# Learning Nearest Neighbor Graphs from Noisy Distance Samples

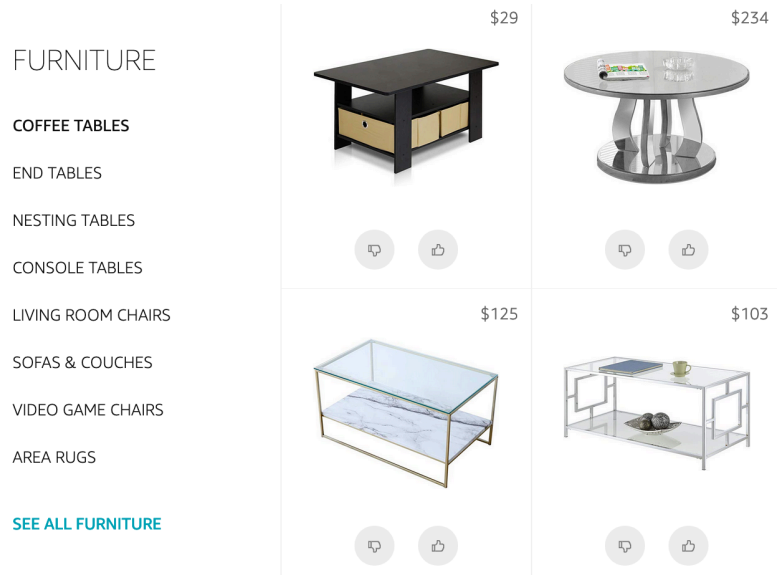
Blake Mason, Ardhendu Tripathy, Robert Nowak

WISCONSIN
INSTITUTE FOR DISCOVERY
AT THE UNIVERSITY OF WISCONSIN-MADISON

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Motivation

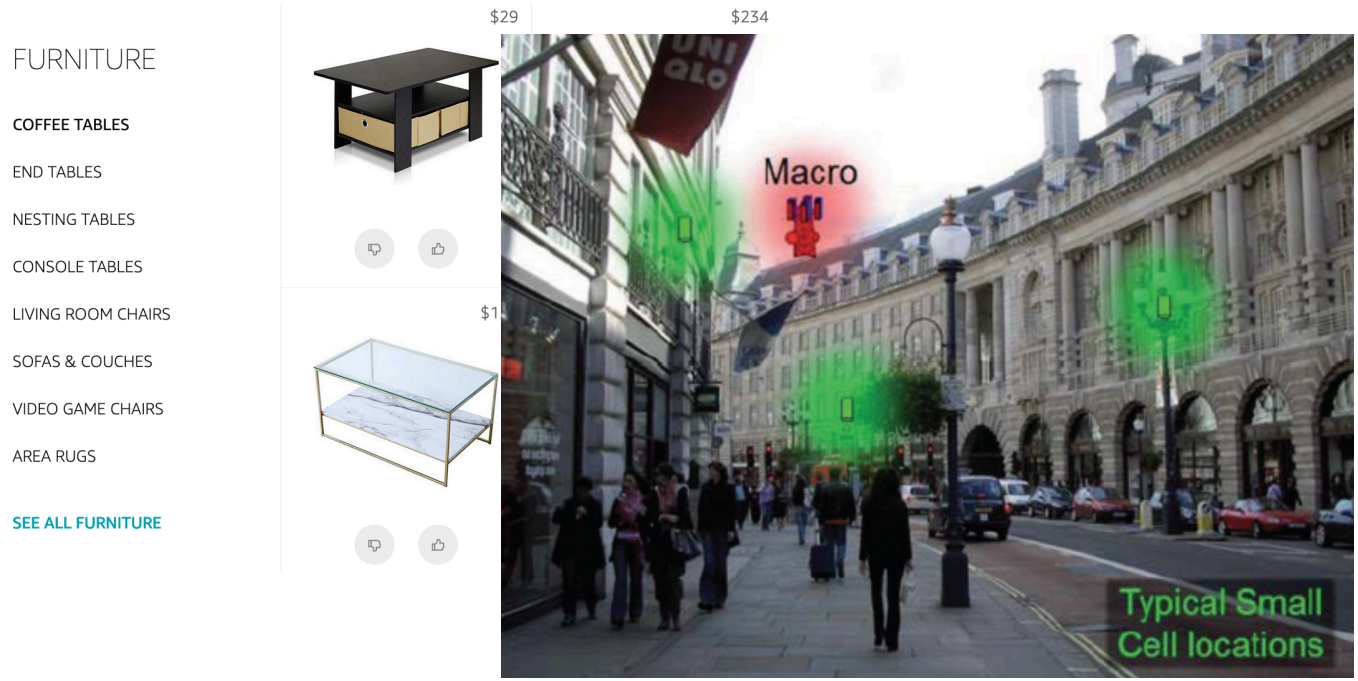
Wish to learn '*most similar*' or '*closest*' *items to a given from noisy measurements*




