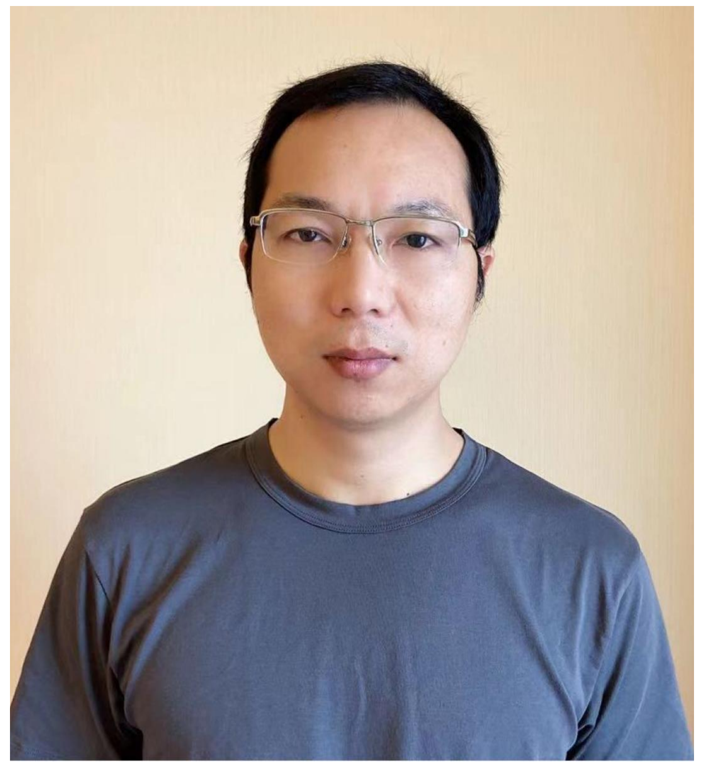
Learning-Augmented Streaming Algorithms for Correlation Clustering

Yinhao Dong
University of Science and Technology of China (USTC)

Joint work with


Shan Jiang
USTC


Shi Li
Nanjing University


Pan Peng
USTC

Correlation Clustering

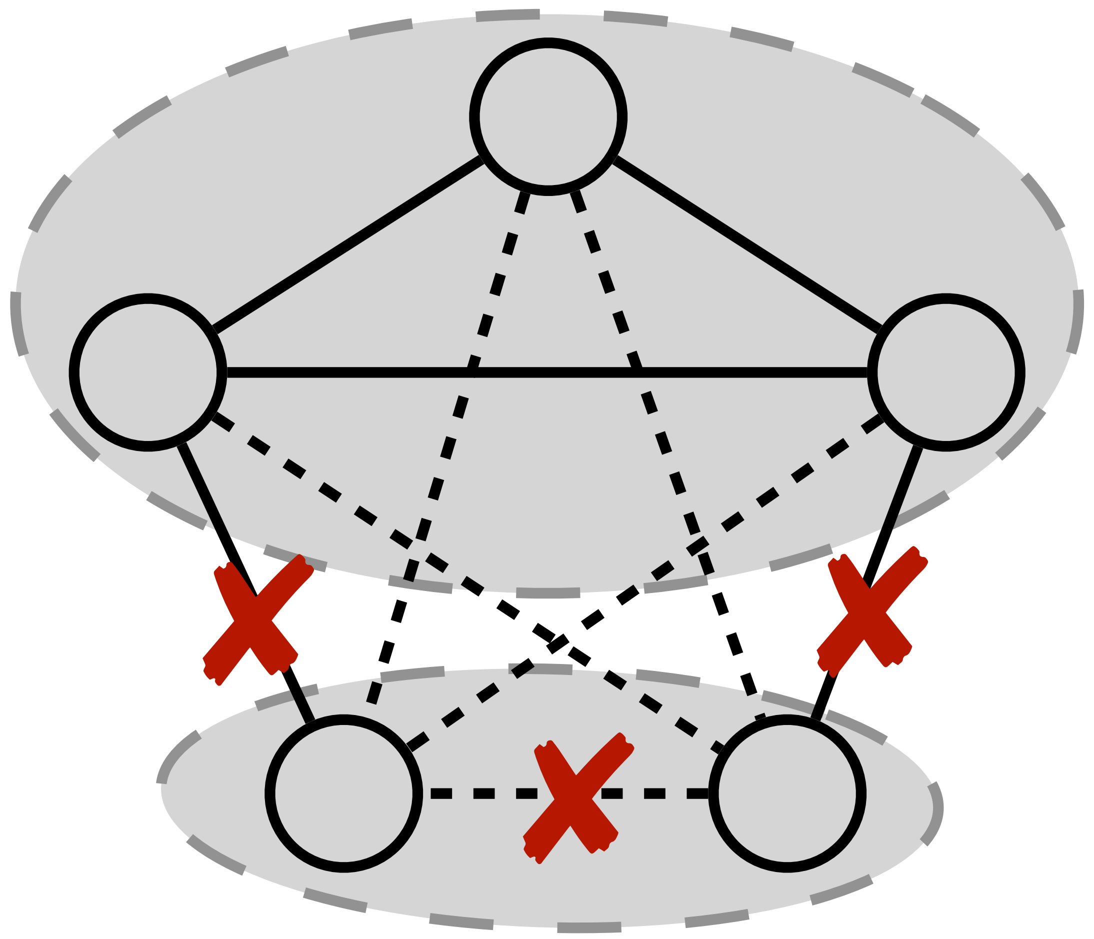
Input: graph $G = (V, E = E^+ \cup E^-)$

