Cross-linguality and machine translation without bilingual data

Eneko Agirre
@eagirre

Joint work with: Mikel Artetxe, Gorka Labaka

IXA NLP group – University of the Basque Country (UPV/EHU)

Motivation

Previous work on **cross-lingual word representations:**

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Motivation

Previous work on **cross-lingual word representations:**

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: **extend mappings to any pair of languages**

- Most language pairs have **very few bilingual resources**
- Key research area for **wide adoption** of NLP tools

Motivation

Previous work on **cross-lingual word representations:**

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: **extend mappings to any pair of languages**

- Most language pairs have **very few bilingual resources**
- Key research area for **wide adoption** of NLP tools

In particular: **no bilingual resources at all**

- **Unsupervised** embedding mappings
- **Unsupervised** neural machine translation

Overview

Arabic monolingual corpora

Arabic embeddings

Chinese monolingual corpora

Chinese embeddings

Overview

Arabic monolingual corpora

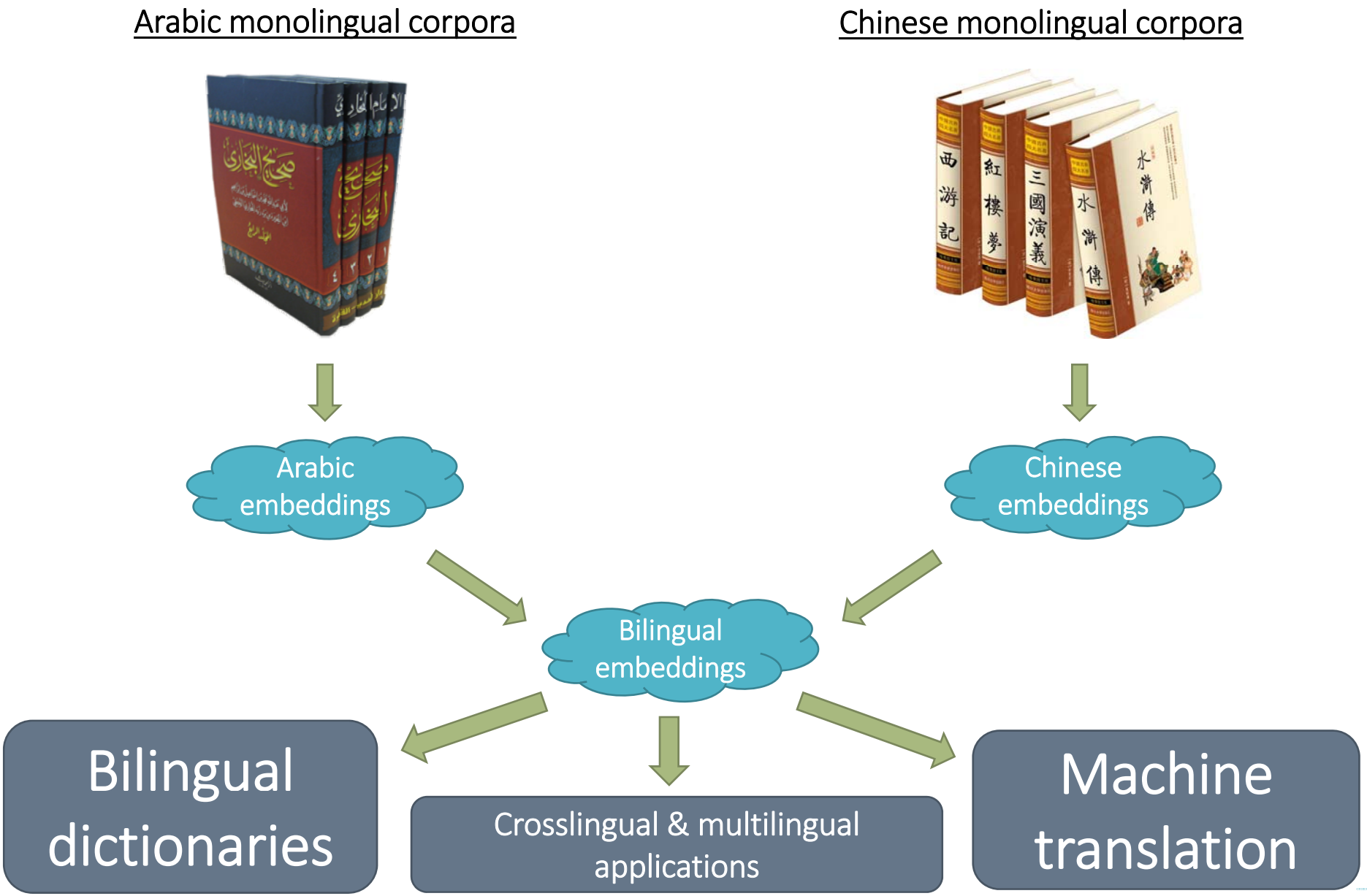
Chinese monolingual corpora

Arabic embeddings

Chinese embeddings

Bilingual embeddings

Overview


Arabic monolingual corpora
Chinese monolingual corpora
Arabic embeddings
Chinese embeddings
Bilingual embeddings
Bilingual dictionaries
Crosslingual & multilingual applications
Machine translation

Overview

Arabic monolingual corpora

No bilingual resource

Chinese monolingual corpora

Arabic embeddings

Chinese embeddings

Bilingual embeddings

Bilingual dictionaries

Crosslingual & multilingual applications

Machine translation

Outline

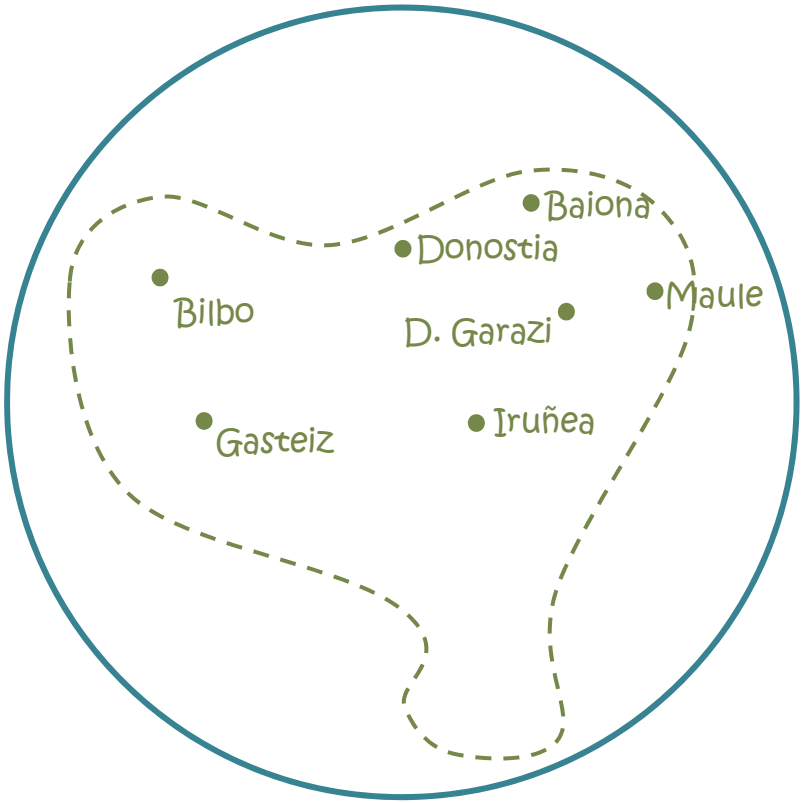
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Introduction to bilingual embedding mappings*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Conclusions*

Outline

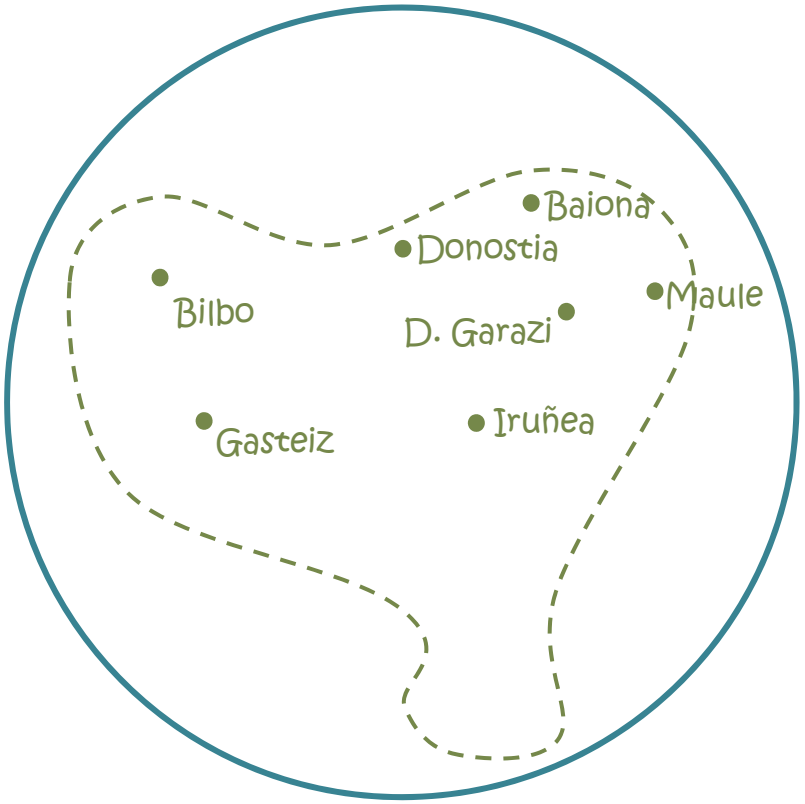
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Introduction to bilingual embedding mappings*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Conclusions*

Introduction to vector space models

Introduction to vector space models

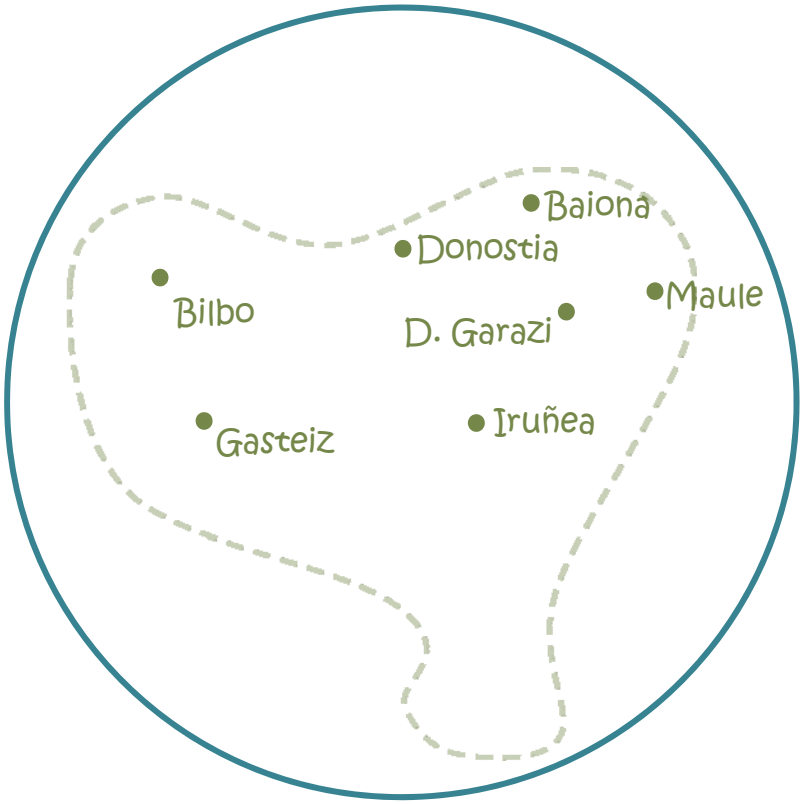


Introduction to vector space models



Geographical space

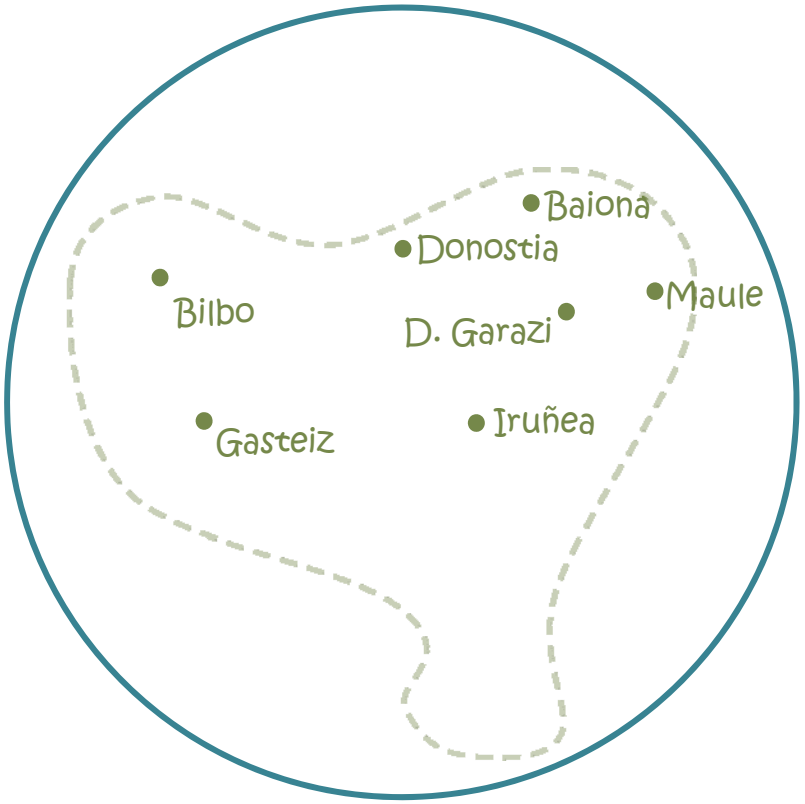
Introduction to vector space models



Geographical space

- Cities

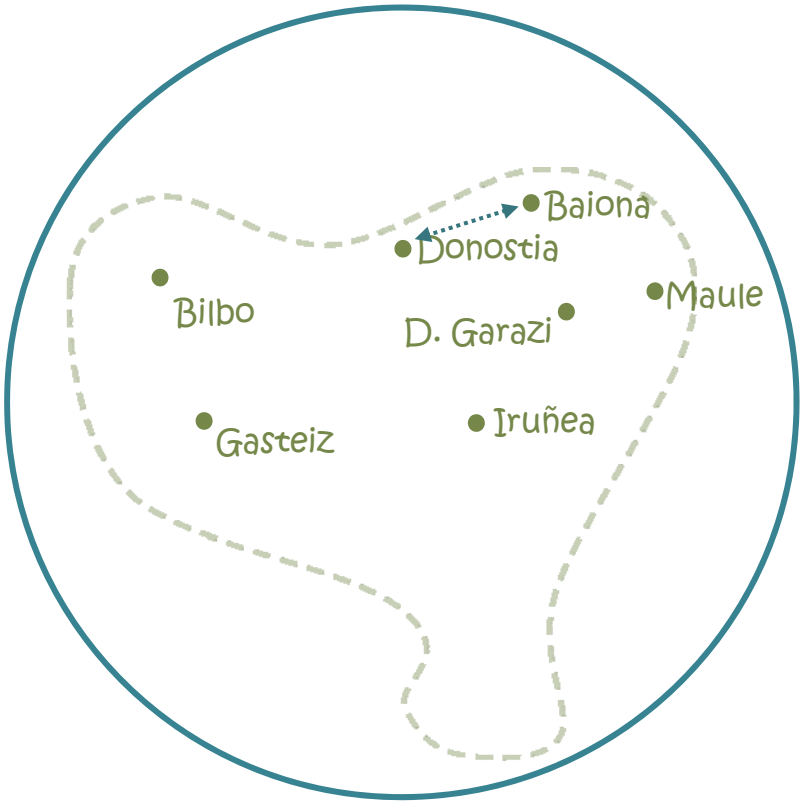
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances

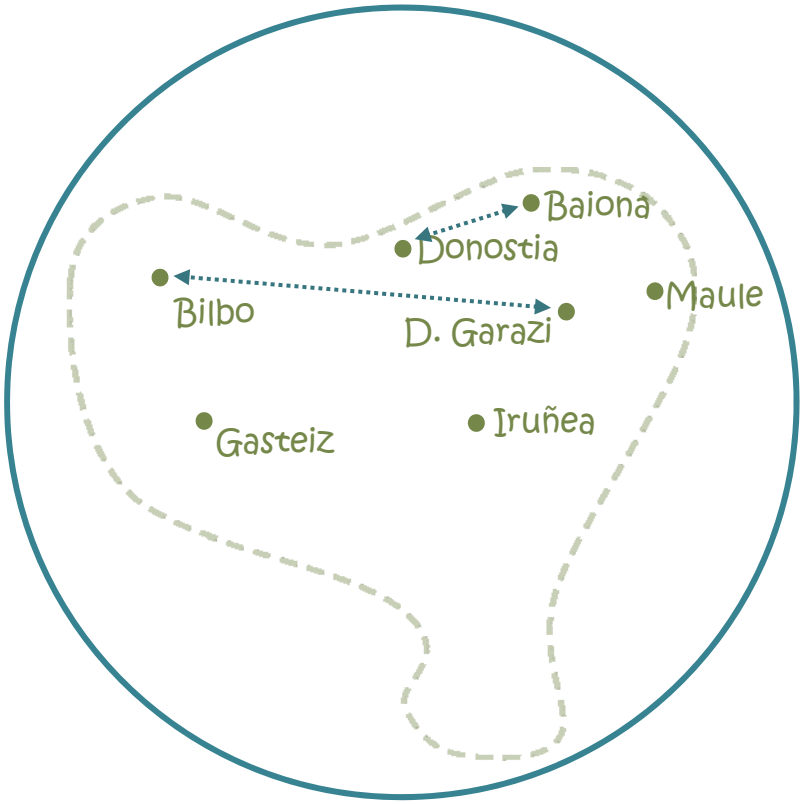
Introduction to vector space models



<u>Geographical space</u>

- Cities
- Meaningful distances

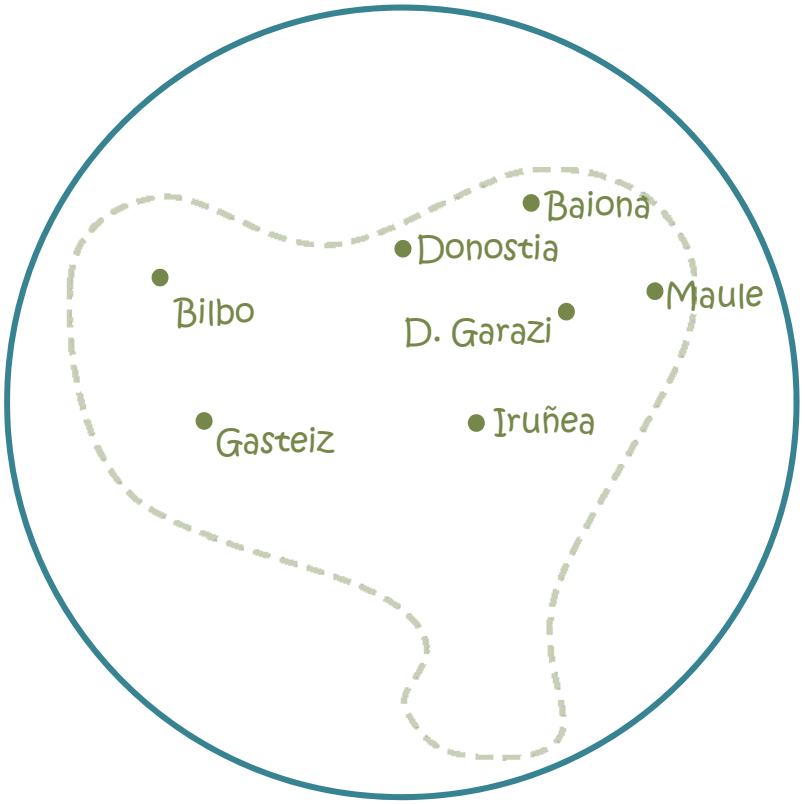
Introduction to vector space models



<u>Geographical space</u>

- Cities
- Meaningful distances

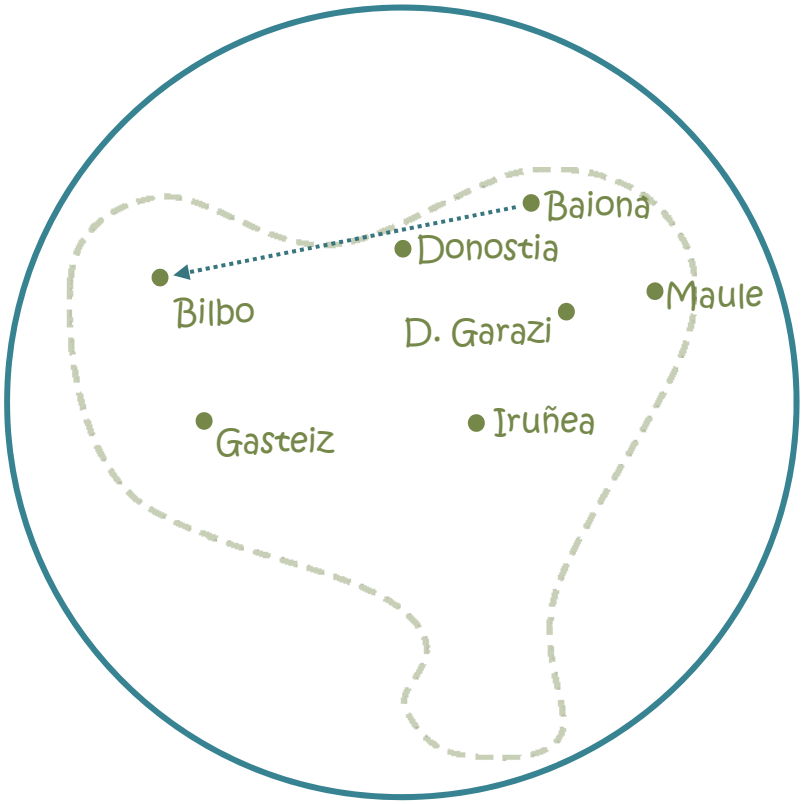
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances
- Meaningful relations

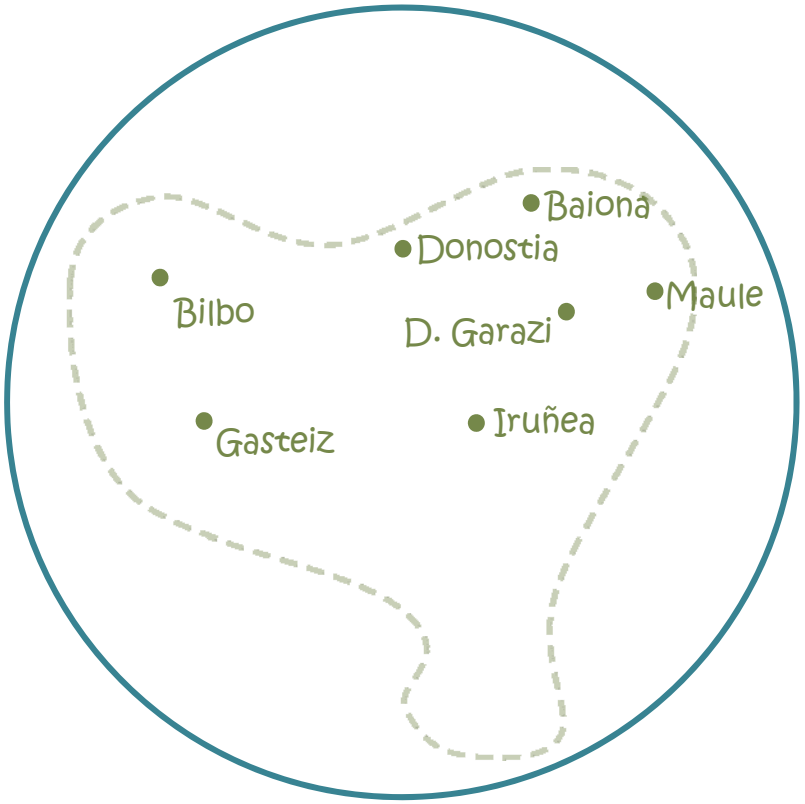
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances
- Meaningful relations

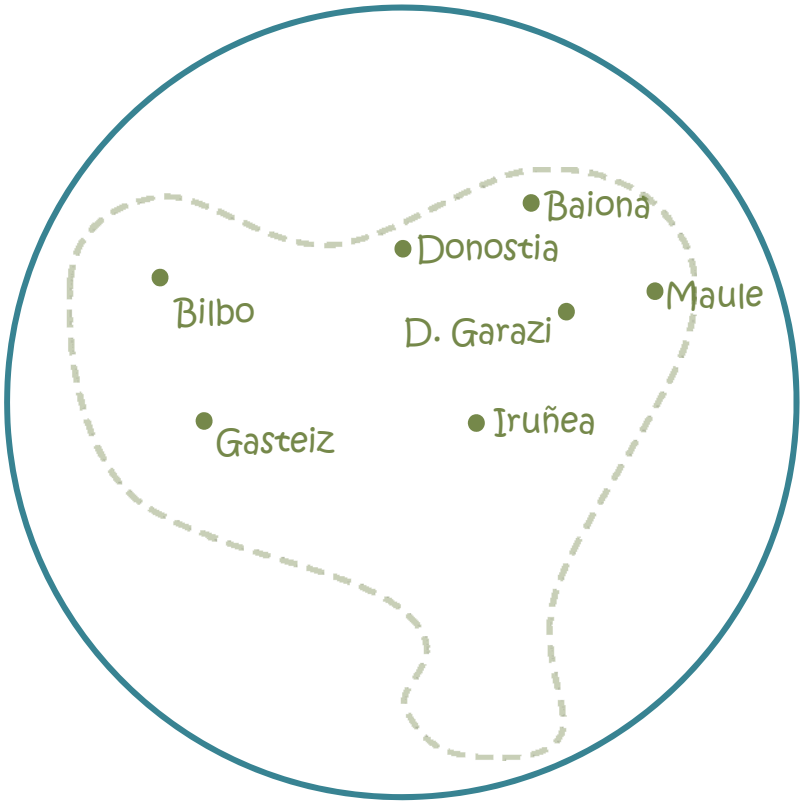
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions

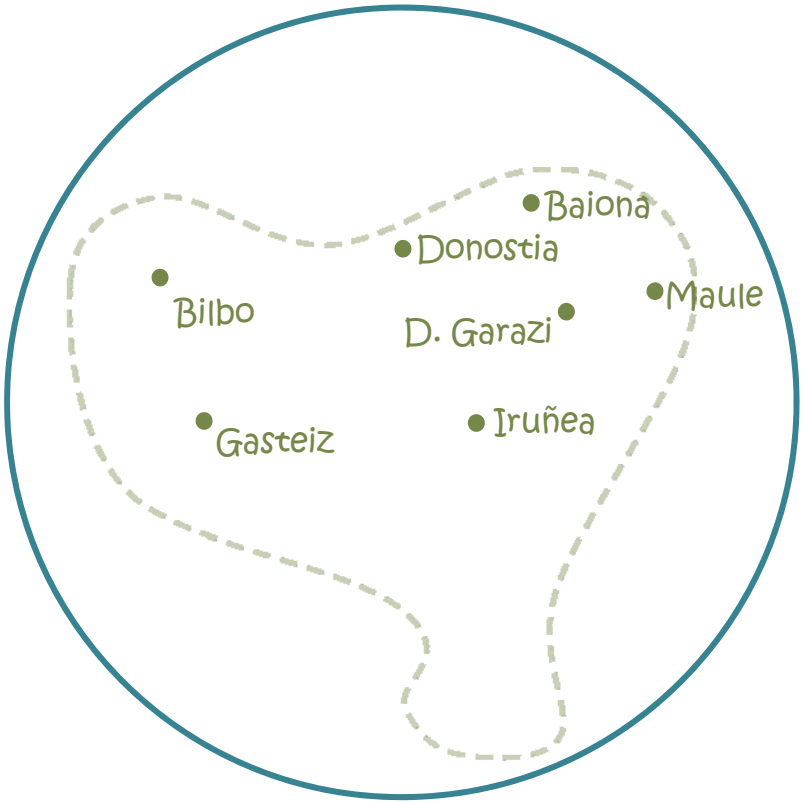
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Introduction to vector space models

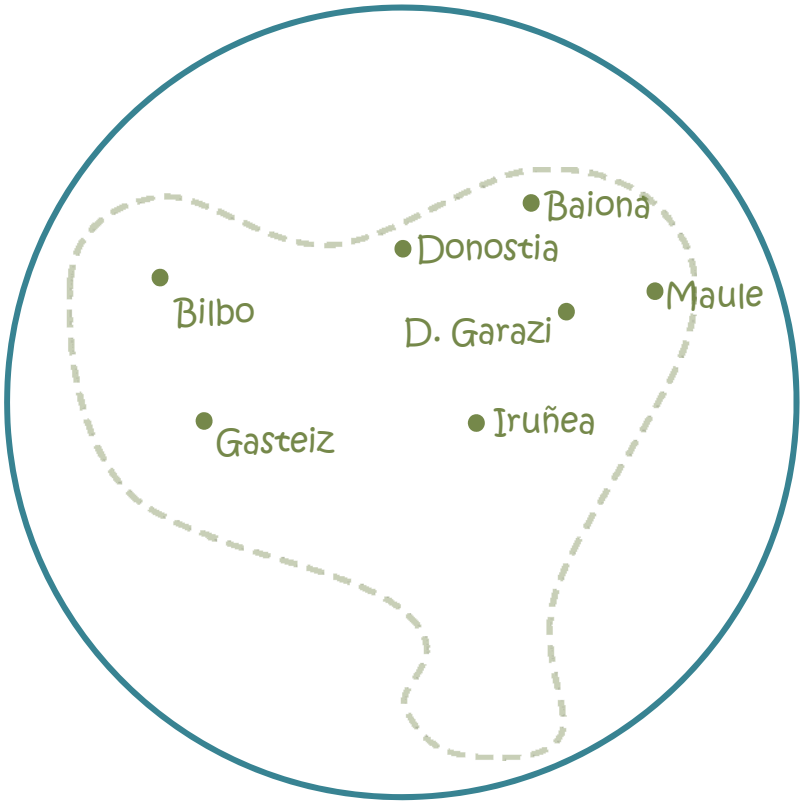


Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

Introduction to vector space models



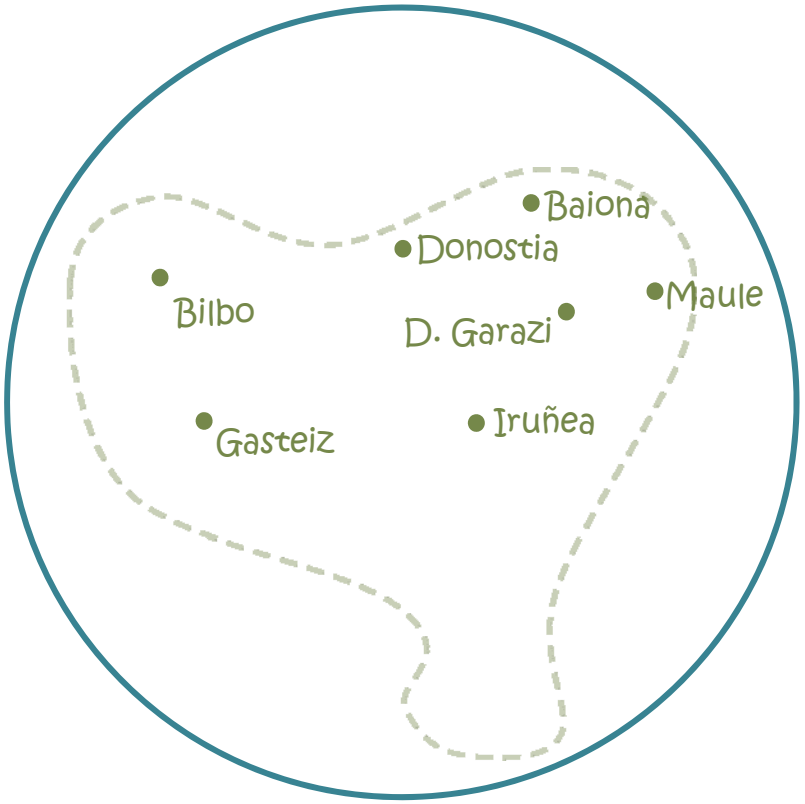
Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words

Introduction to vector space models

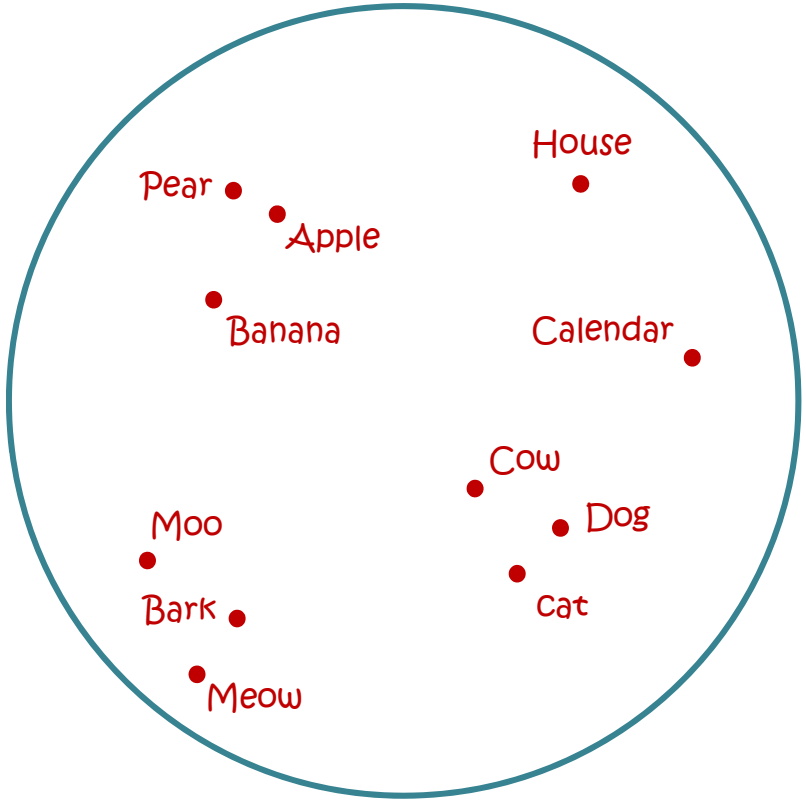


Geographical space

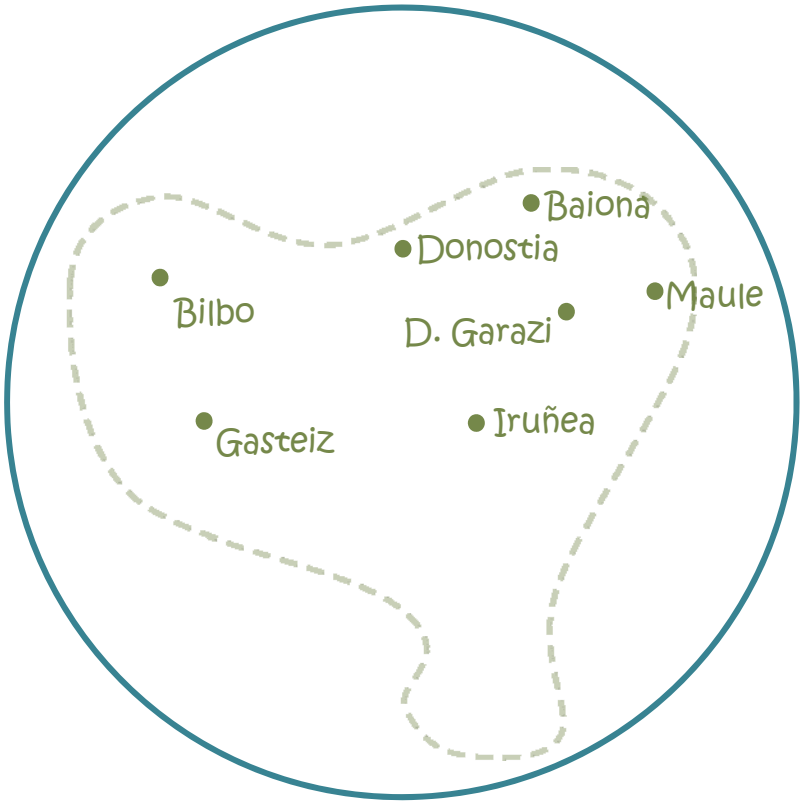
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words



Introduction to vector space models

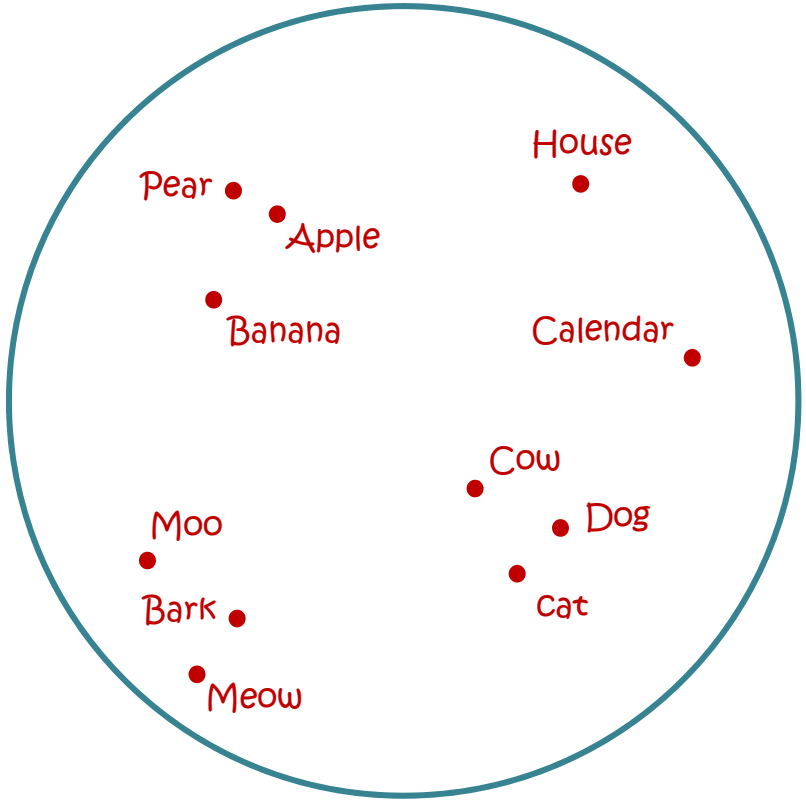


Geographical space

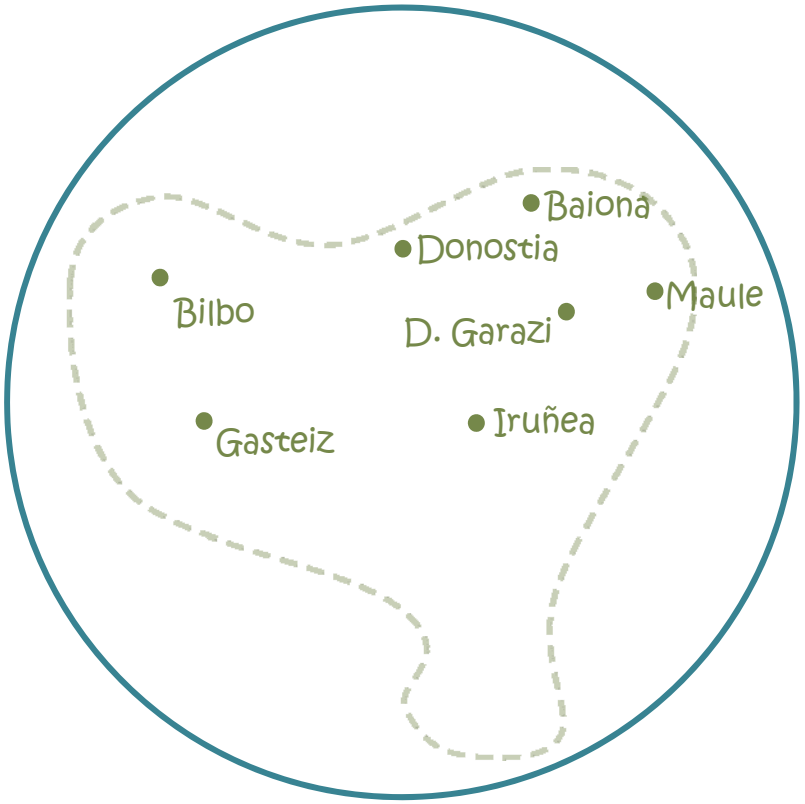
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Introduction to vector space models

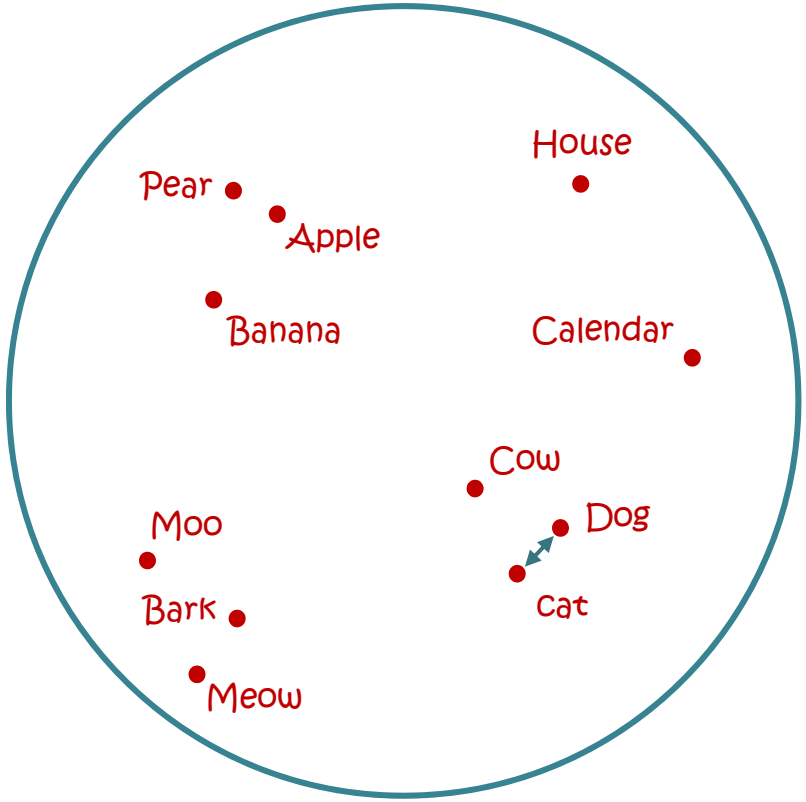


Geographical space

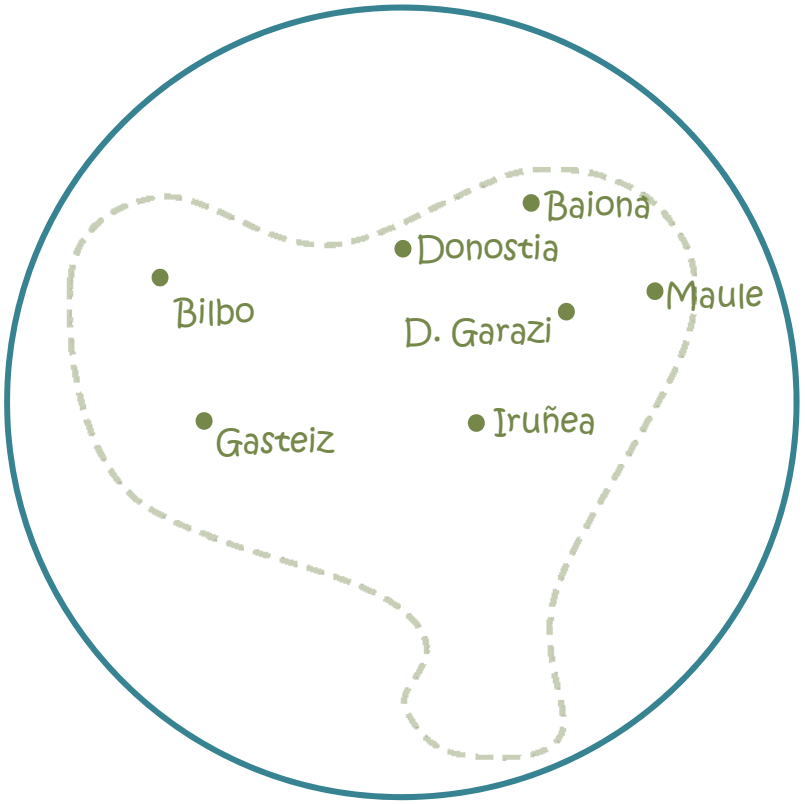
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Introduction to vector space models