amazon.com/discover

# Motivation

Wish to learn '*most similar*' or '*closest*' *items to a given from noisy measurements*




Fujitsu white paper

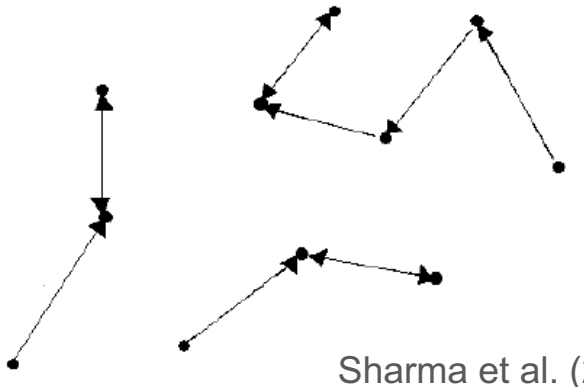
We don't know the given a priori. We want to answer *'closest'* queries for any item quickly!

# The Nearest Neighbor Graph Problem

$\mathcal{X} = \{x_1, \ldots, x_n\}$ is a set of $n$ points with unknown distance function $d(\cdot, \cdot)$. In *few* queries to a noisy distance oracle $\mathrm{Q}(\cdot, \cdot)$, learn

$$x_{j^*} := \arg \min_{x \in \mathcal{X} \setminus \{x_j\}} d(x_j, x) \qquad \forall j \in [n],$$

that are all correct with probability at least $1 - \delta$.



Sharma et al. (2015)

# Preliminaries and Notation

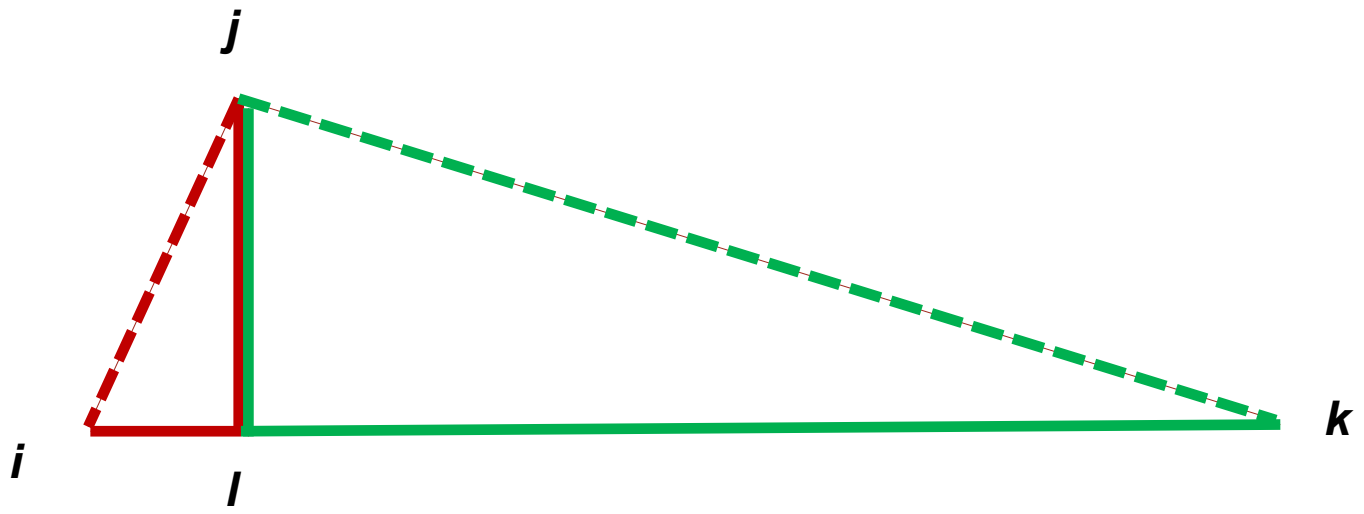
- $\mathcal{X} := \{x_i\}_{i=1}^n$ and $x_{j^*} := \min_{x \in \mathcal{X} \setminus x_j} d(x_j, x)$
- $\mathrm{Q}(i, j)$ yields a realization of $d(x_i, x_j) + \eta$ where $\eta$ is a $1$-sub-Gaussian random variable
- $\Delta_{i,j} := d(x_i, x_j) - d(x_i, x_{i^*})$

