The Complexity of Approximating Earth Mover's Distance

Lorenzo Beretta (IBM Cambridge)

Oct 8, 2025

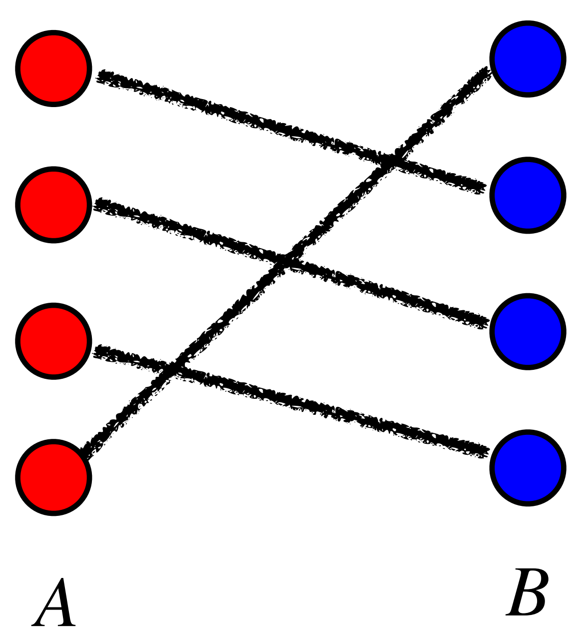
Based on joint work with: Vincent Cohen-Addad (Google), Rajesh Jayaram (Google) and Erik Waingarten (UPenn).

Earth Mover's Distance

(a.k.a. Wasserstein-1 Distance, Optimal Transport Distance)

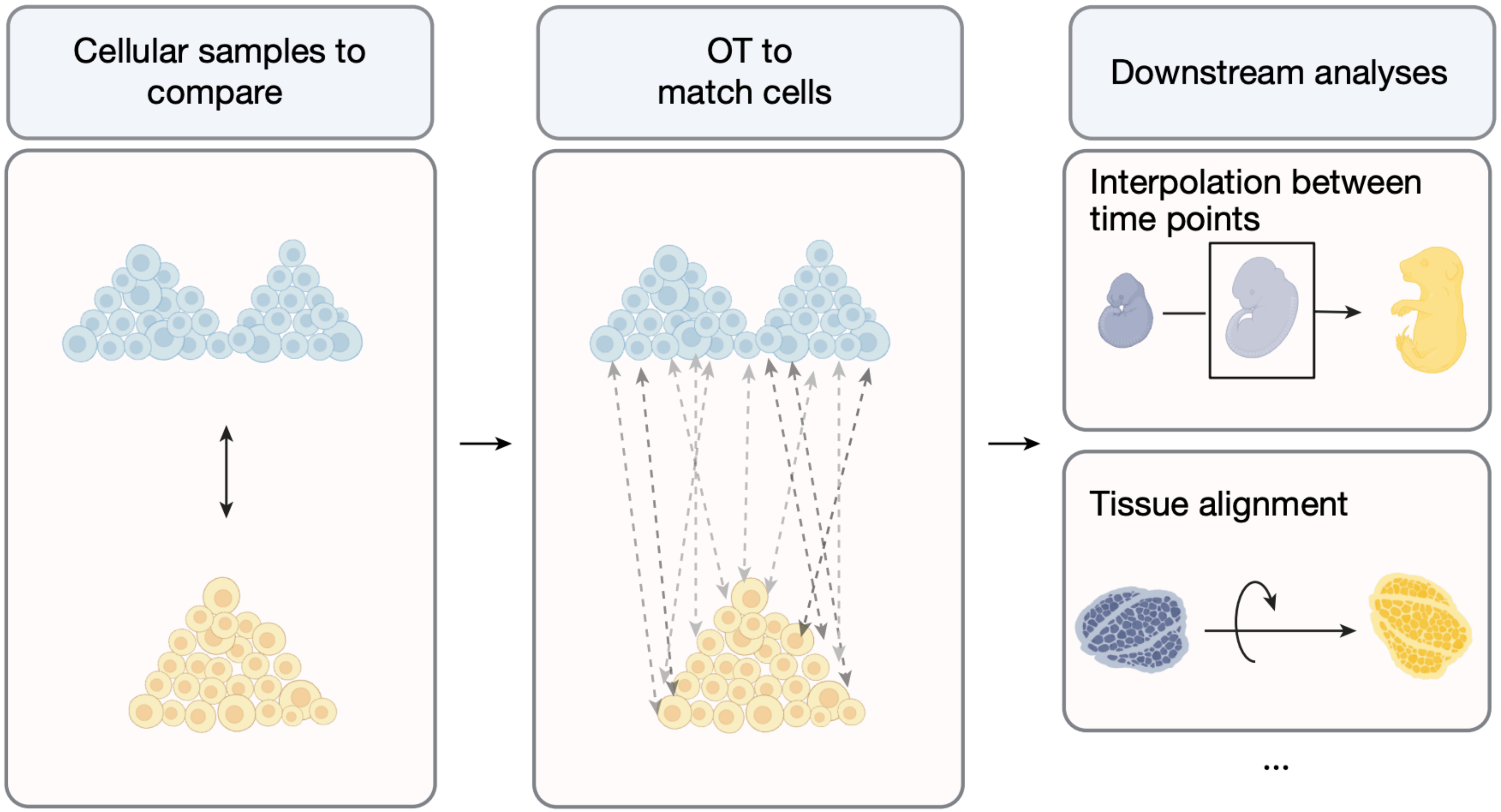
A, B size-n sets in $(\mathbb{R}^d, \ell_2)$ for $d \approx \log n$

$EMD(A, B)$ is the **minimum cost** of a perfect bipartite **matching**



Applications

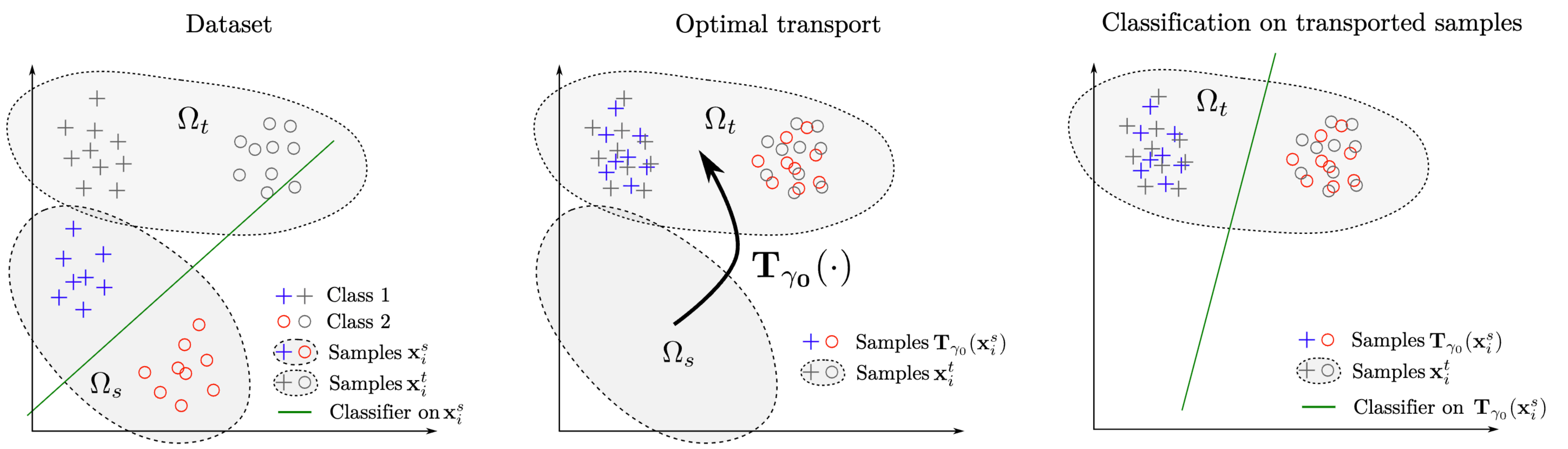
Biology: Single-Cell Analysis



“Mapping cells through time and space with moscot”. Klein et. al. 2025.

