研究紹介

大阪大学大学院情報科学研究科
天方 大地

結び
MUSUBI

大阪大学
OSAKA UNIVERSITY

自己紹介

- Name: **Daichi Amagata**（天方 大地）
- Job: Assistant Professor at Osaka University
  - BE: Osaka University
  - MSc: Osaka University（佐々木先生はsenior studentだった）
  - Ph.D. (information science) Osaka University（鬼塚先生は博士論文の副査担当）
    - DC1
- Love:
  - Football (watching w-cup, EURO, and champions league)
  - Reading comics (>1800 e-comics ≫ #papers read ever)

研究テーマ：データベース，データマイニング，および機械学習のための高速アルゴリズム

問い合わせ

結果の出力

- **ここの計算をとにかく高速に！**
  Google & Microsoft も自社内で精力的に開発
  Meta は良さそうなアルゴリズムをFaiss lib. に導入
- **実際に動く & 理論的性能保証！**

研究テーマ：データベース，データマイニング，および機械学習のための高速アルゴリズム

巨大な入力

X

= n 個のデータの集合

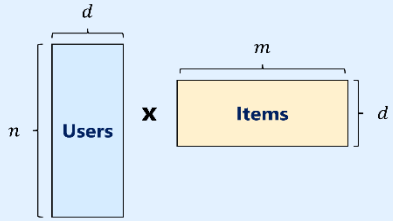
～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

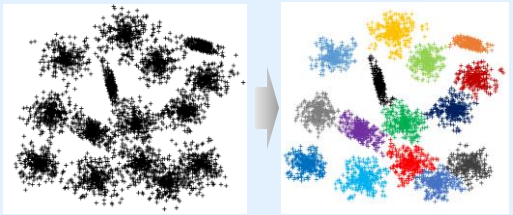
Y

探索

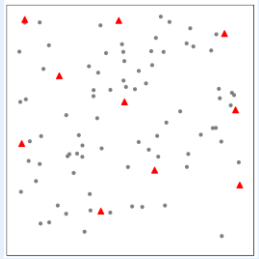
	p_1	p_2	p_3	p_4	p_5
u_1	4.1	**5.0**	2.5	1.8	3.0
u_2	**4.9**	4.3	2.9	4.5	2.1
u_3	2.3	**4.6**	4.1	3.1	1.5
u_4	3.3	**4.1**	3.4	3.6	1.9
u_5	2.1	3.9	2.6	**4.3**	3.1



