Efficient Triplet Based Perceptual Embeddings

Serge Belongie

Part I: Multidimensional Scaling

TABLE 14.3. *Data from a political science survey: values are average pairwise dissimilarities of countries from a questionnaire given to political science students.*

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Given an input of $N \times N$ dissimilarity matrix D, MDS provides us a set of N k-dimensional points that could have given rise to D. In principle, one could approach this problem by a succession of triangulation steps to solve for the relative location of all the points. The SVD provides us a more elegant solution.

As described in Ripley (1995), we start by converting the dissimilarity matrix into an inner product matrix. For any symmetric matrix $\mathbf{T}$, we can define $\mathbf{T}'$ as follows:

$$\mathbf{T}' = -\frac{1}{2}\left[\mathbf{T} - \frac{(\mathbf{T}\vec{1})\vec{1}^\top}{N} - \frac{\vec{1}(\mathbf{T}\vec{1})^\top}{N} + \frac{\vec{1}^\top\mathbf{T}\vec{1}}{N^2}\right]$$

where $\vec{1}$ is a length N column vector of $1s$. Pre- or post-multiplication by a vector of 1s is a linear algebraic trick for summing the rows or columns of a matrix, respectively. Sandwiching $\mathbf{T}$ in the form $\vec{1}^\top\mathbf{T}\vec{1}$ simply adds up all its entries.

The resulting matrix $\mathbf{T}'$ has the following properties:

1. If $\mathbf{T}$ was formed by computing pairwise Euclidean distances on the x_is, then $\mathbf{T}'$ contains the inner product between the x_is, i.e. $\mathbf{T}' = \mathbf{X}\mathbf{X}^\top$.

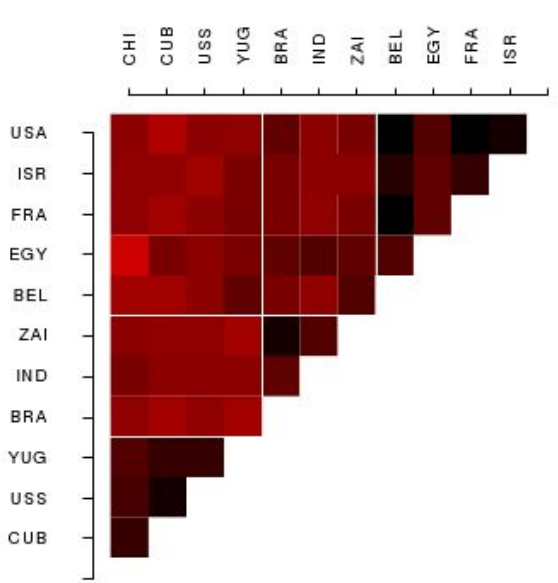
2. We can use the SVD to find the matrix square root of $\mathbf{T}'$, i.e., solve for $\mathbf{X}$ such that $\mathbf{T}' = \mathbf{X}\mathbf{X}^\top$, and the rows of $\mathbf{X}$ will contain the coordinates we seek.

The proof of this is based on the following observation. Recalling that $\|a\|^2 = a^\top a$, the squared distance between two points x_i and x_j is expressed as

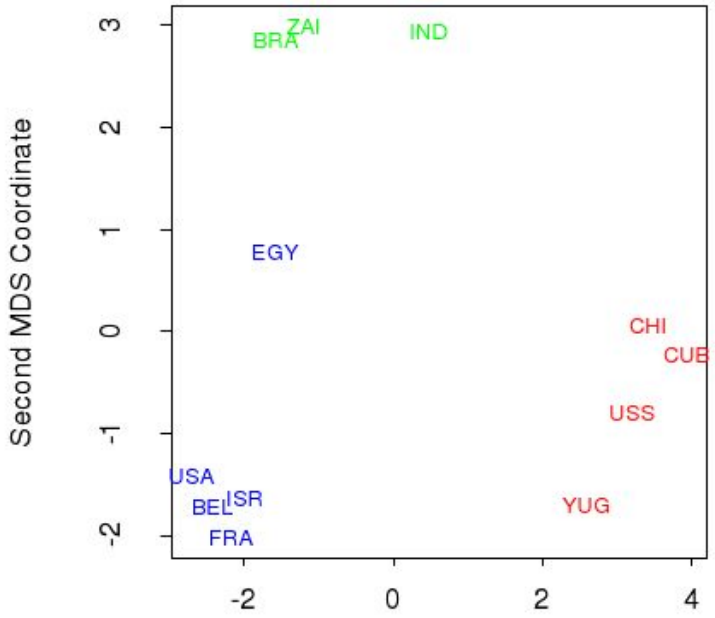
$$\|x_i - x_j\|^2 = (x_i - x_j)^\top (x_i - x_j) = \|x_i\|^2 + \|x_j\|^2 - 2x_i^\top x_j$$

The transformation of $\mathbf{T}$ into $\mathbf{T}'$ effectively subtracts off the two terms corresponding to the norms of x_i and x_j and just leaves us with the inner product term, $x_i^\top x_j$.

Because $\mathbf{T}'$ can be expressed as $\mathbf{X}\mathbf{X}^\top$, it is positive semidefinite, which means all of its eigenvalues are nonnegative. This also means we can interpret it as a covariance matrix. As a result, MDS has strong conceptual links to PCA.



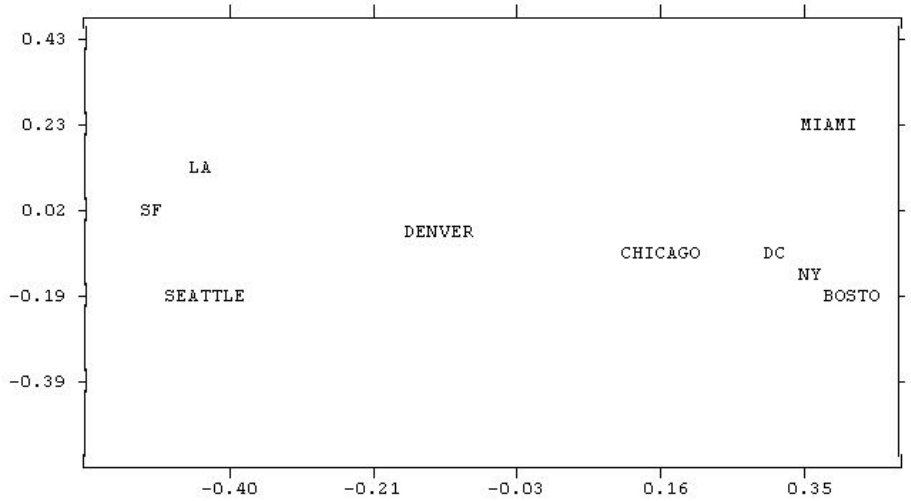
Reordered Dissimilarity Matrix



First MDS Coordinate

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

For instance, given the matrix of distances among cities shown above, MDS produces this map:



[http://www.analytictech.com/borgatti/mds.htm]