Output: clustering $\mathscr{C}$ of V

Goal: minimize the number of edges in disagreement

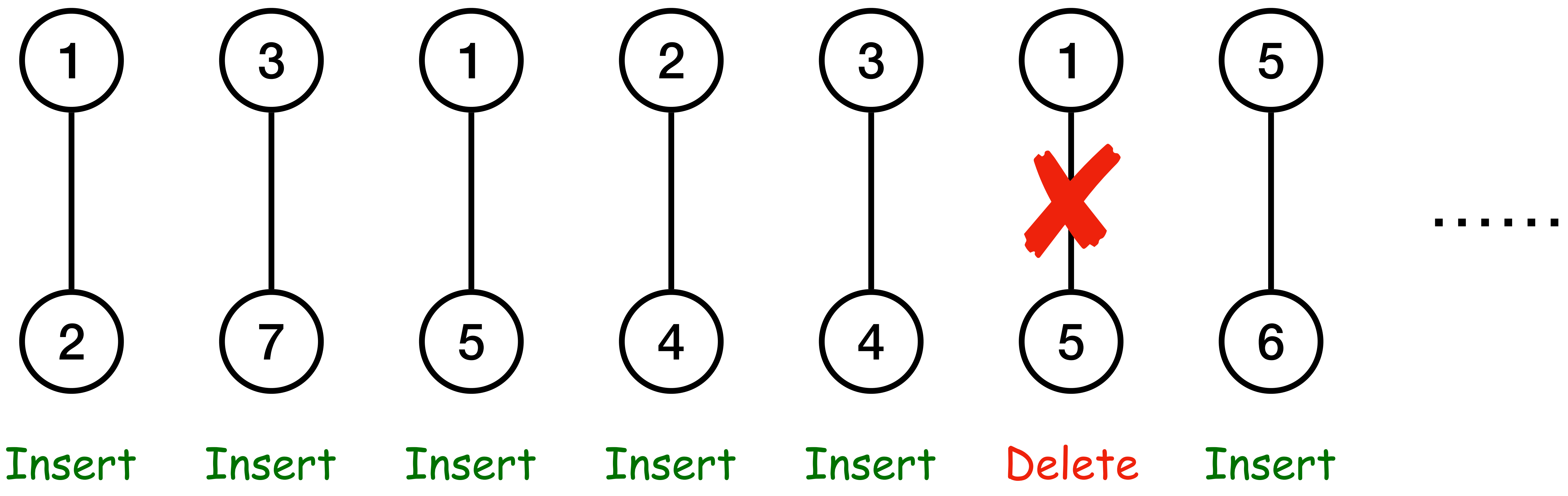
	u, v in same cluster of $\mathscr{C}$	u, v in different clusters of $\mathscr{C}$
$(u, v) \in E^+$	agreement	*disagreement*
$(u, v) \in E^-$	*disagreement*	agreement

- Most commonly studied version: G is a complete graph, i.e., $E = \binom{V}{2}$
- We consider both complete and general graphs



Streaming Model

- **Graph Stream:** The input graph is presented as a sequence of edge insertions and deletions.
  - *insertion-only* stream: contains only edge insertions
  - *dynamic* stream: contains both edge insertions and deletions
- **Goal:** scan the stream in (ideally) **one pass**, and find the solution at the end of the stream **using small space**



Correlation Clustering in Dynamic Streams

- Since outputting the clustering requires $\Omega(n)$ space, we consider semi-streaming model: $\tilde{O}(n)$ space is allowed