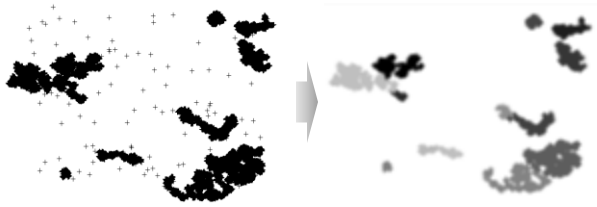
クラスタリング



要約



外れ値検出



探索：条件に合致したデータを出力

巨大な入力

X

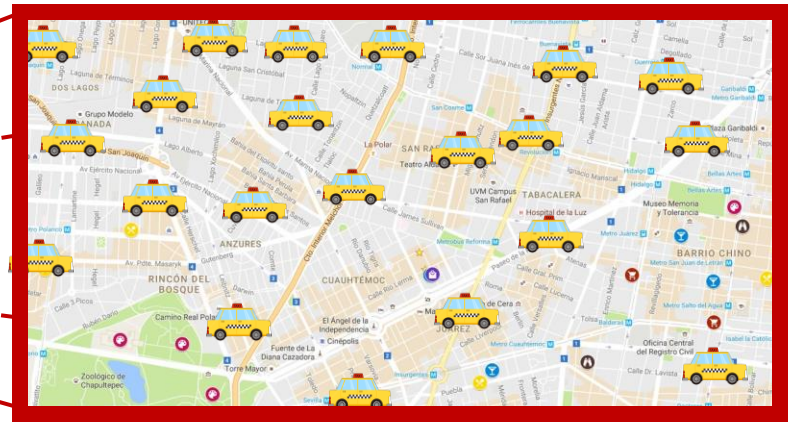
= n 個のデータの集合

フィルタ

出力

Y

Y に関わるデータにだけアクセスすることで高速化を実現



クラスタリング：Xをグループに分ける

巨大な入力

X

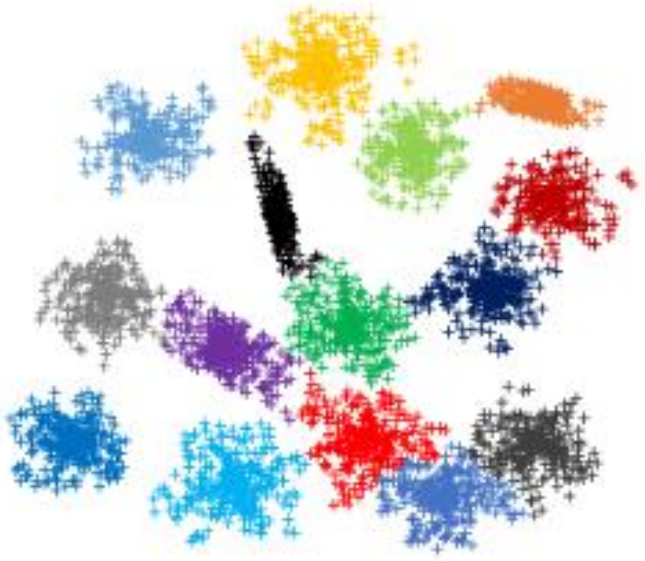
$= n$ 個のデータの集合

分類

出力

Y

クラスタとはこうである，を定式化したとき，
そのクラスタを出力するために必要な処理
のみを行うことにより高速化を実現



要約 : $X \to X'$ s.t. $|X'| \ll |X|$

巨大な入力

X

= n 個のデータの集合

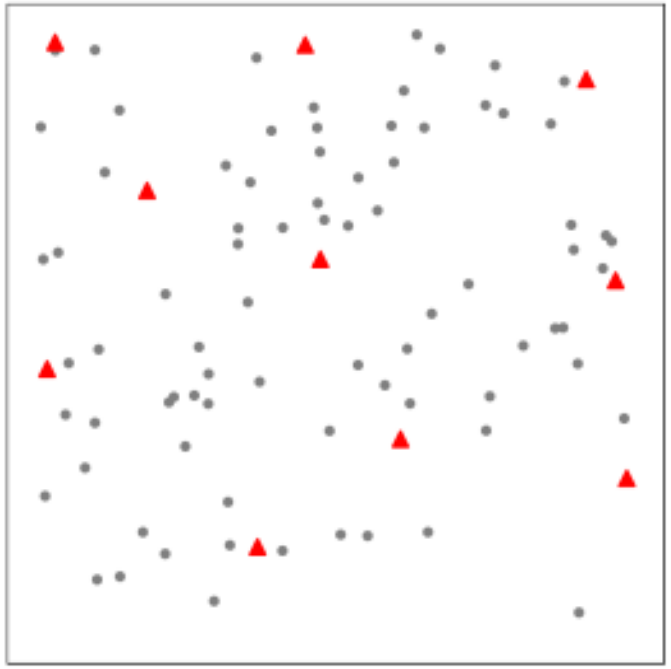
選択

出力

Y

$$S^* = \arg\max_{S \subseteq X, |S|=k} f(S)$$

$$f(S) = \min_{x,x' \in S} dist(x, x')$$



要約とはこうである，を定式化し，その要約を出力するために
