UNIVERSITY OF
TORONTO

ORDER-EMBEDDINGS OF IMAGES AND LANGUAGE

Ivan Vendrov, Ryan Kiros, Sanja Fidler, Raquel Urtasun

Semantic Image Search

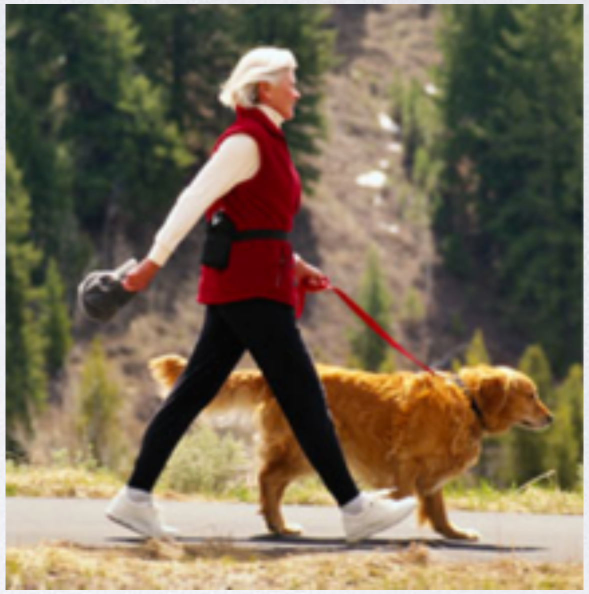
- Given a database of images and a natural language query, identify which images it accurately describes

Semantic Image Search

- Given a database of images and a natural language query, identify which images it accurately describes

Semantic Image Search

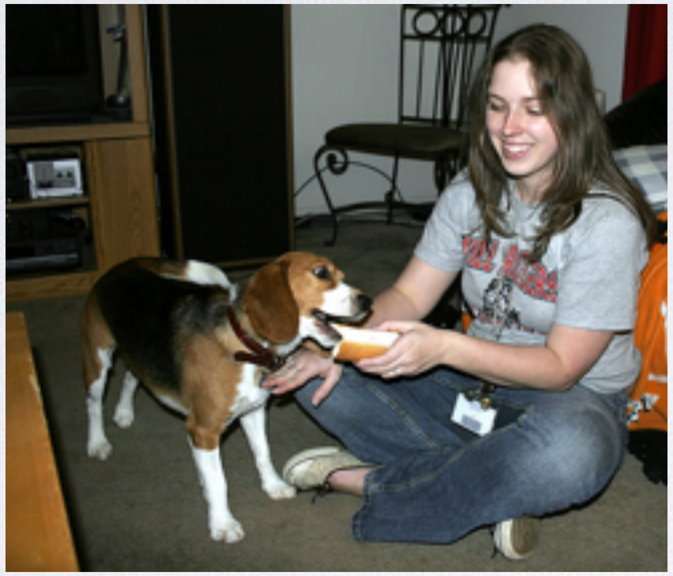
- Given a database of images and a natural language query, identify which images it accurately describes

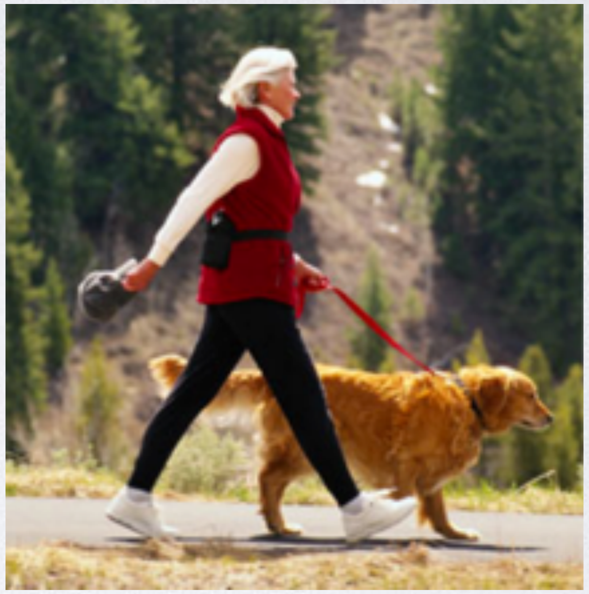


“a woman walking her dog in a park”

Semantic Image Search

- Given a database of images and a natural language query, identify which images it accurately describes

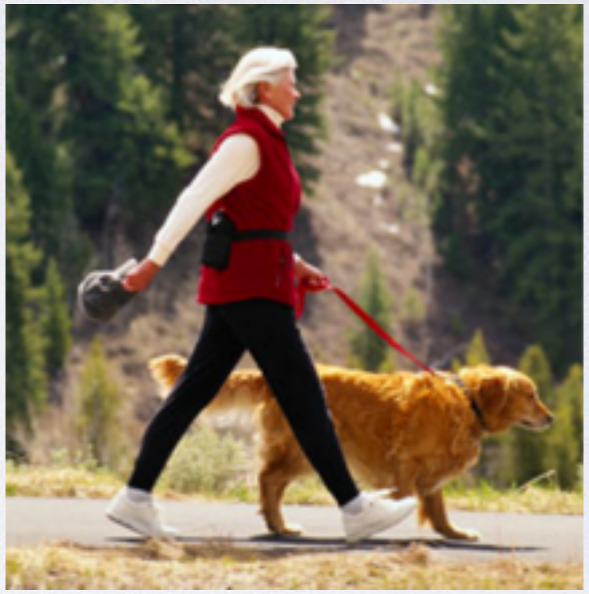




“a woman walking her dog in a park”

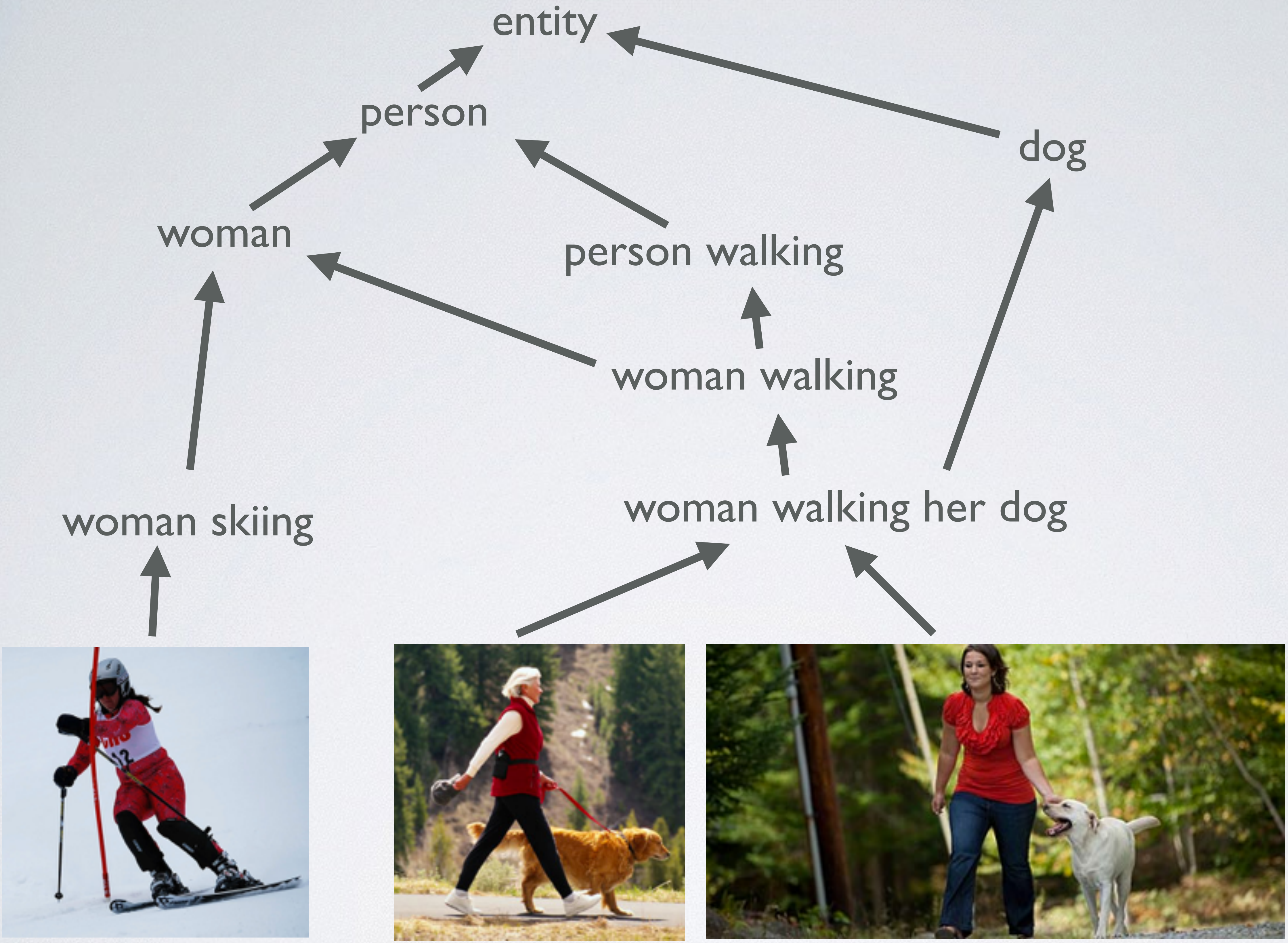
Semantic Image Search

- Given a database of images and a natural language query, identify which images it accurately **describes**



“a woman walking her dog in a park”

What is the relationship between images and the language we use to **describe** them?



entity

person

dog

woman

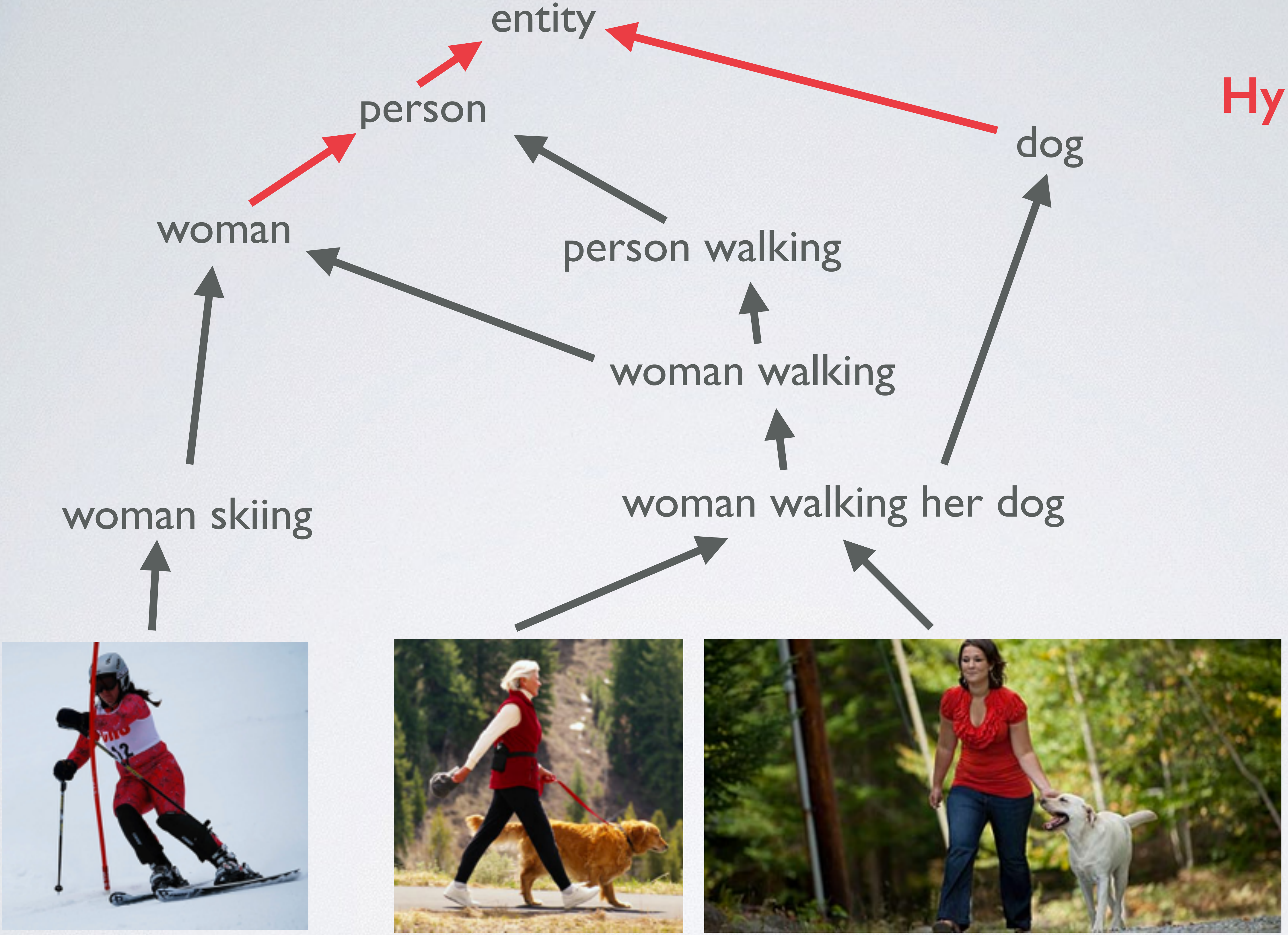
person walking

woman walking

woman skiing

woman walking her dog

Hypernymy



entity

person

dog

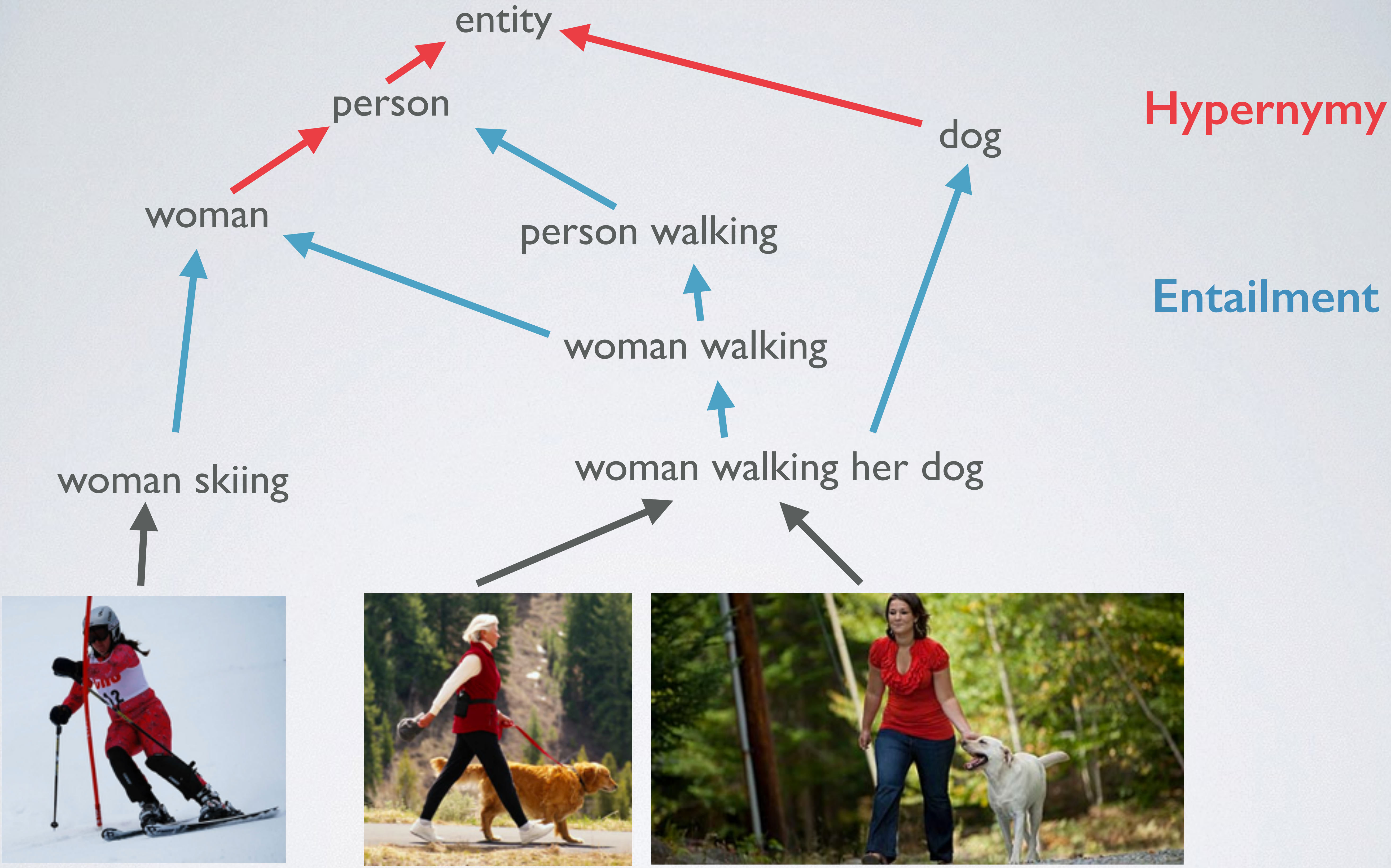
woman

person walking

woman walking

woman skiing

woman walking her dog



entity

person

woman

person walking

dog

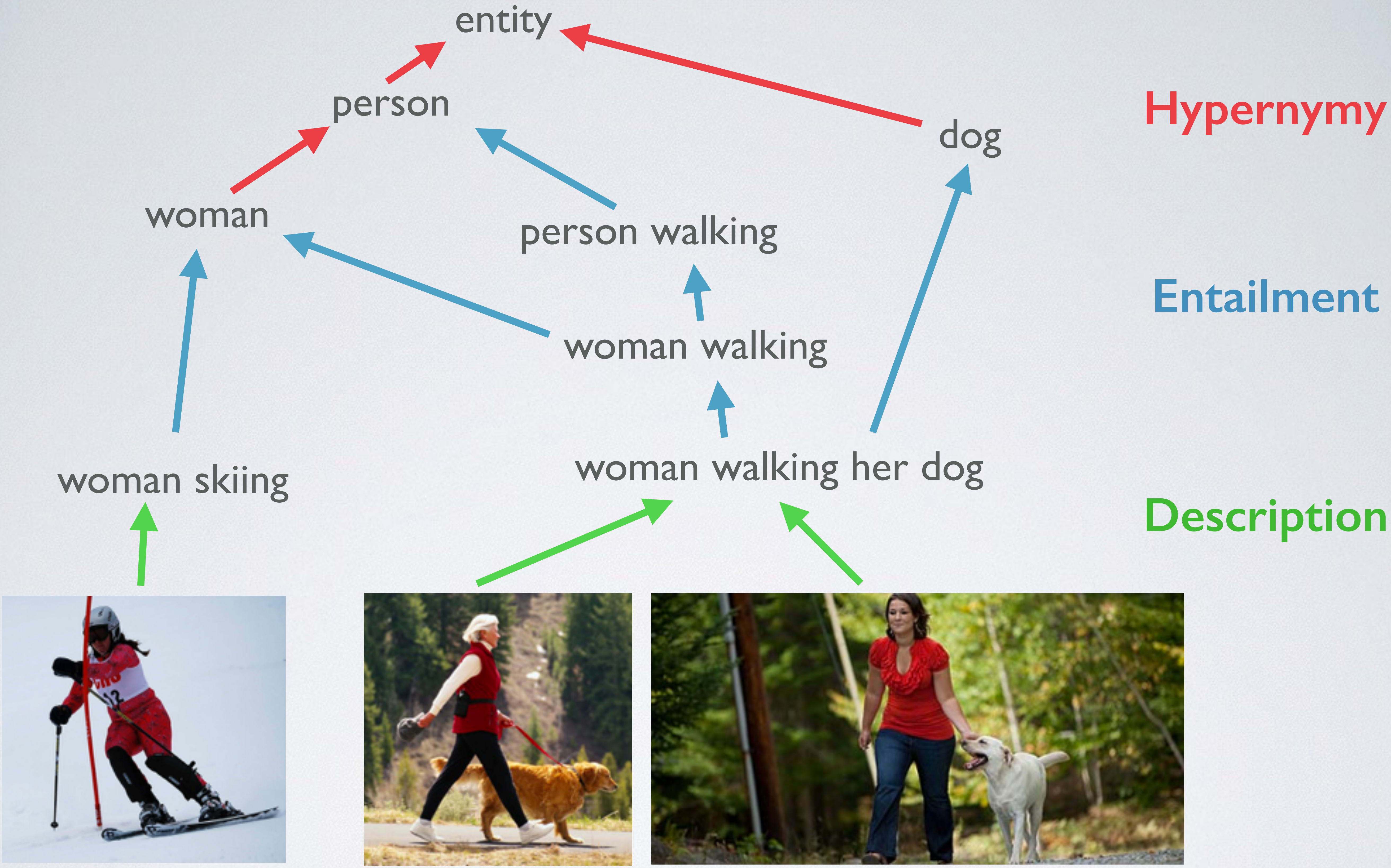
woman walking

woman skiing

woman walking her dog

Hypernymy

Entailment



entity

person

dog

woman

person walking

woman walking

woman skiing

woman walking her dog

Hypernymy

Entailment

Description

Modeling the Visual-Semantic Hierarchy

Modeling the Visual-Semantic Hierarchy

Most previous approaches with learned representations either

Modeling the Visual-Semantic Hierarchy

Most previous approaches with learned representations either

- Use symmetric similarity in the embedding space
  e.g. (Frome et al, 2013; Socher et al, 2014; Karpathy and Li, 2015)

