Approximate *k*-Nearest Neighbor Query over Spatial Data Federation

Kaining Zhang[1], Yongxin Tong[1], Yexuan Shi[1],
Yuxiang Zeng[1,2], Yi Xu[1], Lei Chen[2], Zimu Zhou[3],
Ke Xu[1], Weifeng Lv[1], Zhiming Zheng[1]

[1]Beihang University
[2]The Hong Kong University of Science and Technology
[3]City University of Hong Kong

Outline

- Background
- Problem Definition
- Our Solution
- Experiment
- Conclusion

Data fragmentation and isolation

- Data is the new oil
- Data exists in the form of isolated islands

“The world’s most valuable resource is no longer oil, but data” [1].

“In most industries, data exists in the form of isolated islands” [2].

[1] The World’s Most Valuable Resource Is No Longer Oil, but Data. The Economist. 2017
[2] Qiang Yang, et al. Federated Machine Learning: Concept and Applications. ACM TIST 2019

Data fragmentation and isolation

- Due to industry competition, data privacy, etc., "it is difficult to integrate the data scattered around institutions, or the cost is prohibited" [2]

2018 — General Data Protection Regulation (GDPR)

2020 — California Consumer Privacy Act (CCPA)

2021

Personal Information Protection Law (PIPL)

Strict Data Protection Laws

[2] Qiang Yang, et al. Federated Machine Learning: Concept and Applications. ACM TIST 2019

What is Data Federation?

- Data federation[3]: a set of data owners support a common schema, and each holds a horizontal partition (i.e., a subset of rows) [4] of the table

An illustration example

Virtual Database

Attr. #1	Attr. #2	Attr. #3	Attr. #4	Attr. #5	Attr. #6
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,1}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,1}$
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,1}$

D_1 Owner #1

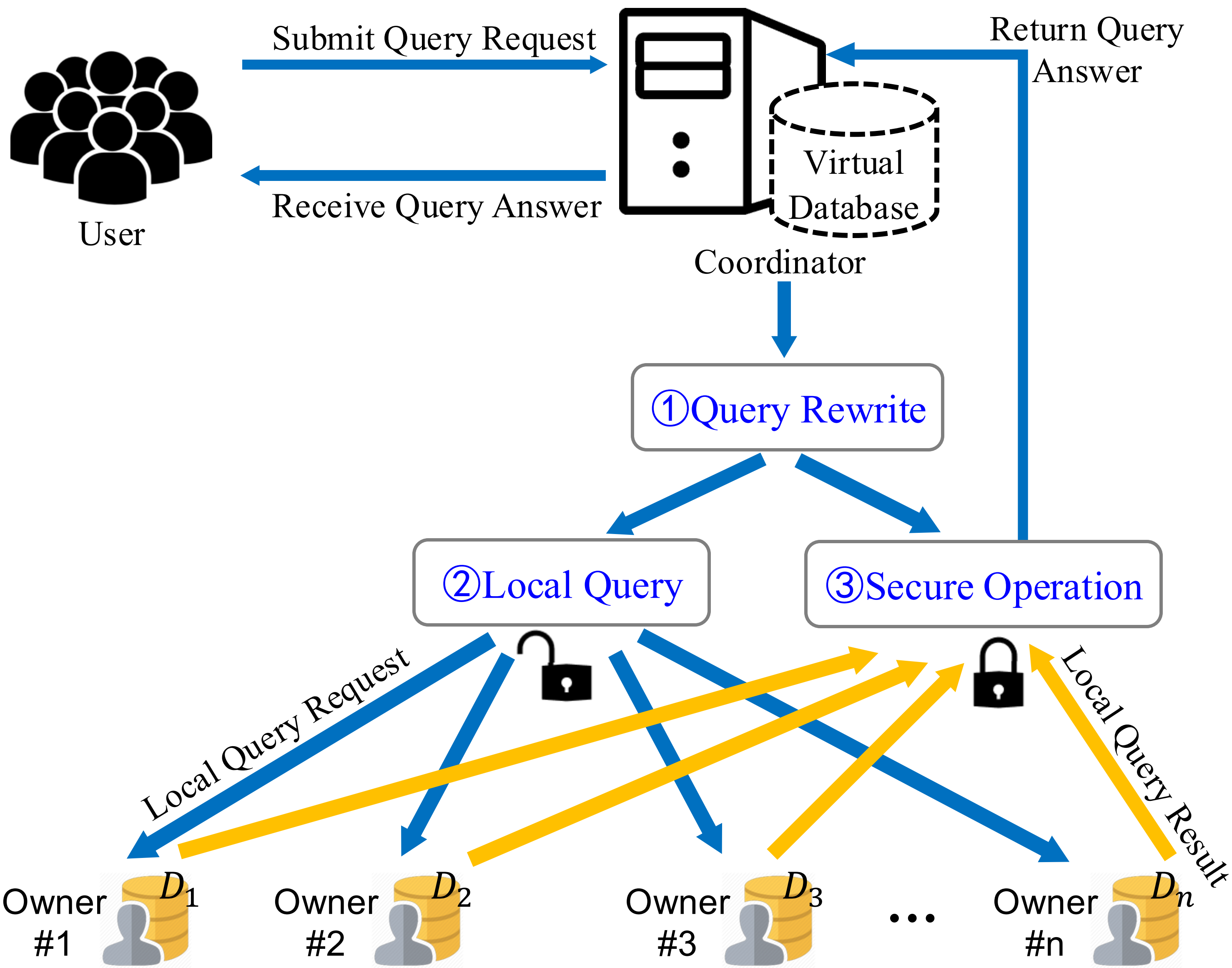
Attr. #1	Attr. #2	Attr. #3	Attr. #4	Attr. #5	Attr. #6
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,1}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,1}$

D_2 Owner #2

Attr. #1	Attr. #2	Attr. #3	Attr. #4	Attr. #5	Attr. #6
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,1}$

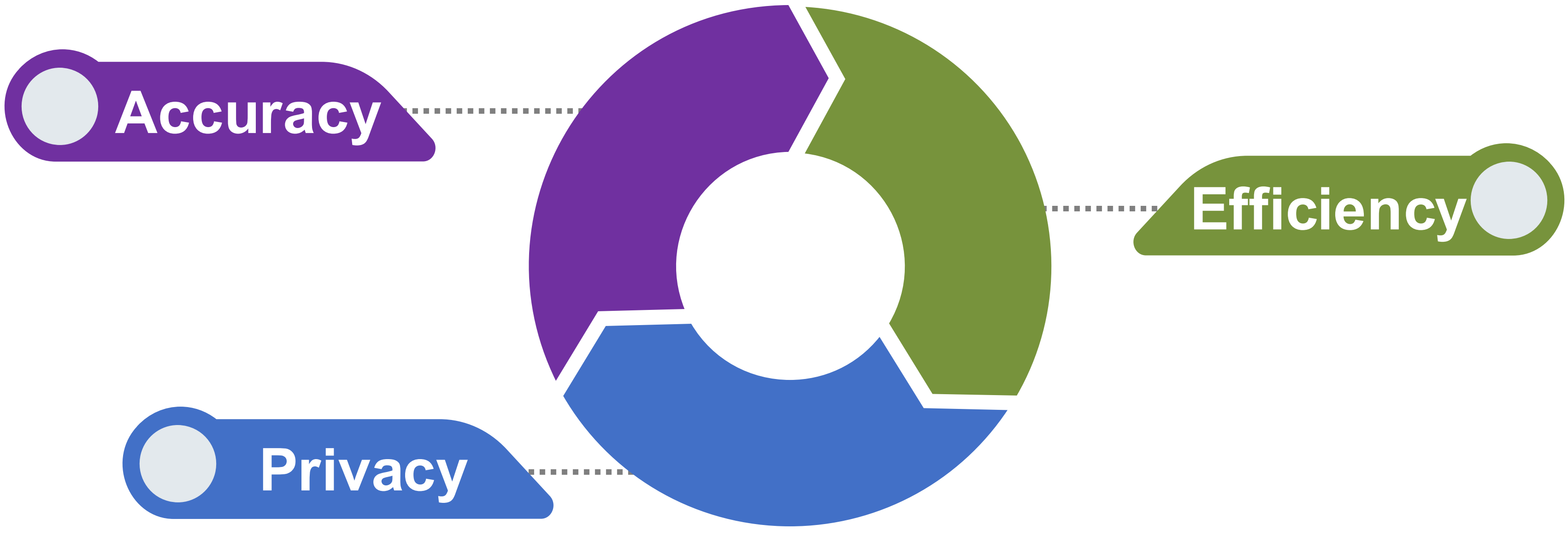
[3] Johes Bater, et al. SMCQL: Secure Query Processing for Private Data Networks. PVLDB 2017
[4] Akash Bharadwaj, Graham Cormode. An Introduction to Federated Computation. SIGMOD 2022

