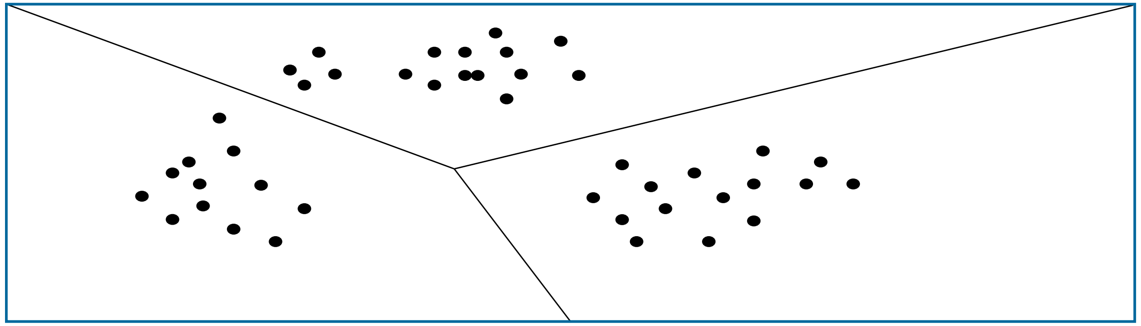
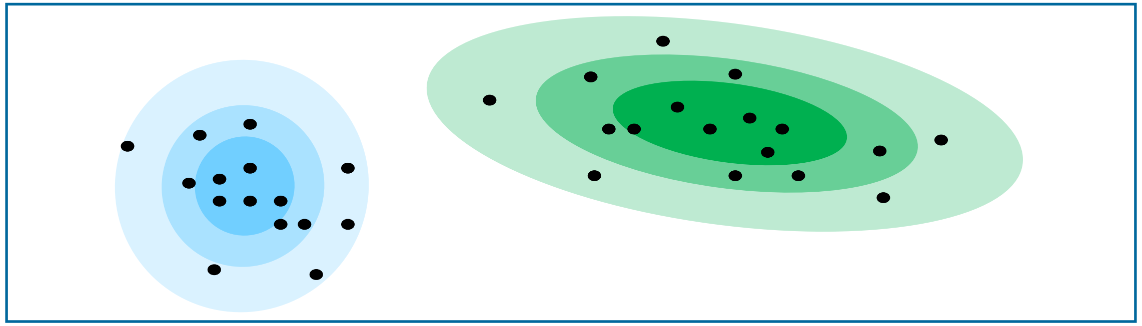
SoK: Efficient Privacy-preserving Clustering

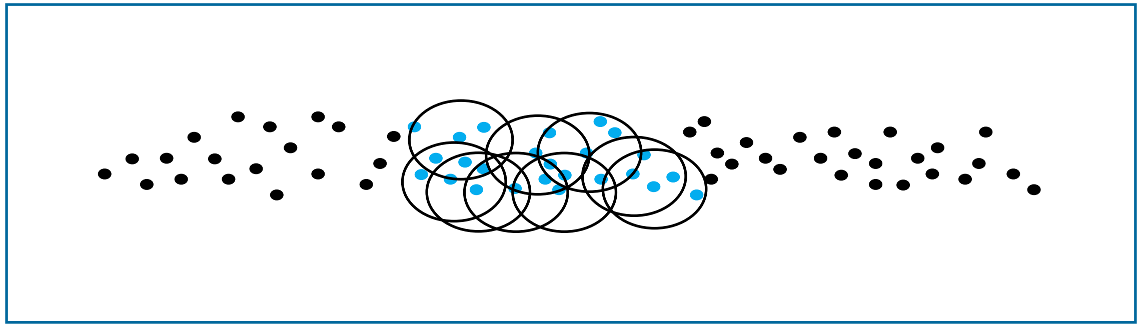
Aditya Hegde, Helen Möllering, Thomas Schneider, Hossein Yalame



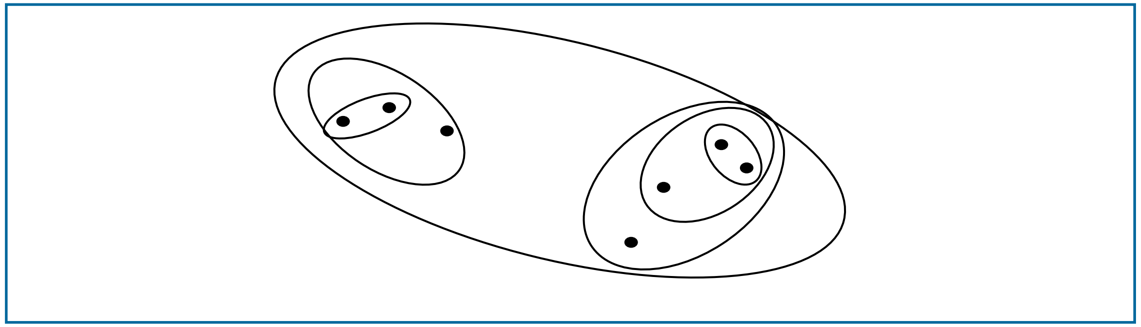
Partitioning-based Clustering



Distribution-based Clustering



Density-based Clustering



Hierarchical Clustering

Agenda

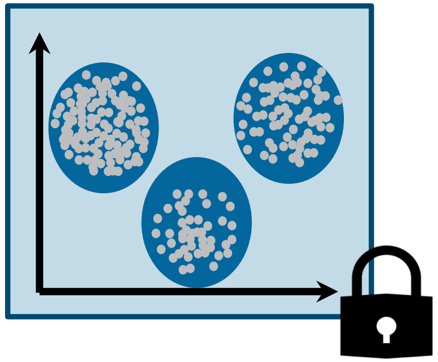
1. Motivation and Preliminaries
2. Survey of Private Clustering
3. Evaluation of State-of-the-Art Protocols
4. Challenges to Real-life Application

Clustering is applied on highly sensitive information

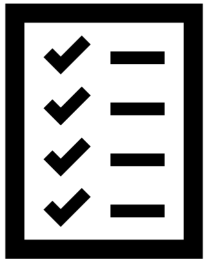
ENCRYPTO
CRYPTOGRAPHY AND PRIVACY ENGINEERING

Our Contributions

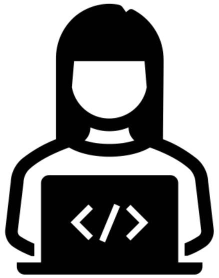
- **First comprehensive review and analysis of private clustering protocols**



- **Guideline on how to choose an appropriate private clustering protocol for concrete applications**



- **Open-source implementation and benchmark of four most efficient, fully private clustering schemes: [CKP19], [MPO+19], [MRT20], [BCE+21]**



