



Core-sets for Fair and Diverse Data Summarization

Sepideh Mahabadi

Microsoft Research

Stojan Trajanovski

Microsoft

Diversity Maximization

Diversity Maximization

Given a set of objects, how to pick **a few** of them while maximizing **diversity**?

Applications

- **Summarization (e.g. User's Feed, Video, Documents, Images)**

Applications

- **Summarization**

- User's feed Generation

- A set of users

- Each with a set of messages

- People who they interact with

- The channels they are part of

- ...

- Which messages to show in their feed?

- Relevant messages are shown to the users based on user's likes and replies

- Need to have diversity in the retrieved summary

Applications

- Summarization (e.g. User's Feed, Documents, Images)
- Searching

The image shows a Google search interface for the term "Jaguar". The search bar contains the word "Jaguar" and is surrounded by navigation tabs for "All", "Images", "Shopping", "Videos", "News", and "More". Below the search bar, there are several suggested search terms: "car", "cat", "f type", "f pace", "drawing", "leopard", "wallpaper", "logo", "wild", "panther", "price", and "convertible".

The search results are divided into two main sections:

- Jaguar Animal:** This section features a grid of images of jaguars in various settings. The first image is from Wikipedia, showing a jaguar on a rock. Other images are from National Geographic, Britannica, St. Louis Zoo, World Wildlife Fund, Zurich Insurance Group, and The Living Desert Zoo.
- Jaguar Cars:** This section features a grid of images of Jaguar vehicles. The first image is from Jaguar USA, showing a dark Jaguar SUV. Other images are from Car.USNews (2019 Jaguar XF Review), Jaguar Manhattan (The Jaguar Symbol | History of th...), Cars.com (2021 Jaguar XF Specs, Price,...), Carscoops (Jaguar Confirms 4-Door EV GT W...), and Kelley Blue Book (2024 Jaguar F-TYPE Pric...).

Applications

- **Summarization (e.g. User's Feed, Documents, Images)**
- **Searching**
- **Recommendation Systems**
 - **Movies, News articles**
 - **Shopping**
 - **Hiring Candidates e.g. for LinkedIn**

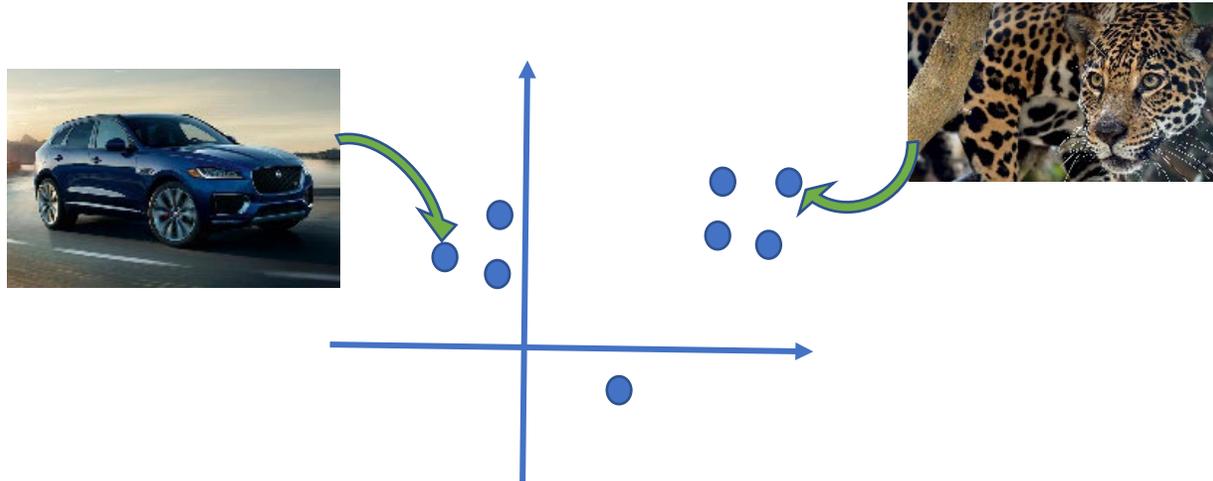
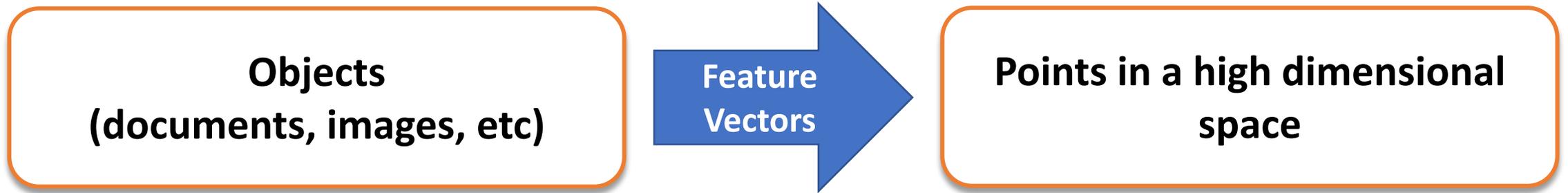