必要な処理のみを行うことにより高速化を実現

外れ値検出：条件に合致したデータを出力

巨大な入力

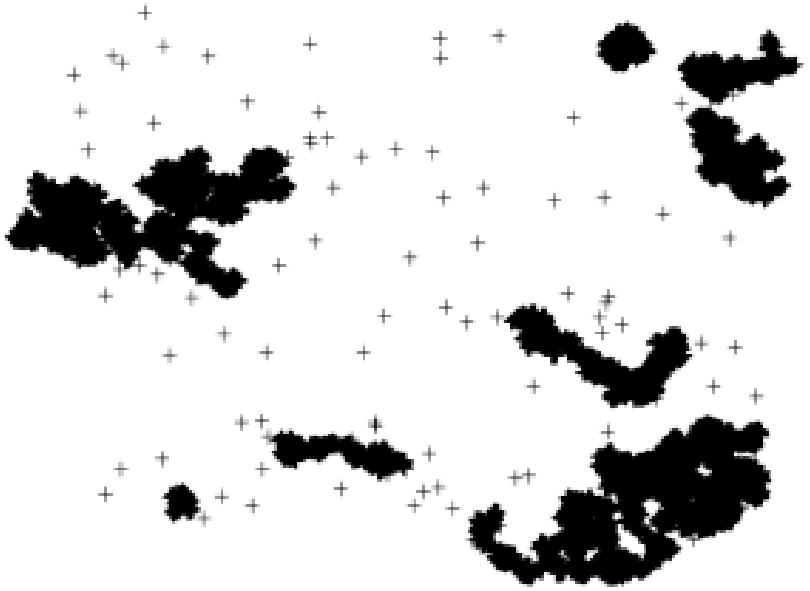
X

$= n$ 個のデータの集合

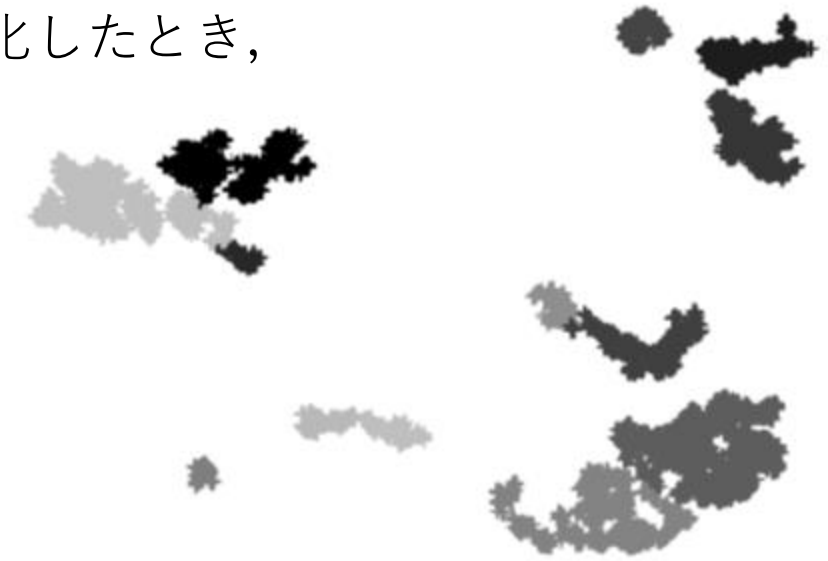
分類

出力

Y



外れ値とはこうである，を定式化したとき，
各データに対して
外れ値？ or not？ （← 探索）
を高速に判定する仕組みを設計



研究テーマ：データベース，データマイニング，および機械学習のための高速アルゴリズム

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

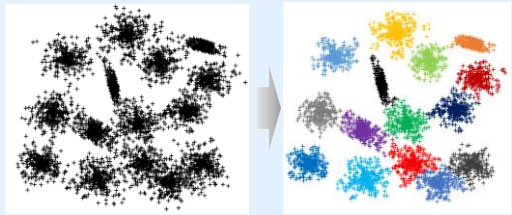
Y

探索

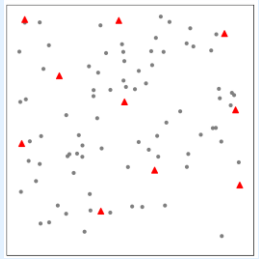
	p_1	p_2	p_3	p_4	p_5
u_1	4.1	**5.0**	2.5	1.8	3.0
u_2	**4.9**	4.3	2.9	4.5	2.1
u_3	2.3	**4.6**	4.1	3.1	1.5
u_4	3.3	**4.1**	3.4	3.6	1.9
u_5	2.1	3.9	2.6	**4.3**	3.1

