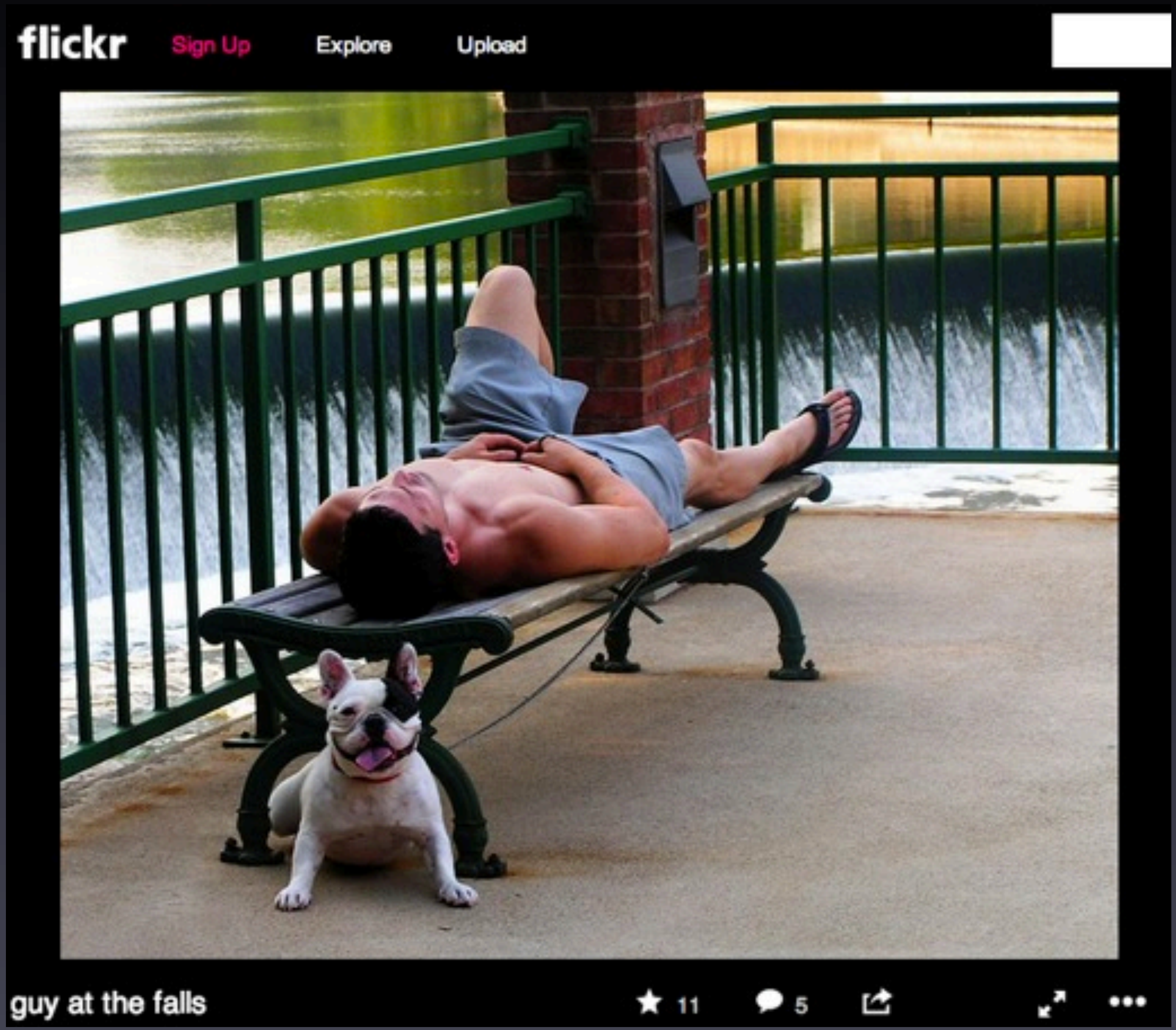
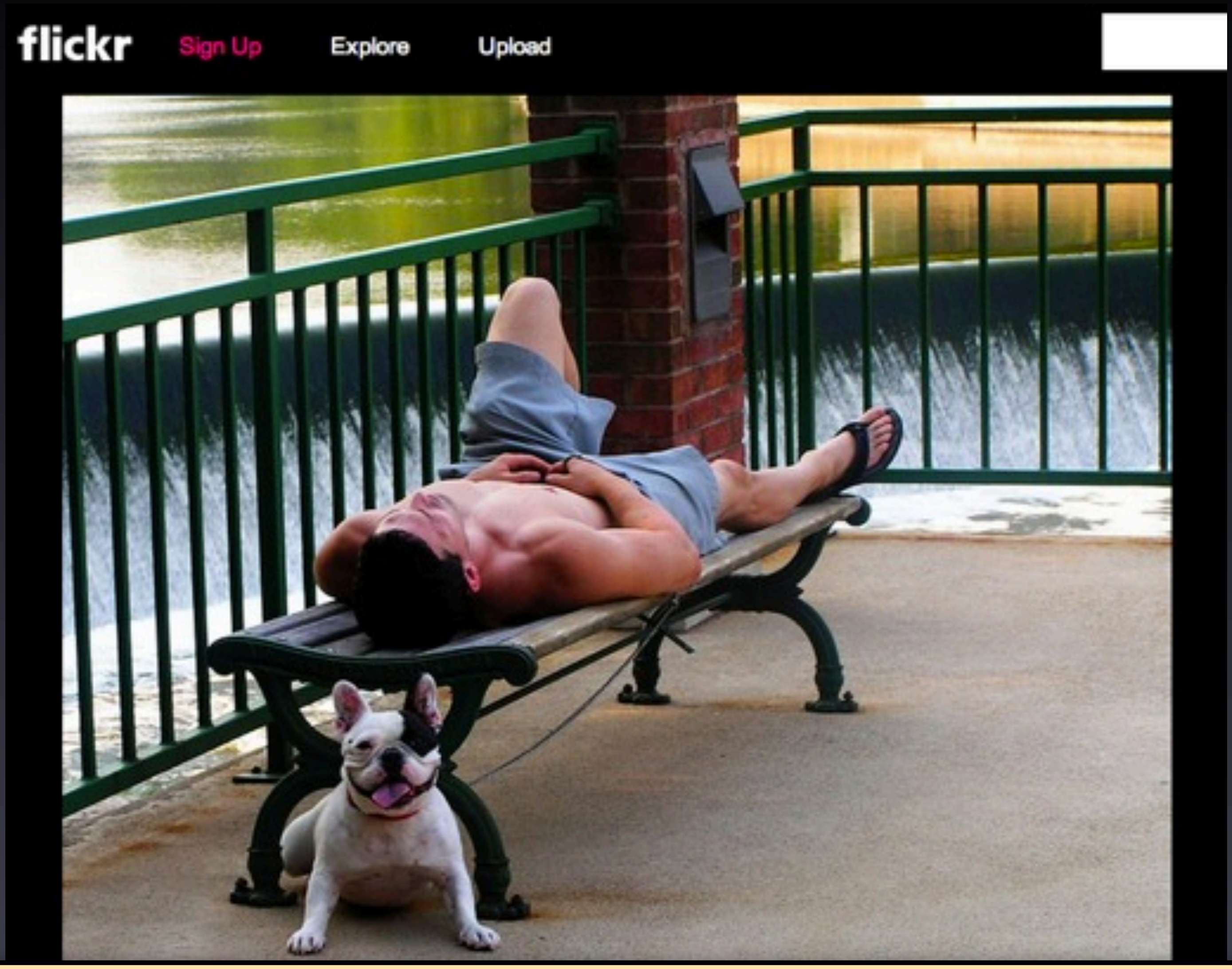
Sentence-based image description with scalable, explicit models

Micah Hodosh
University of Illinois at Urbana-Champaign
mhodosh2@illinois.edu
with Julia Hockenmaier

How would you succinctly describe this image?

