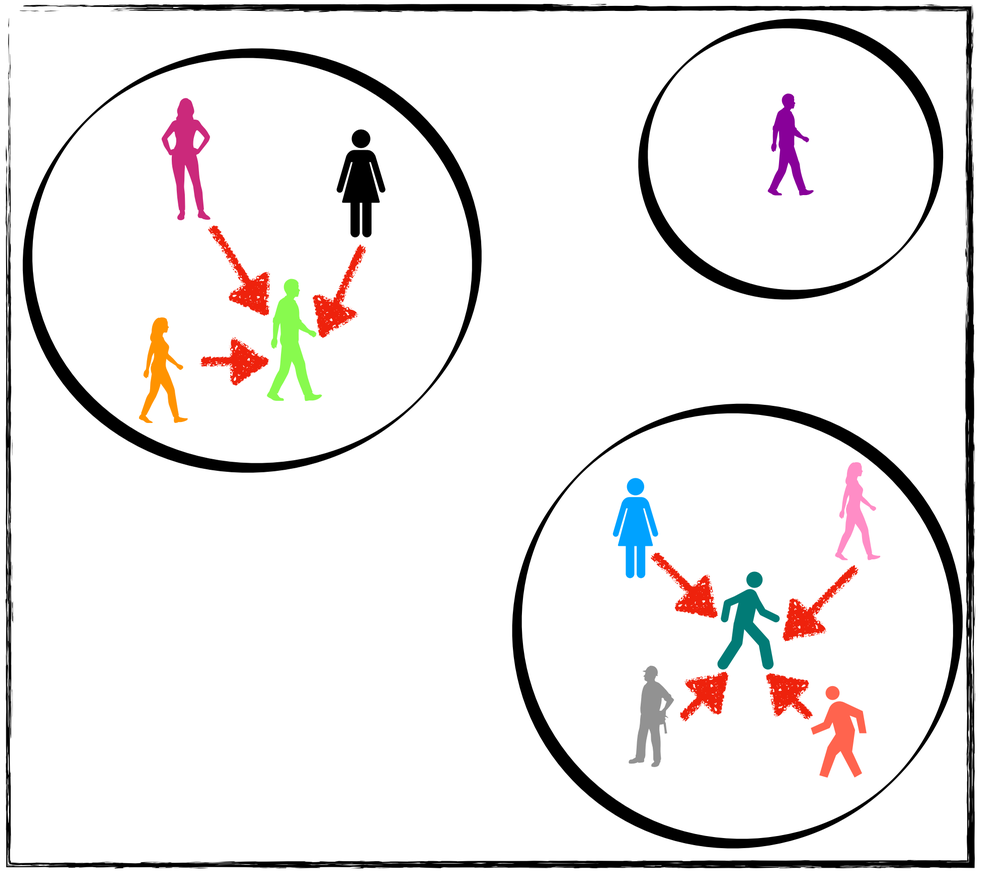
Diversity-aware k-median:
Clustering with fair centre representation

Suhas Thejaswi[1], Bruno Ordozgoiti[1], Aristides Gionis[1,2]

[1] Department of Computer Science,
Aalto University, Finland

[2] Division of Theoretical Computer Science,
KTH Royal Institute of Technology, Sweden