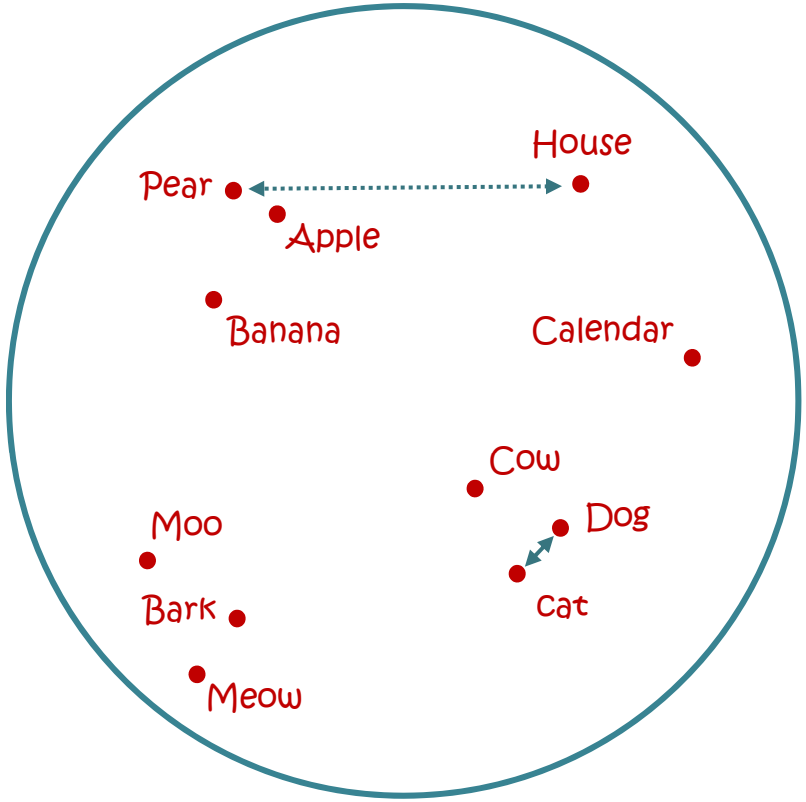


Geographical space

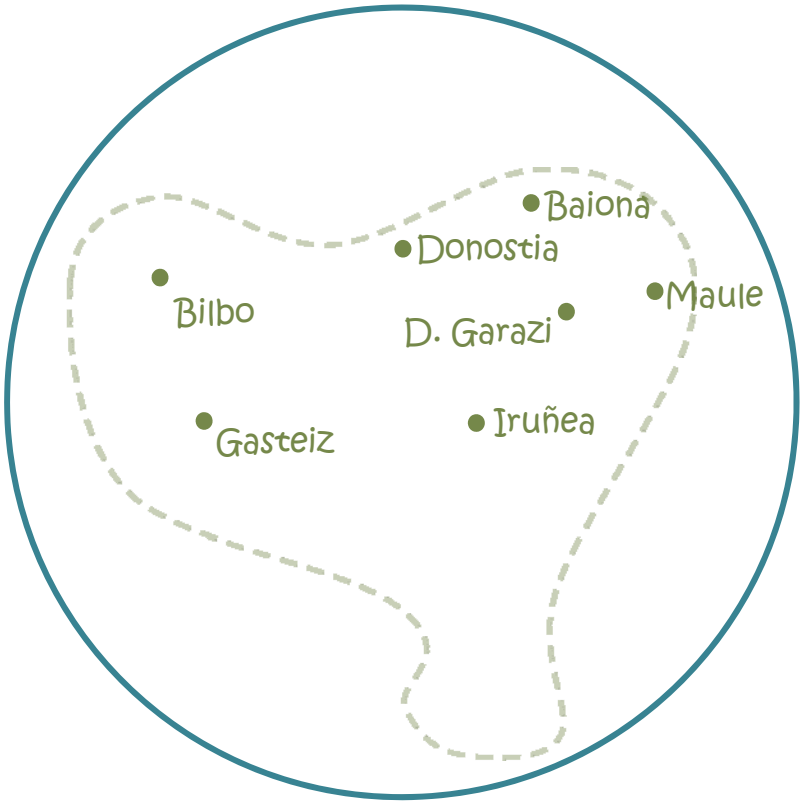
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Introduction to vector space models

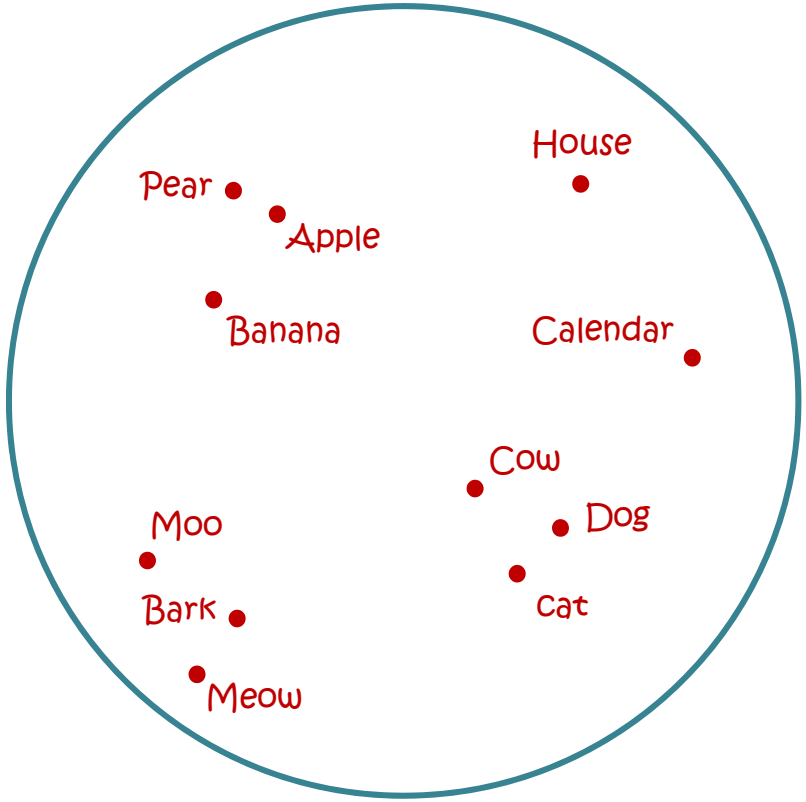


Geographical space

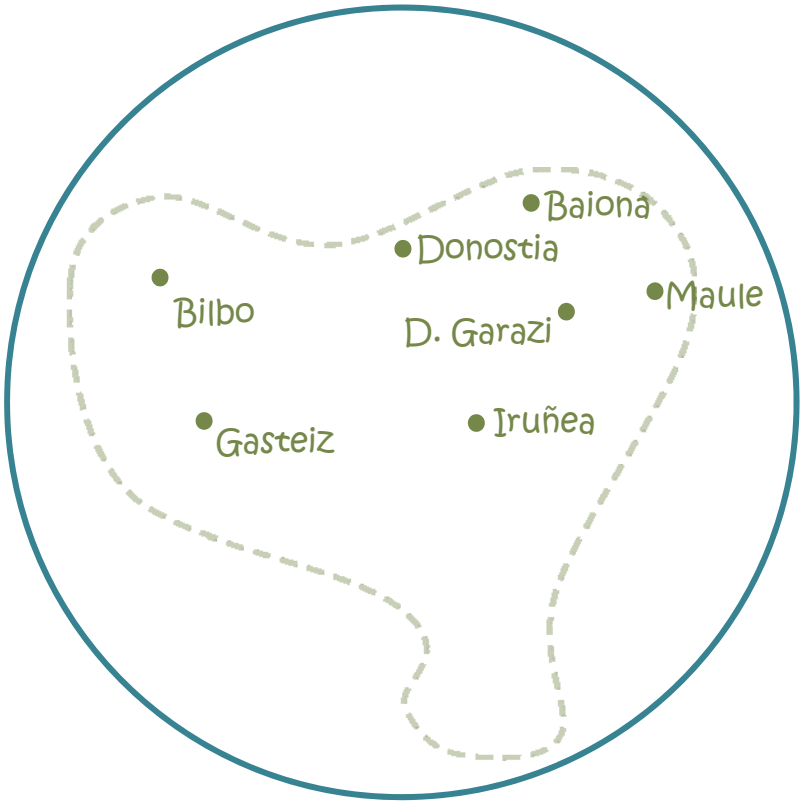
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

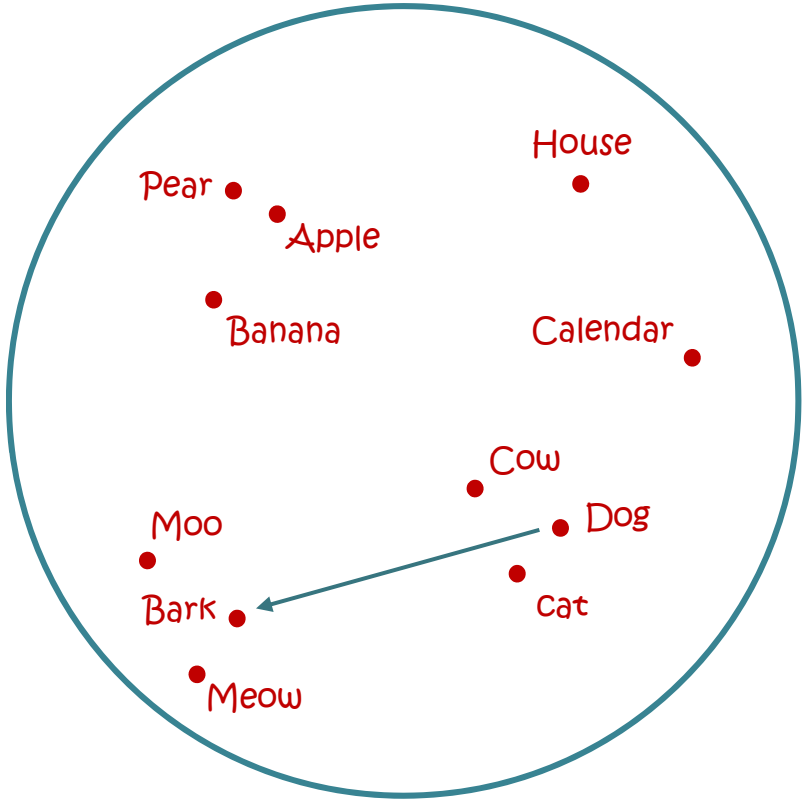


Geographical space

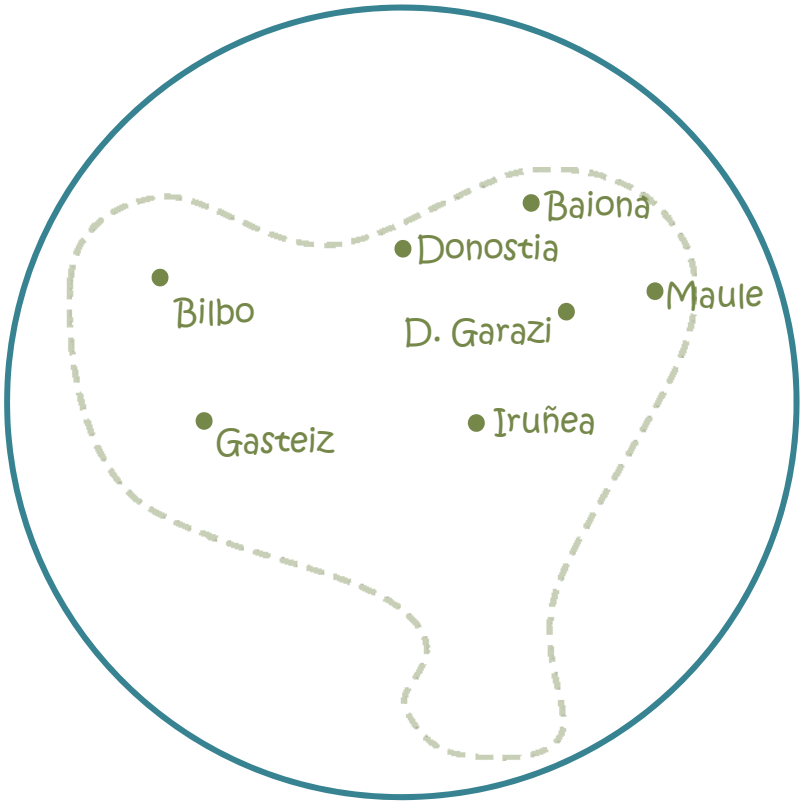
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

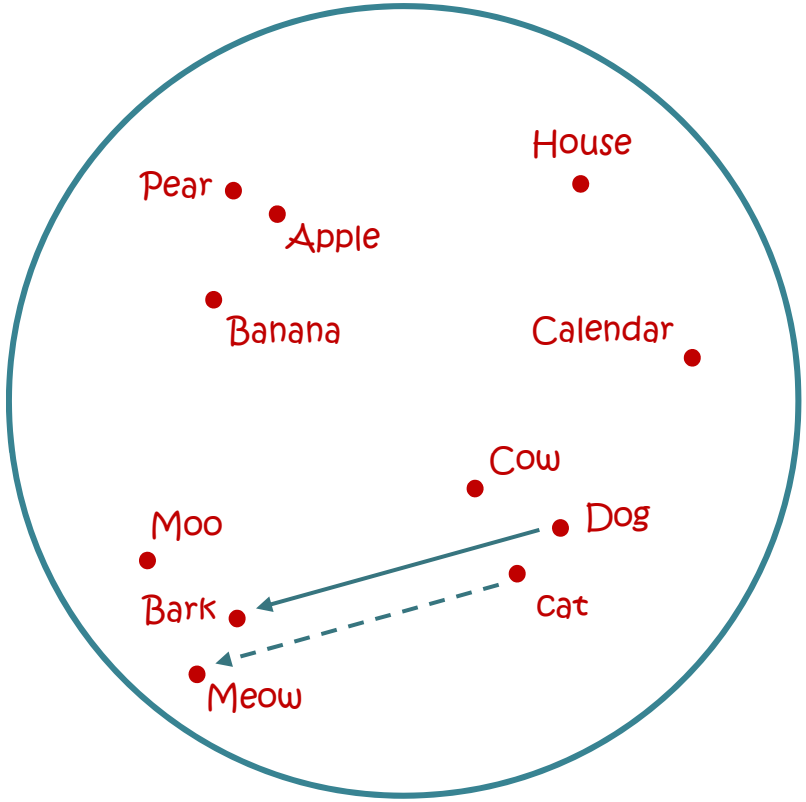


Geographical space

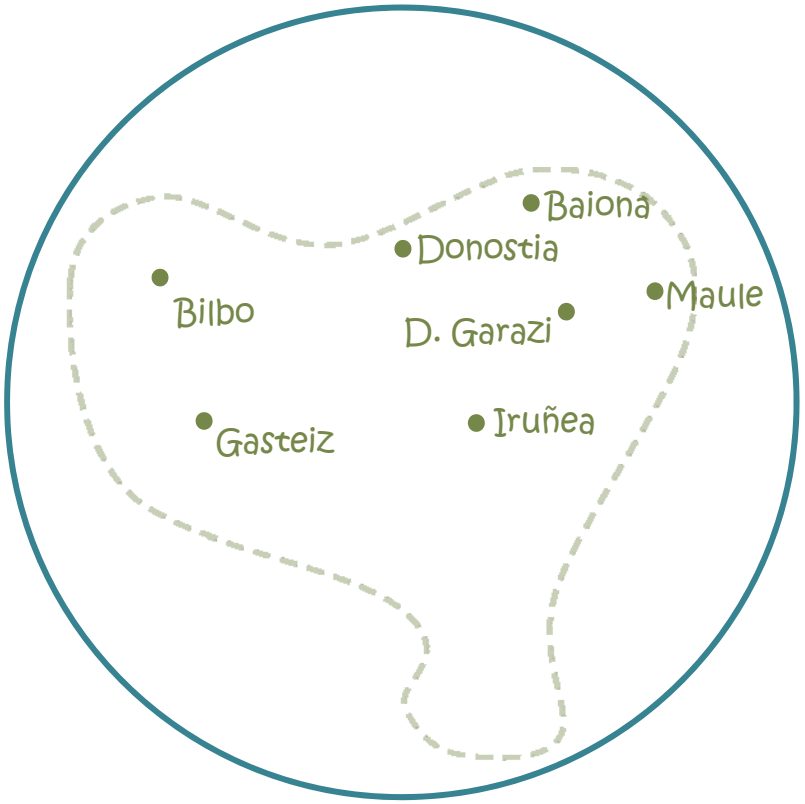
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

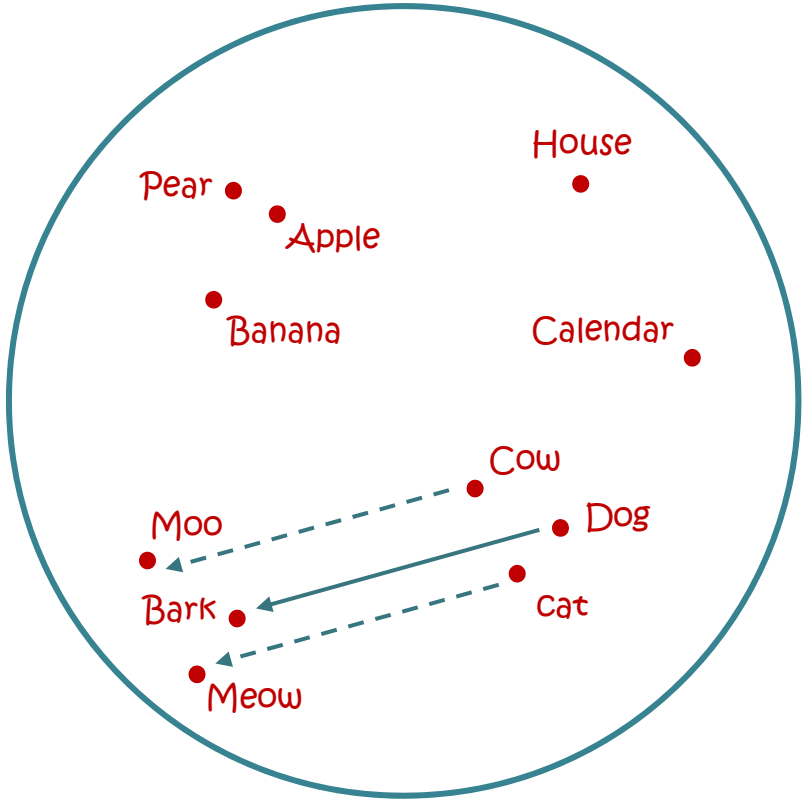


Geographical space

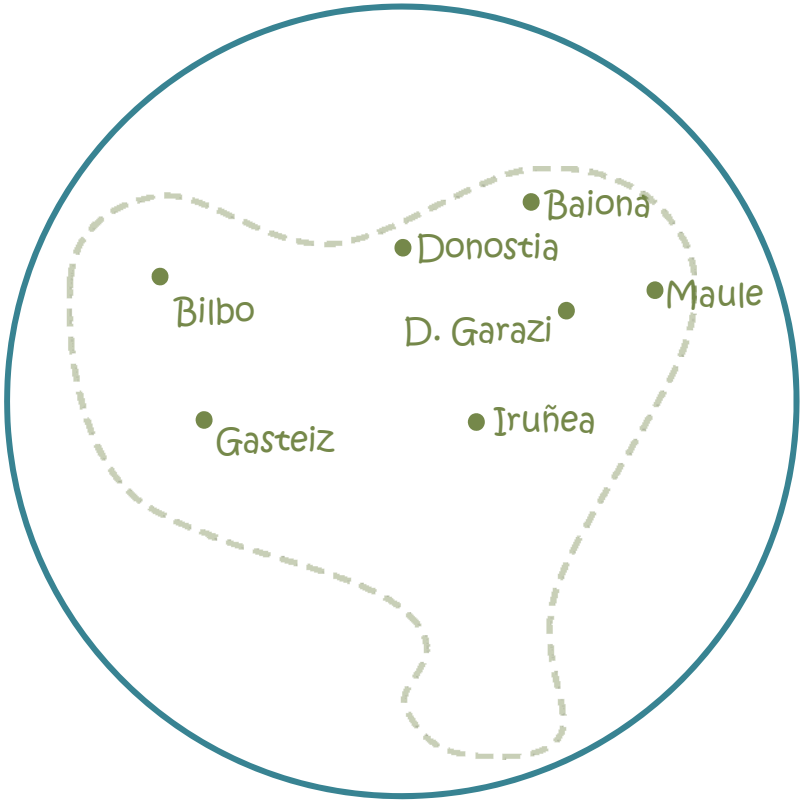
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

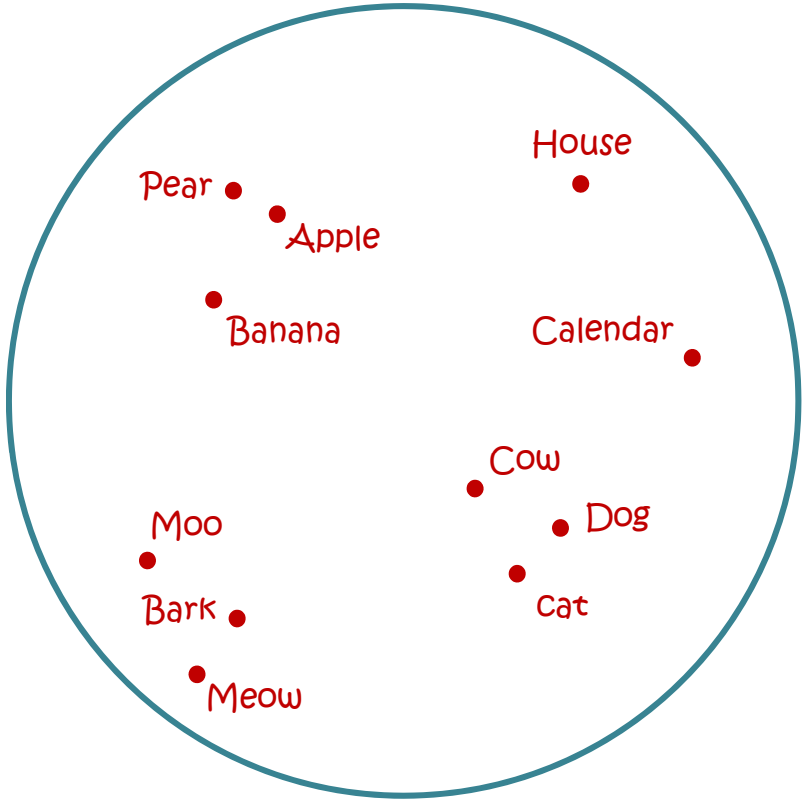


Geographical space

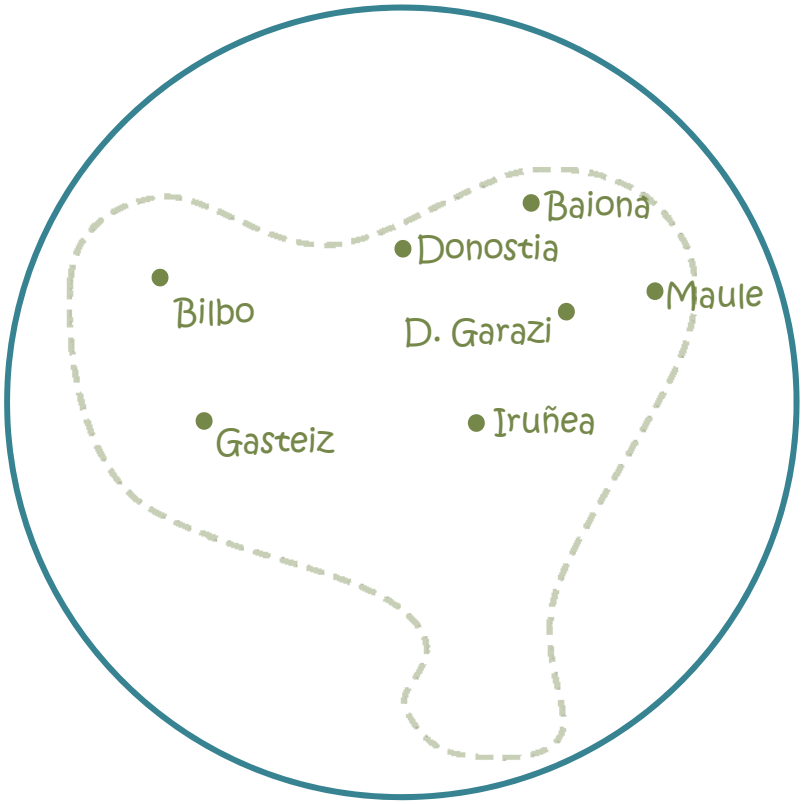
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions



Introduction to vector space models

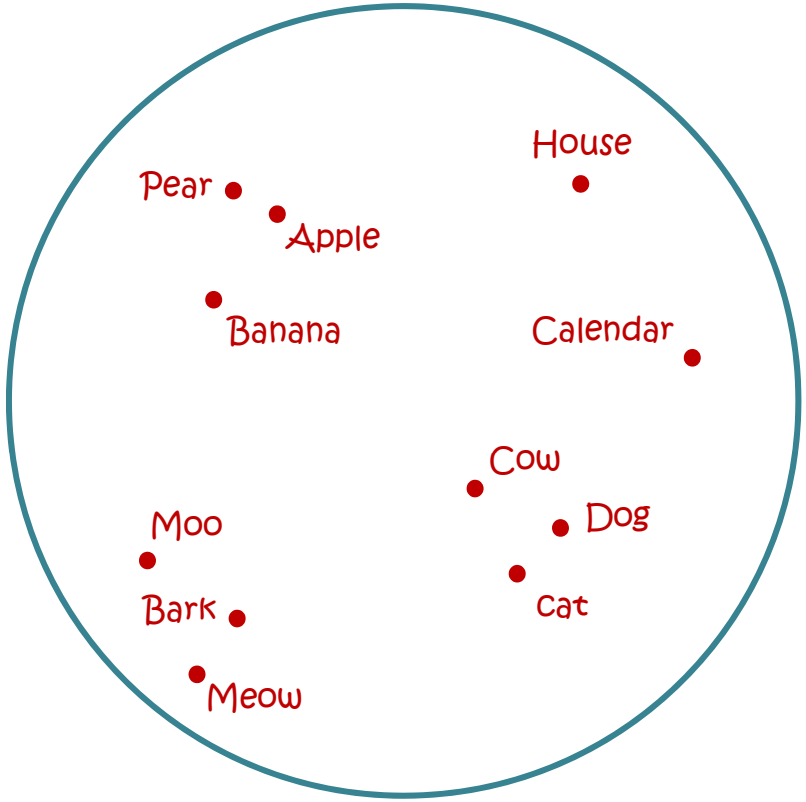


Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

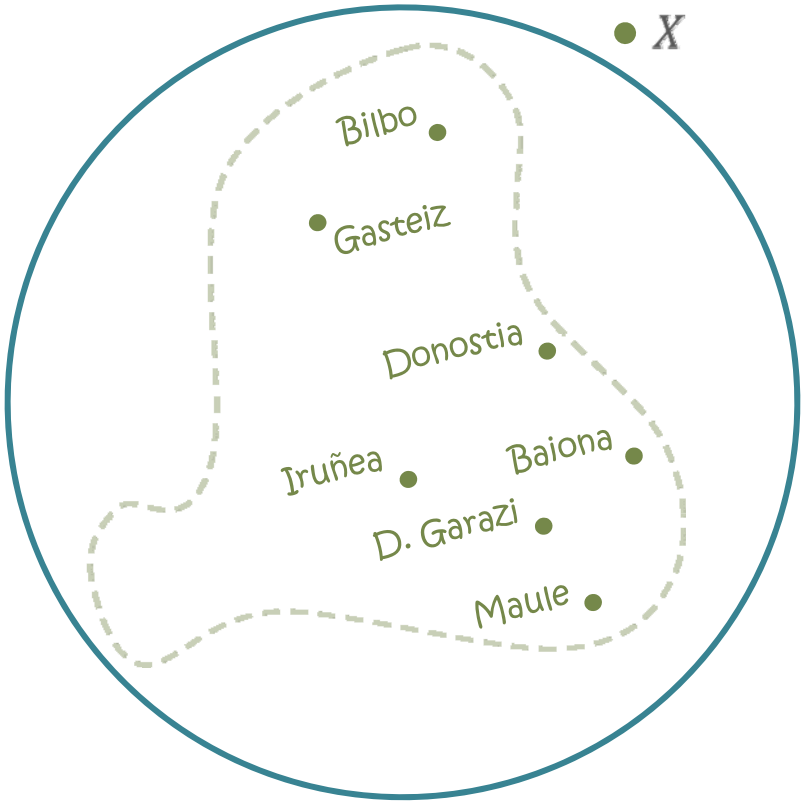
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Neural networks / linear algebra from co-occurrence counts

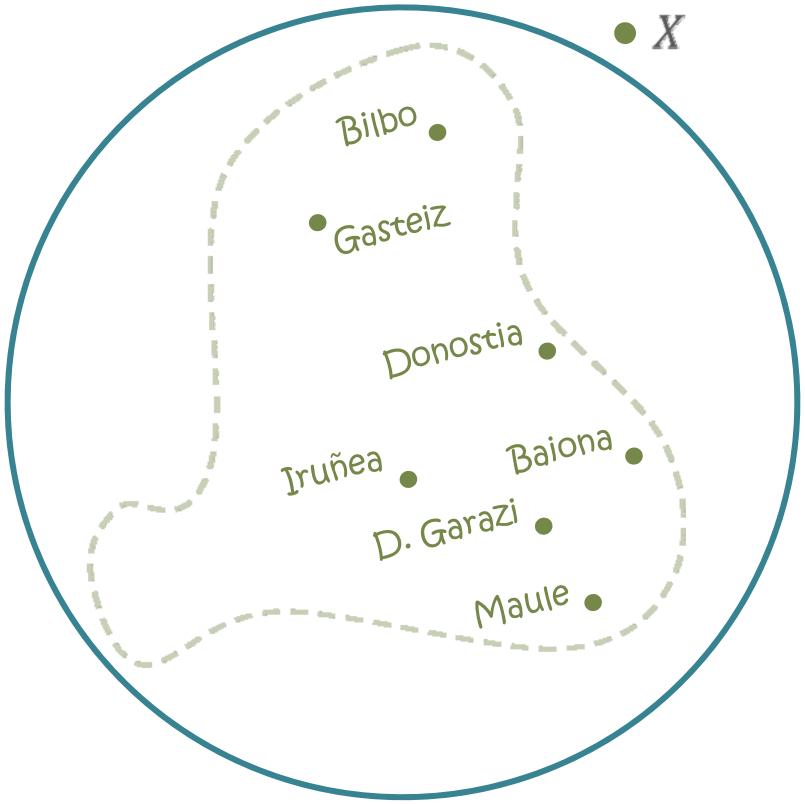


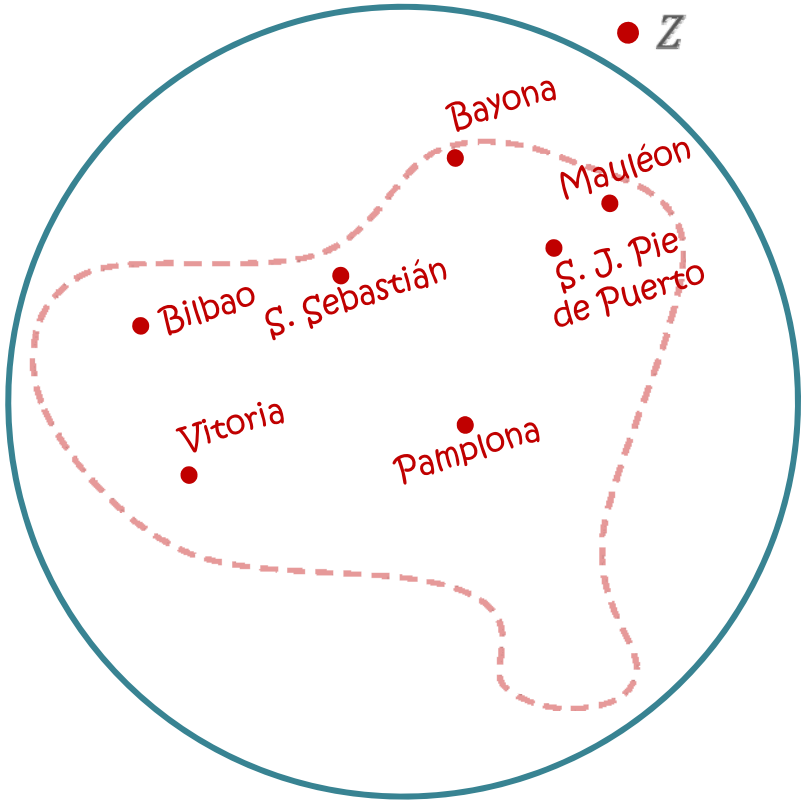
Introduction to embedding mappings

Introduction to embedding mappings

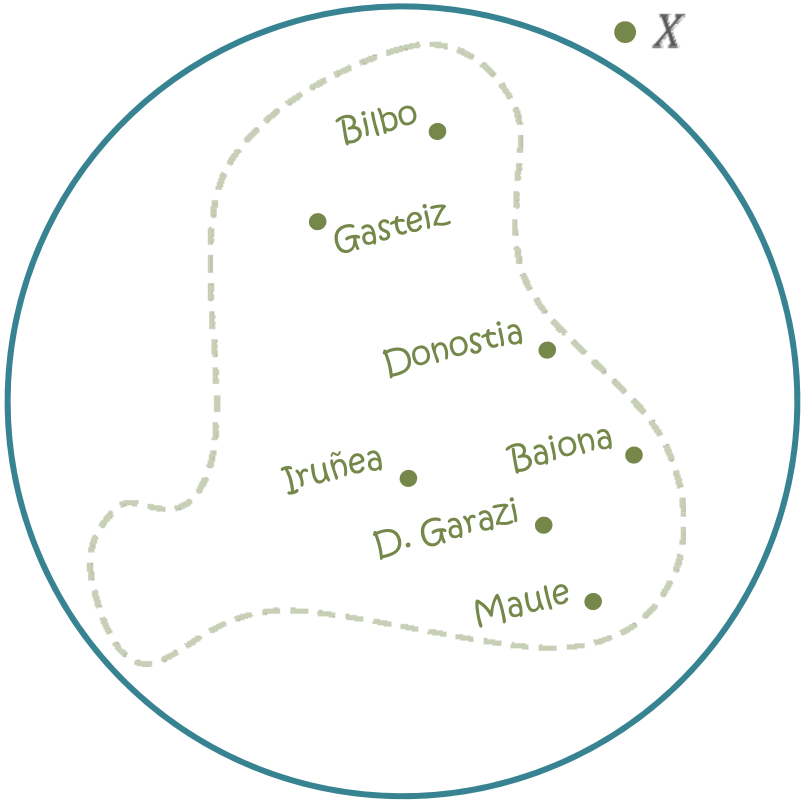


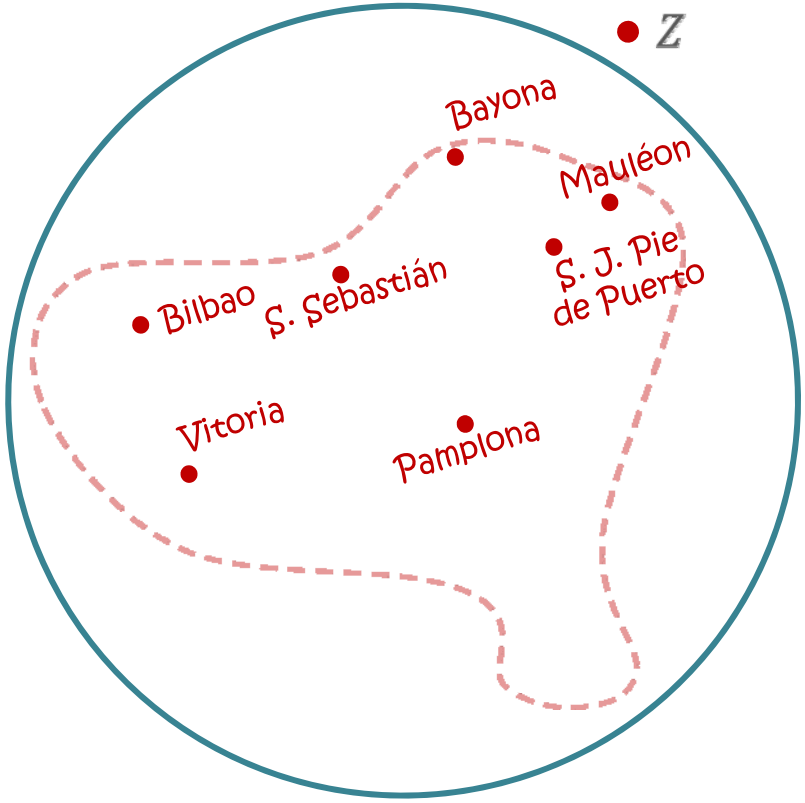
Introduction to embedding mappings





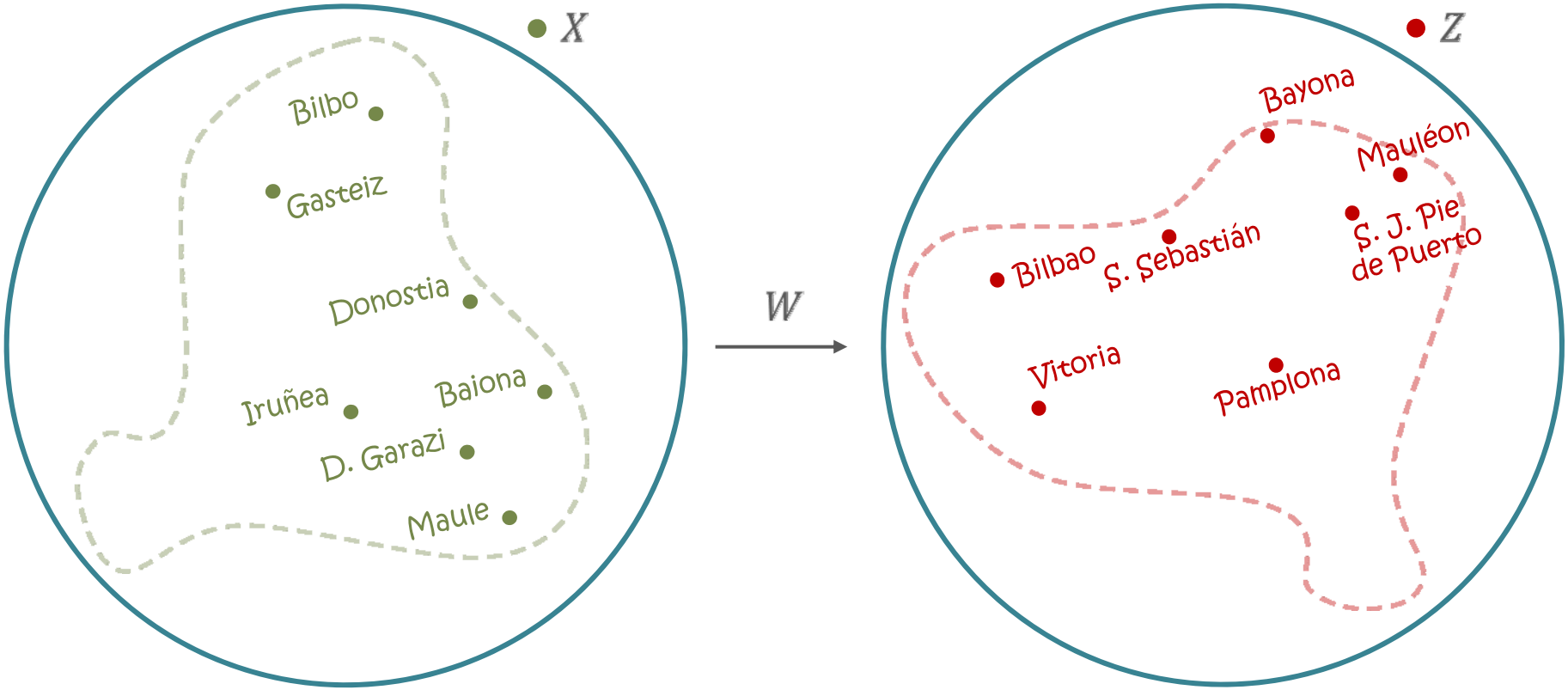
Introduction to embedding mappings





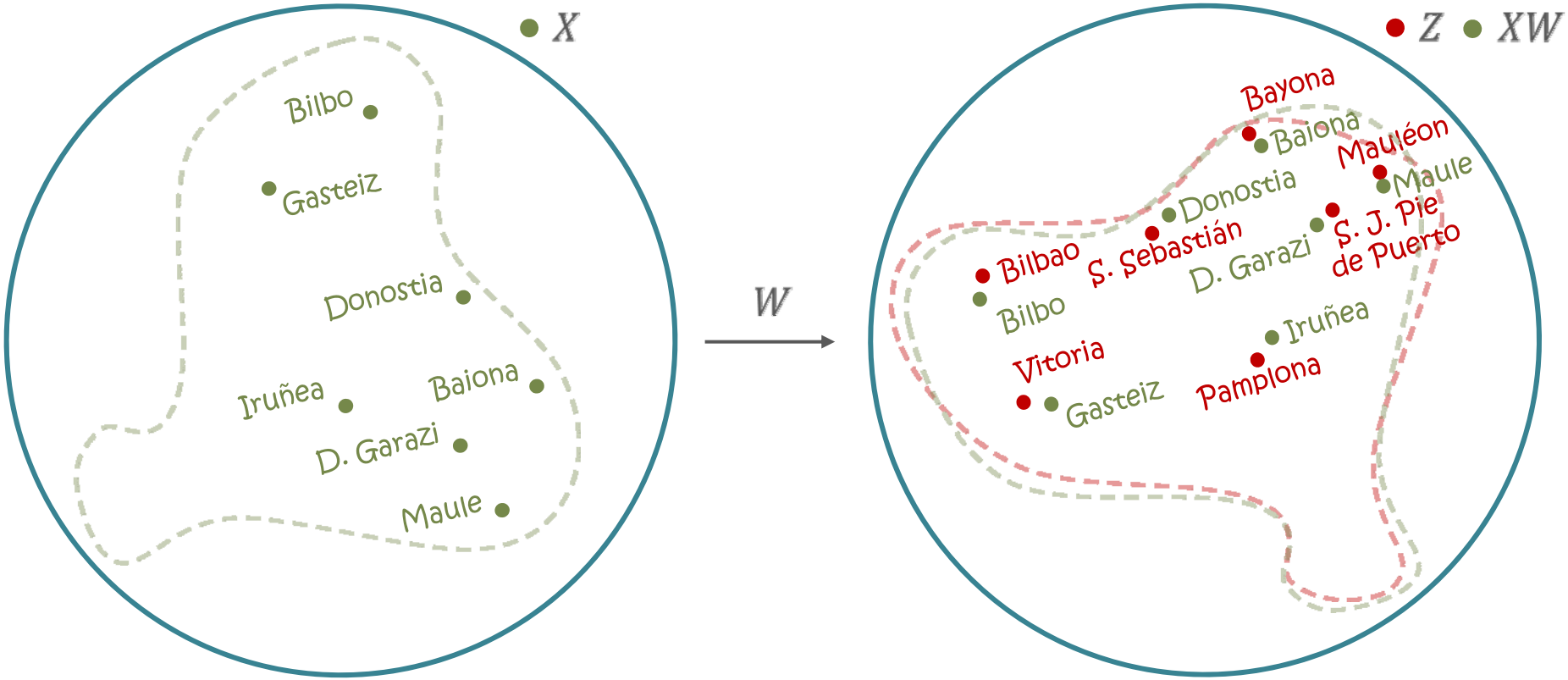
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