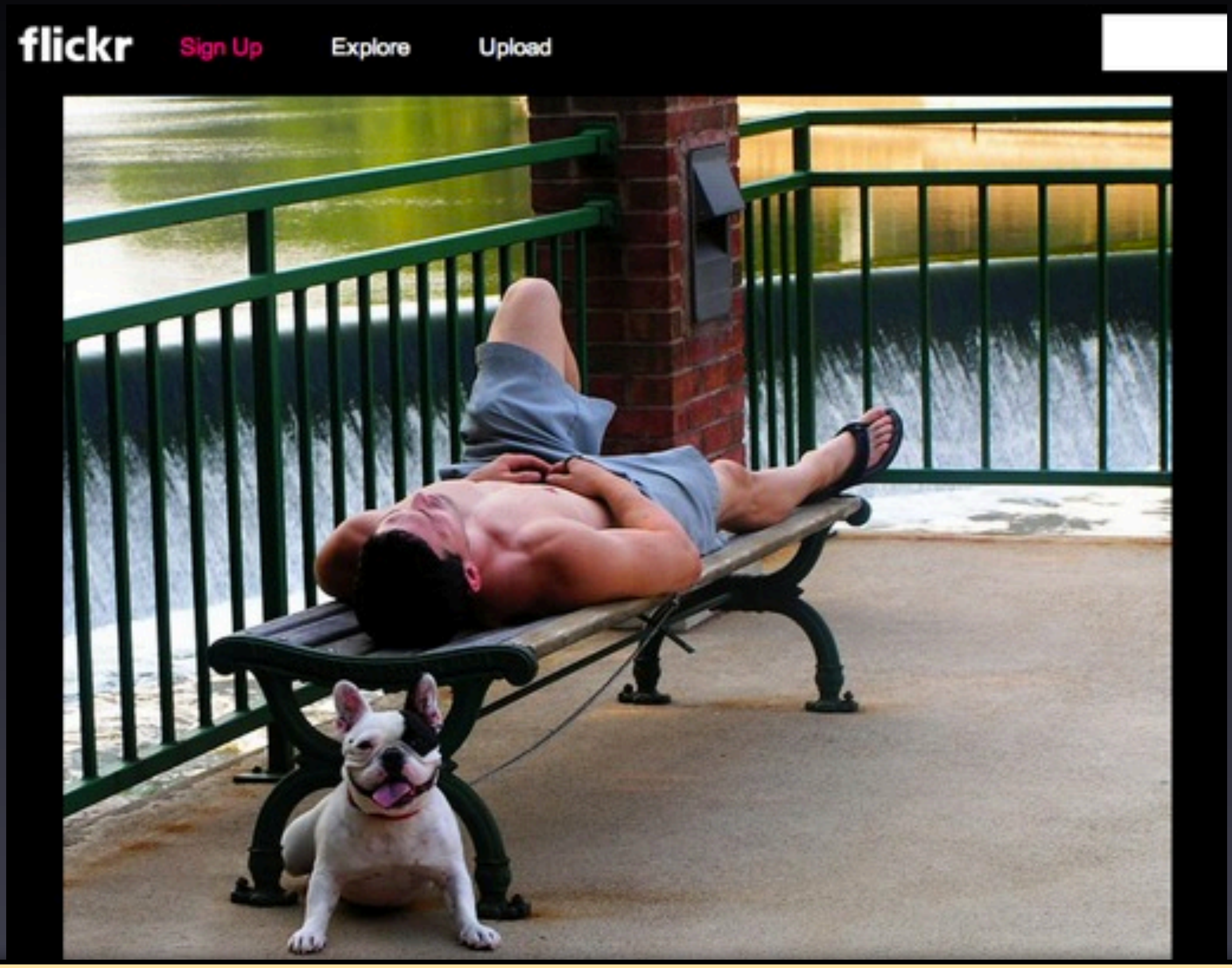


How would you succinctly describe this image?



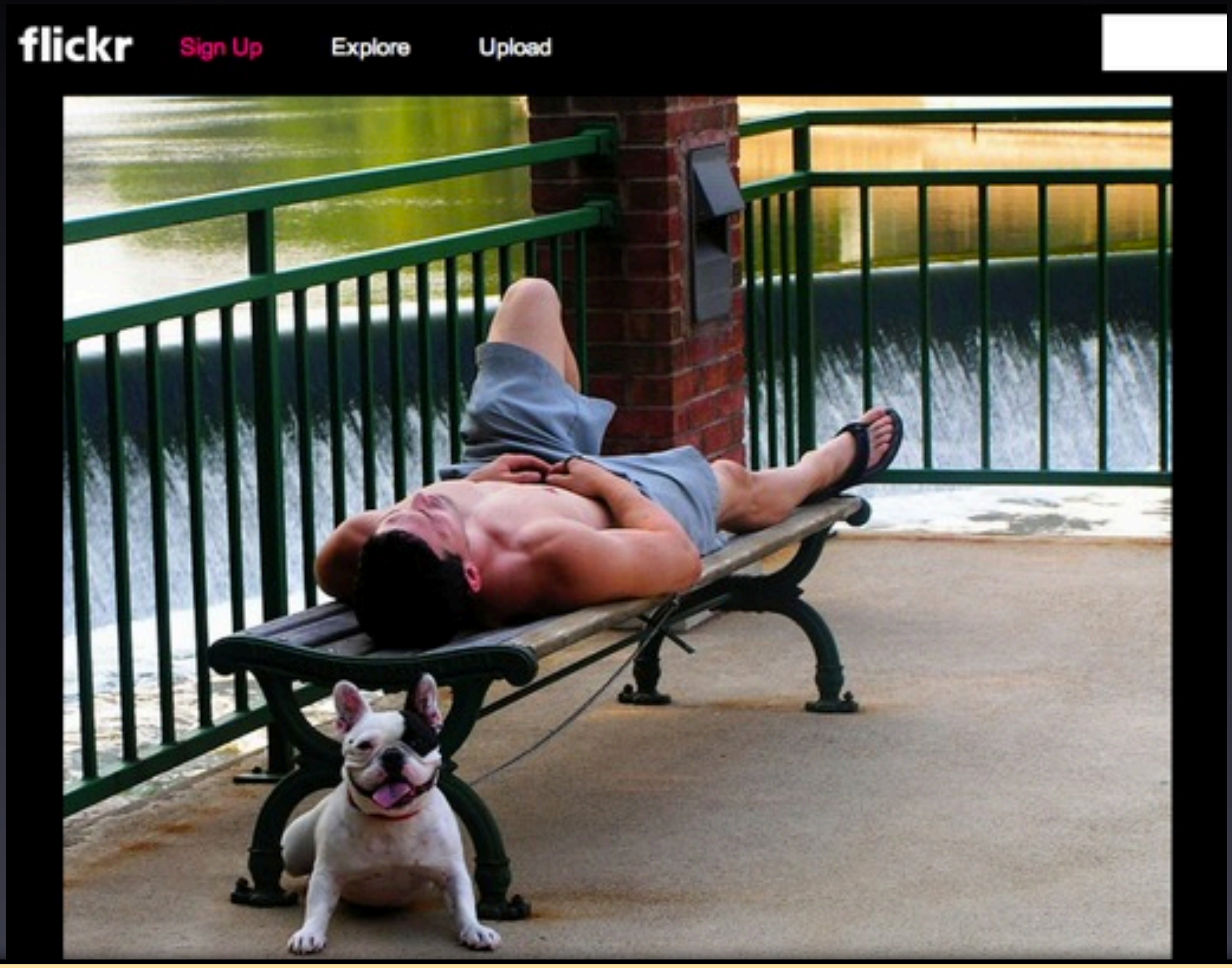
A shirtless guy lies on a park bench with his dog .

How would you succinctly describe this image?



A man lays on a bench while his dog sits by him .
A shirtless guy lies on a park bench with his dog .
A white dog is tied to a bench while its owner sleeps

How would you succinctly describe this image?



Description:

Guy at the falls

I went to the falls to check out the wildlife, and look what I found.

Talk Outline

Talk Outline

Data and the Task: (Hodosh et al. 2013)

Talk Outline

Data and the Task: (Hodosh et al. 2013)

Motivating Related Work: KCCA (Hodosh et al. 2013)

Talk Outline

Data and the Task: (Hodosh et al. 2013)

Motivating Related Work: KCCA (Hodosh et al. 2013)

Alternative Model: Ranking SVM

Talk Outline

Data and the Task: (Hodosh et al. 2013)

Motivating Related Work: KCCA (Hodosh et al. 2013)

Alternative Model: Ranking SVM

Experiments and Results

Talk Outline

Data and the Task: (Hodosh et al. 2013)

Motivating Related Work: KCCA (Hodosh et al. 2013)

Alternative Model: Ranking SVM

Experiments and Results

Representational Issues of Image Descriptions

Our Datasets

1,000 PASCAL Images (2010)
8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)

Our Datasets

1,000 PASCAL Images (2010)
8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)

Mostly people "doing things"

Our Datasets

1,000 PASCAL Images (2010)
8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)

Mostly people “doing things”