Workflow of data federation system



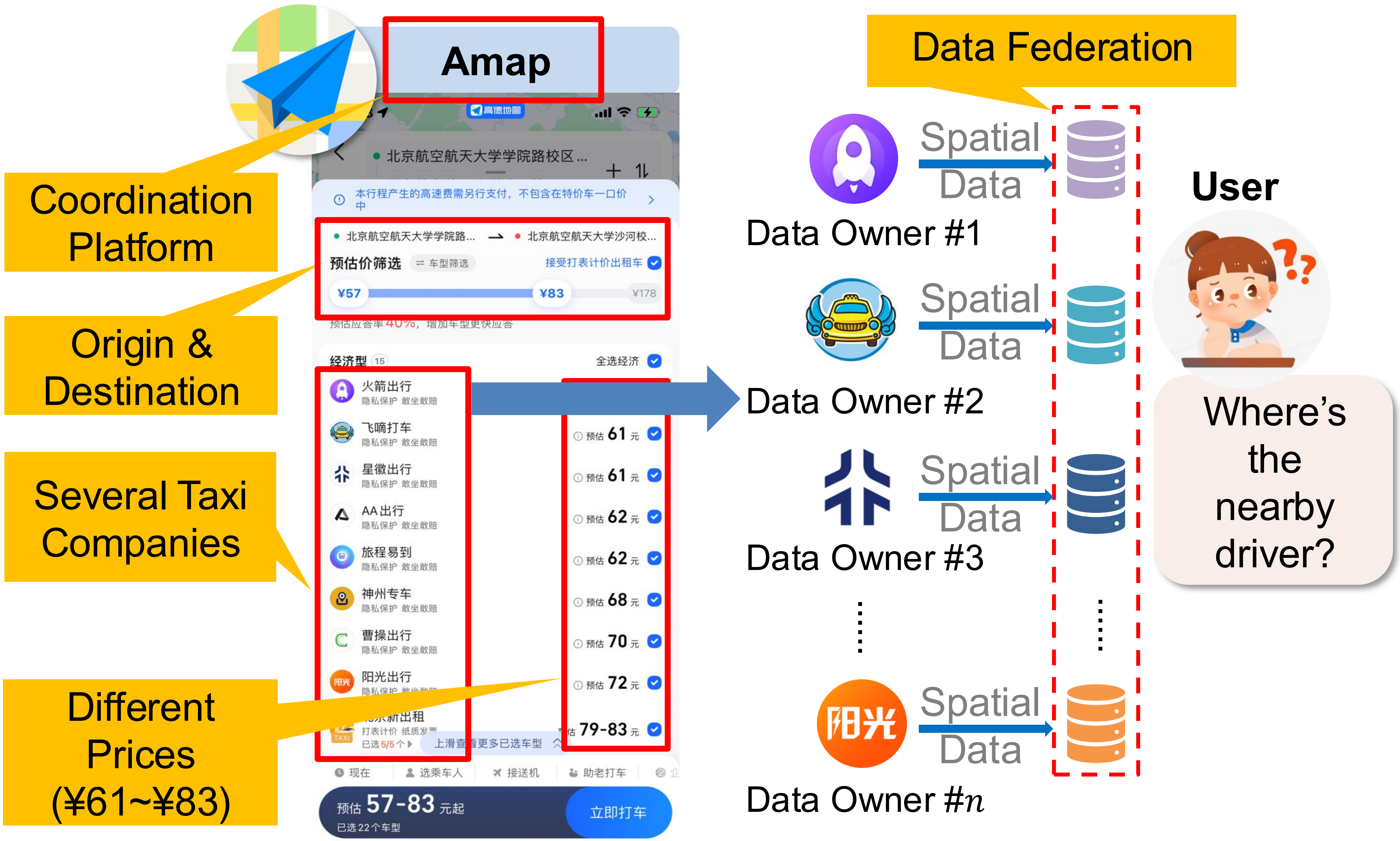
Challenges in Data Federation

- Accuracy
  - E.g., whether the query result is accurate enough
- Efficiency
  - E.g., whether the query efficiency can be real time
- Privacy
  - E.g., whether the private information is well protected



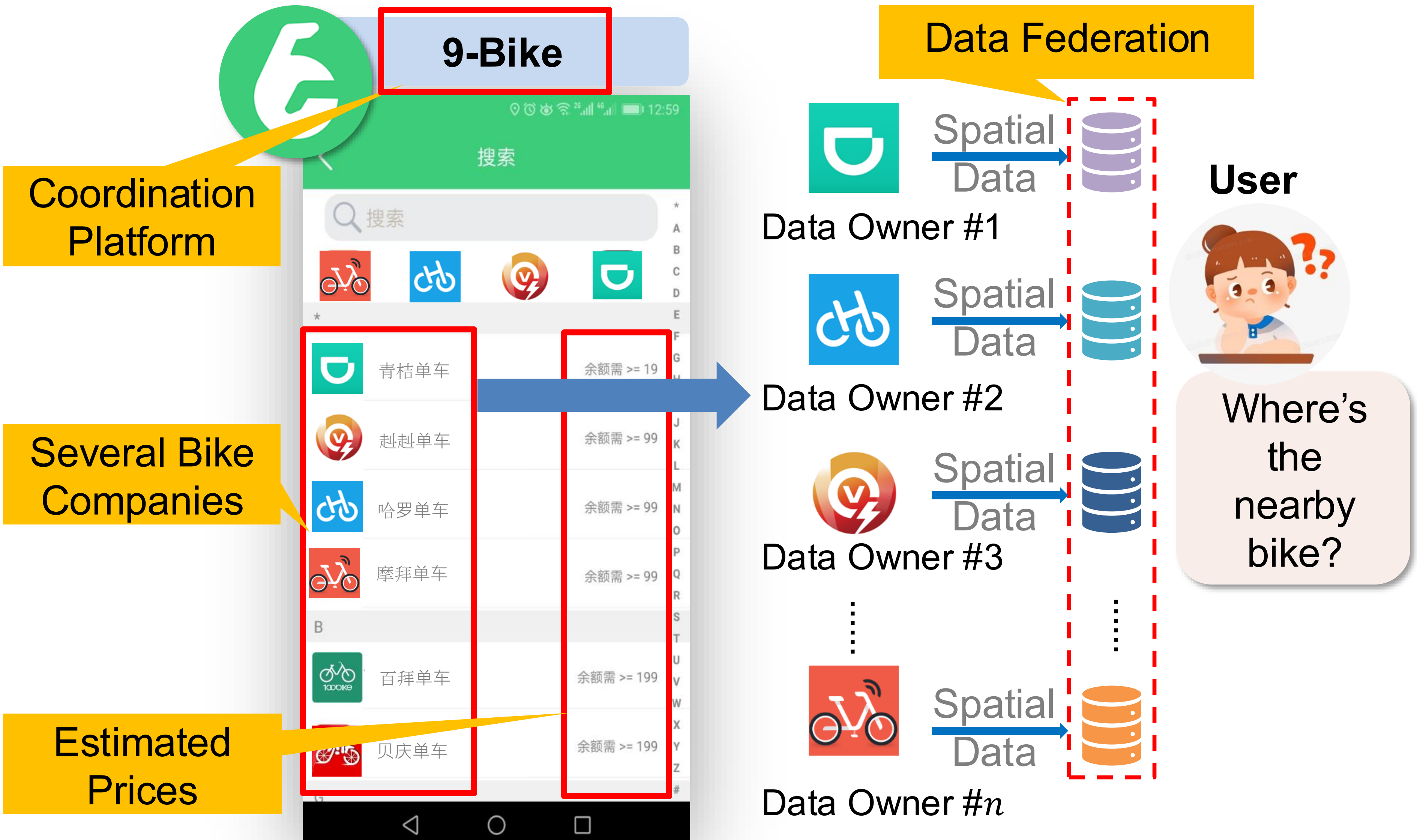
Spatial Data Federation

- Real applications: taxi calling, bike sharing, etc.



Spatial Data Federation

- Real Applications: taxi calling, bike sharing, etc.



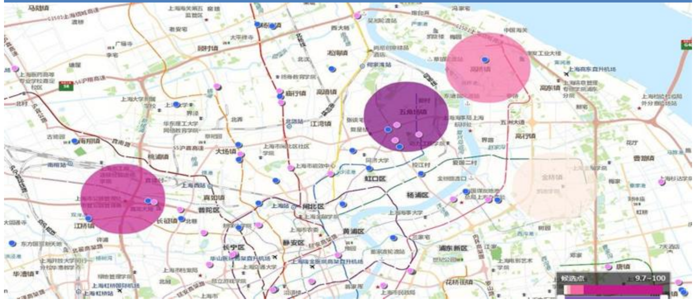
Approximate kNN Query

- Approximate kNN query plays an important role in applications of data federation

Taxi Calling

Bike Sharing

Map Navigation



Social Media

Outline

- Background
- Problem Definition
- Our Solution
- Experiment
- Conclusion

Basic Concepts

- Data owner S_i
  - Hold a local dataset D_i
    - Contains spatial objects $d_1, d_2, \dots, d_{|D_i|}$
  - He is not willing to share his raw data with other data owners S_j $(j \neq i)$ directly

- Spatial data federation F = $\{S_1, S_2, \dots, S_n\}$
  - Virtual dataset $D = D_1 \cup D_2 \cup \cdots \cup D_n$
    - Local dataset D_i is a horizontal partition
  - Coordinate data owners to answer spatial queries

Problem Definition

- Approximate kNN Query over Spatial Data Federation ("federated approximate kNN")
  - Given:
    - Spatial data federation F, query object l_q, integer k
  - Goal:
    - A set res that contains k spatial objects from F to maximize the accuracy δ (a popular metric [5,6])
      - $\delta = \frac{|res \cap res^*|}{k}$, where res^* is the answer of exact kNN
  - Constraint:
    - Security constraint: a data owner cannot infer any sensitive information from others except for res
    - Assumption: attackers are s

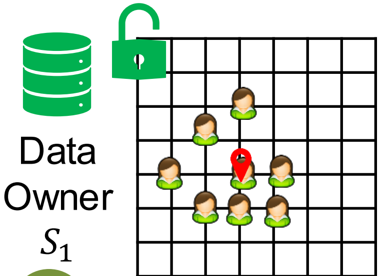
E.g., locations of other spatial objects

[5] Wen Li, et al. Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement. IEEE TKDE 2020
[6] Mengzhao Wang, et al. A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search. PVLDB 2021

Toy Example

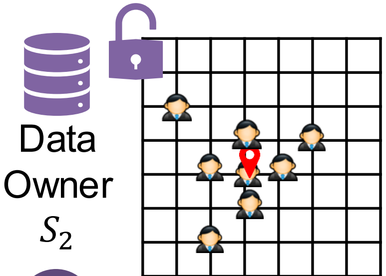
Spatial Data Federation F

Data Owner S_1



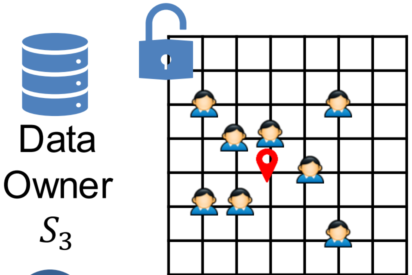
D_1	x	y
d_1	1	3
d_2	2	2
d_3	2	4
d_4	3	2
d_5	3	3
d_6	3	5
d_7	4	2
d_8	4	3

Data Owner S_2



D_2	x	y
d_1	1	5
d_2	2	1
d_3	2	3
d_4	3	2
d_5	3	3
d_6	3	4
d_7	4	3
d_8	5	4

Data Owner S_3



D_3	x	y
d_1	1	2
d_2	1	5
d_3	2	2
d_4	2	4
d_5	3	4
d_6	4	3
d_7	5	1
d_8	5	5

Query object $l_q = (3,3)$

Integer $k = 8$

res

	x	y
d_1	3	3
d_2	3	3
d_3	3	2
d_4	3	2
d_5	3	4
d_6	4	3
d_7	2	4
d_8	2	1

res^*

	x	y
d_1	3	3
d_2	3	3
d_3	3	2
d_4	3	2
d_5	3	4
d_6	4	3
d_7	4	3
d_8	4	3

$$res \cap res^* = \{d_1, d_2, d_3, d_4, d_5, d_6\}$$

$$\delta = \frac{|res \cap res^*|}{k} = \frac{6}{8} = 75\%$$

Existing Work: Exact Solution

SMCQL[3]/Conclave[7]