- **Best-known approximation-space trade-offs on complete graphs**
  - $(3+\epsilon)$-approx., $\tilde{O}(\epsilon^{-1}n)$ total space [Cambus, Kuhn, Lindy, Pai, Uitto, 2024]

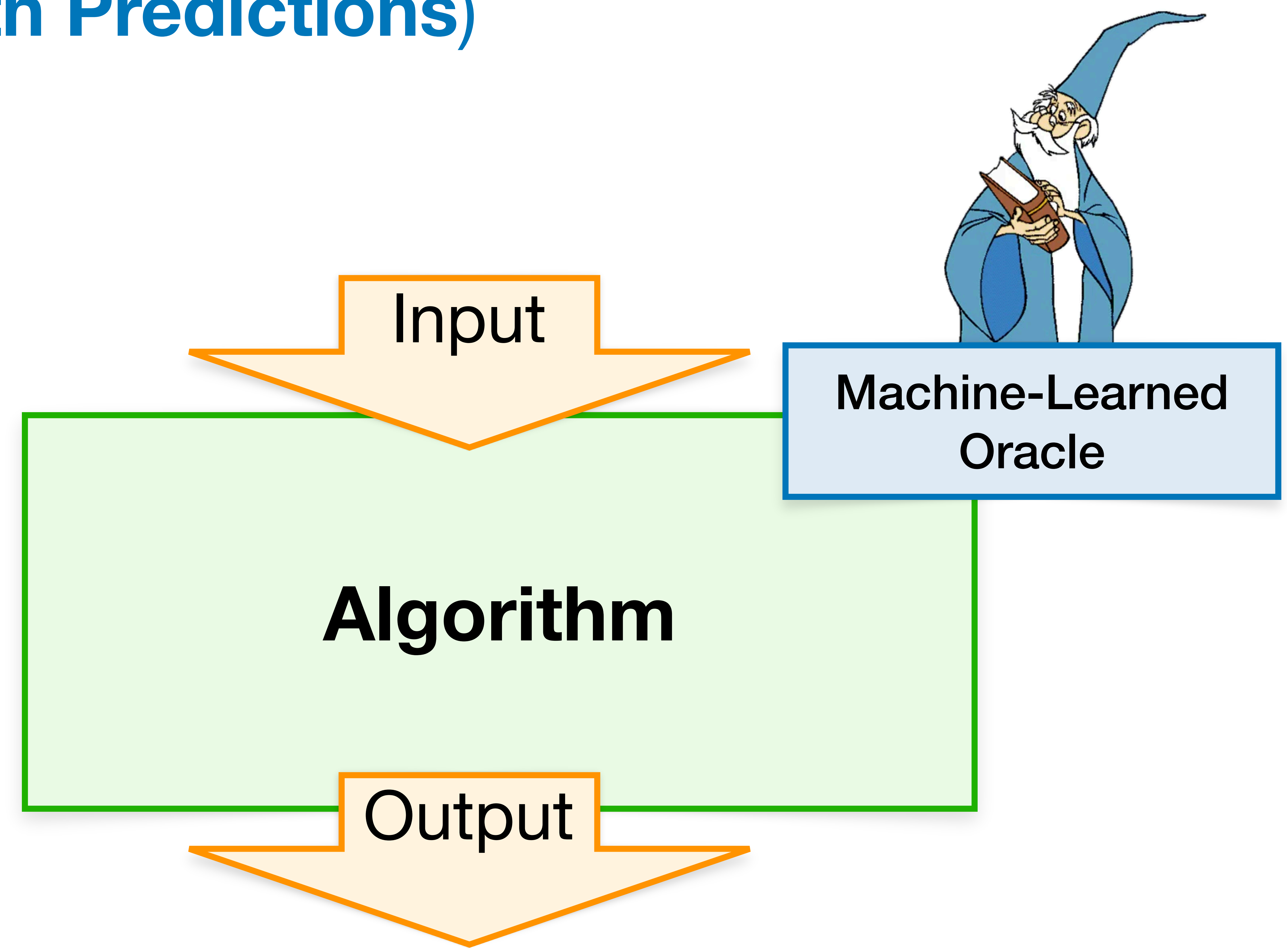
    best approx. ratio of any poly-time classical algorithm
  - $(\alpha_{\text{BEST}} + \epsilon)$-approx., $\tilde{O}(\epsilon^{-2}n)$ space during the stream, $\text{poly}(n)$ space for post-processing [Assadi, Khanna, Putterman, 2025]

- **Best-known approximation-space trade-off on general graphs**
  - $O(\log|E^-|)$-approx., $\tilde{O}(\epsilon^{-2}n + |E^-|)$ total space [Ahn, Cormode, Guha, McGregor, Wirth, 2015]