Modeling the Visual-Semantic Hierarchy

Most previous approaches with learned representations either

- Use symmetric similarity in the embedding space
  e.g. (Frome et al, 2013; Socher et al, 2014; Karpathy and Li, 2015)
- Learn an unconstrained binary relation, e.g. (Socher et al, 2013)

Modeling the Visual-Semantic Hierarchy

Most previous approaches with learned representations either

- Use symmetric similarity in the embedding space e.g. (Frome et al, 2013; Socher et al, 2014; Karpathy and Li, 2015)
- Learn an unconstrained binary relation, e.g. (Socher et al, 2013)

Our approach:

Modeling the Visual-Semantic Hierarchy

Most previous approaches with learned representations either

- Use symmetric similarity in the embedding space
  e.g. (Frome et al, 2013; Socher et al, 2014; Karpathy and Li, 2015)
- Learn an unconstrained binary relation, e.g. (Socher et al, 2013)

Our approach:

- Impose a partial-order prior by embedding into an **ordered** space.

Ordered Embedding Space

Ordered Embedding Space

Space: $\mathbb{R}_{+}^{N}$

Ordered Embedding Space

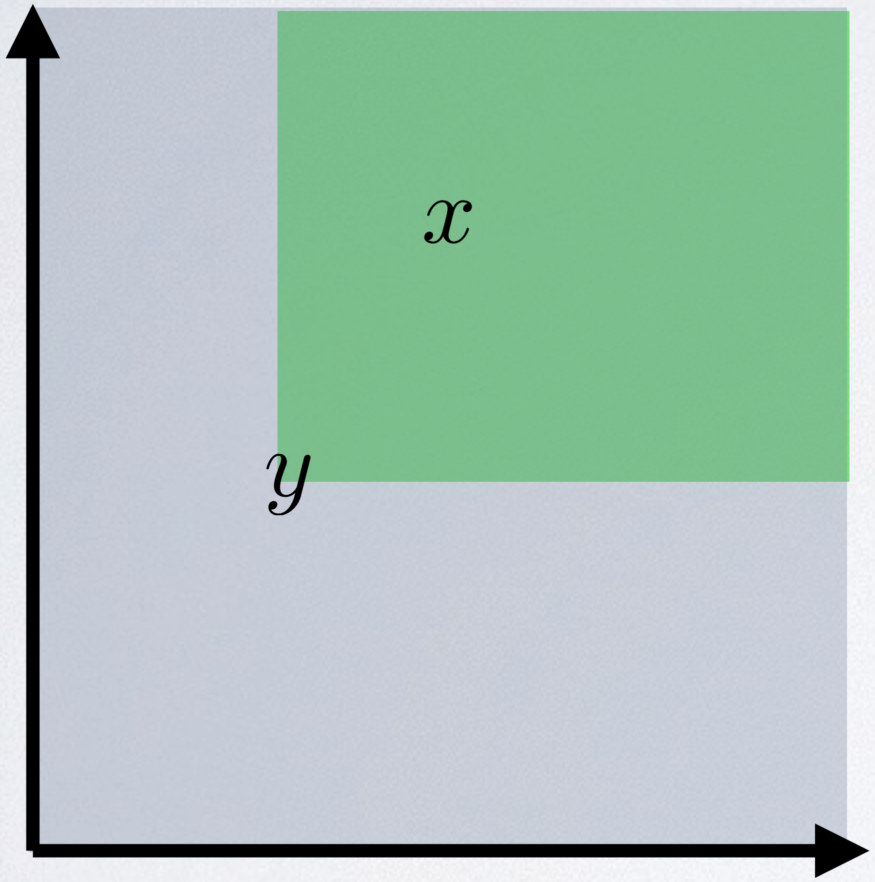
Space: $\mathbb{R}_+^N$

Order: $x \preceq y$ if and only if $\forall i, x_i \geq y_i$

Ordered Embedding Space

Space: $\mathbb{R}_{+}^{N}$

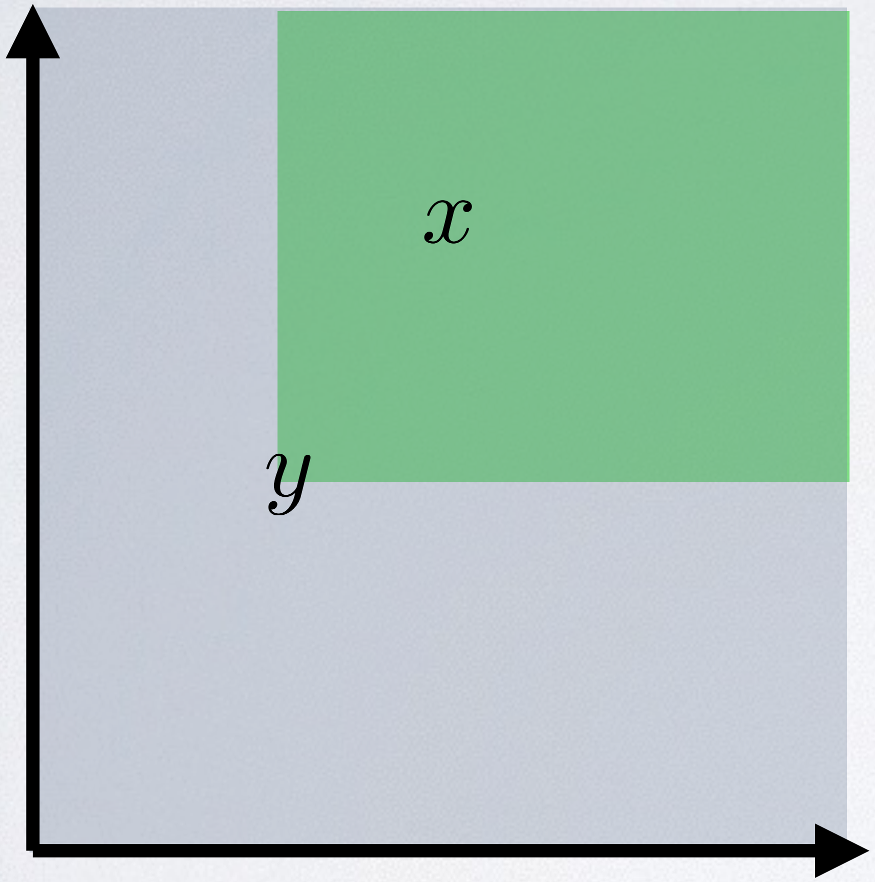
Order: $x \preceq y$ if and only if $\forall i, x_i \geq y_i$

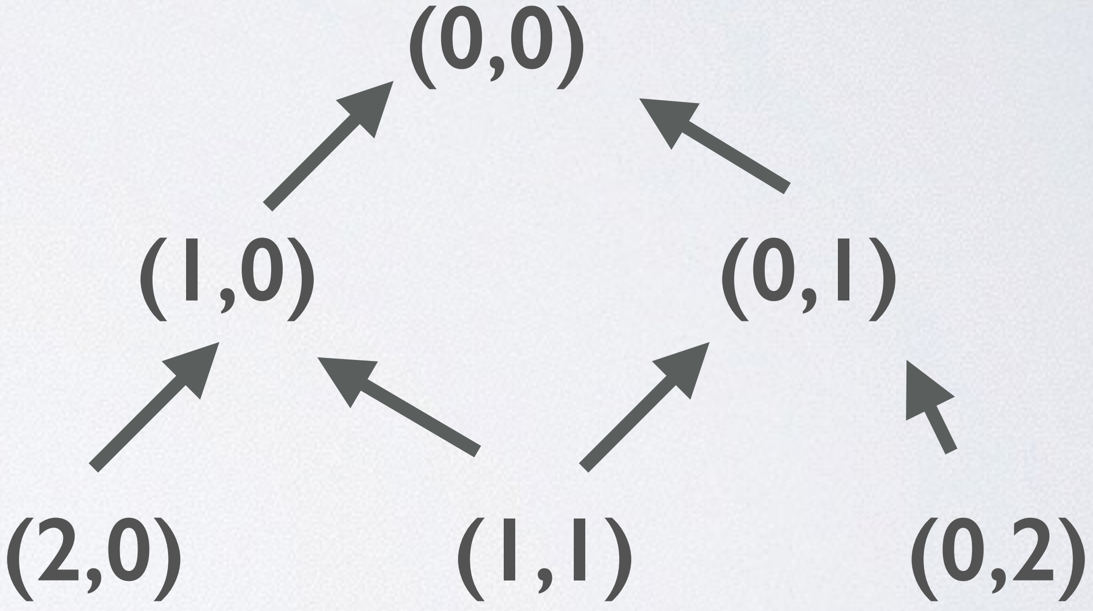


Ordered Embedding Space

Space: $\mathbb{R}_+^N$

Order: $x \preceq y \text{ if and only if } \forall i, x_i \geq y_i$




(0,0)

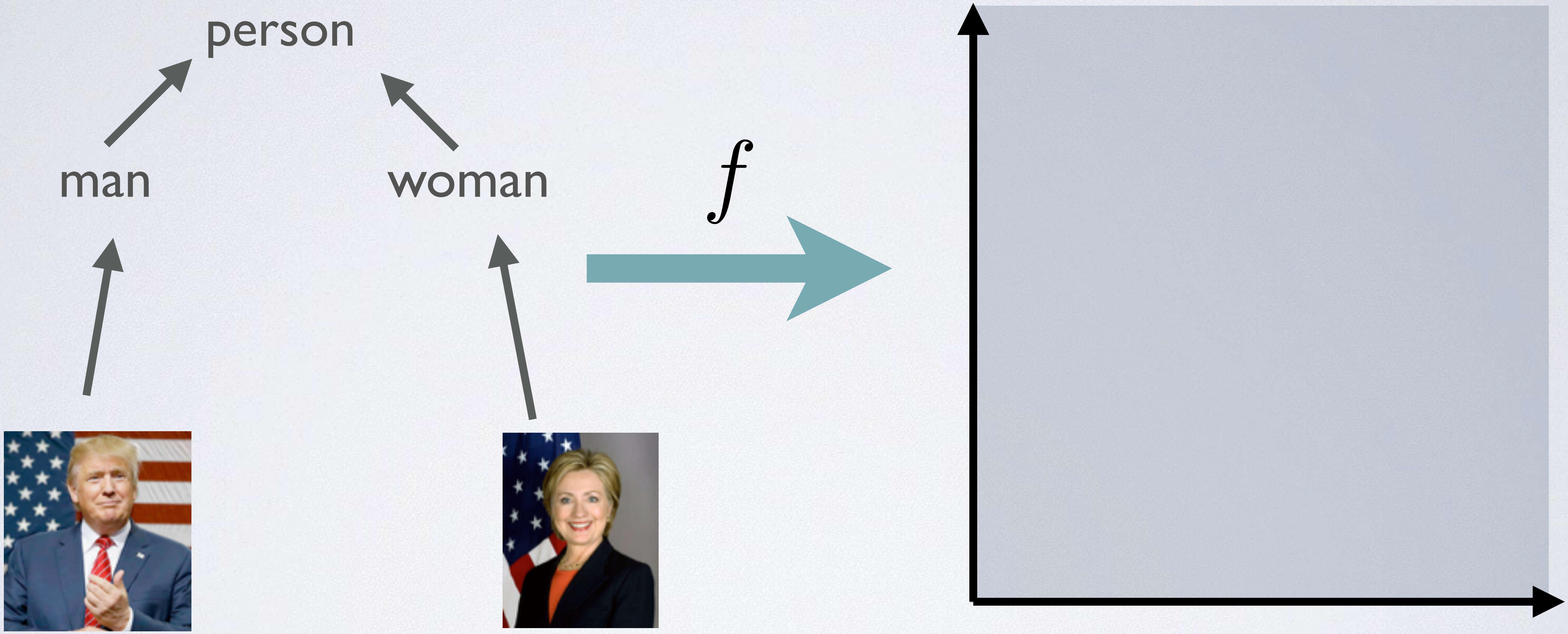
(1,0)

(0,1)

(2,0)

(1,1)

(0,2)



