Ridge regression, hubness, and zero-shot learning

Yutaro Shigeto[1] Ikumi Suzuki[2] Kauzo Hara[3]
Masashi Shimbo[1] Yuji Matsumoto[1]

1: Nara Institute of Science and Technology
2: The Institute of Statistical Mathematics
3: National Institute of Genetics

Zero-shot learning [Larochelle+, '08]

Active research topic in ML, CV, NLP

Many applications:

- Image labeling
- Bilingual lexicon extraction

+ Many other cross-domain matching tasks

ZSL is a type of multi-class classification

...but classifier has to predict
labels **not** appearing in training set

Standard classification task

$$Y_{\text{train}} = \{\text{gorilla, lion, tiger}\}$$
$$Y_{\text{test}} = \{\text{gorilla, lion, tiger}\}$$

$\rightarrow$ $Y_{\text{train}} = Y_{\text{test}}$

ZSL task

$$Y_{\text{train}} = \{\text{gorilla, lion, tiger}\}$$
$$Y_{\text{test}} = \{\text{chimpanzee, leopard}\}$$

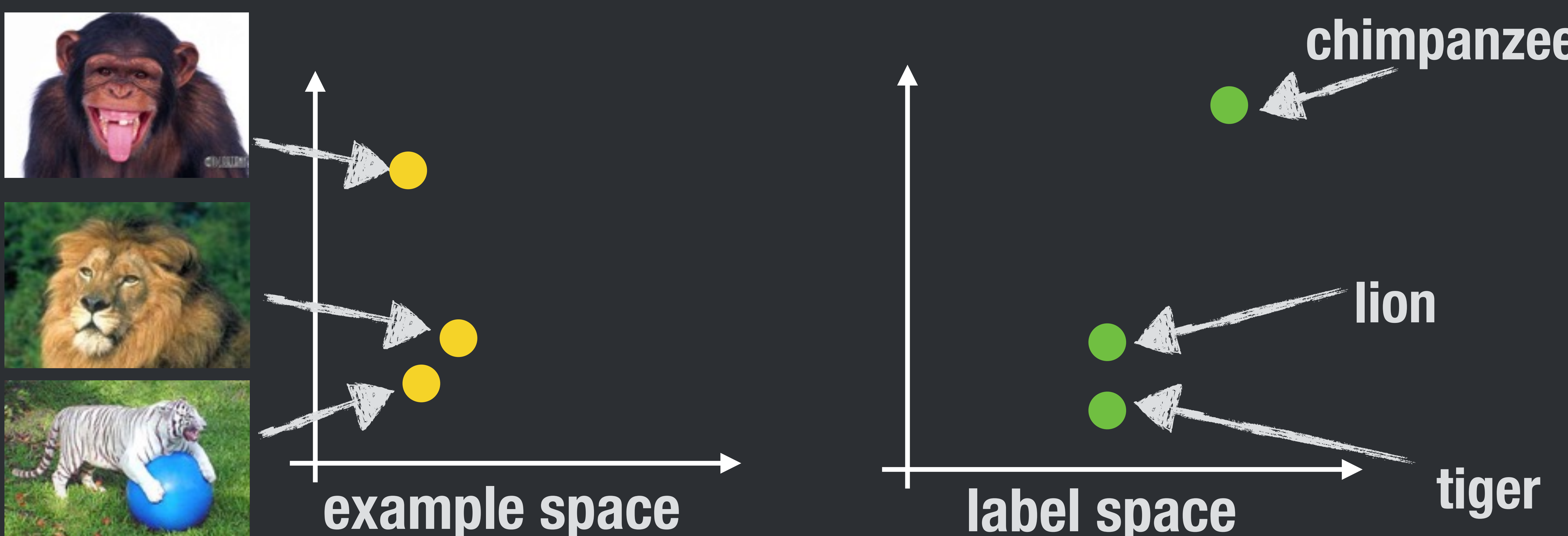
$\rightarrow$ $Y_{\text{train}} \cap Y_{\text{test}} = \emptyset$

Pre-processing: Label embedding

Labels are embedded in metric space

$$(\mathbf{x}_i, \mathbf{y}_i), i = 1 \cdots, N$$

Examples and labels = both vectors

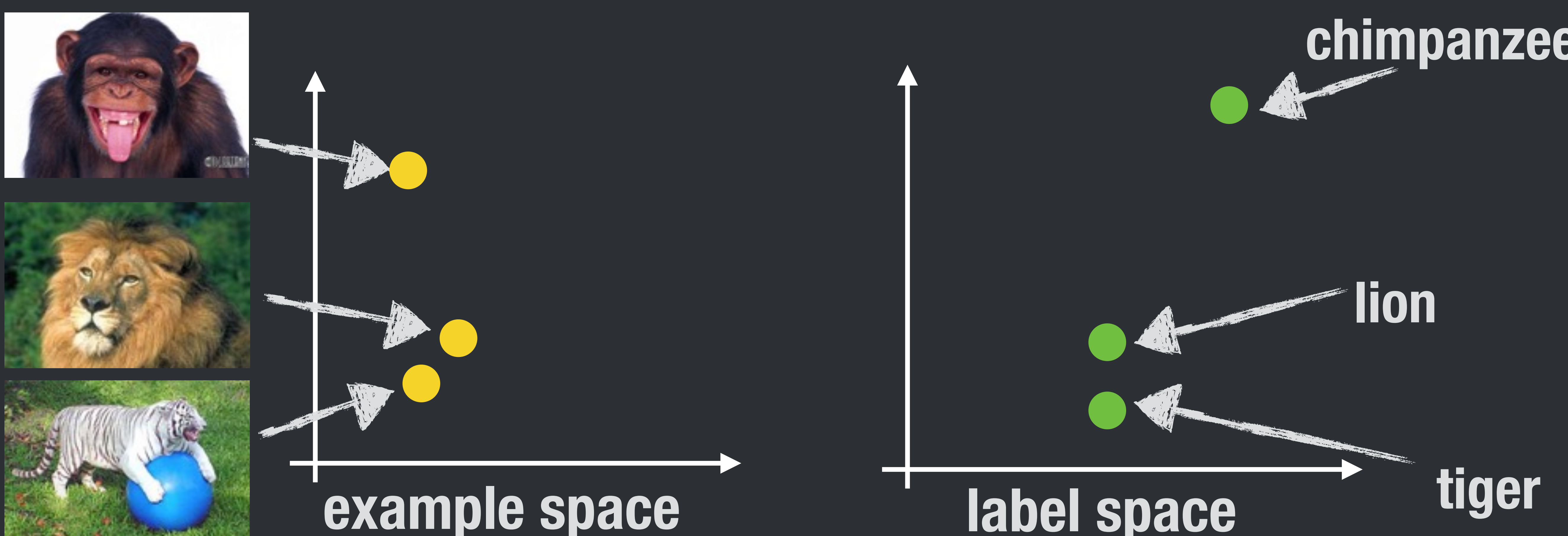


Photos provided by Animal with Attributes

Regression-based ZSL: Training

Find a matrix **M** that projects examples into label space

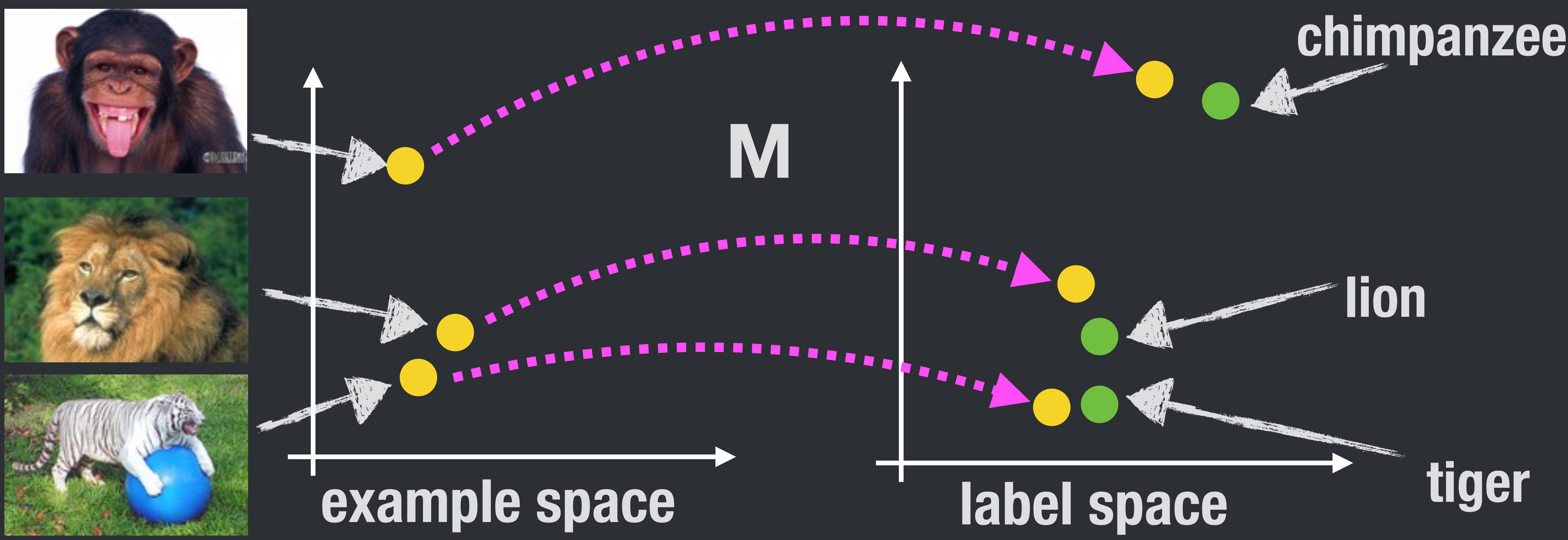
$$\min_{\mathbf{M}} \sum_{i=1}^{N} \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|^2 + \lambda \|\mathbf{M}\|_F^2$$