Idea: filter and refine

Local Exact kNN

↓

Secure Sort

↓

Secure Top-k

Hu-Fu[8]

Idea: binary search k^{th}NN distance

Local Exact Range Count

↓ ↑

Secure Comparison (to k)

↓

Local Range Query

↓

Secure Set Union

Efficiency bottleneck hinders real-time applications

Existing Work: Extension

SMCQL[3]/Conclave[7]		SMCQL[3]/Conclave[7]
Idea: filter and refine		Idea: filter and refine
Local Exact kNN	Extension →	**Local Approx. kNN**
↓		↓
Secure Sort		**Secure Sort**
↓		↓
Secure Top-k		**Secure Top-k**

Efficiency can be improved but with a very small margin

Existing Work: Extension

Hu-Fu[8]

Idea: binary search k^{th}NN distance

Local
Exact Range Count

Secure
Comparison (to k)

Secure
Set Union

Extension

Hu-Fu[8]

Idea: binary search k^{th}NN distance

Local
Approx. Range Count

Secure
Comparison (to k)

Local
Range Query

Secure
Set Union

Counter example (k = 5):

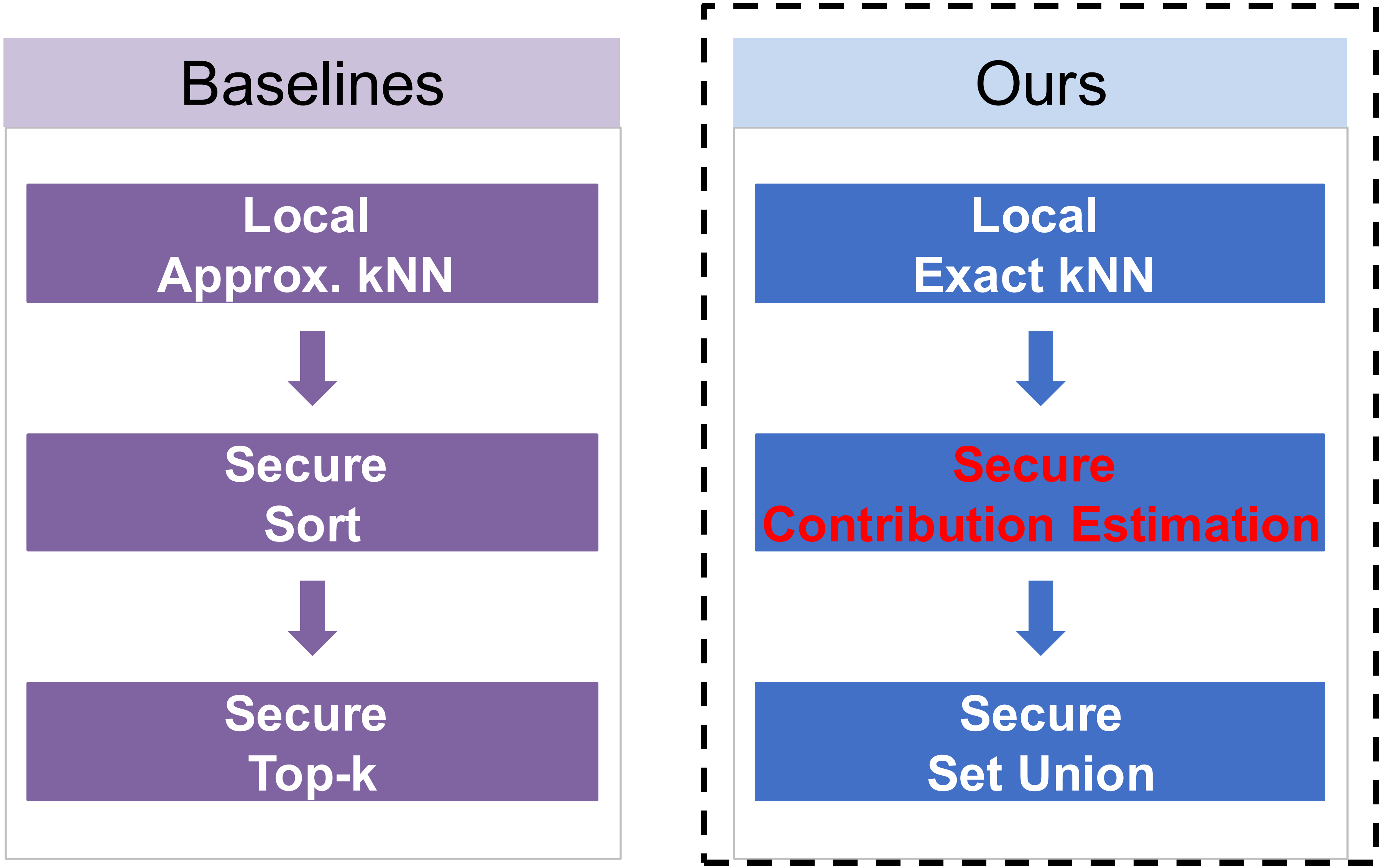
1. When kth distance is 10, sum of local approx. range counts is 7 (exact is 5).
2. When kth distance is 5, sum of local approx. range count is 6 (exact is 4).

Convergence of k^{th} NN distance becomes inaccurate

Outline

- Background
- Problem Definition
- Our Solution
- Experiment
- Conclusion

Our Framework

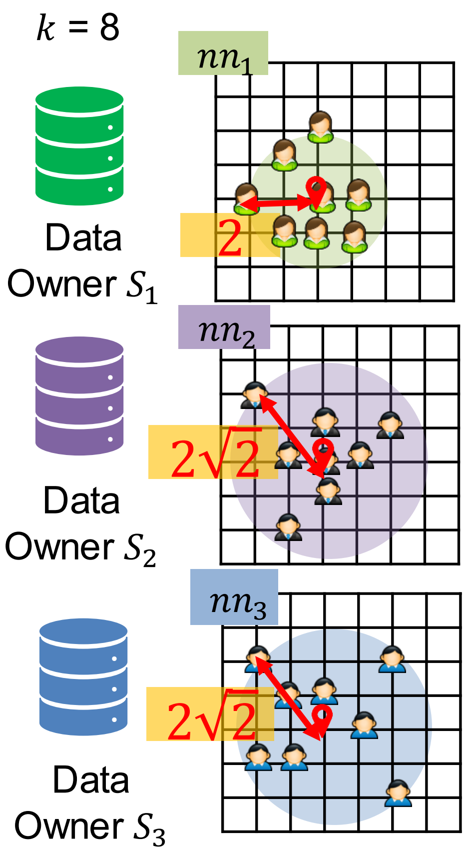


Our One-Round (OR) Algorithm

- Main Idea:
  - Warm-up: if we know the contribution of each data owner to the final answer, say $\{k_i\}$, then perform local k_iNN and securely unite partial results
  - Challenge: how to securely estimate contribution
    - Intuition: a data owner, whose kth nearest neighbor is closer, he tends to take more contributions
    - S1. Get local exact kNN and k^{th}NN distance r_i
    - S2. Compute each data owner's density
    - S3. Compute contribution proportional to density
    - S4. Collect final answer by secure set union

Toy Example

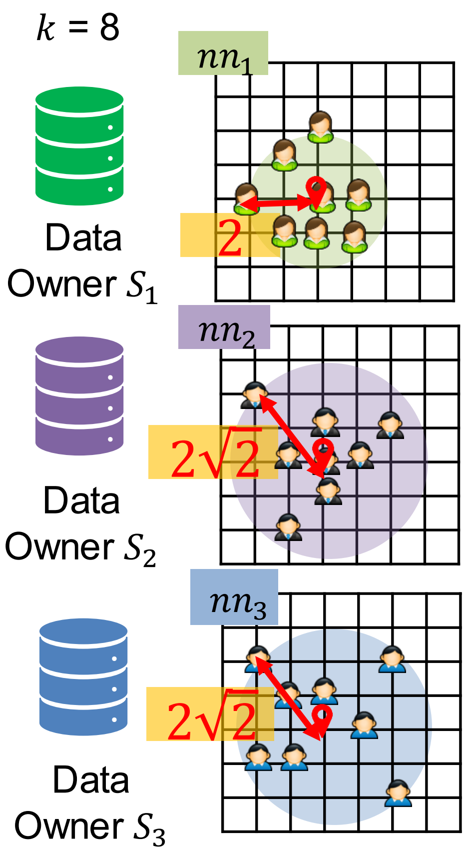
- Get local exact kNN and k[th]NN distance r_i
  - $nn_1 \leftarrow$ local exact kNN of S_1, $r_1 \leftarrow max_{j\in[1,k]} dis\left(l_{nn_1[j]}, l_q\right)$
  - $nn_2 \leftarrow$ local exact kNN of S_2, $r_2 \leftarrow max_{j\in[1,k]} dis\left(l_{nn_2[j]}, l_q\right)$
  - $nn_3 \leftarrow$ local exact kNN of S_3, $r_3 \leftarrow max_{j\in[1,k]} dis\left(l_{nn_3[j]}, l_q\right)$



	r_i
S_1	2
S_2	$2\sqrt{2}$
S_3	$2\sqrt{2}$

Toy Example

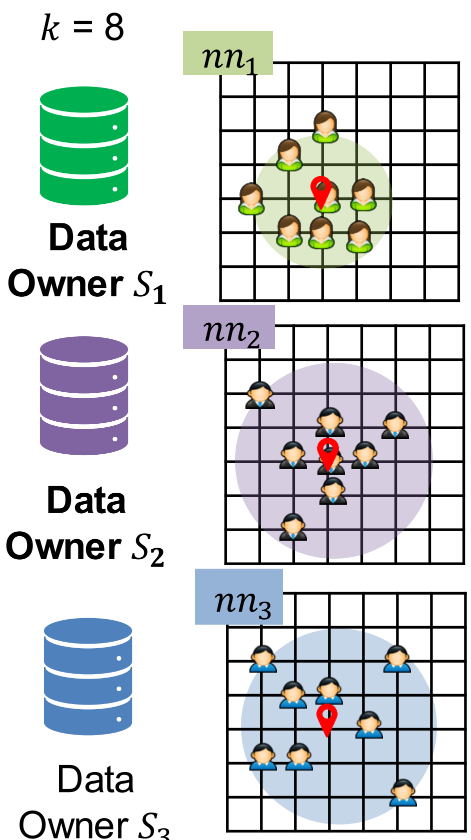
- Compute each data owner's density $k/area_i$
  - $area_1 \leftarrow \pi(r_1)^2$, $density_1 \leftarrow k/4\pi$
  - $area_2 \leftarrow \pi(r_2)^2$, $density_2 \leftarrow k/8\pi$
  - $area_3 \leftarrow \pi(r_3)^2$, $density_3 \leftarrow k/8\pi$



	r_i	$area_i$	$density_i$
S_1	2	4π	$k/4\pi$
S_2	$2\sqrt{2}$	8π	$k/8\pi$
S_3	$2\sqrt{2}$	8π	$k/8\pi$