Part II: Triplet Embeddings

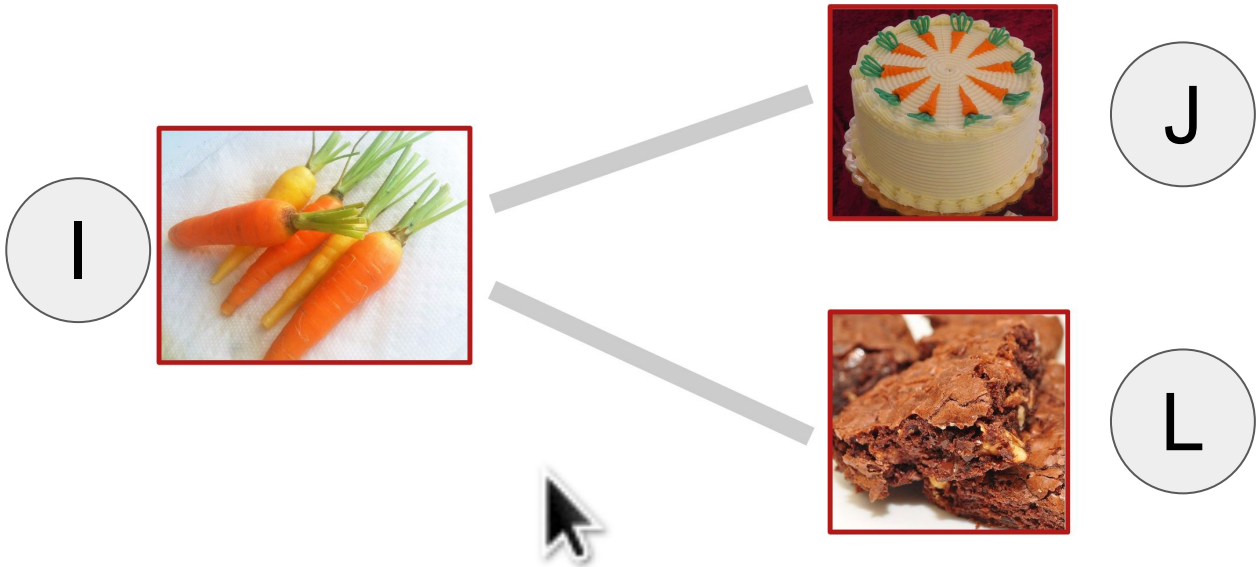
Triplets are a special case of “Paired Comparisons”

Problem 2 (Paired Comparisons). *Given a set $\mathcal{S}$ of quadruples, find $X \in \mathbb{R}^{d \times n}$ such that*

$$(i, j, k, l) \in \mathcal{S} \iff \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 < \|\mathbf{x}_k - \mathbf{x}_l\|_2^2 \qquad (2)$$

[Agarwal et al. 2007]

Triplets Formulation



$$\mathcal{T} = \{(i, j, \ell) | \, \mathbf{z}_i \text{ is more similar to } \mathbf{z}_j \text{ than } \mathbf{z}_\ell\} .$$

[van der Maaten & Weinberger]

Goal: find vector embedding that satisfies the underlying pairwise similarity function s()

$$\|\mathbf{x}_i, \mathbf{x}_j\|_2 < \|\mathbf{x}_i, \mathbf{x}_\ell\|_2 \iff s(\mathbf{z}_i, \mathbf{z}_j) < s(\mathbf{z}_i, \mathbf{z}_\ell). \quad (2)$$

For notational simplicity, we define the $r \times N$ design matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and the kernel matrix $\mathbf{K} = \mathbf{X}^\top \mathbf{X}$.

Generalized Non-metric Multidimensional Scaling

$$\min_{\mathbf{K}} \ \mathrm{trace}(\mathbf{K}) + C \sum_{\forall (i,j,\ell) \in \mathcal{T}} \xi_{ij\ell} \ \text{subject to:}$$

$$(1) \quad k_{jj} - 2k_{ij} - k_{\ell\ell} + 2k_{i\ell} \leq 1 + \xi_{ij\ell}$$

$$(2) \quad \xi_{ij\ell} \geq 0$$

$$(3) \quad \mathbf{K} \succeq 0.$$

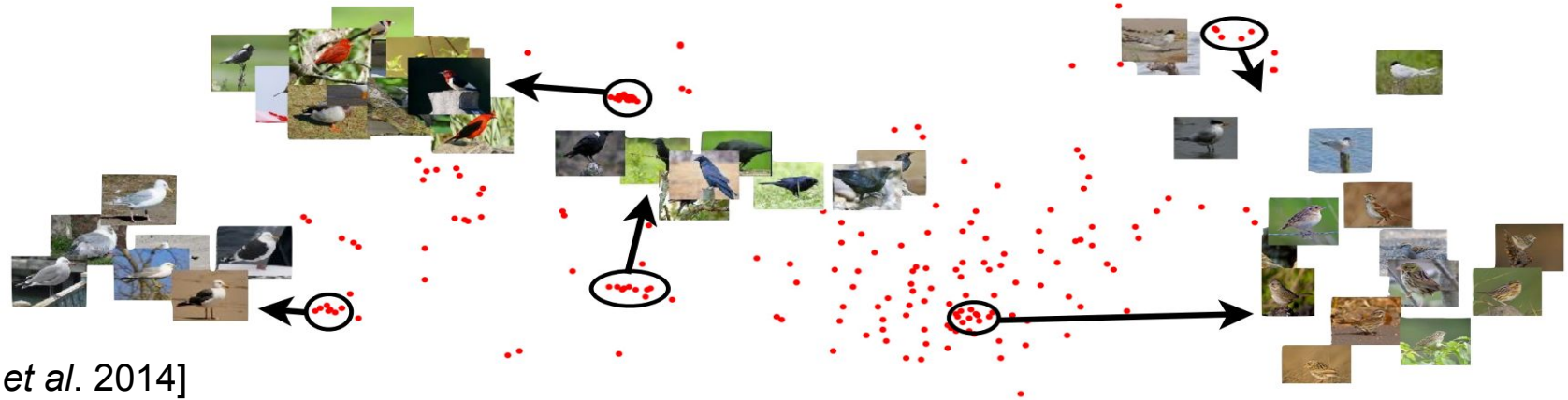
Part III: Grid-Based Triplets

Perceptual Embeddings

- A **perceptual embedding** is a space where distance corresponds to **intuitive similarity**.