Motivation

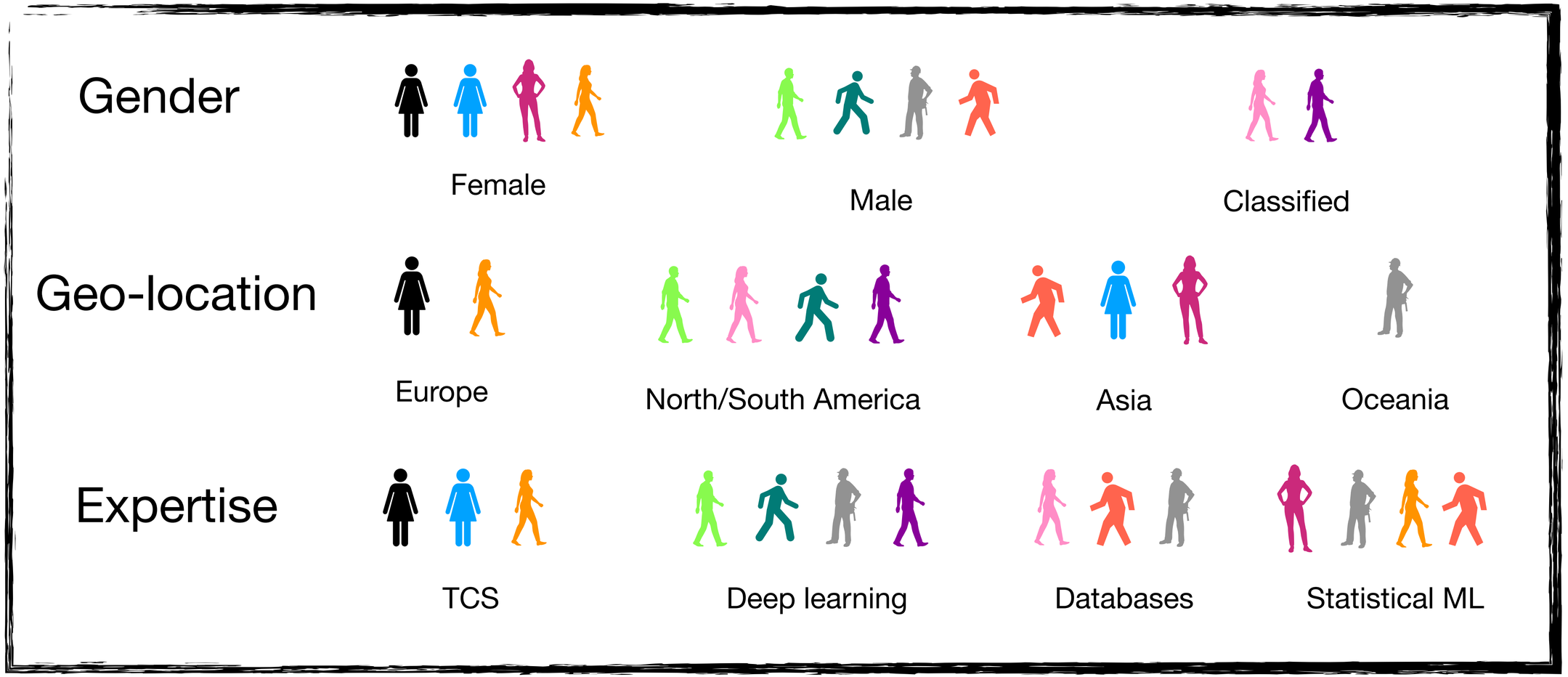


- **task**: form a committee of three representatives over a set of individuals
- **distance** — value measuring the strength of ties between individuals, for example, agreement over a set of issues
- **possible solution**: form three clusters that reduces the total distance between selected representatives (cluster centres) and individuals

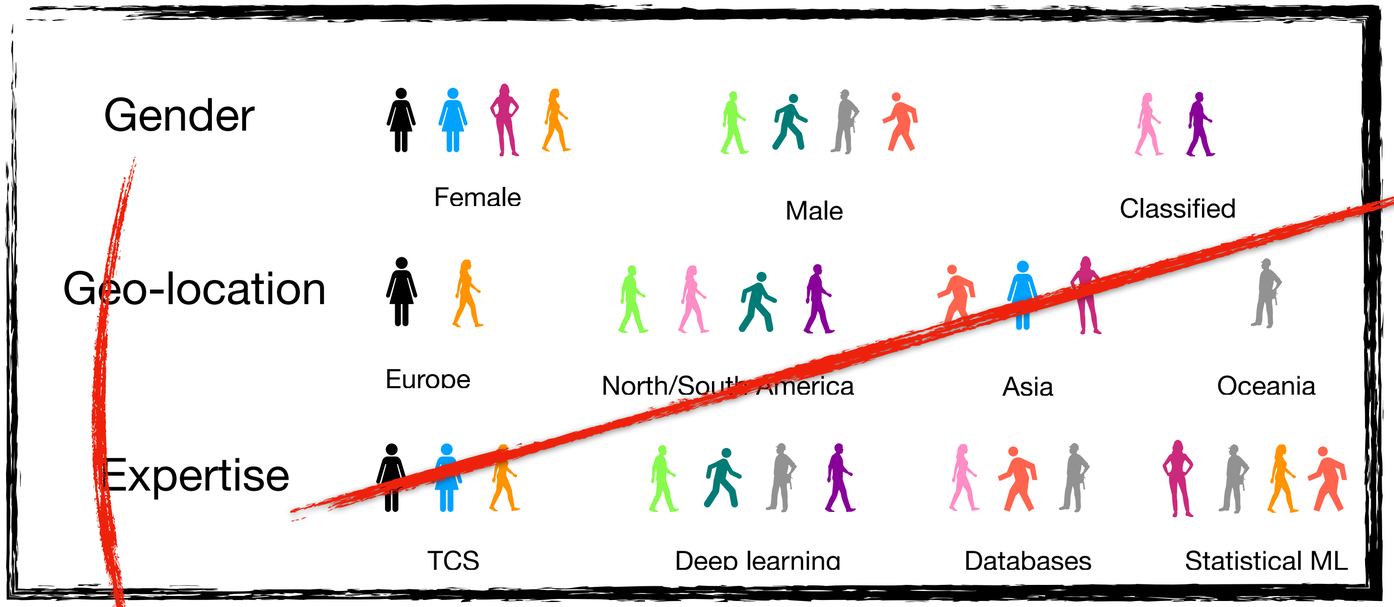
chosen committee:

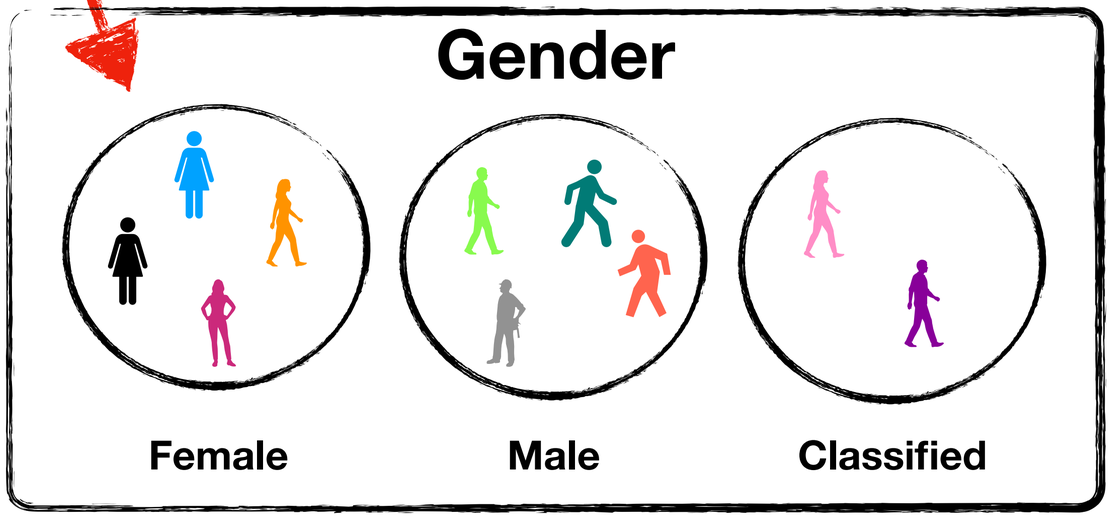
Is the solution diverse?