Toy Example

- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]
    - $sum = SecureSum\left(\frac{k}{4\pi}, \frac{k}{8\pi}, \frac{k}{8\pi}\right) = \frac{k}{2\pi}$



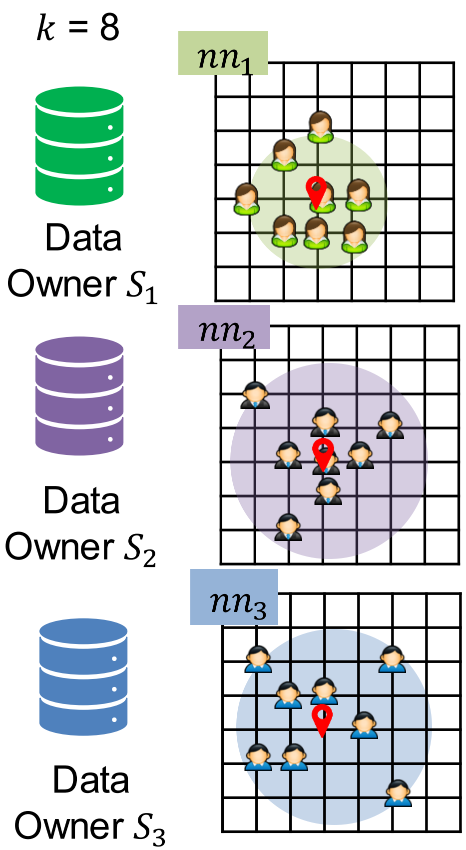
	r_i	$area_i$	$density_i$
S_1	2	4π	$k/4\pi$
S_2	$2\sqrt{2}$	8π	$k/8\pi$
S_3	$2\sqrt{2}$	8π	$k/8\pi$

$$sum = SecureSum\left(\frac{k}{4\pi}, \frac{k}{8\pi}, \frac{k}{8\pi}\right) = \frac{k}{2\pi}$$

[9] Kallista A. Bonawitz et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning. CCS 2017

Toy Example

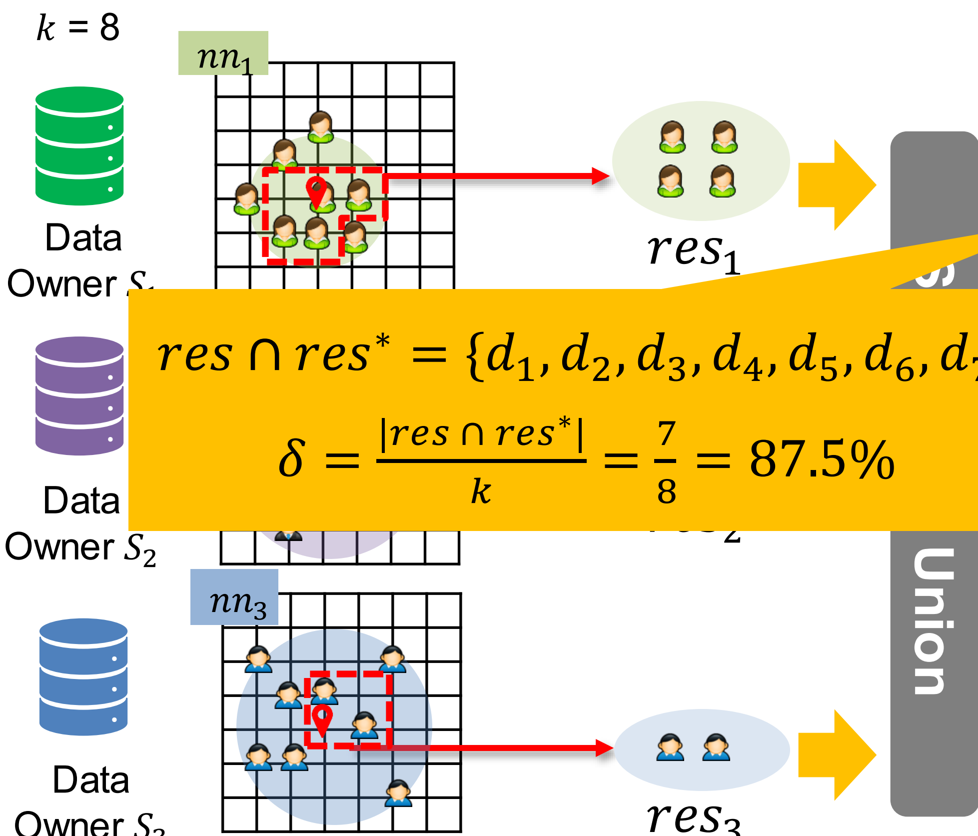
- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]
  - Compute contribution $k_i = k \times \frac{density_i}{sum}$, $sum = \sum_i density_i = \frac{k}{2\pi}$



	r_i	$area_i$	$density_i$	k_i
S_1	2	4π	$k/4\pi$	$k/2$
S_2	$2\sqrt{2}$	8π	$k/8\pi$	$k/4$
S_3	$2\sqrt{2}$	8π	$k/8\pi$	$k/4$

Toy Example

- Collect final answer by secure set union
  - Each data owner pick k_iNN as the partial answer
  - Collect partial answers by secure set union protocol in [10]



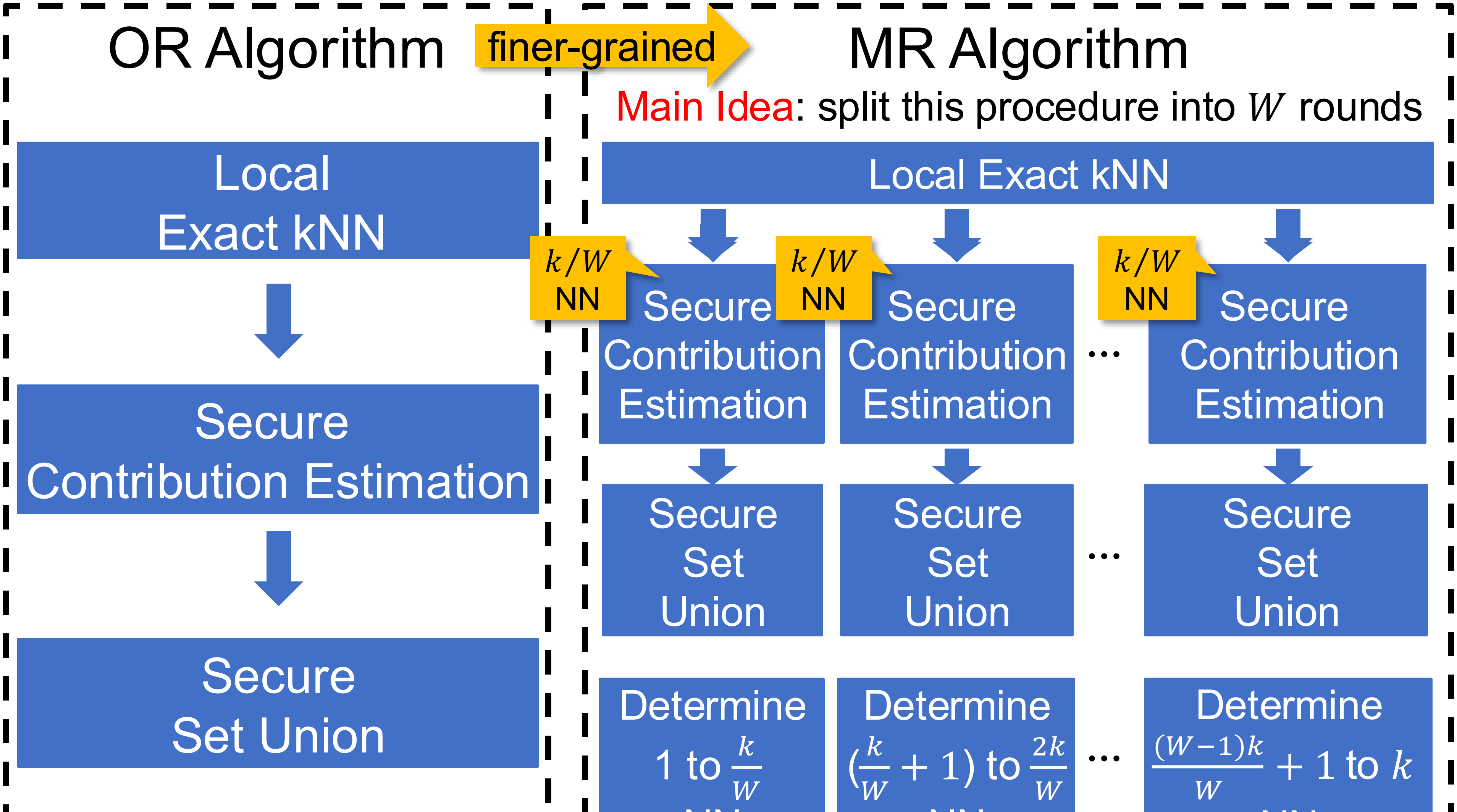
$$res \cap res^* = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$$

$$\delta = \frac{|res \cap res^*|}{k} = \frac{7}{8} = 87.5\%$$

Our Answer	x	y	Exact Answer	x	y
d_1	3	3	d_1	3	3
d_2	3	3	d_2	3	3
d_3	3	2	d_3	3	2
d_4	3	2	d_4	3	2
d_5	3	4	d_5	3	4
d_6	4	3	d_6	4	3
d_7	4	3	d_7	4	3
d_8	2	2	d_8	4	3

[10] Pawel Jurczyk et al. Information Sharing across Private Databases: Secure Union Revisited. SocialCom/PASSAT 2011

Optimization: Our MR algorithm



Approx. Guarantee: $Pr(\delta < 1 - \varepsilon) \le 2\exp(\frac{-2W\varepsilon^2}{n})$