5 independently written captions from Amazon Mechanical Turk

Image description as a ranking task

Test Images

●●●

Test Captions

Dogs are running on a wet beach

A snowboarder is sitting on a mountain

The footballer is tackling the other football player

●●●

Image description as a ranking task

Test Images

•••

Test Captions

Dogs are running on a wet beach

A snowboarder is sitting on a mountain

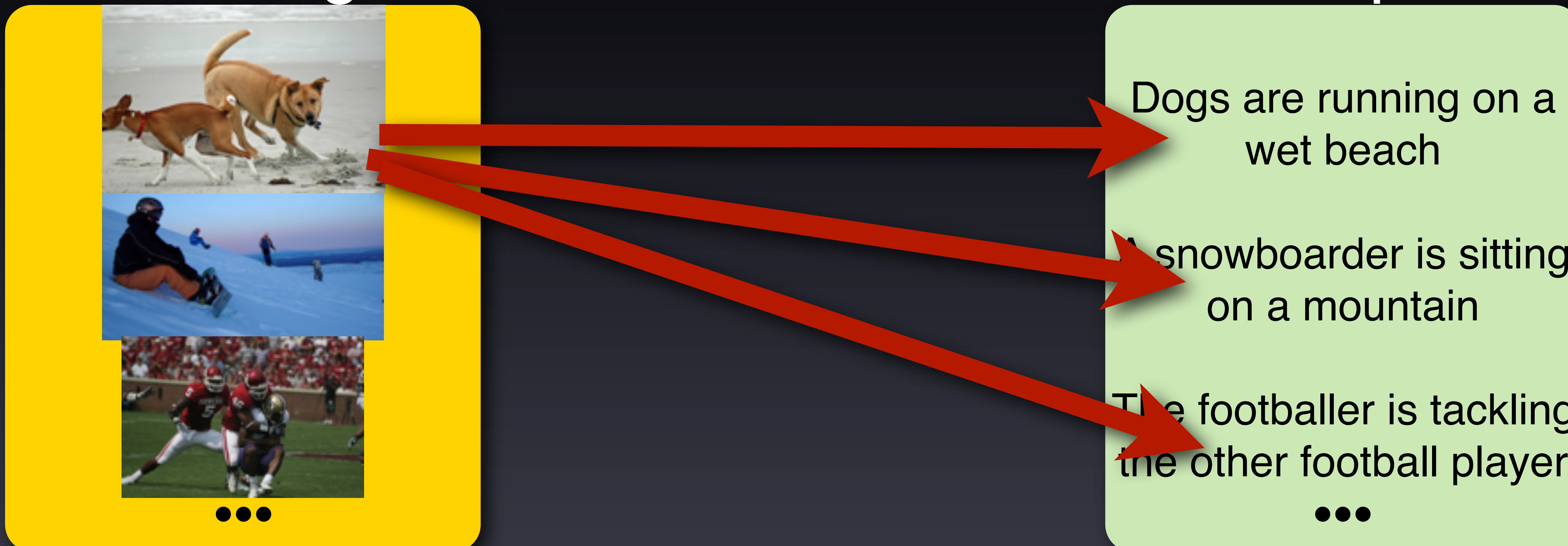
The footballer is tackling the other football player

•••

Image description as a ranking task

Test Images

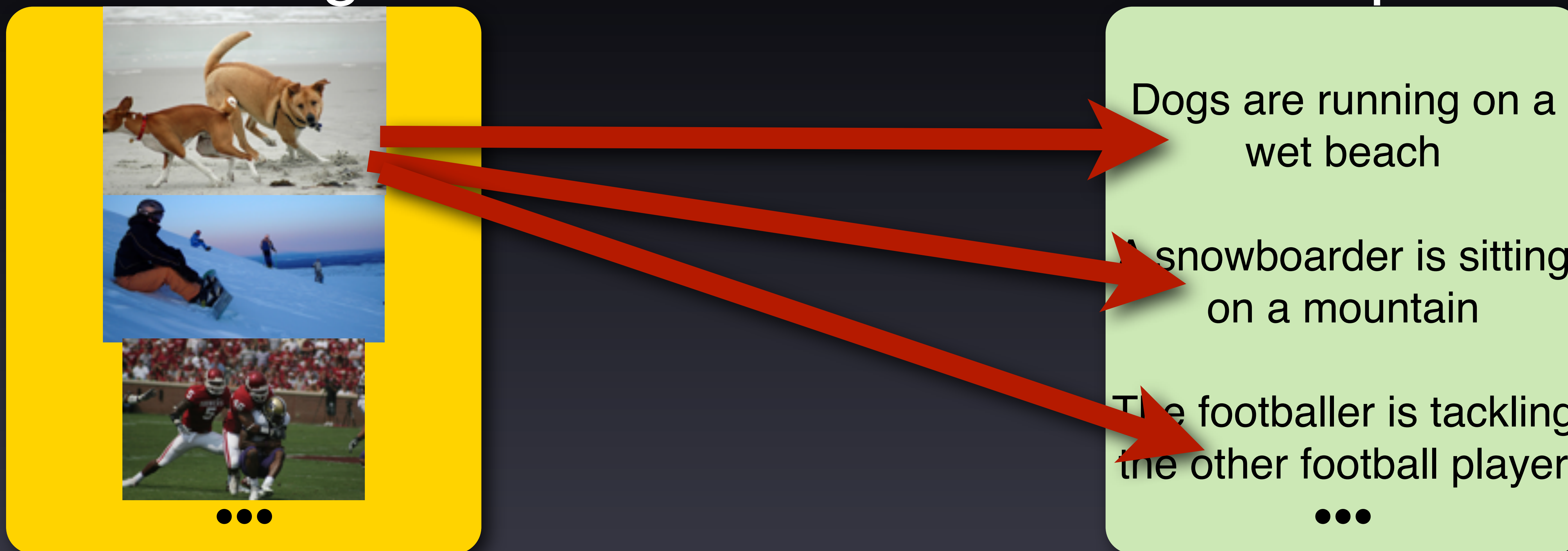
Test Captions



For each test image, rank the pool of test captions

Image description as a ranking task

Test Images · Test Captions



For each test image, rank the pool of test captions

Evaluation: rank of the test image's original caption

Image description as a ranking task

Test Images

Test Captions

Dogs are running on a wet beach

A snowboarder is sitting on a mountain

The footballer is tackling the other football player

•••

For each test image, rank the pool of test captions

Evaluation: rank of the test image's original caption

Can also augment the data with relevance judgments

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Would someone **actually say it**?

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Would someone **actually say it**?

An outside picture with some blue and blue-green ✗

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Would someone **actually say it**?

An outside picture with some blue and blue-green ✗

Generation: semantically correct but **grammatically unsound**?

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Would someone **actually say it**?

An outside picture with some blue and blue-green ✗

Generation: semantically correct but **grammatically unsound**?

Tennis woman play ?

Why evaluate against human captions?

Do the **underlying semantics** of the image and description line up?

A woman is playing tennis. ✓

Would someone **actually say it**?

An outside picture with some blue and blue-green ✗

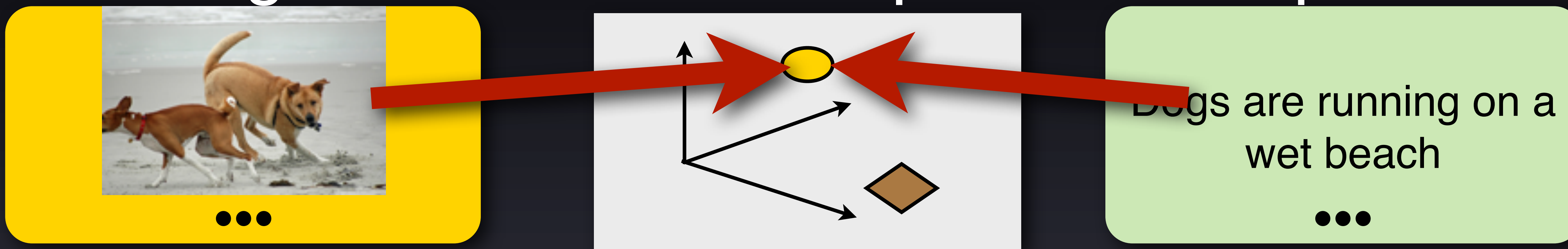
Generation: semantically correct but **grammatically unsound**?

Tennis woman play ?