Image from: <http://news.mit.edu/2017/better-recommendation-algorithm-1206>

Applications

- **Summarization (e.g. User's Feed, Documents, Images)**
- **Searching**
- **Recommendation Systems**
- ...

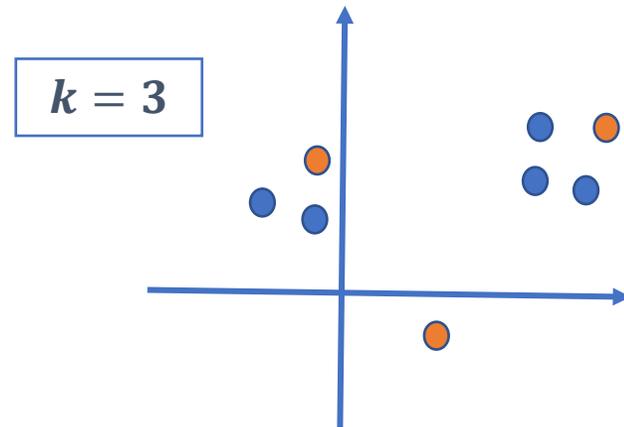
Modeling the Objects



The Diversity Maximization problem

Given: a set of n points P in a metric space and a parameter k ,

Goal: pick a subset $S \subseteq P$ of k points while maximizing “diversity”.



Diversity Notions

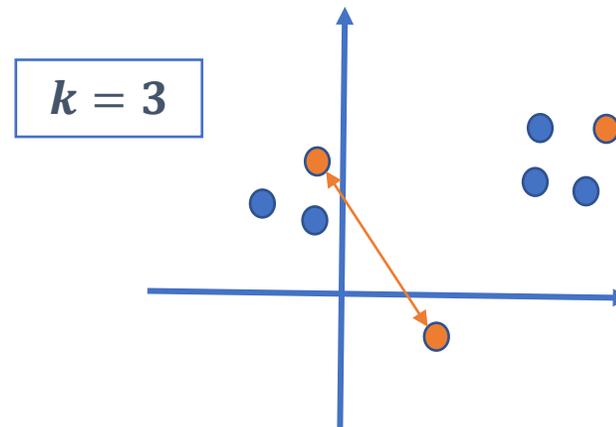
Diversity I: Minimum Pairwise Distance

Input: a set of n vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

Goal: pick k points s.t. the **minimum pairwise distance** of the picked points is maximized.

$$\min_{p, q \in S} \text{dist}(p, q)$$

- **$O(1)$ -approx Greedy Algorithm**
[RRT'94]



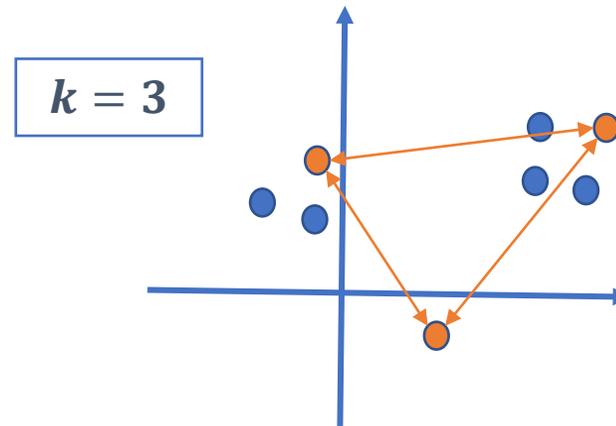
Diversity II: Sum of Pairwise Distances

Input: a set of n vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

Goal: pick k points s.t. the **sum pairwise distances** of the picked points is maximized.