Outline

- Background
- Problem Definition
- Our Solution
- Experiment
- Conclusion

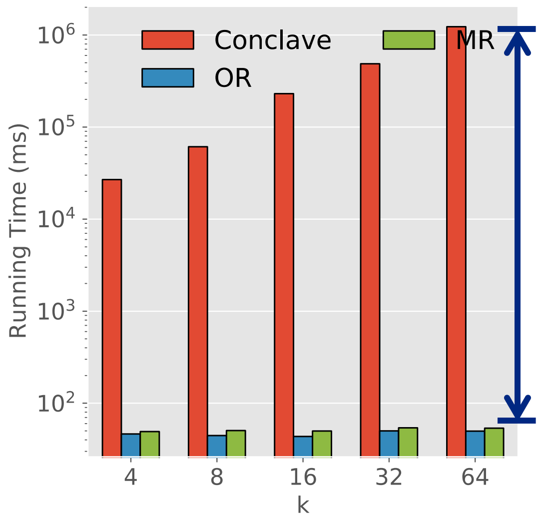
Experimental Setup

- **Datasets**
  - **Real Dataset:** Multi-company Spatial Data in Beijing (MBJ)
    - 10 data owners
    - Up to 10^6 spatial objects
  - **Synthetic Datasets:** OpenStreetMap (OSM)
    - 6 data owners
    - Up to 10^8 spatial objects
- **Baselines**
  - SMCQL: extended from [3] that only supports 2 data owners
  - Conclave: extended from [7] that supports ≥2 data owners
  - Both baselines have security and approximation guarantees

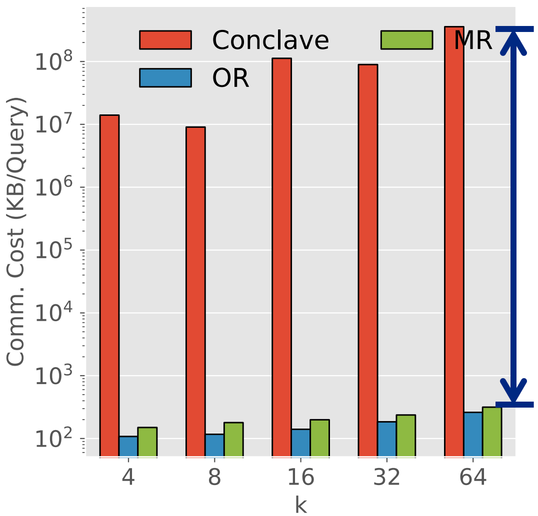
[3] Johes Bater, Gregory Elliott, et al. SMCQL: Secure Query Processing for Private Data Networks. PVLDB 2017
[7] Nikolaj Volgushev, Malte Schwarzkopf, et al. Conclave: Secure Multi-party Computation on Big Data. EuroSys 2019

Experimental Result

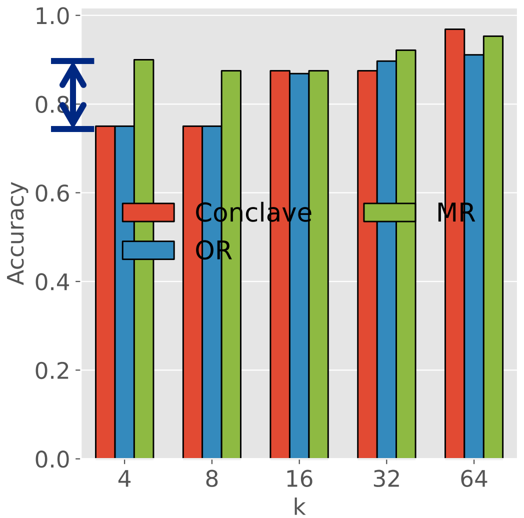
- Varying parameter k on MBJ dataset
  - Running time
    - 2～4 orders of magnitude shorter
  - Communication cost
    - 4～6 orders of magnitude lower
  - Accuracy
    - Can be 13% higher



(a) Running Time



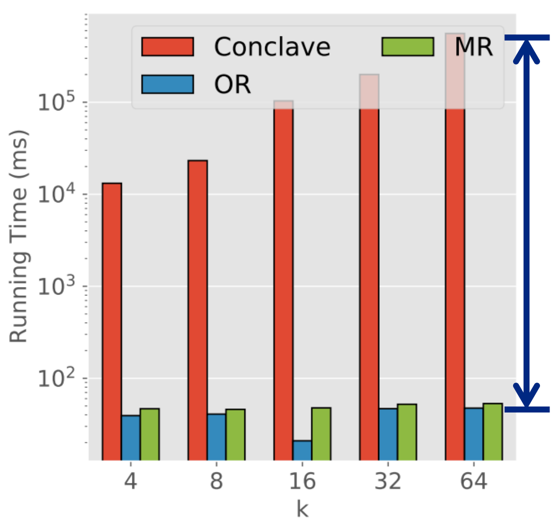
(b) Communication Cost



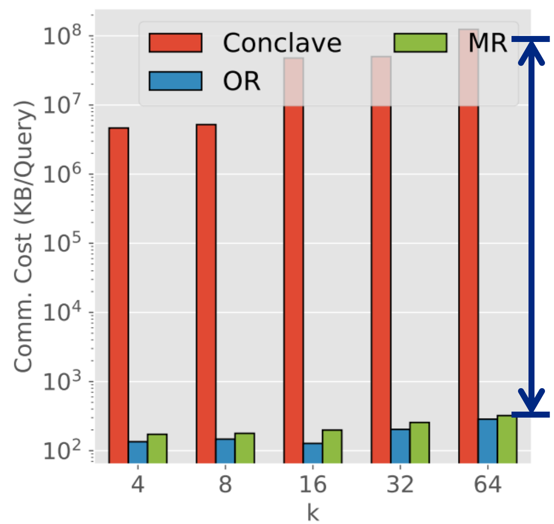
(c) Accuracy

Experimental Result

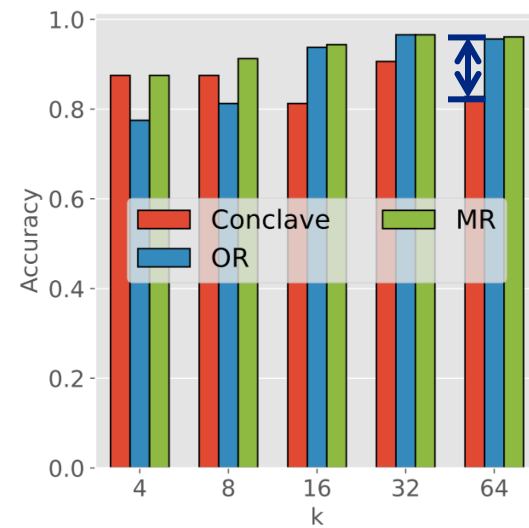
- Varying parameter k on OSM dataset
  - Running time
    - Up to 3 orders of magnitude shorter
  - Communication cost
    - Up to 5 orders of magnitude lower
  - Accuracy
    - More robust



(a) Running Time



(b) Communication Cost



(c) Accuracy

Outline

- Background
- Problem Definition
- Our Solution
- Experiment
- Conclusion

Conclusion

- Data federation is one of the most popular solutions to data fragmentation and isolation
- We studied approximate kNN query over large-scale spatial data federation
- We proposed solutions with theoretical analysis on complexity and approximation guarantee
- Extensive experiments demonstrate superiority performance of our solution in terms of efficiency and accuracy

Q & A

Thank You !

References

[1] The World's Most Valuable Resource Is No Longer Oil, but Data. The Economist. 2017

[2] Qiang Yang, et al. Federated Machine Learning: Concept and Applications. ACM TIST 2019

[3] Johes Bater, et al. SMCQL: Secure Query Processing for Private Data Networks. PVLDB 2017.

[4] Akash Bharadwaj, Graham Cormode. An Introduction to Federated Computation. SIGMOD 2022

[5] Wen Li, et al. Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement. IEEE TKDE 2020

[6] Mengzhao Wang, et al. A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search. PVLDB 2021

[7] Nikolaj Volgushev, et al. Conclave: Secure Multi-party Computation on Big Data. EuroSys 2019

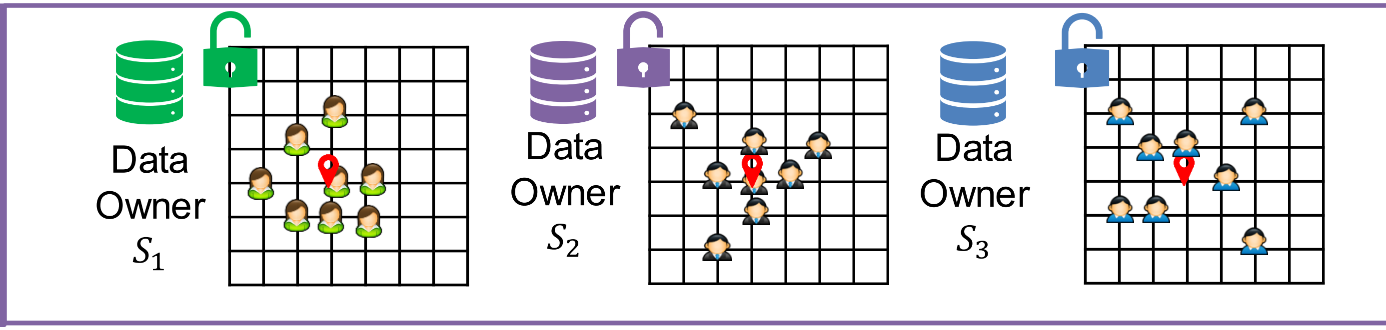
[8] Yongxin Tong, et al. Hu-Fu: Efficient and Secure Spatial Queries over Data Federation. PVLDB 2022

[9] Kallista A. Bonawitz et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning. CCS 2017

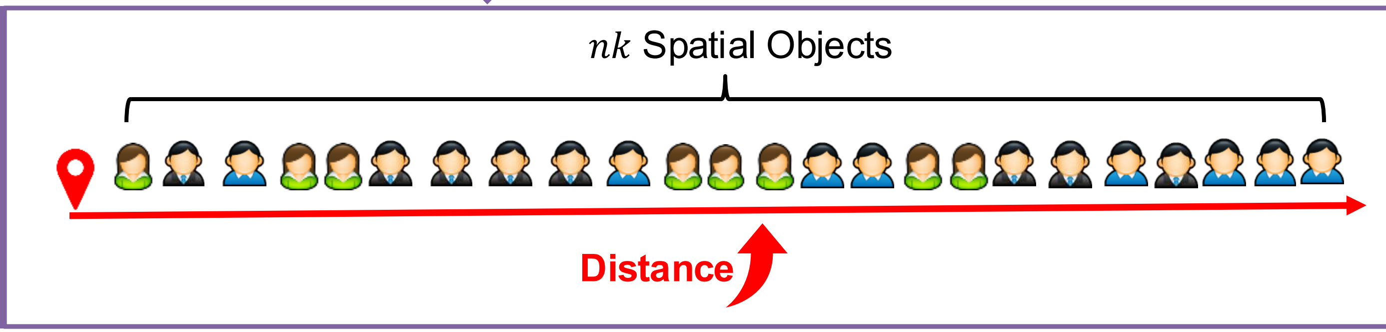
[10] Pawel Jurczyk et al. Information Sharing across Private Databases: Secure Union Revisited. SocialCom/PASSAT 2011

