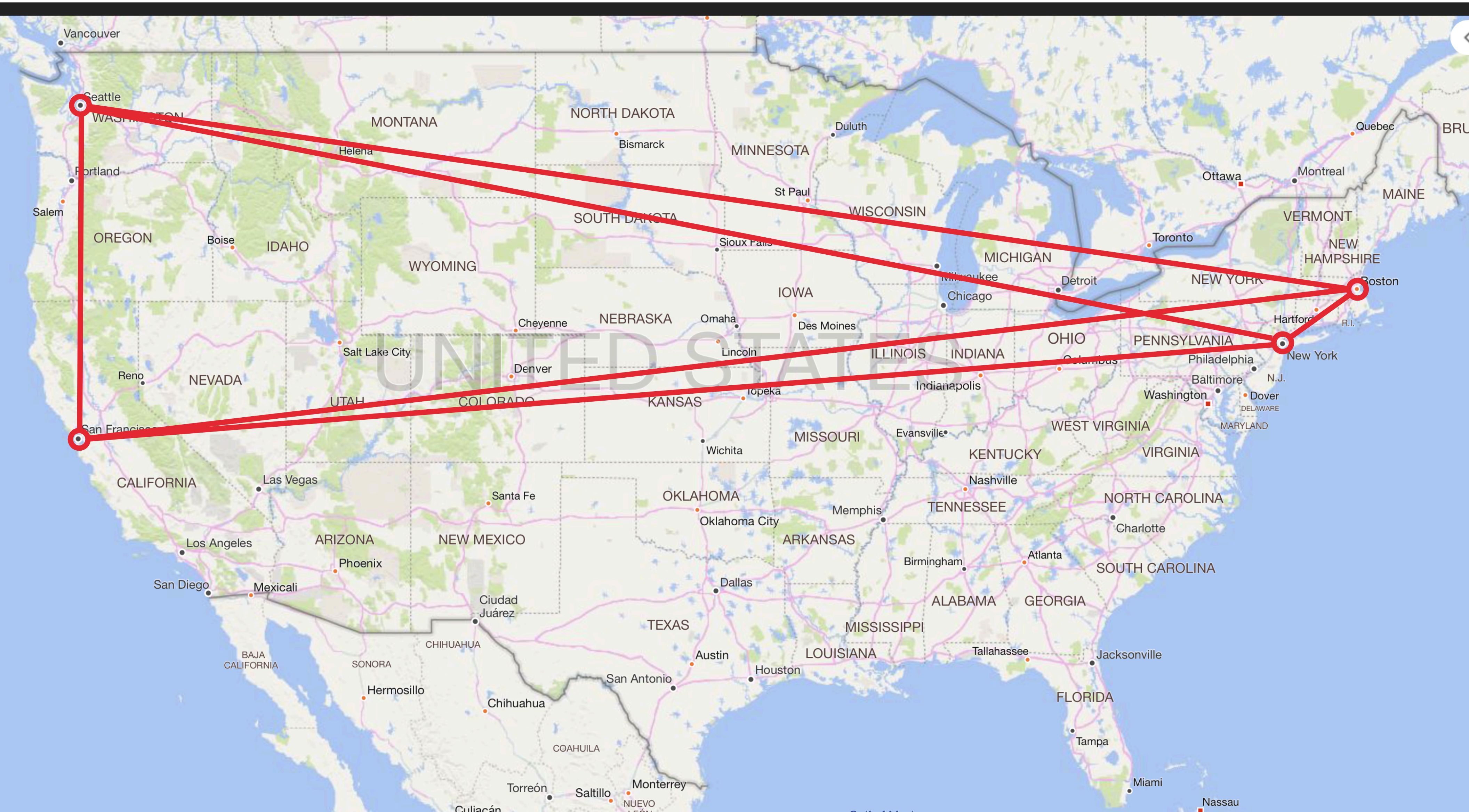
JESSE ANDERTON

ADVISOR: JAVED ASLAM
COMMITTEE MEMBERS: FERNANDO DIAZ, DAVID SMITH, BYRON WALLACE

SCALABLE ORDINAL EMBEDDING TO MODEL USER BEHAVIOR


UNITED STATES
Vancouver
Seattle
WASHINGTON
Portland
Salem
OREGON
Boise
IDAHO
MONTANA
Helena
NORTH DAKOTA
Bismarck
MINNESOTA
Duluth
St Paul
WISCONSIN
SOUTH DAKOTA
Sioux Falls
WYOMING
IOWA
Omaha
Des Moines
Cheyenne
NEBRASKA
Lincoln
Salt Lake City
Denver
Reno
NEVADA
UTAH
COLORADO
KANSAS
Topeka
Wichita
San Francisco
CALIFORNIA
Las Vegas
Santa Fe
OKLAHOMA
Oklahoma City
Los Angeles
ARIZONA
NEW MEXICO
Phoenix
San Diego
Mexicali
Ciudad Juárez
TEXAS
Dallas
Austin
San Antonio
Houston
BAJA CALIFORNIA
SONORA
CHIHUAHUA
Hermosillo
Chihuahua
COAHUILA
Torreón
Saltillo
Monterrey
NUEVO LEÓN
Culiacán
Gulf of Mexico
MISSOURI
ARKANSAS
Memphis
LOUISIANA
MISSISSIPPI
ALABAMA
Birmingham
GEORGIA
Atlanta
Tallahassee
FLORIDA
Jacksonville
Tampa
Miami
Nassau
TENNESSEE
Nashville
KENTUCKY
Evansville
ILLINOIS
INDIANA
Indianapolis
Chicago
Milwaukee
MICHIGAN
Detroit
OHIO
Columbus
WEST VIRGINIA
VIRGINIA
NORTH CAROLINA
Charlotte
SOUTH CAROLINA
PENNSYLVANIA
Philadelphia
Baltimore
Washington
Dover
DELAWARE
MARYLAND
N.J.
NEW YORK
New York
Hartford
R.I.
Boston
NEW HAMPSHIRE
VERMONT
MAINE
Toronto
Ottawa
Montreal
Quebec
BRU


Vancouver
Seattle
Portland
Salem
OREGON
Boise
IDAHO
MONTANA
Helena
NORTH DAKOTA
Bismarck
MINNESOTA
Duluth
St Paul
WISCONSIN
SOUTH DAKOTA
Sioux Falls
WYOMING
MICHIGAN
Milwaukee
Detroit
Chicago
Toronto
Ottawa
Montreal
Quebec
MAINE
VERMONT
NEW HAMPSHIRE
NEW YORK
Boston
Hartford
R.I.
IOWA
Des Moines
Omaha
NEBRASKA
Cheyenne
Salt Lake City
UNITED STATES
Lincoln
ILLINOIS
INDIANA
OHIO
PENNSYLVANIA
New York
Philadelphia
Reno
NEVADA
Denver
UTAH
COLORADO
KANSAS
Topeka
Indianapolis
Baltimore
N.J.
Washington
Dover
DELAWARE
MARYLAND
San Francisco
WEST VIRGINIA
MISSOURI
Evansville
KENTUCKY
VIRGINIA
Wichita
CALIFORNIA
Las Vegas
Nashville
Santa Fe
OKLAHOMA
NORTH CAROLINA
Memphis
TENNESSEE
Oklahoma City
Charlotte
Los Angeles
ARIZONA
NEW MEXICO
ARKANSAS
Phoenix
Birmingham
Atlanta
SOUTH CAROLINA
San Diego
Mexicali
Dallas
Ciudad Juárez
ALABAMA
GEORGIA
TEXAS
MISSISSIPPI
CHIHUAHUA
Austin
LOUISIANA
Tallahassee
Jacksonville
BAJA CALIFORNIA
SONORA
Houston
San Antonio
Hermosillo
Chihuahua
FLORIDA
Tampa
COAHUILA
Miami
Torreón
Saltillo
Monterrey
NUEVO LEÓN
Nassau
Culiacán
Gulf of Mexico

PAIRWISE CITY DISTANCES

	Boston	NYC	Seattle	SF
Boston	–	190	2,485	2,692
NYC	–	–	2,401	2,565
Seattle	–	–	–	679
SF	–	–	–	–

TOTAL DISTANCE ORDER

	Boston	NYC	Seattle	SF
Boston	–	1st	4th	6th
NYC	–	–	3rd	5th
Seattle	–	–	–	2nd
SF	–	–	–	–

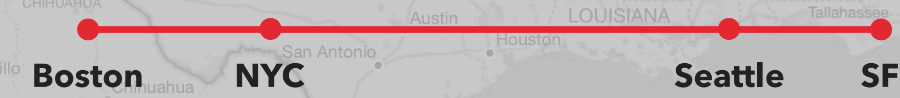
DISTANCE RANKINGS

Anchor	Boston	NYC	Seattle	SF
Boston	–	1st	2nd	3rd
NYC	1st	–	2nd	3rd
Seattle	3rd	2nd	–	1st
SF	3rd	2nd	1st	–

DISTANCE RANKINGS

Anchor	Boston	NYC	Seattle	SF
Boston	–	1st	2nd	3rd
NYC	1st	–	2nd	3rd
Seattle	3rd	2nd	–	1st
SF	3rd	2nd	1st	–

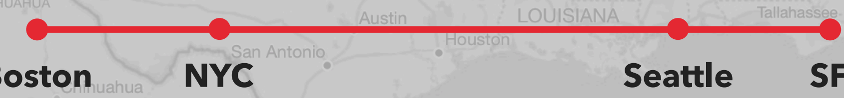
Perfect?



DISTANCE RANKINGS

Anchor	Boston	NYC	Seattle	SF	Dallas
Boston	–	1st	3rd	4th	2nd
NYC	1st	–	3rd	4th	2nd
Seattle	4th	3rd	–	1st	2nd
SF	4th	3rd	1st	–	2nd
Dallas	3rd	1st	4th	2nd	–

Perfect? No!



ASSIGNING ORDER-PRESERVING POSITIONS

- An *embedding* positions a set of objects within some vector space (like $\mathbb{R}^d$) to satisfy some objective.
- An *ordinal embedding* focuses on satisfying some given ordering constraints.
- Constraints can be expressed as triples like:

“Boston is closer to New York City than to Seattle”

“The Matrix is more like Star Wars than it is like La La Land”

“People who like steak tend to prefer chicken over tofu”

EVALUATE BY RANK CORRELATION

Mean Kendall's τ – Mean rank correlation across anchors

Mean τ_{AP} – Mean top-heavy rank correlation across anchors

GROUND TRUTH RANKINGS

Anchor	Boston	NYC	Seattle	SF
Boston	–	1st	2nd	3rd
NYC	1st	–	2nd	3rd
Seattle	3rd	2nd	–	1st
SF	3rd	2nd	1st	–

EMBEDDING RANKINGS

	Boston	NYC	Seattle	SF
Boston	–	1st	3rd	2nd
NYC	1st	–	2nd	3rd
Seattle	1st	2nd	–	3rd
SF	3rd	1st	2nd	–

HUMAN-BASED PREFERENCE/SIMILARITY

- Easier for assessors to say "The Matrix is more like Star Wars than it is like La La Land."
- Focus on lab studies/crowdsourcing limits research interest in scalability.
- Limited scalability prohibits focus on similarity expressed through logged user behavior.

ORDINAL EMBEDDING OF FACES
TAMUZ ET AL., ICML 2011

[3] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai, "Adaptively Learning the Crowd Kernel," ICML, 2011.

IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

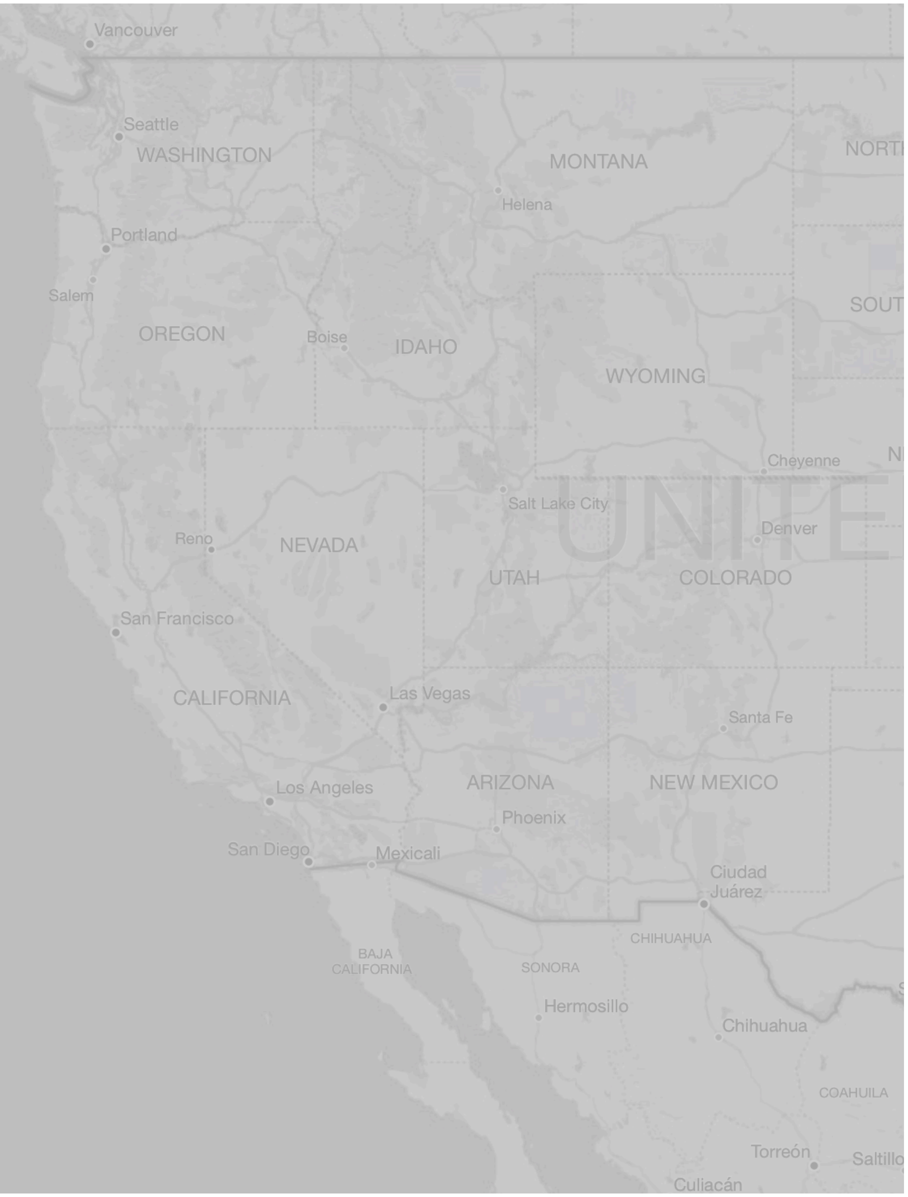
HOW MANY COMPARISONS TO LEARN ALL RANKINGS?

