Rule Induction from Monolingual Continuous Representation

Kai Zhao

Graduate Center
City University of New York

Hany Hassan

Microsoft Research

Michael Auli

Facebook AI Research

the explosive dog

$\boxed{\text{搜爆}}$犬

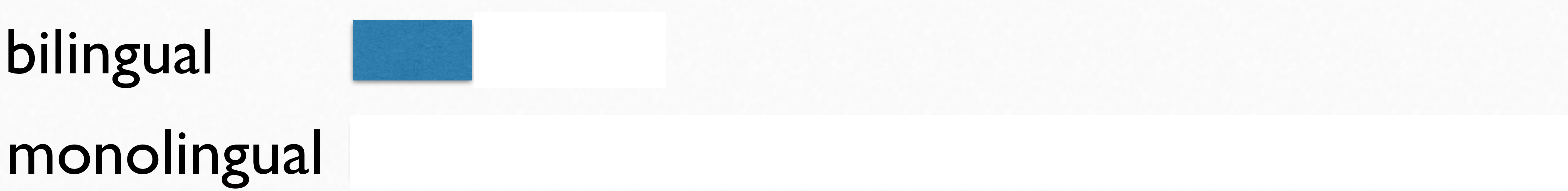
rare

the explosive dog

the explosive detection dog

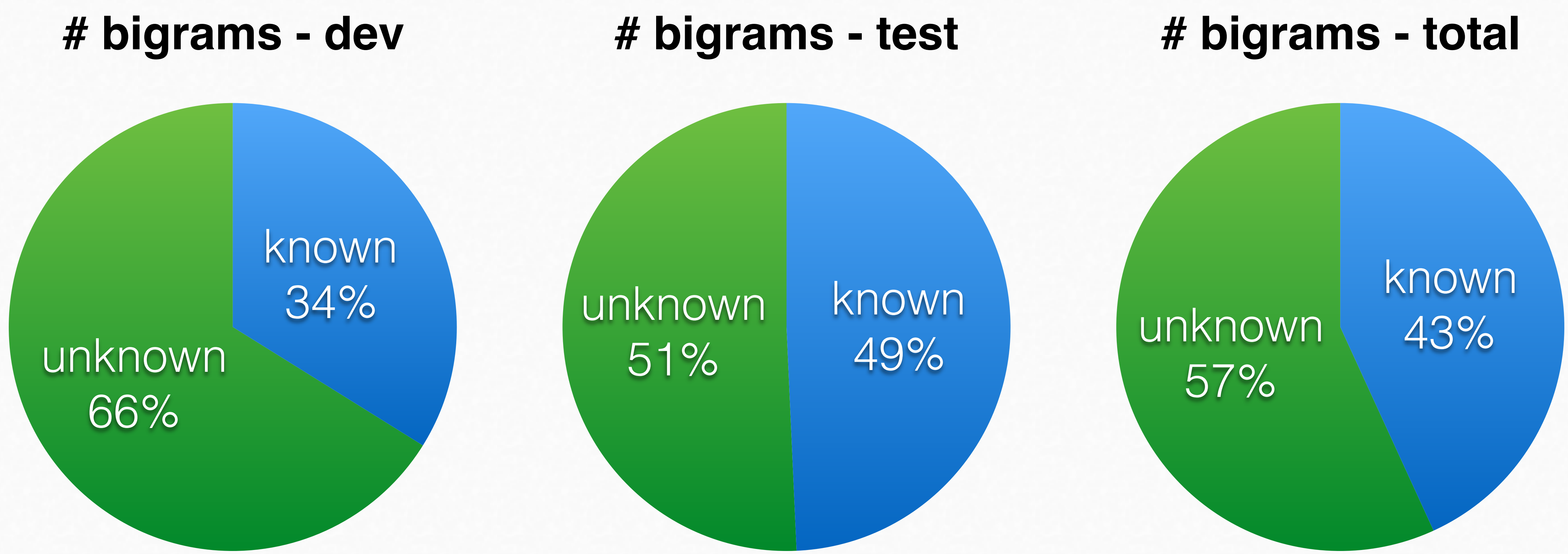
Monolingual Rule Induction

- Translation models are often learned from ***small*** bilingual corpora
  - for some language pairs, there are no large bitext corpora
  - unable to handle infreq./unseen phrases in dev/test set
- We have ***huge*** monolingual corpora
  - can be used to help improve translation models when combined with bilingual data



Unknown Phrases from Bilingual Data

- Rules induced from bilingual data
  - lots of unknown phrases in dev/test set
  - # of unknown bigrams

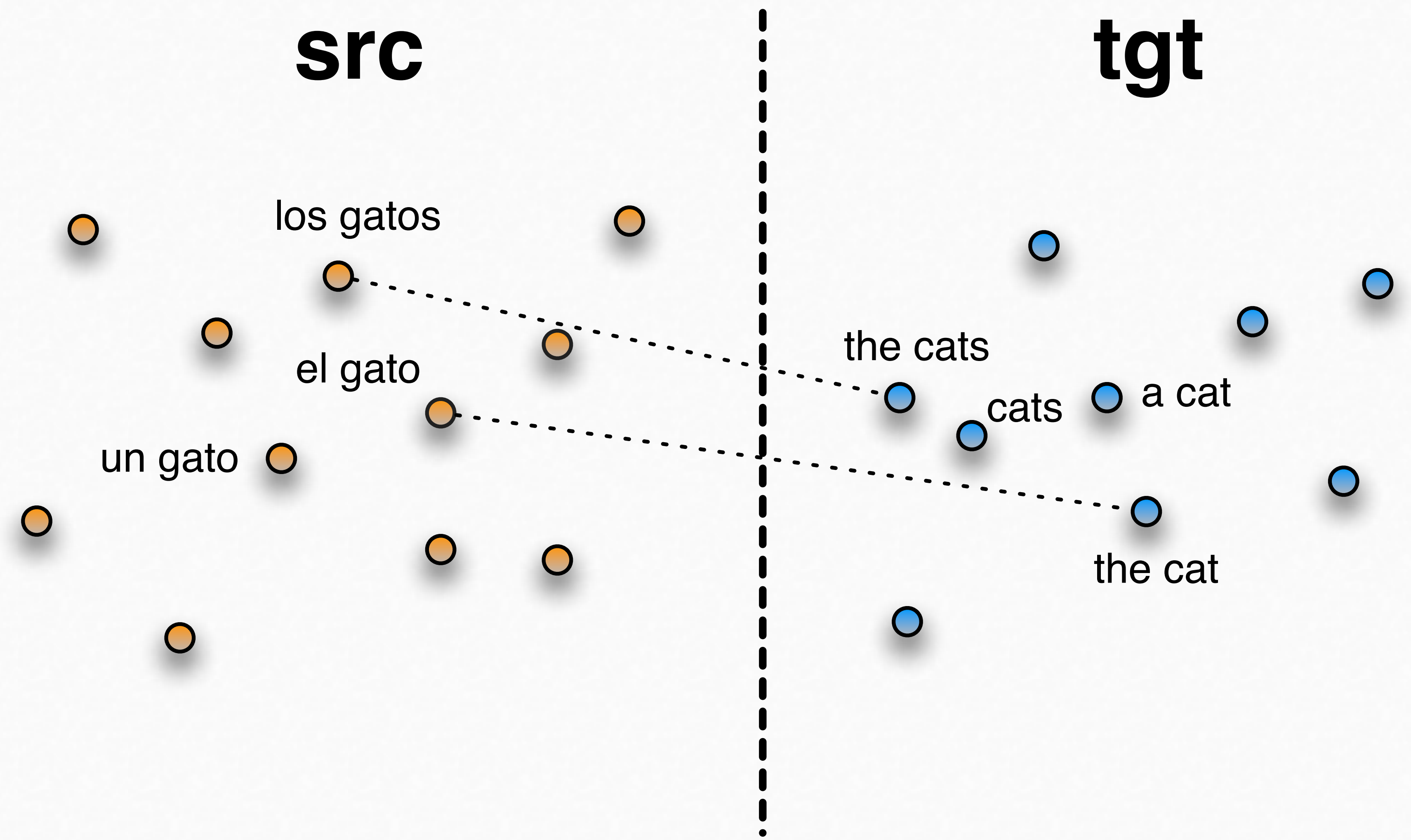


(Arabic - English; Saluja et al., 2014)

Monolingual Rule Induction

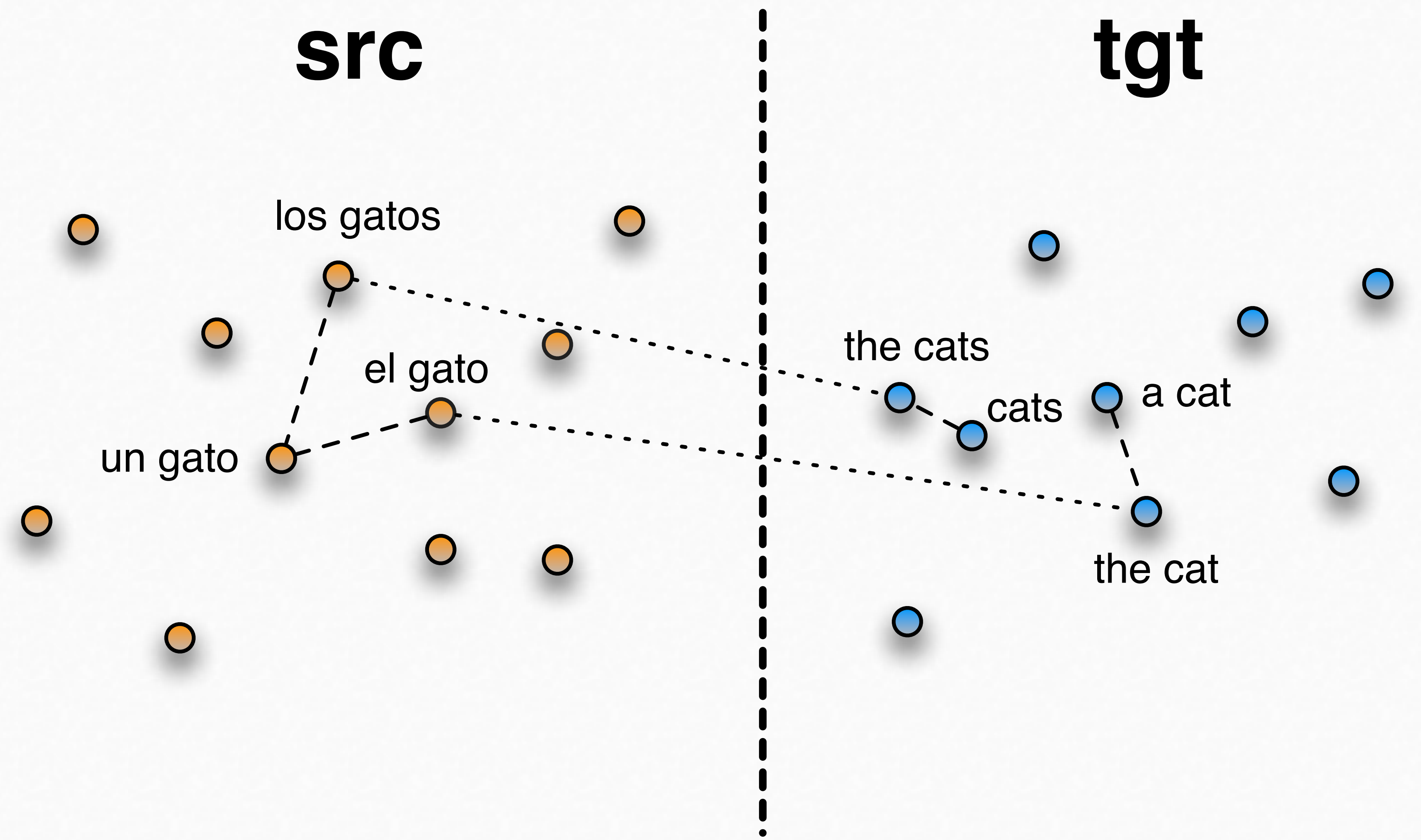
- Translation rule induction for infrequent phrases using **similar** phrases which we have a translation
- Infrequent phrases should be frequent enough in monolingual corpora
- How to model meaning **similarity**
  - phrases which occur in similar contexts (Saluja et al., 2014)
    - computationally expensive
  - continuous representations
    - e.g., Word Embeddings (Mikolov et al., 2013)