# Outline of ANNTri

- Iterate over $\{x_1, \ldots, x_n\}$ in order and find $x_{j^*}$ correctly w.p. $1 - \delta/n$ in the $j^{\text{th}}$ round.
- $x_{j^*}$ is found by successive elimination (`SETri`).
- For any $\ell < j$, both $\mathsf{Q}(\ell, j)$ and $\mathsf{Q}(j, \ell)$ follow the same law, so *reuse* samples (and associated bounds) from $\ell^{\text{th}}$ round while finding $x_{j^*}$.
- Using confidence intervals for $d(x_\ell, x_j)$, $d(x_\ell, x_k)$ and *triangle inequality*: we can bound $d(x_j, x_k)$.

# Elimination via the triangle inequality

$$d_{i,j} \leq d_{i,l} + d_{j,l} \text{ and } d_{j,k} \geq |d_{k,l'} - d_{j,l'}|,$$
$$\text{so } d_{i,l} + d_{j,l} < |d_{k,l'} - d_{j,l'}| \implies x_k \neq x_{j^*}$$

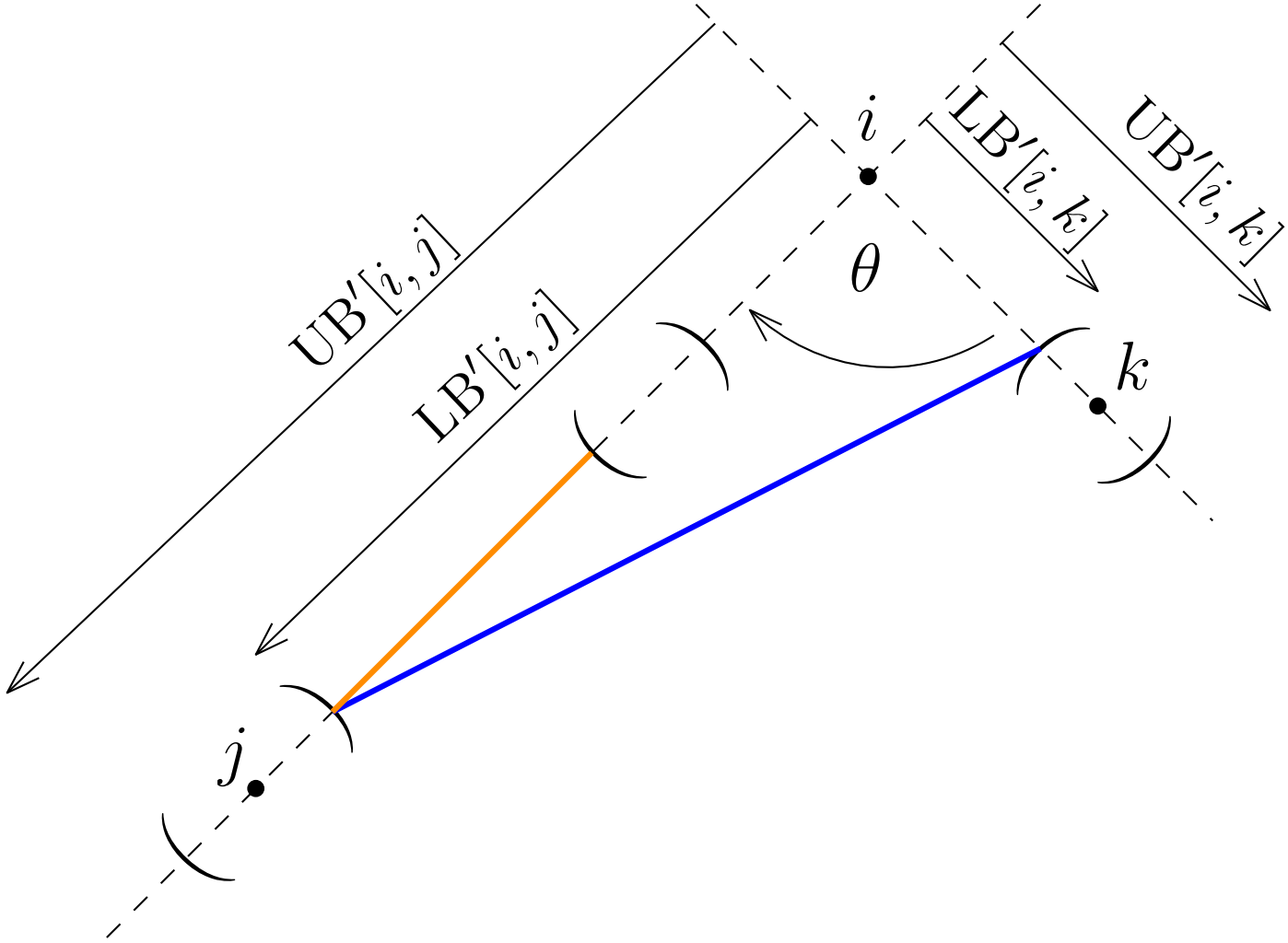


# Triangle Inequality: Upper Bounds

$$d_{i,k} \leq d_{i,l} + d_{l,k}$$

$$U_{i,k}^{\triangle \ell}(t) := \min_{\max\{\ell_1,\ell_2\}<\ell} \Big( \min\{U_{\ell,i}(t), U_{\ell,i}^{\triangle \ell_1}(t)\} + \min\{U_{\ell,k}(t), U_{\ell,k}^{\triangle \ell_2}(t)\} \Big)$$

# Triangle Inequality: Lower Bounds



# Theoretical Results

**Theorem 1.** *In the good event when all the conf. intervals are valid, which occurs w.p.* $1-\delta$, *simplified* `ANNTri` *learns the* NN *graph by making*

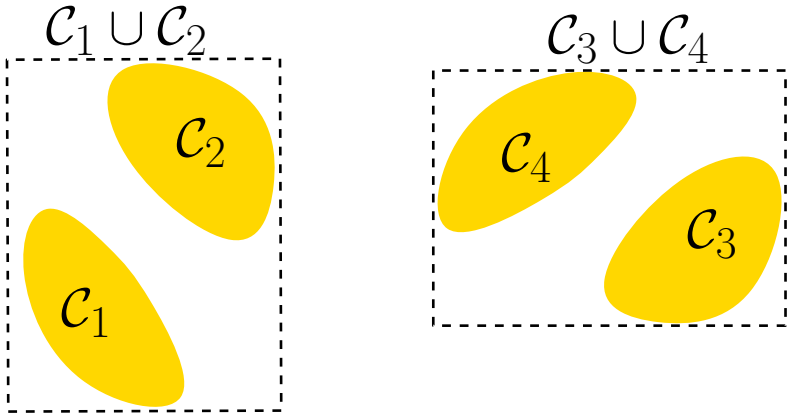
$$\mathcal{O}\left(\sum_{j=1}^{n}\sum_{k>j}\mathbb{1}_{[A_{j,k}]}H_{j,k}+\sum_{k<j}\mathbb{1}_{[A_{j,k}]}(H_{j,k}-\mathbb{1}_{[A_{k,j}]}H_{k,j})_{+}\right)$$