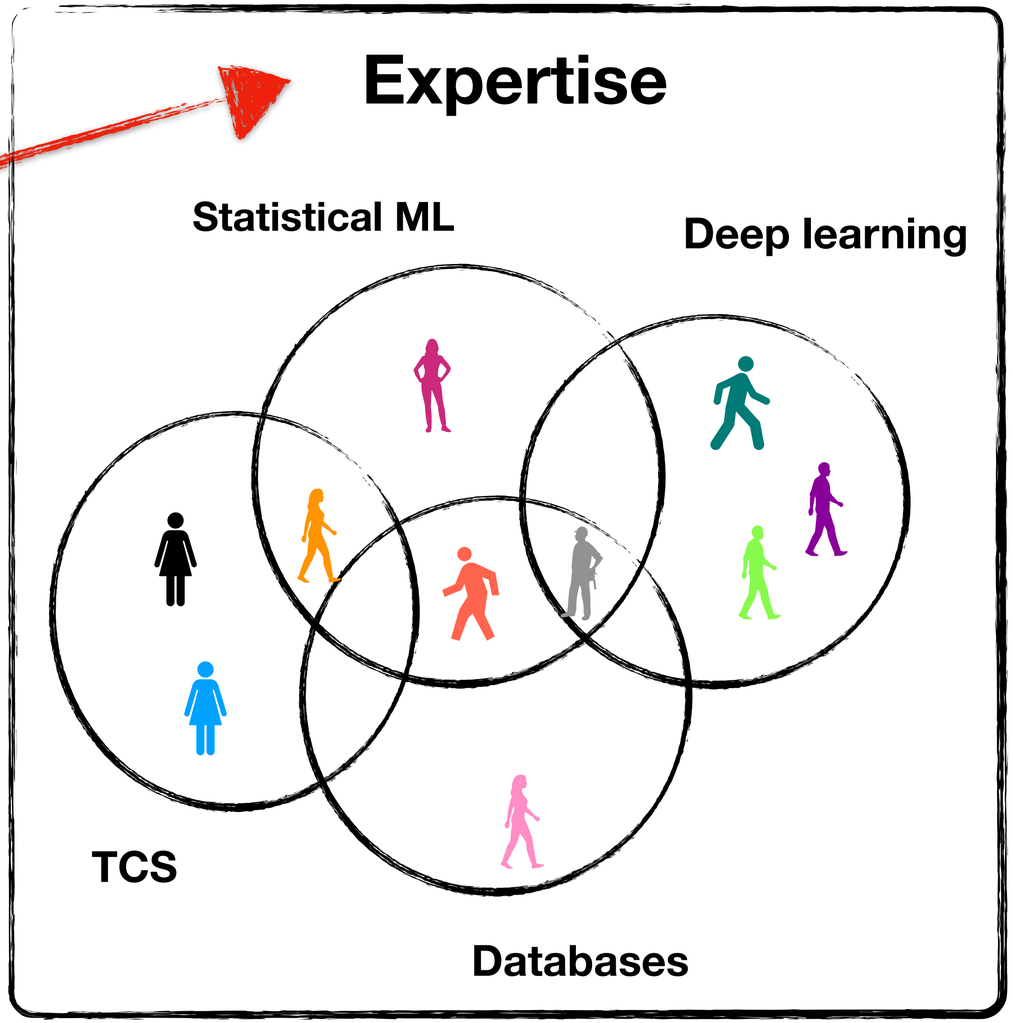
Motivation


Gender
Female
Male
Classified
Geo-location
Europe
North/South America
Asia
Oceania
Expertise
TCS
Deep learning
Databases
Statistical ML

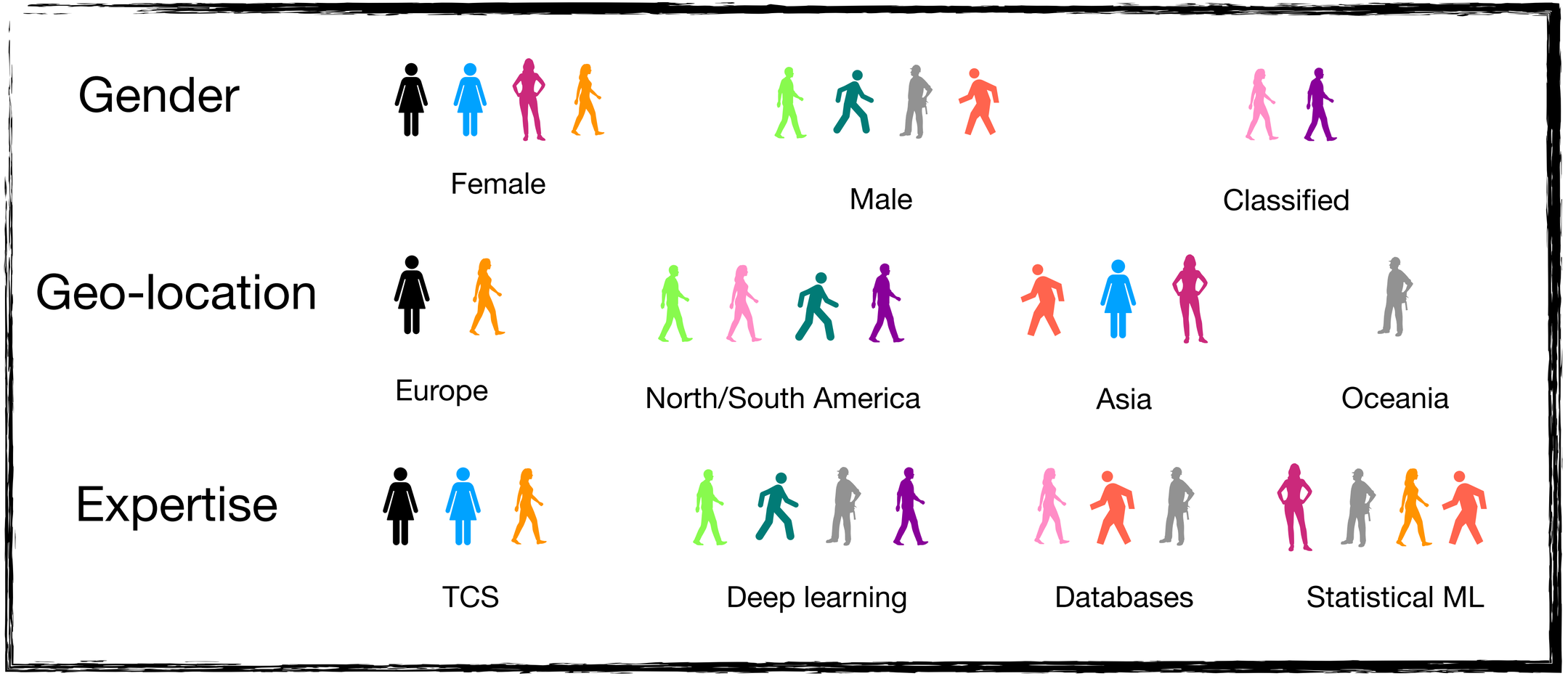
Motivation


Gender
Female
Male
Classified
Geo-location
Europe
North/South America
Asia
Oceania
Expertise
TCS
Deep learning
Databases
Statistical ML


Gender
Female
Male
Classified


Expertise
Statistical ML
Deep learning
TCS
Databases

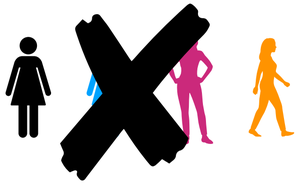
Motivation



Is each group represented in the committee?