Monolingual Phrasal Embedding



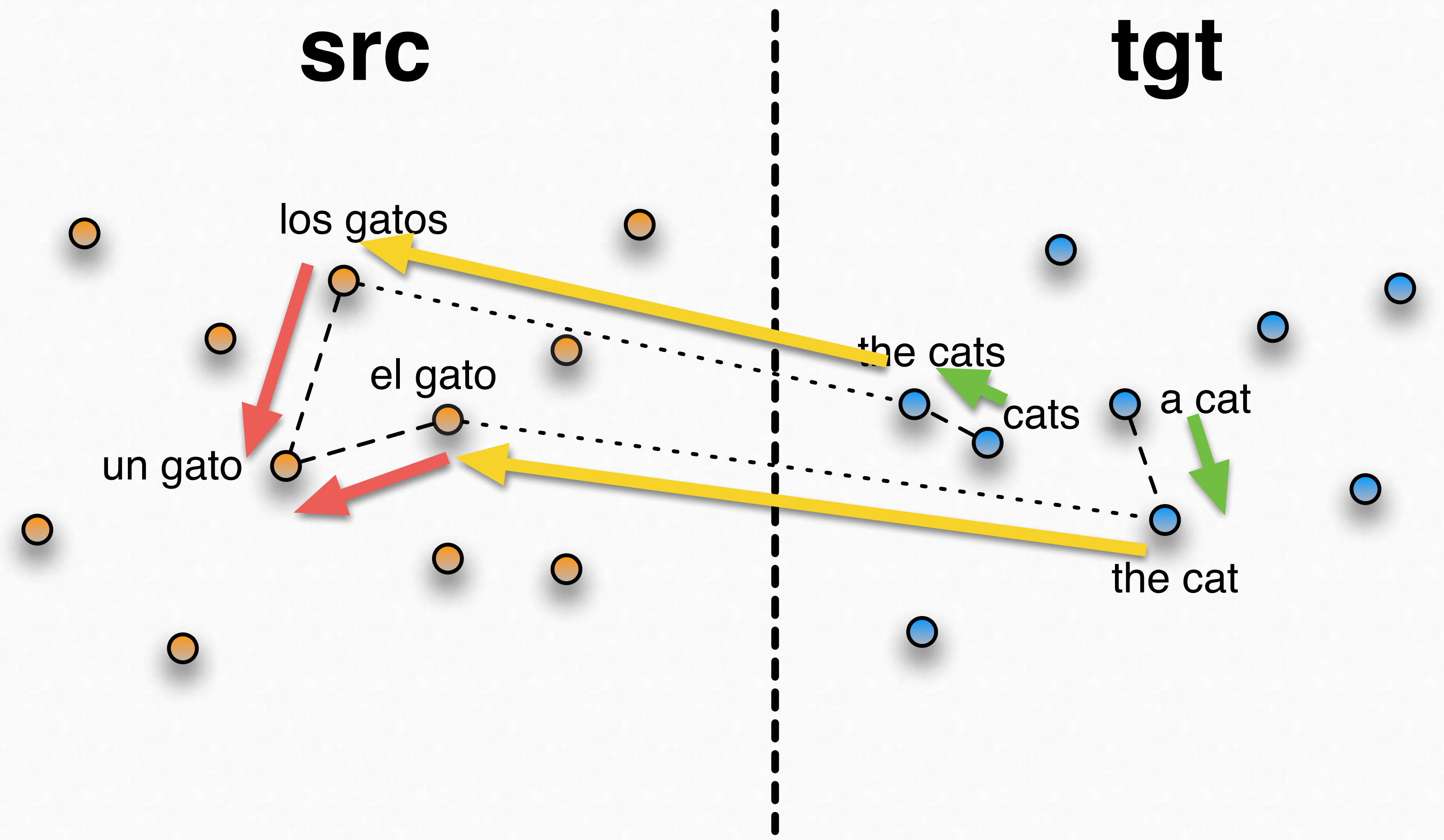
- Phrasal embeddings from monolingual corpora
  - combine word vectors via component-wise addition (Mitchell & Lapata, 2010)

Phrasal Embedding + SLP



- Structured Label Propagation (Saluja et al., 2014)
  - propagates correct translation candidates through labeled neighbors

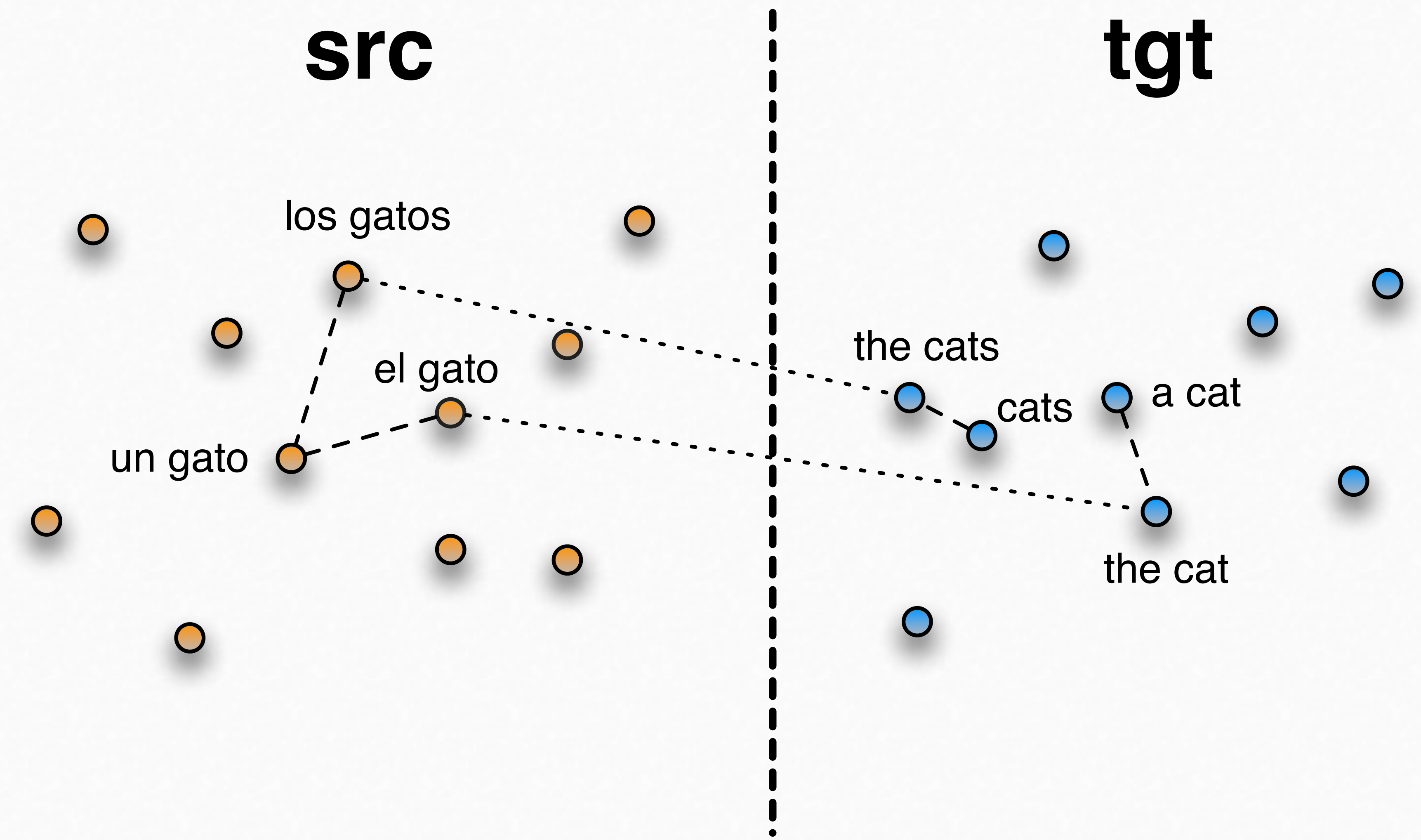
Phrasal Embedding + SLP



$$P^{t+1}(e|f) = \sum_{j \in N(f)} T_s(j|f) \sum_{e' \in H(j)} T_t(e'|e) P^t(e'|j)$$

