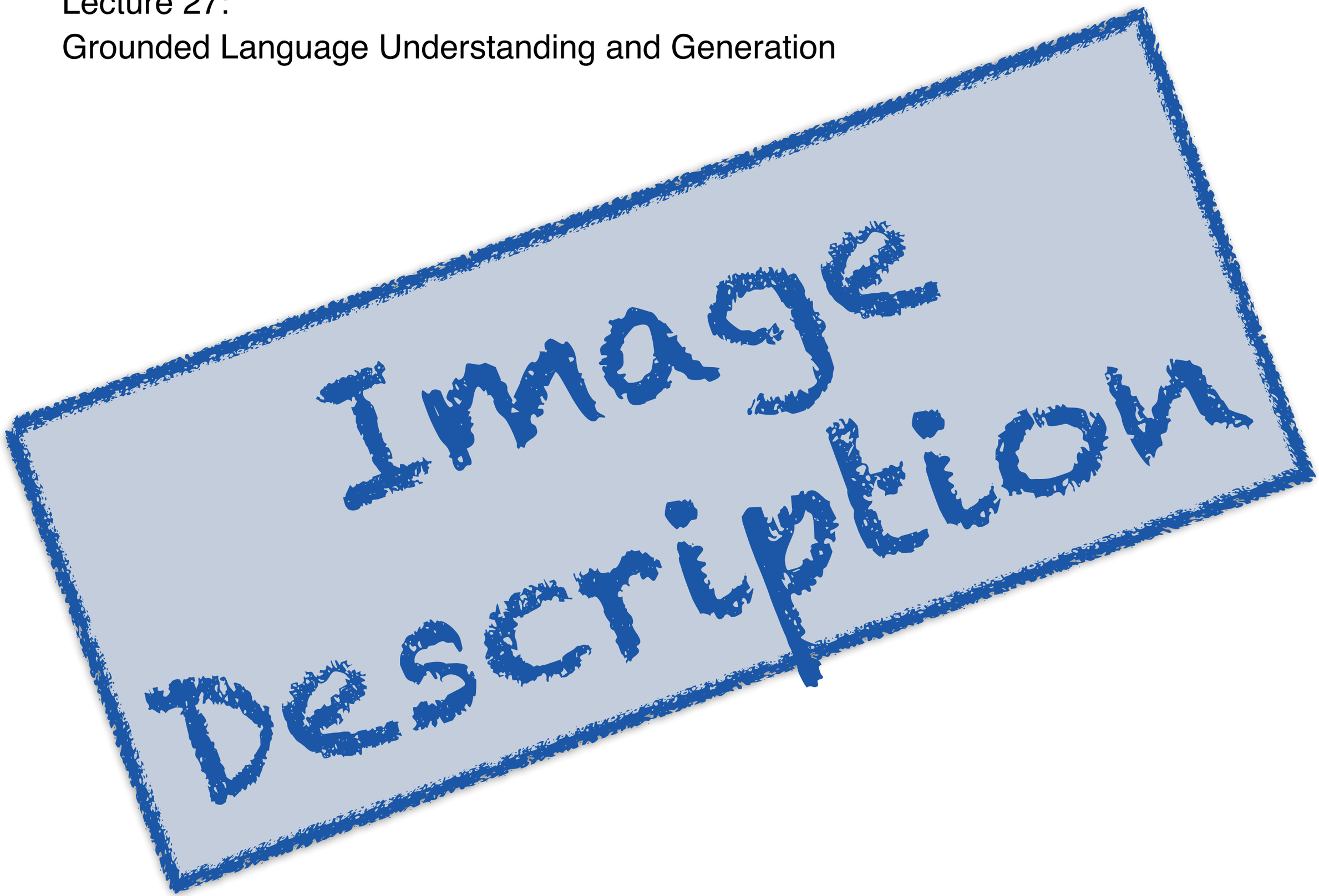
Lecture 27:
Grounded Language Understanding and Generation



So, we can go home, right?

No, not yet...

How do you get a computer to describe images?

Where do image description models break down?

How can we improve image description models?

How do you get a computer to describe images?

What task do you use
to develop and evaluate image description systems?

What model do you use
to score how well a sentence describes an image?