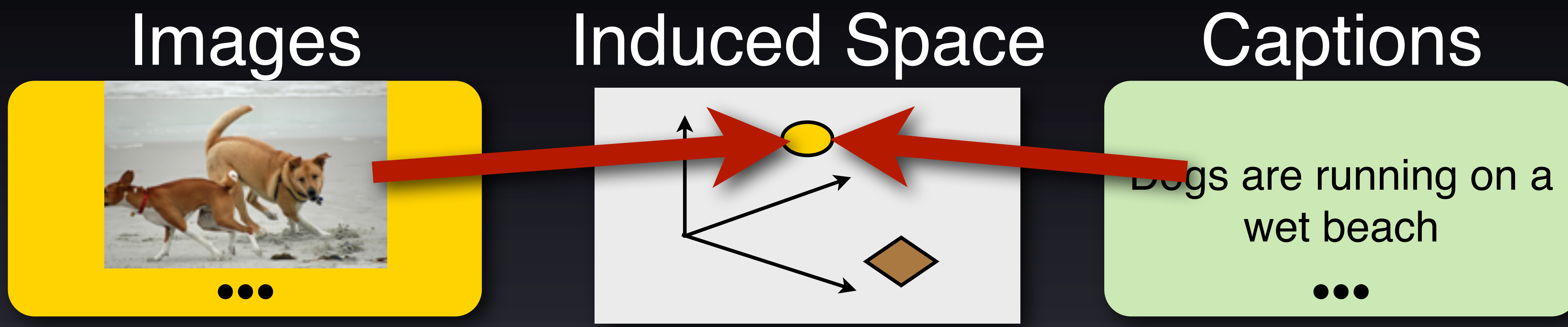
Correlates better with human judgments than BLEU/ROUGE (recall/precision) (Hodosh et al. '13)

KCCA approach (Hodosh et al. 2013)

Images | Induced Space | Captions

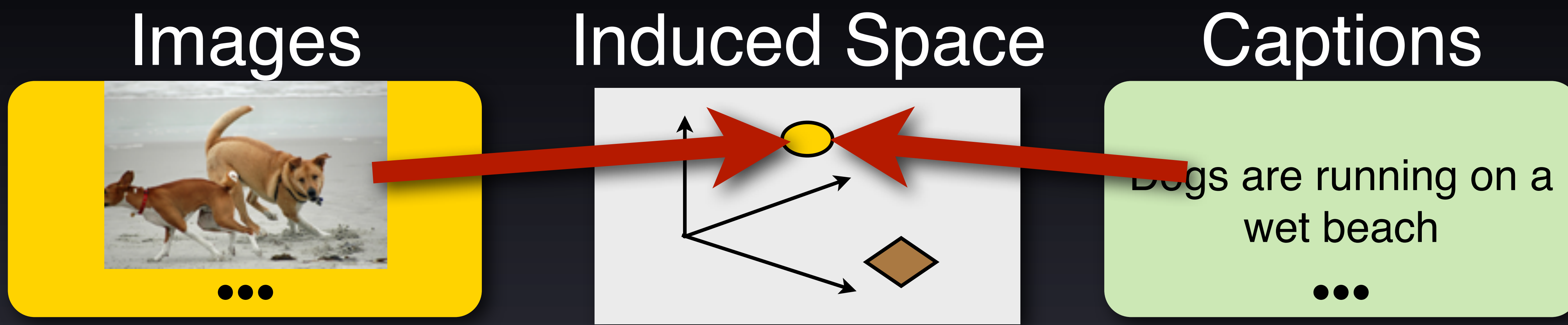


KCCA approach (Hodosh et al. 2013)



Induced space: Linear projection on implicit feature spaces to maximize correlation (KCCA)

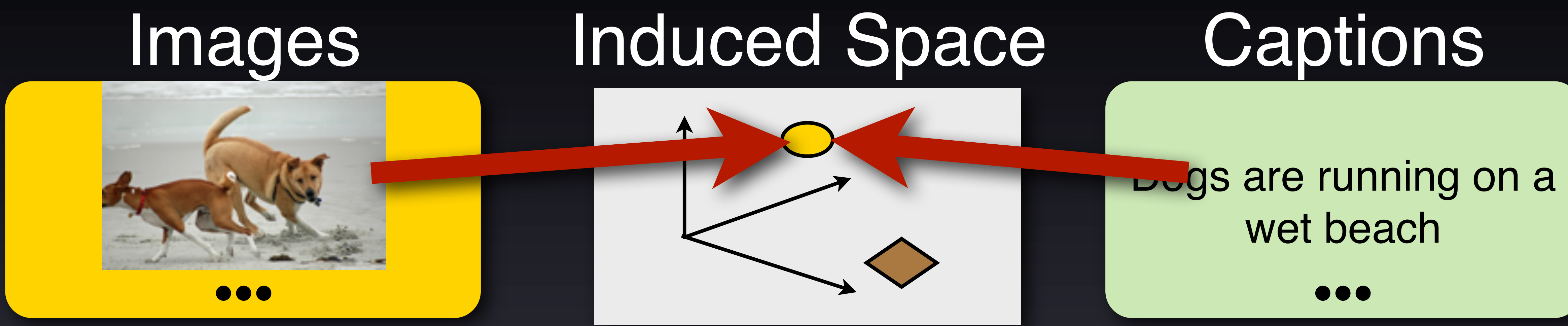
KCCA approach (Hodosh et al. 2013)



Induced space: Linear projection on implicit feature spaces to maximize correlation (KCCA)

Image: Spatial Pyramid with Color, SIFT, Texture
(Intended as a baseline for future work)

KCCA approach (Hodosh et al. 2013)



Induced space: Linear projection on implicit feature spaces to maximize correlation (KCCA)

Image: Spatial Pyramid with Color, SIFT, Texture
(Intended as a baseline for future work)

Text: Bag of words and beyond
(Increases in complexity increase performance)

Text kernel of (Hodosh et al. 2013)

A boy *does a skateboard trick* off a metal plank
A young man *jumps in the air* on a skateboard
Skateboarder on a rail
A skater *does a trick* on a rail

Text kernel of (Hodosh et al. 2013)

A boy *does a skateboard trick* off <u>a metal plank</u>
A young man *jumps in the air* on a skateboard
Skateboarder on a <u>rail</u>
A skater *does a trick* on a <u>rail</u>

Sequence kernel: Beyond BoW

Text kernel of (Hodosh et al. 2013)

A boy *does a skateboard trick* off a metal plank
A young man *jumps in the air* on a skateboard
Skateboarder on a rail
A skater *does a trick* on a rail

Sequence kernel: Beyond BoW

Similarity kernel(s): Partial matches

Text kernel of (Hodosh et al. 2013)

A boy *does a skateboard trick* off a metal plank
A young man *jumps in the air* on a skateboard
Skateboarder on a rail
A skater *does a trick* on a rail

Sequence kernel: Beyond BoW

Similarity kernel(s): Partial matches

Alignment: Translation modeling on our corpus

Text kernel of (Hodosh et al. 2013)

A boy *does a skateboard trick* off a metal plank
A young man *jumps in the air* on a skateboard
Skateboarder on a rail
A skater *does a trick* on a rail

Sequence kernel: Beyond BoW

Similarity kernel(s): Partial matches

- **Alignment:** Translation modeling on our corpus
- **Distributional:** Co-occurrence to capture topic info

Qualitative KCCA Examples*:

*See JAIR paper (Hodosh et al'13) for more discussion

Qualitative KCCA Examples*:

Query
Top Result
Human Judgement*

*See JAIR paper (Hodosh et al'13) for more discussion

Qualitative KCCA Examples*:

Query	
Top Result	A girl wearing a yellow shirt and sunglasses smiles.
Human Judgement*	4 out of 4

*See JAIR paper (Hodosh et al'13) for more discussion

Qualitative KCCA Examples*:

Query		
Top Result	A girl wearing a yellow shirt and sunglasses smiles.	A child jumping on a tennis court.
Human Judgement*	4 out of 4	3 out of 4

*See JAIR paper (Hodosh et al'13) for more discussion

Qualitative KCCA Examples*:

Query			
Top Result	A girl wearing a yellow shirt and sunglasses smiles.	A child jumping on a tennis court.	A boy in a blue life jacket jumps into the water.
Human Judgement*	4 out of 4	3 out of 4	2 out of 4

*See JAIR paper (Hodosh et al'13) for more discussion

Tractability

Datasets are **growing rapidly**

8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)
1 Million+ (Ordonez et al 2011)

Tractability

Datasets are **growing rapidly**

8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)
1 Million+ (Ordonez et al 2011)

KCCA Memory: O(n^2) (kernels & learned weights)

Tractability

Datasets are **growing rapidly**

8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)
1 Million+ (Ordonez et al 2011)

KCCA Memory: $\mathbf{O(n^2)}$ (kernels & learned weights)

KCCA Running Time: $\mathbf{O(n^3)}$ (exactly)

Tractability

Datasets are **growing rapidly**

8,000 Flickr Images (2010)
31,000+ Flickr Images (2013)
1 Million+ (Ordonez et al 2011)

KCCA Memory: **$O(n^2)$** (kernels & learned weights)

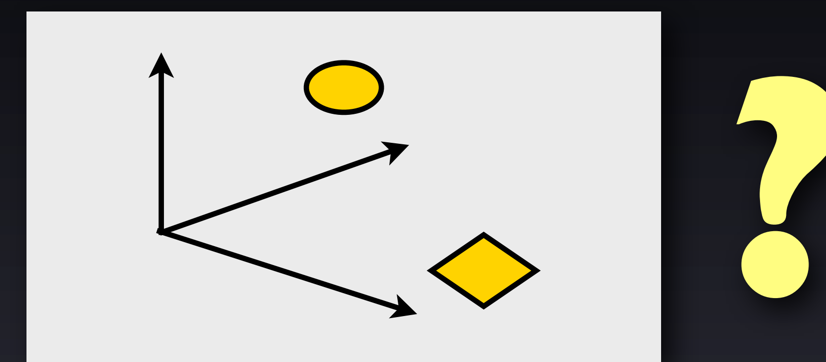
KCCA Running Time: **$O(n^3)$** (exactly)