59 works were analyzed

ENCRYPTO
CRYPTOGRAPHY AND PRIVACY ENGINEERING

Algorithm	Scheme	Privacy	Security	PETs	L1	L2	L3	L4	O1	O2	O3	Interactivity (Scenario)	Data	Other issues
K-means	[82, KDD'03]	✗	◐	HE+blinding	(✗)[1]	✗	✗	✗	✗	✓	✗	all data owners (≥ 3)	v	
	[83, KDD'05]	✗	◐	HE+ASS+GC	✓	✓	✗	✗	✓	✓	✗	2PC	a	wrong division
	[84, ESORICS'05]	✗	◐	HE *or* OPE	✗	✓	✓	✗	✗	✓	✗	2PC	h	
	[12, CCS'07]	✓	◐	HE+ASS	✓	✓	✓	✗	✗	✓	✗	2PC	a	
	[85, SECRYPT'07]	✗	◐	blinding	✗	✓	✗	✗	✓	✓	✗	all data owners	v/h	
	[86, AINAW'07]	✗	◐	HE+ASS+OPE	✓	✗	✗	✗	✓	✓	✗	2PC	h	
	[87, PAIS'08]	✗	◐	ASS	✓	✓	✗	✗	✓	✓	✗	all data owners (≥ 4)	v	
	[88, WIFS'09]	✗	◐	HE	✗	✓	✗	✓	✓	✗	✗	data owners + 1 server	h	
	[89, KAIS'10]	✗	◐	HE+ASS	✓	✓	✗	✗	✓	✗	✗	all data owners	h	
	[90, PAISI'10]	✗	◐	SS	✓	✗	✗	✗	✓	✓	✗	Outsourcing ≥ 3 servers	a	
	[91, ISPA'10]	✗	◐	HE	✓	✓	✗	✗	✗	✓	✗	all data owners	v/h	
	[92, WIFS'11]	✗	◐	HE+GC	✓	✗	✓	✓	✓	✗	✗	Outsourcing, 3 servers	h	
	[93, ISI'11]	✗	◐	HE+ASS	(✗)[1]	✗	✗	✗	✓	✗	✗	2PC	v	
	[94, TM'12]	✗	◐	SSS	✗	✗	✓	✗	✗	✓	✗	all data owners	h	distance calculation unclear
	[95, JIS'13]	✗	◐	HE	✗	✓	✓	✗	✓	✗	✗	data owners + 2 servers	h	
	[96, ICDCIT'13]	✗	●	SSS+ZKP	✗	✗	✓	✗	✗	✓	✗	all data owners	h	
	[97, ASIACCS'14]	✗	◐	HE	✗	✗	✗	✗	✓	✓	✗	outsourcing, 1 data owner + 1 server	–	insecure HE [107]
	[98, MSN'15]	✗	◐	HE	✗	✗	✗	✓	✗	✗	✗	outsourcing, data owners + 1 server	h	insecure HE [107]
	[99, IJNS'15]	✗	◐	HE	✗	✗	✗	✗	✗	✓	✗	all data owners	h	
	[13, CIC'15]	✓	◐	HE	✓	✓	✓	✗	✗	✓	✗	Outsourcing, 2 servers	h	
	[100, ICACCI'16]	✗	N/A	SS	✗	✗	✗	✗	✓	✗	✗	arbitrary number of servers	a	
	[101, ISPA'16]	✗	◐	blinding	✗	✗	✗	✓	✗	✓	✗	all data owners (≥ 3)	h	
	[102, SecComm'17]	✗	◐	HE	✓	✗	✗	✓	✗	✓	✗	outsourcing, ≥ 4 servers	h	
	[103, TII'17]	✗	◐	HE	✗	✗	✗	✗	✗	✗	✗	data owners + 1 server	h	
	[14, SAC'18]	✓	◐	HE	✓	✓	✓	✓	✗	✓	✗	Outsourcing, 1 server	–	
	[15, CLOUD'18]	✓	◐	HE	✓	✓	✓	✗	✗	✓	✗	Outsourcing, 2 servers	–	distance calculation unclear
	[108, CCPE'19]	✗	N/A	HE	✗	✗	✗	✗	✗	✓	✗	Outsourcing, 2 data owners + 1 server	h	insecure HE [107]
	[104, TCC'19]	✗	◐	HE	✓	✗	✗	✓	✓	✗	✗	Outsourcing, 1 data owner + ≥ 1server(s)	–	
	[105, Inf. Sci.'20]	✗	◐ (●)[2]	HE+GC	✗	✗	✗	✗	✗	✓	✗	Outsourcing, 2 data owners + 1 server	h	
	[106, SCN'20]	✗	◐	HE+SKC	✓	✗	✗	✓	✗	✓	✗	Outsourcing, 3 servers	h	
	[11, PETS'20]	✓	◐	GC	✓	✓	✓	✗	✗	✓	✗	2PC/Outsourcing	h	
	[8, TKDE'20]	✗	◐	HE	✓	✗[3]	✓	✗	✗	✓	✗	Outsourcing, 2 servers	a	
Kernel K-means	[58, KAIS'16]	✗	N/A	PKC	✓	✗	✗	✗	✓	✗	✗	Outsourcing, 1 server	–	security model
Possibilistic C-means	[43, TBD'17]	✗	N/A	HE	✗	✗	✗	✗	✓	✓	✗	Outsourcing, 1 data owner + 1 server	–	
K-medoids	[57, SMC'07]	✗	N/A	HE+blinding	✓	✗	✗	✓	✗	✗	✗	all data owners	v	exhaustive search
	[71, CCSEIT'12]	✗	N/A	HE+blinding	✓	✗	✗	✓	✗	✗	✗	all data owners	v	exhaustive search
GMM	[45, KAIS'05]	✗	◐	blinding	✓	✓	✗	✗	✓	✗	✗	all data owners	h	
	[44, DCAI'19]	✗	◐	ASS	✓	✓	✗	✗	✓	✗	✗	all data owners (> 2)	v/h	
Affinity Propagation	[81, INCoS'12]	✗	◐	HE + blinding	✓	✓	✗	✓	✓	✗	✗	all data owners	v	
	[16, SECRYPT'21]	✓	◐/●	ASS+GC	✓	✓	✓	✓	✓	✗	✗	all data owners/Outsourcing	a	
Mean-shift	**[9, SAC'19]**	✓	◐	HE	✓	✓	✓	✓	✓	✗	✗	Outsourcing, 1 server	–	
DBSCAN	[72, ISI'06]	✗	◐	blinding	✓	✓	✗	✓	✗	✗	✗	all data owners	v	lack of complete protocol
	[73, ADMA'07]	✗	◐	HE+blinding	✓	✗	✗	✓	✓	✗	✗	2PC	v/h	
	[74, IJSIA'07]	✗	◐	PKC+blinding	✓	✓	✗	✓	✓	✗	✗	all data owners	v	
	[75, ITME'08]	✗	◐	HE+blinding	✓	✗	✗	✓	✓	✗	✗	data owners + 1 server	h	
	[22, TDP'13]	✗	◐	HE+blinding	✓	✗	✗	✓	✓	✗	✗	2PC	a	
	[17, S&P'12]	✓	◐/●[5]	GC	✓	✓	✓	✓	✓	✓	✗	2PC	h	
	[46, SIBCON'17]	✗	◐	HE+PKC	✓	✓	✗	✓	✓	✗	✗	all data owners	v	cluster expansion missing
	[47, PRDC'17]	✗	◐	HE	✓	✗	✗	✓	✗	✗	✗	outsourcing, all data owners + 1 server	h	
	[76, AI'18]	✗	◐	HE	✓	✗	✗	✓	✓	✗	✗	data owners + 1 server	a	uses absolute distance
	[18, ASIACCS'21]	✓	◐	ASS+GC	✓	✓	✓	✓	✓	(✓)[4]	✗	2PC/Outsourcing	a	
HC	[77, SDM'06]	✗	◐	HE+ASS+GC	✓	✓	✗	✓	✗	✓	✗	2PC	h	
	[50, TKDE'07]	✗	◐	blinding *or* SKC	✓	✓	✗	✓	✓	✗	✗	data owners + 1 server	h	SKC not semantically secure
	[49, TDP'10]	✗	◐	HE+GC	✓	✓	✗	✓	✓	✓	✗	2PC	h	
	[48, ISI'14]	✗	N/A	HE	✓	✗	✗	✓	✓	✓	✗	2PC	v	
	[78, ISCC'17]	✗	◐	HE	✓	✓	✗	✓	✗	✗	✓	2PC	v/h	
	[19, ArXiv'19]	✓	◐	HE & GC	✓	✓	✓	✓	✗	✓	✓	2PC	h	
BIRCH	[79, SDM'06]	✗	◐	HE+ASS	✓	✓	✗	✓	✗	✗	✗	2PC	v	
	[80, ADMA'07]	✗	◐	HE+ASS	✓	✓	✗	✓	✗	✗	✗	2PC	a	

