Maybe, $\frac{1}{2}$ hour from now you can insert the following in your small talk:

- Differential Privacy
- Coresets
- Private Coresets
- Private Data Structures (not only coresets)
- Private $\varepsilon$ nets
- Private Bi Criteria

New
Why Privacy?

- $r_i \in \{0, 1\}$: indicator variable $= 1$ if $i$ Republican
Why Privacy?

Indicator variable:

\[ r_i = 1 \]

\[ r_i = 0 \]

\[ r = r_1 + \cdots + r_n \]
If everyone has known political opinion but for voter \( n \):

\[
r_n = r - \sum_{i=1}^{n-1} r_i
\]
Differential Privacy [DMNS06]

Algorithm $A$ is $\alpha$-differentially private if:

- for every two sets $P$ and $P'$ that differ by a single item:
- for every set $S$ of possible outputs:

$$\frac{\Pr[A(P) \in S]}{\Pr[A(P') \in S]} \leq e^\alpha \approx 1 + \alpha$$
Private Counting

We publish \( \tilde{r} = r + \text{Noise} \)

\[
\Pr[\tilde{r} \in r + \text{Noise} \pm \epsilon] \approx \epsilon \frac{\alpha}{2} \cdot e^{-\alpha|\text{Noise}|}
\]
Example: $r = 18$

We publish $\tilde{r} = 18 + \text{Noise}$

$$\Pr[\tilde{r} \in 20 \pm \epsilon] \approx \epsilon \frac{\alpha}{2} \cdot e^{-2\alpha}$$

(Noise = 2)
**Example:  \( r = 17 \)**

We publish \( \tilde{r} = 17 + \text{Noise} \)

\[
\Pr \left[ \tilde{r} \in 20 \pm \epsilon \right] \approx \epsilon \frac{\alpha}{2} \cdot e^{-3\alpha}
\]

(Noise = 3)
Private Counting

\[
\frac{\Pr \left[ \tilde{r} \in 20 \pm \epsilon \mid r = 18 \right]}{\Pr \left[ \tilde{r} \in 20 \pm \epsilon \mid r = 17 \right]} = \frac{e^{-2\alpha}}{e^{-3\alpha}} = e^{\alpha} \approx 1 + \alpha
\]

\(\tilde{r} = r + \text{Noise} \) is \(\alpha\)-differentially private
Strong Notion of Privacy

The attacker learns little "useful" Prior Information does not help

Because of $\varepsilon$ leakage: Cannot be used to answer many queries
Strong Notion of Privacy

Want to answer not one query privately
But many queries privately

Leak $\varepsilon$ only once, create Sanitized Data Set/Data Structure
k-Median Queries

- Input: $P \subseteq [0, 1]^d$

No privacy
**k-Median Queries**

- **Input:** \( P \subseteq [0, 1]^d \)
- **Query:** A set \( Q \) of \( k \) points

\[
k = 2 \\
Q = \{q_1, q_2\}
\]
**k-Median Queries**

- **Input:** $P \subseteq [0, 1]^d$
- **Query:** A set $Q$ of $k$ points
- **Output:** $\sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \|p - q\|$

$k = 2$

$Q = \{q_1, q_2\}$
Motivation
Comparing alternatives:
How good is this placement?

No privacy

Party Branch
Party Faithful
Coresets

- Coresets: “Clever Sample”
- Answer approximate queries from reduced representation (Coreset)
- Often leads to PTAS, FPTAS
- Many, many, papers, surveys
- Many problems: median, mean, flats, projective clustering, regression
- **Intuition**: Coresets give privacy **on average**

No privacy
(k, \varepsilon)-Median Coreset

Answer k-median queries in sub-linear time

Key Idea: Replace many points by one weighted representative:
(k, \varepsilon)-Median Coreset

Answer k-median queries in sub-linear time

Key Idea: Replace many points by one weighted representative:
(k, \varepsilon)-Median Coreset