Motivation

Gender



Female

Male

Classified

Geo-location

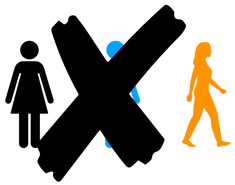
Europe

North/South America

Asia

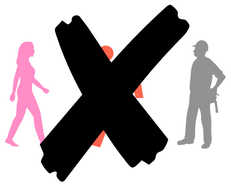
Oceania

Expertise



TCS

Deep learning



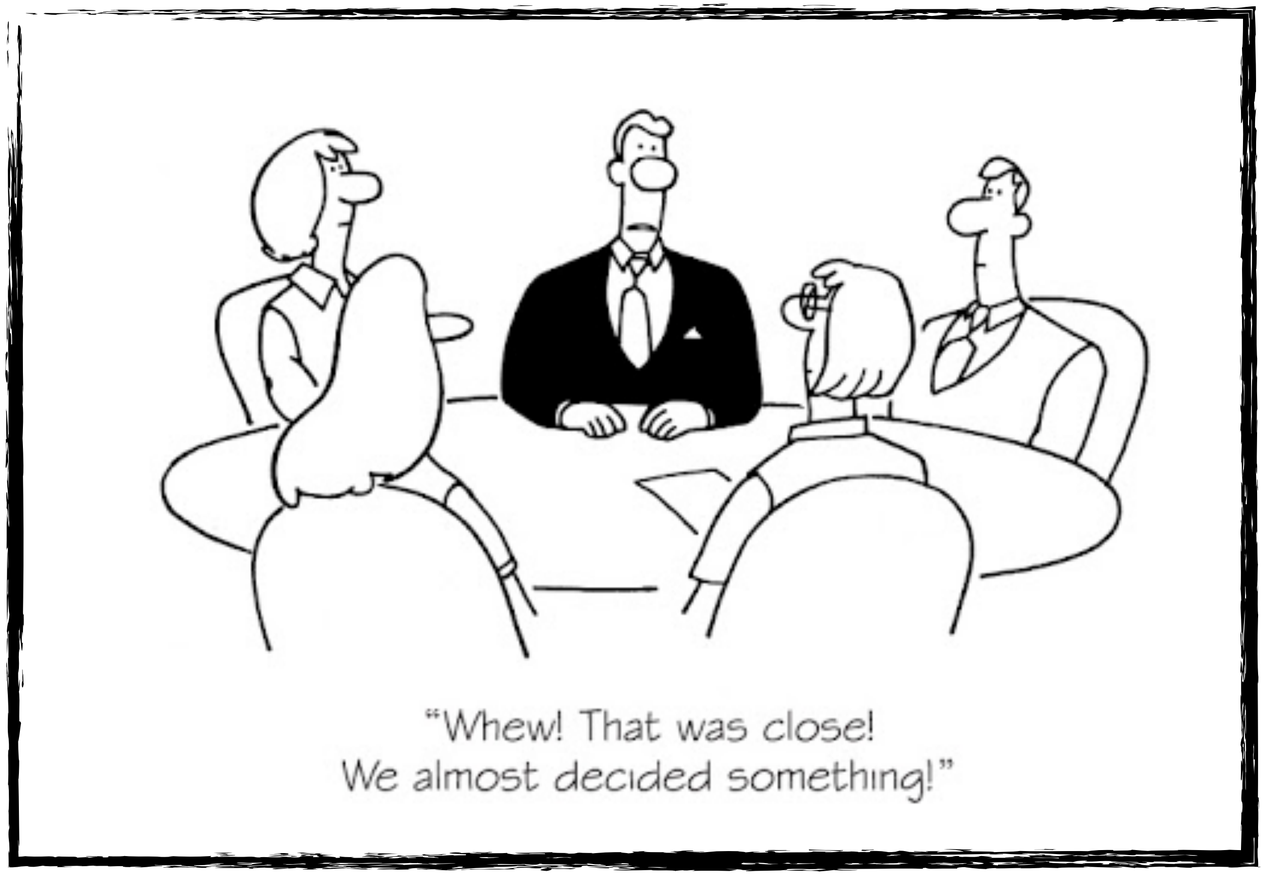
Databases

Statistical ML

chosen solution:

(naive clustering)

