Differentially Private k-Means with Constant Multiplicative Error

Uri Stemmer

Ben-Gurion University

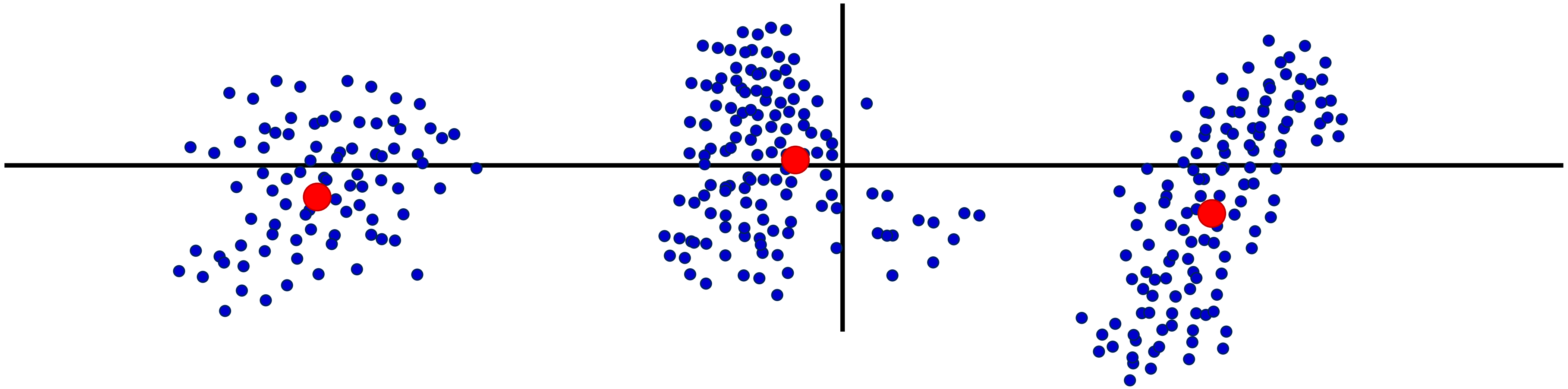
joint work with

Haim Kaplan

What is k-Means Clustering?

<u>Given</u>: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{d}\right)^{n}$ and parameter $\boldsymbol{k}$

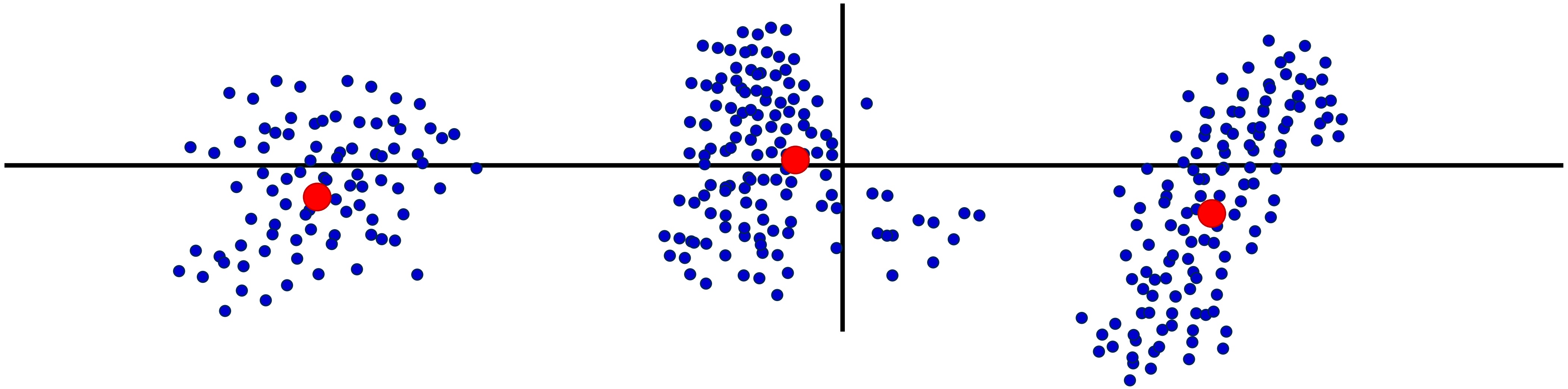
Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_i \mathbf{min}_\ell \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$



What is k-Means Clustering?

Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}) \in \left(\mathbb{R}^d\right)^n$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \ldots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_i \mathbf{min}_\ell \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

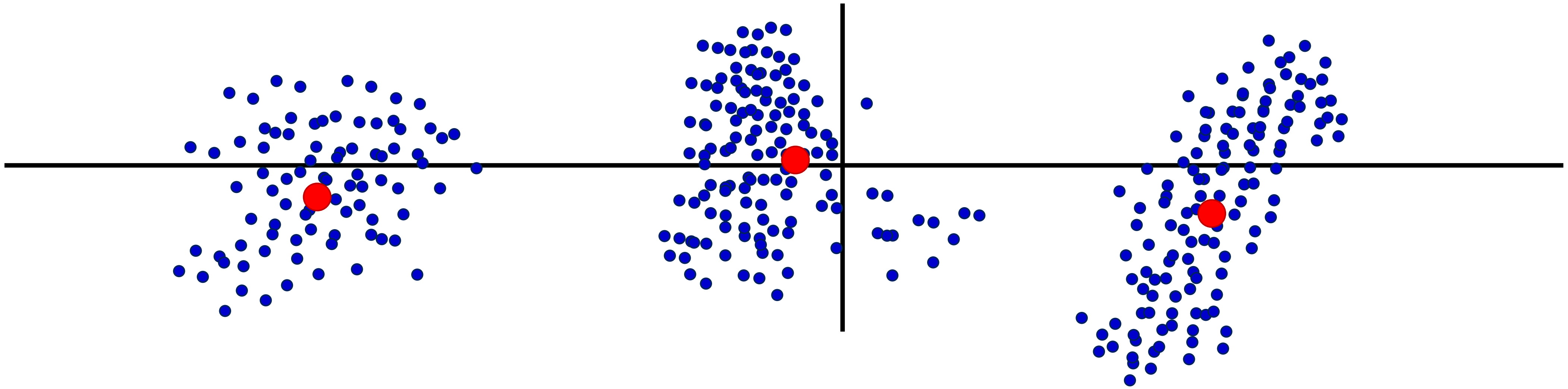


✓ Probably the most well-studied clustering problem
✓ Tons of applications
✓ Super popular

What is k-Means Clustering?

Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_{\boldsymbol{i}} \mathbf{min}_{\ell} \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$



What is Differentially Private k-Means?

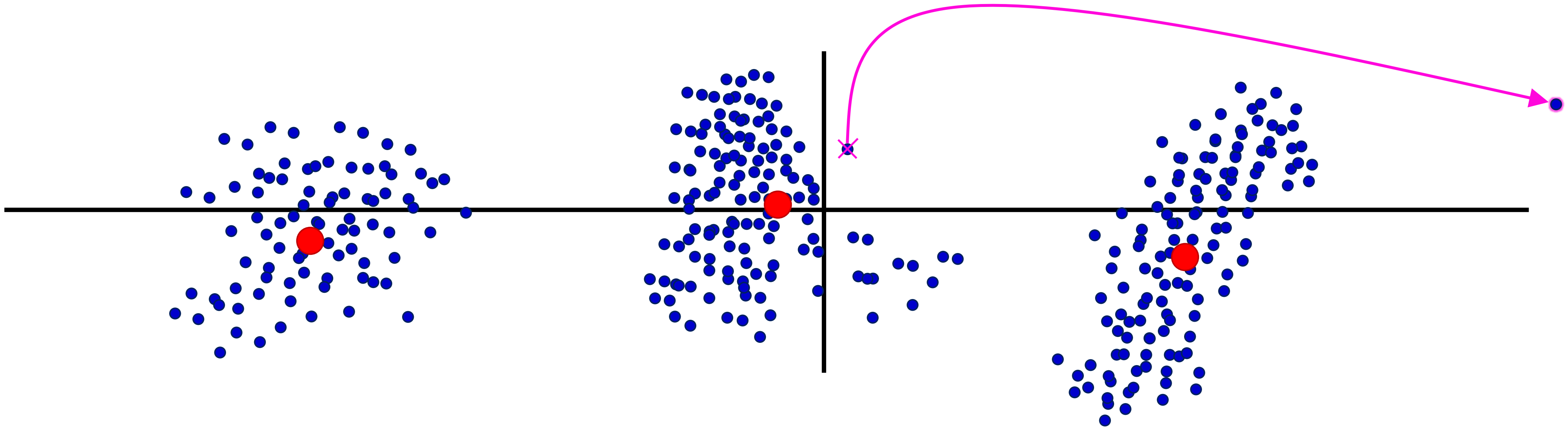
[Dwork, McSherry, Nissim, Smith 06] (informal)