"a IS MORE LIKE b THAN LIKE c" $\Rightarrow \delta_{ab} < \delta_{ac} \Rightarrow$ TRIPLE (a, b, c)

- $O(n^3)$ total triples (with n total objects).
- $O(n^2 \log n)$ triples to get all rankings.
- $O(d\, n \log n)$ triples if a perfect embedding exists in $\mathbb{R}^d$ (we think)
- On a limited budget, we want to adaptively pick next triples to improve the embedding the most.

DISTANCE RANKINGS

Anchor	Boston	NYC	Seattle	SF
Boston	-	1st	2nd	3rd
NYC	1st	-	2nd	3rd
Seattle	3rd	2nd	-	1st
SF	3rd	2nd	1st	-



RELATED WORK

CROWD KERNEL
ICML 2011

ICML 2011: "ADAPTIVELY LEARNING THE CROWD KERNEL" [T,B,S,K]

- By "kernel" they mean "embedding."
- Assumes that assessors disagree more when similar distances are compared.
- They pick triples that (approximately) maximize expected information gain.
- Model uses an intermediate embedding to find triples where (a,b,c) and (a,c,b) are both likely.

Prob. that assessor says $\delta_{ab} < \delta_{ac}$

$$Pr((a,b,c)|X) = \frac{\lambda + \delta_{ac}^2(X)}{2\lambda + \delta_{ab}^2(X) + \delta_{ac}^2(X)}$$

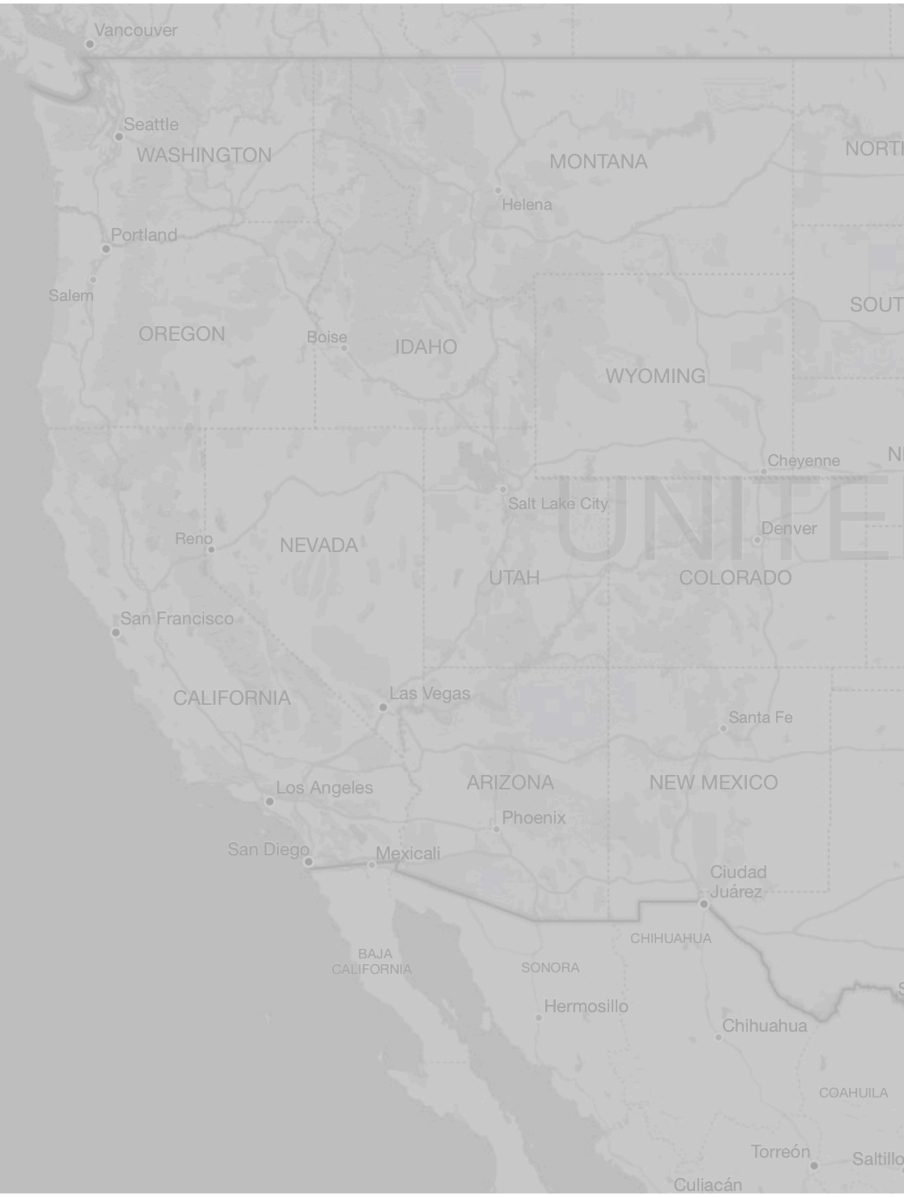
$\delta_{ab}(X)$	$\delta_{ac}(X)$	$Pr((a,b,c)\|X)$
1	2	0.75
2	1	0.25
1.4	1.5	0.53
1.5	1.5	0.50

[3] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai, "Adaptively Learning the Crowd Kernel," ICML, 2011.

SCORE CARD: CROWD KERNEL

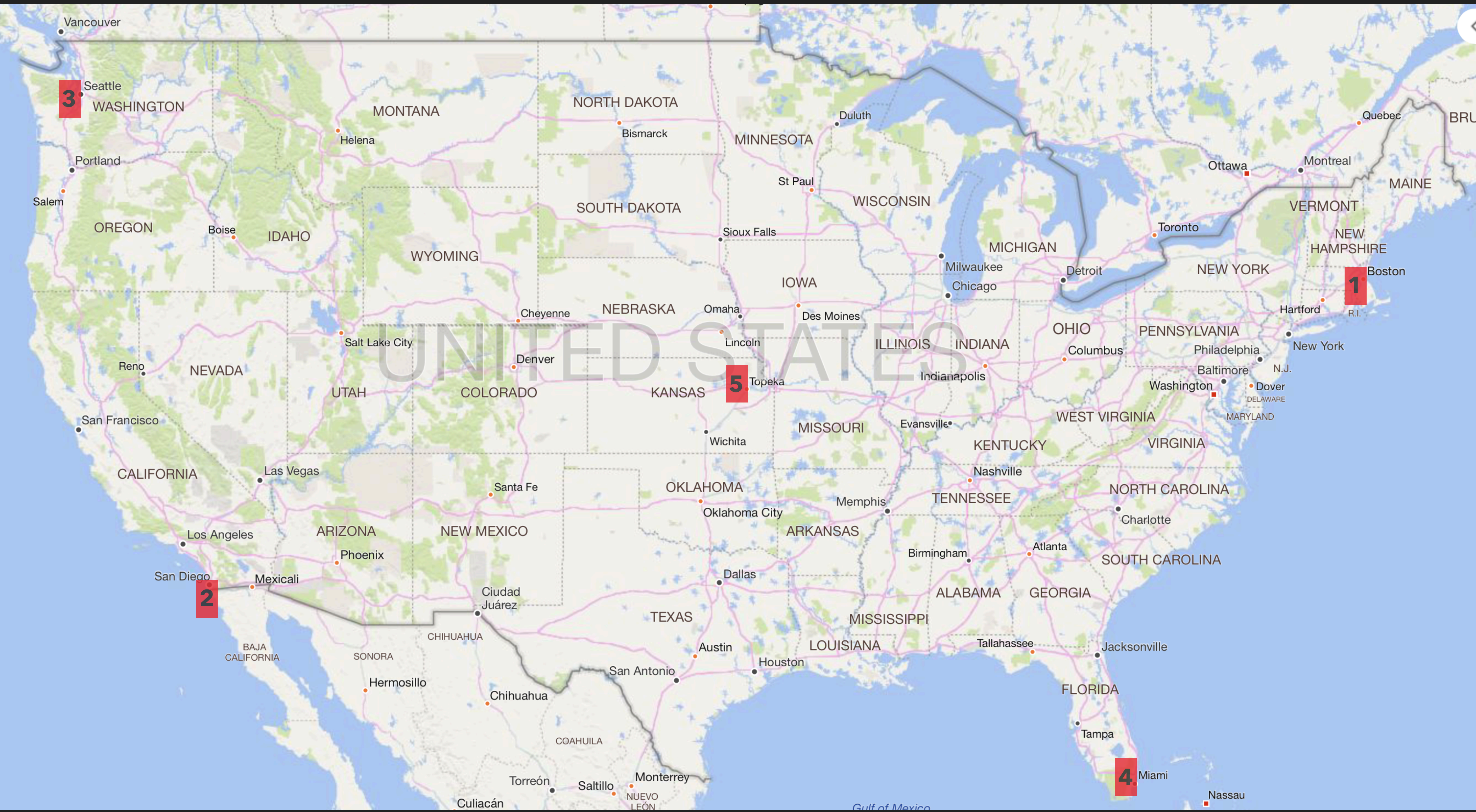
After a year trying to use this tool, I decided to write a thesis on better tools.

	CK
Active Learning	🥉 Good for small budgets
Num. Objects	🥉 Hundreds
Num. Dimensions	🥉 <10
Accuracy	🥉 Medium
Speed	🐌 Prohibitively Slow



MY METHOD

FRFT ADAPTIVE SORT



1
2
3
4
5
UNITED STATES
Vancouver
Seattle
WASHINGTON
Portland
Salem
OREGON
Boise
IDAHO
MONTANA
Helena
NORTH DAKOTA
Bismarck
MINNESOTA
Duluth
St Paul
WISCONSIN
SOUTH DAKOTA
Sioux Falls
WYOMING
IOWA
Des Moines
NEBRASKA
Omaha
Lincoln
Cheyenne
Salt Lake City
Denver
NEVADA
Reno
UTAH
COLORADO
KANSAS
Topeka
Wichita
MISSOURI
ILLINOIS
Chicago
Milwaukee
MICHIGAN
Detroit
INDIANA
Indianapolis
Evansville
OHIO
Columbus
Toronto
Ottawa
Montreal
Quebec
MAINE
VERMONT
NEW HAMPSHIRE
NEW YORK
Boston
Hartford
R.I.
New York
PENNSYLVANIA
Philadelphia
N.J.
Baltimore
Dover
DELAWARE
MARYLAND
Washington
WEST VIRGINIA
VIRGINIA
KENTUCKY
Nashville
TENNESSEE
NORTH CAROLINA
Charlotte
SOUTH CAROLINA
Atlanta
GEORGIA
ALABAMA
Birmingham
MISSISSIPPI
Memphis
ARKANSAS
OKLAHOMA
Oklahoma City
TEXAS
Dallas
Austin
San Antonio
Houston
LOUISIANA
Tallahassee
Jacksonville
FLORIDA
Tampa
Miami
Nassau
Gulf of Mexico
San Francisco
CALIFORNIA
Las Vegas
Los Angeles
San Diego
Mexicali
ARIZONA
Phoenix
NEW MEXICO
Santa Fe
Ciudad Juárez
CHIHUAHUA
Chihuahua
BAJA CALIFORNIA
SONORA
Hermosillo
COAHUILA
Torreón
Saltillo
Monterrey
NUEVO LEÓN
Culiacán