*queries, where* $\Delta_{j,k}:=d(x_j,x_k)-d(x_j,x_{j^*})$, $H_{j,k}:=\log(n^2/(\delta\Delta_{j,k}))\Delta_{j,k}^{-2}$, *and* $\mathbb{1}_{[A_{j,k}]}=0$ *if* $\exists i<j$ *such that* $6C_{\delta/n}(1)\le d_{i,k}-2d_{i,j}$ *and*

$$\{j,k\}\cap(\cup_{m<i}\{\ell:2d_{m,i}<d_{m,\ell}\})=\emptyset.$$

# Theoretical Results

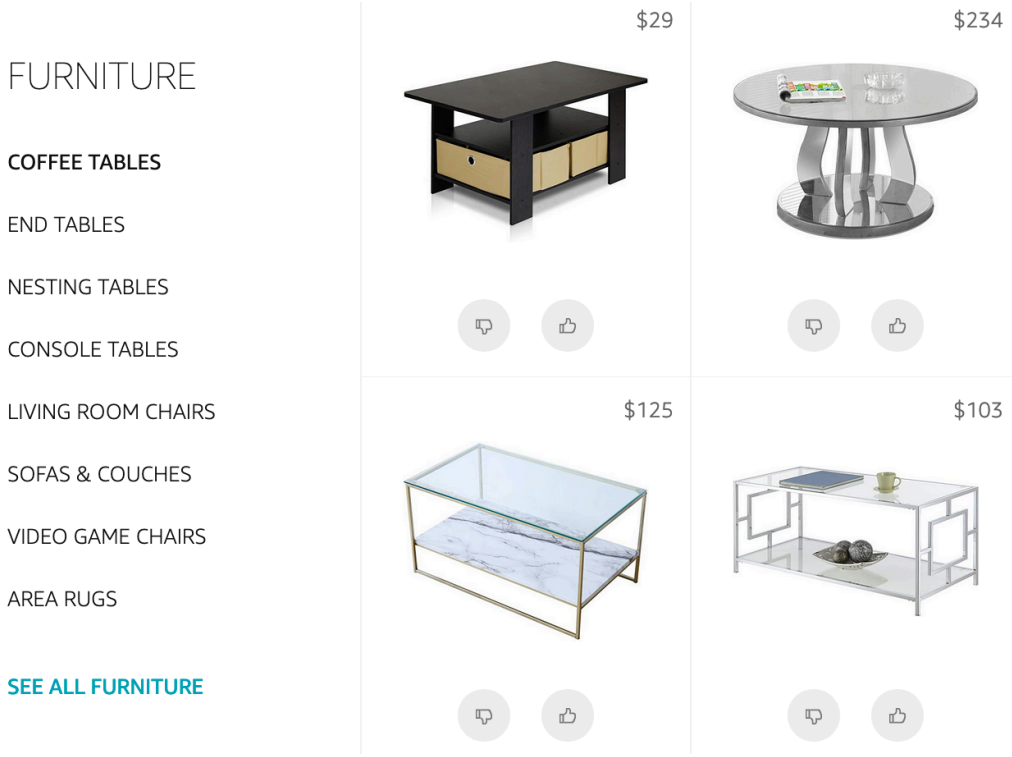
- Often, we can do better:

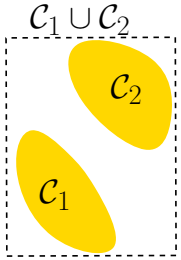


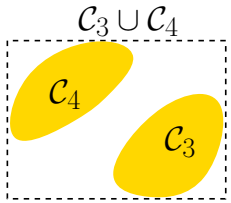

$$\{x_k : d_{i,k} < 6C_{\delta/n}(1) + 2d_{i,j}\} \subseteq \mathcal{C}_m \; \forall i, j \in \mathcal{C}_m$$

# Theoretical Results

- An example of separation:



# Theoretical Results

**Theorem 2:** For hierarchical datasets of $\nu$ clusters, `ANNTri` learns the correct nearest neighbor graph in

$$\mathcal{O}(n \log(n) \overline{\Delta^{-2}})$$

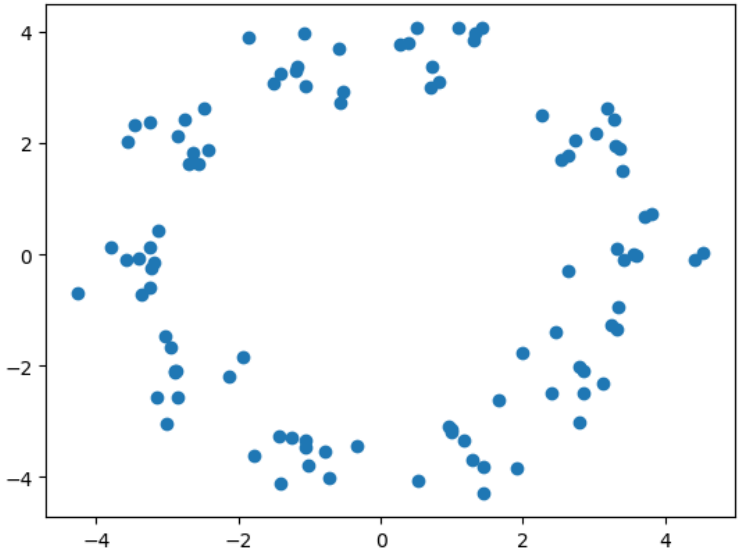
samples where

$$\overline{\Delta^{-2}} := \frac{1}{n\nu} \sum_{i=1}^{n/\nu} \sum_{j,k \in \mathcal{C}_i} \log(n^2/(\delta \Delta_{j,k})) \Delta_{j,k}^{-2}$$

is the average number of samples between nearby points and is due to the noise.