Pre-computation: **$O(n^2)$** kernel operations

Understandability

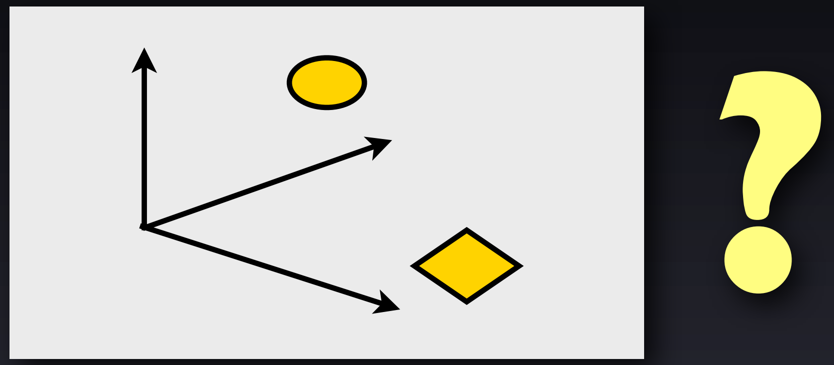
Understandability

Interpreting the **why** of induced implicit spaces can be difficult



Understandability

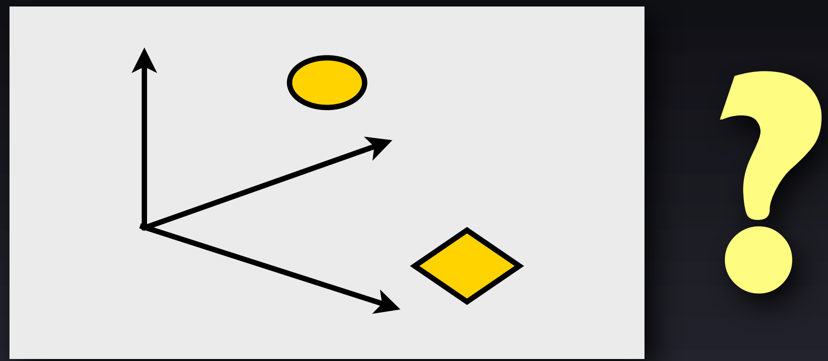
Interpreting the **why** of induced implicit spaces can be difficult



How does one feature or component affect the much larger kernel?

Understandability

Interpreting the **why** of induced implicit spaces can be difficult



How does one feature or component affect the much larger kernel?

How does one change in a kernel effect the space KCCA learns?

Appropriate loss metric

KCCA's loss isn't the same as the task's loss

Appropriate loss metric

KCCA's loss isn't the same as the task's loss

$$\operatorname{argmax}_{\mathbf{w}_{\mathcal{A}},\mathbf{w}_{\mathcal{B}}} \frac{\langle \mathbf{A}\mathbf{w}_{\mathcal{A}}, \mathbf{B}\mathbf{w}_{\mathcal{B}} \rangle}{\|\mathbf{A}\mathbf{w}_{\mathcal{A}}\| \|\mathbf{B}\mathbf{w}_{\mathcal{B}}\|}$$

Appropriate loss metric

KCCA's loss isn't the same as the task's loss

$$\operatorname{argmax}_{\mathbf{w}_{\mathcal{A}},\mathbf{w}_{\mathcal{B}}} \frac{\langle \mathbf{A}\mathbf{w}_{\mathcal{A}}, \mathbf{B}\mathbf{w}_{\mathcal{B}} \rangle}{\|\mathbf{A}\mathbf{w}_{\mathcal{A}}\| \|\mathbf{B}\mathbf{w}_{\mathcal{B}}\|}$$

: A woman hiding her face behind an umbrella > A man is running in a city park

Rank-SVM*

A woman hiding her face behind an umbrella

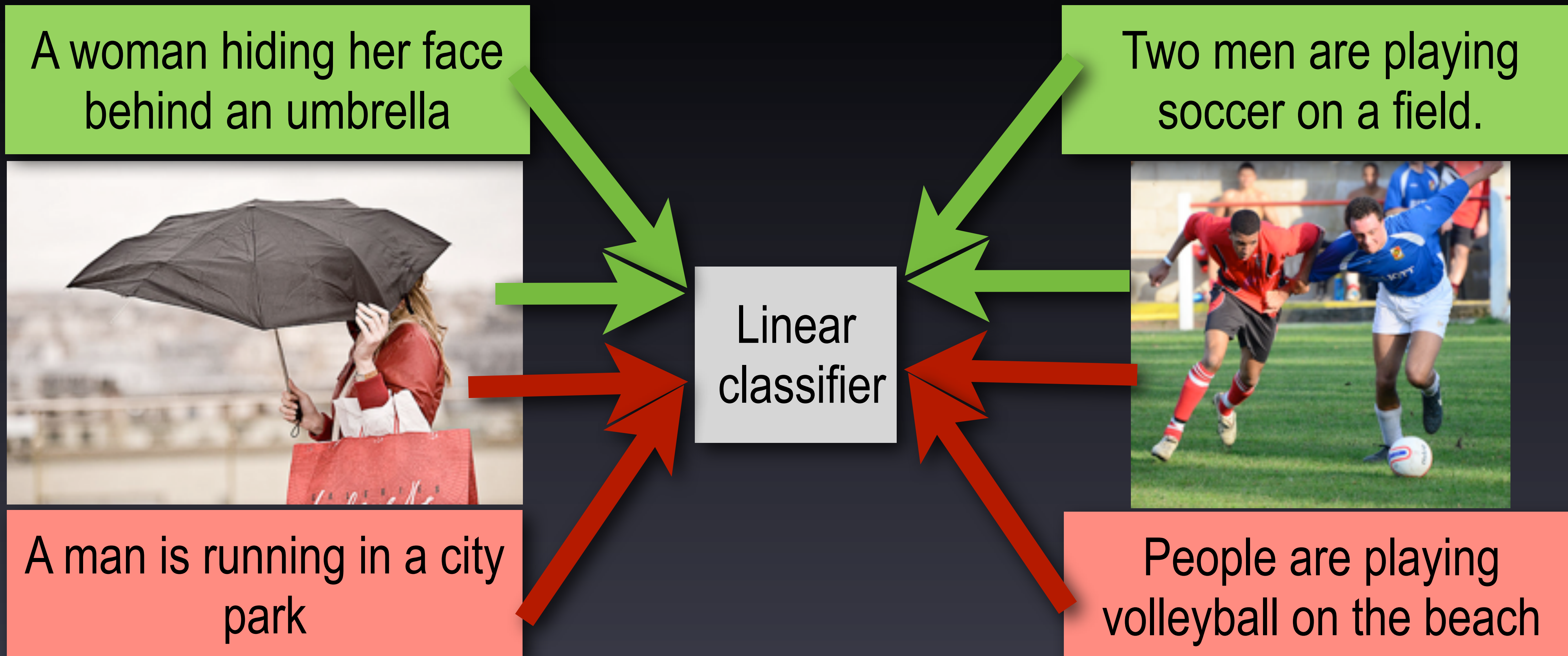
A man is running in a city park

Two men are playing soccer on a field.