Learning-Augmented Algorithms

(a.k.a. **Algorithms with Predictions**)

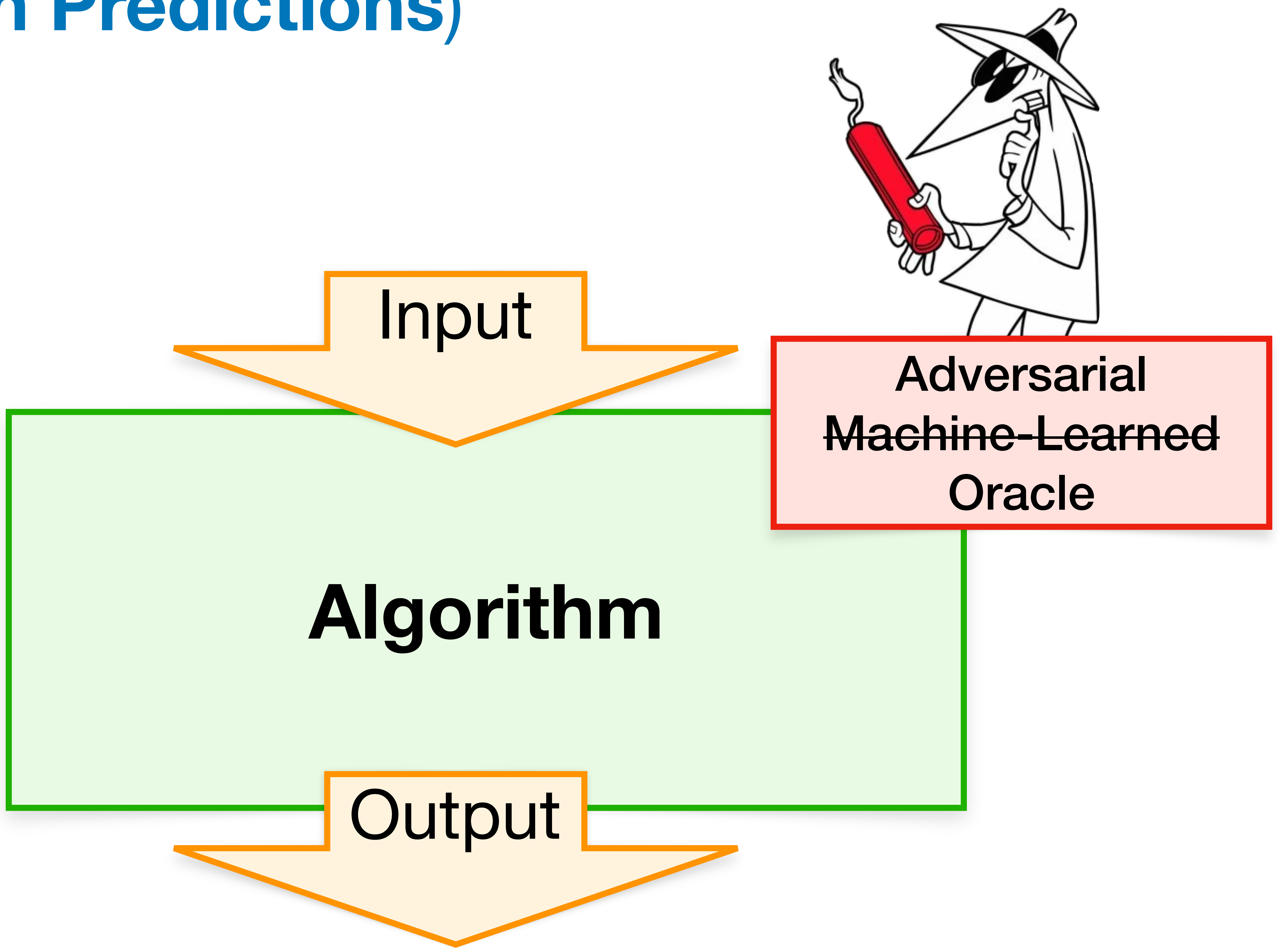
- **Motivation:** Use ML techniques in classical algorithms to improve their performance beyond *worst-case* bounds
- **Assumption:** The algorithm has oracle access to an (untrusted) predictor
- **Goals:**
  - High prediction quality $\Longrightarrow$ significantly outperforms the best-known classical (worst-case) algorithm



Learning-Augmented Algorithms

(a.k.a. **Algorithms with Predictions**)