FARTHEST-RANK-FIRST TRAVERSAL ADAPTIVE SORT

1. Pick an anchor far from all previous anchors (first time: use a point on boundary).

2. Guess the anchor's ranking using an embedding of data collected so far.

3. Sort the guessed ranking adaptively: O(n) triples if guess was good, O(n log n) if guess was bad.

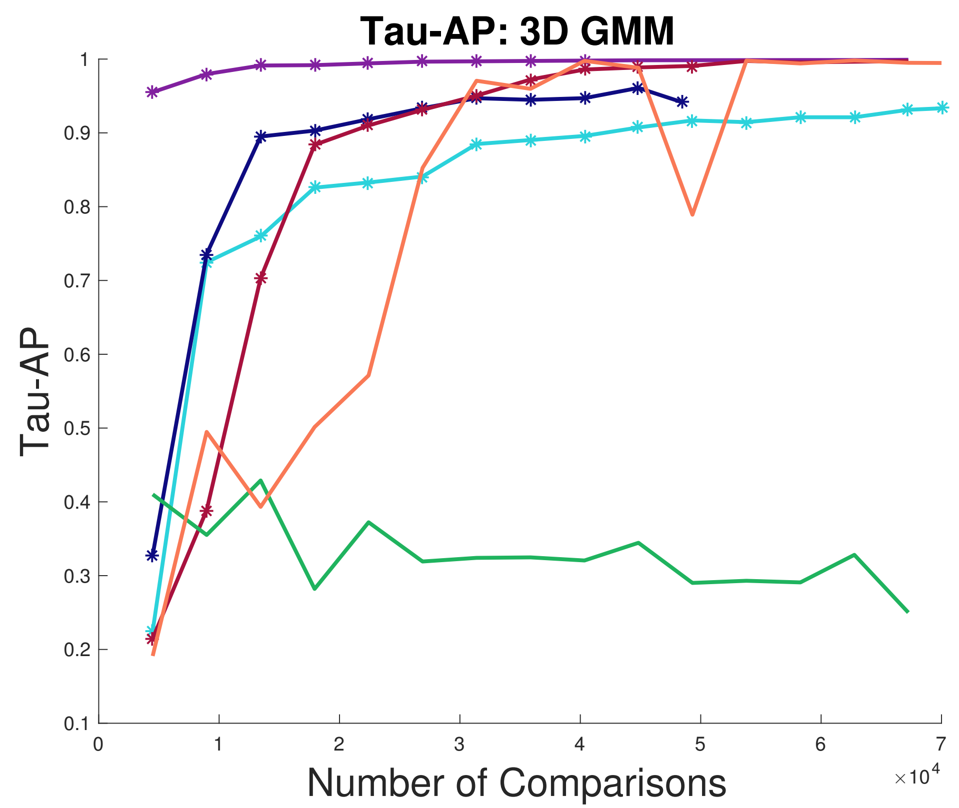
4. If guess was very good, stop; else, go to 1.

[8] J. Anderton, V. Pavlu, J. Aslam, "Triple Selection for Ordinal Embedding," unpublished, 2016.

EMPIRICAL COMPARISON

τ_{AP} IS A TOP-HEAVY RANK CORRELATION MEASURE

- FRFT Ranking – My algorithm, using rankings from features – O(n) triples per ranking.
- FRFT Adaptive Sort – My algorithm, using no prior knowledge – O(n log n) then O(n).
- Crowd Kernel – Active learning baseline.
- Random Tails – Random baseline.
- kNN – Gradually add next NN for each obj.
- Landmarks – Gradually add objects to all rankings.

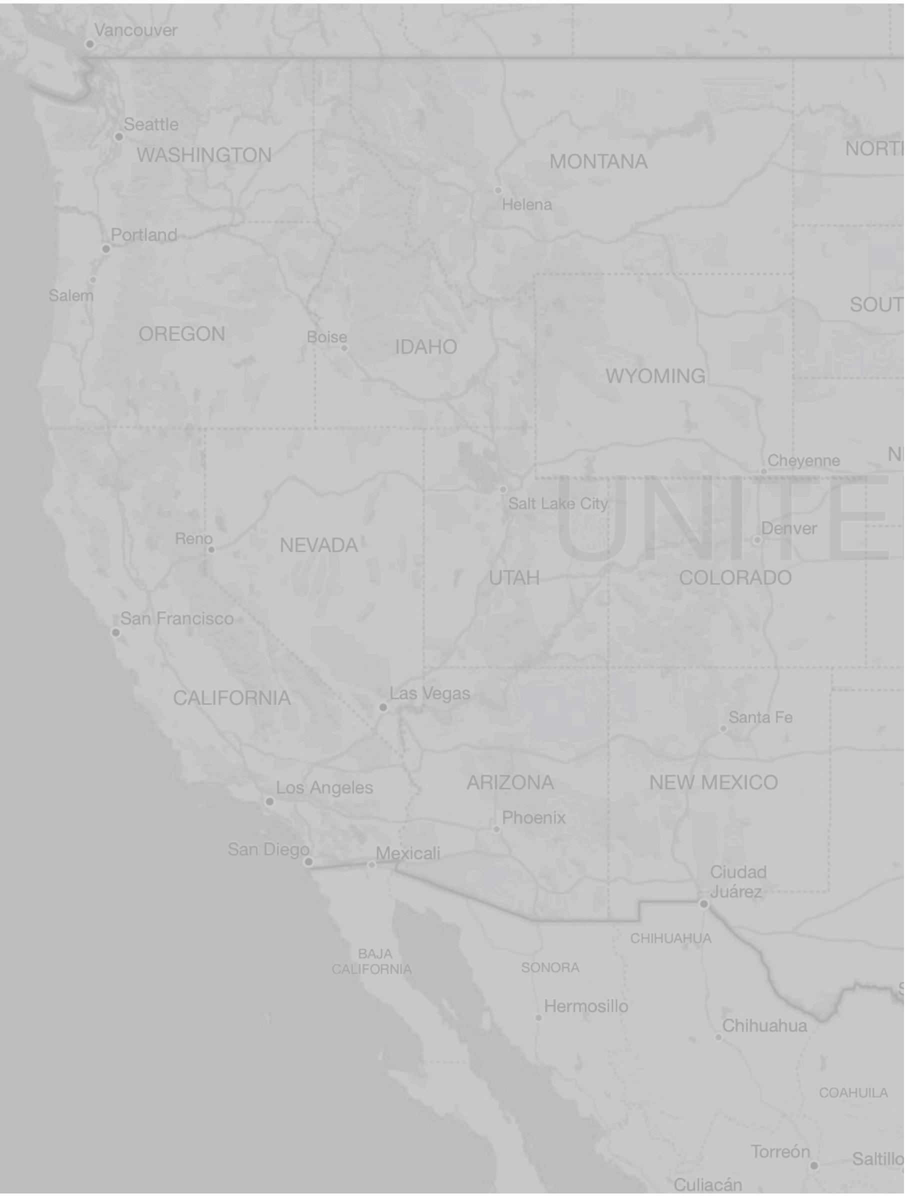


[8] J. Anderton, V. Pavlu, J. Aslam, "Triple Selection for Ordinal Embedding," unpublished, 2016.

SCORE CARD: FRFT ADAPTIVE SORT

Active learning beats CK, but we still have work to do.

	CK	AS	
Active Learning	🥉	🥈	Approaches lower bound
Num. Objects	🥉	🥈	10,000’s
Num. Dimensions	🥉	🥉	<10
Accuracy	🥉	🥈	Very good
Speed	🐌	🐇	Medium



PROPOSED WORK

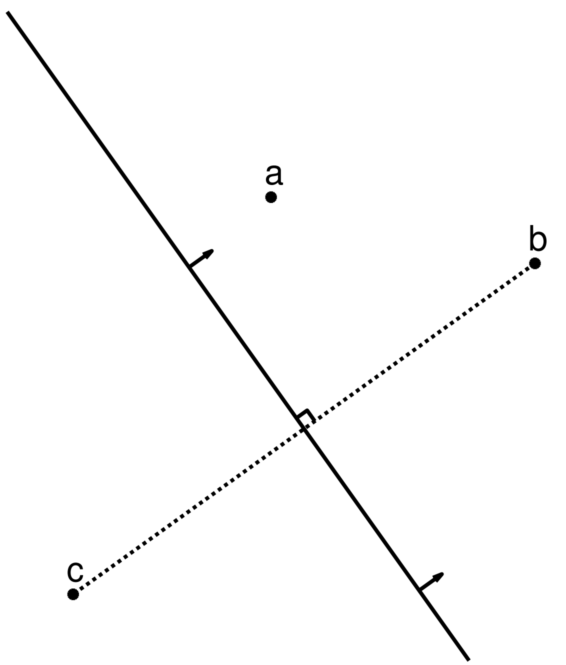
CAN WE DO BETTER?

- Empirically, FRFT Adaptive Sort approaches the lower bound [4] of $\Omega(\mathrm{d\ n\ log\ n})$.
- Intermediate embedding step is slow and error-prone.
- When our guess is already correct, we still waste (?) triples to confirm it.
- I believe we can avoid the embedding step and reduce redundancy using the geometry implied by the triples.

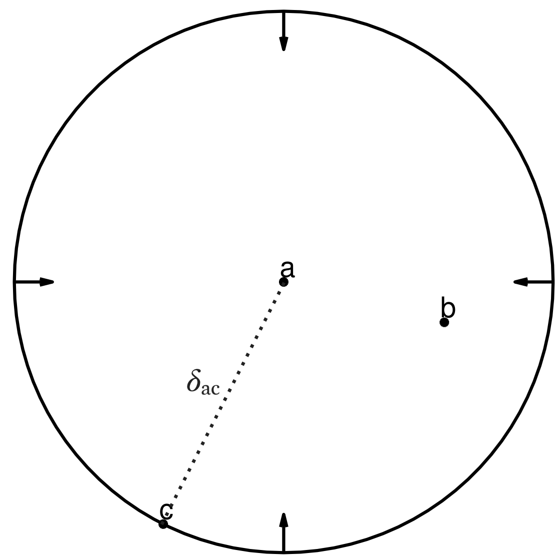
[4] K. G. Jamieson and R. D. Nowak, Low-dimensional embedding using adaptively selected ordinal data. IEEE, 2011, pp. 1077–1084.

THE THREE VIEWS OF A “TRIPLE CONSTRAINT” a IS MORE LIKE b THAN c: (a,b,c)

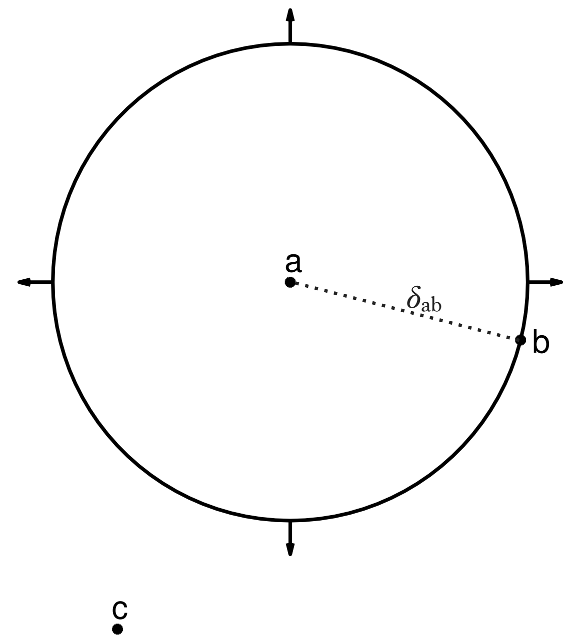
$$\delta_{ab} < \delta_{ac}$$



a IS INSIDE A HALF-SPACE



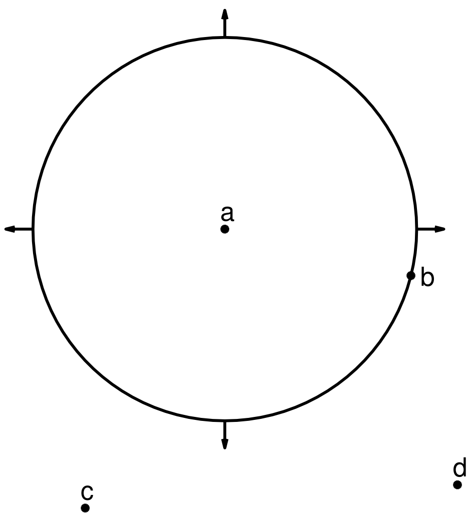
b IS INSIDE A SPHERE



c IS OUTSIDE A SPHERE

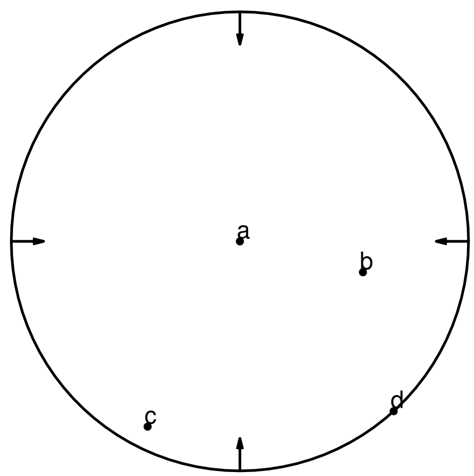
COMBINING TRIPLE CONSTRAINTS

$$\delta_{ab} < \delta_{ac} < \delta_{ad}$$



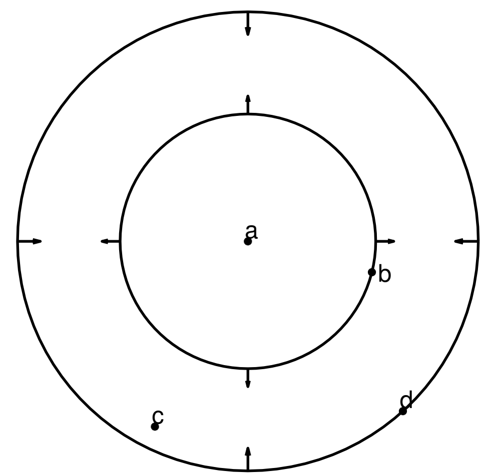
$\mathbf{c}$, $\mathbf{d}$ ARE OUTSIDE A SPHERE

