Data-Driven Similarity-based Worker Recruitment Towards Multi-task Data Inference for Sparse Mobile Crowdsensing

En Wang*, Zijie Tian*, Yongjian Yang*, Wenbin Liu*, Baoju Li*, Nan Jiang‡ and Jie Wu§

*Jinlin University, ‡East China Jiaotong University , § Temple University

Outline

I. Background

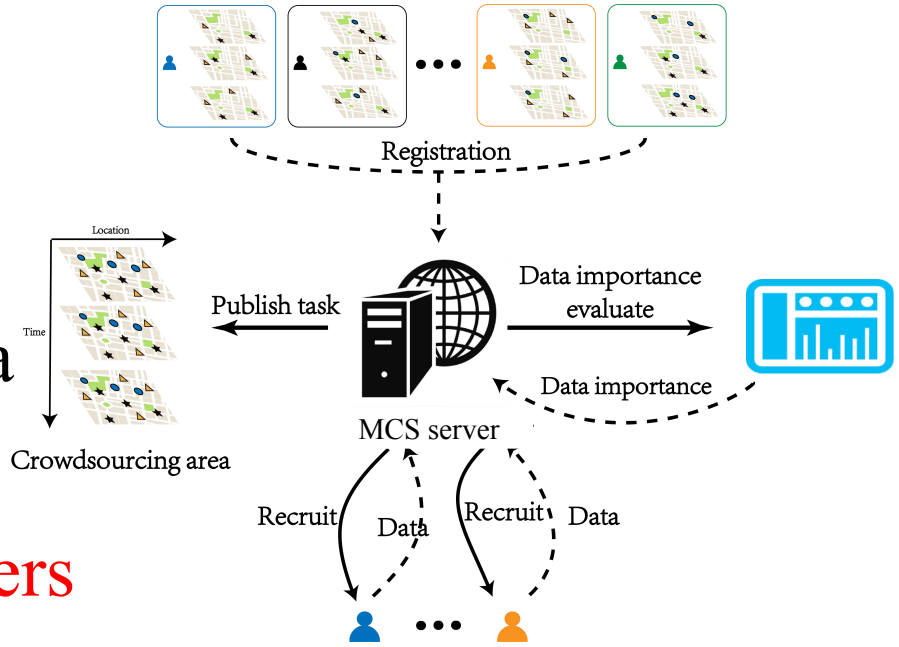
II. Challenge

III. Method

IV. Experiments

Background

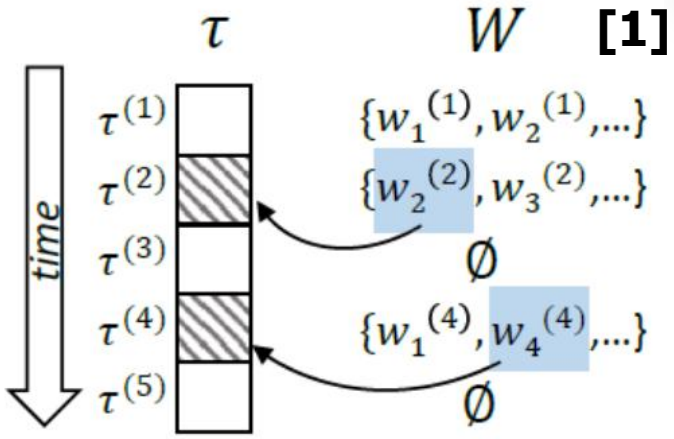
- Crowdsensing recruits workers to collect various data
- The limitation of budget & workers
- Inferring unsensed areas using a little sensed data
- Finding important data > Recruiting useful workers



Background

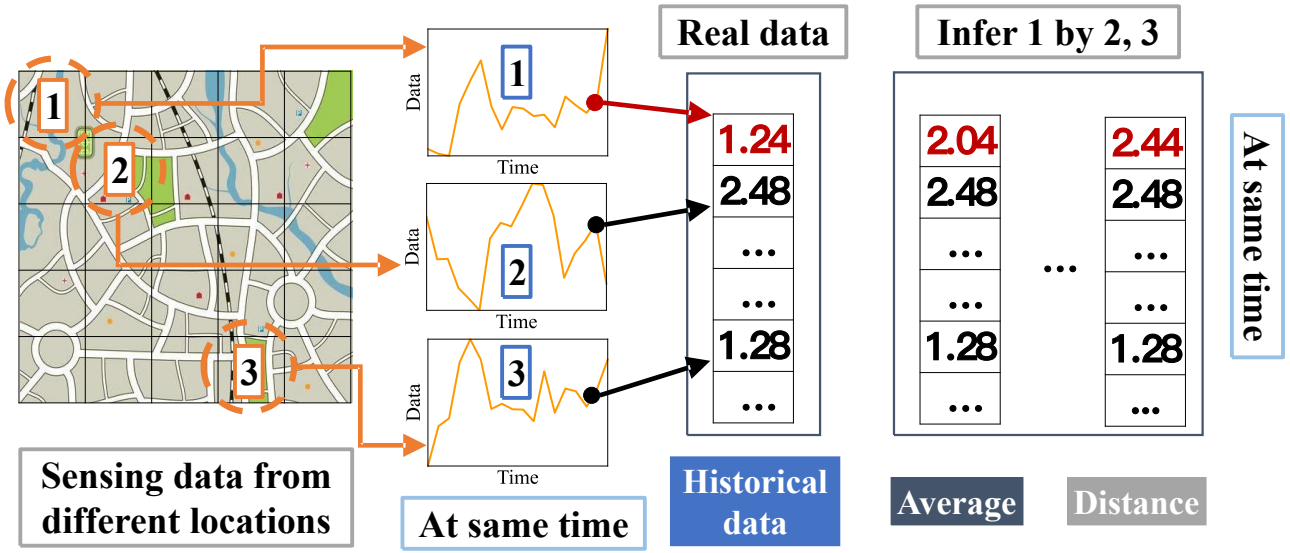
Existing work:

- Spatial-temporal distance between sensed data.
- Using entropy to evaluate importance and select workers.



Disadvantage:

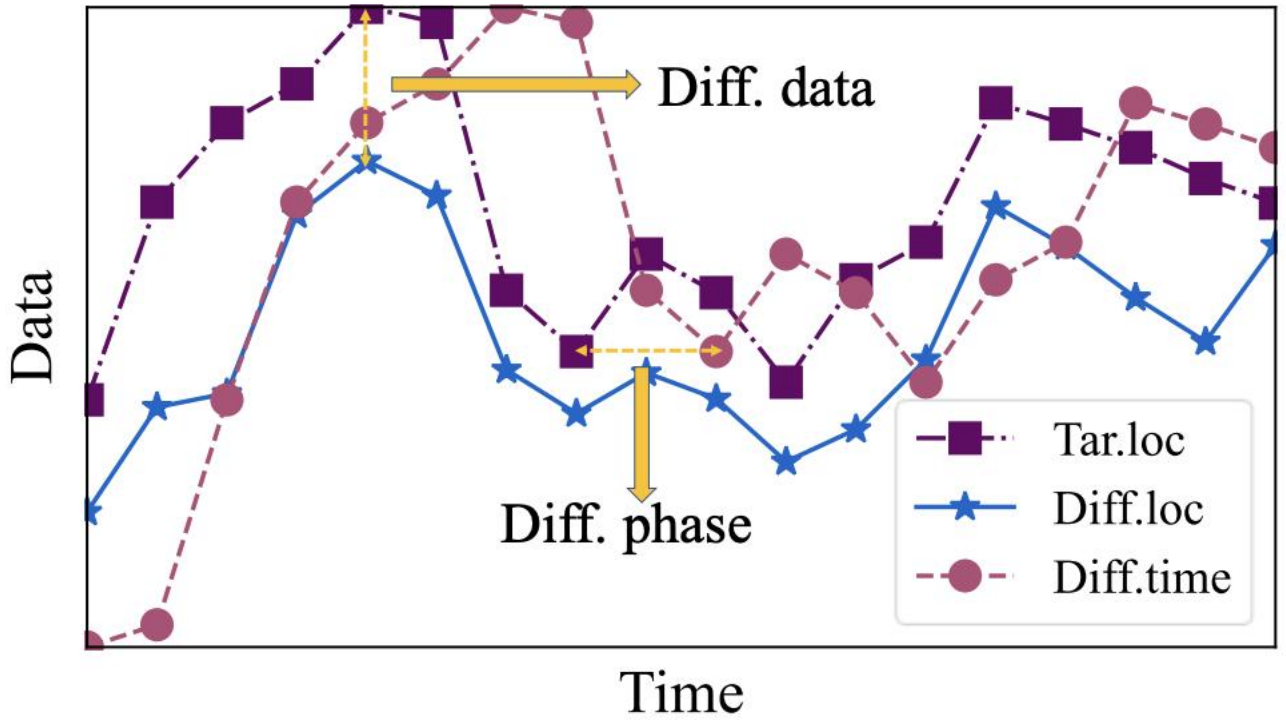
- Distance vs. Historical data knowledge



[1] On Efficient and Scalable Time-Continuous Spatial Crowdsourcing

Challenge

- Diff. data & Diff. phase
- Data reliability



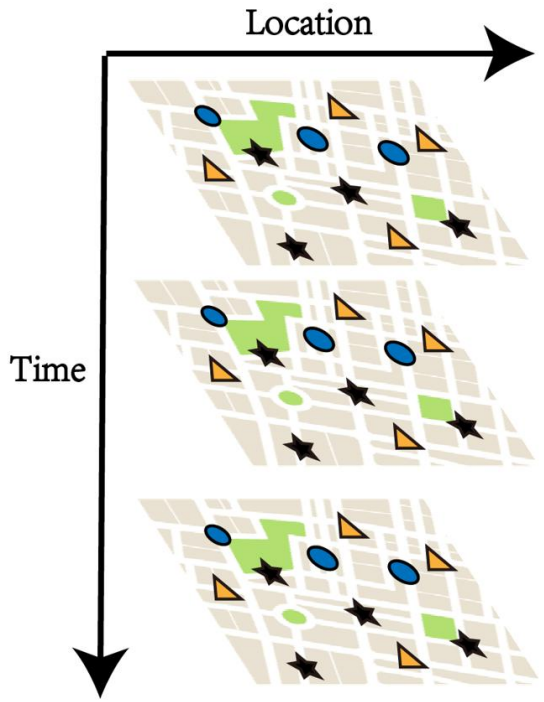
- How to calculate similarity according to the historical data?