n Users d × Items m d

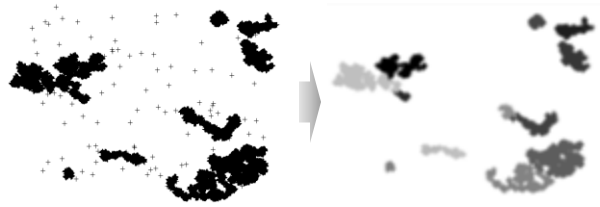
クラスタリング



要約



外れ値検出



公平性の概念を組み込んだ定式化

この分野における「公平性」ってなに？？？

（たぶん）定義は沢山ありますが，私が扱う研究では

公平性 ≒ 平等

「探索」，「クラスタリング」，および「要約」における平等ってなに？？？

この分野における「公平性」ってなに？？？

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

Y

公平性 ≒ 平等

「探索」，「クラスタリング」，および「要約」における平等ってなに？？？

この分野における「公平性」ってなに？？？

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

要件を満たす集合

Z

出力

Y

公平性 ≒ 平等

「探索」，「クラスタリング」，および「要約」における平等ってなに？？？

探索問題における「公平性」ってなに？？？

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

要件を満たす集合

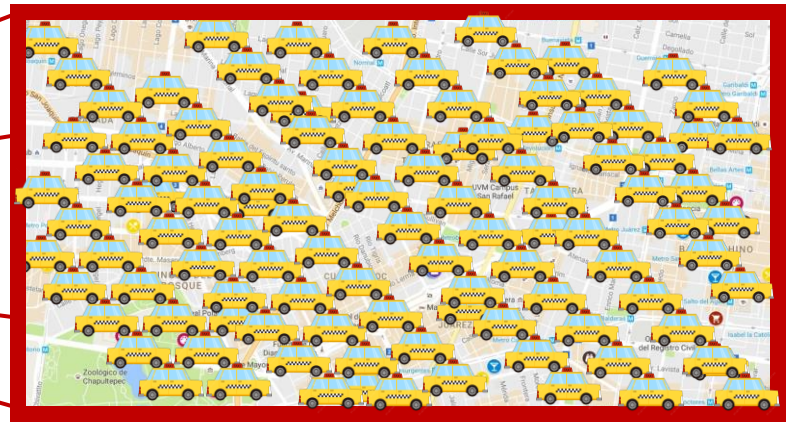
Z

= m 個のデータの集合

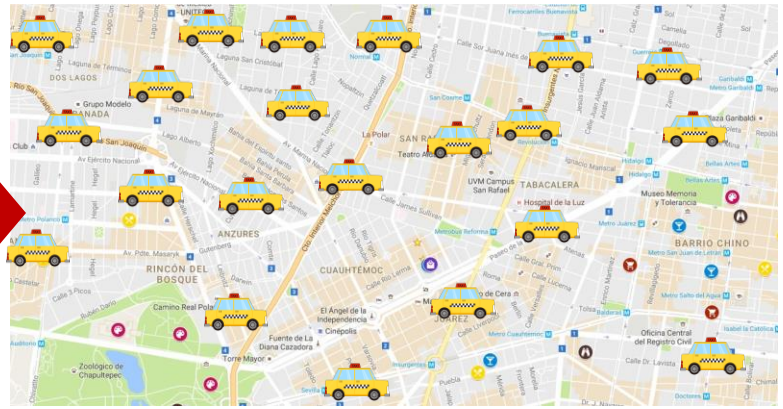
出力

Y

$\forall x \in Z, \mathbf{Prob}[x \in Y] = 1/|Z|$：統計的に重要な条件



Z: 条件を満たす全てのデータ



Y: Zからランダムサンプルされたデータ

公平な探索問題の何が難しい？？？

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

要件を満たす集合

Z

= m 個のデータの集合

出力

Y

X から Z を計算

ランダムサンプリング

→ $|Z|$が大きい
→ 少なくとも大きい数のデータにアクセスする必要有
→ 遅い（計算コスト大）

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

Y

公平な探索問題で何を達成した？

内積探索問題 [1]