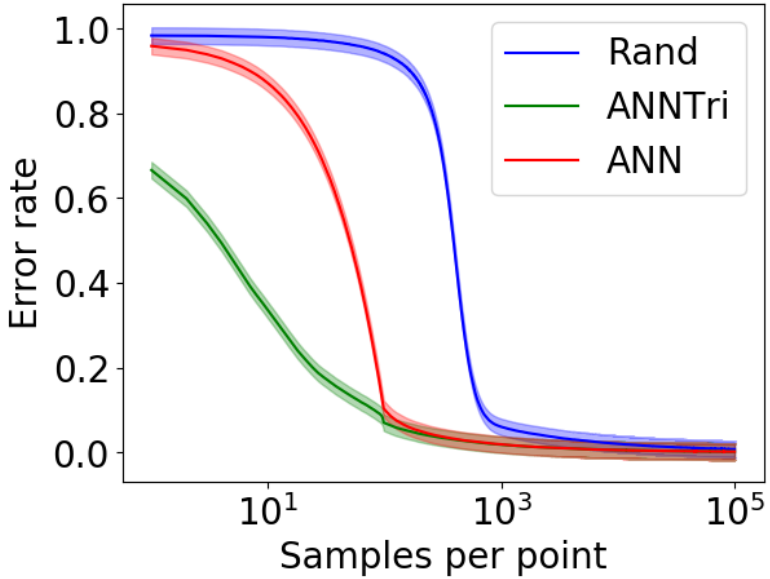
# Experimental Results

- Simulated data



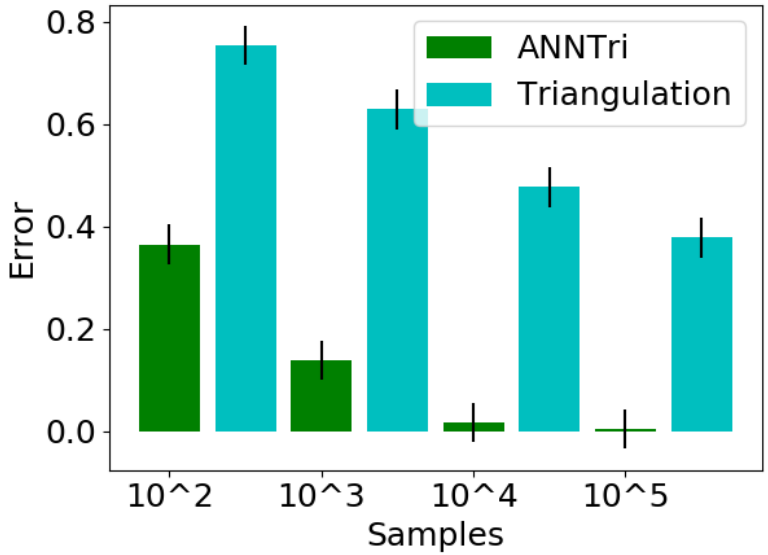

- 100 points in $\mathbb{R}^2$
- 10 clusters of 10 points
- Euclidean distance
- Gaussian noise, $\sigma^2 = 0.1$

# Experimental Results




- Compare against Random sampling
- Test effect of triangle inequality

# Experimental Results




- The metric is (2d) Euclidean
- We can compare against (distance) matrix completion
- With a distance matrix, the graph can be computed easily

# Experimental Results

- What shoes are most similar?
- 85 images from UTZappos50K dataset
- Human judgements collected by Heim et al., (2015).

# Experimental Results

Click on the two most similar shoes

# Experimental Results

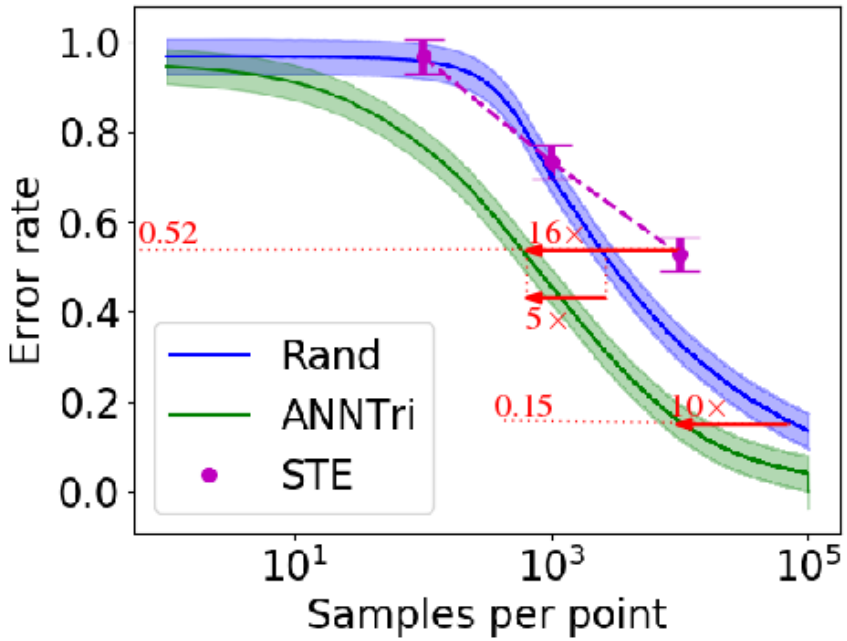
- What shoes are most similar?
- 85 images from UTZappos50K dataset
- Human judgements collected by Heim et al., (2015).

Click on the two most similar shoes

$d_{i,j} := \mathbb{E}_{k \sim \text{Unif}(\mathcal{X} \setminus \{i,j\})} \mathbb{E}[\mathbf{1}_{E^j_{i,k}} | k]$,
the probability that $i, j$ are *not* chosen when queried with random $k$.

# Experimental Results

- What shoes are most similar?
- 85 images from UTZappos50K dataset
- Human judgements collected by Heim et al., (2015).

# Main takeways for ANNTri

1. ANNTri finds the nearest neighbor graph for general metrics using the triangle inequality

2. Only requires access to noisy oracle

3. In favorable settings, requires $\boldsymbol{O(nlog(n)\,\Delta^{-2})}$ queries versus $\boldsymbol{O(n^2\Delta^{-2})}$ needed by brute force!