Challenge

- Multiple data-types in practice

	LVTH			
	Light	**Voltage**	**Humidity**	**Temperature**
Country/Region - City	Intel Berkeley Research Lab - California			
Subarea	33 subareas			
Time slots	384 time slots each with half an hour			
Mean ± Std.	546.96 ± 649.31 (lux)	2.62 ± 0.08 (v)	35.74 ± 7.06 (%)	22.68 ± 7.08 (°C)

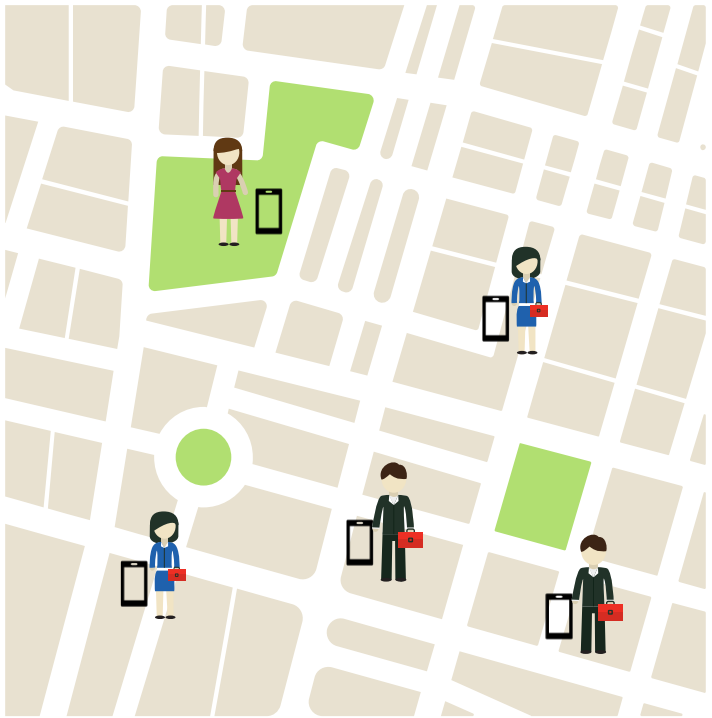
	PM		HT	
	PM2.5	**PM10**	**Humidity**	**Temperature**
Country/Region - City	Chinese mainland - Beijing		Switzerland - Lausanne	
Subarea	36 subareas		57 subareas	
Time slots	264 time slots each with an hour		336 time slots each with half an hour	
Mean ± Std.	79.11 ± 81.21 ($\mu g/m^3$)	63.12 ± 48.56 ($\mu g/m^3$)	84.52 ± 6.32 (%)	6.04 ± 1.87 (°C)

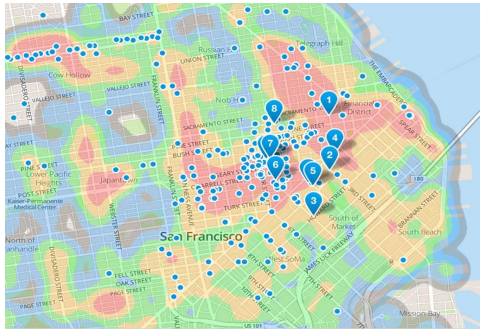


- How to measure similarity and use it to evaluate data importance in three dimensional (spatial, temporal and data-type) ?

Challenge

- Worker unreliability[2]
- Multiple equipped sensors
- Platform budget limitation





Thermometer

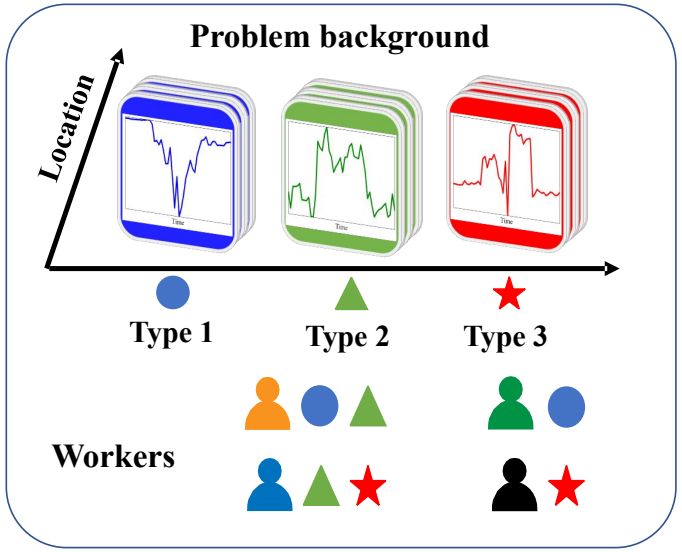
Air Quality Sensor

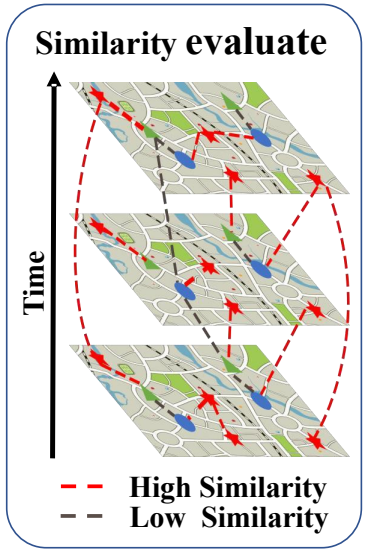
.....