Applications

ML: Domain Adaption



"Optimal Transport for Domain Adaptation". Courty, Flamary, Tuia. 2017.

Outline

Part I: Introduction

- EMD and Applications
- **Algorithms and Complexity**
- First Sub-quadratic Algorithm

Part II: Reducing EMD to Closest Pair

- Complexity of Closest Pair
- Techniques
- Heuristic Fantasies

Epilogue: Future Directions

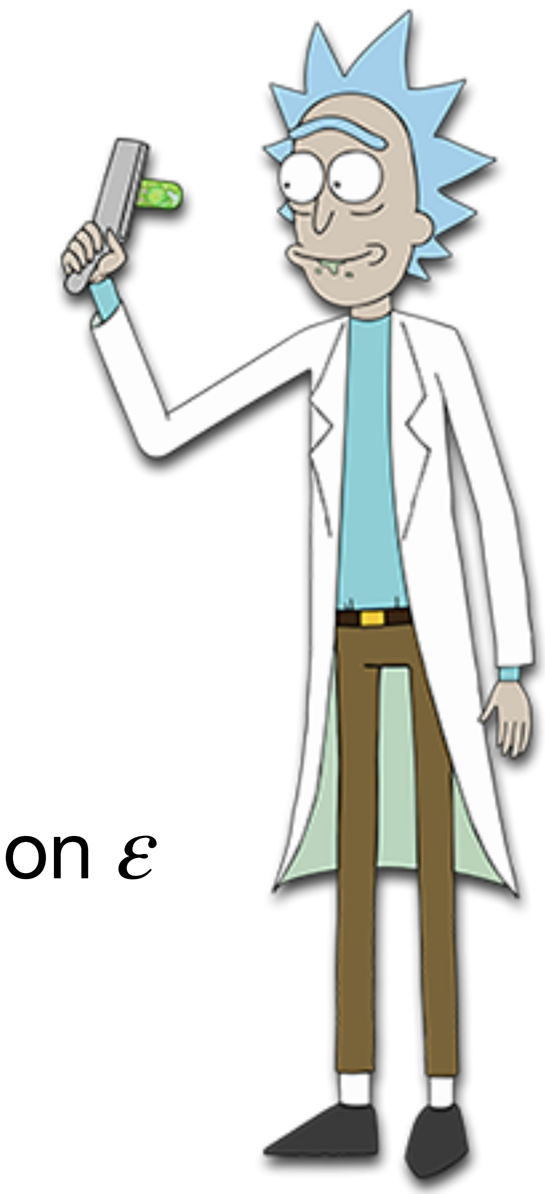
Practical Algorithms

POT: Python Optimal Transport

Exact

- Network Simplex algorithm $O(n^3)$, exact

Sinkhorn is GPU-friendly, parallelizable and differentiable



Approximation Schemes

- Sinkhorn algorithm $O(n^2/\varepsilon)$, additive approximation ε

“Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. Cuturi. 2013.

Complexity (before FOCS '23)

Exact

- Network Simplex algorithm $\cancel{O(n^3)}$, exact
- Min-Cost Flow in near-linear time $n^{2+o(1)}$
- No $n^{2-\delta}$ exact algorithm, under OVH

"Maximum Flow and Minimum-Cost Flow in Almost-Linear Time". Chen, Kyng, Liu, Peng, Gutenberg, Sachdeva. 2022.

"Conditional Hardness of Earth Mover Distance". Rohatgi. 2019.

Approximation Schemes

- Sinkhorn algorithm $O(n^2/\varepsilon)$, additive approximation ε

"Sinkhorn Distances: Lightspeed Computation of Optimal Transport". Cuturi. 2013.

Bad for $d \approx \log n$

- Curse of dimensionality: $\tilde{O}(n \cdot \varepsilon^{-d})$ for $(1+\varepsilon)$-approximation

"A deterministic near-linear time approximation scheme for geometric transportation". Fox. Lu. 2023.

Outline

Part I: Introduction

- EMD and Applications
- Algorithms and Complexity
- **First Sub-quadratic Algorithm**

Part II: Reducing EMD to Closest Pair

- Complexity of Closest Pair
- Techniques
- Heuristic Fantasies

Epilogue: Future Directions

FOCS'23: Sub-quadratic $(1+\varepsilon)$-approx. EMD

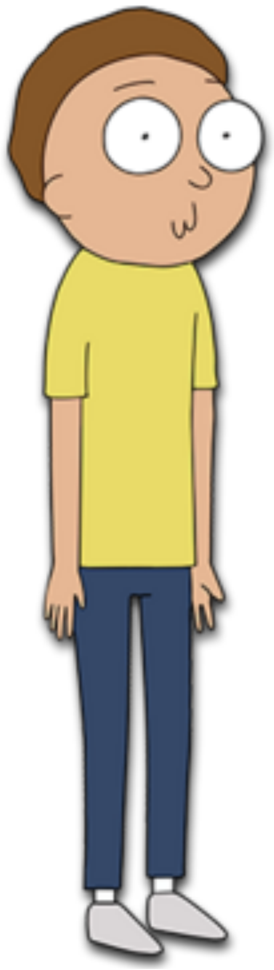
Theorem (Andoni-Zhang 2023):

$\boxed{(1+\varepsilon)\text{-Approx. Euclidean EMD can be computed in } n^{2-\Omega(\varepsilon^2)} \text{ time}}$

Techniques?

"Sub-quadratic $(1+\varepsilon)$-approximate Euclidean Spanners, with Applications". Andoni, Zhang. 2023.

Interlude: Approximate Nearest Neighbor



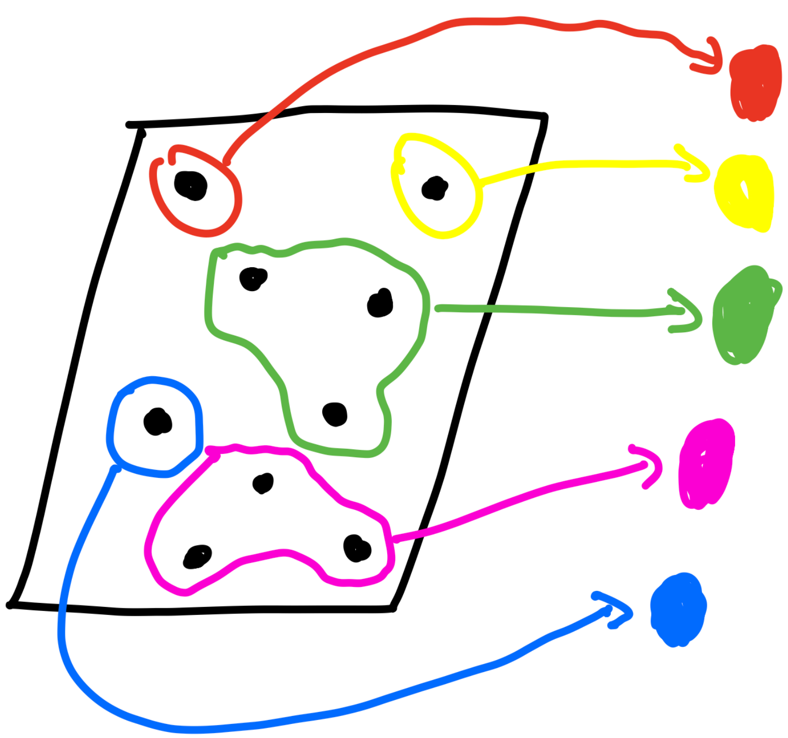
Approximate Nearest Neighbor (ANN):