Embeddings for bird identification

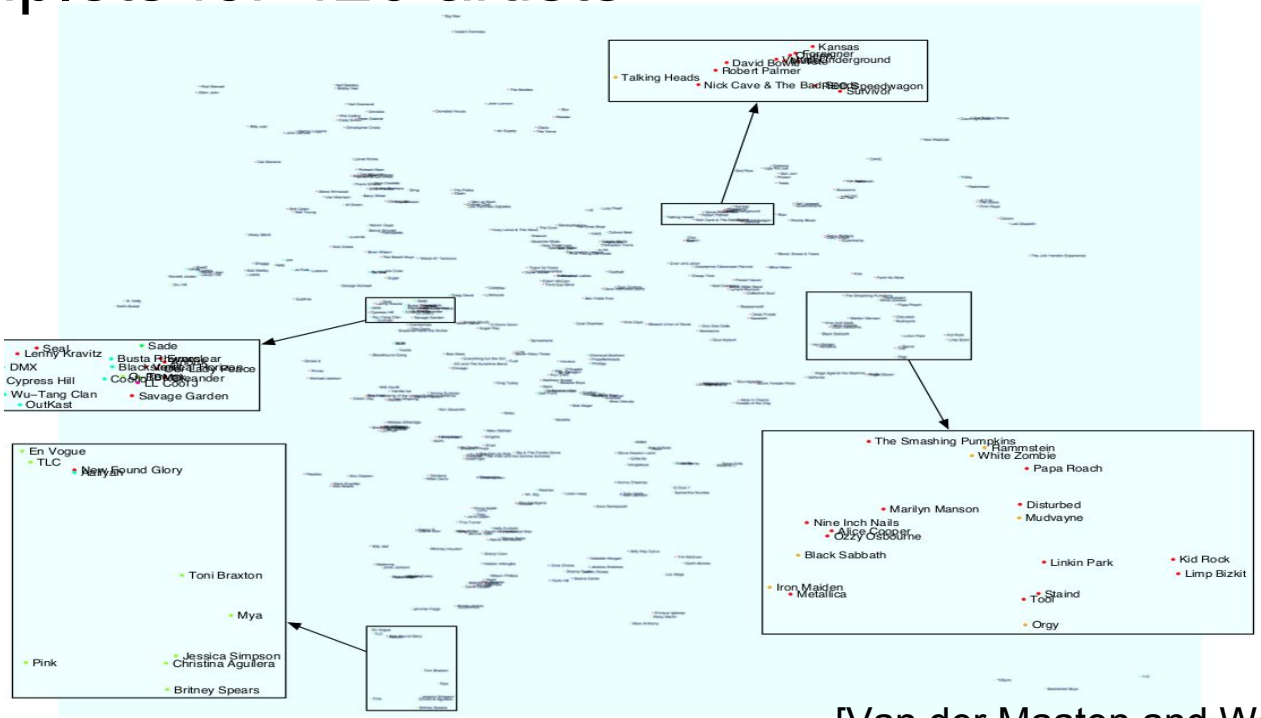
- Wah et al, 2014 built a bird ID system that used human appearance similarity as a prior
- In this space, **bird species that look similar are close together**.
- Asked **93,000 questions** to embed N=200 points



[Wah *et al*. 2014]

Embeddings for musician similarity

- http://homepage.tudelft.nl/19j49/ste
- 22,310 triplets for 426 artists



[Van der Maaten and Weinberger, 2012]

How to create perceptual embeddings

- **Step 1**: Collect dataset
- **Step 2**: Ask the crowd tens of thousands of questions
- **Step 3**: Construct the embedding by maximizing some optimization objective
- **Step 4**: Use the embedding in applications like
  - Bird identification (Wah et al. 2014)
  - Semantic clusters (Gomes 2011)
  - Inferring music similarity (McFee 2012)
  - Training better neural networks (Wang et al. 2014)

How to create perceptual embeddings

- **Step 1**: Collect dataset
- **Step 2**: Ask the crowd tens of thousands of questions
- **Step 3**: Construc embedding by maximizing some optim
- **Step 4**: Us ns like
  - Bird identi
  - Semantic
  - Inferring r
  - Training bette al. 2014)

Bottleneck!

There are too many possible questions to ask!

Can we ask questions that are **more informative?**

What kinds of questions could we ask?

- Hardcoded **labels?**
  - But not all pizzas taste alike.
  - Humans are bad at fine-grained classification.
  - Taxonomies are imprecise.

What kinds of questions could we ask?

- ~~Hardcoded **labels?**~~
- Pairwise comparisons?
  - Needlessly **quantized**, which discards information
  - May be **inconsistent** across humans
  - Not informative for extremely similar or extremely dissimilar answers
  - Metric assumptions are **violated** in human perceptual judgments (Tversky, 1977)

What kinds of questions could we ask?

- ~~Hardcoded **labels?**~~
- ~~Pairwise comparisons?~~
  - ~~Needlessly **quantized**, which discards information~~
  - ~~May be **inconsistent** across humans~~
  - ~~Not informative for extremely similar or extremely dissimilar answers~~
  - ~~Metric assumptions are **violated** in human perceptual judgments (Tversky, 1977)~~

Alternative: **Triplet questions**

- “Which food tastes **more similar** to food A?”

A

B

C

Can we do better?

- **Key question:** How does the design of the HIT task influence the time, cost, and quality of our triplet embeddings?

- Our contribution: **Grid questions**

Our design: **Grid questions**

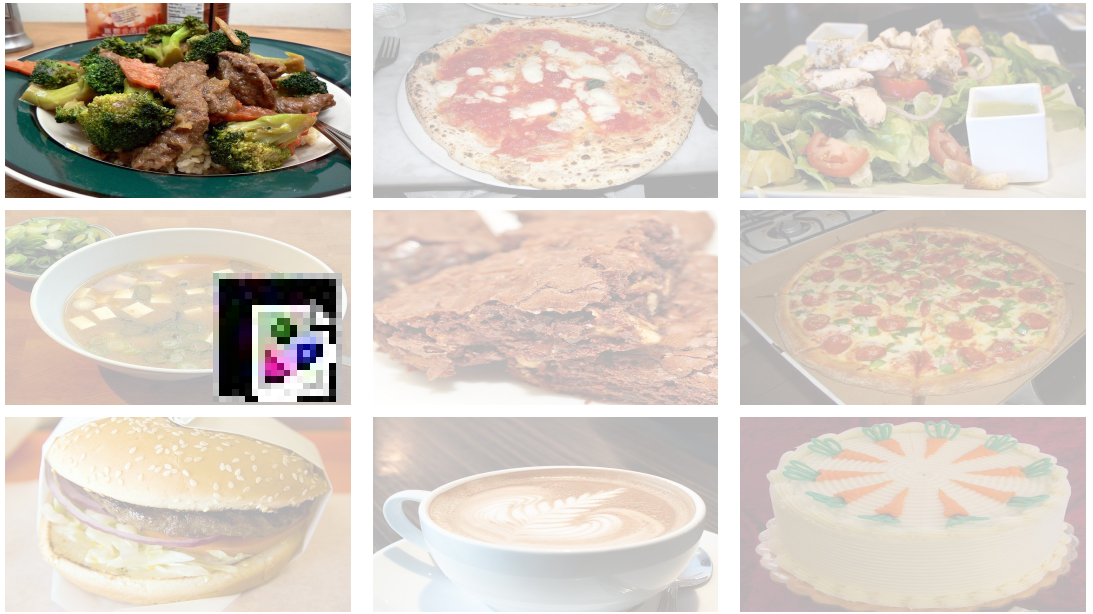
- “Please select the **four foods** that taste similar to the one on the left.”

Our design: **Grid questions**

- “Please select the **four foods** that taste similar to the one on the left.”