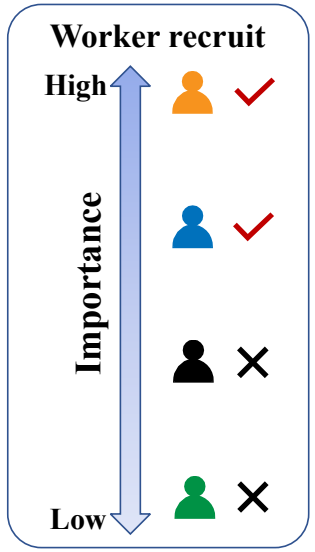
- How to select a group of suitable workers?

[2] Reliable Diversity-Based Spatial Crowdsourcing by Moving Workers

Method







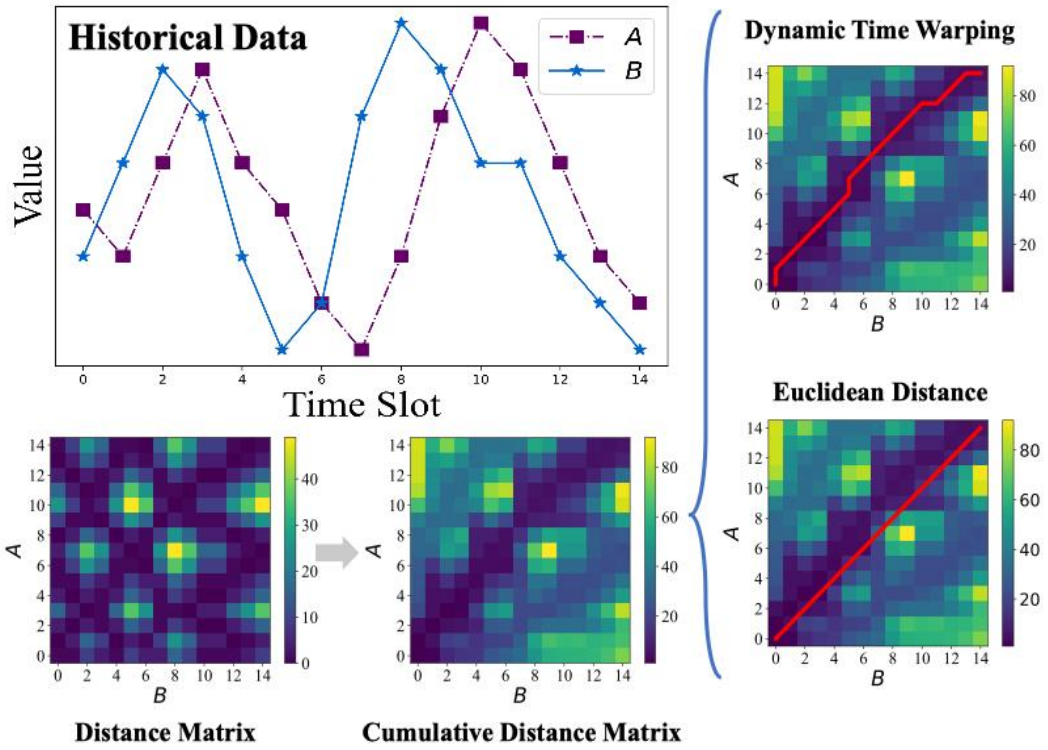
- SWDTW to evaluate time-series similarity
- Entropy-Weighted method to calculate similarity in three dimensional
- WRGSA to recruit workers according reliability and equipped sensors

Method

- Similarity-Weighted Dynamic Time Warping (SWDTW)
  - Dynamic Time Warping (DTW)