$\wedge$



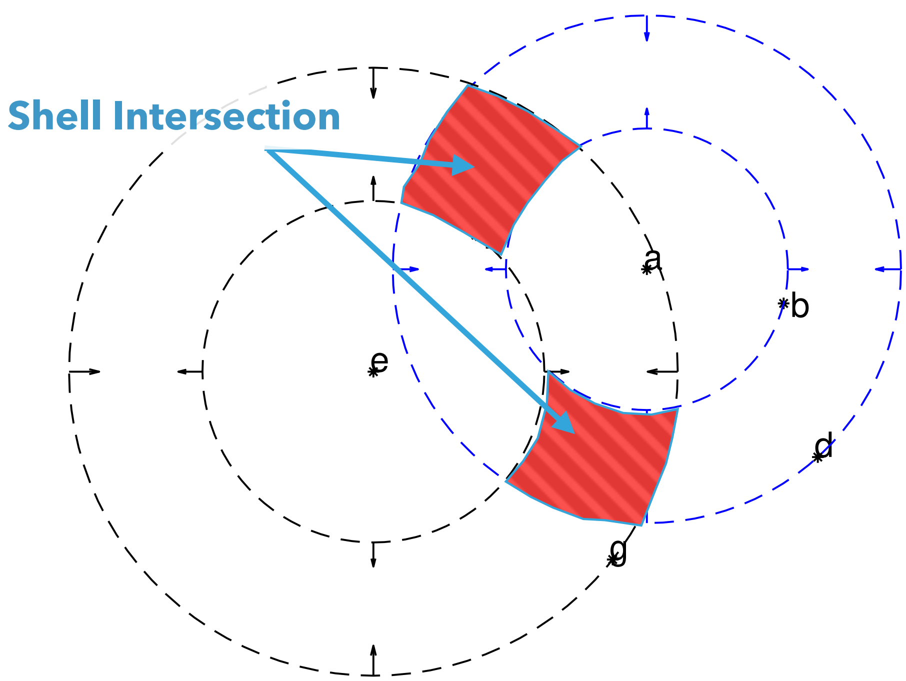
$\mathbf{b}$, $\mathbf{c}$ ARE INSIDE A SPHERE

$\Rightarrow$

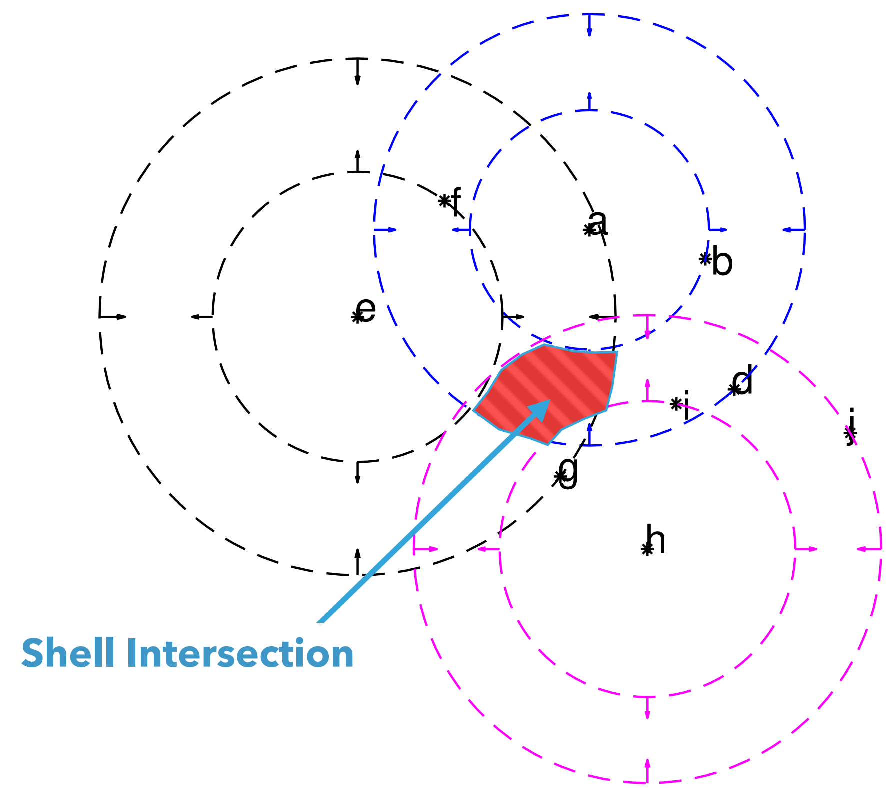


$\mathbf{c}$ IS INSIDE A SPHERICAL SHELL

COMBINING SPHERICAL SHELLS

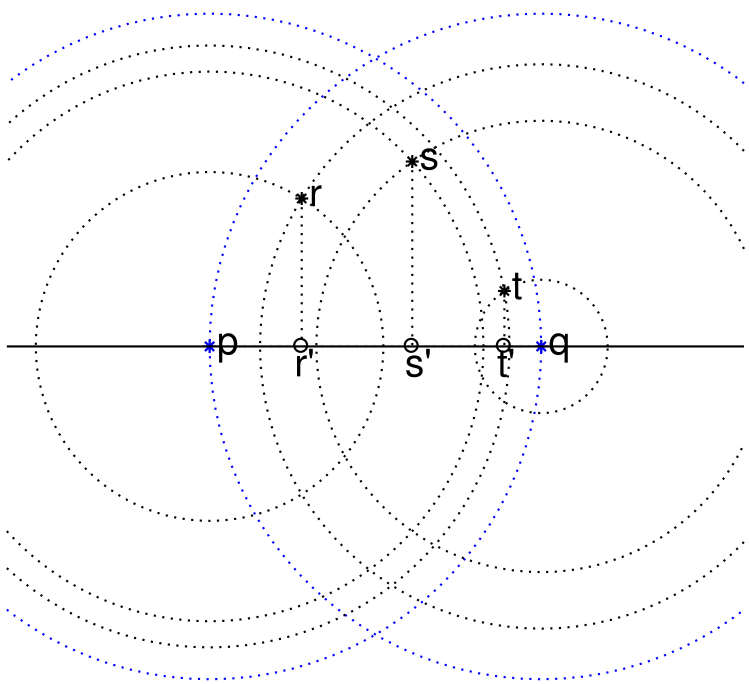


TWO SHELLS IN R^2

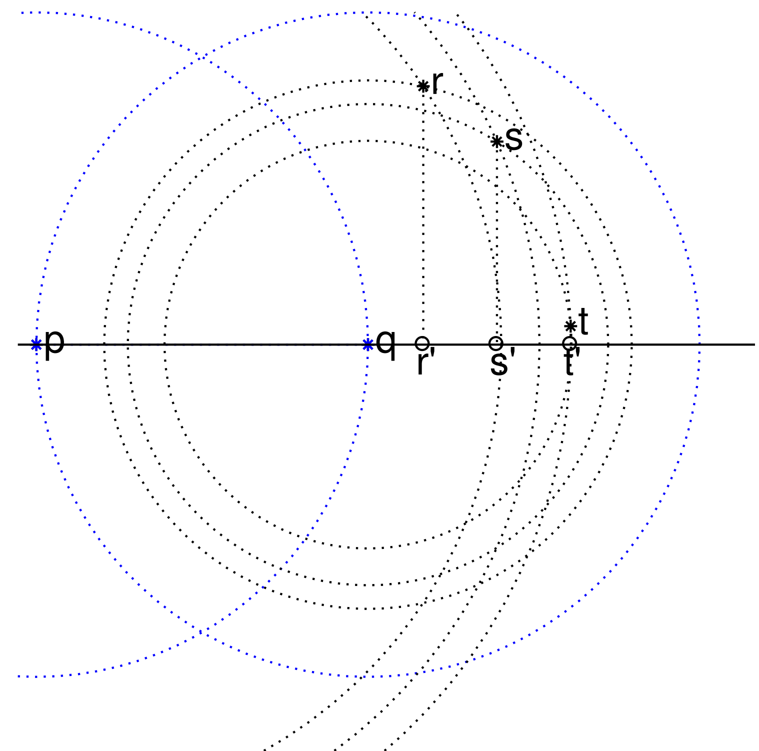


THREE SHELLS IN R^2

PARTIAL ORDERING ON VECTOR PROJECTIONS



INFERRING ORDER IN BLUE BALL INTERSECTION
P, R', S', T', Q



INFERRING ORDER NEAR BLUE BALL INTERSECTION
P, Q, R', S', T'

GUESSING ORDER WITH LINE PROJECTION

- Line projection preserves approximate order.[6]
- Rankings for a pair of points gives partial order of projections onto their connecting line.
- Idea: Don't waste time on intermediate embedding; guess order by majority vote of partial orders!

[6] K. Li and J. Malik, "Fast k-Nearest Neighbour Search via Dynamic Continuous Indexing," ICML, 2016.

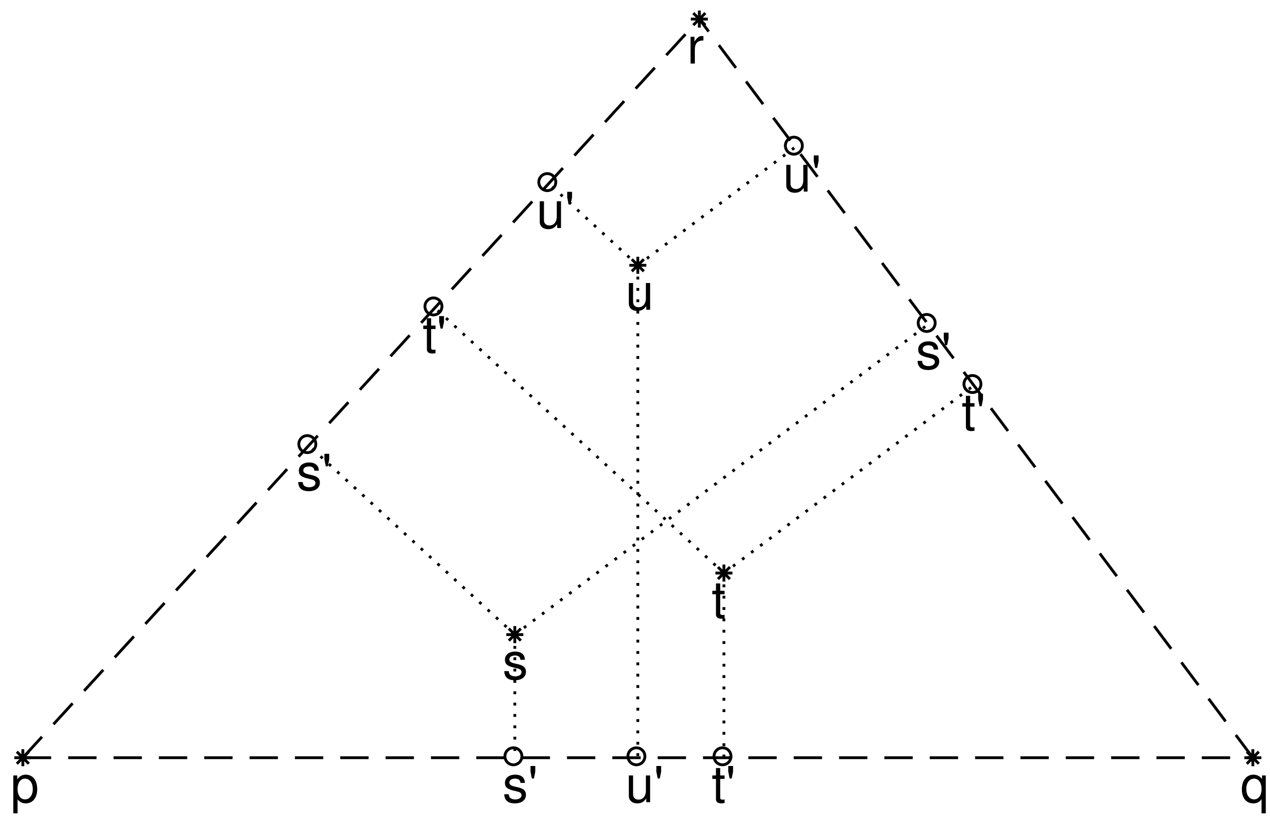
GUESSING ORDER WITH LINE PROJECTION

TWO RANKINGS

Point	NN	Maj. Vote
s	t	u (1/1)
t	s	u (1/1)
u	t	t (1/1)

THREE RANKINGS

Point	NN	Maj. Vote
s	t	t (2/3)
t	s	u (2/3)
u	t	t (2/3)



IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

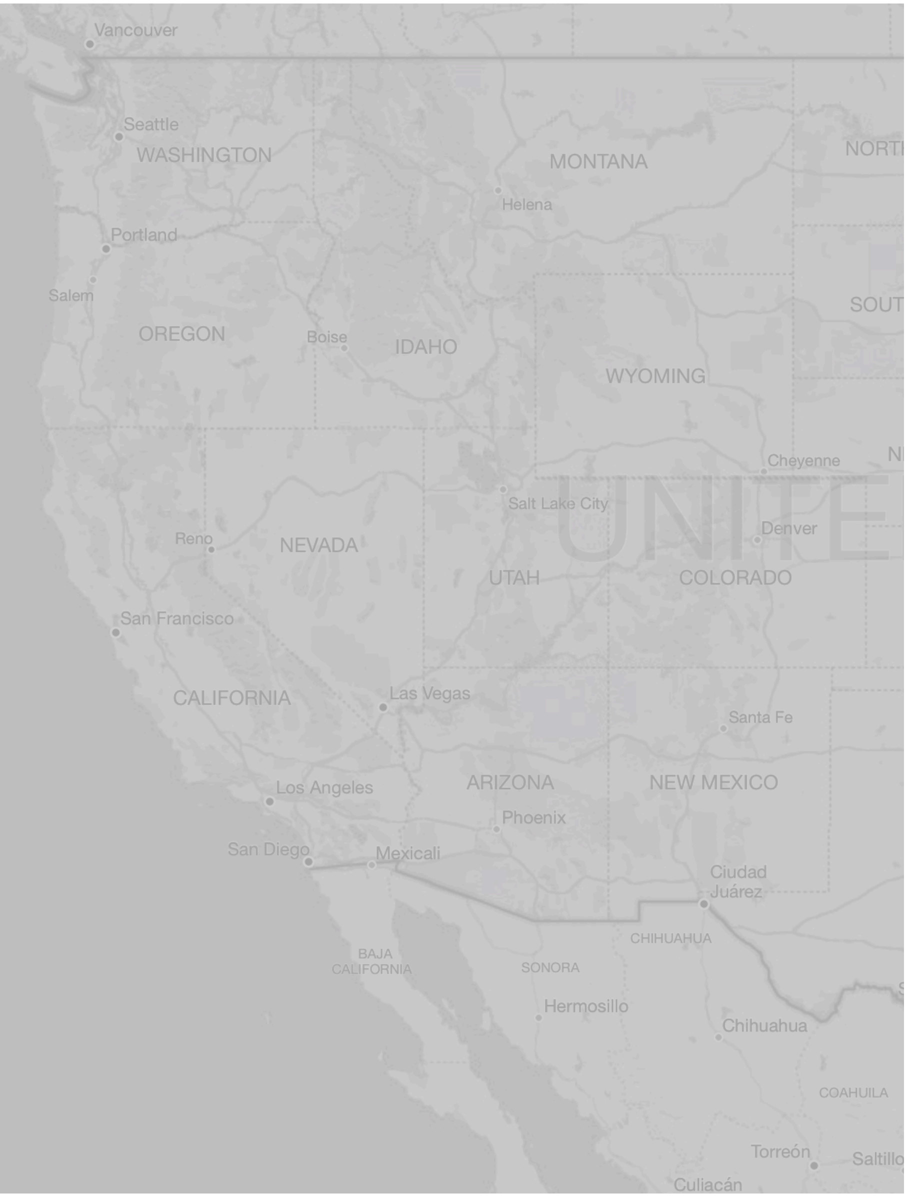
IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

FROM TRIPLES TO EMBEDDINGS

- Given a set of triples and target space $\mathbb{R}^d$, how can we find an embedding?
- A hard non-convex optimization problem.
- No known algorithm for large, high dimensional datasets.
- State-of-the-art example is Soft Ordinal Embedding[5].
- Larger sets can be handled by merging SOE embeddings[7].

[5] Y. Terada and U. von Luxburg, "Local ordinal embedding," ICML, 2014.
[7] M. Cucuringu and J. Woodworth, "Point Localization and Density Estimation from Ordinal kNN graphs using Synchronization," arXiv.org, 2015.



RELATED WORK

SOFT ORDINAL EMBEDDING ICML 2014

ICML 2014: SOFT ORDINAL EMBEDDING [T,VL]

- A triple (a,b,c) means $\delta_{ab} + \lambda < \delta_{ac}$; $\lambda > 0$ sets scale and prevents degenerate solutions.
- Can be minimized using standard optimizers.
- Works until n × d gets large (e.g. >100,000).

When embedding violates $\delta_{ab} + \lambda < \delta_{ac}$

$$Err_{soft}(X|d,\lambda) := \sum_{(a,b,c)\in T} \max\left[0, \delta_{ab}(X) + \lambda - \delta_{ac}(X)\right]^2$$

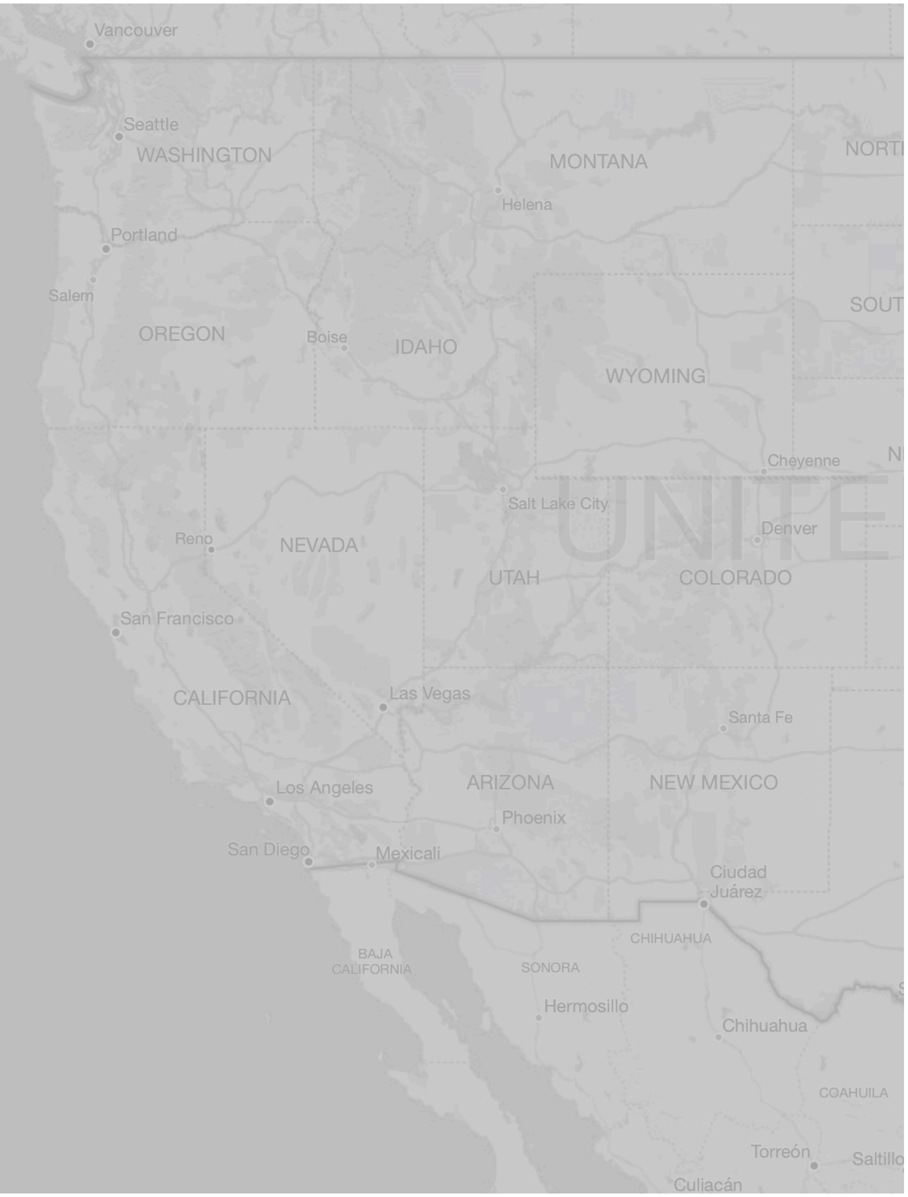
δ_{ab}	δ_{ac}	Err_{soft}
1	2	0.00
2	1	1.44
1.4	1.5	0.01
1.5	1.5	0.04

[5] Y. Terada and U. von Luxburg, “Local ordinal embedding,” ICML, 2014.

SCORE CARD: SOFT ORDINAL EMBEDDING

Current state-of-the-art, but requires restarts and can't handle high dimension.

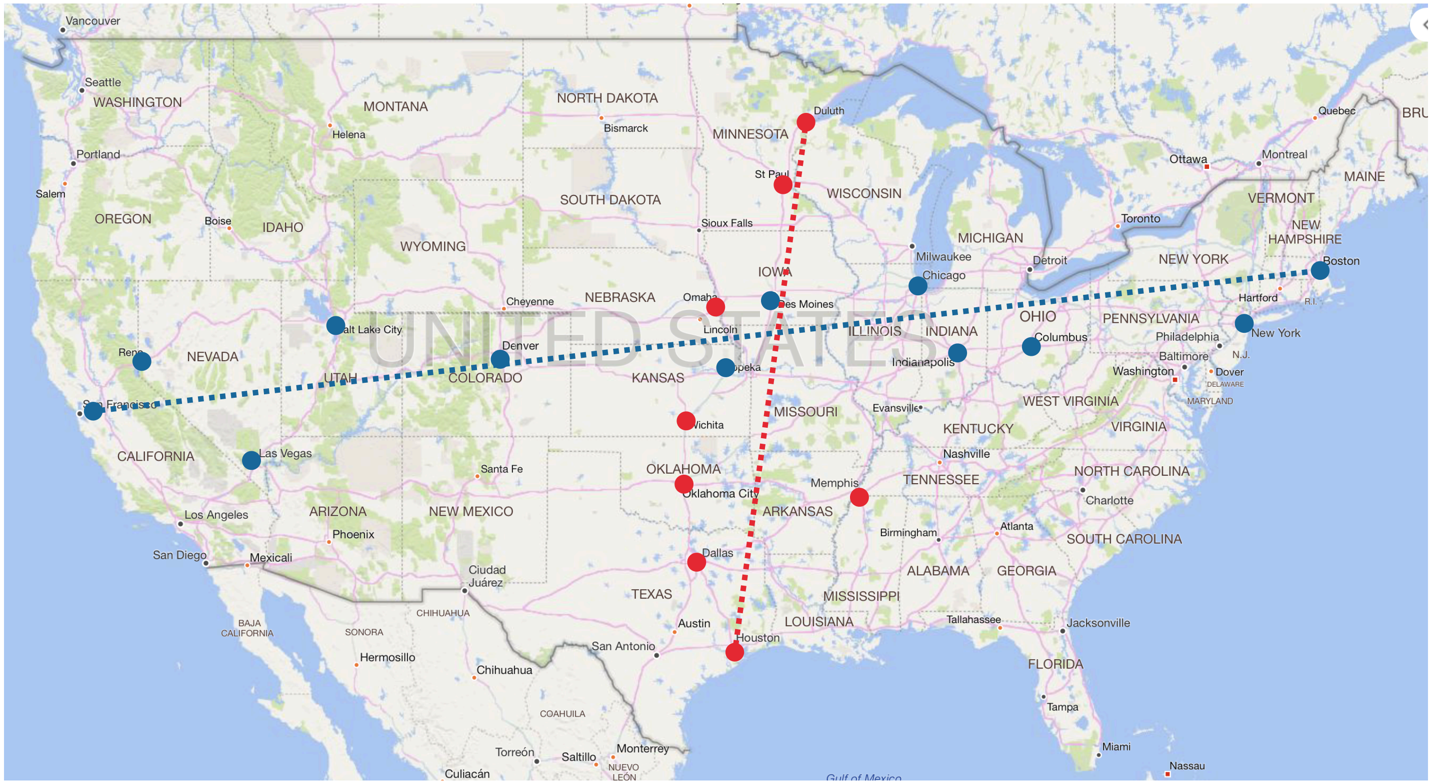
	CK	AS	SOE	
Active Learning	🥉	🥈	😅	N/A
Num. Objects	🥉	🥈	🥈	10,000's
Num. Dimensions	🥉	🥉	🥉	<10
Accuracy	🥉	🥈	🥈	High
Speed	🐌	🐇	🐇	Medium



MY METHOD

BASIS EMBEDDING

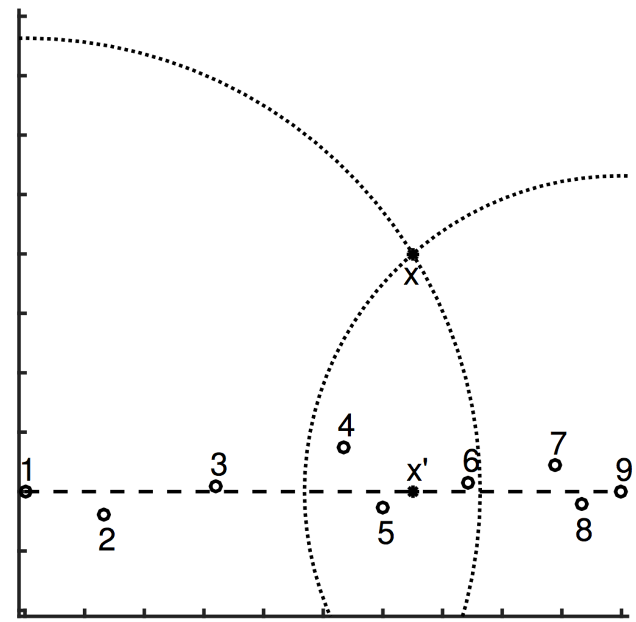
BASIS EMBEDDING (SUMMARY)



CHOOSING COORDINATES

- Pick line connecting pair of points as an "axis;" use points near line as "coordinates."
- The median "coordinate" point beneath a given point is its (approximate) position on the axis.
- We add axes until we can't find a point orthogonal to the existing axes.