person

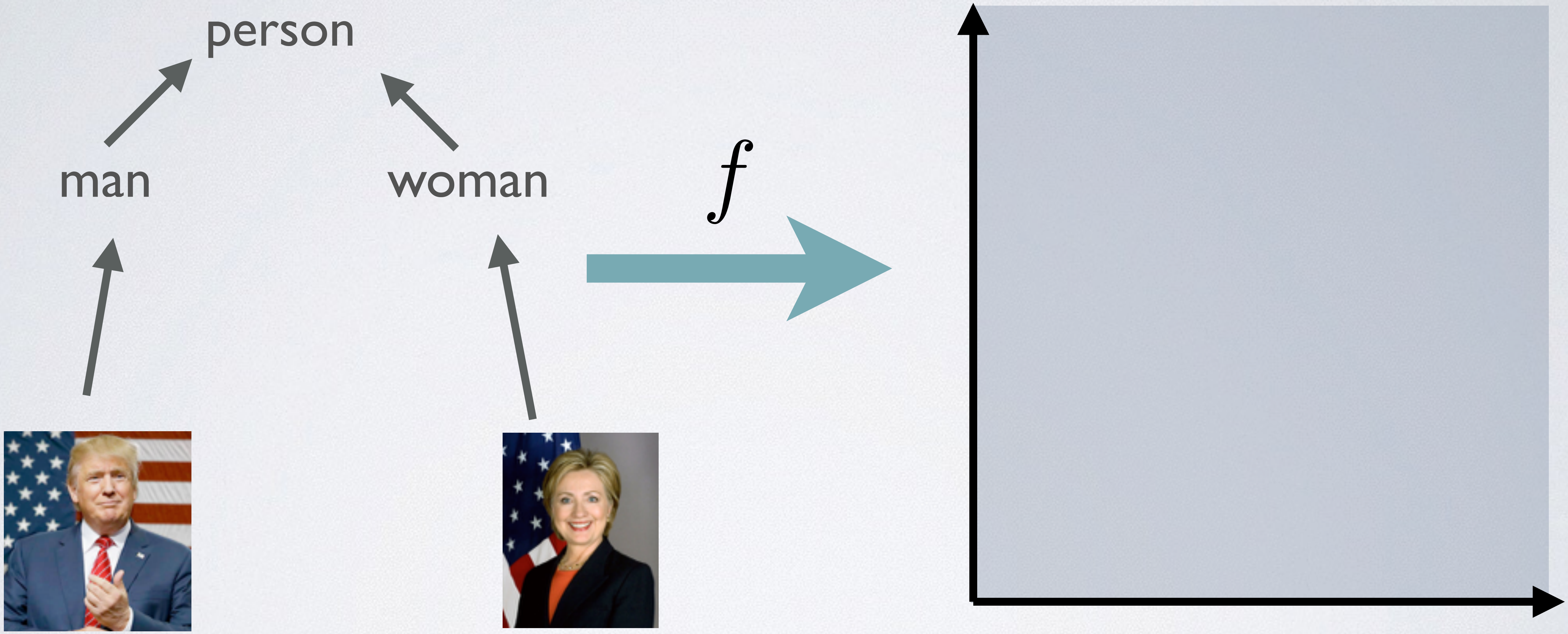
man

woman

f

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

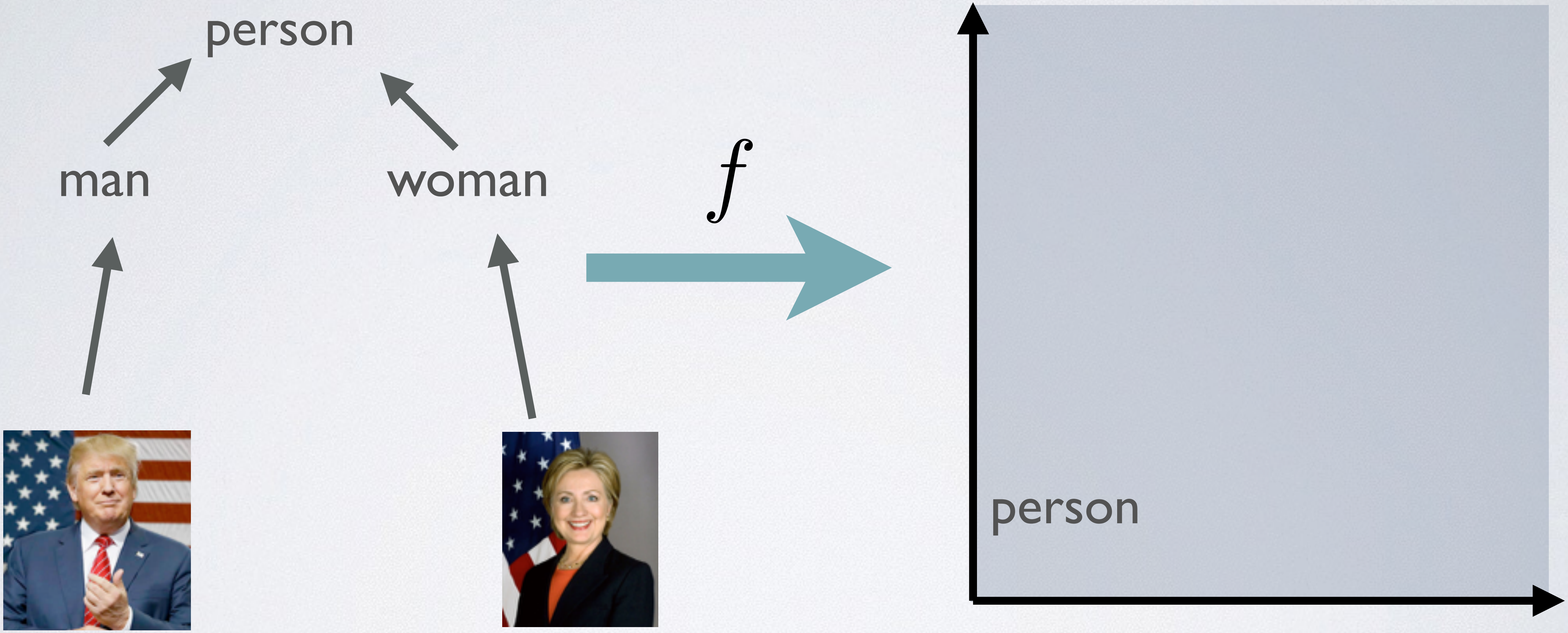
man

woman

f

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

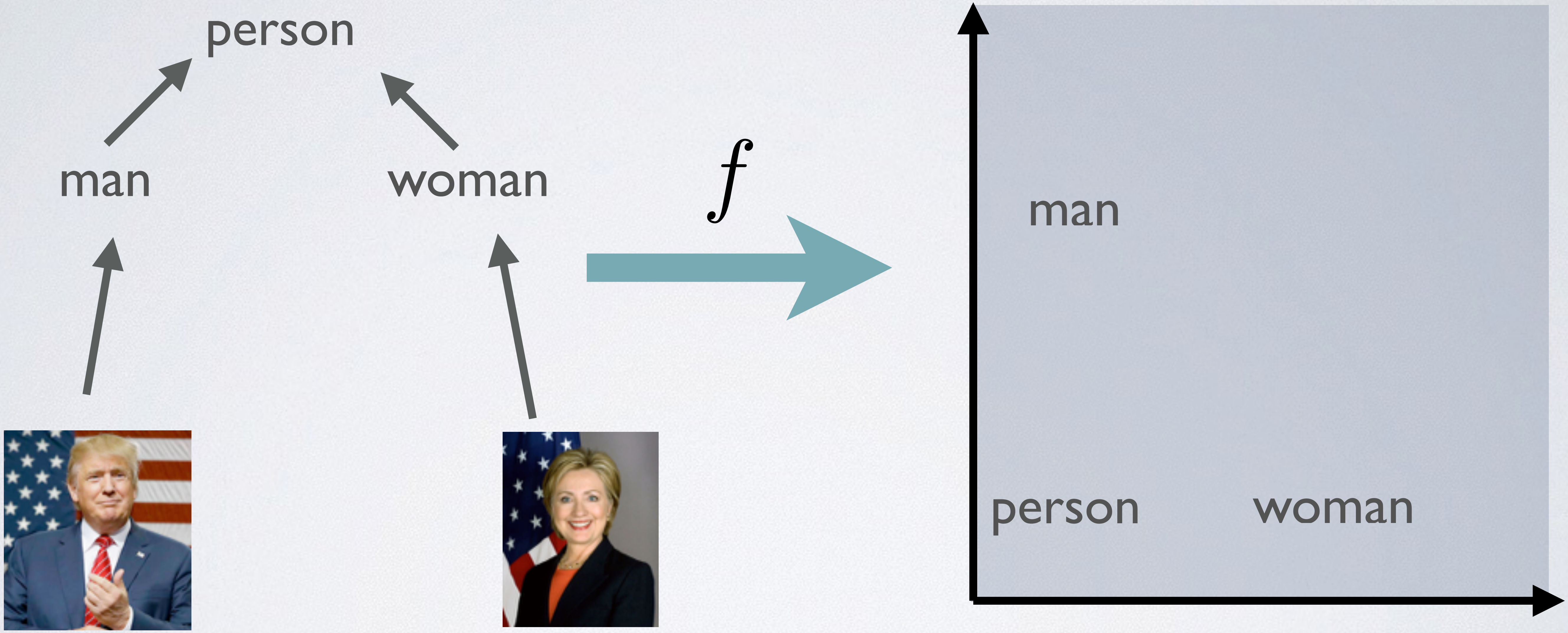
woman

f

person

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

f

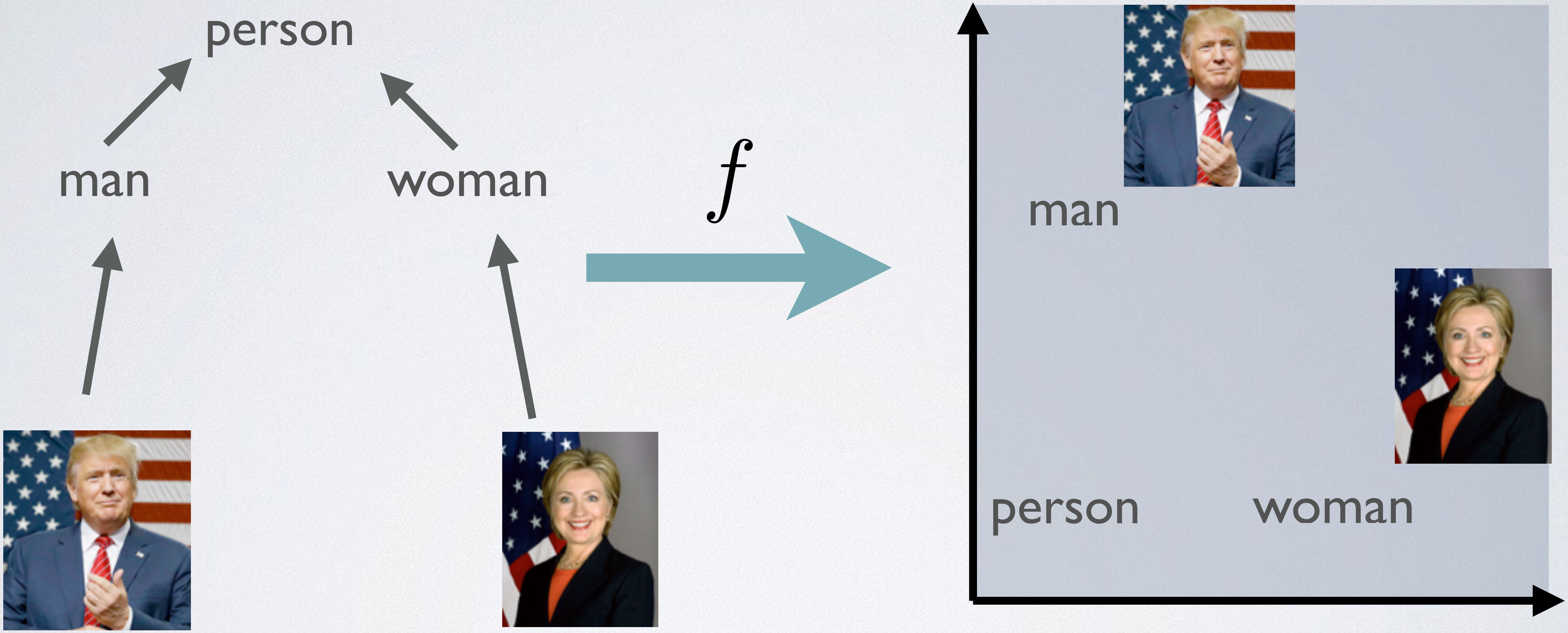
man

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

f

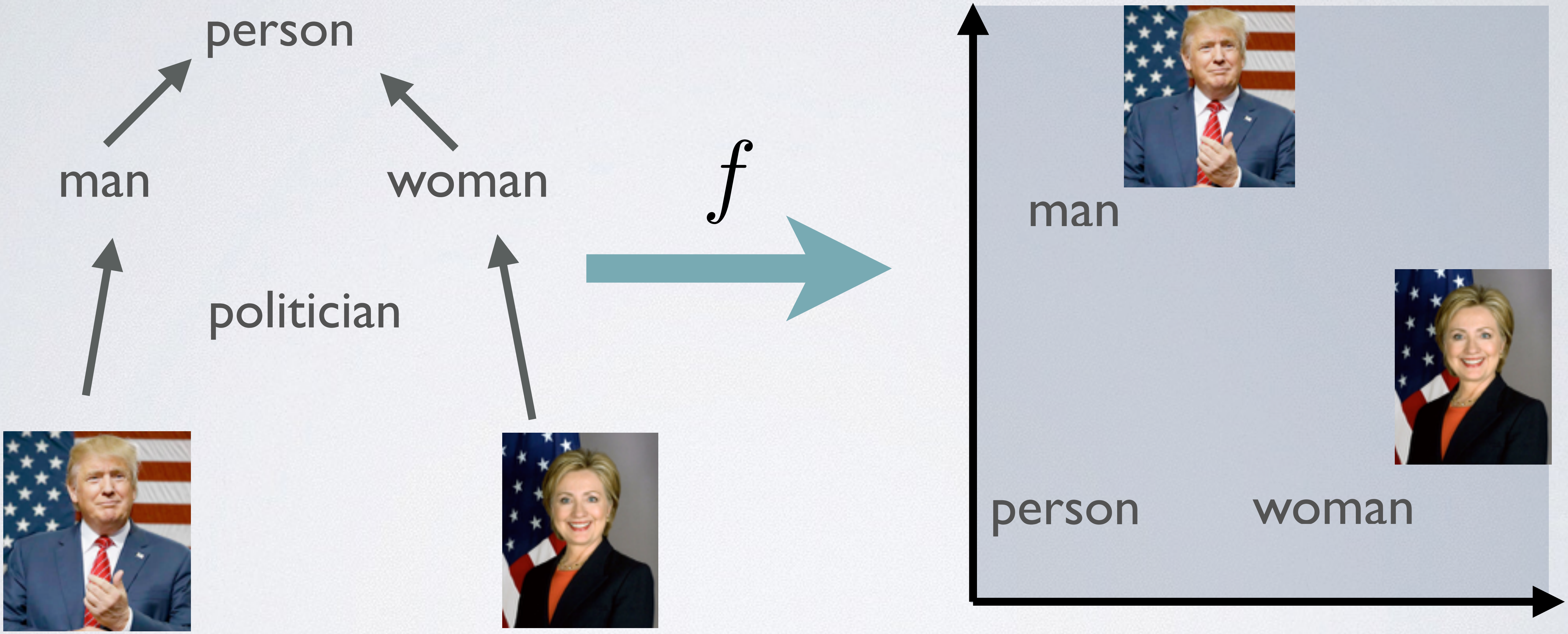
man

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

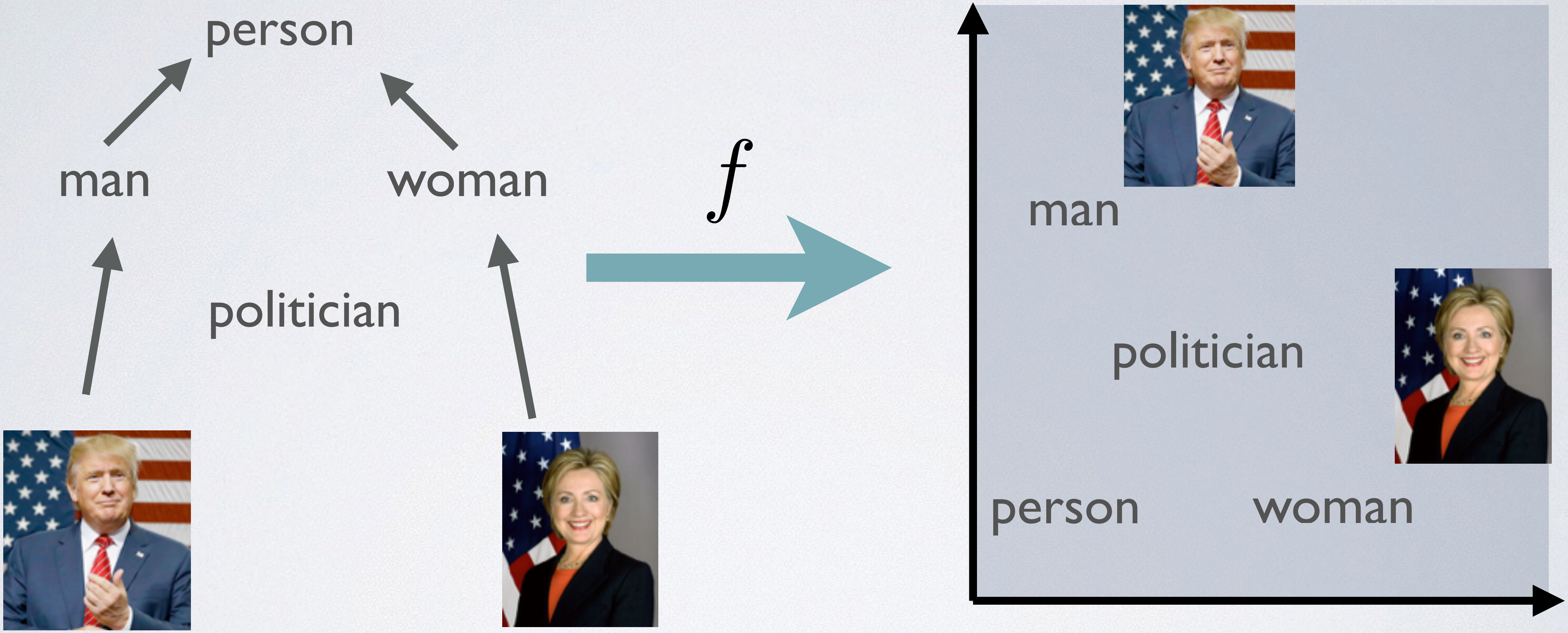
man

person

woman

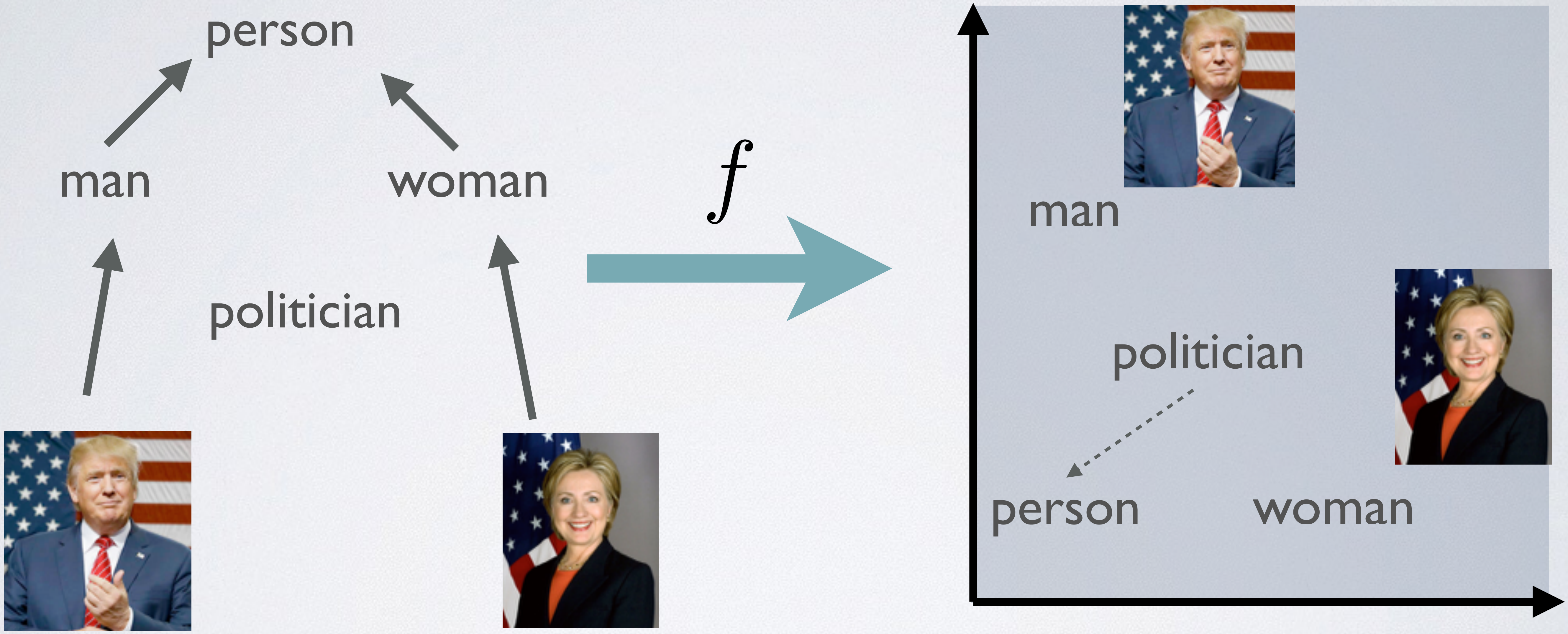
f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