What data do you use
to learn this scoring function from?

What kind of image descriptions do we want to produce?

How would you describe this image?

A boy in a yellow uniform carrying a football is blocking another boy in a blue uniform.

yes

perhaps

Two boys are playing

A dog is running on the beach.

no

How would you describe this image?

Jake tackled Kevin really hard.

perhaps

Last Sunday's game was really rough.

probably not

Image descriptions...

... should describe the depicted
entities, events, scenes

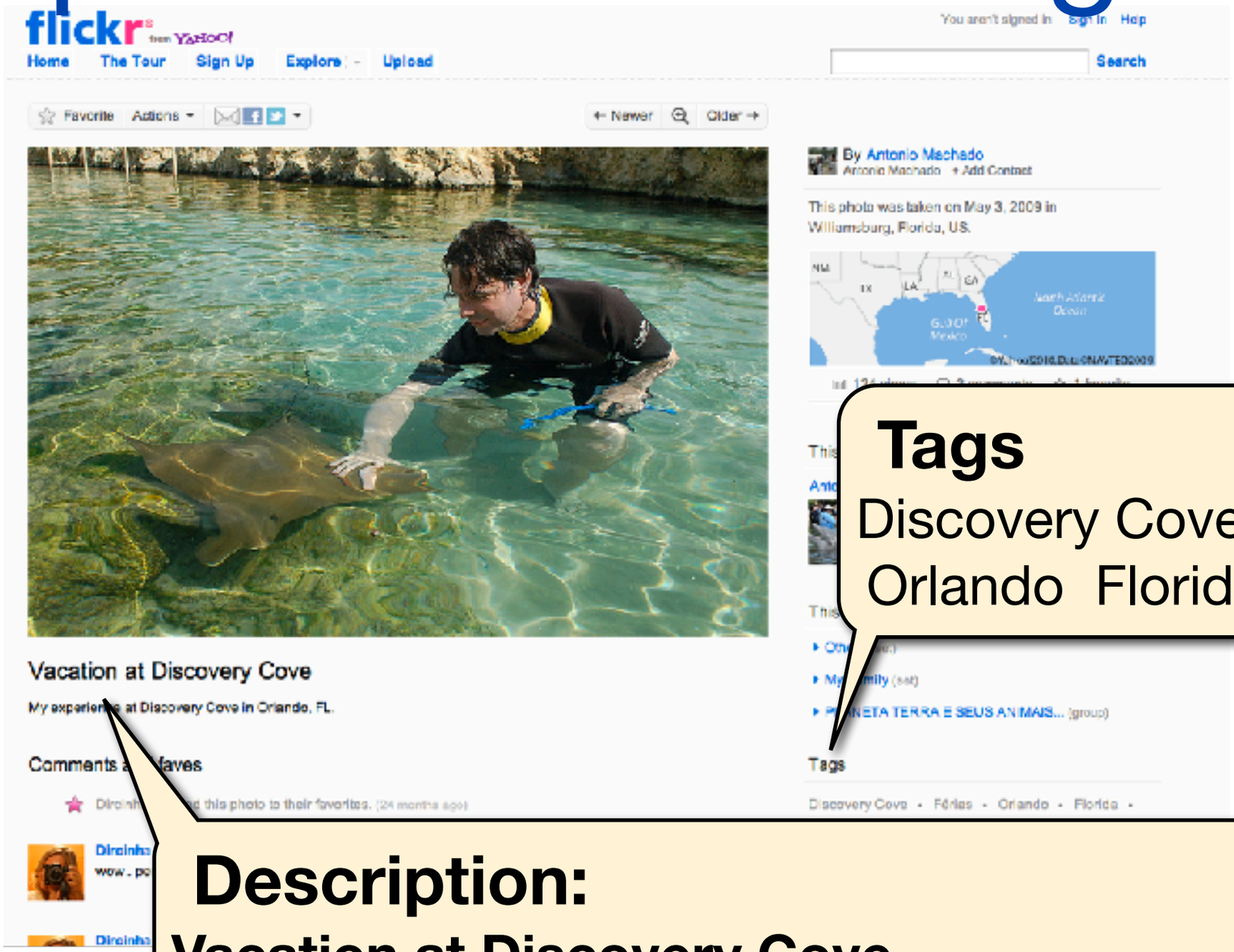
... should only describe
what can be seen from the image

... may differ in the amount of detail
provided

Image Description

Data

On photo-sharing sites, people describe images...