Regression-based ZSL: Training

Find a matrix **M** that projects examples into label space

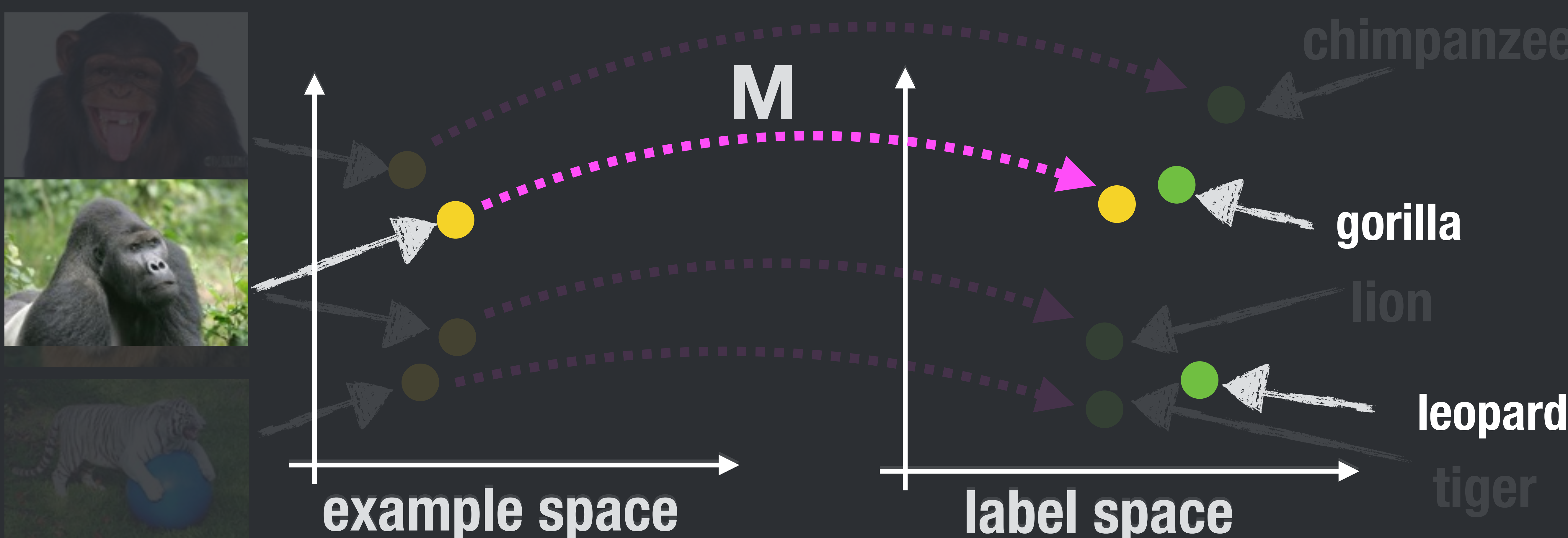
$$\min_{\mathbf{M}} \sum_{i=1}^{N} \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|^2 + \lambda \|\mathbf{M}\|_F^2$$



Regression-based ZSL: Prediction

To predict the label of a test example,

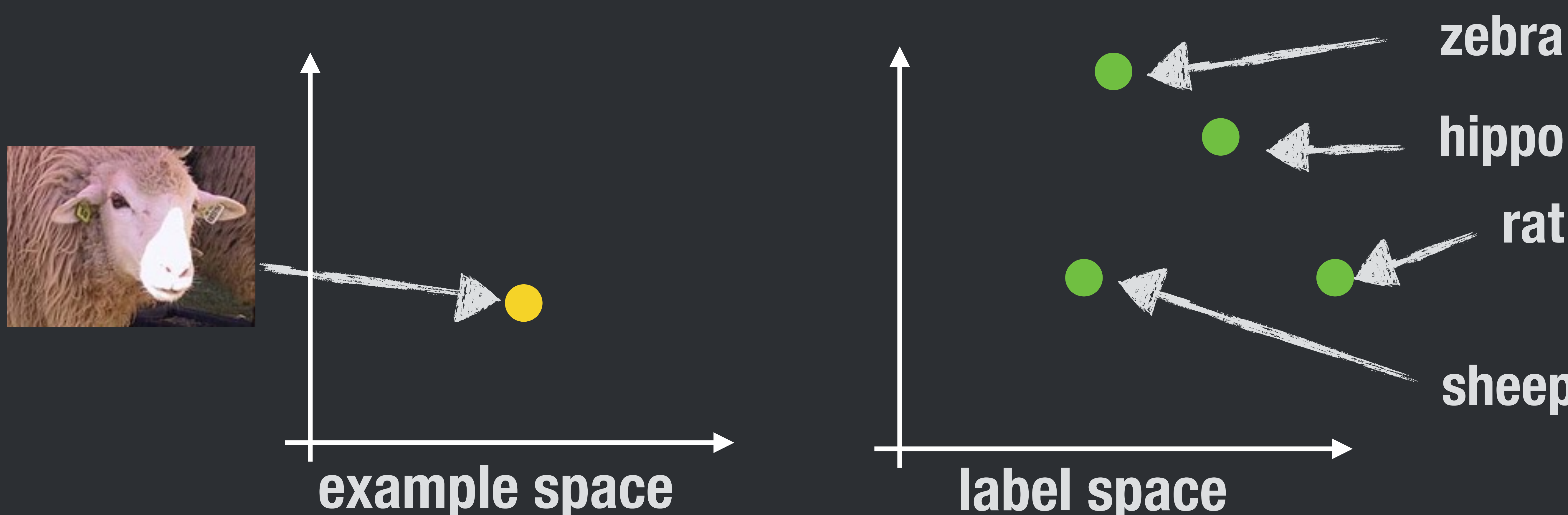
1. project the example into label space, using matrix **M**
2. find the nearest label



Hubness: Problem in current approach

[Dinu and Baroni 15; see also Radovanovic 10]

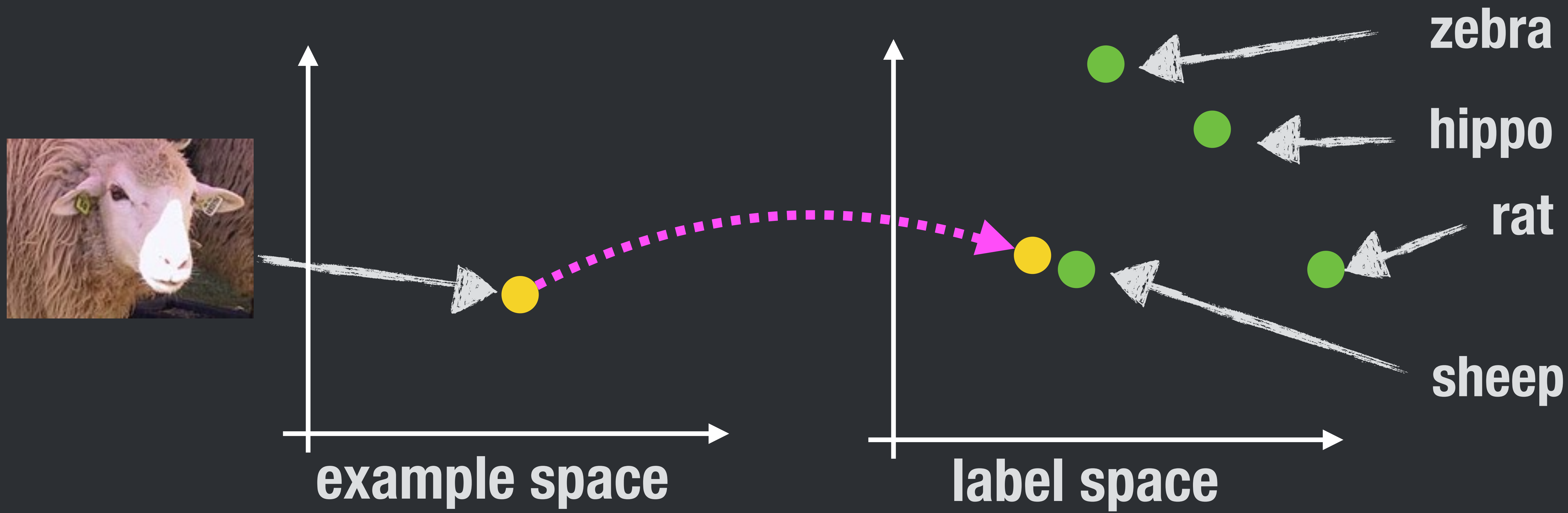
Classifier frequently predicts the same labels (“hubs”)



Hubness: Problem in current approach

[Dinu and Baroni 15; see also Radovanovic 10]

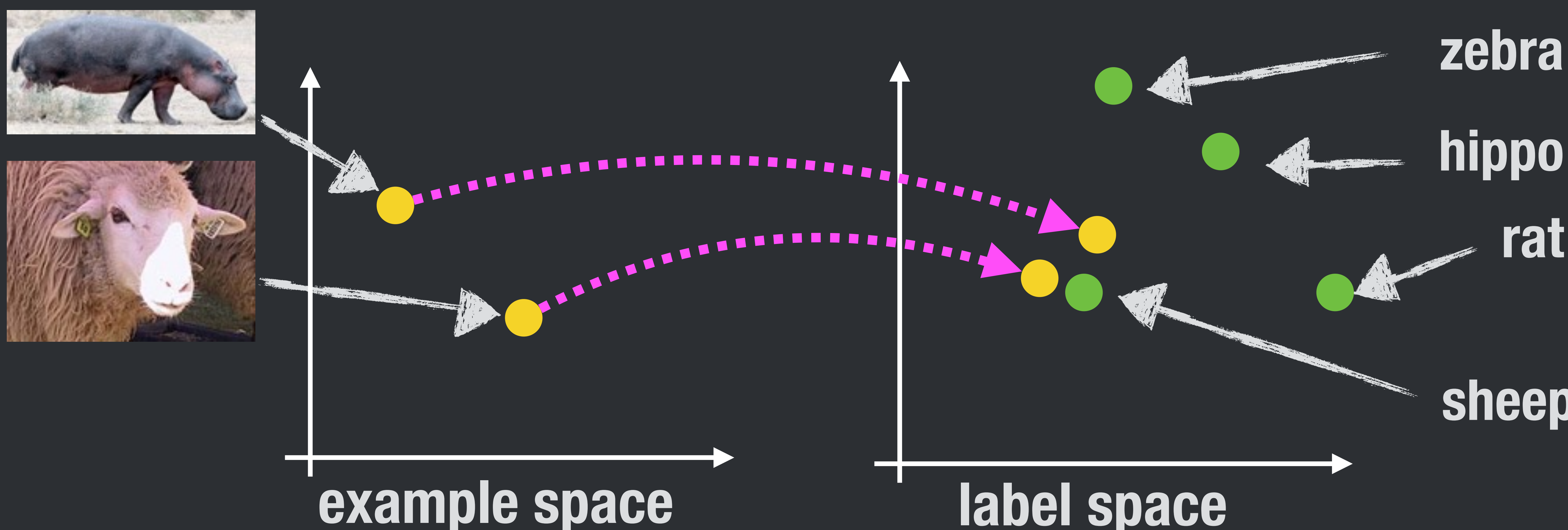
Classifier frequently predicts the same labels (“hubs”)



Hubness: Problem in current approach

[Dinu and Baroni 15; see also Radovanovic 10]

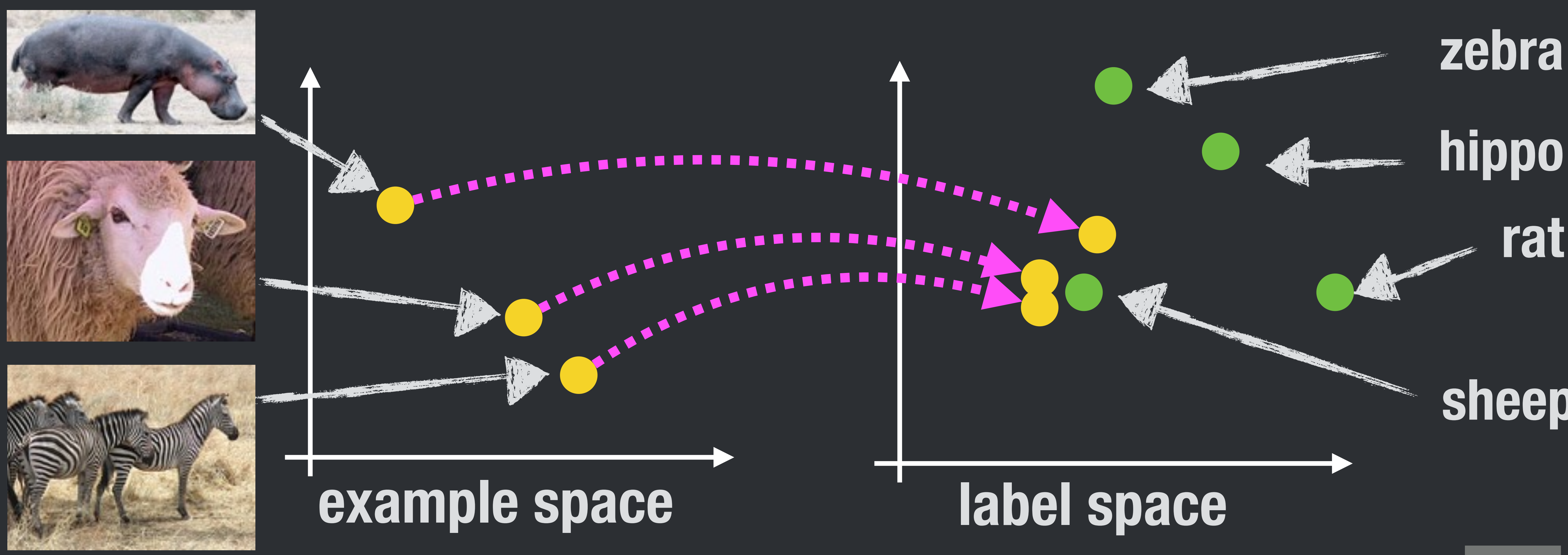
Classifier frequently predicts the same labels (“hubs”)