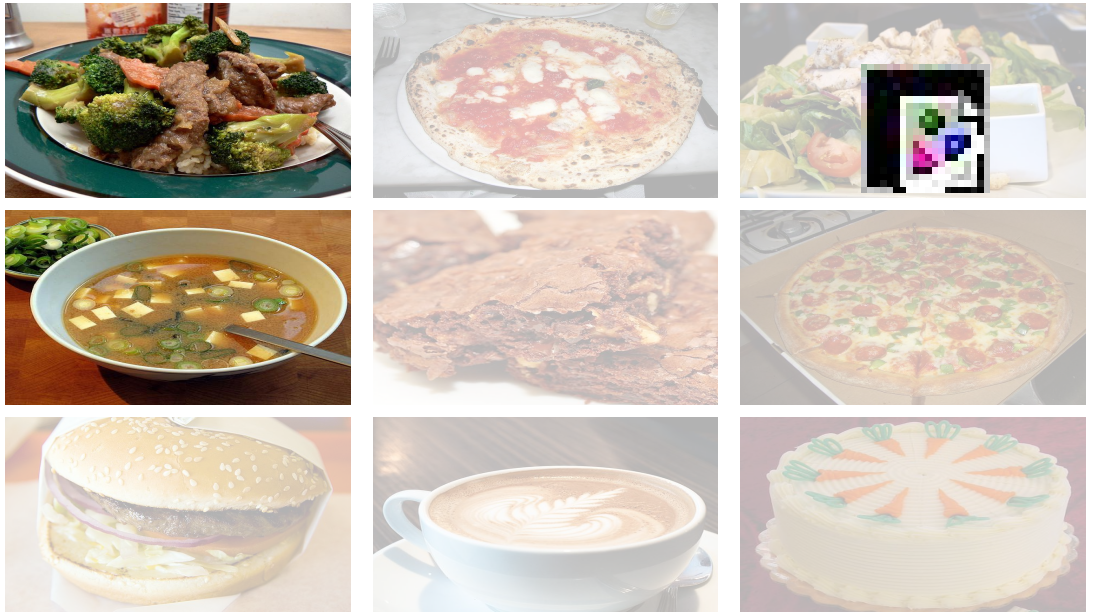
Our design: **Grid questions**

- “Please select the **four foods** that taste similar to the one on the left.”



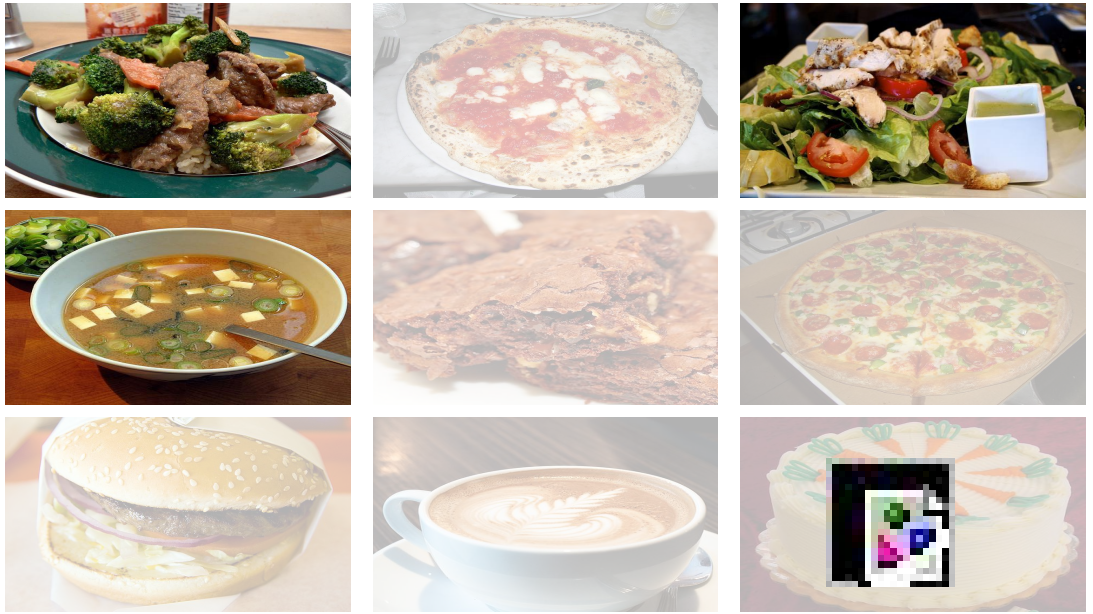
Our design: **Grid questions**

- “Please select the **four foods** that taste similar to the one on the left.”



Our design: **Grid questions**

- “Please select the **four foods** that taste similar to the one on the left.”



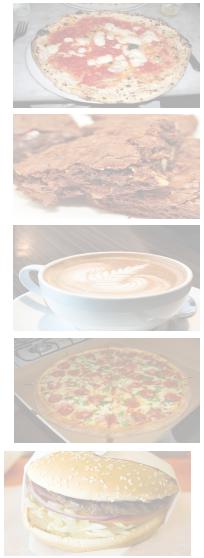
Our design: **Grid questions**

- “Please select the **four foods** that taste similar to the one on the left.”