Tags

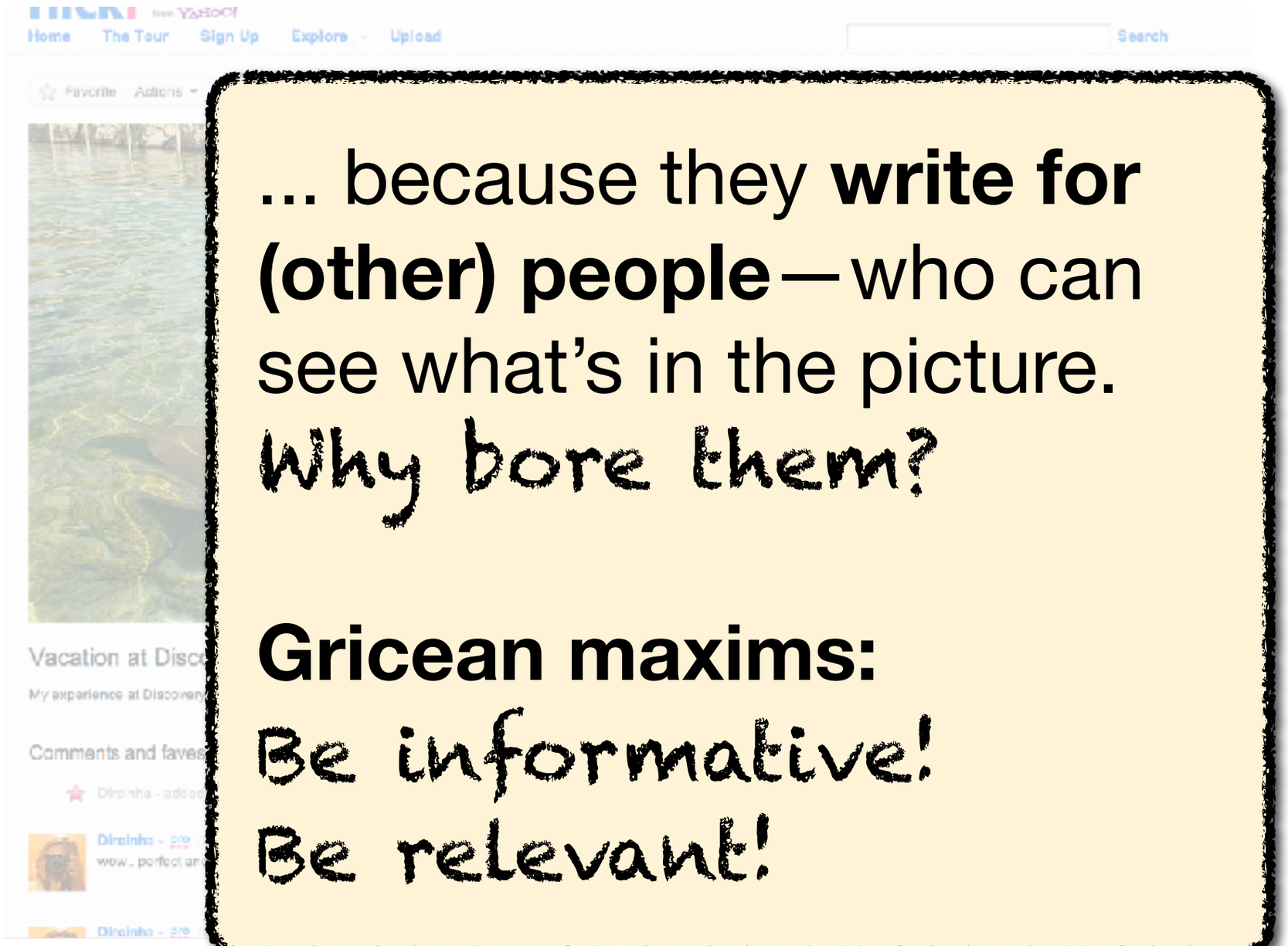
Discovery Cove Férias
Orlando Florida USA EUA

Description:

Vacation at Discovery Cove

My experience at Discovery Cove in Orlando,

… but they don't provide conceptual descriptions…



… because they **write for (other) people**—who can see what's in the picture.