Hubness: Problem in current approach

[Dinu and Baroni 15; see also Radovanovic 10]

Classifier frequently predicts the same labels (“hubs”)



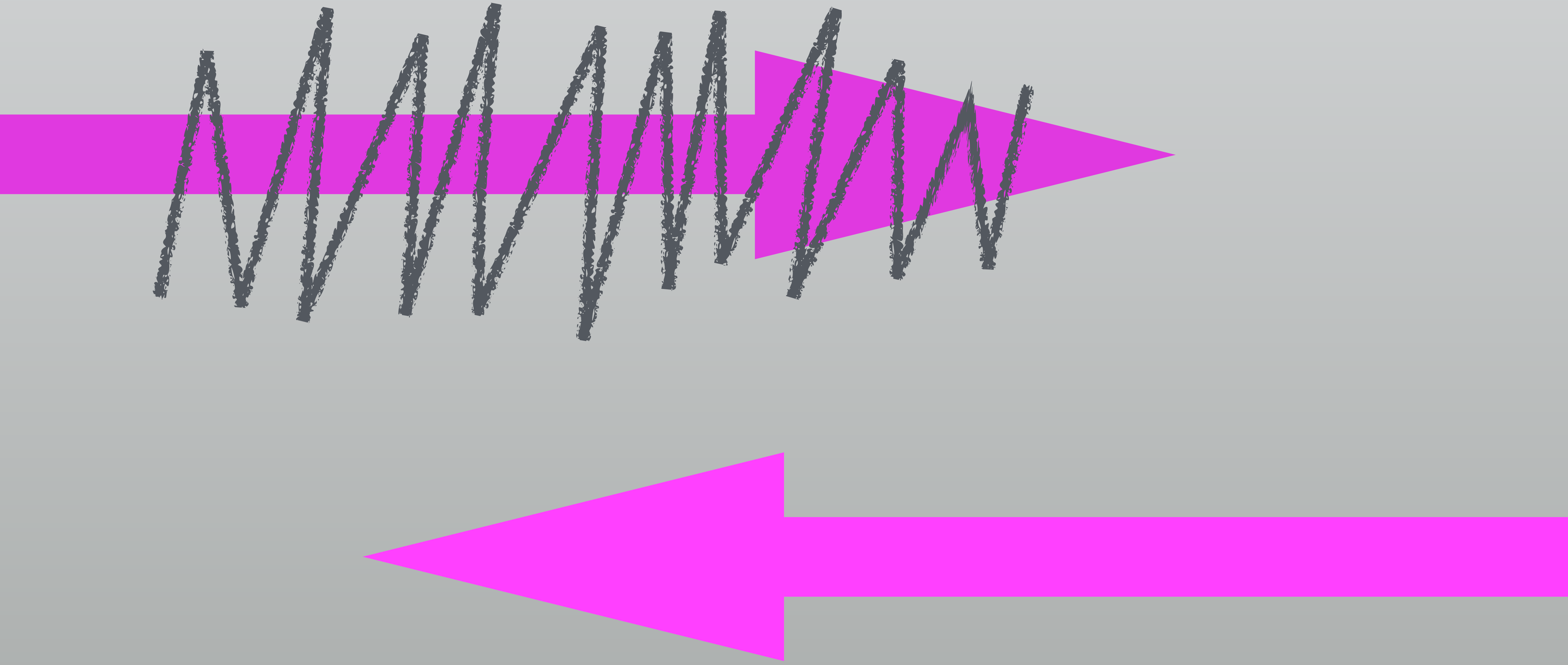
Problem with current regression approach:

Learned classifier frequently predicts the same labels (Emergence of “hub” labels)

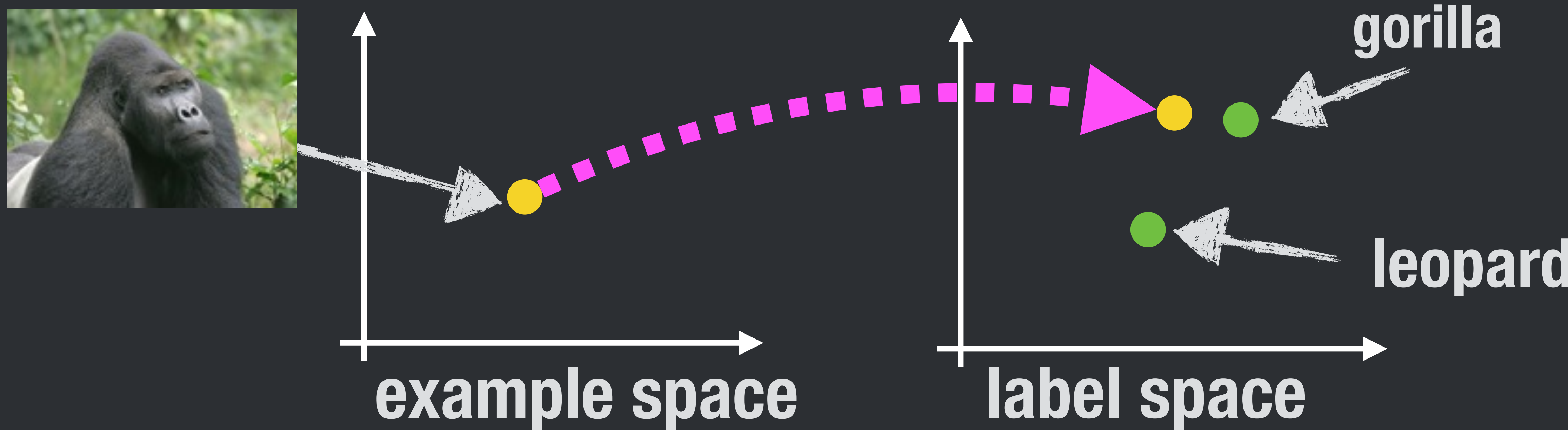
Research objective:

Investigate how to reduce hubness in regression-based ZSL, and to improve classification accuracy

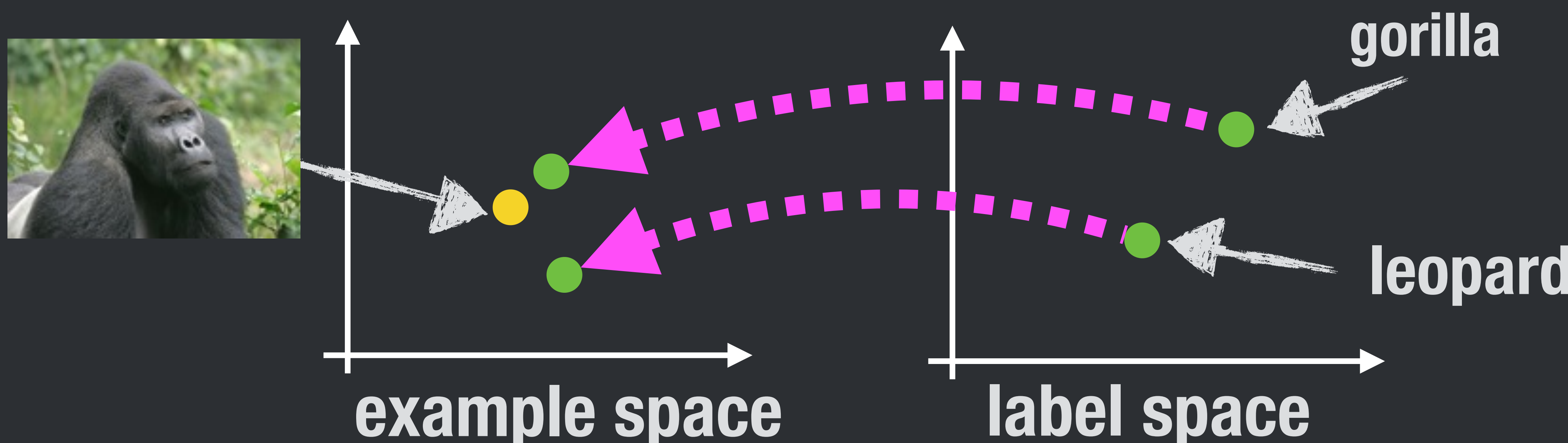
Proposed approach



Current approach: $\min_{\mathbf{M}} \sum \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|^2 + \lambda \|\mathbf{M}\|_F^2$



Proposed approach: $\min_{\mathbf{M}} \sum \|\mathbf{x}_i - \mathbf{M}\mathbf{y}_i\|^2 + \lambda \|\mathbf{M}\|_F^2$



Synthetic data result

	Current		Proposed
Hubness (N_1 skewness)	**24.2**	→	**0.5**
Accuracy	**13.8**	→	**87.6**