People are playing volleyball on the beach

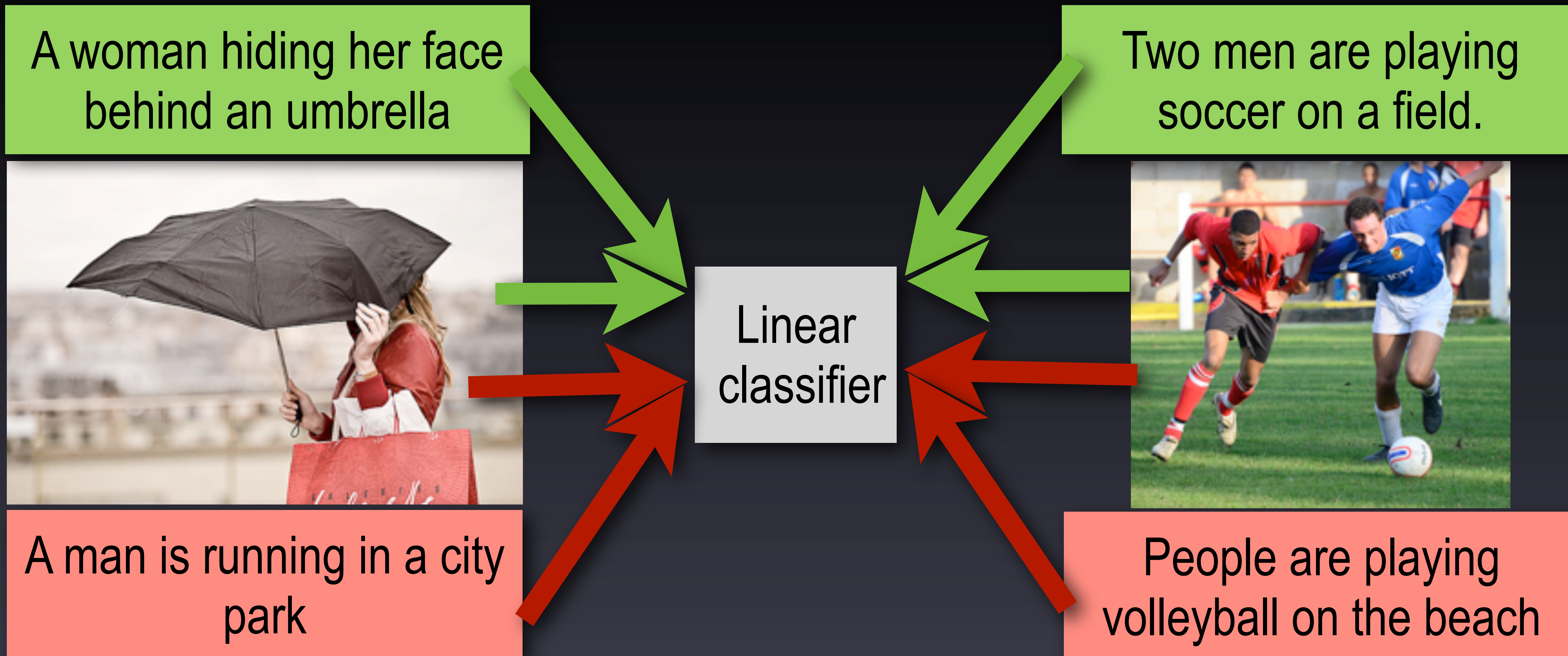
*Related to Grangier et al (2008) (PAMIR)

Rank-SVM*



*Related to Grangier et al (2008) (PAMIR)

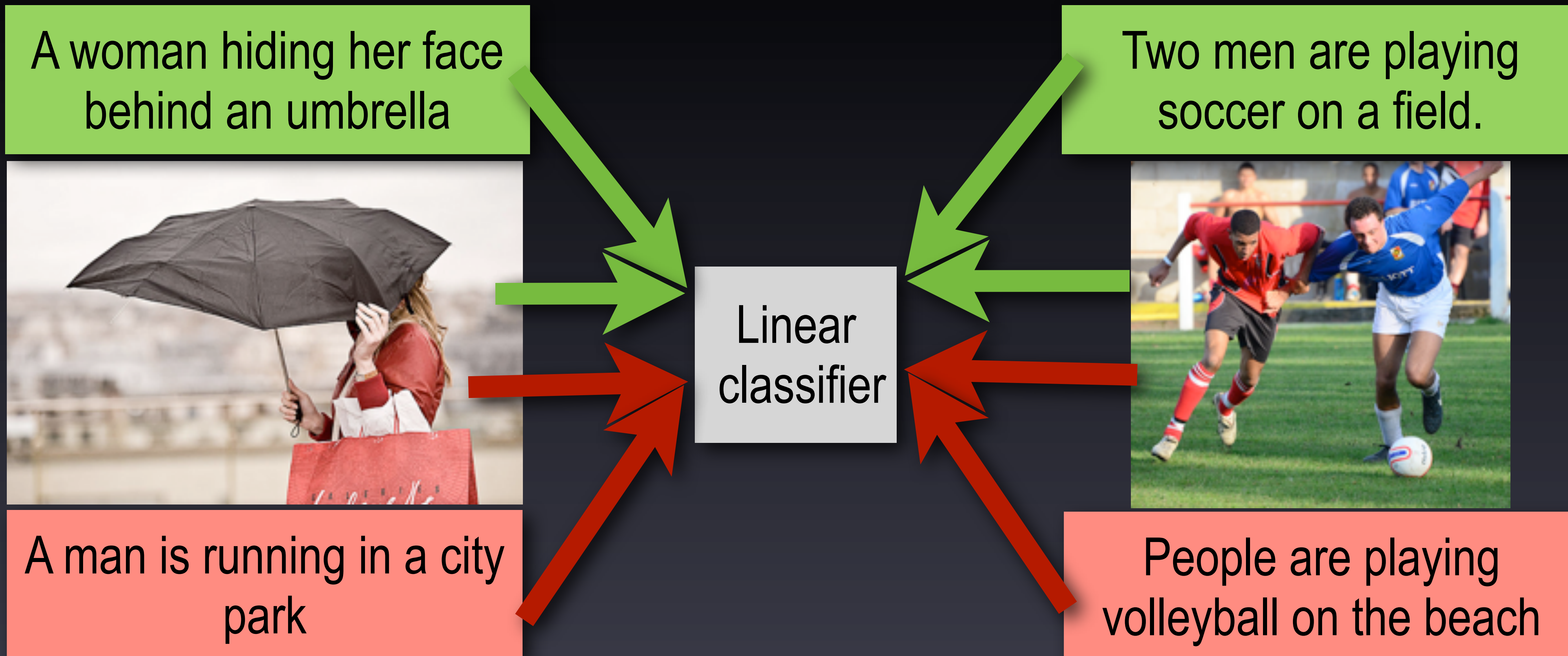
Rank-SVM*



Representation: Simple cross-product of active image and text features.

*Related to Grangier et al (2008) (PAMIR)

Rank-SVM*

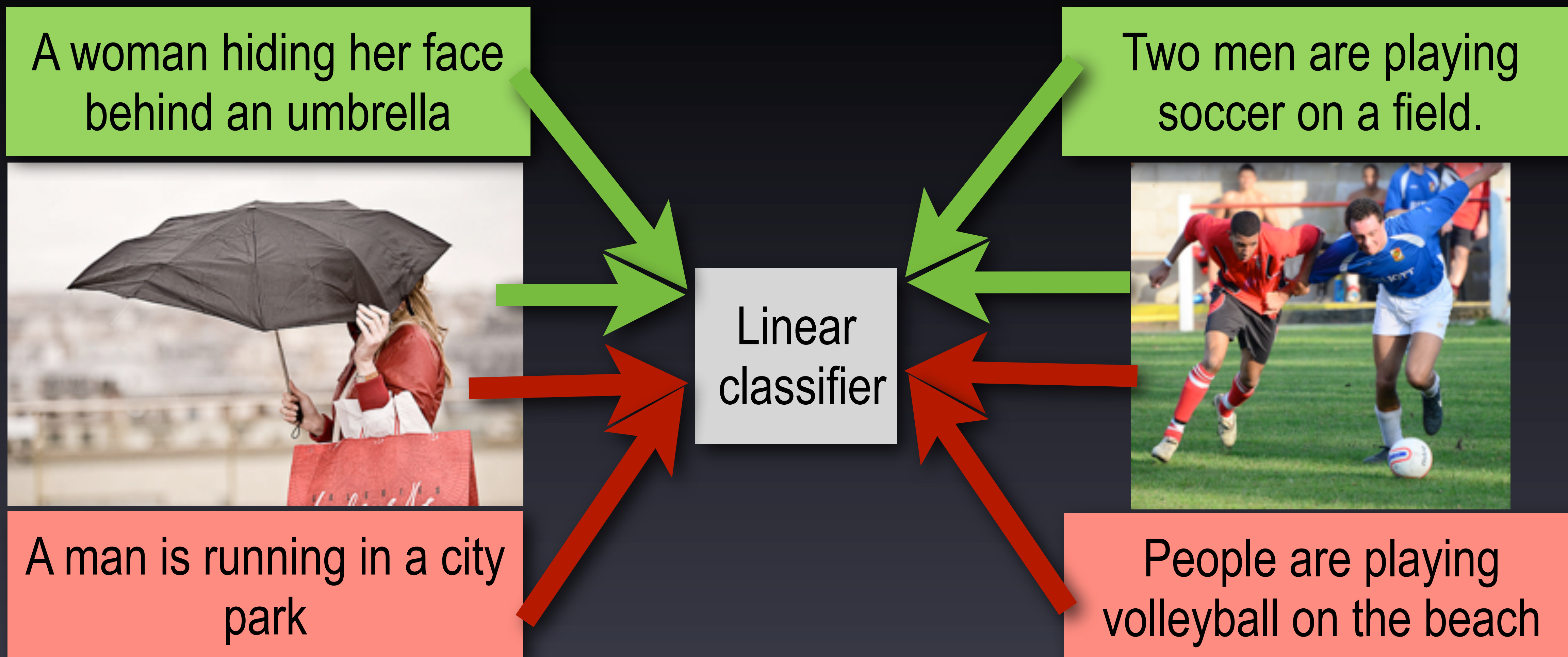


Representation: Simple cross-product of active image and text features.

