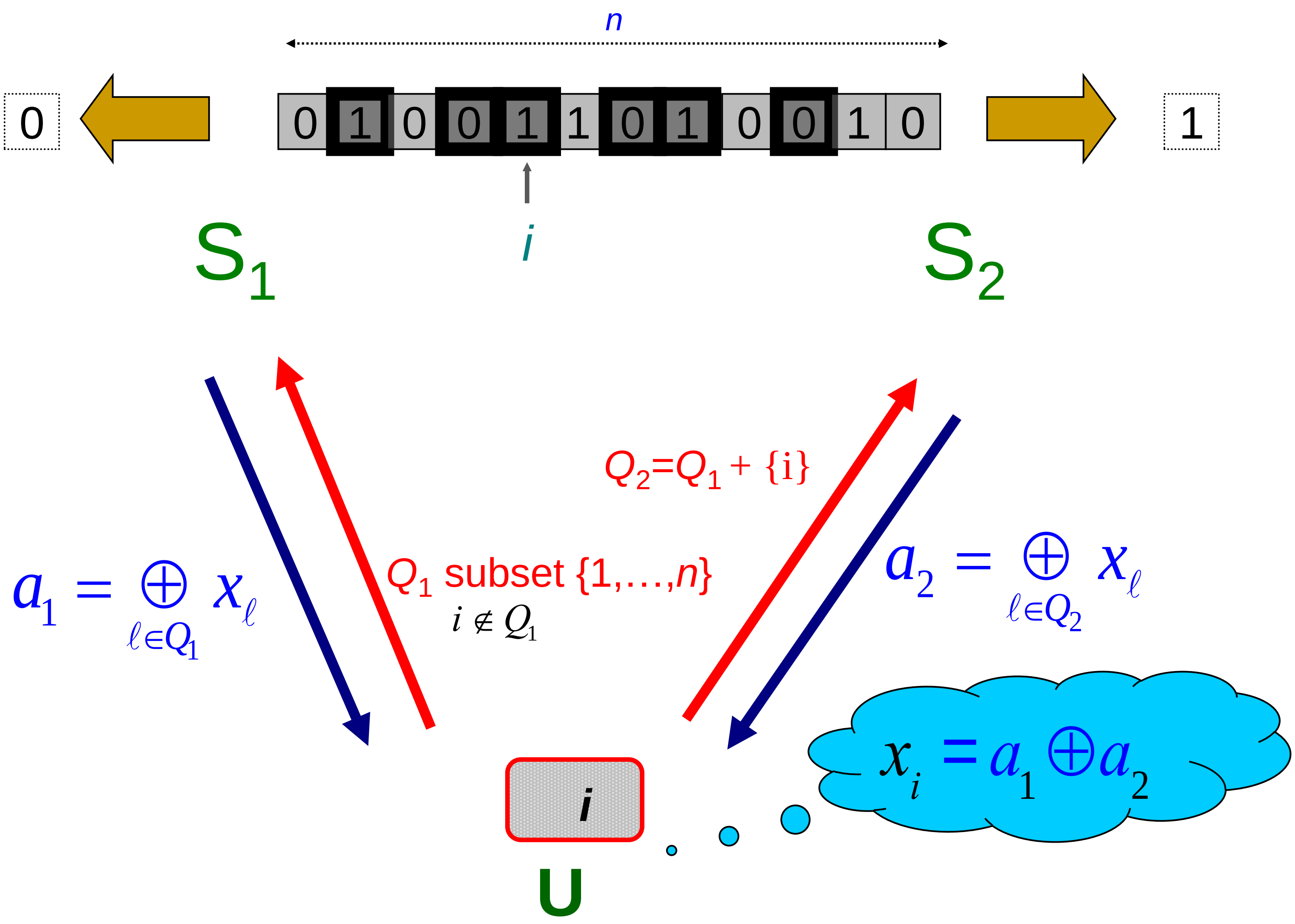
$a_2 = \bigoplus_{\ell \in Q_2} x_\ell$

i

U

Weakness: Servers should not collude!

PROTOCOL I: 2-SERVER PIR



n

0 1 0 0 1 1 0 1 0 0 1 0

0

1

i

S_1

S_2

Q_2=Q_1 + {i}

Q_1 subset {1,…,n}

$i \notin Q_1$

$$a_1 = \bigoplus_{\ell \in Q_1} x_\ell$$

$$a_2 = \bigoplus_{\ell \in Q_2} x_\ell$$

$$x_i = a_1 \oplus a_2$$

i

U

Weakness: Servers should not collude!

APPROACH II: COMPUTATIONAL PIR

Only one server, no need to trust

Based on cryptographic assumptions

Downside: Server has to run over the whole database, otherwise leaks information

- High computation load on the server