- ✓ Every data point $\boldsymbol{x_i}$ represents the (private) information of one individual
- ✓ **Goal:** the output (the set of centers) does not reveal information that is specific to any single individual

What is k-Means Clustering?

Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in (\mathbb{R}^d)^n$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_i \min_\ell \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$



What is Differentially Private k-Means?

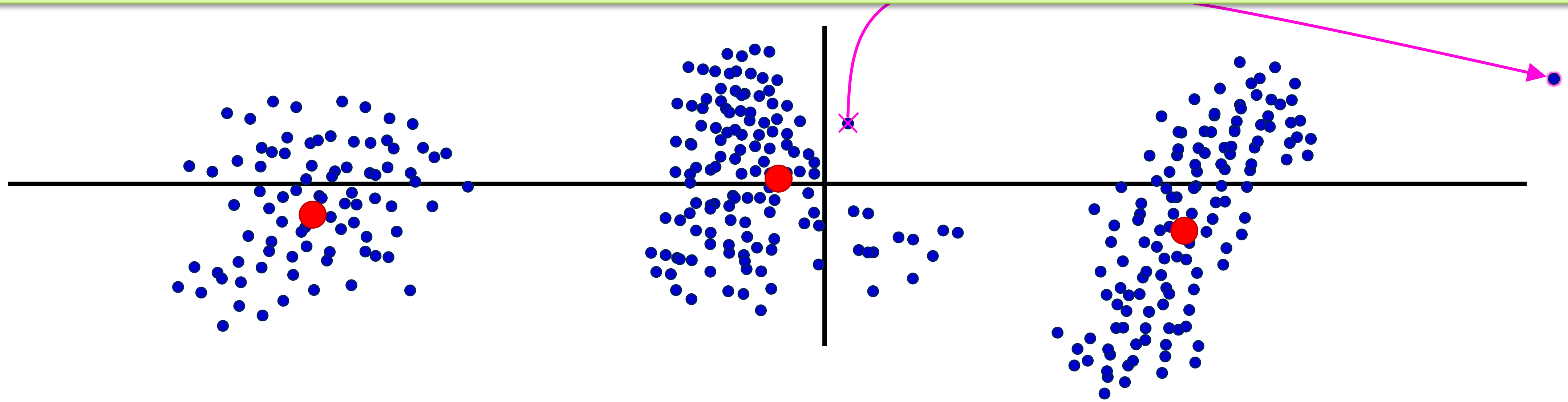
[Dwork, McSherry, Nissim, Smith 06] (informal)

✓ Every data point $\boldsymbol{x_i}$ represents the (private) information of one individual

✓ **Goal:** the output (the set of centers) does not reveal information that is specific to any single individual

✓ **Requirement:** the output distribution is insensitive to any arbitrarily change of a single input point
(an algorithm satisfying this requirement is ***differentially private***)

Why is that a good privacy definition?
Even if an observer knows all other data point but mine, and now she sees the outcome of the computation, then she still cannot learn “anything” on my data point



What is Differentially Private k-Means?