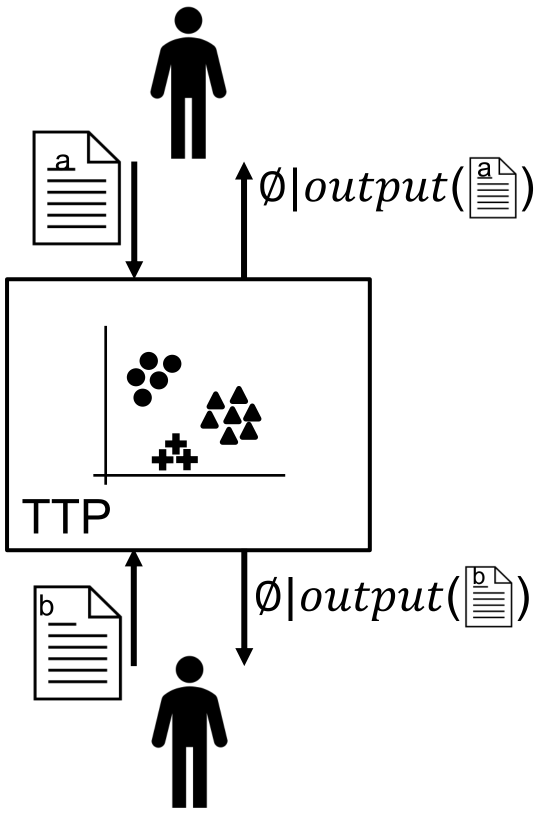
1 Of the parameters hold by the respective data owner.
2 Assuming max. 1 party deviates from the protocol.
3 Leaks partial information about cluster sizes.
4 Not implemented, but possible.
5 Can be used with any security model of GCs.

TECHNISCHE UNIVERSITÄT DARMSTADT

Fully private clustering does not leak anything beyond the output

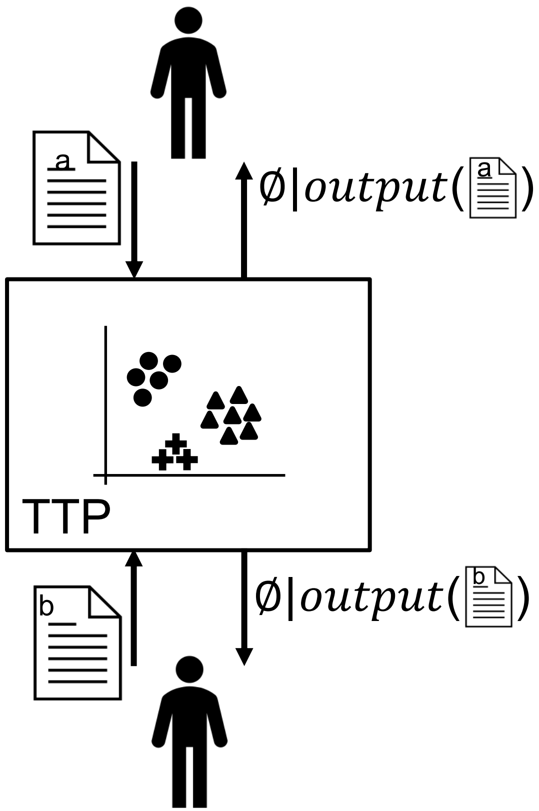
ENCRYPTO
CRYPTOGRAPHY AND PRIVACY ENGINEERING

Ideal Functionality

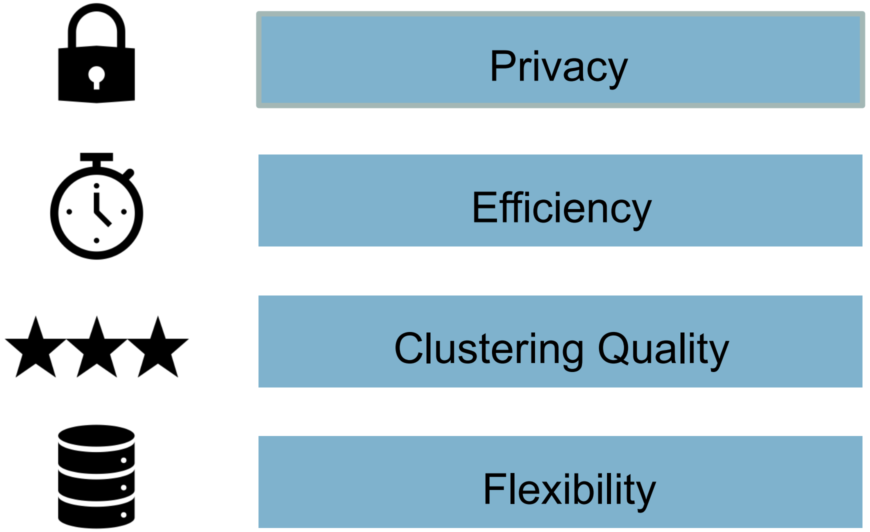


Fully private clustering does not leak anything beyond the output

Ideal Functionality



Requirements



Agenda

1. Motivation and Preliminaries
2. **Survey of Private Clustering**
3. Evaluation of State-of-the-Art Protocols
4. Challenges to Real-life Application