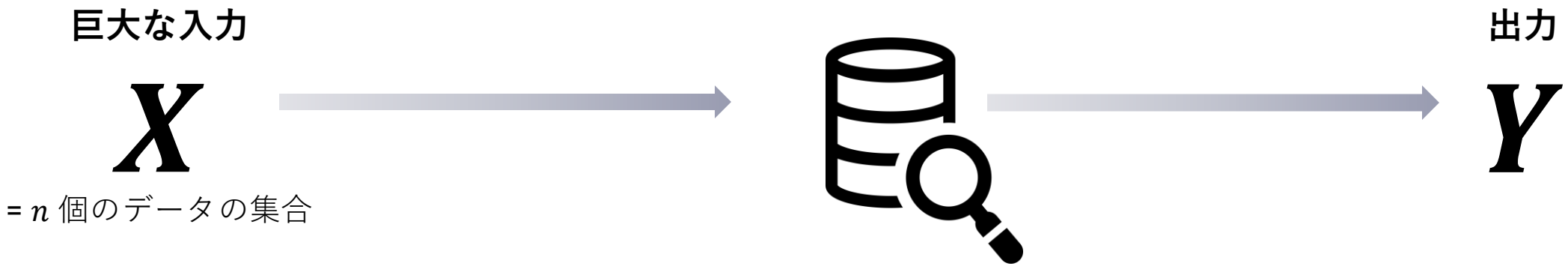
$\{x \mid x \in X,\ x \cdot q \geq \tau\}$ から，ランダムな t 個のデータをサンプルする時間（期待値）：$\log n + t$ に比例

インターバル探索問題 [2]

$\{x \mid x \in X,\ x \cap q \neq \emptyset\}$ から，ランダムな t 個のデータをサンプルする時間：$\log^2 n + t$ に比例

空間ジョイン問題

$\{(x, y) \mid (x, y) \in X \bowtie Y\}$ から，ランダムな t 個のデータをサンプルする時間（期待値）：$n + m + t$ にほぼ比例



[1] K. Aoyama et al., “Simpler is Much Faster: Fair and Independent Inner Product Search,” In *SIGIR*, 2023.
[2] D. Amagata, “Independent Range Sampling on Interval Data,” In *ICDE* 2024.

研究テーマ：データベース，データマイニング，および機械学習のための高速アルゴリズム

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

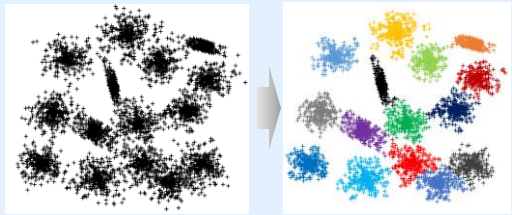
Y

探索

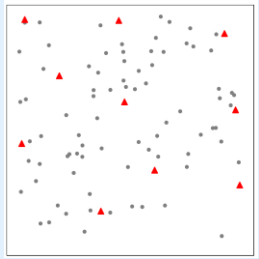
	p_1	p_2	p_3	p_4	p_5
u_1	4.1	**5.0**	2.5	1.8	3.0
u_2	**4.9**	4.3	2.9	4.5	2.1
u_3	2.3	**4.6**	4.1	3.1	1.5
u_4	3.3	**4.1**	3.4	3.6	1.9
u_5	2.1	3.9	2.6	**4.3**	3.1