X IS "ABOVE" 4, 5, AND 6;
WE CHOOSE 5 AS X'S COORDINATE ON THIS AXIS.



BASIS EMBEDDING: RESULTS

Table 2: Embedding Quality
* indicates global optimum was not found; means procedure computationally too expensive

Method	Dataset	d	$\hat{d}$	# Cmp.	τ
Basis	3dgmm	3	3	38K	0.71
Basis+SOE	3dgmm	3	3	38K	0.99
Extra+SOE	3dgmm	3	3	61K	**0.99**
Rand+SOE	3dgmm	3	3	38K	0.95
CK	3dgmm*	3	3	38K	-0.01
Basis	5dcube	5	3	39K	0.49
Basis+SOE	5dcube	5	6	39K	0.88
Extra+SOE	5dcube	5	6	61K	**0.94**
Rand+SOE	5dcube*	5	6	39K	0.61
CK	5dcube*	5	5	39K	0.01
Basis	5dgmm	5	3	39K	0.68
Basis+SOE	5dgmm	5	6	39K	0.94
Extra+SOE	5dgmm	5	6	62K	**0.98**
Rand+SOE	5dgmm*	5	6	39K	0.01
CK	5dgmm*	5	5	39K	-0.01

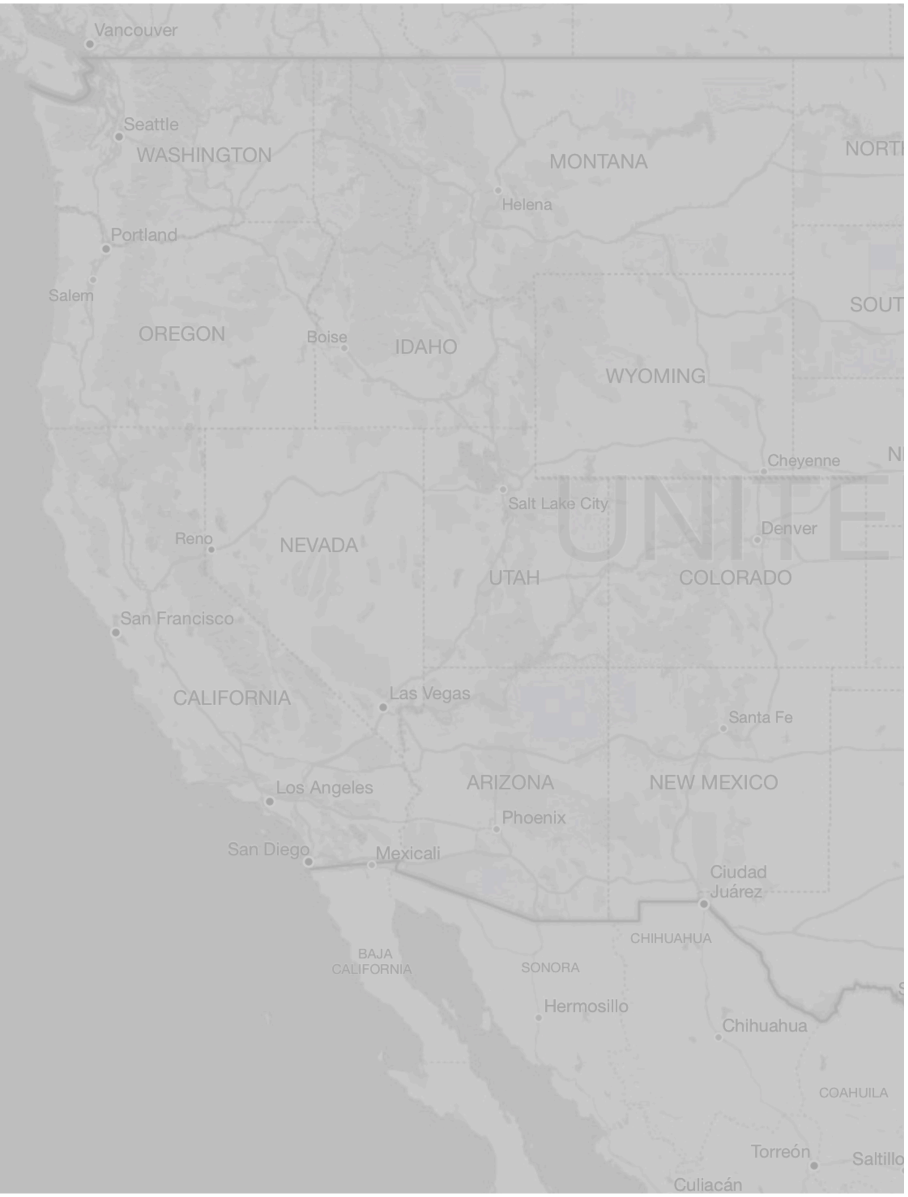
Method	Dataset	d	$\hat{d}$	# Cmp.	τ
Basis	20news	34K	3	186K	**0.11**
Basis+SOE	20news*	34K	6	186K	0.01
Extra+SOE	20news*	34K	6	310K	-0.01
Rand+SOE	20news*	34K	3	186K	0.01
CK	20news	34K	16	—	—
Basis	cities	3	2	28K	0.37
Basis+SOE	cities	3	4	28K	0.89
Extra+SOE	cities	3	4	50K	**0.96**
Rand+SOE	cities*	3	4	28K	0.01
CK	cities*	3	3	28K	0.01
Basis	digits	784	6	159K	0.52
Basis+SOE	digits*	784	12	159K	0.01
Extra+SOE	digits*	784	12	211K	0.01
Rand+SOE	digits*	784	12	159K	**0.73**
CK	digits	784	10	—	—
Basis	spam	57	3	85K	0.85
Basis+SOE	spam*	57	6	85K	-0.01
Extra+SOE	spam*	57	6	138K	0.01
Rand+SOE	spam	57	3	85K	**0.94**
CK	spam	57	10	—	—

[9] J. Anderton, V. Pavlu, J. Aslam, "Revealing the Basis: Ordinal Embedding through Geometry," unpublished, 2016.

SCORE CARD: BASIS EMBEDDING

First purely-geometric approach. Fast, reliable medium-quality embeddings.

	CK	AS	SOE	Basis	
Active Learning	🥉	🥈	😅	🥇	Meets lower bound
Num. Objects	🥉	🥈	🥈	🥇	Unlimited
Num. Dimensions	🥉	🥉	🥉	🥈	Nontrivial for high-dim
Accuracy	🥉	🥈	🥈	🥈	Medium but reliable
Speed	🐌	🐇	🐇	🚀	Very fast

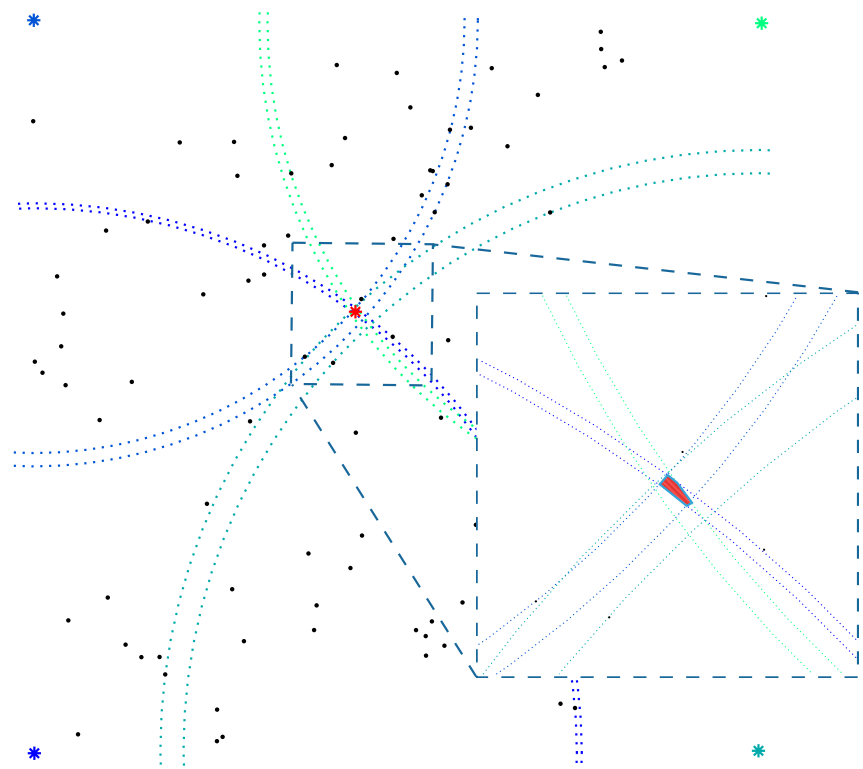


MY METHOD

SUBSET EMBEDDING

SUBSET EMBEDDING

- SOE can accurately embed small sets.
- Easy to embed with distances to known positions.
- So: embed a random subset with SOE, then use approximate distances to quickly embed remaining points.
- Makes an approximate embedding of a large set from a good embedding of a small set.



FAST APPROXIMATE EMBEDDING FROM A SUBSET

SUBSET EMBEDDING: EARLY RESULTS

- O(d n log m) when subset size m ≪ n: linear in n, and beats active learning lower bound!
- Needs further testing to explore limitations of method (noise sensitivity, insufficient dim.?)
- Want to prove quality bounds and explain quality theoretically.

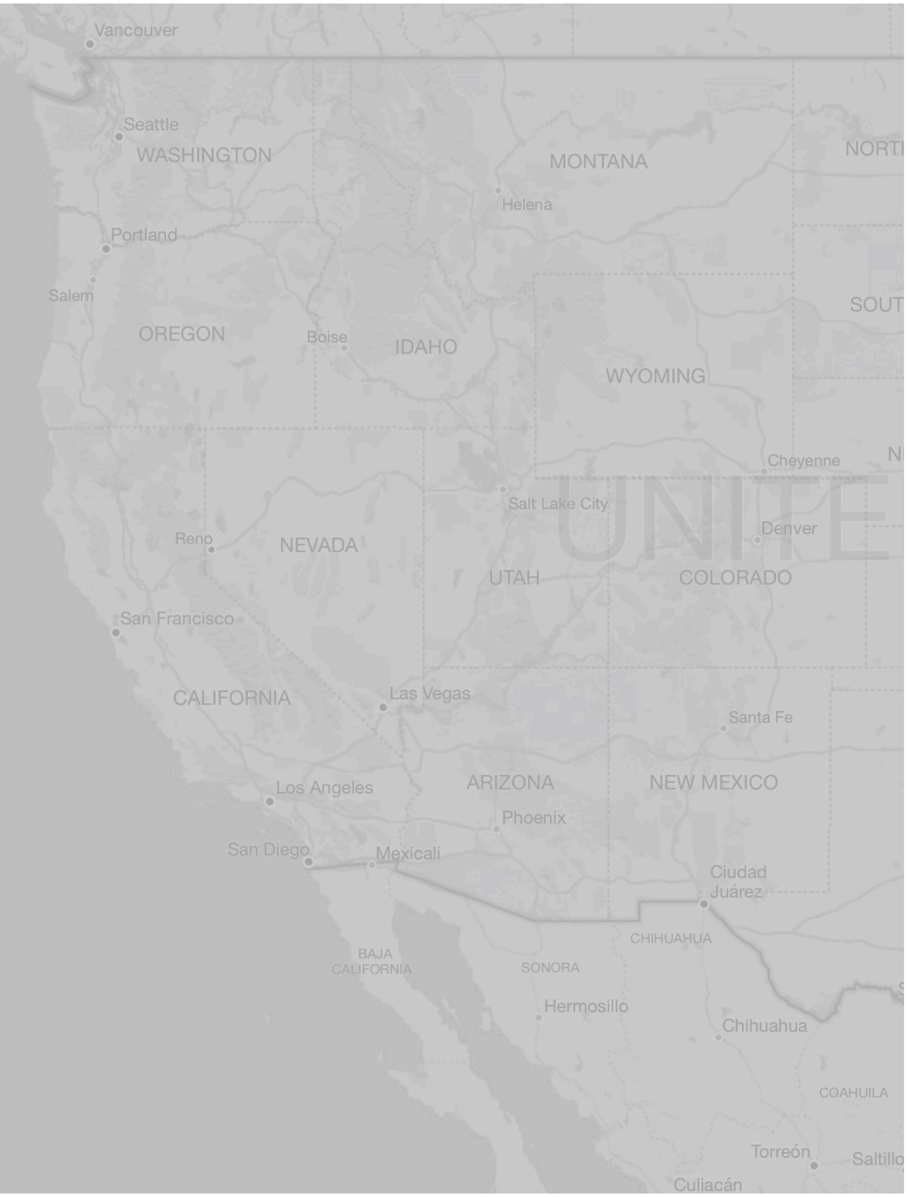
RESULTS ON SIMULATED AND REAL DATASETS. MEDIAN OF 10 RUNS.