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious
Scenarios	2PC/MPC, Outsourcing

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious
Scenarios	2PC/MPC, Outsourcing
Data Partition	horizontal (**h**), vertical (**v**), arbitrary (**a**)

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious
Scenarios	2PC/MPC, Outsourcing
Data Partition	horizontal (**h**), vertical (**v**), arbitrary (**a**)
PETs	Homomorphic Encryption (**HE**, [GB09]), Public Key Cryptography, Garbled Circuits (**GC**, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious
Scenarios	2PC/MPC, Outsourcing
Data Partition	horizontal (**h**), vertical (**v**), arbitrary (**a**)
PETs	Homomorphic Encryption (**HE**, [GB09]), Public Key Cryptography, Garbled Circuits (**GC**, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])
Privacy	Fully privacy-preserving, Leakage

Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm	K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Security Model	Semi-honest, Malicious
Scenarios	2PC/MPC, Outsourcing
Data Partition	horizontal (**h**), vertical (**v**), arbitrary (**a**)
PETs	Homomorphic Encryption (**HE**, [GB09]), Public Key Cryptography, Garbled Circuits (**GC**, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])
Privacy	Fully privacy-preserving, Leakage
Efficiency	Computation, Communication, Memory

There are only 10 fully private clustering schemes

Algorithm	Paper	PETs: HE	PETs: GC	PETs: MIX	Scenario: MPC	Scenario: Out	Data: h	Data: a	Output	Efficiency
K-means	[BO07]			✓	✓			✓	final centroids	✗
	[RSB+16]	✓				✓	✓		final centroids	✗
	[JA18]	✓				✓			final centroids	✗
	[KC18]	✓				✓			cluster sizes	✗
	[MRT20]		✓		✓	✓	✓		final centroids	✓
Mean-shift	[CKP19]	✓				✓			final centroids	✓
Affinity Prop.	[KMS+21]			✓	✓	✓		✓	final clusters	✗
DBSCAN	[ZE13]	✓			✓		✓		Cluster labels	✗
	[BCE+21]			✓	✓	✓		✓	Cluster labels	✓
HC	[MPO+19]			✓	✓		✓		Final dendrogram	✓

There are only 10 fully private clustering schemes

Algorithm	Paper	PETs			Scenario		Data		Output	Efficiency
		HE	GC	MIX	MPC	Out	h	a		
K-means	[BO07]			✓	✓			✓	final centroids	✗
	[RSB+16]	✓				✓	✓		final centroids	✗
	[JA18]	✓				✓			final centroids	✗
	[KC18]	✓				✓			cluster sizes	✗
	[MRT20]		✓		✓	✓	✓		final centroids	✓
Mean-shift	[CKP19]	✓				✓			final centroids	✓
Affinity Prop.	[KMS+21]			✓	✓	✓		✓	final clusters	✗
DBSCAN	[ZE13]	✓			✓		✓		Cluster labels	✗
	[BCE+21]			✓	✓	✓		✓	Cluster labels	✓
HC	[MPO+19]			✓	✓		✓		Final dendrogram	✓

Agenda

Performance is the decisive metric

HE-Meanshift
[CKP19]

PCA/OPT
[MPO+19]

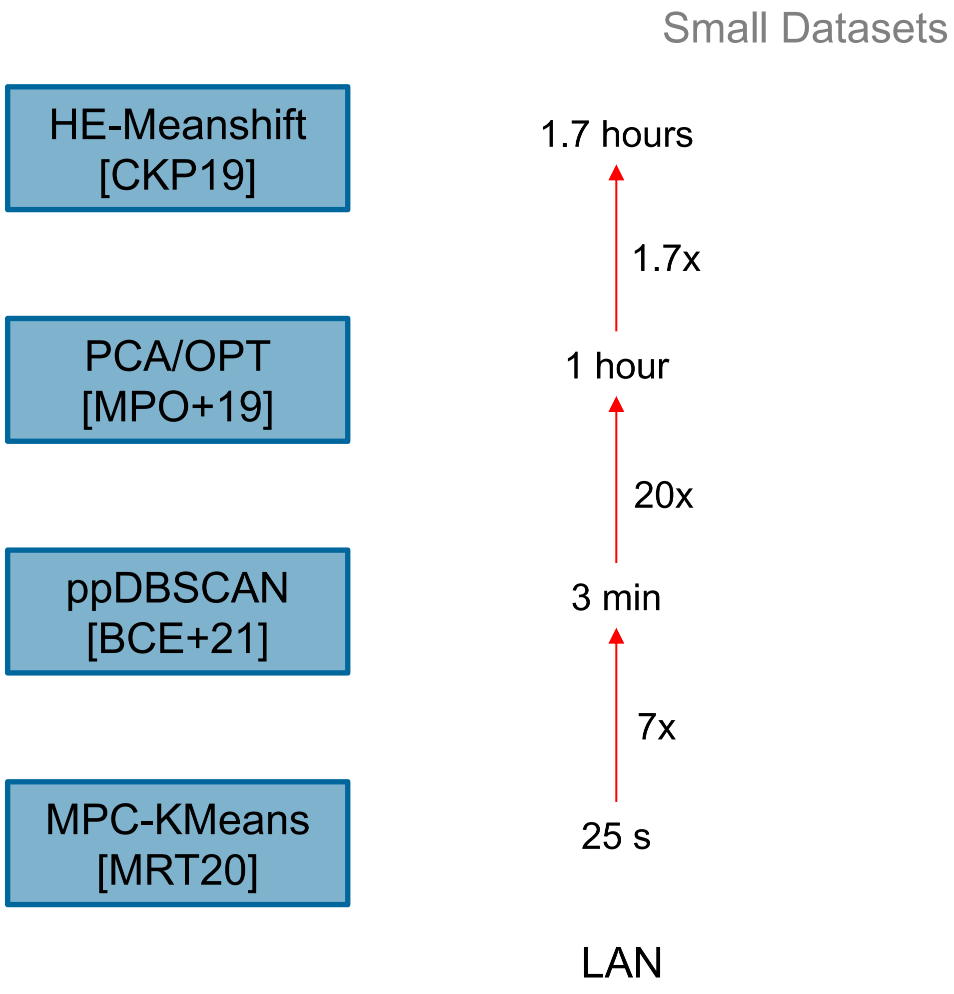
ppDBSCAN
[BCE+21]

MPC-KMeans
[MRT20]

Small Datasets:

- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

Performance is the decisive metric



Small Datasets:

- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

Performance is the decisive metric

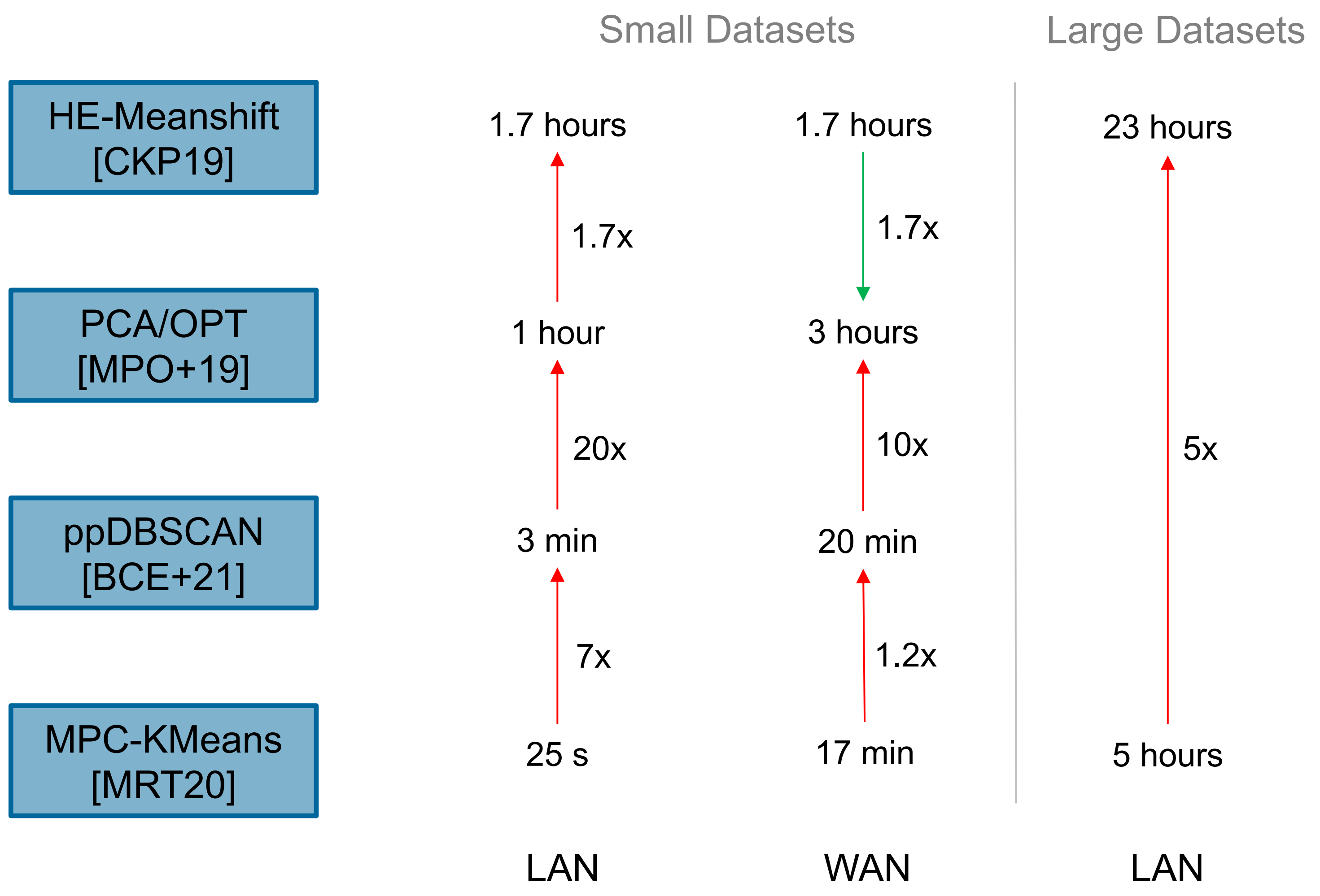
Small Datasets

	LAN	WAN
HE-Meanshift [CKP19]	1.7 hours	1.7 hours
	1.7x	1.7x
PCA/OPT [MPO+19]	1 hour	3 hours
	20x	10x
ppDBSCAN [BCE+21]	3 min	20 min
	7x	1.2x
MPC-KMeans [MRT20]	25 s	17 min

Small Datasets:

- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

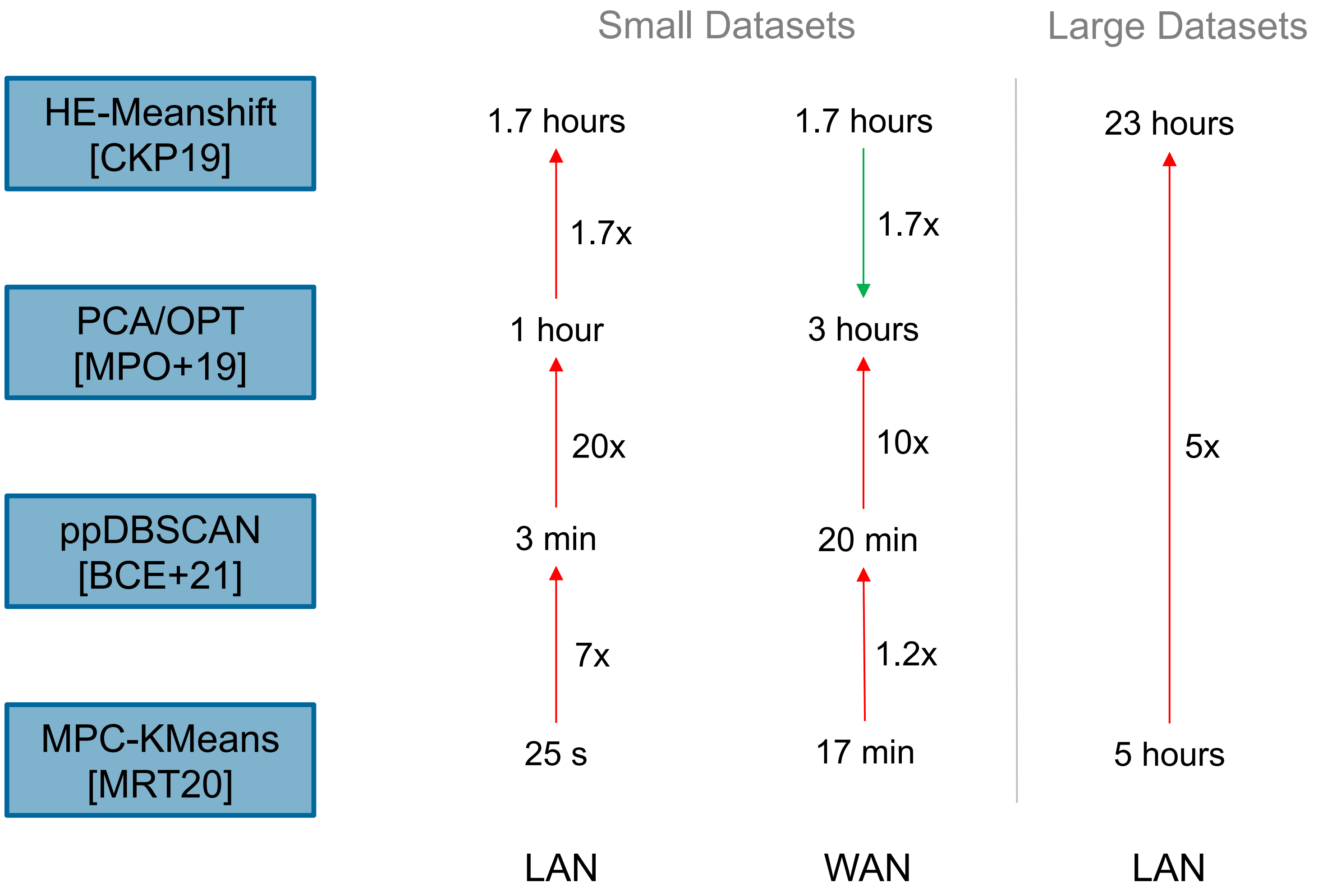
Performance is the decisive metric



Small Datasets:
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

Large Datasets:
- Number of points: $2^{13} \leq N \leq 2^{16}$
- Dimension: $1 \leq d \leq 16$
- Number of clusters: $2 \leq K \leq 20$

Performance is the decisive metric



Small Datasets:

- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

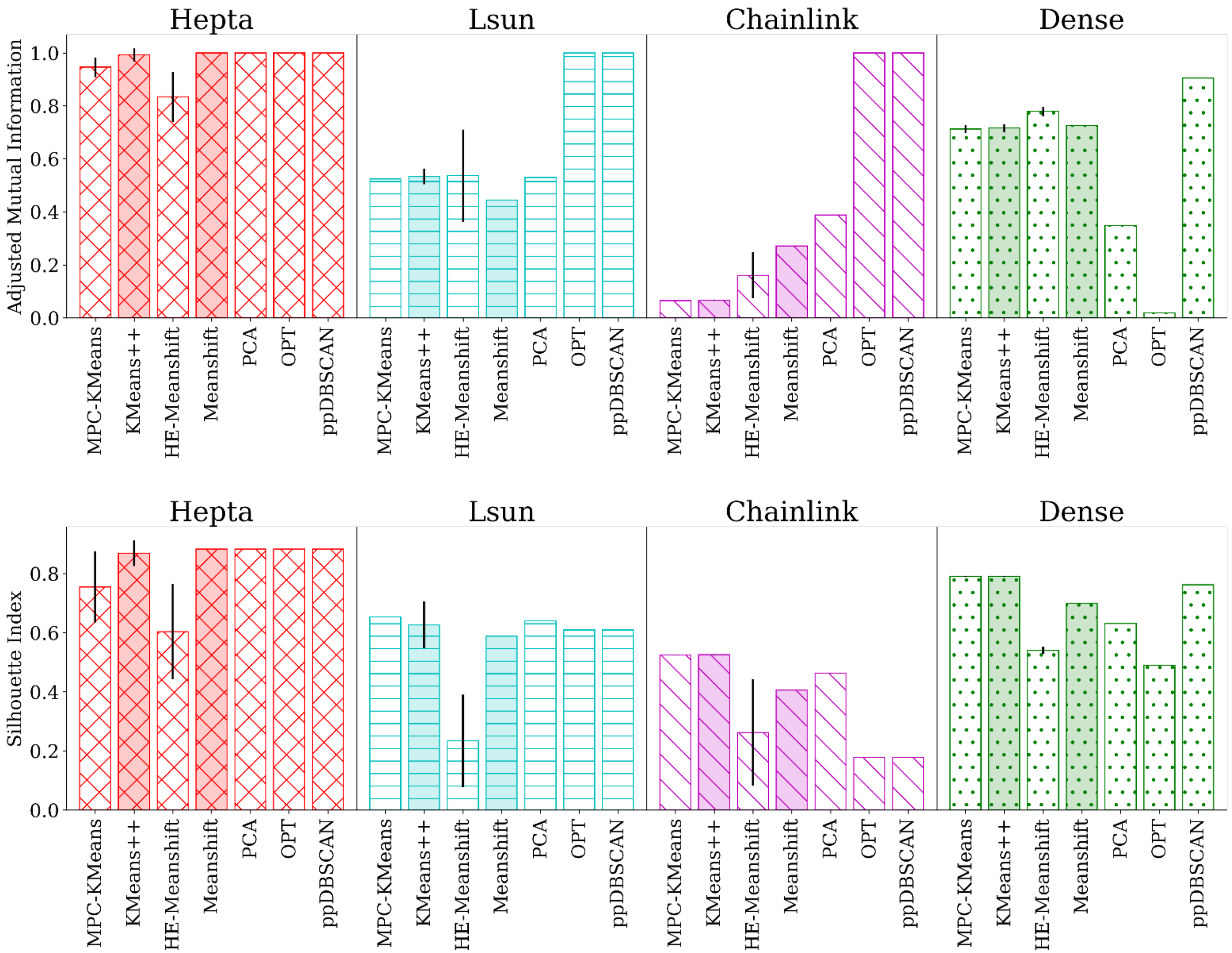
Large Datasets:

- Number of points: $2^{13} \leq N \leq 2^{16}$
- Dimension: $1 \leq d \leq 16$
- Number of clusters: $2 \leq K \leq 20$

Performance strongly affects choice of protocol.

Several factors affect clustering quality

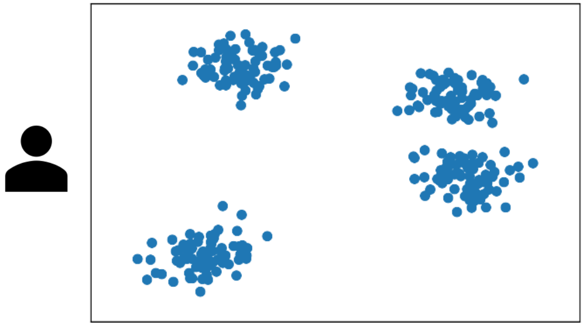
- Protocol/Algorithm
- Parameters
- Randomness



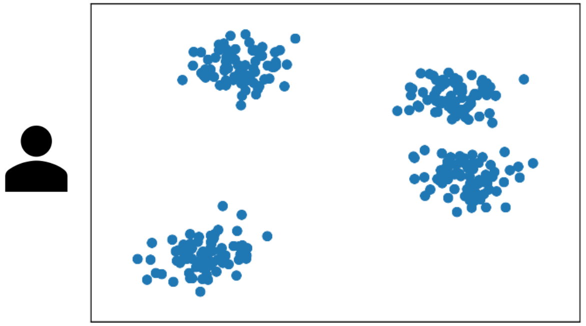
Agenda

1. Motivation and Preliminaries
2. Survey of Private Clustering
3. Evaluation of State-of-the-Art Protocols
4. **Challenges to Real-life Application**