Existing Work: SMCQL/Conclave

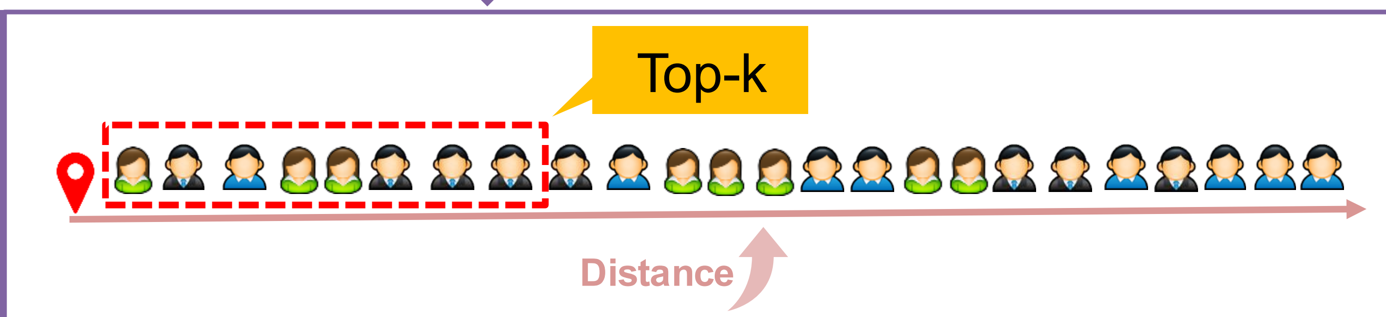
Local Exact kNN (E.g., k=8)



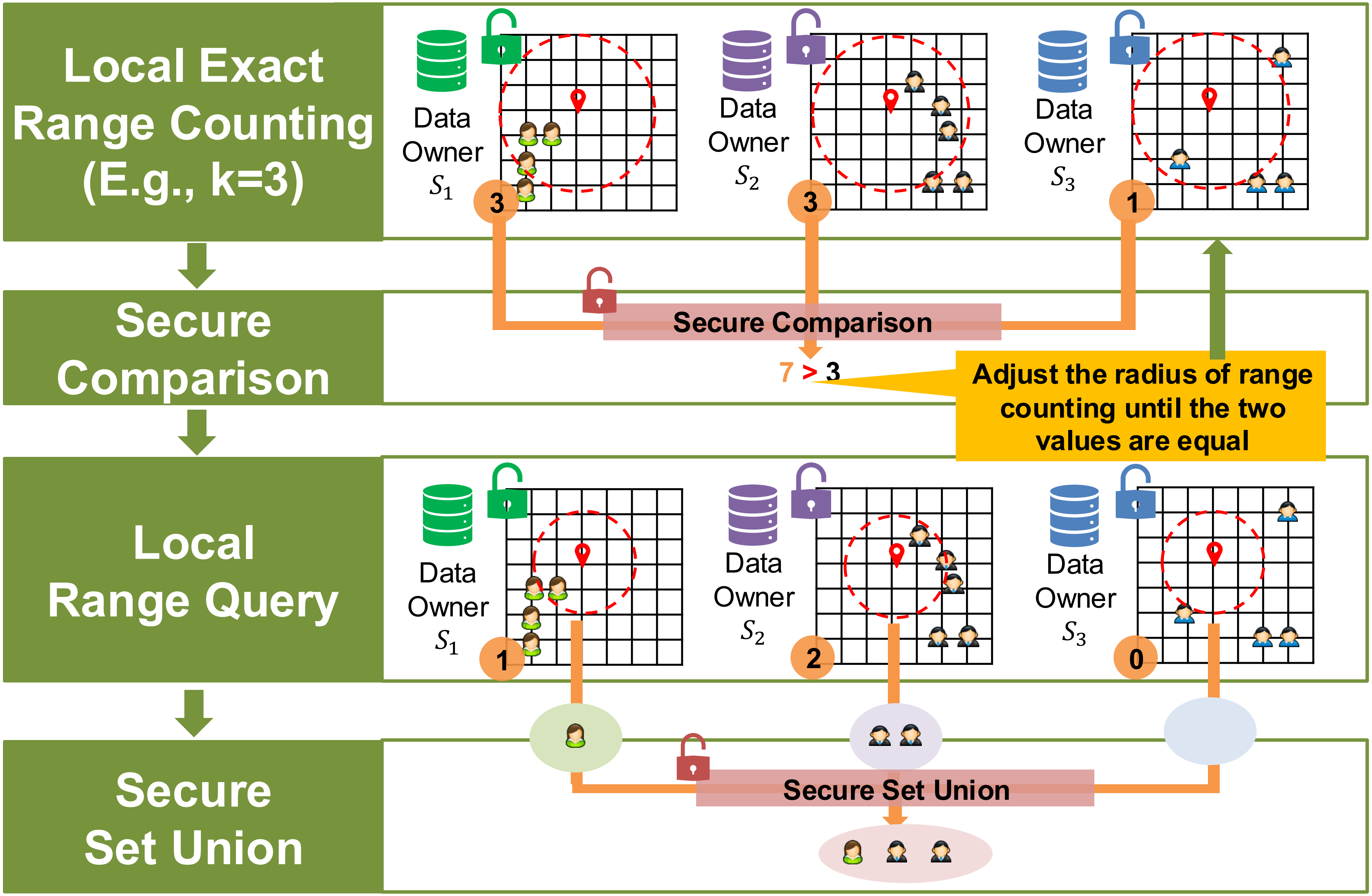
Secure Sort



Secure Top-k



Existing Work: Hu-Fu



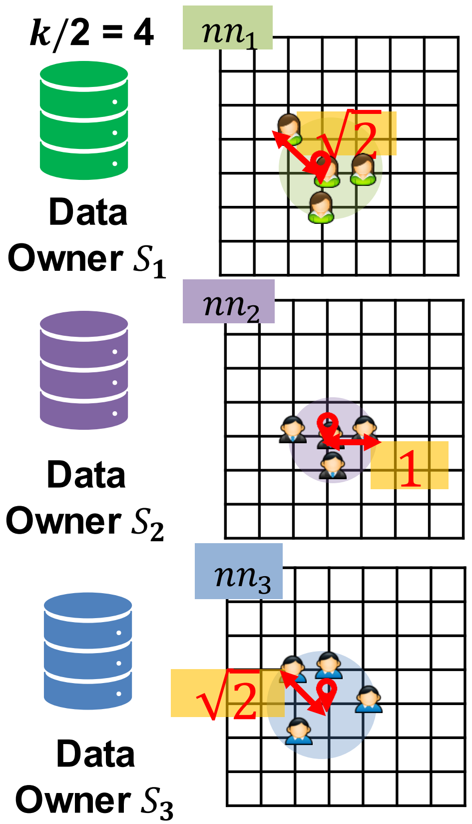
Toy Example

- Get local (k/2)NN and (k/2)th nearest distance r_i
  - $nn_1 \leftarrow$ local exact (k/2)NN of S_1, $r_1 \leftarrow max_{j\in[1,k/2]} dis(l_{nn_1[j]}, l_q)$
  - $nn_2 \leftarrow$ local exact (k/2)NN of S_2, $r_2 \leftarrow max_{j\in[1,k/2]} dis(l_{nn_2[j]}, l_q)$
  - $nn_3 \leftarrow$ local exact (k/2)NN of S_3, $r_3 \leftarrow max_{j\in[1,k/2]} dis(l_{nn_3[j]}, l_q)$

1st Round

k = 8

$k/2$ = 4



	r_i
S_1	$\sqrt{2}$
S_2	1
S_3	$\sqrt{2}$

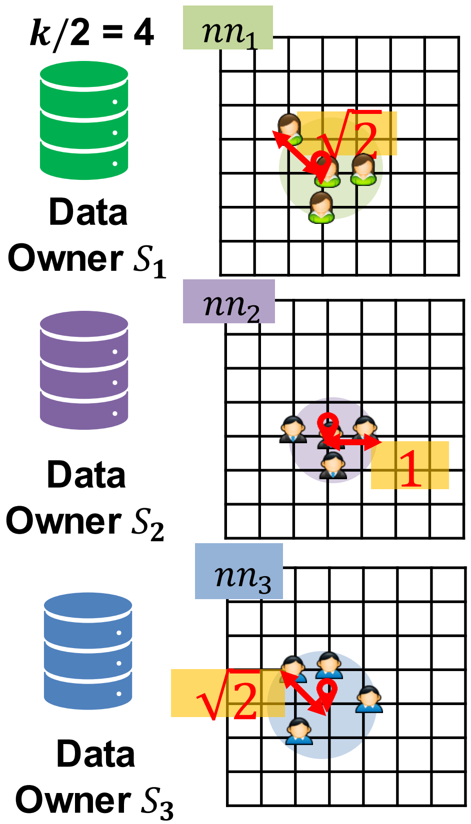
Toy Example

- Compute each data owner's density $({}^{k}/_{2})/area_i$
  - $area_1 \leftarrow \pi(r_1)^2$, $density_1 \leftarrow ({}^{k}/_{2})/2\pi$
  - $area_2 \leftarrow \pi(r_2)^2$, $density_2 \leftarrow ({}^{k}/_{2})/\pi$
  - $area_3 \leftarrow \pi(r_3)^2$, $density_3 \leftarrow ({}^{k}/_{2})/2\pi$

1st Round

$\boldsymbol{k}$ **= 8**

$\boldsymbol{k/2}$ **= 4**



	r_i	$area_i$	$density_i$
S_1	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$
S_2	1	π	$({}^{k}/_{2})/\pi$
S_3	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$

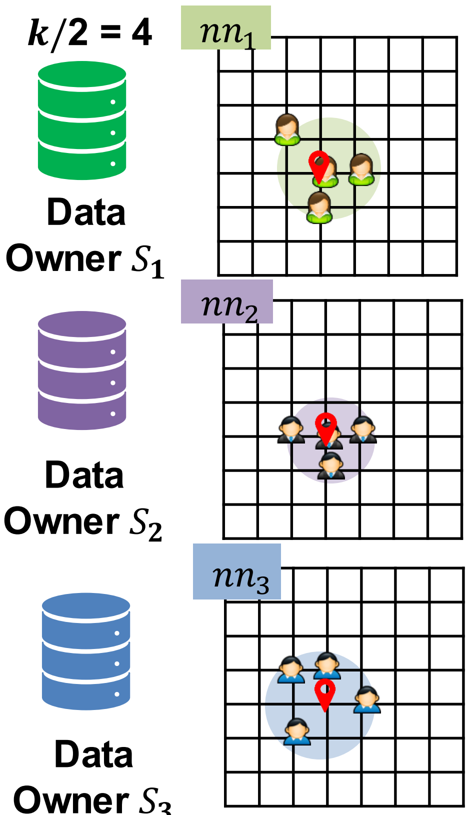
Toy Example

- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]
    - $sum = SecureSum\left(\frac{k/2}{2\pi}, \frac{k/2}{\pi}, \frac{k/2}{2\pi}\right) = \frac{k}{\pi}$