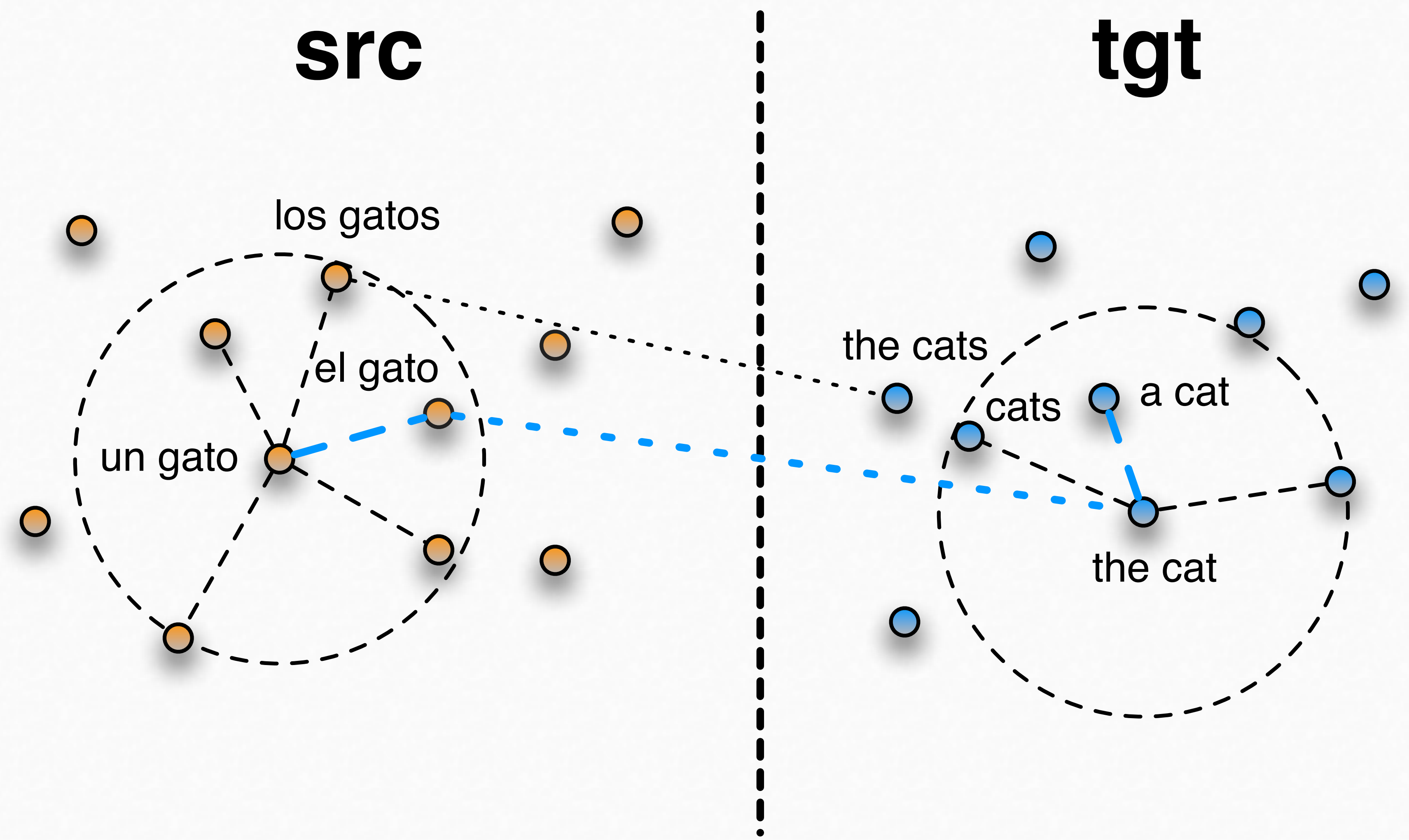
- Structured Label Propagation (Saluja et al., 2014)
  - propagates correct translation candidates through labeled neighbors

Phrasal Embedding + SLP



- How to define neighbors? How to find them?
  - Saluja et al., 2014: distributional similarity, contextual bag of words, PMI
  - $O(n^2)$ time linear search over the whole phrase space

Phrasal Embedding + SLP



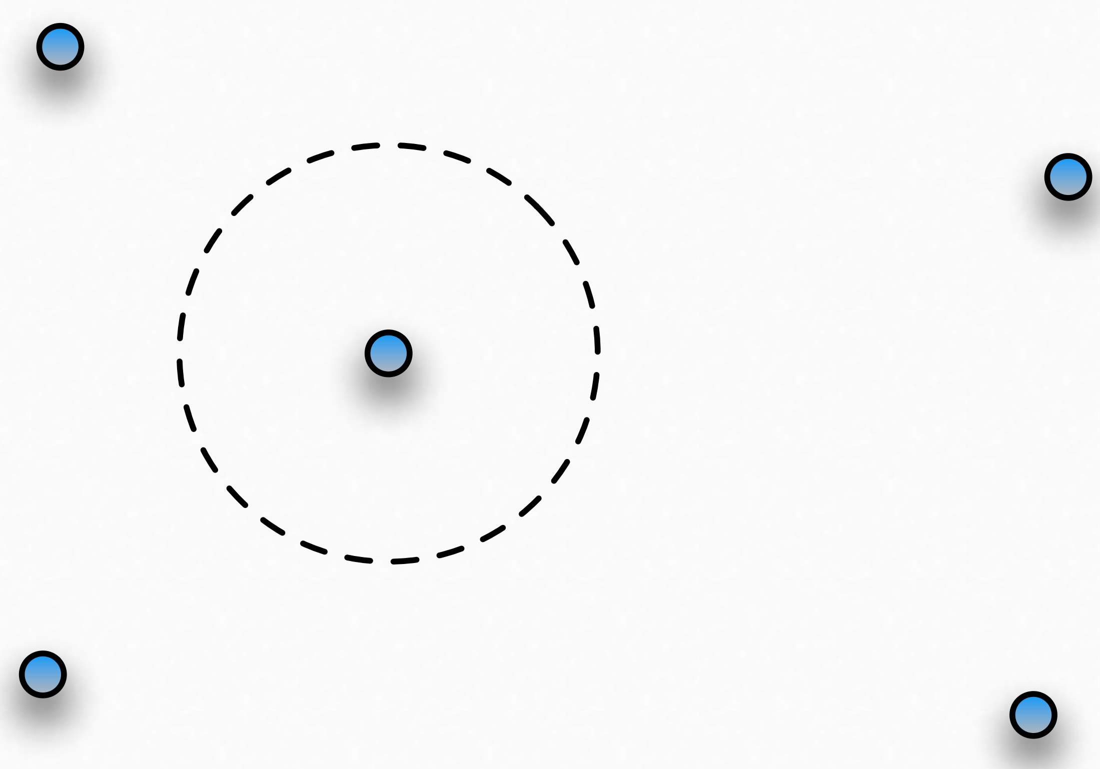
- How to define neighbors? How to find them?
  - Phrases with similar meanings are close in the continuous space
  - (Approximated) K nearest neighbor query

Approximated *k*-NN

- Locality sensitive hashing, LSH (Indyk & Motwani, 1998)
  - based on random projections
- Redundant bit vectors, RBV (Goldstein et al., 2005)
  - designed for computer vision tasks
  - split each dimension into slices, mark overlapping points w/ bit vectors
  - use bitwise *and* to fetch close points