man

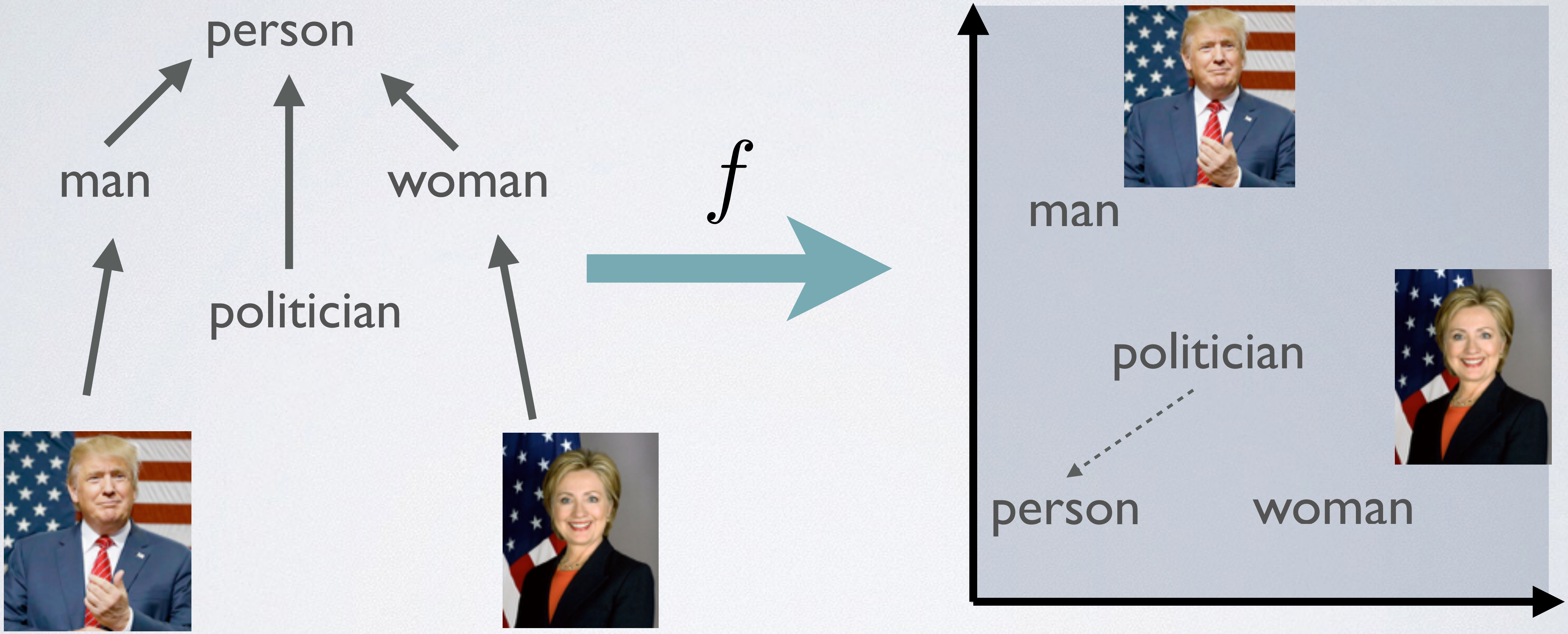
politician

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

man

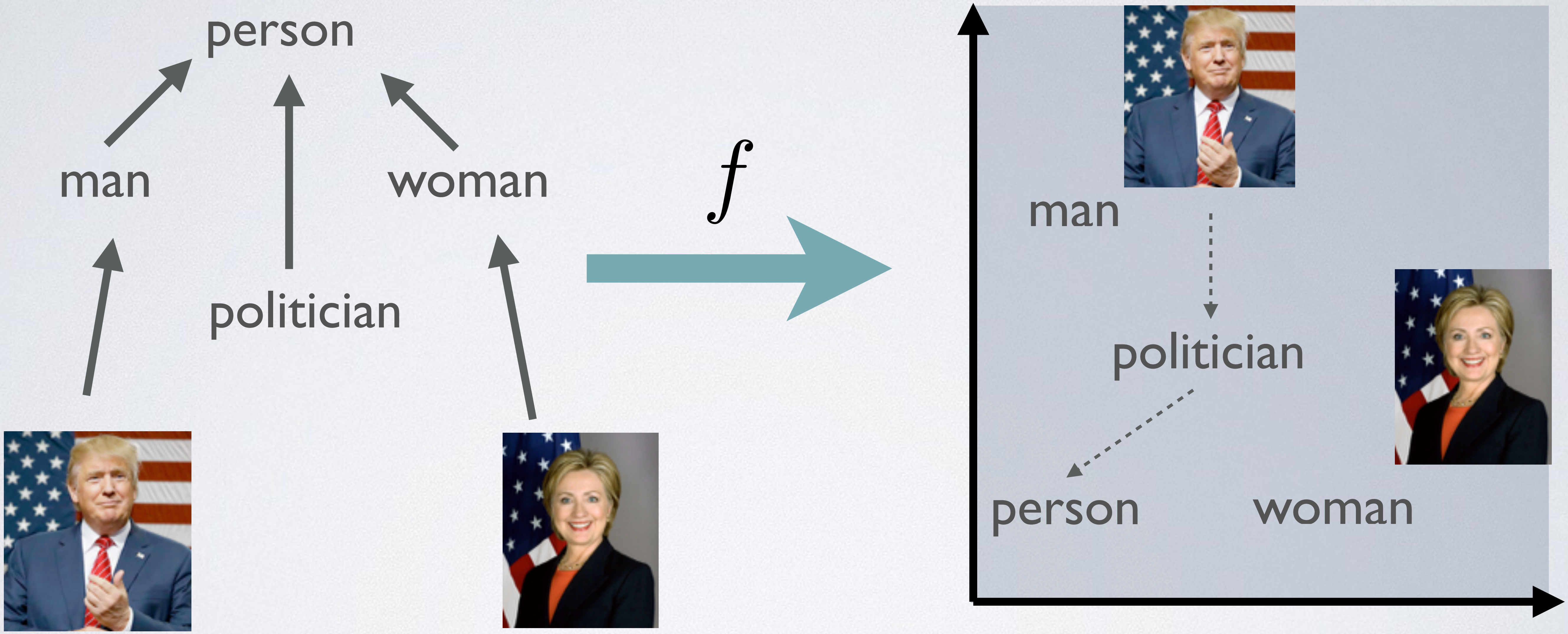
politician

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

man

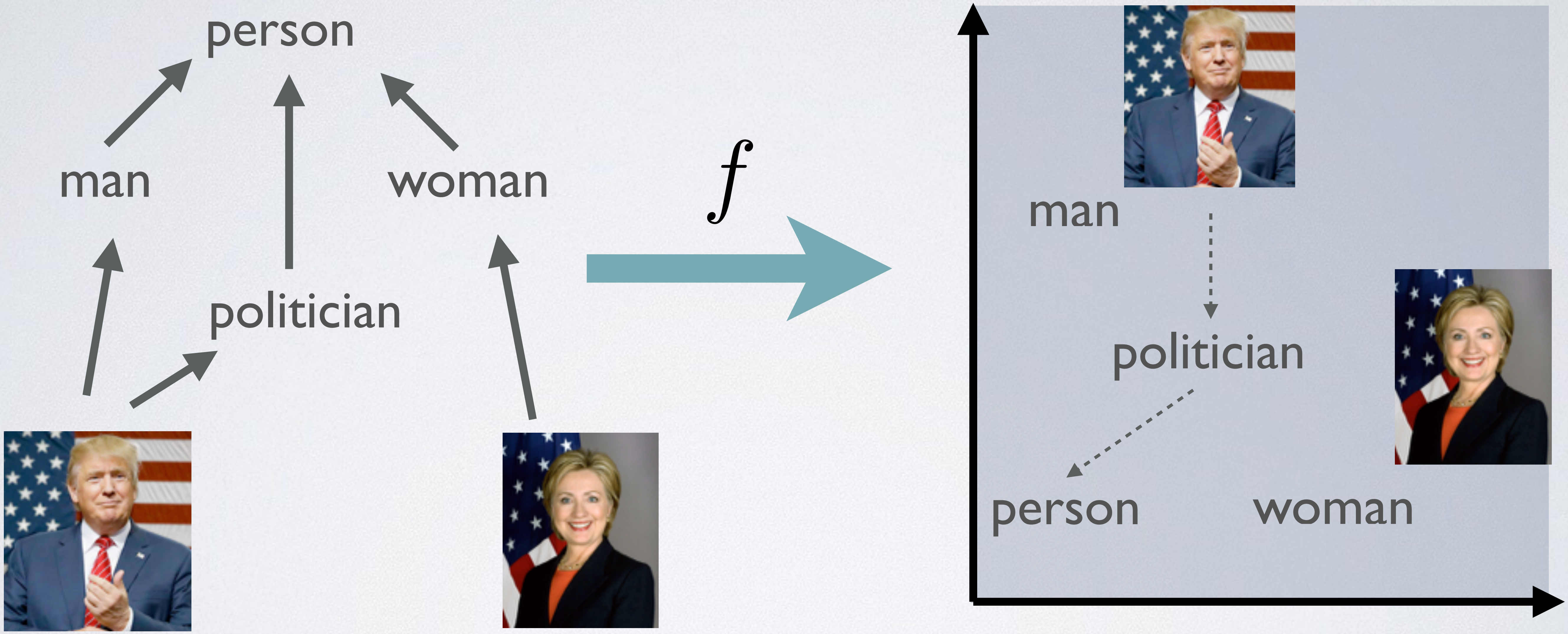
politician

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

man

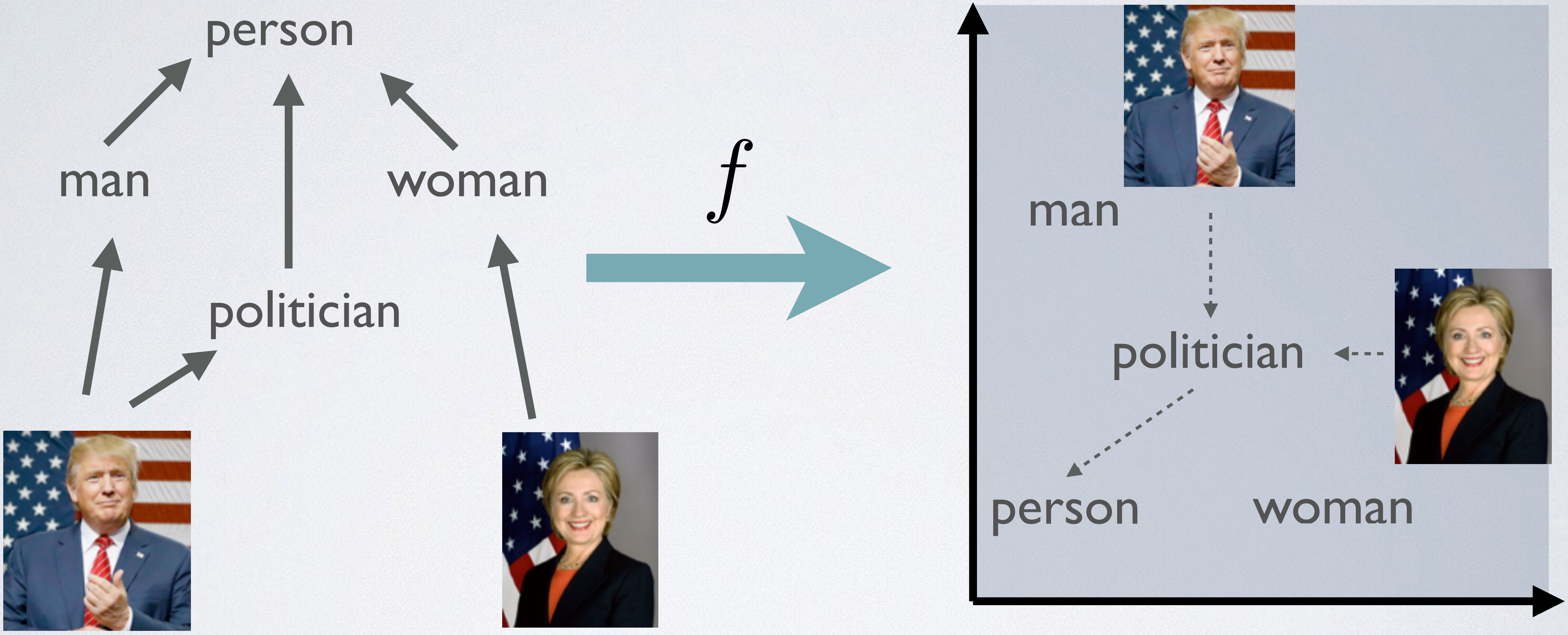
politician

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

f

man

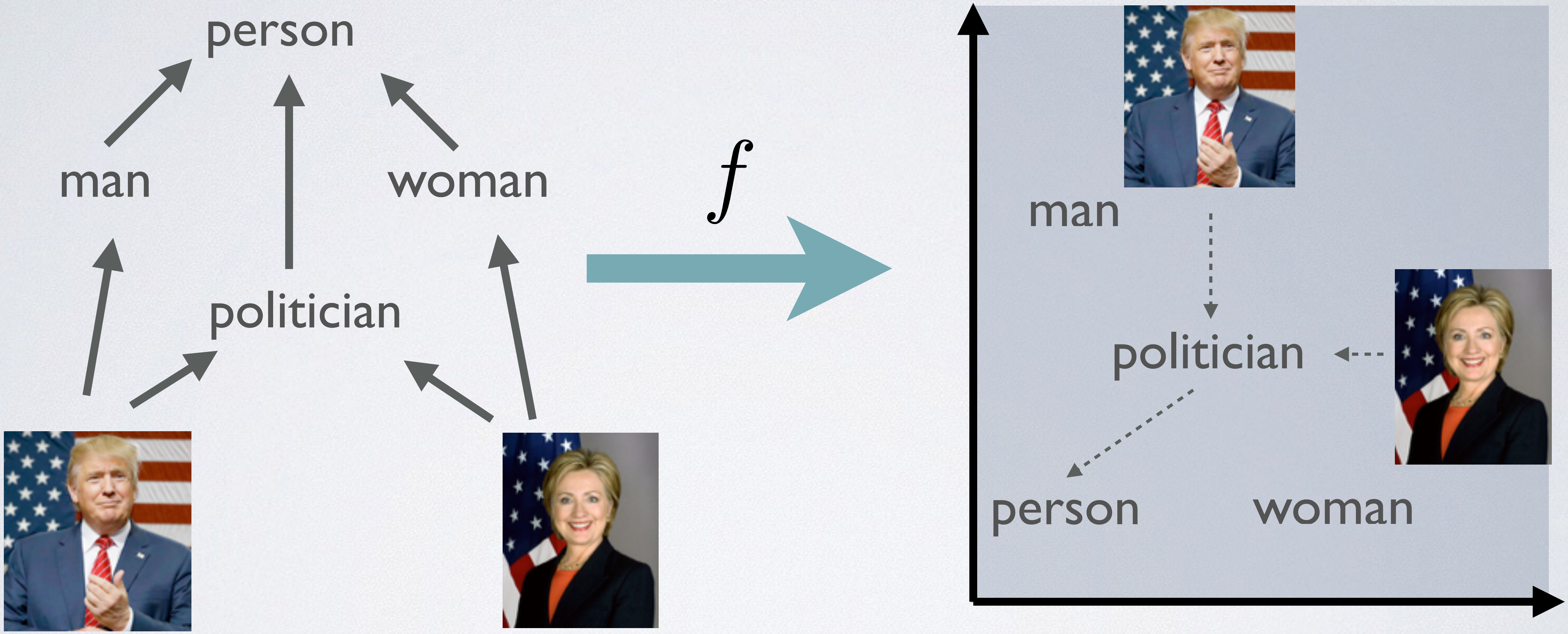
politician

person

woman

f is an **order-embedding** if

$$u \preceq v \iff f(u) \preceq f(v)$$



person

man

woman

politician

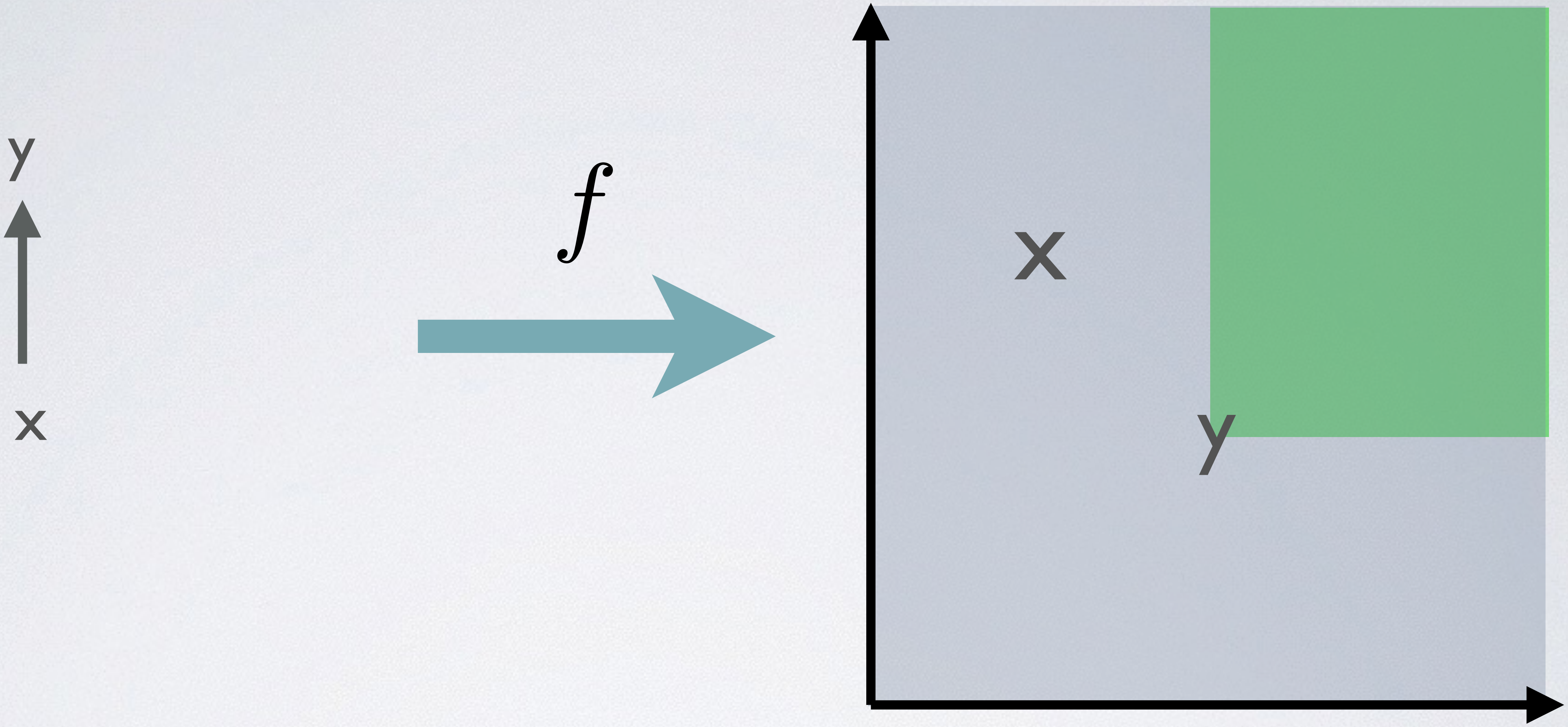
f

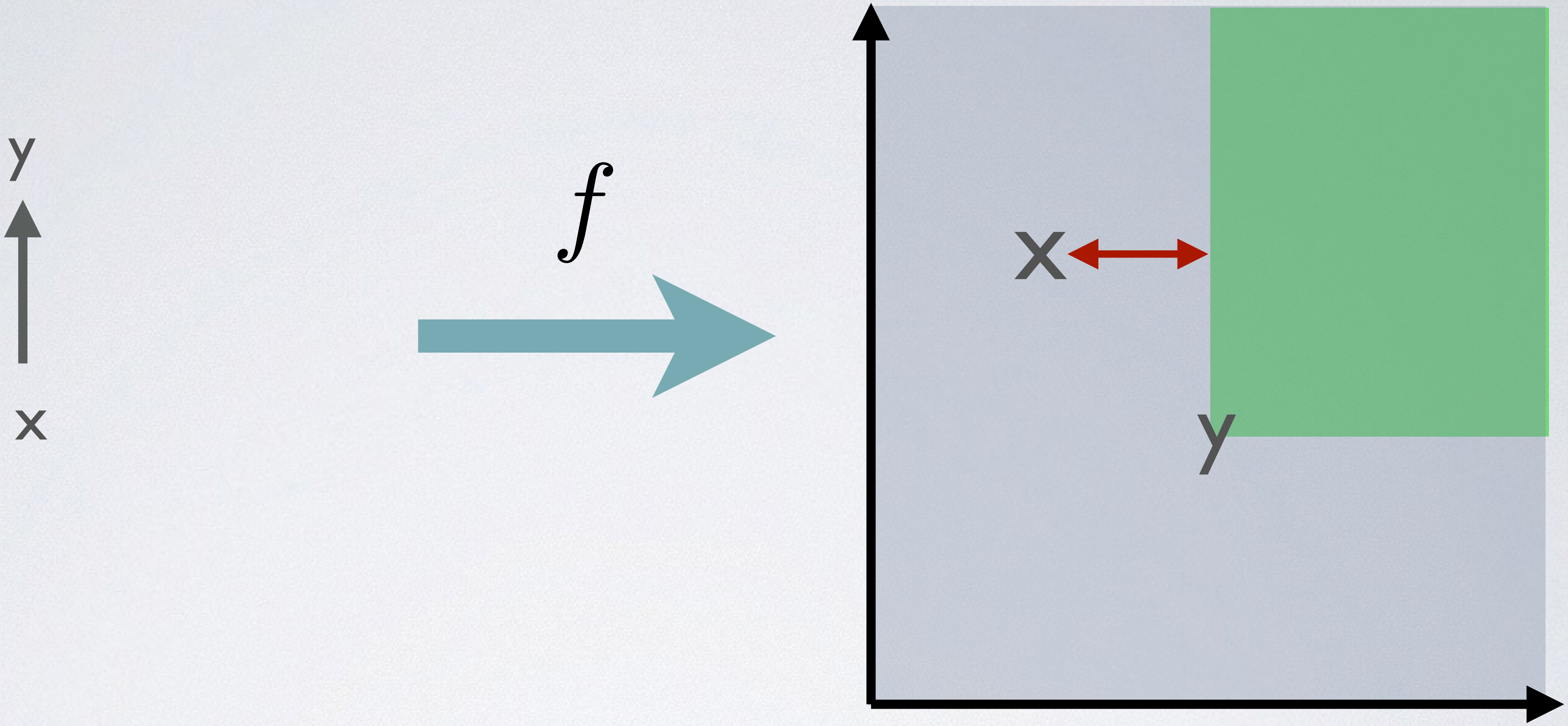
man

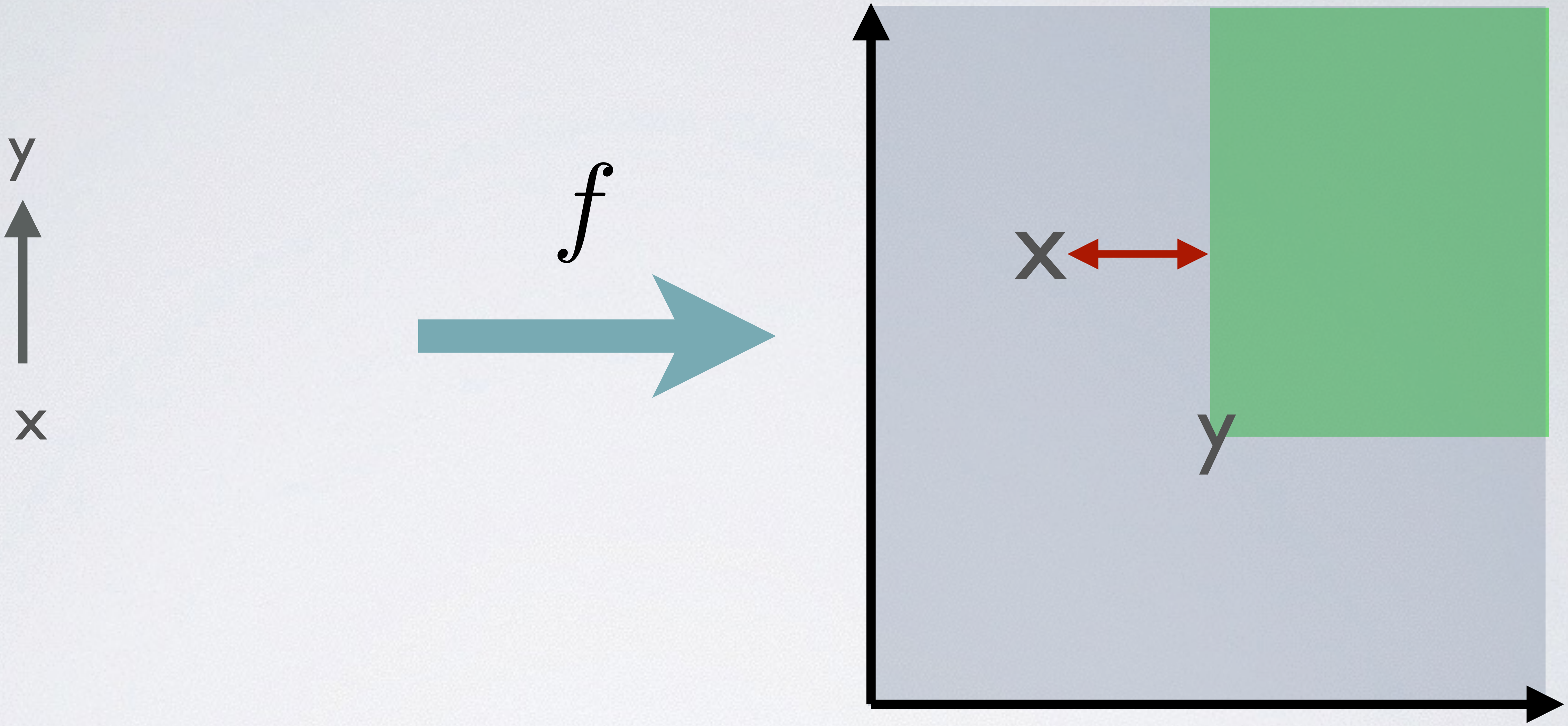
politician

person

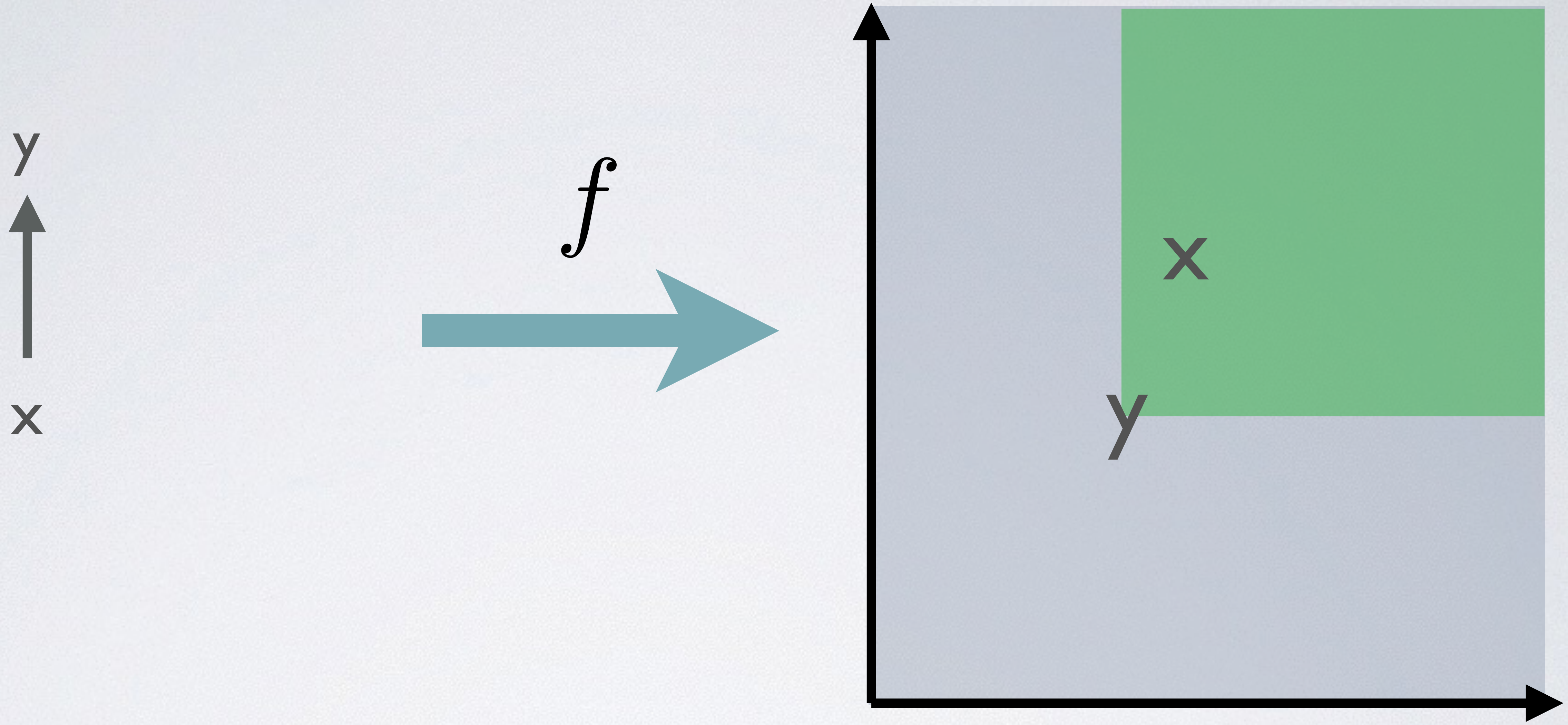
woman


y
x
f
x
y


y
x
f
x
y



Order violation error: $E(x, y) = ||\max(0, y - x)||^2$

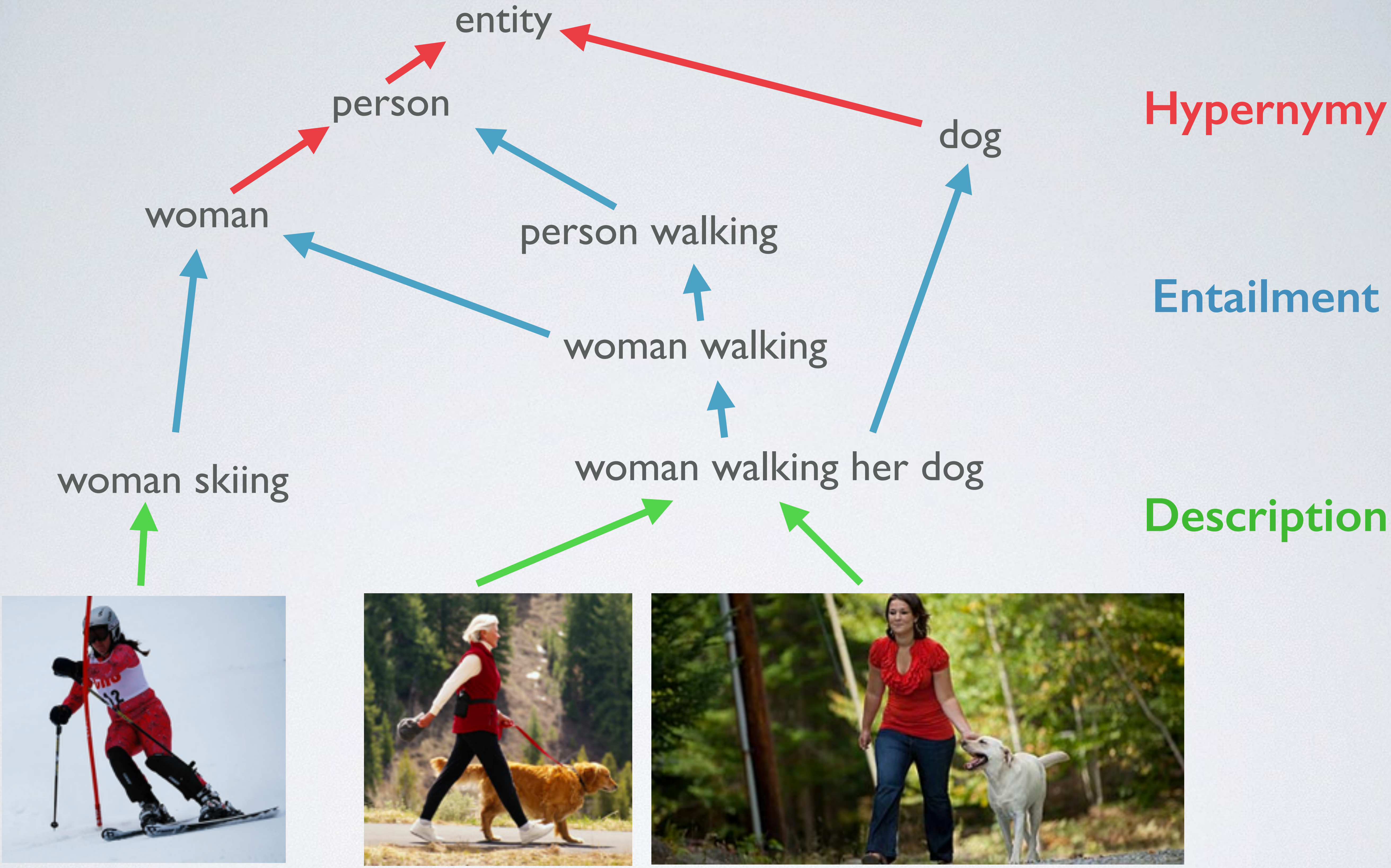


Order violation error: $E(x, y) = ||\max(0, y - x)||^2$

Order violation error: $E(x, y) = || \max(0, y - x)||^2$

Order violation error: $E(x, y) = ||\max(0, y - x)||^2$

Experiments



entity

person

dog

woman

person walking

woman walking

woman skiing

woman walking her dog

Hypernymy

Entailment

Description

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

$$S(c, i) = f_c(c) \cdot f_i(i)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