Given $A \subseteq \mathbb{R}^d$ and $q \in \mathbb{R}^d$ return $\bar{a} \in A$ such that

$$||\bar{a} - q|| \leq (1 + \varepsilon) \cdot \min_{a \in A} ||a - q||$$

Locality-Sensitive Hashing

- $n^{2-\Omega(\varepsilon)}$ preprocessing time
- $n^{1-\Omega(\varepsilon)}$ query time

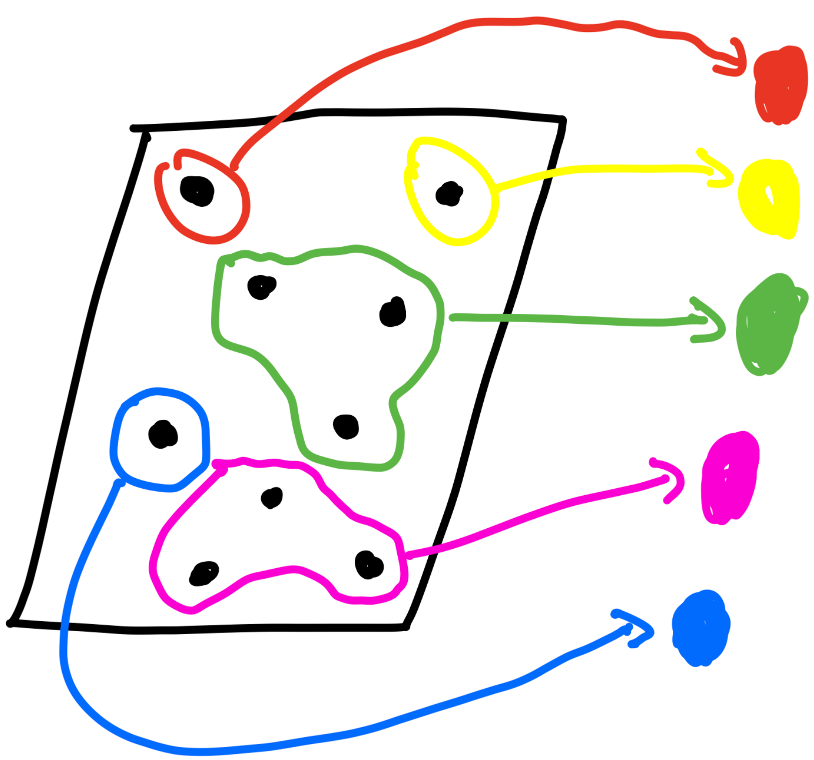


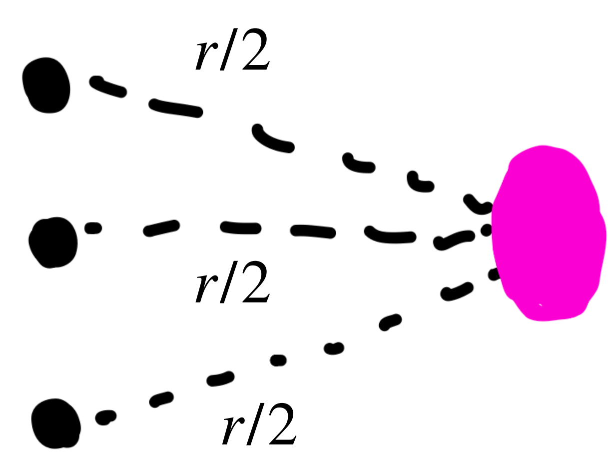
$$Pr[h(x) = h(y)] = \begin{cases} Small & \text{if } ||x - y|| > (1 + \varepsilon)r \\ Large & \text{if } ||x - y|| \leq r \end{cases}$$

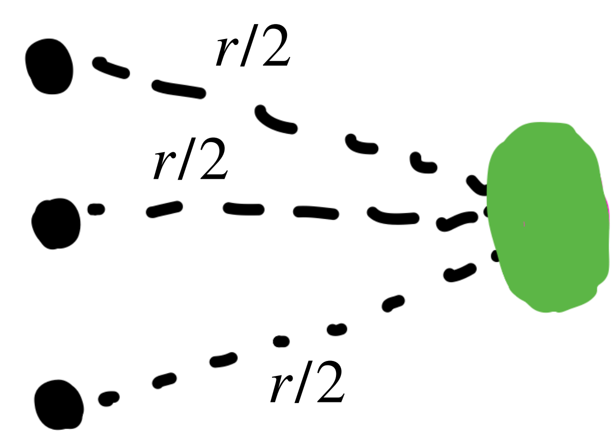
AZ'23 Techniques: Spanners via LSH

Def: A $(1+\varepsilon)$-Spanner Graph G is a weighted graph which shortest-path metric $(1+\varepsilon)$-approximates the ground metric.

Construct $(1+\varepsilon)$-Spanners with $n^{2-\Omega(\varepsilon^2)}$ edges via LSH







$$Pr[h(x) = h(y)] = \begin{cases} Small & \text{if } ||x-y|| > (1+\varepsilon)r \\ Large & \text{if } ||x-y|| \le r \end{cases}$$

AZ'23 Techniques: Spanners via LSH

Def: A $(1+\varepsilon)$-Spanner Graph G is a weighted graph which shortest-path metric $(1+\varepsilon)$-approximates the ground metric.

Construct $(1+\varepsilon)$-Spanners with $n^{2-\Omega(\varepsilon^2)}$ edges via LSH

Small Spanners + Min-Cost Flow in near-linear time =

$(1+\varepsilon)$-approximate EMD in time $n^{2-\Omega(\varepsilon^2)}$

Are the complexity of EMD and ANN related?

Outline

Part I: Introduction

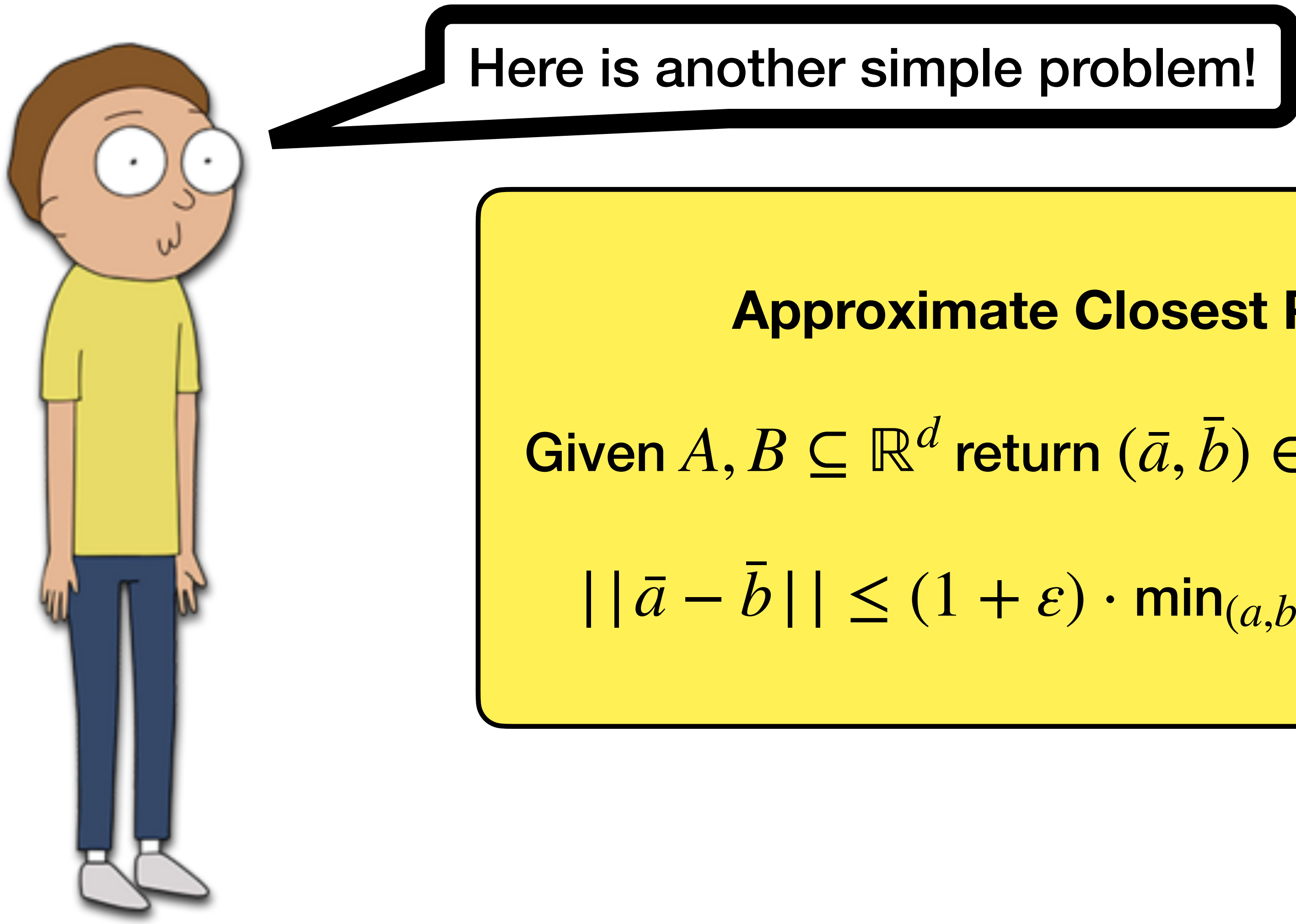
- EMD and Applications
- Algorithms and Complexity
- First Sub-quadratic Algorithm