Euclidean Distance

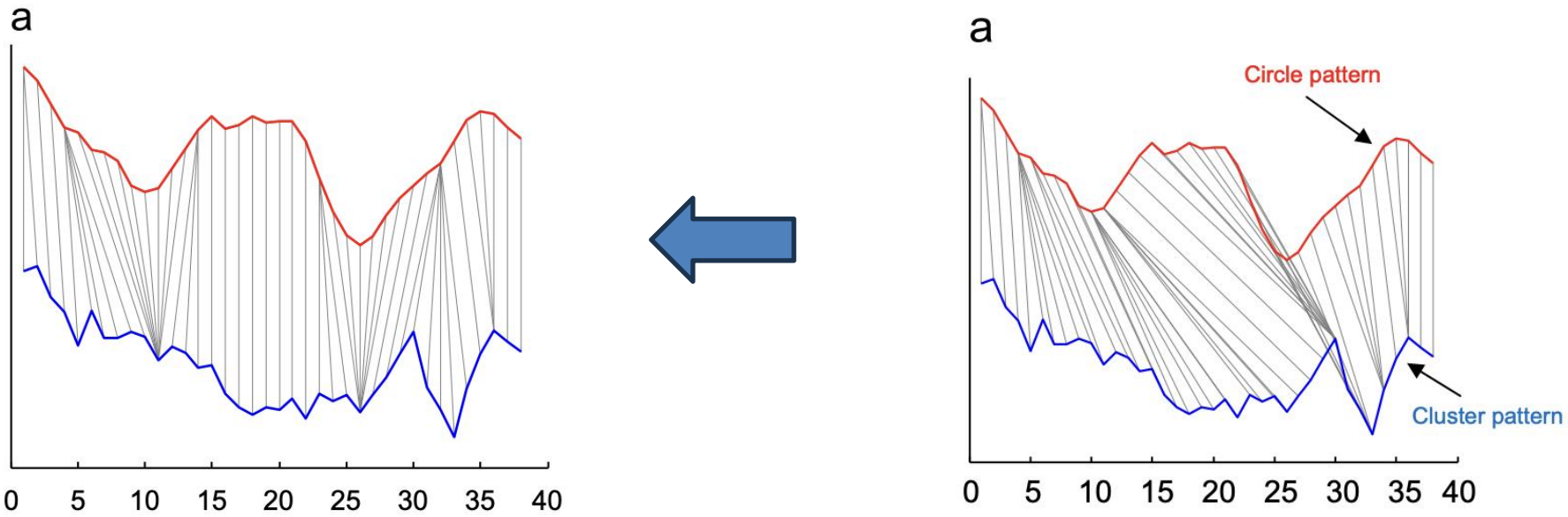
$$\Upsilon_{n_1,n_2} = dis_{n_1,n_2} + \min\{\Upsilon_{m,n-1}, \Upsilon_{m-1,n}, \Upsilon_{m-1,n-1}\}$$



Method

[3]

- Similarity-Weighted Dynamic Time Warping (SWDTW)
  - Weighted Dynamic Time Warping (WDTW)



$$\mathrm{dis}_{n_1,n_2} = \frac{\omega_{max}}{1+\exp(-g(|n_1-n_2|-m_c))} \times (a_{n_1}-b_{n_2})^2$$

[3] Weighted dynamic time warping for time series classification

Method

- 三维空间相似性

$$\alpha + \beta + \gamma = 1$$

$$\mathcal{G}_{i,j} = \alpha \cdot \mathcal{G}_{m_{x_i}, m_{x_j}} + \beta \cdot \mathcal{G}_{t_{x_i}, t_{x_j}} + \gamma \cdot \mathcal{G}_{l_{x_i}, l_{x_j}}$$

Data-type Similarity　　Time Similarity　　Location Similarity

Method

- 子任务选取

$$err\left(x_i\right)=\frac{\sum_{j\in S_k(x_i)}\left(1-\mathcal{G}_{i,j}\right)}{k}$$

$$kr_i=\frac{1}{|X|}\left(1-err\left(x_i\right)\right)$$ The knowledge of x_i for inferring. $[0, \frac{1}{|X|}]$

$$Q(X|\zeta)=-\sum_{\forall i\in X} kr_i \log_2 kr_i$$ The data quality[1] of X with selected datapoints[1].

For $x \in X-\zeta$:

$$\triangle_x=Q(X|\zeta+x)-Q(X|\zeta)$$

[1] On Efficient and Scalable Time-Continuous Spatial Crowdsourcing

Method

- 参与者招募
  - 参与者集合: $w \in W = (p_w, c_w, \mathbf{S}_w)$
  - 子任务可靠性: $rel(x_i, W_{x_i}) = 1 - \prod_{\forall w_j \in W_{x_j}} (1 - p_{w_j})$

$$d'_{i,j} = (1 - \mathcal{G}i, j) \cdot (1 - rel(j, W_j))$$

$$err(x_i) = \frac{\sum_{j \in S'_k(x_i)} d'_{i,j}}{k}$$

$$kr_i = \frac{1}{|X|} \left(\frac{\sum_{j \in S'_k(x_i)} rel\,(j, W_j)}{k} - err\,(x_i) \right)$$

Method

- Worker Recruitment

Algorithm 3 WRGSA

Input: $B, W, X, T, T_{max}, \alpha$;
Output: ε : the set of recruited workers
$R = \{rel(x_1, W_{x_1}), rel(x_2, W_{x_2}), ..., rel(x_n, W_{x_n})\}$: the reliabilities of all datapoints

Initialize ε^* by algorithm 2
while stop condition not met **do**
  generate a new solution ε' from ε^* by Algorithm 4
  if $Q(X|\varepsilon') > Q(X|\varepsilon^*)$ **then**
    $\varepsilon^* \leftarrow \varepsilon'$
  else
    $\varepsilon^* \leftarrow \varepsilon'$ with probability $\mathcal{J}_{\varepsilon^*, \varepsilon'}$
  end if
  if $Q(X|\varepsilon^*) > Q(X|\varepsilon)$ **then**
    $\varepsilon \leftarrow \varepsilon^*, T_b \leftarrow T$
    calculate R according Eq.(1)
  end if
  $T \leftarrow \alpha \times T$
  if $T < 0.01$ **then**
    $T_b \leftarrow 2 \times T_b, T \leftarrow \min\{T_b, T_{\max}\}$
  end if
end while