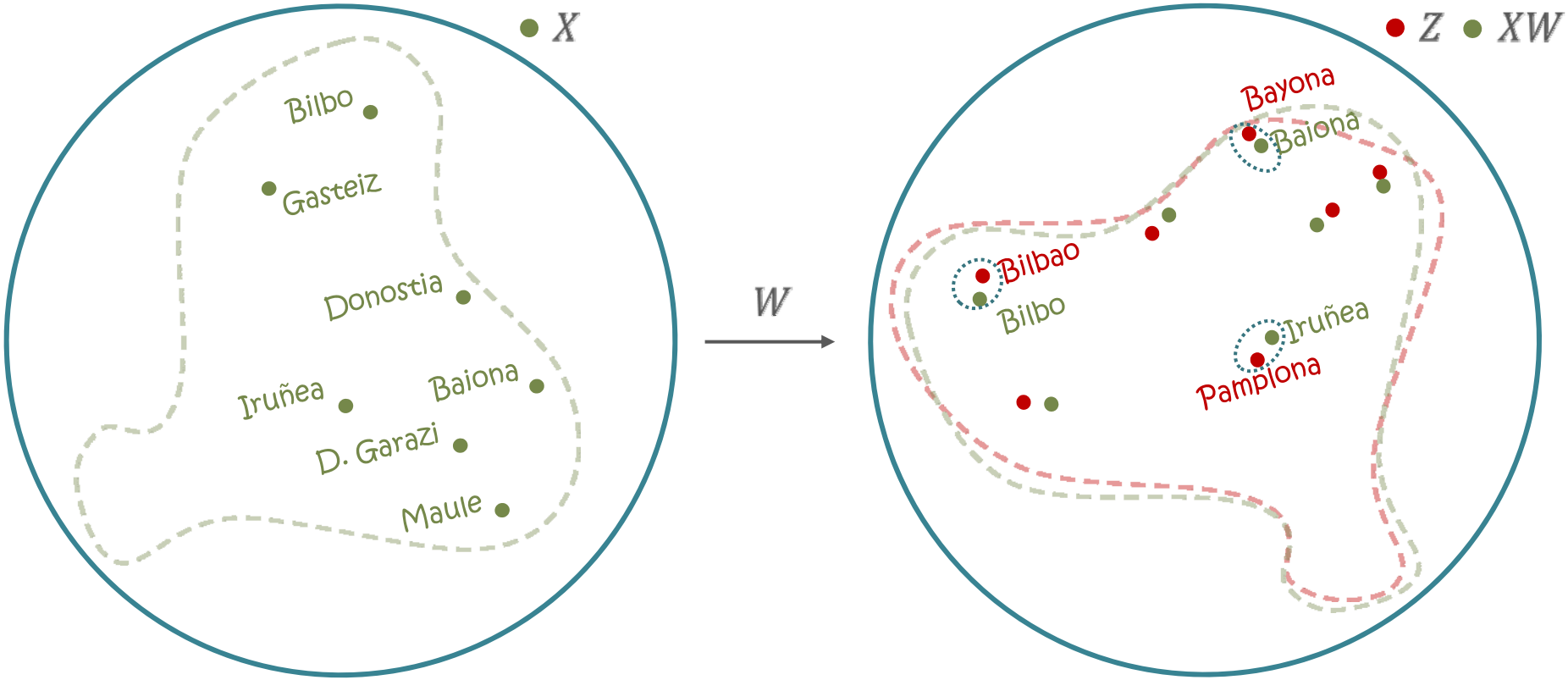
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



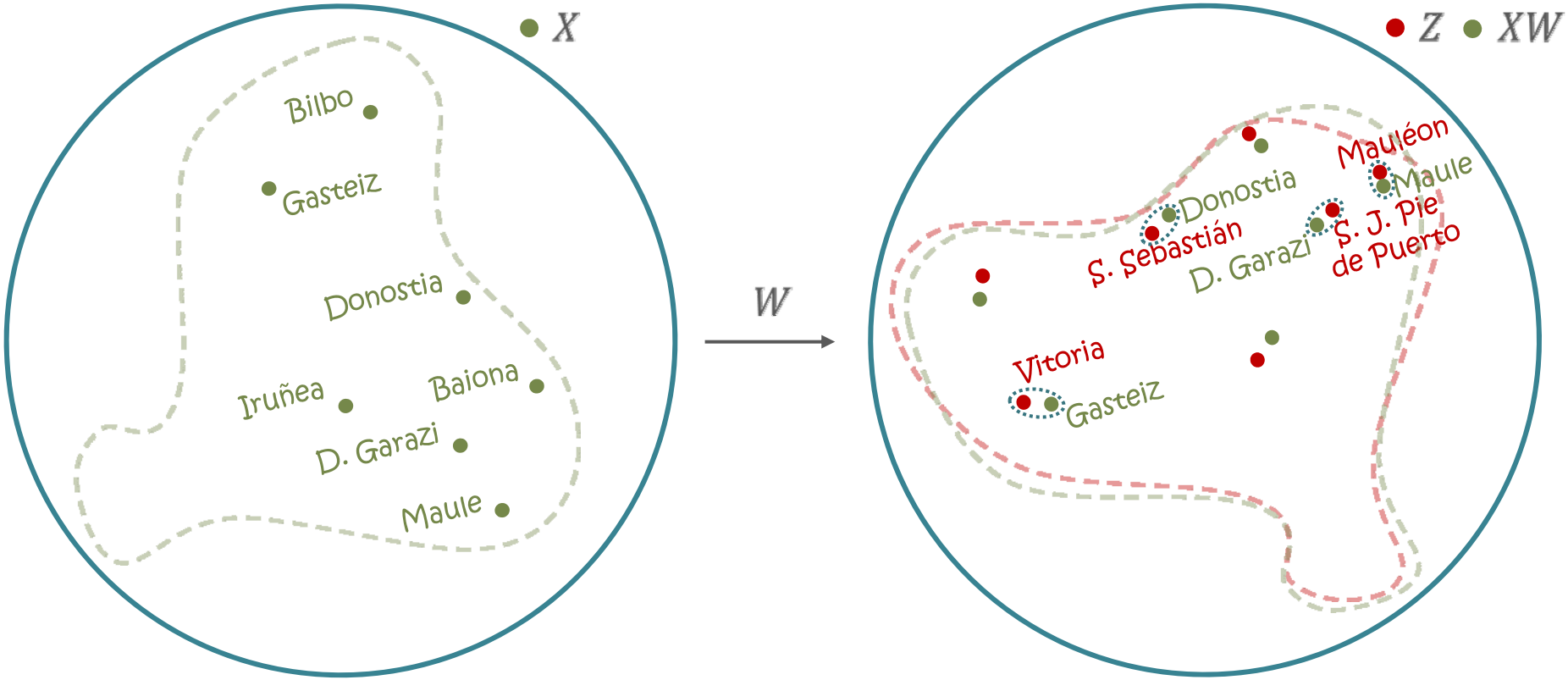
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

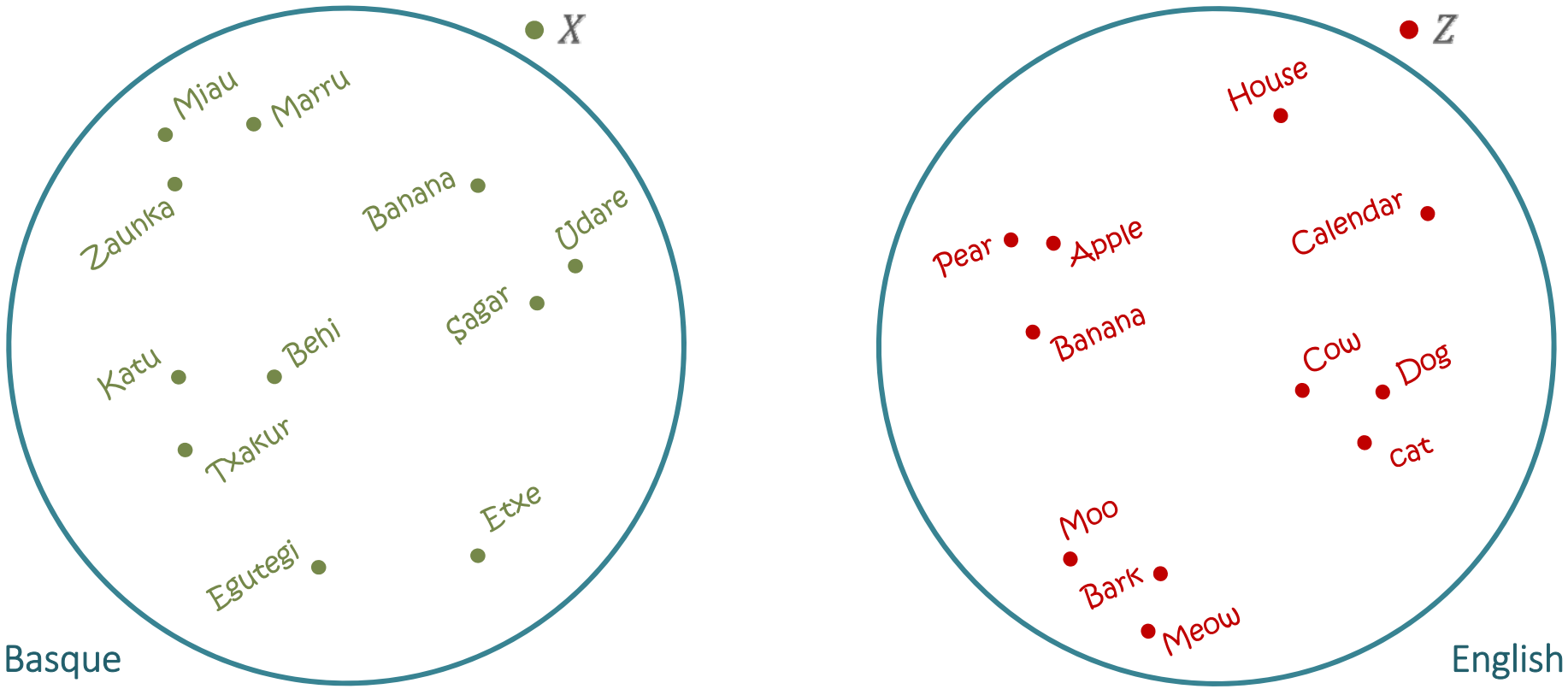
Introduction to embedding mappings



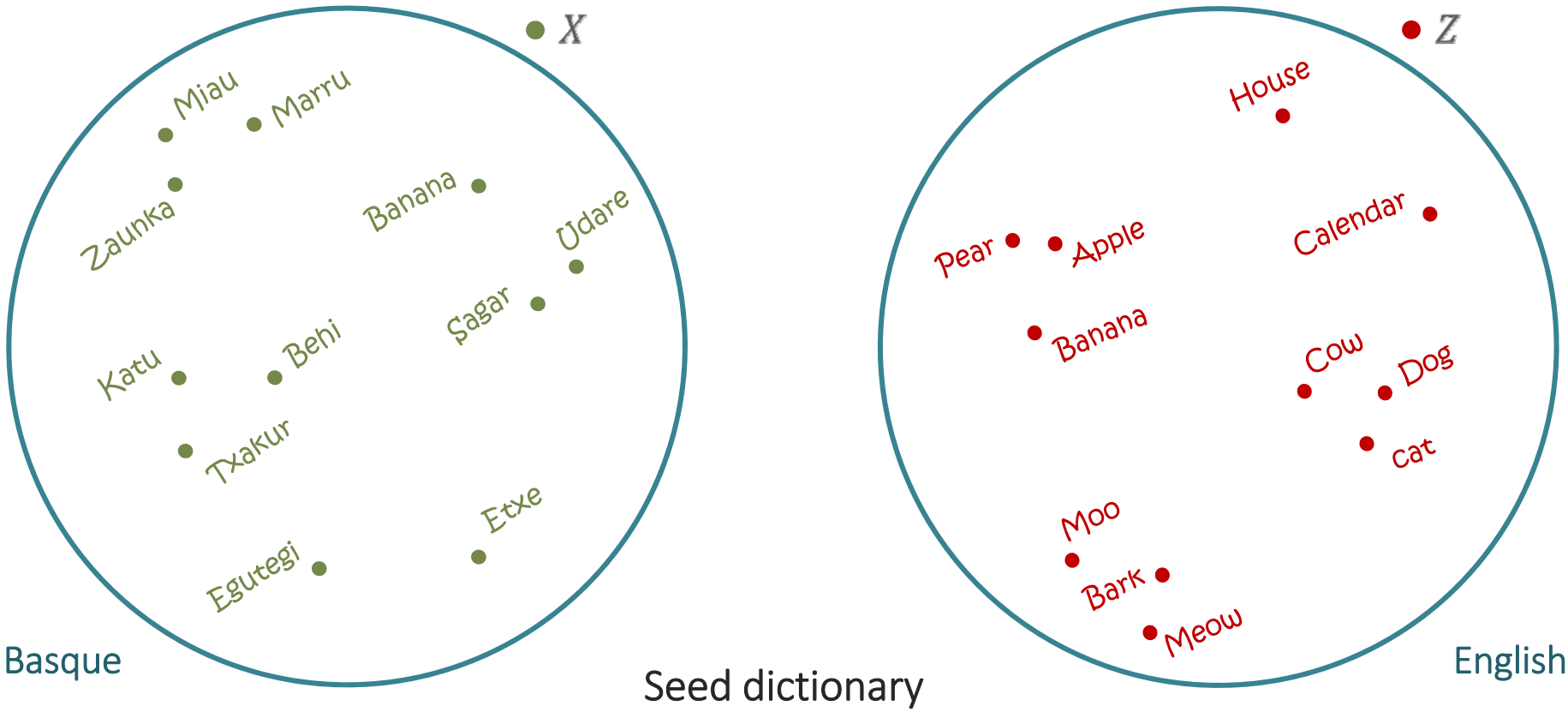
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings

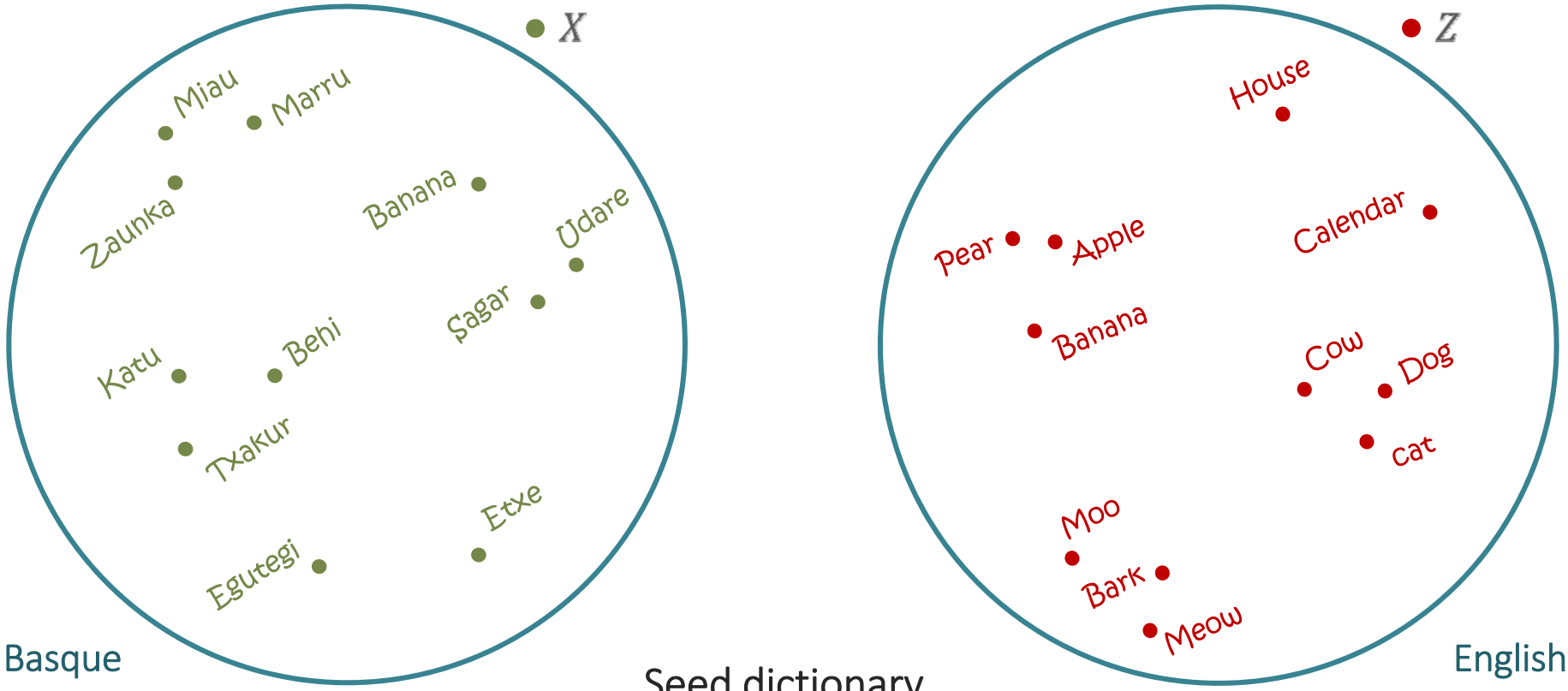
Introduction to embedding mappings



Introduction to embedding mappings



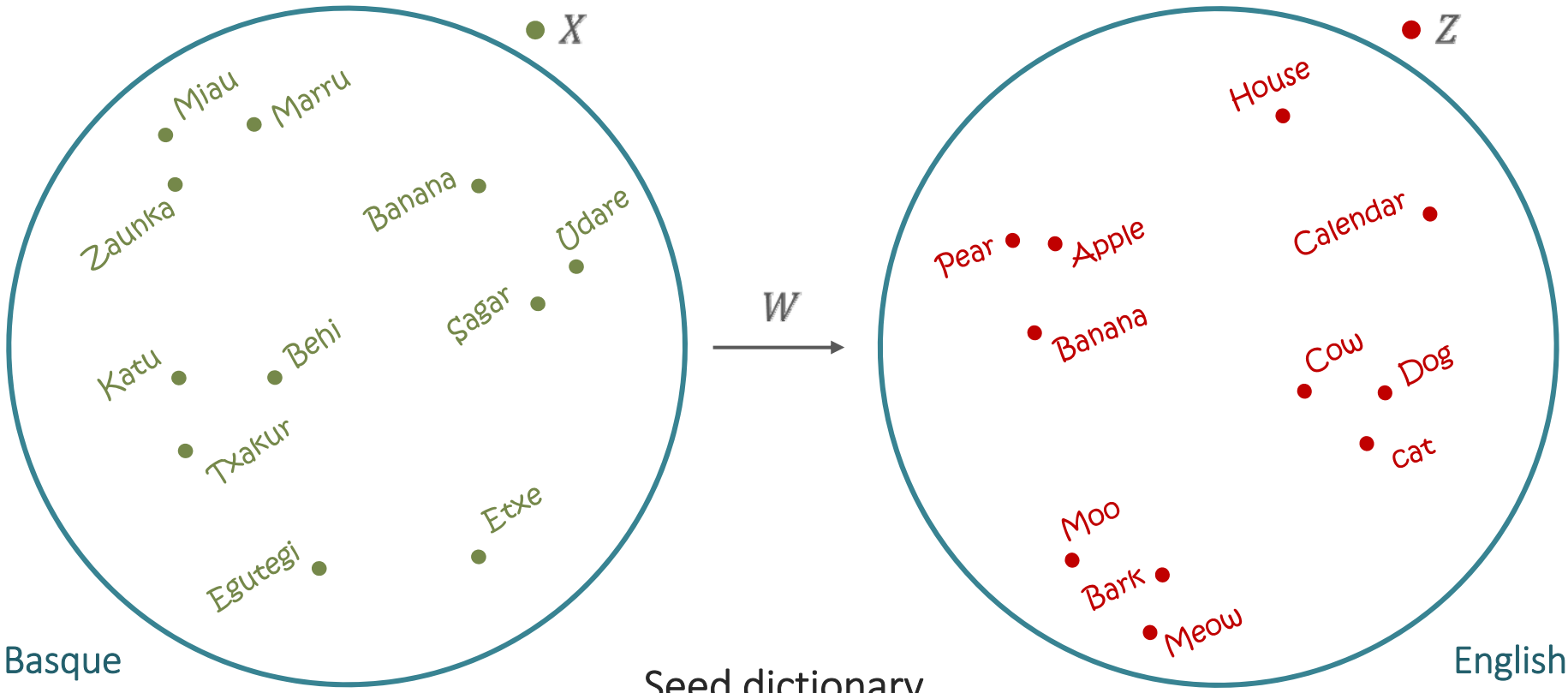
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

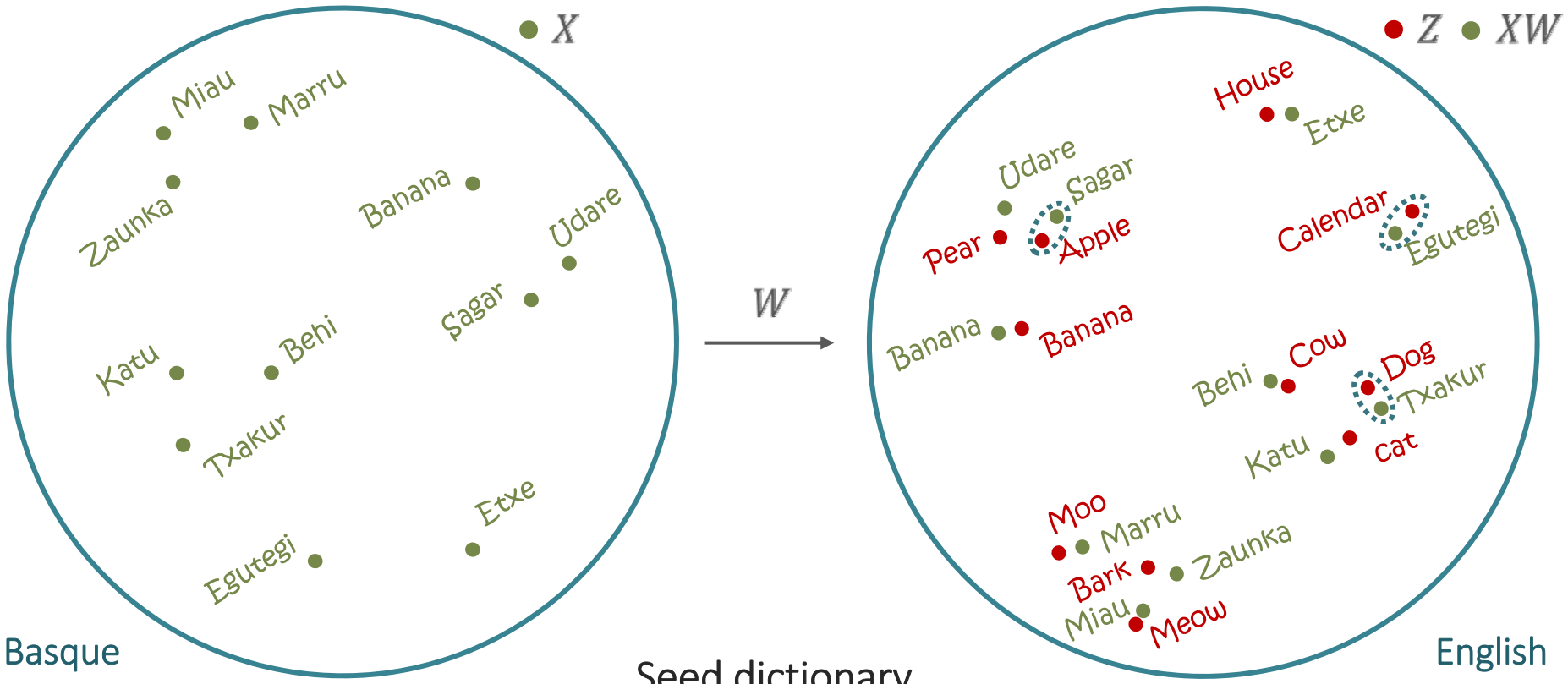
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

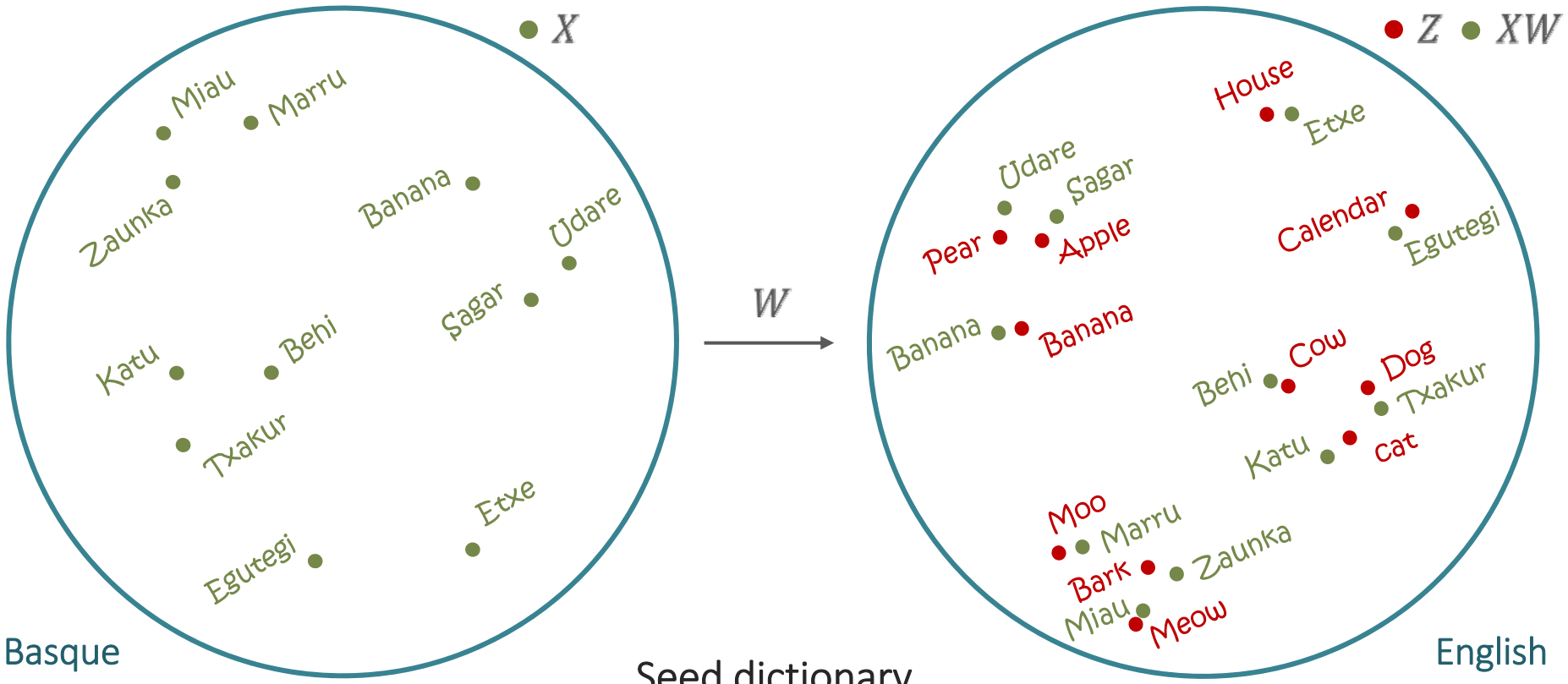
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

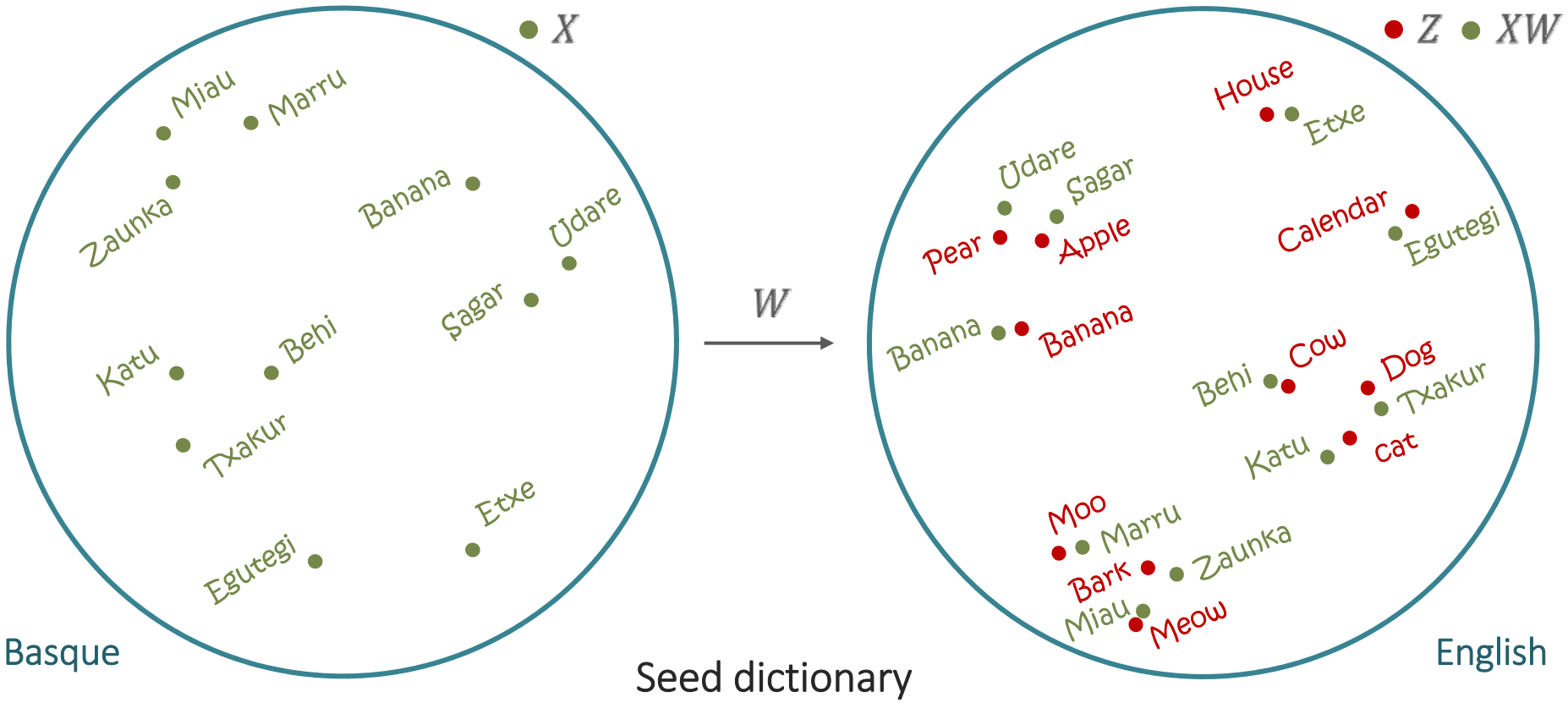
Introduction to embedding mappings



Seed dictionary

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} \quad \Bigg| \quad \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

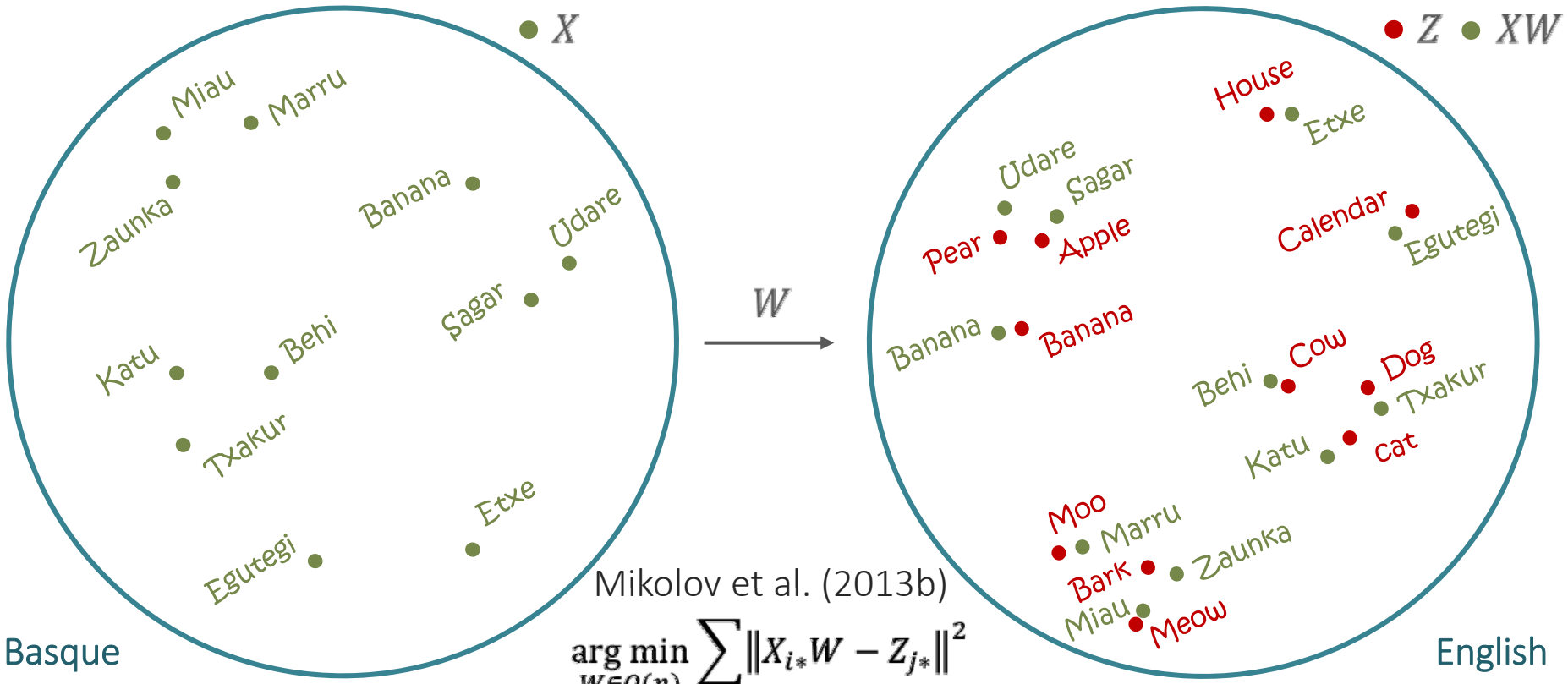
Introduction to embedding mappings



Seed dictionary

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

Introduction to embedding mappings

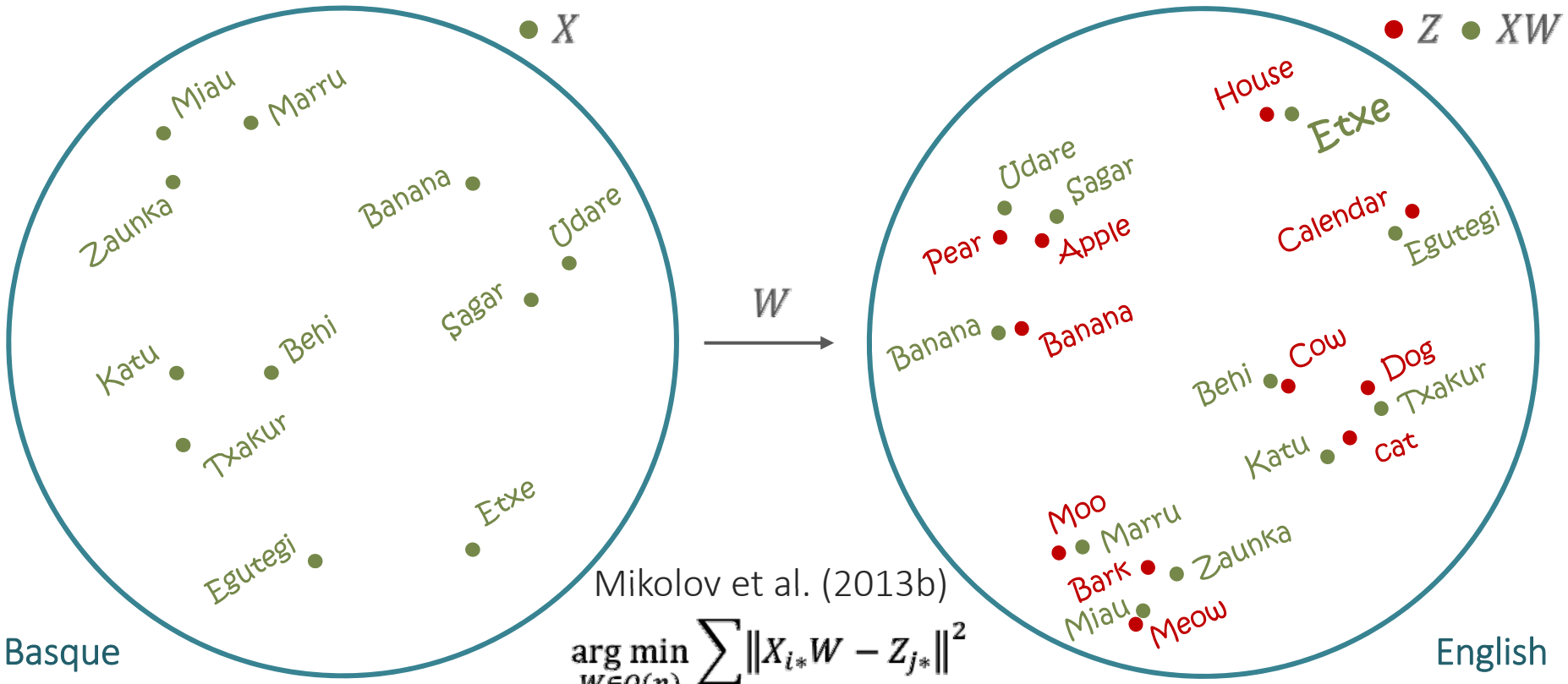


Mikolov et al. (2013b)

$$\underset{W \in O(n)}{\arg\min} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

Introduction to embedding mappings

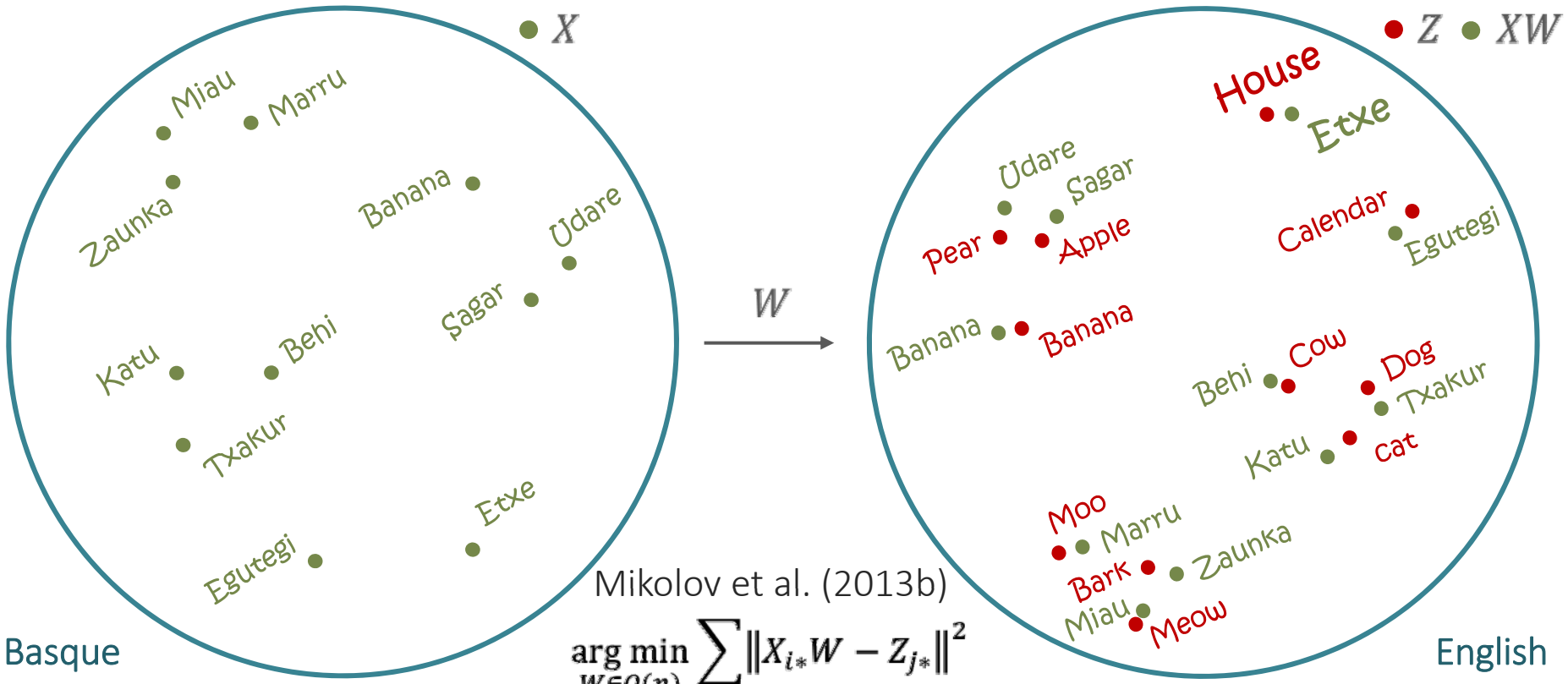


Mikolov et al. (2013b)

$$\underset{W \in O(n)}{\arg\min} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