Part II: Reducing EMD to Closest Pair

- **Main Theorems**
- Techniques
- Heuristic Fantasies

Epilogue: Future Directions

Approximate Closest Pair



Approximate Closest Pair (CP):

Given $A, B \subseteq \mathbb{R}^d$ return $(\bar{a}, \bar{b}) \in A \times B$ such that

$$||\bar{a} - \bar{b}|| \leq (1 + \varepsilon) \cdot \min_{(a,b) \in A \times B} ||a - b||$$

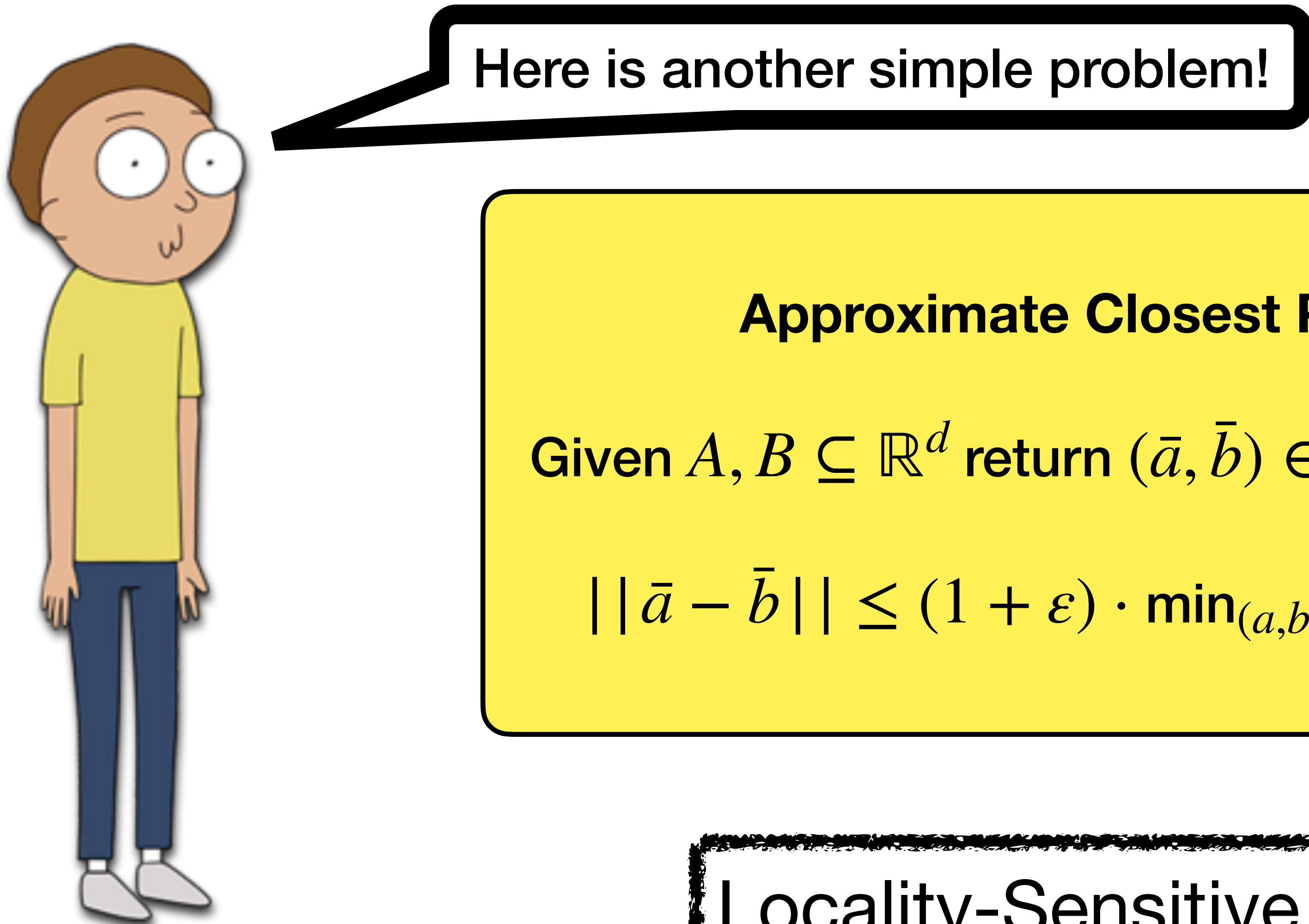
EMD reduces to Closest Pair

Theorem (B., Cohen-Addad, Jayaram, Waingarten ’25):

Given an algorithm for $(1+\varepsilon)$-approximate Closest Pair that runs in time $n^{2-\phi}$, there exists an algorithm for $(1+O(\varepsilon))$-approximate EMD that runs in time $n^{2-\Omega(\phi)}$

“Approximating High-Dimensional Earth Mover’s Distance as Fast as Closest Pair”. Beretta, Cohen-Addad, Jayaram, Waingarten. 2025.

Approximate Closest Pair



Approximate Closest Pair (CP):

Given $A, B \subseteq \mathbb{R}^d$ return $(\bar{a}, \bar{b}) \in A \times B$ such that

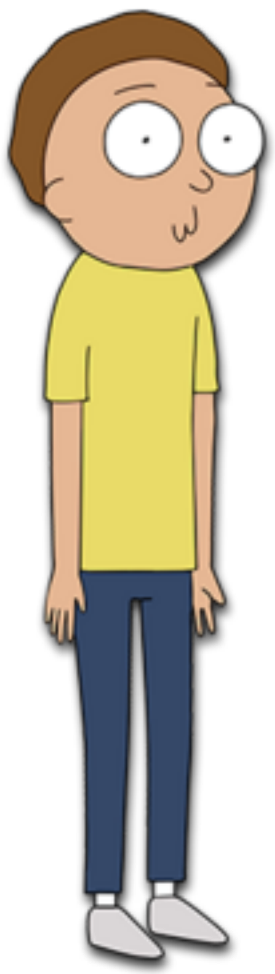
$$||\bar{a} - \bar{b}|| \leq (1 + \varepsilon) \cdot \min_{(a,b) \in A \times B} ||a - b||$$

Locality-Sensitive Hashing

- $n^{2-\Omega(\varepsilon)}$ preprocessing time
- $n^{1-\Omega(\varepsilon)}$ query time

n ANN queries yield complexity $n^{2-\Omega(\varepsilon)}$

Approximate Closest Pair



Approximate Closest Pair (CP):

Given $A, B \subseteq \mathbb{R}^d$ return $(\bar{a}, \bar{b}) \in A \times B$ such that

$$||\bar{a} - \bar{b}|| \leq (1 + \varepsilon) \cdot \min_{(a,b)\in A\times B} ||a - b||$$

CP can be solved in time $n^{2-\Omega(\varepsilon^{1/3})} \ll n^{2-\Omega(\epsilon)}$

Via polynomial method and fast matrix multiplication

"Finding Correlations in Subquadratic Time, with Applications to Learning Parities and Juntas". Valiant. 2012.