$$S(c, i) = f_c(c) \cdot f_i(i)$$

$$f_c(c) = RNN(c)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c, i) + S(c', i)\} + \sum_{i'} \max\{0, \alpha - S(c, i) + S(c, i')\} \right)$$

$$S(c, i) = f_c(c) \cdot f_i(i)$$

$$f_c(c) = RNN(c)$$

$$f_i(i) = W_i \cdot CNN(i)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

$$S(c, i) = \cancel{f_c(c) \quad f_i(i)}$$

$$f_c(c) = RNN(c)$$

$$f_i(i) = W_i \cdot CNN(i)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c,i)\}$

Learn a caption-image similarity $S(c,i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

$$S(c,i) = \cancel{f_c(c) \quad f_i(i)} \; -E(f_i(i), f_c(c))$$

$$f_c(c) = RNN(c)$$

$$f_i(i) = W_i \cdot CNN(i)$$

Image Search w. Visual-Semantic Embeddings

(Kiros et al, 2014)

Given a dataset of caption-image pairs $\{(c, i)\}$

Learn a caption-image similarity $S(c, i)$

By minimizing the pairwise ranking objective

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c,i) + S(c',i)\} + \sum_{i'} \max\{0, \alpha - S(c,i) + S(c,i')\} \right)$$

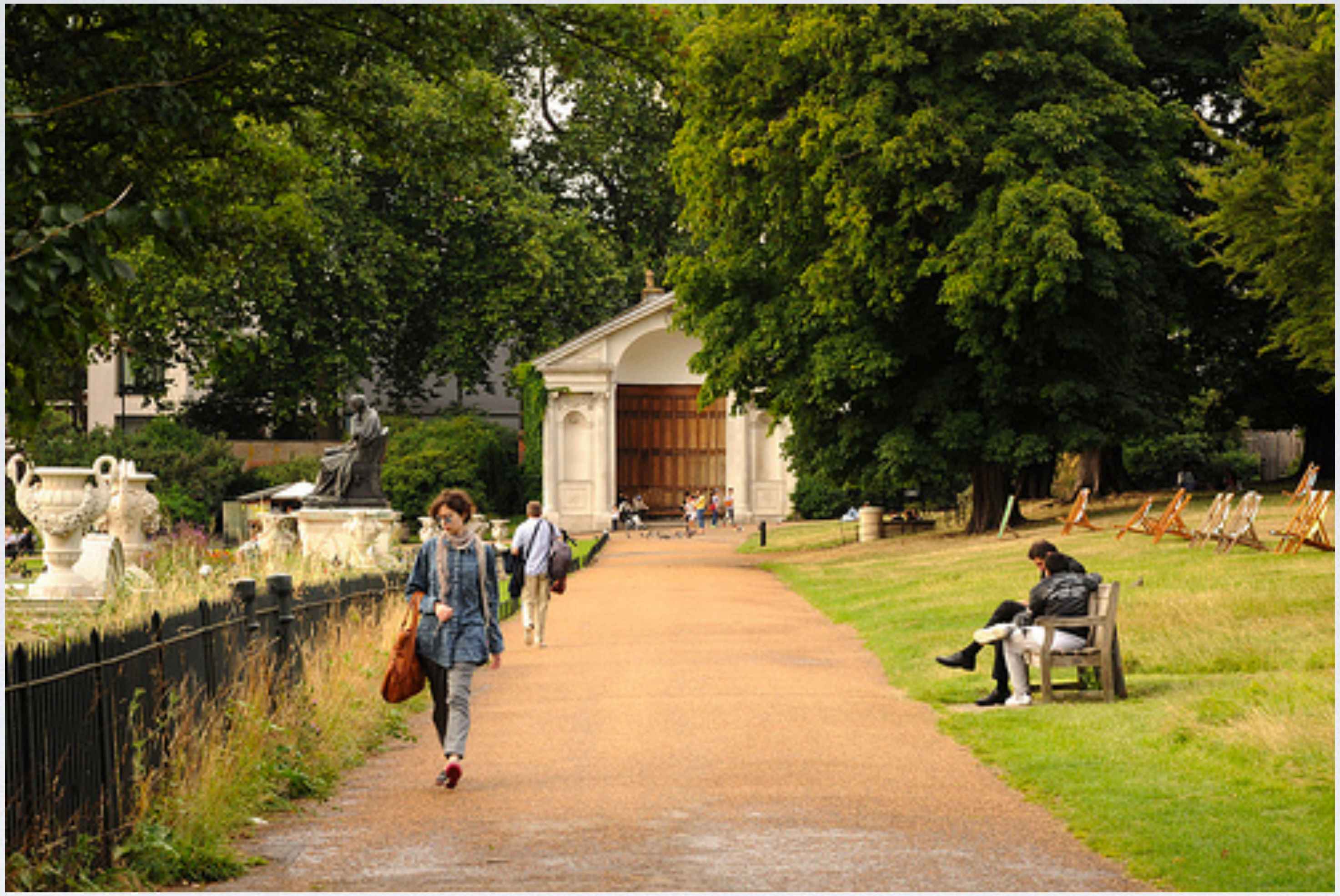
$$S(c, i) = \cancel{f_c(c) \cdot f_i(i)} \; -E(f_i(i), f_c(c))$$

$$f_c(c) = |RNN(c)\,|$$

$$f_i(i) = |W_i \cdot CNN(i)|$$

MS-COCO Ranking Benchmark

- 120k images
- Each image has 5 human-written captions.
- We use 110k images for training, 5k for validation and test.