Introduction to embedding mappings

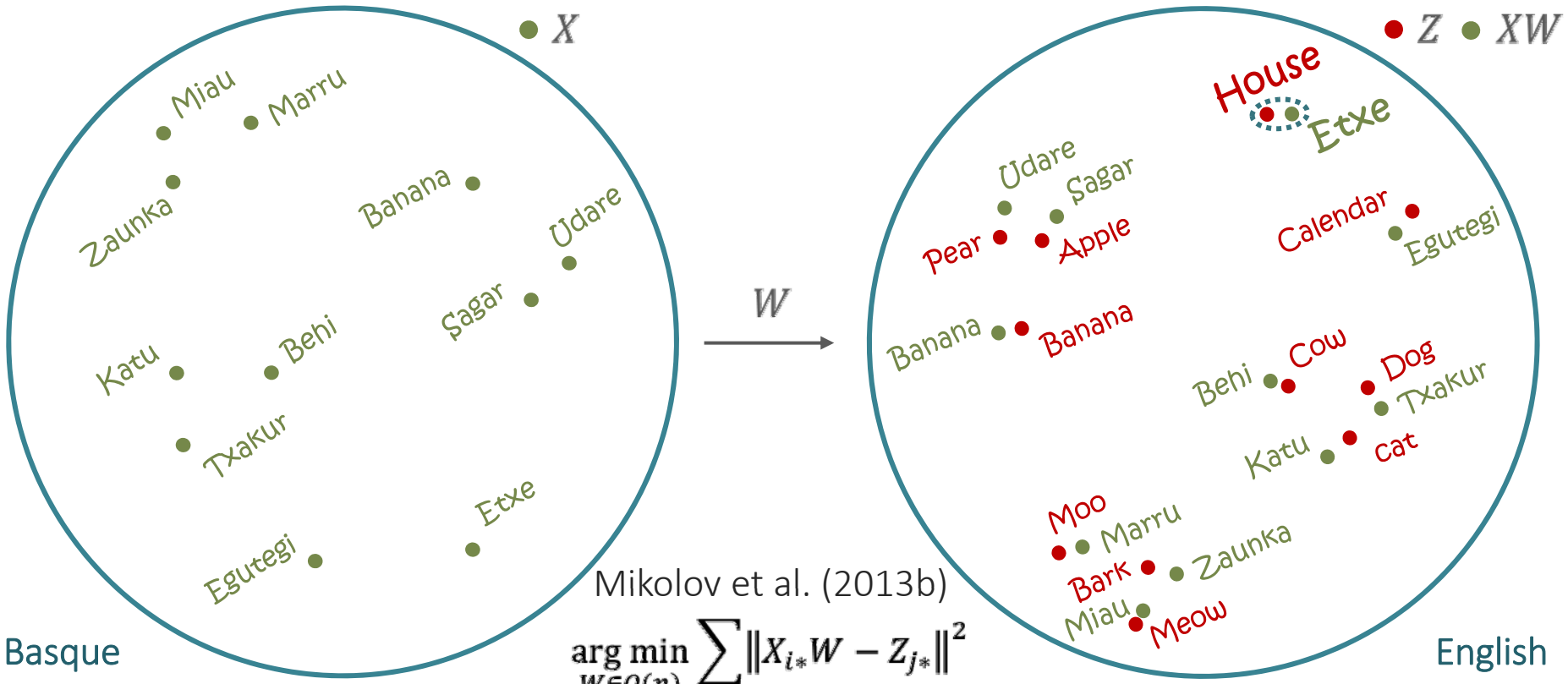


Mikolov et al. (2013b)

$$\underset{W \in O(n)}{\arg\min} \sum_i \left\| X_{i*} W - Z_{j*} \right\|^2$$

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

Introduction to embedding mappings



Mikolov et al. (2013b)

$$\underset{W \in O(n)}{\arg\min} \sum_i \left\| X_{i*} W - Z_{j*} \right\|^2$$

$$\begin{array}{r} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{array} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{array}{l} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{array}$$

State-of-the-art in **supervised** mappings

Artetxe et al. AAAI 2018

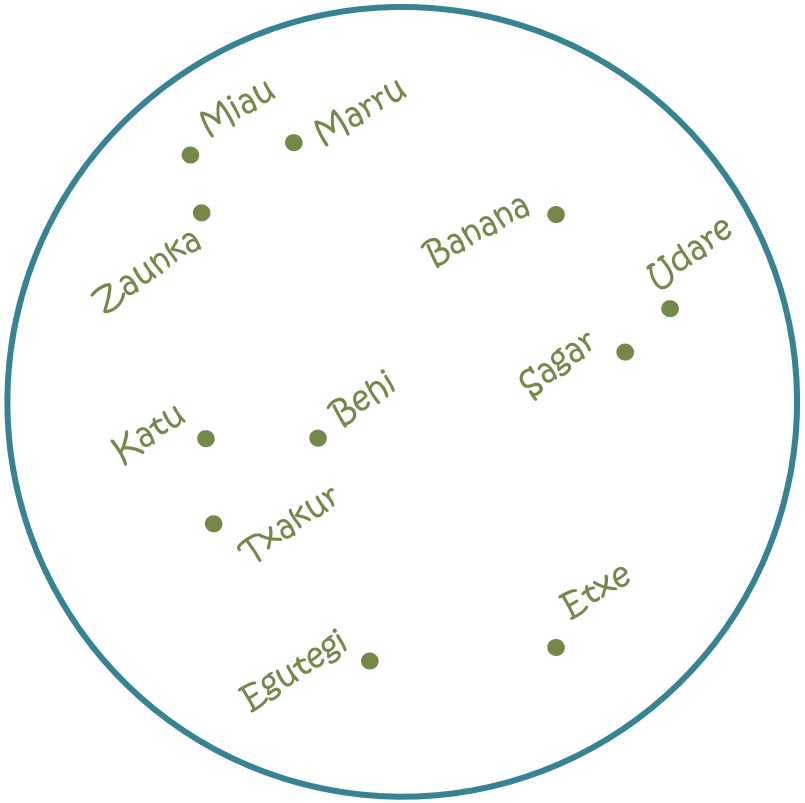
- Use 5000 sized seed bilingual dictionary
- Framework subsuming previous work,
  a **sequence of (optional) linear mappings**:

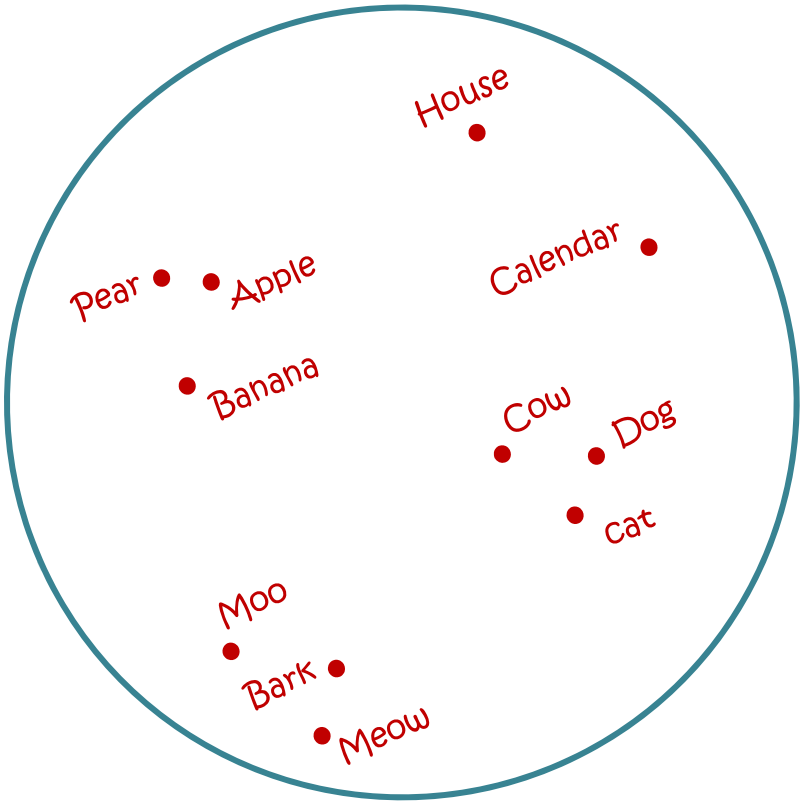
  (opt.) Pre-process: Normalize length, mean centering

  1. *(opt.) Whitening*
  2. ***Orthogonal mapping**, solved with SVD (Procrustes)*
  3. *(opt.) Re-weighting*
  4. *(opt.) De-whitening*
- Optional steps, properly combined,
  bring up to 5 points improvement

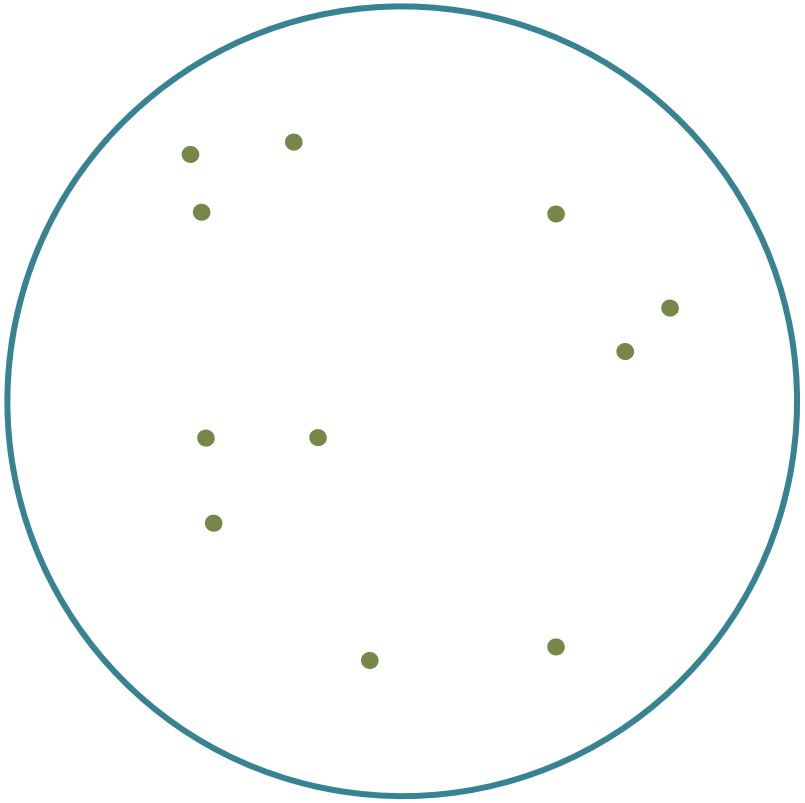
Why does it work?

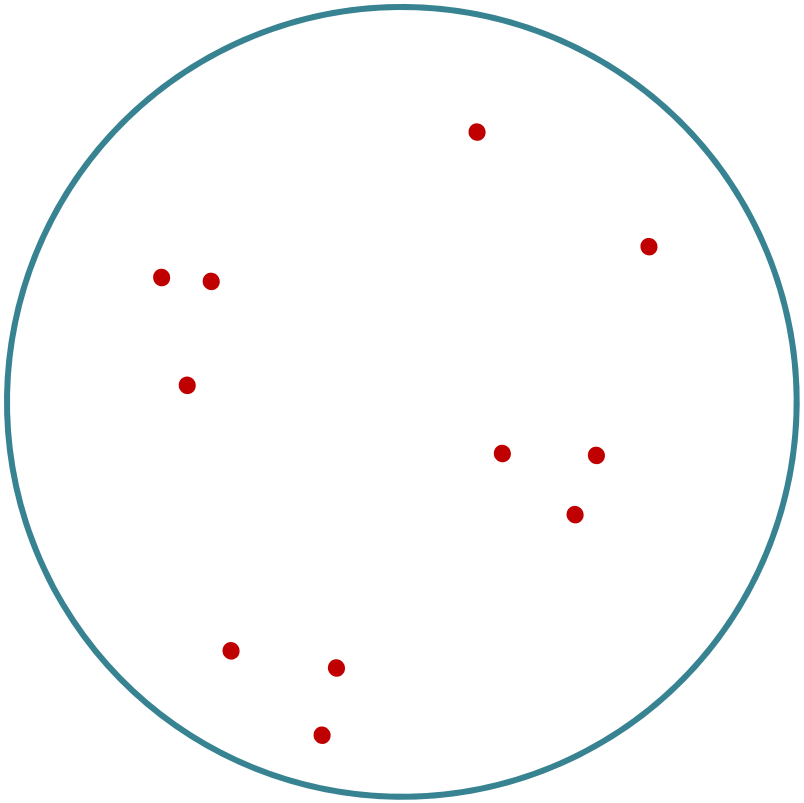
Why does it work?



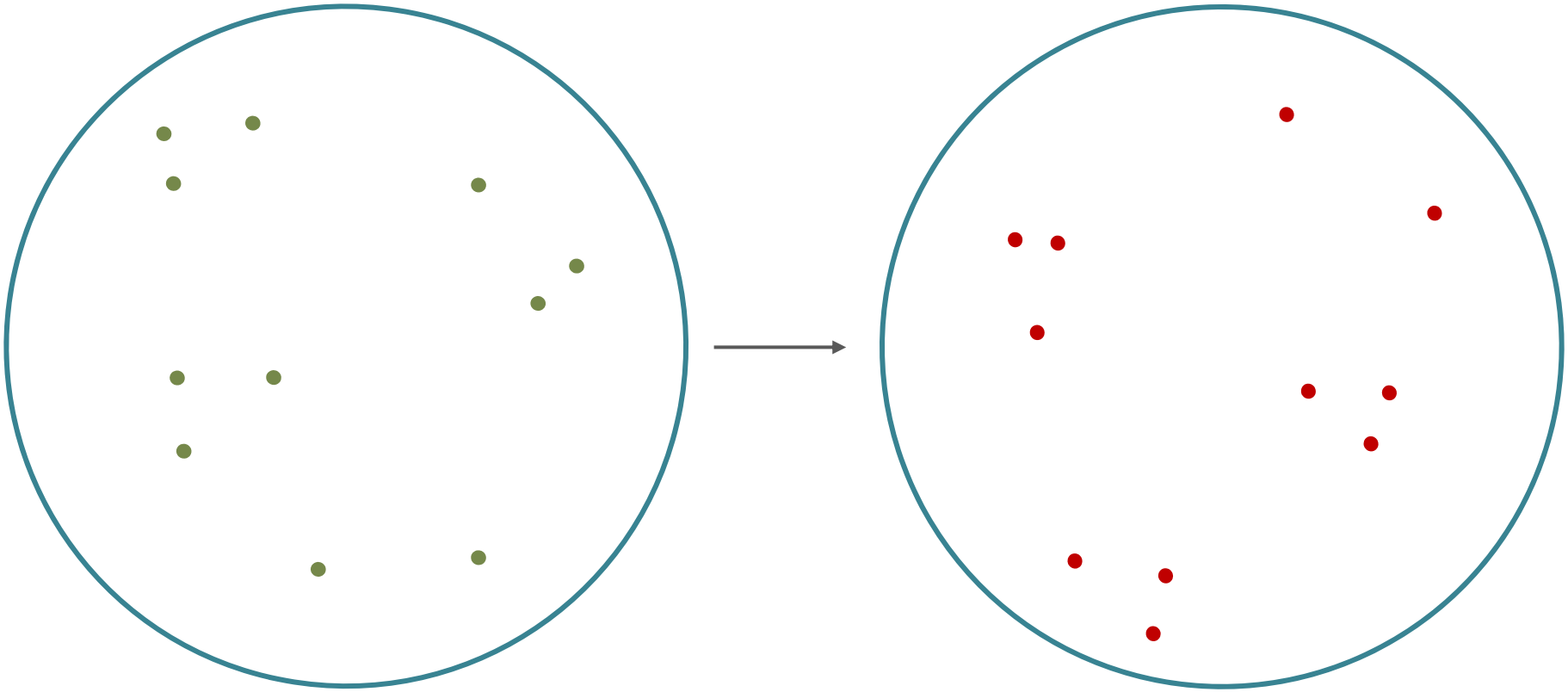


Why does it work?

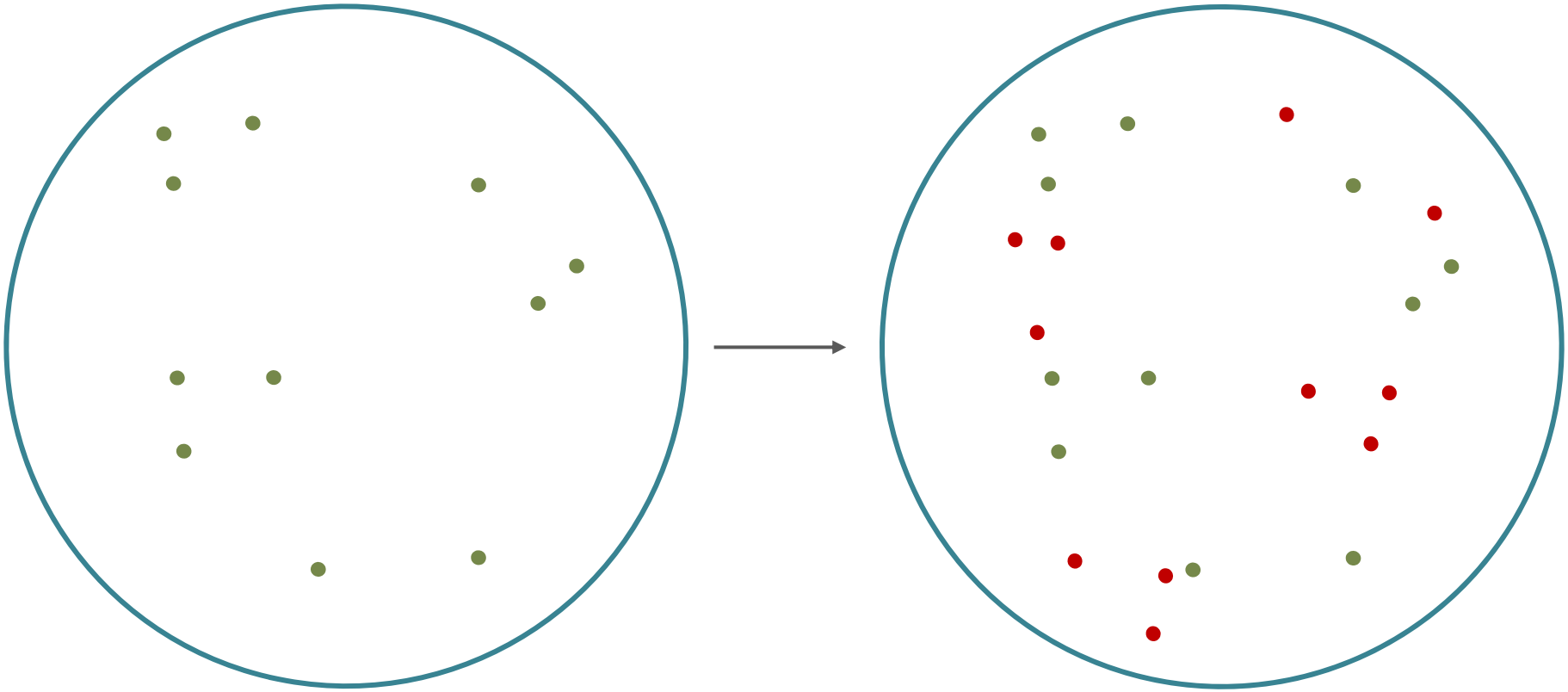




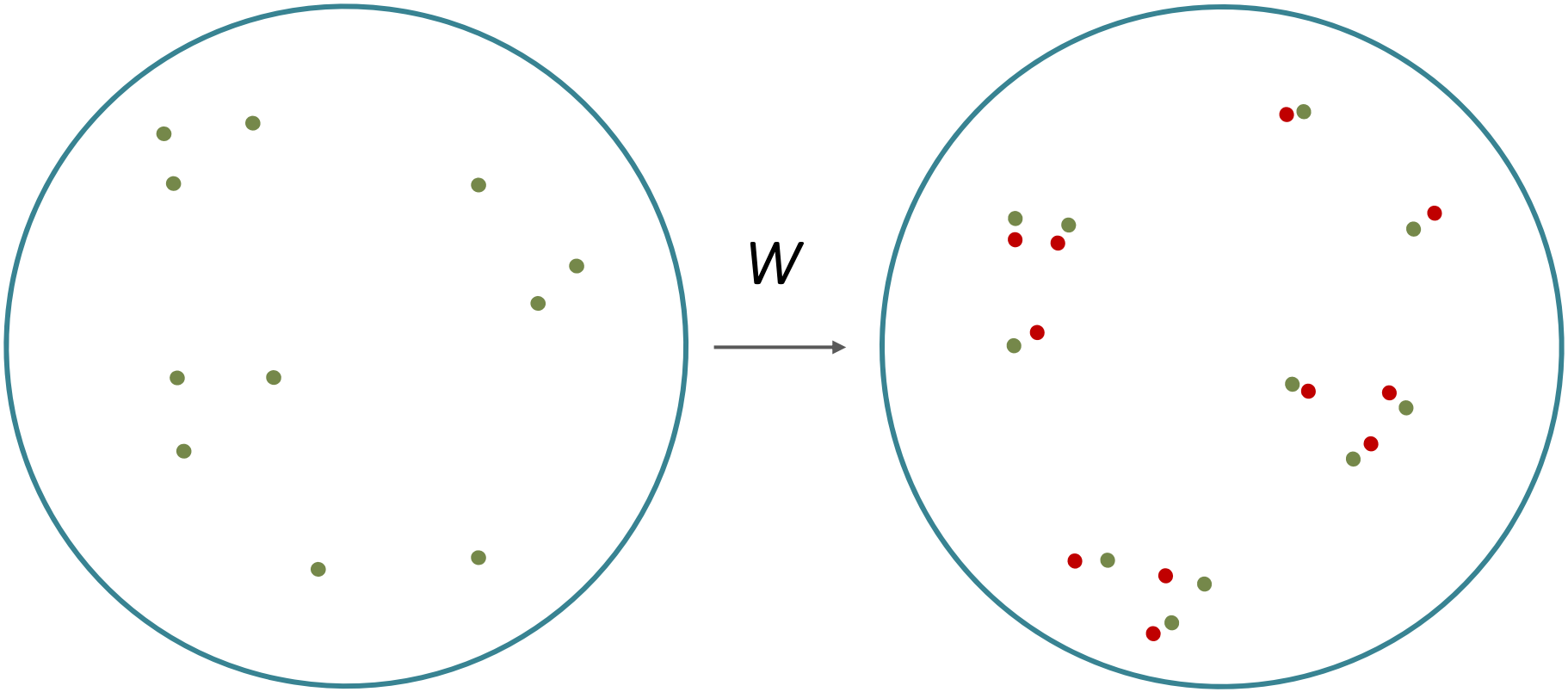
Why does it work?



Why does it work?

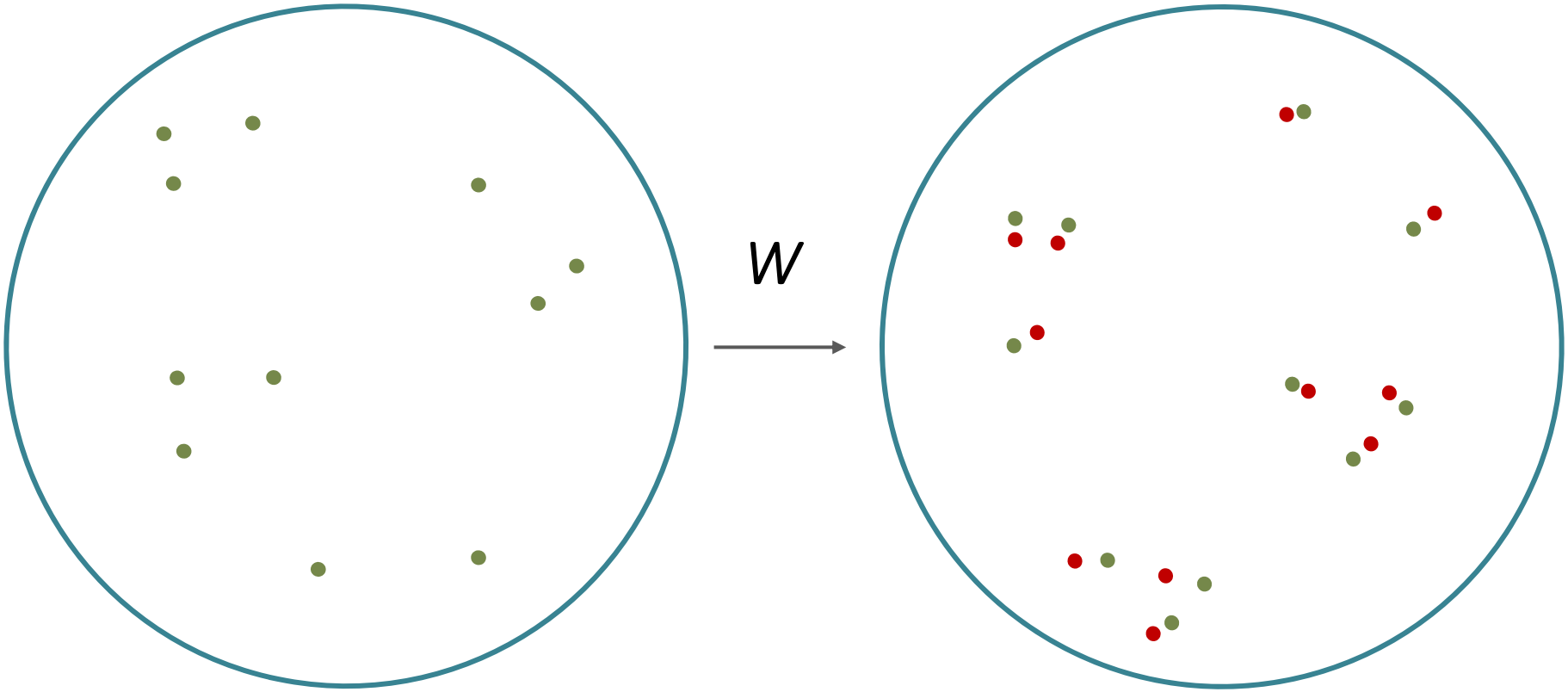


Why does it work?



Why does it work?

Languages are largely isometric
in embedding space (!)

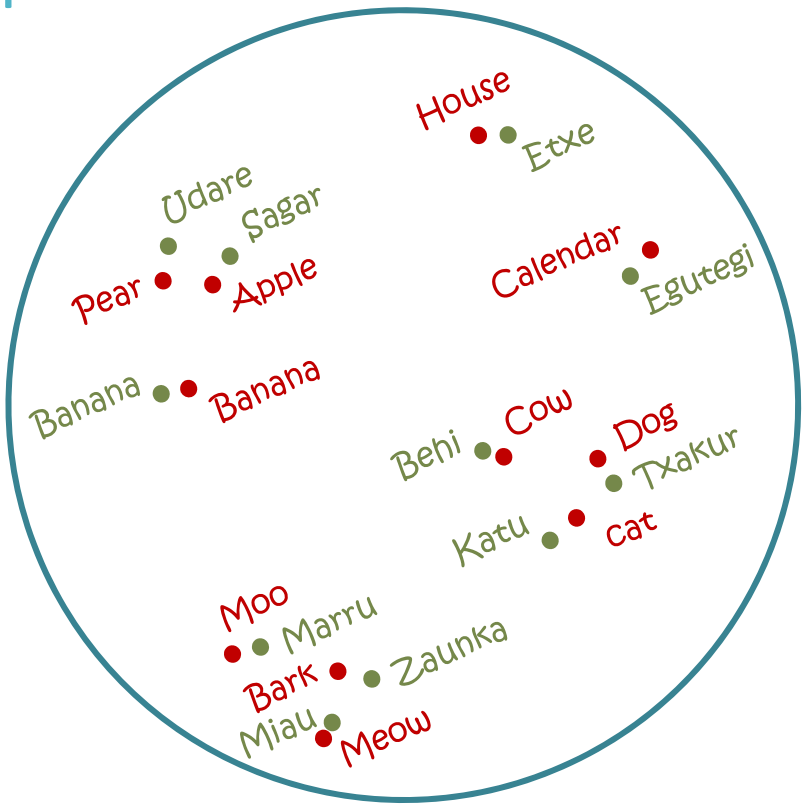


Outline

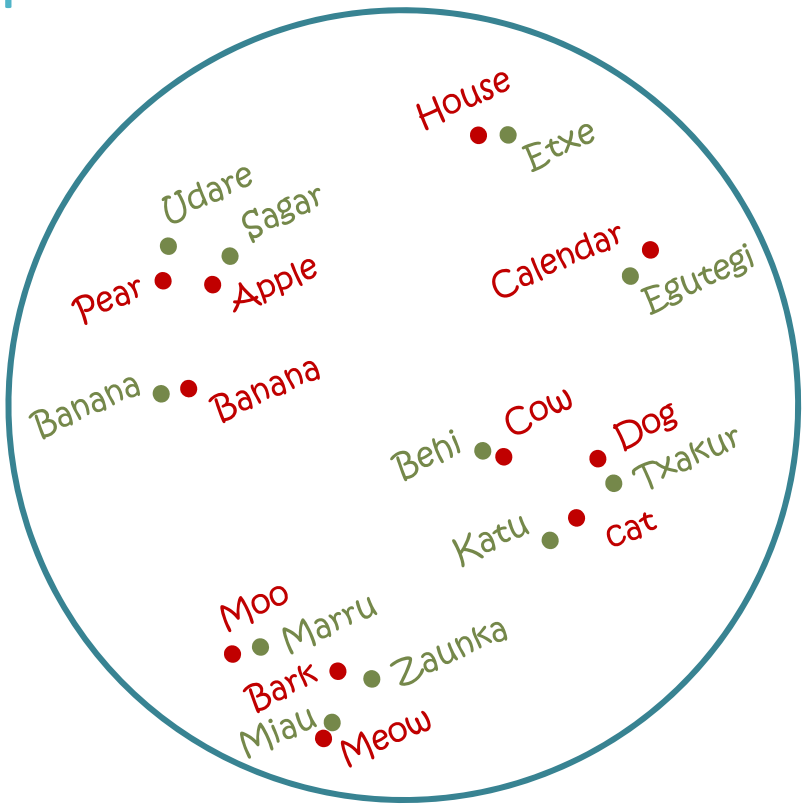
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Introduction to bilingual embedding mappings*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Conclusions*

Reducing supervision

Reducing supervision



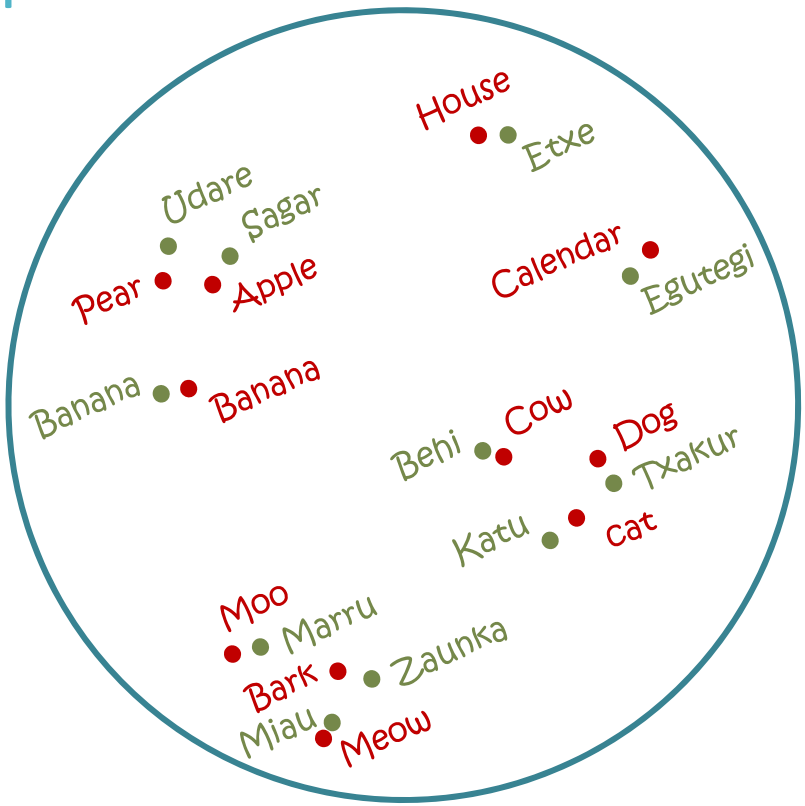
Reducing supervision



Previous work

bilingual signal
for training

Reducing supervision

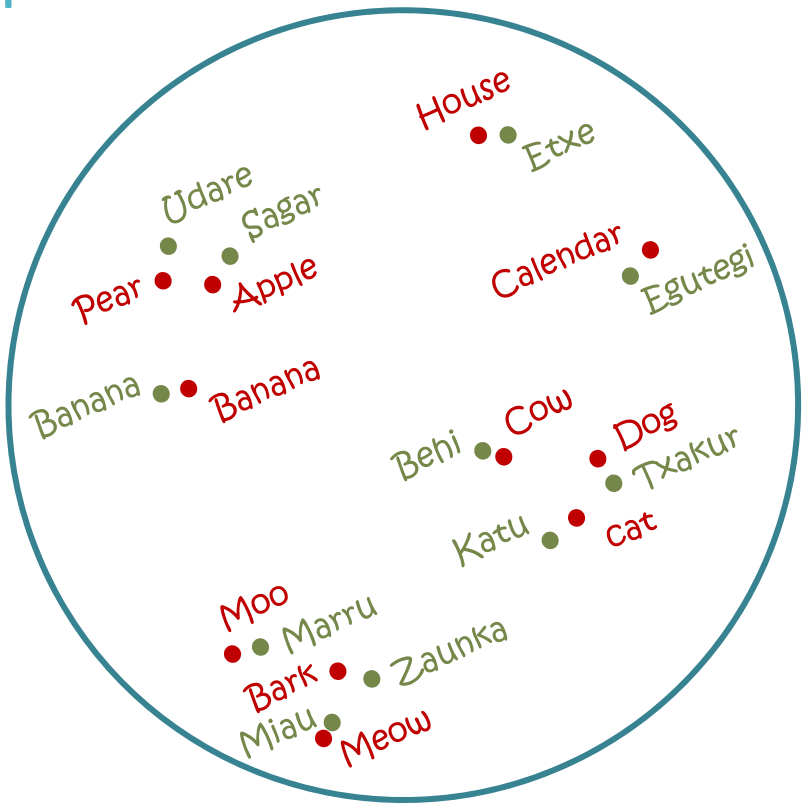


bilingual signal for training

<u>Previous work</u>

- parallel corpora
- comparable corpora
- (big) dictionaries

Reducing supervision

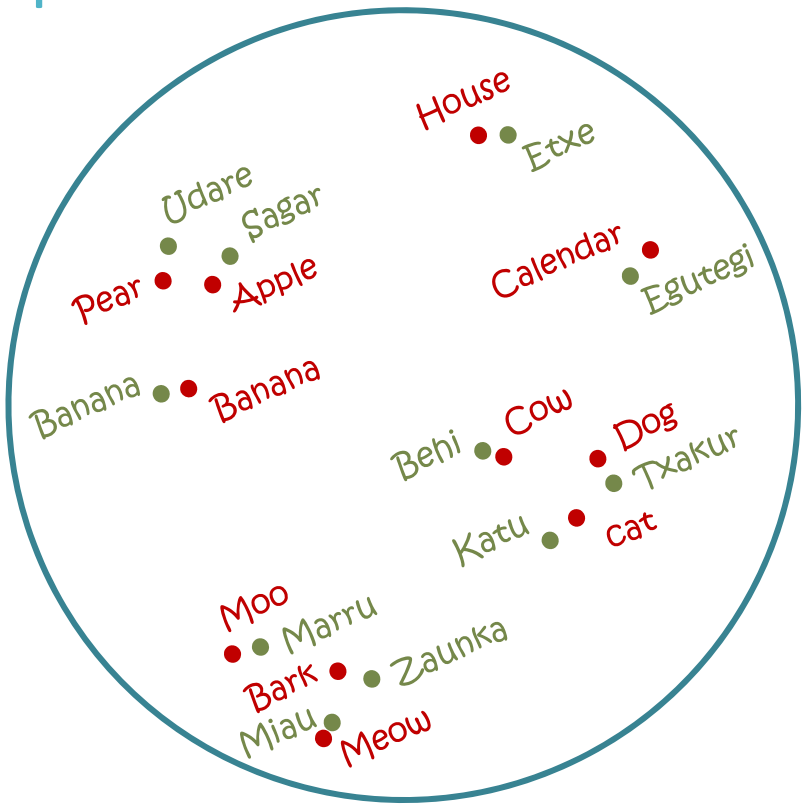


bilingual signal for training

Previous work

- parallel corpora
- comparable corpora
- (big) dictionaries

Removing supervision



bilingual signal for training

Previous work	Our work
- parallel corpora	- 25 word dictionary
- comparable corpora	- numerals (1, 2, 3…)
- (big) dictionaries	- nothing