"Polynomial Representations of Threshold Functions and Algorithmic Applications". Alman, Chan, Williams. 2016.

CP algorithms must use at least $n^{2-f(\varepsilon)}$ time with $f(\varepsilon) \to 0$

"Hardness of Approximate Nearest Neighbor Search". Rubinstein. 2018.

EMD reduces to Closest Pair

Theorem (B., Cohen-Addad, Jayaram, Waingarten '25):

Given an algorithm for $(1+\varepsilon)$-approximate Closest Pair that runs in time $n^{2-\phi}$, there exists an algorithm for $(1+O(\varepsilon))$-approximate EMD that runs in time $n^{2-\Omega(\phi)}$

Corollary

We solve $(1+\varepsilon)$-approximate EMD in time $n^{2-\Omega(\varepsilon^{1/3})} \ll n^{2-\Omega(\varepsilon^2)}$, improving over AZ '23 and breaking the LSH barrier.

"Approximating High-Dimensional Earth Mover's Distance as Fast as Closest Pair". Beretta, Cohen-Addad, Jayaram, Waingarten. 2025.

Outline

Part I: Introduction

- EMD and Applications
- Algorithms and Complexity
- First Sub-quadratic Algorithm

Part II: Reducing EMD to Closest Pair

- Complexity of Closest Pair
- **Techniques**
- Heuristic Fantasies

Epilogue: Future Directions

EMD Linear Program

Primal

Minimize $\sum_{a,b} x_{a,b} \cdot ||a-b||$

subject to

$$\sum_{a} x_{a,b} = 1 \quad \forall b \in B$$

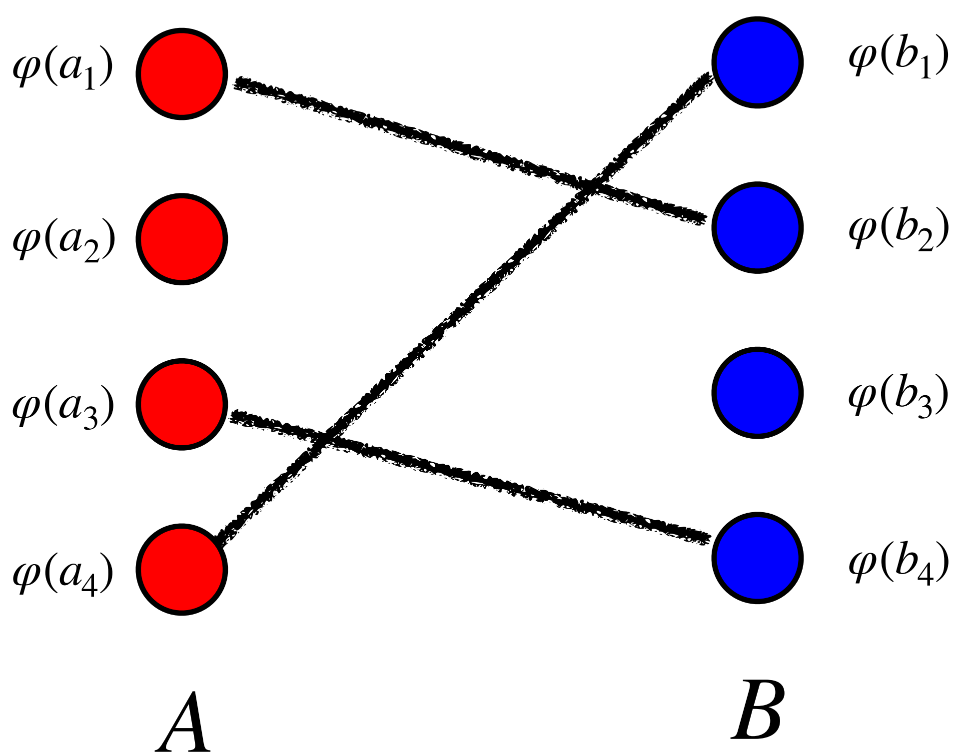
$$\sum_{b} x_{a,b} = 1 \quad \forall a \in A$$

$$x_{a,b} \geq 0 \quad \forall a, b$$

Dual

Maximize $\sum_{a} \varphi(a) + \sum_{b} \varphi(b)$

subject to $\varphi(a) + \varphi(b) \leq ||a-b|| \quad \forall a, b$



Approximate Dual via MWU

Dual

$$\begin{array}{ll} \text{Maximize} & \sum_a \varphi(a) + \sum_b \varphi(b) \\ \text{subject to} & \varphi(a) + \varphi(b) \leq ||a-b|| \quad \forall a, b \end{array}$$

- Update the dual variables $\varphi^{(t)}$ over time $t = 1 \ldots O(\log n)$
- Maintain a distribution $\lambda_{a,b}^{(t)}$ over $A \times B$
- Define $\varphi^{(t+1)}$ that satisfies the $\lambda_{a,b}^{(t)}$-average of constraints

To this end, $\tilde{O}(n)$ samples suffice!

- $$\lambda_{a,b}^{(t+1)} \propto \lambda_{a,b}^{(t)} \cdot \exp\left(\frac{\varphi^{(t)}(a) + \varphi^{(t)}(b)}{||a-b||}\right) \propto \exp\left(\sum_{s \leq t} \frac{\varphi^{(s)}(a) + \varphi^{(s)}(b)}{||a-b||}\right)$$

Sampling via Closest Pair

$$\lambda_{a,b} \propto \exp\left(\frac{K}{||a-b||}\right)$$

Approximate Closest Pair (CP):
Given $A, B \subseteq (\mathbb{R}^d, \ell_1)$,
find $a \in A$ and $b \in B$ that minimize
$||a-b||_2$ up to a factor $1+\varepsilon$.

With some work, we can reduce to the case

$$\sum_{s<t} \varphi^{(s)}(a) + \varphi^{(s)}(b) = K_{a,b} = K$$

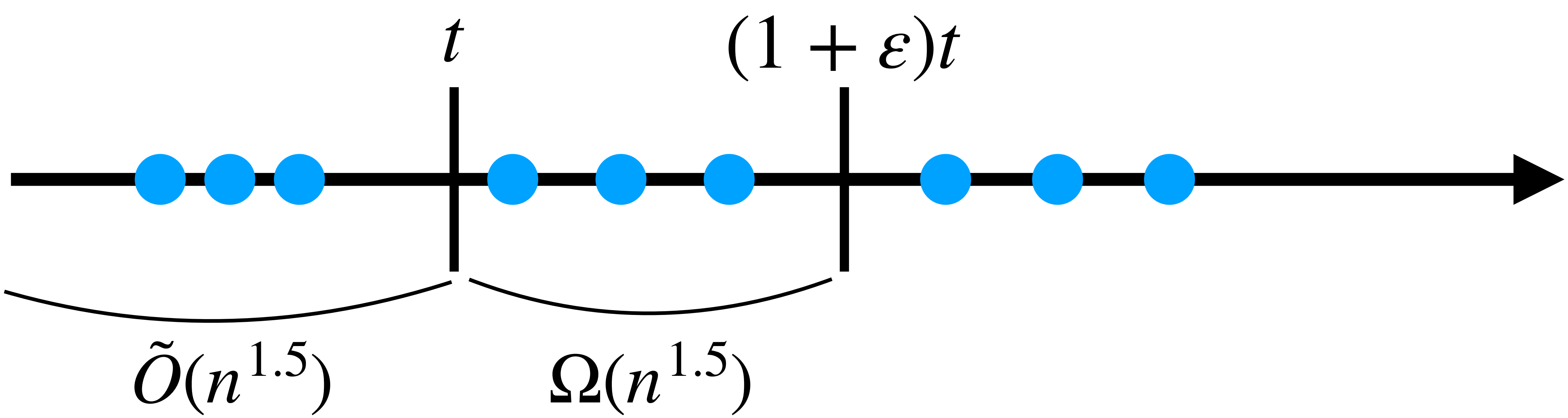
For $K \approx \varepsilon^{-1} \cdot \log n$, CP reduces to sampling from $\lambda_{a,b}$