Why bore them?

Gricean maxims:

Be informative!

Be relevant!

Image description with Amazon Mechanical Turk

Instructions:

Describe the objects and actions;

Use adjectives;

Be brief

5 captions per image

Illinois Flickr8k/30k data sets

A goalie in a hockey game dives to catch a puck as the opposing team charges towards the goal.
The white team hits the puck, but the goalie from the purple team makes the save.
Picture of hockey team while goal is being scored.
Two teams of hockey players playing a game.
A hockey game is going on.

A group of people are getting fountain drinks at a convenience store.
Several adults are filling their cups and a drink machine.
Two guys getting a drink at a store counter.
Two boys in front of a soda machine.
People get their slushies.

32k images of people (and dogs) from Flickr with 5 crowdsourced captions

Rashtchian et al. 2010, Hodosh et al. 2013, Young et al. 2014

Image Description

Tasks

Generation-Based Image Description

A little girl is enjoying the swings

Let the computer produce a sentence.

What should we say about an image?
How do we produce a grammatical sentence?
How do we evaluate generated descriptions?

Ranking-Based Image Description

A little girl is enjoying the swings

Two boys are playing football.

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Let the computer rank a pool of captions.

Evaluation is straightforward:
One caption in the pool was written for the image

Ranking-Based Image Search

A little girl is enjoying the swings

Two boy

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Image Description

Models

Image description models: The affinity function *f*(I,S)

As a probability: *f*(I,S) = *P*(S | I)

Generate S conditioned on I

As similarity/distance: *f*(I,S) = *sim*(I,S)

Map I and S to a common (vector) space

Find the closest S for each I (or vice versa)

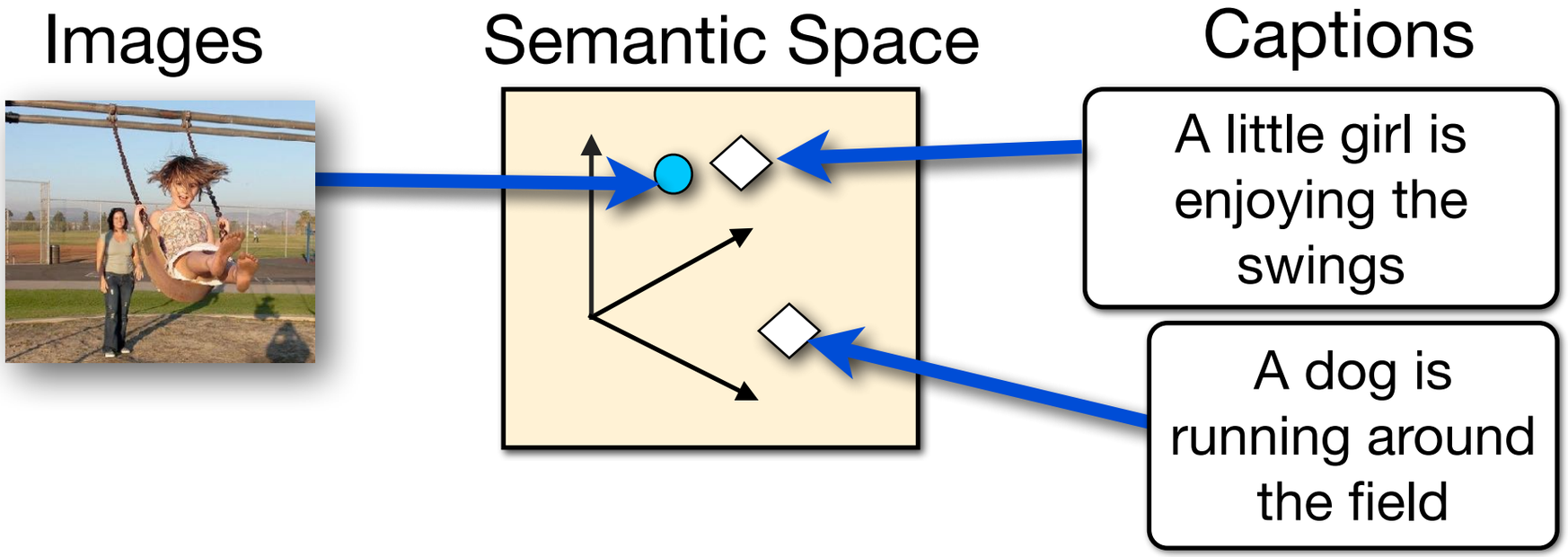
f(S,I) may or may not be mediated by an explicit (symbolic) semantic representation of S and I