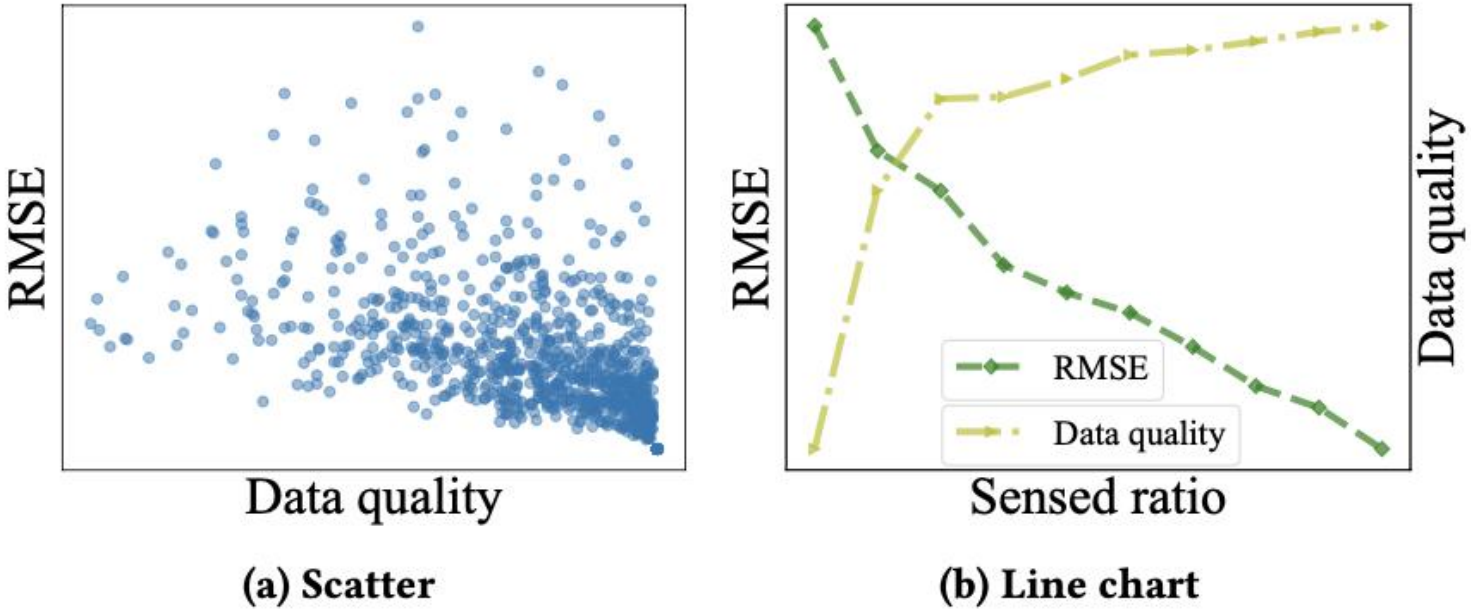
Algorithm 4 Generate New Solution

Input: $B, W, X, cost$
R: the reliabilities of all Data
ε^*: a set of already recruited workers
Output: ε' : the new set of recruited workers

$\varepsilon' \leftarrow \varepsilon^*, S_{min} \leftarrow \emptyset, \mathcal{L} \leftarrow \emptyset$
for each $\omega_i \in \varepsilon^*$ **do**
  compute $\frac{Q(X|\varepsilon' - \omega_i) - Q(X|\varepsilon')}{c_{\omega_i}}$ as Δ_{ω_i}
  $S_{min} \leftarrow S_{min} \cup (\omega_i, \Delta_{\omega_i})$
end for
sort S_{min} by ascending order of Δ_ω
for each $\omega \in S_{min}$ with its rank i **do**
  if ω is removed with $\mathcal{P}_i$ **then**
    $\varepsilon' \leftarrow \varepsilon' - \omega, cost \leftarrow cost + c_\omega$
    add ω in $\mathcal{L}$
  end if
end for
Select workers $\mathcal{W}$ from $\subseteq W - \varepsilon' - \mathcal{L}$ by greedy
Add $\mathcal{W}$ in ε