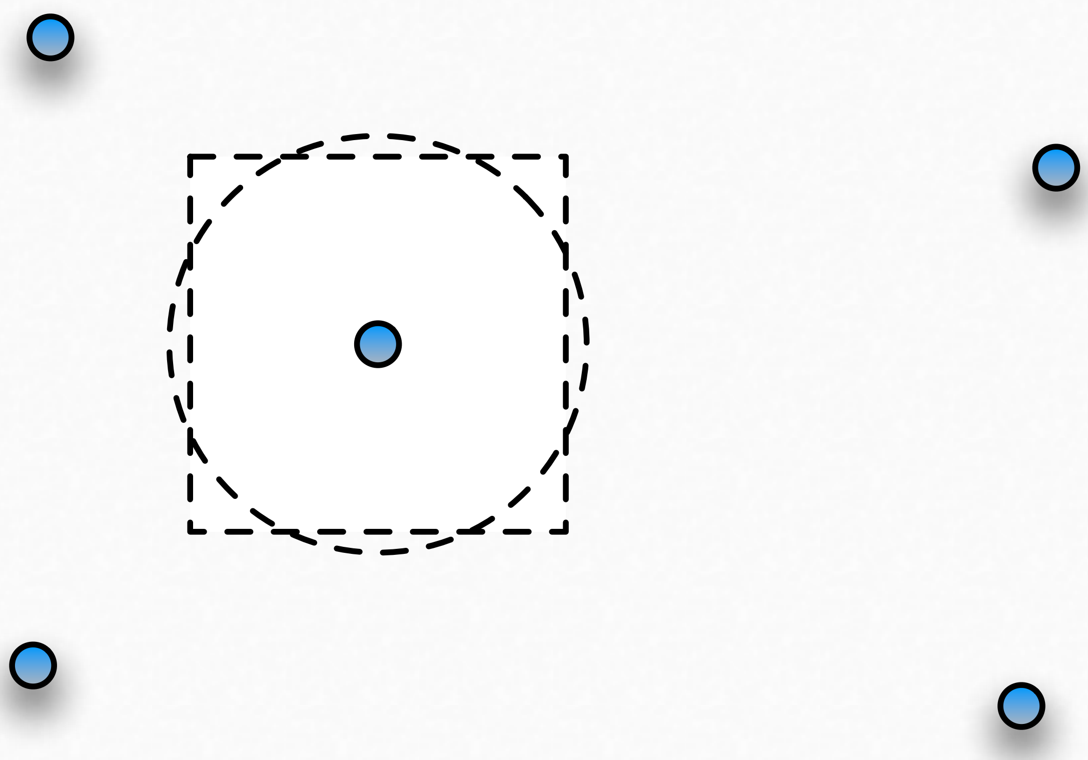
Approximated k-NN

- RBV



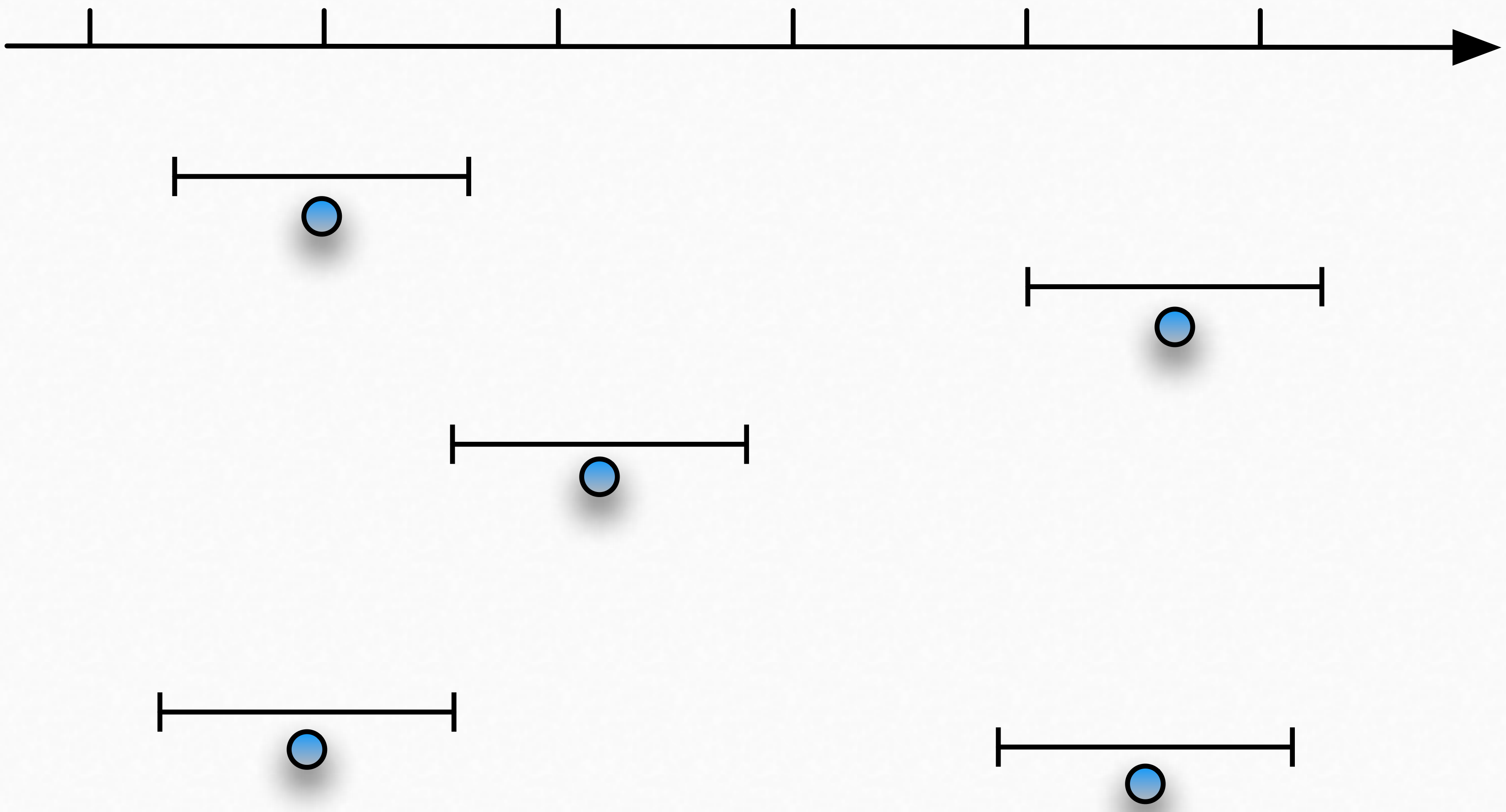
Approximated *k*-NN

- RBV: Hypercube instead of Hypersphere
- To cover 99% of the hypersphere, hypercube has smaller r
    - For 256d, hypercube only needs 1/3 r to cover 99% of the hypersphere
- neighboring test: in hypersphere => in hypercube => distance on each dim



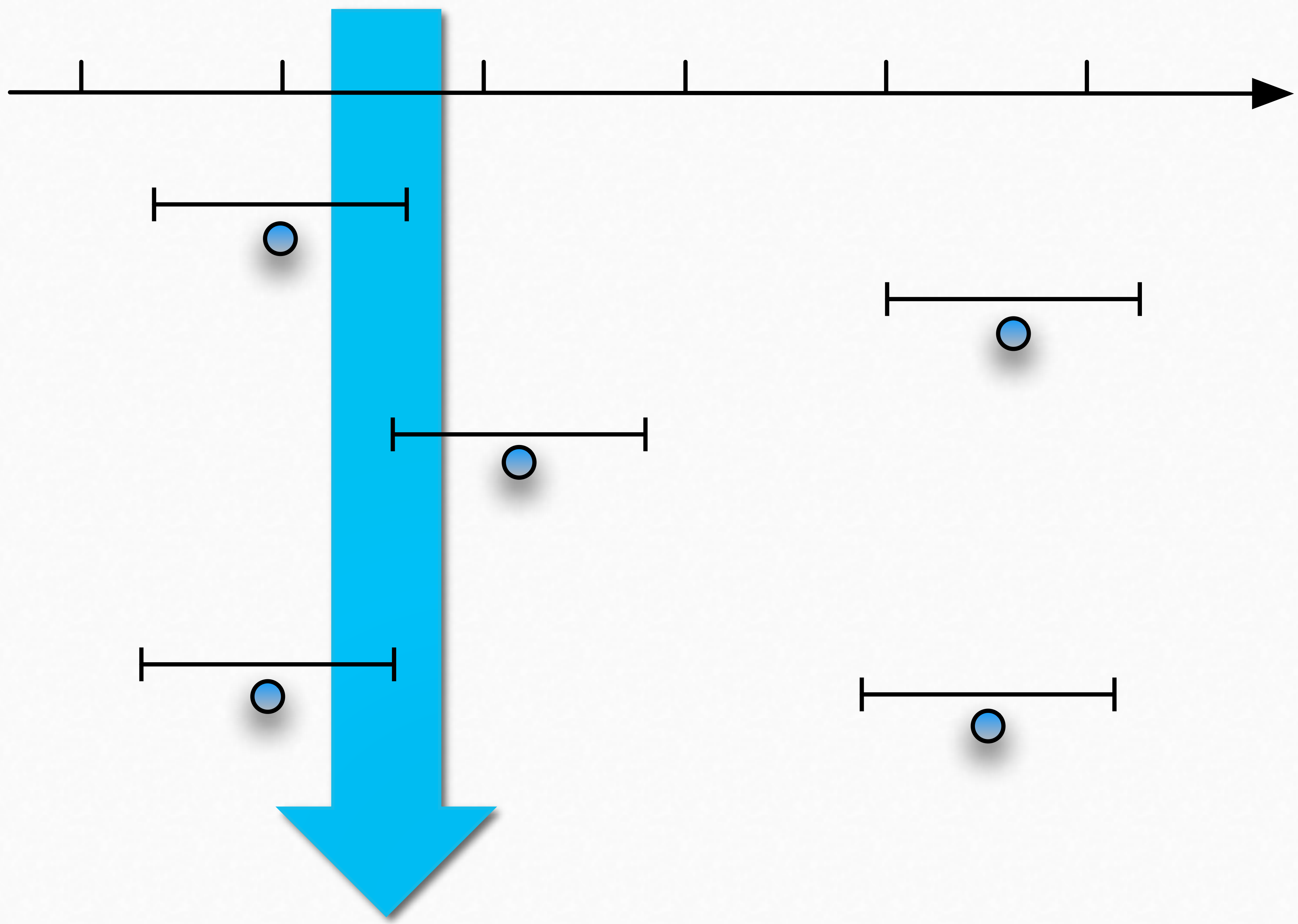
Approximated *k*-NN

- RBV: Split each dimension into slices



Approximated *k*-NN

- RBV: Querying by bitwise *and* over dimensions

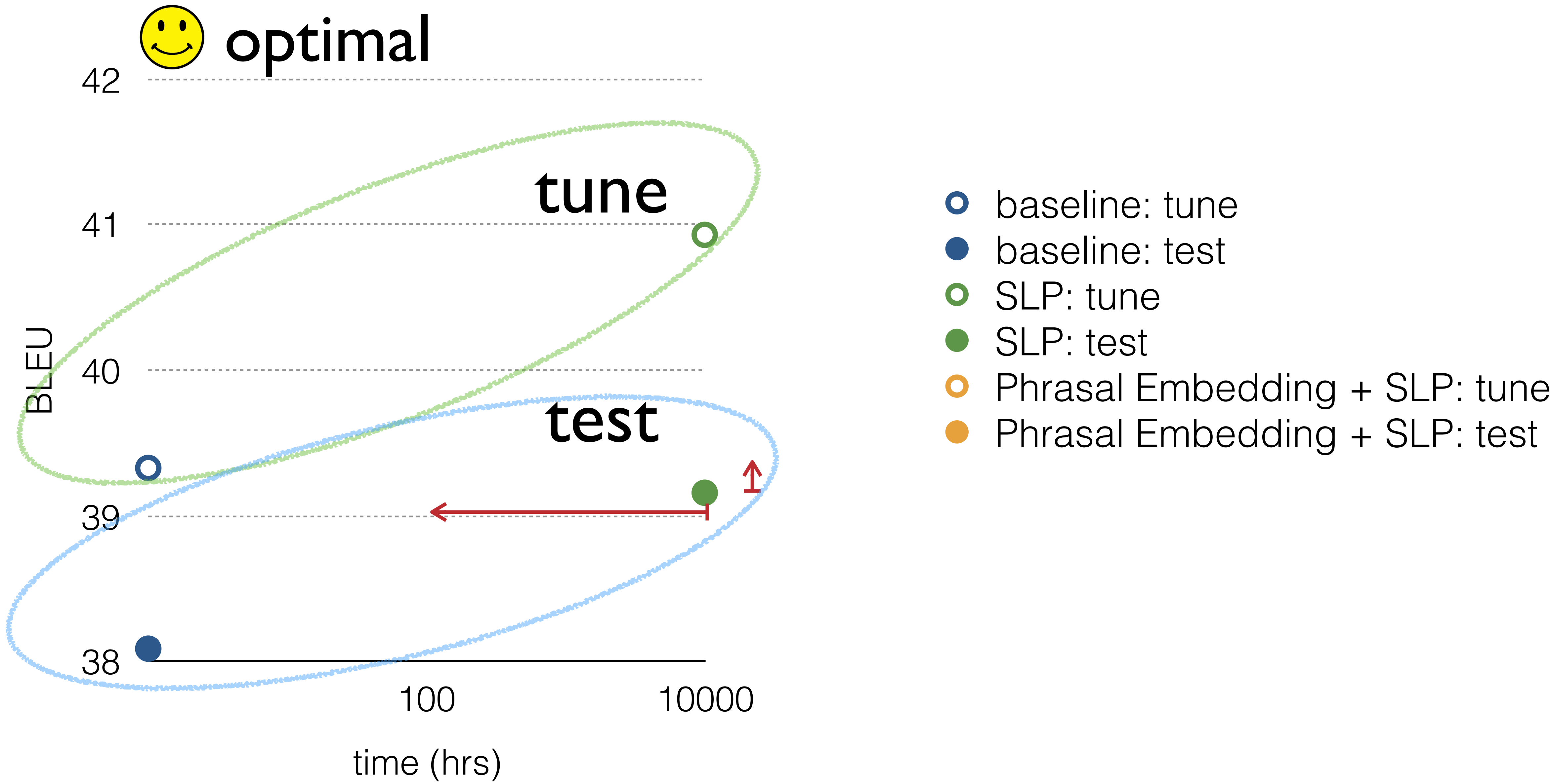


Approximated *k*-NN: Performance

- 951,453 word embedding vectors
- 200 dimensions
- Test on 100 words, k = 200 nearest neighbors
- False Negative Rate
  - true neighbors missed by *k*-NN
  - correct translations missed