Generation-Based Image Description

Earlier approaches:
Using traditional NLG techniques
(templates, grammars) with explicit detectors

Current approaches:
Using recurrent neural nets (LSTMs)
as language models, with deep-learning based
image features (possibly with explicit detectors)

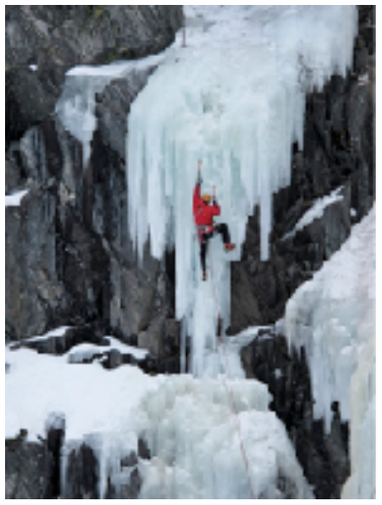
Mapping images and sentences to a semantic vector space



Hodosh, Young, Hockenmaier 2013:
Map images and sentences to a shared vector space, e.g. by (Kernel) Canonical Correlation Analysis, (K)CCA.
Rank sentences by their distance to the query image.

Image annotation examples

A girl wearing a yellow shirt and sunglasses smiles.



A man climbs up a sheer wall of ice.

A child jumping on a tennis court.

Basketball players in action.

Hodosh, Young, Hockenmaier 2013
No object/scene detectors,
No neural nets/deep learning,
just pyramid kernels over low-level visual features (SIFT, texture, color)

Do these models actually “understand” images (or language)?

Why does image description work as well (or as poorly) as it does?

Where do image description models break down?

Can we construct some tasks that allow us to analyze the behavior of various off-the-shelf image description models?

For more details: see Hodosh (2015)

Binary Forced-Choice Tasks

GOLD

A. There is a woman riding a bike down the road and she popped a wheelie.

DISTRACTOR

B. Two men in jeans and jackets are walking down a small road.

Task: Pick one of two captions for a given image.

Evaluation: How often does the system choose the gold caption over the distractor caption?

Binary Forced-Choice Tasks

GOLD

A. There is a woman riding a bike down the road and she popped a wheelie.

DISTRACTOR

B. Two men in jeans and jackets are walking down a small road.

In each task, the gold and distractor items **differ systematically**.

This allows us to **focus the evaluation on specific aspects** of image description.

What did our results show?

Vision-language models that were close to state of the art in 2015 **did not perform any better at choosing the correct caption** as a simple **bigram language model that had no access to the image**.

So, are we done?

Learning to associate images with simple sentences that describe them seems to be a **much easier task** than we might have thought a few years ago.

But we're fooling ourselves if we think this means that these systems 'understand' images or simple sentences.

So, are we done?

Current **evaluation metrics hide real weaknesses** of image description models.

This may matter as we move beyond image captioning to more complex task that require deeper image and language understanding.

For current tasks, simple BOW models may do as well as LSTMs.

Lots remains to be done!

Flickr30K Entities

[Plummer, Wang, Cervantes, Caicedo, Hockenmaier, Lazebnik, 2015]

Flickr30k Entities augments Flickr30k
with **267,000 bounding boxes**
and **244,000 coreference chains**
for all mentioned entities.

Annotation was done via crowdsourcing.

Flickr30K Entities

[Plummer, Wang, Cervantes, Caicedo, Hockenmaier, Lazebnik, 2015]

A man with pierced ears is wearing glasses and an orange hat.

A man with glasses is wearing a beer can crocheted hat.

A man with gauges and glasses is wearing a Blitz hat.