- a group of people walking down a small walkway.
- a girl walking on a path near a person on a bench.
- young lady walking down a path on the right is a couple setting on a park bench.
- a woman with a large, brown purse walks down a path while two people sit on a bench.
- people walking and sitting along a road dividing a green park and a cemetery.

Evaluation (Image Search)

- Take each caption from the test set, and rank all test images by decreasing S(c,i) (i.e. increasing order-violation error E)

Recall@k: % of captions for which the GT image was in the first **k**
Mean r: mean rank of first ground-truth image
Med r: median rank of first ground-truth image

Quantitative Results

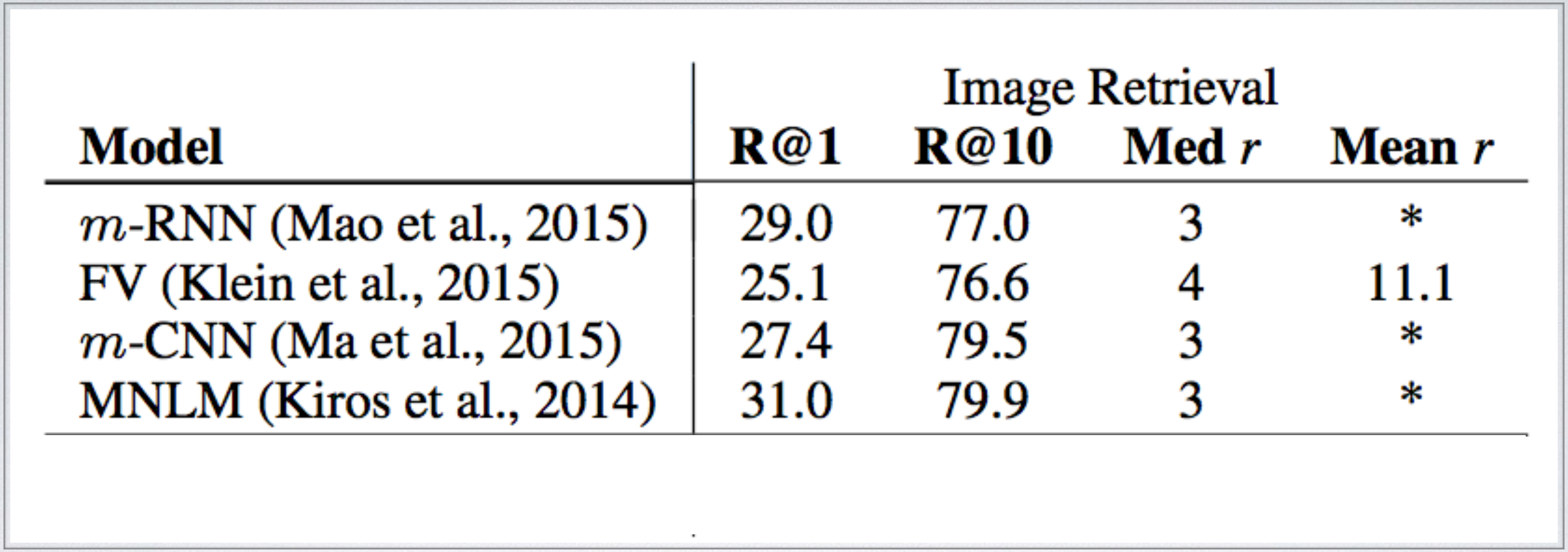


	Image Retrieval			
Model	**R@1**	**R@10**	**Med *r***	**Mean *r***
m-RNN (Mao et al., 2015)	29.0	77.0	3	*
FV (Klein et al., 2015)	25.1	76.6	4	11.1
m-CNN (Ma et al., 2015)	27.4	79.5	3	*
MNLM (Kiros et al., 2014)	31.0	79.9	3	*

Table 1: Results on COCO test with test set of1k images.

Quantitative Results

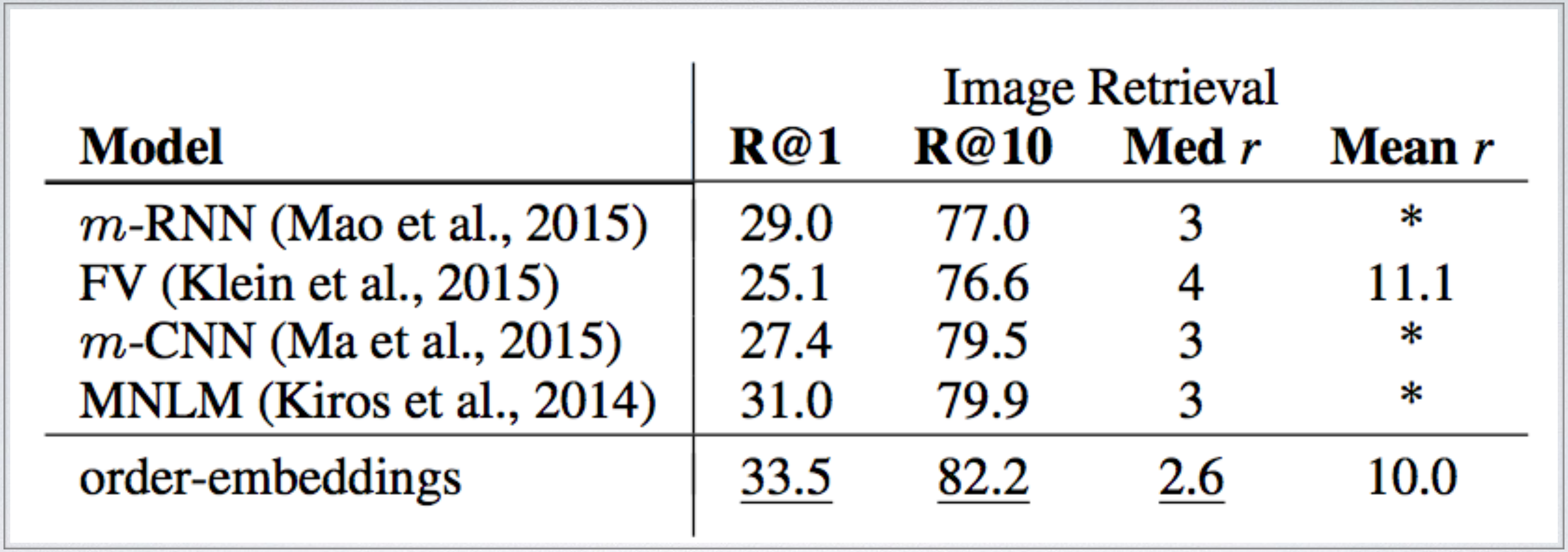
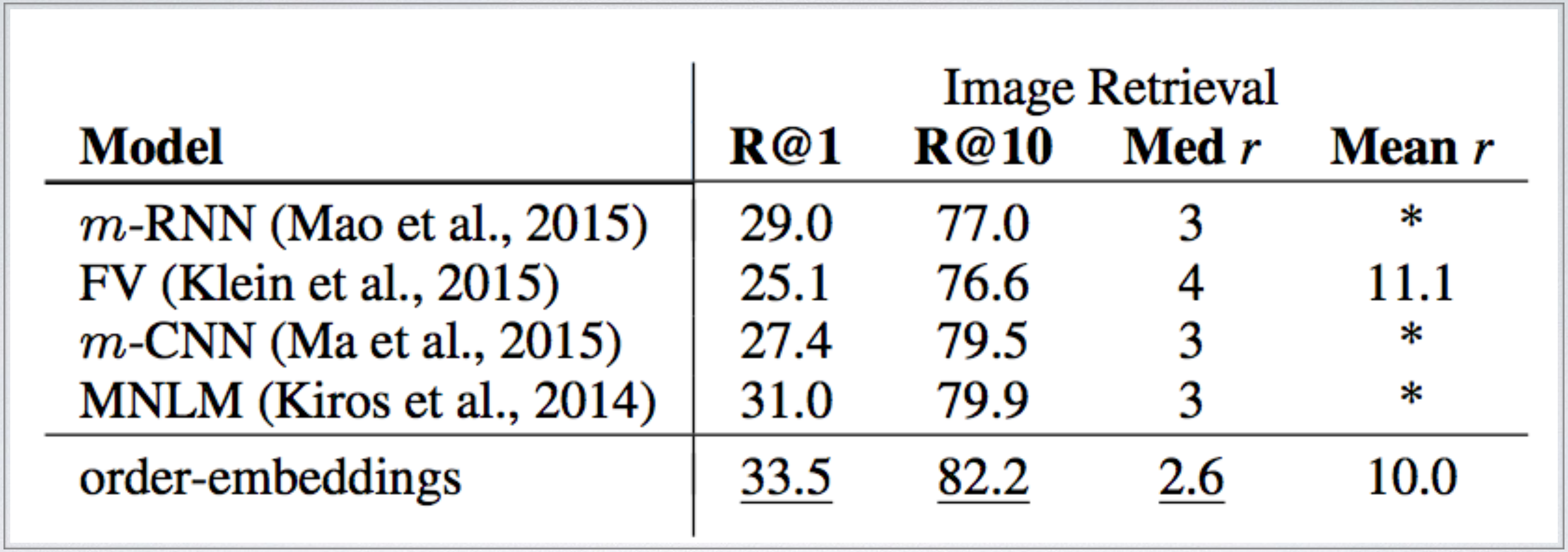


	Image Retrieval			
Model	**R@1**	**R@10**	**Med *r***	**Mean *r***
m-RNN (Mao et al., 2015)	29.0	77.0	3	*
FV (Klein et al., 2015)	25.1	76.6	4	11.1
m-CNN (Ma et al., 2015)	27.4	79.5	3	*
MNLM (Kiros et al., 2014)	31.0	79.9	3	*
order-embeddings	33.5	82.2	2.6	10.0

Table 1: Results on COCO test with test set of1k images.

Quantitative Results



Model	Image Retrieval			
	R@1	**R@10**	**Med *r***	**Mean *r***
m-RNN (Mao et al., 2015)	29.0	77.0	3	*
FV (Klein et al., 2015)	25.1	76.6	4	11.1
m-CNN (Ma et al., 2015)	27.4	79.5	3	*
MNLM (Kiros et al., 2014)	31.0	79.9	3	*
order-embeddings	33.5	82.2	2.6	10.0

Table 1: Results on COCO test with test set of1k images.

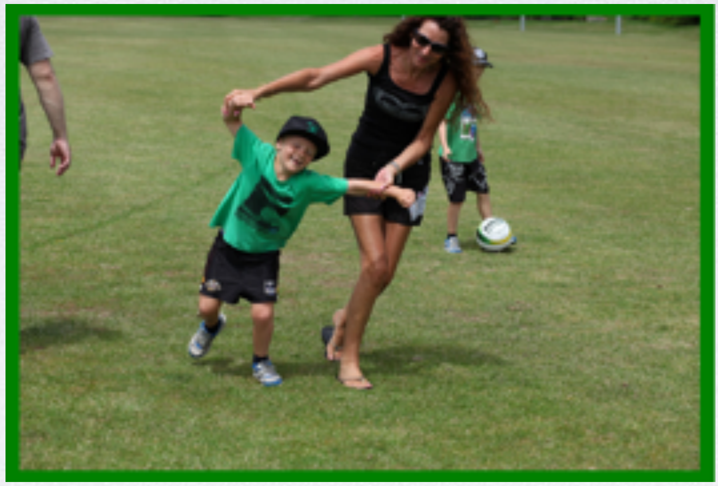
(See our paper for full results)

Image Search Examples: Success

Query

Top Images

"a woman and little boy are walking and holding arms on a soccer field"



"a man in a mask is holding an umbrella"

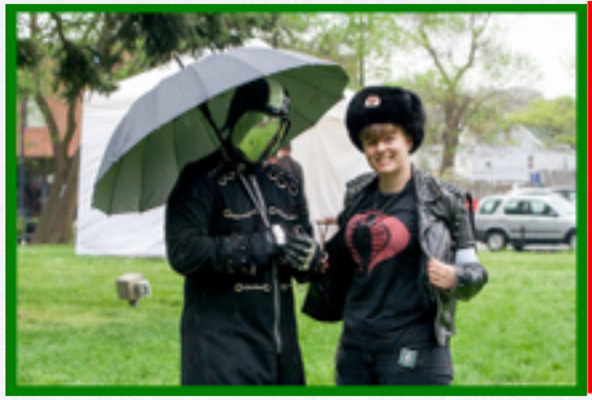


Image Search Example: Failure

Query

"the man is trying to eat three hot dogs are the same time"

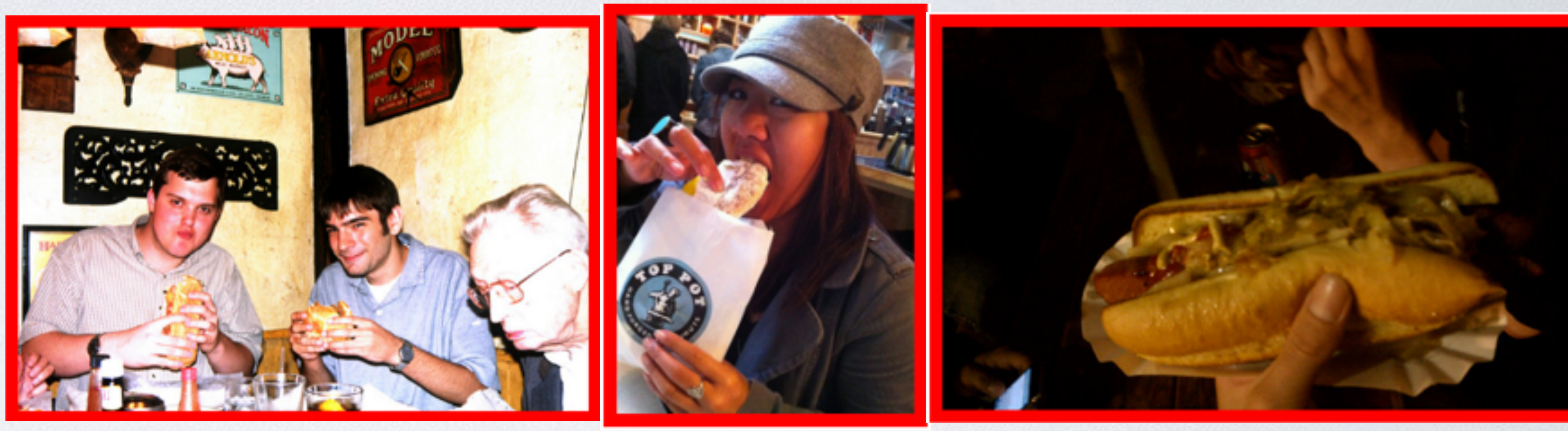
Top Images

Image Search Example: Failure

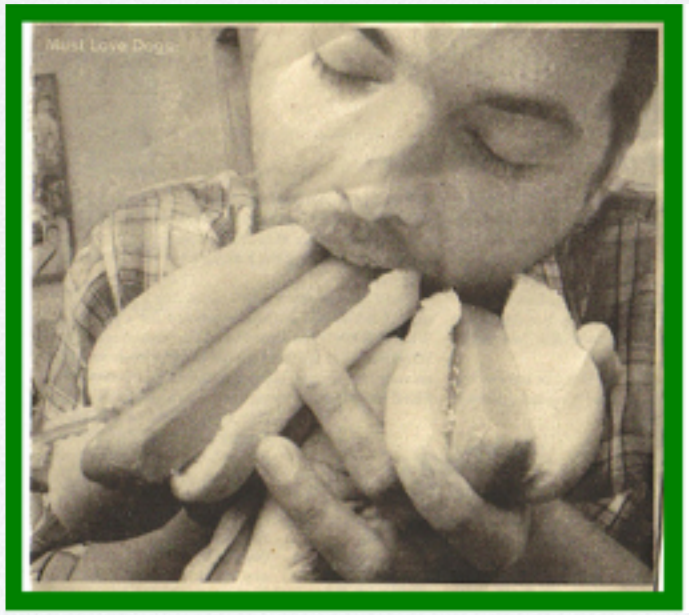
Query

“the man is trying to eat three hot dogs are the same time”

Top Images



GT Image:



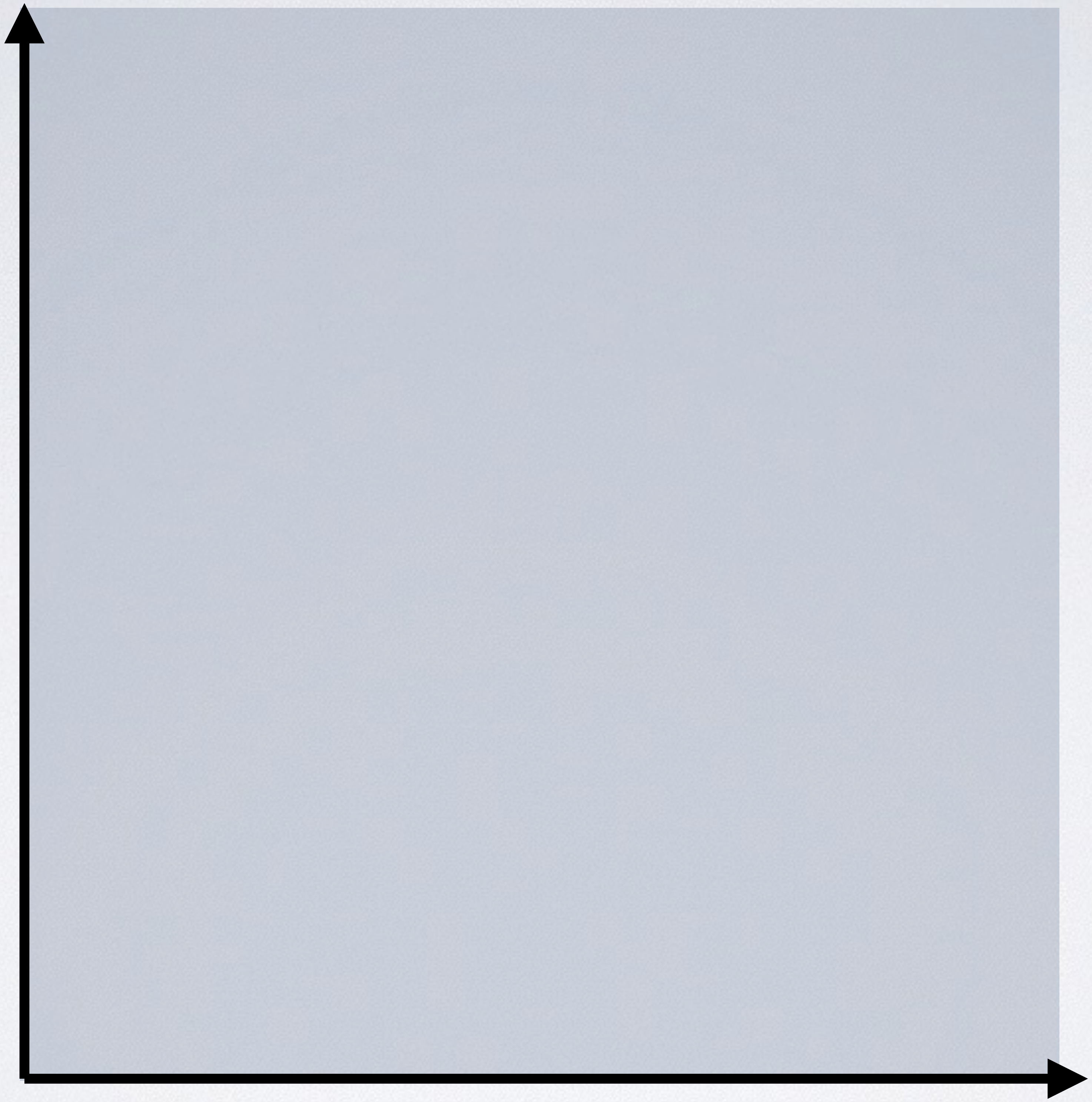
Multimodal Regularities

Multimodal Regularities

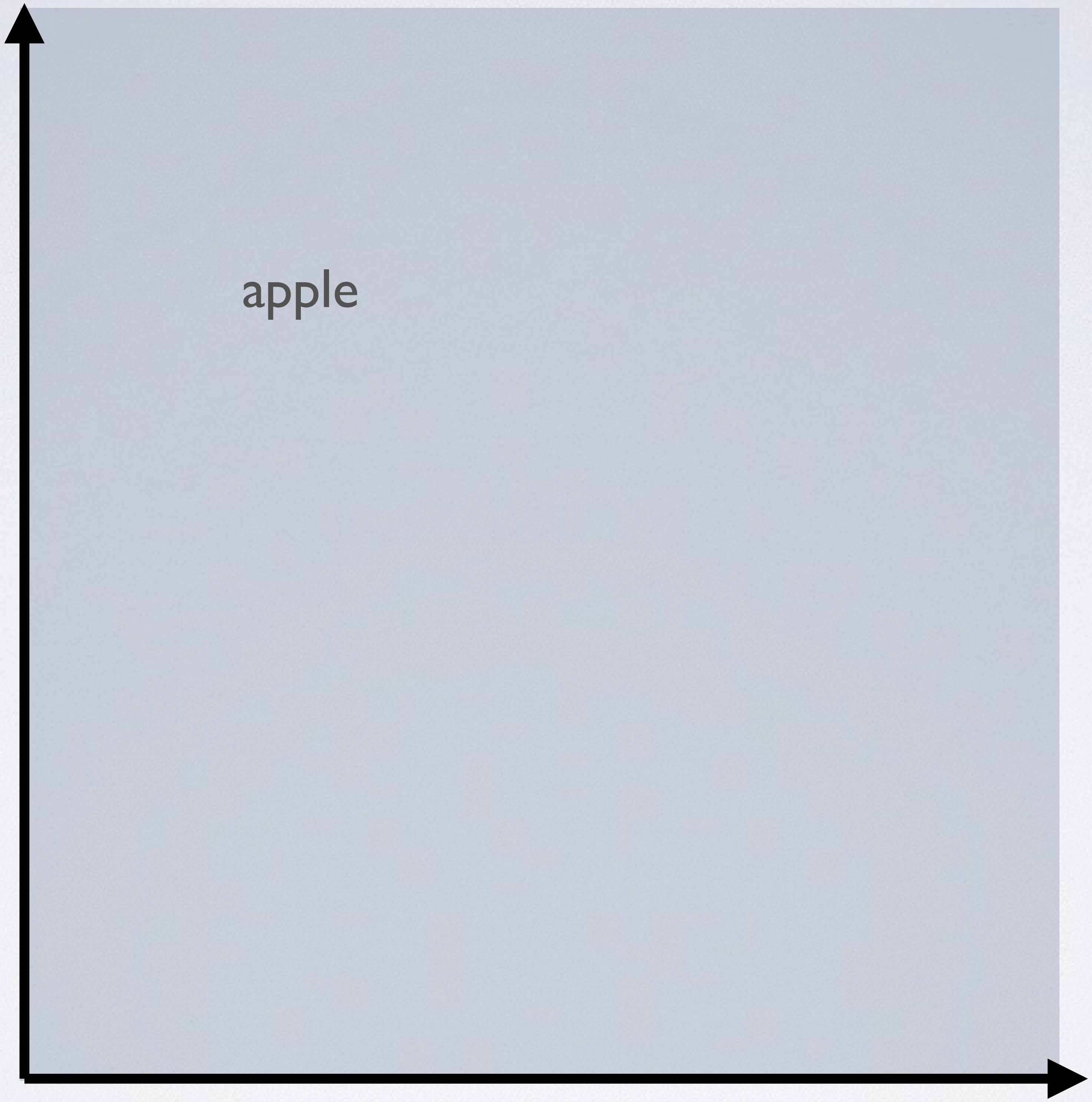
“king” - “man” + “woman” ~ queen

Multimodal Regularities

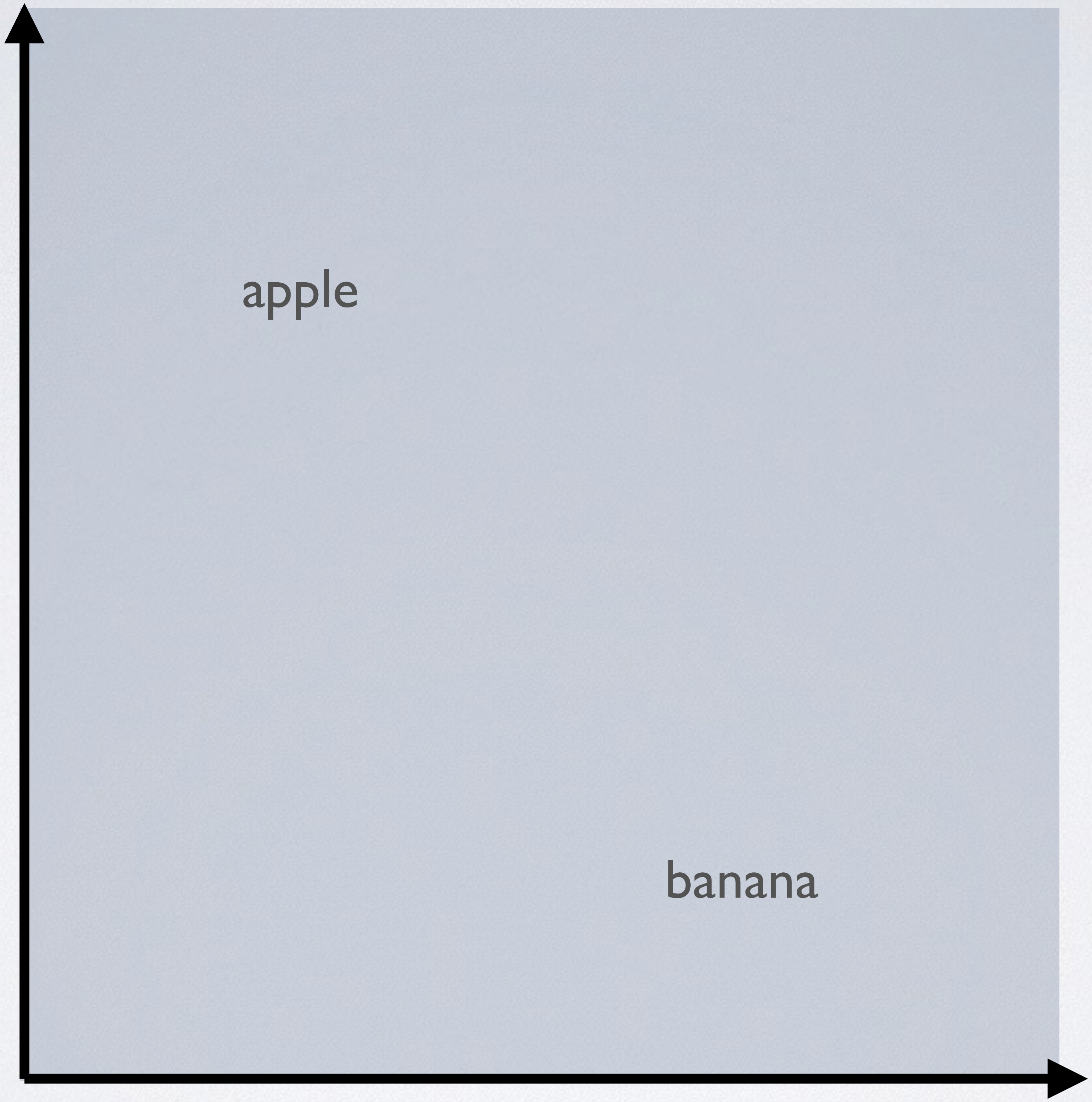
Multimodal Regularities



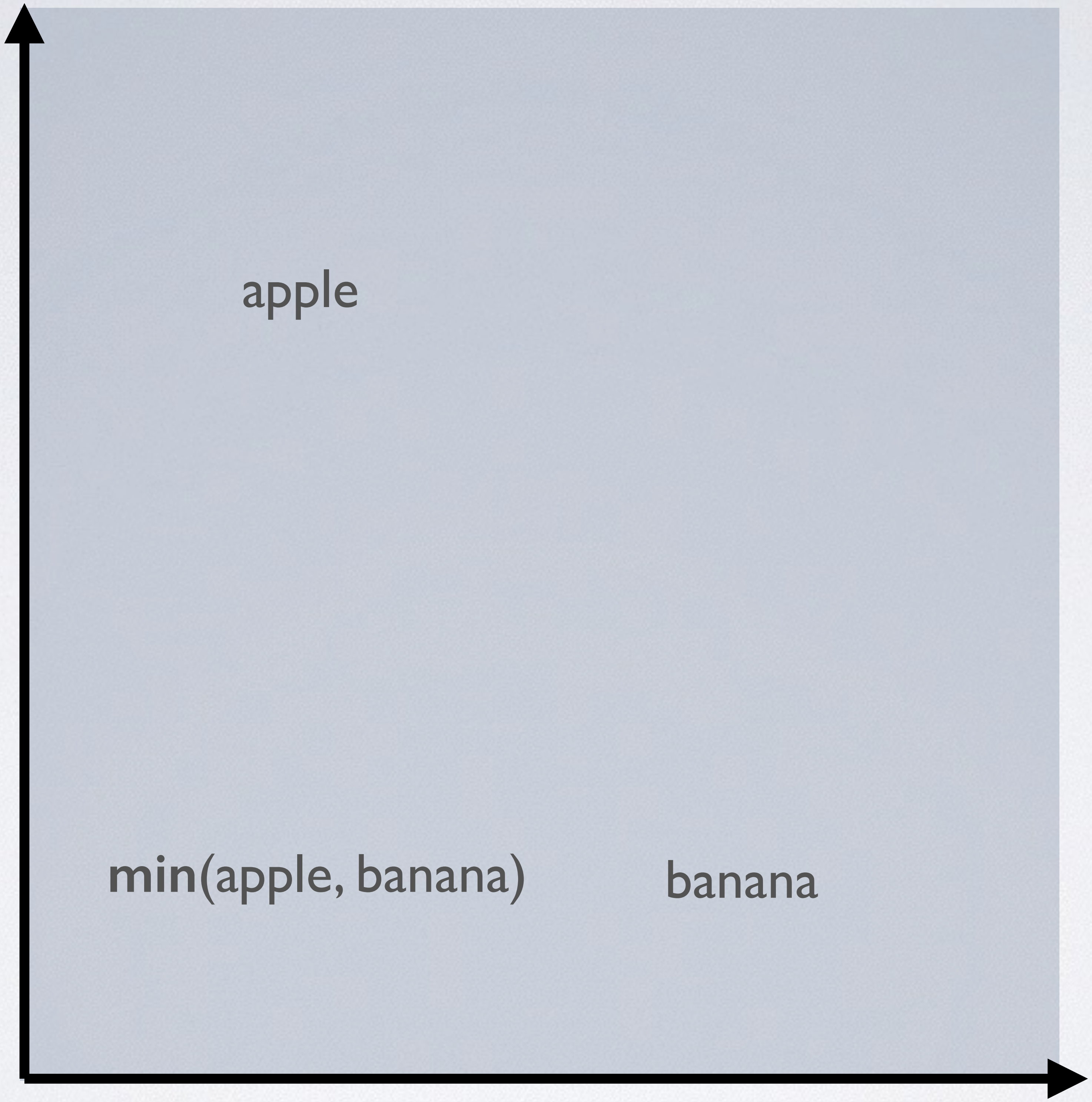
Multimodal Regularities



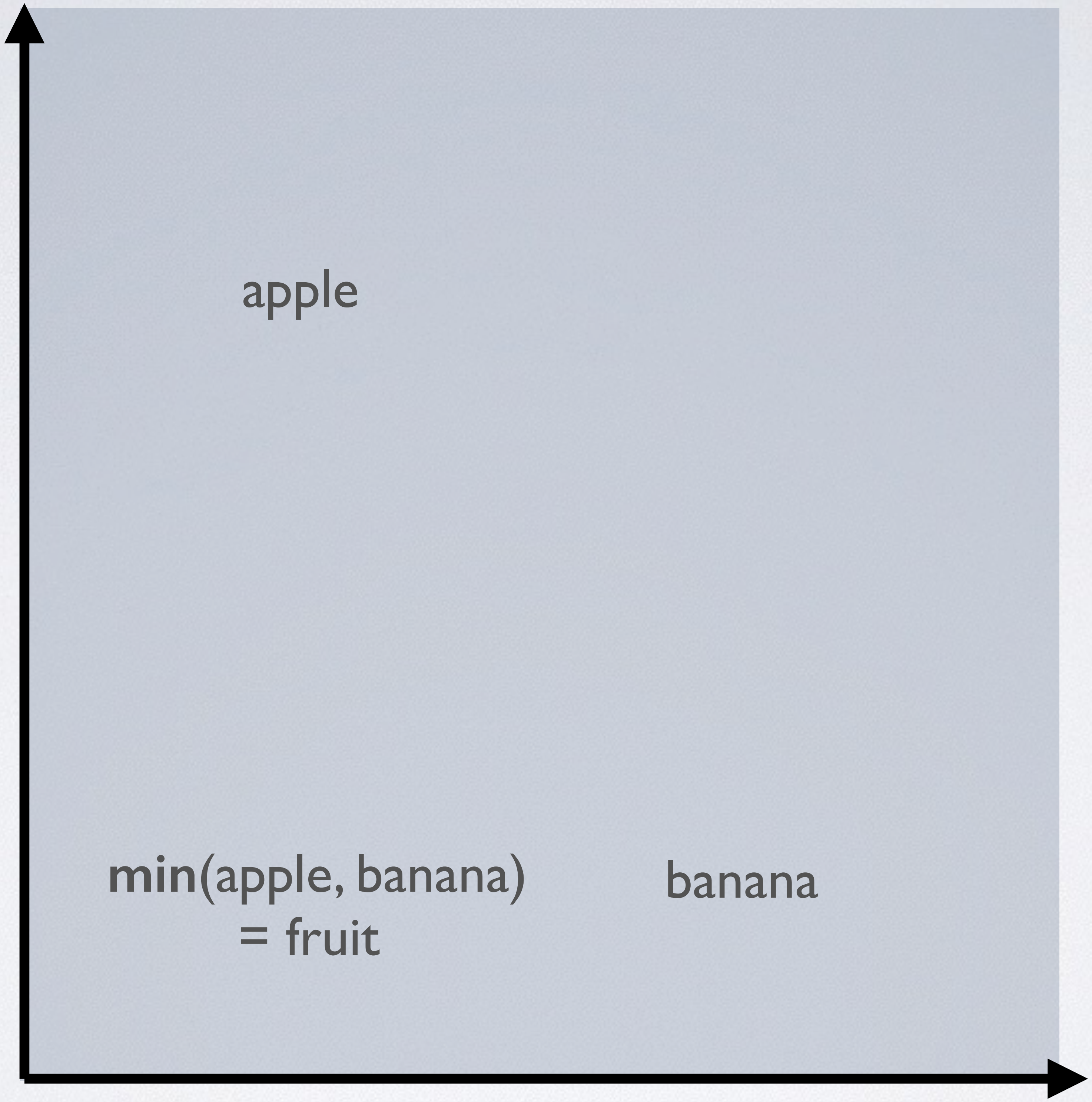
Multimodal Regularities



Multimodal Regularities



Multimodal Regularities

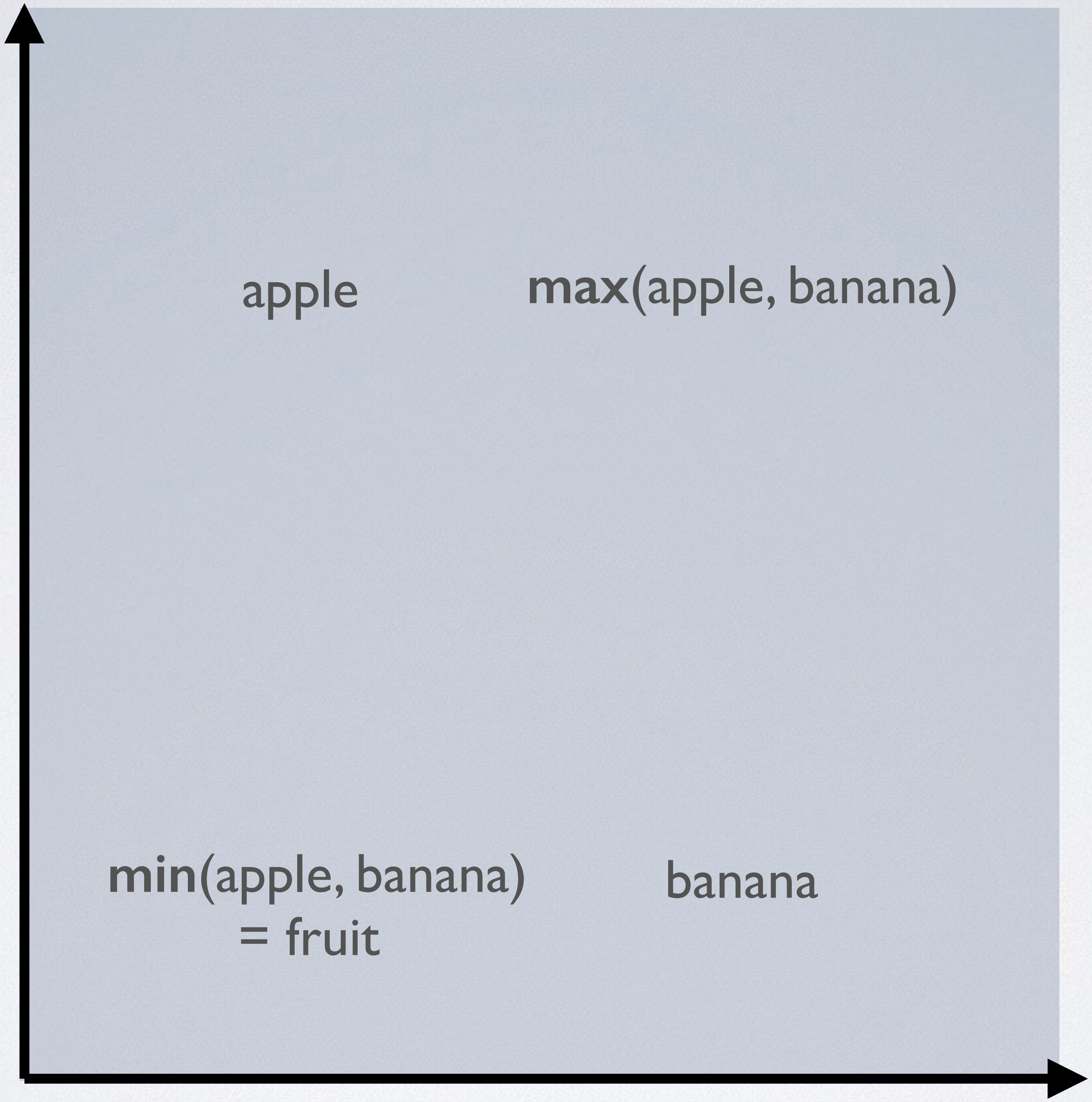


apple

min(apple, banana)
= fruit

banana

Multimodal Regularities



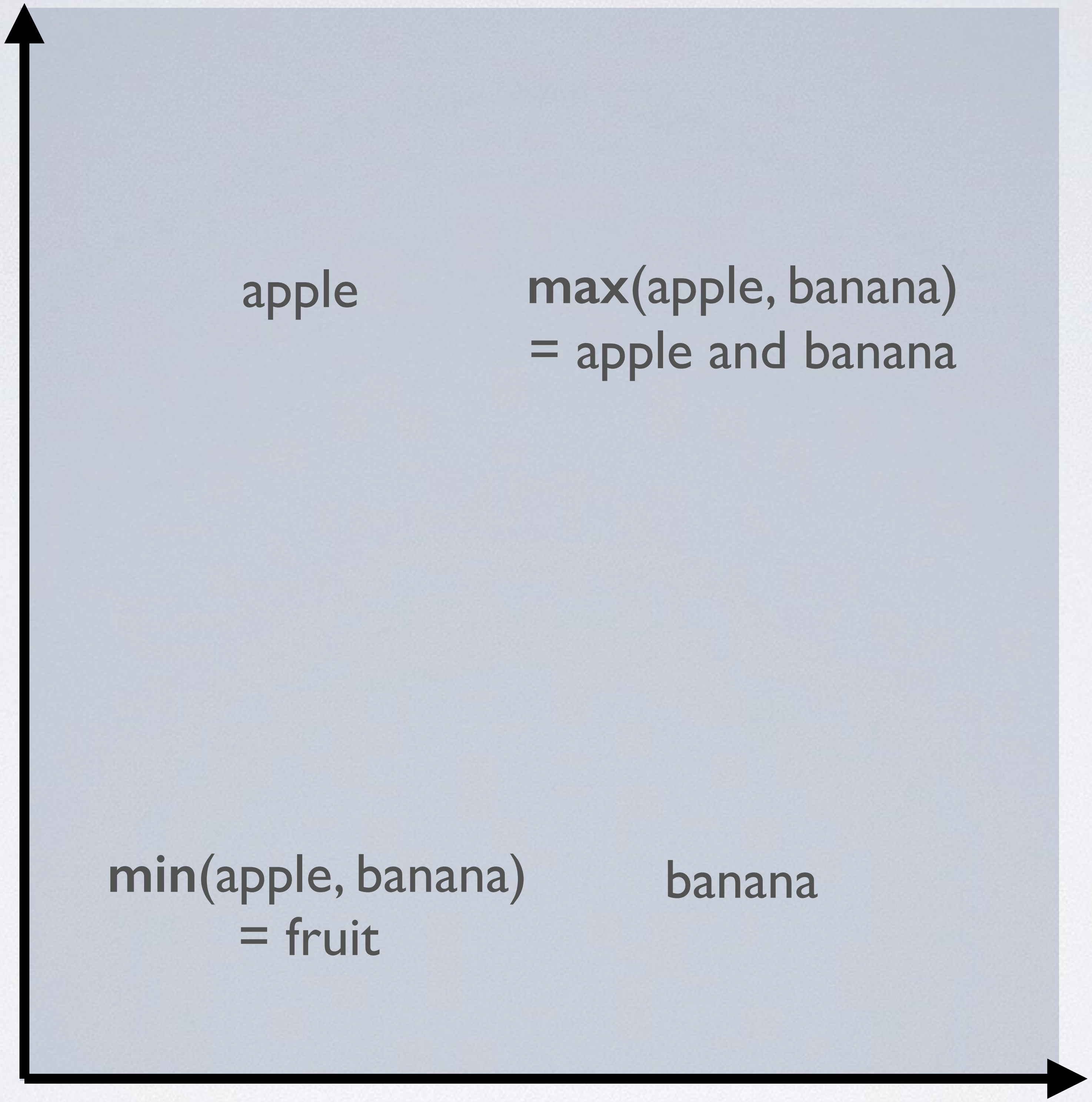
apple

max(apple, banana)

min(apple, banana)
= fruit

banana

Multimodal Regularities



apple

max(apple, banana)
= apple and banana

min(apple, banana)
= fruit

banana

Query

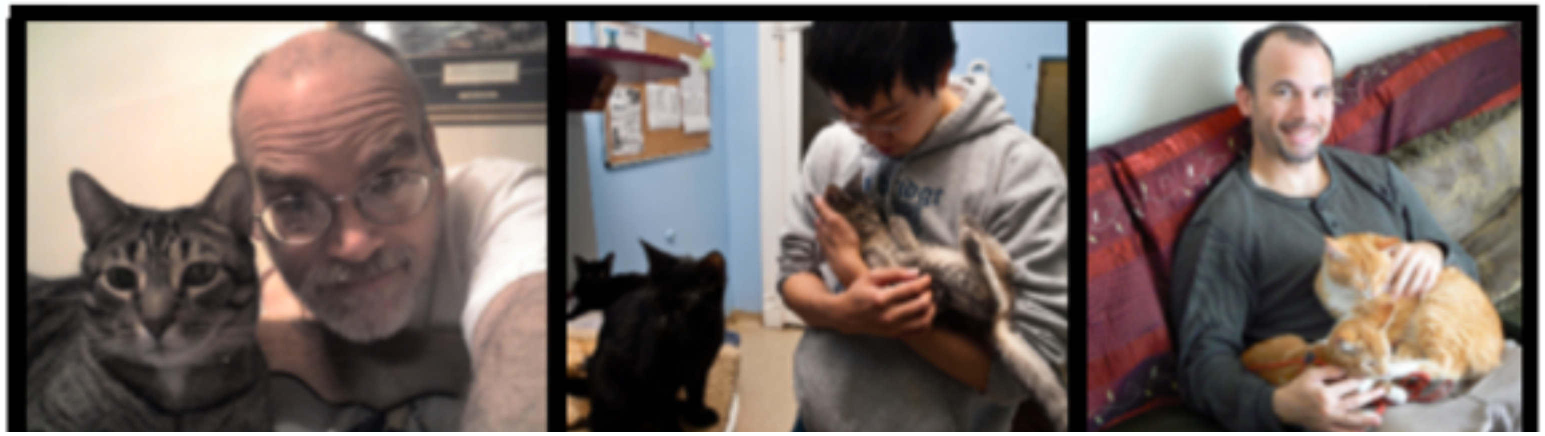
Nearest non-query images in COCO train

max(“man”, “cat”)

Query

max(“man”, “cat”)

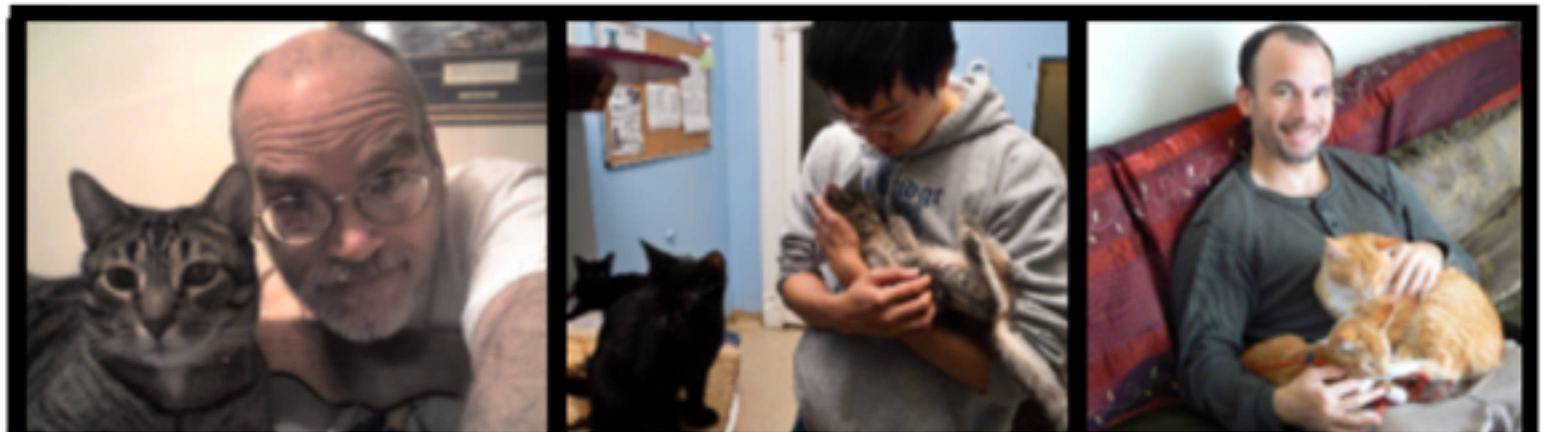