Self-learning

Self-learning

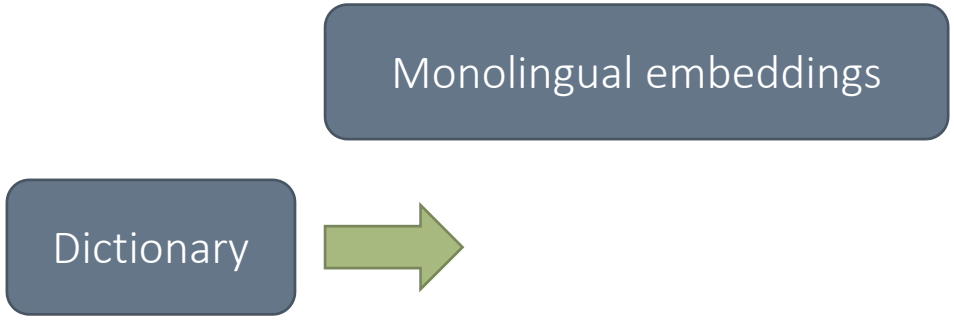
Monolingual embeddings

Self-learning

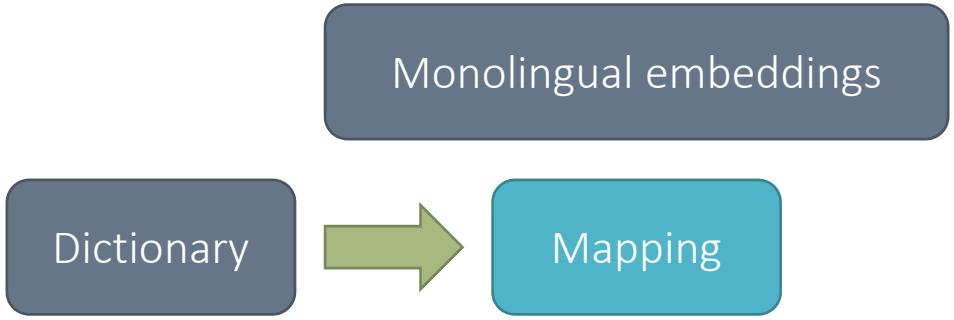
Monolingual embeddings

Dictionary

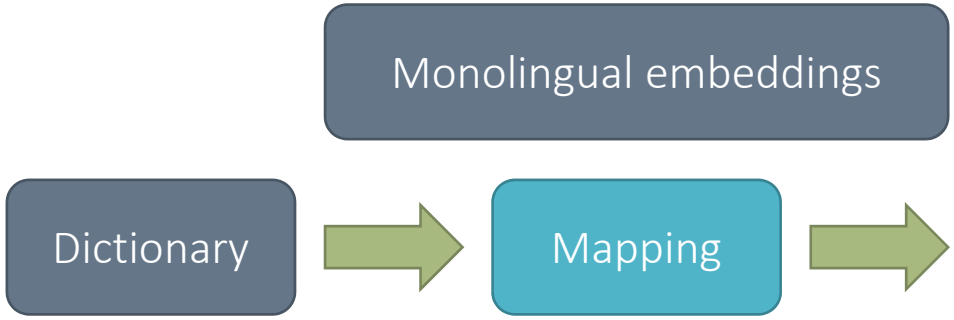
Self-learning



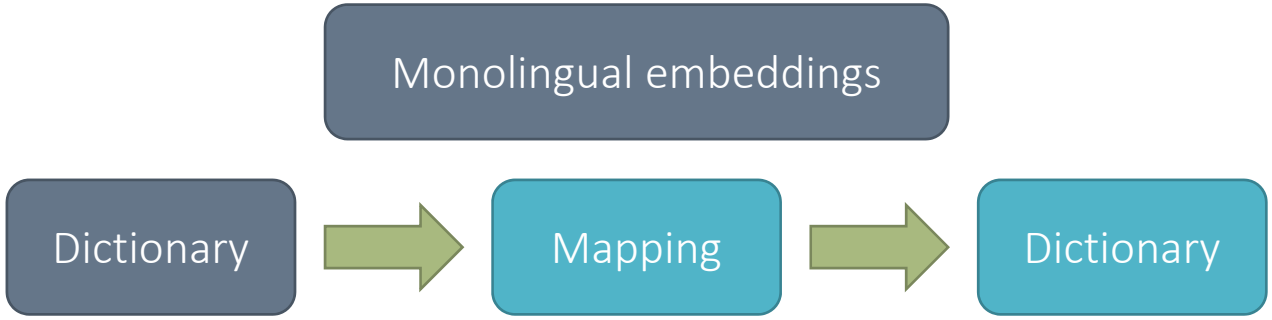
Self-learning



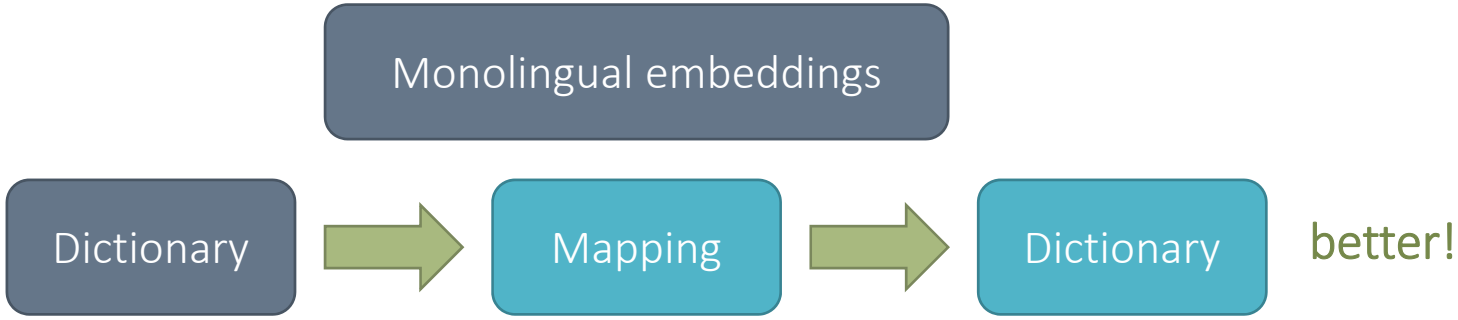
Self-learning



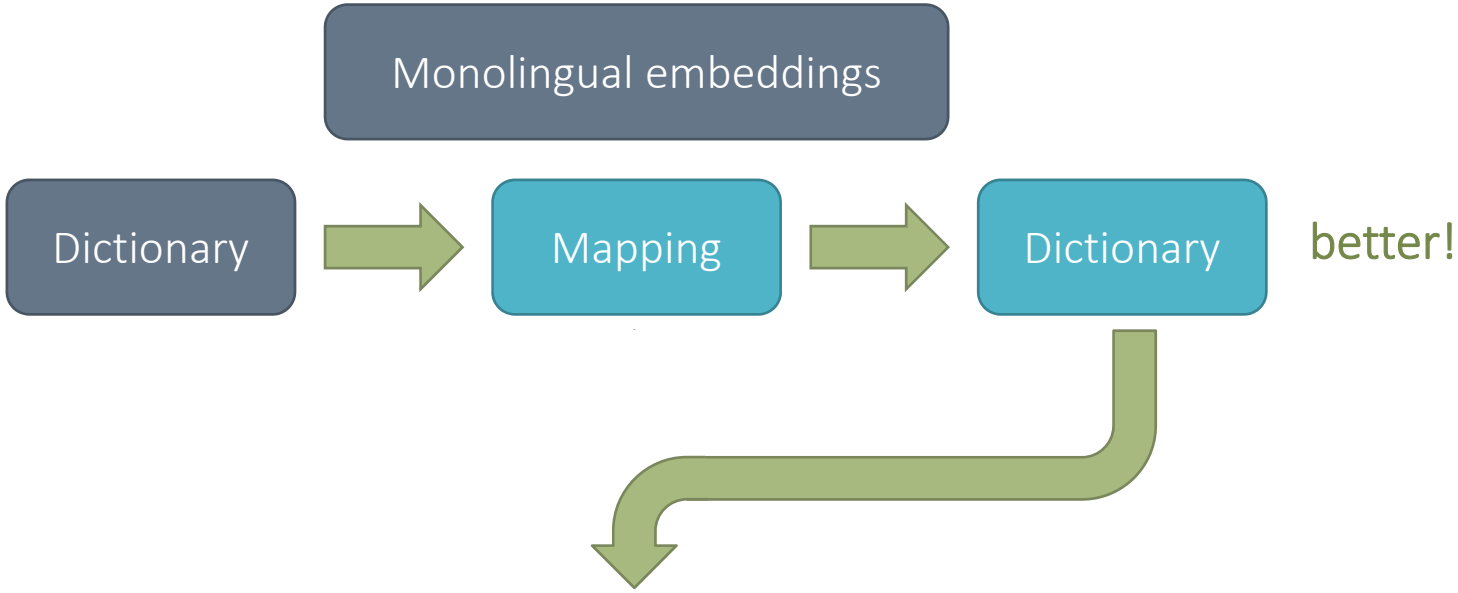
Self-learning



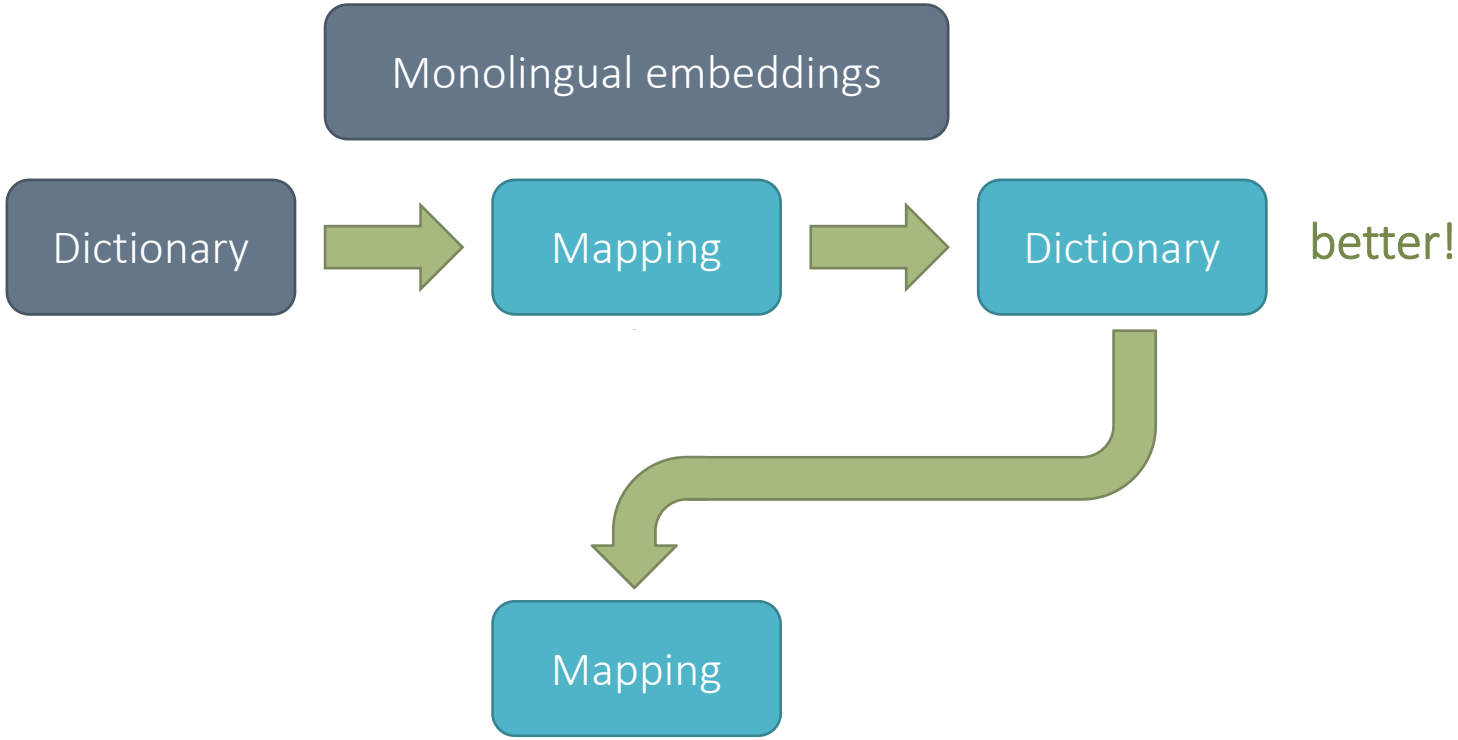
Self-learning



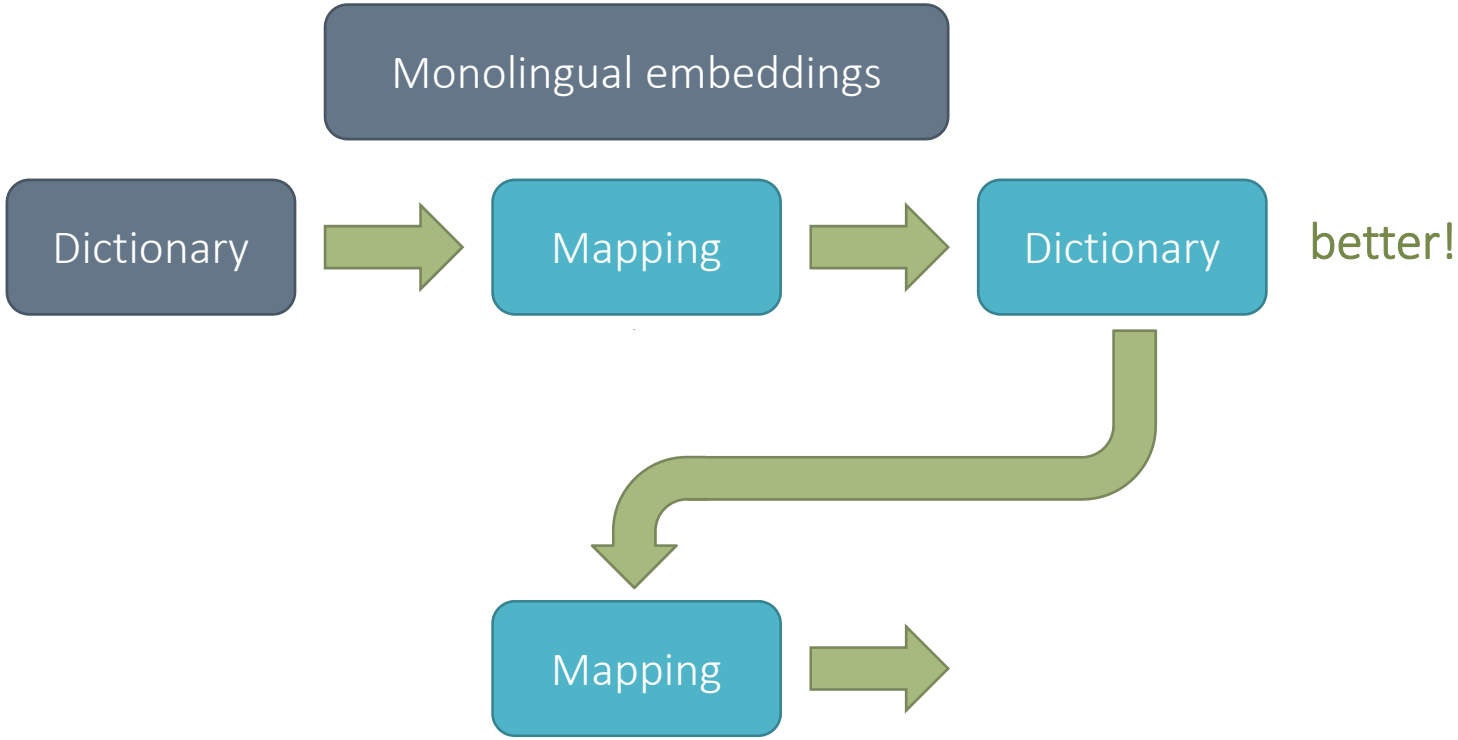
Self-learning



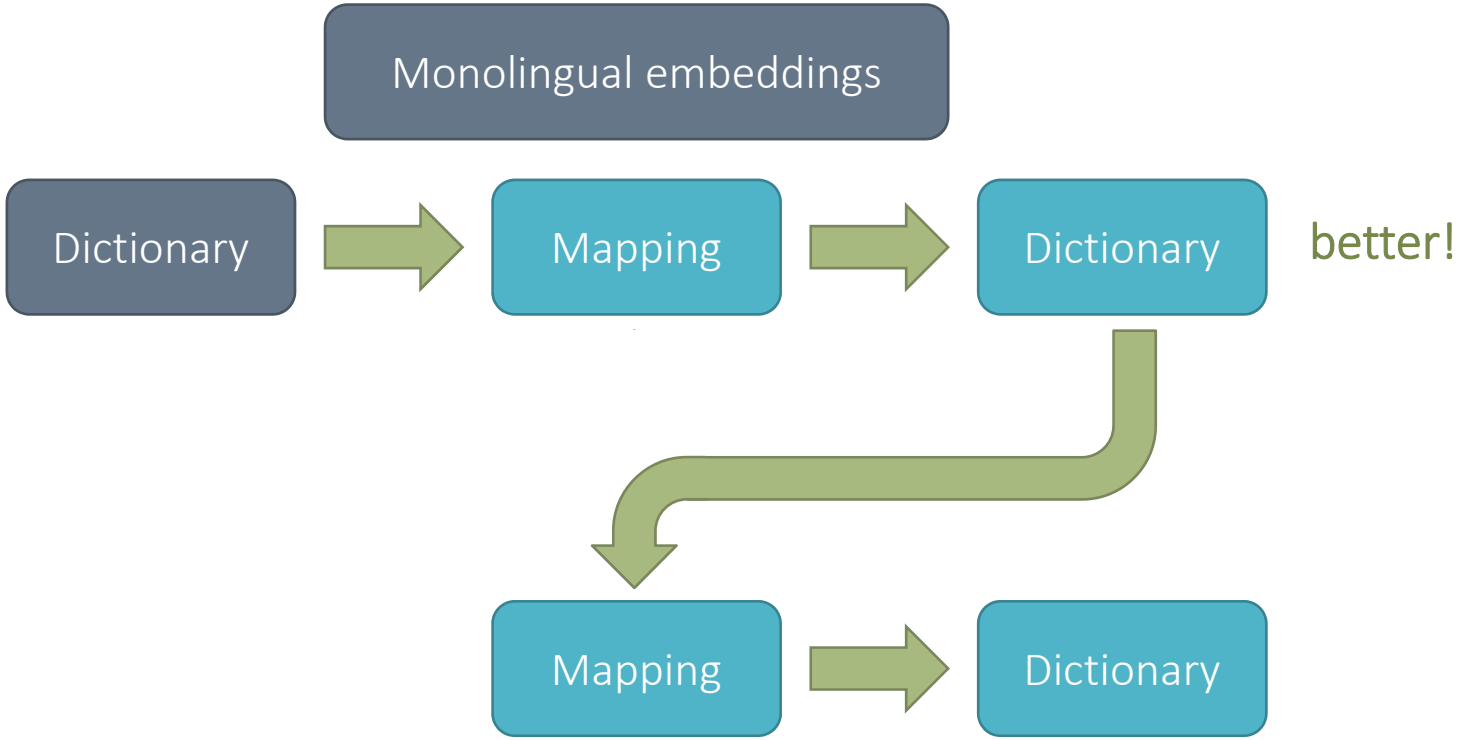
Self-learning



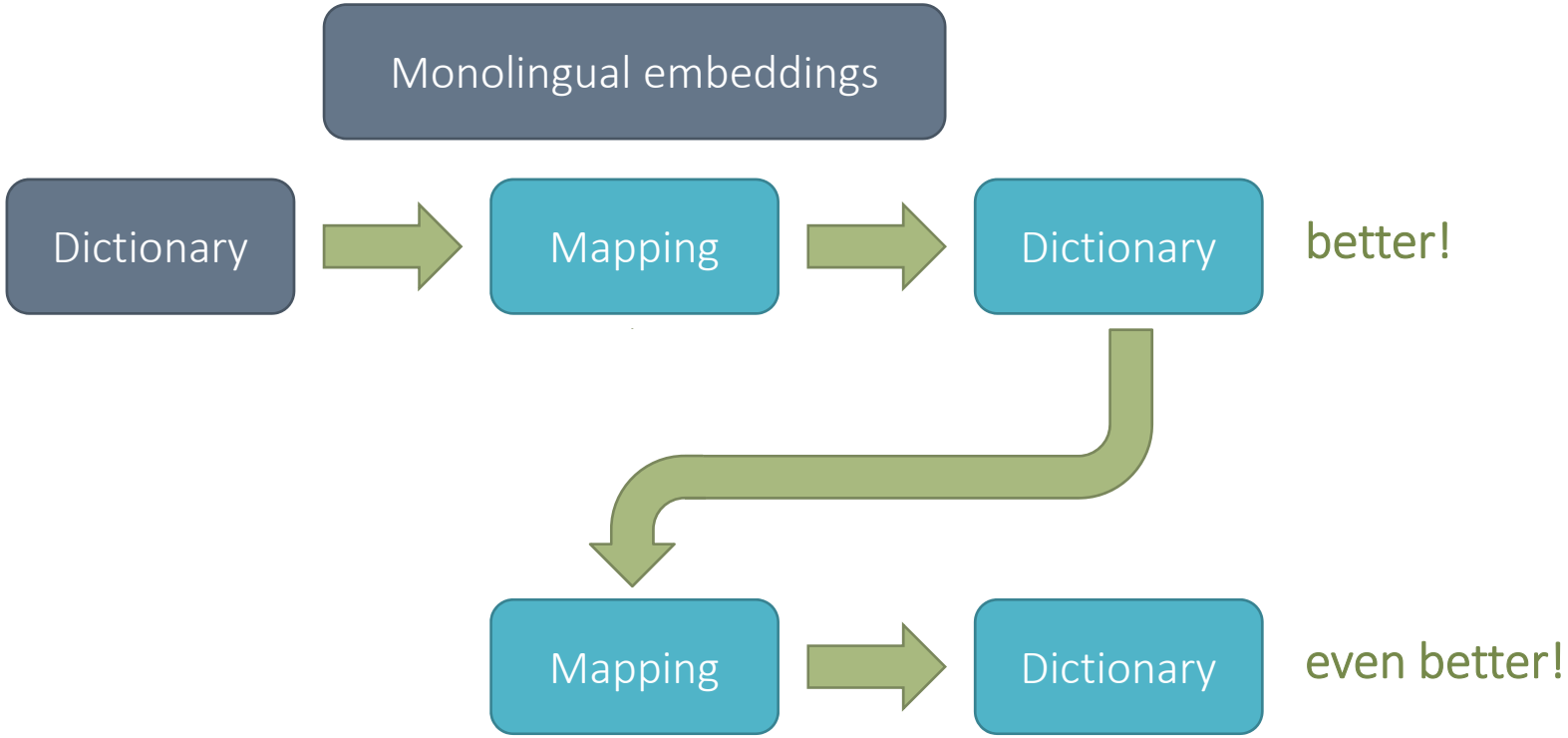
Self-learning



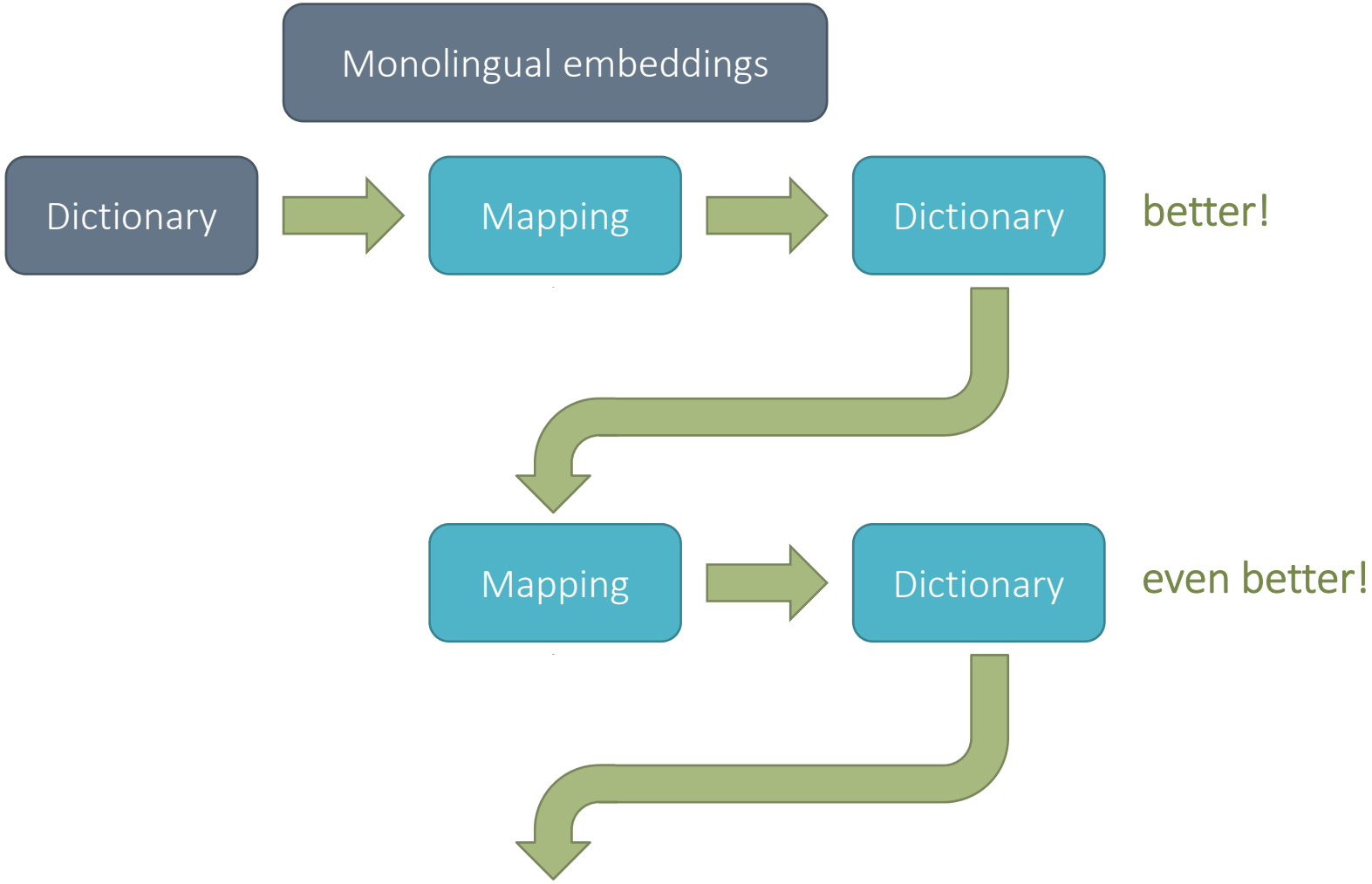
Self-learning



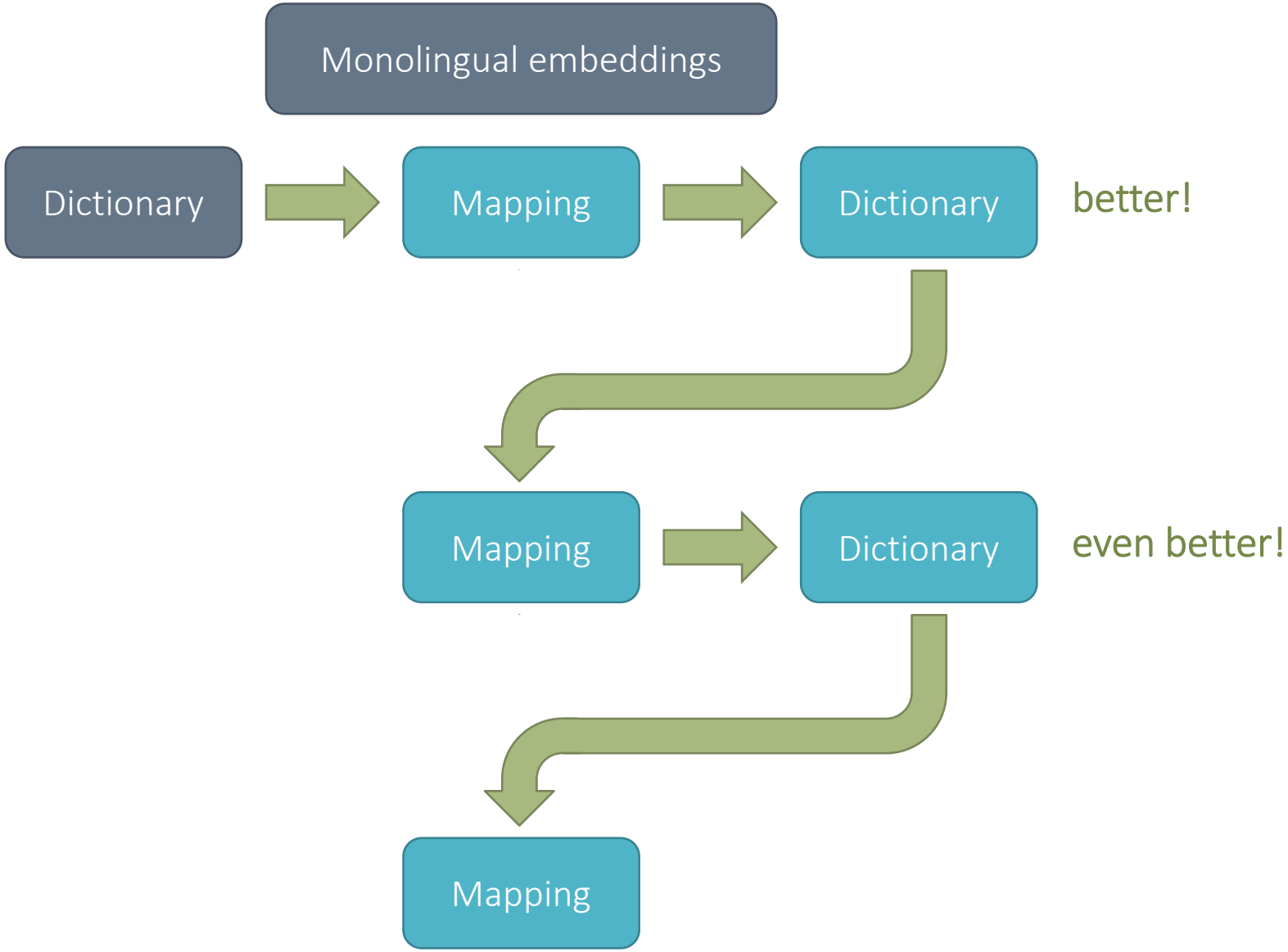
Self-learning



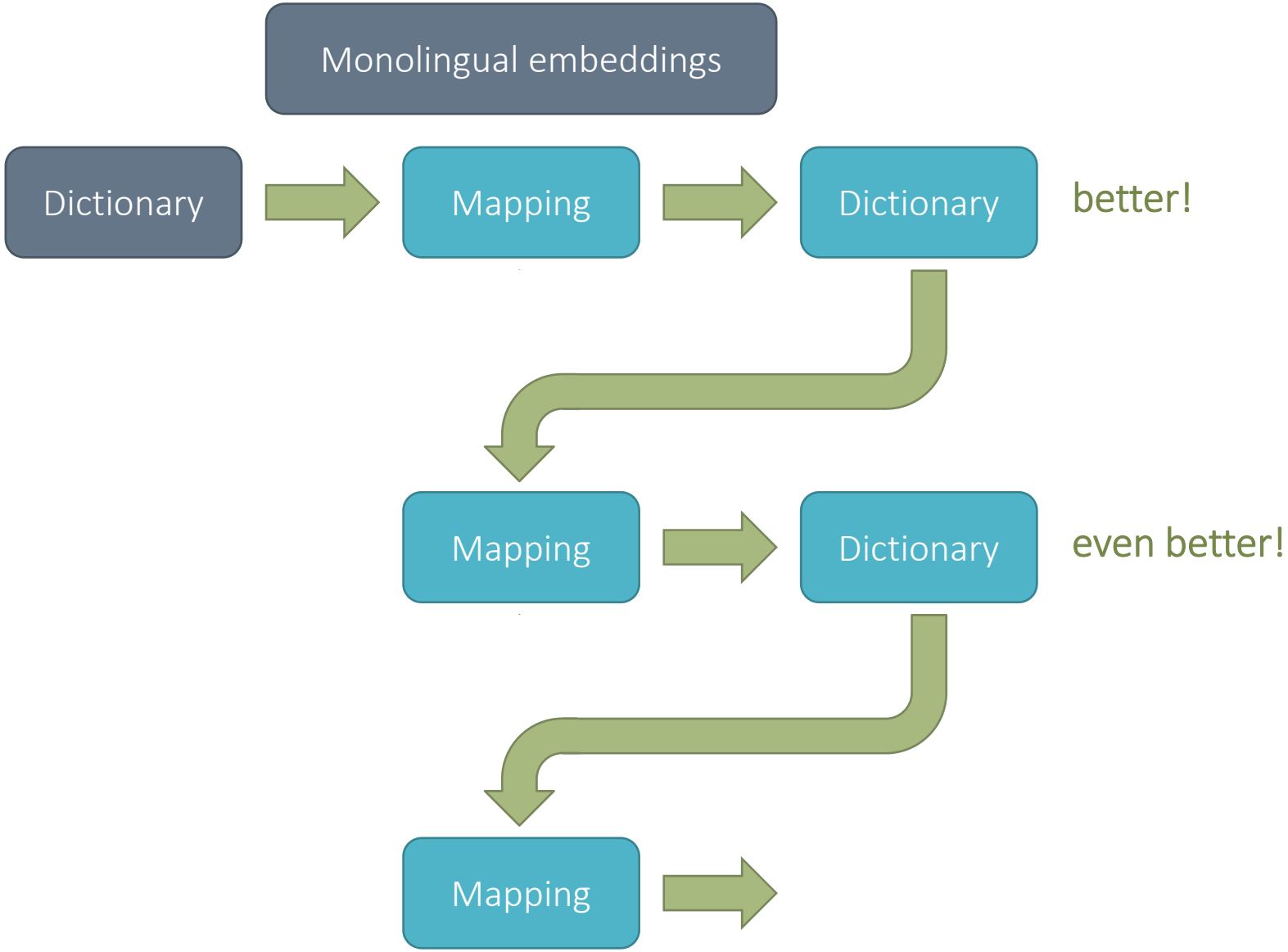
Self-learning



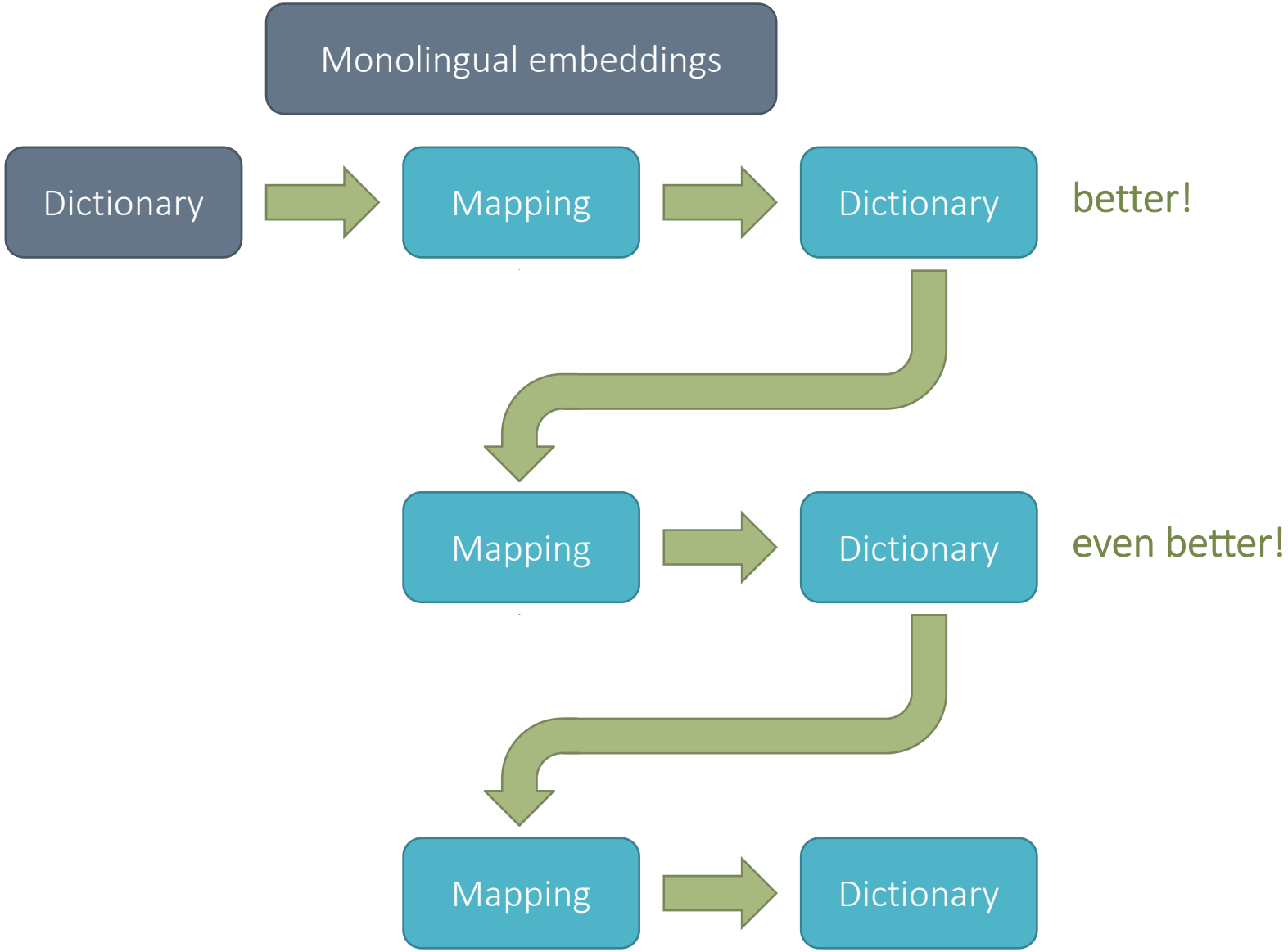
Self-learning



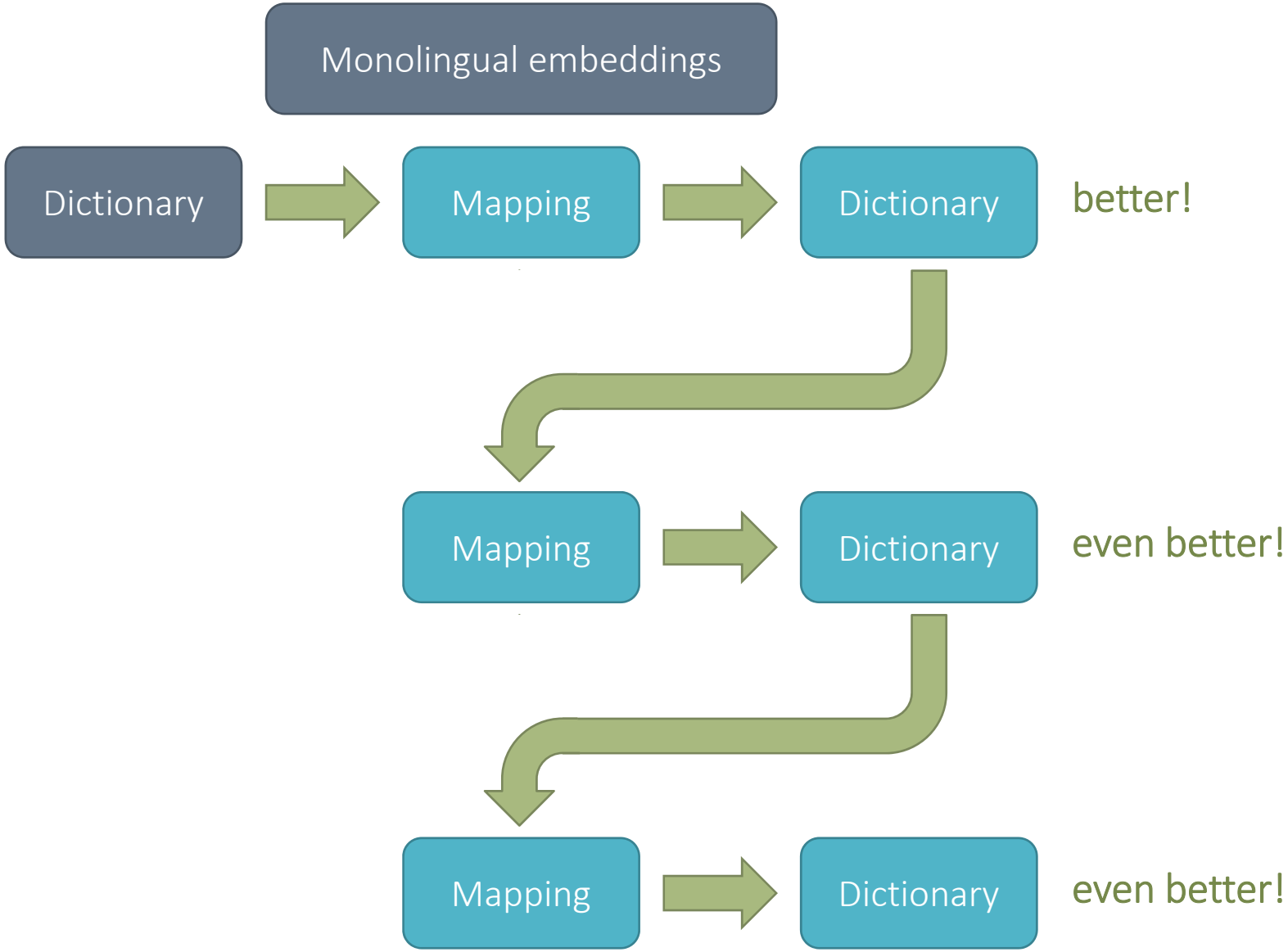
Self-learning



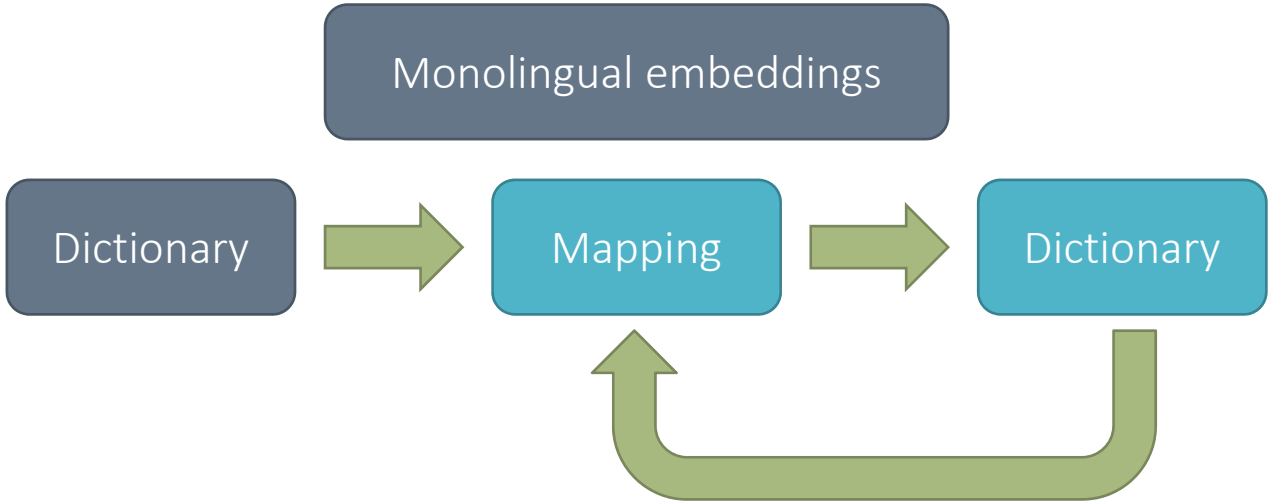
Self-learning



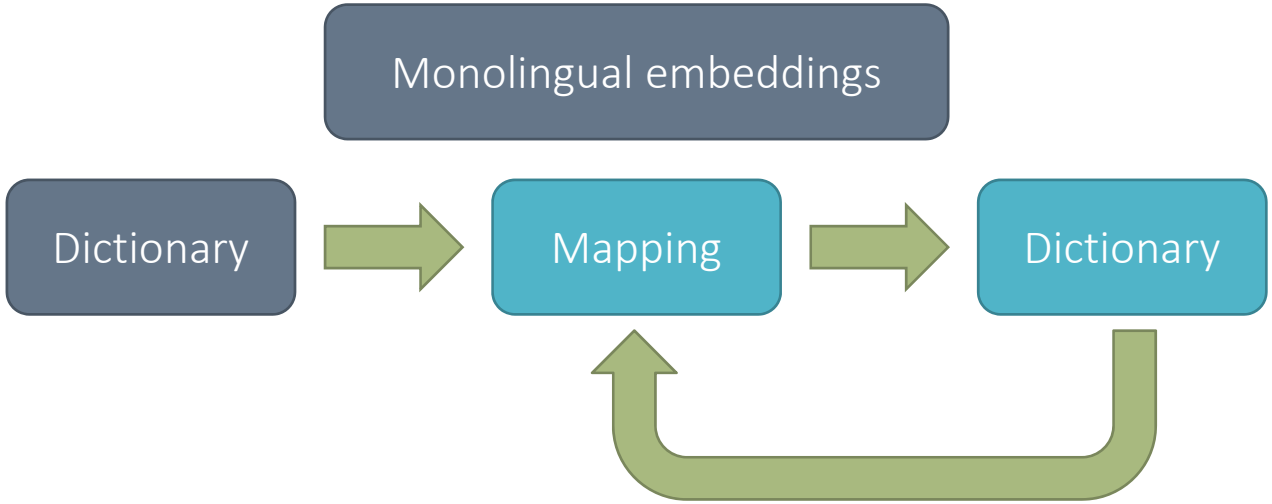
Self-learning



Self-learning



Self-learning



proposed self-learning method

Too good to be true?

Semi-supervised experiments (ACL17)

Semi-supervised experiments (ACL17)

- Given monolingual embeddings
  plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals

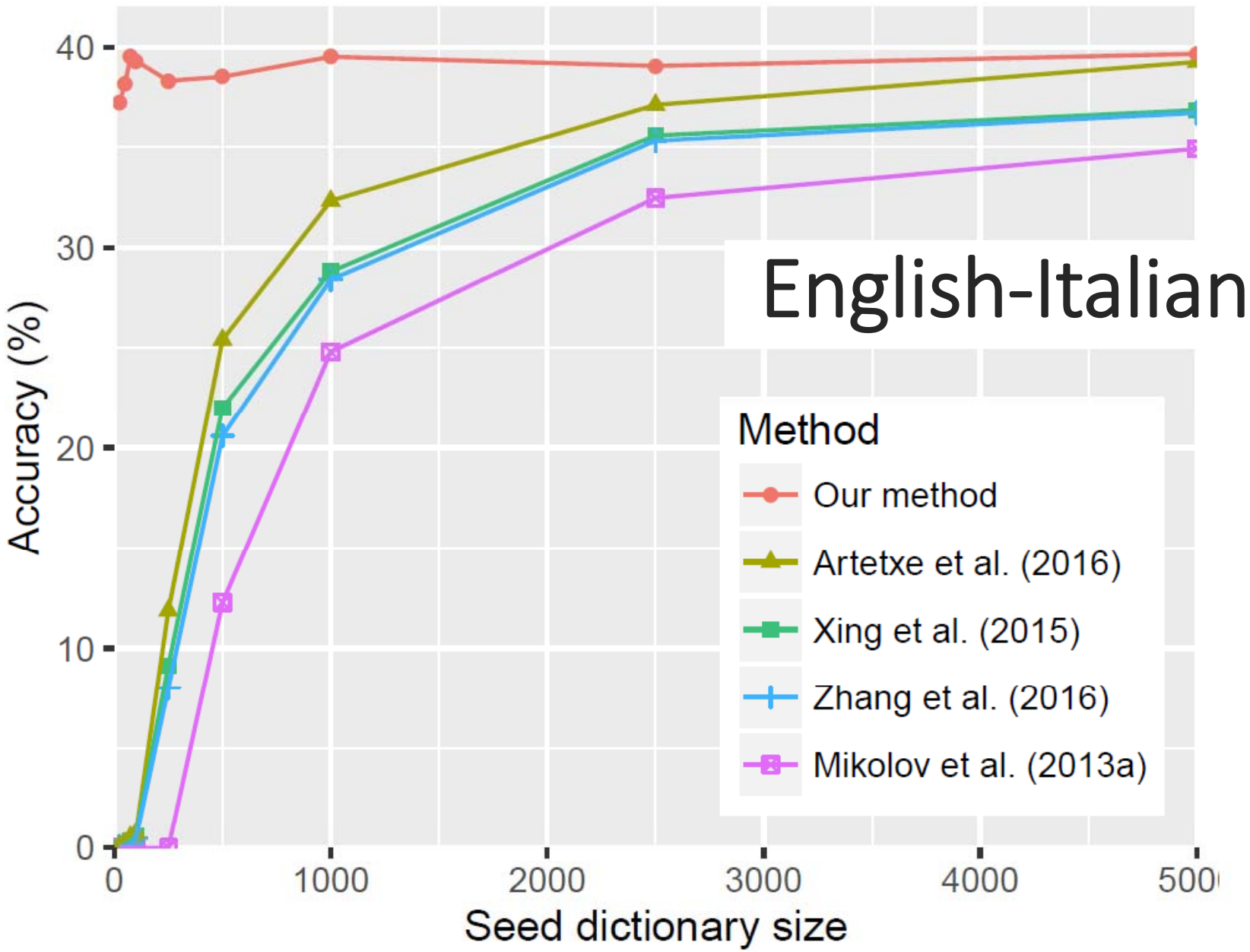
Semi-supervised experiments (ACL17)

- Given monolingual embeddings
  plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning
  for full vocabulary

Semi-supervised experiments (ACL17)

- Given monolingual embeddings plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning for full vocabulary
- Evaluation
  - Compare translations to existing bilingual dictionary (*test* dictionary)
  - Accuracy

Semi-supervised experiments (ACL17)



Why does it work?