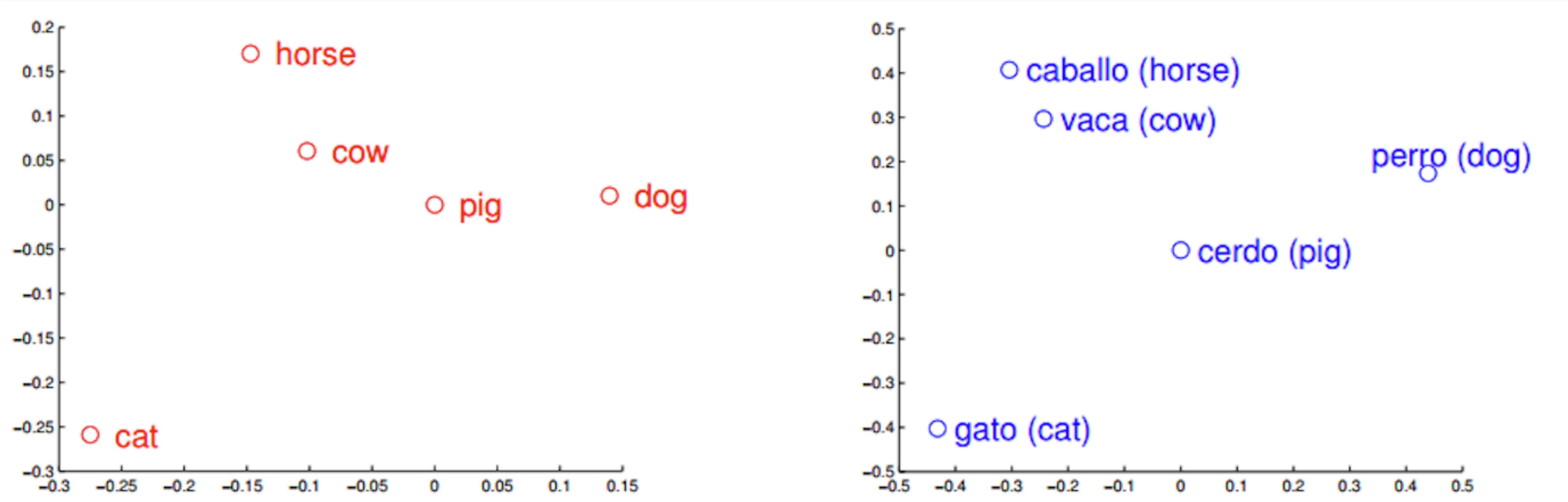
	False Negative	Time
Linear Search	0	342s
LSH	14.29%	69s
RBV	**9.08%**	**19s**

Phrasal Embedding + SLP: Performances



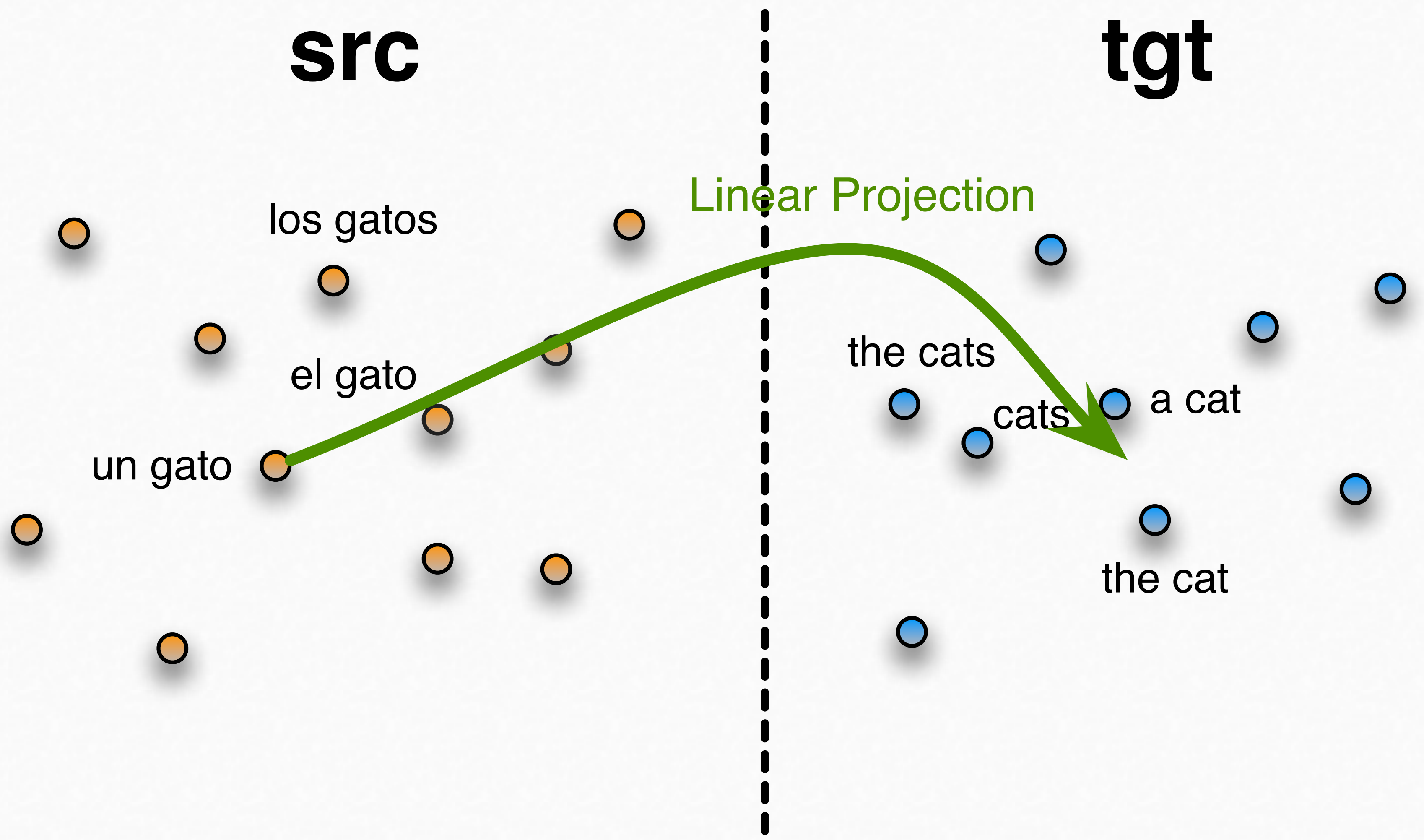
- Phrasal Embedding + SLP
  - 100 times faster than vanilla SLP
  - slightly better in translation quality than vanilla SLP

Direct Projection



- The relative positions of different words are similar between different languages (Mikolov et al., 2013)
  - trained on most frequent words
  - Linear Projection?

Direct Projection

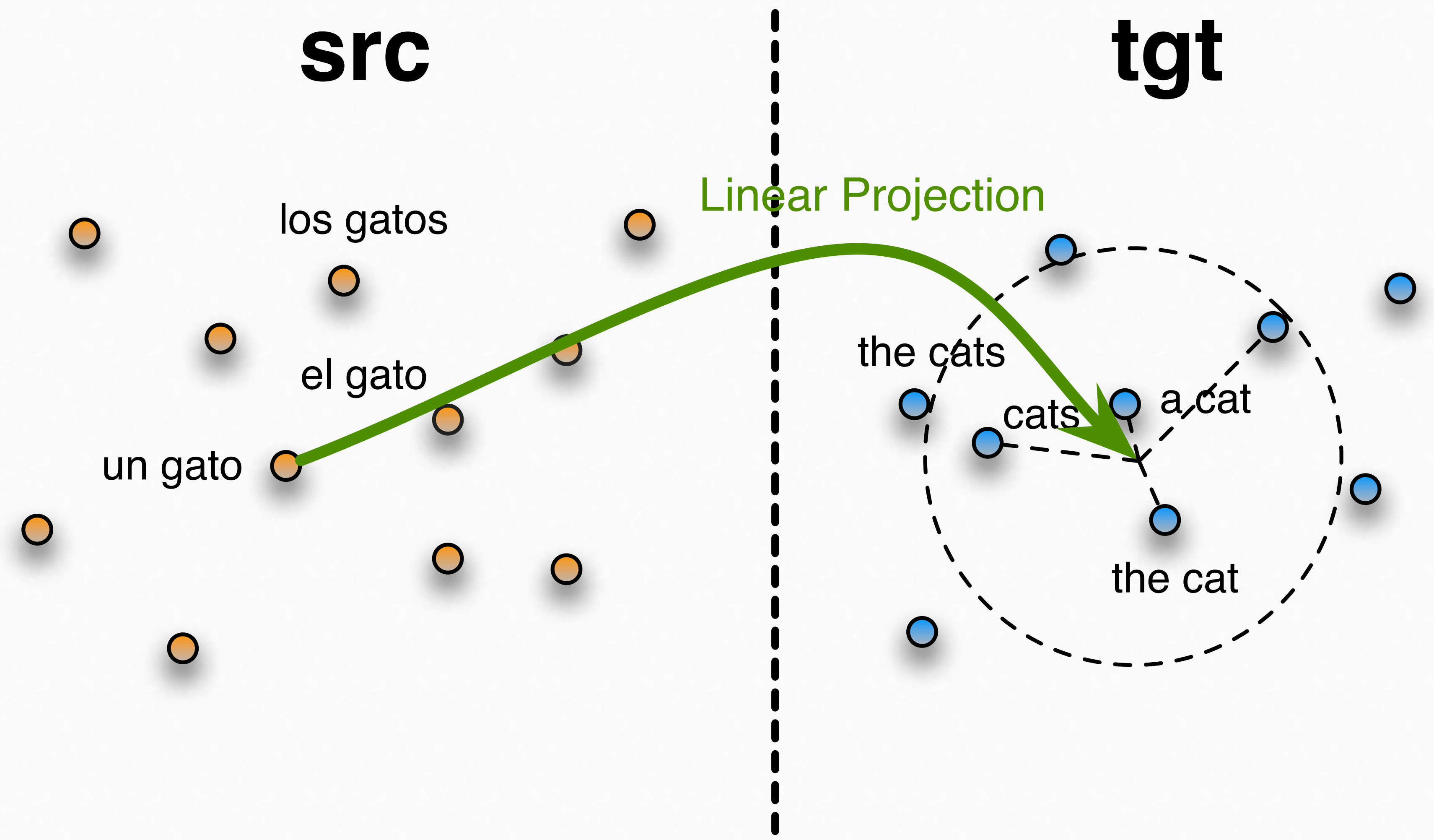


- Project embeddings of infrequent phrase to the target space
  - Projection can be learned by solving linear system