Dataset	n	d	$\hat{d}$	τ
Ball	10K	3	3	0.99
Sphere	10K	3	3	0.99
Swiss Roll	10K	3	3	0.99
GMM	10K	3	3	0.99
Spambase	4.6K	57	3	0.89
Cities	15K	3	3	0.97
MNIST Digits	1K	784	12	0.57

SCORE CARD: SUBSET EMBEDDING

Fast, reliable high-quality embeddings. Sensitive to noise and limited dimensionality.

	CK	AS	SOE	Basis	Subset	
Active Learning	🥉	🥈	😅	🥇	🎖️	Beats lower bound!
Num. Objects	🥉	🥈	🥈	🥇	🥇	Unlimited
Num. Dimensions	🥉	🥉	🥉	🥈	🥉	Constrained by SOE
Accuracy	🥉	🥈	🥈	🥈	🥇	Highest; “approximate”
Speed	🐌	🐇	🐇	🚀	🚀	Linear in n!



PROPOSED WORK

CAN WE DO BETTER?

- Subset embedding is amazing but does not work in high dimension.
- Can we replace SOE in subset embedding with something more robust?
- Basis embedding is geometry-based but not great…
- Proposal: try to improve basis embedding using random vectors instead of "axes."

EMBEDDING WITH RANDOM VECTORS

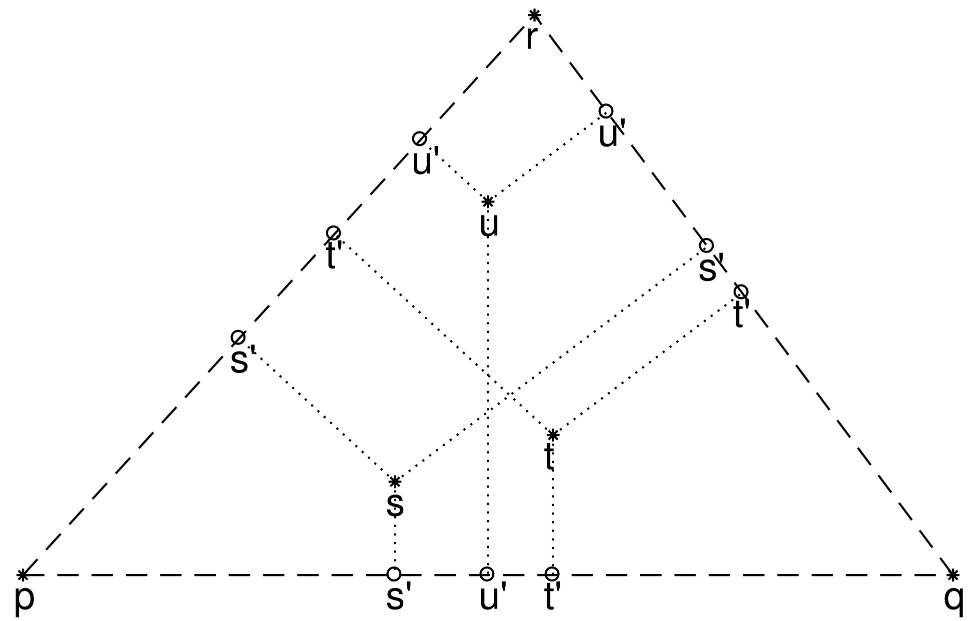
Each “orthogonal axis” in Basis Embedding is a vector upon which points are projected. So:

1. Choose many vectors (not necessarily orthogonal) and partially order points’ projections along them.

2. Solve constrained optimization problem to preserve projected order along each axis.

-or-

2. Solve for point positions geometrically.



WITH ENOUGH POINTS, PROJECTED ORDERS CONSTRAIN EMBEDDING

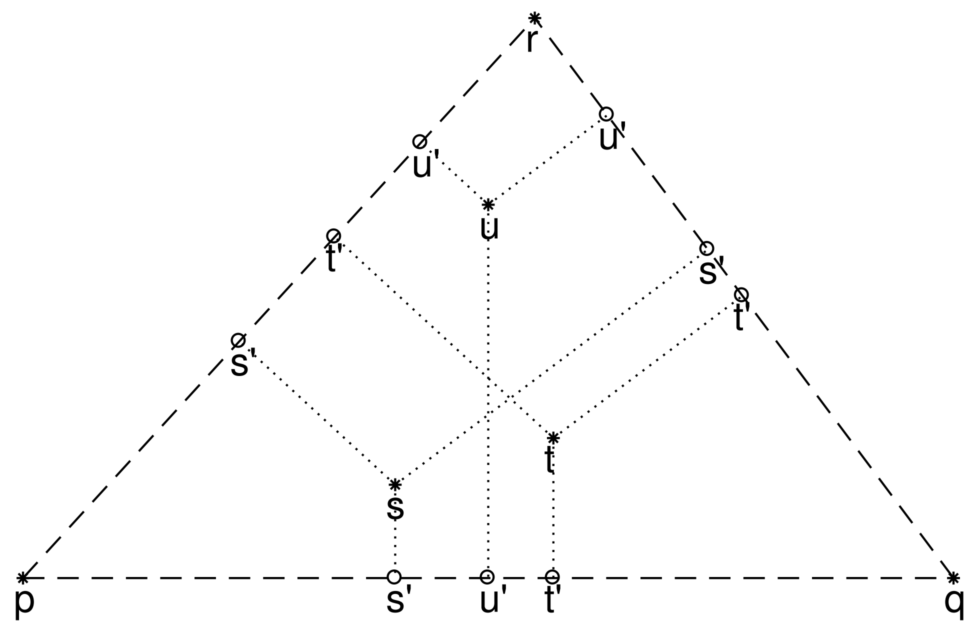
EMBEDDING WITH RANDOM VECTORS: OPTIMIZATION IDEA

a' precedes b' on vector from p to q ⇒ (a,b,p,q) ∈ PO

Objective:

$$\mathcal{L}(X; PO) = \sum_{(a,b,p,q) \in PO} \max\left[0, (X_a - X_b) \cdot (X_q - X_p) + \lambda\right]^2$$

- Incur loss when vector from a to b has negative component on "axis" from p to q.
- Boundaries are hyperplanes, not spheres, so easier objective; may be convex.
- Steepest decent for X_a, X_b is parallel to "axis!"



WITH ENOUGH POINTS, PROJECTED ORDERS CONSTRAIN EMBEDDING

IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

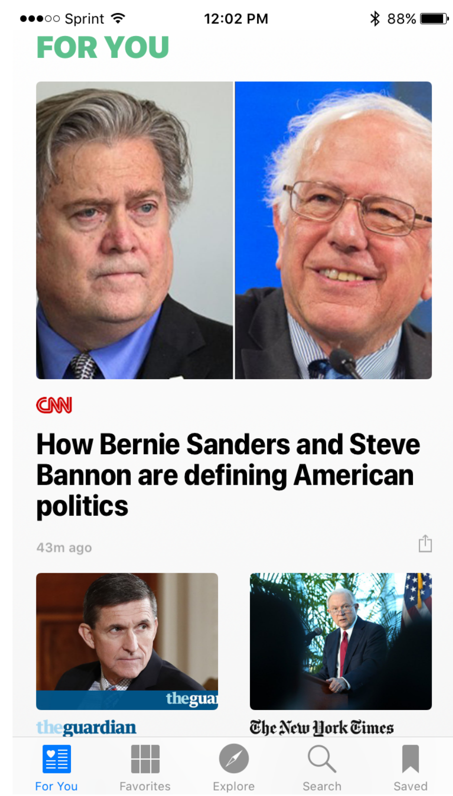
Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

EMBEDDINGS FOR RECOMMENDATIONS

- We often try to predict future user preferences using their past behavior.
- Can use embeddings: users showing interest in some object may have interest in other “nearby” objects.
- Could embed entities from news articles by inferring triples from user behavior, e.g. articles a user reads/skips.
- Is this mathematically valid?



ARTICLES RECOMMENDED BY APPLE NEWS APP.

INCONSISTENT COMPARISONS

> "A flame is similar to the moon because they are both luminous, and the moon is similar to a ball because they are both round, but in contradiction to the triangle inequality, a flame is not similar to a ball." – William James, 1890.

- The similarity function changed!
- An embedding would conflate "luminosity similarity" with "roundness similarity" and not quite capture either.

SAME ENTITY, DIFFERENT CONTEXTS

- People care about different features in different contexts.
- Different features ⇒ different similarity fn
- But different similarity function ⇒ different neighbors ⇒ different other entities in the article…
- The context should tell us this is happening!

A VARIETY OF CONTEXTS FOR ENTITY "JESSE VENTURA" — WRESTLER, GOVERNOR, AND ACTOR

MODELLING OPTIONS

Want to parameterize embedding by context.

- Discrete form: Data uses k different similarity functions, $sim_1, \ldots, sim_k \Rightarrow k$ embeddings of all n objects; learn sim_i and prob. in sim_i given context.
- Continuous form: $sim_i(x, y) = X_x C^{(i)} X_y^T$ with $C(i) \in \mathbb{R}^{d \times d}$ an affine transformation of global embedding $X \in \mathbb{R}^{n \times d}$.

IMPROVE ORDINAL EMBEDDING TECHNIQUES FOR TEXT SIMILARITY APPLICATIONS

Active Learning	Which triples should we collect?
Embedding	How can we embed accurately, at scale?
Contextual Embeddings	Can we make embeddings that adapt to context?

TIME LINE

Fall 2017

- Vector projection active learning; prove all-rankings problem is Θ(d n log n).
- Vector projection embedding; high-dim. subset embedding.

Spring 2018

- Contextual Embeddings for recommendation.

Summer 2018

- 👨🏻‍🎓, ✈️, 🍹

THANK YOU!

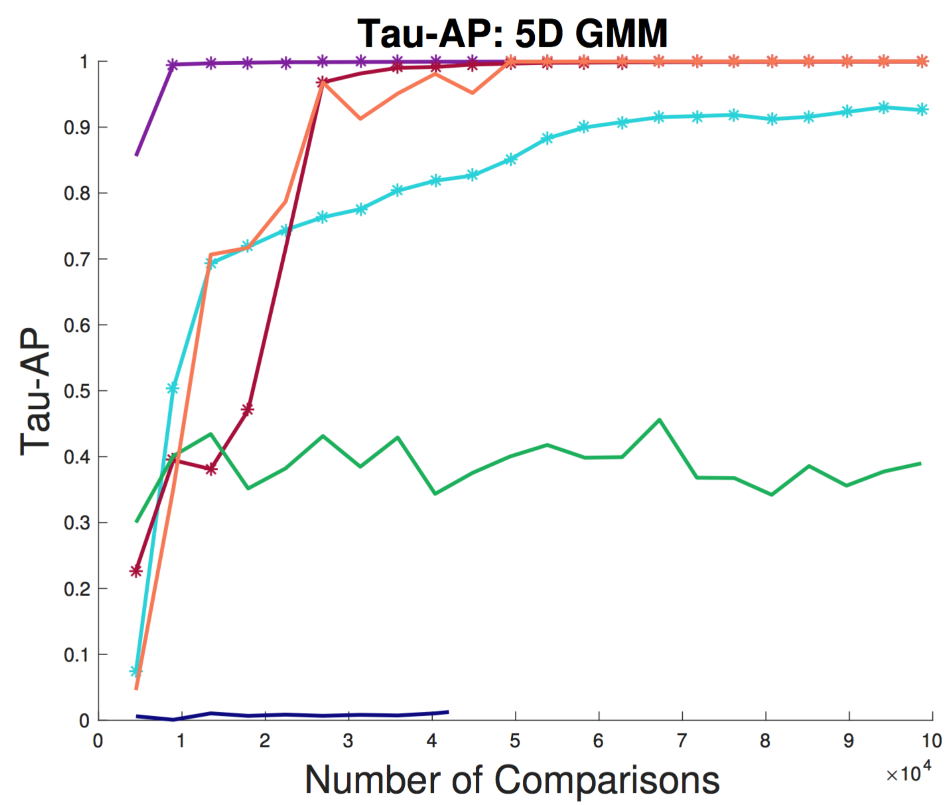
🍹

[1] M. Kleindessner and U. von Luxburg, "Uniqueness of Ordinal Embedding.," COLT, 2014.
[2] E. Arias-Castro. Some theory for ordinal embedding. Bernoulli 23 (2017), no. 3, 1663--1693. doi:10.3150/15-BEJ792.
[3] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai, "Adaptively Learning the Crowd Kernel," ICML, 2011.
[4] K. G. Jamieson and R. D. Nowak, Low-dimensional embedding using adaptively selected ordinal data. IEEE, 2011, pp. 1077–1084.
[5] Y. Terada and U. von Luxburg, "Local ordinal embedding," ICML, 2014.
[6] K. Li and J. Malik, "Fast k-Nearest Neighbour Search via Dynamic Continuous Indexing," ICML, 2016.
[7] M. Cucuringu and J. Woodworth, "Point Localization and Density Estimation from Ordinal kNN graphs using Synchronization," arXiv.org, 2015.
[8] J. Anderton, V. Pavlu, J. Aslam, "Triple Selection for Ordinal Embedding," unpublished, 2016.
[9] J. Anderton, V. Pavlu, J. Aslam, "Revealing the Basis: Ordinal Embedding through Geometry," unpublished, 2016.
[10] J. Anderton, P. Metrikov, V. Pavlu, J. Aslam, "Measuring Human-Perceived Similarity in Heterogeneous Collections," unpublished, 2014.