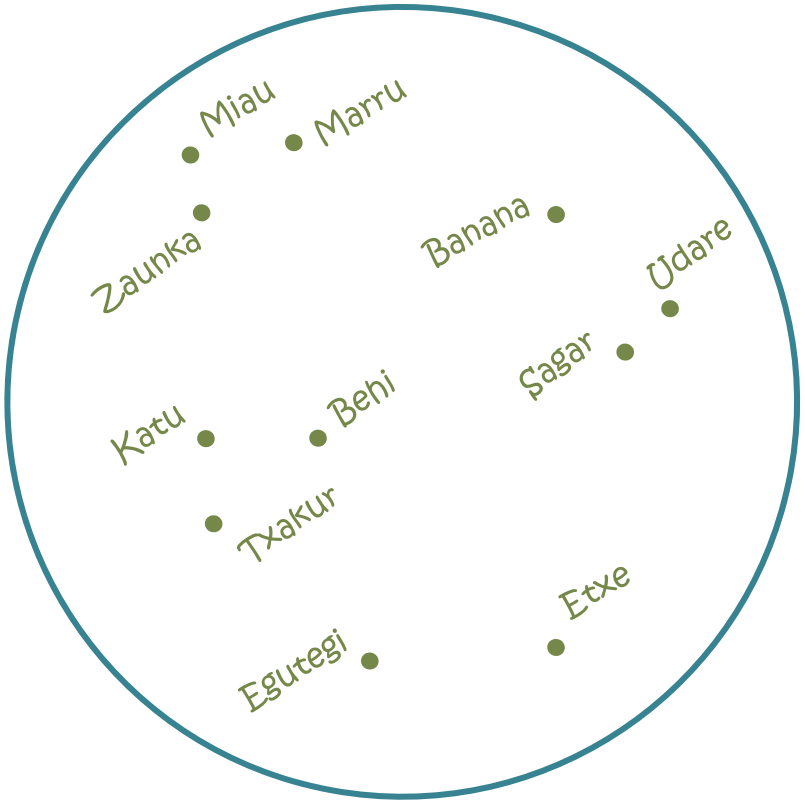
Why does it work?

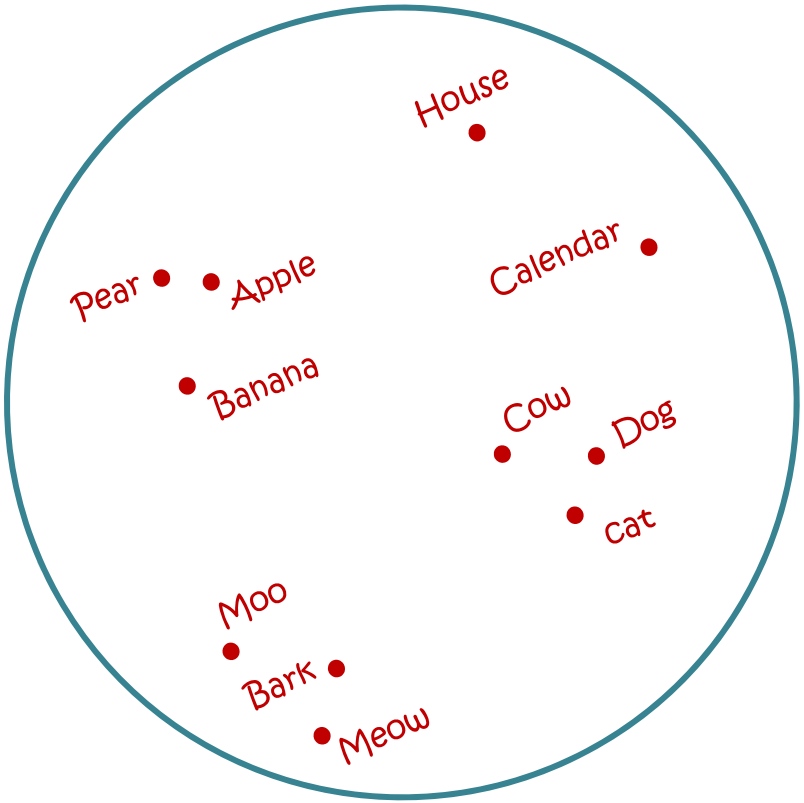
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$

Why does it work?

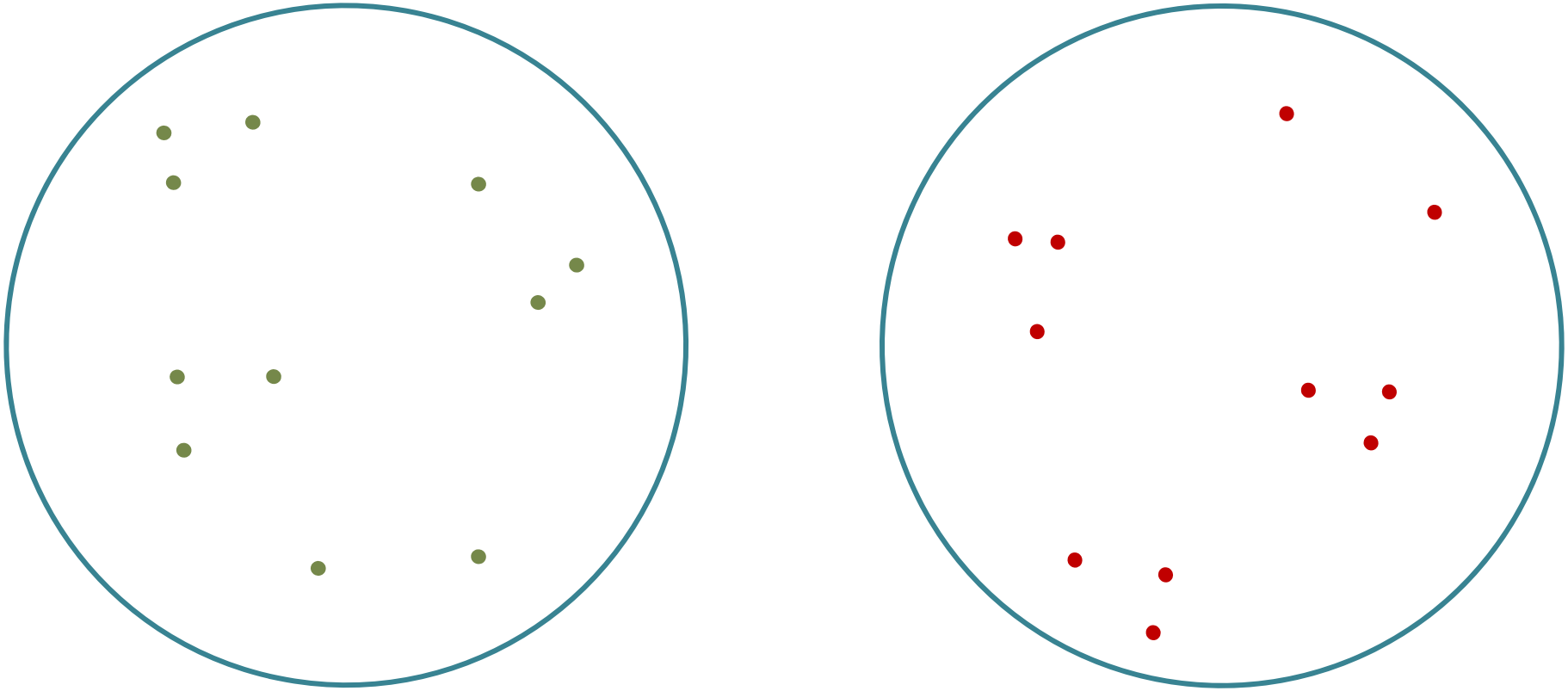
<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$





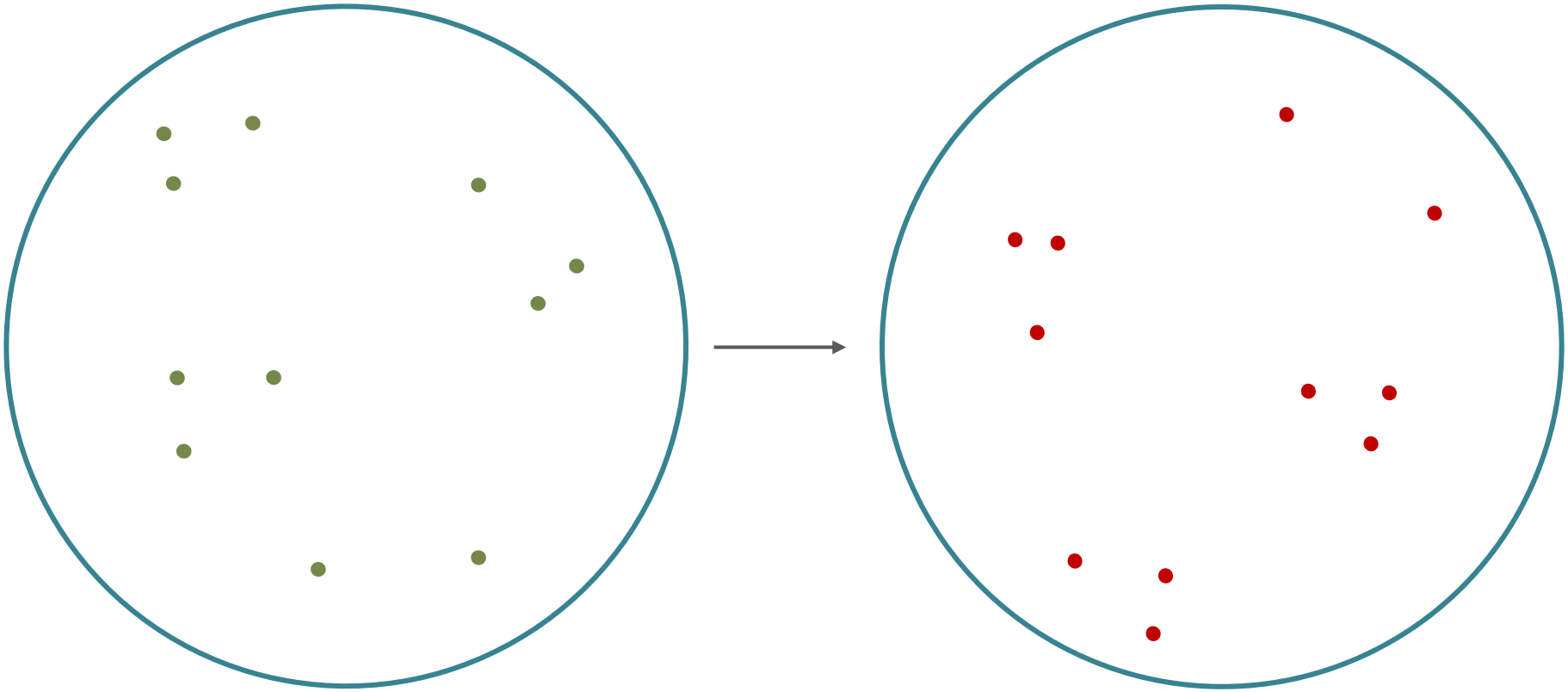
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



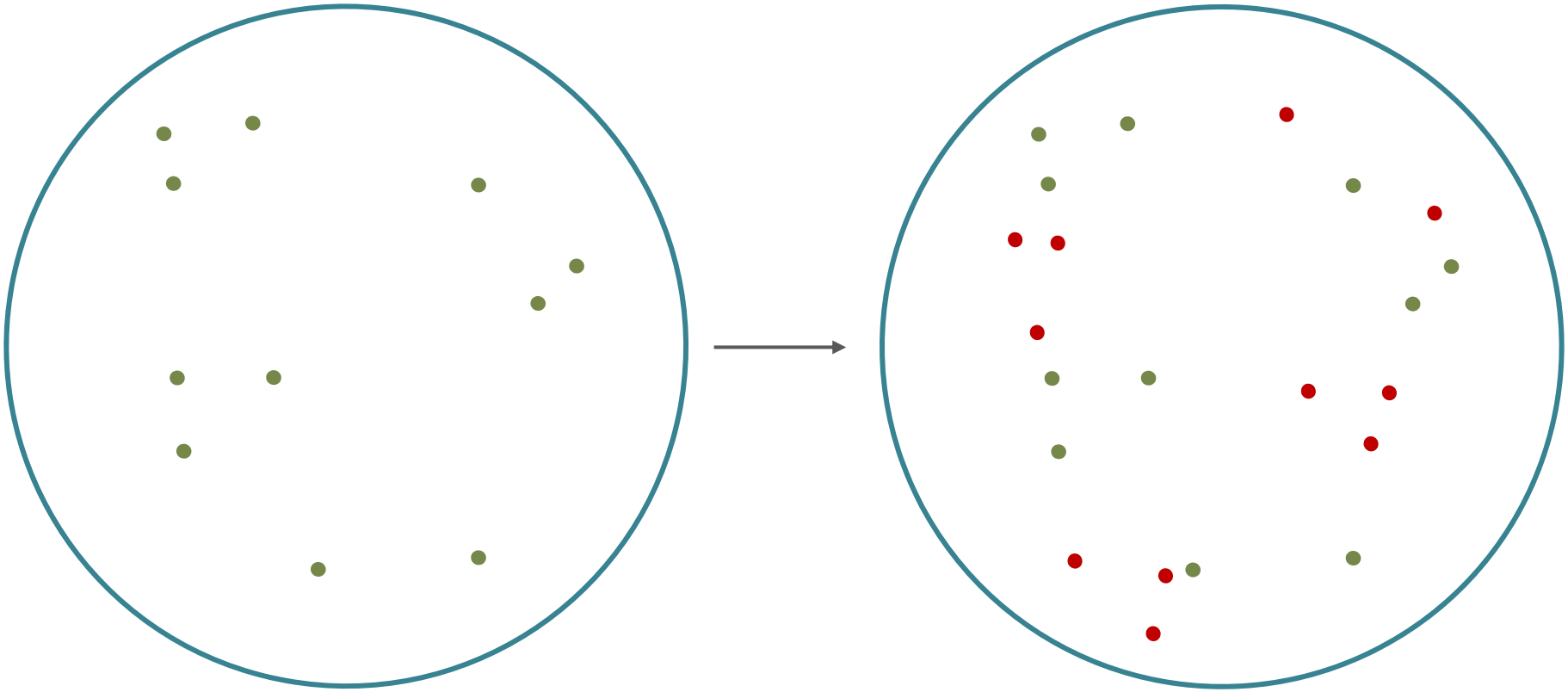
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



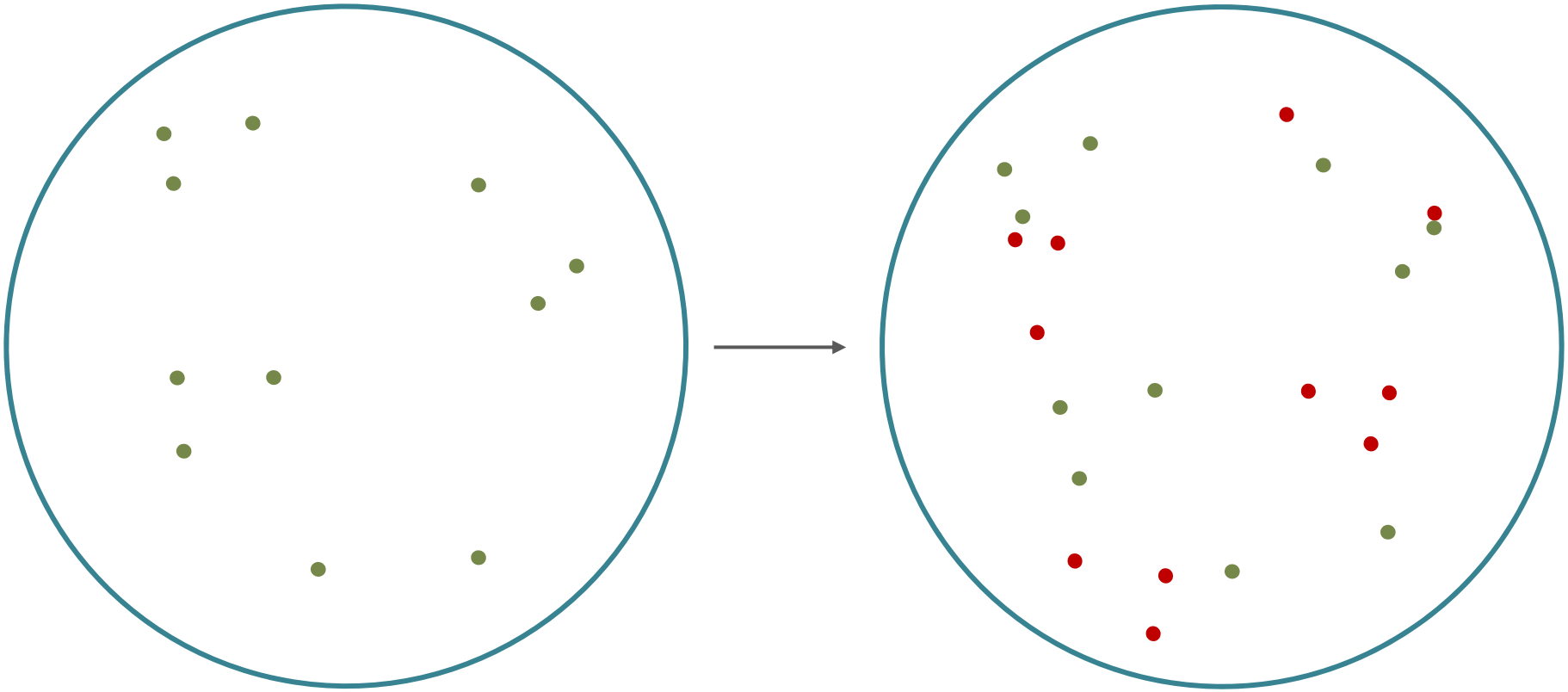
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



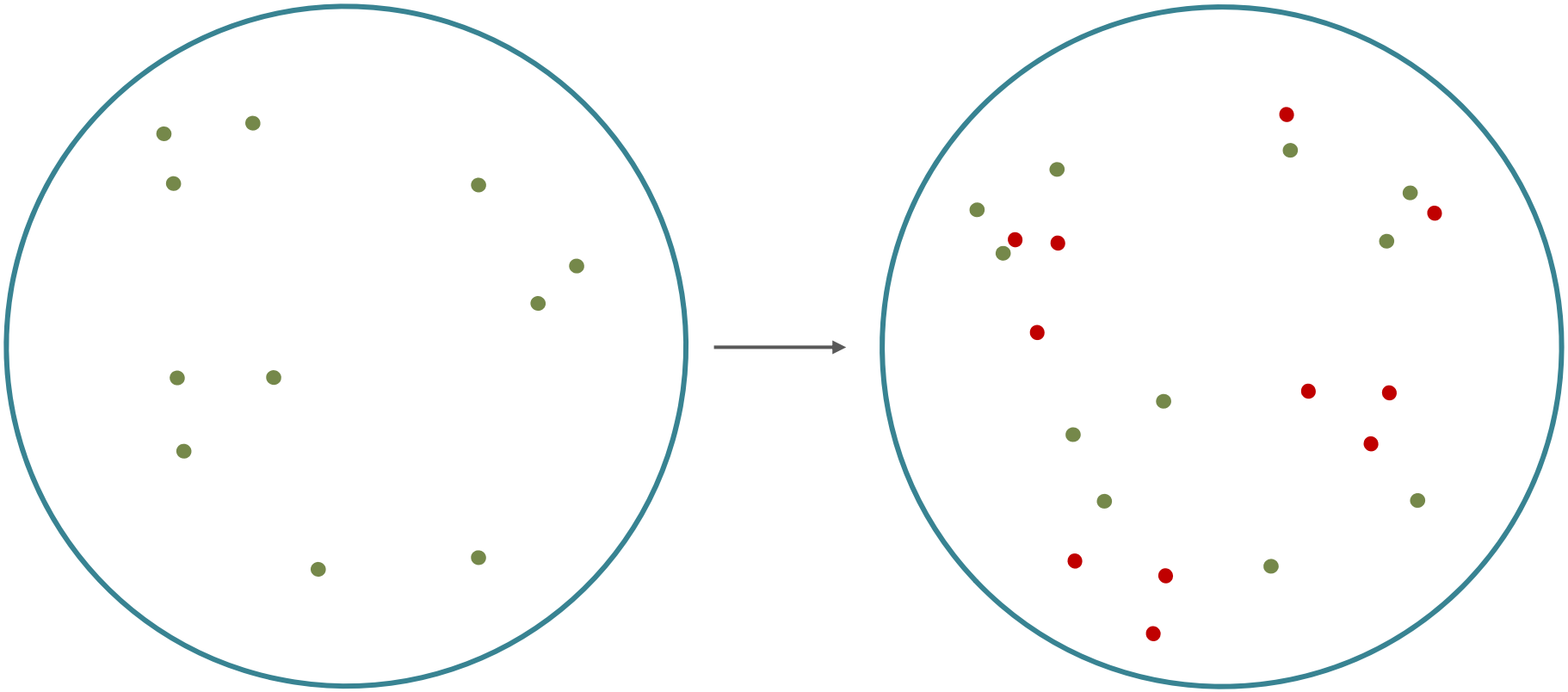
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



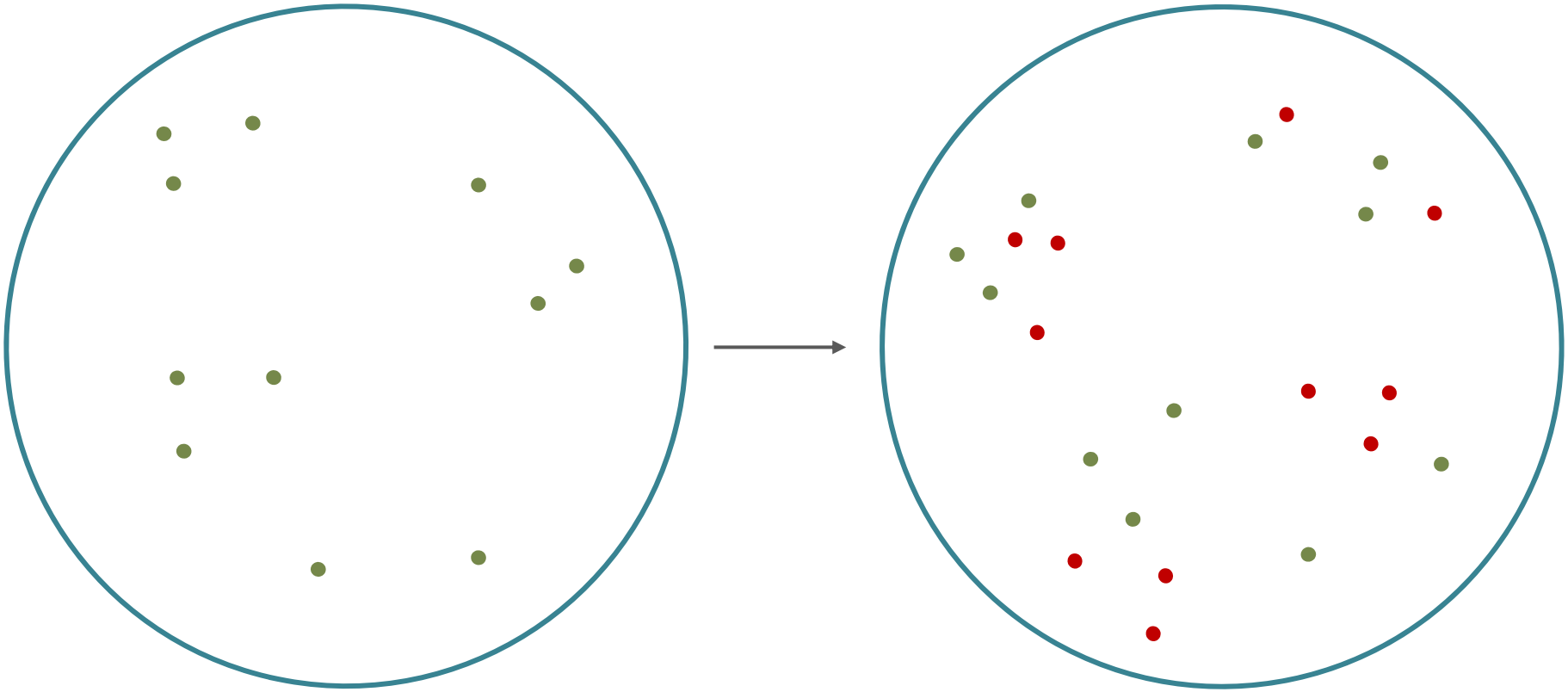
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



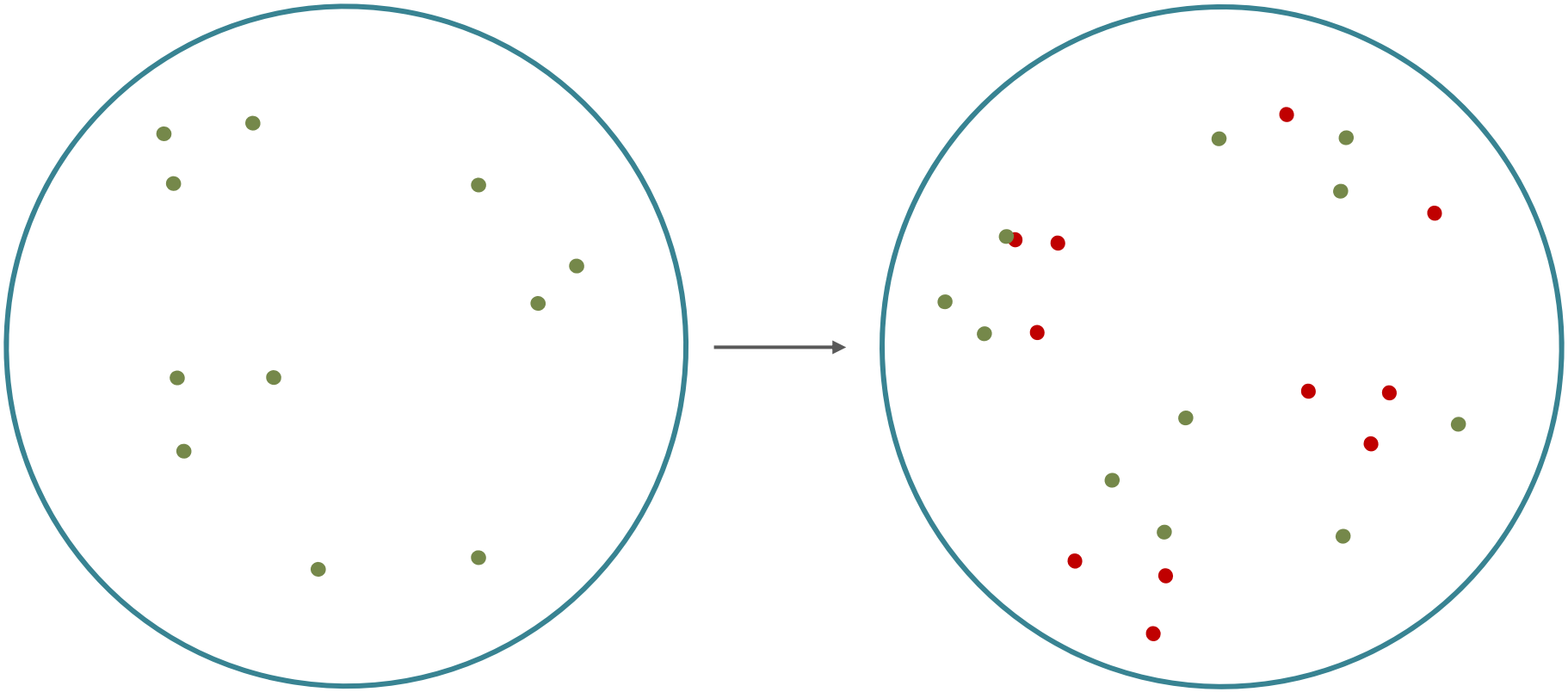
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



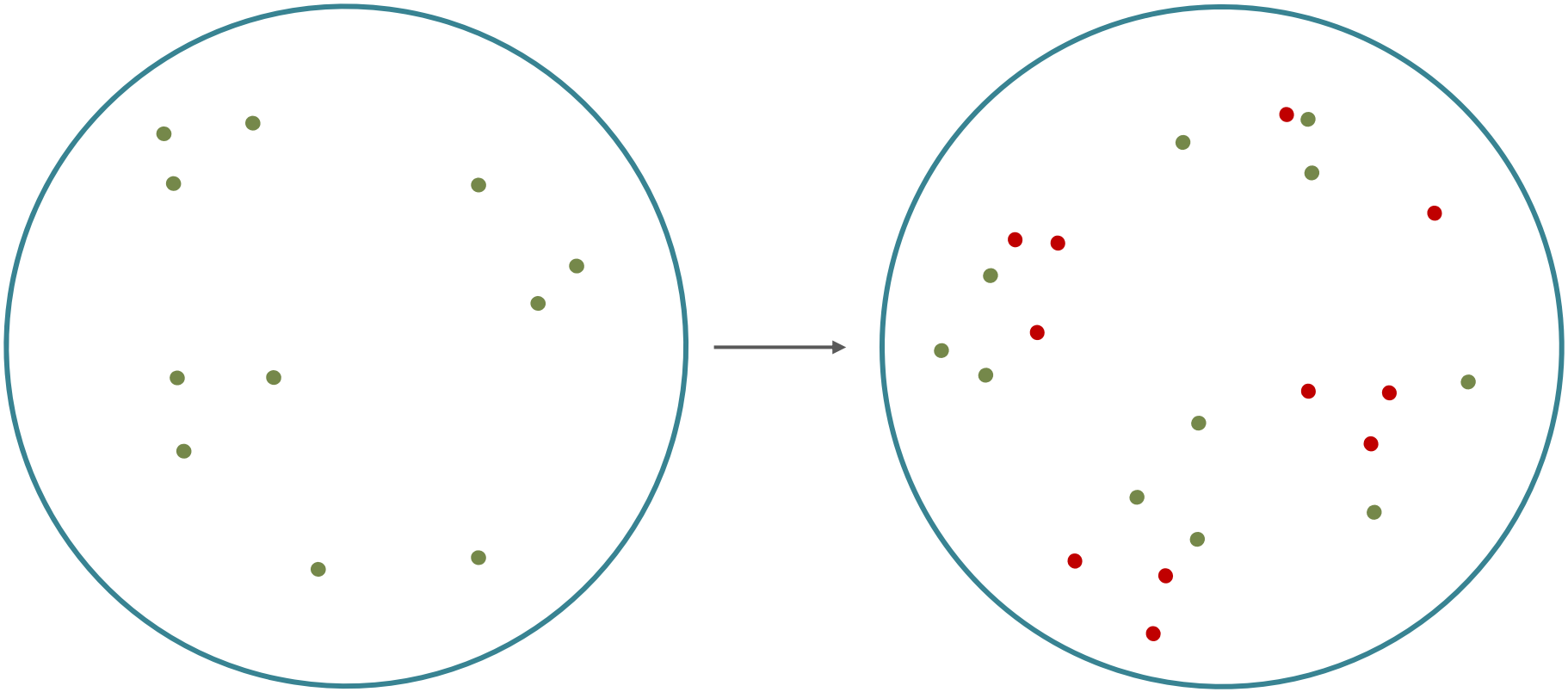
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



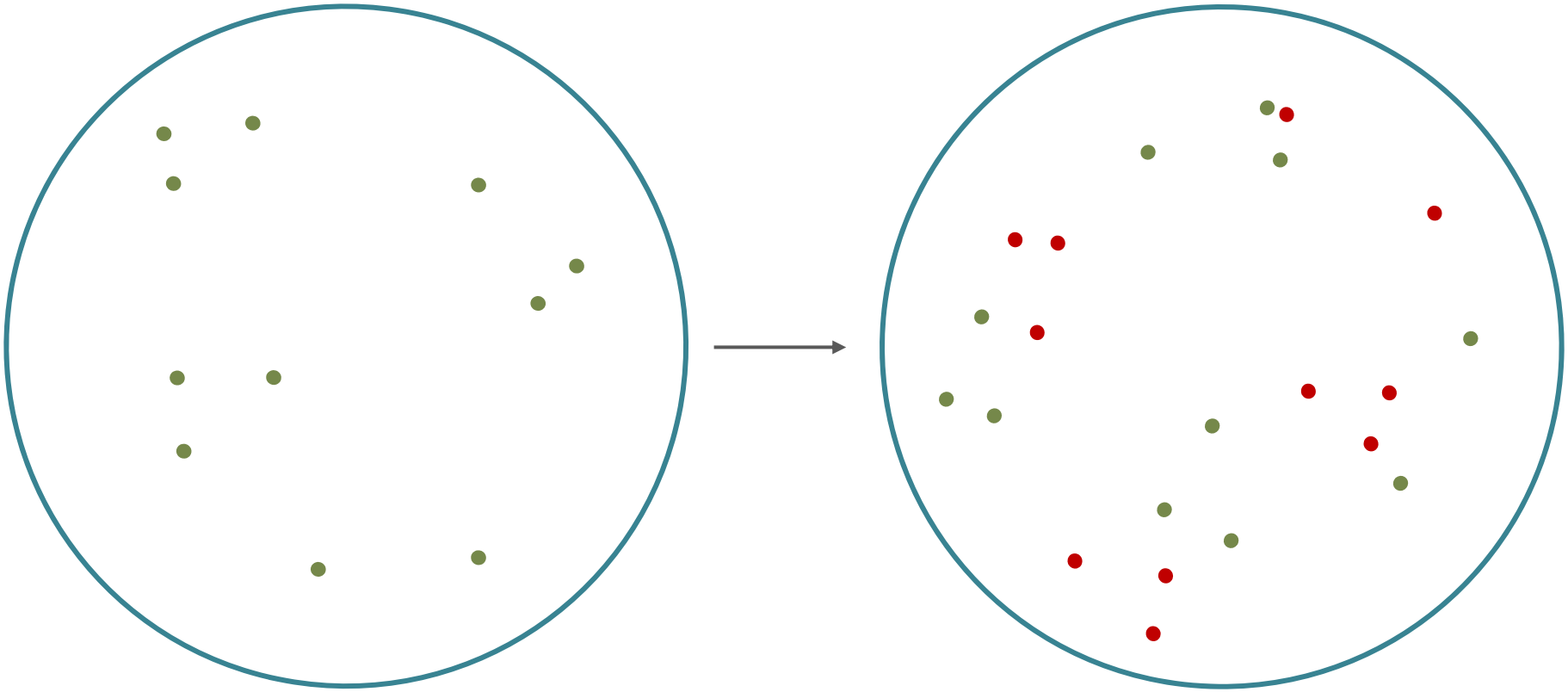
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



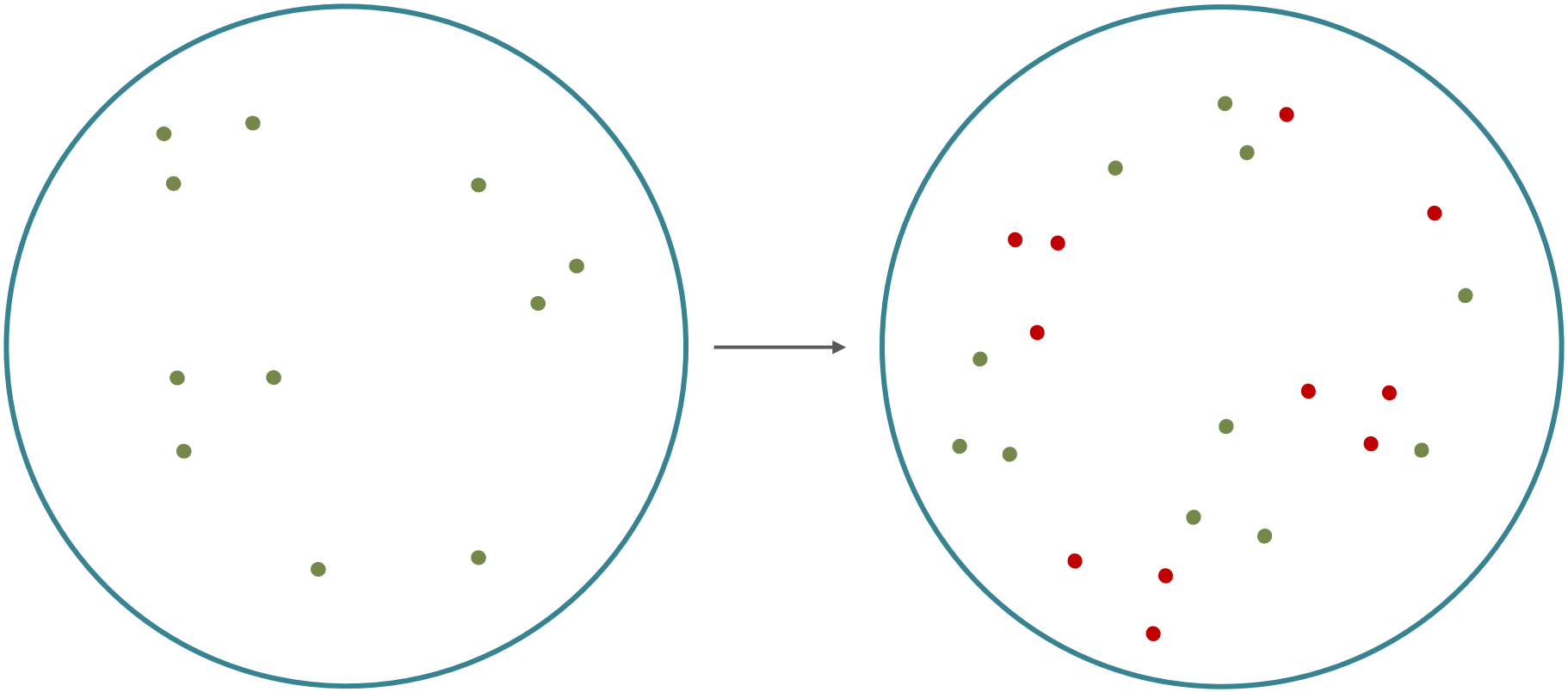
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



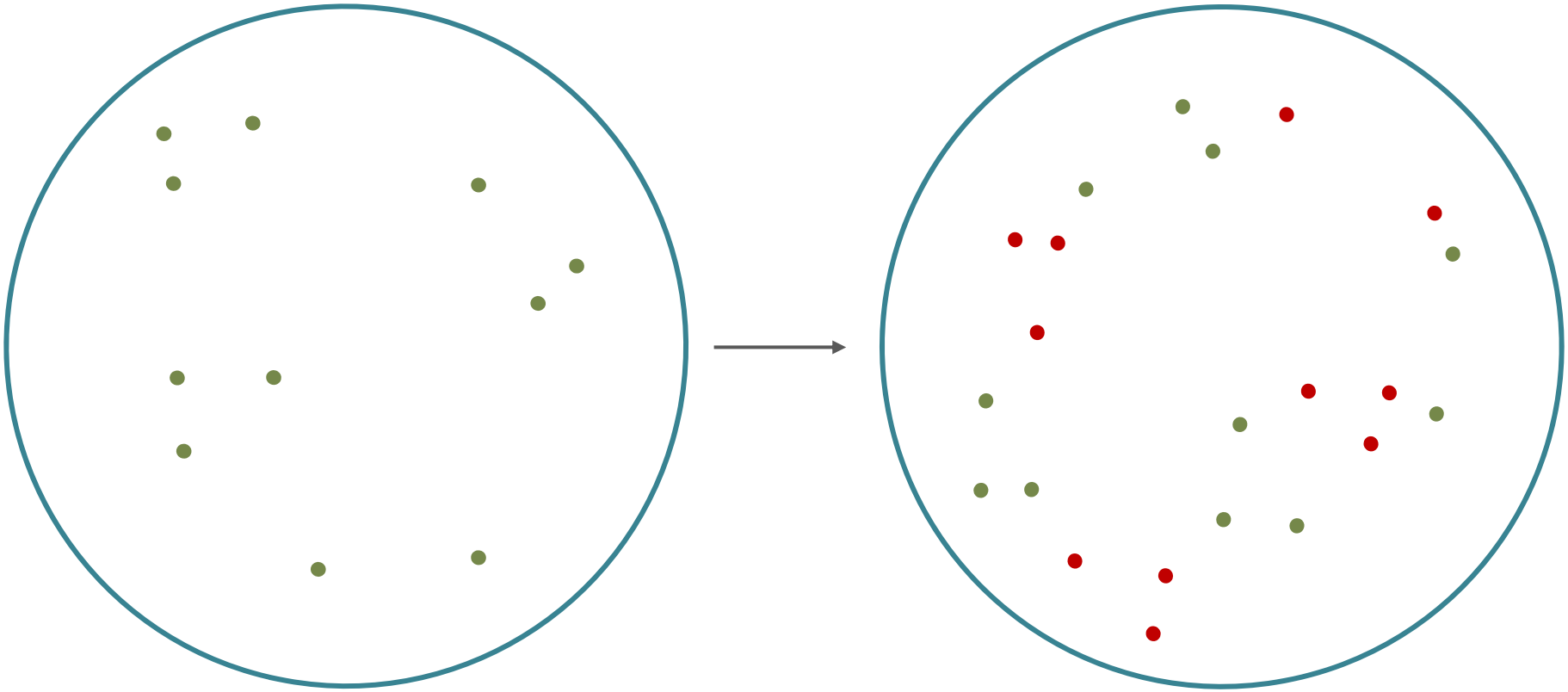
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