Our design: **Grid questions**

- Grid questions → 20 triplets at once

Near



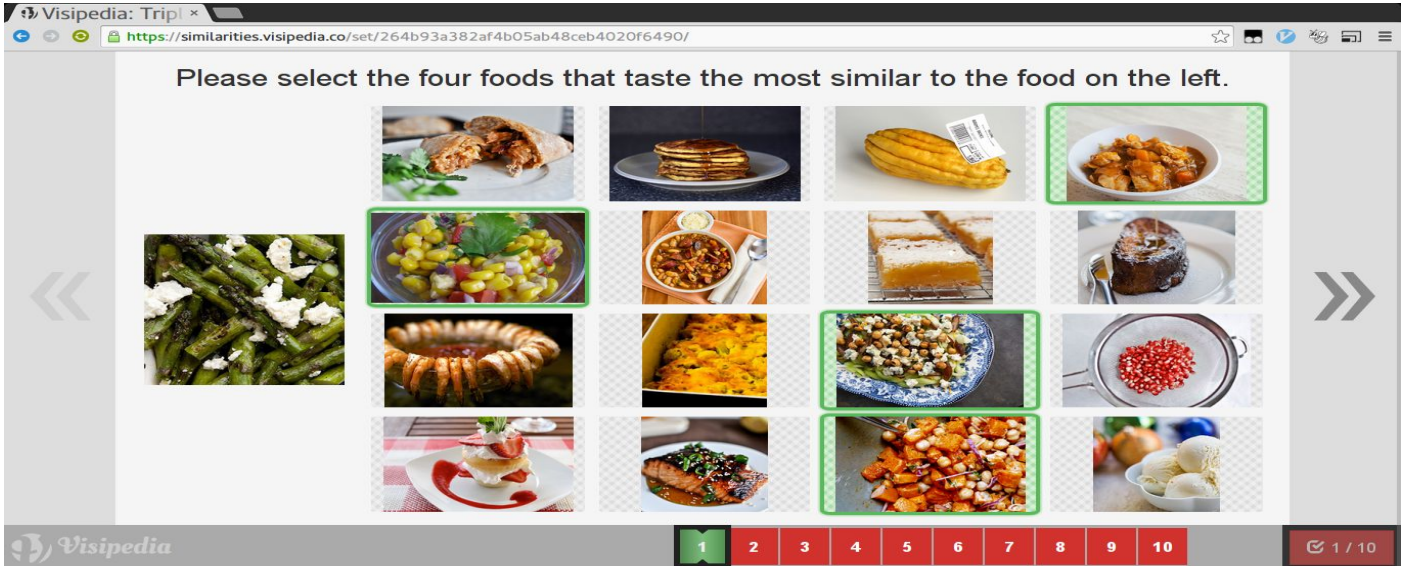
Far

Crowdsourced food experiments

- **Dataset**: 100 Yummly food images

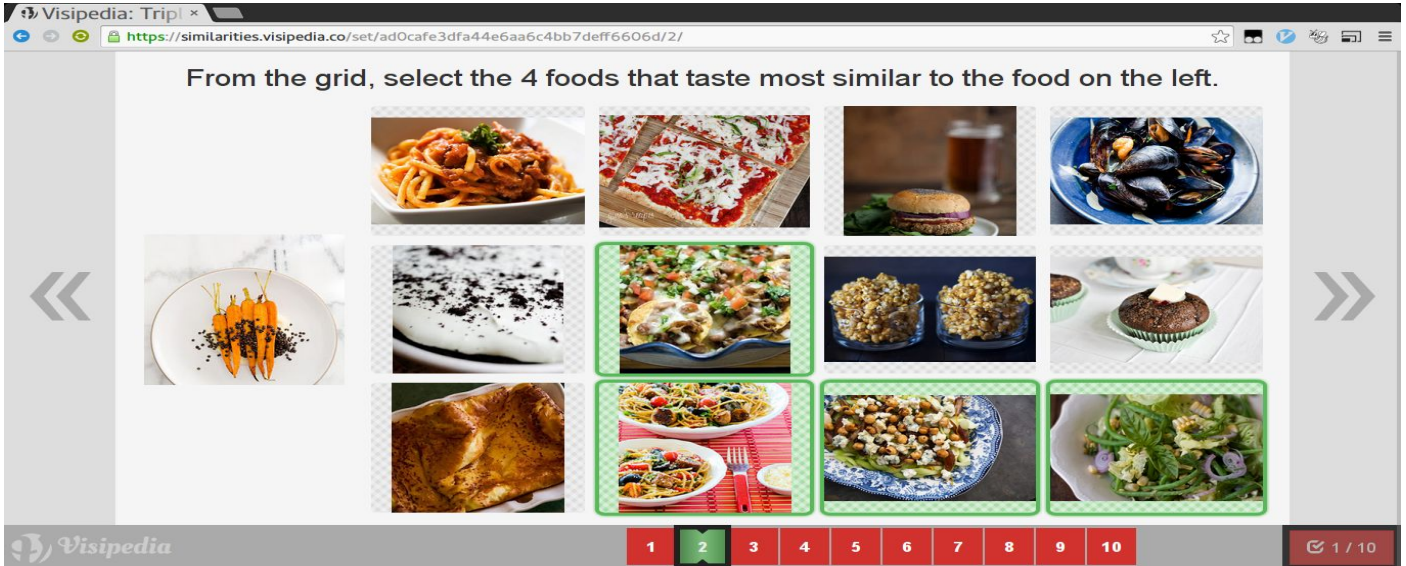
Crowdsourced food experiments

- **Experiments**: We sampled 14,088 grid questions, which gave us 189,519 triplets.
- **Grid sizes**: We tried several grid sizes:
- Select 4 out of 16 images



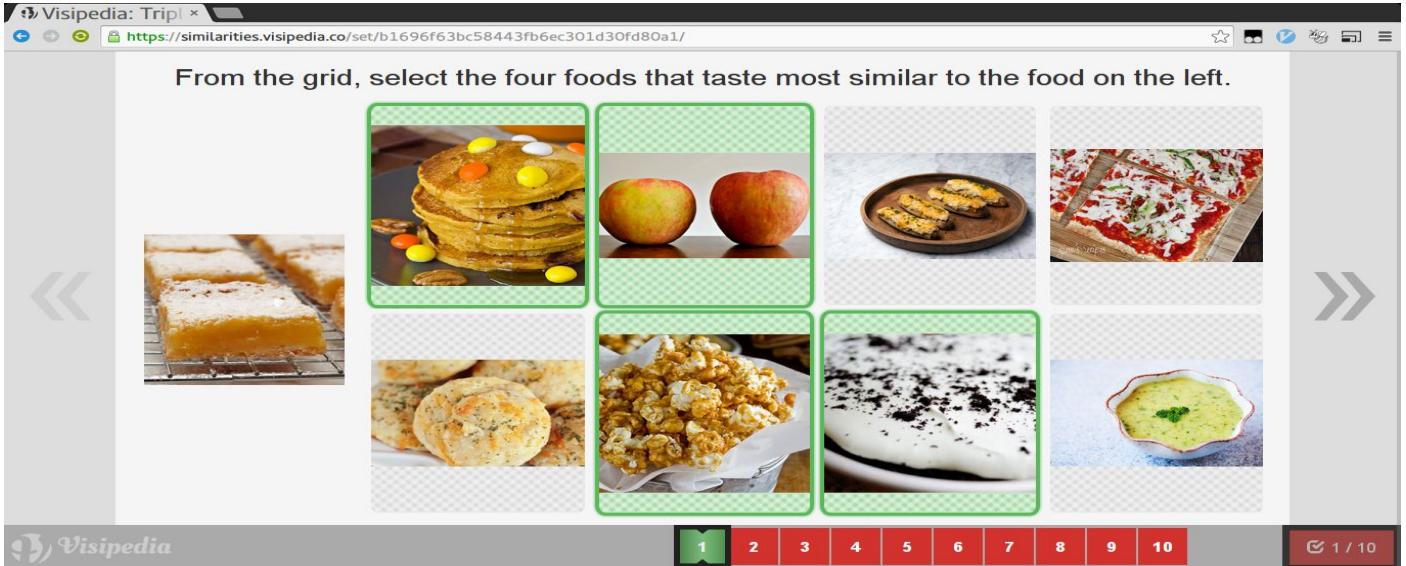
Crowdsourced food experiments

- **Experiments**: We sampled 14,088 grid questions, which gave us 189,519 triplets.
- **Grid sizes**: We tried several grid sizes:
- Select 4 out of 12 images