Sampling via Closest Pair

$$\lambda_{a,b} \propto \exp\left(\frac{K}{||a-b||}\right) \rightarrow$$

All Close Pairs

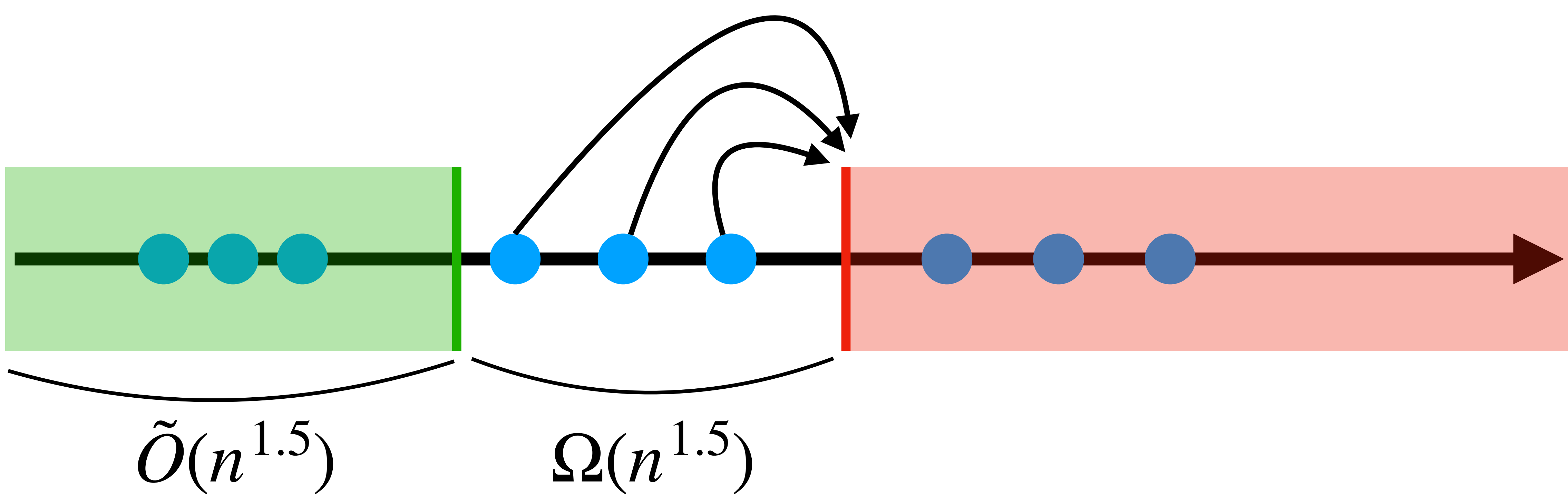
Given $A, B \subseteq (\mathbb{R}^d, \ell_1)$ and $t > 0$, return all pairs at distance $\leq t$ if there are less than $n^{1.5}$ pair at distance $\leq (1+\varepsilon)t$



Sampling via Closest Pair

$$\lambda_{a,b} \propto \exp\left(\frac{K}{||a-b||}\right)$$

All Close Pairs
Given $A, B \subseteq (\mathbb{R}^d, \ell_1)$ and $t > 0$, return all pairs at distance $\leq t$ if there are less than $n^{1.5}$ pair at distance $\leq (1+\varepsilon)t$. Else, return $\perp$.



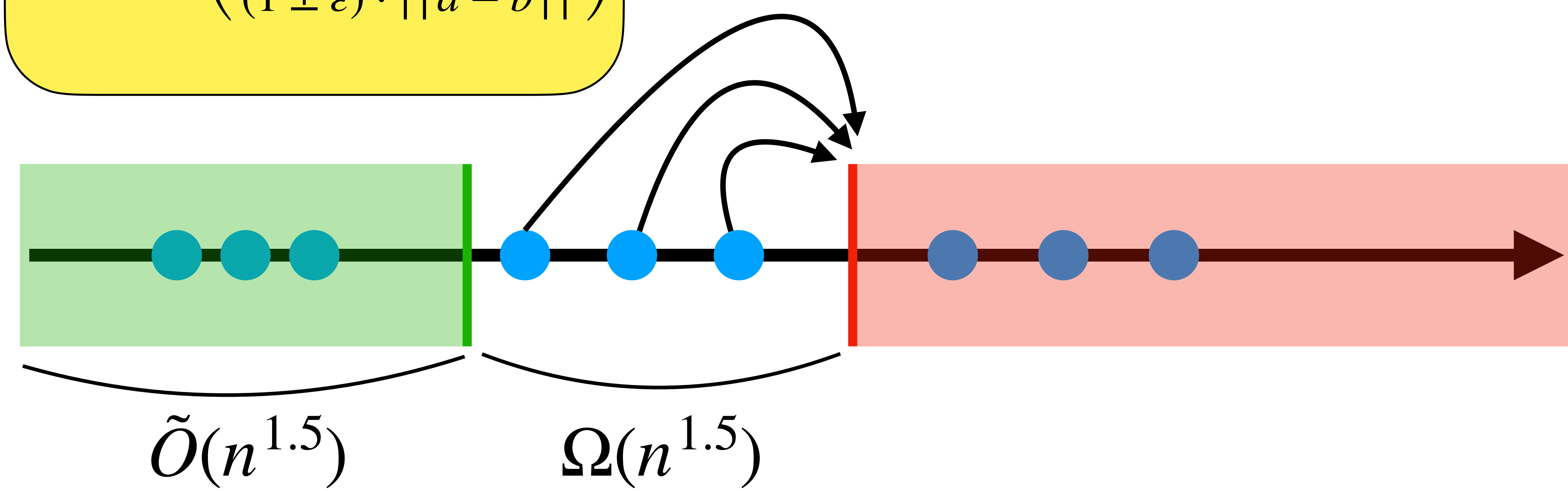
Sampling via Closest Pair

$$\cancel{\lambda_{a,b} \propto \exp\left(\frac{K}{||a-b||}\right)}$$

All Close Pairs

Given $A, B \subseteq (\mathbb{R}^d, \ell_1)$ and $t > 0$, return all pairs at distance $\leq t$ if there are less than $n^{1.5}$ pair at distance $\leq (1+\varepsilon)t$

$$\lambda_{a,b} \propto \exp\left(\frac{K}{(1 \pm \varepsilon) \cdot ||a-b||}\right)$$



Outline

Part I: Introduction

- EMD and Applications
- Algorithms and Complexity
- First Sub-quadratic Algorithm