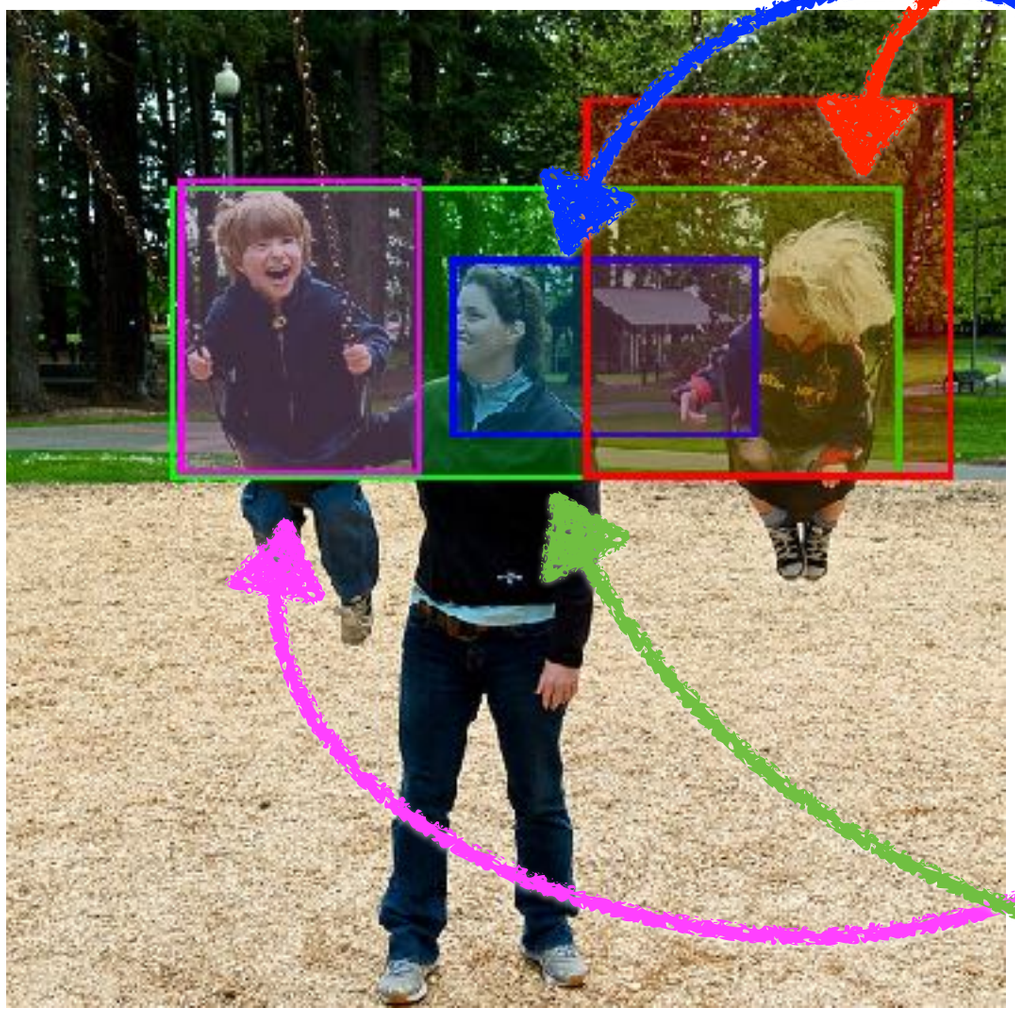
A man in an orange hat starring at something.

A man wears an orange hat and glasses.

Entity grounding

A woman pushes a child on a swing while another child looks on.

Entity grounding gone wrong



A woman pushes a child on a swing while another child looks on.

Using image descriptions to learn entailments

We can leverage the fact that we have multiple independent descriptions of each image (scene) to learn entailments e.g. (Young et al. 2013):

p(VP_1 | VP_2)

p(talk | engage in conversation) = 0.79
p(play tennis | swing racket) = 0.82
p(stand | wait for subway) = 0.58
p(stand | lean against building) = 0.53
p(shave | look in mirror) = 0.41
p(dig hole | use shovel) = 0.38
p(make face | stick out tongue) = 0.38


Grounded
Dialogue