1st Round

k = 8

$k/2$ = 4



	r_i	$area_i$	$density_i$
S_1	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$
S_2	1	π	$({}^{k}/_{2})/\pi$
S_3	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$

$$sum = SecureSum\left(\frac{k/2}{2\pi}, \frac{k/2}{\pi}, \frac{k/2}{2\pi}\right) = \frac{k}{\pi}$$

[9] Kallista A. Bonawitz et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning. CCS 2017

Toy Example

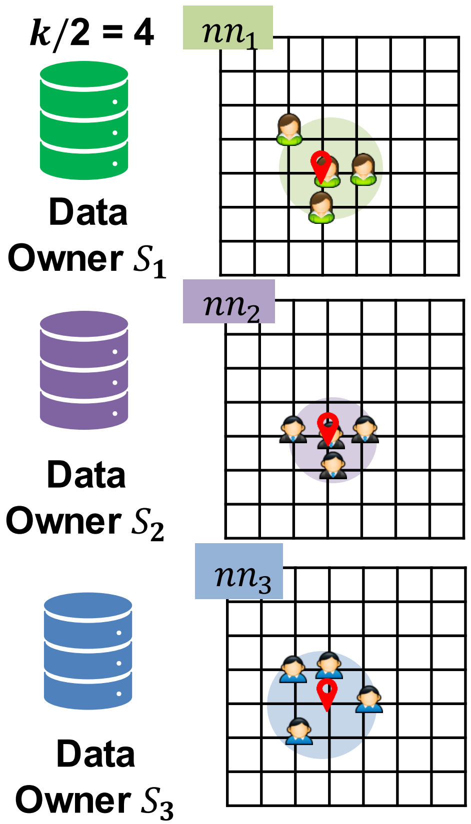
- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]

1st Round

Compute contribution $k_i = density_i \times \frac{k/2}{sum}$, $sum = \frac{k}{\pi}$

$k = 8$

$k/2 = 4$

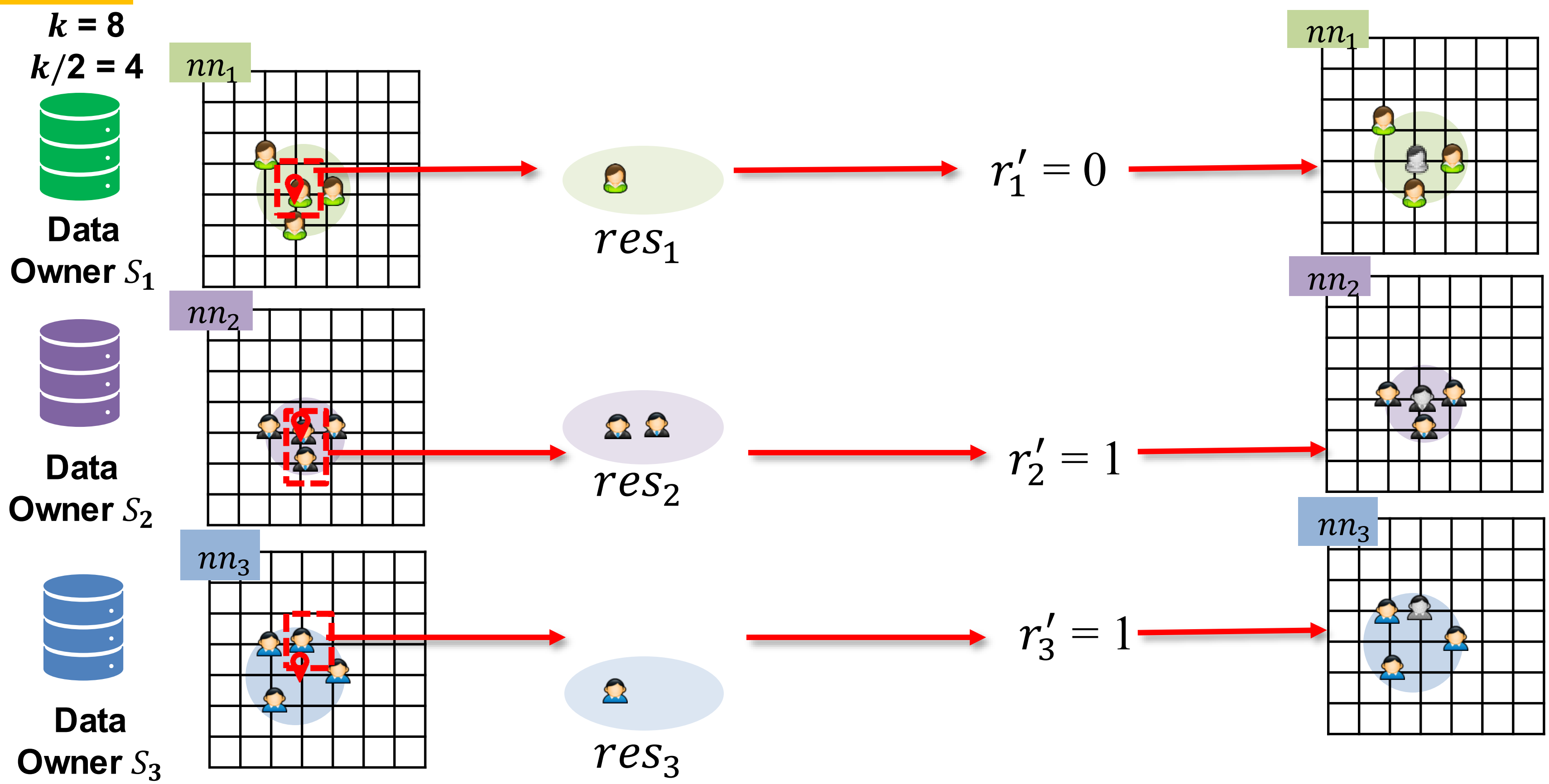


	r_i	$area_i$	$density_i$	k_i
S_1	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$	$k/8$
S_2	1	π	$({}^{k}/_{2})/\pi$	$k/4$
S_3	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$	$k/8$

Toy Example

- Each data owner pick k_iNN as the partial answer

1st Round



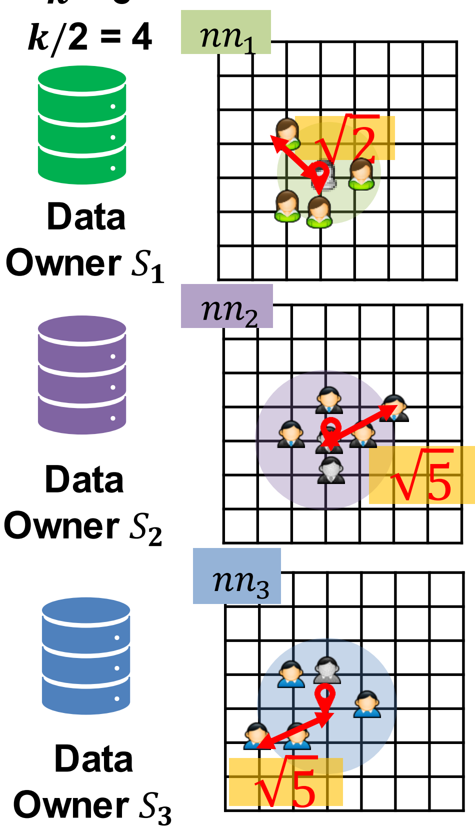
Toy Example

- Get local (k/2)NN and (k/2)th nearest distance r_i
  - $nn_1 \leftarrow$ local exact (k/2)NN of S_1, $r_1 \leftarrow max_{j\in[1,k/2]} dis\left(l_{nn_1[j]}, l_q\right)$
  - $nn_2 \leftarrow$ local exact (k/2)NN of S_2, $r_2 \leftarrow max_{j\in[1,k/2]} dis\left(l_{nn_2[j]}, l_q\right)$
  - $nn_3 \leftarrow$ local exact (k/2)NN of S_3, $r_3 \leftarrow max_{j\in[1,k/2]} dis\left(l_{nn_3[j]}, l_q\right)$

2st Round

$\boldsymbol{k}$ **= 8**

$\boldsymbol{k/2}$ **= 4**



	r_i
S_1	$\sqrt{2}$
S_2	$\sqrt{5}$
S_3	$\sqrt{5}$

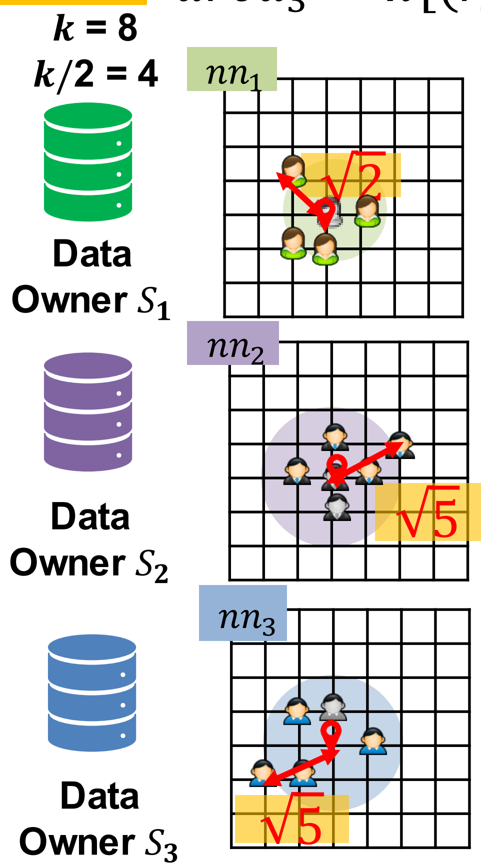
Toy Example

- Compute each data owner's density$({}^{k}/_{2})/area_i$
  - $area_1 \leftarrow \pi[(r_1)^2 - (r_1')^2]$, $density_1 \leftarrow ({}^{k}/_{2})/2\pi$

2st Round

$area_2 \leftarrow \pi[(r_2)^2 - (r_2')^2]$, $density_2 \leftarrow ({}^{k}/_{2})/4\pi$