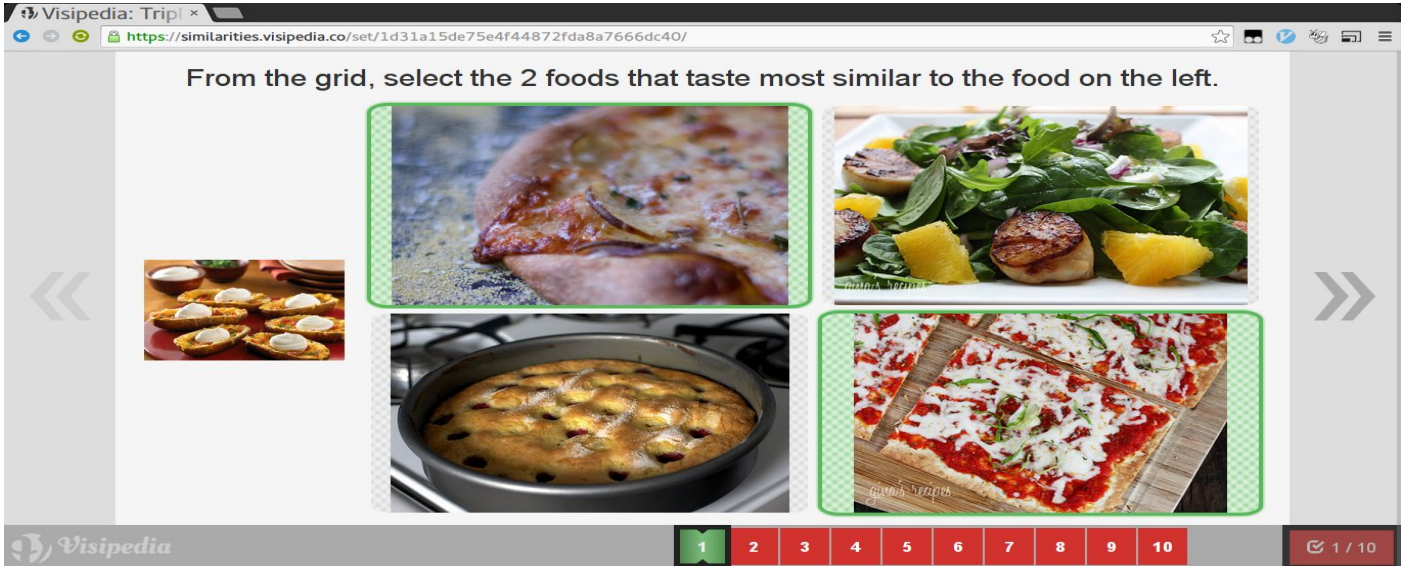
Crowdsourced food experiments

- **Experiments**: We sampled 14,088 grid questions, which gave us 189,519 triplets.
- **Grid sizes**: We tried several grid sizes:
- Select 4 out of 8 images



Crowdsourced food experiments

- **Experiments**: We sampled 14,088 grid questions, which gave us 189,519 triplets.
- **Grid sizes**: We tried several grid sizes:
- Select 2 out of 4 images



Triplet embedding algorithm

- To turn triplets into an embedding, the embedding algorithm places objects at locations that maximize an objective function.

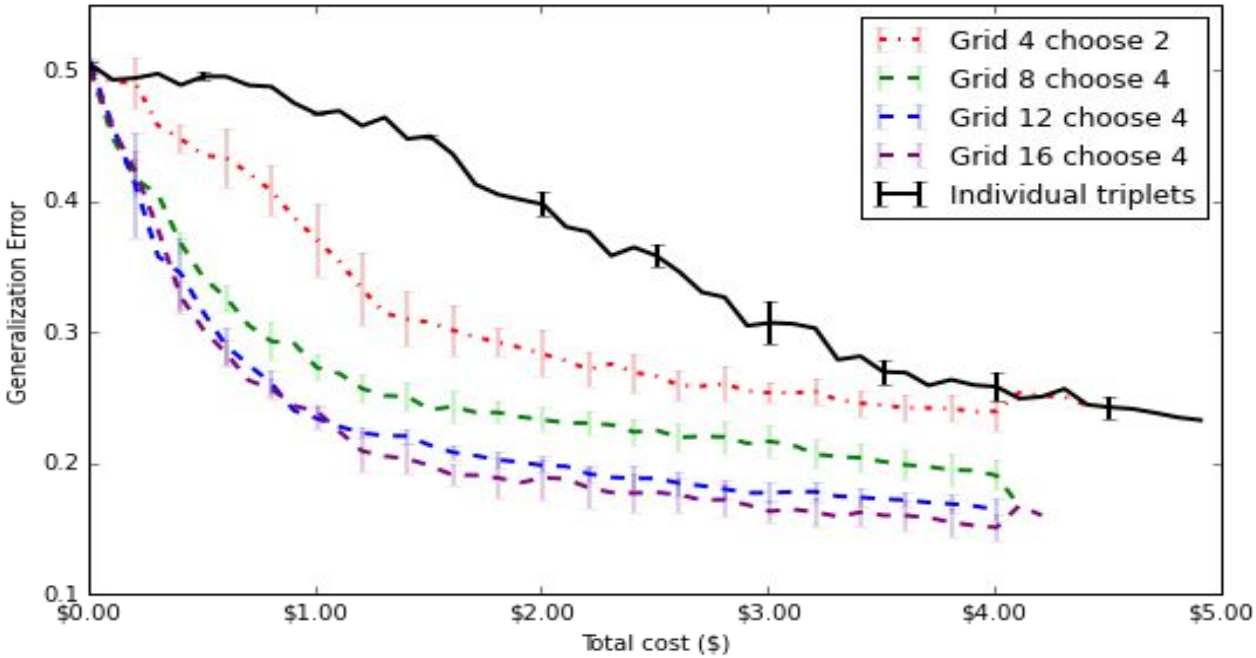
- Our embedding algorithm: t-STE (Van der Maaten et al, 2012), with default parameters.

- **This is not our focus.** We're concerned about question design, not the embedding algorithm.

Quantitative results

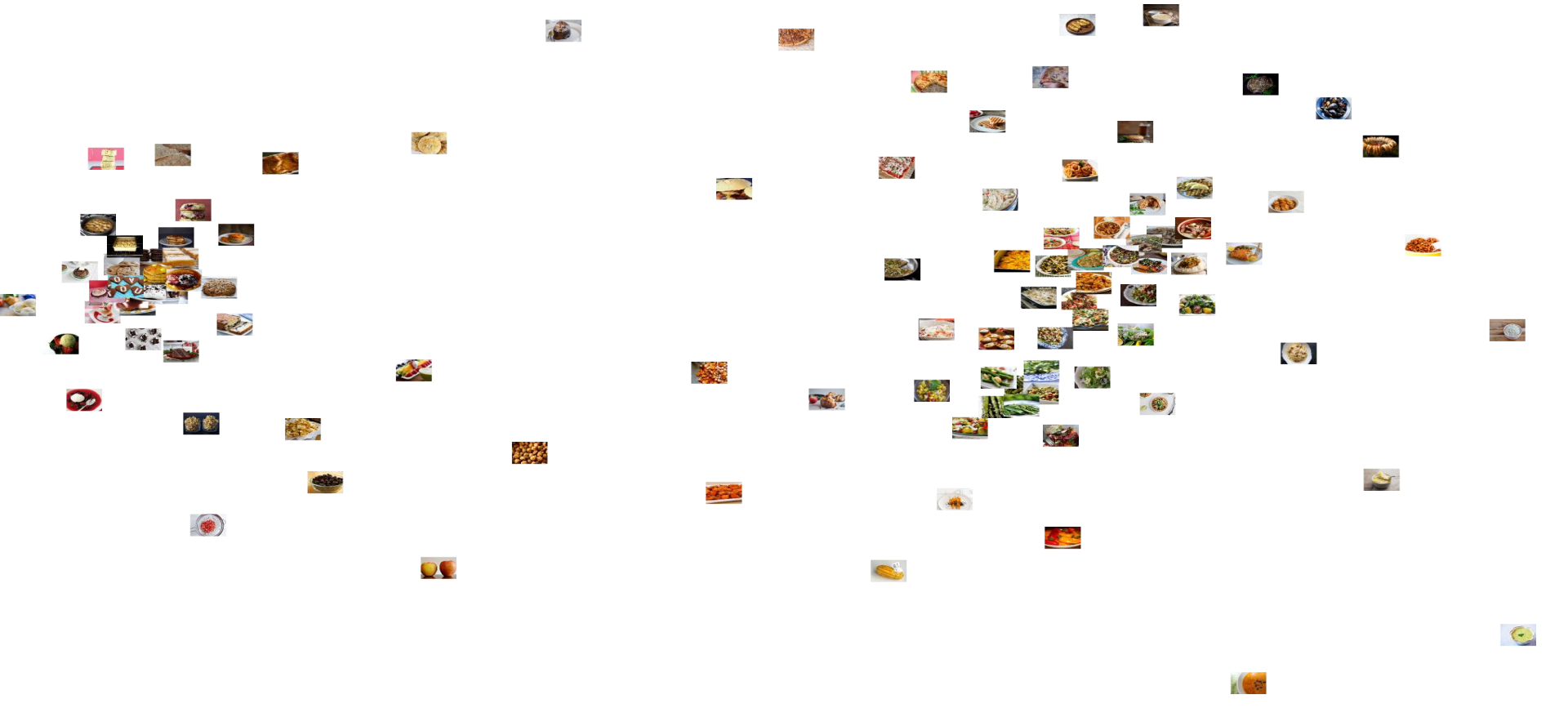
- When we view **embedding quality vs. dollars spent,** grid questions converge *faster*