Image: Binary MetaClass (Bergamo & Torresani '12)

*Related to Grangier et al (2008) (PAMIR)

Rank-SVM*



Representation: Simple cross-product of active image and text features.

Image: Binary MetaClass (Bergamo & Torresani '12)

Text: Currently just binary "BoW"

*Related to Grangier et al (2008) (PAMIR)

Rank-SVM formally

Let D_{train} be a set of pairwise preferences of captions for the training images

$$\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{|D_{train}|} \sum_{(i,c^+,c^-)\in D_{train}} \ell((i,c^+,c^-),\mathbf{w})$$

Rank-SVM formally

Let D_{train} be a set of pairwise preferences of captions for the training images

$$\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{|D_{train}|} \sum_{(i,c^+,c^-)\in D_{train}} \ell((i,c^+,c^-),\mathbf{w})$$

Loss is hinge-loss on each of these preferences

$$\ell((i,c^+,c^-),\mathbf{w}) = \max(0, 1 - \langle \mathbf{w}, \Phi(i,c^+) - \Phi(i,c^-)\rangle)$$

Binary text features

Allows for more **compact** storage in memory

Binary text features

Allows for more **compact** storage in memory

When words **repeat**: more of the concept?

Binary text features

Allows for more **compact** storage in memory

When words **repeat**: more of the concept?

A **man** with a black shirt giving another **man** a tattoo

A **man** wearing jeans gets a new tattoo

Binary text features

Allows for more **compact** storage in memory

When words **repeat**: more of the concept?

A **man** with a black shirt giving another **man** a tattoo	A **man** wearing jeans gets a new tattoo

Incorporating the **KCCA features**?

The quantitative task

The quantitative task

Test Set: 1000 unseen images and captions

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

Metric: Recall of gold (correlates with human)

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

Metric: Recall of gold (correlates with human)

Models: Independent Baseline

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

Metric: Recall of gold (correlates with human)

Models: Independent Baseline

KCCA Model of Hodosh et al. 2013

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

Metric: Recall of gold (correlates with human)

Models: Independent Baseline

KCCA Model of Hodosh et al. 2013

Rank SVM w/ IDF + Extra training

The quantitative task

Test Set: 1000 unseen images and captions

Task: For each image, rank the captions

Metric: Recall of gold (correlates with human)

Models: Independent Baseline

KCCA Model of Hodosh et al. 2013

Rank SVM w/ IDF + Extra training

(See paper for more models / description / etc)

Automatic Evaluation

	Recall at 1	Recall at 5	Recall at 10	Median rank of gold
Independent	4.1	13.2	20.3	51.0
Rank SVM	6.8	19.2	28.7	34.7
KCCA*	**8.3**	**21.6**	**30.3**	**34.0**

See workshop and JAIR paper for more experiments

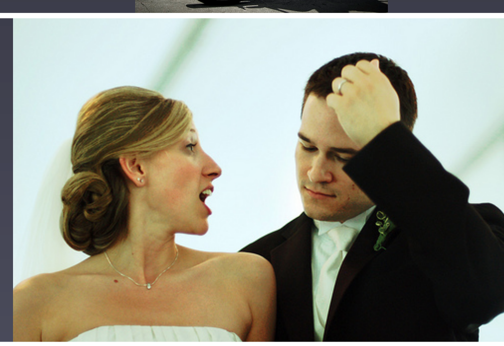
Automatic Evaluation

	Recall at 1	Recall at 5	Recall at 10	Median rank of gold
Independent	4.1	13.2	20.3	51.0
Rank SVM	6.8	19.2	28.7	34.7
KCCA*	**8.3**	**21.6**	**30.3**	**34.0**

See workshop and JAIR paper for more experiments

*Different visual/text features etc, so not directly comparable

What is the Rank-SVM learning?

Crowd			
Table			
Bicycle			
Two			

All descriptions are not created equal

All descriptions are not created equal

A meal is on a table
in a restaurant.

Model responses	
Overall	**0.96**
Meal	0.39
Restaurant	0.34
Table	0.22

All descriptions are not created equal

A meal is on a table in a restaurant.

Model responses	
Overall	**0.96**
Meal	0.39
Restaurant	0.34
Table	0.22

A well lit room, with three glasses on the table and two plates.

Model responses	
Overall	**-0.85**
Three	-0.45
Two	-0.26
Well	-0.25

L2 Normalization

L2 Normalization

L2 normalization **doesn't help**

L2 Normalization

L2 normalization **doesn't help**

	
A man with a black shirt giving another man a tattoo	A man wearing jeans gets a new tattoo

L2 Normalization

L2 normalization **doesn't help**

A man with a black shirt giving another man a tattoo

A man wearing jeans gets a new tattoo

The left image isn't **less of "tattoo"**

L2 Normalization

L2 normalization **doesn't help**

A man with a black shirt giving another man a tattoo	A man wearing jeans gets a new tattoo

The left image isn't **less of "tattoo"**

Without L2 Normalization, **worst case position is bounded**

Annotators miss “salient” information

Annotators miss “salient” information

A man on an orange bike

Annotators miss “salient” information

A man on an orange bike

Might not be able to **localize the color** of the bike

Annotators miss “salient” information

A man on an orange bike

Might not be able to **localize the color** of the bike

If we knew **“in the air” (etc) could be implied** it would push the correct picture closer in the learned space

Annotators miss “salient” information

A man on an orange bike

Might not be able to **localize the color** of the bike

If we knew **“in the air” (etc) could be implied** it would push the correct picture closer in the learned space

Is **“in the woods” more likely** for biking?

Annotators miss “salient” information

A man on an orange bike

Might not be able to **localize the color** of the bike

If we knew **“in the air” (etc) could be implied** it would push the correct picture closer in the learned space

Is **“in the woods” more likely** for biking?

Is “in the woods” more likely to be **implied**? **(less salient)**?

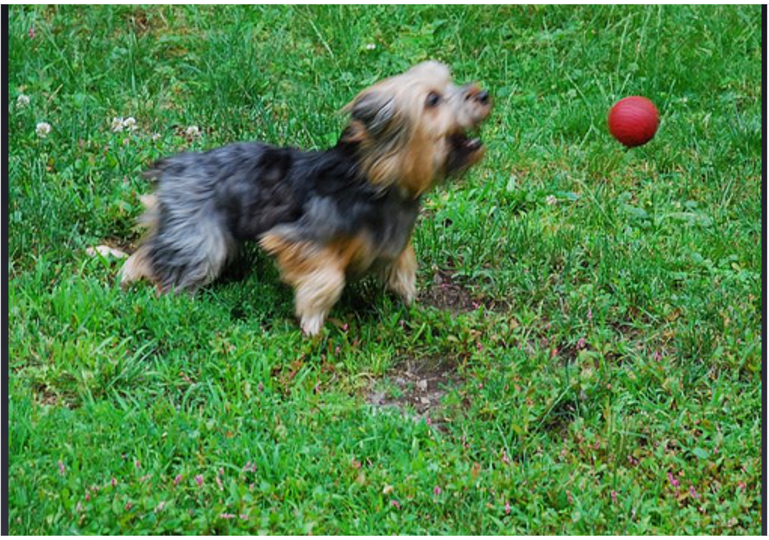
Different saliency: Red

Different saliency: Red

The same word can vary in overall notability for an image

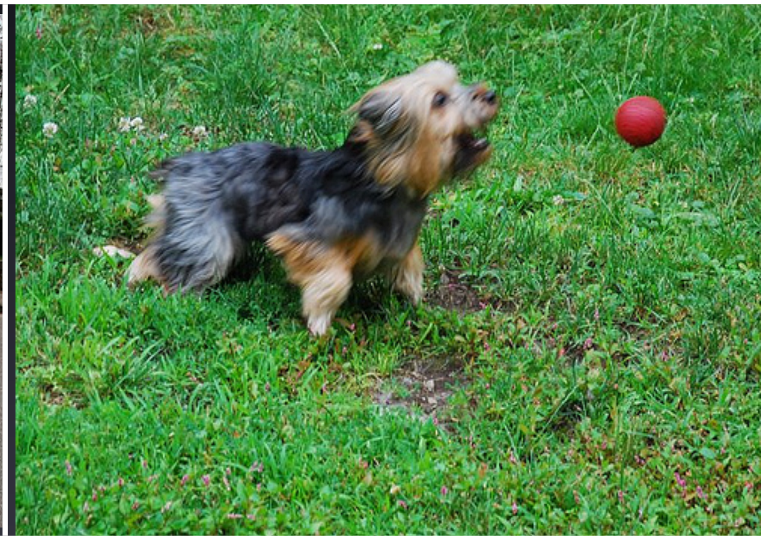
Different saliency: Red

The same word can vary in overall notability for an image

		
A man is getting into a red car	A small dog tries to catch a red ball	A woman in red shoes walking in the street