Key Idea: Replace many points by one weighted representative:

- \( w(c_1) = 3 \)
- \( w(c_2) = 100 \)
- \( w(c_3) = 1 \)
- \( w(c_4) = 2 \)
- \( w(c_5) = 1 \)

No privacy
(k, ε)-Median Coreset

**Key Idea:** Replace many points by one weighted representative:

- $w(c_1) = 3$
- $w(c_2) = 100$
- $w(c_3) = 1$
- $w(c_4) = 2$
- $w(c_5) = 1$
(\(k, \varepsilon\))-Median Coreset

\[
\sum_{p \in P} \text{dist}(p, Q) \sim \sum_{c \in C} w(c) \cdot \text{dist}(c, Q)
\]
Definition

$C$ is a $(k, \varepsilon)$-coreset for $P$, if $\forall Q, |Q| = k$:

$$\sum_{p \in P} \text{dist}(p, Q) \sim \sum_{c \in C} w(c) \cdot \text{dist}(c, Q)$$

Multiplicative error $\leq 1 + \varepsilon$
Additive error $\leq \frac{1}{\varepsilon}$

No privacy
Locating Branch Offices

No privacy
Private (Republican) Coresets

Intuition: Coresets reveal little information

No privacy
Coresets & Privacy

**Good:** Coresets reveal little information on average

**Bad:** Coresets are not differential private
Private Coreset Scheme

An algorithm that:

- is $\alpha$-differentially private.

- for $P \subseteq [0, 1]^d$, outputs a $(k, \varepsilon)$-coreset, w.h.p.
Our Contributions

1. [Simple, non-constructive]:

   \[ k\text{-median coreset} \rightarrow \text{Private } k\text{-median coreset} \]
   \[ k\text{-mean coreset} \rightarrow \text{Private } k\text{-mean coreset} \]
   \[ \ldots \]

Using Exp. Mechanism of \textbf{[MT07]}
Our Contributions

2. [Constructive, linear time]:
   - Private $k$-median coreset
   - Private $k$-mean coreset
Our Contributions

2. [Constructive, linear time]:
   - Private $k$-median coreset
   - Private $k$-mean coreset

3. Lower bound tradeoffs on multiplicative-additive approximation for private coresets
Applications

- Private k-median clustering
- Comparing alternatives privately
- Private streaming algorithms
- Approximately truthful mechanisms [MT07]
Related Work

- Sanitized Database [BLR08]
- (Non-private) coresets for $k$-median [HM04][HK05][FS05][Chen06][FMS07]
- Private clustering [BDMN05][NRS07]
Overview

- Private coreset for 1-median, $P$ on line.
Overview

- Private coreset for 1-median, $P$ on line.
- Private coreset for 1-median, $P$ in $[0,1]^d$
Overview

- Private coreset for 1-median, $P$ on line.
- Private coreset for 1-median, $P$ in $[0, 1]^d$
- Private bi-criteria approximation for $k$-median
Overview

- Private coreset for 1-median, $P$ on line.

- Private coreset for 1-median, $P$ in $[0, 1]^d$

- Private bi-criteria approximation for $k$-median

- Private coresets for $k$-median, $P \subseteq [0, 1]^d$
Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]
Coreset for $P \subseteq [0, 1], k = 1$ [HMO04]

$$2\text{opt} = 2 \sum_{p \in P} \text{dist}(p, \bar{p})$$

$(1 + \varepsilon)^2 \cdot \text{opt}/n$

$(1 + \varepsilon) \cdot \text{opt}/n$

$\text{opt}/n$

$\bar{p}$
Coreset for $P \subseteq [0,1]$, $k = 1$ [HMO04]
Coreset for $P \subseteq [0, 1]$, $k = 1$ [HMO04]

For each interval $I$:

- Choose an arbitrary representative $c \in P \cap I$
- $w(c) \leftarrow |P \cap I|$

$$2\text{opt} = 2 \sum_{p \in P} \text{dist}(p, \bar{p})$$

- $w(c_1) = 3$
- $w(c_2) = 100$
- $w(c_5) = 1$
Main Observation: \(|I| \leq \varepsilon |J|\)

Because the size of the intervals forms a geometric sequence of ratio \((1 + \varepsilon)\)
\[
\text{error}(q) = \left| \sum_{p \in P} \text{dist}(p, q) - \sum_{p \in P} \text{dist}(c_p, q) \right|
\]

\[
\leq \sum_{p \in P} \text{dist}(p, c_p)
\]
error(q) = \left| \sum_{p \in P} \text{dist}(p, q) - \sum_{p \in P} \text{dist}(c_p, q) \right| \\
\leq \sum_{p \in P} \text{dist}(p, c_p) \leq \sum_{p \in P} \varepsilon \cdot \text{dist}(p, \bar{p})
error = \left| \sum_{p \in P} \text{dist}(p, q) - \sum_{p \in P} \text{dist}(c_p, q) \right|

\leq \sum_{p \in P} \text{dist}(p, c_p) \leq \sum_{p \in P} \varepsilon \cdot \text{dist}(p, \bar{p})

\leq 2\varepsilon \cdot \text{opt}

\leq 2\varepsilon \sum_{p \in P} \text{dist}(p, q)
New: Private Coreset

For each interval \( \mathbf{I} \):
- Choose the rightmost point \( c \in \mathbf{I} \)
- \( \tilde{w}(c) \leftarrow |P \cap \mathbf{I}| + \text{Noise} \)
Coreset for $P \subseteq [0,1], k = 1$ [HM04]

$$\sum_{p \in P} \text{dist}(p, q) - \sum_{c \in C} w(c) \cdot \text{dist}(c, q) \leq \varepsilon \sum_{p \in P} \text{dist}(p, q)$$

New: Private Coreset

$$\sum_{p \in P} \text{dist}(p, q) - \sum_{c \in C} w(c) \cdot \text{dist}(c, q) \leq \varepsilon \sum_{p \in P} \text{dist}(p, q) + \mathcal{O}\left(\frac{1}{\varepsilon}\right)$$
Generalization for $P \subseteq [0, 1]^d$
Generalization for $P \subseteq [0,1]^d$
Generalization for $P \subseteq [0,1]^d$
Generalization for $P \subseteq [0,1]^d$
Generalization for $P \subseteq [0, 1]^d$
Generalization for $k > 1$
The $k$-Median of $P$

$$\text{opt} = \min_{|\text{OPT}|=k} \sum_{p \in P} \text{dist}(p, \text{OPT})$$
The $k$-Median of $P$

$$\text{opt} = \min_{|\text{OPT}| = k} \sum_{p \in P} \text{dist}(p, \text{OPT})$$
Constant Approximation

\[ |\widetilde{OPT}| = k, \quad \sum_{p \in P} \text{dist}(p, \widetilde{OPT}) \leq c \cdot \text{opt} \]
Bi-Criteria Approximation

$|B| = O(k \log n)$, 

$\sum_{p \in P} \text{dist}(p, B) \leq c \cdot \text{opt}$
Generalization for $k > 1$
Compute Private Bi-Criteria Approx.
Based on [FFSS07]
On Each Cluster:

Apply construction for $k = 1$
Weak $\varepsilon$-Net $N$ for $P \subseteq [0, 1]$
Weak $\varepsilon$-Net $N$ for $P \subseteq [0, 1]$

For every interval $I$:

$$|I \cap P| \geq \varepsilon n \implies |I \cap N| \geq 1$$
Weak $\frac{1}{4}$-Net $N$ for $P \subseteq [0,1]$

For every interval $I$:

$$|I \cap P| \geq n/4 \implies |I \cap N| \geq 1$$
Weak $\varepsilon$-Net for $P \subseteq [0, 1]^d$
Weak $\varepsilon$-Net for $P \subseteq [0,1]^d$