Error:
Total number of unsatisfied constraints.
Lower is better.



Qualitative Results

Cost: **$5.10**, collected **19,199** triplets



Qualitative Results

Cost: **$5.10**, collected **19,199** triplets

Dessert foods are clustered together

Qualitative Results

Cost: **$5.10**, collected **19,199** triplets

Entrees are close, with salads below

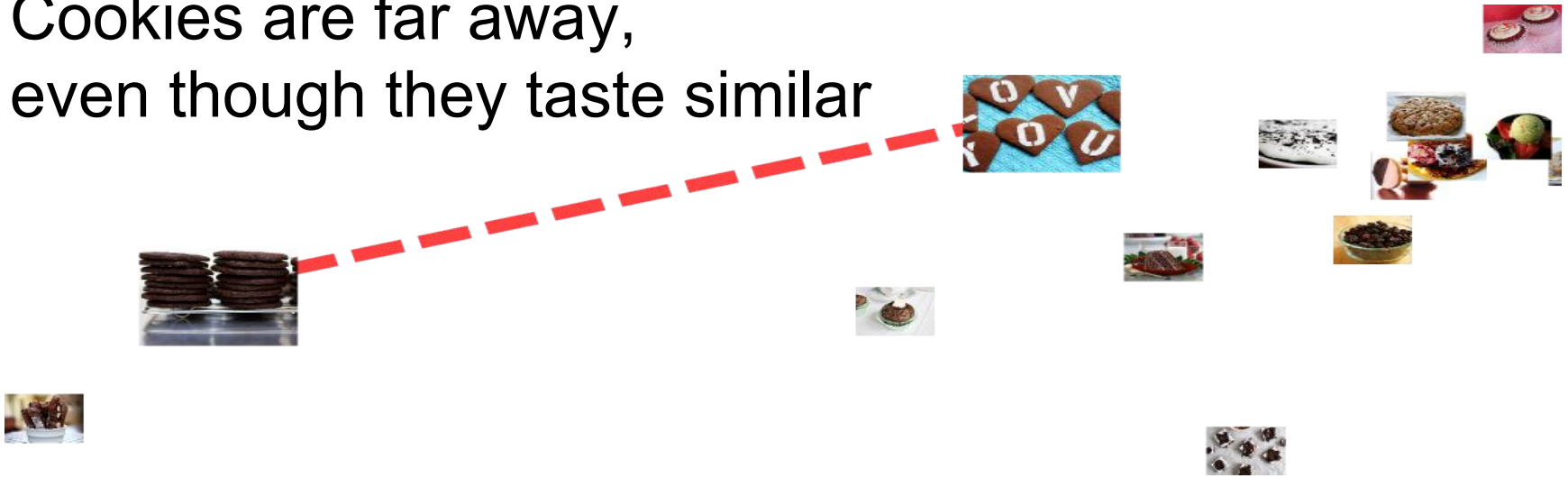
Qualitative Results: Individual triplets

Cost: **$5.10**, collected **408** triplets

Qualitative Results: Individual triplets

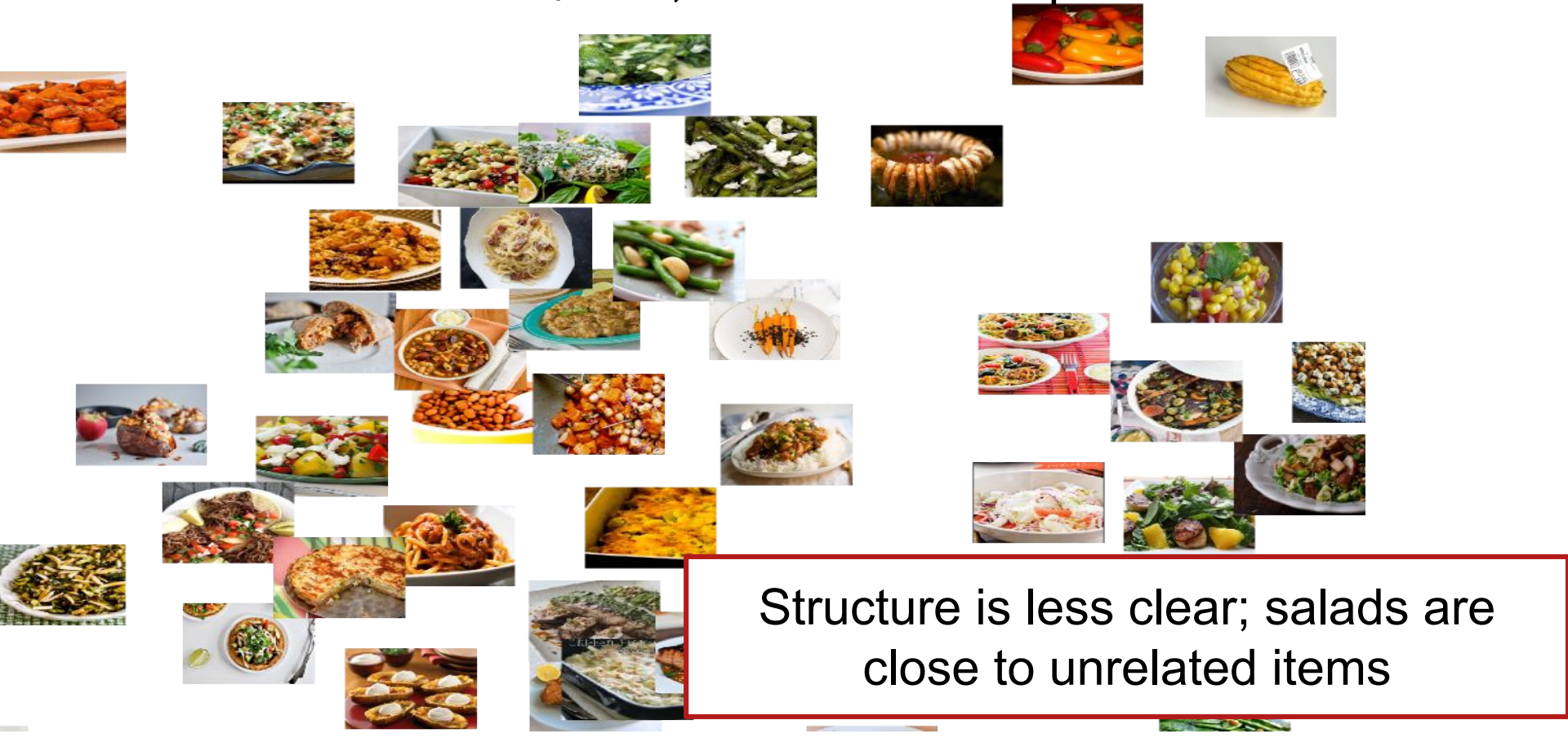
Cost: **$5.10**, collected **408** triplets