n Users d × Items m d

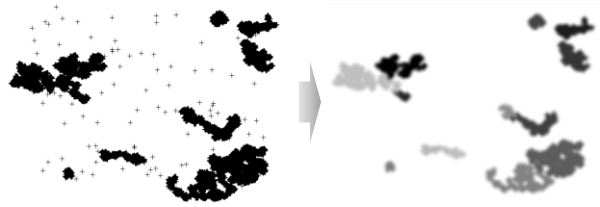
クラスタリング



要約



外れ値検出

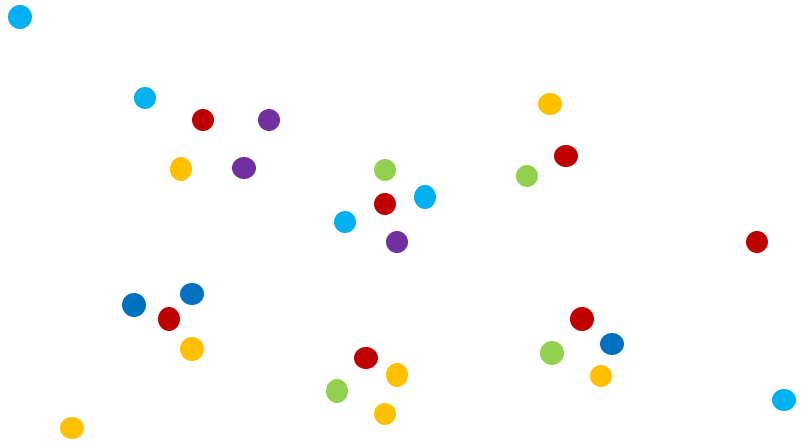


公平性の概念を組み込んだ定式化

クラスタリングにおける「公平性」ってなに？？？

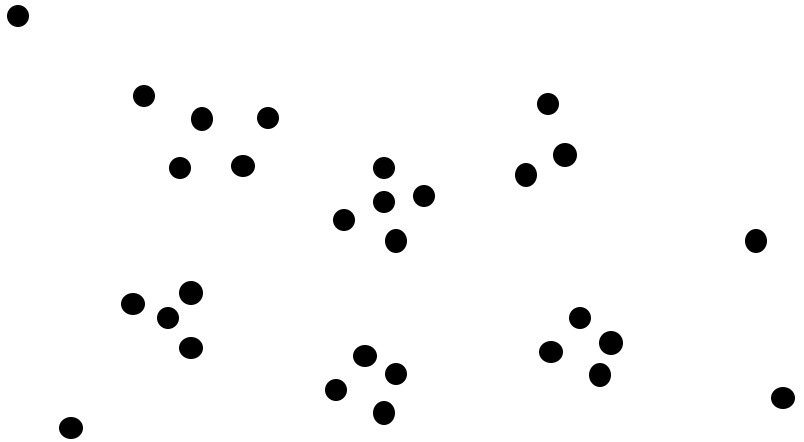
グループ公平性

各データの属性を考慮



個人公平性

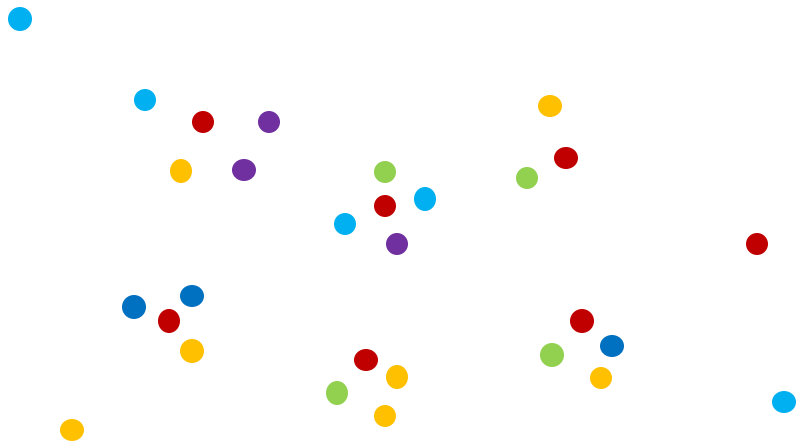
類似したデータ → 類似した結果



クラスタリングにおける「グループ公平性」ってなに？

グループ公平性

各データの属性を考慮



w/o over- & lower-representation

$= "\alpha_j \leq \left|C_i \cap P_j\right| \leq \beta_j"$

→ 各クラスタC_iには属性jのデータがα_j以上β_j以下属する.