Motivation

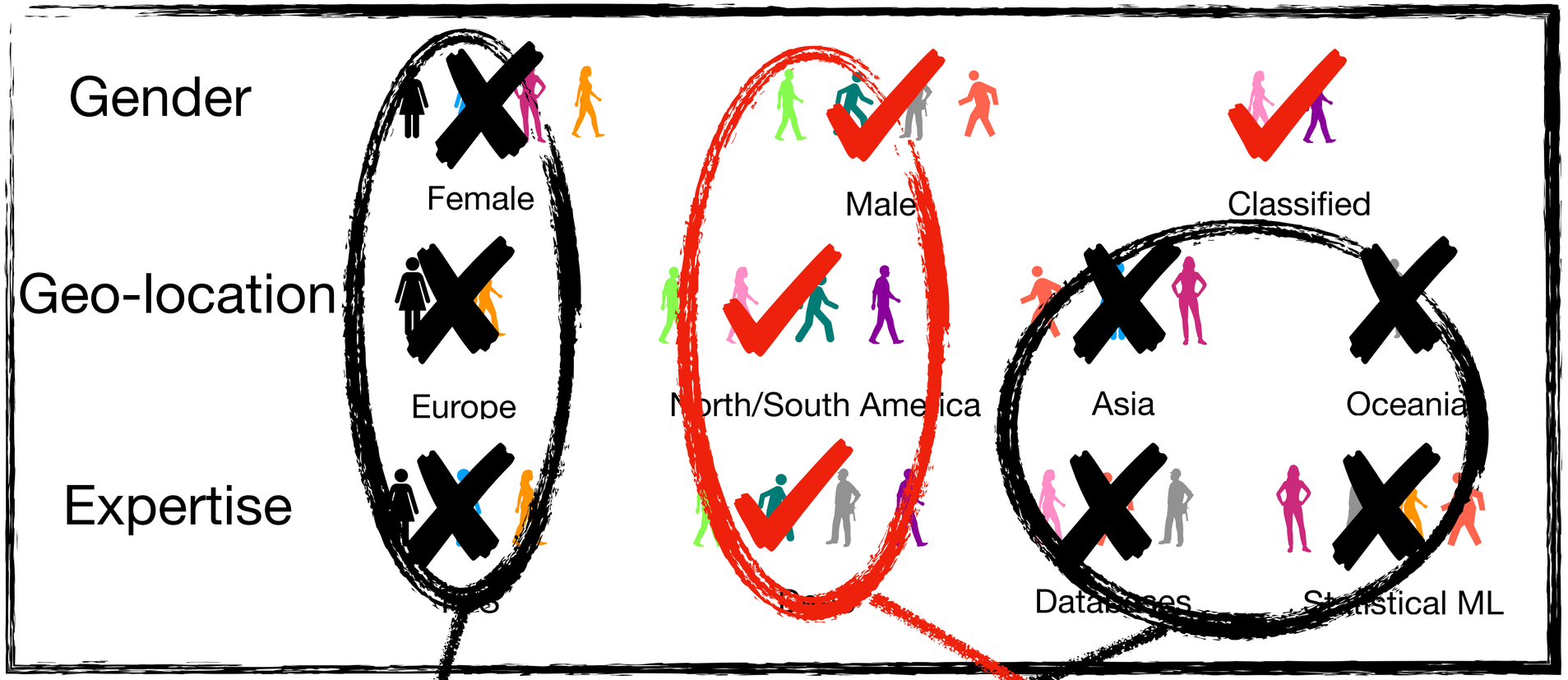


task: form a diverse-committee such that

- experts from different fields
- representation of genders
- different geographical location
- minimise clustering cost, as measured by k-median clustering objective

Image source: clipartsuggest.com

Two sides of the problem



chosen solution:

(naive clustering)