$$XW \approx Y$$

$$W \approx (X^T X)^{-1} X^T Y$$

Global Linear Projection



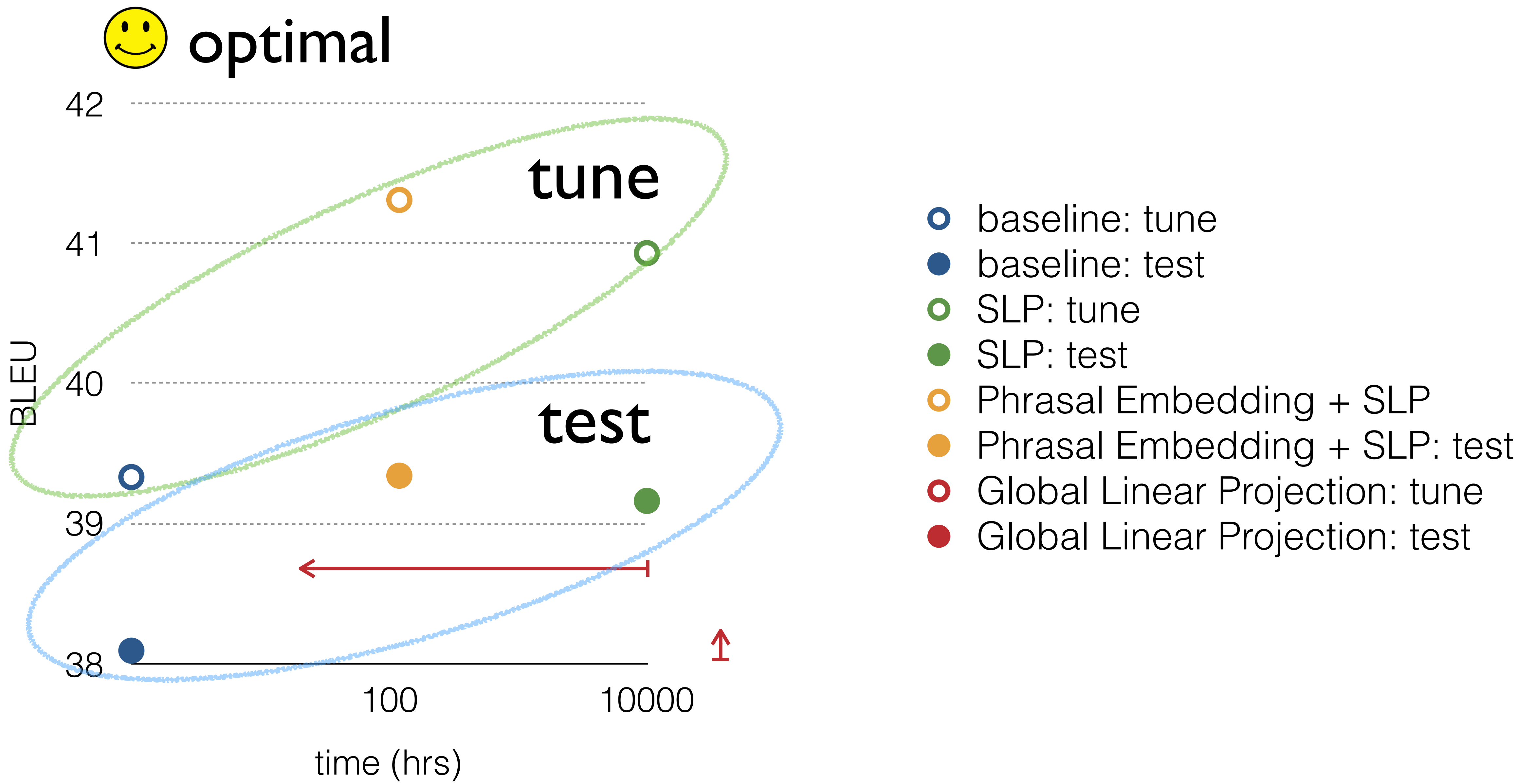
- Project embeddings of infrequent phrase to the target space
  - Projection can be learned by solving linear system

$$XW \approx Y$$

$$W \approx (X^T X)^{-1} X^T Y$$

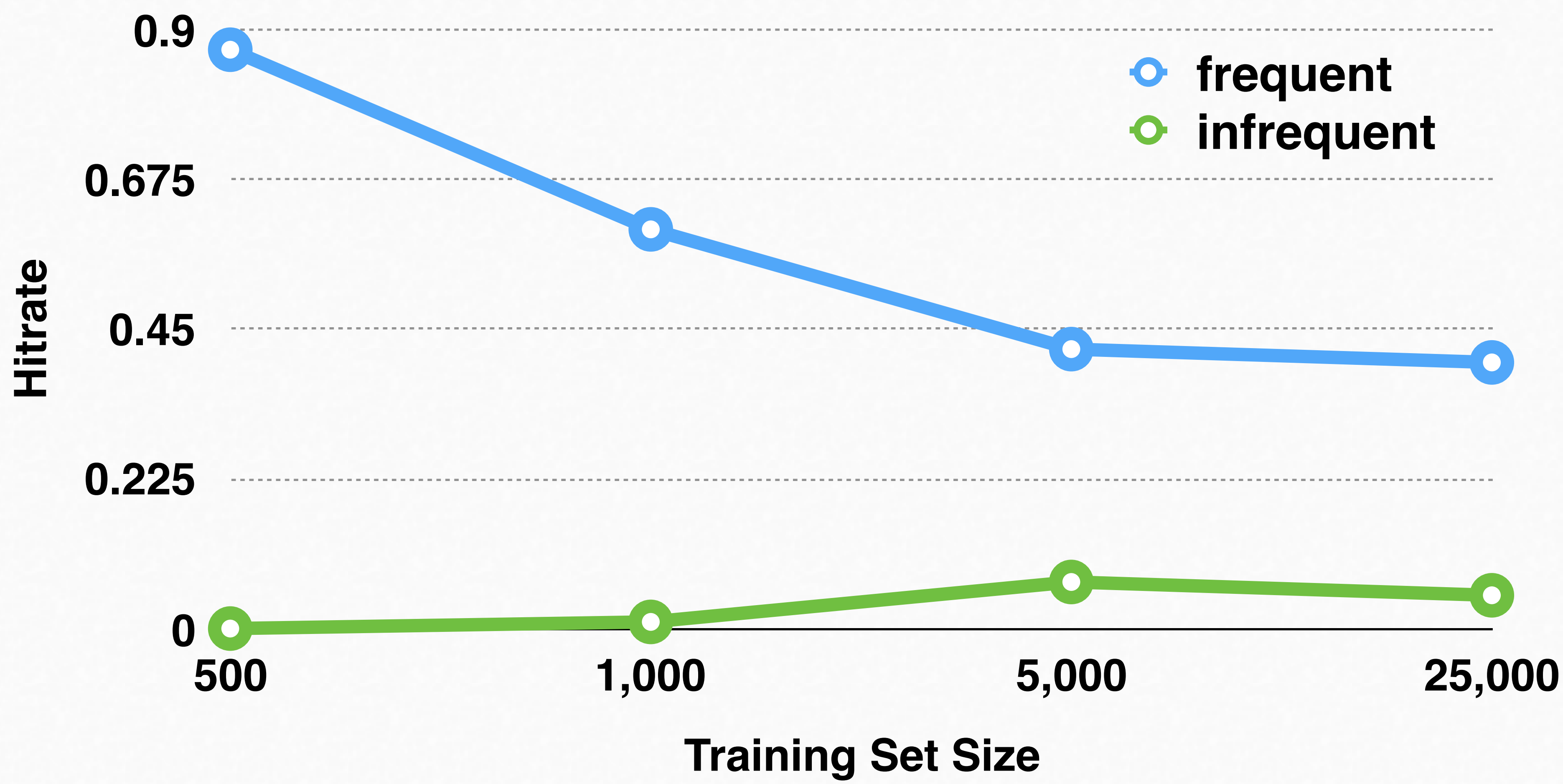
- Query *k*-NN as translation candidates

Global Linear Projection: Performance



- Global Linear Projection
  - 500 times faster than vanilla SLP
  - only slightly better in translation quality than baseline