Part II: Reducing EMD to Closest Pair

- Complexity of Closest Pair
- Techniques
- **Heuristic Fantasies**

Epilogue: Future Directions

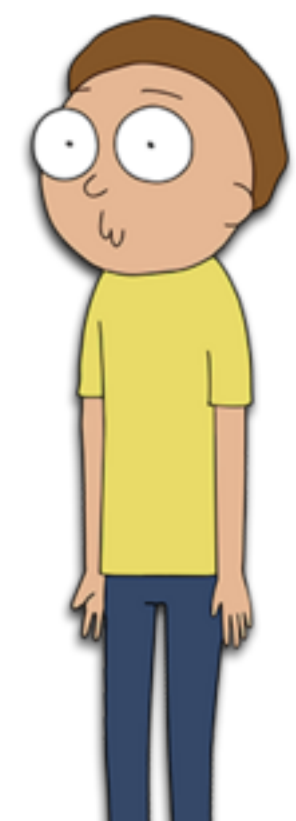
EMD reduces to Closest Pair

Theorem (B., Cohen-Addad, Jayaram, Waingarten ’25):

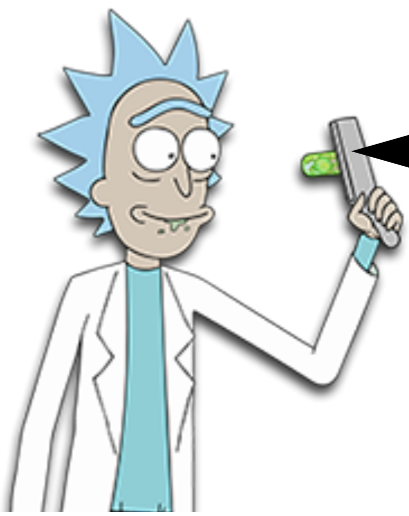
Given an algorithm for $(1+\varepsilon)$-approximate Closest Pair that runs in time $n^{2-\phi}$, there exists an algorithm for $(1+O(\varepsilon))$-approximate EMD that runs in time $n^{2-\Omega(\phi)}$

Reducing EMD to CP is pointless, as we have a lower bound of $n^{2-f(\varepsilon)}$ with $f(\varepsilon) \to 0$ for CP

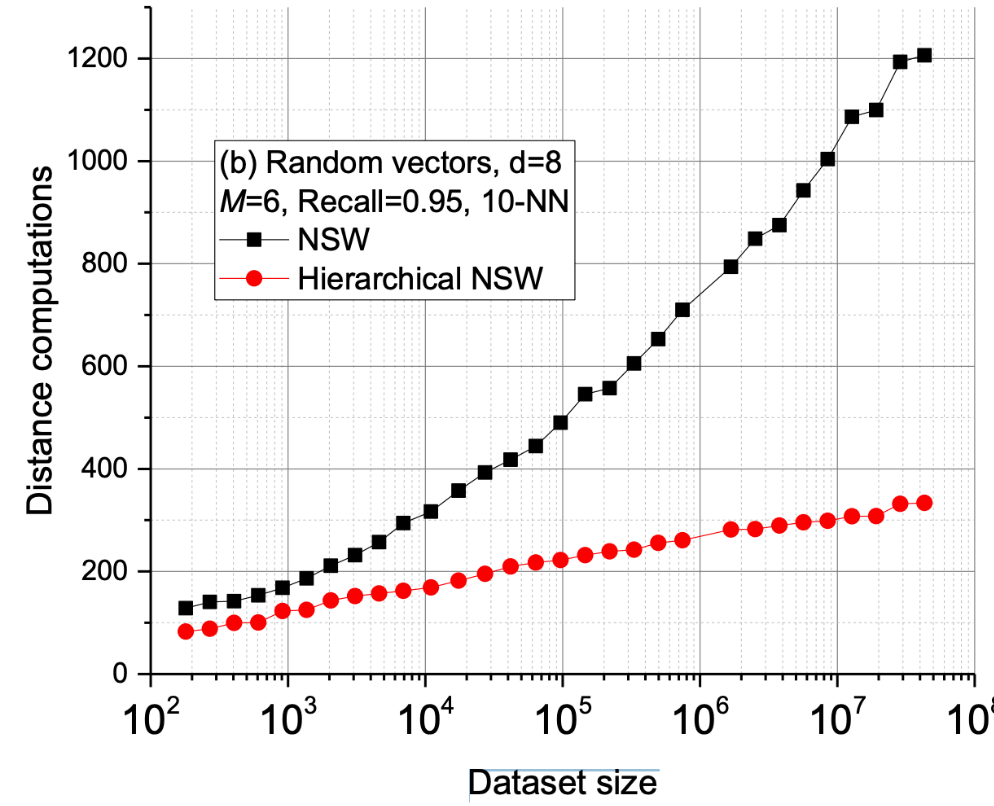
Fast heuristics for ANN / CP might give fast algorithms for EMD!

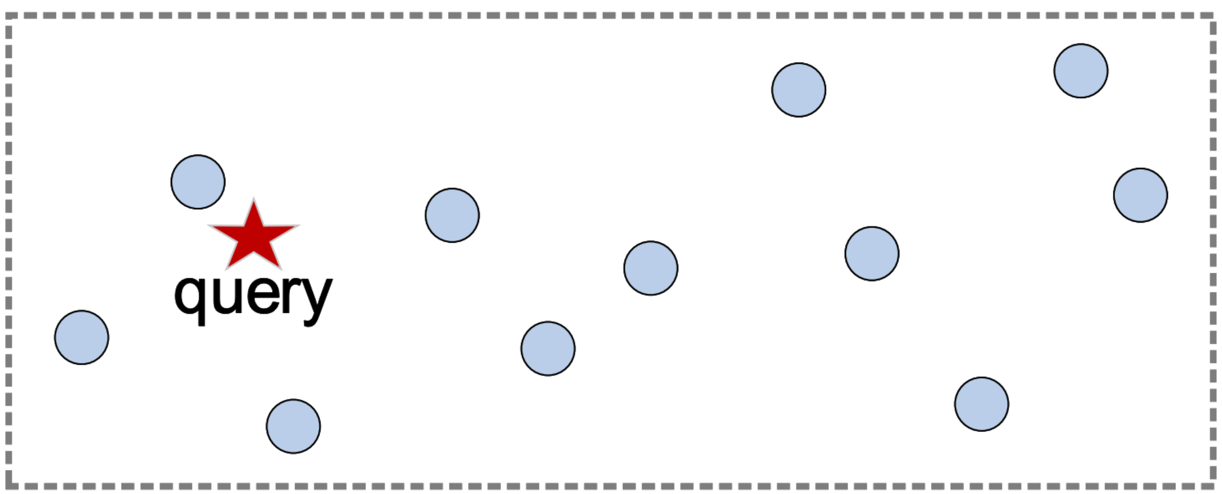


Heuristics for ANN

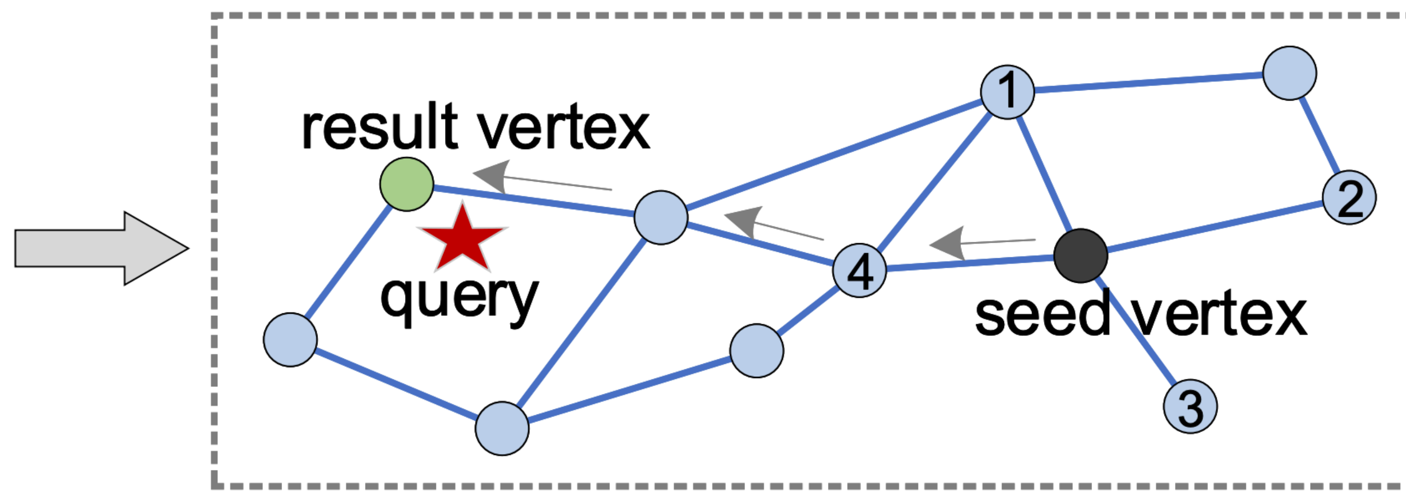


These ANN heuristics take $\approx \log n$ time empirically!