Cookies are far away,
even though they taste similar



Qualitative Results: Individual triplets

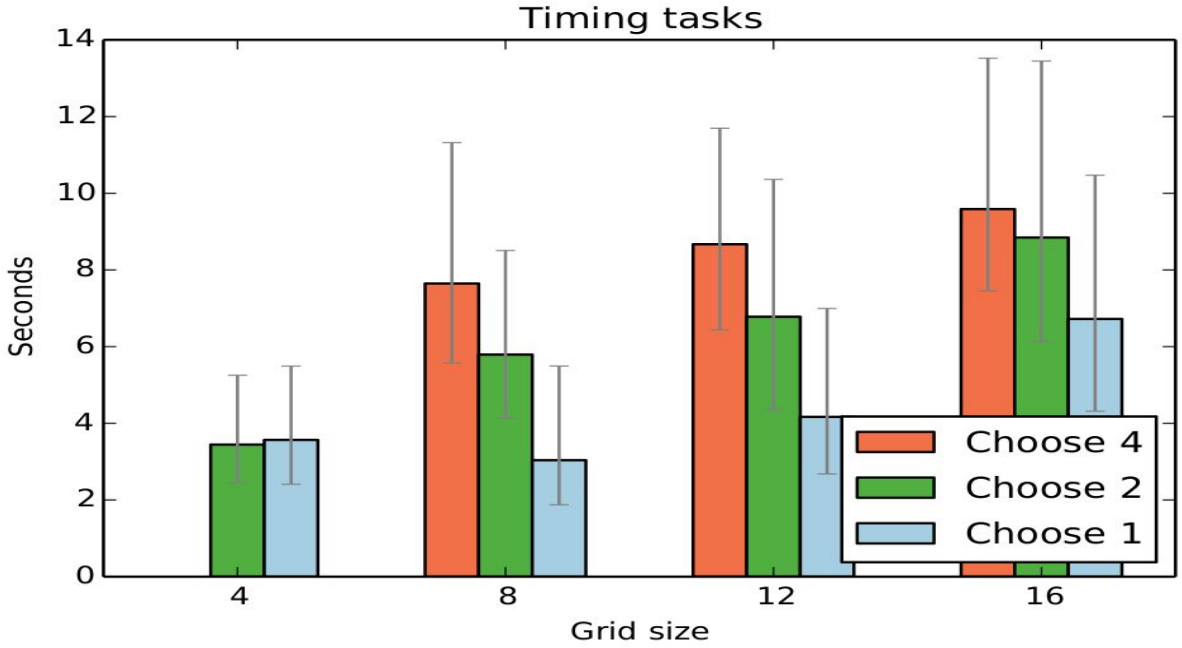
Cost: **$5.10**, collected **408** triplets



Structure is less clear; salads are close to unrelated items

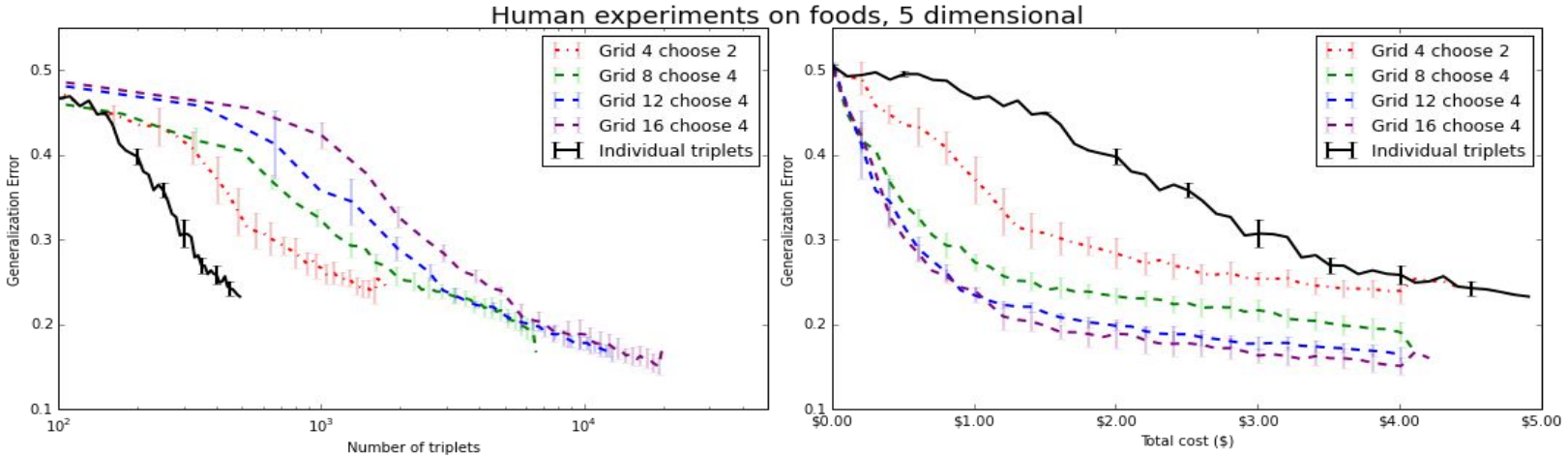
Results: Worker satisfaction

- Workers felt they were reasonably compensated. Wages ranged from $4-$10/hour.



Interesting result: Distribution of triplets from grid questions

- When viewed in terms of **quality per triplet,** triplets sampled via grid questions appear to do worse. However, the sheer quantity outweighs quality.



Takeaway 1.

Question design and embedding algorithm are complementary!

- We can get pretty far without changing the embedding algorithm.

- Are you asking the right thing? **Bad questions leave information on the table.**

Takeaway 2.

If you're collecting triplets, try using grid questions!

- Consider the trade-off between grid size and effort
- **Strategy**: Pick the largest grid size that workers are comfortable with at your price point, then ask them to select about half the items

Thanks!

- Mike Wilber, Sam Kwak, Jan Jakes, Tomas Matera, Edward Cheng, Vicente Malave
- Explore our food embeddings, and download the dataset!

http://vision.cornell.edu/n2h3g