EMPIRICAL COMPARISON

τ_{AP} IS A TOP-HEAVY RANK CORRELATION MEASURE

- FRFT Ranking – My algorithm, using rankings from features – O(n) triples per ranking.
- FRFT Adaptive Sort – My algorithm, using no prior knowledge – O(n log n) then O(n).
- Crowd Kernel – Active learning baseline.
- Random Tails – Random baseline.
- kNN – Gradually add next NN for each obj.
- Landmarks – Gradually add objects to all rankings.

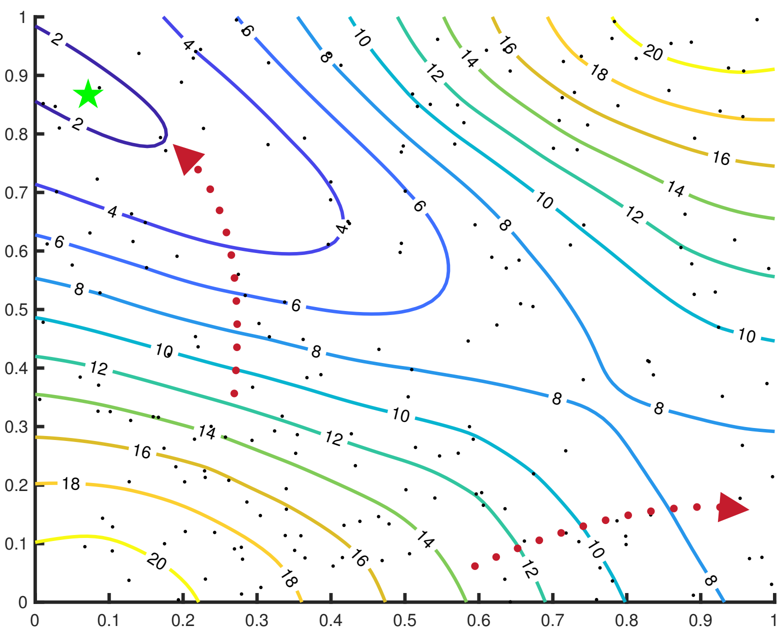


[8] J. Anderton, V. Pavlu, J. Aslam, "Triple Selection for Ordinal Embedding," unpublished, 2016.

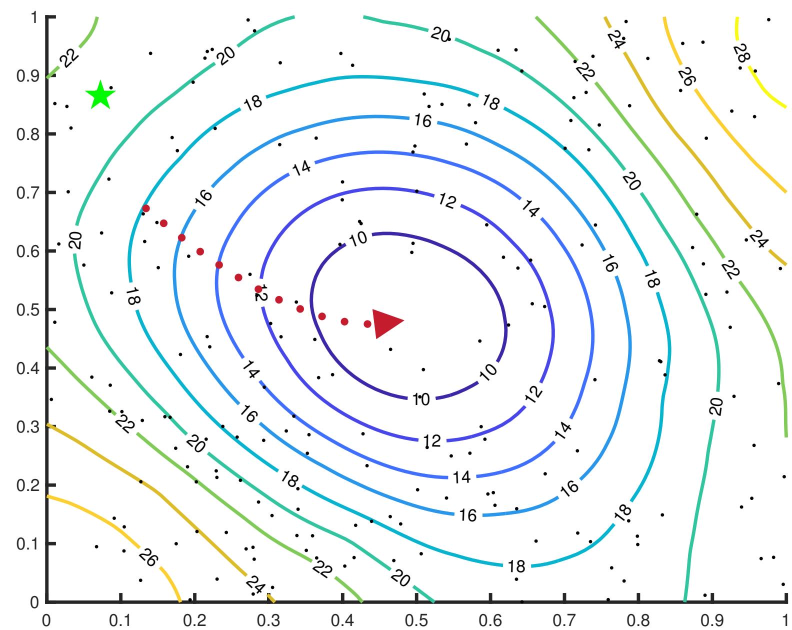
OPTIMIZATION AT SCALE IS DIFFICULT

With random initialization, the gradient is misleading. This is harder to fix as n and d increase.

SOE LOSS OF SINGLE POINT: OTHER POINTS IN CORRECT POSITIONS

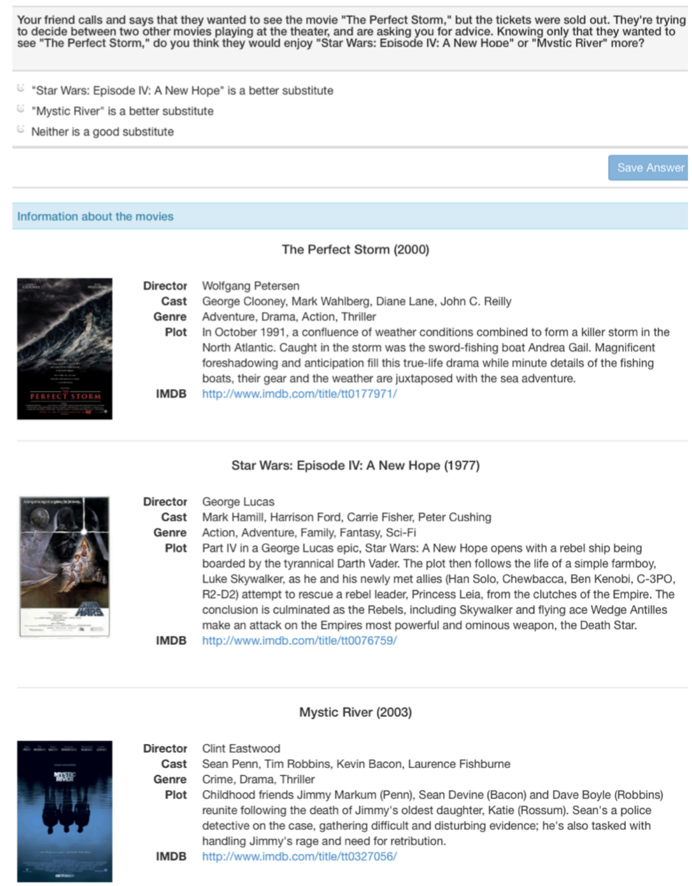


SOE LOSS OF SAME POINT: OTHER POINTS IN RANDOM POSITIONS



PER-USER CONTEXTS FOR CROWDSOURCING

- We tried a simple first approach using crowdsourced triples.
- For two datasets (movies and foods), users were asked, "would a person who likes object a prefer b or c?"
- We attempted to train a global embedding of all objects and a per-user transformation of that embedding.



CROWDSOURCING INTERFACE

[10] J. Anderton, P. Metrikov, V. Pavlu, J. Aslam, "Measuring Human-Perceived Similarity in Heterogeneous Collections," unpublished, 2014.

PER-USER CONTEXTS FOR CROWDSOURCING

Given an embedding matrix $X \in \mathbb{R}^{n \times d}$, the standard similarity function is the Gram matrix,

$$K = XX^T$$

For each user k, we learn a per-user weight for each feature in a diagonal matrix $U^k \in \mathbb{R}^{d \times d}$. This gives a new similarity,

$$K = XU^kX^T$$

We chose questions adaptively using the Crowd Kernel method adapted to our model, and embedded the result using a Newton-Rhapson method.

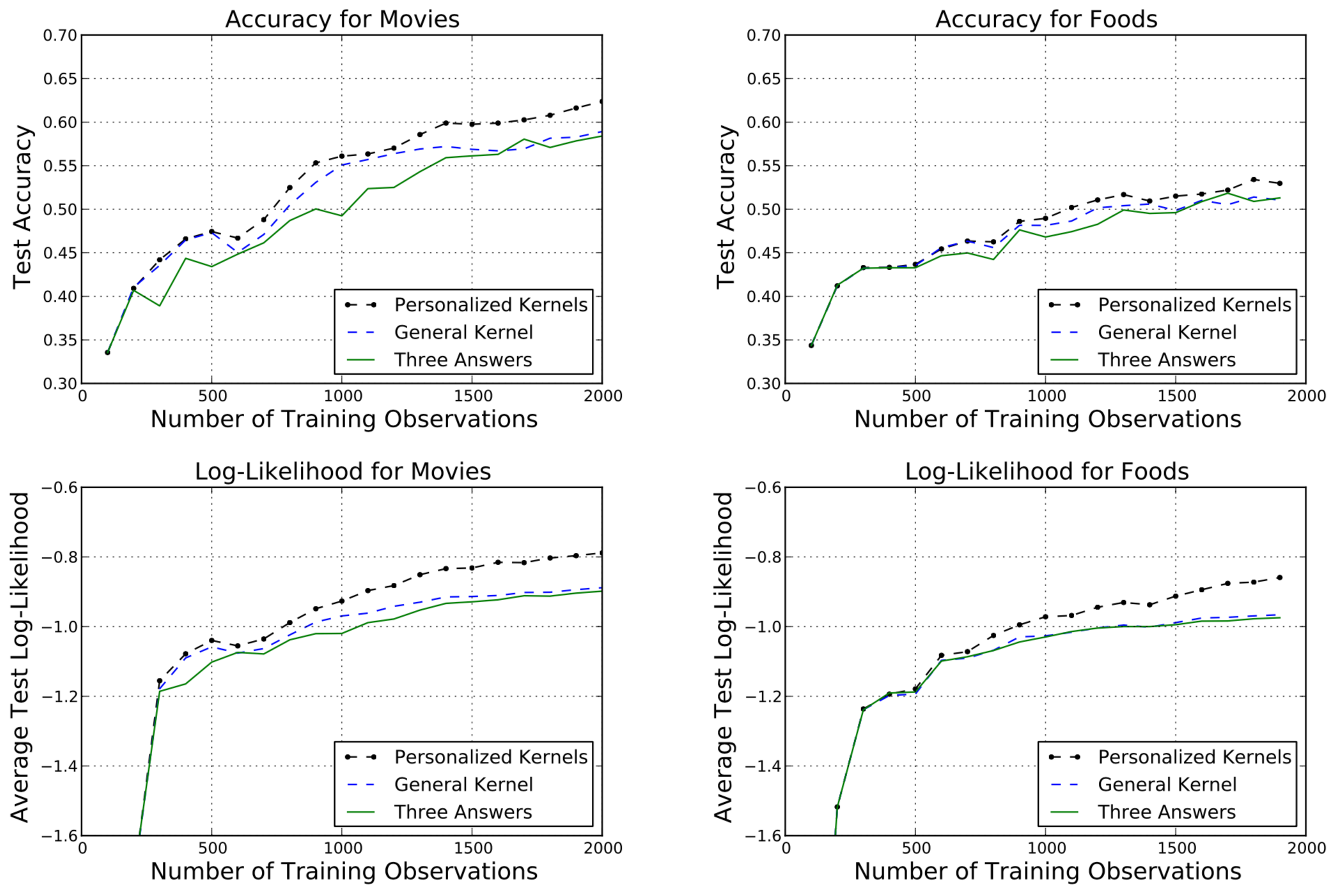
USER RESPONSE MODEL

$$\delta_{ab}{}^{k} = \|X_a \cdot \mathrm{diag}(U^k) \cdot X_b\|^2$$

Answer prob.	**Kernel two answers**	**Kernel three answers**
$\hat{p}^a_{bc}$	$\frac{\lambda + \delta_{ac}}{2\lambda + \delta_{ab} + \delta_{ac}}$	$(1 - \hat{p}_{neither}) \cdot \frac{\lambda + \delta_{ac}}{2\lambda + \delta_{ab} + \delta_{ac}}$
$\hat{p}^a_{cb}$	$\frac{\lambda + \delta_{ab}}{2\lambda + \delta_{ac} + \delta_{ab}}$	$(1 - \hat{p}_{neither}) \cdot \frac{\lambda + \delta_{ab}}{2\lambda + \delta_{ac} + \delta_{ab}}$
$\hat{p}_{neither}$	0 (N/A)	$\frac{\mu + \delta_{ab}}{\mu + d^2 + \delta_{ab}} \cdot \frac{\mu + \delta_{ac}}{\mu + d^2 + \delta_{ac}}$

[10] J. Anderton, P. Metrikov, V. Pavlu, J. Aslam, "Measuring Human-Perceived Similarity in Heterogeneous Collections," unpublished, 2014.

PER-USER CONTEXTS FOR CROWDSOURCING: RESULTS



[10] J. Anderton, P. Metrikov, V. Pavlu, J. Aslam, "Measuring Human-Perceived Similarity in Heterogeneous Collections," unpublished, 2014.