Proposed approach **reduces hubness** and **improves accuracy**

Why proposed approach reduces hubness

Argument for our proposal relies on two concepts

Spatial centrality
of data distributions

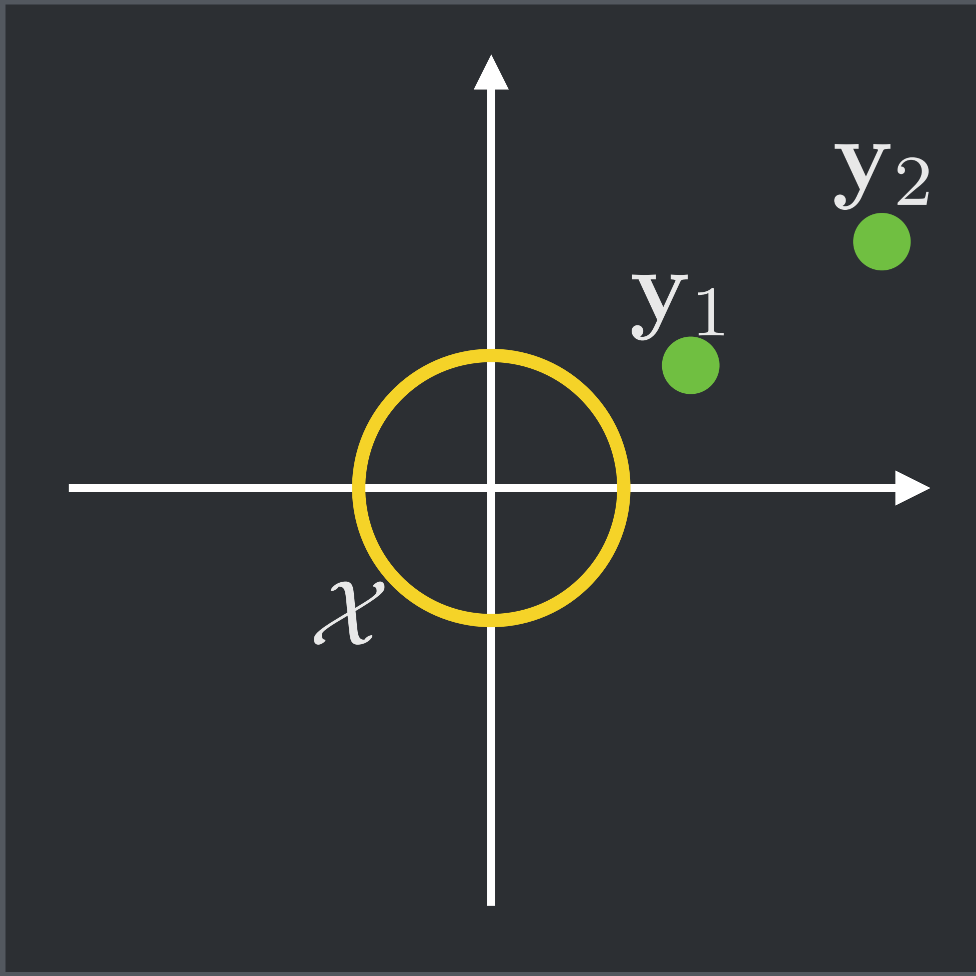
Shrinkage
in regression

“Spatial centrality” [Radovanović+ 10]

$\mathcal{X}$: query distribution (zero mean)

Fixed objects $\mathbf{y}_1$, $\mathbf{y}_2$ with

$$\|\mathbf{y}_1\|^2 < \|\mathbf{y}_2\|^2$$



“Spatial centrality” [Radovanović+ 10]

$\mathcal{X}$: query distribution (zero mean)

Fixed objects $\mathbf{y}_1$, $\mathbf{y}_2$ with

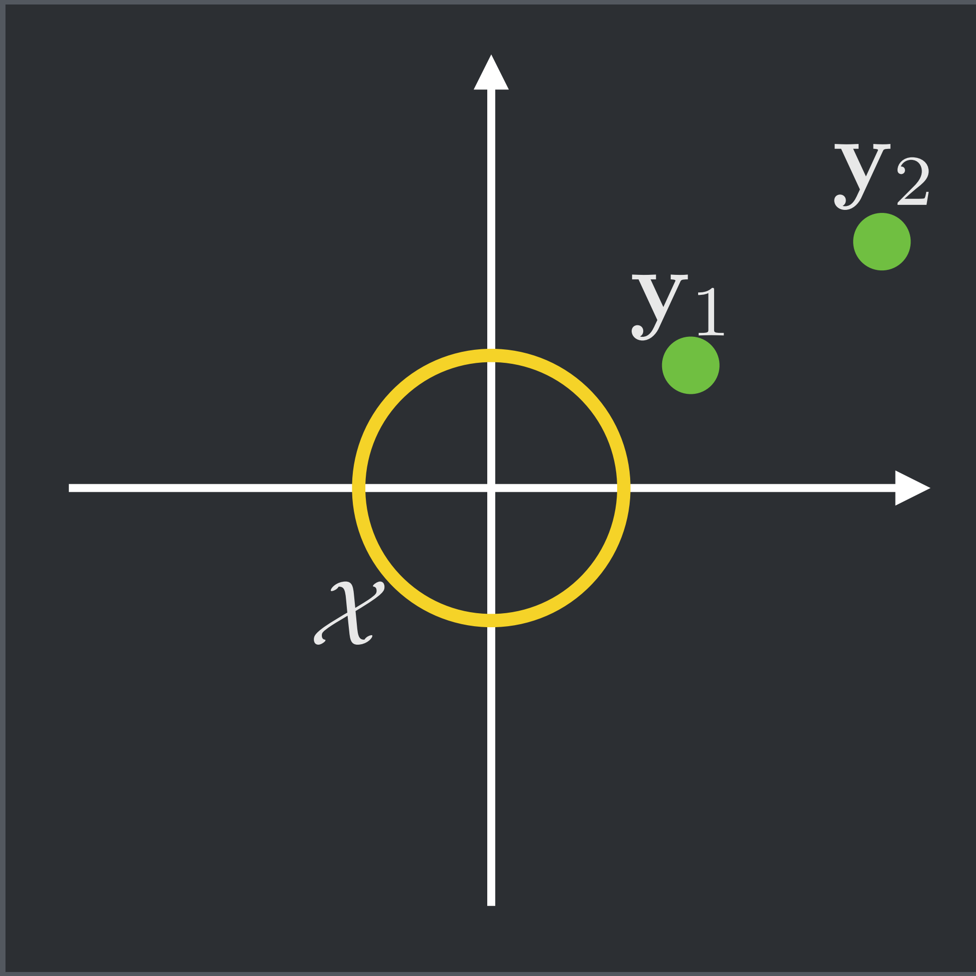
$$\|\mathbf{y}_1\|^2 < \|\mathbf{y}_2\|^2$$

Then it can be shown that

$$\mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_1\|^2] < \mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_2\|^2]$$

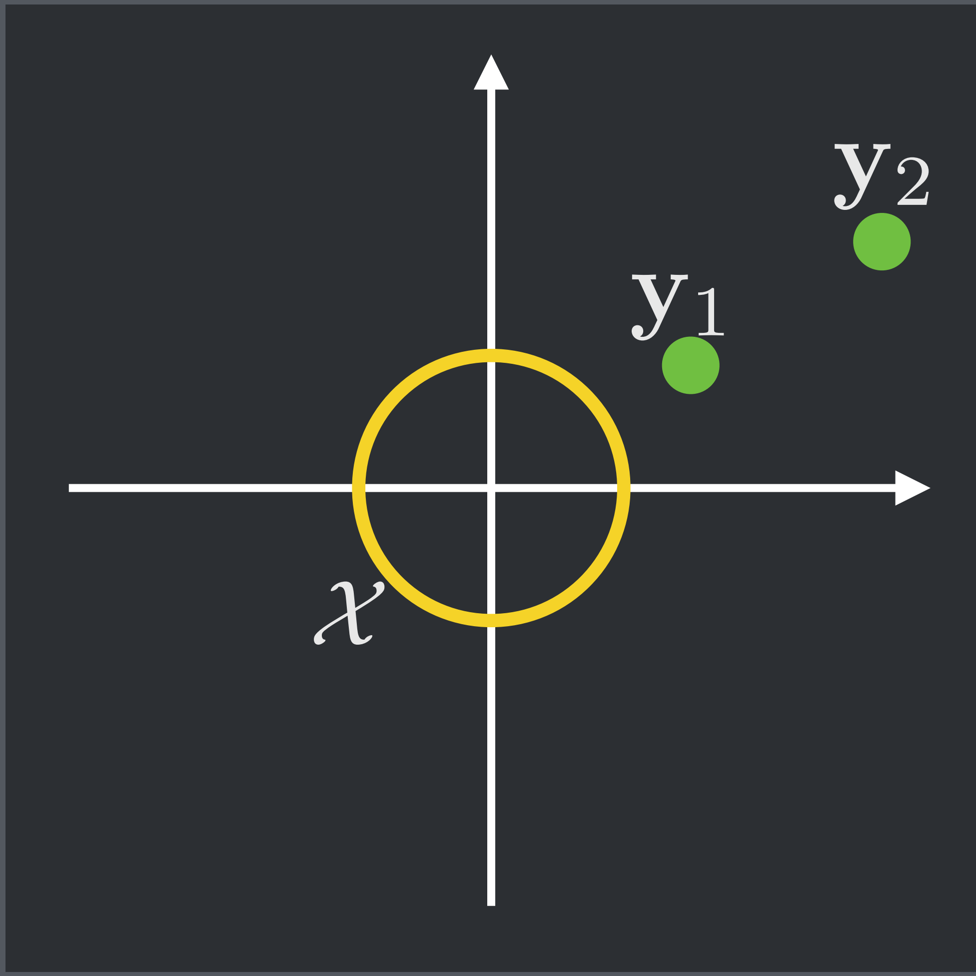
$\mathbf{y}_1$ is more likely to be closer to $\mathbf{x} \sim \mathcal{X}$

i.e. $\mathbf{y}_1$ more likely to be a hub



“Spatial centrality” [Radovanović+ 10]

$\mathcal{X}$: query distribution (zero mean)



Fixed objects $\mathbf{y}_1$, $\mathbf{y}_2$ with

$$\|\mathbf{y}_1\|^2 < \|\mathbf{y}_2\|^2$$

Then it can be shown that

$$\mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_1\|^2] < \mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_2\|^2]$$

Because this holds for any pair $\mathbf{y}_1$ and $\mathbf{y}_2$, objects closest to the origin tend to be hubs

This bias is called “spatial centrality.”

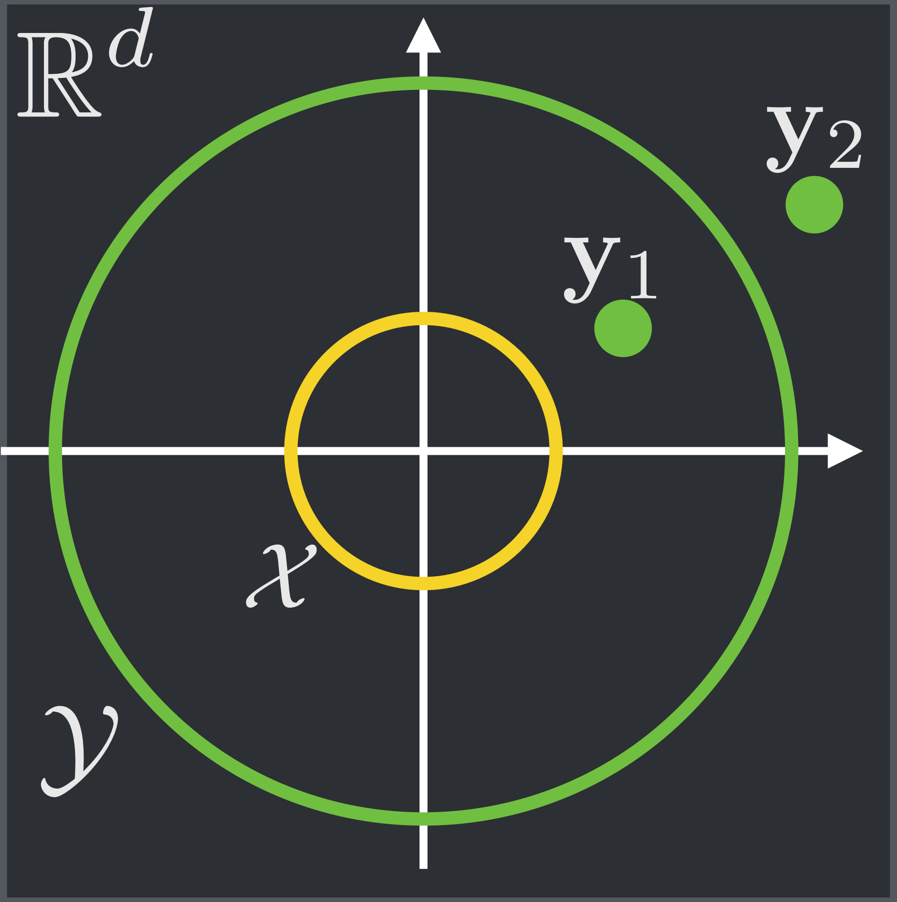
Degree of spatial centrality

Further assume distribution of $\mathcal{Y}$

$$\mathcal{Y} = \mathcal{N}(\mathbf{0}, s^2\mathbf{I}_d)$$

and

$$\|\mathbf{y}_2\|^2 - \|\mathbf{y}_1\|^2 = \gamma\sqrt{\mathrm{Var}_{\mathcal{Y}}[\|\mathbf{y}\|^2]}$$