$|B \cap P| \geq \varepsilon n$

$|B \cap N| \geq 1$
Weak $\varepsilon$-Net for $P \subseteq [0,1]^d$
Weak $\varepsilon$-Net for $P \subseteq [0, 1]^d$
Weak $\varepsilon$-Net for $P \subseteq [0, 1]^d$
Private $\varepsilon$-Net for $P \subseteq [0, 1]^d$

Add noise to each representative
Input

A set of $n$ points $P \subset \mathbb{R}^d$, $k \geq 1$. 
Output

$N$ : a small bicriteria approximation to the $k$ median of $P$
The Bicriteria Algorithm

1) $t \leftarrow 1$
2) $N \leftarrow \emptyset$
3) Construct a weak $(\frac{1}{8k})$-net $N_t$ for $P$
4) $N \leftarrow N \cup N_t$
5) Discard $P_t$: $P/2$ pts closer to $N_t$
6) $t \leftarrow t + 1$
7) Repeat steps 3 to 6 until no more pts
8) Return $N$
Thank You!
A point $b \in P$ is bad for $N_t$, if:

$$\text{dist}(b, N_t) > 2 \text{dist}(b, N^*)$$
A point $g \in P$ is good for $N_t$ otherwise:

$$\text{dist}(g, N_t) \leq 2 \text{dist}(g, N^*)$$
Main Technical Theorem

We can map every bad point $b \in P_t$ to a distinct good point $g \in P_{t+1}$.
\[ \text{dist}(b, N) \leq \text{dist}(b, N_t), \text{ because } N \supseteq N_t. \]

Since \( b \in P_t \) and \( g \in P_{t+1} \):

\[ \text{dist}(b, N_t) \leq \text{dist}(g, N_t) \]

Since \( g \) is good for \( N_t \):

\[ \text{dist}(g, N_t) \leq 2 \text{dist}(g, N^*) \]
\[ \text{dist}(b, N) \leq \text{dist}(b, N_t), \text{ because } N \supseteq N_t. \]

Since \( b \in P_t \) and \( g \in P_{t+1} \):

\[ \text{dist}(b, N_t) \leq \text{dist}(g, N_t) \]

Since \( g \) is good for \( N_t \):

\[ \text{dist}(g, N_t) \leq 2 \text{dist}(g, N^*) \]

\[ \text{dist}(b, N) \leq 2 \text{dist}(g, N^*) \]
Bi-Criteria for $k$-Median

$$\sum_{p \in P} \text{dist}(p, N) = \sum_{g} \text{dist}(g, N) + \sum_{b} \text{dist}(b, N)$$

$$\leq \sum_{g} 2 \text{dist}(g, N^*) + \sum_{g} 2 \text{dist}(g, N^*)$$

$$\leq 4 \sum_{p \in P} \text{dist}(p, N^*)$$
Open Questions

- Private coresets for $k$-median in high dimensional spaces
- Private coresets for $k$ subspaces of $\mathbb{R}^d$
- Private coresets for other shapes.
- Private dynamic Coresets
Bi-Criteria Approximation Algorithm [FFS07]
Initialization

1) \( t \leftarrow 1 \)

▷ Counter for iterations

2) \( F \leftarrow \emptyset \)

▷ The output set of \( j \)-flats
3) Construct a weak $\left(\frac{1}{8k}\right)$-net $N_t$ for $P$

$t = 1$

ADS 2009 Bertinoro
4) \( N \leftarrow N \cup N_t \)

\((t = 1)\)
5) $\forall p : \text{Compute } \text{dist}(p, N_t)$

$(t = 1)$
6) Remove $P_t$: the half of $P$ that is closer to $N_t$

$(t = 1)$
6) Remove $P_t$: the half of $P$ that is closer to $N_t$

$t = 1$
7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6:
3) Construct a weak \((1/k)\)-net \(N_t\) for \(P\)

\((t = 2)\)
4) \( N \leftarrow N \cup N_t \)

\((t = 2)\)
\[ 5) \ \forall p : \text{Compute } \text{dist}(p, N_t) \]

\[(t = 2)\]
6) Remove $P_t$: the half of $P$ that is closer to $N_t$

$(t = 2)$
6) Remove $P_t$: the half of $P$ that is closer to $N_t$

$(t = 2)$
6) Remove $P_t$: the half of $P$ that is closer to $N_t$(\hspace{1cm} \textcolor{red} {t = 2})
7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6 till there are no more input points.