(a) Original dataset

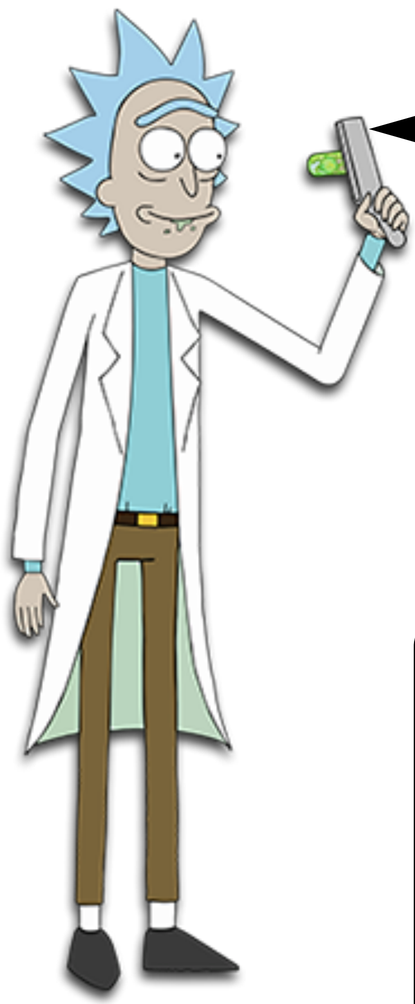


(b) Graph index

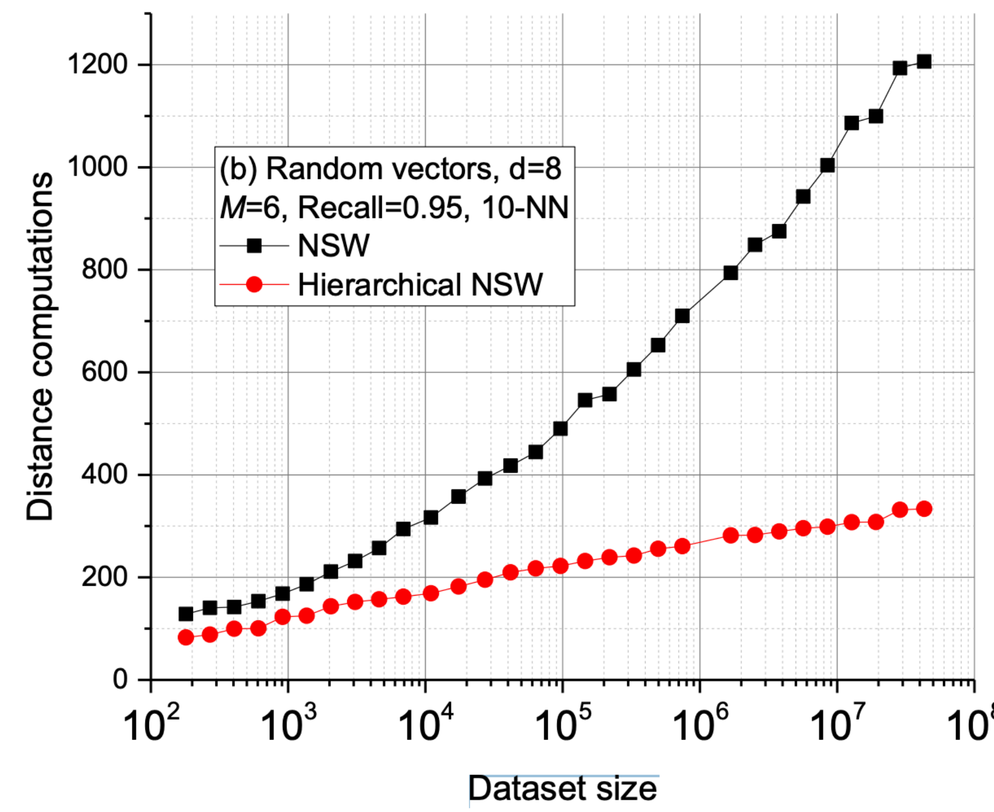
"A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search". Wang, Xu, Yue, Wang. 2021.

"Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs". Malkov, Yashunin. 2018.

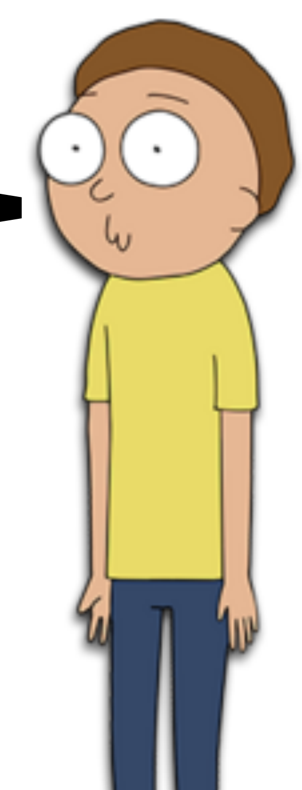
Heuristics for ANN



These ANN heuristics take $\approx \log n$ time empirically!



Rick, the reduction from EMD to CP is highly impractical!



Keep dreaming, boy!

Heuristics for EMD through ANN

Pros

- ANN algorithms seem to exploit the structure of data
- ANN is highly studied and engineered

Cons

- Current reduction is impractical
- Unlike Sinkhorn, Graph-based ANN Algorithm are sequential

Outline

Part I: Algorithms and Complexity for EMD

- Practical Algorithms
- Complexity
- First Sub-quadratic Algorithm

Part II: Reducing EMD to Closest Pair

- Complexity of Closest Pair
- Techniques
- Heuristic Fantasies

Epilogue: Further Directions

Further Directions

Likely a HARD question!

$\boxed{\text{Is } (1+\varepsilon)\text{-approximate EMD in } n^{1.99} \text{ time possible?}}$

Can we prove a fine-grained lower bound on $(1+\varepsilon)$-approximate EMD similar to the one proved for CP?

CP algorithms must use at least $n^{2-f(\varepsilon)}$ time with $f(\varepsilon) \to 0$

"Hardness of Approximate Nearest Neighbor Search". Rubinstein. 2018.

Further Directions

LSH-based algos for c-approximate EMD run in $n^{1+\Theta(1/c)}$ time.

Can we break the LSH barrier also for c larger than $1 + \varepsilon$?

Can we extend our techniques to more geometric problems such as Kernel Density Estimation?

Further Directions

Design heuristics for EMD that leverage fast ANN heuristics

Theory: Find simpler EMD-to-CP reductions

Practice: Explore ANN-based EMD heuristics empirically

Summary

- We proved a reduction between two central problems in high-dimensional computational geometry: EMD and CP
- Algorithmically, we broke the LSH barrier for approximate EMD
- In practice, ANN is easier than worst-case bounds
- Our reduction might inspire heuristics for EMD via ANN heuristics

Thanks! Any Questions?