under-represented

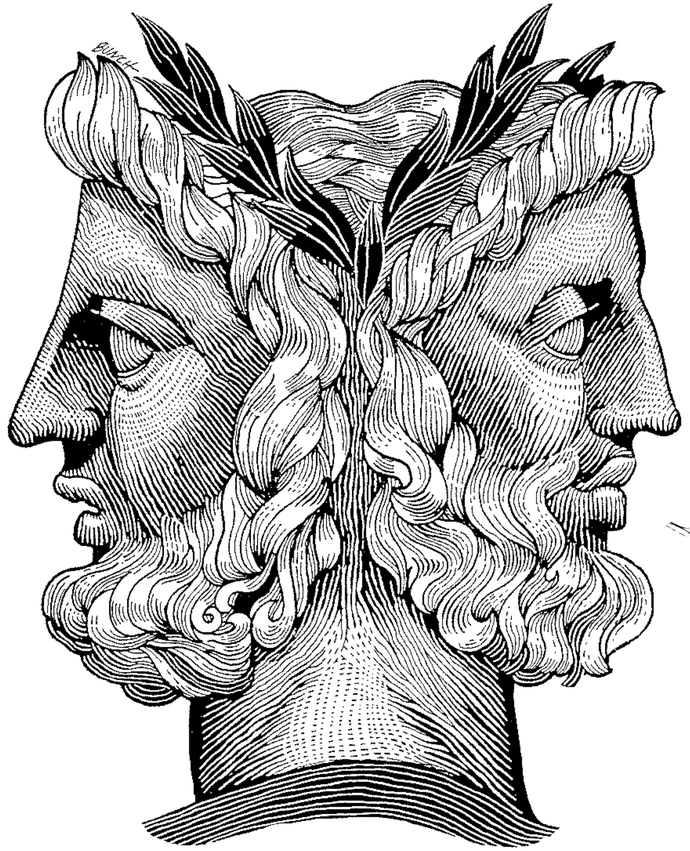
over-represented

Two sides of the problem

Avoid over-representation

Avoid under-representation

Red-blue median problem
Hajiaghayi et al. 2010, 2012



This paper

Matroid-median problem
Krishnaswamy et al. 2010
Chen et al. 2016

The problems are mathematically different and lead to problems with different complexity

Are these problems the same?

Image source: pinterest.com

Diversity-aware k-median

- given
  a set of clients C and a set of facilities $F \subseteq C$,
  a distance function $d : C \times F \to R_+$,
  a collection $\mathcal{F} = \{F_1, \ldots, F_t\}$ of facility groups $F_i \subseteq F$,
  a set $R = \{r_1, \ldots, r_t\}$ of lower-bound thresholds,
  an integer $k \leq |F|$

- find a subset $S \subseteq F$ of facilities such that
  size $|S| = k$,
  constraint $|S \cap F_i| \geq r_i$ is satisfied for all $i \in [t]$,
  minimises the cost function $cost(S) = \sum_{c \in C} \min_{s \in S} d(c, s)$

avoid under-representation
$|S \cap F_i| \geq r_i$

avoid over-representation
$|S \cap F_i| \leq r_i$
red-blue median (Hajiaghayi et al.)
matroid median (Krishnaswamy et al.)

k-median clustering with lower-bound constraints

Hardness results (Diversity-aware k-median)

- **NP-hardness**
  - NP-hard to find a feasible solution
  - NP-hard to find optimal solution

- **inapproximability results**
  - inapproximable to any multiplicative factor
  - inapproximable even if the underlying distance is tree metric
  - inapproximable even all facility groups are size two

- **fixed parameter intractability (FPI)**
  - FPI with respect to parameter k, size of the solution sought

Tractable cases (Diversity-aware k-median)

Problem	**NP**-hard	FPT(k)	Approx. factor	Approx. method
Intractable case: intersecting facility groups				
General variant	✓	✗	inapproximable	
Tractable cases: disjoint facility groups				
$t > 2$, $\sum_{i \in [t]} r_i = k$	✓	open	8	LP
$t > 2$, $\sum_{i \in [t]} r_i < k$	✓	open	8	$\mathcal{O}(k^{t-1})$ calls to LP
$t = 2$, $r_1 + r_2 = k$	✓	open	$3 + \epsilon$	local search
$t = 2$, $r_1 + r_2 < k$	✓	open	$3 + \epsilon$	$\mathcal{O}(k)$ calls to local search