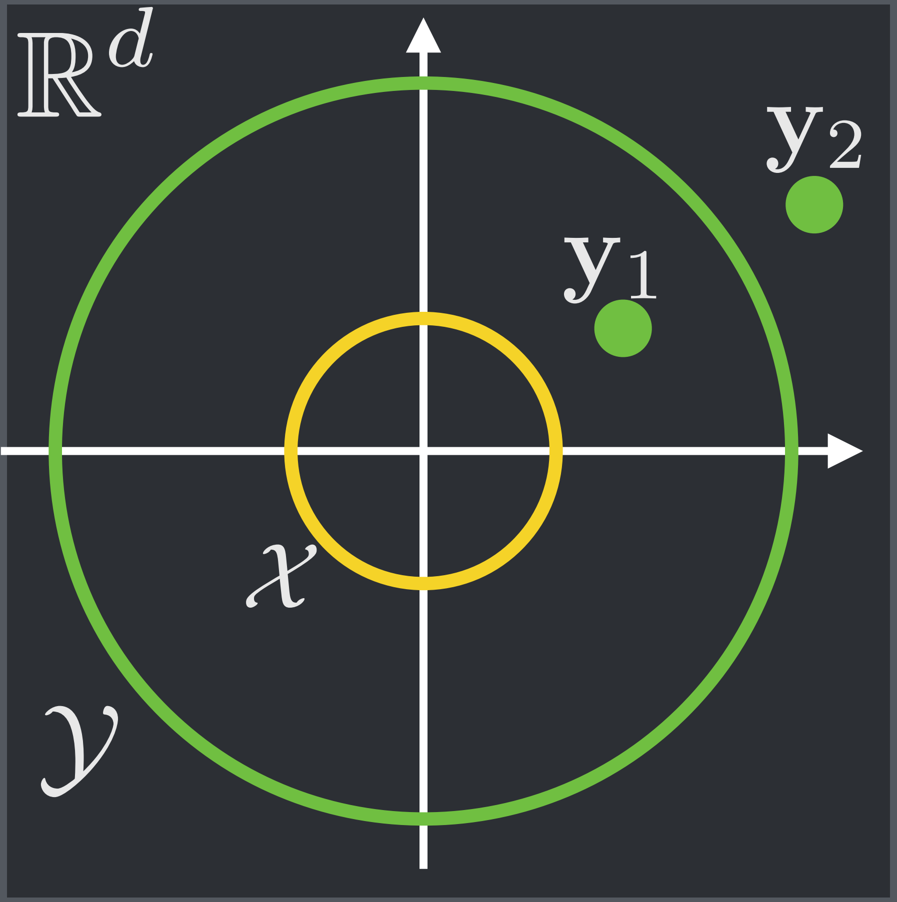
Degree of spatial centrality

Further assume distribution of $\mathcal{Y}$

$$\mathcal{Y} = \mathcal{N}(\mathbf{0}, s^2\mathbf{I}_d)$$

and

$$\|\mathbf{y}_2\|^2 - \|\mathbf{y}_1\|^2 = \gamma\sqrt{\mathrm{Var}_{\mathcal{Y}}[\|\mathbf{y}\|^2]}$$



We have

$$\mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_2\|^2] - \mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_1\|^2] = \gamma s^2\sqrt{2d}$$

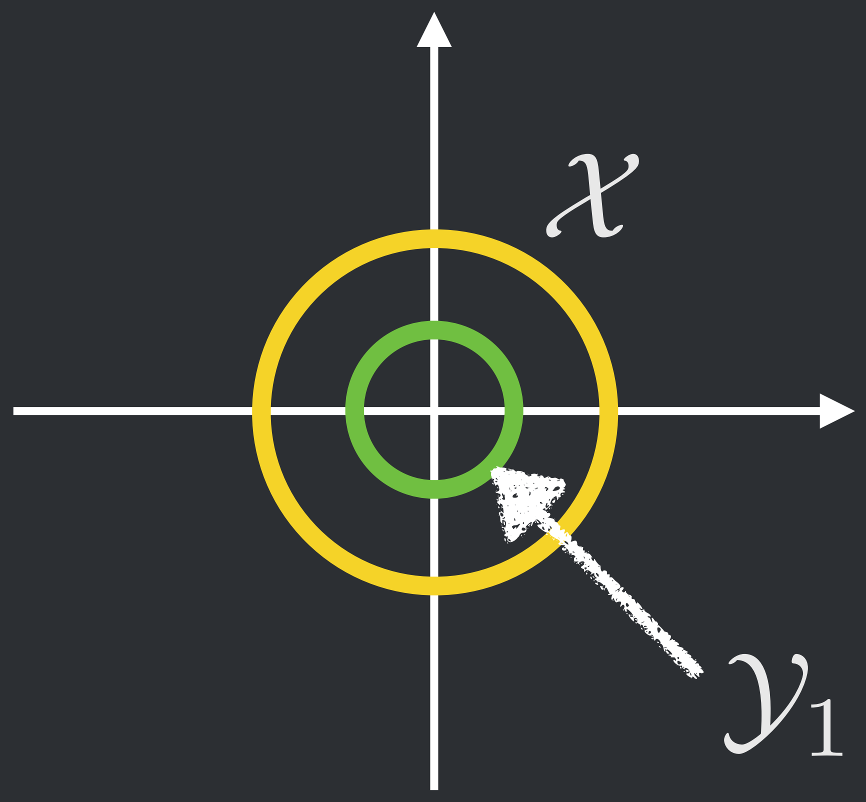
This formula quantifies the degree of spatial centrality:

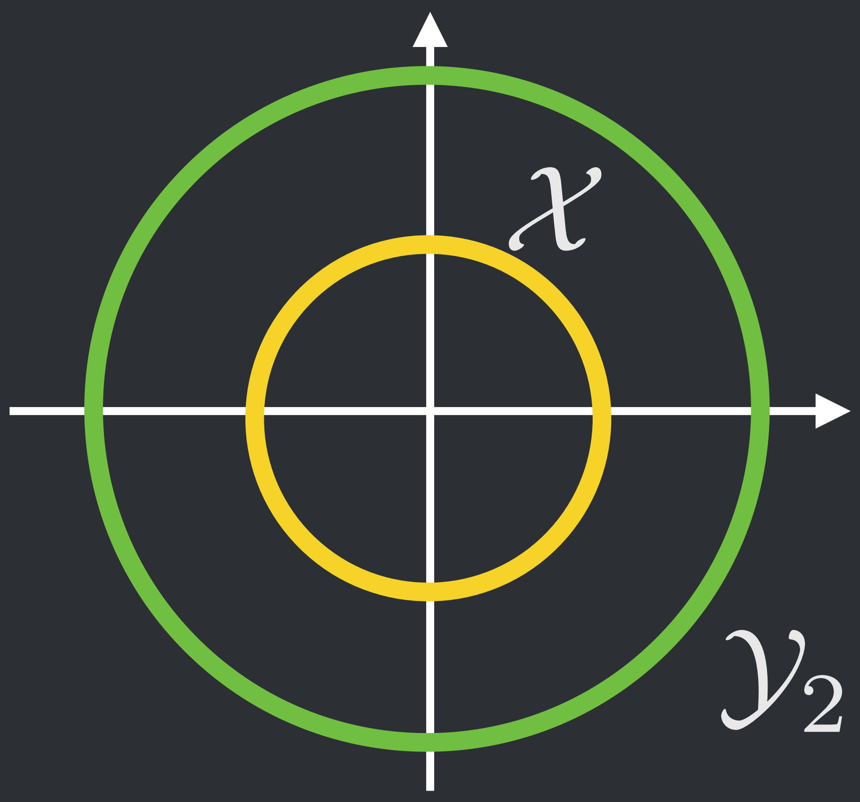
The smaller the variance s^2 of label distribution, the smaller the spatial centrality (= bias causing hubness)

Takeaway

To reduce hubness, label distribution Y with smaller variance should be preferred