- **Motivation:** Use ML techniques in classical algorithms to improve their performance beyond *worst-case* bounds
- **Assumption:** The algorithm has oracle access to an (untrusted) predictor
- **Goals:**
  - High prediction quality $\implies$ significantly outperforms the best-known classical (worst-case) algorithm
  - Low prediction quality $\implies$ performs no worse than the best-known classical (worst-case) algorithm



Our Prediction Model

- Oracle access to pairwise distance $d_{uv} \in [0,1]$ between any $u, v \in V$
- **Arises in many scenarios: multiple graphs on the same vertex set**
  - Healthcare: disease network, provider network, clinical trial network
  - Biology: protein-protein interaction network, gene co-expression network, signaling pathway network
  - Temporal graphs: same vertices, different edges over time
- **Observation:** Two vertices similar in one network are likely similar in another — cluster structure can thus be extracted

Our Prediction Model

β-level predictor ($\beta \geq 1$): predicts pairwise distance $d_{uv} \in [0,1]$ between any $u, v \in V$ such that

(1) $d_{uv} + d_{vw} \geq d_{uw}$ for all $u, v, w \in V$ (triangle inequality)

(2) $\displaystyle\sum_{(u,v)\in E^+} d_{uv} + \sum_{(u,v)\in E^-} (1 - d_{uv}) \leq \beta \cdot \mathrm{OPT}$

- Inspired by the metric LP formulation of Correlation Clustering
- Smaller $\beta \implies$ higher quality
- Can be implemented in practice!

$$
\begin{aligned}
\min \quad & \sum_{(u,v)\in E^+} x_{uv} + \sum_{(u,v)\in E^-} (1 - x_{uv}) & \\
\text{s.t.} \quad & x_{uw} + x_{wv} \geq x_{uv} & \forall u, v, w \in V \\
& x_{uv} \in [0,1] & \forall (u,v) \in \binom{V}{2} \\
& x_{uu} = 0 & \forall u \in V
\end{aligned}
$$

Our Results

Setting	Best-known approx.-space trade-offs (without predictions)	Our results (with predictions)
Complete graphs, Dynamic streams	$(3+\epsilon)$-approx. $\tilde{O}(\epsilon^{-1}n)$ total space [Cambus, Kuhn, Lindy, Pai, Uitto, 2024] $(\alpha_{\mathrm{BEST}}+\epsilon)$-approx. $\tilde{O}(\epsilon^{-2}n)$ space during the stream $\mathrm{poly}(n)$ space for post-processing [Assadi, Khanna, Putterman, 2025]	$(\min\{2.06\beta, 3\}+\epsilon)$-approx. $\tilde{O}(\epsilon^{-2}n)$ total space [**D.**, Jiang, Li, Peng, 2025] better approx.-space tradeoff
General graphs, Dynamic streams	$O(\log\|E^-\|)$-approx. $\tilde{O}(\epsilon^{-2}n+\|E^-\|)$ total space [Ahn, Cormode, Guha, McGregor, Wirth, 2015]	$O(\beta\log\|E^-\|)$-approx. $\tilde{O}(\epsilon^{-2}n)$ total space [**D.**, Jiang, Li, Peng, 2025] better space complexity

Our Streaming Algorithm for Complete Graphs

1. **During the stream:**
   - Maintain a truncated subgraph G' of G (refer to [Cambus, Kuhn, Lindy, Pai, Uitto, 2024]).
2. **After the stream:**
   - Run the 3-approx. combinatorial algorithm (PIVOT) on G', then assign unclustered vertices and obtain clustering $\mathscr{C}_1$ on G.
   - Run the 2.06-approx. LP rounding algorithm on G' (**use predictions d_{uv} to replace metric LP solution x_{uv}**), then assign unclustered vertices and obtain clustering $\mathscr{C}_2$ on G.
   - **return** the clustering with the lower cost between $\mathscr{C}_1$ and $\mathscr{C}_2$

Theorem [**D.**, Jiang, Li, Peng, 2025]**:**

β-level predictor

↓ w.p. $\geq 1 - 1/n^2$

$(\min\{2.06\beta, 3\} + \epsilon)$-approx.

$\tilde{O}(n)$ words of **total space**,

works in dynamic streams

Remarks:

- Better than 3-approx. under good prediction quality
- Simple and efficient
- Do not consider the space for the predictor

What I Skipped

- An algorithm for general graphs with pairwise distance predictions
  - Better space complexity than its non-learning counterpart
- Extensive experiments on synthetic and real-world datasets
  - Our algorithm performs much better in practice than the theoretical guarantee suggests.

Check out our paper and poster!