$$\sum_{p, q \in S} \text{dist}(p, q)$$

- $O(1)$ -approx Local Search
Algorithm [HRT'97][AMT'13]



Diversity III: Sum of Nearest Neighbor Distances

Input: a set of n vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

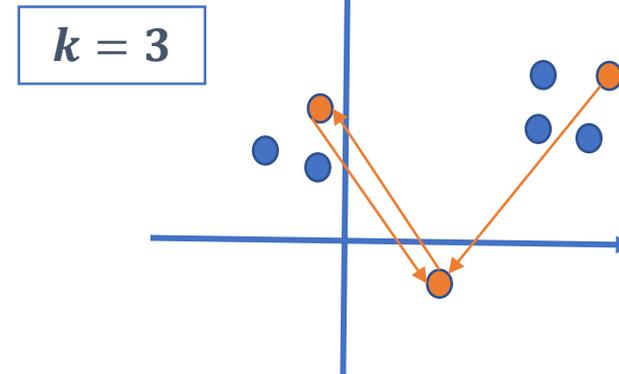
Goal: pick k points s.t. the **sum of NN distances** of the picked points is maximized.

$$\sum_{p \in S} \min_{q \in S \setminus \{p\}} \text{dist}(p, q)$$

□ Between Min-Pairwise Dist and Sum of Pairwise Dists

□ $\mathcal{O}(\log k)$ -approx Alg [CH'01]

□ $\mathcal{O}(1)$ -approx Alg [BGMS'16]



Diversity Notions

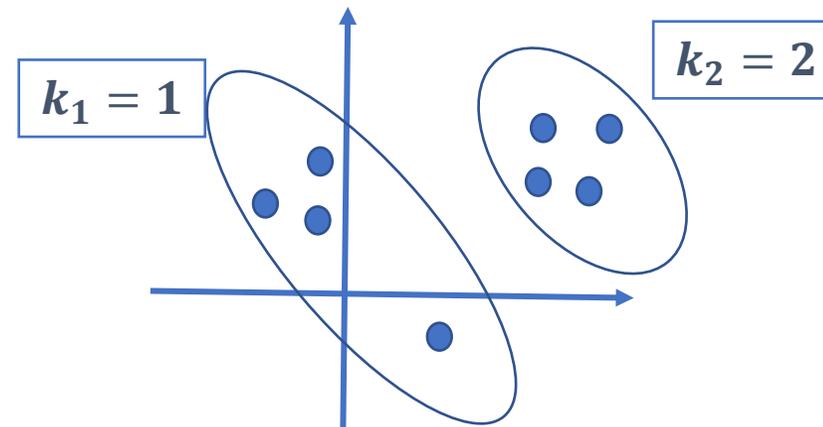
Diversity Notion		Offline
Min Pairwise Distance	$\min_{p,q \in S} \text{dist}(p, q)$	$\theta(1)$ [Ravi et al 94]
Sum of Pairwise distances	$\sum_{p,q \in S} \text{dist}(p, q)$	$\theta(1)$ [Hassin et al 97]
Sum of NN Distances	$\sum_{p \in S} \min_{q \in S \setminus \{p\}} \text{dist}(p, q)$	$\theta(1)$ [BGMS'16]
...