Plaintext clustering eases parameter selection

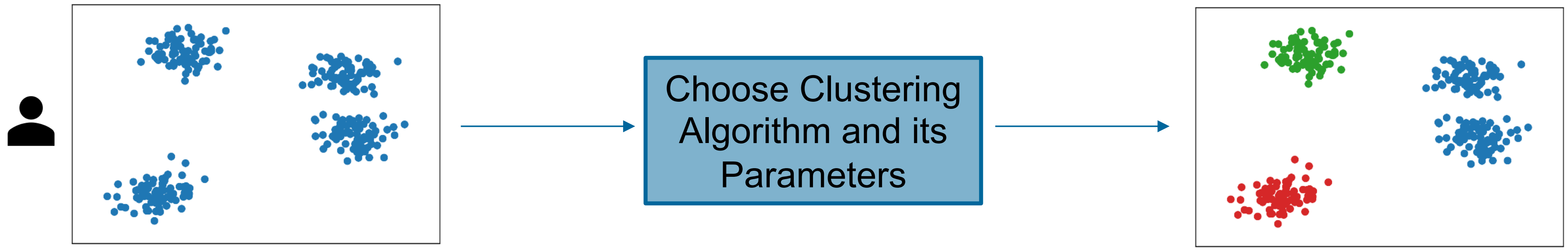


Plaintext clustering eases parameter selection

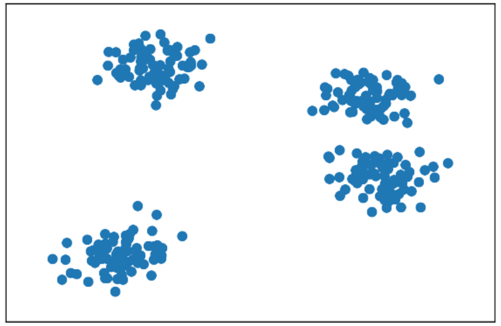


Choose Clustering Algorithm and its Parameters

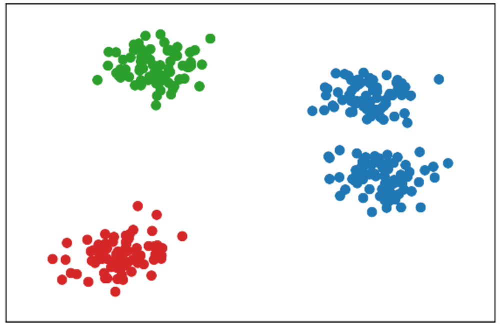
Plaintext clustering eases parameter selection



Plaintext clustering eases parameter selection

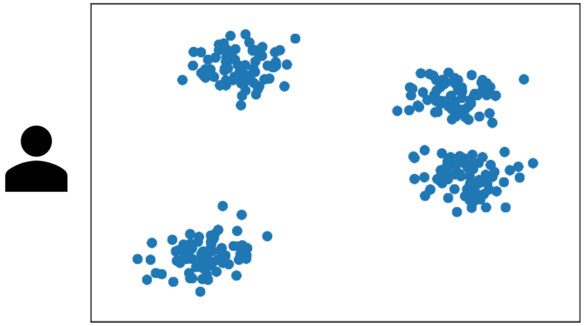


Choose Clustering Algorithm and its Parameters



<u>Evaluate</u>
- Visual
- Clustering indices

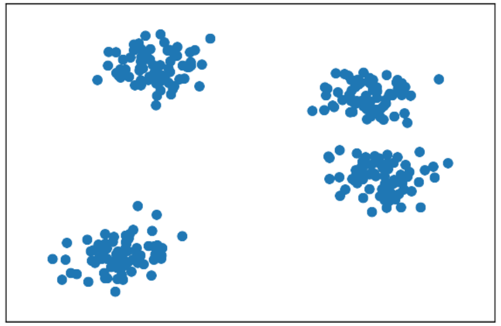
Plaintext clustering eases parameter selection



Choose Clustering Algorithm and its Parameters

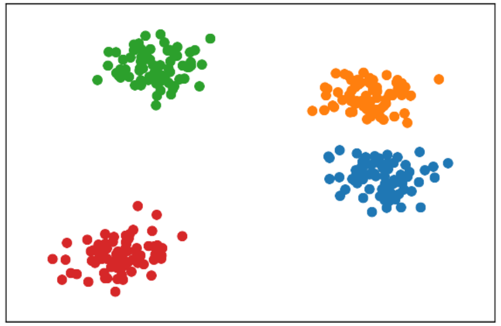
Different Random Seed

Plaintext clustering eases parameter selection



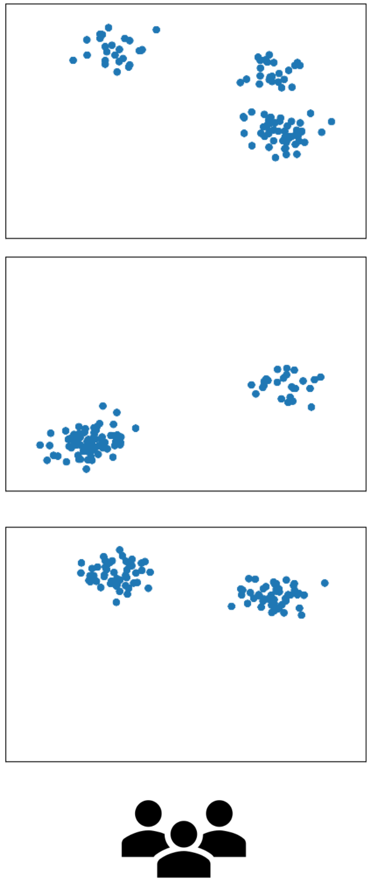
Choose Clustering Algorithm and its Parameters