Desirable	Not desirable

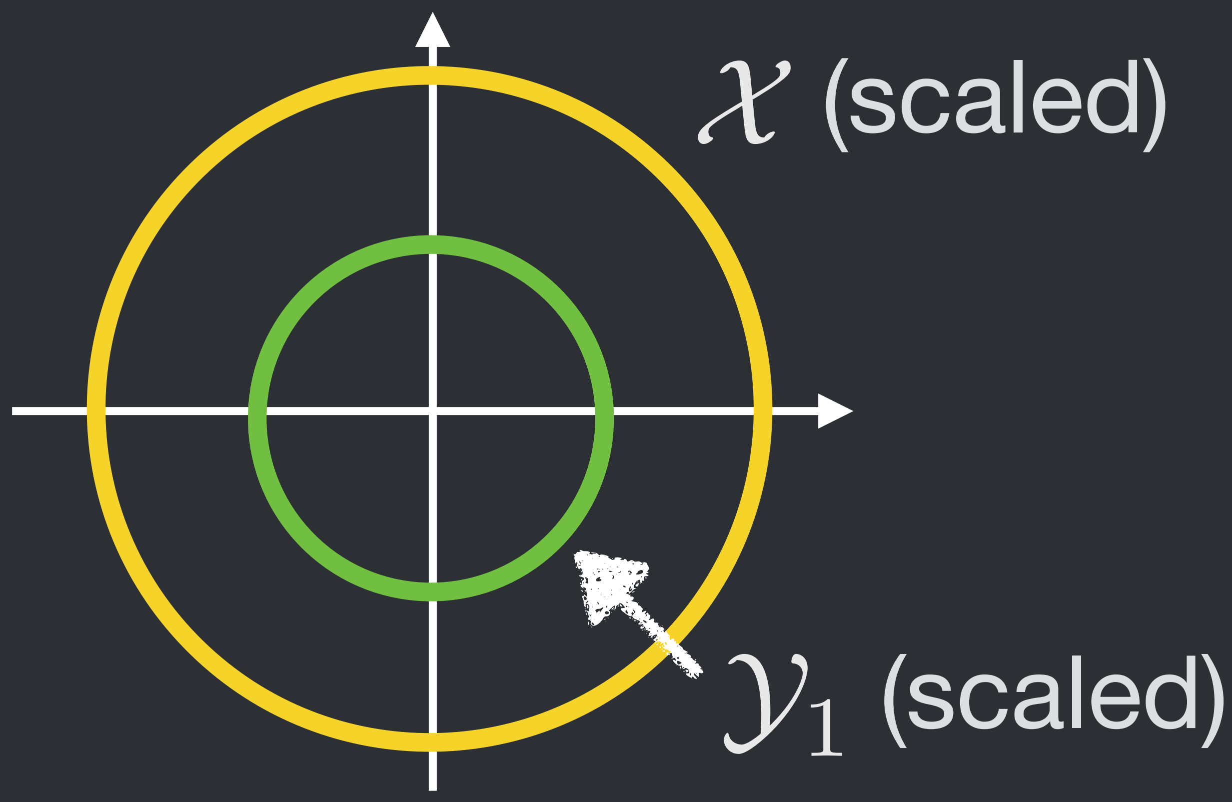




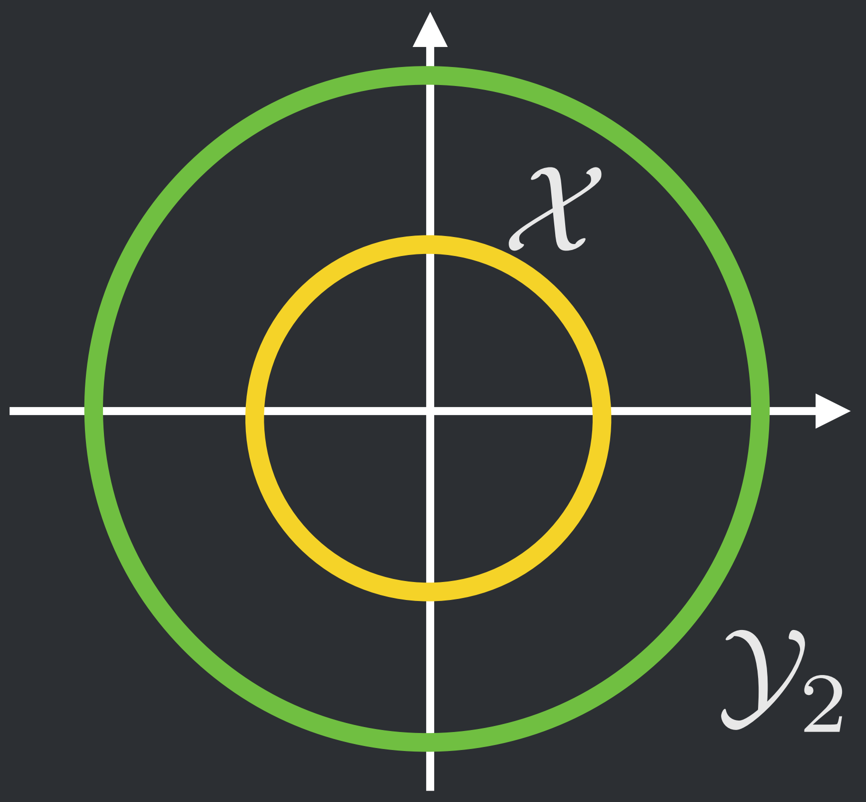
Takeaway

To reduce hubness, label distribution Y with smaller variance should be preferred

Desirable



Not desirable



Why proposed approach reduces hubness

Argument for our proposal relies on two concepts

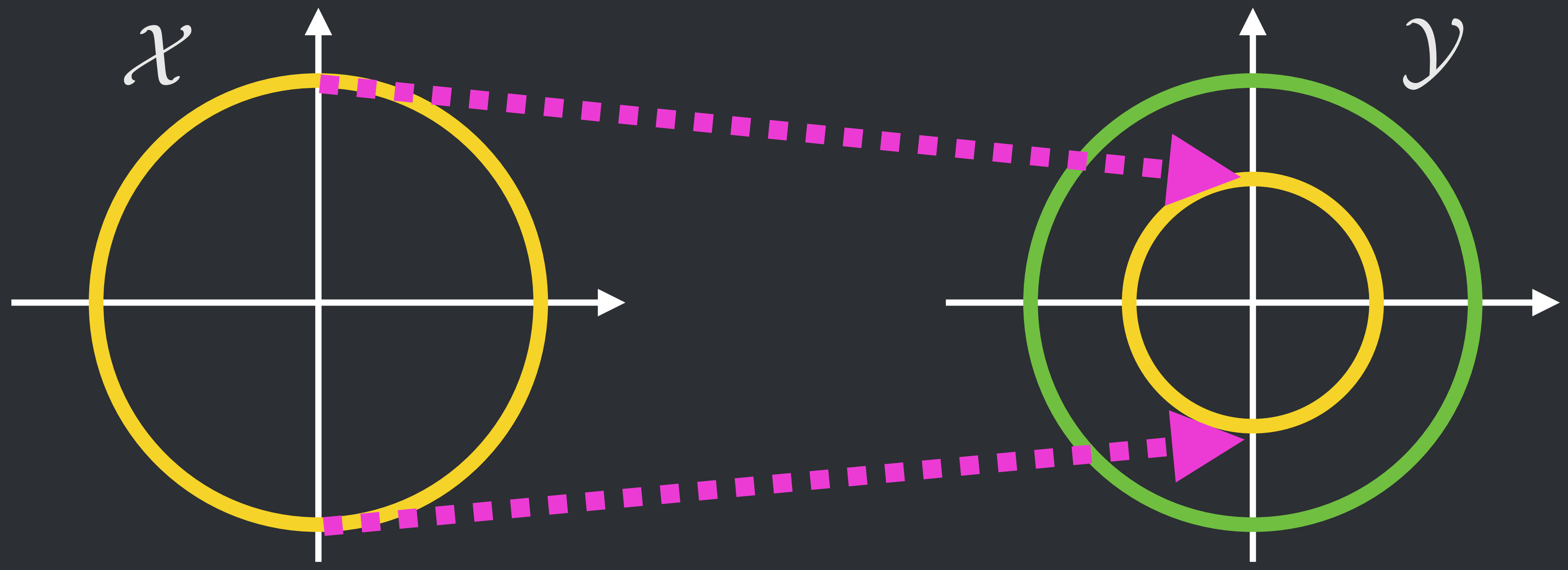
Spatial centrality
of data distributions

Shrinkage
in regression

“Shrinkage” in ridge/least squares regression

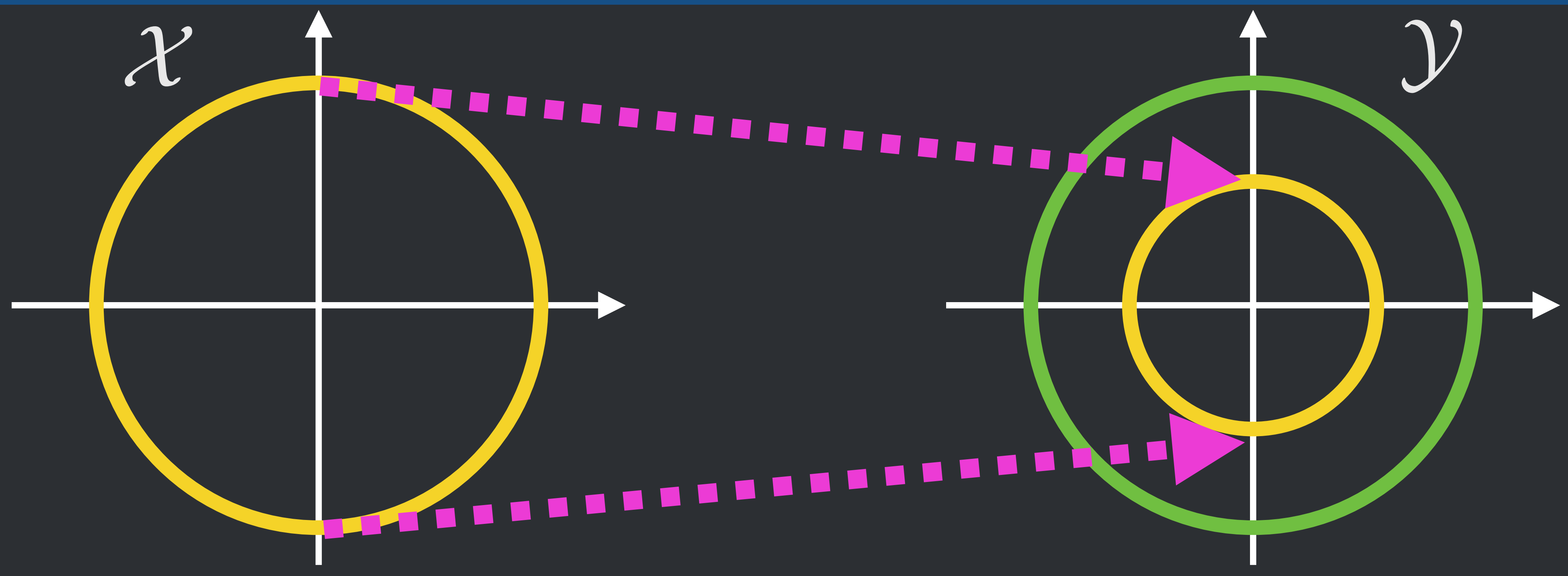
If we optimize $$\min_{\mathbf{M}} \|\mathbf{M}\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{M}\|_F^2$$