→ あるクラスタが特定の属性に独占されないようにする.

Matroid constraint

$= "\left|S \cap P_j\right| = k_j \; \forall j \in [1, m], |S| = k"$

→ クラスタ中心には属性jからk_j個選ばれる.

→ クラスタ中心が特定の属性に独占されないようにする.

Social fairness

$= \min \max_m \frac{\Delta(C, P_i)}{|P_i|}$

→ 各属性に対して遠いクラスタセンタを抑制

→ クラスタ中心が特定の属性に対して遠い，を防ぐ.

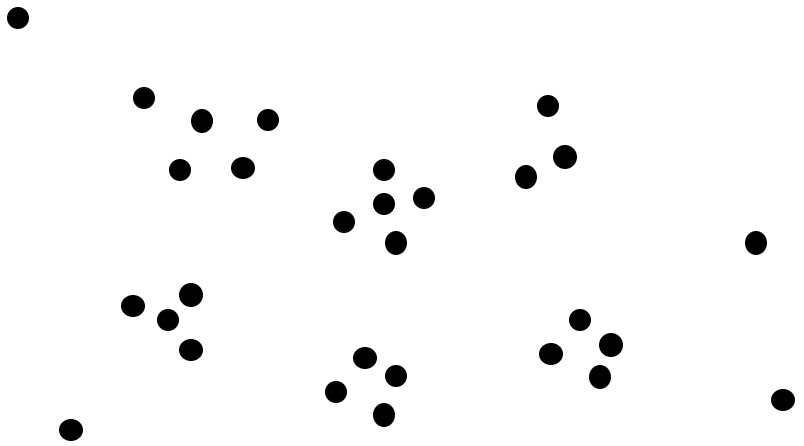
クラスタリングにおける「個人公平性」ってなに？？？

定義

各データに対して、クラスタセンタが自身のn/k-最近傍に含まれている.

個人公平性

類似したデータ → 類似した結果



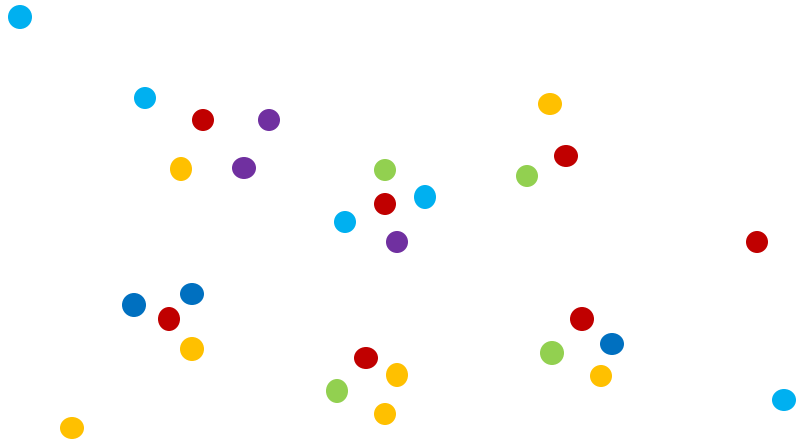
公平なクラスタリングの何が難しい？

k-クラスタリング ：$S^* = \arg\min_{|S|=k} f(P,S)$

S はクラスタセンタの集合であり，

S^*の計算はNP困難（多項式時間では解が得られない）

→ S^*っぽい解を高速に探したい



w/o over- & lower-representation

= "$\alpha_j \leq |C_i \cap P_j| \leq \beta_j$"