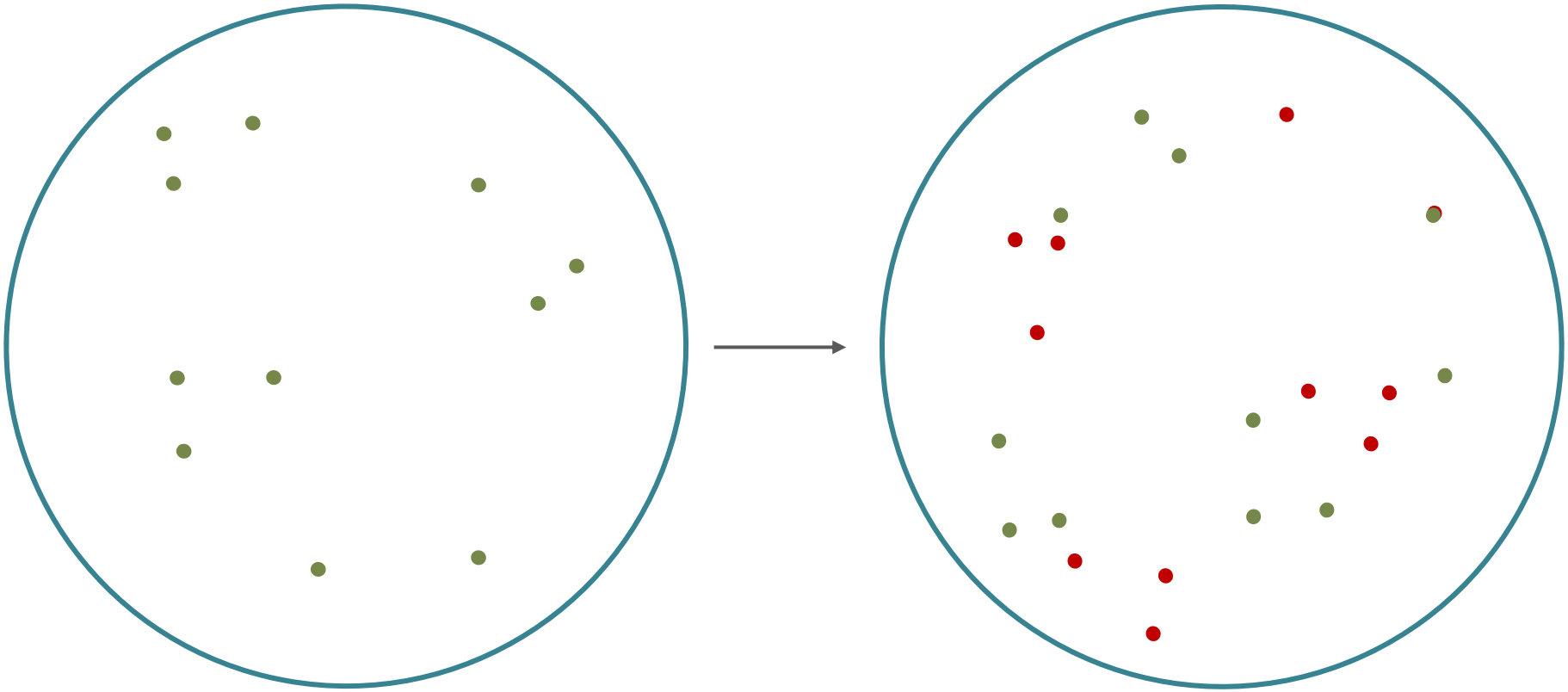
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



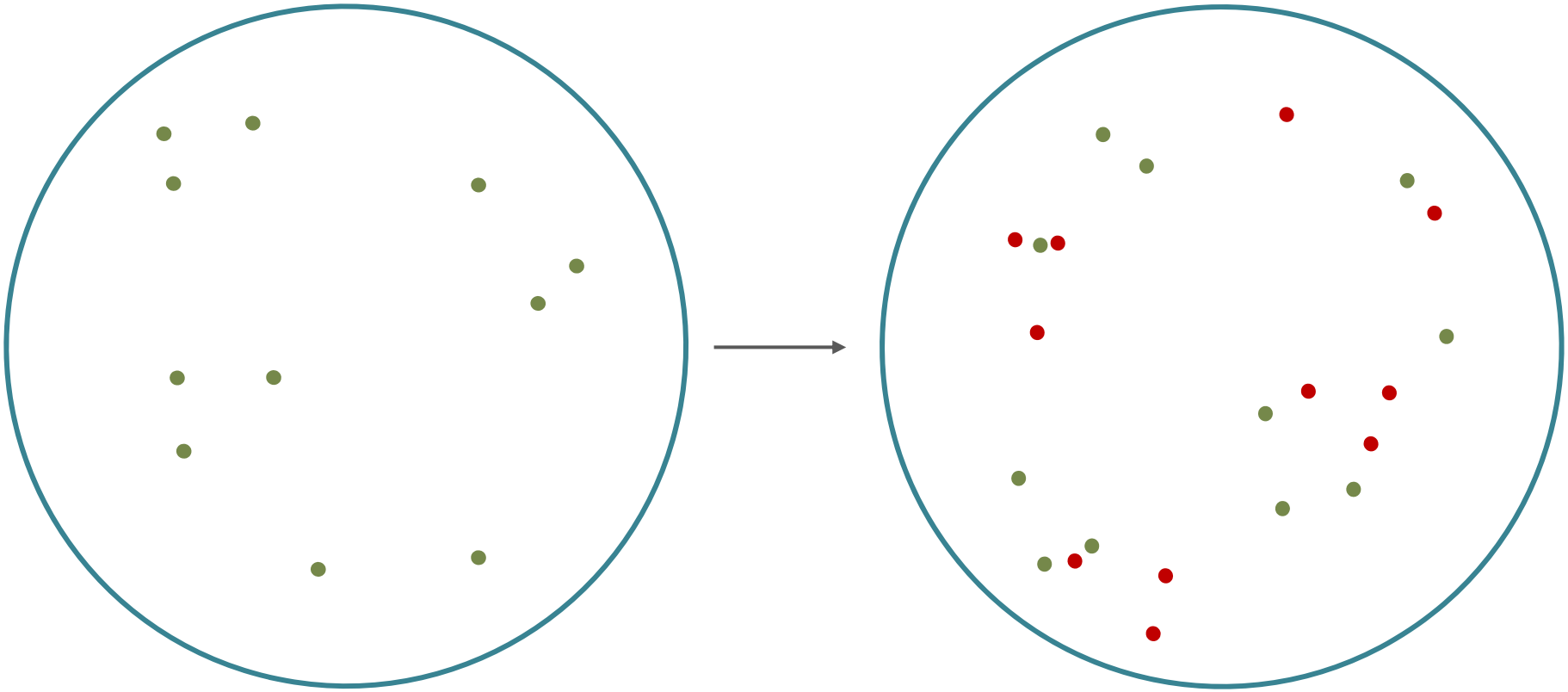
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



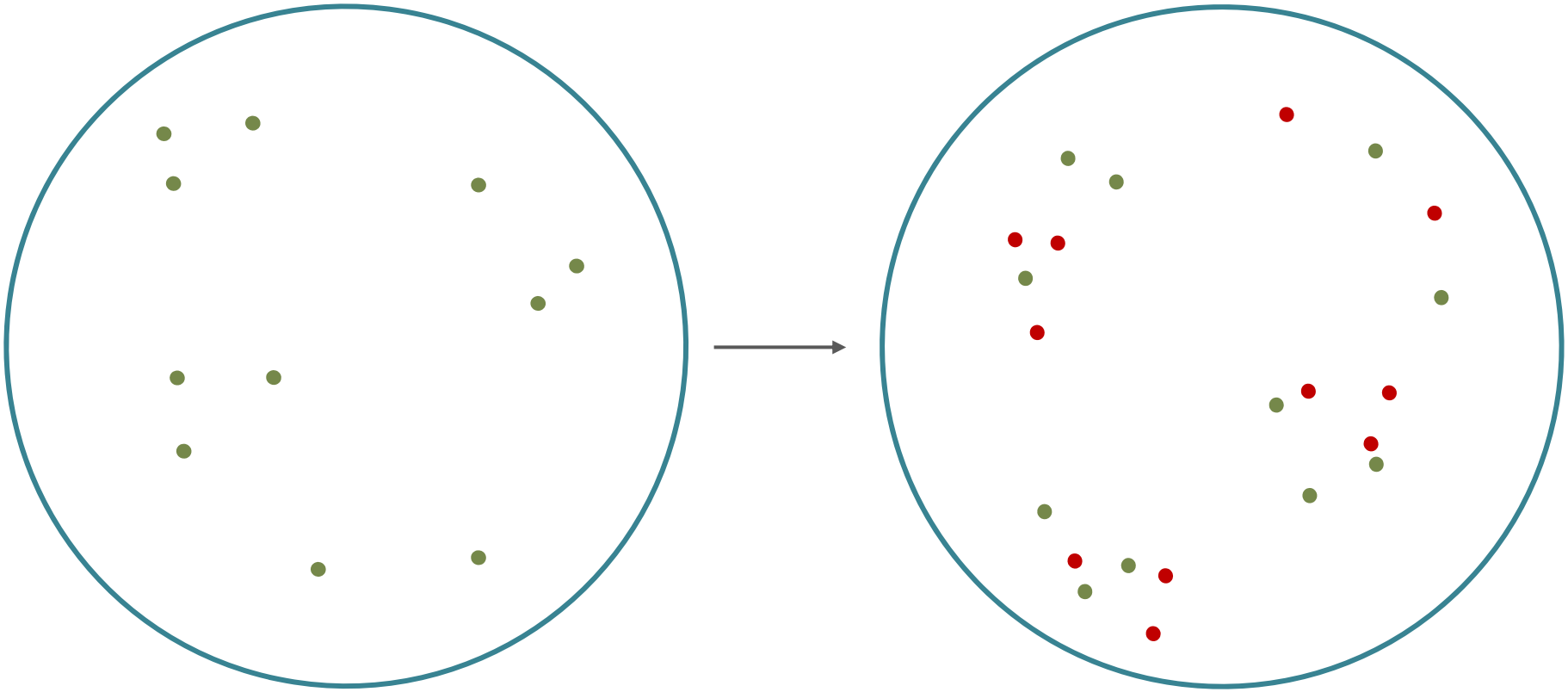
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



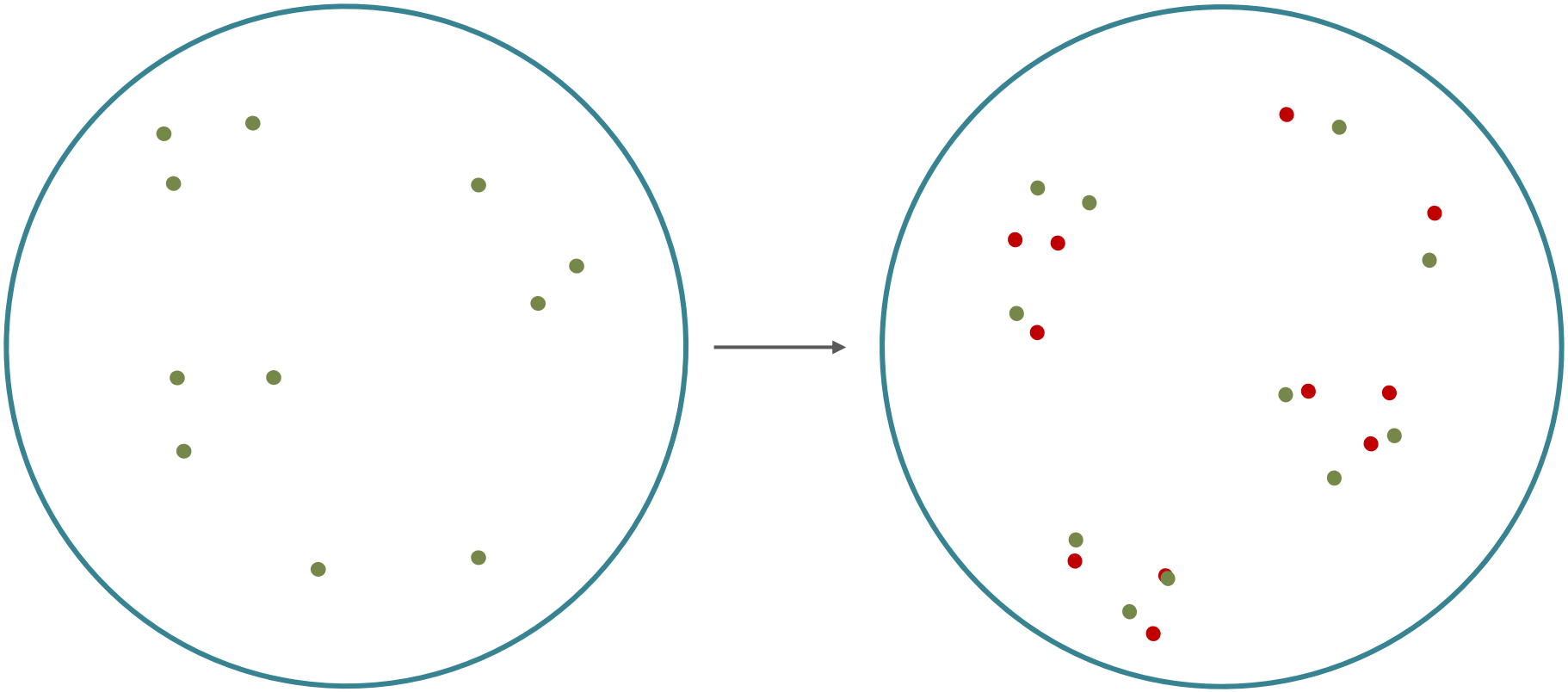
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



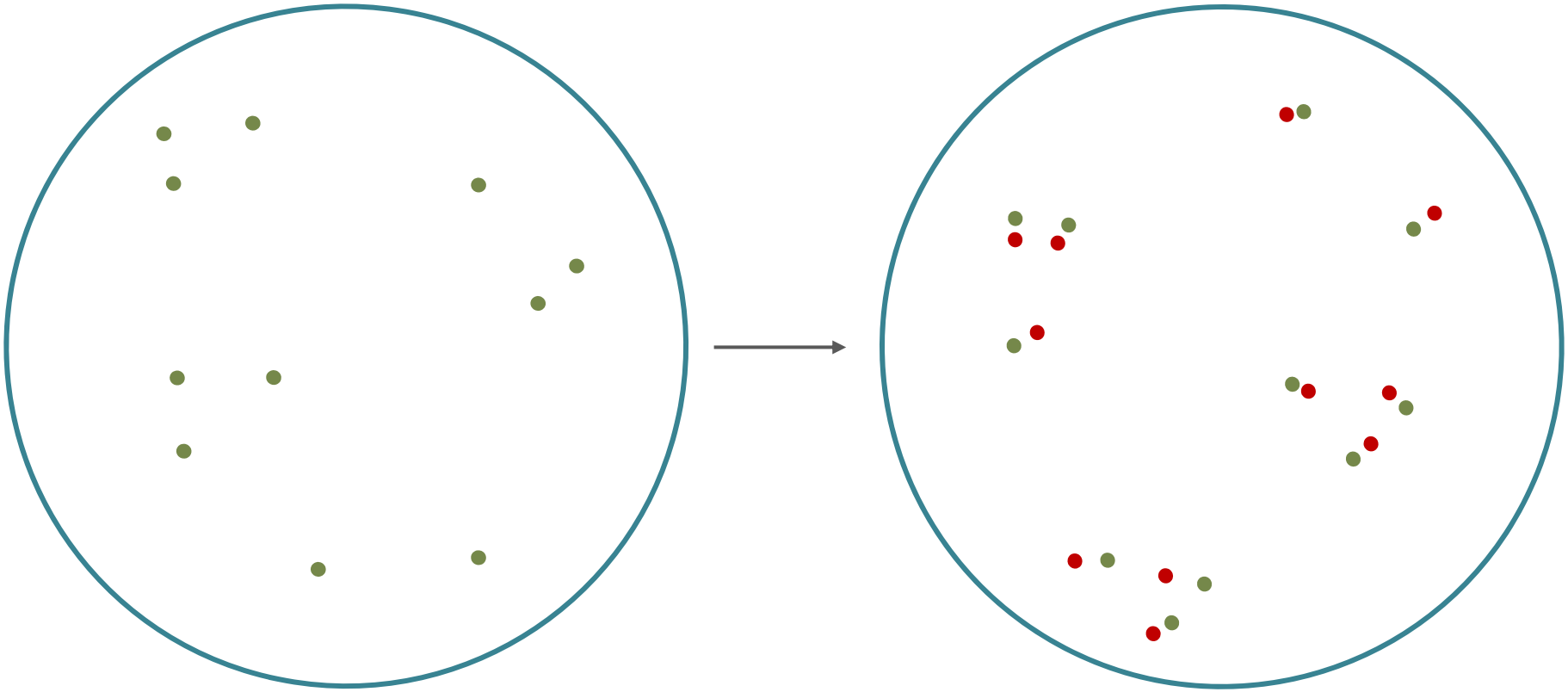
Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



Why does it work?

<u>Implicit objective</u>: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*}$ s.t. $WW^T = W^TW = I$



Why does it work?

<u>Implicit objective</u>: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t. } WW^T = W^TW = I$$

Independent from seed dictionary!

Why does it work?

<u>Implicit objective</u>: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t. } WW^T = W^TW = I$$

Independent from seed dictionary!

So why do we need a seed dictionary?

Why does it work?

<u>Implicit objective</u>: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t. } WW^T = W^TW = I$$

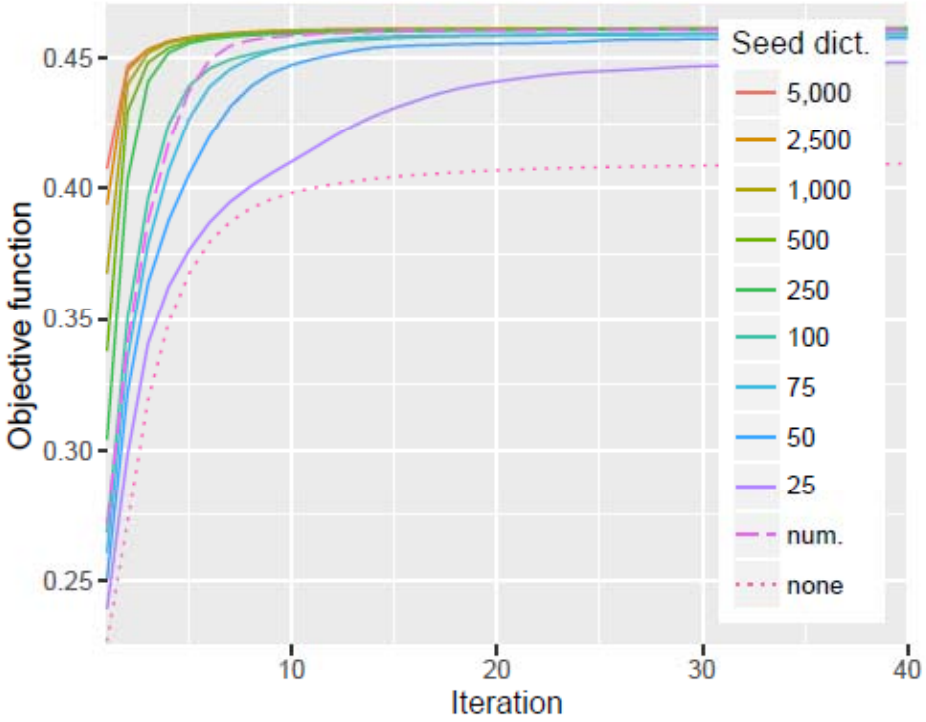
Independent from seed dictionary!

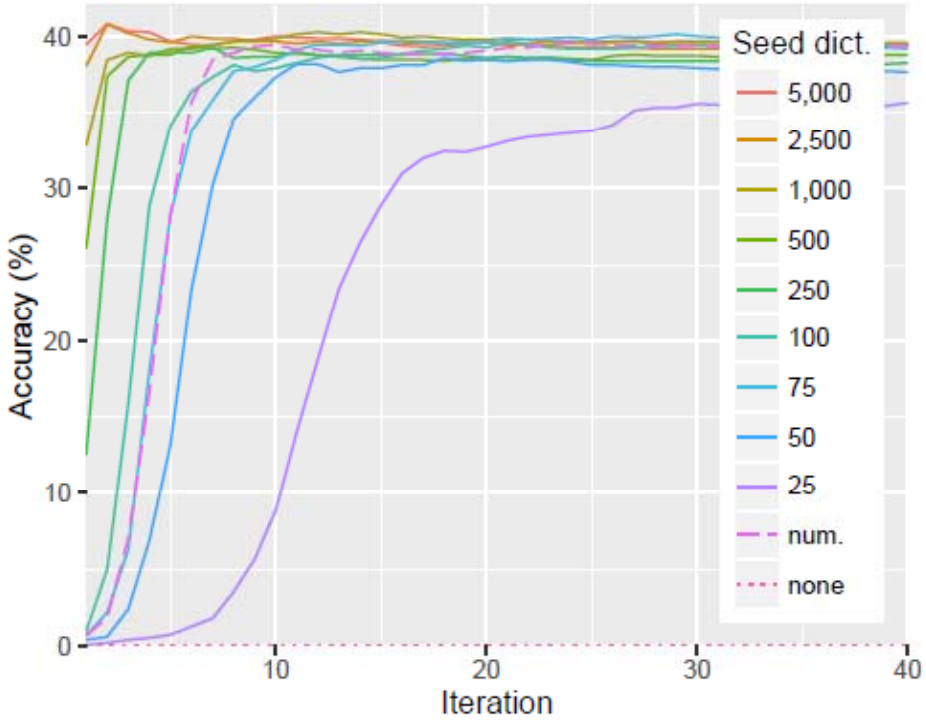
So why do we need a seed dictionary?

Avoid poor local optima!

Why does it work?

<u>Implicit objective</u>: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t. } WW^T = W^TW = I$$





Next steps

Is there a way we can avoid the seed dictionary?

Would an initial noisy initialization suffice?

Unsupervised experiments (ACL18)

Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other

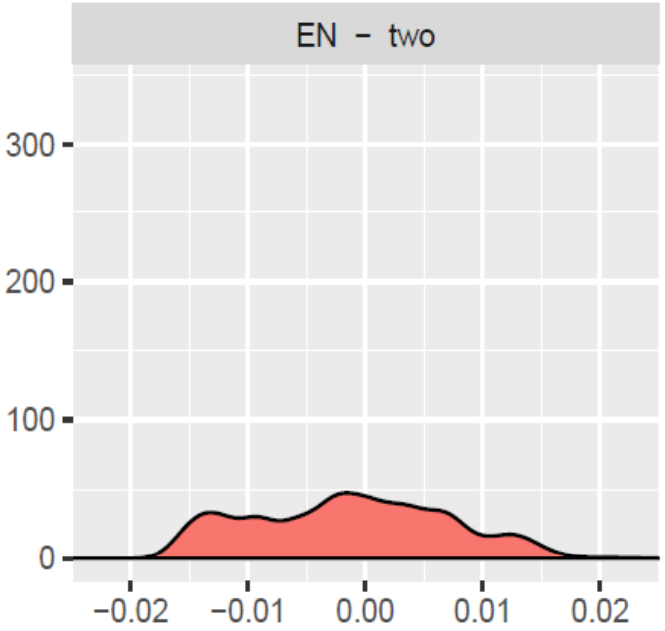
would have analoguous similarity histograms (isometry hyp.)

Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other

would have analoguous similarity histograms (isometry hyp.)

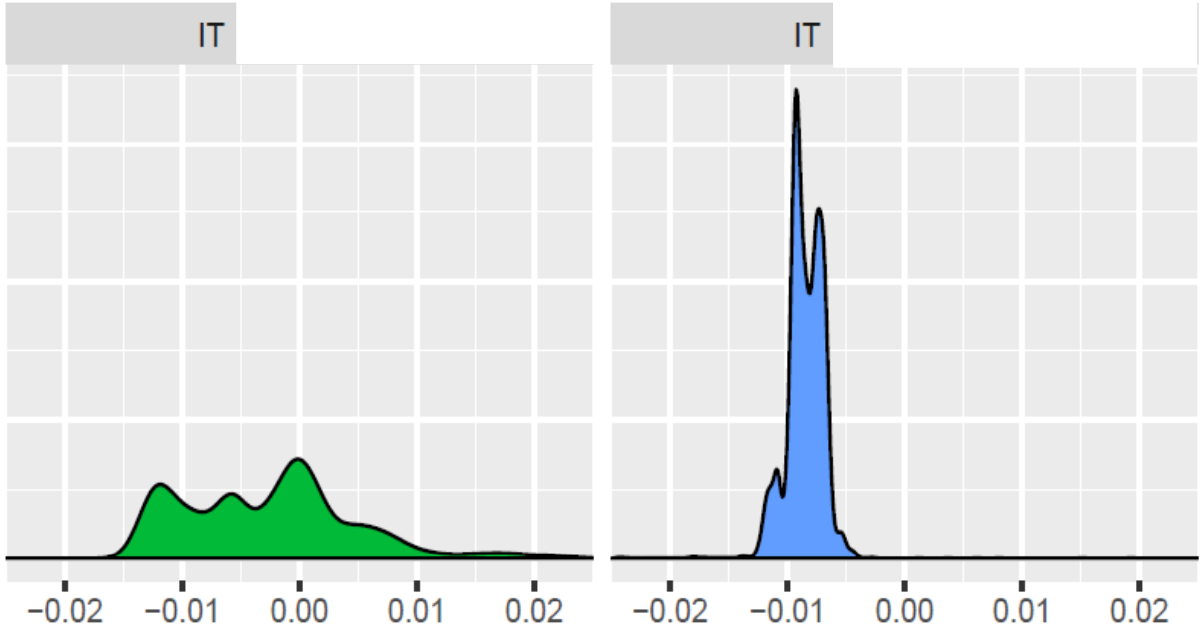


Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other

would have analoguous similarity histograms (isometry hyp.)

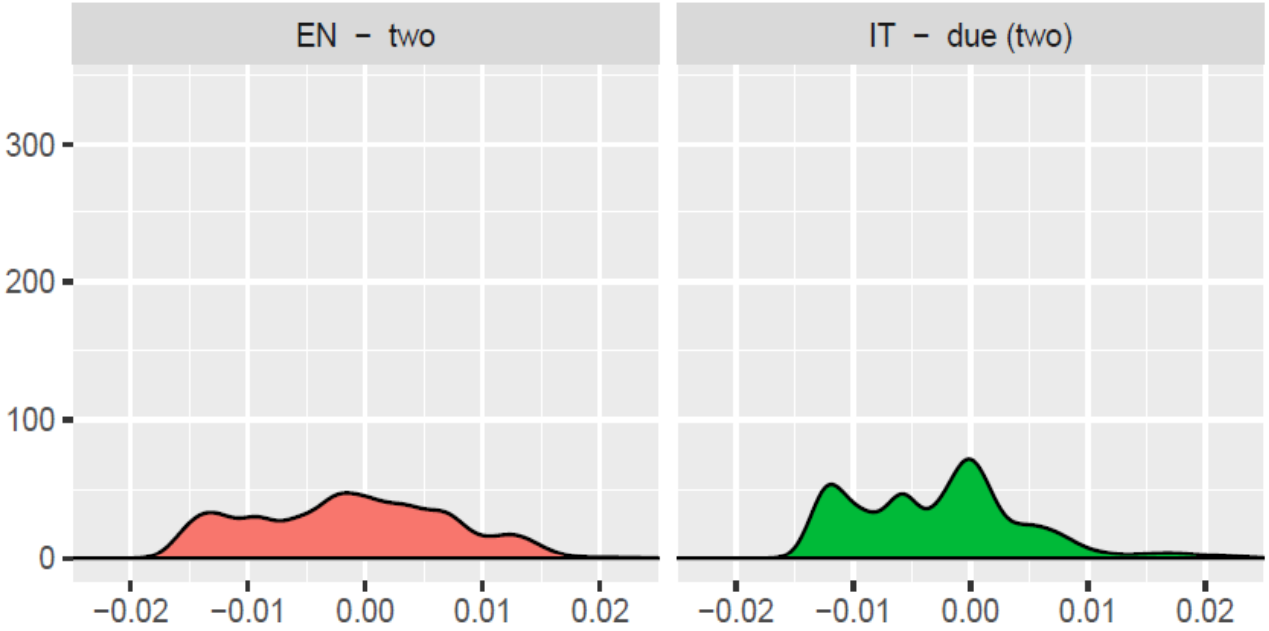


Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other

would have analoguous similarity histograms (isometry hyp.)



Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other

  would have analoguous similarity histograms (isometry hyp.)

- It works, but very weak: Accuracy 0.52%

Unsupervised experiments (ACL18)

Initial dictionary:

- Compute intra-language similarity
- Words which are translations of each other
  would have analoguous similarity histograms (isometry hyp.)

- It works, but very weak: Accuracy 0.52%

- For self-learning to work we added:
  1) Stochastic dictionary induction
  2) Frequency-based vocabulary cut-off
  3) Instead of inducing dictionary with nearest-neighbour
     use CSLS (Lample et al. 2018), due to hubness problem

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015)

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015)

Supervision	Method	EN-IT

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) , extended German, Finnish, Spanish

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	**30.62***	31.40*
	Smith et al. (2017)	**43.1**	**43.33†**	29.42†	**35.13†**
	Our method (AAAI18)				
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	**45.27**	**44.13**	**32.94**	**36.60**
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	**39.27**	**41.87***	**30.62***	31.40*
25 dict.	Our method (ACL17)				
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	**39.27**	**41.87***	**30.62***	31.40*
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	**45.27**	**44.13**	**32.94**	**36.60**
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-

None

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	**45.27**	**44.13**	**32.94**	**36.60**
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-
None	Zhang et al. (2017)				
	Conneau et al. (2018)				
	Our method (ACL18)				

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	45.27	44.13	**32.94**	36.60
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-
None	Zhang et al. (2017)	0.00	0.00	0.01	0.01
	Conneau et al. (2018)	13.55	42.15	0.38	21.23
	Our method (ACL18)	**48.13**	**48.19**	32.63	**37.33**

Conclusions

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Full reproducibility (including datasets):
  `https://github.com/artetxem/vecmap`

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Full reproducibility (including datasets):
  `https://github.com/artetxem/vecmap`

- Shows that languages share “semantic” structure to a large degree

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Matches results of supervised methods with no supervision
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Full reproducibility (including datasets):
  `https://github.com/artetxem/vecmap`

- Shows that languages share “semantic” structure to a large degree
- No need of magic

Conclusions

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. **Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations**. In *AAAI-2018.*

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. **Learning bilingual word embeddings with (almost) no bilingual data**. In *ACL-2017*.

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. **A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings**. In *ACL-2018*.

Outline

- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Introduction to bilingual embedding mappings*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Conclusions*

Introduction to (supervised) NMT

Introduction to (supervised) NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

Introduction to (supervised) NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Recurrent Neural Nets

  return all hidden states, encoding input $x_1 \ldots x_n$

Introduction to (supervised) NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Recurrent Neural Nets

  return all hidden states, encoding input $x_1 \ldots x_n$

- Train a **decoder** based on Recurrent Neural Nets

  - based on hidden states and last word in translation y_{i-1}
  - plus an **attention** mechanism
  - classifier guesses next word y_i

Introduction to (supervised) NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Recurrent Neural Nets

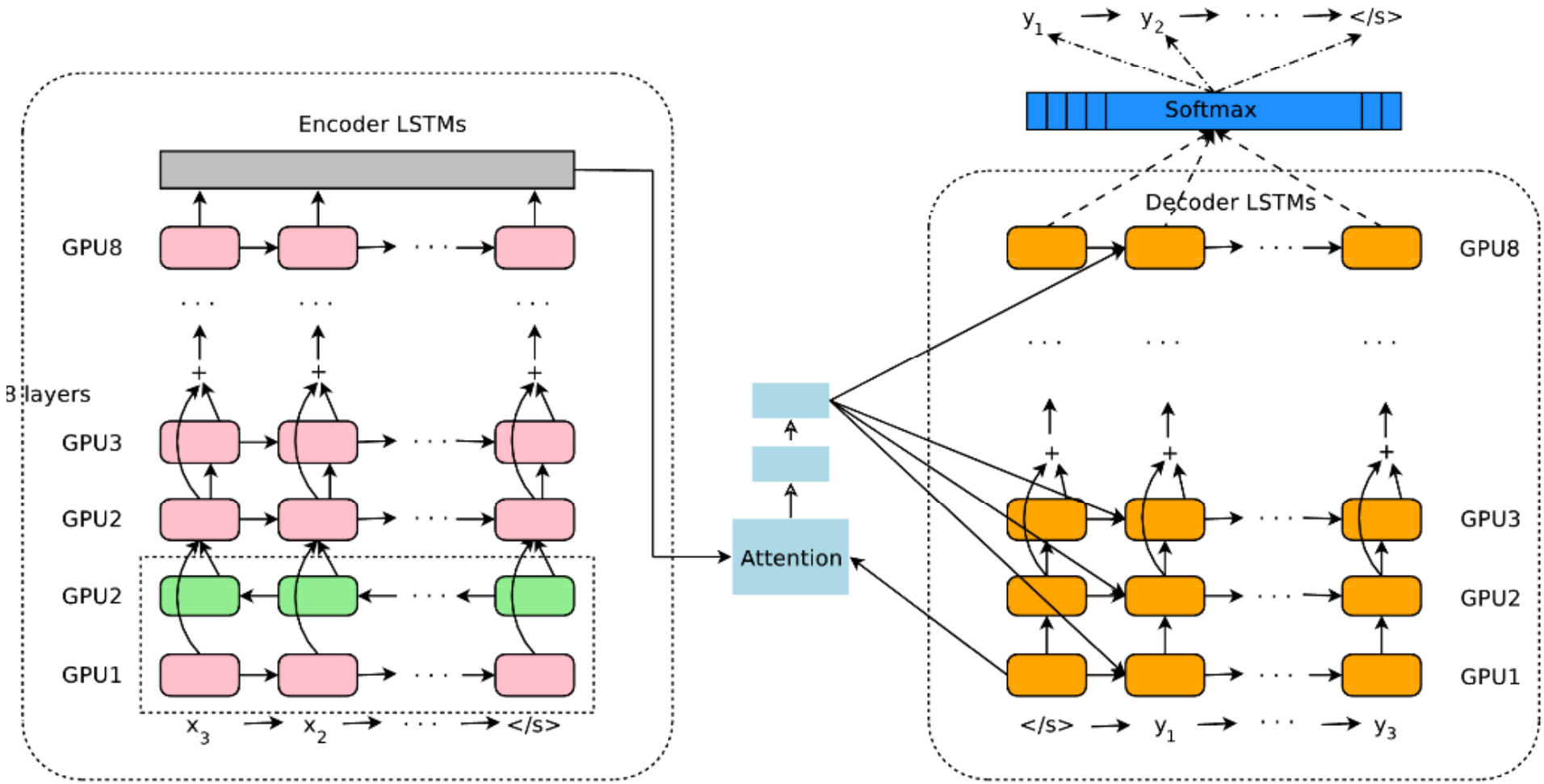
  return all hidden states, encoding input $x_1 \ldots x_n$

- Train a **decoder** based on Recurrent Neural Nets

  - based on hidden states and last word in translation y_{i-1}
  - plus an **attention** mechanism
  - classifier guesses next word y_i

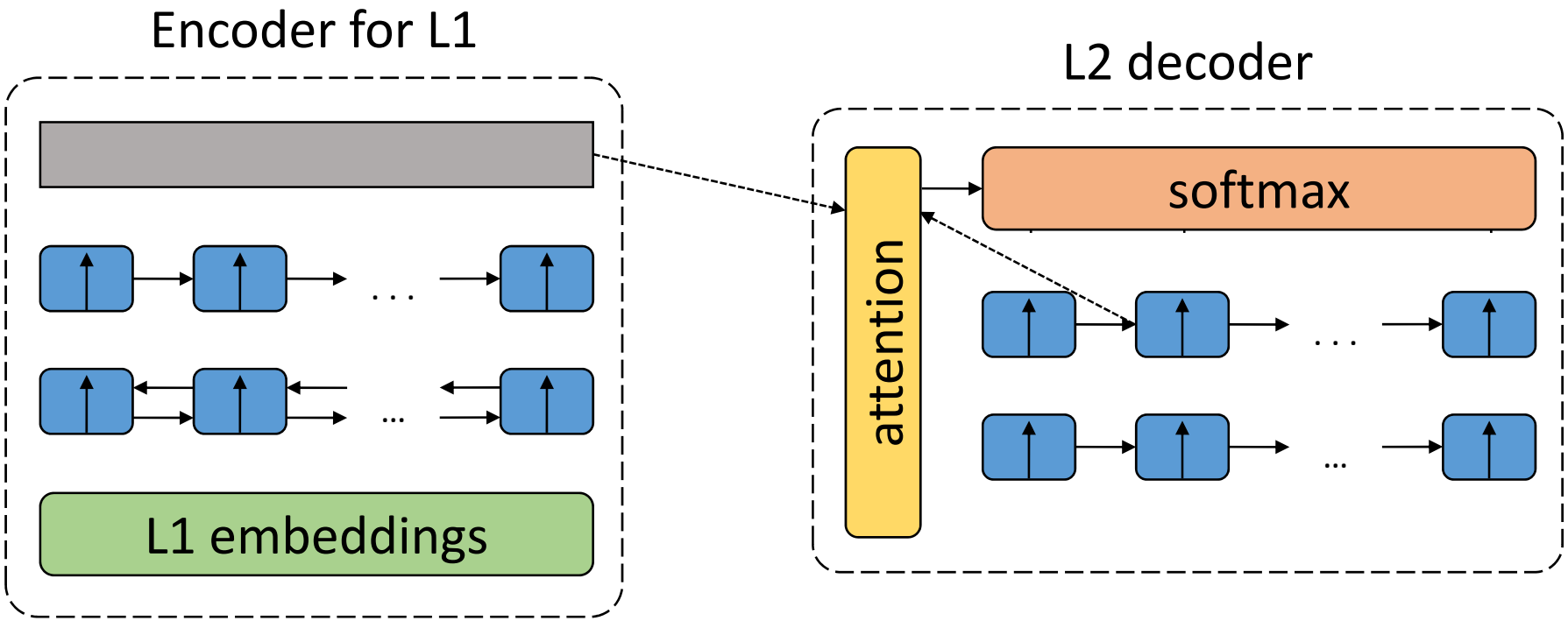
End-to-end training

Introduction to (supervised) NMT



Source: Wu et al. 2016 (~ 30 authors – Also known as Google NMT)

Introduction to (supervised) NMT



Unsupervised neural machine translation

- Now that we **can represent words in two languages in the same embeddings space** without bilingual dictionaries…

  … what can we do?

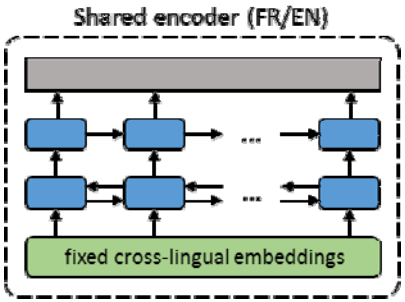
Unsupervised neural machine translation

- Now that we **can represent words in two languages in the same embeddings space** without bilingual dictionaries…

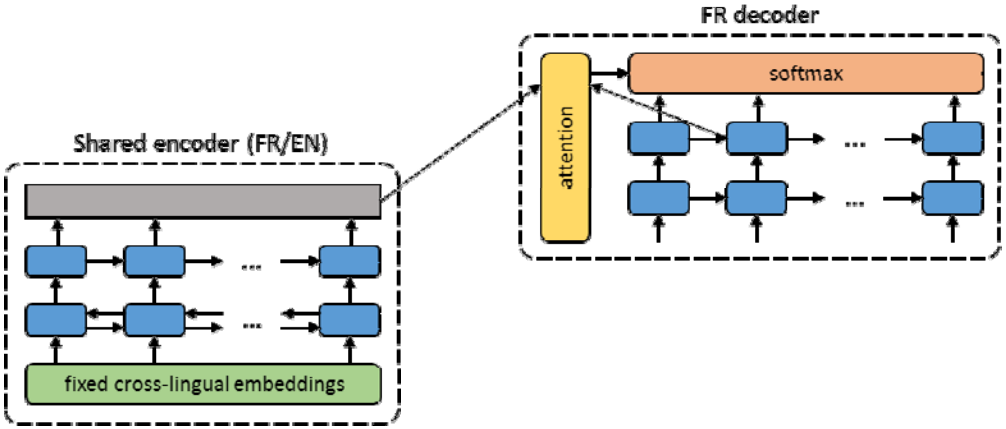
  … what can we do?