Then, we have $$\|\mathbf{M}\mathbf{X}\|_2 \leq \|\mathbf{Y}\|_2$$

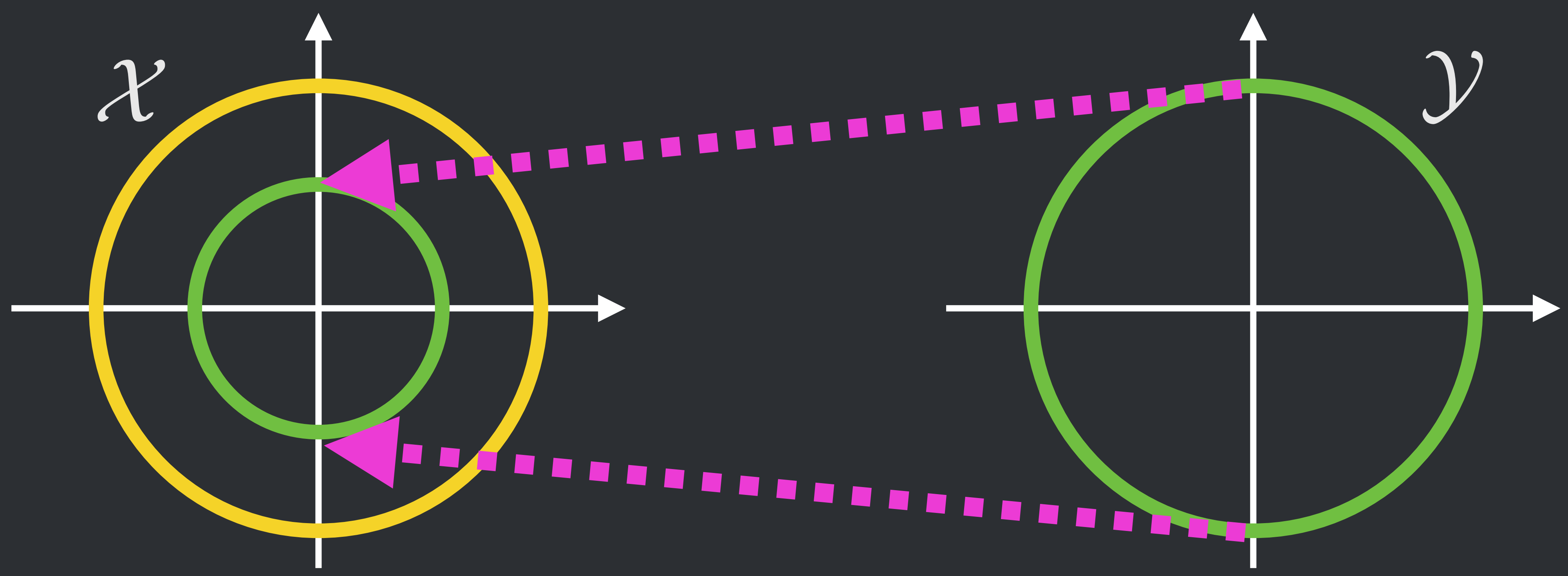


For simplicity, projected objects ae assumed to also follow normal distribution

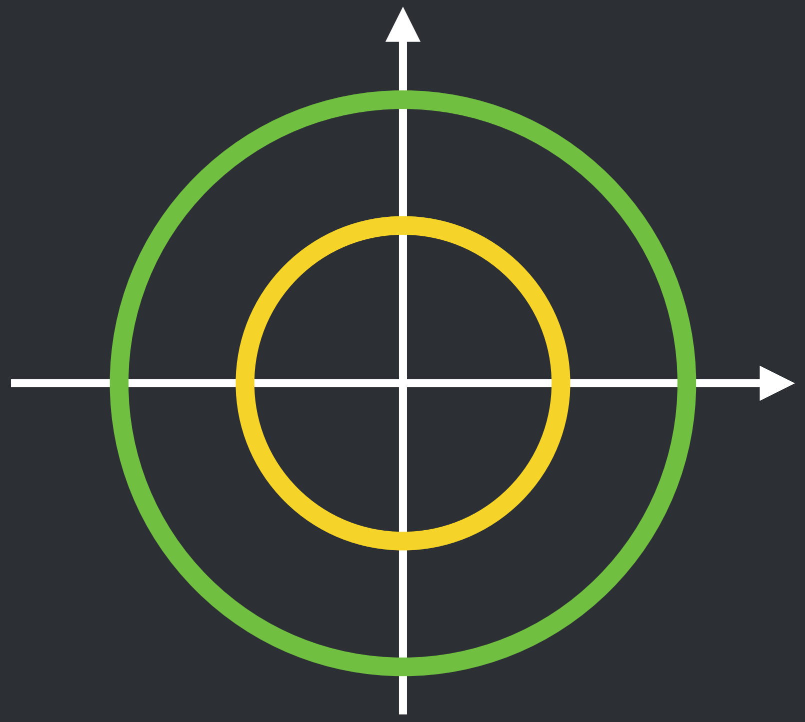
Current approach: map X into Y



Proposed approach: map Y into X



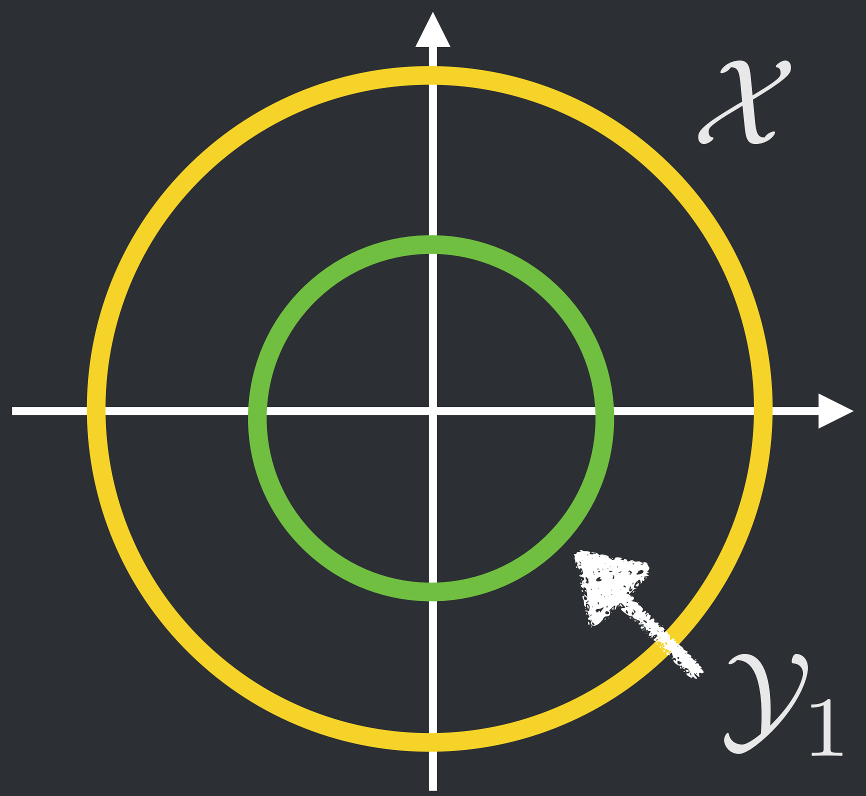
Current approach: map X into Y

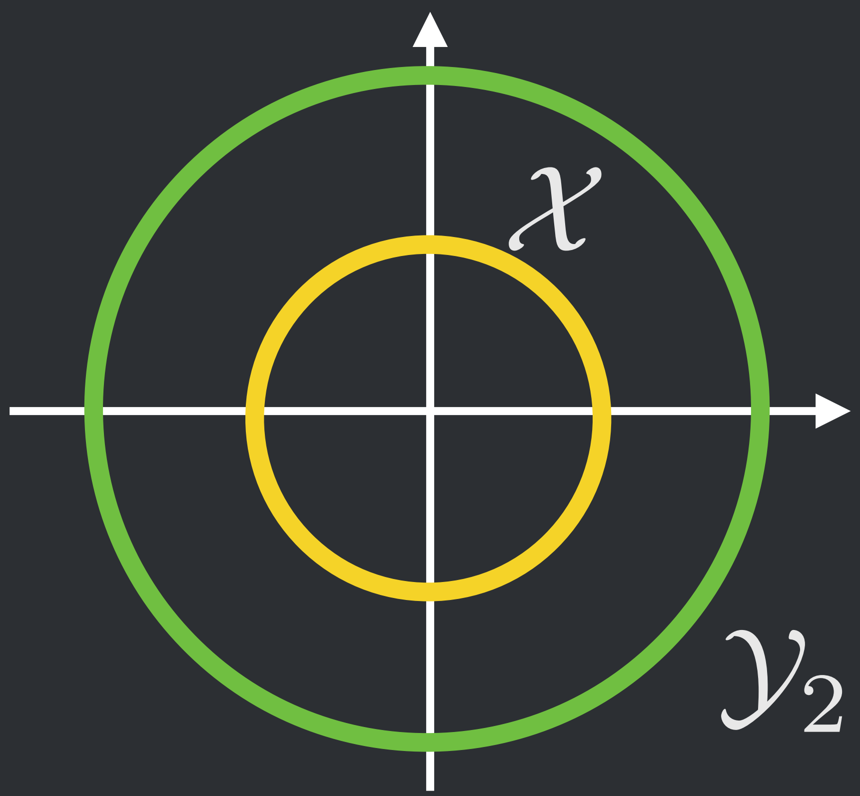


Proposed approach: map Y into X

To reduce hubness, label distributions with smaller variance is more desirable

Desirable	Not desirable





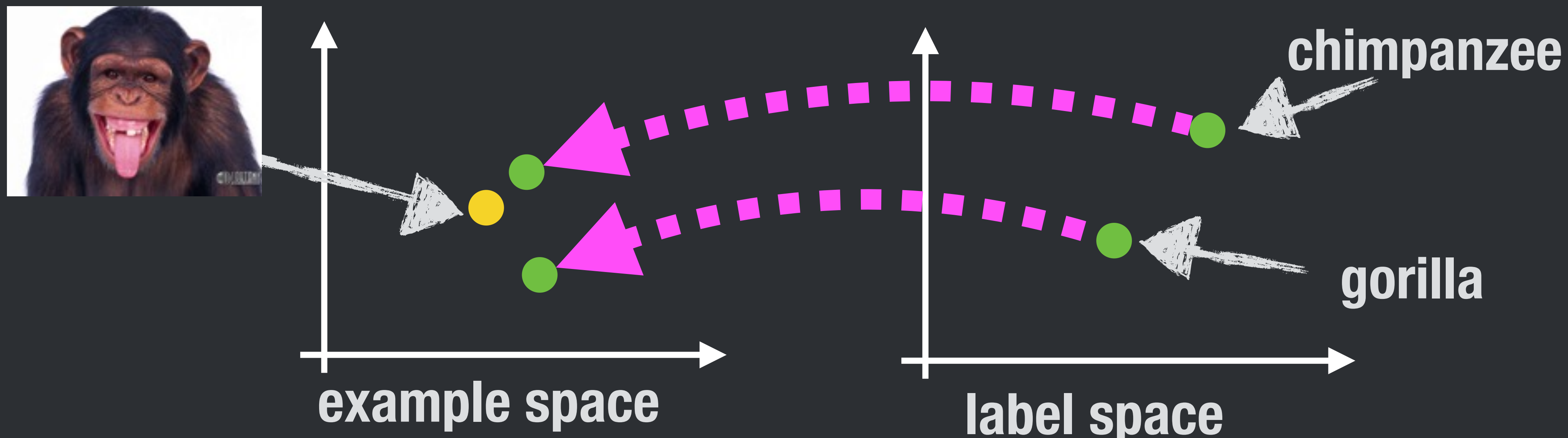
Summary of our proposal

Spatial centrality Label distribution with smaller variance is desirable to reduce hubness