Experiment

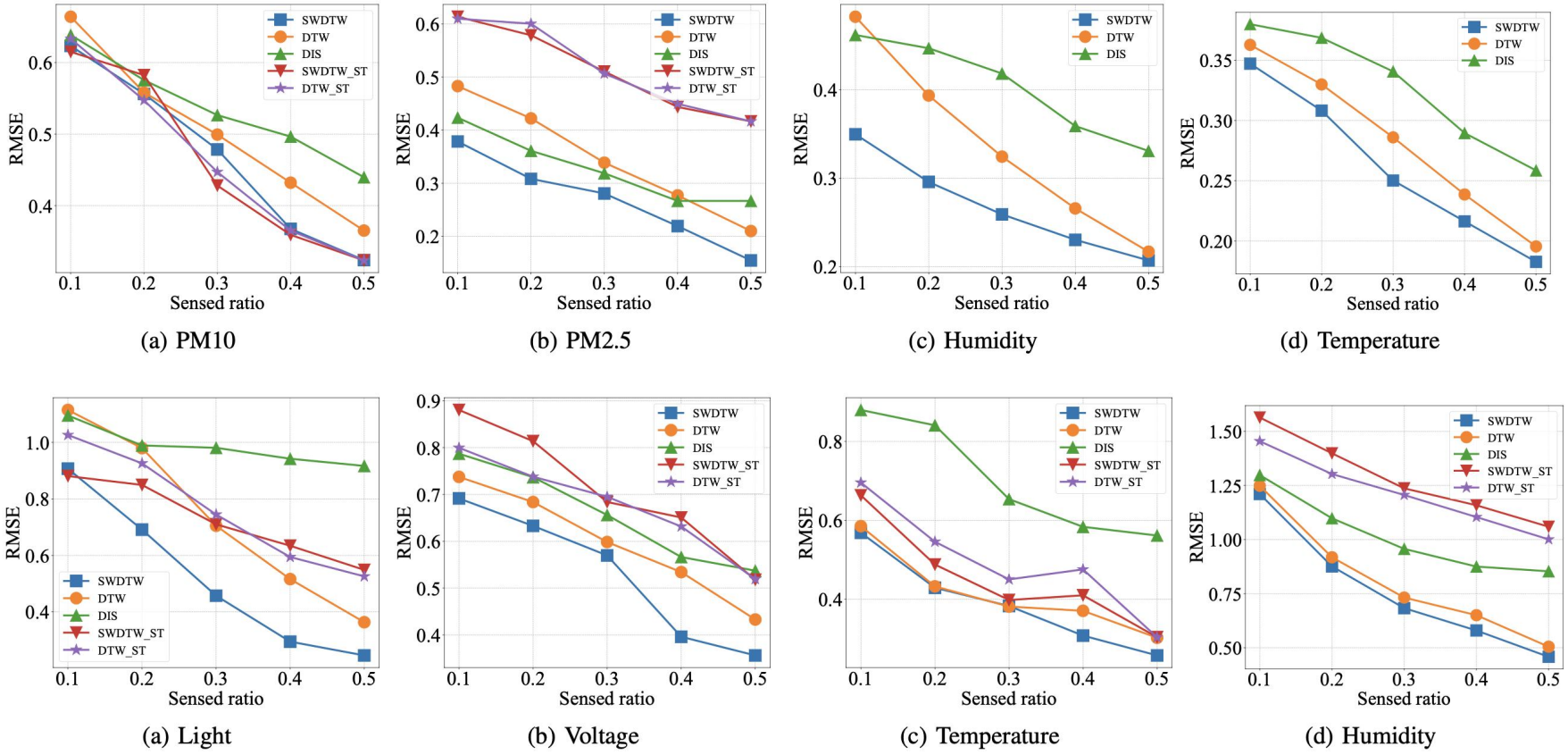
- Datapoint Selection
  - Similarity based data quality vs. Inference error