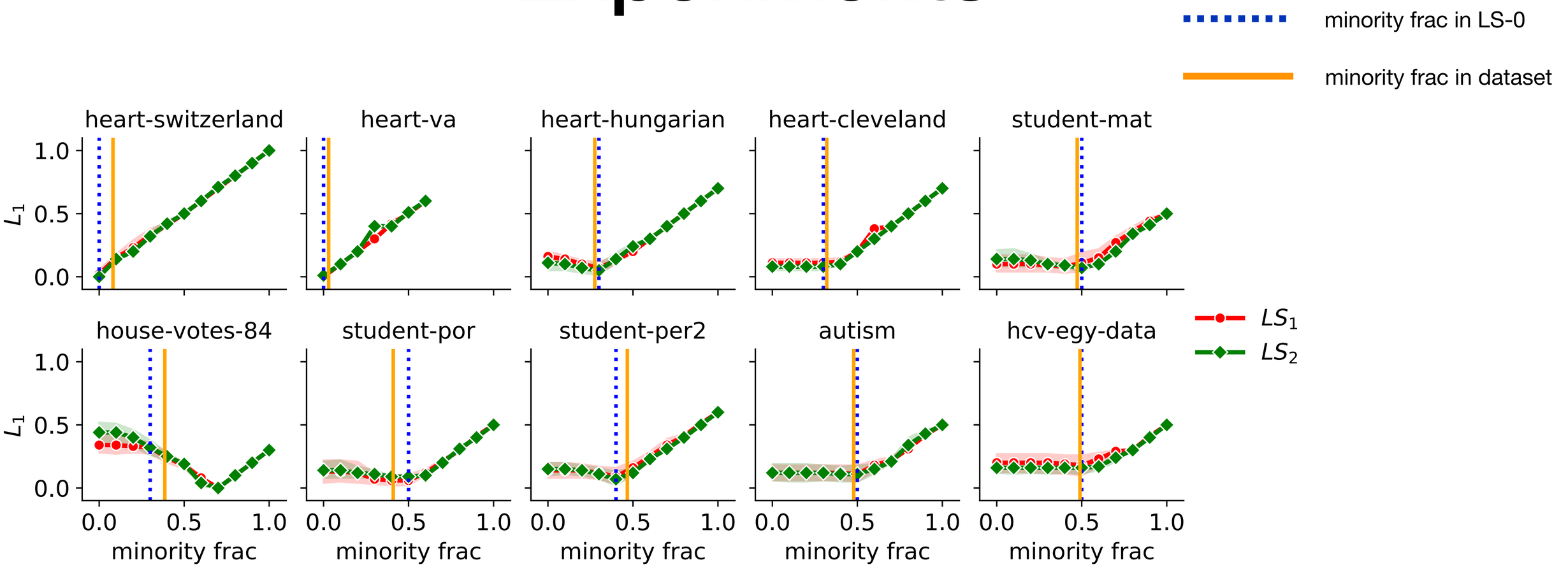
Experiments

- **datasets :** UCI machine learning repository
- **baseline (LS-0) :** local-search with no constraints
- **local-search with constraints (proposed scalable solutions)**
  - LS-1 : single swap local search with constraints
  - LS-2 : multi-swap local search with constraints
- **minority fraction :** ratio of smallest group in the dataset
- **price of diversity (POD) :** ratio of increase in the cost of a solution

$$POD(LS-1) = \frac{l_1 - l_0}{l_0} \qquad POD(LS-2) = \frac{l_2 - l_0}{l_0}$$

l_0 : cost of solution from LS-0
l_1 : cost of solution from LS-1
l_2 : cost of solution from LS-2

Experiments



groups : male, female (non-intersecting)

size of solution (k) : 10

minority fraction : fraction of facilities in the solution that belong to minority group

Conclusions

- a novel way of introducing fairness in clustering problems
- complexity results follow to most clustering formulations with under-representation constraints
- future work: introduce under-representation constraints that are approximable to a multiplicative factor
- future work: fixed parameter algorithms using other parameters

Diversity-aware k-median: Clustering with fair center representation*

Suhas Thejaswi[1], Bruno Ordozgoiti[1], and Aristides Gionis[1,2]

[1] Aalto University, Espoo, Finland.
su.thejaswi@gmail.com, bruno.ordozgoiti@aalto.fi
[2] KTH Royal Institute of Technology, Stockholm, Sweden.
argioni@kth.se

Abstract. We introduce a novel problem for diversity-aware clustering. We assume that the potential cluster centers belong to a set of groups defined by protected attributes, such as ethnicity, gender, etc. We then ask to find a minimum-cost clustering of the data into k clusters so that a specified minimum number of cluster centers are chosen from each group. We thus require that all groups are represented in the clustering solution as cluster centers, according to specified requirements. More precisely, we are given a set of clients C, a set of facilities $\mathcal{F}$, a collection $\mathscr{F} = \{F_1, \ldots, F_t\}$ of facility groups $F_i \subseteq \mathcal{F}$, a budget k, and a set of lower-bound thresholds $R = \{r_1, \ldots, r_t\}$, one for each group in $\mathscr{F}$. The *diversity-aware k-median problem* asks to find a set S of k facilities in $\mathcal{F}$ such that $|S \cap F_i| \geq r_i$, that is, at least r_i centers in S are from group F_i, and the k-median cost $\sum_{c \in C} \min_{s \in S} d(c, s)$ is minimized. We show that in the general case where the facility groups may overlap, the diversity-aware k-median problem is **NP**-hard, fixed-parameter intractable with respect to parameter k, and inapproximable to any multiplicative factor. On the other hand, when the facility groups are disjoint, approximation algorithms can be obtained by reduction to the *matroid median* and *red-blue median* problems. Experimentally, we evaluate our approximation methods for the tractable cases, and present a relaxation-based heuristic for the theoretically intractable case, which can provide high-quality and efficient solutions for real-world datasets.

Keywords: Algorithmic bias · Algorithmic fairness · Diversity-aware clustering · Fair clustering.

Thank you

source code : github.com/suhastheju/diversity-aware-k-median