→ 各クラスタC_iには属性jのデータがα_j以上β_j以下属する．

→ あるクラスタが特定の属性に独占されないようにする．

Matroid constraint

= "$|S \cap P_j| = k_j \ \forall j \in [1,m], |S| = k$"

→ クラスタ中心には属性jからk_j個選ばれる．

→ クラスタ中心が特定の属性に独占されないようにする．

Social fairness

= $\min \max_m \frac{\Delta(C,P_i)}{|P_i|}$

→ 各属性に対して遠いクラスタセンタを抑制

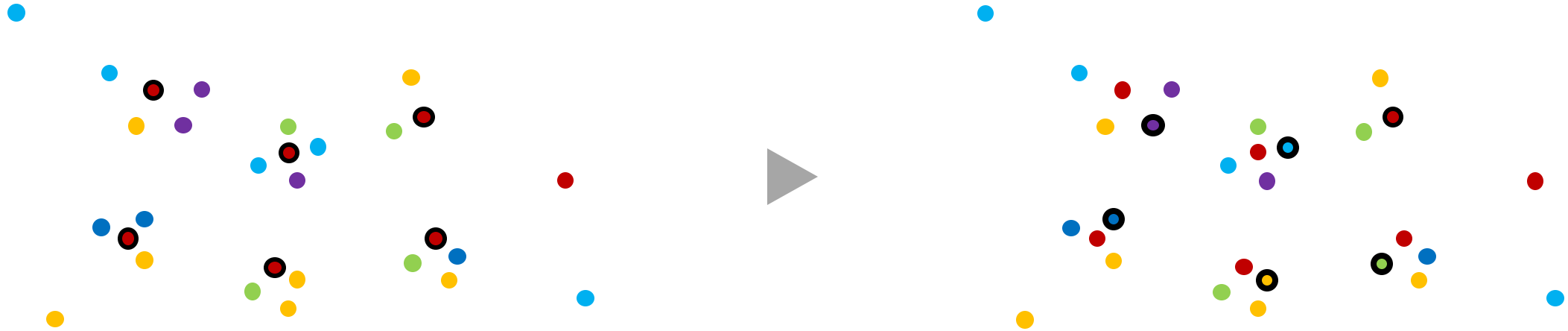
→ クラスタ中心が特定の属性に対して遠い，を防ぐ．

公平なクラスタリング問題で何を達成した？

Fair k-center clustering with outliers [3]

$$S^* = \mathrm{argmin}_{S \subseteq P \setminus P_{out}, |S|=k, \forall i \in [1,m], |S \cap P_i| \leq k_i} \max_{p \in P \setminus P_{out}} dist(p, S)$$

There is an $O(nk)$ time algorithm that needs $O(n)$ space and yields a $(3+\gamma)$-approximation result for the fair $(k, (1+\epsilon)z)$-center clustering problem with probability at least $\left(1 - \frac{z}{n}\right)\left(\frac{\epsilon}{1+\epsilon}\right)^{k-1}$, by returning exactly k centers and removing at most $(1+\epsilon)z$ points.



[3] D. Amagata, “Fair k-center Clustering with Outliers,” In *AISTATS* 2024.

まとめ：この分野における「公平性」とは

巨大な入力

X

= n 個のデータの集合

～2010年代前半：n = million は大きい
2010年代後半～：n = billion も珍しくない

出力

Y

公平性 ≒ 平等

探索における公平性：解に入る確率が平等（同じ）

クラスタリングにおける公平性：グループ公平性（属性が平等に扱われる）と個人公平性（似たデータがクラスタセンタ）

共同研究っぽい何かは絶賛大募集

巨大な入力

X

$= n$ 個のデータの集合

公平な出力

Y

提供できます！

- データ処理関係で困ってることの解決

是非教えて下さい！

- 実はこんな公平性が求められています、的な話
- 〇〇な処理は不公平な結果を出しがち、的な話