[Dwork, McSherry, Nissim, Smith 06] (informal)

- ✓ Every data point x_i represents the (private) information of one individual
- ✓ **Goal:** the output (the set of centers) does not reveal information that is specific to any single individual
- ✓ **Requirement:** the output distribution is insensitive to any arbitrarily change of a single input point
  (an algorithm satisfying this requirement is ***differentially private***)

Differentially Private k-Means Clustering

<u>Given</u>: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_{\boldsymbol{i}} \mathbf{min}_{\ell} \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

Requirement: the output distribution is insensitive to any arbitrarily change of a single input point

Differentially Private k-Means Clustering

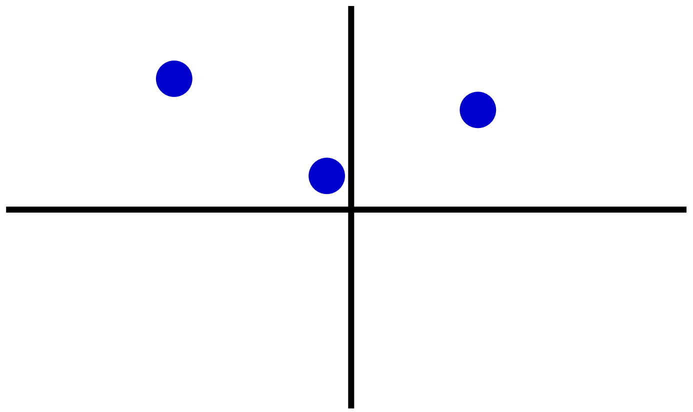
Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \ldots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_{\boldsymbol{i}} \mathbf{min}_{\ell} \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

Requirement: the output distribution is insensitive to any arbitrarily change of a single input point

Observe: With privacy we must have additive error

- Assume $\boldsymbol{k} = \boldsymbol{n} = \mathbf{3}$
- OPT's cost $= 0$



Differentially Private k-Means Clustering

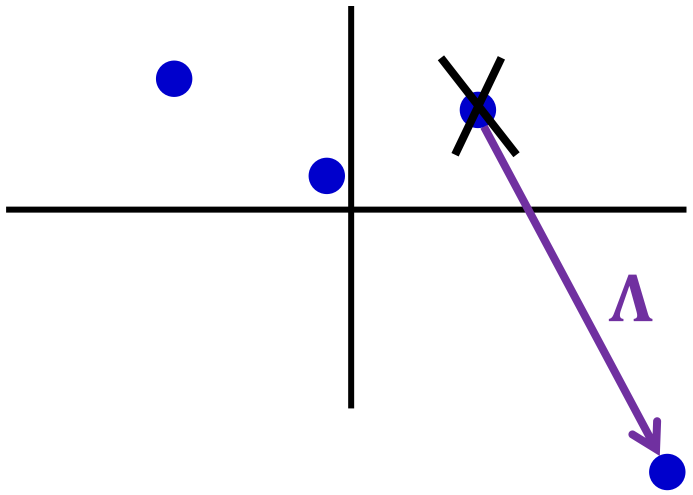
<u>Given</u>: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_{\boldsymbol{i}} \mathbf{min}_{\ell} \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

Requirement: the output distribution is insensitive to any arbitrarily change of a single input point

<u>Observe:</u> With privacy we must have <u>additive</u> error

- Assume $\boldsymbol{k} = \boldsymbol{n} = \mathbf{3}$
- OPT's cost $= 0$
- Move one point
- OPT's cost $= 0$



Differentially Private k-Means Clustering

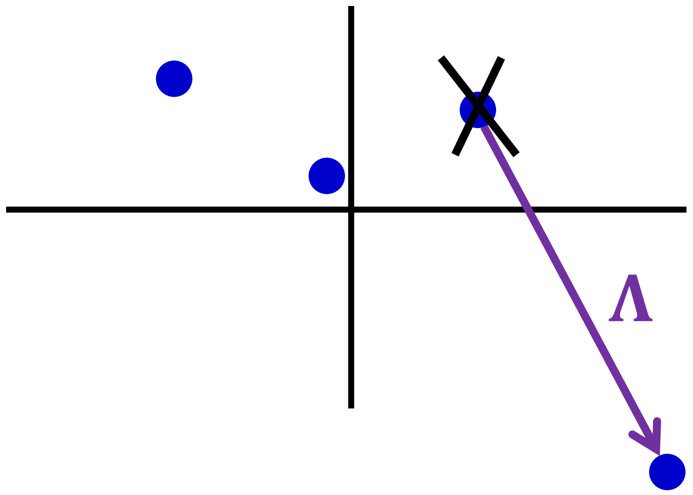
Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^d\right)^n$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_i \mathbf{min}_\ell \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