Nearest non-query images in COCO train



Query

Nearest non-query images in COCO train

max(“man”, “cat”)

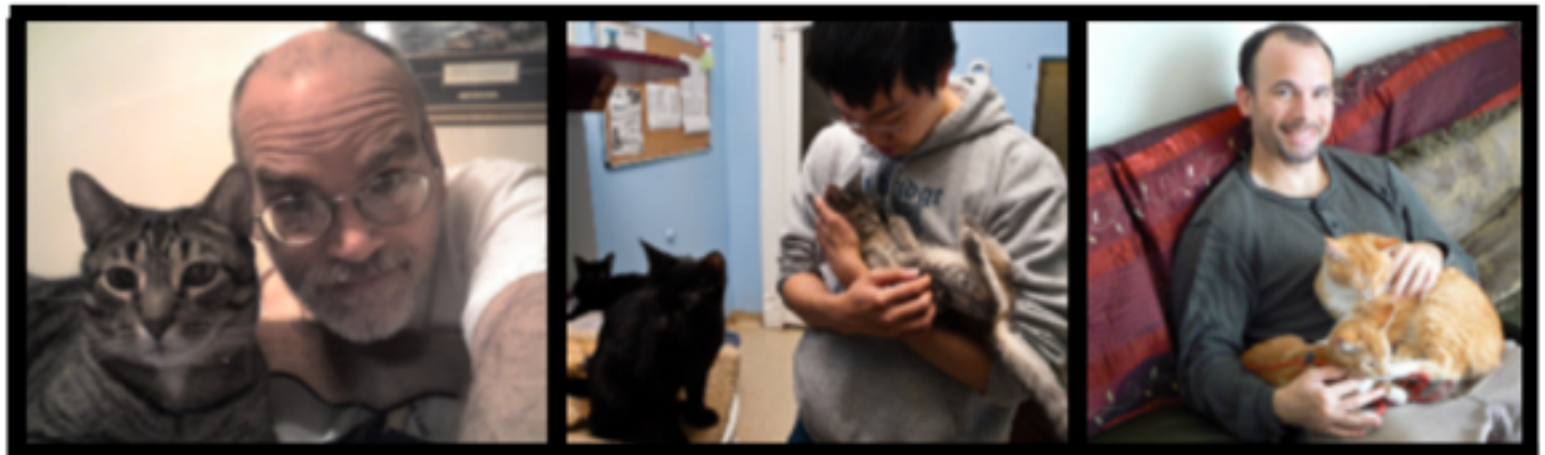


max(“black dog”, “park”)

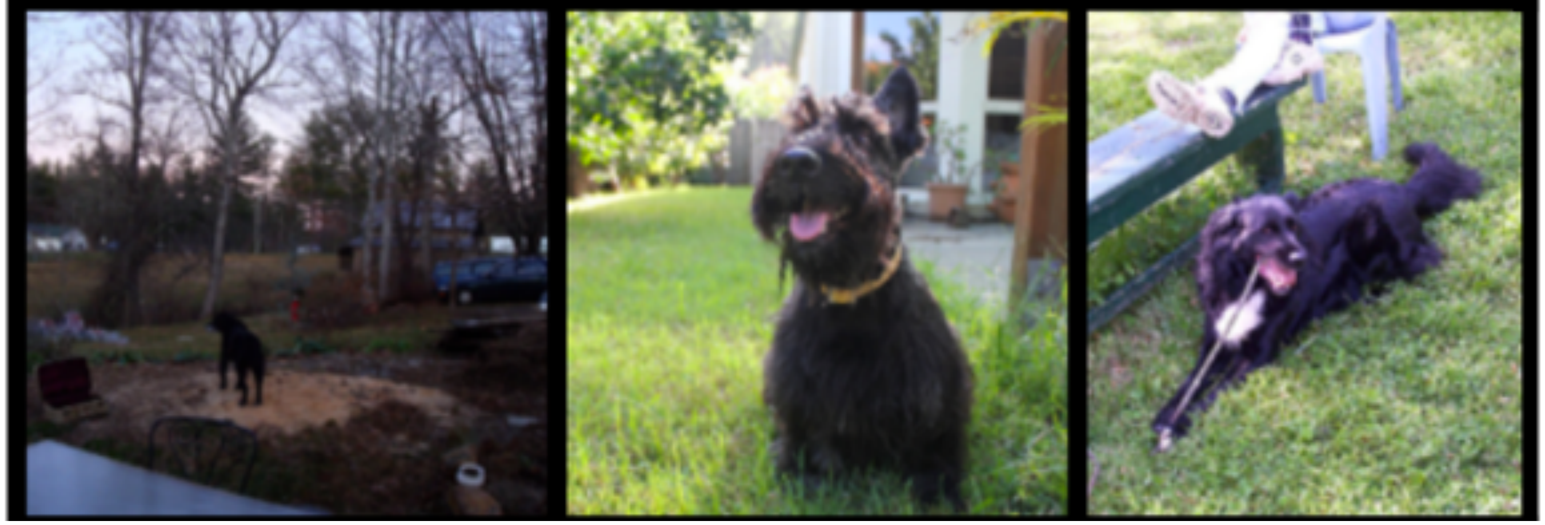
Query

Nearest non-query images in COCO train

max("man", "cat")

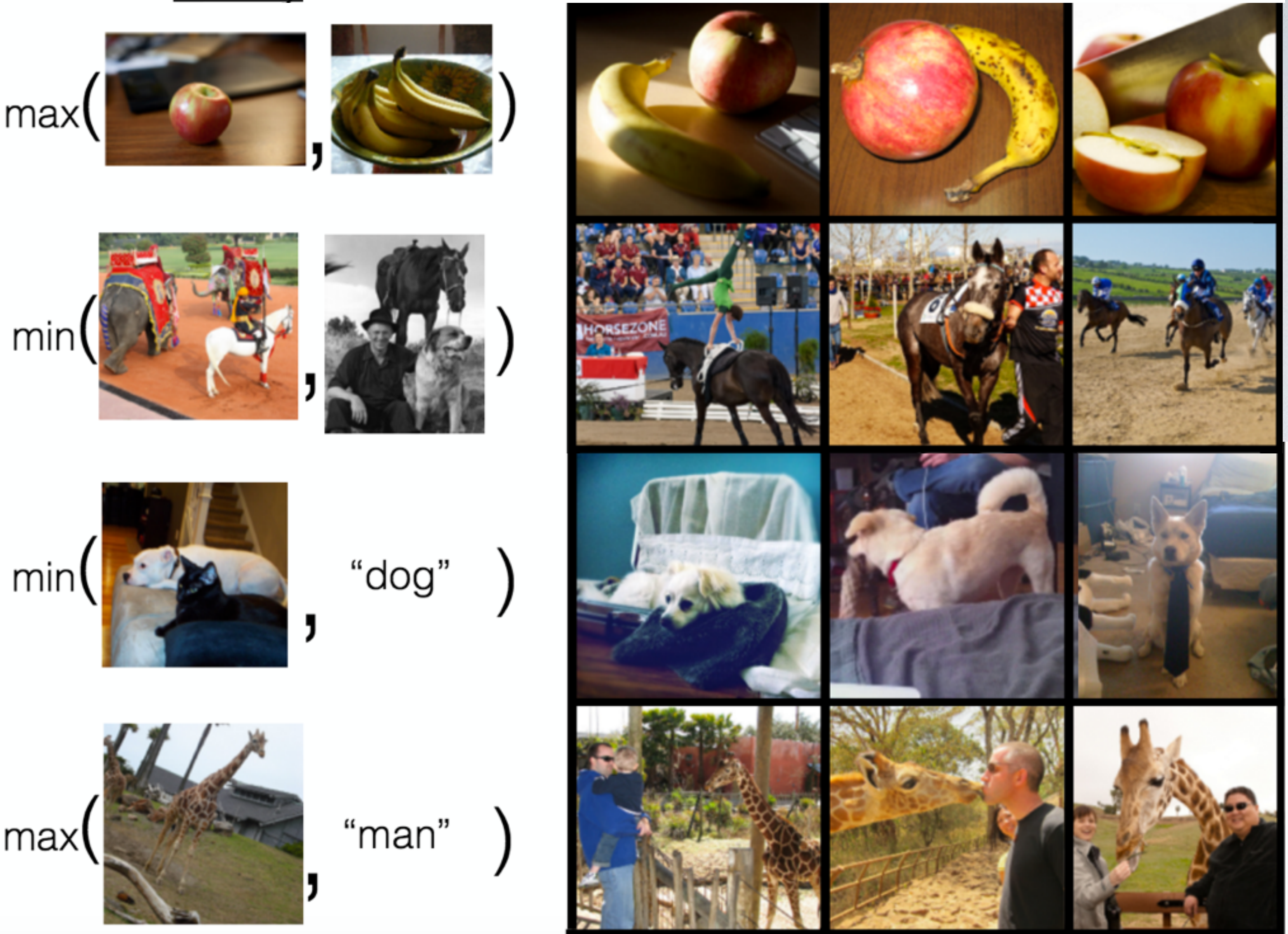


max("black dog", "park")



Query

Nearest non-query images in COCO train



Conclusions

Conclusions

- The relationship between images and language forms a partial order.

Conclusions

- The relationship between images and language forms a partial order.
- To efficiently learn partial orders from data, use order-preserving mappings between the domain and an ordered vector space.

Conclusions

- The relationship between images and language forms a partial order.
- To efficiently learn partial orders from data, use order-preserving mappings between the domain and an ordered vector space.

Code available at github.com/ivendrov/order-embedding