- We change the **architecture of the NMT system**:
  - Handle both directions together (L1 -> L2, L2 -> L1)
  - Shared encoder for the two languages (E)
  - Two decoders for each language (D1, D2)
  - Fixed embeddings

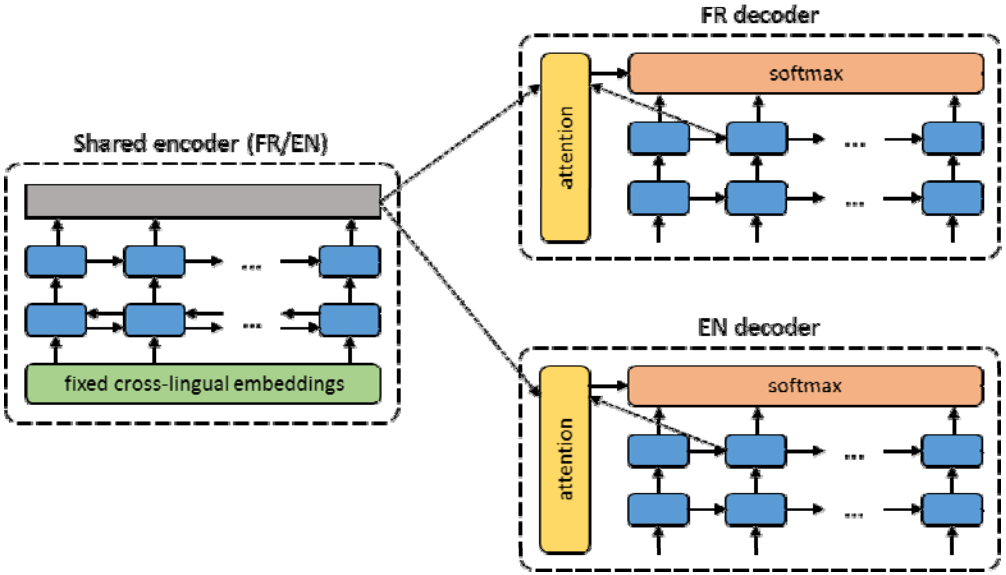
Unsupervised neural machine translation



Unsupervised neural machine translation



Unsupervised neural machine translation



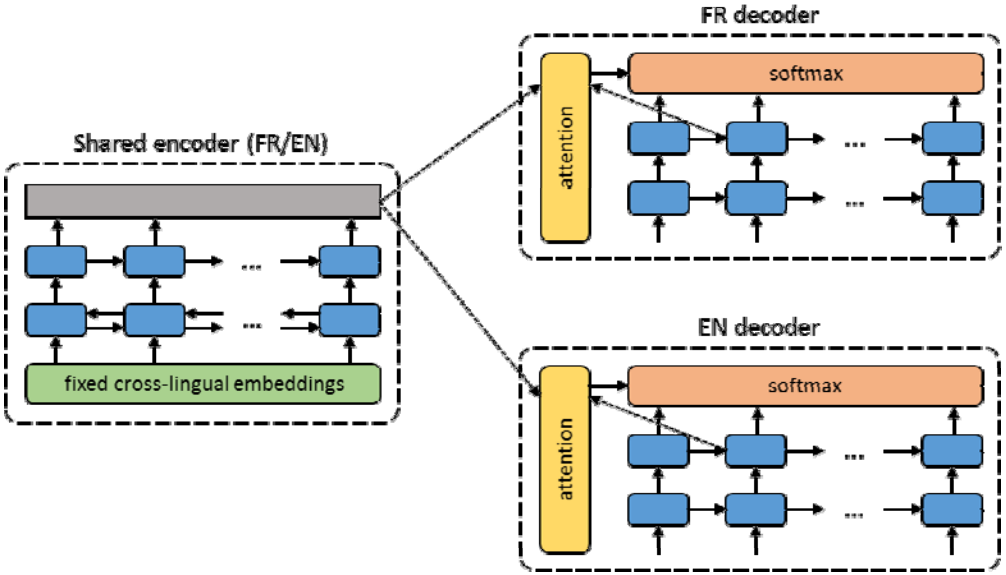
Unsupervised neural machine translation

- We change the **architecture of the NMT system**:
  - Handle both directions together (L1 -> L2, L2 -> L1)
  - Shared encoder for the two languages (E)
  - Two decoders for each language (D1, D2)
  - Fixed embeddings

- We change the **training regime, mixing mini-batches**:
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2

Unsupervised neural machine translation

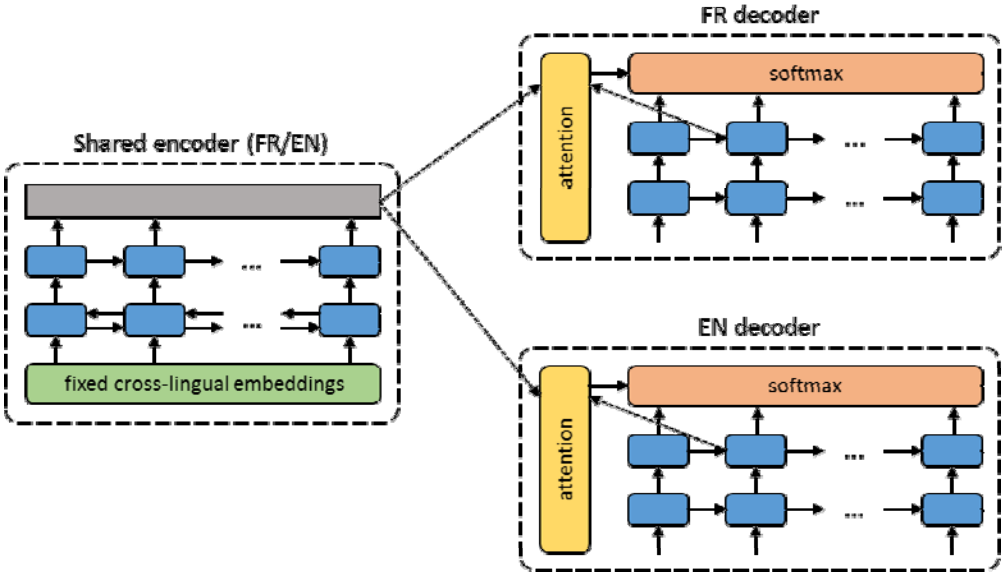
Training



Unsupervised neural machine translation

Training

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

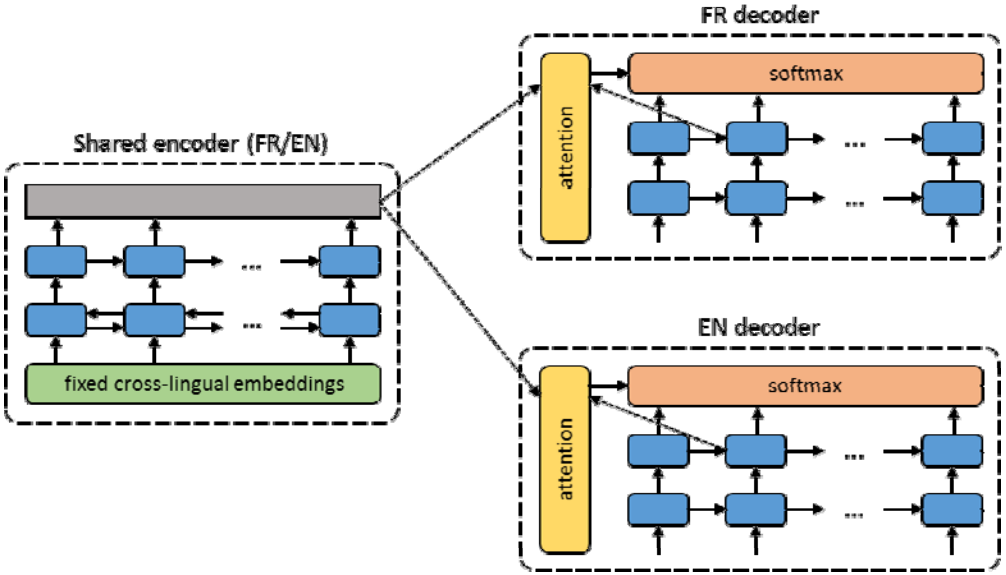


Unsupervised neural machine translation

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

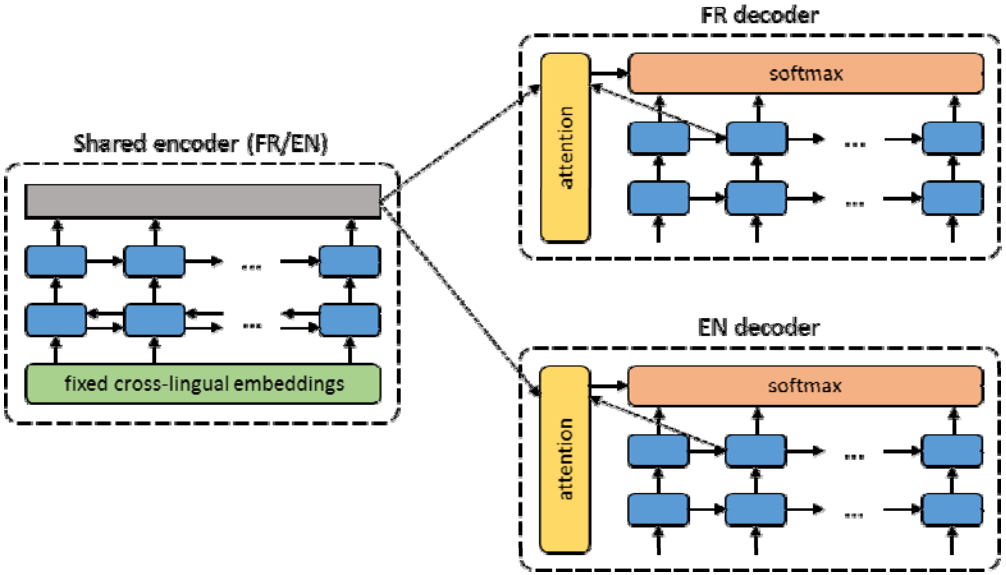


Unsupervised neural machine translation

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

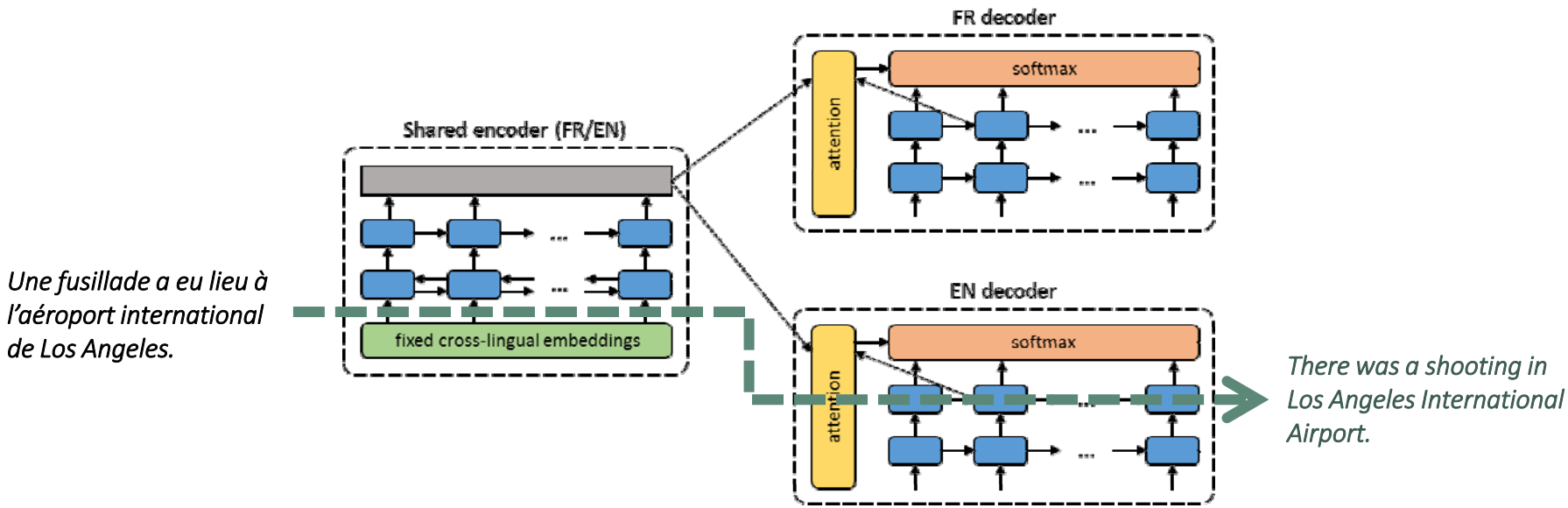


There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

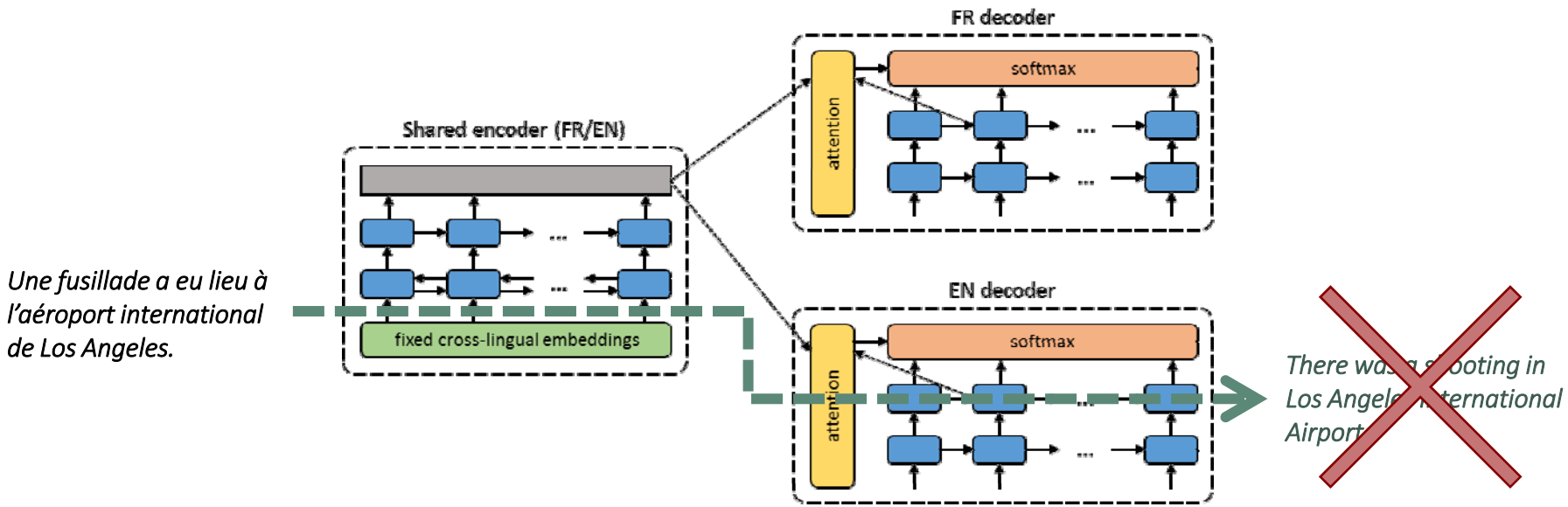
- Supervised



Unsupervised neural machine translation

Training

- Supervised

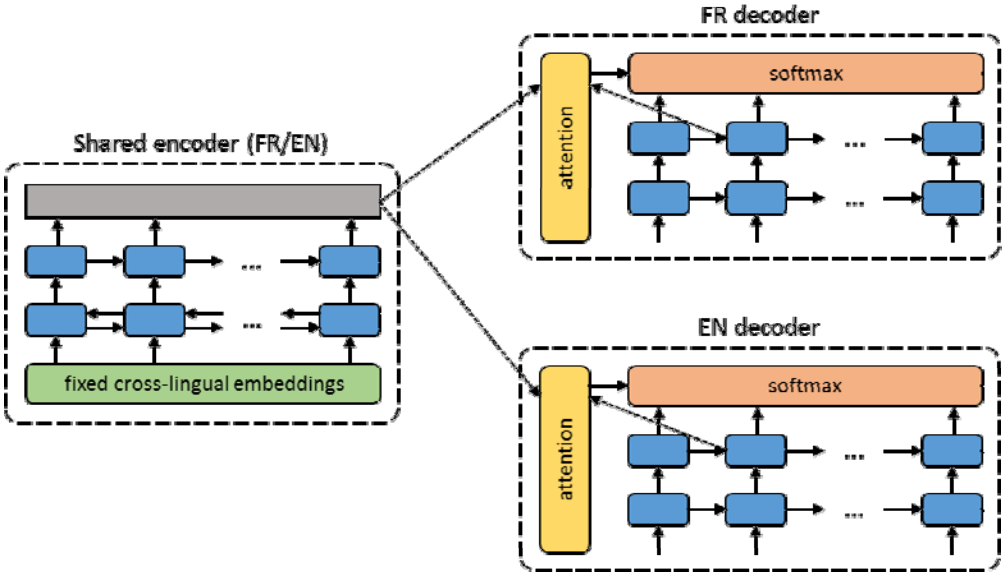


Unsupervised neural machine translation

Training

- ~~Supervised~~

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

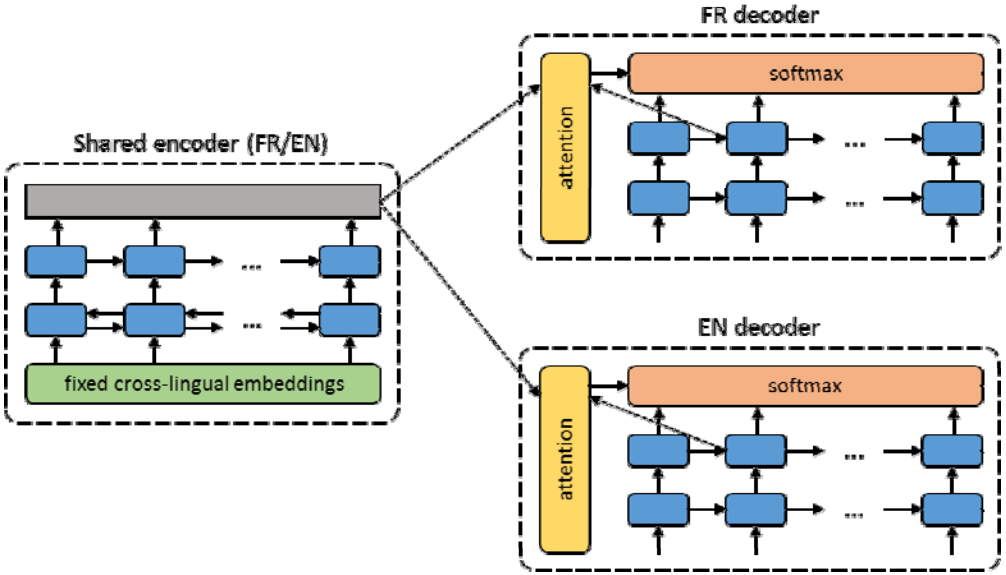


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Autoencoder

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

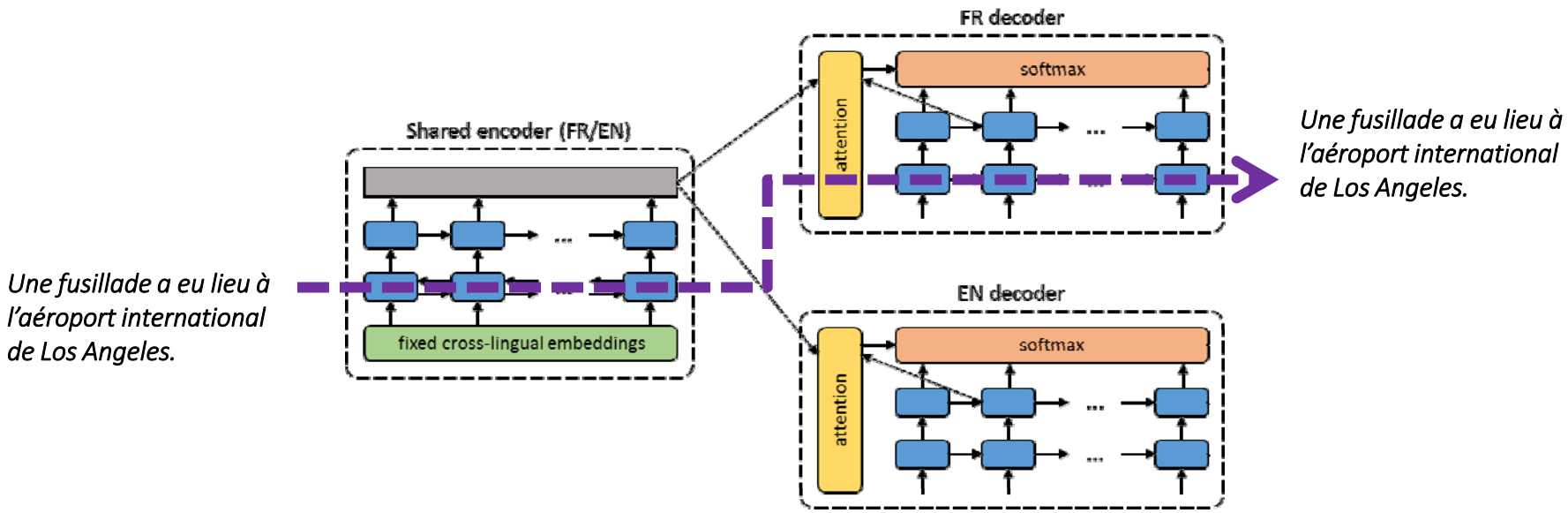


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Autoencoder

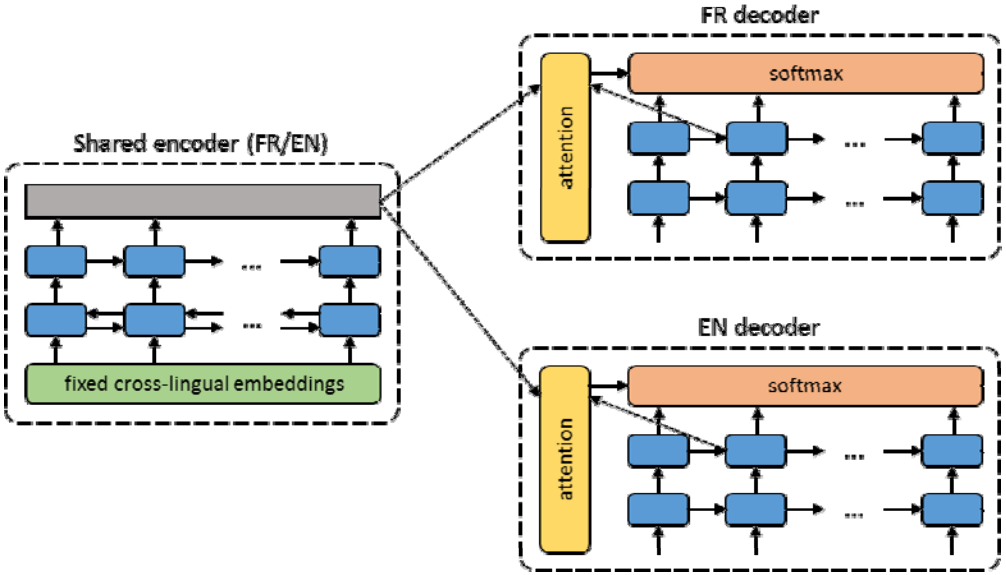


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

Une lieu fusillade a eu à l'aéroport de Los international Angeles.

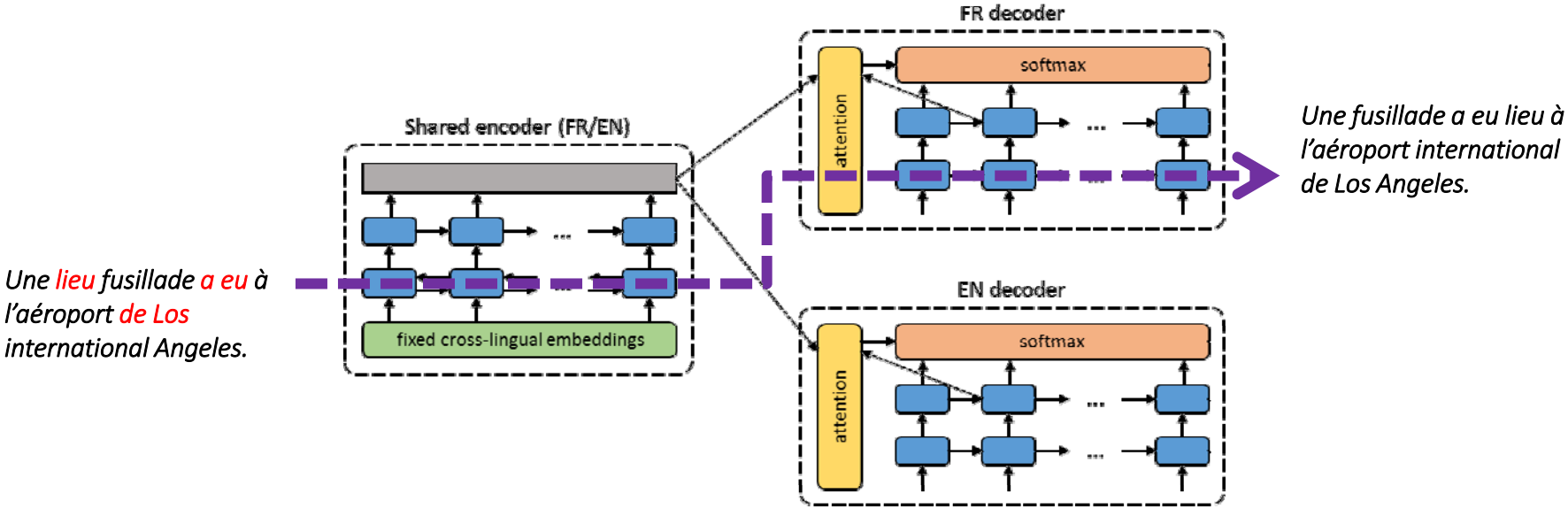


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

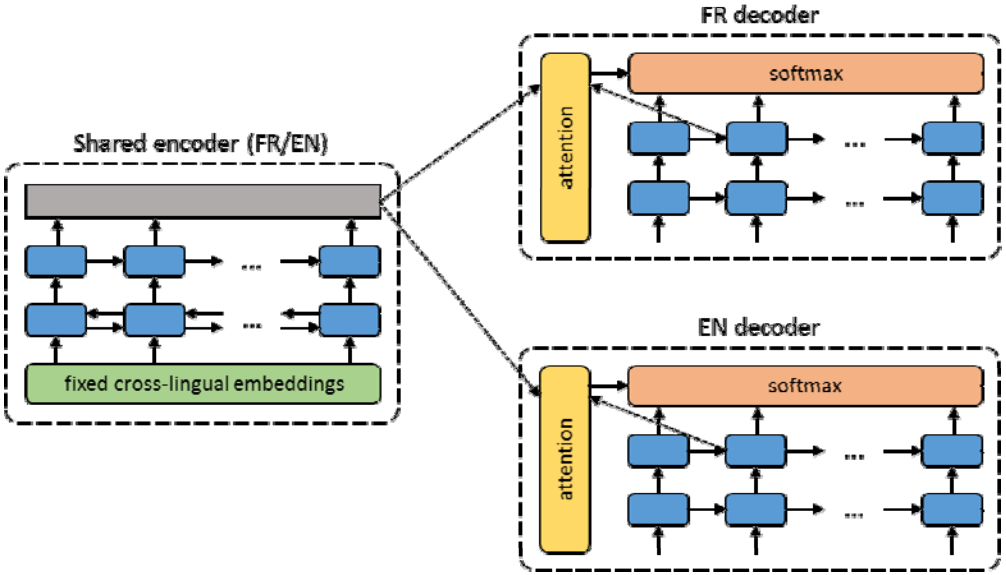


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

There a shooting was in Airport Los Angeles International.



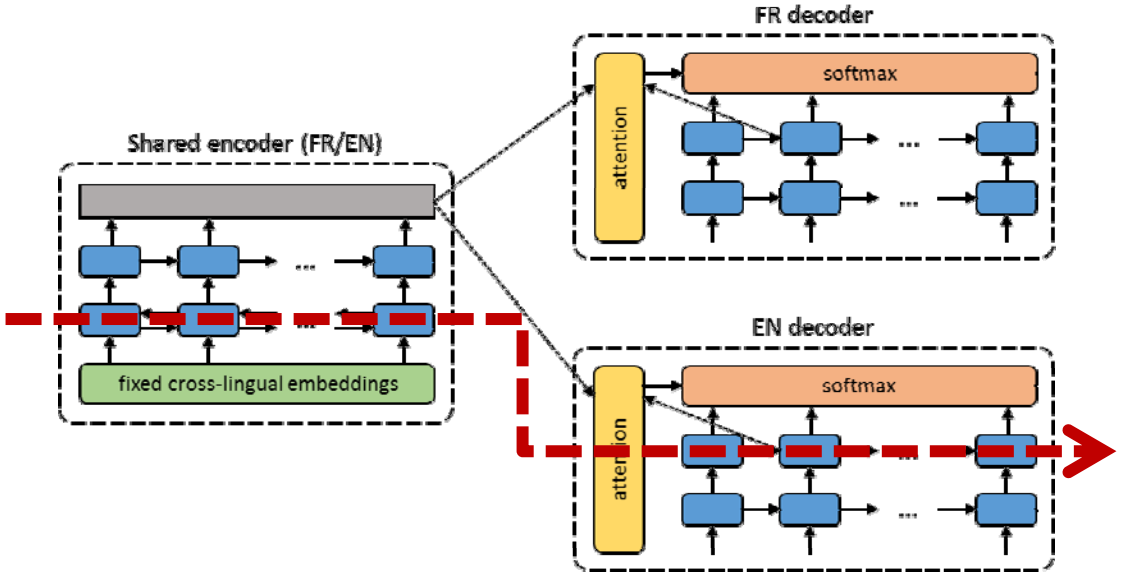
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

There a shooting was in Airport Los Angeles International.



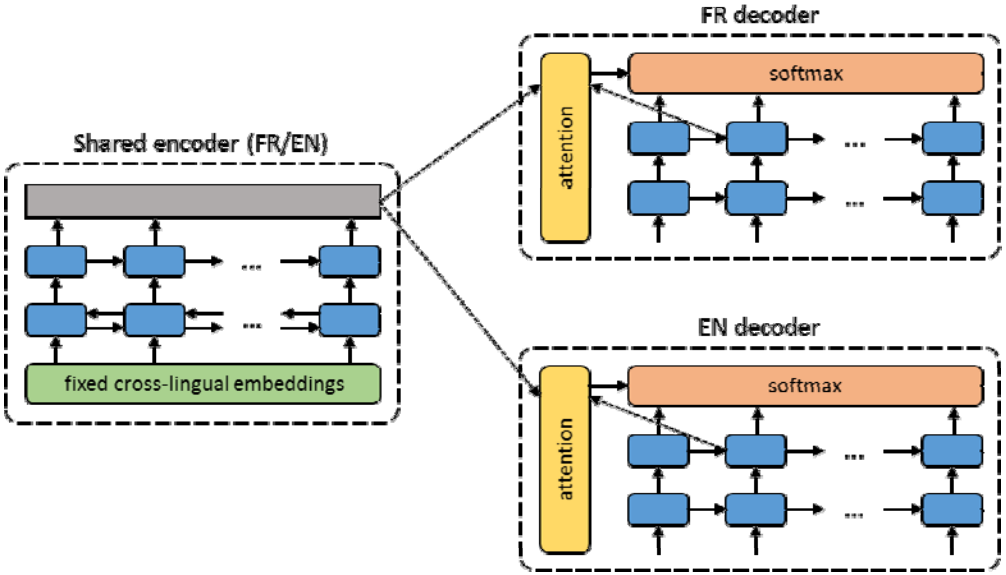
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

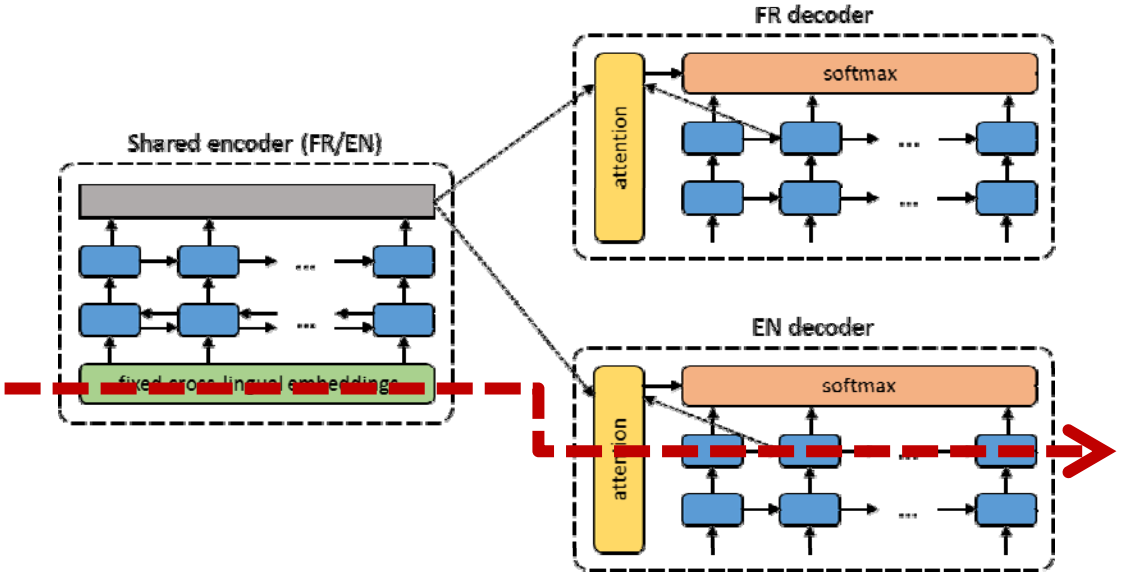


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

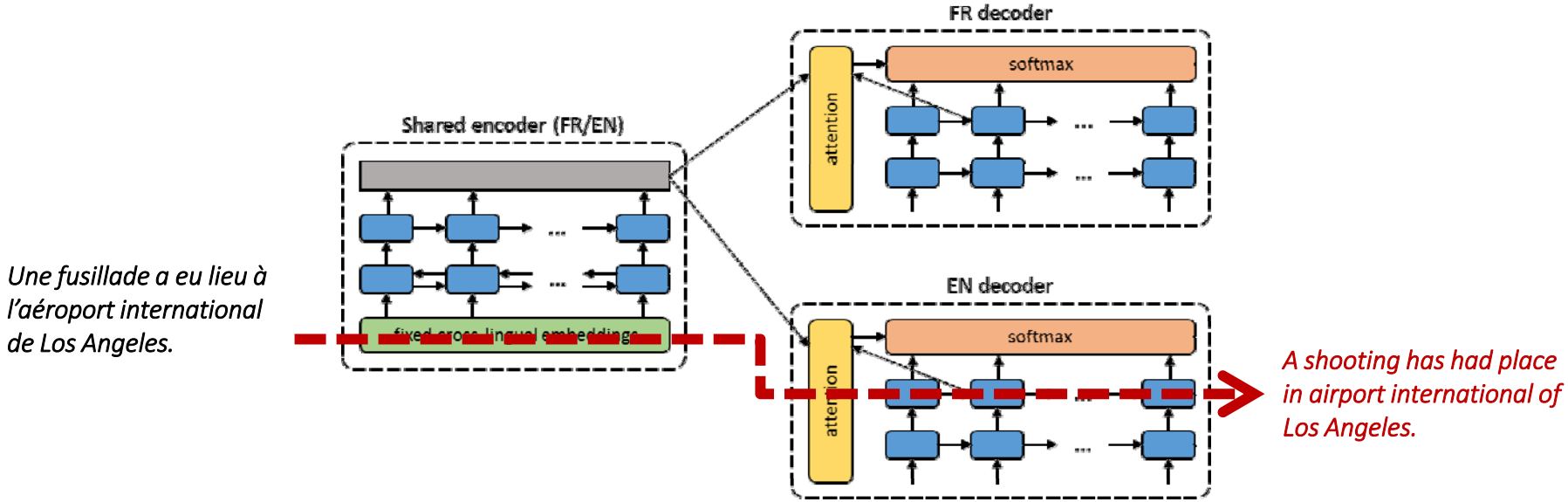
Une fusillade a eu lieu à l'aéroport international de Los Angeles.



Unsupervised neural machine translation

Training

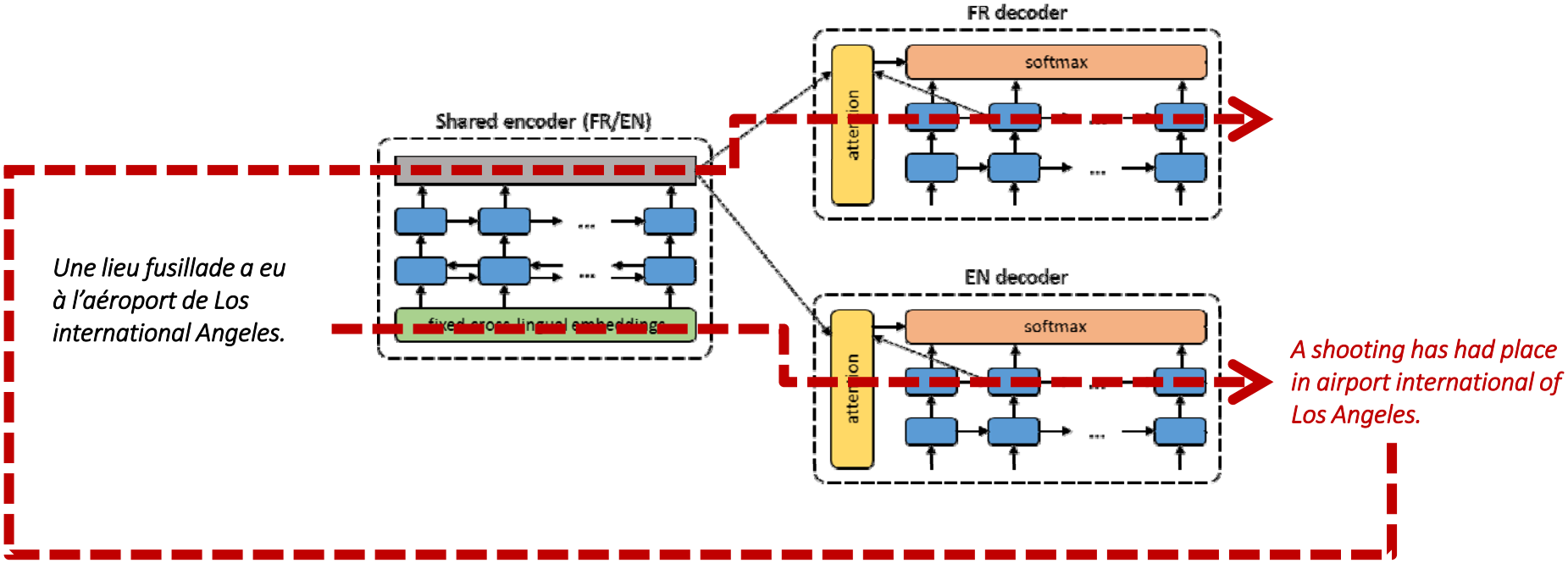
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

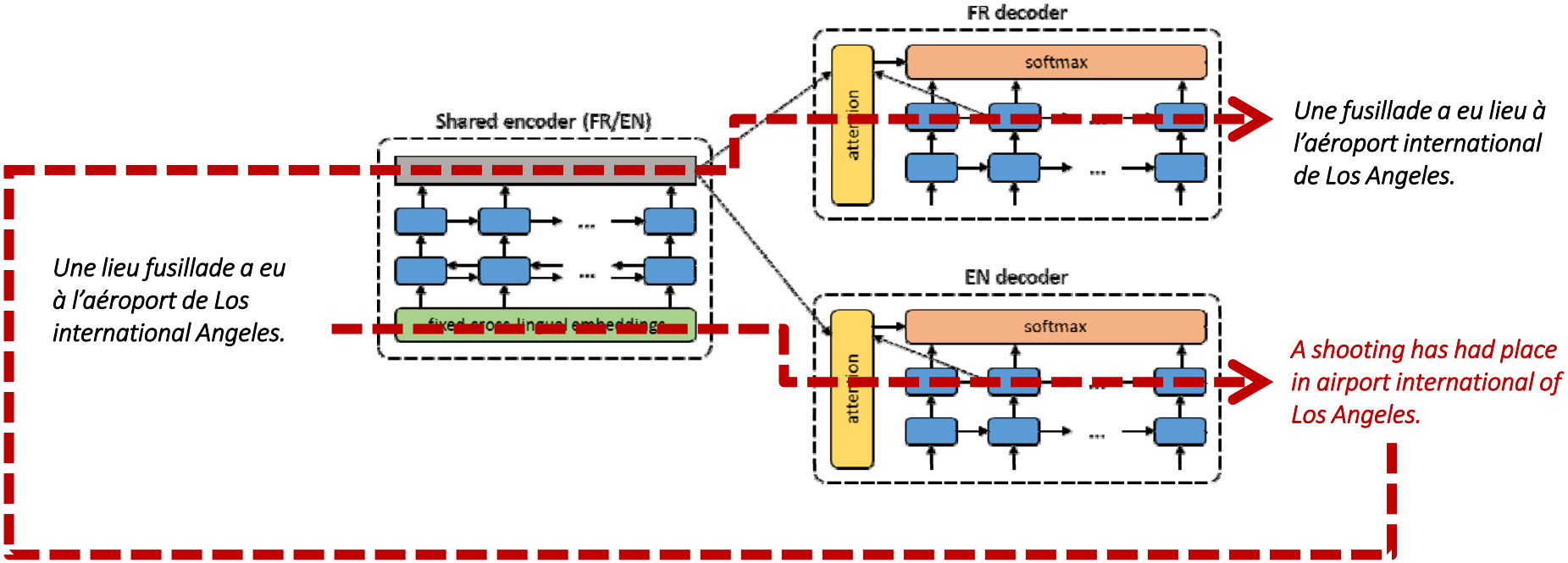
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

- We change the **architecture of the NMT system**:
  - Handle both directions together (L1 -> L2, L2 -> L1)
  - Shared encoder for the two languages (E)
  - Two decoders for each language (D1, D2)
  - Fixed embeddings

- We change the **training regime, mixing mini-batches**:
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

	FR-EN	EN-FR	DE-EN	EN-DE

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised				

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
Unsupervised					

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55

It works!

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55

Semi-supervised

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised	Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised	Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised	Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised (upperbound)					

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised					
Supervised (upperbound)	Comparable NMT	20.48	19.89	15.04	11.05

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised					
Supervised (upperbound)	Comparable NMT	20.48	19.89	15.04	11.05
	GNMT (Wu et al., 2016)	-	38.95	-	24.61

Unsupervised neural machine translation

Source	Translation

Unsupervised neural machine translation

Source	Translation
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	A shooting occurred at Los Angeles International Airport.

Unsupervised neural machine translation

Source	Translation
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beacoup de speculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-operation ciblée.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.

Unsupervised neural machine translation

Source	Translation
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beacoup de speculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-operation ciblée.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.
Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.	The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.

Unsupervised neural machine translation

Source	Translation
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beacoup de speculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-operation ciblée.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.
Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.	The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.
À l'exception de l'opéra, la province reste le parent pauvre de la culture en France.	At an exception, opera remains of the state remains the poorest parent culture.

Why does it work?

Why does it work?

Early to say… but intuition:

Why does it work?

Early to say… but intuition:

- Mapped embedding space provides information for k-best possible translations
- Encoder-decoder figures out how to best “combine” them

Why does it work?

Early to say… but intuition:

- Mapped embedding space provides information for k-best possible translations
- Encoder-decoder figures out how to best “combine” them

- No need of magic

Conclusions

- New research area – unsupervised Machine Translation

  The main Machine Translation competition (WMT18) has now an **unsupervised track**

- New papers are coming out, reporting 25 BLEU

- Code for replicability
  `https://github.com/artetxem/undreamt`

- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. **Unsupervised neural machine translation**. In *ICLR-2018*.

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**
- **Cross-lingual unsupervised mappings** enabled breakthroughs in
  - Bilingual dictionary induction
  - Unsupervised machine translation
  - Also (Conneau et al. 2018; Lample et al. 2018)

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**
- **Cross-lingual unsupervised mappings** enabled breakthroughs in
  - Bilingual dictionary induction
  - Unsupervised machine translation
  - Also (Conneau et al. 2018; Lample et al. 2018)
- Unexplored area in its **infancy**
  - Potential for **MT in low resource languages and domains**
  - Potential for **transforming the NLP landscape**
    - From monolingual NLP (e.g. English) to multilingual tools
    - Universal sentence representations

Thank you!

@eagirre
http://ixa2.si.ehu.eus/eneko
https://github.com/artetxem/vecmap
https://github.com/artetxem/undreamt