Different saliency: Red

The same word can vary in overall notability for an image

		
A man is getting into a red car	A small dog tries to catch a red ball	A woman in red shoes walking in the street

Localization alone isn't all that is needed
(People **don't mention every adjective or color**)

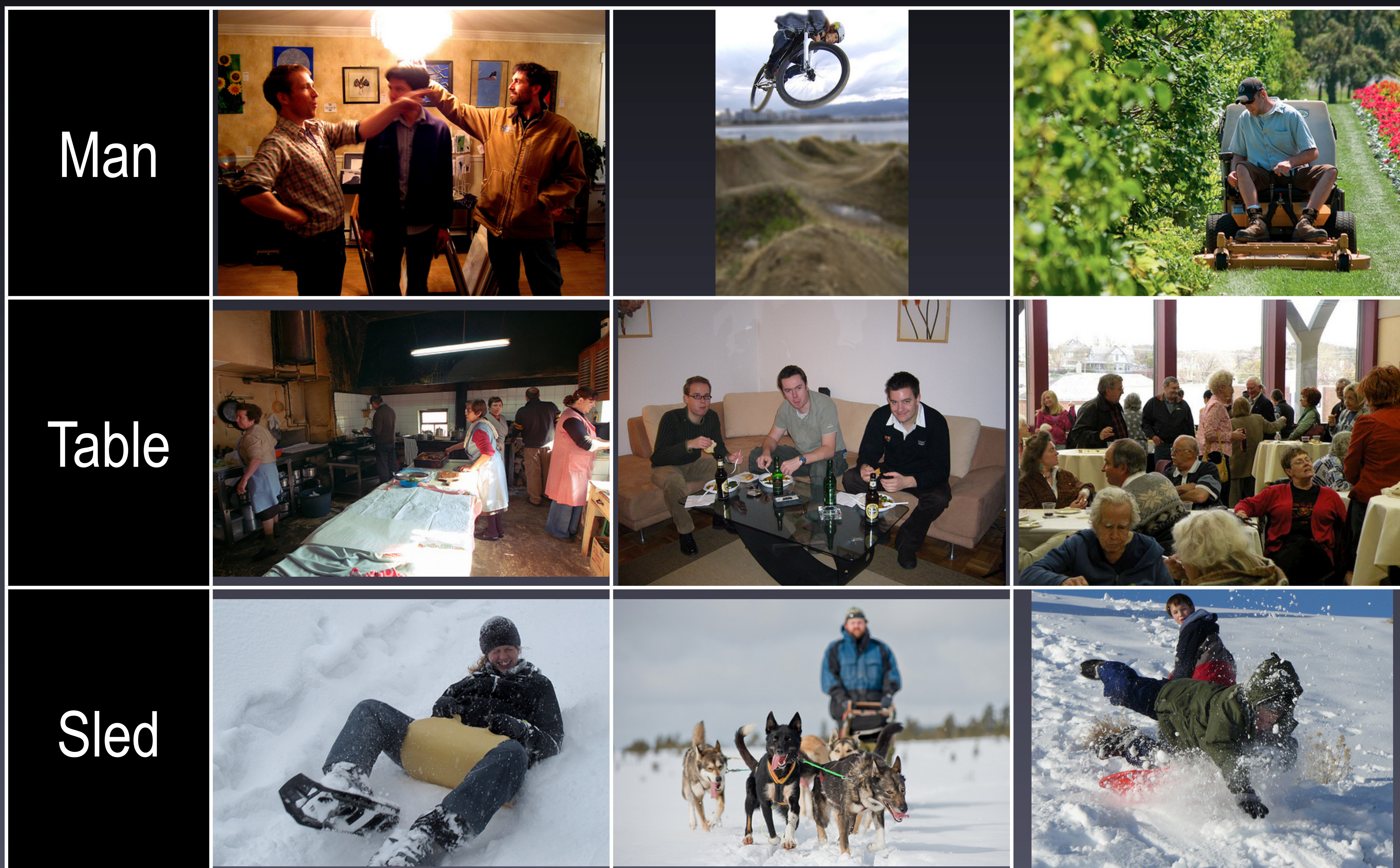
“Scene” words

“Scene” words

Certain words inherently constrain the image better than others

“Scene” words

Certain words inherently constrain the image better than others



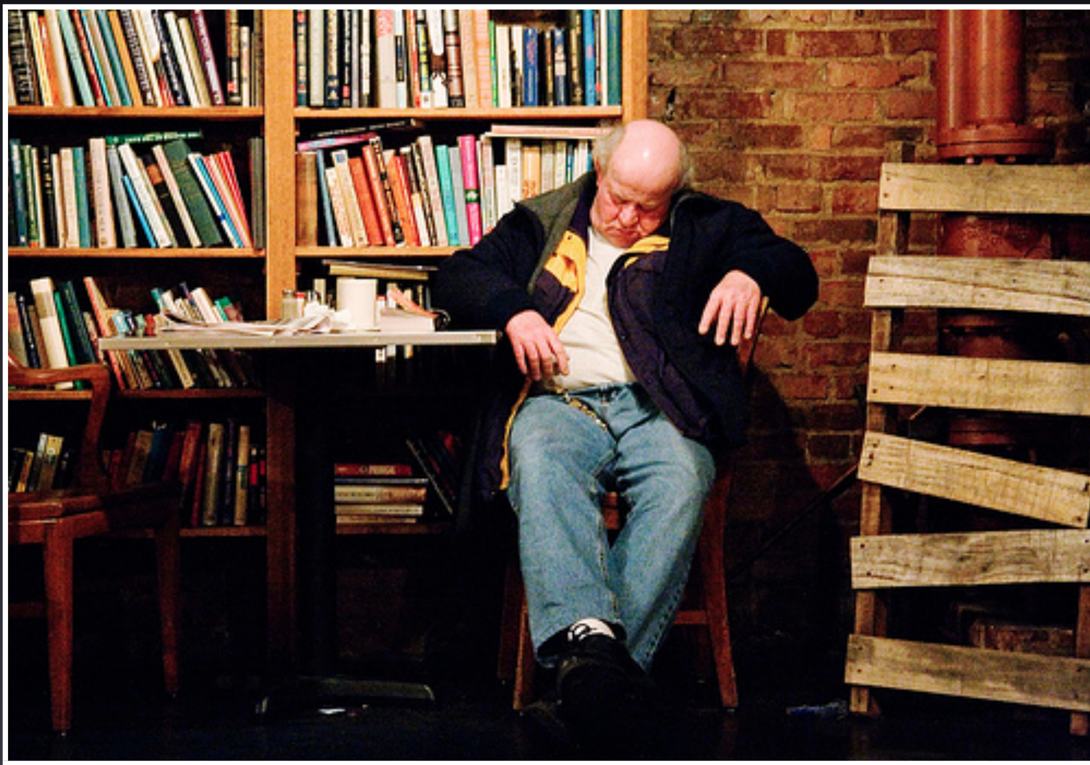
Different context: Sleeping

Different context: Sleeping

Is having **one representation** for a word appropriate?

Different context: Sleeping

Is having **one representation** for a word appropriate?



A man asleep in a chair in front of a full bookshelf

A woman in a red shirt is sleeping on a tan couch.

Non-visual words: “Two”

Non-visual words: “Two”

“Two” by itself is rather **meaningless a priori**

Non-visual words: "Two"

"Two" by itself is rather **meaningless a priori**

		
Two women laughing together at a table.	Two soccer players are going after the ball.	A well lit room, with three glasses on the table and two plates.

“Repetition” of concepts

“Repetition” of concepts

A little boy at a lake
watching a duck

Model responses	
Overall	**1.21**
Lake	0.89
Duck	0.17
Boy	0.13

"Repetition" of concepts

A little boy at a lake watching a duck

Model responses	
Overall	**1.21**
Lake	0.89
Duck	0.17
Boy	0.13

A man standing on a deck above a lake or river

Model responses	
Overall	**1.81**
Lake	0.89
River	0.70
Deck	0.17

In Conclusion

In Conclusion

Image description as an retrieval task simplifies cross-model comparison

In Conclusion

Image description as an retrieval task simplifies cross-model comparison

Important to consider models that will scale to increasingly large datasets

In Conclusion

Image description as an retrieval task simplifies cross-model comparison

Important to consider models that will scale to increasingly large datasets

In order to make progress, the linguistic issues of English need to be considered