9) Return $N$:
Let $N^*$ be any set of $k$ points in $\mathbb{R}^d$. 
Let $N^*$ be any set of $k$ points in $\mathbb{R}^d$. 
Let $N^*$ be any set of $k$ points in $\mathbb{R}^d$.

Consider $N_t$ that is constructed during the $t^{th}$ iteration.
A point $b \in P$ is bad for $N_t$, if:

$$\text{dist}(b, N_t) > 2 \text{dist}(b, N^*)$$
A point $g \in P$ is good for $N_t$ otherwise:

$$\text{dist}(g, N_t) \leq 2 \text{dist}(g, N^*)$$
Main Technical Theorem

We can map every bad point $b \in P_t$ to a distinct good point $g \in P_{t+1}$.
dist(b, N) \leq dist(b, N_t), \text{ because } N \supseteq N_t.

Since \( b \in P_t \) and \( g \in P_{t+1} \):

\[
dist(b, N_t) \leq dist(g, N_t)
\]

Since \( g \) is good for \( N_t \):

\[
dist(g, N_t) \leq 2 \cdot dist(g, N^*)
\]
\[ \text{dist}(b, N) \leq \text{dist}(b, N_t), \text{ because } N \supseteq N_t. \]

Since \( b \in P_t \) and \( g \in P_{t+1} \):

\[ \text{dist}(b, N_t) \leq \text{dist}(g, N_t) \]

Since \( g \) is good for \( N_t \):

\[ \text{dist}(g, N_t) \leq 2 \text{dist}(g, N^*) \]

\[ \text{dist}(b, N) \leq 2 \text{dist}(g, N^*) \]
Bi-Criteria for \( k \)-Median

\[
\sum_{p \in P} \text{dist}(p, N) = \sum_{g} \text{dist}(g, N) + \sum_{b} \text{dist}(b, N) \\
\leq \sum_{g} 2 \text{dist}(g, N^*) + \sum_{g} 2 \text{dist}(g, N^*) \\
\leq 4 \sum_{p \in P} \text{dist}(p, N^*)
\]
Proof of the Technical Theorem

• The number of bad points is at most

\[ |B| = \frac{|P_t|}{8} \]

• \[ |P_{t+1}| = \frac{|P_t|}{2} \]

The number of good points in $P_{t+1}$ is at least

\[ |P_{t+1}| - |B| \geq \frac{|P_t|}{2} - \frac{|P_t|}{8} \geq |B| \]
Claim: Only $B_0 = \frac{|P_t|}{8k}$ points are bad for $q \in N_t$.

$\text{dist}(p, q) \leq 2 \text{dist}(p, q^*)$
$B_0$: the $\frac{|P_t|}{8k}$ closest points to $q^*$
$B_0$: the $\frac{|P_t|}{8k}$ closest points to $q^*$

$B_0$ contains $q \in N_t \left( \frac{1}{8k} \text{-net} \right)$
For every yellow point \( p \in P \setminus B_0 \):

\[
\text{dist}(p, q) \leq \text{dist}(p, q^*) + \text{dist}(q^*, q) \leq 2 \text{dist}(p, q^*)
\]
All the yellow points are good for $N_t$

$$\text{dist}(p, q) \leq 2 \text{dist}(p, q^*)$$
Only the black points $B_0$ are bad for $N_t$

\[ |B_0| = \frac{|P_t|}{8} \]