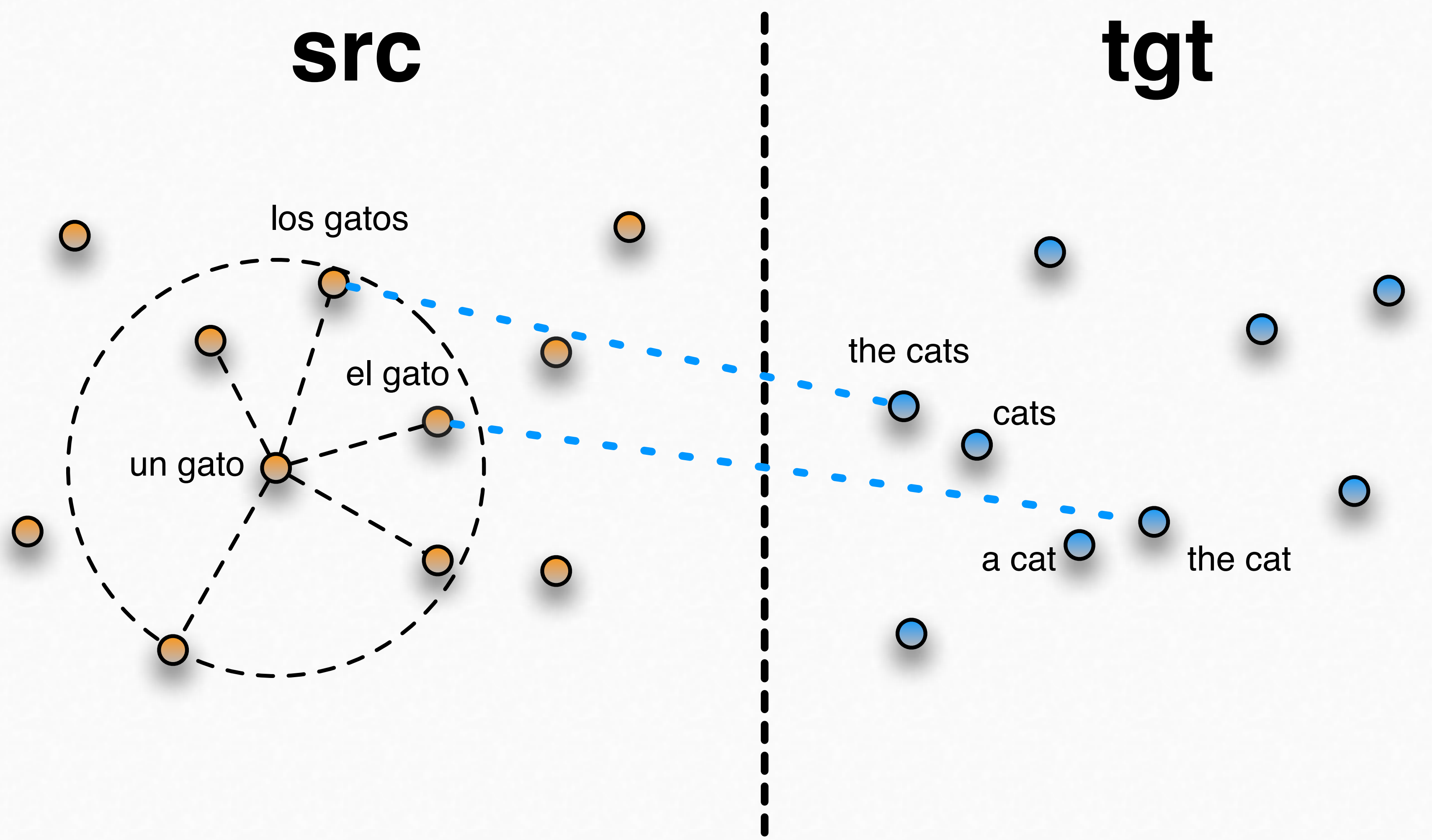
Global Linear Projection: Projection Quality



- Optimal Linear Projection trained on most frequent words
- Quality of the projection is evaluated on two sets: frequent & infrequent
- Hit rate: probability that the correct translation is fetched by *k*-NN of the projected point (k = 200)

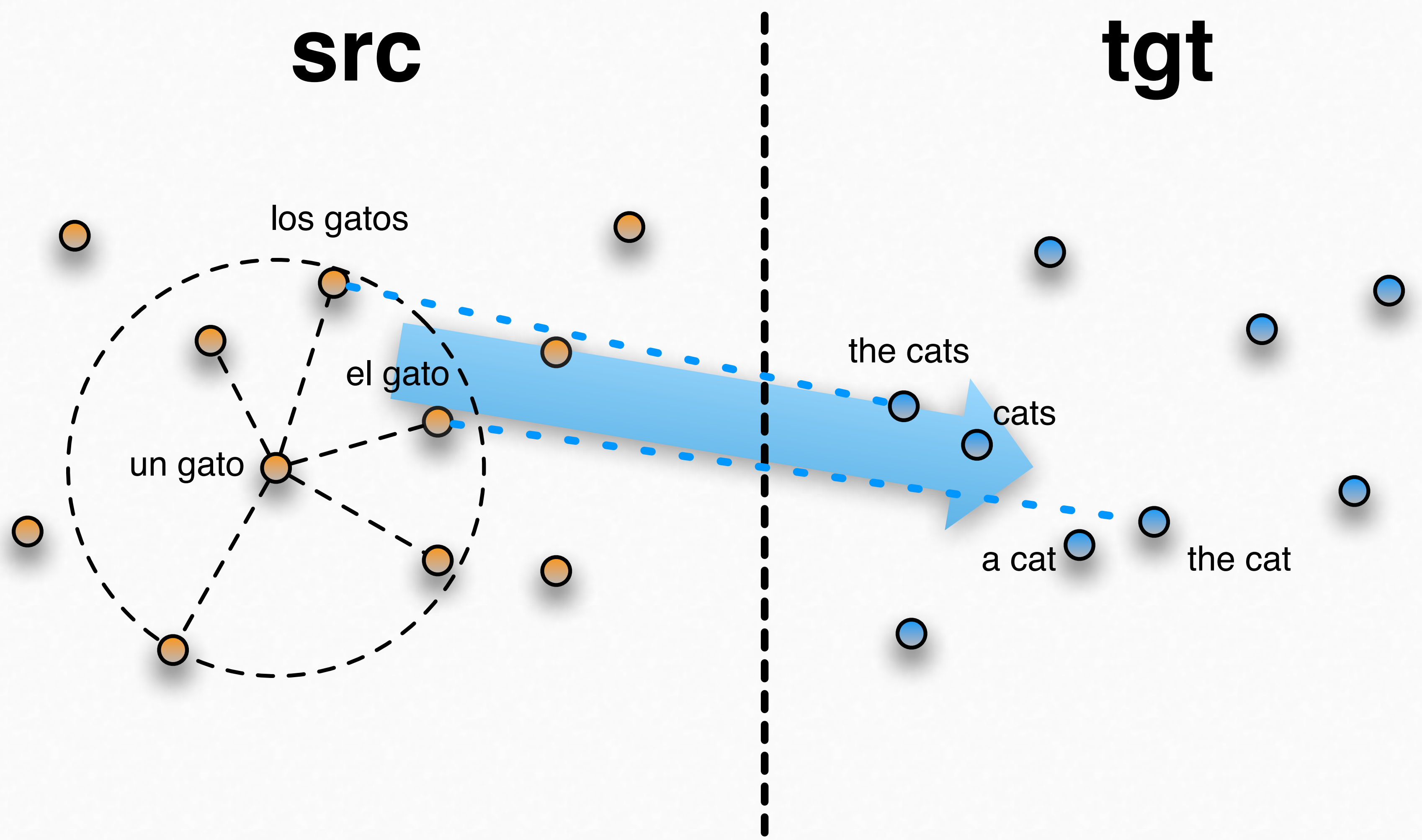
Direct Projection: Global => Local

- Global linear projection is noisy for infrequent phrases
- Linear projection likely to be more accurate for the subsets of the data
  - idea: use many **local** projections instead of a single global projection
  - analogous to Locality Perserving Projections (He & Niyogi, 2004)



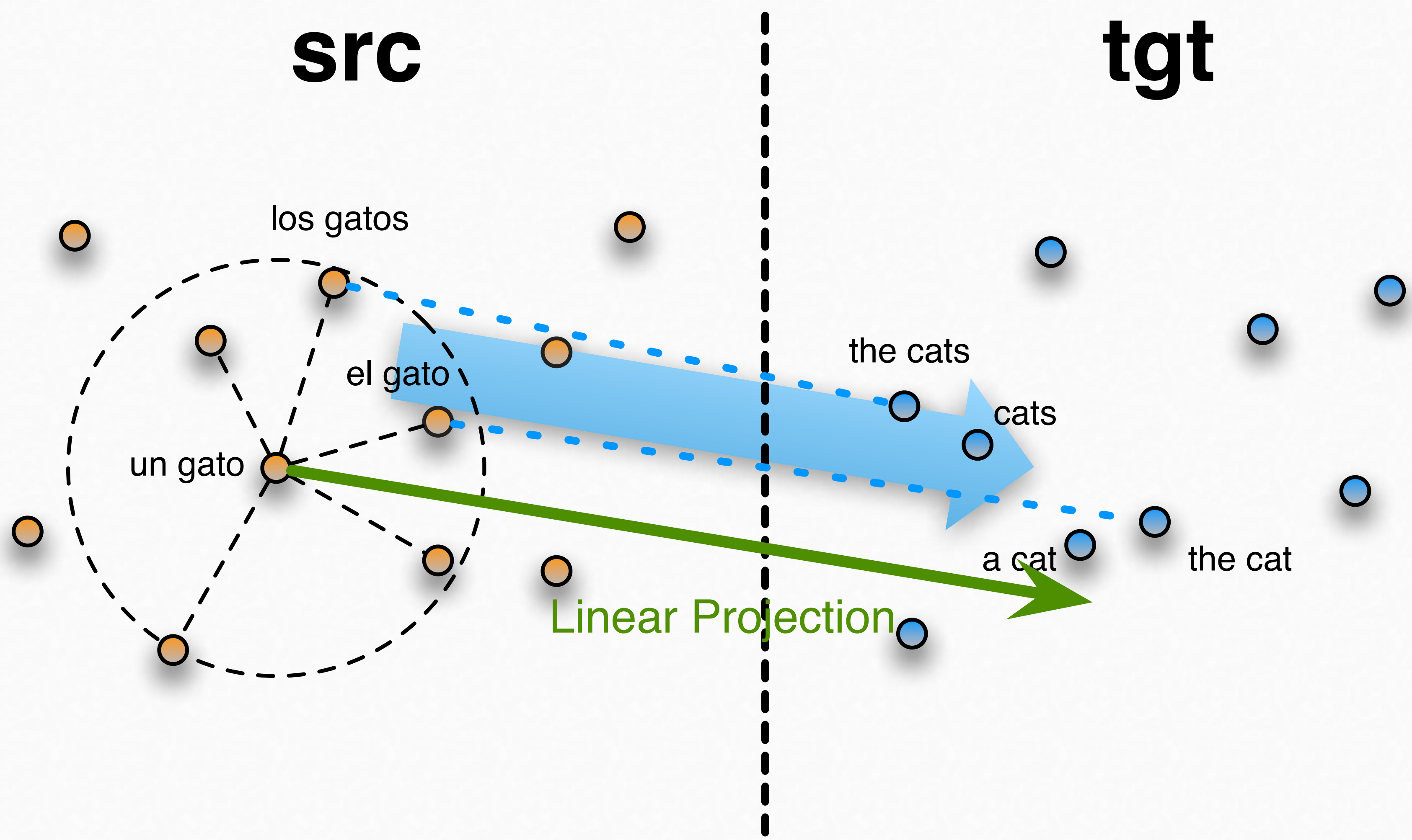
Local Linear Projection

- Global linear projection is noisy for infrequent phrases
- Linear projection likely to be more accurate for the subsets of the data
  - idea: use many **local** projections instead of a single global projection
  - analogous to Locality Perserving Projections (He & Niyogi, 2004)