Different Random Seed

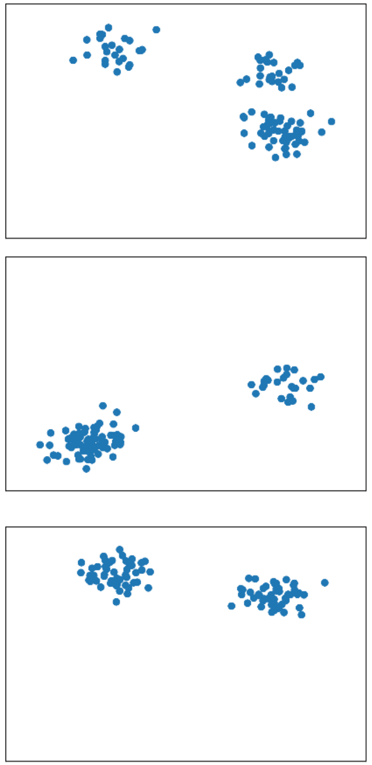


Evaluate
- Visual
- Clustering indices

Distributed data and protocol efficiency are the main challenges



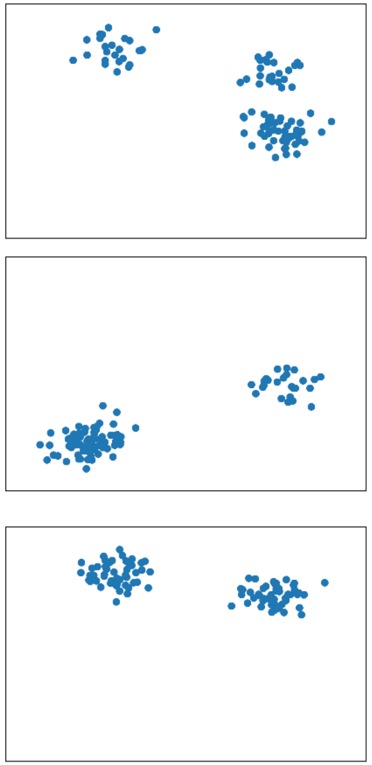
Distributed data and protocol efficiency are the main challenges



<u>Choose Clustering Protocol and Parameters</u>

- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice

Distributed data and protocol efficiency are the main challenges

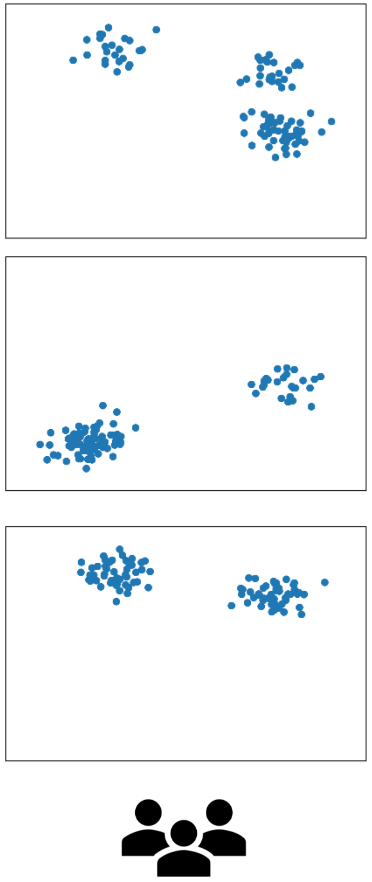


Choose Clustering Protocol and Parameters

- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice

Overhead of secure computation

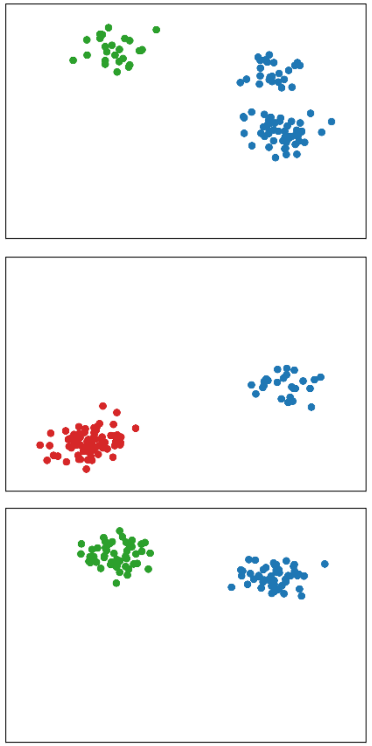
Distributed data and protocol efficiency are the main challenges



Choose Clustering Protocol and Parameters

- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice

Overhead of secure computation



Evaluate

- Securely computing clustering indices
- Handling outliers and noise

Future research directions for private clustering

- Efficiency: runtime, communication, and memory
- Parameters that can be set independent of input data
- Protocols that handle outliers and noise
- Techniques to securely evaluate clustering output

THANKS FOR YOUR ATTENTION!

Contact: https://encrypto.de/moellering
Code: https://encrypto.de/code/SoK_ppClustering

References (1)

[BCE+21] B. Bozdemir, S. Canard, O. Ermis, H. Möllering, M. Önen, T. Schneider, “Privacy-preserving density-based clustering,” in ASIACCS, 2021.

[BO07] P. Bunn and R. Ostrovsky, “Secure two-party K-means clustering,” in CCS, 2007.

[CKP19] J. H. Cheon, D. Kim, and J. H. Park, “Towards a practical cluster analysis over encrypted data,” in SAC, 2019.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise” in International Conference on Knowledge Discovery and Data Mining, 1996.

[ELLS11] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, “Cluster analysis” in Wiley, 2011.

[FD07] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” Science, 2007.

[FH75] K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition” in TIT, 1975.

[GB09] C. Gentry and D. Boneh, A fully homomorphic encryption scheme. Stanford university Stanford, 2009.

[GMW87] O. Goldreich, S. Micali, and A. Wigderson, “How to play any mental game,” in STOC, 1987.

[JA18] A. Jäschke and F. Armknecht, “Unsupervised Machine Learning on Encrypted Data,” in SAC, 2018.

References (2)

[JA18] A. Jäschke and F. Armknecht, “Unsupervised Machine Learning on Encrypted Data,” in SAC, 2018.

[KC18] H. Kim and J. Chang, “A privacy-preserving k-means clustering algorithm using secure comparison protocol and density-based center point selection,” in International Conference on Cloud Computing, 2018.

[KMS+21] H. Keller, H. Möllering, T. Schneider, and H. Yalame, “Balancing quality and efficiency in private clustering with affinity propagation,” in SECRYPT, 2021.

[MPO+19] X. Meng, D. Papadopoulos, A. Oprea, and N. Triandopoulos, “Private two-party cluster analysis made formal & scalable,” arXiv:1904.04475v2, 2019.

[MRT20] P. Mohassel, M. Rosulek, and N. Trieu, “Practical privacy preserving K-means clustering,” in PETS, 2020.

[RSB+15] F.-Y. Rao, B. K. Samanthula, E. Bertino, X. Yi, D. Liu, “Privacy-preserving and outsourced multi-user K means clustering,” in CIC, 2015.

[Steinhaus56] H. Steinhaus, “Sur la division des corp materiels en parties” in Bulletin L’Académie Polonaise des Science, 1956.

[XW05] R. Xu and D. Wunsch, “Survey of clustering algorithms” in TNN, 2005.

[Yao86| A. C.-C. Yao, “How to generate and exchange secrets,” in FOCS, 1986.

[ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: An efficient data clustering method for very large databases” ACM SIGMOD, 1996.

[ZE13] S. Zahur and D. Evans, “Circuit structures for improving efficiency of security and privacy tools,” in IEEE S&P, 2013.