Shrinkage Regression shrinks variance of projected objects

Proposal Project labels into example space

➥ reduces variance of labels, hence suppresses hubness



Experiments

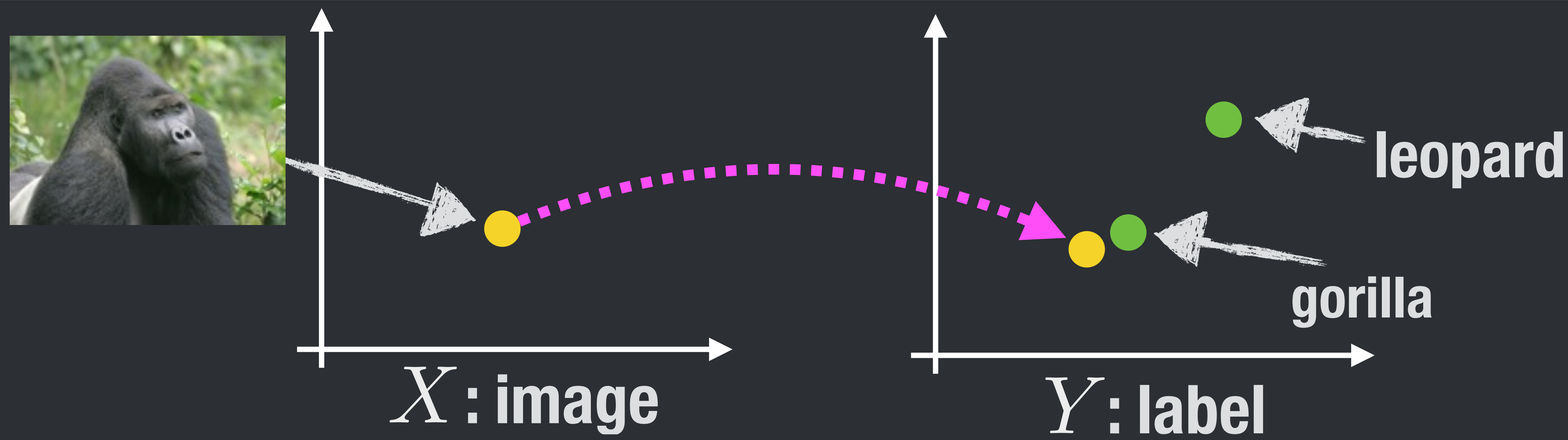
Experimental objective

We evaluate proposed approach in real tasks

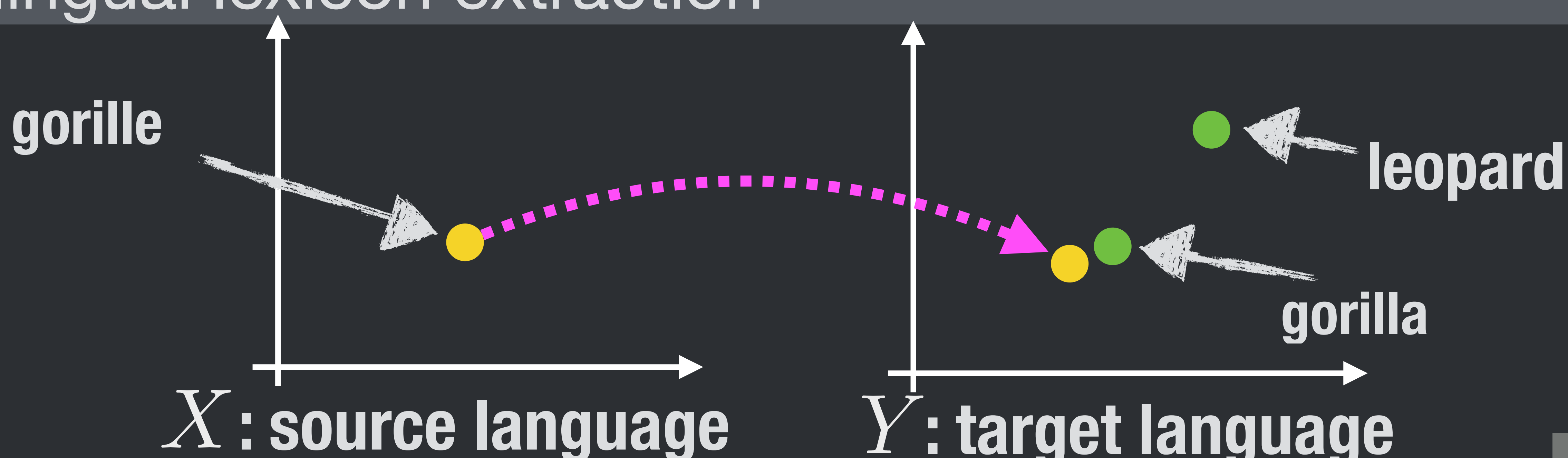
- Does it suppress hubs?
- Does it improve the prediction accuracy?

Zero-shot tasks

- Image labeling



- Bilingual lexicon extraction



Compared methods

Current

$$X \longrightarrow Y$$

Proposed

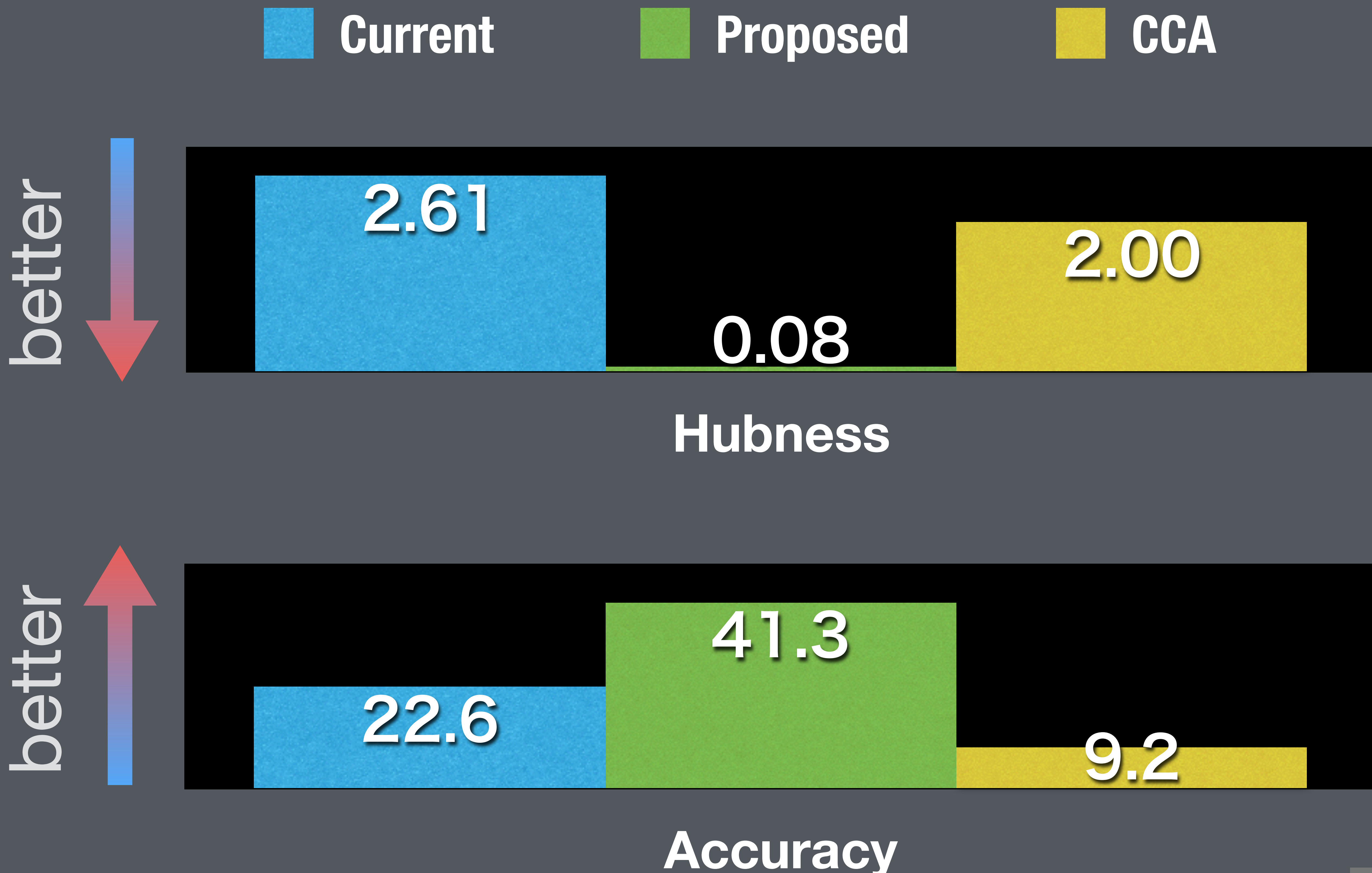
$$X \longleftarrow Y$$

CCA

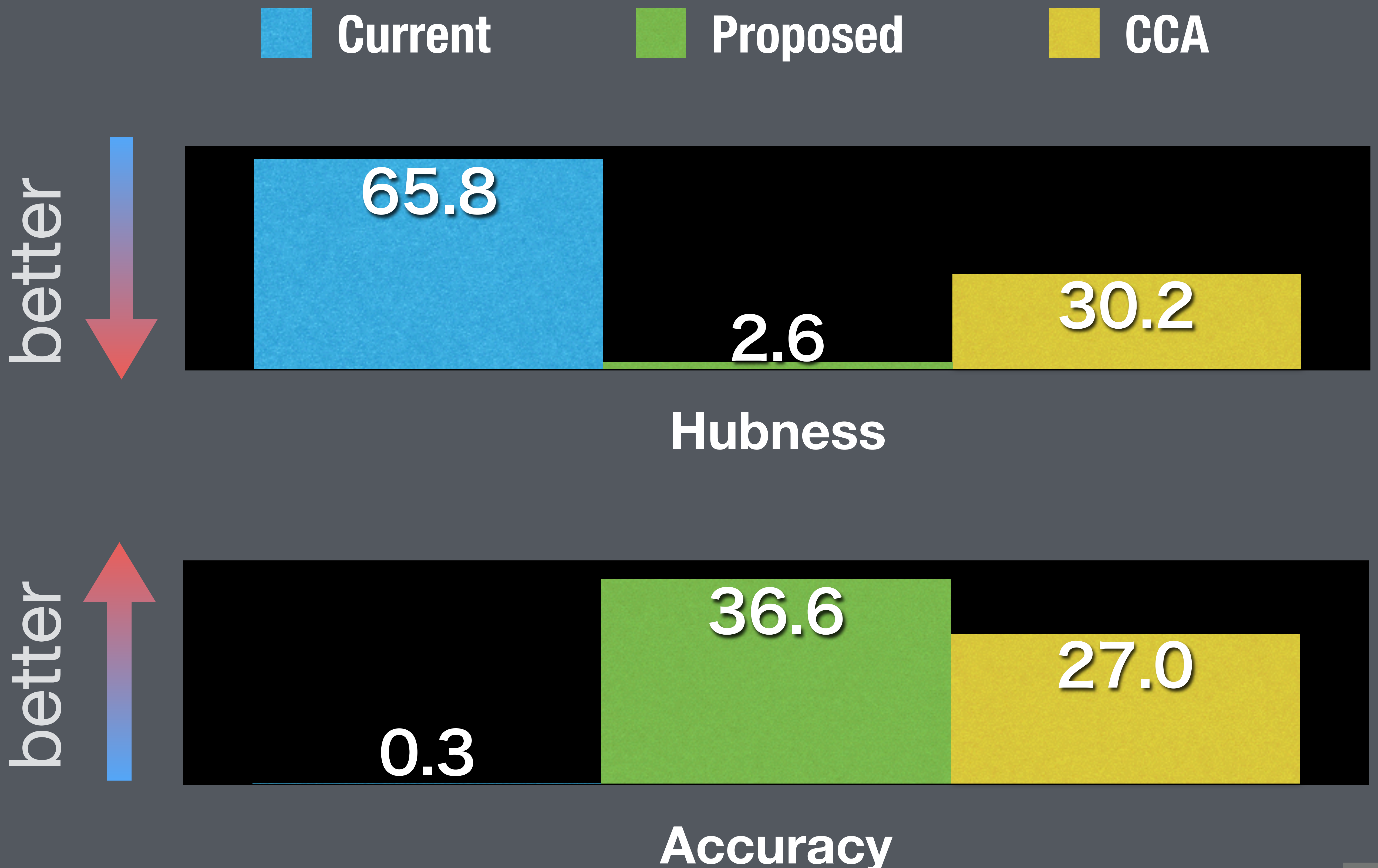
$$\begin{tikzcd} X \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

We used Euclidean distance as a distance measure for finding the nearest label

Image labeling



Bilingual lexicon extraction: fr→en



Conclusions

- Analyzed why hubs emerge in current ZSL approach
  - Variance of labels greater than examples
- Proposed a simple method for reducing hubness
  - Reverse the mapping direction
- Proposed method reduced hubness and outperformed current approach and CCA in image labeling and bilingual lexicon extraction tasks