Requirement: the output distribution is insensitive to any arbitrarily change of a single input point

Observe: With privacy we must have additive error

- Assume $\boldsymbol{k} = \boldsymbol{n} = \mathbf{3}$
- OPT's cost $= 0$
- Move one point
- OPT's cost $= 0$
- Each solution must remain approx. equally likely
- On at least one of these inputs our cost is $\approx \boldsymbol{\Lambda}^{\mathbf{2}}$



Differentially Private k-Means Clustering

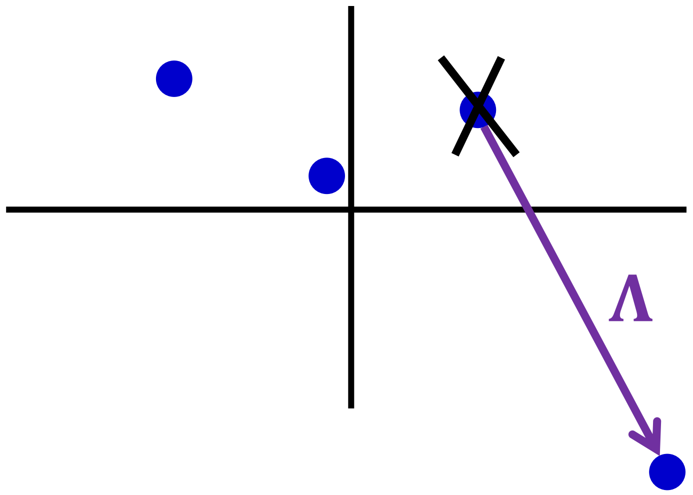
Given: Data points $\boldsymbol{S} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Identify $\boldsymbol{k}$ centers $\boldsymbol{C} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_k})$ minimizing $\mathbf{cost}(\boldsymbol{C}) = \sum_{\boldsymbol{i}} \mathbf{min}_{\ell} \|\boldsymbol{x_i} - \boldsymbol{u_\ell}\|^2$

Requirement: the output distribution is insensitive to any arbitrarily change of a single input point

Observe: With privacy we must have additive error

- Assume $\boldsymbol{k} = \boldsymbol{n} = \mathbf{3}$
- OPT's cost $= 0$
- Move one point
- OPT's cost $= 0$
- Each solution must remain approx. equally likely
- On at least one of these inputs our cost is $\approx \boldsymbol{\Lambda}^{\mathbf{2}}$



$\Longrightarrow$ **We assume that input points come from the unit ball**

Previous and New Bounds

Ref	Model	Runtime	Bounds
GLMRT'10	differential privacy	n^d	$\mathrm{O}(1) \cdot \mathrm{OPT} + \tilde{O}(k^2 \cdot d)$
NCBN'16	differential privacy	poly	$\mathrm{O}(\log k) \cdot \mathrm{OPT} + \tilde{O}(n)$
FXZR'17	differential privacy	poly	$O(k \log n) \cdot \mathrm{OPT} + \tilde{O}\left(k^{3/2} \cdot \sqrt{d}\right)$
BDLMZ'17	differential privacy	poly	$O(\log^3 n) \cdot \mathrm{OPT} + \tilde{O}(k^2 + d)$
NS'18	differential privacy	poly	$O(k) \cdot \mathrm{OPT} + \tilde{O}(k^{1.51} \cdot d^{0.51})$
New	differential privacy	poly	$O(1) \cdot \mathrm{OPT} + \tilde{O}(k^{1.01} \cdot d^{0.51} + k^{3/2})$