$area_3 \leftarrow \pi[(r_3)^2 - (r_3')^2]$, $density_3 \leftarrow ({}^{k}/_{2})/4\pi$



	r_i	$area_i$	$density_i$
S_1	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$
S_2	$\sqrt{5}$	4π	$({}^{k}/_{2})/4\pi$
S_3	$\sqrt{5}$	4π	$({}^{k}/_{2})/4\pi$

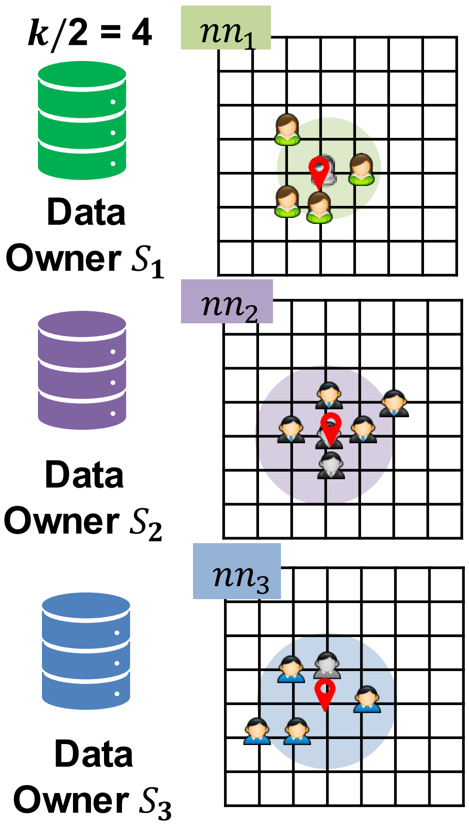
Toy Example

- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]
    - $sum = SecureSum\left(\frac{k/2}{2\pi}, \frac{k/2}{4\pi}, \frac{k/2}{4\pi}\right) = \frac{k/2}{\pi}$

2st Round

$\boldsymbol{k}$ **= 8**

$\boldsymbol{k/2}$ **= 4**



	r_i	$area_i$	$density_i$
S_1	$\sqrt{2}$	2π	$(k/2)/2\pi$
S_2	$\sqrt{5}$	4π	$(k/2)/4\pi$
S_3	$\sqrt{5}$	4π	$(k/2)/4\pi$

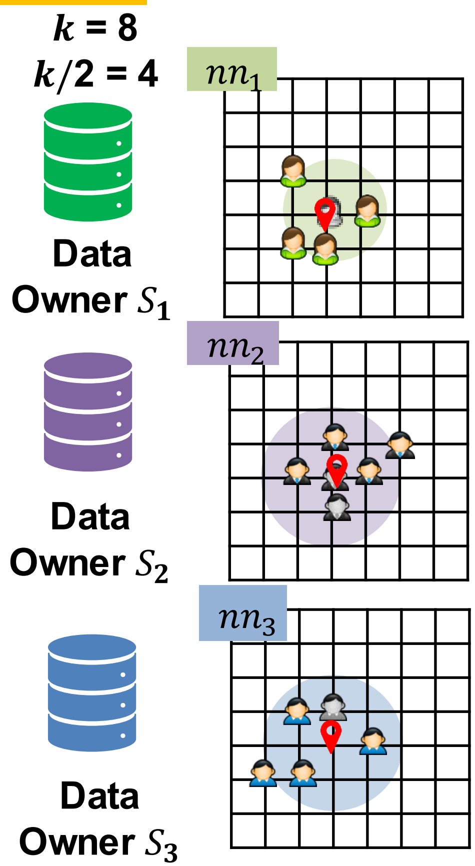
$$sum = SecureSum\left(\frac{k/2}{2\pi}, \frac{k/2}{4\pi}, \frac{k/2}{4\pi}\right) = \frac{k/2}{\pi}$$

Toy Example

- Compute contribution proportional to density
  - Compute sum of density by secure summation protocol in [9]

2st Round

Compute contribution $k_i = density_i \times \frac{k/2}{sum}$, $sum = \frac{k/2}{\pi}$



	r_i	$area_i$	$density_i$	k_i
S_1	$\sqrt{2}$	2π	$({}^{k}/_{2})/2\pi$	$k/4$
S_2	$\sqrt{5}$	4π	$({}^{k}/_{2})/4\pi$	$k/8$
S_3	$\sqrt{5}$	4π	$({}^{k}/_{2})/4\pi$	$k/8$

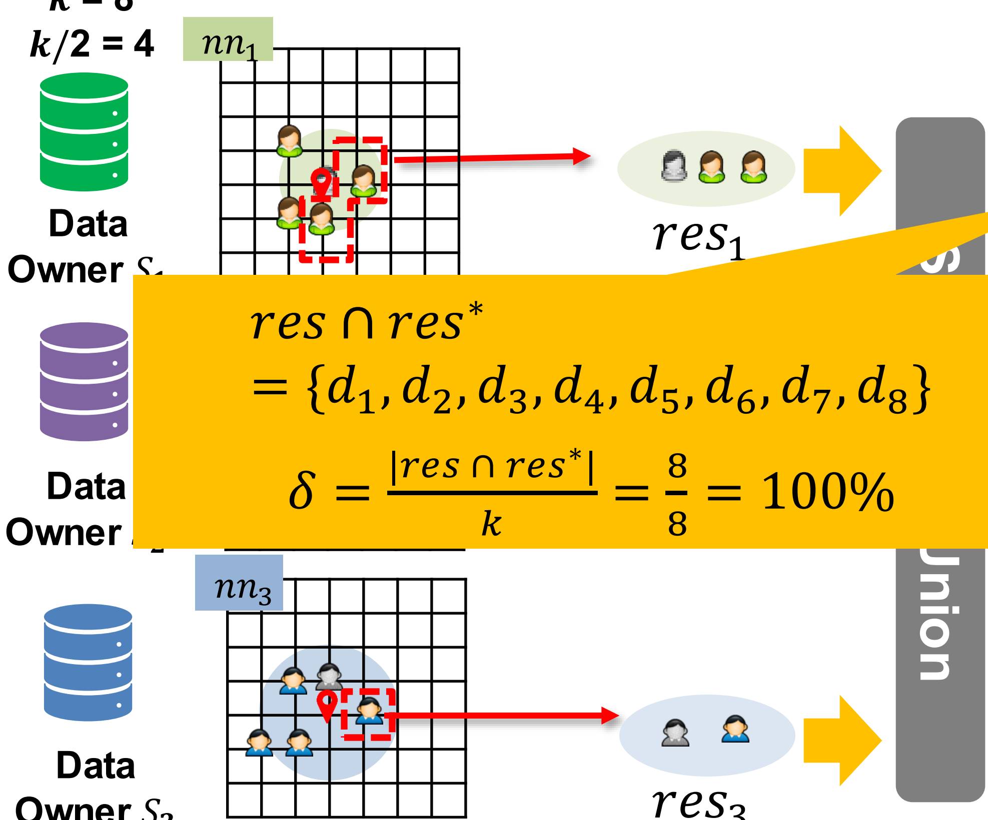
Toy Example

- Collect final answer by secure set union
  - Each data owner pick k_iNN again as the partial answer
  - Collect partial answers by secure set union protocol in [10]

2st Round

$k = 8$

$k/2 = 4$



$$res \cap res^* = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8\}$$

$$\delta = \frac{|res \cap res^*|}{k} = \frac{8}{8} = 100\%$$

Final Answer

	x	y
d_1	3	3
d_2	3	3
d_3	3	2
d_4	3	2
d_5	3	4
d_6	4	3
d_7	4	3
d_8	4	3

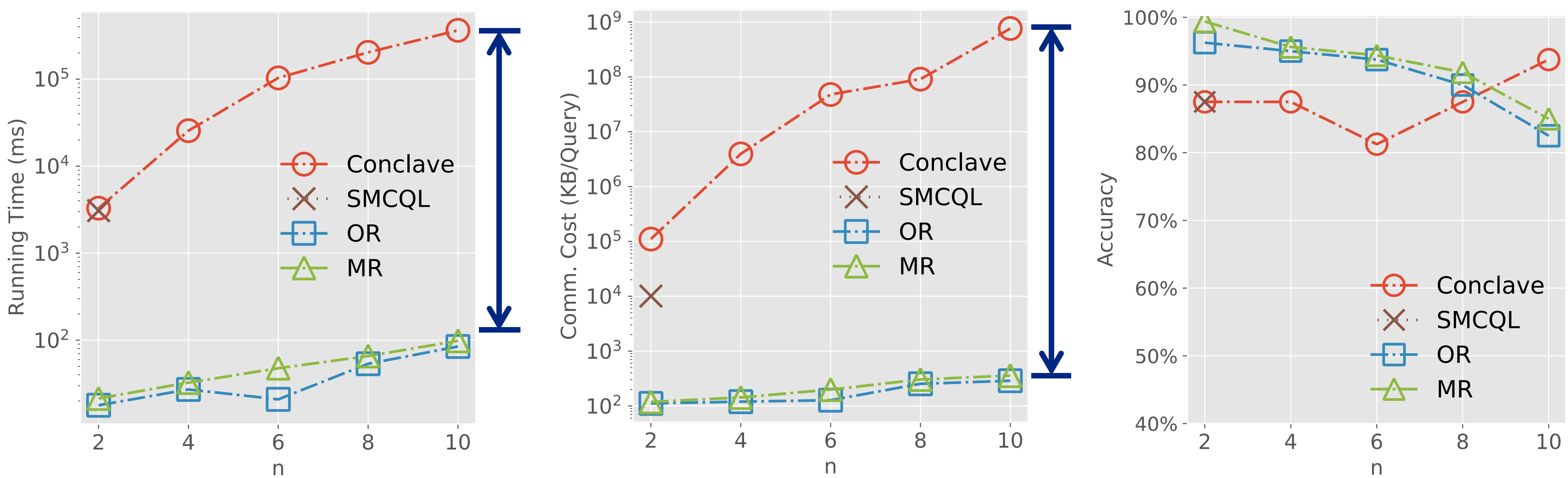
Exact kNN $res*$

	x	y
d_1	3	3
d_2	3	3
d_3	3	2
d_4	3	2
d_5	3	4
d_6	4	3
d_7	4	3
d_8	4	3

[10] Pawel Jurczyk et al. Information Sharing across Private Databases: Secure Union Revisited. SocialCom/PASSAT 2011

Experiment: Result

- Varying #(data owners) n on OSM dataset
  - Running time
    - Up to 3 orders of magnitude shorter
  - Communication cost
    - Up to 5 orders of magnitude higher
  - Accuracy
    - More robust



(a) Running Time (b) Communication Cost (c) Accuracy