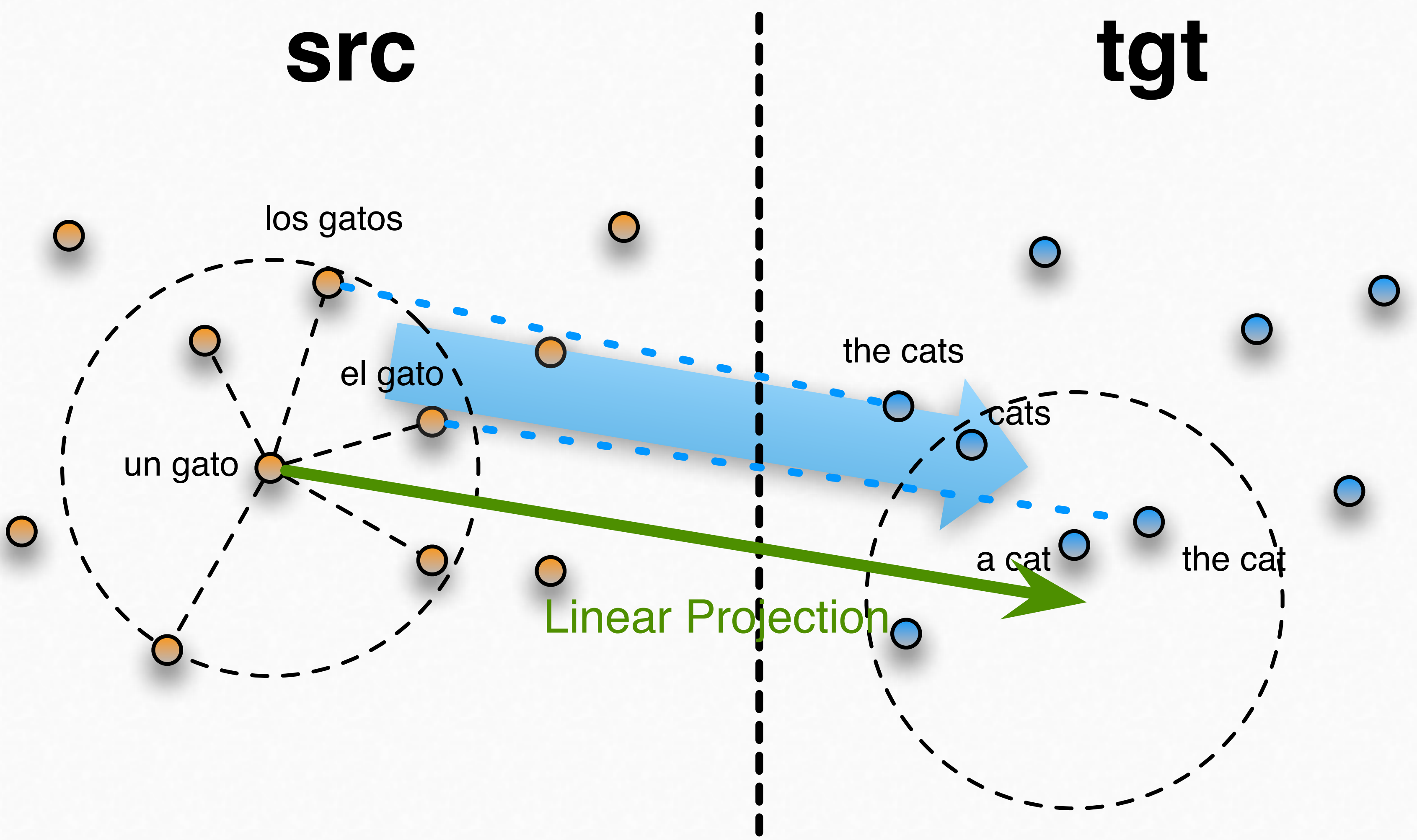
Local Linear Projection

- Global linear projection is noisy for infrequent phrases
- Linear projection likely to be more accurate for the subsets of the data
    - idea: use many **local** projections instead of a single global projection
    - analogous to Locality Perserving Projections (He & Niyogi, 2004)

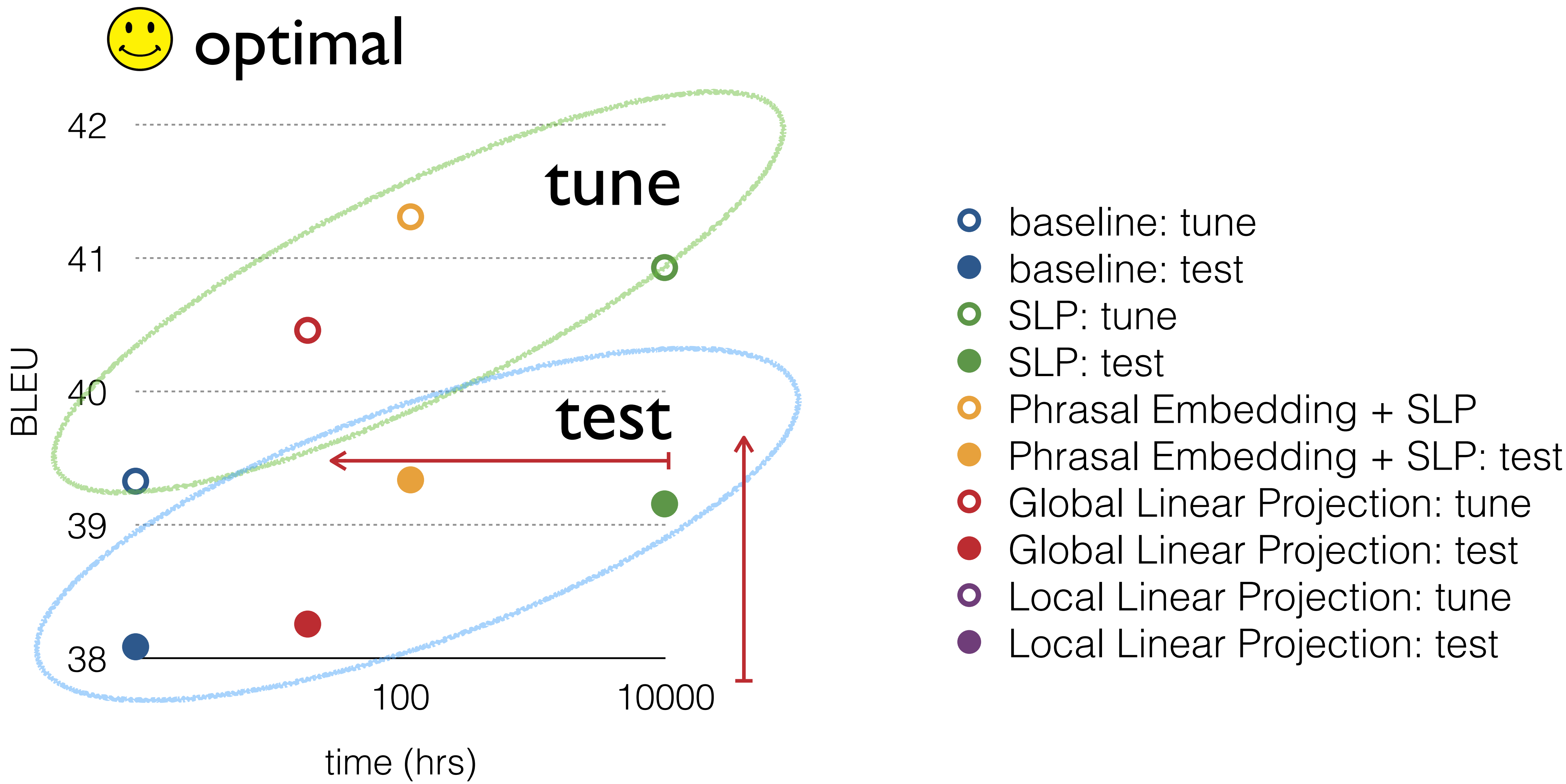


Local Linear Projection

- Global linear projection is noisy for infrequent phrases
- Linear projection likely to be more accurate for the subsets of the data
  - idea: use many **local** projections instead of a single global projection
  - analogous to Locality Perserving Projections (He & Niyogi, 2004)



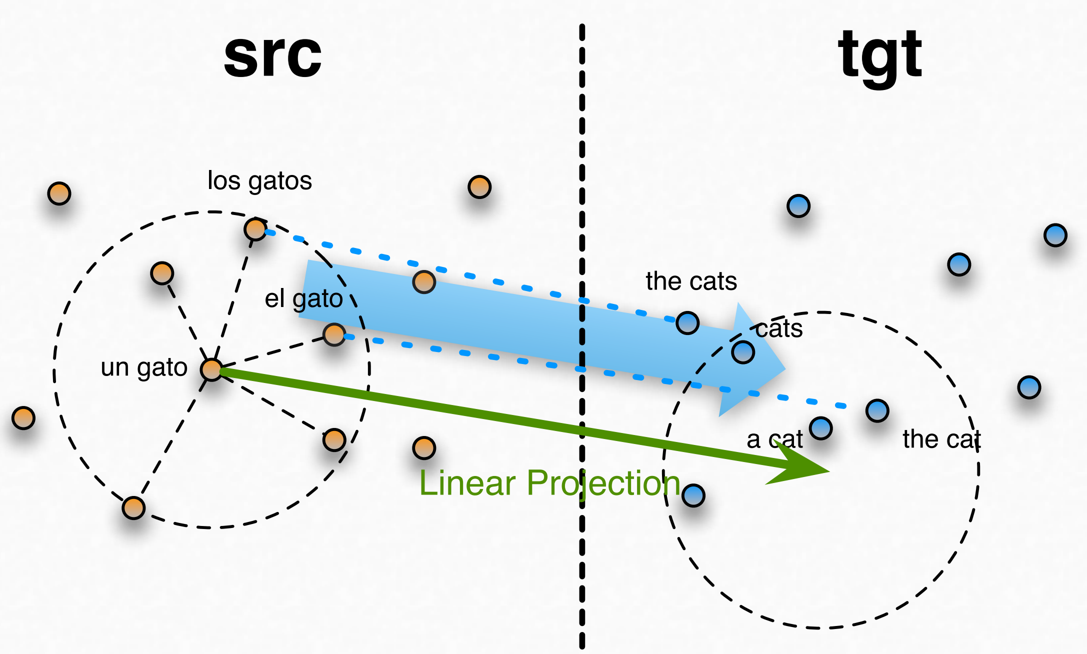
Local Linear Projection: Performances



- Local Linear Projection
  - 400 times faster than vanilla SLP
  - best performance over all

Conclusion

- Introduced a simple set of linear projections to learn new translations
- Projections 400x times faster than SLP at the same accuracy
- A single global projection is vulnerable to noise
- Demonstrated RBV as a fast and accurate alternative to LSH
- Non-Linear Projection? Contextual Information?



FIN

Thank you!
Questions?