Constrained(Fair) Diversity Maximization

Constrained/Fair Diversity Maximization

Input:

- sets of vectors P_1, \dots, P_m , $P = \cup_i P_i$
- and k_1, \dots, k_m , $k = \sum_i k_i$

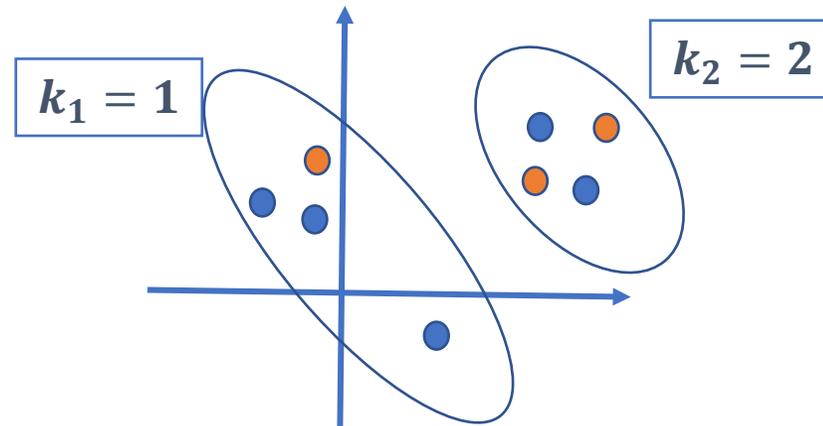


Constrained/Fair Diversity Maximization

Input:

- sets of vectors P_1, \dots, P_m , $P = \cup_i P_i$
- and k_1, \dots, k_m , $k = \sum_i k_i$

Goal: pick k_i points $S_i \subset P_i$ s.t. the diversity of the picked points $S = \cup_i S_i$ is maximized.



Prior Work: Fair Diversity Maximization

Diversity Notion	FDM
Min Pairwise Distance	$\theta(m)$ [MMM20, AMMM'22]
Sum of Pairwise distances	$\theta(1)$ [AMM'13]
Sum of NN Distances	$\theta(1)$ [BGMS'16]

Application I: in User's Feed Generation

- Each message has a posted time
- Goal: show more recent messages and less old ones
- Still need diversity
- Modeling **Recency**
 - Divide the messages in a month into four groups based on the week they have been posted
 - Set k_i to be higher for more recent weeks
- Data Set: Reddit Messages
 - Messages of a single month (~21000 messages) and divide it into four groups based on the week they appear in

Application II: Movie Recommendation

- Task: Movie recommendation
- Goal: assign budgets for each genre, e.g. comedy, action, drama, ...
- MovieLens Data Set
 - Collection of 4000 movies
 - Group based on the movie genre into 18 groups (e.g. “documentary”, “crime”, “drama”, “action”, ...)