(a) Scatter (b) Line chart

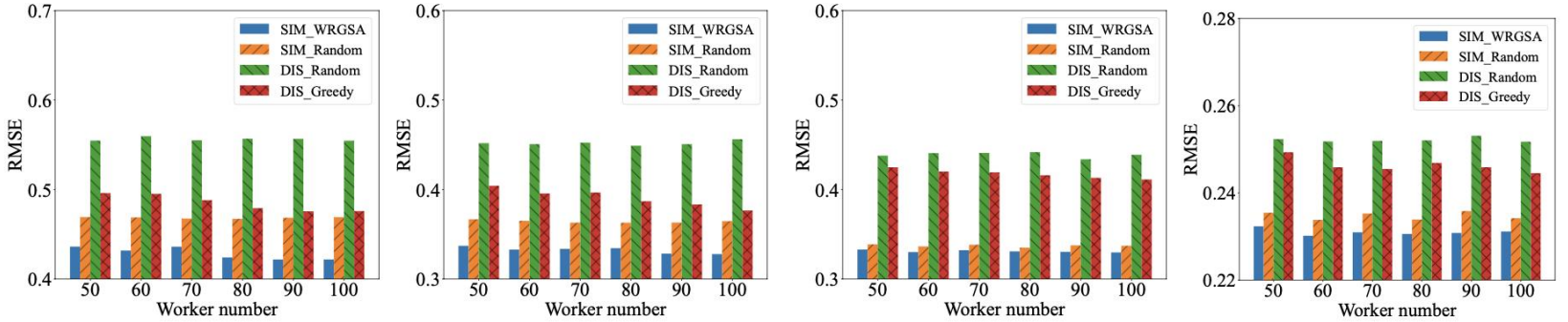
Experiment

- Datapoint Selection
  - The performance in datasets

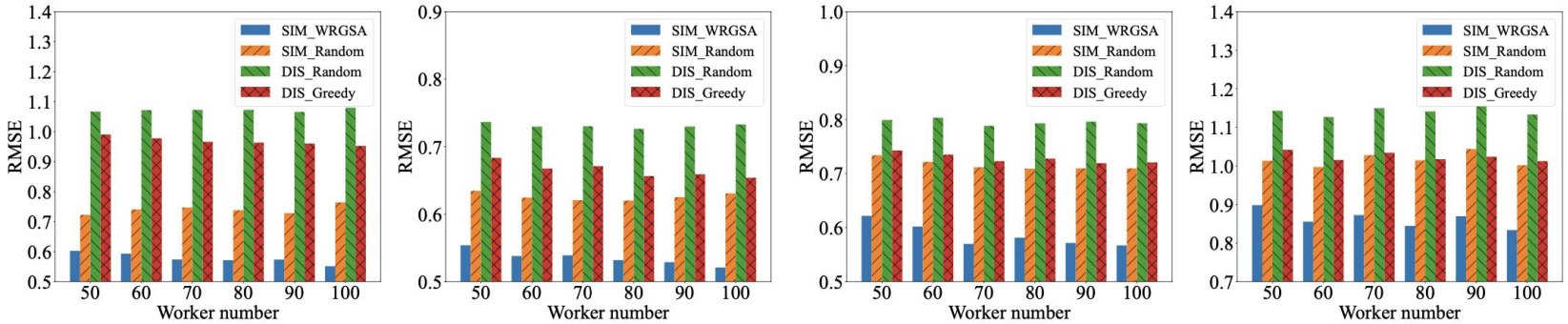


Experiment

- Worker selection
  - Increasing worker number



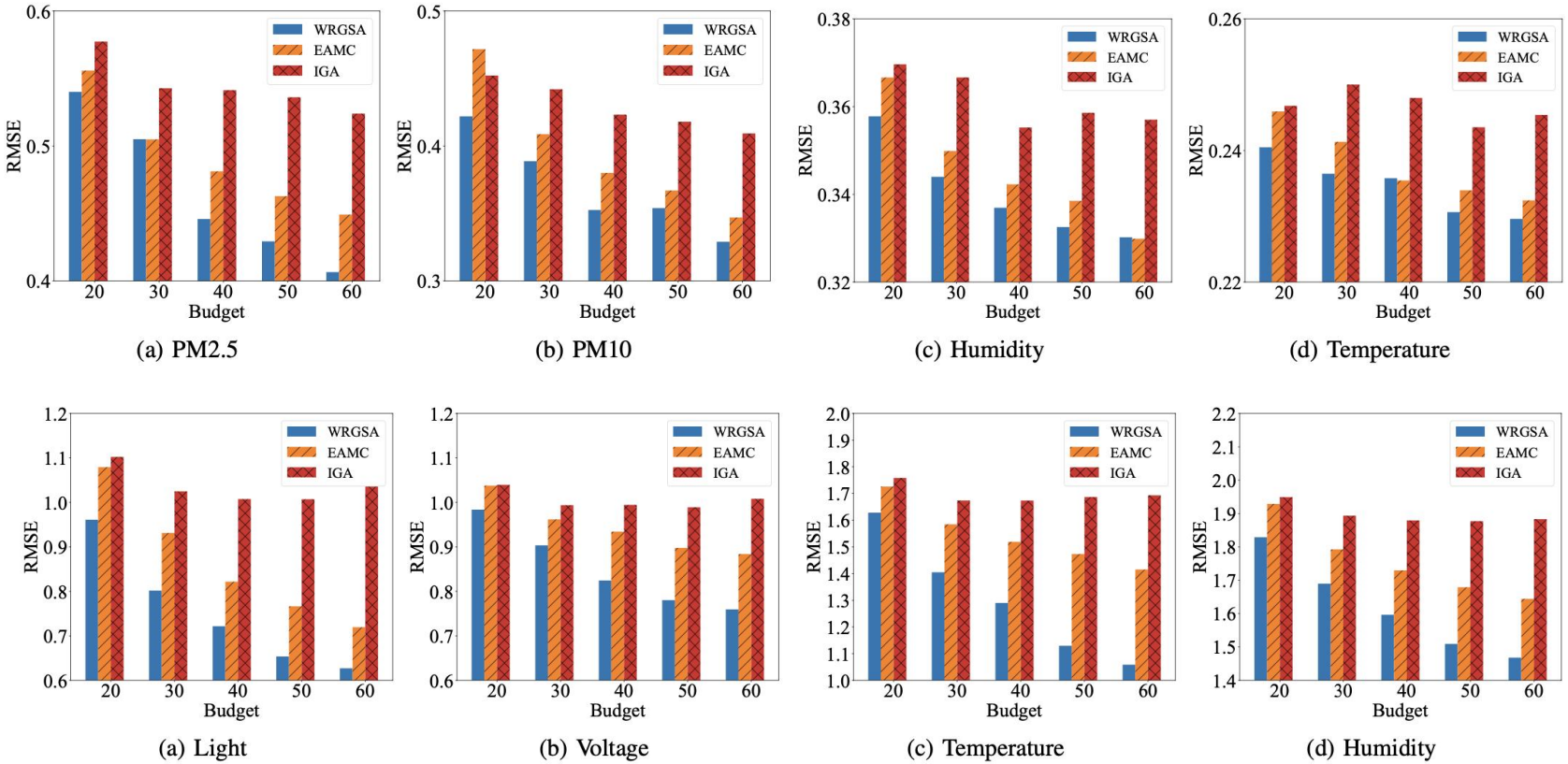
(a) PM2.5 (b) PM10 (c) Humidity (d) Temperature



(a) Light (b) Voltage (c) Temperature (d) Humidity

Experiment

- Worker selection
  - Increasing budget



Thank you!