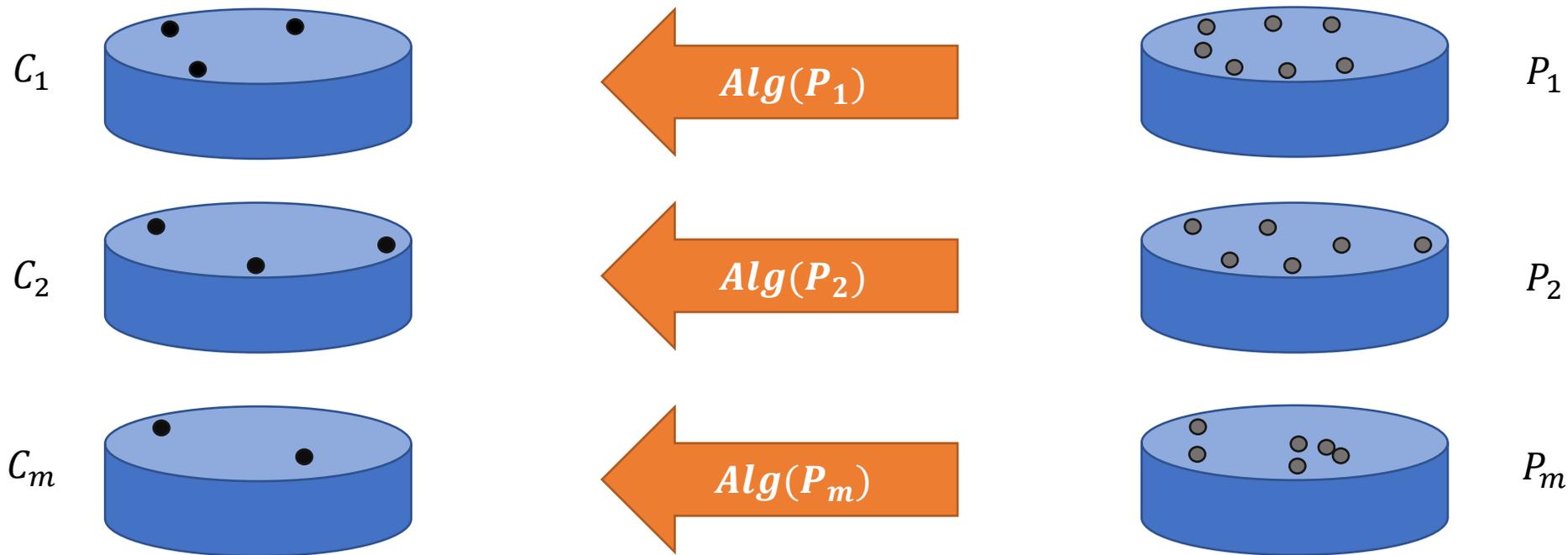
Experimental Results

- 1. Need for FDM:** As expected, in the unconstrained version, the recency is not preserved

Experimental Results

- 1. Need for FDM:** As expected, in the unconstrained version, the recency is not preserved
- 2. Price of Balancedness:** diversity loss by resorting to FDM
 - 1%, for sum-of-pairwise distances
 - 20%, for sum of NN-distances
 - 50%, for minimum pairwise distance

FDM under Big Data Model: Coresets



$$div_{k_1, \dots, k_m} \left[\begin{array}{c} \text{Cylinder with 6 black dots} \\ C \end{array} \right] \geq \frac{1}{\alpha} \cdot div_{k_1, \dots, k_m} \left[\begin{array}{c} \text{Cylinder with 12 gray dots} \\ P \end{array} \right]$$

Theoretical Results

- ✓ Algorithms are simple to implement
- ✓ Show a new offline algorithm for FDM under Sum-of-NN-Distances

Diversity Notion	FDM	Coreset Setting		
		Approx.	Coreset Size	Reference
Min Pairwise Distance	$\theta(m)$ [MMM20, AMMM'22]	$O(1)$	$O(k)$ <i>per group</i>	[MMM20]
Sum of Pairwise distances	$\theta(1)$ [AMM13]	$(1 + \epsilon)$	Depends on n or aspect ratio	[CPP18]
		$O(1)$	$O(k_i^2)$ <i>per group</i>	[This work]
Sum of NN Distances	$\theta(1)$ [BGMS'16]	$O(m \cdot \log k)$	$O(k^2)$ <i>per group</i>	[This work]

Experimental Results

- 1. Need for FDM:** As expected, in the unconstrained version, the recency is not preserved
- 2. Price of Balancedness:** diversity loss by resorting to FDM
 - 1%, for sum-of-pairwise distances
 - 20%, for sum of NN-distances
 - 50%, for minimum pairwise distance
- 3. Using Coresets**
 - The runtime of the algorithm improves by a factor of **100x**
 - The diversity is only lost by a **few precents**.
 - No **need to recompute the summary** of old messages.

Experimental Results

- 1. Need for FDM:** As expected, in the unconstrained version, the recency is not preserved
- 2. Price of Balancedness:** diversity loss by resorting to FDM
 - 1%, for sum-of-pairwise distances
 - 20%, for sum of NN-distances
 - 50%, for minimum pairwise distance
- 3. Using Coresets**
 - The runtime of the algorithm improves by a factor of **100x**
 - The diversity is only lost by a **few precents**.
 - No **need to recompute the summary** of old messages.
- 4. Show superiority of our coreset construction algorithm over Prior work**

Summary

Diversity Notion	FDM	Coreset Setting		
		Approx.	Coreset Size	Reference
Min Pairwise Distance	$\theta(m)$ [MMM20, AMMM'22]	$O(1)$	$O(k)$ per group	[MMM20]
Sum of Pairwise distances	$\theta(1)$ [AMM13]	$(1 + \epsilon)$	Depends on n or aspect ratio	[CPP18]
		$O(1)$	$O(k_i^2)$ per group	[This work]
Sum of NN Distances	$\theta(1)$ [BGMS'16]	$O(m \cdot \log k)$	$O(k^2)$ per group	[This work]

- Algorithms are simple to implement
- Showed effectiveness of coresets

THANK YOU!